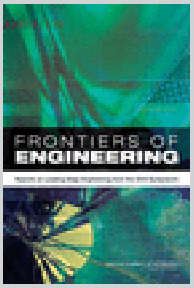


Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2013 Symposium

ISBN
978-0-309-29603-8

184 pages
6 x 9
PAPERBACK (2014)

National Academy of Engineering

FRONTIERS OF ENGINEERING

Reports on Leading-Edge Engineering from the 2013 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, NW • 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 DuPont, The Grainger Foundation, Microsoft Research, Defense Advanced Research Projects Agency, Air Force Office of Scientific Research, Department of Defense ASD(R&E)–Research Directorate-STEM Development Office, Cummins Inc., and the Greater Wilmington Convention and Visitors Bureau. This material is also based upon work supported by the National Science Foundation under Grant No.1305854. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. In addition, the content of this publication does not necessarily reflect the position or the policy of the government, and no official endorsement should be inferred.

International Standard Book Number-13: 978-0-309-29603-8
International Standard Book Number-10: 0-309-29603-X

Additional copies of this report are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; 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. C. D. Mote, Jr., 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. C. D. Mote, Jr., are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

ORGANIZING COMMITTEE

KRISTI ANSETH (Chair), Howard Hughes Medical Institute Investigator and Distinguished Professor of Chemical and Biological Engineering, University of Colorado, Boulder
HALIL BERBEROGLU, Assistant Professor, Department of Mechanical Engineering, University of Texas at Austin
TANZEEM CHOUDHURY, Associate Professor, Department of Information Science, Cornell University
ELIZABETH HOEGEMAN, Cummins Fuel Systems World Wide Manufacturing Leader, Cummins Inc.
SCOTT KLEMMER, Associate Professor, Department of Computer Science, University of California, San Diego
YUEH-LIN (LYNN) LOO, Professor, Department of Chemical Engineering, Princeton University
J. RHETT MAYOR, Associate Professor, George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology
TSE NGA (TINA) NG, Senior Research Scientist, Electronic Materials and Devices Laboratory, Palo Alto Research Center
STUART THOMAS, Global Technology and Licensing Manager, DuPont Cellulosic Ethanol, Industrial Biosciences, DuPont

Staff

JANET R. HUNZIKER, Senior Program Officer
VANESSA LESTER, Program Associate

Preface

This volume presents papers on the topics covered at the National Academy of Engineering's 2013 US Frontiers of Engineering Symposium. Every year the symposium brings together 100 outstanding young leaders in engineering to share their cutting-edge research and innovations in selected areas. The 2013 symposium was held September 19–21 and was hosted by DuPont in Wilmington, Delaware. The intent of this book is to convey the excitement of this unique meeting and to highlight innovative developments in engineering research and technical work.

GOALS OF THE FRONTIERS OF ENGINEERING PROGRAM

The practice of engineering is continually changing. Engineers must be able not only to thrive in an environment of rapid technological change and globalization but also to work on interdisciplinary teams. Today's 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.

At the annual 2½-day US Frontiers of Engineering Symposium, 100 of this country's best and brightest engineers—ages 30 to 45, from academia, industry, and government and a variety of engineering disciplines—learn from their peers about pioneering work in different areas of engineering. The number of participants 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 symposium is designed to foster contacts and learning among promising individuals who would not meet in the usual round of professional meetings. This networking may lead to collaborative work, facilitate the transfer of new techniques and approaches, and produce insights and applications that bolster US innovative capacity.

The four topics and the speakers for each year's meeting are selected by an organizing committee of engineers in the same 30- to 45-year-old cohort as the participants. Speakers describe the challenges they face and communicate the excitement of their work to a technically sophisticated but nonspecialist audience. They provide a brief overview of their field of inquiry; define the frontiers of that field; describe experiments, prototypes, and design studies (completed or in progress) as well as new tools and methods, limitations, and controversies; and assess the long-term significance of their work.

THE 2013 SYMPOSIUM

The topics covered at the 2013 symposium were (1) designing and analyzing societal networks, (2) cognitive manufacturing, (3) energy: reducing our dependence on fossil fuels, and (4) flexible electronics.

The session on societal networks addressed opportunities and challenges posed by the large-scale adoption of social technologies such as social networks, smart mobile devices, digital health, and online education. The first speaker covered the modeling of large-scale networks based on mobility data, examining how users' data profiles alone, without any connectivity information, can be used to infer their connectivity with others. The next presenter described several prototype crowd computing systems that harness the power of people on the Web to do tasks such as editing or reviewing computer code that may be too difficult or time-consuming for one person to do alone. The next talk, about how people on the ground during a disaster use mobile devices and social media platforms to share information about unfolding events, considered the challenge of getting real-time information to the right people at the right time. The session ended with a presentation on computational social science, the study of complex social systems through computational modeling and related techniques; noting that there has been little progress on "big" questions such as the dynamics of epidemics, the speaker described reasons for and suggested ways to address this problem.

Cognitive manufacturing refers to production systems that utilize "cognitive reasoning" engines or distributed intelligence agents that are capable of autonomous operation and require only high-level supervisory control. These systems can, for example, perceive changes in processes and effect adaptations to maintain target ranges of metrics such as production cost, rate, and energy consumption. The first presentation concerned the use of distributed anomaly detection agents to detect and remedy flaws without the need for fault-specific recognition, thus expediting correction and dramatically reducing factory downtime and associated costs. This was followed by a description of the adaptation and incorporation of business process management techniques to enhance decision making and system development in manufacturing. The third speaker covered the deployment of computer-enabled decision making at the production system logistics level to optimize global logistics methods and supply chain design. The final speaker discussed the application of computer-enabled cognitive manufacturing systems to support sustainable production systems.

The focus of the third session was energy issues, specifically efforts to reduce dependence on fossil fuels and to develop technical solutions for diversifying the fuel production infrastructure to meet energy needs. An overview talk surveyed technical, economic, environmental, social, and policy issues associated with dependence on fossil fuels as well as challenges for technology innovation. The second presenter reviewed advances and challenges in biofuel production technologies (e.g., for terrestrial biomass feedstock), infrastructure, and transportation. This was followed by an industry perspective on efforts to include nonfossil fuels in the fuel infrastructure. In the last talk the speaker described the mechanics and components of artificial solar fuel generation and the research needed to advance its use.

The symposium concluded with a session on flexible electronics, in which conventional fabrication processes have been transformed to incorporate electronic control and power sources into diverse materials, including surfaces that are soft, pliant, and easily damaged. The first speaker reported materials developments that have enabled the fabrication of optoelectronic devices and the design and processing strategies that have advanced their performance. This was followed by a presentation on the mechanics, materials, and biointegration of tissue-like electronics that can conform to and deform with living organisms for physiological sensing and stimulation. The session's final talk addressed next-generation flexible electrode arrays and optoelectronic neural scaffolds that minimize tissue damage and maintain high-quality neural recordings over longer timescales, which could facilitate monitoring and repair of damaged neural tissues.

In addition to the plenary sessions, the attendees had many opportunities for informal interaction. On the first afternoon, they gathered in small groups for "get-acquainted" sessions during which they presented short descriptions of their work and answered questions from their colleagues. This helped them get to know more about each other relatively early in the program. On the second afternoon attendees joined one of six technical tours—on biofuels, biomaterials, solar innovations, tire testing, automotive lightweighting, and performance polymers characterization—at DuPont's Experimental Station and Chestnut Run Plaza facilities.

Every year a distinguished engineer addresses the participants at dinner on the first evening of the symposium. The 2013 speaker was Dr. Douglas Muzyka, senior vice president and chief science and technology officer at DuPont. He provided historical background on the company and described its current focus on integrated science and engineering to develop solutions to challenges in food, energy, and safety of people and the environment. DuPont chair and CEO Ellen Kullman welcomed the group via a video message before Dr. Muzyka's presentation.

The NAE is deeply grateful to the following for their support of the 2013 US Frontiers of Engineering symposium:

- DuPont
- The Grainger Foundation
- Defense Advanced Research Projects Agency
- Air Force Office of Scientific Research
- Department of Defense ASD(R&E)–Research
- National Science Foundation (this material is based on work supported by the NSF under grant number 1305854)
- Microsoft Research
- Cummins Inc.
- Individual contributors

We also thank the members of the Symposium Organizing Committee (p. iv), chaired by Dr. Kristi Anseth, for planning and organizing the event.

Contents

DESIGNING AND ANALYZING SOCIETAL NETWORKS

COGNITIVE MANUFACTURING

ENERGY: REDUCING OUR DEPENDENCE ON FOSSIL FUELS

FLEXIBLE ELECTRONICS

APPENDIXES

DESIGNING AND ANALYZING SOCIETAL NETWORKS

Designing and Analyzing Societal Networks

Tanzeem Choudhury
Cornell University

Scott Klemmer
University of California, San Diego

Computing is increasingly woven into the fabric of everyday life. Many remarkable societal changes emerge as social technologies are adopted at massive scale—in social networks, smart mobile devices, digital health tools, online education. Socially and physically embedded computing yields new possibilities for increased sensing and data mining and personalized information. The speakers in this session focus on the opportunities and challenges of this quickly growing scale.

One dramatic change over the past decade is the number of people who carry Internet-connected sensing devices. Smart mobile devices make it possible to monitor health, capture rich media, and access vast information repositories at the touch of a button. Perhaps more than any other technology, these devices are ushering in new techniques for massive data science by correlating sensing and behavior—and also raising concerns about privacy in a "transparent" society.

Another important change is massive, socially connected online media. Social networks open up new avenues for communication and civic engagement. Online health platforms enable people to share health information. Online education has enrolled millions of students in just the past year and is causing many universities to rethink their long-term strategy.

All of us—as citizens, people, and educators—are affected by these changes. How should citizens, families, and universities think about massive online social interaction? How does it change the services we provide, the science we conduct, even our very conversations? This panel explored these issues.

Tony Jebara (Columbia University) led off the session with a presentation on modeling large-scale networks based on mobility data. Although most network growth models are based on incremental link analysis, he explored how to draw on users' data profiles alone—without any connectivity information—to infer their connectivity with others. For example, in a class of incoming college freshmen with no known friendship connections, is it possible to predict which pairs will become friends at the end of the year using only information such as their dorm or relationship status? Similarly, based only on the location history of a population of mobile phone users, can an observer predict which pairs of users are likely to communicate with each other?

Rob Miller (Massachusetts Institute of Technology) followed with remarks on the use of crowd computing to harness the power of people for tasks that are hard for individual users or computers to do alone.[1] He described prototype crowd-computing systems that he and his colleagues have built: a Word plugin that crowdsources text editing tasks, an app that helps blind people see using a crowd's eyes, and a system for code reviewing by a crowd of programmers. Crowd computing raises new challenges at the intersection of computer systems and human-computer interaction, to improve quality of work, minimize latency, and provide the right incentives to the crowd.

In the third talk, Kate Starbird (University of Washington) examined the crowdsourcing phenomenon during natural disasters and other crisis events. Armed with mobile devices and connected through social media platforms, people at the site of a disaster event are newly enabled to share information about unfolding events. This real-time information could be a vital resource for affected people and responders, but it remains difficult to transmit the right information to the right person at the right time. She described various ways that the crowd works to process data during disaster events and suggested future directions for leveraging "crowd work" to improve response efforts.

In the session's final presentation, Duncan Watts (Microsoft) reviewed exciting progress and challenges in the new field of "computational social science." He cited three obstacles to the widespread use of this resource at the convergence of the social sciences and the computer sciences. First, social scientific problems are almost always more difficult than they seem. Second, the data required to address many problems of interest to social scientists remain hard to assemble. And third, thorough exploration of complex social problems often requires the complementary application of diverse research traditions. He described some ideas for addressing these challenges.

---

[1] Dr. Miller's presentation is not included in this volume.

Modeling Large-Scale Networks

TONY JEBARA
Columbia University

Many real-world social networks involve topological connectivity information such as a set of *edges* between pairs of nodes that indicate a relationship between the individuals represented by the nodes. Furthermore, social networks also involve *attributes* associated with each node, such as a vector of demographic information describing the individual. Such networks are formed from the vast social data, mobile data, and location data being generated by large populations of users. Although most network growth models are based on incremental link analysis (Adamic and Adar 2003), it is also informative to consider how users' data profiles alone (after censoring the connectivity information) can be used to accurately predict the connectivity information. For example, in a class of incoming freshmen students with no known friendship connections, is it possible to predict which pairs will become friends at the end of the year using only their demographic profile information? Similarly, can colocation be used to predict communication—that is, using only the location history of a population of mobile phone users, can an observer predict which pairs of users are likely to communicate with each other?

To algorithmically reconstruct these networks, two methods—structure-preserving metric learning (SPML) (Shaw and Jebara 2009; Shaw et al. 2011) and degree-distributional metric learning (DDML) (Huang et al. 2011)—have been proposed. Both automatically recover a Mahalanobis distance metric from network and profile information. The simpler method, SPML, ensures that a connectivity algorithm (such as k-nearest neighbors or b-matching; Huang and Jebara 2007) yields the correct connectivity when applied on the distances computed using the learned Mahalanobis distance metric. The more sophisticated method, DDML, also estimates the degree of each node in addition to the Malahanobis distance metric.

The SPML approach begins with a known network and known attributes for the nodes, such as a friendship network of n students as they graduate from college. Figure 1 depicts such a network in the box on the left side. Nodes are adjacent to each other when the people they correspond to are friends (in other words, an edge links the two nodes). On the right side of Figure 1 is a representation of this connectivity information represented as an adjacency matrix ***A*** containing n rows and n columns of binary entries. For each individual in the network, it is also possible to observe static demographic attributes (e.g., age, height, weight, hometown, income bracket, favorite music, dorm room assignment), represented by a matrix ***X*** containing d rows and n columns of real entries as shown in the right side of Figure 1. Assume that ***A*** and ***X*** are training data to be used to determine the distance metric.

After training, the goal is to predict the adjacency matrix for a new set of incoming students on their first day of college by observing only their demographic attributes (***X'***). This prediction should closely match the true ***A'*** adjacency matrix that would be formed at the time of graduation for this new set of students. More specifically, ***A*** and ***X*** permit the estimation of an appropriate distance metric that can then be used to compute the distance between user *i*'s demographic vector $\boldsymbol{x}_i$ and user *j*'s demographic vector $\boldsymbol{x}_j$. This distance metric helps quantify how various attributes compete in the formation of friendship links, for example, how much does an age difference matter relative to a height difference when computing the affinity or distance between a pair of users? Given a good distance metric that balances all the multivariate demographic attributes, it eventually becomes

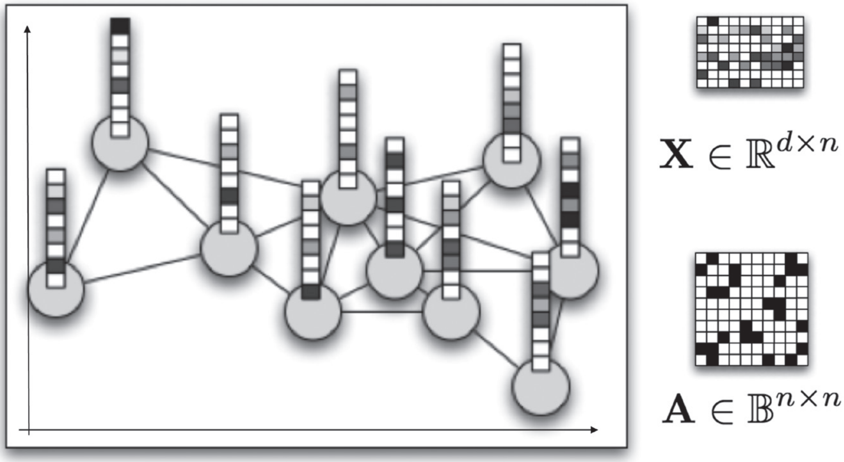


FIGURE 1 A social network with connectivity information (*A*) and attributes (*X*) to be used as training data for recovering a distance metric.

possible to reliably reconstruct ***A*** from ***X*** alone. To then test how well this method performs and generalizes, one uses the learned distance metric to reconstruct an unseen network ***A'*** from only new ***X'*** demographic data.

To evaluate the performance of such a framework, it is standard to use the area under the receiver operating characteristic curve. In a variety of synthetic and real-world experiments, SPML was capable of predicting links and edges from node attributes more accurately than standard techniques (Huang et al. 2011; Shaw et al. 2011). In particular, the approach outperforms simple naïve distance metrics (like Euclidean distance), relational topic models (Chang and Blei 2010), and support vector machine classifiers (Boser et al. 1992).

Furthermore, the SPML approach is efficient and quickly processes large network datasets. Computationally, it performs the optimization of a cost function using stochastic gradient descent. This conveniently eliminates the running-time dependency on the size of the network and makes it possible to scale to networks with hundreds of thousands of nodes and millions of edges. SPML has been used to reconstruct networks from Facebook data, Wikipedia data, FourSquare data, and mobile phone call detail records. Figure 2 shows the results of reconstructing Facebook data (Traud et al. 2011). Some interesting interpretable findings emerge. For instance, through Facebook social networks it appears that when students form friendships, those at Harvard are relatively more attentive to differences in relationship status, those at Stanford and Columbia are relatively more sensitive to differences in graduation year, and for those at MIT differences in dorm assignments are most important. Thus, social network structure helps tease apart demographic attributes and determine which are more or less relevant.

Like Facebook data, mobile phone data is also well suited for the SPML framework. Using a large dataset of location-augmented call detail records (CDRs) from a mobile phone carrier, it is possible to use phone calls and text messages between pairs of users to establish the existence of a friendship. Let the attributes of each user be their individual location history (the places they visited as measured by GPS or tower triangulation of the user's mobile device). One broad goal is to estimate distance metrics that reveal the colocations or meeting places that are most likely to correlate with (or predict) friendship (or communication) in the calling network. Are people more likely to be friends if they spend time together in high- or low-population-density regions? If they colocate in a coffee shop or at a subway station? The findings from this study are currently under review in a forthcoming article.

Finally, once a network is reconstructed (for instance using SPML), it is straightforward to do a variety of interesting things with it, such as visualize the network (Shaw and Jebara 2009) or predict missing attributes for some users (e.g., marital status, income) (Jebara et al. 2009 and Wang et al. 2013).

In conclusion, the graph topology of a social network helps define metrics of similarity and dissimilarity and reshapes the axes of demographic attributes. Novel algorithms such as SPML can be used to predict which networks could

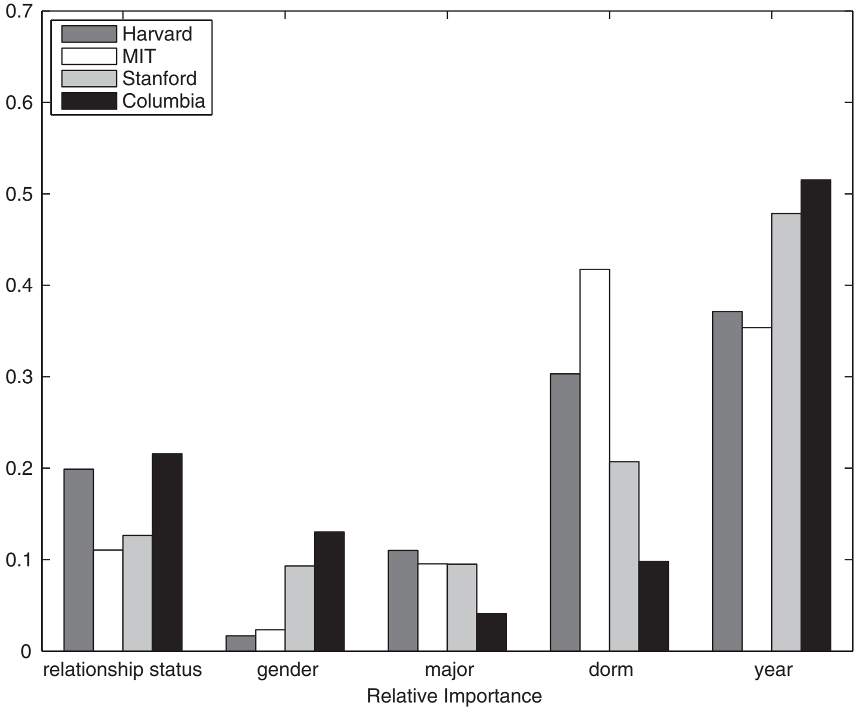


FIGURE 2 The relative importance of various demographic attributes (on the y-axis) for reconstructing friendship edges. Note that the relative importance of the demographic attributes varies across the four different universities in the study.

likely form between new users and elucidate which specific aspects of user profiles and demographics are predictive of communication and social interaction (Lazer et al. 2009; Newman 2003).

REFERENCES

Adamic LA, Adar E. 2003. Friends and neighbors on the Web. Social Networks 25:211–230.

Boser BE, Guyon IM, Vapnik VN. 1992. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory.

Chang J, Blei D. 2010. Hierarchical relational models for document networks. Annals of Applied Statistics 4:124–150.

Huang B, Jebara T. 2007. Loopy belief propagation for bipartite maximum weight b-matching. Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATs), San Juan, Puerto Rico, March.

Huang B, Shaw B, Jebara T. 2011. Learning a degree-augmented distance metric from a network. Proceedings of a Neural Information Processing Systems (NIPS) workshop, Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity, Sierra Nevada, December.

Jebara T, Wang J, Chang SF. 2009. Graph construction and b-matching for semi-supervised learning. Proceedings of the International Conference on Machine Learning (ICML), Montreal, June.

Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis D, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M. 2009. Computational social science. Science 323(5915):721–723.

Newman M. 2003. The structure and function of complex networks. SIAM Review 45:167–256.

Shaw B, Jebara T. 2009. Structure preserving embedding. Proceedings of the International Conference on Machine Learning (ICML), Montreal, June.

Shaw B, Huang B, Jebara T. 2011. Learning a distance metric from a network. Proceedings of a Neural Information Processing Systems (NIPS) workshop, Beyond Mahalanobis: Supervised Large-Scale Learning of Similarity, Sierra Nevada, December.

Traud A, Mucha P, Porter M. 2011. Social structure of Facebook networks. arXiv 1102.2166 [cs.SI].

Wang J, Jebara T, Chang SF. 2013. Semi-supervised learning using greedy max-cut. Journal of Machine Learning Research (JMLR) 14(Mar):771–800.

Crowds, Crisis, and Convergence: Crowdsourcing in the Context of Disasters

Kate Starbird
University of Washington

On October 29, 2012, Hurricane Sandy slammed into the US eastern seaboard, becoming one of the deadliest and costliest storms in US history. "Superstorm Sandy" caused 72 direct fatalities in the United States and tens of billions of dollars of damage, due in large part to a catastrophic storm surge that flooded hundreds of thousands of homes and businesses (Blake et al. 2013). The aftermath of the storm brought major disruptions to transportation systems, long-term power outages, and gas rationing in parts of New York and New Jersey.

THE SOCIAL MEDIA SURGE DURING HURRICANE SANDY

Like other disaster events in recent years, Sandy precipitated a huge surge in social media use. Twitter reported that it hosted more than 20 million tweets with search terms related to the event during a six-day window around the US impact. Instagram, a popular photo-sharing site, announced that users posted more than ten photos per second as Sandy came ashore.

Research suggested that a large portion of this content would have come from users outside affected areas and that much of it would have been "derivative"—i.e., reposted and remixed content (Starbird et al. 2010). But in this instance these platforms facilitated real-time information sharing that effectively informed response efforts. Residents of affected areas shared first-hand reports of actionable information—photos of flooded streets, videos of trees falling and houses catching fire, and tweets reporting stranded people. Emergency responders turned to social media to broadcast storm warnings and to quell rumors.

Problems with the propagation of misinformation drew widespread attention on social and mainstream media. One Twitter user reported, among other dubious claims, that the New York Stock Exchange had been flooded, and this misinformation spread rapidly before being called out by other Twitter users in what one writer called a "savage self-correction" (Herrman 2012).

Social media platforms and other online forums hosted self-organized community response efforts and other forms of volunteerism. The latter included the establishment of a Twitter hashtag to share information about open gas stations and a related project by a group of high school students who created and maintained a live "gas map"—an online map that displayed in real time where gas was available.

The role of social media during Sandy's lead-up, impact, and response generated considerable media attention. Some claimed the event marked a significant shift in the use of these services for emergency response, and at least one journalist suggested (in personal communication) that Sandy was the first "social" disaster. But social media were already becoming an established feature of disaster events—after the 2010 Haiti earthquake, for example, and the 2011 Japan tsunami. And disasters have always been inherently social, since well before the emergence of social media.

SOCIOLOGY OF DISASTER MEETS WEB 2.0: CHALLENGES AND OPPORTUNITIES

Sociologists of disaster have long known that people "converge" on the scene of disaster events (Fritz and Mathewson 1957; Kendra and Wachtendorf 2003). Fritz and Mathewson (1957) explained that, though this convergence is often physical, it can also be *informational* as people use available channels to seek and share information. Palen and colleagues (Hughes et al. 2008; Palen et al. 2010) connected this phenomenon to what now occurs online, whereby disaster events act as catalysts for massive "digital" convergence—of the kind that can generate 20 million tweets in six days.

This digital convergence carries considerable promise for improving disaster response. First-hand observations of events from citizen reporters on the ground can increase situational awareness both for other affected people and for responders. Social media can also be used for formal crisis communications, and emergency responders are increasingly turning to these platforms for outgoing messaging during and between disaster events.

Challenges

As the examples from Hurricane Sandy suggest, there remain several significant challenges in using social media as a real-time information source. The first is volume. Clearly, it is difficult for an individual to make sense of tens of tweets and photos per second. Similarly, if the focus is on finding actionable information coming from the site of the events, a vast quantity of social media data can be considered *noise*—some portion is completely off-topic, and another large percentage contains repeated, retweeted, or otherwise "derivative" information (Starbird et al. 2010).

Another particularly vexing issue is the problem of lost context, as information loses the connection to its original author, time, or place. For example, a tweet sent at 4:00 pm indicating a voluntary evacuation for a fire-affected neighborhood could become dangerous misinformation if reposted a few hours later, after the evacuation has become mandatory.

Misinformation and intentional disinformation are also major concerns. And the unstructured nature of social media content represents a challenge for those trying to make sense of it in aggregate form.

Automatically Filtering and Classifying the Flood of Data

Purely computational solutions for filtering and otherwise processing social media streams show promise, but have some limitations. Although terms of service and protocols continually change, accessing social media data is often the easiest part of the problem, because many social media platforms provide application programming interfaces for collecting public data.

Storing and searching these massive datasets presents a more complex challenge, one addressed in broader conversations about dealing with "big data." Moreover, because the textual content of social media streams is not quite the "natural language" for which traditional natural language processing techniques have been designed and tested, new approaches for computational content analysis are needed. Additionally, accuracy is extremely important in time- and safety-critical environments like those of a disaster, and currently even the best automatic classification techniques along relatively simpler data dimensions (e.g., identifying situational awareness information) achieve only about 80 percent accuracy (Verma et al. 2011).

Harnessing the Power of the Crowd

Another solution for filtering the flood of data during disasters involves *human computation* or *crowdsourcing*, using a large number of people, connected via the Internet, to manually process the data. In considering the use of these techniques, researchers are very much following the crowd. During recent disaster events, people have appropriated social media platforms and other available online tools such as Skype and shared Google Documents to improvise response efforts, often in the form of informational assistance (e.g., the New Jersey students and their gas map).

The new *digital volunteer* behavior aligns with another long-recognized disaster phenomenon, *spontaneous volunteerism*, whereby people make themselves available to help in various capacities, often by improvising to fill gaps in formal response efforts (Kendra and Wachtendorf 2003). During the 2009 Red

River floods in North Dakota and Minnesota, for example, volunteer programmers created algorithms that automatically tweeted river heights at various locations (Starbird et al. 2010). After the Haiti earthquake, a group of self-named "voluntweeters" used the Twitter platform to help coordinate aid efforts, eventually connecting with each other to form a new organization (Starbird and Palen 2011). In another highly publicized effort during that event, students at Tufts University created and maintained a public map of humanitarian needs, translating and geolocating thousands of reports arriving from Haitian people via an SMS short code (Meier and Munro 2010). During the impact of Hurricane Irene in the Catskills in September 2011, a group of journalists served as "crowdsourcerors," organizing a community information-sharing and response effort through a combination of a Liveblog, Facebook, Twitter, and even radio broadcasts and phone calls from landline phones in more remote areas (Dailey and Starbird, forthcoming).

Although each event spawns new crowd-powered solutions to newly recognized needs, a number of ongoing virtual volunteer organizations have been established (e.g., the Standby Task Force, Humanity Road, Crisis Commons, and several Virtual Operation Support Teams connected to emergency responders). These groups use available online tools to respond to disaster events all over the world. However, questions remain about how they will sustain committed participation and how they can connect both the products of their work and this new information-processing capacity more broadly to the established work practices of formal responders.

One research opportunity lies in understanding the work of digital volunteers and designing tools and platforms to support their efforts—for example, by developing crowdsourcing solutions that align with the motivations of disaster volunteers, initial altruism that soon becomes augmented by social and reputation "capital."

Using the Noise to Find the Signal

The collective behavior of the crowd can be leveraged to address information-processing challenges. Social media users, intentionally and not, shape the information space through their behavior within it. Instead of viewing crowd activity as simply noise, it is possible to consider every repost, "like," "follow," and user mention as productive crowd work and to use this "noise" to find the signal. For example, algorithms can be designed to identify misinformation through features of crowd behavior—i.e., sensing the "savage self-correction" of dozens of voices publicly questioning false information. Alternatively, the crowd itself could serve as a "sensor" for other (e.g., actionable) kinds of information. It has been demonstrated that retweet and follow patterns on Twitter can be used to home in on users tweeting from the site of an event, but there is still work to be done in designing solutions that function in real time, and questions remain about how best to communicate these solutions to decision makers during an event.

Integrating Machine and Human Computation

The most powerful solutions in the social media space may depend on the integration of machine- and human-powered approaches. These would involve machine learning algorithms that learn from volunteers' and other crowd members' online actions and then feed processed data back to volunteers who verify and synthesize the output before forwarding it to responders and affected citizens. Along these lines, it will be important to design solutions that both align with the values and motivations of digital volunteers and fit into formal emergency response processes.

CONCLUSION: THE NEED FOR HUMAN-CENTERED DESIGN IN THE CONTEXT OF DISASTER EVENTS

Massive online convergence is now an established feature of crisis events and carries with it great potential for improving outcomes during response efforts—*if* the right information can be transmitted to the right people at the right time and in the right form. The challenges at this intersection of crowds and crises are both technical and social. Solutions will likely benefit from a human-centered approach to understand and support the informational needs and goals of the people affected, responders, volunteers, and the broader public during disaster events. The most effective solutions will probably integrate the social media–based work of the crowd with computational algorithms that can scale up with the ever increasing size and complexity of the information-processing needs.

REFERENCES

Blake ES, Kimberlain TB, Berg RJ, Cangialosi JP, Beven JL II. 2013. Tropical Cyclone Report, National Hurricane Center (AL182012). Silver Spring MD: US National Oceanic and Atmospheric Administration's National Weather Service. Available at www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf.

Dailey D, Starbird K. Forthcoming. Journalists as crowdsourcerers: Responding to crisis by reporting with a crowd.

Fritz CE, Mathewson JH. 1957. Convergence Behavior in Disasters: A Problem in Social Control. Washington DC: National Academy of Sciences.

Herrman J. 2012. Twitter is a truth machine. Blog on Buzzfeed, October 30. Available at http://gofwd.tumblr.com/post/34623466723/twitter-is-a-truth-machine.

Hughes A, Palen L, Sutton J, Liu S, Vieweg S. 2008. "Site-seeing" in disaster: An examination of on-line social convergence. Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM), Washington DC, August.

Kendra JM, Wachtendorf T. 2003. Reconsidering convergence and converger: Legitimacy in response to the World Trade Center disaster. Terrorism and Disaster—New Threats, New Ideas: Research in Social Problems and Public Policy, vol 11, ed. Bingley CL. UK: Emerald Group Publishing. pp 97–122.

Meier P, Munro R. 2010. The unprecedented role of SMS in disaster response: Learning from Haiti. SAIS Review 30(2):91–103.

Palen L, Anderson KM, Mark G, Martin J, Sicker D, Palmer M, Grunwald D. 2010. A vision for technology-mediated support for public participation and assistance in mass emergencies and disasters. Proceedings of the 2010 ACM-BCS Visions of Computer Science Conference, Edinburgh, April 14–16. Swinton UK: ACM-BCS Visions of Computer Science, British Computer Society. pp 1–12.

Starbird K, Palen L. 2011. "Voluntweeters": Self-organizing by digital volunteers in times of crisis. Proceedings of the 2011 ACM Conference on Human Factors in Computing Systems, Vancouver, May. New York: ACM. pp 1071–1080.

Starbird K, Palen L, Hughes A, Vieweg S. 2010. Chatter on the Red: What hazards threat reveals about the social life of microblogged information. Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, Savannah, February. New York: ACM. pp 241–250.

Verma S, Vieweg S, Corvey W, Palen L, Martin J, Palmer M, Schram A, Anderson K. 2011. Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, July. pp 17–21.

Computational Social Science: Exciting Progress and Future Directions

Duncan J. Watts
Microsoft Research

The past 15 years have witnessed a remarkable increase in both the scale and scope of social and behavioral data available to researchers. Over the same period, and driven by the same explosion in data, the study of social phenomena has increasingly become the province of computer scientists, physicists, and other "hard" scientists. Papers on social networks and related topics appear routinely in top science journals and computer science conferences; network science research centers and institutes are sprouting up at top universities; and funding agencies from DARPA to NSF have moved quickly to embrace what is being called *computational social science.*

Against these exciting developments stands a stubborn fact: in spite of many thousands of published papers, there has been surprisingly little progress on the "big" questions that motivated the field of computational social science—questions concerning systemic risk in financial systems, problem solving in complex organizations, and the dynamics of epidemics or social movements, among others.

Of the many reasons for this state of affairs, I concentrate here on three. First, social science problems are almost always more difficult than they seem. Second, the data required to address many problems of interest to social scientists remain difficult to assemble. And third, thorough exploration of complex social problems often requires the complementary application of multiple research traditions—statistical modeling and simulation, social and economic theory, lab experiments, surveys, ethnographic fieldwork, historical or archival research, and practical experience—many of which will be unfamiliar to any one researcher. In addition to explaining the particulars of these challenges, I sketch out some ideas for addressing them.

WHY IS SOCIAL SCIENCE HARD?

By definition, "social" phenomena are less about the behavior of individuals than of collections of individuals in groups, crowds, organizations, markets, classes, and even entire societies, all of which interact with each other via networks of information and influence, which in turn change over time. As a result, social systems—like complex systems in physics and biology—exhibit "emergent" behavior, meaning that the behavior of entities at one "scale" of reality is not easily traced to the properties of the entities at the scale below (Anderson 1972). Firms, for example, can exhibit highly stable identities and cultures even as the particular employees who work in them change completely over time, just as you remain you even as the cells in your body turn over during the course of your lifetime. Conversely, the stock market, the economy, or a political regime can collapse suddenly and unexpectedly even as the various players and background conditions remain the same.

Complicating matters further, emergent properties can be both the cause and the effect of social change. Sometimes, that is, the decisions of corporations or even governments may depend critically on the personal interests of a handful of executives, whereas at other times the behavior of those same individuals may be powerfully constrained by the corporate or political culture to which they belong.

Nor is emergence as simple as one scale of reality aggregating to another. Rather, in many problems of interest to social scientists, the actions of individuals, firms, regulatory and government agencies, markets, and political institutions may all play important roles. Moreover, because these different types of actors not only exist at different scales (firms comprise individuals, markets comprise firms and individuals, etc.) but also may interact with each other in important ways, problems of this type require one to consider events, actors, and forces across multiple scales simultaneously.

Given the unavoidably multiscale, complex, and emergent nature of social phenomena, it is not surprising that theories of social behavior and change have been difficult to work out in any realistic detail. Compounding this theoretical difficulty are two separate but related empirical difficulties. First, it has been impossible to collect observational data on the scale of hundreds of millions, or even tens of thousands, of individuals. Second, because cause and effect can be difficult to infer from observational data alone, experimental studies are also necessary. Yet experiments involving, say, the performance of an organization with a particular structure, or the popularity of songs in a single instance of a cultural market, represent the collective behavior of hundreds or even thousands of individuals, designs that are clearly impossible to implement in a physical lab (Zelditch 1969).

THE EMERGENCE OF COMPUTATIONAL SOCIAL SCIENCE

In light of these three interrelated difficulties—(1) the complexity of the theoretical issues confronting social science, (2) the difficulty of obtaining the relevant observational data, and (3) the difficulty of manipulating large-scale social organizations experimentally—it is hardly surprising that progress in social science has been slow relative to that in the physical, engineering, and biological sciences, in particular over the past century. But the computing revolution of the past two decades—a revolution that has dramatically increased not only the speed and memory of computers themselves but also the scale and scope of social data that can now be analyzed—has the potential to revolutionize traditional social science, leading arguably to a new paradigm of "computational social science"[1] (Lazer et al. 2009).

The most prominent strand of research in computational social science leverages communication technologies—including email, social networking and microblogging services, and cellphones—as well as online games, ecommerce sites, and other Internet-enabled services. All of these devices and services generate digital signals, often referred to as *digital exhaust* or *digital breadcrumbs*, from which inferences can be made about individual and/or collective behavior. In this way, it is increasingly possible to observe the actions and interactions of hundreds of millions of individuals in real time as well as over time.

Data derived from instant messaging services and social networking sites, for example, have been used to construct networks of hundreds of millions of nodes, analysis of which (Leskovec and Horvitz 2008; Ugander et al. 2011) has confirmed earlier conjectures about the topology of large social networks (Newman 2003; Watts and Strogatz 1998). Other studies have mined email data to estimate the microlevel rules describing new tie formation (Kossinets and Watts 2006) or used blog networks to measure the propensity to join new groups (Backstrom et al. 2006). Others still have mapped the diffusion of online content (Bakshy et al. 2009; Dow et al. 2013; Goel et al. 2012; Leskovec et al. 2007; Sun et al. 2009) or conducted massive randomized field experiments to estimate the causal effects of social influence on adoption (Aral and Walker 2011), voter turnout (Bond et al. 2012), or likelihood to share content (Bakshy et al. 2012).

A less well explored but also important strand of research uses the web to create "virtual labs": controlled environments for the conduct of "macro-sociological" experiments (Hedstrom 2006). Although early efforts relied on volunteers (Dodds et al. 2003; Salganik et al. 2006), an important recent development in this field has been the use of crowdsourcing sites such as Amazon's Mechanical Turk to recruit and pay subjects, analogous to the longstanding tradition in behavioral science of recruiting from college student populations (Mason and Watts 2009).

---

[1] "Computational social science" is a contested label, referring in some quarters to simulation of agent-based models (see, for example, http://computationalsocialscience.org/) and in others strictly to the analysis of computationally challenging datasets (http://research.microsoft.com/en-us/groups/cssnyc/). Here I use the term somewhat liberally to refer to the emerging intersection of the social and computational sciences, an intersection that includes analysis of web-scale observational data, virtual lab–style experiments, and computational modeling.

One important advance due to crowdsourced virtual labs has been resolution of the synchronicity problem to ensure that N subjects will arrive contemporaneously and remain engaged in the experiment for its duration (Suri and Watts 2011), thereby allowing for networked experimental designs. Another advantage is that experiments can be designed, launched, and executed on a much shorter timescale than has been historically feasible, and on a lower cost basis (Wang et al. 2012). Finally, by shrinking the hypothesis-testing cycle—the lag between analyzing one set of experimental results and running the next set of experiments—from months or years to days or even hours, crowdsourced virtual lab experiments can dramatically expand the range of conditions that can be studied.

CHALLENGES AND OPPORTUNITIES FOR COMPUTATIONAL SOCIAL SCIENCE

As impressive as its recent accomplishments have been, computational social science faces a number of pressing challenges if it is to address the important questions of social science in a meaningful way. For example, organizational and interorganizational problem solving, collective action and decision making, the relationship between deliberation, governance, and democracy, the emergence of disruptive technologies, and the rise of new political or cultural movements are all core social scientific questions, but they have received little attention from computational social science largely on the grounds of limits to current data sources, platforms, or methods. In the following sections I describe three challenges and suggest some directions for future progress.

Creating a Social Supercollider

First, the dominant digital exhaust model of data collection imposes important limitations on the type of research questions that can be answered.

Consider, for example, the problem of measuring how friends influence each other's purchase behavior, a question that is of great interest to social scientists as well as to marketers, policymakers, and other change agents. Answering such a question requires the ability to observe both the complete friendship network (already a difficult task) and the shopping behavior of everyone in the network. Using existing systems, one might obtain an approximation of the friendship network by using, say, Facebook data or mining email logs, while ecommerce sites or retailer databases may show how much individuals are spending on particular products. Currently, however, it is extremely difficult to combine even two such sources of data, and of course there are many different modes of communication and many different places to make purchases.

Generalizing beyond this one example, many questions of interest to social scientists require studying the relation between different modes of social action and interaction—for example, search data to infer intent, network data to infer relationships, ecommerce data to infer choices, and social media data to infer opinions—but these modes are generally recorded and stored separately, often by different companies. A major breakthrough for computational social science, therefore, would be a "social supercollider": a facility that combines multiple streams of data, creating richer and more realistic portraits of individual behavior and identity, while retaining the benefits of massive scale.

Against this considerable promise stands the equally pressing concern of protecting individual privacy. Privacy is already an important issue for all industries that collect digital information about their consumers; however, for the same reason that the social supercollider would be so powerful a scientific tool—namely that it would put all the pieces together—it raises far more serious questions about individual privacy even than are posed by existing commercial platforms. These questions have in fact already been raised by recent revelations of the National Security Agency's Prism project, which also appears to be an attempt to combine data from multiple sources. Construction and management of anything like a social supercollider would therefore have to proceed under the strictest scrutiny, with respect to both governance and the end uses of the data.

Expanding Virtual Labs

A second challenge for computational social science concerns the continued development of experimental macrosociology. Perhaps surprisingly, the major limitation to existing experimental designs is not technical but rather logistical—namely, the difficulty of recruiting large numbers of subjects in a reliable and cost-effective manner. For example, the largest synchronous virtual lab experiments to date have not exceeded N=36, largely because of the practical difficulty of recruiting more than that number at any single time.

One potential solution to this problem would be to construct a large, persistent, and well-documented panel of subjects—potentially hundreds of thousands of individuals—who might participate in many experiments over months or years. Increasing the scale of experiments from dozens to thousands of simultaneous participants would fundamentally alter the types of experiments that could be run—making it possible to, for example, study the proverbial army in a lab (Zelditch 1969). By allowing researchers to specify their sampling frame in advance, another advantage of a large persistent subject set would be to facilitate investigations of variation in behavior by demographic, national, or racial group.

Such a panel would also enable the study of entirely novel questions about the connections between individual attributes and behavior as well as between different elements of behavior itself. Do people who contribute generously to public goods games behave in any characteristic way when participating in a collaborative problem-solving exercise or in an exchange network?

Finally, beyond virtual lab experiments, a panel of this scale and duration could be of great value for survey research and randomized field experiments.

Putting the "Social" in Computational Social Science

A final challenge for computational social science is that, in spite of many thousands of papers published on topics related to social networks, financial crises, crowdsourcing, influence and adoption, group formation, and so on, relatively few are published in traditional social science journals or even attempt to engage seriously with social scientific literature. The result is that much of computational social science has effectively evolved in isolation from the rest of social science, largely ignoring much of what social scientists have to say about the same topics, and largely being ignored by them in return.

It is unclear who is to blame for this state of affairs—computer scientists for being presumptuous, social scientists for being defensive, or both—and even whether it is a bad thing. Perhaps all interdisciplinary fields start out as ugly ducklings and have to become swans on their own, not by making friends with existing fields but by outcompeting them. My view, however, is that meaningful progress on important problems will require serious engagement between the communities, each of which has much to offer the other: computer scientists have technical capabilities that are of great potential benefit to social scientists, and the latter's deep subject matter knowledge is essential in order to ask the right questions and to formulate even simple models in ways that address these questions.

NEW INSTITUTIONS FOR COMPUTATIONAL SOCIAL SCIENCE

Unfortunately, harnessing the complementary strengths of multiple research communities is easier said than done. Consider, for example, the problem of managing systemic risk in financial systems. On the one hand, simple and elegant models of financial crises that are inspired by the analogy of contagion in networks (Delli Gatti et al. 2012; Gai and Kapadia 2010; May and Arinaminpathy 2010; Nier et al. 2007) turn out to omit certain features of real banking systems—for example, that banks "create" money by expanding their balance sheets or that prices must adjust so that markets will clear—that are critical to understanding recent crises. On the other hand, descriptively accurate accounts of real financial crises tend to be so complex and multifaceted (Brunnermeier 2009; Financial Crisis Inquiry Commission 2011; Gorton 2012; Hellwig 2009) that it is difficult even for experts to agree on which mechanisms are the most important and therefore what features are critical to include in even a simple model.

The existence of diverse and even incommensurate literatures on the same topic is a surprisingly common problem in social science, and resolving it requires substantial investment in time as well as considerable diversity of expertise. As no one individual is likely to satisfy this requirement, interdisciplinary teams of researchers who have both the resources and the incentives to engage in long-term, high-risk collaborations seem increasingly necessary. Such collaborations are also challenging, however, in light of the cultural and language differences that separate disciplines like computer science from the social sciences, as well the wide variations in publication norms and timescales. Finally, to be successful computational and experimental research designs must be coordinated with methods drawn from the theoretical, survey, and ethnographic traditions of social science.

Deep and significant progress in social science, in other words, will require not only new data and methods but also new institutions that are designed from the ground up to foster long-term, large-scale, multidisciplinary, multimethod, problem-oriented social science research. To succeed, such an institution will require substantial investment, on a par with existing institutes for mind, brain, and behavior, genomics, or cancer, as well as the active cooperation of industry and government partners.

The current and justifiable excitement surrounding computational social science presents an opportune moment to engage in such an undertaking.

CONCLUSION

Driven by new sources of data, ever-increasing computing power, and the interest of computer scientists, social science is becoming a computational discipline much as biology did in the late 1990s. As exciting and important as this development is, however, social science is not and should not become a subfield of computer science or "data science." Just as in computational biology, the computational element of computational social science should remain in service to the substantive and substantial questions of social science.

Achieving this goal will require significant investments in new sources of data, new platforms for organizing existing data, and new institutional arrangements for fostering team-based interdisciplinary research. Although somewhat novel for social science, which has long operated on the model of the single-authored book or paper, the research lab model is familiar from the biological and medical sciences, and with the appropriate commitment could revolutionize social science in the 21st century.

REFERENCES

Anderson PW. 1972. More is different. Science 177:393–396.

Aral S, Walker D. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. Management Science 57(9):1623–1639.

Backstrom L, Huttenlocher D, Kleinberg J, Lan X. 2006. Group formation in large social networks: Membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia. pp. 44–54.

Bakshy E, Karrer B, Adamic LA. 2009. Social influence and the diffusion of user-created content. In: Proceedings of the 10th ACM Conference on Electronic Commerce, Association of Computing Machinery, Stanford, California.

Bakshy E, Rosenn I, Marlow C, Adamic L. 2012. The role of social networks in information diffusion. In Proceedings of the 21st International ACM Conference on World Wide Web, Lyon, France. pp. 519–528.

Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH. 2012. A 61-million-person experiment in social influence and political mobilization. Nature 489:295–298.

Brunnermeier MK. 2009. Deciphering the liquidity and credit crunch 2007-2008. Journal of Economic Perspectives 23:77–100.

Delli Gatti D, Gallegati M, Greenwald B, Stiglitz J, Battiston S. 2012. Liaisons dangereuses: Increasing connectivity, risk sharing and systemic risk. Journal of Economic Dynamics and Control 36:1121–1141.

Dodds PS, Muhamad R, Watts DJ. 2003. An experimental study of search in global social networks. Science 301:827–829.

Dow PA, Adamic LA, Friggeri A. 2013. The anatomy of large Facebook cascades. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, Cambridge, MA.

Financial Crisis Inquiry Commission. 2011. Financial Crisis Inquiry Report: Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States. Washington DC: Government Printing Office.

Gai P, Kapadia S. 2010. Contagion in financial networks. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 466:2401–2423.

Goel S, Watts DJ, Goldstein DG. 2012. The structure of online diffusion networks. In Proceedings of the 13th ACM Conference on Electronic Commerce, Valencia, Spain. pp. 623–638.

Gorton GB. 2012. Misunderstanding Financial Crises: Why We Don't See Them Coming. New York: Oxford University Press.

Hedstrom P. 2006. Experimental macro sociology: Predicting the next best seller. Science 311(5762):786–787.

Hellwig MF. 2009. Systemic risk in the financial sector: An analysis of the subprime-mortgage financial crisis. De Economist 157:129–207.

Kossinets G, Watts DJ. 2006. Empirical analysis of an evolving social network. Science 311:88–90.

Lazer D, Pentland A, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M. 2009. Computational social science. Science 323(5915):721–723.

Leskovec J, Adamic LA, Huberman BA. 2007. The dynamics of viral marketing. ACM Transactions on the Web 1(1):1–46.

Leskovec J, Horvitz E. 2008. Planetary-scale views on a large instant-messaging network. In Proceedings of the 17th International World Wide Web Conference, Beijing, China.

Mason W, Watts DJ. 2009. Financial incentives and the performance of crowds. Proceedings of the ACM SIGKDD Workshop on Human Computation, Paris. pp. 77–85.

May RM, Arinaminpathy N. 2010. Systemic risk: The dynamics of model banking systems. Journal of the Royal Society Interface 7:823–838.

Newman MEJ. 2003. The structure and function of complex networks. SIAM Review 45:167–256.

Nier E, Yang J, Yorulmazer T, Alentorn A. 2007. Network models and financial stability. Journal of Economic Dynamics and Control 31:2033–2060.

Salganik MJ, Dodds PS, Watts DJ. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. Science 311:854–856.

Sun ES, Rosenn I, Marlow CA, Lento TM. 2009. Gesundheit! Modeling contagion through Facebook news feed. In International Conference on Weblogs and Social Media, AAAI, San Jose.

Suri S, Watts DJ. 2011. Cooperation and contagion in web-based, networked public goods experiments. PLoS One 6(3):e16836.

Ugander J, Karrer B, Backstrom L, Marlow C. 2011. The anatomy of the Facebook social graph. arXiv 1111.4503.

Wang J, Suri S, Watts DJ. 2012. Cooperation and assortativity with dynamic partner updating supporting information. Proceedings of the National Academy of Sciences U S A 109(36):14363–14368.

Watts DJ, Strogatz SH. 1998. Collective dynamics of "small-world" networks. Nature 393:440–442.

Zelditch M. 1969. Can you really study an army in the laboratory? In: Etzioni A, ed. A Sociological Reader on Complex Organizations. Boston: Holt, Rinehart and Winston. pp. 528–539.

COGNITIVE MANUFACTURING

Cognitive Manufacturing

ELIZABETH HOEGEMAN
Cummins Inc.

J. RHETT MAYOR
Georgia Institute of Technology

Consider the degree of computer-enabled technology penetration in everyday life, with self-parking cars and smartphones that present locale-specific information through augmented reality displays. Given this increased use of computer-enabled decision making, is it plausible to consider the near-term realization of science fiction notions of autonomous production systems with "machines making machines"?

Manufacturing as an industry has been pervasively impacted by the rapid adoption of information technology (IT). Modern manufacturing systems execute highly sophisticated IT-enabled operations and control infrastructure that track production metrics, quality metrics, and component status in real time. The state of practice in the field exhibits the characteristics associated with "smart" systems, as distributed processors feature embedded low-level logic systems that trigger alerts in response to single value break points, or level-based go/no-go indicators, and report these alerts to supervisory human operators through IT-enabled communication channels. Decisions about how to respond to such alerts are made by human operators based on their knowledge of the process and reasoned judgment. That is, the cognitive process is performed by human intelligence and remains the primary function of the operator.

Cognitive manufacturing is an evolutionary step in computer-enabled production system control that pushes beyond smart technologies, in which the intelligence and reasoning are retained by the human user, and endows the manufacturing system with capabilities of perception and judgment to enable the autonomous operation of the system based on embedded cognitive reasoning, reliant only on high-level supervisory control.

Cognitive manufacturing systems perceive changes in the production process and “know” how to respond to these dynamic fluctuations by adapting the production to stay within target ranges of production cost and rate, and, as are increasingly important, sustainability indices such as energy intensity and carbon footprint. The embedded cognitive capability can be accomplished through the development of cognitive reasoning engines, or distributed intelligence agents, deployed throughout the production system at three hierarchical levels: (1) the manufacturing process level, (2) the manufacturing system or factory level, and (3) the supply chain or production system logistical level.

The speakers in this session introduce and explore cognitive manufacturing as an emerging frontier of engineering science that integrates domain knowledge from industrial and systems engineering, manufacturing process science, computer learning, information technology, adaptive control theory, biologically inspired system design, and environmentally cognizant design and sustainability. The presentations cover the deployment of computer-enabled cognitive reasoning at the three levels of production systems and the application of computer-enabled cognitive manufacturing systems to achieve sustainable production systems and mass sustainability.

The first speaker, Dragan Djurdjanovic (University of Texas at Austin), discussed the development of distributed anomaly detection agents to recognize and address unprecedented faults (i.e., those that the system could not have been programmed to recognize). He illustrated the application of such an approach to dramatically reduce downtimes—and significant costs—associated with fault remediation. In the second presentation Chris Will (Apriso/Dassault Systèmes) traced the emergence of business process management technologies to accelerate process improvement, standardization, and excellence programs by translating process modeling results into an executable form that limits or eliminates the need to code or customize a core application. Next, Steve Ellet (Chainalytics) demonstrated that, by using sophisticated new modeling techniques and tools, companies are making better, faster, fact-based decisions that require fewer resources to make and move their products to market. The session’s final speaker, Steven Skerlos (University of Michigan), explained how cognitive systems can advance the state of the art in sustainable manufacturing, stressing the importance of integrating sustainability objectives into the product design and describing the application of various life cycle assessment methods to clarify the link between manufacturing systems and their environmental and social consequences.

Distributed Anomaly Detection for Timely Fault Remediation in Modern Manufacturing

DRAGAN DJURDJANOVIC
University of Texas at Austin

Decades of academic and industry-based research on traditional condition monitoring led to precedent-based approaches, which recognize faulty behavior modes whose indications have already been established based on engineering knowledge or training data. But for the highly sophisticated systems used in manufacturing today, the precedent-based approach is cumbersome, time consuming, and costly. First, the condition monitoring procedures have to be trained to recognize a large number of potential faults, some of which cannot be anticipated during the design stage. Second, because modern manufacturing equipment can perform a variety of operations, it displays highly dynamic behavior. Finally, faults may manifest very differently under different operating conditions, so training of diagnostic units for all possible conditions and all possible faults is practically impossible. Research is needed to better understand the applications of such tools and support their implementation to promote more efficient and cost-effective manufacturing.

COMPLICATIONS AND COSTS OF UNPRECEDENTED FAULTS

One does not have to look far for evidence of manufacturers scrambling to deal with unprecedented faulty situations. For example, connections in communication networks among stations in automotive assembly lines are commonly interrupted because of contacts compromised by moving robots and workpieces, coolants, and improper installation (Lei et al. 2010, 2011). They are inherently unprecedented faults because, with different network configurations, usage patterns, and fault severities, every intermittent connection in every plant is different, which means that there are no fault signatures available a priori. In one case an intermittent connection on a DeviceNet network in a major automotive assembly plant resulted in a 4½-hour downtime because of an inability to diagnose the problem and find the node that caused it. One minute of downtime on an automotive assembly line can cost more than $20,000 (Spiewak et al. 2000).

With more sophisticated manufacturing processes and systems, consequences of the inability to cope with unprecedented situations become more frequent and have greater impacts. For example, analysis of data for more than 6 months of operation of a plasma-enhanced chemical vapor deposition (PECVD) in a major domestic 300 mm semiconductor fabrication plant found two downtime instances that each lasted more than a week and each cost close to $1 million just in scrapped wafers (lost production damage is probably even larger) (Cholette 2012). The tool kept messing the wafers while emitting signatures that were totally new, and no one could fix the problem. One of those downtimes was finally resolved after a teleconference with a quantum physics PhD, who led the control design and development of the PECVD tool in question. (I have seen the same type of tool cause similar problems in two other semiconductor manufacturing companies, again because it took a long time to find the root causes and do appropriate repairs.)

In the examples described above, the manufacturers were stymied by unknown situations, searching blindly for the sources of the problems, devising ad hoc repair procedures, and wasting significant resources. Such "learning by doing" is inexcusably frequent, especially in highly sophisticated, high-value manufacturing areas, and leads to expensive equipment downtimes and ineffective repairs, costing US manufacturers hundreds of billions of dollars annually (Heng et al. 2009).

It seems evident that the traditional precedent-based diagnostic paradigm has reached its limits and a radically new approach is needed to deal with the ever increasing complexity of modern manufacturing. The planning and scheduling of manufacturing operations in an environment plagued by unknown, unprecedented situations represent an exciting research opportunity with potentially enormous positive impact.

COPING WITH UNPRECEDENTED CONDITIONS

The traditional precedent-based diagnostic approach involves a database of fault models based on elaborate training data and/or modeling efforts describing system behavior in the presence of those faults. Each time an abnormality is detected, the system behavior is checked against the fault models to identify the cause and determine the appropriate maintenance actions. Faults whose models do not exist in the database required lengthy efforts to identify the cause, relying heavily on human expertise and costly trial and error.

Distributed Anomaly Detection

Maintenance personnel in a manufacturing facility need to be able to identify, or localize, the source of a fault and the corresponding field replaceable unit (FRU); the specific character of the fault is usually secondary. For example, a maintenance worker on a PECVD tool needs to know whether a pendulum valve on the tool is anomalous or not. Information about whether the anomaly was caused by a faulty actuator or accumulation buildup or any other reason is secondary; the remedial action is the same: replace the pendulum valve. Hence, instead of identifying various faulty behavior regimes, the more useful focus would be on localizing the source of the anomalous behavior, using the paradigm of *distributed anomaly detection*. Essentially, monitoring can be realized through a set of anomaly detectors (ADs) covering the target system, with each detector that perceives an anomaly splitting into a set of ADs monitoring the pertinent subsystems. This cascading "proliferation" of ADs continues until the FRUs that caused the abnormality are identified (Cholette and Djurdjanovic 2012a,b; Djurdjanovic et al. 2010). Thus, rather than a database of specific faulty behavioral modes, this approach requires a database of *only* normal behavior models of the target system and its subsystems. ADs use these models to detect anomalies as statistically significant departures.

Figure 1 shows how this approach enables the identification of subsystems causing anomalies in the exhaust gas recirculation (EGR) system of an automotive diesel engine (from Cholette and Djurdjanovic 2012b). Initially, an AD monitors the entire EGR system based on the dynamic model of its normal behavior, as indicated in plot (a). Once it detects an anomaly, five ADs are distributed, as shown in plot (b), each using the relevant model of normal behavior of the target subsystem to monitor its behavior.

The "normalness" of behavior of each system is assessed through confidence values (CVs), which express the overlap of modeling residuals observed during normal behavior and those most recently observed. This quantity fluctuates between 0 and 1, with 1 indicating a perfect match (performance identical to normal) and small values indicating anomalies. Using the terminology of the literature on artificial immune systems, CVs are analogous to each detector's "affinity" to the normal behavior of its respective target system(s) (Forest et al. 1994).

Figure 2 shows CVs from the relevant ADs when progressively more severe clogging was simulated in the EGR valve. Interpretation of the CVs clearly points to the culprit subsystem (a valve), even though no fault signatures were collected, only models of normal behavior of the EGR system and its subsystems. The same approach was used to isolate faults in the controller and EGR cooler (Cholette and Djurdjanovic 2012b).

But precedent-free fault localization is not yet used in manufacturing, primarily because of the immense complexity of modern manufacturing machines in which such an approach is really needed. One problem concerns the lack of observability of the underlying phenomena in such machines; for example, the state of plasma in a PECVD or etching tool is inherently unobservable, as will be elaborated below. Another challenge lies in the fact that distributed anomaly detection requires an understanding of interactions among subsystems—which

Figure 1 Plot a

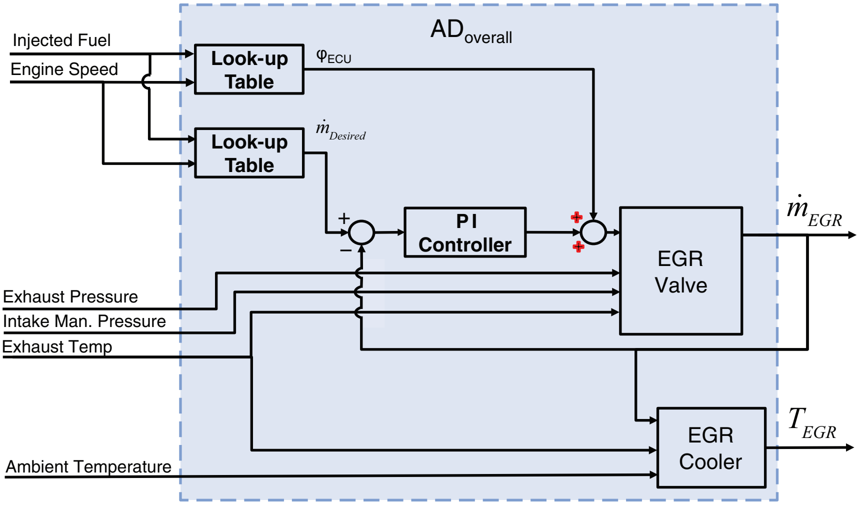


Figure 1 Plot b

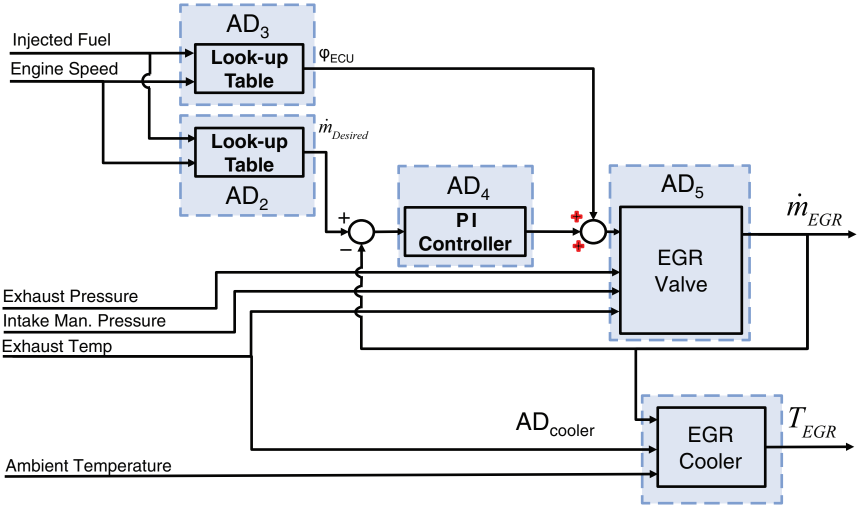


FIGURE 1 Distribution of anomaly detectors (ADs) in a diesel engine exhaust gas recirculation (EGR) system. Once an anomaly is detected, the overall AD, depicted in plot (a), splits into ADs that monitor pertinent subsystems, as illustrated in plot (b). ECU = electronic control unit; Man. = manifold; PI = proportional-integral; T = temperature; $\dot{m}$ = mass flow.

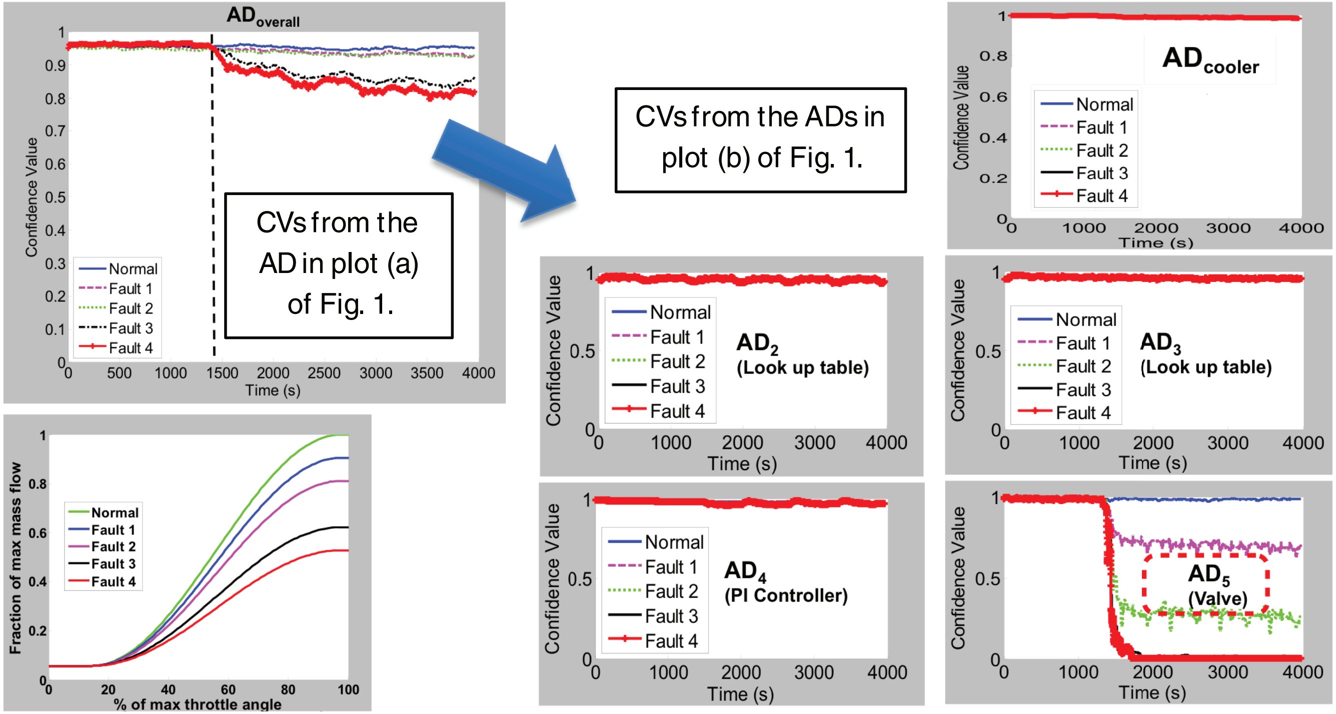


FIGURE 2 Confidence values (CVs) from anomaly detectors (ADs) that proliferate through the exhaust gas recirculation (EGR) system as progressively severe faults are inserted into the EGR valve. Interpretation of CVs clearly identifies the valve as the source of the problem. PI = proportional-integral. Figure can be viewed in color at http://www.nap.edu/catalog.php?record_id=18558.

variables affect the performance of a given system as inputs and which act as outputs, potentially affecting other systems. In an automotive engine, this information can be known from the design of the control system (Cholette and Djurdjanovic 2012b). But in a modern lithography tool or etcher, hundreds of subsystems and components operate in very different physical domains and at very different timescales, emitting thousands of signals representing physical variables whose interactions are not well understood even by the domain experts.

Degradation Modeling and Hidden Markov Models

Anomaly detection is particularly challenging in the case of inherently unobservable phenomena, such as plasma used in CVD, etching, and lithography processes. Plasma is described by 3-dimensional fields of magnetic induction, pressures, and temperatures. But it is possible to detect only a few temperature and pressure points, a few characteristics of the magnetic field at one or two selected points. The state of plasma between these points is inherently invisible, although it can be inferred with more or less confidence from the sensor readings.

In addition, the variable operating conditions of modern manufacturing machines in highly flexible and reconfigurable environments mean that degradation dynamics and fault signatures—which, as mentioned above, are often only probabilistically visible—change with the operating regime. An assumption that the underlying condition of a system is deterministically related to one or several sensor readings in this context is inadequate to cope with such complexity. Advanced signal processing, statistical analysis, and time-series modeling, which worked so well with rotating machinery and traditional manufacturing, do not help in this case.

A new mathematical construct was recently devised for degradation modeling and anomaly detection in inherently unobservable processes in variable operating regimes. The new method models the degradation process through a collection of operating regime–specific hidden Markov models (HMMs) (Cholette 2012; Rabiner 1989). In this context, the equipment conditions are hidden states (equipment conditions) that are stochastically related to the observable variables (based on the sensor readings). The observable variables, hidden states, and dynamics of progression of the hidden states are made to be operating regime dependent, thus enabling context-dependent, operating regime–specific degradation modeling.

Figure 3 illustrates degradation modeling via interconnected HMMs, each of which corresponds to an operation executed on the monitored system.[1] The state dynamics of each of the underlying HMMs is modeled to be unidirectional,

---

[1] On the fictitious machine shown in Figure 3, operation 1 is executed first, followed by operation 2, then operation 1, followed by a maintenance operation and then a totally new operation 3. Each of these operations has its own HMM, with hidden states representing degradation states, while sensor readings are modeled as observable HMM variables.

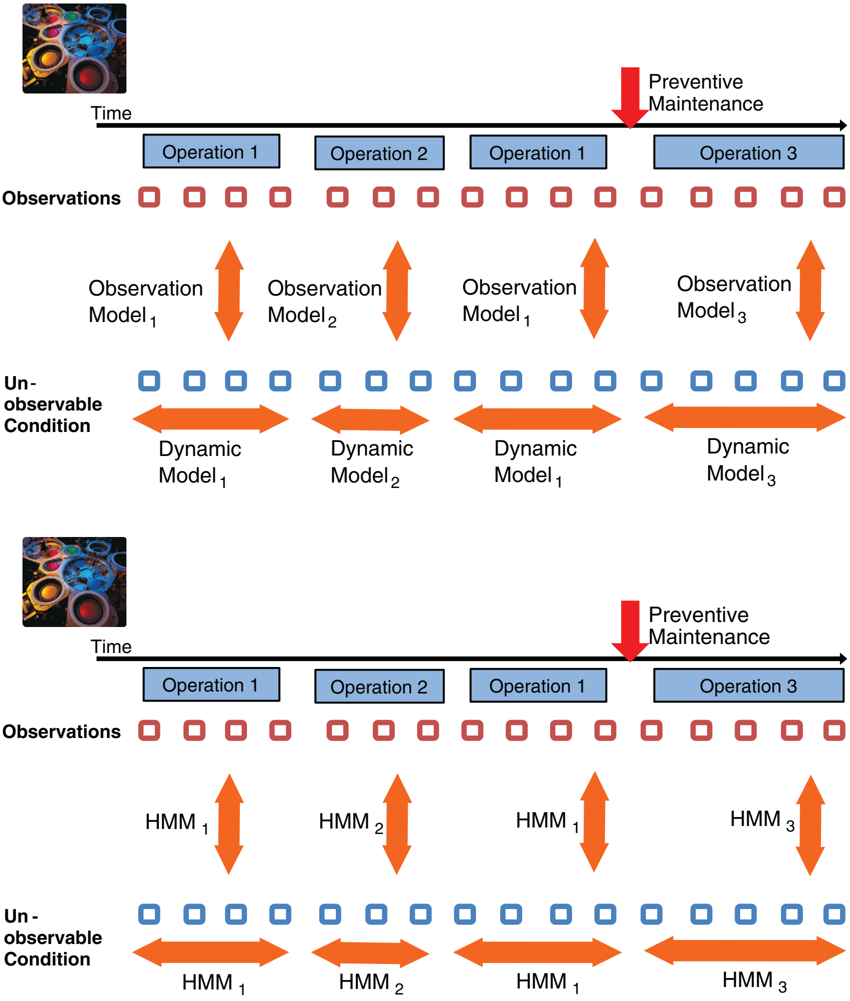


FIGURE 3 Illustration of degradation modeling based on the concept of interconnected, operating regime–specific hidden Markov models (HMMs) of equipment degradation.

which means that without maintenance, the condition (hidden HMM state) of the monitored system increasingly degrades. In addition, following the continuity of degradation, the probability distributions of hidden states at the end of one operation become initial state probabilities for the next. Based on these assumptions, Cholette (2012) introduces a genetic algorithm (GA)–based procedure to identify parameters for the degradation-describing HMMs using sensor readings emitted by a system in arbitrarily mixed operating regimes.

Once the operating regime–specific degradation HMMs are identified, conditional log-likelihoods of the new sensor observations can be used to detect abnormalities. These log-likelihoods drop linearly with the length of an observation sequence, with slopes corresponding to the HMM dynamics. Normalizing these slopes by removing the means of operating mode–dependent slopes and scaling them with operating mode–dependent variances enables the detection of unusual slopes (sequences inconsistent with the degradation HMMs observed during normal system behavior) using simple statistical process control methods, such as an exponentially weighted moving average (EWMA) chart. Detailed information about methods for identifying operating mode–specific degradation HMMs and for detecting anomalies based on such models of degradation are available in Cholette (2012).

The HMM-based degradation modeling and monitoring methods were applied to a PECVD tool at the semiconductor manufacturing facility mentioned above. Over a period of more than 6 months the tool deposited films on more than 100,000 standard 300 mm silicon wafers of the same chemistry but of four varying thicknesses. Signals were collected at a 10 Hz sampling rate from sensors mounted on the tool to gauge gas flow, pressure, radio-frequency (RF) power generation, and chamber pressure. A set of features, such as signal rise times, overshoots, time-durations, and amplitudes of various events during the deposition process were extracted from the sensor signals.[2]

The EWMA chart of normalized log-likelihoods of observations and the corresponding 4-σ control limits are shown in Figure 4. The dashed vertical line labeled “Training” demarcates the data used to identify the parameters of the operating regime–specific HMMs. Two major downtimes caused by severely unacceptable tool behavior were observed during this period and are labeled in Figure 4 as “Big particle event” and “Coulomb crystals.” In addition, based on the analysis of particle counts obtained via particle monitoring wafers (special non-production wafers occasionally passed through the system to assess particle contamination in the system), several minor particle events occurred and are also labeled in Figure 4.

The first minor particle excursion coincides with the cluster of out-of-control points visible somewhat before the first preventive maintenance. During this period particle-monitoring wafers[3] showed a significant increase in particle counts. These anomalous events lasted several days and were consistent with particle contamination from within the chamber itself, leading to a number of scrapped dies on the contaminated wafers.

---

[2] For details on the feature extraction procedure, the reader is referred to Bleakie and Djurdjanovic (2011).

[3] Nonproduction wafers “sense” particle contamination in the chamber by essentially undergoing a deposition process, after which they are sent to a metrology tool that counts the number of particles on the wafer.

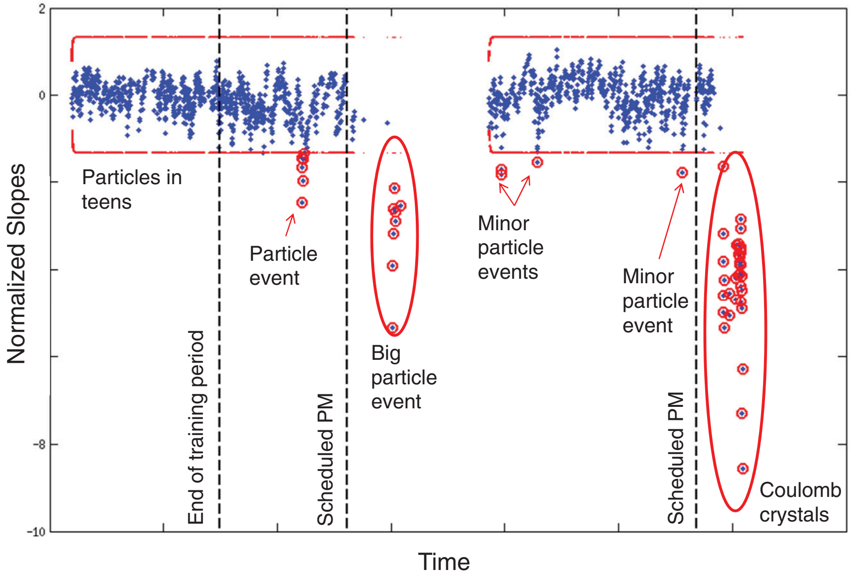


FIGURE 4 Exponentially weighted moving average (EWMA) control chart of normalized log-likelihood slopes k_T, with indications of the times of abnormal behavior events and particle excursions observed during the monitoring period. Each dot in the figure corresponds to a batch of 25–100 wafers.

The first major downtime was more than a week long and was characterized by particle counts two orders of magnitude higher than normal. The length of the downtime was due to the inability to find the source of the particles; various subsystems of the tool were overhauled until the problem simply went away (suspicion is that preventive maintenance that occurred just a bit before this downtime, may have been done poorly and led to the particle failures).

Between the two major downtimes, several minor particle excursions were observed, during which particle monitoring wafers had significantly elevated counts (slightly less severe levels than what was seen during the first minor particle excursion and order of magnitude less severe than what was seen during the first major downtime). As can be seen from Figure 4, each one of those events was flagged by the out of control (unusually low) normalized likelihood slopes produced by the method described in this paper. No downtime occurred due to any of those excursions simply because the nature of the production process was such that particle levels observed during those periods could be tolerated (the product produced in this factory was not requiring the highest possible levels of particle control).

Finally, the second major downtime occurred at the end of the dataset shown in Figure 4. During this event, a number of particle failures and "plasma formations" were noted in the chamber. Again, days of downtime ensued, with fruitless attempts to clear the chamber of the source(s) of particles. After consultation with experts from the company that made the tool, the events were found to be the result of improper evaporation of the deposition product, which caused a phenomenon known as Coulomb crystals (Chu and Lin 1994). In other words, the source of the problem was not even close to the chamber: it was in the gas delivery system, and the failure there led to symptoms that looked as if the problem were in the chamber.

Potential Approaches for Localizing the Sources of Problems in Complex Manufacturing Machines

Although degradation modeling enables anomaly detection even in partially observable, complex processes, the ability to automatically pinpoint the subsystems and components that led to the anomaly remains a challenge. HMM-based anomaly detection could warn the manufacturer using the PECVD tool in the example described above, thus preventing the loss from wasted wafers, but the duration of downtimes could not be reduced because the manufacturer would still need to go through a trial and error procedure until the root cause of the problem was found.

To automate identification of the root cause using distributed anomaly detection, it is necessary to understand the causal interactions between various subsystems of the monitored system—what variables describe each subsystem and FRU of the PECVD tool, and through which variables and in what ways these systems interact while the tool operates.

In the case of the automotive EGR system, all relevant variables were adequately sensed, and it was possible to see exactly how various subsystems interacted (what were the inputs and outputs of each subsystem and component). But with many types of modern manufacturing equipment, such causal topology (what affects what) of interactions between various subsystems may be much harder to determine. Seemingly, everything could affect everything, especially in highly complex and integrated systems, such as PECVD or lithography tools.

Figure 5 illustrates causal interactions for a PECVD tool. However, even the tool experts who designed and manufactured the tool are not sure whether this graph effectively represents all its interactions. And a lithography tool is even less understood.

Formal and systematic identification of causalities in manufacturing equipment is essential to understand how FRUs interact with each other. An optimal model may emerge from a metaheuristic topological search, such as GA or Tabu search, and evaluations using the Akaike information criterion (Akaike 1974) or minimal description length criterion (Rissanen 1978). Such model discovery methods were previously attempted in hot rolling (Li and Jin 2010; Lin et al. 2008), but never in anything remotely as complex as a semiconductor manufac-

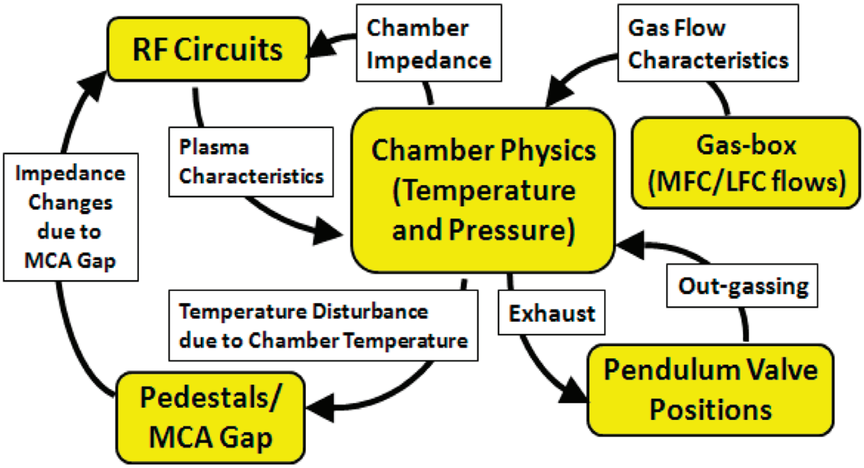
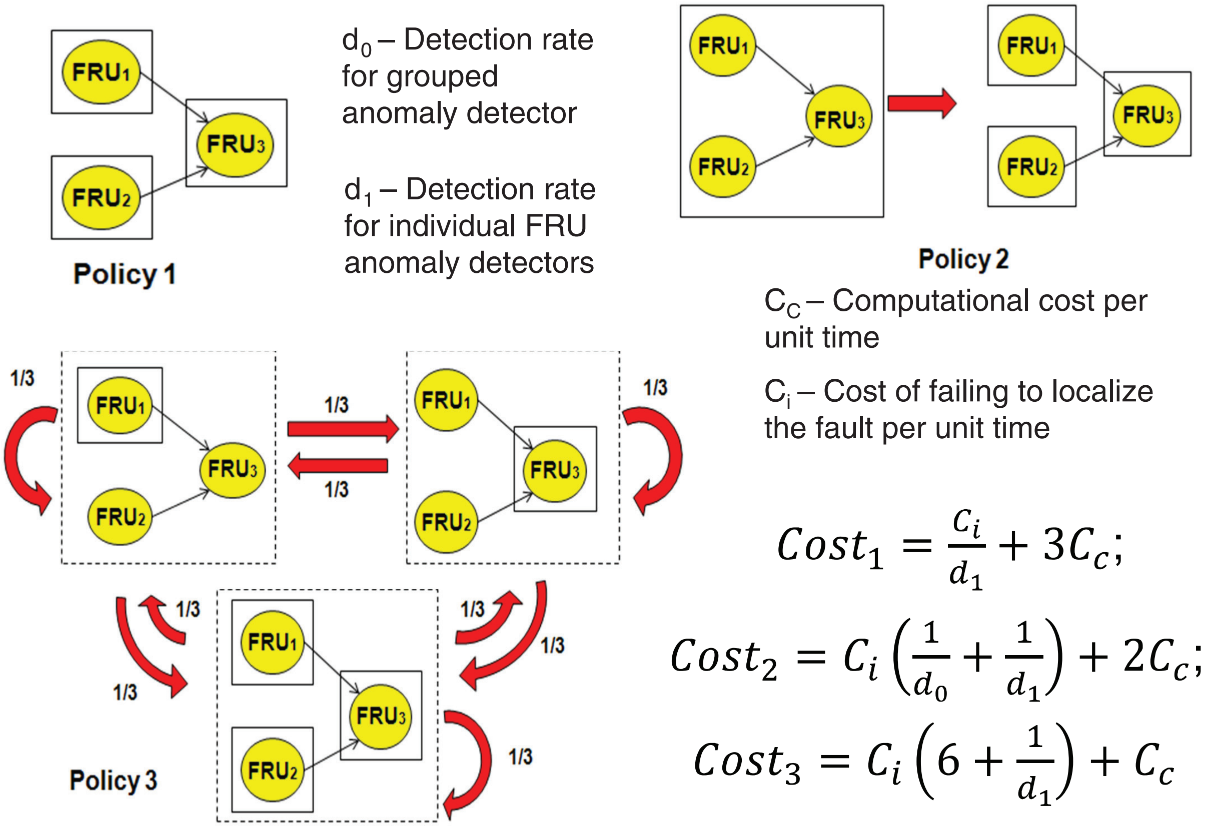


FIGURE 5 Causal graph of plasma-enhanced chemical vapor deposition (PECVD) tool field replaceable units (FRUs), with brief explanations of interaction mechanisms. LFC = liquid flow controller; MCA = minimum contact area; MFC = mass flow controller; RF = radio frequency.

turing tool, where the number of variables is several orders of magnitude greater, modes of operation are much more diverse, and boundaries between (definitions of) FRUs and subsystems are much more blurred.

In addition, research is needed to explore agent distribution policies that optimize tradeoffs between computational resources needed for each AD, their sensitivity to anomaly detection, and the speed of localization of the sources of anomalous behavior. A simple example in Figure 6 illustrates how the use of computational resources can increase fault sensitivity and speed of localization. In that figure, it is assumed that there are only three admissible AD distribution and proliferation policies. Policy 1 is based on always having three ADs monitoring all three FRUs. Policy 2 relies on a single AD monitoring the entire machine, which proliferates into three pertinent ADs as soon as it detects an anomaly (similar to what we see in Figure 2). Policy 3 uses only one AD at a time, and this AD can appear above any FRU with equal probability (implicit assumption that each FRU is equally likely to fail). Let us also assume that only one fault can develop in each of the FRUs, and let d_0 and d_2 be the detection rates within some unit time interval T for the overall AD and each of the FRU ADs, respectively. Also, let C_i be the cost of not reaching the decision within the time interval T, and let C_c be the computational cost of running each of these ADs during the time interval T (realistically, AD monitoring FRU_3 should be more complicated because it has to take into account behavior of FRU_1 and FRU_2, but for the sake of analytic



$$Cost_1 = \frac{C_i}{d_1} + 3C_c;$$

$$Cost_2 = C_i\left(\frac{1}{d_0} + \frac{1}{d_1}\right) + 2C_c;$$

$$Cost_3 = C_i\left(6 + \frac{1}{d_1}\right) + C_c$$

FIGURE 6 Example of tradeoffs in agent distribution policies for distributed anomaly detection. FRU = field replacement unit.

tractability, they are all assumed to be equally complex). Then, simple (but somewhat tedious) algebra gives the following costs for each policy (false alarm costs were not included, but it can be done easily):

$$Cost_1 = \frac{C_i}{d_1} + 3C_c;\ Cost_2 = C_i\left(\frac{1}{d_0} + \frac{1}{d_1}\right) + 2C_c;\ Cost_3 = C_i\left(6 + \frac{1}{d_1}\right) + C_c$$

Obviously, as we go from policy 1 to 2 to 3, the cost of detection grows because more time is needed to isolate the source of the problem. At the same time, the computational cost drops because we are less and less aggressive in terms of AD usage. Thus, policy 1 becomes optimal when the time to isolate the fault is very costly (C_i is high) and/or computational resources are cheap (C_c is low). Conversely, policy 3 is optimal when the time to isolate the fault is cheap and/or computational resources are expensive. Unfortunately, a study like this in any realistic setting (more FRUs, more complex policies, more realistic diagnostic models) has never been conducted.

HUMANS AS DISTRIBUTED AGENTS FOR REMOVAL OF FAULTS (“ANTIGENS”) IN A MANUFACTURING SYSTEM

The unprecedented-fault localization process described above resembles to a degree that of a natural immune system, which identifies and labels an antigen by coating it with appropriately generated antibodies: the diagnostic system described here uses ADs to identify and label faulty subsystems and FRUs.

Once the immune system labels an antigen, leukocytes (white blood cells) dispose of the intrusion by killing anything coated with antibodies. The job of “antigen removal” in a manufacturing system is performed by maintenance practitioners who effectively act as leukocytes. However, unlike the natural leukocytes programmed to kill anything antibodies label as “nonself,” manufacturing “leukocytes” (people) can think, learn—and forget over time. There is a tremendous need for innovative methods to model and match dynamically evolving human skills with maintenance and operational jobs, including (especially) those corresponding to unprecedented situations and faults.

A tree-based representation of machine faults and human skills (Figure 7) may be a way to ensure that a dispatched practitioner has the appropriate skills, even when confronted with an unprecedented problem. An optimization procedure for the joint scheduling of operations and dispatch of operators would also take into account the dynamics of evolving operator skills and interactions between various components of the manufacturing system.

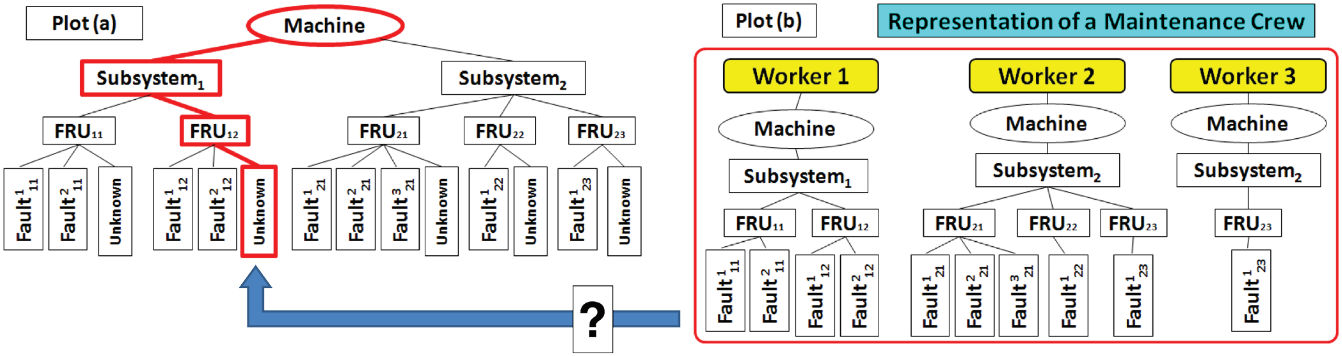


FIGURE 7 Plot (a) illustrates the proposed tree-based representation of a machine and its subsystems and field replacement units (FRUs). It also shows a representation of a maintenance job as a portion of the machine tree "illuminated" (with heavy lines) by the diagnostic system (here, it points to an unknown fault in FRU_{12}). Plot (b) shows the assignments and skills of a three-member maintenance crew, enabling ready identification of the worker best suited for the particular job.

CONCLUDING THOUGHTS

It is clear that a fruitful focus of research pertains to the need and possibility to realize equipment diagnostics and the entire realm of manufacturing system operations through the concept of distributed agents capable of dealing with new, unprecedented situations. Incorporating such agents, architecture of autonomous process and operations control—cognizant and aware of "self" and "nonself"—may be dearly needed as the complexity of future manufacturing systems grows into the realm of cognitive manufacturing.

REFERENCES

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723.

Bleakie A, Djurdjanovic D. 2011. Dynamic feature monitoring technique applied to thin film deposition processes in an industrial PECVD tool. ASME 2011 International Manuacturing Science and Engineering Conference, Corvallis, OR.

Cholette ME. 2012. Performance monitoring and fault-tolerant control of complex systems with variable operating conditions. PhD dissertation, Department of Mechanical Engineering, University of Texas at Austin.

Cholette M, Djurdjanovic D. 2012a. Immunity-inspired monitoring of systems of interacting dynamic systems. Applied Intelligence 37(1):60–79.

Cholette M, Djurdjanovic D. 2012b. Precedent-free fault isolation in a diesel engine EGR system. Journal of Dynamic Systems, Measurement and Control 134(3):031007-1–031007-11.

Chu JH, Lin I. 1994. Direct observation of Coulomb crystals and liquids in strongly coupled RF dusty plasmas. Physical Review Letters 72(25):4009–4012.

Djurdjanovic D, Liu J, Marko KA, Ni J. 2010. Immune systems inspired approach to anomaly detection, fault localization and diagnosis in automotive engines. In: Applications of Neural Networks in High Assurance Systems, ed. Liu Y, Schumann J. Springer-Verlag Berlin Heidelberg. pp. 141–164.

Forest S, Perelson AS, Allen L, Cherukuri R. 1994. Self-nonself-discrimination in a computer. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Los Alamitos, CA: IEEE Computer Society Press.

Heng A, Zhang S, Tan ACC, Mathew J. 2009. Rotating machinery prognostics: State of the art, challenges and opportunities. Mechanical Systems and Signal Processing 23:724–739.

Lei Y, Djurdjanovic D, Barajas L, Workman GC, Ni J, Biller S. 2011. Network health monitoring for DeviceNet using physical layer parameters. Journal of Intelligent Manufacturing 22:289–299.

Lei Y, Djurdjanovic D, Ni J. 2010. DeviceNet reliability assessment using physical and data link layer parameters. Quality and Reliability Engineering International 26:703–715.

Li J, Jin J, Shi J. 2008. Causation-based T2 decomposition for multivariate process monitoring and diagnosis. Journal of Quality Technology 40(1):46–58.

Li J, Jin J. 2010. Optimal sensor allocation by integrating causal models and set covering algorithms IIE Transactions on Quality and Reliability Engineering 42(8):564–576.

Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2):257– 286.

Rissanen J. 1978. Modeling by shortest data description. Automatica 14(5):465–471.

Spiewak SA, Duggirala R, Barnett K. 2000. Predictive monitoring and control of the cold extrusion process. CIRP Annals of Manufacturing Technology 49(1):383–386.

Business Process Management Systems to Optimize Manufacturing

Christian Will
Dassault Systèmes

Challenging economic trends, rising value chain complexities, and intensified global competition are driving the manufacturing industry to upgrade its execution systems. And advances in cloud computing, big data, social collaboration technologies, and mobility are prompting society in general toward the digitally connected enterprise and value chain, which must ultimately satisfy the demands of a better-educated and socially aware consumer. These market dynamics and technology advances pose challenges but also offer opportunities to those who successfully leverage and incorporate them into their mainstream. For manufacturers, the challenges are encouraging a fundamental reassessment of their current and future factories.

This paper introduces key concepts behind a manufacturing execution system built on a process-centric software architecture designed to meet these challenges. After introducing the concept of business process management (BPM), I explain how it supports business intelligence by incorporating assisted and automatic decision making into the manufacturing processes. I then explore opportunities for embedding emerging technologies in the BPM approach.

MODEL-DRIVEN DEVELOPMENT

For business information technology (IT) groups, the use of models has played a key role in enabling both technical and nontechnical professionals to work together to debate and define the business processes and requirements of a system. Such models motivated these groups to create new applications, refactor existing ones, or help guide vendor selection and potential customization of an off-the-shelf application. Modeling languages and tools proliferated during the

1990s, with various methods springing into popularity only to be replaced by others. Today, the most notable surviving method is the unified modeling language (UML). Since 2000 the market for such tools has seen a steady decline, because of UML's complexities as well as the transition to agile, lightweight development methods such as extreme programming.

But the complexity of systems grew with the advent of XML and Web services, the shift toward componentizing application functionality into service-oriented architectures, and the seemingly endless possibilities of transforming them into solutions. Whereas HTML or Hyper Text Markup Language served the first-generation of the Internet well by enabling the creation of Web pages and other content to be displayed in a Web browser, XML, or the Extensible Markup Language, brought on the next generation by introducing a set of rules for encoding documents in a format that is both human-readable and machine-readable. XML also formed the basis for describing Web service interfaces, which businesses and applications would soon use to exchange information over the Web. The increased complexity in turn led to a renewed popularity in model-driven approaches. With advances in graphical software modeling tools, models became the de facto standard for IT programmers and users to develop and maintain applications and the primary vehicles for managing systems throughout their life cycles.

FROM MODELING PROCESSES TO EXECUTING THEM: INTRODUCING BPM

Around the time when model-driven development tools based on the UML reached their peak (ca. 2002), a competing camp emerged seeking to transform models to a machine-readable form that could be executed at run time. This camp sought to prevent programmers from touching the underlying code once a process was authored, thus greatly reducing the need for specialized programming skills to manipulate a solution. The result was the emergence of a now large number of software vendors who deliver packaged software known as *business process management suites* (BPMS). BPM opened the door to nonprogrammers such as business analysts, industrial engineers, and even business users to participate in system development from design through implementation and throughout the life cycle. The rest is mostly history; virtually all of the largest platform houses and dozens of start-up firms currently compete in this segement of software industry.

As BPM vendors sought to model primarily processes instead of entire system architectures, specialized markup languages began to take form. To represent and make models transportable across tools, BPM established the business process model and notation (BPML). And to represent an executable process and make it transportable, XML Process Definition Language (XPDL) soon became the de facto standard.

As the BPM market grew, a few vendors chose to offer a core set of embedded technical and functional components aligned to a specific business context, targeting particular roles, processes, and associated workflows. Apriso's FlexNet BPM product, for example, targets manufacturing operations and brings together engineers from material control, quality, maintenance, and production with IT to configure solutions.

HOW BPM HELPS ACCELERATE MANUFACTURING EXCELLENCE PROGRAMS

Evolution of Organizational Development Methods and the Role of IT

Those familiar with manufacturing excellence approaches such as Lean and Six Sigma, both still in widespread use, may recall the intentional absence of any dependence on IT systems such as those for enterprise resource planning (ERP). Others may remember the hugely popular albeit short-lived wave from the early 1990s known as *business process reengineering* (BPR). While IT played an important role in BPR, it too was short lived.

In many respects BPR was largely a shock-wave approach to push Western companies to quickly respond to threats from overseas competitors that exhibited superior performance in a number of key manufacturing performance metrics and were eroding the US manufacturing base. Those who "survived" the BPR wave by revolutionizing their manufacturing methods soon had to focus on managing and evolving them. By the turn of the millennium, BPR gave way to continuous improvement methods—and business process management was born (Fingar 2006).

BPM readily adopted the main tenets of BPR, which called for using models of current business processes as a starting point for business analysis, redesign, and continuous improvement. With a solid foundation as an organizational development method predicated on evolution, not revolution, BPM advocates knew they had to rely on current IT infrastructures for years to come; IT budgets were severely constrained after 2000 and many companies had yet to show a return on investment on their huge ERP investments.

From Enterprise Resource Planning to Business Process Management

The search for technologies that could integrate with existing systems, shape functionality to the needs of the as-is environment, support global coordination with local operations within the factories, and serve as the vehicle for monitoring performance and targeting process improvement, led to BPM (Figure 1). The BPM approach also aligned well with emerging technologies such as Web services and integration tools that could readily incorporate the functionality of ERP and other traditional applications as well as describe data and interfaces in a neutral manner using XML technologies.

BPM technologies are ideally suited to accelerate process improvement, standardization, and excellence programs because they can translate the outcome

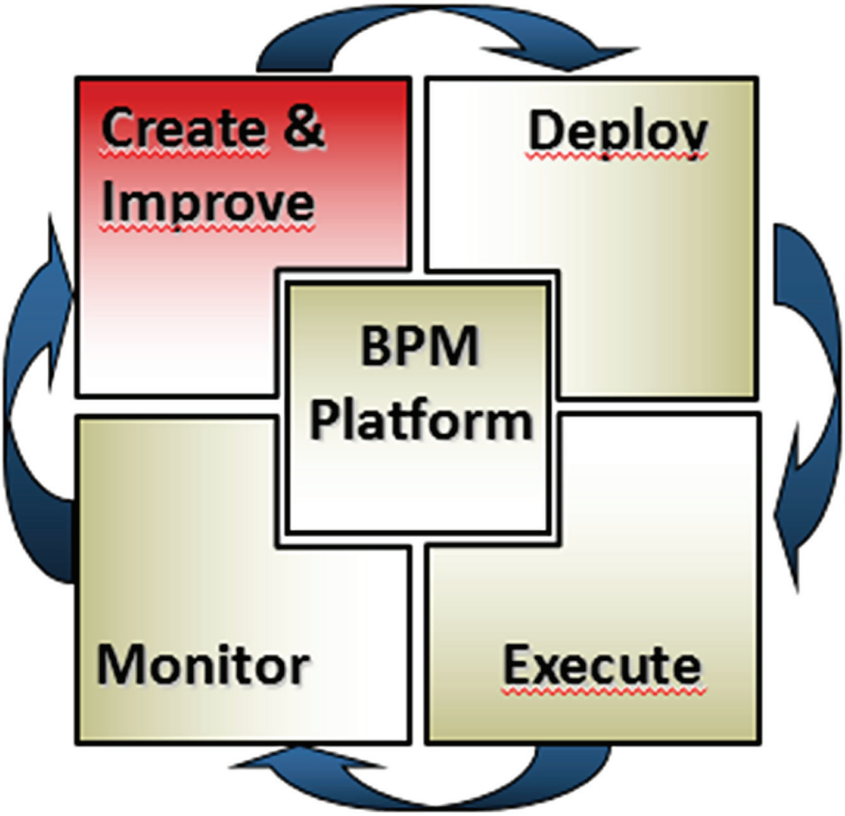


FIGURE 1 Continuous process improvement on a business process management (BPM) platform.

of process modeling efforts into an executable form that limits or eliminates the need to code or customize a core application. Process knowledge that is captured in the model is thus kept current relative to the system's implementation, whereas in almost all other approaches models are retired in stacks of paper or on bookshelves after an initial implementation and become artifacts for a firm's historians.

Recognizing that a generic BPM technology is not an application or a solution in a particular business context, the opportunity exists to further accelerate manufacturing excellence programs by preconfiguring and prepackaging many of the core elements needed for a particular industry. Such packages are integrated in the BPM software and contain a unified model of the business entities involved, a core set of fine-grained business services to accommodate the business entities, and a collection of one or more process fragments applicable to the various business areas in that industry (e.g., production, material handling, machine maintenance, and human capital management). A library of preconfigured, process-centric assets eliminates the need to start from an empty page and provides a solid starting point from which the organization can accelerate its manufacturing excellence programs.[1]

MAKING PROCESSES INTELLIGENT

Processes configured to run in a BPM system can exhibit operational intelligence in one of two ways: by guiding users through the complexities of their process or by autonomously taking some (presumably optimal) action. Both are based on the system's "knowledge" of the particular problem domain to which it is applied. Such knowledge must necessarily be configured into a process. Although the same approach is used in traditional applications, what differentiates a BPM application is that such intelligence is not coded into the solution.

With BPM, intelligent behavior may be added to a single process controlling a work center or automated manufacturing cell, or an end-to-end chain of processes spanning the entire factory, enterprise, or value chain. An example of the former might be the highlighting of a particular sequence for processing work orders to meet competing objectives such as minimizing material consumption, maximizing machine utilization, or focusing on the most important customer. An example that spans the entire factory floor would enhance the just-in-time supply of component materials to production in a manner that accommodates the production schedule while ensuring the efficient use of floor space.

The extent to which intelligence can be enabled is limited only by what the manufacturing excellence team incorporates in its processes and by the kinds of optimization and intelligence tools available in a BPM platform. Often, the BPM can integrate popular tools that are available in the marketplace, such as business intelligence (e.g., analytics, data mining), simulation, statistical analysis, artificial intelligence (e.g., genetic, inference/rules-based), and operations research (e.g., smart math or constraint-based optimization). Such tools can be tightly integrated into the BPM's design and run-time engine, or loosely coupled through a service bus or other means. Hybrid approaches are also possible.

For decision support scenarios, the user may receive guidance in the form of inline feedback. For example, a make-to-order engine manufacturer might configure its assembly operation to conduct background checks of an engine's digital configuration to alert an operator about deviations from an authorized bill of material (e.g., substitutions or engine configuration options) or other checks that need to be performed. Or the system might retrieve specific fail-safe data for an engine

---

[1] Some have referred to the approach of combining BPM technologies with manufacturing excellence programs as outlined by McClellan (2012). This is in stark contrast to traditional approaches, which are data-driven and application-centric.

configuration to provide additional context to both the automation layer and the operator. Fail-safe data might range from the specification of the items and order of the component materials to be consumed, to the tool calibrations and settings for each step of the assembly operation. Standard work instructions might be augmented to display configuration-specific deviations through multimedia visual aids, thereby minimizing the need for operator intervention on the computer terminal.

Decision support may also manifest in the form of supervisorial dashboards that provide data on the throughput and timeliness of a process, end-to-end performance across the entire factory, or process abnormalities. In addition, the system may alert production staff through email, texting, and other notification channels. Automated system actions can also be implemented by invoking secondary business processes (subordinate business processes that are utilized to accomplish a more narrowly focused human-system or system-system function in a common manner), external Web services or applications, or triggering action in a plant's automation layer (e.g., smart machines or programmable logic controllers).

The ability to identify and handle exceptions and to incorporate intelligent decision making in BPM-controlled manufacturing processes is essential. By revealing recurring process exceptions, this capability becomes the core driver for prioritizing and engaging in continuous improvement efforts and opens the door to real-time adjustments of process parameters to improve outcomes. Such adjustments range from changing product data (e.g., to authorize substitute materials when a shortage occurs) to eliminating tasks, adding approval steps, redirecting a work item to another work center, or applying alternative business rules. The timescale for such exception handling may involve long-running transactions that last hours, days, or even months.

OPPORTUNITIES FOR BPM TECHNOLOGY

Connecting the Virtual and the Real: The Role of Simulation

As market dynamics dictate shorter product life cycles, increasingly complex supply chains, and rising costs to bring products to market, simulation may be useful to replace physical prototyping of new product introductions and manufacturing processes. While the cost to physically prototype and test new products and processes increases, at a minimum at the rate of inflation, the cost and capabilities of the virtual will continue to decrease.

For the use of BPM technologies in manufacturing, a natural place to begin is by looking at the simulation capabilities of product life cycle management (PLM) systems. Although there are many applications of PLM simulation, from design through production to sustenance, those applicable to plant operations include the ability to simulate (1) a sequence of actions and alternatives for a single work center, (2) ergonomics associated with the physical actions of an operator, and (3) the interaction among processes throughout the factory.

Because PLM systems can manipulate models of processes, they are ideal for simulating how individual workstations and cells perform activities, from the consumption of component materials to the productive steps performed by an operator or machine to the movement of materials out of a cell. In addition, PLM systems can increasingly simulate human ergonomics to ensure safe and efficient movements in keeping with work instructions.

If the representation of process models is standardized or based on industry standards, the sharing of PLM simulation capabilities can be readily integrated into BPM technologies. In particular, simulation and optimization tools can be used in tandem to improve machine control systems, factory scheduling, and decision support.

The challenge of introducing simulation involves not so much technological capacity as the fact that organizations implementing BPM have been busy for quite some time implementing it and using the more mature analytical frameworks available in the marketplace. So the introduction of simulation is a work in progress.

From Analytics to Big Data: Gaining Insights from Both Structured and Unstructured Data

Analysis of the manufacturing history of a product to search for the root causes of a quality issue, or clues on how to improve production for a part, or any other pattern of interest requires the storage of large amounts of data. The leading database vendors on the market have done a great job of introducing relational database and business intelligence technologies that structure data in a suitable format for viewing and analysis. But these technologies store data in a predefined format. Thus, although large volumes of data can be stored for decades, the data are structured and the repository is usually a relational database, providing limited ability to search and detect patterns in unstructured data that comprises much of what is stored by an organization or across the Internet.

In the case of manufacturing operations data, details about the as-built product, process, or manufacturing history are often dropped for a number of reasons—storage costs, or ERP was deemed the "system of record" and its database design could not accommodate such details, or the manufacturing execution system (MES) transactional database could not retain data for any length of time because of performance concerns. Such data would eventually be archived or purged—as would the ability to analyze or gain insight from it later.

With the maturity of today's business intelligence frameworks, it is possible to extract details from the MES's transactional database and retain them for years in a reporting and analytics data store at a reasonable total cost of ownership. Yet even relational databases and business intelligence data stores cannot keep up with the rapidly growing volumes of data generated by sensors, machines, and other devices in a factory's automation layer. In other cases, the data are highly varied (i.e., unstructured) and contained in various textual document types, log files, blog entries, and other content stored in collaboration portals or email systems.

Enter the world of big data. Until recently only the likes of Google, Facebook, Yahoo, and Microsoft could afford the systems necessary for storing and performing searches of such data. However, big data technologies have become both affordable and usable enough for the average organization and IT skillset. And the big software platform vendors are rushing to introduce, or have already introduced, search engine technologies that work with both unstructured and relational data stores (e.g., Microsoft's SharePoint FAST and Semantic Search, or Apache Lucene), making it possible to store "live," unstructured data right in the database engine and query it at will.

Inclusion of the capacity to search unstructured data in a BPM platform is of immediate benefit to users. It is truly rare to find a process that does not link to at least several unstructured document types. In the aerospace and defense industries, the collection of volumes of unstructured data for as-designed, as-built, as-tested, or as-maintained products has long been a mainstay, as customers demanded such artifacts and were willing to pay large sums to have them. With today's advances in big data technologies, and the low cost of storage and computing power, keeping such details about products and processes is possible for all industries—and they are beginning to see the value and opportunities in the ability to analyze data they couldn't dream of touching a few years ago.

From Processes to Practices: Encouraging Behavior to Drive Innovation and Agility

As many businesses become process-centric and excel at sharing, implementing, and improving processes across their global manufacturing base, one might ask what happens when a business becomes too efficiency-focused? What if it wants to intentionally deviate from standard processes to discover more effective ways of manufacturing a new product? To do this, the business might want to encourage open debate between engineering and manufacturing, between purchasing and quality, and among other groups.

A key capability of next-generation manufacturing is the ability to respond promptly to changes in demand or consumer and market trends. Grieves (2011), in his book *Virtually Perfect*, mentions innovation and other unstructured "practices" (e.g., whiteboard, blogging, crowd sourcing, and gamification) that an organization should encourage to achieve corporate goals and develop an agile culture.

To help organizations engage in unstructured practices, IT groups often introduce social collaboration and content management tools. The opportunity here is to determine the role of BPM in supporting such practices. Could it help accelerate the transition from new engineering or manufacturing innovations to well-documented and cost-effective processes during the ramp-up to a new product introduction?

Today, there are significant opportunities for introducing novel BPM approaches to content and human-interaction management. Under content management is the ability to retrieve, create, update, modify, and correlate unstructured content around the context of a process; support video, audio, text, and social streams; and enable content organization around the processes to which content relates. Under human-interaction management is the ability to manage shared work queues and enable advanced visualization, individual and group collaboration, support for virtual communities, generating user experiences based on a user's role within the context of his current activity, the ability to show the on-line presence of users related to the current activities and initiate voice, messaging, and other forms of interactive communication, and managing off-line notifications to affected users—also within the context of an active and managed process.

WHERE THE FUTURE OF BPM MAY LEAD

With BPM technologies generating billions of dollars annually in revenue, one can readily conclude that they pose a significant and growing threat to the market for traditional software development tools. And although the technology can be considered distinct from developer tools because it targets nondevelopers, this is a state yet to be achieved. Most BPM products automatically generate a user interface in some form, but delivery of a responsive user interface that satisfies today's demanding user on the wide range of fixed and mobile devices available requires developers skilled in Web 3.0 technologies.

But the technology has certainly delivered on one of its key promises: the ability to take a continually changing model of a business's processes and keep it synchronized with an executable form. When that executable form incorporates business logic from a library of preconfigured service-oriented architecture components and a consistent model of business data, BPM can deliver solutions that satisfy a wide range of business contexts, including manufacturing operations.

The inclusion of business intelligence, big data, social, and other emerging technologies in BPM software will make this technology an enduring foundation for any process-centric business that has embarked on the journey from being single-plant and efficiency-focused to demand-sensing with the ability to adapt promptly as markets change.

REFERENCES

Fingar P. 2006. Business Process Management: The Third Wave. Tampa FL: Meghan Kiffer Press.

Grieves M. 2011. Virtually Perfect: Driving Innovative and Lean Products through Product Life Cycle Management. Brevard County FL: Space Coast Press.

McClellan M. 2012. Improving manufacturing excellence: managing production processes across the value chain. White paper. Vancouver WA: Collaboration Synergies, Inc.

The Rise of Computer-Enabled Supply Chain Design

STEVE ELLET
Chainalytics

Engineers across all practices, industries, and applications are dealing with increasing system complexity, pushing the limits of engineering and human ability to grasp the large and complex. Whether it is designing an iPhone, a semi-autonomous rover to land on Mars, or a modern, fast, and efficient supply chain, an increasing level of sophistication is required. In supply chain design, this growth is being driven by increasing business complexity, access to "big data," and Moore's Law.

THE ROLE OF SUPPLY CHAIN DESIGN

Like all fields of design, engineers in supply chain design sift through a vast quantity of options to arrive at the best design—one that meets the needs of a business and its customers with minimal cost, risk, and environmental impact. Decisions about where to manufacture and stock products, which transportation modes to use, and what service levels to provide can either give a company a competitive advantage or leave it vulnerable to competitors and service disruptions.

Today's supply chain designers increasingly use large-scale mathematical programming models (with the help of optimization- and simulation-based software tools) to evaluate tradeoffs between cost and performance. These tools enable the sophisticated modeling of end-to-end supply chains to evaluate a large number of alternatives, suggest new configurations, and test the robustness of the alternatives before proceeding with costly implementation.

Supply chain design has become a respected area of industrial engineering, with dedicated academics, practitioners, software vendors, and consultants. Over the past 20 years the field has transformed from spreadsheets and a few early heuristic-based tools to modern tools and techniques that have become the standard for supporting critical design decisions in leading companies.

HISTORICAL OVERVIEW: LINEAR AND MIXED INTEGER PROGRAMMING

Mathematical historians credit Leonid Kantorovich, a Soviet economist, with developing linear programming (LP) in 1939, applying it first to the lumber industry and soon thereafter to the war effort. In 1947 an American, George Danzig, published the Simplex method for solving LP problems, leading to broader applications of the approach.

In supply chain design, the first LP models were used to answer network flow questions—for example, to determine the amount of volume for each node (i.e., facility) or arc (i.e., transportation lane) in a given network configuration. These analyses became commonplace in the 1980s and early 1990s and were typically used to identify the best locations for distribution centers. At this time, it was common to refer to this type of analysis as *network optimization*; the term *supply chain design* came much later once the design capabilities were more robust.

Linear programming led to the development of mixed integer programming (MIP), which has been substantially more useful in supply chain design. It allows for the direct consideration of on-off decisions and step functions (e.g., deciding whether a facility or manufacturing line should be active or inactive and how large it should be). MIP is highly effective, but it creates an enormous amount of mathematical complexity and can break down badly in real-world applications, as explained in Box 1.

Like an efficient searching algorithm, MIP looks across the millions of options and cuts off entire sections of the solution space that can be proven to be worse than the current best solution. This way, only a fraction of the network configurations actually have to be solved to determine a global optimum. But even then, MIP requires large amounts of memory and time, often more than can be accommodated with current hardware in a business-reasonable amount of time.

RECENT ADVANCES IN SUPPLY CHAIN DESIGN

Over the past decade improvements in processor performance and the correlated drop in cost as described by Moore's Law (see Figure 1) have supported a dramatic increase in the speed, complexity, and size of supply chain design models. The most important advances have allowed for the consideration of additional detail and accuracy, thus increasing confidence that the model represents the actual state. Like AutoCAD, the more detailed and accurate the model the better (to a point of diminishing returns, which is still pretty far off in supply chain design). Key recent advances have occurred in computer hardware, big data systems, and modeling tools.

BOX 1
MIP vs. Enumeration in the Real World

Mixed integer programming (MIP) involves a massive amount of mathematical complexity but is useful in solving otherwise intractable problems. Consider a small system of three warehouses (A, B, and C), in which each facility can be either active or inactive. The number of alternative network configurations to evaluate is 2^3 or 8 (A, B, C, AB, AC, BC, ABC, or none of these). If there were 10 warehouses, there would be 2^{10} or 1,024 configurations. But, even then, it is still not impossible to simply test each one by running an LP 1,024 times and choosing the lowest cost (which is what mathematicians like to call enumeration). Now consider Coca-Cola's US finished goods warehouse network, with more than 400 warehouses. The number of options—2^{400} or around 2.6×10^{120}—is more than the estimated number of atoms in the universe (»10^{80}). This single example vividly shows that enumeration has serious limitations in real-world applications.

Computer Hardware

Computer hardware moved to 64-bit Windows. Under a 32-bit system, the MIP solver was limited to 2 GB of memory (up to 3 GB in some configurations). The removal of this constraint, coupled with the low incremental price, has led to an explosion of model complexity in recent years. Also, the common availability of multicore and multiprocessor hardware in recent years has produced a step change in modeling capability. Solving an MIP model generates many subproblems and is therefore well suited to a multithreaded approach.

Cloud-based solving technology is further removing barriers to large-scale modeling. Some vendors have built this capability directly into the software: the user can opt to connect to a remote solution-focused server maintained by the software vendor to solve one or more larger problems. Services like Amazon Cloud also enable users to push the limits of high-end hardware with a fraction of the hardware investment.

Big Data Systems

Big data systems make it possible to access and manipulate the large datasets that underlie supply chain design models. A company's records of orders, shipments, and production, all at the transaction level, are the preferred inputs to the modeling process to ensure an accurate and unbiased model. The rise of business

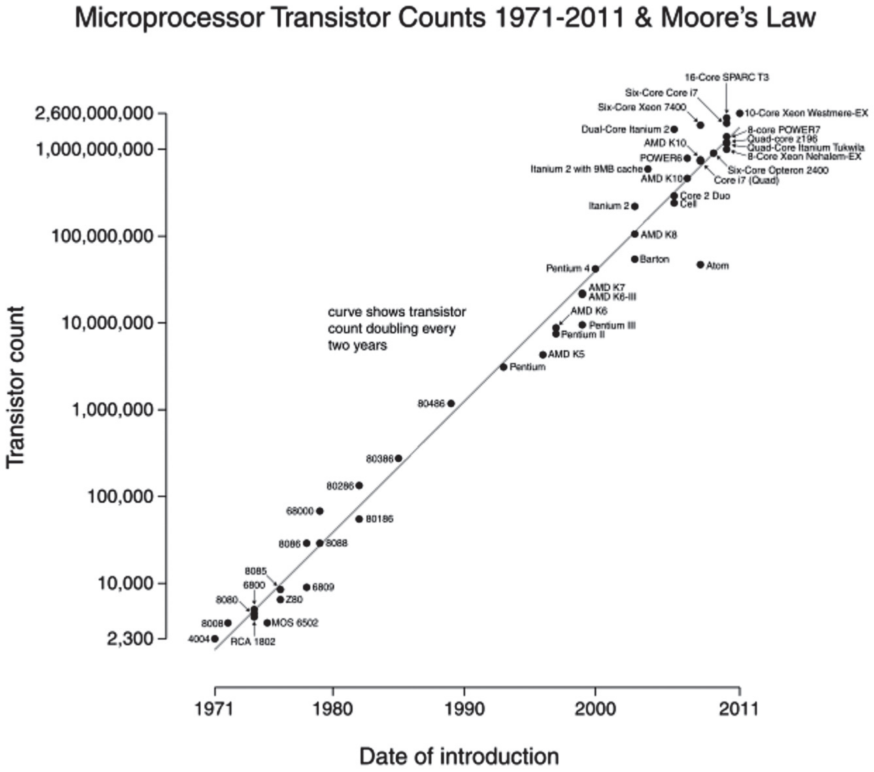


FIGURE 1 The increasing availability of computing power as described by Moore's Law has significantly enhanced the complexity, accuracy, and adoption of computer-enabled supply chain design. Source: WG Simon.

data warehouses, which provide easy access to these data, has increased the use of historical data for analytical purposes, in turn increasing the focus on data accuracy.

In addition to historical data, supply chain design models require what is called *design data,* information about options that did not exist in the historical network, such as candidate transportation lanes and potential facilities. Design data play an important role in supply chain design, and it is essential to guard against bias in such data. For example, consider a model in which all existing transportation lanes use actual costs (i.e., heavily negotiated rates) while new or potential lanes just have a carrier's general estimate and are not negotiated at all. The model would choose all the existing lanes (because they are cheaper), incorrectly reinforcing the current state. The team members might be happy because the model says they are doing a great job and have an optimal supply chain, but this conclusion would be very wrong.

Accurate and unbiased "market data," such as freight costs, are also critical to the success and credibility of the analysis. Big data systems empower large-scale, multicompany econometric models, such as Chainalytics' Freight Market Intelligence Consortium, that produce the required inputs from the market.

Modeling Tools

Modeling tools have become sophisticated but remain easy to use (Figure 2). Some of the important strides in this area are support for multiple objective functions, the coupling of optimization and simulation in a single application, automated sensitivity analysis, math formulation and solver improvements, and usability improvements that have reduced the barriers to entry for inexperienced users.

These advances have pushed the boundaries of supply chain design beyond network optimization into far more complex and valuable analyses of specific manufacturing lines, near-shoring or reshoring (i.e., whether to move production closer to the point of consumption), seasonal production plans, omnichannel distribution, the building of seasonal inventory, global tax strategies, item-specific flow path design, and the consideration of greenhouse gas emissions and other sustainability factors.

CONCLUSION

The world is becoming more complex. Change is accelerating. Companies need to be able to formulate strategies that deal with external factors such as changing oil prices, natural and man-made disasters, and customers' increasing service expectations. Using sophisticated modeling techniques and tools such as MIP, companies are making better, faster, fact-based decisions that require fewer resources to make and move their products to market.

In the world of supply chain design, being more efficient means not only cheaper but also greener. The more accurate and detailed supply chain design models become, the easier it will be to reduce cost and waste.

And this is only the beginning. These tools and techniques have made tremendous strides in the past few decades, but they are still in their infancy. Software companies and practitioners are pushing the envelope on the size of model that can be solved, addressing more and more complexity.

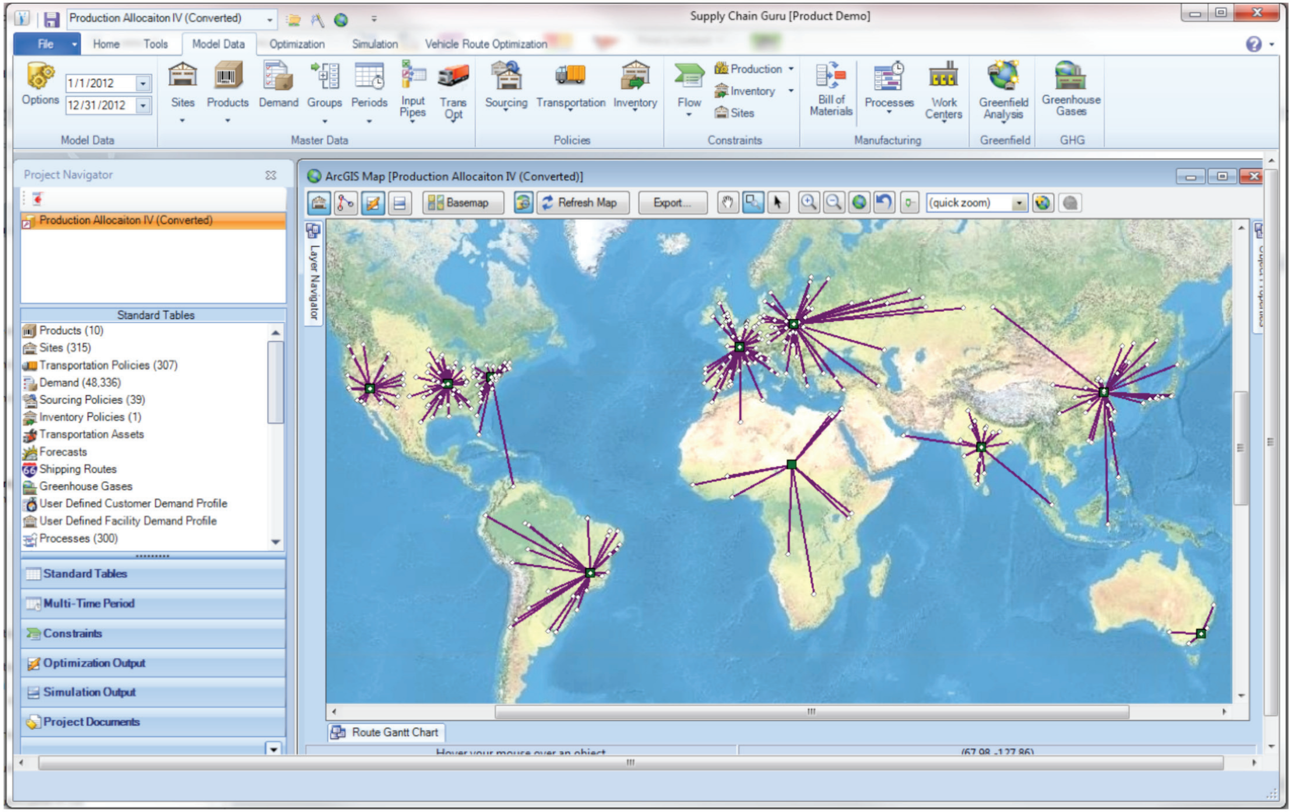


FIGURE 2 Screen image of a modern supply chain design software depicting a global distribution network. Source: LLamasoft.

Advancing Sustainable Manufacturing with the Use of Cognitive Agents

STEVEN J. SKERLOS
University of Michigan

The emerging field of "cognitive" manufacturing is characterized by capabilities and visions for moving beyond "smart" manufacturing toward systems that have the capacity to monitor and evaluate manufacturing performance and then propose process and operations improvements based on sensor and multifaceted data, optimization techniques, and advances in machine learning. In this paper I consider how cognitive systems can advance the state of the art in sustainable manufacturing.

After providing a definition of sustainable and cognitive manufacturing, I explain the importance of an optimization framework for sustainable manufacturing, discuss research needed using different knowledge systems to assess sustainability impacts, and illustrate several applications of cognitive agents to advance sustainable manufacturing.

DEFINING SUSTAINABLE AND COGNITIVE MANUFACTURING

Sustainable manufacturing has been defined by the US Department of Commerce as the creation of manufactured products using processes that minimize negative environmental impacts, conserve energy and natural resources, are safe for employees, communities, and consumers, and are economically sound (Haapala et al. 2013). Implicit in this definition is the support of sustainable development, which is defined as meeting the needs of humanity today without compromising the ability of future generations to meet their own needs (WCED 1987).

Sustainable manufacturing efforts are viewed as incomplete unless they address concerns in three core dimensions: economic, environmental, and social. These dimensions are now commonly referred to as the *triple bottom line*. Sustainable manufacturing can thus be thought of as a multidimensional challenge with components that must be evaluated in a temporal, geographical, and cultural context.

But sustainable manufacturing is challenged by incomplete data, knowledge, metrics, and supporting systems. It must therefore increasingly rely on cognitive manufacturing, which involves the implementation and interaction of a variety of sensors and machine learning techniques that not only provide real-time monitoring but also can perceive performance and suggest alternatives to reduce cost and environmental impacts. These machine-based interpretive systems, called *cognitive agents*, are embedded in the processes of design, supply, production, control, and procurement, to name a few.

The application of cognitive agents can help identify and navigate sustainability tradeoffs in manufacturing decision making. Toward this end, I review research advances needed to help manufacturers establish their sustainability targets. I then suggest that cognitive agents, inspired by early advances in carbon/energy management, can help a manufacturer maximize its profit while coordinating achievement of the company's sustainability targets across its forward/reverse supply chains, manufacturing processes and systems, facility operations, product designs, and, potentially, the influence of future regulations.

OPTIMIZATION FRAMEWORK FOR SUSTAINABLE MANUFACTURING

Translated into an optimization statement, the Department of Commerce definition of sustainable manufacturing would aim to simultaneously maximize economic, environmental, and social performance. But such an approach is inconsistent with how most firms operate. A more realistic formulation would pose economic performance as the objective with the environmental and social dimensions posed as constraints, which could be conceptually written as follows:

$$\begin{array}{ll} \text{MAX} & \textit{Profit = (unit revenue – unit cost) * production volume} \\ \text{Subject to} & \textit{Environmental targets} \\ & \textit{Social targets} \end{array} \tag{1}$$

Equation 1 is a corporationwide execution system problem based on setting environmental and social improvement targets as constraints. The distinction between targets within the objective function and constraints is important because the introduction of environmental and social targets alongside economic targets would require the inherently arbitrary task of monetizing sustainability targets against the firm's primary objective of profit. The inclusion of environmental and social targets in the constraint set is a more transparent expression of what is valued by the firm and makes clear what steps are being taken to address sustainability issues beyond compliance with applicable legislation.

The embodiment of Equation 1 as a design driver for the life cycle influence of manufacturing firms is represented in Figure 1. At the highest level, the manufacturing system would not only optimize product design to maximize profit but also predict environmental and social impacts across the product life cycle to meet environmental and social constraints, using deviations between predictions and data to improve subsequent predictions. The master system would coordinate subsystems at the factory level, and these subsystems would select manufacturing processes and orchestrate facility operations in concert with firm objectives to minimize costs, waste, and negative impacts on workers, communities, and the environment.

Given the complexity of Figure 1 it is appropriate to ponder the development of cognitive agents to support the achievement of sustainable manufacturing objectives. However, as discussed in the next section, the challenges are manifold. Nonetheless, real-world applications of optimization-driven sustainable manufacturing are emerging in narrower contexts, such as factory operations, supply chain design, and manufacturing process planning, as discussed below.

KNOWLEDGE SYSTEMS TO ENABLE SUSTAINABLE MANUFACTURING: RESEARCH NEEDS

The first challenge in achieving the vision of Figure 1 is establishing the link between manufacturing systems and their life cycle environmental and social consequences. Research in environmental and social impact assessment methods is needed to understand these consequences and then, based on the results, set meaningful targets for social and environmental improvement of manufacturing systems.

I discuss below research advances needed in the areas of environmental impact assessment, consequential life cycle assessment, and social life cycle assessment to achieve these ends. Then I conclude this section with a discussion of how firm-level targets can be met with the assistance of coordinated cognitive agents.

Environmental Impact Assessment Methods

Because not all reductions in air/water emissions are equivalent, emissions constraints in Equation 1 cannot be set rationally without some understanding of their real consequences. The need for such knowledge has prompted research to understand the transport, fate, and damages caused by specific product and manufacturing emissions. There is also a need to understand where and when pollutants are emitted throughout the supply chain to connect manufacturing system decisions to their real environmental impacts. As a result researchers in the life cycle assessment (LCA) community are working to resolve LCA data spatially and temporally as well as to interpret the impacts of emissions in terms of ecology and toxicology (e.g., Pennington et al. 2006; Reap et al. 2008). This research

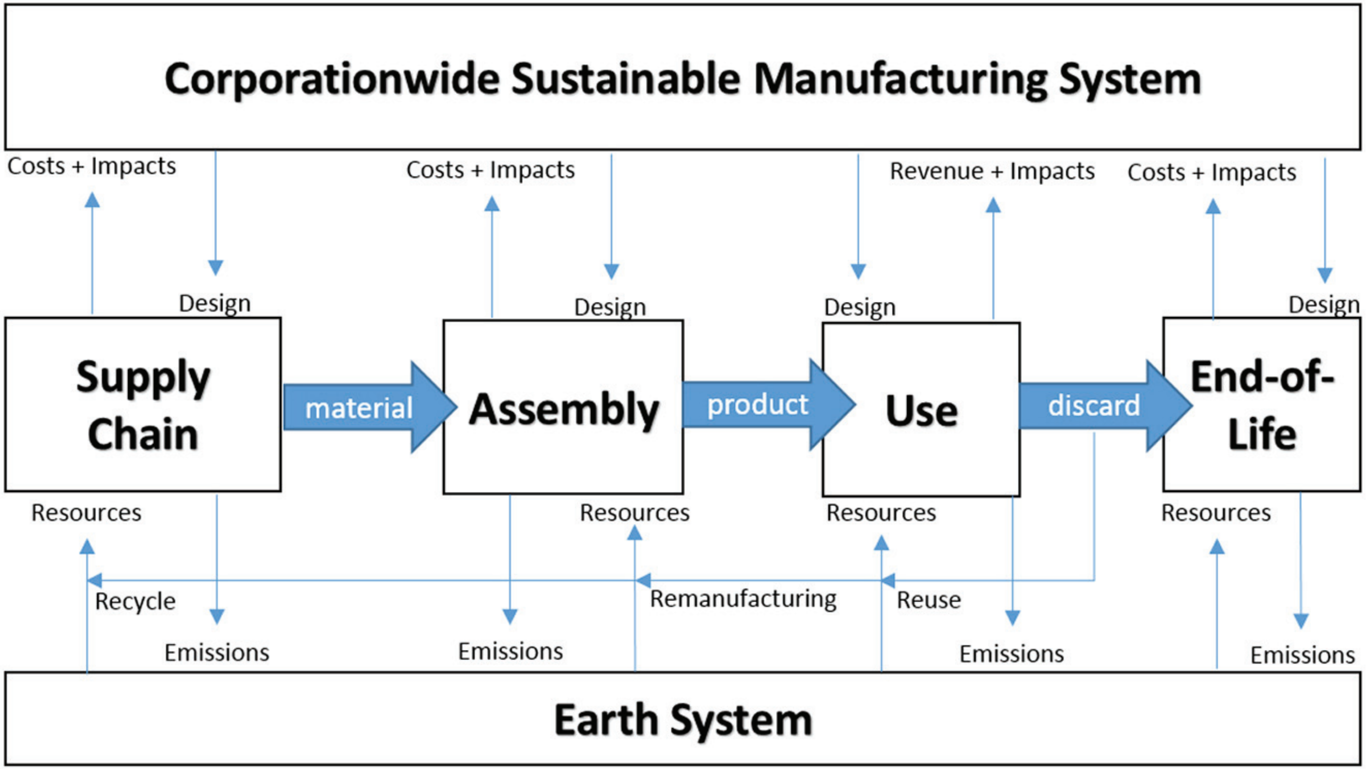


FIGURE 1 Scope of consideration for cognitive agents advancing sustainable manufacturing systems. The top-level system coordinates operational subsystem agents and product/supply chain design choices toward the achievement of corporationwide sustainability targets.

is important to help manufacturing firms provide their stakeholders with relevant and unbiased data as sustainability targets are established.

Consequential Life Cycle Assessment (cLCA)

Life cycle assessment entails holistic consideration of the environmental impacts of a product or process, starting from material acquisition and continuing to the product's end of life, accounting for all the unit processes in the product system (Curran 2006; ISO 2006). cLCA developed from the need to expand the system boundaries of LCA beyond a single product to the interactions between one life cycle and another (Ekvall and Weidema 2004; Finnveden et al. 2009; Hertwich 2005).

cLCA is a technique that can identify expected changes to the environmental performance of interacting manufacturing systems when new technologies are proposed to advance sustainability in factories. Because sustainable manufacturing exists at the nexus of multiple product and process life cycles, advances in cLCA methods and data are necessary to help make decisions that may be counter-intuitive but are actually better for the environment and society. For instance, carbon dioxide may be viewed either as a greenhouse gas pollutant to be avoided in manufacturing or, less intuitively and more correctly, as an environmentally conscious process fluid that is a waste from other industrial processes and that should be recovered for use in manufacturing to eliminate health hazards and water pollution while reducing life cycle greenhouse gas emissions. cLCA methods can distinguish between surface-level rules of thumb and the deeper causalities that actually drive environmental impact.

Social Life Cycle Assessment (sLCA)

sLCA is an effort to fold social aspects of a product or system into environmental life cycle assessment. Jørgensen and colleagues (2008) reviewed sLCA methods and summarized many midpoint indicators (e.g., in the areas of human rights, working conditions, labor practices, job creation, community communication, corruption) and endpoint indicators (e.g., mortality, morbidity, autonomy, safety, security, opportunity, influence) that manufacturers and their stakeholders can consider in establishing sustainability targets. sLCA is an emerging field, and accelerated efforts to develop datasets, metrics, and interpretation methods are needed to advance it for seamless use with other LCA tools in the context of sustainable manufacturing.

From Targets to Action

Advances in LCA will help firms establish targets that are well aligned with the ecological and societal sustainability requirements most relevant to the firms' activities. Once targets are set, they need to be rationally distributed among products and operations. For instance, a firm may set a goal to reduce its carbon footprint in addition to reducing water pollution and improving workplace health and safety. Today it is common for such goals to be applied uniformly across the firm's business units or factories. But such an approach is not cost optimal and could limit opportunities for breakthrough improvements. For instance, it might be more practical and advantageous to increase the carbon footprint in one factory if it means eliminating worker health risks and water pollution at that factory. In such a case the firm's other factories might reduce their carbon footprint so that the firm can meet its overall sustainability goals.

Given the complexity of operations and the multidimensional nature of sustainability, a quantitative approach consistent with the goals of cognitive manufacturing can be useful to coordinate activities in a manner that achieves all firm-level sustainability goals at least cost.

Analytical target cascading is an optimization method that decomposes a system into a hierarchy of subsystems and coordinates their optimization problems such that the solutions are consistent with the overall optimization solution for the top-level system (e.g., Kim 2001; Nyström et al. 2003). Sustainability applications of target cascading can help firms determine sustainability targets for specific products and manufacturing processes to ensure that the firm meets all its goals without unintentionally compromising some (e.g., worker exposure to process chemicals) while pursuing others (e.g., reductions in carbon footprint).

TOWARD COGNITIVE AGENTS TO ADVANCE SUSTAINABLE MANUFACTURING

Advanced optimization techniques are now being applied in supply chains to minimize fuel costs and carbon emissions by proposing alternative transportation modes and routes. Such efforts are being encouraged by the US Environmental Protection Agency (EPA) SmartWay program[1] and adopted by large corporations such as Walmart. These supply chain design tools can be easily modified to include additional metrics. A top priority should be to minimize the harm caused by other air pollutants emitted alongside carbon dioxide, because air emissions from transportation systems are a significant cause of disability and premature mortality (e.g., Caiazzo et al. 2013).

In factories, systems are emerging to help reduce energy consumption and carbon emissions from manufacturing operations. These systems range from control systems for lighting and HVAC based on occupancy to the timing of machine warm-up and standby assignments based on production schedules. More advanced

---

[1] According to the SmartWay website (www.epa.gov/smartway/), "SmartWay® is an EPA program that reduces transportation-related emissions by creating incentives to improve supply chain fuel efficiency."

systems are being applied to optimize production schedules based on time of day and peak demand electricity charges.

As automated systems "learn about" their own energy consumption relative to alternatives available in the marketplace, they will be able to generate suggestions for capital purchases of equipment such as motors and pumps to increase manufacturing process efficiency and eliminate waste. This takes the "energy treasure hunt" concept[2] and embeds it in the factory's cognitive control.

Cognitive agents applied to sustainable manufacturing would extend energy/carbon considerations to material and water consumption, air and water pollutant emissions, and long-term health impacts on workers. These metrics can be constantly evaluated relative to firm-level sustainability objectives to yield suggestions generated by cognitive agents for changes to facility operation or manufacturing process selection.

Two hypothetical examples below illustrate how cognitive agents could begin to influence manufacturing process selection.

Energy Consumption of Alternative Manufacturing Pathways

In the first case a cognitive agent is endowed with models of energy consumption for alternative processes to make dies and molds, with both additive and subtractive manufacturing pathways. The cognitive agent considers the following conceptual problem:

MIN *Production cost* (2)
Subject to *Reduce life cycle energy per unit product*

Morrow and colleagues (2007) built energy consumption models for tool and die production based on subtractive and additive pathways, establishing criteria for the selection of additive manufacturing in Equation 2 over conventional milling. They found that products with high cavity percentage in the total volume were viable candidates for sustainable manufacturing via an additive pathway and that additive manufacturing created the possibility of new mold and die systems with lower life cycle energy consumption (e.g., in systems with conformal cooling channels, heat sinks, protective coatings, and remanufacturing).

In this case the cognitive agent facilitates process selection to include additive, subtractive, and combination pathways that minimize production cost while reducing life cycle energy consumption. This concept is illustrated in Figure 2, which shows the modification of a mold die from a capital *M* to a lowercase *m* using the combination of milling (subtractive process) and direct metal deposition (DMD; additive process).

---

[2] Energy treasure hunts, developed by Toyota, are an extension of the concept of lean manufacturing, aiming to eliminate energy waste.

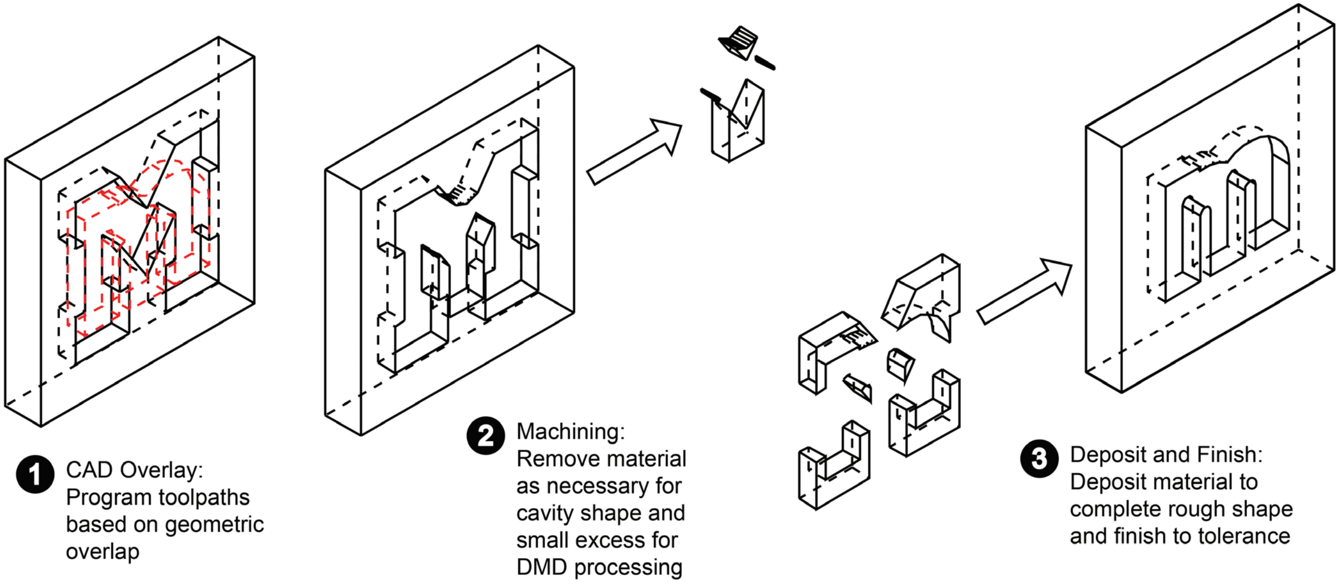


FIGURE 2 A hypothetical application of combining additive and subtractive manufacturing to convert mold tooling with minimal material investment and environmental emissions. CAD = computer-aided design; DMD = direct metal deposition.

Gas-Based vs. Water-Based Cutting and Grinding Fluids

Aqueous metalworking fluids are significant polluters of water and cause long-term health risks for workers (e.g., Skerlos et al. 2008). In this case we suppose that a cognitive system is aware of alternative metalworking fluids, such as gas-based minimum quantity lubrication systems, and considers the following variant on Equation 1.

MIN *Production cost* (3)
Subject to *Reduce disability adjusted life years (DALYs) for workers*
Reduce water consumption
Reduce fats, oils, and grease emissions to water
Quality, throughput not reduced

Based on the materials being machined and on process operating parameters, cognitive agents could advise on the availability of more sustainable metalworking fluid alternatives. They also could adjust process parameters and process operations to enable the accommodation of environmental and health constraints while minimizing cost. Using productivity and quality metrics fed back to the system, the cognitive system can make decisions about metalworking fluid applications that maximize productivity while minimizing the generation of waste and health hazards. To maximize their effectiveness, the agents would be connected to complementary agents in the wastewater treatment system, occupational health system, tool/fluid/material procurement system, and others such that total system costs to the firm are factored into the decision.

Cognitive Agents Beyond the Factory Walls

Future generations of cognitive systems may link decisions made within the firm to its forward and reverse supply chains. The efforts of large manufacturing firms to understand upstream carbon emissions have already led to Internet-based systems to provide information about supplier carbon performance to centralized databases. Networks of cognitive agents could perform this task in an automated manner while offering, for instance, recommendations to procurement regarding supply chain design (Seuring and Müller 2008) and recommendations to product design for enabling "reverse" supply chains through targeted design for remanufacturing actions (Hatcher et al. 2011).

The notion of cognitive agents working collaboratively to achieve sustainability objectives is fundamentally different from simply linking data systems containing environmental performance metrics. The linked cognitive agents would automatically generate opportunities for firms to collaborate toward reducing emissions via strategies that would yield greater profit for both firms than they could achieve if they acted alone.

This concept would not need to stop at communication between firm-level cognitive agents: these could connect with similar agents at the community and national levels to explore new opportunities for mutual gain. For instance, where regulators aim to reduce the environmental impact of manufacturing firms, cognitive agents at the policy level could connect with those at the firm level to inspire novel solutions such as funding mechanisms for clean technology that could benefit manufacturers by overcoming financial hurdles and benefit society by achieving environmental improvements at less cost than traditional "command and control" regulation.

Research has already begun to demonstrate how cLCA frameworks could support efforts to tackle such challenges by enhancing understanding of the interactions shown in Figure 3 that lead from regulation to production/consumption and ultimately to social and environmental impact (Whitefoot and Skerlos 2012).

SUMMARY

The complexity of sustainable manufacturing demands the creation of new knowledge and systems to set targets for social and environmental improvement and achieve them at least cost. This effort can start with today's nascent

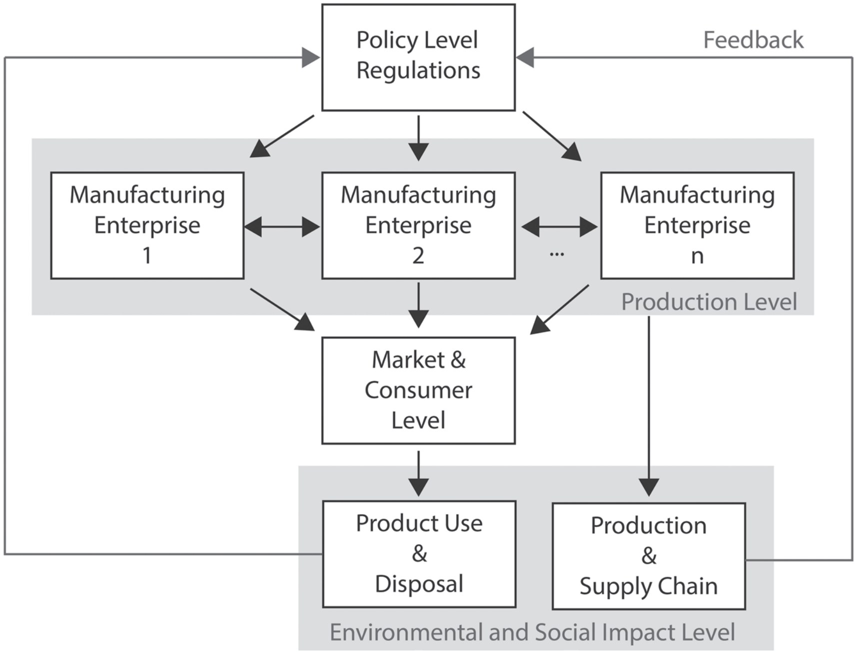


FIGURE 3 Interaction of systems that influence environmental and social impacts.

systems for energy and carbon management and can be extended to a broader set of environmental and health metrics. As cognitive manufacturing systems emerge and gain access to LCA data from the supply chain, they can influence the design of forward/reverse supply chains, factory siting decisions, and broader aspects of manufacturing process selection.

Cooperation between cognitive agents that influence both product design and manufacturing creates opportunities to improve product environmental performance and expand remanufacturing activity. In addition, the cooperation of firm- and government-level cognitive systems can lead to new strategies for achieving sustainability objectives at lower societal cost than permitted by legacy regulatory frameworks.

ACKNOWLEDGMENTS

The author is grateful for the substantial contributions to the concepts described in this paper that were made by J. Rhett Mayor (associate professor of mechanical engineering at Georgia Institute of Technology), W. Ross Morrow (assistant professor of mechanical engineering and economics at Iowa State University), Katie Whitefoot (senior program officer for manufacturing, design, and innovation at the National Academy of Engineering), Panos Papalambros (Donald C. Graham Professor of Mechanical Engineering at the University of Michigan), and Jyoti Mazumder (Robert H. Lurie Professor of Mechanical Engineering at the University of Michigan).

REFERENCES

Caiazzo F, Ashok A, Waitz I, Yim S, Barrett SRH. 2013. Air pollution and early deaths in the United States. Part I: Quantifying the impact of major sectors in 2005. Atmospheric Environment 179:198–208.

Curran MA. 2006. Life Cycle Assessment: Principles and Practice, EPA/600/R-06/060. Cincinnati: US Environmental Protection Agency.

Ekvall T, Weidema B. 2004. System boundaries and input data in consequential life cycle inventory analysis. International Journal of LCA 9(3):161–171.

Finnveden G, Hauschild M, Ekvall T, Guinee J, Heijungs R, Hellweg S, Koehler A, Pennington D, Suh S. 2009. Recent developments in life cycle assessment. Journal of Environmental Management 91:1–21.

Haapala KR, Zhao F, Camelio J, Sutherland JW, Skerlos SJ, Dornfeld DA, Jawahir IS, Zhang HC, Clarens AF. 2013. A review of engineering research in sustainable manufacturing. Journal of Manufacturing Science and Engineering, Transactions of the ASME 135(4):041013-1–041013-16.

Hatcher GD, Ijomaha WL, Windmill JFC. 2011. Design for remanufacture: A literature review and future research needs. Journal of Cleaner Production 19(17):2004–2014.

Hertwich EG. 2005. Life cycle approaches to sustainable consumption: A critical review. Environmental Science and Technology 39(13):4673–4684.

ISO (International Organization for Standardization). 2006. ISO 14040, Environmental Management–Life Cycle Assessment–Principles and Framework. Geneva.

Jørgensen A, Le Bocq A, Nazarkina L, Hauschild M. 2008. Methodologies for social life cycle assessment. International Journal of LCA 13(2):96–103.

Kim HM. 2001. Target Cascading in Optimal System Design. PhD thesis, Department of Mechanical Engineering, University of Michigan, Ann Arbor.

Morrow WM, Qi H, Kim I, Mazumder J, Skerlos SJ. 2007. Environmental aspects of laser-based and conventional tool and die manufacturing. Journal of Cleaner Production 15(10):932–943.

Nyström M, Larsson T, Karlsson L, Kokkolaras M, Papalambros PY. 2003. Linking analytical target cascading to engineering information systems for simulation-based optimal vehicle design. International Conference on Engineering Design, Stockholm, Sweden, August 19-21.

Pennington DW, Margni M, Payet J, Jolliet O. 2006. Risk and regulatory hazard-based toxicological effect indicators in life-cycle assessment (LCA). Human and Ecological Risk Assessment 12(3):450–475.

Reap J, Roman F, Duncan S, Bras B. 2008. A survey of unresolved problems in life cycle assessment, Part 2: Impact assessment and interpretation. International Journal of LCA 13(5):374–388.

Seuring S, Müller M. 2008. From a literature review to a conceptual framework for sustainable supply chain management. Journal of Cleaner Production 16(15):1699–1710.

Skerlos SJ, Hayes KF, Clarens AF, Zhao F. 2008. Current advances in sustainable metalworking fluids research. International Journal of Sustainable Manufacturing 1(1):180–202.

WCED (UN World Commission on Environment and Development). 1987. Our Common Future. Oxford: Oxford University Press.

Whitefoot KS, Skerlos SJ. 2012. Design incentives to increase vehicle size created from the US footprint-based fuel economy standards. Energy Policy 41:402–411.

ENERGY:
REDUCING OUR DEPENDENCE ON FOSSIL FUELS

Energy:
Reducing Our Dependence on Fossil Fuels

HALIL BERBEROGLU
University of Texas at Austin

STUART THOMAS
DuPont

Fossil fuels have served humans well over the past two centuries, improving quality of life and advancing civilization. But we have become overly dependent on this singular energy inheritance that took millions of years to accumulate and are consuming the available reserves at an accelerating pace. A number of serious issues are now associated with this dependence, affecting energy security, national security, air quality, and global climate change. The objective of this session was to provide a broad perspective on these issues and to present and promote technical solutions for diversifying fuel production infrastructure to meet the energy needs of a growing population.

Laura Díaz Anadón (Harvard University) opened with a historical perspective on the development and adoption of fuel resources and then surveyed technical, economic, environmental, social, and policy considerations as well as challenges for technology research and innovation. The second speaker, Joyce Yang (Department of Energy), reviewed advances and hurdles in biofuel production technologies, with a focus on biomass feedstock and processes. In addition to logistical factors (e.g., transportation, infrastructure), she considered these developments in the context of support from federal policies and funding. Willem Rensink (Shell USA) followed with an industry perspective on the need to adapt infrastructure and technology to achieve scale-up and economies of scale in biofuel production from biomass.[1] The session's final speaker, Rachel Segalman (University of California, Berkeley), reviewed artificial photosynthesis research focused on

---

[1] Mr. Rensink's presentation is not included in this volume.

directly harnessing solar energy for fuel production.[2] She explained the mechanics of solar fuel generation as well as materials and design considerations, while acknowledging questions that require further research.

---

[2] Dr. Segalman's paper was coauthored by Dr. Miguel Modestino.

Energy from Fossil Fuels: Challenges and Opportunities for Technology Innovation

Laura Díaz Anadón
Harvard University

Energy has critical impacts on the economic, environmental, and socioeconomic dimensions of human well-being. In the United States more than 82 percent of primary energy consumed comes from fossil fuels (the total for the world is close to 80 percent), which are expected to continue to dominate given the capital intensity, longevity, and incumbent advantages of fossil-based energy systems. But neither the United States nor the world can afford to depend on an energy system that is so heavily reliant on fossil fuels. This paper presents a review of major energy challenges, the role of technology innovation, drivers of previous energy transitions, implications for research, and US policy needs to incentivize innovation.

MAJOR US ENERGY CHALLENGES IN NUMBERS

Accounting for 86 percent of total US greenhouse gas (GHG) emissions (5.7 Gt of CO_2 eq. in 2011), the energy sector is the largest contributor to what is increasingly recognized as the most intractable and dangerous environmental challenge posed by human activity: global climate change.[1] Most energy-related US GHG emissions stem from the use of coal for power and of oil for transportation: 38 percent of the emissions are in the form of CO_2 from the power sector—three quarters come from coal—and 30 percent from the combustion of oil in the

---

[1] In 2005 the United States contributed one-sixth of global GHG emissions (data for later years are not available) and in 2012 one-sixth of energy-related CO2 emissions. The US energy sector is also a major contributor to climate change globally, accounting for more than two-thirds of total GHG emissions in 2012.

transportation sector (EPA 2013). As shown in Figure 1, coal fuels 46 percent of the electric sector and oil 93 percent of the transportation sector.

The consequences for human well-being of GHG emissions from the fossil fuel–based US energy system are already being felt and at a faster rate than expected. A 2007 Intergovernmental Panel on Climate Change (IPCC) report states that it is more than 50 percent likely that human activities have contributed to a rise in the number of heat waves, floods, droughts, and wildfires; hurricanes and typhoons of greater power; and higher risks to coastal property from the surging seas.

Fossil fuel–based energy systems also emit substantial amounts of other pollutants such as sulphur dioxide (SO_2), nitrogen oxide (NO_x), and particulate matter, all of which cause significant health, ecosystem, and economic damages. A National Research Council study estimated that in 2005 energy-related US health costs were substantial: $62 billion related to the effects of coal power, $56 billion related to oil used in transportation, and $0.7 billion related to gas power (NRC 2010).

In economic terms, the fraction of US GDP devoted to oil imports has been rising, oscillating between 1.5 percent and 2.5 percent in 2005–2012, the highest level since 1983 even though crude oil imports have decreased by 16 percent since 2005 (this number is small compared to the 18 percent devoted to healthcare, but

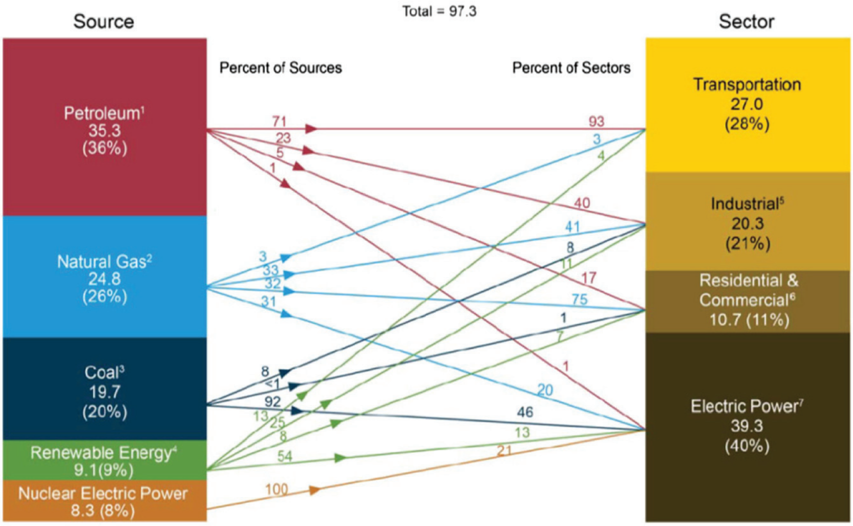


FIGURE 1 Primary US energy consumption in 2011 by source and sector in quadrillion Btu (Quads). Total consumption = 97.3 Quads. Source: EIA 2012.

not so small compared to the 5 percent devoted to defense[2]). The high US dependence on foreign oil is undesirable because it exposes the country to oil supply disruptions and price shocks that could, for instance, be caused by terrorist attacks.[3]

THE NEED FOR AND COMPLEXITY OF ENERGY TECHNOLOGY INNOVATION

There is widespread agreement that innovation in energy-supply *and* end-use technologies is necessary to overcome the major energy challenges associated with US dependence on coal for power and oil for transportation.[4] Partly as a result of the inherent uncertainty and complexity of the technology innovation process, there is more disagreement about the specific role of different technologies.

Innovation starts (but does not end) with discovery and invention (Narayanamurti et al. 2013)—what we engineers call research, development, and demonstration (RD&D). To have an impact, discoveries and inventions need to progress through other "stages": demonstration, market development, and widespread deployment. Of course, technologies do not move through these stages in a linear fashion; stages often take place in parallel and there are feedbacks between them. And the pace and direction of technology innovation are shaped by a multiplicity of actors (e.g., governmental, private, not-for-profit, consumers) and institutions (e.g., norms, policies, culture).

The complexity and enormous size, ubiquity, interconnectedness, and commodity nature of most of the energy system make the transition away from fossil fuels difficult. The experience of previous energy transformations offers some insights about what may be required.

HISTORY OF ENERGY TRANSITIONS

Two main energy transitions have taken place since the Industrial Revolution. The first, starting in the late 18th century and going through the 1920s, was the emergence of steam power from coal, a development that replaced ovens,

---

[2] http://www.whitehouse.gov/administration/eop/cea/TheEconomicCaseforHealthCareReform; http://www.cfr.org/defense-budget/trends-us-military-spending/p28855.

[3] On a global scale, the challenge of providing access to modern energy sources to billions of people to enable economic development is another (and perhaps even greater) challenge. An estimated 2.6 billion people around the world still rely on traditional biomass for cooking, and 1.3 billion have no access to electricity (IEA 2013). Limited access to modern sources of energy is an important contributor to poverty levels worldwide, and a major cause of the 3.5 million premature deaths per year from indoor air pollution (Lim et al. 2012).

[4] In the electricity sector studies variously project more prominent roles for nuclear power, fossil power with carbon capture and storage, renewable electricity, and increased end-use efficiency. In the transportation sector, vehicle electrification, different types of biofuels, compressed natural gas, and increased efficiency could all contribute to reducing oil consumption. Some of these technologies can work as complements, others as substitutes.

boilers, furnaces, horses, and water power, and overcame the limited availability of mechanical power, low-energy densities, and the lack of ubiquitous and cheap transport systems. Stationary steam engines, which largely displaced wind and water power, were first introduced to dewater coal mines and then spread to mechanized textile manufacturing facilities and agriculture and to mobile applications in railways and ships (Grübler 2012).

The second transition was the replacement of coal steam by electricity and petroleum-based technologies, which started slowly in the late 19th century and continues today (see Figure 2). The diffusion of gasoline engines (e.g., automobiles) and electric appliances (e.g., light bulbs) was the driving force behind the second transition. It took about 100 years for steam engines and electric drives to reach 50 percent of market penetration.[5]

In both cases, the transition was driven not by either resource scarcity or lower prices but by niche end-use markets that were willing to pay a premium for performance based on the initial (and crucial) technology improvements. These markets reduced costs through economies of scale, standardization, and learning by doing, and provided time for complementary technologies to emerge and for parallel RD&D to further improve them.

Thus, the historical transitions suggest that a single technology cannot transform the energy system. Rather, transformations often require the formation of technology clusters (e.g., end-use technologies, distribution systems) and new applications of the original technology, both of which take time. Conversely, existing technology clusters and organizations that support the status quo create a path dependency (Arthur 1989), meaning that there is a lack of institutional (e.g., regulatory, social) and physical infrastructures needed to enable the deployment of new technologies.

Now, however, most of the technologies with the potential to significantly displace coal in the power sector and/or oil in the transportation sector are not expected to offer significant comparative advantages in terms of services provided to consumers (with the exception of some of the environmental externalities, which are still not priced in the United States). In addition, they are not expected to offer reduced costs in the short term. And niche markets (e.g., the use of biofuels or biofuel coproducts for the high-value chemicals sector, or of CO_2 or other waste products as input to biofuels or other chemical synthesis) may be insufficient to drive down costs quickly enough. The commodification of energy, the path dependency of large systems, the fact that early versions of "hardware" energy technologies are usually risky and expensive, and the fact that the energy system is large, capital intensive, and long-lived all suggest that government policies to

---

[5] Technology improves during the diffusion process. For example, it took about 100 years for steam engine efficiency to increase from 1 percent to 20 percent and almost another century to increase from 20 percent to 40 percent.

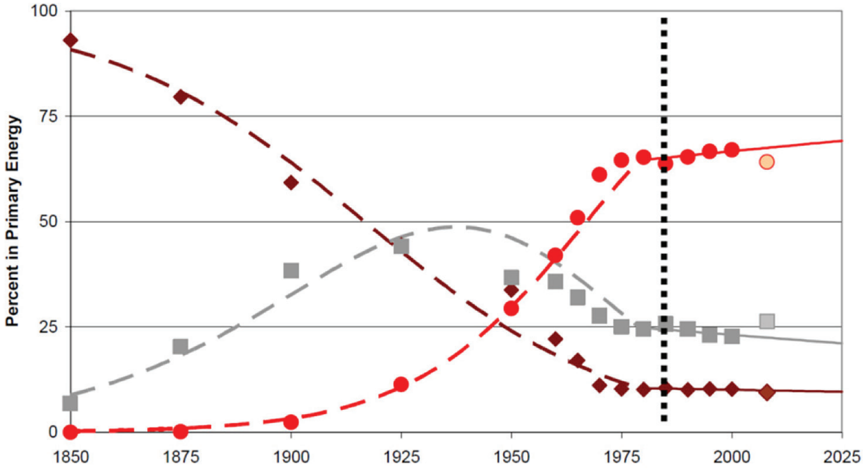


FIGURE 2 Major transitions in global energy systems measuring market shares in total primary energy use. Lines show traditional fuels such as wood, horses, water power (small diamonds), coal (squares), and modern energy sources such as oil, gas, nuclear, and modern renewables (circles). The market dynamics are approximated by a set of coupled logistic equations over the 1850–1975 period (dashed lines). The vertical dotted line denotes the end of the logistic fits and a significant stagnation of the evolution of the penetration of modern energy sources after the mid 1970s. Source: Grübler 2012. Reproduced with permission from the International Institute for Applied Systems Analysis (IIASA).

encourage experimentation, scale-up, and learning may be required over a period of decades to enable this transition.[6]

The sugarcane ethanol program started in 1975 in Brazil and the US shale gas program are both illustrative. Without judging whether the Brazilian government's plan was cost effective, its continuity and comprehensiveness (addressing yields, refineries, and vehicles—end-use technologies) contributed to ethanol's achievement of cost competitiveness with gasoline in 27 years, replacing 40 percent of all the gasoline that would be consumed in Brazil.[7] It took shale gas production

---

[6] While some policies creating markets for some of these technologies are in place today—e.g., federal production tax credits for wind, investment tax credits for solar, a renewable fuel standard for biofuels, vehicle fuel economy standards, loan guarantees for nuclear power, and renewable portfolio standards in 29 states—many studies have questioned both their cost effectiveness and their ability to serve as a guide for long-term investments by firms.

[7] Ethanol from sugarcane was already produced at small scales before World War I to stabilize sugar prices and later to address oil scarcity during both World Wars, but 1975 marks the year of the start of the sustained government push to develop the sugarcane ethanol industry. A recent study estimated the direct costs of the program between 1975 and 2000 at $42.5 billion in 2013 dollars (Meyer et al. 2012), and the benefits in terms of forgone oil imports evaluated at international prices between 1975

in the United States a similar amount of time, with stable R&D funding from the Gas Research Institute (an industry-government partnership) for 24 years, a tax credit for 12 years (MIT 2011), and the persistence of a visionary entrepreneur, George Mitchell.[8]

IMPLICATIONS FOR RESEARCH

It is important to consider that different technology pathways pose different challenges from a commercialization perspective. For example, even though commercial biofuels are available today,[9] their further expansion is not desirable because of competition with food, limited environmental benefits, and their true cost given subsidies. Alternative processes are at different stages of development and rely on different types of cellulosic biomass and algae (Figure 3 shows major biofuel production routes). However, none of these new processes is demonstrated at scale, partly as a result of high costs and the uncertainties surrounding existing regulations and physical infrastructure.

The development of biofuels that can more easily fit in with existing infrastructure (so-called drop-in fuels) should be an important factor driving research, without undermining research on alternatives that may require significant infrastructure changes but could in the long run result in significant cost reductions. Research should be underpinned by an analysis of the materials and energy involved, to focus on areas with the potential to be cost competitive in the long term. The possible impacts of different pathways are also contingent on crucial improvements in crop productivity and waste availability to reduce feedstock costs, expand the supply, and minimize other impacts, making this a particularly important research area.

POLICY NEEDS

Although there is growing government support for energy RD&D (Anadón 2012), research shows that even greater support for RD&D is needed in related

---

and 2002 at $64.9 billion (Goldemberg et al. 2004) (these are just the two most prominent benefits and costs).The program is also using 2.9 million hectares of land (about 29,000 km^2), about the surface area of Massachusetts. The complementary end-use technology—flex-fuel vehicles capable of running on gasoline or on a blend of up to 85 percent ethanol—now dominates the automobile market, accounting for 81 percent of light-duty vehicles in 2008.

[8] In the case of shale gas, changes required in the physical infrastructure and end-use technologies were less significant.

[9] Biofuels in the United States and abroad are largely produced from food crops. In 2012, 211 US ethanol plants produced 13.3 billion gallons of ethanol from corn through hydrolysis and fermentation and sold it mainly as E10 (gasoline with 10 percent of ethanol in volume) to meet the requirements of the Renewable Fuel Standard. Also in 2012, 114 US biodiesel plants produced almost 1 billion gallons of biodiesel, mainly from the transesterification of soybean oil, and sold mainly as B20 (diesel with 20 percent volume of biodiesel).

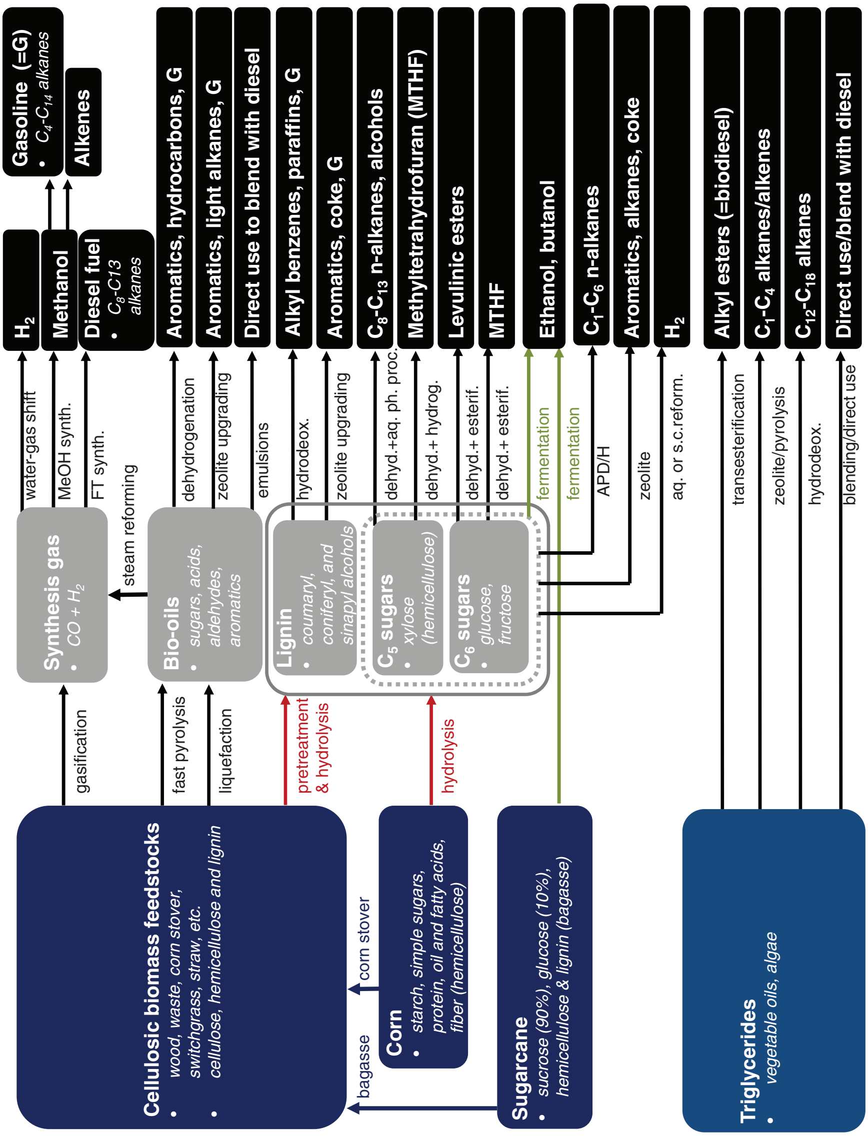


FIGURE 3 Routes to making different types of biofuels from different feedstocks (slightly adapted from Huber et al. 2006 and NSF 2008). Arrows: the fermentation steps are biological conversions, the pretreatment and hydrolysis processes refer to chemical and biological processes, and the rest of the arrows represent various chemical processes. Source: Adapted from NSF, 2008. APD/H = aqueous phase dehydration/hydrogenation; aq. = aqueous; C = carbon; FT = Fischer-Tropsch; hydrodeox. = hydrodeoxygenation; MeOH = methanol; ph. proc. = phase processing; s.c. = supercritical.

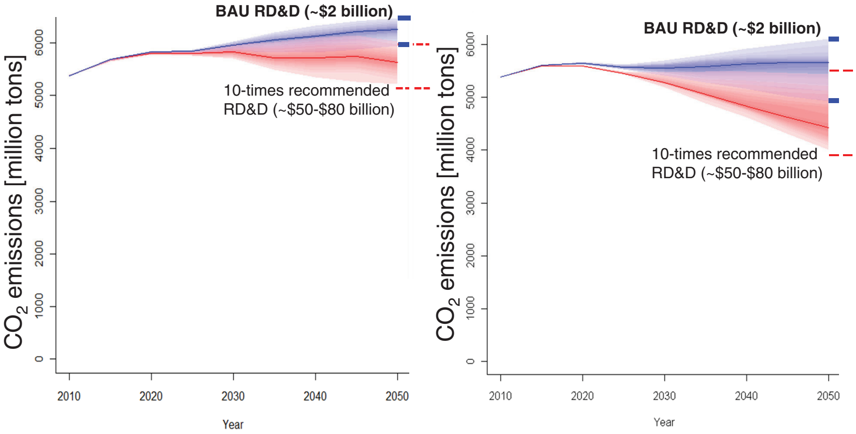


FIGURE 4 US energy-related CO_2 emissions 2010–2050, under business-as-usual federal energy RD&D investment with no additional demand-side policies (top band, 2050 range marked by thick lines) and 10 times the experts' average recommended federal energy RD&D investments (between $50 and $80 billion per year) (lower band, 2050 range of emissions marked by thin discontinuous lines), with no additional demand-side policies, using (left) middle-of-the-road and (right) optimistic experts' technology cost projections (Anadón et al., in press). Figure can be viewed in color at http://www.nap.edu/catalog.php?record_id=18558.

technology areas (Chan and Anadón, forthcoming). Research in solar power, biofuels, and utility-scale energy storage may result in the greatest returns on investment in terms of economic impact with a 2030 timeframe.[10] Experience also suggests that such RD&D needs a stable, long-term, and diverse set of research institutions working in close collaboration with industry (Anadón et al., forthcoming).

Yet even if the US government increased its federal RD&D investments in a wide range of technologies from about $2 billion a year to $80 billion a year between 2010 and 2030, CO_2 emissions from the energy sector are likely to remain far from the targets set by the IPCC (about 1–2 Gt/year) (Figure 4) (Anadón et al. forthcoming). Additional demand-side policies are very likely to be necessary to catalyze the transition from a modeling and historical perspective. A sufficiently high price on carbon (either through a tax or a cap-and-trade arrangement) is likely to result in the most efficient outcome, particularly because

---

[10] Anadón et al. (forthcoming) covered 25 technologies spanning solar photovoltaics, nuclear, bioenergy, carbon capture and storage, various vehicle technologies, and utility-scale energy storage and focused on incorporating uncertainty surrounding technical change.

there is uncertainty about the options that will be most successful (Anadón et al., forthcoming).[11]

REFERENCES

Anadón LD. 2012. Missions-oriented RD&D institutions in energy: A comparative analysis of China, the United Kingdom, and the United States. Research Policy 41(10):1742–1756.

Anadón LD, Bunn M, Narayanamurti V. In press. Transforming US Energy Innovation. Cambridge: Cambridge University Press. The November 2011 report forming the basis of this book is available at http://belfercenter.ksg.harvard.edu/publication/21528/.

Arthur BW. 1989. Competing technologies, increasing returns and lock-in by historical events. Economic Journal 99(294):116–131.

Chan G, Anadón LD. Forthcoming. Utilizing expert assessment to inform allocating government energy RD&D investment portfolios.

EIA (US Energy Information Administration). 2012. Annual Energy Review 2011. Washington DC: Department of Energy. Available at www.eia.gov/totalenergy/data/annual/pdf/aer.pdf.

EPA (US Environmental Protection Agency). 2013. Inventory of US Greenhouse Gas Emissions and Sinks: 1990–2011. EPA 430-R-13-001. Washington DC. Available at www.epa.gov/climatechange/Downloads/ghgemissions/US-GHG-Inventory-2013-Main-Text.pdf.

Goldemberg J, Teizeira Coelho S, Nastari PN, Lucon O. 2004. Ethanol learning curve: The Brazilian experience. Biomass and Bioenergy 26:301–304.

Grübler A. 2012. Grand Designs: Historical Patterns and Future Scenarios of Energy Technological Change. In: The Global Energy Assessment: Toward a Sustainable Future, Chapter 24: Policies for the Energy Technology Innovation System—Historical Case Studies of Energy Technology Innovation. Cambridge: Cambridge University Press and International Institute for Applied Systems Analysis.

Huber GW, Iborra S, Corma A. 2006. Synthesis of transportation fuels from biomass: Chemistry, catalysts, and engineering. Chemical Reviews 106:4044–4098.

IEA. 2013. World Energy Outlook. International Energy Agency, Organization for Economic Cooperation and Development. Paris, France.

IPCC (Intergovernmental Panel on Climate Change). 2007. Climate Change 2007: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Fourth Assessment Report of the IPCC. Cambridge: Cambridge University Press.

Lim S.S., Vos T., Flaxman AD, Danaei G, et al. 2012. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: A systematic analysis for the Global Burden of Disease Study 2010. The Lancet. 380(9859):2224–2260.

Meyer D, Mytelka L, Press R, Dall'Oglio EL, de Sousa PT Jr, Grübler A. 2012. Brazilian Ethanol: Unpacking a Success Story of Energy Technology Innovation. In: The Global Energy Assessment: Toward a Sustainable Future, Chapter 24: Policies for the Energy Technology Innovation System—Historical Case Studies of Energy Technology Innovation. Cambridge: Cambridge University Press and International Institute for Applied Systems Analysis.

MIT (Massachusetts Institute of Technology). 2011. The Future of Natural Gas, Appendix 8A: Natural Gas RD&D Background. The MIT Energy Initiative. Available at http://mitei.mit.edu/system/files/NaturalGas_Appendix8A.pdf.

---

[11] The aggressiveness and existence of waivers in the Renewable Fuel Standard reduce its ability to promote innovation and fail to encourage fuel efficiency, which can be achieved by fuel or carbon taxes.

Narayanamurti V, Odumosu T, Vinsel L. 2013. RIP: The basic/applied research dichotomy. Issues in Science and Technology XXIX.2 (Winter).

NRC (National Research Council). 2010. Hidden Costs of Energy: Unpriced Consequences of Energy Production and Use. Washington DC: National Academies Press. Available at www.nap.edu/catalog.php?record_id=12794.

NSF (National Science Foundation). 2008. Breaking the Chemical and Engineering Barriers to Lignocellulosic Biofuels: Next Generation Hydrocarbon Biorefineries. Ed. George W. Huber, University of Massachusetts Amherst. National Science Foundation. Chemical, Bioengineering, Environmental, and Transport Systems Division. Washington DC. Available at www.ecs.umass.edu/biofuels/Images/Roadmap2-08.pdf.

Bioenergy Technologies and Strategies: A New Frontier

JOYCE C. YANG
Department of Energy

> *O beautiful for spacious skies, for amber waves of grain,*
> *for purple mountain majesties, above the fruited plain!*
> – Katharine Lee Bates (1904)

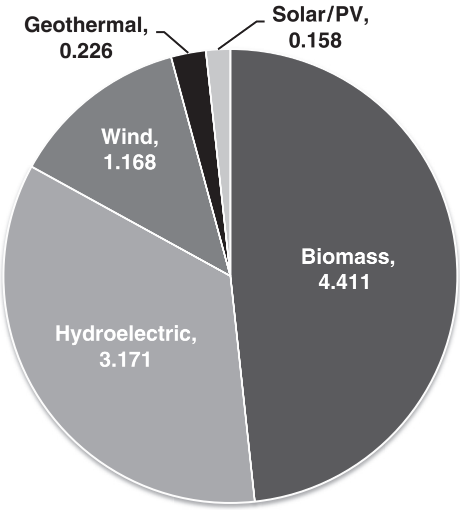
The bounties of American ingenuity, climate, and soil not only inspire the opening verse of a patriotic song but also establish the United States as the world leader in agriculture (USDA 2013) and forestry[1] productivity. Thus it is not surprising that researchers, engineers, industrialists, and policymakers have turned to the nation's abundant biomass resources to reduce consumption of fossil fuel, be it coal, natural gas, or petroleum. In fact, of all forms of renewable energy consumed in the United States, none rivals that produced from biomass (Figure 1). Furthermore, beyond the total illustrated in Figure 1, a recent report estimates an additional renewable resource of 1 billion dry tons of agricultural residues, woody biomass, and new energy crops that can be sustainably harvested every year (DOE 2011).

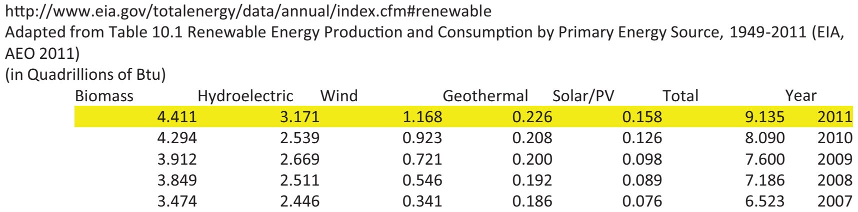
BACKGROUND

Terrestrial biomass feedstocks are typically composed of three major types of polymers: cellulose (homogeneous polymer composed of six-carbon sugars, or C_6s), hemicellulose (heterogeneous polymer but predominantly composed of five-carbon sugars, or C_5s), and lignin (heterogeneous polymer composed of a

---

[1] According to the United Nations Food and Agriculture Organization forest products statistics for 2011, available at www.fao.org/forestry/statistics/80938@180723/en/ (accessed November 25, 2013).




http://www.eia.gov/totalenergy/data/annual/index.cfm#renewable
Adapted from Table 10.1 Renewable Energy Production and Consumption by Primary Energy Source, 1949-2011 (EIA, AEO 2011)
(in Quadrillions of Btu)

Biomass	Hydroelectric	Wind	Geothermal	Solar/PV	Total	Year
4.411	3.171	1.168	0.226	0.158	9.135	2011
4.294	2.539	0.923	0.208	0.126	8.090	2010
3.912	2.669	0.721	0.200	0.098	7.600	2009
3.849	2.511	0.546	0.192	0.089	7.186	2008
3.474	2.446	0.341	0.186	0.076	6.523	2007

FIGURE 1 US renewable energy consumption in 2011 by primary energy source, in quadrillion British thermal units (Btus). Total US renewable energy consumed exceeded 9 quadrillion Btus. PV = photovoltaic. Source: EIA 2011.

significant component of aromatic molecular units). Biofuels derived from terrestrial feedstocks are often referred to as “cellulosic” because of their principal biomass component. In contrast, “conventional” biofuels are grain-based (e.g., corn ethanol) and may compete with food and feed markets. Aquatic biomass, such as algae and cyanobacteria, can be a mixture of C_5s and C_6s polysaccharides along with other classes of biopolymers such as proteins and lipids.

Biomass is usually transformed into biofuel through one of two types of processing: biochemical or thermochemical. In biochemical processing, biomass is typically pretreated with mechanical, chemical, and/or thermal forces to open up the plant cell wall and structure, thus exposing the partially depolymerized material to microbial enzymes (cellulases and hemicellulases) that attack the chemical bonds to yield monosaccharides. These dilute sugar intermediates are usually fed to a microbe to produce fuels or more refined chemicals.

In thermochemical processing the biomass is typically mechanically preprocessed to specific sizes, inorganic contents, and moisture levels, and then subjected to moderate to high pressures and temperatures (with or without catalysts) to generate syngas or bio-oil intermediates. These process intermediates are cleaned or stabilized and then exposed to fuel synthesis catalysts to either regenerate the bonds between the C_1 units into longer-chained hydrocarbons or hydrocrack larger biomass thermal derivatives to generate fuel blendstocks.

The mission of the US Department of Energy (DOE) Bioenergy Technologies Office (BETO) is to transform available domestic biomass resources into fuels, chemicals, and power. It achieves its mission through a diverse and comprehensive set of applied research and development (R&D) programs and first-of-a-kind technology demonstrations called *integrated biorefineries* (IBRs). The BETO strategy is to reduce the risk of biofuel technologies by demonstrating feasibility, process robustness, process control, and scalability to attract private capital for commercialization and market entry. BETO partners are encouraged to use biomass feedstocks that do not compete with food or feed uses, and to develop a suite of versatile conversion technologies that can be deployed in as many regions of the United States as possible to maximize both national and regional benefits.

BETO currently focuses on technologies that seek to use cellulosic or algal biomass feedstocks because of more favorable environmental benefits as demonstrated by a life cycle analysis of greenhouse gas (GHG) emissions and lower water consumption (Wang et al. 2011; Wu et al. 2009). In fact, to qualify as a cellulosic biofuel for incentives, a 60 percent GHG reduction must be achieved relative to gasoline.

RECENT PROGRESS

The US Department of Energy (DOE) announced the completion of several major R&D programs on cellulosic ethanol at the end of 2012. Achievements on both the biochemical and gasification routes to cellulosic ethanol corresponded with a dramatic reduction in the modeled minimum ethanol selling price from more than \$9/gallon, when the program began in 2002, to \$2.15/gallon or less in 2012. The many technical performance improvements include better feedstock quality and logistics, pretreatment technologies, more productive cellulolytic enzymes, gas cleanup technologies, and the development of robust microbial and inorganic fuel synthesis catalysts, not to mention a wealth of enabling knowledge gains and breakthroughs contributed by grantees of the DOE Office of Science, National Science Foundation, National Institute of Standards and Technology, and US Department of Agriculture.

Concurrent with the R&D achievements that helped reduce key biofuel cost factors, four first-of-a-kind IBRs for cellulosic ethanol were established in the United States and either began producing fuel or will begin to produce it in 2014, and one facility has begun production of cellulosic hydrocarbon fuels (Table 1).

TABLE 1 Commercial-scale US integrated biorefineries, constructed or under construction, focused on cellulosic biofuels.

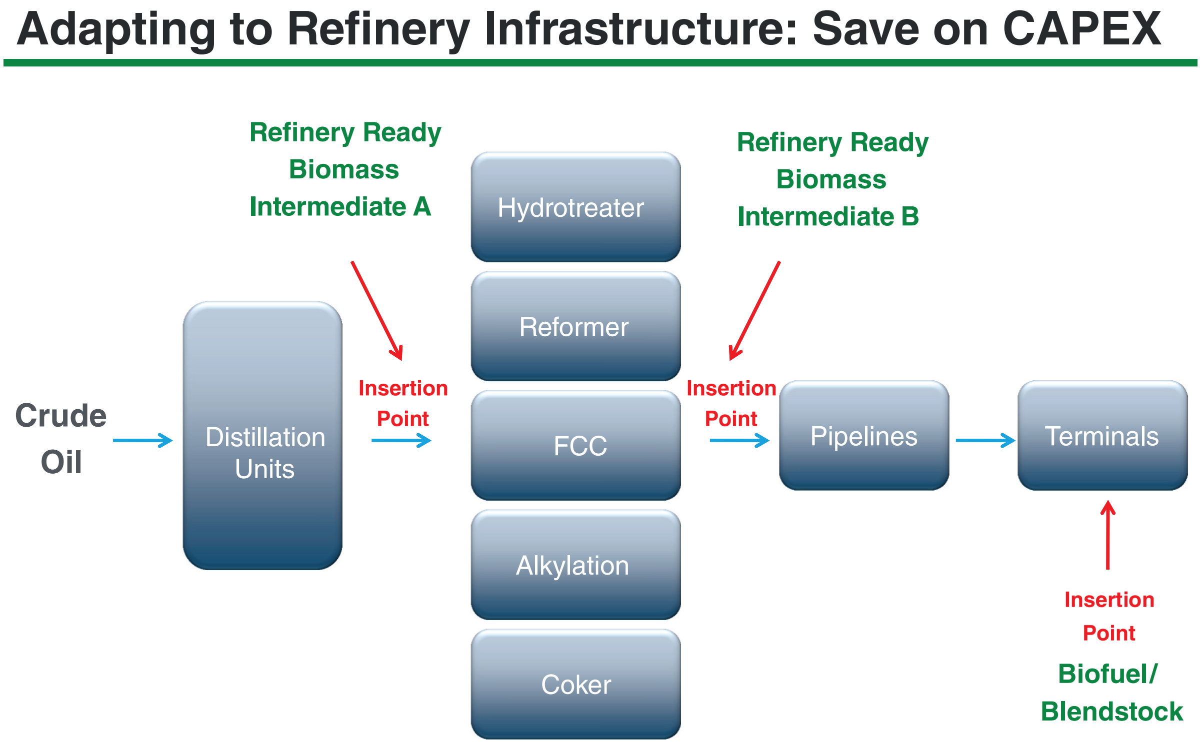
Company	Start of construction	Feedstock	Target product	Process route	Location	DOE role
DuPont	2012Q4	Ag residue	Cellulosic ethanol	Biochemical	Nevada, IA	R&D
POET-DSM	2012Q1	Ag residue	Cellulosic ethanol	Biochemical	Emmetsburg, IA	R&D, IBR
Abengoa	2011Q4	Ag residue	Cellulosic ethanol	Biochemical	Hugoton, KS	IBR
KiOR	2011Q2	Southern pine	Cellulosic gasoline, diesel, and jet	Thermo-chemical	Columbus, MS	None
INEOS-Bio	2011Q1	MSW, citrus waste, yard waste, woody biomass	Cellulosic ethanol	Hybrid	Vero Beach, FL	R&D, IBR

NOTE: IBR = integrated biorefinery; MSW = municipal solid waste; Q = quarter.

These IBRs represent far more than their technological components: each is the result of successful process integration, scale-up, and construction as well as critical success elements such as feedstock contracts, project management, fuel off-take agreements, seasoned senior management, regulatory clearance, and financing. (Financing these biorefineries has been particularly challenging because the economics are as yet unproven.)

MOVING FORWARD

Although ethanol can displace the gasoline used for light-duty passenger cars, it cannot be blended with other transportation fuels. One particularly interesting variation of the hydrocarbon fuel strategy is to produce an “intermediate” that can be inserted at various processing units in traditional petroleum refineries (Figure 2). The key advantage of this strategy is that several units of operations might be avoided by leveraging existing assets of the petroleum refinery, thus significantly lowering capital costs. There is also a fuel distribution advantage with the biomass-derived blendstock strategy. Accordingly, BETO began in 2010 to shift away from a singular focus on cellulosic ethanol to embrace a more holistic biofuels strategy aimed at replacing an entire barrel of oil by targeting



Adapted from the National Advanced Biofuels Consortium Website

FIGURE 2 Proposed insertion points of biomass-derived fuel intermediates into existing petroleum refinery units of operation. Source: Adapted from the National Advanced Biofuels Consortium website, www.nabcprojects.org/biofuels.html; accessed November 25, 2013.

the production of hydrocarbon, or "drop-in," fuels that are compatible with the current infrastructure.

Stoichiometry presents a major technical challenge for hydrocarbon biofuels: biomass is a relatively oxygen-rich carbon feedstock, whereas hydrocarbons lack oxygen. When the target molecule was ethanol, biomass was an advantaged feedstock compared to petroleum, in terms of basic stoichiometry; but when the target molecule is a longer carbon chain with no oxygen, a biomass feedstock is disadvantaged. This basic chemical balancing act, illustrated in Table 2, will be the key challenge to progress, requiring innovations across the biomass-to-biofuel supply chain.

The removal of oxygen in the biomass fuel intermediate will be essential for compatibility with existing crude oil processing streams; but it will also mean a significant loss of yield in the form of either water (requiring a hydrogen source) or carbon monoxide or dioxide (resulting in greater loss of yield from the original biomass). Hydrogen can be derived from natural gas using processes like methane reforming; however, the impact on the GHG reduction for this option should be considered. On the other hand, losses of carbon as CO_2 are also unpalatable and negatively impact the GHG profile.

At least one partial solution is to diversify the product slate. If hydrocarbon fuels cannot contain oxygen molecules, then it is possible that a marketable coproduct that is "oxygen rich" (defined here as having a carbon:oxygen, or C:O, ratio less than 1) can be made alongside the fuel. It is also likely that such a coproduct could enhance the economics of the overall conversion process. The Department of Energy has identified several such value-added chemicals, including sorbitol, xylitol, aspartic acid, and diacids (Holladay et al. 2007; Werpy et al. 2004).

TABLE 2 Stoichiometric relationship between biomass and petroleum feedstocks and different fuel products.

Relative Elemental Distribution by Mass (AFDW Ultimate Analysis)	Fuel Products			Feedstocks	
	Gasoline	Diesel	Ethanol	Petroleum	Biomass
C	0.86	0.84	0.52	0.84	0.50
H	0.13	0.13	0.13	0.11	0.06
S	0.01	0.04	0.00	0.05	0.00
N	0.00	0.00	0.00	0.00	0.00
O	0.00	0.00	0.35	0.00	0.44
Total	1.00	1.00	1.00	1.00	1.00

NOTE: AFDW = ash free dry weight; C = carbon; H = hydrogen; S = sulfur; N = nitrogen; O = oxygen.
Source: Adapted from Morvay and Gvozdenac 2008.

The imbalance in C:O ratio in feedstock and product also requires ever more efficient use of the biomass resource itself. Losses that can occur either under open storage systems (e.g., bale yards or wood laydown yards) to support year-round cellulosic biorefinery operations or from natural disasters (e.g., droughts, flooding) will mean even greater operation expense losses to the hydrocarbon biorefinery versus the ethanol biorefinery. Commoditizing biomass feedstocks can be an effective mitigation strategy; a version of this advanced feedstock concept has been proposed by Idaho National Laboratory (Figure 3) (Hess et al. 2009).

A key aspect of commodity-based biomass feedstocks is that different feedstocks can be collected and transported to regionally distributed depots or terminals where they undergo preprocessing and are blended to predefined physio-

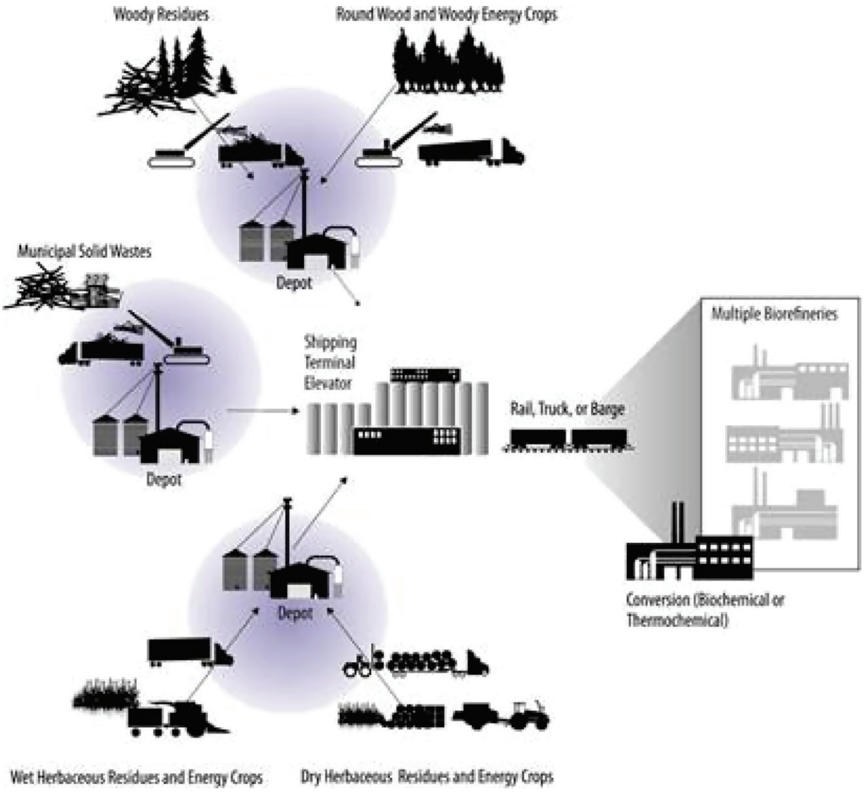


FIGURE 3 Commodity-based biomass feedstock supply logistics system. This design follows the model of the current commodity grain supply system, which manages crop diversity at the point of harvest and/or the storage elevator and thus allows all subsequent feedstock supply system infrastructure to be similar for all biomass resources. Source: Image courtesy of Idaho National Laboratory; Hess et al. 2009.

chemical specifications and then densified to facilitate storage and handling. Although this method will increase costs because of the additional processing, it is modeled after the existing agricultural grain commodity system, raising the interesting possibility of leveraging the grain distribution network as another infrastructure cost reduction opportunity. This advanced system would also take advantage of a variety of technologies while at the same time continuing to use conventional agriculture and forestry equipment where possible.

An intriguing possible alternative is to create an advanced biomass feedstock or feedstock component that changes the overall C:O ratio in vivo to favor hydrocarbon fuel formation. A study published in 2007 suggested that natural plant and microbial oils, such as algal lipids, can be readily converted into hydrocarbon fuels or blendstocks using existing petroleum refinery units (Huber and Corma 2007). Current work on algae suggests that it could soon exceed palm oil (the best terrestrial oilseed crop) (Davis et al. 2012), but the associated cultivation and processing costs result in a fuel product that exceeds $18 per gallon.

The relative advantages of using modified biological feedstocks as a means to achieve refinery-ready intermediates versus other approaches will need to be carefully evaluated in terms of both theoretical yields and practical considerations.

CONCLUSION

Over the past two decades the United States has consistently pursued strategy that involves simultaneously funding research development, demonstrating biofuels technologies, and establishing favorable national policies to incentivize biofuels production to reduce dependence on fossil fuels. Through policies such as the Energy Independence and Security Act of 2007, the Energy Policy Act of 2005, and the Biomass R&D Act of 2000 (Title III), the government has supported innovators across the supply chain, culminating in the first US commercial production of cellulosic ethanol in 2013.

The United States is positioned to benefit not only from an abundance of renewable biomass but also from the use of existing infrastructures across the country. The new frontier of biofuels RD&D will no doubt be full of significant challenges, but scientific and engineering innovators can overcome them by building on a solid foundation of knowledge and leveraging advances already realized in first- and second-generation biofuels.

REFERENCES

Davis R, Fishman D, Frank ED, Wigmosta MS. 2012. Renewable Diesel from Algal Lipids: An Integrated Baseline for Cost, Emissions, and Resource Potential from a Harmonized Model. Golden CO: National Renewable Energy Laboratory.

DOE (US Department of Energy). 2011. US Billion-Ton Update: Biomass Supply for a Bioenergy and Bioproducts Industry. Oak Ridge TN: Oak Ridge National Laboratory.

EIA (Energy Information Administration). 2011. Annual Energy Outlook. Washington DC: US Department of Energy.

Hess JR, Kenney KL, Ovard L, Searcy EM, Wright CT. 2009. Uniform-format Bioenergy Feedstock Supply System Design Report: Commodity-scale Production of an Infrastructure Compatible Bulk Solid from Herbaceous Lignocellulosic Biomass. Idaho Falls: Idaho National Laboratory.

Holladay JE, White JF, Bozell JJ, Johnson D. 2007. Top Value-Added Chemicals from Biomass: II. Results of Screening for Potential Candidates from Biorefinery Lignin. Richland WA: Pacific Northwest National Laboratory.

Huber GW, Corma A. 2007. Synergies between bio- and oil refineries for the production of fuels from biomass. Angewandte Chemie International Edition 46(38):7184–7201.

Morvay Z, Gvozdenac D. 2008. Applied Industrial Energy and Environmental Management. West Sussex UK: John Wiley & Sons, Ltd.

USDA (US Department of Agriculture). 2013. World Agricultural Production. Foreign Agricultural Service, WAP 11-13. Washington DC. Available at www.fas.usda.gov/wap/current/default.asp.

Wang MQ, Han J, Haq Z, Tyner WE, Wu M, Elgowainy A. 2011. Energy and greenhouse gas emission effects of corn and cellulosic ethanol with technology improvements and land use changes. Biomass and Bioenergy 35(5):1885–1896.

Werpy TA, Holladay JE, White JF. 2004. Top Value Added Chemicals from Biomass: I. Results of Screening for Potential Candidates from Sugars and Synthesis Gas. PNNL-14808. Richland WA: Pacific Northwest National Laboratory.

Wu M, Mintz M, Wang M, Arora S. 2009. Water consumption in the production of ethanol and petroleum gasoline. Environmental Management 44(5):981–997.

Artificial Solar Fuel Generators

Miguel A. Modestino
École Polytechnique Fédérale de Lausanne

Rachel A. Segalman
University of California, Berkeley
Lawrence Berkeley National Laboratories

There has been significant interest in increasing the share of renewable energy sources in the world energy landscape (Chu and Majumdar 2012). Associated technologies support the generation or capture of energy from carbon-neutral sources, storage so that the energy can be used when and where needed, and more efficient use. In discussions of alternatives available for power generation from renewable sources, solar energy conversion is prominent, given the vast amount of energy it can yield (peak irradiation of 7.5 kWh/m^2/day, mean annual global irradiation of 8,372 terawatt hours [TWh]/yr), making it a potentially important candidate for a sizable portion of US energy.

This potential contrasts with its current relatively small portion of the global energy portfolio: 0.06 percent (IEA 2012; Zhang and Shen 2012). While economic factors account for a substantial part of the barrier to implementation, considerable technological challenges stem from the inherently intermittent nature of the solar generation process. Adoption of solar power generation will entail significant changes in operation of the power grid, as classical power generation plants will need to respond not only to changes in consumer demand but also to noncontrollable variations in energy generation.

BACKGROUND

One option to mitigate the intermittency of solar energy generation is the incorporation of energy storage capacity into the grid, so that fluctuations in energy generation are buffered and do not affect the operation of the electricity distribution channels. But large-scale implementation of energy storage faces both technological and economic hurdles requiring significant research and development.

Alternatively, one could take inspiration from nature, where energy is stored in the form of chemical bonds. In the case of artificial photosynthesis, this means the generation of fuels directly from solar energy.

Types of Solar Fuel Generators

Integrated energy capture and storage solutions such as solar fuel generators have the potential to increase the fraction of renewables in the mix of energy sources, and can apply to all sectors of energy consumption (industrial, commercial, residential, and transportation) (Bard and Fox 1995; Chu and Majumdar 2012; Concepcion et al. 2012; Faunce et al. 2013; Lewis and Nocera 2006; Nocera 2012). Integrated solar fuel generators are photoelectrochemical (PEC) cells that can capture solar energy and catalytically convert low-energy reactants into energy-dense fuels.

One category of solar fuel generators, water-splitting systems, take water as a feed and produce hydrogen fuel and oxygen as byproduct. A general representation of these systems is shown in Figure 1. Practical systems take water and solar energy as inputs and produce output streams of hydrogen and oxygen in a safe and scalable manner. In this way pure fuel streams can be collected and used in electrochemical energy conversion devices (i.e., fuel cells) or in chemical processes to synthesize or enhance the energy content of liquid fuels (e.g., the Fischer-Tropsch process).

The concept of solar fuel generation can be extended to the electrochemical reduction of CO_2, which can yield carbon-containing fuels but represents a closed cycle from the carbon perspective, making these new fuels truly carbon neutral. Solar-driven CO_2 reduction poses greater technical challenges because the number of electron transfer steps in the reactions is higher, the concentration of CO_2 in electrolytes is generally low, and the diversity of products makes the necessary separation more difficult.

Mechanics of Solar Fuel Generators

As shown in Figure 1, solar fuel generation begins with the absorption of light to form charges that are used to drive oxidation and reduction reactions. These three processes can be done in separate units—for example, using a photovoltaic cell to generate electricity that powers an electrolyzer, which incorporates the catalysts—or in a fully integrated device. Practical comparisons between these two scenarios are largely dependent on possible gains in terms of economics and flexibility of deployment.

A fully integrated water-splitting or solar hydrogen generator, as shown in Figure 1, would consist of interconnected photovoltaic and catalytic units. Ideally, the oxidation and reduction sites are physically separated so that the product (H_2 in the case of water splitting) is generated in a space different from the byproduct

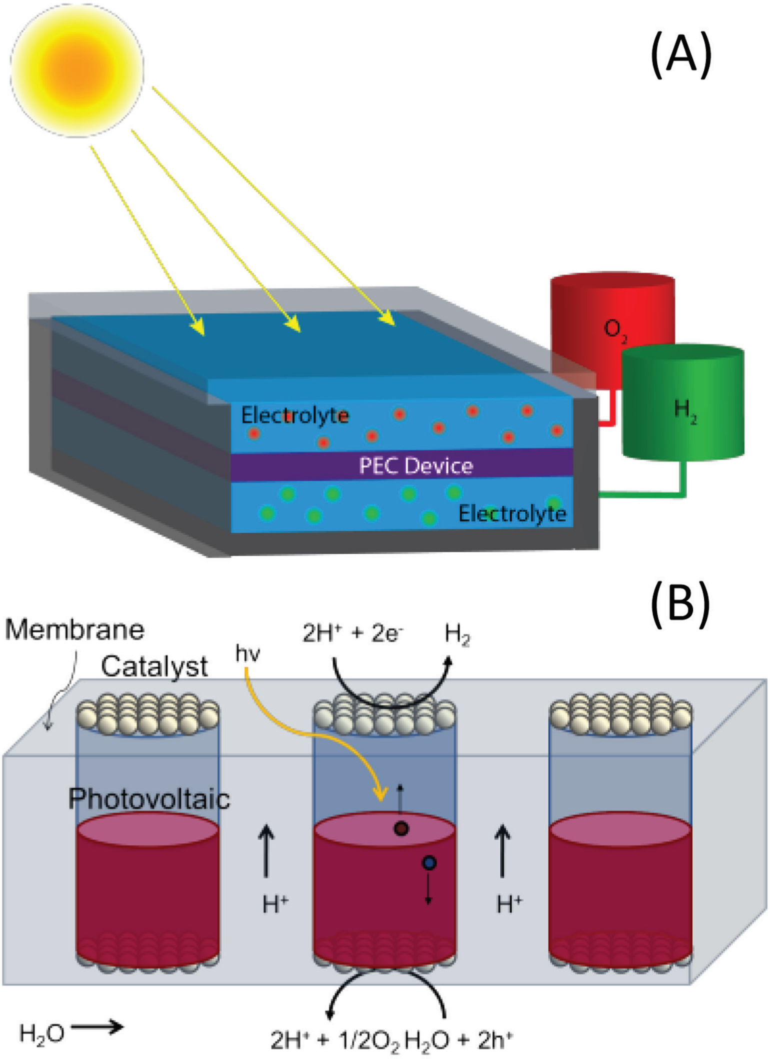


FIGURE 1 (A) Solar fuel generators are composed of photoelectrochemical (PEC) devices that can generate separate streams of fuels directly from sunlight (H_2 and O_2 in the case of water splitting). (B) A PEC device contains photovoltaic units that absorb light and move charges to catalytic centers, where electrochemical hydrogen and oxygen evolution reactions take place. In the diagram, the membrane component is used as the matrix for PEC units and allows for both ion conduction (i.e., H^+) and gas separation.

(O_2 in this case). The size of the photovoltaic unit is generally set by the solar absorption depth of the material and is on the order of microns to millimeters.

For a number of reasons it makes sense to have an array of photovoltaic units held together by a mechanically robust membrane that separates the product gases and shuttles ions from one catalyst site to the other. Under acidic conditions, water will be dissociated into O_2 and protons on the oxidation side of the membrane. The protons will then be transported through the membrane to the reduction side, where H_2 will be evolved. In this way oxidation and reduction products will be generated in separate regions of the membrane, preventing the need for further separation.

In the case of operation under basic electrolytes, the processes are analogous and the ionic current in the system is carried by OH^- ions at steady state. The incorporation of ion-conducting membranes is crucial for this type of operation, as they provide transport pathways for charged intermediaries between the oxidation and reduction sites and at the same time serve as a barrier for gas diffusion, allowing the production of fuel in its pure form. Achieving this configuration can be simple for macroscopic units, but for micrometer- to nanometer-scale (i.e., mesoscale) systems significant advances are required in terms of both membrane and PEC unit self-assembly.

Progress on these technological options depends on research and development currently under way in academic, government, and industrial laboratories. In this article, we discuss aspects of an integrated system that are the focus of current exploration and development. The sections below touch on some of the advances and challenges in achieving practical solar fuel generators, implications for mesoscale assemblies and for membranes used in these systems, and overall system design considerations. Throughout, we describe current research as well as specific areas that require further study to enable progress in this area.

SOLAR FUEL GENERATION SYSTEMS

Since the first demonstration of solar-driven water splitting by Fujishima and Honda (1972), the prospect of using PEC cells for solar fuel generation has motivated the quest for components and integrated systems that can continuously and robustly produce hydrogen fuels directly from sunlight. Over the past 40 years many studies have attempted to tackle parts of the problem, and fuel-generating systems have reached solar hydrogen generation efficiencies of up to 18 percent (Peharz et al. 2007).

But solar hydrogen generation units fall short in satisfying stability and cost-effectiveness requirements. Some high-efficiency systems rely on III-V multijunction photovoltaic components that have prohibitively high costs and serious photocorrosion challenges at the interface between the semiconductor and the electrolyte (Khaselev and Turner 1998; Khaselev et al. 2001; Peharz et al. 2007).

Other systems, based on silicon light-absorbing components, including earth-abundant catalysts, face significant stability problems when operated under basic or acidic electrolytes. Recently, however, Nocera's group at the Massachusetts Institute of Technology demonstrated integrated systems that incorporate earth-abundant components that can stably operate under buffered electrolytes at moderate pH (Reece et al. 2011). This promising demonstration can open avenues for the implementation of cost-effective solar hydrogen generators, but important challenges for the management of ion and mass transport remain, largely based on the need to separate the gaseous products while providing pathways for steady-state ion conduction (Haussener et al. 2012; Hernandez-Pagan et al. 2012).

When systems are operated at moderate pH regimes, the low concentration of proton or hydroxide conduction results in high solution resistance for these ions, and most of the ionic current is carried by supporting ions present in the solution (i.e., ions dissociated from buffer molecules). Under these circumstances, as the conducting ions are not part of the electrode reactions, concentration gradients will evolve and the overall system will not be able to operate continuously.

Efficient solar hydrogen generation would represent a large step to increase the share of renewable fuel sources but implementation would be challenging as current infrastructure is based on liquid carbon–based fuels.

An alternative to solar water splitting lies in the direct reduction of CO_2 for the generation of liquid carbon–containing fuels (Gattrell et al. 2007; Kondratenko et al. 2013; Lewis and Nocera 2006; Olah et al. 2008). Notwithstanding considerable research in this field, challenges persist because requirements for catalyst selectivity, CO_2 absorption, and product separation are quite stringent.

Last, the technoeconomic aspects of solar fuel generators are crucial for the achievement of deployable systems. The US Department of Energy has set the price of hydrogen produced at less than \$4/kg, which imposes bounds on the material systems and configurations that are implementable (Saur and Ainscough 2011). Few reports have tackled these aspects or provided guidance to achieve this price point (e.g., James et al. 2009; Pinaud et al. 2013).

As both the scientific and engineering aspects of artificial photosynthesis devices mature, a better understanding of the challenges to fabricate cost-effective solar fuel generators will be critical for their deployment.

MESOSCALE BUILDING BLOCKS FOR ARTIFICIAL PHOTOSYNTHESIS SYSTEMS

The examples cited above represent initial attempts at developing integrated devices that can produce hydrogen fuels directly from the sun, and they all rely on macroscopic PEC units arranged such that ion transport involves a liquid electrolyte. Under concentrated electrolyte conditions (~1 M), ion transport does not provide significant resistance if the ionic pathway is less than a few centimeters

(Haussener et al. 2012). Furthermore, if the ionic conductivity of electrolyte is lowered, or for operation of systems under water vapor (Spurgeon and Lewis 2011), it is highly desirable to develop PEC units with dimensions in the micro- or nanometer range so that ions have to migrate only small distances.

Several mesoscale building blocks for PEC units have been developed. Complex nanocrystal structures (e.g., nanorods, nanowires) can be synthesized in solution (Amirav and Alivisatos 2010; Dukovic et al. 2008; Sun et al. 2011, 2013) and have shown promising performance in terms of hydrogen evolution. Methods for arranging these nanostructures into architectures that enable oxidation and reduction reactions to occur at separate locations depend on the shape, dimensions, and self-assembly characteristics of the particles.

For long semiconducting nanowire systems, large surface area mats permit a percolated network of wires to act as a self-standing water-splitting membrane (Sun et al. 2011). And for nanorod-based systems, self-assembly techniques are required to achieve architectures resembling that shown in Figure 1 (Baker et al. 2010; Baranov et al. 2010; Gupta et al. 2006; Ryan et al. 2006).

Although these self-assembly techniques have demonstrated the fabrication of large-scale vertically aligned nanorod arrays from solution, it is not clear how to obtain preferential directionality of the ends of asymmetric water-splitting nanorods (Amirav and Alivisatos 2010).

As an alternative to solution-based methods, photocatalytic units can be directly grown via vapor-liquid-solid deposition methods so that the resulting arrays have the desired directionality. The development of silicon-based microwire arrays is an example of such a strategy and can lead to large-area coverage of the photoactive components that can then be incorporated into ion-conducting membranes (Boettcher et al. 2010; Maiolo et al. 2007; Plass et al. 2009; Spurgeon et al. 2011). These systems have many advantages over planar PEC devices because they can absorb nearly all the incident light with only a small fraction of areal coverage (Kelzenberg et al. 2010) and each microwire in the arrays acts as an independent unit, largely alleviating stability constraints.

The incorporation of mesoscale PEC units into fully functional solar hydrogen generators represents a promising alternative to overcome the technological challenges that prevent deployment, and so a great deal of research is being conducted in this area.

MEMBRANE MATERIALS FOR ARTIFICIAL PHOTOSYNTHESIS

Membranes in solar hydrogen generators serve two basic functions: to provide pathways for ion conduction and to keep gaseous products separated (as shown in Figure 2). Ion-conducting membranes, which have been investigated for several decades, are important components not only in artificial photosynthesis applications but also in a variety of energy conversion devices (Walter et al. 2010; Zhang and Shen 2012). The fundamental similarities in membrane requirements

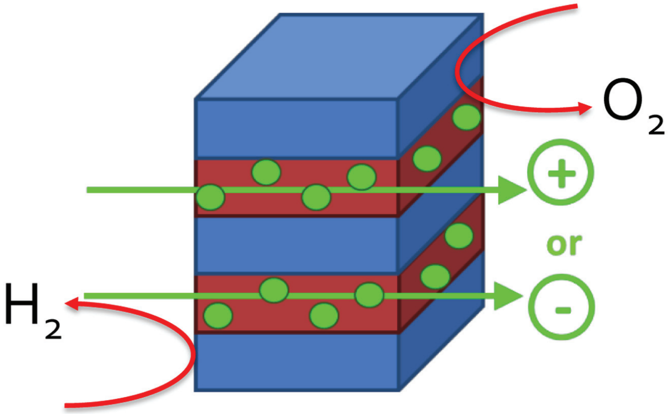


FIGURE 2 Diagram of membrane material used for solar water splitting. These materials contain conductive domains capable of transporting ions across the membrane (positive or negative), while preventing crossover of the gases produced.

between solar fuel devices, hydrogen fuel cells, and electrolyzers suggest an existing set of candidate materials.

In the case of artificial photosynthesis applications, the operating current density is dictated by the solar absorption rate and is relatively low when compared to the requirements for other similar devices, but very sensitive to crossover due to the relatively small quantity of product. Moreover, the presence of a large number of interfaces between the polymer and inorganic PEC components can severely affect the structure and transport properties of common nanostructured fuel cell membrane materials. Perfluorosulfonic acid (PFSA) ionomer membranes (e.g., Nafion®) are the most prominent alternatives for proton conduction, given their high ionic conductivity and remarkable chemical and structural stability.

With artificial photosynthesis membranes, high levels of conductivity are not required and the emphasis should instead be on the balance between the ionic and gas transport properties of materials. The development of ion-conducting block copolymers (BCPs) represents a promising route to decouple these two properties, as different blocks can be designed to provide complementary structural and gas barrier properties as well as ionic conductivity.

Furthermore, properties of BCP systems can be easily tuned and optimized by altering the molecular weight and volume fraction of each phase (Peckham and Holdcroft 2010). BCP membranes based on blends with ionic liquids (ILs) and polymerized ILs (PILs) are characterized by good ionic conductivity and tenability (Bara et al. 2008, 2009; Gu and Lodge 2011; Gwee et al. 2010; Hoarfrost and Segalman 2011, 2012; Lu et al. 2009; Mecerreyes 2011; Simone and Lodge 2009). Recent work has demonstrated the potential of PIL BCP materials for tuning transport properties in membranes used for solar fuel applications (Schneider et al. 2013; Sudre et al. 2013).

SYSTEM DESIGN CONSIDERATIONS

All the components of a solar fuel generator system need to operate stably and perform efficiently under the same conditions (i.e., temperature, electrolyte selection). Additionally, the photovoltage generated by the light-absorbing units needs to be sufficient to support the water-splitting reaction (1.23 V), the catalyst overpotential requirement, the ohmic drop associated with transporting both electrons and ions across the device, and any additional overpotential that may arise from chemical potential differences (i.e., concentration overpotential). Furthermore, all the transport processes in the system need to occur in parallel so that the electronic current matches the ionic current across reaction sites.

Several electrochemical modeling studies provide some guidance on the optimal arrangements and dimensions of each of the components in an integrated solar hydrogen generator (Berger and Newman 2013; Haussener et al. 2012; Surendranath et al. 2012; Winkler et al. 2013). The output from the photovoltaic component must match the electrochemical load from the catalytic and ion transport components of the device. By controlling the dimensions and component architecture, it is possible to optimize the performance of the device so that it operates at near maximum possible efficiency (Jacobsson et al. 2013; Peharz et al. 2007; Winkler et al. 2013).

LOOKING AHEAD

Optimizing the topology of the components in a device can help overcome some stability limitations, achieve operations under a wide range of conditions, and increase overall efficiency. As new components become available, significant work in this design area is necessary to understand what shape and form will lead to optimization of cost, efficiency, and stability.

ACKNOWLEDGMENTS

This material is based on work performed at the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, through the US Department of Energy Office of Science under Award No. DE-SC0004993.

REFERENCES

Amirav L, Alivisatos AP. 2010. Photocatalytic hydrogen production with tunable nanorod heterostructures. Journal of Physical Chemistry Letters 1(7):1051–1054.

Baker JL, Widmer-Cooper A, Toney MF, Geissler PL, Alivisatos AP. 2010. Device-scale perpendicular alignment of colloidal nanorods. Nano Letters 10(1):195–201.

Bara JE, Hatakeyama ES, Gin DL, Noble RD. 2008. Improving CO_2 permeability in polymerized room-temperature ionic liquid gas separation membranes through the formation of a solid composite with a room-temperature ionic liquid. Polymers for Advanced Technologies 19(10):1415–1420.

Bara JE, Carlisle TK, Gabriel CJ, Camper D, Finotello A, Gin DL, Noble RD. 2009. Guide to CO_2 separations in imidazolium-based room-temperature ionic liquids. Industrial and Engineering Chemistry Research 48(6):2739–2751.

Baranov D, Fiore A, van Huis M, Giannini C, Falqui A, Lafont U, Zandbergen H, Zanella M, Cingolani R, Manna L. 2010. Assembly of colloidal semiconductor nanorods in solution by depletion attraction. Nano Letters 10(2):743–749.

Bard AJ, Fox MA. 1995. Artificial photosynthesis: Solar splitting of water to hydrogen and oxygen. Accounts of Chemical Research 28(3):141–145.

Berger A, Newman JS. 2013. Photoelectrochemical modeling of a water-splitting membrane. In preparation for Journal of Electrochemical Society.

Boettcher SW, Spurgeon JM, Putnam MC, Warren EL, Turner-Evans DB, Kelzenberg MD, Maiolo JR, Atwater HA, Lewis NS. 2010. Energy-conversion properties of vapor-liquid-solid-grown silicon wire-array photocathodes. Science 327(5962):185–187.

Chu S, Majumdar A. 2012. Opportunities and challenges for a sustainable energy future. Nature 488(7411):294–303.

Concepcion JJ, House RL, Papanikolas JM, Meyer TJ. 2012. Chemical approaches to artificial photosynthesis. Proceedings of the National Academy of Sciences U S A 109(39):15560–15564.

Dukovic G, Merkle MG, Nelson JH, Hughes SM, Alivisatos AP. 2008. Photodeposition of Pt on colloidal CdS and CdSe/CdS semiconductor nanostructures. Advanced Materials 20(22):4306–4311.

Faunce TA, Lubitz W, Rutherford AW, MacFarlane D, Moore GF, Yang P, Nocera DG, Moore TA, Gregory DH, Fukuzumi S, Yoon KB, Armstrong FA, Wasielewski MR, Styring S. 2013. Energy and environment policy case for a global project on artificial photosynthesis. Energy and Environmental Science 6(3):695–698.

Fujishima A, Honda K. 1972. Electrochemical photolysis of water at a semiconductor electrode. Nature 238(5358):37–38.

Gattrell M, Gupta N, Co A. 2007. Electrochemical reduction of CO_2 to hydrocarbons to store renewable electrical energy and upgrade biogas. Energy Conversion and Management 48(4):1255–1265.

Gu Y, Lodge TP. 2011. Synthesis and gas separation performance of triblock copolymer ion gels with a polymerized ionic liquid mid-block. Macromolecules 44(7):1732–1736.

Gupta S, Zhang Q, Emrick T, Russell TP. 2006. "Self-corralling" nanorods under an applied electric field. Nano Letters 6(9):2066–2069.

Gwee L, Choi JH, Winey KI, Elabd YA. 2010. Block copolymer/ionic liquid films: The effect of ionic liquid composition on morphology and ion conduction. Polymer 51(23):5516–5524.

Haussener S, Xiang C, Spurgeon JM, Ardo S, Lewis NS, Weber AZ. 2012. Modeling, simulation, and design criteria for photoelectrochemical water-splitting systems. Energy and Environmental Science 5(12):9922–9935.

Hernandez-Pagan EA, Vargas-Barbosa NM, Wang T, Zhao Y, Smotkin ES, Mallouk TE. 2012. Resistance and polarization losses in aqueous buffer-membrane electrolytes for water-splitting photoelectrochemical cells. Energy and Environmental Science 5(6):7582–7589.

Hoarfrost ML, Segalman RA. 2011. Ionic conductivity of nanostructured block copolymer/ionic liquid membranes. Macromolecules 44(13):5281–5288.

Hoarfrost ML, Segalman RA. 2012. Conductivity scaling relationships for nanostructured block copolymer/ionic liquid membranes. ACS Macro Letters 1(8):937–943.

IEA (International Energy Agency). 2012. Key World Energy Statistics. Paris. Available online at www.iea.org/publications/freepublications/publication/kwes.pdf.

Jacobsson JT, Fjallstrom V, Sahlberg M, Edoff M, Edvinsson T. 2013. A monolithic device for solar water splitting based on series interconnected thin film absorbers reaching over 10% solar-to-hydrogen efficiency. Energy and Environmental Science, doi: 10.1039/C3EE42519C.

James BD, Baum GN, Perez J, Baum KN. 2009. Technoeconomic Analysis of Photoelectrochemical (PEC) Hydrogen Production. Arlington VA: Directed Technologies.

Kelzenberg MD, Boettcher SW, Petykiewicz JA, Turner-Evans DB, Putnam MC, Warren EL, Spurgeon JM, Briggs RM, Lewis NS, Atwater HA. 2010. Enhanced absorption and carrier collection in Si wire arrays for photovoltaic applications. Nature Materials 9(3):239–244.

Khaselev O, Turner JA. 1998. A monolithic photovoltaic-photoelectrochemical device for hydrogen production via water splitting. Science 280(5362):425–427.

Khaselev O, Bansal A, Turner JA. 2001. High-efficiency integrated multijunction photovoltaic/electrolysis systems for hydrogen production. International Journal of Hydrogen Energy 26(2):127–132.

Kondratenko EV, Mul G, Baltrusaitis J, Larrazabal GO, Perez-Ramirez J. 2013. Status and perspectives of CO_2 conversion into fuels and chemicals by catalytic, photocatalytic and electrocatalytic processes. Energy and Environmental Science 6:3112–3135.

Lewis NS, Nocera DG. 2006. Powering the planet: Chemical challenges in solar energy utilization. Proceedings of the National Academy of Sciences U S A 103(43):15729–15735.

Lu J, Yan F, Text J. 2009. Advanced applications of ionic liquids in polymer science. Progress in Polymer Science 34(5):431–448.

Maiolo JR, Kayes BM, Filler MA, Putnam MC, Kelzenberg MD, Atwater HA, Lewis NS. 2007. High aspect ratio silicon wire array photoelectrochemical cells. Journal of the American Chemical Society 129(41):12346–12347.

Mecerreyes D. 2011. Polymeric ionic liquids: Broadening the properties and applications of polyelectrolytes. Progress in Polymer Science 36(12):1629–1648.

Nocera DG. 2012. The artificial leaf. Accounts of Chemical Research 45(5):767–776.

Olah GA, Goeppert A, Prakash GKS. 2008. Chemical recycling of carbon dioxide to methanol and dimethyl ether: From greenhouse gas to renewable, environmentally carbon neutral fuels and synthetic hydrocarbons. Journal of Organic Chemistry 74(2):487–498.

Peckham TJ, Holdcroft S. 2010. Structure-morphology-property relationships of non-perfluorinated proton-conducting membranes. Advanced Materials 22(42):4667–4690.

Peharz G, Dimroth F, Wittstadt U. 2007. Solar hydrogen production by water splitting with a conversion efficiency of 18%. International Journal of Hydrogen Energy 32(15):3248–3252.

Pinaud BA, Benck JD, Seitz LC, Forman AJ, Chen Z, Deutsch TG, James BD, Baum KN, Baum GN, Ardo S, Wang H, Miller E, Jaramillo TF. 2013. Technical and economic feasibility of centralized facilities for solar hydrogen production via photocatalysis and photoelectrochemistry. Energy and Environmental Science 6(7):1983–2002.

Plass KE, Filler MA, Spurgeon JM, Kayes BM, Maldonado S, Brunschwig BS, Atwater HA, Lewis NS. 2009. Flexible polymer-embedded Si wire arrays. Advanced Materials 21(3):325–328.

Reece SY, Hamel JA, Sung K, Jarvi TD, Esswein AJ, Pijpers JJH, Nocera DG. 2011. Wireless solar water splitting using silicon-based semiconductors and earth-abundant catalysts. Science 334(6056):645–648.

Ryan KM, Mastroianni A, Stancil KA, Liu H, Alivisatos AP. 2006. Electric-field-assisted assembly of perpendicularly oriented nanorod superlattices. Nano Letters 6(7):1479–1482.

Saur G, Ainscough C. 2011. US Geographic Analysis of the Cost of Hydrogen from Electrolysis. Technical Report NREL/TP-5600-52640. Golden CO: National Renewable Energy Laboratory, US Department of Energy. Available at www.nrel.gov/hydrogen/pdfs/52640.pdf.

Schneider Y, Modestino MA, McCulloch BL, Hoarfrost ML, Hess RW, Segalman RA. 2013. Ionic conduction in nanostructured membranes based on polymerized protic ionic liquids. Macromolecules 46(4):1543–1548.

Simone PM, Lodge TP. 2009. Phase behavior and ionic conductivity of concentrated solutions of polystyrene-poly(ethylene oxide) diblock copolymers in an ionic liquid. ACS Applied Materials and Interfaces 1(12):2812–2820.

Spurgeon JM, Lewis NS. 2011. Proton exchange membrane electrolysis sustained by water vapor. Energy and Environmental Science 4(8):2993–2998.

Spurgeon JM, Walter MG, Zhou J, Kohl PA, Lewis NS. 2011. Electrical conductivity, ionic conductivity, optical absorption, and gas separation properties of ionically conductive polymer membranes embedded with Si microwire arrays. Energy and Environmental Science 4(5):1772–1780.

Sudre G, Inceoglu S, Cotanda P, Balsara NP. 2013. Influence of bound ion on the morphology and conductivity of anion-conducting block copolymers. Macromolecules 46(4):1519–1527.

Sun J, Liu C, Yang P. 2011. Surfactant-free, large-scale, solution-liquid-solid growth of gallium phosphide nanowires and their use for visible-light-driven hydrogen production from water reduction. Journal of the American Chemical Society 133(48):19306–19309.

Sun Y, Sun J, Long JR, Yang P, Chang CJ. 2013. Photocatalytic generation of hydrogen from water using a cobalt pentapyridine complex in combination with molecular and semiconductor nanowire photosensitizers. Chemical Science 4(1):118–124.

Surendranath Y, Bediako DK, Nocera DG. 2012. Interplay of oxygen-evolution kinetics and photovoltaic power curves on the construction of artificial leaves. Proceedings of the National Academy of Sciences U S A 109(39):15617–15621.

Walter MG, Warren EL, McKone JR, Boettcher SW, Mi Q, Santori EA, Lewis NS. 2010. Solar water splitting cells. Chemical Reviews 110:6446–6473.

Winkler MT, Cox CR, Nocera DG, Buonassisi T. 2013. Modeling integrated photovoltaic-electrochemical devices using steady-state equivalent circuits. Proceedings of the National Academy of Sciences U S A 110(12):E1076–E1082.

Zhang H, Shen PK. 2012. Advances in the high-performance polymer electrolyte membranes for fuel cells. Chemical Society Reviews 41:2382–2394.

FLEXIBLE ELECTRONICS

Flexible Electronics

Yueh-Lin (Lynn) Loo
Princeton University

Tse Nga (Tina) Ng
Palo Alto Research Center

One of the frontier goals in electronics research is to transform conventional fabrication processes to meet the demands of soft, pliant, and often easily damaged surfaces. Research in new materials and patterning technologies has enabled flexible electronics that push the boundaries of how electronics are made and used toward the possibility of incorporating electronic control and power sources into any object.

Unlike conventional silicon electronics that are limited to rigid wafers, flexible electronic devices have been demonstrated on plastics, paper, fibers, and even biological tissues. These flexible devices enable a wide range of applications, in fields ranging from energy sustainability to smart sensor networks to bioelectronics. Some specific examples are energy-efficient, stretchable lighting, lightweight photovoltaics, smart-sensing wallpaper, and dissolvable electronic implants.

To make flexible electronics that are compatible with delicate surfaces, low-temperature processing is required. This need has led to the development of materials such as organic conductors and semiconductors as well as advanced solution-based techniques that enable low-temperature processing. Thus flexible electronics not only enable novel applications but also promote the use of alternative manufacturing technologies, such as roll-to-roll printing for electronics. Because the materials, fabrication process, and applications are interrelated, the speakers in this session touched on all three aspects to provide an overview of the rich and exciting field of flexible electronics.

Antonio Facchetti (Polyera Inc.) began with a discussion of the materials development that has enabled the fabrication of optoelectronic devices, such as displays, circuits, and solar modules, on unconventional substrates, such as flexible foils. He described materials design and processing strategies that have greatly advanced the performance of printable organic devices.

In the second talk Nanshu Lu (University of Texas at Austin) discussed the fabrication and biointegration of tissue-like electronics that can conform to—and deform with—living organisms for physiological sensing and stimulation. She explained the mechanics of thin films, microfabrication, and biointegration of bendable and stretchable electronics.

In the third presentation, we heard from Polina Anikeeva (Massachusetts Institute of Technology), who has created a new generation of flexible electrode arrays and optoelectronic neural scaffolds that aim to minimize tissue damage and maintain high-quality neural recordings over the course of several months. She reviewed the potential of these devices as a platform to investigate neuronal viability and potentially to facilitate repair of damaged neural tissues.

Materials and Process Engineering for Printed and Flexible Optoelectronic Devices

ANTONIO FACCHETTI
Polyera Corporation and Northwestern University

Printed optoelectronics is a revolutionary technology to fabricate mechanically flexible, low-cost, lightweight, and large area electronic devices by using electronic inks and processes borrowed from the graphic arts industry (Kim et al. 2012; Loo and McCulloch 2008).[1] Devices such as light emitters, light harvesters, circuits, and sensors will be based on a new materials set for the semiconductor, dielectric, electrical contact, emitter components (Facchetti 2007). Applications such as flexible displays, plastic radio frequency identification tags, disposable diagnostic devices, rollable solar cells, and simple consumer products and games represent a future multibillion-dollar market. Smart objects (e.g., packages that integrate multiple printed devices) are further examples of printed optoelectronics.

Companies, startups, research institutions, and government agencies are investing in research and development in this field. Progress will also depend on close collaboration among chemists, materials scientists, and engineers—whether they are material providers, equipment makers, producers, or system integrators—to ensure the success of printed electronics in the marketplace.

This paper provides an overview of the materials and process requirements and the applications of this technology.

---

[1] *Printed electronics* and *optoelectronics* are used interchangeably in this paper, although printed electronics generally refers to electronic circuits, whereas optoelectronics embraces photonic devices such as light-emitting diodes, light-emitting transistors, and solar cells. Also, because all printable materials were originally organic, *organic electronics* was used instead of *printed electronics*; several inorganic materials are now printable.

BACKGROUND

The goal of printed electronics is not to replace the conventional inorganic-based electronic industry. Rather, it offers opportunities for new products and/or reduces the cost of certain devices (Table 1).

Traditional thin film materials deposition is accomplished with chemical vapor deposition, physical vapor deposition, and sputtering. Although these processes are performed in a vacuum, they are not intrinsically "low speed." For instance, polymer webs over 2 m in width are metallized at 18 m/s for food packaging at the cost of pennies per square meter. Conventional printing presses used in the graphic art industry commonly run at speeds of m/s, with webs several meters wide, and are used to deposit several different types of color inks. Similarly, at the highest degree of sophistication liquid crystal display production is based on processing large glass plates (>4 m^2) with a takt time of a few minutes.

Film patterning on glass (for display) and on silicon wafers uses conventional photolithography. In this subtractive process active film is deposited over the entire substrate area and then selected regions of it are removed by coating the film with photoresist, exposing the photoresist film to contact or projection optics (or electrons for e-beam lithography), developing the photoresist, etching the underlying layer, and stripping the resist. This process produces the resolution and reliability required for the high-tech integrated circuit industry. But photolithography is costly, using extremely expensive equipment (a new FAB line, the infrastructure to fabricate electronic devices, costs \$2–3 billion), and requiring several batch-to-batch steps.

TABLE 1 Conventional versus organic electronics.

	Conventional Electronics	Organic/Printed Electronics
Advantage or disadvantage	High performance	Low performance
	Small area/feature size	Large area/feature size
	High cost/unit area	Low cost/unit area
	High capital investment	Low capital investment
	Long production run	Short production run
	Durable	Disposable
	Rigid	Flexible
	Selected markets	Everywhere
	Photolithography	Printing
Materials	Semiconductor, conductor, dielectric, passive, substrate	
Devices/applications	Transistors, circuits, memory, diodes, sensors, displays, batteries, photovoltaics, conductive traces, antennas, resistors, capacitors, inductors	

The advantage of using printable materials, most of which are organic (see below), is to replace conventional processes for device fabrication. The formulation of organic materials into inks (active material + solvent + additives) would enable roll-to-roll printing or printing-like processes. Furthermore, the additive process of printing minimizes material waste. If similar processes could be used for functional materials, high-volume inexpensive devices could be fabricated. This is a goal of printed (or organic) electronics.

The main obstacle to the realization of this technology is on the materials side, particularly the semiconductor (charge-transporting material), because highly processable semiconductors exhibit poor charge transport performance. Furthermore, it is unlikely that the same printing presses used for newspapers and magazines can be used for processing functional materials, so modification and optimization on the processing side are also necessary. Finally, device design architectures and circuit engineering are necessary to address the poorer performance of organic materials and/or take advantage of their unique properties.

DEVICES AND APPLICATIONS

Transistors and Circuits

The field-effect transistor (FET) is essential in almost all electronic devices (Facchetti 2007), and it is the building block of the circuits (a collection of connected FETs) necessary for logic operations, memory functions, displays, and sensors.

FETs based on organic (OFET) or other printable semiconductors have the structure of a thin film transistor (Figure 1), a three-terminal device composed of source, drain, and gate electrodes, a dielectric layer, and a semiconducting layer. The transistor is essentially an electronic valve or switch, and the flow of current between the source and drain electrodes (for a given source-drain bias) is controlled by the magnitude of the source-gate voltage (or electric field dropped through the dielectric layer). The charge flow in the transistor channel can be dominated by positive charges or negative charges (electrons), which define whether the semiconductor is p- or n-type, respectively.

The two most important transistor performance parameters are the charge carrier mobility (μ, how fast electrons move) and the current on-off ratio (I_{on}:I_{off}, how efficiently the current can be modulated by the source-gate bias). To maximize the transistor speed, the carrier mobility should be as high as possible and the distance between the source and drain electrodes (channel length) as small as possible. The carrier mobility of printable semiconductors is about two orders of magnitude lower than that of crystalline inorganic materials, and typical resolution for the OFET channel length in printed devices is larger by the same order of magnitude. Thus OFET circuit speeds cannot compete with those based on silicon or gallium arsenide and fabricated using photolithographic processes. However,

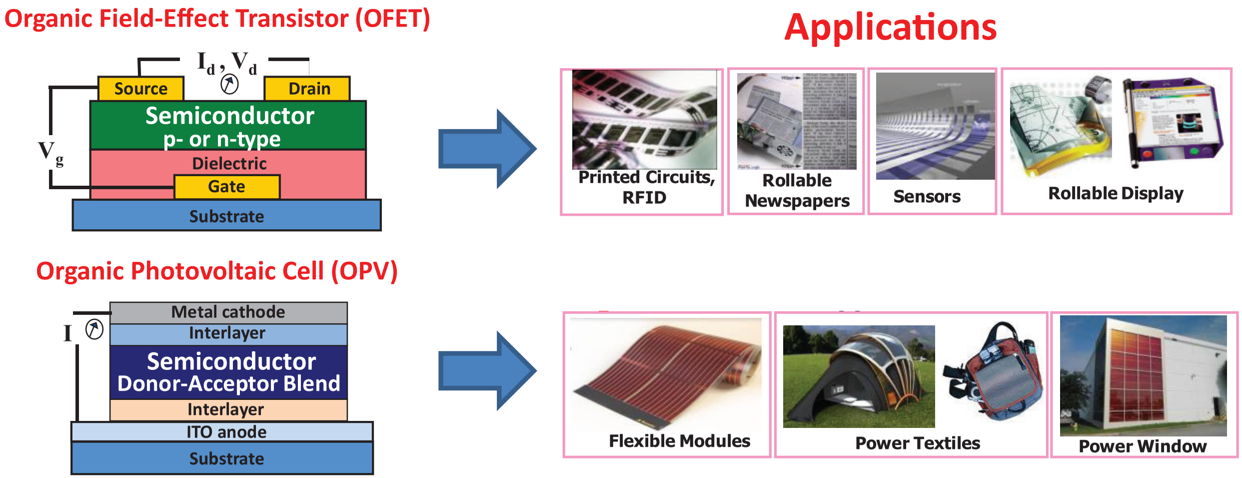


FIGURE 1 Structure of an organic field-effect transistor and photovoltaic cell and corresponding applications. ITO = tin-doped indium oxide; RFID = radio frequency identification; Id = Drain current; Vd = drain voltage; Vg = gate voltage; I = Photocurrent.

when the performance requirements are relaxed and/or it is necessary to add device functions (e.g., flexibility, easy integration) and/or reduce costs, OFETs may be very competitive.

Displays

One great area of opportunity for printed electronics is in the fabrication of backplane circuits for flexible displays. Electrophoretic displays, in which the image is formed by black and white charged particles, are well suited for printed transistors because of the slow switching time and minimal current flow needed to drive them. Furthermore, they are bi-stable, meaning that the image is retained without power (power is required only during refresh). In these displays the contrast is independent of viewing angle and significantly better than newsprint. Polymer-dispersed liquid crystal and electrochromic-based displays can also be driven by printed transistors.

In addition to transistors and backplanes, organic electronic materials, some of which can be solution processed, are used to fabricate emissive devices such as organic and polymer light-emitting diodes (OLEDs and PLEDs, respectively).

Radio Frequency Identification Tags

Radio frequency identification (RFID) tags use radio frequency transmissions to identify, track, sort, and detect persons and items based on communication between a reader (interrogator) and a transponder (a chip connected to an antenna, often called a tag). RFID tags can be either active (powered by battery) or passive (powered by the reader field) and come in various forms such as smart cards, tags, labels, watches, and even embedded in mobile phones. The communication frequencies used range from 125 KHz to 2.45 GHz, depending to a large extent on the application. Regulations imposed by most countries control emissions and prevent interference with other industrial, scientific, and medical equipment.

RFID is not expected to replace bar codes in the supply chain because tags are still too expensive, even though their prices have fallen to around 20 cents in volume (versus 0.2 cents for a bar code label). Adoption is therefore likely first for expensive items, then as technology advances and costs decline tags will probably appear on more and more products. It is commonly thought that the only way to reduce the price sufficiently, and produce billions or trillions of tags per year, is by printing both the circuitry (using solution-processable materials) and the antenna in an integrated process. The major obstacle to printed RFID applications is to achieve high circuit frequency operation and enable efficient rectification (convert the ac-voltage detected and generated by the antenna at the targeted base carrier frequency to a dc voltage).

Sensors

Another important opportunity for printed electronics lies in the fabrication of sensors, which can be used to detect a variety of stimuli such as temperature, pressure, radiation, and chemical identity. An integrated temperature and pressure sensor array has been utilized as an artificial skin, and organic actuators have been fabricated. Sensors are also used in tamper-detecting packaging, data-logging pill dispensers, chemical sensors, electronic noses and tongues, photodiodes, and light scanners.

Photovoltaics

Another exciting application of printed electronics is in photovoltaics, a field currently dominated by silicon. The vision is to fabricate inexpensive, lightweight, flexible, conformal, and efficient production of energy from the sun.

Photovoltaic devices are composed of a charge-transporting donor-acceptor semiconductor blend sandwiched between two electrical contacts (Figure 1) (Facchetti 2011). Light exposure produces free carriers that are collected at the electrodes as electrical energy. The most important metric is that of power conversion efficiency (showing how efficiently the solar energy is converted into electrical energy).

To be more widely adopted, the photovoltaic semiconductor needs to provide high power conversion efficiencies (solar to electrical energy efficiency) while remaining stable and inexpensive.

PRINTING TECHNOLOGIES

Enabling the use of roll-to-roll and high-throughput device fabrication technologies is the most relevant aspect of printed electronics. Throughputs greater than 1 m^2/s are considered “high volume”; most printing methods fall in this category. There are several considerations to determine what process can be used based mainly on the viscoelastic properties of materials and the desired feature sizes (lateral resolution, film thickness, surface morphology, surface energy) required for device assembly. Table 2 shows some of the most important printing techniques and relevant specifications for use in electronics; probably the most used processes are inkjet, screen, and gravure printing.

Inkjet Printing

This process uses a stepper motor to control the position of the print head along a stabilizer bar; as the print head slides back and forth along the bar, ink drops are ejected from the nozzle to create a pattern on the substrate. There are two primary mechanisms for ejecting the ink: in a thermal inkjet, evaporation of

TABLE 2 Typical ink requirements, printing features, and throughputs of conventional printing methods.

Technique	Ink Viscosity (mPas)	Film Thickness (μm)	Resolution (μm)	Registration (μm)	Throughput (m^2/s)	Features/ Issues	Electronic Materials Printed
Lithography	$(10–4) \times 10^2$	1.5–0.5	50–10	>10	0.1	High quality; Need for additives	Conductor
Screen	$(5–0.5) \times 10^2$	100–30	100–30	>25	2–3	Wide range of inks; Medium quality	Conductor
Flexography	5000–50	2.5–0.8	70	<200	10	Wide range of substrates; Medium quality	Conductor Semiconductor Dielectric
Gravure	200–50	5–0.5	<10	>5	60	Large run length; High quality	Dielectric Semiconductor
Pad	>50	2–1	20	>10	0.1	Nonplanar objects	Conductor
Laser/thermal	N/A	<1	5	~10	0.002	N/A	Conductor Semiconductor
Inkjet	30–1	<0.5	<10	20–5	0.5–0.01	Digital data; Local registration	Conductor Semiconductor

N/A = not applicable

a small portion of the ink solvent forces ink out of the nozzle; in a piezoelectric inkjet, voltage applied to a piezoelectric material causes it to expand, forcing ink out of the nozzle.

Inkjet printing is now one of the most widespread graphic arts printing methods, and in recent years has received attention as a technique to deposit functional materials with specific electrical, optical, chemical, biological, or structural functionalities at well-defined locations on a substrate.

One of the most useful features of inkjet printing is its capacity to vary the printing pattern without the need to make a new printing plate. Using a camera and image analysis software, the printed image can be adjusted "on the fly" to compensate for many of the registration errors that plague other types of printing process. Organic semiconductors are commonly inkjet patterned for transistor and OLED fabrication. Recently a variation on inkjet printing, self-aligned printing, was used to pattern features as small as 60 nm.

Inkjet printing has some limitations, such as susceptibility of the printing head to corrosion from aggressive solvents, a liability to high mechanical shears in piezoelectric print heads, and high temperatures in thermal inkjet heads. In addition, fluctuations in droplet volume or trajectory can adversely affect film uniformity (creating a "coffee-stain" effect) and materials performance.

Screen Printing

Screen printing consists of three key elements—the screen, which is the image carrier; the squeegee; and the ink—and uses a porous polyester or stainless steel mesh stretched tightly over a wood or metal frame. A stencil, produced on the screen either manually or photochemically, defines the image to be printed (in other printing technologies this is called an image plate). The technique usually produces relatively thick films (thinner films fabricated using less viscous inks often result in poorly defined printed patterns); the resolution is poor and limited by the screen size, although recent presses perform much better.

In organic electronics this technique has been used to fabricate top-level interconnects and contacts where the thickness of the printed film is not a critical factor. It can also be used to print thick dielectric layers and passive materials for device encapsulation. Indeed, the first report of a "printed organic thin film transistor" described screen-printed carbon paste electrodes for source, drain, and gate contacts (Horowitz et al., 1996).

Gravure Printing

In gravure printing the image areas consist of honeycomb-shaped cells that typically are etched or engraved on a copper cylinder. The cylinder rotates into an ink pan and, as it turns, excess ink is removed by a flexible steel doctor blade so that the ink remaining in the recessed cells forms the image by direct transfer to the substrate as it passes between the plate cylinder and the impression cylinder. The process uses fluid inks with relatively low viscosity.

Because it is one of the highest-volume printing processes, gravure printing is often used commercially to produce high-quality graphic materials such as magazines. Depending on the nature (depth, surface energy, shape, etc.) of the engraved cells, different amounts of material used to fabricate the devices (semiconductor, conductor, dielectric) can be deposited in different substrate regions, enabling tuned film thicknesses. The drawback is that the printed pattern edges may be wavy.

Gravure has recently been used to fabricate dielectric and semiconducting layers for organic transistors.

ELECTRONIC MATERIALS

Printed electronic devices need a set of core materials for charge accumulation, injection, and transport as well as specific materials to enable particular device functions (Facchetti 2013; Facchetti et al. 2005; Usta et al. 2011).

Every type of electronic device needs memory capacity and a control for current flow, both of which are based on FETs, which in turn need three fundamental materials: a conductor, dielectric, and semiconductor (Table 3). Depending on the specific device functions, additional active materials may be needed. OLEDs, for example, require an emissive material for efficient conversion of electricity to light, whereas organic photovoltaics need a photosensitizer and/or efficient light absorber for photon absorption and dissociation (in addition to the materials needed for efficient charge transport).

Displays may be based on different technologies for pixel fabrication, such as organic emitters, electrophoretic inks (proper dyes are necessary), liquid crystals (LC molecules are used), and electrochromic compounds. Many other types of chemicals/materials may also be necessary for device fabrication, such as small molecules for interfacial layers to ensure efficient charge injection or surface energy match, additives for use as dopants or stabilizers, and polymers for encapsulation.

The properties of conductors, semiconductors, and dielectrics are summarized below.

Semiconductors

The semiconductor is the most important material in optoelectronic devices, although its key function varies depending on the device application, and different electronic devices use different types of semiconductors. When used in organic transistors, it is the material where, at the interface with the dielectric, charges are accumulated and transported. The semiconductor must satisfy a number of requirements for charge injection and transport related to the device structure.

TABLE 3 Properties of current-generation solution-processed materials for printable electronics.

	Materials						
	Conductor		Semiconductor			Dielectric	
Performance/ needs	Metal/metal oxides	Polymers	Small molecules	Polymers	Inorganics	Polymers	Inorganics
Current performance	$\sigma > 10^4$ S/cm	$\sigma > 1$ S/cm ($\sim 10^4$ S/cm)	$\mu = 1–30\ \text{cm}^2/\text{Vs}$; $I_{on}:I_{off} > 10^6$	$\mu = 0.1–10\ \text{cm}^2/\text{Vs}$; $I_{on}:I_{off} > 10^6$	$\mu = 1–100\ \text{cm}^2/\text{Vs}$; $I_{on}:I_{off} > 10^6$	$J < 10^{-8}\ \text{A/cm}^2$; BF > 6 MV/cm	$J < 10^{-7}\ \text{A/cm}^2$; BF > 5 MV/cm
Current materials	Ag, Au, Cu nanoparticles; ITO	PEDOT:PSS; PANI; Polymer + CNT; Graphene	Fused thiophenes; Heteroarenes; Perylenes	Polythiophenes; Naphthalene diimides; DPPs	In2O3, ZnO, IZO, IGZO	PMMA; P-UV; Cross-linked PVP	Sol-gel oxides; Oxides-epoxy
Advantages	Good processability; High conductivity	Good processability; Sufficient conductivity	Easy purification; Facile scale-up	Easy ink formulation	High mobility	Easily printable	Tunable permittivity
Limitations	Costly	Low-speed application	Difficult ink formulation for printing; Ambient stability	Difficult purification; Few n-channels available; Limited ambient stability	High processing temp; Few p-channels available; Ambient stability	Low permittivity; High permeability; Thick films	Leaky; Rough surface; Thick films
Next-generation targets	Reduce cost	Enhance conductivity	Increase μ	Increase μ	Reduce T; enhance uniformity	Reduce film thickness	Enhance printability

NOTE: σ = conductivity; μ = charge carrier mobility; Ag = silver; Au = gold; BF = breakdown field; CNT = carbon nanotube; Cu = copper; DPP = diketopyrrolopyrrole; IZO = indium-zinc-oxide; IGZO = indium-gallium-tin-oxide; ITO = tin-doped indium oxide; J = leakage current density; PANI = polyaniline; PMMA = poly(methylmetacrylate); PEDOT:PSS = polyethylenedioxothiophene:polystyryl sulphonate; P-UV: A UV-vis crosslinkable polymer; PVP = poly(vinylphenole); T = temperature.

The design of highly efficient and easily printable organic semiconductors for OFETs is the key challenge in roll-to-roll electronics. Optimization of charge transport (and thus carrier mobility) requires that the semiconductor molecules be planar so that the molecular orbitals can overlap efficiently. However, this molecular design usually leads to poorly soluble materials, which are very difficult to print. Obviously, for roll-to-roll printing fabrication, the organic semiconductor must be solution processable so that it can be formulated into an ink.

But good molecular design and solution processability are only two factors. The charge transport in organic semiconductor films is highly dependent on the film deposition conditions—printing process, solvent used in formulating the ink, active/additive component concentrations, deposition temperature, substrate morphology, and surface energy. Environmental conditions during film deposition can also affect materials performance, although some organic semiconductors are air stable and do not require a controlled environment during film processing.

Several organic semiconductors for OFETs have been synthesized, including those based on small molecules and polymers. P-type organic semiconductors have been shown to have carrier mobilities of ~10 cm^2/Vs as thin film and up to 35 cm^2/Vs for single crystals. These mobility numbers are greater than that of amorphous silicon (~ 0.1-1 cm^2/Vs), which is commonly used for large area display backplanes, but very few semiconductors exhibit good charge transport and solution processability. In this respect, probably the most promising semiconductor families are those based on thiophene-containing polymers.

A common drawback in organic electronics is the limited availability of electron-transporting (n-type) semiconductors, which are needed for complementary circuit applications; very few air-stable n-type organic semiconductors are known. Sol-gel and nanoparticulate inorganic semiconductors or hybrid organic-inorganic semiconductor materials have been investigated; these materials promise both the superior carrier mobility of inorganic semiconductors and the processability of organic materials.

Conductors

All electronic devices have electrical contacts, which should satisfy a number of requirements—high conductivity, appropriate work function, chemical stability, and appropriate surface energy characteristics and morphology.

Materials used as conductors are metals/metal oxides and conducting π-conjugated polymers. Metallic features can be fabricated by thermal evaporation, sputtering, and printing (the latter is essential for organic electronics). Printing metal is usually achieved by using inks that contain metal particles, which may differ substantially in morphology, size, and type/amount of stabilizers; gold, silver, and other noble metals are typically used.

Alternatively, nanoparticle-based inks can be printed and subsequently sintered at temperatures (<150°C) compatible with inexpensive plastic foils, or metal precursors can be used, sometimes in combination with other materials, and similarly thermally cured. Metal oxides are another class of conductive materials often used for electrodes (tin-doped indium oxide is by far the most commonly used).

Even though certain π-conjugated polymers are highly conductive, they typically exhibit far lower conductivity than metals. The most common conducting polymers used in printing conducting lines are polyaniline, polythiophenes, and polypyrroles; of these, PEDOT:PSS (a polythiophene-based polymer) is the most widely used because it is commercially available and exhibits good conductivity (<400 S/cm).

Dielectrics

The dielectric film is extremely important for OFETs because it enables the creation of induced charges in the semiconductor upon application of the gate voltage. A good dielectric material should exhibit high dielectric strength, low leakage current, and high capacitance. The latter allows higher charge density to be induced at lower gate voltages and thus reduces power consumption.

The dielectric layer capacitance can be enhanced by using thinner films and/or by increasing the material dielectric constant. Unfortunately, when the layer becomes too thin, film defect density and leakage current increase. Furthermore, the dielectric film surface in contact with the semiconductor should be very smooth. Because charge transport in organic semiconductors is confined in the semiconductor within a few nanometers of the semiconductor-dielectric interface, rough interfaces generate charge scattering and reduce carrier mobility.

Again, for printed electronic applications, the dielectric material must be solution processable. Various dielectric materials have been used to fabricate OFETs; absurdly, they have mostly been inorganic films such as silica, alumina, and titanate, which are not generally printable. A variety of organic polymers—polypropylene, polyvinyl alcohol, poly(vinyl phenol), poly(methyl methacrylate)—have been used as dielectrics; most of them are widely used for other purposes and available in bulk quantities quite inexpensive.

More recently more complex formulations have been printed to fabricate OFETs. These dielectric films are usually cured either thermally or photochemically to enhance mechanical strength and improve dielectric properties.

SUMMARY AND OUTLOOK

Printed electronics has the potential to become a significant industry within the next decade and, contrary to common opinion, many of the forecasted applications will be new and not created by eroding today's electronics markets.

The challenges for this technology are mainly related to materials performance for circuits and proper engineering of printing processing methods for electronic device assembly. The semiconductor is the weakest of the current generation of materials. The most important limitations are the lack of high carrier mobility, environmentally stable semiconductors (particularly n-type), high-performance solution-deposited semiconducting films over a large area, and temporal performance stability.

Despite these difficulties, initial important successes in device fabrication using organic materials and roll-to-roll processes are encouraging and several companies strongly believe that organic electronics is already a reality.

REFERENCES

Facchetti A. 2007. Semiconductors for organic transistors. Materials Today 10(3):28–37.

Facchetti A. 2011. π-Conjugated polymers for organic electronics and photovoltaic cell applications. Chemistry of Materials 23:733–758.

Facchetti A. 2013. Polymer donor-polymer acceptor (all polymer) solar cells. Materials Today 16:123–132.

Facchetti A, Yoon M-H, Marks TJ. 2005. Gate dielectrics for organic field-effect transistors: New opportunities for organic electronics. Advanced Materials 17:1705–1725.

Horowitz G, Kouki F, Spearman P, Fichou D, Nogues C, Pan X, Garnier F. 1996. Evidence for n-type conduction in a perylene tetracarboxylic diimide derivative. Advanced Materials 8:242–245.

Kim J, Ng TN, Kim WS. 2012. Highly sensitive tactile sensors integrated with organic transistors. Applied Physics Letters 101:103308-1–103308-5.

Loo Y-L, McCulloch I. 2008. Progress and challenges in commercialization of organic electronics. MRS Bulletin 33:653–662.

Usta H, Facchetti A, Marks TJ. 2011. n-Channel semiconductor materials design for organic complementary circuits. Accounts of Chemical Research 44:501–510.

Mechanics, Materials, and Functionalities of Biointegrated Electronics

NANSHU LU
University of Texas at Austin.

Robust bioelectronic interfaces present unlimited potentials in wearable health monitors, implantable devices, and human-machine interfaces. But conventional high-performance electronics, which are based on planar and rigid silicon wafers, are intrinsically incompatible with curvilinear and deformable natural organisms. This challenge is being approached with a mechanics-based strategy involving the use of neutral planes and filamentary serpentine networks. The resulting structural-electrical design has enabled flexible and stretchable electronics to conform to—and deform with—biological tissues for physiological sensing, programmable stimulation, and on-demand therapeutics. This article summarizes the mechanics, materials, and functionalities of such biointegrated electronics and concludes with a discussion of future directions.

BACKGROUND

Research on flexible electronics started nearly two decades ago (Bao et al. 1997; Garnier et al. 1994) with the demand for macroelectronics (i.e., large-area electronics), such as paperlike flexible displays (Rogers et al. 2001). Early research focused on organic semiconductors and conducting polymers because their intrinsic deformability, light weight, and low manufacturing cost are appealing for large-area flexible electronics, especially when merged with roll-to-roll processes (Forrest 2004). Methods to synthesize, pattern, and passivate organic electronic materials (Forrest and Thompson 2007; Menard et al. 2007) were then developed and applied to the manufacture of devices such as organic solar cells (Kaltenbrunner et al. 2012; Lipomi et al. 2011) and artificial electronic skins for robotics (Mannsfeld et al. 2010; Someya et al. 2004; Takei et al. 2010), and flexible displays based on organic light-emitting diodes are nearing commercial reality.[1]

But the chemical instability of organic semiconductors and difficulties associated with low electronic performance have somewhat limited their application in high-speed, low-power, or long-lasting electronics. In contrast, inorganic semiconductors exhibit high carrier mobility and on-off ratio as well as excellent chemical stability in ambient environments (Service 2006). Furthermore, the material and electronic properties of inorganic semiconductors and metals have been well defined and the manufacturing processes well established after more than 100 years of research and applications. Thus flexible electronics based on rigid but high-quality monocrystalline inorganic semiconductors started to emerge in the mid-2000s (Khang et al. 2006).

To overcome the rigidity of inorganic electronic materials, thin film mechanics has been applied to enhance the deformability of polymer-bonded metallic and ceramic membranes.

MECHANICS: BENDABILITY AND STRETCHABILITY OF INORGANIC ELECTRONIC MATERIALS

Inorganic materials such as silicon and metals are stiff and readily rupture or yield when their intrinsic strain exceeds even very small values, such as 1 percent. But the mechanical limit of a structure can be offset by the geometry of the construction even for intrinsically fragile materials.

Basic beam theory predicts that the bending-induced maximum strain of a membrane is proportional to the product of film thickness and bending curvature. If the maximum strain is limited to a critical strain to rupture of the material (e.g., 1 percent), then the maximum allowable bending curvature will be inversely proportional to the thickness of the silicon plate/membrane, as shown in the log-log plot of Figure 1. As the membrane thickness decreases from millimeters to tens of nanometers, the attainable bending curvature can be enhanced by five orders of magnitude. As a result, although bulk silicon wafers are rigid plates, silicon nanomembranes (with a thickness of ~100 nm) can be readily arched to the radius of a folded paper (~0.1 mm) without rupture, as shown in Figure 1C.

Building on this unprecedented bendability, silicon nanomembranes can be made stretchable by applying two prevailing design strategies. One strategy calls for bonding flat nanoribbons to a prestretched elastomeric substrate to produce wrinkled nanoribbons (represented in Figure 2A) (Khang et al. 2006; Kim et al. 2008a; Sun et al. 2006). When the prestretch is released, the elastomeric substrate fully retracts, inducing out-of-plane sinusoidal buckling in the nanoribbons in a mechanism similar to the Euler buckling of an elastic rod under axial compres-

---

[1] As evidenced in a promotional Samsung video, www.youtube.com/watch?v=N3E7fUynrZU, presented at the International Consumer Electronics Show (CES), January 8–11, 2013.

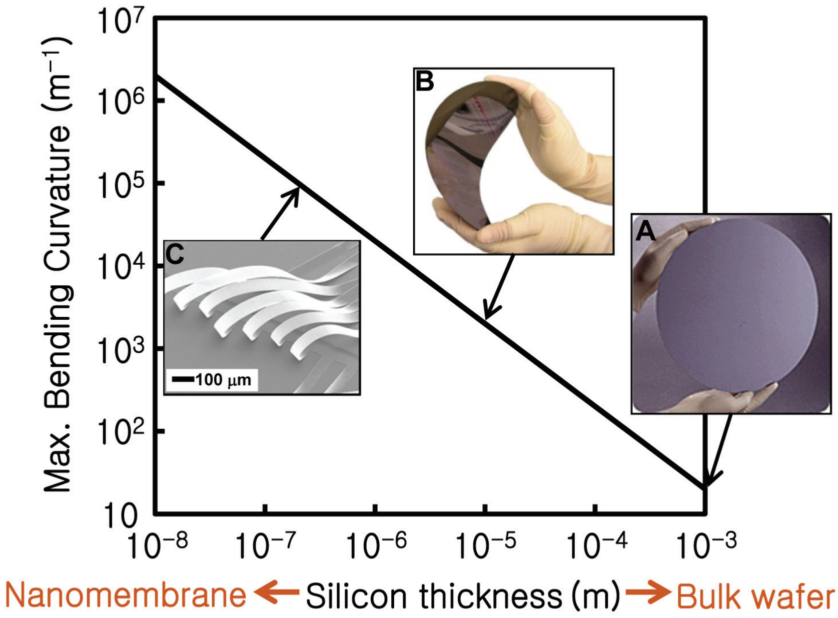


FIGURE 1 Maximum allowable bending curvature is plotted as a function of silicon plate/membrane thickness, with insets showing the bendability of (A) a bulk silicon wafer, (B) a silicon thin film, and (C) silicon nanoribbons. Reprinted from Kim et al. (2012a) with permission from Nature Publishing Group.

sion. Nanomembranes bonded to biaxially prestretched elastomeric substrates form two-dimensional wrinkled patterns as shown in Figure 2B (Choi et al. 2007). Buckling instabilities involving large displacement but small strains are the desired outcome in stretchable electronics.

With the other strategy, isolated rigid islands linked by buckled linear metallic ribbons (Figure 2C) can be stretched up to 40 percent without mechanical failure (Kim et al. 2008b; Ko et al. 2008; Lee et al. 2011). When serpentine ribbons (Figure 2D) are used instead of linear ribbons, stretchability of the system can vary from 10 percent to 300 percent depending on the serpentine tortuosity (Kim et al. 2011a,b; Xu et al. 2013).

Both wrinkling and serpentine strategies have proven effective in keeping strains in inorganic semiconducting or metallic materials below 1 percent when the polymer substrate is subjected to significant deformation (e.g., of orders of magnitude).

Furthermore, when substrate materials are too stiff to stretch but thin enough to bend (e.g., plastic sheets, paper, leather, fabric), electronics fabricated on the surface of such substrates have to survive tensile strains induced by bending

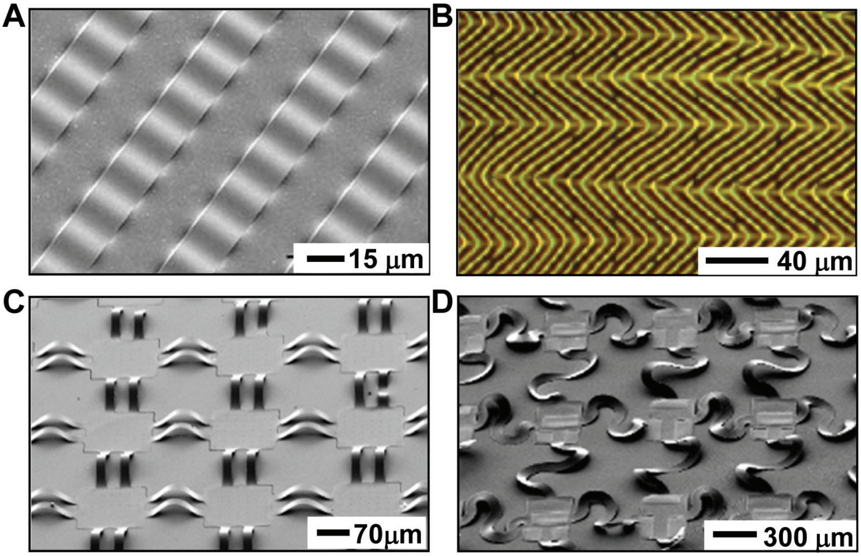


FIGURE 2 Design strategies of stretchable electronics enabled by the mechanics of film-substrate interaction. (A) Silicon nanoribbons buckled on uniaxially prestretched soft elastomer. Reprinted from Khang et al. (2006) with permission from the American Association for the Advancement of Science. (B) Silicon nanomembrane buckled on biaxially prestretched soft elastomer. Reprinted from Choi et al. (2007) with permission from the American Chemical Society. (C) Isolated device islands interconnected by popped-up linear metallic ribbons. Reprinted from Kim et al. (2012b) with permission from the Materials Research Society. (D) Isolated device islands interconnected by serpentine-shaped metallic ribbons. Reprinted from Kim et al. (2008b) with permission from the National Academy of Sciences.

curvatures. A thin compliant layer laminated between the substrate and the active device islands has been found to greatly reduce tensile strain in the islands through large shear deformation (Sun et al. 2009). Such a strain isolation mechanism has enabled bendable and even foldable electronics on a lot of unconventional substrates, such as printing papers, fabrics, and aluminum foils (Kim et al. 2009).

Exciting discoveries such as these offer ways to overcome the intrinsic brittleness and stiffness of inorganic semiconductors and open the door for their applications in flexible and stretchable electronics.

MATERIALS PROCESSING: MICROTRANSFER PRINTING

Microtransfer printing technology developed for single crystal inorganic semiconductors (Kim et al. 2010c; Meitl et al. 2006; Yoon et al. 2010) has enabled the integration of high-performance electronics on deformable substrates such as flexible displays (Park et al. 2009), high-efficiency flexible solar cells (Yoon et al. 2008, 2010), bioinspired electronic eye cameras (Ko et al. 2008; Song et al. 2013), and biointegrated electronics (Kim et al. 2012a,c,d).

Figure 3 illustrates the generalized two-step microtransfer printing method. The fabrication begins with the high-temperature process of doping silicon nanomembranes on silicon-on-insulator (SOI) wafers. Preprocessed monocrystalline silicon nanomembranes are then released from the SOI wafer and printed onto the polyimide (PI)-coated rigid handle wafer using elastomeric stamps; the precoated PI layer serves as a support and encapsulation layer for the functional metal and semiconducting nanomembranes. Conventional microfabrication processes (e.g., low-temperature sputter or electron beam deposition, photolithography, and wet or dry etching) can then be readily performed on the PI-coated wafer. The circuit is eventually patterned into stretchable open mesh networks and transfer printed from the wafer onto a wide variety of deformable substrates, again using elastomeric stamps to render a fully functional flexible/stretchable system.

Because high-quality monocrystalline silicon is used as the semiconductor and low-resistance gold wires are used as the conductor in these devices, their electronic performance and long-term chemical reliability are on par with wafer-based electronics while high flexibility and/or stretchability is incorporated through the structural design. Similar fabrication strategies are applicable to the fabrication of stretchable AlInGaP[2] optoelectronics (Kim et al. 2010b) and gallium arsenide (GaAs) photovoltaics (Lee et al. 2011).

FUNCTIONALITIES

Epidermal and in Vivo Sensing

With the maturity of the enabling technology for microtransfer printing, flexible and stretchable electronics found their exemplary applications in the late 2000s with the emergence of biointegrated electronics, a field that has greatly facilitated epidermal and in vivo sensing (Rogers et al. 2010).

For epidermal sensing, physiological electrodes are mounted on the skin (via adhesive tape, mechanical straps, or needles) with terminal connections to separate boxes that house collections of rigid circuit boards, power supplies, and communication components (Gerdle et al. 1999; Webster 2009). These systems have many important capabilities, but they are poorly suited for practical application outside of research labs or clinical settings.

The development of novel electronic systems with matching form factors and the mechanical properties of biotissues is essential for long-term, intimate bioelectronic interfaces. To that end, the application of serpentine structural

---

[2] Aluminum gallium indium phosphide.

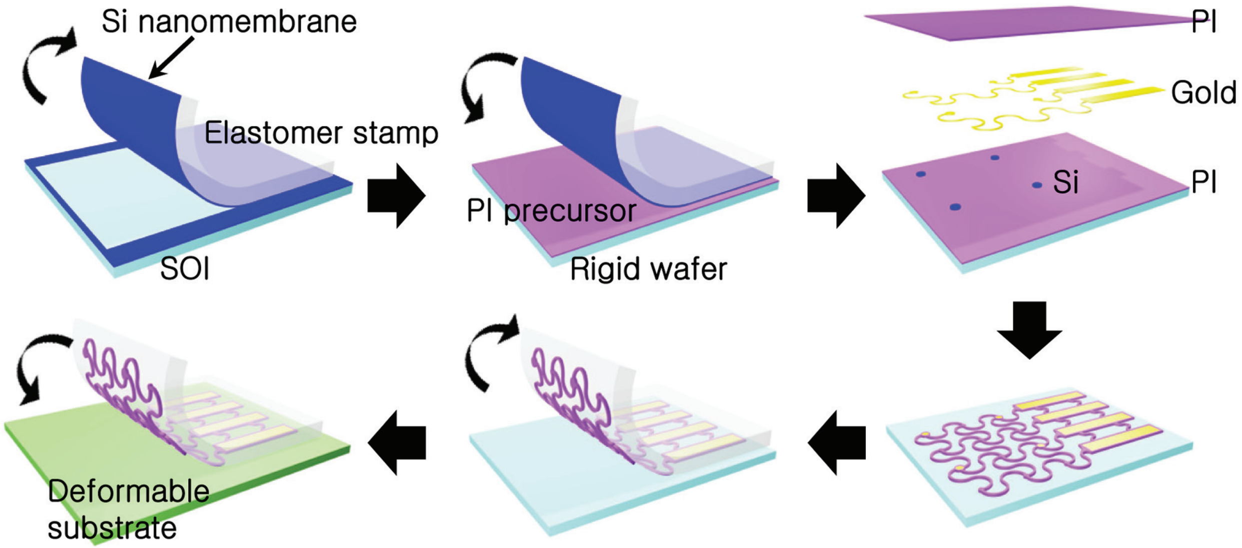


FIGURE 3 Schematics of the fabrication procedures of stretchable electronics: transfer doped silicon (Si) nanomembranes from silicon-on-insulator (SOI) wafers onto polyimide (PI)-coated rigid handle wafer with an elastomer stamp. Silicon patterning and metallization are followed by PI encapsulation. Dry etching of PI defines the serpentine open mesh structure. Finally the well-fabricated stretchable circuit is transferred from rigid wafer to a deformable substrate.

designs and transfer-printing methods has enabled the development of ultrathin, ultrasoft electronics composed of high-performance inorganic materials. Such biointegrated electronics have in turn led to exciting applications such as epidermal electronics for vital sign monitoring (Huang et al. 2012; Kim et al. 2011b; Yeo et al. 2013), brain-computer interfaces (Kim et al. 2010a; Viventi et al. 2011), electrocardiogram (ECG) mapping devices (Kim et al. 2012b; Viventi et al. 2010), and smart or minimally invasive surgical tools (Kim et al. 2011a, 2012e).

Figure 4 illustrates the use of biointegrated electronics for epidermal and in vivo physiological sensing. Electroencephalograph (EEG) measurements are

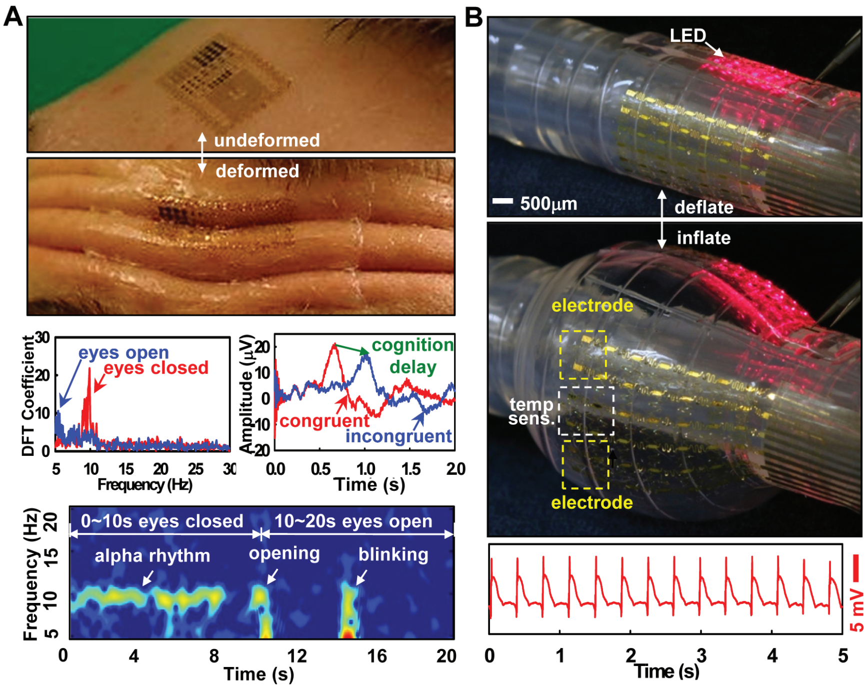


FIGURE 4 Biointegrated sensors based on stretchable electronics. (A) Ultrathin, ultrasoft epidermal electronic system laminated on a human forehead to read human electroencephalograph (EEG) (upper frames). Discrete Fourier transform coefficients of EEG alpha rhythms (middle left), demonstration of Stroop effects in EEG (middle right), and spectrogram of alpha rhythm (bottom). Reprinted from Kim et al. (2011b) with permission from the American Association for the Advancement of Science. (B) Multifunctional "instrumented" balloon catheter incorporating stretchable electrophysiological and radio frequency ablation electrodes, temperature sensors, pressure sensors, flow sensors, and arrays of microscale inorganic light-emitting diodes (μ-ILEDs) performing electrocardiogram recording of a rabbit heart. Reprinted from Kim et al. (2011a) with permission from Nature Publishing Group.

shown in Figure 4A, based on epidermal electronic systems laminated on a human forehead in a manner much like a temporary transfer tattoo, mechanically invisible to the wearer (Kim et al. 2011b). Because the attachment is enabled solely by van der Waals force without any conductive gels, these systems can function for more than two weeks at the exact same position without decomposition of the adhesives. Depending on where the electronic tattoo is placed, EEG, ECG, and EMG (electromyogram) measurements are possible with very high signal-to-noise ratio, thanks to the low impedance enabled by the intimate interface.

In addition to electrophysiological sensing, studies have successfully demonstrated the monitoring of skin temperature, mechanical deformation (strain), and hydration (Huang et al. 2012; Kim et al. 2011b; Yeo et al. 2013). Wireless power and data transmission coils as well as a stretchable battery (Xu et al. 2013) and stretchable memory patches (Son et al. 2013) further contribute to the standalone operation of wearable physiological sensors.

Soft electronics can integrate with not only human skin but also internal organs for in vivo monitoring. As an example, Figure 4B shows a multifunctional, “instrumented” balloon catheter that maintains a small initial diameter to travel through human veins and then inflates by 200 percent in cardiovascular cavities to perform minimally invasive surgeries such as the deployment of coronary stents. Electrodes and temperature, contact, and flow sensors integrated on the balloon skin provide in vivo endovascular and endocardial information, which used to be very difficult to obtain (Kim et al. 2011a).

Studies have also shown the effectiveness of other in vivo functionalities, such as epicardial ECG and beating amplitude sensing (Kim et al. 2012b; Viventi et al. 2010) as well as the mapping of brain activities (Kim et al. 2010a; Viventi et al. 2010, 2011).

Stimulation and Treatment

The most sophisticated version of biointegrated electronics will be a fully automated, closed-loop sensing-diagnosis-feedback device; the “feedback” that the device transmits will be information (e.g., a reminder to take medicine) or therapeutics (e.g., a pacemaker adjustment). Although the development of closed-loop biointegrated electronics is not yet fully realized, several types of stimulation and treatment are available.

One type involves the administration of a modulated electrical current to human skin to excite cutaneous mechanoreceptors, which provide instantaneous electrotactile feedbacks to the wearer in an acute and time-controlled manner (Warren et al. 2008). Figure 5A features a wearable finger tube that integrates high-performance inorganic electronics to sense finger-tip motion and provide electrotactile stimulation. The voltage-frequency combination to enable electrotactile sensation is shown in the right frame of Figure 5A (Ying et al. 2012).

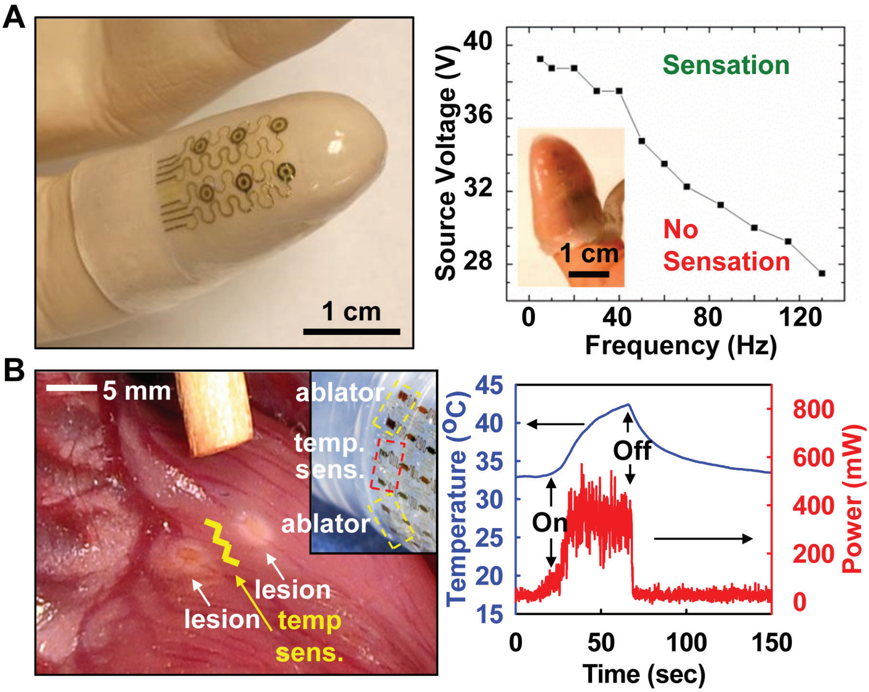


FIGURE 5 Biointegrated electrotactile stimulation and treatment tools based on stretchable electronics. (A) Wearable, conformable finger tube generates electrotactile sensation on human fingertip with suitably modulated current. Reprinted from Ying et al. (2012) with permission from IOP Publishing. (B) Lesions on a rabbit heart created by radio frequency (RF) ablators integrated on a balloon catheter (left frame). The supplied RF power and in situ tissue temperature measured by adjacent temperature sensors are shown in the right frame. Reprinted from Kim et al. (2011a) with permission from Nature Publishing Group.

As an in vivo example, Figure 5B shows lesions on a live rabbit heart, treated by radio frequency (RF) ablation as a therapeutic procedure to stop heart arrhythmia. The ablation was performed using stretchable electrodes on an inflatable balloon catheter (Kim et al. 2011a). Lesion size and depth can be determined with the use of in situ temperature monitoring during RF ablation (right frame). In vivo pretreatment sensing can provide critical information to guide treatment, and in situ posttreatment sensing can provide immediate data to evaluate treatment results and help guide the next treatment if any.

OUTLOOK

In the past decade, studies on mechanics, materials, and microfabrication techniques have advanced the design and manufacture of flexible and stretchable electronics, and it is likely that biointegrated electronics will soon revolutionize personal health care and human-machine interaction.

Further progress will likely depend on advances in the following areas. Maximization of the application potentials of wearable and implantable electronic systems will require the development of mechanically compatible and electronically sufficient microcontrollers, memory, power supply, and wireless data transmission modules. Multifunctional compliant systems that incorporate optical and biochemical tools would also be desirable. Another frontier of biointegrated electronics concerns transient electronics (Hwang et al. 2012). Roll-to-roll transfer printers for the deterministic assembly of inorganic semiconductors on polymer substrates hold the key for large-volume, low-cost manufacture of biointegrated electronics (Yang et al. 2012). More detailed discussion on the mechanics, materials, and functionalities of biointegrated electronics is available in several recent review articles (Kim et al. 2012a,c,d; Lu and Kim 2013).

REFERENCES

Bao ZN, Feng Y, Dodabalapur A, Raju VR, Lovinger AJ. 1997. High-performance plastic transistors fabricated by printing techniques. Chemistry of Materials 9:1299–1301.

Choi W, Song J, Khang D, Jiang H, Huang Y, Rogers J. 2007. Biaxially stretchable “wavy” silicon nanomembranes. Nano Letters 7:1655–1663.

Forrest SR. 2004. The path to ubiquitous and low-cost organic electronic appliances on plastic. Nature 428:911–918.

Forrest SR, Thompson ME. 2007. Introduction: Organic electronics and optoelectronics. Chemical Reviews 107:923–925.

Garnier F, Hajlaoui R, Yassar A, Srivastava P. 1994. All-polymer field-effect transistor realized by printing techniques. Science 265:1684–1686.

Gerdle B, Karlsson S, Day S, Djupsjöbacka M. 1999. Acquisition, processing and analysis of the surface electromyogram. In: Windhorst U, Johansson H, eds. Modern Techniques in Neuroscience Research. Berlin: Springer Verlag. pp. 705–755.

Huang X, Yeo WH, Liu YH, Rogers JA. 2012. Epidermal differential impedance sensor for conformal skin hydration monitoring. Biointerphases 7:1–9.

Hwang SW, Tao H, Kim DH, Cheng HY, Song JK, Rill E, Brenckle MA, Panilaitis B, Won SM, Kim YS, Song YM, Yu KJ, Ameen A, Li R, Su YW, Yang MM, Kaplan DL, Zakin MR, Slepian MJ, Huang YG, Omenetto FG, Rogers JA. 2012. A physically transient form of silicon electronics. Science 337:1640–1644.

Kaltenbrunner M, White MS, Glowacki ED, Sekitani T, Someya T, Sariciftci NS, Bauer S. 2012. Ultrathin and lightweight organic solar cells with high flexibility. Nature Communications 3:770.

Khang DY, Jiang HQ, Huang Y, Rogers JA. 2006. A stretchable form of single-crystal silicon for high-performance electronics on rubber substrates. Science 311:208–212.

Kim DH, Ahn JH, Choi WM, Kim HS, Kim TH, Song JZ, Huang YGY, Liu ZJ, Lu C, Rogers JA. 2008a. Stretchable and foldable silicon integrated circuits. Science 320:507–511.

Kim DH, Song JZ, Choi WM, Kim HS, Kim RH, Liu ZJ, Huang YY, Hwang KC, Zhang YW, Rogers JA. 2008b. Materials and noncoplanar mesh designs for integrated circuits with linear elastic responses to extreme mechanical deformations. Proceedings of the National Academy of Sciences U S A 105:18675–18680.

Kim DH, Kim YS, Wu J, Liu ZJ, Song JZ, Kim HS, Huang YGY, Hwang KC, Rogers JA. 2009. Ultrathin silicon circuits with strain-isolation layers and mesh layouts for high-performance electronics on fabric, vinyl, leather, and paper. Advanced Materials 21:3703–3709.

Kim DH, Viventi J, Amsden JJ, Xiao JL, Vigeland L, Kim YS, Blanco JA, Panilaitis B, Frechette ES, Contreras D, Kaplan DL, Omenetto FG, Huang YG, Hwang KC, Zakin MR, Litt B, Rogers JA. 2010a. Dissolvable films of silk fibroin for ultrathin conformal bio-integrated electronics. Nature Materials 9:511–517.

Kim DH, Lu NS, Ghaffari R, Kim YS, Lee SP, Xu LZ, Wu JA, Kim RH, Song JZ, Liu ZJ, Viventi J, de Graff B, Elolampi B, Mansour M, Slepian MJ, Hwang S, Moss JD, Won SM, Huang YG, Litt B, Rogers JA. 2011a. Materials for multifunctional balloon catheters with capabilities in cardiac electrophysiological mapping and ablation therapy. Nature Materials 10:316–323.

Kim DH, Lu NS, Ma R, Kim YS, Kim RH, Wang SD, Wu J, Won SM, Tao H, Islam A, Yu KJ, Kim TI, Chowdhury R, Ying M, Xu LZ, Li M, Chung HJ, Keum H, McCormick M, Liu P, Zhang YW, Omenetto FG, Huang YG, Coleman T, Rogers JA. 2011b. Epidermal electronics. Science 333:838–843.

Kim DH, Ghaffari R, Lu NS, Rogers JA. 2012a. Flexible and stretchable electronics for bio-integrated devices. Annual Review of Biomedical Engineering 14:113–128.

Kim DH, Ghaffari R, Lu NS, Wang SD, Lee SP, Keum H, D'Angelo R, Klinker L, Su YW, Lu CF, Kim YS, Ameen A, Li YH, Zhang YH, de Graff B, Hsu YY, Liu ZJ, Ruskin J, Xu LZ, Lu C, Omenetto FG, Huang YG, Mansour M, Slepian MJ, Rogers JA. 2012b. Electronic sensor and actuator webs for large-area complex geometry cardiac mapping and therapy. Proceedings of the National Academy of Sciences U S A 109:19910–19915.

Kim DH, Lu NS, Ghaffari R, Rogers JA. 2012c. Inorganic semiconductor nanomaterials for flexible and stretchable bio-integrated electronics. NPG Asia Materials 4:e15.

Kim DH, Lu NS, Huang YG, Rogers JA. 2012d. Materials for stretchable electronics in bioinspired and biointegrated devices. MRS Bulletin 37:226–235.

Kim DH, Wang SD, Keum H, Ghaffari R, Kim YS, Tao H, Panilaitis B, Li M, Kang Z, Omenetto F, Huang YG, Rogers JA. 2012e. Thin, flexible sensors and actuators as "instrumented" surgical sutures for targeted wound monitoring and therapy. Small 8(21):3263–3268.

Kim RH, Kim DH, Xiao JL, Kim BH, Park SI, Panilaitis B, Ghaffari R, Yao JM, Li M, Liu ZJ, Malyarchuk V, Kim DG, Le AP, Nuzzo RG, Kaplan DL, Omenetto FG, Huang YG, Kang Z, Rogers JA. 2010b. Waterproof AlInGaP optoelectronics on stretchable substrates with applications in biomedicine and robotics. Nature Materials 9:929–937.

Kim S, Wu JA, Carlson A, Jin SH, Kovalsky A, Glass P, Liu ZJ, Ahmed N, Elgan SL, Chen WQ, Ferreira PM, Sitti M, Huang YG, Rogers JA. 2010c. Microstructured elastomeric surfaces with reversible adhesion and examples of their use in deterministic assembly by transfer printing. Proceedings of the National Academy of Sciences U S A 107:17095–17100.

Ko HC, Stoykovich MP, Song JZ, Malyarchuk V, Choi WM, Yu CJ, Geddes JB, Xiao JL, Wang SD, Huang YG, Rogers JA. 2008. A hemispherical electronic eye camera based on compressible silicon optoelectronics. Nature 454:748–753.

Lee J, Wu JA, Shi MX, Yoon J, Park SI, Li M, Liu ZJ, Huang YG, Rogers JA. 2011. Stretchable GaAs photovoltaics with designs that enable high areal coverage. Advanced Materials 23:986–991.

Lipomi DJ, Tee BCK, Vosgueritchian M, Bao ZN. 2011. Stretchable organic solar cells. Advanced Materials 23:1771–1775.

Lu N, Kim DH. 2013. Flexible and stretchable electronics paving the way for soft robotics. Soft Robotics 1:53–62.

Mannsfeld SCB, Tee BCK, Stoltenberg RM, Chen CVHH, Barman S, Muir BVO, Sokolov AN, Reese C, Bao ZN. 2010. Highly sensitive flexible pressure sensors with microstructured rubber dielectric layers. Nature Materials 9:859–864.

Meitl MA, Zhu ZT, Kumar V, Lee KJ, Feng X, Huang YY, Adesida I, Nuzzo RG, Rogers JA. 2006. Transfer printing by kinetic control of adhesion to an elastomeric stamp. Nature Materials 5:33–38.

Menard E, Meitl MA, Sun YG, Park JU, Shir DJL, Nam YS, Jeon S, Rogers JA. 2007. Micro- and nanopatterning techniques for organic electronic and optoelectronic systems. Chemical Reviews 107:1117–1160.

Park SI, Xiong YJ, Kim RH, Elvikis P, Meitl M, Kim DH, Wu J, Yoon J, Yu CJ, Liu ZJ, Huang YG, Hwang K, Ferreira P, Li XL, Choquette K, Rogers JA. 2009. Printed assemblies of inorganic light-emitting diodes for deformable and semitransparent displays. Science 325:977–981.

Rogers JA, Bao Z, Baldwin K, Dodabalapur A, Crone B, Raju VR, Kuck V, Katz H, Amundson K, Ewing J, Drzaic P. 2001. Paper-like electronic displays: Large-area rubber-stamped plastic sheets of electronics and microencapsulated electrophoretic inks. Proceedings of the National Academy of Sciences U S A 98:4835–4840.

Rogers JA, Someya T, Huang YG. 2010. Materials and mechanics for stretchable electronics. Science 327:1603–1607.

Service RF. 2006. Materials science: Inorganic electronics begin to flex their muscle. Science 312:1593–1594.

Someya T, Sekitani T, Iba S, Kato Y, Kawaguchi H, Sakurai T. 2004. A large-area, flexible pressure sensor matrix with organic field-effect transistors for artificial skin applications. Proceedings of the National Academy of Sciences U S A 101:9966–9970.

Son D, Lee J, Qiao S, Ghaffari R, Kim J, Lee JE, Song C, Kim SJ, Lee DJ, Jun SW, Yang S, Park M, Shin J, Do K, Lee M, Kang K, Hwang CS, Lu N, Hyeon T, Kim D-H. 2013. Nanoparticle-embedded wearable memory, sensors and actuators for diagnosis and therapy of movement disorders. Submitted to Nature Nanotechnology.

Song YM, Xie Y, Malyarchuk V, Xiao J, Jung I, Choi KJ, Liu Z, Park H, Lu C, Kim RH, Li R, Crozier KB, Huang Y, Rogers JA. 2013. Digital cameras with designs inspired by the arthropod eye. Nature 497:95–99.

Sun JY, Lu NS, Yoon J, Oh KH, Suo ZG, Vlassak JJ. 2009. Inorganic islands on a highly stretchable polyimide substrate. Journal of Materials Research 24:3338–3342.

Sun YG, Choi WM, Jiang HQ, Huang YGY, Rogers JA. 2006. Controlled buckling of semiconductor nanoribbons for stretchable electronics. Nature Nanotechnology 1:201–207.

Takei K, Takahashi T, Ho JC, Ko H, Gillies AG, Leu PW, Fearing RS, Javey A. 2010. Nanowire active-matrix circuitry for low-voltage macroscale artificial skin. Nature Materials 9:821–826.

Viventi J, Kim DH, Moss JD, Kim YS, Blanco JA, Annetta N, Hicks A, Xiao JL, Huang YG, Callans DJ, Rogers JA, Litt B. 2010. A conformal, bio-interfaced class of silicon electronics for mapping cardiac electrophysiology. Science Translational Medicine 2:24ra22.

Viventi J, Kim DH, Vigeland L, Frechette ES, Blanco JA, Kim YS, Avrin AE, Tiruvadi VR, Hwang SW, Vanleer AC, Wulsin DF, Davis K, Gelber CE, Palmer L, Van der Spiegel J, Wu J, Xiao JL, Huang YG, Contreras D, Rogers JA, Litt B. 2011. Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. Nature Neuroscience 14:1599–1605.

Warren JP, Bobich LR, Santello M, Sweeney JD, Tillery SIH. 2008. Receptive field characteristics under electrotactile stimulation of the fingertip. IEEE Transactions on Neural Systems and Rehabilitation Engineering 16:410–415.

Webster JG. 2009. Medical Instrumentation: Application and Design. New York: Wiley.

Xu S, Zhang YH, Cho J, Lee J, Huang X, Jia L, Fan JA, Su YW, Su J, Zhang HG, Cheng HY, Lu BW, Yu CJ, Chuang C, Kim TI, Song T, Shigeta K, Kang S, Dagdeviren C, Petrov I, Braun PV, Huang YG, Paik U, Rogers JA. 2013. Stretchable batteries with self-similar serpentine interconnects and integrated wireless recharging systems. Nature Communications 4:1543.

Yang SY, Carlson A, Cheng HY, Yu QM, Ahmed N, Wu J, Kim S, Sitti M, Ferreira PM, Huang YG, Rogers JA. 2012. Elastomer surfaces with directionally dependent adhesion strength and their use in transfer printing with continuous roll-to-roll applications. Advanced Materials 24:2117–2122.

Yeo W-H, Kim Y-S, Lee J, Ameen A, Shi L, Li M, Wang S, Ma R, Jin SH, Kang Z, Huang Y, Rogers JA. 2013. Multifunctional epidermal electronics printed directly onto the skin. Advanced Materials 25(2):2773–2778.

Ying M, Bonifas AP, Lu NS, Su YW, Li R, Cheng HY, Ameen A, Huang YG, Rogers JA. 2012. Silicon nanomembranes for fingertip electronics. Nanotechnology 23:344004.

Yoon J, Baca AJ, Park SI, Elvikis P, Geddes JB, Li LF, Kim RH, Xiao JL, Wang SD, Kim TH, Motala MJ, Ahn BY, Duoss EB, Lewis JA, Nuzzo RG, Ferreira PM, Huang YG, Rockett A, Rogers JA. 2008. Ultrathin silicon solar microcells for semitransparent, mechanically flexible and microconcentrator module designs. Nature Materials 7:907–915.

Yoon J, Jo S, Chun IS, Jung I, Kim HS, Meitl M, Menard E, Li XL, Coleman JJ, Paik U, Rogers JA. 2010. GaAs photovoltaics and optoelectronics using releasable multilayer epitaxial assemblies. Nature 465:329–333.

Biocompatible Materials for Optoelectronic Neural Probes: Challenges and Opportunities

POLINA ANIKEEVA
Massachusetts Institute of Technology

The ability to understand and treat debilitating neurological conditions such as Parkinson's disease, spinal cord injury, and chronic pain is limited largely by the lack of materials and devices that can seamlessly interface with neurons and restore or bypass the malfunctioning neural circuits (Cogan et al. 2008; Gilja et al. 2011; Normann 2007). But the technology involved in deep brain and spinal cord stimulation devices used to treat such conditions dates back to the 1950s (Hamani and Temel 2012; Kringelbach et al. 2007). Even cutting-edge experiments that enable tetraplegic patients to control robotic aids (Hatsopoulos and Donoghue 2009; Hochberg et al. 2012) depend on devices invented more than 20 years ago (Campbell et al. 1991). These devices do not take into account the fundamental materials properties of neural tissue, and so their reliability and long-term effectiveness are diminished (Lee et al. 2005; Polikov et al. 2005).

Flexible organic and hybrid electronics offers a compelling solution to the elastic and surface chemistry mismatch between neural probes and neural tissues, while enabling novel approaches for neural interrogation. Recent developments in materials chemistry and fabrication methods make flexible electronics ripe for tailored, biointegrated neuroprosthetics.

In this article I review challenges and opportunities in the materials selected for neural probes and the role of flexible electronics and optoelectronics at the frontier of neural engineering.

BACKGROUND

Methods of Neural Stimulation and Recording

Devices for neural recording and stimulation interact with neural tissues with different degrees of precision and invasiveness (Buzsáki et al. 2012). For example, electroencephalography (EEG) is performed noninvasively through the skull and thus offers a low-resolution map of smoothed field potentials associated mainly with the neural activity of the whole cortical surface. Electrocorticography (ECoG), involving devices placed directly on the cortical surface, yields higher temporal and spatial resolution and is routinely used to identify seizure loci in epilepsy patients.

Neural systems exchange information in the form of action potentials—voltage spikes that propagate along neuronal membranes—and fluctuations in local field potentials (LFPs) averaged across a neuronal subnetwork or even an entire structure in the nervous system. Detailed mapping of neural activity is clinically relevant not only in the cortex but also in deep brain regions (e.g., the subthalamic nucleus in Parkinson's patients), the spinal cord, and peripheral nerves (e.g., in trauma patients or those in chronic pain). Moreover, many neurological disorders are associated with abnormal activity of specific types of neurons, and hence single-neuron resolution is essential to the development of effective therapies. I focus here on penetrating neural recording devices, designed to interface with individual cells in a particular region of the nervous system.

As with neural recording, neural stimulation offers varying degrees of precision and invasiveness. Noninvasive transcranial magnetic stimulation (TMS) allows for interrogation of cortical circuits via initiation of local flows of ions, which are hypothesized to cause changes in LFPs (Allen et al. 2007; Ridding and Rothwell 2007). However, there is currently no strategy for extending this approach to deep brain regions or targeting it to specific neuronal types because of the nonspecific nature and limited penetration depth of the low-frequency magnetic fields used in TMS.

In deep brain stimulation (DBS), an approved treatment for Parkinson's and essential tremor patients, high-voltage pulses (1–10 V; as compared to membrane voltages, ~30–100 mV, or LFPs, ~1–5 mV) are used to stimulate the neural tissue surrounding the electrodes (Perlmutter and Mink 2006). But although the DBS therapeutic effect is well documented, its underlying mechanisms remain unclear; both electrically induced excitation and inhibition of neural activity have been proposed (Kringelbach et al. 2007). Furthermore, nonspecific interrogation of large tissue volume often yields undesirable side effects such as depression or compulsive behaviors (Frank et al. 2007; Temel et al. 2007).

Epidural electrical stimulation (in the spinal cord of chronic pain patients) is essentially equivalent to DBS, with the key difference that the electrode leads are placed on top of the dura (the thin barrier that isolates nerves from other tissues) rather than deep in the neural tissue.

Development of Optogenetics

With the development of optogenetics it became possible to excite or inhibit specific neuronal types with millisecond precision (Boyden et al. 2005; Zhang et al. 2007). This method uses genetic targeting of light-sensitive proteins, opsins (of algal, archaeal, and bacterial origin), to establish neuronal sensitivity to a variety of visible light wavelengths. Opsins can be roughly categorized as excitatory (used for evoking action potentials; e.g., cation channel channelrhodopsin 2, ChR2) or inhibitory (used for inhibiting action potential firing; e.g., modified chloride pump halorhodopsin, eNpHR3.0, and modified proton pump archaerhodopsin, eArch3.0) (Zhang et al. 2011).

Optogenetics is a powerful tool for scientific investigation of the behavioral correlates of neural dynamics, but its genetic and mechanical invasiveness impedes its clinical translation (Yizhar et al. 2011). As mammalian tissues are highly scattering and absorptive in the visible light range, implantation of optical waveguides or light-emitting devices is necessary for implementation of optogenetics. Thus, optical stimulation technologies face materials design and biocompatibility challenges similar to those of tissue-penetrating neural recording and stimulation electrodes.

RELIABILITY CHALLENGES OF IMPLANTABLE NEURAL PROBES

Neural recording and stimulation devices have traditionally been fabricated out of hard materials with elastic moduli (Young's modulus E~10s–100s GPa[1]) exceeding those of neural tissues (E~kPa–MPa) (Borschel et al. 2003; Green MA et al. 2008) by many orders of magnitude. For example, neural recording and electrical stimulation electrodes (Figure 1) are often based on silicon (silicon multielectrode or "Utah arrays"; Bhandari et al. 2008; Campbell et al. 1991), multitrode probes (Blanche et al. 2005; Kipke et al. 2003; Seymour et al. 2011), silica (cone electrodes; Bartels et al. 2008; Kennedy et al. 1992), or metals (individual microwires of tungsten, gold, platinum, or platinum-iridium alloys; tetrodes and stereotrodes of nickel-chromium alloys; Gray et al. 1995; Jog et al. 2002; McNaughton et al. 1983). Similarly, optical stimulation in optogenetic experiments is most often performed with standard commercially available silica optical fibers (E~50–90 GPa) implanted directly into neural tissue.

It is hypothesized that this mismatch in stiffness contributes to tissue damage and the resulting encapsulation of devices in dense scars composed of glial cells, leading to a decrease in recording quality (Lee et al. 2005; Polikov et al. 2005). It is reasonable to assume that the probe insertion itself produces a certain amount of initial damage as well (destruction or displacement of cells in the path of the implant), an assumption that is supported by the commonly observed improve-

---

[1] GPa = gigapascals; kPa = kilopascals; MPa = megapascals.

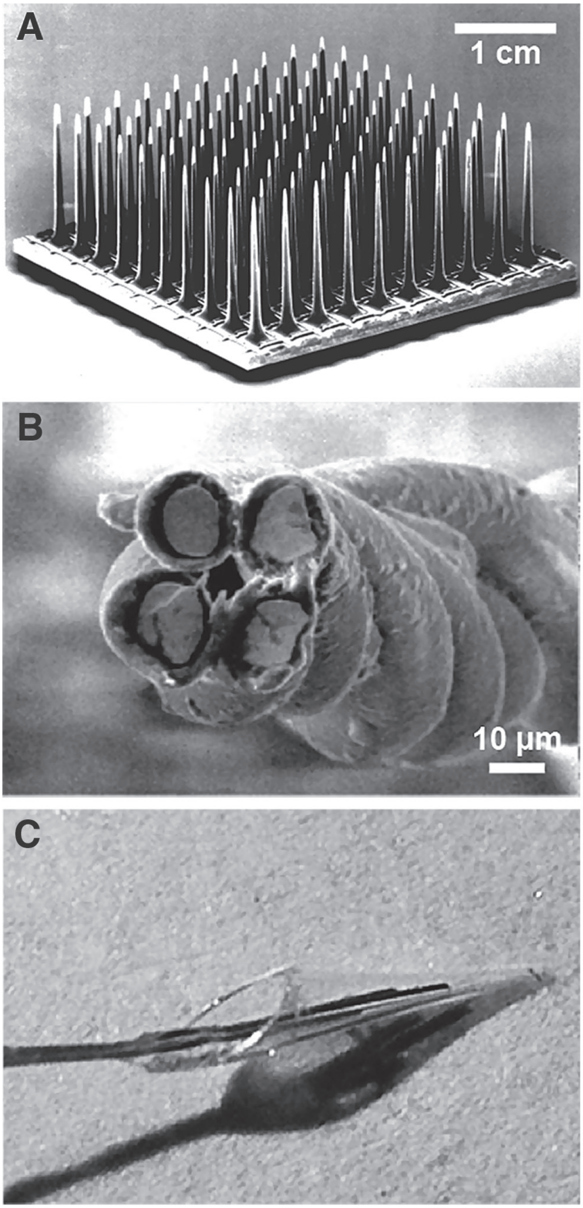


FIGURE 1 Examples of single unit and local field potential (LFP) recording devices commonly used in research. (A) Silicon multielectrode array (reprinted with permission from Blackrock Microsystems). (B) Tetrode microwire bundle (reprinted with permission from the University of Queensland). (C) Silica cone electrode. Source: Reprinted with permission from Bartels et al. (2008).

ment in recording quality approximately two weeks after implantation. However, the signal-to-noise ratio (SNR) and the total number of recorded neurons then decay steadily over the course of the implant lifespan.

Several mechanisms have been proposed to explain the neuronal death and glial scarring that compromise the probe's effectiveness. One hypothesis is that, as neural probes are generally at least partially fixed to the skull/vertebrae, their motion is constrained, whereas the neural tissues may shift by tens to hundreds of micrometers due to movement, heartbeat, and respiration (Britt and Rossi 1982; Muthuswamy et al. 2003). This micromotion of soft neural tissues around the hard implants is thought to introduce additional tissue damage.

Another theory is that the disruption of glial networks by devices larger than an average cell (i.e., >10 µm) may increase astrocytic and astroglial responses that lead to thickening of the scar tissue around the device (Seymour and Kipke 2007). Furthermore, devices with particularly sharp edges have also been shown to be disruptive to the blood-brain barrier, inducing an inflammatory response that raises glial activity and the likelihood of scarring (Saxena et al. 2013).

MATERIALS AND METHODS FOR FLEXIBLE SUBSTRATES

Flexible organic and hybrid electronics and optoelectronics offer opportunities to address the elastic, geometric, and chemical compatibility challenges of neural recording and stimulation devices.

Combining traditional metal and semiconductor technologies with flexible substrates provides a first transitional step toward stealthy bioinspired neural probes. Over the past decade polymer substrates have been used as a backing for metal and silicon-based neural recording electrodes. As illustrated in Figure 2, electrode arrays have been developed (using lithographic MEMS-inspired processing[2]) on silicone resins (poly(dimethylsulfoxane); PDMS), polyimide, and parylene C, to name a few (Kim BJ et al. 2013; Minev et al. 2012; Stieglitz et al. 2009; Viventi et al. 2011). Because these devices exhibit high flexibility and conformability to complex landscapes, they found immediate application in high-density microstructured cortical arrays (micro-ECoG) and nerve cuffs.

Contact printing methods developed by Rogers and colleagues have enabled highly innovative neural probes. This technology takes advantage of mature semiconductor-based (opto)electronics and combines it with flexible interconnects that enable transfer of circuit elements that are several microns thick onto polyimide and silk fibroin backing (Kim et al. 2010a, 2011). These flexible and foldable devices were recently introduced deep into the brain with the use of resorbable microneedles (Kim TI et al. 2013).

Meng and colleagues have taken an alternative approach by using a thermal molding process to produce soft cone electrodes based on parylene C, with active

---

[2] MEMS = microelectromechanical systems.

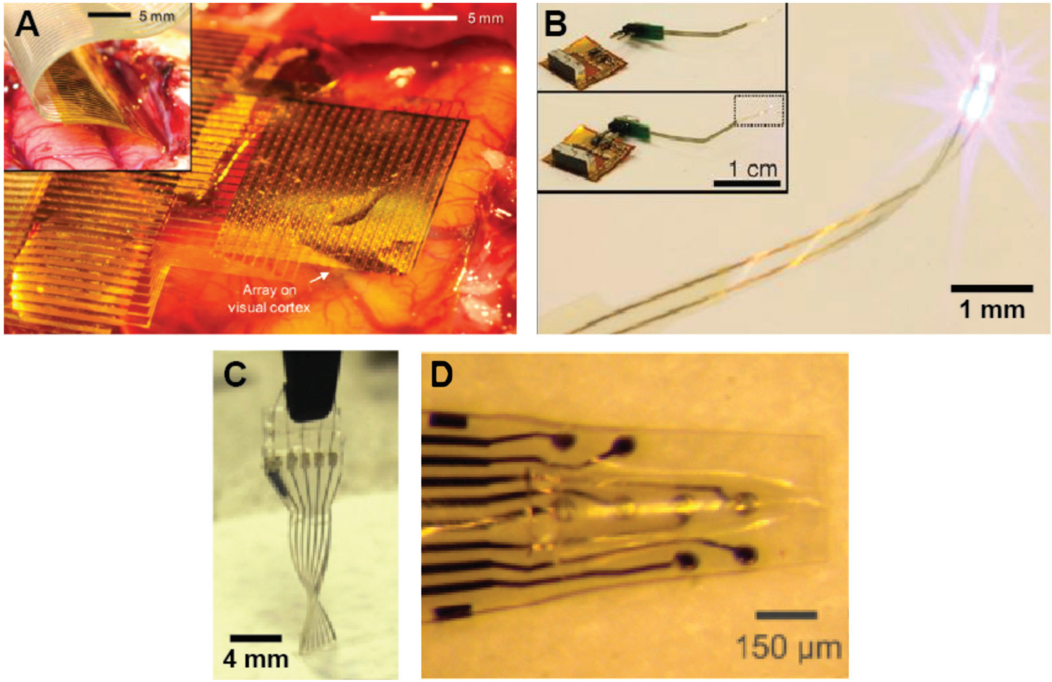


FIGURE 2 Examples of neural probes on flexible substrates. (A) Microprinted high-resolution microelectrocorticography array on polyimide substrate for cortical LFP mapping. Reprinted from Viventi et al. (2011) with permission from MacMillan Publishers Ltd. (B) Microprinted optoelectronic device on a polyimide substrate incorporating a gold electrode, a gallium nitride–based light-emitting diode, a silicon photodetector, and a resistor for temperature monitoring. For insertion device is adhered onto a silicon microneedle with silk fibroin. Reprinted from Kim TI et al. (2013) with permission from the American Association for the Advancement of Science. (C) Flexible microelectrode array on a PDMS substrate. Reprinted with permission from Minev et al. (2012). (D) Parylene C sheath electrode. Reprinted with permission from Kim BJ et al. (2013).

electrode pads facing inside the cone (Kim BJ et al. 2013; Tooker et al. 2004). This creative technology relies on earlier findings by Kennedy and colleagues (1992), who used silica capillaries seeded with nerve fragments to attract neuronal growth into the capillary containing an electrode, thus making a truly biointegrated device.

Yet there remain a number of challenges in the fabrication of neural probes on flexible substrates, such as relatively low resolution (dictated by contact printing methods), inability to scale to the high number of channels necessary for comprehensive mapping of brain activity, and inadequate capacity to interface with optical or drug delivery elements essential for neural interrogation (and potentially cell type identification). Robust reproducible manufacturing of probes suitable for use in human patients presents another challenge, as MEMS-style processing offers relatively low yield and is currently constrained to standard wafer sizes (several inches as compared to the several feet needed for a spinal cord).

SURFACE MODIFICATION AND ENCAPSULATION OF NEURAL PROBES

Because materials interfaces between devices and neural tissues play a critical role in both tissue response and the quality of neural recording, surface engineering is an important aspect of neural probe design. With their tunable chemical properties and low elastic moduli, organic materials offer a compelling toolbox for the engineering of intimate electrically and optically active interfaces between neurons and neural probes.

Surface Engineering

Polymers such as (poly(3,4-ethylenedioxythiophene); PEDOT) (Blau et al. 2011; Ludwig et al. 2011; Richardson-Burns et al. 2007), polylysine (Boehler et al. 2012; Hai et al. 2010), and polypyrrole (Abidian et al. 2010; George et al. 2005) have been shown to boost the reliability and SNR of neural recording electrodes by promoting cell adhesion and reducing the impedance of equivalent circuits between the devices and the neuronal membranes.

Hydrogels based on polymers and polymer blends of natural (agarose, alginate, xyloglucan, hyaluronan, methylcellulose, chitosan, and matrigel) and synthetic (methacrylate, polyethylene glycol (PEG), poly(vinyl alcohol), and poly(acrylic acid)) origins are used for most neural regeneration scaffolds (Frampton et al. 2011; Hanson Shepherd et al. 2011; Jhaveri et al. 2008; Nisbet et al. 2008; Seliktar et al. 2012; Shin et al. 2012) and have recently found application in surface modification of neural probes (Jun et al. 2008; Kim et al. 2010b; Lu et al. 2009). The advantages of hydrogels include elastic moduli comparable to those of the neural tissues as well as high permeability for nutrients and oxygen.

However, the electrical and optical properties of these soft gels have not yet been engineered for improved neural recording and stimulation. Consequently their application in neural probe engineering has been restricted to providing low-modulus biocompatible buffers, which may reduce the damage associated with micromotion.

Encapsulation

Encapsulation is another form of surface modification routinely used during deep-tissue implantation of flexible neural probes. As mentioned above, flexible substrates make it possible to overcome the elastic modulus mismatch between an electronic or optoelectronic probe and the surrounding neural tissue. But it is difficult to target soft devices to a specific region of the nervous system because they are prone to buckling, which hampers straight-line penetration.

Dissolvable encapsulation temporarily stiffens the probe to permit targeted implantation. Organic and biopolymeric materials such as PEG, sugar, tyrosine-based polymers, and silk fibroin are often used because of their adjustable dissolution speed in aqueous environments as well as their versatile chemistry and biocompatibility. Silk fibroin (also used as a biocompatible adhesive) enables the introduction of PDMS-backed probes through silicon microneedles that are retracted shortly after implantation upon dissolution of the silk fibroin.

OPPORTUNITIES WITH POLYMER OPTOELECTRONICS

Two decades of advances in materials chemistry have propelled small-molecule organic optoelectronics into commercial applications in the display industry and beyond, but the sensitivity of these materials to environmental moisture and oxygen hinders their application in the human body. In contrast, environmentally stable polymers and polymer composites with versatile chemical and electronic properties and low elastic moduli present a promising materials system for the development of multifunctional tissue interfaces.

Despite their wide adoption throughout the medical community (orthopedic implants, encapsulation materials for stimulation electrodes, porous scaffolds for soft tissue regeneration), polymers have yet to be fully explored with respect to their applications in neural probes (Green RA et al. 2008). Pioneering studies by Martin and Kipke, among others, illustrate the potential of PEDOT, polypyrrole, and polymer-carbon composites (Abidian et al. 2010; Kozai et al. 2012) (Figure 3) to solve the elastic mismatch of neural recording devices while reducing the overall electrode impedance and thus increasing SNR. In parallel, Capadona and Tyler have applied biologically inspired design principles to create polymer composites with controllable elastic properties that mimic sea cucumber dermis (Capadona et al. 2008; Harris et al. 2011).

Despite growing evidence of the utility of polymers in neural probe design, various engineering challenges prevent widespread adoption of these materials systems by neuroscientists and clinicians. For example, polymer probes are pri-

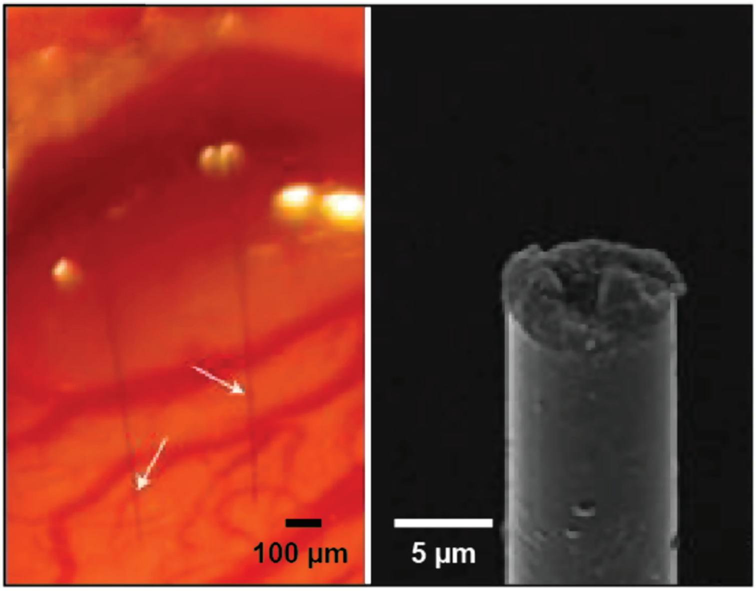


FIGURE 3 Examples of polymer and organic/inorganic composite neural recording electrodes. Left: Carbon-composite microelectrodes chemical vapor deposition (CVD)-coated with polyxylene. Right: Tip electrochemically coated with poly(3,4-ethylene-dioxythiophene) (PEDOT). Source: Reprinted from Kozai et al. (2012) with permission from Nature Publishing Group.

marily fabricated by electrospinning, chemical vapor deposition, thin film spin-casting, and lithography. The first two methods offer relatively low throughput and require painstaking postsynthesis assembly if multiple electrodes are desired, which is true for most neuroprosthetic applications. Furthermore, these methods currently do not allow for integration of optical elements, which are essential for neural stimulation applications. Although well-developed lithographic methods allow for integration of multiple functional elements, they are limited by the flat substrate geometry, which is not ideal for applications in deep brain regions.

In my laboratory we have recently explored a thermal drawing process (TDP) inspired by optical fiber production as a means of fabrication for multifunctional neural probes. During TDP a macroscale preform, which can be fabricated using low-end mechanical processing, is drawn into a fiber with microscale features (Abouraddy et al. 2007; Bayindir et al. 2004; Goff 2002; Varshneya 1994). The lateral dimensions are scaled by as much as 10,000-fold using, if necessary, multiple drawing steps, enabling the creation of structures on the nanometer scale without the need for high-resolution fabrication technology (Kaufman et al. 2011; Yaman et al. 2011). At the same time, the length is stretched by a factor of ~100, yielding hundreds of meters of fibrous devices with a conserved cross-sectional pattern.

Because TDP faithfully reproduces the cross-sectional geometry of the macroscopic preform, it enables the creation of sophisticated multifunctional structures on the microscale. In addition, it is compatible with a wide range of materials with varying optical and electrical properties, permitting, for example, the combination of waveguide core and cladding materials, conductive polymer composites, and low-melting-temperature metal microwires in a single device.

We have used TDP to produce a range of fiber-inspired neural probes (FINPs), from high-channel-count neural recording arrays of arbitrary lengths to multifunctional devices incorporating waveguides, drug delivery channels, and neural recording electrodes (Figure 4).

Our preliminary in vivo evaluation of FINPs suggests that TDP may provide a scalable fabrication tool for flexible optoelectronic devices compatible with implantation in a variety of regions of the nervous system. Furthermore, this process may complement recent materials discoveries by Martin, Capadona, Kipke, and others as it may not only enable the integration of these innovative polymer systems into multifunctional probes but also offer a pathway toward their high-throughput production.

CONCLUSION

High-fidelity recording and stimulation of neural activity are essential to the development of neuroprosthetic devices as well as to the mapping of neural circuits involved in neurological and neuromuscular disorders. While mature semiconductor technologies provided initial promise for neural probe design, recent tissue engineering studies illustrate the need for alternative biocompatible materials platforms. In this article I have reviewed the challenges of established neural probe technologies and the opportunities of flexible organic and hybrid materials platforms for improvements in the biocompatibility and longevity of these sensors. I have also emphasized the importance of integration of optical neural stimulation modules and discussed fabrication approaches that may enable flexible multifunctional neural prosthetics.

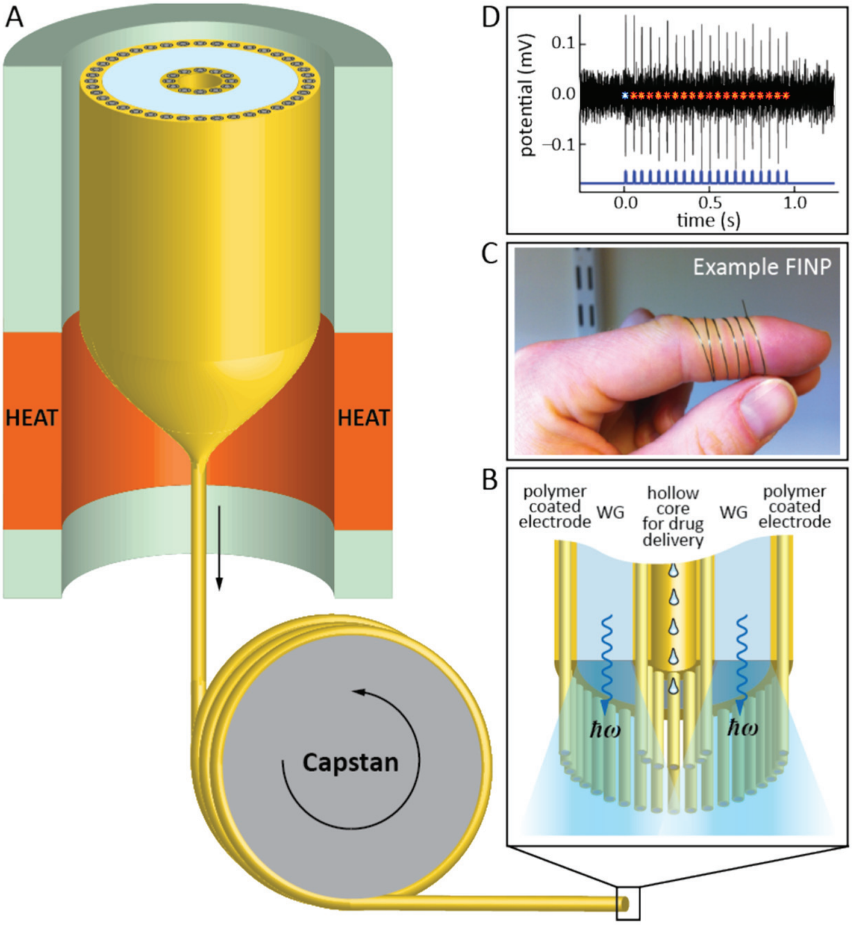


FIGURE 4 (A) Thermal drawing process (TDP) applied to fiber-inspired neural probe (FINP) fabrication. (B) Longitudinal cross section of a FINP for recording, optical stimulation, and drug delivery. (C) Sample FINP. (D) Optically evoked action potentials recorded with a FINP in the medial prefrontal cortex of a transgenic Thy1-ChR2-YFP mouse expressing ChR2 in a broad neuronal population. WG = waveguide.

REFERENCES

Abidian MR, Corey JM, Kipke DR, Martin DC. 2010. Conducting-polymer nanotubes improve electrical properties, mechanical adhesion, neural attachment, and neurite outgrowth of neural electrodes. Small 6(3):421–429.

Abouraddy AF, Bayindir M, Benoit G, Hart SD, Kuriki K, Orf N, Shapira O, Sorin F, Temelkuran B, Fink Y. 2007. Towards multimaterial multifunctional fibres that see, hear, sense and communicate. Nature Materials 6:336–347.

Allen EA, Pasley BN, Duong T, Freeman RD. 2007. Transcranial magnetic stimulation elicits coupled neural and hemodynamic consequences. Science 317(5846):1918–1921.

Bartels J, Andreasen D, Ehirim P, Mao H, Seibert S, Wright EJ, Kennedy P. 2008. Neurotrophic electrode: Method of assembly and implantation into human motor speech cortex. Journal of Neuroscience Methods 174(2):168–176.

Bayindir M, Sorin F, Abouraddy AF, Viens J, Hart SD, Joannopoulos JD, Fink Y. 2004. Metal-insulator-semiconductor optoelectronic fibres. Nature 431(7010):826–829.

Bhandari R, Negi S, Rieth L, Normann RA, Solzbacher F. 2008. A novel method of fabricating convoluted shaped electrode arrays for neural and retinal prostheses. Sensors and Actuators A: Physical 145–146:123–130.

Blanche TJ, Spacek MA, Hetke JF, Swindale NV. 2005. Polytrodes: High-density silicon electrode arrays for large-scale multiunit recording. Journal of Neurophysiology 93:2987–3000.

Blau A, Murr A, Wolff S, Sernagor E, Medini P, Iurilli G, Ziegler C, Benfenati F. 2011. Flexible, all-polymer microelectrode arrays for the capture of cardiac and neuronal signals. Biomaterials 32(7):1778–1786.

Boehler MD, Leondopulos SS, Wheeler BC, Brewer GJ. 2012. Hippocampal networks on reliable patterned substrates. Journal of Neuroscience Methods 203(2):344–353.

Borschel GH, Kia KF, Kuzon WM Jr, Dennis RG. 2003. Mechanical properties of acellular peripheral nerve. Journal of Surgical Research 114(2):133–139.

Boyden ES, Zhang F, Bamberg E, Nagel G, Deisseroth K. 2005. Millisecond-timescale, genetically targeted optical control of neural activity. Nature Neuroscience 8(9):1263–1268.

Britt RH, Rossi GT. 1982. Quantitative analysis of methods for reducing physiological brain pulsations. Journal of Neuroscience Methods 6(3):219–229.

Buzsáki G, Anastassiou CA, Koch C. 2012. The origin of extracellular fields and currents: EEG, ECoG, LFP and spikes. Nature Reviews Neuroscience 13(6):407–420.

Campbell PK, Jones KE, Huber RJ, Horch KW, Normann RA. 1991. A silicon-based, three-dimensional neural interface: Manufacturing processes for an intracortical electrode array. IEEE Transactions in Biomedical Engineering 38:758–768.

Capadona JR, Shanmuganathan K, Tyler DJ, Rowan SJ, Weder C. 2008. Stimuli-responsive polymer nanocomposites inspired by the sea cucumber dermis. Science 319(5868):1370–1374.

Cogan SF. 2008. Neural stimulation and recording electrodes. Annual Review of Biomedical Engineering 10(1):275–309.

Frampton JP, Hynd MR, Shuler ML, Shain W. 2011. Fabrication and optimization of alginate hydrogel constructs for use in 3D neural cell culture. Biomedical Materials 6(1):015002.

Frank MJ, Samanta J, Moustafa AA, Sherman SJ. 2007. Hold your horses: Impulsivity, deep brain stimulation, and medication in parkinsonism. Science 318(5854):1309–1312.

George PM, Lyckman AW, LaVan DA, Hegde A, Leung Y, Avasare R, Testa C, Alexander PM, Langer R, Sur M. 2005. Fabrication and biocompatibility of polypyrrole implants suitable for neural prosthetics. Biomaterials 26(17):3511–3519.

Gilja V, Chestek CA, Diester I, Henderson JM, Deisseroth K, Shenoy KV. 2011. Challenges and opportunities for next-generation intracortically based neural prostheses. IEEE Transactions on Biomedical Engineering 58(7):1891–1899.

Goff D. 2002. Fiber Optic Reference Guide. Woburn MA: Focal Press.

Gray CM, Maldonado PE, Wilson M, McNaughton B. 1995. Tetrodes markedly improve the reliability and yield of multiple single-unit isolation from multi-unit recordings in cat striate cortex. Journal of Neuroscience Methods 63:43–54.

Green MA, Bilston LE, Sinkus R. 2008. In vivo brain viscoelastic properties measured by magnetic resonance elastography. NMR in Biomedicine 21(7):755–764.

Green RA, Lovell NH, Wallace GG, Poole-Warren LA. 2008. Conducting polymers for neural interfaces: Challenges in developing an effective long-term implant. Biomaterials 29(24):3393–3399.

Hai A, Shappir J, Spira ME. 2010. In-cell recordings by extracellular microelectrodes. Nature Methods 7:200–202.

Hamani C, Temel Y. 2012. Deep brain stimulation for psychiatric disease: Contributions and validity of animal models. Science Translational Medicine 4(142):1–12.

Hanson Shepherd JN, Parker ST, Shepherd RF, Gillette MU, Lewis JA, Nuzzo RG. 2011. 3D microperiodic hydrogel scaffolds for robust neuronal cultures. Advanced Functional Materials 21(1):47–54.

Harris JP, Capadona JR, Miller RH, Healy BC, Shanmuganathan K, Rowan SJ, Weder C, Tyler DJ. 2011. Mechanically adaptive intracortical implants improve the proximity of neuronal cell bodies. Journal of Neural Engineering 8(6):066011.

Hatsopoulos NG, Donoghue JP. 2009. The science of neural interface systems. Annual Review of Neuroscience 32(1):249–266.

Hochberg LR, Bacher D, Jarosiewicz B, Masse NY, Simeral JD, Vogel J, Haddadin S, Liu J, Cash SS, van der Smagt P, Donoghue JP. 2012. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. Nature 485(7398):372–375.

Jhaveri SJ, Hynd MR, Dowell-Mesfin N, Turner JN, Shain W, Ober CK. 2008. Release of nerve growth factor from HEMA hydrogel-coated substrates and its effect on the differentiation of neural cells. Biomacromolecules 10(1):174–183.

Jog MS, Connolly CI, Kubota Y, Iyengar DR, Garrido L, Harlan R, Graybiel AM. 2002. Tetrode technology: Advances in implantable hardware, neuroimaging, and data analysis techniques. Journal of Neuroscience Methods 117:141–152.

Jun SB, Hynd MR, Dowell-Mesfin NM, Al-Kofahi Y, Roysam B, Shain W, Kim SJ. 2008. Modulation of cultured neural networks using neurotrophin release from hydrogel-coated microelectrode arrays. Journal of Neural Engineering 5(2):203.

Kaufman JJ, Tao G, Shabahang S, Deng DS, Fink Y, Abouraddy AF. 2011. Thermal drawing of high-density macroscopic arrays of well-ordered sub-5-nm-diameter nanowires. Nano Letters 11(11):4768–4773.

Kennedy PR, Mirra SS, Bakay RA. 1992. The cone electrode: Ultrastructural studies following long-term recording in rat and monkey cortex. Neuroscience Letters 142(1):89–94.

Kim BJ, Kuo JT, Hara SA, Lee CD, Yu L, Gutierrez CA, Hoang TQ, Pikov V, Meng E. 2013. 3D Parylene sheath neural probe for chronic recordings. Journal of Neural Engineering 10(4):045002.

Kim DH, Viventi J, Amsden JJ, Xiao J, Vigeland L, Kim YS, Blanco JA, Panilaitis B, Frechette ES, Contreras D, Kaplan DL, Omenetto FG, Huang Y, Hwang KC, Zakin MR, Litt B, Rogers JA. 2010a. Dissolvable films of silk fibroin for ultrathin conformal bio-integrated electronics. Nature Materials 9:511–517.

Kim DH, Wiler JA, Anderson DJ, Kipke DR, Martin DC. 2010b. Conducting polymers on hydrogel-coated neural electrode provide sensitive neural recordings in auditory cortex. Acta Biomaterialia 6(1):57–62.

Kim DH, Lu N, Ma R, Kim YS, Kim RH, Wang S, Wu J, Won SM, Tao H, Islam A, Yu KJ, Kim TI, Chowdhury R, Ying M, Xu L, Li M, Chung HJ, Keum H, McCormick M, Liu P, Zhang YW, Omenetto FG, Huang Y, Coleman T, Rogers JA. 2011. Epidermal electronics. Science 333:838–343.

Kim TI, McCall JG, Jung YH, Huang X, Siuda ER, Li Y, Song J, Song YM, Pao HA, Kim RH, Lu C, Lee SD, Song IS, Shin G, Al-Hasani R, Kim S, Tan MP, Huang Y, Omenetto FG, Rogers JA, Bruchas MR. 2013. Injectable, cellular-scale optoelectronics with applications for wireless optogenetics. Science 340(6129):211–216.

Kipke DR, Vetter RJ, Williams JC, Hetke JF. 2003. Silicon-substrate intracortical microelectrode arrays for long-term recording of neuronal spike activity in cerebral cortex. IEEE Transactions on Neural Systems and Rehabilitation Engineering 11:151–155.

Kozai TD, Langhals NB, Patel PR, Deng X, Zhang H, Smith KL, Lahann J, Kotov NA, Kipke DR. 2012. Ultrasmall implantable composite microelectrodes with bioactive surfaces for chronic neural interfaces. Nature Materials 11(12):1065–1073.

Kringelbach ML, Jenkinson N, Owen SLF, Aziz TZ. 2007. Translational principles of deep brain stimulation. Nature Reviews Neuroscience 8(8):623–635.

Lee H, Bellamkonda RV, Sun W, Levenston ME. 2005. Biomechanical analysis of silicon microelectrode-induced strain in the brain. Journal of Neural Engineering 2:81–89.

Lu Y, Wang D, Li T, Zhao X, Cao Y, Yang H, Duan YY. 2009. Poly(vinyl alcohol)/poly(acrylic acid) hydrogel coatings for improving electrode-neural tissue interface. Biomaterials 30(25):4143–4151.

Ludwig KA, Langhals NB, Joseph MD, Richardson-Burns SM, Hendricks JL, Kipke DR. 2011. Poly(3,4-ethylenedioxythiophene) (PEDOT) polymer coatings facilitate smaller neural recording electrodes. Journal of Neural Engineering 8(1):014001.

McNaughton BL, O'Keefe J, Barnes CA. 1983. The stereotrode: A new technique for simultaneous isolation of several single units in the central nervous system from multiple unit records. Journal of Neuroscience Methods 8:391–397.

Minev IR, Chew DJ, Delivopoulos E, Fawcett JW, Lacour SP. 2012. High sensitivity recording of afferent nerve activity using ultra-compliant microchannel electrodes: An acute in vivo validation. Journal of Neural Engineering 9(2):026005.

Muthuswamy J, Gilletti A, Jain T, Okandan M. 2003. Microactuated neural probes to compensate for brain micromotion. Engineering in Medicine and Biology Society, Proceedings of the 25th Annual International Conference of the IEEE.

Nisbet DR, Crompton KE, Horne MK, Finkelstein DI, Forsythe JS. 2008. Neural tissue engineering of the CNS using hydrogels. Journal of Biomedical Materials Research Part B: Applied Biomaterials 87B(1):251–263.

Normann RA. 2007. Technology Insight: Future neuroprosthetic therapies for disorders of the nervous system. Nature Clinical Practice Neurology 3(8):444–452.

Perlmutter JS, Mink JW. 2006. Deep brain stimulation. Annual Review of Neuroscience 29:229–257.

Polikov VS, Tresco PA, Reichert WM. 2005. Response of brain tissue to chronically implanted neural electrodes. Journal of Neuroscience Methods 148(1):1–18.

Richardson-Burns SM, Hendricks JL, Martin DC. 2007. Electrochemical polymerization of conducting polymers in living neural tissue. Journal of Neural Engineering 4(2):L6.

Ridding MC, Rothwell JC. 2007. Is there a future for therapeutic use of transcranial magnetic stimulation? Nature Reviews Neuroscience 8(7):559–567.

Saxena T, Karumbaiah L, Gaupp EA, Patkar R, Patil K, Betancur M, Stanley GB, Bellamkonda RV. 2013. The impact of chronic blood-brain barrier breach on intracortical electrode function. Biomaterials 34(20):4703–4713.

Seliktar D. 2012. Designing cell-compatible hydrogels for biomedical applications. Science 336(6085):1124–1128.

Seymour JP, Kipke DR. 2007. Neural probe design for reduced tissue encapsulation in CNS. Biomaterials 28:3594–3607.

Seymour J, Langhals N, Anderson D, Kipke D. 2011. Novel multi-sided, microelectrode arrays for implantable neural applications. Biomedical Microdevices 13(3):441–451.

Shin Y, Han S, Jeon JS, Yamamoto K, Zervantonakis IK, Sudo R, Kamm RD, Chung S. 2012. Microfluidic assay for simultaneous culture of multiple cell types on surfaces or within hydrogels. Nature Protocols 7(7):1247–1259.

Stieglitz T, Rubehn B, Henle C, Kisban S, Herwik S, Ruther P, Schuettler M. 2009. Brain-computer interfaces: An overview of the hardware to record neural signals from the cortex. Progress in Brain Research 175:297–315.

Temel Y, Boothman LJ, Blokland A, Magill PJ, Steinbusch HW, Visser-Vandewalle V, Sharp T. 2007. Inhibition of 5-HT neuron activity and induction of depressive-like behavior by high-frequency stimulation of the subthalamic nucleus. Proceedings of the National Academy of Sciences U S A 104(43):17087–17092.

Tooker A, Meng E, Erickson J, Tai YC, Pine J. 2004. Development of biocompatible parylene neurocages. Engineering in Medicine and Biology Society: 26th Annual International Conference of the IEEE, San Francisco.

Varshneya AK. 1994. Fundamentals of Inorganic Glasses. San Diego: Academic Press.

Viventi J, Kim DH, Vigeland L, Frechette ES, Blanco JA, Kim YS, Avrin AE, Tiruvadi VR, Hwang SW, Vanleer AC, Wulsin DF, Davis K, Gelber CE, Palmer L, Van der Spiegel J, Wu J, Xiao J, Huang Y, Contreras D, Rogers JA, Litt B. 2011. Flexible, foldable, actively multiplexed, high-density electrode array for mapping brain activity in vivo. Nature Neuroscience 14:1599–1605.

Yaman M, Khudiyev T, Ozgur E, Kanik M, Aktas O, Ozgur EO, Deniz H, Korkut E, Bayindir M. 2011. Arrays of indefinitely long uniform nanowires and nanotubes. Nature Materials 10:494–501.

Yizhar O, Fenno LE, Davidson TJ, Mogri M, Deisseroth K. 2011. Optogenetics in neural systems. Neuron 71(1):9–34.

Zhang F, Wang LP, Brauner M, Liewald JF, Kay K, Watzke N, Wood PG, Bamberg E, Nagel G, Gottschalk A, Deisseroth K. 2007. Multimodal fast optical interrogation of neural circuitry. Nature 446(7136):633–639.

Zhang F, Vierock J, Yizhar O, Fenno LE, Tsunoda S, Kianianmomeni A, Prigge M, Berndt A, Cushman J, Polle J, Magnuson J, Hegemann P, Deisseroth K. 2011. The microbial opsin family of optogenetic tools. Cell 147(7):1446–1457.

APPENDIXES

Contributors

Polina Anikeeva is the AMAX Assistant Professor in Materials Science and Engineering at the Massachusetts Institute of Technology. She takes a materials science approach to neural prosthesis by developing implantable and minimally invasive devices for direct recording and stimulation of neural activity. She employs principles of optoelectronics and magnetics to create functional micro- and nano-structured interfaces with individual neurons.

Kristi Anseth is a Howard Hughes Medical Institute Investigator and distinguished professor of chemical and biological engineering at the University of Colorado at Boulder. Her research interests lie at the interface between biology and engineering where she designs new biomaterials for applications in drug delivery and regenerative medicine. She is an elected member of the National Academy of Engineering and the Institute of Medicine.

Halil Berberoglu is an assistant professor in the Department of Mechanical Engineering at the University of Texas at Austin where his work focuses on developing technologies for producing renewable fuels from solar energy, in particular algal biofuel and solar hydrogen production technologies. He performs fundamental studies on photosynthesis, transport phenomena, and concentrated solar energy, as well as systems-level analysis for energetic, economic, and environmental sustainability of these technologies.

Tanzeem Choudhury is an associate professor in Computing and Information Sciences at Cornell University. She directs the People Aware Computing group, which develops mobile sensing and learning systems that track behavioral and vocal indicators of mental and physical health.

Laura Díaz Anadón is an assistant professor of technology innovation and public policy at the John F. Kennedy School of Government at Harvard University. Her research in innovation in energy and environmental technologies seeks to quantify the benefits that derive from policies designed to promote innovation, map the complex factors that contribute to it, and create tools to manage its uncertainties. She also focuses on the design of innovation institutions and the connections between water and energy systems.

Dragan Djurdjanovic is an associate professor in the Department of Mechanical Engineering at the University of Texas at Austin. His work involves functions of anomaly detection, isolation of the source of anomalous behavior, fault characterization, and fault compensation in systems of interacting dynamic systems. Applications include multistage manufacturing systems such as automotive assembly lines or semiconductor lithography overlay processes, automotive engine systems, and engine-generator sets for power generation.

Steve Ellet is the vice president for supply chain design at CHAINalytics. He applies optimization and simulation to large-scale supply chain design for industries with complex networks such as consumer packaged goods, process, retail, and high-tech.

Antonio Facchetti is the chief technology officer of Polyera Inc. in Skokie, Illinois. His research interests include organic semiconductors and dielectrics for thin-film transistors, conducting polymers, molecular electronics, organic second- and third-order nonlinear optical materials, and organic photovoltaics.

Elizabeth Hoegeman is the fuel system worldwide manufacturing leader at Cummins Inc. Her areas of research include product and process design of cost-effective and sustainable green precision machining, assembly, and testing, and product performance.

Tony Jebara is an associate professor in the Department of Computer Science at Columbia University and chair of the Foundations of Data Science Center in the Institute for Data Science and Engineering. His primary interests are machine learning, social networks, graphs, spatio-temporal data, mobile data, and computer vision. He has advised and co-founded startup companies that focus on big data.

Scott Klemmer is an associate professor of cognitive science and computer science and engineering at the University of California, San Diego where his research is on human-computer interaction. His group's research tools gather and synthesize examples to empower more people to design interactive systems, program, learn new skills, and work creatively.

Yueh-Lin (Lynn) Loo is Theodora D. '78 & William H. Walton III '74 Professor in the Department of Chemical Engineering at Princeton University. Her interests are in materials chemistry and physics of electrically active polymeric and molecular materials. Her group elucidates the fundamental processing-structure-property relationships of these materials in order to develop innovative processing and patterning technologies for low-cost, lightweight, and mechanically flexible thin-film devices such as organic transistors and solar cells.

Nanshu Lu is an assistant professor in the Department of Aerospace and Engineering Mechanics at the University of Texas at Austin. Her research focuses on the mechanics, materials, and microfabrication of flexible, stretchable, and bio-integrated electronics through theoretical, numerical, and experimental approaches.

Rhett Mayor is an associate professor in the George W. Woodruff School of Mechanical Engineering at the Georgia Institute of Technology. His work is primarily in manufacturing and heat transfer, combustion, and energy systems; micro-factories and micro/meso-scale manufacturing processes; integrated micro-mechatronics; and micro-engines and micro-power generation.

Rob Miller is a professor in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. His work lies at the intersection of programming and human computer interaction (HCI): making programming easier for end users (web end-user programming) and more productive for professionals (HCI for software developers), and making people part of the programming system itself (crowd computing and human computation).

Miguel Modestino is a postdoctoral fellow in the School of Engineering at the École Polytechnique Fédérale de Lausanne.

Douglas Muzyka is senior vice president and chief science and technology officer at DuPont. He joined the company in 1985 as a research scientist and has held a variety of research and research management roles in North America and abroad. He was named to his current position in 2010.

Tse Nga (Tina) Ng is a senior research scientist in the Electronic Materials and Devices Laboratory at Palo Alto Research Center (PARC). She develops mechanically flexible electronics such as large-area photosensor arrays for x-ray medical imaging and nonvolatile memory for sensors. Prior to PARC, her research involved the development of pico-newton force measurement techniques such as cantilever magnetometry and electric force microscopy in order to study phenomena in organic semiconductors.

Willem Rensink is part of the Shell GameChanger team in the Innovation and R&D Department of Shell Projects and Technology (US). GameChanger identifies and incubates revolutionary ideas to proof of concept by partnering with Shell employees and external entrepreneurs or university researchers and funding early stage, revolutionary ideas. The ideas address challenges in the traditional oil and gas industry or create opportunities for future energy operations.

Rachel Segalman is a professor in the Department of Chemical Engineering at the University of California, Berkeley. She develops polymers for energy applications, including ion-conducting membranes for solar fuel devices as well as thermoelectric and electronic active layers for a variety of devices.

Steven Skerlos is the Arthur F. Thumau Professor in the Department of Mechanical Engineering at the University of Michigan where his research focuses on knowledge and technology systems that minimize the environmental consequences of engineering design and manufacturing systems. This objective is being pursued primarily in the automotive, electric power, water/wastewater treatment, and manufacturing sectors.

Kate Starbird is an assistant professor in the Department of Human Centered Design and Engineering at the University of Washington. Her research, which is in the field of human-computer interaction, examines online interaction and emergent collaboration during crisis events (e.g., natural disasters, mass protests, etc.). Specifically, she studies ways in which the "crowd" contributes to emergency and humanitarian response efforts.

Stuart Thomas is global technology and licensing manager in industrial biosciences at DuPont. His research is focused on technology development and commercialization of cellulosic ethanol.

Duncan Watts is a principal researcher at Microsoft where he examines the collective dynamics of large-scale social systems applied to issues such as cooperation, organization problem-solving, diffusion of information, and systemic risk.

Christian Will is the chief technology officer for the DELMIA brand at Dassault Systèmes. His work brings business intelligence, social media, mobility, and cloud computing technologies into a process-centric "application composition" platform. The goal is to merge the virtual world of modeling and simulation with the real world of execution, thereby accelerating business transformation to new product-based economics, flexible factories, and demand-driven value networks.

Joyce Yang is a technology manager in the Bioenergy Technologies Office of the US Department of Energy. Her work is in microbial deconstruction and biotransformation of biomass components into fuels, chemicals, and power.

Program

NATIONAL ACADEMY OF ENGINEERING

2013 US Frontiers of Engineering Symposium
September 19-21, 2013

Chair: Kristi Anseth, University of Colorado, Boulder

DESIGNING AND ANALYZING SOCIETAL NETWORKS

Organizers: Tanzeem Choudhury and Scott Klemmer

Modeling Large-Scale Networks
Tony Jebara

Crowd Computing
Rob Miller

Crowds, Crisis, and Convergence: Crowdsourcing in the Context of Disasters
Kate Starbird

Computational Social Science: Exciting Progress and Future Directions
Duncan Watts

***

COGNITIVE MANUFACTURING

Organizers: Elizabeth Hoegeman and Rhett Mayor

Distributed Anomaly Detection for Timely Fault Remediation in Modern Manufacturing
Dragan Djurdjanovic

Business Process Management Systems to Optimize Manufacturing
Christian Will

The Rise of Computer-Enabled Supply Chain Design
Steve Ellet

Advancing Sustainable Manufacturing with the Use of Cognitive Agents
Steven Skerlos

***

ENERGY: REDUCING OUR DEPENDENCE ON FOSSIL FUELS

Organizers: Halil Berberoglu and Stuart Thomas

Energy from Fossil Fuels: Challenges and Opportunities for Technology Innovation
Laura Díaz Anadón

Bioenergy Technologies and Strategies: A New Frontier
Joyce Yang

Drivers for Successful Biofuel Production Scale-up
Willem Rensink

Artificial Solar Fuel Generators
Rachel Segalman

***

FLEXIBLE ELECTRONICS

Organizers: Yueh-Lin (Lynn) Loo and Tse Nga (Tina) Ng

Materials and Process Engineering for Printed and Flexible Optoelectronic Devices
Antonio Facchetti

Mechanics, Materials, and Functionalities of Biointegrated Electronics
Nanshu Lu

Biocompatible Materials for Optoelectronic Neural Probes: Challenges and Opportunities
Polina Anikeeva

***

DINNER SPEECH

Doug Muzyka

Participants

NATIONAL ACADEMY OF ENGINEERING

2013 US Frontiers of Engineering Symposium
September 19-21, 2013

Hal Alper
Assistant Professor
Department of Chemical Engineering
University of Texas at Austin

Polina Anikeeva**
AMAX Assistant Professor in Materials Science and Engineering
Department of Materials Science and Engineering
Massachusetts Institute of Technology

Kristi Anseth*
Distinguished Professor of Chemical and Biological Engineering, Professor of Surgery, and HHMI Assistant Investigator
Department of Chemical and Biological Engineering
University of Colorado, Boulder

Burcin Becerik-Gerber
Assistant Professor and Stephen Schrank Early Career Chair in Civil and Environmental Engineering
Department of Civil and Environmental Engineering
University of Southern California

Halil Berberoglu*
Assistant Professor
Department of Mechanical Engineering
University of Texas at Austin

Irene Beyerlein
Technical Staff Member
Theoretical Division
Los Alamos National Laboratory

Marcie Black
Director Chief, Technology Officer, and President
Bandgap Engineering

---

*Organizing Committee
**Speaker

Mark Borden
Assistant Professor, Nicholas Rome Faculty Fellow
Department of Mechanical Engineering
University of Colorado, Boulder

Bryan Boudouris
Assistant Professor
School of Chemical Engineering
Purdue University

John Carpenter
Research Chemical Engineer
Center for Energy Technology
RTI International

Julie Champion
Assistant Professor
School of Chemical and Biomolecular Engineering
Georgia Institute of Technology

Yiran Chen
Assistant Professor
Department of Electrical and Computer Engineering
University of Pittsburgh

Tanzeem Choudhury*
Associate Professor
Department of Information Science
Cornell University

Karen Christman
Associate Professor
Department of Bioengineering
University of California, San Diego

Jason Clevenger
Principal
Polymers Science and Materials Chemistry
Exponent, Inc.

Baratunde Cola
Assistant Professor
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Douglas Densmore
Assistant Professor
Departments of Biomedical Engineering and Electrical and Computer Engineering
Boston University

Laura Díaz Anadón**
Assistant Professor in Technology Innovation and Public Policy
John F. Kennedy School of Government
Harvard University

Dragan Djurdjanovic**
Associate Professor
Department of Mechanical Engineering
University of Texas at Austin

Andreas Dreher
Section Head
Corporate Function R&D
Procter & Gamble Company

Jennifer Dy
Associate Professor
Department of Electrical and Computer Engineering
Northeastern University

Steve Ellet**
Vice President
Supply Chain Design
CHAINalytics

Carolyn Ellinger
Device Physicist
Kodak Technology Center
Eastman Kodak Company

Laura Espinal
Materials Research Engineer
Material Measurement Laboratory
National Institute of Standards and Technology

Antonio Facchetti**
Chief Technology Officer
Polyera Inc.

Philip Feng
Assistant Professor
Department of Electrical Engineering and Computer Science
Case Western Reserve University

Matthew Gaston
Director, SEI Emerging Technology Center
Software Engineering Institute
Carnegie Mellon University

Dennice Gayme
Assistant Professor
Department of Mechanical Engineering
Johns Hopkins University

Elizabeth Gerber
Assistant Professor
Department of Mechanical Engineering
Northwestern University

Ramon Gonzalez
Associate Professor
Departments of Chemical and Biomolecular Engineering and Bioengineering
Rice University

Angela Harris
Research Engineer
Materials Engineering
Ford Motor Company

Davion Hill
Principal Engineer
Research and Innovation
DNV

Elizabeth Hoegeman*
World Wide Manufacturing Leader
Cummins Fuel Systems
Cummins Inc.

Prashant Jain
Assistant Professor
Department of Chemistry
University of Illinois at Urbana Champaign

Tony Jebara**
Associate Professor of Computer Science
Institute for Data Sciences and Engineering
Columbia University

Sham Kakade
Senior Research Scientist
Microsoft Research New England

Mandakini Kanungo
Member, Research and Technology Staff
Xerox Innovation Group
Xerox Webster Research Center

Karen Kennedy
Process Engineer
Emerging Technology
Air Products and Chemicals

Darby Kimball
Deputy Facility Manager, Plutonium Facility
Nuclear Materials Technology Program
Lawrence Livermore National Laboratory

Scott Klemmer*
Associate Professor
Department of Computer Science and Engineering
University of California, San Diego

Mykel Kochenderfer
Technical Staff
Lincoln Laboratory
Massachusetts Institute of Technology

Ken Laberteaux
Senior Principal Scientist
Future Mobility Research Department
Toyota Research Institute North America

Gert Lanckriet
Associate Professor
Department of Electrical and Computer Engineering
University of California, San Diego

Robert Larsen
Process Engineering Specialist
Science and Technology
Dow Corning Corporation

Janice Li
Principal Project Manager
New York Office–Transportation
CH2M HILL

C. Michael Lindsay
Technical Advisor
Energetic Materials Branch
Air Force Research Laboratory

Jennifer Logue
Research Scientist/Engineer
Environmental Energy Technologies Division
Lawrence Berkeley National Laboratory

Yueh-Lin (Lynn) Loo*
Professor
Department of Chemical Engineering
Princeton University

Nanshu Lu**
Assistant Professor
Department of Aerospace Engineering and Engineering Mechanics
University of Texas at Austin

James Luedtke
Assistant Professor
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

Carmel Majidi
Assistant Professor
Department of Mechanical Engineering
Carnegie Mellon University

J. Rhett Mayor*
Associate Professor
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

John McCloy
Associate Professor
School of Mechanical and Materials Engineering
Washington State University

Thomas McGuire
Aeronautical Engineer, Compact Fusion Inventor and Team Lead
Advancement Development Program
Lockheed Martin Aeronautics

Rob Miller**
Professor
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Meredith Morris
Senior Researcher
Microsoft Research

Philipp Mueller
Technical Supervisor
Titanium Technologies
DuPont

Samrat Mukherjee
Chemical Engineering Consultant
DuPont Engineering Research and Technology

Vidhya Navalpakkam
Research Scientist
Strategic Technologies
Google

Tse Nga (Tina) Ng*
Senior Research Scientist
Electronic Materials and Devices Laboratory
Palo Alto Research Center

Alexander Orlov
Assistant Professor
Department of Materials Science and Engineering
Stony Brook University

John Owens
Associate Professor
Department of Electrical and Computer Engineering
University of California, Davis

Bo Pang
Senior Research Scientist
Strategic Technologies
Google

Ah-Hyung (Alissa) Park
Lenfest Junior Professor in Applied Climate Science
Departments of Earth and Environmental Engineering and Chemical Engineering
Columbia University

Leila Parsa
Associate Professor
Department of Electrical, Computer and Systems Engineering
Rensselaer Polytechnic Institute

Marco Pavone
Assistant Professor
Department of Aeronautics and Astronautics
Stanford University

Ashley (Ash) Peterson
Principal R&D Engineer
Aortic-Thoracic Development, Endovascular Therapy
Medtronic

Mina Rais-Zadeh
Assistant Professor
Department of Electrical Engineering and Computer Science
University of Michigan

Sundar Ramamurthy
General Manager, Corporate Vice President
Metal Deposition Products
Applied Materials

Narayan Ramesh
Associate R&D Director
Dow Solar Solutions
Dow Chemical Company

Jose Emmanuel Ramirez-Marquez
Associate Professor
School of Systems and Enterprises
Stevens Institute of Technology

Patrick Reed
Professor
School of Civil and Environmental Engineering
Cornell University

Travis Reine
Process Engineering Supervisor
R&S Refining Americas
ExxonMobil Corp.

Willem Rensink**
GameChanger
Innovation and R&D
Shell Projects and Technology (US)

Carol Rose
Engineering Specialist
Transportation and Infrastructure Division
STV

Eric Ruggiero
Manager
Turbine Heat Transfer Technologies Laboratory
GE Global Research Center

Tuhin Sahai
Staff Research Scientist
Systems Dynamics and Optimization
United Technologies Research Center

Jay Sayre
Director of Internal Research and Development
Energy, Health, and Environment
Battelle Memorial Institute

Jon Schuller
Assistant Professor
Department of Electrical and Computer Engineering
University of California, Santa Barbara

Rachel Segalman**
Professor
Department of Chemical Engineering
University of California, Berkeley

Hae-Jong Seo
Senior Engineer
Corporate R&D
Qualcomm Technologies, Inc.

Jian Sheng
Associate Professor and Whitacre Endowed Chair
Department of Mechanical Engineering
Texas Tech University

Steven Skerlos**
Arthur F. Thurnau Professor
Department of Mechanical Engineering
University of Michigan

Weidong Song
Fabrication and Structural Analysis Engineer
Boeing Commercial Airplanes
Boeing Company

Kate Starbird**
Assistant Professor
Department of Human Centered Design and Engineering
University of Washington

Daniel Steingart
Assistant Professor
Department of Mechanical and Aerospace Engineering
Princeton University

Stuart Thomas*
Global Technology & Licensing Manager-DuPont Cellulosic Ethanol
Industrial Biosciences
DuPont

Bret Thomson
Mechanical Engineer
Maritime Systems Division
Space and Naval Warfare Systems Center-Pacific

Brian Thurow
W. Allen and Martha Reed Associate Professor
Department of Aerospace Engineering
Auburn University

Alberto Valdes-Garcia
Research Staff Member, Manager
Communication Technologies
IBM T.J. Watson Research Center

Brett Van Horn
Global Project Leader for Refrigerants Development
Fluorochemicals R&D
Arkema Inc.

Subhas Venayagamoorthy
Associate Professor and Borland Chair of Hydraulics
Department of Civil and Environmental Engineering
Colorado State University

Subramaniam Venkatraman
Research Scientist
R&D
Fitbit

Duncan Watts**
Principal Researcher
Microsoft Research NYC

Christian Will**
Chief Technology Officer
DELMIA
Dassault Systèmes

Pawel Woelke
Senior Associate
Applied Science
Weidlinger Associates, Inc.

Robert Worl
Lead Electrophysics Engineer
Phantom Works
Boeing Company

Chongjin Xie
Distinguished Member of Technical Staff
Transmission Systems and Networking Research
Bell Labs, Alcatel-Lucent

Joyce Yang**
Technology Manager
Bioenergy Technologies Office
US Department of Energy

Chong Wing Yung
Research Scientist
Agilent Labs-Molecular Tools
Agilent Technologies

Xuanhe Zhao
Assistant Professor
Department of Mechanical Engineering and Materials Science
Duke University

Dinner Speaker

Doug Muzyka
Senior Vice President and Chief Science & Technology Officer
DuPont

Guests

William Hayden
Vice President
The Grainger Foundation

Sohi Rastegar
Director
Emerging Frontiers in Research and Innovation
National Science Foundation

Phil Szuromi
Senior Editor
Science Magazine

National Academy of Engineering

C. Daniel Mote, Jr.
President

Lance A. Davis
Executive Officer

Janet Hunziker
Senior Program Officer

Vanessa Lester
Program Associate