DIVERSE DISCIPLINES, ONE COMMUNITY

BiomedicalComputation REVIEW

Published by Simbios, an NIH National Center for Biomedical Computing

Big Data Analytics in Biomedical Research

PLUS:

Privacy and Biomedical Research: Building a Trust Infrastructure

INSIDE:

Tapping the Brain: Decoding fMRI

Personalized Cancer Treatment

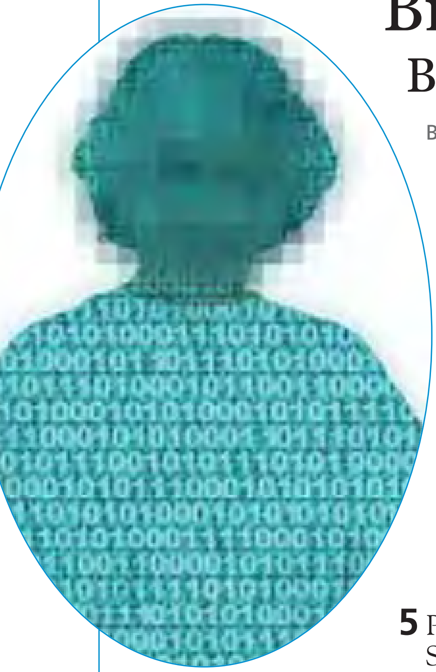
Leveraging Social Media For Biomedical Research

And more

Winter 2011/2012

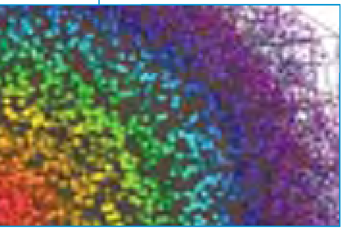
contents

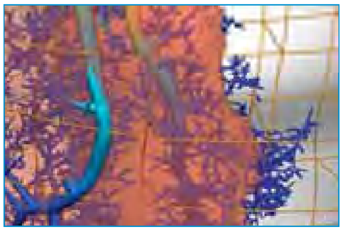
ContentsWinter 2011/2012



DEPARTMENTS

Cover Art: Digital profile image © Vectomart | Dreamstime.com. **Page 14:** Created by Rachel Jones of Wink Design Studio using: digital profile image © Vectomart | Dreamstime.com, and word clouds created on http://www.tagxedo.com. **Page 17, 19, 20:** word clouds created on http://www.tagxedo.com. **Page 22:** Created by Rachel Jones of Wink Design Studio using: in-house images and patient image © Wavebreakmedia Ltd | Dreamstime.com.





Winter 2011/2012
Volume 8, Issue 1
ISSN 1557-3192

Executive Editor Russ Altman, MD, PhD

Advisory Editor David Paik, PhD

Associate Editor Joy Ku, PhD

Managing Editor Katharine Miller

Science Writers
Alexander Gelfand, Louisa Dalton Hudock, Katharine Miller, Kristin Sainani, PhD

Community Contributors
Joy Ku, PhD, Oren Freifeld, Justin Foster, Paul Nuyujukian, Ton van den Bogert, PhD

Layout and Design
Wink Design Studio

Printing
Advanced Printing

Editorial Advisory Board
Russ Altman, MD, PhD, Brian Athey, PhD, Dr. Andrea Califano, Valerie Daggett, PhD, Scott Delp, PhD, Eric Jakobsson, PhD, Ron Kikinis, MD, Isaac Kohane, MD, PhD, Mark Musen, MD, PhD, Tamar Schlick, PhD, Jeanette Schmidt, PhD, Michael Sherman Arthur Toga, PhD, Shoshana Wodak, PhD, John C. Wooley, PhD

For general inquiries, subscriptions, or letters to the editor, visit our website at www.biomedicalcomputationreview.org

Office
Biomedical Computation Review
Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

Biomedical Computation Review
is published by:

Simbios
The NIH National
Center for Physics-
Based Simulation of
Biological Structures

Publication is made possible through the NIH Roadmap for Medical Research Grant U54 GM072970. Information on the National Centers for Biomedical Computing can be obtained from http://nihroadmap.nih.gov/bioinformatics. The NIH program and science officers for Simbios are:

Peter Lyster, PhD (NIGMS)
Grace Peng, PhD (NIBIB)
Jim Gnadt, PhD (NINDS)
Peter Highnam, PhD (NCRR)
Jennie Larkin, PhD (NHLBI)
Jerry Li, MD, PhD (NIGMS)
Nancy Shinowara , PhD (NICHD)
David Thomassen, PhD (DOE)
Janna Wehrle, PhD (NIGMS)
Jane Ye, PhD (NLM)

BY TON VAN DEN BOGERT, PhD

Slaying Villains Outside the Ivory Tower

Just over a year ago, I left academia. I had been in that realm for 25 years, working in musculoskeletal biomechanics and human movement analysis. It was a move that might have surprised anyone who knew me as a child, as well as many who watched me through my career. Yet the transition has been smooth and rewarding, and harbors hints of a lesson for others.

The story begins at age eight with my stated intent to become a professor. What I really meant was the stereotypical absent-minded professor of European comic books: I dreamed of inventing cool gadgets to help defeat the villains.

Eventually I became a real professor. I enjoyed the mix of research and teaching, and didn't mind the additional reality of "publish or perish." But as this evolved into "get more grants or perish," I began to wonder whether input or output was more important to our academic institutions. And when would I get to defeat the villains anyway?

I became more interested in an alternative lifestyle after doing some private consulting. A rather typical scenario would proceed as follows: The client explains a need; I propose a solution and write a two-page proposal with a budget; a contract is signed; work begins; and in the next year, my software is used in a product that disrupts the industry.* This was tremendously refreshing, compared to waiting nine months for a (possibly negative) decision on a grant proposal. And it felt much more like my dream of inventing cool gadgets. But small disruptive projects do not fit well within the current business model of academic research, where we are encouraged to have large teams, large grants with high indirect costs, and graduate students who do the work over five years.

So when I left academia to start a small business in computational biomechanics, the decision was easier than you might expect. Many things wouldn't change: I could still get NIH funding (as a small business); teach graduate students (as an adjunct); and publish (if I avoided contracts with unreasonable restrictions and set my fees high enough to have time to write). And I figured I could out-compete others in this niche (mostly graduate students and postdocs such as those I used to train) while remaining competitive through continuous improvement of my capabilities.

In biomedical computing, it's actually possible for an individual or small business to be at the cutting edge of research. To some extent, that's because they can collaborate with academic investigators, as I've continued to do (mostly through the Simtk.org collaborative platform).

In biomedical computing, it's actually possible for an individual or small business to be at the cutting edge of research. To some extent, that's because they can collaborate with academic investigators, as I've continued to do (mostly through the Simtk.org collaborative platform). Sometimes academics have expertise or laboratory facilities that I do not have, or they need my expertise. Funding for such collaborations is available, and I believe they provide an increasingly viable model for biomedical research, resulting in high-quality work and a win-win situation for all parties. Most importantly, the independent "comic book" scientist can enjoy the freedom to quickly and efficiently invent cool gadgets and slay a few villains. □

* In 2005, I received a Technical Achievement Award from the Academy of Motion Picture Arts and Sciences for work with Motion Analysis Corp. (Santa Rosa, CA). This was a consulting project that resulted in software to generate realistic 3-D human animations from motion capture data.

DETAILS

Antonie (Ton) J. van den Bogert is President of Orchard Kinetics, a startup company doing computational musculoskeletal biomechanics R&D, based in Cleveland, Ohio. Previous positions were at the University of Calgary and the Cleveland Clinic Lerner Research Institute. Ton is well-known as the moderator of Biomch-L, a mailing list and social network for biomechanists, which has existed since 1988 and now has more than 10,000 members. He co-founded the Technical Group on Computer Simulation of the International Society of Biomechanics (ISB) in 1989, and is currently the president of the ISB. As a pioneer in musculoskeletal modeling and simulation, he has held a collaborating R01 grant with Simbios, and currently serves on the Scientific Advisory Board of Simbios.

BY JOY P. KU, PhD, DIRECTOR OF SIMBIOS

Grand Challenge Competition Provides Rich Data Set to Improve Joint Contact Force Predictions

There are numerous modeling methods available to make predictions of muscle and joint contact forces. While such predictions can help improve treatments for movement-related disorders such as stroke or osteoarthritis, there's a problem: choosing which one to use. The research community hasn't had a way to easily evaluate different approaches. The Grand Challenge Competition to Predict *In Vivo* Knee Loads—now in its third year—changes that by motivating researchers to critically evaluate their simulations of contact forces in the knee, and by providing a wealth of experimental data to do so.

"Through the competition, I hope that we as a research community will be a little more critical of ourselves—in a good way—so that we can really gain confidence in what we're predicting and our models can become much more useful clinically," says **B.J. Fregly**, **PhD**, professor of mechanical and aerospace engineering at the University of Florida. He and **Darryl D'Lima**, **MD**, **PhD**, at Scripps Clinic are principal investigators of the grand challenge grant.

Each fall, the competition organizers make available a comprehensive data set for one subject who has received an instrumented, force-measuring knee implant. The data includes pre-surgery MRI and CT data, surface marker trajectory data, electromyography signals, and dynamic x-ray images of knee motion, just to list a few. But the competition's participants are not given the *in vivo* contact force measurements. These they must predict and submit to the American Society of Mechanical Engineers (ASME) and the competition organizers. Presentations and discussions of the competition submissions and the announcement of the winner occur at ASME's Summer Bioengineering Meeting.

While other research areas have held similar competitions, "this grand challenge competition is the only one I'm aware of in biomechanics," says **Grace Peng**, **PhD**, a program director at the NIH's National Institute of Biomedical Imaging and Bioengineering, which funds the knee grand challenge grant. "In that way, it's really unique."

The government has recently been pushing the use of prizes and challenges, especially for science and technology. "Challenges like this help benchmark some of the many techniques available and determine best practices," says Peng. "It is helpful to discuss the methods openly and publicly and have the technology converge."

That's what **Stephen Piazza**, **PhD**, associate professor of kinesiology at Pennsylvania State University, and his then-graduate student **Michael Hast**, **PhD**, discovered when they participated in (and won) last year's competition. "When multiple groups are working on the same problem, the same data, there's a potential for sharing and learning that can't occur otherwise," says Piazza.

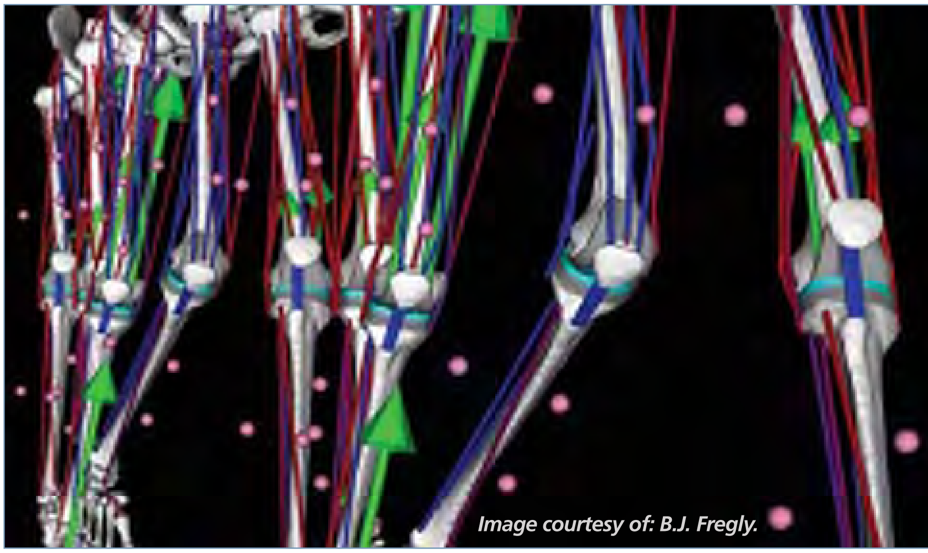


Image courtesy of: B.J. Fregly.

Hast agrees: "Sometimes it feels like you're in a cave working on these problems. Being able to see how others approached the same problem was a good experience."

Piazza and Hast actually waited one year after learning about the competition before entering. "We couldn't do it the first year. Our simulation wasn't where we wanted it to be," says Piazza. The competition motivated them to extend their simulations, which had previously been used to predict forces for a highly constrained mechanical knee simulator, to work with the more variable human data.

For Hast, the competition was the capstone to his PhD. The competition's data set, which is synchronized and captures subjects performing different activities and using different walking motions, was much more extensive than he had the time or opportunity to collect for his thesis.

"It is really an embarrassment of riches as far as the data is concerned," says Hast, who plans to use the data in his future research. As for the competition, he says, "It's cost-effective and helps the greater orthopedic community. It's a really great cause." □

DETAILS

The publication "Grand Challenge Competition to Predict *In Vivo* Knee Loads" in the *Journal of Orthopaedic Research* provides a detailed description of the project and data. To access the data, visit http://simtk.org/home/kneeloads.

TAPPING THE BRAIN: Decoding fMRI

By Louisa Dalton Hudock

Revealing the brain's hidden stash of pictures, thoughts, and plans has, until recently, been the work of parlor magicians. Yet within the last decade, neuroscientists have gained powerful methods for delving into the contents of brain activity, allowing them to predict specific thoughts—including images, memories, and intentions—from brain activity.

"When these pattern recognition techniques came out, it gave the field a big boost. People realized that now we can really get at content," says **John-Dylan Haynes**.

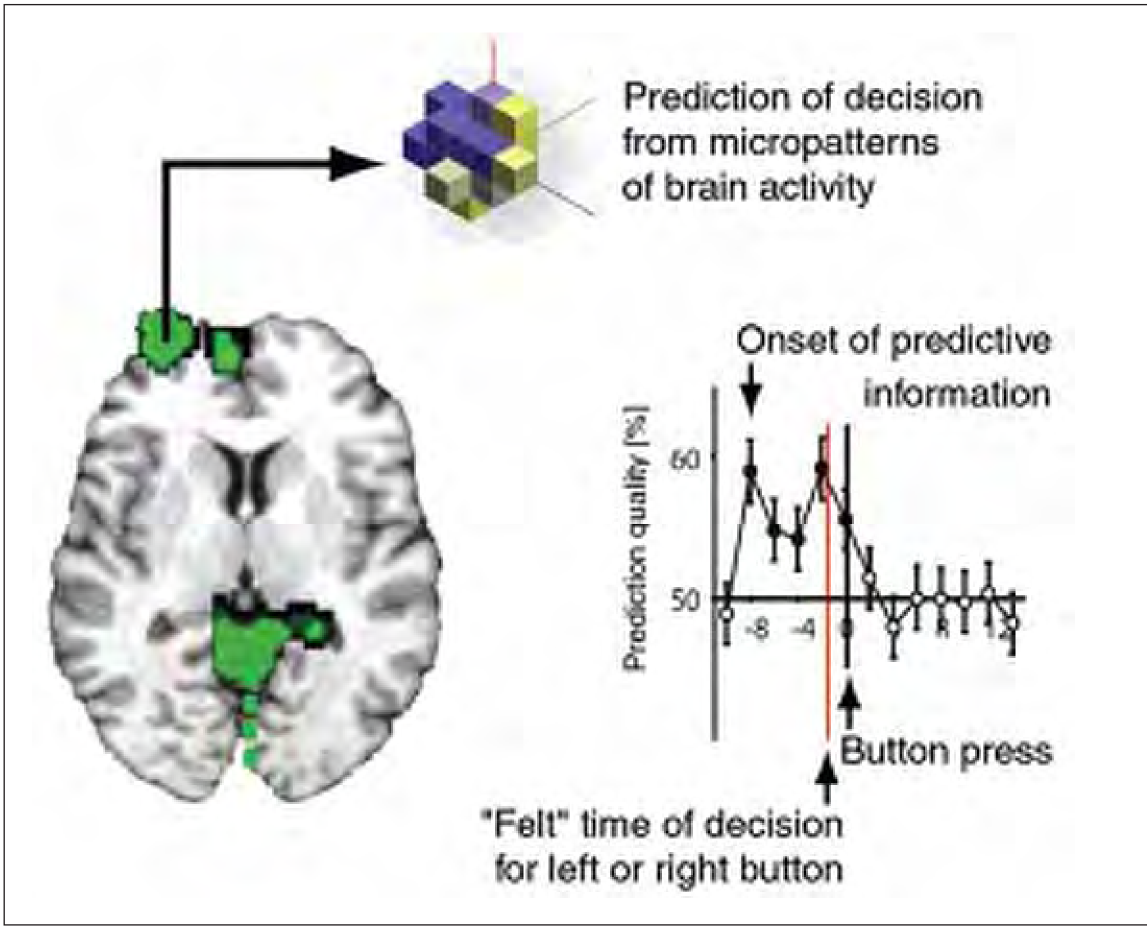
"When these pattern recognition techniques came out, it gave the field a big boost. People realized that now we can really get at content," says **John-Dylan Haynes, PhD**, at the Bernstein Center for Computational Neuroscience in Berlin.

Since the 1990s, functional magnetic resonance imaging (fMRI) has been used to track the flow of blood and oxygen in the brain, thus showing which spots in the brain are busy. Around 2005, neuroscientists discovered that computational multi-voxel pattern analysis (MVPA)—a technique used in other fields such as fingerprint identification—could help them do more than just pinpoint what part of the brain is active. It could help them read meaning in the patterns of activity. If fMRI takes pictures of the brain's hidden bar codes, MVPA decodes them. Two studies below show the power of MVPA applied to intentions and memories. The final study below shows how modeling can recreate the images in the mind's eye.

Decoding Intentions

Some of the early studies using MVPA showed how scientists could use a volunteer's brain patterns to train a linear classifier to predict whether the volunteer was looking at, say, a face or a chair. Then researchers expanded to predict other mental content: emotions, sounds, and memories, for example.

Intentions, Haynes found, are also predictable. In a study published first in 2008 in *Nature Neuroscience* and repeated using ultra-high field fMRI in June 2011 in *Public Library of Science One*, Haynes and his coworkers asked volunteers to enter an MRI machine with one button near their left finger, and one button near their right finger. Relax, they told the volunteers. Watch this stream of letters in front of your eyes, choose which button you will push, remember the letter in front of you at the moment you



***A Brain's Choice to Make.** Before a volunteer clicked on either a left or right button, Haynes and his group collected the volunteer's brain activity using fMRI. The scientists found regions (in green) that, analyzed using pattern recognition techniques, could predict the button the volunteer would push. Those regions revealed the predicted decision up to seven seconds before the volunteer felt he or she had made a decision. Credit: John-Dylan Haynes and Chun Siong Soon.*

choose, and then press the button. The volunteers told the researchers which letter was in front of their eyes when they chose which button to press. Haynes and his coworkers found that not only could the classifier predict which button a volunteer was about to press, it could predict it far earlier than volunteer was able to—up to seven seconds before the volunteer reported consciously choosing which button to push.

With this study, Haynes established that our brains know some things before we do. "Not all of our decisions are made consciously," he says.

Haynes was also one of the first to use MVPA to explore the brain for patterns. Because Haynes didn't know exactly where he might find patterns of decision-making in the brain, or how many seconds before the decision he might be able to see those patterns, he used MVPA as an initial tool for exploration. He used a procedure called searchlight decoding to search the whole brain without making any assumptions ahead of time, and pinpoint the areas and times in the brain that forecast decision-making. More recently, Haynes and his group have found they can predict not just button pushing, but abstract thought as well, such as deciding whether to add or subtract two numbers.

Decoding Memories

Others, including **Frank Tong, PhD**, at Vanderbilt University, have grappled with decoding memories. Tong and his group designed an experiment using MRI to test for working memory (published in *Nature* in April 2009) in the early visual areas of the brain—areas that perceive basic visual features. They showed a volunteer first one set of parallel lines, then another set in a different orientation. They told the volunteer to remember one of the sets. Then after an 11-second pause, they asked the volunteer to recall the orientation they kept in mind.

During that pause, the overall brain activity in the early visual areas of the brain often returned to normal. However, even at that low level of activity (no greater than would be expected when viewing a blank screen), Tong found that pattern classifiers could pick out subtle shifts in brain patterns associated with each orientation, and could still be trained to predict which orientation the volunteer held in memory. "We can see these faint echoes of what they saw before, what they are actively holding in mind," Tong says. "That would be invisible if we didn't do multivoxel pattern analysis."

Tong's study shows that working memory resides in the early visual areas of the brain, a zone where few expected to find such a higher thought process. And it helps establish that images held only in the mind are robust enough that pattern algorithms can decode and predict them.

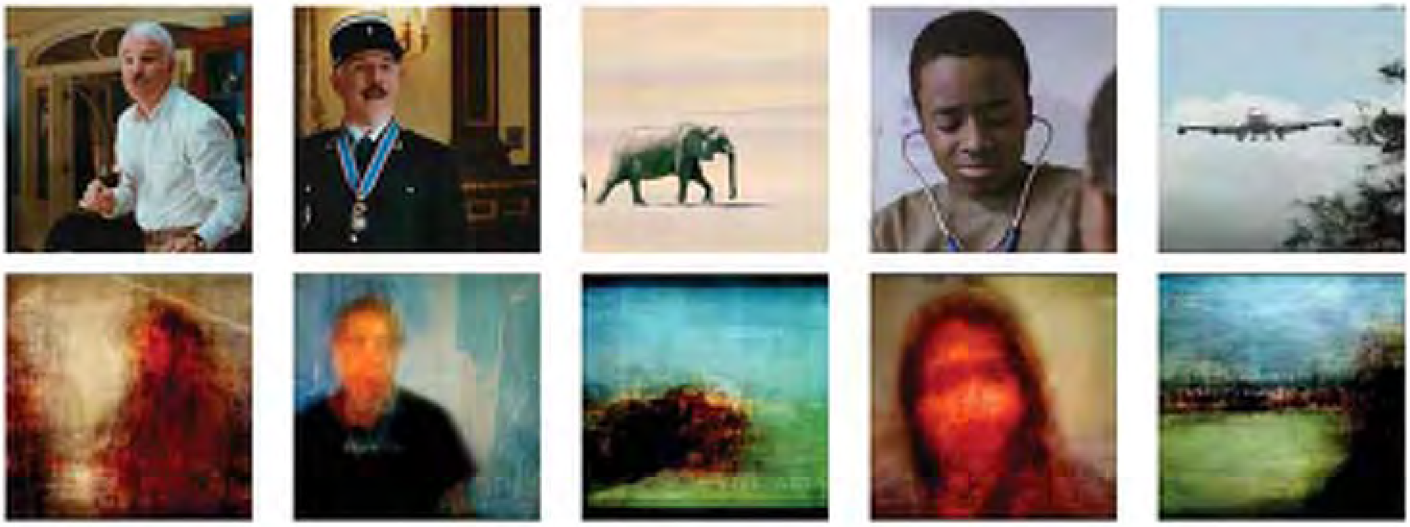
Encoding Models

Researchers have also started decoding novel brain patterns—something a linear classifier cannot do because it can only predict patterns it has already seen. In a study published in *Current Biology* in October, computational neuroscientist **Jack Gallant, PhD,** and post-doctoral researchers **Shinji Nishimoto, PhD**, and **Thomas Naselaris, PhD**, and their group at the University of California, Berkeley used a combination of brain modeling and decoding to reconstruct a primitive version of a movie someone is watching.

After showing a series of movie clips to a volunteer lying inside an MRI machine, the scientists ran the data through a two-step process, Naselaris says. They first created an encoding model: in essence, a virtual brain that generates signals when images are presented to it. In the early visual region that Gallant studies, the brain processes images as pieces of moving edges. Each two-millimeter cube of brain space (also called a voxel) processes the moving edges in a unique way, and Gallant's group created a model for each cube. Each model maps various features of the movie input—motion, color, spatial frequency, and so on—into an output signal that matches as closely as possible the brain activity seen in each cube of the volunteer's fMRI brain data. "It is the signature of that one individual voxel," Naselaris says.

Then, using a brand-new set of movie clips, Gallant's group collected new brain activity data, and set the model going the other direction. They input the brain activity to see if the model could now predict movie scenes from brain data. A novel form of Bayesian decoding helped them match images, taken from a massive database of 1-

Reconstructing Movies via the Brain. Using brain activity of volunteers watching a movie (frames in top row), Gallant and his coworkers created a model that predicts brain patterns from movie clips. They then used that model and a database of 18 million seconds of random movie clips from the Internet to predict the brain activity that would be expected from each clip. They then averaged together the 100 clips in the random library that were most likely to have produced brain activity similar to what was actually observed (bottom row). See Nishimoto, S et al., Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies, **Current Biology** ***21:1641-1646 (2011).***



second movies, to brain activity patterns. Their dark, blurry movie reconstructions are an average of the top 100 short movie clips with a predicted brain pattern that fits best to the actual brain activity patterns.

Naselaris cautions that they can't recreate dreams or other mind's-eye images with this model. Because they focused on an early visual brain region, their model "has everything to do with the light that is hitting your eye," Naselaris says.

Yet Gallant and his group's strategy is powerful for a number of reasons. With it, they can make surprisingly accurate predictions about images that the model has not seen before. In addition, their strategy of first creating an encoding model before decoding gives researchers a method for testing theories about how the brain processes information. For other areas of the brain, the strategy will likely have to be modified. "The brain is a pretty complicated place," Gallant says, "so there is no one grand approach that will solve everything. Instead, there are thousands of neuroscientists using thousands of different approaches to try to move ahead." □

> "The brain is a pretty complicated place," Gallant says, "so there is no one grand approach that will solve everything. Instead, there are thousands of neuroscientists using thousands of different approaches to try to move ahead."

PERSONALIZED CANCER TREATMENT: Seeking Cures Through Computation

By Kristin Sainani, PhD

Personalized cancer therapy is now a reality. A handful of tumor-classifying tests and targeted drugs are in widespread clinical use; and early attempts are underway to match high-risk cancer patients to experimental drugs based on genetic testing of their tumors.

But progress has been incremental and successes have been measured. Cancer is complex and insidious; knock out one bad player with a drug and the system evolves resistance. Patients may live longer, but still die of their disease. To take personalized cancer medicine to the next level—to achieve cures—computational approaches are needed. "We are at a crossroads where it's becoming increasingly difficult to do anything of value without a heavy element of computation," says **Andrea Califano, PhD**, professor of biomedical informatics at Columbia University and director of the Columbia Initiative in Systems Biology.

> "We are at a crossroads where it's becoming increasingly difficult to do anything of value without a heavy element of computation," says **Andrea Califano.**

Bioinformatics and computing are helping to make advances on several fronts, including: cataloging the full spectrum of genomic defects in cancer; identifying the defects that drive malignancy; efficiently translating these discoveries to patient care; and improving the tools that are already in clinical use.

Mapping the Landscape

Several cancer genome initiatives are cataloging the array of molecular defects that define different cancers and cancer subtypes. The NIH's The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC) have already collected multiple layers of data—including sequencing, mutational, copy number alteration, DNA methylation, microRNA, and gene expression data—on an unprecedented number of tumors.

"These different layers will be mapped out, overlaid, and integrated, so we can see the complete genomic picture of the tumor," says **Douglas A. Levine, MD**, head of the Gynecology Research Laboratory at the Memorial Sloan-Kettering Cancer Center and a TCGA investigator.

This is the first time we've had all these different data types on exactly the same samples, Califano says. "And simply amazing findings are coming out."

The TCGA reported results for its second complete genome—for high-grade serous ovarian cancer (a common and aggressive form of this cancer)—in *Nature* in June 2011. The project analyzed data from 489 tumors, including the complete exome sequences of 316, the most ever reported to date for any solid tumor, Levine says.

Among the most exciting findings, about half the patients had defects—inherited or acquired mutations or epigenetic silencing—in the tumor suppressor genes BRCA1 and BRCA2 or in related DNA repair genes. These tumors might respond to PARP inhibitors, which improve survival in women with inherited BRCA1 and BRCA2 mutations, Levine says.

Also, seven percent of tumors had homozygous deletions in the tumor suppressor gene PTEN, a defect not previously reported in this subtype of ovarian cancer. Levine and others have already begun clinical trials to treat this subset of patients with a new class of drugs called PI3 kinase inhibitors, which target the PI3K/AKT/mTOR pathway that PTEN regulates.

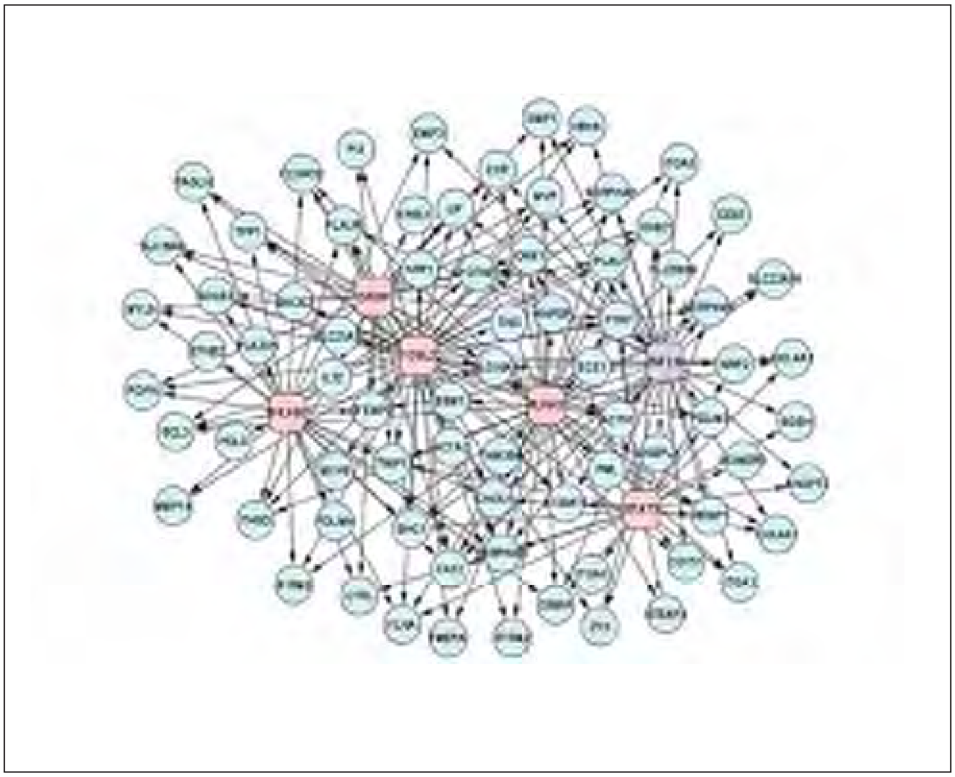
"All these things need to be tested. But we now have the landscape and the roadmap laid out," Levine says. "Molecular medicine is not really being used at all today in ovarian cancer. I hope it can be used in the near future to make better treatment decisions."

Identifying the Driving Defects

Sequencing cancer genomes is only the first step in understanding this disease; the next step is to sort out which genetic changes are driving the cancer—and thus will make robust biomarkers and drug targets—and which are merely incidental. This is where systems biology comes in handy, Califano says. "The idea is to computationally interrogate regulatory networks of the cancer cell to find out what are the genes that are actually causally related to the presentation of a specific tumor phenotype," Califano says.

For example, in a 2010 paper in *Nature*, Califano's team used network analysis to identify two master regulators of a particularly aggressive subtype of glioma brain cancer. These two transcription factors (C/EBPβ and Stat-3) don't appear in the gene expression signature for this subtype, as they are about the 500th and 1300th most differentially expressed, Califano says. "But, if you look at them with these network analyses, they stand out as being the most significant genes in terms of their activity in regulating the signature." Inactivating these genes in mouse xenografts blocked tumor development or reduced malignancy.

In an October 2011 paper in *Cell*, Califano's team used a novel algorithm to identify a "hidden" network of mRNAs and microRNAs that together control PTEN expression. They showed that 13 genes from this network are commonly mutated in glioma. Deleting any of these 13 genes—which had never previously been linked to glioma—suppresses PTEN expression even if the PTEN gene is intact. "So this gives you a very strong hint that mTOR or AKT inhibitors, in combination with other drugs, may actually work in patients that have absolutely no detectable genetic alteration of PTEN," Califano says.



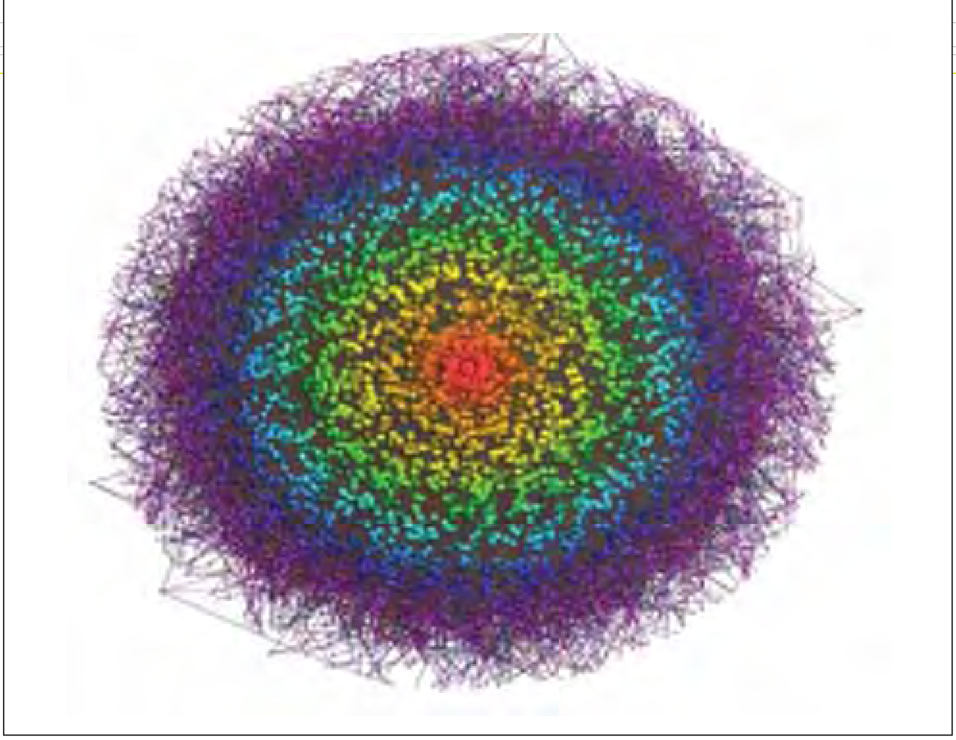
Master Regulators.* *Six transcription factors, including C/EBPβ and Stat-3, control most of the genes in the gene signature of an aggressive subtype of high-grade glioma. Reprinted by permission from Macmillan Publishers Ltd: Carro MS, et al., The transcriptional network for mesenchymal transformation of brain tumours. **Nature.** ***2010;463:318-25.***

Bringing the Data to Biologists and Physicians

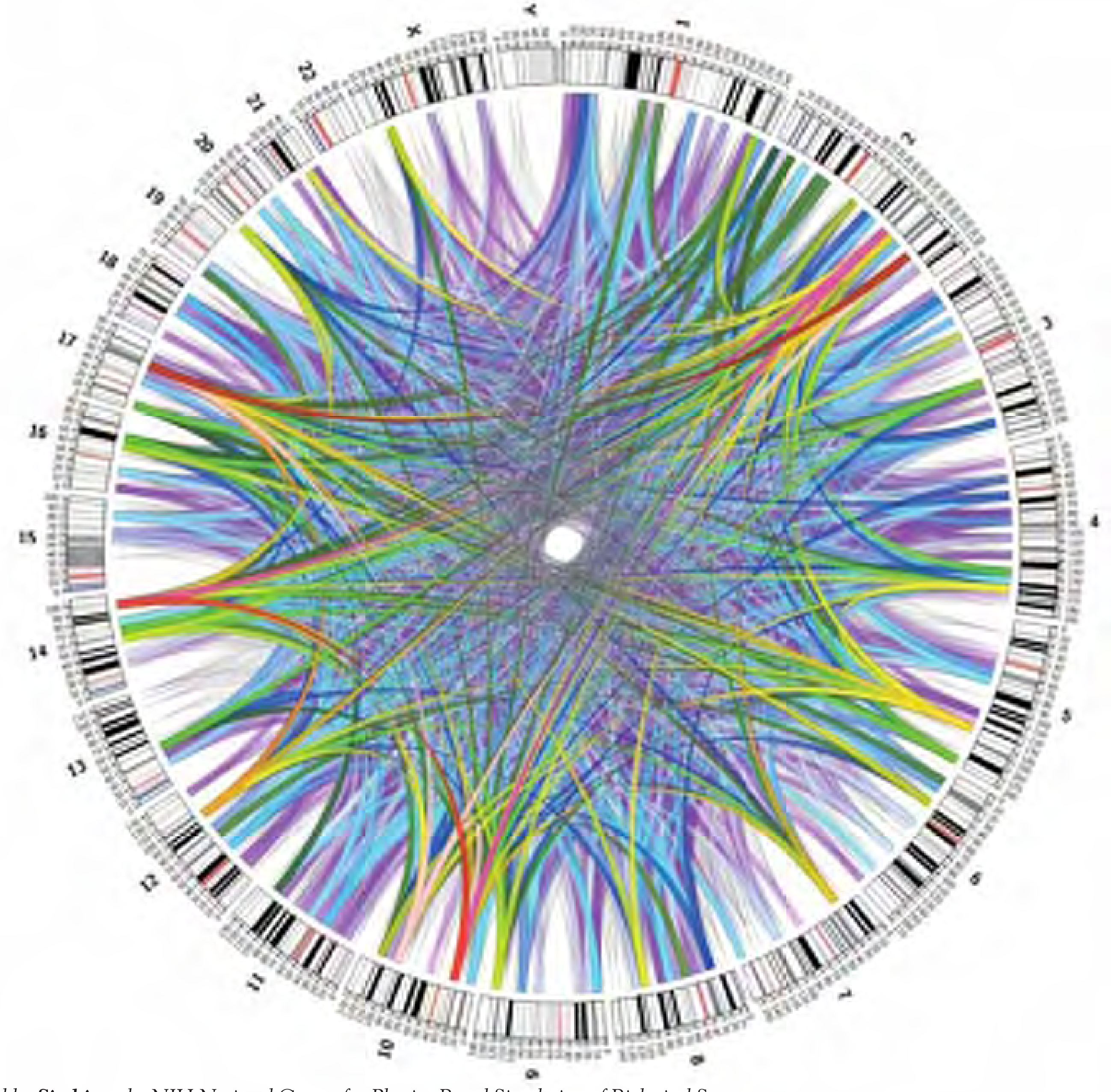
Experimental biologists and clinical researchers are in the best position to translate cancer genome findings into meaningful advances in patient care. But they often lack the expertise needed to access and make sense of the data. A team of scientists at Georgetown University is trying to bridge this gap by creating a user-friendly integrated database called G-DOC (Georgetown Database of Cancer), which they describe in a September 2011 paper in *Neoplasia*.

"We are unique in the way that we provide the data mining as well as the analytics in one environment," says **Subha Madhavan, MD**, director of clinical research informatics at the Lombardi Comprehensive Cancer Center at Georgetown. G-DOC integrates patient data, genomic data, and small molecule data ("for matchmaking molecular targets with drugs," Madhavan says) with popular tools for analyzing and visualizing these data, including GenePattern, Pathway Studio, and Cytoscape.

Continued on page 8



***Hidden Network**. Left: Graphic visualization of a complex network of RNA-RNA interactions in glioma. The interactions regulate the expression of oncogenes and tumor suppressor genes, including PTEN. Nodes represent individual RNAs and edges represent RNA-RNA interactions. Nodes near the center are more tightly regulated. Below: Close-up of the densest 564-node subgraph shown in red at the center of the network. Reprinted Sumazin, et al, An Extensive MicroRNA-Mediated Network of RNA-RNA Interactions Regulates Established Oncogenic Pathways in Glioblastoma,* Cell *147:2:370-381 (2011) with permission from Elsevier.*



Users (both internal and external to Georgetown) have contributed clinical data on more than 3000 patients with breast or gastrointestinal cancers. “We were just lucky to work with investigators who had de-identified clinical data that we could leverage,” Madhavan says. Madhavan’s team has also imported a wealth of data from public databases and published articles. “We bring in the raw data and standardize them, so there’s a lot of value added to that data,” Madhavan says. They will add TCGA data for breast and colon cancer when they become available, she says.

> “Right now ... the vast majority of labs can only access computational analyses through collaborations,” Califano says. “There has to be some kind of connective tissue.”

Researchers can use G-DOC to generate or test hypotheses; run *in silico* experiments; learn about the newest types of data—including next generation sequencing, metabolomics, DNA copy number abnormalities, and microRNA expression—as well as about systems biology; and speed up the pace of their research. For example, it took one person using G-DOC one month to complete a colon cancer analysis that would otherwise have taken a team of people at least a year to complete, Madhavan says.

Overcoming Computing Barriers

To fully realize the vision of personalized cancer therapy, more labs will need to become computationally savvy, Califano concludes. “Right now there are a few computationally empowered labs, but the vast majority of other labs can only access computational analyses through collaborations,” Califano says. “There has to be some kind of connective tissue.” He points to models such as the National Centers for Biomedical Computing, which have fostered collaborations between computer specialists and biologists and physicians. “I think this is an example of the way things can go forward.” □

ADVANCING Gene Expression Signatures

Gene expression signatures that stratify patients into likely and unlikely treatment responders are already in clinical use for certain cancers. But these “first generation” tests have severe limitations, says **W. Fraser Symmans, MD**, professor of pathology at MD Anderson Cancer Center. He and his colleagues are using state-of-the-art bioinformatics and biostatistics techniques to develop the next generation of gene expression tests for breast cancer.

Symmans and his colleagues discovered a paradox with some first generation tests for breast cancer. The tests accurately separate patients into “good” and “poor” responders to chemotherapy; but the “good responders” have worse survival. (The tests misclassify certain aggressive tumors that initially respond vigorously to chemotherapy but tend to relapse.) His team developed a second-generation test, described in the May 2011 issue of *JAMA*, that overcomes this issue and accurately predicts survival.

The test comprises a series of gene signatures (from the tumor) that sequentially predict: (1) response to hormonal treatment; (2) resistance to chemotherapy; and (3) sensitivity to chemotherapy. “We realized that one predictor was not going to be enough to capture the complexity,” says **Christos Hatzis, PhD**, who led the computational aspects of the project; Hatzis is founder and vice president of technology at Nuvera Biosciences Inc., which has commercial rights to the technology. The team used a multivariate approach to identify the key genes that define the signature; univariate approaches yield too many redundancies because genes work in pathways, Hatzis says.

The test accurately identifies patients who will respond to therapy about twice as often as standard methods. “It doesn’t completely solve the problem but it’s a big step forward,” Hatzis says.

FOLLOW THE MONEY:
Big Grants in Biomedical Computing

By Kristin Sainani, PhD

Several biomedical computing projects received multi-million dollar funding in the fall of 2011, including efforts to: simulate the cardiac physiology of the rat; build a state-of-the-art DNA simulation toolkit; build an artificial pancreas; and mine data for clues about psychiatric disease. The initiatives will bring together diverse experts, datasets, or models to accomplish ambitious goals.

Modernizing the Lab Rat

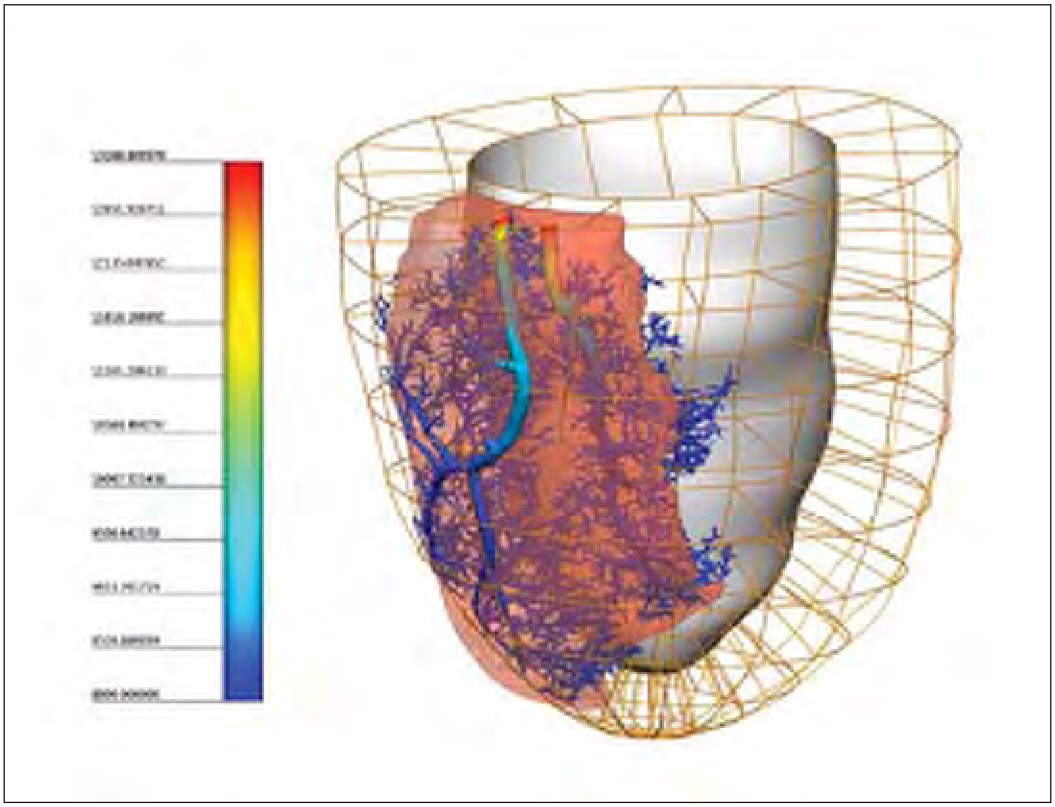
A new type of lab rat—one simulated on a computer—will help scientists tease out the multifactorial causes of cardiovascular disease, thanks to a $13 million grant from the National Institutes of Health. The grant establishes a new National Center for Systems Biology.

"The goal of the Virtual Physiological Rat Project is to understand how high level traits, such as hypertension, arise from multiple inputs at multiple levels, including genetic variation and environmental perturbations," says principal investigator **Daniel A. Beard, PhD**, professor of physiology at the Medical College of Wisconsin.

Beard's team will build detailed computational models of the rat's heart, vasculature, and kidneys. "In some cases, we're drilling down to the individual cells and individual transporters and pumps that are involved in the operation of the organ and integrating all the way up to the whole organ," Beard says. Though sophisticated models already exist for some of the pieces—for example, the heart is the best modeled of all organs—they will have to adapt these models for rats in general and for particular genetic strains of rats.

"The biggest novelty and the biggest challenge is this cycle of making a prediction and then making a rat," Beard says. "We're doing this on an unprecedented scale."

Once the models are refined, Beard's team will breed new strains of rats *in silico* and subject them to different environmental stressors to predict which combinations of genes and environment lead to hypertension, heart failure, kidney failure, and other cardiovascular diseases. Then his team will test the most interesting predictions with experiments in real rats. "The biggest novelty and the biggest challenge is this cycle of making a prediction and then making a rat," Beard says. "We're doing this on an unprecedented scale."



***Virtual Rat Heart:** A computational model of the rat heart that will be incorporated into the Virtual Physiological Rat Project. Courtesy of: Daniel Beard, Medical College of Wisconsin.*

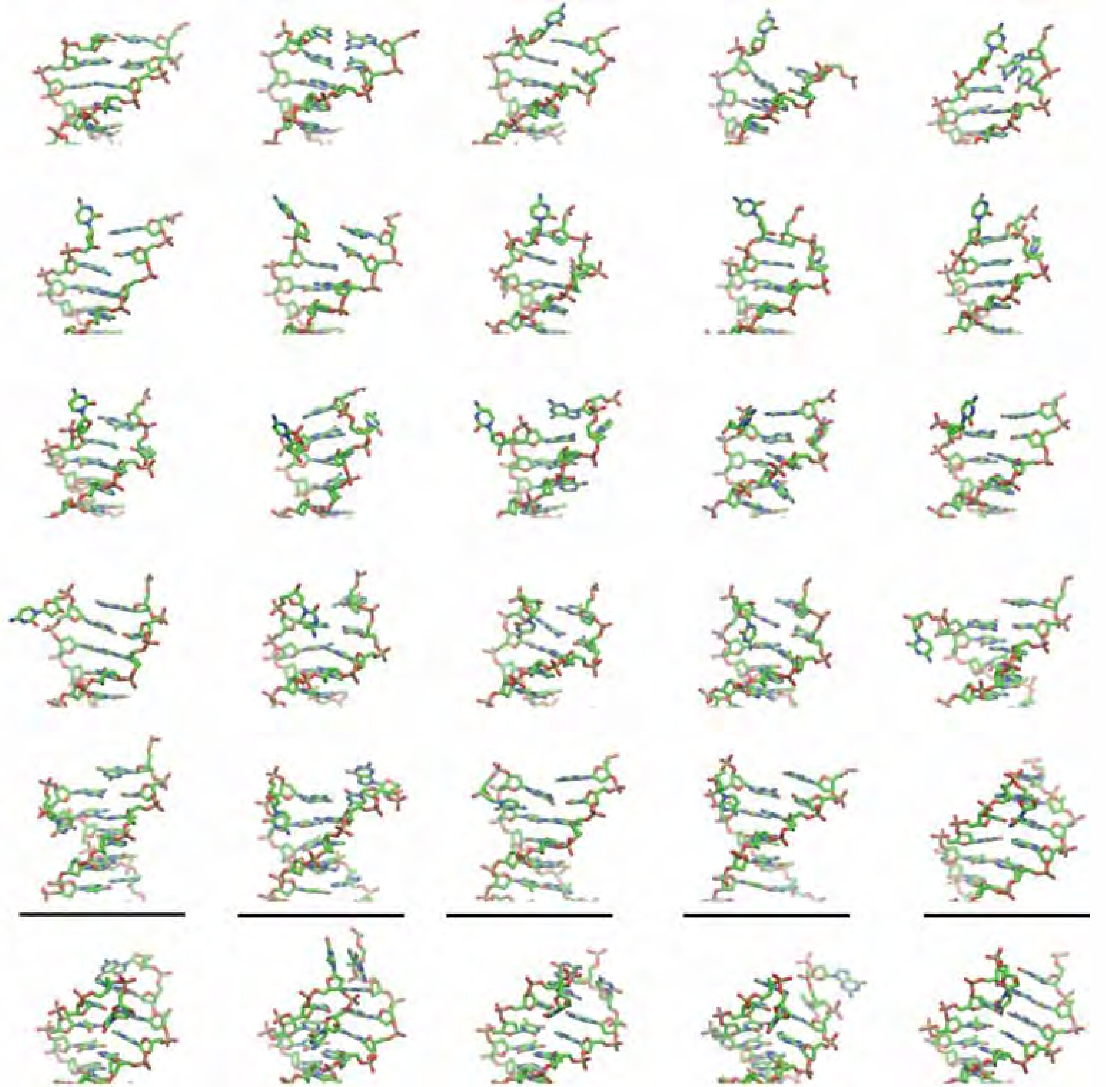
Simulating DNA at All Levels

A new DNA simulation toolkit will be the first to span all levels of resolution. Funded by a $3 million grant from the European Research Council, the toolkit will help scientists gain new insights into how DNA functions, including how genes are regulated and how they interact with the environment (epigenetics).

"The overall goal is to develop a complete theoretical and computational framework to be able to simulate DNA in a multiscale manner, from atomistic to chromatin scale," says principal investigator **Modesto Orozco, PhD**, senior professor and head of IRB Barcelona's Molecular Modeling and Bioinformatics group and director of life sciences at the Barcelona Supercomputing Center. "If we are successful, we will define a unique series of tools covering the entire time and size scale of DNA."

The toolkit will help researchers answer questions about how nucleic acids work "from the point of view of the first principles

"The overall goal is to develop a complete theoretical and computational framework to be able to simulate DNA in a multiscale manner, from atomistic to chromatin scale," says principal investigator Modesto Orozco.



***DNA Unveiled:** This computer simulation (which precedes from left to right and top to bottom) gives insights into the mechanism by which DNA starts to unfold. Courtesy of: A. Pérez, IRB Barcelona.*

of physics"— for example, how the physical properties of chromatin impact DNA regulatory mechanisms. "We will have to overcome major challenges at the frontier between the different simulation levels," Orozco says. "It is a very ambitious project."

Advancing the Artificial Pancreas

A device that mimics the pancreas' job may soon be a reality, thanks in part to a $4.5 million grant from the National Institutes of Diabetes, Digestive, and Kidney Diseases. The device—which features a state-of-the-art control algorithm—would free patients with type I diabetes from their current regimen of manual glucose monitoring and insulin injections.

"We've assembled an international dream team," says principal investigator **Frank Doyle**, **PhD**, professor of chemical engineering and of electrical engineering at the University of California, Santa Barbara. His team includes scientists from Italy, the Mayo Clinic, the University of Virginia, and the Sansum Diabetes Institute. Several groups have prototypic artificial pancreas devices in testing. But Doyle's team aims to bring the first sophisticated device into real-world use. "The exciting part of this grant is the possibility of getting beyond in-clinic prototyping," Doyle says.

The devices consist of an ipod-sized computer, glucose sensor, and insulin pump (which can be attached to the arms, legs, or stomach). A simple version (made by Medtronic) is already on the market in England: the system monitors glucose and shuts off the pump automatically when blood sugar drops too low. But Doyle's team is going beyond such a simple feedback loop, using an advanced algorithm called "model predictive control" (which is also used in aerospace controls). "We forecast and anticipate insulin needs," rather than simply responding to current glucose levels, Doyle says. The algorithm will even adapt to a patients' individual patterns, such as the timing of exercise and meals, as well as to individual variation in insulin metabolism, Doyle says. "It won't be a one-size fits all algorithm; it will be tailored and customized to the individual patient."

Synthesizing Data on Mental Illness

A new center at the University of Chicago will explore the origins of psychiatric disease by integrating existing data from diverse disciplines and across multiple sites. The center is the newest Silvio O. Conte Center for Neurosciences Research and the first with a computational focus. It received $14 million in grants from the National Institutes of Mental Health and the Chicago Biomedical Consortium.

"We have a lot of datasets from different communities that have never been analyzed within the same model before. It's an exciting research opportunity," says principal investigator **Andrey Rzhetsky**, **PhD**, professor of medicine and human genetics at the University of Chicago Medical Center.

Rzhetsky will head a consortium of 15 lead investigators from seven schools that will bring together clinical data, genetic linkage data, gene pathway data, functional data on genes and proteins, drug data, and drug-gene interaction data. "The main premise of the center is to get together wonderful specialists in different disciplines; make them talk to each other; design models that span all datasets; and make predictions that can be tested experimentally."

Rzhetsky's team will attempt to unearth novel connections between genes, environment, and disease phenotypes, as well as between the disorders themselves. For example, Rzhetsky and colleagues have previously shown that autism, schizophrenia and bipolar disorder have considerable genetic overlap. "You can get a lot more from joint analysis of several phenotypes than from a single phenotype," Rzhetsky says. □

LEVERAGING SOCIAL MEDIA: For Biomedical Research

By Katharine Miller

It has become commonplace for people to use social media to share their healthcare stories, seek a community of individuals with the same diseases, and learn about treatment options. All this Internet activity also produces data that can be used for research.

"In the networked world, who cures cancer? We all do," says Paul Wicks of PatientsLikeMe.

"In the networked world, who cures cancer? We all do," says **Paul Wicks**, **PhD**, director of research and development at PatientsLikeMe, a site where people diagnosed with serious life-changing illnesses can record and share information.

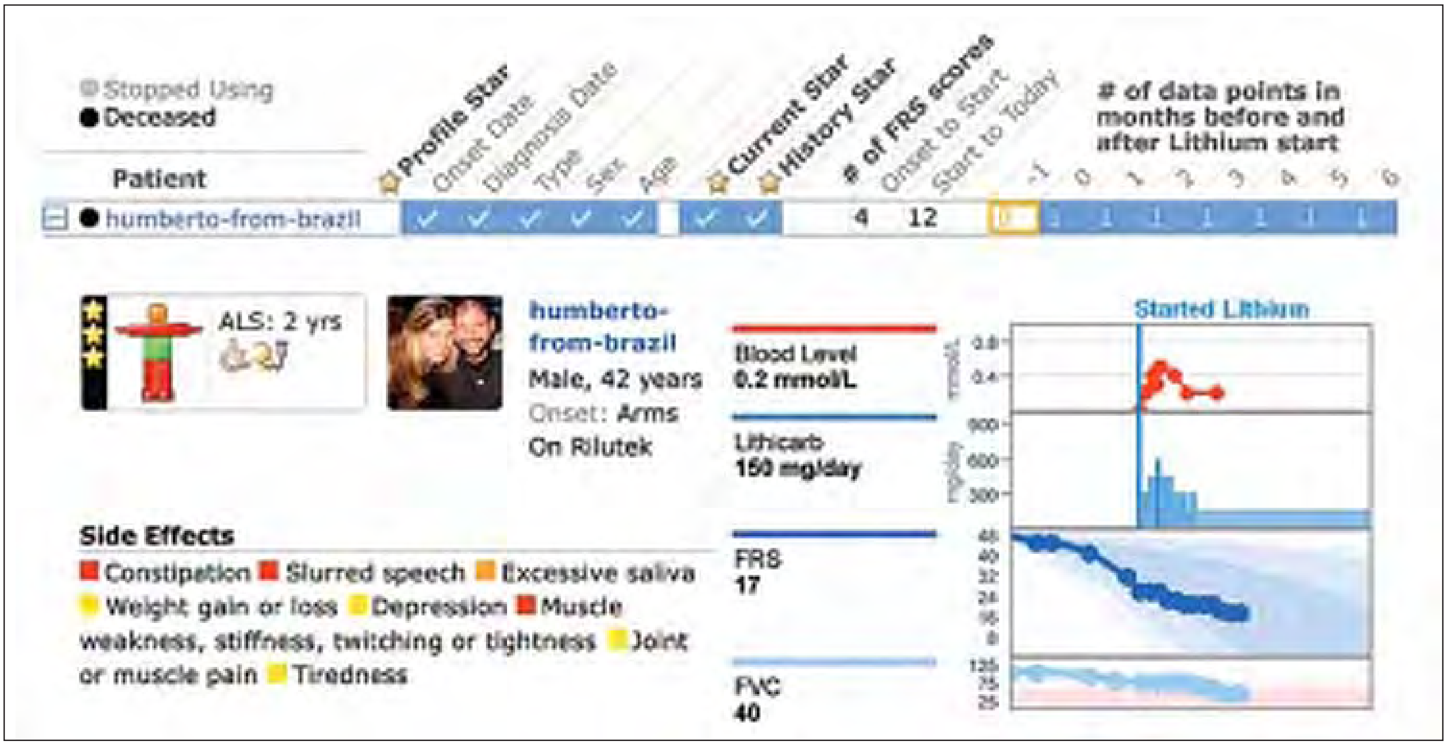
For PatientsLikeMe and a number of other sites, doing biomedical research using data gathered online is part of the business plan. With names such as 23andMe, MedHelp, TUDiabetes, myMicrobes.eu, CureTogether, these sites blend community building with information gathering. They then turn to computational approaches, such as data mining and natural language processing (NLP), to analyze the information gathered.

This crowd-sourced research often reaches into realms that otherwise wouldn't or couldn't be studied, due to a lack of either appropriate information or financial support. Moreover, with their access to large populations of both cases and controls, these sites are rapidly producing clinical research results. That they function in a landscape of ever-changing and growing data just makes the process that much more interesting.

Doing Research That Others Can't or Won't

On social media healthcare sites such as PatientsLikeMe, people record and share information about their diseases. This self-reported data may have some inherent biases, says Wicks, who hopes that those issues will disappear as they get to a large enough scale. But it also has some inherent strengths: Online, people talk about issues they might not raise with a physician, and they can report on and track their conditions more frequently.

To take advantage of this, PatientsLikeMe set out to "do research that's new and novel… and not just cheaper than a survey by mail," Wicks says. Moreover, he says, "Our inclination is to do work that reflects the needs of patients." So, for example, PatientsLikeMe studied the incidence of compulsive gambling among people with Parkinson's disease (PD) because people on the site were concerned about the phenomenon. In their sample—assembled in the course of just a week—they found that gambling was twice as common among PD patients as would be expected from physician notes—suggesting that patients don't necessarily share certain embarrassing information with their doctors (although bias in the sample could also be an issue)—and that compulsive gambling was not associated with being on a dopamine-agonist drug (as previous studies had suggested).



humberto-from-brazil
Male, 42 years
Onset: Arms
On Rilutek

Blood Level
0.2 mmol/L

Lithicarb
150 mg/day

FRS
17

FVC
40

Side Effects
🟥 Constipation 🟥 Slurred speech 🟧 Excessive saliva 🟨 Weight gain or loss 🟨 Depression 🟥 Muscle weakness, stiffness, twitching or tightness 🟨 Joint or muscle pain 🟨 Tiredness

This screenshot of the ALS tracking tool for an individual patient in the PatientsLikeMe lithium study shows how the patients entered their disease characteristics, demographics, blood levels, dosage, ALSFRS-R score (a measure of disease progression), forced vital capacity, and side effects. Reprinted from supplemental figure 1, Wicks, P, et al., Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm, **Nature Biotechnology *29, 411–414 (2011).***

PatientsLikeMe has also looked at off-label drug use. "People don't want to fund research of off-label drugs, especially generics," Wicks says. "Our platform provides a way to capture data that no one else has the bandwidth to look at."

Access to Large Populations for Clinical Trials

At PatientsLikeMe, 23andMe and MedHelp, researchers are finding that online communities offer a huge benefit to clinical research: A vast treasure trove of cases and an even vaster population of controls.

Launched in 2005, PatientsLikeMe has 115,000 users and covers about 1300 conditions. For about twenty of those conditions, PatientsLikeMe collects patient data in a structured way, requesting information on specific outcomes—the kinds of things typically used for clinical trials. "We build it so we can prepare for future research studies," Wicks says.

For example, early on, the site created a community and several surveys for people with ALS (amyotrophic lateral sclerosis). This meant they already had lots of valuable background data when, in 2008, the community clamored for treatment with lithium. A small (16-person) study in Italy had shown that lithium could slow the progress of the disease. But PatientsLikeMe researchers were wary. "Many studies of ALS treatments kill patients faster than placebo," Wicks says. "You want to be sure it's not harmful." So PatientsLikeMe immediately spent a year gathering data on off-label lithium use by 150 eager ALS patients in their community. And they matched cases to controls in a rigorous way—using an algorithm that considered data on both ALS onset and the shape of the disease progression curve, key traits that vary in significant ways among ALS patients. This was possible, Wicks says, because they had lead-in data describing the patients' status before taking the drug. Preliminary results announced in December 2008 (just nine months after the Italian research was published) showed that lithium was not effective in slowing disease progress. Since then, this result was confirmed in randomized clinical trials. The PatientsLikeMe research was published in *Nature Biotechnology* in April 2011.

The genotyping service 23andMe does research using data they gather from people who provide not only saliva samples but also phenotype information gathered through online surveys. And the company leverages social media such as Twitter and Facebook to recruit communities of individuals with a particular disease. "Recruiting is not done through a clinical center," says **Chuong (Tom) Do, PhD**, a research scientist at the company. "It's done entirely online."

For many communities, Do says, "we actually have the genotyping process completely sponsored, making the financial barriers to participation in the research as low as possible." For example, using a private donation from Google founder Sergey Brin, 23andMe was able to sponsor most of the genotyping costs for PD cases in a recent study. But for controls, Do says, 23andMe has the advantage of being able to use data from the population of people who pay for the service. For a less common disease like PD, Do says, the small proportion of misclassified cases mixed in with that population would have a negligible effect on the results of the study. "We just need to be sure to get enough cases," he says. "Controls come for free. It's actually a huge help for us." Indeed in a recent study of PD (published in *PLoS Genetics* in June of 2011) involving roughly 3400 cases and 29,000 controls, they were able to identify two novel genes contributing to the risk of developing PD. Because of the control group's size, Do says, "We could wring a lot of statistical power from our dataset."

23andMe has also launched initiatives to study several rare disorders, namely sarcoma and myeloproliferative neoplasms. While recruitment for these conditions can be difficult and expensive in the setting of a traditional research center, 23andMe's system allows for aggregation of individuals at low overhead to the company and without regard for geographic barriers, Do says.

With over 12 million users, MedHelp is the largest online health community. The business focuses on helping people track their diseases as well as connecting them with appropriate communities and physicians. In addition, though, they work in partnership with academics, physicians and others to extract useful knowledge from MedHelp's accumulating data. For example, several physicians examined data on lens implant failures pulled from the eyecare forums on MedHelp (forums that were sponsored by the American Academy of Ophthalmology). The researchers found that multi-focal implants had a much higher failure rate than other types—information that was very valuable to the ophthalmology community.

Rapid Turnaround Time

Compared to clinical research centers, those who leverage social media web sites can conduct clinical research very quickly. The PatientsLikeMe study of lithium use in ALS was completed in just twelve months—before a randomized clinical trial even began recruitment. In another example, when members of the site's ALS community raised a question about excessive yawning, PatientsLikeMe published research on the problem in just three months.

"The ability to accelerate the pace of research through social media is exciting to me," says Tom Do of 23andMe.

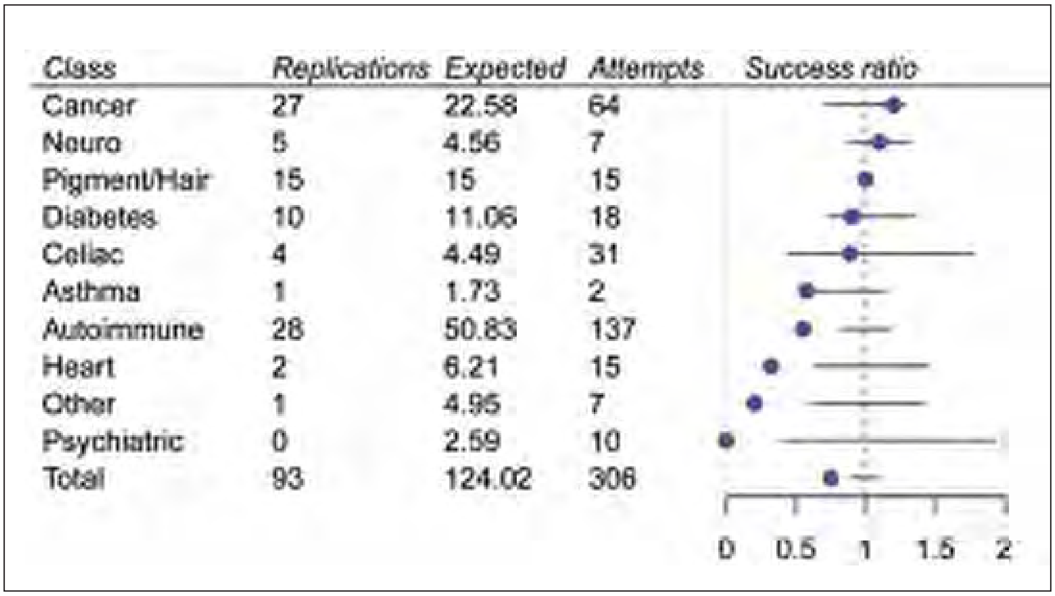
"The ability to accelerate the pace of research through social media is exciting to me," Do says. Part of the acceleration comes from the immediate ability to amass and access large cohorts, he says. But it goes beyond that. For example, when 23andMe set out to determine whether its data was reliable enough to replicate published genome-wide association studies (GWAS), they completed the task at lightning speed compared to a typical GWAS. Indeed, it took 23andMe less than one year to replicate and present results from a PD GWAS that had taken the previous researchers almost six years from hypothesis to publication.

Flexibility

If data initially collected online is incomplete or even wrong, it is easily amended by going back to the users with revised surveys. For example, when 23andMe first attempted to replicate a GWAS for celiac disease, they did not find the expected associations. Because their survey had asked "Have you ever been diagnosed with celiac disease," they believed their study might include some false positives. So they re-worded the question to ask: "Have you ever been diagnosed with celiac disease, as confirmed by a biopsy of the small intestine." And with the newly (and rapidly) acquired answers, they were able to replicate four of the six expected associations.

Dealing with changes of this kind also means re-running the GWAS. "Many times based on research results, we'll ask new questions," Do says. "So we end up with a very fluid dataset and the need for tools that allow us to work with the data as it constantly changes." They often run the same GWAS studies repeatedly. "We have over 1000 that we run on a regular basis, culled from the 50 plus surveys," Do says. That is a unique computational aspect of the work: custom-built software to conduct parallelized GWAS on the same dataset.

Today, 23andMe has more than 120,000 peoples' genotypes in its database. "We look forward to the day when we have one million plus," Do says. "We can only imagine the types of discoveries that will be possible with a database that size." □



Class	Replications	Expected	Attempts
Cancer	27	22.58	64
Neuro	5	4.56	7
Pigment/Hair	15	15	15
Diabetes	10	11.06	18
Celiac	4	4.49	31
Asthma	1	1.73	2
Autoimmune	28	50.83	137
Heart	2	6.21	15
Other	1	4.95	7
Psychiatric	0	2.59	10
Total	93	124.02	306

***23andMe successfully replicated previous GWAS for a number of diseases as shown here in a chart of success rate (versus total power) by disease class. Expected = number of associations they expected to replicate. Attempts = number of associations they attempted to replicate. The blue dot represents the success ratio (number of successful replications divided by number of expected replications). The black line represents the 95 percent prediction interval for the success ratio. Reprinted from Tung, J, et al., Efficient Replication of Over 180 Genetic Associations with Self-Reported Medical Data,* PLoS ONE 6(8) (2011).**

Big Data Analytics

in Biomedical Research

By Katharine Miller

Biomedical infrastructure for big data analytics lags well behind the curve. Yet the examples described here suggest possible pathways to the dream of an intelligent healthcare system with big data at its core.

"We have recommendations for you," announces the website Amazon.com each time a customer signs in.

This mega-retailer analyzes billions of customers' purchases—nearly $40 billion worth in 2011 alone—to predict individuals' future buying habits. And Amazon's system is constantly learning: With each click of the "Place your order" button, the company's databank grows, allowing it to both refine its predictions and conduct research to better understand its market.

These days, this sort of "Big Data Analytics" permeates the worlds of commerce, finance, and government. Credit card companies monitor millions of transactions to distinguish fraudulent activity from legitimate purchases; financial analysts crunch market data to identify good investment opportunities; and the Department of Homeland Security tracks Internet and phone traffic to forecast terrorist activity.

Where is Amazon's equivalent in healthcare and biomedical research? Do we have a "learning healthcare system" that, like Amazon.com, can glean insights from vast quantities of data and push it into the hands of users, including both patients and healthcare providers? Not even close.

It's a situation that frustrates and inspires **Colin Hill**, CEO, president, chairman and cofounder of GNS Healthcare, a healthcare analytics company. "When I go to my doctor for some treatment, he's kind of guessing as to what drug works," he says. With the data currently being captured and stored, he says, there's now an opportunity to take a broader view of the problem. “We need to make this system smarter and use data to better determine what interventions work,” he says.

And there is hope, says **Jeff Hammerbacher**, who formerly led the data team at Facebook and is now chief scientist at Cloudera, a company that provides businesses with a platform for managing and analyzing big data. “I believe that the methods used by Facebook and others—commodity hardware, open source software, ubiquitous instrumentation—will prove just as revolutionary for healthcare as they have for communications and retail,” he says.

Others agree: “We have to create an infrastructure that allows us to harvest big data in an efficient way,” says **Felix Frueh**, **PhD**, president of the Medco Research Institute.

> “I believe that the methods used by Facebook and others (commodity hardware, open source software, ubiquitous instrumentation) will prove just as revolutionary for healthcare as they have for communications and retail,” says Jeff Hammerbacher.

Right now, biomedical infrastructure lags well behind the curve. Our healthcare system is dispersed and disjointed; medical records are a bit of a mess; and we don’t yet have the capacity to store and process the crazy amounts of data coming our way from widespread whole-genome sequencing. And then there are privacy issues (see “Privacy in the Era of Electronic Health Information,” a story also in this issue). Moreover, while Amazon can instantly provide up-to-date recommendations at your fingertips, deploying biomedical advances to the clinic can take years.

Despite these infrastructure challenges, some researchers are plunging into biomedical Big Data now, in hopes of extracting new and actionable knowledge. They are doing clinical trials using vast troves of observational health care data; analyzing pharmacy and insurance claims data together to identify adverse drug events; delving into molecular-level data to discover biomarkers that help classify patients based on their response to existing treatments; and pushing their results out to physicians in novel and creative ways.

Perhaps it’s asking too much to expect that the complexities of biology can be boiled down to Amazon.com-style recommendations. Yet the examples described here suggest possible pathways to the dream of an intelligent healthcare system with big data at its core.

DEFINING BIG DATA IN BIOMEDICINE

Big data in biomedicine is coming from two ends, says Hill: the genomics-driven end (genotyping, gene expression, and now next-generation sequencing data); and the payer-provider end (electronic medical records, pharmacy prescription information, insurance records).

On the genomics end, the data deluge is imminent. With next-generation sequencing—a process that greatly simplifies the sequencing of DNA—it is now possible to generate whole genome sequences for large numbers of people at low cost. It’s a bit of a game-changer.

“Raw data-wise, it’s 4 terabytes of data from one person,” says **Eric Schadt**, chair of genetics at Mt. Sinai Medical School in New York City. “But now imagine doing this for thousands of people in the course of a month. You’re into petabyte scales of raw data. So how do you manage and organize that scale of information in ways that facilitate downstream analyses?”

For now, as we wait for next-gen sequencing to work its magic, genomics data matrices remain long and thin, with typically tens to hundreds of patients but millions or at least tens of thousands of variables, Hill notes.

“But on the payer-provider data side,” Hill says, “we’re dealing now with large longitudinal claims data sets that are both wide and deep.” A data matrix might have hundreds of thousands of patients with many characteristics for each—demographics, treatment histories, outcomes and interventions across time—but typically not yet thousands or millions of molecular characteristics.

To a great degree, the two sides of biomedical big data have yet to converge. Some researchers work with the clinical and pharmaceutical data; others work with the biomolecular and genomics data. “The bottom line is,” says **Eric Perakslis**, **PhD**, chief information officer at the U.S. Food and Drug Administration, “the large body of healthcare data out there has yet to be truly enhanced with molecular pathology. And without that you’re really not getting at mechanisms of action or predictive biology.” Where there is data, he says, “It’s almost this random thing: Molecular data is collected at a few time points but that’s it.”

Nevertheless, Schadt believes that a world where these biomolecular and clinical datasets come together may arrive soon.

> “In maybe ten years time,” says Eric Schadt, “all newborns and everyone walking through the door will have his or her genome sequenced and other traits collected and that information will all be crunched in the context of their medical history to assess the state of the individual.”

“In maybe ten years time,” he says, “all newborns and everyone walking through the door will have his or her genome sequenced and other traits collected and that information will all be crunched in the context of their medical history to assess the state of the individual.”

THE TOOLS OF BIG DATA ANALYTICS

If you have data sets with millions or tens of millions of patients followed as a function of time, standard statistics aren't sufficient, especially if you are looking for associations among more than two variables, or data layers. "This is not about genome-wide association studies (GWAS)," Hill says. Such studies typically seek to connect genomic signatures with disease conditions—essentially looking at only two layers of data. "When people start doing this from multiple layers of data, that's where it becomes non-trivial," Hill says. "That's where in my mind it gets to big data analytics rather than biostatistics or bioinformatics."

"When people start doing this from multiple layers of data, that's where it becomes non-trivial," Colin Hill says. "That's where in my mind it gets to big data analytics rather than biostatistics or bioinformatics."

Many of the tools of big data analytics are already being used in other fields, says Schadt. "We're almost latecomers to this game but the same sorts of principles applied by Homeland Security or a credit card fraud division are the kinds of approaches we want to apply in the clinical arena."

The U.S. Department of Homeland Security, for example, examines such things as cell phone and email traffic and credit card purchase history in an attempt to predict the next big national security threat. They want to consider everything together, letting the data speak for itself but looking for patterns in the data that may signify a threat, Schadt says. They achieve this using machine learning in which computers extract patterns and classifiers from a body of data and use them to interpret and predict new data: They know when a prior threat occurred, so they look for features that would have helped them predict it and apply that looking forward. In a clinical setting, that could mean looking at not only which molecular or sequencing data predicts a drug response but also what nurse was on duty in a particular wing during specific hours when an event occurred. "You just want all this information and then crunch it to figure out what features turn out to be important," Schadt says.

"We're almost latecomers to this game," says Schadt, "but the same sorts of principles applied by Homeland Security or a credit card fraud division are the kinds of approaches we want to apply in the clinical arena."

In addition to machine-learning, Hill says, there is a need for approaches that scale up to the interpretation of big data. In his opinion, this means using hypothesis-free probabilistic causal approaches, such as Bayesian network analysis, to get at not only correlations, but cause and effect.

He points to strategies developed by **Daphne Koller, PhD,** professor of computer science at Stanford University, as an example of what can be done. Much of her work involves the use of Bayesian networks—graphical representations of probability distributions—for machine learning. These methods scale well to large, multi-layered data sets, he says. Hill's company, GNS Healthcare, has developed its own variation, which they call "reverse engineering and forward simulation" (REFS). "We break the dataset into trillions of little pieces, evaluating little relationships," he says. Each fragment then has a Bayesian probabilistic score signaling how likely the candidate relationship is as well as the probability of a particular directionality (an indication of possible cause and effect). After scoring all of the possible pair-wise and three-way relationships, REFS grabs the most likely network fragments and assembles them into an ensemble of possible networks that are robust and consistent with the data. That's the reverse engineered part. Next comes forward simulation to predict outcomes when parts of each network are altered. This procedure allows researchers to score the probability that players in the ensemble of networks are important and to do so in an unbiased way across a large dataset.

Schadt agrees that such data-driven approaches are essential, and he uses them in his own work. But he says big data analytics covers a vast computational space ranging from bottom-up dynamical systems modeling to top-down probabilistic causal approaches—whatever approach (including hypothesis-driven), he says, "can derive meaningful information to aid us in understanding a disease condition or drug response or whatever the end goal is." Essentially, he says, it's not the approach that defines big data analytics, but the goal of extracting knowledge and ultimately understanding from big data.

Perakslis views the problem somewhat differently. "In order to get translational breakthroughs, you have to start out with an intentional design, which starts with intentional sampling," he says. "And to be honest, I don't think it works well yet hypothesis free." In a GWAS, he says, of course you look at everything because you don't know where to look. But the answers you seek can be lost in the noise.

> In a GWAS, Perakslis says, of course you look at everything because you don't know where to look. But the answers you seek can be lost in the noise.

Perakslis is more interested in broadening the types of data that are brought to bear in focused clinical trials. "Too often, when people make decisions, they are only looking at part of the story," he says. So, for example, if at the end of a Phase III clinical trial a drug doesn't produce the degree of success needed for approval, the database should be rich with information to help figure out why and where to go from there. TranSMART, a clinical informatics database that Perakslis helped assemble when he worked at J&J, does just that: It integrates different types of data into one location.

CLINICAL & PHARMACEUTICAL BIG DATA: ALREADY ABUNDANT

These days, for certain large healthcare organizations, large quantities of data simply accrue as an inevitable part of doing business. This is true of most hospitals, health maintenance organizations (HMOs), and pharmacy benefits managers (also known as PBMs). In these settings, "You really are getting at big data on the scales of LinkedIn, Amazon, Google, eBay and Netflix," says **Yael Garten**, **PhD**, a senior data scientist at LinkedIn who received her doctorate in biomedical informatics from Stanford University.

For example, Kaiser Permanente (KP), an HMO, has a 7 terabyte research database culled from electronic medical records, says **Joe Terdiman**, **MD**, **PhD**, director of information technology at Kaiser Permanente Northern California's Division of Research. That doesn't include any imaging data or genomics data. This special research database has been pre-cleaned and standardized using SNOWMED CT, an ontology of medical terms useful for research. "By cleaning and standardizing the data and making it easily accessible, we hope to do our research faster and more accurately," Terdiman says.

Medco, a PBM, accumulates longitudinal pharmacy data "because we are who we are and do what we do," Frueh says. As a large PBM that covers about 65 million lives in the United States, Medco manages the pharmaceutical side of the healthcare industry on behalf of payers. Their clients are health plans and large self-insured employers, state and governmental agencies, as well as Medicare. The company has agreements with some of these clients who provide large sets of medical claims data for research purposes. From the claims data, Medco can extract patient indications, treatments, dates of treatment, and outcomes (for example, whether the patient was hospitalized or not). Putting this multi-layered data together, Medco can search for associations between drug use, patient characteristics, and clinical impact (good, bad or indifferent) in order to determine whether a drug works the way it should.

And at Medco, big data analytics has already reaped dividends by uncovering drug-drug interactions. For example, clopidogrel (Plavix™) is a widely used drug that prevents harmful blood clots that may cause heart attacks or strokes. However, researchers were concerned that certain other drugs—proton-pump inhibitors used to reduce gastric acid production—might interfere with its activation by the body. Using their database, Medco looked for differences in two cohorts: those on one drug and those on the two drugs that potentially interact. The study revealed that patients taking both Plavix and a proton-pump inhibitor had a 50 percent higher chance of cardiovascular events (stroke or heart attack).

A similar study showed that antidepressants block the effectiveness of tamoxifen taken to prevent breast cancer recurrence. Patients taking both drugs were twice as likely to experience a recurrence.

"Both of these studies are prototypical of the kinds of questions we can ask in our database where we can correlate pharmacy data with clinical outcome data," Frueh says.

GOING HYPOTHESIS-FREE

Though Medco's outcomes are impressive, they have thus far relied on fairly straightforward statistical and epidemiological methods that were nevertheless quite labor intensive. "The hands-on analytics time to write the SAS code and specify clearly what you need for each hypothesis is very time-consuming," Frueh says. In addition, the work depends on having a hypothesis to begin with—potentially missing other signals that might exist in the data.

To address this limitation, Medco is currently working with Hill's GNS Healthcare to determine whether a hypothesis-free approach could yield new insights. So in the Plavix example, rather than starting with the hypothesis that proton-pump inhibitors might interact with drug activation, Frueh says, "We're letting the technology run wild and seeing what it comes up with."

> Using hypothesis-free approaches, Frueh says, "We're letting the technology run wild and seeing what it comes up with."

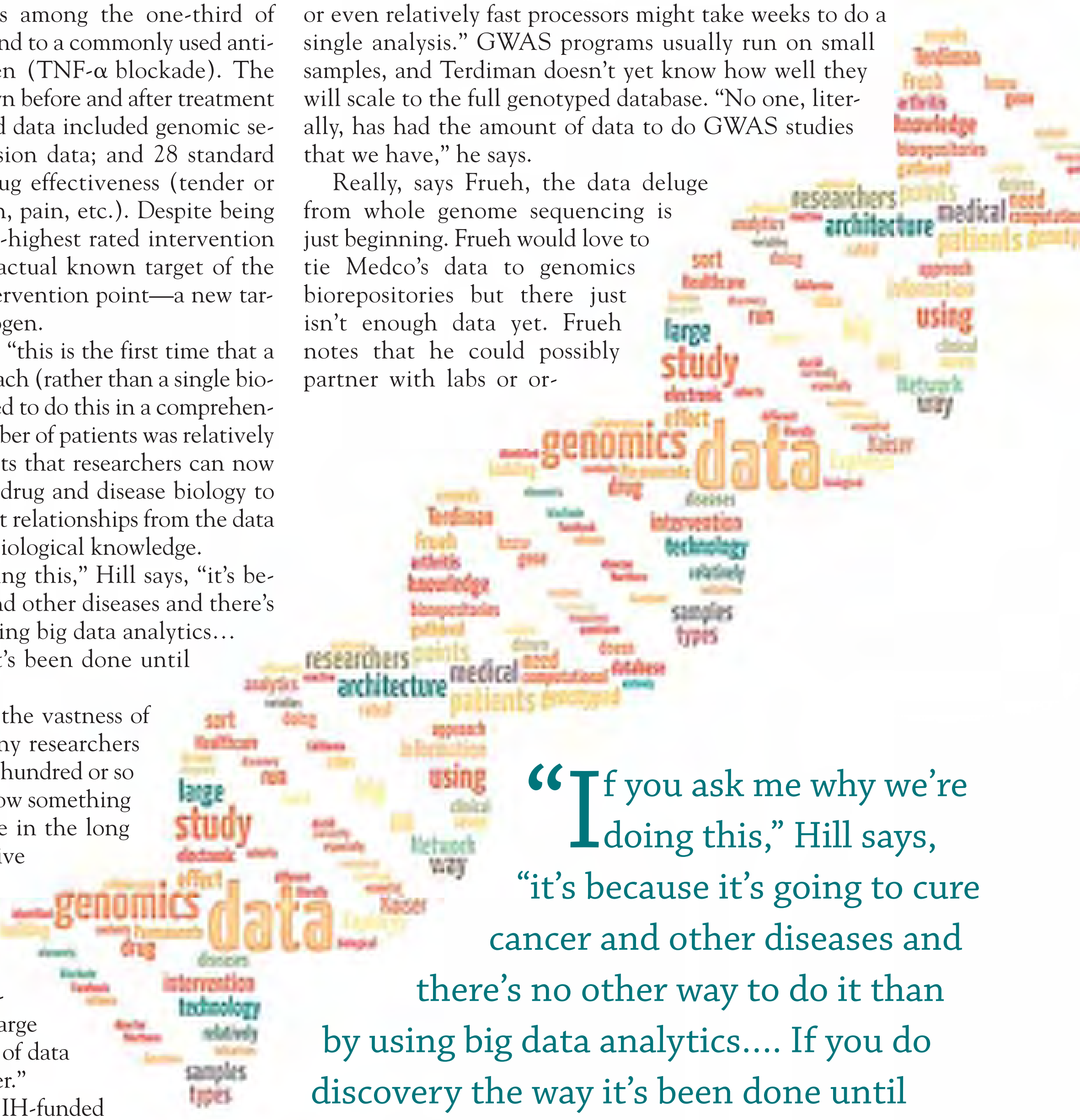
Because GNS Healthcare's REFS platform automates the process, he says, Medco can take the strongest signals from the data and avoid wasting time on hypotheses that don't lead to anything. Right now they are confirming whether the strongest findings identified by applying the REFS platform to the Plavix database actually hold up to more in-depth analysis.

ADDING GENOMICS TO THE MIX

The REFS platform developed by GNS Healthcare also functions in contexts that include genomic data. For example, in work published in *PLoS Computational Biology* in March 2011, GNS Healthcare and Biogen identified novel therapeutic intervention points among the one-third of arthritis patients who don't respond to a commonly used anti-inflammatory treatment regimen (TNF-α blockade). The clinical study sampled blood drawn before and after treatment of 77 patients. The multi-layered data included genomic sequence variations; gene expression data; and 28 standard arthritis scoring measures of drug effectiveness (tender or swollen joints, c-reactive protein, pain, etc.). Despite being entirely data driven, the second-highest rated intervention point they discovered was the actual known target of the drug. The first-highest rated intervention point—a new target—is now being studied by Biogen.

"To my knowledge," Hill says, "this is the first time that a data-driven computational approach (rather than a single biomarker approach) has been applied to do this in a comprehensive way." And although the number of patients was relatively small, Hill says, the study suggests that researchers can now interrogate computer models of drug and disease biology to better understand cause and effect relationships from the data itself, without reliance on prior biological knowledge.

"If you ask me why we're doing this," Hill says, "it's because it's going to cure cancer and other diseases and there's no other way to do it than by using big data analytics… . If you do discovery the way it's been done until now, it just doesn't cut it."



Today, rather than deal with the vastness of genomics data, Schadt says, many researchers distill it down to look only at the hundred or so gene variants they think they know something about. But this will be a mistake in the long run, Schadt says. "We need to derive higher level information from all of that data without reducing dimensionality to the most naïve level. And then we need the ability to connect that information to other large data sources such as all the types of data gathered by a large medical center."

The eMERGE Network, an NIH-funded collaboration across seven sites, is taking a running start at doing this. They are linking electronic medical records data with genomics data across seven different sites. Researchers will be able to study cohorts extracted from this "big data" without having to actively recruit and gather samples from a study population.

To a great extent, though, the eMERGE Network is still building its repository and confirming that it can repeat known results. The analytics are only now getting underway.

Kaiser Permanente, like the eMERGE network, is currently building what will be one of the largest biorepositories anywhere, with genotype data from 100,000 patients. "We hope to reach 500,000," says Terdiman of Kaiser Permanente.

But Kaiser is still sorting through what sort of platform to use for the data. They are looking at Hadoop—an up-and-coming open-source distributed-computing framework for storing and managing big data—as well as other possibilities. "With 100,000 patients genotyped, and each one has 700,000 SNPs, that's a pretty big matrix," Terdiman says. And then when you associate that with phenotypic data from the electronic medical record, he points out, "there's a combinatorial effect of all these variables such that simple or even relatively fast processors might take weeks to do a single analysis." GWAS programs usually run on small samples, and Terdiman doesn't yet know how well they will scale to the full genotyped database. "No one, literally, has had the amount of data to do GWAS studies that we have," he says.

Really, says Frueh, the data deluge from whole genome sequencing is just beginning. Frueh would love to tie Medco's data to genomics biorepositories but there just isn't enough data yet. Frueh notes that he could possibly partner with labs or organizations that have done large GWAS but, he says, unless you're asking the same questions as the GWAS, you won't get a lot of depth in those studies, especially after matching people to the pharmacy database. "You go from large to small numbers very quickly," he says.

> "If you ask me why we're doing this," Hill says, "it's because it's going to cure cancer and other diseases and there's no other way to do it than by using big data analytics.... If you do discovery the way it's been done until now, it just doesn't cut it."

Stephen McHale, CEO of Explorys, a big data bioinformatics company based in Cleveland Ohio, says that traditional relational data-warehousing technology can't efficiently handle the 30 billion clinical elements in their dataset. So Explorys implemented the type of data architecture that supports Google, Yahoo and Facebook. “Our technology is all column store and using MapReduce and those kinds of architectures,” he says, referring to approaches that use large numbers of computers to process highly distributable problems across huge datasets. He says that, to his knowledge, it’s a first in the medical space. “We needed that sort of architecture to support this much data.” And with genomics coming their way, it seems even more essential to use these types of architecture, McHale says. Explorys is now working on some pilot initiatives to integrate genomic data with observational data.

MAKING BIG DATA ACTIONABLE

Extracting knowledge from big data is a huge challenge, but perhaps a greater one is ensuring that big data infrastructure will form the backbone of an effort to push Amazon.com-style recommendations to practitioners and patients.

Garten notes that implementing an Amazon or LinkedIn style recommendation system in biomedicine will be tough. Such systems use machine learning and natural language processing to, in a sense, bucket customers into groups. “But the ability to bucket people together is harder in biomedicine,” Garten says. “The slightest variations can matter a lot in terms of how we metabolize drugs or respond to the environment, so the signal is harder to find.” The stakes are also higher for getting a false result.

But Medco’s experience suggests such bucketing is already possible, at least to some extent. For example, in the Plavix example described above, Medco was in a position to immediately effect a change: “We can pull a switch and say that each and every pharmacist on our list needs to be told about this,” Frueh says. After implementing the rule, Medco saw a drop of about one third in co-use of the interacting drugs. “This is one example where the use of big data in this stepwise process has cut down on the time it takes to get changes into clinical practice,” Frueh says.

“This is one example where the use of big data in this stepwise process has cut down on the time it takes to get changes into clinical practice,” Frueh says.

In another example, Medco was able to use its infrastructure to increase uptake of a genotyping test for warfarin dosing. First, however, they had to show payers that the test was cost-effective. In a clinical trial conducted in collaboration with Mayo Clinic, Medco showed that genotyping reduced the rate of hospitalizations among warfarin-dosed patients by 30 percent. Armed with that information, payers became supportive of Medco reaching out to physicians to suggest they use the genotyping test before prescribing warfarin. Because of Medco’s big data infrastructure, this outreach could be easily accomplished: Each time a physician prescribed warfarin, a message was routed back through the pharmacy to the physician, suggesting use of the test. The result: an increase in uptake of the test from a rate of 0.5 percent or so

in the general physician population up to approximately 20 to 30 percent by physicians in the network.

"This has to do with creating an environment and the operational infrastructure to be proactive," Frueh says. And Frueh suspects that uptake of the test will continue to grow. "We're probably at the beginning of what we hope will be a hockey-stick shaped uptake of this test." The lesson: Big data, and the connectedness of big data to the real world, provides the opportunity to take advantage of teachable moments at the point of care.

"The practices and experience from the corporations with large amounts of data (i.e., LinkedIn, Amazon, Google, Yahoo) will propagate back to the academic and research setting, and help accelerate the process of organizing the data," Yael Garten says.

As we go from data generation to knowledge about what it means, to making that knowledge actionable, Schadt says, "It will impact clinical decisions on every level."

PLAYING CATCH UP AND THEN SOME

To some extent, big data analytics in biomedicine lags finance and commerce because it hasn't taken advantage of commercial methods of handling large datasets—like Hadoop and parallelized computing. "These allow data analytics in an industry-level manner," Garten says. "That's something that LinkedIn, Amazon and Facebook have already nailed, and bioinformatics is lagging behind those industries."

Bioinformatics researchers still spend a lot of time structuring and organizing their data, preparing to harvest the insights that are the end goal, says Garten. By contrast, the private sector has completed the phase of structuring and collecting data in an organized fashion and is now investing more and more effort toward producing interesting results and insights. Eventually, Garten says, "the practices and experience from the corporations with large amounts of data (i.e., LinkedIn, Amazon, Google, Yahoo) will propagate back to the academic and research setting, and help accelerate the process of organizing the data."

At the same time, bioinformatics actually has something to offer the broader world, Garten says. She and others with a bioinformatics background who have moved into other arenas bring to the table an ability to handle messy data that is often incomplete. The expertise in integrating various datasets in creative ways to infer insights from this data, as is done in translational bioinformatics, is useful for extracting business insights in other industries.

Hill also sees biomedical approaches filtering outward. "REFS is data-agnostic," he says. It can work on genomic data as easily as clinical data—or, for that matter, financial data. Hill's company recently created a financial spinoff called FINA Technologies. He also spun off Dataspora, which is focused on consumer ecommerce. "We've created a technology that goes all the way from unraveling how cancer drugs work to predicting financial markets," Hill says. "This technology is applicable to how complex systems work in different industries, and there's something profound about that." □

PRIVAT
you already received for your condition? ☑ Medicatio
□ None □ Other
you for your conditi
Spinal X-Ray
Chest X-Ray
3/2009
MRI, CT Scan
if you have had any of the follow
□ Yes ☒ No
□ Yes ☒ No
□ Yes ☒ No
☒ Yes □ No
□ Yes ☒ No
□ Yes ☒ No
□ Yes ☒ No
□ Yes ☒ No
□ Yes ☒ No
□ Yes ☒ No
Whooping Cough □ Yes
Other
Arthritis □ Yes ☒ No
Packs/Day 0
Drinks/Week 5
Cups/Day 2
Reason
Date
1/2006
2, 1/2008
MINERALS

By Alexander Gelfand

Privacy & Biomedical Research: Building a Trust Infrastructure

Trust. It's the basis of every patient/physician interaction: Shared personal health information is kept confidential and used only for the patient's benefit. It's a tradition that started before the time of Hippocrates, endured through the era of records stored in filing cabinets, and persists today as we move to electronic patient records. And it's codified in the form of HIPAA, the federal Health Insurance Portability and Accountability Act, which ensures the privacy of health records.

But as sensitive personal health data accrue in ever larger databases, concerns over privacy breaches are on the rise. And as researchers perceive the potential usefulness of this vast data trove, they seek strategies to access it without violating HIPAA. In response, data privacy experts are developing ever more sophisticated methods to protect electronic health data from unwanted exposure. And while many of these experts have raised alarms about the vulnerabilities of the privacy-protection schemes currently in place, they have also begun talking about the possibility of implementing far more powerful technologies in the near future.

"This is the start of the golden age of privacy research," says **Dan Kifer, PhD**, a computer scientist at Pennsylvania State University who has investigated privacy-preserving techniques for applications ranging from biomedical research to the U.S. Census.

Privacy Fears Drive Innovation

The rapid progress taking place in privacy research in the biomedical arena is driven in large part by fear—namely, fear that the vast warehouses of biomedical data now being assembled could be vulnerable to the same kinds of privacy

NAME: [redacted] DOB: [redacted]

De-identification protocols suppress or modify bits of data that might allow an attacker to determine precisely to whom a particular record belongs.

breaches that have in recent years plagued such information aggregators as Google and Facebook.

Granted, the evidence of such breaches in the realm of medical records, clinical data, and genomic information remains slim. The most alarming incidents to date have involved simple failures of security, or of access control, such as the theft or loss of unsecured computers containing electronic medical records, rather than the unintentional leakage of sensitive information from large biomedical databases; and as yet, no one has reportedly been harmed by the unauthorized release of their biomedical information. Instead, the most impressive privacy breaches to date have been perpetrated by academic researchers who were trying to find weaknesses in the systems they were attacking: identifying the medical records of a particular individual in a hospital system, for example, or identifying participants in a genome-wide association study (GWAS) designed to link particular diseases to specific genetic variations.

> "This is such a critical piece of the puzzle, that we need to address privacy concerns before we plan for other activities," says **Lucila Ohno-Machado**.

Yet anxiety over the possibility of more public, and more harmful, privacy breaches continues to build, the principal concern being that insurers and employers might use biomedical data to discriminate against policyholders and employees. Such fears are not unfounded. "Insurers have historically used data to make coverage determinations," says **Deven McGraw**, director of the health privacy project at the Center for Democracy and Technology, a nonprofit public interest group in Washington, DC. **Carl Gunter, PhD**, a computer scientist at the University of Illinois who studies the health information exchanges that enable hospitals to share electronic medical records, emphasizes the "dreadful risks" posed by medical identity theft, in which one person assumes the identity of another when seeking medical care, and the medical histories of both victim and thief become dangerously entangled. And there is rising concern over the privacy risks associated with genomic data in particular. As **Brad Malin, PhD**, director of the Health Information Privacy Lab at Vanderbilt University, points out, genomic data is highly distinguishable, extremely stable, and can in certain situations be used to predict the likelihood that an individual might fall prey to this disease or that one—information that could be used to deny coverage or a job.

Some of these scenarios might seem unlikely at the present time. It's doubtful, for example, that many people outside of a university computer science department would have the technical wherewithal to pick a single individual out of the mass of statistics associated with a GWAS. But technological progress has a way of closing the gap between the improbable and the probable. "There's nothing to say that what's unreasonable now won't be unreasonable in the near future," Malin says. And even the smallest risk of a privacy violation can be enough to scare a patient away from participating in a clinical trial, or persuade an institution to withhold data from researchers due to ethical or legal concerns. According to Gunter, some health information exchanges have already prohibited the sharing of electronic medical records for research purposes. "There is such fear that we need to address it before we can make full use of this data for research purposes," says **Lucila Ohno-Machado**, **PhD**, founding chief of the division of biomedical informatics and associate dean for informatics at the University of California at San Diego (UCSD) and principal investigator for iDash (integrating Data for Analysis, anonymization, and SHaring), a National Center for Biomedical Computing.

All of this unease—over the privacy rights of individuals, over potential discrimination, and over the chilling effect that privacy concerns can have on research—has prompted a great deal of innovation amongst the mathematicians, cryptographers, and computer scientists who are working to develop mechanisms that will allow researchers to analyze biomedical data without compromising privacy.

Data-Driven Privacy Measures

Staal Vinterbo, **PhD**, a computer scientist in the division of biomedical informatics at UCSD, distinguishes between two broad classes of privacy-protecting mechanisms currently under development: "data-driven" mechanisms that define privacy in terms of the data itself, and "process-driven" mechanisms that define privacy in terms of how they access the data. Both are intended to let researchers perform meaningful analyses without disclosing sensitive personal information, but they stem from very different concepts of data privacy, and they often work via very different means. The most common data-driven approaches, for example, modify raw data so that it can be released without revealing sensitive information, while most process-driven approaches leave the underlying data alone and instead build privacy protections into the algorithms they use to extract it. Data-driven mechanisms came first, but process-driven ones may offer better protection—albeit at a price.

De-identification, or anonymization, is the most commonly employed privacy measure, and one that lies very much on the data-driven side of the divide. Rather than freely sharing all of the data in a group of records, de-identification protocols suppress or modify the bits that might allow an attacker to determine precisely to whom a particular record belongs. Some of these protocols, especially the more experimental ones, can be quite sophisticated. Spectral anonymization, which Vinterbo has investigated with **Thomas Lasko**, **MD**, **PhD,** a researcher at Vanderbilt, manipulates data in mathematically complex ways so that useful correlations can be maintained for research purposes even after the information has been randomized to prevent specific individuals from being identified. Synthetic data generation, which Kifer has explored, creates new data that statistically mimics the real stuff, but shields the actual participants in the original data set. But the anonymization schemes that are most often used to de-identify electronic health records in the real world simply delete or truncate specific data fields containing identifiable information like proper names and ZIP codes.

The advantage of de-identification is that it allows analysts to examine the raw data itself, albeit in altered form, rather than running queries against it from behind a privacy-preserving interface. The disadvantage is that it does not always work.

The first and most widely publicized demonstration of the weaknesses of de-identification occurred in 1997, when **Latanya Sweeney**, **PhD**, linked the supposedly de-identified health records released by the Massachusetts state insurance commission to the state's voter-registration rolls and re-identified the personal medical records of then-Governor William Weld. (Sweeney, who was a graduate student at MIT at the time, is now director of the Data Privacy Lab at Harvard University.)

Sweeney's successful re-identification attack helped prompt the adoption of the HIPAA Privacy Rule in 2000. The Privacy Rule imposes restrictions on the release of "individually identifiable health information." These federally legislated constraints on disclosure are waived, however, if the data has been de-identified by applying the so-called "safe harbor" method, which involves removing 18 identifiers, including names, dates, and Social Security numbers. Since data that has been de-identified under the safe harbor method is no longer considered to be individually identifiable, it is no longer covered by the Privacy Rule, and can be freely shared.

Yet there is a growing sense among data privacy experts that no form of de-identification will ever be good enough to meet the highest standards of privacy pro-

> **Staal Vinterbo, PhD,** distinguishes between two broad classes of privacy-protecting mechanisms: "data-driven" ones that "perturb" or modify the data in some way so that it can be released without revealing sensitive information; and "process-driven" ones that leave the underlying data alone and instead build some kind of privacy protection into the algorithms that are used to analyze it.

> There is a growing sense among data privacy experts that no form of de-identification will ever be good enough to meet the highest standards of privacy protection, and that the entire approach will only grow less reliable over time.

Differential privacy is achieved by introducing some random noise into the query responses. An analyst can only see the blurry answers provided by the algorithms, never the raw data itself.

tection, and that the entire approach will only grow less reliable over time. The reason for this is simple: as more and more data is collected and stored about us all—in online databases and on social networking sites, in publicly available government repositories and elsewhere—it becomes easier and easier to launch the kind of linkage attack that allowed Sweeney to re-identify William Weld's medical records. As the computer scientists **Arvind Narayanan**, **PhD**, and **Vitaly Shmatikov**, **PhD**, noted in a 2010 article in *Communications of the* ACM (Association for Computing Machinery), "any attribute can be identifying in combination with others." In other words, no matter how many fields are deleted from an individual's record, as long as there is something left for researchers to work with, there will also be enough left for re-identification.

Moreover, because de-identification techniques have for the most part been designed to protect specific kinds of data from specific kinds of attack, they lack the flexibility needed to deal with a rapidly changing data landscape. The task of anonymization will only become harder, for example, as more and more categories of information, from handwritten clinical notes to genetic sequences, are gathered and stored in digital repositories. The *ad hoc* nature of such data-driven methods also means that they tend to lack a rigorous mathematical basis; and by their very nature they can only access a very limited amount of data. As a result, it can be difficult to quantify just how much privacy protection they truly offer. Scientists can only estimate their efficacy by trying to break them—thereby proving their limitations, but not their strengths.

This is not to say that de-identification and other data-driven approaches do not have their uses. Malin and his colleagues at Vanderbilt, for example, have for several years used de-identification strategies such as k-anonymization to help protect the privacy of electronic medical records used for research purposes. "We have de-identified more than 1.5 million medical records from the Vanderbilt University Medical Center," he says. K-anonymization works by suppressing or modifying enough data to make a certain number of records (k) in a database appear to be identical. If done appropriately, the data can still be used for research, but the individuals who provided it become lost in a crowd of look-alikes. Several data privacy experts have pointed to flaws in k-anonymization—for example, an attacker who possesses sufficient background knowledge about someone can break k-anonymity and re-identify that individual– but the risks of re-identification remain low. And like any de-identification scheme, k-anonymization allows researchers to examine the raw data.

"Eventually, all of us realized that this is just a never-ending cycle: find a way of perturbing the data, find weaknesses, try to fix them, find more weaknesses, try to fix them," says Kifer.

Process-Driven Privacy Measures

Still, many scientists have begun moving away from data-driven approaches and toward more mathematically rigorous process-driven ones. "Eventually, all of us realized that this is just a never-ending cycle: find a way of perturbing the data, find weaknesses, try to fix them, find more weaknesses, try to fix them," says Kifer. While a graduate student at Cornell University, Kifer helped develop a more robust refinement to k-anonymity known as l-diversity—a refinement that was soon shown to have flaws of its own. Recognizing the apparent limitations of de-identification in general, he and his fellow researchers began seeking a different path: one that involves devising ways of querying statistical databases while providing privacy guarantees that can be expressed as mathematical and statistical statements.

The first and still the most promising of these approaches, differential privacy, was proposed in 2006 by **Cynthia Dwork**, **PhD**, and colleagues at Microsoft Research, Ben-Gurion University, and the Weizmann Institute of Science. Dwork recognized the cycle described by Kifer from the history of cryptography. She also knew that modern cryptographers only liberated themselves from that cycle when they developed formal, provable definitions of information security that could be quantified. So Dwork and her collaborators did the same in the realm of privacy, formulating a mathematical definition of the concept that amounts to a promise to the data subject that his life will not, in Dwork's words, "change substantially for the better or the worse as a result of a computation on the data." Precisely how that is achieved is more or less up for grabs; any solution that satisfies the basic definition, which in its true form looks more like a mathematical proof than a verbal guarantee, will necessarily be differentially private.

Scientists like Dwork and Kifer are still working out how to implement differential privacy in the real world, and few applications have moved beyond the lab. In general, however, differential privacy is achieved by writing special algorithms that sit between a statistical database and an analyst who wishes to run queries against it. If an algorithm is differentially private, then the results it produces should be essentially the same independent of whether any single person is included in the database or not. One important consequence of this is that no matter what an analyst knows—no matter what background knowledge they might possess—they still cannot learn anything more about a specific individual just because they happen to be in the database. Conversely, even if an analyst were to know everything about each individual represented in the data except for one, they still should not be able to learn much about that one remaining person. No de-identification scheme can make those kinds of guarantees.

If an algorithm is differentially private, then the results it produces should be the same regardless of whether any single record-holder is included in the database or not.

In practice, differential privacy is achieved by introducing some random noise into the query responses. For example, if an analyst were to ask how many people in a database were over 5 feet tall, and the true answer was 56, then a differentially private algorithm might grab a random variable from a probability distribution and add it to the true answer, spitting out 57 instead. The noise is the difference between the response (57) and the true answer (56). "Our choice of randomness," Dwork writes in an e-mail, "makes responses close to the truth much more likely than answers that are far from the truth (which is what we want for accuracy)."

Nonetheless, attentive readers will have noticed that differential privacy does in fact work by providing slightly inaccurate results; as Dwork says, it uses probability to introduce "a little bit of uncertainty." This has two significant consequences.

First, in a differentially private setting, an analyst can only see the blurry answers provided by the algorithms; he can never examine the raw data itself. Dwork and several colleagues are currently investigating the possibility of allowing trusted individuals to view the underlying data—a situation Dwork describes as "differential privacy with a human in the loop"—but that is still very much under development. At least for now, researchers who need to see the innards of the data sets they are working with must look elsewhere for privacy protection.

Second, the question of how much noise is enough noise, and how much noise is too much noise, is a rather thorny one. The trick is to add just enough randomness to the query answers to protect the privacy of the individuals whose records lie in the database, but not so much that an analyst can no longer learn accurate or meaningful things from a statistical perspective about the sample population they comprise. This is the price of privacy, or the trade-off between privacy and utility; and it may be the most serious challenge facing those who are trying to bring differential privacy out of the lab. "We can always find wildly inaccurate ways of computing something that ensures a given level of privacy," says Dwork. The science lies in finding ways of ensuring privacy that do not destroy utility.

"We can always find wildly inaccurate ways of computing something that ensures a given level of privacy," says Dwork. The science lies in finding ways of ensuring privacy that do not destroy utility.

As it turns out, some queries, and some databases, are more "sensitive" than others, meaning that they are more prone to leak information. As a result, they require more noise. According to Vinterbo, more fine-grained infor-

NAME: [redacted] DOB: [redacted]

mation also requires more noise—a situation that may prove challenging in the case of genomic databases, which contain enormous amounts of incredibly detailed data. Since you wouldn't want to add more noise than is absolutely necessary, matching the appropriate amount of noise to the sensitivity of the query and of the database—in effect, figuring out how to balance privacy against utility—is crucial. And as **Kamalika Chaudhuri**, **PhD**, an expert on machine learning at UCSD, says, it also turns out to be "fairly technical and complicated."

Which is not to say that it can't be done. A number of researchers are investigating ways of relaxing differential privacy so that it still offers strong privacy protection without requiring excessive amounts of noise, while others are trying to find novel methods of adding noise that won't degrade accuracy.

Chaudhuri, for example, is interested in using algorithms called "classifiers" that can be used to trawl through large collections of medical records in order to predict things like whether a particular individual might require hospitalization. Classifiers must be trained on standard data sets, however, and the training process can leak sensitive information about the training samples. A differentially private approach would typically involve adding a bit of noise to the results coming out of the classifier—a technique known as "output perturbation." This protects privacy, but also makes the classifier more error-prone. Chaudhuri has figured out a way to insert the noise earlier in the process, injecting it into the classifier itself—a technique she calls "objective perturbation." The latter still ensures differential privacy, but the results are more accurate.

Efforts like these bode well for the adoption of differential privacy in the coming years. But even its supporters agree that differential privacy alone cannot be counted upon to solve the privacy problem once and for all. "There is no single solution that will suit every possible scenario," says Vinterbo. In a recent paper, Kifer pointed toward a few specific weaknesses of differential privacy, most notably some limitations on its ability to protect privacy in social networks and in circumstances where some statistics have already been released into the wild. "Differential privacy works," Kifer says. "But nothing works all the time."

"Differential privacy works," Kifer says. "But nothing works all the time."

Finding Integrated Solutions

As a result, many experts are beginning to envision a more integrative and contextual approach to biomedical data privacy—one that would offer a menu of technical solutions backed up by policy measures, the precise mixture of which would depend on the nature of the data, the needs of the researchers, and the concerns of the data subjects themselves.

Haixu Tang, **PhD**, and **XiaoFeng Wang**, **PhD**, at the Indiana University Bloomington School of Informatics and Computing, advocate for what they call a "hierarchical method of data release" that would consider the kind of analysis researchers wish to perform on a particular data set, the level of privacy risk involved, and the degree of utility required before deciding on a particular privacy mechanism. (The two recently won the 2011 Award for Outstanding Research in Privacy Enhancing Technologies for their work demonstrating that individuals could be identified in a GWAS even when the precision of the published statistics was low and some of the data were missing. They are currently investigating ways of introducing miniscule amounts of noise in order to guard against such attacks without sacrificing utility.)

Vinterbo, for his part, thinks that a comprehensive solution to the privacy problem will require a "trust infrastructure" that includes not only technical solutions, but also "legal frameworks that efficiently combine technology and law." "The needs that will be met by technical measures alone are a minority," he says.

Vinterbo thinks that a comprehensive solution to the privacy problem will require a "trust infrastructure" that includes not only technical solutions, but also "legal frameworks that have some actual teeth."

Similarly, Malin would like to see a holistic, risk-based approach that draws on the pooled expertise of technologists, legal experts, and ethics review boards, all of whom would have a say in determining how best to safeguard privacy in a particular context—whether that meant implementing the most rigorous technical scheme possible, or applying something less formal and backing it up with carefully crafted use agreements and legal sanctions. Only then, he believes, will the biomedical community have the kind of flexible, nuanced tools needed to address the challenges of protecting its data.

"We have developed great technical solutions, and more are coming down the pipeline," says Malin, echoing Kifer's prediction of a golden age. "But we have to keep the bigger picture in mind." □

BY OREN FREIFELD, JUSTIN FOSTER AND PAUL NUYUJUKIAN

Calculating Statistics of Arm Movements

Suppose 20 friends live in the same city and want to meet for dinner. They should be able to identify a unique spot that minimizes the squared distance everyone needs to travel by taking the arithmetic mean of each starting location. However, if these 20 friends were spread across the world rather than in one city, the mean of all the starting locations would be in the interior of the Earth! In the latter example, where we cannot approximate the friends' locations in a 2-D plane, it is necessary to impose a geometry constraint: the meeting spot must be on the 2-D surface of Earth, a subset of the 3-D world. Since the arithmetic mean does not incorporate geometric constraints into its calculation, it yields a nonsensical answer.

Geometric constraints influence even very simple calculations and arise in many contexts. In a biomechanics lab, researchers take measurements of limb movements during reaching and walking. In the same way that individuals cannot have arbitrary locations in the 3-D world (lest they be found inside Earth), limbs cannot have arbitrary positions. Limbs have bones that have fixed sizes and joints that can only rotate in certain directions, creating the "space of rotations." Think of this space as a low-dimensional surface embedded in a higher-dimension Euclidean space, similar to how the 2-D surface of Earth is embedded in a 3-D world. Classical statistical computations don't make sense in this situation. However, having a mathematical framework that can calculate quantities similar to the arithmetic mean under these inherent geometric constraints would be very useful.

Geometric constraints influence even very simple calculations and arise in many contexts. In a biomechanics lab, researchers take measurements of limb movements during reaching and walking.

One solution is to choose a valid meeting spot or arm pose that also minimizes the distance to the calculated arithmetic mean. This point, called the exterior mean, is not necessarily unique, but is within the set of points that satisfy the geometric constraints and are closest to the arithmetic mean. For instance, if the average meeting spot were the exact center of the earth, then Rome, San Francisco, and Tokyo would all be exterior means. In order to calculate the exterior mean, we first can find the mean of all locations and then project it onto the set of all points that meet the geometric constraints. Depending on the details of the geometric constraints, the exterior mean might be found analytically or through iterative methods.

One issue with the exterior mean is that it ignores the route taken from the starting points, so it could produce a non-optimal solution. In our earlier example, the distance a friend travels to the exterior mean location may turn out to be shortest only if he walks through the center of Earth, since the calculation doesn't account for the path the friend takes to get there.

Measurements of distances along paths as dictated by the constrained geometry of a surface are called geodesic distances, and a better analog of the arithmetic mean should minimize the sum of the squared geodesic distances. This solution is referred to as the interior mean. For our restaurant example, the interior mean would be the location such that the total squared distance each friend travels along the Earth's surface is minimized.

For the arm movement problem, the interior mean would identify an arm orientation that minimizes the average squared distance of the path (measured in the space of rotations—recall our surface analogue) that all other arm poses must go through to get to that pose.

Finding the geodesic distance between two points can be difficult in many geometries. Fortunately for the study of limb movements, which are geometrically constrained to rotations, there exists a simple and elegant iterative algorithm with rapid convergence towards the interior mean. More complicated statistics can be computed similarly based on geodesic distances. □

DETAILS

Oren Freifeld, a graduate student in Michael Black's Vision lab at Brown University, is an exchange scholar working in Krishna Shenoy's Neural Prosthetics Systems Laboratory at Stanford University. Paul Nuyujukian and Justin Foster are graduate students in Shenoy's lab. For additional information see Karcher, H, Riemannian center of mass and mollifier smoothing, *Communications on Pure and Applied Mathematics*, 30:509–541 (1977).

DIVERSE DISCIPLINES, ONE COMMUNITY
BiomedicalComputation REVIEW
Published by Simbios, an NIH National Center for Biomedical Computing

Stanford University
318 Campus Drive
Clark Center Room S231
Stanford, CA 94305-5444

seeing science

SeeingScience

BY KATHARINE MILLER

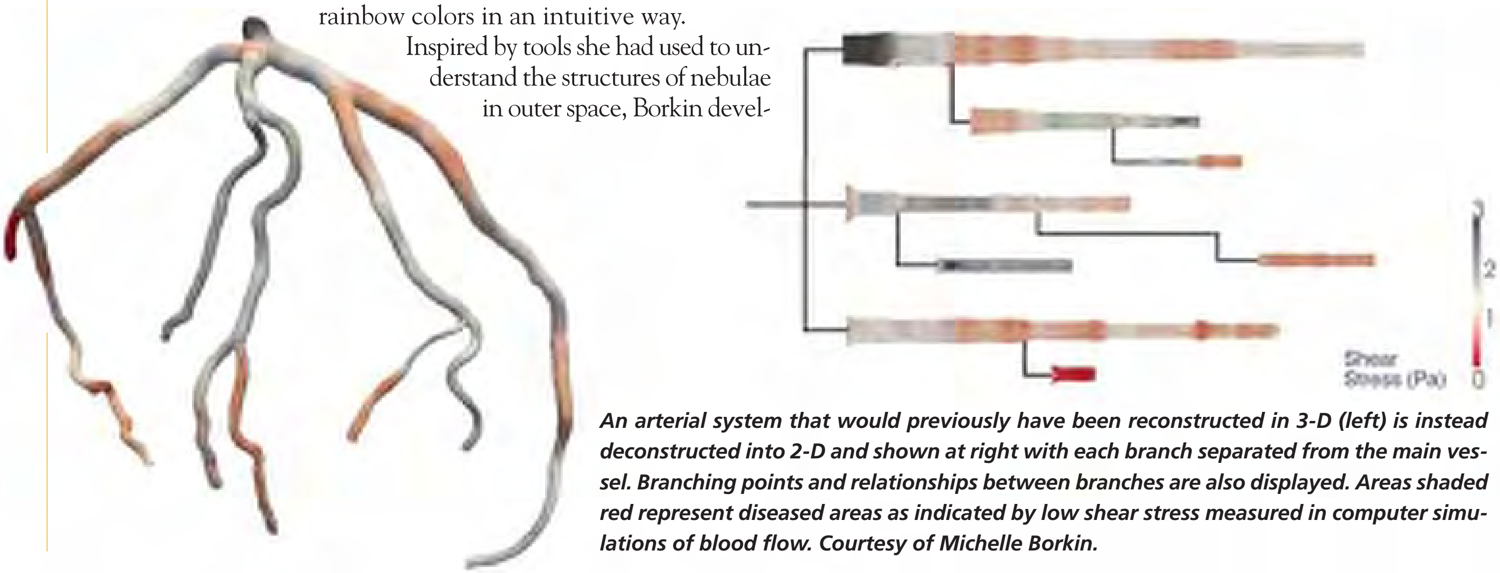
Busting Assumptions about Rainbows and 3-D Images

To diagnose heart disease noninvasively, scientists combine 3-D visualizations of the heart and blood vessels (reconstructed from CT scans) with computer simulations of blood flow. Typically, a palette of rainbow colors is used to help identify areas of low shear stress—trouble spots of low friction or stagnant blood flow that weaken vessel walls—a good indicator of disease progression. But 3-D rainbows aren't as useful as our instincts suggest, says **Michelle Borkin**, a PhD candidate in applied physics at Harvard University.

After observing and interviewing cardiologists, Borkin realized that interacting with and rotating 3-D images took time and sometimes meant interrupting a procedure. And research into the psychology of visualization suggests that humans do not read rainbow colors in an intuitive way.

Inspired by tools she had used to understand the structures of nebulae in outer space, Borkin developed software called HemoVis that visualizes simulated blood flow in two dimensions, splays or "butterflies" vessels open in a tree diagram, and colors the areas of low shear stress in gradations from red to gray. In a test of the software, the 2-D visualizations (compared with 3-D) led to much more accurate and efficient diagnoses by 21 medical students; the same was true for the red/gray palette when compared to rainbow. "At a single glance," Borkin says, "they get a quick and accurate diagnosis." The work was published in *IEEE Transactions on Visualization and Computer Graphics*.

"This paper shows that making smart choices about how you display your data in dimensionality and color not only can help doctors see the data better and help them make discoveries," Borkin says, "but might also save lives." □



An arterial system that would previously have been reconstructed in 3-D (left) is instead deconstructed into 2-D and shown at right with each branch separated from the main vessel. Branching points and relationships between branches are also displayed. Areas shaded red represent diseased areas as indicated by low shear stress measured in computer simulations of blood flow. Courtesy of Michelle Borkin.