Frontiers of Engineering 2011: Reports on Leading-Edge Engineering from the 2011 Symposium

National Academy of Engineering of the National Academies

ISBN
978-0-309-22143-6

174 pages
6 x 9
PAPERBACK (2012)

FRONTIERS OF ENGINEERING

Reports on Leading-Edge Engineering from the 2011 Symposium

NATIONAL ACADEMY OF ENGINEERING
OF THE NATIONAL ACADEMIES

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

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

Funding for the activity that led to this publication was provided by Google, The Grainger Foundation, Air Force Office of Scientific Research, Department of Defense ASD(R&E)–Research, Microsoft Research, and Cummins Inc. This material is also based upon work supported by the National Science Foundation under Grant No. 1108785. 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.

International Standard Book Number-13: 978-0-309-22143-6
International Standard Book Number-10: 0-309-22143-9

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

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

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

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

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

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

www.national-academies.org

ORGANIZING COMMITTEE

ANDREW M. WEINER (*Chair*), Scifres Family Distinguished Professor of Electrical and Computer Engineering, Purdue University
TIMOTHY DENISON, Director of Neuroengineering, Medtronic
ALEKSANDAR KUZMANOVIC, Lisa Wissner Slivka and Benjamin Slivka Chair in Computer Science and Assistant Professor, Department of Electrical Engineering and Computer Science, Northwestern University
ANNIE PEARCE, Associate Professor, Myers-Lawson School of Construction, Virginia Tech
CAROLYN SEEPERSAD, Assistant Professor, Department of Mechanical Engineering, University of Texas at Austin
MICHAEL SIEMER, President, Mydea Technologies
AMARNAG SUBRAMANYA, Research Scientist, Google
JUSTIN WILLIAMS, Associate Professor, Departments of Biomedical Engineering and Neurological Surgery, University of Wisconsin-Madison
ZHIQIANG (JOHN) ZHAI, Associate Professor, Department of Civil, Environmental, and Architectural Engineering, University of Colorado at Boulder

Staff

JANET R. HUNZIKER, Senior Program Officer
ELIZABETH WEITZMANN, Program Associate

Preface

This volume highlights the papers presented at the National Academy of Engineering's 2011 U.S. Frontiers of Engineering Symposium. Every year, the symposium brings together 100 outstanding young leaders in engineering to share their cutting-edge research and technical work. The 2011 symposium was held September 19-21, and hosted by Google at their headquarters in Mountain View, California. Speakers were asked to prepare extended summaries of their presentations, which are reprinted here. The intent of this book is to convey the excitement of this unique meeting and to highlight cutting-edge developments in engineering research and technical work.

GOALS OF THE FRONTIERS OF ENGINEERING PROGRAM

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

At the 2-½-day U.S. Frontiers of Engineering Symposium, 100 of this country's best and brightest engineers, ages 30 to 45, have an opportunity to learn from their peers about pioneering work being done in many areas of engineering. The symposium gives early career engineers from a variety of institutions in academia, industry, and government, and from many different engineering disciplines, an opportunity to make contacts with and learn from individuals they would not meet in the usual round of professional meetings. This networking may lead to collaborative work and facilitate the transfer of new techniques and approaches. It is hoped that the exchange of information on current developments in many fields of engineering will lead to insights that may be applicable in specific disciplines and thereby build U.S. innovative capacity.

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

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

THE 2011 SYMPOSIUM

The four general topics covered at the 2011 meeting were: additive manufacturing, semantic processing, engineering sustainable buildings, and neuroprosthetics. The additive manufacturing session described how technologies such as stereolithography, fused deposition modeling, 3D printing, selective laser melting, laser-engineered net shape processes, ultrasonic consolidation, and selective laser sintering enable layer-wise fabrication of complex parts directly from CAD files without part-specific tooling. The presentations included an overview of additive manufacturing processes and their impact on industrial practice and academic research, a description of applications in the aerospace and medical fields, and a discussion of the challenges and application frontiers of additive manufacturing.

The explosion of content on the Internet and its growth as a source of information requires a deep understanding of Web content. Semantic processing, the topic of the second session, refers to high-level information understanding tasks such as inferring author sentiment in a particular piece of writing; searching through documents, images, and videos; or translating text into different languages. Because natural language and images constitute the majority of the data on the Internet, presenters described semantic processing algorithms that advance understanding of word and sentence meaning, relationships, and sentiment; use collaboratively generated content to represent the semantics of natural language; and improve search for images and video as well as plots, graphs, and diagrams.

The engineering of sustainable buildings was the focus of the third session. This is an area where architects, engineers, and those in the construction industry work together to create buildings that are more energy-efficient, have fewer adverse environmental impacts, and provide healthier indoor environments. The four speakers focused on cutting-edge benchmarking for building performance and life-cycle-cost assessment, tools that execute more efficient and effective design processes, multi-scale modeling for the design of new or renovation of old buildings with sustainability in mind, and use of location-based services and social networks to drive market transformation for sustainable building.

The symposium concluded with the session on neuroprosthetics, which are devices that interface with the nervous system. These technologies are used to stimulate the nervous system in order to restore sensory function or to elicit motor intention from the brain for artificial prostheses. Talks covered state-of-the-art research and clinical studies of retinal implants and how the emerging field of optogenetics may address some of the challenges of bioelectronic approaches, the current status of brain-computer interfaces (BCI) and how research in cortical physiology may enhance BCI-based control, and efforts to understand the biology of how the brain processes information in order to apply that knowledge to next-generation neuroprosthetic applications such as cochlear implants.

In addition to the plenary sessions, the participants had many opportunities to engage in informal interactions. On the first afternoon of the meeting, participants broke into small groups for "get-acquainted" sessions during which individuals presented short descriptions of their work and answered questions from their colleagues. This helped attendees get to know more about each other relatively early in the program. On the second afternoon, there were presentations by Google staff on current work in translation, speech recognition, optical character recognition, machine perception, and audio signal processing.

Every year, a distinguished engineer addresses the participants at dinner on the first evening of the symposium. The speaker this year was Dr. Alfred Z. Spector, vice president of research and special initiatives at Google, who gave a talk on the evolution of computer science. This talk can be viewed at *www.naefrontiers.org*.

NAE is deeply grateful to the following organizations for their support of the 2011 U.S. Frontiers of Engineering Symposium:

- Google
- The Grainger Foundation
- Air Force Office of Scientific Research
- Department of Defense ASD(R&E)-Research
- National Science Foundation. This material is based upon work supported by the National Science Foundation under grant number EFRI-1108785.
- Microsoft Research
- Cummins Inc.

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

Contents

APPENDIXES

ADDITIVE MANUFACTURING

Introduction

CAROLYN SEEPERSAD
University of Texas at Austin

MICHAEL SIEMER
Mydea Technologies Corp.

Additive manufacturing technologies enable layer-wise fabrication of complex parts directly from CAD files without part-specific tooling. Examples of additive manufacturing technologies include stereolithography, fused deposition modeling, 3D printing, selective laser melting, laser engineered net shape processes, ultrasonic consolidation, and selective laser sintering. Selective laser sintering, for example, fabricates parts in a layer-wise manner by selectively fusing powdered material in regions defined by the part's cross-sectional geometry. Additive manufacturing offers many strategic advantages, including increased design freedom for building complex internal and external part geometries that cannot be made in any other way and the abilities to rapidly iterate through design permutations, build functional parts in small lot sizes for end-user customization or bridge manufacturing, and repair expensive parts for aerospace and other industries.

Brent Stucker (University of Louisville) begins with an overview of additive manufacturing processes and their impact on industrial practice and academic research. He provides insight on the basic principles of additive manufacturing, the frontiers of our capabilities for fabricating functional parts, and the impact that additive manufacturing is having on design and manufacturing. Next, Brett Lyons (Boeing) describes a number of additive manufacturing's applications in the aerospace industry. Some of its industry-changing capabilities include lightweighting via part reduction and honeycomb-like structures and development of flight-ready materials with aerospace levels of repeatability and reliability.

Additive manufacturing has the potential to revolutionize the medical industry by fabricating implants, prosthetics, orthotics, and other devices that are customized for an individual user's body. For example, Walter Reed Army Medical

Center is currently fabricating customized cranial implants for injured soldiers, with titanium implants customized for a specific patient's cranial profile and injury. In the third paper, Andrew Christensen (Medical Modeling Inc.) discusses medical applications of additive manufacturing.

The final paper by Hod Lipson (Cornell University) describes the challenges involved in designing parts for additive manufacturing, including the need for advanced design techniques and tools that can tailor not only the shape of additively manufactured objects but also their composition and functionality. He discusses some of the application frontiers of additive manufacturing, including biological, culinary, and mechatronics applications.

Additive Manufacturing Technologies: Technology Introduction and Business Implications

Brent Stucker
University of Louisville

INTRODUCTION

Additive manufacturing (AM) technologies have finally hit the mainstream. After 25 years of development as "rapid-prototyping" techniques, the industry is transforming into a manufacturing-focused enterprise. Today, when you walk the trade show floors and attend industry conferences, there are more than just engineers and manufacturers walking around. You also bump shoulders with investment bankers, venture capitalists, and do-it-yourself enthusiasts. A recent *National Geographic* video has gone viral on the Internet, showing the potential for 3D printing of a crescent wrench. So many people have seen the video and forwarded its link via email that some recipients assumed it must be a viral hoax, and *snopes.com* (the email hoax-tracking website) had to confirm the reality of "3D printing." Within the past year the *Economist*, the *New York Times*, *Standard & Poor's*, *Wired* magazine, and others have all written about 3D printing techniques.

So what is the hype all about? Many believe additive manufacturing is revolutionary and has the potential to transform manufacturing in the same way that Web 2.0 transformed cyberspace (the transition of the Internet to user-driven content, such as Facebook, Twitter, LinkedIn, etc.). As consumers begin to see the potential for creating their own manufactured goods using additive manufacturing, Factory 1.0 (production of physical goods by companies and manufacturing experts) will transform to Factory 2.0 (production of physical goods by consumers), thus permanently transforming the manufacturing landscape.

ADDITIVE MANUFACTURING

Additive manufacturing techniques are a collection of manufacturing processes that join materials to make physical 3D objects directly from virtual 3D computer data. These processes typically build up parts layer by layer, as opposed to subtractive manufacturing methodologies, which create 3D geometry by removing material in a sequential manner. In 2009, after more than 20 years of confusing terminology, the ASTM International F42 Committee on Additive Manufacturing Technologies defined additive manufacturing as the "process of joining materials to make objects from 3D model data, usually layer upon layer, as opposed to subtractive manufacturing methodologies." These technologies were also called rapid prototyping, direct digital manufacturing, solid freeform fabrication, additive fabrication, additive layer manufacturing, and other similar technology names over the years. In the technical community, an international consensus has coalesced around the use of "additive manufacturing," whereas in the popular press the technologies are known as "3D printing."

Every existing commercial AM machine works in a similar way. First a 3D computer-aided design (CAD) file is sliced into a stack of two-dimensional planar layers. These layers are built by the AM machine and stacked one after the other to build up the part (Figure 1). Today, there are seven different approaches to AM, and dozens of variants of these approaches. As most of these approaches were first patented in the late 1980s and early 1990s, in many cases the fundamental process patents have expired or are expiring soon—thus opening up the marketplace for significant competition in a way that was impossible over the past 20 years due to intellectual property exclusivity.

The remainder of this paper provides an overview on the seven different approaches to AM, followed by a discussion of the business trends and opportunities afforded by AM techniques. The breakdown of AM processes into the following seven categories is based upon the work of the Terminology Subcommittee of the ASTM F42 committee. At the time of the writing of this paper this categorization is being balloted, and thus the final names of these categories are subject to change as the standards-development consensus process proceeds.

Material Jetting

Material jetting is the use of inkjet printers or other similar techniques to deposit droplets of build material that are selectively dispensed through a nozzle or orifice to build up a three-dimensional structure. In most cases these droplets are made up of photopolymers or wax-like materials to form parts or investment casting patterns, respectively. These processes are truly 3D-printing machines, as they use inkjet and other "printing" techniques to build up three-dimensional structures.

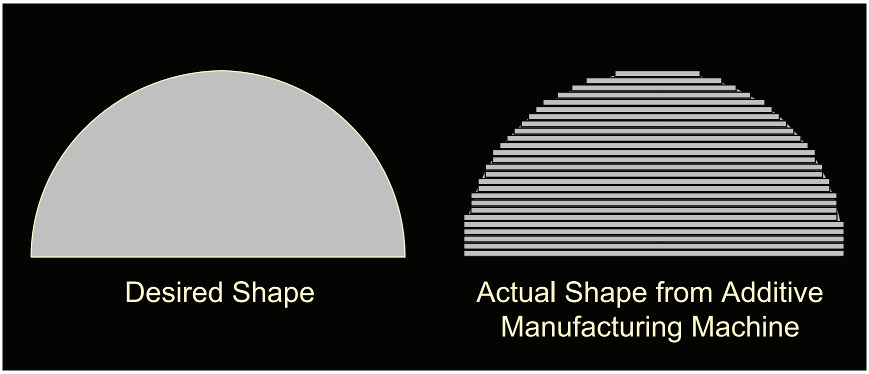


FIGURE 1 Layered approach to AM.

Photopolymers are useful materials for material jetting because they transform from a liquid to a solid in the presence of light. Photopolymers can be tuned to cross-link and harden in response to different wavelengths of light, and for AM they typically transform in the visible or ultraviolet wavelength ranges.

Material-jetting techniques often use multiple arrays of printheads to print different materials. The most common reason for printing two materials is for one of the materials to be used as the "build" material, while the second material is used as a "support" material. For 3D geometry that includes channels, voids, or overhanging structures, a support must be built below any overhanging surfaces (as droplets have to land on something to keep them in a fixed location; see Figure 2). When a secondary support material is used, a water-soluble material is commonly used so that the supports can be removed by immersing the part in a water-based liquid.

Material jetting is capable of printing multimaterial and gradient-material structures. Applications of multimaterial parts range from parts with controlled hardness and flexibility to parts with differing electrical properties in various regions to tissue-engineered structures with different biological properties in different regions of the part.

Binder Jetting

Binder-jetting techniques also use nozzles to print material, but instead of printing with the build material, the printed material is "glue," which holds powder together in the desired shape. A binder-jetting process starts by first depositing a thin layer of powder. A printhead is then used to print a glue pattern onto the powder, thus forming the first layer. A new layer of powder is deposited and glue is printed again. This pattern is repeated until the part is completed.

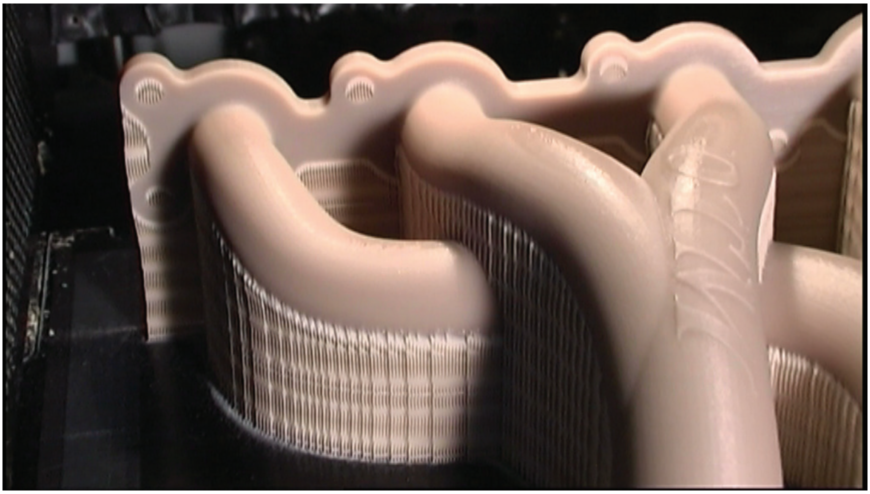


FIGURE 2 Thermojet wax manifold. Note the supports.

Two benefits of binder jetting are its speed and its lack of need for secondary support materials. Since the majority of the volume of the part is made up of the powder material, only a small fraction of the volume of the part needs to be deposited from the printheads. As a result, a layer can be formed very quickly (using arrays of printheads)—often in a matter of seconds. The powder that surrounds the part being formed will naturally act as a support for any subsequent overhanging geometries, and no secondary support materials are necessary.

The only commercially available full-color 3D printing machines are binder-jetting machines. A binder-jetting machine can be set up in such a way that a complete color spectrum can be printed layer-by-layer. This enables assembled parts to be produced in the intended colors (for marketing purposes) and for graphics, labels, and other visual features to be directly printed onto a part as it is being produced (Figure 3).

Photopolymer Vat

Photopolymer vat processes involve selective curing of predeposited photopolymers using some type of light source. Stereolithography, the first patented and commercialized AM process, works by scanning a laser across the surface of a vat of photopolymer. A platform is raised to just one layer thickness below the surface of the liquid. The laser scans the first cross-sectional layer, attaching the layer to the platform. The platform is lowered within the vat one layer thickness

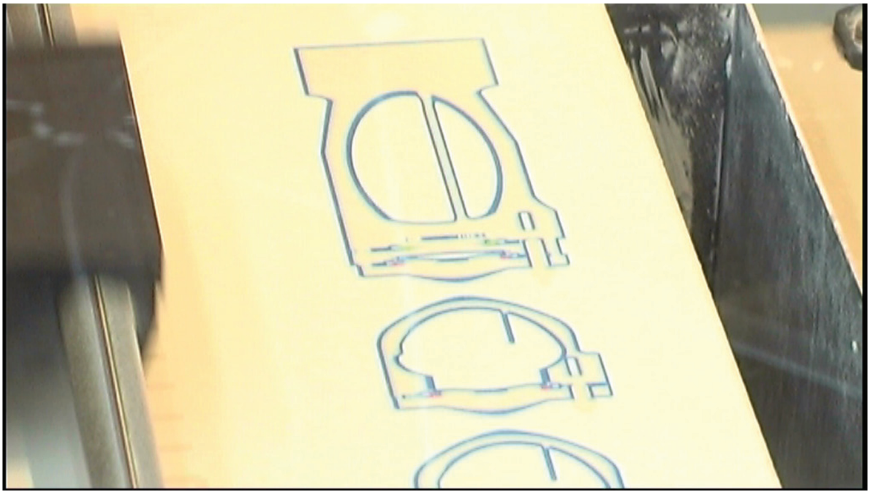


FIGURE 3 ZCorp 3D printed layer. Glue is printed inside the blue contours of the part layer. Color figure available online at *http://www.nap.edu/catalog.php?record_id=13274*.

(and material flows over top of the previously formed layer, or is formed over the previous layers using a blade or deposition device), and the process repeats until the part is completed. Photopolymer vat techniques give some of the best accuracies and surface finishes of any AM process.

Some photopolymer vat technologies have been developed to use digital light processing projectors to project an image of the layer on the surface of the vat, thus cross-linking the photopolymer and converting the entire layer from a liquid to a solid simultaneously.

Photopolymer vat technologies require a support network to be built for overhanging structures, otherwise these structures are subject to breaking or deforming. These supports are made from the same material that the part is made from, so these supports must be cut away after the part is completed.

Material Extrusion

The largest installed base of AM techniques is based upon material extrusion. Material-extrusion machines work by forcing material through a nozzle in a controlled manner to build up a structure. The build material is usually a polymer filament that is extruded through a heated nozzle—an automated version of the hot-glue gun used for arts and crafts. After a layer of material is deposited by the nozzle onto a platform, the platform either moves down or the nozzle moves up, and then a new layer of material is deposited.

In instances where two nozzles are installed in a machine, one of the nozzles is typically used to deposit a water-soluble support material. Three or more nozzles are sometimes used in machines designed for tissue-engineering research, so that scaffolds and other biologically compatible materials can be deposited in specific regions of the implant.

The simplicity of material-extrusion machines makes them suitable for in-office and home environments. Today, thousands of inexpensive extrusion machines, ranging in price from $700 to $3,000, are sold each year to do-it-yourself enthusiasts, small companies, educational institutions, and hobbyists. This proliferation of low-cost AM machines is a major reason for the current mainstream interest in 3D printing.

Powder Bed Fusion

Powder-bed-fusion machines work in a manner similar to binder jetting; however, instead of printing glue onto a layer of powder, thermal energy is used to melt the powder into the desired pattern. In most machines a laser is used to melt polymer or metal powders to build up three-dimensional objects. Another common variant is the use of an electron beam to melt metal powders.

In the case of polymer powders, the powder surrounding the part being built makes possible the creation of complex three-dimensional objects without supports. However, for metal powders, the thermal shrinkage of metal parts during solidification causes the parts to warp; supports are used to attach the part to a thick metal baseplate to maintain the accuracy of the parts and restrain them from warping. These metal supports are machined off after the part is completed.

Powder-bed-fusion machines are the most commonly used AM machines for the creation of end-use parts for highly engineered products. Polymer and metal parts made using these techniques are becoming widely used in aerospace, defense, and other highly engineered systems.

Directed Energy Deposition

Directed-energy-deposition machines melt material with a laser or other energy source as material is being deposited. These machines work similarly to material-extrusion machines except that, instead of melting the material with a nozzle, the wire or powder feed material is melted as it is being deposited onto a part (Figure 4).

In order to make parts with overhangs, directed-energy-deposition systems need to use either a five-axis deposition system (so that material can be deposited from any orientation) or a secondary support material. Because these systems are typically used to make metal parts or metal-ceramic composite structures, any supports that are used require machining to remove them.

Directed-energy-deposition processes are used primarily to add features

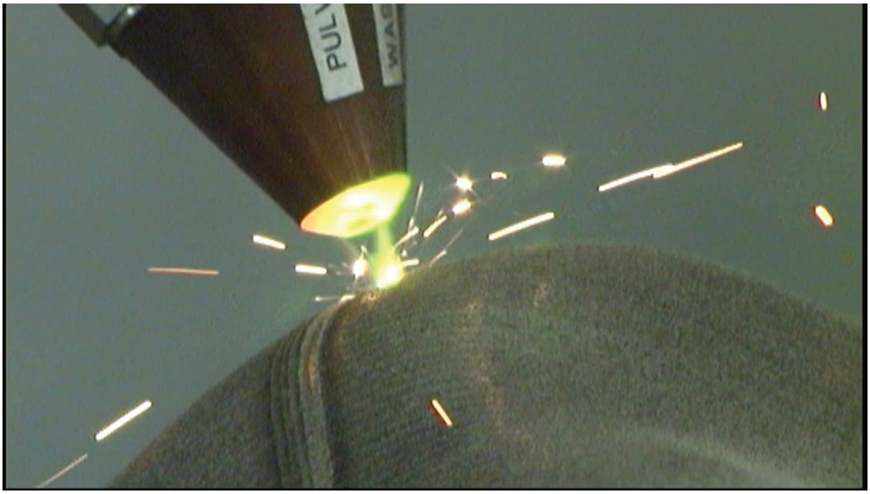


FIGURE 4 Trumpf machine depositing metal from powder (across a curved surface).

to an existing structure (such as adding strengthening ribs onto a plate) or for repair of damaged or worn parts. In most cases these processes are used to build up metal structures, and thus they are commonly referred to as metal-deposition AM machines.

Sheet Lamination

Sheet lamination techniques work by cutting and stacking sheets of material to form an object. This approach has been used with paper, plastic, and metal sheets to build up wood-like, plastic, and metal parts, respectively. A binder is typically used to bond paper and plastic sheets, whereas welding (either thermal brazing or welding, or ultrasonic welding) or bolting of sheets together is typically used for metals. Additionally, sheet lamination has been used with ceramic and metal green tapes (e.g., powder held together by a polymer binder in the form of a sheet of material) to build up structures that are later fired in a furnace to achieve a dense part.

Hybrid and Direct-write AM

In some instances multiple AM techniques are combined within the same machine or AM is combined with subtractive techniques such as computer numerical control milling or laser cutting. For instance, by combining a simple horizontal milling head into a material-jetting machine, Solidscape has created a popular line of high-precision wax printing machines with layer thicknesses of 0.0005 in. that are heavily used in the jewelry industry to create wax patterns for custom jewelry.

Several AM techniques have been modified to work at a small scale to deposit passive electronic structures (conductors, insulators, resistors, antennas, etc.). These techniques are often known as direct-write techniques and, for instance, use electronic "inks" that contain nanoparticles or other additives that result in electronic properties after drying, thermal decomposition, or other post-treatment. By combining direct-write techniques with other AM techniques it becomes possible to create multifunctional 3D-embedded electronic structures on a layer-by-layer basis that combine structural, thermal, electronic, and other functions into a single component.

BUSINESS IMPLICATIONS OF AM

AM has made a significant impact on manufacturing design over the past 25 years as rapid-prototyping techniques. The ability to create three-dimensional objects directly from CAD data allows designers to print out 3D representations of their designs for form, fit, and functional testing during the design process. This reduces the time needed to prototype new designs and enables more optimized designs to propagate through to manufacturing, catching errors before committing to costly mass-production tooling and assembly lines.

The layer-by-layer approach of AM is particularly well suited for highly complex geometries, including geometries with internal passageways, undercuts, and features that are difficult or impossible to make using traditional manufacturing techniques. For small, highly complex objects, AM powder-bed-fusion processes are sometimes cheaper than injection molding for volumes of parts approaching 100,000 components. However, the larger the part or the simpler its geometry, the more cost-effective traditional manufacturing processes become, compared to AM.

AM is well suited for parts with complex geometries that are made in low volumes or for personally customized geometries. For instance, AM is used today in many aerospace systems, including the Space Station, microsatellites, F-18 fighter jets, Boeing 787 aircraft, and more, and for custom dental aligners and hearing aids. AM techniques are commonly used as assembly jigs and fixtures by BMW and others to enable more ergonomic, accurate, and quick assembly of automobiles.

One of the key benefits of AM techniques is that there is no tooling or specialized hardware needed for any individual component. Each and every part can be unique from the part created just before or after it. A manufacturer can thus distribute AM machines near to the point of assembly or to the end customer, dramatically changing the logistics and distribution models for manufacturing rather than concentrating manufacturing where tooling is located. Since AM involves little human involvement to create a part, there is no longer an incentive to offshore manufacturing to a region of low-cost labor. As a result, onshoring is the standard for AM.

One particularly good opportunity for AM techniques is in the area of spare part production. If an out-of-production component is needed, a scanning technique can be used to capture the geometry of a broken part, modifications can be made to the CAD file, and a replacement part can be printed out. This is increasingly being used by people who restore automobiles, aircraft, defense systems, and other machines that were designed and built decades ago for which there are no CAD drawings or tooling available to make a replacement part and/or where the original manufacturer is no longer in business.

PERSONAL MANUFACTURING

AM is today where personal computing was in the late 1970s and early 1980s. AM is starting to move from the "mainframe" manufacturing arena into the "personal" manufacturing arena. In 2010, more than 5,978 "personal" AM machines were sold (all material-extrusion machines, most of which are sold for between $700 and $1,500), as compared to an estimated 6,164 other AM systems. In 2007, when personal AM machines first became available, only 66 personal AM machines were sold compared to 4,938 other AM systems (Wohlers, 2011). Thus, even though mainframe AM systems are growing at a respectable rate, the personal AM market has grown explosively. From an economic standpoint, as personal AM machines are sold at a cost of between 0.1 and 10 percent of the cost of other AM machines (with a commensurate lack of speed, accuracy, materials flexibility, etc.), the personal market is still a small portion of the overall AM market. However, explosive growth has caught the eye of many investors and companies, who are starting to make plans to enter the market.

ENTREPRENEURSHIP

A comparison of 3D printing to 2D printing is commonly used to understand the potential for AM. For 2D printing, there is a relatively stable mix of high-volume printers (e.g., newspapers, magazines, advertisements, etc.), medium-volume printers (e.g., digital photocopiers), and low-volume printers (e.g., laser and inkjet printers). In the same way that a consumer can purchase a mass-produced poster, print their own photos at Walmart, or use their own inkjet printer at home, 3D printing will likely be used by major manufacturers in factories, local suppliers in neighborhood 3D print shops, and within peoples' homes in the near future.

For AM, entrepreneurial opportunities include the development of new AM machines, supplying AM-produced components to major manufacturers, printing designs for consumers, selling unique designs for people to print at home, and more.

Thus, AM offers entrepreneurs the opportunity to develop new hardware, software, and services that either supplant existing products or introduce completely new products. For instance, entrepreneurs have started companies that develop and sell personal AM machines (*www.makerbot.com*), provide web portals for designers and consumers to sell and buy parts (*www.shapeways.com*; *i.materialise.com*), provide customized dental aligners (*www.aligntech.com)*, sell unique artistic (*www.bathsheba.com*) and consumer products (*www.freedomofcreation.com*), and more.

AM enables designers and entrepreneurs to start selling products without their own brick-and-mortar infrastructure. This is a new paradigm in manufacturing. Instead of requiring investors to provide startup capital, a creative person can create and sell unique goods without ever buying a manufacturing machine or paying for the development of a mold or tool. This means that the barrier to market for entrepreneurial activity in AM is very low and the distinguishing factor between successful and unsuccessful entrepreneurs may often have more to do with their ability to create marketing momentum through social media than their ability to secure venture capital or other financing. This means that entrepreneurs who are creative and understand the capabilities, benefits, and limitations of AM will have an edge over others.

CONCLUSIONS

AM technologies have the potential to create a new type of industrial revolution. AM is moving far beyond rapid prototyping in many industries. In the same way that the Internet has democratized the creation and distribution of information, AM has the potential to democratize the creation and distribution of physical goods. The ability to create customized, geometrically complex parts without tooling lowers the barriers for entrepreneurial activity and gives designers never-before-available opportunities for optimized design. Today AM is crossing the barrier from a specialized set of manufacturing and design tools into the regime of "personal" manufacturing. Those who are able to capitalize upon new business models involving physical goods might very well be the "Internet billionaires" of the future.

SUGGESTED SOURCES FOR FURTHER INFORMATION

Gibson, I, D.W. Rosen, and B. Stucker. 2010. Additive Manufacturing Technologies: Rapid Prototyping to Direct Digital Manufacturing. New York: Springer (an engineering textbook).

Rapid Prototyping Journal, Emerald Publishing (the leading AM-related journal, in its 16th year of publication).

Solid Freeform Fabrication Symposium proceedings (the premiere academic conference on AM since 1990). All papers are available for download at *http://utwired.utexas.edu/lff/symposium/.*

Wohlers Report 2011: Additive Manufacturing and 3D Printing State of the Industry. 2011. Fort Collins, Colo.: Wohlers Associates (a detailed industry analysis, updated annually).

www.additive3d.com (the well-known "Castle Island" website. Its terminology is a bit dated, but it has a lot of information for people who are new to the industry.)

Additive Manufacturing in Aerospace: Examples and Research Outlook

Brett Lyons
The Boeing Company

This paper presents an industrial view on the use of additive manufacturing for production of aircraft components and provides research examples that show the direction of related development. The advantages of additive manufacturing are becoming broadly recognized, and the stringent requirements found in the aerospace industry provide the context required to develop these complex processes to the level of robust performance established by traditional manufacturing methods.

INTRODUCTION TO AEROSPACE REQUIREMENTS FOR ADDITIVE MANUFACTURING

Additive manufacturing (AM) processes are unique in their ability to form the final part desired without any intermediate tooling. Additive processes such as selective laser sintering (SLS) begin with a computer-generated three-dimensional design of a given part. The part is then digitally segmented into very thin layers, which are selectively solidified in the machine, layer by layer. This ability to "grow" parts allows for designs with such complexity that they cannot viably be built with other processes. This approach to manufacturing removes the need, cost, and delay associated with tooling. Even with increasing rates of aircraft production (The Boeing Company, 2011), aerospace companies have numerous parts that are produced in very low quantities, making these tool-less processes attractive from an economic perspective (Ruffo et al., 2006). The use of AM provides a host of benefits, many of which are being recognized even in general media (The Economist, 2011).

From an aerospace and defense (A&D) design perspective, the weight of the part is often the deciding design factor for choice of material and manufacturing process use. Also, the uncompromised need for safety in air travel adds a long list of complex requirements, even for the simplest part. To consistently produce parts with identical and understood properties, the material and the process used to form it must be understood to a very high level. This complicated aerospace manufacturing context, which blends low-volume economics with acute weight sensitivity and the need for highly controlled materials and manufacturing processes, has led to the development of knowledge within The Boeing Company required to safely transition AM from the laboratory and model shop onto the factory floor.

To begin to understand the foundation of requirements placed on a commercial aircraft part, one can look to the U.S. Federal Aviation Regulations, which must be met before a Type Certification can be issued for a given aircraft series, required for entry into service with an airline (Federal Aviation Administration, 2011). While this set of regulations is very extensive and detailed, the single most pertinent language within the context of an AM review can be found in Title 14, Section 25, Subpart D, Subsection 25.605: "The suitability and durability of materials used for parts, the failure of which could adversely affect safety, must (a) Be established on the basis of experience or tests; (b) Conform to approved specifications (such as industry or military specifications, or Technical Standard Orders) that ensure their having the strength and other properties assumed in the design data; and (c) Take into account the effects of environmental conditions, such as temperature and humidity, expected in service." This brief but clear requirement is one of many that leads to the incredible safety record of commercial air transportation and also provides the impetus to rigorously study new fabrication methods such as SLS. Each A&D manufacturer will have internal specifications or will look to established standards organizations for data that allow accurate design of components from a given material, based on minimum allowable performance. Examples of material performance factors that are considered for even the simplest of components include specific strengths, fatigue, creep, use temperature, survival temperature, several tests of flammability, smoke release and toxicity, electric conductivity, multiple chemical sensitivities, radiation sensitivity, appearance, processing sustainability, and cost.

USE OF ADDITIVE MANUFACTURING IN AEROSPACE

Within Boeing, both military (Hauge and Wooten, 2006) and commercial (Lyons et al., 2009) programs use SLS to produce lightweight, highly integrated systems and payload components, as seen in Figure 1, that eliminate nonrecurring tooling costs and provide for life-cycle production flexibility. Since the first implementations on Boeing aircraft, the use of SLS has grown organically within a large number of programs. This is primarily due to its ability to produce thermoplastic parts that are lightweight, nonporous, thin-walled, and highly complex

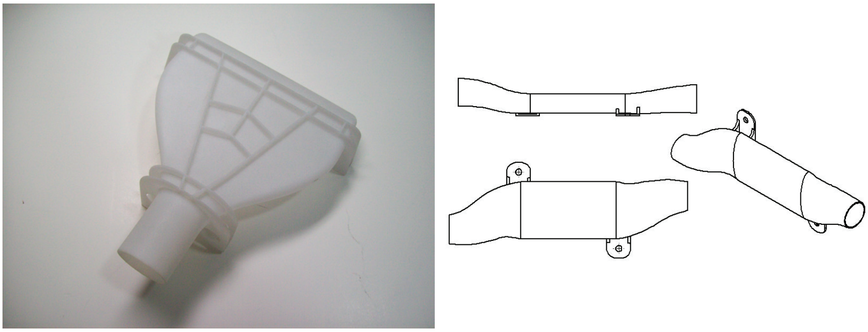


FIGURE 1 Photograph and illustration of laser-sintered air ducts.

geometrically and to do so in an economical fashion. These properties also led to the frequent use of SLS within the burgeoning field of unmanned aerial vehicles.

With weight being a critical factor, the very thin walls and complex designs possible in SLS are attractive for replacement of parts typically made through established processes such as rotational, injection, or polymer matrix composite molding. To take advantage of SLS, one must have a firm understanding of the extreme four-dimensional energy input gradients that exist during processing. For example, typical SLS machines use 75-watt CO_2 lasers that have a 500-µm spot size. That laser spot moves at up to 10 meters per second, over layers of nylon powder only 100 µm thick, with each layer being completed in approximately 60 seconds. The thorough description and efforts to simulate the details of the energies present in the SLS process can be found in the literature (Franco et al., 2010). This unique manufacturing context requires that any aerospace company develop in-depth knowledge of the materials and process used in order to draft commercially efficient specifications.

EXAMPLES OF AEROSPACE-DRIVEN RESEARCH IN AM

To build parts with repeatable mechanical properties and dimensional control, the temperature distribution across the part-building platform must be held at as even a temperature as possible. In order to accomplish this and reduce scrap rates, The Boeing Company and its partners at the University of Louisville and at Integra Services International (Belton, Texas) developed a patented method for zonal control of the part bed temperature in SLS equipment (Huskamp, 2009). The multizone, near-infrared (IR) wavelength heating elements, seen in Figure 2, provide the fast response and spatial resolution required to maintain even part bed temperature. This invention, when paired with real-time IR imaging, provides a significant improvement in thermal control. This level of thermospatial control

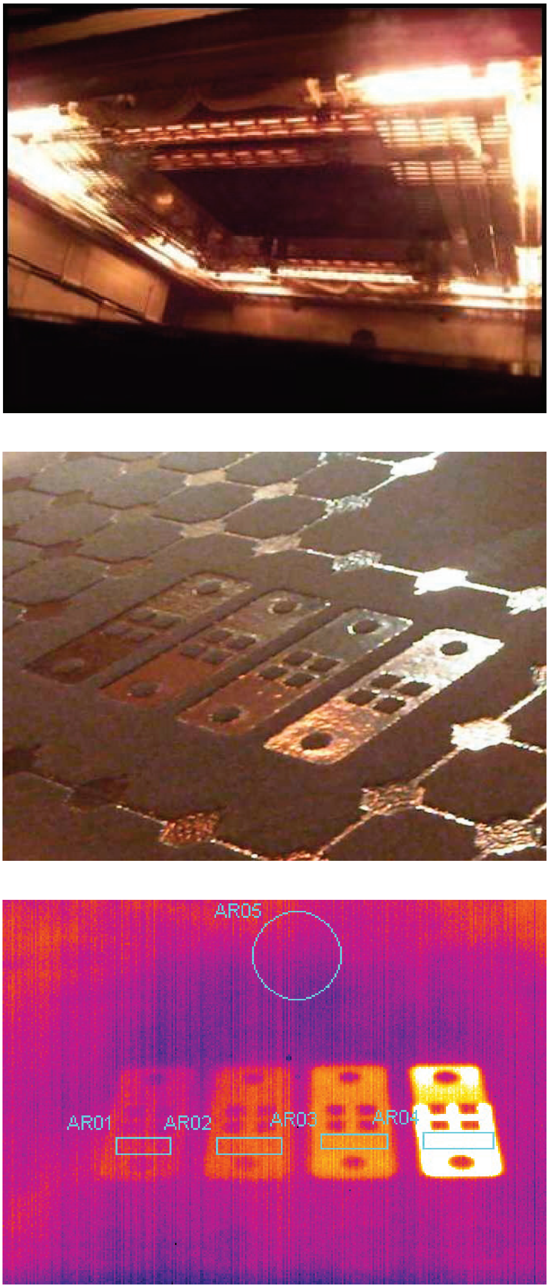


FIGURE 2 MZ heating (top), SLS part bed (middle), and same parts seen via infrared thermography during laser scanning (bottom).

has had to become more advanced than that found in most other thermoplastic processing methods.

Another example of an aerospace-driven AM need being met by researchers can be seen in the emergence of flame-retardant polyamides. When considering that many polymers are derived from fossil fuel-based hydrocarbon feedstocks, the concept of a related chemistry being self-extinguishing when exposed to flame is impressive. That characteristic is required, to a greater or lesser degree, of all polymer materials used on the interior of commercial aircraft. To gain the weight and manufacturing benefits provided by SLS on its commercial aircraft, Boeing collaborated with its suppliers to develop the first material that could be laser processed and that passed the required flammability tests (Booth, 2010).

CURRENT AREA OF DEVELOPMENT

Progress has been made to increase the number of AM applications in aerospace, which has identified three new performance challenges for SLS polymer materials. The operating requirements of programs such as F-35 and 787 have put requirements on the AM community to develop materials that can (a) operate at higher temperatures, (b) have significantly better flame resistance, and (c) offer an adjustable degree of electrical conductivity (Shinbara, 2011). These new physical performance targets must be met while maintaining as many of the attributes already established by SLS polyamide materials as possible. Those attributes include mechanical toughness, resistance to chemical attack, ultraviolet radiation resistance, dimensional fidelity, and viable economics. Such a material, if developed successfully, could have a wide range of applications within and beyond aerospace. Two of the most notable non-aerospace applications for new high-performance polymers in SLS are the potential use in medicine for implants and devices (Schmidt et al., 2007) and low-volume automotive production.

As is historically the case when a new technology is enabled and near transition to useful service, numerous parties from many industrialized nations can be seen working on the same technical problem simultaneously. Researchers in the United States, Germany, Japan, and the United Kingdom have made the deepest investigations into developing high-performance polymers for SLS (Hesse et al., 2007; Kemmish, 2010). There are many high-performance polymers that are attractive for development from a cost perspective; however, the cost of testing required by aerospace makes multiple, simultaneous material development efforts cost prohibitive. Because of the known performance of the polyaryletherketone (PAEK) family of materials, they are viewed as the lowest-risk option for current development. The PAEK family includes different chemistries such as polyetherketone, polyetheretherketone, and polyetherketoneketone. The choice of a PAEK as the next material family to be developed is based on factors inherent to its chemistry, including very good flammability and chemical resistance, low moisture sorption, good mechanical performance, good resistance to creep and fatigue, compatibility with several methods of sterilization, and numerous material grades and suppliers to choose from.

Owing to the comparatively small size of the SLS market and the cost of developing new polymer chemistries, raw materials for SLS are typically selected from commercial-off-the-shelf (COTS) grades. These COTS materials have been designed for other applications such as coatings, films, or rotational molding. While injection molding and other methods of polymer processing can make use of both heat and pressure to form a part against the surface of a mold, AM processes have to rely primarily on thermal energy input. Viable materials for additive processes must have very specific viscosity and other properties to be successfully processed. To begin generating material performance test data, a processable material form must first be developed.

The development of a viable PAEK-SLS material is an area of competitive, industrial research at this paper's time of writing, so specific information from any one party is generally not published. However, the comparison of established, well-understood polyamides (PAs) to the PAEK family shows the problem space of engineers working in this field. Table 1 provides comparative thermal properties of lower-temperature PA and PAEK materials, as found in the literature and in manufacturer-published information (Kemmish, 2010; Kohan, 1995).

By understanding and comparing the bulk thermal properties of these two material families, one can begin to understand how differently they will behave within the SLS process. One such comparison can be seen when the amount of energy required to heat the material is considered. To process a PA powder, the lower melt temperature combined with the lower specific heat (the amount of energy required to heat a given mass of material one degree Kelvin) indicates that the effort required to achieve a given viscosity with heat input is much lower for a PA than for a PAEK. The PAEK must be heated to twice the temperature just to approach melt and will require almost twice the energy per degree of heating. This is further complicated in SLS processing, as the transient heating retirements of each layer must not change drastically.

A second comparison is the ratio of specific heat to the heat of fusion (the energy flux exhibited in the transition from solid to liquid, and vice versa). In the SLS process, the polymer powder is heated in stages from ambient condi-

TABLE 1 Comparison of PA and PAEK Family Thermal Properties

Material	Melt Temp °C	Glass Trans. Temp °C	Specific Heat J/g K	Heat of Fusion (100% crys.) J/g	Thermal Conductivity W/m K	Thermal Expansion (ppm/Tg °C)	Specific Gravity g/cc (crystalline)
PA	180-186	42-55	1.26	226	0.19	85	1.03
PAEK	300-375	145-165	2.20	130	0.26	60	1.30

tions to very near the melt point. PA has a lower ratio of specific heat to heat of fusion than the PAEK family of materials (1.26/226 compared to 2.20/130). This is important because, despite best efforts, there is some gradient of energy input and temperature across the building area at any given time. If a material has a very gradual transition into melt, such as fully amorphous polymers, it can be difficult to feed smoothly onto the machine's part-building area.

This ratio of specific heat to heat of fusion gives an indication of how easily a given material can be heated to near the melt point, across the whole part bed, without fusing particles together. The closer to the melt point the material can be fed into the machine, the lower the energy input requirements are on the laser for heat input that transitions the material into the melt region. The lower the requirement put on the laser for energy input, the lower the risk of polymer degradation. This is because within the CO_2 laser spot there is a roughly Gaussian distribution of energy, the peak of which can cause degradation.

Also tied to this comparison is the speed at which the laser draws each layer of the part. With a given energy input requirement put on the laser per the above comparison, the layer can be drawn with faster or slower laser scan speeds. The scan speed affects the overall per-layer time which, in addition to the proportion of preheat to laser energy required, results in a variable temperature distribution and cooling rate across a part's cross section, per given layer. If too long a time has passed between the start and stop of a given layer and too high of an energy demand is put on the laser, sections of that layer will have cooled faster than others and, in turn, will have shrunk nonuniformly. Dimensional distortion can result if too high a cooling gradient exists relative to recrystallization temperature, thermal conductivity, coefficient of thermal expansion, and a host of other material factors.

A thorough description of the interaction between just the properties shown in Table 1 is beyond the scope of this paper, but the three examples give a window into the problems currently being solved by AM researchers. Thankfully, despite

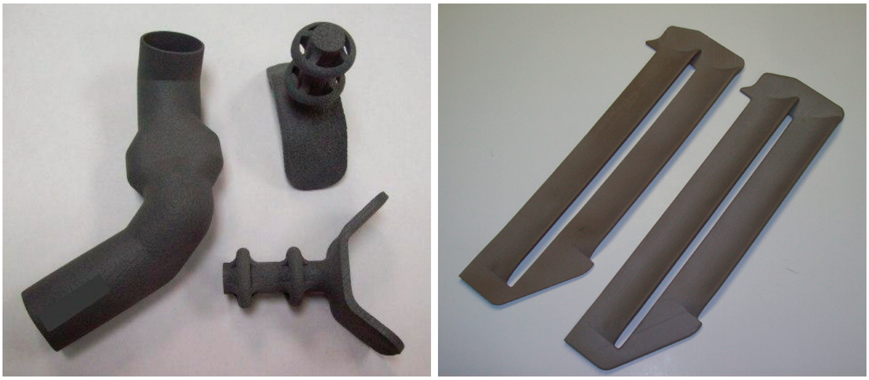


FIGURE 3 Examples of parts of PEAK materials processed via SLS.

all of these complexities, multiple parties are reporting success with the processing of PAEK materials via SLS, as can be seen in Figure 3.

MOVING FORWARD

Beyond new higher-performance polymers, numerous research frontiers exist within AM. Separate from the specific material development identified earlier, AM equipment must transition from comparable low-reliability laboratory-grade equipment to hardened, cost-effective, high-temperature industrial-grade machines. The AM industry can look to its predecessors in injection molding and computer numerical control machining for examples of how to establish new manufacturing technology and the supporting business case. The unique material requirements posed by AM processing have, during the technology's infancy phase, tied machine manufacturers to material and even part sale activity. While this has provided a good revenue source to support the new companies, it has also impeded new applications by making new material development difficult for all but the largest of users and material companies. This dependence on material and part sales, and nonproductive patent litigation, has also distracted the machine manufacturers from improving upon their equipment with an eye toward higher-volume, economical industrial manufacturing. Equipment manufacturers such as Toshiba, Haas Automation, MAG, Husky, and Arburg do not rely on material or part sales to bolster their equipment business, and if the AM industry is to grow successfully it might look to their business models and history for reference.

Beyond polymers the use of metals in AM for aerospace is equally as complex and exciting. Leading the way in direct metal part manufacturing have been engine manufacturers. While direct part manufacturing is a highly dynamic field, the leveraging of AM's ability to create highly complex shapes is very applicable to tooling for both metal and composites components. New tooling-focused machines, processes, and materials are being actively developed that leverage the process benefits of AM while delivering the performance of cast metals (Halloran et al., 2010) or long-fiber-reinforced composites (Wallen et al., 2011).

Independent of material or processing conditions, the analysis of complex geometries that can be built only with additive methods is also an active and important area of research. Even with material test data generated, the types of structures that AM can build, such as the trussed airfoils seen in Figure 4, are difficult to analyze for predictive behavior. This field of study is generating new software tools for the generation and predictive analysis of complex structures, such as three-dimensional trusses (Engelbrecht et al., 2009).

These descriptions of current research areas, along with examples such as Boeing's use of SLS on commercial and military aircraft, show that the aerospace industry has the opportunity to lead, and responsibility to contribute to, this revolutionary field of manufacturing technology.

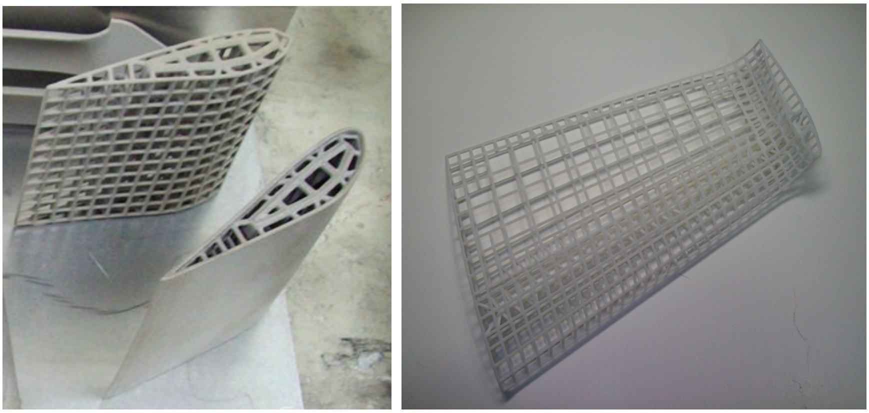


FIGURE 4 Two complex truss examples that indicate the difficulty of predictive analysis.

REFERENCES

The Boeing Company Press Release. 2011. Available at *http://boeing.mediaroom.com/index.php?s=43&item=1426*.

Booth, R., October 7, 2010. Methods and systems for fabricating fire retardant materials. U.S. Patent Application 2010/0255327.

The Economist. "Print me a Stradivarius." February 12, 2011. p. 11.

Engelbrecht S., L. Folgar, D. Rosen, G. Schulberger, and J. Williams. 2009. Cellular structures for optimal performance. Proceedings of the University of Texas at Austin Solid Freeform Fabrication Symposium, Austin, Texas, 2009.

Federal Aviation Administration. 2011. Available at *http://www.faa.gov/regulations_policies/faa_regulations/*.

Franco, A., M. Lanzetta, L. Romoli. 2010. Experimental analysis of selective laser sintering of polyamide powder: An energy perspective. Journal of Cleaner Production 18:1722–1730.

Halloran, J.W., V. Tomeckova, S.P. Gentry, S. Das, P. Cilino, D. Yuan, R. Guo, A. Rudraraju, P. Shao, T. Wu, T. Alabi, W. Baker, D. Legdzina, D. Wolski, W. Zimbeck, and D. Long. 2010. Photopolymerization of powder suspensions for shaping ceramics. Journal of the European Ceramic Society 31(14): 2613–2619.

Hauge, R., and J. Wooten. 2006. Rapid Manufacturing: An Industrial Revolution for the Digital Age, Chap. 15, p. 233. Chichester, U.K.: John Wiley & Sons Ltd.

Hesse P., P. Tillmann, R. Weiss. November 22, 2007. Powder for rapid prototyping and associated production method. U.S. Patent Application 2007/0267766.

Huskamp, C. April 7, 2009. Methods and systems for controlling and adjusting heat distribution over a part bed. U.S. Patent 7,515,986.

Kemmish, D. 2010. Update on the Technology and Applications of Polyaryletherketones, Shawbury, U.K.: iSmithers.

Kohan, M. 1995. Nylon Plastics Handbook. Cincinnati, Ohio: Hanser/Gardner Publications.

Lyons, B., E. Deck, and A. Bartel. 2009. Commercial Aircraft Applications for Laser Sintered Polyamides. Society of Automotive Engineers Technical Paper 09ATC-0387.

Ruffo, M., C. J. Tuck, and R. J. M. Hague. 2006. Cost estimation for rapid manufacturing–laser sintering production for low to medium volumes. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 220(9):1417–1427.

Schmidt, M., D. Pohle, and T. Rechtenwald. 2007. Selective laser sintering of PEEK. CIRP Annals–Manufacturing Technology 56(1):205–209.

Shinbara, T. 2011. High temperature laser sintered components. Presented at the Society of Manufacturing Engineer's AeroDef Manufacturing 2011 Event, Anaheim, California, April 7, 2011.

Wallen, M., J. Rossfeldt, C. Aune, and Z. Wing. 2011. Washout Tooling for Hollow Composites: Toward Rapid, Low Cost Production Technology. Society of Manufacturing Engineers: Composite Manufacturing, Talk #6793, Dayton, Ohio.

Additive Manufacturing Is Changing Surgery

Andrew M. Christensen
Medical Modeling Inc.

ABSTRACT

Additive manufacturing (AM) is changing how surgery is performed. Over the past two decades relatively small, continuous advances have pushed forward the concept of using AM as part of the surgical treatment of conditions ranging from arthritis of the hip or knee to ablation of malignant tumors of the head and neck. Historical uses of AM in medicine have been primarily limited to custom anatomical models based on medical imaging data and prototyping instruments for new designs (Figure 1). Today's trend for treatment is toward surgery that is not "one size fits all" but surgery that is now tailored to the exact needs of each patient. AM plays a key role in making this happen in a time- and cost-efficient manner based on its built-in ability to create objects of almost unlimited complexity. This is evidenced by new applications for patient-specific surgical instrumentation whose design is driven by that patient's surgical plan and AM techniques that allow for direct output of implantable metals.

BACKGROUND

Historically the use of additive manufacturing as applied to surgery was mostly related to manufacturing of surgical planning models based on patient-specific medical image data (Mankovich et al., 1990). These "3D x-rays" allow surgeons to more closely evaluate complex anatomical structures before they enter the operating room. Most times these models depict hard-tissue anatomy (i.e., bone structures) and are used (a) to visualize anatomy preoperatively, (b) to allow for surgical simulation such as creation of bone cuts, or (c) for prebending

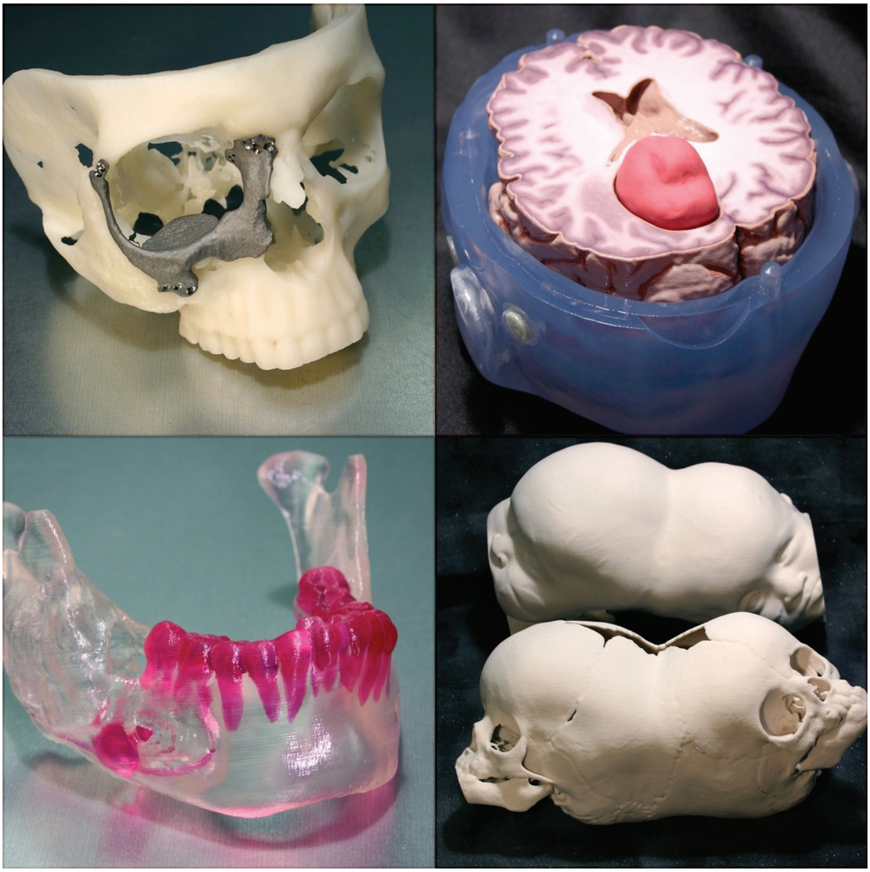


FIGURE 1 Images of typical medical-image-based models used in planning for surgery of the head and neck. Source: Courtesy of Medical Modeling Inc.

or sizing of off-the-shelf implant components, which can be "customized" to fit the bone model.

In the late 1980s came the introduction of the first commercial additive manufacturing process called stereolithography (SL or sterolithography apparatus [SLA]). Stereolithography offered immediate access to creating models of almost unlimited complexity from a three-dimensional computer-aided design (CAD) file in a translucent resin material. Immediately those surgeons who had previously been using subtractively manufactured models created from computed-tomography (CT) scans started experimenting with this new additive fabrication technology. SLA allowed not only the creation of the external anatomy but also a precise depiction of internal anatomy, including bony voids. Case reports immediately surfaced and head and neck surgical applications became the mainstay of use of this technology. Historically many have asked why head and neck surgery was leading this area and why not orthopedics. The answer seemed to lie in the fact that for orthopedic procedures, which were far more numerous (by two orders of magnitude, give or take), the surgeons cared more about *function* and less about *form*. If you imagine a hip joint reconstruction, the primary focus surrounds its biomechanical function and optimization of concentric wear from ball to cup. The reconstruction could look awful from an aesthetic standpoint but be perfectly functional from a biomechanical perspective. Reconstructive surgery of the head and neck, by contrast, has probably more to do with form than function. There are exceptions, such as reconstruction of the temporomandibular joint, but for the most part surgery of the facial skeleton has to do with restoring or creating symmetry. This pushes surgeons to agonize about fairly insignificant details such as forehead projection, symmetry between the left and right sides of the face, and interaction between the upper and lower teeth. This is probably why anatomical modeling technology by additive techniques found a home supporting surgical specialties such as oral and maxillofacial surgery, neurosurgery, and craniofacial surgery.

ADDITIVE MANUFACTURING MATERIALS

Additive manufacturing methods exist today allowing one to directly output CAD models in many different types of materials. For surgical applications the interesting materials of late fall into one of two categories: (1) strong and durable plastics for use as instruments in surgery and (2) implantable metals. Strong and durable plastics are now used in surgical settings as one-time-use instruments on a daily basis around the globe. These have found great application in surgeries such as total knee replacement and reconstruction of the mandible using virtual surgical planning techniques. The main technologies that play a role in providing durable plastic parts are laser sintering (or selective laser sintering) and SL or SLA. A key to these materials being usable is the fact that they are biocompatible for limited in vivo use during surgery and can be sterilized for use in a sterile field, including contact with the patient's body fluids and tissues. All materials for these applications need to be tested to some portion of the standard for materials biocompatibility, ISO 10993.

Implantable metals created using additive manufacturing have only been available for the past five years or so and represent a major step forward for the industry. Now, not only can one mock up a prototype of the implant or create a mold to create the implant in an indirect fashion, but one could directly output the implant in an implantable metal. There are a few companies making additive manufacturing equipment that can directly output parts in implantable metals including titanium, titanium alloys, cobalt-chrome alloy, and other metals. If the parts can be certified to a surgical implant material standard (many ASTM and ISO standards exist), then it is possible for these materials to be used in the body. In Europe, parts created by Electro Optical Systems direct metal laser sintering process (*www.eos.info*) and Arcam AB's electron beam melting process (EBM; *www.arcam.com*) are CE (Conformité Européenne) marked for sale. One of those, an acetabular cup system offered by Italian orthopaedic company Lima Orthopaedics, has been used clinically (i.e., implanted) in more than 15,000 patients in the past 3 years (Marin et al., 2010). This still represents a very small number of the total of hip-replacement surgeries used—the 2010 number worldwide is 1.4 million hip replacements—but for additive manufacturing it is an impressive number. The U.S. market has been slower to adopt this new technology as a mainstream manufacturing technique, but in the past year there are at least four devices that have been cleared by the U.S. Food and Drug Administration that involve metallic implants for spine and hip reconstruction made using additive manufacturing methods. The known clearances are all for implants created using the EBM process. Several custom implant concepts have been reported, including work at Walter Reed Army Medical Center and National Naval Medical Center on reconstruction for war veterans with large cranial defects. In this case a customized porous titanium implant is fabricated with integrated fixation for adherence to the surrounding skull with standard titanium screws.

VIRTUAL SURGICAL PLANNING

The advent of easier access to medical images has given rise to a whole new area of medicine that involves virtual planning in a CAD-type environment for the upcoming surgery. Imagine planning for the surgery of your knee (Figure 2) or hip or jaw with your surgeon before surgery to allow for optimization of the surgical plan and creation of custom instruments to help precisely guide this plan into reality. This sounds like science fiction but across the globe today thousands of patients per year are getting this "customized" treatment (Davis et al., 2010; Hirsch et al., 2009).

The major step forward here, in this author's opinion, has been the combination of surgical planning software with online collaborative environments such as those popularized by WebEx and GoToMeeting. Two methods exist for virtual surgical planning today and many methods are in the gray area between. The first is the "off-line method," whereby the surgeon sends a service company his patient's medical images and an order form with detailed instructions and awaits the surgical plan for approval. The surgical plan is developed by an engineering team using the inputs provided by the surgeon in conjunction with their proven workflow for treating the particular type of surgical situation. The second approach involves the surgeon sending a service company the medical images and the same detailed instructions. The difference is that the design of the surgical procedure is done in real-time collaboration between the surgeon and the engineering team during a

FIGURE 2 Customized, disposable instrument developed from a CT scan to guide total knee arthroplasty. Source: Courtesy of DePuy Orthopaedics.

web meeting conducted over the Internet (the "real-time" method). If we imagine the surgery being planned is a total knee replacement, the result of either technique is the development of a surgical plan for how the surgery will be performed, where bone cuts will be placed, what bone will be moved around or taken out, and where the prosthetic implant(s) will be placed. Once the surgical plan is created, the output of that surgical plan could go down one of two major pathways: (1) creation of a one-of-a-kind custom implant based on the surgical plan, or (2) creation of custom, disposable surgical instrument(s) that would facilitate placing an off-the-shelf implant into its desired, predetermined position.

THE FUTURE

If we can provide "designer" surgeries today and create fully custom surgical interventions, where might we go next? The answer probably lies in two major areas, both having to do with the implants themselves. Additive manufacturing plays a key role in the first next logical step: *functional elements driving design*. Today's implants, while custom-fit or truly custom, are only custom in form. Their shape mimics the patient's shape and they are meant to fit in a more stable manner than "off-the-shelf" interventions. It is commonly known that the more stable the implant, the longer it will last because loosening is the number one cause of failure for certain artificial joints like the hip or knee. Apart from something that fits well, however, we need something that is designed just for you, something that takes your form and your function requirements and marries the two to create an implant that takes the shape that it needs to take, not the shape that an engineer predetermines it should take. As an example, we look at a hip stem implant. Typically this implant fills the intramedullary canal (marrow space) of the femur with a large amount of solid titanium material, either cemented or press-fit for stability. The issue with the current designs involves a one-size-fits-all mentality where this works for most, but people at the ends of the bell curve end up with problems with the stem being either too large or too small. Furthermore, working in conjunction with the size issue is a biomechanical issue called "stress shielding," whereby the bone surrounding this large mass of metal is not stressed the way it was originally. There is a risk that this bone will resorb because it is not under appropriate levels of stress. In this instance, if we let form follow the functional requirements, we end up with a stem that fits the intramedullary canal but instead of being a large solid mass of metal it might look more like a combination of solid areas and patches of metallic mesh that vary based on the surrounding bone and the function required at that particular point along the canal. This idealized hip stem might allow just the right amount of flexural modulus so that the bone surrounding the implant not only does not die off but actually thrives. This concept is powerful, and additive manufacturing is ideal for its implementation based on the ability to create objects of almost unlimited complexity with a lot size of one being just as easy to fabricate as a lot size of 1,000.

Today's alloplastic (man-made) biomaterials have not changed much in many years. Sure, there are newer variations of things and new categories such as ceramics that are becoming more and more a mainstay for articulating surfaces. The next frontier for biomaterials involves moving toward regenerative medicine, whereby the materials we put into the body eventually help regenerate the body's own natural tissues. Much work is going on today to look at creating artificial tissues of many types, from simple bone structures all the way up to complex organ systems like the kidney and lung (Mironov et al., 2009). Additive manufacturing is known to be an ideal fit for the future of these projects, such as printing a kidney, because of the way that you can control the internal structure of an object built up in three dimensions by creating it in thin, cross-sectional two-dimensional layers.

CONCLUSIONS

Major advances in applications and materials have led to additive manufacturing (AM) playing a key role shaping the way surgery is performed today. Output of customized surgical instruments via AM and direct AM-fabricated implants will continue to expand into many other anatomical areas and surgical specialties. The future is bright for direct metallic implant production via AM technologies because they are fundamental to eventually creating truly "customized" devices that will focus on functional elements and restoring biomechanical stability in replaced joints. Additive manufacturing plays a key role in progressing these advanced treatments, based on its core abilities to produce complex objects and to produce these objects with a lot size of one just as easily as producing a lot size of 100. This flexibility will allow for more personalization of care in the areas of joint replacements and, farther in the future, organs as the numbers of people needing these therapies will continue to grow dramatically over the coming decades.

REFERENCES

Davis, J. J., J. V. Bono, and B. G. Lindeque. 2010. Surgical strategies to achieve a custom-fit TKA with standard implant technique. Orthopedics 33(8):569–576.

Hirsch, D. L., E. S. Garfein, A. M. Christensen, K. A. Weimer, P. B. Saddeh, and J. P. Levine. 2009. Use of computer-aided design and computer-aided manufacturing to produce orthognathically ideal surgical outcomes: A paradigm shift in head and neck reconstruction. Journal of Oral and Maxillofacial Surgery 67(10):2115–2122.

Mankovich, N. J., A. M. Cheeseman, and N. G. Stoker. 1990. The display of three-dimensional anatomy with stereolithographic models. Journal of Digital Imaging 3(3):200–203.

Marin, E., S. Fusi, M. Pressacco, L. Paussa, and L. Fedrizzi. 2010. Characterization of cellular solids in Ti_6Al_4V for orthopaedic implant applications: Trabecular titanium. Journal of the Mechanical Behavior of Biomedical Materials 3(5):373–381.

Mironov, V., R. P. Visconti, V. Kasyanov, G. Forgacs, C. J. Drake, and R. R. Markwald. 2009. Organ printing: Tissue spheroids as building blocks. Biomaterials 30(12):2164–2174.

The Shape of Things to Come: Frontiers in Additive Manufacturing

HOD LIPSON
Cornell University

Additive manufacturing technologies—machines that can automatically fabricate arbitrarily shaped parts, pixel by pixel, layer by layer, from almost any material—have evolved over the past three decades from limited and expensive prototyping equipment in the hands of few, to small-scale commodity production tools available to almost anyone. It has been broadly recognized that this burgeoning second industrial revolution will transform every aspect of our lives. Clearly, we will see the usual improvements we have come to expect of any burgeoning technology: more material options, better resolution, faster printing, easier and more reliable operation, and lower costs. But where will this technology go next?

3D printing is only the tip of a much larger phenomenon. We can look at the evolution of additive manufacturing technologies' past, present, and future as a series of milestones in humans' increasing control over physical matter.

PRINTING FORM: PROGRAMMING THE SHAPE OF MATTER

The first episode of this journey, maturing today, has been the unprecedented control over the shape of objects. Machines today can fabricate objects of almost any material—from nylon to glass, from chocolate to titanium—and with any complex geometry. This ability is transforming many fields to a degree that few inventions ever have—not just engineering, but many others, from biology to archeology, and from education to culinary arts.

Complexity for Free

Perhaps the most dramatic impact of 3D printing is the ability to manufacture objects without factoring their complexity. Fabricating a solid block costs almost the same as fabricating an oddly shaped object with curved surfaces and notches. In either case, the printer will still scan back and forth, depositing one layer at a time, and only the material deposition pattern will change. In the same way that printing a picture of a circle takes no longer than printing a picture of the map of the world, printing a mousetrap does not take more time, resources, or skill than printing a paperweight. Regardless of how you want to measure cost—in production time, material weight, energy waste, or in production planning effort—adding more features hardly changes the cost. In some cases, added complexity can even reduce cost; printing a block with a hole is cheaper and faster than a solid block without a hole. The marginal cost of added complexity is therefore near zero, in stark contrast with conventional manufacturing where every new feature—every additional hole, surface, protrusion, and corner—takes more planning effort, requires longer production time, and consumes more energy and possibly more raw material.

Why do we care about the marginal cost of complexity? It is easy to understand that lowering the cost of manufacturing complex products is a good thing, but the reason is more profound. Industrial revolutions are triggered when a fundamental cost associated with production drops dramatically, essentially taking that factor out of the equation. The industrial revolution of the 19th century occurred when the cost of power dramatically declined, as steam engines replaced horses and waterwheels. As a consequence, these power sources were not only replaced, but the range and types of work that machines could perform greatly expanded, leading to a cascade of innovation, such as railroads and factory automation. The Internet similarly reduced the cost of disseminating information and, as a consequence, expanded the range and types of media that could be distributed—not just online newspapers, but also Wikipedia, blogs, and user-generated content. One could argue that 3D printing has drastically reduced the cost of making complexity. Initially, this simply implies that this technology will gradually replace the old way of making the things we once made in more expensive ways. But in the longer term, the range and types of objects being manufactured will greatly expand.

The Personalization of Manufacturing

Alongside the vast new design possibilities, however, is the personalization of manufacturing. This trend has profound economic implication on how we will design and consume products in the future, who designs them, and where they are made (Lipson and Kurman, 2010). Most important, the ability of anyone to design and make complex products without the barriers of resources and skills of traditional manufacturing will democratize innovation and unleash the long tail of human creativity. Look online right now and you will see thousands of objects for sale ready to be printed on demand, from custom-shaped hearing aids, to flapping-hovering micro air vehicles, to authentic-looking replicas of ancient cuneiforms (Figure 1).

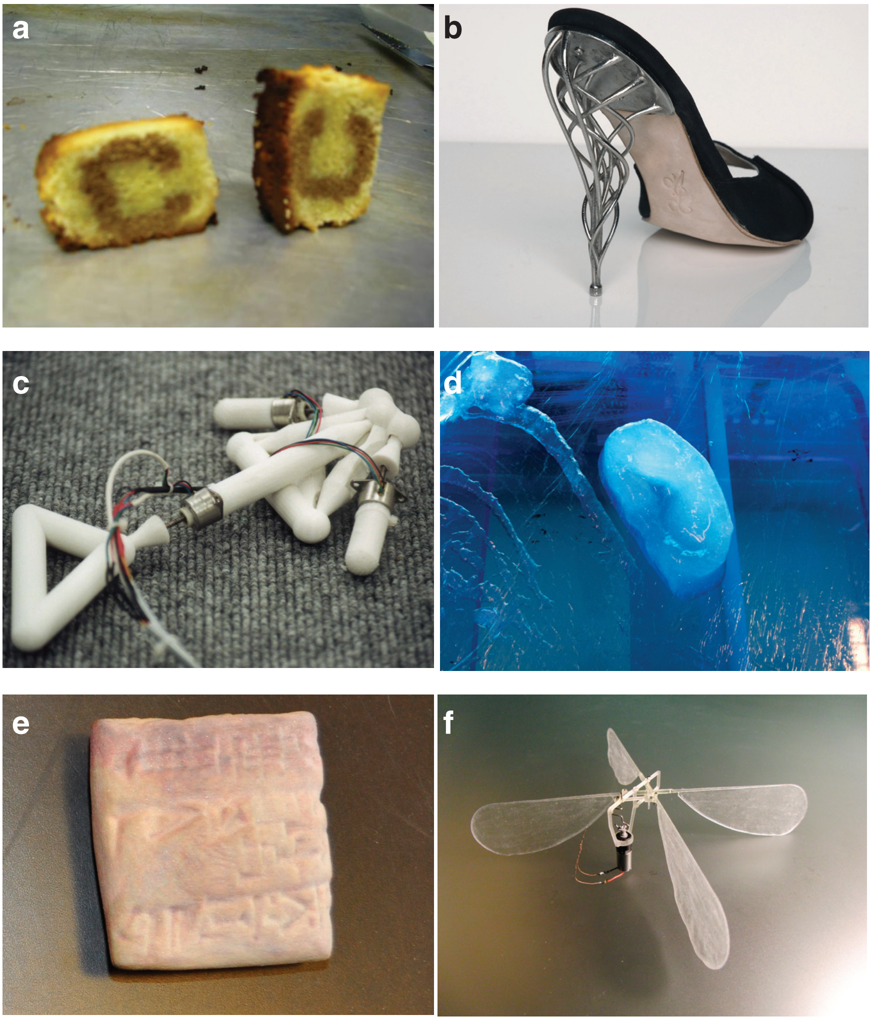


FIGURE 1 From chocolate to titanium, from biology to archeology, and from robotics to avionics, 3D printing is a universal technology that will impact every industry and discipline. Sources: Figures 1a, 1c, 1d, 1e, and 1f courtesy of Cornell University and http://creativemachines.cornell.edu. Figure 1b courtesy of WTWH Media.

PRINTING COMPOSITION: SHAPING THE INTERNAL STRUCTURE OF MATERIALS

The second episode of this journey, which we are beginning to experience now, is the control over the composition of matter—going beyond shaping just the external geometry to shaping the internal structure of materials with unprecedented fidelity. Using multimaterial additive manufacturing technologies, we can make materials within materials, embed and weave multiple materials into complex patterns, and cofabricate entangled components. For example, we can print hard and soft materials in patterns that create bizarre and new structural behaviors, like materials that expand laterally when pulled longitudinally. We are shedding the traditional limitations imposed by conventional manufacturing where each part is made of a single material. Instead, microstructure can be specified with micron-scale precision. With such possibilities, you will be able to print a custom tennis racket that cleverly enhances your unique backhand, or a replacement spinal disc implant tailored for your aching back. While the possibilities are vast, however, few theories can predict properties of these new materials and few designers can exploit the new design space, so new design tools will be needed to augment human creativity.

From Bioprinting to Food Printing

There is literally no discipline that will remain unaffected by this unprecedented control over the shape and composition of matter. It is easy to imagine the implication for any field that involves the design and fabrication of physical objects, from mechanical engineering to art and architecture. But the technology can also change fields that have no immediate connection to engineering or to manufacturing. Consider an example: the emerging field of 3D printing for medical applications.

Custom-shaped prostheses. One of the earliest applications of 3D printing to health-related applications was the fabrication of custom-shaped, complex prosthetic limbs and devices. A key challenge in many prosthetic devices is the interface point with the body. For example, the ability to match the socket of a limb prosthesis to a patient's bony prominences or muscular tissue can greatly improve both functionality and comfort. In many cases where the aesthetic appearance of a prosthetic is also a factor, the ability to custom shape the exterior to blend with its surrounding or match symmetrical features is an important advantage as well. The ability to further shape internal cavities, add strengthening girders, and shave material off non-load-bearing components can also help improve the weight-to-strength ratio of the device, making it both stronger and lighter. With more sophisticated multimaterial fabrication, one could imagine that the mechanical performance of such devices could be tailored even further, improving their elasticity, shock bearing, and energetic performance. A second example of shaped prostheses is hearing aids. A major factor in the comfort and effectiveness of hearing aids is their fit in the aural canal. With 3D printing, the patient's canal can be optically scanned and a tailored soft prosthesis fabricated almost instantly.

Custom implants. One size fits all is not a good compromise when it comes to your hip-replacement implant. The idea of using 3D printing to fabricate custom-shaped implants dates back to the earliest days of free-form fabrication. Custom-shaped titanium or platinum implants can be fabricated at exactly the right size and shape, either parametrically scaled to fit the patient or even produced directly from a computed tomography (CT) scan geometry of the original bone to be replaced or a symmetrical healthy bone. Since complexity is free, printed implants can go beyond the standard bone shape and include various cavities and connection points that make the bonding with existing tissue more compatible, reliable, and effective.

Bioprinting. Instead of using titanium and other engineering materials, bioprinting involves fabricating implants (and other constructs) from biological materials directly. Early experiments involved 3D printing of biocompatible scaffolds, which were later infused with live cells and incubated before implantation. The live cells gradually replaced the scaffold, resulting in a custom-shaped live tissue-engineered implant. While the use of printed scaffolds is still prevalent, a further advance involves printing with biological cells directly with no scaffold at all. In this case, the cells are first immersed in a biocompatible hydrogel ink—a *bioink*—and then printed into their target form. The bioink has two key, contradicting properties that make it tricky to develop: On the one hand, it needs to be fluid enough to be printed—to flow through the nozzle while not harming the cells that experience severe shear forces as they come out of the printhead. On the other hand, the bioink needs to be stiff enough to hold its shape after printing or else the material will ooze into a shapeless mass. Solving that challenge using a variety of chemical and optical cross-linking agents, we were able to fabricate cartilage implants in the shape of a meniscus directly from CT data. Unlike scaffold infusion techniques, however, the ability to print with live cells directly opens the door to fabrication of heterogeneous tissue implants. Imagine the fabrication of something as complex as a spinal disk or a heart valve, which involves multiple cell types in a complex spatial arrangement that is critical to proper functionality (Cohen et al., 2006).

Drug screening models. The ability to fabricate complex, multicell heterogeneous tissue arrangements in three dimensions has many applications beyond implants. One such exciting opportunity is the fabrication of models for drug screening. Traditionally, drugs and other treatments are tested first in simple petri dish-like environments that tries to replicate the anticipated target environment where the actual drug or treatment will be in use. These petri dishes or test tubes are relatively simple compared to the real environment, both in the range of cell types and in their spatial distribution. This mismatch often necessitates the use of animal models and other more sophisticated high-throughput testing procedures. The use of 3D bioprinting, however, opens the door to fabrication of 3D, spatially heterogeneous tissue models that more closely resemble the target application of a drug. One could imagine, for example, that cancer cells in the shape and distribution of a tumor could be fabricated directly and then the effectiveness of treatments tested directly in vitro. Such experiments could provide a more realistic prediction of drug effectiveness and shorten the cycle of drug development.

Surgical planning. A somewhat less obvious but equally important application of 3D printing is for preparation for surgery. Nonroutine, complex surgical procedures often involve manipulation of tissue and tools through an intertwined and unknown environment. Some operations also require the preparation and attachment of permanent or temporary plates and other devices within the patient's body. Just as it is easier to put a puzzle together the second time around, surgeons can substantially reduce the time and improve the reliability of an operation if they get a chance to practice it beforehand. This is especially true when a complex and nonroutine procedure is executed for the first time. For example, we have often been requested to three-dimensionally print a set of shattered or deformed bones produced from a CT scan of an injured dog coming into emergency surgery at the veterinary school. The surgeons will receive the printed bones and practice the surgery—learning to identify the fragments, optimize the reconstruction process, and prepare any necessary plates and jigs in advance. Informally, surgeons have reported that surgery time can be reduced by half, not to mention the increase in quality of care.

Surgical training. Surgeons in training often get access to state-of-the-art surgical equipment and tools, but rarely do they get a chance to practice on realistic cases. It is unlikely that a training surgeon will get to practice removing a brain tumor, for example, because it is unlikely that an animal model or a human cadaver will be found with such a case or that a synthetic training model happens to available for purchase. With 3D printing, however, relevant and typical cases from real patients can be recorded and reproduced on demand for practice. With some modeling, it is even possible to combine cases and adjust the severity of cases on demand in order to challenge training surgeons at exactly the right level for the best learning experience. With multimaterial bioprinting, the training models could be fabricated with biological materials to provide an even more realistic training experience that also provides the feel and the responsiveness of the real wet tissue.

Custom medication. An excellent example of an application of 3D printing where composition control is key is the fabrication of medical pills on demand.

A growing challenge for both doctors and patients is the administration of multiple medications simultaneously. Instead of keeping track of a dozen pills daily, a printer can now fabricate a single, custom-made pill for each patient. That pill contains the exact amount, combination, and arrangement of medication that a patient needs that day. The pill could have a special identification marking, eliminating much of the confusion and uncertainty associated with conventional delivery methods.

Food printing. Perhaps the unforeseen killer app, food printing is for 3D printing as video gaming is for computers. Based on the very same technology as bioprinting, food printing involves the fabrication of edible items from raw edible inks. The term "edible inks" does not quite do justice here: We are talking about chocolate and peanut butter, cookie dough and frosting. And it does not have to be all bad: You can even print with organic pesto and locally made goat cheese. Download the recipe, load in the frozen food cartridges, and hit print. The recipe dictates which material goes where and what in-line cooking procedure is applied during deposition and after. We have printed with all the materials above and more, creating chocolate confections with frosted decorations and vanilla cookies with vertical text lettering in chocolate inside. Imagine a cookie printer with sliders to adjust the crispiness, flavor, color, and texture in any arbitrary way. How about a checkerboard lasagna? Your imagination is the limit. While some find the idea of printed foods the epitome of processed foods, others find the whole idea fascinating in its vast new opportunities to innovate. Download and share recipes, tailor variants to your liking, and have freshly made dishes just for you.

PRINTING FUNCTION: PROGRAMMING BEHAVIOR OF ACTIVE MATERIALS

The third and final episode of this journey, of which we are only beginning to see early signs, is the control over behavior. In this section we go beyond controlling just the shape of matter, past controlling just its composition; we will now be able to program these materials to function in arbitrary ways—to sense and react, to compute and behave. This last step entails a blurring between material and code (Gershenfeld, 2005), leading to what is essentially *programmable matter*—moving from an object's mechanical functionality to controlling how it processes information and energy as well. When this day comes, you will be able to print virtually anything, from a cellphone to a robot that will walk out of the printer, batteries included. But that robot will not look at all like today's robots because it will not be limited by the constraints imposed by conventional manufacturing, nor will it be designed directly by humans. The ability to manufacture such arbitrary active systems comprising both passive and active substructures will have opened the door to a new space of designs and a new paradigm of engineering, one that is not unlike biology.

THE NEW CAD

For the past four decades, computer-aided design (CAD) tools have played a critical role in the product design process, but to a large extent, CAD tools have remained relatively unchanged in their role and format. User interfaces have improved, geometric manipulations have become faster and more reliable, and graphics have become three-dimensional and photorealistic. But conceptually, CAD software remains today merely a passive 3D drawing board that records our intentions but offers little insight or ideas of its own.

As 3D printing technologies become more abundant, the traditional barriers of resources and skill for manufacturing have all but vanished. The limit is now our imagination, and our imagination is, unfortunately, limited. I have seen over and over how fresh students faced with the blank page of CAD and the unlimited capability of a 3D printer design nothing more than a rectangular object with a few linear notches. To a large extent, this is a cultural blindness evolved out of years of observing mass-produced objects made subject to traditional manufacturing constraints. But on the other hand, this is also due to the design thinking imposed by conventional CAD tools and the lack of new design tools that take advantage of the vast new design space offered by 3D printing capabilities.

It is clear that while the classical CAD paradigm will remain dominant for the foreseeable future, new paradigms for design tools are beginning to emerge (Figure 2).

Function Representations

As our ability to control the shape, composition, and behavior of materials advances, it becomes more appropriate to think about geometry and material specification as programming rather than as drawing. For example, say you would like to fabricate a helical surface with elliptical perforations at some semiperiodic pattern. Making something like that would be a complicated and expensive nightmare using traditional manufacturing techniques, so that kind of capability is buried deep down the advanced CAD options menu, if it can be found at all. It would involve multiple steps and would bring the CAD software to its knees. But unencumbered by traditional manufacturing constraints, a growing number of designers would want to explore exactly those kind of geometries, easily and quickly. And why not? There are, in fact, many more crazy shapes than regular shapes, but they are just currently difficult to explore. The helical perforated shape might be more easily described using a procedural construction process rather than a target geometry (Pasko et al., 2011)—in other words, algorithmic rather than descriptive geometry—much like biology describes phenotypes using a developmental process, or like a software engineer describes the appearance of a dynamic web page.

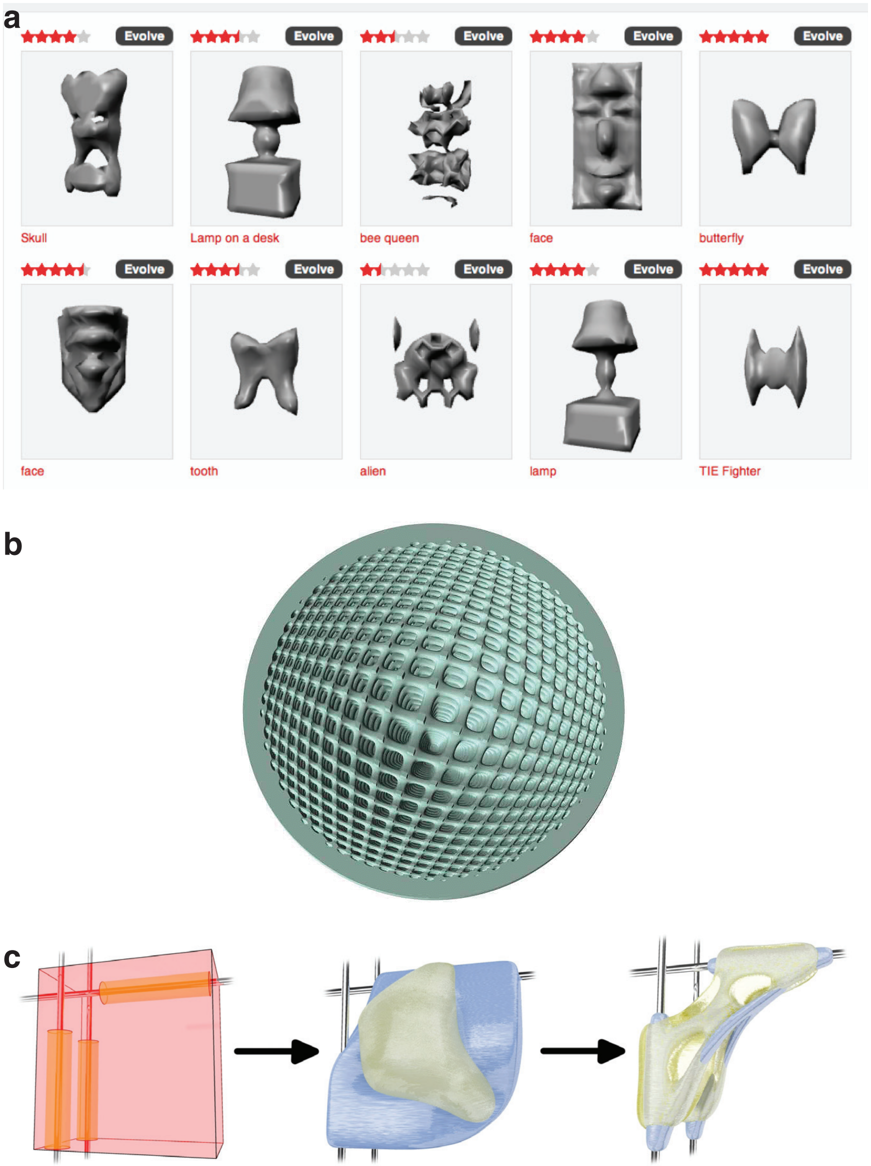


FIGURE 2 The New CAD. Beyond geometric modeling, products are described by interactive evolution, through procedural construction algorithms, or by compiling high-level requirement and constraints automatically. Sources: Figure 2a courtesy of Cornell University and http://endlessforms.com. Figure 2b: Pasko et al., 2011. Courtesy of Turlif Vilbrandt. Figure 2c designed by Jon Hiller. Reprinted courtesy of Cornell University and http://creativemachines.cornell.edu.

Matter Compilers

An alternative approach to design is to specify what the design needs to accomplish rather than what it needs to look like and then to let the machine compile the design to your specification. Some products, especially those that have a purely functional role, would fit well into such a design process. Consider, for example, the task of designing a supporting bracket. We know the geometric constraints of the load and support contact points, the weight to be carried, and the material properties. Specify those requirements, hit the design button, and watch how the optimal design emerges automatically. The optimal design will not be a block with rectangular notches and holes. Instead, an organic-looking optimal structure with beautifully shaped cavities will emerge—something that would take a human designer years to come up with manually. Now say the bracket cannot be installed because of a pipe protruding into the bracket's location. No problem—add the pipe protrusion constraint and recompile. Programming by specifying the target behavior and constraints, then compiling it into functional geometry, can address complex requirements as well as optimally exploit the new manufacturing capabilities afforded by multimaterial 3D printing as they become available. In a way, the designer becomes the customer, and the CAD software becomes the designer.

Interactive Evolution

What happens when the desired object cannot be described quantitatively? Such situations occur frequently in the design of objects that have an aesthetic component that defies quantification. Imagine you need to design a perfume bottle. While some aspects of the design can be specified quantitatively, such as the desired volume and size, others, such as the look and feel, are more difficult to describe. Moreover, those skilled at the art of assessing perfume bottles may have no inclination to dabble in CAD speak. Instead, a new type of CAD software can display a range of initial design concepts and can allow an expert to indicate which ones they like, or to put it in perhaps less gentle terms, which they least dislike. Armed with this information, the machine can begin to infer automatically a sense of aesthetic that the designer has in mind and can generate a new set of solutions. These solutions will still obey the quantitative requirements—the bottle volume and size, say—but they will more closely match the designer's aesthetic preferences. Repeating this process through a sequence of solutions and selections, the designer might be able to design what might be a complex perfume bottle with relatively little CAD expertise. Importantly, the designer might actually come across new ideas and be provoked into new corners of the design space where no person has been before. Using this process, visitors to the EndlessForms website (Clune and Lipson, 2011) have collaboratively designed objects from furniture to faces, and from bottles to butterflies.

FabApps

What if I want to design a toothbrush to print on my 3D printer today? It is unlikely that I will be able to design a good, ergonomic, and safe toothbrush. Even though a toothbrush seems like a simple product, it takes years of experience and know-how to design a successful one. Yet with the advent of 3D printers, it is likely that people will want to do exactly that—design their own products with almost no experience and little patience to learn. The solution may be in the form of simple CAD applications dedicated to a narrow product and which encapsulate all the relevant knowledge, yet expose just the right level of flexibility to the user. Such FabApps (term coined by Daniel Cohen and Jeffrey Lipton), similar to the iPhone apps that you can download for 99 cents, specialize in the design of one particular product. In contrast with general-purpose CAD, FabApps use extensive built-in know-how to guide a user through the design of one item and ensure the resulting product will be successful. A toothbrush app can ask for dimensions of your hand and mouth, process pictures of your face and palm, walk you through fifty different options and ask twenty more questions, and then produce the perfect toothbrush that uniquely fits your needs and is guaranteed to be a success.

CONCLUSION

3D printing is only the tip of a series of milestone in humans' increasing control over physical matter. If humans distinguished themselves from their evolutionary ancestors by making tools, then additive manufacturing represents the ultimate tool—perhaps changing human culture forever in ways we can hardly anticipate.

REFERENCES

Clune, J., and H. Lipson. 2011. Evolving three-dimensional objects with a generative encoding inspired by developmental biology. Proceedings of the 2011 European Conference on Artificial Life, Paris, France, August 8–12, 2011. Available at *http://EndlessForms.com.*

Cohen, D. L., E. Malone, H. Lipson, and L. Bonassar. 2006. 3D direct printing of heterogeneous tissue implants. Tissue Engineering 12(5):1325–1335.

Gershenfeld, N. 2005. FAB: The Coming Revolution on Your Desktop—From Personal Computers to Personal Fabrication. New York: Basic Books.

Lipson, H., and M. Kurman. 2010. Factory@Home: The emerging economy of personal fabrication. White House Office of Science and Technology Policy. Available at *http://web.mae.cornell.edu/lipson/FactoryAtHome.pdf.*

Pasko, A., O. Fryazinov, T. Vilbrandt, P.-A. Fayolle, and V. Adzhiev. 2011. Procedural function-based modelling of volumetric microstructures. Graphical Models 73(5):165–181.

ADDITIONAL READING

Carrabine, L. 2009. Explore the range of DMLS of titanium. Available at *http://www.makepartsfast.com/2009/12/828/explore-the-range-of-dmls-of-titanium/*.

Hiller, J., and H. Lipson. 2009. Design automation for multi-material printing. Paper presented at the Solid Freeform Fabrication Symposium (SFF'09), Austin, Texas, August 3–5, 2009. Available at *http://creativemachines.cornell.edu/sites/default/files/SFF09_Hiller2.pdf*.

Knapp, M., R. Wolff, and H. Lipson. 2008. Developing printable content: A repository for printable teaching models. Paper presented at the 19th Annual Solid Freeform Fabrication Symposium, Austin, Texas, August 2008. Available at *http://creativemachines.cornell.edu/papers/SFF08_Knapp.pdf*.

Lipson, H., and J. B. Pollack. 2000. Automatic design and manufacture of robotic lifeforms. Nature 406:974-978. Available at *http://creativemachines.cornell.edu/papers/Nature00_Lipson.pdf*.

Lipton, J.I., D. Arnold, F. Nigl, N. Lopez, D.L. Cohen, N. Noren, and H. Lipson. 2010. Multi-material food printing with complex internal structure suitable for conventional post-processing. Paper presented at the 21st Solid Freeform Fabrication Symposium (SFF'10), Austin, Texas. Available at *http://creativemachines.cornell.edu/sites/default/files/69-Lipton-Mutlimaterial%20food%20printing%20Final.pdf*.

Richter, C., and H. Lipson. 2010. Untethered hovering flapping flight of a 3D-printed mechanical insect. 12th International Conference on Artificial Life (Alife XII), Odense, Denmark, August 2010, Pp. 797–803. Available at *http://creativemachines.cornell.edu/sites/default/files/Alife10_Richter.pdf*.

SEMANTIC PROCESSING

Introduction

Aleksandar Kuzmanovic
Northwestern University

Amarnag Subramanya
Google Research

Semantics is the study of meaning. A large number of naturally occurring phenomena follow certain semantic rules, for example, the semantics of human speech, semantics associated with an image of a scene, and the semantics of natural language. Accurate semantic processing is required for a number of high-level information-understanding tasks such as inferring author sentiment given a blog or review; searching through a collection of documents, images, and videos; and translating text from one language to another. For example, it may be hard to infer the positive sentiment expressed by the statement, "*The Prince of Egypt* succeeds where other movies have failed," without the aid of semantics-based inference.

In the past few years, there has been an explosion in the amount of human-generated content on the Internet and exponential growth in the number of times a user turns to the Internet to perform a daily activity. It is estimated that we create about 1.6 billion blog posts, 60 billion emails, 2 million photographs, and 200,000 videos on the Internet every day. These days, users read the news, watch television, and stay connected to their friends and family via the Internet, yet users' need for Internet-based applications is now greater than ever before. Satisfying these ever-increasing demands requires a deeper semantic understanding of all the content on the Web. This session focuses on semantics processing algorithms for natural language and images since they constitute a large majority of the data on the Internet.

In the context of natural language, there are many different levels of semantic processing, ranging from word- and sentence-level analysis to more complex analysis of discourse. The task of understanding the meaning of words and their relationships falls under the former; whereas, the ability to infer the meaning of pronouns (e.g., he, she) and inferring sentiment expressed by a paragraph are examples of the latter. Ani Nenkova (University of Pennsylvania) begins with a survey of some of the techniques that have been successfully applied to automatic text understanding and will point out some of the outstanding challenges. She also sheds light on the impact that text quality has on semantic processing algorithms.

The proliferation of Internet use has led to the creation of large bodies of knowledge such as Wikipedia. Furthermore, the social aspect of the Web has resulted in collaboratively generated content (e.g., Yahoo! Answers). Accurate semantic processing of such sources of knowledge can lead to knowledge-rich approaches to information access that go far beyond the conventional word-based methods. Evgeniy Gabrilovich (Yahoo! Research) describes using collaboratively generated content for representing the semantics of natural language and presents new information retrieval algorithms enabled by this representation.

Images and video form a key component of the overall Internet experience. Accurate semantic understanding of images and video can lead to faster and better search. Samy Bengio (Google Research) discusses algorithms that learn how to “embed” images and their descriptions (labels or annotations) within a common space. Such a space can be used to find the nearest annotations to a given image. He shows how one can construct a “visio-semantic” tree from such annotations.

Tables, plots, graphs, and diagrams are yet another way information is represented on web pages. These data-driven images are complicated objects that have a close relationship with the surrounding text. For example, they may be used to illustrate the text’s conclusions or provide additional data. Unfortunately, state-of-the-art algorithms treat diagrams in the same way as photos or illustrations. As a result, searching for a relevant diagram online often yields very poor quality results. Michael Cafarella (University of Michigan) covers smart semantic processing algorithms for plots, graphs, and diagrams. He also discusses ways such data can be summarized to make it easier for end-user consumption.

Automatic Text Understanding of Content and Text Quality

Ani Nenkova
University of Pennsylvania

Reading involves two rather different kinds of semantic processing. One is related to understanding what information is conveyed in the text and the other to appreciating the style of the text—how well or poorly it is written. For people, text content and stylistic quality are inextricably linked. For machines, robust understanding of written material has become feasible in many contexts but text quality has been out of reach so far. The mismatch matters a great deal because people rely on machines to locate and navigate information sources and increasingly read machine-generated text, for example as machine translations or text summaries.

In this presentation I discuss some of the simple and elegant intuitions that have enabled semantic processing in machines, as well as some of the emerging directions in text quality assessment.

TEXT SEMANTICS (MEANING)

Reading and Understanding the Web

A single insight about language semantics has led to successes in a variety of automatic text understanding tasks. Words tend to appear in specific contexts and these contexts convey rich information about the type of word, its meaning, and connotation (Harris, 1968). Computers can learn much semantic information without human supervision simply by collecting statistics of (hundreds of) thousands of texts.

The context of a target word, consisting of other phrases or words that occur nearby in texts more often than expected by chance, is accumulated over large text collections. For example, the word `tea` may be characterized by the context [drink:60, green:55, milk:40, sip:30, enjoy:10, . . .]. Each entry shows a word that appeared five words before or after tea, and the number of times the pair was seen in a large text collection. Taking just the number of occurrences of context words makes the representation even more convenient, because various standard (geometric) approaches exist for comparing the distance between numeric vectors. In this manner, a machine can compute the similarity between any two words.

Here is an example from Pantel and Lin (2002) of the 15 words most similar to wine computed by this approach:

> **Wine**: beer, white wine, red wine, Chardonnay, champagne, fruit, food, coffee, juice, Cabernet, cognac, vinegar, Pinot noir, milk, vodka, . . .

The list may not look immediately useful but is certainly impressive if one considers how little similarity there is in the sequence of letters wine, beer, Chardonnay.

Building upon these representations, it has become possible to automatically discover words with multiple senses by clustering words similar to them (plant: (plant, factory, facility, refinery) (shrub, ground cover, perennial, bulb)), finding synonyms and antonyms. To aid analysis of customer reviews, researchers at Google developed a large lexicon of almost 200,000 positive and negative words and phrases, identified through their similarity to a handful of predefined positive or negative words such as excellent, amazing, bad, horrible. Among the positive phrases in the automatically constructed lexicon were cute, fabulous, top of the line, melt in your mouth; negative examples included subpar, crappy, out of touch, sick to my stomach (Velikovich et al., 2010).

Another line of research in semantic processing exploits the stable meaning of some contexts. For example, patterns like "X such as Y," if occurring often in texts, is very likely an indicator that Y is a kind of X (i.e., "Red wines such as Cabernet and Pinot noir . . ."). Similarly a phrase like "The mayor of X" is a good indicator that X is a city. NELL (Never Ending Language Learning, *http://rtw.ml.cmu.edu/rtw/*) is a system that constantly learns unary and binary predicates, corresponding to categories and relations such as isCity(Philadelphia) and playsInstrument(George_Harrison, guitar). The learning of each type of fact starts with minimal supervision in the form of several examples of category instances or entities between which a relation holds, given by the researchers. Then the system starts an infinite loop in which it finds web pages that contain the examples, finds phrase patterns that typically occur with the examples, selects the best patterns that indicate the predicate with high probability, and then applies the patterns to new texts to discover more instances for which the predicate is true. Different flavors of this approach to machine understanding have been developed to help search and question answering (Etzioni et al., 2008, Pasca et al., 2006).

Reading and Understanding a Text

In the semantic processing I have discussed so far, the computer reads numerous textual documents with the objective to learn representations of words, come up with a lexicon of phrases with positive or negative connotation, or learn category instances and relations. A more difficult task for a computer is to understand a specific text.

Much traditional research related to computer processing of a single text has relied on supervised techniques. Researchers invested effort to prepare collections in which human annotators marked positive and negative examples of a semantic distinction of interest. For example, they could mark the different senses of a word, the part of speech of words, or would mark that Roger Federer is a person, Bulgaria is a country. Then features describing the context of the categories of interest would be extracted from the text, and a statistical classifier would use the positive and negative examples to combine the features and predict the same type of information on unseen text. More recently it has become clear that the unsupervised approach in which computers accumulate knowledge and statistics from large amounts of text and the supervised approach can be combined effectively and result in better systems for semantic processing.

When reading a specific text, computers also need to resolve what entity in the document is referred to by pronouns such as "he/his," "she/her," and "it/its." Systems are far from perfect but are getting better at this task. Usually pronouns appear in the text near noun phrases, i.e., "the *professor* prepared *his* lecture," but in other situations gender and number information is necessary to correctly resolve the pronoun, as in, "*John* told Mary *he* had booked the trip." Machines can rather accurately learn the likely gender of names and nouns, again by reading large volumes of text and collecting statistics of co-occurrence. Statistics about the co-occurrence of a pronoun of a given gender and the immediately preceding noun or honorifics and names (Mr. John Black, Mrs. Mary White), collected over thousands of documents, give surprisingly good guesses about the likely gender of nouns (Bergsma, 2005).

TEXT QUALITY (STYLE)

Automatic assessment of text quality, or style, is a far more difficult task compared to the acquisition of semantics, or at least considerably less researched. Much of the effort in my lab has been focused on developing models of text quality. I will discuss two successful endeavors: prediction of general and specific sentences and automatic assessment of sentence fluency in machine translation and summary coherence in text summarization.

A well-written text contains a balanced mix of general overview statements and specific detailed sentences. If a text contains too many general sentences it will be perceived as insufficiently informative, and too much specificity can be confusing for the reader.

To train a classifier, we exploit a resource of 1 million words of *Wall Street Journal* text with discourse annotations (Louis and Nenkova, 2011a). The discourse annotations, among other things, specify the way two adjacent sentences in the text are related. There could be an implicit comparison between two statements (John is always punctual. Mary often arrives late.), or a contingency (causal) relation (I hurt my foot. I cannot go dancing tonight.), or temporal relations.

One of the discourse relations annotated in the corpus is instantiation. It holds between two adjacent sentences where the second gives a specific example of information mentioned in the first, as in, "He is very smart. He solved the problem in five minutes." We considered that the first sentence is general while the second is specific in all instances of instantiation relation. We computed a number of features that according to our intuition would distinguish between the two categories. We expected that the presence of opinion or evaluative statements would characterize the general sentences as well as unusual use of language that would later be interpreted or clarified in a specific sentence. Among the features were

- the length of the sentence.
- the number of opinion or subjective words, derived from existing dictionaries.
- the specificity of words in the sentences, derived from corpus statistics as the fraction of documents in one year of *New York Times* articles that contain the word. The fewer documents contain the word, the more specific it is.
- mentions of numbers and people, companies, and geographical locations; such mentions are detected automatically.
- syntactic features related to adjectives, adverbs, verbs, and prepositions.
- probabilities of sequences of one, two, or three consecutive words computed over one year of *New York Times* articles.

A logistic regression classifier, trained on around 2,800 examples of general and specific sentences from instantiation relations, learned to predict the distinction incredibly well. On a completely independent set of news articles, five different people were asked to mark each sentence as general or specific. For sentences in which all five annotators agreed about the class, the classifier can predict the correct class with 95 percent accuracy. For examples on which only four out of the five annotators agreed, the accuracy is 85 percent. For all examples, which included sentences for which people found it hard to classify in terms of general and specific, the accuracy of prediction was 75 percent. Moreover, the confidence of the classifier turned out to be highly correlated with annotator agreement, so it was possible to identify which sentences would not fit squarely into one of the classes. The degree of specificity of a sentence given by the classifier gives an accurate indication of how a sentence will be perceived by people.

Applying the general-or-specific classifier to samples of automatic and human summaries of clusters of news articles has demonstrated that machine summaries are overly specific and has indicated ways for improving system performance (Louis and Nenkova, 2011b).

Word co-occurrence statistics and subjective language have also been successful in automatically distinguishing implicit comparison, contingency, and temporal discourse relations (Pitler et al., 2009). Identification of such relations is not only necessary for semantic processing of text, it is also required for robust assessment of text quality (Pitler and Nenkova, 2008). Finally, statistics on types, length, and distance between verb, noun, and prepositional phrases, as well as probabilities of occurrence and co-occurrence of words, are highly predictive of the perceived quality of summaries (Nenkova et al., 2010).

REFERENCES

Bergsma, S. 2005. Automatic acquisition of gender information for anaphora resolution. Proceedings of the 18th Conference of the Canadian Society for Computational Studies of Intelligence, Victoria, Canada, May 9–11, 2005, pp. 342–353.

Etzioni, O., M. Banko, S. Soderland, and D. S. Weld. 2008. Open information extraction from the web. Communications of the ACM 51(12):68–74.

Harris, Z. S. 1968. Mathematical Structures of Language. New York: Wiley.

Louis, A., and A. Nenkova. 2011a. Automatic identification of general and specific sentences by leveraging discourse annotations. Proceedings of the International Joint Conference in Natural Language Processing, Chiang Mai, Thailand, November 8–13, 2011.

Louis, A., and A. Nenkova. 2011b. Text specificity and impact on quality of news summaries. Proceedings of the Workshop on Monolingual Text-To-Text Generation, Portland, Oregon, June 24, 2011, pp. 34–42.

Nenkova, A., J. Chae, A. Louis, and E. Pitler. 2010. Structural features for predicting the linguistic quality of text: Applications to machine translation, automatic summarization and human-authored text. Empirical Methods in Natural Language Generation, edited by E. Krahmer and M. Theune. Lecture Notes in Artificial Intelligence, Vol. 5790. Berlin: Springer-Verlag. pp. 222–241.

Pantel, P., and D. Lin. 2002. Discovering word senses from text. Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2002, Edmonton, Canada, July 23–26, 2002, pp. 613–619.

Pasca, M., D. Lin, J. Bigham, A. Lifchits, and A. Jain. 2006. Names and similarities on the web: Fact extraction in the fast lane. Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia, July 2006.

Pitler, E., and A. Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. Proceedings of the Conference on Empirical Methods in Natural Language Processing, Waikiki, Hawaii, October 25–27, 2008.

Pitler, E., A. Louis, and A. Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, August 2–7, 2009, pp. 683–691.

Velikovich, L., S. Blair-Goldensohn, K. Hannan, and R. McDonald. 2010. The viability of web-derived polarity lexicons. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, Calif., June 1–6, 2010, pp. 777–785.

Advancing Natural Language Understanding with Collaboratively Generated Content

Evgeniy Gabrilovich
Yahoo! Research

Proliferation of ubiquitous access to the Internet enables millions of Web users to collaborate online in a variety of activities. Many of these activities result in the construction of large repositories of knowledge, either as their primary aim (e.g., Wikipedia) or as a by-product (e.g., Yahoo! Answers). In this paper, we discuss how to use the cornucopia of world knowledge encoded in the repositories of collaboratively generated content (CGC) for advancing computers' ability to process human language.

Prior to the advent of CGC repositories, many computational approaches to natural language employed the WordNet electronic dictionary (Fellbaum, 1998), which covers approximately 150,000 words painstakingly encoded by professional linguists over the course of more than 20 years. In contrast, the collaborative Wiktionary project (*www.wiktionary.org*) includes more than 2.5 million words in English alone. Encyclopaedia Britannica, published since 1798, has approximately 65,000 articles, while Wikipedia has over 3.7 million articles in English and over 15 million articles in over 200 other languages. Ramakrishnan and Tomkins (2007) estimated the amount of user-generated content produced worldwide on a daily basis to be 8-10 gigabytes, and this amount has likely increased considerably since then.

REPOSITORIES OF COLLABORATIVELY GENERATED CONTENT AS AN ENABLING RESOURCE

The unprecedented amounts of information in CGC enable new, knowledge-rich approaches to natural language processing, which are significantly more powerful than the conventional word-based methods. Considerable progress has been made in this direction over the past few years. Examples include explicit manipulation of human-defined concepts and their use to augment the bag of words in information retrieval (Egozi et al., 2011), or using Wikipedia for better word sense disambiguation (Bunescu and Pasca, 2006; Cucerzan, 2007).

One way to use CGC repositories is to treat them as huge additional corpora, for instance, to compute more reliable term statistics or to construct comprehensive lexicons and gazetteers. They can also be used to extend existing knowledge repositories, increasing the concept coverage and adding usage examples for previously listed concepts. Some CGC repositories, such as Wikipedia, record each and every change to their content, thus making the document authoring process directly observable. This abundance of editing information allows us to come up with better models of term importance in documents, assuming that terms introduced earlier in the document life are more central to its topic. The recently proposed Revision History Analysis (Aji et al., 2010) captures this intuition to provide more accurate retrieval of versioned documents.

An even more promising research direction, however, is to distill the world knowledge from the structure and content of CGC repositories. This knowledge can give rise to new representations of texts beyond the conventional bag of words and allow reasoning about the meaning of texts at the level of concepts rather than individual words or phrases. Consider, for example, the following text fragment: "Wal-Mart supply chain goes real time." Without relying on large amounts of external knowledge, it would be quite difficult for a computer to understand the meaning of this sentence. Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2009) offers a way to consult Wikipedia in order to fetch highly relevant concepts such as "Sam Walton" (the Wal-Mart founder); "Sears," "Target," and "Albertsons" (prominent competitors of Wal-Mart); "United Food and Commercial Workers" (a labor union that has been trying to organize Wal-Mart's workers); and "hypermarket" and "chain store" (relevant general concepts). Arguably, the most insightful concept generated by consulting Wikipedia is "RFID" (radio frequency identification), a technology extensively used by Wal-Mart to manage its stock. None of these concepts are explicitly mentioned in the given text fragment, yet when available they help shed light on the meaning of this short text.

In the remainder of this article, I first discuss using CGC repositories for computing semantic relatedness of words and then proceed to higher-level applications such as information retrieval.

COMPUTING SEMANTIC SIMILARITY OF WORDS AND TEXTS

How related are "cat" and "mouse?" And what about "preparing a manuscript" and "writing an article?" Reasoning about semantic relatedness of natural language utterances is routinely performed by humans but remains challenging for computers. Humans do not judge text relatedness merely at the level of text words. Words trigger reasoning at a much deeper level that manipulates concepts—the basic units of meaning that serve humans to organize and share their knowledge. Thus, humans interpret the specific wording of a document in the much larger context of their background knowledge and experience.

Prior work on semantic relatedness was based on purely statistical techniques that did not make use of background knowledge (Deerwester et al., 1990), or on lexical resources that incorporate limited knowledge about the world (Budanitsky and Hirst, 2006). CGC-based approaches differ from the former in that they manipulate concepts explicitly defined by humans, and from the latter in the sheer number of concepts and the amount of background knowledge. One class of new approaches to computing semantic relatedness uses the structure of CGC repositories, such as category hierarchies (Strube and Ponzetto, 2006) or links among the concepts (Milne and Witten, 2008). Given a pair of words whose relatedness needs to be assessed, these methods map them to relevant concepts, (e.g., articles in Wikipedia) and then use the structure of the repository to compute the relatedness between these concepts. Gabrilovich and Markovitch (2009) proposed an alternative approach that uses the entire content of Wikipedia and represents the meaning of words and texts in the space of Wikipedia concepts. Their method, ESA, represents texts as weighted vectors of concepts. The meaning of a text fragment is thus interpreted in terms of its affinity with a host of Wikipedia concepts. Computing semantic relatedness of texts then amounts to comparing their vectors in the space defined by the concepts, for example, using the cosine metric.

Subsequently proposed approaches offer ways to combine the structure-based and concept-based methods in a principled manner (Yeh et al., 2009). Beyond Wikipedia, Zesch et al. (2008) proposed a method for computing semantic relatedness of words using Wiktionary. Recently, Radinsky et al. (2011) proposed a way to augment the knowledge extracted from CGC repositories with temporal information by studying patterns of word usage over time. Consider, for example, an archive of the *New York Times* spanning 150 years. Two words such as "war" and "peace" might rarely co-occur in the same articles, yet their patterns of use over time might be similar, which allows us to better judge their true relatedness.

CONCEPT-BASED INFORMATION RETRIEVAL

Information retrieval systems traditionally rely on textual keywords to index and retrieve documents. Keyword-based retrieval may return inaccurate and incomplete results when different keywords are used to describe the same concept in the documents and in the queries. Furthermore, the relationship between those related keywords may be semantic rather than syntactic, and capturing it thus requires access to comprehensive human world knowledge. Previous approaches have attempted to tackle these difficulties by using manually built thesauri, by relying on term co-occurrence data, or by extracting latent word relationships and concepts from a corpus. ESA introduced in the previous section, which represents the meaning of texts in a very high-dimensional space of

Wikipedia concepts, has been shown to offer superior performance over the previous state-of-the-art algorithms. In contrast to the task of computing semantic relatedness, which usually deals with short texts whose overlap is often empty, information retrieval usually deals with longer documents. It is noteworthy that in such cases optimal results can be obtained by extending the bag of words with concepts rather than merely relying on the conceptual representation alone.

Intuitively, one might expect domain-specific knowledge to be key for processing texts in terminology-rich domains such as medicine. However, as Gabrilovich and Markovitch (2007) showed, it is the general purpose knowledge that leads to much higher improvements in text classification accuracy. In the follow-up article (Gabrilovich and Markovitch, 2009), the authors also showed that using larger repositories of knowledge,(e.g., later Wikipedia snapshots) leads to superior performance as more knowledge becomes available.

Potthast et al. (2008) and Sorg and Cimiano (2008) independently proposed CL-ESA, a cross-lingual extension to ESA. Using cross-language links available between a growing number of Wikipedia articles, the approach allows to map the meaning of texts across different languages. This allows, for example, formulating a query in one language and then using it to retrieve documents written in a different language.

CONCLUSION

Publicly available repositories of collaboratively generated content encode massive amounts of human knowledge about the world. In this paper, we showed that the structure and content of these repositories can be used to augment representation of natural language texts with information that cannot be deduced from the input text alone.

Using knowledge from CGC repositories leads to double-digit accuracy improvements in a range of tasks, from computing semantic relatedness of words and texts to information retrieval and text classification. The most important aspects of using exogenous knowledge are its ability to address synonymy and polysemy, which are arguably the two most important problems in natural language processing. The former manifests itself when two texts discuss the same topic using different words, and the conventional bag-of-words representation is not able to identify this commonality. On the other hand, the mere fact that the two texts contain the same polysemous word does not necessarily imply that they discuss the same topic, since that word could be used in the two texts in two different meanings. We believe that concept-based representations are so successful because they allow generalizations and refinements, which partially address synonymy and polysemy.

REFERENCES

Aji, A., Y. Wang, E. Agichtein, and E. Gabrilovich. 2010. Using the past to score the present: Extending term weighting models with Revision History Analysis. Proceedings of the 19th ACM Conference on Information and Knowledge Management, Toronto, Canada, October 26–30, 2010.

Budanitsky, A., and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics 32(1):13–47.

Bunescu, R., and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April 3–7, 2006.

Cucerzan, S. 2007. Large-scale named entity disambiguation based on Wikipedia data. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, June 28–30, 2007.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman, 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6):391–407.

Egozi, O., E. Gabrilovich, and S. Markovitch. 2011. Concept-based information retrieval using Explicit Semantic Analysis. ACM Transactions on Information Systems 29(2):1–34.

Fellbaum, C. 1998. WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press.

Gabrilovich, E., and S. Markovitch. 2007. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. Journal of Machine Learning Research 8:2297–2345.

Gabrilovich, E., and S. Markovitch. 2009. Wikipedia-based semantic interpretation for natural language processing. Journal of Artificial Intelligence Research 34:443–498.

Milne, D., and I. A. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. Proceedings of the 2008 AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy, Chicago, Ill., July 13, 2008.

Potthast, M., B. Stein, and M. Anderka. 2008. A Wikipedia-based multilingual retrieval model. Proceedings of the European Conference on Information Retrieval, Glasgow, Scotland, March 30–April 3, 2008.

Radinsky, K., A. Agichtein, E. Gabrilovich, and S. Markovitch. 2011. A word at a time: Computing word relatedness using Temporal Semantic Analysis. Proceedings of the 20th International World Wide Web Conference, Hyderabad, India, April 16–20, 2011.

Ramakrishnan, R., and A. Tomkins. 2007. Toward a PeopleWeb. IEEE Computer 40(8):63–72.

Sorg, P., and P. Cimiano. 2008. Cross-lingual information retrieval with Explicit Semantic Analysis. Working Notes for the CLEF Workshop, Aarhus, Denmark, September 17–19, 2008.

Strube, M., and S. P. Ponzetto. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. Proceedings of the 21st National Conference on Artificial Intelligence, Boston, Mass., July 16–20, 2006.

Yeh, E., D. Ramage, C. D. Manning, E. Agirre, and A. Soroa. 2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. Proceedings of the 2009 Workshop on Graph-Based Methods for Natural Language Processing. Singapore, August 7, 2009.

Zesch, T., C. Mueller, and I. Gurevych. 2008. Using Wiktionary for computing semantic relatedness. Proceedings of the 23rd AAAI Conference on Artificial Intelligence, Chicago, Ill., July 13–17, 2008.

Large-Scale Visual Semantic Extraction

Samy Bengio[1]
Google Research

ABSTRACT

Image annotation is the task of providing textual semantic to new images, by ranking a large set of possible annotations according to how they correspond to a given image. In the large-scale setting, there could be millions of images to process and hundreds of thousands of potential distinct annotations. In order to achieve such a task, we propose to build a so-called embedding space, into which both images and annotations can be automatically projected. In such a space, one can then find the nearest annotations to a given image or annotations similar to a given annotation. One can even build a visiosemantic tree from these annotations that corresponds to how concepts (annotations) are similar to each other with respect to their visual characteristics. Such a tree will be different from semantic-only trees, such as WordNet, which do not take into account the visual appearance of concepts.

INTRODUCTION

The emergence of the Web as a tool for sharing information has caused a massive increase in the size of potential data sets available for machines to learn from. Millions of images on web pages have tens of thousands of possible annotations in the form of HTML tags that can be conveniently collected by querying search engines (Torralba et al., 2008), tags such as in *www.flickr.com*, or human-curated

[1] This work summarizes the following papers: Weston et al. (2010) with Samy Bergio and Nicolas Usunier, and Bengio et al. (2010) with Jason Weston and David Grangier.

labels such as in *www.image-net.org* (Deng et al., 2009). We therefore need machine learning algorithms for image annotation that can scale to learn from and annotate such data. This includes (i) scalable training and testing times and (ii) scalable memory usage. In the ideal case we would like a fast algorithm that fits on a laptop, at least at annotation time. For many recently proposed models tested on small data sets, it is unclear if they satisfy these constraints.

In the first part of this work, we study feasible methods for just such a goal. We consider models that learn to represent images and annotations jointly in a low-dimension embedding space. Such embeddings are fast at testing time because the low dimension implies fast computations for ranking annotations. Simultaneously, the low dimension also implies small memory usage. To obtain good performance for such a model, we propose to train its parameters by learning to rank, optimizing for the top annotations in the list, for example, optimizing precision at k (p@k).

In the second part of this work, we propose a novel algorithm to improve testing time in multiclass classification tasks where the number of classes (or labels) is very large and where even a linear algorithm in the number of classes can become computationally infeasible. We propose an algorithm for learning a tree structure of the labels in the previously proposed joint embedding space, which, by optimizing the overall tree loss, provides a superior accuracy to existing tree labeling methods.

JOINT EMBEDDING OF IMAGES AND LABELS

We propose to learn a mapping into a feature space where images and annotations are both represented, as illustrated in Figure 1. The mapping functions are therefore different but are learned jointly to optimize the supervised loss of interest for our final task, that of annotating images. We start with a representation of images $x \in \Re^d$ and a representation of annotations $i \in \Upsilon = \{1,\ldots,Y\}$, indices into a dictionary of possible annotations. We then learn a mapping from the image feature space to the joint space $\Re^D$,

$$\Phi_I(x) : \Re^d \rightarrow \Re^D,$$

while jointly learning a mapping for annotations,

$$\Phi_W(i) : \{1,\ldots,Y\} \rightarrow \Re^D.$$

These are chosen to be linear maps, i.e., $\Phi_I(x) = Vx$ and $\Phi_W(i) = W_i$, where W_i indexes the *i*th column of a $D \times Y$ matrix, but potentially any mapping could be used. In our work, we use sparse high-dimensional feature vectors of bags of visual terms for image vectors *x* and each annotation has its own learned representation (even if, for example, multiword annotations share words). Our goal is, for a given image, to rank the possible annotations such that the highest ranked

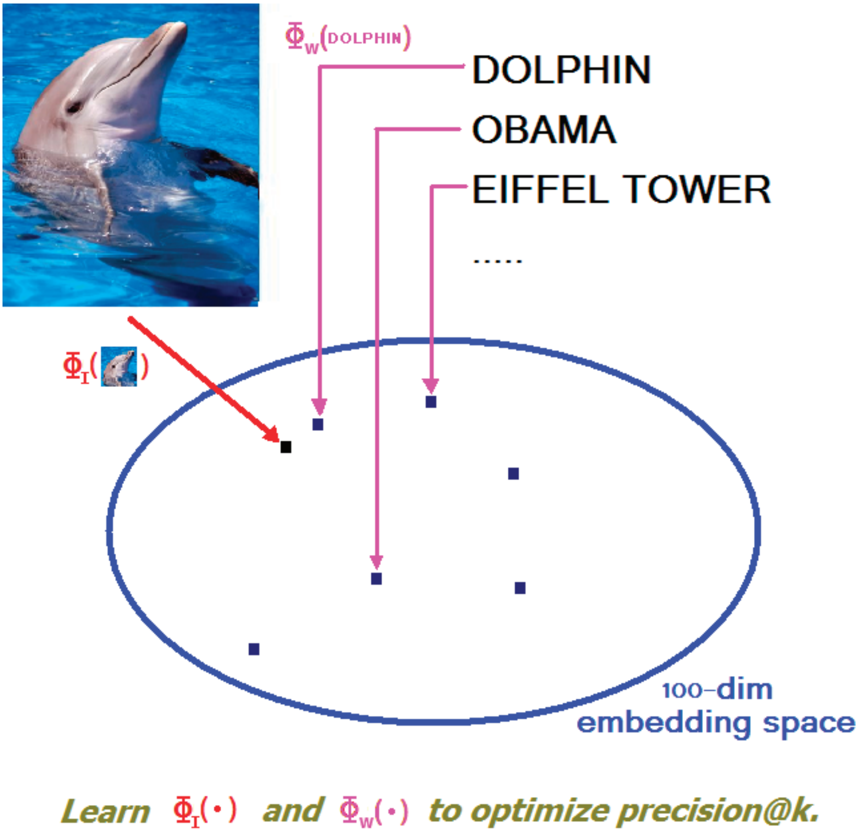


FIGURE 1 Joint embedding space.

annotations best describe the semantic content of the image. We consider the following model:

$$f_i(x) = \Phi_W(i)^T \Phi_I(x) = W_i^T V x,$$

where the possible annotations i are ranked according to the magnitude of $f_i(x)$, largest first.

Image Labeling as a Learning-To-Rank Task

Labeling an image can be viewed as a ranking task where, given an image, one needs to order labels such that the top ones correspond to the image while the bottom ones are unrelated to it. Various learning-to-rank methods have been proposed in the machine learning literature over the years, some of which can scale to large data sets while others cannot. The simplest scalable approach is the following. One can decompose the ranking task as a large sum of several smaller tasks:

$$loss = \sum_{x} \sum_{y^+} \sum_{y^-} \left| 1 - f_{y^+}(x) + f_{y^-}(x) \right|_+ ,$$

where for each training image x, we want the score of each good label y^+ to be higher than the score of any bad label y^- by a margin of at least 1, otherwise we pay the corresponding price. This loss can be trained very efficiently on very large data sets using stochastic gradient descent. However, a better alternative would be a loss that concentrates on the top of the ranking, instead of considering every triplet (x,y^+,y^-) uniformly. In a previous study (Weston et al., 2010), we proposed the weighted approximate-rank pairwise (WARP) loss, which can weigh each of the triplets according to the estimated rank of the good labels and still yields an efficient implementation. The resulting model is much faster to train and obtains a much better performance at the top of the ranking.

Large-Scale Learning

We trained an embedding model with the WARP loss on a very large data set, containing more than 10 million training images and more than 100,000 labels, where labels correspond to queries uttered on Google Image Search and images attributed to these labels were images often clicked for these queries. That meant a very noisy data set, where queries are in several languages, with many spelling mistakes and many apparently similar queries.

An interesting side effect of training such a model is that it provides a natural way of organizing labels among themselves, by looking at the nearest labels of a given label in the embedding space. Table 1 shows some examples of the nearest labels of some labels, where we see several misspellings, translations, and semantically similar labels.

Finally, we show in Table 2 examples of images from our test set (there were more than 3 million of them, different from the 10 million training images), as well as the nearest 10 labels in the embedding space.

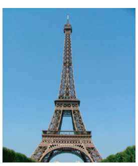
LEARNING LABEL TREES

Labeling images when the number of labels is large (in the section Joint Embedding of Images and Labels, we had on the order of 100,000 labels) can be prohibitive for real-time applications. We thus proposed (Bengio et al., 2010) a novel approach to learn a label tree, where each node makes a prediction of the subset of labels to be considered by its children, thus decreasing the number of labels at a logarithmic rate until a prediction is reached. Existing approaches (Beygelzimer et al., 2009a, 2009b; Hsu et al., 2009) typically lose accuracy com-

TABLE 1 Nearest labels in the embedding space learned on the Web data.

Target Label	**Nearest Labels**
Barack obama	barak obama, obama, barack, barrack obama, bow wow, george bush
david beckham	beckham, david beckam, alessandro del piero, del piero, david becham
Santa	santa claus, papa noel, pere noel, santa clause, joyeux noel, tomte
dolphin	delphin, dauphin, whale, delfin, delfini, baleine, blue whale, walvis
Cows	cattle, shire, dairy cows, kuh, horse, cow, shire horse, kone, holstein
Rose	rosen, hibiscus, rose flower, rosa, roze, pink rose, red rose, a rose
pine tree	abies alba, abies, araucaria, pine, neem tree, oak tree, pinus sylvestris
Mount fuji	mt fuji, fujisan, fujiyama, mountain, zugspitze, fuji mountain
eiffel tower	eiffel, tour eiffel, la tour eiffel, big ben, paris, blue mosque, eifel tower
Ipod	i pod, ipod nano, apple ipod, ipod apple, new ipod, ipod shuffle
f18	f 18, eurofighter, f14, fighter jet, tomcat, mig21, f 16

TABLE 2 Examples of the top 10 labels obtained for some test images.

Target Image	**Nearest Labels**	**Target Image**	**Nearest Labels**
	delfini, orca, dolphin, mar, delfin, dauphin, whale, cancun, killer whale, sea world		barrack obama, barack obama, barack hussein obama, barack obama, james marsden, jay z, obama, nelly, falco, barack
	eiffel tower, statue, eiffel, mole antoneliana, la tour eiffel, londra, cctv tower, big ben, calatrava, tokyo tower		ipod, ipod nano, nokia, i pod, nintendo ds, nintendo, lg, pc, nokia 7610, vino

Source: Adapted from Weston et al., 2010.

pared to naive linear time approaches. Instead, we apply the following two steps: (a) learning a label tree and (b) learning predictors for each of the nodes of the tree.

Learning the Label Tree Structure

In order to learn a label tree such as the one in Figure 2, we proceed as follows. Given a set of labels in a node, we look for a partition of that set into subsets such that, inside a subset, labels are difficult to distinguish with classifiers trained

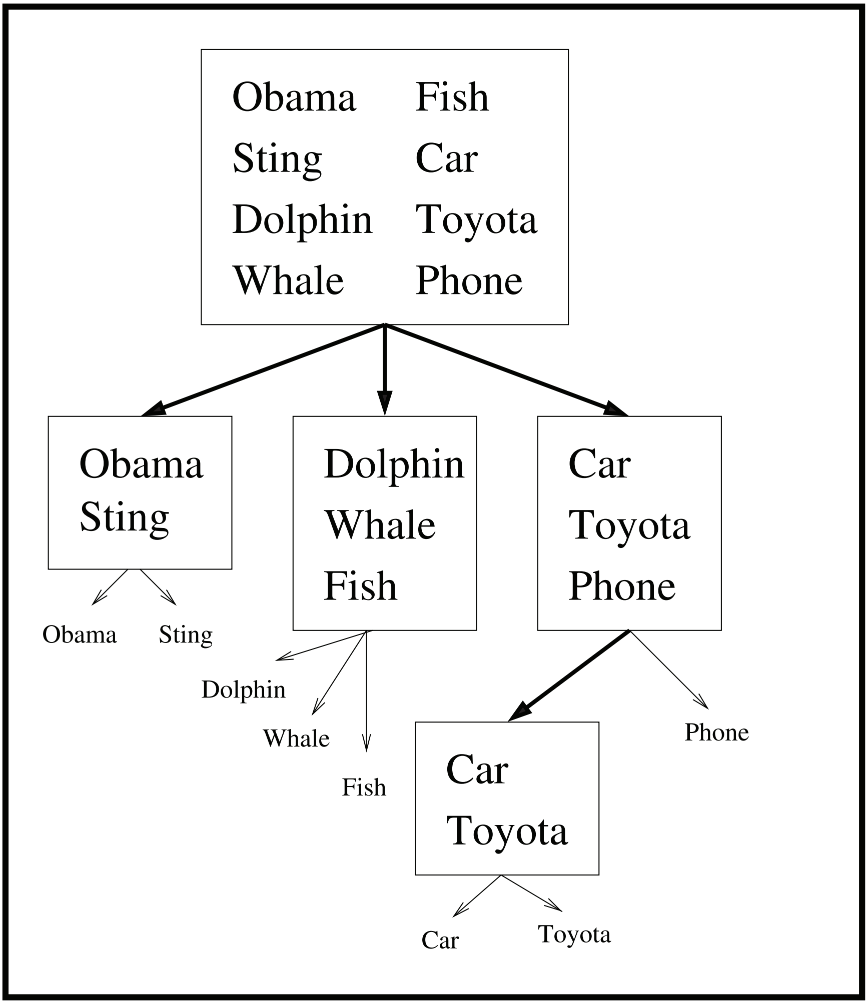


FIGURE 2 Example of a label tree.

on their corresponding images, whereas it is easier to distinguish images belonging to labels of a subset from images belonging to labels of another subset. We do so by computing the confusion matrix between all labels, where we count the number of times our classifiers confuse class *i* with class *j*, and use this matrix to apply spectral clustering (Ng et al., 2002). This procedure can then be applied recursively to obtain a complete label tree. Table 3 gives an example of labels that were clustered together thanks to that technique.

TABLE 3 Examples of obtained clusters of labels.

great white sharks, imagenes de delfines, liopleurodon meduse, mermaid tail, monstre du loch ness, monstruo del lago ness, oarfish, oceans, sea otter, shark attacks, sperm whale, tauchen, whales	apple iphone 3gs, apple ipod, apple tablet, bumper, iphone 4, htc diamond, htc hd, htc magic, htc touch pro 2, iphone 2g, iphone 3, iphone 5g, iphone app, iphone apple, iphone apps, iphone nano	chevy colorado, custom trucks, dodge ram, f 250, ford excursion, ford f 150, mini truck, nissan frontier, offroad, pickup, toyota tundra

Learning a Label Embedding Tree

Once labels are organized into a tree, one can retrain jointly an embedding model (using the algorithm described in the section Joint Embedding of Images and Labels) where each image can now be labeled either with its original labels, or with any of the nodes of the tree that contains them. Moreover, whenever an internal node is selected as a positive label for a given image during training, we select a competing negative label as a sibling node in the label tree, as this corresponds to how the tree would then be used at test time.

The result provides a structure of labels based on both semantic and visual similarities. Furthermore, the performance of a label embedding tree is not only faster at test time, it is also better on average, as can be seen in the literature (Bengio et al., 2010).

REFERENCES

Bengio, S., J. Weston, and D. Grangier. 2010. Label embedding trees for large multi-class tasks. Advances in Neural Information Processing Systems 23: Proceedings of the 24th Annual Conference on Neural Information Processing Systems, Vancouver Canada, December 6–9, 2010.

Beygelzimer, A., J. Langford, and P. Ravikumar. 2009a. Error-correcting tournaments. Proceedings of the International Conference on Algorithmic Learning Theory, Porto, Portugal, October 3–5, 2009, pp. 247–262.

Beygelzimer, A., J. Langford, Y. Lifshits, G. Sorkin, and A. Strehl. 2009b. Conditional probability tree estimation analysis and algorithm. Proceedings of the Conference in Uncertainty in Artificial Intelligence, Montreal, Canada, June 18–21, 2009.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, Fla., June 20–25. 2009.

Hsu, D., S. Kakade, J. Langford, and T. Zhang. 2009. Multi-label prediction via compressed sensing. Advances in Neural Information Processing Systems 22: Proceedings of the 23rd Annual Conference on Neural Information Processing Systems, Vancouver, Canada, December 7–10, 2009.

Ng, A. Y., M. I. Jordan, and Y. Weiss. 2002. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 15: Proceedings of the 16th Annual Conference on Neural Information Processing Systems, Vancouver Canada, December 12–15, 2002.

Torralba, A., R. Fergus, and W. T. Freeman. 2008. 80 million tiny images: A large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11):1958–1970.

Weston, J., S. Bengio, and N. Usunier. 2010. Large scale image annotation: Learning to rank with joint word-image embeddings. Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Barcelona, Spain, September 20–24, 2010.

Searching for Statistical Diagrams

Shirley Zhe Chen, Michael J. Cafarella, and Eytan Adar
University of Michigan

INTRODUCTION

Statistical, or data-driven, diagrams are an important method for communicating complex information. For many technical documents, the diagrams may be readers' only access to the raw data underlying the documents' conclusions.

Unfortunately, finding diagrams online is very difficult using current search systems. Standard text-based search will only retrieve the diagrams' enclosing documents. Web image search engines may retrieve some diagrams, but they generally work by examining textual content that surrounds images, thus missing out on many important signals of diagram content (Bhatia et al., 2010; Carberry et al., 2006). Even the text that is present in diagrams has meaning that is hugely dependent on their geometric positioning within the diagram's frame; a number in the caption means something quite different from the same number in the x-axis scale (Bertin, 1983).

There has been growing commercial interest in making data-driven diagrams more accessible, with data search systems such as SpringerImages (*http://www.springerimages.com/*) and Zanran (*http://www.zanran.com/q/*). While there is a huge amount of research literature on search and image-related topics, diagram search per se is largely unexplored.

In this paper we propose a Web search engine exclusively for data-driven diagrams. As with other Web search engines, our system allows the user to enter keywords into a text box in order to obtain a relevance-ranked list of objects. Our system addresses several challenges that are common among different search engines but that require solutions specifically tailored for data-driven diagrams.

Diagram Corpus Extraction

Obtaining the text of a Web document is usually as easy as downloading and parsing an HTML file; in contrast, statistical diagrams require special processing to extract useful information. They are embedded in PDFs with little to distinguish them from surrounding text, the text embedded in a diagram is highly stylized with meaning that is very sensitive to the text's precise role, and, because diagrams are often an integral part of a highly engineered document, they can have extensive "implicit hyperlinks" in the form of figure references from the body of the surrounding text. Our *Diagram Extractor* component attempts to recover all of the relevant text for a diagram and determine an appropriate semantic label (caption, y-axis label, etc.) for each string.

Ranking Quality

All search engines must figure out how to score an object's relevance to a search query, but scoring diagrams for relevance can yield strange and surprising results. We use the metadata extracted from the previous step to obtain search quality that is substantially better than naive methods.

Snippet Generation

Small summaries of the searched-for content, usually called snippets, allow users to quickly scan a large number of results before actually selecting one. Conventional search engines select regions of text from the original documents, while image search engines generally scale down the original image to a small thumbnail. Neither technique can be directly applied to data-driven diagrams. Obviously, textual techniques will not capture any visual elements. Figure 1 shows that image scaling is also ineffective: although photos and images remain legible at smaller sizes, diagrams quickly become difficult to understand.

This paper describes DiagramFlyer, a search engine for finding data-driven diagrams in Web documents. It addresses each of the above challenges, yielding a search engine that successfully extracts diagram metadata in order to provide both higher-quality ranking and improved diagram "snippets" for fast search result scanning.

The techniques we propose are general and can work across diagrams found throughout the Web. However, in our current testbed we concentrate on diagrams extracted from PDFs that were discovered and downloaded from public Web pages on academic Internet domains. Our resulting corpus contains 153,000 PDFs and 319,000 diagrams. We show that DiagramFlyer obtains a 52% improvement in

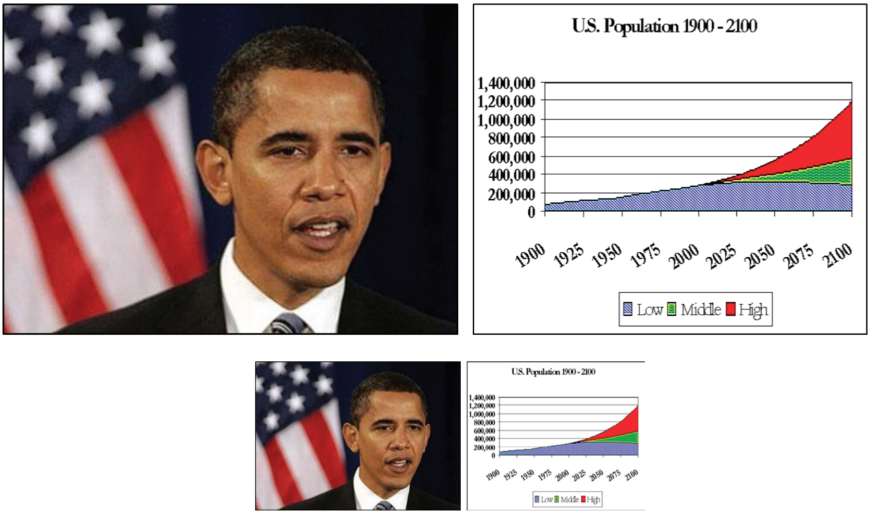


FIGURE 1 Scaling down images works well when generating visual snippets for photos, but diagrams can quickly become illegible.

search quality over naive approaches. Furthermore, we show that DiagramFlyer's hybrid snippet generator allows users to find results 33% more accurately than with a standard image-driven snippet. We also place DiagramFlyer's intellectual contributions in a growing body of work on *domain-independent information extraction*—techniques that enable retrieval of structured data items from unstructured documents, even when the number of topics (or domains) is unbounded.

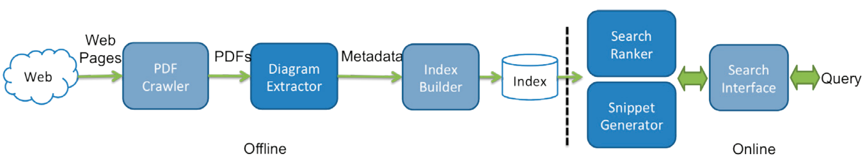


FIGURE 2 The data processing pipeline. An offline component crawls the Web for PDFs, extracts the statistical graphics, and constructs an inverted text index over the resulting extracted metadata. This index and the diagrams are then fed to an online system that ranks diagrams according to users' queries and generates query-appropriate search snippets. The dark blue boxes signify research components described in this paper.

SYSTEM OVERVIEW

As with a traditional Web search engine, DiagramFlyer employs a pipeline of offline corpus-processing steps that produce output then used by an online search query system. The system architecture is seen in Figure 2.

The offline pipeline has three components:

1. The *PDF Crawler* attempts to download a large number of Web-hosted scientific papers for diagram search.
2. The *Diagram Extractor* receives the resulting stream of nearly 153,000 papers. This extractor attempts to identify all the diagrams in the corpus and then annotate the text in each diagram with an appropriate semantic role. As seen in Figure 3, the Diagram Extractor identifies eight roles within the diagram (legend, caption, title, etc.). It also looks for any surrounding text that mentions the figure, labeling the relevant sentences as

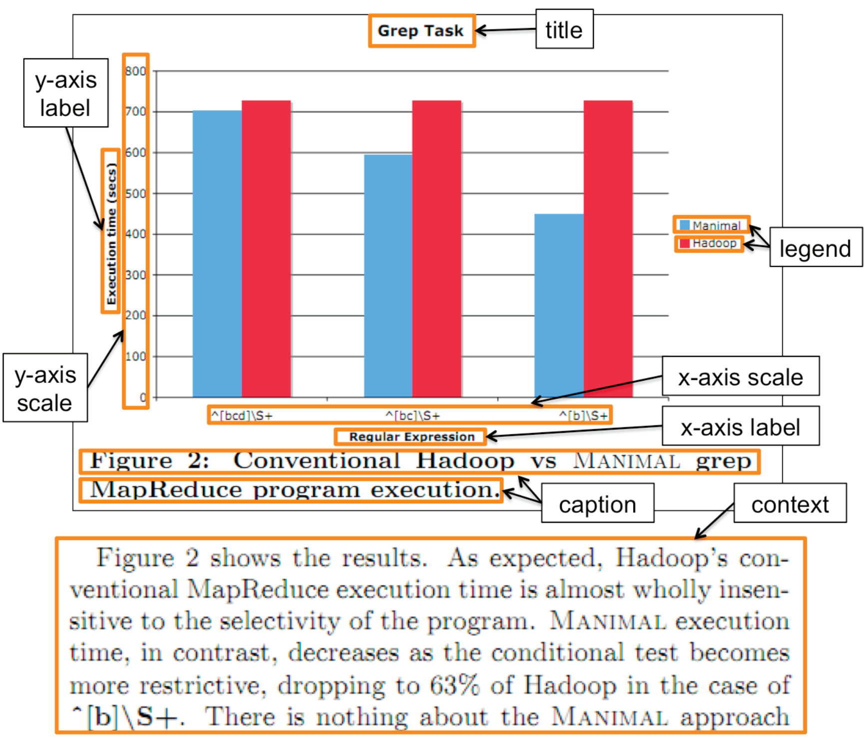


FIGURE 3 Diagram metadata labels for a sample diagram. Some labels, such as title and legend, are found in different places for different graphics.

"context." For the testbed system, we used two-dimensional data-driven plots (including scatterplots, time series, and bar plots).

3. The *Index Builder* constructs a search index over the extracted and annotated diagrams. The index tracks each extracted field separately so that keyword matches on individual parts of the diagram can be weighted differently during ranking.

As seen in Figure 4, DiagramFlyer's online search system is similar in appearance to traditional Web search engines. Answering an online query requires two additional components:

1. The *Search Ranker* assesses the relevance of each diagram in its index. Our system's main advantage over a standard search ranker is its access to the textual features generated by the Diagram Extractor.
2. Finally, the *Snippet Generator* generates a brief summary of each search hit, ordering them according to the Search Ranker. DiagramFlyer's Snippet Generator creates special diagram-specific snippets that contains both graphical and textual elements.

ALGORITHMS

As mentioned above, our system has three main novel components. Because of space limitations, this version of the paper only discusses the Diagram Extractor component.

The Diagram Extractor uses a PDF extractor to obtain all text strings from the document. It then employs a four-stage processing sequence to recover groups of labeled text strings that correspond to real data-driven diagrams:

1. A trained text-centric classifier gives strings an initial label based strictly on textual features, such as the number of words in a string, whether a string is capitalized, distribution of parts of speech, and so on.
2. We then group labeled strings together into geometrically neighboring sets that loosely correspond to diagrams. Sets without critical labels, such as relevant x- and y-axis data, are thrown away. This filters out a huge number of strings that are not relevant to any diagram.
3. We then recompute labels for each string, using the initial labels to compute a series of position-sensitive features. For example, one important feature is a text string's distance to the nearest x-axis scale. This round of classification substantially improves label precision and recall.
4. Finally, we group the resulting labeled strings into sets that represent the final diagram estimates. This step relies heavily on the semantic label applied above; for example, a caption string should always be part of the lower portion of a diagram.

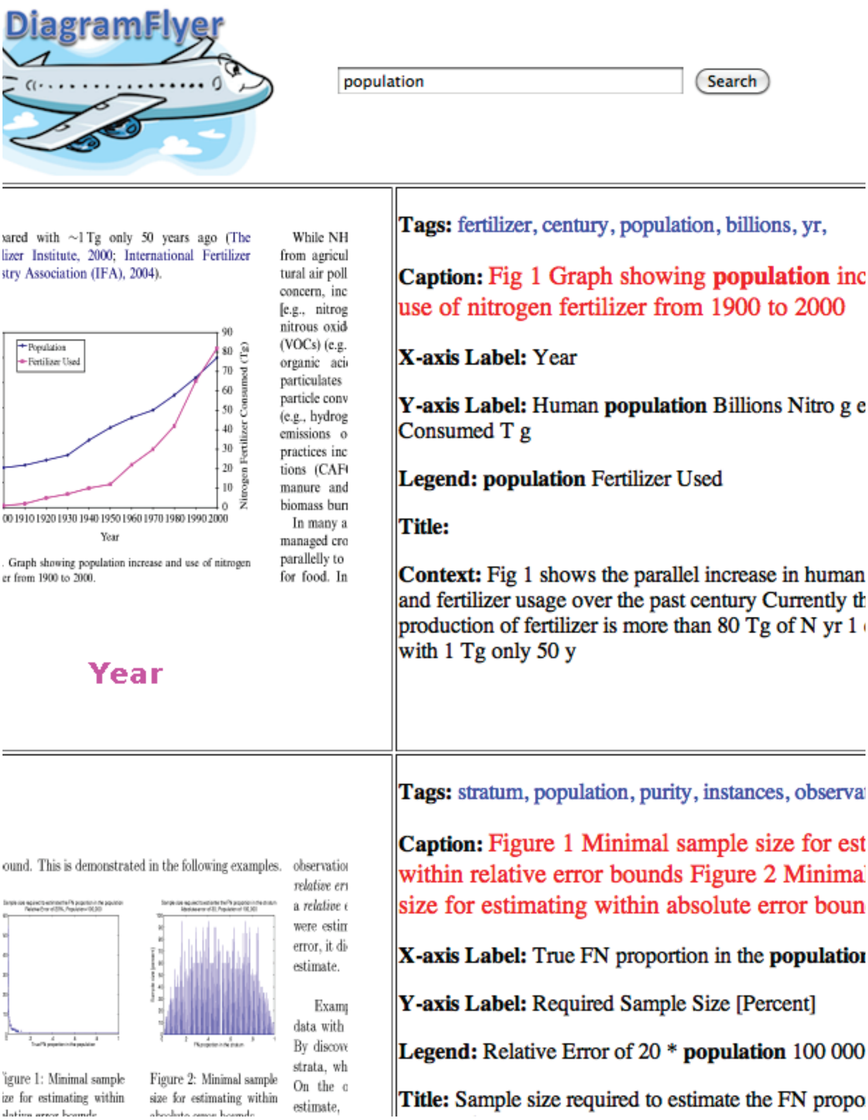


FIGURE 4 A screenshot of the DiagramFlyer search system.

This output is then fed to the Search Ranker and Snippet Generator components.

EXPERIMENTAL RESULTS

The Diagram Extractor is a query-independent component, and so it can be evaluated strictly using our downloaded corpus of scientific papers. We started with 4.7 billion URLs from the English segment of the ClueWeb09 data set (*http://lemurproject.org/clueweb09.php/*). Of these, we retained those pointing to PDF documents. To target PDFs that are more likely to contain diagrams, we further restricted the crawl to the .edu domain. A query workload is critical for evaluating our Search Ranker and Snippet Generator components, but we do not discuss them in this abbreviated paper.

To determine the best implementation for the Diagram Extractor, we evaluated three slightly different versions:

- *text-only*: Just the simple textual classifier.
- *all-classifiers*: The textual classifier and the position-sensitive classifier, without filtering.
- *full*: All steps.

We trained these classifiers using all of the text segments derived from more than 260 data-driven diagrams that were randomly chosen from the PDF corpus; the segments were labeled by hand. We tested the results using another set of 180 similarly generated and labeled diagrams. The evaluation results are shown in the following table (the best scores for any task are shown in bold).

	Recall			Precision		
	text-only	*all*	*full*	*text-only*	*all*	*full*
Title	0.256	0.651	**0.674**	0.344	0.609	**0.617**
y-scale	0.782	**0.796**	0.754	0.889	0.843	**0.900**
y-label	0.835	0.864	**0.874**	0.775	0.752	**0.797**
x-scale	**0.903**	0.835	0.835	0.616	**0.915**	0.896
x-label	0.241	**0.681**	**0.681**	0.340	**0.842**	0.835
Legend	0.520	0.623	**0.656**	0.349	0.615	**0.631**
caption	0.952	**0.887**	0.839	0.450	0.887	**0.929**
nondiagram	0.768	**0.924**	0.313	0.850	**0.909**	0.838

The precision gain of *full* over *all-classifiers* is due to the diagram group filter. On a set of 449 candidate diagram groups, this filter removed 165 bad ones and just 11 good ones. For most labels, this filter does not influence recall much. However, it dramatically reduces recall of non-diagram text in the full case, from 0.9239 to 0.3126. In the case of non-diagram, a reduction in recall is actually a good sign: Most of the "bad candidates" arise from diagrams that are pictorial or otherwise not data driven and would not make sense in the downstream search engine. Reducing recall for this label means that strings that are unnecessary for any diagram are being removed from the output and possibly downstream diagram detection. Although *all-classifiers* has a comparable overall performance, we chose *full* in DiagramFlyer to emphasize precision over recall.

It is clear that *title* and *legend* are the metadata items that are most difficult to classify. In some ways, the result is not surprising: *title* is not always presented, and *legend* can appear in several different locations. Finally, we also evaluated our method for diagram regrouping. We successfully reconstructed 89 diagrams out of a potential 119, with just 20 incorrect ones. These incorrect outputs arose from splitting a single diagram or joining two distinct diagrams.

RELATED WORK

There is a vast literature on text search, snippet generation, image search, and image processing; much of it is not relevant to the unusual demands of searching statistical diagrams. There has been some work in specialized diagram understanding, for example, in processing telephone system diagrams (Arias et al., 1995), but this work is extremely tailored to a narrow diagram type and is not suitable for a general search application.

Only a few pieces of work process diagrams in ways suitable for Web-style search. Huang et al. (2003) proposed an automatic mechanism to recover actual numerical quantities from diagrams' graphical components; it may be usable at large scale. Huang et al. (2005) attempted to label regions of chart text, similar to the Diagram Extraction phase, albeit with fewer labels and somewhat lower accuracy, and it is unclear whether their technique can handle multidiagram images. The most relevant is work from Kataria et al. (2008) and Lu et al. (2009). They extract information from paper-embedded diagrams, recovering both text labels and graphical elements; their text recovery component focuses on recovering OCR text, with some amount of label recovery as a side effect of the technique. Their system uses some of the same features as DiagramFlyer's Diagram Extractor, though it is not clear how much their technique can be extended to yield more fine-grained labels, and they do not focus on any tasks downstream from the extraction stage.

CONCLUSION

We have shown that domain-independent diagram extraction is possible. In the full presentation we also present evidence that shows how this system enables higher-quality search relevance and snippet generation than is possible using standard techniques.

REFERENCES

Arias, J. F., C.P. Lai, S. Surya, R. Kasturi, and A.K. Chhabra. 1995. Interpretation of telephone system manhole drawings. Pattern Recognition Letters 16(4):355–369.

Bertin, J. 1983. Semiology of Graphics: Diagrams, Networks, Maps. Madison, Wisc.: University of Wisconsin Press.

Bhatia, S., P. Mitra, and C. L. Giles. 2010. Finding algorithms in scientific articles. Proceedings of the 19th International World Wide Web Conference, Raleigh, N.C., April 26–30, 2010.

Carberry, S., S. Elzer, and S. Demir. 2006. Information graphics: An untapped resource for digital libraries. Proceedings of the 29th Annual International ACM SIGIR Conference, Seattle, Wash., August 6–11, 2006.

Huang, W., C. Tan, and W. Loew. 2003. Model-based chart image recognition. Fifth International Workshop on Graphics Recognition (GREC), Barcelona, Spain, July 30–31, 2003.

Huang, W., C. L. Tan, and W. K. Leow. 2005. Associating text and graphics for scientific chart understanding. Eighth International Conference on Document Analysis and Recognition, Seoul, Korea, August 29–September 1, 2005.

Kataria, S., W. Browuer, P. Mitra, and C. L. Giles. 2008. Automatic extraction of data points and text blocks from 2-dimensional plots in digital documents. Proceedings of the 23rd AAAI Conference on Artificial Intelligence, Chicago, Ill., July 13–17, 2008.

Lu, X., S. Kataria, W. J. Brouwer, J. Z. Wang, P. Mitra, and C. L. Giles. 2009. Automated analysis of images in documents for intelligent document search. International Journal on Document Analysis and Recognition 12(2):65–81.

Engineering Sustainable Buildings

Introduction

Annie Pearce
Virginia Tech

John Zhai
University of Colorado Boulder

Buildings account for 40 percent of the primary energy usage and 70 percent of all the electricity consumption in the United States.[1] Construction, operation, and demolition of buildings generate tremendous pollution that directly and indirectly causes urban air quality problems and climate change. Poor design of buildings and systems not only wastes resources and energy and causes adverse impacts on the environment but also creates uncomfortable and unhealthy indoor environments. In addition, as the impact of humans on the environment at both local and global scales becomes increasingly apparent, sustainable development of buildings has emerged as an important goal throughout the entire life span of a building project.

Sustainability implies the ability of a system to maintain itself or be maintained over time without threatening the stability of other systems upon which it depends. However, just like their ecological counterparts, complex human-designed systems such as buildings sometimes exhibit emergent behaviors that make their sustainability difficult to model and evaluate, and thus design and optimize, especially in situations where the performance of those systems depends on dynamic interactions among nature, humans, and systems. Modern design concepts of high-performance buildings, associated with the usage of new building materials and advanced mechanical and electrical systems, result in an increased need for understanding the integration of building elements and systems, including humans who design and operate them. To reach the net-zero-energy building goal by 2030 will require highly multidisciplinary efforts from many collaborators such as policy makers, architects, urban planners, material scientists, civil engi-

[1] U.S. Energy Information Administration, 2010.

neers, mechanical engineers, environmental engineers, those in the construction industry, social scientists, and even public health practitioners throughout the long life cycle of buildings.

This session introduces the emerging integration and transformation effort of the architecture/engineering/construction industry to increase social, economic, and environmental benefits via sustainable building development. John Ochsendorf (Massachusetts Institute of Technology [MIT]) introduces the current challenges and opportunities for low-carbon buildings by presenting the cutting edge in benchmarking building performance and building life-cycle cost assessment. Using case studies of ultra-low-carbon buildings designed by his team at MIT, he discusses the best integrated design strategies and future research and industry needs.

Next, John Haymaker (Design Process Innovation) uses industry case studies and surveys to summarize the difficulties that building design teams have defining and searching through solution spaces and how this results in unsustainable designs. He presents an emerging platform of industrial and academic tools that are helping professional and student teams execute far more efficient and effective design processes.

Jelena Srebric (Pennsylvania State University) discusses the challenges in modeling the energy and environmental performance of an entire building. She defines key questions that multiscale modeling can address for an engineer facing a design of new or renovation of old buildings with sustainability in mind. She also analyzes the strengths and weaknesses of existing multiscale modeling opportunities and concludes with a discussion on future needs in developing new building multiscale models.

Chris Pyke (U.S. Green Building Council) wraps up the session with an industry perspective that covers the use of location-based services and social networks to drive market transformation for sustainable building. He demonstrates an innovative Geographic Information System-based platform for conducting dynamic, multicriteria benchmarking and facilitating the collection and analysis of unprecedented information about the experience of occupants in and around green buildings. He discusses how these new tools will drive continuous performance in a number of specific areas, including greenhouse gas emissions reduction, water conservation, and public health. It is evident that the ability to identify, compare, and reward high-achieving projects and individuals is central to green building's success.

Challenges and Opportunities for Low-Carbon Buildings

John Ochsendorf
Massachusetts Institute of Technology

ABSTRACT

Buildings are responsible for higher greenhouse gas emissions than any other sector of society. Furthermore, the portion of emissions due to building construction and operation has been increasing in recent decades. These emissions are primarily due to heating, cooling, and lighting, though the embodied emissions in materials are also significant. Major new initiatives are under way to reduce the energy consumed by buildings, but there are numerous technical, economic, and policy barriers. This paper summarizes some of the key challenges for the design and implementation of future low-carbon buildings in the United States and identifies opportunities for the engineering profession. More efficient building design represents one of the most cost-effective opportunities for large-scale carbon reductions on a national and global scale. A greater emphasis on integrated building design for the full life cycle can lead to dramatically improved building performance. Finally, a series of recent projects demonstrates successful life-cycle design for low-carbon buildings.

INTRODUCTION

This paper outlines some of the key challenges facing the engineering of low-carbon buildings. Though buildings have major environmental impacts due to water use, raw material consumption, and other natural resource depletion, it is useful to focus on greenhouse gas emissions due to the central role of buildings in mitigating climate change. In particular, carbon dioxide equivalent (CO_2e) provides a simple metric for building environmental performance. As the

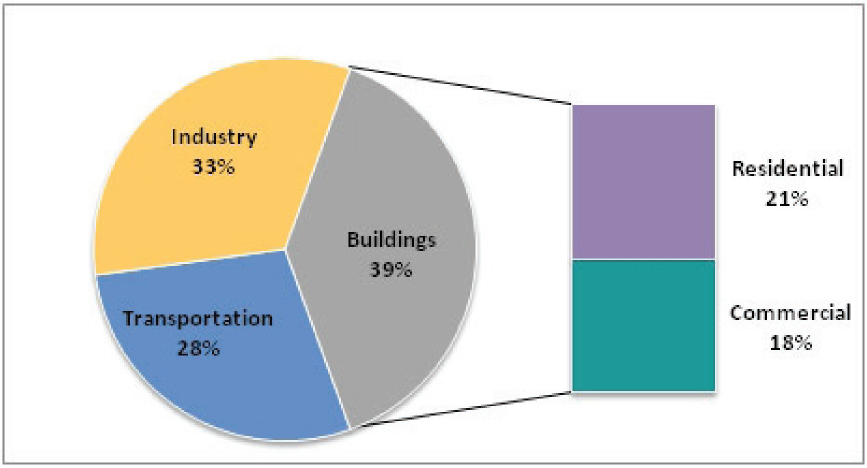


FIGURE 1 U.S. energy consumption by sector. Source: U.S. DOE, 2010.

largest source of carbon emissions in the United States, buildings also represent a significant opportunity (Figure 1). The Intergovernmental Panel on Climate Change (IPCC) has identified buildings as the sector offering the greatest potential for carbon reductions (IPCC, 2007). Similarly, the World Business Council for Sustainable Development (WBCSD) has demonstrated that global energy use of buildings could be reduced 60% by 2050 using existing technologies (WBCSD, 2009). McKinsey Consulting has identified the building sector as the most cost-effective strategy for carbon abatement, with most building improvements providing carbon reductions at a negative cost (McKinsey Consulting, 2007). In short, there are a range of strategies in the building sector to save both money and carbon emissions. Clearly engineers have a major role to play in helping to reduce the carbon emissions of existing and new buildings, though a number of challenges exist.

Several current initiatives in the United States provide targets for improved carbon performance of buildings. The 2030 Challenge establishes targets for carbon reductions of new buildings, with increasing standards in the coming two decades (Architecture2030, 2011). The current aim is to design buildings to use 60% less energy than average for the building type. The reduction target increases by 10% every five years (e.g., lowering to 70% reductions by 2015), until carbon-neutral buildings are the target in the year 2030. These design goals are to be met through three primary approaches: (1) improved design strategies, (2) more efficient technologies and systems, and (3) off-site renewables (up to a maximum of 20%). While the 2030 Challenge is voluntary at present, many leading engineering firms have committed to pursuing its goals, and a number of architecture firms are now tracking the energy consumption of their new buildings in relation to the 2030 Challenge goals. Other countries in the world have binding legal requirements to achieve dramatic carbon reductions in new buildings in the coming decade, though the closest legal target in the United States is California's Assembly Bill 32, which mandates reductions to 1990 emissions levels by 2020 (Assembly Bill 32, 2011). Such policies, as well as more stringent building codes, can serve as important motivators for engineers and architects.

CHALLENGES FOR LOW-CARBON BUILDINGS

While engineers play a crucial role in the design of more sustainable buildings, they often get involved too late in the design process. Many of the key decisions are taken in the earliest design stages, when building orientation, glazing ratio (ratio of area of glass to area of opaque wall), and the overall form of the building are decided. Once these critical decisions have been made, engineers can attempt to optimize a poor design, but it is difficult to arrive at low-carbon design without having engineers involved in the initial design. The challenge is to integrate engineering analysis in a manner that provides rapid feedback to architects and the rest of the design team early in the process. This requires engineers who are trained as designers and can pose multiple solutions to open-ended problems. In short, the design of high-performance buildings requires integrated, systems thinking from the earliest conceptual design stage.

There is a need to train more engineers in the fields of building science and sustainable design. In particular, most engineering schools do not directly address the design and operation of buildings, since the field of sustainable building design falls among the disciplines of mechanical engineering, civil engineering, and architecture. Though there are a small number of programs in architectural engineering, the number of architectural engineering graduates nationally is much lower than traditional engineering disciplines, and the coming decades will require more engineers in this field. Despite the dramatic economic and environmental impact of buildings, the engineering of sustainable buildings is missing from most schools of engineering in the United States.

Finally, there is an acute lack of spending on research and development (R&D) for sustainable buildings. Compared to other sectors such as the automobile or electronics industries, the construction industry does not spend nearly as much on R&D. It has been estimated that only 0.25% of gross sales in construction are spent on R&D in the United States (Gould and Lemer, 1994). Furthermore, only 0.2% of federal research funding is spent on topics related to green buildings (USGBC, 2008). The scarcity of research funding means that fewer researchers are working on the vital topic of sustainable building technology, despite the pressing urgency of climate change and the favorable economics for carbon reductions through improved buildings.

OPPORTUNITIES FOR LOW-CARBON BUILDINGS

The key to improved building performance is integrated design, which incorporates systems thinking at the conceptual design stage. Decisions about building orientation, façade systems, heating and cooling strategies, and glazing ratios play a crucial role in the final energy performance of the building. Building design must be climate specific and should be tailored to the regional climate. The choice of materials plays a central role in the embodied carbon of any building, and it also has implications for the carbon emissions due to the operation of buildings, through thermal mass. However, most traditional architectural design software does not include the capacity to analyze energy or environmental performance. In the past decade, several tools have been developed to aid architects and engineers in the early design stage. One example is the MIT Design Advisor, which allows for the rapid estimation of building energy use based on massing, orientation, climate, glazing ratios, and more (MIT, 2011). More recently, the DIVA platform has been developed to allow architects to run basic performance simulations in the early design stage within existing architectural design software (*www.diva-for-rhino.com*). In order to influence the initial design phase, these programs provide rapid feedback to architects. Though such programs are developing quickly, there is still an urgent need for improved design software as well as engineers capable of systems-level building design.

A multidisciplinary education focusing on sustainable design can attract a new generation of engineers to the profession. In addition to engineering expertise in heat transfer, thermal science, materials engineering, and other traditional building science disciplines, sustainable buildings require expertise in design thinking and creative problem solving. Furthermore, there is demand in the market for increased literacy in rigorous sustainability metrics and for expertise in life-cycle environmental and economic performance of buildings. Integrated graduate education in the built environment, such as the Solving Urbanization Challenges by Design program at Columbia, which combines engineering, architecture, and planning, can help to provide the broad education needed for sustainable building designers and researchers (IGERT, 2011).

To increase R&D funding for sustainable buildings, new partnerships are required between industry, academia, and government. The Concrete Sustainability Hub at MIT is one example of such a partnership, which is advancing our knowledge of the environmental performance of buildings through rigorous life-cycle assessment of the carbon emissions of buildings. From federal sources, the National Science Foundation (NSF) is a particularly appropriate venue for greater research funding in sustainable buildings. A number of NSF Graduate Research

Fellowships could be established for students working in building science, which would help to attract more researchers to the field.

CASE STUDIES

Many significant low-carbon buildings have been constructed in recent years, which can serve as inspiration for engineers and architects. Some noteworthy examples for a range of different building types include

- Residential building: BedZED, Beddington, England
- School: Richardsville Elementary School, Richardsville, Kentucky
- Office building: National Renewable Energy Laboratory, Golden, Colorado
- Cultural building: Mapungubwe Visitor's Center, South Africa

In these and most other cases of sustainable buildings, the crucial design strategies were developed by a collaborative team of architects and engineers early in the design process. By setting clear targets for energy consumption, these designs were achieved at similar costs as conventional construction, with significant life-cycle economic benefits due to reduced operating energy requirements. These and other projects demonstrate that the technology for carbon-neutral buildings exists today.

CONCLUSIONS

Buildings are widely recognized as the largest opportunity for cost-effective carbon reductions in the United States and around the world. While engineers have a vital role to play in transforming the built environment into a low-carbon future, there are numerous areas for improvement. In particular, a new emphasis on conceptual design, life-cycle thinking, and innovative research partnerships can help to advance the field, reduce carbon emissions, and train a new generation of engineers.

REFERENCES

Architecture2030. 2011. The 2030 Challenge. Available at *http://www.architecture2030.org/2030_challenge/the_2030_challenge*.

Assembly Bill 32. 2011. California Global Warming Solutions Act. Available at *www.arb.ca.gov/cc/ab32/ab32.htm*.

Gould, J.P., and A.C. Lemer. 1994. Toward Infrastructure Improvement: An Agenda for Research. Washington, DC: National Academy Press.

IGERT (Integrative Graduate Education and Research Traineeship). 2011. Columbia University IGERT Program, Solving Urbanization Challenges by Design. Available at *http://www.civil.columbia.edu/igert/index.html*.

IPCC (Intergovernmental Panel on Climate Change). 2007. IPCC Fourth Assessment Report: Climate Change 2007 (AR4). Cambridge, U.K.: Cambridge University Press. Available at *http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr.pdf*.

MIT (Massachusetts Institute of Technology). 2011. Design Advisor, Building Technology Program. Available at *http://designadvisor.mit.edu*.

McKinsey Consulting. 2007. Reducing U.S. Greenhouse Gas Emissions: How Much at What Cost? U.S. Greenhouse Gas Abatement Mapping Initiative. Available at *http://www.mckinsey.com/en/Client_Service/Sustainability/Latest_thinking/~/media/McKinsey/dotcom/client_service/Sustainability/PDFs/Reducing%20US%20Greenhouse%20Gas%20Emissions/US_ghg_final_report.ashx*.

U.S. DOE (Department of Energy). 2010. Buildings Energy Data Book. Available at *http://buildingsdatabook.eren.doe.gov/*.

USGBC (U.S. Green Building Council). 2008. A National Green Building Research Agenda, United States Green Building Council (USGBC) Research Committee, April 2007, Revised February 2008. Available at *http://www.usgbc.org/ShowFile.aspx?DocumentID=3402*.

WBCSD (World Business Council for Sustainable Development) 2009. Energy Efficiency in Buildings: Transforming the Market. Available at *http://www.wbcsd.org/DocRoot/Ge2Laeua8uu2rkodeu7q/91719_EEBReport_WEB.pdf*.

Expanding Design Spaces

John Haymaker
Design Process Innovation

ABSTRACT

Current sustainability design practice is struggling to create design spaces that leverage advances in building technology, computer science, and human collaboration. But the methods for and interest in creating such spaces exists. This paper describes an evolving platform of tools to improve the design spaces that will enable a more sustainable built future.

INTRODUCTION

Design research in the twentieth century oscillated between a view of design as a discipline and as science (Cross, 2001). In the former, design synthesis is an innately human-centric process aided by technology for documentation and analysis. In the latter, technology is at the center of both synthesis and analysis—design is a technical process from which the best designs emerge. In both cases, design is broadly seen as a process in which stakeholders and computer tools construct interrelated design spaces of objectives, alternatives, impacts, and value in search of the best design (Clevenger and Haymaker, 2011a).

The previous century also saw a dramatic increase in awareness about sustainability—the ability of a designed system to be maintained over time without threatening the stability of other systems upon which it depends (Zhai and Pearce, 2011). In recent years, scientific and practical dialogue has actively sought design methods to enable the integrated sustainability of social, environmental, and economic systems at multiple scales and dimensions. In response, significant progress and convergence in design methods for sustainability are being made. New organizational and contractual arrangements and physical and virtual collaboration spaces make it feasible to assemble large, multidisciplinary teams. New rating systems, guidelines, and codes help define objectives. Building information models and parametric modeling help generate alternatives. Automated analysis tools can analyze energy, structure, schedule, cost, and many other performance objectives. Designers have at their disposal a rich and growing set of human and technological processes for constructing design spaces and understanding and maximizing multistakeholder value.

These developments come hopefully just in time. In the next decades, project teams will design unprecedented amounts of new and refurbished buildings and infrastructure. As resources become increasingly scarce, building technologies more numerous, and the impacts of the built environment more apparent, architecture, engineering, and construction (AEC) professionals will need to make profoundly complex and difficult decisions. To find the best designs, they will need to construct and communicate very large design spaces, which help them fully understand the objectives, exhaustively search the alternatives, systematically do the analysis, and explicitly communicate the value of their projects. Even as government, industry, and academic leaders are calling for professionals to apply these new methods, however, successful adoption is proving difficult. There is still a large gap between the available and emerging methods and the ability of design teams to reliably and successfully and consistently apply them in practice. The traditional precedence-based design processes still taught and in use today construct and communicate small design spaces that leave better designs undiscovered. Project teams lack sociotechnical platforms and methodologies that allow them to relate all of these new methods and information to effectively and efficiently construct and explore larger, better design spaces.

Many research fields are relevant to developing this new platform, including organizational theory and social networking, design theory and methodology, building information and process modeling, model-based analysis, multidisciplinary design optimization, decision science, human-computer interaction, economics, and artificial intelligence. Numerous research methods involving ethnography, theory building, tool building, data analysis, and action research are relevant to understanding where existing methods fail and how to better fit and improve this platform. This presentation discusses one ongoing effort to develop such a platform for collaboratively constructing sustainable building and infrastructure design spaces.

The next section summarizes methods to describe the quality and clarity of design spaces and the efficiency and effectiveness of the processes used to construct them. These definitions are then used to illustrate how design spaces are constructed and communicated today. Subsequently, the paper summarizes work to construct an integrated platform of tools. Tests in the laboratory and in practice illustrate the ways elements of the platform can individually and collectively improve design space quality and clarity considerably over current methods.

Nevertheless, significant opportunity remains to improve the performance of this platform. The presentation concludes with a discussion of future work.

HOW TO MEASURE DESIGN PROCESSES

To improve design processes, it is necessary to model and measure them. Design theory and methodology, process modeling, lean construction, and decision analysis provide the foundation for methods and metrics for describing and measuring design processes.

Design integrates numerous processes, organizations, and products. To help designers and researchers model and understand this complexity, we develop graphical process communication methods to help teams formally describe and relate these components (Haymaker, 2006).

Design requires communication of processes, but it is necessary to know how to efficiently communicate them. The Mock Simulation Design Charette measures and compares the efficiency and effectiveness of process communication methods (Senescu and Haymaker, 2011).

Design involves the exploration of design spaces. The Design Exploration Assessment Methodology (Clevenger et al., 2010) provides metrics and procedures to measure and compare design processes in terms of the challenge addressed, the strategy implemented, and the exploration achieved.

Design is about making and communicating decisions. The Rationale Clarity Framework (Chachere and Haymaker, 2011) helps teams measure how clearly they communicate decision rationale.

These methods for capturing and measuring the efficiency, quality, and clarity of AEC design processes, spaces, and decisions enable the assessment of current practice and provide key insights as to how to improve these processes.

ASSESSMENTS OF CURRENT AEC PROCESSES

My research group employs ethnographic-action research (Hartmann et al., 2009) and laboratory design charettes to understand the performance of projects. We build detailed process models that help us understand the design and analysis tasks that teams perform. We find that designers' tacit knowledge alone is insufficient to guide them for the challenges they face today, and yet they still perform a narrow search of small design spaces (Clevenger and Haymaker, 2011b). We find design teams struggle to communicate process (Senescu et al., 2011a) and design rationale (Haymaker et al., 2011).

We use surveys to test our observations. For example, we found that leading high-rise design firms, consisting mainly of architects, spend many hours generating few options and analyze them principally in terms of architectural and economic criteria (Gane and Haymaker, 2010). A survey of a leading multidisciplinary engineering firm confirms these observations and finds that design teams spend over 50 percent of their time on non-value-adding information management tasks (Flager and Haymaker, 2007).

In summary, we find in practice today that underrepresented teams develop inadequate statements of objectives and analyses and rely on potentially invalid precedent knowledge to perform limited and superficial searches of poorly defined and communicated design spaces.

PLATFORM TO IMPROVE AEC PROCESSES

To address these limitations, we have developed and tested a collaborative platform of tools that assist building design teams to generate, evaluate, and develop consensus around far larger and better formulated design spaces than achieved in practice today.

The Process Integration Platform (PIP) provides highly visual and interactive tools to help teams communicate design processes (Senescu and Haymaker, 2011). PIP helps to communicate processes within a project team to improve collaboration between project teams to improve process sharing and, across a firm or industry, to promote understanding and to drive innovation and process improvement.

Collaborative design generation processes help teams create and manage alternative spaces. Perspectors help teams assemble graphs of geometric transformations that generate and manage dependent geometric designs (Haymaker et al., 2004). Multi Attribute Interaction Design helps teams conceptualize and relate parameters to discover synergistic interactions that help them tunnel through cost barriers and maximize multidisciplinary value (Ehrich and Haymaker, 2012). Design Scenarios helps them transform these initial parameters into parametric geometric design spaces suitable for multidisciplinary analysis (Gane et al, 2011).

Analysis processes help teams understand the impacts of their designs. Members of my research group develop, test, and industrialize advanced analysis and optimization processes. For example, ThermalOpt integrates a parametric Building Information Model (Digital Project) with energy (EnergyPlus) and day lighting (Radiance) simulation engines using an open data model (Industry Foundation Classes) (Welle et al., 2011) to provide an integrated thermal design optimization environment. Biopt is a method for shape and member sizing optimization of steel frame structures (Flager et al., 2011).

Management processes are needed to help teams set up and explore design spaces. For example, the Attribute Management Methodology for Multidisciplinary Optimization gives designers control over how attributes of building objects are transferred and varied in optimization processes (Welle and Haymaker, 2010). The ability to represent and process so much information requires efficient ways to organize and leverage it to make complex multistakeholder decisions. After a design space is constructed, importance analysis reveals key parameters on building performance, helping design teams focus their exploration (Clevenger et al., 2008). Multi-Attribute Collaborative Design Analysis and Decision Integration helps teams integrate objectives, alternatives, analyses, and values to efficiently develop consensus around decisions (Haymaker and Chachere, 2006).

Platforms for bringing multiple stakeholders and their tools together and to orchestrate their interaction are emerging. For example, Filter Mediated Design investigates the feasibility of multiple designers and agents collaborating using various interaction protocols around a central model (Haymaker et al., 2000), whereas Process Integration and Design Optimization enables the wrapping and orchestration of commercial tools in a workflow (Flager et al., 2009). Collaborative spaces, like the iRoom at Stanford and platforms like PIP, help teams configure an entire socio-technical infrastructure to meet project-specific challenges.

We actively test the tools and platform in the classroom, in the laboratory, and in practice. Building and communicating high-quality design spaces can efficiently lead to improved design. Through this applied research, we are investigating how best to synthesize and use these tools and to train designers in the use of them. In an afternoon, it is feasible to work with a large group of diverse stakeholders to explicitly capture and communicate project objectives; investigate several hundred alternatives for energy, structural, daylight, and other implications; document qualitative analyses on a subset of these options; and clearly communicate which alternatives perform best and why, as well as provide instantaneous data about process efficiency and effectiveness and design space clarity and quality. This stands in stark contrast to the undersized and poorly communicated design spaces designers are able to explore today.

CONCLUSION

This presentation summarized one ongoing effort to develop an integrated process platform for efficient and effective construction and communication of design spaces. This, and research in other labs referenced in our papers, represents an emerging and exciting new paradigm in design for sustainable buildings. The two views of design research from the past century—discipline and science—are merging, but there are many cultural and technical challenges ahead. Multidisciplinary engineering needs a common language, a platform of fully realized tools that allow designers to design and broadly communicate processes and results. I see several steps to creating such a platform: incentive structures that enable academic researchers and the software industry to invest the time needed to develop the platform, training, and time for designers to use and perfect the platform, and case studies to validate the need for and further improve the platform. My hope is that the power of such a platform can transcend the current political, economic, and technical barriers that keep us from making sustainable decisions.

REFERENCES

Chachere, J., and J. Haymaker. 2011. Framework for measuring the rationale clarity of AEC design decisions. ASCE Journal of Architectural Engineering 17(3):86–96.

Clevenger, C., and J. Haymaker. 2011a. Metrics to assess design guidance. Design Studies 32(5):431–456, doi:10.1016/j.destud.2011.02.001.

Clevenger, C., and J. Haymaker. 2011b. The Need to Measure the Guidance Afforded by Design Strategies. Center for Integrated Facility Engineering, Technical Report No. 190.

Clevenger, C., J. Haymaker, and S. Swamy. 2008. The importance process: Enabling creativity in performance-based design through systematic, model-based search of multidisciplinary impacts. Proceedings of the 2008 International Scientific Committee World Sustainable Building Conference, Melbourne, Australia. Eds. G. Foliente, T. Luetzkendorf, P Newton, and L. Paevere, ISBN/ISSN: 978-0-646-50372-1.

Clevenger, C., J. Haymaker, and A. Ehrich. 2010. Design Exploration Assessment Methodology: Testing the Guidance of Design Processes. Center for Integrated Facility Engineering, Technical Report No. 192.

Cross, N. 2001. Designerly ways of knowing: Design discipline versus design science. Design Issues 17(3):49–55.

Ehrich, A., and J. Haymaker. 2012. Multiattribute interaction design: An integrated conceptual design process for modeling interactions and maximizing value. Artificial Intelligence for Engineering Design, Analysis and Manufacturing 26(1):85–101.

Flager, F., and J. Haymaker. 2007. A comparison of multidisciplinary design, analysis and optimization processes in the building construction and aerospace industries. Proceedings of the 24th International Conference on Information Technology in Construction, June 26–29, 2007, Maribor, Slovenia, edited by I. Smith. pp. 625–630.

Flager, F., B. Welle, P. Bansal, G. Soremekun, and J. Haymaker. 2009. Multidisciplinary process integration and design optimization of a classroom building. ITcon 14:595–612.

Flager, F., K. Shea, M. Fischer, and J. Haymaker. 2011. BiOPT: A Method for Shape and Member Sizing Optimization of Steel Frame Structures. Center for Integrated Facility Engineering, Technical Report No. 203.

Gane, V., and J. Haymaker. 2010. Benchmarking conceptual high-rise design processes. Journal of Architectural Engineering 16(3):100–111.

Gane, V., J. Haymaker, M. Fischer, and V. Bazjanac. 2011. Application of Design Scenarios Methodology to Evaluate the Effectiveness of Transparent Parametric Design Spaces. Center for Integrated Facility Engineering, Technical Report No. 199.

Hartmann, T., M. Fischer, and J. Haymaker. 2009. Implementing information systems with project teams using ethnographic–action research. Advanced Engineering Informatics 23(1):57–67.

Haymaker, J. 2006. Communicating, integrating, and improving multidisciplinary design narratives. International Conference on Design Computing and Cognition 2006. Amsterdam: Springer. pp. 635–653.

Haymaker, J., and Chachere, J. 2006. Coordinating goals, preferences, options, and analyses for the Stanford Living Laboratory Feasibility Study. Intelligent Computing in Engineering and Architecture 13th EG-ICE Revised Selected Papers, Lecture Notes in Computer Science, Vol. 4200/2006, Ian Smith (ed.), New York: Springer-Verlag.

Haymaker, J., P. Keel, E. Ackermann, and W. Porter. 2000. Filter mediated design: Generating coherence in collaborative design. Design Studies 21(2):205–220.

Haymaker, J., J. Kunz, B. Suter, and M. Fischer. 2004. Perspectors: Composable, reusable reasoning modules to construct an engineering view from other engineering views. Advanced Engineering Informatics 18(1):49–67.

Haymaker, J., Chachere, J., and Senescu, R. 2011. Measuring and improving rationale clarity in a university office building design process. ASCE Journal of Architectural Engineering 17(3):97–111.

Senescu, R., and J. Haymaker. 2011. Communicating Design Processes Effectively and Efficiently. Center for Integrated Facility Engineering, Technical Report No. 198.

Senescu, R., J. Haymaker, and M. Fischer. 2011a. Design Process Communication Methodology: Improving the Efficiency and Effectiveness of Collaboration, Sharing, and Understanding. Center for Integrated Facility Engineering, Technical Report No. 197.

Senescu, R., J. Haymaker, and G. Aranda-Mena. 2011b. Relationships Between Project Complexity and Communication. Center for Integrated Facility Engineering, Technical Report No. 196.

Welle, B., and J. Haymaker. 2010. Attribute Management Methodology for Multidisciplinary Optimization. Center for Integrated Facility Engineering, Technical Report No. 195.

Welle, B., J. Haymaker, and Z. Rogers. (2011. ThermalOpt: A Methodology for BIM-Based Passive Thermal Multidisciplinary Design Optimization. Center for Integrated Facility Engineering, Technical Report No. 200.

Zhai, J., and A. Pearce. 2011. Engineering Sustainable Buildings. National Academy of Engineering Frontiers of Engineering Symposium, Mountain View, Calif., Sept. 19–21, 2011.

Opportunities and Challenges for Multiscale Modeling of Sustainable Buildings

Jelena Srebric
Pennsylvania State University

Existing urban settlements are composed of buildings that use approximately 40 percent of the total primary energy consumption in the United States, and as a result are the major contributors to greenhouse gas emissions based on an International Energy Agency (IEA) report (EIA, 2011). In fact, buildings use more energy than either the transportation or industry sectors. This intense energy demand is projected to increase in the next couple of decades based on IEA projections through the year 2035. As a result, building infrastructure has become an important research area and funding agencies have launched new initiatives such as (1) the Department of Energy's Energy Innovation HUB on Building Energy Efficiency and (2) the National Science Foundation's Emerging Frontiers in Research and Innovation on Science in Energy and Environmental Design in 2010. Furthermore, the issue of energy-efficient and environmentally friendly buildings was also addressed in the National Academy of Engineering's report entitled, *The Grand Challenges for Engineering* in the chapter "Restore and Improve Urban Infrastructure." One technology that can support addressing this grand challenge in engineering is the predictive multiscale modeling of transport processes in and around buildings.

Contemporary approaches to multiscale modeling of buildings in urban settlements are limited to isolated case studies on energy consumption of building systems and resultant projected CO_2 emissions. The full integration of results into a comprehensive understanding of system behavior does not exist or is based on simplified, linear approximations of various system components. Only recently have models and simulation platforms emerged that implement comprehensive modeling of buildings in urban settlements. At present, novel approaches to addressing simulation challenges are derived from two disciplinary domains, each informed by its own respective fields of expertise:

1. Meteorologists and climatologists have introduced constraining anthropogenic sources in prognostic weather and climate models to better understand urban heat islands as well as outdoor air quality and contaminant dispersion in cities.
2. Building scientists have explored weather and climate forcing for airflow, energy, and contaminate simulations in and around buildings to understand building energy consumption, ventilation/infiltration, and contaminant dispersion.

The connection among all of these disparate fields of study is the universal system of transport equations solved numerically for appropriate temporal and spatial scales. The solution of transport equations typically includes mass and momentum equations to solve the airflow field, while the addition of partial differential equations to represent scalars, such as temperature and contaminant concentrations, or solid phase for particles are problem specific. The required computational power to directly solve these partial differential equations is enormous. For example, the fastest petaflops supercomputers allow up to approximately 10^{12} grid resolution that is only sufficient to solve simple indoor airflows in a single room, where a typical Reynolds number is 10^5. Directly solving an outdoor airflow problem is impossible, as Reynolds numbers are on the order of 10^7. Therefore, the required grid resolution for a simple outdoor airflow problem would be close to 10^{16}. As a compromise, building simulations have to be based on accurate physical models that can be successfully implemented and solved with the available computational power.

For the past couple of decades, modeling of buildings was accomplished using several approximations that were quite important for understanding physical transport processes in and around buildings even as we were gaining access to unprecedented computational power. More recently, those models are being coupled in unifying simulation platforms, and novel methods for leveraging different models are being discovered. For example, multizone modeling (MZ), energy simulations (ES), and computational fluid dynamics (CFD) based on Reynolds-averaged Navier-Stokes equations all have their strengths and weaknesses in modeling building transport processes. MZ can predict infiltration rates, bulk flow, and contaminant transport; ES can predict building energy consumption; while CFD can predict detailed airflow, temperature, and contaminant concentrations. For the same simulation domain of a single building, MZ typically requires seconds, ES takes minutes, and CFD needs hours to run a model on a personal computer. This is due to different levels of model complexity, which correspond to the level of details that each model provides. No matter how simple or complex, each of these models has supported development of sustainable building solutions, such as natural and hybrid ventilation, advanced building enclosure materials, and unconventional mechanical systems.

Initial approaches to sustainable building solutions have focused on simulation models of a single building and the environmental impacts at the occupancy and building scales based on past climatic conditions. This historic overlay was partially due to both the limited computational capacity and our limited understanding of driving transport processes for buildings. Simulations at the single-building scale are now widely used even though there are still important issues to be resolved with the accuracy of these models. Contemporary work is focused on improving the accuracy of existing single-building models via in situ validation studies and assimilation of in situ measured data into simulation platforms. Energy simulation models can produce results that closely track measured data or can produce errors as large as an order of magnitude when compared to the actual building energy consumption. During recent validation efforts, it was found that one of the largest sources of error in physical models was due to modulating factors driven by human behavior. Those factors include building systems management and maintenance, occupancy rates, and how occupants use building systems. In our recent study, we have also found that human behavior factors can become insignificant for building energy consumption when the outdoor and indoor air temperatures are very different, such as during typical winter days in the northeastern United States. Overall, as our understanding of physical transport processes and their modulating factors improve, so should model accuracy.

In the next couple of decades, it is expected that with ever-increasing numbers of buildings and renovations of existing structures, the impact on augmenting local microclimates will be on the order of a 10^3-meter radius, which will also amplify large-scale transport processes in the boundary layer and lower troposphere. Therefore, buildings should not be simulated based on past climatic conditions, and sustainable building design concepts have to be conceptualized on much larger spatial and temporal scales than the comparatively miniscule footprint of a single building and its associated annual energy consumption. This presents a unique opportunity for unprecedented synthesis of simulation models from meteorology and engineering, where a range of scales encapsulating relevant processes should be integrated, including (1) mesoscale predictions at a scale of $\sim 10^5$ m, (2) weather forecasting ($\sim 10^3$ m), (3) microscale outdoor transport processes ($\sim 10^0$ m), and (4) indoor transport processes ($\sim 10^{-1}$ m). The range of relevant spatial and temporal scales is substantial, but this challenge represents a profound opportunity for innovations in simulation technology that will enable progress in the development of sustainable urban settlements. There are several outstanding efforts that are attempting to address this exciting and daunting research problem globally, including those led by scientists from Japan (Yamaguchi and Shimoda, 2010), Europe (Rasheed et al., 2011), the United States (Chen et al., 2011), as well as the building science research group led by the author. If successful, these efforts can be extended from performance predictions to sustainable engineering of urban settlements at an unprecedented scale.

REFERENCES

Chen, F., H. Kusaka, R. Bornstein, J. Ching, C. S. B. Grimmond, S. Grossman-Clarke, T. Loridan, K. W. Manning, A. Martilli, S. Miao, D. Sailor, F. P. Salamanca, H. Taha, M. Tewari, X. Wang, A. A. Wyszogrodzki, and C. Zhang. 2011. The integrated WRF/urban modelling system: Development, evaluation, and applications to urban environmental problems. International Journal of Climatology 31(2):273–288.

EIA. 2011. Annual Energy Outlook 2011: With Projections to 2035. U.S. Department of Energy, Energy Information Administration (EIA), Report, DOE/EIA-0383.

Rasheed, A., D. Robinson, A. Clappier, C. Narayanan, and D. Lakehal. 2011. Representing complex urban geometries in mesoscale modeling. International Journal of Climatology, 31(2):289–301.

Yamaguchi, Y., and Y. Shimoda. 2010. District-scale simulation for multi-purpose evaluation of urban energy systems. Journal of Building Performance Simulation 3(4):289–305.

Accelerating Green Building Market Transformation with Information Technology

Christopher Pyke
U.S. Green Building Council

ABSTRACT

The green building movement seeks to transform the way that built environments are designed, constructed, and operated. Over the past decade, the tools of this transformation have included market interventions such as professional training and accreditation, project rating systems, and the third-party certification processes. These interventions have made a demonstrable difference in the industry with levels of participation exceeding 10% of new commercial construction in leading metropolitan areas. Today, the movement is envisioning the interventions it will need to dramatically scale up and extend this impact. The foundation for these new approaches will rest on information technology and analytics—tools that will provide unprecedented insights into market activity and allow near-real-time comparison and benchmarking. These emerging capabilities will create new dimensions for market competition, competitive advantage for high-performing projects, and increasing risks for low performers. Taken together, these approaches will accelerate and intensify the movement toward high-performance, green buildings and communities

INTRODUCTION

The green building movement seeks to advance the design, construction, and operation of built environments to promote human health, well-being, and the restoration of the natural environment. The contemporary green building movement began two decades ago with a powerful mental image and a simple idea. The image was a classic curve—the distribution of practice across the industry ranging from a few scofflaws, through the average-performing majority, to a small group of innovators—a variation on a pattern recognized across many industries (Roger, 1962). The idea was to use strategic market interventions to permanently shift this distribution toward higher performance. At the time, very little information existed to define this conceptual distribution of practice, and there was little experience with specific market interventions.

The lack of experience or data did not deter the movement. The nascent green building industry set course and went to work with passion. The early areas of focus included efforts to create a broad-based industry coalition, grow a trained workforce, create assessment tools, and reward buildings based on performance and achievement. There are clear signs of success in each of these areas. Here, we will focus on new opportunities pertinent to project-based information and analytics.

Green building practice rests on tools and processes to design and assess high-performance green buildings and communities (i.e., projects). These tools and processes allow practitioners to identify and communicate about relative merits of green building strategies (e.g., integrative design, energy efficiency, or water conservation), the achievement of milestones (e.g., facilities management policies), and, ultimately, the performance of whole systems ranging from interior spaces to neighborhoods (e.g., whole-building energy performance).[1] These tools and processes are codified in building rating systems, such as the U.S. Green Building Council's Leadership in Energy and Environmental Design (LEED™)[2] and a number of analogous systems around the world (Cole, 1999).

LEED provides practitioners with a platform to advance the consideration of issues related to location and transportation, design and engineering processes, construction activity, site planning, energy, water, materials, indoor environmental quality, and innovation. One of LEED's fundamental benefits to the market is greater transparency about the achievements and performance of buildings with regard to these previously invisible characteristics.

Over the past decade, the day-to-day tools underlying LEED have been a simple paper scorecard and, at the end of the process, a glass plaque displayed in a building lobby. It is remarkable to consider the impact that these simple elements have had on the industry. Today, we have the opportunity to build on these fundamental goals and concepts with information technologies that can vastly accelerate and scale up their impact. This paper describes one vision for this new phase of information-powered, analytically driven market transformation.

[1] In this context, the term "performance" refers to a measurable, typically quantitative metric, such as energy efficiency, renewable energy generation, water consumption, or occupant satisfaction. The term "achievement" refers to binary or qualitative activities, such as policies, procedures, or discrete choices (e.g., green cleaning, commissioning, or the use of third-party certified building products). The terms are often used together as "performance and achievement" to reflect the typical range of green building practice.

[2] See *www.usgbc.org/leed* for more information.

MARKET EFFICIENCY

Classical economics assumes that market participants have equal and immediate access to information (Fama, 1970; Malkiel, 2003). Markets use this information to set prices and value assets. Today, real estate markets have developed sophisticated tools to provide information on the financial aspects of individual buildings and portfolios. Commercial information services provide data and benchmarking related to capital cost, sales price, tenancy, and a myriad of other factors. Markets for this information are sophisticated and highly segmented. However, there are no readily accessible, consistent, or comprehensive resources to address the non-financial dimensions of assets, such as energy use, water consumption, or occupant experience in or around the property. The absence of this information contributes to inefficient markets, impairs innovation, and, in some cases, contributes to market failure (Gillingham et al., 2009).

The most direct remedy to this situation is to create public and private mechanisms to provide information on the non-financial aspects of assets, that is, the green dimensions of homes and commercial buildings. This can be accomplished through public labeling programs and private efforts to create asset dashboards and key performance metrics. The development of these programs is accelerating, witnessed by the success of building-level Energy Performance Certificates in the European Union, green building certification, and, in a few major metropolitan areas, municipal energy benchmarking (IEA, 2010). However, these efforts only scratch the surface. Ultimately, we need to connect information about energy performance with detailed information about project attributes (e.g., technologies, management strategies, etc.) and utilization (e.g., occupancy schedules and occupant density). These data must then be embedded in tools and services explicitly designed to foster constructive competition and accelerate market transformation.

ACCELERATING THE DIFFUSION OF INNOVATION

Information about outcomes and performance are the foundation and currency for the next generation of green buildings. However, by itself, this is not sufficient to propel market transformation. Information alone does not drive the change. It needs to be interpreted and attached to mechanisms that create clear market opportunities for high-performing projects and, by extension, competitive risks for low performers. This is where our interest diverges from agnostic market analytics. The green building movement seeks to use this information to drive permanent, self-sustaining change. Simply providing richer reporting on the status quo is inadequate. Our success will ultimately be measured by the rate and magnitude of change.

This means that we seek to use information technology to actively accelerate the diffusion of innovation. This concept refers to the rate with which new practices are taken up by market participants. Some industries have a long tradition of embracing diffusion theory and working across research, development, and deployment to accelerate change. For example, programs to increase appliance efficiency have raised the bar repeatedly over the past decades and achieved notable success (Gillingham et al., 2006; Nadel et al., 2003).

The building sector as a whole has not traditionally embraced these concepts, particularly for whole buildings or real estate portfolios. Yet, information technologies create opportunities for new, scalable market interventions. We recognize and address three key dimensions in our work:

- Define outcomes—dimensions for performance and evaluation. Green building is not, in isolation, an outcome. It is a framework, and we have developed new approaches to define and evaluate specific outcomes expected from green buildings. These outcomes provide the basis for market competition and differentiation.
- Understand and reward relatively high performers. These new performance dimensions can be used to sort and rank projects, discover their underlying practices, products, and services, and create performance-based reward systems.
- Inspire and assist relatively low performers. Conversely, this information creates the opportunity for low-performing projects to identify higher-performing peers and potential solution providers.

Outcomes

Over the past decade, green building has been rooted in a single, simple perform dimension: the total number of points a project achieves with respect to the criteria of a rating system. This dimension is segmented into categories, such as LEED's Certified, Silver, Gold, and Platinum. The act of certification and, at times, the level certification became a goal in itself.

Over the past several years, we have explored new approaches to expand this traditional focus, including developing and implementing a multidimensional framework linking green building outcomes and practices. Our terminology has evolved with our understanding. An initial version of this framework was released with LEED 2009 (USGBC, 2008). In this framework, every green building "credit" (a.k.a. strategy) is quantitatively associated with 13 environmental "impact categories," such as greenhouse gas emissions, resource depletion, and smog formation (Figure 1). This is used to assign weights (points) to individual credits. It also allows credit achievement to be used to track specific outcomes—literally unpacking information collected during certification.

In LEED 2012, these categories will be adapted to include seven core green building outcomes (e.g., greenhouse gas emissions reduction) supported by a set of more than 30 metrics (e.g., energy efficiency, renewable energy production). The bottom line is that the design of the next generation of ratings systems will

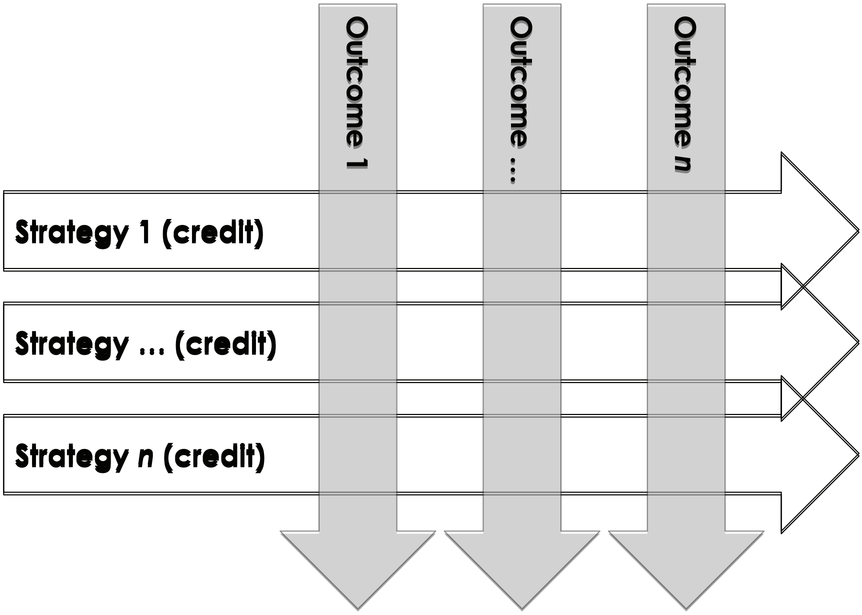


FIGURE 1 LEED is an outcome-oriented rating system. Points (relative weights) are assigned based on the association between credits (a.k.a., strategies) and outcomes (e.g., greenhouse gas emissions reductions).

be closely tied to specific outcomes. In turn, these design tools will create opportunities for advanced analytics relating project performance and achievement to specific outcomes. We can unpack certification-based data to focus on specific outcomes and their associated strategies.

Taken together, a paradigm is rapidly emerging that will allow green buildings to be defined and analyzed across a set of well-defined, sometimes standardized, performance dimensions or outcomes (e.g., UNEP SBCI, 2010). These outcomes or performance dimensions can be as "simple" as energy use intensity (e.g., annual energy use per square unit of floor space) or much more complicated, synthetic measures, such as the 29 weighted factors included in the LEED 2009 GHG Index. Each of these metrics provides a new dimension to rank and sort green building projects with respect to different goals and outcomes.

Higher Performers

Each performance dimension is populated with real projects using third-party verified data collected during the certification process. In every case, we have the opportunity to identify and reward high performers. Simply scoring based on performance (e.g., an Energy Star score) is a first step. However, technology allows us to create and share more valuable information. We seek to understand the factors that contribute to a level of performance and achievement (Figure 2). Fundamentally, we want to understand how projects achieve a given level of relative performance. This means identifying and tracking relationships between people, organizations, practices, technologies, and a myriad of other factors. Each high-performing project has value as a model for lower-performing projects and a milestone for those that achieved it.

Today, we can use a demonstration information system called the Green Building Information Gateway (*www.gbig.org*) to begin to identify and explore high-performing projects across multiple outcome dimensions. For example, Table 1 illustrates the performance and achievement of an exemplary office building in Chicago, Illinois, across six categories. The accompanying density plots compare the selected project (the dark triangle) with others certified using the same rating system, in this case LEED for Existing Building: Operations & Maintenance (more information is available from *http://www.gbig.org/projects/10049661*).

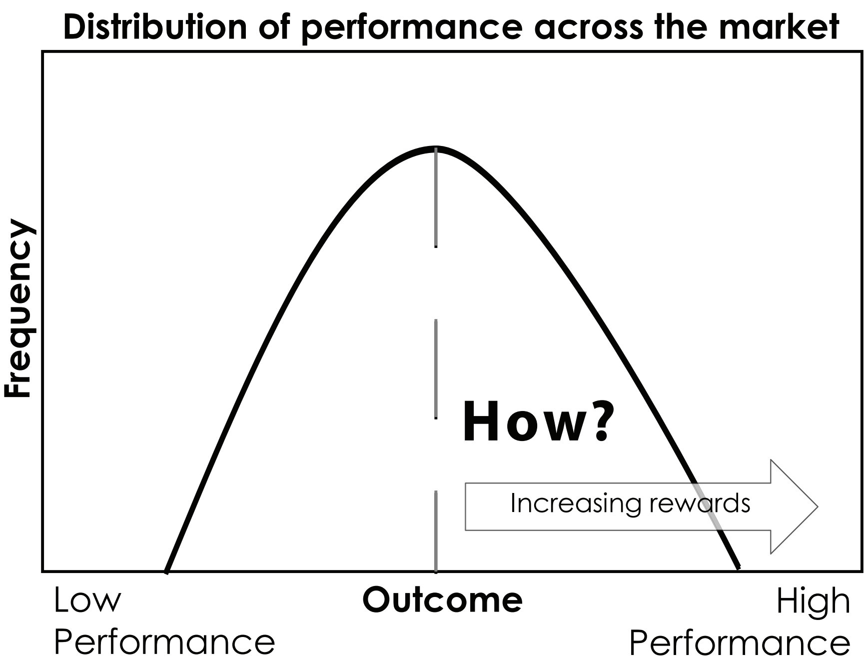


FIGURE 2 The Green Building Information Gateway (GBIG) provides data on the distribution of performance across populations of green building projects. This information can be used to understand "how" high-performing projects deliver above average results. This provides the basis for recognition and competitive advantage.

TABLE 1 A sample "nutrition" label for an exceptionally high-achieving LEED-certified project.

CATEGORY	POINTS	RANKING ?
Energy and Atmosphere	25/30	
Materials and Resources	10/14	
Indoor Environmental Quality	17/19	
Sustainable Sites	8/12	
Water Efficiency	7/10	
Innovation in Operations	7/7	
Total	74	

Source: Green Building Information Gateway (*www.gbig.org*), URL: *http://www.gbig.org/projects/10049661*.

Our ambition is to use this type of data and information technology to shorten cycles between innovation, market uptake, operational performance, and positive recognition. This means creating highly-scalable information systems to collect data on performance, practice, and technology in near real time and provide dynamic, context-relevant benchmarking and recommendations. This will provide decision makers with clear and timely information for their market; green "comps" not currently available in the real estate industry.

Lower Performers

The green building industry has always been comfortable recognizing high performers. The preceding approach to high performers accelerates this process and increases the timeliness and relevance of information flows as tools for market transformation. However, for every high performer there are a commensurate number of underperformers. Outside of Lake Woebegone, such underperformers are statistically inevitable.

Yet, we have been less aggressive in rigorously searching them out and trying to understand and assist them. We must find a way around our inhibitions regarding underperformance and low achievement. We must pursue an understanding of these projects that is equal to or greater than our energies devoted to recognizing and rewarding high performance.

Fortunately, we can adapt the same foundation of information technologies to identify projects that underperform or achieve less than their peers (Figure 3). We can then dive deeper to understand why these projects lag their peers and recommend specific strategies for improvement based on practices used by comparable higher-performing projects. We want to understand the challenges and, if necessary, create new or improved interventions to barriers such as technology limitations, technical understanding, or first costs.

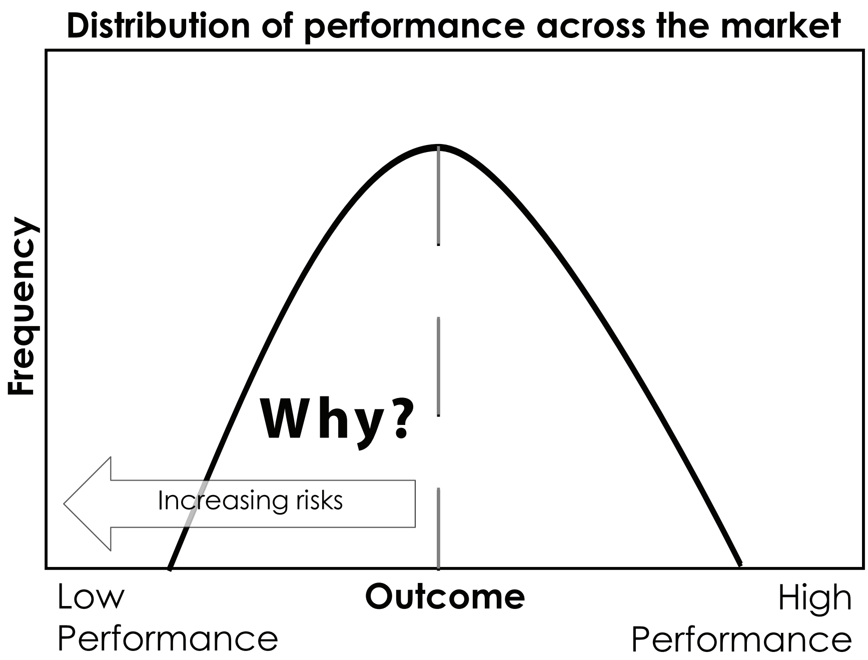


FIGURE 3 Information in GBIG can also be used to understand “why” relatively low-performing projects deliver below average result. This provides the basis for critical evaluation and improvement.

Again, we can use the Green Building Information Gateway (*www.gbig.org*) to begin to identify and explore relatively low-achieving projects across multiple outcome dimensions. Table 2 illustrates selected metrics for a LEED for New Construction (version 2.2) project in Washington, DC (more information is available from *http://www.gbig.org/projects/10100317*).

Fundamentally, this is a simple mirror image of our approaches to high performers. We seek to use information technologies to "unpack" projects, identify similar, higher-performing projects, and use data analysis to flag potential problems. We have the opportunity to use information technologies to highlight strategies used by comparable higher-performance projects.

TABLE 2 A sample "nutrition" label for a relatively low-achieving LEED-certified project.

CATEGORY	POINTS	RANKING ?
Energy and Atmosphere	1/17	
Materials and Resources	7/13	
Indoor Environmental Quality	5/15	
Sustainable Sites	7/14	
Water Efficiency	2/5	
Innovation in Design	5/5	
Total	27	

Source: Green Building Information Gateway (*www.gbig.org*), URL: *http://www.gbig.org/projects/10100317*.

CONCLUSION

The success of green building over the past decade attests to the ability for relatively simple interventions to produce demonstrable market transformation. The coming decade requires new tools and approaches to bring these concepts to scale and to generate the pace of change needed to achieve our mission of creating sustainable, healthy, high-performance built environments.

My belief is that this change will be powered by a new generation of information technologies specifically designed and deployed to promote market-based competition across multiple dimensions, to understand and learn from high performers, and to recognize and improve low performers. Every performance dimension we track provides an opportunity for competitive differentiation. Every high-performing project we identify and rank provides an opportunity to learn, recognize, and reward. Every low-performing project we touch provides an opportunity for education, investment, and improvement.

The critical technologies are in hand or rapidly emerging, including search, recommendation engines, distributed sensors, social media, service-based software architectures, and cloud solutions. We will engage orders-of-magnitude more projects and, ultimately, move from an episodic "certification event" to a regime of continuous performance and real-time monitoring. We can see the contours of this new world and envision its sweeping implications for green building practice.

REFERENCES

Cole, R. J. 1999. Building environmental assessment methods: Clarifying intentions. Building Research and Information 27(4/5):230–246.

Fama, E. 1970. Efficient capital markets: A review of theory and empirical work. Journal of Finance 25(2):383–417.

Gillingham, K., R. Newell, and K. Palmer. 2006. Energy efficiency policies: A retrospective examination. Annual Review of Environment and Resource Management 31:161–192.

Gillingham K., R. G. Newell, and K. Palmer. 2009. Energy Efficiency Economics and Policy. Resources for the Future Report No. RFF-DP-09-13.

IEA (International Energy Agency). 2010. Energy Performance Certification of Buildings. International Energy Administration Policy Pathway Report. Available at *http://www.iea.org/papers/pathways/buildings_certification.pdf*.

Malkiel, B. G. 2003. The efficient market hypothesis and its critics. Journal of Economic Perspectives 17(1):59–82.

Nadel, S., J. Thorne, H. Sachs, B. Prindle, and R. N. Elliott. 2003. Market Transformation: Substantial Progress from a Decade of Work. American Council for an Energy-Efficient Economy, Report No. A036.

Roger, E. M. 1962. Diffusion of Innovations, 1st edition. New York: The Free Press.

UNEP SBCI (United Nations Environment Programme, Sustainable Buildings and Climate Initiative). 2010. Common carbon metric for measuring energy use and reporting greenhouse gas emissions from building operations. Available at *http://www.unep.org/sbci/pdfs/Common-Carbon-Metric-for_Pilot_Testing_220410.pdf*.

USGBC (U.S. Green Building Council). 2008. LEED 2009 Credit Weighting. Washington, DC: USGBC. Available at *http://www.usgbc.org/DisplayPage.aspx?CMSPageID=1971*.

Neuroprosthetics

Introduction

Timothy Denison
Medtronic

Justin Williams
University of Wisconsin

The brain has always been attractive to engineers. Neurons and their connections, like tiny circuit elements, process and transmit information in a dramatic way that is intimately curious to researchers in the computer science and engineering fields. Neurons are amazing computational devices capable of both robust response to widely varied inputs and adaptability to changing conditions. Our most advanced computing systems are still dwarfed by the computational power of the human brain. Even small groups of neurons are capable of intricate interactions that produce basic mechanisms of learning and memory, highly parallel processing, and exquisite sensing capabilities.

Science has made great strides in the past few decades toward uncovering the basic principles underlying the brain's ability to receive sensation and control movement. These discoveries, along with revolutionary advances in computing power and microelectronics technology, have led to an emerging view that neural prosthetics, or electronic interfaces with the brain for restoration or augmentation of physiological function, may one day be possible. While the creation of a "six million dollar man" may still be far into the future, neural prostheses are rapidly becoming real potential treatments for a broad range of patients with injury or disease of the nervous system.

This session focuses on the types of engineering technology used to interface with the nervous system. This includes technology for stimulating the nervous system for restoration of sensory function as well as methods for extracting motor intention from the brain for use in artificial prostheses. In addition, we consider how lessons learned about the way the nervous system processes information can also be applied to circuit design—both for prosthetics and consumer circuits in general.

The papers in this session give perspectives from both academia and industry. Clinical studies are presented that span both basic research and commercial applications. Finally, discussion of emerging technologies that combine genetic and optical approaches provide a glimpse into the state of the art in neural interfacing technology.

James Weiland (Doheny Eye Institute, University of Southern California) covers the historical use of electrical stimulation of the nervous system and then focuses on recent clinical development of retinal implants to restore sight. He also gives a brief overview on the emerging field of optogenetics. Eric Leuthardt (Washington University) discusses the use of neural recording devices to extract motor command signals for applications as communication aids and brain machine interfaces for disabled populations. Finally, Rahul Sarpeshkar (Massachusetts Institute of Technology) presents new paradigms of "neuromorphic" processing—how we can learn from the brain's amazing processing properties and apply that knowledge in next-generation applications like cochlear prostheses.

Retinal Prosthetic Systems for Treatment of Blindness

James D. Weiland and Mark S. Humayun
Doheny Eye Institute, University of Southern California

Most functions of the human body are controlled by small electrical signals delivered via nerves. Thus, it is no surprise that electrical signals applied to the body from external sources can modulate physiological activity. In fact, reports of such physiological electromodulation date back to the eighteenth century, but scientists of that time lacked the understanding of neurophysiology to understand the basic mechanism behind electricity's ability to modulate biological activity. Advances in neurobiology, medicine, and engineering have allowed informed design of clinically beneficial implantable neurostimulation devices that are now used for a number of debilitating neurological diseases.

Implantable neural stimulators activate nerve cells. Nerves are a class of cells responsible for processing and communicating information between the brain and other parts of the body (Kandel et al., 1991). Sensory receptors are specialized nerve cells that convert physical stimuli into electrical signals that can be relayed by other nerve cells to the brain. Nerve cells are polarized, meaning an electrical potential can be measured across the cell membrane. Transient changes in membrane potential signal to other cells that an event has occurred. Neural networks composed of connected neurons determine if that event, along with inputs from others cells, requires action by other parts of the nervous system. Diseases or injuries that damage nerve cells, particularly sensory cells, can result in significant disability for the affected individual, including loss of sensory input, diminished capability to process information, or reduced motor function.

How can an electrical signal generated by an implanted device activate a nerve cell? It is not as simple as two wires being connected. Part of the complexity arises from the difference in electrical charge carriers; in metals, electrons carry charge, whereas in the body, ions carry charge. The conversion from electrons to ions occurs at the electrode, typically a metal or metal oxide, in direct contact with the extracellular fluid. An electrical signal applied to the electrode will cause current flow in the tissue via movement of charged ions, which include ions of sodium, chloride, and potassium, among others. The end effect of this charge movement is to depolarize the nerve cell membrane, after which the natural signaling mechanism of the nerve cell takes place and the brain, and person, observe an electrically elicited sensation or modulated function.

A number of successful neurostimulators are in widespread use. Cochlear implants stimulate the auditory nerve to give hearing to deaf people. Using these implants allows deaf people to hear so well that they can talk on a telephone. Unrelenting pain sensations sometimes result from nerve damage or disease. Implantable devices that stimulate the lower spinal cord are known to decrease or even eliminate the feeling of pain. One of the most dramatic successes of electrical stimulation is in the treatment of Parkinson's disease. By stimulating a part of the brain called the thalamus, many symptoms of Parkinson's subside almost immediately. Parkinson's patients with uncontrollable tremor or rigidity that severely limits motor function show improved coordination within minutes of commencing stimulation.

TREATING BLINDNESS WITH ELECTRICAL STIMULATION

The retina is the light-sensitive, multilayer tissue that lines the interior surface of the back of the eye (Figure 1) (see *http://webvision.med.utah.edu/*). Photoreceptors are the light-sensing cells of the retina, while the other cells process photoreceptor signals and send information to the brain via the optic nerve. When photoreceptors degenerate due to disease, the retina can no longer respond to light. However, the other nerve cells of the retina remain in sufficient numbers that electrical stimulation of these cells results in the perception of light. Blindness often results from progressive degeneration of the photoreceptors, which are the sensory nerves of the eye. Diseases like retinitis pigmentosa and age-related macular degeneration blind millions (Gehrs et al., 2010; Hartong et al., 2006). At first, symptoms are subtle, including difficulty seeing at night or blurred central vision. Ultimately, these conditions result in blindness. Given that vision is the sense by which people obtain most of the information about their environment, blindness has an extremely detrimental impact on the afflicted.

Electrical stimulation has been proposed as a treatment for blindness for decades, but only recently have systems been implemented that are consistent with clinical deployment. This first documented use of electrical stimulation to create visual perception dates to 1755, when Charles LeRoy discharged a large capacitor through the head of a blind person, who described "flames descending downwards" (Marg, 1991). Over time, as science, medicine, and engineering progressed, feasibility of a permanent implant to stimulate the retina was established.

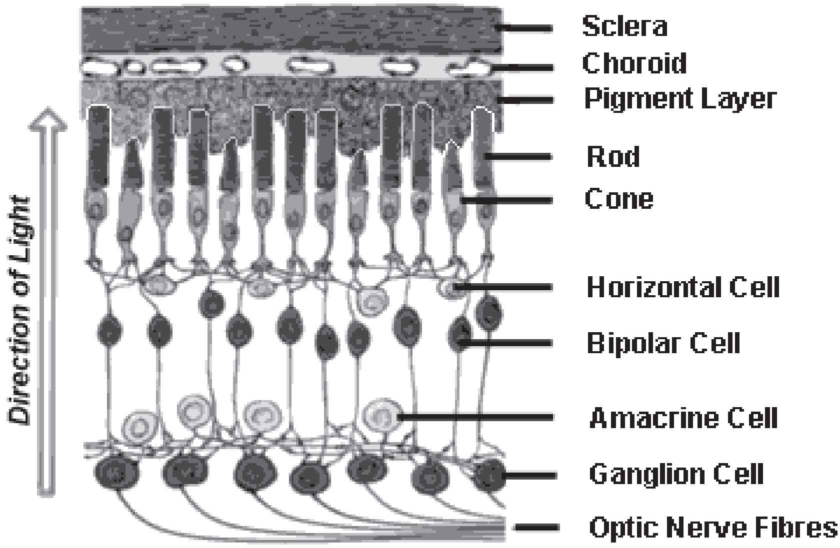


FIGURE 1 Cross section of the retina and other layers of the eye (choroid and sclera). The rods and cones are the photoreceptors that sense light. The other cells process this information and transmit electrical impulses via the optic nerve fibers to the brain. Source: Image from *http://www.catalase.com/retina.gif*.

RETINAL PROSTHESES: GENERAL DESCRIPTION AND CURRENT CLINICAL SYSTEMS

A retinal prosthesis consists of several components that perform specific functions: a camera to convert photons to digital data, a processing unit to generate stimulus commands based on the image, analog drivers to produce stimulus current, and an array of stimulating electrodes to deliver stimulus current to the retina (Weiland et al., 2005). As shown in Figure 2, the electrode array can be positioned in two locations in the eye, the epiretinal surface and the subretinal space, and these anatomical locations have come to define the two basic approaches that are being pursued for retinal prostheses. An epiretinal implant will rest on the inner limiting membrane of the retina, whereas a subretinal implant would be inserted in the space occupied by photoreceptors in a healthy retina.

Systems have been tested in blind humans in several clinical trials. For brevity, the two main trials that have produced the most significant results will be discussed. Retina Implant, GmbH has developed an externally powered, subretinal microphotodiode array (Figure 3-top). This device has been tested in 12 subjects (Zrenner et al., 2011). The device has 1,500 repeating units on a single silicon chip. Each unit has the following elements: a microphotodiode to sense light,

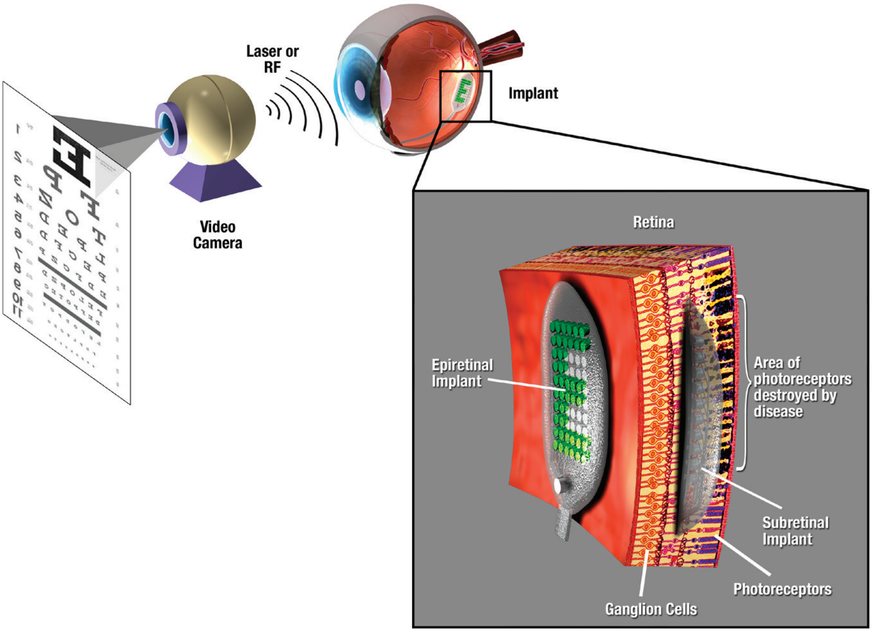


FIGURE 2 Retinal prosthesis concept. An image is captured by a camera and transmitted to an implant. The implant applies a patterned, complex stimulus to the retina via an array of electrodes on the surface of the retina (epiretinal) or underneath the retina (subretinal). Source: Weiland et al., 2005.

digital and analog circuitry to scale a voltage stimulus based on sensed light, and a microelectrode to apply the stimulus to the retina. The voltage stimulus pulse is supplied by a source that is outside the eye. The best test subject was able to read large letters (although it took a long time to do so) and demonstrated visual acuity of approximately 20/1000. The ARGUS II retinal prosthesis (Second Sight Medical Products, Inc.) has been implanted in 30 subjects (Figure 3-bottom). The ARGUS II has 60 electrodes. An external camera unit delivers image information wirelessly to the implant. The best visual acuity result to date is 20/1200 (Humayun et al., 2011). Letter reading was demonstrated in 22 of 30 subjects. Like the subretinal device, reading letters took much longer than reading with natural vision.

These results have generated considerable excitement in ophthalmology, vision research, and biomedical engineering. The possibility of restoring vision captures the imagination of many, and the reports from the human implantees testify to the promise of this approach. In controlled tests, improved mobility is clearly evident, and the subjects report more confidence during ambulation. How-

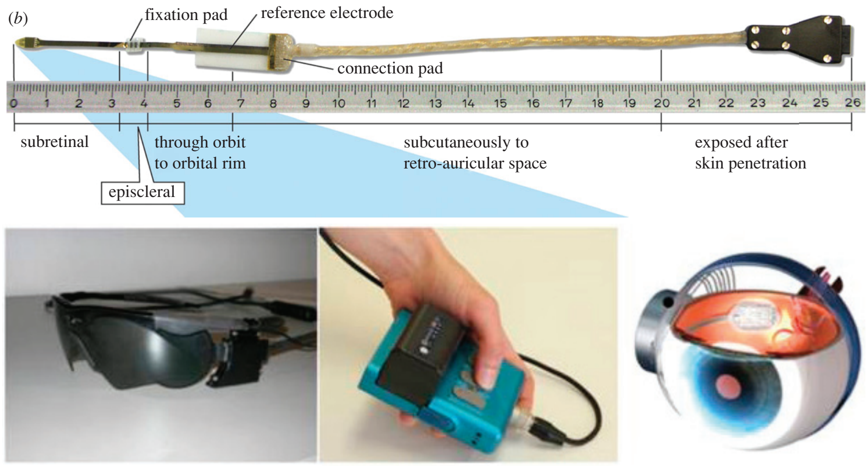


FIGURE 3 (top) A subretinal implant developed by Retina Implant GmbH has been tested in 11 subjects. The microphotodiode array that is implanted in the subretinal space is to the far left, while the remainder of the implant is cabling to allow connection to external test equipment. The epiretinal implant system developed by Second Sight Medical Products, Inc. includes (bottom left) a videocamera mounted in a pair of glasses, (bottom center) a wearable image processing system, and (bottom right) an epiretinal implant with components inside and outside the eye. Stimulus commands generated by the wearable image processing system are transmitted wirelessly to the implant via an inductive coil in the glasses frame. Source: (top) Zrenner et al., 2011. (bottom) Ahuja et al., 2011.

ever, from an objective analysis of the data as a whole, one must conclude that there remains a long road before the claim of "vision restoration" can be considered valid. By all measures, the individuals are still considered blind, even when the devices function well. The best reported visual acuity was 20/1000, whereas the cutoff for legal blindness is 20/200 (normal vision is 20/20). The field of view is limited. The ARGUS II device extends to 20 degrees (which is at the limit for legal blindness), but natural vision has a 180-degree field. The subretinal approach is limited in this regard, with current implants less than 10 degrees.

CHALLENGES FOR ARTIFICIAL VISION

To increase visual acuity with future implants, both technical and biological advances are needed. Simulations of artificial vision project that about 1,000 individual pixels are needed to allow visual function such as reading (at near normal speed) and face recognition. Improvements in electronic packaging (the materials and assembly techniques that protect the circuits from saline) will be needed to allow 1,000 signals from the circuits to connect to the retina. Low-power integrated circuits must be developed to generate a complex stimulus pattern to evoke form perception. Even if these barriers are overcome, more needs to be understood about how to connect the device to the retina. The subretinal device described above has 1,500 electrodes, but it does not achieve the visual acuity possible if each of these electrodes were acting independently. As electrode arrays become denser with smaller electrodes, better positioning of individual electrodes is needed to increase electrode-retina proximity. This will allow each electrode to activate a small part of the retina, thereby increasing visual acuity. Better stimulation strategies are needed to make the vision appear more natural and less artificial. Currently, perceptions fade within seconds as neural adaptation mechanisms attenuate the artificial input. Research questions should focus on optimizing the stimulus protocols to maximize user performance.

A new approach to artificial vision has potential to address some of the challenges facing electronic retinal prosthesis. The "optogenetic" technique modifies individual neurons to incorporate light-sensitive ion channels into the cell membrane, the most common light-sensitive channel being channelrhodopsin2 (ChR2) (Gradinaru et al., 2010). Ion channels are the means by which ions pass through the cell membrane, and whether channels are open or closed contributes to membrane potential. When light of a specific wavelength is shone on the cell, ChR2 ion channels open, resulting in depolarization of the cell. Bi et al. (2006) first demonstrated that this could be used to modify retinal ganglion cells, showing that light-evoked neural responses were present in a mouse model of retinal degeneration when the mouse retina cells contained ChR2, and others have extended this work. The optogenetic approach has some significant advantages over the bioelectronic approach. By making each cell light sensitive, vision can potentially be restored to near-normal acuity. Also, using light as the activating signal allows the optics of the eye to focus an image on the retina. In other words, the optogenetic technique can come much closer to restoring natural vision, versus artificial vision provided by bioelectronic approaches. However, artificial vision based on optogenetics has its own set of challenges that preclude clinical use. The main issue relates to sensitivity. Currently, modified cells require bright, blue light (460 nm) to be activated, roughly 7 orders of magnitude above the light sensitivity threshold in normally sighted people. It is not clear how such intense light would interact with a diseased retina, with remnant light sensitivity. Also, it is not known if cells can be modified permanently.

SUMMARY

These are interesting times for retinal prostheses. Clinical trials have shown both the promise and limitations of electrical stimulation as a treatment for blindness. Working in this field, one is constantly reminded of the wondrous sense of vision available to most people, how reliant we are on vision, and the devastating impact of vision loss. Given the complexity of vision, it is clear that prosthetic vision systems will, for the foreseeable future, provide vision that is artificial in appearance and below the resolution needed for complex, visually guided tasks. However, even a slight improvement in vision can have a large impact on quality of life. Restoring the ability to see large objects and detect motion can provide more confidence for someone as they navigate through unfamiliar environments. In addition, the brain has an amazing ability to adapt to new input and the individual, through experience, can use other contextual and sensory information to better understand what they are "seeing" via an artificial vision system. Although important breakthroughs have been achieved in this field, it is still early, and the journey toward restoration of high-acuity vision will involve concerted efforts by scientists, engineers, clinicians, and, most importantly, blind patients.

Financial Disclosures

Dr. Humayun has a financial interest in Second Sight Medical Products, Inc.

REFERENCES

Ahuja, A.K., J.D. Dorn, A. Caspi, M.J. McMahon, G. Dagnelie, L. daCruz, P. Stanga, M.S. Humayun, R.J. Greenberg, Argus II study group. Blind subjects implanted with the Argus II retinal prosthesis are able to improve performance in a spatial-motor task. British Journal of Ophthalmology 95(12):539–543.

Bi, A., J. Cui, Y.-P. Ma, E. Olshevskaya, M. Pu, A.M. Dizhoor, and Z.-H. Pan. 2006. Ectopic expression of a microbial-type rhodopsin restores visual responses in mice with photoreceptor degeneration. Neuron 50(1):23–33.

Gehrs, K.M., J.R. Jackson, E.N. Brown, R. Allikmets, and G.S. Hageman. 2010. Complement, age-related macular degneration and a vision of the future. Archives of Opthalmology 128(3):349–358.

Gradinaru, V., F. Zhang, C. Ramakrishnan, J. Mattis, R. Prakash, I. Diester, I. Goshen, K.R. Thompson, and K. Deisseroth. 2010. Molecular and cellular approaches for diversifying and extending optogenetics. Cell 141(1):154–165.

Hartong, D.T., E.L. Berson, and T.P. Dryja. 2006. Retinitis pigmentosa. The Lancet 368(November):1795–1809.

Humayun, M.S., L. da Cruz, G. Dagnelie, J.A. Sahel, P.E. Stanga, E. Filley, D. Eliott, J.L. Duncan, and R.J. Greenberg. 2011. Interim performance results from the Second Sight Argus II retinal prosthesis study. Investigative Ophthalmology & Visual Science 52:E-Abstract 2594.

Kandel, E.R., J.H. Schwartz, and T.M. Jessel. 1991. Principles of Neural Science, 3rd edition. Norwalk, CT: Appleton & Lange.

Marg, E. 1991. Magnetostimulation of vision: Direct noninvasive stimulation of the retina and the visual brain. Optometry and Vision Science 68(6):427–440.

Weiland, J.D., W. Liu, and M.S. Humayun. 2005. Retinal prosthesis. Annual Review of Biomedical Engineering 7:361–401.

Zrenner, E., K.U. Bartz-Schmidt, H. Benav, D. Besch, A. Bruckmann, V.P. Gabel, F. Gekeler, U. Greppmaier, A. Harscher, S. Kibbel, J. Koch, A. Kusnyerik, T. Peters, K. Stingl, H. Sachs, A. Stett, P. Szurman, B. Wilhelm, and R. Wilke. 2011. Subretinal electronic chips allow blind patients to read letters and combine them to words. Proceedings of the Royal Society of London, Series B: Biological Sciences 278(1711):1489–1497.

The Evolution of Brain-Computer Interfaces

Eric C. Leuthardt
Washington University

ABSTRACT

The notion that a computer can decode brain signals to infer the intentions of a human subject and then enact those intentions directly through a machine is emerging as a realistic technical possibility. These types of devices are known as brain-computer interfaces or BCIs. The evolution of these neuroprosthetic technologies could have significant implications for patients with motor disabilities by enhancing their ability to interact and communicate with their environment. Classically, the cortical physiology that has been most investigated and utilized for device control has been brain signals from the primary motor cortex. To date, this classic motor physiology has been an effective substrate for demonstrating the potential efficacy of BCI-based control. Emerging research in cortical physiology, however, now stands to further enhance our understanding of the cortical physiology underpinning human intent and provides further signals for more complex brain-derived control. In this review, we discuss the current status of BCIs and detail the emerging research trends that stand to further augment clinical application in the future.

INTRODUCTION

The notion that the brain can be directly accessed to allow a human being to control an external device with thoughts alone is emerging as a real option for patients with motor disabilities. This area of study, known as neuroprosthetics, has sought to create devices, known as *brain-computer interfaces* (BCIs), that acquire brain signals and translate them to machine commands such that they reflect the intentions of the user. In the past 20 years, the field has progressed rapidly from fundamental neuroscientific discovery to initial translational applications. Examples are seen in the seminal discoveries by Georgopoulus and Schwartz that neurons in the motor cortex, when taken as a population, can predict the direction and speed of arm movements in monkeys (Georgopoulos et al., 1982, 1986; Moran and Schwartz, 1999a). In the subsequent decades, these findings were translated to increasing levels of brain-derived control in monkeys and to preliminary human clinical trials (Hochberg et al., 2006; Taylor et al., 2002). Fundamental to the evolution of neuroprosthetic application, this brain-derived control is dependent on the emerging understanding of cortical physiology as it encodes information about intentions. In recent years, an emerging understanding of how the cortex encodes motor and nonmotor intentions, sensory perception, and the role that cortical plasticity plays in device control have led to new insights in brain function and BCI application. These new discoveries stand to further expand the potential of neuroprosthetics in regards to both control capability and patient populations that will be served. In this review, we provide an overview of current BCI modalities and of emerging research on the use of nonmotor areas for BCI applications, and we assess their potential for clinical impact.

BRAIN-COMPUTER INTERFACE: DEFINITION AND ESSENTIAL FEATURES

A BCI is a device that can decode human intent from brain activity *alone* in order to create an alternate communication channel for people with severe motor impairments. More explicitly, a BCI does not require the "brain's normal output pathways of peripheral nerves and muscles" to facilitate interaction with one's environment (Wolpaw et al., 2000, 2002). A real-world example of this would entail a quadriplegic subject controlling a cursor on a screen with signals derived from individual neurons recorded in the primary motor cortex without the need of overt motor activity. It is important to emphasize this point. A true BCI creates a completely new output pathway for the brain.

As a new output pathway, the user must have feedback to improve the performance of how one alters one's electrophysiological signals. Similar to the development of a new motor skill (e.g., learning to play tennis), there must be continuous alteration of the subject's neuronal output. The output should be matched against feedback from intended actions such that the subject's output (swinging the tennis racket or altering a brain signal) can be tuned to optimize performance toward the intended goal (getting the ball over the net or moving a cursor toward a target). Thus, the brain must change its signals to improve performance, but additionally the BCI may also be able to adapt to the changing milieu of the user's brain to further optimize functioning. This dual adaptation requires a certain level of training and a learning curve, both for the user and for the computer. The better the computer and subject are able to adapt, the shorter the training that is required for control.

There are four elements essential to the practical functioning of a BCI platform (Figure 1):

1. Signal acquisition, the BCI system's recorded brain signal or information input;
2. Signal processing, the conversion of raw information into a useful device command;
3. Device output, the overt command or control functions that are administered by the BCI system; and
4. Operating protocol, the manner in which the system is altered and turned on and off (Wolpaw et al., 2002).

All of these elements play in concert to manifest the user's intention in his or her environment.

Signal acquisition is some real-time measurement of the electrophysiological state of the brain. This measurement of brain activity is usually recorded via electrodes, but this is by no means a theoretical requirement. These electrodes can be either invasive or noninvasive. The most common types of signals include electroencephalography (EEG), electrical brain activity recorded from the scalp

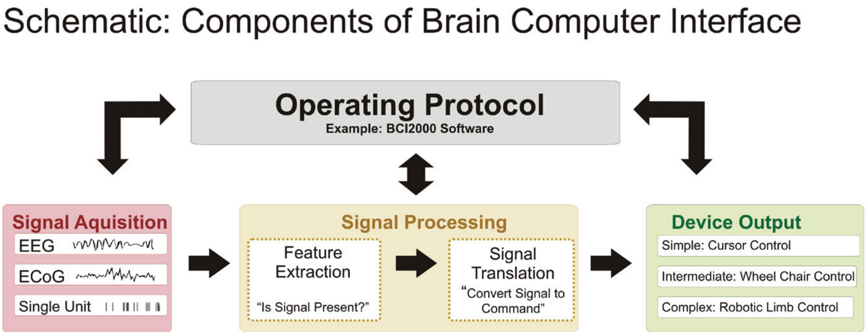


FIGURE 1 Essential features and components of a BCI. There are four essential elements to the practical functioning of a brain computer interface platform: (1) signal acquisition, the BCI system's recorded brain signal or information input; (2) signal processing, the conversion of raw information into a useful device command; (3) device output, the overt command or control functions that are administered by the BCI system; and (4) operating protocol, the manner in which the system is turned on and off, and the way the user or a technical assistant adjusts parameters of the previous three steps in converting intentions to machine commands. All of these elements play in concert to manifest the user's intention in his or her environment (Schalk et al., 2004b). Source: Leuthardt et al. (2009).

(Elbert et al., 1980; Farwell and Donchin, 1988; Freeman et al., 2003; Pfurtscheller et al., 1993; Sutter, 1992; Vidal, 1977); electrocorticography (ECoG; Leuthardt et al., 2004, 2005), electrical brain activity recorded beneath the skull (Leuthardt et al., 2004, 2005; Schalk et al., 2004a); field potentials, electrodes monitoring brain activity from within the parenchyma (Andersen et al., 2004); and "single units," microelectrodes monitoring individual neuron action potential firing (Georgopoulos et al., 1986; Kennedy and Bakay, 1998; Laubach et al., 2000; Taylor et al., 2002). Figure 2 shows the relationship of the various signal platforms in terms of anatomy and population sampled. Once acquired, the signals are then digitized and sent to the BCI system for further interrogation.

In the signal-processing portion of BCI operation, there are two essential functions: feature extraction and signal translation. The first function extracts significant identifiable information from the gross signal, and the second converts that identifiable information into device commands. The process of converting a raw signal into one that is meaningful requires a complex array of analyses. These techniques can vary from assessment of frequency power spectra, event-related potentials, and cross-correlation coefficients for analysis of EEG or ECoG signals to directional cosine tuning of individual neuron action potentials (Levine et al., 2000; Moran and Schwartz, 1999a; Pfurtscheller et al., 2003). The impetus for these methods is to determine the relationship between an electrophysiologic event and a given cognitive or motor task. As an example, after recordings are made from an ECoG signal, the BCI system must recognize that a signal alteration has occurred in the electrical rhythm (feature extraction) and then associate that change with a specific cursor movement (translation). As mentioned above, it is important that the signal processing be dynamic such that it can adjust to the

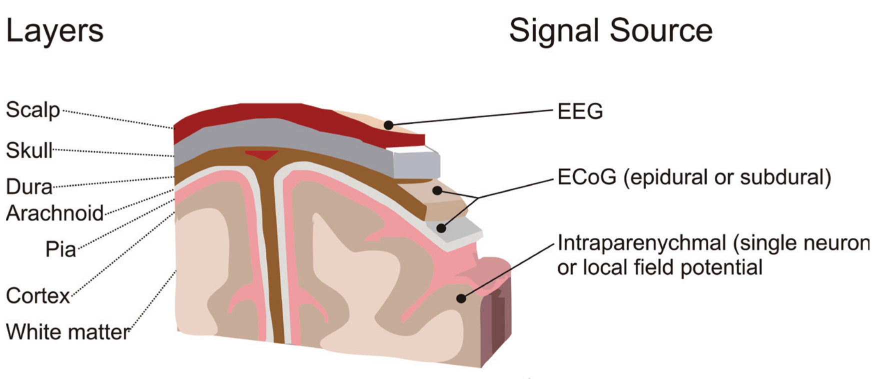


FIGURE 2 Signals for BCI. Three general categories of signals that are used for BCI application and their anatomic location relative to the brain and its respective covering layers. EEG, electroencephalography; ECoG, electrocorticography. Source: Leuthardt et al. (2009).

changing internal signal environment of the user. In regards to the actual device output, this is the overt action that the BCI accomplishes. As in the previous example, this can result in moving a cursor on a screen; other possibilities are choosing letters for communication, controlling a robotic arm, driving a wheelchair, or controlling some other intrinsic physiologic process such as moving one's own limb or controlling one's bowel and bladder sphincters (Leuthardt et al., 2006a).

An important consideration for practical application is the overall operating protocol. This refers to the manner in which the user controls *how* the system functions. The "how" includes such things as turning the system on or off, controlling the kind of feedback and how fast it is provided, how quickly the system implements commands, and switching between various device outputs. These elements are critical for BCI functioning in the real-world application of these devices. In most current research protocols, these parameters are set by the investigator. In other words, the researcher turns the system on and off, adjusts the speed of interaction, or defines very limited goals and tasks. These are all things that the user will need to be able to do by him- or herself in an unstructured applied environment.

CURRENT BCI PLATFORMS

There are currently three general categories of BCI platforms that have been put forward as possible candidates for clinical application. These categories are primarily determined by the source from which the controlling brain signal is derived. The first category uses EEG, which involves brain signals acquired from the scalp. The second category, referred to as "single-unit systems," uses intraparenchymal microelectrodes that detect action potential firings of individual neurons. The third is an intermediate modality in which electrodes acquire signal from the cortical surface directly (either above or below the dura). The current status of each of these platforms is briefly reviewed in terms of level of control, surgical considerations, and current clinical populations served.

EEG-Based Systems

EEG-based BCIs use electrical activity recorded from the scalp (Birbaumer et al., 1999; Blankertz et al., 2006; Farwell and Donchin, 1988; Kübler et al., 2005; McFarland et al., 1993, 2008a; Millan et al., 2004; Muller et al., 2008; Pfurtscheller et al., 1993, 2000; Sutter, 1992; Vaughan et al., 2006; Wolpaw and McFarland, 1994, 2004; Wolpaw et al., 1991). Most BCI studies in humans used EEG, probably because this recording method is convenient, safe, and inexpensive.

EEG has a relatively poor spatial resolution. This is because a large brain area must be involved to generate the necessary detectable signals (Freeman et al., 2003; Srinivasan et al., 1998). Despite this limitation, signals relevant to BCI research can still be found in the EEG. This includes modulations of *mu*

(8-12 Hz) or *beta* (18-25 Hz) rhythms produced by sensorimotor cortex. These rhythms show nonspecific changes (typically decreases in amplitude) related to movements and movement imagery. They do not contain specific information about the details of movements, such as the position or velocity of hand movements. This may be an important limitation because signals associated with specific movement parameters are typically used in BCI systems based on action potential firing rates. Another issue of EEG recordings is that the detected amplitudes are very small. This makes them susceptible to artifacts created by sources outside the brain such as electromyographic signals produced by muscle contractions. Despite these potentially limiting issues, EEG-based BCIs have been shown to support higher performance than is often assumed, including accurate two-dimensional (McFarland et al., 2008a; Wolpaw and McFarland, 2004) and even three-dimensional control of a computer cursor (McFarland et al., 2008b). To date, the large majority of clinical application of BCI technologies for people with severe motor disabilities have been demonstrated using EEG (Kübler et al., 2005; Nijboer et al., 2008; Vaughan et al., 2006). Ultimately, this intrinsic lack of signal robustness may have important implications for chronic application of BCI systems in real-world environments. BCI systems based on EEG typically require substantial training (Birbaumer, 2006; Wolpaw and McFarland, 2004) to achieve accurate one- or two-dimensional device control (about 20 or 50 thirty-minute training sessions, respectively), although some reports have reported training requirements that are shorter (Blankertz et al., 2006). These shortcomings of noise sensitivity and prolonged training are fundamental limitations in the scalability of widespread clinical application of EEG-based BCIs.

In summary, EEG has been shown to support much higher performance than previously assumed and is currently the only modality that has been shown to actually help people with paralysis. However, because of its important limitations, it is currently not clear to what extent EEG-based BCI performance, in the laboratory and in clinical settings, can be further enhanced.

Single-Neuron-Based Systems

From a purely engineering point of view, the optimal method to extract electrical information from the brain would be to place a series of small recording electrodes directly into the cortical layers (1.5–3 mm) to record signals from individual neurons. This, in essence, is what single-unit action potential BCI systems do, and they have been very successful for limited time periods in both monkeys (Carmena et al., 2003; Serruya et al., 2002; Taylor et al., 2002; Velliste et al., 2008) and humans (Hochberg et al., 2006; Kennedy and Bakay, 1998). To extract single-unit activity, small microelectrodes having ~20-micron-diameter tips are inserted in the brain parenchyma where relatively large (e.g., 300 microvolt) extracellular action potentials are recorded from individual neurons from 10-100 microns away. These signals are usually bandpass filtered from 300 to 10,000 Hz and then passed through a spike discriminator to measure spike time occurrences. The firing rates of individual neurons are computed in 10- to 20-millisecond bins and "decoded" to provide a high-fidelity prediction to control either a computer cursor or robot endpoint kinematics (Georgopoulos et al., 1986; Moran and Schwartz, 1999b; Wang et al., 2007). Given its high spatial resolution (100 microns) as well as its high temporal resolution (50-100 Hz), this modality arguably provides the highest level of control in BCI applications.

Unfortunately, there are two major problems with single-unit BCIs. First, the electrodes must penetrate into the parenchyma, where they cause local neural and vascular damage (Bjornsson et al., 2006). Second, single-unit action potential microelectrodes are very sensitive to encapsulation. Insertion of penetrating devices in the brain parenchyma damages neurons and vasculature, which can initiate a cascade of reactive cell responses, typically characterized by activation and migration of microglia and astrocytes toward the implant site (Bjornsson et al., 2006). The continued presence of devices promotes formation of a sheath composed partly of these reactive astrocytes and microglia (Polikov et al., 2005; Szarowski et al., 2003). This reactive sheath can have numerous deleterious effects, including neural cell death and an increased tissue resistance that electrically isolates the device from the surrounding neural tissue (Biran et al., 2005; Szarowski et al., 2003; Williams et al., 2007). Research into novel biomaterial coatings and/or local drug delivery systems that may reduce the foreign body response to implanted electrodes is ongoing but, to date, is far from clinical application (Abidian and Martin, 2008; Seymour and Kipke, 2007; Spataro et al., 2005). Until these issues are solved, this remains a limitation for developing a long-term BCI system based on single-unit activity.

Electrocorticography-Based Systems

Over the past 5 years, the use of ECoG as a signal platform for BCI has gained mounting enthusiasm as a more practical and robust platform for clinical application. As detailed above, both EEG- and single-unit-based systems have been impeded for large-scale clinical application. This is either due to prolonged user training and poor signal-to-noise limitations with EEG, or due to inability to maintain a consistent signal with current single-unit constructs (Bjornsson et al., 2006; Szarowski et al., 2003; Wolpaw and McFarland, 2004). The use of ECoG has been posited to be an ideal trade-off for practical implementation (Leuthardt et al., 2004). When compared to EEG, the signal is substantially more robust. Its magnitude is typically five times larger, its spatial resolution is much greater (0.125 versus 3.0 cm for EEG), and its frequency bandwidth is significantly higher (0-500 Hz versus 0-40 Hz for EEG) (Boulton et al., 1990; Freeman et al., 2003; Srinivasan et al., 1998). Of particular note, the access to higher-frequency bandwidths carries particularly useful information amenable to BCI operation (Gaona et al., 2011). Many studies have demonstrated that different frequency bands carry specific and anatomically distinct information about cortical processing. The lower-frequency bands known as *mu* (8-12 Hz) and *beta* (18-25 Hz), which are detectable with EEG, are thought to be produced by thalamocortical circuits and show broad anatomic decreases in amplitude in association with actual or imagined movements (Huggins et al., 1999; Levine et al., 1999; Pfurtscheller et al., 2003; Rohde et al., 2002). The higher frequencies only appreciable with ECoG, also known as *gamma* band activity, are thought to be produced by smaller cortical assemblies. Gamma activity shows close correlation with action potential firing of tuned cortical neurons in the primary motor cortex in monkey models (Heldman et al., 2006). Additionally, these high-frequency changes have been associated with numerous aspects of speech and motor function in humans (Chao et al., 2010; Crone et al., 1998, 2001a, 2001b; Gaona et al., 2011; Leuthardt et al., 2004; Schalk et al., 2007). Beyond higher information content, since the ECoG signal is recorded from larger electrodes that do not penetrate the brain, these constructs should have a higher likelihood for long-term clinical durability. This expectation of good long-term stability of ECoG sensors is supported by some pathologic and clinical evidence. For example, in cat, dog, and monkey models, long-term subdural implants showed minimal cortical or leptomeningeal tissue reaction while maintaining prolonged electrophysiologic recording (Bullara et al., 1979; Chao et al., 2010; Loeb et al., 1977; Margalit et al., 2003; Yuen et al., 1987). Additionally, preliminary work in humans using the implantable NeuroPace device for the purpose of long-term subdural electrode monitoring for seizure identification and abortion has also been shown to be stable (Vossler et al., 2004).

The use of ECoG for BCI applications has been primarily studied in motor-intact patients with intractable epilepsy requiring invasive monitoring. Similar to EEG-based BCI systems, the ECoG approach has primarily focused on the use of changes in sensorimotor rhythms from motor cortex. What has been distinct, however, has been the access to the higher-frequency gamma rhythms with ECoG. Use of these higher-frequency rhythms has provided a significant advantage in regards to training requirements and multidimensional control. In 2004, Leuthardt et al. demonstrated the first use of ECoG in closed-loop control in a one-dimensional cursor control task with minimal training requirements (under 30 minutes). In additional experiments, the same group and others have demonstrated that specific frequency alterations encode very specific information about hand and arm movements (Leuthardt et al., 2004; Pistohl et al., 2008; Sanchez et al., 2008; Schalk et al., 2007). In 2006, Leuthardt et al. further demonstrated that ECoG control using signal from the epidural space was also possible (Leuthardt et al., 2006b). Schalk et al. (2008) showed that ECoG signals can be used for two-dimensional control whose performance was within the range of that shown before using invasive single-unit systems. Because the electrode arrays cover broad regions of cortex, several groups have begun to explore alternate cognitive modalities and their cortical physiologies to expand BCI device control. Felton et al. (2007)

showed that, in addition to motor imagery, sensory imagery could also be used for device control. The same group also demonstrated that the auditory cortex could be trained to acquire simple control of a cursor (Wilson et al., 2006). Ramsey et al. (2006) showed that higher cognitive functions, such as working memory in the dorsal lateral prefrontal cortex, can also be used for effective device operation. Recently, Leuthardt et al. (2011) demonstrated that phonemic content taken from speech networks could also be used for simple device control.

Taken together, these studies show that ECoG signals carry a high level of specific cortical information, and that these signals can allow a user to gain control rapidly and effectively. It is worth noting that these control paradigms have not been extended to motor-impaired subjects thus far. How these cortical signals will be affected in the setting of a spinal cord injury or ALS has not been explicitly tested.

CONCLUSIONS

The field of neuroprosthetics is growing rapidly. The cortical physiology that underpins the manner in which a human brain encodes intentions is beginning to be understood. This will have a significant impact in augmenting function for those with various forms of motor disabilities. As research stretches beyond motor physiology, the field of neuroprosthetics now stands to further expand in capability and in diversity of clinical populations served. The evolving understanding of cortical physiology as it relates to motor movements, language function, and plasticity could all provide higher levels of complexity in brain-derived control. Given the rapid progression of these technologies over the past decade and the concomitant swift ascent of computer processing speeds, signal analysis techniques, and emerging ideas for novel biomaterials, neuroprosthetic implants will hopefully in the near future be as common as deep brain stimulators are today. The clinical advent of this technology will usher in a new era of restorative neurosurgery and new human-machine interfaces.

ACKNOWLEDGEMENTS

The author would like to acknowledge the generous support from the following: the James S. McDonnell Foundation, the Department of Defense (W911NF-07-1-0415 & W911NF-08-1-0216 – Leuthardt), the National Institutes of Health (NINDS NIH R01-EB000856-06), and the Children's Discovery Institute.

Financial Disclosures

Dr. Leuthardt owns stock in the company Neurolutions.

REFERENCES

Abidian, M.R., and D.C. Martin. 2008. Experimental and theoretical characterization of implantable neural microelectrodes modified with conducting polymer nanotubes. Biomaterials 29(9):1273–1283.

Andersen, R.A., J.W. Burdick, S. Musallam, B. Pesaran, and J.G. Cham. 2004. Cognitive neural prosthetics. Trends in Cognitive Sciences 8(11):486–493.

Biran, R., D.C. Martin, and P.A. Tresco. 2005. Neuronal cell loss accompanies the brain tissue response to chronically implanted silicon microelectrode arrays. Experimental Neurology 195(1):115–126.

Birbaumer, N., N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. 1999. A spelling device for the paralysed. Nature 398(6725):297–298.

Birbaumer, N. 2006. Brain-computer-interface research: Coming of age. Clinical Neurophysiology 117(3):479–483.

Bjornsson, C.S., S.J. Oh, Y.A. Al-Kofahi, Y.J. Lim, K.L. Smith, J.N. Turner, S. De, B. Roysam, W. Shain, and S.J. Kim. 2006. Effects of insertion conditions on tissue strain and vascular damage during neuroprosthetic device insertion. Journal of Neural Engineering 3(3):196–207.

Blankertz, B., G. Dornhege, M. Krauledat, K.R. Muller, V. Kunzmann, F. Losch, and G. Curio. 2006. The Berlin Brain-Computer Interface: EEG-based communication without subject training. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2):147–152.

Boulton, A.A., G.B. Baker, and C.H. Vanderwolf. 1990. Neurophysiological Techniques, II: Applications to Neural Systems. Clifton, N.J.: Humana Press.

Bullara, L.A., W.F. Agnew, T.G. Yuen, S. Jacques, and R.H. Pudenz. 1979. Evaluation of electrode array material for neural prostheses. Neurosurgery 5(6):681–686.

Carmena, J., M. Lebedev, R. Crist, J. O'Doherty, D. Santucci, D. Dimitrov, S. Patil, C. Henriquez, and M. Nicolelis. 2003. Learning to control a brain-machine interface for reaching and grasping by primates. PLoS Biology 1(2):193–208.

Chao, Z.C., Y. Nagasaka, and N. Fujii. 2010. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Frontiers in Neuroengineering 3:3.

Crone, N.E., D.L. Miglioretti, B. Gordon, and R.P. Lesser. 1998. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. Brain 121(Pt 12):2301–2315.

Crone, N.E., D. Boatman, B. Gordon, and L. Hao. 2001a. Induced electrocorticographic gamma activity during auditory perception. Clinical Neurophysiology 112(4):565–582.

Crone, N.E., L. Hao, J. Hart, Jr., D. Boatman, R.P. Lesser, R. Irizarry, and B. Gordon. 2001b. Electrocorticographic gamma activity during word production in spoken and sign language. Neurology 57(11):2045–2053.

Elbert, T., B. Rockstroh, W. Lutzenberger, and N. Birbaumer. 1980. Biofeedback of slow cortical potentials. I. Electroencephalography and Clinical Neurophysiology 48(3):293–301.

Farwell, L.A., and E. Donchin. 1988. Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. Electroencephalography and Clinical Neurophysiology 70(6):510–523.

Felton, E.A., J.A. Wilson, J.C. Williams, and P.C. Garell. 2007. Electrocorticographically controlled brain-computer interfaces using motor and sensory imagery in patients with temporary subdural electrode implants. Report of four cases. Journal of Neurosurgery 106(3):495–500.

Freeman, W.J., M.D. Holmes, B.C. Burke, and S. Vanhatalo. 2003. Spatial spectra of scalp EEG and EMG from awake humans. Clinical Neurophysiology 114(6):1053–1068.

Gaona, C., M. Sharma, Z. Freudenburg, J. Breshears, D. Bundy, J. Roland, D. Barbour, G. Schalk, and E. Leuthardt. 2011. Nonuniform high-gamma (60-500 Hz) power changes dissociate cognitive task and anatomy in human cortex. Journal of Neuroscience 31(6):2091–2100.

Georgopoulos, A.P., J.F. Kalaska, R. Caminity, and J.T. Massey. 1982. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. Journal of Neuroscience 2(11):1527–1537.

Georgopoulos, A.P., A.B. Schwartz, and R.E. Kettner. 1986. Neuronal population coding of movement direction. Science 233(4771):1416–1419.

Heldman, D.A., W. Wang, S.S. Chan, and D.W. Moran. 2006. Local field potential spectral tuning in motor cortex during reaching. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2):180–183.

Hochberg, L.R., M.D. Serruya, G.M. Friehs, J.A. Mukand, M. Saleh, A.H. Caplan, A. Branner, D. Chen, R.D. Penn, and J.P. Donoghue. 2006. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. Nature 442(7099):164–171.

Huggins, J.E., S.P. Levine, S.L. BeMent, R.K. Kushwaha, L.A. Schuh, E.A. Passaro, M.M. Rohde, D.A. Ross, K.V. Elisevich, and B.J. Smith. 1999. Detection of event-related potentials for development of a direct brain interface. Journal of Clinical Neurophysiology 16(5):448–455.

Kennedy, P.R., and R.A. Bakay. 1998. Restoration of neural output from a paralyzed patient by a direct brain connection. Neuroreport 9(8):1707–1711.

Kübler, A., F. Nijboer, J. Mellinger, T.M. Vaughan, H. Pawelzik, G. Schalk, D.J. McFarland, N. Birbaumer, and J.R. Wolpaw. 2005. Patients with ALS can use sensorimotor rhythms to operate a brain-computer interface. Neurology 64(10):1775–1777.

Laubach, M., J. Wessberg, and M.A. Nicolelis. 2000. Cortical ensemble activity increasingly predicts behaviour outcomes during learning of a motor task. Nature 405(6786):567–571.

Leuthardt, E.C., G. Schalk, J.R. Wolpaw, J.G. Ojemann, and D.W. Moran. 2004. A brain-computer interface using electrocorticographic signals in humans. Journal of Neural Engineering 1(2):63–71.

Leuthardt, E.C., K.J. Miller, G. Schalk, R.N. Rao, and J.G. Ojemann. 2005. Electrocorticography-based brain computer interface—the Seattle experience. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2):194–198.

Leuthardt, E.C., G. Schalk, D. Moran, and J.G. Ojemann. 2006a. The emerging world of motor neuroprosthetics: a neurosurgical perspective. Neurosurgery 59(1):1–14; discussion 1–14.

Leuthardt, E.C., K.J. Miller, G. Schalk, R.P. Rao, and J.G. Ojemann. 2006b. Electrocorticography-based brain computer interface—the Seattle experience. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2):194–198.

Leuthardt, E.C., G. Schalk, J. Roland, A. Rouse, and D.W. Moran. 2009. Evolution of brain-computer interfaces: Going beyond classic motor physiology. Neurosurgical Focus 27(1):E4.

Leuthardt, E.C., C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, and G. Schalk. 2011. Using the electrocorticographic speech network to control a brain-computer interface in humans. Journal of Neural Engineering 8(3):036004.

Levine, S.P., J.E. Huggins, S.L. BeMent, R.K. Kushwaha, L.A. Schuh, E.A. Passaro, M.M. Rohde, and D.A. Ross. 1999. Identification of electrocorticogram patterns as the basis for a direct brain interface. Journal of Clinical Neurophysiology 16(5):439–447.

Levine, S.P., J.E. Huggins, S.L. BeMent, R.K. Kushwaha, L.A. Schuh, M.M. Rohde, E.A. Passaro, D. A. Ross, K.V. Elisevich, and B.J. Smith. 2000. A direct brain interface based on event-related potentials. IEEE Transactions on Neural Systems Rehabilitation Engineering 8(2):180–185.

Loeb, G.E., A.E. Walker, S. Uematsu, and B.W. Konigsmark. 1977. Histological reaction to various conductive and dielectric films chronically implanted in the subdural space. Journal of Biomedical Materials Research 11(2):195–210.

Margalit, E., J.D. Weiland, R.E. Clatterbuck, G.Y. Fujii, M. Maia, M. Tameesh, G. Torres, S.A. D'Anna, S. Desai, D.V. Piyathaisere, A. Olivi, E. de Juan, Jr., and M.S. Humayun. 2003. Visual and electrical evoked response recorded from subdural electrodes implanted above the visual cortex in normal dogs under two methods of anesthesia. Journal of Neuroscience Methods 123(2):129–137.

McFarland, D.J., G.W. Neat, and J.R. Wolpaw. 1993. An EEG-based method for graded cursor control. Psychobiology 21:77–81.

McFarland, D., D. Krusienski, W. Sarnacki, and J. Wolpaw. 2008a. Emulation of computer mouse control with a noninvasive brain-computer interface. Journal of Neural Engineering 5(2):101–110.

McFarland, D.J., W. Sarnacki, and J.R. Wolpaw. 2008b. Electroencephalographic (EEG) control of three-dimensional movement. Paper presented at the Society for Neuroscience Annual Meeting, Washington, DC, November 15–19.

Millan, J.R., F. Renkens, J. Mourino, and W. Gerstner. 2004. Noninvasive brain-actuated control of a mobile robot by human EEG. IEEE Transactions on Biomedical Engineering 51(6):1026–1033.

Moran, D.W., and A.B. Schwartz. 1999a. Motor cortical representation of speed and direction during reaching. Journal of Neurophysiology 82(5):2676–2692.

Moran, D.W., and A.B. Schwartz. 1999b. Motor cortical activity during drawing movements: Population representation during spiral tracing. Journal of Neurophysiology 82(5):2693–2704.

Moran, D.W., J.R. Wolpaw, and E.C. Leuthardt. 2008. Two-dimensional movement control using electrocorticographic signals in humans. Journal of Neural Engineering 5(1):75–84.

Muller, K.R., M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, and B. Blankertz. 2008. Machine learning for real-time single-trial EEG-analysis: From brain-computer interfacing to mental state monitoring. Journal of Neuroscience Methods 167(1):82–90.

Nijboer, F., E.W. Sellers, J. Mellinger, M.A. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D.J. Krusienski, T.M. Vaughan, J.R. Wolpaw, N. Birbaumer, and A. Kubler. 2008. A P300-based brain-computer interface for people with amyotrophic lateral sclerosis. Clinical Neurophysiology 119(8):1909–1916.

Pfurtscheller, G., D. Flotzinger, and J. Kalcher. 1993. Brain-computer interface—a new communication device for handicapped persons. Journal of Microcomputer Applications 16:293–299.

Pfurtscheller, G., C. Guger, G. Mueller, G. Krausz, and C. Neuper. 2000. Brain oscillations control hand orthosis in a tetraplegic. Neuroscience Letters 292(3):211–214.

Pfurtscheller, G., B. Graimann, J.E. Huggins, S.P. Levine, and L.A. Schuh. 2003. Spatiotemporal patterns of beta desynchronization and gamma synchronization in corticographic data during self-paced movement. Clinical Neurophysiology 114(7):1226–1236.

Pistohl, T., T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. 2008. Prediction of arm movement trajectories from ECoG-recordings in humans. Journal of Neuroscience Methods 167(1):105–114.

Polikov, V.S., P.A. Tresco, and W.M. Reichert. 2005. Response of brain tissue to chronically implanted neural electrodes. Journal of Neuroscience Methods 148(1):1–18.

Ramsey, N.F., M.P. van de Heuvel, K.H. Kho, and F.S. Leijten. 2006. Towards human BCI applications based on cognitive brain systems: An investigation of neural signals recorded from the dorsolateral prefrontal cortex. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2):214–217.

Rohde, M.M., S.L. BeMent, J.E. Huggins, S.P. Levine, R.K. Kushwaha, and L.A. Schuh. 2002. Quality estimation of subdurally recorded, event-related potentials based on signal-to-noise ratio. IEEE Transactions on Biomedical Engineering 49(1):31–40.

Sanchez, J.C., A. Gunduz, P.R. Carney, and J.C. Principe. 2008. Extraction and localization of mesoscopic motor control signals for human ECoG neuroprosthetics. Journal of Neuroscience Methods 167(1):63–81.

Schalk, G., E.C. Leuthardt, D. Moran, J. Ojemann, and J.R. Wolpaw. 2004a. Two-dimensional cursor control using electrocorticographic signals in humans. Proceedings of the Society for Neuroscience Annual Meeting, San Diego, Calif., October 23–27, 2004.

Schalk, G., D.J. McFarland, T. Hinterberger, N. Birbaumer, and J.R. Wolpaw. 2004b. BCI2000: A general-purpose brain-computer interface (BCI) system. IEEE Transactions on Biomedical Engineering 51(6):1034–1043.

Schalk, G., J. Kubánek, K.J. Miller, N.R. Anderson, E.C. Leuthardt, J.G. Ojemann, D. Limbrick, D. Moran, L.A. Gerhardt, and J.R. Wolpaw. 2007. Decoding two-dimensional movement trajectories using electrocorticographic signals in humans. Journal of Neural Engineering 4(3):264–275.

Schalk, G., K.J. Miller, N.R. Anderson, J.A. Wilson, M.D. Smyth, J.G. Ojemann, D.W. Moran, J.R. Wolpaw, and E.C. Leuthardt. 2008. Two-dimensional movement control using electrocorticographic signals in humans. Journal of Neuroscience Methods 5(1):75–84.

Serruya, M.D., N.G. Hatsopoulos, L. Paninski, M.R. Fellows, and J.P. Donoghue. 2002. Instant neural control of a movement signal. Nature 416(6877):141–142.

Seymour, J.P., and D.R. Kipke. 2007. Neural probe design for reduced tissue encapsulation in CNS. Biomaterials 28(25):3594–3607.

Spataro, L., J. Dilgen, S. Retterer, A.J. Spence, M. Isaacson, J.N. Turner, and W. Shain. 2005. Dexamethasone treatment reduces astroglia responses to inserted neuroprosthetic devices in rat neocortex. Experimental Neurology 194(2):289–300.

Srinivasan, R., P.L. Nunez, and R.B. Silberstein. 1998. Spatial filtering and neocortical dynamics: Estimates of EEG coherence. IEEE Transactions on Biomedical Engineering 45:814–826.

Sutter, E.E. 1992. The brain response interface: Communication through visually-induced electrical brain responses. Journal of Microcomputer Applications 15:31–45.

Szarowski, D.H., M.D. Andersen, S. Retterer, A.J. Spence, M. Isaacson, H.G. Craighead, J.N. Turner, and W. Shain. 2003. Brain responses to micro-machined silicon devices. Brain Research 983(1-2):23–35.

Taylor, D.M., S.I. Tillery, and A.B. Schwartz. 2002. Direct cortical control of 3D neuroprosthetic devices. Science 296(5574):1829–1832.

Vaughan, T.M., D.J. McFarland, G. Schalk, W.A. Sarnacki, D.J. Krusienski, E.W. Sellers, and J.R. Wolpaw. 2006. The Wadsworth BCI Research and Development Program: At home with BCI. IEEE Transactions on Neural Systems and Rehabilitation Engineering 14(2):229–233.

Velliste, M., S. Perel, M.C. Spalding, A.S. Whitford, and A.B. Schwartz. 2008. Cortical control of a prosthetic arm for self-feeding. Nature 453(7198):1098–1101.

Vidal, J.J. 1977. Real-time detection of brain events in EEG. IEEE Proceedings. Special Issue on Biological Signal Processing and Analysis 65:633–664.

Vossler, D., M. Doherty, R. Goodman, L. Hirsch, J. Young, and D. Kraemer. 2004. Early safety experience with a fully implanted intracranial responsive neurostimulator for epilepsy. Paper presented at the Annual Meeting of the American Epilepsy Society (AES). New Orleans, La., December 3–7, 2004.

Wang, W., S.S. Chan, D.A. Heldman, and D.W. Moran. 2007. Motor cortical representation of position and velocity during reaching. Journal of Neurophysiology 97(6):4258–4270.

Williams, J.C., J.A. Hippensteel, J. Dilgen, W. Shain, and D.R. Kipke. 2007. Complex impedance spectroscopy for monitoring tissue responses to inserted neural implants. Journal of Neural Engineering 4(4):410–423.

Wolpaw, J.R., and D.J. McFarland. 1994. Multichannel EEG-based brain-computer communication. Clinical Neurophysioogy 90(6):444–449.

Wolpaw, J.R., and D.J. McFarland. 2004. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. Proceedings of the National Academies of the USA 101(51):17849–17854.

Wolpaw, J.R., D.J. McFarland, G.W. Neat, and C.A. Forneris. 1991. An EEG-based brain-computer interface for cursor control. Electroencephalography and Clinical Neurophysiology 78(3):252–259.

Wolpaw, J.R., N. Birbaumer, W.J. Heetderks, D.J. McFarland, P.H. Peckham, G. Schalk, E. Donchin, L.A. Quatrano, C.J. Robinson, and T.M. Vaughan. 2000. Brain-computer interface technology: A review of the first international meeting. IEEE Transactions on Rehabilitation Engineering 8(2):164–173.

Wolpaw, J.R., N. Birbaumer, D.J. McFarland, G. Pfurtscheller, and T.M. Vaughan. 2002. Brain-computer interfaces for communication and control. Clinical Neurophysiology 113(6):767–791.

Yuen, T.G., W.F. Agnew, and L.A. Bullara. 1987. Tissue response to potential neuroprosthetic materials implanted subdurally. Biomaterials 8(2):138–141.

Ultra Low-Power Biomedical and Bio-Inspired Systems

Rahul Sarpeshkar
Massachusetts Institute of Technology

Biological systems are incredibly energy efficient and compute with unreliable and noisy components to perform reliable and precise computations. For example, the brain computes with a staggering energy efficiency of approximately 0.2 femtojoules (fJ) per floating-point operation, and the cell, which is even more energy efficient, uses only 20 kT of energy per active biomolecular operation, i.e., only 8×10^{-20} Joules or 20 units of thermal energy (Sarpeshkar, 2010). Similarly impressive numbers for the energy efficiency of the eye, ear, and organs of the body may be found in the literature (Sarpeshkar, 2010). The creation of circuits that are inspired by biology can lead to novel architectures and systems that have applications outside of biology. Such circuits can also be immediately applied to repair biological systems when they do not work (i.e., in neural prosthetics). This talk illustrates how the synergy between biological and electronic circuits has led to ultra-low-power and noise-robust systems for people who are deaf, blind, and paralyzed and to advanced ear-inspired radio receivers. It concludes with a discussion of *cytomorphic* or cell-inspired systems that establish an important bridge between electronics and chemistry (Sarpeshkar, 2010). Such systems lay a rigorous foundation for an analog-circuits approach to systems and synthetic biology, fields highly important in the future of medicine and biological engineering.

A rigorous comparison of the pros and cons of analog versus digital computation (Sarpeshkar, 1998) reveals that analog computation, which exploits freely available physical basis functions in the underlying technology that are not necessarily logical or linear to compute, is more energy efficient than digital computation at low precision and vice versa (Sarpeshkar, 1998). Biology exploits this insight to compute in a novel and highly energy-efficient collective analog or hybrid fashion that is not purely digital or purely analog but an intimate combination of both (Sarpeshkar, 2010). Engineering systems can take inspiration from biology to also compute in this fashion, and they can improve energy efficiency by delaying digitization after an optimal amount of analog preprocessing (Sarpeshkar, 2010).

One example of a bio-inspired collective analog system is the radio frequency (RF) cochlea (Mandal et al., 2009), an electronic chip that takes inspiration from the spectrum analysis of the inner ear or cochlea to create an energy-efficient and ultrafast broadband RF spectrum analyzer. This chip exploits the fact that the ear's spectrum-analysis architecture is the fastest and most hardware efficient known to man—faster than a digital fast Fourier transform or an analog filter bank. It efficiently maps the partial differential equations that describe fluid membrane-hair cell interaction in the biological cochlea at kilohertz audio frequencies to inductor-capacitor-amplifier interaction in the RF cochlea at gigahertz frequencies. The resulting broadband RF cochlea chip operates with 20-fold lower hardware cost than a traditional analog filter bank or with 100-fold lower power than a system that directly digitizes its RF input to perform spectrum analysis. The RF cochlea is useful as a front end in advanced cognitive or software radios of the future (Sarpeshkar, 2010).

The use of analog circuits to perform energy-efficient spectrum analysis is also useful in bionic ear or cochlear implant processors for people who are profoundly deaf. Cochlear-implant processors compress spectral information present in a microphone signal in a nonlinear fashion such that it is suitable for charge-balanced tonotopic current stimulation of a cochlear electrode array implanted near the auditory nerve. For example, a digitally programmable analog cochlear-implant processor described in the literature (Sarpeshkar et al., 2005) lowered power consumption by 20-fold over a conventional design that performs analog-to-digital conversion followed by digital signal processing; it enabled flexible 86-parameter programming in a patient who understood speech with it on her first try (Sarpeshkar, 2006); it was highly robust to several sources of noise including transistor mismatch, *1/f* or pink noise, power-supply noise, RF crosstalk, thermal noise, and temperature variations; and it is at or near the energy-efficient optimum even at the end of Moore's law. Thus, this processor is amenable to fully implanted and low-cost systems of the future: Its 251-μW power consumption enables it to function on a small 100-mAh battery with 1,000 wireless recharges for 30 years. A more advanced 357-μW bio-inspired asynchronous interleaved sampling cochlear-implant processor uses a novel bio-inspired method of nerve stimulation similar to that present in the auditory nerve. It enables fine-time encoding of phase information in a signal without requiring a high sampling rate (Sit and Sarpeshkar, 2008). Hence, it enables music information to be encoded in an energy-efficient fashion without requiring a high number of electrodes or requiring high-stimulation power consumption, a bottleneck in the field of cochlear implants. It is also important for improving speech understanding in noise. Similarly, a companding algorithm inspired by tone-to-tone suppression and gain control in the cochlea has led to improved speech performance in noise (Oxenham et al., 2007; Turicchia and Sarpeshkar, 2005).

Recent work has reported an ultra-energy-efficient adiabatic energy-recycling neural stimulator that can lower power dissipation of nerve stimulation in implants for people who are deaf, blind, or paralyzed, or in other neural, cardiac, or muscle-stimulation applications by a factor of at least 2 to 3 (Arfin and Sarpeshkar, 2011). Such work can be combined with state-of-the-art micropower neural amplifiers that operate near the fundamental limits of physics (Wattanapanitch et al., 2007), with (1) 1 nJ/bit near-field RF telemetry systems that enable transcutaneous wireless bidirectional data transmission in implants (Mandal and Sarpeshkar, 2008), (2) energy-efficient wireless recharging circuits that operate at the limits of physics set by coil quality factors (Baker and Sarpeshkar, 2007), (3) novel highly area and power efficient battery-recharging circuits (Do Valle et al., 2011), (4) highly energy-efficient bio-inspired processors for neural decoding (Rapoport and Sarpeshkar, 2010), (5) highly energy-efficient imagers and novel cochlear-implant-inspired image-processing algorithms (Turicchia et al., 2008), (6) bio-inspired analog vocal tracts for speech and hearing prostheses that perform well in noise (Wee et al., 2011), and (7) blocking-capacitor-free highly miniature precision neural stimulation (Sit and Sarpeshkar, 2007). The integration of several such ultra-low-power and bio-inspired innovations can enable ultra-low-power, low-cost, highly miniature, and fully implanted neural prosthetics for people who are deaf, blind, or paralyzed or for people who have other conditions (Sarpeshkar, 2010). Examples of complete working systems that successfully stop a bird from singing via wireless neural stimulation (Arfin et al., 2009), perform wireless recording from a monkey (Wattanapanitch and Sarpeshkar, 2011), and summarize system aspects of design (Sarpeshkar et al., 2008) are given in the literature. Sarpeshkar (2010) discusses the practical engineering constraints needed in such devices.

As these examples illustrate, analog and bio-inspired circuits have enabled and are continuing to enable noise-robust, highly miniature, and ultra-low-power operation in neural prosthetics, a necessity to reduce advanced research to practical clinical applications (Sarpeshkar, 2010). In fact, the deep and astounding mathematical similarities between a form of electronics termed *subthreshold electronics* and chemistry (Sarpeshkar, 2010) suggest that the impact of electronics on the future of medicine may not be confined to neural, cardiac (Turicchia et al., 2010), or muscular prosthetics but in fact could be much broader: It could encompass a whole new way of thinking about biological circuits–simulating them, designing them, and fixing them.

The average 10-μm cell is a marvel of nanotechnology, performing 1×10^7 energy-consuming biochemical operations per second in its stochastic, nonlinear, feedback 30,000-node gene-protein and protein-protein network with just 1 pW of power (Sarpeshkar, 2010). Efficient precise computation with noisy components is achieved via clever nonlinear, feedback, and hybrid analog-digital strategies in cells in biology (Hahnloser et al., 2000) as it is in the most advanced ultra-low-power analog electronic systems of today. Circuits in cell biology and circuits in electronics may be viewed as being highly similar, with biology using molecules, ions, proteins, and DNA rather than electrons and transistors. The striking mathematical similarities between chemical reaction dynamics and electronic current flow in the subthreshold regime of transistor operation, including the Boltzmann stochastics of current flow (Sarpeshkar, 2010), imply that one can mimic and model large-scale chemical-processing systems in biological and artificial networks very efficiently on an electronic chip at time scales that could potentially be a million times faster. This key idea has been built on to show how to create current-mode subthreshold transistor circuits for modeling arbitrary chemical reactions in protein-protein (Mandal and Sarpeshkar, 2009a) and gene-protein networks (Mandal and Sarpeshkar, 2009b; Sarpeshkar, 2010).

The latter work shows that we can potentially attempt to simulate cells, organs, and tissues with ultrafast highly parallel analog and hybrid analog-digital circuits, including molecular stochastics and cell-to-cell variability on large-scale "supercomputing" electronic chips. Such molecular dynamics simulations are extremely computationally intensive, especially when the effects of noise, nonlinearity, network-feedback effects, and cell-to-cell variability are included. Stochastics and cell-to-cell variabililty are highly important factors for predicting a cell's response to drug treatment (e.g., the response of tumor cells to chemotherapy treatments). In turn, analog circuit-design techniques can also be mapped to design and create synthetic-biology circuits that have been shown to be in accord with biological data (Danial et al., 2011). Thus, they can affect the treatment of gene therapies in diseases like cancer and diabetes, or affect the understanding of how such circuits malfunction, thus leading to better drug therapies.

The deep links between energy and information allow one to articulate information-based principles for ultra-low-power design that apply to biology or to electronics (Sarpeshkar, 2010). Engineering can aid biology through analysis, instrumentation, and repair (medicine). Biology can aid engineering through bio-inspired design. The positive-feedback loop created by this two-way interaction can amplify and speed progress in both disciplines and shed insight into both (Sarpeshkar, 2010).

REFERENCES

Arfin, S.K., and R. Sarpeshkar. 2011. An energy-efficient, adiabatic electrode stimulator with inductive energy recycling and feedback current regulation. IEEE Transactions on Biomedical Circuits and Systems. In press.

Arfin, S., M. Long, M. Fee, and R. Sarpeshkar. 2009. Wireless neural stimulation in freely behaving small animals. Journal of Neurophysiology 102(1):598–605.

Baker, M.W., and R. Sarpeshkar. 2007. Feedback analysis and design of RF power links for low-power bionic systems. IEEE Transactions of Biomedical Circuits and Systems 1(1):28–38.

Danial, R., S.S. Woo, L. Turicchia, and R. Sarpeshkar. 2011. Analog transistor models of bacterial genetic circuits. Paper presented at the 2011 IEEE Biological Circuits and Systems (BioCAS) Conference, San Diego, Calif., November 10–12, 2011.

Do Valle, B., C. Wentz, and R. Sarpeshkar. 2011. An area and power-efficient analog Li-ion battery charger circuit. IEEE Transactions on Biomedical Circuits and Systems 5(2):131–137.

Hahnloser, R., R. Sarpeshkar, M. Mahowald, R. Douglas, and S. Seung. 2000. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature 405(6789):947–951.

Mandal, S., and R. Sarpeshkar. 2008. Power-efficient impedance-modulation wireless data links for biomedical implants. IEEE Transactions on Biomedical Circuits and Systems 2(4):301–315.

Mandal, S., and R. Sarpeshkar. 2009a. Log domain circuit models of chemical reactions. Proceedings of the 2009 IEEE International Symposium on Circuits and Systems (ISCAS 2009), Taipei, Taiwan, May, pp. 2697–2700.

Mandal, S., and R. Sarpeshkar. 2009b. Circuit models of stochastic genetic networks. 2009 IEEE Symposium on Biological Circuits and Systems (BioCAS), Beijing, November, pp. 109–112.

Mandal, S., S. Zhak, and R. Sarpeshkar. 2009. A bio-inspired active radio-frequency silicon cochlea. IEEE Journal of Solid-State Circuits 44(6):1814–1828.

Oxenham, A.J., A.M. Simonson, L. Turicchia, and R. Sarpeshkar. 2007. Evaluation of companding-based spectral enhancement using simulated cochlear-implant processing. Journal of the Acoustical Society of America 121(3):1709–1716.

Rapoport, B.I., and R. Sarpeshkar. 2010. A biomimetic adaptive algorithm and micropower circuit architecture for implantable neural decoders. In System and Circuit Design for Biologically-Inspired Learning, edited by T. Temel. IGI Global. 38 pp.

Sarpeshkar, R. 1998. Analog versus digital: Extrapolating from electronics to neurobiology. Neural Computation 10:1601–1638.

Sarpeshkar, R. 2006. Brian power: Borrowing from biology makes for low-power computing. IEEE Spectrum 43(5):24–29.

Sarpeshkar, R. 2010. Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications,and Bio-Inspired Systems. Cambridge, U.K.: Cambridge University Press.

Sarpeshkar, R., C. Salthouse, J.J. Sit, M. Baker, S. Zhak, T. Lu, L. Turicchia, and S. Balster. 2005. An ultra-low-power programmable analog bionic ear processor. IEEE Transactions on Biomedical Engineering 52(4):711–727.

Sarpeshkar, R., W. Wattanapanitch, S.K. Arfin, B.I. Rapoport, S. Mandal, M. Baker, M. Fee, S. Musallam, and R.A. Andersen. 2008. Low-power circuits for brain-machine interfaces. IEEE Transactions on Biomedical Circuits and Systems 2(3):173–183.

Sit, J.-J., and R. Sarpeshkar. 2007. A low-power, blocking-capacitor-free, charge-balanced electrode-stimulator chip with less than 6nA DC error for 1mA full-scale stimulation. IEEE Transactions on Biomedical Circuits and Systems 1(3):172–183.

Sit, J.-J., and R. Sarpeshkar. 2008. A cochlear-implant processor for encoding music and lowering stimulation power. IEEE Pervasive Computing, Special Issue on Implantable Systems 1(7):40–48.

Turicchia, L., and R. Sarpeshkar. 2005. A bio-inspired companding strategy for spectral enhancement. IEEE Transactions on Speech and Audio Processing 13(2):243–253.

Turicchia, L., M. O'Halloran, D.P. Kumar, and R. Sarpeshkar. 2008. A low-power imager and compression algorithms for a brain-machine visual prosthesis for the blind. Proceedings of the SPIE, Vol. 7035, San Diego, Calif., August 10–14, pp. 703510-1–703510-13.

Turicchia, L., B. do Valle, J. Bohorquez, W. Sanchez, V. Misra, L. Fay, M. Tavakoli, and R. Sarpeshkar. 2010. Ultra low power electronics for cardiac monitoring. IEEE Transactions on Circuits and Systems 57(9):2279–2290.

Wattanapanitch, W., and R. Sarpeshkar. 2011. A low-power 32-channel digitally programmable neural recording integrated circuit. IEEE Transactions on Biomedical Circuits and Systems 5(6):592–602.

Wattanapanitch, W., M. Fee, and R. Sarpeshkar. 2007. An energy-efficient micropower neural recording amplifier. IEEE Transaction on Biomedical Circuits and Systems 1(2):136–147.
Wee, K.H., L. Turicchia, and R. Sarpeshkar. 2011. An articulatory silicon vocal tract for speech and hearing prostheses. IEEE Transactions on Biomedical Circuits and Systems 5(4):339–346.

Appendixes

Contributors

Samy Bengio is a research scientist at Google where his interests are centered on large-scale machine learning, in particular for ranking tasks, with applications to image processing, music, text processing, and combinations of these.

Michael Cafarella is an assistant professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. His research interests include databases, information extraction, and data mining, and he is particularly interested in applying data mining techniques to data management problems encountered on the Web and in the natural social sciences.

Andrew Christensen is president of Medical Modeling, Inc. in Golden, Colorado. His work is in the application of medical imaging technologies to patient treatment and in additive manufacturing of medical implants.

Timothy Denison is a technical fellow at Medtronic and director of neural engineering where he helps oversee the design of next-generation sensor, algorithm, and actuation technologies for the treatment of chronic neurological disease.

Evgeniy Gabrilovich is a senior research scientist at Yahoo! Research where he focuses on information retrieval, machine learning, computational linguistics, and enriching artificial intelligence applications with human knowledge.

John Haymaker is founder of Design Process Innovation in San Francisco. He investigates formal models for design and construction processes; builds computational methods to help engineers better communicate, integrate, and optimize these processes; and looks for opportunities to apply these enhanced processes in industry to improve sustainability.

Aleksandar Kuzmanovic is Lisa Wissner Slivka and Benjamin Slivka Chair in Computer Science and assistant professor in the Department of Electrical Engineering and Computer Science at Northwestern University. His work is in computer networking and distributed systems with emphasis on design, measurements, analysis, denial-of-service resiliency, and prototype implementation of protocols, algorithms, and large-scale systems for the Internet.

Eric Leuthardt is a neurosurgeon and assistant professor in the Departments of Neurological Surgery and Biomedical Engineering at the Washington University School of Medicine. His research focuses on neuroprosthetics—devices linked to the brain that may restore neurologic function. He studies the use of surface cortical electrophysiology, known as electrocorticography, as a signal platform for brain computer interfaces. He integrates multiple domains of expertise ranging from engineering to neurosurgery to complex signal analysis.

Hod Lipson is an associate professor in the School of Mechanical and Aerospace Engineering at Cornell University. His research focuses on understanding the synthesis and function of complex engineering systems and their relationships to similarly complex systems in biology. He has been developing methods for automating the synthesis and analysis of a variety of systems across engineering and biological domains using principles inspired from biological evolution and co-evolution.

Brett Lyons is a materials and process engineer at Boeing Research and Technology. His area of focus is material and process development for additive manufacturing and polymer composites.

Ani Nenkova is an assistant professor in the Department of Computer and Information Science at the University of Pennsylvania. She works on natural language processing where her main interests are in summarization, evaluation, discourse processing, and automatic prediction on text quality.

John Ochsendorf is an associate professor in the Departments of Architecture and Civil and Environmental Engineering at the Massachusetts Institute of Technology. His research focuses on the structural assessment of existing buildings and on the sustainable design of new structures. His research group has particular expertise in the static and dynamic assessment of historical masonry buildings and also has participated in the engineering design of new low-carbon buildings.

Annie Pearce is an associate professor in the Myers-Lawson School of Construction at Virginia Tech. Her areas of interest include metrics of sustainability for built facilities, green building materials and systems, cost modeling to support sustainability implementation, and in situ performance of sustainable facility technologies.

Christopher Pyke is vice president of research at the U.S. Green Building Council in Washington, D.C. He directs a diverse research portfolio that includes next-generation green building rating systems, the assessment of building performance and occupant experience, and the study of market trends and dynamics. He has a particular interest in applications of advanced information technologies to accelerate market transformation and the diffusion of innovation.

Rahul Sarpeshkar is an associate professor in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology. His research areas are bioelectronics—biomedical and bio-inspired electronics (electronics inspired by cell biology and neurobiology); systems biology, synthetic biology, and analog circuit design of molecular and cellular circuits; ultra low-power, ultra miniature, and ultra energy-efficient circuits and systems; medical implants for the deaf, blind, paralyzed, and cardiac-impaired; and brain-machine interfaces.

Carolyn Seepersad is an assistant professor in the Department of Mechanical Engineering at the University of Texas at Austin. Much of her work focuses on design for additive and freeform manufacturing with an emphasis on products with customized meso- and micro-structures. She also conducts research in design innovation, sustainable design, and systems design optimization.

Michael Siemer is president of Mydea Technologies in Orlando, Florida, which focuses on improving the utilization of rapid prototyping, 3D printing, and additive manufacturing technologies for low-volume production. Specific areas include rapid tooling, mass customization via parametric 3D modeling, and direct digital manufacturing.

Alfred Spector is vice president of research and special initiatives at Google. Prior to that, he was vice president of strategy and technology and chief technical officer of IBM's software business and vice president of services and software at IBM Research. He also was founder and CEO of Transarc Corporation, a pioneer in distributed transaction processing and wide area file systems, and an associate professor of computer science at Carnegie Mellon University (CMU). While at CMU he did fundamental work in a number of areas, including the Andrew File System that changed the face of distributed computing.

Jelena Srebric is a professor in the Department of Architectural Engineering at Pennsylvania State University. Her research seeks to understand physical transport processes in and around buildings and their modeling, including simulations and validation approaches. These efforts enable discoveries in novel building systems and engineering of urban settlements.

Brent Stucker is a professor in the Department of Industrial Engineering at the University of Louisville. His area of research is additive manufacturing technologies and their applications, including advanced materials development and multifunctional structures.

Amarnag Subramanya is a research scientist at Google Research. His expertise is in the area of machine learning and natural language processing, and his interests include semi-supervised learning, speech signal processing, multi-sensory fusion, and probabilistic models. Currently he is focusing on the applications of machine learning methods to solve problems in natural language processing.

James Weiland is an associate professor in the Department of Ophthalmology at the Doheny Eye Institute at the University of Southern California. He is interested in technology to assist the blind, and the main focus of his research is an implantable retinal prosthesis. His research group investigates the interface between the retina and the implantable stimulator in order to optimize the visual abilities of patients. They are also developing wearable computer vision systems for aiding blind people.

Justin Williams is an associate professor in the Departments of Biomedical Engineering and Neurological Surgery at the University of Wisconsin-Madison. His research is focused on developing new technology for interfacing with the nervous system to study and treat neurological disease. He applies microtechnology to study the basic processes by which neurons and glial cells interact with their micro-environment in order to develop new implantable devices for extracting information from the nervous system to help treat disabilities.

Zhiqiang (John) Zhai is an associate professor in the Department of Civil, Environmental, and Architectural Engineering at the University of Colorado, Boulder. His research and teaching interests are in integrated building systems, sustainable building technologies, and indoor environmental quality.

Program

NATIONAL ACADEMY OF ENGINEERING

2011 U.S. Frontiers of Engineering Symposium
September 19–21, 2011

Chair: Andrew M. Weiner, Purdue University

ADDITIVE MANUFACTURING

Organizers: Carolyn Seepersad and Michael Siemer

Additive Manufacturing Technologies:
Technology Introduction and Business Implications
Brent Stucker

Additive Manufacturing in Aerospace: Examples and Research Outlook
Brett Lyons

Additive Manufacturing is Changing Surgery
Andrew Christensen

The Shape of Things to Come: Frontiers in Additive Manufacturing
Hod Lipson

***

SEMANTIC PROCESSING

Organizers: Aleksandar Kuzmanovic and Amarnag Subramanya

Automatic Text Understanding of Content and Text Quality
Ani Nenkova

Advancing Natural Language Understanding with Collaboratively Generated Content
Evgeniy Gabrilovich

Large-Scale Visual Semantic Extraction
Samy Bengio

Searching for Statistical Diagrams
Michael Cafarella

***

ENGINEERING SUSTAINABLE BUILDINGS

Organizers: Annie Pearce and John Zhai

Challenges and Opportunities for Low-Carbon Buildings
John Ochsendorf

Expanding Design Spaces
John Haymaker

Opportunities and Challenges for Multiscale Modeling of Sustainable Buildings
Jelena Srebric

Accelerating Green Building Market Transformation with Information Technology
Christopher Pyke

***

NEUROPROSTHETICS

Organizers: Timothy Denison and Justin Williams

Retinal Prosthetic Systems for Treatment of Blindness
James Weiland

The Evolution of Brain-Computer Interfaces
Eric Leuthardt

Ultra Low-Power Biomedical and Bio-Inspired Systems
Rahul Sarpeshkar

***

DINNER SPEECH

The Evolution of Computer Science
Alfred Spector

Participants

NATIONAL ACADEMY OF ENGINEERING

2011 U.S. Frontiers of Engineering Symposium
September 19–21, 2011

Maneesh Agrawala
Associate Professor
Department of Electrical Engineering and Computer Science
University of California, Berkeley

Kristi Anseth
Distinguished Professor, Professor of Surgery, and HHMI Assistant Investigator
Department of Chemical and Biological Engineering
University of Colorado Boulder

Shorya Awtar
Assistant Professor
Department of Mechanical Engineering
University of Michigan

Timothy Barckholtz
Laboratory Director
Petroleum Sciences Lab
ExxonMobil

Robert Barsotti
Research Scientist
Altuglas International Research and Development
Arkema Inc.

Jennifer Benaman
Principal Engineer
Environmental Sciences
Anchor QEA, LLC

Samy Bengio**
Research Scientist
Google Research
Google

*Organizing Committee
**Speaker

Christopher Bettinger
Assistant Professor
Departments of Biomedical Engineering and Materials Science and Engineering
Carnegie Mellon University

Chad Bouton
Research Leader
Advanced Engineering for Life Sciences
Battelle Memorial Institute

Shawn Brueggemeier
Senior Research Investigator II
Late Phase Chemical Development
Bristol-Myers Squibb Company

Karen Burg
Hunter Endowed Chair and Professor
Department of Bioengineering
Clemson University

Michael Cafarella**
Assistant Professor
Department of Electrical Engineering and Computer Science
University of Michigan

Luca Carloni
Associate Professor
Department of Computer Science
Columbia University

Ioannis Chasiotis
Associate Professor
Department of Aerospace Engineering
University of Illinois at Urbana-Champaign

Kevin Chen
Renewable Energy Project Manager
Google Research
Google

Li-Te Cheng
Research Staff Member
Collaborative User Experience Group
IBM

Andrew Christensen**
President
Medical Modeling Inc.

Whitney Colella
Senior Research Engineer
Energy and Environment Directorate
Pacific Northwest National Laboratory

Neiyer Correal
Distinguished Member of the Technical Staff
Chief Technology Office
Motorola Solutions

Todd Danko
Lead Research Scientist
Intelligent Robotics Laboratory
Lockheed Martin Advanced Technology Laboratories

Virginia Davis
Associate Professor
Department of Chemical Engineering
Auburn University

Michael Degner
Senior Technical Leader, Electric Machine Drives
Electrification Research and Advanced Engineering
Ford Motor Company

Tao Deng
Senior Scientist
NanoStructures and Surfaces Laboratory
GE Global Research

Tim Denison*
Director of Neuroengineering
Neuromodulation
Medtronic

Jaydev Desai
Associate Professor
Department of Mechanical Engineering
University of Maryland

Douglas Eck
Research Scientist, Music and Machine Learning
Google Research
Google

John Ehrgott
Research Civil Engineer
Explosion Effects Branch/ Geotechnical and Structures Laboratory
U.S. Army Engineer Research and Development Center

Bryan Eisenhower
Associate Director
Center for Energy Efficient Design
Department of Mechanical Engineering
University of California, Santa Barbara

Audrey Ellerbee
Assistant Professor
Department of Electrical Engineering
Stanford University

Nick Feamster
Associate Professor
School of Computer Science
Georgia Institute of Technology

Adam Fontecchio
Associate Professor
Department of Electrical and Computer Engineering
Drexel University

Evgeniy Gabrilovich**
Senior Research Scientist
Yahoo! Research
Yahoo!

Sam Geckler
Technical Advisor
Advanced Systems Integration, Research and Technology
Cummins Inc.

Claudia Gunsch
Assistant Professor
Department of Civil and Environmental Engineering
Duke University

Craig Hamer
Division Scientist
TPDC Manufacturing Technology
3M

Jeffrey Hansbro
Global Applications Technology Leader, Building Envelope
Dow Construction Chemicals
Dow Chemical Company

Alexander Harkness
Fellow Engineer
New Plants Engineering
Westinghouse Electric Company

John Haymaker**
Founder
Design Process Innovation

Kim Hazelwood
Associate Professor
Department of Computer Science
University of Virginia

Jamie Hestekin
Associate Professor
Ralph E. Martin Department of Chemical Engineering
University of Arkansas

Robert Hopkins
Principal Hardware Engineer
VLSI Research
Oracle Labs

Jonathan Hui
Lead Engineer
Smart Grid, Connected Energy Group
Cisco Systems

Masaya Iwamoto
Compound Semiconductor Device R&D Engineer
High Frequency Technology Center
Agilent Technologies, Inc.

Madhusudan Iyengar
Senior Engineer/Master Inventor
Advanced Thermal Laboratory, Systems and Technology Group
IBM

Alan Jacobsen
Senior Scientist
Sensors and Materials Lab
HRL Laboratories, LLC

Anurag Jain
Principal
Applied Sciences and Investigations
Weidlinger Associates, Inc.

Dan Jiao
Associate Professor
School of Electrical and Computer Engineering
Purdue University

John Johnson
Civil Engineering Group Leader, Senior Associate
Transportation and Infrastructure Division
STV Inc.

Steven Johnson
Senior Signal Processing Engineer
Intelligence Systems
SAIC

Nirmal Keshava
Principal Member Technical Staff
Fusion, Exploitation, and Inference Technologies
Draper Laboratory

Ali Khademhosseini
Associate Professor
Harvard-MIT Division of Health Sciences and Technology
Harvard University

Tracy Kijewski-Correa
Leo E. and Patti Ruth Linbeck Associate Professor and Associate Department Chair
Department of Civil Engineering and Geological Sciences
University of Notre Dame

Nitin Kolhapure
Biofuels Technology Lead
Industrial Biosciences
DuPont Company

Dana Komin
Director, Global Automation Strategy and Execution
Global Manufacturing Assembly and Automation Center
General Motors

Elisa Konofagou
Associate Professor
Department of Biomedical Engineering
Columbia University

Jeffrey Koplow
Principal Member of Technical Staff
Thermal/Fluid Science and Engineering
Sandia National Laboratories

Katherine Kuchenbecker
Skirkanich Assistant Professor of Innovation
Department of Mechanical Engineering and Applied Mechanics
University of Pennsylvania

Aleksandar Kuzmanovic*
Lisa Wissner Slivka and Benjamin Slivka Chair in Computer Science and Assistant Professor
Department of Electrical Engineering and Computer Science
Northwestern University

Eric Leuthardt**
Neurosurgeon and Assistant Professor
Departments of Neurological Surgery and Biomedical Engineering
Washington University School of Medicine

Hod Lipson**
Associate Professor
School of Mechanical and Aerospace Engineering
Cornell University

Mitchell Loescher
Senior Principal Engineer Lead
Conceptual Engineering—R&D
ConocoPhillips

Christopher Loose
Chief Technology Officer and Director of Research
Semprus BioSciences

Lonnie Love
Group Leader
Robotics and Energetic Systems Group
Oak Ridge National Laboratory

Brett Lyons**
Materials and Process Engineer
Boeing Research and Technology
The Boeing Company

Michael McAlpine
Assistant Professor
Department of Mechanical and Aerospace Engineering
Princeton University

Kenneth McGuire
Principal Scientist
Package Development
Procter & Gamble

Mark Meyer
Graphics Software Engineer
Research
Pixar Animation Studios

Amit Misra
Co-Director, Energy Frontier Research Center
Materials Physics and Applications Division
Los Alamos National Laboratory

Kristi Morgansen
Associate Professor
Department of Aeronautics and Astronautics
University of Washington

Steven Murray
Principal Engineer
Electrical and Semiconductor Practice
Exponent, Inc.

Sanjeev Naik
Engineering Group Manager
Advanced Engine Controls
General Motors

Ani Nenkova**
Assistant Professor
Department of Computer and Information Science
University of Pennsylvania

John Ochsendorf**
Associate Professor
Departments of Architecture and Civil and Environmental Engineering
Massachusetts Institute of Technology

Aydogan Ozcan
Associate Professor
Department of Electrical Engineering
University of California, Los Angeles

Cesar Palerm
Principal Scientist
Diabetes Reliability and Manufacturability
Medtronic

Satinderpall Pannu
Section Leader
Center for Micro- and Nano-Technology Engineering
Lawrence Livermore National Laboratory

Annie Pearce*
Associate Professor
Myers-Lawson School of Construction
Virginia Tech

Suzanne Plummer
Fellow, Director of Design Engineering
Low-Power Microprocessor Cores
AMD

Maurizio Porfiri
Associate Professor
Department of Mechanical and Aerospace Engineering
Polytechnic Institute of NYU

Michelle Povinelli
Assistant Professor
Ming Hsieh Department of Electrical Engineering
University of Southern California

Christopher Pyke**
Vice President, Research
U.S. Green Building Council

Ramaswamy Ramaswamy
Principal Research Chemical Engineer
Eastman Research
Eastman Chemical Company

Arnaud Robert
Senior Vice President
Technology
Walt Disney Studios

John Sample
Geospatial Computing Section Head
Geospatial Sciences and Technology
Branch
Naval Research Laboratory

Rahul Sarpeshkar**
Associate Professor
Department of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology

Kevin Schlichting
Staff Engineer, Materials/Structures
Materials and Process Engineering
Pratt & Whitney

Carolyn Seepersad*
Assistant Professor
Department of Mechanical Engineering
University of Texas at Austin

Nicoleta Serban
Assistant Professor
School of Industrial and Systems
Engineering
Georgia Institute of Technology

M. Scott Shell
Assistant Professor
Department of Chemical Engineering
University of California, Santa Barbara

David Sholl
Michael Tennenbaum Family Chair
and GRA Eminent Scholar
School of Chemical and Biomolecular
Engineering
Georgia Institute of Technology

Rishi Shukla
Global Technology Leader
James R. Randall Research Center
Archer Daniels Midland

Michael Siemer*
President
Mydea Technologies

Armin Sorooshian
Assistant Professor
Department of Chemical and
Environmental Engineering
University of Arizona

Jelena Srebric**
Professor
Department of Architectural
Engineering
Pennsylvania State University

Roman Stocker
Associate Professor
Department of Civil and
Environmental Engineering
Massachusetts Institute of Technology

Karin Strauss
Researcher
Microsoft Research
Microsoft

Brent Stucker**
Professor
Department of Industrial Engineering
University of Louisville

Amarnag Subramanya*
Research Scientist
Google Research
Google

Andrea Sudik
Research Engineer
Fuel Cell and Electric Vehicle Research
Ford Research and Advanced Engineering

Yu An Sun
Technical Manager
Xerox Innovation Group
Xerox Research Center-Webster

Siddharth Suryanarayanan
Assistant Professor
Department of Electrical and Computer Engineering
Colorado State University

Kathleen Vaeth
Senior Principal Scientist
Fluidics Department
Kodak Research Labs
Eastman Kodak Company

Richard Vaia
Technology Advisor, NanoStructured and Biological Materials
Materials and Manufacturing Directorate
Air Force Research Laboratory

Cyrus Wadia
Assistant Director, Clean Energy and Materials R&D
Office of Science and Technology Policy
Executive Office of the President

Haiyan Wang
Associate Professor
Department of Electrical and Computer Engineering
Texas A&M University

James Weiland**
Associate Professor
Department of Ophthalmology
Doheny Eye Institute
University of Southern California

Andrew Weiner*
Scifres Family Distinguished Professor
School of Electrical and Computer Engineering
Purdue University

Brad White
Detection Solutions Team Lead
Chemistry Technology, Automation Engineering Team
Monsanto Company

Justin Williams*
Associate Professor
Departments of Biomedical Engineering and Neurological Surgery
University of Wisconsin-Madison

Oliver Williams
Researcher
Research Department
Microsoft

Zhiqiang (John) Zhai*
Associate Professor
Department of Civil, Environmental, and Architectural Engineering
University of Colorado Boulder

Lisa Zhang
Member of Technical Staff
Enabling Computing Technologies
Bell Labs Research, Alcatel-Lucent

Xiaojing (John X.J.) Zhang
Associate Professor
Department of Biomedical Engineering
University of Texas at Austin

Dinner Speaker

Alfred Spector
Vice President of Research and Special Initiatives
Google

Guests

Laura Adolfie
Director, STEM Development Office
Office of Assistant Secretary of Defense (Research and Engineering)

William Hayden
Vice President
The Grainger Foundation

Tekla Perry
Senior Editor
IEEE Spectrum Magazine

Sapun Parekh
AAAS Fellow
Division of Chemical, Bioengineering, Environmental, & Transport Systems
National Science Foundation

Sohi Rastegar
Division Director
Office of Emerging Frontiers in Research and Innovation
Engineering Directorate
National Science Foundation

Robert H. Wagoner
George R. Smith Chair and Professor
Department of Materials Science and Engineering
Ohio State University

Google

Urs Hoelzle
Senior Vice President, Technical Infrastructure

National Academy of Engineering

Charles M. Vest
President

Lance A. Davis
Executive Officer

Janet Hunziker
Senior Program Officer

Elizabeth Weitzmann
Program Associate