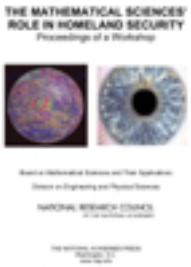
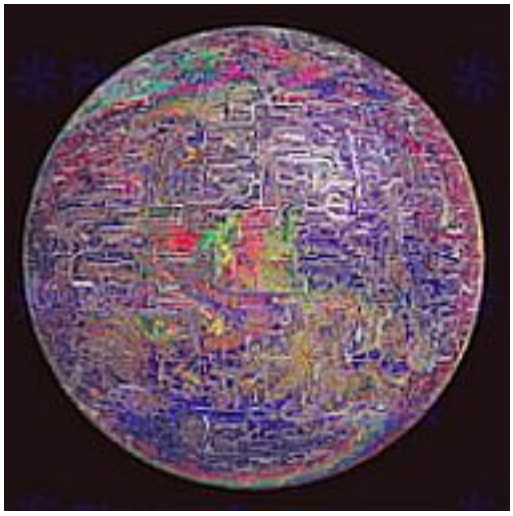


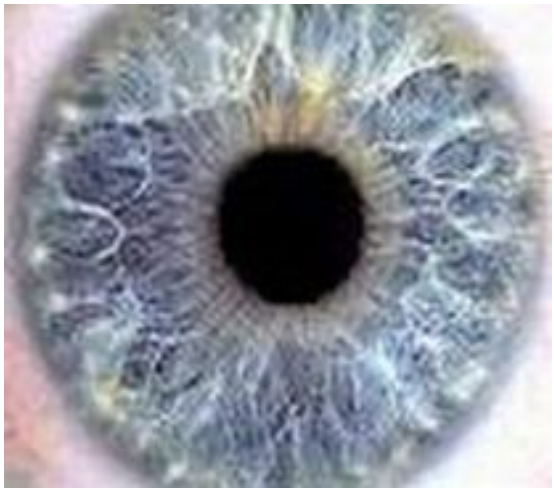
The Mathematical Sciences' Role in Homeland Security: Proceedings of a Workshop

National Research Council
ISBN: 0-309-53149-7, 576 pages, 8 1/2 x 11, (2004)

THE MATHEMATICAL SCIENCES' ROLE IN HOMELAND SECURITY

Proceedings of a Workshop





Board on Mathematical Sciences and Their Applications

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by the Army (Grant No. DAAD 190210066), the Air Force (PR# fq8671-0200746), the National Security Agency (Contract MDA904-02-01-0104), Microsoft Corporation (Award # 2327100), the National Science Foundation (Grant No. DMS-0215714), and the Navy (Grant No. N00014-02-1-0366). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-0935-0 (POD)
International Standard Book Number 0-309-53149-7 (PDF)

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu

Printed in the United States of America

COVER ILLUSTRATIONS: On the left is "circuit earth," a conceptual image of the earth as being as interconnected as a complex semiconductor circuit. While the worldwide interconnectivity is indisputable, scientists and engineers are a long way from being able to map, characterize, and analyze this complex network. That task will require many new developments in the mathematical sciences. The image on the right is of a human iris, the patterns of which are sometimes used in biometric security systems. Realization of the full potential of such biometric technologies also depends on future mathematical developments. The "circuit earth" illustration is courtesy of PhotoDisc, Inc., while the iris image is courtesy of John Daugman, University of Cambridge. Both images are reprinted with permission.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Bruce M. Alberts is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Wm. A. Wulf is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Bruce M. Alberts and Dr. Wm. A. Wulf are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

.

ACKNOWLEDGEMENTS

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council (NRC). The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

George Casella, University of Florida,
David Ferguson, The Boeing Company,
Valen Johnson, University of Michigan, and
Jon Kettenring, Telcordia Technologies.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. Responsibility for the final content of this CD report rests entirely with the authoring committee and the institution.

Preface

On April 26-27, 2002, the Board on Mathematical Sciences and Their Applications (BMSA) of the National Research Council organized a workshop on the role of the mathematical sciences in homeland security. The workshop was developed to illustrate contributions of mathematical sciences research to important areas of homeland security. The workshop drew over 100 researchers and focused on five major areas of research: data mining, detection and epidemiology of bioterrorist attacks, image analysis and voice recognition, communications and computer security, and data fusion.

The goal of this CD report is to help mathematical scientists and policy makers understand the connections between lines of research and important problems of national security. Included in this report are video presentations from most of the speakers at the workshop, as well as transcripts and summaries of the presentations, and any presentations materials used, such as power point slides. The presentations represent independent research efforts from academia, the private sector, and government agencies, and as such they provide a sampling rather than a complete examination of the interface between the mathematical sciences and the complex challenge of homeland security.

Each presenter identified numerous avenues of mathematical sciences research necessary for progress in homeland security. By design, none of the presentations provides a broad outline connecting the five major areas of research. However, common threads did emerge, such as the need for non-parametric methods, data visualization, understanding verification and validation of models and simulations, the need to deal with high-dimensional data and models, and the value of basing actions on sound mathematical analyses.

This proceedings represents the viewpoints of its authors only and should not be taken as a consensus report of the BMSA or of the National Research Council. We are grateful to the following individuals who reviewed this report: George Casella, University of Florida; David Ferguson, The Boeing Company; Valen Johnson, University of Michigan; and Jon Kettenring, Telcordia Technologies.

Funding for the workshop, its videotaping, and resulting report was provided by the National Science Foundation, Defense Advanced Research Projects Agency, Microsoft Corporation, Office of Naval Research, Air Force Office of Scientific Research, and the Army Research Office.

Board on Mathematical Sciences and Their Applications

THE MATHEMATICAL SCIENCES' ROLE IN HOMELAND SECURITY

National Research Council
Washington, D.C.

April 26-27, 2002

Welcome and Overview of Sessions, April 26

Peter Bickel, Chair, Board on Mathematical Sciences and Their Applications
Department of Statistics, University of California at Berkeley

Data Mining, Unsupervised Learning, and Pattern Recognition

James Schatz, *National Security Agency,* Session Chair
Introduction

Jerry Friedman, *Stanford University*
Role of Data Mining in Homeland Defense

Diane Lambert, *Bell Laboratories*
Statistical Detection from Communications Streams

Rakesh Agrawal, *IBM-Almaden*
Data Mining: Potentials and Challenges

Review and Discussion

Remarks, **Donald McClure,** *Brown University*

Remarks, **Werner Stuetzle,** *University of Washington*

Detection and Epidemiology of Bioterrorist Attacks

Claire Broome, *Centers for Disease Control and Prevention,* Session Chair
Introduction

Kenneth Kleinman, *Harvard Medical School*
Ambulatory Anthrax Surveillance: An Implemented System,
with Comments on Current Outstanding Needs

Stephen Eubank, *Los Alamos National Laboratory*
Mathematics of Epidemiological Simulations for Response Planning

Sally Blower, *University of California at Los Angeles*
Predicting the Unpredictable in an Age of Uncertainty

Review and Discussion

Remarks, **Simon Levin,** *Princeton University*

Remarks, **Arthur Reingold,** *University of California at Berkeley*

Image Analysis and Voice Recognition

Roberta Lenczowski, *Technical Directory (Bethesda), National Imagery and Mapping Agency,* Session Chair
Introduction

Jitendra Malik, *University of California at Berkeley*
Computational Vision

Ronald Coifman, *Yale University*
Mathematical Challenges for Real-Time Analysis of Imaging Data

Larry Rabiner, *AT&T Laboratories*
Challenges in Speech Recognition

Review and Discussion

Remarks, **David McLaughlin,** *New York University*

Remarks, **David Donoho,** *Stanford University*

Opening Remarks and Discussion, April 27

Communications and Computer Security

Howard Schmidt, *President's Critical Infrastructure Board,* Session Chair
Introduction

Dorothy Denning, *Georgetown University*
A Security Challenge: Return on Security Investment

Kevin McCurley, *IBM Almaden*
Talk omitted at speaker's request

David Wagner, *University of California at Berkeley*
A Few Open Problems in Computer Security

Review and Discussion

Remarks, **Andrew Odlyzko,** *University of Minnesota*

Remarks, **Michael Freedman,** *Microsoft Research*
Remarks omitted at speaker's request

Data Integration and Fusion

Alexander Levis, *U.S. Air Force,* Session Chair
Introduction

Tod Levitt, *IET, Inc.*
Reasoning About Rare Events

Kathryn Laskey, *George Mason University*
Knowledge Representation and Inference for Multisource Fusion

Valen Johnson, *Los Alamos National Laboratory*
A Hierarchical Model for Estimating the Reliability of Complex Systems

Review and Discussion

Remarks, **Arthur Dempster,** *Harvard University*

Remarks, **Alberto Grunbaum,** *University of California at Berkeley*

Funding for this workshop, its videotaping, and the resulting report was provided by the National Science Foundation, the Defense Advanced Research Projects Agency, Microsoft Corporation, the Office of Naval Research, the Air Force Office of Scientific Research, and the Army Research Office. The summaries of the presentations were derived from the transcripts by science writer Steven J. Marcus and BMSA staff.

Peter Bickel
"Opening Remarks," April 26

Transcript of Remarks
Summary of Remarks
Video Presentation

Peter Bickel's research spans a number of areas. In his work on semiparametric models (he is a co-author of the recent book *Efficient and Adaptive Estimation for Semiparametric Models*), he uses asymptotic theory to guide development and assessment of such models. He has also studied hidden Markov models, which are important in such diverse fields as speech recognition and molecular biology from the point of view of how well the method of maximum likelihood performs. Recently he has become involved in developing empirical statistical models for genomic sequences. He is a co-author of the text *Mathematical Statistics: Basic Ideas and Selected Topics*. He is past president of the Bernoulli Society and of the Institute of Mathematical Statistics, a MacArthur fellow, and a member of the American Academy of Arts and Sciences and of the National Academy of Sciences.

DR. BICKEL: Ladies and gentlemen, I think we had better get started. As usual we are running a little bit late, but that is fine.

Welcome. I am Peter Bickel. I am Chair of the Board on Mathematical Sciences and Applications of the National Research Council, a Board whose charge in part is to see what the mathematical sciences can do for the nation and of course, also, what the nation can do for the mathematical sciences.

When we proposed this workshop it was just 2 months after September 11, and we and most mathematical scientists in common with our fellow citizens want to help protect our country against terrorism.

We realized that although the mathematical sciences play a key role in most of the defenses that could be mounted against terrorism this was not appreciated by most of our fellow mathematical scientists, as well as many of the agencies working on homeland security.

So, we decided to try to do something about the situation, and with the generous assistance of the DOE, the National Science Foundation, National Security Agency, the Air Force Office of Scientific Research, Office of Naval Research and DARPA and Microsoft we have convened this workshop.

So, what is the element that makes the mathematical sciences have a central role? Preventing or minimizing the effect of terrorist activity requires learning as much as possible about future terrorist plans or about the nature of a current terrorist attack.

Mathematics, operations research statistics and computer science provide the fundamental tools for extracting relevant information from the flood of data of all types that our senses receive.

We want to know about people moving, money moving, materials moving and so on. This data may be in the form of numbers and text that is then subject to the techniques known under the names of data mining or unsupervised learning or pattern recognition.

That is the subject of this first session. The data may be in the form of voice messages or video images corrupted by noise, and it is then subject to mathematical techniques which extract the true sounds or images from the corrupted form in which it is received. That is largely the subject of Session Three.

Information bearing on what may happen often comes from several sources simultaneously, for instance, modern communications, zone images, intelligence data. Putting such data together is the subject of Session Five.

Terrorists can encrypt their communications, and we need to decrypt them. Conversely you want them not to know about our knowledge and actions. That is the subject of Session Four.

Finally in the domain of biothreats we may be faced not with isolated cases but with full-blown epidemics of unknown origin. Stopping these and tracing their origin as quickly as possible on the basis of information gleaned from the pathogens, geography, speed of transmission, etc., is the subject of Session Two.

Of course, there is a great deal of overlap between the subjects in these sessions. The synergy achieved by bringing these different groups of researchers together is also one of our goals.

We have been fortunate that a collection of distinguished speakers with expertise in each of these areas agreed to give expository talks, that a number of distinguished mathematical scientists not working in these areas have agreed to act as discussants, and that a number of distinguished officers of the federal agencies concerned with homeland security have agreed to chair these sessions.

As you see we have left plenty of time for discussion. We hope that these discussions will lead to

future research and contributions by the mathematical science community to these questions.

I encourage the various federal agency representatives, I am glad to see are in attendance to help point the way towards this future work.

I hope and expect to see two goals achieved, appreciation by mathematical scientists of the important contributions that they make to homeland security and appreciation by the federal agencies of the important role of the mathematical scientists in the campaign for homeland security.

I will now turn to the chair of the first session, Jim Schatz.

Welcome and Overview of Sessions

Peter Bickel

Dr. Bickel welcomed the participants to the workshop and spoke about how important it is for both the mathematics community and the security community to understand and appreciate the role that the mathematical sciences play in security. Mathematics, operations research, statistics, and computer science provide the fundamental tools for extracting relevant information from the flood of data of all types that our senses receive. The techniques to be discussed are used to search and sort data, stop and trace disease outbreaks, extract true data from a message complicated by noise, encrypt and decrypt messages, and pull together data from disparate sources.

Dr. Bickel mentioned how fortunate we were to have such a distinguished group of participants and said that he hoped the discussions would lead to future research and contributions by the mathematical science community to questions dealing with security.

James Schatz
”Introduction by Session Chair”

Transcript of Presentation
Summary of Presentation
Video Presentation

James Schatz is the chief of the Mathematics Research Group at the National Security Agency.

DR. SCHATZ: Thank you, Peter. We had a session in February down at the Rayburn Building in Washington to talk about homeland security that the American Mathematical Society sponsored and I thought I would like to just as an introduction to our first session here give some of the remarks we made down there which I think are relevant here.

It is a wonderful privilege to be here today. In these brief remarks I would like to describe the critical role that mathematics plays at the National Security Agency and explain some of the immediate tangible connections between the technical health of mathematics in the United States and our national security.

As you may know already the National Security Agency is the largest employer of mathematicians in the world. Our internal mathematics community is a dynamic professional group that encompasses full-time agency employees, three world-class research centers at the Institute for Defense Analyses that work exclusively for NSA and a network of hundreds of fully cleared academic consultants from our top universities.

As the Chief of the Mathematics Research Group at NSA, and the executive of our mathematics hiring process I

have been the agency's primary connection to the greater US mathematics community for the past 7 years.

The support we have received from the mathematicians in our country has been phenomenal. Our concern for the technical health of mathematics in our country is paramount.

Perhaps the most obvious connection between mathematics and intelligence is the science of cryptology. This breaks down into two disciplines, cryptography, the making of codes and cryptanalysis, the breaking of codes.

All modern methods of encryption are based on very sophisticated mathematical ideas. Coping with complex encryption algorithms requires at the outset a working knowledge of the most advanced mathematics being taught at our leading universities and at the higher levels a command of the latest ideas at the frontiers of research.

Beyond cryptology the information age that is now upon us has opened up a wealth of new areas for pure and applied mathematics research, areas of research that are directly related to the mission of the National Security Agency.

While advances in telecommunications science and computer science have produced the engines of the information age, that is the ability to move massive

amounts of digital data around the world in seconds, any attempt to analyze this extraordinary volume of data to extract patterns, to predict behavior of the system or recognize anomalies quickly gives rise to profound new mathematical problems.

If you could visit the National Security Agency on a typical work day you would see many, many groups of mathematicians engaged in lively discussions at blackboards, teaching and attending classified seminars on the latest advances in cryptologic mathematics, arguing, exchanging and analyzing wild new ideas, mentoring young talent and most importantly pooling their knowledge to attack the most challenging technical problems ever seen in the agency's history.

You would hear conversations on number theory, abstract algebra, probability theories, statistics, combinatorics, coding theory, graph theory, logic and Fourier analysis.

It would probably be hard to imagine that out of this chaotic flurry of activity and professional camaraderie anything useful could emerge.

However, there is a serious sense of urgency underlying every project, and you would soon realize that the mathematicians of NSA are relentless in their pursuit

of tangible, practical solutions that deliver critical intelligence to our nation's leadership.

The mathematicians of NSA, the Institute for Defense Analyses and our academic partners are the fighter pilots in a way that takes place in the information and knowledge layer of cyberspace.

As Americans you would be very proud of their achievements in the war on terrorism. The National Security Agency's need for mathematicians is extreme right now.

Although we hire approximately 50 highly qualified mathematicians per year, we actually require more than that, but the talent pool will not support more.

Over 60 percent of our hires have a doctorate in mathematics, about 20 percent a master's and 20 percent a bachelor's degree.

We are very proud of the fact that 40 percent of our mathematics hires are women and that 15 percent are from under represented minority groups.

Of course, the agency depends solely on the greater US mathematics community to educate each new generation of students, but we, also, depend on the professors at universities across the country to advance the state of mathematics research.

If the US math community is not healthy the National Security Agency is not healthy, and I always like to use an occasion like this to thank everybody here for all they have done for math in this country because our agency benefits so greatly.

Okay, this first session here is on data mining, unsupervised learning and pattern recognition. This is a very exciting, very active area of research for my office and for the agency at large.

We attended just recently the Siam Conference on Data Mining about, when was that, just about a week ago here in Washington, and had a great presence there.

It is a wonderful topic. There is absolutely nothing going on in this conference that isn't immediately relevant to NSA and homeland security for us, and this first topic is an area of research that I think we had a bit of a head start on. We have been out there doing this for a few years, but there is a whole lot to learn. It is a young science.

So, let me without further ado bring up our first speaker for this session, and that is Professor Jerry Friedman from Stanford University.

Introduction by Session Chair

James Schatz

Perhaps the most obvious connection between mathematics and intelligence is the science of cryptology. This breaks down into two disciplines—cryptography, the making of codes, and cryptanalysis, the breaking of codes. All modern methods of encryption are based on very sophisticated mathematical ideas. Coping with complex encryption algorithms requires at the outset a working knowledge of the most advanced mathematics being taught at our leading universities and at the higher levels a command of the latest ideas at the frontiers of research.

Beyond cryptology the information age that is now upon us has opened up a wealth of new areas for pure and applied mathematics research. While advances in telecommunications science and computer science have produced the engines of the information age—that is, the ability to move massive amounts of digital data around the world in seconds—any attempt to analyze this extraordinary volume of data to extract patterns, to predict behavior of the system, or to recognize anomalies quickly gives rise to profound new mathematical problems.

Although the National Security Agency hires approximately 50 highly qualified mathematicians per year, it actually requires more than that, but the talent pool will not support more. Of course, the agency depends on the greater U.S. mathematics community to educate each new generation of students, and it also depends on the professors at universities across the country to advance the state of mathematics research. If the U.S. math community is not healthy, the National Security Agency is not healthy.

Jerry Friedman
"Role of Data Mining in Homeland Defense"

Transcript of Presentation
Summary of Presentation
PDF Slides
Video Presentation

Jerry Friedman is a professor in the Statistics Department at Stanford University and at the Stanford Linear Accelerator Center.

PROF. FRIEDMAN: Jim asked me to talk about the role of data mining in homeland defense, and so in a weak moment I agreed to do so, and in looking it over I discovered that unlike any other areas of the mathematical sciences there is a well-perceived need among decision makers for data mining on homeland defense.

So, I have a few examples. Here is an excerpt from a recent speech by Vice President Cheney, and he said, "Another objective of homeland defense is to find connections with huge volumes of seemingly disparate information. This can only be done with computers and only then with the very latest in data linkage analysis."

So, that was in a recent speech by Vice President Cheney. Here is a slightly higher decision maker. This is from the President's Office on Homeland Security Presidential Directive 2, and it is a section on the use of the best opportunities for data sharing and enforcement efforts. It says, "Recommend ways in which existing government databases can best be utilized to maximize the ability of the government to identify and locate and apprehend terrorists in the United States. The utility of advanced data-mining software should be addressed."

Here is the trade journal, the Journal of Homeland Security. Technologies such as data mining, as

well as regular statistical analysis can be used at the back end of biodefense communication networks to help analyze implied signatures for covert terrorist events.

In addition data mining applications might look for financial transactions occurring as terrorists prepare for their attacks.

Okay, here is the popular press. Computer databases are designed to become a prime component of homeland defense. Once the databases merge the really interesting software kicks in, data mining programs that can dig up needles in gargantuan haystacks.

Okay, as many of you know DARPA has set up an information awareness office and they were charged among other things to look into biometric speech recognition and machine translation tools, data sharing among the agencies for quick decisions and knowledge discovery technology; knowledge discovery is another code word for data mining, that uncovers and displays links among people, content and topics.

Here is my favorite. It is not quite germane but this is a comment by Peter W. Hoover, not Peter Hoover the statistician but the engineer from MIT, and he said that in this new era of terrorism it will be their sons versus our silicon, a rather startling point, but I think part of our

silicon will be data mining algorithms running our computers, and the data mining bureau, Interpol and the DARCY(?) coined the phrase MacInt for machine intelligence. It sounds like something that might come from either Apple computer or a hamburger chain, but we need a MacInt or machine intelligence capabilities to provide cuing or early warning from data pertaining to national security.

So, there doesn't seem to be a need to convince decision makers that data mining is relevant to national security issues. Many think it is central for national security issues.

So, I think the problem here is not in convincing decision makers of the need for data mining technology but to live up to the current expectations of its capabilities, and that I think is going to be a big job.

Now, what is data mining? Well, data mining is about data and about mining. Okay, let us talk about data. What are the kinds of data we are going to see in homeland security applications? Well, there would be real-time high-volume data streams, massive amounts of archived data, distributed data that may or may not be centrally warehoused; hopefully it will be centrally warehoused, but you can't centrally warehouse it all and of course many different data types that will have to be merged and

analyzed simultaneously like signals, text, images, transactional data, streaming media, computer-to-computer traffic, web data and of course biometric data.

So, that is the kind of data we will be seeing. I will have more to say about that in a moment. Where is that data going to come from? Well, there is a big effort as most of you know in the government for data information sharing. The buzz word is to tear down the information stovepipes between the various agencies who apparently guard rather carefully all their own databases and information. Now, the big effort is to merge them.

So, here is a quote from an unnamed official in the Global Security Newswire. It says, "There are many community-wide data mining architectures that are being looked at to allow information sharing among intelligence and law enforcement communities," and a big example of that is going on right now. It is the merging of the CIA and the FBI databases for forensics, and that will be made available to state and federal agencies that they can query that and this year the Federal Government is spending $155 million for this effort of tearing down information stovepipes and next year they have requested over $700 million for that.

So, that is where the data is going to come from. Here is another example again that I like. I got this from www.privacy.org, another example of where the data is going to come from.

Federal aviation authorities and technology companies will soon begin testing a vast air security screening system linking every reservation system in the US to private and government databases designed to instantly pull together every passenger's travel history and living arrangements, plus a wealth of other personal and demographic information.

The network reviews data mining and prediction software to profile passenger activity and intuit obscure clues about potential threats even before the scheduled date of the flight. That is a tall order, and now, www.privacy.org doesn't necessarily consider this a positive development, but it certainly is an example of the kind of data that we are going to be seeing.

Now, what are the applications? Again, when you talk about data mining the first thing you have to think about is where is the data going to come from and second, what do you want to do with it; what are the applications?

Okay, well, the applications read like the titles of the sessions of this conference, obviously bioterrorism,

detect early outbreaks of diseases and look for characteristic signatures of bioterrorism versus some other natural outbreak.

Cybersecurity is a big area that of course is being studied on the commercial front quite a lot, intrusion detection, protection of physical infrastructure with surveillance systems. Lots of data are going to come from security cameras everywhere and/or transfer the infrastructure like transportation, water, energy, food supplies. Of course, document classification, as was mentioned before we have to recognize coded messages.

There are some more applications. Again, they are all going to be discussed here and mathematical sciences are relevant to all of them.

Behavior recognition, travel behavior, financial behavior, immigration behavior, automatic screening of people, mail, cargo, and of course forensics, post-apprehension. If there is an attack you want to catch the perpetrators and all those who planned it so they can't plan another one.

Now, what all of these problems have in common are they are needle in haystack problems, namely, you are looking for a very, very small signal in a very, very big background.

Now, what is kind of the current state of the art in data mining? Well, to talk about that in 20 minutes is kind of hard, but let me give you just a very rough overview. The most successful applications of data mining have been commercial applications which you can basically characterize as behavior recognition using transactional data.

Classic example are fraud detection, fraudulent phone calls, fraudulent credit card usage, tax evaders, those are classic applications of the commercial.

Churn analysis, where you look for customers who are likely to quit your services or your competitor,a nd this is very important among competitors who offer virtually the same service like telephone companies. They do a lot of that. Purchasing patterns, recommender systems have been reasonably successful and other ways of trying to detect consumer purchasing patterns so you can promote very special things and hopefully make more money. Direct marketing, who do I send my brochures to, and in general customer relation management have many texts on data mining, actually have commercial customer relation management in their title. That has been the main focus and the main success so far in data mining, and of course, stretching a little bit of course document search over

large corpora like web and library search engines have obviously been very successful.

Now, homeland security problems if you look at them are very similar qualitatively to many of the commercial applications. They are of the same kind but in my view at least are different in degree because I think they are much harder and here are the reasons why I think they are much harder.

First of all the haystacks are bigger. We are looking for needles in haystacks. The government database and although data mining has been applied to some very large commercial databases the government databases are going to be even larger and they are going to be more diffuse in the kinds of information, and the needles are going to be smaller. I mean hopefully there are fewer terrorist than people who are trying to make fraudulent phone calls or trying to cheat on their taxes, and so the needles are going to be really small, and the stakes are much higher.

If your data mining algorithm on a consumer database doesn't quite get all the customers who might be attracted to your product, well, your client is going to realize slightly less profit and won't even probably know it.

Here if we make a mistake the mistakes are highly visible and can lead to catastrophes, and so the stakes are very much higher. We need fast response. We can't take time to print up our brochures. We don't have any luxury of time. We have to worry about what the machine learning people call concept drift, namely, the underlying joint distribution of the variables if you like is changing with time for both the signal because the terrorists aren't stupid and the background. People's shopping patterns are likely to change or they are changing because of our intervention in homeland security, and we have an intelligent adversary and I think that is one of the things that makes these kinds of problems unique, not that consumers aren't intelligent but they generally don't try to evade you.

So, the data that we are trying to see I characterize as maliciously observational data. First of all it is observational. You can't design experiments, and all data mining is applied to observational data, but more importantly they know we are watching. They are going to try to disguise their transactions and as they realize we are detecting them they are going to change the signatures in real time as we detect them, and they can also jam us with deliberate decoys. They can go through the motions of

planning an attack, do all the things we know how they would and then not attack and do that repeatedly just like car alarms that you ignore now until finally you know you say, "Well, this isn't working. We keep predicting attack and it is not happening. Obviously we are having many errors of a false positive nature. So, something is wrong with our system," and maybe nothing is wrong with our system.

So, this maliciously observational data I think is a new twist that certainly existing data mining algorithms may be valuable to tell us about fraudulent phone calls but I think that is a really new twist to this kind of thing.

So, what are the properties that our methodology is going to have to have? It is going to have to be fast in both training, that is in developing the data mining detection algorithms and it is going to also have to be fast in predicting and that will leave out a lot of traditional data mining because the computing requirements are going to be intense. They are going to have to be real time, of course, timely, accurate, robust and adaptive because everything will be changing with time.

What are some of the other attributes of the methodology that we are going to have to have? It is going

to have to be dynamically updatable. Again, lots of data mining algorithms you run them on the data; if you change the data you are going to have to run them on the data again, and we can't tolerate that.

You are going to have to model complex changing, surely non-linear relationships and those relationships as I mentioned before will be changing both of the signal and of the background. To the extent we can do any inference at all, it is going to have to be assumption free and I think most important is we will have to effectively incorporate the new knowledge which in most traditional data mining we haven't done that with the possible exception of the Bayesian-belief networks which may actually if we can ever get them fast enough might really have a lot to offer here.

In my view we are not there yet. Now, of course, if all you have is a hammer every problem looks like a nail, and I think that all of us working on homeland defense issues will bring their own favorite set of tools to these problems.

In the data mining area popular techniques are decision trees, rule induction, support vector machines and neural networks and a lot of other things, and it is not clear at least in my mind which if any of these will provide success. They all have advantages and

disadvantages, and none of them have enough of the advantages that are going to be required for the kinds of accuracy, speed and other properties that will be needed, and all of them if they are used will need a big increase in their performance both in their computational performance and in their statistical performance as well.

I don't think all is lost. This allows us to do research into new methodology and often new good things can happen. First of all as I said, we will have new kinds of data. The data itself won't be new kinds but the fact that we have to merge it and analyze it simultaneously I think is fairly new, the fact that a single database will have transactional data, audio, video and text data all in it and we will have to treat it all simultaneously. Nearly all of everything we measure will be irrelevant. The signatures,for instance the reservation database, the travel database, of all the things that they are going to pull out of there about the nature of their trip very few of them will be really relevant.

Very few attributes will be relevant to detect the terrorist attack and the performance requirements as I mentioned before both computational and statistical are very high, at best severely challenging or probably exceeding the present state of the art.

So, there are huge expectations. A lot of these databases are being merged and a lot of money spent on the expectation that the data mining algorithms will find things in them, and that is their raison d'etre for merging all of this data, spending all that money, and so I think that we are going to have to perform and figure out ways to increase computational and statistical performance so that we really do something with these very large and complex databases when we see them.

However, I think that as is often the case more time requirements often lead to scientific advances and in World War II there were a ton of them in the mathematical area. Of course, it played a huge role and many aspects of World War II led also to advances in mathematics that led to the invention and advances in computers which of course have changed our lives in many areas other than code breaking, and advances are certainly needed in data mining if we are going to successfully address the expectations of what data mining can do in homeland security, and if we are successful, even remotely so we will also impact a wide variety of other areas which are not quite so difficult.

So, I think it is a wonderful opportunity and of course very important.

Thank you.

(Applause.)

DR. SCHATZ: Thank you, Jerry.

Technical note. We requested that the speakers speak here at the podium and if assistance is needed with the changing of transparencies, etc., we have an assistant who will do that. For the camera though they would like you up here.

Just a couple of points I wanted to make about our last talk. A wonderful talk. It is very much to the point of my daily life right now. The government problems are much bigger typically than what you are seeing anywhere else, and the needles are very tiny.

We talk about volume, variety and velocity where I come from, and one of the good points, also, made there was the robustness issue. In certain databases the data is very clean. It is very uniform. Certainly for the problems we deal with we have fragmented data, very badly garbled data, data from different sources that we are trying to pull together and the scalability of algorithms is absolutely a critical point to be made in all this.

Certainly other fine points we have here are trees, rule based algorithms, support vector machines, neural nets; we are using all those techniques right now. They all are useful. They all help to a degree. They are

all not enough. So, there is really a wonderful opportunity here I think to bring in some advanced statistics and mathematics into these problems. Also, a wonderful point about the fact that wartime has certainly in the past inspired some amazing advances in science, the Manhattan Project, of course being one and the cryptography effort that really began, modern cryptography really began of course during World War II in the serious sense, and I think I agree with Jerry's point there exactly.

I think if we look at this war on terrorism, the big advance is going to be in information processing and data extraction because that is the edge we need at this point.

Okay, well, we are a little ahead of schedule which is fine, and I think Diane is ready. So, our next speaker is Diane Lambert form Bell Labs.

The Role of Data Mining in Homeland Defense

Jerry Friedman

The problem is not in convincing decision makers of the need for data-mining technology but in living up to their expectations. We must be able to handle a variety of data formats, including real-time, high-volume data streams, massive amounts of archived data, and distributed data that may not be centrally warehoused. We also must be able to handle various data types, including signals, text, images, transactional data, streaming media, computer-to-computer traffic, Web data, and biometric data. In addition, data-mining methodologies must operate in real-time and be accurate, robust, and adaptive.

In an effort to bring compatibility to various government databases, the government has begun work on a project to merge the CIA and FBI forensic databases. The product will be made available to other agencies, federal and state, for querying. In 2002, the federal government invested $155 million in this effort, and it plans to increase the 2003 budget to more than $700 million.

Data mining has several applications to homeland security: air travel, characteristic signatures of bioterrorism versus natural outbreaks, cybersecurity, intrusion detection, surveillance, and travel, financial, or immigration behavior.

Currently, the federal government plans to create a vast air-security screening system linking every reservation network in the United States to private and government databases. The system is designed to instantly pull together every passenger's travel history and living arrangements, plus a wealth of other personal software to profile passenger activity and intuit obscure clues about potential threats.

Diane Lambert
"Statistical Detection from Communications Streams"

Transcript of Presentation
Summary of Presentation
PDF Slides
Video Presentation

Diane Lambert is head of the Statistics Research Department of Bell Laboratories, Lucent Technologies, and a Bell Labs fellow. She is also a fellow of the American Statistical Association and the Institute of Mathematical Statistics and has served on NSF and National Academy of Science panels and as editor of the *Journal of the American Statistical Association*. She has also served on ACM and SIAM program committees for conferences on data mining. Her research for the past several years has focused on ways to analyze and model the kinds of massive, complex data streams that arise in telecommunications.

DR. LAMBERT: It is actually a little overwhelming to be here since I don't work in security and so trying to guess what someone's problems are when you don't work in the area always seems a little bit dangerous.

So, what I am going to stick with is communications data. So, one thing that we have a lot of at Bell Labs is communications data, and you can imagine that communications data has a lot of information in it that would be of interest to people who are trying to track what is going on.

So, there are all kinds of communications records that are kept now about all different kinds of communications, much more than maybe the average person is aware of. Every call of course has to be billed. So, for a very, very long time you have had a record of every call that is placed. You know who placed the call, when the call was placed, who was called and if it is a mobile phone you know where they were when the call was placed, at least generally because you know which cell site handled the call.

You know how they paid for it. There is just a lot of information in this record. Okay, now, there is actually a lot of information about online communications, too.

If you request a file, if you want to download a file, then in a log file there is a record of that transfer. It says, "Who" where here Who depends on whether you allow cookies to be used or not and Who might be just an IP address. It is not too meaningful or it may actually be meaningful, but you know when and what was downloaded, which images and everything else.

Going beyond that a bit online chatrooms the way they are structured there is actually a record kept in a log file of everything that is posted. That tells you who did the posting where again who might be an IP address. Who might partially a user name that somebody uses consistently. Who might be an actual person in some cases.

You know which room they were posting in. You know when they posted and what they posted. All of this is kept in the file.

So, there is this wealth of information about what people are doing both generally and in particular. You know can you get any useful information out of that. If you want to learn something about users and about people or about groups, you know, what can you actually do with this data?

Now, to be believed that if you have data then all you have to do is think about it enough and you will

be able to extract information from it, and I am not sure we have made a lot of progress here yet, but let me tell you what we have done.

Next slide, please?

The first question is what kind of information do we want, and you know, it is always a little hard to do particular examples because in any particular example you can say, "I don't want that information," but we will start with the easy one which is the easiest flows of all are really just call records.

So, they are records of the calls. So, this for example, is one user. Now, you don't get this user or this caller's, in this case it is a cell phone, you don't get the call phone records all nicely delineated like this. They come all mixed and interleaved together, but for the moment imagine you have got a process which isn't too hard to imagine that sort of filters them all for use. You get them all together. Maybe you have just pulled them all off, and you look at them.

Now, what you can see here is each one of these vertical lines represents a call. Longer calls are represented by longer lines. You have time going on the horizontal axis, and then you get sort of categorical information by color.

Here it is terminating number, and you get things like which calls are incoming, which calls are outgoing, and then from that you can think of describing what the typical behavior of that caller is, and then you can also think of how do I find something that is atypical; how do I detect there is a change in the pattern?

Now, if we look at it you can see that all of a sudden on the right there is all this purple stuff that starts happening that has longer lines and there is more of it.

So, it looks in fact if you look at it as if there is a change, but actually that is a month's worth of data in the purple part over there, and you don't want to wait a month in order to say, "Oh, yes, there is something interesting going on."

Also, you don't get to look at all the past data when you want to make a decision about this one. So, as soon as it turns purple you would like to say something interesting, and you would, also, like to be able to do that without looking at all the past data.

So, the question is how do you describe behavior in an ongoing way? You can keep updating and you can start from new callers, okay, people you have never seen before

because that happens all the time. In such a way you can tell when something interesting is happening.

Now, as Jerry said, you can't hope to be perfect here. All we can hope in some sense is to provide a filter and to give a rich set of calling, of numbers to some human analyst that then will look at it to figure out what really is interesting.

So, all we are thinking of here is sort of filtering the data to look for interesting behavior but even in that one it is fairly hard.

So, what do we have to think about behavior? If you are going to think about behavior and you are going to do it all automatically, all right, you can't think about behavior as I don't know, we can't think about behavior as a picture. We have to think about it as something more structured than that.

So, how are we going to represent behavior? What we are going to do is say that behavior is just a probability distribution, that what it means is what are you likely to do and what are you unlikely to do and both of them may be of interest to you.

In some cases you want to know what typical behavior is just to know how something is being used, but if you want to detect a change in behavior then you, also,

have to know what is typical because you have to know its baseline.

If something is typical all the time, then you can't actually, you know, if what you are looking for is typical then there is no hope of finding it basically. So, we are looking for this change in behavior.

Now, what we have are records for people. Either a record is the phone call record or it may be the posting in the chatroom, but basically you have got records and so that record just has some information in it that you can then extract and put into variables.

In some cases it is pretty easy to see what a variable might be. So, for example, a variable might be the person who posted it, the time the post was made, which room it was posted in and then you can, also, get content information, too.

You get this record that you have and then that has some distribution that varies widely across people. So, what people, the kinds of posts people make, how often they do it, what the different rooms look like all that is changing. Now, if it is changing too quickly you have no hope of tracking it, but on the other hand most things don't change too quickly. So, the hope is that you can evolve the probability distribution as time goes on.

So, what we want to do is to track those probability distributions. Okay, now, these bulletin boards and chatrooms, there are lots of them out there. There are at least 35,000 web sites that support specialty forums, and some of those have thousands of forums, aol.com, yahoo, right? There are all kinds of forums people can join, and then you know some of them are new. Some of them are technical, just anything that you can imagine and these discussions can be fractured and free flowing.

So, most of the time people will post something and no one will answer. It will just go flat, and then about half of the time if there is an answer it is something completely unrelated to anything that has been said before, and probably about 80 percent of it is 18-year-old males seeking something or other.

(Laughter.)

DR. LAMBERT: So, there is a lot of it that may or may not be interesting.

Next slide, please?

So, this is just a small example of what a chat record would look like with stripping off who is making the post, that this is September 17. This is the day that the stock market opened after September 11, from the www. financialchat.com, and you can see there are two different

posters here. One is purple, and one is white, and they are sort of talking to each other in some sense.

So, the question if someone all of a sudden starts saying something interesting are you going to be able to pick it up. We don't actually know that, but what we are actually trying to do first is to understand whether we can understand what, you know, kind of represent what people are doing because if you can't do that, it is going to be much harder to find anything that is interesting.

Next slide, please?

Okay, so what we are actually doing here, and I am not trying to say that this is the best way to do it at all. I am just trying to say that this is the way we have implemented it, and actually Mark has implemented this. This is beyond my understanding, unfortunately, but I can at least explain this picture.

So there are all these discussions going on and these discussions use something called the IRC, the Internet Relay Chat protocol and so they are just writing files kind of like you would write in a log file for a web download.

Okay, so, we have a fire wall and on one side of the fire wall --

Next?

On one side of the fire wall we have our system which is doing the data collection which is just a bunch of Linnex(?) machines basically and then on the other side of it --

Can you hit the next slide, please?

On the other side of it is all the chatrooms which are all running different kinds of protocols. So, basically you have to be able to deal with all that kind of stuff, but you know that is a hard problem, but that is something that you can at least come up with solutions for.

So, at the moment what we are doing is on the right where the chatrooms are we are monitoring somewhere 5 to 10 thousand chatrooms.

These chatrooms, to monitor a chatroom doesn't mean that you are necessarily taking everything form it all the time because most of the time chatrooms are actually inactive except for very large ones, and so what you might do is you have to decide what you want to watch.

So, if there is only, you know, if there is only a little bit of activity like five posts in the past hour then you may decide only two posters are active right now. You may decide that you don't want to sample that and move on to the next one.

So, there is some question about where you want to do it. Now, suppose you actually had some chat streams that you have collected. You bring them up beyond the fire wall and they are all in sort of a funny kind of time order, not exactly time order but they are not in the order of users. They are not in the order of topic, and they are not in the order of even room anymore. It is kind of all mashed together.

So, what you have to do if that is what you are interested in is a different unit other time, and typically time is not interesting when you aggregate over everything because you have to somehow extract from this information about users, information about the group itself.

So, let me just go on. So, some of it we are going to get here are things that are variables that you read out directly and some of it is things that you have to extract. So, for example, rate, you have to get a drive for knowing the time of this post and the time of the last post and just dynamically adapting a rate estimate over time.

Okay, and other parts of it you are going to have to get out topic information.

Next slide, please?

Okay, so now what we want to do is to take each user for example or each room and what we are going to say

is what is the probability distribution that would allow you to predict what is going to happen next for that person or for that room.

Now, you are not going to be able to predict exactly. So, that is where we end up that you need to do this probability distribution.

Now, since everybody has a probability distribution, every room has a probability distribution and every topic has a probability distribution all over different kinds of variables what you have to do is to make sure that you are not trying to be too aggressive here. So, you have to have a distribution that is non-parametric because the variables vary so much from one user to another there is no hope of choosing one family with a small number of parameters and just estimating that.

It, also, has to be small because you are keeping so many of these. You don't get to have much space or much detail on any one person. I mean maybe something that is not quite so obvious is you want to have it simple and that is because it turns out it is often useful to let people know if you are going to show them something what the state of knowledge about the person that you are showing them was at the time you made your decision.

So, it can't be so complicated that someone can't look at it and make sense out of it, and then the last two are some technical problems that arise.

New users are coming into the system all the time. Somehow you have to be able to start off their probability distribution in a reasonable way and so this makes kind of odd clustering kinds of problems in a way or indexing problems because you could look at a few records for an individual and say,"Oh, this individual looks like they are going to behave like that group of people," and take an average for that group of people to start them off with, and then once you initialize you know that is not very good.

It is better than just sort of guessing average over everything in the world, but it is not very good. So, you have to have ways to update it, and this has to be done record by record because you never get to go back and look at all the old records.

Once you see them you have to move on. It just takes you too long to go back.

Can you go back, please?

So, there are many ways to do this. I am not saying that this is the best by any means but just to let you know what we have done. We say that this is a big joint

distribution. We are going to model it by a set of marginal and conditional histograms so that they don't take up too much space and if you had all of the marginals and conditionals then you could always get back the full joint distribution, but clearly there are some of these that are not so relevant.

At any rate, you can come up with ways for choosing marginal and conditional histograms and maybe I will say a bit more about that later, and again you can do this for topics well. I mean not well. You can do these topics as though, you know, you have a topic; you have just standard information retrieval processes that get you started here.

Next slide, please?

Okay, so, one problem that comes up here that we run into is how are we going to visualize all of this stuff. So, you make a probability distribution up for everyone and then you are trying to track everyone. Is there any useful information you get out about what is going on out there in the world using all those probability distributions that you have collected? You have been just sort of using all that data. Well, you know if we have just the call records we can visualize that, right? You can make that plot and then you can look at it, and you can say,

"Oh, yes, I can see what is going on here. This person makes calls in between 1 and 2 minutes about such and such a rate. They never make them in the middle of the night."

You can look at that and you can visualize that and then you can, also, get a sense of whether you think there is a reasonable pattern there. You can, also, if it spits it out and says, "This is something where fraud is being committed," you can then go back and look at it and say, "Oh, yes, the fraud started here. I believe this," or also equally likely probably the fraud, there is no fraud here. This is just that the system has been confused.

Now, in this chatroom thing how are we going to actually visualize because it is much more complex than in the call record world.

First of all there is there is this topic, this text stuff. Well, there is kind of a history in the Bell system of using sound, maybe because of the phone. I don't know. There is kind of a history that comes back from time to time about using sound as well as visualization in order to understand high-dimensional data.

So, going back to the 1960s Tukey was involved in verifying atomic test ban treaties and one thing they did was they discriminated earthquakes and atomic blasts using sound, what they called seisometer sounds and so then

they put on a map, okay, so you take the data; you would map that to a sound which either represented an earthquake of a certain magnitude or it represented an atomic blast, but you didn't know which. You just had the sound. Okay, then you played the sound that went with that. So, you took this time series of seismometer readings. You map that into sound and then you play that back and then the analyst can tell. I didn't bring it with me, but it is pretty easy to tell which is the atomic blast and which is the earthquake.

So, that was one case where sound was very successful. There are other works about a decade later which used going back to Max Matthews who is a musician among other things and John Chambers about using sound to represent point clouds. Now, more recently Mark Hanson who is Bell Labs and Ben Rubin who is something called the Ear Studio had a grant from Lucent(?) in order to see how sound could be used to visualize web data and so what they do is they use sight and sound to represent the evolution of chatrooms. They do this with a huge number of speakers, and little displays.

Next one, please?

Okay, so this would be like one of what a little display would be. So, basically you have go this text coming and right now you can think of doing it in terms of

how the topics are evolving in time and so, as these things come in you can cluster the topics because these topics are represented by a bag of words, and you can cluster those, and if you use some kind of online algorithm which is again, you know, there are ways to do it but not necessarily good ways, you can do dynamic clustering and see what is coming out over time.

You can then map those to displays that are high dimensional, and so this is actually one which I had no idea what it was about, but I think it is about a gecko is what I have been told and so people, there is a bulletin board or a chatroom for people who have geckos as pets and this is a topic for them which I guess is the only thing on about that time of day.

Next slide, please?

So, what we would like to do is to be able to visualize what is the typical activity. Then what you can do is you can tabulate these posts by their length so that longer posts appear as bright pixels in the display. You can look for typical posts. Okay, so before we might look for long calls. Now, we are looking for sort of long posts or long trains of posts.

You can look for what is a typical post. You can look for what is a typical topic. The clustering can be

used as the display changes as you keep updating these cells and as they change. That means a new topic has come in or the topic has changed somewhat and what is shown in that display is you pick a cluster and then you pick a representative phrase from that cluster.

You can also do this for users.

Next slide, please?

Now, that is really as far as we have gotten to be honest, but now that you have these probability distributions at least in principle you can think about trying to detect interesting changes in behavior.

Now, things that are subtle like Jerry said there is no, you know, it is not really going to be too easy to find it. I suppose it is too pessimistic maybe to say that there is no hope, but what we would like to do is to just find out when there has been an interesting change here.

So, we have again this huge distribution. You have got a record. You have put it into, you know, now it has defined variables into it, whether those are topics in terms of clusters of words, whatever or just who did it. You have got all these variables. You can make that up into a probability distribution.

You are going to keep all the marginals and some of the conditionals and the first technical step that you

come into here is deciding what to keep because you can't keep everything and the twist here from maybe the usual data mining procedure is that what you want to do, you would like it to be the best possible choice of variables for each individual.

So, maybe time of day is important for somebody but it isn't important for someone else, but what is the best possible set of variables changes a lot from one person to another.

So, you can't actually optimize for everyone at the same time, but you can take some large set of users and train from them and then what you can try to do is to figure out a set of variables which isn't too hard that is useful for a large fraction of users and then you would say that if is useful for a large fraction of people then you will keep it, but I think this is one place where things get a little murky, one of many, many places, but now if you have that you get probability distribution for what the person usually does say or what the topic usually is in this chatroom.

Now, at the same time you can have a probability distribution for what you are looking for. So, if you are looking for fraud you have some kind of training data; usually it is not very good training data but you have

some, and you can use that to make another probability distribution which is what is sort of unlikely under a fraud scenario and what is unlikely under a fraud scenario and you can use that to direct what is an interesting departure from normal behavior because not all departures from normal behavior are interesting.

If someone always makes a 10-minute call and then all of a sudden they make a 3-minute call well, that is change, but it is just not very interesting most likely. So, what we want to do here is to direct it by comparing the probability distributions of what usually goes on compared to the probability distribution of what you are looking for.

Now, if you do this in a log likelihood ratio then big changes in either one should be flagged here.

Now, as I said, this is easy to do in a call record case. So, you know, just giving you a sense here what happens here is just plotting the cumulative log likelihood ratio of the positive contributions.

You don't want to put the negative ones in because there is some fraud superimposed over the legitimate behavior and you don't want legitimate behavior to cancel out the fraud. So, at any rate that is just what the phone picture is.

Next one, please?

So, I think that is kind of all I have to say about communication data screens, that they are large; they are evolving; they have incredibly complex structure, but in fact it is probably not impossible to get some useful information.

Now, whether you can detect really small signals I don't know, but you can certainly understand what is typical. You can take this information and you can formalize it and structure it in such a way that it is actually useful, but we don't necessarily know how to do it well. There are processing challenges. There is topic classification, evolution challenges. There are algorithm challenges for updating these things, and then for filtering how you decide whether something is worth showing to a human analyst because none of these systems as yet as Jerry said are foolproof, already just set off some long lists of actions based on an automated detection scheme.

Thank you.

(Applause.)

DR. SCHATZ: Are there any questions?

We are running ahead a little bit. If anybody had any questions for Diane and we could even get Jerry back up

here I am sure if there are questions. Does anyone have anything right at the moment?

PARTICIPANT: I was intrigued by your comment about using sound data. Generally we find little response; they feel it is uninteresting. Do you agree with that decision?

DR. LAMBERT: I think it is too harsh to say that it is not interesting. I think that when you have, you know, I think for the point cloud work it is hard to find examples where it adds a lot of value.

I think here where there is actually text and words it may in fact help. Somehow your ear can pick out different words in conversations better than maybe your eyes can pick different words out of a page. So, I don't know whether it is an answer to anything, but I think it helps here.

PARTICIPANT: Let me ask a follow-up question on that. This is not necessarily a question for the speaker but maybe someone in the audience. My impression is that there is at least an order of magnitude difference between sight and sound. That might have a significant bearing on the data.

DR. LAMBERT: I think you don't want to replace visualization with sound but you might want to augment it.

PARTICIPANT: Presumably that would be much more useful. Presumably the phone company of course would have conversations in English. Now can you tell from the patterns what certain; you know, do patterns change in different ways?

DR. LAMBERT: I have no idea. This system, quote, system, I wouldn't call it a system. This software we wrote, that Mark wrote, doesn't know anything about English. What it has is a dictionary, right, and then it has to pre-process the chat stream to get rid of things like pronouns which have very low content, with conjunctions which have very low content and some other words.

Then it gets basically a bag of words and it doesn't really care whether they are Urdu words. It gets rid of these symbols you know and kind of the web speak stuff like LOL, but it doesn't really have any sense of meaning, not at the moment. It probably would be reasonable to put that in, but we did something that is much cruder than that.

PARTICIPANT: I. T. Good had a suggestion years ago. He was imagining the Allies in World War II built a broadcast tower and broadcast 24 hours a day pure noise except at predetermined times something came through in

something that is fairly plain text. In the modern communications era it seems to me that terrorists could publish noise on the Internet all the time and just overwhelm the detection capabilities except for at predetermined times, you know, 12:03 a.m. on Fridays there is something that is actually important and that is how communication happens.

DR. LAMBERT: If there is, I mean that depends. If somehow the nonsense is pure nonsense and the vocab changes, the vocabulary changes when there is something important to say then you have some hope, but if everything has been seen before then you are right. You are going to need something more sophisticated than this.

DR. SCHATZ: A couple of other points I wanted to make on Diane's presentation. She made a very good point for our work which is the idea that a database isn't just a pile of stuff that is sitting there statically.

There is really a very important challenge in tracking behavior over time and really trying to keep track of things. Some bit of data may have arrived a year ago that is relevant today, and that is a big challenge. It is not just what came in in the last 5 minutes.

The other thing which may be covered here in the conference at some point but I couldn't exactly see where

it fit in which is very important to us and might not seem all that critical to the high-end math community when you think of it at the beginning is data visualization tools.

At the end of the day we do all of our analysis and there could be some huge algebra computation or mark-up models, who knows what is going on behind the scenes, but the people who are the first line customers of our algorithms; at our place we call them analysts, they need to be able to get what they need to get out of these things very quickly, and the data visualization aspect of this is critical for us, and so, I just point that out as an area where we might not have thought about that aspect of our work, but that presentation at the end, somebody has to act on something.

The other thing of course that came up in Diane's talk, a couple of points was just good old fashioned clustering algorithms, very important to us, and I think when we start talking about data sets of these magnitudes that we are all up against change point algorithms, clustering algorithms very, very important.

Okay, I think we are all set. Our next speaker is Rakesh Agrawal from IBM-Almaden.

Statistical Detection from Communications Streams

Diane Lambert

At Bell Labs, records are kept about many types of communications, including phone calls and online communications. One useful way to analyze behavior from communications data is to derive probability distributions that indicate what a user is likely and unlikely to do, either of which may be interesting to the analyst. First, it is important to define "typical behavior." Then, it is possible to detect changes relative to the baseline.

A current research project at Bell Labs involves monitoring several thousand chat rooms in order to extract from the resulting data a probability distribution for each user and each room. It is also possible to devise a joint distribution and model it by a set of marginal and conditional histograms. The value of organizing the information this way is that we can compare the probability distribution of what usually goes on with the probability distribution of what we are looking for.

Processing challenges include topic classification, algorithms for rapidly and accurately updating information, and decisions regarding what information should be filtered to an analyst.

Rakesh Agrawal
"Data Mining: Potentials and Challenges"

Transcript of Presentation
Summary of Presentation
Video Presentation

Rakesh Agrawal is an IBM fellow whose current research interests include privacy technologies for data systems, Web technologies, data mining, and OLAP. He leads the Quest project at the IBM Almaden Research Center, which pioneered key data mining concepts and technologies. IBM's commercial data mining product, Intelligent Miner, grew out of this work. His research has been incorporated into other IBM products, including DB2 Mining Extender, DB2 OLAP Server, and WebSphere Commerce Server. His technical contributions have also influenced several external commercial and academic products, prototypes, and applications. He has published more than 100 research papers and he has been granted 47 patents. He is the recipient of the ACM-SIGKDD First Innovation Award, ACM-SIGMOD 2000 Innovations Award, as well as the ACM-SIGMOD 2003 Test of Time Award. He is also a fellow of IEEE.

Dr. Agrawal received the M.S. and Ph.D. degrees in computer science from the University of Wisconsin-Madison in 1983. He also has a B.E. degree in electronics and communication engineering from the University of Roorkee, and a 2-year postgraduate diploma in industrial engineering from the National Institute of Industrial Engineering (NITIE), Bombay. Prior to joining IBM Almaden in 1990, he was with Bell Laboratories from 1983 to 1989.

DR. AGRAWAL: Good morning. This is what I thought I would do. So, this is my piece of the talk. Basically it is going to equal what Jerry said earlier. In fact, Jerry and I were going to share a cab ride and somehow we missed each other from the airport.

It is good we missed each other because my talk is going to be like Jerry's talk in some sense. So, my first feeling is that data mining has started to live up to its promise in the commercial world, particularly applications involving structure data.

By structure data I mean things which can be put in nice relational or object oriented databases. You know, they have fields. They have columns and so on, and Jerry gave a lot of examples of such applications.

My sense is that at this time we are beginning to see data mining applications in non-commercial domains and I guess homeland defense definitely falls into this category of non-commercial domain, and these applications involve use of both structured and unstructured data, but my post hoc sense is that we are just seeing the beginning of it.

I am pretty hopeful that if we have further research in this particular topic we might be able to have

bigger successes in this area, and this is going to be very important as we go in future.

So, what line is that? I think of somebody saying that you know we have off-the-shelf data mining technology available which can solve all the homeland defense problem but have got, you know, that is too much of an exploration at this stage, but there is promise here, and it is worth exploring that promise.

Okay, so, this is what I am going to do. I am going to give some examples of some applications of non-commercial data mining applications and sort of completely ignore all the commercial applications which you know Jerry gave a very nice introduction of that and you might have read that in two data mining conferences and so on, and I will sort of conclude by pointing out some of the important things as I see them.

I like the talk to be interactive. So, we have some time. So, feel free to interrupt me at any time to ask questions.

Next?

I will begin by giving the first example of an application. So, this essentially is showing, you know, I guess Jerry mentioned identifying social links or linkage analysis kind of things.

So, this is showing some person and what does this person's social network look like, and in some sense you mentioned about data visualization. This is using fish eye kind of visualization which is showing that you know in some sense these are the people who are closely linked with this person, and this is the second layer of association and so on.

You can sort of work it out and zoom on any person and get to see what these relationships look like, how these social linkages look.

The interesting question is, you know, this is just an example we created, and the interesting question is how this was done, and basically this was just based on about the caller for about a million pages and using underneath it our data mining algorithm, and so let me just tell you what that algorithm is and then I will come back to this page and show how this was done.

So, this is something in data mining work. You can call association rules and basically you know you have transactions. Transactions is just a set of literal and database construct of a set of such transactions.

So, in this particular example if you look into a page a page would be called a transaction, and you run some sort of a spotter on that and you find out the names in

that particular page. So, those names would become the items of that particular transaction.

Okay, so, a page is a transaction. Names appearing in that page are the items in that particular transaction, and if you are trying to find out sort of you know when ABC happens, then UVW, also, happens, and then there is the social support and confidence and people on this Committee have spent a whole bunch of time trying to figure out computationally what are the good ways of doing this.

Can we do it, and I think you mentioned earlier, can we do this task efficiently because it is clear it is kind of important to obtain. Number of items is extremely large. Number of transactions is extremely large. We are talking in the millions or billions of pages, and number of items again you are talking of all the possible names that can become items. You are trying to find out relationships that might exist. So, it is a huge combinatorial effort and the question is can we do this thing really well and really fast.

If you look into the data mining processes people have spent the last few years figuring out how to do this computation very well.

So, once you have found these rules then let us go back to the previous one. This is simply a visualization which has been put on top of these rules of what is called a frequent item set. So, this is just a visualization to a known problem.

So, this sort of reminds me of what Jerry was earlier saying that a lot of these problems kind of look like major problems but we might have things to identify about them.

So, here is an example, and here what we are trying to do is essentially give a profile for a web site. What does a web site look like, and in this case this is asgonvenga.com(?). This is a site in Pakistan and what is happening here is you know these are sort of different characteristics that we are interested in. It has things like financial support, Islamic leaders and so on and the idea is the following that can we by looking at a site give quote, unquote, a profile of this particular site in terms of these features that we are interested in.

Okay, and you can sort of see how the profile looks like from these particular sites. So, how is this done?

Again, next slide?

This is done by what Jerry earlier was alluding to. This has been done using classification which again is a technique that, you know, Jerry was one of the persons who initially wrote the first decision tree classification, and what it does essentially is this is how the input looks like; you know, this could be in this case records coming from some sort of a, you know, this is the variable. This is people who have high credit risk and these are the data and they are trying to sort of develop can we sort of develop rules for who are the people who are high credit risks, people who would be low credit risk and this is how a decision tree might look like, and the idea is that once you have printed this decision tree a new person comes in and we don't know whether this person is a high risk or a low credit risk and we again run this person's record through this decision tree and we are able to make a prediction.

So, this is what people have done in the past and can we just go back to the previous page? This is simply an application in the same way here. What is happening is that we got pages which are examples of pages which sort of fall into a particular category.

So, we got a good training set and once we have got a good training set we are building the classifier and

once we have built the classifier we can take any site and pump it to that, you know, sort of give the pages of that particular site to develop this kind of a profile.

Yes?

PARTICIPANT: I am just curious whether your example is realistic. It brings up the question of whether you wanted an effective algorithm or a legal algorithm.

DR. AGRAWAL: This is an illustrative example basically.

PARTICIPANT: Of course in market data that question will repeatedly arise in one form or another.

DR. AGRAWAL: Yes.

Okay, next one?

This is the third example I wanted to give which is about discovering trend and in this case I have chosen to show what can be done with some sort of a, you know, again, this is a small application built on top of database and what is happening is the following. So, the input here is on the patent applications which have been filed form 1990 through 1994, and the idea is that you might be interested in finding out sort of specifying I am interested in this kind of a trend. This is an example of kind of a resurgence trend where technology was popular.

Then it started losing popularity, and it is coming back into popularity again.

So, the input in this case is to change your interest and the original input is all the documents in this case patents and you want to sort of find out some sort of a result which looks like this.

Notice in this case we did not provide things like e-tools and pooling as inputs. This was something which was sort of figured out by the underlying system.

So, what is happening?

Next slide, please?

In this case the essentially the underlying technology is sort of two things. One is what we call sequential patterns, and the other thing is what we call chip queries. Sequential pattern the idea is that database consists of database of sequences. A transaction is again as I said earlier a set of items, sequenced in order of this for transactions and you will find out all the sequential patterns which are presented in this database.

This lets you take these patterns and support for them over a period of time and then clearly the shift does support history for that, okay? So, this is the underlying technology beneath it.

Let us go to the previous page, and these things were essentially the sequential patterns which were found in the data and for these patterns of what these sequential factors there is a history and that is what is being shown here.

Next slide?

The next thing I want to mention is the idea of given all the web pages we certainly know the big communities that exist, things like AOL and so on but the interesting question was can we find out microcommunities that might exist, and the idea was that these microcommunities would be some of these types, if you will, and frequently pages are related. Pages with large geographicals were not related and these are some of the sites that you get by looking or trying to find out these microcommunities, and there are things like you find Japanese elementary schools, Australian fire brigade and so on.

So, I mean the labeling of the community was done outside but you could find out that there were these communities that existed. You could look at it and then you could provide label to it, and again underneath it if you think about it the input was large number of pages with the links on them and trying to find out these trends.

Initially there was enough to use the standard algorithm which has been done but then a variation of it was used to discover these microcommunities from the web.

Okay, next one?

Okay, this is the third application I wanted to mention which is again a very non-conventional application that is similar to what Diane was mentioning earlier except that this is not using chatrooms but this is using the news group postings, and the idea was the following, that in some sense we were trying to sort of do the similar thing, can we find out from the news group postings, quote, unquote, pulse on a particular topic.

So, the idea was that in fact a lot of these if you go to groups.google they have all the usenet postings. So, basically on any particular topic you can find for the last 20 years what has been posted and is available to you.

So, the idea was that you take this posting and you have some of a response analyzer, and this response analyzer will show you the results on a particular topic over time, and it turns out without going into details that if you are trying to use standard data mining algorithms they don't work very well because if you are trying to let us say find out people who are against a topic or for a topic than just say let us say finding out about

immigration, how the people are feeling about it, then the positive and the negative postings tend to have almost identical vocabulary.

So, if you use standard classifiers they won't do well. So, the analyzer works essentially just using some sort of computation of a class which is on the scent of what people and hits are doing there.

Let me just quickly show you some results. This is, you know, we took some sort of a news group posting from State Farm and we tried to essentially see if we can from these news reports answer the following question. What do people think of State Farm's handling of claims? And you can sort of see that there is person here who doesn't like State Farm, but a lot of people, a surprising large number of people just comment in defense of this company and sort of answer every person. We found it pretty interesting. You can click on any of these things, see what these people are saying and so on.

DR. CHAYES: Can you tell where they come from?

DR. AGRAWAL: Sorry?

DR. CHAYES: Where these positive people come from, do they come from statefarm.com?

(Laughter.)

DR. AGRAWAL: So, clearly some of them might be agents, also. On the other hand that is the kind of the noise and to stress a point which was made earlier that somebody can use and go in and so it is a classic misinformation kind of a thing, that you can go and try to post things to confuse the thing. So, I mean that is a classical thing. All that you are trying to do was to sort of say technically can something be done, and there is a lot which has to be done on top of it. I mean somebody just takes these and tries to sort of use it, and that is going to be problematic.

I just wanted to show you another one. State Farm decided to leave New Jersey and there was this concern how people were going to react to that particular fact, that they are quitting New Jersey, and again it is kind of interesting which was the point which Diane mentioned about how in these news groups you know there is a topic that lots of people respond off topic and so on, but you can sort of see that really what happened in this case it was very easy to see it once you went through this particular algorithm that in this instance people didn't think State Farm was at fault. It was either New Jersey they blamed or they blamed no-fault insurance which was just kind of pretty interesting to see that again, you know, it is

opinion and of course it can all be biased. It can all be done in different ways but it was interesting to see just how at least by looking at it. There is a tool available for you that you can get a good sense of what is happening.

Next?

PARTICIPANT: All these data mining algorithms we have seen involve large numbers. Are there algorithms where people can be solicited in several states and processed?

DR. AGRAWAL: That is an excellent question, and can I just defer that question because I want to come back to that question? There is a lot of interesting work that has happened and a lot of people are thinking in that direction exactly, but that is a very good point.

So, what I tried to do in this part was to give you an example of some sort of interesting applications and I thought I would mention at least two things which I think are kind of important to, also, think through, and I call them technical taboos(?) because they might really kind of destruct what people think data mining can do and one is sort of what I broadly call privacy concern, and there is some idea of the word "privacy," in data mining that I want to mention here and the second thing is where is this data going to come from. This is what Jerry was calling the

stovepipe essentially. You know how are you going to do data mining over compartmentalized databases?

I, personally, don't believe that you ever have one place where all the data will be collected. It just can't be done, and I personally believe it is not even desirable to do that.

So, how do we do sort of data mining over compartmentalized databases? That is going to be an interesting technical challenge, and in both of these things, again, I try to point what people have been thinking and then again think of the solutions that are sort of meant to illustrate the direction.

So, this is the area of privacy in data mining and there is a clue from the fact that people were sort of really concerned that data mining is too powerful, too invasive. I think Jerry mentioned the example of privacy.org where what people are trying to do is not the best thing that should happen.

So, here is a thought again. You can disagree with it, but that is okay. The idea here was that here is an example of one person's records. This is this person's age and this is this person's salary and here is another person whose age is 50. His salary is 40K and the goal is that using this particular data you want to build some kind

of classification model. We want to put a decision tree, and the question is can we do that particular task without violating individual privacy, and if you think about it what is really private about our being in the values, and what we tried to do here was essentially capture the intrusion of what happens anywhere on the web today.

Basically somebody asks you to fill out a form. People go and lie. You know, they take the age, put that value as an age, you know as the question, and I will pick a number and put that. So, we basically said, "Hey, can we somehow institutionalize this lying? Can we in some sense made the lying scientific?"

That is all we were trying to do. So, instead of you putting up an arbitrary number what you do is you take your true value and you add to it a random value. So, you throw a coin and this coin could be coming from some other distribution and you throw that coin and whatever value you get you add to the true value.

So, the value seen here is extra. Okay, so this person's age was 30.

PARTICIPANT: R is positive or negative or only --

DR. AGRAWAL: In this case this was negative. So, what is going to happen is the following. So, here is an

example showing bimodal distribution once you do this randomization; you know, things might look pretty bad. So, you can sort of say, "What have you done?"

Next, please?

So, just in practice, you can take this randomized distribution and you can factor in how this was randomized and using these pieces of information you can try to reconstruct how the original distribution might have looked like.

So, my idea of deconstructing distribution you are not reconstructing. You will never be able to sort of reconstruct precisely what this looked like, but you might be able to reconstruct the distributions, and once you have reconstructed distributions most of the data mining algorithms require really working at distribution levels. So, you can use these reconstructed distributions to go and build the models.

Next?

This is the reconstruction problem. You know the appropriate distribution of Y and you want this to lead to redistribution and so, I will skip down and skip this, please?

I just want to show quickly that this seems to work quite well. I mean this is how the distribution will

look and this is what this might look like, and this green one is how the reconstructed distribution looked like.

Next?

And if you use these reconstructed distributions this is an example where maybe sort of showing that for a particular task different level of randomization you are losing a certain amount but you might include more privacy and there is this notion of how much of randomization is there. If you can say a value scheme came, the age came from the range here of 200 and you tell me your age is 65, I can't move through on that estimate except to say that your age is between 0 and 100 somewhere, and you have 100 percent privacy. That is the notion of privacy. So, it is basically to say that for randomization you don't lose much privacy.

Okay, next?

So, I just want to sort of quickly again mention that I talked to you about how to do this for numeric data. This is how you do with categorical or nominal data. Here would be a transaction. You replace this transaction with this item, with an item which is not given in the transaction. So, this is the sort of books you have read. You take a book and replace it with some other title which

is not present in the transaction. So, this is when you are doing randomization.

This is another nice way of selecting the complex randomization but it might be a slightly more complex randomization but this one addresses the problem of privacy. I look at the result based on the result and I can infer something more.

Next, please?

Here are some results again here that this is two different data sets, and for something like this exercise these frequent items are what are the things which come together. This is things we have taken the true data. This is how the true item sits on there. After randomization using the random data, I just say how many true positives you could recover. You know, what was the number of false positives, and then you can see that you can still do extremely well and you get fairly high level of privacy.

Okay? So, once again another topic and so I will talk a little bit about this computation of compartmentalized databases. This is again this frequent travel example that Jerry was pointing out. You know, people are very frustrated with the system, that some sort of a frequent traveler rating model for this is going to come from all kinds of different sources, and I personally

believe it is not even desirable that these sources go and share the data, and you know there are lots of reasons for not sharing the data, and I firmly believe they should not share the data unless it is necessary to do it.

So, whatever is necessary, so there has to be some notion of on demand minimal sharing of data which is needed here. It is sort of relevant to the topic that you have with fire walls and so, there is a bunch of data, and if you think about it you know the idea that I presented to you about randomized data kind of applies here also because that would be one way of sharing information by randomizing before we share the information.

The other approach would be that each database builds some sort of local models and then you try to combine the models. So, you do partial computations and then try to combine partial results and third would be to do some sort of on-demand data shifting and data composition kind of thing.

I will briefly point out that this can be done. I just want to sort of cite some work which is not done in IBM-Almaden, but I will try. I think it is sort of worth mentioning.

Next one?

Here in this particular model this is for people who do cryptographic kind of stuff. Here are two parties, a phone conversation here. This is a decision tree type thing on the union of the databases without revealing any unnecessary information.

Next?

This is a two-party protocol, basically how to compute effects of X of Y without revealing what X and Y is.

Next?

And basically decision tree, the key boils down to particularly if they are using ID3 kind of classifier, the key step is to sort of compute information and sort of do this kind of computation and in this case B1 and B2 could be coming from different sources, and using the protocol you can combine that information. That is basically what was done in that paper.

The question, the challenges, you know do these results generalize and can they be applied for the data mining task, and there is sort of interesting work to be done here.

Okay, so, this is sort of my sense of what I thought would be some interesting kind of new directions in data mining towards attacking these so-called "non-

conventional" problems and I just thought I would mention some of the problems which I, personally, find extremely hard. I really personally find it, you know, when I think about them I don't know how to even approach this sometimes.

Most of the data mining is kind of based on the assumption that the past is connected to the future and what if history doesn't repeat? History changes abruptly. It is not a gradual kind of change which is happening, but you know there are abrupt changes, and how do we sort of detect those changes and know that whatever models we have will be totally applicable here, and if you think about it this is exactly what has happened, you know, on September 11, and since then.

The profile of suicide bombers has completely changed from the examples of the current suicide bombers. So, anybody who was building data mining model using the previous examples collected from the past was fighting an old war. So, how do we learn? How do we sort of detect that we are fighting an old war now, and the war has completely changed?

Reliability and quantity of data, this point is extremely relevant. We all know that. The kind of model that you are going to build depends on the quality of data

you have got, and how are we going to sort of detect which one of them is part of the data? Actually it is kind of interesting.

Let me give you an example of what data mining values do? Clustering. Intelligent miner is an IBM product. When it was being developed basically people have taken an algorithm which was used in some place and trying to recode this algorithm and they found out that the results they were getting after this reimplementation were different from the results they were getting before.

So, the question arose what is happening here and they found out that the reason the results were different was that the algorithm had a bug in it and so that is one part of this. The interesting part of the story is the following, that the group which was providing the algorithm and more important the consultants who were using that algorithm in engagements with real companies insisted that we should put that bug back in because they had convinced customers that what they were getting as a result of these algorithms gave good results. Customers were happy. So, think about that.

(Laughter.)

DR. AGRAWAL; So, the other part is the following, that I personally believe that data mining

because these truths are so general can only provide first of all the information and they have to be fed into something afterwards. So, how do we take these patterns and make them actionable? Some way we have to bring them in and i still don't know how, you know, principal way to bring the main knowledge into this whole process. I still don't know that.

I think everybody has said that data mining is like finding needles in haystacks. In fact, it is not. I mean most of the data mining right now doesn't find needles in haystacks. It finds patterns which are dominantly present and how do we sort of save ourselves from overfeeding in our quest for finding real nuggets?

There is a huge danger that we might be just finding noise, and all the things I know of people say that it is about finding really the real events and generally most of the data mining algorithms at this stage today are extremely weak in finding real events.

Patterns? We can find some sets. We can find some sequences. I look at medical applications and there people really wanted to find out deaths or at least some sort of problem of deaths, because that is the kind of pattern they are looking at, and we don't know how to do this. Data types, I mean I have given you examples of text

kind of data, structured data, but you have much better data types on which you want to do mining, when to use which algorithm, you know, I just resonated with what Jerry said. You have a hammer and everybody looks like a nail and IBM, it was very interesting. You know, I work with a bunch of consultants. This consultant was a specialist in Uronet(?). Every problem he will convert into a Uronet problem, right? This guy knew accession. He will take every problem and convert it into accession problem. That was the best algorithm. That was the solution for every problem, and we still don't know which algorithm to use there, and these algorithms are tricky. A lot of them have a lot of wrinkles and you know, is there hope of getting this thing done without the help of something behind them and tuning these algorithms? Is there a hope of building things which in some sense go and figure out in a data dependent way what parameters to use for what particular algorithm?

So, this is just a set of some things I thought I would like to mention, and this is sort of my summary slide saying that I do believe that data can be a strong ally. There is no doubt about that. I do believe that data mining has just shown some promise. I don't think it has matured enough to be used. It requires lots of work, but it

has shown promise and it is worth investing in it, and if we do invest in it, we might be able to realize its potential.

That is it. Thank you.

(Applause.)

PARTICIPANT: Several months ago Karen Casadar suggested that instead of profiling passengers for airplanes what one might do is profile airplanes. It seems to me that you have offered an idea in another domain. I gather you can do social network analysis to find cliques and you have a very long list of cliques based on Internet data. The next step would be to try to characterize or profile the cliques. There would probably be more information at the clique level than at the individual level.

DR. AGRAWAL: I cannot agree with what you just said. Jerry and I were coming. To stop us, you know, we were talking together. They had to cycle some people, right? Who would be the best two people, two nice people talking about this conference. They are going together. Come on, you know.

PARTICIPANT: Sometimes common knowledge points to people who act as a bomb or something like that. Are

there any kind of data mining analyses to discover these kinds of common knowledge?

DR. AGRAWAL: I think that is a really good point. I was talking to someone who was saying exactly what you are saying that, hey, you guys are looking too hard. If you look for common knowledge you improve the quality.

DR. SCHATZ: Okay, thanks.

DR. AGRAWAL: I forgot to answer a question that was raised. So, I think there is something happening right now. In fact, you know there is this point that we don't track a big model in one shop. You take some data, bring some models, and then refine that and that is how the model gets framed up.

DR. SCHATZ: Thanks. A couple of points I just wanted to make before we break here,in Rakesh's talk, which is that the randomization and privacy issue I think is very important for us, too, and the challenges of working across what you are referring to as compartmentalized databases is a big one. It is hard to know how to get a grasp on those, but they are very big problems, important for us, too.

Okay, we have got a 15-minute break. We are back on schedule, and we will see you her at ten-thirty-five. Save your questions up for Don McClure and Werner Steutzle.

(Brief recess.)

Data Mining: Potentials and Challenges

Rakesh Agrawal

There are several potential applications of data-mining techniques to homeland security. These include linkage analysis, Web site profiling, trend discovery, and response analyzer algorithms.

Linkage analysis consists of forming a graph that illustrates a person's social network, storing this information in a database, and recording "association rules" and "transactions" for each relationship in order to permit the use of data-mining algorithms. Web site profiling uses a decision-tree-based technique called "classification." A classifier is built in order to classify a set of people and may be used to distinguish people who are high credit risks from those who are low credit risks. The technique of trend discovery is used to exploit a database of sequences by recording sequential patterns and chip queries. For example, trend discovery could be used to review patent applications filed over a number of years in order to detect a possible resurgence of a certain technology's popularity for inventions. Response analyzer algorithms can be used to identify trends in news group postings on a particular topic over time. However, it is often difficult to distinguish between responses for and against a particular topic, given that positive and negative postings tend to have almost identical vocabulary.

There are, however, several concerns. It is impractical and undesirable to merge different organizations' databases. Therefore, it is important to create local models from independent databases and then develop techniques to combine these models into an overall model. There are also concerns about privacy: a known distribution can be added to a data set, resulting in a randomized distribution from which the analyst may preserve the original distribution without knowing the original data values. Randomized distributions should not affect data-mining algorithms since most data-mining algorithms just require working at distribution levels.

One challenge facing the field is the fact that we cannot always rely on data from the past. For example, the profile of a suicide bomber has completely changed from what it once was. Another challenge is that data mining requires data that are reliable and of good quality.

Donald McClure
"Remarks on Data Mining, Unsupervised Learning, and Pattern Recognition"

Transcript of Presentation
Summary of Presentation
Video Presentation

Donald McClure is a professor of applied mathematics at Brown University.

DR. MC CLURE: Thank you. I would like to remind people of one point that Peter Bickel made at the beginning. The discussants are in general people who come from outside the areas of the main presenters in these sessions, and I am certainly an outsider to the area of data mining. It is not an area in which I have been active.

On the other hand, I do have a fair understanding of problems in the area of trying to extract information from massive amounts of data, generally image data. I will mention a couple of examples of that type in the comments that I make, and some experience with pattern recognition.

I had no idea exactly what the content of the talks this morning was going to be. I enjoyed them all very much. I did take a peak at a couple o papers that Jerry and Diane had written recently. So, I had some idea of what their perspective was going to be.

So, there are three or four points that I wanted to make that related directly to the talks. The first two were sort of related, but an observation that Jerry Friedman made in his talk that advances in the area of data mining will rely on a combination of expertise, a coordination among disciplines and things that different people can bring to bear on these problems, I don't know whether data mining is regarded as an area that is owned by

computer science or owned by statistics, and it should probably be neither. It should probably be an area that attempts to draw on the things that people from different disciplines can bring to bear.

I, personally, feel that one of the challenges in new areas that are poised for advances is technology transfer and this works in both directions. How do we stimulate cross fertilization between scholarly research that is going on on one hand and the expertise of people who are designing and integrating systems on the other hand, and this really has to occur in both directions.

I felt very strongly in my academic career that people doing scholarly research, I am an applied mathematician; so, that the applied part is part of my own discipline, that the research, the scholarly basic research benefits greatly from having an understanding of what the real problems are.

So, there needs to be more, that the academic disciplines can benefit from a better understanding of what the real problems are and in the other direction when good research is done how do we then do what most people refer to as technology and see that that gets integrated into the systems that are being developed so that systems design and integration is not a process of people taking things from

standard toolboxes and assembling those to assist them but instead drawing on new advances in research.

There were two remarks made that relate to a third dimension, my own personal bias relating to the need or use of models, Jerry commented that, and I may have the words wrong. I was jotting down notes during the talk but one of the needs and important aspects of the way in which statistics might have a bearing on data mining would be in the area of assumption free inference. We need inference methods that are assumption free.

I don't know if we completely agree or if we perhaps are saying the same things in different words or in different perspectives. Diane made the comment to pick up from the comments from the talks first that behavior should be modeled as a probability distribution, and I emphasize the word "modeled," that it is important to have a model when we are trying to develop decision procedures.

She, also, emphasized that the models need to be non-parametric or distribution free. So, they are very flexible and can be adapted to all kinds of environments and I think distribution free models may be in the direction of what Jerry had in mind in referring to assumption free inference, models that are very flexible as opposed to ones that are parametric and more restrictive.

My own personal bias that relates to these is that if we are looking for either a needle or a nugget in a haystack we need to do something about modeling the needle or the nugget.

We need to know something about what it is we are looking for. What can mathematics and statistics contribute to problems of identifying and extracting information from massive amounts of data?

I believe that there are many contributions that mathematical sciences can make. Let me, I know I am a discussant, but I couldn't resist bringing a couple of pictures. So, let me put up a couple of pictures to show you?

This is a picture provided Yali Ami of the Department of Statistics at the University of Chicago. He and Donald Gieman have cooperated for several years on research on trying to identify generic object classes in digital images and still images and as a prototype of the problem of detecting generic object classes, not ones where i have a fixed rigid model for what I am looking for but where there is something that in different instances can take many different forms.

For a specific problem to focus on they focused on the problem of detecting faces in video imagery. So, the

triangles up here are hits and in this image these are the faces that have been identified.

This is not a particularly complicated example. It is not a representative example but their methods are really quite impressive and what they have achieved and they have relied heavily on ideas and methods from probability and statistical inference, and I can give people, I don't have titles of papers, but I can give people pointers to some of this work.

Last Sunday morning CBS news had a segment on, the key point was the new surveillance cameras being installed in Washington, DC, and how were these going to be used and concerns that should be debated in public debate about the needs to balance surveillance protection and privacy.

They commented that in London where cameras are already in very widespread use on average every person in London has their face on camera 300 times a day. Now, when we think about trying to identify faces and video imagery this is I think a problem of trying to identify image information from massive amounts of data.

The data rate for standard resolution, digital video, I mean like MTSE or Powell(?) video is, for color imagery is about 25 megabytes per second. Now, all of that

isn't interesting, but if we are going to find anything interesting in it, I believe it is going to be based on having a model for what it is that we are looking for.

There is a line under each face pointing at it, and in this picture it says that it is probably good at recognizing brain wave membrane ties.

(Laughter.)

DR. MC CLURE: It is actually looking for facial features. So, this is sort of a postage stamp that I have blown up in order to present it at that size. So, here is one that doesn't look terribly exciting in terms of scientific content but the problem is to find and read the license plate. This is actually work that was going on in connection with developing a system to be used in parking lots at Logan Airport. The work has been going on for a couple of year.

License plate reading is a problem in computer vision that has been studied for decades. In the sixties people were developing systems to read license plates. I took a look last week at what is available today in commercial systems, and I found something like 50 different numbers of systems on the market for reading license plates, and my feeling for the reason that there are so many systems is that none of them is really a superior

solution, and there is still room for improvement in this problem.

This is a hard problem as a pattern recognition problem to find the plate and to read it. Simple instances of the problem, it is not rocket science to specify a system that is going to be able to do it, but even in this image we can find the plate very easily but there is a complication in the image and it is not -- there is illumination for example and a shadow creates a difficulty in terms of a vision problem, but the real challenges come from just the enormous variety of ways that this problem can manifest itself.

So here are some challenging license plates to read. These come from a web page, www.photocomp.com where there is actually a fairly interesting summary or review from a consultant system integrator about systems that will read plates.

License plates as an optical character recognition problem are challenging because of the many different forms in which the problems can be presented. So, the small characters with large characters, this is actually pretty common if you watch plates as much as I do when your car is riding down the street. The graphics are very common on plates and license plate frames that have a

habit of invading the bottom, strokes of the characters so that we can't simply say that the characters are dark characters on a light plate background.

License plate reading is a more difficult problem than a character recognition problem. We can report that statistical methods from statistical inference have contributed to a very successful algorithm for reading these characters.

This is relative to the plates. This particular idea is an easy problem. It is a very easy problem to contrast the light characters on a dark background. It is not like the noise. It isn't white noise in this image. The background has structure in it, and the characters come from a fixed font, from a font specification. So, we know exactly what the model for a character is before we start trying to do the reading, but in these problems these are problems, these are characters that are etched on wafers and as wafers go through the manufacturing process things happen to make this a more interesting vision problem.

So, there is according to the specs supposed to be clear space around the identifying marks on a wafer but the semiconductor fabs try to use every square millimeter of space on that wafer so they will etch over the region where the ID occurs.

Now, I won't put up any more examples of this, but I will return to my general remarks for a minute just about how mathematics can play a role in a problem like this more generally how mathematics and statistics can play role in problems of extracting information from non-conventional massive data sets.

In addressing this problem, an approach that we have used the problem is first of all putting on our statistician's hat regard it basically as a hypothesis testing problem. I want to decide that every location in this image in the whole field of view is there a character present at that location, character present versus character not present.

There are typically 36 different characters, alphabetic characters and the 10 digits. So I can refine that process or that hypothesis of character into it is a composite hypothesis formed from the different characters that might be present.

So, at any rate we view this as a hypothesis testing problem and try to model what it is we are looking for and try to model the variation and try to model the probability distributions.

This involves identifying in variants of what really defines what it is that we are looking for and using

distribution free methods, using non-parametric methods in order to have robust methods for forming a basis for decision tree procedures, articulating the model in the form of probability distributions, so that when we need to actually make a decision among the hypotheses we can use methods that are based on time-honored principles of statistics such as like the good ratio tests in particular in this problem.

At any rate I will conclude my comments. I found all three presentations this morning to be very, very interesting.

DR. SCHATZ; Thank you, Don.

Remarks on Data Mining, Unsupervised Learning, and Pattern Recognition

Donald McClure

Advances in data mining rely on coordination among disciplines, with dialogue between people doing scholarly research and those who are designing and integrating systems. Models used in data mining should be very flexible as opposed parametric and restrictive. If we are looking for a needle in a haystack we need to do something about modeling the needle. We need to know something about what it is we are looking for. In problems of extracting information from massive data sets, nonparametric procedures in particular, we should permit robust decision-tree analyses that articulate the model in the form of probability distributions. In that way, when we need to make a decision among the hypotheses, we can use methods that are based on time-honored principles of statistics such as likelihood ratio tests.

Several applications of data mining are being used in computer-vision research related to homeland security. One example is finding and reading license plates. Although there are many commercial systems to do this, there is still room for improvement in the areas of illumination and its resulting shadows; character size, type, and color; graphics; contrasts between characters and background; and the obscuring of information by license plate frames. In addition, data mining will help in the detection of patterns on silicon wafers as they go through the manufacturing process. Although there is supposed to be clear space around the identifying marks on a wafer, semiconductor makers try to use every square millimeter of space on that wafer, often etching over the regions where the identification would occur.

Werner Stuetzle
"Remarks on Data Mining, Unsupervised Learning, and Pattern Recognition"

Transcript of Presentation
Summary of Presentation
Video Presentation

Werner Stuetzle was born on September 14, 1950, in Ravensburg (Germany). He completed his undergraduate degree at Heidelberg University and at the Swiss Federal Institute of Technology (ETH) in Zurich. Dr. Stuetzle completed a master's degree (Diplom) in mathematics (1973) and a Ph.D. in mathematics (1977), both from ETH. He studied estimation and parameterization of growth curves under the direction of P.J. Huber.

Dr. Stuetzle was an assistant professor in the Department of Statistics at Stanford University from 1978 to 1983, with a joint appointment in the Computation Research Group of the Stanford Linear Accelerator Center. In 1981, he was a visiting professor in the Department of Applied Mathematics and Center for Computational Research in Economics and Management Science at MIT. In 1983 and 1984 he was a research staff member at the IBM Zurich Research Lab.

Since 1984, Dr. Stuetzle has served on the faculty of the Statistics Department, University of Washington, Seattle, with an adjunct appointment in the Department of Computer Science and Engineering. He was chair of the Statistics Department from 1994 to 2002 and spent 1999 and 2000 on sabbatical in the Research Division of AT&T Labs.

DR. STEUTZLE: All right, I just made a couple of slides here. So, one thing that struck me this morning or when I actually looked at the transparencies for Jerry's talk that he sent me ahead of time, one thing that certainly struck me is that there are very high expectations for the usefulness of data mining for homeland security.

So, for example, when the Presidential Directive says, "Using existing government databases to detect, identify, locate and apprehend potential terrorists," so not people who did something, you know, who might do something in the future, that is certainly I think an extremely ambitious goal, the same way with locating financial transactions occurring as terrorists prepare for their attack. I mean given that it doesn't take a lot of money actually to commit terrorist activities and then finally, the use of data mining to pick out needles in a gigantic haystack.

So, as to how realistic that all is I remember when in the late sixties, early seventies there was a terrorist group in Germany called the Baader-Meinhof gang. So, Germany is a very highly regulated society compared to the United States.

For example, the government knows where everybody is. I mean when you live someplace you have to go to city hall and you have to be registered. When you rent a room you have to go there. So, Germany has much more complete control over its citizenry than the US has or I am sure that US citizens would ever tolerate. That is point one.

Point two is the bottom line is knowing who these people were. So, they wouldn't go looking for people who were going to commit activities. They had actually done stuff. I mean they had murdered leading bankers, politicians, judges and so on, blown up things and so even despite all these, besides those two factors, the highly regulated society and it was already known who the perpetrators were it took years to actually track these people down.

So that is one thing. I think that is an important thing to keep in mind. They basically looked for certain characteristics. So, for example, when landlords reported the rental of apartments they would check for certain things like how old was the individual. Do they drive BMWs or other fast cars? Did they make the deposit in cash, etc. So, that is in a sense a data mining technique but despite all it actually took years to track these people down.

So, I am not saying that it is impossible. I am just saying that even under very highly regulated circumstances when you know who you are looking for this is not a trivial matter at all.

Okay, so then when you think about applying data mining to this problem there is first of all what I would call the simple although not easy. Some of the ideas have the super big databanks. So, it is actually not physically one database. It is the collection of all information available so that it could be travel records, bank records, credit card records, previous travel patterns, demographics of the people, etc.

So, you have properties of people and you also might have network data like who interacts with whom. Who calls whom. Who sends e-mail to whom, etc., and so forth. So, you have this large collection of information. In principle you could have that amongst all the people in the US and then the idea is to apply data mining tools and then from that you estimate the probability that the person is a terrorist given all these properties.

So, that is a little part of what one would do, and so you look at all data. You apply data mining tools and you basically classify people into potential terrorists or not potential terrorists.

So, this is actually, so something related to that was what Diane was talking about. You take the calling card fraud. So, that is what AT&T does to detect calling card fraud. They know the people who have AT&T long distance and then they collect call records of these people. For every call in the long distance network there is a call record which gives you the calling number, the number that was called, who the call was billed to, how long the call was, etc. They have a long record of that. They build up a profile for all the users and then based on that profile and some training data they try to make a rule that sort of flags people who might or who are suspected of committing calling card fraud, and that is a much easier problem.

First of all, you have a universe of people that you are interested in and those people have AT&T long distance. So, that is the first thing why it is easier.

The second reason why it is easier is the consequences. So, the consequences of missing somebody or erroneously identifying somebody are quite different. So, if they think that somebody might be committing calling card fraud or if they think somebody is abusing my calling card they will just give me a call and say, "Well, did you really call Nigeria 15 times the last week?" And if I say,

"No," when then they are confident they have found something and if I say, "Yes," well, then that is not a big loss and it is different from being apprehended at the airport. So, if they miss somebody they will just lose some money which is painful, you know, but not like an airplane that has gone wrong.

So, this is a much easier problem but even that is already quite, I mean people have put a lot of work into that. So, now getting back to the original problem of estimating that someone is a terrorist given the properties, the problems are first of all you have a very small training set.

Jerry alluded to that. Actually you don't have a small training set. You have a small number of bonds in the training set. You have 280 million zeroes and 16 ones in that training set. I mean one example, Ted Kaczinski and Tim McVeigh you don't really see the attackers. Mohammed Atta and accomplices, but say you have the useful data in the training sample. So, that is one problem.

Then the second problem is what you know from anybody who has ever taken any basic statistics course knows the problem with screening. If you have a tiny incidence with the population which we are going to have you need methods with very high specificity. Otherwise all

your alarms are going to be false alarms. So, that is another.

So, that was something that Rakesh pointed out. You are not looking for things which are dominantly present. That is what data mining is good at, but you find out the people, that most people who buy beer also buy potato chips. So, that is something that is just dominantly present. So, you are looking for things which have a tiny incidence and so unless you have very high specificity you get a huge number of false hits, and finally, another problem is that the good predictors might not necessarily be legal predictors.

Good predictors might be if you listen in on everybody's phone calls and automatically process those, and you can look at everybody's e-mail, etc., and that might not be possible or you might not want to do that. It might be illegal or just infeasible because of the sheer mass of the data.

Okay, so, even in its highly idealized form where you have this universe of people and you want to make this rule and if you could put all these databases together and run these data mining algorithms, even in the idealized form this is very difficult.

Now, however, this is much too simplified. This is much too simplistic. First of all there is going to be a lot of individuals outside the system. So, the system, meaning the people who are in the US and are in these databases, well, first of all there are foreign travelers, people who come in as travelers who are basically not in the system of immigrants. As far as I know the US doesn't even know how many millions of illegal Mexicans you have. I mean it is a matter of millions of people that one is not sure about. So, it is not at all clear, but that universe is not clearly delineated.

PARTICIPANT: They have attempted to build a system in the US.

DR. STEUTZLE: I don't know much. I just hear rumors that it is not that effective. I might be wrong.

Okay, so then there is the other issue of identity. So, even if you had the universe of people and you can run that decision rule and then I show up at the airport then the rule says that I am not a potential terrorist. Now, that only makes sense if I am really who the system thinks I am. So, therefore, if you can't prevent identity theft that kind of system is not going to be very useful because you have to be sure that the person that you are making the prediction about really is the

person standing at the gate. If you can't guarantee that then this prediction isn't going to be very useful.

Then of course there is terrorism on foreign soil which seems very hard to prevent. So, now, I am not a national security expert but realistically it seems to me what you can hope to achieve is some access control. So, you can hope to deny access to some critical places, and places i put in quotation marks because there could be virtual things like access to certain machines on the network, etc. So, you can maybe hope to deny access to some individuals who are either outside of the database, so they are not in your universe; you don't know anything about them. You could just say, "Well, I am not going to let them do X or let them go in." So, that is something you might realistically be able to do or you could deny access to people who are inside the system on whom you have data but you have bad indicators. So, you can run the specification procedure on your universe and you can try to see, well, you can try to make such a prediction, and of the likelihood of being a terrorist and if you think that is high you might deny access.

So, finally, it seems to me that for any such thing establishing identity is critical and so it is not easy it seems to me and these biometric identification

methods are really crucial, really have to be a crucial part of any strategy because that is the only way to totally reliably establish identify.

All right, that is all I have to say.

Thanks.

(Applause.)

DR. SCHATZ: Don, Werner, thanks very much.

We have a little time for discussion. I do want to comment that the biometrics problem is very important to us at NSA, too. We have a large group of people who have been studying that recently.

Just to get some discussion going here, if I could take one second before we jump in at NSA we have an activity we call the advance research and development activity and the director of that activity, Dean Collins is here, and I know that no one in this room is the least bit interested in funded or money or anything like that, but Dean actually provides funding for external research, and if he could just take a second to talk about his program I think you will be interested for a number of points of view.

So, Dean, would you just give us a couple of minutes here on that?

DR. COLLINS: Thank you, Jim.

My name is Dean Collins, and I am the director of a very small organization called ARGA(?) and has anybody ever heard of us? Oh, a couple of people, okay, my colleagues.

(Laughter.)

DR. COLLINS: So, there is a comment that I think that Jim is doing a very nice job of describing after each talk the things that NSA is interested in, and I represent not only NSA but our little activity is an intelligence community activity. So, to a certain extent I go across all the three letter agencies.

The thing that Jim mentioned was that his customers are the analysts and so after all the data comes in then the analyst looks at it and I have been in this business, this particular job for a little over 2 years, and I have spent a lot of time with analysts, and I would like to tell you that a lot of the tools that are provided to them are not used, and so, we have only four particular areas we work in and one area that we are starting up is an area called novel intelligence from massive data, and that kind of clicks back in here, and it is hopefully of some interest to you not necessarily for the funding although if you would like to participate in the funding we would certainly welcome that, but we have spent about 6 months

working with analysts and saying, "What do you want?" as opposed to what can we give you, and I thought that you know if you have a hammer, you know, every problem is a nail. It is sort of like what are the problems in the form that you would like, and so, one of the things that we spent in the 6 months is a formulation of what are the analysts looking for, and they are quite different than you might think.

Since I think the patriotic duty which a lot of us ascribe to in the fact that you are here you might be interested in looking at the, you know, what the analysts really want.

This program is ongoing as we speak. It is unclassified. It was put out in what I used to call a Commerce Business Daily but it is now called Fed Business Ops and I will write down the URL. You can look at it. As of Wednesday there were 145 white papers. Obviously not a lot of people in this audience are aware of this opportunity and maybe you decided you didn't want to do it.

I would just like to bring that to your attention but primarily from this standpoint. Look at the questions. You know, that is the important thing, you know, the funding put aside. Look at the questions about what they are really interested in. You may get a totally different

viewpoint of what is important to an analyst, and they are looking for that rare occurrence and so that the gentleman from IBM I think put it very succinctly. That is a very difficult problem, and all of the work is unclassified, and there will be conferences involved.

So, if you don't get into the program you might like to at least attend some of the conferences.

Jim, thank you very much.

DR.CHAYES: I have got one question for you. Who are the people you are working with?

DR. COLLINS: We are not working with anybody at the present time.

DR. CHAYES: But in universities there are people, I mean what areas are they from, math or --

DR. COLLINS: We are problem focused not discipline focused specifically. We can't tell you who we are working for because nobody is under contract.

DR. MC CLURE: Who is submitting the white papers? Where are they coming from?

DR. COLLINS: I do not know. I have not looked at any of them. They are just coming in and that is something I couldn't tell you if I did know because that would be interfering with the procurement process.

DR. CHAYES: And this URL has the description of what you are looking at?

DR. COLLINS: It refers you to an announcement for the questions.

DR. AGRAWAL; Can you give us one example?

DR. COLLINS: You mean one of the problems? Okay, I think that one of the problems for the analysts, and this is something that is a very big problem, this issue of passive knowledge. If you do a database, a data mining search and you come out and say, "Gee, there is a strong correlation between the White House and George Bush," this is not very interesting to an analyst and he is probably going to take your software and throw it in the trash. So, now this is a very difficult problem but it is a very crucial one to an analyst.

DR. SCHATZ: Okay, Dean, thanks.

DR. COLLINS: I will write this down afterwards so I don't interfere.

DR. SCHATZ: So, I will open the discussion up here to anybody who would like to start.

PARTICIPANT: I thought the issues about essentially screening which is what we are talking about were very important issues and we can apply all the mathematical techniques we want to but if there are only a

half dozen positive cases in a massive database you have to use your domain knowledge. You have to figure out how the domain knowledge, but we are also going to have to figure out policy and so I think that one of the things that this community should be, we are looking at the question of how to find the terrorists but we also have to bring to bear knowledge about, or we have also have to bring to bear policy issues like screening programs. If we have the metal detectors that keep people from doing things, how does that play in? Are we reducing the incidence of these things? We have to bring to bear behavioral analysis. Behavioral and social science research has to be brought to bear on this. How can we quantify that and put it into the same framework so that it can be integrated in the statistical models?

DR. SCHATZ: Good points. Did I miss a question in there?

PARTICIPANT: It was a point that I wanted to make and I wanted to ask people to make comments on that if you would.

DR. STEUTZLE: I really have got not much to say to that. I mean I agree that that is true, and that is a problem I think that classification procedures are not necessarily, it is not always so clear, for example, how to

integrate networking information into those. So, that probably would not be properties of a single person that allow you to make diagnosis but properties of how a person relates to other people in the database, and that is not always totally obvious and how you would integrate that, how you would use current methodology to do that.

DR. CIMENT: I would like to hear a little bit more from people with a vision about the future that relates to possibilities of using mathematics in the context of new architectures not just looking at the world the way it is today but the way the world might be in years to come based on for example, the proliferation of supply chains based on computer integrated systems, sort of the WalMartization of the world you might say, right? The story I heard relative to 9/11 which makes it significant is that certain citizens understood immediately after the impact of 9/11 that there were certain requirements.

Motorola sent truckloads of batteries for New York City with policy escort, understanding that people with cell phones were going to go out more because they weren't recharging them.

Home Depot I think was the one that shut down 20 of its stores understanding that people were going to need their appliances and services and things like that.

What I am getting at is the extreme events of the future which we never know what they will be. They won't repeat the past, but it seems like they will require all sort of integration, not just from governmental agencies but from private sources, private organizations, industry, etc.

The challenge for us in our society as Werner well pointed out that will not tolerate privacy invasion is to develop ideas that will preserve the privacy, allowing data sharing and I think mathematicians are probably the most suitable people to think these abstractions through and not worry about what the lawyers tell you you can't do, what the policy people tell you you can't do, create these models and show that there are maybe possibilities here if the policy would change and if the architecture would change, and we might have a society that preserves both security and privacy.

DR. MC CLURE: I would reiterate what Jim jus said. That is an interesting point. I have given absolutely no thought to this. It sounds like you have given a good deal of thought to this. So, I don't really have anything to add.

DR. SCHATZ: These are certainly critical issues, the privacy issue versus trying to actually do something

for homeland defense, and they are hard issues that we at NSA face every day because we go through extremes to ensure privacy for citizens. We are there for the citizens and yet you are trying to do these data mining things. We have had more discussions with lawyers at our agency in the past 8 months than probably in the preceding 40 years, but I think you are right. I mean just the randomization ideas that Rakesh pointed out are very important for this kind of work. I mean so much of the time, especially in the industry setting the information people want you don't need to have individual names. You are looking for trends and so forth.

PARTICIPANT: This is both a comment and a question and I think related to one of the points that Don McClure mentioned. This issue of data mining I think is sometimes too general. Much of the action I think has to be at the level of domains. A particular example he mentioned which is what I know about, to say, "Oh, yes, this is a classification problem. It gets you maybe 10 percent of the way or 20 percent of the way," and there are studies showing that we are not suggesting taking a general off-the-shelf thing and running it. You have to look at the domain. There are particular questions how things are found, what the fusions are, what life forces, etc. It

seems to me that this theme may actually be doing a little bit of this. I just take an extreme position here, by forcing people to concentrate at this very broad level, whereas if they came down to the more specific level they might actually make more progress.

DR. MC CLURE: This relates to the point that I mentioned that Jerry Friedman made also about problems in this domain just requiring the coordination of many different kinds of expertise. It is the missions of data mining but when you get down to the specifics it doesn't seem like everything fits within the kind of highest level definitions.

DR. SCHATZ: Certainly a good general point. I mean at the National Security Agency the one thing we have that not everyone has is data and because of that it really forces us to, we have many conferences where we try to think big picture, but we have many very specific problems and there is nothing like very specific problems to just get you down to business in designing algorithms and pushing forward and you are right. There is only so much progress you can make at an abstract level and data mining the concept is vague and broad and so on.

Yes, Peter?

DR. BICKEL: I would like to follow up on that in a somewhat facetious way. Jerry had two things. He had assumption free, and then he had domain knowledge. There is only really one assumption-free statistical model, that is that you have all of the things that you observed on certain -- I am talking unknown distribution and that doesn't get you very far. So, domain knowledge already implies that you are starting to put in structure and of course the more you know, the more you can put in.

DR. PAPANICOLAOU: I am George Papanicolaou of Stanford University, and just about the comment that Diane Lambert made about during the sixties the sonification of seismograms to determine whether they come from natural earthquakes for from explosions and I know a little bit about that problem. I would like to see how you feel about this issue. In fact, in the early sixties the discrimination problem was very much an unsolved problem. It was a very large community of maybe a couple of hundred seismologists and applied mathematicians and others who were involved in this, and the problem wasn't solved until about 20 years later in the eighties, and it turned out it was a very complex problem. It wasn't an issue that you could hear, that the seismograph people hear whether you could discriminate with your ear, but it was an

understanding of how waves that are generated, the surface of the earth over long distances. These are important. Yes, we are collecting information about seismograph activity down to a very small earthquake level, natural or otherwise.

This information is collected. We have huge databases. International agencies are collecting. There is a data mining issue there, but the real issue is when you go and use data how do you discriminate; what is the basic science that goes in there that would tell you what to do with the data that you have, and speaking also in the direction of some of the criticisms mentioned earlier exposing data mining in such very broad terms hides some extremely important long-term issues in basic science like computational wave propagation and various other issues for example that have to do with the imaging in the complex environment, and these are very bit as important let us say for the detection problems that are going to nuclear proliferation. Very small tests are going to be made and have to be discriminated as to the role the data mining problem starts to extract some general information, some popular information and I think that that is something that this community, this small group here should attempt to put a fix on. Exactly what will the general, in order to better

fix, what will the general ideas do; how far will they go and that they are applied and that the underling basic science needs to be emphasized very seriously.

DR. SCHATZ: Disagreement with that? Good point. Yes?

PARTICIPANT: We talked in an earlier session about techniques and terrorists changing their tactics, and focusing on false negatives. What about terrorists who try to force the system to make findings in people? Any comment about what current techniques and that kind of thing? What kind of techniques might there be?

DR. MC CLURE: I don't know. I think about the problem as it has been approached over the years of automatic or aided target recognition. It is always easier to try to foil the system than it is to design it to be effective, and I think that is a problem in this area.

DR. SCHATZ: Yes, you know one of the problems about that that struck me at the Siam data mining conference a couple of weeks ago is that our community of which we have a cross section here; we have academia; we have industry; we have the government and among ourselves we would like to share information about how we are doing this data mining, but in truth it is a little bit difficult because companies have proprietary information. They are

trying to put out products and academic folks need to be tenured and government agencies keep what they are doing secret, and it is all for very valid reasons.

If people know how AT&T is going to detect phone fraud, they will get around that. So, it makes it difficult to exchange the science sometimes I think because of all the proprietary information and that is a difficult situation here, but at the same time if we don't keep our sources and methods quiet when we need to they won't be effective either.

DR. STEUTZLE: Getting around things is only one problem and once you know how the system works then you can also flood the system so you have both options that you can flood it or get around it, you know. So, that is I guess why the airlines don't want to tell you exactly what they are looking for when they profile you at the gate.

PARTICIPANT: The reason I asked the question is because even with many false negatives you feel the positives are still useful. With many false positives it is untenable and the terrorists may force us to remove the system entirely and that is why I asked the question.

DR. SCHATZ: Good point.

PARTICIPANT: I don't think it is a question. It is more of a comment. It is not so hard to find a needle in

a haystack. The point is that you often find sometimes the thing that you are looking for is a special feature that occurs in a population not necessarily for a single individual. The project that we had with fraudulent access to a computer system, sometimes you can just ignore the data and look for some movements or command that is not typical. Very much to what Werner mentioned your suspicions of the sixties and seventies looking for a robust method, sort of trying hard to avoid, how to evolve more for extreme events, sort of reverse thinking in trying to find these things in the bulk of the data and then from there on you can sort of try to find the individual. In fact, even in a synthetic model if you look at Diane's data if you take Diane's data and say, "Can I detect fraudulent usage of phones?" having the phone bills of individuals for say a year or two, taking just random streams will give you this phone and my phone and somebody else's phone bill but just for a week you can observe for a week there is ongoing data. Could you find fraud in that type of data, and the answer is probably you could because like Werner said if you see calls to Nigeria that may be a good start.

DR. LAMBERT: Maybe I should say something. I wasn't actually trying to say that you can use these for terrorists. This is far beyond what we ever try to do.

The other thing is maybe we are focusing too much on trying to accomplish the final goals whereas it might be useful just to give people a filtered set of information so that they have less than actually puts that by hand, which is you know it is not that we are trying to accumulate analysts. We are so far from that; we are not trying to do that at all. Another thing is that even you know, actually in detecting fraud you don't have long histories on people because if people are going to commit fraud they don't go into the system that they have long distance service with for example. They make a call and access somebody else's system where you have no history whatsoever. So, you are right, being able to handle people is very important, and I will have to defer to the comment about earthquakes. That is just some math that I had which had little symbols on it. I actually don't know. It could have been that they developed the signals and used the signal extraction from 20 years after the original application. You do have to take all the information you have. The trick is to figure out how do you handle it.

DR. CHAYES: Just as a summary I think I am not someone who knows about any of this but what I am hoping is that what we are going to get out of all of these sessions are some questions that mathematicians can approach, and so

I have just been writing down some more mathematical questions that have been coming up. That is also one of the things that we want the final report to do, to come out with a list of questions that mathematicians can look at, and I guess the one that has been coming up the most is how do we focus on extreme events and what I heard from everybody is that we really have to know how to model extreme events properly. So, I am not sure how much of that is the mathematical question.

DR. AGRAWAL: That assumes that you know something.

DR. CHAYES: That assumes that you know something. So, on a general level how do you get extreme events and it sounds like we are very far away from that.

Another one that Jim mentioned was if you have a lot of data how do you visualize the data and I know that there are people working on this. I am certainly not one of them. I am not sure if there is anyone here who can speak to the question of how do you visualize data.

PARTICIPANT: Not just visualize.

DR. CHAYES: Yes, I mean in a metaphorical sense how do you visualize data and then there is also the question that seems to me the one that we are furthest along on which a number of people talked about which is how

do we randomize data to try to ensure privacy along with security. However being further along doesn't mean that we are very far. So, it struck me that those are three areas that could set a mathematical agenda and if anybody has any comments on any of those?

PARTICIPANT: I have one comment. If we are talking about addressing terrorism are we talking about preventing a small number of events and data mining to prevent these events to make sure that you have got every single individual and on the other hand there are a number of organizations like Al Qaeda and maybe we could concentrate more on the structures of these organizations and then you are not talking about identifying every individual, identifying every possible conspiracy but identifying plans of the organization.

DR. KARR: I am Alan Karr from the National Institute of Statistical Sciences and I would just like to point out that there is a wealth of techniques associated with preserving privacy in data other than randomization. Randomization has some well-recognized shortcomings in other cases, but I think this point is a lot broader and there is a whole area of statistical disclosure that ought to be brought into this.

DR. LASKEY: With these issues that have been brought up I would like to add one more which is combining, by the way, I am Kathy Laskey from George Mason University and combining human expertise with statistical data and that does in fact have mathematical issues associated with it because of methods where you represent the human knowledge and ability distributions to combine them to data, and there are lots of important innovations in that area.

I would, also, like to point out on the varied events the importance of outliers of rare events have been mentioned a lot, but the importance of multivariate outliers, data points that are not particularly unusual on any one feature. It is in combination that they are unusual and in fact in the events leading up to September 11, these people blended in with the society, but if you look at the configuration of their behaviors if somebody had actually been able to home in on those individuals and say, "Okay, you know, they paid cash for things, plus they were taking flying lessons, plus, they were from the Middle East, plus, this, plus," and then you discover that an Al Qaeda cell was planning to use airplanes as bombs there were enough pieces that could have been put together ahead of time, not that I am saying that it would be easy, but pieces were not

individually significant enough to set off anybody's warning system. It was the combination that was the issue.

DR. KAFADAR: I am Karen Kafadar from University of Denver. I think I heard someone from the FAA say that actually the airlines did identify something unusual about at least one of those. The response was to recheck the check level, rescreen the check level. There was another variable there. They didn't know that.

DR. SCHATZ: I don't want to cut off any discussions although it looks like we are getting into the lunch break. We will take a couple more quick ones and then I am sure there will be lots of time later to talk.

PARTICIPANT: I am trying to put together a couple of things that seem related. One is that we classify this and we can see this and this and this, and that ought to be intuitively meaningful and it is information that ought to be the model used, but I also have a sense that we are looking for the kinds of things you can see looking back but not forward so easily. In retrospect every newscaster would know what was coming. So, what is the potential for these folks who are analysts who are in the business of knowing how the targets are changing? Are we talking about being experts in real time participating in the development a system that might have

to change in time as well and what are the odds that the system can say, "Is this interesting?" and have them say, "Yes," or "No," and then the system from what the analysts thought of it with the perspective of the analyst looking at it Tuesday of this week instead a month ago and with all the complexity that you are not going to be able to deal in rules no matter how careful you are; so, is that, I know analysts are probably overworked like everybody else, but maybe you could participate in something like this.

DR. SCHATZ: We do a fair amount of analyzing the analysts if that is what you are asking. I get in trouble at our agency when I talk about rebuilding the analysts because they don't like that, but we do; a lot of our activities and algorithms have to do with on the one hand helping them prioritize data for them that we think they are interested in based on what they have been doing, try to predict things they should have looked at that they are not getting time to get to but modeling analyst behavior is something that we do all the time and will be more and more important for us, absolutely.

PARTICIPANT: The third time that a rep came to us and said, "Bush is linked to the White House," you know the system should be one because the analyst knows well that that is not interesting.

DR. SCHATZ: Yes, there certainly is for us again we enjoy a population of people to study in that regard that other people don't have access to, but certainly when we do have access to it, and we do, a lot of what we do is studying analyst behavior and trying to correlate did they pull the document; did they look at a document; did they act on a document and try to maximize our advantage there because at the end of the day no matter how many individuals you have it is a minuscule epsilon number compared to the data size. So, what they actually do and act on is critically important.

One more, Rakesh?

DR. AGRAWAL: It is not a question. Many times I like to go and look at things, but sometimes I think I wish there was more computational aspect to it. So, in decisions and so on the interesting thing is like the combination of things. There is a lot of very interesting work happening and it is interesting for somebody in this Committee to understand what happened and to look at it, to understand the computational people and something I very strongly believe that we don't have hope for doing some of the massive common warehouse kind of things that somebody would pay for. I don't have the experience to look at the kind of data you have they are critical for commercial testing

in the field and they can be done. So, how would you solve all the complications that you have which essentially assume that there is one data source but think how would you do all the computations you wanted to do where you have these data sources which are kind of ready to share something through a mode of computation and these are some of the kinds of data points here which would be useful.

DR. SCHATZ: Very relevant, absolutely.

Okay, Andrew?

DR. ODLYZKO: If you look at the broad technology what we have to know in the next few years is storage capacity.

DR. SCHATZ: Good wrap up. Thanks, everybody. Thanks to the speakers for the morning.

(Applause.)

DR. SCHATZ; Twelve-thirty, here.

(Thereupon, at 11:50 a.m., a recess was taken until 12:40 p.m., the same day.)

Remarks on Data Mining: Unsupervised Learning, and Pattern Recognition

Werner Stuetzle

There appear to be unrealistically high expectations for the usefulness of data mining for homeland security. When a Presidential Directive refers to "using existing government databases to detect, identify, locate, and apprehend potential terrorists," that is certainly an extremely ambitious goal. For example, pinpointing the financial transactions that occur as terrorists prepare for their attack is difficult given that it doesn't take a lot of money to commit terrorist acts.

Using data-mining systems to combat counterterrorism is more difficult than applying data mining in the commercial arena. For example, to flag people who may be committing calling card fraud, a long-distance company has extensive records of usage. As a result, there are profiles of all users. However, such convenient data are nonexistent when detecting people who might be terrorists. In addition, errors and oversights in the commercial arena are, in general, not terribly costly, whereas charging innocent people with suspected terrorism is unacceptable.

Biometrics will have to be a crucial part of any strategy in order to combat attempted identity theft.

Claire Broome
"Introduction by Session Chair"

Transcript of Presentation
Summary of Presentation
Video Presentation

Dr. Claire Broome serves as the senior advisor to the director for integrated health information systems at the Centers for Disease Control and Prevention (CDC). Dr. Broome oversees the development and implementation of CDC's National Electronic Disease Surveillance System, one of the highest priorities of CDC and the administration.

Dr. Broome served as deputy director of the CDC and deputy administrator of the Agency for Toxic Substances and Disease Registry (ATSDR) from 1994 to 1999; as CDC's associate director for science from 1990 to 1994; and as chief of the Special Pathogens Branch in the National Center for Infectious Diseases from 1981 to 1990.

Her research interests include epidemiology of meningitis and pneumonia; meningococcal, pneumococcal, and *Haemophilus b* vaccines; observational methods for vaccine evaluation; and public health surveillance methodology.

Dr. Broome has received many professional awards, including the PHS Distinguished Service Medal, the Surgeon General's Medallion, the Infectious Disease Society of America's Squibb Award for Excellence of Achievement in Infectious Diseases, and the John Snow Award from the American Public Health Association. She was elected to membership in the Institute of Medicine in 1996.

She graduated magna cum laude from Harvard University and received her M.D. from Harvard Medical School. She trained in internal medicine at the University of California, San Francisco, and in infectious diseases at Massachusetts General Hospital.

DR. BROOME: Good afternoon. Let me just go ahead and get started with the afternoon's first session. I am Claire Broome, medical epidemiologist at the Centers for Disease Control and Prevention where I have been involved in actually getting the data that you all would like to have as part of the targets for your modeling and simulations, and I have worked closely with a lot of the folks working on bioterrorism preparedness, and I will be moderating this session and what we will do is go through the first presentations and then try to pull some of this together and have a more interactive session during the discussant time period.

We hope there will be some time for questions to the presenters as we go along but that will depend on how much the presenters keep to time. So, the first presenter is Ken Kleinman from the Harvard Medical School, and he will be talking ambulatory anthrax surveillance, an implemented system.

Introduction by Session Chair

Claire Broome

Dr. Broome introduced herself as a medical epidemiologist at the Centers for Disease Control and Prevention and said that she is involved in obtaining a good amount of the data used in bioterrorism models and simulations. She is also involved with scientists who are working on bioterrorism preparedness.

Kenneth Kleinman
"Ambulatory Anthrax Surveillance: An Implemented System, with Comments on Current Outstanding Needs"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Kenneth Kleinman is assistant professor in the Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care. He serves as the main biostatistician on three CDC-funded projects to implement surveillance of health care system utilization in the Boston area and nationally, and he works with the national BioSense project. His interests include the analysis of longitudinal and other clustered data, epidemiologic methods, and missing data problems.

DR. KLEINMAN: Thank you. After the generality of this morning's talk I feel like this subtitle here is kind of grandiose, but I am a biostatistician and I am going to actually going to show you real data and describe a system that is running now in Boston, and there are a lot of people involved with getting the system running. It is actually a collaboration between the academic department, an HMO care provider and the State Department of Public Health.

So, all these folks are involved in various aspects from those places. My two co-authors on the statistical part of what I am going to talk about today are Ralph Plasers and Rich Plott who are infectious disease epidemiologists.

So, here is the outline for the talk. I am going to talk about why surveillance is important especially for anthrax and then I am going to talk about the data that we have and where we are, and I have tried to organize the talk around problems and our approaches to those problems that we encounter while trying to set up the surveillance system and where we should go in the future.

So, why is surveillance for anthrax important? Anthrax is what they call a biphasic disease and what happens is you get exposed. You have no symptoms for a

while. Then you have symptoms that are very non-specific and they resemble a cold or flu and that happens within a couple of days and actually we don't know a whole lot about anthrax because there have only been about 40 cases in the United States this century including last October in humans I should say.

So, there is not a whole lot known about it but it is supposed that most people have symptoms within a couple of days and then a day or two after you have those symptoms you start having really severe symptoms like severe sweating and breathing problems and eventually shock, and if you are not treated then there is death in 98 percent of cases and there is even death in some cases where there is treatment.

So, what are you going to do if you have anthrax? Well, nothing when you get exposed and when you start having symptoms you might go see your doctor and if you don't go see your doctor then or if you don't get diagnosed correctly then when you start having the more severe symptoms in the second phase you probably go to the hospital and get ciprofloxacin or another approved treatment.

So, there are a bunch of surveillance systems that are up and running now in the country that are based

on hospital surveillance, and what they do is they wait around in the hospital and they see if there are too many emergency room visits or too many diagnoses of anthrax and what can they do if they detect anthrax? Well, there are lots of people already in Phase II of the illness, and they are very sick and they are going to be arriving in the hospital in large numbers. So, at least you can be ready for them to come. So, you introduce some good if you do surveillance on that basis, but it would be better if we could detect people when they visit their doctor instead because that would be the earliest time that we would know about it, and also then we could just break out the drugs and prevent people from even entering Phase II and probably save them a lot of discomfort and problems and even lives.

So, our data, our study, as I mentioned we are trying to do this. Our position is located between a health maintenance organization and a provider group and also we have, we are in an academic department. I am in an academic department. So, the people who are part of this group, this care group and HMO, there are about 250,000 people in a certain area of Massachusetts, and that is about 10 percent of the population in that area.

So, what is very nice about this is that we can actually attempt to do that surveillance on the doctor's

visit before people get to the hospital because the provider group uses ambulatory medical records data and what that means is that every time you go to your doctor's office they actually have a PC in each examining room, and they will type in information about you including an ICD-9 diagnosis, and that is just some coding system for a diagnosis if you are not familiar with it, but it is symptoms and confirmed diagnosis, and they are continuously updated. They are centrally stored by the provider group, and we don't have to do anything different from what they are already doing, and it is standard practice to record all the diagnoses they make for each person who comes to see them,and that includes phone calls and nurse practitioners and physicians, and the system is actually a commercial system.

So, it is relatively easy for us to take our system which is in Eastern Massachusetts and transport it to some other locale where they happen to use the same medical records system and there are a bunch of them.

PARTICIPANT: Does this system store all the medical record?

DR. KLEINMAN: It stores all the medical record. Things like pharmacy data and test results aren't always updated in the same system. So, it is all the direct

patient contact between the physician and the patient. Does that answer your question?

PARTICIPANT: Lab results?

DR. KLEINMAN: Lab results are not --

PARTICIPANT: No, would have been a good answer to my question.

DR. KLEINMAN: Okay, I guess I wasn't clear enough on you question. Now, I am going to start talking about the problems we encountered when trying to set up the surveillance, and the first one is that it actually uses adult standard test for anthrax, and you have to get a chest x-ray, and I don't know enough about medicine to know what about this x-ray says that there is anthrax here, but if we actually waited for chests x-rays we wouldn't do any better than hospital surveillance because the physicians don't order chest x-rays for people who come in with cold symptoms.

So, we can't actually do surveillance for diagnosis of anthrax, and so what we do instead is we define symptom clusters or syndromes, and what we do is we say that anything that your doctor might say if you came in with the first phase of anthrax problems we are going to try to collect that, and we call that lower respiratory illness or I might say lower respiratory infection later

because those are the symptoms that characterize the first phase of anthrax infections.

That includes, I listed cough, pneumonia and bronchitis because about 90 percent of the diagnosis falls in one of those categories in this syndrome which includes about, I think it wrote it down here, 119 ICD counts go in there and that syndrome we actually borrowed from a Department of Defense product that was doing this before we were in a slightly different context.

So, in our data set in about a 4-year period there are about 120,000 visits that found this syndrome, and if you think about a natural disease unlike anthrax if it doesn't get fixed the first time you go to your doctor you are going to go back again, and when we looked at our records very nearly about one-third of them were repeat visits that were shortly after other visits.

So, based on the clinical expertise of the MDs working with us we were able to say that if it was within 6 weeks of a previous visit, the previous visit we weren't interested in it. It was probably the same illness, and so we throw those away, and what is left we call episodes of lower respiratory infection.

PARTICIPANT: How many don't visit with the same illness?

DR. KLEINMAN: I have no idea. This is a very small portion of the doctor's visits. I don't know that information. It is a completely different order of magnitude. I don't know the answer.

So, here is actual data and each dot here represents the count of visits for their respiratory infection on a given day in between January 1, 1996 and December 31, 1999 or January 1, 2000, and I just want to point out a couple of interesting features here that will come into play later, and the first is that you can see that during the winter there are big peaks.

People tend to go to their doctors for respiratory complaints more in the winter than they do in the summer presumably because they are more likely to have those symptoms and the other interesting feature is that you can see there are two trends here, and our clinics are actually open on the weekends.

So, this lower band is the weekend visits and the upper band is the week day visits, and finally, I don't know how clear this is from here but you can see these points, really low here in winter. That is Christmas and New Year's. So, people don't go to the doctor on holidays but they will go as often on the weekends, and they tend to go an awful lot more in the winter than they do in summer.

PARTICIPANT: What are the dots?

DR. KLEINMAN: I am sorry. Each dot is the count of the number of visits for LRI on a given day. So, this is about 130. So, there are 130 visits among our 250,000 numbers for lower respiratory illness on I don't know what day this is, probably like May 1, 1999.

PARTICIPANT: Oh, I see. So, only one doctor per vertical slice.

DR. KLEINMAN: Yes, that is right, at least in theory that is right. I am not sure that the resolution on the screen is up to that, but --

PARTICIPANT: Presumably the double banding begins in May?

DR. KLEINMAN: Yes. So, if can see data like this you might be thinking I should do a time series analysis and I need to include a covariant for weekend versus week day and maybe some kind of yearly cycle for seasons, and I think there are a couple of problems with this, and the first one is that you are kind of constrained to get a threshold for the whole area under surveillance, and the real big problem is you can send cipro to 2-1/2 million people. It is going to break the bank and plus there are bad side effects of cipro so you don't want to dose that many people who don't need it, and there are other problems

with dosing that many people, and the other problem is it is not very sensitive at all, and the best, I know people who are doing this, and the best that I have heard that they can do is, you know, if there are 10 additional cases on a day they might detect it, and that is an awful lot.

You don't want to have to have 10 initial people if you can avoid it. So, what we do to solve the problem is we actually geo code everyone, and geo coding just means that we take their address and we find out where on the map it lies, and it is actually remarkably easy to do, and accurate and cheap. We outsource it, and it costs us, I don't know 4 or 5 thousand dollars to do 250,000 addresses, and it comes back within a week. So, it is pretty slick. We only know the billing address. So, if people are exposed at the place where they work we are not going to know anything about that. We are working on trying to solve that problem, but it is kind of a feasibility issue because we don't really know where people work.

If they work for IBM they could be anywhere in the country instead of at some particular location, and that is the kind of work address that we would get is who employs them.

So, the coding changes a little bit, and so it is not perfect but it is a lot better than not knowing where in the area they live.

So, this is the kind of thing I get on a given day and here I took the plot of the number of people who had LRI on a given day and their location and I used some kind of three-dimensional smootherometer(?) or two-dimensional smootherometer and so, I am looking at these peaks, and you know there are probably one or two cases at this location.

PARTICIPANT: What are the axes?

DR. KLEINMAN: Sorry, this is latitude and longitude. So, it is an arbitrary grid on space but these actually, if you were to look at this latitude and longitude point you would be able to find it in the middle of Massachusetts Bay.

PARTICIPANT: Is it adjusted for population?

DR. KLEINMAN: This is not adjusted for populations, no. This is just raw counts, but if I were adjusting for population the question would be then, if I got five bioterrorism I would say, "Is this point too high? Is this bump too big?" and that is kind of the job that we are trying to do now.

PARTICIPANT: This was one day?

DR. KLEINMAN: That was just one day, yes. I have more slides that I could show about what happens. So, we have this data that is spread out in space and there is a whole bunch of spatial techniques. Most of them are developed for mining or for epidemiological exposures and so they don't actually think of time as an interesting dimension. They tend to summarize over time and say, "Is there a bump somewhere near say a contaminated well or near some sort of a high concentrated world where they are looking for something in it?" But we really need to know whether there is an event now as much as we need to know where it was, and there is a couple of so-called "special" techniques that are kind of intended for surveillance, and this touches on the same issue of the World Statistical Society, these two references and they are really designed to say, "Is there a bump right now?" and i don't think that either of these is really important for what we want to do either because they both assume that the time periods are similar to one another. So, they wouldn't be able to fit a covariate saying that this count happened during the winter; this count happened on Monday; this count happened on Christmas and be able to adjust for that, and they are, also, both designed for smaller data sets although I think that is probably not a big problem.

Now, the really big problem with these things though is that they are looking for clusters now and if you recall the problem that we have we are trying to say that we know there are clusters of flu because flu is contagious and flu symptoms are what we count. So, we know there is going to be a cluster on any given day especially during winter, but we know that there is going to be a cluster and we want to know whether the cluster we are seeing is bigger than the cluster we expect as opposed to whether there is a cluster, not whether there is a bump but is the bump bigger than the bump we expect based on the clustering there has been in the past.

PARTICIPANT: So, the difference is anthrax is not transmitted.

DR. KLEINMAN: That is true, yes. So, anthrax won't spread from person to person. So, what we have done here and I want to echo what some of the speakers were saying this morning, I am not saying that this is the best way to do this, but it is a way that seems to work, and that I was able to do and I think is easy to transport and those were qualities that I was looking for.

What I am going to do is I am going to again more or less arbitrarily break up space into localities that make sense to me and the ones that make sense to me are

census tracts. Census tracts are defined by the Census Bureau, but they are meant to be somewhat meaningful. You know, they don't cross a highway in general. They are kind of neighborhoods as you might define a neighborhood near your house and they are actually defined by people locally, and so, what we have now is a bunch of longitudinally repeated data.

We have a denominator and numerator for every census tract in our area over time, and what you do is use the generalized linear mixed model approach, logistic regression, and of course I could use a Poisson(?) regression instead, too, but I didn't do that. It is really just logistic regression, but it takes into account the correlations of counts on the same tract on different days.

So, I could fit the model using this GLIMMIX macro that actually is distributed with SAS. It is not officially part of SAS. So, it is somewhere in limbo there, but it is part of your distribution. If you use SAS you can find it there and that means that it is easy to explain to someone else how to do it and see that they get the same results you would on a test data set for example, and with this kind of data of course I am not sure, as an aside but I could use generalized nesting equations to fit the data but I would to get a separate estimate for each tract for

purposes of explaining what the results are, and just because I am a statistician I have to show some notation and really what I am doing is just a logistic regression. I am trying to model the binomial distribution for tract I in time T and I take a model with some covariates usually on the time and a random effect that applies to each census tract, and usually assume that those random effects are distributed normally in some variance.

That is all I am going to show you about that really. There is a lot of validity to the model that we get out of it. The observations for the winter months are much bigger than the ones for the summer months. They are bigger on Mondays and lowest on the weekends and the observations for holidays are less. So, what we are doing makes sense and the random effect might be features of each tract. So, if there are tracts where there are more people or there is more people that are prone to sickness or fewer people who are prone to sickness we are going to find that through these random effects and when we test for variance of those random effects we find out it is not zero. There really is a variability between the communities. So, it is worth doing that in the model, and if you are interested in interpretation you can actually estimate the random effect

that goes with each tract and those things are the odds ratios relative to the average tract.

Now, another question that was brought up this morning is what are you going to do with the data once you have got it and what are you going to do with your model once you have it, and we are working with folks at the Mass Department of Public Health and we want to be able to communicate with them when we think something strange is happening and what we do with that is we can actually invert the estimate for each census tract and turn it into a PHAT(?) for each census tract on each day and then we can calculate the probability of seeing as many cases as we saw or more and the P value for the null hypothesis of the data comes from the binomial distribution we got from the model.

Now, there is a problem with that. The big problem is that we have 520 census tracts in our area and the estimated P value for each one each day and that means we have 180,000 tests each year. So, that is a lot. The folks who were talking about data mining this morning, I don't know enough about data mining to know whether doing innumerable tests is a problem there. We are going to do, our proposal is to do a Bob Ferroni-like thing and what we do is actually we record the estimated number of years of testing required to see one P value as small as the one we

saw or smaller which is just the inverse of the number of tests we are doing times the P value that we got, and that is of course assuming that all the tests are independent which is not true, but it is better than nothing and one of the advantages of doing it this way is that when you get a big number it is bad, and we don't have to rely on people remembering that a small P value means something unusual.

We can say that we would only expect this to happen once every 100 years. We think people will probably understand that pretty quickly.

So, this is the point that actually we distribute it by means of a web site every day. We report the name of the town and the census tract number and I didn't mention earlier that those things are actually not determinants. Census tracts can bridge more than one town and of course multiple census tracts would be included entirely in one census tract. We also report the number of cases in that town or that census tract on that day, the number of insured people who live there and then the statistical -- I am actually going to work years between those counts and report the five most unusual counts and if it happens, if we expect those counts to happen once a day we just say that it has got a score of zero. It is really nothing to worry about at all.

Now, .004 years is like a day and one-half. So, we found a very interesting example, but there aren't more interesting days than that and actually the last time I did it it was something like this, I think it was March 21. So, I was able to say that a week ago we had this event that happens about once every 55 days, and that was infrequent enough that our department of public health is very interested and they sent a team, but that is a motion that we are trying to shorten where they contact people at the health plan or the care provider and figure out whether there is something really unusual about those cases or you know, if they all happen to be people in the same family with the same cold you probably wouldn't worry about it. So, they are able to investigate that.

What I really want to point out about this is that this is the number of cases we saw in that census tract on that day and if there are only four of them we can be very sensitive here. We don't have to wait for 10 cases to show up the way you would have to do in a regular risk analysis.

So, I am close to the end, and what I think is a really big unsolved problem for us and I have a very easy solution to it, but an event can take place over more than one census tract, and when it does our models aren't going

to know about that and so we rely on people looking at the following plot and this is automatically generated by the SAS program but we can say, "I think that these two census tracts are close enough together that I am concerned that the cause of that disease might be the same exposure to anthrax." It relies on someone actually looking at the plot every day. So, it is not a very good solution.

So, now I am on to what I think are four things to do here and again this is a very different scale of importance compared to the coin toss but I think this question of what to do about cases in adjacent tracts is a big problem and I think that one approach would be to incorporate some kind of secondary initial test and I think that it would be easy to incorporate spatial correlation in the census tracts but you are not going to change the answer much.

Flu is bad in some years and not bad in other years and it would be in theory easy to incorporate some measure of whether this season is a bad season for flu or a not so bad season for flu. Logistically it is hard to do that for various reasons, but I am not a cryptographer so I am sure that it is a great idea.

We are using census tracts as our neighborhoods. We could use census block groups. We could use a truly

arbitrary grid on space that wouldn't take into account anything special about the boundaries and this is something that definitely should be assessed and I think that it would be good if we could figure out how to incorporate individual level covariates because when a 20-year-old person goes to their doctor complaining of respiratory infection it means something a lot different than when an 80-year-old person does that.

Finally, I want to disclaim any collusion between me and the next speaker but I think that the biggest problem is figuring out how to simulate data because there is really no way to test the model sensitivity right now and we need to be, what we really need to be able to do is simulate the background noise which is flu and see whether we can detect and kind of solve the simulated data with anthrax and see whether we can detect additional cases. So, that is the end. I just have a summary which has the -- I don't think what we are doing is great. I think it solves some of the problems that other approaches that have already been developed don't solve, and I think that the most important thing to do is to try to assess the sensitivity of the model and we need to simulate a background in order to do that.

Thank you.

(Applause.)

PARTICIPANT: I just want to follow up on an earlier question about anthrax is not contagious. It didn't seem that the model incorporates that knowledge.

DR. KLEINMAN: I think you are probably right.

DR. BROOME: But what the model is trying to do is detect an increase that is different from your background expected. It doesn't matter, and in fact it is a good model for anthrax because if you got a point source for leads you will see cases all over the place and you want to pick them up as rapidly as possible. It doesn't matter whether they have then self-propagate.

PARTICIPANT: But the lack of self-propagation is a significant signature that would help statistically distinguish anthrax attack from just a coincidental --

DR. BROOME: If you have to wait until you are picking up second waves that would be the case, but in fact the idea is to try to pick it up from that first release.

PARTICIPANT: Oh, so, this is something I don't know. When people have the flu what is the cycle?

DR. BROOME: One of the problems with using flu for modeling is it doesn't have, you know, well, I will tell you what. Rather than get into all the details why don't we wait until we have heard from Stephen and Sally

because I think they raise a number of the issues around disease modeling and we can get back to this question which is interesting but I think trying to put it in the context of what are we trying to do with these different models would be valuable.

So, why don't we move ahead?

Any other sort of clarification questions for Ken? Why don't we try to do that at this point.

PARTICIPANT: Do you report your results to the disease center?

DR. KLEINMAN: We don't do any of the reporting ourselves, meaning when I say, we, I mean the academic department.

What happens to the data is it goes on the web site and I look at the plots and I send out an e-mail if I think there is something interesting about the plots, and the people at the Mass DPH decide what to do about it. So, they could decide to report it to the CDC or other organizations at that point.

DR. BROOME: The real question is have you notified alerts to the health department and has anything been found that had any public health significance. What we are talking about here is just a statistical signal and assessing the meaning of that requires investigation by

usually either the health care providers or by the public health department neither of which has a lot of excess people sitting around waiting to investigate, and that is why the point Ken was making about being sensitive to small numbers of cases but at the same time minimizing the number of false alarms is really a critical characteristic.

DR. KLEINMAN: And that is, also, why we don't send for the false alarm model, but I understand your question, but the answer to your question is there have been several times when a relatively unusual kind of event happened and that once every 55 days may have been the most extreme that has happened since we started running this in late October. That one may have been the most extreme but none of the ones that we have looked at have actually had public health ramifications, fortunately. So, we see about the number, approximately the number we expect of once-a-month events and because they are the expected events they are not interesting for public health.

PARTICIPANT: The communication is actually not the focus indicated.

DR. KLEINMAN: The communication has flaws. You know that is the most difficult part because it involves people on both ends and it requires someone at the DPH to be answering their e-mail and contacting someone at the

health care provider and that person to contact someone else to get the data on individual people. So, that is definitely something that we are working on, but it is not perfect yet.

DR. BROOME: Does that answer your question?

PARTICIPANT: No.

DR. BROOME: What are you --

PARTICIPANT: I am looking at the time that an incident would take place for it to reach the Communicable Disease Center because if you saw this happening in one location you might be getting information from other locations. So, I am interested in this time line from a suspicious event.

DR. BROOME: Well, to me that is an important question but it is secondary to knowing that this is actually useful information, and I think there is some very serious need for evaluation of whether this kind of very non-specific signal going off all the time is a useful thing to communicate. So, I mean we definitely want to hear about true signals, but in fact the critical need is to say does this have any potential real relevance. I mean just because you have got a whole bunch of flu patients who show up in a census tract, that happens a lot.

DR. LEVIN: So, if you are really interested in bioterrorism incidents wouldn't it be important to look at the genetic distance of the strains that are appearing from the conventional case from the year before because if the strains are only slight deviations from last year's in the area then it is more likely to have arisen naturally but if it has been an engineered strain --

DR. BROOME: I think there is a misperception here. We are not trying to find genetically engineered flu. We are trying to detect a non-specific febrile illness which might be misdiagnosed as flu. So, you are trying to separate out an anthrax, a febrile smallpox rash from the background noise of a normal influenza season and there is a whole lot more background noise than there is anthrax. Is that a fair summary?

DR. KLEINMAN: Yes, obviously a fascinating question and one that we wouldn't be able to say anything about.

DR. BROOME: I think people are really engaged and interested. Why don't we go ahead and get our presentations laid out because I think there is some very interesting information that will help frame the discussion, and Stephen Eubanks is going to discuss the

mathematics of epidemiologic simulation for response planning. He is at Los Alamos National Laboratory.

Ambulatory Anthrax Surveillance: An Implemented System, with Comments on Current Outstanding Needs

Kenneth Kleinman

Because the first symptoms of anthrax are similar to those of influenza, bronchitis, and other common illnesses, patients may or may not see their doctors at the onset of symptoms. However, with more severe symptoms occurring in the second phase of the disease, patients often seek treatment at a hospital. Currently, there are surveillance systems in hospitals to track anthrax occurrences.

However, if patients could be diagnosed with anthrax upon seeing their doctors with the first symptoms, they could be given treatment immediately, be saved a good deal of discomfort, and never have to enter the second phase of the disease. The Harvard Medical School, an HMO, a care provider, and the Massachusetts Department of Public Health have implemented a system to detect outbreaks of anthrax in the doctor's office instead of in the emergency room.

In this system, medical records are gathered by the provider for every patient visit and are continuously updated and centrally stored for use by the provider and the surveillance-system operators. The system tracks symptom clusters for lower respiratory illness, symptoms that characterize the first phase of anthrax infections. Of course, this includes the symptoms of other, much more common illnesses as well. The symptom clusters are compared with symptom cluster records of the same census tract from the past to assess whether a given cluster is larger than what is expected. A report is issued daily on the Internet that gives each census tract, the number of patients who live there, and the number of years between patient counts of the given magnitude. For example, if a particular count is so common it can be expected to happen once a day, it gets a score of zero. On the other hand, if we say that we would only expect this to happen once every 100 years, people will probably understand that pretty quickly. In this way, the doctors communicate with the Department of Public Health when they think something strange is happening.

There are a few ways in which the system could be improved. One improvement would be to determine whether results from several census tracts are really just manifestations of the same overall exposure event. Another would be to find the optimal type of basic geographical unit—census block groups, for example, or even an arbitrary grid—and compare it to the census-tract groupings that are currently used. It would be good if we could figure out how to incorporate individual-level covariance, because when a 20-year-old person goes to the doctor complaining of respiratory infection, it means something a lot different than when an 80-year-old person does that.

The biggest problem is figuring out how to simulate data, because there is still no other way to test a model's sensitivity. It is designed to detect outcomes that differ from the expected background, so the thing to do is to simulate the background noise—that is,

cases of flu and the like—and see whether the model detects any simulated anthrax data embedded in that noise.

Stephen Eubank
"Mathematics of Epidemiological Simulations for Response Planning"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Stephen Eubank received his B.A. in physics from Swarthmore College in 1979 and his Ph.D. in theoretical particle physics from the University of Texas at Austin in 1986. He has worked in the fields of fluid turbulence (at La Jolla Institute); nonlinear dynamics and chaos (at the Los Alamos Center for Nonlinear Studies); financial market modeling (as a founder of Prediction Company); and natural language processing (at ATR in Kyoto, Japan). He returned to Los Alamos as a staff member in 1987, where he has played a leading role in development of the traffic microsimulation for the Transportation Analysis and Simulation System (TRANSIMS) and has led the Epidemiology Simulation (EpiSims) project. His current interests include developing advanced technology for the study of large sociotechnical systems and understanding the dynamics and structure of social networks.

DR. EUBANK: Although I am very interested in providing a flu background for biosurveillance I haven't thought enough about the problem.

What I am going to talk about is an individual simulation that is rather different from the classical kinds of epidemiological models. I am using simulation here not in the sense of solving ODEs or PBEs but as a more agent based kind of notation.

I will describe the simulation that I am building. I will talk a little bit about why we use simulation in the first place and then I am going to try to give some mathematical questions that arise not just from epidemiological simulation but from any simulations of social technical systems we do, and then I would like to give an existence proof for a math program that has a very definite role in homeland defense in particular.

So, to dive right in the kind of epidemiological simulation I am talking about requires three components. One is an estimate of the social network of a large region.

Another is a model for the within-host progress of the disease once someone becomes infected and a third is a transmission model between people.

So, very briefly I will go over where these things come from. I am not going to spend a lot of time on

the process of simulation because once you understand what the components are it is pretty obvious how you can use it to build an epidemiologic simulation.

The reason we have gone in this direction of individual based simulation for response planning is that we have been asked to detect anomalous patterns which means looking at geographic distributions of disease by time for demographic distributions.

With that kind of information we should be able to identify the critical path a disease takes through a population either geographically or demographically and if you can identify a critical path then you should be able to evaluate the effectiveness of countermeasures that are targeted geographically and demographically.

So, we need demographic information and it just so happens that because of a transportation model that was developed at Los Alamos we have an estimate for every household in a large city, in particular Portland, Oregon, 640,000 households and what everyone in the family does every day. So, this is a particular person. This parent takes a kid to child care, goes shopping, picks the kid back up and goes home.

This is a completely trivial activity set. The actual activity sets we have are much more detailed. Since

we have this for every household in the city and there are only a limited number of locations where these activities can happen we can take a time slice in this picture and understand or gain an estimate of who is in contact with whom for how long, where, and what kind of activities they are doing.

I can talk a little bit about where those estimates come from because I think it is interesting and very different from the kinds of data sets other people have been developing.

Let me go ahead and explain the rest of the simulation models first. We have a very simple disease progression model, all we really need to know about the disease or what the effects on the individual who is sick are.

So, this is a kind of a cartoon of viral load or how many spores you might have or something like that. It is time dependent. We can vary the rate to growth rates according to an individual's demographics and we can list the times what is happening to the person by place these thresholds on the disease.

So, in this case someone can clear the disease completely if they don't receive a big dose. They become symptomatic at some level. You can change these thresholds.

You already have the thresholds around. It is very important for the actual dynamics of the epidemic and someone can either die or recover.

This is to reinforce the notion that this is not some point estimate of what the disease looks like.

PARTICIPANT: Is it true that the threshold is the same as the actors underneath? What is the load?

DR. EUBANK: No, this is not really the actual load of virus in the body. So, if the threshold is not necessarily the same for entering as it is for leaving this is just the time rates so that you become recovered at the right time regardless of what the actual load in your body is.

We can assign different growth rates as I said for different people so we can get a distribution of effects of disease on a population. With a representation like this I can add in the effects say of post-exposure vaccination by shifting some of the margins around. This particular disease has several different, well, it is smallpox. So for a very old major infection you could have several different presentations of the disease and one of them is an early hemorrhagic one which often involves death much sooner than most.

PARTICIPANT: So, I guess the water supply is appropriate.

DR. EUBANK: That is a lake in this water supply. That is the social network and the disease in the host. So, there is a transmission component that is very important. The traditional models for epidemiology require something called a reproductive number. I will talk about this a little bit more in a minute but it is basically what happens if you introduce an infected person into a susceptible population. How many cases is that index case going to cause?

The reproductive number is an output of our simulation. It is very important to understand that. The probability is that given two people are in contact and one of them is infected you need to know the probability that the other person becomes infected, and it can be a function o how close the contact is, how long it takes, the demographics of the people involved, what kind of activity they are performing.

Now, if you go talk to an epidemiologist they are going to want to know what this is in your simulation. They is quite a body of research into what that number should be, but if you talk to a clinician or more importantly if you talk to a patient this is what they want to know. I

want to know when I can send my kid back to school if he is still infectious. Should I go get a chickenpox vaccination, something like that?

Then you compare a little bit more from these two approaches. Traditional epidemiological models, the next speaker may correct me on this but I class them all as basically coupled rate equations. The attempt to break down a population into very coarse groups, subgroups based on one or two demographic variables, and here I have shown age, and we assume some sort of mixing between these subgroups which is not really very well understood and then you make this assumption about the number that you have so many susceptible and so many infected, it is the rate.

I am going to argue that reproductive number is --

DR. CHENEY: Why are they broken up? I mean this is a basic question but why are they broken up into those age groups? What if you broke them up into different age groups, would it not be --

DR. EUBANKS: I picked these completely at random without looking at traditional models, but for a flu model, for instance kids may transfer flu among themselves more rapidly than older people.

PARTICIPANT: The mixing rate is simulated?

DR. EUBANKS: The mixing rate would be, yes, how many times infected kids have an influence on the parents, but that means you know something about the structure of the population and how many times kids are in contact with older people.

The reproductive number, it really involves two different things. It is person-to-person transmission characteristics and the disease and social mixing patterns, and you can't observe it in isolation. You can't go to a population and just say, "What will the reproductive number for this population be?" without an understanding of how the population mixes.

So, in contrast to this approach what we are doing is to represent individuals each of whom carries some set of demographics which is much more finely resolved than in the other models. We are estimating contact rates between the individuals and the whole population and what may or may not be fairly important here is that these contact rates are estimated independently from the disease spreading problem.

So, in the traditional epidemiological model you could say that we will fit the mixing rates and we will fit the reproductive numbers and we will reproduce a given epidemic that we have data for and generalize from that.

These contact rates were estimated for a transportation problem which means that the validation and verification issues are very different, but most importantly the reproductive number emerges as it does in the real world from transmission characteristics between individuals and the mixing of those individuals.

PARTICIPANT: If I have a disease that only gets transmitted by kisses, I don't --

DR. EUBANK: There are some diseases this won't work for. This is not very good for HIV, but for something like influenza, colds and so on. I will tell you where the contacts come from and then you can make your own judgment about how well they represent the disease.

PARTICIPANT: But I thought that this model would get at others. That was my question. The contact rate must depend on the disease.

DR. EUBANK: Transmission during the contact depends on the disease.

DR. LEVIN: But what constitutes a contact?

DR. EUBANK: That is what I am about to tell you. What constitutes contact depends on the disease. Very often it depends on duration and proximity of physical contact.

The next few slides are actually fairly apropos of the previous session's data mining, I think because they

illustrate a way to take data from very disparate sources and build some information of use to other processes like epidemiology.

So, this is a description of where we get our contact patterns from. We take census data down to the block group level and we build a synthetic population. The property of this population that is important is that if you were to do a census on the synthetic population you would match the results of the actual census in an actual city.

Also, there are no real people in our population. We have no problems with privacy because we don't have any real people to complain about it.

If you did have, say, the NSA had data on everyone's movements during the course of a day and maybe we could get German data on everyone we could plug that in here just as easily, but we felt that it wasn't politically acceptable for Los Alamos to go tell people they wanted to know what they did all day.

The next piece of data we used is activity surveys that cities tend to collect on their population. They send out diaries and ask people what they did every day.

We have about 2000 activity surveys for Portland, Oregon. We use those as templates and we matched households to households in the activity survey to find out which of the templates we want to use for that household. A template might be something like that cartoon I showed at the beginning of the talk.

We remove from the survey all geographic information because 2000 templates is not enough to tell us where everyone is going. So, we make the assumption that transportation infrastructure constrains how people move through the city. It seems like a fairly reasonable assumption. We take travel time estimates for the transportation network and generate possible locations for each of these activities to be carried out.

The final step, what we set up then is a game. Everybody in the simulation is trying to minimize their travel time to fulfill all the constraints that their activity structure imposes on them, but they don't know what everyone else is doing except that everyone else is trying to minimize their travel time. So, we play this game on the computer and we simulate the traffic and results which updates the travel times which can update the activities and certainly update the routes that people choose.

When all this settles down we have an estimate for the contact panels. So, we put all that together and take the census data activity surveys and network the infrastructure data and we produce epidemic curves. Whether the curves are realistic or not depends on the assumptions we put in about the disease transmissions. This particular one shows vaccination response. We vaccinate people who are known to have been in contact with infected people. We can get rates of vaccination, numbers of people you would need to go around giving vaccinations, numbers of people you would need to go around as contact persons,numbers of people who become infected on a daily basis, all those kinds of things that you would expect have epidemiologically.

DR. CHAYES: So, how come the vaccination is bimodal?

DR. EUBANKS: That is probably just a reflection of the incubation period. The disease has about a 10-day incubation period..

In addition to those kinds of traditional responses and results we can produce a demographically distributed picture. So, for instance this is the probability of distribution with age, given that you were

infected during a specific week after an incident occurs, and this doesn't tell you as much as you might think.

It is saying that in the first week the age distribution depends directly on exactly what the release scenario was. We happen to have a scenario where we release at a shopping mall. So, there are not many school-age kids there. There are some infants and there is a bunch of people you might expect to be at a shopping mall, and the distribution itself is relaxing as the weeks go by to a marginal distribution for the population, but there are some interesting question you could ask about this.

One is what is the rate at which this distribution is relaxing and is it the same for all the different initial conditions. That would tell you something about when it is possible to apply these mixing models and just ignore all this detailed simulation.

You could say that within 4 weeks the simulation is indistinguishable from a model assuming uniform mix here of the population and that gives you a good time development on use of the simulation.

We are, also, hoping to use this for targeting strategies for vaccinations, and it is not clear yet exactly how you would do that.

So, I won't spend more time on the simulation. I will briefly go through some of these points. Why is it that we are doing the simulation aside from the fact that my group is called basic and applied simulations lab. I think it is a very natural thing for homeland defense in particular.

If you are talking about defending critical infrastructure you are really talking about having people and resources and interactions among people and resources are local. Even given telecommunications somebody has to interact locally with their telephone on their desk in order t o place that call to someone else.

People move around and the resources often move with them and moreover defensive actions can change the way people move. So, what you really need to know of social networks is where the resources and people are and the function of time, possibly in hypothetical circumstances.

So, just going out and measuring the way things are today is not going to help you understand what might happen if someone takes your advice and institutes a policy for defending the infrastructure.

Another reason to use simulation I think is that it is a way to fuse, you know everyone knows about the data fusion problem, but there is another problem of fusing

information that is completely different in time. I am thinking here of something like demographic tables with procedural constraints that could be very vague like quality of life influences the decision or fairly specific like vehicles aren't going to physically pass through each other and it needs to take into account things that are completely asymmetric, irregular. The only reason for them being the way t hey are is some, it is contingent on some completely unknown past history.

As an example I use the development of the social network that I went through briefly before. Suppose you have these demographic data tables and you had a land use network. I would like to know what the implication of this constraint that vehicles can't pass through one another is, and I really don't know of any statistical modeling technique that can tell you that. It is because the representation of the problem isn't quite large enough whereas this individual based simulation is a very natural representation and I can tell you that instead of the, like this represented the unweighted demographic tables. This represents the transportation infrastructure and the implication of this is that that person is going to arrive at a particular place at a particular time.

Of course, it is not exactly clear what that means and so this is a simulation and that person doesn't exist. Simulation is really a deductive process that lets you untangle the entailments of hypotheses you make that may seem to be very unrelated. It is a very powerful technique, not in the statistical sense but in the sense of just what it can do, but the nature of the estimation that it makes is not so clear.

What does it mean that that person on simulation arrived at that place at that time? I don't know. I think this is a very interesting mathematical question and what I would like to describe for you briefly is a few other interesting mathematical questions that arise in this simulation and then a program that attempts to address some of them.

So, in addition to that kind of philosophical question about what simulation means there are very specific problems arising in our simulations of sociotechnical systems that I would to just very briefly mention.

The first is we get questions in the form of well, here is what, you guys know what the social network looks like. Here is what the disease looks like. Tell me, how am I going to prioritize vaccination? To me that means

I have got some graph or network, and I need to identify the critical paths on this graph, and analyze what happens if I cut those patterns.

I tell that to my computer scientist friends and they say, "Oh, that means this particular problem that I know about in computer science, although I don't know exactly how to apply it to your network." So, then we have to have a little discussion about how to take results that he knows about and produce something that is relevant to my needs.

On the theoretical side if you think computer science is too much to worry about but I would like to know about results about very structured random graphs, we have a long history of theory on random graphs that have say a Poisson degree distribution where you pick links at random to turn on and off, and recently there has been a lot of work on small world methods that have power law degree distributions and these arise in places where first order clustering statistics are very important.

The graphs that I am working with have a very different degree distribution, number of neighbors of each. This is a log-log plot and there is certainly some structure there but it is not structure that people have studied before.

So, now, very briefly I am going to describe a math program. This is my existence. The first thing we had to understand was an axiomatic framework for what a simulation is. We derived a three-element object which is local functions on a network with a particular order of evaluation of the functions.

So, we went through the whole network evaluating this. There is a computational theory associated with this and I am just glancing over this because again, like the other speakers have said I am not proposing this as the end all and be all of mathematical science.

I am just trying to give you the flavor of a math program that has a very direct influence on homeland security. The kinds of research that go on in this program are characterization of dynamical systems, deriving morphisms between different dynamical systems, between different objects, inductions from one to another, algorithm development, specification simulations and extensions to other things as required by the applications that we are developing.

One reason I agree on this point is that the morphisms and the relative relations between simulations is that it is very hard to go about validating these kinds of simulations and what we are trying to do is approach it at

a more basic level, understanding when two simulations are the same even if you are exploring different parts of the phase base. These things operate in huge dimensional probabilities bases and you can't possibly generate enough instances to explore the space base. So, you need to understand something about the structure of the simulation itself before you can examine your validation.

So, focusing between the systems allows you to address knowledge and also to generalizability. The mathematical underpinnings let us know when it is even reasonable to suspect that we can build a simulation to do some of the things we want to do. A lot of our sociotechnical development depends on composing different simulations like the traffic simulation with an epidemiology simulation.

We can understand something about the process of composing simulations with this approach, and it ha direct applications to some of the network and drafting problems that I mentioned.

I think I am completely out of time. I can give you a long list of applications and theory to go with it but just to reassure you that it is possible to do this kind of work and not be completely shunned by the mathematical community this is a list of where we publish

SDS papers but I would like to mention on the computer science side and the math side we have papers come back from a journal saying,"That is a wonderful paper, great results. Please remove all reference to applications." And if you want my opinion as to what mathematics can do for participating i homeland security change the culture a little bit.

Math and society can have something to do with each other but the research must be driven by application, and those speakers before me have commented if you are just doing research in a vacuum it is not ever going to work, and the results have to be communicated back to people like me who can understand what the implications of your research are for very applied problems.

Thank you.

(Applause.)

DR. BROOME: Any quick clarification questions?

Thank you, Stephen, and I think that certainly the point of this workshop is to try to get some discussion amongst both mathematicians and those who might use their research.

Mathematics of Epidemiological Simulations for Response Planning

Stephen Eubank

An epidemiological simulation can be used to detect anomalous patterns. It requires three components: an estimate of the social network of a large region; a model for the within-host progress of a disease once someone becomes infected; and a model of transmission between people. Dr. Eubank then spoke about each component of his simulation dealing with disease spread in Portland, Oregon, emphasizing the uniqueness of the transmission model that he uses.

Dr. Eubank and his colleagues essentially inherited the first component—a database of the 640,000 households of Portland, Oregon, with estimates of what everyone in every family does every day—from a transportation model that was developed at Los Alamos.

For the second required component, the researchers have a very simple disease-progression model that incorporates essentially all they really need to know about the disease or what the effects are on the individual who is sick. Thresholds that can be varied according to individuals' demographics include pathogen-growth rates in the hosts, whether or not they clear the disease (if the dose is small enough), at what point they become symptomatic at some level, and whether or not they ultimately recover or die.

The third component—transmission between people—is quite different from traditional models for epidemiology, which require as input a "reproductive number." This number quantifies what happens if you introduce an infected person into a susceptible population: how many cases is that index case going to cause?

The reproductive number is an *output* of the simulation. The reproductive number involves two different things—person-to-person transmission characteristics of the disease and social-mixing patterns—and the latter can't be determined in isolation. So in contrast to the traditional approach, researchers are estimating contact rates between individuals and the whole population, and what may or may not be fairly important is that these contact rates are estimated independently from the disease-spreading problem. What is important is that the reproductive number emerges as it does in the real world—from transmission characteristics between individuals and the mixing of those individuals.

Dr. Eubank noted that while the simulation he'd been describing obviously has a very direct influence on homeland security, it is also a vehicle for research. The elements of this program are characteristic of dynamical systems, and from the program can be derived morphisms between different dynamical systems, between different objects, reductions from one to another, algorithm development, formal specifications of simulations, and extensions to other things as required by the applications being developed.

In closing, Dr. Eubank noted that he and his colleagues have had papers returned to them from mathematics and computer-science journals: " 'This is a wonderful paper, with great results,' editors would say. 'But please remove all reference to applications.'. . . So if you want my opinion as to what mathematics can do for homeland security," he said, "change the culture a little bit so that it's not bad to have applications. The speakers before me have also said that if you are just doing research in a vacuum, it is not ever going to work."

Sally Blower
"Predicting the Unpredictable in an Age of Uncertainty"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Sally Blower, Ph.D., is a professor of biomathematics at the University of California at Los Angeles. She is a mathematical and evolutionary biologist whose research focuses on developing models of transmission dynamics. She uses these models as health policy tools: to design epidemic control strategies for a variety of infectious diseases, to understand and predict the emergence of antibiotic and antiviral drug resistance, and to develop vaccination strategies. The main focus of her research is to develop the study of infectious diseases into a predictive science. Recently her work has focused on HIV, tuberculosis, and genital herpes. She has also pioneered the application of innovative uncertainty and sensitivity techniques (based upon Latin hypercube sampling) to the analysis of transmission models. These techniques enable transmission models to be used to predict the future with a degree of uncertainty and to identify which parameters are critical in determining which future outcome will actually occur.

DR. BLOWER: Thank you. So, I have been told to stand rigidly behind here and not move an inch.

Okay, so in overview the material I want to cover then is to say a little bit about transmission models and their uses and to spend most of the time talking about uncertainty and sensitivity analysis to explain to you what it is, what it can be used for, some examples from work that we published on HIV and then to talk about the problem with risky vaccines, again, some work that we have done on HIV identifying vaccine perversity points, and you will have to wait to find out what those are and then talk a little bit about how that is related to smallpox, as we have a somewhat similar problem as we have currently a risky vaccine to deal with smallpox, and then I will go through as far as I know the complete literature of mathematical bioterrorism which is four publications at the moment.

Next?

So, what is a transmission model? It is a series of mathematical equations that specify the transmission dynamics of an infectious agent, and you can use either stochastic and obviously have probabilities in that versus deterministic, and both are useful.

You can either have simple or complex and also one person's simple is another person's complex. The only thing to remember is that a complex model should always reduce down to a simple model. So, you end up with more results when you have a complex model, but they shouldn't be totally different results.

So, the way that I use models in my group at UCLA is basically as health policy tools, and what we do is we predict the future with a degree of uncertainty by using uncertainty analysis. So, mathematical models can be used to look at the effects of current treatment conditions.

PARTICIPANT: Could you speak a bit louder?

DR. BLOWER: Okay, right, I will pretend I am talking to my children. So, to look at current treatment conditions and to look at what if conditions. So, what if we did something and what if we had a vaccine?

Second use of them is to define perversity thresholds based upon drug resistance. When you treat individuals you get drug resistance. So, treatment has a good effect. It reduces drug sensitive strains. It has a bad effect. It increases drug resistance. So, you can look at trade-offs, and that is essentially what I mean by perversity thresholds when you make things worse than before you started.

So, you can look at changes in treatment and risk behavior and also risky vaccines, and then we can, also, identify the most effective prevention strategy for an epidemic by using sensitivity analysis.

So, models can predict the future and mathematicians can write down equations fairly easily with a great deal of certainty. The problem becomes for all of these models we need specific parameter estimates and once you start to actually go to the literature and to talk to your infectious disease friends that is when you realize there is a problem and there is a great deal of problem with parameter estimation uncertainty even with diseases that we have been studying for hundreds of years like tuberculosis.

So, when it comes to new agents there is going to be even more uncertainty. So, how we can deal with this is by using uncertainty and sensitivity analysis based upon something called Latin hypercube sampling and this allows us to predict the future with a degree of uncertainty, and I will show you some examples, but essentially all you need to know about this is the parameter -- can we go back? The parameter estimation uncertainty is being translated into prediction estimation uncertainty, okay?

So, just to talk a little bit about models and perversity thresholds when you are using treatment or vaccines there are approved by the FDA because they will help individuals. So, they work at the individual level but you can have individual good but population level harm and this is just referring to a paper that was published in 1996 in Science looking at tuberculosis and drug resistance and showing that the more you treat, if you have a high degree of drug resistance which actually happens in some parts of the world you are actually making things worse.

Next?

Another thing to be aware of, another trade-off is basically risk behavior and perversity. Again, this is another paper I published with Angele McLean in 1994, in Science and a paper we published in 2000 in Science, both looking at HIV and trade-offs between if risk behavior increases and you have got treatment or vaccine how you can actually end up making the epidemic worse.

So, you need to have your epidemic interventions coordinated. So, what is uncertainty analysis? Uncertainty analysis is used to assess the variability which you can also call prediction imprecision in the outcome variable that is due to the uncertainty in our input parameters.

So, uncertainty analysis can be based upon Latin hypercube sampling. So, the question becomes what is Latin hypercube sampling, and that is a type of stratified Monte Carlo sampling. So, it is a statistical sampling technique that allows you to simultaneously vary all of the values of all of the input parameters.

So, you don't just vary one at a time as with traditional methods or keep them all constant. You can vary everything at once and Latin hypercube sampling was first proposed by McKay, Conover and Beckman in 1979, to aid in the analysis of nuclear reactor safety. So, they were concerned with a problem of nuclear reactor meltdown, obviously a very important problem and they had very large models and the parameters were uncertain, so, a somewhat similar problem to what we are thinking about today.

PARTICIPANT: When you mention hypercube and matching the color, is that the same thing?

DR. BLOWER: Exactly but in hyperspace. So you just think in more dimensions. It is exactly the same conceptual principle.

So, what is sensitivity analysis? We then couple the uncertainty analysis with sensitivity analysis and use it to identify the parameters in the model that are

increasing the outcome variables that we are most interested in.

For example, if you are looking at the probability and the severity of an outbreak that is our outcome variable, and then by doing a sensitivity analysis we could tell which of the parameters are going to increase this.

DR. CHAYES: Can we ask you why would you be using a Latin hypercube because in a Latin hypercube you don't want to have the same color occur in --

DR. BLOWER: Exactly.

DR. CHAYES: -- direction. What does that correspond to?

DR. BLOWER: It is actually a very efficient way of sampling the parameter space. So, instead of doing a full factorial in which you look at all the possible values of all the parameters and all the possible combinations it allows you as Latin squares does to do a much fewer number of simulations but makes sure that you have sampled the entire parameter space, and if you just did it randomly you could have very funny subset.

So, it is a stratified random sampling method, okay?

Could we go back?

So, to talk then about some specific examples so you can get a more concrete idea of this one thing that we have done is to predict the probability and the size of an outbreak occurring, and basically this is based on the basic reproductive number or R zero or R naught depending what side of the Atlantic you are from and so we have done an uncertainty analysis of R zero for HIV and basically if R zero is greater is than one an epidemic outbreak occurs and obviously for HIV this has already happened.

So what we have done is basically the reverse of this saying that if we had enough treatment and could we actually make R zero less than one,. So, it is basically the same principle, okay?

So, to do this uncertainty analysis we first of all used an HIV transmission model. You write down equations for the transmission of HIV. Then you analyze that to calculate an analytical expression for R zero. Then you estimate probability density functions for all of the parameters in the model. You basically get maximum and minimum upper and lower bounds for them. You then use Latin hypercube sampling and then calculate 1000 values for R zero and therefore you then have a measure for R zero as an uncertainty estimate.

Next?

And this shows you what is on the Y axis here. So, on the Y axis is the value of R zero that we calculated and this is for three different conditions and these are box plots.

So, the line in the middle shows you the median and then the main part of it is 50 percent the predicted data and then the whole thing is the 100 percent of it.

So, you end up with an estimate of the outcome variable, but you can see exactly the prediction imprecision, the uncertainty in your estimate. Okay?

Then we did sensitivity analysis to determine what is driving the value of R zero, and in this particular example it is the change in risk behavior is the most important parameter in the model. So, again the Y axis is R zero, the value of it and on the X axis this shows part of the sensitivity analysis. The parameter here is the change in risk behavior and you can see as that increases and decreases the value of R zero changes and that this is again uncertainty analysis. So, you have basically got a predicted cloud of data.

Next?

So, another example for you to see how uncertainty analysis can be used is what we did here is the title of the paper was predicting the unpredictable, the

transmission of drug-resistant HIV, and we published that in Nature Medicine last year.

The idea here is that currently combination antiretroviral therapy is being used widely to treat the HIV epidemic in the United States. There is, also, a lot of drug resistance arising.

At the moment there are very few data sets to actually say what is actually happening and so we did a model to predict what is likely to happen and therefore called it predicting the unpredictable.

So, we predicted a variety of things, the number of HIV infections prevented over time, the number of new cases of drug-resistant HIV arising over time, the prevalence of drug-resistant cases over time, overall prevalence and then the number of infections prevented per drug-resistant case that arose. So, this is sort of a biological cost/benefit analysis, how many things that you have done that are good, the number of infections prevented versus sort of the number of things you have done bad which is generate drug resistance.

So, you can use these models to work out biological cost/benefit, and then obviously you could add economics on top and do an economic cost/benefit analysis.

So, to see some of these results, this is the result showing number of HIV infections prevented if you started treating an HIV epidemic.

So, on the Y axis is the number of new HIV infections that are prevented. On the X axis is the fraction that you treat and it is the uncertainty analysis. So, this parameter is an uncertain variable parameter and we allow this to vary between 50 and 90 percent, and therefore we predicted 1000 different epidemics over time.

The blue predicted data are after 1 year. The yellow are after 5, and the red are after 10 years, and here you can see the more you treat, if you go along the X axis, the more HIV infections you prevent, and since it is an uncertainty analysis you get again the predicted clouds of data.

This shows the predicted transmission that we predicted going from 1996 up to 2005, for what we predicted how much drug resistant HIV would be arising and as I said there are no data sets, no published data sets. We did this for San Francisco, and the red lines are the median. The predictions are box plots. So, you can see 50 percent of our predictions lie in the rectangles.

After we made these predictions there was one datum point that arose which is the green. So, that

actually occurred after we had made the prediction. So, you could say that this made the predictions look fairly good, but there is only one piece of data, but it was collected afterwards.

Anyway, next?

This then shows again uncertainty analysis predicting the prevalence of HIV. Again, the treatment rate is an uncertain parameter varying between 50 and 90 and this is what would happen if you treated an HIV epidemic after one year of combination antiretrovial therapy.

The drug sensitive strains come down, the white data at the top. The drug-resistant strains come up which are the red strains at the bottom, and it is a function of treatment and you can see these predictions are with uncertainty.

This is what happens after 5 years. You can see that the treatment rate has the same effect but more so over time, and then finally after 10 years.

Next?

And actually this is what we are predicting the HIV epidemic will be like in the San Francisco gay community in 2005 which is pretty bad.

Okay, this shows predictions for the prevalence of the HIV prevalence again as a function of the treatment

rate and you can see again what happens over time, year one, after 5 years, after 10.

Then these are the results of the biological cost/benefit analysis in which we have predicted what would be again the effects of combination antiretroviral therapy treating an HIV epidemic, the number of HIV infections you prevent for each drug-resistant case you generate, so basically a trade-off, a biological trade-off, and these effects change over time.

So, these are our predictions from 1996 to 2005. So, then the other part of this was to do the sensitivity analysis and this uses the results of the uncertainty analysis that we generate and then we calculate partial correlation coefficients and this basically shows that one of the key factors in increasing the amount of drug resistance was the treatment rate which is not surprising.

This shows the unadjusted data but the partial correlation ren(?) coefficient is .9. So, these are the unadjusted data and then the other key factor revealed by the sensitivity analysis was the actual biological fitness of the drug-resistant strains and again these are the predicted unadjusted data.

When you adjust this statistically for it then you get partial ren correlation coefficients of about .9.

So, how good does a vaccine have to be? If we are worried about smallpox and other biological agents we can fairly easily, this is sort of a classical result in the literature, fairly easily work out how good a vaccine has to be, and the three important components are the vaccination coverage, how many people you need to vaccinate. The E is the efficacy of the vaccine, how good is it and R zero is the basic reproduction number. So, basically how severe is the problem; how severe is the epidemic?

So, if you have good estimates for those or again you can do an uncertainty analysis on it and work out for the size a basic reproduction number how good the vaccine has to be and what kind of coverage, how much, what percentage of the population do you need to get to take the vaccine.

Live attenuated vaccines are actually used, what we have available for smallpox, polio and measles. They have many advantages. They generate a very high protective efficacy because they are essentially generally the same pathogen but attenuated. They are low cost and they are simple immunization schedules.

So, live attenuated vaccines are very nice and are used a lot, but let us just think about some of the

problems for that. The problem with risky vaccines, live attenuated vaccines is that they can actually cause the disease in some people. Therefore they have a benefit in which they are very, very effective, but they also have a risk.

So, these type of vaccines, we call them risky vaccines because they are risky, and they can under some circumstances do more harm than good which is why people are worried about the smallpox vaccine at the moment. If you use mass vaccination with a smallpox vaccine there would be some very detrimental effects in large numbers of immunocompromised individuals, HIV infected individuals or immunocompromised because they have cancer in many of the urban centers, and there are hundreds of thousands of the people in the United States. So, it is a big problem.

We looked at this in terms of HIV vaccine and this is a paper we published in PNAS last year looking at live attenuated HIV vaccines. Basically this is an ordinary differential equation model and allows us to look at the effect of a vaccine that would be very effective and protect against infection but in some people because they are vaccinated will actually cause AIDS though at a much slower rate than if they had gotten infected with the wild type.

So, if you had this live attenuated HIV vaccine would it be a good thing to use in a public health campaign? This was the question we addressed and you could actually, we have got web bush(?) into this model running and it is very user friendly. My children can actually run this, and so if you want to you can go to our web site and put in parameters and see how the model behaves for different coverage levels and different safety levels. So, that is something you may want to do.

What we did is predict what would happen in Zimbabwe where there is a very severe HIV epidemic; 25 percent of the population are currently infected and in Thailand where very much fewer, much less severe epidemic and compare and contrast between these two countries using exactly the same hypothetical HIV vaccines to see what would happen.

So, again, we used time-dependent uncertainty analysis and we specified the parameters such as efficacy, such as coverage levels by probability distribution function and then we made some predictions.

If you actually look at analytical results, and we also looked at analytical results what you find is that all of the, we tested 1000 different HIV vaccines theoretically and on the computer. We tested them and found

if you looked at it analytically all of these thousand vaccines would lead to HIV eradication, and it would be a two-step process. You would replace the virulent strain, the wild type with the avirulent vaccine strain and then you would have to stop the vaccination program and you would have eradicated HIV.

So, that is what you find out if you look at it analytically, and so equilibrium results are very important but it is important to look at the transient dynamics, too.

So, we did that, and we found something that we have called the vaccine perversity point. Now, the vaccine perversity point is defined in terms of the vaccine safety level. So, in the case of HIV this is the fraction of vaccinated individuals who progress to AIDS as a result of the vaccine strain. So, it would be the same thing if you were looking at a smallpox vaccine.

At the vaccine perversity point then the annual AIDS death rate with a live attenuated HIV vaccine in place is equivalent to the annual AIDS death rate without a live attenuated HIV vaccine and this shows our predicted results for what would happen if you put an HIV vaccine out in Thailand. On the Y axis is the total annual AIDS deaths for 8 per 100,000. On the X axis is the safety of the vaccines that we were looking at for HIV for them causing anywhere

between 1 percent and 10 percent of the vaccinated individuals to progress to AIDS in 25 years. So, these are fairly safe vaccines, and then again it is an uncertainty analysis. So, you have got predictive clouds of data.

Sorry, can you go back?

So, at time zero we start off with the sort of light green data, after 10 years the pink data and then after 50 years the yellow data and 200 years turquoise. We have just plotted the 200 years to show you that basically you are almost at the equilibrium after 50 years. Things don't change that much. What you can see here is there is very clearly a vaccine perversity point that is the vaccines that cause 5 percent or more of vaccinated individuals to progress to AIDS in 25 years actually raised the death rate. They make things worse. So, these vaccines in Thailand if you used them would actually make the HIV epidemic worse. If you had live attenuated vaccines that were 5 percent or less they would actually reduce the death rate and be a good idea.

The same vaccines in Zimbabwe as you see there is no vaccine perversity point. The time zero data are much higher. We have a much more severe AIDS epidemic in Zimbabwe and these vaccines behaving in exactly the same way actually because the epidemic is so severe, their risky

effects basically aren't seen and therefore they are just seen to be beneficial, though obviously the safer the vaccine the more you lower the death rate, but there is no perversity point. You can't make things, frankly, any worse in Zimbabwe than they are already.

DR. REINGOLD: Are you assuming this for now?

DR. BLOWER: Yes, but you could also put it at any point and do the same analysis.

DR. REINGOLD: So, you could make the epidemic worse without treatment over, without vaccine over time?

DR. BLOWER: You can, yes, definitely. This is just to assess the effects of the vaccine.

So, in Zimbabwe none of the vaccines resulted in perversity. In Thailand vaccine strains that caused more than 5 percent of vaccinated individuals to progress to AIDS in 25 years led to perversity and if you just want to think about testing these vaccines, HIV vaccines in clinical trials that is an enormous problem.

So, whether or not a vaccine will cause a perverse effect will be situation specific and depend upon the

risk of becoming infected and this really then is the big problem with small pox. Since we have a risky, dangerous vaccine, if it was used now in a mass vaccination campaign

in the United States, it would obviously do more harm than good. I think everyone is in agreement about that.

It should only be used basically if there has been a major release of small pox and we are convinced that the number of death, therefore, are going to be greater than the number of deaths that would occur without using the vaccine.

So, to go through the mathematical bioterrorism literature, actually the first paper is back in 1760 by Daniel Bernoulli. I think it is little read paper. It is written in French, but you can get a translation. I haven't read it in the French. I have read it in the English.

So, it is an attempt at new analysis of the mortality caused by small pox and of the advantages of inoculation to prevent it. So, Bernoulli was very interested in doing health policy research back in 1760 and outcomes research and all the buzz words that are now used. The idea was to use a model to actually show that using vaccination against small pox was worthwhile.

So, we are basically now doing -- it is always worthwhile to look back in the literature.

The next paper that I came across -- there may be more out there, but these are the only ones I am aware of -

- is in 2001 by the CDC, Modeling Potential Responses to Small Pox as a Bioterrorism Weapon. Then there are two more papers. The paper by Ron Brookmeyer and Blade in Science in 2002, Prevention of Inhalation Anthrax in the U.S. Outbreak, and a recent paper that just went on line in PNAS 2002, looking at early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales.

So, just to summarize, even in an age of uncertainty, I believe prediction is possible.

[Applause.]

MS. BROOME: Clarification questions for Dr. Blower?

MR. KAPER: I am Hans Kaper, National Science Foundation.

You haven't given any specifics about your models, but I gather they are systems of ordinary --

DR. BLOWER: Yes, they are systems of ordinary -- and if you want actually any of the papers, they are downloadable from my web site and all the papers have the equations in them.

MR. KAPER: The reason why I am asking is that in the mathematics community there is quite some experience

now that large systems -- I was wondering whether anything has been applied to your models?

DR. BLOWER: To my models? Maybe Simon knows about that.

MR. LEVIN: The person I would turn to Mac Kliman(?) because otherwise I haven't seen anything like that applied to any of these models.

MR. KAPER: It seems to me that there is a fruitful area for applied mathematics to get involved.

MR. LEVIN: I have a question, Sally, which is when you are looking at these cost benefit tradeoffs in terms of -- is there any notion of discounting them in terms of how long in the future you push the -- or how late in one's life?

DR. BLOWER: I think that is very interesting and actually there was some work done actually by Hans Waller back in the seventies, who worked for the WHO until he got fired for doing some of those such things. But basically then you have to decide -- I mean, the reason I keep away from economics is you have to say basically how much life is worth, first of all, and then you have to discount, and future people are worth less than current people.

So, you definitely could do that. But we haven't yet done that.

MR. LEVIN: Not just future people, but maybe delaying to later in one's life when the death occurs.

DR. BLOWER: Oh, increasing life expectancy.

MR. LEVIN: Yes. You may have the same number of deaths, but it may be when people last longer and die --

DR. BLOWER: We have looked at that, that basically that you avert because, yes, you don't prevent -- basically in the Science 2000 paper for HIV, you can show that you avert AIDS deaths and that that is a significant effect and, therefore, very worthwhile, as well as the number of infections prevented.

Predicting the Unpredictable in an Age of Uncertainty

Sally Blower

Mathematical transmission models, whether deterministic or stochastic, can be valuable in predicting the dynamics of an infectious agent. Dr. Blower and her group at UCLA use them as health-policy tools in two basic ways: to assess the effects of particular treatments and vaccines on current or what-if conditions and to define "perversity thresholds" beyond which the use of a nominally beneficial technology actually makes things worse, as in some cases of drug-caused drug resistance. In the latter circumstance, you can have individual good but population-level harm.

For all these models, mathematicians can write down equations fairly easily with a great deal of certainty. But there is a problem with parameter-estimation *un*certainty even with diseases that we have been studying for hundreds of years, like tuberculosis; and with new agents there is going to be even more uncertainty.

Analysts deal with this situation by using uncertainty analysis and sensitivity analysis based upon something called Latin hypercube sampling, which is a type of stratified Monte Carlo sampling. Essentially all you need to know about this technique is that parameter-estimation uncertainty is being translated into *prediction*-estimation uncertainty.

Dr. Blower described some applications of these techniques in her group's work on HIV infection. These include determining the probability and size of outbreak, the prevalence of disease and of drug-resistant cases, and the perversity point of a live vaccine under different circumstances.

She summarized her talk tersely: "Even in an age of uncertainty, I believe prediction is possible."

Simon Levin
"Remarks on Detection and Epidemiology of Bioterrorist Attacks"

Transcript of Presentation
Summary of Presentation
Video Presentation

Simon A. Levin is the George M. Moffett Professor of Biology and was the founding director of the Princeton Environmental Institute at Princeton University.

Dr. Levin is a member of the National Academy of Sciences and the American Philosophical Society and a fellow of the American Academy of Arts and Sciences and the American Association for the Advancement of Science.

Dr. Levin has also served as president of the Ecological Society of America and has won its MacArthur Award and Distinguished Service Citation. He was the founding editor of the journal *Ecological Applications* and has edited numerous journals and book series. He is past chair of the board of the Beijer International Institute of Ecological Economics and is also a member of the science board of the Santa Fe Institute and the vice-chair for mathematics of the Committee of Concerned Scientists.

MR. LEVIN: I will say, Sally, that the one thing that I was amused by was the notion that user friendly was something my kids could do on the computer because that was by far no guarantee that it would be something that I could do. What I would like to do, if we can get the slides on, is to first of all touch on some of the points we heard, review them and put them into context, maybe try to identify some challenges, including some topics that haven't been hit and finishing up with sort of a pet hobby horse of mine, which has to do with system level aspects, indeed, the notion of developing a system level immune system.

As we have heard, especially in the last two lectures, there is a large classical literature in epidemiology and, indeed, in much of mathematical modeling and infectious diseases, the focus has been on the sorts of things that Sally just talked about and we have talked about before that; namely, developing predictive models that will help us to understand what the spread of an infection will be once introduced into the population.

There are some model aspects of this part of the problem that are associated with bioterrorism. You have also heard -- this is sort of the classic framework, the

compartmental approach, in which the population is broken into classes, such as susceptible, infected, recovered and then maybe into age groups or risk groups.

Some compartmental analysis, but as you also heard, one can develop individual-based models. The unique aspects are that in the case of bioterrorism, one may not be dealing with a single introduction. Typically, we would ask what happens if a disease is introduced at low levels in a particular environment. What is the potential for spread, et cetera? But if bioterrorism could involve multiple introductions several different places, it could also involve the introduction of multiple agents. Diseases that might interact with each other might affect our response capabilities by moving resources one place and then hitting the system with resources somewhere else.

So, one is going to have to extend these approaches to the modeling of multiple agents. Indeed, they don't even have to all be infectious disease agents. It could be different kinds of terrorist attacks coupled with bioterrorism. So, these are different sorts of notions that is in the same framework, but involves our understanding much more complex systems.

But as we have been hearing, especially in this afternoon session, prediction is not the only thing that we

need to be concerned about. There are all sorts of challenges for mathematical models, from dealing with either chemical or biological threats, ranging from prevention to detection, from prediction to what we do after an introduction, remediation. How do we respond?

So, I will identify at least four different categories dealing with infectious disease modeling and I am sure there are others, most of which we have heard about, but they relate to how do you detect the surveillance systems about which we heard this morning, prediction. The response systems once a disease appears in the population and tracing, that is, how do you trace back to identify where the source is, something which we have seen doesn't work very well in the case of the anthrax attacks, but has been used, for example, in the foot and mouth disease in the U.K. to trace back to what were the original sources.

That introduces inverse problems, sometimes very simplistic inverse problems, but imagine anyway that you have a model for the spread of a disease, how do we run it backwards to try to identify the sources or can we do that? Is that a much harder problem?

Now, all of these things and I will focus here briefly on prediction, at least two different aspects, what

I will call a priori and a post priori, namely, there are things that you could do in advance to prepare you before there has even been an attack and these would include running scenarios using mathematical models. A very important use of mathematical models is in exercises like Tupoff(?) or Dork(?) Witter(?), which have been exercises in which one uses a model to drive a theoretical epidemic and allow people who in response systems to do what they would do if there were a real attack.

It is like a flight simulator. It is a learning experience by which you learn about where the weaknesses are in the system and you give people experience with responding to it. So, models not only are going to be useful when there has actually been an attack, but they can be extremely useful in training people or helping you to identify where you ought to be putting the resources and what is the best way to set up response systems. There have been a few exercises of that sort, but not a lot and typically the models that have been used have been fairly unsophisticated in driving the epidemics in those systems.

So, detailed models such as Steve told us about, for example, could be extremely useful in these training exercises. Then there are ones after the fact. The

difference between single threads and multiple threads or single focus and multiple ones, I have already mentioned.

I will just shove these in after listening to Sally because first of all, I wanted to emphasize what is sort of a central concept of the use of models of this sort and then to identify some of the particular new problems that we may need to be thinking about. Sally talked about r_0. That is the number of secondary cases per primary case in a naive population. Notice I called it r-naught. I speak British.

So, the criterion usually here is that r_0 is greater than 1, then the introduction will spread and, therefore, to control it, as she showed us, you are going to have to reduce r_0 below 1. In the sorts of models that she talked about, r_0 typically can be broken up into the mean, infectious time per individual, multiplied times the average number of secondary infections per unit time while an individual is infectious. So, one can look at this. Beta would be the probability of transmission per individual and it is the population size. Nu(?) would be with death rate and gamma would be the recovery rate and, therefore, understand that there are various ways to control the disease by focusing either on improving the recovery rate -- you could also increase the death rate,

but that is probably not a good way to do it, but some sort of removal rate that would reduce the infectious time per individual. Isolation would be one way to do it.

Secondly, to reduce the mean infectious transfer per individual. For example, we do that with sexually transmitted diseases by the use of condoms and other ways by isolating individuals who are infectious and in itself is important, that is, the population size is important and one way we get at that is through vaccination to reduce the size of potential infectious individuals.

But this very useful concept has some problems associated with it. It is a deterministic concept. So, for stochastic models, it is only an approximate threshold. If there are heterogeneities in the population, it may be only an estimate and in particular if there is heterogeneity in terms of the number of infectious transfers per individual that -- if some individuals are more sexually active than others or more likely to transmit the disease, then every individual in a sense has her own r_0. Some individuals are much more likely to spread the disease and, therefore, where you introduce the disease becomes important.

This is an average quantity and how to deal with that in a heterogeneous population, such as Steve told us

about, is a challenge. Another challenge in the individual-based models, which relates to this question is how do you reduce the dimensionality of the system? Can you take that large individual-based model and collapse it through some sort of moment closure or other techniques, through a system of equations that can be dealt with analytically.

So, that is another mathematical challenge, which is to take these large scale, individual-based models and try to develop essentially a statistical mechanics that would allow us to do some analysis.

MS. CHAYES: Are you looking at non-stationary, as well as stationary r_0 there? You said that r_0 need not be constant, that it can vary from one individual to the next.

MR. LEVIN: Absolutely. It varies spatially and it varies over time. That is certainly the case.

MS. CHAYES: So, you study non-stationary models.

MR. LEVIN: Non-stationary models are studied and are -- the people who do this sort of modeling study -- but certainly the predominant literature focuses on the computation of one r_0, but you are absolutely right. There are many examples of diseases which have been introduced and didn't take, depending on the season. So, r_0 is very

much a function of when the introduction takes place. In the case of vector diseases, for example, ones that are transmitted by mosquitoes, r_0 depends a great deal on the season of the year and whether the mosquito population is at high density.

The classic example or a classic example is the introduction of mix(?) mitosis into Australia to control rabbit populations, which didn't take for a number of years and then suddenly after a number of wet years, it just took off and spread and temporarily wiped out the rabbit population. So, r_0, certainly can be temporarily --

Now, in addition to the spread of infection over time, there is also the problem of spatial spread and there are two dimensions to that. One is the spread of the agent. For example, in the case of anthrax, there are lots of models that have been used, essentially variance on calcium plume models for how an agent with particular characteristics could be spread.

But secondly there is the sort of spread that we have heard about today, which has to do with the infectious transfer of the disease, not applicable to anthrax, but to many other diseases, from one individual to the next. Most of them, near as I can tell, most of the modeling of spatial spread that is in use in our response agencies has

to do with the first element here, the spread of agents, the transport of materials in the environment.

But there is obviously great potential for understanding the temporal and spatial spread due to infectious transfers. There are a couple of different dimensions to this. For example, in the case of foot and mouth disease, which has been a current hot topic for the last year and a half in the U.K. Issues of spread relate both to the spread of the agent, the possibility that the agent itself could be spread by airborne methods over very large distances and, secondly the spread due to the movement of individuals, due to the movement of cattle or sheep around the country, therefore, cutting down -- therefore, spreading the disease.

The first thing that goes into effect when a new case is discovered is or would be shutting down movement of animals around the country because it is that rapid movement that spreads it. So, understanding spatial spread is crucial. That has multiple scales to it and the same explanations may not be true on all scales.

For example, influenza typically could be approximated by a kind of diffusive spread once it is introduced into a particular community, but as the Russian Rovotchev(?) and Irolin Jenie(?) and others showed some

years ago, entirely different models would be needed in order to understand the spread among cities and that would relate on the airline traffic. Once one built that model, it might be useful for other sorts of diseases that would spread.

You have got to keep updating that model as airline schedules change, however. So, there are multiple scales of transmission.

That was prediction. We heard a good deal about detection methods, largely statistical methods in which they are involved, surveillance and how to discriminate a signal from the background. The use of inverse methods, tracing, how do you run models backwards, as I said, in order to identify, for example, what the sources are or in order to identify, as we have heard in the talks this afternoon, there is also the problem of identifying what the parameters are, which brings me to response and some of the -- as I said, the hobby horses I have.

Again, these are responses that can be done before the fact or after the fact, but, in particular, in dealing with the responses to terrorist attacks, we need to recall that we are not dealing with the same sorts of problems as we are with accidental introductions. We are first of all dealing with a game theoretic problem in which

the terrorists are trying to outsmart us and understanding what our strategies of a response will be.

Obviously, even work such as this that we are talking about that gets done informs potentially terrorists. So, it is really a game theoretic situation in which you are dealing potentially with multiple agents about which -- whose introduction may be unpredictable and, therefore, it is crucial to develop responses to have systems that are adaptive, that have the capability to respond to unpredictable attacks.

In other words, to develop essentially an immune system for the whole system. So there are a whole class of operations research questions that relate among other things to how do you deploy individuals, how do you respond to the system, how do you seal it off. For example, modularity is an important part of the transmission system. We heard about small world networks. The particular typology of interactions has a great deal to say about the potential spread. How do we reduce modularity? Could we do targeted vaccinations, et cetera?

Can we reduce contacts in certain places in order to make the system less susceptible to propagate an unexpected attack?

Why don't I stop there.

MS. BROOME: Thank you. I think we are trying to stick more or less close to time so there will be some opportunity for further discussion.

Remarks on Detection and Epidemiology of Bioterrorist Attacks

Simon Levin

There are many challenges for modeling disease prevention, detection, prediction, and response in the case of bioterrorism. Typically, we would ask what happens if a disease is introduced at low levels in a particular environment. But in the case of bioterrorism, one may not be dealing with a single introduction. Bioterrorism could involve multiple introductions in several different places, and it could also involve the introduction of multiple agents. Indeed, they don't even have to all be infectious-disease agents—different kinds of terrorist attacks could be coupled with bioterrorism. These diverse possibilities are all more or less in the same framework, but they require an understanding of much more complex systems.

One important tool to prepare for an attack is running scenarios using the mathematical models. In this way, they are like flight simulators. These are experiences by which you learn about where the weaknesses are in the system—where to put resources—and they can be extremely useful in training people.

Particular areas that should be studied include these:

1. Spatial spread, related to the spread of disease over time. Most modeling of spatial spread involves modeling the spread of the agent, although analyses and actions relating to the mobility of individuals have proven effective in such epidemics as foot and mouth disease

2. Outbreaks related to terrorist attacks as opposed to accidental introductions. These are game-theoretic problems, in which the terrorists are trying to outsmart us.

Arthur Reingold
"Remarks on Detection and Epidemiology of Bioterrorist Attacks"

Transcript of Presentation
Summary of Presentation
Video Presentation

Arthur Reingold, M.D., M.P.H., is professor of epidemiology and head of the Division of Epidemiology at the School of Public Health, University of California, Berkeley (UCB). He holds concurrent appointments in the Departments of Medicine and Epidemiology and Biostatistics at the University of California, San Francisco (UCSF). He has devoted the past 20-plus years to the study and prevention of infectious diseases in the United States and in various countries in Africa, Asia, and Latin America, initially at the Centers for Disease Control and Prevention (CDC) for 8 years and at UCB since 1987.

Current activities include directing the National Institutes of Health-funded UCB/UCSF Fogarty International AIDS Training Program, now in its 14th year, and co-directing the CDC-funded California Emerging Infections Program, now in its 8th year. Dr. Reingold's current research interests include prevention of transmission of HIV in developing countries; the intersection of the HIV/AIDS and tuberculosis epidemics; malaria in Uganda; emerging and reemerging infections in the United States and globally; sexual transmission of hepatitis C virus; vaccine-preventable diseases; and respiratory infections in childhood.

DR. REINGOLD: Thank you. We are going to go from power point to slides, back to overheads.

I am going to actually focus my comments primarily around the context in which Ken is doing the work he is doing about detection of bioterrorist events and outbreaks. This is a topic that I first had the opportunity to think about and work on when I was at CDC working with Claire and we were thinking about how to detect meningitis outbreaks in the meningitis belt as soon as possible in order to rush in with vaccines and reduce morbidity and mortality.

The first point I would make, I think, is that detection and rapid response to outbreaks is more or less the same whether or not the outbreaks are naturally occurring or intentionally introduced. At least from an epidemiologic perspective, most of the issues are the same.

What I would like to do in the few minutes I have is try and give you some context for the complexity of what Ken is trying to do in Massachusetts and what others are trying to do. You may have seen the picture of George Bush looking under a microscope in Pittsburgh a couple of weeks ago, looking for outbreaks in Pittsburgh. This is a challenging thing to do for a variety of reasons.

So, let me just first of all give you a quick definition of surveillance in terms of what we are trying to do in terms of detecting diseases or illnesses. Those of us in public health generally refer to surveillance as a systematic ongoing collection analysis and dissemination of health data with findings linked to actions in the decision-making process.

The implication here fundamentally in terms of bioterrorism is we would like to have a system in place that would permit us to detect illnesses quickly and then be in a position to respond rapidly to keep morbidity and mortality to a minimum. Now, in terms of surveillance needs for bioterrorism, there really are two phases, I think, one can look at. One is pre-event and the other is post-event or release. In the case of pre-release or pre-event, obviously, what we are talking about is can we detect the event and can we detect it earlier than we would have in the absence of our system.

Then, of course, in terms of post-release, there are some important issues around being able to map the event and guide the response, make projections about the numbers that will be affected and other things of that kind. Now, the reality is most of the emphasis has been on the first phase of this, the detection of the event. I am

going to spend most of my time talking about that, rather than about the second.

Now, I think it is worth pointing out since many of you are not physicians and not working in public health, how it is we in public health generally do detect outbreaks in the absence of fancy systems and computers and software and hardware. The reality is that generally we detect many outbreaks because an astute patient or family member notices that, gee, aren't a lot of us who ate at the church supper or who all went to the same wedding all vomiting more or less at the same time.

It doesn't take an enormous lot of smarts to figure that out and frequently the health department learns because someone calls and says I think we have an outbreak here among people who were in such and such a place or participate in such and such an event. Similarly, we learned a lot of outbreaks from an astute clinicians, who recognize the one or more cases of an illness are more than they should be seeing and they are smart enough to let the health department know about that.

If you could say, there is an intrinsic comparison of what the expected is, but it is primarily because that individual or that individual and someone they talked to on the elevator come to the conclusion that they

are seeing too much of something that this whole process has set in motion. Sometimes it happens because an astute laboratorian looking at specimens in a clinical microbiology laboratory notices the same thing.

On rare occasions, it is actually those of us in public health who are collecting and analyzing surveillance reports or laboratory reports, who detect outbreaks that otherwise had not been detected yet or might have gone undetected. Clearly, what we would like to do with these fancier systems we are developing is see if we can be a larger part of this detection process, as opposed to waiting for clinicians and patients to tell us about the problem.

Now, I want to point out, this is the epidemic curve. These are the cases. The X axis here is the date. The Y axis is number of cases and for epidemiologists this is what is called an epidemic curve, showing the cases of anthrax from last fall.

What I would like to point out is a couple of things about this actual bioterrorist event that we now have experience with. The first is that these cases were primarily brought to our attention by astute clinicians; that is to say, a doctor was able to look at a patients and say I wonder if this person has anthrax based on the

clinical diagnosis and then report that to someone and collect appropriate specimens.

I am sorry to say that I think the systems that many people are building to try and detect various syndromes would never have detected this particular outbreak because there was one case here and one case there, because in this case, the cases were related to mailing of letters and the anthrax spores ended up in various places.

So, if you are looking to detect increases in patients with febrile illnesses, this outbreak probably would never have been detected by that kind of system. I think it is fair to say that a lot of the activity around detection of outbreaks in terms of anthrax in particular is assuming that some plane will fly over New York or Washington, release a cloud of spores or an individual with small pox will walk through Washington Dulles Airport and infect various people and then a whole thing will unfold and large numbers of people will become ill.

This particular event obviously had different features, which would not have been readily detectible by the systems that many people are building. So, what are the various approaches people are taking to improving surveillance to detect bioterrorist events? Well,

obviously, as Ken has already said, one approach to this is trying to monitor visits to health care providers and while we do a lot of that already through hospital admissions for various diseases, many people are trying to add a component that looks at outpatient visits and intelligently doing it in places like HMOs, where people already collect and enter these data, where the hardware and software things you need to add to the system are minimal or non-existent.

People are interested in looking at energy department visits, clinical microbiology laboratories, indicators, things such as 911 calls, over-the-counter drug sales, absenteeism at work and at school. Some people are even considering direct monitoring of samples of the population through Nielson rating type setups, in which people are routinely answering questionnaires over the Internet, a whole host of different, creative approaches. I think we need to give some attention to the question of how likely is it that these kinds of systems will actually improve our early detection of bioterrorist events.

Well, Ken has already referred to the fact that one of his problems is influenza and these are the kind of monitoring data we have on influenza and have had for a number of years. This just simply looks at the number of -- the proportion of all deaths due to pneumonia and

influenza in the 121 largest cities in the United States and you can see here that that fluctuates with season because influenza is a seasonal disease.

Now, one thing I would take issue with Ken a little bit about is influenza is not our only problem in this regard. There are many other viruses that can cause similar illnesses, parainfluenza, adenoviruses, microplasma, which is a bacterium. So, it is not just the problem with influenza, but there are many things perturbing the background that we can monitor, but certainly will make it more challenging to detect something like inhalational anthrax at an early stage or small pox before the rash erupts.

This is also just to point out that, in fact, we currently do have surveillance systems in place and have had for a long time for influenza, not just on what proportion of deaths are due to pneumonia, but an outpatient visits to sentinel physicians and to a number of other indicators that all pretty much tell us that influenza is a seasonal disease and that all tend to peak at approximately the same time.

I would really like to point out and you may not be able to read this very well is people in France a number of years ago were interested in the question of how well do

these various indicators allow us to say when an outbreak of a febrile respiratory illness is occurring. So, what they looked at here, what we epidemiologists call the sensitivity, the specificity and the predictive value of a positive of these various indicators to say when has an influenza epidemic occurred.

They were looking at everything from emergency visits to sick leave reported to the National Health Service, sick leave reported by companies, visits to general practitioners, hospital fatalities, a whole host of drug consumption, many of the things that we are considering looking at now and measuring how sensitive, specific or the positive predicted value. By sensitivity we mean of all the outbreaks that occur, what proportion will be predicted by this indicator.

By specificity we mean of all the times there is not an outbreak, what proportion of the time will the system tell us there is not an outbreak and most importantly, the predicted value of a positive of all the times the system tells us there is an outbreak, what proportion of the time is it right and there really is an outbreak.

That is important because if we envision rushing in with vaccines or very labor and time and other intensive

things in response to some indicator, we would like to be doing it in response to a real problem most of the time, rather than response to a false alarm.

You can see that, in fact, some of these various indicators have reasonably good sensitivity and specificity and predicted value positives and others don't. I think these are the kind of indicators that we need to look at as we are trying to judge how good a job our detection methods are doing.

So, when we think about detecting outbreaks, these, it seems to me, are the things we need to think about. I have mentioned three of them. The sensitivity of the system to detect a bioterrorist event, the specificity in predicting a bioterrorist event, the predictive value of a positive, what kind of response will we make to a false alarm. What I would point out to you is if you set up these systems in many counties or many parts of the country, in any given county, in any given week, there will be zero bioterrorist attacks. Therefore, every positive -- everything above the baseline that stimulates a response, if all you are looking for is bioterrorist events, a hundred percent of the events you detect will be false positives in most counties most of the time.

So, response you are going to make requires a lot of thought because most of the time it will be response to something that is not the bioterrorist event. The timeliness is obviously critical because we want something that will tell us about something sooner than we would learn about it otherwise and I think that is the real challenge in this. Certainly cost used to be an issue. Maybe it is not so much anymore.

So, let me just end -- I think this is the last one -- there are a lot of things we need to think about in figuring out how to affect bioterrorist events. Ken has referred to monitoring people who have a febrile respiratory illness and that is how things like anthrax and small pox will present initially. But botulism and a host of other things would present with different things. So, we need to consider what syndromes to monitor, not just febrile respiratory illness, but possibly other syndromes.

What kind of case definitions to use, in order to determine who does and doesn't have the illness, whether we are looking at inpatient or outpatient or other kinds of data. Who will enter the data may be less of an issue in an HMO, but it is a big issue in other health care provider settings? How will the data be transmitted to people like Ken in order to analyze them? How often will the data be

examined? What statistical parameters will be used to signal a possible event and what actions will be taken in response to that signal?

So, I think these are all some of the complexities that we need to, those of us working in public health need to think about and have to have some pretty concrete answers for it. So, let me stop there so we can proceed to general questions and discussion.

MS. BROOME: Thank you.

Any clarification questions for Dr. Reingold?

[There was no response.]

We can then open the session for more general discussions. I would like to just remind folks that the overall session was on detection and epidemiology of bioterrorist attacks. So, even though Ken and Art have focused on the detection issue and I think there is a number of threads there that we could pursue, I do hope that we also pay attention to the epidemiology; that is, what is the distribution of an outbreak once it occurs and what is the strategy and effectiveness of the response to that outbreak. I think some of the work that Steve and Sally discussed are relevant in that area.

So, I would like to direct folks thinking to both how can mathematics be useful on the detection side, but

also what are some of the contributions that we get from simulation and modeling in these other areas.

So, let me see if we have some general comments or questions for the panel or among the panel. I know you all have been thinking about this a great deal.

PARTICIPANT: The question is, I guess, for Sally.

To what extent are the models that you are constructing -- what kind of data is there that can actually --

DR. BLOWER: For the HIV models --

PARTICIPANT: Yes.

DR. BLOWER: For the HIV vaccine models, all we can do is make sure that they fit the current data because there are no vaccines out there and nothing has been done. So, those you can't test and those models are used more as what could happen and this was not to show whether it would be a good idea or not.

For the HIV ones, we are actually just finally getting some data from San Francisco. They have been collecting it, but they haven't actually analyzed it or released it. So, we will actually be -- I am working with Rob Weiss at UCLA. He is, you probably know, in biostatistics and his thing is longitudinal data analysis.

So, we will actually be fitting the models to data and trying to do so using new statistical methods, not just, you know, minimize squares and things. So, we haven't yet done it because of the data.

PARTICIPANT: Also a question for Sally.

In your sensitivity analysis and uncertainty analysis, how large are your systems and how many parameters do --

DR. BLOWER: How large are the systems -- I am not quite -- the system, there are about five to eight, depending on which model we are using, ordinary differential equations and they have probably about 12 parameters. Obviously, the different models have different numbers of parameters. When we do a sensitivity analysis, we put all of the parameters in. What you generally find, though, are two or three parameters that are driving the whole system.

PARTICIPANT: This is a question for Steve. In social network theory, one of the key difficulties they have always had with that literature is a goodness of fit test to determine whether or not the models that they have had for social network formation are, in fact, adequate descriptions of the data. It seems to me that what you have would be a true or empirical goodness of fit

evaluation that could be useful for them and then also useful for evaluating whether or not those sort of simple models can be applied to the -- structures that -- disease spread in populations. Would that be something fun to do?

MR. EUBANK: Yes. Basically, I agree. It is not just epidemiology. As I mentioned we can take the implications of the network structure that we have, things like traffic or mobile communications or energy use, all these different areas and validate against them all.

One slide I was going to show, but didn't was showing the results of the traffic simulation, which estimates traffic counts you would observe on different -- on particular streets in the city. We can verify those against collected counts or just do simple hypothesis testing. Does our traffic look like observed traffic? But it turns out that depending on what we constrain for, we can fit to -- we can have a better fit or a much looser fit. So, we could fit to traffic counts. Works fine, but we might not fit to demographics of the people who are traveling down those roads or it might not fit to the purpose of the trip, the data that is also collected.

Part of the problem we have though is the observations are so difficult to come by. In the case of epidemiology, for example, I would love to work on flu and

get a typical background of flu season in the city, but the reporting requirements on flu are loose and then we need a model for who went to a doctor, who even let it be know that they had flu to anyone.

It is kind of hard to convolve all that into the validation of these models.

PARTICIPANT: I have a question both to you and to the panel in general.

When you had your contact graph, you compared it to the random graph and to the social network graphs. We know how to -- general random graph. We also know how to generalize the social network graphs, using -- attachment. We don't really have good mathematical models, it seems, to generate this kind of degree distribution for your contact graph. So, my question is, one, are there ideas how one could generate those and, two, would it be of any use for antiterrorism to do so?

MR. EUBANK: To answer your second question, yes, I think there would be a big use for antiterrorism to do so.

How do you generate them? I think what I would need to do is understand what structural properties in the network are important. Maybe it is not even a degree distribution that is important, but something else. Then

turn the mathematicians and figure out to generate those kind of graphs.

For instance, for the small world graphs, Strogatz and Watts decided to look at triangles, numbers of triangles in a graph compared to what you might expect in a random graph. Once you have done that, it is easy to understand how to generate particular kinds of small world structures.

PARTICIPANT: So, what would be the structure?

MR. EUBANK: I don't know yet. But that is an open question that I would hope mathematicians would be able to give me some advice on.

MS. CHAYES: You would probably also be very interested in some type of vertex cover because if you had a vertex cover, then you could go in and knock out the disease at those points and wipe it out. Right? So, that would be a minimal -- the cheapest, most effective way of wiping it out.

MR. EUBANK: These graphs -- because we have both people and locations, they are bipartite graphs. If you could find a cover of the locations, that would guarantee you caught every person in the city or every person who was infected, it would be wonderful.

MS. BROOME: Although I think one of the issues is also separating out the different modes of disease transmission. If you have got a common source outbreak where you have got contaminated food or contaminated water, you know, there are some diseases where a person-to-person spread and contact is important or respiratory droplet spread is important, but there are a number of others where that is not the issue.

I have a question for Ken. I was interested -- you know, you are applying your aberration detection tool to influenza-like illness and there is an obvious reason for that and you also do have a lot of historical comparable data. But another approach that is taken for early detection is to try to pick markers that might be more specific. For example, somebody who shows up in an emergency room with a high fever and has a blood culture taken might select for more high probability, more specific likelihood that this is at least a severe disease. There are lots of other things that can present that way, but it might increase the yield.

But obviously, you would have much less substantial background and I wondered how good your statistical approach was, you know, in trying to pick up other types of signals.

MR. KLEINMAN: Actually, I presented the talk as if all we were doing was anthrax surveillance. We also actually run a very similar system for upper respiratory complaints and upper and lower gastrointestinal complaints and we get a lot more visits for upper respiratory complaints than for the lower respiratory and a lot fewer for either of the gastrointestinal things.

My sense is from looking at the way the models worked in the past six months when we have been running it, that it doesn't work as well for the rarer events. The events that you are talking about would be rarer still. We have some ideas about how we might do that kind of surveillance anyway, but they haven't been tested yet. So, I don't want to say anything more about them.

MS. BROOME: But I mean I do think as a general statement the issue of what types of aberration detection approaches will be useful is going to be highly dependent on, you know, what you are looking for, the specificity, the kind of -- basically the historical data if you are doing time series or other sort of signal to noise ratio issues.

So, you know, we have had a lot of success with detecting a particular kind of food borne disease due to salmonella, but in that case we actually have a bacterial

fingerprint of the particular outbreak organism and we digitize it and we can look across a national database to say there is an excess of that particular organism. That is sort of the opposite, where you are taking a highly specific signal, even though you are trying to pick it out from a huge background.

MR. KLEINMAN: Yes. We made the decision to focus on less specific traits because the turnaround time is so much faster. We could wait for a lab analysis, assuming a lab analysis was going to be done on many people who showed up, which is not the case. But we could make the decision to wait for that, but then we have to wait for the lab to process the data and if they have a different data entry system, we have to wait for them to actually enter the data. Then it gets back to us. It is three days after the visit. By then, many, many more people are in Phase 2 of anthrax.

So, we could tailor the system to work that way, to be more sensitive.

MS. BROOME: But the test orders are in some ways also attractive because the timeliness may be there, but there are a lot of issues that are concealed behind all of the good data that you showed.

Yes.

PARTICIPANT: I wanted to comment that the social networks graphs and the kind of distribution of connectivity that you show in your graphs is very closely related to the kind of linkage you see in the worldwide web pages and there is a heuristic that motivates that that applies to the same kind of situation you are trying to model. So, it is something to go on line to look at, I think.

But there is a question I wanted to ask really was I work more in information security. So, when I was listening to this talk, these talks, I was trying to cast them in the light of how can I take what you guys are studying and abstract that away and apply it to the situation of spread of infection among computer systems. This has been attempted by several people in the literature and I see some differences of how some of the assumptions and some of the things you have to deal with simply don't carry over, like the geographic context.

It doesn't really carry over in computer systems. Basically everything is connected to everything from our point of view, but there are lots of other things that do carry over and this is the thing, I think, that mathematicians are particularly adept at doing is trying to

abstract a way and see what commonality is there between these situations.

But think about it. It is a very similar problem that is also based in probably homeland security. But, of course, the rate of infection spread is --

MS. BROOME: Yes.

PARTICIPANT: I wanted to just take a step backwards and ask about the social organization of a concerted effort to work on some of the very difficult problems that I think arise in this context. We have been having meetings of this type for several months now. What is clear from these is that the epidemiological and surveillance questions pose serious methodological questions that have to be dealt with and I think that they largely have to be dealt with by interdisciplinary teams of researchers. At least that is my impression.

And that to really get started, you have to think about how these teams are going to be put together and whether it is, indeed, feasible to have people who are doing very successful research on their own topics, have them break away from that for awhile and work seriously on this very important issue or do we think that perhaps going on doing business as usual with individual researchers

communicating sort of by e-mail and phone will be sufficient to meet the challenge.

Where I am coming from is the fact that what I have seen is a lot of communication like this, but no urgency to move forward and really put something together fast if it hasn't happened right away. So, I just want to hear from the panel and from, indeed, the rest of the participants how we might begin thinking about getting something really going and what structures we might try to work on.

DR. BLOWER: I mean, I think that is very interesting and I do think there are lessons from -- Simon was talking about foot and mouth that occurred in the U.K., that the British Government response was to get scientists involved and to have teams and different teams involved directly with the government and focus on that.

Obviously, you don't want to do that once small pox is actually happening.

MR. EUBANK: There is some problem with time scales. My group is computer scientists, mathematicians and whatever I am, some sort of applications engineer, I suppose. But the mathematicians cannot work on a 60 day turnaround.

MR. LEVIN: Well, in fact, there has been discussion of something of that sort. There is an Academy report that will be coming out in a week or two and there is likely to be some mention of the development of new structures for that. But I am not -- that is not going to have a very mathematical focus to it. But, I think, indeed, there is a need to take people out of their usual context and put them together for a Manhattan Project approach.

MS. CHAYES: But there are some structures for that on a small scale, like the new Bampf Research Center has focused research groups, where people, you know -- something on the order of five or six people can come together for a couple of weeks and work together away from their institutions, if you can manage to get away from your labs and your students and your families, but, you know -- which is obviously a difficult thing, but there are structures in the research community now, I mean, there is the infrastructure to respond to that in a way that there wasn't a few years ago. So, maybe with the impetus of September 11th --

MS. KELLER-McNULTY: Let me chime in here. I am Sally Keller-McNulty on the board and also chair of CATS. I want to follow directly up on what Tom said. Part of the

goal of this workshop is to get exactly at that question is that, you know, these are important problems and, you know, this morning we heard in the data mining session that there is a lot of proprietary work that is going on clearly and for good reason and things that are going on in different sectors and how do we share, but if we really want to energize the mathematical sciences community to try to help address the homeland security problems and vice-versa, we have to begin to address exactly what Tom said.

How do we do this quickly? You know, how do we find and mobilize the people to try to do these problems and to pull them away from their really successful research? So, anyway, so hopefully, during the course of this workshop some good ideas will come out.

MS. BROOME: I would like to suggest that the way the research -- the workshop is set up has the potential in terms of you brought in folks like me, who are not mathematicians, but I actually spend a lot of time thinking about what do we need to have an effective surveillance and response infrastructure.

But I must say this is only going to work if we, I think, think about cross disciplinary teams but think about what are the most likely questions or challenges to set them where there is likely to be a payoff. I will pick

one where I tend to doubt it, the suggestion that maybe we do inverse modeling to find out the source of an outbreak. You know, to an epidemiologist, that is an intriguing thought, but I would much rather send out a team of epidemiologists to interview a bunch of people and find out what is going on. We do that and we generally find the answer.

I am not saying that --

PARTICIPANT: [Comment off microphone.]

MS. BROOME: Talk to the FBI. I don't think modeling is going to do that one either. On the other hand, I have heard some -- you know, there is clearly a whole area around detection aberration, detection issues that Ken is addressing, but there is, as I was trying to indicate a much broader range of possible syndromes, as Art laid out, which are going to have different challenges, let alone as was alluded to in the data mining session, you know, you have got to have the data to apply these to. So, there is some highly applied questions in having data available electronically so that these wonderful tools can be used.

We are sort of working on that side by trying to get down to the nitty-gritty with clinical information technology systems to say what data can we get tomorrow

that then might be used for these kinds of -- to validate how useful these different approaches might be. I think, obviously, Sally has given a very concrete example of how policy decisions on vaccine usage could be approached with modeling.

There is a very active debate on the use of small pox on the vaccinia vaccine or other new vaccines that benefit from that kind of a quantitative approach. So, those are just some things I would throw out as focus areas that are -- you know, would benefit from --

MR. TONDEUR: I want to return to the idea of the infrastructure for this and I want to say two avenues within the Division of Mathematical Science. One is focused research groups, which are specifically to address such issues and the other is an institute we plan to fund attached to the American Institute for Mathematics, which is specifically targeted to have group focused workshops, which actually do the work, where you can assemble teams from different disciplines.

MS. CHAYES: I have got one more area, which kind of came up, I think, in Simon's talk, which is games theory. What I have seen happening in network research is that for awhile people were just looking at the structure of networks like the Internet or the worldwide web and now

they are overlaying game theory on that. So, some cost benefit analysis or some protocol analysis on top of that and if you want to try to implement some of the things that Sally is talking about, you might want, rather than just looking at the differential equations, to take one of Steve's networks and then, you know, put on a game theory functional that would give you some cost benefit analysis on top of that and see how that -- what the results of that are.

I think that that is an area of mathematics that people are just starting to look at for networks for the Internet and the worldwide web and it would probably also be very useful in this context.

MR. LEVIN: The fact that the networks are in some sense adaptive, that as you make interventions, the networks will change in some of the directions. So, for example, if you remove a focal vertex that in the case of a sexually transmitted disease or prostitute another note becomes crucial.

MR. TONDEUR: When you say games theory, I sort of -- if I were a terrorist, I would do the following game theoretic approach. I would say how would I achieve the maximum damage. But then that means we should probably play that same game and say, okay, so if I were a

terrorist, how would I achieve the maximum damage, then how can we protect against.

PARTICIPANT: They don't always think that way. Remember, the main thing, how do they achieve the maximum terror, not the maximum damage. There is a difference between those.

PARTICIPANT: That could be a measure of damage.

MS. CHAYES: It is a different function.

MR. LEVIN: I just don't want people to think about this problem in just the way, you know, how many people are going to die. That is not the only way --

MS. CHAYES: Right. Well, the anthrax certainly didn't kill a lot of people, but it terrorized people, but that is just a different -- I mean, if you want to model, that is a different functional for your game theoretic.

MR. LEVIN: That is right, but the point I was making is that for either side, the point would be how do I achieve the maximum damage or terror given that the response is likely to be this. Do I assume that -- or how do I, if I am interested in response system, create a response system based on the fact that the terrorists are going to respond to my response system. So, those are the -- what makes it a game theoretic problem.

DR. BLOWER: I think the thing -- if somebody said that they had released small pox and alerted all the news media that Washington and New York and San Francisco had now been contaminated and the cases were going to -- that could bluff the government into doing a mass vaccination campaign that would do more harm than good. So, they wouldn't actually have to do anything, just alert the media and the response could be worse than -- I think this needs to be thought about.

MR. MC CURLEY: I would like to ask one other question on the detection, only speaking about detection at the level of people getting sick. There are more automated systems that are looking at more pathogenic agents. We can feed into the data mining thing. I don't know how practical it is. My understanding was at the Olympics that we did have some sort of detectors going for anthrax and other things.

MR. KLEINMAN: I know the Department of Defense is doing that sort of thing. They have sniffing machines.

MS. BROOME: I think, again, it comes back to the specificity and the predictive value of a positive. I mean, just during the Gulf War, there were enumerable alerts of gas that worked as false positives. You know, I think there are some fundamental parameters defining the

characteristics of these tools that we have to pay tremendous attention to.

MR. MC CURLEY: -- way of merging the first session on data mining -- get rid of these delays.

MS. BROOME: There is a lot of interest in that and I think a lot of research ongoing, but it is actually a real challenge. I mean, in many of these settings, as has been noted before, you basically have to have a specificity of a hundred percent to have a useful tool.

MR. EUBANK: If I could make a comment about the practicalities of getting mathematicians involved in these research areas, we have a curiosity-driven research model. So, that means that my goal is trying to convince mathematicians that the problems we have are interesting for them to work and that they should be curious about them. But if there is some way to drive research based on our problems, I think it would be worth exploring because it is not -- I think the problems are interesting, but I don't always get agreement from the people I am trying to convince.

MS. BROOME: Okay. We are over time. I am looking at folks in terms of -- take a couple more questions? Need to break? Quickly.

MR. MC CURLEY: I wanted to ask how many people here have a degree in mathematics?

DR. BLOWER: Mine is in biology.

MR. MC CURLEY: So, this is a question that you are really -- how do you get mathematicians to interact. This is actually very rare for mathematicians to interact this way with other scientists.

MR. TONDEUR: [Comment off microphone.]

[Multiple discussions.]

MR. AGRAWAL: [Comment off microphone.]

MS. CHAYES: Those are the things you -- that is the mathematics you want to look at.

MR. AGRAWAL: [Comment off microphone.]

MS. CHAYES: Collaboration of whom?

MR. AGRAWAL: [Comment off microphone.]

MS. CHAYES: Surveillance.

PARTICIPANT: [Comment off microphone.]

MR. AGRAWAL: [Comment off microphone.]

MS. BROOME: I think one of the issues that Art and I could spend a lot of time talking about is the complexities and the varieties of surveillance. For traditional public health surveillance because it includes individual identities, it doesn't lend itself to wide open, although certainly the project that I am managing actually

is just getting us on the web for doing traditional surveillance and interfacing with hospitals. But there are also other applications where we do, for example, we are looking at doing kind of survey stuff we do over the Internet so that there is group participation.

Then there is some fairly substantial data sets that are put together, for example, of pharmacy data that is de-identified. But most of these are -- you know, they rely on a number of collaborators. I don't think it is the sort of mass computing platform you are thinking about.

MR. AGRAWAL: [Comment off microphone.]

MS. BROOME: We have thought about trying to -- for example, for our web site, trying to also have two-way communication, where we can record incoming information from physicians or the public. So, you know, I think there is a lot of interest in seeing what could be done with that. But, again, there is complexities when you get down to individually identifiable data that we have to be very conscious of.

MS. CHAYES: Do you want the ten minute break now?

MS. BROOME: Okay.

MS. CHAYES: Also, one other thing, everyone who is giving a presentation, please give us a copy of your transparencies or your power point presentations.

Remarks on Detection and Epidemiology of Bioterrorist Attacks

Arthur Reingold

Disease surveillance is a systematic ongoing collection, analysis, and dissemination of health data, with findings linked to actions in the decision-making process. With regard to bioterrorism, Dr. Reingold and his colleagues would like to have a system in place to permit them to detect illnesses quickly and then be in a position to respond rapidly to keep morbidity and mortality to a minimum.

While the 2001 anthrax attacks generated a lot of concern about detection and rapid response to outbreaks, the public-health community comes to bioterrorism detection and response with a good deal of relevant experience. Generally, it detects outbreaks because an astute patient or family member notices them: “Gee, aren’t a lot of us who ate at the church supper all vomiting at more or less the same time?” Similarly, the community learns a lot about outbreaks from astute clinicians who recognize they’re seeing more cases of an illness than they should be seeing and are smart enough to let the health department know about that.

However, public health workers rarely detect outbreaks. Therefore, the goal is to see if they can be a larger part of this detection process, as opposed to waiting for clinicians and patients to tell them about the problem.

People are working in various ways to improve surveillance for the detection of bioterror events. One approach is to monitor visits to health-care providers, particularly outpatient visits, emergency-department visits, clinical-microbiology laboratories, and indicators such as 911 calls, over-the-counter drug sales, and absenteeism at work and at school. Some people are even considering direct monitoring of samples of the population through Nielsen rating-type setups, in which a large group of individuals is routinely answering questionnaires over the Internet.

Three criteria will help us judge a proposed system:

- Sensitivity. Of all the outbreaks that occur, what proportion of them will be predicted by this system?
- Specificity. Of all the times there is not an outbreak, what proportion of the time will the system tell us there is not an outbreak?
- The predictive value of a positive. Of all the times the system tells us there is an outbreak, what proportion of the time is it right?

Roberta Lenczowski
"Introduction by Session Chair"

Transcript of Presentation
Summary of Presentation
Video Presentation

Roberta E. Lenczowski is technical director, National Imagery and Mapping Agency (NIMA), in Bethesda, Maryland. Ms. Lenczowski earned her B.A. in philosophy from Creighton University in 1963. She completed her M.A. in philosophy from St. Louis University in 1970. In November 1977, Ms. Lenczowski began her professional career with the Defense Mapping Agency, now the National Imagery and Mapping Agency. She has held numerous technical and supervisory positions before her appointment as technical director in 2001.

DR. LENCZOWSKI: To try to get this session started, albeit a little bit late, but I know that the discussions have been very enthusiastic, very, very interesting.

One of our speakers, Dr. Rabiner, has to leave by quarter to 6:00. So, I don't want to do a whole lot of introduction with respect to my background. Suffice it to say, I am from the National Imagery and Mapping Agency. As a result, I have a very vested interest in understanding imagery analysis.

So, based upon the short abstracts I have read for each of these speakers, I know that things they are going to say will be very relevant to work that we do within the agency. So, I will introduce each of the speakers. I will give you a little bit of an insight, a tickler, if you will, in terms of what their topic is.

When they have completed then, of course, we will open this to a discussion and an exchange.

Our first speaker is Dr. Malik. He, in fact, received his Ph.D. in computer science from Stanford University in 1985. In 1986, he joined the faculty of the Computer Science Division in the Department of Electrical Engineering and Computer Science at the University of California in Berkeley, where he is currently a professor.

For those of you who are not familiar with his background, his research interests are in computer vision and computational modeling of human vision. His work spans the range of topics in vision, including image segmentation, in grouping texture, stereopsis, object recognition, imagery-based modeling and rendering, content-based imagery querying and intelligent vehicle highway systems.

He has authored or co-authored more than a hundred research papers on these topics. So, he is an incredible asset to the discussion as we continue.

I need to point out that in my own background since the very early eighties, I have been following the work that has gone on in the community that I am predominantly from, that being the mapping side of the National Imagery and Mapping Agency and have watched as we have attempted to have insights in terms of some of the research issues of automatic or what we refer to now as assisted feature recognition or automated or assisted target recognition.

So, one of the things that Dr. Malik plans to talk about in terms of recognizing the objects and the actions is analyzing images with the objective of recognizing those objects or those activities in the scene that is being

imaged. As we get into some of the later discussion period, hopefully I will have an opportunity to provide some more descriptive information about why this is so relevant to the business that we are in and how that, in fact, is of great support with respect to the core topic here of homeland security.

Introduction by Session Chair

Roberta Lenczowski

Dr. Lenczowski introduced herself as a scientist from the National Imagery and Mapping Agency who has a very vested interest in understanding imagery analysis. Recognizing objects and actions means analyzing images with the objective of recognizing those objects or those activities in the scene that is being imaged. Such analysis is important with respect to the core topic of homeland security.

Jitendra Malik
"Computational Vision"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Jitendra Malik was born in Mathura, India, in 1960. He received a B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1980 and a Ph.D. degree in computer science from Stanford University in 1985. In January 1986, he joined the University of California at Berkeley, where he is currently the Arthur J. Chick Endowed Professor of EECS and the associate chair for the Computer Science Division. He is also on the faculty of the Cognitive Science and *Vision Science groups.*

His research interests are in computer vision and the computational modeling of human vision. His work spans a range of topics in vision, including image segmentation and grouping, texture, stereopsis, object recognition, image-based modeling and rendering, content-based image querying, and intelligent vehicle highway systems. He has authored or *coauthored more than a hundred research papers on these topics.*

He received the gold medal for the best graduating student in electrical engineering from IIT Kanpur in 1980, a Presidential Young Investigator Award in 1989, and the Rosenbaum Fellowship for the Computer Vision Programme at the Newton Institute of Mathematical Sciences, University of Cambridge, in 1993. He received the Diane S. McEntyre Award for Excellence in Teaching from the Computer Science Division, University of California at Berkeley, in 2000. He was awarded a Miller Research Professorship in 2001. He serves on the editorial boards of the *International Journal of Computer Vision* and the *Journal of Vision* and on the scientific advisory board of the Pacific Institute for the *Mathematical Sciences.*

DR. MALIK: Thank you, Dr. Lenczowski.

Let's begin at the beginning. How do we get these images and what do I mean by images? So, the first thing to realize is that images arise from a variety of sources and they could be volumetric images, just 2D images, images over time. So, of course, we have our economical example of optical images, which are getting to be extremely cheap and extremely high resolution. There are going to be zillions of these devices everywhere, whether we like it or not.

But I also wanted to mention other modalities, which are for their own advantages. So, x-ray machines so when you walk through the airport, but those are not tomography machines, but they increasingly will be and thus they will give access to 3D data. This may be of great relevance and automated techniques for detecting weapons, explosives and so on.

There are rain sensors, which give you not just information about the brightness coming back, but also directly depths. We refer to this as 2.5(b) data, because it is not the full volume. You could have infrared sensors and no doubt many other exotic techniques are constantly being developed.

I won't talk about the reconstruction part of the problem. I assume that we have the images and let's process them further. So, this is the coal problem. We start off with -- it is just a collection of big -- with some attributes and some spatial, XY, XYZ, XYZT, whatever. But we want to able to say in -- and I am going to use examples from optical images but many of the techniques do carry over. That is a problem. How do we do this?

That is a recognition problem. So, when we look at this image, we all recognize it. Maybe it is because you have been to the Louvre. Maybe it is because this image is splattered over a zillion places. But probably you have not seen him before. So, I wish to argue that -- to identify that something is a face, doesn't rely on you having seen that specific face and, of course, you have varieties.

So, the key point to note is that recognition is at weighting levels of specificity. There is at a category level, faces, cards. There is specific process, Mona Lisa, or it could be a specific person in a specific expression.

The second key point and this is a very important one is that we want to recognize objects in spite of certain kinds of radiation, radiation such as changes in pose. You can look at the person from different

viewpoints. Changes in lighting and -- images, you would have occlusion. There would be clutter and there will be objects hiding each other.

This is why just thinking of it as a traditional statistical classification problem is not quite adequate because in that framework, some of these radiations due to geometry and lighting, which are very systematic, which result from the -- of physics, they are treated as noise and that may not be the right thing to do. So, that is about objects. But we can talk about action recognition as well, particularly when we have video data, whatever.

Just a collection of common English words that are really associated with actions, so you can read the list there. They involve movement and posture change, object manipulation, conversational gesture and moving the hands about, sign language and so on. Certainly, we can recognize these as well.

So, how shall we do this? So, what I am going to do in this talk is sort of -- our way through this area and at a later part, I am going to mostly talk about work and approaches taken in my group, but I wanted to make the survey sort of inclusive. So, initially, I am going to talk about face recognition. I want to mention -- why do I pick faces? Faces are really just another class of object.

So, in principle, the same techniques apply. However, it is a very special and very important domain. It has been argued that a baby at the age of one day -- and there have been experiments which have argued that at age 15 -- can distinguish a face from a non-face. It may, in fact, be already hard-wired into the vision system.

There has been a lot of work in the last ten years and the implications for surveillance and so forth are obvious. So, it is worth taking that specialty. So, this is work from Carnegie Mellon University. So, there are now pretty good techniques for doing face detection. That has to be the first problem. You start out with some image and you want to find that set of pixels, which are -- to a face. So, that is the face detection problem.

I am sure they picked apropos examples here. You can try this detecting yourself. You can select any photograph you like and see the results of the algorithm. I think this is currently the best algorithm in the business because it works for faces and a variety of views and does it reasonably well.

This gives you sort of an idea of what the performance is like. So, the performance is 94 percent detection rate on this with false detection every two

images. This is the current state of the art. No doubt things could be improved.

Okay. So, that is what you can do to try to find faces. But now you want to say whose face is it. Is it Nixon's face or Clinton's face? So, here what I am going to do is to just report results from a study, which was carried out a couple of years ago and the person who did this study or was a leader in the study is Jonathan Phillips, who is actually back in the room. So, any detailed questions, he is the one to ask.

This -- become commercial, which unfortunately always complicates things because now scientific claims get mixed up with commercial interests and so on and so forth. Anyway, here is a study, which -- systems from different vendors.

So, the key issue here is you will take some picture of a person at some time. You are not going to act when you finally try to identify that person in a crowd. The person is not going to obligingly present his or her face in exactly the same pose, exactly the same lighting with exactly the same hairdo, with exactly the same state of wrinkledness or lack thereof. So, you have to consider this recognition in spite of these variations.

So, here the experimenter -- a gallery of 200 people. So, even though you can have exactly the same pose, at which point the accuracy -- pretty good. You can see that the face is slightly angled in each of these. At 25 degrees -- performance has dropped to 68 percent. And at 45 degrees, it is at 30 percent. So, basically it is useless at this stage.

So, basically what we are talking -- radiation plus, minus, 20, 25 degrees. This is starting from frontal views. People at CMU(?) have done a -- they tell me that the results from side views are even worse. The degree for the variation that you can tolerate becomes even smaller. So, this is the reality and the reason is that the techniques are primarily two dimensional. So, they don't deal with a -- variation in any serious way.

PARTICIPANT: How about if you have --

DR. MALIK: If you have multiple views. Yes. So, the CMU people claim that -- because the frontal view is actually the most generalizable view. That gives you this essentially 45 degree cone. From the side view, the generalization cone is smaller. We can get into in more detail later.

So, this is a question about illumination radiation and here again there is a drop here, from 91

percent to 55 percent with different kinds of lighting. This is radiation -- distance, but let's move on.

So, here, this is sort of a summary slide, which shows how the performance -- so, this is performance, how it drops with radiation and pose. Radiation and indoor versus outdoor lighting, different differences and then the same person over differing time intervals. You know, the same person -- one year, two years.

So, I think the moral of the story is that face recognition is not a solved problem. If any propaganda had told you, so please disabuse yourself of this notion. But there has been concrete progress, without quantifiable measurable results, but the issues of generalization, in particular, generalization to -- pose, dealing with the quality of data you might expect with surveillance cameras. These are all real issues and there is plenty of scientific work to be done on this problem. It is not a solved problem yet.

Okay. Let's move on. So, now I want to talk about general object recognition and this is, again, a huge topic and I am going to use the typical caveat, which is -- approach and not necessarily the final approach, so on and so forth.

Okay. So, here is the philosophy first. How do we deal with object recognition? I think the key issue is dealing with shape availability. That is the bottom line. How do we recognize an object? These are -- faces -- or fish. We identify them as the same category, even though they are not identical. The person who -- and now how do we think about this. There is a standard statistical way of thinking about this. You construct a feature vector. You have some training examples and then you sort of model the variation. So that would require that you -- the problem into that framework, which means you take shapes and make them into feature vectors.

This is not necessarily a good thing because you are losing the -- structure of the problem and you will come up with some feature vectors, like moments or whatever and they will often not capture all the aspects of the problem. There will be arbitrariness.

How should we think about this? D'Arcy Thompson, nearly a hundred years ago, had a very good insight about this problem and he said that while it may be difficult to characterize shape, variability and shape may be easier if we think of it this way, that there is this grid here and what this grid is supposed to indicate is that there are spatial transformations, which transform one of these

shapes to this other shape. You can see that with the face. And this transform is obviously a non-linear transform. I mean, you could imagine simple rotation and translation, but we have to go beyond that. Now, this transform could be -- we then considered this variation model in terms of transformation and when we want to say is this possibly the same -- in the same category, we put probability distributions on the space of transformations.

These will be around an operating point, which is the -- shape and this gets us out of this fundamental issue of how do we define shape in this general way. Now, I think mathematicians have studied shape for quite some time, but they basically have this notion of equivalence plus and we really need this notion of similar but not the same. That needs a fair amount of work.

MS. CHAYES: So, what are the kinds of transformations that you allow? I mean, obviously, you are not allowing everything.

DR. MALIK: I will talk about it.

So, this actually -- so, this philosophy is broadly called the deformable template philosophy and there has been a fair amount of work on it and -- has sort of pioneered this approach, but in his work, there are two issues in all of this, which is one -- statistical

efficiency, optimality estimation. There are questions like that. There are questions of geometry and -- questions of computational efficiency and algorithms. All of these need to be sort of factored and tracked at the same time. Otherwise you don't have a workable approach.

So, a key issue here is the issue of correspondence. If I want to estimate a transformation, so these two A's are the same, but to estimate the transformation, I must know what each point was. How do I find that out? Okay. So, that is the so-called correspondence problem and that -- if we can do that, then the second problem is this transformation problem.

So, the first problem essentially is more a discrete -- algorithm kind of problem. The second problem is a more traditional continuous mathematics kind of problem and this is often the case that there is this mixture of discrete type stuff and continuous type stuff and we have to play in both these games.

PARTICIPANT: It is not clear to me that you have to -- I mean, I might be wrong, but tell me why -- why can't I just say I look at such information such that the image -- is that --

DR. MALIK: Computationally it could be -- basically you will wind up doing gradient -- space, which is non-convex and you will get stuck at a local minimum.

Suppose we represent a shape by a collection of -- I basically won't tell you how we do this, but it ultimately comes down to attaching to each point some descripter which sort of captures a relative arrangement of the rest of the shape with respect to that point and then trying to -- so, shape is essentially relative configuration of the rest of the points with respect to you and two points are to be matched if they have similar relative configurations. So, you can sort of formalize that in some way.

Based on what is called an assignment problem or a bipartite matching problem and we will skip the details here. Okay. So, let's say we have done that. Now we talk about the problem of estimating the transformation. So, here is where -- now I can answer your question. So, estimating the transformation, the transformation cannot be arbitrary. You basically wanted to -- you wanted to be plus but not an arbitrary -- you want to somehow penalize ones which are non-smooth and -- but that smoothness has to be -- MS. CHAYES: So, there is a cost --

DR. MALIK: Yes. So, the next slide will tell you what the cost function is. So, essentially what we use is a templex(?) prime model. So, there will be some set of points at which you know the correspondence and now which -- how do you interpolate? Well, a very natural model is a tenplex prime model, in which case there is a certain cost function, which is -- energy and you can solve this in a fairly natural way.

So, the next slide actually will give you an example of this. So, here is what I am trying to do. I have this unknown shape. I don't know what it is. My rule is going to be if I can take one of my model shapes, which I know. I know that kind of fish. I will try to deform it to line up with this. If I succeed, well, then I am probably working.

MS. CHAYES: How sensitive is this to the cost function? If I change the cost function, will I -- a little bit, will I dramatically change the output or --

DR. MALIK: There is continuity, yes.

PARTICIPANT: How do you know to choose the model?

DR. MALIK: I will try each model in turn and then whichever one will match the best is what I would

declare to be the result. You don't want a linear search through all the models. But I won't get into that now.

The grid here is trying to indicate the transform. The grid is like a gauge figure to show you the transform. It is really two steps. One is a discrete -- step. One is a continuous transform estimation step. As you see in the

-- you can see that it essentially stabilizes effectively of the formed shape to be like this one.

PARTICIPANT: So, is it corresponds that you should match, that two points should have the same --

DR. MALIK: No, it is a bit more complicated than that. You don't want it to be just an XY thing. You want it to somehow be more in some way between geometric and topological but that is a meaningless statement, I can see.

MR. AGRAWAL: [Comment off microphone.]

DR. MALIK: No, that separation of the -- and you can do it in either version. You can essentially define a local frame from the direction of -- and you can orient everything with respect to that. But that will be with a middle reflection problem, but it will be with small rotations.

We took this approach and -- with this approach we showed that we can do this for a variety of different

domains and this is important because, again, the tradition of object recognition work has been that it takes a lot of engineering and hard work to really make it succeed in any domain. So, digit recognition, you work on it for ten years. You want to do printed circuit board inspection, the same thing.

Now, this is not a scalable solution because if we want to approach the level of 10,000 objects, hundred thousand objects, you really need one general technique, which you could then clean up with appropriate data to work in many different settings. This approach seems to have that characteristic.

Digit recognition is a classic problem in this domain, that people have worked on it for, you know, many years and there is this data set where every technique known to man or at least to statisticians has been tried. We had nothing special on this and we -- an error rate which is probably not significantly better but at least as good with a completely -- technique and I think tuned for digits with much less training data.

These are actually all the errors and if you look through them carefully, you will find that some of these even you as a human will get drunk. This data set is digits written by high school students and -- two sets.

There is a subset from USPS employees and a subset from high school students. The USPS employees, everybody can do them. It is the high school students who create problems.

Here is how to do 3D objects. A crude way of doing it is to store multiple views and do the matting. There are techniques by which you can select automatically how many views that you need. Here the performance can be judged this way. The tradeoff is between having many views, in which you can do the problem easily and having few views and yet having a good error rate. The point is that even with an error rate with -- with those full view for objects, we could get an error rate of 2.5 percent.

MS. CHAYES: How come that is not monotone?

DR. MALIK: Well, there is some sampling issue there. There is some stochastic processes involved in the selection of the view.

So, here is an example in a more difficult type of domain. So, now we are moving towards talking about action recognition. You want to identify people and what action they perform. Now, identifying people really -- if you have enough resolution on them really implies knowing their body parts, meaning knowing all their drugs. So, on this figure you want to -- you want to be able to -- right shoulder, left shoulder, et cetera, et cetera. If you can

do that, then you can -- it turns out that the same approach works, except that you must use an appropriate information model.

This shows some results and this is from a -- this is from a Japanese, this tradition of comic books. So, they have these -- a variety of causes -- these artists to copy from -- a stick figure model from a single image and the reason you can do this is because of anthropometrical strengths. You know all the names of the body parts, essentially provide you an equivalent of a second view.

This is an example of -- if you move the mouse -- so, this is a sequence of -- what is being done in -- a stick figure is being identified. This -- the question of recognizing actions. This is what we mean by action recognition.

I want to point out that actions are -- the -- action is in a very early stage. I mean, there has been work done by others in the community, but usually you do distinguish between two classes of action. Actions are to -- a hierarchy and this is an area which really needs to be explored. Think of the action of withdrawing money from an ATM. It is in fact a sequence of steps, but the sequence of steps need not always be executed in exactly the same

order. It is really there is a partial order plan, which is -- here and multiple agents could be involved in this.

So, this is very relevant because if we want to talk about situation assessment, that is really this -- I want to just remind you that the problems are actually quite isomorphic. More or less -- idea really applies in space time as well as in space.

The cues are both static as well as dynamic. So, in some sense the -- and the dynamics and when you want to recognize action, this is all just based on the movement of my body but what objects are manipulated. So, object recognition and activity recognition is coupled. The fact that I am writing with a pen -- so, the philosophy that we can try to do this by recovering -- I will conclude here. This is my final slide.

So, it should be obvious that there are zillions of challenges for mathematicians and statisticians in this business. It is a fundamental problem. Thirty percent of the brain is devoted to -- it has evolved through a million years. There is an arms race between animals, trying to camouflage animals, trying to detect and break through camouflage.

Just a few of the problems. Shape, modeling shape radiation. Shape is a fundamental problem. The

field of geometry and mathematics is all about shape. But here is a question that is similarity. How do we define similarity in that. This is our fundamental problem. The statistician viewpoint together with the algorithmic viewpoint because, for example -- had a great -- construct on how to think about -- because you claim that you have a optimization problem and then you can't solve it. Doesn't get you too far.

I should, again, return to my -- I should return to the beginning. I -- optical images, but that is not the only imaging modality. The more exotic imaging modalities you have more fundamental questions of how to produce -- the question of reconstruction. Thank you.

[Applause.]

Computational Vision

Jitendra Malik

Recognition by computational vision may be carried out at varying levels of specificity. It can involve searching for any face, a type of face, a specific person's face, a specific person with a specific expression, or a person performing a certain activity. Major challenges in recognition relate to detecting objects despite "radiation," which includes changes in pose, hair, lighting, viewpoint, distance, and age.

These are two approaches to the problem of shape variability. One possibility would be to divide a particular shape into components and perform comparisons between two shapes component-wise. A second would be to view the two shapes of interest as shapes on a basic grid that indicates spatial transformations and determine if there is a transformation mapping one shape to the other.

For the future, researchers are moving toward the problem of action recognition—that is, identifying people and the action they are performing. Current approaches draw from the above techniques, but the work is in a very early stage.

Ronald Coifman
"Mathematical Challenges for Real-Time Analysis of Imaging Data"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Ronald Coifman joined the mathematics and computer science faculty at Yale University in 1980. His research interests include nonlinear Fourier analysis, wavelet theory, singular integrals, numerical analysis and scattering theory, real and complex analysis, and new mathematical tools for efficient computation and transcriptions of physical data, with applications to numerical analysis, feature extraction recognition, and denoising. He is currently developing analysis tools for spectrometric diagnostics and hyperspectral imaging.

Dr. Coifman is a member of the American Academy of Arts and Sciences, the Connecticut Academy of Science and Engineering, and the National Academy of Sciences. He is a recipient of the 1996 DARPA Sustained Excellence Award, the 1996 Connecticut Science Medal, the 1999 Pioneer Award of the International Society for Industrial and Applied Science, and the 1999 National Medal of Science.

Dr. Coifman was the Phillips Professor of Mathematics License en Sciences Mathematiques in 1962 and received his Ph.D. from the University of Geneva in 1965.

DR. LENCZOWSKI: While we do the conversion to the next speaker's presentation, let me introduce Dr. Coifman. He is the professor of -- the Phillips Professor of Mathematics at Yale University. He received his Ph.D. from the University of Geneva in 1965. Prior to coming to Yale in 1980, he was professor of Washington University in St. Louis, one of my alma maters. So, great regard for his work.

Professor Coifman is currently leading a research program to develop new mathematical tools for the efficient transcription of physical data with applications to future extraction, recognition and denoising.

He was chairman of the Yale Mathematics Department from 1986 until 1989. Today his topic, the mathematical challenges for real time analysis of imaging data, is caught in the first statement of his abstract, where he points out that the range of task to be achieved by imaging systems designed to monitor, detect and track objects require a performance comparable to the perception capabilities of an eye. Such vision systems are still beyond our capabilities.

A couple of days ago, I had a couple of contractors come in and we were doing a briefing and describing the fact that despite all the work we have done

with knowledge management and knowledge engineering and artificial intelligence, that to date, even best computer is only comparable to the brain of a cat. So, clearly, you have the challenge.

DR. COIFMAN: I would like to actually describe an experiment that I am involved in but -- and raise and describe the various mathematical challenges that come along the way. You will see that most of what we discussed this morning about data mining occurs immediately in the context of imaging and understanding imaging and trying to extract features from the images. Images you might think of are just what the camera gets but they are really as we have seen in the preceding talk are really quite a bit more complex.

So, as we heard, I mean, what we really want to be able to do with an imaging system is to detect objects, track them, monitor for objects and basically we want to do what we do with our eye. Unfortunately, a camera is not an eye, as I said. A camera measures something for each individual pixel. Then you hand it on to a computer and the computer would -- hopefully you will have a decent algorithm to get something out of it.

This is really not what we do with our eyes and I will indicate a system that allows you to resemble more to

an eye than to a camera and the system is basically mathematics converted to hardware. That is, I think, another challenge that will happen. I think that future devices will have the mathematics much more integrated in them because of existence of new technology like MEMS(?), which allows you to control on a micro level analogue data. So, we will see. We will get to that.

The main obstruction to everything is what we heard this morning, but I would like to maybe describe a little bit the mathematical problems or basically our inadequacy to handle the kind of data streams that we are collecting.

So, problem No. 1 is we really don't have any tools to describe in any efficient fashion high dimension geometric structures. So, we heard this morning about having a variety of parameters, which have to be looked at as a multivariate system of parameters and then we want to have a probability distribution describing probabilities of various events, but we have no tool to describe the actual support of those probability distributions or just have an efficient way of describing the geometry underlying that.

In fact, we don't even know what the word "geometry" means. You know, just a few years ago it was Komolgorov, who made the observation that there are other

objects in traditional geometries, which are very useful. In fact, I think our task is really to follow that path and identify structures as they arise in nature as we confront them and the structures are not that simple.

The second problem -- the second issue is that it might seem a little silly, but we really -- all the algorithms that we have are really exponential in their dimension. So, if you are already in Dimension 10, you are dead. We deal with things, which are much higher dimensions.

Even simpler than that, we don't know how to approximate ordinary functions, which are given to us as empirically, as functions of some parameters. So, to give an example, you have a blood test. You will get ten chemicals being measured and you want to approximate the function, which tells you the state of your health in any sort of way. I am not aware of any mathematical algorithm capable of doing this efficiently and giving you any sort of information.

What we need is, of course, structuring clusters, a structural clustering tool. We have to know what the word "structure" means and I will try and indicate how we get into that. Again, what I just said, in notions of geometric structures are really limited. Our imagination

is rather dull and we have to have nature sort of inspire us.

Last but not least, probably the most important part of this at this point in terms of practical application in order to take those ideas and have them work is we need to have efficient digital representations of the massive data sets that we have been talking about.

So, what I would like to -- there is just one idea I want to convey today in some sense. That compression, which is basically an efficient representation of data is a key ingredient in everything that one wants to do. I mean, you may think of compression as something that multimedia people do in order to transmit something on the line, but actually compression is basically an outcome razor-type principle. If you can compress and describe data very efficiently, you have a model. Having a model means that you basically are way ahead of this.

So, a good test is actually -- so, in many cases, right, the fact that we cannot store the data is really an indication that we are just -- we just don't understand it. It is a real paradigm -- it is a real guidance. Over the last few years we have seen a lot of -- there have been many developments about trying to represent objects with minimum complexity.

David Donoho recently wrote a beautiful paper showing that if you have a minimum complexity description of a set of data, it provides you essentially an optimal statistical estimator. So, the two things are tied. Complexity is -- good compression and noise and denoising are very tied together. Noise is a generic problem everywhere. The problem is that, of course, this is just nice goals. We need to know how to reach them.

So, let me describe -- so, the next few slides, some of you may have seen them, but they sort of fit the themes that we are describing here. It is basically a way of taking a complex structure and starting to -- trying to orchestrate it the way you would orchestrate a piece of music, as a collection of scores, which describe the effect of each instrument. Here, the structure is an image.

So, you may have seen this. One of my favorite. So, here is a complex image, which has a lot of stuff in it and just trying to compress it, immediately unravels it as a combination of different structures, which are layered one on top of the other. So, here, this is just compressing it with basically taking one-third of 1 percent of the number of parameters and what you see happened is that it is blurred. It looks like a water color painting of the same thing.

All of the hairy features have disappeared, but all the coloring information, which may be useful for a given task, say, for example, then the color might be all we need in order to determine the age or the state of health of this particular mandrill and then that is fine. We don't need the rest of the stuff.

This is a residual between the original image and the ones that we had. So, it looks like a lousy impressionistic thing. So, we took this residual and now we compress it by painting -- taking the best description mode of it. Now you would have a Van Gogh type impressionistic image. Right? It means it has been synthesized with brush strokes essentially.

The other one was water colors. But bear in mind there is no human intervention in this business. You will see -- and this is a residual, okay, which is completely -- so, the original image is the sum of those three layers set upon each other, each one synthesized with a different instrument and it is completely like an orchestration for music. I mean, each musical score corresponds to an instrument. The score is given by a collection of -- by a sort of alphabet or language that describes it. Exactly the same thing has happened here.

Bear in mind this guiding principle that I have described is a principle of just complexities. The ability with the tools we have at our disposal to actually just compress this object to some accuracies, so to speak, and it gives a pretty decent compression, much better than if you were to do it once directly without -- but it is an indication that a lot of the things that we are trying to measure or assess are really the result of a super possession of structures, one on top of the other. If we had the ability to see that somehow or to unravel it into those structures, in other words, lift it to higher dimensions in order to do that, we are really doing a substantial job.

So, let's return to the eye topic. We heard in the proceeding talk what kind of processing and how difficult it is to actually do very simple operations that we won't even think about them, seeing that there is a face down here.

So, how does one deal with that directly without having to measure each individual pixel, make up an image and then start to fool around with the pixels and do things. In other words, we want to really proceed as an eye. The eye that I am involved in building at this point is an eye, which is -- should be superior to our eye. It

is designed to not only see the colors of the image, but actually measure the spectrum, the electromagnetic spectrum, the absorption at different, at maybe a hundred frequencies or 200 frequencies, the point being that this would allow you to detect a chemical plume.

This would allow you to tag, to look at somebody if you want to tag or track him. You will look at the spectral signature and you don't have to worry about shapes and stuff like that. Then you measure and you follow. We know that it is much harder to do any operation on a gray level image than it is on a color image. If I want to discriminate object by color, it is much easier. When I look at the -- looking for a green spot here, I find it on the spot. I don't have to measure -- these are in green there or there or there. I can just look at it and my eye immediately goes to that location.

So, the question is how can we do that and what is the role of the mathematician in doing that? So, problem No. 1 is how does one integrate the processing with the actual measurement event? So, you don't want to measure each individual pixel. You just want to extract some maybe potentially global information and you want to analyze it in some fashion. Okay. So, the analogy, it is just like the guy -- like we heard this morning, if I am

looking for a needle in a haystack, if I have a good magnet, I bypass a lot of image processing.

So, here the magnet is to do the spectroscopy in some sense, I mean, and just look at the color cubes. I will indicate how you integrate the computation or the mathematical number into the measurement. So, first of all, just the fact that understanding the -- or measuring the spectrum, and here is an example of spectra, is very important information about the material component of ingredient that you are looking at. You see, there is a camouflage nest over there and grass and -- and you see that those curves are different.

So, to each pixel now you have associated a full curve. So, now we are getting into the data mining glut here. It is a vector, which may have hundreds of components and if we wanted to do -- by the way, everything I am telling you has been around for 50 years. Hyper-spectral imaging has been around. The reason it is not commonly used by everybody here is mathematics and data glut. I mean, you can't deal with the monster amounts of measurement that you are capable of measuring if you are capable of measuring it.

So, each pixel, you may have hundreds of data. So, it does give you a beautiful signature, but if I am

looking in a scene for some chemical spill or maybe for the presence of some camouflaged object or somebody, which is really different from the environment or maybe it is melanoma on my skin, then the issue is how do I find this on the fly, so to speak.

So, I will give you a recipe in a second. So, by the way, this is something that the people in the U.S. Geological Survey are doing. This is an image of the Nevada Desert. What you really want to do is find out what minerals are present at various locations. What you want to do is get this image.

That tells you exactly what is present at various places. Unfortunately, this takes a long time to do. You can collect an image and it can take you days or months to actually get this kind of analogy, depending on the image. You may want to look at vegetation, see if it comes here or not or you may look at tissue and see if it helps. Or you may want to track somebody and discriminate that somebody from everybody else in the environment.

So, please go to the next one. Mathematics and the tools being used are very primitive. Most people -- what most people do is this is the nine first -- the first nine components of a hyper-spectral image of the arm in which what is being done is they do here a -- analysis of

the maybe a hundred components that they measured and then display the first three components in red, green and blue, the next three in red, green and blue, the next three in red, green and blue. So, you get those blurred images, which may or may not be useful. But you have now taken everything and mixed it altogether and done what most people do is use single -- composition and use that to find your coordinator, which is not really the -- a particularly good tool for that kind of activity.

So, let me tell you about the -- so, the camera we have is based on a digital mirror array. This is a device that Texas Instrument manufactures for projectors. It is an array that has half a million little mirrors. Each one of the mirrors has an on and off situation. It is being used usually to project on the screen by aiming the pixels that need to be aimed onto the screen and eliminating for a length of time corresponding to the intensity and running a color -- what they have in their hand is a computer that they are using as a paperweight essentially.

The point is that you can take -- and this is just an example of how more smarts can use this device. So, you have an image here, which is this red plume over there and behind the missiles, you image the whole scene,

if you wish, on this square array of mirrors and you ask yourself is there any red spot in that region. So, the way it goes is you aim all the mirrors in the region everywhere, all of them together, into a special analyzer, which is also run by mirrors.

So, because it found a red spot, the question is where did it find it. Well, it found it -- it breaks it into four quarters and checks for each one of them and then keeps on doing and in a few measurements has a binary coordinates of the location and did not do a lot of measurements. So, the next few slides will show you -- we took a chemical plume and we tried to find it exactly doing that. So, this is just an image of the mirror system and this is the image of our camera. So, this is not just science fiction at this point.

That tells us how this spectroanalyzer goes. Light comes in and then it is being -- goes through a prism or something . Different colors go to -- this is a mirror array up here. Different colors will go to the mirror. You can select any combination of colors you want to measure whether they are present or not and then going to your detector.

So, you both analyze spectrally and physically with the same thing. So, this is a scene in which there is

a chemical plume in which we measured for each individual pixel, whether we could detect a plume or not. But the level of the plume is so faint that the pixel was not capable of actually seeing anything. So, on the other hand, there is a method of multiplexing the light called Hadamard multiplexing, which allows you to actually detect the plume as you can see it. As you can see, it is quite noisy and the plume is rather faint, but in order to achieve that, we needed to do 250,000 measurements because that is how many pixels you have.

If you follow the scheme I just described, well, in the white squares you find it. In the black ones, you don't. You keep doing it until you are finished. So, basically in 20 measurements you did it, as opposed to measuring the whole thing and then analyzing -- first of all, measuring the whole thing was garbage anyway. It was below the threshold of the system.

Secondly, we really needed to combine the intensity of light coming from different locations in order to get this thing to work. So, this is just a tiny example of what you can do with this combined spectral -- combined spatial spectral mix and the switching of the mirror. So, the mirror operates as a computer, basically doing mathematical operations on the light as it comes in before

it is being converted to digital as opposed to the usual way, which is you convert it to digital. You are dead and that is it, I mean, basically because you have too much data.

PARTICIPANT: [Comment off microphone.]

DR. COIFMAN: Yes, I mean, this particular system of GI would go up to 2 1/2 microns. That is because they have a cover plate. If we change the cover plate, you can go all the way up. The cover plate is not letting the light go through it.

So, the point being that I think this is going to be a paradigm for a lot of the sensing systems that are coming and DARPA has -- in fact, the mathematics of DARPA have a serious program in pushing it forward and that is that you really want to -- if you want to reduce the data flow that comes into a system, you really have to do it as you measure it, which is really what a biological system is doing. We never collect everything and then -- I mean, this kind of -- this sort of forces the issues because if you wanted to do this in video -- for a real time thing, you are going to collect for each image several megabytes of data if you are lucky, if you already reduced it.

It just won't flow through the system. It doesn't really matter. The transmission is a problem. The

storage is a problem. Computation is also -- so in all of that, where does mathematics enter? It enters in -- doing this sort of -- the data mining of the spectra, which is very similar to the data mining occurring, say, in gene expression, for example, the same kind of activity if there is a dictionary going back and forth between the two activities.

The issue is -- understanding the internal geometries of the spectra as they relate to the spatial distribution, this is critical if you are looking at objects of a given size. Say it is an artificial object in the natural environment or if you are looking at somebody, for example, who has makeup on his face, you will be able to immediately recognize that or basically you want to just track and object.

Again, there are lots of theoretical issues and a lot of practical issues and I don't think you can separate one from the other.

[Applause.]

Mathematical Challenges for Real-Time Analysis of Imaging Data

Ronald Coifman

Researchers would like to develop an imaging system that will detect objects and track them–not pixel-by-pixel, as in a camera, but how we do it with our eye. One challenge includes the fact that researchers don't have tools to describe, in any efficient fashion, high-dimensional geometric structures. In addition they do not know how to approximate ordinary functions, which are given to us empirically, as functions of some parameters. Also, in order to make data-mining techniques work, researchers need to have efficient digital representations of mass data sets.

Dr. Coifman is currently involved in building an eye not only to see the colors of the image but to actually measure the electromagnetic spectrum, which could enable the detection of, say, a chemical spill, a melanoma on the skin, or the presence of particular vegetation in the Nevada desert. Looking at the spectral signature eliminates the need to look for shapes, and we can perform mathematical operations on light as it comes in as opposed to storing the information in digital form.

With too much data, transmission is a problem, storage is a problem, and computation is a problem. For that reason, the data-mining of the spectra, which is very similar to the data-mining occurring, say, in gene expression, is going to be a paradigm for a lot of the sensing systems that are coming.

Larry Rabiner
"Challenges in Speech Recognition"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Larry Rabiner was born in Brooklyn, New York, on September 28, 1943. He received an S.B. and an S.M. simultaneously in June 1964 and a Ph.D. in electrical engineering in June 1967, all from the Massachusetts Institute of Technology.

From 1962 through 1964, Dr. Rabiner participated in the cooperative program in electrical engineering at AT&T Bell Laboratories. During this period he worked on designing digital circuitry, issues in military communications problems, and problems in binaural hearing. Dr. Rabiner joined AT&T Bell Labs in 1967 as a member of the technical staff. He was promoted to supervisor in 1972, department head in 1985, director in 1990, and functional vice president in 1995. He joined the newly created AT&T Labs in 1996 as director of the Speech and Image Processing Services Research Lab and was promoted to vice president of research in 1998, where he managed a broad research program in communications, computing, and information sciences technologies. Dr. Rabiner retired from AT&T at the end of March 2002 and is now a professor of electrical and computer engineering at Rutgers University and the associate director of the Center for Advanced Information Processing (CAIP) at Rutgers.

Dr. Rabiner is coauthor of the books *Theory and Application of Digital Signal Processing* (Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Prentice Hall, 1978), *Multirate Digital Signal Processing* (Prentice-Hall, 1983) and *Fundamentals of Speech Recognition* (Prentice-Hall, 1993).

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, the National Academy of Engineering, and the National Academy of Sciences and a fellow of the Acoustical Society of America, the IEEE, Bell Laboratories, and AT&T. He is a former president of the IEEE Acoustics, Speech and Signal Processing Society, a former vice president of the Acoustical Society of America, a former editor of the Associated Students of Seattle Pacific's (ASSP's) *Transactions*, and a former member of the *IEEE Proceedings* editorial board.

DR. LENCZOWSKI: The third speaker in our session this afternoon takes us beyond the issue of the visual to the issue of the sense of hearing. Dr. Rabiner is an individual who has an experience base in the private sector. He received his Ph.D. from the Massachusetts Institute of Technology in 1967, but he created or moved to the newly created AT&T labs in 1996 as the director of speech and image processing services research lab. He was promoted to the vice president of research in 1998, where he managed the broad research program and communications, computing and information science technologies.

He retired from AT&T at the end of March 2002. Now he has pioneered a number of novel algorithms for digital filtering and digital spectral analysis and in the area of speech processing, he has made contributions to the fields of speech synthesis and speech recognition.

I am not sure what many of your personal experiences are with speech recognition systems, but I recall being assigned to a lab in 1981, where we have the first system with a computer-based activity that was to allow us to some speech recognition so that we could give commands to the system for some of our mensuration work.

It was my experience that voice recognition did not work very effectively. In fact, the only time that I

knew that we truly had voice recognition was whenever we were scheduling tours in the lab. Somehow the equipment always knew, recognized the term "tour" and we invariably would have crashes. So, that is my background experience.

Now, what Dr. Rabiner is going to talk about today is the maturing, in fact, of our speech recognition to the point where it is not -- now it is very widely applied to a broad range of applications.

However, as he points out, although the technology is often good enough for many of the current applications, there remain key challenges in virtually every aspect of voice recognition that prevent this technology from being ubiquitous in every environment for every speaker and for even broader ranges of applications.

DR. RABINER: Thank you. You took away my entire introduction.

Actually, the question that people ask me a lot is after 40 years of working in this field, you know, isn't this a solved problem. You know, you can actually -- there is actually a system we put out in 1992 does 1.2 billion recognitions a year for operator services and it is essentially flawless. It is about 99.8 percent accurate, but it only works on a five word vocabulary over the phone.

It actually works in the environments you are in. But the whole problem is that it is very task specific. It is very environment specific. It is very well-trained to what it can do. It just can't get to where we want to go.

So, what I am going to try to use is the next half hour to tell you what are the current challenges in speech recognition. Now, for the most part, the services people talk about are broader than just automatic speech recognition. So, I will concentrate on that first box in what we call the speech circle, but in order for you to maintain a real service, a real voice-enabled service with a customer or with somebody who wants to use the system, you have really got to go around the whole circle.

So, the first thing is, you know, the customer speaks something and I should recognize the words. That is a tough problem and we will talk about what are the challenges there.

But even when you get the answer, I dialed the wrong number, in order for you to do something with that, okay -- this is a customer care application -- you have to go out to the next level, which is spoken language understanding and get the meaning.

Now, I dialed the wrong number might mean I don't want to pay for that. But it might also mean could you

please help me get to the number. Something is wrong or there is something going funny here. What do I do? So, meaning, which is something that is completely unclear because the exact same input has lots of meaning. Okay? Once you try to guess at that meaning, you have got to go do a dialogue management, which means take some action.

Okay. So, you might have to do a database dip. You might have to go out and look at person's records and you have to do spoken language generation, create some message to get that, to keep this speech circle working. Then you might want to say what number were you trying to dial. Okay. So, I can give you credit so that when you dial that number, you don't get charged twice or I help you dial it.

We use a synthesizer to speak it. What number did you want to call and now you are in a loop. Hopefully, that system, that entire speech circle will enable you to keep using the system and these are the kind of services people try to build today.

So, we are going to concentrate on that first box a lot because there really are a lot of challenges there. The speech recognition process after a very long time is very mathematically oriented. It starts off with your infant speech, you know, 10 kilohertz, 16 kilohertz, 8

kilohertz, whatever you want to have and does three processing steps. The first one is feature analysis. You really don't want to process that entire 64 kilobyte signal since it is laden with redundancy and it has information that is useful to you as a human, you know, to talk to somebody, to have a conversation, but it absolutely has nothing to do with the meaning of the words.

So, things like the pitch and your inflection and your emotion and things like that, which, you know, are very important for human communication, but for recognition, we try to eliminate them because they are confusing variables.

So, we do a mathematical analysis. We will talk about that. Then we do the heart of the work, statistical decoding, based on statistical models and, of course, this is the box that has three attached models. Those attached models are the training environment and we will get into that.

Then, finally, we do a little bit of cleanup and there is the classical computer science, *helloworld* program, you know. If you have got that one, you may even have a chance of going to a little more sophisticated one.

Let's look at each of the boxes there. The feature -- box is all this to go out and get robust

features, things that are irrelevant or invariant to the environment, to the noise, to the transmission, to the transducer used, to the background situation and make them be relevant for -- and they did have the information. So, that was spectral information that characterizes the sounds of speech, the phoning, the context of -- et cetera. So, over the last 40 years we have learned how to do it through a spectrum analysis. We are using something like linear predictive coding, which let's derive parameters that we think are extremely efficient representation to the signal.

In fact, we code speech from 64 kilobytes down to about 5 kilobytes. Now, if I have to give you kind of a guidance point, your cell phone uses an 8 kilobyte representation. So, it is below that, but it is fundamentally -- it has got some of the same parameters that the cell phone uses for actually operating in the poor environment of cell -- but it has got about 30 or -- that you use in cell phones. So, it is a little bit higher rate. The first challenge you come across is the fact that -- as I said, there are at least a dozen major systems out there. You can call Charles Schwab and get stock price quotes. You can call AT&T's customer care and I will show you, for example, that system. They all work nicely, but they are not robust.

They fall apart in noisy environments, when you have office noise, when you try to operate it in the middle of an airport lounge, when you are operating in your car at 60 miles an hour and you have road noise and you have all the background and the radio blaring. When you try to use it over a speaker phone or a cell phone, every system just about falls apart. Get somebody with an accent or a dialect or a speaking defect, real problem. Of course, noise and echo environments.

So, the first problem and the major problem is robustness. You might say to yourself, why don't you make parameters that are robust to everything. What happens is then you have taken a distribution which is finely tuned and works real well and made it broad and it works terribly all the time. That is your choice. So, we have got to do something.

The second thing, of course, is that the speaker set is one that we devised probably about 10 to 15 years ago and maybe there is some place for it to go. So, let me play a couple of examples.

The mismatch between when you train the system, so you train a system in an environment for a set of speakers, for a set of inputs, et cetera, and when you test them, on the whole, people don't obey those rules. You

know, they really want it to work where they want, the way they want it to work. I will show you examples of how it can result in performance degradation.

So, the concept is let's try to make it a little bit more like humans would do it perceptually. So, here comes a nice little model of what goes on. When you train your system, you had a training signal. People smoke in the environment. They are going to use it. You extracted your features and you build models.

Those statistical models are really good representations of that data. Now, all of the sudden you hand that system off to somebody and they start speaking in a new environment and now the signal is mismatched to that signal. So, of course, what you are saying is, well, I want to add to the new environment and retrain or after the fact, since I can't really do that in many cases, I have three paths I can try. One is I can say let me try to enhance the signal so that even though you are in -- well, let's enhance it. Let's get rid of the noise or let's get rid of the reverberation or let's make it less sensitive. So, enhancement is kind of intense. Make this signal statistical characteristics look like the original one.

The second thing you can do is -- and try to normalize them so that the average statistical properties

of these features match those. And, finally, the last thing you can do is when you finally have this model, you can say this model is a bad mismatch. Let me try to go at it, adapt that model. All of those have been tried and clearly with varying degrees of success.

The acoustic model, though, now we are at the second stage and before -- there were three major models. There is one that is based on physical speech data, called the acoustic model and two that are based on -- so, the acoustic model is basically how do you take frames, which come on a regular basis every 10 milliseconds, but sounds don't come regularly. Some sounds are long vowels, some sounds are very brief, like stock consonants, very short. So, we want to map the acoustic features, which occur on a regular frame-by-frame basis into events. The events are sounds of the English language and you can think of these as being the phonings of the language.

Now, that is just a very simplified view, because we don't use phonings. We used context dependent phonings. Instead of having about 50 sounds, we have 50 times 50 times 50 sounds. That is where mathematics really is nice because you can handle 50 times 50 times 50 sounds. Computers are pretty fast and we have some pretty good techniques.

So, you characterize each of these sounds by the -- Markov model, which is a statistical method for characterizing the special properties of the signal and it is not the right model for speech, but it is so powerful that you can get into the model and represent virtually any distribution. Now, the problem, of course, is you are over-training it in some sense, which makes you non-robust. You are training it to a distribution, which it doesn't actually follow. But it actually in that environment, that overtraining works extremely well.

So, it is a powerful statistical model and, of course, it is a classification model. You give it only exemplars of the sound you are trying to train. So, it does a terrible job. So, for example, let's say your vocabulary with the letters of the alphabet and you are spelling a name. So, you have a letter like "e" and, you know, it is a set of b, c, d, e, g, p, t, d, z. Listen over the phone and you are almost random. When I spell my name, I say R-a-b as in boy, because if I don't say that, it will come out R-a-v, R-a-g, R-a-d. It is whatever somebody hears because they are almost indistinguishable.

So, that is one of the challenges is if you only train models on similar events, how does it every know how

to discriminate that model from other models that are in the system to give you better performance. Okay?

The next thing is the word lexicon. So, we have got sounds and now we want to form words because not every sound is followed by every other sound and we know that. We know there are only a certain number of consonants that can follow each other before a vowel has to appear. So, the concept of the word lexicon is it is a dictionary. How do you map legal phone sequences into words according to rules? So, a name like David, it is da a va id a. That is kind of nice for a name like David. How about a word like d-a-t-a. Is it data or is it data? Completely different. So, you have to put in at least two representations. We all know of words with multiple -- either, either, et cetera. Lots of pronunciations. So, here is a challenge.

We can pick up a dictionary, you know, take -- you know, use it all, but how do you generate a word lexicon automatically. What happens if there are scientific terms or there are new names that you haven't seen before. How do you edit various dialects and word pronunciations, mathematically complicated problems.

The next level of the language model is what word sequences are now valid in the task you are doing, that so-

called language models. How do you map words and phrases into sentences that are valid, based on the syntax of the recognition task you are doing? So, there are two ways. You can hand craft them. So, you literally write out a graph of what is allowed and people do this for a very long time or, of course, you go back to statistics. So, you take, you know, a few hundred million words of text and build a trigram, an engram statistic of, you know, what is the probability of this word followed and preceded by other words and this word or combinations of words.

Of course, the problem is that you get the text from somewhere and that assumes that people will speak according to that text model, much as we assume they would speak according to the accoustic model and of course there are problems of how you build a new one for a new task. Still real challenges there.

Finally comes the heart of the work, the place where mathematics has really helped a lot and that is the pattern classification. So, we have got these features coming in. We want to determine the proper sequence of sounds, words and the sentence that comes out. We want to combine all of those, those three models, to generate the optimal word sequence, the highest probability, the maximum likelihood event given a task, which defines the syntax,

given a set of words that could be spoken, which is the lexicon and given the sounds of the language.

So, we have an -- search through every possible recognition choice. So, the real issue is how do you build an efficient structure, a finite state network for decoding and searching large vocabularies. Let's assume we have a vocabulary of a million words, okay, with 50 times 50 times 50 units, with all of the complexity of the language model. So, we have features, cross units, cross phones, cross words, cross sentences and some of our big recognizers have 10 to the 22nd states that have to be searched.

Well, that is difficult. The goal is to take this structure, which even sequentially you are going to 10 to the 22nd states. No one can do that, even with nice fast computers. So, we have very, very fast approximation algorithms, like the A Star Decoder or the fast linear search, but instead we figured out methods that can compile the network down from 10 to the 22nd to 10 to the 8th states. That is 14 orders of magnitude of deficiency.

We don't know what the limit is. And I will show you some examples of what we have got. So, the concept is for a 64,000 word Wall Street Journal vocabulary -- this is the one that went down -- 10 to the 22nd, 10 to the 8th. The way it does it is by using -- weighted finite state

transducer at the HMM level for the sound, at the phone level, at the word level, at the phrase, they all get combined. They get optimized. They get determinized and they get minimized using very, very fancy computational linguistic techniques and out comes a search network with significantly lower states.

So, here is an example of just a simple word transducer under the word "data" and basically the transducer just has arcs, which are the sounds and states where it winds up. So, from the da we wrote an a or an a with certain probabilities, with certain weights. These weights can be optimized and they can be minimized for the whole graph and the end result of it is that we have been looking at these techniques for the better part of six years and here is Moore's law, for those who are familiar with what -- you know, every 18 months, computational capability doubles. This is what we call Morey's law, from the name of the person who has been spearheading the effort at AT&T, Nario(?) Morey, and he has doubled the performance of search every single year. So, it is almost a factor of 2 to the 5th in five years of searching.

Using the best approximation techniques in the speech community, it is off by a factor of about 2 to 1.

So, the community has done pretty well, too. So, another interesting problem.

Finally, the last stage in recognition, utterance verification. The concept here is that you are asking the machine to give you the best sequence of words according to a grammar and a task. The answer is how do you know if that sequence of words is reasonable or the right ones. So, we do what we call utterance verification. So, somebody says credit please in a somewhat noisy environment. The machine finds an equally valid sentence credit fees. When you all done, you can ask the machine to tell you what is the likelihood or the confidence score of the word "credit" and it is very high. That is great.

But fees was not. It was certainly better than please. So, you would take that one and say I am not so sure that is the output. Okay. You can reject it or you could go out and ask for some clarification. So, here are some examples.

Somebody -- a baby cries in the background of this one. Those are both in this database, which I will show you the result on in a minute. So, the rejection problem is how do you extract -- how do you eliminate extraneous acoustic events, noise, background, get a

confidence measure and you do the best you can with it. So, what is the performance?

I have three databases we have used over the last 15 years for just connected digit recognition, simple cases; you know, reading a telephone number or reading a calling card number or a credit card number. The first one is recorded by TI in a really antiseptic environment and the word error rate for 11 digits, 0 through 9, plus 0, is 3/10ths of a percent. We have another one in a customer care application where you can speak freely and during it, people will list numbers, you know, like they want credit for numbers and there the word error rate is 5 percent. That is a 17 to 1 increase in digit error rate.

It is a robustness problem. We are using pretty much the same models, but the good news as you work your way up is we can do a thousand word vocabulary that is -- for a number of years around 2 percent when they stop the work at a thousand. Even as you work your way up to reading spontaneous speech for airline travel information, reading text from Wall Street Journal and business publications, broadcast news, listening to anchors on the radio and translating them, switchboard is sort of what you think of, literally listening to conversations off of the

telephone switchboard. These are cases of people literally calling home.

Now, the error rates certainly go way, way up. You say to yourself, my God, 40 percent error rate. It doesn't work. Well, we have a system I am going to show you later that started off with a 50 percent word error rate and the task error rate was virtually zero. Okay? Because it all depends on what words it kind of doesn't do too well on. Here is the connected digit. That is 3/10ths of a percent. It is clean. It is beautiful speech. Here is North American Business News. Here is some from the switchboard. Very folksy, but, you know.

Here is an example of -- this one is hard to see, I guess, because it is not blown up enough. But fundamentally today we can operate 50, 60, 70,000 word vocabularies in essentially real time as you see there.

What we have been able to show as a result of all of the mathematics we have applied and all the signal processing is that whatever task we look at, we can make its word accuracy start approaching natural performance, but only approaching it. We haven't gotten there. We handle vocabularies -- in 1970, there was a two word vocabulary with a yes/no application. Today we are handling vocabularies in excess of a million words, a

million names, et cetera. So, there is no limitation of that form at all.

But if you want to kind of figure out how we are doing compared to human performance, here is the range where machines outperform humans and you notice there are not many data points in this one, practically none. Within the 10 to 1 range, there is only a couple that get on the edge of 10 to 1. Then digit recognition, here is digit recognition, the humans outperformed the machine by a factor of about 30 or 40 to 1, including those great error rates. Humans doa real -- we had operators for a long time who got their pay based on never misrecognizing the telephone numbers because that would annoy a customer. Remember, we used to have a lot of operator services.

The best ones, if you look at what those are, like The Wall Street Journal is within about a factor of 10. That is because humans don't -- you know, they do well, but it is deadly dull stuff and it is not the kind of thing you really want to translate.

So, let me take the last five or so minutes on the rest of it. Spoken language understanding is the next stage. Interpret the meaning of words and phrases and take action. The key thing here is that you can get accurate understanding without correctly recognizing every word.

Okay? So, I am going to show an application where we started with a 50 percent word error rate, but almost always got the key words that told us what the meaning was. That we actually have as a nationwide service. It is called the "How May I Help You Service." It is AT&T customer care for its long distance. It is operating nationwide today.

Also spoken -- makes it possible to let a customer speak naturally, which is why we don't understand all the words. We don't even know what words they are going to say. We just happen to know that certain words key certain things to happen and we do really, really well on those words. So, we know the ones that key the things.

So, we find what is the salient words and phrase. They map high information, word sequences to meaning and we do that really, really well. We do performance evaluation and we do these operator assisted tasks. They are real hard problems. There are challenges there.

The next phase, of course, is in order to do that, you need dialogue management and spoken language generation and that is -- the goal there is determine the meaning and what you could do based on the whole interaction, not just on that local loop. I really should have made that clear, that what you say depends on how

often you have gone through the loop. The methodology is you have a model of dialogue to guide the dialogue forward so it is a clear and well-understood goal and, obviously, evaluation of speed and accuracy of obtaining the goal, like booking an airline reservation, renting a car, purchasing a stock, obtaining help, and we use that internally.

The challenge here is how do you get a science of dialogue. How do you keep it efficient in terms of -- how do you obtain goals, get answers and is there is an -- how does the user interface play into the art and science of dialogue. We will see that there are examples where it is better to use multimodal interactions. Sometimes it is better to point than it is to --

Here is an example of a before and after the system. This was our customer care, touch tone hell system, as we call it. I will play a little bit of this. I am not going to let you go through this. It took two minute and 55 seconds to get to a reverse directory assistance.

When we started playing with -- 28 seconds right there all the way. So, the way this thing works -- we call it the "How May I Help You." The prompt is simple. You

can say anything you want. If you say pure gibberish, you are going to get pure gibberish out.

From listening to about two or three million conversations between humans and our real attendants, we figured out that people call us to find out about account balances or about our calling plans or local plans or if there is an unrecognized number on their bill or whatever. There are between 15 and 50 things and guess what? It is easier to map all of those two million and figure out what words they say that cues you into what they wanted. Okay? So, we made it work.

So, here is -- I am going to play one example from our field test labs here. This guy was irate. He bleeped out in the beginning and in the end he got exactly what he wanted. It really worked well.

The system has been in use nationwide since last November, a 37 percent decrease, repeat calls that weren't happy and 78 percent decreased customer complaints. Customers love this system.

[Multiple discussions.]

Here is a little example of how you might do it on the web-based customer care agent. This is basically how you would use multimodal to basically get you in situations where it might be -- the concept is that

sometimes better actually to point and sometimes it is better to speak and the combination is a more powerful paradigm. We play a few examples of things that would generate totally from text, never spoken by a human.

That is not a human speaking. That is a machine. Now, here is some Spanish, same text. Now, for those who prefer a more natural face, this is the exact same synthesis lined up with the face. Now, this last example is a broad band one where we use 16 kilohertz rather than -- so, again, in 40 years, what have we done?

We have put out a lot of services. There are desktop applications like dictation and command an d control of the desktop and control of document properties, stock quotes and traffic reports and weather and voice dialing. There are 20 million cell phones that are voice enabled today, access to messaging, access to calendars. Voice works as a concept of converting any web page to a voice enabled site. So, you can answer any question on line and

-- other key protocols into making this happen.

Telematics, which is command and control of automotive features, not the car, but, you know, the comfort system, the radio, the windows, the sun roof and finally smart devices, PDAs and cell phones.

So, we have come a long way. Mathematics has driven us from small vocabulary to medium vocabulary in the sixties to -- and we got to the mathematics and statistical based in the seventies and early eighties. We started going to from tens and hundreds of words to tens of thousands and hundreds of thousands and eventually a million words in the early nineties, we had a multimodal and dialogue in this time frame and finally the last one as we go forward.

The dialogue systems are going to be very large vocabulary, limited task and control. That is where we are today. In three years, we are going to be able to go through an arbitrary environment. We are going to have the robustness problem pretty much in hand and in about another three or four years after that, we will go to unlimited tasks. So, mathematics has played a key role in that 40 year evolution. Got a long way to go.

Thanks.

[Applause.]

PARTICIPANT: Why did it take 40 years?

DR. RABINER: First of all, Moore's law is the key thing. Okay? Until you get that computation, you know, I run the starting and, you know, a two word vocabulary in 1980, we had a real application and we had to

build specialized hardware for it on an operator handset. Okay? Today, you know, we can do in real time on your basic PC hundreds of thousands of words. It is a big difference when the computation in this storage is not --

PARTICIPANT: Did you in 1980 or -- did you really know exactly what you wanted to do but you couldn't do it fast enough or you didn't have the right -- the conceptual framework, how did it evolve?

DR. RABINER: You know, it is a chicken and an egg thing. In 1980, we had a sense of what we wanted to do, but it was computationally infeasible.

You learn what works and then as machines get faster, you -- gee, that is not exactly what I want to do. Do I want to do mixture density? Do I want to play games with -- we had a sense of what we wanted to do but not the power to do it.

PARTICIPANT: How did this know-how occur that you have obtained?

DR. RABINER: Carnegie Mellon particularly made their same -- they were one of the key pioneers in this. In England, Cambridge University -- and they made that public. They first were going to sell it. Not many people wanted to buy it. So, they made it open source, public domain. So, anybody, any university today can go out and

just get it all for research purposes only and use it. So, in a sense it is the most open community you can imagine.

DARPA also helped to make it open because -- almost all of the projects over the last decade have been based on DARPA responses initiative. They would issue a challenge. All of the DARPA work is open.

MR. AGRAWAL: [Comment off microphone.]

DR. RABINER: There is nothing inhibiting us. I mean, I think we know what to do. Okay? One of the things that I didn't mention is a challenge. In synthesis, the reason synthesis -- what we realized is that if you really want to make synthesis good, you know, the easiest way to do it? Record a trillion sentences that somebody says and decompose them into their basic unit and store them. It was inconceivable when this was proposed in the 1980s. We store them all and we have extremely fast retrieval.

We literally look through every one of them and say what is the best unit, in combination with every other best unit and -- in recognition there ought to be -- and that is called data driven.

So, obviously, in theory, we actually store about 50 megabytes worth. But, you know, that is just the compression technology. Get that same database. Store across a few thousand people. Just do recognition by just

looking up the units. That is one of the things we would love to do. We haven't actually done it. Why haven't we done it? Because the stuff works so well in limited environments. Synthesis has a problem because it never worked well. It always was intelligible but it sounded like -- we used to call it non-customer quality. If you listen for 20 minutes, you -- asleep. It would put you to sleep. But when you do this, you can listen to that forever. You know, it is absolutely perfect.

PARTICIPANT: Is there any intelligence that can be placed on the hand set that the receiver can be trained so that the recognition -- the server area can be --

DR. RABINER: Certainly in cell phone technology, there is a standard called Aurora(?). The goal is to put intelligence in the hand set, so you can make speech recognition in cell phones much less dependent on the transition you use.

All of the robustness techniques have that capability.

PARTICIPANT: It seems like the next doable challenge will be probably to do real time translation because it seems like you have --

DR. RABINER: Real time translation is a much harder problem. It is a much harder problem because even

if you do the recognition in one language really well, the text to text translation itself is a real problem. So, there has been a lot of studying on how do you do text to text translation. There are old classic examples of -- in time, you know, idiomatic sentences that are almost impossible -- but the answer is there are limited task domains. We have done it with Japanese and with Chinese, English. Certainly, translation is probably about a decade away because of the text to text problems, not because of the recognition or the synthesis.

PARTICIPANT: [Comment off microphone.]

DR. RABINER: You are in an environment right now -- again, if -- piece of cake. Why not?

PARTICIPANT: [Comment off microphone.]

DR. RABINER: People practice that art and who do it really very well. There are certain pieces of things -- user interface is an art. Dialogue management is an art. So, we have to make it into a science.

PARTICIPANT: -- speaker recognition?

DR. RABINER: That is a completely different topic. Okay? Speaker recognition is something that will probably never be done. Speaker verification is what you really want to ask about. Okay?

The answer is -- obviously, you can do it with some performance -- whether you use retinal scan, whether you use fingerprints, all of these have various -- false alarm rates. The best systems -- achieve 98, 99 percent. Verification accuracy with something on the order of a half to 1 percent false projection. Okay? Is that good enough? Well, certainly it screens out some things. It is very easy to go at in some systems and just record your voice. You can't use a fixed, you know, verification phrase. You are dead as soon as somebody records you.

You have to change it all the time and once you change it, that performance starts edging down a little bit because you are a variable yourself. But the technology that as not good enough -- you can say I am going to guarantee to keep all these bad guys -- but it is a pretty good return technology.

MS. CHAYES: Why won't it be done? I mean, we can do it ourselves, right? You hear a song, you know who is singing it. You mean it won't be done on a short time scale or it won't be done on a --

DR. RABINER: Can you listen to somebody you have never heard before, listen to then two or three or four times and then -- I will let you listen to 30 people you

have never heard before five times each and then a week later I am going to give you one of those 30 and then --

MS. CHAYES: The computer has a better memory than I do.

DR. RABINER: Not much.

MR. MC CURLEY: [Comment off microphone.]

DR. RABINER: Right now, what this whole concept of this broadcast news is to say can you record all of these and use that as a searching index.

MR. MC CURLEY: I guess my question was would you convert it into text or would you convert it into something else.

DR. RABINER: No, you convert it into text. Okay? What it would do is it would do real well on the significant words. Okay? So, if you listen to that North American Business News or whatever, the key words were all perfect. "For" or "it" or "this" gets wrong a lot of the time. So, the bottom line to this is you would align that text with the speech. Find all occurrences of Saddam Hussein. Okay? And guess what? You will find most of them absolutely perfectly. It will miss a few. And it will have a couple of false alarms, but you can check. You click on each one. That is not what I want. That is not what I want.

So, it is not perfect, but it gets you a long, long way to where you want to be.

You come back and you have 30 voice mail messages. You don't go through them sequentially. Okay? We have a system which we call ScanMail. It was actually written up in The New York Times yesterday. ScanMail converts every one of your voice messages, to text and gives you a chance to scan it as though it is text mail. More than that, if you are really, really good at finding numbers, call me back at 917-3211-2217. It will find that and it aligns it with it. You can click on it. So, it gives you a synopsis. Synopsis isn't perfect, but most people find that you actually get through that and what it does is it gives them the triaging capability.

So, if you are a sales person with 30 -- I get one voice mail message every month. E-mail, I get 75 a day. But there are sales people who get 30, 40, 50 voice messages a day and very little e-mail and they are the people who love this kind of capability.

PARTICIPANT: [Comment off microphone.]

DR. RABINER: They come down to making things more efficient, so you can try more things.

PARTICIPANT: [Comment off microphone.]

DR. RABINER: Well, the text-to-text translation is a really tough problem. Okay. And very likely, most terrorists don't give you those free --

[Laughter.]

Obviously, homeland defense is -- you know, the speaker recognition, the verification. They go out in the surveillance. Having something that converts it into text and looking for some key words. They are not going to be the key words you wanted just for the reason that you can't really listen to everything that goes on. No one can do that.

MR. EUBANK: You talked a lot about robustness, but the kind of robustness you were describing is noise and things like that and not an adversarial --

DR. RABINER: What is adversarial in speech recognition?

MR. EUBANK: [Comment off microphone.]

DR. RABINER: Yes, probably, but, you know, you would to be much more specific. I just can't give you an easy answer on that one. Sorry.

I am going to turn this back to you.

DR. LENCZOWSKI: I want to thank you for very stimulating discussions. With that, I was going to turn this over to Dr. McLaughlin, whose responsibility during

this session was to listen and to capture some of the challenges, the complexities that we should be thinking about in terms of the presentations.

Challenges in Speech Recognition

Larry Rabiner

There are two main challenges in speech recognition. To begin, many speech recognition systems are very task-specific, with limited vocabularies. In addition, speech recognition systems are often very environment-specific, performing poorly when tested with speaker phones or cell phones, or in noisy environments such as restaurants, airport lounges, or moving cars.

It is easy to take a well performing, finely tuned system and make it so broad that it works terribly all the time. Further, when a system is designed in an environment with a set of speakers and a set of inputs, then in a new environment, the signal is mismatched.

There are three models for speech recognition:

1. *The acoustic model*. This model is based on physical speech data and the numerous types of sounds it must distinguish. One powerful approach in this domain is a Markov model, which is a statistical method for characterizing the special properties of the signal. The system works well in the given environment, but is nonrobust.

2. *The word lexicon*. This is a dictionary bridging legal phenome sequences and words, according to rules. The mathematical challenge is in generating a word lexicon automatically, given that new words are often introduced to a language and the lexicon must handle a range of pronunciations.

3. *The language model*. This model builds valid sequences from words and phrases, based on the syntax of the recognition task. Like the acoustic model, it assumes that people will speak according to a certain text model, making this model nonrobust.

The goal is to determine the most probable word sequence, given a task (which defines the syntax), given a set of words that could be spoken (which is the lexicon), and given the sounds of the language. The current state of the art is a decoder that searches 10^8 possible recognition choices whereby combinations are weighted and minimized. The final stage is "utterance verification," where key words are identified. Currently, speech recognition systems have very large vocabularies and limited tasks. By 2006, systems should be able to perform effectively in arbitrary environments, and by 2010, systems should be able to perform unlimited tasks.

David McLaughlin
"Remarks on Image Analysis and Voice Recognition"

Transcript of Presentation
Summary of Presentation
Video Presentation

David McLaughlin is a provost and a professor of mathematics and neural science at New York University. His research interests are in visual neural science, integrable waves, chaotic nonlinear waves, and mathematical nonlinear optics.

DR. MC LAUGHLIN: So, I am going to try to be very brief and just make a few remarks. The first remark I wanted to make concerned comments this morning about how do -- does the government or any interested set of parties get mathematicians and scientists interested in homeland defense. Rather than try to answer that, I just want to -- I want to state what I think are three of the major problems that have to be overcome.

The first is that in our guts as scientists and mathematicians, we don't see the immediacy and the urgency of the problem in our guts. Intellectually, we say it is a big problem. But we all hope that we have had the last terrorist attack. That is combined with two other things that we realize.

The second is that this isn't one problem, homeland defense. It is thousands of problems and as mathematicians or as scientists, we tend to attack one problem or one area of problems, if you like. This is pretty -- a pretty broad list of things for us to sink our teeth into.

You combine that with how we know we work, we work over 40 year time frames and we are being -- and we think if we do think about the immediacy of the problem, we are worried about what is going to happen in the next six

months or a year. It seems to me that you have to face those three realities and I certainly don't have any solutions, but it does seem to me that those sorts of realities have to be faced by people concerned with getting scientists and mathematicians immediately involved in homeland defense as they were in the Manhattan Project, where some of those problems were not necessary to overcome or else the scientific leaders realized they were not necessary to overcome. I don't know the history of the time.

That was the first point I wanted to make. A second point I wanted to make concerned the small needle in a large haystack that has been referred to many times today and it refers to it somewhat indirectly, but I wanted to share with you a story that I have heard several times in recent months.

There is a man by the name of Steve Flynn. Steve is a high officer in the Coast Guard, perhaps a lieutenant commander, but in any case a high officer, who this year is on leave from the Coast Guard in one of the government think tanks and his task is to think about border security. He is concerned with two things. That he wants to protect our borders;on the other hand, he doesn't want to close the Canadian border unnecessarily for financial reasons.

So, he tells this story, which I will try to abbreviate it because it is late, but it is rather disturbing. He talks about one problem of preventing a bomb, perhaps a dirty bomb, from being brought into the country through shipping. He talks about how in Pakistan, crates are loaded onto something and they get to Hong Kong and in Hong Kong, they are stored more or less in the open, after which they are loaded onto boats and come to San Diego.

In San Diego they are loaded onto trucks. There is no inspection of this process yet. They are loaded onto trucks and nobody monitors where those trucks go. They know where they are going to end up. Perhaps they are going to end up in Newark at the -- near the Newark Airport in a storage facility, where they are -- where whoever has bought this equipment will pick them up.

It is there that Customs will pay some attention to them. But in the meantime, they have gone through Chicago. They have taken detours totally unmonitored, depending on the truck drivers. Once they get in Newark, it would be conceivable for a phone call to detonate.

The question is how do you discover which, if any, of the crates has a bomb in it? Now, Flynn's response -- so this is definitely a small needle in a large haystack

and his response is that it probably cannot be done at that point in time. In fact, he is urging people to monitor much more closely the packing of the boxes in -- at the entry point.

The reason I mention that is -- and it is an obvious point, but when you are dealing with a small needle in a large haystack, it seems to me you want to carefully define that needle. Flynn is defining it as entry point, but in each case, I think we want to very carefully define the needle that we are seeking. So, that is the point two that I wanted to make.

The last point I wanted to make is slightly more scientific and it returns to a point that George Papanicolaou made this morning, when he was -- when he mentioned a comment about general data mining and he was urging people when possible to use science in addition to statistics.

He used the example of the physics of earthquakes versus sound detonation. In what we talked about today, this afternoon, I want to focus on a few remarks about computer vision, although I believe similar remarks would apply to speech recognition and so forth. But as Professor Malik knows and if you read his work, he uses, it is very wise in computer vision to attempt to use biological vision

to guide our thinking and construction of computer vision algorithms.

The problem is we know very little about real vision. But when you think about how visual information is processed in the cortex, there are a few things that we do know and I just want to close by listing some of them and I am sure that the computer vision experts (a) know these, (b) undoubtedly try to use them, but I also believe that since we don't understand how they are used in the human and mammalian visual system, we haven't yet unveiled how they should be used in the computer visual system.

In any case, when you think about the cortical processing of visual information, the first thing that is completely clear is that that region of the brain is a massively complex feedback system. There are extensive feedbacks between regions of the brain and many of these regions of the brain are themselves layered structures. There is massive feedback between the layers.

So, if you like, you can almost think about feedback within feedback within feedback. You can ask questions like how is the brain making use of the fact that the architecture is layered? People don't know the answer to that. People do know which regions are coupled to which. They know that is based on anatomy and they know

very well which layers are coupled to which, also based on anatomy. They don't have much feeling for the strengths of those couplings.

For example, the visual system is often modeled as a feed forward network where information enters through the retina and it goes on to a filter called the LGN and on into the visual cortex; two specific layers of the primary visual cortex, 4 and 6. Now, it is known anatomically that there is a huge amount of feedback from 6 back to the LGN. In fact, there is more feedback anatomically from 6 back to the LGN than there is from the LGN forward to 6.

Yet, nobody knows within the visual system what that feedback is doing. It must be doing something because there is so much of it anatomically. Another issue is time scales. People are beginning to wonder is there temporal delays between these different regions for which the feedback might be slightly slower or might be slightly delayed and, in fact, it could be delayed within delayed within delayed when you think about regions within regions within regions, which seems to be the architecture.

It is quite controversial as to how important these delays are. David Mumford and LeMay and Lee have been concentrating on delays for several years now. Other neuroscientists focus upon the fact that there are

precursors in these delay processes that are almost instantaneous.

But in any case, when you think about those features of the real visual system, namely, layered structures, massive feedbacks and the importance of temporal dynamics, some of the more static single layer or two dimensional image reconstruction that is done in computer vision, there is a chance that by better understanding the real visual system or even if we don't understand the real visual system by thinking about possible uses of this architecture and delay might well improve the computer vision system.

Now, this is saying nothing to Professor Malik, who does this in his work continually, but it is a fact that I think that the computer vision community doesn't talk with the biological vision community enough.

So, I think that that is the only three points that I really wanted to make and it is late and I think I will conclude with that.

Remarks on Image Analysis and Voice Recognition

David McLaughlin

How can the government, or any interested set of parties, get mathematicians and scientists interested in homeland defense? Some of the challenges are these:

1. As mathematicians and scientists, we don't see the immediacy and the urgency of the problem. Intellectually, we *say* it's a big problem, but we all *hope* that we've had the last terrorist attack.

2. Homeland defense is not a single problem. It's thousands of problems. But mathematicians and scientists tend to attack one problem, or one area of problems, at a time, and there is a pretty broad list of things for us to sink our teeth into.

3. We work over 40-year time frames, and if we do think about the immediacy of the problem, we worry about what's going to happen in the next 6 months or a year.

Many of the problems in homeland security involve finding a small needle in a large haystack. One way to deal with this problem is to put more effort into carefully defining that needle. Dr. McLaughlin illustrated his comment with a hypothetical example, attributed to Steve Flynn, a high-ranking officer in the U.S. Coast Guard, whose current task is to think about border security.

It is necessary—say, in data mining—to use not only statistics but also insights from whatever other scientific realms inform the problem, even if they are distant from the analyst's own field. In computer vision, for example, it would be wise to attempt to use biological vision to guide our thinking and construction of algorithms.

David Donoho
"Remarks on Image Analysis and Voice Recognition"

Transcript of Presentation
Summary of Presentation
Video Presentation

David Donoho is the Anne T. and Robert M. Bass Professor in the Humanities and Sciences at Stanford University. He has worked on wavelet denoising, on visualization of multivariate data, on multiscale approaches to ill-posed inverse problems, on multiscale geometric representations, such as ridgelets, curvelets, and beamlets, and on manifolds generated by high dimensional data. He received his A.B. in statistics at Princeton and his Ph.D. in statistics at Harvard. He is a member of the National Academy of Sciences and the American Academy of Arts and Sciences.

DR. LENCZOWSKI: Once again setting up electronically here -- wouldn't it be nice to have voice recognition and simply tell it to connect.

With respect to the voice recognition, that there were only two words there in 1980 because I know in 1981, that system we had recognized or purportedly recognized 10 if properly enunciated.

The next discussant that we have here to make commentary on, again, the complexity of the various topics is David Donoho. Interestingly enough, he was also a reference in one of the previous presentations. So, he has the opportunity to determine whether or not he was represented properly.

MR. DONOHO: Well, the cardinal difficulty I have to face here is I am standing between you and a reception. So, be brief, right?

I wanted to put in context an idea that I think it is implicit. Everybody already accepts, but I might as well just say it out loud. Why do we think that mathematics is relevant to homeland security? It is because we are entering a data saturated era. We are seeing a one time transition to ubiquitous measurement and exploitation of data and we believe that mathematics is relevant to that.

So, I just want to go into that and talk about some of the obstacles not yet addressed and that I think were exemplified in some of these talks and isolate a few mathematical challenges, but I can't possibly be exhaustive.

So, while world hunger has been with us for a long time and will continue to be with us sadly, we are at a point now where we will never be like Sherlock Holmes. We will be inundated with so much data that we will never feel that we don't have what we need to make decisions. The era that we are entering we have got the prodigious efforts of the best and brightest in our society are dedicated to create huge data repositories and exploit it all sorts of ways. The information technology business is a huge part of our economy. It dominates corporate America. It is like 50 percent of Fortune 500 budgets in information technology.

So, it is clear that this is an important phenomenon but it is really, let's say, pushing the thinking of everybody about how we ought to go about things, at least Internet he United States. So, if you look at the definition of an era, of course, there were, you know, great minds in the Renaissance and in the Middle Ages that defined the era for them by, for example,

building wonderful structures or creating great works of art.

What we are doing instead is creating networks -- pervasive networking, gargantuan databases and promulgating an ethic of data rich discourse and decisions. So, homeland security that we can contribute to, I guess, everyone is assuming it is all about data gathering, data analysis and then somehow we are going to find in this mountain of data things that are going to be useful.

For example, if we had a national I.D., there would be some way to codify that into a process that would let us keep people from doing bad things to our citizens. Okay? So, it is important to say how far this really goes. I mean, anything can be quantified and data really can be gathered about everything and we really will be able to encode all sorts of things about human entity or behavior.

People would say smells are an example of something that is merely poetry, but actually now we have electronic noses, arrays that have excellent abilities to distinguish between let's say roses and garbage or more to the point, they can distinguish between, for example, different hydrocarbons like in this case methane and benzene and so on. They could do a job that wouldn't be so safe for humans to do.

As Professor Coifman pointed out, we have hyperspectral photography that goes considerably beyond the abilities of the human eye, getting thousands of spectral bands and able to identify what things are made of, not merely what they looked like, simply by looking at them. Human motion tracking, there are now sensors that can record gestures, pose and the articulation over time and so on. So, all this data is being gathered and I could only sketch here, just mention a few examples, but it goes on and on.

So, certainly all aspects of human identity, date and posture will be captured and digitized completely. This goes to the point that the NBA is currently -- has a web site where they have nine weekly games and talk about amazing moves that people made that were atypical for that player.

Okay. Now, we all know that there are people who are naysayers about the data era and they would say that there is going to be a violation of privacy or morals and then there are people who say that calculating all this data and compiling it is just wasting time and you are just going to be lost in this glut and be paralyzed by the fact that you have the information.

So, if we look at the issues for homeland security, I guess I see the following things. So, first of all, any time you talk about compiling data, for example, around imagery or around speech, something like that, there is going to be in this city in particular a big reaction, which is based around political maneuvering and there will be always issues of -- there will always be issues of civil liberties and so on.

Secondly, and this was mentioned before, there is also the issue that in our system we do things by private enterprise and so there is going to be a lot of trade speech claiming benefits and workability, which is quite different from the actual experience of some of the customers who tried to do these things.

In addition, there are massive engineering issues -- there are massive engineering problems associated with this, that just merely to have something like a national idea or to use biometrics in a consistent way that would really close off from undocumented people the ability to get on airplanes and so on, just to create that system is very challenging and there would be a lot of discussion about whether you could actually build the infrastructure.

But none of those things are science. So, it is really not Ralph Nader or the Pope or the data glut

arguments. Suppose all of those things are solved, we really should just face what the scientific issues are in using all this data. I see them as being -- if we wanted broad issues identified, generalizability, scalability, robustness or lack of fundamental understanding of the structure of the data.

Generalizability would mean you can build these small systems in a research lab, but then you can't deploy them through the level of dealing with a quarter billion people in a satisfactory way. Robustness would be, again, work in perfect conditions, but you can't deal with articulated conditions, where something is buried away from the research lab.

A lack of fundamental understanding has to do with the fact that a lot of what we are doing is using what mathematics gave us 40 or 50 years ago and not developing mathematics that tries to understand what the underlying issues out of the data that we are trying to deal with is. We are just getting what -- how can we deploy the stuff that is out there.

So, it seems to me that all of those things are due to the fact that mathematics hasn't been going where the data resolution is pushing us, at least not enough to make a difference in the directions that we like. Where

the data revolution is pushing us is dealing with very, very high dimensional space. So, if you look at an utterance or an image, you are talking about something that is a vector up in a thousand or a million dimensional space.

So, what we really need to think about is what is going on in high dimensions. I just mention here some examples, human identity, even highly decimated . It might be 65,000 variables just to take raw imagery, hyperspectral image. For a chemical, you have a thousand spectral lines for a sample. If you went to an image, you would have a thousand times the number of pixels.

So, what are the needs with high dimensional data? I see a couple of things that came out in the talks today. Coping with dimensional glut, understanding high dimensional space and representing high dimensional data in qualitatively new ways. I think in various ways if mathematicians make major progress in these areas, there will be long term payoffs. Again, it might be just 40 years from now when the real problems are something else. But these things will get ultimately inserted in deliverables.

One issue about dimensional glut, which underlay a little bit, when you have too many dimensions, often what

we are seeing in modern data sets is very, very high dimensions about each individual under study. So, an image is like maybe a million dimensions about one particular thing you are looking at.

Traditional statistical theory says that the number of dimensions compared to the number of observations should be such that you have a very favorable ratio; namely, D over N. It is small. But where Moore's law is pushing us is D over N is going to infinity in some sense. We are getting more and more complex articulated data about just a few individuals and we are not getting more and more data per se about independent observations.

As a result, we find it very difficult to train from this enormously high dimensional stuff and then get something that generalizes. Just as an example, I went out on the web and found a class project on Eigenfaces. So, you use principle components to find the underlying structure of face space in some sense. So, here is a class and there are their Eigenfaces. The thing is that what they have is each person is like a 16,000 dimensional vector. Right?

Then there is only, I don't know, 30 kids in the class or something. So, it is exactly the situation that I was talking about. When you look at Eigen analysis on

this, what you find is that basically each person is equal to one Eigenface, take away the average. So, actually what happens is that the Eigen analysis just gives you back the individual identities and so on.

Now, Eigen analysis is being used all the time throughout science and people are exactly going down this road all the time. It is not well-understood that if you go to high dimensions and you have relatively few examples, you have these distortions. The whole area of what distortions there are and what systematic issues you have are not very well understood.

Another area of dimensional glut is that you have the cursive dimensionality. Algorithms don't run well in high dimensions. A simple example in the context of the talks that we have had is say I have to search through a database for nearest neighbors. That gets slow as you go up into high dimensions and there is a lot of theoretical research that has tried to crack the problem, but basically it has not been really solved in any practically effective way. So, basically you just have to go through a significant fraction of the data set to find the nearest neighbor.

No ideas of hierarchical search really work in these way high dimensions. For example, Jerry Friedman in

the audience pioneered the idea of using hierarchical search. It works great in low dimensions, like ten dimensions or something, but you know, when N is much smaller than D, it is not a viable method.

Well, the other issues we just don't understand enough about what our data structures are that we are gathering. They are these things up in high dimensional space. There isn't enough terms of reference for mathematics that has been provided to us to describe it. We lack the language. It is really the responsibility of math to develop that language.

So, a lot of these things are about point clouds and high dimensions, just individual points are pictures or sounds or something. They are up in high dimensions. If we collect the whole data base, we have a cloud of points and high dimensions. That has some structure. What is it? We believe that there is something like a manifold or complex or some other mathematical object behind it.

We also know that when we look at things like the articulations of images as they vary through some condition like the pose, that we can actually see that things follow a trajectory up in some high dimensional space notionally, but we can't visualize it or understand it very well.

So, people are trying to develop methods to let us understand better what the manifolds of data are like. Tell me what this data looks like. Josh Tannenbaum, who I guess is now going to M.I.T., did this work with EisoMap(?), which given an unordered set of pictures of the hands, okay, figures out a map that shows what the underlying parameter space is. So, there is a parameter that is like this and there is -- that is this parameter and then there is a parameter that goes like that and that is discovered by a method that just takes this unstructured data set and finds the parameterization.

We need a lot more understanding in this direction. The kinds of image data that you see here, this is just one articulation, one specific issue and one methodology for dealing with it. But we could go much further.

Representation of high dimensional data, Professor Coifman mentioned this and in some of Professor Malik's work it is also relevant. What are the right features or the right basis to represent these high dimensional data in so that you can pull out the important structures? So, just as an example, if you look at facial gestures, faces articulating in time and you try to find how do you actually represent those gestures the best, one

could think of using a technique like principal components analysis to do that. It fails very much because the structure of looking at facial articulation is nothing like the linear manifold structure that, you know, 70 years ago was really, you know, cutting edge stuff, when Hoteling(?) came up with principal components.

It is a much more subtle thing. So, independent components analysis deal much more with the structure as it really is and you get little basis functions, each one of which can be mapped and correlated with an actual muscle that is able to do an articulation of the face. You can discover that from data.

It is also nice to come up with mathematically defined systems that correlate in some way with the underlying structures of the data. So, through techniques like multi-resolution analysis, there are mathematicians designing, you know, bases and frames and so on that have some of the features that correlate in a way with what you see in these data. So, for example, through a construction of taking data and band pass filtering it, breaking it into blocks and doing local redon(?) transform on each block, you can come up with systems that have the property that they add in little local ridges and so on. So, you can represent with low dimensions structures like fingerprints

because you have designed special ways of representing data that have long -- elongated structures and it is also relevant to the facial articulation because the muscles and so on have exactly the structure that you saw.

Okay. So, just some of the points I have made, we are entering an era of data, the generalizability, the robustness, the scalability that have come up in some of these things are issues that will have to be addressed and in order for that data to deliver for any kind of homeland defense, where mathematicians can contribute is understanding fundamentally the structure of this high dimensional data much better than we do today. It can provide a language to science and a whole new set of concepts that may make them think better about it.

Our challenge is to think more about how to cope with dimensional glut, understand high dimensional space and represent high dimensional data. The speakers have brought up the issues that I have said in various ways and trying to -- you know, under one umbrella, cover all of this is pretty hard, but certainly in Professor Malik's talk was brought up the issue of robustness to articulation. We don't understand the articulation manifold of the images well enough. As an object in the world is articulated, the images that it generates go

through a manifold in high dimensional space. It is poorly understood. IsoMap is an example of a tool that goes in the right direction.

But to get robustness to articulations, it would be good to understand them better. The structure of image manifolds, as Professor Malik mentioned, working directly with shape, that is again trying to get away from a flat idea of the way data are in high dimensional space and come up with a completely different way to parameterize things.

Professor Coifman, well, he basically like we are co-authors, so I am just extending his talk, but, anyway, the connections there are pretty obvious.

Finally, I think what Rabiner showed is first of all the ability to be robust to articulations was mentioned. So, again, to understand I think how those strange noises in the data -- what is the structure of the manifold of those things up in high dimensional space. So, the structure is speech and confusers would be interesting to know.

I think he has shown that if you work very hard for a long time, you can take some simple math ideas and really make them deliver, which also is a very important thread for us to not lose track of. He mentioned a number of math problems that he thought would be interesting. I

wish he had been able to stay and we could have discussed those at greater length.

There was a questioner in the audience who said, well, but you said you had just been carrying out an implementation for 20 years. Are there any new math problems that we really need to break now? It would be very interesting to hear what he had to say.

I think all the speakers have pointed out that a lot of progress has been stimulated by -- in math, I think, from the study of images and -- associated multimedia data and I think that will continue as well and eventually we will see the payoff in homeland security deliverables, but perhaps a long time.

Thank you.

[Applause.]

PARTICIPANT: [Comment off microphone.]

MR. DONOHO: IsoMap is just a preprocessor for multidimensional scaling, for example. So, for instance, you didn't have this standard method multidimensional scaling, we wouldn't be able to -- just provide a special set of distances.

Has any of this played out? I mean there are 10,000 references a year to that methodology. I don't think I could track them all down and make a meta-analysis

of how many were just, you know, forgotten publications. But it just seems like millions of references --

PARTICIPANT: [Comment off microphone.]

MR. DONOHO: We are right now getting data in volumes that were never considered 20 years ago. We have computation in ways that were not possible 20 years ago; whereas, the methodology of 80 years ago, just with the data and computations that we have is starting to deliver some impressive results. So, if we are talking about more recent methodology, well, let the data take hold and let some computations take hold.

MR. MC CURLEY: [Comment off microphone.]

MR. DONOHO: Your earlier question about what features would help to index, I think it is connected with that. I think that the idea of looking for nearest neighbors in a database of all the multimedia experiences in the whole unit, it is sort of where this is going. Right? That is not going to happen. There ought to be a more intelligent way to index, to compress and so on.

Can we make a decision right now how much data to gather? I don't know. I think economics will drive everything.

PARTICIPANT: [Comment off microphone.]

MR. DONOHO: It is an excellent question. I have some opinions on the matter but I don't know if I have really the right to stand between people and liquid refreshment.

Yes, I think that most of the political things that I have seen are between non-technical people who are playing out the dynamics of where they are coming from rather than looking at really solving the problem. So, you know, I am a little bit -- so, for example, a national I.D. card wouldn't have any privacy issues if all it was used for is for -- you know, it contained data on the card and it was yours and it was digitally signed and encrypted and all of that stuff and there was no national database that big brother could look through.

Actually, the point was that it was the card and that was it. Right? What people assumes it means there is that big brother has all the stuff there and they can track their movements and stuff like that. So, there are all those kinds of things. There is nothing that says the I.D. card can't just be used for face verification or identity verification. You are that person and that is it. They are not communicating with anybody. It is illegal to communicate with anybody. That can all be done.

But people get together and butt heads is sort of an Orwell characterization I guess. The science is absolutely clear, that you could do it without having all of those -- violating personal liberty and so on.

DR. LENCZOWSKI: Are there other questions for any of the speakers? I guess that everyone is ready to take their break.

I want to thank you very much for your participation this afternoon.

[Applause.]

Remarks on Image Analysis and Voice Rcognition

David Donoho

Our era has defined itself as a data era by creating pervasive networks and gargantuan databases that are promulgating an ethic of data-rich discourse and decisions. Some scientific issues that arise in working with massive amounts of data include these:

- Generalizability–building systems that can efficiently handle data for billions of people
- Robustness–building systems that work under articulated conditions, and
- Understanding the structure of the data–techniques for working with data in very high dimensional space.

Take, for example, the task of coping with high-dimensional data in body gesture recognition. If you look at an utterance or an image, you are talking about something that is a vector in a 1,000- or 1,000,000-dimensional space. For example, human identity, even highly decimated, might require some 65,000 variables just to produce raw imagery.

Characterization of facial gestures is a subtle thing, and independent-components analysis (as opposed to principal-components analysis) deals much more with the structure as it really is. You get little basis functions, each one of which can be mapped and correlated with an actual muscle that is able to do an articulation of the face.

Opening Remarks and Discussion, April 27

Transcript of Presentation

DR. CHAYES: I'm not exactly sure what we are going to do in this session. We have a great session coming up at 9 o'clock.

Let me tell you a couple of things about the room. We are filming this for the MSRI website, so that it will be archived. This is the only mike which we will hear in the room, but all the speakers should attempt to speak into those mikes, even though they won't seem as if they are being miked, because it is being picked up for the film. If you can remember, when you ask a question, it would be great if you would go to that mike over there and speak into that, just so that this is archived better.

The second thing is that the temperature in the room -- I know that some people were freezing yesterday and other people claimed to be quite warm. It is considerably warmer on that side of the room, so if you are freezing, you can do your own temperature control by just moving from one side of the room to the other, if we haven't gotten the heating or cooling system working properly in here, which it seems like we haven't, actually, given the way the room feels.

The purpose of this session, which I was just told a few minutes ago that I am chairing, is just to try to summarize what we talked about yesterday. This is a

very unusual workshop, as anyone who was here yesterday realizes. We have got people who are actually practitioners in various fields, which are related to homeland security, although the speakers themselves may not have thought in much detail about homeland security in the past. We have also got mathematicians, some of whom are really core mathematicians.

What we hope to come out of this is that a lot of mathematicians who have not really thought about doing applied work in the past or certainly who have not thought about doing anything like homeland security, are now thinking about it. A lot of mathematicians who haven't made contributions to these kinds of efforts in the past want to make contributions in the future, so we need a research agenda and we don't want to do what mathematicians sometimes do, which is just make up a problem and then write in the beginning of your NSF proposal that this is relevant to such-and-such. We really want to talk to people in the field to find out what their real problems are, and set a research agenda for the community that people who are interested can get involved in. Hopefully, it will be interesting enough that many people will want to get involved.

I know that there were some comments yesterday that we didn't have time for, so I am hoping that some of you who had general comments to make -- I know that Andrew made some interesting general comments, but if others of you have general comments on homeland security, on the role of mathematics in homeland security, and if you are prepared to make those comments at this hour of the morning, this is your big chance.

Does anybody have comments? I'll start calling on people. Sally. And, Sally, can you speak into that so that we can record you for posterity?

DR. KELLER-McNULTY: I don't need to, because I was going to point to Kathy, who was making some important comments outside during coffee.

DR. CHAYES: Oh, Kathy Laskey, wonderful.

DR. KELLER-McNULTY: So I call on her to stand up and do that.

DR. LASKEY: As I said to a few of you at the reception before and also in comments outside, I think we really have to take a systems approach to the homeland security issue. We think of ourselves as being part of a system that involves equipment, people, processes, and we want to improve our overall security process, and

mathematics plays a role in that. There are important mathematics challenges.

One of the things that I want to avoid is people thinking I have this mathematical algorithm that I am going to use to solve problems of homeland security, or this mathematical theorem that I have just proved is going to solve homeland security. What you want to do is look for aspects of that system that can be improved.

What we want to look at is the critical aspects. We want to be able to analyze the whole system and say, what are the bottlenecks, what are the problems. It may be that I can fix this piece of the system and it wouldn't do anything for overall system performance. It may be that the driver is something else. So we have to look at how all of the components of the system interact, and that in itself poses mathematical challenges.

We were talking about gain theory. We were talking about economic and gain theoretic models of the actors in the system playing against each other. In order to solve those kinds of challenges, it requires the mathematicians to work with the political scientists, the anthropologists, the organizations, the psychologists, to analyze behaviors of organizations and what happens when we

do this and they do that, and the gaming aspects of it. But all those things have to go together as a system.

DR. CHAYES: Let me ask you a question about that. What about mathematics? How does the system approach involve mathematics? Or are you simply saying that mathematicians must interact with these other groups?

DR. LASKEY: There are mathematical systems. There are mathematical challenges in analyzing a system, breaking it into sub-systems, issues of modelling pieces of the system and how they interact, and different resolutions. I can model the economy at the macro level by flows of currency, or I can look at individual micro -- I forget the name of the person who gave the talk yesterday, where he was talking about the agent simulations. I can make the connection between the micro and the macro behavior.

There are definitely mathematical challenges in that. There are system architectures we can design. Suppose we design our airport security architecture this way. This is what happens when people walk in the door. They go up to the counter and they present their security, and then they go through the lines and they do this. You can build a simulation of that, a mathematical model of it, and then simulate it, and then analyze different changes in

the system architecture and how that will impact on airport security. If I increase the sensitivity of this sensor when I am putting my luggage through, how is that going to impact on overall security?

DR. CHAYES: And the economic implications of doing that also.

DR. LASKEY: It is benefit tradeoffs, right.

DR. CHAYES: Yes, cost-benefit.

DR. LASKEY: But the issues of looking at these things not as a simulation that gives as an answer, because I don't think we can build a giant simulation of our security apparatus and then say, let me change this parameter and see what the classic implications are and the security implications. But we can analyze pieces of it. We can try to think globally as a system.

DR. CHAYES: Peter?

DR. BICKEL: I think what has to distinguish the short term and long term effects and the interaction with mathematics. I think Kathy has described the short term interactions which could benefit directly homeland security.

On the other hand, we had yesterday a longer term question, which would be called for not only by the homeland security concerns, but more generally by society.

I am referring to Dave Donoho's presentation, or the discussion from Coifman. There you have these large problems, contributions to which will hopefully, in fact, I think almost surely, will move back possibly to directly affect how one can deal with problems of homeland security. So I think one has to distinguish between --

DR. CHAYES: Any other people volunteering to make comments before I call on someone? I would actually like to hear from someone from one of the funding agencies, or one of the agencies that would potentially fund.

I see people almost heading towards the door at this point, looking away. This is what I do; I look away when I go through the airport security, so that they won't choose me. So I see all the people I know from the agencies, looking away. It is effective sometimes, but I know your names.

But seriously, I think that one of the ways in which the mathematics community moves, since we are a fairly conservative community, is that we are energized by some of the federal agencies putting funds in various places. So I was wondering if anybody from the agencies would like to speak to the question of, if people here want to start doing work on some of the problems that we have discussed, how do you deal with very high dimensional

systems and some of the other problems that we have talked about, where would this fit in, in NSF and DARPA and DoD? Where would somebody apply to do this, and what is the infrastructure that exists already, or that we might want to implement to support these kinds of efforts?

Deborah? You knew I was going to call on you.

DR. LOCKHART: If I had to describe NSF's mission, it is to support basic research in science and education. So we respond to proposals that come. We don't necessarily put out special calls for proposals in a particular area, although sometimes we do.

What I would say at this point on the best thing someone could do if they wanted to make a proposal in this area is simply to submit a proposal. We have a number of programs in our division that would certainly welcome such proposals.

My own program in applied mathematics, I can see a number of the issues that we talked about yesterday being relevant to that. We have a program called computational methods for statistics and probability that I think would also be very, very response to proposals in these areas. So there already exists this venue for individuals who want to do research.

But I want to follow up on something Felipe Hondure said yesterday. There is another vehicle that we started two or three years ago that I think can be very, very useful for those of you who are seeking support. That is a relatively new vehicle, research groups in mathematical sciences.

We are currently beginning the process of recommending awards in the third round. The purpose of this program is support groups of researchers to work on what they think are important problems. These are the kinds of problems that require the collective expertise of either a group of mathematicians working together or a group of statisticians working together or mathematicians and statisticians working together with people in other disciplines. So the proposals can either be multidisciplinary or not, as the problem is described, and in terms of what is demanded in terms of expertise.

The duration of such grants is three years, and the funding varies from $150 K to $350 K per year. So a number of the grants approach on the order of close to a million dollars over three years, which can support students, postdocs, et cetera. What is important is the timeliness of the problem, and I can't think of things more timely than this, of course the scientific quality, the

fact that a project has to make the case that the results will reflect -- that the group will be more than the sum of its parts.

So that is certainly a vehicle. Now, in terms of when proposals come in, we are in the process of putting together our new solicitation for that right now. I would expect that the required letters of intent would be coming in sometime in August, and the proposals would be due sometime in mid-September, so there would be sufficient time. So it is not a hurry up, get this in tomorrow kind of program, but something that could reflect these middle and longer term issues that so many of you have talked about over the last day or so.

So I think right now there are ways we can use our existing venues. I don't know if there is going to be a special kind of money available at NSF to broaden these kinds of things. That will be up to the President and Congress. But we don't have anything right now that -- but I suspect that if we do, we will be hearing about it.

DR. KELLER-McNULTY: Jennifer, I'm going to pick on another person. I am going to make Sally get up and reiterate some of the things we were talking about at the reception in terms of trying to think of how --

DR. CHAYES: This is not being picked up, so perhaps you can come up here or go over there.

DR. KELLER-McNULTY: To me, part of the issue is how do we actually start to mobilize an effort in certain areas to try to look at the complexity of the problems we are talking about, both short term and long term.

At the reception we were talking about that a little bit, and Sally actually brought up some really good examples of historically where this has been done in the areas that started out with some problems, probably equally vague to some degree as what we are talking about in the homeland security.

So Sally, do you want to comment on some of that? I don't know how one gets funding for this; that is sort of a secondary issue to me in terms of figuring out what are some models to jump start activity.

DR. BLOWER: What we were talking about were problems that academics have, that we are all trying to work on specific problems. We can't suddenly switch because your funding depends on it.

But one thing that perhaps would be a good idea is to put a team of people together for a month and give them a specific question and a specific mandate. They would spend a week talking about something and then two

weeks actually doing it, and the final week actually writing the manuscripts and getting them out. So you are taking people out of their normal environments but giving them a specific problem.

For example, since I am on infectious diseases, that is immediately what comes to mind, to say with smallpox, and what could happen at different centers. So you might get teams of people, and they end up with a variety of different approaches, or you get a group effort at the end. So that would be a month's worth of work, but you could do that in a relatively short term.

That is what happened with the foot and mouth epidemic in the U.K. People were switched onto that full time, and they came out with it relatively quickly.

DR. LOCKHART: I just want to mention -- I can't really say any specifics right now, but there are a number of things that we are working on that would facilitate the formation of such groups. Watch this space; we may be able to tell you more in a month or two.

DR. CHAYES: Right. I am allowed to say a little bit, since I am not bound. I don't know anything for certain, but it may be that there is an institute that may be funded which is designed for this type of program, which is designed to bring people together to solve a particular

problem. If such an institute were to be funded, I would assume -- and I am asking this of the granting agency that might do such funding -- whether it could be done on a short time scale.

This is not clear. I now that the NSF can sometimes with relatively small amounts of money respond on a small time scale. I think that for these particular issues, it is very important. But the foot and mouth disease, if in England they had to wait for a new funding cycle, a lot of animals would have died. So hopefully there will be a way of responding to this quickly. If these unsubstantiated rumors are true, there will be a venue.

Also, Bamf, which has been funded, has some monies set aside for small research groups. Bamf is also in a very nice area. So if you like to ski or you like the mountains, you might like it even more than Santa Fe. They certainly do have money for -- I think they are also called focused research groups.

I'm sure many of you know, there is a new math institute at Bamf that is a collaboration of PIMS, which is the Pacific Institute of Mathematical Sciences, and MSRI. It has some NSF funding, highly leveraged NSF funding. There are going to be on the order of 40 weekly workshops

there. If somebody has longer term plans, one cold apply for a workshop on one of these issues, but you can also bring together small groups of people to work on a particular problem. That has been funded.

So any issues other than funding issues? I really liked what we heard at the end of the day yesterday, about high dimensional problems. I think that brings together not just statistics. There is a huge element of statistics in that, but I think that a lot of us realize that stat has a huge place in homeland security.

One of the things that I would like to see more discussion of is how other areas of mathematics, especially core mathematics, are necessary to solve some of these problems. I think it is clear that they are necessary. When you get these large amounts of data, you need a statistical analysis, but you also need some geometry, anthropology.

So I would like to hear either from somebody now in the next few minutes, or hopefully in the talks that we will hear today, how areas of mathematics that we generally think of as core mathematics are necessary to solve these problems, working hand in hand with statisticians and with applied people in the particular field.

I guess I am not going to hear that now, which is fine. Is Howard here? Howard should be here soon. Is anybody here from the 9 o'clock session?

PARTICIPANT: Howard is outside.

DR. CHAYES: Howard is outside. I will go get Howard, and then that session will start.

(Brief recess.)

DR. CHAYES: I am very excited about this next session. We are very lucky to have Howard Schmidt from the White House here. He is from the Office of Cyber Security, which is part of the White House.

He was at Microsoft, running the security for Microsoft until he went to the White House a few months ago. It is Microsoft's loss, but hopefully it is cyber security's gain.

DR. SCHMIDT: Thank you very much. It is great to be here this morning.

I think back to the first e-mail I got from Jennifer about this meeting. The title itself is somewhat very ominous as far as mathematics and sciences and homeland security. At first, I thought she was joking with me, in all honesty.

As I started thinking about it, as I got the e-mail that said this really wasn't a joke, I started looking

into it and thinking, what better way to solve some of the issues that we have got that aren't going to be solved with guns, gates and guards and fences and stuff like that. So I thank you for the opportunity to be here.

I am joined by some extremely distinguished folks on the panel and the follow-on review and discussion. I am just going to say their names for right now. After I asked everyone give me some little talking points on your bios and everything, I am probably going to let them do it, because I think I would not do them justice by introducing them my way, so I would like to have them do it themselves.

In that vein, as we go through the session this morning, I would like to start out just framing some of the things we are looking at from the White House perspective in this area. As I have talked to the panelists -- and their presentation -- in the back of your mind, and I'm sure they will point it out specifically, look at some of the correlations between some of the things we are looking to accomplish in creating a national strategy in defending cyberspace, where is where one of our key focuses is, and some of the things that the panelists are going to be talking about.

Going back to my previous comment about thinking this might be a joke, in reality this makes a lot of sense.

Listening to what they are going to say, you will see whether there is so much potential in using the talent to solve some of the key problems we have got.

I'm not sure if it was Dorothy or someone at one point talked about the big encryption debate that was going around. The comment was made, if you think encryption is the answer to security, you understand neither security nor encryption. So when you look at the picture from the things we are trying to solve, it is just as complex as that.

So with that, let me talk about some of the things that the President's Critical Infrastructure Protection Board is looking at as priorities, and then turn it over to my distinguished colleagues here.

First and foremost, one of the things that we find to be in short supply is awareness. As we have gone around the country, we have talked to government leaders, we have talked to industry leaders. If you get outside that small sphere of security and you talk about security, you get the deer in the headlight look, so people start to drool, going, what are you talking about? Why do I care about this?

So there is this component about the awareness and the education we really need to focus on, and build

that piece up. One of the ways we are looking at this right now through the education component is, we have created a scholarship program called Scholarships for Service. The National Science Foundation administers it. I think our biggest customer thus far has been the Department of Defense, where they allocate funds through NSF to scholarships to people in advanced degree programs in information assurance, information security. They do a one-for-one; if we pay for one year of tuition, they come back and do one year of government service, two years and two years, et cetera.

The intent is to build the cadre of expertise that we have internal to the government, because we lose it regularly. Many go back and forth between the private sector.

The discussion also goes, though, if we train these people and they come back and do two years of government service, they are going to be prime candidates to go in the public sector. My answer is, wonderful, because who are the owners and operators of the critical infrastructure that we care about? The private sector. So it is a win-win situation.

We have a couple of years to beef up the government stuff, which we need desperate help on. At the

same time, we have the opportunity for those folks to get some real, live, on the job training, move out into the private sector and then continue to proliferate the wonderful things they have learned.

The other priority is the information sharing part. This is a wonderful forum for that as well. There is this pace of activity that goes on that you see in the newspaper all the time. I read one last night. There was a bunch of computer sites in Korea, in which the ill-intended people are doing things and using those to launch attacks on other systems around the world.

That is a bad thing. But when you try to get details and you try to get some information, it is generally a standoff approach. We are not privy to a lot of the details. We are not privy to a lot of the things that could help us better protect ourselves. So this sharing amongst professionals, and there is no group that does it better than academia, and sharing that information and saying, let's figure out how this is going on, let's figure out the defenses to make it work accordingly.

The other one is the R&D component. There is a true belief, at least in the government circle, and I think it is shared by some of my colleagues, I know when I was in the private sector, many of us talked about it, that there

is some wonderful R&D being done in the buildings where the walls that have no windows and being done in the venue of national security.

There is some really great stuff being done by the researchers in the private sectors to generate things that can be used to bring to market to benefit the public. But there is some space in the middle that we are not sure what that space is. We think there is some really hard-core, thoughtful R&D that needs to be done that is not being funded.

So we have asked the Congress to give us a boatload of money, in the tens of millions of dollars, to fund some key programs. People come to us and say, gee, I think we can do this, and this will help the overall package and we can help fund these things on the front end. So the R&D is extremely important.

I want to touch on another thing that is a priority for us, and that is some pure technology things, the way the Internet was built. That is the domain name servers and border gateway protocols.

If you are not familiar with this aspect of it, the domain name servers are those things, when you type in a name, it is converted to a number, when then identifies your address on the Internet. There are about 14 of them

out there. So if I wanted to disrupt activity in the online world, be it commercial or be it telecommunications, that is where I will go, because I can knock out those fairly handily because they are addressable from the Internet. They are addressable in spaces where they have to be able to have an in-band address to be able to communicate. So consequently, we have some real concerns about that.

I don't think redundancy is the answer. In the border gateway protocols, the language they talk in is insecure. Many times it is done in unclear text. We see in this, particularly going back to the illustration I mentioned about career -- one of the things I cited was being able to create denial of service attacks as a result of it.

Then there is the priority we have about standards and best practices. Many of you -- and Dorothy and I were just talking about this in the lobby, about the old Orange Book that effectively said, here is the standard to which you design things. Then no one can meet the standard, so consequently they start to give exemptions. Then exemptions led to almost total obliteration of the standard and say don't worry about it anymore, because nobody can meet it.

We have got to find a meaningful scientific way to say, we can bring this up. We can raise the standard so we can use the procurement power that we have both within government and outside of government to make sure that the development process meets what we need in the areas of security.

Let me broaden security for just a moment, because I am almost fanatical in some cases about this. I want to use the word trust, because security is only a component of it. I will qualify that right now. You have got the security, you've got the privacy, you've got the availability component. There has been a lot of discussion of late -- this is a little bit of a digression from my notes here, but there has been a lot of discussion that security is going to trump privacy.

I oftentimes get asked, where we are going to level? I don't know. We are still in this aftershock mode after what happened last year. So am I willing to give up a little bit of my privacy for security? I don't know that I will be six months from now, so I don't know what I'll feel. But I think fundamentally, the issue always comes across as an issue of trust. You have to have the security, you have to have the privacy, you have to have

the availability. So we are talking about the standards and best practices that we look to; those all play into it.

The next one is something that is extremely worrisome to me as well as many of my colleagues, and that is digital control systems. Last year, there was an incident where a disgruntled employee left a company in Australia, went back in in an unauthorized manner, broke into the systems and reversed the flow of raw sewage. Instead of going into the raw sewage treatment plant, they went into one of the local parks. It is all because of accessibility to digital control systems.

Look what we are seeing today. We are seeing a lot of these digital control systems being accessible or addressable from the Internet. It makes business sense, but it doesn't make security sense. Not only do we have directly accessible from the Internet, but we are finding some that are saying, no, we don't have any addressable space on the Internet, and you find out that they have digital control systems connected to an internal administrative LAN which is then connected to the Internet on the other side, which translates into, they are addressable from the Internet.

That is very worrisome. It controls the power grid, it controls the water supplies in many instances. It

controls the water flowing over many dams to generate electricity. There is a whole bunch of things that are being controlled by digital control devices right now.

When we talk to some of the people that are involved in the technology designing some of these things -- this is something that maybe you all can collectively help with -- they say, we would like to do more. But what happens is, even if we are looking to do a simple thing like authentication a digital control system, when we are talking nanosecond switching time, there is no way to authenticate something and still do the switching in an appropriate manner. So we need to figure out a scientific way to be able to do the authentication without losing the gating factor, that we have to do switching of these things.

It is a complex problem, and it is only going to be solved by some of the activity that you all are doing.

DR. BORGS: I don't understand that. If I can go in from the Internet to reverse the flow of the sewage system, this is not necessarily to make it --

PARTICIPANT: An example. There are other examples.

DR. SCHMIDT: Yes, that was a very broad example from something that was very public in the news.

DR. BORGS: But where you are worried about this outside the controls, that should not --

DR. CHAYES: The electricity, for example.

DR. SCHMIDT: For example, last year there was a storm in the Pacific Northwest. A tree blew down in Oregon and the lights went out in Tucson, Arizona, 1500 miles away. It is all because of the switching controls.

Many of the switching controls, for example, in the power grid are based on very, very slight fluctuations in electrical usage that would cause the entire system to switch over to another grid to provide power. Those are the sort of instantaneous controls that need to be switched, but there has also got to be the ability to do them on an authenticated mechanism. That is what I am referring to.

Lastly, and by all means no less importantly, would be the issues around securing the future systems. I love wireless. I don't know how many of you use it in here, but I couldn't live without it. I did it when I was at Microsoft, I use it at my home now, and I love it, but it is not the most secure environment right now because it hasn't been designed as such. We have grave concerns about it.

Many agencies are talking about outlawing the use of it. So consequently, there are issues around the authentication piece, about the encryption piece, and other future generation systems that we are looking at.

So with that, I took this opportunity to talk about what we are concerned about in framing a broad perspective, before I turn it over to my distinguished colleagues to talk about their concerns.

Thank you very much. I'd like to start out by asking Dorothy Denning to step up and give us her thoughts on it. Thank you.

Howard Schmidt
"Introduction by Session Chair"

Transcript of Presentation
Summary of Presentation
Video Presentation

Howard A. Schmidt was appointed by President George W. Bush as a Special Assistant to the President and the Vice Chair of the President's Critical Infrastructure Protection Board in December 2001. The Cyber Security Board supports Dr. Condoleezza Rice, National Security Advisor and Tom Ridge, Secretary of Homeland Security. The Cyber Security Board focuses on building a specialized group of senior government and private sector leaders to focus on cyber security issues and coordination of security related incidents.

Previously, Mr. Schmidt was chief security officer for Microsoft Corp., overseeing the Security Strategies Group, which was responsible for ensuring the development of a trusted computing environment via auditing, policy, best practices and incubation of security products and practices. Before working at Microsoft, Mr. Schmidt was a supervisory special agent and director of the Air Force Office of Special Investigations (AFOSI), Computer Forensic Lab and Computer Crime and Information Warfare Division. While there, he established the first dedicated computer forensic lab into the government. AFOSI specialized in investigating intrusions into government and military systems by unauthorized persons in counterintelligence organizations and criminals. Before working at AFOSI, Mr. Schmidt was with the FBI at the National Drug Intelligence Center, where he headed the Computer Exploitation Team. He is recognized as one of the pioneers in the field of computer forensics and computer evidence collection. Before working at the FBI, Mr. Schmidt was a city police officer from 1983 to 1994 for the police department in Chandler, Arizona.

Mr. Schmidt served with the U.S. Air Force in various roles from 1967 to 1983, both on active duty and in the civil service. He has served in the military reserves since 1989 and currently serves as a credentialed special agent in the U.S. Army Reserves, Criminal Investigation Division. He has testified as an expert witness in federal and military courts in the areas of computer crime, computer forensics, and Internet activity.

Mr. Schmidt had also served as the international president of the Information Systems Security Association (ISSA) and the Information Technology Information Sharing and Analysis Center (IT-ISAC). He is a former executive board member of the International Organization of Computer Evidence and served as the cochair of the Federal Computer Investigations Committee. He is a member of the American Academy of Forensic Scientists. He served as an advisory board member for the Technical Research Institute of the National White Collar Crime Center and is a distinguished special lecturer at the University of New Haven in Connecticut, teaching a graduate certificate course in forensic computing. He served as an augmented member of the President's Committee of Advisors on Science and Technology in the formation of an Institute for Information Infrastructure Protection. Mr. Schmidt was one of 29 industry leaders called to the White House to meet with President Clinton on cybersecurity. He has testified before a joint committee on computer security and has been instrumental in the creation of public and private partnerships and information-sharing initiatives.

Mr. Schmidt holds a bachelor's degree in business administration and a master's degree in organizational management.

DR. SCHMIDT: Thank you very much. It is great to be here this morning.

I think back to the first e-mail I got from Jennifer about this meeting. The title itself is somewhat very ominous as far as mathematics and sciences and homeland security. At first, I thought she was joking with me, in all honesty.

As I started thinking about it, as I got the e-mail that said this really wasn't a joke, I started looking into it and thinking, what better way to solve some of the issues that we have got that aren't going to be solved with guns, gates and guards and fences and stuff like that. So I thank you for the opportunity to be here.

I am joined by some extremely distinguished folks on the panel and the follow-on review and discussion. I am just going to say their names for right now. After I asked everyone give me some little talking points on your bios and everything, I am probably going to let them do it, because I think I would not do them justice by introducing them my way, so I would like to have them do it themselves.

In that vein, as we go through the session this morning, I would like to start out just framing some of the things we are looking at from the White House perspective in this area. As I have talked to the panelists -- and

their presentation -- in the back of your mind, and I'm sure they will point it out specifically, look at some of the correlations between some of the things we are looking to accomplish in creating a national strategy in defending cyberspace, where is where one of our key focuses is, and some of the things that the panelists are going to be talking about.

Going back to my previous comment about thinking this might be a joke, in reality this makes a lot of sense. Listening to what they are going to say, you will see whether there is so much potential in using the talent to solve some of the key problems we have got.

I'm not sure if it was Dorothy or someone at one point talked about the big encryption debate that was going around. The comment was made, if you think encryption is the answer to security, you understand neither security nor encryption. So when you look at the picture from the things we are trying to solve, it is just as complex as that.

So with that, let me talk about some of the things that the President's Critical Infrastructure Protection Board is looking at as priorities, and then turn it over to my distinguished colleagues here.

First and foremost, one of the things that we find to be in short supply is awareness. As we have gone around the country, we have talked to government leaders, we have talked to industry leaders. If you get outside that small sphere of security and you talk about security, you get the deer in the headlight look, so people start to drool, going, what are you talking about? Why do I care about this?

So there is this component about the awareness and the education we really need to focus on, and build that piece up. One of the ways we are looking at this right now through the education component is, we have created a scholarship program called Scholarships for Service. The National Science Foundation administers it. I think our biggest customer thus far has been the Department of Defense, where they allocate funds through NSF to scholarships to people in advanced degree programs in information assurance, information security. They do a one-for-one; if we pay for one year of tuition, they come back and do one year of government service, two years and two years, et cetera.

The intent is to build the cadre of expertise that we have internal to the government, because we lose it

regularly. Many go back and forth between the private sector.

The discussion also goes, though, if we train these people and they come back and do two years of government service, they are going to be prime candidates to go in the public sector. My answer is, wonderful, because who are the owners and operators of the critical infrastructure that we care about? The private sector. So it is a win-win situation.

We have a couple of years to beef up the government stuff, which we need desperate help on. At the same time, we have the opportunity for those folks to get some real, live, on the job training, move out into the private sector and then continue to proliferate the wonderful things they have learned.

The other priority is the information sharing part. This is a wonderful forum for that as well. There is this pace of activity that goes on that you see in the newspaper all the time. I read one last night. There was a bunch of computer sites in Korea, in which the ill-intended people are doing things and using those to launch attacks on other systems around the world.

That is a bad thing. But when you try to get details and you try to get some information, it is

generally a standoff approach. We are not privy to a lot of the details. We are not privy to a lot of the things that could help us better protect ourselves. So this sharing amongst professionals, and there is no group that does it better than academia, and sharing that information and saying, let's figure out how this is going on, let's figure out the defenses to make it work accordingly.

The other one is the R&D component. There is a true belief, at least in the government circle, and I think it is shared by some of my colleagues, I know when I was in the private sector, many of us talked about it, that there is some wonderful R&D being done in the buildings where the walls that have no windows and being done in the venue of national security.

There is some really great stuff being done by the researchers in the private sectors to generate things that can be used to bring to market to benefit the public. But there is some space in the middle that we are not sure what that space is. We think there is some really hard-core, thoughtful R&D that needs to be done that is not being funded.

So we have asked the Congress to give us a boatload of money, in the tens of millions of dollars, to fund some key programs. People come to us and say, gee, I

think we can do this, and this will help the overall package and we can help fund these things on the front end. So the R&D is extremely important.

I want to touch on another thing that is a priority for us, and that is some pure technology things, the way the Internet was built. That is the domain name servers and border gateway protocols.

If you are not familiar with this aspect of it, the domain name servers are those things, when you type in a name, it is converted to a number, when then identifies your address on the Internet. There are about 14 of them out there. So if I wanted to disrupt activity in the online world, be it commercial or be it telecommunications, that is where I will go, because I can knock out those fairly handily because they are addressable from the Internet. They are addressable in spaces where they have to be able to have an in-band address to be able to communicate. So consequently, we have some real concerns about that.

I don't think redundancy is the answer. In the border gateway protocols, the language they talk in is insecure. Many times it is done in unclear text. We see in this, particularly going back to the illustration I mentioned about career -- one of the things I cited was

being able to create denial of service attacks as a result of it.

Then there is the priority we have about standards and best practices. Many of you -- and Dorothy and I were just talking about this in the lobby, about the old Orange Book that effectively said, here is the standard to which you design things. Then no one can meet the standard, so consequently they start to give exemptions. Then exemptions led to almost total obliteration of the standard and say don't worry about it anymore, because nobody can meet it.

We have got to find a meaningful scientific way to say, we can bring this up. We can raise the standard so we can use the procurement power that we have both within government and outside of government to make sure that the development process meets what we need in the areas of security.

Let me broaden security for just a moment, because I am almost fanatical in some cases about this. I want to use the word trust, because security is only a component of it. I will qualify that right now. You have got the security, you've got the privacy, you've got the availability component. There has been a lot of discussion of late -- this is a little bit of a digression from my

notes here, but there has been a lot of discussion that security is going to trump privacy.

I oftentimes get asked, where we are going to level? I don't know. We are still in this aftershock mode after what happened last year. So am I willing to give up a little bit of my privacy for security? I don't know that I will be six months from now, so I don't know what I'll feel. But I think fundamentally, the issue always comes across as an issue of trust. You have to have the security, you have to have the privacy, you have to have the availability. So we are talking about the standards and best practices that we look to; those all play into it.

The next one is something that is extremely worrisome to me as well as many of my colleagues, and that is digital control systems. Last year, there was an incident where a disgruntled employee left a company in Australia, went back in in an unauthorized manner, broke into the systems and reversed the flow of raw sewage. Instead of going into the raw sewage treatment plant, they went into one of the local parks. It is all because of accessibility to digital control systems.

Look what we are seeing today. We are seeing a lot of these digital control systems being accessible or addressable from the Internet. It makes business sense,

but it doesn't make security sense. Not only do we have directly accessible from the Internet, but we are finding some that are saying, no, we don't have any addressable space on the Internet, and you find out that they have digital control systems connected to an internal administrative LAN which is then connected to the Internet on the other side, which translates into, they are addressable from the Internet.

That is very worrisome. It controls the power grid, it controls the water supplies in many instances. It controls the water flowing over many dams to generate electricity. There is a whole bunch of things that are being controlled by digital control devices right now.

When we talk to some of the people that are involved in the technology designing some of these things -- this is something that maybe you all can collectively help with -- they say, we would like to do more. But what happens is, even if we are looking to do a simple thing like authentication a digital control system, when we are talking nanosecond switching time, there is no way to authenticate something and still do the switching in an appropriate manner. So we need to figure out a scientific way to be able to do the authentication without losing the

gating factor, that we have to do switching of these things.

It is a complex problem, and it is only going to be solved by some of the activity that you all are doing.

DR. BORGS: I don't understand that. If I can go in from the Internet to reverse the flow of the sewage system, this is not necessarily to make it --

PARTICIPANT: An example. There are other examples.

DR. SCHMIDT: Yes, that was a very broad example from something that was very public in the news.

DR. BORGS: But where you are worried about this outside the controls, that should not --

DR. CHAYES: The electricity, for example.

DR. SCHMIDT: For example, last year there was a storm in the Pacific Northwest. A tree blew down in Oregon and the lights went out in Tucson, Arizona, 1500 miles away. It is all because of the switching controls.

Many of the switching controls, for example, in the power grid are based on very, very slight fluctuations in electrical usage that would cause the entire system to switch over to another grid to provide power. Those are the sort of instantaneous controls that need to be switched, but there has also got to be the ability to do

them on an authenticated mechanism. That is what I am referring to.

Lastly, and by all means no less importantly, would be the issues around securing the future systems. I love wireless. I don't know how many of you use it in here, but I couldn't live without it. I did it when I was at Microsoft, I use it at my home now, and I love it, but it is not the most secure environment right now because it hasn't been designed as such. We have grave concerns about it.

Many agencies are talking about outlawing the use of it. So consequently, there are issues around the authentication piece, about the encryption piece, and other future generation systems that we are looking at.

So with that, I took this opportunity to talk about what we are concerned about in framing a broad perspective, before I turn it over to my distinguished colleagues to talk about their concerns.

Thank you very much. I'd like to start out by asking Dorothy Denning to step up and give us her thoughts on it. Thank you.

Introduction by Session Chair

Howard Schmidt

Mr. Schmidt discussed some of the priorities of the President's Critical Infrastructure Protection Board.

One priority is to build awareness of various homeland security issues and build the cadre of expertise that we have internal to the government. One way the government is addressing this problem is through a program called Scholarships for Service, which is administered by the National Science Foundation and funded in large part by the Department of Defense. Scholarships go to people in advanced-degree programs in information security, who then do a one-for-one payback—if government has paid for *N* years of tuition, recipients return *N* years of government service.

Another priority is information-sharing. We are not privy to a lot of the things that could help us better protect ourselves. Government needs to be included in the kind of information-sharing among professionals that is done so well in academia.

While wonderful R&D is being done by the government directly in the area of national security and by researchers in the private sector, there is some hard-core, thoughtful R&D that needs to be done but that is not being funded, because it falls between the traditional interests of national security and those of commercial entities. Therefore, the administration has asked Congress for funds to support competitive R&D programs that fill this gap.

Other priorities include improving Internet security, and reviewing security standards for digital control systems, such as power grids or water supplies, some of which are currently accessible from the Internet. Having such systems accessible from the Internet makes business sense, but it doesn't make security sense.

Dorothy Denning
"A Security Challenge: Return on Security Investment"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Dorothy Denning is a professor in the Department of Defense Analysis at the Naval Postgraduate School. At the time of the workshop, she was the Patricia and Patrick Callahan Family Professor of Computer Science and Director of the Georgetown Institute of Information Assurance at Georgetown University. Dr. Denning has published 120 articles and four books, her most recent being *Information Warfare and Security*. She is an Association for Computing Machinery (ACM) fellow and recipient of several awards, including the Augusta Ada Lovelace Award and the National Computer Systems Security Award.

DR. DENNING: Thanks. I was wondering if the slides would magically appear, and they did. So I can relax now.

DR. SCHMIDT: Vince Serf and I were at a meeting recently, and Vince in his opening comment was having a similar problem and said, power corrupts, Power Point corrupts absolutely.

DR. DENNING: What I thought I would do is just start out with a little bit of data about the state of security currently.

This is the data that is reported to the Computer Emergency Response Team at Carnegie-Mellon University. They receive incident reports from people who have had some kind of security incident, generally through the Internet.

As you can see in the last couple of years, the number of incidents reported has gone up rather dramatically. On top of that, each incident that is reported to CERT can actually represent an attack against thousands or even hundreds of thousands of host computers. So the Code Red worm, for example, which hit hundreds of thousands of computers is one data point on that chart. It is also true that not all data is reported to CERT. So it doesn't represent all attacks.

This is some more data. This is what was collected by MI 2G in London. This is just web defacements, and again, in the last few years a dramatic increase.

This is one that was collected by Riptech. They are a local company which is a managed security firm that monitors and networks their clients. This represents data collected over a six-month period of 300 of their clients.

What it shows is the average number of network attacks against each of their clients over that period. One of the things that is very interesting here is that each company on average is experiencing a greater intensity of attacks, or more attacks. So the increase in attacks on the Internet cannot just be attributed to the fact that maybe there are more computers on the Internet and he Internet is growing. The attacks on a particular company also are increasing.

DR. CHAYES: Is that a seasonal thing?

DR. DENNING: Who knows? I don't know. We will have to look at the next six months that they get. I hope they will continue this, particularly with those 300 clients, because it will be interesting to see what happens on a long term basis.

DR. MC CURLEY: You might also look at the release cycle of Outlook, to see if they are correlated.

DR. DENNING: One of the things that is interesting is that contrary to a lot of statements that were made, attacks did not go down after September 11.

DR. CHAYES: What do you mean? Some people don't know what you mean by an individual attack.

DR. DENNING: What is an attack on here?

DR. CHAYES: Yes.

DR. DENNING: It could be anything from people just scanning their networks to actually getting in the networks, denial of service attacks, anything that involved either penetration or an attempt to penetrate a computer network.

PARTICIPANT: (Comments off mike.)

DR. DENNING: Ping? I don't remember exactly what --

DR. SCHMIDT: Some people considering pinging as a base -- for example, if you throw a thousand packets against someone, they consider that an attack.

DR. DENNING: That would be definitely, in that case.

DR. BICKEL: When you break it down, do you get consistent trends? So if you look at the attacks that potentially can cause the most damage, --

DR. DENNING: They did that, too. They broke it down into different categories of attacks. I don't think I have the slide in here for that. I just put a couple of slides in, just to give you a general idea.

Their whole report is on the web. If you go to riptech.com, you can get a copy of their report for free.

Finally, also more vulnerabilities are being reported in products. It is not just Microsoft, it is Linux and Solaris and everything else. But the vulnerability reports have gone up. When you just think in terms of the raw numbers, over 2,000 last year, 2500, I guess it was, last year, when you think of the impact of trying to deal with all those, if you are a network administrator, that is pretty horrendous. You are looking at several vulnerabilities per day with the patches that you might have to respond to. Responding to patches is not all that straightforward, because you can start patching things up, and it changes your network, and something else might break. So it is a non-trivial exercise. Yes?

DR. CHAYES: Do you think it is due to the complexity of the programs, they actually are more vulnerable?

DR. DENNING: Yes.

DR. CHAYES: Or do you think that people are just seeing more of the vulnerabilities?

DR. DENNING: I don't think I have the slide in here, either, but if you just look at the software itself, it is getting more complex. It is getting bigger. Like, Windows XB is 40 million lines of code, whereas if you go back a few generations, there is less millions of lines of code in it. Then there are more interactions and so on, so it is a difficult problem.

So what can we conclude from all that? This is just a statement of the obvious. Either what we have installed in the way security is optimal -- and I want to throw that out as a possibility, even though no one believes that, but it could be that given what the products are that are out there, and what we know about security, right now today, we are doing the best we can, and we need to learn to live with the risk.

The other alternative is that it is really not optimal, that we could be doing better. We could be reducing the risk, reducing losses from attacks at a cost

that would be lower, so we wouldn't have to spend ten million dollars in order to prevent one million dollars worth of losses.

What I want to do is look at the question of return on security investment. A lot of the security that has been developed in the past is based on what we think works, intuition, wisdom, things like this. These are not bad things, but a lot of it is not really based on any empirical data.

The goal. Risk security has to be cost effective. It has got to be that the cost that you spend on the security has got to be less than the expected losses that you are protecting against, because you don't have an infinite budget to spend on security, and so you have to put it in places where you are going to reduce risk in a cost effective way.

DR. CHAYES: Dorothy, you are talking about the kinds of things that Howard was talking about, like the protection of our water supply and that kind of thing. So actual personal injury resulting. How hard is it to quantify?

DR. DENNING: Absolutely hard to measure. Let me just keep going.

Return on security investment would measure in some way the bang that you got for your buck, so to speak. The comment like you just mentioned is that it is very difficult to measure this, because a lot of it is intangibles. Not only that, you have also got dependencies, so that a loss that you might suffer as your company may also translate into losses to other people who are depending on your business or whatever it is that you offer. So when you go down, they go as well, so there can be a lot of impact.

The next slide actually is the same example is the same example that Howard used, so you can get a little more data there. One of the, I thought, very interesting aspects of that case was that this guy tried 46 times apparently to do this. It wasn't until the 46th try that he succeeded. So something was very wrong in whatever security they were deploying there that they didn't pick this up and abort it before it happened.

There is some good news there, too. I think what it does suggest is, maybe these attacks aren't all that easy. Here we had an insider; he had a copy of the software, he had helped develop the software and everything else. So while he did pull it off, it wasn't something

that just anybody anywhere in the world could have done remotely.

One other comment back on that. This particular attack actually led to loss of life. It wasn't human life, but there was wildlife that was killed by that, and it had a pretty severe impact on the environment.

To me, the security challenge that has to be met is developing security better than we are doing now, but it has to be the case that it is also cost effective. What that means is that as we go forward, we have to base it on sound mathematical models so that we can do good analysis. Also, we have to base it on empirical data. A lot of the work is not based on empirical data.

I can tell you, my colleagues here, one of the things that I have admired about their work is that it has in fact been -- I remember, the first paper I ever read about David Wagner's was the one where he showed that the 40-bit key that Netscape was using was -- he could break it in 30 seconds or something phenomenal. He had the data to show that he could do it, as well.

That is the kind of thing we need, so that we actually can support scientifically the best practices and the standards and so on that we develop for security. When

you take some of the existing standards, they are not based on any empirical data.

Let me just give you a couple of examples of the kind of thing that I have in mind. One of my big beefs is the systems that make me change my password every 30 days or 60 days or something. I hate these things. I would like to see somebody do a study and tell me that if you do that, you are lowering the risk by such and such and that this is a cost effective thing to do. I have never seen any studies to support that, or any of the other quote best practices that are stated for security.

There are a lot of questions about what we can expect from a return on security investment if we do certain things. I would like to see research that goes more in that direction.

Let me give you some very specific examples. This is an example that was done by researchers in the lab at Stake, which is a security consulting company. A lot of people have said, you should harden your server, and you should shut down the ports you are not using and get rid of the services you are not using and so on.

So what they did was, they went out and measured to see what is the effect of doing all that. They looked at several different kinds of networks -- this is just the

data from one of them -- and what they showed was that you actually get better throughput on your network. So here is a great example of return on security investment. Spend a little bit of time hardening that network, and now your network performs better. So this is the kind of data I had in mind.

Another example. We know that the software is vulnerable. What would be the impact if more effort is put into developing software, so that there are fewer holes in it from the beginning. We know Microsoft now is committed to doing that, trying to develop their software with less bugs in it.

Here is some data that shows that in fact, the impact in terms of the cost of responding to a flaw that is found in the software is actually rather dramatic. This is just accounting for the effort required to fix the software on the part of the vendor. This is not even taking into effect the impact on the people who are using the software, and what losses they will experience as a result of the flaws in the software.

They have this cost multiplier factor. If you find and fix the flaw in the design stage before you start coding, it costs unit one. When you get to the implementation stage, it is 6.5, testing 15, and when it is

all the way up to the maintenance, you are up potentially to a cost increase of a hundred fold.

Then this slide here just shows again the return on security investment if you can remediate them at different stages. So if you remediate at the design stage, you have a 21 percent increase, and so on.

If you want to get the details of these, these were both published in articles that were published in the Secure Business Quarterly. They are on the web. If you type in Secure Business Quarterly into Google or provide, you will probably come up with a website.

Then finally, this is on a totally different topic. Since this is math sciences, I started thinking about what kinds of mathematics have been used in security over the years, and came up with a rather lengthy list. But what gave me despair about all this is that we don't teach any of this, except for probability and statistics, in most computer science curriculum, as something that anybody would get.

If you take some artificial intelligence courses, you will get some of the stuff related to that, like maybe neural nets and pattern recognition and stuff, but generally, the math itself is not something that most of the students would pick up.

I don't know if there is something that needs to be done there, or if it is just appropriate to teach the math and the courses where you drop in IT cryptography and you obviously need number theory. So I teach the number theory in the course, because none of the students otherwise would have picked up the number theory from their basic curriculum.

The other interesting thing is, when I started out working in security, which was around 1972, 30 years ago, none of this math was used in security at all, with the exception of probability and statistics and number theory starting to show up in cryptography. So we have in many respects come a long way in applying math to the security problems.

I think that is all I've got, so thank you.

A *Security Challenge: Return on Security Investment*

Dorothy Denning

Our goal is to develop better security than we have now, in a cost-effective way. One current obstacle is that some existing practices and standards are not based on scientific data. For example, many systems require users to change their password every 30 or 60 days, but there is no scientific study supporting this as a best security practice. In addition, there is a lack of formal mathematics in computer science curricula, with the exception of probability and statistics.

Security should be based on empirical data and sound mathematical models as opposed to intuition. We need to scientifically support the best practices and standards. There are two examples of improved return on security investment based on empirical data:

- Studies show that by hardening a server, shutting down ports, and removing unused services, not only does security cost-effectiveness improve but the network's throughput increases as well.
- Studies show that developing code with fewer bugs from the beginning is 100 times as cost-effective as fixing bugs during the maintenance stage.

Kevin McCurley
Session on Communications and Computer Security

Dr. Kevin McCurley is a computer scientist in the Theory Group at IBM Almaden Research Center. During his career he has worked in several different fields, including number theory, cryptography, information retrieval, and Web data mining.

Dr. McCurley received a Ph.D. in mathematics from the University of Illinois in 1981, under the direction of Paul T. Bateman.

David Wagner
"A Few Open Problems in Computer Security"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

David Wagner is a professor of computer science at the University of California, Berkeley. His research interests include computer security, especially the security of large-scale systems and networks, applications of static and dynamic program analysis to computer security, theory of cryptography, design and analysis of symmetric-key cryptosystems and operating systems, and computer science theory. He is currently working on software security, wireless security, sensor network security, cryptography, and other topics.

DR. SCHMIDT: While we are doing the mike switch, I'll just take this opportunity to make a couple of quick comments about one of the things that Dorothy had said when we talk about the analysis of the activity in the research we are seeing.

One of the challenges that I think we need some help on is looking at developing a methodology to find out where the noise level is. You were asking about pinging. Pinging is basically as far as we are concerned, because there is not a real structure to attack. It can be problematic in a larger sense, but right now we see a lot of noise out there, but not a real good ability to differentiate between what we have to be concerned about and what we have to react to.

David, did I give you enough time?

DR. WAGNER: While we are bringing the slides up, I am Dave Wagner, and I am from UC-Berkeley.

I have to apologize that I wasn't here yesterday, so for my benefit I wonder if I can take a poll just to get a sense for what the audience is like. Here is your chance to stretch. How many of you would identify yourselves as mathematicians, defined however you would like to define it? Any computer scientists?

PARTICIPANT: I think some people are putting their hands up twice.

PARTICIPANT: I'd like to speak up for the unrepresented minority here. I am a statistician.

DR. WAGNER: Okay, thanks. I wanted to show you a couple of problems that I had hoped might prove for some fruitful collaboration between the computer security folks and the mathematicians. So I think it is great to see a sprinkling here from both camps. I am not going to be nearly as prolific as Kevin McCurley was. I have only two sets of problems.

Let me start off with some general remarks. I am going to say something later in this talk about cryptography, but I would like to remind you that computer security is about a lot more than cryptography.

Here is a way to think about the security problem. There are two computers. They want to talk to each other, and yet they don't trust each other. They want to engage in some collaborative computation, possibly in the presence of an adversary, where they may not even have aligned interests or aligned incentives.

There are two problems here for security people to solve. First, the communication channel between the two in general may not be secure, as is often the case with the

Internet. The second challenge is the end points may not be secure.

If you want to think at a very high level, we can think of two techniques in computer security to deal with these problems. To deal with the insecure channel problem we often use cryptography and to deal with the insecure end point problem, we often use other techniques. So let me just warn you that cryptography can't solve both problems. Cryptography isn't the silver bullet that can solve all your problems. If I have got an authenticated connection to Joe Hacker, it doesn't help much, the fact that that connection is encrypted, if he is sending me the latest virus.

I'm going to move on to the first of two problem areas that I thought might be interesting to look at. One of the really nice things about this workshop setting is, I thought I could talk about problems and questions where I don't know the answers, or I don't know how to solve them. So I am hoping this will be a discussion, not just a one-way presentation.

Since the topic of this workshop is homeland defense, I thought that critical infrastructure protection was an obvious problem area to talk about. If one looks at the infrastructures in the nation, one finds a number of

critical infrastructures, from the utilities like power, water, oil, gas, to the telecomm industries, your public switch telephone network, Internet and other communications mechanisms, and the financial sector and others.

We have become very dependent on these systems. These systems are evolving legacy systems that weren't necessarily designed for security when they were first deployed, and continue to change on the fly. They are becoming increasingly reliant on IT, information technology, which I find troubling, because the security of information technology is not at the level that I would want to put much reliance in.

So I think the defining characteristic of the critical infrastructure problem is we have got systems here that are very large scale and are tightly interdependent. If we look at the banking system, the correct functioning of the banking system depends first of all on the correct functioning of many banks, but also on a number of other critical infrastructures. If the power is not on, the banks and the ATMs are probably not going to be working. If the phones are out, then your ATM probably isn't going to work. So we have got tightly interdependent systems, and that means that security is a challenge.

So let's move on. I wanted to give an example in a little bit of detail, which is one of the critical infrastructures, the power grid. Let me warn you up front that I am definitely not an expert on the power grid. I wanted to share with you some of my recent explorations to learn more, and just be warned to take this with a big grain of salt. This is not my specialty.

A quick introduction to the power grid. There are a number of key elements. First of all, there are the loads. That is my laptop, plugged into the wall over there, and your refrigerator, and so on and so forth, and the loads on the power system can vary from time to time, both predictably, seasonally, and unpredictably.

There is what we might think of as the power grid -- the transmission system can be broken down into two parts, the local distribution mechanisms from your local area substation to your home, and there is the long distance transmission lines that span the country. Then of course, there are the generators which actually supply the power to the transmission lines, which make their way to the distribution system and finally to your home.

On top of that, it is important to remember that the power grid is not a passive system. It is a system with active feedback loops. There are a number of control

centers. In fact, the picture here on the lower left is showing the control centers that try to coordinate the power grid. There is a higher level financial settling and bidding and coordination between power providers.

So there are a number of issues here. All the layers of the power grid, for instance, the eligible receiver operation run by the NSA in 1997, I believe, is one of the more troubling exercises I am aware of. The NSA put together a number of computer security experts, and their charter was to take off the shelf software they could download off the Net, nothing but what they could download off the Net, without breaking any laws, in a simulated attack try to see what damage could be done in the United States.

A number of critical infrastructures were ruled to be at severe risk. The referees ruled that if they had wanted, they could have taken down part of the power grids, substantial portions of the emergency 911 system and other, so that is quite troubling.

PARTICIPANT: (Comments off mike.)

DR. WAGNER: Unknown.

DR. DENNING: They did not actually attack the power grid. This was a paper exercise.

PARTICIPANT: (Comments off mike.)

DR. WAGNER: I don't know. Details are sparse.

PARTICIPANT: (Comments off mike.)

DR. DENNING: No, they did assume in their paper exercise that anybody with Internet access and just using tools that you could download from the Internet, so it didn't require any special access.

DR. SCHMIDT: When we did that, it was just basically an escalation process, where we had a system administrator of a relatively insignificant system, which was connected to another system, connected to another system. Through escalation of privileges, it gives you the ability to reach into this backbone type of environment.

DR. WAGNER: So our own vulnerability assessments have reported systems connected by modems or connected to the Internet, their administrators didn't know about, and that is an obvious risk. So I frankly don't know these kinds of details.

One of the really problematic features of the power grid that makes life hard for securing the power grid is this feature of cascading filters. It is a tightly interdependent system, and a single failure can cascade and have a very significant effect, too, to affect large parts of the system.

Just a couple of examples. In 1989, solar storms caused a failure in Quebec. That is in yellow, the picture on the left. This failure in Quebec affected the power grid throughout all of the United States. You can see these little circles are protective breakers that were tripped by this, that if not for the protective equipment would have caused damage. In fact, some equipment was also damaged on the East Coast. So this map illustrates that local failures can cause global effects.

Maybe an even better example is the 1996 blackouts in July and August, particularly the ones in August. There were two line failures that occurred nearby each other. I believe a tree fell on the line or something like this, and this led to oscillations in the power grid that caused eventually blackouts in 13 states and affected millions of folks on the West Coast.

In fact, I am told, if I got my sources right, that it is not even completely understood why this caused the failure. So that is quite troubling. That means that our understanding of the power grid is really not at the level we would like it to be.

This is caused by the interdependencies in the power grid. If anything, all the evidence suggests that things are going to get better, not worse. For instance,

the capacity margin, the redundancy built into the system, the over capacity in generation is built into the system. It was about a 25 percent in 1980 and now is about 12 percent, and we can expect this is going to be going down, because projected demand over the next decade is growing at 20 percent, while the projected building of new generation capacity is expected to only grow about three percent. So these are becoming very, very slim margins, and the combination of these trends and deregulation is something to be concerned about.

I tried to put together some very simplistic examples that I thought would help give some intuition for how cascading failures can occur. Here is a very simple transmission system with five nodes and seven lines. The idea is that each line may have a different capacity. The lines on the left have a 100,000 volt capacity; anything above that and they fry. The orange lines have a 75,000 volt capacity, and the blue ones have a 25,000 volt capacity -- completely made-up numbers, just to illustrate an example.

We can imagine that the resistance in all the lines is the same, so that the amount of current flowing through the line is proportional to the voltage difference across the line.

Here is a steady state flow through the network. We are trying to transmit 160,000 volts equivalent in current from the left-most node to the right-most node. You can see that due to symmetry, it is going to split evenly up and down. If you work out V equals IR, that tells you you will get the following flows.

You can see that the flow is splitting evenly between up and down. In this node it is splitting three to one for the large line. You can check that all these flows are well within capacity. There is plenty of slack here.

So we have got about 20,000 volts of capacity over here, and spare capacity may be another 15 dozen volts of spare capacity there. So there is certainly 10,000 or 15,000 or 20,000 volts of spare capacity in the system.

Now let's imagine we have a line failure. A tree falls, you have lost this line right here, so this line is going to go away, it fried, it went to zero kilovolts. The load is now redistributed through the rest of the network, again following Ohm's law. We see that the load is now distributed so that more of the load flows down along the bottom half of the network. Here we have some lines in parallel.

Now you expect there is no problem. We had plenty of slack laying around, plenty of spare capacity,

and we lost a line that was carrying 20,000 volts, but we had 20,000 volts of spare capacity laying around, in fact more than that, so there should be no problem. If this were a flow problem, all would be well. You could arrange the flows to send this much capacity through, even though you lost the line, because there is plenty of redundancy. But this isn't a flow problem. It is constrained to follow the equipotential Ohm's law and so on and so forth. You find if you take this solution, you find the following flow numbers along the links.

You notice that most of these links are well within capacity except that this poor guy right here has capacity for 25,000 volts, and now has gotten 29,000. So that is what happens. He is overloaded, he is going to fry. We lost that link, so that link is now not going to be able to carry any more current, so we have a new solution. The current redistributes itself through the network, nice and symmetric.

Now we have got a problem. No problem in this link, we are well within capacity, but now this link has got too much current flowing through it. So we have got an overload there. Once again, those lines are now overloaded, and we have killed our network.

So even though a priori it looked like we had plenty of spare capacity, nonetheless losing one line causes the network to become completely disconnected. So this is a very simplistic example, only showing you with capacities. In practice, I believe there are a number of effects you would want to model, not just the capacity. Moreover, this is only the static behavior of the network. In practice, there are also dynamic behaviors as well. When a line is lost the system oscillates a while before it hits the steady state, and those oscillations can be quite important, and I think are poorly understood.

So I hope this gives some idea of what we might like to understand about a power network.

Let me pose a few research questions for you. Stepping up the level now, thinking not just about power grids, but about principles that might apply to infrastructure protection in general.

We have got a large scale system here. It is very interdependent. There are a number of natural questions to ask here. One natural question is, pick a system, can we build a useful predictive model? Can we build a mathematical model that allows us to analytically express some of this properties, and in particular, can we measure security against malicious attack?

I showed here a system that failed if a tree fell on one line, then the whole system failed. If we are worried about malicious attack on our power system, then we can imagine that a plausible threat model might be that an attacker who is able to figure out where that critical line is will be able to fry a line or fry three carefully chosen lines.

So if you have a model, an algorhithmic question might be, is there an efficient way to detect whether there exists any three lines in the power system whose failure will cause you to become disconnected.

Right now, I believe what the industry uses is the exhaustive search algorithm. They try all possible thoughts, and their focus so far has been on not so much security against malicious attack as robustness against random failures. So the premise is that they only need to worry about a single failure and they exhaustively look at all possible single failures to see whether any of them could cause a cascading failure. So I think there is room for a lot of progress here.

More generally, you might ask, rather than looking at specific systems, you might raise the level of abstraction just a bit more and look at class of systems and ask about the structural properties of such systems.

We can imagine that the attackers' capabilities give us some cost metric or some distance metric on systems. The attacker can perturb the system a little it, but not too much. Then we can ask, what are the parameters of these drafts that determine their robustness properties. Can we design systems that are inherently self stabilizing, inherently robust. Moreover, are there any local control rules that insure global stability with only local evolution.

I mentioned that the power grid is an adaptive feedback loop. In fact, it is crucial right now that the industry is relying on the ability to have global knowledge of the whole grid. That doesn't scale, and it is going to run into problems with deregulation. So there are a number of research problems that I think might be exciting in this area.

I wanted to tell you a little bit about some problems in cryptography that I think might be interesting, particularly to folks interested in the algebraic side of things. Let me tell you about block ciphers, which are useful in cryptography.

What is a block cipher? A block cipher is a black box like this. It takes two inputs. It takes a key K and it takes a plain text X and if you fix the key, then

this is a bijective map on some space. Moreover, we insist that it be bijective for all keys. This just means -- the bijectivity means that it can be decrypted, and should be efficiently computable in both directions.

This picture illustrates a common design theme for how you build block ciphers in practice. You build a number of simple transformations, call them rounds, and you compose them, just take the composition.

The obvious question to ask is, when is a block cipher secure. Let me give you the definition that turns out to be useful. On the left, we have a block cipher keyed by some key K that is chosen randomly and kept secret by the parties to the conversation. On the right, I want you to imagine pi, which is permutation chosen uniformly at random from the set of all bijective maps on this space.

I want you to consider the distributions on this. The block cipher is secure if intuitively it approximates the behavior of a random permutation. Let me give this to you in a very operational sense. Imagine that behind my back I am going to either decide I'm going to give you a block cipher or I am going to decide I'm going to give you a random permutation. You don't know which. If it is the block cipher I am going to pick a key at random, if it is the permutation I am going to pick a permutation at random.

I am going to build an implementation of them. I am going to stick them inside a black box, and now I hand you the black box.

The black box has an input and an output. You can feed any inputs you like and observe the outputs and interact with it adaptively, and your role is to guess how do I substantiate the black box. This is a secure primitive if you can't guess. That is the object.

DR. BORGS: Using one instance or several instances?

DR. WAGNER: Using one instance. We want to know what is the probability you can correctly identify what kind of black box you have been given. So you can flip a coin and be right with a probability of half, and this is secure if you can't do any better than that, if you have only negligible advantage compared to flipping a coin.

Here is an example of a block cipher that has been standardized, that is going to take over the world. It is called the AES. It is the new standard. It is fantastic, it is great. It follows this product cipher methodology, so it iterates the number of rounds.

Here is a round. I wanted to show you this in detail, because I wanted to show you just how much

algebraic structure is present in this structure that is being used in practice.

What does a round do? This operates on 16 bytes. So you can think of 16 elements from GF-256. A round we have exclusive or key material, so those are just additions in the finite field. Then we apply 16 S boxes, they are called. They are simply bijective maps. It is effectively just the inversion map in the finite field. Then we hit it with some linear transformations, which are the bottom two layers of the round. We do that about ten times, using different keys in every round.

The claim is, as far as we know, the conjecture is that this is a secure block cipher.

PARTICIPANT: It is secure in the prayer theoretic model.

DR. WAGNER: In the prayer theoretic model. We have no proof whatsoever. In fact, there are good reasons to believe that it is not likely going to be easy to provide a proof.

So you can see there is a lot of mathematical structure here. In fact, as I described it so far, this is working entirely in the finite field. That leaves so much structure that that is probably not good for security. So the S box is tweaked a little bit by adding a linear map,

composing with a linear map that -- anyhow, the details aren't too important.

Let's move on. I also wanted to show you a candidate public key encryption cipher that has been proposed. Here is one such candidate. The private key is a pair of linear maps, L and L', general linear maps. They are linear over the finite field with 256 elements.

This middle layer consists again of a row of S boxes, but in this case the S boxes will be a bijective map. It is just a cubing map in the finite field. That looks like it is order three; you can think of it as being close to -- remember, the squaring map is the Frobenius map, it is linear, so this is the first power map that is not a linear map.

The public key will be the composition of these three layers. Given explicitly, in other words, we write this as a multivariate polynomial with 16 variables over the finite field, and then we write down the coefficients. It is easy to see there can't be too many coefficients, because this is essentially a quadratic function, so the number of coefficients is something like 16 or something.

This it turns out is not a public key encryption scheme, but there are schemes much like this that are conjectured to be secure.

DR. CHAYES: So is this being used now?

DR. WAGNER: Not being used yet, because cryptographers are conservative, but it is a promising candidate. It has a number of promising features.

Let's move on to the next slide, which will be the conclusion of this area. Let me suggest a problem that I think will be very interesting to know something about. The problem that seems to come up in both these private key and public key settings is the following one.

We have a system of N equations. An equation is a multivariate polynomial, N unknowns over some finite field. Typically the kind of finite fields we care about are the two element finite field or the 256 element finite field, or small finite fields of characteristic two. Often the multivariate polynomials are sparse and they are low degree, and we may frequently have a very over-defined system of equations, so many more equations than unknowns.

The problem is to efficiently find a solution to this. Here is what is known about this problem. First you might think there is some reason to believe this is a hard problem. If you have the same number of equations as unknowns, even in the case where you have only got quadratic equations, this is empty. So there are unlikely to be efficient algorithms for it.

Yet, on the other hand, if the number of equations exceeds the square of the number of unknowns, then there is a polynomial time algorithm, linearization. You simply replace all the terms, XI, XJ. There can only be N squared such terms, and choose two such terms. Replace each one by a new linear formal unknown and you solve the linear system of equations with Gaussian elimination.

Now the interesting question is, what is in between the two. This linearization led to a successful attack on some of these systems. Since then in the crypto literature there has been a number of improvements and refinements discovered. It has been found that you can replace this constant half by any other constant and you still have a polynomial time algorithm by a recursive application of linearization.

By some ideas inspired from Gröbner bases, there is a conjecture that if the number of equations exceeds the number of -- this is an error over here. This is supposed to say, if N exceeds N plus C. C is the number of unknowns plus a small constant. There is an algorithm that the authors have conjectured to be run in sub-exponential time. So that is a significant improvement over the naive algorithm of just guessing a solution.

The obvious question --

PARTICIPANT: By conventional computer?

DR. WAGNER: On a conventional computer. So the obvious question is, why not use Gröbner basis algorithms. Let me just say, to my understanding, they are not competitive for these kinds of parameters, because they are exponential in practice and the constants are so huge, that more than 15 unknowns is infeasible, which compares very poorly to the naive exhaustive search attack.

So here is a problem that I think might be very interesting to look at, because I believe there would be a number of applications for cryptography.

I have used up more than my time, so I'll end here.

DR. CHAYES: This is a semi-efficient algorithm, so what are you looking for exactly here?

DR. WAGNER: Yes. The research question I think is to understand the complexity of this problem.

DR. MC CURLEY: Actually, I think I proposed something like this as a meta problem when Abelman and I wrote this paper on open problems, because it has many facets to it.

DR. SCHMIDT: Jennifer, do we take a break now and come back for discussion?

DR. CHAYES: Yes, we'll have about a 15-minute break, and then we'll have the discussants.

A Few Open Problems in Computer Security

David Wagner

Two topics might lead to fruitful collaboration between computer security people and mathematicians:

1. *Critical infrastructure protection*. Infrastructures such as electric power, water, oil, gas, and telecommunications were not necessarily designed for security when they were first deployed, and they continue to evolve. They are increasingly dependent on information technology, which is troubling because the security of IT is not reliable enough.

- "Can we build a mathematical model that allows us to analytically express some of the system's properties? In particular, can we measure security against malicious attack?
- Is there an efficient way to detect whether there exist any lines in the power system whose single failure will produce a cascading failure?
- Can we (re)configure the system to eliminate or bolster these weakest links?
- More abstractly, can we design systems that are inherently self-stabilizing—that is, robust?

2. *Enhancing security for block ciphers*.

- We should investigate the AES standard for secure block ciphers.
- We should investigate a candidate public-key encryption cipher that is conjectured to be secure.
- We should investigate a certain polynomial-time algorithm produced by a recursive application of linearization.

Andrew Odlyzko
"Remarks on Communications and Computer Security"

Transcript of Presentation
Summary of Presentation
Video Presentation

Andrew Odlyzko is director of the Interdisciplinary Digital Technology Center, holds an ADC professorship, and is an assistant vice president for research at the University of Minnesota. Prior to assuming that position in 2001, he devoted 26 years to research and research management at Bell Telephone Laboratories, AT&T Bell Labs, and AT&T Labs, as that organization evolved and changed its name.

Dr. Odlyzko has written more than 150 technical papers in computational complexity, cryptography, number theory, combinatorics, coding theory, analysis, probability theory, and related fields, and has three patents. He has an honorary doctorate from the Université de la Marne la Vallée and serves on the editorial boards of over 20 technical journals, as well as on several advisory and supervisory bodies.

He has managed projects in such diverse areas as security, formal verification methods, parallel and distributed computation, and auction technology. In recent years he has also been working on electronic publishing, electronic commerce, and the economics of data networks, and he is the author of such widely cited papers as "Tragic loss or good riddance: The impending demise of traditional scholarly journals," "The bumpy road of electronic commerce," "Paris Metro pricing for the Internet," "Content is not king," and "The history of communications and its implications for the Internet." He may be known best for an early debunking of the myth that Internet traffic would double every three or four months.

Andrew Odlyzko's e-mail address is odlyzko@umn.edu, and all his recent papers as well as other information can be found on his home page at http://www.dtc.umn.edu/~odlyzko.


Andrew Odlyzko

DR. SCHMIDT: Thanks, everyone, for coming back so rapidly. We are going to move to the next section on review and discussion. Michael and Andrew are going to be joining us from the University of Minnesota and Microsoft Research. Michael will make his comments first. Well, would you like to go first?

DR. ODLYZKO: When you talk about the difficulty with big secure systems, just think how hard it is to do a very simple coordination system like this one. That is why software is hard.

Let me just comment on some of the talks here, and maybe also a bit more generally. Kathy Laskey had some very good comments about general issues, that we have to think about security at a systems level in general, and the issue of what matters to people.

When we do that, we also have to think about the general questions of what it is that we mean by security, or what kind of risks we are willing to accept, and look at a whole range of possibilities.

Just to make it very clear that we do have a wide range, let me tell you a little joke. The story goes that back in the old days of the Soviet regime, a Western group was visiting the Soviet Union. They are being driven through the Siberian tundra, drive for miles, not a soul in

sight, deep forest, et cetera. They come to a clearing, and they see a pile of gold bricks, and not a soul in sight, and nothing else. Their tour guide says, how come you have this gold here, totally unprotected? This is a Communist regime. Gold is nothing. The real treasure of the Communist regime are the people; those we watch night and day. So there are different ways to achieve security. The question is, how do you want to do it, and what kind of security do we want.

Something that Werner Stuetzle explained yesterday is that we have societies which are much more regimented than ours, which have suffered from terrorism and have managed to live with it. Indeed, while 9/11 was a striking event for us, you see many societies, some quite democratic ones such as the British dealing with the IRA, the German dealing with the Bader-Meinhof gang, the Spanish dealing with the Basques, having certain levels of insecurity and terrorism.

So in many ways, one could actually say the task is not necessarily eradicating terrorism, which seems to be hard -- everybody wants to do it, but it seems to be essentially impossible with the limits of some societies -- but keeping it to a tolerable level.

You may also look at some other risks that we put up with. 40,000 people die on the roads each year, after all. Now there is a big debate about the double nickel, 55 mile per hour speed limit, what effect it would have. If you go for a single nickel, five mile per hour speed limit, you could eliminate those deaths. Well, we are not willing to do that, which says that we are willing to accept a certain level of risk.

This then goes back to some of the comments people made, that we have to look at the whole system. We have to look at economics, sociology, politics, general public policy questions that are involved here.

Now, to come back to the presentations in the session, Dorothy Denning's presentation was largely at the large systems level, where she is explaining that we should be looking at the economic questions, return on investment. One way I might phrase some of what she said is that one could think about these issues in terms of insurance.

Other people have made the same point. I think Ross Anderson may have been first almost a decade ago when he said, when you think about security issues and if you think about what is the right level of security, ask your insurance company.

Unfortunately, that doesn't always work very well. The problem is that insurance is unsuccessful when you are dealing with well understood risks. Indeed, all insurance policies that I am aware of exclude war risk, and many of them increasingly are excluding terrorist risks, too. When you are talking about rapidly changing technologies, insurance may not be the right approach.

There is also the issue of market failures. The general trend has been for very good reasons to rely increasingly on the markets for resource allocations, but there are market failures. We do not rely on markets to provide police protection, et cetera. There is a question whether the commercial industry is behaving optimally for society, given the incentives they face.

That goes to the question of exactly what kind of assurance do we really want as a society, do we want government to either bribe or coerce companies like Microsoft and IBM into producing more secure systems. These are very important questions.

These are all very high level questions, and they are also the kind of questions which go to what we might call the integrationist line of thought in science and technology. Mathematics and physics has been more the reductionist approach. This gets to the question of very

uncomfortable cultural transition that many mathematicians and computer scientists as well have to undergo when faced with these questions. These mathematicians have tended to like nice, neatly posed problems.

It is also true of physicists. The theory of gravitation has been cited as one beautiful example, then of course Einstein's theory of relativity, and now we have the search for the ultimate unified theory of physics.

On the other hand, if you look at where resources are going or what is happening, they are going to other areas. There is a huge shift in general funding of research and development at the federal level, but also in the commercial sector towards the biomedical sciences. What happens in the search for elementary particles or unified field theory is essentially irrelevant for those areas. Even if physicists succeed beyond their wildest imaginations, nobody can figure how that is going to impact on the bulk of the research that is going on right now not for the next few decades.

So we can look at different levels. Most of the problems that society cares about, like reducing the risks of terrorist attacks, seems to be at a system level, the kind of things that Kathy Laskey has been talking about. But there are problems which are more congenial perhaps to

the traditional mode of operations of mathematicians and computer scientists, and we heard quite a few examples that she cited here.

Kevin talked about a variety of problems such as detection of covert -- here we are talking about a much more manageable problem. We can perhaps model it more easily, and can talk about applying a variety of mathematical tools in that situation. General issues of limits on security, exactly how much information is protected by different kinds of crypto systems. This is a very comfortable mathematical question we can attack.

David Wagner then posed a variety of questions having to do with algebraic crypto systems, and this is straight mathematics; we understand exactly what it is. But even in David's presentation we also had questions of the other variety, the integrationist approach, namely, questions about infrastructure security. There are some questions, very nicely posed mathematical questions, that do suggest themselves very easily. Or he talked about some others I can see coming out of his presentation. A more general question is, do you build networks which are ultra reliable but maybe with centralized controls and very efficient, or do you go for redundancy.

We see in the 9/11 events much of the success we have seen in communication was due to the fact that we had cell phones, we had wired phones and we have the Internet. Not a single one of them was faultless, not a single one of them operated as well as we might have wished, but altogether they produced quite a satisfactory response.

So there is quite a variety of different problems that mathematical scientists can investigate there.

DR. SCHMIDT: Thank you very much, Andrew. Michael?

Remarks on Communications and Computer Security

Andrew Odlyzko

It is important to realize that people are willing to accept some level of risk and that while it is not necessary (or possible) to eradicate risk, it is possible to keep it at a tolerable level. In addition, when one thinks of security issues, one often thinks of insurance. However, insurance usually deals with well-understood risks. The risks posed by war, terrorism, and changing technology are, in general, poorly understood.

These types of questions are high-level questions and go to what we might call the intergrationist line of thought in science and technology. In the past, however, mathematics often took reductionist approach; mathematicians tended to like nice, neatly posed problems. Many mathematicians, and computer scientists as well, will have to undergo a very uncomfortable cultural transition when faced with these questions. After all, most of the problems that society cares about, like reducing the risks of terrorist attacks, seem to be at a system level.

Nevertheless, many problems are much more congenial to the traditional mode of operation of mathematicians and computer scientists. Some examples proposed by other participants include the detection of covert channels and questions dealing with algebraic crypto systems.

The basic question is, Do you build networks that are ultra-reliable and efficient but maybe with centralized controls, or do you go for redundancy? The events of September 11 seem to indicate the latter, as cell phones, wired phones, and the Internet all worked imperfectly but well enough in the aggregate so that communication was satisfactory.

There is quite a variety of different problems that mathematical scientists can investigate without having to change their traditions, though they will have many more to consider if they adopt, at least some of the time, a more integrationist line of thought.

Michael Freedman
Session on Communications and Computer Security

Michael Freedman is a member of the Theory Group at Microsoft Research. Before working at Microsoft, Dr. Freedman was the Charles Lee Powell Professor of Mathematics at the University of California at San Diego.

The work for which Dr. Freedman is best known is the solution of the long-standing Poincaré conjecture in four dimensions, for which he received the Fields Medal. He has received numerous other awards and honors, including Sloan and Guggenheim Fellowships, a MacArthur Fellowship, and the National Medal of Science. He is an elected member of the National Academy of Sciences, the American Academy of Arts and Sciences, and the New York Academy of Sciences.

Dr. Freedman's current research focuses on fundamental problems in the theoretical computer science, in particular on the P/NP question and nonstandard models of computation.

Alexander Levis
"Introduction by Session Chair"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Alexander Levis is chief scientist of the U.S. Air Force, Washington, D.C. He serves as chief scientific adviser to the Chief of Staff and Secretary of the Air Force and provides assessments on a wide range of scientific and technical issues affecting the Air Force mission.

Dr. Levis received his professional education at the Massachusetts Institute of Technology. Prior to his current position, he was University Professor of Electrical, Computer and Systems Engineering at George Mason University in Fairfax, Virginia, and head of the System Architectures Laboratory of the C3I Center. For the last 20 years, Dr. Levis has conducted basic and applied research in and taught many aspects of command and control, from organization design for command centers, to operational and system architectures, to decision support systems.

Dr. Levis has served as senior officer in national and international professional societies, is on the editorial board of several professional journals, and on the Board of Directors of the AFCEA Education Foundation. He has held two appointments to the Air Force Scientific Advisory Board, where he participated in several summer studies and several ad hoc studies.

DR. LEVIS: Just on a serious note on that, the term stegonography was mentioned in the first session, and people asked what is it. As far as I know, Al-Queda has used it already. So they are fairly sophisticated.

Good afternoon. My name is Alex Levis, and I will be chairing this session. My current position -- I have never worn a uniform in my life, but my current position is with the Air Force. I am chief scientist. It is the best job in the world. It has no specification. What I am supposed to do is give advice to the chief of the Air Force and the Secretary of the Air Force. The Secretary of the Air Force did his Ph.D. on decision analysis at Harvard. He does not need any advice from me on matters technical.

The chief has a war to fight, he needs no advice from me. So I am having a great time coming to workshops, and I really appreciate the invitation.

Somewhere in my distant path actually, I do have a degree in mathematics, but it is an undergraduate one. I noted when David earlier -- there was a question asked yesterday, how many people have a degree in mathematics, and four or five people raised their hands. The question today that David asked is, how many mathematicians, and

two-thirds of the people raised their hands. I wonder what the significance of this is.

I would like to make some comments first. I have learned a lot of interesting things yesterday and today. Information security I am a little more familiar with. But the thing that hasn't come up very much is the problem. Those who want to contribute in mathematics, and I believe strongly in that, and I'll try to show a couple of examples, mathematics to the problem.

I am an engineer, and I need to understand the problem before I apply solutions to it. The problem has not been defined in this workshop. We all took it for granted. From the discussion it is apparent that our knowledge is what we all hear from the newspapers, et cetera. I don't think even the White House has defined homeland security, let alone -- I don't know the difference between homeland defense and homeland security, by the way, but different people use them in different ways.

So let me try to define in non-technical terms the problem. We don't really know who the adversaries are. We know some individuals, we know some organizations, but we don't have a complete knowledge of the adversaries. We don't know where they are. You see every night on TV that we really don't know where they are, and we don't know

where they are going to attack, whether it is going to be in the U.S. The call just went out again, after being prepared for a year and a half.

We don't know when they will attack. We do not know what they will attack, and we don't know how they are going to attack. We know bits and pieces of that, but this is the classic journalism, the five questions. So when we talk about homeland security, we have to look at all those aspects.

You can characterize the uncertainty. We talked a lot about probability yesterday and a little bit today, but there are different kinds of uncertainty that are associated with this. You have temporal uncertainties, you have locational uncertainties, you have all sorts of things that somehow have to mix and match together. This session, remember, is about integration and fusion.

Now, the strategies that you can consider fall in three categories: reactive strategies, --

PARTICIPANT: By the way, you forgot one question. We don't know why they are attacking us.

DR. LEVIS: No, I'm sorry, that we know. They don't like us. We can start a debate, but I come from that part of the world, and I can tell you in great length why they don't like us.

We have concentrated mostly in terms of what people have been discussing and the approaches they are indicating. We are looking primarily at the reactive problem, what happens if somebody does that. Much of the work that was presented in epidemiology would go into the reactive part: once there is an attack, what can we do about it.

There is also the anticipatory part. Some of the work yesterday relates to that. We know that somebody is crossing the broader. We don't know where, but we are going to look for it, we are expecting an attack on the bridges of San Francisco during that week. This is anticipatory.

The other one is the proactive; get them before they get a chance to get us, which is what we are doing right now, what the Air Force is actually doing. The term that the Air Force uses for this kind of a notion, that is, the last one, is predictive battle space awareness. It doesn't really mean anything. This is a bumper sticker to put at the end of the car. It tells you it is great to know what they have done. It is great to have all the sensors to tell me what happened, but that really is of very limited value. What I really want to know is what

they are going to do next, so I can go and get them before they do it. That is the hard problem.

With the expansion of sensors, information technology, et cetera, as many of you indicated, we have a wealth of data, but our difficulty is to take that data and start determining intent, which takes us away from just projection of the data. We have to understand what is going on, because you need to know intent, to understand intent, to make the predictions.

The first part I have already mentioned. Yesterday we talked a lot about searching for a needle in a haystack. I am the last person to tell you how to do that. I have no idea now to do that. But there are alternative approaches, how to think about the problem.

First of all, somebody mentioned yesterday getting a magnet or changing the properties of the needle so that you can find it more easily, given the kind of sensors that you have. A lot of our intelligence community is worrying about things like that. There are ways of trying to mark things that we eventually may want to find. That is one way of approaching it.

The last one is, find the needle before it gets in the haystack. As an engineer, this appeals to me very much. I am lazy. I want to solve the easy problem, I

don't want to solve the hard one. We need to keep this in mind. There is a whole spectrum of activities over here.

To get to this predictive integration notion, given that there are large gaps in knowledge, because you can get to intent -- intent is human, and you can't really model all that stuff. You cannot go to databases to find those things. You need to do some modeling and simulation.

This is now a case study to respond to a question that was raised later, you have all those good solutions, how do you go and give them for people to use. That is a very hard problem. You cannot just say, I have a wonderful solution. I think it was mentioned, you have to understand the science of the problem, we have to understand the application of the problem, because nobody is going to pay attention. You expect somebody else to understand the solution and then convert it and apply it. Not anymore; nobody has the patience for that. One needs to go out.

One example over here very quickly which is relevant to the presentations today, that is why I chose it, has to do with influence nets. After 9/11, actually a couple of weeks later in September, I went to the Air Force Studies and Analysis. These are the people who do the actual math and do modeling and so forth for the Air Force. Originally they were the whiz kids. Some of you who are my

age will remember McNamara; this is the whiz kids group of McNamara's. They still exist, with a fancy name.

I gave them some software from the lab, research software, to start modeling Al-Queda. They did it, and it has been used since then to do a number of analyses. That software was certainly not for prime time. It was homemade by my master's level students. Since then, we have got people from the Air Force Research Laboratory, this is from Rome Lab up in upstate New York, the information directorate, and they were developing software over there which had some relation to mathematics, but it was not very explicit what the relationship between the software and the mathematics was. It was very heuristic in a way, but it did approximately the same things, but it is much more user friendly.

Once we established from the laboratory the validity of the approach using more rigorous tools, a heuristic approach was being used and is used currently.

Now we need the data. Some of the data came from databases, as described yesterday, but some of the data had to do with judgments. We brought in the analysts from many of the three-letter agencies that were discussed yesterday to provide that information.

The moment the analysts got that stuff, after they learned within two or three days how to use the software and how to model in that way, immediately they started asking for more. All this does is, you wiggle the inputs to see what the outputs are.

This is good for planning, but how do I go to the execution? This is what we showed to the generals. This way they understood the influence nets and Bayesian networks, except for the Secretary, who does know Bayesian networks. Fortunately, I did my homework, and before I discussed those things I knew that he knew Bayesian networks, and I didn't say silly things to him.

But the idea of how to use them is a little bit different here. You have to put yourself in the mind of the adversary and define a set of effects that you want to achieve. We want Milosovich to change his mind, we want Saddam Hussein to stop bothering us and go on vacation or something like this. You put that stuff over here, and then you start looking, what are the things that influence his decisions.

This blob here is where the modeling takes place. You want to bring them to the point that you have what is called actional events, the things that we can do -- we are

the good guys, the blue -- that will eventually have an effect here.

This is the kind of models -- now we have made a bunch of models of those things. Actually we have been using this kind of stuff in DoD since 1994, in the intelligence part of DoD. They use it that way.

But now the question came up -- and I wish I had mathematicians working with me to solve the problem, because I don't know -- and my students don't know the answer, and this is not the field that we work in. Suppose now that they start the engagement. Bombs start dropping or things are happening in Afghanistan. I start having observations that come over here from various intermediate nodes; this thing occurred, that thing occurred.

Can I propagate forward? Sure, I can propagate forward except I have to update first and update the priors and do all those things, and do it. That is brute force, one can do it. It takes a few weeks, you get it done, you implement. But then they are a more sophisticated question. If I look at this node and I see what has happened and that improves my information over here, how well am I doing in reaching the outcomes, can I solve the interest problem? Can you tell me where I should put sensors to see what is happening, so I can get a better

understanding of whether I am achieving my goals or not. That is a lot harder problem, but from what I know, mathematicians will know how to do them correctly, as opposed to trying to ad hoc them, the way it is occurring right now.

But you cannot do it in two years. You need to be ready to go and work and do it in a couple of weeks and put it in. I hope it will be over before it is actually needed.

This is a real transition, but it took a lot of people. That is one of the good things when you are chief scientist; I could call people and cash in all my IOUs. This is how it has been established now, studies and analysis, now has the capability to do modeling, using Bayesian nets. We never had that capability before. The research laboratory, Air Force Office of Scientific Research, indeed, some of you in the audience probably know it, that is the basic research component of the Air Force research establishment. This is the applied research, the Air Force Research Laboratory. They are providing the algorithms and so forth.

Here is the alphabet soup of the intelligence community, providing the data and the updates. We are using it, responding to decision makers. What we need here

is better math, because I am math limited in this area. So are the analysts. It is not their job to be mathematicians; their job is to understand that and to provide the answers.

Opportunities, every day. But how do you exploit -- because I have good people like Kathy and Tod that I have known a long time who know about those areas. A long time ago, 1994, I started learning about this stuff; I could make the transition. It is not my technology, but at least I could see the possibility of a transition. It takes a lot of work and it takes involvement to get to know the problem. You need to understand the science. The devil is in the details, and you have got to understand.

I think I have said enough as an introduction to this session. We have three speakers. We are going to continue straight through with two discussers, and then an open discussion. The intent is to leave here by 6 o'clock. No reaction by anybody? You are all awake, I see, we are doing fine.

With this, I would like to ask the first presenter, Tod Levitt. I will ask each of the presenters to make appropriate comments about themselves. Although I know them, they do not want me to say what I know about

them. It is all good stuff, but it will take a long time to describe their accomplishments.

Introduction by Session Chair

Alexander Levis

There are three types of strategies for handling an attack:

1. Reactive. Once there is an attack, how is it handled?

2. Anticipatory. What are the procedures when we are expecting an attack?

3. Proactive. What will happen next, so that it can be prevented?

In a proactive strategy, mathematics is used to create models and simulations of the problem. One idea, from the Air Force's Studies and Analysis group, is to use Bayesian networks to put ourselves in the mind of the adversary, and to define a set of effects that we then want to achieve. The model helps identify the things that influence our decisions, as well as the actional events—the things we can *do*. We can also "propagate forward" to see how well we're doing in reaching the desired outcomes, and test the outcomes' sensitivities to what happened at key nodes along the way.

Tod Levitt
"Reasoning About Rare Events"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Tod Levitt's research has focused on the development of advanced capabilities for multisensor fusion, SAR, IR, and EO image understanding, ground robot vision, air-to-ground surveillance systems, and C4ISR systems supporting multiple military intelligence, planning, and command and control applications. He has authored over 50 publications in the fields of multisensor fusion, image understanding, robotics, and artificial intelligence. Dr. Levitt has a track record of unique developments in evidential reasoning in large-scale, high-dimensional data, qualitative navigation for mobile robots, encryption of algorithms, and the foundations of model-based software system design.

Dr. Levitt has led the development of a diverse family of advanced information software systems built to handle real-world data under complex operating conditions. The breadth of these applications reflects his unique ability to transfer technology capabilities across sensor modalities and to demonstrate the application of advanced fusion techniques to dynamically meld the resulting multisensor observations into actionable intelligence.

These systems include a fully automated middle-Eastern armor unit detector for the U.S. Army that was evaluated to perform at expert imagery analyst levels on wide-area, low-resolution Desert Storm SAR; a system for automated diagnostic measurement from digital x-rays of the hand that was employed in clinical care at the San Francisco Veterans Administration Medical Center; a system for semiautomated, three-dimensional construction of neural processes from multiple hundreds of two-dimensional cross-sectional confocal microscope images that was used in developmental anatomy studies at Stanford University's Department of Neurosciences; and development and installation of an automated hot steel slab inspection system at U.S. Steel in Gary, Indiana.

Dr. Levitt founded Information Extraction and Transport, Inc. (IET) in September 1991, aiming to achieve large-scale, breakthrough technology development of automated systems supporting scientific discovery. IET provides products and services that manage uncertainty to add bottom-line value in enterprise computing across business, defense, and engineering industries. IET brings together a world-class team of technologists with a shared technology vision who produce best-of-breed tools and state-of-the-art solutions for dynamic Bayesian inference and decision making.

Prior to founding IET, Dr. Levitt was employed as a research scientist at Advanced Decision Systems from 1983 to 1991 and at the Honeywell Systems and Research Center from 1978 to 1983.

Dr. Levitt is a co-founder and member of the board of directors of the Association for Uncertainty in Artificial Intelligence. He received his Ph.D. in mathematics from the University of Minnesota.

DR. LEVITT: Thanks, Alex. I have an advantage, in that I have few accomplishments, because I have spent the last 12 years running a company that deals with applied research, largely in information fusion for tactical battle space applications, a lot of work with DARPA and other military agencies. As such, we have a wealth of experience in knocking our heads on problems that the government typically -- impossible problems the government states vaguely on purpose to try and get smart people like yourselves to come forward with answers to the questions they can't even pose.

Certainly anything called homeland defense fulfills this in spaces.

The title is reasoning about rare events. I meant the term rare in two senses. One is unlikely, the traditional sense that we simply don't expect any given thing to happen with any great likelihood, and the other is more technical, a concept due to David Schumm. In fact, everything I will talk about is somebody else's work.

In that sense of rareness, we have the concept of rare evidence, that can occur even when a rare event is happening with certainty. Rare evidence turns out to be very important for detecting anomaly. In this way, we can hope to model things that we might expect and still detect

that something else is going on within the context of those models. I will get to that eventually.

But first, to wave hands at what homeland defense might actually mean, any notion of it has to immediately acknowledge that it is a stupendous panorama, any way you look at it. This is one little cut Alex gave on it. Here is another, to attempt to trot out taxonomies of what might be, and multiply them down and come up with possibilities for what we are trying to defend against. So you are talking about people in organizations that want to hurt the United States, of all varieties there, where they might come from. They don't necessarily come from overseas. They can be recruited here, of course, all kinds of different attacks, especially conventional ones. TNT is readily available in the United States. You can talk about hitting all sorts of supply chains from electrical grid to information to the waterways and actual transportation, agriculture, livestock is a huge issue. It would be very easy to cause a terrific economic impact, not to mention effective terror, and political, information itself. Information is very easily spread and can cause a lot of problems in the United States.

I had an experience in December. I heard a noted scholar remark that since we had a machine that could play

chess at a grandmaster level, doing automated intelligence assessment for tactical warfare surely must be solvable now. I was kind of shocked, having spent most of my life trying to do that problem. But it struck me after the fact, a lot of people think -- when you talk about automation which I insist, given the size of the problem, is where we want to focus, at least from the point of view of where mathematical breakthroughs are needed, and brilliant applications are needed in order to make any progress at all.

You start looking at what chess looks like versus tactical warfare, which I am very familiar with -- these numbers are just wags. They can be substantiated, but I won't take your time. If you just casually look at what you are talking about here, you see two orders of magnitude in size basically as you move to the right each time.

Most importantly, you notice that there is no uncertainty in what you observe in chess. It is all right there in front of you all the time. That of course is spectacularly not the case. It is the source of what makes data fusion necessary. There is terrific uncertainty in what is out there, where it is at any given point in time, what it is doing, or as Alex said, the buzz word du jour, predictable awareness, what it will do.

Then when you actually do see something, typically 65 percent is the typical number that turns up as to how certain you are of what you are looking at. That is in tactical warfare. That is very constrained. You might be a little surprised that organizations in tactical warfare -- random number, just pick one, eight is typical. It is not commonly known when you go looking at these current types of tactical conflicts that go on.

For instance, in Bosnia there were three major armies, the Serbian, the Croatian Muslim and the Croatian Christian were all more or less conventional armies, meaning they had big armor and artillery and things that they had to haul around with big logistics trails. There were 28 paramilitary groups, including mujahadeen coming in from all over the world, mercenaries, all kinds of different religious factions, all doing crazy things for strange reasons, joining up with the army, switching sides. This is pretty typical of war situations. There are many actual organizations to consider.

When you go to homeland defense, everything is much worse. You are talking about the territory of the United States including Alaska and Hawaii and such. You don't even know what to look for. What are we looking for?

If the answer is that you have to know about box cutters, this is a little frightening.

Is it completely impossible? No, I don't think so. Mostly impossible, maybe, but there is actually a great deal of information that is being collected all the time, and that has been. There are many standing organizations, Centers for Disease Control, the USDA, Plum Island spends all its time, as has been discussed at the workshop. The Treasury Department is always auditing and tracking all sorts of things. And of course, the intelligence agencies are doing these.

One can imagine trying to -- these are efforts going on more or less. This is a cartoon of what could be. You look on the right-hand side, you have different sources of expertise on the left, different areas gathering statistics. These can be the same as they are at CDC, but they don't have to be.

If you start asking to consciously integrate these, so that the information is being automatically sifted and compared, and regionally focused and then hierarchically fused up from regions, it is conceivable that a great deal more might be known. It makes it obvious, I hope, that automation is absolutely critical to sift and compare in an attempt to match against models, and

that this needs to be done from a strong scientific perspective, with any hope of not just generating spaghetti and garbage.

Quite basically, if we are going to have automated reasoning about what is going on when one formulates hypotheses on the basis of evidence, just to nail things down a little bit, if we had a hypothesis that was anthrax attack at a location in the United States, and that is either true or false, there is evidence of the sort of thing the Centers for Disease Control gets. They get reports from physicians that have to report certain occurrences such as anthrax. So they get anthrax diagnosis incidents at a location. It is baseline, whatever that may be, zero most places, but it is not zero actually in the absence of attack; anthrax is naturally occurring. Then anything else would be continued high.

Then we have a standard one-stage Bayesian inference here. It is important to notice that in modeling this, typically the way we think of this in building actual applications, it need not be causal, but it is a powerful way to think about the model. If there is an attack going on, then obviously one should expect to see an incidence of disease. That is after all the purpose of the attack.

The inference goes the other way. You want to observe the incident and infer whether or not an attack is going on. The Bayesian rules are built to do that. In a one-stage inference like this, the likelihood ratio has the information. So the likelihood of any -- once you observe the state of evidence, in this case it would be either high incidence or normal, then the ratio has all the relevant information in a one-stage inference.

As you build complex networks, one-stage inference is no longer so simple. We change to real-life modeling in terms of all kinds of webs. There is actually a very large literature since the 1980s, focused in the uncertainty in the artificial intelligence community. There has been a proceedings, an annual conference since 1985, and those are available since 1990 through Morgan Kaufman; I highly recommend it to any mathematician who wants to get quickly up with the state of the art in what is understood in automated Bayesian learning and in computation.

There are two basic algorithms that have been discovered in 1988. Spiegelhalter developed the so-called joint triagramen, which solves any Bayesian network that is well structured automatically. Then in 1991, Bruce d'Ambrosia and Brendon Claverill developed the

probabilistic inference algorithm, which also does the exact solution for any Bayesian network, based on an approach of heuristic rearrangement of polynomials, all probabilities here being modeled as discrete. That is the website of the organization.

Fortunately, Professor Laskey is going to talk a great deal about this, so I won't spend any more time on it.

The next one, to get back to how we get to these complex models that require these sorts of algorithms, as we look at what really goes on and what the evidence comes from, this is a little more realistic here.

In this case, what you actually observe is a report at the Centers for Disease Control. That report may or may not accurately reflect the diagnosis that is supposed to be referred by a doctor because errors happen when reporting goes on, especially when it is done on a massive scale. Then of course, the doctor's diagnosis may or may not be correct.

So the observability -- was the report made, was it made accurately -- the credibility of the doctor, doctors who have no experience. There is an attack in the heartland or something, it is very likely that most of the physicians that are seeing ill patients have never seen a

case of anthrax. There can be delays, there can be all kinds of different sorts of things. Together, if we look at the hypothesis I know, it is that people are actually ill from anthrax, and that is hidden. It is not observed at all.

But together, we see that we are modeling the veracity of that illness, based on factoring over the credibility of the physicians who are diagnosing, and the accuracy and timeliness often of the reporting.

I stuck another random variable in there, the dispersion. This is to get the sense of what is relevant. One might think off the top of one's head that the weather is not relevant to whether an anthrax attack is going to happen, but it certainly is if a crop duster is being used to disperse it. It is going to be timed so that the wind is blowing and is likely to continue to be blowing and such, to cause the most damage. So what one might think of as ancillary factors might become critical first-order variables. This issue of what is relevant is place modeling, especially for something as broad as homeland defense might be defined to be.

So in a deep hierarchy like this, odd things happen. Most of the approaches to information fusion that have been used for automation focus from the tradition of

what we might call a standard probability Pascalian type probability, and focus on conditional independence, whether or not it is true, between random variables, and the weight of evidence that a piece of evidence carries in its pre-post area to an intermediate hypothesis or another piece of evidence, or the target's hypotheses about what is going on that one is interested in.

But as this example suggests, there is actually a lot of other issues that come into play when one goes to do large scale modeling of these scarce events, as opposed to, we are going to drive tanks into somebody, and we have a pretty good idea of how they are going to respond when we do that. They are not going to like it, and they are either going to run away or they are going to shoot back. Whereas here, things are much more wide open.

So the issue, as suggested here, is evidential semantics and how that applies to how one should model things mathematically.

The granularity of reasoning, how fine is it, how does one define the relevance of information, how fine does it have to be in order to get sufficient accuracy to draw conclusions to take action on? Again, even naive reasoning, I hope, when one looks over the breadth of the land suggests that any action taken is going to be

enormously expensive. So you want to be careful about what you actually choose to do.

How finely do you have to reason in order to get good enough probabilities, whatever calculus might be used, so that your conclusions are likely to be robust? Is that possible?

That leads to the question of completeness, whether you are spanning enough of the world or not. These are all different approaches which are amenable, most of them explicitly so, to graphical representation. The importance of that is that it leads naturally from the mathematics supporting the fusion of the information to algorithms that software engineers can actually build and compute.

By the way, I might mention that the exact Bayesian inference solutions for arbitrary directed acyclic graphs of random variables is NP hard. It has been proven so by Greg Cooper in 1992. It is in the uncertainty AI literature. That raises the issue of approximations. For instance, variational mean field approximation is an explicit attempt by Tommy Jeckle and Michael Jordan -- not the basketball player -- to come up with a robust computable approximation to very large scale Bayesian networks, et cetera.

On the bottom here, the so-called belief functions, also called Dempster-Schaeffer functions and networks, and fuzzy set theory, are two examples that deviate from the Pascalian tradition in explicitly being motivated by other modeling considerations.

In particular, we look at what some people in philosophy have probably called non-Pascalian, modeling ignorance. When you look at homeland security, this raises its head as a big issue. We simply don't know an awful lot. It is forbidding to think of actually modeling everything one might. So this suggests looking at three logics, where you have a supporting, denying, don't know.

Then work pioneered by, among others, Professor Dempster, upper-lower probabilities and inter-value probabilities motivated by some of the same issues. Fuzzy theory, motivated by lack of precision in being able to express, lack of ability to obtain calibratable estimates. Of course, none of these calculi are particularly miscible with each other, and there are other issues.

The one I would point at from a computational perspective and an automation perspective that has plagued the use of these theories is, in order to take action, ultimately we end up having to rank things. And certainly in automation you do. Interval valued or pair valued

calculi of course are not rankable, and it is easy to prove that if you do have a mathematically consistent way to rank them, then they have to flatten down to be equivalent to probability. Nevertheless, these are important issues to look at.

Something that is much less well known than the work of Professors Dempster and Schaeffer and Zotti is work by Jonathan Cohen, and more recently in causality with Glen Schaeffer and Yuta Proboth making very significant contributions.

These approaches, Baconian they have been called, the Humeian approaches, address in the first case, instead of weight of evidence approach to reasoning, eliminative reasoning, attempting to come up with exhaustive explanations that essentially deny a hypothesis on logical grounds, and counting them.

This may seem silly when you say that of all the possible explanations. But in fact, that is more or less what you do when you model in computer systems. You have got to go build models, and when you are done, there is at least a numerable number of them in some sense. They may be parameterized, but there typically is a finite number of models. So even with the parameterization, you are only modeling so much of the world.

John Cohen's calculi addresses the issue, not sufficiently for automation, but it is certainly an excellent place to start; how do you know when you've got enough, certainly a critical issue in homeland defense.

Can we infer causality? Going back to the Reichenbach and Fisher, there was a tradition in the 20th century in science of saying that causality could not be inferred, only correlation. Now both Schaeffer and Pearl have challenged that, and have algorithms and approaches that address this question of, can we from the observable infer the causes of doing them? This has the possibility then of discovery, without having to do so much explicit modeling.

All the systems that we build at IT are based on hierarchical Bayesian inference, because there is tremendous power there to exploit, and there are fascinating issues that arise in attempting to do that exploit in the weight of evidence reasoning.

One of the things that happens when you look at homeland defense is a wide area of fusion. You are going to talk about evidence arising from all over at computers that are going to then do fusion. In any national scheme, one would have to have a hierarchy of those or a web of them. Evidence will come not necessarily in the temporal

order of occurrence or in relevance, and it may not be conditionally independent from evidence -- in fact, it is unlikely to be conditionally independent from evidence that has already been fused in models that have been spawned, or hypotheses that have been generated to this point.

I would strongly point you at David Schumm's book, Evidential Foundations of Probablistic Reasoning, which gives the first scientific treatise of these issues in toto. I am going to borrow heavily from his results in the following. He distinguishes weak versus rare evidence, and this is evidence of which I was speaking at the beginning.

Weak evidence is when a piece of evidence does not either strongly confirm or deny the hypothesis that it purports to support. That means that the likelihood of that evidence relative to hypothesis is close to one.

Rare evidence on the other hand is evidence where the absolute likelihood does not provide strong support or denial. That is to say, it is close to zero. In other words, what that is saying is that evidence doesn't occur very often, no matter what the state of the world. The significance of rare evidence is that one could choose to model it on purpose, because when it occurs, it is going to

suggest that the hypothesis to which it might be attached in a model based sense is probably not the right one.

So by explicitly attempting to model rare evidential states relative to hypotheses that we expect, we can build in anomaly detection and hope to essentially have an inference power that is far greater than the model space that we can explicitly spawn. Kind of a subtle point, actually.

What is interesting is that as Schumm points out in his book, that is not as widely known as it ought to be, when you go to multi-stage hierarchical inference, the likelihood ratio no longer tells the story. In fact, the difference of the likelihoods is evidential in two or more states, so rare evidence becomes very salient. You can have weak evidence that is rare or weak evidence that is not rare, and rare evidence that is not weak, et cetera.

What is interesting is that weak evidence in hierarchical based inference and solution algorithms and Bayesian nets and such, the order that you accrue that evidence, given a static set to accrue, the order of that accrual will not significantly affect the results except up to precision. In rare evidence, the inference order varies a lot.

This is just a conceptual example, or an explicit way that I stuck in this idea that we could have a medium amount, whatever that would mean, of illness in the population. One would expect that in an effective anthrax attack, you are going to have a high incidence, where high doesn't have to be very many, of course, in the human population in the United States.

The normal baseline -- by the way, those baseline statistics are available for all regions all over the United States from the CDC, so these are things that exist off the shelf -- and you can expect to see low or none. So explicitly modeling a medium amount of anthrax is a way to build in an anomaly detector.

For instance, what actually happened with the anthrax being distributed by mail turned up exactly that kind of behavior. It was not like you would expect with an anthrax attack, but it was much higher than the baseline. It was like a medium amount. That should suggest right away that it is not like an aerosol attack or something, that one would get.

The weak evidence here is, if you had a doctor for instance who consistently misdiagnosed anthrax, then him saying things are normal is not particularly evidential to whether or not an attack is going on. That is an

example of weak evidence. So the credibility of a doctor to introduce that.

This is an example again drawn from Schumm and applied to this example. The weak evidence here is, if you have a report here of a given type, if you look down at the lower right, at the red numbers, it is saying that the probability of seeing those reports given on the doctor diagnosing is low and the probability of seeing the report if he is not diagnosing it is of course much lower. In that case, this would be modeling that as rare evidence, not that it necessarily is; the one I pointed out as rare is in case three. But the point here is just to show how the numbers work, if you insert them in and just do hierarchical Bayesian accrual up.

You see down at the bottom here, we have the likelihood of the report, being that you do receive the reports for anthrax, given the concatenation of all the hypotheses above it, which is more or less the same as asking what is the force of this report on the final hypothesis at the top.

As you can see, what is fascinating here is that when the rare evidence is close to the base of the argument, of the inference, then you get the same result as when there is not any rare evidence, in terms of the force

of that, the likelihood of it, on the final hypothesis, if you will.

But as the rare evidence is moved up the chain, that force of the base report, the thing you actually observe, becomes weaker and weaker. In other words, the whole argument weakens. It is a very unexpected result.

What this says is that when you start saying, okay, we are going to start distributing some kind of sensors around -- those are the green things. This could be people, it could be the current CDC, it could be hardware, whatever it is, any kind of way to make some kind of observations. You are going to distribute them spatially. They are going to report based on some protocol. They are going to report some machine that is going to choose that with other information.

The first thing to note is that when you distribute this sort of thing, this issue of rare evidence now when it arrives there is going to affect your conclusion. That is a problem. There is no answer to that. That is a math problem, if you will.

Now I am changing direction here. The evidence itself for the sorts of things that are often modeled for homeland attack as being the critical large destruction type of issues, is often diffusive itself. So you have

aerosols, you have disease incidents. Foot and mouth for instance can be trivially transmitted. It doesn't hurt people, you can carry it in your pocket, you can walk it right through Customs. None of the things detect foot and mouth when you go through the airport. They will say, you have a lot of grease in your pocket. You can rub it on an animal, and that herd will be infected in a day.

It is a little-known fact that I found out, little-known to me, anyway, that there are only about 22 facilities in the U.S. where all live processing from animals goes to stuff on the shelf. So if you go rub foot and mouth all over those facilities all at once, you are going to have a trillion dollar impact on the livestock industry in the United States. It is not that hard to do. That is a simple example; how could you detect this?

So you get these examples which have to do with combinatorial problems. Optimal placement. Again, the happy medium here is cost. You absolutely want to minimize any active deployment, because most of the time it is not going to be doing you any good.

Optimal placement and the assessors. So I have the computers here, stuff that is fusing. How do you do that, where do you put them? You have the issues of going over telecommunications networks, you have the issue of the

latency of flight. You are trying to infer here -- this gets to some of the issues Alex raised of space and time -- you are trying to infer whether you are getting a hypothesis that this is an aerosol or some other thing, a subway release, for mail if it was anthrax. But this also applies to other distributed sorts of things like worms and viruses in computers. You want to do this in such a way that you get a rapid convergence of such evidence when something is going on.

I'm going to blitz through this. This is an example of the kind of modeling work science that I teach. Jorgensen has been doing this as a reference in e-mail. The idea is to apply population ecology models, mathematical models, dynamical systems models, to detecting worms and viruses in computer environments, and also to proactively attempt to intervene to design them.

In this case, you look at a computer network and look at what goes on in terms of a work flow model, which corresponds to the growth and death of things in a population. You model that. You then can do a standard differential equation model and do the standard dynamical systems analysis. You can play with that in a what-if scenario to say, what if we changed that, what if we change the protocol, so that you get different effects in terms of

stability, in terms of how it would be taken over in the introduction of a worm.

So this is an example, just one in a fairly tractable space, of the sort of thing that I mean when you talk about modeling how evidence is gathered, and what can be done about it.

DR. BORGS: (Comments off mike.)

DR. LEVITT: This is not a Bayesian network. This is a diagram that indicates the work flow.

DR. CHAYES: But if it has loops on it, you can't solve it.

DR. LEVITT: No, it's not a Bayesian network. It is a work flow diagram that is used to model the differential equations for the varying quantities that are being exchanged in the work going on in the LAN. Then when you start having worms and things, that affects the parameters in the equations. That is not a fusion technique, it is a modeling technique for the evidence itself.

Just a quick left turn here. If you now take a proactive defense view from a building point of view, instead of looking at the diffusion outdoors or whatever, or through computer systems, if you look at a building that is purposefully being asked -- and they do this with State

Department buildings and such now. You say, if there was an aerosol release outside, how would it penetrate, and how would we deploy sensors to do that. What you get is amenable to art gallery theorems. You have vents, windows and doors and things where this can come in. That is a completely different approach, but very valid in the sense of defending.

Finally, these two things have to come together. The evidence has to match ultimately into the inference chains. There is no such math. These are different bodies of mathematics that simply aren't miscible at the current time.

This is the punchline. Where should we focus? Well, certainly calculi for automated inference for fusion, issues in accounting for granular completeness and causality. We need to solve the problem of the asynchronous arrival of rare evidence, the optimal placement combinatorial problem of sensors and assessors, especially from the point of view -- and this is the twist that hasn't been done, the rapid convergence of evidence in the fusion algorithms that are used, not just the placement to meet some objective function.

Finally, the integration of these mathematical representations, or the interoperation of them, if you

will, to transforation, from the continuous dynamical systems, boundary value type problems that you get when you look at diffuse phenomenon, to the inference webs and chains that have to occur if you are going to do fusion on machines.

Thank you.

Reasoning About Rare Events

Tod Levitt

The goal of data fusion is to allow us to consciously integrate data from organizations such as the Center for Disease Control, the U.S. Department of Agriculture, the Treasury Department, and the intelligence agencies, so that the information is automatically sifted and compared, regionally focused, and then hierarchically fused up from the regions.

It is absolutely critical to build an automated system in order to sift and compare collected data with highly complex mathematical models. Automated reasoning is a function of how observed data may depart from baseline conditions.

Researchers are faced with several challenges in the area of data fusion. First, it is crucial to verify the accuracy and timeliness of data. Also, given the breadth of the United States, any action taken is going to be enormously expensive.

Some areas of future work include the following:

- Developing calculi for automated inference for fusion;
- Determining when information is relevant and how complete it must be in order to give a level of accuracy necessary for drawing conclusions;
- Handling the problem of asynchronous arrival of rare evidence; and
- Transitioning from the mathematical models that arise in studying diffuse phenomena to inference webs and chains that are required for data fusion.

Kathryn Laskey
"Knowledge Representation and Inference for Multisource Fusion"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Kathryn Laskey is a professor in the Systems Engineering and Operations Research Department at George Washington University. She received a B.S. in mathematics from the University of Pittsburgh, a master's in mathematics from the University of Michigan, and a Ph.D. in statistics and public affairs from Carnegie Mellon University. Her research interests include Bayesian inference and decision theory, multisource fusion, uncertainty in artificial intelligence, and situation assessment.

DR. LEVIS: Thank you. We will proceed right away with Kathryn Laskey, who is on the faculty of the systems engineering operations research department at George Mason University. That is where I also am in my real life. Kathy.

DR. LASKEY: I am going to talk about multi-source integration.

DR. LEVIS: While Kathy is working on the hardware problem, I mentioned before some of the work my lab is doing on Bayesian network, et cetera. I send my students to Kathy to learn that stuff. That gets me off the hook.

DR. LASKEY: Before I plunge into my talk, I want to stop and say, these days it is hard to talk about multi-source information fusion without talking about Bayesian networks. The inventor of Bayesian networks was the father of Danny Pearl, so I would like to dedicate my talk to Danny Pearl and his father, and to say that I hope that Danny Pearl's memory will live on in those of us who apply his father's research to the problem of homeland security.

I am going to talk about some basic requirements for inference and decision support for homeland security which you already heard some of in my comments at the microphone earlier. Then I am going to talk about -- most

of my talk is focused on knowledge representation, which is my area of research.

In particular, I liked David's remarks yesterday about the cathedrals. I tell freshmen who are coming into my university, in our information technology engineering school, that they are living in a very exciting time. If you think about the course of human history, there was the agricultural revolution, where we learned to apply technology to food production. It enabled the emergence of cities. Then there was the Industrial Revolution, where we learned to apply technology to the development of the production of physical systems. Now we are moving into information technology, where we are learning to apply technology to information processing. I believe that the impact on human society will be every bit as fundamental as in the agricultural and industrial revolutions.

So I want to talk about representation of knowledge for the ability to do information technology. Then I will talk briefly about open challenges.

This is a quote from the Washington Post on September 12, which pretty much encapsulates what the challenges are in front of us. We want to be proactive rather than reactive. I'm not going to spend much time on this Vu-graph, because everybody said this already, but my

favorite phrase that I coined to speak about the issues that confront us with this needle and haystack problem is, data, data everywhere and not the time to think.

The Joint Directors of Laboratories has developed a taxonomy of information fusion, broken it into levels.

Level zero is pre-detection. That is, we are observing things that are observable in the environment, but we haven't yet said there is something out there, and can we fuse different sources to enable us to do a better job of detecting things that are out there.

Level one is detecting individual entities, like, I see a tank there, or I saw an animal with anthrax there. Level two is to support situation assessment, which is militarily significant for the Joint Directors of Laboratories, but for our problem, a significant configuration of entities in the environment that has some meaning. So for example, I have a company of vehicles, about 30 of them, or I have a pattern of unusually high incidence of anthrax.

Level three is support of threat assessment, which is, what is the intention of my adversary, so there is a bioterrorist attack.

Level four is sensor management, tasking with collection assets. That is the question of where do I put

my sensing resources so that I can do a better job of detecting what is out there.

What is really needed in practice and what mathematicians can help with is a unified, theoretically based technology that spans all five of these levels. Right now, it is pretty ad hoc and things are separate. We need to be able to span all the levels, and we need to be able to have a theoretically justified, unified approach.

If we are going to do a good job of this problem, we need to support and not replace the human. That means that we need to combine expert judgment with prior knowledge, apply expertise, but what is important is that we need to be able to apply expertise. In the homeland security problem, it is essential to be able to fuse both hard and soft types of knowledge. So not just the biology of disease transmission, but also, we need the knowledge of, for example, the political scientist about how conflicts evolve, the anthropologist, the social psychologist, the clinical psychologist on the mind of the terrorist. We need to look at -- a lot of different kinds of knowledge have to be brought to bear, and they are not always easily quantifiable, but they have to be brought together in a common framework.

We need to be able to present results so that humans can use them. As had been mentioned before, we need huge volumes of data. We need to span the five levels of fusion. We need to perform effectively. Although there is uncertainty, there are types of threats that we haven't seen before. Configurations of things that are meaningless in isolation, but when you put them together, they are meaningful, and huge numbers of possibilities to consider.

We also need to be able to learn from our experience. Not just humans learning and not just data mining, but we need to be able to combine the human strengths of pattern recognition and the computer strengths of sifting through large quantities of data to have an effective learning capability that makes use of both human expertise and automated processing.

Let me talk a little bit -- this is where I am going to talk about Danny Pearl's father. I am going to talk a little bit about knowledge representation, but first let me talk about a paradigm shift that I see occurring in computing technology.

The old paradigm is, you think of a computer as an algorithm running on a machine, executing deterministic steps, and it either gets the right or the wrong answer to the problem that you are solving. We are using logic, but

a computer is an automaton that processes preprogrammed steps.

There is a new paradigm that is occurring, actually. We hear these buzz words like agent-based systems. In fact, there is a recipe in recent years for getting a Ph.D thesis, I have noticed. You take your problem, some very complex and high dimensional optimization or statistical inference problem and you map your objective function if it is an optimization problem, or your measure of your log likelihood if it is a statistics problem. You call that energy or action, and then you invent a fictitious physical system, and you have these particles wandering around in the physical system, and you import methods from statistical physics to solve your problem. We have got the explosion of micro chain Monte Carlo methods and statistics and variational methods. This is a recipe for getting a Ph.D thesis. The Hotfield nets were one of them, Radford-Neal was another one.

If you think about this agent-based computing, we think of agents making decisions to achieve their objectives, and instead of writing programs, we are designing dynamic systems which by the laws of physics will evolve in such a way that we will optimize whatever we want to optimize, or infer whatever we want to infer. Then

we can put these in hardware. People are building neural net chips in hardware. In decision theory, we maximize expected utility or minimize expected loss. In physics, we minimize action. So let's think about building agent-based systems, and this paradigm shift actually is occurring. Instead of programming computers, we are designing economies of interactive agents. Just like any paradigm shift, the old paradigm is a limited case of the new paradigm.

Let me tell you very briefly what a knowledge based system is, because we are talking about representing knowledge. It is a computer program that is supposed to behave intelligently, so if you are going to design intelligent agents, you have got to use knowledge based systems.

So you need to have a representation language, which has a syntax as its grammar. The ontology defines what I am allowed to talk about in my domain. It is like my vocabulary, and it has the semantics, which defines the meaning.

I now have a knowledge base and an inference engine. The knowledge base is the main specific knowledge. The basic idea with the knowledge based system is to separate representation of a problem from solution of a

problem. And mathematicians are good at transforming from one representation to another, so if I can represent a problem in terms that are meaningful to an analyst, but put it into a computable representation, it might take until the end of the universe to compute, but it is mathematically well specified, then a mathematician can take that representation and either approximate the solution or come up with a different way of representing it that is more efficient to solve it, and this is the kind of thing that we need to be able to do. We need to have a unified theory for knowledge based systems.

We see a synthesis occurring between traditional artificial intelligence, which looked at structured representations for knowledge, and complex search algorithms, probability and decision theory, which can deal with uncertainty and can deal with objectives and values. Bayesian statistics can bring observations to bear as a theory of belief dynamics, of updating with evidence, and then database management, which taught us to separate the logical view of the data from the physical representation on the computer, methods of access update, administration security and shared data repositories. We can put all these things together into a theory of knowledge based systems.

I am going to talk now about a Bayesian network. Then I am going to talk about how Bayesian networks are evolving into graphical modeling language for specifying agent-based systems.

This is an example that has nothing to do with homeland security, but I have used it as a classroom example that illustrates the basic ideas.

Maria starts sneezing. This is a Bayesian network about Maria's sneezing problem. She might be sneezing because she is having an allergic reaction, or because she has a cold. She wants to go visit her grandmother -- this is the utility value -- but she cares about the health of her grandmother, and she cares about pleasure from the visit. If she has a cold, she doesn't want to go because she cares about the health of her grandmother.

This is called an influence diagram or decision graph. The probability part of it is called a Bayesian network. So she starts sneezing, and she draws a plausible inference, so her probability of having a cold went from about eight percent to about 50-50 from sneezing. So the utility shifts here, and she shouldn't go and see her grandmother.

But then she sees scratches on the furniture, which as we have already talked about is distributed propagation algorithms. It makes her think there is a cat nearby, which makes her think she is having an allergic reaction, which explains away the sneezing, so she doesn't need cold as an explanation anymore, so now it is okay to visit her grandmother again.

This is a trivial example. We have built much more complicated systems than this on much more interesting applications. But it shows you how you can use these Bayesian networks to represent commonsense knowledge.

Let's talk about what happened here when Maria was sneezing. The decision graph is both a knowledge representation and a computational architecture. So it formulates somebody's model in terms that are meaningful to a human being. We have worked with domain experts, and we can build a fairly complex domain model within a couple of days with an expert. The expert can play with a model and make it in a day or two.

I have had junior high kids playing with Bayesian networks, and they understand them and they like them. I can get students after one or two classes to build passable Bayesian networks. So it really is human understandable language for specifying very complex, hundreds of

variables, probability and decision models, and the output accords with human intuition.

If there is just the probability nodes, that is a probabilistic inference problem. You can also put decision nodes, and it is a decision theoretic optimization model. There are inference algorithms for computing the marginal distribution of one variable, given evidence on the other variables. So we learn the probability. We updated it, learned that Maria was sneezing and that changed our probability, and that was Bayes rule, and it was a distributed algorithm for applying Bayes rule.

There are also what they call learning algorithms, which is parameter estimation, where you use data sets to update the parameters. Most of the methods that are applied are getting increasingly sophisticated, but in the artificial intelligence field, most of the algorithms that are applied are fairly straightforward from a statistics perspective. They tend to be very simple models and independent identically distributed observations with no missing data, things like that. They are becoming more sophisticated now.

From a mathematician's perspective, there are a lot of fascinating problems there. In geometry, in inference, for example, if you take a Bayesian network and

you don't know the structure so you don't know what the arcs are, then you don't know what dimension your manifold is.

If you do these mixture models, where I have one probability of this model, one probability of that model, one probability of that model, there may be hundreds of thousands of them, you get something called a stratified exponential family, which is an interesting mathematical structure. It would be interesting to do some theory on the properties of those things and the estimators that you get.

What I just showed you was pretty much the state of the art up until about 1990, and is really exploring out now and being applied all over the place. There are hundreds of papers on Bayesian networks now coming out every year. But that was a one size fits all model, in the sense that the types of applications people do. You might have a clinic where you have thousands upon thousands of patients coming in, but you have exactly the same symptoms for every patient, exactly the same background variables, exactly the same diseases that you are worried about.

That is not good enough for this kind of problem. We don't know how many terrorists there are out there. There are lots of different stockyards that they might be

infecting. There are lots of different plans that they might have. There are just too many things to have one grand Bayesian network to cover them all.

There is also the temporal dimension which makes things much more complex.

Let me talk about Maria again. I am going to from the sublime to the ridiculous. Our first variation is a very trivial variation. We have Tran. He is sneezing and he saw scratches, but he was recently exposed to a cold. Unlike Maria, he is not allergy prone. That is very easy.

The second variation is, he is the one who saw the scratches, Maria didn't see the scratches, and he is in the room with her. That is a little bit more complex. The final variation I am going to talk about is that they are both sneezing, and this time they are both allergy prone, and we don't know about their cold history, and they both saw scratches, but they are a continent apart.

Those three variations are going to illustrate slightly more complex types of reasoning that come up in the homeland defense area.

Variation one I can do just by adding a background variable, that somebody is or is not allergy

prone. Maria is, and we don't know about Tran, so this one is not gray, and it is not either zero or one.

Then there is exposure to cold. We don't know when Maria and -- we do know it on Tran, these guys are mirror images of each other. It turns out that Maria should visit her grandmother and Tran shouldn't visit his. All we have done is added a few extra variables to cover all these situational factors. That is what people have been doing with Bayesian networks. You add variables that cover all the situational factors that you might want to model.

In variation two though, that doesn't quite work. We have repeated sub-structures. I have replaced cat nearby with Maria being near a cat and Tran being near a cat, and there is a cat here to the location of Maria and the location of Tran. Maria is near Tran, so we find out if Maria is near Tran, then they both have to be near the cat.

I've got two copies of the same model, and I have pasted them together like Legos. That is the direction that things are moving. Notice as a statistician, there is replication here for learning that this piece and that piece are the same, or close to the same or whatever. It

is an interesting statistical model and problem to learn these kinds of models.

This is what happens if we try to apply that model, but then we try to put them a continent apart, and have them both see scratches. What happens is, they are both 50-50 near a cat. That illustrates something that in the fusion literature is called the hypothesis management problem. We haven't enumerated enough cats to cover the situation, so we have the hypothesized entities.

Here I have all this spaghetti up here, where I have got two possible cats, one of which might be the cat that is near Maria and one of which might be the cat that is near Tran. It turns out that this model is essentially the same as variation one, variation two and variation three done right. In other words, I can use this model for everything, but I don't need it if I only have one cat, and I don't need it if Maria and Tran aren't related at all. So the question is, how much of the model do I construct and how much do I prune away. That is something that has got very interesting mathematical challenges.

I like to let my experts specify the model in conceptually meaningful units. So I've got my cats and allergies fragment, that talks about allergic reactions. I've got a spatial reasoning fragment, which is rather

rudimentary for this problem. This just says that if I am near you and you are near somebody else, then they are near me, too.

I've got my evolution of colds and time fragment, which is a very simple markup chain here. I've got my value fragment, which says I care about my grandmother's health and I care about my pleasure. I can paste these things together, and from a logician's standpoint this kind of model has first order expressive power, whereas the previous kind of model had propositional expressive power. This is the way things are moving in artificial intelligence.

A simpler model give the same results as the more complex model. From a mathematician's standpoint, we are building an infinite dimensional Bayesian network implicitly in our knowledge base. We need to think about the mathematical properties of these algorithms.

This is an architecture for a system that retrieves Bayesian network or decision graph fragments and pastes them together. I will just show you that there is this architecture.

Let's get to the challenges. This is a bunch of applications that I have been involved in with various students and colleagues. I can make these Vu-graphs

available to you. I can let you read through some of these, but I want to move to a few Vu-graphs on challenges, and I am running out of time, so let's go ahead here.

Model life cycle management as a systems engineer. We want to be building reusable models. So we need to have a library of fragments. We need to have something that the database people call meta data, different describes the model's capabilities, inputs and outputs, and models that pass information between them. From that perspective, the modeling language provides a universal representation language that describes generic decision and inference problems. We need software tools that are theory based.

Inference and learning technology is probably more meaty for mathematicians. Efficient solution methods, temporal reasoning is intractable. Decisions over time is even more intractable. Value of information as Tod mentioned is even more intractable. Value computation. Standard probability theory assumes that I know all the logical consequences of my beliefs, but I don't know the optimum computation before I have done it.

People have modeled -- Eric Corbett from Microsoft has done a lot of work in value of computation,

applying value of information to decide whether to perform a computation or not.

We need addual modeling. That is a buzz word that came from a DARPA program in addual modeling, the idea of being able to take these model pieces and break out when we get new problems that we haven't seen before. Some parts of the world don't change at all, and we want to leave those fixed, and we want to change only those parts that we need to change, and how do we structure out models so that they are easily changeable.

Multi-resolution modeling. I want to model at different resolutions because of computational efficiency or other factors. The really important thing is deception. If they know I am modeling them, how do I model the fact that they know I am modeling them. Game theory is relevant there.

Support for human-computer interaction. We want to combine expert knowledge with designed experiments and observational data. Most of the learning methods cannot handle anything more complex than independent, identically distributed observations and a conjugate prior. You have got to get better than that.

Human-computer interaction design for model input, model output, sensitivity and uncertainty analysis

and collaboration among heterogenous and geographically dispersed analysts.

I think that is the end. Summary, one more.

We are moving from hand-crafted special purpose models to reconfigurable model pieces that humans work on parts of the problem and then piece together with other parts of the problem. It is something that we really need theory to understand what is going on. That is the direction things are going. There is lots of useful application experience.

By the way, Clippy in your Microsoft office products is a Bayesian network, in case you didn't know.

DR. CHAYES: But the interface was not written by the Bayesian.

DR. LEVITT: Eric does not like the obnoxious little -- he says if they had used his decision theory on when to bother you, as well as the Bayesian stuff on what do you want to be doing, that people would like Clippy a lot better.

DR. CHAYES: Eric is very embarrassed by this. This is used as the example of his work, and he didn't have anything to do with the obnoxious part.

Knowledge Representation and Inference for Multisource Fusion

Kathryn Laskey

Excellent inference and decision support for homeland security requires supporting expert, human judgment with data; extracting key conclusions from volumes of heterogeneous data; and effective performance in the presence of uncertainty and ambiguity.

There are many recent advances in knowledge representation and inference technology, such as dynamic computing systems, whose solutions improve over time.

We need to have a unified theory for knowledge-based systems, and in that spirit it was noted that a synthesis is occurring between traditional artificial intelligence (which looks at structured representations of knowledge) and complex search algorithms involving probability and decision theory (which can deal with uncertainty, as well as with objectives and values). The main links between the two arenas are Bayesian statistics, which can bring observations to bear in updating evidence, and database management.

Dr. Laskey discussed Bayesian networks, which essentially are causal graphs associated with underlying probability distributions. She offered some examples to illustrate Bayesian networks' usefulness for representing common-sense knowledge and assisting in decision-making. The decision graph is both a knowledge representation and a computational architecture, so it formulates somebody's model in terms that are meaningful to a human being. It can be used for specifying very complex probability and decision models, with as many as hundreds of variables, and the output will still accord with human intuition.

Valen Johnson
"A Hierarchical Model for Estimating the Reliability of Complex Systems"

Transcript of Presentation
Summary of Presentation
Power Point Slides
Video Presentation

Valen Johnson is a professor of biostatistics at the University of Michigan. He received a Ph.D. in Statistics from the University of Chicago in 1989 and was a professor of statistics and decision sciences at Duke University prior to moving to Ann Arbor in the fall of 2002. He is a fellow of the American Statistical Association, past treasurer of the International Society of Bayesian Analysis, has served as an associate editor for the *Journal of the American Statistical Association* and *IEEE Transactions for Medical Imaging*, and is a Savage Award winner. He is coauthor of *Ordinal Data Modeling* with James Albert and author of *Grade Inflation: A Crisis in College Education*. His research interests include ordinal and rank data modeling, Bayesian image analysis, Bayesian reliability modeling, convergence diagnostics for Markov chain Monte Carlo algorithms, Bayesian goodness-of-fit diagnostics, and educational assessment.

DR. JOHNSON: The title of the talk is Estimating the Reliability of Complex Systems. This is joint work with Ty Graves and Mike Yamada at Los Alamos.

Before I get into the talk, I would like to acknowledge some of the people we have been working with. A lot of the work started as an application of the nuclear weapons program at Los Alamos. We were trying to model the reliability of nuclear weapons that can't be tested anymore. But we have also been applying some of these applications through some Air Force problems, the F-22 safety program. We have been working with the Army, and the assessment division of the engineering directorate has provided some funding for this. We are also trying to apply this type of methodology to the Ballistic Missile Defense -- or I guess, the Missile Defense Agency now.

Here is a non-classified example of what we are doing. This represents the system diagram for an anti-aircraft missile. This particular weapons system has 17 components. The fault diagram -- in this case, the way these fault diagrams work, this notion means that for the guided missile on this weapons system to work, all of the sub-systems down here work.

The idea here is that we are going to in general, and in particular for nuclear weapons, we are going to have

prior expert opinion about the probability that any one of these particular sub-systems works. The prior expert opinion can be very important in the nuclear weapons system and in the anti-aircraft missiles case, because you can't test or you can't get as much information from actual data as you would like. So we need to use expert opinion.

The expert opinion is going to come at different system levels. You may have a physicist who can give you information at the systems level. You might have an engineer who can give you information at a valve or something level. So we need to incorporate all these different sources of expert opinion. We also may have binomial data that has also been collected at different levels in the system. It may also be collected from different systems. So we want to incorporate that into estimation of system reliability.

We need to handle sparse data. I'll show you why that is important in a little bit. Then in the weapons program, we may have five or six different weapons systems that are all similar, and we want to model the similarity across these different systems.

My talk is going to be a little bit different from some of the previous talks, in that I am going to become somewhat specific in how we are modeling this

reliability. Just a little bit of notation. In some of the slides I am going to show in a minute, C_i is going to denote the component or the sub-component I have in the system. The p_i will denote the probability that that sub-component functions. These are the numbers that we are really interested in. We want to do inference on those.

When we have data for a system, x_i will denote the number of successful trials, and n_i's will denote the total number of trials. So different types of information we might have.

PARTICIPANT: (Comments off mike.)

DR. JOHNSON: That is the probability that that component functions. We are not going to know that, we are going to estimate that. Generally you don't know that.

DR. CHAYES: So it is the real problem.

DR. JOHNSON: Real, but unknown. The first sort of information is specific prior expert opinion concerning p_i's. So you go to an engineer and you ask him, what is the probability that this valve functions. The engineer in this case is going to tell me, the probability that that valve functions is A. So maybe it is .93.

We model this in a beta type density function, but we haven't fixed the K. So A is the maximum value of this beta density. We are going to model the precision of

this beta density. As K gets bigger, the beta density collapses around this estimate. We are going to see how consistent the expert is with the other data we have and with the other experts.

For specific prior information, let me just emphasize that the A is fixed, but we are going to estimate the K, and we are going to use the beta density to do that.

PARTICIPANT: (Comments off mike.)

DR. JOHNSON: We have flexibility on how we can do that, but we have been assuming that the K is the same for the same expert. So if an expert comes along and gives us guesses that are not consistent with the data, we down weight his guesses for all of the components, whether we have data on them or not.

Generic expert opinion, if we go back to the fault diagram, what we want to do is, we want to allow experts to come in and say, maybe this control assembly and this guide assembly have similar success probability or similar reliabilities, but we don't know what they are. He won't say it is .93, but he will say these two things are similar.

If we do a similar thing here, again we are going to use a beta density to estimate all the probabilities within a group of components that an expert has said should

be similar. The difference now is that we don't fix C. So we are going to have a precision parameter to say how tightly these components are grouped together, and we are going to have a C which we are going to also estimate from the data.

So we generally will put a Jeffreys prior on C and we'll put a gamma prior on M. But we are going to estimate both of those parameters.

Then finally what we do -- and we don't have data on all the nodes in this system. We can't for example compute a system-wide probability if we only have data on five or six of these terminal nodes. So we are going to assume another grouping type prior on the terminal nodes in our fault tree. So that is the final piece of the puzzle until we get to the data. Here we assume a Bayesian distribution as well.

So we take a hierarchical specification, a lot of priors on the terminal nodes. For all of the terminal nodes, those are the nodes on the tree that don't have any sub-components. We take this type of beta. We say we don't know what the parameters of the beta are, but we have a parameter B which we estimate, which is the mean reliability of all the terminal nodes. The M is going to

estimate how tightly grouped they are together, whether they are all precisely the same or very much different.

PARTICIPANT: (Comments off mike.)

DR. JOHNSON: What we do here, we say the terminal nodes have some probabilities. We don't know what the mean probability is. We assume that B is drawn. In the next stage of the hierarchy, we assume it is drawn from a beta density. Generally what we assume here is a Jeffreys prior. So C and D are taken to be one half.

PARTICIPANT: (Comments off mike.)

DR. JOHNSON: M is also a pre-primer.

PARTICIPANT: M is not drawn from -- M would be determined by --

DR. JOHNSON: No. We take a gamma prior typically for M. As Kathryn mentioned, this is sort of like a Bayesian network. A lot of these conjugate priors -- we use conjugate priors because they are convenient, but it is not necessary. The data terms of course are coming in as binomial observations. I'll mention at the end how we are extending those also to time-bearing cases for this anti-aircraft missile data.

Then finally, if you look in the reliability literature, you will find that there has traditionally been a consistency problem in these type of nets, because people

have introduced different probabilities in different levels in the system. Then they use different priors and combine data in ways to give you incoherent results.

We have avoided this problem in our model by using the fault tree specification to relate the parameters. So for example, for component four, I am not going to model the piece of four, because with the fault tree -- let me first say, if I had ten successes and one failure, then the binomial probably at p_4 would be this, except from the fault tree. When you look at p_4, p_4 only works if component 8 and component 9 work. So we just re-parameterized. We said p_4 is identically equal to p_8 times p_9. We are developing diagnostics to see when that actually holds as well.

That is sort of an important point, because if you have one system level test, all this information is transferring itself all the way up and down the fault tree now. So for example, if I have N trials at the system levels, the likelihood that comes from those N trials takes this form, where we have the product over the probabilities of all the terminal nodes minus that same probability. So now we have information about all of the terminal nodes from one system test.

The joint posterior distribution which is the thing that we are going to use to do inference on all these probabilities, is just obtained by multiplying the specific expert opinion, the generic expert opinion, the hierarchical specification on the terminal nodes and the data terms all together, so we are assuming essentially independence there. We get a joint probability density on all the nodes in our system. Ten years ago, that would have been a hard problem to address, but now of course with Monte Carlo algorithms, we can go through and sample these probabilities very easily. We can update all the precision parameters, all the probability parameters, and even the hierarchical means and the generic means.

Let me go through a quick example of how this works. We had some very unique data for testing this model. We had data from an anti-aircraft missile system, approximately 350 component level tests were performed on the system. I can't show you the actual data because it is proprietary.

But it also turns out that they have done 1400 system level tests. So I have gone through and I have done the analysis just using the 350 component level tests, and then also doing it using the 350 component level test plus

the 1400 system level test to see how the two results compare.

Here is what we ended up with. We do have the 1400 system level test for component one, where I have drawn X's through components where we had no component local data. You can see where we get rid of the entire weapon test.

There may be problems in predicting the missile round and various other components if you look through this. For example, component 8 and component 9 may not be identified because we just have the test up here. So it is a fairly complicated problem, and it is also representative of a lot of the nuclear weapons system problems that we have, where we don't have test data for certain components.

Here are the posterior distributions. For the reliability of the system as a whole -- and again, I apologize for not being able to display the actual axes here. These axes are between zero and one.

We are using the 1400 system level test and the component level test. We get a posterior distribution that looks like this black line, so that is the posterior distribution. When we throw away all of the system level test and just use the component level tests, we get the red line. As you expect, the posterior distribution based on

just the component level test is quite a bit broader than it is using system level data, but it is reassuring, in that it seems to cover the posterior distribution that we would have gotten using all of the system level tests. So we were heartened by this.

We had expected a priori that when we did this analysis, the posterior default system would be down here somewhere, just because of assembly problems, the fact that the components might not interact together. We had anticipated having to put an assembly node into our default diagram to account for that. But in this case, the data just didn't support that.

We tried this to see what the effects were on components where we have a lot of data. For one of the sub-systems that had a lot of test data, the posterior didn't change much when we got system level tests and when we incorporated that information, whereas we were looking for components where there was no data, the posteriors did change somewhat, because the system level tests did provide that additional information.

DR. BORGS: (Comments off mike.)

DR. JOHNSON: That is a good question, and we wondered why that was happening. It turns out, when we talked to the people who provided us with this data, they

had made improvements in the manufacture of the weapons system over time. If you actually go through and do logistic regression just on a component level, where you have the time for each of these things.

Did I answer that question? I think I answered a different question.

Some extensions that we are currently modeling. Some missile test data that we had, we had a time associated with all the binomial data, so we are modeling now -- instead of just having a beta distribution at each node where we have data, we are looking at a logistic regression model.

We are incorporating data in other applications from computational physics models, where they run computer code, for example, to predict outputs of systems. We have to incorporate the uncertainties in the input parameters to these codes into this. Material degradation models, and then different types of engineering data.

I think an important point to make is also that the model provides guidance where additional information can best be obtained. So we can look at the posterior distributions and different components and see where we would like to get more data if we wanted to better estimate the reliability of the system as a whole.

So I'll stop there. There was a question that you asked.

DR. BORGS: (Comments off mike.)

DR. JOHNSON: That issue is involved in something we are working on actively. That is fitting these time varying models. It turns out the component level tests were done in a different time than the system level tests. So there was a time effect on these curves.

But I think the more appropriate answer here is just to say that there is uncertainty in the estimate of these probabilities. Based on just component level tests in general, we are going to have a broader distribution for the system, and when we get the system level information, it is not always going to fall right in the middle of that distribution.

There are certain diagnostics we would like to do, for example, omitting some of the data in the nodes that we have data, and see if we predict well what that data would have done. But the assembly error in putting these components together is an important thing to be considered, as is the independence of the expert opinions from each other and from data that they have already seen.

PARTICIPANT: I have got a practical question that is less related to the --

DR. LEVIS: Will you please go to the microphone?

DR. FREEDMAN: (Comments off mike.) Are there other experiences in our society or in the world which are -- that is, something was tested for some period of time and then for a long period of time thereafter, we didn't really test them. We had to rely in some way on expert opinion, that we then model and put together to try to infer how a system worked. My apprehension is that the farther away we get from actually testing something, the more we are deluded into believing that the aggregation of expert opinion and wonderful methodology still speaks truth. What is the empirical basis to give us confidence that this speaks truth about how a system will perform?

DR. JOHNSON: I think your question is basically, if the test data and the expert opinion was all collected before 1992 when we stopped doing nuclear tests, do we really want to use it now. And of course, at the lab that is a major focus of investigation, what is happening with these weapons systems as they age and they begin to exceed their intended lifetime.

DR. CHAYES: He is saying that the test data itself is --

DR. JOHNSON: That is what I am trying to answer. There is a major effort in trying to tie the output from

computer codes, for example, that have been designed to run very accurate simulations of what happens in a nuclear reaction, to tie that back to the test data, as well as to tie it to any other data that they can collect now. So some critical experiments where they may do hydro tests.

It is an extremely important problem, and people are not naive in thinking that they can just simply use data collected 15 years ago and put this into the model and think it is going to work. But they are trying to use the data collected 15 years ago to validate their codes.

So in the last slide where I said we are trying to incorporate different sources of information into this model, this binomial system, we are really trying to look at some fairly sophisticated ways of putting information into these different nodes that isn't really in a conjugated form.

A Heirarchical Model for Estimating the Reliability of Complex Systems

Valen Johnson

Engineering problems may involve estimating the reliability of a system that cannot be tested. For example, at Los Alamos, efforts are under way to model the reliability of nuclear weapons. This issue also arises in handling Air Force projects, such as the F-22 safety program, and working with the Army and the Missile Defense Agency.

Consider a multicomponent system in which every component must work in order for the system to work. In order to estimate the probability that any one of these particular subsystems will work, as in the case of nuclear-weapons systems, prior expert opinion can be very important. This is because we can't test, or we can't get as much information from actual data as we would like, so we *need* to use expert opinion.

Dr. Johnson described in detail his method for using different sources of expert opinion in the estimation of system reliability.

Arthur Dempster
"Remarks on Data Integration and Fusion"

Transcript of Presentation
Summary of Presentation
Video Presentation

Arthur Dempster is a professor of theoretical statistics at Harvard University. His research interests include the methodology and logic of applied statistics; computational aspects of Bayesian and belief function inference; modeling and analysis of dynamic processes; statistical analysis of medical, social, and physical phenomena.

DR. DEMPSTER: That's okay. I notice that the talks have been getting shorter and shorter. I think it has something to do with Saturday afternoon. The schedule says I have five minutes, then I pass it on to Alberto. I don't know what he is going to do.

I have a few topics I wanted to touch on very briefly that I planned in advance. Then I can come back perhaps to the three talks, again very briefly.

First of all, on defense, I am completely non-expert in this area. I have heard lots of wonderful things, beginning with our chairman. I am glad to hear that there are positive activities as well as possible ones, seeking out and pre-empting possible dangers and thinking about strategies for doing this.

A big theme is of course modeling the whole complex system and all its actors and characteristics and environments and dynamics and so on and so forth. I perhaps would have liked to hear a little bit more about gaming, the opposite to people are developing strategies, and we have to cope with their strategies and back and forth, multiple feedbacks and so on. Anyway, it goes much beyond passive description of the situation.

One thing about complexity. Jimmy Savage used to say, make your models as big as elephants. Another thing

that I used to hear, Bill Cochran was my colleague in the late '50s, the late '70s, would quote Ari Fischer with the advice to make your causal thinking complicated, bring in all the causal mechanisms before you try to understand what is going on.

Obviously, the last question occurred to me when I listened to it, the idea of, we are well past the period of testing. There are all kinds of dynamic changes that have gone on in the active weapons and the newly constructed ones and so on. One would have to think of all the causal mechanisms that were operating there and build them into the mathematical models as they get more and more dependent on scientific expertise and such things. So that is one set of comment on defense.

A second comment that isn't perhaps very much on today's topic, but I'll back up a little bit -- I think discussants have been allowed to do this by the chairs -- that is to say that data are dumb. What I mean by that is, in themselves, if they are just a string of bits, that means nothing at all and it is obvious.

But even if you have lots of structures that are set up wonderfully from the computer science point of view for access and information flow and manipulation, that is not getting at what is really going on here. So the

necessary complement is scientific understanding and meaning.

This aspect I think was brought up by George Papanicoulou in the seismic problem, where the databases of course are huge. But the real understanding came about through working through the science and improving that. So in all these homeland security models, I think the science is somehow primary, and we should always keep that in mind.

This session has been on fusion, and that is what I should try to address a little bit. We are talking about fusing different measurement sources that are measuring objective things with error. That is the kind of model that is important to develop and to try to pool.

The other aspect of it though is that we need to be fusing the models that refer to measurement, the models of the underlying phenomenon, the models that represent the science. That is an even bigger task. I think again, we should be deliberately focusing on those kinds of issues.

As to mechanisms for fusion, the way I look at it and am conditioned to look at it is that we have had 250 years now since Bayes presented the formal tool of conditional probability. It was 200 years ago roughly that Gauss deliberately used that tool in order to propose Lee's squares for models with formal errors.

Interestingly, it has been 150 years roughly since Bool set out the basic ideas of propositional logic, which is essentially a formal fusion tool of deterministic information. About 35 years ago, I presented what seemed to me to be the natural combination of these two ideas in a single mechanism that seemed to me to be totally natural, and then 25 years ago Ben Schaeffer gave a new name theory of belief functions in his Theory of Evidence book to that conception.

For me, it has always been just that, the fusion of Boolean logic with probability, so it is not different from probability. But one does have to emphasize the word logic here, logic in a broad sense of trying to reason. I do want to emphasize that I think of it as a set of tools for reasoning about a scientific phenomena that include deterministic models and uncertainty and probabilistic uncertainty and so on.

One comment about it though is that this fusion idea rests very heavily on a concept of independence. So the different information sources of evidential sources do have to have within them some representation that is independent from source to source before either Boolean logic or probabilistic combination or whatever you are doing, Bayesian combination, can operate.

If we take a specific example like Val Johnson's, one should be thinking carefully about whether the evidence that comes in at higher levels in the tree are independent of one another. He and I have discussed this. He is well aware of course of that kind of issue. So that is a technical thing that should be of great concern.

I'll say a little bit about graphical structures. It seems to me that this technology, certainly the one I am familiar with, the belief function technology, has been closely tied to join trees, tree structures, that kind of thing, since the mid-80s when Schaeffer and Schnoy wrote a paper about it, and my student, Augustine Kahm, did a thesis in 1986 developing that subject. That is not different from the better known Bayesian network theory.

In fact, it subsumes it. In some ways, these conditional independence ideas go back to Phil David's '79 paper. I regard Judea Pearl as a good friend and very stimulating, and certainly has done a lot to move this theory out into the computer science world. I am a little surprised that Kathy ascribes it to him, since some of us have known about it rather longer than that.

A little bit about mathematics. I don't profess to be a mathematician, although I do have a math stat degree from the Princeton math department way back. I like

mathematics, and I especially like beautiful mathematics, which I think is part of its great power.

I don't have a good sense that mathematical thinking has come to grips with computing and the data revolution and all those complexities and so on. Maybe it is just the reductive aspect of the way mathematicians work, but to me, what I would like to see mathematics do is have new conceptions or frameworks for addressing scientific issues. I include all of these homeland security issues as scientific issues, in the sense that they need to be described, and what you know about them set up and formalized as much as possible and so on.

Anyway, there is certainly plenty to do for understanding computations. You can set up a mathematical model and define an algorithm that will do it by brute force, but one big thing mathematics can do is provide clever ways around the difficulties and gain you orders of magnitude of speed.

This MCNC approach that I am familiar with and that Val has used in his example is another area where there is a tremendous amount of mathematical development that can go on. Mathematicians in mathematical statistics have emphasized properties and procedures, and there is continuing to be a lot of new work going on there, very

interesting, very important, as to the efficiency of methods not only of calculation, but also of scientific methods, issues of robustness and all that kind of thing.

So perhaps I should turn a little to the speakers. I think Alexander gave a very nice outline of one set of ways of thinking about these things, and Tod and Kathy the same. I find that they are presenting very complex stories. I didn't have them in advance. Maybe that was a blessing.

One reaction I have, I would have to make a mathematical model of each of their presentations, and then I could analyze it and draw some conclusions about what I thought. But at the moment, I am left with trying to react to a very few different things.

I certainly thought Tod's talk was extremely interesting. He threw out all these different approaches to uncertainty. I'm not so sure that what Lad Visati has been trying to do and what I am interested in doing are not miscible. I think they may be miscible in certain ways.

One kind of thing that I am thinking about these days is a mathematical model of object recognition, very abstract, away from the real problem or what features, the complexities or dynamics or this kind of thing. But there was an example there of a medium characteristic. This

object I am looking at might have a membership value of .9 for being medium. I can treat that as a simple support function in the Dempster-Schaeffer theory, and I can discuss with Lad Visati, as I have done and will do more of in the next few months, exactly what the interactions are, since he is more interested in getting probability mixed in with his technologies. So there is some hope I think of some of these different approaches perhaps coming more together.

I did have the advantage of being at Los Alamos last week and talking to Tod Graves and Val Johnson on the paper, and I had a chance to look at it. I think of it as a very nice case study type thing that can scale up to very complex examples. It does make use of the graphical model formulation in a very beautiful and non-trivial way.

It brings in several different kinds of modeling. One is the traditional binomial sampling model. What I found when I thought about it was that the Bayesian aspect -- it has got a Bayesian label on it, but what the Bayesian would do, the standard Bayesian, would be to take the samples, the binomial samples and the 13 unknown probabilities and say, how am I going to assign a prior to those 13 probabilities that I can mix with the data. But those authors didn't do that. They put in prior

information here and there, and I regard that as very much in the spirit of the belief function approach to things. So I don't think of it as belonging to Bayes in the classical sense. It is much more moving towards the belief function approach.

I thought the way they used beta priors with those capital K and capital N values in them was quite ingenious, although it seemed to me to mix a little bit the notion of expert opinion, the U of I.J. Good, the expert as a set of data. Whereas, the other aspect of it, the hierarchical thing, was just shrinking towards a common mean, that aspect of it gets buried in the beta priors and it was a little hard to see, unless you are very familiar with what is going on there. I misinterpreted it the first time I read the paper.

That assumption, that the true values of P are drawn from some population, that is a judgment, not an expert opinion, not a sampling of likelihood term, but it is a kind of judgment on the part of the analyst who is constructing the model, to treat these things at that level as exchangeable. That is a judgment by what I.J. Good called "U", in quotes, or Jimmy Savage someplace else called thou.

So there is a kind of a philosophical attitude going on here, whereby there is objective science, something that I think Glen Schaeffer called nature, nature as the source of causation in the book, The Art of Causal Conjecture, that Tod mentioned. There is objective science of that nature. There is expert opinion, which is almost like data, and there is the U or the thou that makes some judgment that puts it all together into a model that you can have some belief in. So you need to have some kind of a working philosophy there, which I think very few people have.

Perhaps one of the reasons that what I regard as this very simple combination of probability and Boolean logic, Bayes and Bool, is understood is because what it is really talking about is this U performing some kind of logic. If people started to think that way, they would understand this technology a little better.

Remarks on Data Integration and Fusion

Arthur Dempster

Dr. Dempster spoke on a number of related topics relevant to homeland security.

Game theory. In order to devise an effective method for seeking out and preempting possible dangers, perhaps there should be more focus on game theory. We have to cope with the attacker's strategies, and then back and forth with multiple feedbacks. It goes much beyond passive description of the situation.

The primacy of science. Data by themselves are just strings of bits that mean nothing at all without the necessary complement of scientific understanding and meaning. The science is primary, and we should always keep that in mind.

Fusion. Dr. Dempster spoke of the need to fuse together measurement sources that measure objective quantities with some amount of error. In addition, there was a need to develop ways of fusing together the models that refer to measurement with the models of the underlying phenomena—the models that represent the science. That is an even bigger task, but one researchers should be deliberately focusing on.

Mathematics. "I don't have a good sense that mathematical thinking has come to grips with computing and the data revolution," Dr. Dempster said. "Maybe it's just the reductive aspect of the way mathematicians work, but what I would like to see mathematics do is have new conceptions or frameworks for addressing scientific issues—including homeland security issues." Although one can set up a model and define an algorithm that will do a computation problem by brute force, one big thing mathematics can do is provide clever ways around the difficulties and gain you orders of magnitude of speed.

Alberto Grunbaum
”Remarks on Data Integration and Fusion”

Transcript of Presentation
Summary of Presentation
Video Presentation

Alberto Grunbaum is a professor of mathematics at the University of California, Berkeley. His research interests include analysis, probability, integrable systems, and medical imaging. Dr. Grunbaum’s current research is in medical imaging. He is studying the use of an infrared laser in place of X-rays.

DR. GRUNBAUM: Being the last speaker in this fantastic couple of days is a mixed thing. On the one hand, you are going to remember everything that I say the very best, because there is nobody speaking after me. On the other hand, the real question you have is, can you get me a taxi to go to the airport. So I am not sure.

I really enjoyed this tremendously. I wouldn't claim to know anything about information fusion, even before or after the great talks. But I am very, very interested in the issue that I think brought some of these people together here, coming from so many different parts. The issue is, is there any way of bringing mathematical types in general, that includes all sorts of different things, to think about this new unfortunately national effort of homeland defense.

I share fully what George mentioned yesterday and Dave McLaughlin and some other people, that it is very important to keep in mind the physics, the biology, the chemistry, the computational background that are part of these problems. Even when we formulate them as mathematical problems, that should be a very, very important element all the time.

On the other hand, mathematics has this amazing, almost poetic ability of looking at a problem in a certain

context and then taking a step back, playing with data, and all of a sudden turning around into some other problem and saying, by the way, maybe I can use these tools in this different area. That is the feeling that I have. I will be pushing some of the community that I think I represent now, which is the community of inverse problems.

There are a number of people that have worked in medical imaging, in geophysics. Those are two good success stories of uses of mathematics in areas where, if you go back say 20, 25, 30 years ago, nobody would believe that math could contribute at all. To convince a medical doctor 30 years ago that mathematics could be useful you would be laughed out of any such medical meeting. Things have changed a lot.

Now, as we move in this poetic fashion from one publication to a different one, I think it is very important to remind ourselves that we have to go back and start from scratch and talk to the experts in this new field, because the problem may look the same from the mathematical point of view, but they have all very different features.

So I was actually very, very pleased when I heard our chairman talking about inverse problems, because I had prepared a slide about that ahead of time. What I will do

-- I'm sure you all want to stay here until 9 p.m., but I want to get out before that, so what I will do is use the principle of the hammer and the nail and talk a little bit about a problem that I have been playing with for awhile, and hopefully the audience will tell me if it has any meaning whatsoever in terms of looking at a very complicated network and solving what from my point of view is actually an inverse problem.

I already mentioned that I could have picked another application, but this is one that I am more familiar with. A number of pieces have been mentioned in these past two days that have appeared from the very beginning in the area of medical imaging, the question of high dimensionality being the main one.

Back in the '70s, maybe even earlier than that, we finally realized something that by now everyone realizes is real, but it wasn't real at all back then. If you have some function of two variables, we think about the information density of human tissue when you expose them to X-rays, if you measure all the line integrals and put in sources and detectors in the regimen that I have here, you have an expert source here, and measuring over there, you realize that they are not the same. So there is a model that tells you -- that you can use to change the intensity

here to an intensity there, is the integral of this unknown function along the line A. Do that, and then eventually by doing this with all different lines, you recover the function.

Nowadays this is part of the culture. An element that was mentioned yesterday in the very first session has to do with the training of analysts at NSA. The issue of training radiologists is actually very similar; people look at the same set of pictures and draw very different conclusions. So this is just an example that some of these things -- what is my thesis? I am trying to say, if you are going around trying to look for a bunch of mathematicians that have already looked at some of these issues.

I have been fooling around for the last few years, trying to use laser instead of X-rays. Now you have to be able to perform a discovery. The phenomenon is that instead of having to begin with linear equations, from the very beginning, the thesis is so complicated -- again, you cannot ignore the thesis that determines everything. The source is one, that is one to go, that is applied here repeatedly, and you put goggles to be able to see any of this, you are going to see light coming through your hand, except on the other side.

So I want to take this as a study problem. I want to make propaganda for this journal published in England. I do this because I happen to be the editor right now. We are finding that a collection, a whole community of people coming from many different areas, some of them even call themselves mathematicians, that have dealt with a number of issues that are very, very -- the way somebody was talking about at what level should we move in matters of analysis. This is a very classical problem in medical imaging; should we go for higher spatial resolution or quantum resolution. The truth is that we would like to do both of them at the same time.

In a very rare event, there is a difference in attenuation, in the order of one and two percent. At the same time you want to go to reconstruction, you have fantastic spatial resolution and you want to be able to use it in an effective way. Some of you may have seen a movie called the Brain Attack Team. Nowadays if you suffer a stroke, they actually into your brain and they clean you up. That requires very good time and spatial -- what I want to do in the last slide, let's start from the beginning.

Here is a problem that I want to share with you in the hope that this may be of some use in some of these

extremely complicated networks. What I have drawn here is supposed to be a Markov chain with finite states base. The states are of three different kinds, some of them in code. S is one. Demographically this would mean these are the source positions. There are all these states that are inside, there is an arrow connecting two states. That means that you can make in one unit of time a transition. Then finally there are some detective states that detect those sites.

What is it that I actually consider here? I imagine that I have the seven one-step transition probability. I didn't want to clutter this any further. This probability which is not known, which is -- then in the next state I keep on going, wandering around. I can only make boundary measures. So in the columns that we had in the very first transparency, we determine what are the new ones, the good ones, the red ones, or orange, on the other side. Think of the source positions as the new ones. There is nothing that you can look inside. Not assuming the sum of the probabilities from one age to all the -- that is a way of modeling absorption.

The fact is, one can solve the forward problem completely, which is not entirely trivial, but that can be done. What I have managed to do in some very simple cases

is describe the complete class of all solutions for the inverse problem. If we don't make any assumptions -- of course, everything depends on the number of things that you have, but even the simplest cases are complicated like this, actually a little bit more complicated than this one, is data source positions, vector ones. In that case there are 64 parameters, measuring the probability from going from every incoming to every outgoing and measuring the time of light, you cans love analytically in terms of explicit formulas, for 56 out of 64.

So I want to stop here. Thank you for the fantastic two days. If somebody would tell me, I have a problem that looks a little like that, please do it soon. That's all.

DR. LEVIS: Thank you very much. Could you join us here at the table? I believe the schedule is now to open the discussion. But I first thank the speakers and the discussers for the last session.

There are two kinds of discussions, but my suggestion is that we merge the two, both specific and general discussion, right now, if that is agreeable. So please come to the microphone, identify yourself. I am working with a handicap; those two things are blinding me,

so I can't even see you. So please come to the microphone right there.

PARTICIPANT: I had a general question about your introduction in the first talk. We saw the diagram for the possible threats to the United States and the five questions to which we knew none of the answers. It makes it sound like homeland defense is not a really tenable proposition. Is defense the right strategic paradigm? After all, the number of people who seem to be committed to doing these things is very small. You are not talking about millions of people, perhaps a few thousands or a few tens of thousands. Maybe the whole idea of trying to wall the country in is not strategically the right approach, and one should look for instance at the possibilities of an offense.

DR. LEVIS: I will answer diplomatically on that one. I think I already gave the answer. Many people cal this homeland security, as opposed to homeland defense. I said, there are different meanings to the word defense and security. But we are doing all of the above as far as I know, but I don't know everything that we do.

PARTICIPANT: Certainly the emphasis in the talks that we had today was very much passive waiting, lying

around and waiting to be hit next, and trying to prevent attacks.

I'm not deprecating any of that. But I think looking at the diagram that the first speaker showed us, an unimaginable number of threats coming from an unimaginable number of places, and looking at all the vulnerabilities of a complex society, I have to say, didn't look very promising. I just wanted to posit that as a possibility, that that is maybe just the wrong strategic thing about it.

DR. LEVIS: I will give a very brief response and then refer it to the other speakers or members of the audience.

To begin with, and this is from limited knowledge, I am not an academic in real life, we are doing very well, actually. If you consider how many things have not happened -- this is a issue of metrics.

Let me put it in a different way. We have the police. The real metric for the police, if I am happy with my police department, is not how many people they arrested, it is how many crimes they prevented by their mere existence. When you go to the arrest, the crime has already occurred. But it is very difficult to get measures, as you understand very well about this, about the things that have not happened.

So the problem is very complex, it is very difficult, it is a large country. There is no way that you can border it totally, disallow anybody to enter it and so forth, the way of life that has been discussed. We have a huge intelligence establishment for which we pay a lot of taxpayers' dollars. Part of their job is to do the anticipatory and not even suggest, just bring the data up front for the appropriate decision makers to decide whether to be proactive or not.

I am speaking as an individual, but I don't think we are doing that badly, when you consider how many real instances of terrorism we have had in this country. It is the most open country in the world that I know.

PARTICIPANT: I wasn't suggesting that we have done badly. It was just a concern of mine towards this sort of strategy around a lot of what we heard today. There was a sense that we could not --

PARTICIPANT: You are not by the microphone.

PARTICIPANT: There was just a very defensive kind of mindset, and I was just concerned that this feeling of vulnerability could itself -- I was just concerned about the strategy part.

DR. DEMPSTER: It seems to me to be the new thing that there are events of extremely low probability that

have extremely high costs. That is what maybe we should be focusing on a bit.

DR. LEVITT: Alex mentioned something in his opening. He said, I could tell you, but then I'd have to kill you. I think the problem with the offensive stuff is that it tends to be handled by the intelligence community. I think there is an enormous amount going on, but I don't know much of it, and therefore I wouldn't know, as far as germane to this workshop goes, where the mathematical challenges lay in offensive activity, which I think are highly focused.

However, I completely agree with the scope of homeland defense. There is the newspaper version, or what I consider the public political version. On the other hand, where I see the real payoff is that there is enormous amount, and there has been for man years, of data in automation, collection, and critical things going on, that really can be exploited. I think that we don't know what the limits of automation might be, we don't have to have people, you are talking PCs and memory. If we could deal with the inferential issues, so that we didn't generate false alarms all the time, that there might be a tremendous payoff.

DR. LASKEY: I wanted to comment on that, and say that I agree thoroughly with the points you made about not sitting around and just passively observing. In fact, I'm sorry if my talk came across that way, because it wasn't how it was intended. Inference can be used to passively observe, but you notice that level five of the five levels of fusion that I talked about was placing your sensors. Part of what you are doing is, you are not just passively observing, but you are watching and anticipating what is going to happen and preparing yourself to respond to it, but also at the same time, if we discover that if there is an Al-Queda cell that is planning something, I'm sure that our military people would go in and stop them before they had the chance to do anything about it.

But it is not just waiting until the horse is gone to close the barn door. The technologies that I was talking about would apply much more broadly.

DR. LEVIS: Other comments? Questions to the speakers? The timetable for the airplanes?

I'd like to make one comment and a question, and then we can adjourn, if you like. Earlier today there was a question about funding. This is Washington, so I would like to address it very briefly.

For mathematics, the Air Force has the Office of Scientific Research. There is the Mathematics and Computer Science Division within it. I think one person was present here today. They are looking into those problems, and it is basic research.

However, things have changed a little bit, and you have to write more than one paragraph or one opening sentence. The old joke, everything looks like a worm. You have to write a little more and show some understanding of how that work will have an impact.

Of course, good mathematics has a good impact, but that is not sufficient. There is a lot of scrutiny for relevance. So understanding the problem makes also a much better proposal, even though you are doing the mathematics that you want.

People mentioned DARPA. I have known DARPA for many, many years. For DARPA you have to know the problem. There is a very small part of the DARPA budget that is in basic research. Most of it is in applied research and demonstrations. In the process of using those large three to five year demonstrations, a lot of research is being done, and I am sure some of you have participated. But those programs are fairly focused in trying to solve a particular problem and make a demonstration. So in order

to play, you don't have to know every aspect of the problem. They are usually large themes, and you spend an awful lot of time going to meetings and coordinating on those things.

But there is a substantial amount of funding that allow the basic research to be done in the context of the applied research. But you have to know -- the quick kill three program is, if you are going to have a new algorithm that is going to do it faster, cheaper, better and all the other things. There is a substantial amount of funding.

The Defense Information Systems Agency together with DARPA, they have a joint program office called JPO. That office has kicked off a major homeland security initiative. It is not exactly new money yet, there will be new money in the future, but right now they are earmarking existing money everywhere that could be pointed in that direction.

I have been through that exercise, I write a lot of proposals. I can see how we can take some of the work that we are doing, dress it up appropriately, and look at that problem. But in order to get funding for that, I will have to be credible that at least I understand some aspects that are relevant to homeland security, not leave it to the reviewer to make the connection. I have seen enough from

the inside, the paperwork, there is somebody that will have to write a paragraph that is that long on a form, explaining why your work is relevant to the problem. So you are passing the buck to somebody else. When you are doing that, things don't work very well. It is much better if you write it yourself, if I write it myself and submit it and discuss it. It helps the proposal along.

But there is funding that is developing within the various -- I haven't checked with the Office of Naval Research, but I am sure they are looking at that problem in the mathematics section, but one has to be a little more closely aligned to the problem than was discussed in here, at a fairly abstract and high level.

Thank you. I don't have any money, by the way. Whoever would like to close --

DR. KELLER-McNULTY: I guess I was voted in on this. I am the only other board member besides you two here.

I just want to first of all thank everybody for coming to the workshop. I think it has been pretty exciting. I know that I have learned a lot, and I have tried to listen carefully.

I think that there are clearly great challenges mathematically for us to try to address that can be very

supportive to homeland security. I also think that there is a fair amount of what we already do that can almost be directly applied to some of the immediate problems, and it behooves us to seek out those opportunities and try to do that.

We heard a lot of things in the last couple of days. Some of the things that resonated with me were the following. First of all, we have to keep the human in the loop. No one is talking about -- and some of our colleagues when we talk to them, some of our scientific colleagues, they get really nervous that we are trying to develop methods that somehow take them out of the decision making loop. I think it was pretty clear from everything we heard that the human has to be in the loop, the science has got to be incorporated into the problem, the domain knowledge has got to be included in the solutions and in the frameworks we build to solve these problems.

Which then of course points to one of the very first things that Alex said when he opened the final session, which was that you have got to know the problem. Clearly, if we want to get collective funding to work in this area, we had better darn well be willing to admit to knowing that we are trying to solve a real problem, and try to figure out how to pull the pieces together.

The other thing that came out really clear to me is -- and something that we sometimes forget to talk about when we are talking to people about the problems in the work we are doing, is that quantification underlies almost everything we are doing here. We tend to forget about it, because it is part of our being as mathematical scientists and as computer scientists. But everything is about uncertain quantification.

The question that came up about the nuclear weapons certification, it is not just about predicting a point value, whether or not we think these systems will work. It is all about uncertainty quantification and knowing when we have to go back and start testing again, or collect other types of information. That is true throughout all the homeland security problems.

A lot of people talked about how we are in a stage and in an era of being swamped with data, we just have so much of it. What was it that Art said? That data is dumb, which is true. Kathy commented that we have all this data, data, data, we don't have time to think. That is true on a certain level, but as soon as we ratchet these problems up to the incredible dimensions that they are -- and Tod showed us what the dimensionality of the space is, there actually is a lot of data sparsity.

In the last session, I think that Val's talk tried to show how we can try to put things together in the presence of a lot of data when you take some slices through the problem, but not a lot of data or information when you take other slices through that problem space. So that is really important that we remember as well, that it is not just about massive data mining, but it is about how to integrate all of this information.

The final thing I want to say is that what we are talking about is modeling complex systems. Perhaps we do have to put hats on a bit like engineers and take a systems approach to looking at these problems and figuring out how we can build and develop the mathematical frameworks to actually make progress.

I don't think the problems are impossible. I actually disagree with one of the last comments that was made from the floor, that a lot of what I heard at this workshop is about prediction and forecasting, how do we take what we have learned, what we are seeing, and project ahead and forecast ahead, to try to understand what the next thing is that needs to be done or where the next vulnerability is.

These are not easy problems. We have to come together and work on them. I hope that we do. I have been

pretty energized by this, and I look forward to continuing to interact with many of you on these.

That's it. Safe travel.

Remarks on Data Integration and Fusion

Alberto Grunbaum

Mathematics has anmazing, almost poetic, ability to look at a problem in a certain area and then take a step back and suddenly realize that maybe we can use these tools for some other problem in a different area. For example, there are a number of people who have worked first in geophysics and then in medical imaging!

He cautioned, however, that as he and his colleagues moved in this poetic fashion from one area of application to a different one, it was very important to remind themselves that they had to go back and start from scratch and talk to the experts in this new field. The problems might look the same from the mathematical point of view, but they have all very different features.

This article can be viewed online at
http://www.businessweek.com/technology/content/apr2002/tc20020430_3678.htm

APRIL 30, 2002

By Stephen H. Wildstrom

Enlisting Math to Defend the Homeland

America's number whizzes say their science has a key role to play -- and they just met in D.C. to show their stuff

At first glance, the notion of a conference on "The Mathematical Sciences' Role in Homeland Security" looks like a shameless attempt to channel some of the mighty stream of security funding that has been rolling out of Washington since September. 11. "When I got the e-mail invitation," says Howard Schmidt, vice-chair of the President's Critical Infrastructure Protection Board, "I thought at first it was a joke."

No joke. In fact, the Apr. 26-27 workshop sponsored by the National Research Council's Board on Mathematical Sciences & Their Applications was deadly serious. It turns out that many of the diffuse and complex problems of homeland security are deeply mathematical in nature, and even some of the science's most abstruse branches, such as topology and high-dimension geometry, can be brought to bear on security problems.

The most obvious application is the encryption techniques used to protect data from prying eyes. But encryption issues are so well known that they went largely unmentioned. Even crypto specialist David Wagner of the University of California at Berkeley devoted much of his presentation to other topics, from the mathematics of power-grid reliability to the design of "inherently self-stable systems."

DATA EXTREMES. Not surprisingly, much of the discussion focused on the use of advanced statistical techniques to deal with two almost opposite problems. First, the growing use of cameras and other surveillance techniques is overwhelming analysts with more data than they can hope to make use of. Mathematicians can help by developing data-mining techniques that help spot patterns in an ocean of seemingly random information.

At the other extreme, epidemiologists chasing, say, an outbreak of anthrax must figure out whether they're dealing with a terrorist attack or a random, natural event based on extremely scanty evidence. A situation that presents a large number of variables and a small number of data points is very poorly handled by traditional statistical analysis.

Michael H. Freedman of Microsoft Research pointed out that techniques developed in the field of high-dimensional geometry are relevant here. Geometers have found ways that a space with a large number of dimensions can be approximated using a much smaller dimensional field of numbers that are far easier to work with. In statistics, this is

generally analogous to reducing the number of variables.

WORKING BACKWARDS. Freedman, a winner of the most prestigious prize in mathematics, the Fields Medal, added a touch of levity to an otherwise serious session by describing "the general world view of mathematicians." On his way from Seattle to Washington, he was selected for a secondary search at SeaTac airport. He pulled out his itinerary and said he was on his way to a conference on mathematics and national security. "The guard was very skeptical," Freedman said. "She asked, 'Are you a mathematician?' I said 'yes." She replied, 'Then God help us.'"

Alexander H. Levis, chief scientist for the Air Force, suggested that mathematicians might find ways to apply to domestic security the statistical techniques that the military has developed for analyzing threats. One, using an approach called a Bayesian inference network, works backwards from a set of possible events to assign probabilities to the potential actions that opponents might take.

In the end, mathematicians don't suffer illusions that math alone is going to make the nation significantly more secure. Workshop organizer Jennifer Chayes, director of Microsoft Research's Theory Group, says she chose the topic for the annual workshop simply because it seemed natural and relevant. For example, Freedman's research specialty, the mathematics of quantum computing, could one day enable the solution of problems that today remain dauntingly complex.

Practical applications remain, at best, years away. "I don't think," Freedman says, "that we can stop terrorism in time by building quantum computers." All in all, however, the conference put the relationship between math and national security front and center.

Mathematical Sciences in Homeland Security Links

[verified September 2003]

Air Force Office of Science Research (AFOSR) http://www.afosr.af.mil

ANSER Institute for Homeland Security http://www.homelandsecurity.org/

Army Research Office (ARO) http://www.arl.army.mil/aro/index.htm
- ARO Broad Agency Announcement: Institute for Collaborative Biotechnologies http://www.arl.army.mil/aro/biotech/biotechfinalbaa.pdf
- ARO Computing and Information Sciences Division http://www.arl.army.mil/aro/computer/cis.htm
- ARO Mathematics Division http://www.arl.army.mil/aro/mcsc/math.htm
- ARO Mechanical and Environmental Sciences Division http://www.arl.army.mil/aro/ees/eng.htm

Centers for Disease Control and Prevention (CDC) Emergency Preparedness and Response http://www.bt.cdc.gov

Defense Advanced Research Projects Agency (DARPA) http://www.darpa.mil/

Department of Energy (DOE) National Security http://www.doe.gov/engine/content.do?BT_CODE=NATIONALSECURITY

Department of Homeland Security (DHS) http://www.dhs.gov/dhspublic/
- Department of Homeland Security Fellowship and Scholarship Program http://www.orau.gov/dhsed/

Devlin, Keith. Devlin's Angle: Mathematics and Homeland Security. MAA Online, April 2002. http://www.maa.org/devlin/devlin_04_02.html

Environmental Protection Agency (EPA) Chemical Emergency Preparedness and Prevention http://yosemite.epa.gov/oswer/ceppoweb.nsf/content/homelandSecurity.htm?OpenDocument

National Imagery and Mapping Agency (NIMA) http://www.nima.mil/

National Institutes of Health (NIH) Human Immunity and Biodefense Research Awards http://www.nih.gov/news/pr/sep2003/niaid-17.htm
- NIH Application of Use of Advanced Network Infrastructure Technology in a Health Related Environment. http://grants2.nih.gov/grants/guide/notice-files/NOT-LM-02-001.html

National Security Agency (NSA) http://www.nsa.gov/

National Science Foundation (NSF) Awards in Response to the September 11, 2001 Attacks http://www.nsf.gov/od/lpa/news/media/01/nsf_response_awards.htm
- NSF and Homeland Security http://www.nsf.gov/od/lpa/news/media/01/nsf_response.htm
- NSF Data Mining and Homeland Security Applications http://www.nsf.gov/od/lpa/news/03/fact030124.htm
- NSF Grants to Boost Homeland Security Research http://www.nsf.gov/od/lpa/news/02/ma0231.htm
- Colwell, Rita. "Math Matters." Address at SIAM Meeting, July 10, 2002. http://www.nsf.gov/od/lpa/forum/colwell/rc020710siam.htm

Office of Naval Research (ONR) http://www.onr.navy.mil/default.asp

Senate Subcommittee on Science, Technology, and Space: Hearing on Homeland Security and the Technology Sector. http://www.nsf.gov/od/lpa/congress/107/hs_042402cybersecurity.htm

“Statement for the Record Before the Governmental Affairs Subcommittee on International Security, Proliferation, and Federal Services Hearing on Critical Skills for National Security and the Homeland Security Federal Workforce Act.” http://www.nsa.gov/releases/20020312.pdf

US Department of Agriculture (USDA) Keeping America’s Food and Agriculture Safe http://www.usda.gov/homelandsecurity/homeland.html