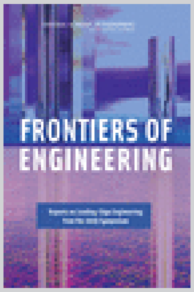


Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2009 Symposium

National Academy of Engineering

ISBN
978-0-309-14818-4

176 pages
6 x 9
PAPERBACK (2010)

FRONTIERS OF ENGINEERING

Reports on Leading-Edge Engineering from the 2009 Symposium

NATIONAL ACADEMY OF ENGINEERING
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: This publication has been reviewed according to procedures approved by a National Academy of Engineering report review process. Publication of signed work signifies that it is judged a competent and useful contribution worthy of public consideration, but it does not imply endorsement of conclusions or recommendations by the NAE. The interpretations and conclusions in such publications are those of the authors and do not purport to represent the views of the council, officers, or staff of the National Academy of Engineering.

Funding for the activity that led to this publication was provided by The Grainger Foundation, Arnold O. and Mabel Beckman Foundation, Air Force Office of Scientific Research, Defense Advanced Research Projects Agency, Department of Defense (DDR&E), National Science Foundation, Microsoft Research, and Cummins, Inc.

International Standard Book Number-13: 978-0-309-14818-4
International Standard Book Number-10: 0-309-14818-9

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

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. Ralph J. Cicerone 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. Charles M. Vest 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. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

ORGANIZING COMMITTEE

ANDREW M. WEINER (*Chair*), Scifres Family Distinguished Professor of Electrical and Computer Engineering, Purdue University
STEPHANIE GUERLAIN, Associate Professor, Department of Systems and Information Engineering, University of Virginia
SETH GUIKEMA, Assistant Professor, Department of Geography and Environmental Engineering, Johns Hopkins University
LOUISE HAMLIN, Technical Manager, Next-Generation Image Spectrometers, NASA Jet Propulsion Laboratory
EVA K. LEE, Associate Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology
MICHAL LIPSON, Associate Professor, Department of Electrical and Computer Engineering, Cornell University
J. CHRISTOPHER LOVE, Texaco-Hangeldorf Career Development Assistant Professor, Department of Chemical Engineering, Massachusetts Institute of Technology
PATRICK O'MARA, Senior Transportation Engineer, STV, Inc.
NAREN RAMAKRISHNAN, Professor and Associate Head, Department of Computer Science, Virginia Tech
NELSON TANSU, Associate Professor, Department of Electrical and Computer Engineering, Lehigh University

Staff

JANET R. HUNZIKER, Senior Program Officer
VIRGINIA BACON TALATI, Program Associate
JACQUELINE M. MARTIN, Senior Program Assistant

Preface

In 1995, the National Academy of Engineering (NAE) initiated the Frontiers of Engineering Program, which brings together about 100 young engineering leaders at annual symposia to learn about cutting-edge research and technical work in a variety of engineering fields. The 2009 U.S. Frontiers of Engineering Symposium was held at The National Academies' Arnold O. and Mabel Beckman Center on September 10-12. Speakers were asked to prepare extended summaries of their presentations, which are reprinted in this volume. The intent of this book is to convey the excitement of this unique meeting and to highlight cutting-edge developments in engineering research and technical work.

GOALS OF THE FRONTIERS OF ENGINEERING PROGRAM

The practice of engineering is continually changing. Engineers today must be able not only to thrive in an environment of rapid technological change and globalization, but also to work on interdisciplinary teams. Cutting-edge research is being done at the intersections of engineering disciplines, and successful researchers and practitioners must be aware of developments and challenges in areas that may not be familiar to them.

Every year at the U.S. Frontiers of Engineering Symposium, 100 of this country's best and brightest engineers, ages 30 to 45, have an opportunity to learn from their peers about pioneering work being done in many areas of engineering. The symposium gives early-career engineers working in academia, industry, and government in many different engineering disciplines an opportunity to make contacts with and learn from individuals they would not meet in the usual round of professional meetings. This networking may lead to collaborative work and facilitate the transfer of new techniques and approaches. It is hoped that the exchange of information on current developments in many fields of engineering will lead to insights that may be applicable in specific disciplines and thereby build U.S. innovative capacity.

The number of participants at each meeting is limited to 100 to maximize opportunities for interactions and exchanges among the attendees, who are chosen through a competitive nomination and selection process. The topics and speakers for each meeting are selected by an organizing committee of engineers in the same 30- to 45-year-old cohort as the participants. Different topics are covered each year, and, with a few exceptions, different individuals participate.

Speakers describe the challenges they face and communicate the excitement of their work to a technically sophisticated audience with backgrounds in many disciplines. Each speaker provides a brief overview of his/her field of inquiry; defines the frontiers of that field; describes experiments, prototypes, and design studies that have been completed or are in progress, as well as new tools and methodologies, and limitations and controversies; and then summarizes the long-term significance of his/her work.

THE 2009 SYMPOSIUM

The four general topics covered at the 2009 meeting were: engineering tools for scientific discovery, nano/micro photonics and new applications, engineering the health care delivery system, and resilient and sustainable infrastructure. The Engineering Tools for Scientific Discovery session described how advances in technologies and tools provide the foundation for scientific advances. Talks in the session provided examples of this at various scales—from self-assembly at the micro-scale through tools for studying marine mammals and planetary systems to computational sustainability that can answer questions about the interactions of the environment, economics, and societies. The Nano/Micro Photonics and New Applications session focused on the development of optical materials structured on a length scale comparable to the wavelength of light. Presentations in this session covered optical antennas for enhanced light-matter interactions, light forces in guided-wave nanostructures, intersubband optoelectronics, and light-emitting diode technology for solid-state lighting. Health care delivery was the topic of the third session, which included presentations on health information technology and its role in diagnosis and treatment advances, patient safety and detection of adverse events, and effective disease management. The symposium concluded with talks on resilient and sustainable infrastructure that included an overview of the state of U.S. infrastructure, methods for assessing the vulnerability of urban infrastructure systems to natural disasters, and life-cycle assessment modeling.

In addition to the plenary sessions, the participants had many opportunities to engage in informal interactions. On the first afternoon of the meeting, participants broke into small groups for "get-acquainted" sessions during which individuals presented short descriptions of their work and answered questions from their colleagues. This helped attendees get to know more about each other relatively early in the program. On the second afternoon, attendees had another opportunity for informal interaction by joining "salons" that were discipline- or topic-based, e.g., materials, product/process development, energy/environment, information/communication, etc. There were also groups on engineering education issues and a movie and discussion about infrastructure that was a lead-in to the following day's plenary session.

Every year, a distinguished engineer addresses the participants at dinner on the first evening of the symposium. The speaker this year was Dr. Bradford W. Parkinson, Edward C. Wells Professor of Aeronautics and Astronautics, Emeritus, at Stanford University, who gave a talk on the challenges of developing the global positioning system (GPS) as well as its military and civilian applications and future.

NAE is deeply grateful to the following organizations for their support of the 2009 U.S. Frontiers of Engineering Symposium:

- The Grainger Foundation
- Arnold O. and Mabel Beckman Foundation
- Air Force Office of Scientific Research
- Department of Defense–DDR&E Research
- Defense Advanced Research Projects Agency. This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U.S. Army Research Office under contract number W911NF-09-1-0098.
- National Science Foundation. This material is based upon work supported by the National Science Foundation under grant number EFRI-0903867.
- Microsoft Research
- Cummins Inc.

NAE would also like to thank the members of the Symposium Organizing Committee (p. iv), chaired by Dr. Andrew M. Weiner, for planning and organizing the event.

Contents

ENGINEERING TOOLS FOR SCIENTIFIC DISCOVERY

NANO/MICRO PHOTONICS AND NEW APPLICATIONS

ENGINEERING THE HEALTH CARE DELIVERY SYSTEM

RESILIENT AND SUSTAINABLE INFRASTRUCTURE

APPENDIXES

ENGINEERING TOOLS FOR SCIENTIFIC DISCOVERY

Introduction

LOUISE HAMLIN
Jet Propulsion Laboratory
Pasadena, California

J. CHRISTOPHER LOVE
Massachusetts Institute of Technology
Cambridge, Massachusetts

NAREN RAMAKRISHNAN
Virginia Tech
Blacksburg, Virginia

New technologies and tools often provide the critical foundations for new discoveries in science. Revolutions in instrumentation for observing and measuring intrinsic characteristics and behaviors of natural systems are often followed by periods of prolific scientific activity. For example, telescopes and microscopes are tools that enabled us to look at systems that are much larger and much smaller than those that fit into our conventional frame of reference. They provided the means to develop many fundamental ideas about how matter is organized, how it interacts, and how it empowers the processes of life. Similar to the advent of tools for observing the world around us (i.e., telescopes and microscopes), tools for modeling and simulating processes by computational and mathematical methods have provided a structure for integrating knowledge gained from observational science and predicting future responses or outcomes.

The four presentations in this session highlight recent advances in technologies that have opened new windows into how systems comprising discrete members organize and interact. The speakers will cover systems that range in scale from nanoscale systems to natural oceanic and environmental systems.

In the first presentation, Vinothan Manoharan explains how very simple systems of micro- or nano-particles can assemble themselves into ordered structures.

Understanding the mechanisms of self-assembly is essential to building new kinds of optical materials and photonic devices. At the level of individual organisms, Sean Wiggins describes how situating listening devices at various locations around the world enables scientists to track the behavior of marine mammals and determine how these sound-dependent creatures are affected by their increasingly noisy environment.

On a much different scale, Riley Duren introduces us to the Kepler space telescope, which helps answer fundamental questions about the formation of solar systems and the frequencies of Earth-like planets. Finally, Carla Gomes describes advances in the new field of computational sustainability that can contribute to the health of the environment, the success of our economy, and human well-being. For instance, this work shows how we can stabilize tuna populations by changing the way diminishing natural resources are allocated, and how we can transition to ethanol fuel without destabilizing food production.

The connecting thread among these talks is an understanding of how systems made up of discrete members organize themselves and interact and how engineering tools can support that understanding.

Digital Holographic Microscopy for 3D Imaging of Complex Fluids and Biological Systems

Vinothan N. Manoharan
Harvard University
Cambridge, Massachusetts

In the past decade, three-dimensional optical imaging has emerged as a leading tool for scientific discovery in many fields, including condensed-matter physics, materials science, and biology. Confocal microscopy and related techniques provide quantitative characterizations of both the structure and dynamics of systems of many interacting microscopic components. In materials science and physics, these real-space imaging techniques have revealed in enormous detail the mechanisms of phase transitions and nonequilibrium phenomena such as glass formation—mechanisms that cannot in general be resolved using ensemble-averaging techniques such as scattering (van Blaaderen and Wiltzius, 1995; Weeks et al., 2000; Yethiraj and van Blaaderen, 2003). In addition, these techniques have led to a better quantitative understanding of many biological processes, including cell adhesion (Discher et al., 2005), mechanosensing (Ingber, 2003), and nucleoid formation in bacteria (Bates and Kleckner, 2005; Jun and Mulder, 2006).

Building up a three-dimensional image using optical microscopy requires mechanically scanning the field of view through a relatively thin sample. Thus these techniques are limited to studying processes that occur on time scales slower than the acquisition time, which is on the order of one second. An addition, confocal microscopes cost hundreds of thousands of dollars. This paper describes a promising interferometric technique, digital holographic microscopy (DHM), that is designed to overcome many of the limitations of optical microscopy and enable new experiments and new scientific discoveries.

DIGITAL HOLOGRAPHIC MICROSCOPY

Principles

DHM (Garcia-Sucerquia et al., 2006; Schnars and Juptner, 2002) is based on a technique originally outlined by Denis Gabor in 1949. A simple, in-line holographic microscope (Figure 1) consists of a laser, a camera, and an objective lens. Laser light scatters off structures in a microscopic sample, and the camera images the interference pattern between the scattered and unscattered light. The image is called a hologram.

To see how this technique can lead to a three-dimensional image, consider a coherent plane wave scattering off an idealized point particle. This particle represents anything that scatters light in the microscopic sample (e.g., a nanoparticle in solution or a subcellular structure). If we place a screen directly in front of the particle, we can see a fringe pattern that arises from interference between the scattered wave and the unscattered portion of the plane wave. This interference pattern has concentric circular fringes of varying intensity. The number and diameter of the fringes varies with the distance of the screen from the particle, and therefore, the interference pattern tells us about the location of the particle in all three dimensions.

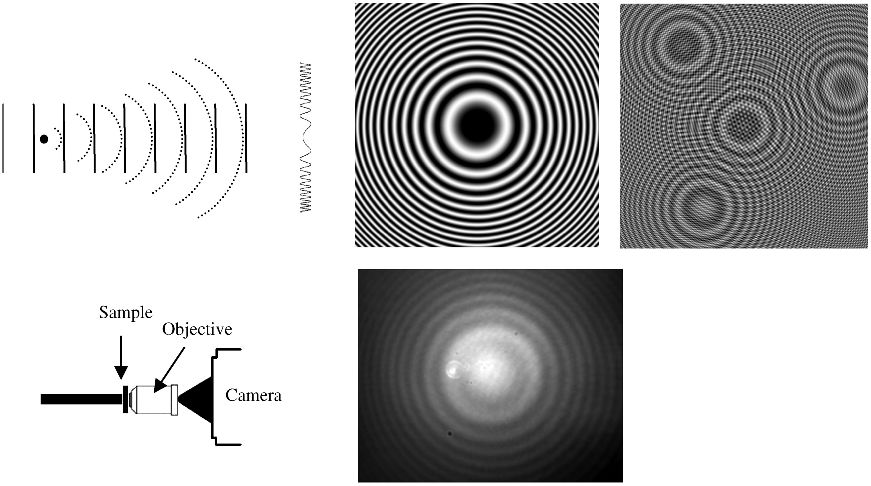


FIGURE 1 Top left: Hologram formation from a point source showing a plane reference wave interfering with a spherical scattered wave to produce (center, calculated) a Gabor zone plate pattern. Right, calculated: A set of point scatterers yields the coherent superposition of zone plates. Bottom center: Diagram of in-line holographic instrument (bottom left); image of hologram from 1 μm diameter polystyrene particle in water.

If the interference pattern is exposed onto photographic film, it produces an amplitude pattern called a Gabor zone plate. The film containing the zone plate now functions as a type of diffractive lens; if we shine a plane wave back through the developed film, it will come to a focus at a point exactly where our point particle was. If the object consists of many point particles distributed through space, then the interference pattern is simply the coherent superposition of Gabor zone plates. As Gabor first noted, when the scattering from the particles is weak, one can shine a plane wave back through the hologram (the photograph of the interference pattern) and recover an accurate, three-dimensional reconstructed image of the object.

In DHM, the hologram is captured on a digital camera, and the reconstruction is done by numerically solving the Fresnel-Kirchhoff diffraction equation—in effect, digitally "shining light" back through the hologram. Today nearly all holography is digital. Because the hologram of a set of point sources is mathematically equivalent to the Fourier convolution of zone plate patterns, one can use Fourier-transform methods for the reconstruction (Kreis, 2002).

Holography, therefore, enables us to obtain a volume rendering of a three-dimensional object or particle distribution from a single two-dimensional image. This means DHM is, in principle, a fast, real-space, three-dimensional imaging technique. Compared to other 3D techniques, such as tomography (Xu et al., 2001) and confocal microscopy (Weeks et al., 2000), DHM is potentially 3 orders of magnitude faster in acquisition time.

Moreover, DHM, especially in the in-line configuration that requires only a laser, an objective, and a camera, is simple and cheap to build. The major cost is the huge increase in processing time required to reconstruct the 3D image. DHM is not a new technique, but until recently its applications have been limited because of the high computational cost of processing holograms. The availability of fast computers for image processing, inexpensive semiconductor lasers, and high-speed CMOS cameras—all made possible by advances in semiconductor technology—have made it possible to build holographic microscopes cheaply and easily in the laboratory.

Quantitative Holographic Microscopy

Because the particles imaged in the hologram, which form the structure of the sample, can be tracked through space and time with high spatial precision, DHM can be used as a quantitative imaging technique. Like many other particle-tracking techniques (Crocker and Grier, 1996), there is some confusion over the word "resolution." In holography, as in all optical techniques, the spatial resolution is limited by diffraction; at best, one can resolve two point sources about a wavelength apart, but not much closer than that. However, it is possible to resolve the center of brightness of a *single spherical particle* to a precision on the order of tens of nanometers, well below the diffraction limit.

For example, using a DHM built with a diode laser and a high-speed CMOS camera, my research group has captured holograms of a 1 μm polystyrene sphere diffusing in water as a function of time. By reconstructing these images and tracking the centroid of the particle, we can measure the diffusion coefficient in all three dimensions to within 5 percent of the value expected from the Stokes-Einstein relation. DHM can accurately measure particle dynamics in real space, in three dimensions, with spatial precision of about 10 nanometers (nm) and temporal resolution of milliseconds. Even smaller time scales (tens of microseconds) are possible with faster cameras.

HOLOGRAPHIC MICROSCOPY FOR STUDYING SELF-ASSEMBLY IN COMPLEX FLUIDS

We use DHM primarily for studying the self-assembly of complex fluids, and in particular colloidal and nanoparticle suspensions; the particles in these systems have diameters of 10 to 1,000 nm. Interactions among these particles can drive them to self-assemble into ordered structures at equilibrium, such as colloidal crystals, which can serve as the basis for advanced functional materials like photonic crystals (Vlasov et al., 2001). Thus, at least in principle, self-assembly represents a cheap and easy way to fabricate advanced materials (Dinsmore et al., 1998, 2002; Klein et al., 2005; Manoharan and Pine, 2004). However, we do not yet understand the dynamic processes involved in self-assembly. DHM is one of the few experimental tools that can resolve the particle positions and dynamics in these small, rapidly changing, isolated systems.

A small self-assembling system consisting of colloidal particles on the surface of a spherical oil droplet is shown in Figure 2 (McGorty et al., 2008). Altogether, there are 10 polymer particles, each with a diameter of about 800 nm, that interact on the surface of the droplet to form an ordered polyhedron. Such small self-assembled structures, which are common in industrial fluids and personal care products, are also useful as building blocks for new materials. But, because of droplet rotation and particle motion, such systems are nearly impossible to probe in three dimensions with confocal microscopy.

With DHM, however, it is possible to capture the full three-dimensional positions of all of the particles at a given time in one hologram (Figure 2). We gather these fully three-dimensional data sets at a rate of 30 per second. Because the system is at equilibrium, we can use Boltzmann statistics to extract parameters, such as the potential of mean force (also shown in the figure), which is the interparticle pair potential averaged over the configurations of all other particles in the system.

The advantage of this technique is that it gives us (in theory) *all the information* about the system at equilibrium. With each hologram we obtain a 3D snapshot of the system at a particular spot in its configuration space. The statistics of these configurations can then be used to derive the interactions and energetics through

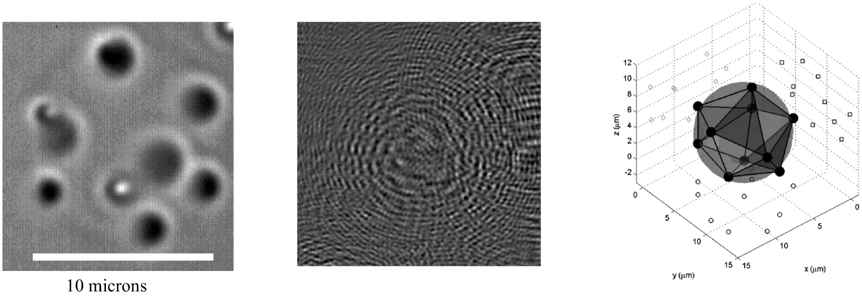


FIGURE 2 Left: Bright field microscope image of 10 particles at the interface of an emulsion droplet (index matched). Center: Hologram of same droplet taken with our DHM instrument. Right: Reconstruction of hologram showing positions of particles on the surface of the droplet and projections of positions onto coordinate planes. Reconstruction is drawn to scale. Source: McGorty et al., 2008.

the Boltzmann distribution (Crocker and Grier, 1994; Lee and Grier, 2007). These measurements are beginning to reveal the mechanism of how these interactions lead to the ordered structure seen on the surface of the droplet.

CHALLENGES AND PROSPECTS

There are still some challenges to using holography as a general-purpose 3D imaging technique. The most severe drawback is the enormous processing power, memory, and storage capacity required for computing holographic reconstructions. Today, with fast Fourier-transform algorithms, we can reconstruct a hologram for a single z-section within a few tenths of a second on a common personal computer. But we generally work with thousands of holograms taken at different times, and for each hologram we reconstruct a volume that contains about a thousand different z-sections. Thus processing a typical time series of holograms takes several days on a personal workstation, and for each 1 gigabyte holographic "movie," we can generate nearly a terabyte of volumetric reconstructions. In the meantime, although work remains to be done before DHM can replace confocal microscopy as a general tool, it can be used to probe specific systems and important scientific questions.

One intriguing possibility is biological imaging. Several studies have demonstrated that holographic microscopy can be used to image subcellular structures in three dimensions in near-real time (Choi et al., 2007; Marquet et al., 2005; Xu et al., 2001). Holographic microscopes could greatly reduce costs and sample preparation times over confocal microscopy. In fact, the most expensive part of a DHM is the computer used to reconstruct the images. Therefore, it becomes possible to use a microscope as an "add-on" to an existing biological apparatus, such as a cell-culture chamber, rather than as a central facility. With this goal in mind, we have started to build inexpensive holographic microscopes from diode lasers and consumer-grade digital cameras. Such instruments, which can be built for less than $1,000, might be useful for studying traction forces (Dembo et al., 1996) and cell rheology (Bursac et al., 2005; Hoffman et al., 2006), as well as for providing a more physical understanding of certain pathologies, such as sickle cell anemia (Chien, 1987).

ACKNOWLEDGMENTS

This work was supported by CAREER award from the National Science Foundation.

REFERENCES

Bates, D., and N. Kleckner. 2005. Chromosome and replisome dynamics in e. coli: loss of sister cohesion triggers global chromosome movement and mediates chromosome segregation. Cell 121(6): 899–911.

Bursac, P., G. Lenormand, B. Fabry, M. Oliver, D.A. Weitz, V. Viasnoff, J.P. Butler, and J.J. Fredberg. 2005. Cytoskeletal remodelling and slow dynamics in the living cell. Nature Materials 4(7): 557–561.

Chien, S. 1987. Red cell deformability and its relevance to blood flow. Annual Review of Physiology 49(1): 177–192.

Choi, W., C. Fang-Yen, K. Badizadegan, S. Oh, N. Lue, R.R. Dasari, and M.S. Feld. 2007. Tomographic phase microscopy. Nature Methods 4(9): 717–719.

Crocker, J.C., and D.G. Grier. 1994. Microscopic measurement of the pair interaction potential of charge-stabilized colloid. Physical Review Letters 73(2): 352–355.

Crocker, J.C., and D.G. Grier. 1996. Methods of digital video microscopy for colloidal studies. Journal of Colloid and Interface Science 179(1): 298–310.

Dembo, M., T. Oliver, A. Ishihara, and K. Jacobson. 1996. Imaging the traction stresses exerted by locomoting cells with the elastic substratum method. Biophysical Journal 70(4): 2008–2022.

Dinsmore, A.D., J.C. Crocker, and A.G. Yodh. 1998. Self-assembly of colloidal crystals. Current Opinion in Colloid and Interface Science 3(1): 5–11.

Dinsmore, A.D., M.F. Hsu, M.G. Nikolaides, M. Marquez, A.R. Bausch, and D.A. Weitz. 2002. Colloidosomes: self-assembled, selectively permeable capsules composed of colloidal particles. Science 298(5595): 1006–1009.

Discher, D.E., P. Janmey, and Y.L. Wang. 2005. Tissue cells feel and respond to the stiffness of their substrate. Science 310(5751): 1139–1143.

Gabor, D. 1949. Microscopy by reconstructed wave-fronts. Proceedings of the Royal Society A Mathematical, Physical and Engineering Sciences 197(1051): 454–487.

Garcia-Sucerquia, J., W. Xu, S.K. Jericho, P. Klages, M.H. Jericho, and H.J. Kreuzer. 2006. Digital in-line holographic microscopy. Applied Optics 45(5): 836–850.

Hoffman, B.D., G. Massiera, K.M. Van Citters, and J.C. Crocker. 2006. The consensus mechanics of cultured mammalian cells. Proceedings of the National Academy of Sciences 103(27): 10259–10264.

Ingber, D.E. 2003. Mechanosensation through integrins: cells act locally but think globally. Proceedings of the National Academy of Sciences 100(4): 1472–1474.

Jun, S., and B. Mulder. 2006. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome. Proceedings of the National Academy of Sciences 103(33): 12388–12393.

Klein, S.M., V.N. Manoharan, D.J. Pine, and F.F. Lange. 2005. Synthesis of spherical polymer and titania photonic crystallites. Langmuir 21(15): 6669–6674.

Kreis, T.M. 2002. Frequency analysis of digital holography with reconstruction by convolution. Optical Engineering 41(8): 1829–1839.

Lee, S.H., and D.G. Grier. 2007. Holographic microscopy of holographically trapped three-dimensional structures. Optics Express 15(4): 1505–1512.

Manoharan, V.N., and D.J. Pine. 2004. Building materials by packing spheres. MRS Bulletin 29(2): 91–95.

Marquet, P., B. Rappaz, P.J. Magistretti, E. Cuche, Y. Emery, T. Colomb, and C. Depeursinge. 2005. Digital holographic microscopy: a noninvasive contrast imaging technique allowing quantitative visualization of living cells with subwavelength axial accuracy. Optics Letters 30(5): 468–470.

McGorty, R., J. Fung, D. Kaz, S. Ahn, and V.N. Manoharan. 2008. Measuring Dynamics and Interactions of Colloidal Particles with Digital Holographic Microscopy. In Digital Holography and Three-Dimensional Imaging, OSA Technical Digest. Available online at *http://www.opticsinfobase.org/DirectPDFAccess/30DA92AA-BDB9-137E-C8A9FB62AAFF878B_156213.pdf?da=1&id=156213&seq=0*.

Schnars, U., and W.P.O. Juptner. 2002. Digital recording and numerical reconstruction of holograms. Measurement Science & Technology 13(9): R85–R101.

van Blaaderen, A., and P. Wiltzius. 1995. Real-space structure of colloidal hard-sphere glasses. Science 270(5239): 1177–1179.

Vlasov, Y.A., X.Z. Bo, J.C. Sturm, and D.J. Norris. 2001. On-chip natural assembly of silicon photonic bandgap crystals. Nature 414(6861): 289–293.

Weeks, E.R., J.C. Crocker, A.C. Levitt, A. Schofield, and D.A. Weitz. 2000. Three-dimensional direct imaging of structural relaxation near the colloidal glass transition. Science 287(5453): 627–631.

Xu, W.B., M.H. Jericho, I.A. Meinertzhagen, and H.J. Kreuzer. 2001. Digital in-line holography for biological applications. Proceedings of the National Academy of Sciences 98(20): 11301–11305.

Yethiraj, A., and A. van Blaaderen. 2003. A colloidal model system with an interaction tunable from hard sphere to soft and dipolar. Nature 421(6922): 513–517.

Engineering Tools for Studying Marine Mammals

Sean M. Wiggins
Scripps Institution of Oceanography
University of California, San Diego

Studying whales and dolphins (cetaceans) can be challenging because they often occupy remote, inhospitable areas, spend much of their lives underwater, and have diverse behaviors and habitats. With ship-based visual observations, the traditional method of studying marine mammals in the ocean, animals can be counted and their behaviors noted. However, because whales and dolphins spend as much as 95 percent of their time below the sea surface and beyond visual range, traditional methods require years to decades of effort to provide enough data for the development of ecological and behavioral models. Fortunately, because these animals often use sound while submerged to sense their environment, communicate with each other, and find food by echolocation, acoustic monitoring methods can provide rich data sets for studying them.

The reliance on sound for their life functions potentially makes marine mammals highly susceptible to acoustic disturbances. For example, the use of sonar by the U.S. Navy has been linked to mass strandings (Cox et al., 2006; Frantzis, 1996), and increased ambient noise levels from large-scale commercial shipping (McDonald et al., 2006) might interfere with their ability to locate food and mates. Adverse effects from these and other anthropogenic activities, especially the use of sonar, have accelerated the development of technologically advanced tools for studying marine mammals.

Two main types of tools—tags and passive acoustic monitors—are used to study the natural behavior of whales and dolphins and their responses to sound beneath the sea surface. Tags are small instrumented devices that are attached to individual animals; they provide detailed behavioral information of the tagged animal. However, because of the difficulty of finding animals in the ocean and of attaching the tags, not all species of marine mammals have been tagged, and the number of tagged animals is relatively small. Despite these limitations, detailed information provided by tags has greatly advanced our understanding of many species.

Acoustic monitoring (which has also been incorporated into some tag devices) is typically used over much longer periods of time and larger distances to provide temporal and spatial patterns of animal and anthropogenic sounds, providing a basis for the development of ecological and behavioral response models. This paper describes technologically advanced devices used for tagging and acoustic monitoring and some of the challenges of making measurements with these tools in the ocean environment.

TAGS

Tags for studying whales and dolphins have various capabilities: pressure sensors for measuring depth; global positioning system (GPS) and Argos satellite receivers for tracking movement on a large scale; compass sensors for identifying headings; multi-axis accelerometers for describing swimming dynamics; acoustic sensors for recording sound; and video cameras for capturing images (e.g., Andrews et al., 2005; Burgess et al., 1998; Goldbogen et al., 2006; Hooker and Baird, 1999; Johnson and Tyack, 2003; Marshall, 1998). Two techniques are used to attach tags. Long-duration (weeks to months) tags use barbed darts that pierce the animal's skin and are delivered by crossbow or air-gun (Figure 1A). Short-duration (hours to days) tags use non-invasive suction cups and are typically placed on the animal using long poles (Figure 1B).

Long-deployment tracking tags, which send data (locations) to scientists through satellite communications and are not recovered, can be packaged in a small form factor because battery requirements are small and data are not stored. Recoverable short-duration suction-cup tags require larger packaging for data storage, additional sensors and batteries, a radio transmitter, and a flotation device. Acoustic recorders incorporated into some short-duration, multi-sensor tags have provided valuable information on the acoustic behavior of large cetaceans (whales). Thus far, however, they have not been used on smaller cetaceans (dolphins), because their large size would create noticeable hydrodynamic drag that would interfere with the animal's movements. Another drawback of using suction-cup tags for small cetaceans is that these animals produce higher frequency (10s–100s kHz) sounds than larger animals, such as baleen whales (10 Hz–1000s Hz). Thus smaller animals require faster sample rates, which in turn require larger storage capacity and additional batteries.

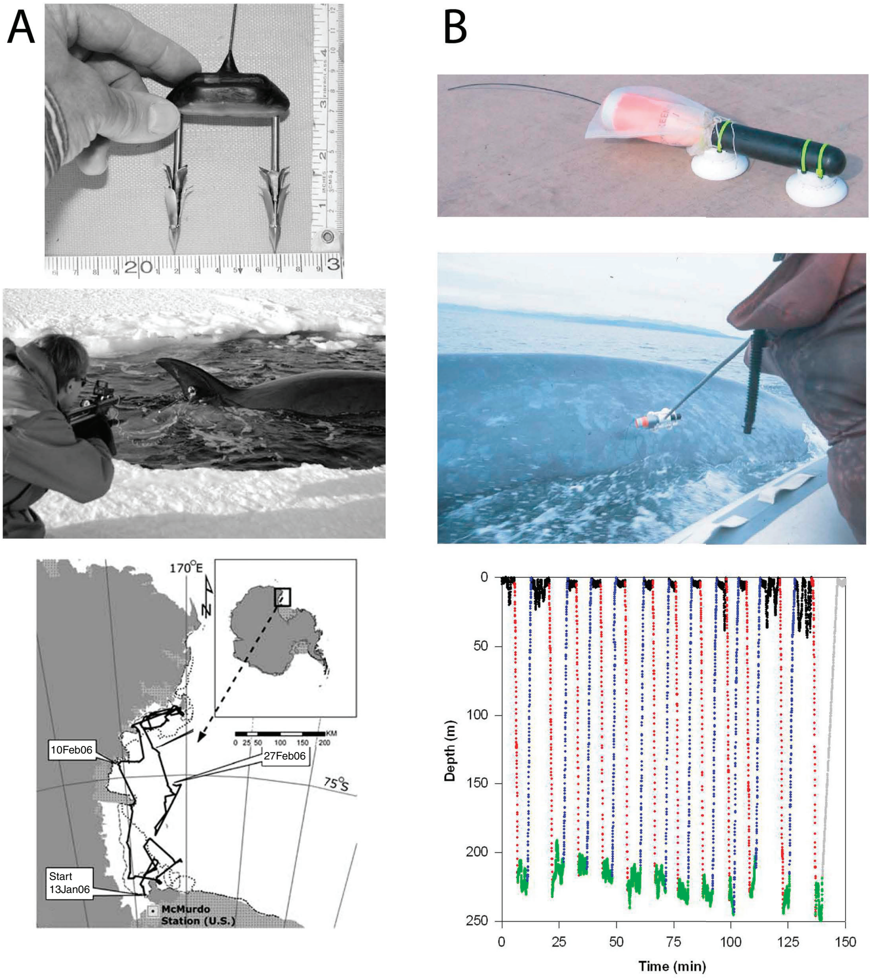


FIGURE 1 Two types of tags, their attachment methods, and results. (A) Long-duration (days to months) barbed darts on satellite tracking tag, air-gun deployment of tag on killer whale, and one month of killer whale tracks. Source: Reprinted with permission from Springer Science+Business Media: Andrews et al., 2008. (B) Short-duration (hours to days) suction cups on B-probe (acoustic, depth, 3-axis acceleration tag), tag attachment on blue whale using pole, and two-hour dive profiles from a tagged fin whale offshore of southern California.. Sources for photos: Erin Oleson; Burgess et al., 1998; Goldbogen et al., 2006.

Fortunately, “loss-less” data compression and small, low power, high capacity storage devices can be used with tags. Currently, the best example is a DTAG, a device based on cell phone technology that can record not only compressed loss-less audio of up to 192 kHz, but also pitch, roll, heading, and depth (Johnson and Tyack, 2003). So far, DTAGs have only been used on animals about 5 meters or longer. The current challenges are to miniaturize the packaging of these electronics further in a container that keeps seawater out at high pressures, has buoyancy and a radio transmitter for recovery, and is small enough to be attached to dolphins without affecting their swimming behavior.

PASSIVE ACOUSTIC MONITORING TOOLS

A variety of tools are used for remote monitoring of free-ranging dolphins and whales over long periods of time. Some of these passive devices provide real-time acoustics via cabled or radio-linked hydrophones (e.g., McDonald, 2004). Others are autonomous devices that record sounds internally (e.g., Clark et al., 2002; Fox et al., 2001; Lammers et al., 2008; Wiggins, 2003).

In many ways, autonomous acoustic recorders are more practical than real-time systems because they can be deployed in remote locations worldwide, have lower costs, and require no supervision for collecting data. In addition, because autonomous recorders have larger data and power storage capacities than acoustic tags, they can monitor for longer periods of time (months–a year). Autonomous acoustic recorders can also be distributed over large areas to provide temporal and spatial patterns and relative abundance estimations of calling animals (e.g., Munger et al., 2008; Oleson et al., 2007; Sirovic et al., 2004). Furthermore, they can be configured into arrays with close sensor spacing for tracking individuals or groups of animals (e.g., Frasier et al., 2009; McDonald et al., 1995; Tiemann et al., 2004).

One of the most capable autonomous systems currently available is the high-frequency acoustic recording package (HARP), which can sample up to 200 kHz and has 2 terabytes (TB) of data storage on 16 laptop-type disk drives (Wiggins and Hildebrand, 2007). I have been developing HARPs (Figure 2) since 2004 at Scripps Institution of Oceanography and deploying them worldwide in deep and shallow waters to monitor and track a variety of marine mammals, from low-frequency (10 Hz) blue whales to high-frequency (100 kHz) dolphins. Continuous sampling at 200 kHz and 16 bits per sample fills up a HARP disk space in about two months (~ 35 GB/day), at which point the instrument must be recovered and refurbished with new batteries and disks. However, servicing instruments every few months requires ship time and personnel, both of which are expensive.

One solution my group is working on is to increase data storage capacity while lowering power consumption by replacing the hard disk drives with solid-state memory (NAND flash), a type of data storage that has been shown with time to decrease in price and increase in capacity. Even as improvements are made, our

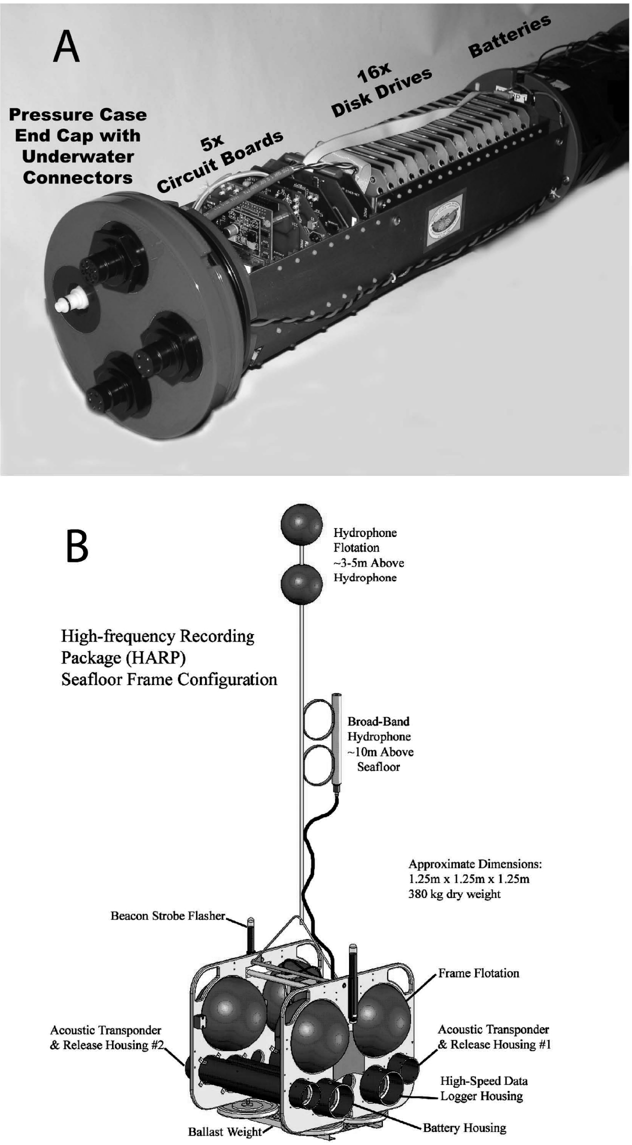


FIGURE 2 High-frequency acoustic recording package (HARP). (A) Data logger attached to end cap of pressure resistant case. The autonomous data logger consists of low-power electronics including 200 kSample/sec analog-to-digital converter, low drift (10^{-8}) clock, and about 2 TB of data storage on 16 laptop-style hard disk drives. Source: Adapted from Wiggins and Hildebrand (2007). (B) HARP instrumentation packaging in a seafloor-mounted frame. Source: Adapted from Wiggins and Hildebrand (2007).

greatest challenge is analyzing the large amounts of acoustic data collected by these instruments. Each instrument records up to 12 TB/year, and the number of instruments is increasing from our current count of 25 (i.e., 300 TB/year).

Acoustic data are measured as time series of pressure, which can be transformed into the spectral (frequency) domain via Fourier transforms and displayed as spectrogram (time-frequency) plots. Because most species and man-made sounds are unique in spectral and temporal character and can be easily differentiated, spectrograms are often used to evaluate acoustic data for animal and anthropogenic sounds (Figure 3A). However, evaluating wide-frequency-band data, such as data collected by HARPs, can only be analyzed in near real-time because of human and computational limitations. Thus, analysis of long-term data sets directly by spectrograms cannot be done in a reasonable amount of time.

Long-term spectral averages (LTSAs, Figure 3B) can provide an efficient overall view of large data sets, as well as a means of searching for and evaluating events of interest (Wiggins and Hildebrand, 2007). LTSAs are essentially spectrograms with each time pixel representing many (1,000s) spectra (e.g., 5 seconds) averaged together rather than just one spectrum (e.g., 5 microseconds) as used in a typical spectrogram.

For more detailed, quantitative analyses, automated detectors can be used on time series, spectrograms, and LTSAs to find specific sounds with known characteristics in large data sets. The resulting detections can be organized by time and location to reveal seasonal, daily, and regional patterns related to species behavior and habitat. The performance of a detection algorithm is based on various algorithm parameters and the data set. Once parameters have been optimized through multiple training tests conducted by analysts, the algorithm can be used in an automated way on a full data set to find sounds with specific characteristics. However, because these long-term data sets have a wide range of sounds, many detection algorithms must be developed, optimized, and applied to the same data sets.

In the future, our approach to the problem of running multiple detectors on large data sets will be to use multiple processors arranged in clusters that can access the same data nearly simultaneously. We believe this will provide more efficient detections of a wide range of animal and anthropogenic sounds.

SUMMARY

Whales and dolphins use a variety of sounds underwater to sense their environment and to communicate. These sounds can be recorded over long durations using passive acoustic monitoring instrumentation and over shorter periods of time for more detailed information with devices attached directly to animals. These tools provide information on marine mammal spatial and temporal distribution and acoustic behavior. As work on these tools continues, they could also potentially provide information about animals' responses to anthropogenic sound

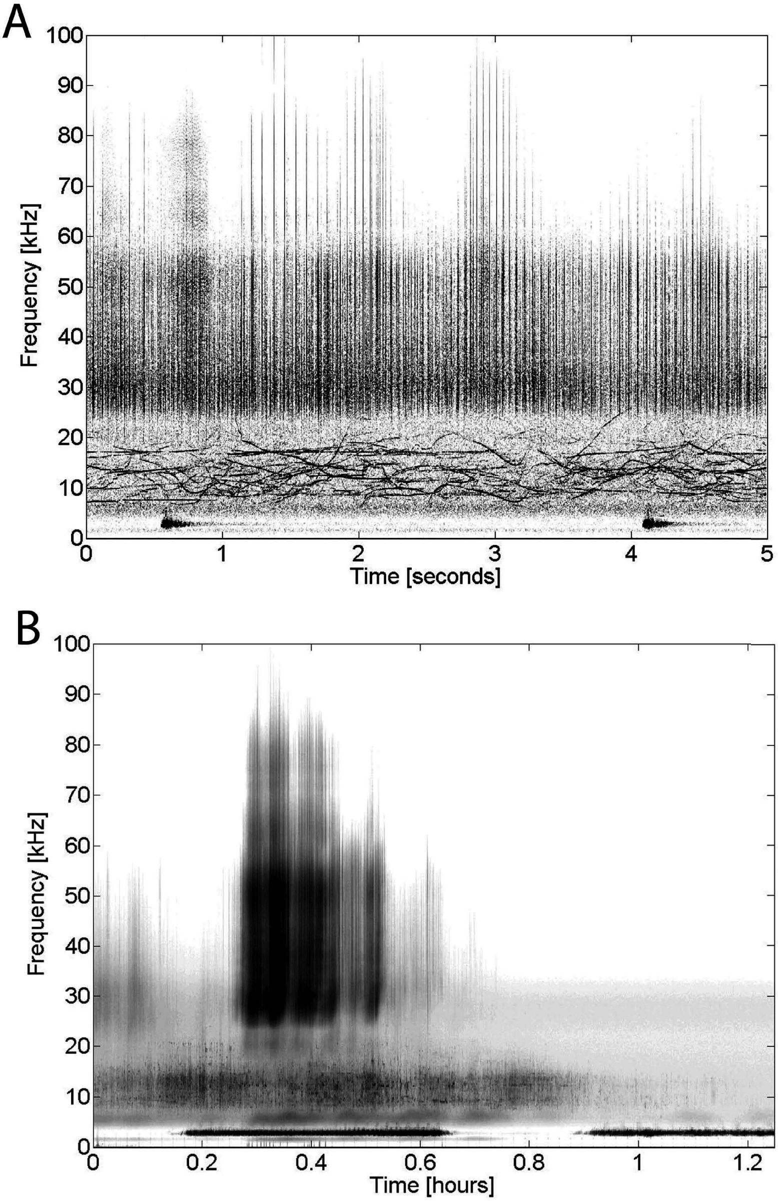


FIGURE 3 Example HARP acoustic data offshore of southern California. (A) Five-second spectrogram shows dolphin clicks from about 25 kHz up to 100 kHz, dolphin whistles from about 8 kHz to more than 20 kHz, and man-made sonar around 3 kHz. (B) Long-term spectral average (LTSA) over 75 minutes shows a bout of dolphin whistles and clicks and sonar.

sources such as sonar and explosions. At this point, these technically advanced tools have created another challenge—the need for technologies for analyzing these very large data sets.

REFERENCES

Andrews, R.D., L. Mazzuca, and C.O. Matkin. 2005. Satellite tracking of killer whales. Pp. 238–248 in Synopsis of Research on Steller Sea Lions: 2001–2005, edited by T.R. Loughlin, D.G. Calkins, and S. Atkinson. Seward, Alaska: Alaska SeaLife Center.

Andrews, R.D., R.L. Pitman, and L.T. Ballance. 2008. Satellite tracking reveals distinct movement patterns for Type B and Type C killer whales in the southern Ross Sea, Antarctica. Polar Biology 31(12): 1461–1468.

Burgess, W.C., P.L. Tyack, B.J. Le Boeuf, and D.P. Costa. 1998. A programmable acoustic recording tag and first results from free-ranging northern elephant seals. Deep-Sea Research Part II: Topical Studies in Oceanography 45(7): 1327–1351.

Clark, C.W., F. Borsani, and G. Notarbartolo-di-Sciara. 2002. Vocal activity of fin whales, Balaenoptera physalus, in the Ligurian Sea. Marine Mammal Science 18(1): 286–295.

Cox, T.M., T.J. Ragen, A.J. Read, E. Vos, R.W. Baird, K. Balcomb, J. Barlow, J. Caldwell, T. Cranford, L. Crum, A. D'Amico, G. D'Spain, A. Fernandez, J. Finneran, R. Gentry, W. Gerth, F. Gulland, J. Hildebrand, D. Houser, T. Hullar, P.D. Jepson, D. Ketten, C.D. Macleod, P. Miller, S. Moore, D.C. Moutain, D. Palka, P. Ponganis, S. Rommelo, T. Rowles, B. Taylor, P. Tyack, D. Wartzok, R. Gisiner, J. Meads, and L. Benner. 2006. Understanding the impacts of anthropogenic sound on beaked whales. Journal of Cetacean Research and Management 7(3):177–187.

Fox, C.G., H. Matsumoto, and T.A. Lau. 2001. Monitoring Pacific Ocean seismicity from an autonomous hydrophone array. Journal of Geophysical Research 106(B3): 4183–4206.

Frantzis, A. 1996. Does acoustic testing strand whales? Nature 392(6671): 29.

Frasier, K.E., S.M. Wiggins, and J.A. Hildebrand. 2009. Acoustic tracking of whistling dolphins offshore of Southern California. Journal of the Acoustical Society of America 125(4): 2548.

Goldbogen, J.A., J. Calambokidis, R.E. Shadwick, E.M. Oleson, M.A. McDonald, and J.A. Hildebrand. 2006. Kinematics of foraging dives and lunge-feeding in fin whales. Journal of Experimental Biology 209(Part 7): 1231–1244.

Hooker, S.K., and R.W. Baird. 1999. Deep-diving behaviour of the northern bottlenose whale, Hyperoodon ampullatus (Cetacea: Ziiphidae). Proceedings of the Royal Society of London B 266: 671–676.

Johnson, M.P., and P. Tyack. 2003. A digital acoustic recording tag for measuring the response of wild marine mammals to sound. IEEE Ocean Engineering 28(1): 3–12.

Lammers, M.O., R.E. Brainard, W.W.L. Au, T.A. Mooney, and K. Wong. 2008. An ecological acoustic recorder (EAR) for long-term monitoring of biological and anthropogenic sounds on coral reefs and other marine habitats. Journal of the Acoustical Society of America 123(3): 1720–1728.

Marshall, G.J. 1998. CRITTERCAM: an animal borne imaging and data logging system. Marine Technology Society Journal 32(1): 11–17.

McDonald, M.A., J.A. Hildebrand, and S.C. Webb. 1995. Blue and fin whales observed on a seafloor array in Northeast Pacific. Journal of the Acoustical Society of America 98(2): 712–721.

McDonald, M.A. 2004. DIFAR hydrophone usage in whale research. Canadian Acoustics 32(2): 155–160.

McDonald, M.A., J.A. Hildebrand, and S.M. Wiggins. 2006. Increase in deep ocean ambient noise in the Northeast Pacific west of San Nicolas Island, California. Journal of the Acoustical Society of America 120(2): 711–717.

Munger, L.M., S.M. Wiggins, S.E. Moore, and J.A. Hildebrand. 2008. North Pacific right whale (Eubalaena japonica) seasonal and diel calling patterns from long-term acoustic recordings in the southeastern Bering Sea. Marine Mammal Science 24(4): 795–814.

Oleson, E.M., S.M. Wiggins, and J.A. Hildebrand. 2007. Temporal separation of blue whale call types on a southern California feeding ground. Animal Behaviour 74(4): 881–894.

Sirovic, A., J.A. Hildebrand, S.M. Wiggins, M.A. McDonald, S.E. Moore, and D. Thiele. 2004. Seasonality of blue and fin whale calls and the influence of sea ice in the Western Antarctic Peninsula. Deep Sea Research Part II: Topical Studies in Oceanography 51(17–19): 2327–2344.

Tiemann, C.O., M.B. Porter, and L.N. Frazer. 2004. Localization of marine mammals near Hawaii using an acoustic propagation model. Journal of the Acoustical Society of America 115(6): 2834–2843.

Wiggins, S.M. 2003. Autonomous acoustic recording packages (ARPs) for long-term monitoring of whale sounds. Marine Technology Society Journal 37(2): 13–22.

Wiggins, S.M., and J.A. Hildebrand. 2007. High-frequency Acoustic Recording Package (HARP) for broad-band, long-term marine mammal monitoring. Pp. 551–557 in Proceedings of the International Symposium on Underwater Technology 2007 and International Workshop on Scientific Use of Submarine Cables & Related Technologies 2007, April 17–20, 2007, Tokyo, Japan.

The Kepler Mission:
A Search for Terrestrial Planets

Riley Duren
Jet Propulsion Laboratory
Pasadena, California

The Kepler mission, which was launched by the National Aeronautics and Space Administration (NASA) on March 6, 2009, is the first mission capable of finding Earth-size planets orbiting in the habitable zones (HZs) of other stars. Kepler will also determine the distribution of these planets near solar-like stars.[1] After a two-month commissioning activity, Kepler science observations began; these observations will monitor more than 100,000 dwarf stars simultaneously over the primary mission life of 3.5 years, with a capability of extending observations for a total of 7 years. Precision differential photometry will be used to detect the periodic signals of transiting planets via small changes in the light of their host stars. Kepler will also support asteroseismology by measuring the pressure-mode (p-mode) oscillations of selected stars.

Transits of a solar-size star by an Earth-size planet causes a reduction in the stare light of 84 parts per million (ppm) or 0.008 percent. For a statistically significant detection, the minimum single-transit signal-to-noise ratio (SNR) is taken to be 4 sigma (σ), leading to a combined average significance of 8σ for 4 transits. The Kepler combined differential photometric precision must therefore be less than 21 ppm binned at 6.5 hours (half the duration of a central transit). The

---

[1] Earth-size planets are terrestrial or rocky planets with masses ranging from 0.5 to 10 Earth masses. Smaller planets cannot maintain a life-sustaining atmosphere, and larger planets retain primordial atmospheres. The habitable zone for a given star is defined as the region in which orbiting planets have the potential for liquid surface water, a key building block of life.

detection threshold for terrestrial HZ planets in the Kepler data pipeline is set at 7σ, yielding a detection rate of 84 percent and controlling the number of expected false alarms to no more than one for the entire experiment. Given this design, the mission will be capable of detecting not only Earth analogs, but also a wide range of planetary types and characteristics, ranging from Mars-size objects and orbital periods of days to gas-giants and decade-long orbits. In fact, the mission is designed to survey the full range of spectral types of dwarf stars.

The criteria for an Earth-size planet transiting a solar-size star—three or more transits of a star with a statistically consistent period, brightness change, and duration—provide a rigorous detection method. The size of the planet can be calculated from the relative change in brightness. The size (semi-major axis) of the orbit can be calculated from the observed orbital period using Kepler's 3rd law of planetary motion and the planet's location relative to the HZ.

A key consideration when looking for planetary transits is the probability that the orbital plane is aligned along the line of sight. The probability of random orbital alignment is simply the ratio of the stellar diameter to the orbital diameter. For example, the probability of orbital alignment for a solar system like our own where the earth's orbital diameter is about 300 million kilometers (km) and the Sun's diameter is 1.4 million km is approximately 0.005 or 0.5 percent. Hence, many thousands of stars need to be monitored before a statistically meaningful result can be drawn. Confirmation of the existence of a planet will also require a sequence of transits with a consistent period, depth, and duration.

The Kepler observatory, or Flight Segment of the mission, consists of a large field-of-view (FOV) photometer and a spacecraft bus. This observatory was launched on a Delta II rocket into an Earth-trailing, heliocentric orbit, which will mean that after 3.5 years the spacecraft will have drifted less than 0.5 astronautical units away from the Earth. The photometer is a Schmidt camera design consisting of a graphite-cyanate ester metering structure, a sunshade, a 95 centimeter diameter Schmidt corrector, a 1.4 meter diameter primary mirror, field-flattening lenses, and an array of 42 charge-coupled devices (CCDs) with an active FOV greater than 100 square degrees.

The CCDs, back-illuminated devices with dual outputs and 1024 × 2200 27-μm pixels, are passively cooled by a radiator panel. The CCDs measure the brightnesses of 103,000 planetary target stars which are each accumulated (summed) for 30 minutes before being stored on the spacecraft's solid-state recorder. Approximately 512 targets assigned to a succession of different target stars are co-added at a one minute cadence to support p-mode asteroseismology and other non-transit science. Since the targets are pre-selected, only the pixels relevant to each star (rather than the entire image) will be stored for downlink. This means that only about 3 percent of the pixels will be stored, thus saving a tremendous amount of on-board storage and communications link time. In addition, a data-compression scheme is used to further reduce storage and transmission requirements. The photometer does not have a shutter, and the only moving parts are three focus and tilt adjustment mechanisms under the primary mirror and a one-time deployable dust cover.

The spacecraft provides necessary support for the photometer, including power, pointing control, and data systems. The spacecraft has three-axis stabilization and fine guidance sensors mounted on the scientific focal plane. The entire mission will be spent viewing a single star field centered near the constellation Cygnus. Attitude maneuvers will only occur every three months when the spacecraft is rotated about the photometer axis by 90 degrees to keep the solar array pointed toward the Sun and the radiator pointed toward deep space.

Given Kepler's heliocentric orbit, no disturbances will be caused by geomagnetic moments, gravity gradients, or atmospheric drag. The only disturbance will be a steady torque due to solar pressure. Stability for precision pointing is provided by reaction wheel assemblies for which periodic momentum-de-saturation maneuvers are provided by the same hydrazine-reaction control system that was used to de-spin and de-tumble the spacecraft following separation from the launch vehicle. No propulsive capability for orbital delta-V correction is required because the spacecraft's drift-away rate was limited by the launch vehicle's injection dispersion errors.

The transmission of uplink commands and downlink real-time engineering data are performed by omni-directional X-band antennas. All stored science and engineering data are downlinked using a high-gain Ka-band system. The solar arrays, which are rigidly mounted on the spacecraft, also provide some shielding of the photometer from the Sun. The only moving parts on the spacecraft bus are the reaction wheels. All spacecraft sub-systems are fully redundant.

During science operations, regular "housekeeping" communications with the spacecraft are planned to occur approximately twice weekly, and playback of the stored science data is downlinked once a month. The antennas of NASA's Deep Space Network (DSN) support Kepler's uplink and downlink communications.

The Kepler Ground Segment includes a collection of facilities, software, processes, and procedures for operating the Flight Segment and analyzing data. Overall mission direction is provided from the Mission Management and Science Offices hosted by the Science Operations Center (SOC) at NASA Ames Research Center in Mountain View, California. Strategic mission planning and target selection are performed at SOC. Target selection to separate smaller solar-type dwarf stars from giant stars,[2] is based on a Kepler-unique input catalog produced by a pre-launch Stellar Classification Program—based on ground-based observations.

Operations management of the Flight Segment, tactical mission planning, sequence validation, and engineering trend analyses are provided by a flight

---

[2] Primary interest is on main-sequence dwarf stars, which are longer lived than giant stars and hence more likely to host habitable planets. In addition, transit signals from Earth-size planets are easier to detect around smaller stars, which have a larger obscuration ratio.

planning center (FPC) at Ball Aerospace Technologies Corporation in Boulder, Colorado. Command and data processing, monitoring of the health and status of the Flight Segment, and scheduling of the DSN, are performed by the Mission Operations Center (MOC) at the Laboratory for Astronomy and Solar Physics (LASP) in Boulder, Colorado. Uplink and downlink telecommunications use NASA's DSN 34 meter antennas located in Goldstone, California; Madrid, Spain; and Canberra, Australia. Navigation and orbit propagation are provided by the Jet Propulsion Laboratory in Pasadena, California.

The Data Management Center (DMC) at the Space Telescope Science Institute (STScI) in Baltimore, Maryland, receives the "raw" telemetry data and calibrates pixel levels. The resulting calibrated data set is archived by DMC and forwarded to SOC for further processing, which includes generating calibrated photometric light curves and transit detection. STScI also provides p-mode analysis.

After an extensive data-validation process, follow-up observations on each planetary candidate will be performed (the Follow-up Observing Program [FOP]) to eliminate intrinsic false positives caused by grazing eclipsing binaries and extrinsic false positives caused by background eclipsing binaries and to discriminate between terrestrial transits of the target star and transits of giant-planet background stars. In some cases, FOP should also be able to use ground-based observations to measure the mass of the largest planets, which together with estimates of planet diameters derived from Kepler transit-depth measurements, can be combined to determine the density of the planets.

ACKNOWLEDGEMENTS

Project management and systems engineering for the Kepler mission development and commissioning was provided by the NASA Jet Propulsion Laboratory. The NASA Ames Research Center, which provides mission management during science operations, led the development of the Ground Segment and is the home of the Science Principal Investigator and Science Operations Center. The Flight Segment was designed and built by the Ball Aerospace Technology Corporation. The research described here was carried out at the Jet Propulsion Laboratory, California Institute of Technology, and was sponsored by NASA.

Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society

Carla P. Gomes
Cornell University
Ithaca, New York

The dramatic depletion of natural resources in the last century now threatens our planet and the livelihood of future generations. *Our Common Future*, a report by the World Commission on Environment and Development published in 1987, introduced for the first time the notion of “sustainable development: development that meets the needs of the present without compromising the ability of future generations to meet their needs” (UNEP, 1987). The concerns raised in that report were reiterated by the Intergovernmental Panel on Climate Change (IPCC, 2007). In the fourth Global Environmental Outlook report published later that same year the authors concluded, “there are no major issues raised in *Our Common Future* for which the foreseeable trends are favorable” (UNEP, 2007).

Key issues in the development of policies for sustainable development will entail complex decisions about the management of natural resources and more generally about balancing environmental, economic, and societal needs. Making such decisions optimally, or nearly optimally, presents significant computational challenges that will require the efforts of researchers in computing, information science, and related disciplines, even though environmental, economic, and societal issues are not usually studied in those disciplines.

In this author’s opinion, it is imperative that computer scientists, information scientists, and experts in operations research, applied mathematics, statistics, and related fields pool their talents and knowledge to help find efficient and effective ways of managing and allocating natural resources. To that end, we must develop critical mass in a new field, computational sustainability, to develop new computational models, methods, and tools to help balance environmental, economic, and societal needs for a sustainable future.

Examples of computational sustainability problems presented in this short paper range from wildlife preservation and biodiversity to balancing socio-economic needs and the environment to the large-scale deployment and management of renewable energy sources.

BIODIVERSITY AND SPECIES CONSERVATION

The reduction and fragmentation of natural habitats as a result of deforestation, agriculture, urbanization, and land development is a leading cause of species decline and extinction. One strategy for improving the chances of species viability is to protect habitats by creating biologically valuable sites or reserves. Examples include the National Wildlife Refuge System, managed by the U.S. Fish and Wildlife Service, national parks, and conservation reserves established by private groups, such as the Nature Conservancy and the Conservation Fund.

Given the limited resources available for conservation, these sites must be carefully chosen. From a mathematical point of view, the site-selection or reserve-design problem involves optimizing certain criteria, such as habitat suitability for species, while simultaneously satisfying one or more constraints, such as limited budgets (e.g., Ando et al., 1998; Moilanen et al., 2009; Polasky et al., 2008).

In recent years biologists attempting to combat habitat fragmentation have promoted so-called "conservation corridors," continuous areas of protected land that link biologically significant zones. The design of conservation corridors is a special aspect of the site-selection problem, and the objective is to create connected corridors made up of parcels of land that will yield the highest possible level of environmental benefit ("utility") (Onal and Briers, 2005; Williams et al., 2005).

At the Institute for Computational Sustainability (ICS) at Cornell University, we recently formulated this problem mathematically as a so-called "connection sub-graph problem" (Conrad et al., 2007; Dilkina and Gomes, 2009; Gomes et al., 2008). The goal was to design wildlife corridors for grizzly bears in the U.S. northern Rockies to enable movement between three core ecosystems—Yellowstone, Salmon-Selway, and Northern Continental Divide Ecosystems—that span 64 counties in Idaho, Wyoming, and Montana. This large-scale optimization problem places significant demands on current computational methods.

To scale up solutions, we needed a deeper understanding of the underlying structure of the problem. To that end, we developed a budget-constrained, utility-optimization approach using hybrid constraint-based mixed-integer programming that exploits problem structure. Our results showed that we can dramatically reduce the cost of large-scale conservation corridors by provably finding corridors with minimum cost. If more than minimum funding for a corridor is available, this approach guarantees optimal utility. For example, for the grizzly bear problem our solutions are guaranteed to be within 1 percent of the optimal solution for budget levels above the minimum cost.

Complexity in site-selection and corridor-design problems increases when different models for land acquisition over different time periods (e.g., purchase, conservation easements, auctions), dynamic and stochastic environments, and multiple species must be considered. For example, preserving bird habitats and designing bird corridors requires a good understanding of hemispheric-scale bird migrations with complex population dynamics across different climate and weather systems and geographic topologies.

Thus modeling complex species distributions and developing conservation strategies requires new large-scale stochastic-optimization methods. Moreover, to obtain the right model parameters and determine current species distribution, machine learning and statistical techniques must be used to analyze large amounts of raw data (Dietterich, 2009; Elith et al., 2006; Kelling et al., 2009; Phillips et al., 2004).

Gathering biological, ecological, and climatic data is essential to studying complex systems, and the deployment of large-scale sensor networks is becoming a key tool for environmental monitoring (e.g., Polastre et al., 2009; Werner-Allen et al., 2006). The National Science Foundation (NSF) supports several cyber-infrastructure initiatives for massive data collection and data analysis based on large-scale autonomous sensor networks, such as the National Ecological Observatory Network (NEON) and the Long-Term Ecological Research Network (LTER).

Designing a large-scale sensor network also presents computational challenges (e.g., network architecture, operating system and programming environments, data collection, analysis, synthesis, and inference) (Akyildiz et al., 2007). For example, when using sensor networks to monitor spatial phenomena, selecting the best placement of sensors to maximize information gain while minimizing communication costs is a complex problem that requires new techniques (Krause and Guestrin, 2009).

Citizen observation networks have several benefits. They help in collecting data and, at the same time, enable the general public to engage in scientific investigation and develop problem-solving skills. Galaxy Zoo,[1] for example, provides access to a large collection of images and engages the general public in classifying galaxy shapes to improve our understanding of their formation. eBird,[2] a joint initiative of the Cornell Laboratory of Ornithology and the National Audubon Society, engages citizen-scientists in observing birds using standardized protocols. Since eBird was released in 2002, it has been visited by more than 500,000 users and has collected more than 21 million bird records from more than 35,000

---

[1] Available online at *http://www.galaxyzoo.org/*.

[2] Available online at *http://ebird.org/content/ebird*.

unique users in more than 180,000 locations across the Western Hemisphere and New Zealand (Sullivan et al., 2009).

MANAGEMENT OF NATURAL RESOURCES

This example concerns the state of marine fisheries. The biomass of top marine predators is estimated to be one-tenth of what it was half a century ago and is still declining (Worm et al., 2006). As a result of overfishing, pollution, and other environmental factors, many important marine species are extinct, with dramatic consequences for the filtration of nutrients by the ocean. Researchers believe that the collapse of major fisheries is primarily the result of mismanagement (Clark, 2006; Costello et al., 2008). Therefore, we must find sustainable ways of managing fisheries.

One approach that has been shown to be effective for counterbalancing the overharvesting of fisheries involves both placing limitations on total allowable catches per species and requiring permits for harvesting specific quantities of fish (individual transferable quotas) (Costello et al., 2008; Heal and Schlenker, 2008; Worm et al., 2009). Complex dynamical models, originally developed as part of dynamical systems theory, can be used to identify the optimal amount of fish that can be harvested annually in a certain fishery, taking into consideration re-generation rates, carrying capacity of the habitat, discount rates, and other parameters.

Dynamical systems theory, which provides tools for characterizing the dynamics and long-term behavior of systems as a function of the system parameters, provides insights into nonlinear system dynamics and identifies patterns and laws, particularly bifurcations (Ellner and Guckenheimer, 2006; Strogatz, 1994). A bifurcation occurs when small changes in the parameter values of a system (e.g., the rate of harvesting fish) lead to an abrupt qualitative change (e.g., the collapse of a fishery). Decisions (e.g., the amount of fish to be harvested) are often based on combinations of continuous and discrete variables. This leads to hybrid dynamical optimization models, which, in principle, provide information on optimal harvesting strategies (Clark, 1976; Conrad, 1999). However, finding such strategies is computationally difficult, especially when considering multiple species.

BALANCING SOCIOECONOMIC AND ENVIRONMENTAL NEEDS

Chris Barrett of ICS has studied the socio-economic interrelationship between poverty, food security, and environmental stress in Africa, particularly links between resource dynamics and the poverty trap in small-holder agrarian systems (Barrett et al., 2007). Barrett's focus has been on pastoral systems in East Africa that involve herds of cattle, camels, sheep, and goats (Luseno et al., 2003). Due to high variability in rainfall, pastoralists must migrate with their herds looking for water and forage, sometimes traveling as much as 500 kilometers.

The purpose of our studies is to develop a predictive model of the migratory patterns and decision models of these pastoralists. To do that, we use machine-learning methods to determine the structure and estimate the parameters of the models, based on field data about households, water sources, and climate patterns.

Ultimately, these models will help policy makers predict the effects of potential policy interventions and environmental changes, with the goal of improving the livelihoods of thousands of pastoralists. The project involves new technical approaches to large, structural-dynamic, discrete-choice problems that will lead to the development of computational models to support both descriptive studies and predictive policy analyses (Toth et al., 2009).

Other computational sustainability topics in this context include automated decision-support tools for providing humanitarian aid in response to catastrophes, famines, and natural disasters in developing countries. The design of such systems will require the development of intuitive, user-friendly interfaces for use by aid workers.

ENERGY-EFFICIENT DATA CENTERS

The implications of climate change for environmental, economic, and social systems have led to major changes in energy policy in many industrial countries, including incentives for increasing energy efficiency. These incentives present tremendous computational opportunities for helping to increase energy efficiency through the design of intelligent or "smart" control systems for energy-efficient buildings, vehicles, and appliances.

According to the World Business Council for Sustainable Development (2008), buildings account for as much as 40 percent of energy use in industrialized countries. Data centers (i.e., computing facilities with electronic equipment for data processing, storage, and communications networking) are especially inefficient users of energy.

In recent years the shift to digital services has led to a major increase in demand for data centers. The Environmental Protection Agency estimates that in the next decade the demand for data-center capacity will grow at a 10 percent compounded annual growth rate (EPA, 2007). In addition, the costs of data centers in the information technology (IT) sector are estimated to increase at an annual rate of 20 percent, compared to an overall increase in IT of 6 percent (Kaplan et al., 2008).

Data centers also have negative environmental impacts. According to a recent report, the amount of carbon dioxide emissions produced by data centers worldwide exceeds the total emissions of both Argentina and the Netherlands (Kaplan et al., 2008). Thus the IT industry is looking to advanced power management hardware, smart cooling systems, virtualization tools, and dense server configurations to reduce energy consumption (Katz, 2009).

These new approaches rely heavily on large amounts of data provided by large-scale sensor networks (e.g., Bodik et al., 2008; Hoke et al., 2006; Patnaik et al., 2009; Shah et al., 2008). Some companies are using containers that integrate computing, power, and cooling systems in one module for data centers, instead of raised-floor rooms. Several IT companies are committed to using alternative energy sources, such as hydropower, solar power, and wind power, to bring the carbon footprint of data centers to zero.

On a larger scale, data centers can contribute to reductions in energy use and carbon emissions by facilitating e-commerce and telecommuting, for example, which can eliminate some of the need for paper printing and for freight and passenger transportation.

THE SMART GRID

Under the Energy Independence and Security Act (EISA) of 2007, the U.S. Department of Energy was charged with modernizing the nation's electricity grid to improve its reliability, efficiency, and security, a concept known as the Smart Grid. Ideally, the Smart Grid will radically transform the industry's business model from a largely non-digital, electromechanical grid to a network of digital systems and power infrastructure and from a centralized, producer-controlled network to a more decentralized system with more interaction between consumers and local producers.

The objectives for the Smart Grid include: enabling active participation by consumers; making possible the easy integration of a variety of generation options (with a focus on renewable sources) and storage options; enabling new products, services, and markets; providing quality power for the digital economy; optimizing assets and operating efficiently; automatically anticipating and responding to system disturbances; and operating resiliently in the event of attacks or natural disasters.

To realize these objectives, the Smart Grid will include smart sensors and controls throughout the transmission and distribution system and a broad communication platform for two-way communications to move data and electricity between utilities and consumers. For example, consumers will have smart meters that can track energy consumption, monitor individual power circuits in the home, control smart appliances, and actively manage energy use.

Planning and operating such a large, complex digital ecosystem will require technological advances in computing and information science related to sensing and measuring technologies, advanced control methods, monitoring and responding to events, support for dynamic pricing, computational aspects of game-theory models and mechanism design, multi-agent based models, improved interfaces, decision-support and optimization tools, and security and privacy tools.

RENEWABLE ENERGY

The development of renewable energy can have an even greater environmental impact than increasing energy efficiency. In recent years technological progress has been made (partly in response to government incentives) in renewable energy sources, such as biofuels and biomass, geothermal, solar, and wind power. For example, EISA set fuel economy standards for vehicles that will require the production of 36 billion gallons of renewable fuels per year by 2022, a fivefold increase over current ethanol production levels.

The logistics and planning of this large-scale domestic-based biofuels production system raise complex stochastic optimization problems—variants of the so-called "facility-location problem"—that must take into consideration feedstock and demand and the dynamics of demand and capacity (Shmoys, 2004). And the stakes are high. Finding good solutions to these problems can make the difference between economic viability and failure. Overall, we will need complex computational models to find the best mix of energy generation and storage technologies.

A larger project will be the development of computational models (Figure 1) that show interactions between different energy sources and the agents directly or indirectly involved (e.g., households, landowners, farmers, ethanol producers,

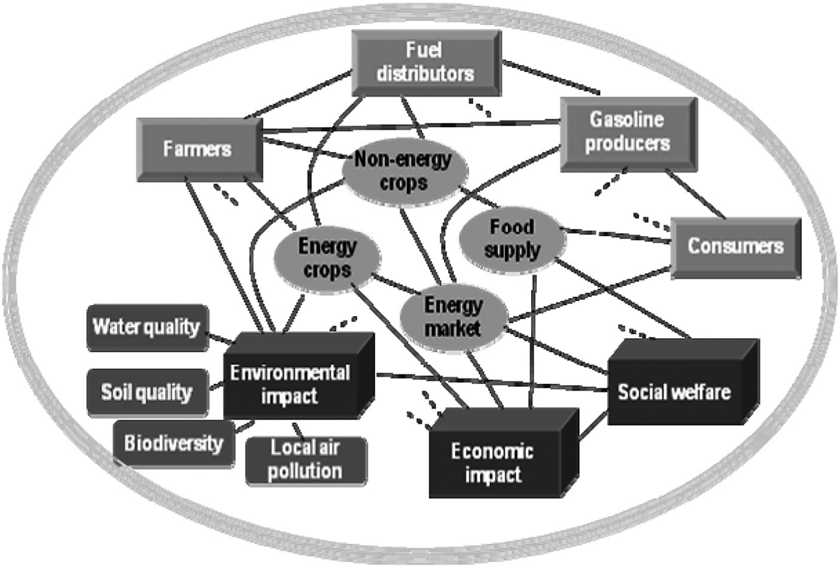


FIGURE 1 Interacting components for biofuel analysis.

gasoline refiners, food producers) and impacts on the environment (e.g., greenhouse gas emissions, water, soil erosion, biodiversity, etc.).

To begin with, the overall impact of biofuels is not well understood. Take, for example, their impact on land use. Traditional life-cycle studies do not take into account emissions from changes in land use, which are difficult to quantify (Seager et al., 2009; Searchinger et al., 2008; Tilman et al., 2009).

Another example is the impact of wind power, a promising renewable energy source that has raised concerns about damage to bird and bat populations. Research will be necessary to provide guidelines for the location of wind farms, especially because most areas with favorable winds are associated with important migratory pathways.

The research challenge is to develop realistic models that capture multiple impacts and interdependencies without imposing strong (unrealistic) assumptions. In traditional approaches, convexity assumptions force unique equilibria, or at the very least, the set of equilibria are themselves convex (Codenotti et al., 2005; Heijungs and Suh, 2002; Ye, 2008). This has made their algorithmic solution possible, but such models do not capture key aspects of systems. Researchers will have to develop more complex decision models through collaboration with resource economists, environmental scientists, and computer scientists.

INDIVIDUAL INTERESTS VS. THE COMMON GOOD

A key issue in environmental policy is balancing individual interests and the common good (e.g., Hardin, 1968). In this area, game-theory models can model the interactions of multiple agents and show the effects of competing interests. In the context of natural resources or climate change on the international level, for example, economic incentives may influence whether a country is motivated to enter an agreement and then abide by it.

Incentive-based policies can also facilitate sustainability challenges on a smaller scale (e.g., the establishment of novel markets for land-conservation activities). To be useful, multi-agent models will have to explore mechanisms and policies for the exchange of goods.

THE RESEARCH CHALLENGES

Research in computational sustainability involves many different areas in computing, information science, and related disciplines. Figure 2 shows some of the areas that are closely related to examples in this article and to the ICS research agenda (ICS, 2010). Figure 3 shows the levels of complexity in computational sustainability, which often addresses large-scale problems based on large volumes of data in highly dynamic and uncertain environments with many interacting components.

Given these complexities, the study of computational sustainability problems

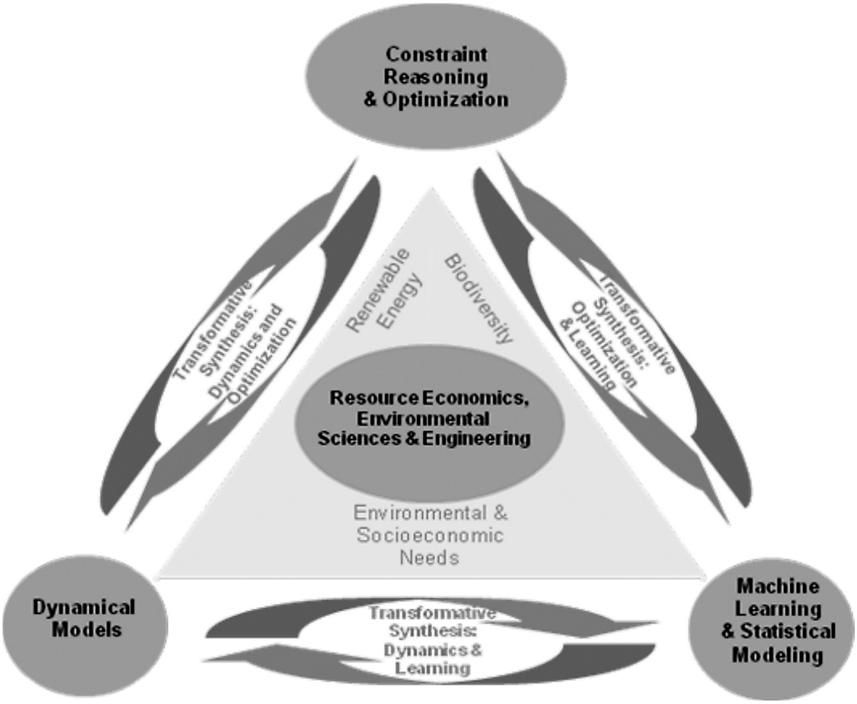


FIGURE 2 Examples of research themes and interactions in computational sustainability that are closely aligned with the research agenda of the Institute for Computational Sustainability at Cornell University.

requires a fundamentally new approach that is unlike the traditional computer science approach (i.e., the science of computation), which is driven mainly by worst-case analyses. From the perspective of computational sustainability, problems are considered "natural" phenomena that are amenable to scientific methodology, rather than purely mathematical abstractions or artifacts. In other words, to capture the structure and properties of complex real-world sustainability problems, principled experimentation is as important as formal models and analysis (Gomes and Selman, 2005, 2007).

SUMMARY

The development of policies for a sustainable future presents unique computational problems in scale, impact, and richness that will create challenges, but also opportunities, for the advancement of the state of the art of computer science and related disciplines. The key research challenges are developing realistic

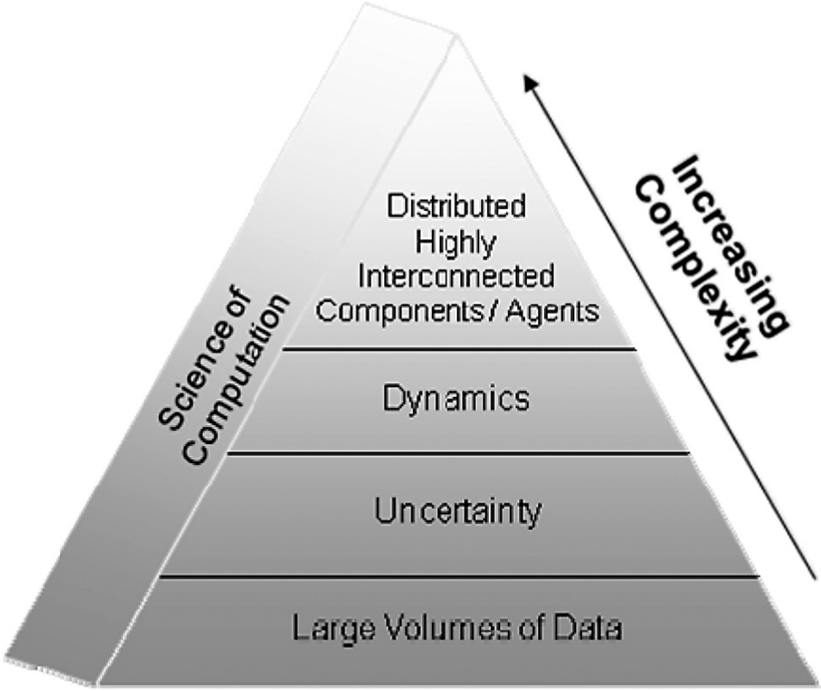


FIGURE 3 Increasing levels of complexity in computational sustainability problems.

computational models that capture the interests and interdependencies of multiple agents, often involving continuous and discrete variables, in a highly dynamic and uncertain environment.

Research in this new field is necessarily interdisciplinary, requiring that scientists with complementary skills work together. In fact, collaboration is an essential aspect of the new science of computational sustainability, an interdisciplinary field that applies techniques from computer science, information science, operations research, applied mathematics, statistics, and related fields to help balance environmental, economic, and societal needs for a sustainable future.

The focus is on developing computational and mathematical models, methods, and tools for making decisions and developing policies concerning the management and allocation of resources for sustainable development. The range of problems encompasses computational challenges in disciplines from ecology, natural resources, economics, and atmospheric science to biological and environmental engineering. Computational sustainability opens up fundamentally new intellectual territory with great potential to advance the state of the art of computer science and related disciplines and to provide unique societal benefits.

ACKNOWLEDGMENTS

The author is the director of the Institute for Computational Sustainability (ICS) at Cornell University and the lead principal investigator of an Expedition in Computing grant on computational sustainability from the National Science Foundation (NSF) (Award Number: 0832782). The author thanks NSF for its support and ICS members for their many contributions to the development of a vision for computational sustainability.

REFERENCES

Akyildiz, I.F., T. Melodia, and K.R. Chowdhury. 2007. A survey on wireless multimedia sensor networks. Computer Networks: The International Journal of Computer and Telecommunications Networking 51(4): 921–960.

Ando, A., J. Camm, S. Polasky, and A. Solow. 1998. Special distributions, land values, and efficient conservation. Science 279(5359): 2126–2128.

Barrett, C., P. Little, and M. Carter, eds. 2007. Understanding and Reducing Persistent Poverty in Africa. Florence, Ky.: Routledge.

Bodik, B., M.P. Armbrust, K. Canini, A. Fox, M. Jordan, and D.A.. Patterson. 2008. A Case for Adaptive Datacenters to Conserve Energy and Improve Reliability. Technical Report No. UCB/EECS-2008-127. Berkeley, Calif.: EECS Department, University of California, Berkeley.

Clark, C. 1976. Mathematical Bioeconomics: The Optimal Management of Renewable Resources. New York: Wiley.

Clark, C. 2006. The Worldwide Crisis in Fisheries. New York: Cambridge University Press.

Codenotti, B., S. Pemmaraju, and K. Varadarajan. 2005. On the polynomial time computation of equilibria for certain exchange economies. Pp. 72–81 in Proceedings of the 16th ACM-SIAM Symposium on Discrete Algorithms. Philadelphia: SIAM.

Conrad, J. 1999. Resource Economics. New York: Cambridge University Press.

Conrad, J., C. Gomes, W.J. van Hoeve, A. Sabharwal, and J. Suter. 2007. Connections in Networks: Hardness of Feasibility versus Optimality. Pp. 16–28 in Proceedings of the Fourth International Conference on the Integration of AI and OR Techniques Constraint Programming, Brussels, Belgium. Berlin, Heidelberg: Springer-Verlag.

Costello, C., S.D. Gaines, and J. Lynham. 2008. Can catch shares prevent fisheries collapse? Science 321(5896): 1678–1681.

Dietterich, T. 2009. Machine Learning in Ecosystem Informatics and Sustainability. Pp. 8–13 in Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, Calif.: IJCAI.

Dilkina, B., and C. Gomes. 2009. Wildlife Corridor Design: Connections to Computer Science. Pp. 3–4 in Proceedings of the Workshop on Constraint Reasoning and Optimization for Computational Sustainability (CROCS-09). Lisbon: CP-09.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudik, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A.Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.McC. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberón, S. Williams, M.S. Wisz, and N.E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2): 129–151.

Ellner, S., and J. Guckenheimer. 2006. Dynamic Models in Biology. Princeton, N.J.: Princeton University Press.

EPA (Environmental Protection Agency) 2007. Report to Congress on Server and Data Center Energy Efficiency. Washington, D.C.: EPA. Available online at: *http://www.energystar.gov/index.cfm?c=prod_development.server_efficiency_study*.

Gomes, C., and B. Selman. 2005. Computation science: can get satisfaction. Nature 435(7043): 751–752.

Gomes, C., and B. Selman. 2007. Science of constraints. Constraint Programming Letters 1: 15–20.

Gomes, C., W.J. van Hoeve, and A. Sabharwal. 2008. Connection in Networks: A Hybrid Approach. Pp. 303–307 in Proceedings of the Fifth International Conference on the Integration of AI and OR Techniques Constraint Programming, Paris, France. Berlin, Heidelberg: Springer Verlag.

Hardin, G. 1968. The tragedy of the commons. Science 162(3859): 1243–1248.

Heal, G., and W. Schlenker. 2008. Sustainable fisheries. Nature 455(7216): 1044–1045.

Heijungs, R., and S. Suh. 2002. The Computational Structure of Life Cycle Assessment (Eco-Efficiency in Industry and Science). Boston: Kluwer Academic Publishers.

Hoke, E., J. Sun, and C. Faloutsos. 2006. InteMon: Intelligent System Monitoring on Large Clusters. Pp. 1239–1242 in Proceeding of Very Large Databases (VLDB'06). Available online at *http://www.vldb.org/conf/2006/p1239-hoke.pdf*.

Luseno, W.K., J.G. McPeak, C.B. Barrett, P.D. Little, and G. Gebru. 2003. Assessing the Value of Climate Forecast Information for Pastoralists: Evidence from Southern Ethiopia and Northern Kenya. World Development 31(9): 1477–1494.

ICS (Institute for Computational Sustainability). 2010. Agenda. Available online at *http://www.cis.cornell.edu/ics/*.

IPCC (Intergovernmental Panel on Climate Change). 2007. Fourth Assessment Report (AR4). Technical Report, United Nations IPCC. Available online at *http://www.ipcc.ch/publications_and_data/publications_and_data_reports.htm#1*.

Kaplan, J., W. Forrest, and N. Kindler. 2008. Revolutionizing the Data Center Energy Efficiency. Technical report produced by McKinsey & Company.

Katz, R.H. 2009. Tech Titans Building Boom. IEEE Spectrum Online. Available online at *http://www.spectrum.ieee.org/feb09/7327*.

Kelling, S., W.M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, and G. Hooker. 2009. Data intensive science: a new paradigm for biodiversity studies. BioScience 59(7): 613–620.

Krause, A., and C. Guestrin. 2009. Optimizing sensing from water to the Web. IEEE Computer Magazine 42(8): 38–45.

Moilanen, A., K. Wilson, and H. Possingham, eds. 2009. Spatial Conservation Prioritization. New York: Oxford University Press.

Onal, H., and R. Briers. 2005. Designing a conservation reserve network with minimal fragmentation: a linear integer programming approach. Environmental Modeling and Assessment 10(3): 193–202.

Patnaik, P., S.R. Marwah, and N. Ramakrishnan. 2009. Sustainable Operation and Management of Data Center Chillers using Temporal Data Mining. In Proceedings of Knowledge Discovery and Data Mining (KDD'09), June 28–July 1, 2009, Paris. France. New York: Association for Computing Machinery.

Phillips, S., M. Dudik, and R.E. Schapire. 2004. A Maximum Entropy Approach to Species Distribution Modeling. P. 83 in Proceedings of the Twenty-First International Conference on Machine Learning, Banff, Alberta, Canada. New York: Association for Computing Machinery.

Polasky, S., E. Nelson, J. Camm, B. Csuti, P. Fackler, E. Lonsdorf, C. Montgomery, D. White, J. Arthur, B. Garberyonts, R. Haight, J. Kagan, A. Starfield, and C. Tobalske. 2008. Where to put things? Spatial land management to sustain biodiversity and economic returns. Biological Conservation 141(6): 1505–1524.

Polastre, J., R. Szewczyk, A. Mainwaring, D. Culler, and J. Anderson. 2009. Analysis of Wireless Sensor Networks for Habitat Monitoring. Pp. 399–424 in Wireless Sensor Networks, edited by C.S. Raghavendra. K.M. Sivalingam, and T. Znati. Dordrecht: Springer Netherlands.

Seager, T., S.A. Miller, and J. Kohn. 2009. Land Use and Geospatial Aspects in Life Cycle Assessment of Renewable Energy. Pp. 1–6 in Proceedings of IEEE International Symposium on Sustainable Systems and Technology. Available online at *http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5156724&isnumber=5156678*.

Searchinger, T., R. Heimlich, R.A. Houghton, F. Dong, A. Elobeid, J. Fabiosa, S. Tokgoz, D. Hayes, and T.H. Yu. 2008. Use of U.S. croplands for biofuels increases greenhouse gases through emissions from land-use change. Science 319(5867): 1238–1240.

Shah, A., V. Carey, C. Bash, and C. Patel. 2008. Energy analysis of data center thermal management systems. Journal of Heat Transfer 130(2): 021401 (10 pages).

Shmoys, D. 2004. The Design and Analysis of Approximation Algorithms: Facility Location as a Case Study. Pp. 85–98 in Trends in Optimization, edited by S. Hosten, J. Lee, and R. Thomas. AMS Proceedings of Symposia in Applied Mathematics, Vol. 61. Providence, R.I.: American Mathematical Society.

Strogatz, S. 1994. Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering. Jackson, Tenn.: Perseus Books Group.

Sullivan, B., C. Wood, M. Iliff, R. Bonney, D. Fink, and K. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142(10): 2282–2292.

Tilman, D., R. Socolow, J. Foley, J. Hill, E. Larson, L. Lynd, S. Pacala, J. Reilly, T. Searchinger, C. Somerville, and R. Williams. 2009. Beneficial biofuels—the food, energy, and environment trilemma. Science 325(5938): 270–271.

Toth, R., C. Barrett, Y. Guo, and C. Gomes. 2009. Pastoral Systems in East Africa. Manuscript in preparation.

UNEP (United Nations Environment Programme). 1987. Our Common Future. Published as annex to the General Assembly document A/42/427, Development and International Cooperation: Environment. Report of the World Commission on Environment and Development. Nairobi, Kenya: UNEP.

UNEP. 2007. Global Environment Outlook 4 (GEO4). Technical Report. Available online at *http://www.unep.org/geo/*.

WBCSD (World Business Council for Sustainable Development). 2008. Energy Efficiency in Buildings: Facts and Trends—Full Report. Available online at *http://www.wbcsd.org/DocRoot/qUjY7w54vY1KncL32OVQ/EEB-Facts-and-trends.pdf*.

Werner-Allen, G., K. Lorincz, M. Welsh, O. Marcillo, J. Johnson, M. Ruiz, and J. Lees. 2006. Deploying a wireless sensor network on an active volcano. IEEE Internet Computing 10(2): 17–25.

Williams, J.C., C.S. ReVelle, and S.A. Levin. 2005. Spatial attributes and reserve design models: a review. Environmental Modeling and Assessment 10(3): 163–181.

Worm, B., E.B. Barbier, N. Beaumont, J.E. Duffy, C. Folke, B.S. Halpern, J.B.C. Jackson, H.K. Lotze, F. Micheli, S.R. Palumbi, E. Sala, K.A. Selkoe, J.J. Stachowicz, and R. Watson. 2006. Impacts of biodiversity loss on ocean ecosystem services. Science 314(5800): 787–790.

Worm, B., R. Hilborn, J.K. Baum, T.A. Branch, J.S. Collie, C. Costello, M.J. Fogarty, E.A. Fulton, J.A. Hutchings, S. Jennings, O.P. Jensen, H.K. Lotze, P.M. Mace, T.R. McClanahan, C. Minto, S.R. Palumbi, A.M. Parma, D. Ricard, A.A. Rosenberg, R. Watson, and D. Zeller. 2009. Rebuilding global fisheries. Science 325(5940): 578–585.

Ye, Y. 2008. A path to the Arrow-Debreu competitive market equilibrium. Mathematical Programming 111(1–2): 315–348.

NANO/MICRO PHOTONICS AND NEW APPLICATIONS

Introduction

NELSON TANSU
Lehigh University
Bethlehem, Pennsylvania

MICHAL LIPSON
Cornell University
Ithaca, New York

Recent successes in the development of optical materials structured on a length scale comparable to the wavelength of light have attracted growing interest from scientists and engineers in many disciplines. Novel concepts of ultrasmall microphotonic devices have the potential to revolutionize communications technologies on all scales, from on-chip data communications in computing to the circuit board level to long-haul communications. Advances in semiconductor nanophotonics and nanostructures have revolutionized solid-state lighting and solar cells, which could have tremendous impacts on future energy applications.

The four presentations in this session provide an overview of photonic technologies and new applications from academic and industrial perspectives. The speakers cover both silicon-based photonics for chip-based applications and III-V-based semiconductors for energy applications.

Optical Antennas: A New Technology That Can Enhance Light-Matter Interactions

LUKAS NOVOTNY
University of Rochester
Rochester, New York

The purpose of optical antennas is to convert the energy of free propagating radiation to localized energy, and vice versa. Although this is similar to what radio wave and microwave antennas do, optical antennas exploit the unique properties of metal nanostructures, which behave as strongly coupled plasmas at optical frequencies. It is hoped that optical antennas can increase the efficiency of light-matter interactions in important applications, such as light-emitting devices, photovoltaics, and spectroscopy.

Electromagnetic antennas, a key enabling technology for devices such as cellular phones and televisions, are mostly used in the radio-wave or microwave regime of the electromagnetic spectrum. At optical frequencies, on the contrary, electromagnetic fields are controlled by re-directing the wave fronts of propagating radiation by means of lenses, mirrors, and diffractive elements. Because this type of manipulation is based on the wave nature of electromagnetic fields, it cannot be used to control fields on the subwavelength scale. In contrast, radio wave and microwave technology predominantly uses antennas to manipulate electromagnetic fields, controlling them on the subwavelength scale and interfacing efficiently between propagating radiation and localized fields.

Recent research in nano-optics and plasmonics has generated considerable interest in optical antennas, and several current studies are exploring ways of translating established radio wave and microwave antenna theories into the optical frequency regime. The introduction of the antenna concept into the optical frequency regime will lead to new technological applications, such as enhancing absorption cross-sections and quantum yields in photo-voltaics, releasing energy efficiently from nanoscale light-emitting devices, boosting the efficiency of photochemical or photophysical detectors, and improving spatial resolution in optical microscopy.

BACKGROUND

The word *antenna* most likely derives from the prefix *an-* (meaning "up") and the Indo-European root *ten-* (meaning "to stretch") (Tucker, 1931; Watkins, 2000). Therefore, from an etymological perspective, an antenna is that which stretches or extends upward (Klein, 1966). Today, we refer to an electromagnetic transmitter or receiver as an antenna, but these were originally called *aerials* in English (Simpson and Weiner, 1989). In 1983, IEEE defined an antenna as a means of radiating or receiving radio waves (IEEE, 1983).

Radio antennas were developed as solutions to a communication problem, whereas optical antennas were developed for use in microscopy. Analogous to its radio wave and microwave counterparts, we define an *optical antenna* as a device designed to efficiently convert free propagating optical radiation to localized energy, and vice versa (Bharadwaj et al., 2009). In the context of microscopy, an optical antenna, which can concentrate external laser radiation to dimensions smaller than the diffraction limit, can effectively replace a conventional focusing lens or objective.

In a letter to Albert Einstein dated April 22, 1928, Edward Hutchinson Synge describes a microscopic method (Figure 1) in which the field scattered from a tiny particle could be used as a light source (Novotny, 2007b). The particle would convert free propagating optical radiation into a localized field that would interact with a sample surface. If we think of the surface as a receiver, the particle can be viewed as an optical antenna. Synge's method was probably inspired by the development of dark-field microscopy, a technique invented at the turn of the twentieth century by Richard Adolf Zsigmondy, an Austrian chemist (Elsevier, 1966).

In 1988, Ulrich Ch. Fischer and Dieter W. Pohl carried out an experiment similar to Synge's proposal, but instead of a solid metal particle, they used a gold-coated polystyrene particle as a local light source (Fischer and Pohl, 1989). They imaged a thin metal film with 320 nanometer (nm) holes and demonstrated a spatial resolution of ~ 50nm. Later, laser-irradiated metal tips were proposed as optical antenna probes for near-field microscopy and optical trapping (Novotny et al., 1997, 1998), and since then various other antenna geometries have been studied (e.g., rods and bowties).

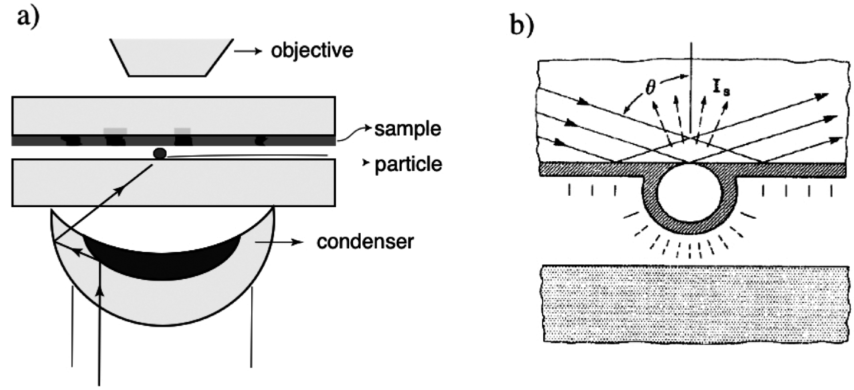


FIGURE 1 (a) Synge's original proposal of near-field optical microscopy based on using scattered light from a particle as a light source. Source: Adapted from Synge's letter to Einstein dated April 22, 1928, cited by Novotny, 2007b. (b) 1988 experiment in which the near-field probe consists of a gold-coated polystyrene particle. Source: Fischer and Pohl, 1989. Reprinted with permission.

HOW OPTICAL ANTENNAS WORK

Although optical antennas are strongly analogous to their radio-frequency (RF) and microwave counterparts, there are crucial differences in their physical properties and scaling behavior. Most of these differences arise because metals are not perfect conductors at optical frequencies, but are strongly correlated plasmas described as a free electron gas. Optical antennas are also not typically powered by galvanic transmission lines; instead, localized oscillators are brought close to the feed point of the antennas, and electronic oscillations are driven capacitively (Pohl, 2000). Moreover, optical antennas can take various unusual forms (e.g., tips or nanoparticles), and their properties may be strongly shape- and material-dependent due to surface plasmon resonances.

Typically, an optical antenna interacts with a receiver or transmitter in the form of a discrete quantum system, such as an atom, molecule, or ion. Because the antenna enhances the interaction between the receiver or transmitter and the radiation field, it may control the light-matter interaction on the level of a single quantum system. On the one hand, the presence of the antenna modifies the properties of the quantum system, such as its transition rates and, in the case of a strong interaction, even the energy-level structure. On the other hand, the properties of the antenna depend on the properties of the receiver/transmitter. Thus, the two must be regarded as a coupled system. The efficiency of the interaction can be expressed in terms of established antenna terminology, such as antenna gain, efficiency, impedance, directivity, and aperture (Bharadwaj et al., 2009).

RADIATION ENHANCEMENT WITH NANOPARTICLE ANTENNAS

A spherical nanoparticle is probably the simplest model antenna (Anger et al., 2006; Bharadwaj and Novotny, 2007; Bharadwaj et al., 2007; Kühn et al., 2006). Although this simple antenna geometry is not very efficient, quantitative comparisons can be made by simple analytical means (Bharadwaj and Novotny, 2007). As shown in the inset of Figure 2a, we can consider a transmitter in the form of a single fluorescent molecule optically pumped by external laser radiation. For weak excitation intensities, the radiation rate Γ_{rad} can be expressed as

$$\Gamma_{\mathrm{rad}} = \Gamma_{\mathrm{exc}}\, \eta_{\mathrm{rad}} \tag{1}$$

where Γ_{exc} is the excitation rate and η_{rad} is the quantum yield. Both Γ_{exc} and η_{rad} depend on the antenna's properties and the separation, z, between antenna and molecule. η_{rad} corresponds to the radiation efficiency, and the rates, Γ_{i}, can be expressed in terms of powers as $P_{\mathrm{i}} = \Gamma_{\mathrm{i}}\, h\nu_{\mathrm{i}}$, with $h\nu_{\mathrm{i}}$ corresponding to the atomic transition energy.

Figure 2a shows the experimentally recorded photon emission rate of a single dye molecule as a function of its separation from an 80nm silver nanoparticle. The superimposed curve is a theoretical calculation based on a simple electromagnetic model in which the molecule is treated as a classical oscillating dipole (Bharadwaj and Novotny, 2007). The data demonstrate that, as the silver particle is brought closer to the molecule, the fluorescence emission rate first increases and then is suppressed. The initial fluorescence enhancement is due to the antenna effect of the silver particle. The excitation rate, Γ_{exc}, increases because of the enhanced local fields near the nanoparticle.

However, for separations shorter than $z = 10$nm, the radiation efficiency η_{rad} decreases rapidly as more and more of the energy is absorbed in the silver nanoparticle. At a distance of $z \sim 3$nm, the rapid decrease of η_{rad} wins over the increase of Γ_{exc}, and the fluorescence of the molecule is quenched. Hence, there is an optimal separation between molecule and antenna.

Figure 2b shows a near-field fluorescence image of single dye molecules dispersed on a flat glass surface. The fluorescence emission rate was recorded pixel by pixel, while the dye sample was raster scanned under a laser-irradiated nanoparticle antenna held at a fixed distance of $z \sim 5$nm above the sample surface by means of a shear-force feedback mechanism (Anger et al., 2006; Höppener et al., 2009; Karrai and Grober, 1995). The resolution achieved in this type of near-field imaging is determined by the antenna size. With an 80nm silver or gold particle, we typically achieve resolutions of ~ 65nm. The different fluorescence patterns in Figure 2b are due to different orientations of the molecular transition dipole axis (Frey et al., 2004; Novotny et al., 2001).

The results of similar experiments performed with other quantum systems,

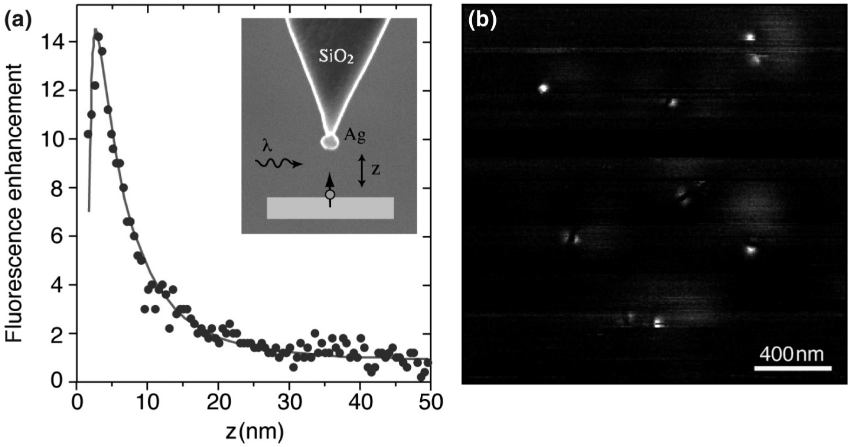


FIGURE 2 Enhancement of the radiation rate of a single molecule with a silver nanoparticle antenna. (a) Normalized fluorescence rate as a function of antenna-molecule separation. Dots are data, and the curve is the result of a theoretical calculation. Inset: scanning electron microscope image of a nanoparticle antenna. The particle is held by a dielectric tip, $\lambda = 488$nm. (b) Fluorescence rate image recorded by raster scanning of a sample with dispersed dye molecules in a plane $z \approx 5$nm underneath a nanoparticle antenna. The different fluorescence patterns are due to different orientations of the molecular transition dipole axis. Source: Adapted from Bharadwaj and Novotny, 2007.

such as quantum dots and carbon nanotubes, are consistent with the results for single fluorescent molecules. An important finding is that for systems with weak intrinsic quantum efficiency (η_i) the radiation efficiency can be enhanced by the optical antenna. In the example discussed here, in which we assumed $\eta_i = 1$, the antenna can only *decrease* the radiation efficiency. However, for poor emitters, such as carbon nanotubes, the antenna can increase the radiation efficiency by more than a factor of 10 (Hartschuh et al., 2005). In general, the lower the η_i, the more the antenna increases the overall efficiency, an effect that was first observed by Wokaun et al. in 1983.

This method of increasing the quantum efficiency of weak emitters might be a promising development that could boost the efficiency of organic light-emitting devices (OLEDs), silicon-based lighting, and solid-state lighting (SSL) in the yellow and green spectral region (Pillai et al., 2007; Wetzel et al., 2004).

The nanoparticle antenna is a model antenna, and its predictions have been tested in various recent experiments. However, much higher efficiencies can be achieved with optimized antenna designs, such as the optical half-wave antenna.

NEAR-FIELD RAMAN SCATTERING

The hallmark of optical antennas, their ability to influence light on the nanometer scale, leads naturally to nano-imaging applications. In the context of nanoscale imaging, an optical antenna represents a near-field optical probe that can interact locally with an unknown sample surface. For a near-field optical image, the optical antenna is guided over the sample surface in close proximity, and an optical response (e.g., scattering, fluorescence, antenna detuning) is detected for each image pixel.

The vibrational spectra provided by Raman scattering define a unique chemical fingerprint for the material under study. Raman scattering involves the absorption and emission of photons, almost identical in energy; thus a nearby antenna can amplify *both* the incoming and outgoing fields. The total Raman scattering enhancement is therefore proportional to the fourth power of the field enhancement (Novotny and Hecht, 2006).

In tip-enhanced Raman scattering (TERS), optical antennas (e.g., metal tips) are used for point-by-point Raman spectroscopy (Hartschuh, 2008; Hartschuh et al., 2003; Stöckle et al., 2000), similar to the original idea of Wessel (1985). Raman enhancements achieved with metal tips are typically in the range of 10^4–10^8, corresponding to field enhancements of 10 to 100.

Our TERS studies are focused on localized states (due to defects and dopants) in carbon nanotubes (Anderson et al., 2005; Maciel et al., 2008). Figure 3 shows (a) the simultaneously recorded topography and (b) near-field Raman image of a single-walled carbon nanotube sample. The image contrast in the near-field

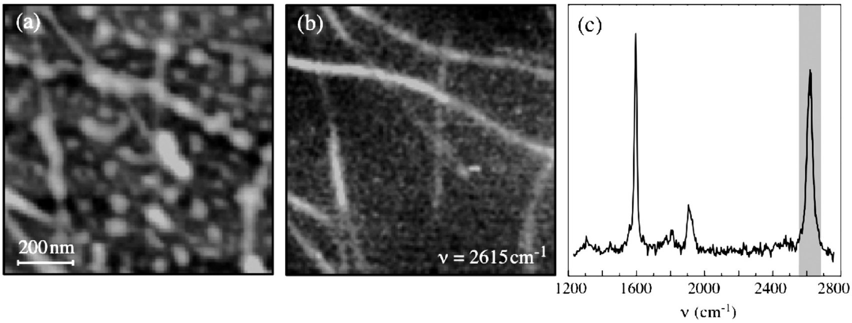


FIGURE 3 Near-field Raman imaging of a single-walled carbon nanotube sample. (a) Topography showing a network of carbon nanotubes covered with small droplets. (b) Raman image of the same sample area recorded by integrating, for each image pixel, the photon counts, which fall into a narrow spectral bandwidth centered around $\nu = 2615\ \mathrm{cm}^{-1}$ (indicated by shading in 3c). (c) Raman scattering spectrum recorded on top of the nanotube. Source: Adapted from Hartschuh et al., 2003.

Raman image (c) is defined by the intensity of the G' line (vibrational frequency of $\nu = 2615\ \mathrm{cm}^{-1}$) highlighted in the spectrum.

WAVELENGTH SCALING

At optical frequencies, electrons in metals have considerable inertia and cannot respond instantaneously to the driving fields. Typically, the skin depth is on the order of tens of nanometers, comparable to the dimensions of the antenna. Traditional design rules that prescribe antenna parameters only in terms of an external wavelength are thus no longer valid. The metal must be rigorously treated as a strongly coupled plasma, which leads to the antenna "seeing" a reduced effective wavelength (Novotny, 2007a). This effective wavelength, λ_{eff}, is related to the external (incident) wavelength, λ, by a surprisingly simple relation

$$\lambda_{\mathrm{eff}} = n_1 + n_2\, [\lambda / \lambda_p] \qquad (2)$$

where λ_p is the plasma wavelength of the metal, and n_1 and n_2 are constants that depend on the geometry and dielectric parameters of the antenna. λ_{eff} is shorter, by a factor of 2 to 6, than the free space, λ, for typical metals (e.g., gold, silver, aluminum) and realistic antenna thicknesses (Bryant et al., 2008; Novotny, 2007a).

The shortening of wavelength from λ to λ_{eff} has interesting implications. For example, it implies that the radiation resistance of an optical half-wave antenna is on the order of just a few Ohms (Alu and Engheta, 2008; Burke et al., 2006; Novotny, 2007a). To see this, we note that the radiation resistance of a thin-wire antenna is roughly $R_{\mathrm{rad}} = 30\, \pi^2\, (L/\, \lambda)^2$, with L being the antenna length. For a half-wave antenna at RF frequencies, $L = \lambda/2$ and $R_{\mathrm{rad}} \sim 73\Omega$. However, for an optical half-wave antenna, $L = \lambda_{\mathrm{eff}}/2$ and hence $R_{\mathrm{rad}} = (30/4)\, \pi^2\, (\lambda_{\mathrm{eff}}/\lambda)^2$. In other words, the radiation resistance at optical frequencies is a factor of $(\lambda_{\mathrm{eff}}/\lambda)^2$ smaller than at RF frequencies. For $\lambda_{\mathrm{eff}} = \lambda/5$ we find $R_{\mathrm{rad}} = 3\ \Omega$.

Figure 4a shows the intensity distribution near a gold half-wave antenna of length L = 110nm and radius R = 5nm resonantly excited at $\lambda = 1170$nm. The effective wavelength is $\lambda_{\mathrm{eff}} = 220$nm. The induced current density $j = i\, \omega\, \varepsilon_o\, [\varepsilon(\omega) - 1]\, E$ evaluated along the axis of the antenna is found to be nearly 180° out of phase with respect to the exciting field.

The notion of an effective wavelength can be used to extend familiar design ideas and rules into the optical frequency regime. For example, the optical analog of the $\lambda/2$ dipole antenna becomes a thin metal rod of length $\lambda_{\mathrm{eff}}/2$. Since λ_{eff} for a silver rod of radius 5nm is roughly $\lambda/5.2$ (Figure 4c), this means that the length of a "$\lambda/2$" dipole antenna is surprisingly small, about $\lambda/10.4$. One can similarly construct antenna arrays like the well established Yagi-Uda antenna developed in the 1920s for the UHF/VHF region (Novotny, 2007a; Taminiau et al., 2008).

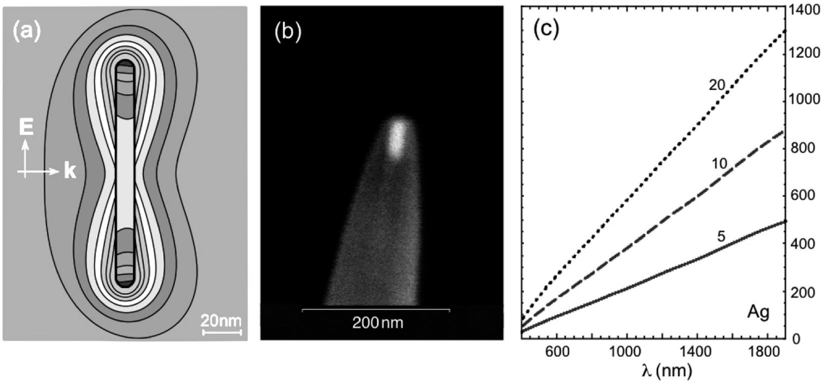


FIGURE 4 Effective wavelength scaling for linear optical antennas. (a) Intensity distribution (E^2, factor of 2 between contour lines) for a gold half-wave antenna irradiated with a plane wave (λ = 1150nm). (b) Scanning electron microscope image of a half-wave antenna resonant at λ = 650nm, fabricated by placing a gold nanorod ~ 65nm long into the opening of a quartz nanopipette. (c) Effective wavelength scaling for silver rods of different radii (5, 10, and 20nm). Source: Novotny, 2007a.

ENHANCED LIGHT-MATTER INTERACTIONS

The localized fields near an optical antenna open up new interaction mechanisms between light and matter, such as higher order multipole transitions and momentum-forbidden transitions. These interactions, which are inaccessible in free space, have the potential to enrich optical spectroscopy and provide new strategies for optical sensing and detection. In free space, the momentum of a photon with energy, E, is $p = E/c$. However, the momentum of an unbound electron with the same energy is two to three orders of magnitude greater, and the photon momentum can be neglected in electronic transitions.

Near an optical antenna, the photon momentum is no longer defined by its free space value. Instead, localized optical fields are associated with a photon momentum defined by the spatial confinement, D, which can be as small as 1 to 10nm. Thus in the optical near field the photon momentum can be drastically increased to a level comparable with the electron momentum, especially in materials with small effective mass, m^*. Hence, localized optical fields can give rise to "diagonal" transitions in an electronic band diagram thereby increasing overall absorption strength, which can be useful for devices such as silicon solar cells. The increase of photon momentum in optical near fields has been discussed in the context of photoelectron emission (Shalaev, 1996) and photoluminescence (Beversluis et al., 2003).

The strong field confinement near optical antennas also has implications for selection rules in atomic and molecular systems. Usually, the light-matter interaction is treated in the dipole approximation where the spatial variation of the fields is much weaker than the spatial variation of quantum wave functions. However, the localized fields near optical antennas give rise to spatial field variations of a few nanometers; hence it may no longer be legitimate to invoke the dipole approximation. This is the case in semiconductor nanostructures, for example, where the low effective mass gives rise to quantum orbitals with large spatial extent.

CONCLUSIONS AND OUTLOOK

Research in the field of optical antennas is currently driven by the need for high field enhancement, strong field localization, and large absorption cross sections. Antennas for high-resolution microscopy and spectroscopy, photovoltaics, light emission, and coherent control are being investigated. In one way or another, optical antennas make processes more efficient or increase the specificity of gathered information.

As in canonical antenna theory, there is no universal antenna design, and optical antennas have to be optimized separately for each application. However, to achieve the highest level of efficiency, the internal energy dissipation of any antenna must be minimized. For a quantum emitter, such as an atom, molecule, or ion, a good antenna yields a low nonradiative decay rate.

New ideas and developments are emerging at a rapid pace, and it is now clear that the optical antenna concept will provide new opportunities for optoelectronic architectures and devices. Today, the building blocks for optical antennas are plasmonic nanostructures that can be fabricated either from the bottom up by colloidal chemistry or from the top down with established nanofabrication techniques, such as electron-beam lithography and focused ion-beam milling. It is also conceivable that future optical antenna designs will draw inspiration from biological systems, such as light-harvesting proteins in photosynthesis.

ACKNOWLEDGMENTS

This work is supported by the U.S. Department of Energy, the National Science Foundation, and the Air Force Office of Scientific Research. I thank all of the students and postdoctoral researchers of the Nano-Optics Group who have contributed to this work.

REFERENCES

Alu, A., and N. Engheta. 2008. Input impedance, nanocircuit loading, and radiation tuning of optical nanoantennas. Physical Review Letters 101(4): 043901.

Anderson, N., A. Hartschuh, S. Cronin, and L. Novotny. 2005. Nanoscale vibrational analysis of single-walled carbon nanotubes. Journal of the American Chemical Society 127(8): 2533–2537.

Anger, P., P. Bharadwaj, and L. Novotny. 2006. Enhancement and quenching of single molecule fluorescence. Physical Review Letters 96(11): 113002.

Beversluis, M.R., A. Bouhelier, and L. Novotny. 2003. Continuum generation from single gold nanostructures through near-field mediated intraband transitions. Physical Review B 68(11): 115433.

Bharadwaj, P., and L. Novotny. 2007. Spectral dependence of single molecule fluorescence enhancement. Optical Express 15(21): 14266–14274.

Bharadwaj, P., P. Anger, and L. Novotny. 2007. Nanoplasmonic enhancement of single-molecule fluorescence. Nanotechnology 18: 044017.

Bharadwaj, P., B. Deutsch, and L. Novotny. 2009. Optical antennas. Advances in Optics and Photonics 1(3): 438–483.

Bryant, G.W., F.J.G. de Abajo, and J. Aizpurua. 2008. Mapping the plasmon resonances of metallic nanoantennas. Nano Letters 8(2): 631–636.

Burke, P.J., S. Li, and Z. Yu. 2006. Quantitative theory of nanowire and nanotube antenna performance. IEEE Transactions on Nanotechnology 5(4): 314–334.

Elsevier. 1966. Nobel Lectures, Chemistry 1922–1941. Amsterdam: Elsevier Publishing Co.

Fischer, U.C., and D.W. Pohl. 1989. Observation of single-particle plasmons by near-field optical microscopy. Physical Review Letters 62(4): 458–461.

Frey, H.G., S. Witt, K. Felderer, and R. Guckenberger. 2004. High-resolution imaging of single fluorescent molecules with the optical near-field of a metal tip. Physical Review Letters 93(20): 200801.

Hartschuh, A. 2008. Tip-enhanced near-field optical microscopy. Angewandte Chemie International Edition 47(43): 8178–8191.

Hartschuh, A., E. Sanchez, X. Xie, and L. Novotny. 2003. High-resolution near-field Raman microscopy of single-walled carbon nanotubes. Physical Review Letters 90(9): 095503.

Hartschuh, A., H. Qian, A.J. Meixner, N. Anderson, and L. Novotny. 2005. Nanoscale optical imaging of excitons in single-walled carbon nanotubes. Nano Letters 5(11): 2310.

Höppener, C., R. Beams, and L. Novotny. 2009. Background suppression in near-field optical imaging. Nano Letters 9(2): 903908.

IEEE (Institute of Electrical and Electronics Engineers Inc.). 1983. Antenna Standards Committee of the IEEE Antennas and Propagation Society, IEEE Std 145-1983. New York: IEEE.

Karrai, K., and R.D. Grober. 1995. Piezoelectric tip-sample distance control for near field optical microscopes. Applied Physics Letters 66(14): 1842–1844.

Klein, E. 1966. Antenna. P. 82 in A Comprehensive Etymological Dictionary of the English Language. Amsterdam: Elsevier Publishing Company.

Kühn, S., U. Hakanson, L. Rogobete, and V. Sandoghdar. 2006. Enhancement of single molecule fluorescence using a gold nanoparticle as an optical nanoantenna. Physical Review Letters 97(1): 017402.

Maciel, I.O., N. Anderson, M.A. Pimenta, A. Hartschuh, H. Qian, M. Terrones, H. Terrones, J. Campos-Delgado, A.M. Rao, L. Novotny, and A. Jorio. 2008. Electron and phonon renormalization at defect/doping sites in carbon nanotubes. Nature Materials 7: 878–883.

Novotny, L. 2007a. Effective wavelength scaling for optical antennas. Physical Review Letters 98(26): 266802.

Novotny, L. 2007b. The History of Near-Field Optics. Pp. 137–180 in Progress in Optics, Vol. 50, edited by E. Wolf. Amsterdam: Elsevier Publishing Co.

Novotny, L., and B. Hecht. 2006. Principles of Nano-Optics. Cambridge: Cambridge University Press.

Novotny, L., R.X. Bian, and X.S. Xie. 1997. Theory of nanometric optical tweezers. Physical Review Letters 79(4): 645–648.

Novotny, L., E.J. Sanchez, and X.S. Xie. 1998. Near-field optical imaging using metal tips illuminated by higher-order Hermite-Gaussian beams. Ultramicroscopy 71(1): 21–29.

Novotny, L., M.R. Beversluis, K.S. Youngworth, and T.G. Brown. 2001. Longitudinal field modes probed by single molecules. Physical Review Letters 86(23): 5251.

Pillai, S., K. Catchpole, T. Trupke, and M. Green. 2007. Surface plasmon enhanced silicon solar cells. Journal of Applied Physics 101(9): 093105.

Pohl, D.W. 2000. Near-field Optics Seen as an Antenna Problem. Pp. 9–21 in Near-field Optics, Principles and Applications, edited by X. Zhu and M. Ohtsu. Singapore: World Scientific.

Shalaev, V.M. 1996. Electromagnetic properties of small-particle composites. Physics Reports 272(2–3): 61–137.

Simpson, J.A., and E.S.C. Weiner. 1989. Antennas. P. 506 in The Oxford English Dictionary. New York: Oxford University Press.

Stöckle, R.M., Y.D. Suh, V. Deckert, and R. Zenobi. 2000. Nanoscale chemical analysis by tip-enhanced Raman spectroscopy. Chemical Physics Letters 318(1–3): 131–136.

Taminiau, T.H., F.D. Stefani, and N.F. van Hulst. 2008. Enhanced directional excitation and emission of single emitters by a nano-optical Yagi-Uda antenna. Optics Express 16(14): 10858–10866.

Tucker, T.G. 1931. Antemna, antenna. P. 19 in A Concise Etymological Dictionary of Latin. Halle/Saale, Germany: Max Niemeyer Verlag.

Watkins, C. 2000. "ten-" P. 90 in The American Heritage Dictionary of Indo-European Roots. Boston: Houghton Mifflin Company.

Wessel, J. 1985. Surface-enhanced optical microscopy. Journal of the Optical Society of America B 2(9): 1538–1541.

Wetzel, C., T. Salagaj, T. Detchprohm, P. Li, and J.S. Nelson. 2004. GaInN/GaN growth optimization for high power green light emitting diodes. Applied Physics Letters 85(6): 866.

Wokaun, A. H.P. Lutz, A.P. King, U.P. Wild, and R.R. Ernst. 1983. Energy transfer in surface enhanced luminescence. Journal of Chemistry and Physics 79(1): 509–514.

Nano-Opto-Mechanics: Using Light Forces in Guided-Wave Nanostructures

MATT EICHENFIELD, RYAN M. CAMACHO, JASPER CHAN, QIANG LIN,
JESSIE ROSENBERG, AMIR H. SAFAVI-NAEINI, AND OSKAR PAINTER
California Institute of Technology
Pasadena, California

Kepler and Newton were the first to propose that light has momentum and can exert a mechanical force. The interaction of light with mechanical vibrations, in the form of Brillouin and Raman scattering, has been known since the 1920s and has many practical applications in the fields of spectroscopy and optoelectronics. With the advent of the laser in the 1960s, it became possible to manipulate micron-scale dielectric particles using optical "tweezers," a technique developed by Art Ashkin (1970). This was also the beginning of the use of laser beams for trapping and manipuling gas-phase atoms, which ultimately led to the demonstration of atomic Bose-Einstein condensates.

More recently, scientists have realized that laser light, with its very low intrinsic noise, can be used to cool a macroscopic mechanical-resonant element, and perhaps reach temperatures suitable for measuring inherently quantum mechanical behavior (Schwab and Roukes, 2005). In addition to the cooling effect, optical amplification from a continuous-wave laser beam can be used to form regenerative mechanical oscillators (Kippenberg and Vahala, 2008). These developments have stimulated interest in the new field of cavity-optomechanics, and a myriad of materials, devices, and techniques are being explored.

It is now widely understood that optical gradient forces, as opposed to the scattering radiation pressure force, can be used in guided-wave nanostructures to generate very large optomechanical couplings to micro- or nano-mechanical motion (Eichenfield et al., 2009a). In contrast to the scattering radiation pressure force, which one can intuit from the reflection of momentum-carrying photons, the gradient force, as its name suggests, results from gradients in the intensity of light, which can be substantial in the near-field.

In a recently developed nanophotonic platform for using this strong gradient optical force, light and sound are manipulated through a common nanostructure (Eichenfield et al., 2009b,c). We call these devices "optomechanical crystals" (OMCs) because of the simultaneous realization of phononic and photonic bandgap states (similar to the electronic bandgaps that form in regular crystalline materials).

OMCs enable the engineering of various integrated functionalities that are not possible in other systems. Ultimately, OMCs may be used in fully integrated planar light and sound wave circuits. In this paper, we introduce the OMC concept and describe a "photon-phonon translator," which may have a variety of applications, from radio-frequency (RF)-over-optical communication to the study of mesoscopic quantum systems.

INTRODUCTION TO OPTOMECHANICAL CRYSTALS

Periodicity applied to the propagation of light gives rise to photonic crystals that can be used to engineer broad- and narrow-band dispersion, to confine optical modes to small volumes with high optical quality factors, and to build planar lightwave circuits (Yablonovitch, 1987). Periodicity applied to mechanical vibrations yields phononic crystals that harness mechanical vibrations the same way photonic crystals harness optical waves (Olson and El-Kady, 2008). The latter raises tantalizing possibilities, such as phononic bandgaps, nonlinear phononics, coherent sources of phonons, and planar sound wave circuits. It has been proposed that periodic structures might be used to simultaneously confine mechanical and optical modes (Maldovan and Thomas, 2006). We endeavor to take this idea one step further by using cavity-optomechanics concepts to marry mechanics and optics in ways that make both more powerful.

The dispersive interaction induced by mechanical motion (Eichenfield et al., 2009c) is responsible for coupling the photonic and phononic crystal properties of the material to yield optomechanical crystals. For the complex motion in these periodic structures, the origin of the optomechanical coupling can be subtle, and in many cases even counter-intuitive. Nonetheless, understanding the nature of the coupling is crucial, because the degree of coupling between different optical-mechanical mode pairs can vary by many orders of magnitude in the same structure. Moreover, subtle changes in the geometry can induce enormous changes in the coupling, which can be used to engineer the coupling if the system is well understood.

A perturbation theory of Maxwell's equations with shifting material boundaries (Johnson et al., 2002) provides a simple, computationally robust method of calculating the optomechanical coupling of these complex motions. We describe below how this theory of perturbation can be used to create an intuitive, graphical picture of the optomechanical coupling of simultaneously localized optical and mechanical modes in periodic systems.

A graphical representation illustrates methodologies for optimizing the coupling of the mechanical and optical modes. Figure 1 shows the band structure for photons and phonons (mechanical vibrations) of a quasi-1D patterned silicon nanobeam. Phonon states can be localized just as localized optical resonances can be formed in "flat-band" regions (regions of low optical dispersion in which the energy velocity of light approaches zero) close to the zone boundaries.

Because the wavelength of the optical mode and the mechanical vibration must be the same (they live on the same 1D lattice), the ratio of the frequency of optical-to-phonon modes can be given as the ratio of the speed of light to that of sound (or more generally whatever type of vibration is involved). It so happens that for the silicon nanobeam shown in Figure 1, which operates at an optical wavelength of 1.5 microns, the mechanical mode frequency is in the 1–5 gigahertz

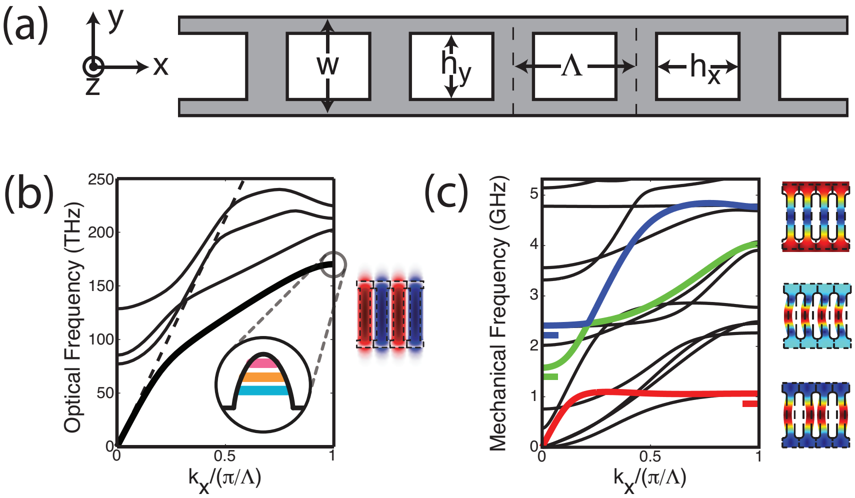


FIGURE 1 (a) General geometry of the structure projection of the periodic nanobeam (infinite structure, no defect). (b) Optical-band diagram of the nanobeam's projection. The band from which all localized optical modes will be derived is shown in thick black, with Ey of the optical mode at the X point shown to the right of the diagram. The harmonic spatial potential created by the defect, along with the first three optical modes are shown as emanating from the X-point band-edge. (c) Mechanical band diagram of the nanobeam's projection. The three bands that form defect modes that are discussed in this paper are colored. The bottom-most mode is from the X point of the red band; the Γ points of the green and blue bands correspond to the middle and top mechanical modes, respectively. The frequencies of the defect modes that form from the band edges are shown as short, horizontal bars. Source: Eichenfield, et al. (2009c). Reprinted with permission. Color figure available online at http://books.nap.edu/openbook.php?record_id=12821&page=59.

(GHz) range (shrinking the width of the nanobeam creates the intriguing possibility of reaching frequencies in the X-band [10–12 GHz] or even higher).

The bands in Figure 1(c) correspond to the bands that have large optomechanical coupling. By forming a "defect" in the periodic lattice through a slow reduction in the hole-to-hole pitch in the center of a patterned beam, one can form localized photon and phonon states (Figure 2). The three primary phononic localized resonances are labeled as pinch, accordion, and breathing modes (see caption).

The optomechanical coupling between mechanical and optical degrees of freedom is given to lowest order by the dispersive term, $g_{\mathrm{OM}} = \delta\omega/\delta\alpha = \omega_c/L_{\mathrm{OM}}$, where ω_c is the unperturbed optical cavity frequency, L_{OM} is an effective length over which a photon's momentum is transferred to the mechanical structure (equal to the cavity length in the case of a Fabry-Perot cavity), and α parameterizes the maximum displacement of the mechanical motion for the mode of interest. Using the perturbation theory of Maxwell's equations with shifting material boundaries (Johnson et al., 2002), one can calculate the derivative of the resonant frequency of a structure's optical modes, with respect to the α-parameterization of a surface deformation perpendicular to the surface of the structure.

An example of the power of the perturbative method to engineer the optomechanical coupling strength, is shown in Figure 3, in which the optomechanical coupling strength of the fundamental accordian mechanical mode at ~1.5 GHz is a function of beam width. The optomechanical coupling length approaches a minimum value close to the wavelength of light for a beam width of 700 nm, whereas it increases to more than 300 times this value for a beam only twice as wide.

THE OPTOMECHANICAL CRYSTAL "TRAVELING PHOTON-PHONON TRANSLATOR"

The OMC concept naturally lends itself to a microchip integration platform for the routing, interaction, and exchange of light and mechanics, with possible applications in photonics and RF-over-optical communication. At the heart of such applications is a device we call the "traveling photon-phonon translator." Shown schematically in Figure 4, this device consists of two optical cavities coupled by a single phonon cavity. An input/output *optical* waveguide is strongly coupled to one of the optical cavities (top cavity, labeled a), as is a *phononic* waveguide to the phononic cavity (labeled b). The second optical cavity (bottom, labeled α_p) is pumped to some large coherent state amplitude via a second optical. For the traveling photon-phonon translator to effectively convert photons to phonons, or vice-versa, several criteria relating to the input and output coupling rates of photons and phonons and internal dissipation rates must be met.

Because of the great flexibility in OMC architecture, however, fulfilling these criteria is rather simple due to the chip-scale platform in which the devices are formed, allowing many degrees of freedom for photon-phonon confinement

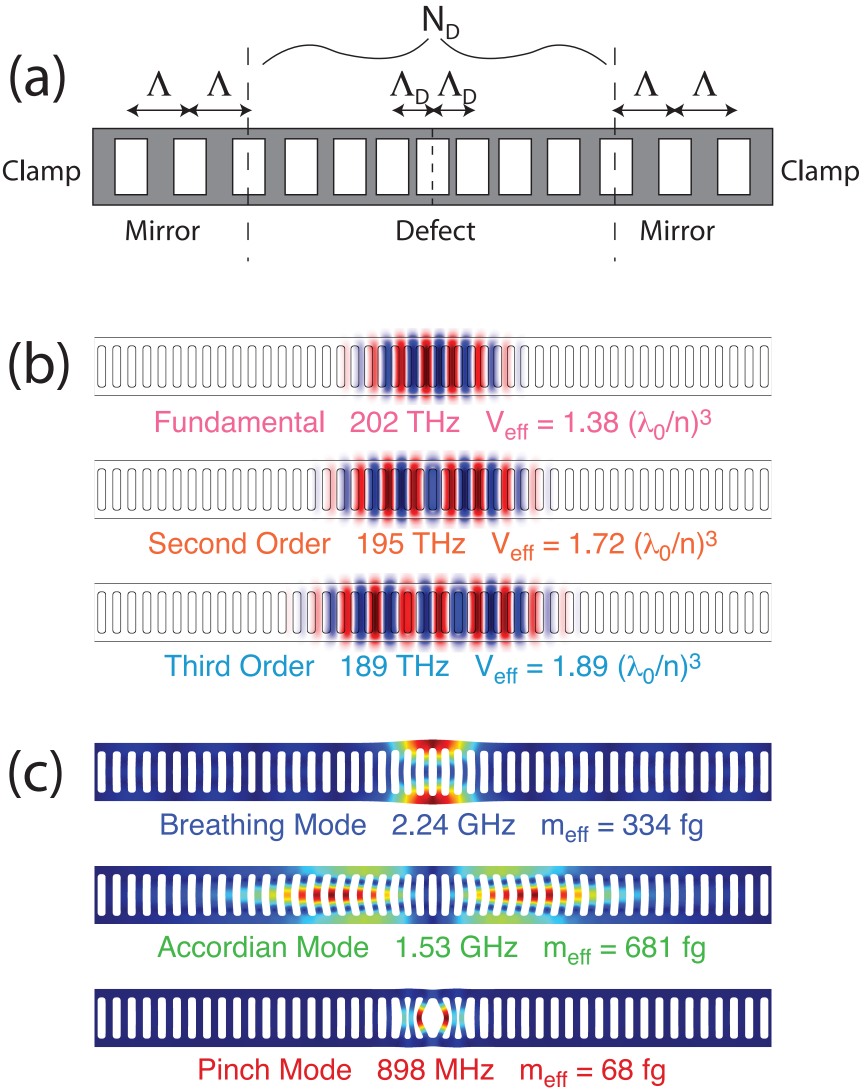


FIGURE 2 (a) Schematic illustration of actual nanobeam optomechanical crystal with defect and clamps at substrate. (b) Localized optical modes of the nanobeam OMC. The colors of the names correspond to the illustration of the inverted potential in Figure 2(b). (c) Localized, optomechanically coupled mechanical modes of the nanobeam OMC. The colors of the names correspond to the colored bands and horizontal bars showing the modal frequencies in Figure 2(c). Source: Eichenfield, et al. (2009c). Reprinted with permission. Color figure available online at http://books.nap.edu/openbook.php?record_id=12821&page=61.

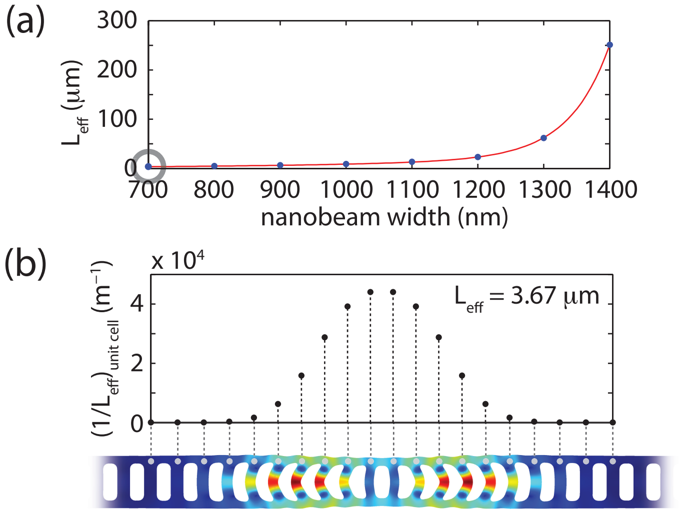


FIGURE 3 For the accordion mode with the fundamental optical mode, (a) the effective length as a function of total beam width, (b) a breakdown of individual unit cell contributions to the total optomechanical coupling for a structure with a beam width of 700 nm (circled in (a)), mode frequency of 3.97 GHz and effective motional mass of 0.334 pg, with accompanying mechanical mode plot. The narrower mechanical mode envelope results in drastically different optomechanical coupling contributions from those with the wider beam. Source: Eichenfield, et al. (2009c). Reprinted with permission.

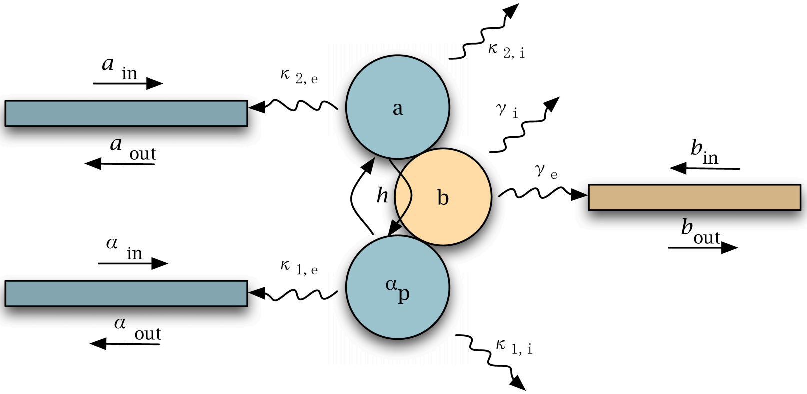


FIGURE 4 Traveling photon-phonon translator. Optical cavities and waveguides are colored blue (a and α_p) while the phononic cavity and waveguide are beige (b). Color figure available online at http://books.nap.edu/openbook.php?record_id=12821&page=62.

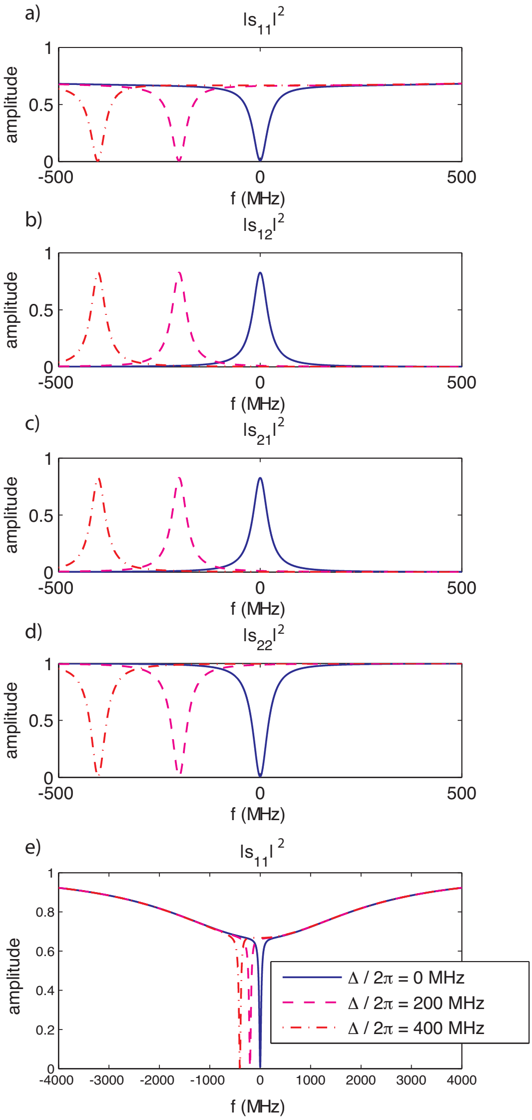


FIGURE 5 Photon-to-phonon scattering matrix amplitudes as a function of laser detuning, Δ. $|s_{11}|$ is the reflected signal from the optical waveguide coupled to cavity a, $|s_{22}|$ is the reflected signal for the phonon input. $|s_{12}|$ and $|s_{21}|$ are the photon-phonon interconversion amplitudes.

and coupling to other modes on the chip. A simulation of the scattering matrix for input and output coupled power (amplitude) for both phonons and photons is shown in Figure 5 for a structure with parameters determined from a numerical modeling of a quasi-2D OMC cavity structure. The efficiency of phonon-to-photon or photon-to-phonon transfer (the system is symmetric, and thus equal for the two conversion efficiencies) can be as high as 75 percent, limited by internal mechanical and optical loss.

These encouraging initial theoretical results indicate that the traveling photon-phonon translator concept can be used to interconvert photons and phonons with high efficiency for applications in optical delay lines (the slower phonon provides the delay), dynamic optical routing/buffering (Lin et al., 2009), and narrow-band RF/microwave filters (Hossein-Zadeh and Vahala, 2008).

Beyond classical RF-microwave photonic applications, the OMC photon-phonon translator works equally well as a quantum translator for individual quanta of phonons or photons if the phononic cavity can be coupled to a low enough bath temperature (100mK). Such a system would be very interesting as a converter of microwave to optical photons when integrated with piezoelectric materials. In the burgeoning field of circuit quantum electrodynamics (Schoelkopf and Girvin, 2008), rapid progress has been made in realizing on-chip coupled qubits via a microwave "quantum bus"; this could enable off-chip coupling via photons for long-distance quantum communication and entanglement between nodes.

REFERENCES

Ashkin, A. 1970. Acceleration and trapping of particles by radiation pressure. Physical Review Letters 24(4): 156–159.

Eichenfield, M., R. Camacho, J. Chan, K.J. Vahala, and O. Painter. 2009a. A picogram- and nanometre-scale photonic crystal optomechanical cavity. Nature 459(7246): 550–555.

Eichenfield, M., J. Chan, R. Camacho, K.J. Vahala, and O. Painter. 2009b. Optomechanical crystals. Nature 462: 78–82.

Eichenfield, M., J. Chan, A.H. Safavi-Naeini, K.J. Vahala, and O. Painter. 2009c. Modeling dispersive coupling and losses of localized optical and mechanical modes in optomechanical crystals. Optics Express, 17(22): 20078–20098.

Hossein-Zadeh, M., and K.J. Vahala. 2008. Photonic RF down-converter based on optomechanical oscillation. IEEE Photonics Technology Letters 20(4): 234–236.

Johnson, S.G., M. Ibanescu, M.A. Skorobogatiy, O. Weisberg, J.D. Joannopoulos, and Y. Fink. 2002. Perturbation theory for Maxwell's equations with shifting material boundaries. Physics Review E 65(6): 066611.

Kippenberg, T.J., and K.J. Vahala. 2008. Cavity optomechanics: back-action at the mesoscale. Science 321(5893): 1172–1176.

Lin, Q., J. Rosenberg, D. Chang, R. Camacho, M. Eichenfield, K.J. Vahala, and O. Painter. 2009. Coherent mixing of mechanical excitations in nano-optomechanical structures. *http://arxiv.org/abs/0908.1128v1*. Submitted to Nature Photonics.

Maldovan, M., and E.L. Thomas. 2006. Simultaneous localization of photons and phonons in two-dimensional periodic structures. Applied Physics Letters 88: 251907.

Olson, R.H. III, and I. El-Kady. 2008. Microfabricated phononic crystal devices and applications. Measurement Science and Technology 20(1): 012002.

Schoelkopf, R.J., and S.M. Girvin. 2008. Wiring up quantum systems. Nature 451(7179): 664–669.

Schwab, K.C., and M.L. Roukes. 2005. Putting mechanics into quantum mechanics. Physics Today 58(July): 36–42.

Yablonovitch, E. 1987. Inhibited spontaneous emission in solid-state physics and electronics. Physical Review Letters 58(20): 2059–2062.

Light-Emitting Diode Technology for Solid-State Lighting

MIKE KRAMES
Soraa, Inc.
Goleta, California.

Light-emitting diode (LED) technology has advanced tremendously since the first demonstration of a practical visible-spectrum LED almost 50 years ago (Holonyak and Bevacqua, 1962). Subsequent LEDs initially used in simple displays (e.g., calculators, watches) and indicator lamps (e.g., clock radios, compact disc players) have been replaced by more powerful, and more sophisticated, devices that produce not only red and green emissions, but also blue and, most important, white. The latter were enabled by the development of the indium-gallium-nitride (InGaN) material system, which was made possible after key breakthroughs in materials technology were made in Japan in the late 1980s (Amano et al., 1986, 1989).

In the early 1990s, efficient blue LEDs based on this material system were demonstrated (Nakamura et al., 1993, 1995), and in combination with a well known yellow-emitting phosphor for scintillators and cathode-ray-tubes, $Y_3Al_5O_{12}:Ce^{3+}$ (“YAG”), these devices demonstrated emission of solid-state white light for the first time (Nakamura and Fasol, 1997). In the late 1990s, Watt-class, high-power LEDs (Höfler et al., 1998) that delivered meaningful levels of light output (from an illumination perspective) were made commercially for the first time.

Since then, InGaN-based LEDs have become more efficient and even more powerful, and the availability of suitable phosphors has increased; today the variety, light output, and quality of LED-based white light has reached the point that it is beginning to unseat conventional lighting technologies in general illumination applications. Indeed, with their high energy efficiency and strong environmental attributes (no lead, no mercury, long operating lifetime), LEDs are certain to be dominant in the future of lighting.

BASICS

In principle LEDs are similar to the simple silicon-based p-n junction diode. Layers of semiconductor material are deposited by an epitaxial method (usually metal-organic chemical vapor deposition [MOCVD]) (Manasevit and Simpson, 1969) on a suitable substrate wafer. The layers are treated (i.e., doped) with extrinsic impurities to form negatively charged (n-type) and positively charged (p-type) regions. The charges induce a built-in electric field at the interface between these regions (the p-n junction).

When sufficient positive external voltage is applied across Ohmic contacts to the p- and n-type regions, the built-in electric field across the p-n junction is reduced, thereby initiating current flow. The current flow is sustained by the recombination of negative charge-carriers (electrons) with positive charge-carriers (holes) in the vicinity of the p-n junction. Each recombination event produces energy approximately equal to the electronic energy band-gap of the semiconductor material at the p-n junction. Since silicon is an *indirect* band-gap semiconductor (Bardeen et al., 1956), electron and hole recombination in silicon requires interaction with the crystal lattice (i.e., the recombination current generates mostly heat).

In other semiconductor materials, especially many III-V compound semiconductor materials, such as gallium arsenide (GaAs), indium phosphide (InP), and gallium nitride (GaN), the transition for an excited electron to the valence band does not require momentum (i.e., lattice interaction), so the released energy is in the form of light. Even in these *direct* band-gap materials, which are used for LEDs, a radiative transition must always compete with crystal lattice imperfections and impurities that produce non-radiative transition pathways. Nevertheless, in very pure material, such as GaAs and InP, the radiative efficiency (internal quantum efficiency) can approach 100 percent. In addition, the external applied voltage is approximately the same as that of the emitted photon. We can thus see that the LED has the potential to generate light with almost 100 percent efficiency, providing a basis for what has been called the "ultimate lamp" (Holonyak, 2000).

Typically, a layer(s) of specific composition is inserted at the p-n junction to allow control over the energy band-gap, and thus over the photon energy (or wavelength), of emission. Today the InGaN-GaN system is used for wavelengths of ~ 365 (ultraviolet, UV-A) to 550 (yellow-green) nanometers (nm). For amber (~ 590 nm) to deep red (~ 650 nm) emission, the most efficient LEDs are based on the (Al,Ga)InP system.

Efficient generation of light alone does not make for an efficient diode. The light must escape the semiconductor crystal into air to be useful. Accomplishing this is less straightforward than one might expect, because the optical refractive indices of most III-V semiconductors is quite high (GaN: n ~ 2.4, InGaP: n ~ 3.5). The high refractive index means that light generated inside the semiconductor must impinge near-normal incidence at the semiconductor/ambient interface in order to escape. Light incident at higher angles is totally internally reflected back into the semiconductor, increasing the chance of absorption (e.g., at metal electrodes, etc.). Various means are used to increase the probably of light extraction, such as chip shaping, texturing, and photonic crystal structures. In the highest performing LEDs today, light extraction efficiency is ~80 percent for InGaN and ~60 percent for AlGaInP (Krames et al., 2007).

For ease of use and to encourage wide adoption in applications, the LED interface for users must be similar to interfaces with other electronic components. Thus means of accessing and contacting (e.g., reflow solderability) the electrodes and, especially for high-power LEDs, removal of heat, are critical factors. The latter is usually accomplished by including a heat-sink element (usually copper) into the primary LED package. Indeed, today's high-power LED packages (Figure 1) have very little resemblance to their "5mm lamp" ancestors, which had small chips

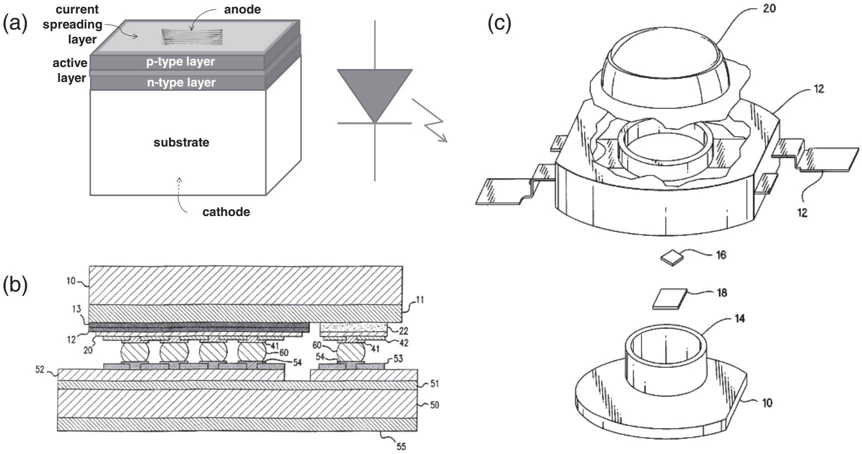


FIGURE 1 (a) Simple LED chip construction (conductive substrate). Typically, the n-type layers are epitaxially deposited, followed by the active layer(s), and then the p-type layers. Ohmic contact electrodes are formed for injecting current into the device. In many cases, current spreading means are employed to provide uniform electrical injection within the thin epitaxial layers. (b) Cross section of a high-power flip-chip LED, showing a considerably more complex structure that is the present requisite for state-of-the-art LEDs for illumination (Krames et al., 2003). (c) Exploded view of a high-power LED package (Carey et al., 2001). High-power chips require sophisticated packaging with good optical efficiency and means for thermal management.

that dissipated less than 100 milliWatts (mW) and produced very little heat. LEDs used today for automotive forward lighting are capable of dissipating up to 10 Watts of power (Dupuis and Krames, 2008).

PERFORMANCE

Sustained improvements in the material quality, diode-layer structure, and overall chip architecture have improved the performance of LEDs dramatically over the last decade or so. Figure 2 shows the best-reported external quantum efficiencies (i.e., the ratio of photons out per electrons injected) for power LEDs for both InGaN and AlGaInP. The best InGaN devices are in the blue-emitting region and have external quantum efficiencies of ~66 percent, meaning that two out of every three electrons injected into the electrical contacts emit a useful photon. Figure 2 also shows the photopic luminosity function, V(λ), which is a measure of the response of the human eye as determined by the Commission Internationale de l'Éclairage (CIE).

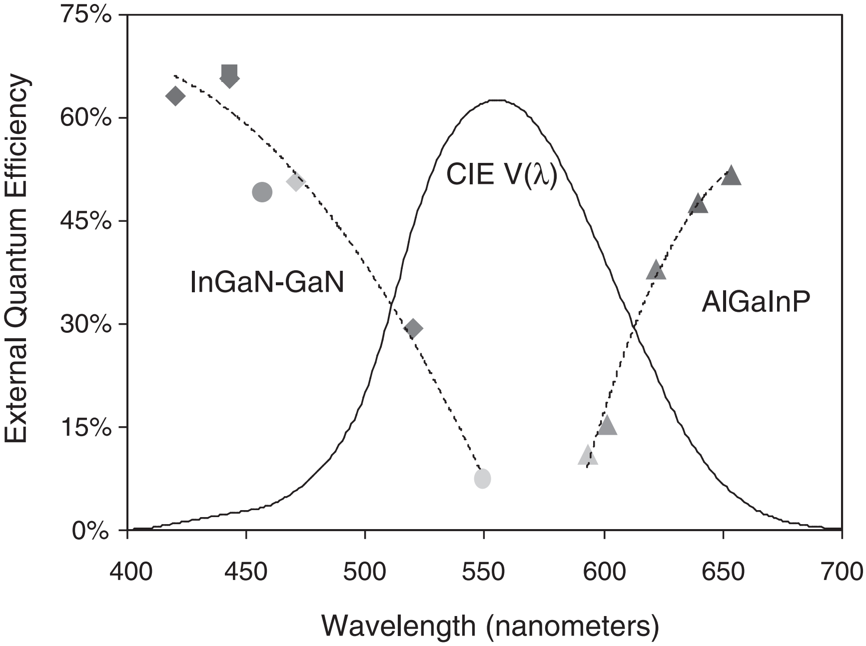


FIGURE 2 Best-reported external quantum efficiencies for high-power InGaN-GaN and AlGaInP LEDs vs. emission wavelength at reasonable operating current densities. Also shown is the human eye responsivity as determined by the photopic luminosity function, V(l), wherein one Watt of optical power at 555 nm corresponds to 683 lumens. Sources: Krames, 1999, 2009; Michiue et al., 2009; Sato et al., 2008, and Vampola et al., 2008.

It is an unfortunate truth that, as Figure 2 shows, the most efficient LED wavelengths are at either side of the visible spectrum (i.e., towards the UV or infra-red). For AlGaInP, the reason for the lower performance at shorter wavelengths is that AlInP is an indirect band-gap semiconductor, and increasing the substitution of Ga by Al for shorter wavelength emission fundamentally reduces the probability of radiative (vs. nonradiative) transitions.

For InGaN, the reason for decreased efficiency at longer wavelengths can be attributed to the miscibility gap between GaN and InN (El-Masry et al., 1998), the increasing strain with higher InN mole fractions, and the fact that this wurtzite (asymmetric) crystal generates polarization-induced built-in electric fields at hetero-interfaces (Bernardini et al., 1997) that perturb the conduction- and valence-band profiles of layer structures and complicate the efficient recombination of electrons and holes.

By working on a non-basal plane of GaN, polarization fields can be reduced, and the 550 nm data point of Figure 2 is, in fact, from a "semi-polar" (11–22) orientation InGaN-GaN LED (Sato et al., 2008) and not the conventional "polar" (0001) orientation. Although considerable improvement is possible (and expected) in the "green gap" region, the present performance of LEDs is nevertheless already very competitive and, in many cases, far superior to, conventional lighting technologies.

WHITE LEDS

Although the combination of separate blue, green, and red LEDs can be tuned to make white light, the most common approach applied in industry is to down-convert blue, violet, or UV light into longer wavelength light by using phosphors, which are excited by the LED primary emission. The phosphors are typically applied in various ways around or on top of the LED chip. The most common method is to use YAG phosphor powder (typically mixed in an organic binder) to overlie a blue LED chip emitting in the range of 440–460nm. The blue LED chip excites the YAG, which produces yellow light. The YAG phosphor powder loading is tuned to allow a precise amount of the primary blue light to "leak" through. When done properly, the combination of the leaked blue light and the yellow phosphor emission yields a white light chromaticity in the 4,000–7,000 Kelvin (K) correlated-color-temperature (CCT) regime with a fairly high conversion efficiency. To generate "warmer" white chromaticities (2,700–4,000K), red phosphors are typically added to the mix (Mueller-Mach et al., 2002).

Figure 3 shows the relative efficacies of white-light generation at ~ 2,900K by: (i) tungsten-filament incandescence; (ii) a tri-phosphor fluorescent lamp (FL); and (iii) an LED with blue-pumped phosphors. The incandescent tungsten (household filament bulb) radiates as a blackbody and at 2,900K generates by far most of its radiation in the infrared range (i.e., heat). The overlap with the visible spectrum is very poor, and convolving the blackbody spectra irradiance with $V(\lambda)$

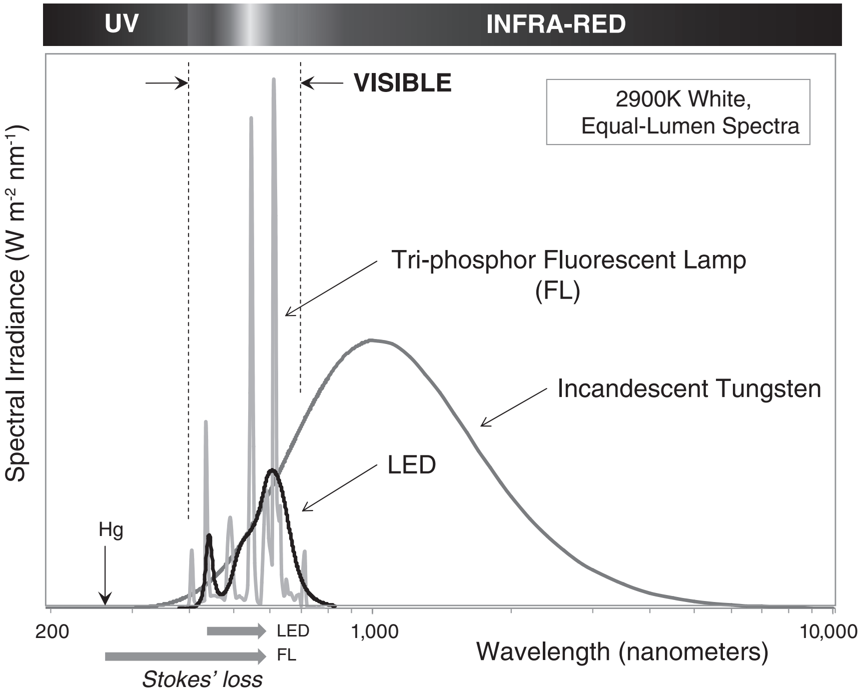


FIGURE 3 Equal-lumen spectra of white light emitters based on (i) incandescent tungsten, (ii) tri-phosphor fluorescence, and (iii) phosphor-converted LEDs, all at a CCT of ~ 2900K. The incandescent bulb radiates most of its energy outside the visible spectrum (400–700nm) as heat. The Stokes energy loss for down-conversion is indicated for the fluorescent lamp (FL) and LED.

gives a maximum luminous efficacy for this source of ~ 16 lumens per electrical Watt (lm/W). In practice, incandescent bulbs perform at a lower level (~ 10–15 lm/W) than this theoretical limit.

The tri-phosphor FL (ii) uses line-emitting phosphors excited by the mercury (Hg) vapor discharge at 254nm. The phosphors are specifically selected so that their emission peaks are in the eye-sensitive region, and indeed the maximum luminous efficacy for the FL spectrum in Figure 3 is quite high, 360 lm/W. However, the enormous Stokes' loss in photon energy from 254nm to ~550–600nm caps the maximum luminous efficacy (in lm/W) to ~ 150 lm/W. In practice, other loss mechanisms also come into play, and typical performance for tube FLs are in the 80–90 lm/W range.

The LED (iii) in Figure 3 uses a blue emitter to pump a combined green and red phosphor mix to obtain the desired CCT of 2,900K. The high intensity of blue light at the chip surface typically requires using fast-decaying phosphors to provide broad emission. The result is an emission spectrum that closely resembles the targeted blackbody curve in the visible-spectrum regime.

One might assume that this provides a more *natural* appearing light, compared to the high-intensity peaks of the FL spectrum, but as of this writing the author is aware of no studies of the effects of smooth vs. spiked spectra on human perception. The smoother spectrum brings a penalty in luminous efficacy, compared to the FL, and the maximum for the LED in this case is 310 lm/W. However, for a blue-pumped LED the Stokes loss is only ~ 20–25 percent, putting the maximum achievable efficacy at ~ 250 lm/W. Thus the obtainable luminous efficacy for the LED is 60–70 percent higher than for FLs and more than 15 times higher than for incandescent lamps.

Figure 4 shows the performance evolution of "warm white" (2,700–4,100K) high-power LEDs and an indication of projected performance from a recent report commissioned by the U.S. Department of Energy (DOE, 2009). Future LED performance is projected to be ~ 160 lm/W by 2018, which is still far from

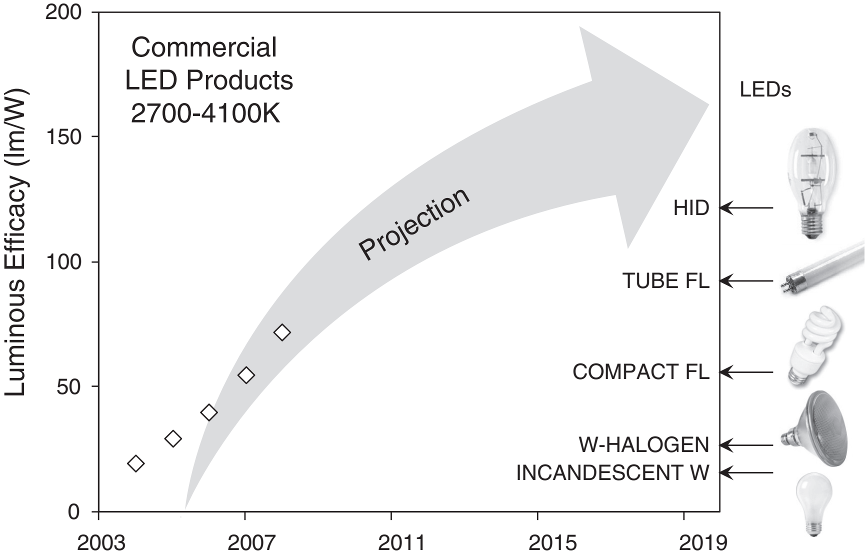


FIGURE 4 Evolution of luminous efficacy (lumens per electrical Watt) for commercial "warm white" (2700–4100K) LED products as well as a projected performance based on information compiled for the U.S. Department of Energy. At right, typical luminous efficacies are indicated for conventional lighting technologies including incandescent tungsten, tungsten-halogen, compact fluorescent, tube fluorescent, and high-intensity-discharge lamps.

the maximum attainable performance described above. Nevertheless, even at the targeted performance level, LEDs would outperform all known technologies for generating white light, including high-intensity-discharge (HID) lamps.

CONCLUSIONS

As LED technology approaches its 50th anniversary, it appears well positioned to penetrate the general lighting market and change the world as we know it. LED-based light sources promise to provide reduced energy consumption, longer operating lifetime (and thus reduced waste), and no generation of materials known to be hazardous to the environment, such as lead or mercury. In addition, the low-voltage drive and fast switching speed of LEDs means lighting for the future could look very different from the lighting we know today. It may include dynamic control features for automatic mood-setting or tuning of intensity and color to increase workforce productivity or simply to elevate people's moods. These additional features, combined with the energy savings and other "green" aspects of LEDs, ensure that LED-based solid-state lighting has a very bright future.

REFERENCES

Amano, H., N. Sawaki, I. Akasaki, and Y. Toyoda. 1986. Metalorganic vapor phase epitaxial growth of a high quality GaN film using an AlN buffer layer. Applied Physic Letters 48(5): 353–355.

Amano, H., M. Kito, K. Hiramatsu, and I. Akasaki. 1989. P-type conduction in Mg-doped GaN treated with low-energy electron beam irradiation (LEEBI). Japanese Journal of Applied Physics Part 2: Letters 28(12): L2112–L2114.

Bardeen, J., F. J. Blatt, and L. H. Hall. 1956. Indirect Transitions from the Valence to the Conduction Bands. Pp. 146–154 in Proceedings of the 1954 Photoconductivity Conference, edited by R.G. Breckenridge, B.R. Russel, and E.E. Hahn. New York: John Wiley and Sons, Inc.

Bernardini, F., V. Fiorentini, and D. Vanderbilt. 1997. Spontaneous polarization and piezoelectric constants of III-V nitrides. Physical Review B 56: R10024–R10027.

Carey, J.A., W.D. Collins, B.P. Loh, and G.D. Sasser. 2001. Surface mountable LED package. U.S. Patent No. 6,274,924.

DOE (U.S. Department of Energy). 2009. Solid-State Lighting Research and Development: Multi-year Program Plan FY'09-FY'15. Report prepared for the U.S. Department of Energy by Navigant Consulting, Inc., Radcliffe Advisors, Inc., and SSLS, Inc.

Dupuis, R.D., and M.R. Krames. 2008. History, development, and applications of high-brightness visible light-emitting diodes. IEEE Journal of Lightwave Technology 26(9): 1154–1171.

El-Masry, N.A., E.L. Piner, S.X. Liu, and S.M. Bedair. 1998. Phase separation in InGaN grown by metalorganic chemical vapor deposition. Applied Physics Letters 72(1): 40–42.

Höfler, G.E., C. Carter-Coman, M.R. Krames, N.F. Gardner, F.A. Kish, T.S. Tan, B. Loh, J. Posselt, D. Collins, and G. Sasser. 1998. Highflux, high-efficiency transparent-substrate AlGaInP/GaP light-emitting diodes. Electronics Letters 34: 1781–1782.

Holonyak, N. Jr. 2000. Is the light emitting diode (LED) an ultimate lamp? American Journal of Physics 68(9): 864–866.

Holonyak, N. Jr., and S.F. Bevaqua. 1962. Coherent (visible) light emission from GaAs P junctions. Applied Physic Letters 1(4): 82–83.

Krames, M.R. 2009. Status and prognosis for solid-state lighting technology. Presentation CM001 at CLEO Conference, Baltimore, Maryland.

Krames, M.R., M. Ochiai-Holcomb, G.E. Höfler, C. Carter-Coman, E.I. Chen, I.H. Tan, P. Grillot, N.F. Gardner, H.C. Chui, J.W. Huang, S.A. Stockman, F.A. Kish, M.G. Craford, T.S. Tan, C.P. Kocot, M. Hueschen, J. Posselt, B. Loh, G. Sasser, and D. Collins. 1999. High-power truncated-inverted-pyramid $(Al_xGa_{1-x})_{0.5}In_{0.5}P/GaP$ light-emitting diodes exhibiting >50% external quantum efficiency. Applied Physics Letters 75(16): 2365–2367.

Krames, M.R., D.A. Steigerwald, F.A. Kish, P. Rajkomar, J.J. Wierer, and T.S. Tan. 2003. III-nitride light-emitting device with increased light generating capability. U.S. Patent No. 6,521,914.

Krames, M.R., O.B. Shchekin, R. Mueller-Mach, G.O. Mueller, L. Zhou, G. Harbers, and M.G. Craford. 2007. Status and future of high-power light-emitting diodes for solid-state lighting. IEEE Journal of Display Technology 3: 160–175.

Manasevit, H.M., and W.I. Simpson. 1969. The use of metalorganics in the preparation of semiconductor materials: I. epitaxial gallium-v compounds. Journal of the Electrochemical Society 116(12): 1725–1732.

Michiue, A., T. Miyoshia, T. Yanamotoa, T. Kozakia, S-I. Nagahamaa, Y. Narukawab, M. Sanob, T. Yamadac and T. Mukai. 2009. Recent development of nitride LEDs and LDs. Proceedings of the SPIE 7216: 561.

Mueller-Mach, R., G.O. Mueller, M.R. Krames, and T. Trottier. 2002. High-power phosphor-converted light-emitting diodes based on III-nitrides. IEEE Journal of Selected Topics in Quantum Electronics 8(2): 339–345.

Nakamura, S., M. Senoh, and T. Mukai. 1993. High-power InGaN/GaN double-heterostructure violet light emitting diodes. Applied Physics Letters 62(19): 2390–2392.

Nakamura, S., M. Senoh, N. Iwasa, and S.I. Nagahama. 1995. High-power InGaN single-quantum-well-structure blue and violet light-emitting diodes. Applied Physics Letters 67(13): 1868–1870.

Nakamura, S., and G. Fasol. 1997. The Blue Laser Diode. Berlin: Springer Verlag.

Sato, H., R.B. Chung, H. Hirasawa, N. Fellows, H. Masui, F. Wu, M. Saito, K. Fujito, J.S. Speck, S.P. DenBaars, and S. Nakamura. 2008. Optical properties of yellow light-emitting diodes grown on semipolar (11-22) bulk GaN substrates. Applied Physics Letters 92: 221110.

Vampola, K.J., N.N. Fellows, H. Masui, S.E. Brinkley, M. Furukawa, R.B. Chung, H. Sato, J. Sonoda, H. Hirasawa, M. Iza, S.P. DenBaars, and S. Nakamura. 2009. Highly efficient broad-area blue and white light-emitting diodes on bulk GaN substrates. Physica Status Solidi A 206(2): 200–202.

ENGINEERING THE HEALTH CARE DELIVERY SYSTEM

Introduction

Stephanie Guerlain
University of Virginia
Charlottesville, Virginia

Eva K. Lee
Georgia Institute of Technology
Atlanta, Georgia

U.S. health care expenditures, $2 trillion in 2005 and 16.2 percent of the U.S. gross domestic product (GDP), are projected to reach $4 trillion by 2015 and 20 percent of GDP. Approximately 31 percent of these costs are administrative, 35 percent are for care of the elderly, 80 percent are related to management of chronic diseases, and 25 percent are related to treating people who engage in risky behaviors. Overall, these expenditures are almost evenly split between public and private funding, and this is expected to continue. Thus, health care expenditures have, and will continue to have, a substantial impact on public spending, private funding, and U.S. competitiveness in the international economy.

The medical and health care industries are fragmented and complex, have multiple stakeholders, and must accommodate dynamic, rapidly changing processes. When compared to other industries, improving health care presents unique challenges. Consider for instance, that one out of 100,000 parcels is misplaced by couriers, but 5 to 10 percent of medical records are reportedly misplaced. Whereas the banking industry has a transaction error rate of 1 in ten million, the error rate in hospital transactions is more than 2 in one hundred (2 percent). The accident rate for airplane landings and takeoffs is on the order of 1 in one million, whereas about 7 in one hundred (7 percent) adverse events are related to the administration of medication.

Because of the enormous cost of care, pressure for optimal decision making on both providers and consumers has grown astronomically, and a variety of engineering tools have been helpful in creating optimal policies for the design and operation of health care delivery systems. As the use of electronic medical records and the availability of data and information increase, we are becoming more aware of how that can be used to help providers *and* patients design and evaluate individual choices of care. However, because of the nascent use of standards for data interchange, privacy and security concerns, and the special interests of insurance, medical, and consumer advocacy groups, a multitude of challenges must be overcome before these tools can be used to their best effect.

The four presentations in this session are focused on health information technology, advances in diagnosis and treatment, patient safety and the detection of adverse events, and effective management of chronic disease. In the first presentation, Elmer Bernstam highlights the status of health information technology (HIT) and describes promising research in biomedical informatics that could potentially improve HIT. Lucila Ohno-Machado, the second speaker, focuses on the technical aspects of calibrating measurements in medical decision support models and the implications of using un-calibrated models to make health care-related predictions.

The third speaker, Genevieve Melton, reviews classes of detection systems for adverse effects, describes some recent advances, and highlights technical challenges. In the fourth presentation, David Dorr describes a care-management model for providing reliable, effective care for older adults with multiple chronic illnesses and preventing unnecessary decline, expensive hospitalizations, and even death.

Why Health Information Technology Doesn't Work

Elmer V. Bernstam and Todd R. Johnson
University of Texas Health Science Center
Houston, Texas

> To improve the quality of our health care while lowering its cost, we will make the immediate investments necessary to ensure that within five years all of America's medical records are computerized. This will cut waste, eliminate red tape, and reduce the need to repeat expensive medical tests. . . it will save lives by reducing the deadly but preventable medical errors that pervade our health care system.
>
> —Barack Obama
> George Mason University, January 8, 2009

Widespread dissatisfaction with health care in America and rapid advancements in information technology have focused attention on information technology, which has dramatically improved efficiency and safety in other industries, as an obvious part of the solution to our health care woes. However, there is increasing evidence that the adoption of health information technology (HIT) will not guarantee comparable benefits in health care.

In fact, unmitigated enthusiasm for HIT may even be dangerous. Similar enthusiasm has repeatedly threatened the field of artificial intelligence (AI), resulting in cycles of excitement and disappointment (referred to as "AI winters"). Motivated by a desire to avoid "HIT winters," we will briefly review the effects of HIT and the "semantic gap," that is, the difference between "health data" and "health information." In addition, we identify significant social and administrative barriers to the adoption of HIT in the context of the technical issues; because

HIT is embedded in a social context, these technical issues must be resolved in a socially and administratively acceptable way. We conclude with research challenges that must be addressed before the full promise of HIT can be realized.

EFFECTS OF HEALTH INFORMATION TECHNOLOGY

HIT is an "easy sell" to an American public increasingly dissatisfied with the U.S. health care system. Indeed, based on some evidence that HIT can improve the quality of health care (Chaudhry et al., 2006), prevent medical errors (Bates et al., 2001), and increase efficiency (Chaudhry et al., 2006), there seem to be some good reasons for optimism.

Unfortunately, many, perhaps most, HIT projects have failed (Littlejohns et al., 2003), and evidence shows that HIT can worsen health care quality in some ways by increasing errors (Koppel et al., 2005; Levenson and Turner, 1993), decreasing efficiency, and perhaps even increasing mortality (Han et al., 2005). The term "e-iatrogenesis" has been coined to describe the unintended deleterious consequences of HIT (Weiner et al., 2007).

Enough negative evidence has accumulated to prompt the Joint Commission (formerly the Joint Commission on Accreditation of Healthcare Organizations) to issue a "Sentinel Event Alert" (defined as "unexpected occurrence[s] involving death or serious physical or psychological injury, or the risk thereof") cautioning health care organizations about potential hazards associated with the implementation and use of HIT (Joint Commission, 2008).

WE'VE BEEN THERE BEFORE: AI WINTERS

During the 1950s, we were faced with a different problem—the cold war. At that time, the government considered IT a promising solution (at least a partial solution) to the problem of tracking Russian communications. It was thought that if researchers could develop automated translation, we would be able to monitor Russian communications and scientific reports in "real time." There was a great deal of optimism about this, and there were ". . . many predictions of fully automatic systems operating within a few years" (Hutchins, 2006).

Although many promising applications were found for the poor-quality automated translations that resulted, the optimistic predictions were not realized. To this day, the fundamental problem of context and meaning remains unsolved, making disambiguation difficult and resulting in some amusing failures. Anecdotal examples include: "The spirit is willing, but the flesh is weak" was translated from English → Russian → English as "The vodka is good, but the meat is rotten," and "out of sight, out of mind" came out as "blind idiot."

In 1966, the influential Automatic Language Processing Advisory Committee (ALPAC) concluded that "there is no immediate or predictable prospect of useful machine translation" (NRC, 1966). As a result, research funding was stopped, and little research was done on automated translation in the United States from 1967 to 1976, when it was revived and supported until 1989 (Hutchins, 2006). Interestingly, disappointment in automated translation in the 1960s was not an isolated event. Similar "AI winters" occurred with respect to connectionism (1970s), expert systems (1990s), and other AI topics.

So, although there is tremendous interest in HIT, and even good evidence that it can be useful, some will certainly be disappointed with the results. A recent report by the National Research Council concluded that ". . . current efforts aimed at the nationwide deployment of health care IT will not be sufficient to achieve the vision of 21st century health care, and may even set back the cause if these efforts continue wholly without change from their present course" (NRC, 2009). Thus, there is also good reason for concern that HIT (and the field of biomedical informatics, in general) may be headed for a bust. However, an "HIT winter" would be unfortunate, because there are real benefits to pursuing research and implementation of HIT.

THE SEMANTIC GAP

Loosely speaking, philosophers who study information draw a distinction between data (syntax) and information, defined as meaningful data (i.e., data + meaning or, alternatively syntax + semantics) (Floridi, 2005). The fundamental problem is that existing technology can store, manipulate, and transmit data but not information. Thus the utility of HIT is limited to the extent to which data approximates meaning, and, unfortunately, there is a large gap between health care data and health care information. Because the difference between data and information is meaning (semantics), we call this the "semantic gap."

Interestingly, Claude Shannon hinted at this issue in 1948 in a seminal paper, "A Mathematical Theory of Communication." This "mathematical theory of communication" came to be known as information theory. Shannon wrote that "[f]requently, the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem." Thus, Shannon's "information theory" explicitly refers to data rather than information in the philosophical sense.

Consider the differences between banking data and health care data, an account at a bank versus a patient's record (Table 1). One difference is that concepts relevant to health are vague compared to banking concepts. The proper interpretation of the symbols relevant to health care requires significant background knowledge. For example, a patient can be "sick" in many ways, including derangements in vital signs (e.g., extremely high or low blood pressure), prognosis associated with a diagnosis (e.g., any patient with myocardial infarction [heart attack] is sick), and other factors. Two clinicians who are asked to describe the same "sick" individual may legitimately focus on different facts or data.

TABLE 1 Comparison of Health "Data" and Banking Data

	Banking Data	**Health Data**
Concepts and descriptions	Precise *Example* Account 123 balance = $15.98	General, subjective *Example* Sick patient
Actions	Usually (not always) reversible *Example* Move money A → B	Often not easily reversible *Example* Give a medication Perform a procedure
Context	Precise, constant, or irrelevant to the task *Example* US $	Vague, variable *Example* Normal lab values differ by lab. No two cells, organs, tumors, or patients are identical.
User autonomy	Well-defined and constrained *Example* What I can do with my checking account = what you can do with yours	Variable and dependent on circumstance *Example* Clinical privileges depend on training, changes over time, and circumstances
Users	Clerical staff, account holder	Varied, including highly trained professionals
Workflow	Well-defined, documented, and explicit	Highly variable, implicit with many undocumented tasks and exceptions

In contrast, the balance in a bank account (e.g., $1,058.93) is relatively objective and is captured by the symbols. If we assume that all transactions (credits and debits) to the account are in the same units (dollars and not pounds or Euros), we need only the numbers and the mathematical operations of addition and subtraction to compute and report the balance. Even though these symbols abstract away the rich semantic complexity of the balance, such as its current purchasing power or that the money can be used to purchase goods and services, this is of no consequence to the successful automation of bank accounts. Thus data-manipulating machines (IT) are much better suited to manipulating bank accounts than they are to manipulating clinical descriptors.

Twenty-five years ago, S. Marsden Blois (1984) argued that the difficulty of using computers in medicine was due to the nature of medical concepts and medical descriptions. Most medical concepts do not have explicit definitions in terms of necessary and sufficient conditions. Thus they are difficult to describe in the formal languages required by computer systems. For instance, a 2000 definition of a myocardial infarction (heart attack) is nine pages long and contains many imprecise terms, such as: "prolonged," "usually," and "experienced observer." Other medical concepts, such as "sharp pain," may be even more difficult to map to formal representations.

SOCIAL AND ADMINISTRATIVE BARRIERS TO THE ADOPTION OF HEALTH INFORMATION TECHNOLOGY

Manipulating data instead of information has many consequences for HIT. The problem for American clinics and hospitals is not usually a shortage of computers. Most hospitals and even small private practices use computers to manage financial and administrative data, and many hospitals have functioning e-mail systems and maintain a Web presence. In addition, many clinicians use personal digital assistants (McLeod et al., 2003), and some communicate with patients via e-mail. In contrast, however, most clinical records are kept on paper.

There are many barriers to the adoption of HIT. Hospitals that have not implemented electronic medical records most frequently cite financial concerns, including the lack of adequate capital for purchasing equipment (74 percent), maintenance costs (44 percent), and unclear return on investment (32 percent) (Jha et al., 2009). Additional barriers include a mismatch between costs and benefits, cultural resistance to change, lack of an appropriately trained workforce to implement HIT, and many others (Hersh, 2004).

To some, clinicians' resistance to computerization appears to be irrational. However, given the mixed evidence regarding the benefits of HIT, caution seems increasingly reasonable. Thus many clinical enterprises are not computerized because of rational skepticism about the costs and benefits of current HIT, not because of an irrational resistance to technological progress.

RESEARCH CHALLENGES

Significant research will be necessary to address serious problems before HIT becomes more attractive to clinicians. Many of these problems are outlined in a recent National Research Council report (NRC, 2009). First, there is a mismatch between what HIT can represent (data) and concepts relevant to health care (data + meaning). This very difficult, fundamental challenge subsumes multiple AI problems (e.g., context or common sense) that have proven very difficult to solve. HIT can manipulate form, but not meaning—hence the term "formal methods." Until we have true information technology, rather than data technology, the benefits of HIT will be limited to applications in which formal methods (i.e., methods that manipulate form) suffice.

A second research challenge is to define appropriate applications for HIT, as well as policies, procedures, and methods of implementation. Clearly, HIT can be helpful in many ways. For example, computerized alerts and reminders can improve compliance with preventive-care guidelines (Shea et al., 1996) and may be cost effective when used in this way (Bernstam et al., 2000). Similarly, examples of reductions in the number of medication errors and other benefits have been published (e.g., Bates et al., 2001). Therefore, in spite of its limitations, current HIT can be useful when applied to suitable problems.

A third research challenge is to evaluate HIT as a clinical intervention. An instructive example is that a commercial electronic health record was associated with increased mortality at one institution (Han et al., 2005), but no such association was found at another institution that implemented the same system in a similar care setting (Del Beccaro et al., 2006). Thus outcomes depend on the interplay among HIT, its implementation at a particular institution, and the nature of the institution (e.g., workflow, patient mix, policies, availability of specialists, etc.).

A computer system cannot be considered in isolation. Its effects must be evaluated in the context of a specific organization. Any system that can affect clinical decisions has the potential to worsen as well as to improve outcomes. Therefore, these systems should be evaluated as clinical interventions, just as drugs, medical devices, and procedures are evaluated.

Fourth, HIT must augment human cognition and abilities. This has been elegantly expressed as the "fundamental theorem of informatics": human + computer > human (Friedman, 2009). In other words, there must be a clear and demonstrable benefit from HIT. Clearly, it can be beneficial in some situations, and in some ways human cognition and computer technology are complementary. Computers excel at precise, efficient manipulation of data, whereas we excel at discovering, storing, and processing meaning. Thus there are tremendous opportunities for effective human-machine collaboration. For example, monitoring (e.g., waveforms) is much easier for computers than for humans. In contrast, reasoning by analogy across domains is natural for humans but difficult for computers.

Defining scenarios, with all relevant parameters, in which HIT is beneficial and demonstrating that using HIT is *reliably* beneficial in these scenarios remains a research challenge. In its present form, HIT will not transform health care the same way that IT has transformed other industries, partly because of the large semantic gap between health data and health information (concepts). In addition, it is worth noting that many problems with health care will only be solved by changes in health care policy, financing, and so forth.

To address the research challenges described above will require unprecedented collaboration among disciplines that have traditionally worked independently and have fundamentally different methods, values, and domains of study. Nevertheless, many promising interdisciplinary approaches have been developed. For example, seemingly simple safety devices, such as checklists, which were pioneered in aviation, have been applied to health care with dramatic results (Pronovost et al., 2006). Statistical process control, simulation, and other engineering methods have also been successfully applied to certain aspects of health care (NAE and IOM, 2005).

HEALTH INFORMATION TECHNOLOGY AND U.S. COMPETITIVENESS

True HIT (i.e., health information technology, not health data technology) is critical for U.S. competitiveness in biomedicine—both for biomedical research and for clinical care. Clinical trials are increasingly being conducted in countries with large populations (i.e., large subject pools) and lower regulatory barriers (compared to those in the United States), such as India (Glickman et al., 2009). Barring substantial changes in our values, privacy concerns, and expectations, we simply cannot compete. For example, because of privacy concerns, the United States has no universal patient identifier. As a result, it is very difficult to identify subjects across clinical trials or patients who move between hospitals and other care settings. In contrast, unique patient identifiers in other countries have greatly facilitated clinical research.

Similarly, the high cost of health care in the United States encourages "medical tourism." Many Americans travel abroad for care that is too expensive for them to obtain in the United States (Wapner, 2008). Some foreign hospitals actually specialize in providing care to Americans who come for a high-cost procedure, such as coronary artery bypass surgery.

True HIT can help address both of these problems. If we can collect clinical information (meaningful data) as a byproduct of routine care, we can then learn from experience, rather than relying solely on clinical trials. In parallel, we can leverage this information to improve care processes. Thus we would fulfill the promise of HIT described by President Obama.

CONCLUSIONS

Clearly we must improve health care in fundamental ways, and HIT will be important in transforming the health care system. However, disappointment seems inevitable, because the promises made on behalf of HIT are not likely to be fully realized in the near future. Historical precedents for such cycles of enthusiasm and disappointment with technology include AI, for which boom and bust cycles appear to be the rule rather than the exception.

Realizing the promise of HIT to improve health care will require an unprecedented level of collaboration among communities that have traditionally had little in common, speak different languages, and have very different world views. Thus we are faced with both challenges and opportunities to find fresh perspectives on fundamental problems in the health care domain. In the process, we may also solve some fundamental information (i.e., computer science) problems related to context and meaning.

REFERENCES

Bates, D.W., M. Cohen, L.L. Leape, J.M. Overhage, M.M. Shabot, and T. Sheridan. 2001. Reducing the frequency of errors in medicine using information technology. Journal of the American Medical Informatics Association 8(4): 299–308.

Bernstam, E.V., H.R. Strasberg, and D.L. Rubin. 2000. Cost-Benefit Analysis of Computer-based Patient Records with Regard to their Use in Colon-Cancer Screening. Presented at the Asia Pacific Medical Informatics Conference, Hong Kong, China. Available online at *http://bmir.stanford.edu/file_asset/index.php/145/BMIR-2000-0847.pdf*.

Blois, M.S. 1984. Information and Medicine: The Nature of Medical Descriptions. Berkeley, Calif.: University of California Press.

Chaudhry, B., J. Wang, S. Wu, M. Maglione, W. Mojica, E. Roth, S.C. Morton, and P.G. Shekelle. 2006. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Annals of Internal Medicine 144(10): 742–752.

Del Beccaro, M.A., H.E. Jeffries, M.A. Eisenberg, and E.D. Harry. 2006. Computerized provider order entry implementation: no association with increased mortality rates in an intensive care unit. Pediatrics 118(1): 290–295.

Floridi, L. 2005. Semantic conceptions of information. In Stanford Encyclopedia of Philosophy. Available online at *http://plato.stanford.edu/entries/information-semantic/*.

Friedman, C.P. 2009. A “fundamental theorem” of biomedical informatics. Journal of the American Medical Informatics Association 16(2): 169–170.

Glickman, S.W., J.G. McHutchison, E.D. Peterson, C.B. Cairns, R.A. Harrington, R.M. Califf, and K.A. Schulman. 2009. Ethical and scientific implications of the globalization of clinical research. New England Journal of Medicine 360(8): 816–823.

Han, Y.Y., J.A. Carcillo, S.T. Venkataraman, R.S. Clark, R.S. Watson, T.C. Nguyen, H. Bayir, and R.A. Orr. 2005. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. Pediatrics 116(6): 1506–1512.

Hersh, W. 2004. Health care information technology: progress and barriers. Journal of the American Medical Association 292(18): 2273–2274.

Hutchins, J. 2006. Machine Translation: History. Pp. 375–383 in Encyclopedia of Language and Linguistics, second edition, edited by K. Brown. Burlington, Mass.: Elsevier.

Jha, A.K., C.M. DesRoches, E.G. Campbell, K. Donelan, S.R. Rao, T.G. Ferris, A. Shields, S. Rosenbaum, and D. Blumenthal. 2009. Use of electronic health records in U.S. hospitals. New England Journal of Medicine 360(16): 1628–1638.

Joint Commission. 2008. Safely Implementing Health Information and Converging Technologies. Available online at *http://www.jointcommission.org/SentinelEvents/SentinelEventAlert/sea_42.htm*.

Koppel, R., J.P. Metlay, A. Cohen, B. Abaluck, A.R. Localio, S.E. Kimmel, and B.L. Strom. 2005. Role of computerized physician order entry systems in facilitating medication errors. Journal of the American Medical Association 293(10): 1197–1203.

Levenson, N.G., and C.S. Turner. 1993. An investigation of the Therac-25 accidents. IEEE Computer 26(7): 18–41.

Littlejohns, P., J.C. Wyatt, and L. Garvican. 2003. Evaluating computerised health information systems: hard lessons still to be learnt. British Medical Journal 326(7394): 860–863.

McLeod, T.G., J.O. Ebbert, and J.F. Lymp. 2003. Survey assessment of personal digital assistant use among trainees and attending physicians. Journal of the American Medical Informatics Association 10(6): 605–607.

NAE and IOM (National Academy of Engineering and Institute of Medicine). 2005. Building a Better Delivery System: A New Engineering/Health Care Partnership, edited by P.P. Reid, W.D. Compton, J.H. Grossman, and G. Fanjiang.. Washington, D.C.: The National Academies Press.

NRC (National Research Council). 1966. ALPAC, Language and Machines: Computers in Translation and Linguistics. Washington, D.C.: National Academy Press.

NRC. 2009. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions, edited by W.W. Stead and H.S. Lin. Washington, D.C.: National Academies Press.

Pronovost, P., D. Needham, S. Berenholtz, D. Sinopoli, H. Chu, S. Cosgrove, B. Sexton, R. Hyzy, R. Welsh, G. Roth, J. Bander, J. Kepros, and C. Goeschel. 2006. An intervention to decrease catheter-related bloodstream infections in the ICU. New England Journal of Medicine 355(26): 2725–2732.

Shannon, C.E. 1948. A mathematical theory of communication. Bell System Technical Journal 27(2): 379–423, 623–656.

Shea, S., W. DuMouchel, and L. Bahamonde. 1996. A meta-analysis of 16 randomized controlled trials to evaluate computer-based clinical reminder systems for preventive care in the ambulatory setting. Journal of the American Medical Informatics Association 3(6): 399–409.

Wapner, J. 2008. American Medical Association provides guidance on medical tourism. British Medical Journal 337: a575.

Weiner, J.P., T. Kfuri, K. Chan, and J.B. Fowles. 2007. "e-Iatrogenesis": the most critical unintended consequence of CPOE and other HIT. Journal of the American Medical Informatics Association 14(3): 387–388; discussion 389.

Calibration in Computer Models for Medical Diagnosis and Prognostication

Lucila Ohno-Machado
University of California, San Diego

Frederic Resnic
Brigham and Women's Hospital and Harvard Medical School
Cambridge, Massachusetts

Michael Matheny
Vanderbilt University
Nashville, Tennessee

Predictive models to support diagnoses and prognoses are being developed in virtually every medical specialty. These models provide individualized estimates, such as a prognosis for a patient with cardiovascular disease, based on specific information about that individual (e.g., genotype, family history, past medical history, clinical findings). Statistical and machine-learning techniques applied to large clinical data sets are used to develop the models, which are used by both health care professionals and patients. However, verification (a critical step in the evaluation of a model) that the probabilities of estimated or predicted events truly reflect the underlying probability for a particular individual is often overlooked.

ASSESSING CALIBRATION

A simplistic type of calibration is calibration-in-the-large or bias. If the outcome is binary (e.g., "0" if a patient is not diseased and "1" if a patient is diseased), the bias corresponds to the average error for the estimates. For example, an estimate of 89 percent for a patient whose outcome is "1" contributes an individual error of 0.11. The average error for all patients is the measure of calibration-in-the-large. Calibration-in-the-large may be appropriate for considering a group of patients, but says little about how calibrated each estimate is. For example, the assignment of the prior probability of an event as the estimate or risk score for every patient, although it would result in a perfectly calibrated-in-the-large model, would provide no individualized information. Hence it would be of limited practical utility for assessing predictive models.

A fundamental problem in evaluating the calibration of a model in a health care setting is the lack of a gold standard against which individual risk estimates can be compared. A gold standard would be based on a sufficient number of exact replicas of the individual, accurately diagnosed or followed without censoring, so that the proportion of observed events would be equal to the "true estimate" for that individual.

Since every individual is unique, meaningful approximations of true probability are only possible for relatively large groups of similar individuals. However, the way the similarity of patient profiles is defined is a critical factor. Currently, calibration is measured by comparing health outcomes in sets of patients with similar estimated risks. That is, given a predicted risk for an individual, a set of neighboring individuals (in the sense of proximity in single dimension of the estimates, or "output space") is assembled and the bias for this set is assessed. This measure of "calibration-in-the-small" is the same as the measure of "calibration-in-the-large," except that it is applied to a smaller set based on individuals who received similar estimates by a given model.

OUTPUT-SPACE SIMILARITY

One of the most widely used indices in assessing calibration of predictive models was developed in the context of logistic regression by Lemeshow and Hosmer (1982). The idea behind the test is simple: if cases are sorted according to their estimated level of risk and the mean estimate for each decile of risk is very close to the proportion of positive cases in the decile, then one cannot reject the hypothesis that the model is correct (Hosmer and Hjort, 2002; Hosmer et al., 1991, 1997). The sum (i.e., the squared differences between the sum of estimates and number of events in each decile divided by the sum of estimates in that decile) for each outcome is reported to follow a χ^2 distribution with 8 degrees of freedom. If $p < 0.05$, we reject the hypothesis that the model fits the data. The H-L-C statistic based on deciles of risk is defined as:

$$C = \sum_{D=0}^{1} \sum_{l=1}^{10} \left[\frac{\left(\pi_{Dl} - O_{Dl}\right)^2}{\pi_{Dl}} \right],$$

where π_{Dl} and o_{Dl} are the sum of estimates in a decile and observed frequencies in the same decile, for cells indexed by group (decile) l and outcome D. Hosmer and Lemeshow showed via simulations that C is approximately distributed as $\chi 2$ with l-2 = 8 degrees of freedom when the fitted model is the correct one and the estimated expected cell frequencies are sufficiently large.

Note that the H-L statistic is model-dependent because the statistic compares the average estimate in each decile of estimated risk with the proportion of events in that decile. To visualize the calibration of a predictive model, it is common to plot the average estimate for groups representing either (1) percentiles of estimated risk against the proportion of events in that group, as described above, or (2) pre-defined ranges of the estimates. The latter is commonly used in clinical predictive models.

INPUT-SPACE SIMILARITY

We described above how output-space similarity can be used to measure calibration in a more refined way. However, output-space similarity is model-dependent and difficult to understand. Similarity at the input-space is much simpler (e.g., calculation of neighborhoods using features obtained directly from data) and may be an equally legitimate way to assess calibration.

We describe a simulation in which we established in advance four tight clusters of "patients" (100 in each cluster) according to two variables, x1 and x2. The purposes of this simulation were to illustrate the H-L "goodness-of-fit" statistic and to check whether differences in calibration can be determined using this statistic. Bi-normal distributions were generated with identical standard deviations (0.1) and centered at (0,0), (0,1), (1,0), and (1,1) for clusters 1 to 4, respectively. The binary outcome for each patient in a cluster was generated from a Bernoulli distribution with probabilities 0.01, 0.4, 0.6, and 0.99 for clusters 1 to 4, respectively. Figure 1 shows the spatial distribution of the clusters. For verification, the four clusters were automatically re-discovered using the Expectation-Maximization algorithm.

The resulting logistic regression model is highly significant. For comparison, we built a neural network with hidden units so that it was capable of finding a non-linear function relating the predictors and outcomes (Figure 1b). An ideal model would assign the true underlying probability for each case (i.e., 0.01, 0.4, 0.6, and 0.99, depending on which cluster the case belonged to). A neural network with enough parameters was able to get closer to that goal than a semi-linear model, such as logistic regression.

Table 1 shows descriptive statistics for the estimates obtained by the two types of models. The H-L-C statistic for the logistic-regression model was 6.43 ($p = 0.59$); hence we would not reject the hypothesis that the model is calibrated. Although the neural network model had a less favorable H-L-C of 11.773 ($p = 0.16$), the overall errors were smaller.

In this example, the neural network provided better approximations of the true underlying probability of the event in clusters 2 and 3, as can be seen in the ranges of estimates in these clusters, as well as in their maximum residuals. However, comparison of H-L-C and the calibration plot (Figure 2) do not indicate that a neural network would be a better model in this case.

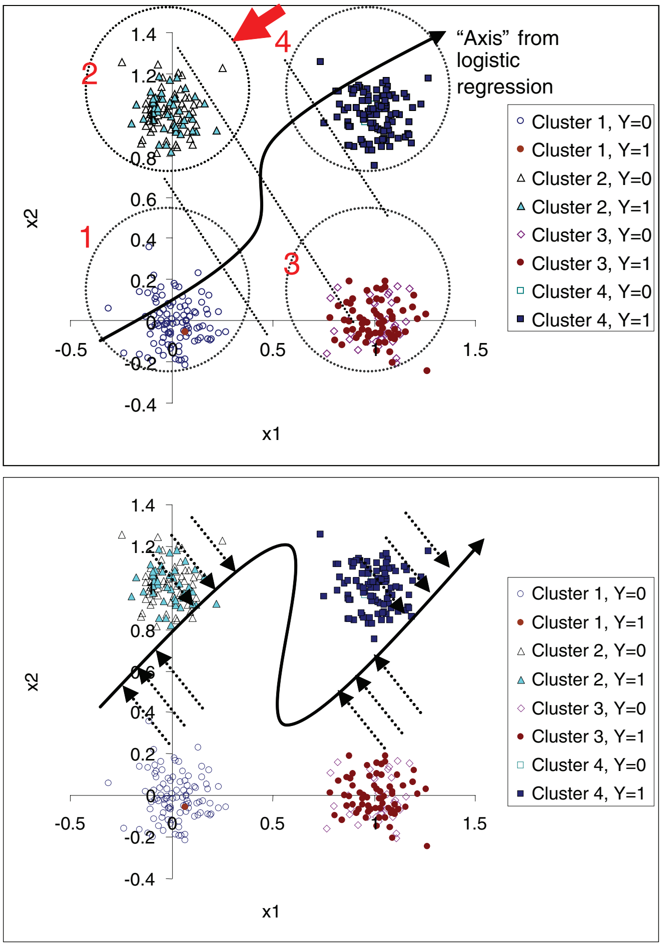


FIGURE 1 Simulation with four predefined non-overlapping bi-normal clusters of individuals with known underlying probability of an event (0.01, 0.40, 0.60, and 0.99 for clusters 1 to 4, respectively). Top panel: Two-dimensional data are projected into one dimension by the logistic-regression model. Dotted diagonal lines divide quartiles of risk. A patient from cluster 2, indicated by the arrow, has an estimate closer to the average estimate for cluster 3 than to the average for cluster 2. Confidence in this estimate should be lower than for a patient in the middle of one of the clusters. The input-space clusters, as opposed to the quartiles of risk, can be explained because patients in cluster 1 have low x1 and low x2, while patients in cluster 2 have low x1 and high x2, and so on. Bottom panel: Projection of the points into an "axis" for neural network estimates. The neural network model comes closer to the true probabilities for clusters 2 and 3 than the logistic-regression model.

TABLE 1 Descriptive statistics of logistic-regression (LR) and neural network (NN) estimates according to input-space clusters. Note that NN estimates do not overlap (i.e., the minimum estimate for cluster 3 is greater than the maximum estimate for cluster 2)

Cluster	Proportion of Events	LR Mean	NN Mean	LR Std Dev	NN Std Dev	LR Minimum	NN Minimum	LR Maximum	NN Maximum
1	0.01	0.0338	0.0219	0.0172	0.0069	0.0066	0.015	0.0949	0.059
2	0.42	0.4129	0.4819	0.1080	0.0207	0.1955	0.431	0.8013	0.584
3	0.64	0.6291	0.6852	0.1149	0.0146	0.2873	0.647	0.8507	0.732
4	0.98	0.9740	0.9908	0.0127	0.0011	0.9301	0.985	0.9954	0.992

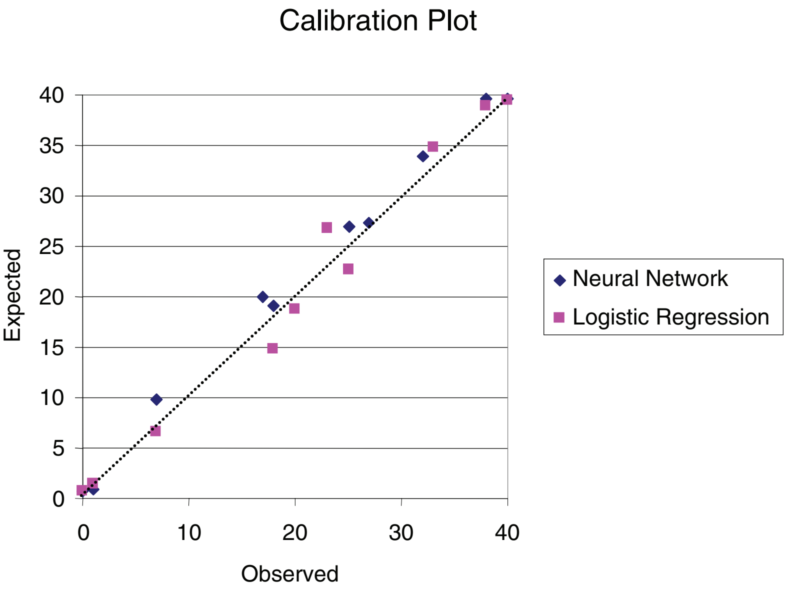


FIGURE 2 Calibration plot for logistic-regression and neural network models based on deciles of risk. There is no apparent superiority of one model over the other.

IMPLICATIONS FOR MEDICAL DECISIONS

In clinical practice, incorrect estimates have significant implications. For example, the widely used clinical practice guideline from the report by the Adult Treatment Panel III (NCEP, 2002) uses cardiovascular risk estimates similar to those available from online calculators to recommend particular treatment regimens. For non-calibrated estimates, this may result in the inappropriate use of medication to manage cholesterol levels.

Computer-based post-marketing tools for the surveillance of new medications and medical devices use models that adjust risk for the population being treated (Matheny et al., 2006). These models depend on the accuracy of the estimates to trigger appropriate alerts for unsafe technologies and drugs. For non-calibrated estimates, risk adjustments may result in a large number of false positives or of false negatives, either of which would incur large costs to the health care system. It is critical, therefore, to assess the calibration of estimates before using models in clinical settings.

We and others have shown, in different domains, that the calibration of medical diagnostic and prognostic models can vary significantly according to the population to which they are applied (Hukkelhoven et al., 2006; Matheny et al., 2005; Ohno-Machado et al., 2006), even though discrimination indices such as the areas under the ROC curves may not vary. Although some efforts are being made to recalibrate models for different populations and study reclassification rates, Web-based calculators that estimate individualized risk do not yet take this issue into account and may present incorrect estimates for particular individuals.

We have proposed methods for taking into account input-space clusters in predictive models (Osl et al., 2008; Robles et al., 2008), but much remains to be done to inform health care workers and the public about the potential shortcomings of this aspect of personalized medicine. As new molecular-based biomarkers for a variety of health conditions are developed and used in multidimensional models to diagnose or prognosticate these conditions, it will become even more important to develop accurate methods of assessing the quality of estimates derived from predictive models.

ACKNOWLEDGMENTS

The authors acknowledge support from the National Library of Medicine, NIH, FDA, and VA grants R01LM009520 (LO), R01 LM008142 (FR), HHSF 223200830058C (FR), VA HSR&D CDA2-2008-020 (MM).

REFERENCES

Hosmer, D.W., and N.L. Hjort. 2002. Goodness-of-fit processes for logistic regression: simulation results. Statistics in Medicine 21(18): 2723–2738.

Hosmer, D.W., T. Hosmer, S. Le Cessie, and S. Lemeshow. 1997. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine 16(9): 965–980.

Hosmer, D.W., S. Taber, and S. Lemeshow. 1991. The importance of assessing the fit of logistic regression models: a case study. American Journal of Public Health 81(12): 1630–1635.

Hukkelhoven, C.W., A.J. Rampen, A.I. Maas, E. Farace, J.D. Habbema, A. Marmarou, L.F. Marshall, G.D. Murray, and E.W. Steverberg. 2006. Some prognostic models for traumatic brain injury were not valid. Journal of Clinical Epidemiology 59(2): 132–143.

Lemeshow, S., and D.W. Hosmer Jr. 1982. A review of goodness of fit statistics for use in the development of logistic regression models. American Journal of Epidemiology 115(1): 92–106.

Matheny, M.E., L. Ohno-Machado, and F.S. Resnic. 2005. Discrimination and calibration of mortality risk prediction models in interventional cardiology. Journal of Biomedical Informatics 38(5): 367–375.

Matheny, M.E., L. Ohno-Machado, and F.S. Resnic. 2006. Monitoring device safety in interventional cardiology. Journal of the American Medical Informatics Association 13(2): 180–187.

NCEP (National Cholesterol Education Program). 2002. Third report of the national cholesterol education program expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (Adult Treatment Panel III) final report. Circulation 106(25): 3143–3421.

Ohno-Machado, L., F.S. Resnic, and M.E. Matheny. 2006. Prognosis in critical care. Annual Review of Biomedical Engineering 8: 567–599.

Osl, M., L. Ohno-Machado, and S. Dreiseitl. 2008. Improving calibration of logistic regression models by local estimates. AMIA Annuual Symposium Proceedings. 2008: pp. 535-539.

Robles, V., C. Bielza, P. Larranaga, S. Gonzales, and L. Ohno-Machado. 2008. Optimizing logistic regression coefficients for discrimination and calibration using estimation of distribution algorithms. TOP: An Official Journal of the Spanish Society of Statistics and Operations Research 16(2): 345–366.

Medical Informatics for Detecting Adverse Events

Genevieve B. Melton
University of Minnesota
Minneapolis, Minnesota

The detection and prevention of adverse events (AEs) in medicine represents a national priority. AEs, defined as injuries that occur in the course of medical management, have important consequences, including higher costs, higher morbidity, and higher mortality. Large-scale focus by the Institute of Medicine (Kohn, 2000) has emphasized most prominently the importance of detecting and preventing AEs to improve patient outcomes. By identifying the AE and analyzing,the context in which it occurred, AE detection can improve system factors and cognitive processes surrounding possible future events and direct resources into *targeted* efforts to prevent AEs in healthcare.

Unfortunately, with traditional voluntary reporting methods, many AEs are not reported. Manual reviews of charts, although effective, are too costly for routine use. Information technology and informatics tools that use data from electronic health records (EHRs) can potentially improve AE detection and have been identified as important tools for creating a "culture of safety."

Several classes of automated AE detection systems have been described, most of which use numeric or coded data from EHRs, such as codes for diagnoses and procedures, records of medication administration, laboratory values, and vital signs. Substantial progress has been made in detecting and preventing adverse drug events (ADEs), particularly with the introduction of electronic prescriptions in both inpatient and outpatient settings. Natural language processing (NLP), a set of automated techniques for converting narrative text into a format appropriate for computer-based analysis, can be used alone or in combination with other automated methods to improve AE detection.

CHALLENGES AND CONSIDERATIONS FOR AUTOMATED SYSTEMS

AE detection can be challenging because of the complexity and lack of standardization of health care EHR systems and associated electronic data, particularly the lack of standards for the quality and formatting of data, standard definitions of AEs, the variable performance of heuristic rule-based systems, and sparse data sets on low-incidence events. Although most automated AE systems provide feedback retrospectively, "active surveillance" systems that alert providers or administrators of events as they occur can potentially identify and investigate AEs much more quickly.

Many robust EHR systems code data for administrative and billing processes (International Disease Classification version 9 codes and Current Procedural Terminology codes), demographics, laboratory results, admission and discharge registrations, medication administration, and computerized physician order entry (CPOE). However, with the exception of coding for administrative processes, even these structured data are often formatted differently for different EHR and hospital systems.

Although some AEs can be found using coded data, a large number of them require supplementary methods and data sources. One reason for this is that administrative and billing data can be incomplete or inaccurate, and they often do not include AEs explicitly. In addition, more sophisticated data necessary for AE detection, such as clinical reasoning, signs and symptoms, clinical summaries, and physical findings are typically not included as structured data.

Standardized definitions of AEs are a fundamental prerequisite for accurate measurement and analysis. However, centralized nomenclatures, or taxonomies, have not been agreed upon for each health care setting. National initiatives will be necessary to reach agreement and bring about consensus. Several promising AE classification systems have been proposed according to setting or discipline, including the JCAHO Patient Safety Event Taxonomy and the Clavien-Dindo Classification of Surgical Complications.

Up to now, rule-based heuristic systems based on data from a variety of sources have been primarily used for AE detection. Although such systems perform well for certain tasks, they rely heavily on "triggers," such as abnormal laboratory values or low blood pressure, to indicate a possible AE. Machine-learning techniques are a promising set of approaches that can help to detect events within datasets. Supervised and semi-supervised techniques are particularly helpful in cases where the AE is not obvious or intuitive, particularly when a large, robust, and well-defined dataset exists to allow for adequate system training and optimized performance.

However, classification systems using machine learning can also perform poorly, particularly for AEs that have a low incidence (< 1 percent), for which data sets may be sparse and unbalanced. Several techniques have been proposed for balancing such data sets, and some have been using sampling techniques with variable success.

Developers of AE detection systems must also be aware of the cost of false negatives and false positives. An important trade-off must be made between the clinical indication and relative cost of screening extra patients to find AEs and the cost of missing AEs. Most AE detection systems are designed to minimize false negatives so as to maximize the overall detection rate; adjunct manual screening is usually used to confirm identified AEs.

ADVERSE DRUG EVENTS: AN EXAMPLE OF IMPROVED DETECTION

Most ADEs occur when drugs are ordered (55 percent), administered (35 percent), transcribed (5 percent), or dispensed (5 percent). In hospitals that use CPOEs, orders for medications and other clinical care are entered directly into the EHR system. CPOE has been the most successful example of an information technology that helps to detect ADEs. Because many CPOE systems now include alerts and reminders about drug prescriptions, they can also prevent many ADEs.

Additional “triggers” in coded data about medication administration or abnormal laboratory values (e.g., supra-therapeutic or sub-therapeutic drug levels, low hemoglobin, or poor renal function) can improve detection or even prevent many ADEs. This has been demonstrated in several clinical trials in both inpatient and outpatient settings.

NATURAL LANGUAGE PROCESSING: A TOOL TO IMPROVE DETECTION

Clinical documents in EHRs are promising data sources for AE detection systems because they often include clinical reasoning, signs and symptoms, clinician’s summaries, and physical findings, all of which may be helpful for AE detection. Although such narratives are rich in content, they present significant challenges to automated systems in the medical domain. Several investigators have tried using “trigger words,” such as “perforation,” “iatrogenic,” or “error,” for AE detection. However, this technique has limited utility because it does not distinguish between a real, potential, or past event or condition.

Serious challenges to using NLP, or medical text-mining, will have to be overcome. Clinical documents are variably formatted with section headers, tabs or other spatial formatting, and transcription errors (i.e., misspellings and grammatical errors). Meaning in medical texts is not straightforward; there are often uncertainties, negations, and questions about timing. In addition, medical terms include synonymy (related or synonymous terms), abbreviations (often redundant), and context-specific meanings.

Several automated text-mining tools have been developed, including open-source tools available through the National Library of Medicine. One widely used proprietary medical NLP application, MedLEE, uses a vocabulary and grammar to extract data from text. Although MedLEE was initially used to extract information from radiographic reports, it has been expanded for application to a wide range of medical texts. When applied to discharge summaries, it has demonstrated a significant improvement in AE detection compared to traditional reporting alone. Thus NLP techniques can potentially improve AE detection systems.

CONCLUSION

Automated AE detection systems with automated informatics techniques have shown promising results for improving the detection and ultimately the prevention of AEs. National initiatives for the adoption of universal EHR systems and advances in informatics techniques for AE detection are likely to increase the use of these systems, which are now widely used only for ADE systems in health care. Addressing technical challenges related to AE nomenclature, machine-learning methods, sampling techniques, and NLP will improve system performance and, ultimately, improve patient safety.

ADDITIONAL READING

Chang, A., P.M. Schyve, R.J. Croteau, D.S. O'Leary, and J.M. Loeb. 2005. The JCAHO patient safety event taxonomy: a standardized terminology and classification schema for near misses and adverse events. International Journal for Quality in Health Care 11(2): 95–105.

Clavien, P.A., J. Barkun, M.L. de Oliveira, J.N. Vauthey, D. Dindo, R.D. Schulick, E. de Santibañes, J. Pekolj, K. Slankamenac, C. Bassi, R. Graf, R. Vonlanthen, R. Padbury, J.L. Cameron, and M. Makuuchi. 2009. The Clavien-Dindo classification of surgical complications: five-year experience. Annals of Surgery 250(2): 187–196.

Handler, S.M., R.L. Altman, S. Perera, J.T. Hanlon, S.A. Studenski, J.E. Bost, M.I. Saul, and D.B. Fridsma. 2007. A systematic review of the performance characteristics of clinical event monitor signals used to detect adverse drug events in the hospital setting. Journal of the American Medical Informatics Association 14(4): 451–458.

Kilbridge, P.M., and D.C. Classen. 2008. The informatics opportunities at the intersection of patient safety and clinical informatics. Journal of the American Medical Informatics Association 15(4): 397–407.

Kohn, L.T., Corrigan, J.M., and Donaldson, M.S., editors. 2000. *To Err is Human: Building a Safer Health System*. Washington, D.C.: National Academy Press.

Melton, G.B., and G. Hripcsak. 2005. Automated detection of adverse events using natural language processing of discharge summaries. Journal of the American Medical Informatics Association 12(4): 448–457.

Managing and Coordinating Health Care: Creating Collaborative, Proactive Systems

David A. Dorr
Oregon Health and Science University
Portland, Oregon

In the last 100 years, huge advances in public health and medical care have resulted in people living longer, healthier lives. These advances have led to a shift from infectious diseases (pneumonia, tuberculosis, and infectious diarrhea) as the top three causes of death to the sequelae of chronic illnesses as the most common causes of death. For instance, heart disease, the most common cause of death in 2000, is hastened by diabetes, hypertension, and high cholesterol (Anderson and Arias, 2003). In addition, as people age, loss of functional ability and increasing disability become primary determinants of increased use of medical services, loss of independence, and death.

It has been shown that increasing disability and multiple conditions near the end of life, rather than single conditions or age alone, are the primary causes of increased hospitalizations and costs (e.g., Shugarman et al., 2009). Nevertheless, the health care system in the United States still focuses on treating individual conditions and meeting acute needs, rather than on ongoing care and overall health. Thus fixing this system will require changing health care delivery to anticipate these needs, teaching and encouraging people and their families to seek help, and providing care that consistently matches medical knowledge.

One possible "fix" is to add care managers—specially trained nurses and social workers who focus on the broad health picture—into primary care clinics. Care managers support a different approach to health care characterized by coordination, prioritization, and protocols for treatment plans assisted by the targeted use of health information technologies (HIT). With these "fixes," the health care system would address ongoing changes in a patient's health as the care manager, supported by HIT, focuses on proactive, collaborative, coordinated care.

COMPLEXITY OF CARE AND THE NEED FOR CARE COORDINATION

In this article, we will consider gaps and potential solutions in the care of a hypothetical elderly patient, Ms. Viera. In this example, we consider two alternate courses in the life of a this hypothetical patient based on differences in care delivery (Figure 1). Ms. Viera is a 75-year-old woman with five common chronic conditions: (1) arthritis in her knees and hips; (2) diabetes, which she has had for the last five years; (3) high blood pressure; (4) moderate kidney problems, which have caused some swelling in her legs; and (5) recent difficulty remembering things day to day. She lives alone and can manage the usual household tasks. Socially, she goes to the senior center once a week, has a part-time professional caregiver, and has a daughter who lives about an hour away by car.

At the start of our hypothetical year, Ms. Viera looks back on last year, during

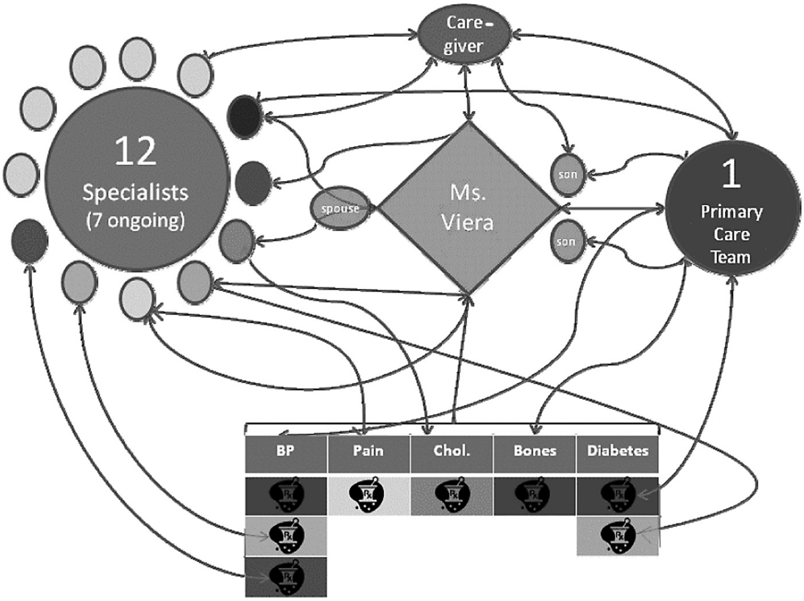


FIGURE 1 An example of average health care and needs for a patient with complex conditions.

which she saw 13 providers, 8 of whom she continues to see regularly. Her regular providers include her primary care provider (PCP), Dr. Smith, a doctor of internal medicine who provides ongoing care with a team of specialists—a rheumatologist for arthritis; a cardiologist; a neurologist, whom she saw in consult for memory loss; a nephrologist for kidney problems and high blood pressure; an orthopedist for her knees; a gynecologist; and an endocrinologist for diabetes. She filled 50 different prescriptions for 8 chronic medications and 4 short-term medications; several of these medications were prescribed by specialists, and some were prescribed by the PCP. She avoided the hospital last year, despite having nearly 90 times the risk of a hospitalization for someone her age with no chronic illnesses (Wolff et al., 2002).

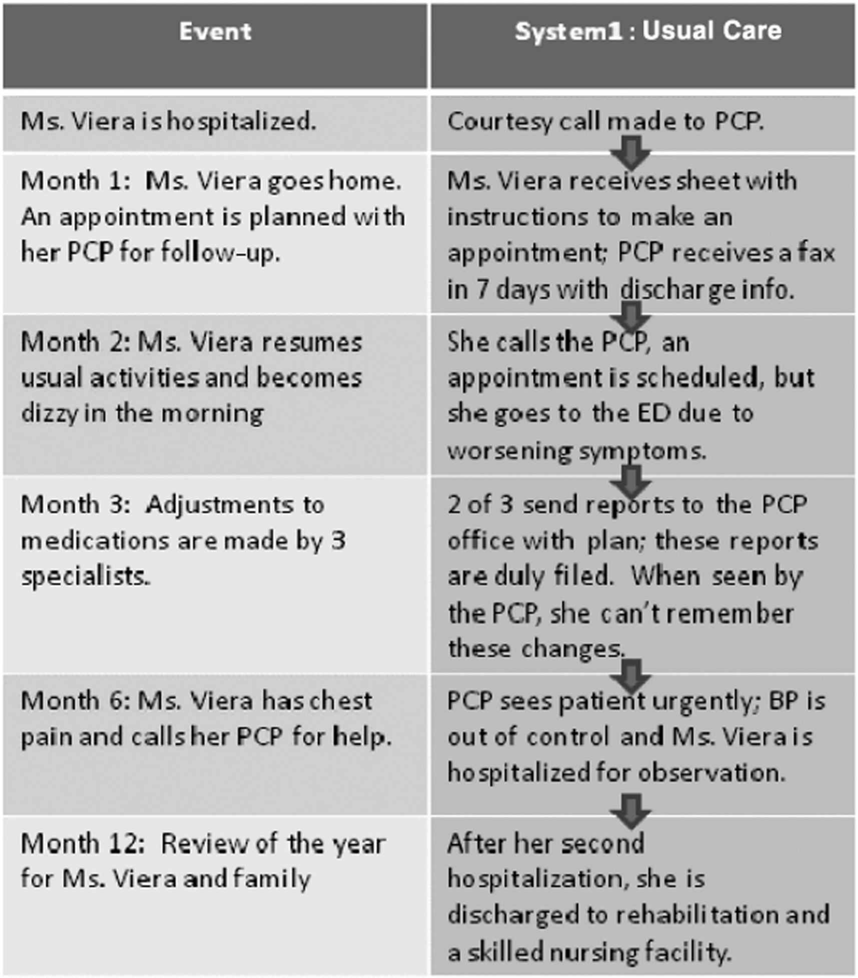
As Figure 1 shows, an enormous number of connections had to be tracked by the patient, the family, and the caregiver—for communication and for changes to medical treatment plans. The coordination of these connections is the primary challenge we address in this paper.

Patients like Ms. Viera represent approximately 5 percent of people over 65, yet they use about 43 percent of all health care resources (Wolff et al., 2002). When we consider re-engineering the system to improve Ms. Viera's care, we must first and foremost consider the benefit of those changes to her. A major hypothesis in care-coordination research is that carefully planning and arranging care can result in higher quality, more efficient care. Society, patients, and insurers will all benefit by avoiding waste from errors and "defects" in the care delivered.

A CRUCIAL JUNCTURE

Let us return to the case of Ms. Viera. At the beginning of our year, she is hospitalized briefly for difficulty breathing and dizziness. After about two days, she is diagnosed with out-of-control blood sugars and some excess fluid on her lungs. The hospital team stabilizes her by adjusting several medications, and she is discharged back to her home.

Figure 2 shows one potential course her post-hospital convalescence could take. Based on the hypothetical events listed on the left, in the next year, she goes home, appointments are planned, she attempts to resume her usual activities, sees specialists, has dizziness, chest pain, and some difficulties with control of her chronic conditions. In the usual course of care, "System 1" (on the right), the care-coordination tasks and their method of completion are highlighted. Studies show that upon discharge from the hospital, one-third of patients have care plans that are not followed or communicated (e.g., instructions to make an appointment with a physician). In addition, calls from the hospital staff to the PCP, although helpful, frequently do not lead to follow-up unless there is further communication—from either the hospital or primary-care team—directly with the patient. By the time the provider reviews the faxed discharge summary, Ms. Viera has about a 10 percent chance of being rehospitalized.



Event	System1 : Usual Care
Ms. Viera is hospitalized.	Courtesy call made to PCP.
Month 1: Ms. Viera goes home. An appointment is planned with her PCP for follow-up.	Ms. Viera receives sheet with instructions to make an appointment; PCP receives a fax in 7 days with discharge info.
Month 2: Ms. Viera resumes usual activities and becomes dizzy in the morning	She calls the PCP, an appointment is scheduled, but she goes to the ED due to worsening symptoms.
Month 3: Adjustments to medications are made by 3 specialists.	2 of 3 send reports to the PCP office with plan; these reports are duly filed. When seen by the PCP, she can't remember these changes.
Month 6: Ms. Viera has chest pain and calls her PCP for help.	PCP sees patient urgently; BP is out of control and Ms. Viera is hospitalized for observation.
Month 12: Review of the year for Ms. Viera and family	After her second hospitalization, she is discharged to rehabilitation and a skilled nursing facility.

FIGURE 2 A year in the life of Ms. Viera with the usual system of health care.

In the next month, Ms. Viera may increase her activities and develop symptoms from her medications. In the usual system, she may call her PCP and, while waiting for the return call, her symptoms may worsen, and she may go to the emergency department. Upon seeing three of her specialists in follow-up, the lack of information sharing among settings and caregivers leads to new medications being prescribed but not remembered by the patient or family and not reconciled against her old list.

Finally, in month six, Ms. Viera may have a serious new problem—worsening chest pain. In the usual system, all of her other issues may have distracted the primary-care team from controlling her blood pressure, which leads to a repeat hospitalization for monitoring. Although she does not have a heart attack, the changes in her medications and the unfamiliarity of her surroundings in the hospital may lead to a fall and a further need for rehabilitation.

In each of these common scenarios, gaps in coordination lead to increased use of health care and worse health for Ms. Viera. Our primary purposes in this paper are to elaborate on the reasons for our failure to create reliable health systems and to provide suggestions for improvement.

DEVELOPING A HEALTH CARE SYSTEM OF SYSTEMS

One way to change the current system of coordination "as usual"—which includes many gaps—is to think of health care as a more reliable and effective system of systems. One challenge in creating a reliable system is that gaps are not uniform; they vary over time and from individual to individual based on a wide range of factors, such as social needs, economic conditions, chronic illnesses, personal preferences, and local system infrastructure. Multiple disciplines, such as cognitive engineering, systems science, industrial engineering, and informatics, must be combined to begin to minimize and then close these gaps.

One way to address these problems is to take a close look at the existing health care delivery system and diagnose the gaps through a structured approach by looking at goals of care, current processes, infrastructure, and participants. We have completed a series of studies of the system of primary care, a subset of the overall health care system that focuses on ongoing, outpatient care by a PCP and a primary-care team (Dorr et al., 2005, 2006a,b, 2007b).

In this system, as shown in Figure 1, coordination of care is crucial to ensure that it is ongoing, comprehensive, and relationship-based. In our studies, we first defined the goals of care coordination and then the crucial processes necessary to attain those goals. These processes are usually nonlinear, are initiated through comprehensive assessments, and require iterative follow-up on care plans and patient needs. Finally, the structure is developed, in terms of the team's abilities, the clinic-based technology that ensures (or at least supports) reliability, and defined roles, all relevant to the patient's needs.

IDENTIFYING GAPS

Because patients have complex needs and there are many potential connections, we first identified the major problems by identifying gaps in the provision of care. We and others have used observations and semi-structured interviews of (1) patients with complex conditions, (2) physicians and nurses, and (3) other health care professionals to identify the most common gaps in care coordination in the primary-care clinic (Bodenheimer, 2008; Dorr et al., 2006b; Wilcox et al., 2007). Principal problems identified in analysis include: (1) a lack of collaboration between patient/family and health-care team; (2) the absence of reliable, complete communication; and (3) failure to prioritize care needs based on both patient input *and* evidence for effective treatment.

Collaboration requires shared decision-making, a process whereby patients are educated about their condition, are offered options, and are provided with tools to help them make decisions. For patients with multiple chronic conditions, decisions must be made frequently and must be coordinated across conditions. Goal-setting by patients—which has been shown to lead to improved health—is done less than 25 percent of the time, and patients report that they do not feel included in decisions more than 50 percent of the time (Bodenheimer and Handley, 2009).

The reasons for frequent gaps in communication include: (1) the patient's need for clear communications that focus on goals and outcomes; (2) the need for multiple inputs (e.g., from specialists and the primary-care team, as well as the patient and family) to complete a communication, which requires cyclical or iterative processes; and (3) the mode of communication either requires more attention than is available (e.g., in-person conversations with the provider) or is not timely (e.g., faxes) (Westbrook et al., 2007). With limited time and attention, failures in communication are common, leading to errors and preventable negative outcomes, such as emergency department visits resulting from unreturned calls or unclear instructions.

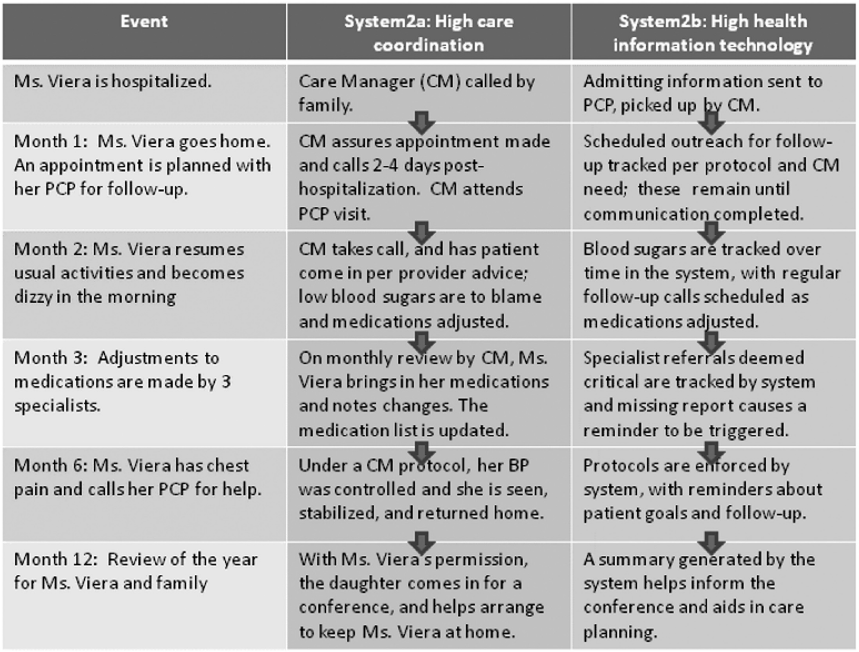
Finally, as the severity and risk factors of the patient's condition increase, the prioritization of needs and next steps is crucial. Systems that remind providers and/or patients about every potential treatment or step in a care plan individually lead to *provider/patient fatigue and distraction* and ultimately fail to improve care. In one study, more than 50 percent of patients did not understand directions given to them by their physicians at the end of a visit (Bodenheimer and Handley, 2009).

COMPONENTS OF A SOLUTION

To understand the components of the solution, we now return to our sample patient, Ms. Viera. Given the same events outlined above over the course of a year, an optimal system, as shown in Figure 3, would address a number of the gaps we identified in the usual system of care.

Reorganizing the Care Team

The first category is team reorganization (Bodenheimer et al., 2002). In other disciplines, such as crew-resource management (e.g., air crews and other teams that work in high-risk, high-attention areas), the crucial requirements for reliable, effective performance include team competencies, thorough training, and



Event	System2a: High care coordination	System2b: High health information technology
Ms. Viera is hospitalized.	Care Manager (CM) called by family.	Admitting information sent to PCP, picked up by CM.
Month 1: Ms. Viera goes home. An appointment is planned with her PCP for follow-up.	CM assures appointment made and calls 2-4 days post-hospitalization. CM attends PCP visit.	Scheduled outreach for follow-up tracked per protocol and CM need; these remain until communication completed.
Month 2: Ms. Viera resumes usual activities and becomes dizzy in the morning	CM takes call, and has patient come in per provider advice; low blood sugars are to blame and medications adjusted.	Blood sugars are tracked over time in the system, with regular follow-up calls scheduled as medications adjusted.
Month 3: Adjustments to medications are made by 3 specialists.	On monthly review by CM, Ms. Viera brings in her medications and notes changes. The medication list is updated.	Specialist referrals deemed critical are tracked by system and missing report causes a reminder to be triggered.
Month 6: Ms. Viera has chest pain and calls her PCP for help.	Under a CM protocol, her BP was controlled and she is seen, stabilized, and returned home.	Protocols are enforced by system, with reminders about patient goals and follow-up.
Month 12: Review of the year for Ms. Viera and family	With Ms. Viera's permission, the daughter comes in for a conference, and helps arrange to keep Ms. Viera at home.	A summary generated by the system helps inform the conference and aids in care planning.

FIGURE 3 A proactive, collaborative system of coordinated health care.

well-defined, well-designed functions (Salas et al., 2006). For care coordination, specific roles—such as the role of a care manager—must be defined to address the need for reliable, effective communication and smooth, efficient workflow (Dorr et al., 2006b).

Evidence based on studies of care managers or care coordinators have increasingly shown that they can be crucial to minimizing the exacerbation of disease (Dorr et al., 2005), reducing the number of hospitalizations (Dorr et al., 2008), and improving patient satisfaction with their care (Wilcox et al., 2007). The essential competencies of care managers include the ability to educate patients and motivate them to set and follow goals and care plans, as well as to communicate effectively with members of the team, the patient, and the patient's family.

Processes tested and implemented for the care of patients with specific conditions have been codified in primary-care team protocols that include identification of common conditions (e.g., elevated blood pressure), a treatment plan, and a flow chart. With the protocol, tasks are disseminated to appropriate team members beyond the beleaguered physician by pre-defining, in sequence, the steps that must be ordered manually under the current system. For protocols to be reliable, however, they must include collaboration, prioritization, and the complexity of the patient's needs. Comprehensive assessments of preferences and goals that include multiple conditions and patients' needs have been shown to improve the health of older adults and to facilitate patient decision-making (Boult et al., 1999).

In our sample case, the care manager would facilitate coordination by receiving the call from the hospital, making the post-hospitalization follow-up call to the patient and family, and following protocols and proactively identifying the patient's needs. Care managers can help close the communication loop because they remain focused on the key communication tasks for at-risk patients, follow up on critical referrals to specialists, and arrange conferences to consolidate communication. In the usual flow of things, the clinical staff must attend to many urgent needs as they arise and therefore have limited time to perform these less urgent, but no less important, tasks. Research has shown that trained care managers can accomplish these tasks and hence greatly improve the effectiveness and efficiency of health care delivery.

Health Information Technology

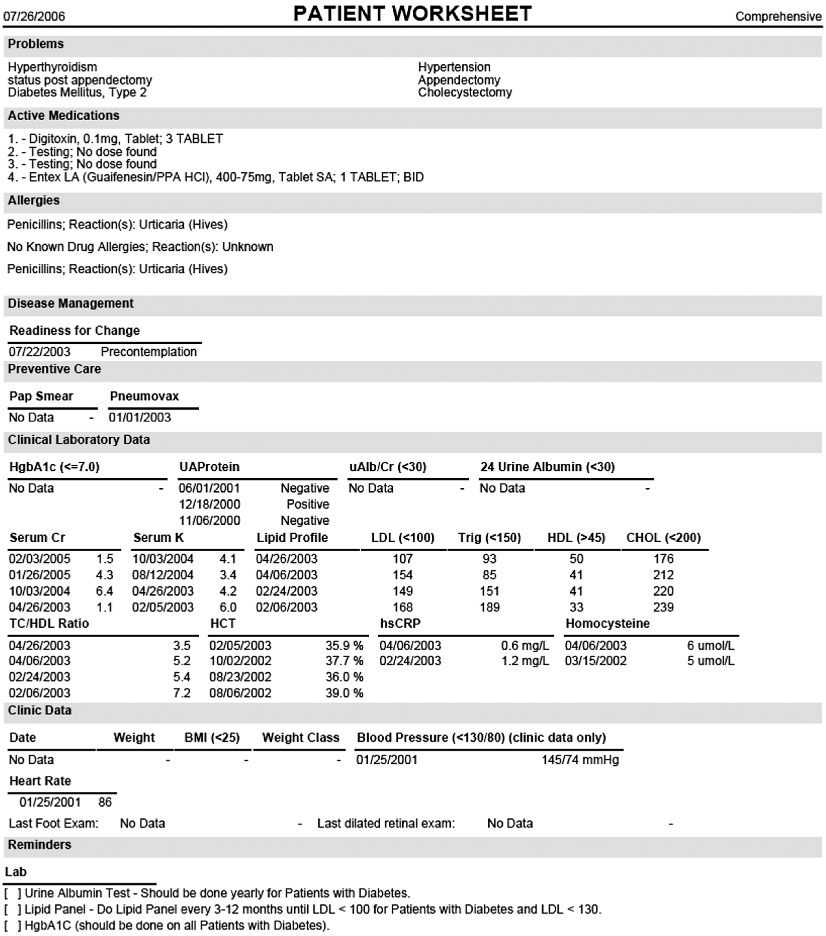
The number of patients that can be followed by a care manager is limited. Studies have shown that 2 to 5 percent of patients in a usual primary care clinic meet the criteria of Ms. Viera's case: an at-risk patient diagnosed with multiple comorbid illnesses in need of ongoing care coordination. In a clinic of seven physicians, more than a thousand patients may meet these criteria, which could easily overwhelm care managers.

The primary goal of care coordination is to monitor, over time, the active care and treatment plan for patients and to take necessary steps to ensure that the plan is completed successfully. For example, health information technology (HIT) can greatly increase the likelihood of success. Key process points can be defined and programmed to remind care managers about crucial tasks. Whereas electronic health record systems usually focus on individual clinic visits and work flow and relegate hundreds of items to unstructured to-do lists, in our example, HIT functions can be adapted to help prioritize tasks by (1) identifying crucial elements that should be shared by members of the primary-care team, (2) ensuring that relevant information is delivered to the correct team members in the appropriate format, and (3) reminding clinicians about uncompleted tasks. To start the process, HIT, using filtering and prioritized data flow, must identify all patients under care management and the state of their current treatment plans and goals.

In our studies, by using HIT, care managers were able to follow an average of 350 patients at a time, approximately 1,000 per year (Dorr et al., 2007a). Patients under care management received prioritized messages about their hospital stay, were given automatic follow-up after sentinel events that persisted beyond an individual call or visit, and were moved to the top of the queue for attention when necessary.

The HIT system can embed protocols, although they must be flexible enough to accommodate the needs of care managers; for example, care managers only need the next step defined and a reminder sent to address a patient's rapidly changing status. Even for individual patients, care managers must identify the highest priority tasks and should be reminded about these first (Dorr et al., 2006a). For Ms. Viera, who sees 12 specialists a year, the care manager would designate which of these referrals are critical and directly affect the care plan; HIT would then remind the caregiver about these elements only.

Finally, summaries of the complex care and needs of patients are crucial to addressing emerging issues quickly and integrating the patient's history with anticipated care needs in one place. Figure 4 shows this summary mechanism, the



07/26/2006 **PATIENT WORKSHEET** Comprehensive

Problems

Hyperthyroidism
status post appendectomy
Diabetes Mellitus, Type 2
Hypertension
Appendectomy
Cholecystectomy

Active Medications

1. - Digitoxin, 0.1mg, Tablet; 3 TABLET
2. - Testing; No dose found
3. - Testing; No dose found
4. - Entex LA (Guaifenesin/PPA HCl), 400-75mg, Tablet SA; 1 TABLET; BID

Allergies

Penicillins; Reaction(s): Urticaria (Hives)

No Known Drug Allergies; Reaction(s): Unknown

Penicillins; Reaction(s): Urticaria (Hives)

Disease Management

Readiness for Change	
07/22/2003	Precontemplation

Preventive Care

Pap Smear		Pneumovax
No Data	-	01/01/2003

Clinical Laboratory Data

HgbA1c (<=7.0)		UAProtein		uAlb/Cr (<30)		24 Urine Albumin (<30)	
No Data	-	06/01/2001	Negative	No Data	-	No Data	-
		12/18/2000	Positive				
		11/06/2000	Negative				

Serum Cr		Serum K		Lipid Profile	LDL (<100)	Trig (<150)	HDL (>45)	CHOL (<200)
02/03/2005	1.5	10/03/2004	4.1	04/26/2003	107	93	50	176
01/26/2005	4.3	08/12/2004	3.4	04/06/2003	154	85	41	212
10/03/2004	6.4	04/26/2003	4.2	02/24/2003	149	151	41	220
04/26/2003	1.1	02/05/2003	6.0	02/06/2003	168	189	33	239

TC/HDL Ratio		HCT		hsCRP		Homocysteine	
04/26/2003	3.5	02/05/2003	35.9 %	04/06/2003	0.6 mg/L	04/06/2003	6 umol/L
04/06/2003	5.2	10/02/2002	37.7 %	02/24/2003	1.2 mg/L	03/15/2002	5 umol/L
02/24/2003	5.4	08/23/2002	36.0 %				
02/06/2003	7.2	08/06/2002	39.0 %				

Clinic Data

Date	Weight	BMI (<25)	Weight Class	Blood Pressure (<130/80) (clinic data only)	
No Data	-	-	-	01/25/2001	145/74 mmHg

Heart Rate	
01/25/2001	86

Last Foot Exam: No Data - Last dilated retinal exam: No Data -

Reminders

Lab

[ ] Urine Albumin Test - Should be done yearly for Patients with Diabetes.
[ ] Lipid Panel - Do Lipid Panel every 3-12 months until LDL < 100 for Patients with Diabetes and LDL < 130.
[ ] HgbA1C (should be done on all Patients with Diabetes).

FIGURE 4 Comprehensive summary sheet.

patient's worksheet (which, by itself, has been shown to improve adherence with evidence-based treatments for chronic and preventive illness by 17 to 30 percent) (Wilcox et al., 2005).

BUILDING A SUSTAINABLE MODEL

Once a patient's needs and potential solutions have been identified, we work to implement them into a model of care. In seven intervention clinics at Intermountain Healthcare, a large, integrated health-delivery system, we installed care managers, trained them, and adapted HIT over a period of two years to develop the protocols and system described above. Over the next four years, patients seen by care managers lived longer, had 24 to 40 percent fewer hospitalizations, and had significantly better control of their conditions than similar patients at clinics without care managers (Dorr et al., 2005, 2008). The clinics with care managers also achieved higher efficiency levels, as measured by clinical output (patients seen and complexity of conditions treated). Lower costs that resulted from greater efficiency covered the costs of the care managers *and* the costs of expanding the program (Dorr et al., 2007b).

Unanticipated effects included variations in referral patterns and care-management patterns that led to some variations in outcomes. For instance, patients with predominantly social or financial problems did not have significantly fewer hospitalizations or emergency department visits, despite the care managers' efforts. A positive unanticipated effect was a result of integrating a set of providers and patients. A number of patients described the care manager as "a life-saver," and a number of providers said they "could not imagine practicing without the care manager."

MAINTENANCE AND SUSTAINABILITY OF THE CARE-MANAGER MODEL

The next step is to determine the maintenance and sustainability of the care-manager model. In our qualitative studies, we defined core aspects of successful care management and embedded these in a training- and HIT-enhancement program. The core components of the model were defined as: (1) a trained care manager; (2) a supportive, trained team; and (3) specialized HIT. We then created a training and HIT support program (for details, see *caremanagementplus.org*). To date, more than 75 clinical teams have participated in the training and have been working on improving their HIT systems. In all, 73 percent of the teams were able to implement the core components of the model.

Further work is being done on sustainability, which can be a problem because many care-management tasks are not specifically reimbursed, despite their value. Changes in the reimbursement system (e.g., payments for the "medical home," a comprehensive model of primary care) or direct payments for care coordination may enable many more primary care teams to adopt these models.

CONCLUSIONS

Successful models of coordinated care that meet identified needs and improve patient health can be created by identifying gaps in current care systems, developing solutions that meet a particular patient's needs, and developing change-management processes. We have shown that one successful model is to use care managers to augment primary-care teams and HIT to remind care managers and clinicians about prioritized tasks. The next steps will be to explore ways to ensure sustainability and to reinforce changes in the current health care system.

ACKNOWLEDGMENTS

Funding for this research comes from The John A. Hartford Foundation, Agency for Healthcare Research and Quality, and National Library of Medicine. This paper does not reflect the official positions of any of these institutions. Thank you also to the Care Management Plus team and collaborators.

REFERENCES

Anderson, R.N., and E. Arias. 2003. The effect of revised populations on mortality statistics for the United States, 2000. National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System 51(9): 1–24.

Bodenheimer, T. 2008. The future of primary care: transforming practice. New England Journal of Medicine 359(20): 2086, 2089.

Bodenheimer, T., and M.A. Handley. 2009. Goal-setting for behavior change in primary care: an exploration and status report. Patient Education and Counseling 76(2): 174–180.

Bodenheimer, T., E.H. Wagner, and K. Grumbach. 2002. Improving primary care for patients with chronic illness. Journal of the American Medical Association 288(14): 1775–1779.

Boult, C., R.L. Kane, J.T. Pacala, and E.H. Wagner. 1999. Innovative healthcare for chronically ill older persons: results of a national survey. American Journal of Managed Care 5(9): 1162–1172.

Dorr, D.A., A. Wilcox, S.M. Donnelly, L. Burns, and P.D. Clayton. 2005. Impact of generalist care managers on patients with diabetes. Health Services Research 40(5 Pt 1): 1400–1421.

Dorr, D.A., H. Tran, P. Gorman, and A.B. Wilcox. 2006a. Information needs of nurse care managers. Pp. 913 in AMIA Annual Symposium Proceedings. Bethesda, Md.: American Medical Informatics Association.

Dorr, D.A., A. Wilcox, L. Burns, C.P. Brunker, S.P. Narus, and P.D. Clayton. 2006b. Implementing a multidisease chronic care model in primary care using people and technology. Disease Management: DM 9(1): 1–15.

Dorr, D.A., A. Wilcox, S. Jones, L. Burns, S.M. Donnelly, and C.P. Brunker. 2007a. Care management dosage. Journal of General Internal Medicine 22(6): 736–741.

Dorr, D.A., A. Wilcox, K.J. McConnell, L. Burns, and C.P. Brunker. 2007b. Productivity enhancement for primary care providers using multicondition care management. American Journal of Managed Care 13(1): 22–28.

Dorr, D.A., A.B. Wilcox, C.P. Brunker, R.E. Burdon, and S.M. Donnelly. 2008. The effect of technology-supported, multidisease care management on the mortality and hospitalization of seniors. Journal of the American Geriatrics Society 56(12): 2195–2202.

Salas, E., K.A. Wilson, C.S. Burke, and D.C. Wightman. 2006. Does crew resource management training work? An update, an extension, and some critical needs. Human Factors 48(2): 392–412.

Shugarman, L.R., S.L. Decker, and A. Bercovitz. 2009. Demographic and social characteristics and spending at the end of life. Journal of Pain and Symptom Management 38(1): 15–26.

Westbrook, J.I., J. Braithwaite, A. Georgiou, A. Ampt, N. Creswick, E. Coiera, and R. Iedema. 2007. Multimethod evaluation of information and communication technologies in health in the context of wicked problems and sociotechnical theory. Journal of the American Medical Informatics Association 14(6): 746–755.

Wilcox, A.B., S.S. Jones, D.A. Dorr, W. Cannon, L. Burns, K. Radican, K. Christensen, C. Brunker, A. Larsen, S.P. Narus, S.N. Thornton, and P.D. Clayton. 2005. Use and impact of a computer-generated patient summary worksheet for primary care. Pp. 824–828 in AMIA Annual Symposium Proceedings. Bethesda, Md.: American Medical Informatics Association.

Wilcox, A.B., D.A. Dorr, L. Burns, S. Jones, J. Poll, and C. Bunker. 2007. Physician perspectives of nurse care management located in primary care clinics. Care Management Journals: Journal of Case Management; The Journal of Long Term Home Health Care 8(2): 58–63.

Wolff, J.L., B. Starfield, and G. Anderson. 2002. Prevalence, expenditures, and complications of multiple chronic conditions in the elderly. Archives of Internal Medicine 162(20): 2269–2276.

RESILIENT AND SUSTAINABLE INFRASTRUCTURE

Introduction

Seth Guikema
Johns Hopkins University
Baltimore, Maryland

Patrick O'Mara
STV, Incorporated
New York, New York

America's infrastructure, which includes highways, bridges, mass transit, fresh water supply, wastewater treatment, telecommunications, energy, dams, and schools to name just some of its components, is often taken for granted and overlooked until something goes wrong. As we have learned in the recent past, the importance of these "invisible systems" becomes immediately apparent when they are not working. Natural and man-made disasters in the last decade have highlighted the vulnerability of our nation's interdependent infrastructure systems.

- Power outages darkened the entire northeastern region in 2003.
- Communication lines have often been disrupted preventing e-mail, cell phone, and telephone transmissions.
- An interstate highway bridge, the I-35W bridge in Minneapolis, collapsed and disrupted traffic patterns and limited access.
- In New York City, steam pipes exploded when an aging power supply system failed, disrupting transportation and energy distribution.
- In 2005, Hurricane Katrina caused levees to collapse in New Orleans and flooded entire neighborhoods and communities in the city and along the Gulf Coast.

The interconnectedness of infrastructure systems increases their susceptibility to failures and compounds recovery efforts. For example, an earthquake may damage buildings, bridges, and roadways, which may hamper emergency access to fires caused by ruptured gas lines. In addition, these fires cannot be extinguished if the transmission lines that provide power to the local water pumping station are down.

The presentations in this technical session will focus on research efforts to improve the long-term resiliency and sustainability of critical infrastructure systems, which we so often take for granted. This research includes investigating ways to develop systems that can withstand natural disasters and/or return to full operation soon after a catastrophe occurs. Similarly, techniques, such as life-cycle assessments, are being studied to create a more sustainable infrastructure by reducing energy requirements, using recycled materials, and/or making better decisions.

In the first presentation, Kristina Swallow presents an overview of the state of our nation's infrastructure as outlined in the American Society of Civil Engineers Infrastructure Report Card, highlights possible causes of infrastructure deterioration, and proposes possible solutions. The second speaker, Stephanie Chang, focuses on research to assess the vulnerability of urban infrastructure systems in areas susceptible to natural disasters with the goal of maximizing preparations and reallocating resources to make recovery and repair operations more efficient. Arpad Horvath, the final speaker, focuses on environmental life-cycle assessment modeling of infrastructure systems and the implications, for the environment and energy use, of using different designs and materials.

America's Infrastructure Report Card: Causes, Costs, and Solutions

Kristina L. Swallow
AVP Consulting, Inc.
Las Vegas, Nevada

Whenever any one of us turns on the water, drives on a road, goes to a park or lake, takes out the trash, or turns on a light, we benefit from the vast complex network of systems that comprises America's infrastructure. Infrastructure is silent, invisible, and rarely thought of unless something goes wrong and it fails to perform. Unfortunately, due to age, poor maintenance, and higher than predicted demands, much of our infrastructure is in danger of failing, and it is critical that it be brought up to current standards.

In 1988, a congressionally chartered commission, the National Council on Public Works Improvement, completed a study on the state of America's infrastructure entitled Fragile Foundations: A Report on America's Public Works. Using a report card format to guide its analysis and publish its results, the commission gave the United States an overall infrastructure grade of C, stating that an annual increase in investment of up to 100 percent was required to improve it (ASCE, 2009, p.9).

Since then the American Society of Civil Engineers (ASCE) has issued four report cards on the nation's infrastructure. In the most recent report card, issued this past March, infrastructure was given an overall grade of D (Table 1) (ASCE, 2009, p. 2). ASCE estimates that infrastructure spending in the next five years will have to be $2.2 trillion to improve its condition of to "good" (ASCE, 2009, p. 6). This represents an investment of $500 billion more than was estimated in

TABLE 1 2009 Report Card for America's Infrastructure

Aviation	D
Bridges	C
Dams	D
Drinking Water	D–
Energy	D+
Hazardous Waste	D
Inland Waterways	D-
Levees	D-
Public Parks and Recreation	C-
Rail	C-
Roads	D-
Schools	D
Solid Waste	C+
Transit	D
Wastewater	D-
AMERICA'S INFRASTRUCTURE G.P.A.	**D**

Note: From *2009 Report Card for America's Infrastructure*, American Society of Civil Engineers, p. 2. Copyright 2009 by the American Society of Civil Engineers. Adapted with permission.

ASCE's 2005 report card and approximately $1.1 trillion more than the U.S. currently invests in infrastructure (Table 2) (ASCE, 2009, p. 7).

The ASCE report card evaluates the condition of infrastructure as determined by its ability to meet current and projected needs and its resiliency and estimates the costs to improve the quality from its current grade to a B. Although long-term maintenance and funding for capital improvements can prolong the life of infrastructure systems, effectively improving the nation's infrastructure will require prioritized spending and an understanding of the underlying causes of failure. ASCE found that infrastructure systems across the nation were (1) either approaching or had already exceeded their overall design life; (2) trying to meet demands that exceeded their design capacity; (3) lacking in redundancy; (4) interdependent on other failing systems; and (5) facing technological obsolescence. By recognizing that the overall decline is attributable to more than one contributing factor, engineers can take a broad approach to understanding the problems and developing solutions.

TABLE 2 Estimated Five-Year Investment Needs, in Billions of Dollars

CATEGORY	5-YEAR NEED	ESTIMATED ACTUAL SPENDING*	AMERICAN RECOVERY AND RESINVESTMENT ACT (P.L. III-005)	FIVE-YEAR INVESTMENT SHORTFALL
Aviation	87	45	1.3	(40.7)
Dams	12.5	5	0.05	(7.45)
Drinking water and wastewater systems	255	140	6.4	(108.6)
Energy systems	75	34.5	11	(29.5)
Hazardous waste and solid waste systems	77	32.5	1.1	(43.4)
Inland waterways	50	25	4.475	(20.5)
Levees	50	1.13	0	(1.13)
Public parks and recreation	85	36	0.835	(48.17)
Rail systems	63	42	9.3	(11.7)
Roads and bridges	930	351.5	27.5	(549.5)
Discretionary grants for surface transportation			1.5	
Schools	160	125	0**	(35)
Transit	265	66.5	8.4	(190.1)
	2.122 trillion***	903 billion	71.76 billion	(1.176 trillion)
Total****	**$2.2 trillion**			

Note: From *2009 Report Card for America's Infrastructure*, American Society of Civil Engineers, p. 7. Copyright 2009 by the American Society of Civil Engineers. Adapted with permission.

* 5 year spending estimate based on the most recent available spending at all levels of government and not indexed for inflation

** The American Recovery and Reinvestment Act included $53.6 billion for a State Fiscal Stabilization Fund for education, as of press time, it was not known how much would be spent on school infrastructure.

*** Not adjusted for inflation

**** Assumes 3% annual inflation

WATER SYSTEMS

Drinking water systems touch the lives of all Americans, and their continued performance is critical to our overall health. At a time when much of the United States is deep into a multi-year drought, it is vital that drinking water systems function continuously, correctly, and efficiently, and that they have the best technology to optimize consumption.

However, ASCE gives the nation's drinking water systems a grade of D-. Despite significant investment at the federal and local levels, ASCE reports that there is still a projected $11 billion shortfall in the funding necessary to replace facilities that are nearing the end of their useful lives (ASCE, 2009, p. 26).

In 2000 the U.S. population consumed an estimated 43 billion gallons of water per day. An additional 7 billion gallons of clean drinking water are lost every day through leaking pipes (ASCE, 2009, p. 24). In addition, drinking water systems are not as resilient as they should be; most of them do not have the redundancy necessary to maintain service in the event of a disruption. A complicating factor is their dependence on a reliable power source to provide continuous service; energy infrastructure earned a D+ rating.

The nation's wastewater systems, facing similar challenges, also received a grade of D-. The facilities in many of these systems have reached the end of their useful lives resulting "in the release of as much as 10 billion gallons of raw sewage yearly" (ASCE, 2009, p. 58). In addition, many are combined storm water/sanitary sewer systems that are incapable of conveying flows generated by major storms or runoff from heavy snowfalls. At a minimum, there must be separate sanitary sewer systems in these locations to prevent uncontrolled discharges of raw sewage. The projected annual funding shortfall for wastewater treatment systems is $6 billion (ASCE, 2009, p. 58).

To meet increasing demands, maintain critical systems, replace aging facilities, and create necessary redundancies to ensure resiliency, ASCE estimates that there is an overall need for $255 billion for water and wastewater systems in the next five years. However, in January 2009, the estimated spending for the next five years was only $146.4 billion, representing a $108.6 billion shortfall. Although there are long-term federal funding programs for national defense and the interstate highway and aviation systems, no such program exists for drinking water and wastewater systems (ASCE, 2009, p. 60).

Long-term funding for drinking water and wastewater systems is critical to maintaining and improving public health and should be coordinated with a variety of other funding mechanisms at the local and state levels. In addition, water is becoming increasingly scarce. Hence water and wastewater utilities will have to work together to identify ways to increase supply to meet growing water needs. Solutions may include developing reclaimed water systems, improving capture of storm water, and using treated wastewater to replenish groundwater supplies.

TRANSPORTATION SYSTEMS

The nation's transportation systems face similar funding and aging challenges. Although bridges received the highest grade on the report card, a C, more than one in four bridges is either structurally deficient or functionally obsolete (ASCE, 2009, p. 74). With a typical design life of 50 years, "the average bridge . . . is now 43 years old" (ASCE, 2009, p. 76).

Although a structurally deficient or functionally obsolete bridge is not necessarily unsafe, it "cannot accommodate current traffic volumes, vehicle sizes, and weights" (ASCE, 2009, p. 76). The resulting restrictions create congestion, delays, and unreliable service for emergency vehicles. The limitations are compounded by growing traffic demands. In the past 20 years truck traffic has more than doubled, and trucks are carrying increasingly large loads.

The condition of bridges is only one contributing factor to congestion on our roadways. ASCE gives the roads themselves a grade of D-. Congestion on roadways results in time wasted and fuel costs of $710 per motorist annually (ASCE, 2009, p. 100). In addition to these wasted resources, the poor quality of roadways, 84.9 percent of which are rated as only adequate, leads to safety problems. In 2007 alone there were 41,059 fatalities from motor vehicle crashes and 2.5 million injuries (ASCE, 2009, p. 100).

Trucks are a significant factor in problems with roadways. In the 10 years ending in 2004, "freight traffic moved by truck grew 33%" (ASCE, 2009, p. 100). This significant increase is a strong indication of "the increased dependency of commerce on the efficiency of the roadways and the added wear and tear [on the roads] caused by trucks" (ASCE, 2009, p. 100).

The resiliency of roadways and bridges is critical. When a bridge or roadway fails, there is often an immediate loss of life and, until the system is restored, increased delays for roadway users. ASCE estimates a total of $930 billion will be required over the next five years to improve the nation's roadways to a grade of B. Of that amount, they anticipate that there will be a funding shortfall of $549.5 billion.

In addition to providing funding at all levels for maintenance and capital improvements, the nation's goals should be not only to improve the condition of roads and bridges, but also to increase research on ways to improve safety, reduce lost time due to construction, and improve overall movement on the roadways.

CONCLUSION

This brief summary highlights some of the challenges facing two types of infrastructure systems included in the ASCE 2009 Report Card for America's Infrastructure. This study provides a snapshot of the current state of America's infrastructure, identifies the reasons for its decay, and offers specific recommendations for addressing the challenges to infrastructure systems in all categories.

In addition, ASCE developed five key solutions applicable to improving the overall state of our infrastructure: (1) increase federal leadership in infrastructure; (2) promote sustainability and resilience; (3) develop federal, regional, and state plans for infrastructure; (4) address life-cycle costs and ongoing maintenance needs; and (5) increase infrastructure investment from all stakeholders (ASCE, 2009, p. 11). Implementing these solutions will result, over time, in an improved infrastructure for the nation that will support the health, safety, and economic security of its citizens.

REFERENCE

ASCE (American Society of Civil Engineers). 2009. 2009 Report Card for America's Infrastructure. Available online at *http://www.infrastructurereportcard.org* (accessed July 2009).

Infrastructure Resilience to Disasters

Stephanie E. Chang
University of British Columbia
Vancouver, British Columbia

Urban societies depend heavily on the proper functioning of infrastructure systems such as electric power, potable water, and transportation networks. Normally invisible, this reliance becomes painfully evident when infrastructure systems fail during disasters. Moreover, because of the network properties of infrastructure, damage in one location can disrupt service in an extensive geographic area. The societal disruption caused by infrastructure failures is therefore disproportionately high in relation to actual physical damage.

Engineers have long tried to design infrastructure to withstand extreme forces, but recently they have begun to address the need for urban infrastructure systems that are *resilient* to disasters (e.g., NIST, 2008). Conceptually, resilience entails three interrelated dimensions: lower probabilities of failure; less-severe negative consequences when failures do occur; and faster recovery from failures (Bruneau et al., 2003). The emphasis on consequences and recovery suggests that improving the resilience of infrastructure systems is not only a technical problem, but it also has societal dimensions.

The consequences of recent disasters have demonstrated that urban infrastructure systems in the United States and other developed countries (not to mention in developing regions of the world) remain highly vulnerable. Moreover, infrastructure failure is often a primary cause of economic and human losses in disasters. Consider, for example, the consequences of infrastructure failures caused by wind, storm surges, and levee failures in New Orleans during Hurricane Katrina.

TABLE 1 Examples of Infrastructure Failures and Consequences in Disasters

Event	Location	Infrastructure Failure and Consequences	Source
1993 Great Midwest floods	Des Moines, Iowa	Businesses suffered greater economic losses from infrastructure outages (water, electric power, and wastewater services) than from physical flooding of their facilities.	Webb et al., 2000
1994 Northridge earthquake ($M_w = 6.7$)	Los Angeles, California	Damage to bridges, which closed portions of four major freeway routes, accounted for \$1.5 billion in losses from business interruption (a quarter of the total).	Gordon et al., 1998
1995 Great Hanshin-Awaji earthquake ($M_w = 6.9$)	Kobe, Japan	Extensive infrastructure failures, including outages of electric power and telecommunications (1 week), water and natural gas (2–3 months), commuter railway (up to 7 months), and highway systems and port infrastructure (approx. 2 years). It took 10 years for the city population to recover. Economic activity, especially at the port, has still not fully recovered.	Chang (in press); Chang and Nojima, 2001

Table 1 provides a few other examples to illustrate the frequency and range of infrastructure failures in disasters.

Because infrastructure failures are clearly a primary cause of disruptions in disasters, strategies for improving the disaster resilience of communities must focus on improving infrastructure resilience. Yet few standards or guidelines have been developed for this, partly because of the complexity of the problem (American Lifelines Alliance, 2006).

RESEARCH ON INFRASTRUCTURE IN DISASTERS

Much of the early work on infrastructure in disasters was on understanding the mechanics of how components of infrastructure systems (e.g., bridge piers, buried pipes, electric power transformers, and other substation equipment) perform when subjected to extreme forces or conditions. This basic understanding

TABLE 1 Continued

Event	Location	Infrastructure Failure and Consequences	Source
September 11, 2001 World Trade Center terrorist attack	New York City	Widespread disruption in lower Manhattan to emergency service facilities, transportation (including subways), telecommunications, electric power, and water.	O'Rourke et al., 2003
August 14, 2003 blackout	Portions of U.S. Midwest, Northeast, and southern Ontario	Power outages began in northern Ohio and cascaded through the electric power grid to cause the largest blackout in North American history (affecting 50 million people). Losses amounted to an estimated $10 billion. Water supply, telecommunications, transportation, hospitals, and other dependent infrastructures were disrupted.	McDaniels et al., 2007; U.S.-Canada Power System Outage Task Force, 2006
2004 Hurricanes Charley, Frances, and Jeanne	Central Florida	Port closures disrupted delivery of fuel and emergency materials. Electric power outages lasted for more than a week. The supply of emergency generators was not large enough to meet demand.	American Lifelines Alliance, 2006

was then extended to the performance of component assemblages (e.g., bridges, pipeline networks, substations). Studies ranged from field work to laboratory simulations with scale models and computer-based analyses.

As a result of these studies, new engineering designs, materials, and retrofitting strategies were developed to improve the ability of infrastructure elements to withstand natural hazards. New technologies were also developed, such as sensors for monitoring structural health and detecting damage and real-time system controls.

While these remain active areas of inquiry, new research themes have emerged to address some of the complexities of infrastructures, which include societal as well as technical issues. How, for instance, will the failure of one bridge affect businesses throughout the urban area that rely on the transportation system? How will the failure of one infrastructure system disrupt other infrastructure systems? How can repairs following a disaster be planned so they minimize social and economic losses? Such questions have prompted research that is, by necessity, interdisciplinary.

CHALLENGES OF INTERDISCIPLINARY RESEARCH

Interdisciplinary inquiry is inherently difficult for many reasons, ranging from intellectual issues, such as differences in communication and attitudes, to organizational issues, such as funding mechanisms and academic structures. Interdisciplinary research at the intersection of engineering and the social sciences is especially challenging (NRC, 2006).

One basic hurdle has been different disciplinary concepts of the term "infrastructure." To structural engineers, for example, infrastructure comprises constructed elements, such as pipes and bridges, described in terms of materials and design properties that condition their responses to physical forces. To economists, infrastructure—often referred to as "public capital"—comprises an input to economic production measured in dollars (e.g., Munnell, 1992) and often quantified at the state or national level. These fundamental differences reflect different ways of conceptualizing and measuring infrastructure.

Overcoming these challenges has required more collaborative, interdisciplinary research than in past engineering studies. It has also required researchers to pay greater attention to issues of time, space, and context. These trends are illustrated below in an example from the field of earthquake engineering.

WATER SYSTEMS IN A LOS ANGELES-AREA EARTHQUAKE

The Los Angeles Department of Water and Power (LADWP), the largest municipal utility in the United States, provides potable water to 3.9 million people through 11,700 kilometers of infrastructure in one of the most seismically active regions of the country. Over the last several years, researchers affiliated with the Multidisciplinary Center for Earthquake Engineering Research (MCEER), a research center funded by the National Science Foundation, have been studying the potential consequences of major earthquakes on the LADWP water system. Highlights of three of these studies[1] illustrate some key challenges and breakthroughs.

Modeling Potential Physical Damage

The first study, conducted by geotechnical engineers, developed a model of potential physical damage to the LADWP network (Romero et al., 2009). Geo-

---

[1] The MCEER-LADWP research program also involved other studies (not described here) that analyzed regional seismicity, modeled performance of the electric power transmission system, and investigated forms of business resilience and resilient behaviors

graphic information system (GIS) technology was used to visualize the spatial dimensions of seismic ground waves, peak ground deformation, fault rupture, soil liquefaction, and landslides, as well as the network itself. The model, the Graphical Iterative Response Analysis for Flow Following Earthquake (GIRAFFE), assesses damage to network components (pipes, tanks, reservoirs, etc.) and performs hydraulic modeling of water flows through the damaged network. GIRAFFE also estimates serviceability—defined as the ratio of post-earthquake to pre-earthquake water flow—for each service area.

In 2008, results for one hypothetical event, a M_W 7.8 earthquake on the southern San Andreas Fault, were used as part of the largest emergency preparedness exercise in U.S. history. In that scenario, overall water serviceability 24 hours after the earthquake was estimated to be as low as 34 percent (after reserves in storage tanks had been depleted). LADWP has now adopted GIRAFFE, trained its personnel to use it, and is applying the results in its system decision making.

Modeling the Post-Earthquake Damage-Repair Process

In a related study, systems engineers modeled the damage-repair process to estimate the duration of water outages (Brink et al., 2009). A discrete-event simulation model was developed that mimics the actual post-earthquake restoration process, including the movements of repair crews over time and their activities, which are subject to personnel and material constraints. Data were derived from extensive consultations with LADWP engineering staff.

The restoration model was then run in tandem with the GIRAFFE damage and water-flow model to simulate serviceability in 12-hour increments as repairs were made over time and space; uncertainties were handled through multiple discrete simulations. The results showed substantial variations in how restoration might proceed, and LADWP concluded the restoration model would be helpful in planning for resource allocations following a disaster.

Modeling the Effects of Water Outages on Businesses

Urban planners used the models described above and other MCEER engineering studies to investigate the consequences of water outages, including impacts on the economy (Chang et al., 2008). For example, an agent-based simulation model accounts for how different types of businesses would be affected by water outages. Inputs include water serviceability ratios and restoration times based on the studies described above, as well as characteristics of businesses per se. Data were derived from surveys of impacts on businesses in previous disasters. Impacts from water outages were estimated in the context of other types of earthquake-related disruptions, specifically damage to buildings and power outages. Results for a M_W 6.9 Verdugo Fault scenario indicated that water outages could account for an estimated $467 million in direct business losses, or about 1.5 percent of the estimated total economic losses.

CRITICAL FACTORS IN INTERDISCIPLINARY STUDIES

The studies described above have demonstrated the feasibility and promise of interdisciplinary research on infrastructure resilience for developing the capability of modeling post-disaster losses and recoveries over time. Based on the experience of developing GIRAFFE, the restoration model, and other models, we have identified factors that promote interdisciplinary research in this area:

- **GIS technology** helps bridge disparate datasets and models by providing a common platform for information sharing and data integration.
- **The concept of infrastructure "services"** is critical for linking physical damage to societal impacts. This concept, which differs from both the traditional engineering concept of infrastructure and the traditional economic concept of infrastructure, reflects an intermediate representation that connects them.
- **A research center approach** makes it possible to address the entire scope of a complex problem through the coordinated efforts of a multidisciplinary team, convened and sustained over several years. This coordination is necessary to identifying critical gaps in knowledge, involving appropriate researchers, and overcoming disciplinary barriers.
- **Collaboration with the end user**, that is, with LADWP, the infrastructure organization itself, is essential. LADWP engineers contributed in important ways to framing questions, developing data, and ultimately, to applying outcomes to decision-making.

Without all of these factors in place, this interdisciplinary research could not have been conceived or conducted.

CHALLENGES ON THE HORIZON

Where is the current frontier in research on infrastructure resilience to disasters? In this author's opinion, much remains to be understood and addressed about the performance of engineered elements and systems. In addition, the nexus between engineering and social sciences has just begun to be explored through interdisciplinary research, and many important questions remain to be answered. In this context, three new challenges have been gaining attention.

Interdependencies

The first challenge is interdependencies—understanding and addressing how failures in one infrastructure system lead to failures in another. Loss of electric power, for example, commonly leads to disruptions in water, transportation, and health care systems, among others (McDaniels et al., 2007). There are several types of interdependencies: physical linkages (e.g., pump stations in water delivery systems that require electricity); cyber linkages (e.g., computerized system controls that rely on telecommunications); geographic linkages (e.g., pipelines located on transportation bridges); and "logical" linkages (e.g., infrastructure elements related through economic markets driven by human decision-making) (Rinaldi et al., 2001).

The technical understanding of these interdependencies is still in its early stages, and many infrastructure organizations have been reluctant to share information about their vulnerabilities for security reasons. Nevertheless, an understanding of infrastructure interdependencies is critical for cities deciding on strategic investments in infrastructure improvements that will have the greatest payoff in terms of resilience.

Multi-hazards

The second new challenge is multi-hazards. Because infrastructure systems are vulnerable to multiple stressors (e.g., wind, ice, flood, earthquake, terrorism, deterioration), it is important to find solutions and support decisions that consider them in that context. Synergies in risk-reduction technologies may reduce the costs of pre-disaster retrofitting and post-disaster repairs. Methods are also needed to assess how the deterioration of infrastructure over time affects disaster risk.

Sustainability

The third challenge, sustainability, is the consideration of infrastructure resilience in a long-term environmental context. It can be argued that disaster resilience is an inherent characteristic of sustainability. On one level, designing and building infrastructure that is able to withstand disasters will reduce their negative environmental impact, such as debris from damaged structures, spills of hazardous materials and other contaminants, and the carbon footprint of reconstruction activities. Infrastructure designers should, therefore, include such life-cycle environmental impacts in their decision-making (Guikema, 2009).

On another level, because infrastructures are long-lived, infrastructure resilience will require the capacity to meet demands that may change drastically over their life cycles. Such changes may include urban growth and increases in populations. Climate change will also be important, for example, through rising sea levels that redefine coastlines and through changes in the occurrence probabilities of hazardous events, including hurricanes, extreme rainfalls, droughts, temperature extremes, landslides, and floods. Climate change will not only put coastal infrastructure, such as port and harbor facilities, at risk. In many cities, climate change will also stress water supplies, wastewater treatment facilities, and transportation systems (Infrastructure Canada, 2006).

In addition, infrastructure systems themselves have substantial effects on the environment. For example, building flood-control levees, paradoxically, encourages development in hazardous floodplains. Thus the vulnerability of New Orleans to Hurricane Katrina was partly attributable to decisions, such as levee construction, that were made over a period of many decades to protect against relatively frequent storms, but that increased the city's vulnerability to very large, albeit rare, catastrophic storms (Kates et al., 2006).

In addition, the capacity of Louisiana's coastal wetlands to help buffer wind and storm surges was substantially degraded over decades by the construction of levees, shipping channels, oil and gas industry facilities, and other infrastructures (e.g., Kousky and Zeckhauser, 2006). Some have suggested that flood protection should not only comprise building levees, but should also be designed to encourage marsh restoration (Guikema, 2009). Others have proposed the decommissioning of existing infrastructures—such as the selective dismantling of dams and levees—along with ecosystem restoration as an approach to addressing the problems of aging infrastructure and the ecological degradation it has caused (Doyle et al., 2008).

Still others have pointed out that compact city designs intended to promote sustainability (e.g., to promote energy efficiency and reduce emissions of greenhouse gases) may actually undermine disaster resilience by putting more people in high-density developments located in floodplains and other hazardous locations (Berke et al., 2009).

UNANSWERED QUESTIONS

How can infrastructure systems be designed to both reduce risk and support more sustainable cities? How can infrastructure systems be designed for disaster resilience—for today, as well as for the future? These questions may be the most difficult, and the most important, to answer. Addressing them will require interdisciplinary research that spans the distances between engineering fields and between engineering and the social sciences.

REFERENCES

American Lifelines Alliance. 2006. Power Systems, Water, Transportation and Communications Lifelines Interdependencies. March. Available online at *http://www.cimap.vt.edu/2DOC/ALA%20Lifeline%20Report%20Final%20Draft%20030606.pdf*.

Berke, P.R., Y. Song, and M. Stevens. 2009. Integrating hazard mitigation into new urban and conventional developments. Journal of Planning Education and Research 28(4): 441–455.

Brink, S., R.A. Davidson, and T.H.P. Tabucchi. 2009. Estimated Durations of Post-Earthquake Water Service Interruptions in Los Angeles. Pp. 539–550 in Proceedings of TCLEE 2009: Lifeline Earthquake Engineering in a Multihazard environment. Reston, Va.: American Society of Civil Engineers.

Bruneau, M., S.E. Chang, R.T. Eguchi, G.C. Lee, T.D. O'Rourke, A.M. Reinhorn, M. Shinozuka, K. Tierney, W.A. Wallace, and D. von Winterfeldt. 2003. A framework to quantitatively assess and enhance the seismic resilience of communities. Earthquake Spectra 19(4): 733–752.

Chang, S.E. In press. Urban disaster recovery: a measurement framework with application to the 1995 Kobe earthquake. Disasters (in press).

Chang, S.E., and N. Nojima. 2001. Measuring post-disaster transportation system performance: the 1995 Kobe earthquake in comparative perspective. Transportation Research Part A: Policy and Practice 35(6): 475–494.

Chang, S.E., C. Pasion, K. Tatebe, and R. Ahmad. 2008. Linking Lifeline Infrastructure Performance and Community Disaster Resilience: Models and Multi-Stakeholder Processes. Technical Report MCEER-08-0004. Buffalo, NY: Multidisciplinary Center for Earthquake Engineering Research.

Doyle, M.W., E.H. Stanley, D.G. Havlick, M.J. Kaiser, G. Steinbach, W.L. Graf, G.E. Galloway, and J.A. Riggsbee. 2008. Aging Infrastructure and Ecosystem Restoration. Science 319(5861): 286–287.

Gordon, P., H.W. Richardson, and B. Davis. 1998. Transport-related impacts of the Northridge earthquake. Journal of Transportation and Statistics 1(2): 21–36.

Guikema, S. 2009. Infrastructure design issues in disaster-prone regions. Science 323(5919): 1302–1303.

Infrastructure Canada, Research and Analysis Division. 2006. Adapting Infrastructure to Climate Change in Canada's Cities and Communities. December. Available online at *http://www.infc.gc.ca/research-recherche/results-resultats/rs-rr/rs-rr-2006-12_02-eng.html*.

Kates, R.W., C.E. Colten, S. Laska, and S.P. Leatherman. 2006. Reconstruction of New Orleans after Hurricane Katrina: a research perspective. Proceedings of the National Academy of Sciences 103(40): 14653–14660.

Kousky, C., and R. Zeckhauser. 2006. JARring Actions that Fuel the Floods. Pp. 59–76 in On Risk and Disaster: Lessons from Hurricane Katrina, edited by R.J. Daniels, D.F. Kettl, and H. Kunreuther. Philadelphia, Pa.: University of Pennsylvania Press.

McDaniels, T., S. Chang, K. Peterson, J. Mikawoz, and D. Reed. 2007. Empirical framework for characterizing infrastructure failure interdependencies. Journal of Infrastructure Systems 13(3): 175–184.

Munnell, A.H. 1992. Infrastructure investment and economic growth. Journal of Economic Perspectives 6(4): 189–198.

NIST (National Institute of Standards and Technology). 2008. Strategic Plan for the National Earthquake Hazards Reduction Program: Fiscal Years 2009–2013. Gaithersburg, Md.: NIST.

NRC (National Research Council). 2006. Facing Hazards and Disasters: Understanding Human Dimensions. Washington, D.C.: National Academies Press.

O'Rourke, T.D., A.J. Lembo, and L.K. Nozick. 2003. Lessons Learned from the World Trade Center Disaster about Critical Utility Systems. Pp. 269–290 in Beyond September 11th: An Account of Post-Disaster Research, edited by J.L. Monday. Boulder, Colo.: Natural Hazards Research and Applications Information Center.

Rinaldi, S.M., J.P. Peerenboom, and T.K. Kelly. 2001. Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE Control Systems Magazine 21(6): 11–25.

Romero, N., T.D. O'Rourke, L.K. Nozick, and C.A. Davis. 2009. Los Angeles Water Supply Response to 7.8 M_W Earthquake. Pp. 1256–1267 in Proceedings of TCLEE 2009: Lifeline Earthquake Engineering in a Multihazard Environment. Reston, Va.: American Society of Civil Engineers.

U.S.-Canada Power System Outage Task Force. 2006. Final Report on the Implementation of the Task Force Recommendations. September. Available online at *http://www.ferc.gov/industries/electric/indus-act/blackout/09-06-final-report.pdf*.

Webb, G.R., K.J. Tierney, and J.M. Dahlhamer. 2000. Businesses and disasters: empirical patterns and unanswered questions. Natural Hazards Review 1(2): 83–90.

The Environmental Footprint of Infrastructure

Arpad Horvath
University of California, Berkeley

Infrastructure, the backbone of everyday life, has contributed immeasurably to human progress. But as we operate, expand, and maintain infrastructure, we must also recognize its huge environmental "footprint": greenhouse gas emissions, toxic discharges, water use and pollution, waste generation, human and ecological health impacts, and so on. In the face of climate change and dwindling resources, we must reduce this environmental footprint and maximize the benefits of our infrastructure.

INTRODUCTION

Operating, expanding, and maintaining infrastructure costs a substantial fraction of our gross national product. Although these costs are enormous, we also reap enormous benefits from our infrastructure. In fact, we cannot function without it. Infrastructure is everywhere, but we tend to notice it only when something goes wrong, when a tap goes dry, when a bridge collapses, when phone lines go silent. Few of us stop to marvel at what infrastructure makes possible.

In fact, most people do not have a correct understanding of the concept of infrastructure. To many, the word means only the transportation and utility systems. But infrastructure is much, much more than that. It includes the buildings we occupy, the telecommunications system and the Internet, the energy economy, and the health care system.

Traditionally, society has worried about the technical, economic, safety, and quality performance of infrastructure systems. Lately, we have become interested in its economic costs, societal costs, and environmental costs.

SUSTAINABILITY—A PARADIGM SHIFT[1]

The U.S. interstate highway system (IHS) launched by the Federal Aid Highway Act in 1956 became the largest infrastructure project the world had ever seen. It was said that IHS would increase the defense readiness of the United States by providing reliable highways for the movement of military personnel and equipment and provide a backbone for interstate commerce. By the time the system was completed some 37 years and more than $400 billion later, it had far exceeded those expectations.

The new highways crisscrossing America from coast to coast and north to south contributed to unprecedented growth in the productivity of U.S. industry (by some accounts, 31 percent in the late 1950s; 25 percent in the 1960s; and 7 percent in the 1980s) (Anonymous, 2008). Interstate highways have made more things possible than anyone predicted. They changed American life forever.

But infrastructure must be periodically reinvented. The flagship investment of the 1990s was the Internet, which has profoundly changed the way we live, work, and play. Like IHS, the Internet has created jobs and eliminated jobs, opened new horizons, spun off new activities, increased productivity, brought many along on virtual road trips, and connected the world in *Second Life* (http://*www.secondlife.com*). The Internet became the world's superhighway much faster than the interstate highways were able to connect the far ends of the United States. And the electronics and telecommunications industries became essential components of all the industries and systems in our infrastructure.

Today, the United States—as well as the rest of the world—needs breakthrough investments in the built environment on the scale and level of relevance of the IHS and the Internet. We need a paradigm shift in our built environment to the paradigm of sustainable development. Using life-cycle assessment (LCA), a systematic methodology that reveals the environmental impacts of every life-cycle stage of every component in the infrastructure, we have learned that buildings, transportation, and water and wastewater systems are responsible for the largest use of energy, water, and raw materials in the world. Physical infrastructure, which was once hailed as a high achievement, now causes people to turn away. Who wants to live near a highway? Or a power plant? Or a water-treatment facility?

Changing the environmental performance of infrastructure will take more knowledge and quicker action. Even as industries try to establish and mitigate their environmental footprint (e.g., industries in the European Union affected by

---

[1] This section first appeared on the web as *http://www.sitra.fi/en/News/articles/Article_2008-03-19.htm*.

carbon constraints and industries in the United States affected by sulfur dioxide (SO_2) regulations), it is high time for infrastructure professionals and industries to get organized and move forward with evidence-based arguments and recommendations for sustainable development.

ENVIRONMENTAL MODELING OF INFRASTRUCTURE

The first things we will need are models of infrastructure. In all models, the real world must be abstracted to make problems understandable, solvable, and manageable. Environmental modeling of infrastructure starts with modeling the complex systems that underlie transportation, energy, water, wastewater, municipal and industrial waste management, telecommunications, and other systems and services. Modeling infrastructure is doubly difficult because most infrastructures are networks of systems with large geographical and temporal dimensions. For example, the electric power system in the United States is more than 100 years old and reaches every corner of this vast country.

Most of these systems, even though they lie within the geographical boundaries of a country, connect with international systems and have supply chains that span the globe. For example, the U.S. civilian air transportation system is closely connected to international civil aviation. Most of the aircraft owned by U.S. airlines were assembled here, but many critical components were produced (or partly produced) in other countries. Mapping these supply chains is often difficult.

Many products and processes, and therefore many economic sectors, are involved in the life cycle (planning and design, construction, operation, maintenance, end of life) of infrastructure systems. Analyzing thousands of products is difficult, but assessing services is even more difficult because descriptions, process models, and data are scarce. In addition, technologies are far from uniform across the United States, let alone internationally, and there are variations in all life-cycle phases. Thus modeling them all is a daunting task. We design, build, and operate infrastructure differently from one state to the next, not to mention from one country to the next, and these differences must be captured in abstracted models.

Once models have been created and tested, they can be translated into basic process-flow diagrams for environmental analyses. A few years ago I attempted to find a comprehensive process-flow diagram for making, delivering, placing, and curing concrete. Although we make and use one billion tons of concrete a year in the United States and 10 billion tons worldwide, I could not find a single model. Mapping the process flow of the Internet and the supply chains that contribute to its (mostly flawless) operation is an even more complex job. Clearly, we have a lot to do in the first step towards environmental analysis.

LIFE-CYCLE ASSESSMENT

Many decision makers are aware of LCAs and their potential to provide infrastructure assessments. In principle, an LCA must quantify all of the resource inputs and environmental outputs of a product, process, or service, not just at the point of manufacture or generation, but through the entire underlying supply chain. Unfortunately, few have used LCA in practice, and it has rarely been applied to real-life situations.

LCA was conceptualized to change practices throughout the entire economy, protect human and ecological health, preserve resources, and support sustainable development. We need to begin using LCA to make infrastructure decisions. Billions of dollars must be invested in infrastructure, but they must be invested to improve environmental quality. Although there is a general consensus among societal stakeholders that this should be done, essential questions remain to be answered by LCA and other kinds of environmentally informed decision making.

Here are some sample questions for which we have no definitive answers:

- Is solar-generated electricity more environmentally friendly than wind-generated electricity?
- Is traveling by high-speed rail more environmentally efficient than flying or driving?
- Should roadways be built to follow natural topographies or should we use cuts, fills, and tunnels wherever needed?
- Does centralized wastewater treatment have lower energy requirements than decentralized treatment?
- Should we build concrete or steel bridges?
- What are the life-cycle environmental emissions of the U.S. telecommunications system?
- Should water reservoirs be built on hills or on flat land?

These are just a few of the infrastructure decisions that must be pondered every day, and answering these questions correctly or incorrectly can have profound environmental implications. Yet for many of these types of questions we do not know whether one alternative is more environmentally friendly than another, and we are even farther from being able to supply robust answers based on absolute numbers. Although the temporal and spatial complexities of these kinds of questions are daunting, we must answer them, and do so correctly, lest we continue making decisions without forethought that may result in further environmental damage. Sadly, we are still many years from answers with an acceptable degree of certainty.

AN EXAMPLE: LIFE-CYCLE ASSESSMENT OF PASSENGER TRANSPORTATION MODES

This example illustrates the complexities and challenges of environmental assessment, especially using LCA. In a recent paper (Chester and Horvath, 2009), we explored the total energy, greenhouse gas (GHG), SO_2, nitrous oxide (NO_x), and carbon dioxide (CO_2) burden of several types of passenger transportation: sedan; SUV; pickup truck; city bus; Bay Area Rapid Transit (BART) (considered a subway); commuter rail (Caltrain); and light rail (MUNI). For comparison, we included small, midsize, and large aircraft, and the Boston Green Line light rail (the oldest line of the Boston subway and the most heavily traveled light-rail line in the United States). The results are presented in Figure 1.

Our analysis was based not only on tailpipe emissions, a typical measure of vehicle emissions, but also on the provision of vehicles, the transportation infrastructure, and fuels, in other words, all of the life-cycle phases of vehicles, the physical infrastructure that makes travel possible (e.g., roads, rails, stations,

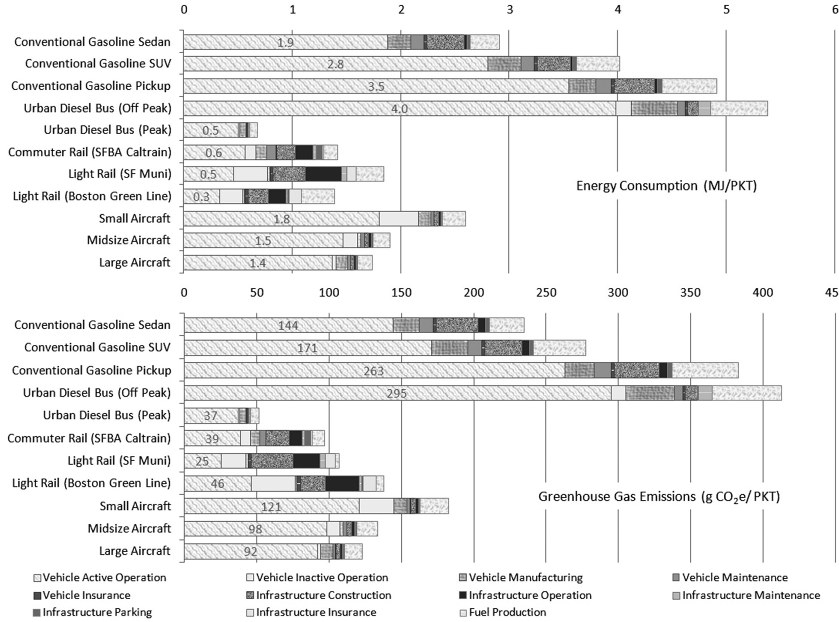


FIGURE 1 Energy use and greenhouse gas emissions per passenger kilometer traveled (PKT) for various modes of passenger transportation. Source: Chester and Horvath, 2009.

airports, etc.), and the stages prior to the combustion of fuels or generation of electricity.

Altogether 79 components associated with these transportation modes were analyzed (about 200 different calculations), representing a great variety of technologies, products, processes, and services in the Bay Area, the state of California, and Boston, as well as nationwide and worldwide, wherever the supply chains for these components reach. Thus assessing the environmental performance of these systems was challenging for many reasons, not least because the data were numerous and spanned many decades and were generally difficult to obtain. For example, components of some systems were built decades ago, and some are shared with freight transportation.

Our analysis showed that physical infrastructure and the provision of fuel contributed significantly to environmental effects. For example, they add 63 percent to road, 155 percent to rail, and 31 percent to air transport's total burden, above and beyond tailpipe GHG emissions, as expressed in passenger kilometers. In addition, usage rates make a big difference in the total environmental burden.

WHAT'S NEXT?[2]

Reducing the environmental footprint of infrastructure presents us with enormous tasks, but also enormous opportunities. Most of the energy we use in constructing, operating, maintaining, retrofitting, and decommissioning infrastructure is based on fossil fuels. We already know that carefully selected, produced, and used biofuels can reduce our fossil-fuel demand. Large-scale efforts, such as the Energy Biosciences Institute at the University of California, Berkeley (*http://www.energybiosciencesinstitute.org*), could lead to breakthrough technologies for infrastructure.

We must work with the supply chains of infrastructure systems to understand how we can reduce the environmental impacts of raw materials, equipment, products, and services before they are built into infrastructures. We also need to educate current students and the leaders of our society in how to achieve the highest environmental standards in the infrastructure organizations they support, run, finance, or regulate.

Environmental assessment of infrastructure was introduced in the scientific literature decades ago, and LCA studies were introduced about 15 years ago. But we must support current researchers and groom future researchers to improve their understanding of the complex, ever-changing set of systems that comprises our infrastructure. To answer even the most burning infrastructure questions, we will need many more studies that can lead to recommendations that can be readily adopted by practitioners.

---

[2] A version of this section first appeared on the web as *http://www.sitra.fi/en/News/articles/Article_2008-03-19.htm*.

Investing in the new paradigm—that the future must be based on principles of sustainability—is the most significant investment we will make in this generation. The question is how far this investment will take us toward reaching our goal.

ACKNOWLEDGMENTS

This material is based on research supported by the UC Berkeley Center for Future Urban Transport (A Volvo Center of Excellence).

REFERENCES

Anonymous. 2008. America's Splurge. *The Economist*, February 14.

Chester, M.V., and A. Horvath. 2009. Environmental assessment of passenger transportation should include infra-structure and supply chains. Environmental Research Letters 4: 024008 (8pp), doi: 10.1088/1748-9326/4/2/024008. Available online at *http://www.iop.org/EJ/abstract/1748-9326/4/2/024008/*.

APPENDIXES

Contributors

Elmer Bernstam is an associate professor of healthy informatics and internal medicine at the University of Texas at Houston. His primary appointment is in the School of Health Information Sciences with a secondary appointment in Internal Medicine. He received an M.D. from the University of Michigan Medical School in 1995, and an M.S. in computer science and engineering from the University of Michigan in 1999. His research focuses on MEDLINE search, consumer health informatics, shareable executable clinical practice guidelines, and informatics in clinical and translational research.
http://www.sahs.uth.tmc.edu/evbernstam/

Stephanie Chang is an associate professor in the Institute for Resources, Environment, and Sustainability at the University of British Columbia. She received a Ph.D. in regional science from Cornell University. Her research and teaching interests are in urban planning, geography, and regional science; natural disasters, risk management, and GIS decision-support systems; human-environmental interactions; transportation and other urban infrastructure systems.
http://www.ires.ubc.ca/?p2=/modules/ires/profile.jsp&id=2

David Dorr is an assistant professor of medical informatics and clinical epidemiology at Oregon Health and Science University School of Medicine. He received an M.D. from Washington University in St. Louis, Missouri, and after completing

his residency, received a master's degree in medical informatics and health science administration from the University of Utah. His research is in care management, coordination of care, collaborative care, chronic disease management, quality, and the requirements of clinical information systems to support these areas. He has broadened into clinical information needs, electronic health record deployment, and health information exchange as ways to expand systems-based approaches to health care.
http://www.ohsu.edu/ohsuedu/academic/som/dmice/people/faculty/dorr.cfm

Riley Duren is the chief systems engineer for the Earth Science and Technology Directorate and a principal staff member at the Jet Propulsion Laboratory in Pasadena, California. From 2002 until 2009, he served as the chief engineer for NASA's Kepler mission, which launched in March 2009 and is conducting a 3.5 year mission to search for habitable planets around other stars. His current focus is applying systems engineering techniques to the challenge of climate change and in this capacity, is leading an interagency effort to define the policy–relevant needs and a design for a global greenhouse gas observing system to support climate treaties, cap-and-trade programs, and adaptation planning.

Claire Gmachl is a professor of electrical engineering at Princeton University. She received a Ph.D. from the Technical University of Vienna in 1995. Her research group is working on the development of new quantum devices, especially lasers, and their optimization for systems applications ranging from sensors to optical communications. Their special focus is on Quantum Cascade (QC) lasers, a novel type of semiconductor injection laser based on electronic intersubband transitions in the conduction band of a coupled quantum well heterostructure.
http://www.ee.princeton.edu/people/Gmachl.php

Carla Gomes is director of the Institute for Computational Sustainability and associate professor with joint appointments in Computer Science, Information Science, and Applied Economics and Management. She obtained a Ph.D. in computer science in the area of artificial intelligence and operations research from the University of Edinburgh. Her central research themes are the integration of concepts from mathematical programming, constraint programming, and algorithm theory for large-scale combinatorial problems; the study of the impact of structure on problem hardness; and the use of randomization techniques to improve the performance of exact (complete) search methods.
http://www.cs.cornell.edu/gomes/

Arpad Horvath is an associate professor in the Department of Civil and Environmental Engineering at the University of California, Berkeley. He received a Ph.D. in civil engineering from Carnegie Mellon University in 1997. His research interests are in developing methods and tools for environmental and economic

analysis of civil infrastructure systems, primarily for the construction industry and the built environment. His research has focused on the environmental implications of the construction, electronics and various service industries, life-cycle assessment modeling using environmentally augmented economic input-output analysis, and environmental performance measurement.
http://www.ce.berkeley.edu/~horvath/

Mike Krames is chief technology officer at Soraa, Inc. in Goleta, California. Previous positions have included director, Advanced Laboratories at Philips Lumileds Lighting Co., and engineering scientist and hardware design engineer in the Optoelectronics Division at Agilent Technologies. He received a Ph.D. from the University of Illinois at Urbana-Champaign in electrical and computer engineering.

Vinothan Manoharan is an assistant professor of physics and chemical engineering at Harvard University. He received a Ph.D. in 2004 from the University of California, Santa Barbara. His research explores light scattering, optical microscopy, spectroscopy, synthesis and other experimental techniques to understand the physics of self-organization. For most of these experiments, his lab uses colloids, suspensions of particles typically about a micrometer in size. They seek to understand how very simple systems of micro- or nano-particles can assemble themselves into ordered structures. By doing so, they might be able to control the process in order to build new kinds of optical materials and photonic devices, and more generally, figure out the physical 'rules' for making nanoscale materials through self-assembly.
http://www.physics.harvard.edu/people/facpages/manoharan.html
http://manoharan.seas.harvard.edu

Genevieve Melton-Meaux is a faculty fellow at the Institute for Health Informatics and an assistant professor in the Department of Surgery at the University of Minnesota. She holds undergraduate degrees in computer science, mathematics, and electrical engineering and received an M.D. from the Johns Hopkins University School of Medicine in 2000. Her research is on improving the use of patient data from electronic health records to enhance the quality, research, and outcomes of health care. She utilizes informatics tools that allow data to be used more effectively for large-scale analyses, such as biomedical ontologies/terminologies, knowledge representation, and text-mining.
http://www.ihi.umn.edu/aboutihi/People/Genevieve%20Melton-Meaux.html

Lukas Novotny is a professor of optics and of biomedical engineering and physics at the University of Rochester. He earned a Dipl. El-Ing (M.S. in electrical engineering) in 1992 and a Dr. sc. techn. (Doctor of Technical Sciences) in 1996, both from the Swiss Federal Institute of Technology (ETH) in Zurich. His research is in

the area of experimental and theoretical optics. He is interested in the application of optical science and technology to the study of nanoscale phenomena ranging from solid-state physics to biology.
http://www.optics.rochester.edu/people/faculty_students_staff/faculty/novotny.html

Lucila Ohno-Machado is professor of medicine and chief of the Division of Biomedical Informatics at the University of California, San Diego. She received a Ph.D. in medical information sciences and computer science from Stanford University. Her research is in the area of validation of breast cancer biomarkers, assessing the quality of individual estimates in decision support systems, and scalable medical alert response technology.
http://dbmi.ucsd.edu/confluence/display/BMI/Lucila+Ohno-Machado

Oskar Painter is an associate professor of applied physics at the California Institute of Technology. He received a Ph.D. from Caltech in 2001. His current research focuses on the interesting ways that light propagates within microfabricated high-contrast periodic dielectric and metallic structures.
http://www.aph.caltech.edu/people/painter_o.html

Bradford Parkinson is Edward C. Wells Professor of Aeronautics and Astronautics, Emeritus at Stanford University. He was the principal advocate and chief architect for the Global Positioning System (GPS) in the early 1970s. He was the first GPS program director and led GPS development through the first satellite launches. He is a graduate of the U.S. Naval Academy and has a masters degree from the Massachusetts Institute of Technology and a Ph.D. from Stanford University in astronautics (1966). As a professor at Stanford University, he led the development of many innovative civil applications of GPS. He was the CEO of two companies and serves on many boards. Among his many awards is the Draper Prize of the U.S. National Academy of Engineering, considered by some to be the "Engineering Nobel."

Kristina Swallow is owner of AVP Consulting, Inc. in Las Vegas, Nevada. She received an M.S. in civil engineering with an emphasis in transportation from the University of Nevada, Las Vegas. AVP Consulting provides flood control, air quality, and transportation studies for public and private sector clients. She has been appointed an American Society of Civil Engineers Congressional Fellow for 2009-2010.

Sean Wiggins is a project scientist and principal development engineer at the Marine Physical Laboratory, Scripps Institution of Oceanography, University of California, San Diego. He received a Ph.D. in oceanography from the Scripps Institution of Oceanography. His research interests are in marine mammal acous-

tics, passive acoustic monitoring, ambient ocean noise, effects of sound on marine mammals, ocean instrumentation and vehicles, and high-resolution, near-surface geophysical imaging.
http://cetus.ucsd.edu/profile_SeanWiggins.html
http://sio.ucsd.edu/Profile/swiggins
http://cetus.ucsd.edu/

Program

NATIONAL ACADEMY OF ENGINEERING

2009 U.S. Frontiers of Engineering Symposium
September 10–12, 2009

Chair: Andrew M. Weiner, Purdue University

ENGINEERING TOOLS FOR SCIENTIFIC DISCOVERY

Organizers: J. Christopher Love, Naren Ramakrishnan, and Louise Hamlin

Digital Holographic Microscopy for 3D Imaging of Complex Fluids and Biological Systems
Vinothan Manoharan

Engineering Tools for Studying Marine Mammals
Sean Wiggins

The Kepler Mission: A Search for Terrestrial Planets
Riley Duren

Computational Sustainability: Computational Methods for a Sustainable Environment, Economy, and Society
Carla Gomes

***

NANO/MICRO PHOTONICS AND NEW APPLICATIONS
Organizers: Nelson Tansu and Michal Lipson

Optical Antennas:
A New Technology That Can Enhance Light-Matter Interactions
Lukas Novotny

Nano-Opto-Mechanics: Using Light Forces in Guided-Wave Nanostructures
Oskar Painter

Engineering Light: Intersubband Optoelectronics
Claire Gmachl

Light-Emitting Diode Technology for Solid-State Lighting
Mike Krames

***

ENGINEERING THE HEALTH CARE DELIVERY SYSTEM
Organizers: Stephanie Guerlain and Eva K. Lee

Why Health Information Technology Doesn’t Work
Elmer Bernstam

Calibration in Computer Models for Medical Diagnosis and Prognostication
Lucila Ohno-Machado

Medical Informatics for Detecting Adverse Events
Genevieve Melton

Managing and Coordinating Health Care:
Creating Collaborative, Proactive Systems
David Dorr

***

RESILIENT AND SUSTAINABLE INFRASTRUCTURE

Organizers: Seth Guikema and Patrick O'Mara

America's Infrastructure Report Card: Causes, Costs, and Solutions
Kristina Swallow

Infrastructure Resilience to Disasters
Stephanie Chang

The Environmental Footprint of Infrastructure
Arpad Horvath

***

DINNER SPEECH

Developing GPS: Conception, Applications, Surprises, and the Future
Bradford W. Parkinson

Participants

NATIONAL ACADEMY OF ENGINEERING

2009 U.S. Frontiers of Engineering Symposium
September 10–12, 2009

Nicholas Able
Senior Development Engineer
Global Product Managment -
Technology Development
Ingersoll Rand

Yves Baeyens
Technical Manager
Physical Technologies Domain
Alcatel-Lucent, Bell Labs

Sumit Basu
Researcher
Microsoft Research

Sandor Becz
Principal Investigator
Thermal and Fluid Sciences
United Technologies Research Center

Elmer Bernstam**
Associate Professor
School of Health Information Sciences
University of Texas, Houston

Dan Belov
Software Engineer
Google

Randall Booker
Senior Engineer
Treatment Process and Planning
Malcolm Pirnie, Inc.

Jason Burdick
Wilf Family Term Assistant Professor
Department of Bioengineering
University of Pennsylvania

---

*Organizing Committee
**Speaker

Ali Butt
Assistant Professor
Department of Computer Science
Virginia Tech

Mark Byrne
Mary & John H. Sanders Associate Professor
Department of Chemical Engineering
Auburn University

Stephanie Chang**
Associate Professor
School of Community and Regional Planning
University of British Columbia

Thomas Chapman
Development Associate, Materials & Process
Diesel Technologies Development
Corning, Incorporated

Yong Chen
Assistant Professor
Epstein Department of Industrial and Systems Engineering
University of Southern California

Tanzeem Choudhury
Assistant Professor
Department of Computer Science
Dartmouth College

Brian Coombe
Project Manager
Bechtel National, Inc.

Dilma da Silva
Research Staff Member
Advanced Operating Systems Group
IBM T.J. Watson Research Center

Michael Degner
Technical Leader, Electric Machine Drives
Hybrid Electric Vehicle & Fuel Cell Vehicle Propulsion Systems
Ford Motor Company

Steven Demetropoulos
Staff Services Traffic Engineer
Engineering Department - Traffic Engineering
Port Authority of New York and New Jersey

David Dorr**
Department of Medical Informatics and Clinical Epidemiology
Oregon Health and Science University

Riley Duren**
Chief Systems Engineer
Earth Science and Technology Directorate
Jet Propulsion Laboratory

Dan Ellis
Associate Professor
Department of Electrical Engineering
Columbia University

David Erickson
Assistant Professor
Sibley School of Mechanical and Aerospace Engineering
Cornell University

Rene-Jean Essiambre
Distinguished Member of Technical Staff
Optical Networks Domain
Alcatel-Lucent, Bell Labs

Eric Furst
Associate Professor
Department of Chemical Engineering
University of Delaware

Ariel Fuxman
Researcher
Microsoft Research

Evgeniy Gabrilovich
Senior Research Scientist
Yahoo! Research

Manuel Garcia-Leiner
Research Scientist
Corporate Research
Arkema, Inc.

Claire Gmachl**
Associate Professor
Department of Electrical Engineering
Princeton University

Carla Gomes**
Associate Professor
Department of Computer Science and Computing
Cornell University

Julia Greer
Assistant Professor
Department of Materials Science and Mechanics
California Institute of Technology

Andrew Guenthner
Head, Polymer Science and Engineering Branch
Weapons Division
Department of the Navy, Naval Air Warfare Center

Stephanie Guerlain*
Associate Professor
Department of Systems and Information Engineering
University of Virginia

Seth Guikema*
Assistant Professor
Department of Geography and Environmental Engineering
Johns Hopkins University

Louise Hamlin*
Technical Manager
Next-Generation Image Spectrometers
NASA Jet Propulsion Laboratory

Jocelyn Hicks-Garner
Research Staff Member
Energy Technologies
HRL Laboratories, LLC

Mark Hilburger
Research Engineer
Research and Technology Directorate
NASA Langley Research Center

Gabriel Hoffmann
Member of Research Staff
Intelligent Systems Laboratory
Palo Alto Research Center

Tracy Honeycutt
Senior Engineer
Nuclear Licensing
Southern Nuclear Operating Company

Arpad Horvath**
Associate Professor
Department of Civil and Environmental Engineering
University of California, Berkeley

Jilei Hou
Senior Staff Engineer/Manager
Corporate R & D
Qualcomm, Inc.

Gareth Jolly
Director
Manufacturing Engineering General Assembly, Global Bill of Process & Bill of Equipment
General Motors, North America

Ravi Kane
P.K. Lashmet Professor
Department of Chemical and Biological Engineering
Rensselaer Polytechnic Institute

Jordan Katine
Research Staff Member
Nano Processing and Recording Heads
San Jose Research Center
Hitachi Global Storage Technologies

David Kerxhalli
Electromechanical Systems Engineer/System Architect
System Engineering & Integration Group
Xerox Research Center

Youngmoo Kim
Assistant Professor
Department of Electrical and Computer Engineering
Drexel University

Gerhard Klimeck
Director and Professor
Network for Computational Nanotechnology
Purdue University

Farinaz Koushanfar
Assistant Professor
Department of Electrical and Computer Engineering
Rice University

Michael Krames**
Chief Technology Officer
Soraa, Inc.

Ramanan Krishnamoorti
Professor and Department Chair
Department of Chemical and Biomolecular Engineering
University of Houston

Terrence Ladau
Senior Project Leader
Strategic Space Operations
The Aerospace Corporation

Jacob Langelaan
Assistant Professor
Department of Aerospace Engineering
Pennsylvania State University

Blake Larson
Executive Engineer, High Horsepower Engineering
Engine Business
Cummins Inc.

Catherine Leatherdale
Research Manager
Projection Systems Department
3M Company

Eva Lee*
Associate Professor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Michal Lipson*
Associate Professor
Department of Electrical and Computer Engineering
Cornell University

J. Christopher Love*
Texaco-Mangelsdorf Career Development Assistant Professor
Department of Chemical Engineering
Massachusetts Institute of Technology

Tonya Love
Area Manager
Xerox Innovation Group
Xerox Research Center, Webster

Hang Lu
Assistant Professor
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Helen Lu
Associate Professor
Department of Biomedical Engineering
Columbia University

Christian Lutkemeyer
Senior Principal Scientist
Office of the CTO / DSP Microelectronics Technology
Broadcom Corporation

SuPing Lyu
Principal Scientist
Medtronic Strategy and Innovation
Medtronic

Vinothan Manoharan**
Assistant Professor
Department of Physics
Harvard University

Sara McComb
Associate Professor
Department of Industrial and Systems Engineering
Texas A&M University

Beverley McKeon
Assistant Professor of Aeronautics
Engineering and Applied Science Division
California Institute of Technology

Genevieve Melton-Meaux**
Assistant Professor of Surgery
Division of Colon and Rectal Surgery
University of Minnesota

Edward Miller, Jr.
Vice President of Research and Development
Applied BioSciences
DuPont Tate & Lyle Bio Products, LLC

Subhasish Mitra
Assistant Professor
Departments of Electrical Engineering and Computer Science
Stanford University

Jeff Moehlis
Associate Professor
Department of Mechanical Engineering
University of California, Santa Barbara

Andreas Molisch
Professor
Department of Electrical Engineering
University of Southern California

William Murphy
Assistant Professor
Department of Biomedical Engineering
University of Wisconsin, Madison

Mangapathirao Mynam
Research Engineer
Research & Development
Schweitzer Engineering Laboratories, Inc.

Vijay Narayanan
Manager, High-K/Metal Gate Process Development
Materials Science
IBM T. J. Watson Research Center

Alshakim Nelson
Research Staff Member
Advanced Organic Materials
IBM Research, Almaden

Ken Nishimura
R&D Program Manager
Agilent Laboratories
Agilent Technologies, Inc.

Lukas Novotny**
Professor
Institute of Optics
University of Rochester

Lucila Ohno-Machado**
Professor and Chief
Division of Biomedical Informatics
University of California, San Diego

Patrick O'Mara*
Senior Transportation Engineer
Transportation & Infrastructure - Planning
STV, Inc.

Stephen Paglieri
Technical Staff Member: R&D Engineer
Weapons Engineering Technology
Los Alamos National Laboratory

Oskar Painter**
Associate Professor
Department of Applied Physics
California Institute of Technology

John Palamara
Senior Principal Process Engineer
Global Engineering - Energy Businesses
Air Products and Chemicals, Inc.

Babak Parviz
Associate Professor
Department of Electrical Engineering
University of Washington

Glenn Paskoff
Biomechanics Engineer
Naval Air Warfare Center, Aircraft Division
Department of the Navy, Naval Air Systems Command

Naren Ramakrishnan*
Professor and Associate Head
Department of Computer Science
Virginia Tech

Meredith Reed
Electronics Engineer/Researcher
Sensors and Electron Devices Directorate
U.S. Army Research Laboratory

William Regli
Professor
Department of Computer Science
Drexel University

Maureen Reitman
Principal
Mechanics and Materials Practice
Exponent - Failure Analysis Associates

Mark Rintoul
Manager
Discrete Math and Complex Systems
Sandia National Laboratories

Nicholas Roy
Associate Professor
Department of Aeronautics and Astronautics
Massachusetts Institute of Technology

Samuel Sadoulet
Vice President
Engineering
Edmund Optics, Inc.

Marc Schaepkens
Engineering Manager
Global X-ray Detectors
GE Healthcare

Fred Schauer
Principal Mechanical Engineer
Propulsion Directorate
Air Force Research Laboratory

Jon Schell
Principal Mechanical Design Engineer
Cardiac Rythm Management
Medtronic

James Schmiedeler
Associate Professor
Aerospace and Mechanical Engineering
University of Notre Dame

Sudipta Seal
Professor
Advanced Materials Processing Analysis Center, Nanoscience and Technology Center
Mechanical, Materials and Aerospace Engineering
University of Central Florida

William Slye
Senior R&D Manager - Primary Metals and Metals Fabrication
Industrial Applications & Healthcare R&D
Praxair, Inc.

Kristina Swallow**
Owner
AVP Consulting, Inc.

Nelson Tansu*
Associate Professor
Department of Electrical and Computer Engineering
Lehigh University

Mark Tompkins
Senior Engineer
Ecosystem Planning and Restoration
CH2M HILL

Eric Tseng
Technical Leader
Powertrain Controls Research & Advanced Engineering
Ford Motor Company

Kevin Uptain
Principal Engineer
Process Solutions Technology Development Center
Cargill, Incorporated

Krystyn Van Vliet
Thomas Lord Associate Professor of Materials Science and Engineering
Departments of Materials Science & Engineering and Biological Engineering
Massachusetts Institute of Technology

James Walker
Institute Scientist
Mechanical and Materials Engineering
Southwest Research Institute

Steven Waller
Associate Professor
Department of Civil, Architectural and Environmental Engineering
University of Texas at Austin

Marvin Warner
Senior Research Scientist
Physical and Chemical Sciences - Chemical and Biological Signature Sciences
Pacific Northwest National Laboratory

Andrew Weiner *
Scifres Family Distinguished Professor of Electrical and Computer Engineering
School of Electrical and Computer Engineering
Purdue University

Sean Wiggins**
Project Scientist
Marine Physical Laboratory
Scripps Institution of Oceanography

Eric Wilhelm
CEO
Development
Instructables

Ronggui Yang
Assistant Professor and Sanders Faculty Fellow
Department of Mechanical Engineering
University of Colorado, Boulder

Michael Yu
Associate Professor
Department of Materials Science and Engineering
Johns Hopkins University

Muhammad Zaman
Assistant Professor
Department of Biomedical Engineering
University of Texas, Austin

Huimin Zhao
Professor
Department of Chemical and Biomolecular Engineering
University of Illinois at Urbana-Champaign

Alan Zimmerman
Director
C4ISR Systems Engineering
Northrop Grumman Electronic System

Dinner Speaker

Bradford W. Parkinson
Edward C. Wells Professor of
Aeronautics and Astronautics,
Emeritus
Stanford University

The United States-Mexico Foundation for Science

Fernando Aboites
The Center on Advanced Technology

Carlos Aquirre
CICATA LEG.IPN

Antonieta Garcia
National Polytechnic Institute

Hilario Hernandez
National Polytechnic Institute

Jose Jimenez
National Polytechnic Institute

Javier Sánchez-Mondragón
National Institute for Astrophysics,
Optics and Electronics

Guests

Randy Avent
Chief Scientist
Basic Science Office
Department of Defense
Defense Research and Engineering

Bill Hayden
Vice President
The Grainger Foundation

Lisa McCracken
Program Manager, Research
Industry Relations
Google

Sohi Rastegar
Director, Emerging Frontiers in
Research and Innovation
National Science Foundation

Robin Staffin
Director, Basic Science
ODDR&E Research Directorate

Phil Szuromi
Senior Editor
Science Magazine

National Academy of Engineering

Lance Davis
Executive Officer

Janet Hunziker
Senior Program Officer

Penny Gibbs
Senior Program Associate