1 TERABYTE

A $200 HARD DRIVE THAT HOLDS 260,000 SONGS

20 TERABYTES

PHOTOS UPLOADED TO FACEBOOK EACH MONTH

THE PETA AGE

460 TERABYTES

ALL THE DIGITAL WEATHER DATA COMPILED BY THE NATIONAL CLIMATIC DATA CENTER

530 TERABYTES

ALL THE VIDEOS ON YOUTUBE

120 TERABYTES

ALL THE DATA AND IMAGES COLLECTED BY THE HUBBLE SPACE TELESCOPE

330 TERABYTES

DATA THAT THE LARGE HADRON COLLIDER WILL PRODUCE EACH WEEK

Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

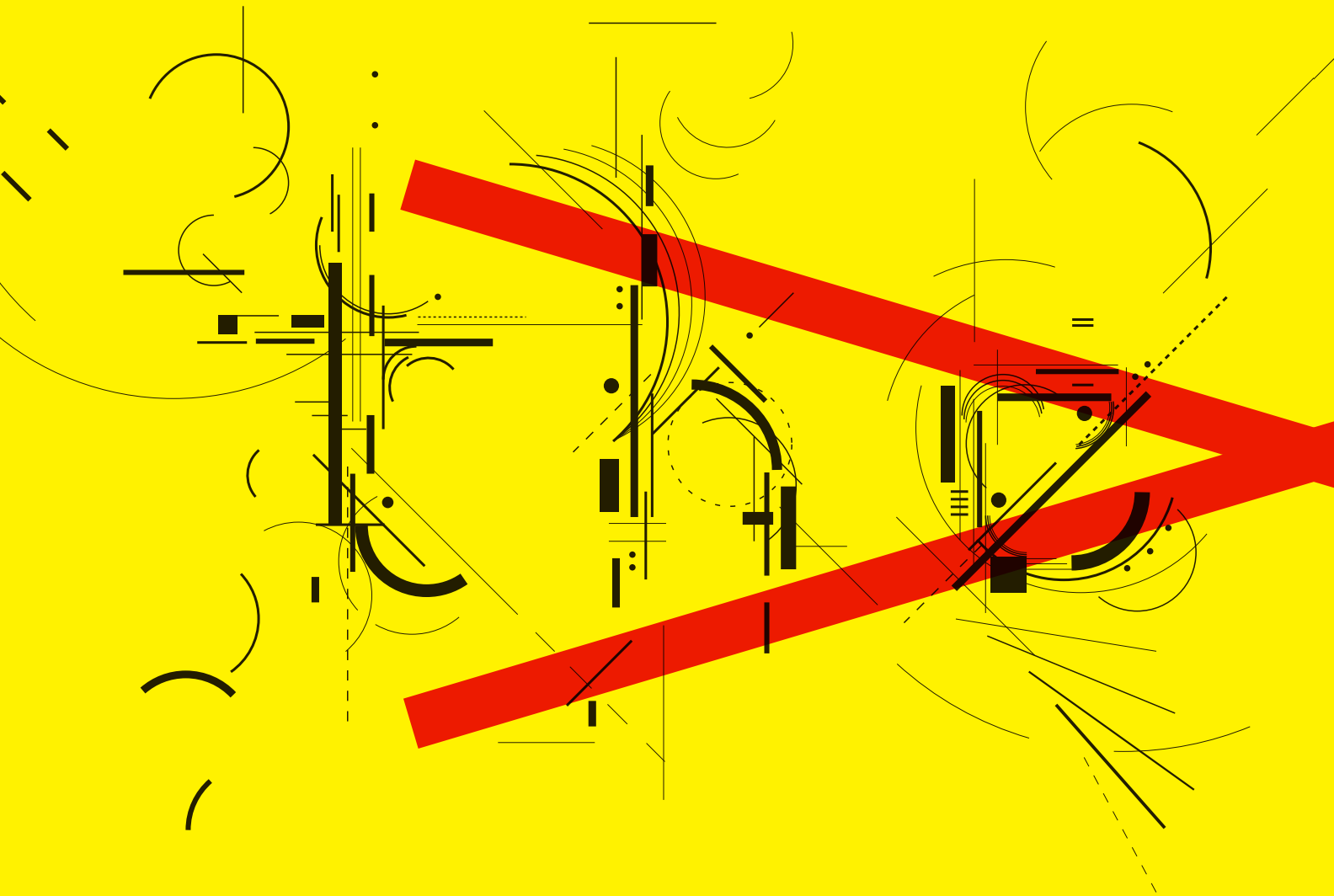
600 TERABYTES

ANCESTRY.COM'S GENEALOGY DATABASE (INCLUDES ALL U.S. CENSUS RECORDS 1790–2000)

1 PETABYTE

DATA PROCESSED BY GOOGLE'S SERVERS EVERY 72 MINUTES

BYTE



THE END OF THEORY

SCIENTISTS HAVE ALWAYS RELIED ON HYPOTHESIS AND EXPERIMENTATION. NOW, IN THE ERA OF MASSIVE DATA, THERE'S A BETTER WAY.

BY CHRIS ANDERSON

"ALL MODELS ARE WRONG, BUT some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age.

The Petabyte Age is different because more is different. Kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to—well, at petabytes we ran out of organizational analogies.

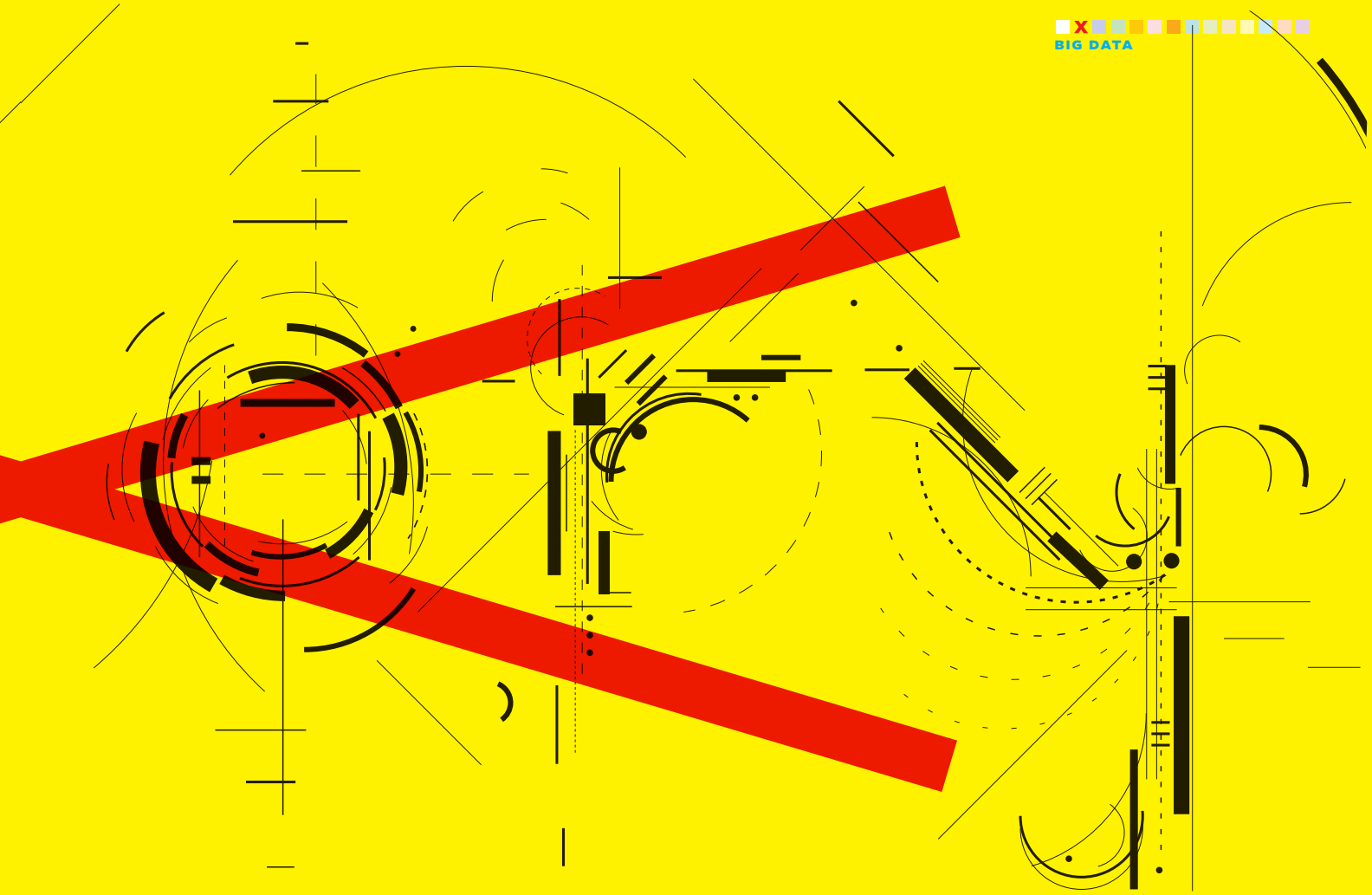
At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising—it just assumed that better data, with better analytical tools, would win the day. And Google was right.

Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them (given equal corpus data, Google can translate Klingon into Farsi as easily as it can translate French into German). And why it can match ads to content without any knowledge or assumptions about the ads or the content.

Speaking at the O'Reilly Emerging Technology Conference this past March, Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."

This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

The big target here isn't advertising, though. It's science. The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm



or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years.

Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

But faced with massive data, this approach to science—hypothesize, model, test—is becoming obsolete. Consider physics: Newtonian models were crude approximations of the truth (wrong at the atomic level, but still useful). A hundred years ago, statistically based quantum mechanics offered a better picture—but quantum mechanics is yet another model, and as such it, too, is flawed, no doubt a caricature of a more complex underlying reality. The reason physics has drifted into theoretical speculation about *n*-dimensional grand unified models over the past few decades (the "beautiful story" phase of a discipline starved of data) is that we don't know how to run the experiments that would falsify the hypotheses—the energies are too high, the accelerators too expensive, and so on.

Now biology is heading in the same direction. The models we were taught in school about "dominant" and "recessive" genes steering a strictly Mendelian process have turned out to be an even greater simplification of reality than Newton's laws. The discovery of gene-protein interactions and other aspects of epigenetics has challenged the view of DNA as destiny and even introduced evidence that environment can influence inheritable traits, something once considered a genetic impossibility.

In short, the more we learn about biology, the further we find ourselves from a model that can explain it.

There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

The best practical example of this is the shotgun gene sequencing by J. Craig Venter. Enabled by high-speed sequencers and supercomputers that statistically analyze the data they produce, Venter went from sequencing individual organisms to sequencing entire ecosystems. In 2003, he started sequencing much of the ocean, retracing the voyage of Captain Cook. And in 2005 he started sequencing the air. In the process, he discovered thousands of previously unknown species of bacteria and other life-forms.

If the words "discover a new species" call to mind Darwin and drawings of finches, you may be stuck in the old way of doing science. Venter can tell you almost nothing about the species he found. He doesn't know what they look like, how they live, or much of anything else about their morphology. He doesn't even have their entire genome. All he has is a statistical blip—a unique sequence that, being unlike any other sequence in the database, must represent a new species.

This sequence may correlate with other sequences that resemble those of species we do know more about. In that case, Venter can make some guesses about the animals—that they convert sunlight into energy in a particular way, or that they descended from a common ancestor. But besides that, he has no better model of this species than Google has of your MySpace page. It's just data. By analyzing it with Google-quality computing resources, though, Venter has advanced biology more than anyone else of his generation.

This kind of thinking is poised to go mainstream. In February, the National Science Foundation announced the Cluster Exploratory, a program that funds research designed to run on a large-scale distributed computing platform developed by Google and IBM in conjunction with six pilot universities. The cluster will consist of 1,600 processors, several terabytes of memory, and hundreds of terabytes of storage, along with the software, including Google File System, IBM's Tivoli, and an open source version of Google's MapReduce. Early CluE projects will include simulations of the brain and the nervous system and other biological research that lies somewhere between wetware and software.

Learning to use a "computer" of this scale may be challenging. But the opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

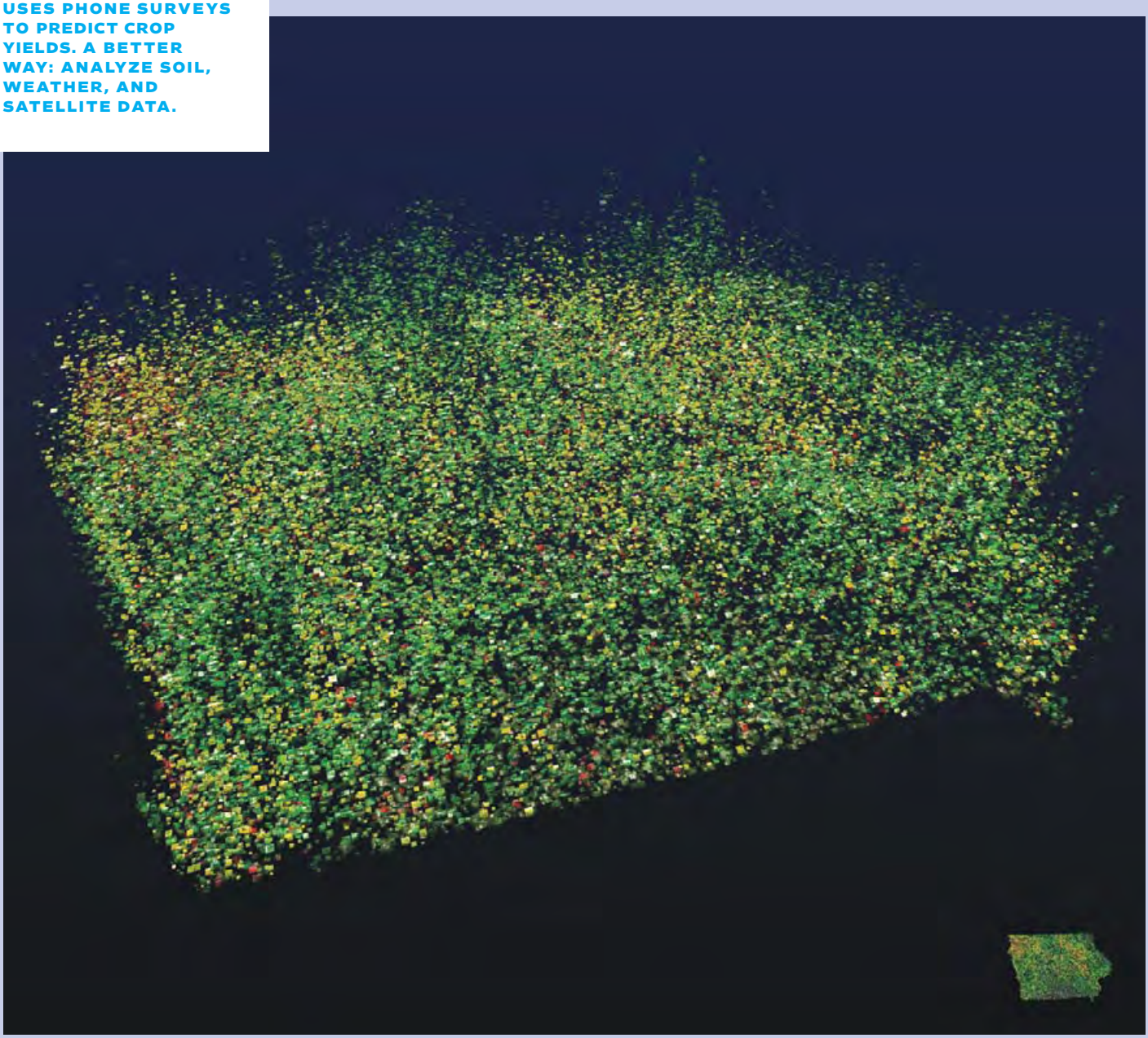
There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?

CHRIS ANDERSON (canderson@wired.com) *is the editor in chief of* WIRED.

TYPE BY MARIAN BANTJES

FEEDING THE MASSES

THE GOVERNMENT USES PHONE SURVEYS TO PREDICT CROP YIELDS. A BETTER WAY: ANALYZE SOIL, WEATHER, AND SATELLITE DATA.



DATA: LANWORTH

THE IOWA AGRICULTURAL LANDSCAPE: GREEN AREAS ARE MORE PRODUCTIVE FOR SOY, CORN, AND WHEAT; RED ARE LEAST.

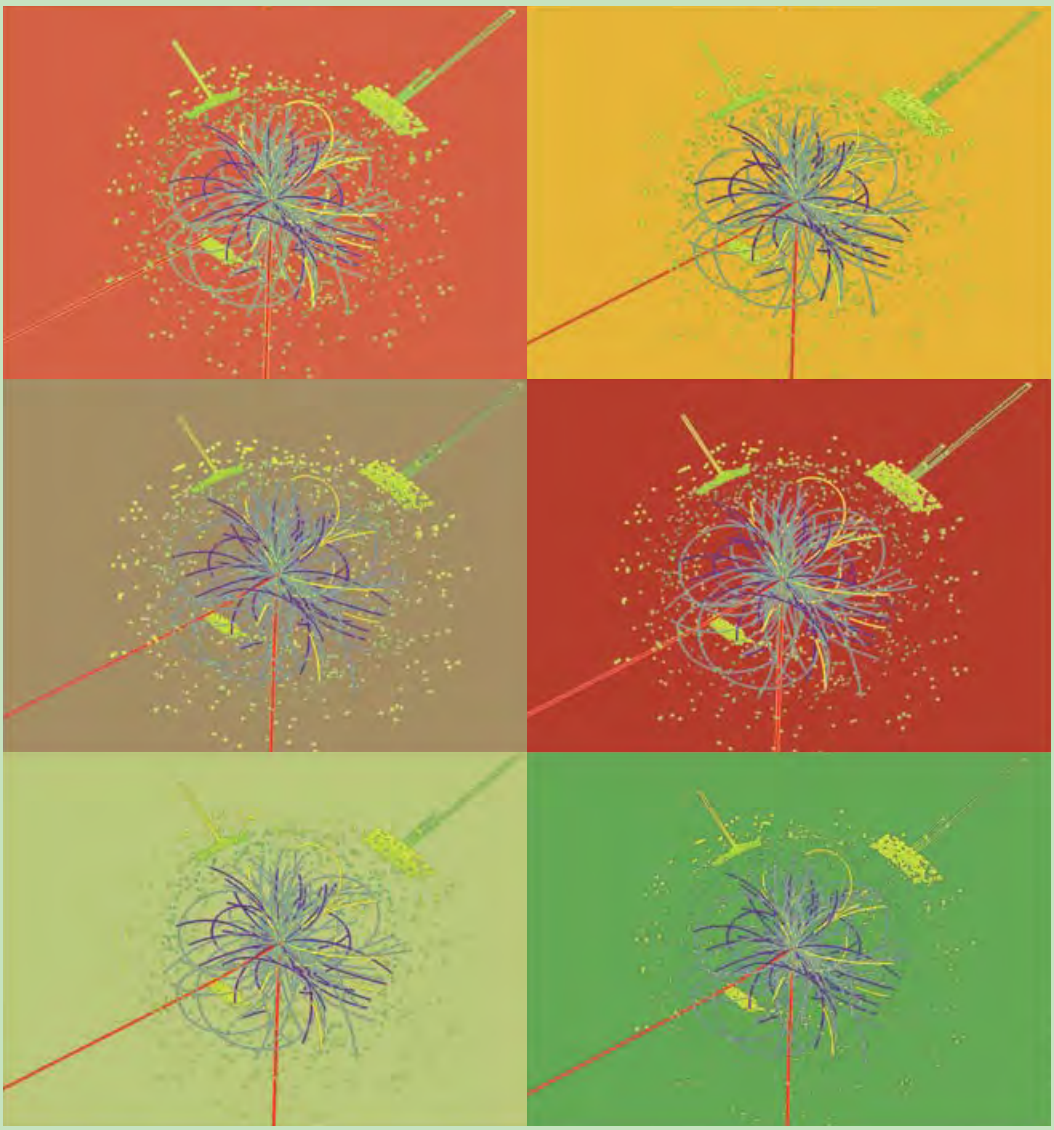
FARMER'S ALMANAC IS FINALLY obsolete. Last October, agricultural consultancy Lanworth not only correctly projected that the US Department of Agriculture had overestimated the nation's corn crop, it nailed the margin: roughly 200 million bushels. That's just 1.5 percent fewer kernels but still a significant shortfall for tight markets, causing a 13 percent price hike and jitters in the emerging ethanol industry. When the USDA downgraded expectations a month after Lanworth's prediction, the little Illinois-based company was hailed as a new oracle among soft-commodity traders—who now pay the firm more than $100,000 a year for a timely heads-up on fluctuations in wheat, corn, and soybean supplies.

The USDA bases its estimates on questionnaires and surveys—the agency calls a sample of farmers and asks what's what. Lanworth uses satellite images, digital soil maps, and weather forecasts to project harvests at the scale of individual fields. It even looks at crop conditions and rotation patterns—combining all the numbers to determine future yields.

Founded in 2000, Lanworth started by mapping forests for land managers and timber interests. Tracking trends in sleepy woodlands required just a few outer-space snapshots a year. But food crops are a fast-moving target. Now the company sorts 100 gigs of intel every day, adding to a database of 50 terabytes and counting. It's also moving into world production-prediction—wheat fields in Russia, Kazakhstan, and Ukraine are already in the data set, as are corn and soy plots in Brazil and Argentina. The firm expects to reach petabyte scale in five years. "There are questions about how big the total human food supply is and whether we as a country are exposed to risk," says Lanworth's director of information services, Nick Kouchoukos. "We're going after the global balance sheet."

BY BEN PAYNTER

VISUALIZATION BY FIRSTBORN



CERN

THE LARGE HADRON COLLIDER MIGHT FIND PARTICLES LIKE THE HIGGS BOSON—SHOWN HERE AS A SIMULATION.

CHASING THE QUARK

TO MAKE INFORMATION USABLE, THROW SOME OF IT AWAY.

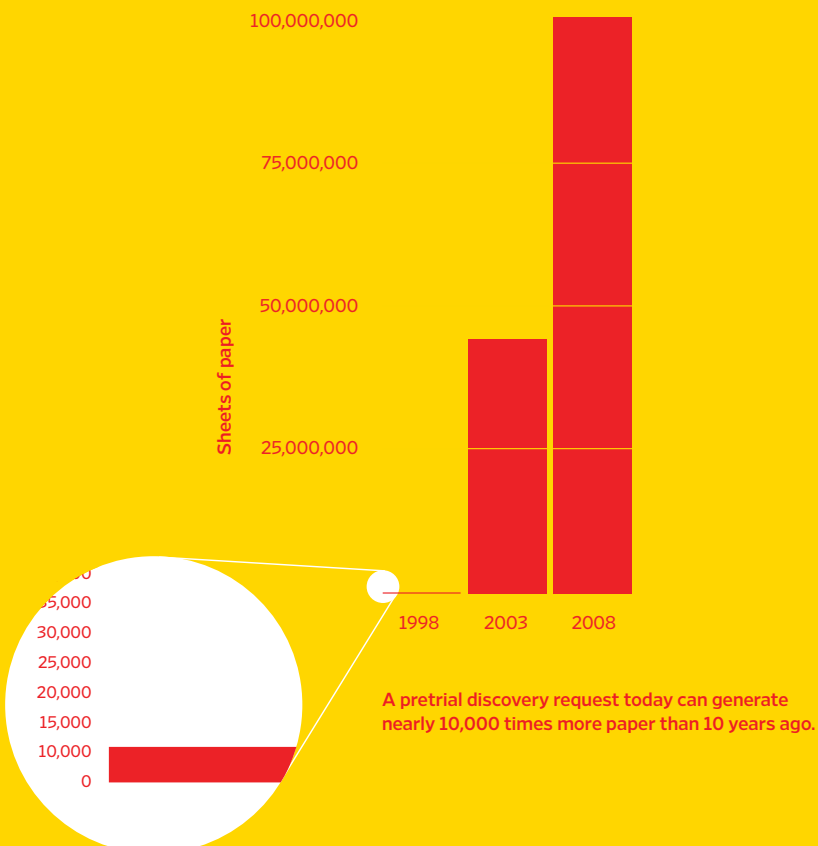
THE ULTIMATE DIGITAL CAMERA will be demo'd at the Large Hadron Collider near Geneva later this year. While proton beams race in opposite directions around a 17-mile underground ring, crossing and recrossing the Swiss-French border on each circuit, six particle detectors will snap a billion "photos" per second of the resulting impacts. The ephemeral debris from those collisions may hold answers to some of the most exciting questions in physics.

The LHC, expected to run 24/7 for most of the year, will generate about 10 petabytes of data per second. That staggering flood of information would instantly overwhelm any conceivable storage technology, so hardware and software filters will reduce the take to roughly 100 events per second that seem most promising for analysis. Even so, the collider will record about 15 petabytes of data each year, the equivalent of 15,000 terabyte-size hard drives filled to the brim. Hidden in all those 1s and 0s might be extra dimensions of space, the mysterious missing dark matter, or a whole new world of exotic superparticles.

BY DAVID HARRIS

WINNING THE LAWSUIT

LAWYERS NOW HIRE PROFESSIONAL DATA-MINERS TO DIG FOR THE DIRT ON CORPORATE AMERICA'S HARD DRIVES.



A pretrial discovery request today can generate nearly 10,000 times more paper than 10 years ago.

SOURCES: FIOS, INC., ATTENEX

WAY BACK IN THE 20TH CENTURY, when Ford Motor Company was sued over a faulty ignition switch, its lawyers would gird for the discovery process: a labor-intensive ordeal that involved disgorging thousands of pages of company records. These days, the number of pages commonly involved in commercial litigation discovery has ballooned into the billions. Attorneys on the hunt for a smoking gun now want to see not just the final engineering plans but the emails, drafts, personal data files, and everything else ever produced in the lead-up to the finished product.

Welcome to e-discovery. Firms like Fios, Attenex, and hundreds of others now specialize in the scanning, indexing, and data-mining of discovery documents. (The industry got a boost in December 2006, when new federal rules went into effect requiring parties to produce discovery documents in electronic format.) E-discovery vendors pulled in $2 billion in 2006—and that figure is expected to double by 2009.

So how has this evidentiary deluge changed the practice of law? Consider that five years ago, newly minted corporate litigators spent much of their time digging through warehouses full of paper documents. Today they're back at their desks, sorting through PDFs, emails, and memos on their double monitors—aided by semantic search technologies that scan for keywords and phrases. In another five years, don't be surprised to find juries chuckling over a plaintiff's incriminating IMs, voice messages, video conferences, and Twitters.

BY JOHN BRINGARDNER

CHART BY BOB DINETZ

SMALL OUTBREAKS OF VIOLENCE, LIKE RECENT FOOD RIOTS IN HAITI, CAN PREFIGURE A LARGER CRISIS.

AP

TRACKING THE NEWS

BY MONITORING ONLINE NEWS FEEDS, GOVERNMENTS CAN PREDICT VIOLENCE AND SPOT DISASTERS.

WHETHER NEWS OF CURRENT EVENTS is good or bad, there is always a lot of it. Worldwide, an estimated 18,000 Web sites publish breaking stories in at least 40 languages. That universe of information contains early warnings about everything from natural disasters to political unrest—if you can read the data.

When the European Commission asked its researchers to come up with a way to monitor news feeds in 2002, all it really wanted was to see what the press was saying about the EU. The commission's Joint Research Center developed software that monitors 1,540 Web sites running some 40,000 articles a day. There's no database per se, just about 10 gigabytes of information flowing past a pattern-matching algorithm every day—3.5 terabytes a year. When the system, called Europe Media Monitor, evolves to include online video, the daily dose of information could be measured in terabytes.

So what patterns does EMM find? Besides sending SMS and email news alerts to eurocrats and regular people alike, EMM counts the number of stories on a given topic and looks for the names of people and places to create geotagged "clusters" for given events, like food riots in Haiti or political unrest in Zimbabwe. Burgeoning clusters and increasing numbers of stories indicate a topic of growing importance or severity. Right now EMM looks for plain old violence; project manager Erik van der Goot is tweaking the software to pick up natural and humanitarian disasters, too. "That has crisis-room applications, where you have a bunch of people trying to monitor a situation," Van der Goot says. "We map a cluster of news reports on a screen in the front of the room—they love that."

EMM gives snapshots of the now. But "the big thing everyone would like to do is early warning of conflict and state failure," says Clive Best, a physicist formerly with the JRC. Other research groups, like the one run by Eric Horvitz at Microsoft Research, are working on that. "We have lots of data, and lots of things we can try to model predictively," says Horvitz. "People think in terms of trends, but I want to build a data set where I can mark something as a surprise—a surprising conflict or surprising turn in the economy."

Horvitz is developing a system that picks out the words national leaders use to describe one another, trying to predict the onset of aggression. EMM has something similar, called tonality detection. Essentially, it's understanding the verbs as well as the nouns. Because once you know how people feel about something, you're a step closer to being able to guess what they'll do next.

BY ADAM ROGERS

SPOTTING THE HOT ZONES

EFFECTIVE DISEASE SURVEILLANCE RELIES ON SPEED AS MUCH AS ON INFORMATION.

IF YOU WANT TO STOP A DISEASE outbreak—or a bioterrorist attack—you have to act fast. But health information typically moves at the pace of the receptionist at your doctor's office. The goal of Essence, the Department of Defense's Electronic Surveillance System for the Early Notification of Community-based Epidemics, is to pick up the tempo. Begun in 1999 to collect health data in the Washington, DC, area, Essence now monitors much of the Military Health System, which includes 400 facilities around the world.

"You don't have to be accurate to detect things," says Jay Mansfield, director of strategic information systems at the Global Emerging Infections Surveillance and Response System, one of the agencies that developed Essence. "But you do need to be precise." Reports from every clinic, doctor, and pharmacy get broken into broad syndrome categories rather than specific diseases. One doctor might diagnose bronchitis and another pneumonia, but Essence doesn't care. It's just looking for similar illnesses and where and when they occur. "It's like a fire alarm," Mansfield says. "It goes off if there's smoke, so you can get in the kitchen and see what's going on."

Because 100 megabytes of data come in every day—the team stores 18 months' worth, about 2.5 terabytes—there's often more smoke than fire. A pharmacy running out of antidiarrheals could signal an outbreak of *E. coli* or just a two-for-one sale. Essence expanded to include new sources (like radiology and laboratory tests) this spring, which means the data issues just got even more complicated. The trick is parsing the data as it comes in so that patterns emerge in hours instead of days. "We detected a gastrointestinal outbreak in Korea," Mansfield says. "I called my boss, and he asked me, 'When did it happen?'"

Korea is 13 hours ahead of Washington. So Mansfield simply answered: "Tomorrow."

BY SHARON WEINBERGER

ILLUSTRATION BY STUDIO TONNE

SORTING THE WORLD

LEAVE IT TO GOOGLE TO FIGURE OUT A BETTER WAY TO MANAGE HUGE DATA SETS.

USED TO BE THAT IF YOU WANTED to wrest usable information from a big mess of data, you needed two things: First, a meticulously maintained database, tagged and sorted and categorized. And second, a giant computer to sift through that data using a detailed query.

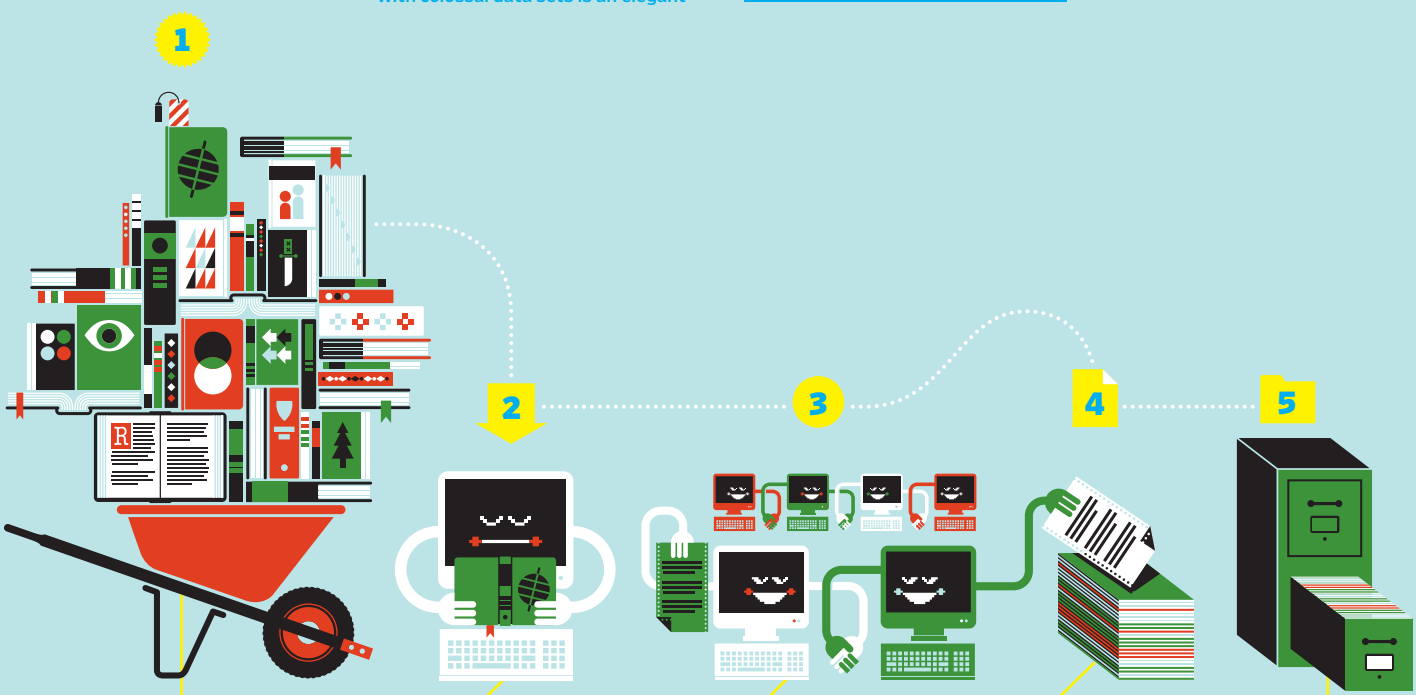
But when data sets get to the petabyte scale, the old way simply isn't feasible. Maintenance—tag, sort, categorize, repeat—would gobble up all your time. And a single computer, no matter how large, can't crunch that many numbers.

Google's solution for working with colossal data sets is an elegant approach called MapReduce. It eliminates the need for a traditional database and automatically splits the work across a server farm of PCs. For those not inside the Googleplex, there's an open source version of the software library called Hadoop.

MapReduce can handle almost any type of information you throw at it, from photos to phone numbers. In the example below, we count the frequency of specific words in Google Books.

BY PATRICK DI JUSTO

INFOGRAPHIC BY OFFICE



HOW GOOGLE CRUNCHES THE NUMBERS

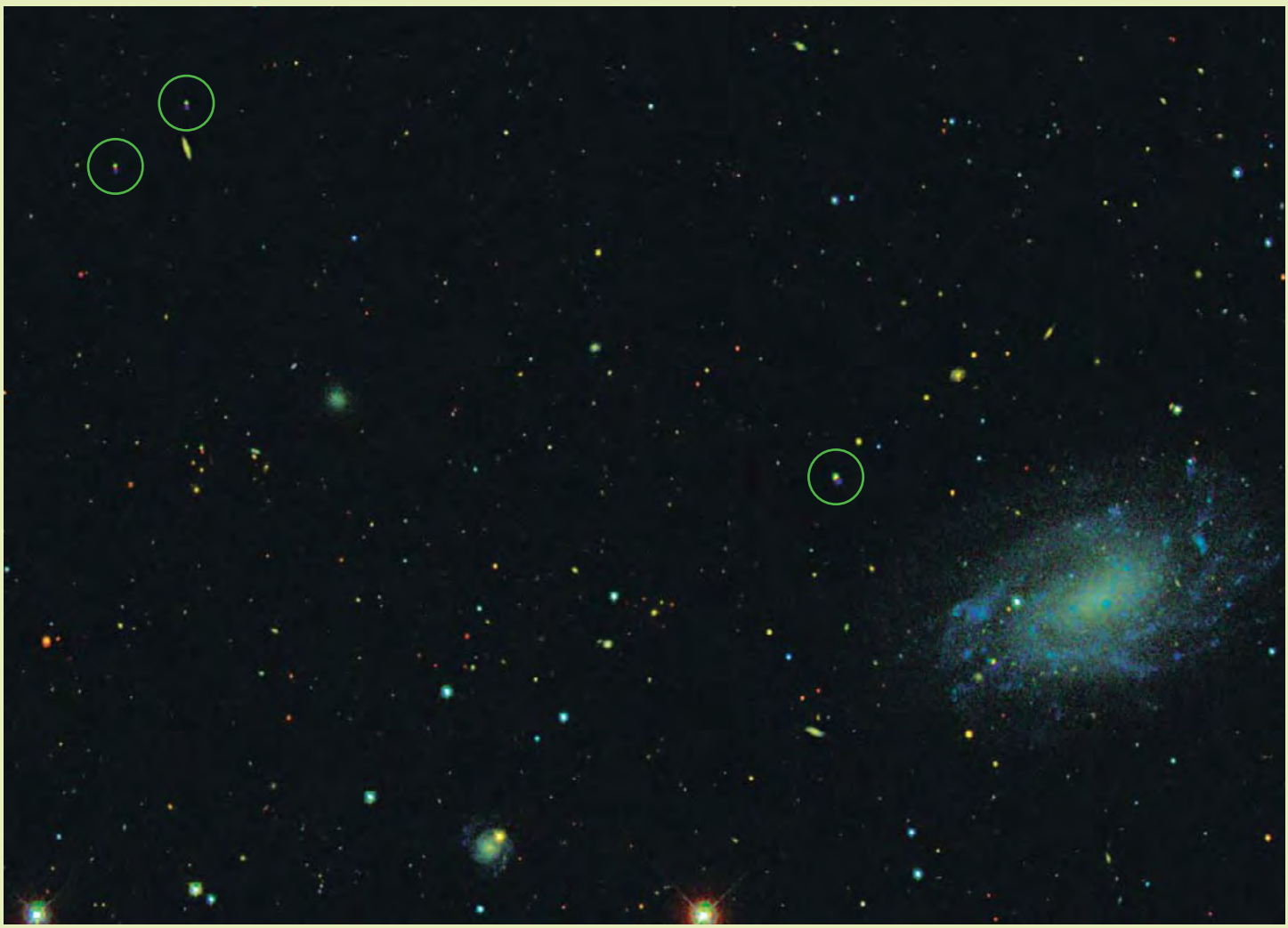
1. COLLECT
MapReduce doesn't depend on a traditional structured database, where information is categorized as it's collected. We'll just gather up the full text of every book Google has scanned.

2. MAP
You write a function to map the data: "Count every use of every word in Google Books." That request is then split among all the computers in your army, and each agent is assigned a hunk of data to work with. Computer A gets *War and Peace*, for example. That machine knows what words that book contains, but not what's inside *Anna Karenina*.

3. SAVE
Each of the hundreds of PCs doing a map writes the results to its local hard drive, cutting down on data transfer time. The computers that have been assigned "reduce" functions grab the lists from the mappers.

4. REDUCE
The Reduce computers correlate the lists of words. Now you know how many times a particular word is used, and in which books.

5. SOLVE
The result? A data set about your data. In our example, the final list of words is stored separately so it can be quickly referenced or queried: "How often does Tolstoy mention Moscow? Paris?" You don't have to plow through unrelated data to get the answer.



IN IMAGES FROM THE SLOAN DIGITAL SKY SURVEY, ASTEROIDS (CIRCLED IN GREEN) APPEAR TO MOVE OVER TIME. GALAXIES LIKE NGC4517A, AT LOWER RIGHT, DON'T.

WATCHING THE SKIES

SPACE IS REALLY BIG—BUT NOT TOO BIG TO MAP.

IN 1930, A YOUNG ASTRONOMER named Clyde Tombaugh found Pluto. He did it with a high tech marvel called a blink comparator; he put two photographs of the same patch of sky taken on different nights into the contraption and flipped back and forth between them. Stars would stay fixed, but objects like comets, asteroids, and planets moved.

Astronomers have since traded photographic plates for massive digital images. But Tombaugh's method—take a picture of the sky, take another one, compare—is still used to detect fast-changing stellar phenomena, like supernovae or asteroids headed toward Earth.

True, imaging the entire sky, and understanding those images, won't be easy. The first telescope that will be able to collect all that data, the Large Synoptic Survey Telescope, won't be finished until 2014. Perched atop Cerro Pachón, a mountain in northern Chile, the LSST will have a 27.5-foot mirror and a field of view 50 times the size of the full moon seen from Earth. Its digital camera will suck down 3.5 gigapixels of imagery every 17 seconds. "At that rate," says Michael Strauss, a Princeton astrophysicist, "the numbers get very big very fast."

The LSST builds on the most ambitious attempt to catalog the heavens so far, the Sloan Digital Sky Survey. Operating from a New Mexico mountaintop, the SDSS has returned about 25 terabytes of data since 1998, most of that in images. It has measured the precise distance to a million galaxies and has discovered about 500,000 quasars. But the Sloan's mirror is just one-tenth the power of the mirror planned for LSST, and its usable field of view just one-seventh the size. Sloan has been a workhorse, but it simply doesn't have the oomph to image the entire night sky, over and over, to look for things that change.

The LSST will cover the sky every three days. And within the petabytes of information it collects may lurk things nobody has even imagined—assuming astronomers can figure out how to teach their computers to look for objects no one has ever seen. It's the first attempt to sort astronomical data on this scale, says Princeton astrophysicist Robert Lupton, who oversaw data processing for the SDSS and is helping design the LSST. But the new images may allow him and his colleagues to watch supernovae explode, find undiscovered comets, and maybe even spot that killer asteroid.

BY MICHAEL D. LEMONICK

SCANNING OUR SKELETONS

SCIENTISTS AGGREGATE MILLIONS OF IMAGES TO UNDERSTAND HOW WE AGE.

WHAT CAN YOU LEARN FROM 80 million x-rays? The secrets of aging, among other things. Sharmila Majumdar, a radiologist at UC San Francisco, is using an arsenal of computer tomography scans to understand how our bones wear out from the inside.

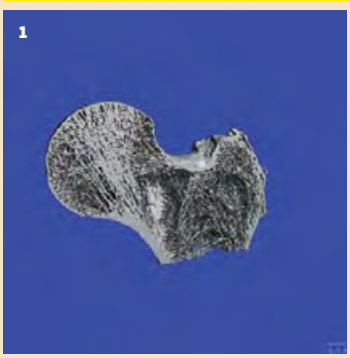
It works like this: A CT scanner takes superhigh-resolution x-rays of a bone, then combines those individual images into a three-dimensional structure. The results are incredibly detailed; a scan of a single segment of bone can run 30 gigs.

Majumdar's method is to churn through the data to identify patterns in how the trabeculae—the material inside bone—changes in people who have diseases like osteoporosis and arthritis. In one day of imaging, it's not uncommon for the lab to generate nearly a terabyte of data. Researchers also aggregate the data from many subjects, putting hundreds of terabytes to work. Majumdar hopes to learn why some patients suffer severe bone loss but others don't. "We don't know the mechanism of bone loss," she notes. "Once we learn that, we can create therapies to address it."

BY THOMAS GOETZ

HOW TO LOOK INSIDE OUR BONES

1



This slice of a human hip joint is 82 microns thick—about half the width of a human hair. Other machines used by the lab can go as fine as 6 microns—the size of a human red blood cell. Each bone is scanned about 1,000 times, creating, in this case, a clear look at osteoporosis in action.

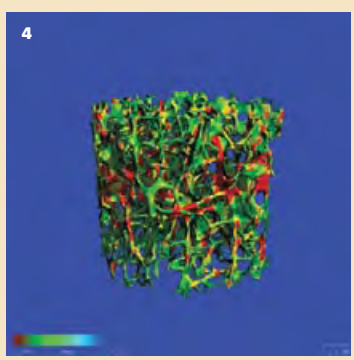
2

Using image processing, the slices are combined into a 3-D model, creating a picture of what the bone looks like from the outside…

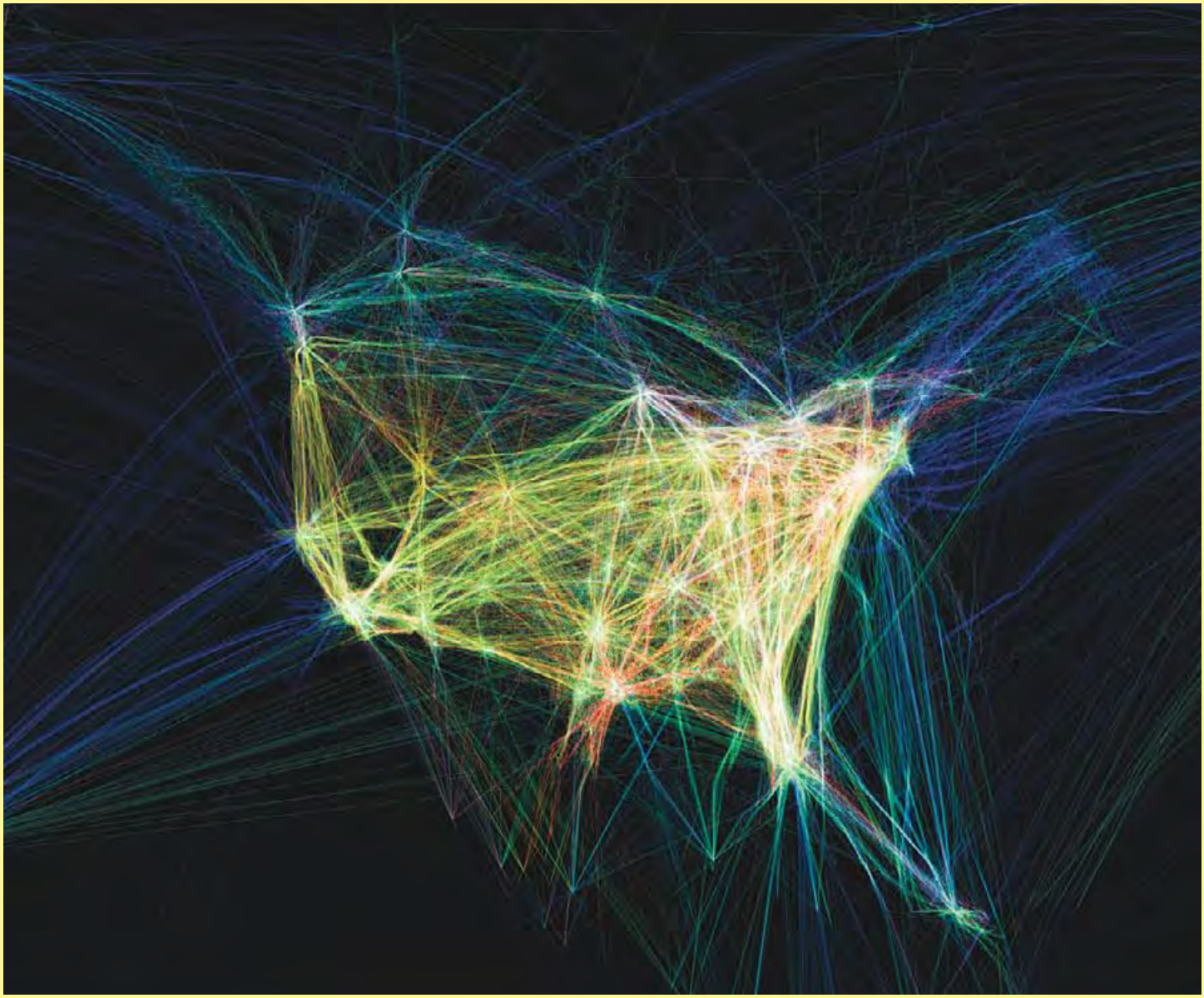
3

…and from the inside. This image of a human vertebra shows the internal microstructure of bone, called the trabeculae.

4



The lab then analyzes the model for weaknesses in density and strength. In this image, the thicker structures are color-coded green, while thinner material is colored red. Majumdar's lab combines hundreds of models to detect bone-loss patterns that help us understand how humans age.



"FLIGHT PATTERNS" SHOWS 141,000 AIRCRAFT PATHS OVER A 24-HOUR PERIOD.

TRACKING AIR FARES

DELVE THROUGH BILLIONS OF TICKET PRICES TO FIND OUT WHEN TO FLY.

IN 2001, OREN ETZIONI WAS ON A plane chatting up his seat mates when he realized they had all paid less for their tickets than he did. "I thought, 'Don't get mad, get even,'" he says. So he came home to his computer lab at the University of Washington, got his hands on some fare data, and plugged it into a few basic prediction algorithms. He wanted to see if they could reliably foresee changes in ticket prices. It worked: Not only did the algorithms accurately anticipate when fares would go up or down, they gave reasonable estimates of what the new prices would be.

Etzioni's prediction model has grown far more complex since then, and the company he founded in 2003, Farecast, now tracks information on 175 billion fares originating at 79 US airports. The database knows when airline prices are going to change and has uncovered a host of other secrets about air travel. Here's a dose of expert advice from the Farecast data vault:

1. COMMON WISDOM IS WRONG …
The lowest price tends to hit between eight and two weeks before departure. Buying tickets farther in advance usually doesn't save money.

2. … EXCEPT WHEN IT'S RIGHT
The rule fails during peak demand: Friday departures for spring break, and Sunday returns during the summer, Thanksgiving, and Christmas. For these, now is never too early.

3. WHEN THE PRICE DROPS, JUMP
Fifty percent of reductions are gone in two days. If you see a tasty fare, snatch it up.

4. IF PRICES SEEM HIGH, HOLD OFF
Behavioral economists call it framing: If last year's $200 flight is now $250, you'll probably find that too dear and won't buy. Everyone else is thinking the same thing. So when airlines hike the price of a route, they often have to cut rates later to boost sales.

5. THE DAY YOU FLY MATTERS
Used to be, you could count on a cheaper fare if you stayed over a Saturday night. But during spring break and summer, weekend trips are in high demand, so flights on Friday, Saturday, and Sunday can easily cost $50 more than those midweek.

6. SO DOES THE DAY YOU BUY
Price drops usually come early in the week. So a ticket bought on Saturday might be cheaper the next Tuesday. That's particularly true outside the summer rush, making fall the best time for a last-minute getaway.

7. MARKUPS VARY BY DESTINATION
Flights to Europe in July can be $350 higher than in May or September. If you want a summer vacation, domestic and Caribbean travel is cheaper to begin with and doesn't rise as high.

8. STAY AN EXTRA DAY
At the end of holidays, there's usually a stampede to the airport. One more day with the in-laws can save you upwards of $100—if you can stand it.

BY CLIFF KUANG

ART BY AARON KOBLIN

PREDICTING THE VOTE

POLLSTERS USE LARGE AND POWERFUL DATABASES TO IDENTIFY POLITICAL NICHES AND TARGET NEW SUPPORTERS.

WANT TO KNOW EXACTLY HOW many Democratic-leaning Asian Americans making more than $30,000 live in the Austin, Texas, television market? Catalist, the Washington, DC, political data-mining shop, knows the answer. CTO Vijay Ravindran says his company has compiled nearly 15 terabytes of data for this election year—orders of magnitude larger than the databases available just four years ago. (In 2004, Howard Dean's formidable campaign database clocked in at less than 100 GB, meaning that in one election cycle the average data set has grown 150-fold.) In the next election cycle, we should be measuring voter data in petabytes.

Large-scale data-mining and microtargeting was pioneered by the 2004 Bush-Cheney campaign, but Democrats, aided by privately financed Catalist, are catching up. They're documenting the political activity of every American 18 and older: where they registered to vote, how strongly they identify with a given party, what issues cause them to sign petitions or make donations. (Catalist is matched by the Republican National Committee's Voter Vault and Aristotle Inc.'s immense private bipartisan trove of voter information.)

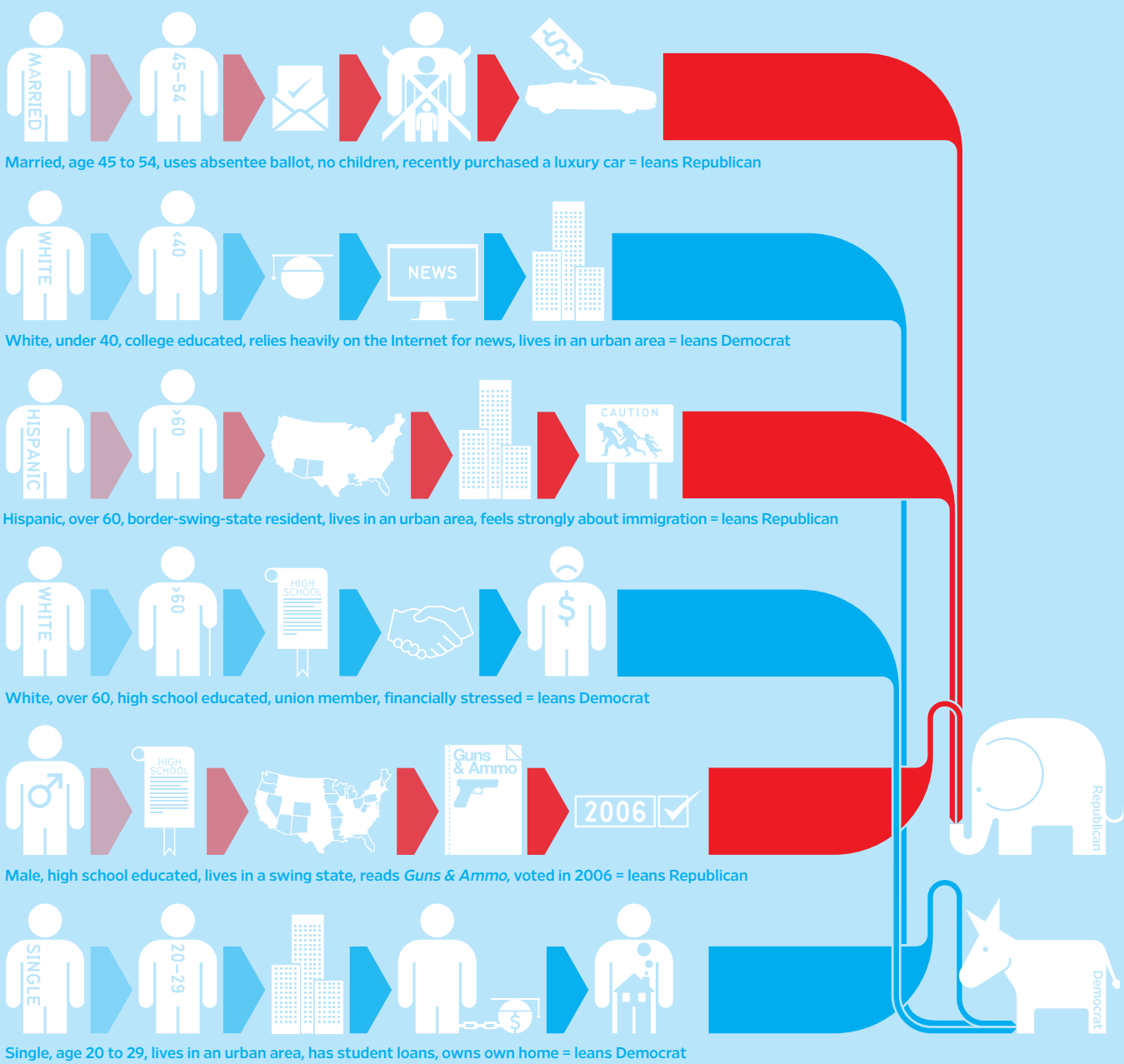
As databases grow, fed by more than 450 commercially and privately available data layers as well as first-hand info collected by the campaigns, candidates are able to target voters from ever-smaller niches. Not just blue-collar white males, but married, home-owning white males with a high school diploma and a gun in the household. Not just Indian Americans, but Indian Americans earning more than $80,000 who recently registered to vote.

Bill and Hillary's pollster, Mark Penn, has been promoting the dream of narrowcasting and microtrends for years (he invented "tech fatales," US women who drive decisions about electronics purchases). Penn was just a cycle or two early. The technology is finally catching up to his theories.

BY GARRETT M. GRAFF

INFOGRAPHIC BY BUILD

HOW WE CAST OUR BALLOTS



SOURCES: CATALIST, REPUBLICAN NATIONAL COMMITTEE

PRICING TERRORISM

EXTREMIST ATTACKS ARE RARE, UNPREDICTABLE, AND DEADLY. BUT INSURERS STILL MUST FIGURE OUT WHAT IT COSTS TO COVER THESE INCIDENTS.

IN THE AFTERMATH OF THE SEP-tember 11, 2001, attacks, Congress passed a law requiring commercial and casualty insurance companies to offer terrorism coverage. That was reassuring to jittery business owners but a major hassle for insurers, who, after all, are in the business of predicting risk. They might not know *when* something like a hurricane or earthquake will hit, but decades of data tell them where and how hard such an event is likely to be. But terrorists try to do the unexpected, and the range of what they might attempt is vast. A recent study published by the American Academy of Actuaries estimated that a truck bomb going off in Des Moines, Iowa, could cost insurers $3 billion; a major anthrax attack on New York City could cost $778 billion.

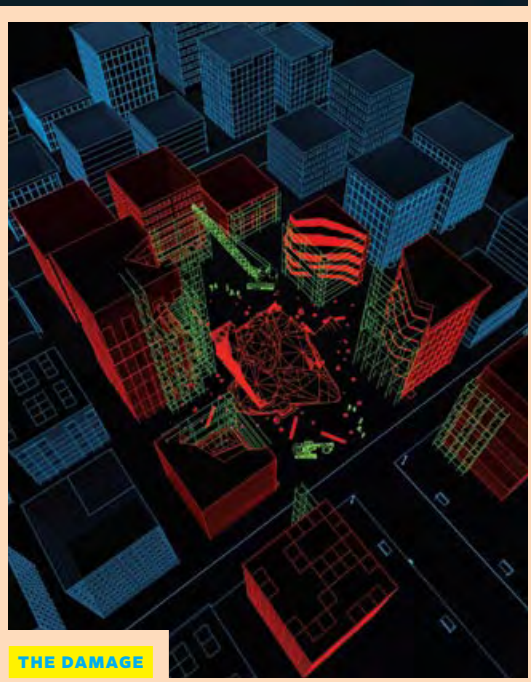
How do you predict a threat that's unpredictable by design? By marshaling trainloads of data on every part of the equation that is knowable. Then you make highly educated guesses about the rest.

BY VINCE BEISER

INFOGRAPHIC BY BRYAN CHRISTIE

THE TARGET

THE RISK



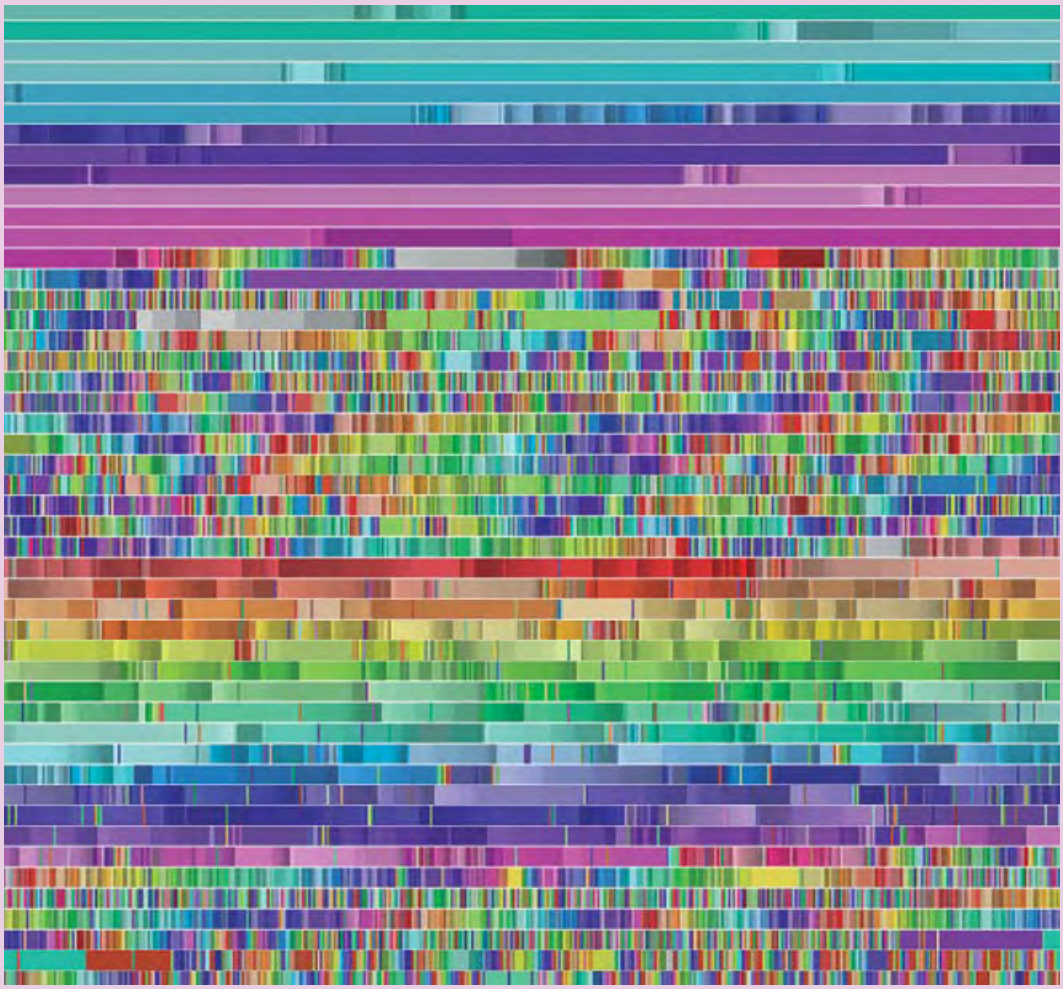
THE DAMAGE

THE TARGET
A random office building isn't likely to be in terrorists' crosshairs, but it could become collateral damage in a strike on, say, a nearby courthouse. To get help quantifying these risks, insurance companies turn to specialized catastrophe-modeling firms like AIR Worldwide. In 2002, AIR enlisted a group of experts formerly with the CIA, FBI, and Energy and Defense departments to brainstorm 36 categories of targets: corporate headquarters, airports, bridges, and so forth. AIR researchers then assembled a database of more than 300,000 actual locations around the US.

THE RISK
AIR's experts estimated the odds of an attack on each target type, taking into account various terror groups and the range of weapons they might use. They figured, for instance, that an animal research lab might have a higher risk of being hit by animal rights extremists than a post office.

THE DAMAGE
A decade ago, insurance companies —and the reinsurance companies that indemnify them—had only a rough idea of what they were covering. Today they have fine-grained details about nearly every property, down to the type of roofing and window glass. Terabytes of this data are run through models that factor in the area near the target, records of industrial accidents, and results of bomb tests. Those calculations yield estimates of casualties and damage, depending on whether the building was the target or a collateral hit.

THE BOTTOM LINE
Actuaries then convert all that mayhem into dollars, figuring out what the insurer will have to pay to repair buildings, replace equipment, and cover loss of life and medical care. What does that mean in terms of premiums? Typical coverage against terrorist attack for a five-story office building in Topeka, Kansas: $5,000 a year. That same building in lower Manhattan? $75,000. Even for mad bombers, it's all about location, location, location.



A VISUALIZATION OF THOUSANDS OF WIKIPEDIA EDITS THAT WERE MADE BY A SINGLE SOFTWARE BOT. EACH COLOR CORRESPONDS TO A DIFFERENT PAGE.

VISUALIZING BIG DATA

IF WE COULD SEE EVERYTHING EVER WRITTEN AT ONCE, HOW WOULD IT LOOK, AND WHAT COULD IT TELL US?

THE BIGGEST CHALLENGE OF THE Petabyte Age won't be storing all that data, it'll be figuring out how to make sense of it. Martin Wattenberg, a mathematician and computer scientist at IBM's Watson Research Center in Cambridge, Massachusetts, is a pioneer in the art of visually representing and analyzing complex data sets. He and his partner at IBM, Fernanda Viégas, created Many Eyes, a collaborative site where users can share their own dynamic, interactive representations of big data. He spoke with WIRED's Mark Horowitz:

HOW DO YOU DEFINE "BIG" DATA?
You can talk about terabytes and exabytes and zettabytes, and at a certain point it becomes dizzying. The real yardstick to me is how it compares with a natural human limit, like the sum total of all the words you'll hear in your lifetime. That's surely less than a terabyte of text. Any more than that and it becomes incomprehensible by a single person, so we have to turn to other means of analysis: people working together, or computers, or both.

WHY IS A NUMBERS GUY LIKE YOU SO INTERESTED IN LARGE TEXTUAL DATA SETS?
Language is one of the best data-compression mechanisms we have. The information contained in literature, or even email, encodes our identity as human beings. The entire literary canon may be smaller than what comes out of particle accelerators or models of the human brain, but the meaning coded into words can't be measured in bytes. It's deeply compressed. Twelve words from Voltaire can hold a lifetime of experience.

WHAT WILL HAPPEN WHEN WE HAVE DIGITAL ACCESS TO EVERYTHING, LIKE ALL OF ENGLISH LITERATURE OR ALL THE SOURCE CODE EVER WRITTEN?
There's something about completeness that's magical. The idea that you can have everything at your fingertips and process it in ways that were impossible before is incredibly exciting. Even simple algorithms become more effective when trained on big sets. Perhaps we'll find out more about plagiarism and literary borrowing when we have the spread of literature before us. We think of our current age as one of intellectual remixing and mashups, but maybe it's always been that way. You can only do that kind of analysis when you have the full spectrum of data.

IS THAT WHY, ON MANY EYES, YOU HAVE VISUALIZATIONS OF WIKIPEDIA USING SIMPLE WORD TREES AND TAG CLOUDS?
Wikipedia also has this idea of completeness. The information there again probably totals less than a terabyte, but it's huge in terms of encompassing human knowledge. Today, if you're analyzing numbers, there are a million ways to make a bar chart. If you're analyzing text, it's hard. I think the only way to understand a lot of this data is through visualization.

BY MARK HOROWITZ

IMAGE BY FERNANDA B. VIÉGAS, MARTIN WATTENBERG, AND KATE HOLLENBACH