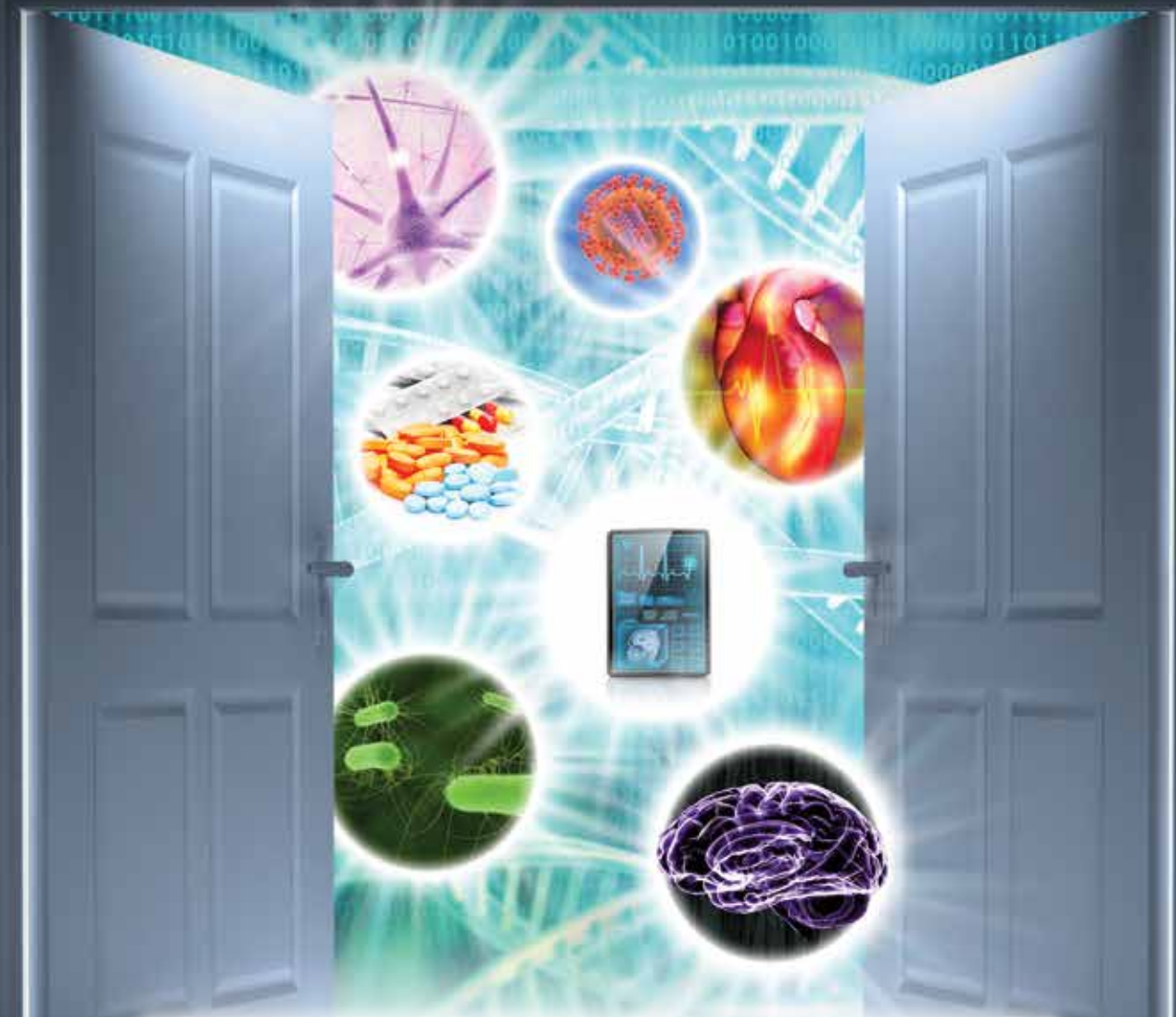


DIVERSE DISCIPLINES, ONE COMMUNITY

# Biomedical Computation

Published by the Mobilize Center, an NIH Big Data to Knowledge Center of Excellence

REVIEW

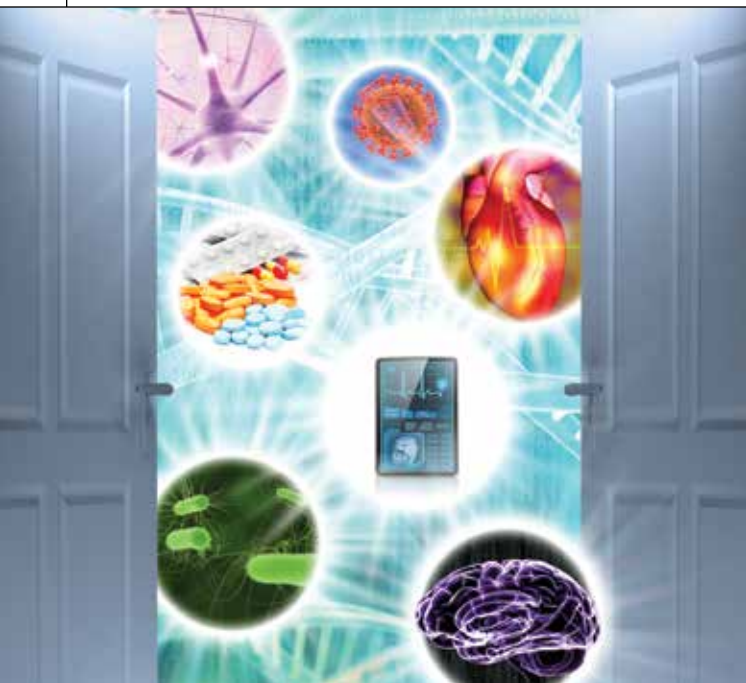


**SPECIAL ISSUE:**

**BD2K CENTERS**

**Open Doors to Discovery**

SUMMER 2017



## 2 BD2K Centers *The Basics*

BY KATHARINE MILLER

## 4 BD2K Centers' Pan-NIH Impact

BY KATHARINE MILLER

## 23 Text Mining:

*How the BD2K Centers are Making Knowledge Accessible*

BY KRISTIN SAINANI, PhD

## 33 The FAIR Data-Sharing Movement:

*BD2K Centers Make Headway*

BY KATHARINE MILLER

**Cover Art:** Created by Rachel Jones of Wink Design Studio using: Neurons © Kts, HIV virus © Sebastian Kaulitzki, Pills © Oleksandr Rozdolyanskiy, Heart © Krishnacreations, Device © Štěpán Kápl,

*E.coli* bacteria © Sebastian Kaulitzki, Human brain © Yakobchuk, DNA © Jezper, all of Dreamstime.com. Doors image is © wavebreak of CreativeMarket.com.

**Pages 2-3:** US map outline © chrupka of 123RF Stock Photo.

**Page 23 Art:** Document sphere © Wuka of Dreamstime.com.

### DEPARTMENTS

#### 1 GUEST EDITORAL

BIG (DATA) SCIENCE MATTERS  
BY SCOTT DELP, PhD

#### 41 BD2K TOOL HIGHLIGHT

EXPLORING PATTERNS IN BIG DATA USING ClusterEnG, A CLUSTERING ENGINE FOR GENOMICS

BY MOHITH MANJUNATH, PhD, AND YI ZHANG

#### 42 SEEING SCIENCE

VISUALIZING HUMAN GENOME VARIATION

BY KATHARINE MILLER

### UNDERCURRENTS

6

**BIG DATA AND DRUGS:**  
BD2K CENTERS SOLIDIFY EMERGING APPROACHES

BY KATHARINE MILLER

10

**BD2K CENTERS SIZE UP BRAIN DISEASE**

BY ESTHER LANDHUIS, PhD

14

**MOBILE HEALTH:**  
BD2K CENTERS HARNESS SENSOR DATA

BY ALEXANDER GELFAND

18

**DISEASE DETECTIVES**

BY JONATHAN WOVEN

#### Summer 2017

Volume 13, Issue 1  
ISSN 1557-3192

#### Co-Executive Editors

Scott Delp, PhD  
Russ Altman, MD, PhD

**Associate Editor** Joy Ku, PhD

**Managing Editor** Katharine Miller

#### Science Writers

Alexander Gelfand, Esther Landhuis, PhD,  
Katharine Miller, Kristin Sainani, PhD,  
Jonathan Wosen

#### Community Contributors

Mohith Manjunath, PhD, Yi Zhang

#### Layout and Design

Wink Design Studio

#### Printing

AMP Printing

#### Editorial Advisory Board

Ivet Bahar, PhD,  
Jeremy Berg, PhD,  
Gregory F. Cooper, MD, PhD,  
Mark W. Craven, PhD,  
Jiawei Han, PhD,  
Isaac S. Kohane, MD, PhD,  
Santosh Kumar, PhD,  
Merry Lindsey, PhD,  
Avi Ma'ayan, PhD,  
Mark A. Musen, MD, PhD,  
Saurabh Sinha, PhD,  
Jun Song, PhD,  
Andrew Su, PhD,  
Paul M. Thompson, PhD,  
Arthur W. Toga, PhD,  
Karol Watson, MD

For general inquiries, subscriptions, or letters to the editor, visit our website at [www.bcr.org](http://www.bcr.org)

#### Office

Biomedical Computation Review  
Stanford University  
318 Campus Drive  
Clark Center Room W352  
Stanford, CA 94305-5444

Publication is supported by NIH Big Data to Knowledge (BD2K) Research Grant U54EB020405.

Information on the BD2K program can be found at <http://datascience.nih.gov/bd2k>.

#### The NIH program and science officers for the Mobilize Center are:

Grace Peng, National Institute of Biomedical Imaging and Bioengineering,  
Theresa Cruz, National Institute of Child Health and Human Development,  
Daofen Chen, National Institute of Neurological Disorders and Stroke

#### Biomedical Computation Review is published by:

The Mobilize Center,  
an NIH Big Data to Knowledge (BD2K) Center of Excellence

[mobilize.stanford.edu](http://mobilize.stanford.edu)



## BIG (DATA) SCIENCE MATTERS



In 1967, when nuclear physicist Alvin Weinberg<sup>1</sup> coined the term Big Science, he was most interested in launching a concentrated effort to develop nuclear technologies. But he anticipated that large-scale approaches to biomedicine would also be productive. In fact, they have been—from the War on Cancer in the 1970s, to the Human Genome Project of the 1980s and 1990s, and The Cancer Genome Atlas and ENCODE projects of the 2000s.

Since Big Science projects require significant funding, there has long been a perceived tension between big research efforts and smaller ones. But I would argue that this is a false dichotomy. Big Science efforts like those listed above have boosted our fundamental understanding of biomedicine (to the benefit of the entire biomedical research community) and produced scientific tools and methods that have had a multiplier effect when distributed and used by the research community in projects big and small.

Biomedical computation has also had some “Big” initiatives, including the National Centers for Biomedical Computing (NCBCs), which flourished from 2004–2014 and the current Big Data to Knowledge (BD2K) Centers of Excellence, which were funded for four years starting in

Special Issue of *Biomedical Computation Review* offers a glimpse at how Big Data Science can transform the biomedical research landscape in ways that benefit the research community and increase our knowledge and understanding of biomedicine.

In *The FAIR Data-Sharing Movement: BD2K Centers Make Headway*, you will read about the ways various BD2K Centers are establishing state-of-the-art methodologies for making data findable, accessible, interoperable and reusable. Fulfilling these goals is essential if biomedical researchers are going to make use of big data sources to advance biomedical knowledge. And the BD2K Centers are at the forefront of making that happen.

In this issue’s other feature story, *Text Mining: How the BD2K Centers are Making Knowledge Accessible*, you will see how top-notch computer scientists are bringing their text-mining tools to bear in biomedicine. From Chris Ré and his team at the Mobilize Center to Jiawei Han and his colleagues at the KnowEnG Center, the level of excellence is nothing short of remarkable.

And then there are the four *UnderCurrents* in this issue, each describing how the BD2K Centers are making a difference in targeted areas of biomedicine. You’ll read about how BD2K Centers are harnessing the vast stream of data coming from wearable sensors to improve health; how large-scale collaborative BD2K projects are deepening our understanding of brain diseases; how the Centers

---

**This Special Issue shows that Big (Data) Science matters in biomedicine. It matters not only to the researchers doing it, but to the entire research community and, more importantly, to the advancement of science and the improvement of health.**

---

the fall of 2015. Compared with the NCBCs, the BD2K Centers are more focused on a single mission: extracting knowledge from big data. At the same time, because there are 13 BD2K Centers rather than 7 NCBCs, their coverage of biomedical data science, and indeed the entire spectrum of biomedicine is more thorough (see pages 4–5 for a graphic showing the Centers’ pan-NIH impact).

Just two and a half years into their four-year grants, the BD2K centers are already proving their value. This


are striving to map the universe of drugs, predict drug responses and adverse reactions, and develop tools for drug repurposing; and how BD2K researchers are using data to detect and predict disease onset and progression.

This Special Issue shows that Big (Data) Science matters in biomedicine. It matters not only to the researchers doing it, but to the entire research community and, more importantly, to the advancement of science and the improvement of health. It’s a perfect fit for the NIH mission: to uncover new knowledge that will lead to better health for everyone. □


<sup>1</sup> Weinberg, A.M. 1967. Reflections on Big Science. *The M.I.T. Press*, Cambridge, MA. 182 pp.

**CEDAR**  
*Center for Expanded Data Annotation and Retrieval*

**MISSION:** To make data submission smarter and faster, so biomedical researchers and analysts can create and use better metadata.




**THE MOBILIZE CENTER**  
*The National Center for Mobility Data Integration to Insight*




**MISSION:** To overcome the challenges of analyzing large data sets that describe human movement and to improve human movement across the wide range of conditions that limit mobility.

**BDTG**  
*Center for Big Data in Translational Genomics*

**MISSION:** A partnership coordinated by UC Santa Cruz to create scalable infrastructure for the broad application of genomics in biomedicine.




**KNOWENG:**  
*A Scalable Knowledge Engine for Large-Scale Genomic Data*



**MISSION:** To transform the way biomedical researchers analyze their genome-wide data by integrating multiple analytical methods derived from the most advanced data mining and machine learning research.

**HEART BD2K**  
*A Community Effort to Translate Protein Data to Knowledge: An Integrated Platform*

**MISSION:** To advance cardiovascular medicine through innovations in data science platforms and tools to enlist community contributions in Big Data computing.




**bioCADDIE\***  
*Biological and HealthCare Data Discovery and Indexing Ecosystem*



**MISSION:** To develop a prototype data discovery index that will enable finding, accessing and citing biomedical big data.

**ENIGMA:**  
*Enhancing Neuroimaging Genetics through Meta Analysis*

**MISSION:** To bring together researchers in imaging genomics to understand brain structure, function, and disease, based on brain imaging and genetic data.



**BDDS**  
*Big Data for Discovery Science Center*



**MISSION:** To take an “-ome to home” approach toward streamlining big data management, aggregation, manipulation, integration, and the modeling of biological systems across spatial and temporal scales.

\* bioCADDIE received a Data Discovery Index Coordination Consortium (DDICC) Award through the BD2K program

# BD2K CENTERS:

## CPCP

### *Center for Predictive Computational Phenotyping*

**MISSION:** To developing methods for modeling and predicting thousands of phenotypes to advance biomedical science and improve human health.



## PIC-SURE

### *Patient-Centered Information Commons: Standardized Unification of Research Elements*

**MISSION:** To create a massively scalable toolkit to enable large, multi-center Patient-centered Information Commons (PIC) at local, regional, and national scale, where the focus is the alignment of all available biomedical data (genetic, environmental, imaging, behavioral, or clinical) per individual.



## CCD

### *Center for Causal Modeling and Discovery of Biomedical Knowledge from Big Data*



**MISSION:** To find meaningful relationships in big data that lead to new insights into health and disease.

## LINCS TRANSCRIPTOMICS

### *Broad Institute LINCS Center for Transcriptomics*

**MISSION:** To develop comprehensive signatures of cellular states that can be used by the entire research community to understand protein function, small-molecule action, physiological states and disease states.



## MD2K

### *Center of Excellence for Mobile Sensor Data-to-Knowledge*



**MISSION:** To develop Big Data solutions to quantify physical, biological, behavioral, social, and environmental factors that contribute to health and wellness in daily life.

## BD2K-LINCS-DCIC

### *LINCS-BD2K Perturbation Data Coordination and Integration Center*



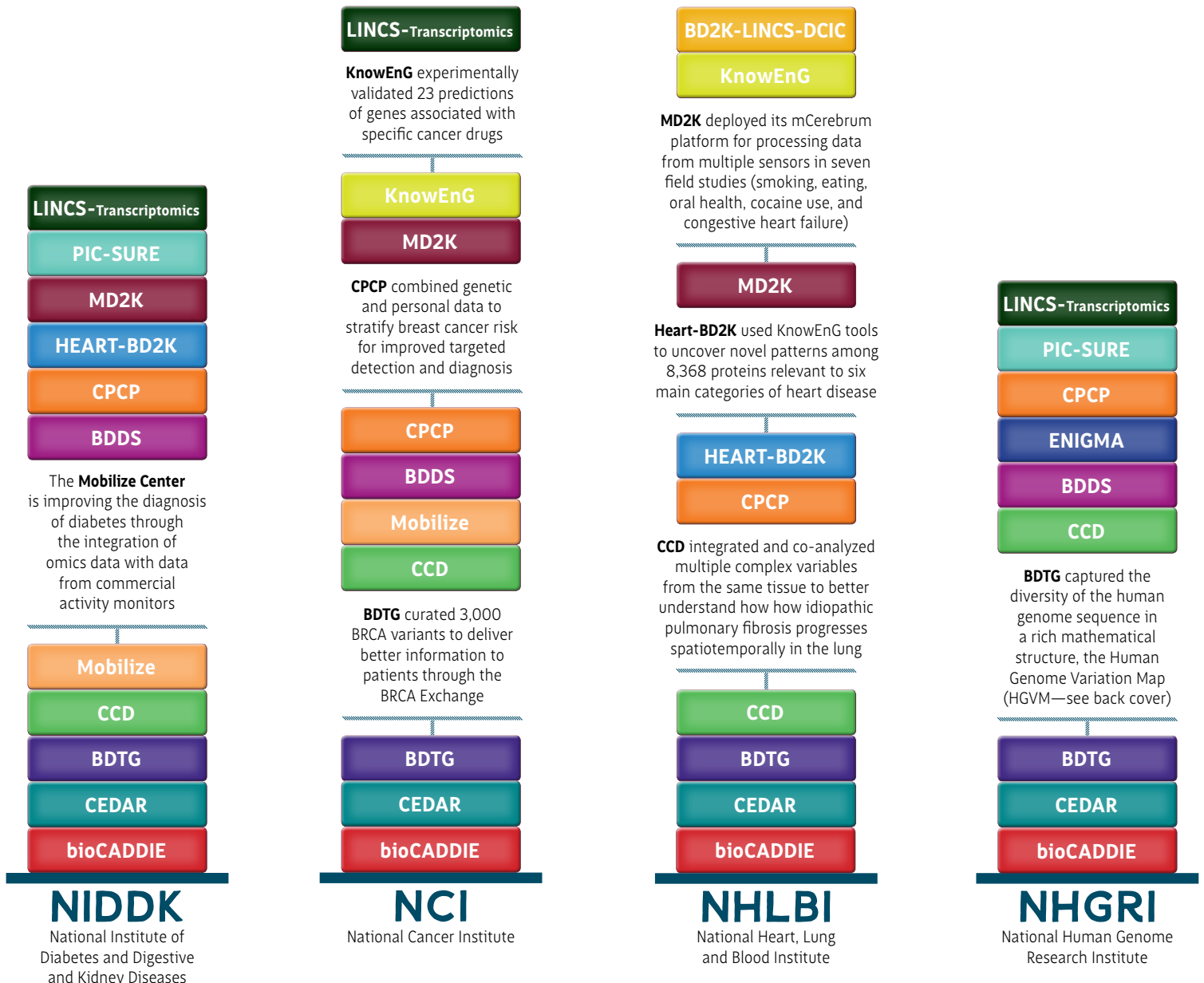
**MISSION:** To construct a high-capacity scalable integrated knowledge environment enabling federated access, intuitive querying and integrative analysis and visualization across all LINCS resources and many additional external data types from other relevant resources.

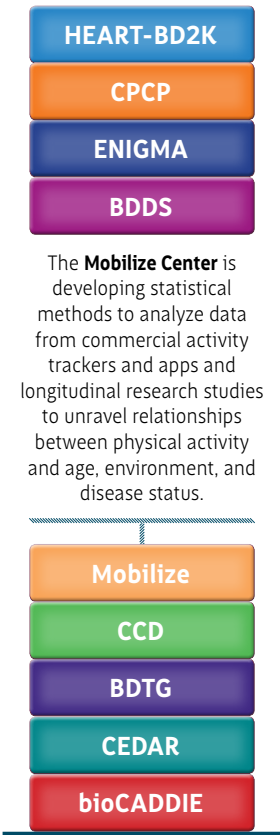
# The Basics

# BD2K CENTERS' PAN-NIH IMPACT

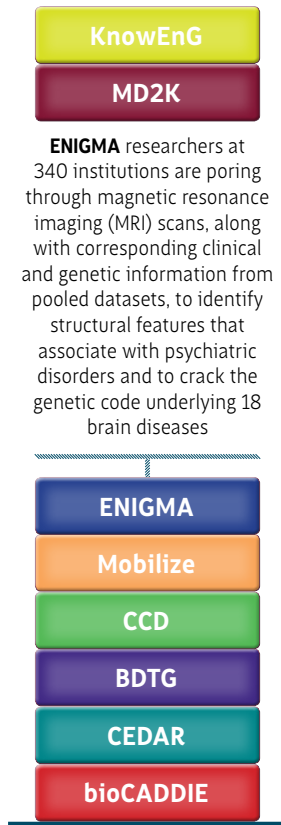
The data science tools and methods developed by the BD2K Centers—from clustering algorithms to data integration approaches, text mining, and image analysis—are valuable across the entire field of biomedicine, benefiting virtually all of the NIH Institutes.

Here we show the impact of the tools and research of the BD2K Centers for 13 of the Institutes. Center names are stacked atop each Institute to which that Center's work relates. A few specific examples are also highlighted in text and more examples are provided in an online interactive graphic at <http://bcr.org/uploads/bd2kimpact.html>.

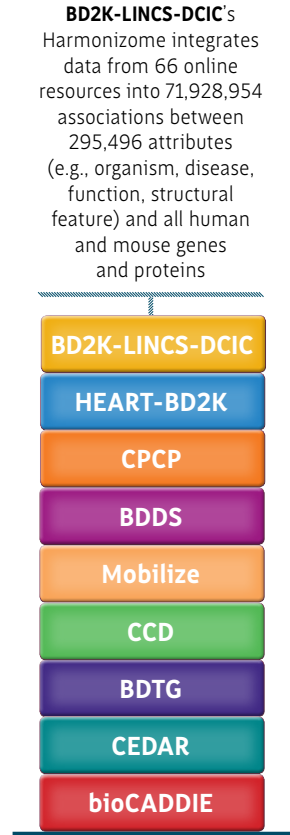




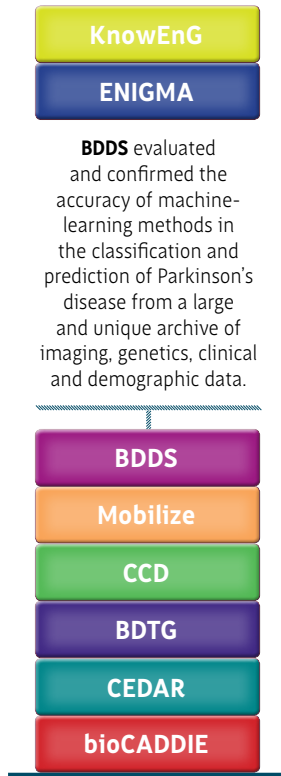
**NIA**  
National Institute on Aging



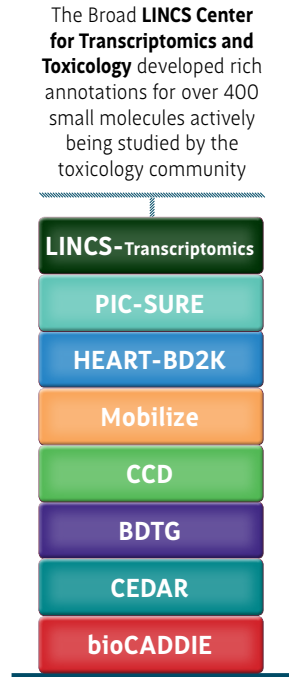
**NIMH**  
National Institute of Mental Health



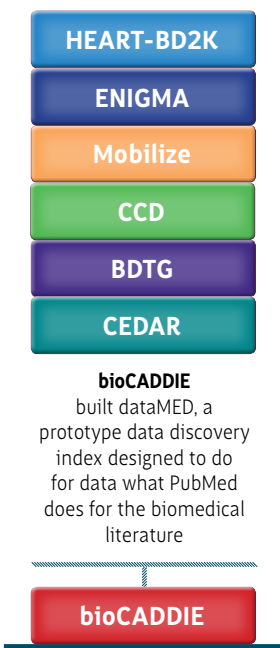
**NIGMS**  
National Institute of General Medical Sciences



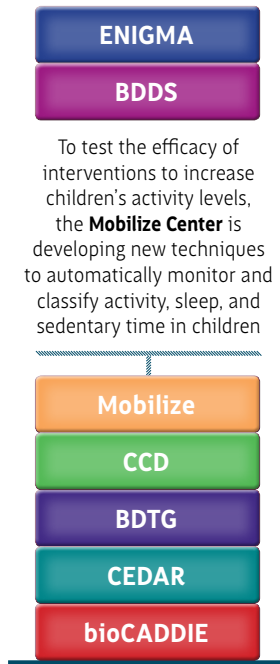
**NINDS**  
National Institute of Neurological Disorders and Stroke



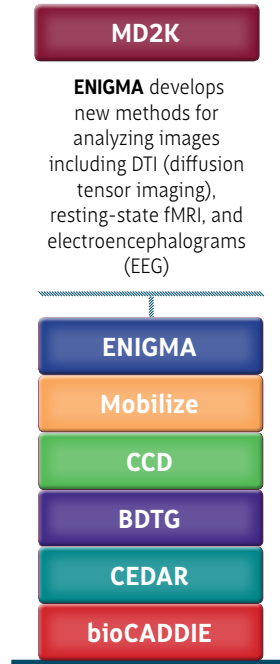
**NIEHS**  
National Institute of Environmental Health Services



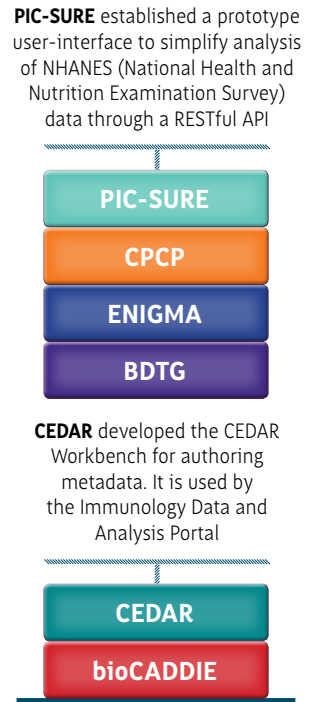
**NLM**  
National Library of Medicine



**NICHD**  
National Institute of Child Health and Human Development



**NIBIB**  
National Institute of Biomedical Imaging and Bioengineering



**NIAID**  
National Institute of Allergy and Infectious Disease

# BIG DATA AND DRUGS: BD2K CENTERS SOLIDIFY EMERGING APPROACHES

The Big Data era in biomedicine offers a grand promise: that by crunching vast quantities of multi-omics data through appropriate statistical analyses, researchers will gain a comprehensive understanding of health and disease that will lead to new, effective, and personalized treatment options.

Current work in systems pharmacology by several BD2K Centers offers a glimpse at that potential. Researchers at

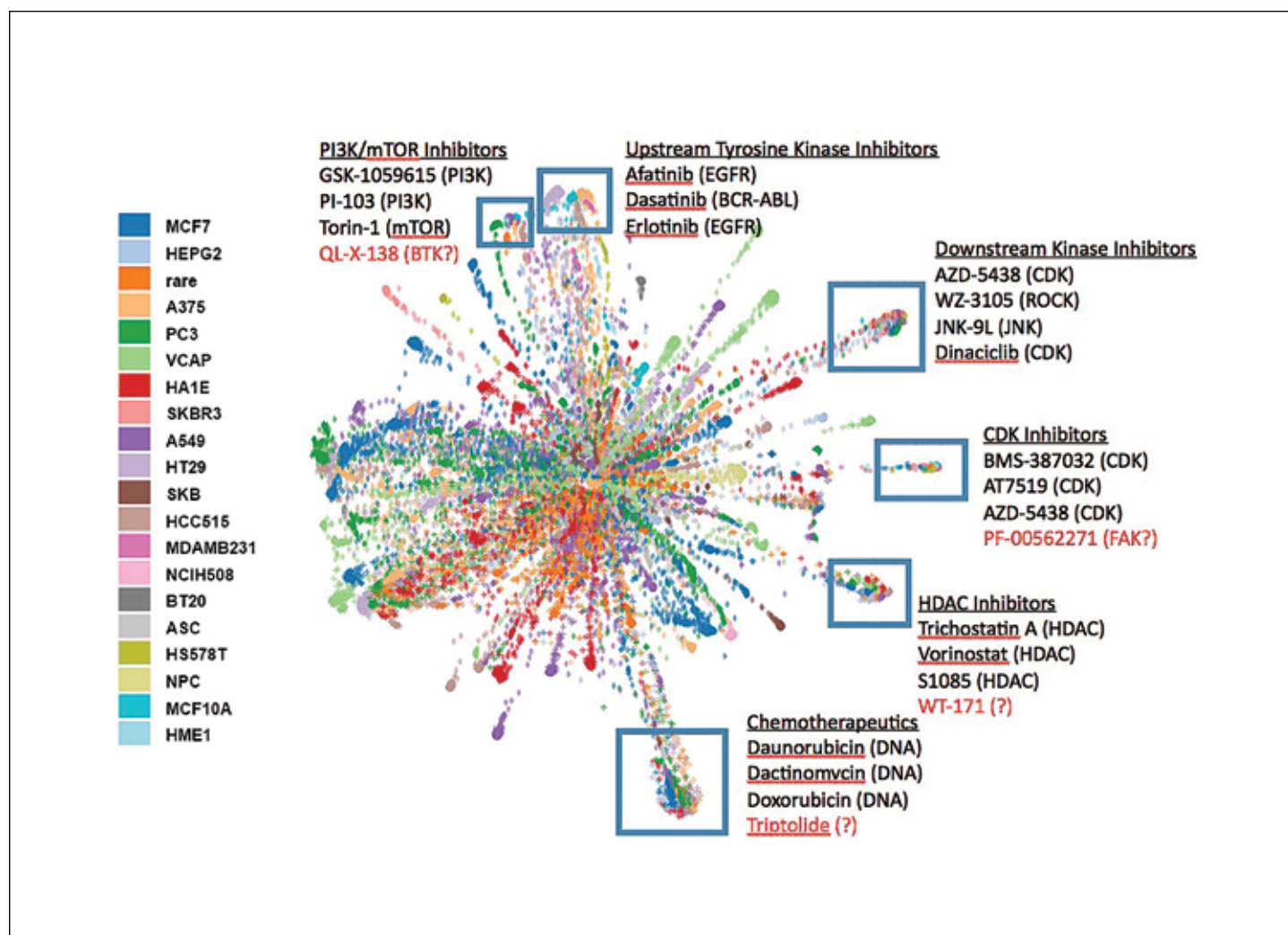
the **BD2K-LINCS-DCIC** are mapping the global space of responses of human cells to many drugs and other small molecules as well as exploring the universe of drug-induced adverse effects. Those at **KnowEnG** are analyzing multi-omics measurements in hopes of understanding drug response in cancer patients. Meanwhile, the **PIC-SURE Center** is standardizing procedures and developing open-source

tools for drug repositioning research.

Taken together, BD2K Centers' systems pharmacology work solidifies a set of high-quality approaches to the field, giving hope that one day the grand promise of big data will be realized.

## Mapping the Drug Universe

By determining how various healthy and diseased cells respond to a wide range



This fireworks plot displays the universe of cellular responses to drugs. Each spot represents one of 17,041 significant drug-induced gene expression signatures for 3,713 drugs and other compounds applied to 63 cell lines in 3 time points and 51 dosages. Colors represent different cell types while the boxes indicate a few of the cellular states induced by specific types of drugs. This visualization enables assigning function and mechanism to new small molecules, suggesting their potential to serve as drugs. Courtesy of the Ma'ayan Lab and the BD2K-LINCS-DCIC.



## RELEVANT NIH INSTITUTES:

**NCI, NHLBI, NIDDK, NINDS and all other disease-focused Institutes**

of perturbations (by drugs and other chemical compounds in varying doses; reagents that mutate, activate and deactivate genes; and changing the micro-environment), researchers could perhaps map the universe of cellular phenotypes and drug responses. That is one goal of the NIH's Library of Integrated Network-based Cellular Signatures (LINCS) program. Now in Phase II, the LINCS program has generated a vast quantity of gene expression, proteomic and epigenetic data, and the LINCS-DCIC (a BD2K Center) is building that map.

"We are interested in mapping the chemical space to the cellular phenotype space through the molecular signature space, and that will give us a global view of cells, all their states, how they respond to small molecules, and then how those match to cellular phenotypes and diseases and drugs," says **Avi Ma'ayan, PhD**, principal investigator of the BD2K-LINCS-DCIC.

At this point, Ma'ayan and his colleagues are building an interactive web

states that can be well-defined and those states can be associated with disease states, and then you can use drugs to manipulate the system in the direction that you want," Ma'ayan says.

This kind of map is a global goal for biomedical research in general, Ma'ayan says. "By pushing cells in different directions, drugs make a perfect case study."

### Predicting Drug Response

Mayo Clinic cancer researchers associated with the KnowEnG Center are also perturbing cells but with a different goal: They are most interested in which cells die in response to chemotherapy drugs. "We know that the same drug given to different patients elicits different responses," says **Saurabh Sinha, PhD**, principal investigator of the KnowEnG Center. "So this is just repeating that observation in a controlled setting in cell

data to phenotype, KnowEnG researchers took several different lines of attack. One example: Even if they couldn't accurately predict the response of each individual, could they at least identify the most important genes whose variation from individual to individual are predictive of the phenotypic differences? "It might not be a 90 to 100 percent accurate final model," Sinha says, "but if we can identify the most significant genes related to the underlying biology, then we can follow up with more targeted biological studies."

As another example, they set out to identify pathways (rather than individual genes) implicated in drug response. Genes tend to work together as part of complicated pathways of interaction. "Are there pathways triggered or not triggered

---

**"We are interested in mapping the chemical space to the cellular phenotype space through the molecular signature space, and that will give us a global view of cells, all their states, how they respond to small molecules, and then how those match to cellular phenotypes and diseases and drugs," says Avi Ma'ayan.**

---

page that will report—for many of the drugs studied by LINCS researchers—the pathways that a drug potentially targets, the genes that are up/down regulated, and other small molecules that are similar to that drug. "We're trying to visualize this space of drug perturbations," Ma'ayan says. The result is a plot of a network that reveals how small molecules in general affect gene expression and cluster into several responses associated with cellular phenotypes. In this global picture of what happens to cells when they are exposed to drugs, the space of responses is not infinite. "It's likely going to be about 100

lines." The resistant cells are the problem: "You'd like to know why they are resistant," he says. So, for each individual cell line, the researchers also sequence the DNA and measure gene expression and methylation patterns before treatment. The goal: to determine whether these high-dimensional data (millions of gene variants and DNA methylation spots as well as tens of thousands of gene expression measurements) can accurately predict whether a particular drug would or would not work on a particular patient.

To tackle the computational and statistical challenges of relating multi-omics

leading to differences in the phenotype?" Sinha says. If so, then follow up studies can confirm findings and perhaps design drugs to target those pathways.

A third strategy traced gene expression back to the transcription factor (TF) responsible for controlling that gene expression. "If we find that a whole bunch of genes are changing their expression levels in a particular individual, then it's reasonable to hypothesize that these changes were regulated by some transcription factor," Sinha says. Instead of predicting individual genes as key players, this approach predicts that one

transcription factor is an important regulator of those key players. “This has the possibility of statistically reducing the noise,” Sinha says. And in fact they found that to be the case. “We were able to identify a small number of TFs for each drug that might play a role in drug response variation,” he says. And they experimentally validated their results for several drugs by knocking down TFs and seeing the expected drug response changes. These results could also help in designing appropriate ways to overcome chemotherapy resistance.

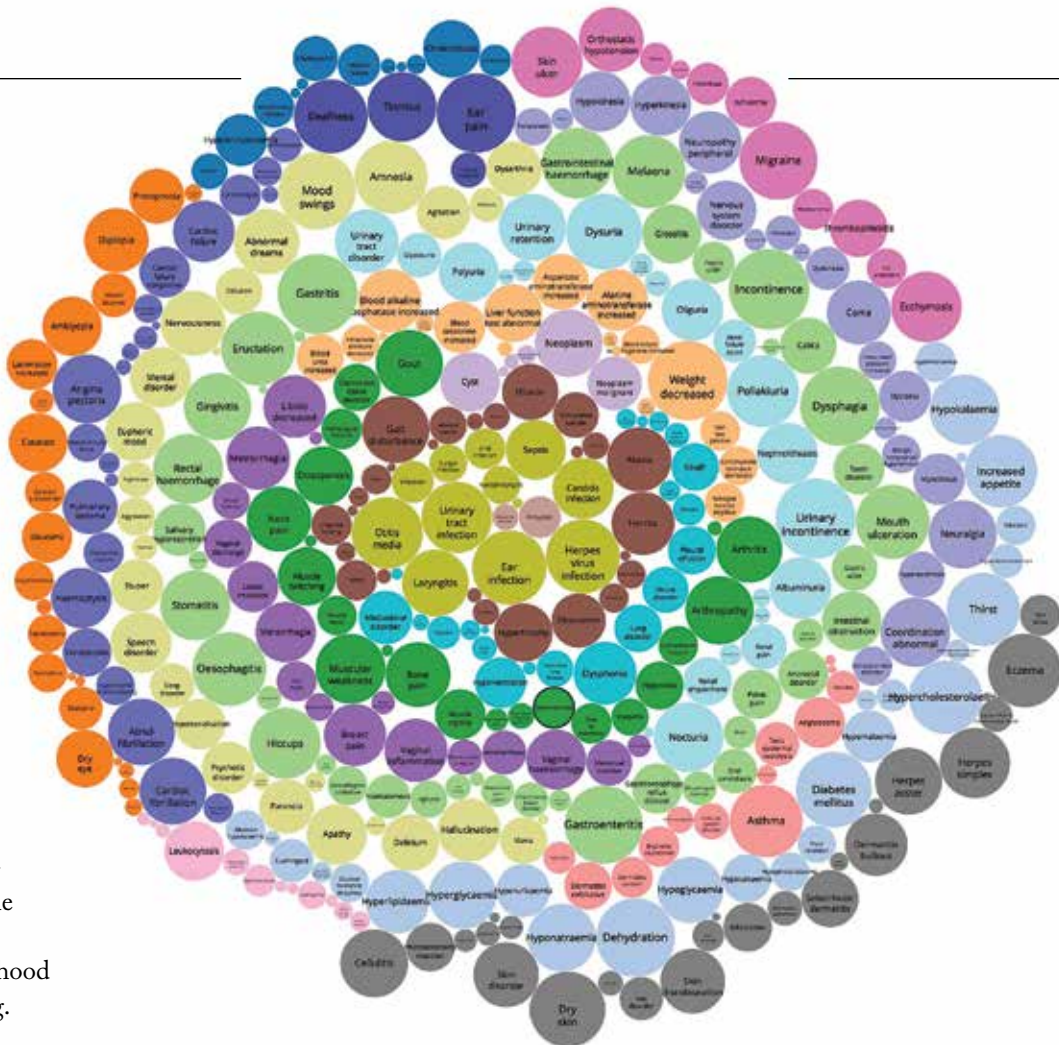
The team is also working on the original problem of building a predictor of drug response levels using all the multi-omics data with the intent of outputting a single number: the likelihood that the patient will respond to a drug.

### Predicting Adverse Drug Reactions

When some friends at the FDA approached Ma’ayan to see if the LINCS’ gene expression data could predict adverse drug reactions for a specified group of drugs, he gave a surprising response: “We can do it for all drugs.”

Other researchers have tried to predict side effects from drug structure alone. Ma’ayan’s group integrated that structural information with LINCS gene expression signatures for 20,000 compounds (including the subset of FDA-approved drugs) and showed that combining these two types of information improved adverse drug event predictions. “This can be helpful to the FDA, which could use computational methods to assess potential toxicity of new compounds,” Ma’ayan says. LINCS-DCIC also developed a web portal for browsing and searching connections between small molecules and adverse drug reactions.

Right now, Ma’ayan says, “This is ready as a suggestive tool, not as a primary approach.” With time, these kinds



of computational approaches will become mainstream, he says.

### Drug Repositioning Tools

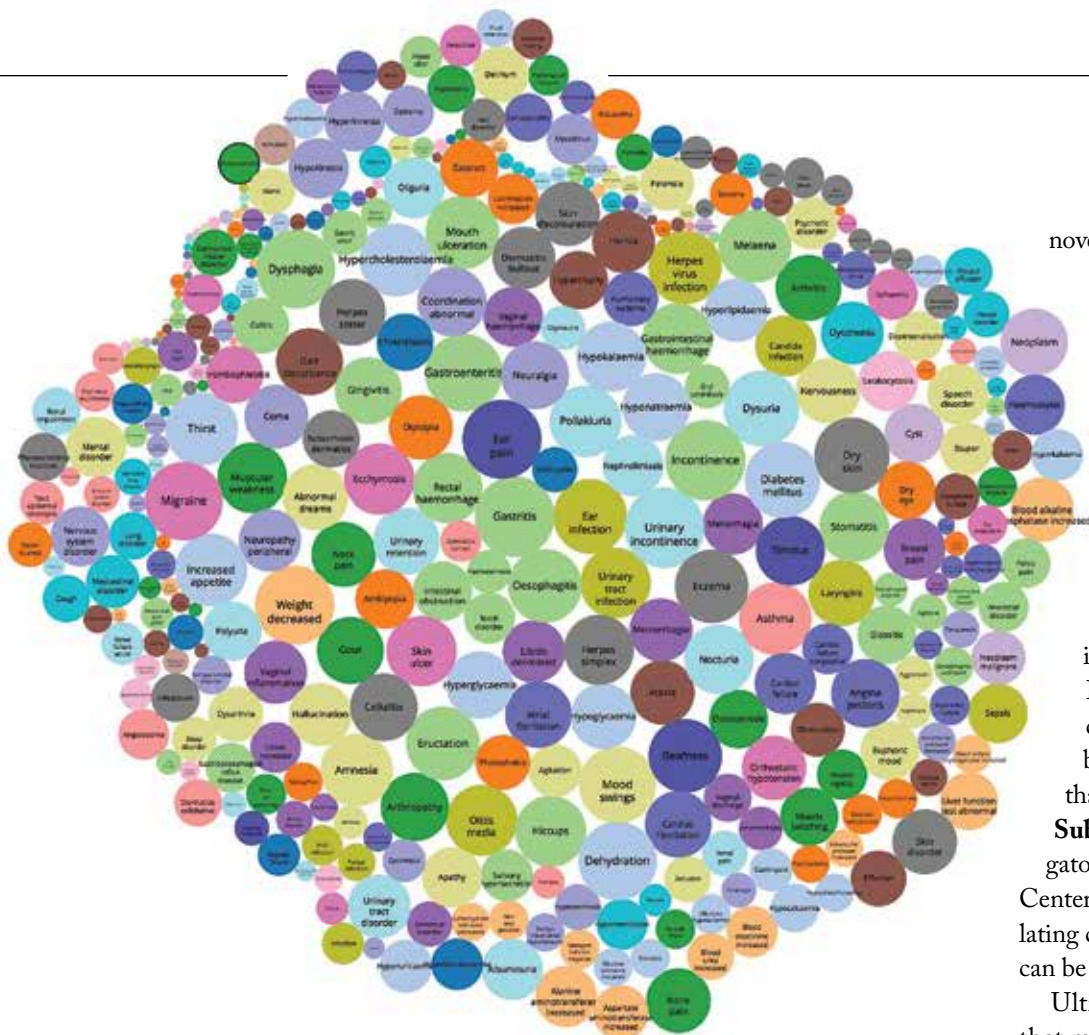
Several BD2K Centers are involved in the computational effort to discover new uses for existing FDA-approved drugs. This makes a lot of sense: Big data will likely prove useful in this effort, and computational drug repositioning can save a lot of money while benefiting many patients. At PIC-SURE, researchers in **Chirag Patel**’s lab have developed tools that will make it easier for anyone to do drug repositioning research.

Frustrated that existing drug repositioning tools required specific data sources or formats, they created a tool called ksRepo that allows researchers to greatly expand the datasets usable to generate predictions about potential drug repositioning candidates.

In addition, concerned that computational researchers were each using a different database to validate their drug repositioning methods, they developed repoDB, a set of standardized drug successes and

*LINCS-DCIC combined drug structural information with gene expression profiles to predict adverse drug reactions for the 20,412 drugs and small-molecule compounds profiled by the LINCS L1000 project. These bubble plots show distinctly different sorting pattern for the side effects when sorted by the system and organ affected (above) versus by drug similarity (opposite). The team also created a freely available web portal at <http://maayanlab.net/SEP-L1000/> where each drug and adverse drug reaction has a dedicated page with a list of the relevant predictions and external links to relevant sites. Courtesy of the Ma’ayan Lab and the BD2K-LINCS-DCIC.*

failures drawn from DrugCentral and ClinicalTrials.gov. “It’s important to have a consistent benchmark set that everyone uses so you can say, ‘my method outperforms this method using this same benchmark,’” says **Adam Brown**, a graduate student in biomedical informatics at Harvard Medical School and member of the PIC-SURE team. “Without that consistency, you just cherry-pick the dataset that fits your story.” Many researchers were also calculating sensitivity and specificity without true negatives (failed drug candidates). RepoDB addressed that problem as



well. “To our knowledge, it’s the only database that includes both approved and failed drugs,” Brown says. “This is something the field has really been missing.”

RepoDB will be particularly useful in studies where researchers are trying to predict associations between

all diseases and all drugs, Brown says. “Hopefully people will use it.”

In another effort to make drug repositioning research more reproducible, the **Broad Institute’s LINCSCenter for Transcriptomics and Toxicology (LINCSCenter)** is creating a

novel comprehensive screening library called the Broad Drug Repurposing Hub. First they identified and created a physical collection of 5,000 compounds, including more than 3,000 drugs of interest, and they curated them as a means of quality control. They then distributed them to anyone interested in screening them in their assays (gene expression, cytotoxicity, proteomics and morphology). But there is a hitch: “It’s a hub to distribute reagents, with the price being contributing the data back so that others can use it,” says **Aravind Subramanian, PhD**, principal investigator for the LINCSCenter. The hub has already begun accumulating curated, quality-controlled data that can be used for drug-repositioning research.

Ultimately, identifying existing drugs that might cure or alleviate symptoms of rare diseases could give patients hope of a treatment. “That’s something I’m pretty passionate about,” Brown says. “It’s important to get good drug/disease pairs into the hands of clinicians.”

## BD2K Systems Pharmacology in Context

Plenty of systems pharmacology research happens beyond the BD2K context, Sinha says. But the BD2K Centers have brought a big picture view to the field as well as a sense of gravitas: Doing this research well and reproducibly requires reliable data such as that generated by the LINCSCenter; well-designed and validated analytical tools such as those BD2K-LINCSCenter-DCIC and KnowENG are building; and quality controls, incentives for data-sharing, and standardized benchmarking and validation procedures such as those being modeled and made publicly available by PIC-SURE and the LINCSCenter. □

## DETAILS

### BD2K Drug Repositioning Tools

#### PIC-SURE:

**ksRepo:** a generalized tool that expands the datasets usable to generate predictions about potential drug repositioning candidates (freely available for download at <https://github.com/adam-sam-brown/ksRepo>)

**RepoDB:** a standard set of drug repositioning successes and failures that can be used to fairly and reproducibly benchmark computational repositioning methods. (freely available for download at <http://apps.chiragjgroup.org/repoDB/>)

**MeSHDD:** uses MeSH-term enrichment to discover literature-based similarities between FDA approved drugs (interactive online app at <http://apps.chiragjgroup.org/MeSHDD/>)

#### Broad-Transcriptomics

**Broad Drug Repurposing Hub:** a best-in-class drug screening collection with more than 3,000 clinical drugs (<https://clue.io/repurposing>)

## BD2K CENTERS SIZE UP BRAIN DISEASE

About a year and a half ago, brain imaging researchers at the University of Southern California (USC) were shooting the breeze over salad and roast beef sandwiches when their lunch conversation took a turn. A skinny guy named **Arthur Toga, PhD**, confessed to his colleagues that he takes Lipitor—a common cholesterol-lowering drug prescribed to prevent heart attacks and strokes. Toga's total cholesterol had climbed above 200, prompting his cardiologist to recommend the statin therapy. "He said, 'this stuff should be in water like fluoride—there's no harm to it. Everybody should take it,'" Toga recalls.

Others at the lunch table weren't so sure: Because cholesterol is vital for brain health, they wondered if reducing cholesterol with statin therapies could lead to cognitive problems—perhaps even increase the likelihood of dementia.

Past studies looked for links between statins and Alzheimer's disease risk but were either too small, with sample sizes in the hundreds, or dealt with limited types of data. And their findings were mixed. "Nothing seemed definitive," says Toga, who runs USC's Laboratory of Neuro Imaging and leads BD2K's **Big Data for Discovery Science (BDDS)** Center there.

of the uncertainties of brain research will fade—helping the field more effectively diagnose and treat brain disease.

### More Data Yield More Definitive Results

Toga and USC colleague **Judy Pa, PhD**, decided to tackle the Alzheimer's question with a big-data approach. They put their computers to work combining through clinical and brain-imaging data from more than 2,100 participants enrolled in various studies at 40 research centers. The goal: look for relationships between statin use, brain structure and Alzheimer's disease status. Since the literature on statins and Alzheimer's is murky, says Pa, "we did not know where the results would take us."

The number crunching revealed a surprise: Statin use does appear to raise Alzheimer's risk but only in women. The findings have been submitted for publication.

This is just one example of the types of analyses made possible by the rise of big data. "If you don't have enough data, you can't possibly do something like this," Toga says. Traditionally, researchers start with a hypothesis and then go collect data to see if it supports the idea. But in the realm of big data, "we have the opportunity to not articulate a hypoth-

trends, relationships and other interesting features emerge. Even when researchers come in with certain ideas they hope to test, adds Pa, there are many more questions that can be asked of the data.

Recently BDDS researchers posed a particularly tough question: Using large quantities of complex, heterogeneous data from multiple centers and studies, can computers learn to identify which people have Parkinson's disease?

To find out, the team used data from the Parkinson's Progression Markers Initiative (PPMI). This \$60 million observational study launched in 2010 to find biomarkers for Parkinson's disease, which afflicts about 10 million people worldwide. PPMI has collected data and samples from nearly 1,000 participants—some with Parkinson's, some without—at 33 clinical sites in 11 countries.

The PPMI has gathered many kinds of data in vast quantities, including brain scans; medication histories; genotypes; and exam results reflecting answers to questions such as whether the person has cognitive issues, difficulty smelling, or the ability to pass a finger tapping test. All that information gets codified. Also, because people join the study at different stages of disease, the computer has to learn to assess differences in disease severity.

---

**With enough images and associated data, perhaps some of the uncertainties of brain research will fade—helping the field more effectively diagnose and treat brain disease.**

---

It's a common problem, as such analyses require huge numbers of brain images. Big data offers a solution: With enough images and associated data, perhaps some

of the uncertainties of brain research will fade—helping the field more effectively diagnose and treat brain disease. "Rather, we say to the data collection, tell me about yourself." Then they let machines sort through huge volumes of data and see what

To further complicate matters, the machines need to recognize different notations for the same information. "Somebody might code sex as

‘M’ or ‘F,’ ‘0’ or ‘1,’ ‘man’ or ‘woman,’ or ‘male’ or ‘female.’ A computer has no idea that those are all the same,” Toga says. “You have to teach it.”

The training seemed to work. Several machine-learning methods in the BDDS study—published August 2016 in *PLoS ONE*—correctly classified people as having Parkinson’s or not with greater than 95 percent accuracy, sensitivity and specificity. Previous studies using machine learning and data-mining methods to recognize Parkinson’s reported just 70 to 90 percent sensitivity.

The ultimate goal is to train computers to predict who’s on the verge of Parkinson’s in advance of symptoms, Toga

says, in order to be able to slow disease progression—similar to how doctors nowadays prescribe statins to people with high cholesterol hoping to prevent future heart disease.

### Cracking the Brain’s Structural and Genetic Code

In the field of neuroimaging, researchers are studying brain scans to identify structural features that associate with neurological and psychiatric disorders. The **Enhancing Neuroimaging Genetics through Meta-analysis (ENIGMA)**

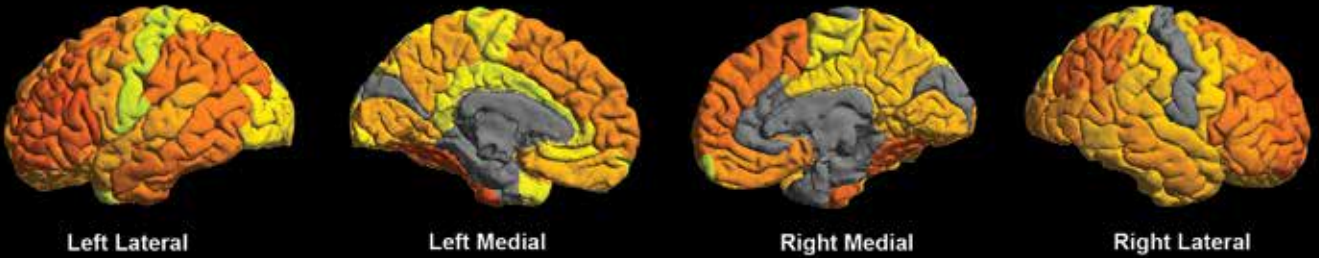
## RELEVANT NIH INSTITUTES:

**NIMH, NINDS, NHLBI, NIA, NIBIB, NIDA, NHGRI, NICHD, and NIAAA**

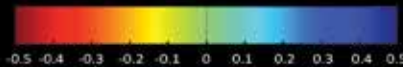
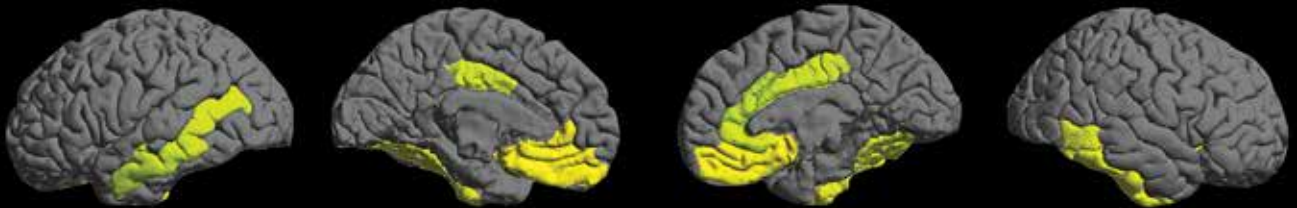
Consortium goes further and tries to look for the genetic underpinnings of these phenotypes and diseases. Because gene effects tend to be subtle, teasing them out requires huge datasets amassed and analyzed with a global team-science approach. Since its launch in 2009, ENIGMA has rallied more than 800 scientists at 340 institutions in 35 countries to crack the genetic code underlying 18 brain diseases.

## ENIGMA Bipolar Disorder and Major Depressive Disorder Cortical Thickness Findings

1,837 bipolar patients and 2,582 healthy controls



1,902 major depression patients and 7,658 healthy controls



The ENIGMA consortium has published the world’s largest neuroimaging studies of bipolar disorder and major depression. In separate studies, ENIGMA researchers observed significant cortical thinning associated with both illnesses. Bipolar disorder (top) was associated with widespread thinning of frontal, temporal, and parietal regions, whereas major depression (bottom) was associated with thinning

in paralimbic regions. Scientists are now working to quantify the similarities and differences between these psychiatric illnesses thanks to harmonized brain measures derived by the ENIGMA consortium. Image courtesy of Christopher Ching, Paul Thompson, and the ENIGMA Bipolar Disorder and Major Depressive Disorder Working Groups.

Poring through magnetic resonance imaging (MRI) scans, along with corresponding clinical and genetic information from pooled datasets, ENIGMA researchers can analyze cohorts 10 to 30 times larger than a typical neuroimaging study.

In some cases, ENIGMA researchers have found that boosting sample size for MRI-based data is enough to gain insight—even without considering genetics. For example, in a January 2017 study in the *American Journal of Psychiatry*, an international team found that children and adults with obsessive-compulsive disorder (OCD) have distinct patterns of subcortical abnormalities. Whereas smaller brain imaging studies in OCD produced mixed results, the conclusions were clear when the ENIGMA team

from 7,957 healthy people and 2,148 depressed patients at 20 sites around the world. Compared to controls, adults with depression—but not children—had thinner cortical gray matter in the orbitofrontal cortex, anterior and posterior cingulate, insula and temporal lobes. Depressed adolescents had different brain abnormalities—namely, lower surface area in frontal regions as well as primary and higher-order visual, somatosensory and motor regions. The large sample size allowed the researchers to distinguish effects in children versus adults, suggesting that depression correlates with brain structure distinctly during different stages of life.

Several recent ENIGMA papers focus more squarely on identifying gene variants that underlie fundamental

“Rather than just download the data... you have a community to help you really dig into a question,” says **Paul Thompson, PhD**, professor of neurology at USC and principal investigator for the ENIGMA BD2K Center.

He compares the situation to wanting to become better at chess. “Let’s say someone says, ‘I really want to be a world-class chess player. I’ve bought all the pieces.’ In fact, my home is full of chess pieces,” Thompson says. But to improve at chess, “I would say they need to be with people who are really active and playing a lot of chess. Really, what’s going to take the science to the next level is working with a large team of experts. The data is a requirement but not the clincher.”

Researchers can propose new stud-

---

**“Really, what’s going to take the science to the next level is working with a large team of experts,” Thompson says. “The data is a requirement but not the clincher.”**

---

pooled 35 sets of structural brain scans from 1,759 healthy controls and 1,830 OCD patients—about a sixth of whom were under age 18. Compared with healthy peers, children with OCD had a larger thalamus, a brain area important for sleep, consciousness and higher-order brain processing. However, in adults with OCD, greater volumes were measured in other brain regions—namely, the hippocampus and the pallidum, an area important for motivating rewards and incentives. These results are in line with the developmental nature of OCD and suggest that further research on neuroplasticity—the brain’s ability to reorganize and form new neural connections throughout life—could be useful.

Combining datasets, as well as separating children and adult subgroups, also proved important in a May 2016 *Molecular Psychiatry* study that revealed cortical differences in people with depression. The analysis pooled MRI scans

brain features and specific diseases. An international team undertook a massive study of more than 32,000 adults at 52 sites. In a paper published October 2016 in *Nature Neuroscience*, the researchers reported identifying seven genes that not only regulate brain volume, memory and reasoning but also seem to influence Parkinson’s disease risk. And in study of people with schizophrenia, ENIGMA scientists found that certain measures of volume and thickness in affected brain regions correlate with gene variants known to confer disease risk. They also found that schizophrenia shares some of these neurogenetic signatures with other psychiatric disorders. These findings appeared October 2016 in *Molecular Psychiatry*.

And it’s not just about pooling data. Each ENIGMA analysis gets vetted by one of 30 working groups—teams of neuroscientists, imagers, geneticists, methods developers and others devoted to a specific disease or subfield of study.

ies by joining the monthly phone calls held by each working group. The calls update members on the group’s ongoing projects and offer a chance for people with new ideas to thrash them out.

### A Networked Brain: Discovering Causal Relationships

Beyond structure and genetics, the brain can also be viewed as a network.

Another BD2K Center—the **Center for Causal Modeling & Discovery of Biomedical Knowledge from Big Data (CCD)**—renders big data as networks. And it connects the network’s nodes, or variables, not with mere lines but arrows. “Our business is computer algorithms that will find causal relations from measured data,” says **Clark Glymour, PhD**, a professor of philosophy at Carnegie Mellon University (CMU) who leads the CCD group focused on the brain.

The basic algorithm was developed

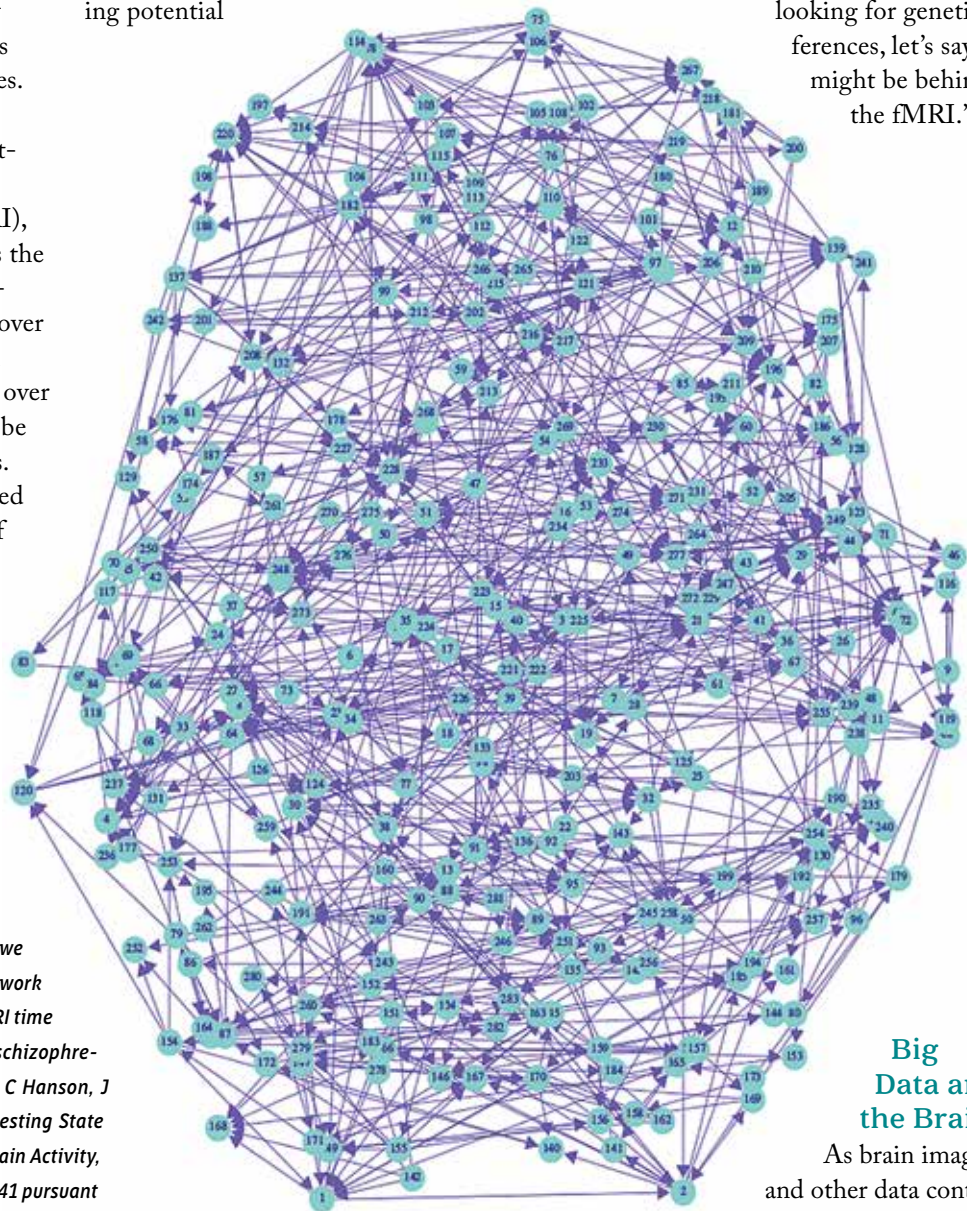
by a CMU graduate student. It could handle 15 to 20 variables—features that take different numerical values over time. About five years ago, with careful programming Glymour’s team got it to run on a few hundred variables. And with further improvements last year the algorithm, called Fast Greedy Equivalence Search (FGES), now runs in about 12 hours on a million variables.

As one test case, CCD decided to apply the FGES algorithm to the resting state brain. Much of their work uses data from functional MRI (fMRI), which approximates neural activity as the amount of energy consumed by thousands of tiny subregions of the brain over time. Generally fMRI produces a full image of the brain every few seconds over a 15 to 20 minute period. And it can be easily performed on many individuals. Highly scalable algorithms are required to make sense of the large quantity of data produced by a typical fMRI-

*CCD researchers used functional MRI (fMRI) data to build causal networks that depict different patterns of brain connectivity in normal versus autistic and schizophrenic individuals. They found that the networks of neuro-atypical individuals exhibited more fractionated and less controllable dynamics in the resting state brain. They even showed that disease severity was revealed by the global topology of the network itself. Here we see the graphical structure of the pattern of network connectivity in the resting state brain (from fMRI time series) aggregated over 10 individuals with schizophrenia. Reprinted from SJ Hanson, D Mastrovito, C Hanson, J Ramsey, C Glymour, Scale-Free Exponents of Resting State are Biomarkers of Neuro-Typical and Atypical Brain Activity, biorxiv.org, doi: <https://doi.org/10.1101/068841> pursuant to Creative Commons license CC-BY-NC-ND 4.0.*

based research study. “What the fMRI work does is give us a really, really hard case for making the best algorithms we can,” says Glymour. In a recent study posted to *bioRxiv* in August 2016, CCD researchers analyzed resting-state fMRI data from one healthy adult, 60 people

with autistic spectrum disorder, and 60 with schizophrenia. Applying the FGES algorithm to the reams of scan data produced causal networks that depict different patterns of brain connectivity in normal versus neuro-atypical individuals. But beyond providing potential



## Big Data and the Brain

As brain images and other data continue to accumulate, the tools developed by the BD2K

There are very likely to be multiple causes and types,” says **Greg Cooper, MD, PhD**, professor of biomedical informatics at the University of Pittsburgh and director of CCD. And if robust patterns were to emerge within the autism group, adds Glymour, “you could start looking for genetic differences, let’s say, that might be behind the fMRI.”

diagnostic information—for example, being able to distinguish healthy individuals from people with autism—the connectivity patterns could help researchers sort autism cases into different subgroups. “These conditions are almost certainly not single monolithic diseases.

Centers are setting the standard for high quality computational neuroscience. Using big data to discover how the brain works in health and disease is becoming routine, as researchers address questions raised over roast beef sandwiches from the comfort of a single workstation. □

# MOBILE HEALTH: BD2K CENTERS HARNESS SENSOR DATA

**H**aving already revolutionized fields ranging from communications to finance, mobile technology and data science are now poised to do the same for healthcare.

That, at least, is the promise of the burgeoning mobile health (mHealth) movement. Thanks to the proliferation of wearable biosensors capable of recording everything from physical activity to blood oxygen levels—and the increasingly sophisticated algorithms used to sift through the mounting pile of data—researchers are finding novel ways of diagnosing illnesses, predicting disease risk, and promoting healthier lifestyles.

Moreover, two NIH Big Data to Knowledge Centers of Excellence—the **Mobilize Center** and the **Center Mobile Sensor Data-to-Knowledge (MD2K)**—are paving the way for the entire mHealth community.

“The research methodologies, algorithms and devices these centers are developing—not to mention the training opportunities they provide—are creating a foundation that will make it easier for others to produce robust mobile health research,” says **Scott Delp, PhD**, professor of bioengineering at Stanford University and principal investigator of the Mobilize Center.

## Disease Detection with Smart Devices

Some Mobilize Center researchers are leveraging consumer products that are already used by large numbers of people. **Jessilyn Dunn, PhD**, a postdoctoral fellow at Stanford University, recently evaluated the possibility of using commercially available wearables to gather and analyze health-related information in ways that aren't normally done in the clinic.

As reported in an article published in January 2017 in *PLoS Biology*, Dunn

and her colleagues, including **Michael Snyder, PhD**, professor of genetics at Stanford and director of the Center for Genomics and Personalized Medicine, performed several different experiments using a variety of wearables. They found that two commonly used tools provided most of the information they needed: a smartwatch capable of detecting heart rate, skin temperature, and activity; and a smartphone capable of reporting activity and location.

When combined with the occasional use of a wearable oxygen sensor, these devices collected much of the same information that would ordinarily be recorded in a doctor's office once every year or so. In this case, however, the data were gathered regularly—often continuously—over a lengthy period: one study participant was monitored for two years, while an additional 43 participants were monitored for an average of 11 months.

Reconciling the different data formats used by competing companies wasn't easy, especially when manufacturers periodically changed the way they packaged the output from their products. As a result, Dunn says, the team spent a lot of their time cleaning the data and “making sure that everything was kosher from one iteration to the next” to ensure they were “comparing apples to apples.”

Participants also underwent blood testing on a regular basis. This allowed the researchers to unearth several interesting findings that were lurking in the sensor data.

For example, the researchers retrospectively detected the onset of a viral infection—in this case Lyme disease—in one subject based solely on elevated skin temperature and unusual heart-rate patterns. This prediction was confirmed by the presence of Lyme bacteria antibodies in his blood. Delving more deeply

into the subject's data, the researchers identified several other periods of illness during which similar abnormalities in temperature and heart rate were accompanied by the presence of an inflammatory biomarker known as high-sensitivity C-reactive protein in his blood.

Based on his data, the researchers developed an algorithm, called Change-of-Heart, that identified instances of illness amongst several other participants before

---

**“The research methodologies, algorithms and devices these centers are developing—not to mention the training opportunities they provide—are creating a foundation that will make it easier for others to produce robust mobile health research,” says Delp.**

---

they reported symptoms, based solely on abnormalities in their heart rates.

Eventually, that kind of predictive capability could allow sensor-based systems to warn people of an impending illness even before they feel sick—enabling an algorithm to tell you to “run to your local pharmacy and pick up some cold medicine, because tomorrow you're going to wake up with a cold,” Dunn says.

In a similar vein, Dunn and her



## RELEVANT NIH INSTITUTES:

**NIMH, NINDS, NHLBI, NIA, NIBIB, NIDA, NHGRI, NICHD, and NIAAA**

colleagues successfully identified sensor-based predictors of insulin resistance, a risk factor for type II diabetes, which they confirmed by testing steady-state plasma glucose (SSPG) levels among a subset of study participants.

The researchers started with clinically measured body mass index (BMI), and added sensor-based reports of both physical activity and heart rate—in particular, differences between day and nighttime heart-rate patterns that they found to be associated with diabetes. While the researchers found that they could best predict insulin resistance if they used all three parameters in combination, variation in heart rate proved to be the strongest biomarker of the lot, and was an effective predictor even in the absence of the other two.

Given their usefulness, Dunn hopes that as wearables become cheaper, they will help expand healthcare access to low-income groups and people in remote rural communities, many of whom cannot easily see a living, breathing doctor. “This is really a fantastic public health opportunity,” she says.

In the meantime, the rich dataset she and Snyder created is available (at <http://hmpdacc.org/data/wearable/stanford.tar>) for others to explore.

### Boosting Health with Games and Social Networks

**Tim Althoff, MS**, a doctoral candidate in computer science who is also affiliated with the Mobilize Center, is

*Mobilize Center researcher Tim Althoff and his colleagues at Microsoft showed that, before starting to play, Pokémon Go users are less active than average users of the leading consumer health apps (A, B, C, and D), but they experience larger increases in physical activity after starting to play (at  $t_0$ ). To determine when a person was actually playing Pokémon, the researchers distinguished between search queries that suggested a user was merely seeking general information about the game, and “experiential queries” that indicated he or she was actively using it. Reprinted from Althoff T, White RW, Horvitz E, Influence of Pokémon Go on Physical Activity: Study and Implications, J Med Internet Res 2016;18(12):e315.*

trying to promote population-scale health benefits as well. But his tool of choice is the online social network. And rather than dealing with dozens of study participants, he’s working with thousands, even millions, of them.

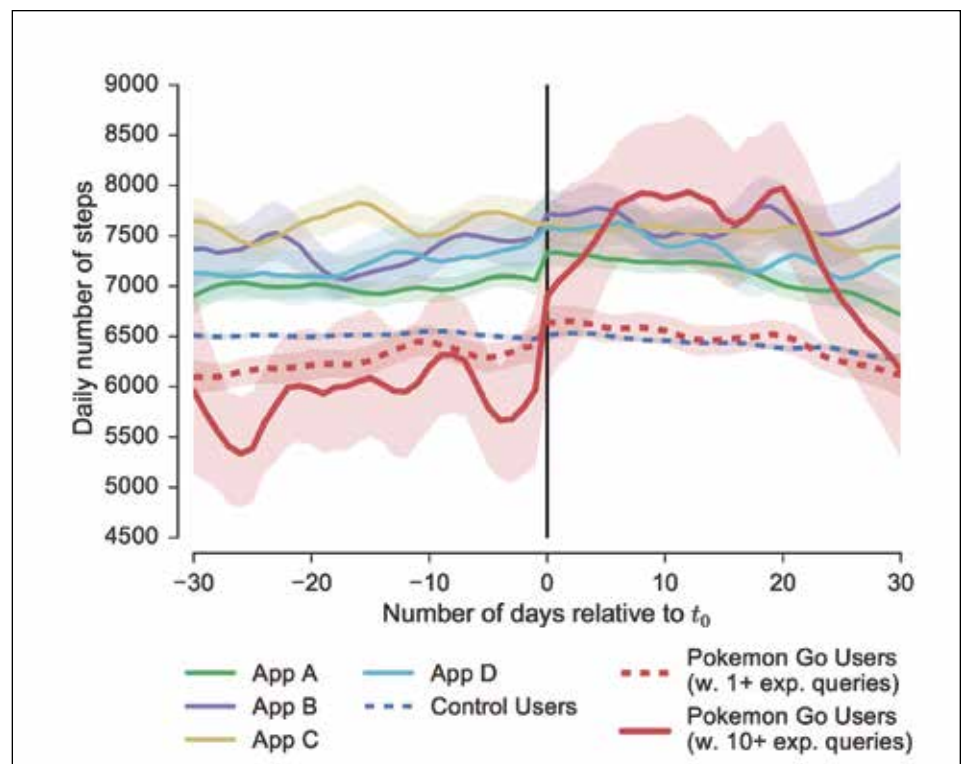
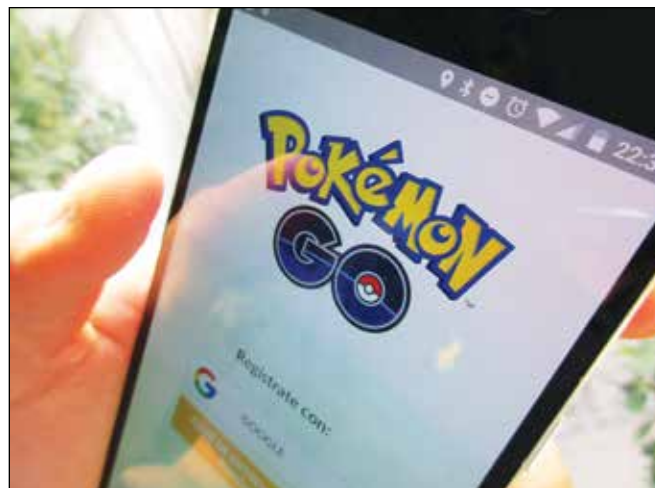
In a paper published in the *Journal of Medical Internet Research* in 2016, Althoff and his colleagues demonstrated that users of the Pokémon Go augmented-reality

game—a group whom Althoff describes as being “way less active than the typical health-app user”—did in fact become more physically active because of their involvement with the app.

Althoff and his co-authors had access to data from nearly 32,000 users of the

Microsoft Band wearable fitness tracker. By algorithmically combing through all the Go-related queries that cohort posted to the Bing search engine, the researchers

*In the summer of 2016, the Pokémon Go app became a huge fad that got people all over the world walking around to capture Pokémon characters at virtual locations in their neighborhoods.*



inferred that just over 1,400 of those Band users were actively engaged with the game. And by examining users' Band data, they determined that Pokémon Go players walked as much as 1,473 extra steps a day—an increase of more than 26 percent over their prior activity levels.

Most of the people who use mobile health apps are already physically active. But Althoff's study suggested that even sedentary, obese, and older users benefited from playing Pokémon Go. And while the benefits of gameplay did fall off after three or four weeks, given the well-established link between physical activity and mortality risk, Althoff and his colleagues suggested that active engagement with the game—which has 65 million monthly active users—could nonetheless have a measurable impact on life expectancy.

The work also has immediate value to the research community. “The Pokémon Go project demonstrated how to do a physical activity study across more than 30,000 people with a very specific treatment,” Althoff says—the treatment being playing Pokémon Go. “It shows how you can contextualize wearable data in a way that allows you to test large-scale interventions like this and provides a model for conducting such studies in the future.”

In a pair of studies published this year, Althoff also exploited wearables to explore the real-world impact of participation in online social networks and app-based competitions.

Althoff and his co-authors, including **Jure Leskovec, PhD**, who spearheads the social and behavioral modeling effort at the Mobilize Center, analyzed data from Argus, a fitness-tracking app developed by the Silicon Valley startup Azumio. The app allows users to create posts about their physical activity (walking, cycling, yoga, etc.), and uses the accelerometers in their smartphones to unobtrusively track their physical activity.

In the first paper, Althoff wanted to see if participation in an online social network organized around fitness would affect physical activity in the real world. Argus

provided the ideal data set: Althoff and his colleagues had access to anonymized information provided by 6 million users from 2011 to 2016—amounting to 10,000 times more users and a million times more activity tracking than most comparable studies—but the app's embedded social network was only added in 2013. This allowed the researchers to observe changes in physical activity among users who joined the network, and to compare their results to those of users who did not join.

Sure enough, people who made new social connections through the app increased their physical activity by approximately 7 percent, or 400 steps per day. Algorithms designed to tease out different kinds of effects, such as changes in internal motivation versus the influence exerted by new social connections, showed that 55 percent of the observed changes in user behavior were due to social influence. And Althoff and his colleagues developed a model that could predict which users would be most influenced by new social network connections—something that could contribute to the design of more effective apps in future.

In the second study, Althoff and his collaborators examined the impact of app-based fitness competitions on users' activity levels. They analyzed the data generated by 3,637 users who participated in 2,432 physical activity competitions over a 10-month period—again, the largest data set of its kind to date. And by considering factors such as age, gender, and prior activity level, the researchers were able to study which features of competition design were most likely to boost participants' activity levels. For example, competitions were most effective when participants shared similar levels of prior activity, and when there was a balanced mix of men and women.

According to Althoff, those kinds of insights could guide the creation of better app-based competitions. And that, in turn, would further his overall goal of optimizing online communities and mobile health apps “to help

people be healthier and happier.”

But the immediate impact of the project will be most acutely felt in the mHealth research community.

“Identifying social influence in observational network data is extremely challenging but very important for interventions,” Althoff says. The Azumio project tested a new causal inference technique based on “delayed friendship acceptance,” he says. “This worked really well and had never been done before.”

### A Comprehensive mHealth Platform and Specialized Sensors

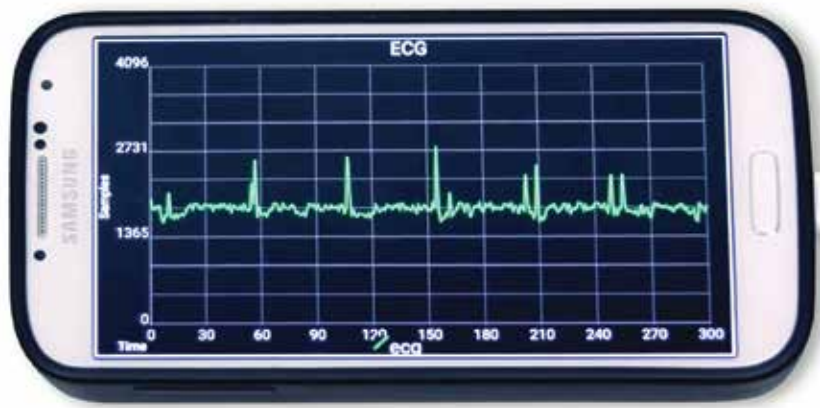
Dunn and Althoff rely on commercially available apps and hardware to supply the data that fuels their algorithms. But the researchers behind MD2K—a Big Data to Knowledge Center of Excellence that brings together experts from 12 universities and the nonprofit startup Open mHealth—are taking a different route.

While they make use of some off-the-shelf wearables, they have also built their own novel sensors. In addition, they have developed an open-source mobile phone-based platform called mCerebrum, which includes more than 20 apps that combine and process the data from those various sources to support the discovery and validation of digital mobile-health biomarkers; and a cloud-based big-data component called Cerebral Cortex, which supports population-scale data analysis, visualization, and modeling.

According to **Timothy Hnat, PhD**, chief software architect for MD2K, mCerebrum can collect up to 70 million data samples per person per day, processing them in real time and running them through predictive models to trigger just-in-time health interventions. And because it's an open-source platform, it can be used by other researchers—with the potential to have a significant impact on mHealth research well beyond the BD2K program.

The mCerebrum sensor data is also synced to Cerebral Cortex, where

MD2K has developed mCerebrum, an open source, real-time software platform for data collection from sensors in smartphones and wearables. It can capture data from multiple sensors simultaneously while continuously evaluating data quality. It also allows for real-time data viewing, such as a live ECG signal. And it uses advanced analytics to convert the data into markers of health, behavior, and risk factors. The gyroscope data from a wrist sensor might, for example, provide insight into smoking or eating behaviors by revealing the telltale hand gestures involved in bringing a cigarette or piece of food to one's mouth. Images courtesy of MD2K.



health science researchers can visualize and interpret the information being gathered from study participants—and where data science researchers can perform large-scale machine-learning exercises to refine their algorithms.

Those refinements, in turn, trickle back down to mCerebrum, where the models that decide whether an intervention is required (and what kind of intervention it should be) can be fine-tuned on an individual basis, opening the door to data-driven personalized medicine.

Since the program's inception in 2015, researchers have used MD2K's best-of-breed wearables and software to track people's stress patterns and smoking behaviors, delivering alerts and behavioral exercises to help calm them or prevent them from taking a puff.

Now they are hoping to use a home-grown device called EasySense to more effectively treat congestive heart failure. This potentially fatal illness characterized by fluid buildup in the lungs affects almost 6 million Americans.

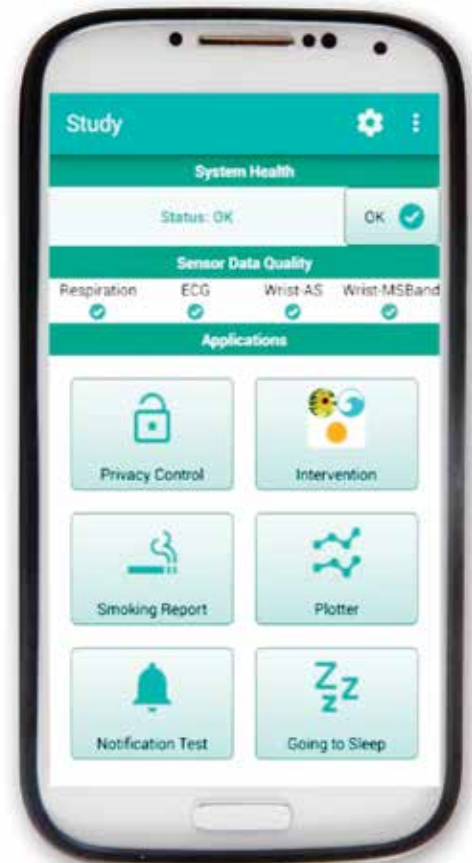
Unfortunately, says **William Abraham, MD**, the cardiologist at Ohio State University College of Medicine who leads the project, the standard method of managing congestive heart failure doesn't work very well.

Patients monitor their own symptoms and body weight (an indicator of fluid buildup), but rarely receive treatment until a significant amount of fluid has accumulated in their lungs—at which point “the horse is already out of the barn.” Relapses are common, and readmission rates are higher than for any other cause of medical hospitalization.

EasySense could help change that. The device, which is roughly the size of a hockey puck, emits pulses of ultra-wideband radio waves and listens to the echoes they create as they bounce off various bodily tissues. According to **Emre Ertin, PhD**, an electrical engineer at Ohio State who leads the development of MD2K's custom-built sensors, this allows EasySense to provide nearly EKG-quality heartbeat detection—and to gauge the fluid content of the lungs.

Abraham recently concluded a 20-person pilot study that successfully demonstrated the device could gather useful data in a hospital setting. He and his collaborators are now beginning a 75-person study in which participants will take the sensors home with them. By analyzing the data provided by EasySense along with the output from other wearables that record parameters such as respiration and oxygen saturation, Abraham hopes to determine which signals are most predictive of relapse and rehospitalization. A third and final study will then use that information to make treatment changes “to see whether or not we can actually keep patients out of the hospital,” he says.

The goal, Abraham says, is to have mCerebrum send alerts and notifications directly to patients and their doctors before things get out of hand. They might, for example, suggest the need to reduce salt intake or prescribe an extra dose of diuretics when fluid levels begin to rise. As the data set grows, he expects patterns will emerge that will allow the team to tailor interventions on an individual basis.



Like the projects at the Mobilize Center, the work being done by Abraham and his MD2K colleagues promises to harness the Big Data generated by wearable biosensors to drive improvements in both personal and public health. And given the ever-increasing ubiquity of wearables (and the ever-increasing sophistication of data science), it's likely that the mobile health revolution is just getting started, spurred on by the methods and devices generated by the BD2K Centers—and available to the entire research community. □

## DISEASE DETECTIVES: BD2K CENTER RESEARCHERS SLEUTH FOR EARLY SIGNS OF DISEASE

**M**ost people don't know that they're sick until they feel, for lack of a better word, *sick*. Like storms, diseases quietly brew and gather strength before wreaking havoc. For the weather, however, you can turn on your local news channel and check next week's forecast. Not so for disease. Not yet, anyway. Researchers across the NIH Big Data to Knowledge (BD2K) Centers are pursuing innovative, data-driven strategies to predict disease and its progression.

Such predictions would help doctors and scientists alike, says **Mark Craven, PhD**, professor of biostatistics and medical informatics at the University of Wisconsin-Madison (UW-Madison) and director of the BD2K **Center for Predictive Computational Phenotyping (CPCP)**.

For many conditions, if you can predict that it's headed your way, Craven says, "that can give clinicians some kind of guidance."

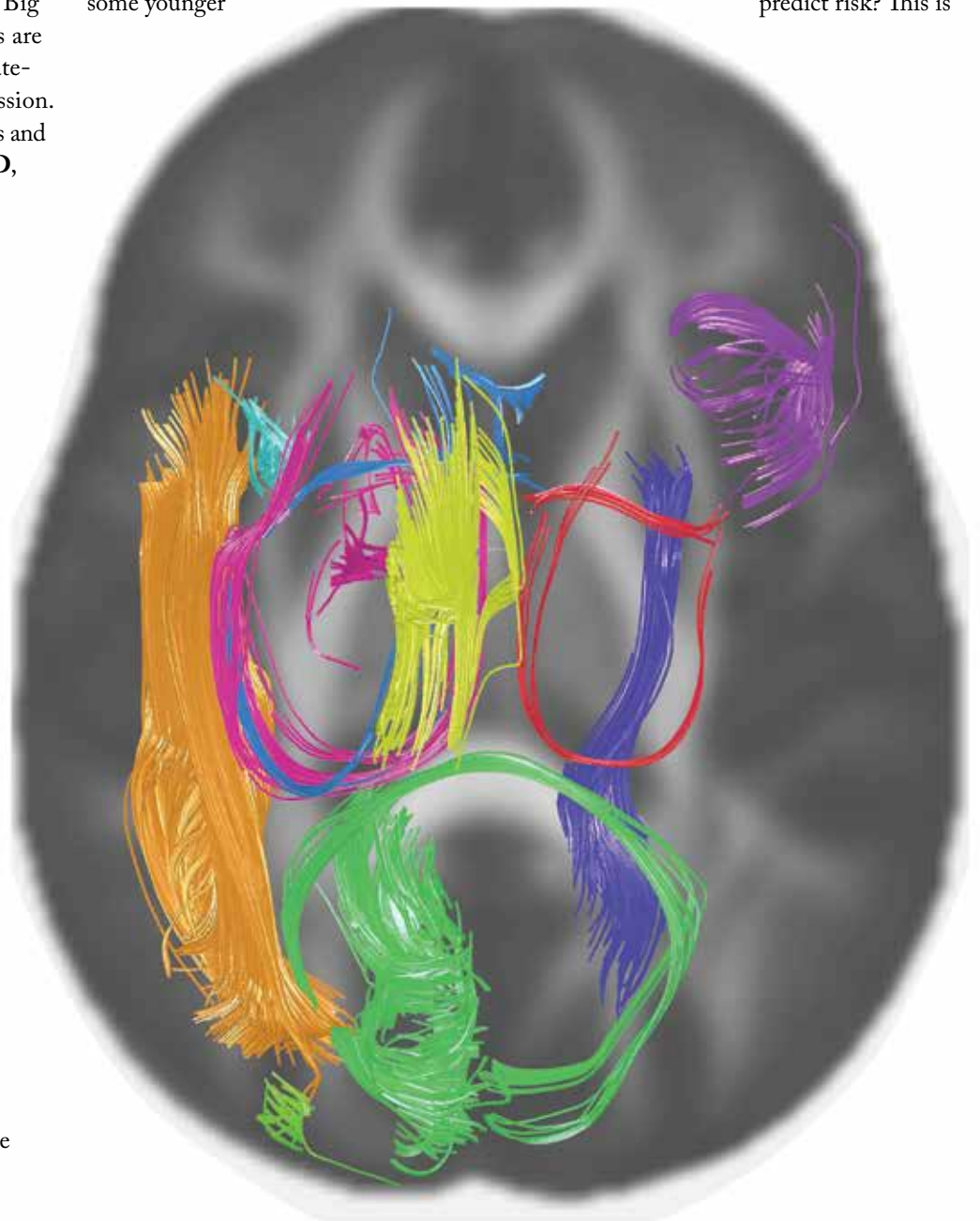
Armed with predictive models, doctors could intervene sooner to improve patient outcomes. Statistical models could also predict whether a patient's disease will progress quickly, slowly, or hardly at all. Identifying who falls into which group may be the key to choosing candidates for clinical trials.

BD2K researchers are using their computational toolkits to study everything from neural changes that presage Alzheimer's disease to rates of osteoarthritis progression. Here, we feature a few stories of exceptional disease detective work with the potential to reshape our understanding of when and why diseases strike.

### Detecting Disease Earlier

For many diseases, there are official guidelines regarding screening patients. Take breast cancer, for instance. The American Cancer Society recommends women age 45 and up receive regular mammograms. But some younger

women are at higher risk than women in older age groups. Should a woman in her early 40s with a family history of breast cancer be screened? A doctor would have to consider family history, demographics, and a patient's medical record to come up with an answer. So why not have an algorithm help predict risk? This is



## RELEVANT NIH INSTITUTES:

**NCI, NHLBI, NIAMS, NIA, NINDS, NCATS, and all other disease-focused Institutes**

the goal of **Elizabeth Burnside, MD**, professor of radiology at UW-Madison.

“What we envision is ... a tailored approach depending on a woman’s risk and values,” Burnside says. “That would hopefully result in better outcomes.”

Burnside’s team has access to nearly 70,000 mammograms collected at UW-Madison’s hospital dating back to 2006, as well as genetic data and personal risk factors (e.g., age, family history, etc.)

machine learning approaches, including support vector machines, neural networks, and deep learning, to predict breast cancer risk.

Breast cancer is caused by genetics and environmental factors that affect estrogen levels, ranging from diet and exercise to breastfeeding

will be useful in clinical settings. “If a patient and physician are going to use a model, they generally want to under-

**“If a patient and physician are going to use a model, they generally want to understand how it’s working,” Burnside says.**

drawn from those patients’ electronic health records (EHRs). In collaboration with CPCP investigators, Burnside’s group combines these data and uses various

history. Burnside believes it is essential to incorporate both nature and nurture into effective, user-friendly models that

stand how it’s working,” she says.

Burnside especially wants increased screening for women at risk of developing aggressive forms of breast cancer.

These include tumors that cannot be treated by hormone therapy, as well as those that break off and spread throughout the body, a process known as metastasis. Her group is analyzing genetic and imaging data to determine what groups of women are at risk for aggressive breast cancer so that they can be screened more intensively.

“What we’re trying to do is to intervene at the right time in the right patient to decrease the chance of poor outcomes,” says Burnside.

Other groups are also capitalizing on the power of imaging to detect subtle phenotypes. A CPCP team led by **Vikas Singh, PhD**, professor of biostatistics and medical informatics at UW-Madison, is

*Diffusion tensor imaging (DTI) is a magnetic resonance imaging technique that can map the bundles of nerve fibers that make up the brain’s white matter. Vikas Singh’s group at UW-Madison applied statistical methods to DTI data to identify individuals at risk of Alzheimer’s years before they show symptoms. Here we see top and side views showing the regions of brain connectivity associated with preclinical Alzheimer’s disease. Image courtesy of Seong Jae Hwang, Singh lab.*

developing statistical algorithms that use diffusion MRI data to map degeneration in brain connections in patients at risk of developing Alzheimer's.

While memory loss and confusion are hallmark Alzheimer's symptoms, they are preceded by a decades-long preclinical phase of the disease. Singh and his collaborators wanted to understand how the brain's intricate web of neural connections change during this early phase of the disease.

"Once you are able to identify or

Both Burnside and Singh have focused their analyses on specific diseases, with the goal of applying their methods to other conditions. **David Page, PhD**, professor of biostatistics and medical informatics at UW-Madison, takes a distinctly broader approach.

"We have lots of EHR data. How well can we predict *every* diagnosis that a patient is going to get?" Page says.

Using 40 years of de-identified EHR data from 1.5 million patients at the Marshfield Clinic in north-central

model accuracy, his 'pan-diagnostic' approach can become a widely used tool that supports both providers and patients.

"We'd like to explore whether some of these models are good enough ... to translate them into use in the clinic—and test whether that has a positive impact," Page says.

### Forecasting Disease Progression and Complications

Once patients learn that they have a disease, they want to know how it will

---

**"Once you are able to identify or predict this future disease course, then you can identify which subset of individuals are most likely to be helped by a new treatment," says Singh.**

---

predict this future disease course, then you can identify which subset of individuals are most likely to be helped by a new treatment," says Singh.

Singh's team, including graduate students **Won Hwa Kim** and **Seong Jae Hwang** and research scientist **Nagesh Adluru, PhD**, analyzed MRI data collected with diffusion tensor imaging, which uses the diffusion of water molecules to reveal tissue architecture. The data, collected at the Wisconsin Alzheimer's Disease Research Center in studies led by UW-Madison professors **Sterling Johnson, PhD**, and **Barbara Bendlin, PhD**, revealed a variety of neural connections that differed in strength between cognitively normal adults with and without a first-degree relative with Alzheimer's. Previous studies have shown that individuals with a family history of Alzheimer's are more likely to develop the disease.

Further research on these connections and the brain regions they encompass may shed light on how Alzheimer's progresses. Singh and his collaborators are now investigating how structural connectivity changes correlate with known protein biomarkers of Alzheimer's, such as beta-amyloid and tau.

Wisconsin, Page's team built predictive models for nearly 4,000 diseases. Their strategy used random forests, a classification algorithm that uses decision trees to guide predictions, and required HT-condor, a high-throughput computing environment that could handle the deluge of data.

---

**Using 40 years of de-identified EHR data from 1.5 million patients at the Marshfield Clinic in north-central Wisconsin, Page's team built predictive models for nearly 4,000 diseases.**

---

The researchers predicted disease from one month to 20 years in advance. All predictions were better than random chance, though, as expected, earlier predictions were less accurate. Page believes that, with further research and improved

progress. Will their symptoms steadily worsen, plateau, or alternate between active and inactive periods? These questions also matter to doctors as they decide the best course of treatment.

At the **Mobilize Center** at Stanford, **Eni Halilaj, PhD**, postdoctoral fellow in the lab of **Scott Delp, PhD**, is studying the progression of osteoarthritis. Halilaj and her Stanford collaborators are analyzing X-rays from the Osteoarthritis Initiative, a multi-center study of knee osteoarthritis progression. Osteoarthritis wears away cartilage over time, which appears on an X-ray as a narrowing of the distance between bones.

Halilaj is building a model that combines information from an initial X-ray with dietary habits, medical histories, joint exam and performance measures, and baseline symptoms to identify slow versus fast progressors—a potentially confounding factor in clinical trials.

"The goal is to predict the kind of progressor that someone will be so that we can balance treatment and control groups in...clinical trials," says Halilaj.

The same statistical tools that predict disease progression can be adapted to other adverse clinical

## MINING FOR PAIN

BY JONATHAN WOSEN

Over a million patients get joint replacements each year in the United States, often due to osteoarthritis, a leading cause of disability. Approximately five percent of replacements fail, according to the American Academy of Orthopedic Surgeons. And postoperative pain can be an indicator that a device is failing.

“We are interested in asking the question, ‘Can we mine electronic health records for device surveillance?’” says **Alison Callahan, PhD**, research scientist in **Nigam Shah’s** biomedical informatics lab at Stanford University. Specifically, she wants to determine whether tracking postoperative pain can provide insight into the effectiveness of specific implant models.

There’s just one problem: Most mentions of pain aren’t neatly coded in a patient’s electronic health record (EHR). Instead, Callahan must dive into the deep, murky waters of unstructured data. Physician notes are a treasure trove of information but are riddled with typos, different ways of referring to pain, negative statements (“the patient did *not* experience pain”), and hypotheticals (“*if* the patient has pain”). It’s a job that, in and of itself, can be quite painful.

To help her mine EHRs from Stanford Hospital and Clinics for, as she puts it, “the type of pain someone’s having and where it hurts,” Callahan needed

labelled training data. Using experts to manually label data would be expensive and time-consuming, so she turned to Snorkel, a tool developed by Mobilize Center researchers in **Christopher Ré’s** lab. Snorkel uses a set of rules, or labeling functions, to create large sets of labelled training data. In Callahan’s case, these rules include whether a clinical note contains pain-related terms and information about sentence structure to ensure a true mention of pain. In collaboration with several Ré lab members (postdoctoral fellow **Jason Fries, PhD**, graduate student **Alex Ratner** and postdoc

**Stephen Bach, PhD**), Callahan used Snorkel to extract mentions of pain and pain location from the notes of roughly 5,000 hip implant patients. She then tested the extraction accuracy with a small subset of data that was manually labeled with the aid of a physician.

---

**“We are interested in asking the question, ‘Can we mine electronic health records for device surveillance?’” says Callahan.**

---

Callahan has presented her work at the 2016 Stanford Data Science Initiative retreat, and her initial extraction results look promising. She now plans to scale up to include larger data sets. Because Snorkel is a general system, Callahan says, it can be used for other research questions as well. “There are other types of experiences which a patient might report which would get captured in a clinical report, [such as] activities of daily living,” Callahan says. As a result, she says, Snorkel has broad applicability for mining unstructured data without the burden of manually labelling large sets of training data.

events, such as postoperative complications. Complications such as infection, heart attack and stroke are major concerns, and studies show that two of every five patients who experience a complication will have more than one. Mark Craven wants to help hospitals understand and predict chains of postoperative complications, which he likens to a snowball effect.

Craven's team utilized a national database of postoperative outcomes known as the American College of Surgeons National Surgical Quality Improvement Program. The researchers considered over 20 different postoperative complications, including infection, heart

---

**Mark Craven wants to help hospitals understand and predict chains of postoperative complications, which he likens to a snowball effect.**

---

failure, and extended use of a ventilator. They used Markov chains, which model changes between states, to predict the complications that occurred each day over a 30-day period post-surgery.

The models were particularly accurate for major complications such as death, heart attack, and kidney failure. Going forward, Craven plans to incorporate additional clinical information from the dataset to make earlier and broader predictions about patient outcomes. "At the time of surgery, how much risk do I think this patient has for having any complications, specific complications, or multiple complications?" Craven says.

His lab has already developed an accurate predictive model for post-hospitalization blood clots using

information from EHRs. Craven plans to test this model in the clinic through a shadow trial—a process of predicting and measuring outcomes without intervening. Predictions will be made about the risk of clots in specific patients as they are monitored over time. If the predictions hold true, doctors may one day use Craven's model to determine who should be given a blood thinner to prevent clotting after hospitalization.

### To Causality and Beyond

Ultimately, predictions for disease progression, outcomes and complications will be more accurate when scientists and doctors understand *why* these events happen. Understanding causation would help researchers design specific therapies that target factors directly involved in disease. **Panayiotis (Takis) Benos, PhD**, professor and vice chair of computational and systems biology at the University of Pittsburgh School of Medicine and project leader for the **BD2K Center for Causal Discovery (CCD)**, wants to develop a causal understanding of chronic lung disease to guide treatment design.

Benos and collaborators are analyzing gene expression and other molecular data together with clinical and histology data from the tissues of patients with idiopathic pulmonary fibrosis. His team uses probabilistic mixed graphical models (MGMs) to combine different data types into a network that reveals direct, causal connections between variables. Using data from the Lung Genomics Research Consortium and new data generated by Benos' team, the researchers have also built MGMs for chronic obstructive pulmonary disease. These models provide insight into how these chronic lung diseases progress and which factors affect the long-term decline of lung function. Knowing these factors, scientists can predict which patients are likely to worsen over the next two to five years, Benos says.

One inherent challenge with this approach is dealing with variables that *aren't* measured. An MGM may

show that a certain gene or measurement is directly associated with a disease, but there could be another untested variable that is in closer association. To bolster his models of lung disease, Benos plans to include larger patient cohorts and incorporate additional variables, including CT scans, biomarkers, and patient symptom questionnaires. In addition, his group is developing algorithms to detect when two variables are controlled by an unmeasured lurking variable. This can help scientists and clinicians recognize when they need to collect additional data. Benos believes this graphical approach can reveal causal relationships in other illnesses, and wants to share his team's analytical tools with the scientific community. "We are planning to apply [MGMs] to cancer, influenza and pneumonia. We also plan to have an R package out soon, so people can easily incorporate our method into their own analysis," he says.

### BD2K Synergies

The BD2K Centers' contribution to the prediction of disease and its progression is still expanding. To provide a fuller picture of changes in a patient's health between doctor's visits, Page would like to supplement EHRs with data from wearable devices that track blood pressure, heart rate, and body temperature. Research out of the **Mobilize Center** and **MD2K (Mobile Sensor Data-to-Knowledge)** could potentially help with that (see "Mobile Health: BD2K Centers Harness Sensor Data," page 10).

In addition, predictive algorithms will be more accurate when built using larger data sets from patients at multiple research centers, which raises the question of how to efficiently share data across centers while also protecting patient privacy. Work out of several BD2K Centers will surely make that a lot easier as well (see "The FAIR Data-Sharing Movement: BD2K Centers Make Headway," page 33).

BD2K has fostered great interactions, Page says. "There's a natural synergy. There's a lot of teamwork." □




# TEXT MINING:

How the BD2K Centers are Making Knowledge Accessible



BY KRISTIN SAINANI, PhD

**S**ince the 1960s, biologists have manually curated data on 6,000 hereditary diseases for the OMIM database (Online Mendelian Inheritance in Man). The database is vital to doctors, who use it for differentially diagnosing genetic conditions; it's much faster and more accurate than asking Dr. Google. But human curators have only been adding about 50 records per month for years, lagging far behind the explosion of information on



gene-disease associations and gene variants currently available in the literature.

What if, instead, computers could curate the literature automatically? What if computers could also scan the millions of papers in PubMed and automatically discover biological networks or predict new uses for existing drugs? These are the many promises of text mining; and some are beginning to come true.

For example, a program called DeepDive, developed by **Christopher Ré, PhD**, assistant professor of computer science at Stanford, and data science lead at the **Mobilize Center**, can now quickly and accurately extract data from the text, figures, and tables of scientific papers. When applied to the paleontology literature as a test case, it extracted 100 times more facts from 10 times more papers than human curators, with an accuracy as good or better than that of

humans. Unlike human curators, DeepDive doesn't get bored or tired, and it can re-read the entire literature any-time to grab new facts of interest.

In biomedicine, the need for high performance text-mining systems like DeepDive has never been more pressing. Most of the collective knowledge of biomedicine is trapped within published papers or buried within the medical notes and images found in electronic health records (EHRs). If researchers could teach computers to make sense

**KnowEnG Center** director **Jiawei Han, PhD**, professor of computer science at the University of Illinois, Urbana-Champaign. "A number of groups are moving text mining along to make it real time and high resolution." Specifically, the community has seen advances in two key text-mining tasks: recognizing entities (e.g., genes and drugs), and extracting relationships between entities (e.g., interactions between genes and drugs). Early systems relied on simple approaches such as matching words to dictionaries; making up simple rules (e.g., the word "kinase" denotes a protein); and assuming that two entities that co-occur in the same sentence are related. Later systems improved accuracy by incorporating machine-learning algorithms.

Now, BD2K researchers at the Mobilize, KnowEnG, and **bioCADDIE** Centers are taking text mining to the next level by leveraging recent advancements in machine learning, such as deep learning and active learning. They are also finding ways to address machine learning's biggest bottleneck: the need for large amounts of hand-annotated data to train the systems. Finally, with tools such as DeepDive, they are putting cutting-edge methods in the hands of users. "It will be exciting to explore how some of these BD2K tools can be combined to form a nice, practical text-mining pipeline," says **Jason Fries, PhD**, a postdoctoral fellow in Stanford's Mobilize Center.

"It will be exciting to explore how some of these BD2K tools can be combined to form a nice, practical text-mining pipeline," says Jason Fries, PhD, a postdoctoral fellow in Stanford's Mobilize Center.

of natural language and pictures, they could unlock this untapped knowledge. But teaching a computer to read is hard; and teaching a computer to read biomedical jargon is even harder. Scientists often write in complicated, convoluted prose; and doctors write in shorthand recognizable only to others in their specialty. Plus, many biological entities have ambiguous names; for example, there are genes named "cheap date", "onion ring", and "pray for elves."

Fortunately, there has been significant progress in biomedical text-mining in the past decade. "There have been a lot of new techniques discovered," says

Even BD2K Centers that are not focusing on text-mining methods per se are getting into the text-mining game by using existing text-mining tools for novel applications, such as cleaning up the metadata in data repositories, an effort that's happening at the **Center for Predictive Computational Phenotyping (CPCP)**.

"What's exciting is that we're moving from just demonstrating that these methods can extract information with reasonably good accuracy to now figuring out ways that this information can be used," says **Mark Craven, PhD**,

director of the CPCP and professor of biostatistics, biomedical informatics, and computer science at the University of Wisconsin-Madison.

## Mobilize: Deep Learning for Text Mining

To achieve state-of-the-art performance in text and image mining, researchers at the Mobilize Center are turning to deep learning. Deep learning models are larger and more complex than traditional machine-learning models, and are driving revolutions in computer vision and speech recognition; for example, they power Apple's digital assistant Siri. "A lot of development has gone into deep learning in the last few years. The community has achieved state-of-the-art and lowered the bar to use," says **Alex Ratner**, a doctoral student in Ré's laboratory at the Mobilize Center.

But there's a catch: Deep learning models need massive amounts of annotated training data from which to learn. "You can label a couple hundred examples and can get a simple model to work, but you can't get one of these deep models to work," Ratner says. "Intuitively it makes sense that a much more complex model—one that has tens of millions of parameters—would need commensurately more data."

It might take weeks or even months for a team of graduate students working around the clock to generate enough training data for one text-mining task. Ré's lab is getting around this problem by having computers label the data. The computer-generated training data are imperfect, but, surprisingly: "You can get really good performance even if you have lower quality labels," Ratner says. Their tools—DeepDive (<http://deepdive.stanford.edu/>) and Snorkel (<http://snorkel.stanford.edu/>)—actually outperform tools that require hand-labeled training data (so-called "supervised" models).

DeepDive automatically annotates training examples with the help of existing knowledge bases, a trick known as "distant" supervision. For example, if an existing database tells us that p53 down-regulates CHK1, then DeepDive would label the sentence: "It was therefore of interest to determine whether p53 affects CHK1," as a positive example of a gene-gene interaction. "You 'lightly' label everything," explains **Emily Mallory**, a doctoral student in the lab of **Russ Altman, MD, PhD**, professor of

### RELEVANT NIH INSTITUTES:

NHGRI, NIBIB, NLM, and all  
disease-focused institutes  
including NCI, NHLBI, NIDDK,  
NINDS, NIAID, and NIAMS

bioengineering, genetics, medicine, and biomedical data science at Stanford. While the labels can be wrong, DeepDive compensates for these inaccuracies with the sheer volume of examples.

DeepDive has been used in applications as far-flung as automatically curating the paleobiology literature to catching sex traffickers by text mining internet ads. In a 2016 paper in *Bioinformatics*, Mallory used DeepDive to automatically extract gene-gene interactions from more than 100,000 full-text articles from *PLoS One*, *PLoS Biology*, and *PLoS Genetics* (see sidebar: Mining for Gene-Gene Interactions, page 26).

It took Mallory a few months to perfect her gene-gene relation extractor because DeepDive requires multiple rounds of iteration and refinement to optimize performance. DeepDive also requires considerable programming knowhow—beyond the skills of a typical biologist. So Ratner and others on Ré's team have developed Snorkel, a successor to DeepDive that is more streamlined, more user-friendly, and achieves better performance.

Snorkel uses even weaker supervision than DeepDive. "Weak supervision is where you say 'I want to use even noisier input streams,'" Fries explains. Rather than relying on just a single knowledge base, you can throw in anything that might contain even a very noisy signal—hundreds of weakly related knowledge bases; training data labeled by lay annotators; or simple, error-prone rules, such as "anytime two chemicals occur in the same sentence label this as a causal relationship"—and Snorkel is able to learn something

Rather than relying  
on just a single  
knowledge base,  
you can throw in  
anything that might  
contain even a very  
noisy signal ... and  
Snorkel is able to  
learn something.

about that signal. Snorkel users input labeling functions via a simple interface that requires only basic programming skills. A typical novice user can write 30 to 40 such labeling functions in hours to days.

The key is that the labeling functions will assign multiple—often conflicting—labels to the same bit of text. Snorkel automatically looks at the patterns of agreement and disagreement to learn which labeling functions are better than others. Labeling functions that mostly agree with other labeling functions are assumed reliable and given the most weight;

labeling functions that mostly run counter to the consensus are considered unreliable and given the least weight. The computer then tallies the votes of the “good” and “bad” labeling functions for a given extraction and assigns it a probability—e.g., there is an 85 percent probability that this sentence contains a gene-gene interaction. By assigning probabilistic rather than yes/no labels, “you’re actually formally acknowledging and modeling the fact that this is weak and inaccurate supervision, not the ground truth,” Ratner says. These training data are then fed to deep-learning

## Mining for Gene-Gene Interactions

In 2016, graduate student Emily Mallory used DeepDive to extract gene-gene interactions from more than 100,000 full-text articles from *PLoS One*, *PLoS Biology*, and *PLoS Genetics*. Mallory first extracted about 1.7 million sentences containing mentions of at least two genes. She labeled sentences as positive for a gene-gene interaction if the gene pair could be found in the BioGRID or ChEA databases and negative if it could be found in the Negatome database (which documents genes and proteins unlikely to interact). This generated a training set with more than 100,000 imperfectly labeled sentences.

Using these training data, DeepDive learned 724 sentence features useful for classification—for example, the presence of the verb “bind” between

two genes. When this model was applied to the 1.6 million unlabeled sentences, it identified 3,356 unique gene pairs where the probability of a true interaction was greater than 90 percent. (To account for uncertainties, including in recognizing gene mentions, DeepDive returns the probability of a true gene-gene interaction rather than a yes/no answer.)

In evaluation against a database of curated protein interactions and manual curation, the system achieved an F1 score of 59 percent, which is on par with state-of-the-art relation extractors that use human-labeled training data. Mallory is now planning to apply the framework to mine gene-gene, gene-disease, and other relations from 500,000 full-text articles available in PubMed Central.

algorithms that can handle uncertainty in the training labels. These algorithms devise a classification model that can be applied to new data for entity tagging or relation extraction.

Snorkel has shown impressive performance. State-of-the-art chemical entity taggers that rely on human-labeled data have achieved an accuracy of 88 percent, as quantified by the F1 score (a common accuracy metric in text mining). On the same task, Fries' team got an F1 score of 87 percent with Snorkel when all they fed it was a dictionary. "We did it completely automatically—we just gave it a dictionary. So, it's completely for free," Fries says. Adding some simple labeling functions improved performance. For a harder task—extracting causal chemical-disease relationships from PubMed abstracts—the top team in the 2015 BioCreAtIvE competition achieved an F1 score of 57 percent using 1000 human-labeled PubMed abstracts. Ratner's team built a Snorkel-based extractor that bested this mark without using any human-labeled data. They used 33 labeling functions applied to hundreds of thousands of unlabeled PubMed abstracts. "We can pour in unlabeled data, and we actually get scaling," Ratner says.

Mallory is now collaborating with the FDA to build a Snorkel-based tool for extracting gut microbiome relationships, such as drug-microbiome and chemical-microbiome interactions, from the biomedical literature. Fries and Ratner are working on Snorkel applications that extract information from the clinical notes of electronic health records. For example, Fries is collaborating with Stanford post-doctoral fellow **Allison Callahan, PhD**, to extract mentions of pain and other symptoms from 500,000 clinical notes for 3500 hip and knee replacement patients. When combined with structured data from the electronic health records, unstructured data from clinical notes may help doctors predict which patients' implants will fail, as well as generate early warnings when specific devices are causing problems (see page 21 for sidebar story about Callahan's work). Snorkel is also being used outside of biomedicine—for example, researchers at the Hoover Institution are using Snorkel to extract data from military combat notes to try to determine what factors cause militants to join or leave insurgencies.

Re's lab is also building tools on top of Snorkel to extract data from images, figures, and tables. Ratner is collaborating with radiologists who study bone tumors to develop a Snorkel-based tool that can accurately classify images of bone lesions as cancerous or not. For images, users write labeling

functions that consider visual features, such as edges, as well as text in titles and captions. Snorkel-based tools that read tables in the biomedical literature are also in development. These tools can help augment manual data curation efforts, such as for the GWAS Catalog (an online catalog of published genome-wide association studies). For example, a computer could extract results from every supplemental GWAS table in the published literature.

## KnowEnG: From Phrases to Relations

Researchers at the KnowEnG Center are also exploiting weak and distant supervision to make state-of-the-art text-mining tools that require minimal labeling from domain experts. KnowEnG's director, Jiawei Han, has developed a suite of text-mining tools that work on everything from tweets, to the New York Times, to the scientific literature. In the past few years, Han's lab has been focusing on applications in biomedicine.

"Jiawei is a mainstream text-mining person. For him to enter bio-text mining is very exciting," says KnowEnG co-director **Saurabh Sinha, PhD**, professor of computer science at the University of Illinois, Urbana-Champaign. "The underlying tools that his lab has developed are making new functionalities that I care about happen."

Han's team first built tools to mine phrases out of text. Text-mining tools need to recognize that certain words go together—such as "congenital heart

Text-mining tools need to recognize that certain words go together—such as "congenital heart disease" or "Obama administration."

disease" or "Obama administration." "Extraction of phrases is critical towards information extraction because many concepts, entities, and relations are manifested in phrases," Han says. The tools are portable across domains and languages, and also

require minimal or no hand labeling. “We worked out a very powerful method using either no training at all or weak training or distant training.”

Han’s team first built ToPMine in 2014, which is an unsupervised method and requires no training data. ToPMine identifies salient phrases using statistical clues, such as how frequently a given string of words appears in the corpus (popularity), how often the words appear together versus apart (concordance), and how often they appear in one docu-

tools. ClusType first uses distant supervision to tag some entities in the corpus. Then it leverages the context clues around the labeled entities to tag additional entities. For example, based on Wikipedia, ClusType may tag ice cream as a food in: “The waiter served ice cream.” When ClusType later comes across an unlabeled phrase in a similar context—for example, “The waiter served pav bhaji”—it is able to predict that pav bhaji is a food. Newly labeled entities give new context clues, and the whole cycle

ClusType may tag ice cream as a food in:

“The waiter served ice cream.” When ClusType later comes across an unlabeled phrase in a similar context—for example, “The waiter served pav bhaji”—it is able to predict that pav bhaji is a food.

ment but not another (distinctiveness). For example, “congenital heart disease” is distinctive because it crops up frequently in some documents but rarely in others, whereas “important problem” is ubiquitous and thus not what Han calls a “quality phrase.”

Han’s team found they could improve performance by adding weak supervision.

Their tool SegPhrase incorporates a machine-learning model that can be trained with a tiny amount of labeled data—just 300 labeled phrases for a 1 gigabyte corpus. The model generates better-quality phrases than completely unsupervised methods such as ToPMine.

Han’s team recently built AutoPhrase, which uses distant supervision to obviate the need for hand-labeled data. Users provide AutoPhrase with a dictionary or knowledge base that can be used to label enough phrases in the corpus to train the machine-learning model. “We like this distantly supervised method because you can get high-quality results without experts,” Han says. “It’s also powerful because it works on many languages. It could also recognize Chinese phrases if we gave it a Chinese Wikipedia, for example.”

Han’s team has also developed an entity tagger called ClusType, which builds on their phrase-mining

repeats until the corpus is adequately labeled. When applied to news stories, Yelp reviews, and tweets, ClusType yielded an average 37 percent improvement over the next best method for tagging entities. Han’s team has extended this to CoType, which works in a similar manner but types both entities and relationships between entities simultaneously.

Han’s suite of text-mining tools are publicly available (at <https://github.com/KnowEnG>) and are being built into the KnowEnG knowledge engine. The KnowEnG Center is also partnering with Heart BD2K to use the tools to solve specific biomedical problems (see sidebar: Ranking Proteins in Heart Disease, opposite).

Han’s team is also working on an exciting new search tool that embeds entity recognition into the search. If you search in PubMed or Google Scholar, these search engines treat genes, proteins, metabolites, and drugs like any other words. But what if the search engine could recognize genes, proteins, metabolites, and drugs as biological entities? “That’s a type of query/response interface to the literature that supports a much richer space of queries,” Sinha says.

Working together with existing biomedical knowledge bases, Han’s tools can tag entities in queries and papers. “His tools can recognize that there are different types of terms in there. They have built in the prior knowledge of what are genes, what are proteins, what are drugs, and so on,” Sinha explains. Now, Han’s team



# Ranking Proteins in Heart Disease

The **Heart BD2K**'s director, **Peipei Ping, PhD**, asked Jiawei Han's team at KnowEnG to help them use the biomedical literature to comparatively rank 250 proteins known to be involved in heart disease. Ping is professor of physiology, medicine/cardiology, and bioinformatics at the University of California, Los Angeles. "The problem is there are millions of papers in cardiology. Nobody can read one million papers in a lifetime. But a computer can," Han says. "What if your computer could read those articles to give you a comparative summary?" Han's and Ping's labs collaborated to build a pipeline called Context-Aware Semantic Online Analytical Processing (caseOLAP), which incorporates SegPhrase.

Ping's team wanted to know which of the 250 proteins were most relevant for each of the six major types of heart disease—cerebrovascular accidents, cardiomyopathies, ischemic heart diseases, arrhythmias, valve disease, and congenital heart disease. They used phrase mining to group abstracts by disease and to discover the predominant proteins for each disease. CaseOLAP calculated a text-mining score for each disease-protein pair based on the quality of the mined phrases, how frequently a given protein appeared in the abstracts of a given

disease, and how distinct a given protein was for one disease versus the other five.

"The thing that I found amazing was just how much information could be processed. This is something that a human being just cannot do," says **David Liem, MD, PhD**, a scientist at UCLA and the project's clinical study coordinator. When they examined the top-ranking proteins, they got some unexpected insights. "Some of the findings were no surprise. For example, we found a lot of inflammatory proteins and proteins involved in hemodynamic regulation," Liem says. "But what was a surprise to us was we found a lot of proteins that are involved in neurodegenerative diseases." This unforeseen link between heart disease and neurodegenerative disease has been confirmed in other recent studies, Liem notes.

The discoveries could help doctors predict new drug targets for heart disease, Ping says. Han and Ping plan to expand the project to explore the role of 8,000 additional proteins and also to rank protein-protein interactions for each type of heart disease. The tool could also be applied to electronic medical records to mine clinical notes. "This collaboration opens possibilities for many other projects," Ping says.

is working on exactly how to use this information to give the most reasonable ranking of papers. “The problem isn’t solved yet, but we have an army of really smart graduate students working on this,” Sinha says.

## bioCADDIE: Customized Pipelines for Text Mining

The bioCADDIE Center is developing dataMED, a search engine for publicly available datasets that does for data what PubMed does for papers. Similar to Han’s search tool, dataMED embeds

“[T]here are millions of papers in cardiology. Nobody can read one million papers in a lifetime. But a computer can,” Han says.

entity recognition into the search. So, it’s not surprising that text-mining expert **Hua Xu, PhD**, was tapped to lead development. Xu is a professor in the School of Biomedical Informatics at the University of Texas Health Science Center at Houston.

Xu’s lab works on text-mining methods, software, and applications for clinical data. “I view it like a circle. You have a new proven method for NLP [natural language processing]; you implement that into software; and then you use the software to extract information for clinical studies,” Xu says. “Then the clinical study actually suggests needs for new technology, and this feeds back to the methods development.”

Like scientists at Mobilize and KnowEnG, Xu wants to reduce demands for costly annotated training data. Rather than turning to weak and distant supervision, however, Xu’s lab is taking a different tack—an approach called interactive machine learning.

Interactive machine learning loops humans into the learning process.

The idea is that if a person injects critical insights at the right time, this can improve efficiency and performance. Normally, human experts label random stretches of training data. But in active learning—a type of interactive machine learning—only the most informative examples are selected for labeling, Xu explains. At the beginning of active learning, a human expert annotates a small amount of randomly selected training data, which is fed into a machine-learning algorithm to build a model. The computer then attempts to classify the unlabeled data using the model—and passes back the ones for which it has the most trouble. The human labels these, and passes them back to the computer. This cycle repeats until the model achieves sufficient accuracy. This approach has the potential to significantly reduce training data for the same performance, Xu’s team has shown.

Xu’s lab has also built several machine learning-based tools for entity tagging and relation extraction that have taken first or second place in major text-mining challenges including the i2b2 NLP Challenge, SemEval (Semantic Evaluation) and BioCreAtIve (Critical Assessment of Information Extraction in Biology). Xu’s team has made these tools available as part of dataMed.

Sometimes, text-mining tools need to be tuned to local data. For example, different hospitals and even different specialties within the same hospital may have idiosyncrasies in how they talk about diseases and traits. “For an institution that doesn’t have a strong NLP team, it could be very challenging,” Xu says. To address this issue, Xu’s team built a user-friendly clinical text-mining system called CLAMP (Clinical Language Annotation, Modeling, and Processing Toolkit, <http://clamp.uth>).





edu/). With CLAMP, users drag and drop ready-made components—such as part-of-speech taggers, dictionaries, and machine-learning modules—to create a customized pipeline. “They can modify each component, including directly annotating local data and clicking a button to train a machine-learning module,” Xu says. “With this interface, users who don’t have much NLP experience can build high-performance NLP pipelines for their own tasks,” Xu says. The tool is freely available to academics, and has been downloaded by about 50 institutions.

For example, a researcher can enter “HIV replication” in gene-search mode and get a ranked list of genes relevant to HIV replication. The system searches millions of abstracts in PubMed to find those relevant to the user’s query, and then applies a named entity recognizer to identify mentions of genes or small molecules. “The challenge is trying to filter out false positives,” says CPCP director Mark Craven. For example, the common English word “cat” could be mistaken for the abbreviation for the catalase gene, CAT. To help avoid errors, the named entity recognizer considers

With CLAMP, Xu says, “users who don’t have much NLP experience can build high-performance NLP pipelines for their own tasks.”

The same cutting-edge text-mining tools that are baked into CLAMP are also being used in dataMED, bioCADDIE’s dataset search tool (<https://datamed.org>). The dataMED development team first indexed all the datasets in the data repositories, using NLP to mine free-text fields for mentions of genes, diseases, and chemical names. Then the team built a search tool that embeds recognized entities into the search. For example, if a user types in breast cancer and NFKappaB, the tool recognizes these as a disease and gene, respectively, and maps them to their standardized ontology concepts. Then it expands the search to include synonyms of these concepts (e.g., “tumor of breast” for “breast cancer”).

dataMED has already indexed more than 63 data repositories that contain a total of 1.4 million biomedical datasets.

## CPCP: Putting Text-Mining Tools to New Uses

Though text-mining methods are not a focus of CPCP, Center researchers have developed some novel ways to use text mining for different kinds of PubMed searches and to clean up metadata in data repositories. GADGET, developed by Craven, uses standard indexing and text-mining algorithms to search PubMed and return, for a given query, genes and metabolites rather than articles.

properties of the word, such as the presence of italics, as well as the context around it. “We look at lots of pieces of evidence like that to decide ‘do I think this is a gene name, yes or no?’” Craven says. The software then ranks the genes based on how many query-specific abstracts mention the gene and how frequently the gene is mentioned in other abstracts.

Biologists at the University of Wisconsin-Madison, are already using the tool to accelerate their science. For example, one stem cell lab searches for genes that might help them steer stem cells to a given fate. Another lab is using the tool to help figure out the networks of host genes involved in HIV replication. “We found that we can get better network models by pulling in this evidence from the literature as identified by GADGET in addition to the genes that are coming directly from experiments,” Craven says. GADGET is freely available here: <http://gadget.biostat.wisc.edu/>.

Another lead investigator of CPCP, **Colin Dewey, PhD**, associate professor of biostatistics and medical informatics at the University of Wisconsin-Madison, is using text mining to clean up the metadata of the Sequence Read Archive (SRA) data repository.

The SRA stores next-generation sequencing reads for 2.1 million samples from 90,000 worldwide studies. Scientists hope to mine these data for new insights; for example, by studying all available lung cancer RNA-seq data, scientists might be able to pinpoint gene expression patterns that characterize the disease. But combining data across different studies is difficult because the samples aren’t labeled consistently. “The metadata are not standardized or normalized,” Dewey says. “People just make up their own names for the attributes, as well as the

values of those attributes.” Attributes (such as “cell line”) and their values (such as “HeLa cells”) vary widely due to misspellings, synonyms, abbreviations, and the use of natural language descriptions.

So, Dewey’s team devised a novel computational pipeline (<https://github.com/deweylab/metasra-pipeline>) that automatically cleans up the metadata in the SRA.

Off-the-shelf text-mining tools yield an unacceptably high false-positive rate when applied to SRA metadata. “There are a lot of cases when there’s an entity mentioned in the metadata that is not actually describing the sample of interest,” Dewey says. For example, a standard named-entity recognizer will extract the word “breast” from “breast cancer” and infer that this is the anatomical source of the sample. But the sample may have been taken from blood rather than breast tissue.

Dewey’s team built a system that’s similar to a named entity recognizer but “with a bunch of heuristics added to remove the errors introduced by such systems,” he says. The system builds a graph, starting with the attribute-value pair from the original metadata. The attribute and value are each mapped to terms from biomedical ontologies (represented as nodes on the graph). But Dewey’s team then subjects the graph to a series of custom-made reasoning rules and operations. For example, one of these heuristics recognizes that “breast” should not be mapped to an anatomical location if the word “breast” in the metadata is part of a larger phrase (e.g., “breast cancer”) that maps to an ontology term. Another rule tells the system that the abbreviations “F” and “f” indicate a female sample when they are paired with an attribute that maps to “sex.” The system also extracts numerical values—such as age—from metadata. “That’s a novel aspect of

95 percent). The MetaSRA database is available at <http://deweylab.biostat.wisc.edu/metasra>. Dewey’s team plans to expand the database in the future.

## Looking Back; Moving Forward

Nine years ago, *Biomedical Computation Review* (Summer 2008) published an article about text mining. One challenge identified then remains a major bottleneck today—data accessibility. Out of 14 million English-language abstracts in PubMed, only about a million are accessible for full-text mining, Mallory says; and when it comes to electronic health records, privacy issues complicate data access. For text mining to realize its full potential, researchers will have to make headway on this issue.

But there have been a lot of wins over the past nine years, thanks in part to work by the BD2K Centers. Text-mining tools have gotten more powerful and, importantly, more usable by doctors and biologists—as evidenced by user-friendly programs such as Snorkel, CLAMP, and GADGET. Text-mining tools are also being used in more real-world applications than ever before—from curating and scanning the literature to making it easier to find and pool publicly available datasets. It might also be possible to build a pipeline from the BD2K tools described here. For example, Fries says, Han’s unsupervised learning tools are great for potentially discovering new patterns and building domain dictionaries, but they are also very noisy. Snorkel could be used to unify these dictionaries into a more robust extraction system. “The different BD2K tools being developed provide complementary ways to tackle the text mining problem,” he says.

“The different BD2K tools being developed provide complementary ways to tackle the text mining problem,” [Fries] says.

our system that will be helpful for doing aggregate analyses using those numerical values as covariates.”

In initial tests on human samples assayed by RNA-seq experiments on the Illumina platform, the system achieved recall rates as good as standard named-entity recognizers (85 to 90 percent) but better false positive rates (precision of 90 to

Nine years from now, perhaps computer curators will have replaced human curators for the OMIM database. As a result, within moments of a new paper hitting PubMed, new knowledge will be deposited in OMIM automatically—making it possible for doctors to instantly use that knowledge to help patients. □

**RELEVANT NIH INSTITUTES:**

NHGRI, NIAID, NIBIB, NLM, NIEHS, NIGMS, as well as all disease-focused Institutes including NCI, NHLBI, NIDDK, and NINDS

findable THE FAIR reusable  
accessible interoperable

# Data-Sharing Movement:

*BD2K Centers Make Headway*

BY KATHARINE MILLER

Science that isn't **reproducible** isn't science at all. And science that relies on big biomedical datasets will only be reliably **reproducible** if those datasets are **FAIR**—**findable**, **accessible**, **interoperable** and **reusable**.

Achieving the laudable goal of **FAIR** datasets requires a shift in scientific culture. Researchers accustomed to storing their data in silos at individual research institutions must become more mindful about how they handle, describe and store their data. In addition, there must be an infrastructure that makes data sharing possible.

When the National Institutes of Health (NIH) funded the twelve Big Data to Knowledge (BD2K) Centers of Excellence in October of 2014, **Philip Bourne, PhD**, the NIH associate director for data science at the time, understood that an emphasis on the **FAIR** standards within the BD2K Centers would seed this cultural change.

"We view this as a virtuous cycle," Bourne told this

magazine. The Centers would generate **FAIR** data and data-sharing tools that others would use to do the same; and this ongoing cycle would serve to simplify and normalize the process. "Sharing the data and the software across the Centers and to other investigators and beyond is key," he said.

Fast forward two and a half years and the Centers are in full swing, propelling the data sharing revolution forward at every level of research and demonstrating that adherence to the **FAIR** principles is an achievable goal.

*Metadata Entry by Humans:  
Achieving **FAIR**ness Up Front*

Biomedical researchers are generating datasets at unprecedented rates. To describe, store and share these datasets in ways that are **FAIR**, the researchers must create

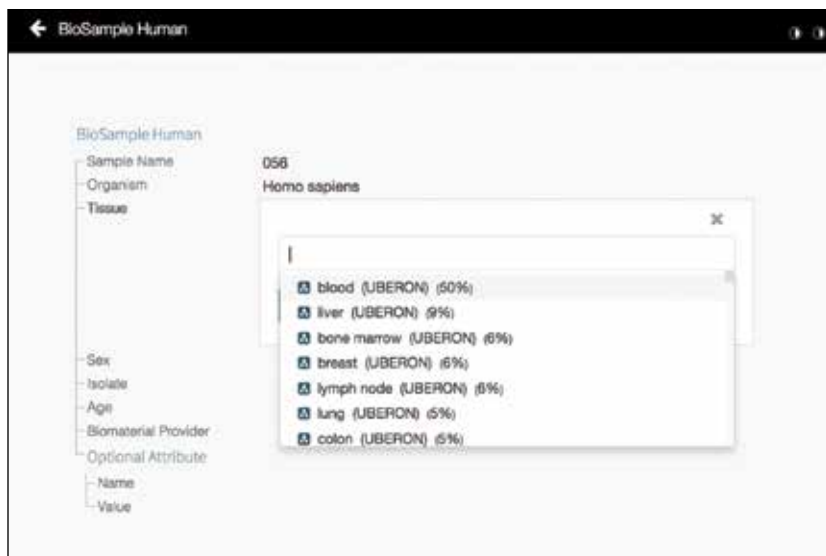
metadata—clear, accurate, computer-readable descriptions of the data. “A lot of metadata are produced by people who are forced to produce them under duress when they have other things they’d rather do,” says **John Graybeal, PhD**, technical program manager for the BD2K Center known as **CEDAR (the Center for Expanded Data Annotation and Retrieval)**. “In the absence of good verification processes and helpful suggestions to such people about what information they need to provide, you get a lot of pretty useless metadata.” And if the metadata are useless, the data itself will never be **FAIR**.

To address that problem, CEDAR has created the CEDAR Workbench, which researchers can use to access libraries of standard templates for defining metadata. The Workbench makes metadata entry easy by suggesting appropriate templates, enabling the use of appropriate terminology from various biomedical ontologies, and suggesting such terms in drop-down menus. In addition to making data **findable** and **accessible**, CEDAR’s Workbench adds a lot of value for **interoperability** and **reproducibility**, Graybeal says, by ensuring that people are using the same terminology in consistent ways.

For example, investigators who use high-throughput B-Cell and T-Cell receptor repertoire sequencing (Rep-Seq) can deposit their data into four different repositories (BioProject, BioSample Sequence Read Archive [SRA] and GenBank) at the National Center for Biotechnology Information (NCBI). CEDAR is working with members of this community to provide a simple and standardized metadata entry process that can be integrated with the data submission process already in place. If these attempts are successful, researchers will be able to submit their data and metadata through the CEDAR Workbench, and it will flow to the appropriate NCBI repositories,

considerably simplifying the submission process.

CEDAR is also collaborating with the Library of Network-Based Cellular Signatures (LINCS) consortium and the **BD2K-LINCS Data Coordination and Integration Center (BD2K-LINCS-DCIC)**. LINCS



*In the CEDAR Workbench, a user selects a metadata template and then fills in the template by selecting values from various dropdown menus. Here the user is selecting a value for the tissue field in the BioSample Human template from candidate tissue types from the UBERON ontology in BioPortal. Courtesy of Mark Musen and CEDAR.*

researchers, who use various methods to disrupt biological pathways and observe the altered phenotypes, have generated huge amounts of data. Their standardized metadata procedure involves several manual steps: They enter information into an Excel spreadsheet, then submit it for manual review by a person who checks it for completion and accuracy and then emails the submitter to request corrections. “Heavily manual processes like this don’t scale well,” Graybeal says. “And Excel spreadsheets are limited in the support they can offer metadata providers.” With a supplemental BD2K grant, CEDAR is helping LINCS researchers develop an integrated workflow for managing metadata in real time, with the system giving users feedback to correct mistakes right away. “From the user’s standpoint and LINCS’ standpoint, that’s a big change,” Graybeal notes. “Curators will be able to review created metadata much faster.” The system is now functioning in a prototype environment and is targeted for production over the summer.

CEDAR is also contemplating how to fix flawed metadata that’s already stashed in data repositories. Looking at the Gene Expression Omnibus (GEO) data repository, for example, researchers on the CEDAR team have noted inconsistent entries for such basic information as age and gender. CEDAR could support workflows to automatically enhance these metadata, or even provide simplified ways for users to correct these issues in a wiki-fied environment. The metadata updates could be forwarded back to the

### Variants of ‘age’ metadata field in Gene Expression Omnibus (GEO) repository

age	age [y]
Age	age [year]
AGE	age [years]
`Age	age in years
age (after birth)	age of patient
age (in years)	Age of patient
age (y)	age of subjects
age (year)	age(years)
age (years)	Age(years)
Age (years)	Age(yrs.)
Age (Years)	Age, year
age (yr)	age, years
age (yr-old)	age, yrs
age (yrs)	age.year
Age (yrs)	age_years

**Metadata entry is difficult and leads to inconsistencies that make data reuse challenging. CEDAR is addressing this problem by creating standardized metadata templates. Information courtesy of Mark Musen, MD, PhD, principal investigator for CEDAR.**

repository, if it had a way to handle these sorts of changes and suggestions. “Ideally, you’d end up with well-reviewed and more accurate documentation,” Graybeal says, which would be a step forward for the **FAIR** principles.

## Finding Accessible Data

Given the huge quantity of biomedical data that has been generated by high-throughput experiments as well as the vast troves of clinical data residing in electronic health records, many researchers hope to address interesting research questions by **finding** and **accessing** existing data rather than generating more. The NIH recognized the potential for **re-use** of existing datasets when, as part of the BD2K program, it funded the **biomedical and healthCare Data Discovery Index Ecosystem (bioCADDIE)** under the leadership of **Lucila Ohno-Machado, MD, PhD**, professor of medicine at the University of California, San Diego.

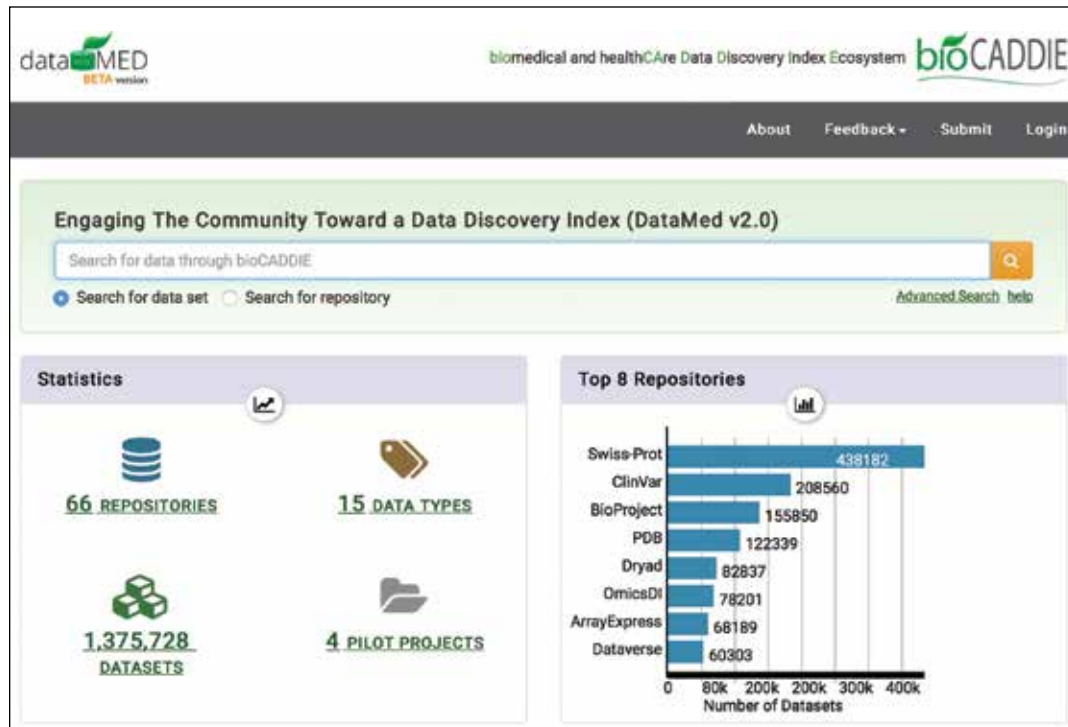
bioCADDIE set out to develop a prototype data discovery index to help people **find** relevant datasets they would otherwise have a hard time **finding**. The result is dataMED, a repository and search engine that does for data what PubMed does for biomedical literature: Rather than having to search individual repositories for relevant data, researchers can search dataMED to **find** what they are looking for. As of March 2017, dataMED had indexed 64 data repositories containing more than 1.3 million datasets, and it is still growing.

Just as scientific journals must meet certain requirements for inclusion in PubMed, repositories included in dataMED must meet certain standards of quality, sustainability and **interoperability**. And their metadata must be capable of being digested into dataMED’s DATS (DAta Tag Suite) metadata system. The core DATS metadata includes information about a dataset’s **accessibility** through an application programming interface (API) as well as whether **access** requires approvals or security clearances.

To test the effectiveness of dataMED, bioCADDIE ran a dataset retrieval challenge competition to see who could develop an algorithm that would identify the best set of data to address a well-defined research question. The top two winning algorithms are now being incorporated into dataMED.

“Achieving **findability** and **accessibility** is just the

beginning of the journey,” Ohno-Machado says. More work will be required to achieve **interoperability** and **reusability**. For now, she says, “it’s critical to **find** the data in the first place.”



Searching for datasets in dataMED is akin to searching for scientific literature in PubMed.

## Making Dataset Recommendations: Findability Goes Deeper

In a separate effort to make datasets more **findable**, the **HeartBD2K Center** created the Omics Discovery Index (OmicsDI). Whereas dataMED indexes a broad array of datasets (including OmicsDI), OmicsDI focuses solely on omics datasets (proteomics, transcriptomics, genomics etc.). It is also searchable at a deeper level than dataMED.

OmicsDI evolved from an even more specifically defined collaboration called the ProteomeXchange, a global network of four proteomics databases that coordinates how they accept data and then centralizes their metadata. The HeartBD2K Center has extended these concepts to multiple omics data types including genomics, transcriptomics, and metabolomics, says **Henning Hermjakob, MSc**, team leader for molecular networks services at the European Bioinformatics Institute (EBI) and co-director of HeartBD2K. “We got off the mark quite fast because we could build on existing experience and infrastructure.”

Like bioCADDIE’s dataMED, OmicsDI can be easily searched to **find** datasets of interest. “But what we offer beyond pure metadata indexing is where it gets interesting,” Hermjakob says. In addition to indexing the metadata, OmicsDI indexes part of the data content. For example, it might index the proteins observed in a proteomics

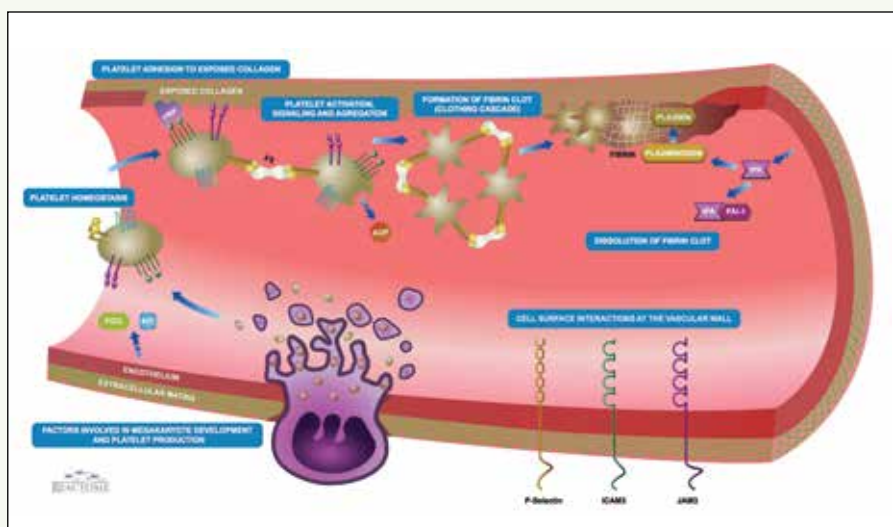
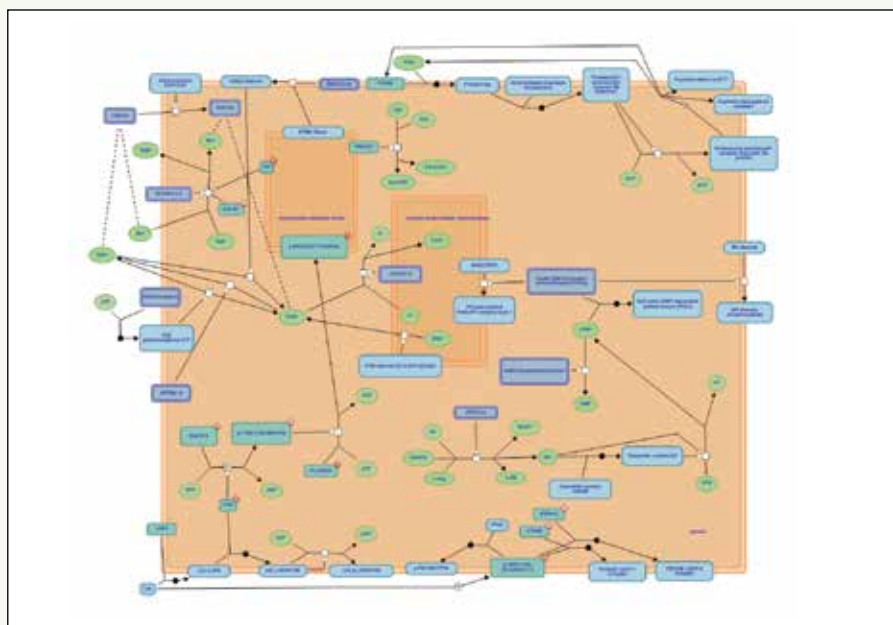
# Extracting Needles from Haystacks: The Reactome

Reactome.org is an open-source curated database of human biological pathways. As of early 2017, it comprised 10,391 human reactions organized into 2,080 pathways involving 10,624 proteins encoded by 10,381 different human genes, and 1,735 small molecules. And under BD2K, the web interface has been re-designed to create a user-friendly tool. Researchers gathering gene expression data about normal and diseased tissue or cells perturbed with a certain drug want to know what the observed expression changes mean. Now they can upload their datasets to Reactome.org and it will show the pathways that are over- or under-represented in their specific gene set. “It’s very helpful to the biologist in reducing large changes in large datasets to something that is much more understandable biologically,” Hermjakob says. It doesn’t necessarily give answers, he says, but it points researchers in the right direction. “It provides a magnet for extracting a needle from a haystack.”

The updates to Reactome make it not only more [accessible](#) but also more [interoperable](#). “We’ve developed computational components allowing Reactome functionality in other websites with very little effort,” Hermjakob says. LINCS-DCIC, for example, provides high-quality new data on systematic perturbations of different cellular systems and Reactome aims to provide

analysis capability for exactly these kinds of data output. “So the two fit together quite nicely,” Hermjakob says. Reactome functionality has

also been incorporated into several non-BD2K projects including the Human Protein Atlas and the Open Targets project.

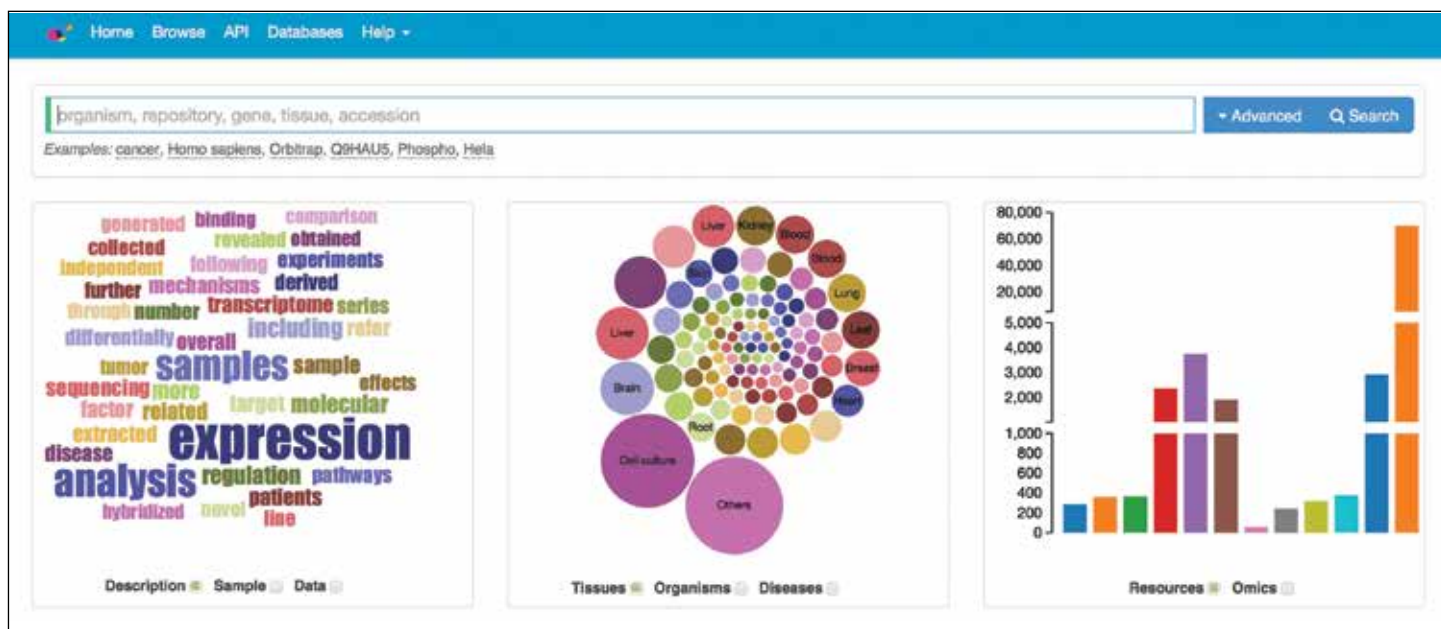


*In the very latest release of Reactome.org, Hermjakob and his colleagues have significantly enhanced the visual presentation of biological pathways. Schematic diagrams (such as the one at top showing platelet homeostasis) have been replaced with textbook style pathway diagrams (bottom, showing homeostasis more generally) drawn by a professional biomedical illustrator. Portions of the diagrams are clickable to dig deeper into the various pathways. “These make Reactome easier to navigate and the diagrams are also released in an editable form that can be used for publications and slides,” Hermjakob says. Courtesy of Reactome.org.*

experiment; or the differentially expressed genes in a transcriptomics experiment. OmicsDI then calculates similarity metrics between all the experiments in a domain. Using this capability, a user can get recommendations for datasets of interest. “The whole functionality is similar to recommendations in Amazon.com: ‘If you’re interested in this dataset

it very easy for someone to do predictions of functions for genes,” Ma’ayan says.

Ma’ayan thinks of the Harmonizome as a prototype that shows what can be done. “The nice thing about the Harmonizome is that it enables search at the data level,” he says. He acknowledges that making it scalable could be



When searching OmicsDI for relevant datasets, the search box offers a dropdown menu of options. When the search is complete, researchers may further refine it by tissue, disease, or organism, and search results can be sorted by relevance—a measure of how closely related the datasets are to the specific query.

you might also be interested in this other one,” Hermjakob says. He’s eager to see if this system leads to more datasets being reused—and OmicsDI is tracking that as well.

## Findability at the Deepest (Data) Level

Avi Ma’ayan, PhD, professor of pharmacological sciences at the Icahn School of Medicine at Mount Sinai and principal investigator of the **BD2K-LINCS Data Coordination and Integration Center (BD2K-LINCS-DCIC)**, decided to take findability to another level. Their creation, the Harmonizome, offers a collection of all the hottest and most exciting databases that everyone is using. “It allows you to find knowledge about genes and proteins that was buried in data silos but now is accessible.”

To create the Harmonizome, Ma’ayan’s team gathered together 66 major online omics resources and processed them into more than 70 million associations between nearly 300,000 attributes and all human and mouse genes and proteins. That processing involved taking either raw data or formatted data from existing databases and mapping it onto common IDs for genes. They also processed the data into simplified formats such as relational tables, making it ready for machine learning. The data are now served online through a user-friendly interface. “It makes

challenging. Still, the Harmonizome has proven popular. Since it became public in 2015, the site has had more than 100,000 unique user visits and 300,000 page views. “We get about 400 users per day now,” Ma’ayan says, with about 40 percent sticking around for a while because they are finding it useful. He’d like to learn more about how others are using the resource. “I’m sure people can think of creative ways to use it that we haven’t thought of,” Ma’ayan says. “That will be the coolest thing.”

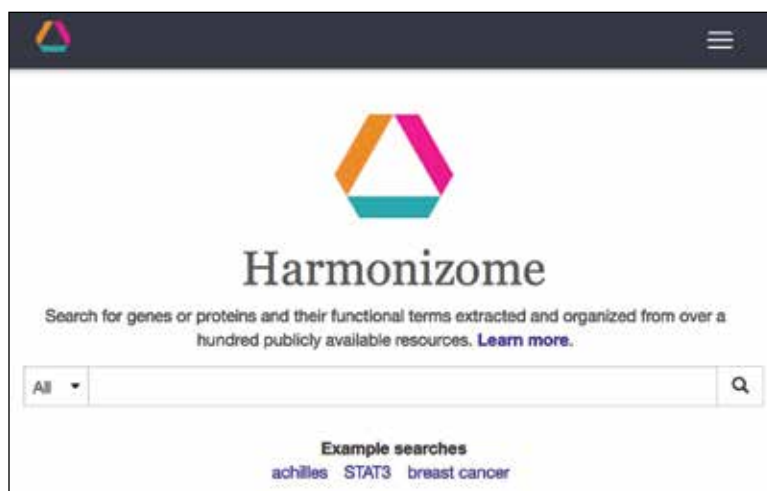
## Accessing Data Where It Lives

For genomics research to achieve its promise of improving human health, vast data resources must be brought to bear. “We absolutely have to share the data so that we can compare cases,” says David Haussler, PhD, principal investigator for the **Center for Big Data in Translational Genomics (BDTG)**, a BD2K Center. But access to genomic datasets is often restricted because of privacy concerns and confidentiality laws established by the countries and institutions where the data reside. This makes it impossible to create a central, unified genomic database. “The only path forward is to create a common API and a common language for containerized workflows so that you can literally ship analysis software off to different countries and medical institutions and let it run on their platforms,” Haussler says.

And that has been one focus of the BDTG Center. They've created an ecosystem of genomics tools and a standard interface for interacting with genomic data—the Global Alliance API. It allows genomics researchers to work together on a global scale, sharing software to achieve **reproducible** results.

As a test case, the Center developed Toil, a portable, open-source workflow, and demonstrated its use in a cloud environment by reprocessing more than 20,000 RNA-seq samples from four major studies. The effort reduced the time and cost of such processing by 30-fold; created a new, publicly available dataset free of batch effects (the statistical problems created by data processed in different ways at different research institutions); and set a precedent for the use of portable software in contemporary cloud workflows

as a path toward allowing research groups to **reuse** data and **reproduce** one another's results.



## Sharing and Integrating Clinical Data

Often, researchers want to study different types of data for the same patient population but the data—much of it privacy protected—are distributed across multiple databases and institutions. In a pilot study called Count Everything, four BD2K Centers of Excellence worked together to create a prototype for integrating various types of data without compromising privacy. Ohno-Machado's Center, bioCADDIE, played the aggregator role using APIs (for genomics, electronic health records, and mobile health data) developed by three other centers: BDTG, **PIC-SURE (Patient-centered Information Commons: Standardized Unification of Research Elements)**, **MD2K (Mobile Sensor Data-to-Knowledge)**. The result: a system that can make simple distributed queries across simulated data from these Centers in a secure and anonymized way. Queries could ask, for example,

*To create the Harmonizome, researchers with LINCS-DCIC distilled information from original datasets into attribute tables that define significant associations between genes and attributes, where attributes could be genes, proteins, cell lines, tissues, experimental perturbations, diseases, phenotypes, or drugs, depending on the dataset. These attribute tables can be searched and integrated to perform many types of computational analyses for knowledge discovery and hypothesis generation.*

## Integrated Transcriptomics: Clue.io

In addition to LINCS-DCIC, the NIH funds six LINCS (Library of Integrated Network-Based Cellular Systems) Data and Signature Generating Centers. They all gather data on perturbed cells, but each center has a different focus (such as transcriptomics, proteomics, the cellular microenvironment, disease, drug toxicity, or the brain).

The LINCS Transcriptomics Center, located at the Broad Institute in Cambridge, Massachusetts, also receives BD2K funding. The aim: to integrate LINCS-generated transcriptomics data (approximately 2 million gene expression profiles) with all the other transcriptomics data in the world (for example, the approximately 1 million profiles that reside in the Gene Expression Omnibus). "That was the premise of our proposal—to

create a unified system across all these transcriptomic sources," says **Aravind Subramanian, PhD**, principal investigator for the **Broad Institute LINCS Center for Transcriptomics and Toxicology**.

The project created clue.io, a website with an API that provides a uniform programmatic interface to all the transcriptomic datasets. Clue.io also offers web applications that can be used to find relationships between genes, compounds and diseases. "BD2K funding allowed us to build all these tools," Subramanian says. "We're hoping that these APIs we are creating to expose the data will be taken up by other BD2K centers and integrated with the other datasets and methodologies they've been developing."



the number of individuals in these datasets who share a clinical phenotype, genomic variant and activity profile. And they could achieve this **interoperability** without any Center seeing another Center's data. According to **Benedict Paten, PhD**, associated research scientist at BDTG, "This is an example of what can happen when big centers with expertise in different areas coordinate and come together."

## Interoperability of Genomics Datasets and Tools

In the good old days, a researcher would upload data to a web server where stand-alone tools would do the analysis. Today, large datasets often reside in specialized cloud environments. Tools must be brought to the data, rather than the other way around.

Researchers at the BD2K Center **KnowEnG**, a collaboration between the University of Illinois at Urbana-Champaign and the Mayo Clinic, realized that the analytical tools they develop have to be more than just **accessible** and downloadable. "Our tools have to be able to talk to other data repositories and other code bases and analysis systems for the user to have a reasonable experience in their analysis pipeline," says **Saurabh Sinha, PhD**, professor of computer science at the University of Illinois at Urbana-Champaign.

For example, Sinha says, The Cancer Genome Atlas (TCGA) is a big data repository that is part of the Stanford Genomics Cloud. "If you want to analyze those data using our kinds of tools, there needs to be a convenient and formal mechanism for these different systems to talk to each other, rather than the researcher making it happen by brute force." The solutions KnowEnG researchers are developing should be available within a year. "We're working on a way by which researchers can invoke our tools from their cloud and analyze the TCGA data right away on that cloud," he says.

Ideally, Sinha says, many such interactions will be possible in the future. "Researchers will be able to say 'get me this slice of the LINCS data, and analyze it with this pipeline in KnowEnG,'" he says. "If this kind of goal can be achieved, it will be fantastic."

## Accessing Data for Reuse

Some large health datasets created over many years remain locked in formats that aren't truly accessible. For example, in the 1960s the Centers for Disease Control (CDC) began conducting surveys and interviews to better understand the health and nutrition status of the American people. Since 1999, this survey, called the National Health and Nutrition Examination Survey (NHANES) has been continuous, covering about 5,000 people each year. The data gathered covers a broad range

of topics and, until recently, was stored in about 250 Excel spreadsheets at the CDC. Anyone hoping to analyze the data could do so only in these discrete subsets.

Researchers at the **PIC-SURE BD2K** Center set out to correct this problem and establish a prototype user-interface that would simplify analysis of NHANES data. First, they integrated the NHANES data, combining thousands of different variables—clinical, environmental, self-reported, and genomic—into one set of data structures. They then loaded the data into a software system called **i2b2/tranSMART**, which makes the data **accessible** to researchers in a web

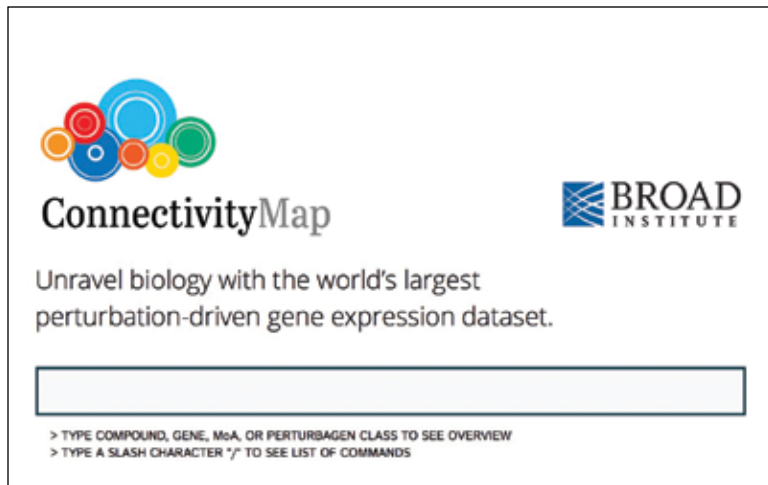
## Real World Data Sharing

Genetic testing sometimes reveals that patients have a heightened risk of breast cancer. Certain variants in the BRCA1 and BRCA2 genes, for example, raise the lifetime risk from about 12 percent to as much as 65 percent. But if a woman has a different variant of these genes (and there are literally thousands of possible variants), doctors don't necessarily know their significance. As a result, clinical tests often reveal variants of uncertain significance, as much as 20 percent of the time, and occasionally different clinical testing companies give women different information about their breast cancer risk. To reduce that uncertainty and inconsistency, BDTG, ENIGMA and many collaborators have created the BRCA Exchange, a public resource that aggregates and unifies data on BRCA variants. The Exchange is now the world's largest aggregation of genetic variance, Paten says.

To curate information for the BRCA Exchange, experts review information about BRCA1 and BRCA2 variants to develop consensus about what they signify. "We started with 1,000 expert curations and now have 3,500," Paten says. Ultimately, he says, all the variations on the site will be curated. The various clinical testing companies will be able to cross-reference the BRCA Exchange and deliver better information to patients. "It's a real-world project that is trying to deal with data sharing right now," Paten says. He also sees it as a model for other diseases for which there are genes of interest and communities of people who want to understand their potential risks.

browser as well as adding a layer of analytical tools. In the user interface, researchers can easily drag and drop different patient characteristics into boxes for comparison using statistical tools that are part of i2b2/transSMART. Furthermore, to

their work. They also added Docker technology (that they then contributed back to the Jupyter open-source hub) to create a protected research environment for each researcher. “We have one Docker container per investigator,” Avillach explains, “so no one can crash the container of another investigator.” It’s a prototype for computational research that Avillach hopes becomes the norm.



*Clue.io is a website with an API that provides a uniform programmatic interface to all the transcriptomic datasets.*

allow large-scale analysis across the entire dataset, the BD2K PIC-SURE team implemented a RESTful API called the PIC-SURE RESTful API “This is a programmatic way of [accessing](#) data for large scale computing,” says **Paul Avillach, MD, PhD**, assistant professor of bioinformatics at Harvard Medical School and member of the PIC-SURE team.

In addition to the NHANES dataset, the API can be used to analyze i2b2 patient electronic health records (with different levels of privacy [access](#) for different types of users), data from the Exome Aggregation Consortium (ExAC) browser at the Broad Institute, or any other dataset a researcher would like to import into the system.

## *Reusability and Reproducibility Using Jupyter*

Often, computational research is done on a postdoc’s laptop. Eventually that person moves on to a different lab or project and leaves an insufficient record of the steps taken or even the location of the computer script. The result: the work is not **reproducible**. But using Jupyter Notebooks, an open-source web application, researchers can detail all the steps of a computational project from input to output using any combination of 40 different computer languages. When researchers publish a paper, the Notebooks can be published alongside the data. “Jupyter Notebooks are a very nice way of doing **reproducible** science for real,” Avillach says. “They allow you to share how you managed to process the data so someone else can **reproduce** the exact same results.”

At the PIC-SURE Center, Avillach and his colleagues established a system for using Jupyter Notebooks to track

## *Piloting the Commons Cloud Credits Model*

In October 2014, Bourne announced plans to create the “NIH Commons” to catalyze the sharing, use, [reuse](#), [interoperability](#) and [discoverability](#) of shared digital research objects, including data and software. The Commons is portrayed as a layered system consisting of three primary tiers: high-performance and cloud computing (at the bottom); data, including both reference datasets and user-defined data (in the middle); and (at the top) services and tools, including APIs, containers, and indexes, as well as scientific analysis tools and workflows and—eventually—an app store and interface designed for users who are not bioinformaticians.

To be eligible for use in the Commons, data and software must meet the **FAIR** principles. For example, the products of all the BD2K centers will be part of the Commons ecosystem, including dataMED from bioCAD-DIE and the CEDAR Workbench. And to incentivize participation in the Commons, the NIH is piloting a plan to offer cloud computing credit vouchers that researchers can use with a provider of their choice, so long as the provider complies with the **FAIR** principles. Several BD2K Centers are participating in the pilot, including KnowEnG, PIC-SURE, and BDTG.

As part of their pilot, PIC-SURE took a HIPAA-compliant research environment (for using privacy-protected patient data) that they had developed for use in the Amazon cloud and added Docker containers to make it cloud-vendor agnostic. “We realized that we didn’t want to be limited to one cloud vendor,” Avillach says. It is now useable across multiple cloud vendors including Amazon, Google and IBM cloud layers.

Hausler strongly supports the Commons effort. “It’s very important that we make it easy for NIH researchers to [access](#) data and compute on the cloud and in so doing share data,” he says. From a financial point of view, having NIH principal investigators each build their own computer facilities for these big data comparisons will end up costing billions more than if researchers can work together in a common computing environment with competitive pricing, he says. But the important thing is the science. “You can’t make progress unless you can share data,” he says. “The technology exists to do it. It’s just the will and the organization. I think we’re at a critical point. We’re very enthusiastic about continuing to work on it.” □

## EXPLORING PATTERNS IN BIG DATA USING ClusterEnG, A CLUSTERING ENGINE FOR GENOMICS



In this age of genomic data deluge, researchers need integrated resources that can efficiently identify hidden structures in data. Researchers often use clustering, a popular machine-learning technique, to explore similarities within data. But they must choose from several clustering algorithms that may yield different results depending on input data and algorithm-specific metrics. Experimental biologists or even bioinformaticians may not be aware of the pros and cons of diverse clustering algorithms that vary in their complexity and ease of use; their ability to handle noisy data, outliers, or datapoints that aren't well separated; their computational expense; and how well they work on non-linear datasets.

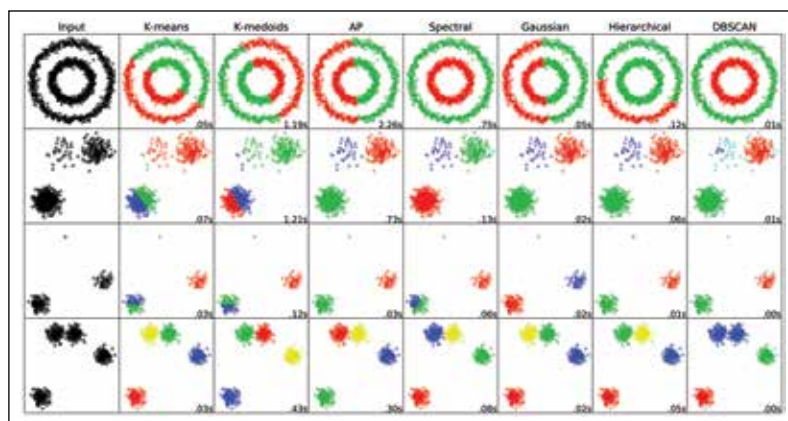


ClusterEnG is a web resource that aims to address that problem. First, it provides a tutorial that defines and describes the pros and cons of various popular clustering algorithms: k-means, k-medoids, affinity propagation, spectral clustering, Gaussian mixture model, hierarchical clustering and DBSCAN. Second, it offers the opportunity for users to upload a file with numeric data in a tabular format and perform clustering analysis on that data. If a user is not sure which algorithm to choose, ClusterEnG offers the option of selecting several algorithms to explore the results. This gives researchers an idea of the structure of their data (see Figure).

ClusterEnG provides visualization of clustering results by performing principal component analysis, a commonly used dimensional reduction technique. The first three principal components of the input dataset are plotted two at a time for 2D visualization, while all three are also plotted together in 3D. The ability to pan,

zoom and select datapoints in the 3D visualization is particularly helpful in revealing the hidden patterns in data.

The user can also explore the similarities and differences between various algorithms by selecting one of the sample datasets available on the ClusterEnG site. One option is the NCI60 gene expression dataset, which provides data for cell lines from 9 different types of cancer tissue of origin. Using this high-dimensional dataset for which the labels of the samples are known, users can



*This table compares clustering results and run-times for six different algorithms (columns) applied across four input datasets (rows) with different structures (non-linear noisy circles, boxes with different densities, data with an outlier, and close boxes). Colors represent cluster labels. Spectral clustering and DBSCAN beat other methods when the dataset has circular structure and boxes with different densities (top two rows). For the second type of dataset, affinity propagation works better than others in most cases. The third dataset includes an outlier not far from the clusters, and Gaussian mixture model clustering does best at finding the outlier. The last row shows a dataset with four clusters, two of them are close to each other. All the algorithms do well, but the ways they partition the two close clusters are different. Image courtesy of the authors.*

see how some algorithms perform better at clustering different cancer types. For this dataset, for example, k-medoids and spectral clustering prove to be better at clustering like with like.

For high-dimensional datasets where the structure is unknown, the process of identifying the best clustering algorithm for the data may require some biological intuition about the data's structure as well as iterative trial and error—i.e., visualizing the principal component plots and experimenting with several algorithms until a meaningful pattern emerges.

Clustering can be a powerful way to explore data, but it is important to understand which methods to use and how to use them correctly. By allowing users to explore their data using multiple clustering algorithms, the ClusterEnG web resource provides much-needed assistance for biomedical researchers dealing with Big Data. □

### DETAILS

Mohith Manjunath is a postdoctoral research associate and Yi Zhang is a graduate student in Jun Song's lab at the Carl R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign. They are part of the development team for ClusterEnG, which was developed as part of the KnowEnG BD2K Center and can be accessed at <http://education.knoweng.org/clustereng>.

Stanford University  
318 Campus Drive  
Clark Center Room W352  
Stanford, CA 94305-5444

### SeeingScience

BY KATHARINE MILLER

## VISUALIZING HUMAN GENOME VARIATION

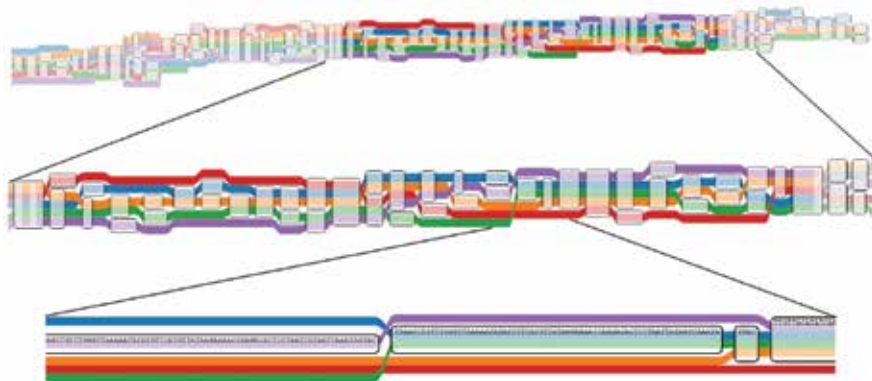
**H**umans share 99.5 percent of their DNA sequence, but that still leaves plenty of variation to go around. To get a handle on which variations contribute to health or disease, researchers typically compare individuals' genomes to a single "reference" genome that represents an assemblage of very high quality human genome sequences.

But now researchers are envisioning a better way to think

about reference genomes by building a genome graph that represents not just a single linear genome but also known variation. "The graph is this comprehensive representation of human variation that allows us to have a discourse that computers can understand about all of the different ways that humans vary," says **Benedict Paten, PhD**, associated research scientist at the **Big Data and Translational Genomics (BDTG) BD2K Center** at the University of California, Santa Cruz.

BDTG is developing tools that Paten says are "head and shoulders better" for understanding human genome variation than what can be done with just a linear reference genome. Alongside the applications that enable analysis of the human genome variation map, Paten's team is creating a visualization tool that shows the way different humans are explicitly represented within the graph. "These are pretty pictures that make intuitive sense," Paten says.

"The work started out largely theoretical but is really moving toward something that we think will have wide-scale practical application in the next few years," Paten says. □



*This prototype visualization of a genomic variation graph zooms in on portions of the NOTCH2 gene, an important gene for development. The colored bands represent 5 different variants of the gene, with rectangular shapes representing nodes (shared DNA sequences) and the colored ribbons between nodes representing paths/edges (not sequences). In the top panel, introns are shaded out (at right and left) while the solid colors represent exons 4 and 5. The exons are shown in increasingly greater detail in the bottom two panels. The visualization tool can also provide an intuitive graphical view of inversions, as shown in the green and red loops in the simulated example to the right. Images courtesy of **Wolfgang Beyer**, software developer for the Computational Genomics Laboratory at the University of California, Santa Cruz.*

