Stanford | Stanford Data Science Initiative

Funded Research Projects in Data Science

OCTOBER 2015

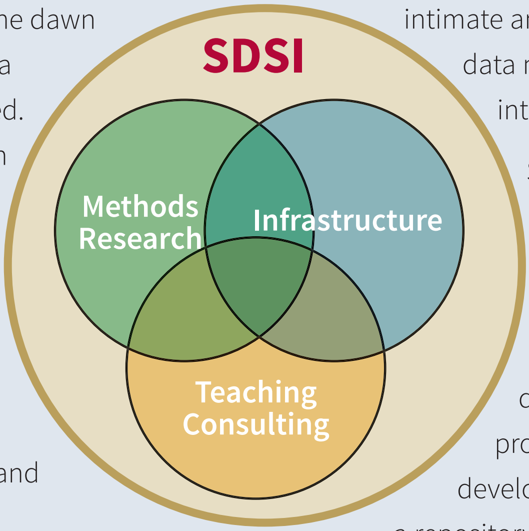
Funded Research Projects

Flagship Projects

Small Projects

What is SDSI?

The Stanford Data Science Initiative (SDSI) is a university-wide organization focused on core data technologies with strong ties to application areas across campus.



Data has supported research since the dawn of time, but there has recently been a paradigm shift in the way data is used. In the past, data was used to confirm hypotheses. Today, researchers are mining data for patterns and trends that lead to new hypotheses. This shift is caused by the huge volumes of data available from web query logs, social media posts and blogs, satellites, sensors, medical devices, and many other sources.

Data-centered research faces many challenges. Current data management and analysis techniques do not scale to the huge volumes of data that we expect in the future. New analysis techniques that use machine learning and data mining require careful tuning and expert direction. In order to be effective, data analysis must be combined with knowledge from domain experts. Future breakthroughs will often require intimate and combined knowledge of algorithms, data management, the domain data, and the intended applications.

SDSI will meet these challenges by striving to achieve a number of goals. The initiative will develop new algorithms and analytical techniques, foster collaboration with domain scientists generating big data, provide a gateway for corporate partners, develop shared data analysis tools, provide a repository of data and software, and develop relevant courses.

The SDSI consists of data science research, shared data and computing infrastructure, shared tools and techniques, industrial links, and education. As an expression of its collaborative approach, the SDSI has strong ties to many groups across Stanford University including medicine, computational social science, biology, energy, and theory.

Letter from the Directors

Steve Eglash

Hector Garcia-Molina

If you are reading this then you already know that we are in the midst of a data revolution. Industry and science have been collecting, analyzing, and acting on data for a very long time, so you might well ask what is new? There is more data than ever before. Due to the proliferation of social media, sensors, and the Internet of Things, there are more sources of data. The cost of storage has plummeted, making it economical to store massive amounts of data. The power of processors, particularly GPU clusters, continues to increase. And mobile communications provides us with more ways to generate, interact with, and use data.

The results of the data revolution are, well, revolutionary. New hardware and software can handle massive data sets and one result is the rise of statistical and probabilistic approaches over deterministic techniques. Modern data science provides us with the opportunity to transition from retrospective analysis to what-if scenario planning, prediction, and discovery. Machine learning and deep learning enable us to uncover subtle and complex relationships, often with dynamic and flexible features. The convergence of machine learning techniques and natural language processing provides us with the ability to search and extract information from structured, unstructured, and semi-structured data. Results are increasingly contextual and human centered. Advances in resolution and segmentation are giving us the ability to act on individuals rather than averages.

The implications for industry, science, and scholarship are dramatic. Products and services can be personalized. Sentiment analysis provides insight on preferences and emotions. Real-time physical data can inform actions based on what is happening in the world right now. The diversity of data sources means that our answers and predictions are more accurate and more robust.

The Stanford Data Science Initiative was formed to connect industry with Stanford's research, faculty, and grad students. The core of SDSI is the research. The articles in this volume describe progress in SDSI's first ten funded research projects. Each of these projects is developing new algorithms and analytical techniques to deal with very large data sets and challenging applications. In many cases the techniques being developed will be useful in other domains. These researchers are creating the future of data science and we can hardly wait to see what happens next.

Steve Eglash
Executive Director, Stanford Data Science Initiative
Executive Director, Artificial Intelligence Lab
Executive Director, Secure Internet of Things Project
Stanford University

Hector Garcia-Molina
Leonard Bosack and Sandra K. Lerner Professor
in ihe School of Engineering and Professor of
Electrical Engineering, Stanford University

Secure Analytics on the Internet of Things

Here's a programming challenge that's not for the faint-hearted: Write a library of cryptography software so secure that it can withstand an attack not only from today's most powerful computers, but also from computers that haven't even been built yet. And one more thing: the program won't be running on some high-end scientific workstation, but instead, inside of an ordinary doorknob.

That's a taste of the challenge faced by the "Secure Analytics on the Internet of Things" program at SDSI. Its mission: to make sure that the next Internet — which will one day include billions of household appliances, home security systems, personal health monitors, and yes, even doorknobs — is more secure than the current one. The Secure Analytics program aims to do this by designing security into the "IOT" from the ground-up, something the architects of the first Internet, who were mainly building a messaging system for themselves, failed to do.

“When you expect things to be installed for 10 or 20 years, that turns out to have big implications for security,” said Phil Levis, the associate professor in Stanford’s Computer Science and Electrical Engineering Departments directing the program.

The “door knob problem” is emblematic of the issues Levis’ program is dealing with. Since a doorknob is expected to last for several decades, he said, its built-in security will need to last just as long, meaning it must anticipate new developments in computing technology, such as the quantum computers that physicists all around the world are racing to build.

Because some current security technology is known to be vulnerable to a quantum computer, Levis said that a secure doorknob — or any other IOT device, for that matter — is going to also need to include “standby” quantum-proof software that would be activated if quantum machines ever become real.

Encryption plays a crucial role in the Secure Analytics program because so much of the data on the Internet of Things will be extremely personal, notably medical information from the portable monitoring devices that are being used more and more in medicine. Levis said the several of the six or eight research programs that are part of Secure Analytics will be taking advantage of recent programming breakthroughs that will allow computer programs to work with encrypted data without first having to decrypt it.

Several of the six or eight research programs that are part of Secure Analytics will be taking advantage of recent programming breakthroughs that will allow computer programs to work with encrypted data without first having to decrypt it.

Levis said the program will also work on developing the advances in data science that will be necessary to allow computers to deal with the torrents of “noisy” data that will be generated by the many things that will be connected via the IOT.

For example, engineers are designing a proof-of-concept program that will put low-cost water monitors on all of the showers in a Stanford dorm, and then upload usage information into the cloud for water conservation monitoring. That sort of data will be common in the Internet of Things, said Levis, but very little of it will have the tidy structures that data scientists are used to working with.

Yet another issue complicates all of the work of Levis and his team: Power. Most of the devices expected to be on the Internet of Things will have minimal power supplies; the batteries in today’s mobile phones are gigantic by comparison. Among other things, that means designing an IOT operating system that places a premium on low power consumption, another task that engineers in the Secure Analytics program are now tackling.

The first Internet rolled out slowly, with university researchers calling all the shots. But the Internet of Things is being deployed apace, as a visit to any hardware store full of smart thermostats and garage door openers will attest. Levis said that now is exactly the right time for this project. On the one hand, technology has just reached the tipping point. On the other hand, it’s still early enough.

“There are going to be billions and billions of devices,” he said. “Sure, there are some devices out there now, but we’re just at the beginning. We’re not even close to the penetration of smart devices that we’ll see in a few years.”

INVESTIGATORS

Philip Levis, Associate Professor of Computer Science and Electrical Engineering, Director of the Secure Internet of Things Project

Noah Diffenbaugh, Associate Professor of Environmental Earth Systems Science

Christopher Ré, Assistant Professor of Computer Science

Dan Boneh, Professor of Computer Science and Electrical Engineering, Co-Director of Stanford Computer Security Lab

Mark Horowitz, Professor of Electrical Engineering and Computer Science

Jure Leskovec, Assistant Professor of Computer Science

Mapping the “Social Genome”

The Human Genome Project has taught us that what emerges from the billions of apparently random ones and zeroes of genetic sequencing is nothing less than the key to understanding life. Now, hopes are high that a vast new source of digital information will shed light on a subject that is equally complex — human behavior.

“Mapping the Social Genome” is a research project in the Stanford Data Science Initiative that aims to make predictions about human behavior by sifting through the Everests of data being generated today, especially on the Internet. It aims to answer questions as diverse as “What’s the best way to design a discussion group inside a company?” to “Which criminal defendants should be allowed to get out on bail?”

Jure Leskovec, the assistant professor of Computer Science heading up the Social Genome project, said the field might be considered "computational social science," and that questions like the bail issue have already yielded promising results. "It's almost as though we are writing down the equations of human behavior," he said.

For example, looking at more than 100,000 records from several Illinois counties, researchers were able to come up with an algorithm that has so far proven to be up to 25% more accurate than human judges at predicting which defendants will either not show up for trial or else commit another crime while out free on bail.

That wasn't the first time that Leskovec and his colleagues have used data science to predict human behavior. Earlier in 2015, he and two other researchers made news when they could predict whether someone would become a disruptive "troll" on an Internet Web site simply by examining the first few posts the person made. Their analysis discovered, for example, that future trolls are much more likely to make comments that are irrelevant to the actual topic at hand.

Looking at more than 100,000 records from several Illinois counties, researchers were able to come up with an algorithm that has so far proven to be up to 25% more accurate than human judges at predicting which defendants will either not show up for trial or else commit another crime while out free on bail.

Leskovec describes the Social Genome project as a continuation of the centuries-long project of making science ever more empirical and data based. Just as physics has been able to turn to ever more power particle accelerators to probe the nature of the atom, Leskovec said social scientists are using the abundance of what Leskovec called "digital traces" of our activities so that they can "understand human behavior at a new level of resolution."

It's well known that retailers, among others, stand to profit from better understanding the online behavior of their users, if only to better target their offers and advertisements. But Leskovec said that all companies can benefit from the sorts of Big Data tools usually associated with Google or Amazon.

For example, he said that techniques are being developed to predict the health of a company by looking at the patterns of how people inside the company are communicating with each other. And Leskovec isn't talking about reading the content of people's private emails.

"It doesn't matter what people are saying. What matters is who is talking to whom," he said. "Companies should be asking, 'Does my communication network look healthy, with different parts of the company talking to each other? Or is my network fragmented, without information spreading?' People just aren't doing this today, because the tools don't yet exist."

This is a situation that is changing. "Because we have lots of data, and are developing solid theories," he said, "we are now beginning to make observations and spot differences that previously would have simply been impossible to see."

INVESTIGATORS

Jure Leskovec, Assistant Professor of Computer Science
Michael Bernstein, Assistant Professor of Computer Science, Co-Director of Stanford Human-Computer Interaction Group
Amir Goldberg, Assistant Professor of Organizational Behavior
Dan Jurafsky, Professor and Chair of Linguistics, Professor of Computer Science
Dan McFarland, Professor of Education, Director of Stanford Center for Computational Social Science
Christopher Potts, Associate Professor of Linguistics, Director of Stanford Center for the Study of Language and Information

Data Science for Personalized Medicine

A petabyte is a lot of ones and zeroes; to hold that much data, you'd need a stack of disk drives, the sort found in an average PC, as tall as a 15-story building. But that is how much medical information Mike Snyder has collected *just about himself*. And now, with the data in hand, Snyder is leading a multi-campus research program to learn how to best store and use that information. It's an effort that puts him at the crossroads where modern medicine meets cutting-edge data science.

Snyder chairs the Genetics Department at the Stanford Medical School, and is principal investigator of the SDSI-funded project, “Data Science for Personalized Medicine.” The general goal of personalized medicine is to use data collected about a patient to tailor a custom-fitted treatment for a particular illness; one of the widely covered efforts in personalized medicine involves using the specific DNA of an individual’s cancer tumor to design a unique chemotherapy agent to combat it.

Snyder’s three-year program is specifically aimed at diabetes, but shares with the cancer program the fact that modern medicine is capable of collecting staggering amounts of data, and the hope that proper use of all that data will make for longer, healthier lives.

Snyder began focusing on diabetes three years ago, when after sequencing his own DNA, he discovered an elevated risk for Type II diabetes. (This despite the fact that Snyder is himself trim and fit.) Snyder began regularly collecting blood samples and in 2013 learned through lab tests that he had developed Type II diabetes, most likely on account of a viral infection.

Anxious to discover if his diabetes could have been more accurately predicted, Snyder expanded his one-person research group, and it now includes about 100 volunteers, many of them fellow researchers, all of whom give blood samples monthly.

Snyder’s three-year program is specifically aimed at diabetes, but shares with the cancer program the fact that modern medicine is capable of collecting staggering amounts of data, and the hope that proper use of all that data will make for longer, healthier lives.

The enormous amount of information this group is producing — Snyder is paying $25,000 a month in data storage fees — is a direct result of a generation of breakthroughs in medical diagnostics. We can, of course, sequence an individual’s genome with its six billion base pairs—that in itself creates a terabyte of data. But we now know the genome is just the start of things.

There is, for example, the equally complex epigenome, which are the changes the genome has undergone since the organism’s birth, as well as the proteome, which is the armory of proteins produced under orders of DNA.

The latest entrant: The microbiome, or the organisms that are, as Snyder put it, “in and on you, but aren’t really you.” Each of us has 10 times as many microbiomic cells as autonomous ones; the microbes in our intestinal system alone would, all by themselves, weigh three pounds.

It is the interactions of all these systems that make us either healthy or sick. That process, now shrouded in mystery, is what Snyder is trying to make explicit. And if that can be accomplished for diabetes, the techniques will be relevant to a host of other diseases.

Snyder’s SDSI effort thus faces the obvious, and age-old, data science problem of figuring out whether a link between two events involves causation or merely correlation. It’s a problem that becomes an order of magnitude more pressing with the sheer amount of data his team is collecting, and the increased chances for “false positives” that result.

But there are computer science-related problems too, Snyder said. Medical information needs to be kept highly confidential. That usually means encryption, but computer scientists have yet to figure out an efficient way of doing the sort of regular, intense work with data that Snyder is performing while still guaranteeing it stays confidential.

Overall, Snyder places his program in the context of the continuing advance of human knowledge about wealth and wellness. “We are about to enter an era where large data sets can be used to better manage health,” he said, “and to move us way from medicine that is hunch-driven, and finally towards medicine that is data-driven.”

INVESTIGATORS

Michael Snyder, Professor and Chair of Genetics, Director of Stanford Center for Genomics and Personalized Medicine

David Tse, Professor of Electrical Engineering

Euan Ashley, Associate Professor of Medicine and Genetics, Director of Stanford Center for Inherited Cardiovascular Disease, Director of Stanford Clinical Genomics Service, and Co-Director of Stanford Research Training Program in Myocardial Biology

Mohsen Bayati, Assistant Professor of Operations, Information and Technology

Dan Boneh, Professor of Computer Science and Electrical Engineering, Co-Director of Stanford Computer Security Lab

Andrea Montanari, Associate Professor of Electrical Engineering and Statistics

Ayfer Ozgur, Assistant Professor of Electrical Engineering

Tsachy Weissman, Professor of Electrical Engineering

DeepDive—a High-Performance Inference and Learning Engine

Texts, graphs, tables, pictures, and illustrations—human beings have a wide range of print-based tools that we can use to convey information, so we have no trouble flipping through a book and learning whatever the author wished to convey.

Computers, unfortunately, are utterly lost in this same terrain; for them, it's all "dark data." Before a computer can "understand" anything, the information usually needs to be highly structured, like the rows and columns in an Excel spreadsheet. The result is that, except via some rudimentary text search tools, the information in the world's billions of books, journals, documents and reports can't be effectively mined for insights by computers.

DeepDive has already proven itself in law enforcement, via a high-profile application involving human trafficking, and in fields as diverse as genomics, clinical medicine and semiconductor manufacturing.

DeepDive is a research program at Stanford University seeking to change that. It uses machine learning techniques to quickly and automatically extract structured data from totally unstructured printed sources; a tedious and error-prone process that now requires human beings — sometimes even human experts. Once dark data is transformed by DeepDive into the kind of well-ordered information that is easily accessible by standard database tools, the opportunities for new insights are endless.

DeepDive has already proven itself in law enforcement, via a high-profile application involving human trafficking, and in fields as diverse as genomics, clinical medicine, and semiconductor manufacturing.

And while DeepDive is now being applied to some of the most advanced areas of business and finance, the technology first proved its mettle with something downright prehistoric: dinosaurs.

Program director Christopher Ré said the DeepDive's computers were able to "read" through nearly one-half million paleontology journal articles and books, and then create a table that listed all the dinosaur fossils along with the location and likely taxonomic classification of each. It did this automatically after only a relatively brief period of human "training," the sort common in all machine learning applications.

Ré said his team chose dinosaurs as their first project because it would be easy to benchmark their work, on account of the existence of the "Paleobiology Database," which has been developed over two decades by a global network of hundreds of experts. DeepDive won the competition hands down. It was able to process 100 times as many data points as the human paleontologists—not only in vastly less time, but also with a 12% greater accuracy rate.

DeepDive created its "synthetic database" by combining disparate information from multiple bits of unstructured data. A particular fossil might have been described in a paragraph of text, but its location found in an accompanying table, and its appearance shown in a nearby picture.

To create Paleo DeepDive, said Ré, essentially two human steps were involved. First, researchers needed to develop the schema into which the data would ultimately be placed, something not much different from designing a database. After DeepDive had done some rudimentary natural language processing, researchers would assign a few hundred "facts" to their proper places in the pre-made schema. DeepDive then figured out the probability that a particular statement belonged in a particular part of the schema.

That emphasis on probability was key to the technology's success. "Our main insight was to treat all problems of understanding as probabilistic inference problems," said Ré. "DeepDive regards everything you show it as an observation about the world. It then takes all of that information, and all the rules the user gives it to understand that data, to predict the information's most likely location in the database."

With the billions of iterations made possible by modern multi-core computers, the results that emerge are remarkably accurate: 94% in the case of Paleo DeepDive. Said Ré, "We were astonished at the way that a small number of relatively low quality training sessions produced remarkable results."

Another pleasant surprise from DeepDive is that each succeeding application of the technology takes less time; several months for the paleontology program, but a few weeks for some of the more recent domains. He said a business application of DeepDive, such as a petroleum company working through shelves full of oil field reports, could conceivably be up and running in a few weeks.

"We've proven that computers can do this in multiple domains," said Ré. "And for us, that is very, very exciting."

INVESTIGATOR

Christopher Ré, Assistant Professor of Computer Science

Michael Cafarella, Assistant Professor of Computer Science and Engineering, University of Michigan

Large-Scale Time Series Analysis of Food Price Spikes and Malnutrition

The world becomes aware of famines when photos of tiny children with distended bellies and discolored hair start appearing on television news shows. Unfortunately, by that time, malnutrition has often taken a permanent toll on the mental and physical development of the children.

Stanford researchers Sanjay Basu and Eran Bendavid are trying to find data that might give an early warning that famine is imminent. Such information might enable governments and international aid agencies to respond to the emergency in time to reduce the impact. Some events like floods and droughts that devastate croplands are visible predictors of food shortages. But most of the political, social and logistical events that cause famine are hard to comprehend as they happen.

The researchers theorize that spikes in food commodity prices might presage food shortages and malnutrition. Food price data is readily available from almost every country. Agricultural agencies closely track prices in regional markets in order to help their farmers price their crops and sell futures. They distribute the data almost as fast as they collect it.

There is also a lot of data available in the public health sphere, but much of it is released months or years after it is collected. Since the 1980s, the United States Agency for International Development and others have conducted regular surveys of individual health measures all over the world. Decreases in arm and leg circumferences and deficiencies in certain micronutrients in blood samples are reliable indicators of malnutrition. The effects are most visible in very young children and the elderly.

The Stanford researchers are performing a historical analysis that would link food price spikes in past years to the evidence of famine collected by public health workers. They are using monthly and weekly price data from 250 local markets in 26 countries, covering 31 staple foods, from 1995 to 2013. They are linking the food data by market to individual health data on 350,000 children. They are analyzing a total of 2 terabytes of information.

Stanford researchers are performing a historical analysis that would link food price spikes in past years to the evidence of famine collected by public health workers. They are using monthly and weekly price data from 250 local markets in 26 countries, covering 31 staple foods, from 1995 to 2013.

Such time-series computations are notoriously computationally intensive. In part because the time-lag between the price spike and the famine aren't known in advance, traditional linear regression techniques can miss correlations. The researchers have based their development on a recently developed statistical technique called convergent cross mapping, and have refined it with a new approach using the Julia language for distributed parallel processing. Their algorithm narrows down the number of relationships that must be examined more closely.

Even with all the data, counterintuitive results appear. For example, in some rural areas price spikes seem to cause improved nutrition. The likely explanation: the rural farmers can sell their crops for more and have more money to buy other food. Malnutrition appears more quickly in cities. Wheat price spikes seem less likely to cause malnutrition because sorghum is usually an available substitute. But rice or maize price spikes are more likely to foreshadow famines.

The researchers anticipate reporting some results about the relationship between price spikes and income levels on nutrition later in 2015. As they refine their work, they plan to try to analyze the effectiveness of various child-nutrition plans, including food-purchasing subsidies, national grain reserves, and timing of import/export programs. They will also publish an open-source package to help scientific researchers use Julia for other research questions.

INVESTIGATORS

Sanjay Basu, Assistant Professor of Medicine at Stanford Prevention Research Center

Eran Bendavid, Assistant Professor of General Internal Medicine

Use of Electronic Phenotyping and Machine Learning Algorithms to Identify Familial Hypercholesterolemia Patients in Electronic Health Records

Familial hypercholesterolemia, or FH, is a genetic disorder caused by very high levels of "bad" cholesterol that sometimes kills adults in their 30's by causing fatal heart attacks. The condition can be treated, but it affects only one in 500 people, so few doctors diagnose it before it causes heart disease.

Nigam Shah, a Stanford Assistant Professor of Medicine specializing in biomedical informatics, is collaborating with Dr. Josh Knowles, an expert in FH, to mine electronic health records (EHRs) to identify the risk in undiagnosed individuals. Even though none of their records contain the diagnosis, using new data-mining techniques these at-risk individuals can be identified via approaches being pioneered by Dr. Shah's team.

Electronic phenotyping is made possible by the exploding volumes of data captured in electronic health records. However, in the real world, EHRs provide "noisy" data. There are myriad forms and codes in different hospital and insurance systems. A single electronic medical record installation can contain thousands of tables and hundreds of thousands of fields. On top of that, busy doctors may accidentally enter incorrect codes or not include keywords that computerized expert systems look for. So just having the data is not enough to find patients at risk. Novel data mining methods are needed to accurately find the individuals that may have a certain condition of interest, such as FH.

Rather than look for keywords, the Stanford researchers decided to train computers by showing them over 100 health records of patients who had been diagnosed with FH by Dr. Knowles. Then the computer was left to figure out what information constituted evidence of the disease. For instance, they might learn that a sibling or parent who had a heart attack in his 30's was a marker. Dr. Shah says it is analogous to the way email systems are taught to detect and delete spam, even though none of the emails contain the word "spam" and spammers constantly change their messages and addresses.

The researchers have been able to prove that their method achieves very high accuracy in identifying known cases of FH in the Stanford hospital's database, with minimal false positives. Now they are contacting other teaching hospitals with FH expertise to examine portability of the method.

Dr. Shah's team already has demonstrated that even with noisy data they can maintain accuracy of phenotyping at around 90% by enlarging the training data set.

Dr. Shah's team already has demonstrated that even with noisy data they can maintain accuracy of phenotyping at around 90% by enlarging the training data set. That means that automated electronic phenotyping can become as accurate as painstakingly developed expert rules--purely by adding more data. Developing expert rules is a chokepoint for exploiting electronic health records, often taking a year or more of work by top clinicians.

Their statistical approach provides a pathway to automating disease detection. They anticipate using the technique to diagnose hundreds of other relatively rare diseases. It is being made available as open-source software for use by other medical researchers.

The Stanford researchers have previously used data-mining techniques to highlight unexpected risks of heart attacks posed by one of the most widely used classes of new pharmaceuticals. In 2013, they analyzed huge databases of EHRs. They found that people who took proton pump inhibitors such as Prilosec to treat acid reflux disease were at increased risk of heart attacks compared to people who didn't take the drugs. The increase in risk was too small to show up in the clinical trials the drug manufacturers had conducted to get FDA approval. However, once millions of people were taking the drugs over a long period, the risks became apparent and significant.

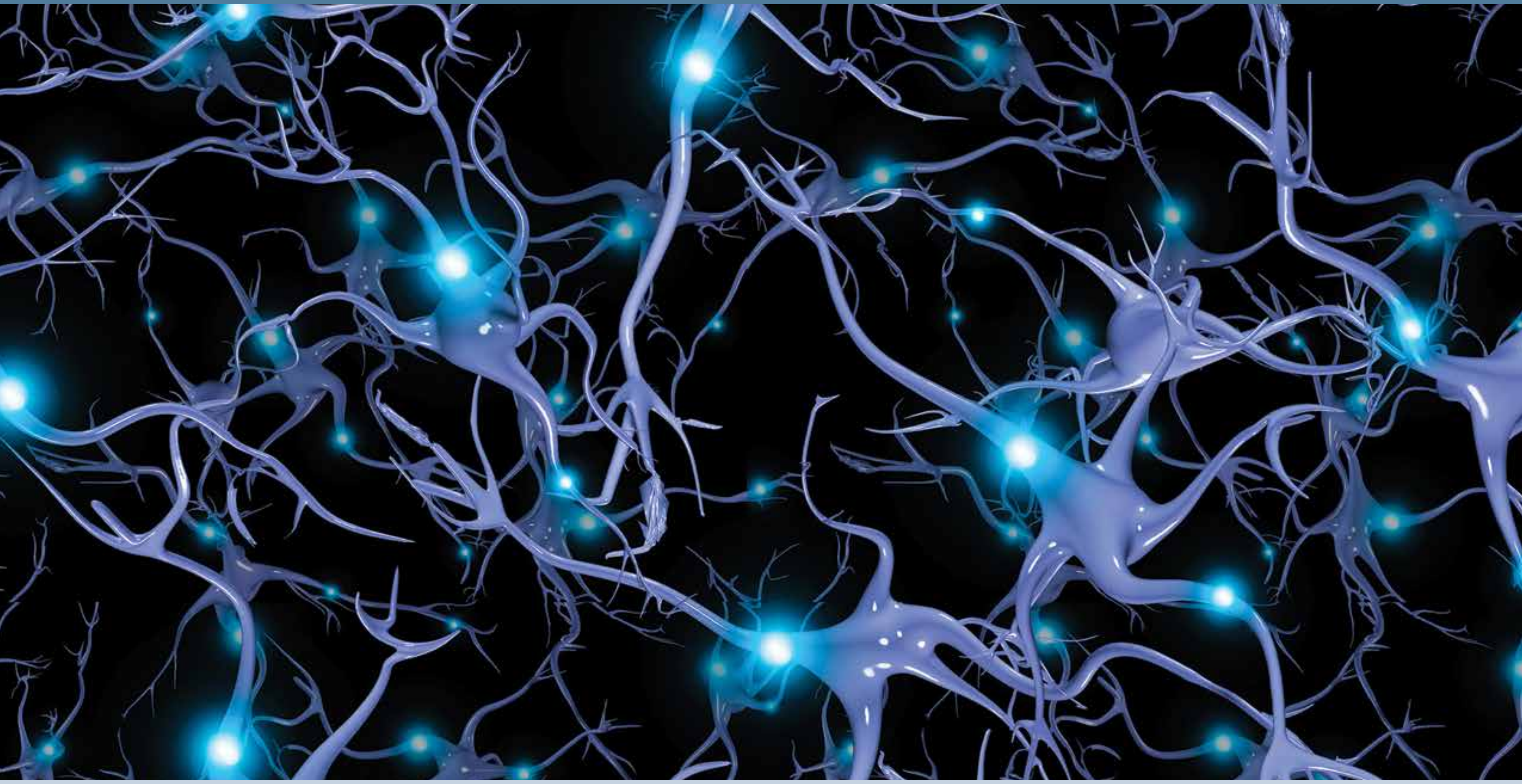
Data-mining based findings do not show cause and effect. But the association is strong enough that doctors may reconsider whether to prescribe that class of drug to patients with cardiovascular risk factors. There are alternative treatments. An older class of heartburn drugs didn't show the same association with heart disease. The ability to cheaply analyze the impact of drugs and especially lightly researched medical devices after they are in widespread use could yield significant benefits and change the practice of medicine for the better.

INVESTIGATORS

Joshua Knowles, Assistant Professor of Cardiovascular Medicine

Nigam Shah, Assistant Professor of Biomedical Informatics Research

Real-Time Large-Scale Neural Identification



The retina of the human eye does far more than record light pixel-by-pixel like a camera's CCD. Each of the 20 different output cell types of the retina processes the light impulses in a unique way before passing them along as electrical impulses to the brain. Learning the cells' function is necessary to understand what signals are sent to the visual centers in the brain. That task is a key step to the dream of building an artificial retina implant that could let blind people see.

E.J. Chichilnisky, a neurobiologist at Stanford, has been studying the cells and neural circuits of the eye for nearly 20 years. Along with his co-investigator Andrea Montanari, an associate professor of Electrical Engineering and Statistics at Stanford, he anticipates that detailed analysis of the terabytes of data that have been collected by the large-scale electrode arrays used in his lab will help to map the functions of the many cell types and how they send visual information to the brain.

Researchers can manually classify about 20 different retinal ganglion cell types. They either define them by their detailed shape as visualized under a microscope, or they use implanted electrodes to measure their electrical output in response to light. But study of these cells, some of whose functions remain unknown, is constrained by the time required for researchers to painstakingly analyze anatomical and functional data in order to classify them.

Interpreting the signals from retinal cells is made more difficult because many cells fire at once in response to visual stimuli, and when they fire, they may be recorded by several electrodes. Dr. Chichilnisky is working with 512-electrode arrays that record the activity of hundreds of retinal cells, identifying the timing of visual responses in the cells with millisecond precision. A mathematical technique called spike sorting identifies which cell has emitted a spike at each moment in time. Having tightly packed electrode arrays, and the vast amounts of data they create, is proving crucial to accurate classification of cell types.

Drs. Chichilnisky and Montanari, along with Electrical Engineering postdoctoral fellows Emile Richard and Georges Goetz, are developing new mathematical algorithms that sort the electrical signatures of retinal cells, such as the electric field produced by a cell firing spikes, and the timing of these spikes, in order to classify the cells. They develop the algorithms by analyzing past experiments using methods from machine learning and statistics. Then they test the validity of the algorithms by applying them to the results of a separate data set.

Dr. Chichilnisky is working with 512-electrode arrays that record the activity of hundreds of retinal cells, identifying the timing of visual responses in the cells with millisecond precision.

This automated classification of the different cells promises to give the researchers a major boost in understanding the neural circuits formed by the retinal cells and the receptors in the brain. As far as researchers understand, each of the 20 or so retinal cell types process the light signals they receive in a unique way. This processing is analogous to a Photoshop filter that changes the original raw photo to highlight specific aspects of the scene.

Classifying the cells in real time will be crucial to making artificial retinas that function the way natural eyes do. In many patients, current artificial retinas only distinguish light from dark over coarse patches of the scene, and in some transplant recipients they hardly work at all. This may be because the signals sent by the different cell types are received by the wrong targets in the brain. Sending a cell's signal to the wrong target is likely to create dissonance in the visual image, the way an orchestra would be disrupted if sheet music for flutes was given to cello players. With knowledge of the cell types stimulated, it is possible that prostheses with much higher fidelity can be developed.

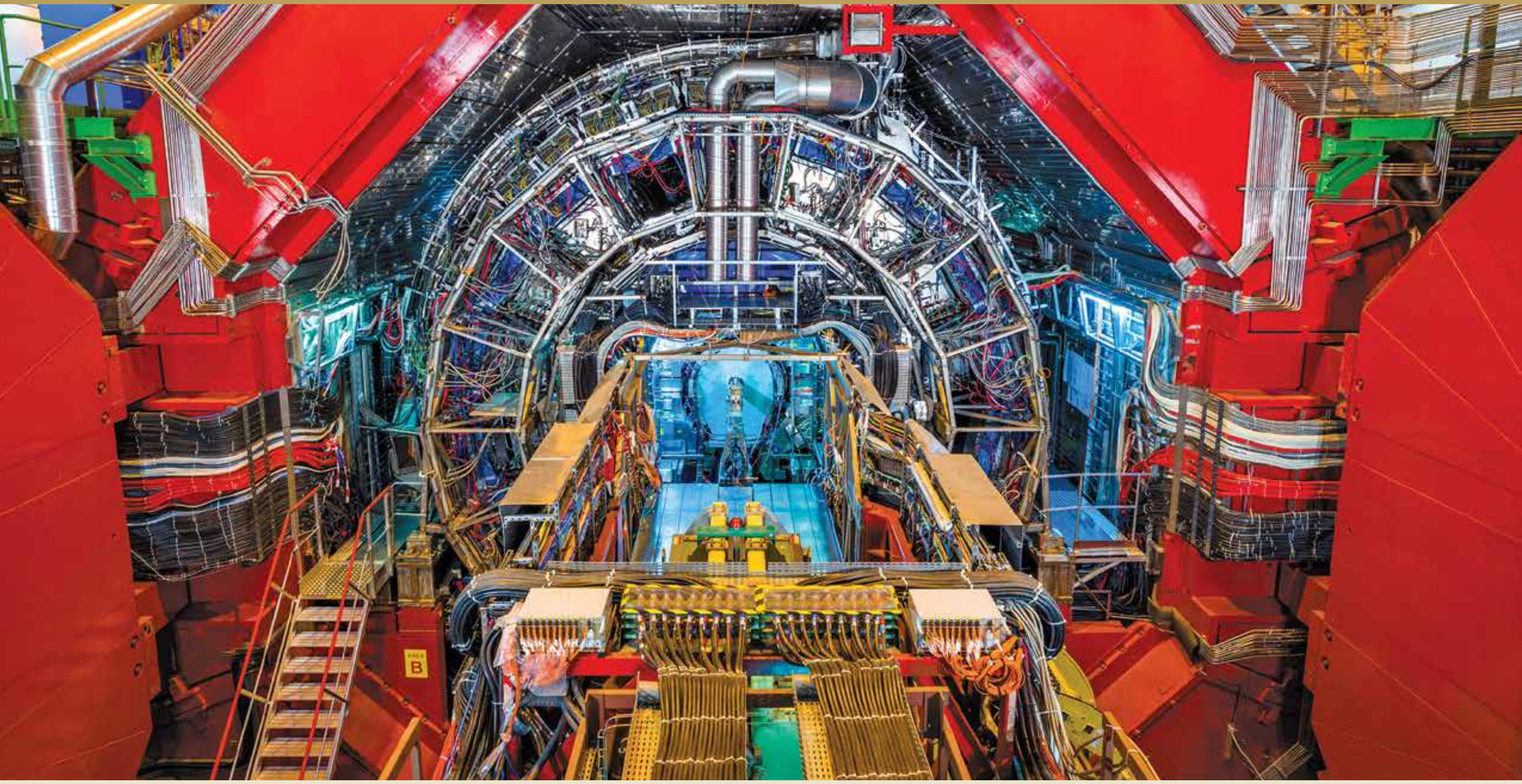
Such large-scale neural recording is transforming how we study and understand the brain and nervous system. The computational ability to analyze and mine such large data sets opens the possibility of understanding the function of entire neural circuits composed of hundreds of thousands of neurons in various parts of the brain.

INVESTIGATORS

E.J. Chichilnisky, Professor of Neurosurgery, Ophthalmology

Andrea Montanari, Associate Professor of Electrical Engineering and Statistics

Physics Event Reconstruction at the Large Hadron Collider



Data scientists and physicists both work with numbers. But they seldom work together. Stanford's Ariel Schwartzman, a particle physicist, and Lester Mackey, a professor of statistics, think that collaborating will give better understanding of the information that is being produced by the Large Hadron Collider (the LHC), the huge particle accelerator in Geneva, Switzerland. It might even help find new subatomic particles.

By applying big data analysis techniques to the petabytes of data that the LHC generates every year, the scientists expect to be able to more accurately identify and differentiate particles like W, Z, and Higgs bosons, as well as top quarks. Bosons and top quarks are important to study because they are predicted by many models of new physics, such as supersymmetry, or models with extra dimensions of space.

The LHC is designed to allow scientists to see what was happening with energy and matter shortly after the universe was created in the Big Bang 13.7 billion years ago. Then, most scientists believe, in a tiny fraction of a second, elemental forces started to expand and formed the protons, neutrons, and electrons that comprise most of the measurable universe.

In the LHC, particle collisions occur 40 million times in a second. Scientists capture the action with a detector—the equivalent of a digital camera with about 100 million pixels. Using information on the energy and number of particles, they try to piece together the frame-by-frame activity to understand what particles were created by the collision, and how they decay. For example, a W boson decays into two quarks.

Currently physicists usually group similar observations into classifications called jets, using particle clustering algorithms. Most jets are uninteresting because they are comprised of quarks or gluons. The ones the physicists want to study closely are the ones originating from different types of bosons and the top quark. More important, they anticipate finding additional particles, the existence of which is predicted theoretically but hasn't been observed. New particles are required to answer some of the most fundamental questions in particle physics today, such as the origin of dark matter. That quest is one of the highest priorities of the LHC experiments.

In the LHC, particle collisions occur 40 million times in a second. Scientists capture the action with a detector—the equivalent of a digital camera with about 100 million pixels.

Prof. Mackey and Prof. Schwartzman are working to make it easier to identify the jets that are of greatest interest. To spot a W, Z, or Higgs boson jet, physicists traditionally use algorithms to identify a two-pronged structure. The Stanford scientists developed a new technique that turns the LHC events into a visualization of a jet and then uses computer vision to look at the images. The technique extracts more information from the LHC data and increases the precision of identifying the jets. With better identification of these rare boson jets, scientists have more opportunity to search for new particles decaying into bosons and the top quark.

A member of their team is currently trying to harness deep neural networks, another computer science technology, to increase the precision of the identification techniques. This will improve the separation of the signal (the boson jets) from the noise (the vast majority of jets that don't originate from bosons).

The goal of the collaboration doesn't end there. The scientists also think that they will be able to define for the first time the limits of information that can be extracted from the LHC. They want to determine whether increasing the granularity of the data — the number of pieces of information that can be extracted from a single collision — will improve the scientific understanding. Among other things, they expect to find the point at which increasing the power used to create collisions will stop producing more useful information.

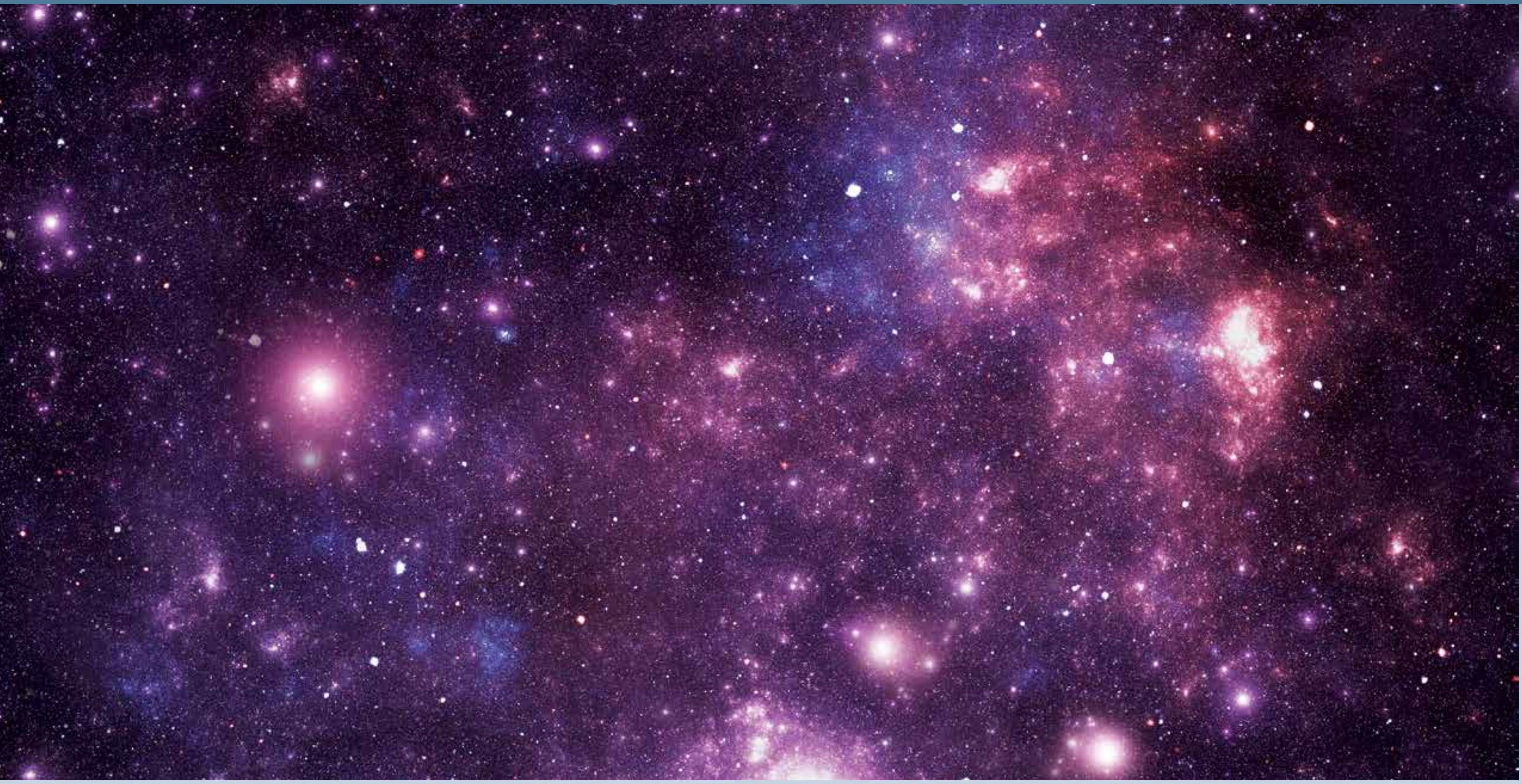
Using machine learning and big data to categorize particles is one of the hottest study areas in particle physics today. The Stanford group is relying on a tight collaboration between experimental physicists and data-chomping statisticians to develop a deeper understanding of the physics of jets, and to develop new and more powerful ways to analyze and interpret LHC events.

INVESTIGATORS

Lester Mackey, Assistant Professor of Statistics

Ariel Schwartzman, Assistant Professor of Particle Physics and Astrophysics

Inferring the Mass Map of the Observable Universe from 10 Billion Galaxies



Sometime in 2021, a new telescope atop a mountain in Northern Chile will start taking pictures of half the sky, encompassing some 10 billion galaxies. The images will be captured by the world's largest digital camera—a 3.2-billion pixel, 3-ton device, about the size of a small car—that is under construction in Menlo Park, California.

As remarkable as the Large Synoptic Survey Telescope (LSST) and its camera are, they won't be able to directly picture most of the mass of the universe. Dark matter is invisible to the camera, even with ultraviolet and infrared filters, because it neither emits nor reflects light.

Stanford researchers Risa Wechsler and Phil Marshall are developing mathematical methods for identifying those masses and mapping them in three dimensions. An accurate map of all of the galaxies in the observable universe and the dark matter contained within them may help us understand the origins of the universe and why it is expanding at an accelerating rate. The color and brightness of a distant galaxy help astronomers determine its distance and, coupled with its location in the sky. provide a good approximation of its location in three dimensions.

Computational analysis of the distortions of light emitted by distant galaxies can then be used to "see" the dark matter in individual galaxies, or at least know how much dark matter is likely to be there.

Even though dark matter doesn't produce light, it distorts light waves. Light emitted from a galaxy doesn't travel straight to the telescope: it is bent by the gravitational fields of the invisible dark matter that it passes, an effect known as "gravitational lensing." Because the gravitational pull is predictably related to the mass of the dark matter that makes up most of a galaxy, it gives astronomers a way to measure the mass of that galaxy.

> ***Computational analysis of the distortions of light emitted by distant galaxies can then be used to "see" the dark matter in individual galaxies, or at least know how much dark matter is likely to be there.***

A very large mass near a light-emitting galaxy will distort the light dramatically. Circular galaxies appear ellipsoid due to these distortions. Even small masses along the light's path will bend the light slightly. Light from distant sources may be bent repeatedly along its billions-of-light-year journey. Analyzing such "weak lensing" effects is expected to give a much more detailed and accurate map of the individual galaxies in the universe.

The researchers are already starting to apply their statistical methods to data sets from the two-year-old Dark Energy Camera which is taking pictures in Chile with a 570 megapixel camera. Its pictures cover 300 million galaxies. LSST will image four times more sky, about 70 times more often, and will require new computational algorithms to take advantage of all this additional data.

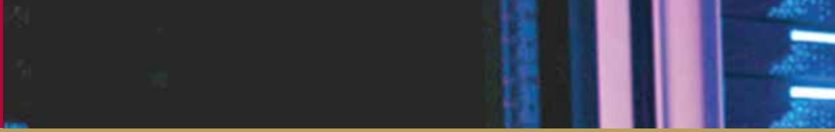
Wechsler's team is also working to build plausible model universes that can be used to simulate the actual one and allow experimental tests of the weak lensing analysis. By running simulations they can study things that cannot be followed directly in real life, such as how galaxies form. The output of the simulations will provide pointers to telltale pieces of evidence that should be present in the vast amounts of data that come out of the Large Synoptic Survey telescope.

Astrophysicists around the world are preparing to analyze the LSST's digital images and related data, anticipating an opportunity to understand the universe at a scale and level of detail never before possible. Some of the results are likely to prove theses that are already widely believed based on sampling and averaging of current, smaller-scale experiments. But in other cases the more detailed picture will reveal unexpected results that alter our understanding of the universe.

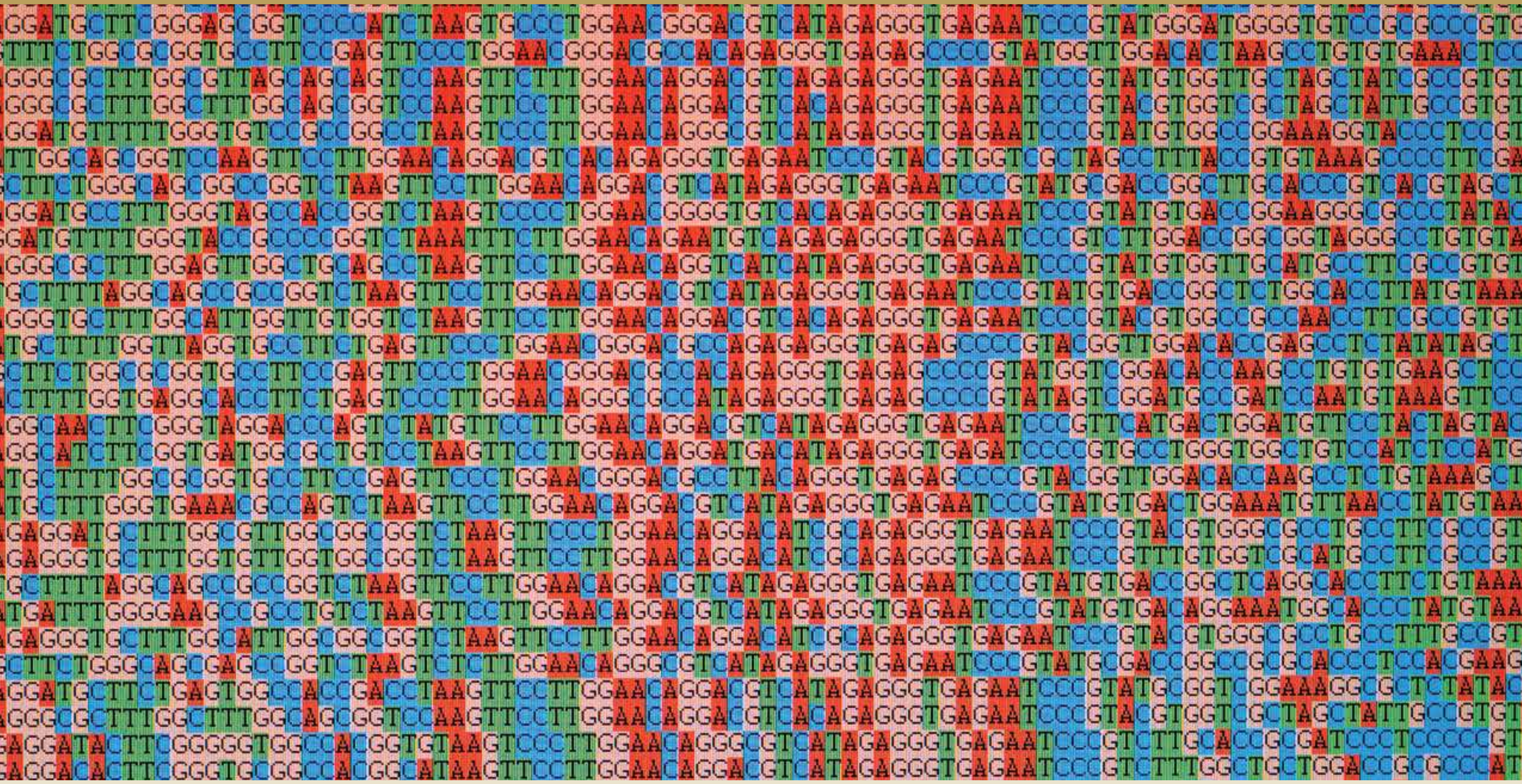
INVESTIGATORS

Risa Wechsler, Associate Professor of Physics

Phil Marshall, Scientist at SLAC National Accelerator Laboratory



The Stanford Resident (Reason-Syndicate-Identify) Project: Toward Knowledge-Driven Medical Genomics



Recently parents of a young boy with troubling signs of developmental disabilities brought him to Stanford's pediatric hospital. The reasons for his issues stumped the doctors who examined him.

Applying cutting-edge technology, the doctors ordered a genetic assay. The researcher who reviewed the assay found a mutation in the boy's genetic circuitry that he suspected might cause problems.

The mutation wasn't common, and it didn't immediately suggest a disease. But when the mutation was entered into a new database of genetic research, the Stanford team discovered a month-old research report. It identified the mutation, and the related disabilities. Even better, it described a drug regimen that has given similar patients some improvement.

Reading the one million medical research reports produced every year is impossible for diagnosing physicians. But advanced machine learning methods can develop computer systems that can read all the research and systematize it so it is accessible to doctors.

The new database is the product of work by Gill Bejerano, a Stanford polymath who is an associate professor in biology, pediatrics and computer science, and Christopher Ré, a Stanford computer scientist who developed an increasingly popular computer program called "DeepDive."

DeepDive can analyze unstructured text and use it to determine the answers to questions; it can then put those answers in a SQL relational database. For example, it can read thousands of newspaper articles about people and determine which ones are married to each other. It's the equivalent of turning a Google search that returns thousands of results into a credit-score inquiry that answers a specific question.

Reading the one million medical research reports produced every year is impossible for diagnosing physicians. But advanced machine learning methods can develop computer systems that can read all the research and systematize it so it is accessible to doctors.

Prof. Bejerano and Prof. Ré are using DeepDive to look for research articles that describe a connection between a patient's genotype and his or her phenotype, that is, the patient's medical symptoms. They are currently building a database that will be accessible to doctors all over the world. While genetic assays of patients aren't routine today, they are likely to become more common. Right now they cost about $10,000 and insurers are reluctant to cover the costs except as a last resort. But costs are dropping steadily and evidence that they can help diagnose mysterious problems is likely to increase demand.

More and more diseases turn out to have a genetic component—often one that can be alleviated by changes in diet or medication. When genomic assays are conducted, and reviewers identify an anomaly in the genetic code, they often don't know the significance. Sometimes the connection between the genotype and the patient's symptoms is unknown. Many times, researchers somewhere have identified a connection, but Stanford doctors don't know it because it's impossible to keep up with all the research.

Learning that a child has a genetic defect can be heartbreaking for parents. But most people want to know the reason for symptoms. It can help families plan for the future if the course of a disease is known. Often there are treatments that alleviate symptoms and this knowledge can help parents understand if a future pregnancy would be likely to result in another child with the condition.

Genetic medicine represents one of the brightest hopes for improving medical care. But currently, too many of the benefits are lost because of the sheer volume of research. Machine learning systems can classify genetic anomalies and the resulting phenotypes. The systems promise to lead to accurate diagnoses for many more patients.

INVESTIGATORS

Gill Bejerano, Associate Professor of Developmental Biology, Computer Science, and Pediatrics (Medical Genetics)

Christopher Ré, Assistant Professor of Computer Science

CORPORATE MEMBERS

We are pleased to acknowledge the generous support of our corporate members.

FOUNDING MEMBERS

AMERICAN FAMILY INSURANCE®

Google

Infosys®

REGULAR MEMBERS

htc

intel®

LIGHTSPEED VENTURE PARTNERS

Swiss Re

STANFORD DATA SCIENCE INITIATIVE (SDSI)

Professor Hector Garcia-Molina, Faculty Director

Dr. Steve Eglash, Executive Director

Kathy Menchaca, Program Manager

Stanford Data Science Initiative

FUNDED RESEARCH PROJECTS IN DATA SCIENCE

Thanks to everyone who contributed to this report.

CONCEPT CREATION
Zach Dammann

RESEARCH AND WRITING
Bill Bulkeley, www.billbulkeley.com
Lee Gomes

EDITORIAL ASSISTANCE
Kathy Menchaca and Marianne Siroker

GRAPHIC DESIGN
Alexander Atkins Design, Inc., www.alexatkinsdesign.com

Stanford | Stanford Data Science Initiative

sdsi.stanford.edu