Why “Big Data”

Data now stream from daily life: from phones and credit cards and televisions and computers; from the infrastructure of cities; from sensor-equipped buildings, trains, buses, planes, bridges, and factories. The data flow so fast that the total accumulation of the past two years—a zettabyte—dwarfs the prior record of human civilization. “There is a big data revolution,” says Weatherhead University Professor Gary King. But it is not the quantity of data that is revolutionary. “The big data revolution is that now we can do something with the data.”

The revolution lies in improved statistical and computational methods, not in the exponential growth of storage or even computational capacity, King explains. The doubling of computing power every 18 months (Moore’s Law) “is nothing compared to a big algorithm”—a set of rules that can be used to solve a problem a thousand times faster than conventional computational methods could. One colleague, faced with a mountain of data, figured out that he would need a $2-million computer to analyze it. Instead, King and his graduate students came up with an algorithm within two hours that would do the same thing in 20 minutes—on a laptop: a simple example, but illustrative.

New ways of linking datasets have played a large role in generating new insights. And creative approaches to visualizing data—humans are far better than computers at seeing patterns—frequently prove integral to the process of creating knowledge. Many of the tools now being developed can be used across disciplines as seemingly disparate as astronomy and medicine. Among students, there is a huge appetite for the new field. A Harvard course in data science last fall attracted 400 students, from the schools of law, business, government, design, and medicine, as well from the College, the School of Engineering and Applied Sciences (SEAS), and even MIT. Faculty members have taken note: the Harvard School of Public Health (HSPH) will introduce a new master’s program in computational biology and quantitative genetics next year, likely a precursor to a Ph.D. program. In SEAS, there is talk of organizing a master’s in data science.

“There is a movement of quantification rumbling across fields in academia and science, industry and government and nonprofits,” says King, who directs Harvard’s Institute for Quantitative Social Science (IQSS), a hub of expertise for interdisciplinary projects aimed at solving problems in human society. Among faculty colleagues, he reports, “Half the members of the government department are doing some type of data analysis, along with much of the sociology department and a good fraction of economics, more than half of the School of Public Health, and a lot in the Medical School.” Even law has been seized by the movement to empirical research—“which is social science,” he says. “It is hard to find an area that hasn’t been affected.”

The story follows a similar pattern in every field, King asserts. The leaders are qualitative experts in their field. Then a statistical researcher who doesn’t know the details of the field comes in and, using modern data analysis, adds tremendous insight and value. As an example, he describes how Kevin Quinn, formerly an assistant professor of government at Harvard, ran a contest comparing his statistical model to the qualitative judgments of 87 law professors to see which could best predict the outcome of all the Supreme Court cases in a year. “The law professors knew the jurisprudence and what each of the justices had decided in previous cases, they knew the case law and all the arguments,” King recalls. “Quinn and his collaborator, Andrew Martin [then an associate professor of political science at Washington University], collected six crude variables on a whole lot of previous cases and did an analysis.” King pauses a moment. “I think you know how this is going to end. It was no contest.” Whenever sufficient information can be quantified, modern statistical methods will outperform an individual or small group of people every time.

In marketing, familiar uses of big data include “recommendation engines” like those used by companies such as Netflix and Amazon to make purchase suggestions based on the prior interests of one customer as compared to millions of others. Target famously (or infamously) used an algorithm to detect when women were pregnant by tracking purchases of items such as unscented lotions—and

Information science promises to change the world.

by Jonathan Shaw

Is a Big Deal

Photographs by Stu Rosner

Reprinted from Harvard Magazine. For more information, contact Harvard Magazine, Inc. at 617-495-5746
offered special discounts and coupons to those valuable patrons. Credit-card companies have found unusual associations in the course of mining data to evaluate the risk of default: people who buy anti-scuff pads for their furniture, for example, are highly likely to make their payments.

In the public realm, there are all kinds of applications: allocating police resources by predicting where and when crimes are most likely to occur; finding associations between air quality and health; or using genomic analysis to speed the breeding of crops like rice for drought resistance. In more specialized research, to take one example, creating tools to analyze huge datasets in the biological sciences enabled associate professor of organismic and evolutionary biology Pardis Sabeti, studying the human genome’s billions of base pairs, to identify genes that rose to prominence quickly in the course of human evolution, deter-
mining traits such as the ability to digest cow’s milk, or resistance to diseases like malaria.

King himself recently developed a tool for analyzing social media texts. “There are now a billion social-media posts every two days...which represent the largest increase in the capacity of the human race to express itself at any time in the history of the world,” he says. No single person can make sense of what a billion other people are saying. But statistical methods developed by King and his students, who tested his tool on Chinese-language posts, now make that possible. (To learn what he accidentally uncovered about Chinese government censorship practices, see http://harvardmag.com/censorship.)

King also designed and implemented “what has been called the largest single experimental design to evaluate a social program in the world, ever,” reports Julio Frenk, dean of HSPH. “My entire career has been guided by the fundamental belief that scientifically derived evidence is the most powerful instrument we have to design enlightened policy and produce a positive social transformation,” says Frenk, who was at the time minister of health for Mexico. When he took office in 2000, more than half that nation’s health expenditures were being paid out of pocket—and each year, four million families were being ruined by catastrophic healthcare expenses. Frenk led a healthcare reform that created, implemented, and then evaluated a new public insurance scheme, Seguro Popular. A requirement to evaluate the program (which he says was projected to cost 1 percent of the GDP of the twelfth-largest economy in the world) was built into the law. So Frenk (with no inkling he would ever come to Harvard), hired “the top person in the world” to conduct the evaluation, Gary King.

Given the complications of running an experiment while the program was in progress, King had to invent new methods for analyzing it. Frenk calls it “great academic work. Seguro Popular has been studied and emulated in dozens of countries around the world thanks to a large extent to the fact that it had this very rigorous research with big data behind it.” King crafted “an incredibly original design,” Frenk explains. Because King compared communities that received public insurance in the first stage (the rollout lasted seven years) to demographically similar communities that didn’t, the results were “very strong,” Frenk says: any observed effect would be attributable to the program. After just 10 months, King’s study showed that Seguro Popular successfully protected families from catastrophic expenditures due to serious illness, and his work provided guidance for needed improvements, such as public outreach to promote the use of preventive care.

King himself says big data’s potential benefits to society go far beyond what has been accomplished so far. Google has analyzed clusters of search terms by region in the United States to predict flu outbreaks faster than was possible using hospital admission records. “That was a nice demonstration project,” says King, “but it is a tiny fraction of what could be done” if it were possible for...
Nathan Eagle, an adjunct assistant professor at HSPH, was bridges to business public health and medicine, fields in which, King says, “People are literally dying every day” simply because data are not being shared.

Bridges to Business

Nathan Eagle, an adjunct assistant professor at HSPH, was one of the first people to mine unstructured data from businesses with an eye to improving public health in the world’s poorest nations. A self-described engineer and “not much of an academic” (despite having held professorships at numerous institutions including MIT), much of his work has focused on innovative uses of cell-phone data. Drawn by the explosive growth of the mobile market in Africa, he moved in 2007 to a rural village on the Kenyan coast and began searching for ways to improve the lives of the people there. Within months, realizing that he would be more effective sharing his skills with others, he began teaching mobile-application development to students in the University of Nairobi’s computer-science department.

While there, he began working with the Kenyan ministry of health on a blood-bank monitoring system. The plan was to recruit nurses across the country to text the current blood-supply levels in their local hospitals to a central database. “We built this beautiful visualization to let the guys at the centralized blood banks in Kenya see in real time what the blood levels were in these rural hospitals,” he explains, “and more importantly, where the blood was needed.” In the first week, it was a giant success, as the nurses texted in the data and central monitors logged in every hour to see where they should replenish the blood supply. “But in the second week, half the nurses stopped texting in the data, and within about a month virtually no nurses were participating anymore.”

Eagle shares this tale of failure because the episode was a valuable learning experience. “The technical implementation was bulletproof,” he says. “It failed because of a fundamental lack of insight on my part...that had to do with the price of a text message. What I failed to appreciate was that an SMS represents a fairly substantial on my part…that had to do with the price of a text message. What I failed to appreciate was that an SMS represents a fairly substantial addition of a simple script let him credit the rural nurses with a small denomination of prepaid air time, about 10 cents’ worth—

Eagle’s next project, based in Rwanda, was more ambitious, and it also provided a lesson in one of the pitfalls of working with big data: that it is possible to find correlations in very large linked datasets without understanding causation. Working with mobile-phone records (which include the time and location of every call), he began creating models of people’s daily and weekly commuting patterns, termed their “radius of generation.” He began to notice patterns. Abruptly, people in a particular village would stop moving as much; he hypothesized that these patterns might indicate the onset of a communicable disease like the flu. Working with the Rwandan ministry of health, he compared the data on cholera outbreaks to his radius of generation data. Once linked, the two datasets proved startlingly powerful; the radius of generation in a village dropped two full weeks before a cholera outbreak. “We could even predict the magnitude of the outbreak based on the amount of decrease in the radius of generation,” he recalls. “I had built something that was performing in this unbelievable way.”

And in fact it was unbelievable. He tells this story as a “good example of why engineers like myself shouldn’t be doing epidemiology in isolation—and why I ended up joining the School of Public Health rather than staying within a physical-science department.” The model was not predicting cholera outbreaks, but pinpointing floods. “When a village floods and roads wash away, suddenly the radius of generation decreases,” he explains. “And it also makes the village more susceptible in the short term to a cholera outbreak. Ultimately, all this analysis with supercomputers was identifying where there was flooding—data that, frankly, you can get in a lot of other ways.”

Despite this setback, Eagle saw what was missing. If he could couple the data he had from the ministry of health and the mobile operators with on-the-ground reports of what was happening, then he would have a powerful tool for remote disease surveillance. “It opened my eyes to the fact that big data alone can’t solve this type of problem. We had petabytes* of data and yet we were building models that were fundamentally flawed because we didn’t have real insight about what was happening” in remote villages. Eagle has now built a platform that enables him to survey individuals in such countries by paying them small denominations of airtime (as with the Kenyan nurses) in exchange for answering questions: are they experiencing flu-like symptoms, sleeping under bednets, or taking anti-malarials? This ability to gather and link self-reported information to larger datasets has proven a powerful tool—and the survey technology has become a successful commercial entity named Jana, of which Eagle is co-founder and CEO.

New Paradigms—and Pitfalls

Willy Shih, Cizik professor of management practice at Harvard Business School, says that one of the most important changes wrought by big data is that their use involves a “fundamentally

* A petabyte is the equivalent of 1,000 terabytes, or a quadrillion bytes. One terabyte is a thousand gigabytes. One gigabyte is made up of a thousand megabytes. There are a thousand thousand—or, a million—petabytes in a zettabyte.

Reprinted from Harvard Magazine. For more information, contact Harvard Magazine, Inc. at 617-495-5746.
different way of doing experimental design." Historically, social
scientists would plan an experiment, decide what data to collect,
and analyze the data. Now the low cost of storage (“The price of
storing a bit of information has dropped 60 percent a year for six
decades,” says Shih) has caused a rethinking, as people “collect
everything and then search for significant patterns in the data.”

“This approach has risks,” Shih points out. One of the most
prominent is data dredging, which involves searching for patterns
in huge datasets. A traditional social-science study might assert
that the results are significant with 95 percent confidence. That
means, Shih points out, “that in one out of 20 instances” when
dredging for results, “you will get results that are statistically
significant purely by chance. So you have to remember that.” Al-
though this is true for any statistical finding, the enormous number
of potential correlations in very large datasets substantially mag-
nifies the risk of finding spurious correlations.

Eagle agrees that “you don’t get good scientific output from
throwing everything against the wall and seeing what sticks.” No
matter how much data exists, researchers still need to ask
the right questions to create a hypothesis, design a test, and use
the data to determine whether that hypothesis is true. He sees
two looming challenges in data science. First, there aren’t enough
people comfortable dealing with petabytes of data. “These skill
sets need to get out of the computer-science departments and
into public health, social science, and public policy,” he says.
“Big data is having a transformative impact across virtually all
academic disciplines—it is time for data science to be integrated
into the foundational courses for all undergraduates.”

Safeguarding data is his other major concern, because “the privacy implica-
tions are profound.” Typically, the owners of huge datasets are very nervous about
sharing even anonymized, population-level information like the cell records
Eagle uses. For the companies that hold it, he says, “There is a lot of downside to
making this data open to researchers. We need to figure out ways to mitigate that
concern and craft data-usage policies in ways that make these large organizations
more comfortable with sharing these data, which ultimately could improve the lives of the millions of people who are
generating it—and the societies in which they are living.”

John Quackenbush, an HSPH professor of computational biology and bioin-
formatics, shares Eagle’s twin concerns. But in some realms of biomedical big
data, he says, the privacy problem is not easily addressed. “As soon as you touch
genomic data, that information is fundamentally identifiable,” he explains. “I can
erase your address and Social Security number and every other identifier, but I can’t anonymize your ge-
nome without wiping out the information that I need to analyze.” Privacy in such cases is achieved not through anonymity but by
consent paired with data security: granting access only to autho-
ized researchers. Quackenbush is currently collaborating with a
dozen investigators—from HSPH, the Dana-Farber Cancer Insti-
tute, and a group from MIT’s Lincoln Labs expert in security—to
develop methods to address a wide range of biomedical research
problems using big data, including privacy.

He is also leading the development of HSPH’s new master’s
program in computational biology and quantitative genetics, which is designed to address the extraordinary complexity of
biomedical data. As Quackenbush puts it, “You are not just you.
You have all this associated health and exposure information that
I need in order to interpret your genomic information.”

A primary goal, therefore, is to give students practical skills in
the collection, management, analysis, and interpretation of genom-
ic data in the context of all this other health information: electronic
medical records, public-health records, Medicare information,
and comprehensive-disease data. The program is a joint venture be-
tween biostatistics and the department of epidemiology.

**Really Big Data**

LIKE EAGLE, Quackenbush came to public health from another
discipline—in his case, theoretical and high-energy experimen-
tal physics. He first began working outside his doctoral field in
1992, when biologists for the Human Genome Project realized
they needed people accustomed to collecting, analyzing, manag-
ing, and interpreting huge datasets. Physicists have been good at
that for a long time.

The first full human genome sequence took five to 15 years
to complete, and cost $1 billion to $3 billion (“Depending on whom
you ask,” notes Quackenbush). By 2009, eight years later, the
cost had dropped to $100,000 and took a year. At that
point, says Quackenbush, “if my wife had a rare, difficult
cancer, I would have mortgaged our house to sequence
her genome.” Now a genome sequence takes a little more
than 24 hours and costs about $1,000—the point at which
it can be paid for “on a credit card. That simple statement
alone,” he says, “underscores why the biomedical scienc-
es have become so data-driven.

“We each carry two copies of the human genome—
one from our mother and one from our father—that
together comprise 6 billion base pairs,” Quackenbush
continues, “a number equivalent to all the seconds in
190 years.” But knowledge of what all the genes en-
coded in the genome do and how they interact to influ-
ence health and disease remains woefully incomplete.
To discover that, scientists will have to take genomic
data and “put it in the context of your health. And
we’ll have to take you and put you in the context of the
population in which you live, the environmental factors
you are exposed to, and the people you come in contact
with—as so as we look at the vast amount of data we can
generate on you, the only way we can effectively inter-
pret it is to put it in the context of the vast amount of
data we can generate on almost everything related to
you, your environment, and your health. We are moving
from a big data problem to a really big data problem.”

Curtis Huttenhower, an HSPH associate professor of com-
putational biology and bioinformatics, is one of Quackenbush’s
really big data collaborators. He studies the function of the hu-
man microbiome, the bacteria that live in and on humans, principally in the gut, helping people extract energy from food and maintaining health. “There are 100 times more genes in the bugs than in a human's genome,” he reports, and “it’s not unusual for someone to share 50 percent or less of their microbes with other people. Because no one has precisely the same combination of gut bacteria, researchers are still learning how those bacteria distinguish us from each other; meanwhile both human and microbial genetic privacy must be maintained.” Not only do microbiome studies confront 100 times more information per human subject than genome studies, that 100 is different from person to person and changes slowly over time with age—and rapidly, as well, in response to factors like diet or antibiotics. Deep sequencing of 100 people during the human microbiome project, Huttenhower reports, yielded a thousand human genomes’ worth of sequencing data—and we could have gotten more. But there is still no comprehensive catalog of what affects the microbiome,” says Huttenhower. “We are still learning.”

Recently, he has been studying microbes in the built environment: from the handstraps of Boston’s transit system to touch-screen machines and human skin. The Sloan Foundation, which funded the project, wants to know what microbes are there and how they got there. Huttenhower is interested in the dynamics of how entire communities of bugs are transferred from one person to another and at what speed. “Everyone tends to have a slightly different version of Helicobacter pylori, a bacterium that can cause gastric cancer and is transmitted vertically from parents to children,” he says. “But what other portions of the microbiome are mostly inherited, rather than acquired from our surroundings? We don’t know yet.” As researchers learn more about how the human genome and the microbiome interact, it might become possible to administer probiotics or more targeted antibiotics to treat or prevent disease. That would represent a tremendous advance in clinical practice because right now, when someone takes a broad spectrum antibiotic, it is “like setting off a nuke,” say Huttenhower. “They instantly change the shape of the microbiome for a few weeks to months.” Exactly how the microbiome recovers is not known.

A major question in microbiome studies involves the dynamics of coevolution: how the bugs evolved in humans over hundreds of thousands of years, and whether changes in the microbiome might be linked to ailments that have become more prevalent recently, such as irritable bowel disease, allergies, and metabolic syndrome (a precursor to diabetes). Because of the timescale of the change in the patterns of these ailments, the causes can’t be genomic, says Huttenhower. “They could be environmental, but the timescale is also right for the kinds of ecological
WHY “BIG DATA” IS A BIG DEAL
(continued from page 35)

cannot be gleaned from connecting
gene sequences, health records, and envi-
ronmental influences? And how can hu-
mans understand the results?

One of the most powerful tools for fa-
ilitating understanding of vast datasets
is visualization. Hanspeter Pfister, Wang
professor of computer science and director
of the Institute for Applied Computational
Science, works with scientists in genomics
and systems biology to help them visual-
ize what are called high-dimensional data
sets (with multiple categories of data being
compared). For example, members of his
research group have created a visualization for use
by oncologists that connects gene sequence
and activation data with cancer types and
stages, treatments, and clinical outcomes.

That allows the data to be viewed in a way
that shows which particular gene expres-
sion pattern is associated with high mortal-
ity regardless of cancer type, for example,
giving an important, actionable insight
for how to devise new treatments.

Pfister teaches students how to turn
data into visualizations in Computer Scien-
tce 109, “Data Science,” which he co-
teaches with Joseph K. Blitzstein, profes-
sor of the practice in statistics. “It is very
important to make sure that what we will
be presenting to the user is understand-
able, which means we cannot show it
all,” says Pfister. “Ag-
eggregation, filtering,
and clustering are
hugely important
for us to reduce the
data so that it makes
sense for a person
to look at.” This is a
different method of sci-
entific inquiry that
ultimately aims to
create systems that
let humans combine
what they are good
at—asking the right
questions and inter-
perting the results—
with what machines
are good at: compu-
tation, analysis, and statistics using large
datasets. Student projects have run the
gamut from the evolution of the American
presidency and the distribution of tweets
for competitive product analysis, to pre-
dicting the stock market and analyzing the
performance of NHL hockey teams.

Pfister’s advanced students and post-
doctoral fellows work with scientists who
lack the data science skills they now need
to conduct their research. “Every collabo-
ration pushes us into some new, unknown
territory in computer science,” he says.

The flip side of Pfister’s work in creat-
ing visualizations is the automated analy-
sis of images. For example, he works with
Knowles professor of molecular and cellular
biology Jeff Lichtman, who is also Ramon
y Cajal professor of arts and sciences, to
reconstruct and visualize neural connec-
tions in the brain. Lichtman and his team
slide brain tissue very thinly, providing Pfis-
ter’s group with stacks of high-resolution
images. “Our system then automatically
identifies individual cells and labels them
consistently,” such that each neuron can be
traced through a three-dimensional stack of
images, Pfister reports. Even working with
only a few hundred neurons involves tens of
thousands of connections. One cubic milli-
meter of mouse brain represents a thousand
terabytes (a petabyte) of image data.

Pfister has also worked with radioas-
tronomers. The head teaching fellow in
his data science course, astronomer Chris
Beaumont, has developed software (Glue)
for linking and visualizing large telescope
datasets. Beaumont’s former doctoral ad-
iser (for whom he now works as senior
software developer on Glue), professor
of astronomy Alyssa Goodman, teaches
her own course in vi-
sualization (Empiri-
cal and Mathematical
Reasoning 19, “The
Art of Numbers”).

Goodman uses visual-
ization as an explor-
atory technique in her
efforts to understand
interstellar gas—the
stuff of which stars
are born. “The data
volume is not a con-
cern,” she says; even though a big telescope
can capture a petabyte of data in a night,
astronomers have a long history of deal-
ing with large quantities of data. The trick,
she says, is making sense of it all. Data vi-
sualizations can lead to new insights, she
says, because “humans are much better at
pattern recognition” than computers. In
a recent presentation, she showed how a

Changes in the
microbiome might be
linked to ailments
that have become
more prevalent
recently, such as
irritable bowel
disease, allergies,
and metabolic
syndrome (a precur-
sor to diabetes).

Discerning Patterns in Complexity

Making sense of the relationships be-
tween distinct kinds of information is an-
other challenge facing researchers. What
insights can be gleaned from connecting

Changes in the
microbiome might be
linked to ailments
that have become
more prevalent
recently, such as
irritable bowel
disease, allergies,
and metabolic
syndrome (a precur-
sor to diabetes).

Discerning Patterns in Complexity

Making sense of the relationships be-
tween distinct kinds of information is an-
other challenge facing researchers. What
insights can be gleaned from connecting

Changes in the
microbiome might be
linked to ailments
that have become
more prevalent
recently, such as
irritable bowel
disease, allergies,
and metabolic
syndrome (a precur-
sor to diabetes).

Discerning Patterns in Complexity

Making sense of the relationships be-
tween distinct kinds of information is an-
other challenge facing researchers. What
insights can be gleaned from connecting
Physicians were able to locate and successfully diagnose arterial blockages only 39 percent of the time. Using Borkin's novel visualization—akin to a linear side-view of the patient's arteries—improved the rate of successful diagnosis to 62 percent. Then, simply by changing the colors based on an understanding of the way the human visual cortex works, Borkin found she could raise the rate of successful diagnosis to 91 percent.

Visualization tools even have application in the study of collections, says Pfister. Professor of romance language and literatures Jeffrey Schnapp, faculty director of Harvard's metaLAB, is currently at work on a system for translating collections metadata into readily comprehensible, information-rich visualizations. Starting with a dataset of 17,000 photographs—trivial by big data standards—from the missing paintings of the Italian Renaissance collection assembled by Bernard Berenson (works that were photographed but have subsequently disappeared), Schnapp and colleagues have created a way to explore the collection by means of the existing descriptions of objects, classifications, provenance data, media, materials, and subject tags.

The traditional use of such inventory data was to locate and track individual objects, he continues. “We are instead creating a platform that you can use to make arguments, and to study collections as aggregates from multiple angles. I can't look at everything in the Fogg Museum's collections even if I am Tom Lentz [Cabot director of the Harvard Art Museums], because there are 250,000 objects. Even if I could assemble them all in a single room,” Schnapp says, “I couldn't possibly see them all.” But with a well-structured dataset, “We can tell stories: about place, time, distribution of media, shifting themes through history and on and on.” In the case of the Berenson photo collection, one might ask, “What sorts of stories does the collection tell us about the market for Renaissance paintings during Berenson's lifetime? Where are the originals now? Do they still exist? Who took the photographs and why? How did the photo formats evolve with progress in photographic techniques?”

This type of little “big data” project makes the incomprehensible navigable and potentially understandable. “Finding imaginative, innovative solutions for creating qualitative experiences of collections is the key to making them count,” Schnapp says. Millions of photographs in the collections of institutions such as the Smithsonian, for example, will probably never be catalogued, even though they represent the richest, most complete record of life in America. It might take an archivist half a day just to research a single one, Schnapp points out. But the photographs are being digitized, and as they come on line, ordinary citizens with local information and experience can contribute to making them intelligible in ways that add value to the collection as an aggregate. The Berenson photographs are mostly of secondary works of art, and therefore not necessarily as interesting individually as they are as a collection. They perhaps tell stories about how works were produced in studios, or how they circulated. Visualizations of the collection grouped by subject are telling, if not surprising. Jesus represents the largest portion, then Mary, and so on down to tiny outliers, such as a portrait of a woman holding a book, that raise rich questions for the humanities, even though a computer scientist might regard them as problems to fix. “We’re on the culture side of the divide,” Schnapp says, “so we sometimes view big data from a slightly different angle, in that we are interested in the ability to zoom between the micro level of analysis (an individual object), the macro level (a collection), and the massively macro (multiple collections) to see what new knowledge allows you to expose, and the stories it lets you tell.”

Data, in the final analysis, are evidence. The forward edge of science, whether it drives a business or marketing decision, provides an insight into Renaissance painting, or leads to a medical breakthrough, is increasingly being driven by quantities of information that humans can understand only with the help of math and machines. Those who possess the skills to parse this ever-growing trove of information sense that they are making history in many realms of inquiry. “The data themselves, unless they are actionable, aren't relevant or interesting,” is Nathan Eagle’s view. “What is interesting,” he says, “is what we can now do with them to make people’s lives better.” John Quackenbush says simply: “From Copernicus using Tycho Brahe’s data to build a heliocentric model of the solar system, to the birth of statistical quantum mechanics, to Darwin's theory of evolution, to the modern theory of the gene, every major scientific revolution has been driven by one thing, and that is data.”