NIH Launches A United Ecosystem FOR BIG DATA

By Katherine Miller

By taming big data, the BD2K ecosystem will enable a deeper understanding of the human organism while at the same time motivating major improvements in the practice of medicine.

Francis Collins, **MD**, **PhD**, Director of the National Institutes of Health (NIH), says he used to feel "data envy" toward the field of physics. In those days, "no one would have predicted that biology would emerge as the biggest challenge in terms of data. But that is now the case."

Last year, under Collins' leadership, the NIH stepped up to take on that challenge by announcing the Big Data to Knowledge (BD2K) program, a wide-ranging plan to enhance biomedical researchers' ability to make effective use of big data.

"The goal is to begin establishing an ecosystem that supports tools, data and best practices for this new expanded way of doing biomedical research," says **Philip Bourne**, **PhD**, NIH Associate Director for Data Science.

The first step toward achieving that goal was the announcement last September that the NIH is establishing 12 BD2K Centers of Excellence, granting each center approximately $2 million a year for four years ($24 million/year total).[1]

At the time of this writing, the Centers are just getting started. But interviews with the principal investigators (PIs) about their data science goals reveal the Centers' potential to alter the landscape of big data science in biomedicine.

Imagine: All health information across an individual's lifetime accessed through a single system and layered with data about lifestyle and environmental exposures; wearable sensors continuously tracking patients' health status and allowing remote interventions that are both effective and reliable; physicians predicting, with a few tests and a few clicks of a mouse, what treatments are appropriate for a specific patient based on his or her unique genetic make-up; and the cooperative analysis of neuroimages worldwide to generate an exponential increase in our understanding of the brain. These imaginings are all part of the future envisioned by the BD2K Centers and supported by the NIH.

But before these visions can become a reality, the Centers have their work cut out for them. They need to find ways to make effective use of the big data that continues to flow from biomedical labs and high-throughput experiments, including patient records, genomics and other –omics data, imaging data, and data from mobile devices and wearable sensors. The Centers are charged with integrating vast amounts of data, connecting data to knowledge, and developing new statistical and analytical approaches that work well with big data. And let's not forget the nuts and bolts of standardizing data, collecting better metadata and building pipelines to bring all of this to bench biologists.

Wisely, NIH has directed that the data science goals be achieved and validated in a biomedical research milieu. "For methods to be broadly applicable, they need to be developed in the context of a particular question," says **Scott Delp**, **PhD**, professor of biomedical engineering and PI of the BD2K-funded **Mobilize Center**. For example, several centers will determine whether wearable sensor data, collected for 24 hours seven days a week, can be used to improve patient health by motivating exercise, detecting when former smokers relapse, or reducing hospital admissions for congestive heart failure.

By taming big data, the BD2K ecosystem will enable a deeper understanding of the human organism while at the same time motivating major improvements in the practice of medicine.

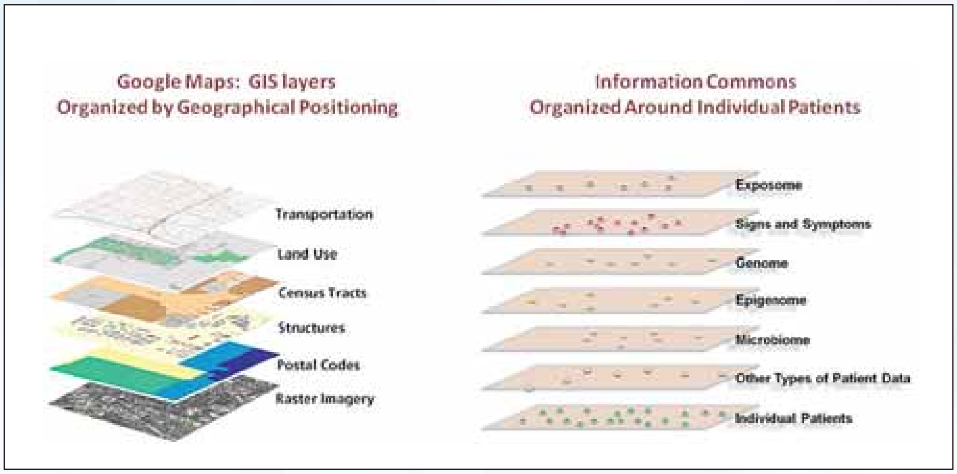
"My view," says **Santosh Kumar**, **PhD**, associate professor of computer science at the University of Memphis and PI for the **Center of Excellence for Mobile Sensor Data-to-Knowledge** (MD2K), "is that after four years, it becomes possible for any researcher to use all of these tools collectively to get a holistic view of the person they are studying." Biomedical discovery will no longer be siloed in any particular data source but instead available to anyone with a computer. "With that, the true power of BD2K will be realized," Kumar says.

PATIENTS AND BIG DATA: *Going Holistic*

Building a Patient-Centered Coordinate System

To gain a holistic view of patients, physicians and researchers need health data to be better organized and more readily accessed. The new BD2K-funded Patient-Centered Information Commons (PIC) envisions a Google maps-like layering of data, with patients as the essential coordinates.

The idea sprang from a National Acad-



PIC proposes developing a Patient-Centered Information Commons (right panel) that is analogous to a layered geographic information system (left panel). This integrated system would include clinical information from electronic medical records as well as genomics, proteomics, and environmental context, thus enabling a much more comprehensive characterization of disease states and health states. Reprinted with permission from **Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease.** ***Washington, DC: The National Academies Press, 2011.***

[1] Additional BD2K initiatives announced at the same time bring the total to $32 million per year for four years. Watch this magazine for future articles about these projects.

emies of Science report "Toward Precision Medicine," says PIC PI **Isaac Kohane**, **MD**, **PhD**, professor of pediatrics at Harvard Medical School. The system would include clinical information from electronic medical records as well as genomics, proteomics, and environmental context—the so-called exposome.

"If we could bring all these disparate data together across a lifetime, we'd have a much better sense of what informs disease state and health state," Kohane says. "That's the information commons. And our goal is to make that a pragmatic reality so others can explore it."

The challenges are numerous. Lacking a national health identifier, it's difficult to determine what medical data belong with a single patient, let alone how to layer that with non-health sources of data, such as shopping behavior, local pollen counts, pollution indices, or social media information, Kohane says. Medical record numbers uniquely identify patients within particular healthcare environments, but distinguishing one John Smith from another across a claims database (without medical record numbers) is another matter. A few identifying characteristics may allow probabilistic assignments of data to specific people, but that becomes non-trivial as the data get farther and farther apart, Kohane says. "All these data are heterogeneous, sparse, biased and noisy," he says. "So how you actually clean them up in a way that can use our standard tool kits against them is an open question."

PIC plans to start by building a virtual sandbox with data uploaded to a secure cloud service. "That will allow us to start playing the games of finding common coordinate systems across the data types," he says.

PIC's first focus will be neurodevelopment, for which several large hospitals have committed genomic and clinical data. The center will layer the data together as a way to not only achieve the center's goals but also answer questions for the neurodevelopment community. The success of a center like this, Kohane says, "lies not just in developing a widely adopted architecture, but in using the architecture to do interesting things."

Several BD2K centers will research the efficacy of using mobile sensors such as mobile phones, wrist sensors and Google glass to monitor health status on a continuous basis and to create more effective interventions. The 24/7 nature of these devices have the potential to radically change the way medicine is done. RisQ mobile smoking detection app and wristband image reprinted from Abhinav Parate, Meng-Chieh Chiu, Chaniel Chadowitz, Deepak Ganesan, Evangelos Kalogerakis, RisQ: Recognizing Smoking Gestures with Inertial Sensors on a Wristband, **Proceedings of the 12th International Conference on Mobile Systems, Applications and Services** ***(MobiSys 2014). Google glass image by Mikepanhu, creative commons license.***

Exploiting Mobile and Wearable Sensor Data

A variety of sensors on wristbands and inside smart phones allow the collection of health data around the clock, potentially enabling a holistic view of patient health in ways never before imagined. Several BD2K centers are exploring this potential. At the Mobilize Center, for example, researchers envision using wearable sensor data to encourage healthy physical activity in people at risk for obesity or to warn runners of impending injury.

At the same time, the team at the MD2K Center of Excellence will investigate whether data from multiple types of mobile sensors can help clinicians monitor—and eventually treat—people with various kinds of chronic illness. The MD2K team will gather immense amounts of data from a population of smokers for two weeks using wristbands and chest bands to track activity levels; mobile phones to track not only movement but also location; and Google glass to track what is in an individual's field of vision. In addition, they will strap these devices as well as radiofrequency sensors to a separate population of heart disease patients. It's a huge amount of data that MD2K will be collecting 24 hours a day at a rate of tens of kilobytes per second, Kumar says.

At first, the researchers will just be trying to convert sensor data into markers of health state, behavior, or environmental exposure. For example, can wrist sensors reveal arm movements indicative of smoking as compared to eating? Can chest sensors signal stress levels in ways that relate to an urge to begin smoking again? Can GPS or Google glass data signal proximity to social cues (such as being in a bar, or in close proximity to cigarettes in a store) that might prompt smoking? And for heart patients, can radiofrequency sensors yield valuable information about fluid accumulation in the lungs that might suggest an adverse health event and be used to reduce hospital readmissions?

Once MD2K finds markers of health state, the team will apply machine-learning approaches to discover associations among the various markers. "The challenge is that any one sensor by itself has some information but not necessarily enough," Kumar says. And data quality issues abound. “We need to be able to figure out when changes in the data are due to something we want to infer or something else entirely.” MD2K’s goal is to make extracting health markers from sensor data feasible as well as reliable enough to trigger an appropriate intervention.

DISEASES, DRUGS, AND THERAPIES: *Integrating Data and Knowledge to Improve Patient Care*

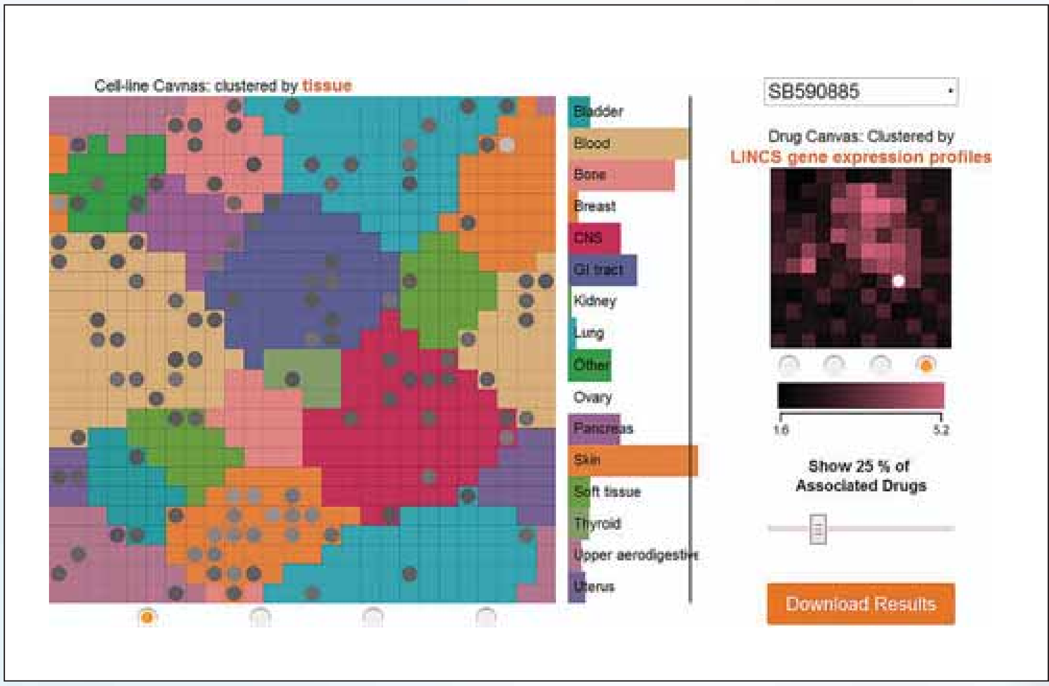
Large public datasets such as GenBank, the Protein Data Bank (PDB) and many others are already widely used by biomedical researchers. But numerous valuable data resources remain dispersed and isolated at institutions around the world. The BD2K Centers will develop various approaches for connecting multiple data types and knowledge resources with one another. Thus, just as PIC is planning to bring together disparate data to gain a much more holistic understanding of patient health, many of the other BD2K Centers plan to integrate multiple types of data and knowledge to achieve a more comprehensive understanding of the human organism. Though the approaches the Centers take to this task may differ, they all have the potential to generate insights that could lead to new drugs, targeted drug regimens, or personalized therapies or surgical interventions.

Data Integration and Cellular Signaling

For bureaucratic and scientific reasons, the pharmaceutical development process is notoriously slow. Many researchers believe drug discovery would be more efficient if we had a better understanding of the relationships between diseases, the drugs that treat them, and the pathways the drugs target in different cells and tissues. Gaining that understanding requires gathering and integrating lots of different kinds of data and knowledge.

That thinking lies, at least in part, behind BD2K support for the **Data Coordination and Integration Center for LINCS** (BD2K-LINCS DCIC). LINCS is the Library of Integrated Network-based Cellular Signatures (LINCS), a group of NIH-sponsored centers, each of which is tasked with characterizing how various cells, tissues and networks respond to disruption by drugs or genetic perturbations. The centers are producing a variety of different data types, including gene expression, epigenetic changes, proteomics, and images.

BD2K-LINCS DCIC, under the leadership of **Avi Ma’ayan, PhD**, associate professor of pharmacology and systems therapeutics at the Icahn School of Medicine at Mt. Sinai, is charged with integrating these diverse data. His team will pull all of the LINCS datasets together, standardize them with ontologies, and also integrate them with data from elsewhere. “We’re organizing the data into networks based mostly on correlations between diseases, side effects, genes and drugs and bringing it all together,” Ma’ayan says.



Ma’ayan and his colleagues have already developed a drug/cell-line browser for LINCS. Users can select a dataset and then visualize the effect of more than 100 drugs on various different cell lines by tissue, mutation, gene expression profile, and drug sensitivity. Shown here: a visualization of cancer cell lines and their sensitivity to the drug SB590885, a Raf inhibitor, with the top 25 percent of the most sensitive cell lines highlighted with circles. The vertical bar graph shows that skin is most sensitive to the drug, which is consistent with the known role of B-Raf in many melanomas. Reprinted with permission from Q Duan et al., Drug/Cell-line Browser: interactive canvas visualization of cancer drug/cell-line viability assay datasets, **Bioinformatics** ***30 (22): 3289–3290 (2014). The browser is freely available at http://www.maayanlab.net/LINCS/DCB/***

For example, the team might have a matrix of merged LINCS experiments where each column is a different experiment—a drug treatment for a single cell or tissue type, say—and the rows are the gene expression responses. This matrix might then be converted into a gene-gene similarity network based on similarities among the drugs’ effects or among groups of genes with a correlated response. Then researchers can look at the overlap between these networks and other networks—for example a network of known drug side effects. “If there is some relationship [between the known side effects and gene expression], it can be very powerful,” Ma’ayan says. “Now you can take new drugs and predict their side effects ahead of time.” Performing this same operation over numerous different cell types and drugs becomes a vast integration task with enormous potential to learn new things.

Ma’ayan hopes the lessons of LINCS data integration will be broadly applicable to the larger BD2K effort. “We want to bring the BD2K effort into LINCS and make LINCS part of the BD2K effort.”

Combining Knowledge with Data for Breast Cancer Pharmacogenomics

To understand why standard chemotherapy works for some breast cancer patients and not others, researchers at Mayo Clinic in Rochester, Minnesota, are sequencing and comparing the genomes of patients’ tumorous and normal tissue. But they'd like to evaluate their patients' tumors in light of what's already known about the various treatment options and the mutations they find. For example, how have those mutations been annotated in another context? This can be difficult because bench biologists typically use algorithms suited to analyzing spreadsheets, whereas huge public datasets have been analyzed using graph or network approaches, says **Saurabh Sinha, PhD**, associate professor of computer science at the University of Illinois, Urbana-Champaign (UIUC). "There's been very little work doing both at the same time."

Bridging that gap is a major goal of the new BD2K center called **KnowEnG, a Scalable Knowledge Engine for Large-Scale Genomic Data** under PI **Jiawei Han, PhD**, professor of computer science, also at UIUC. KnowEnG, when implemented as a cloud-based resource, will enable users to perform spreadsheet analysis in the context of the existing network of genomics data—without downloading the network. Individuals will be able to explore how their data fits in with large networks of public domain datasets, such as the STRING database, which describes protein-protein interactions, and Genemania, which is like a Google search engine for genes, where researchers input a set of genes and retrieve genes that are related in some way, drawing from all available genomics information.

Various KnowEnG team members already have working systems for analyzing large graphs, analyzing spreadsheets in scalable ways, and putting genomics datasets in scalable structures. "The initial one to two years of work will be about connecting these pieces together," says Sinha, who is in charge of the KnowEnG data science core.

Mayo Clinic will use KnowEnG to analyze breast cancer pharmacogenomics data in hopes that community knowledge will shed some light and generate testable hypotheses to improve chemotherapy outcomes. KnowEnG will also be applied to projects as diverse as exploring the relationship between gene expression and human behavior and predicting which bacterial strains are likely to produce novel antibiotic agents. "We're looking forward to having the KnowEnG framework tested on the frontlines," Sinha says.

Combining Data and Mechanics to Enhance Mobility

To evaluate surgical or therapeutic treatments for mobility problems resulting from conditions such as running injuries, osteoarthritis or cerebral palsy (a neurological disorder that affects movement and muscle coordination), the medical profession relies largely on trial and error. The Mobilize Center seeks to establish a different paradigm: the use of big data to optimize treatment.

To learn more from statistical techniques, Delp and his team at the Mobilize Center will bring statistical learning together with mechanistic understanding—knowledge of how something works based on the fundamentals of physics and biology. "By combining these approaches, you simplify your big data problem and gain insights that are meaningful to the biomedical researchers or clinicians," Delp says.

To succeed, they need to find ways to ensure that imperfect data can yield useful information. Motion-capture data to study human mobility is often collected and recorded using different protocols at multiple labs and hospitals around the world. These data can be incomplete, noisy, imprecise, and heterogeneous, and integrating them with notoriously unreliable health records for the same individuals makes things even messier. Throw in the variable of change over time—from periodic clinic visits, for example—and it's a wonder researchers don't just throw up their hands. Fortunately, scientists—including those associated with the Mobilize Center—are becoming quite adept at handling data problems without tossing the baby out with the bathwater. "Part of the big challenge is that if you're trying to gain insight you can't expect to rely on perfect data," Delp says.

The Center will address data imperfection by building on a system called DeepDive developed by **Chris Ré, PhD**, assistant professor of computer science at Stanford and a data science core lead for the Center. Using statistical inference techniques, DeepDive can not only integrate diverse data types but also take imprecision into account and deliver probabilities that an assertion is true. Meanwhile, **Trevor Hastie, PhD**, professor of statistics at Stanford will lead a second data science effort to extract insight from time-varying mobility data spanning from seconds in duration to years.

But even after managing the data imperfection problems, what's left is still just data without any of the advantages of accumulated expert knowledge. The typical big data project looks at all the data and makes inferences based on statistics, says Delp. Sometimes this yields wonderful, insightful surprises, but it can also yield meaningless correlations among bizarre variables no one pays attention to, he says. To learn more from statistical techniques, he and his team at the Mobilize Center will bring statistical learning together with mechanistic understanding—knowledge of how something works based on the fundamentals of physics and biology. "By combining these approaches, you simplify your big data problem and gain insights that are meaningful to the biomedical researchers or clinicians," Delp says.

For example, statistical learning across a large dataset of children with cerebral palsy might identify 23 variables that predict the outcome of a particular surgery intended to improve the patient's ability to walk. But a person with a mechanistic understanding of cerebral palsy gait might be able to select the three variables that can be easily measured and will give surgeons most of what they need to know. "That is so much more powerful to a clinician, when you are getting to the essence of how things work," Delp says. "Finding ways to combine mechanistic understanding with statistical methods is one of the tools the Mobilize Center will develop."

Extracting and Predicting Phenotypes

Researchers would like to be able to look at big data resources—such as electronic health records or collections of brain images—to easily determine a patient's disease status as well as predict how illnesses such as breast cancer or Alzheimer's disease will progress. But "a lot of phenotypes are tricky to interpret or predict," says **Mark Craven, PhD**, professor of biostatistics and medical informatics at the University of Wisconsin, Madison. For example, it can take several months to design algorithms to extract a single phenotype—type 2 diabetes, say—from electronic medical records, just to identify a cohort of cases and controls to study. And even with a three-million-voxel brain image, it's hard to predict whether a patient will progress to Alzheimer's disease.

Accurately modeling causation in a biological system is challenging. The sheer number of variables can raise millions of chicken and egg questions about what aspect of a system caused another, not to mention whether a hidden variable (the rooster next door?) has a causal influence.

The new BD2K-funded **Center for Predictive Computational Phenotyping** (CPCP) under Craven's leadership is hoping to improve the methods for extracting and predicting phenotypes from electronic health records (EHRs), images or other large datasets such as transcriptomic or epigenomic data—as well as combinations of these different data types. "One of the interesting challenges is how you can leverage all of the different data sources," Craven says.

Like the Mobilize Center, CPCP will wrangle with datasets that are sparse, incomplete or untrustworthy, as EHRs typically are. Trying to identify something that should have been explicitly recorded in these records (such as a diagnosis) is surprisingly challenging. But it's even harder to extract information that is not explicitly measured, such as disease duration, risk factors for complications, or the effectiveness of a particular treatment. So CPCP will work on developing improved and streamlined approaches for extracting information from electronic health records, with an initial focus on such illnesses as heart attacks, asthma, and VTE (venous thromboembolism, a type of blood clot). For example, since reduced blood volume is a risk factor for VTE but is not directly recorded in the EHR, they will try to identify a constellation of other information that could be used to infer reduced blood volume.

Through its "Value of Information" lab, CPCP is also interested in using the data they already have to predict, in an optimal way, what information would add the greatest value and produce the greatest increase in predictive power, essential functionality for this information-rich age. Such a capability could help decide the minimum set of tests needed to arrive at a diagnosis for a patient, or which additional experiments a researcher should do to best understand a system. In the context of computational phenotyping, it would also help researchers extract the most predictive information from an EHR or image.

Finding Causation

Accurately modeling causation in a biological system is challenging. The sheer number of variables can raise millions of chicken and egg questions about what aspect of a system caused another, not to mention whether a hidden variable (the rooster next door?) has a causal influence.

The new BD2K-funded **Center for Causal Modeling and Discovery** (CCMD) **of Biomedical Knowledge from Big Data** is dedicated to deriving causal insight from the huge numbers of variables often present in biomedical data. For example, they will look for patterns that suggest causation among millions of variables involved in cancer signaling pathways, including genomics and gene expression data as well as data on cell function or dysfunction. Similarly, for chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis, they will hunt down causal factors in high-throughput data that includes gene expression, DNA methylation, and microRNA data as well as clinical records. "The algorithms can be applied to a wide variety of biomedical data," says **Greg Cooper, MD, PhD**, professor of biomedical informatics at the University of Pittsburgh and contact PI of CCMD. Having highly efficient algorithms to search for likely causal patterns in the vast biomedical data we have today could be a key step toward prevention and/or treatment for a huge range of diseases.

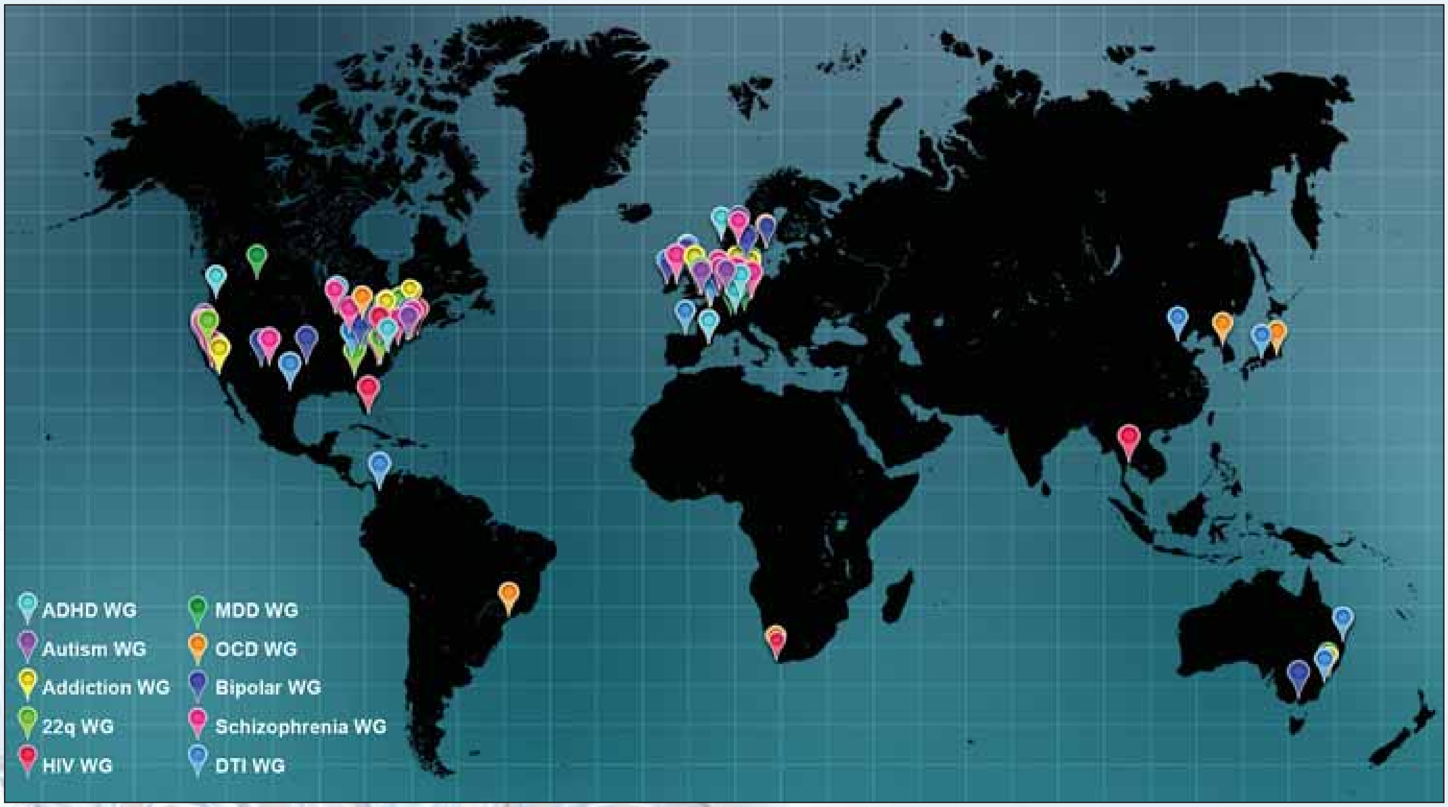
CCMD plans to provide "one-stop shopping" for these types of high quality causal discovery algorithms. Some algorithms search over different possible networks to determine which are most strongly supported by the data while accounting for prior knowledge and belief based on the scientific literature (using Bayesian methods). Others look for patterns of independencies and dependencies among the variables that suggest particular causal relationships.

By applying these methods to diverse biomedical problems and making the algorithms available through application programming interfaces (APIs), the Center will ensure that they are both broadly applicable and available to the research community. "These APIs are one key deliverable of our center," Cooper says.

POWER TO THE PEOPLE: *Distributing Algorithm Development*

In 2009, the ENIGMA Consortium (Enhanced Neuro-Imaging Genetics through Meta-Analysis) was launched to bring together researchers—and their genomic and imaging data—to get a better understanding of brain function and disease. The Consortium, which has now grown to over 300 researchers, has access to genome-wide, neuroimaging, and clinical data from more than 31,000 subjects worldwide. Using this impressive amount of data, they are studying ten major brain diseases, including schizophrenia, major depression, bipolar illness, attention deficit hyperactivity disorder, and autism. They can look for genes associated with these diseases, examine differences in how the brain reacts to different drugs, and trace how different parts of the brain are connected to one another in people with and without brain disease.

The beauty of ENIGMA is that the results are achieved without shipping data



The ENIGMA alliance studies brain scans and DNA at more than 185 sites around the world. They created working groups to pool and compare data from many neuroimaging centers in order to understand the effects on the brain of various conditions, including bipolar disorder, major depressive disorder (MDD), addiction and schizophrenia. The result is a data pool with tens of thousands of subjects. The institutions involved in the working groups are shown on this map from June 2013. Thompson PM et al., The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data, Brain Imaging Behav. *2014. Epub 2014/01/09. doi: 10.1007/s11682-013-9269-5. PubMed PMID: 24399358.*

around the world. Instead, small groups within the Consortium who wish to take a crack at a particular research question form an alliance to help each other out. They develop algorithms and distribute them to others in the Consortium, essentially enabling a meta-analysis across multiple centers. "They can get off the ground quickly with tens of thousands of data points by sending software out," says P**aul Thompson**, **PhD**, professor of neurology at the University of Southern California.

Thompson, who is the director of the ENIGMA Consortium, is now also the PI for the new BD2K-funded **Enigma Center for Worldwide Medicine, Imaging and Genomics**. The Center's work lies squarely in the area of developing analytical and statistical tools for big data, but it is funded as 20 sub-awards to researchers around the globe. "Data is nothing without people," Thompson says.

ENIGMA Center researchers will develop refined algorithms that can analyze brain maps, measures and signals, and relate them to genomic, environmental and epidemiological data as well as clinical outcomes.

"There are some alliances you can form that make it easier for everybody to do science," Thompson says. "We hope to see discoveries on a scale that hasn't been possible," he says.

DATA STANDARDS AND METADATA:
Sorting the Nuts and Bolts

Basic tasks, such as storing and accessing data efficiently, may require little or no attention when researchers work with small datasets. But these foundational issues must be addressed head on when datasets become enormous. And several BD2K Centers are doing just that: creating data structures that allow for efficient storage of and access to big data; and developing ways to easily tag the data with annotations or "metadata" so they carry signatures of where they came from as well as how they've been used through time.

These are not simple problems. They are also not glamorous. "Most people are thinking about the great discoveries they are going to be making with the data and—make no mistake—we are too," says **David Haussler**, **PhD**, professor of biomolecular engineering at the University of California, Santa Cruz, and PI of the new BD2K **Center for Big Data in Translational Genomics** (CBDTG). "But we're also emphasizing the need to get the nuts and bolts right before making discoveries."

Establishing Data Standards

In cooperation with the Global Alliance for Genomics and Health (GA4GH), a nonprofit consortium of genomics researchers worldwide, CBDTG will develop and implement global standards for genomics data.

The Center's effort builds on work begun by the Thousand Genomes Project, which has already pioneered several novel file formats—BAM for storage of large files of DNA reads and VCF for storage of files called variants. GA4GH will make these formats ready for prime time and clinical use, as well as create an additional compressed format, called CRAM, that Haussler says will save millions of dollars in space costs for storing large genomics files. CBDTG will work with them to build abstract data schemas so that the data can be stored efficiently and optimally accessed. “It would be hopelessly inefficient to paw through the coming massive amounts of genomics data to get the information you want if it is stored in the current file formats,” Haussler says.

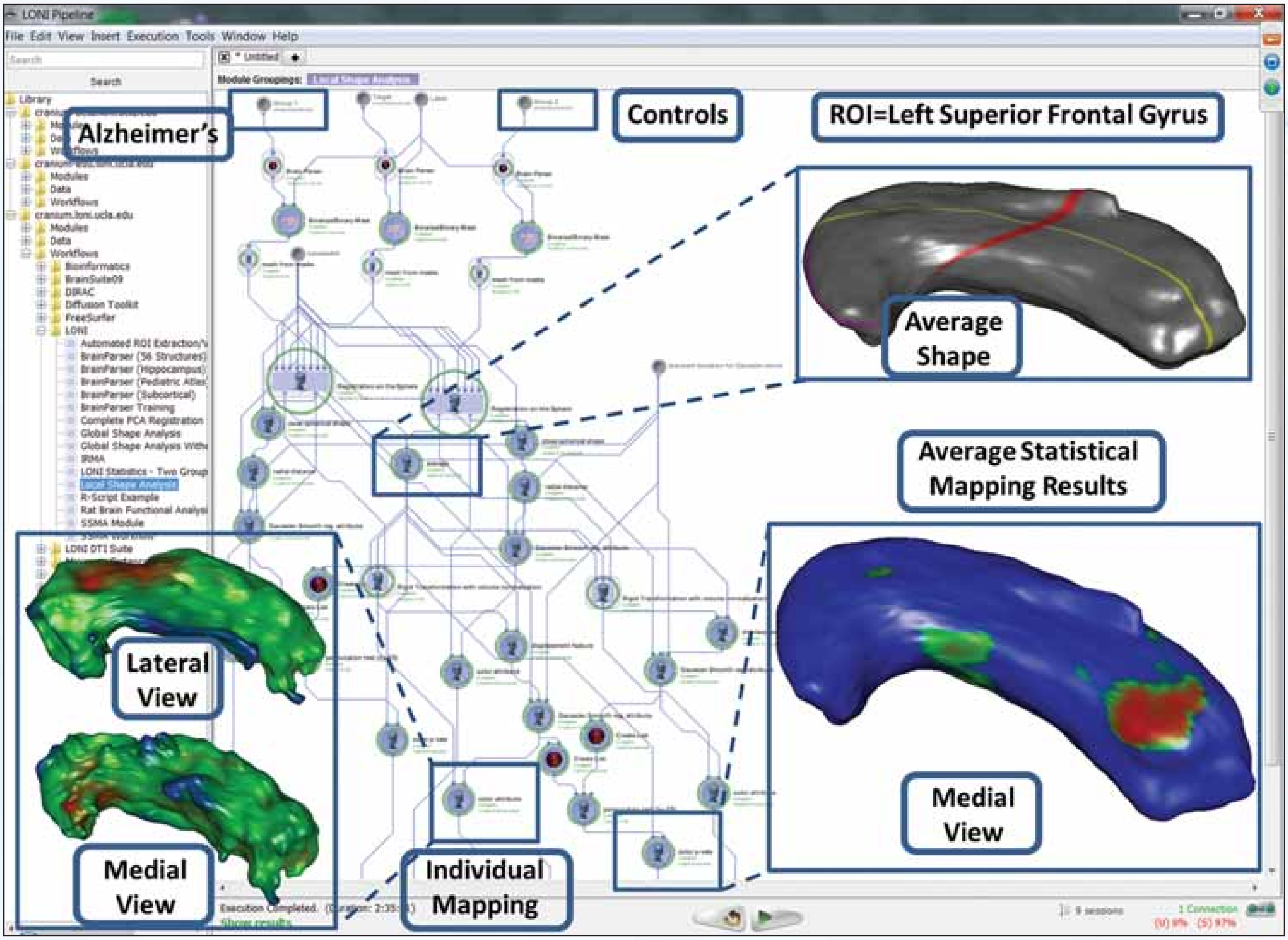
At the same time, Haussler’s team is working with GA4GH on standards for representing genetic variation—not only single nucleotide changes but also rearrangements and duplications. “If we already knew all possible human variations, it would be a lookup problem. We’d have a name for each variation,” Haussler says. “But that’s not the case. Every individual’s genome will reveal new variations.”

CBDTG will also help the other BD2K centers deal with the federation of data across multiple locations. “We need something like URLs that identify the objects you’re looking for,” he says. “You shouldn’t really care where it is.” Haussler’s center is working with Google, Amazon and Microsoft to address these problems at a fundamental level. One solution involves attaching a cryptographic signature to pieces of data that verify what they are, sort of like bar codes on a package. But there are lots of potential stumbling blocks. For example, there can be multiple copies of the same data, and one small thing can change in one copy. “It can become a nightmare in electronic librarianship,” Haussler says. “We need rules of the road for how data are represented, verified, changed, stored back in, never lost, never compromised. It’s great to be working with the biggest and best on that.”

Making Annotation Happen

Ever since the 1980s when researchers were first required to post certain experimental datasets online, they have been required to also create metadata—basic information about how the data were produced. But scientists don’t always do a thorough enough job of it.

“There’s a real ‘what’s in it for me’ problem to overcome,” says **Mark Musen, MD, PhD**, professor of medicine at Stanford University and PI of the new BD2K **Center for Expanded Data Annotation and Retrieval** (CEDAR). And that threatens re-


LONI Pipeline
File Edit View Insert Execution Tools Window Help
Search
Module Groupings: Local Shape Analysis
Alzheimer's
Controls
ROI=Left Superior Frontal Gyrus
Average Shape
Average Statistical Mapping Results
Lateral View
Medial View
Individual Mapping
Medial View
Execution Completed.

One goal of the BD2K program is to provide computational tools in a format that can be easily used by biomedical researchers. This image shows an example of the pipeline workflow for a specific brain-mapping problem—local shape analysis—developed by Toga and his colleagues. The example here starts with the raw magnetic resonance imaging data for 2 cohorts (11 Alzheimer's disease patients and 10 age-matched normal controls), extracts a region of interest (left superior frontal gyrus [LSFG]) for each subject, and generates a 2-D shape manifold model of the regional boundary. Then the pipeline computes a mean LSFG shape using the normal subjects’ LSFG shapes, co-registers the LSFG shapes of all subjects to the mean (atlas) LSFG shape, and maps the locations of the statistically significant differences of the 3-D displacement vector fields between the 2 cohorts. The insert images illustrate the mean LSFG shape (top-right), the LSFG for one subject (bottom-left), and the between-group statistical mapping results overlaid on the mean LSFG shape (bottom-right), red color indicates p-value < 0.01. Reprinted with permission from Dinov, ID et al., Applications of the Pipeline Environment for Visual Informatics and Genomics Computations, BMC Bioinformatics, *12:304 (2011).*

searchers' ability to rely on, replicate, or share one another's data. "Until we make it simple to annotate data in a clear way, we're going to have serious problems for science," Musen says.

The Center's team already has a few tricks up its sleeve for simplifying annotation, including technologies developed through earlier projects such as Protégé, an ontology development system, and BioPortal, an ontology repository developed by the National Center for Biomedical Ontology. CEDAR will leverage these two technologies to automatically create Web-based interfaces for filling out metadata templates. CEDAR will also create text analysis and predictive data tools to fill in portions of the templates automatically.

In addition, CEDAR will develop ways of allowing metadata to evolve as the data are re-analyzed or compared to other data. "Metadata are not a static description of an experiment," Musen says. "They are an evolving record of the conversations researchers have about experiments over time."

The CCMD team also plans to develop tools for annotation—but for application to causal models that are derived from analyzing biomedical data, rather than to the raw data. "Just as meta-information about data (metadata) can have significant value, so too can meta-information about the models derived from that data," says Cooper.

Whether simplifying metadata collection will incentivize researchers to do a better job remains to be seen. "I hope the overall ecosystem CEDAR creates will show researchers that the authoring of metadata need not be onerous, and that science has much to gain from first-rate annotations," Musen says.

Should CEDAR's researcher-driven strategy prove insufficient, **Andrew Su, PhD**, of the Scripps Research Institute has a different idea. As part of **The Heart of Data Science**, a new BD2K center based at the University of California, Los Angeles, under PI **Peipei Ping, PhD**, Su will enlist the help of citizen scientists who will extract information from the biomedical literature and annotate proteomic data for cardiovascular research. To evaluate the quality of these annotations, the crowdsourced work will be compared to the Reactome and Intact databases, which are curated by experts, says **Henning Hermjakob, PhD**, who will head up the center's data science core.

Ping's center will also showcase the value of metadata for discovery by expanding the Proteome Xchange, a consortium that aggregates proteomic metadata into a searchable centralized form. The Center will add more proteomics repositories to the Exchange as well as extend it to include other –omics data types, such as metabolomics. This program will not only motivate better metadata collection but also address "the absolutely nontrivial problem of finding which datasets are relevant to a particular research project," Hermjakob says.

OFFER IT TO THE WORLD: *User-Friendly Interfaces*

Thompson likens big data research to a lengthy relay race. The baton passes from labs into structured datasets; becomes integrated with other data; is subjected to analysis using novel tools and creative mathematics; and finally is presented in a user-friendly interface for others to use. That's the victory lap for the BD2K Centers: providing a means for other researchers or clinical personnel to use big data effectively.

"Great integrative work and scalable computing will amount to nil if the interface isn't immediately appealing to the biologist," says Sinha. The KnowEnG team's interface will allow biologists to identify genes that discriminate between samples as well as probe the literature for relevant information about those genes.

Similarly, CCMD will create a workstation for biomedical scientists so they can easily select datasets, apply causal discovery algorithms, see results graphically, and annotate and store them. And the PIs of other centers have similar plans.

But the new BD2K-funded **Big Data for Discovery Science** (BDDS) **Center** under PI **Art Toga, PhD**, provost professor and director of the Laboratory of Neuro Imaging at the University of Southern California, has interfaces as a focus. They're creating a smart pipeline that offers big data analysis tools for use by non-cognoscenti. It offers drag and drop glyphs in a graphical environment that is layered with expertise to guide users toward the appropriate tool for a given task. "The system is self-aware," says Toga. It will not only show publications about a particular tool, but also offer information about parameter settings based on past experience—acting as a sort of advisor on best practices. Toga says the pipeline has to make tools understandable within five to ten minutes, and also give users feedback when they make a mistake. "If we have to develop complicated user manuals, we've failed," he says.

The pipeline will also include novel ways to interrogate data casually, looking for relationships and providing instantaneous insights to drive hypothesis generation. For example, researchers could peruse large, integrated datasets to look for relationships between two cohorts that differ in only one particular feature. "We built a prototype of such a thing—with the computation attached to the database—and played with it and it was unbelievably useful," Toga says.

Though Toga will test the pipeline using neuroimaging, genomics and proteomics tools, nothing precludes its use in other scientific domains. "The workflow is agnostic as to the tool type," Toga says.

A VIRTUOUS CYCLE

Data science methods can't be developed in a vacuum. "You have to think, what does the method do; what can you learn; what problem are you solving," Delp says. It's an approach designed to ensure the development of big science methods that work—and the first step in assuring that they'll also be useful to others.

As Bourne puts it, "We view this as a virtuous cycle." The researchers are motivated by the biomedical research that gets done, and in the process of doing that work they generate data, use data, and develop tools that all get "virtuously" shared with others to provide further motivation. "Sharing the data and the software across the centers and to other investigators and beyond is key," he says.

So too is cooperation: Where the Centers can work synergistically with one another and the NIH, they will do so. For example, both the Mobilize Center and MD2K will be exploring the effectiveness of using wearable sensor data to change unhealthy behaviors. Ontologies and annotations, the focus of CEDAR, play a role in a number of the Centers, including CCMD and the Heart of Data Science. And nearly all the Centers have to find ways to deal with the sparsity, noisiness, and heterogeneity that so often characterizes big data. By tackling these challenges together, Collins says: "The whole is going to be a lot greater than the sum of its parts." □