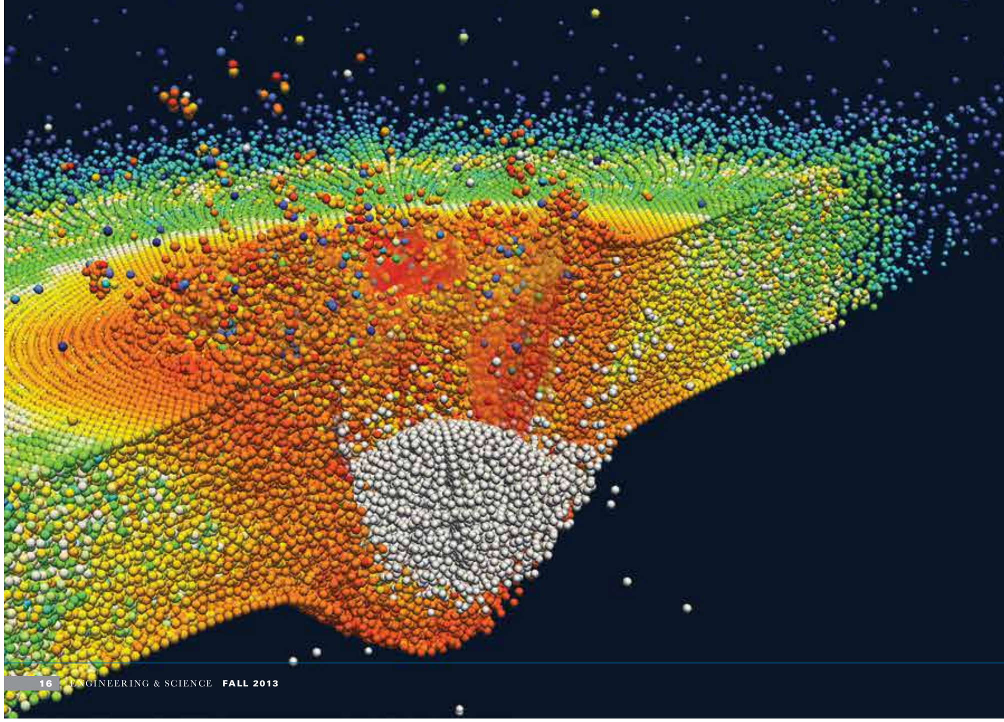
What's the Big Deal about Big Data?

For businesses from high-tech companies to banks, big data can mean big money. For Caltech researchers, however, this new glut of information means a new way of doing science.

by Marcus Y. Woo



THERE'S NO DOUBT that we're awash in more information than at any other point in history. Every time you swipe a credit card, buy something online, do an Internet search, or upload a photo or video, you add to the global flood of data. Thanks to the exponential rise of cheaper and faster computers—as described by Moore's law, in which Gordon Moore (PhD '54) accurately predicted that the number of transistors on a computer chip would double every two years—we can now collect, process, and store more data than we know what to do with. Such facts and figures as stock market fluctuations, financial loan information, people's "likes" on Facebook, and their shopping habits are potential gold mines—but only if these numbers can be turned into tangible, useful knowledge.

Of course, targeted advertising and potential corporate profits are just the tip of the information iceberg. Data is inundating all aspects of society; some experts say we are on the cusp of a transformative shift. We know from experience that we can mine these unprecedented heaps of information to glean insights into everything from medicine to the environment. The Human Genome Project's analysis of all the genes in our DNA, for instance, has revealed the genetic factors that predispose certain people to particular diseases, leading to better diagnoses and treatments. Scientists' constant monitoring of the earth is helping us understand how our climate is changing and how we can best respond to other hazards, like earthquakes and mudslides. Even President Obama's reelection campaign used large amounts of data, combined with sophisticated statistical analysis, to target potential voters like never before, a tactic that's been credited with his victory.

Which may or may not be why, last year, President Obama announced a $200 million Big Data Research and Development Initiative to improve the ways we take advantage of and learn from massive data sets in such areas as health care, the environment, national defense, and education.

Despite its name, when it comes to big data, size (or, as researchers tend to refer to it, volume) isn't everything. There's also velocity and variety—which, added to volume, form the so-called three Vs. After all, the huge quantities of data are being produced, collected, and disseminated so rapidly—a few gigabytes of measurements per second from the Higgs-boson-finding Large Hadron Collider, for instance—that scientists and engineers need to continually create new computer algorithms and techniques to be able to sort the important information from the useless.

In addition, there are often so many different kinds of data involved in studying a single problem that it becomes a real challenge to integrate it all and extract any kind of coherent insight. To monitor the global climate, for instance, scientists need to keep track not only of local temperatures but of sea and ice levels and the presence or absence of a multitude of greenhouse gases to gain an understanding of the system as a whole.

It's this level of complexity that distinguishes big data from the data of the past. While data has been getting bigger for decades, it has now become so abundant, complex, and rich that its underlying meaning is not always self-evident, and conventional approaches to understanding it no longer suffice.

That bigness is changing many areas of science, such as astronomy. Instead of measuring a specific thing, be it a gene or a single galaxy, scientists now grab data on *everything*—the whole genome or large swaths of sky—and only later comb through it for potential discoveries. "There are things you can do now that you couldn't do without this data," says astronomer George Djorgovski. "Data complexity—that's the really interesting part. It's where the new, exciting things happen."

Still, big data isn't just going to deliver scientific breakthroughs on a silver platter. Big data may help answer questions, says molecular biologist Barbara Wold, "but big data itself isn't an answer. It's not magic—nor should anyone expect it to be magic." Hard work, a little ingenuity, and the scientific method will always apply, Wold says.

But the big-data craze isn't all hype, says Mark Stalzer, former executive director of Caltech's Center for Advanced Computing Research; not by a long shot. "There's an underlying truth to it," he says. "There has to be something there. Actually, I think there's a lot of great stuff there."

The next few pages describe not only the impact big data is having on the science going on at Caltech, but also how Caltech computer scientists and engineers are creating the techniques and infrastructure that will be needed for us to navigate our data-intensive future. What follows is in no way a comprehensive look at big-data science at Caltech; there's simply too much of it. Instead, as Stalzer says of the field itself, "We're just scratching the surface."

Left: A frame from a simulation of a ballistic impact, looking at the stress experienced by the more than one million particles involved when a steel spherical projectile 1.778 mm in diameter hits a 1.6 mm-thick aluminum alloy plate at a speed of 2.7 kilometers per second.

Seismic Networks

The Southern California Seismic Network (SCSN)—run by Caltech and the U.S. Geological Survey—operates seismometers at more than 400 sites spanning the region. These sensors monitor every quiver and shake in the ground, recording seismic activity and sending the data via microwave signals, satellite, and the Internet to Caltech, where it's stored and analyzed. Earthquakes are immediately and automatically identified, located, and given a magnitude.

The data rates aren't so big as to be a problem—yet. But they will be a challenge as the SCSN researchers continue to build up the network, says Caltech geophysics professor Robert Clayton. This year alone will see the addition of 100 more sensors, which will add considerably more data for the seismologists to juggle. One possible solution is to upload the data to the cloud—which would also remove the inherent problem of having a data center located in the middle of earthquake country.

But it's not only the SCSN that's a data goldmine: Caltech also operates the Community Seismic Network (CSN), a denser network of about 300 sensors designed for the home or office, centered in the Pasadena region. Data from these sensors is automatically processed and uploaded to the cloud. The eventual goal is to have at least one sensor on almost every block—as well as in schools, hospitals, and on each floor of high rises. That would require expanding the network by a factor of over 100, says Mani Chandy, Simon Ramo Professor and professor of computer science at Caltech, and it would also require new computer architectures to process the increased data flow, in which several thousand signals must be processed every second. "Current algorithms aren't fast enough to deal with that amount of data," he says.

A **Greener** Cloud

"The cloud" sounds like somewhere magical—an ethereal place where much of our computing happens and where our email, photos, and music live, along with a seemingly endless amount of other information. But in reality the cloud isn't quite so perfect.

The cloud consists of massive data centers—enormous warehouses containing thousands and thousands of computers, humming away 24/7. Companies like Google, Microsoft, and Amazon operate tens of thousands of these data centers all around the country. And because the computers need to be on at all times—delays or interruptions are bad for business—they carry a substantial environmental cost.

"These data centers are huge power sucks," says Caltech professor of computer science Adam Wierman, who's working to make such centers environmentally sustainable. Accounting for 2 to 3 percent of the nation's energy use, data centers emit as much carbon as the airline industry, he says. An investigative report last fall by the *New York Times* found that some data centers waste at least 90 percent of the electricity they take from the power grid. And many are in violation of various environmental regulations.

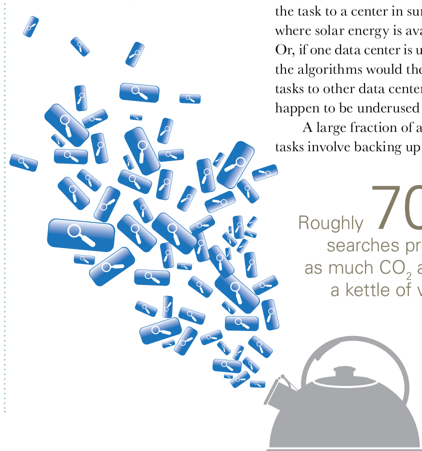
To help remedy the problem, engineers are working to develop more efficient hardware—such as processors that can run at higher temperatures and require less cooling—and some data centers are starting to run on renewable energy. But renewable energy is unpredictable: it's not always sunny, and the wind doesn't always blow. That's where Wierman comes in.

Data centers are managed by software that determines which server should do what when. To help these centers deal with erratic energy sources, Wierman and his colleagues have developed new algorithms that optimize how the centers are used. Say you want to watch a movie online. Instead of having your network access the movie through the nearest data center, even if it happens to be cloudy there, the new algorithms would send the task to a center in sunny Arizona, where solar energy is available. Or, if one data center is unusually busy, the algorithms would then distribute tasks to other data centers that happen to be underused at that time.

A large fraction of a data center's tasks involve backing up data or doing updates and other jobs that don't need to be completed right away. The new algorithms therefore delay nonurgent jobs while prioritizing those that require immediate attention. And if certain servers aren't needed at all at a particular time—say in the middle of the night, when demand is lower—then they will be shut off.

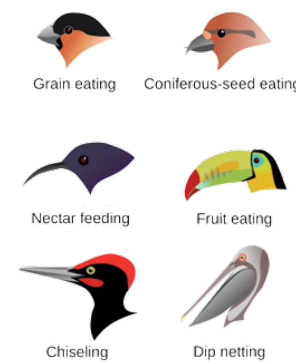
Although companies tend to get nervous when you start shuttling tasks around and turning servers on and off, the researchers have showed—on a fundamental, theoretical level—that their algorithms are indeed reliable and will save companies money in the long run. "We've been able to give really rigorous guarantees on the algorithms," Wierman says. Although the vast majority of data centers have yet to adopt these sustainable approaches, Wierman is partnering with Hewlett-Packard, which supplies server systems to other companies—including Apple—to implement the algorithms.

Wierman is now beginning to apply his algorithms to the integration of renewable-energy-powered data centers into the electrical power grid itself. For instance, when there's high demand on the grid—say on a sweltering summer day—a utility company could pay a renewable-energy-powered data center to lower its energy usage by, for instance, delaying nonurgent computational tasks. The result would be more available energy to be used elsewhere on the grid during those peak times. By providing such a power boost, the data centers would act like a battery that has stored away extra energy to inject into the grid, Wierman explains. "It's a huge win for the grid because batteries are expensive. It's going to be a long time before large-scale batteries are widely available." Using his algorithms in this way, he adds, will hopefully propel us ever closer to a sustainable future.



Roughly 70 Google searches produce as much CO_2 as boiling a kettle of water.

Image **Search**



Images are among the richest forms of data—and the web is full of them. To search for a particular image, computer algorithms target key words associated with the desired picture. But what if you want to look up something whose name you don't know but that you can picture in your head—you know, that broken thingamajig in your car engine or that colorful bird in your backyard?

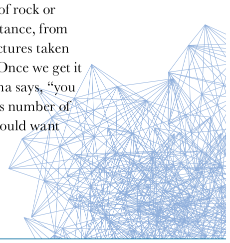
In most cases, images are not adequately cross-referenced, linked, or indexed to make such a search possible, explains Pietro Perona, Caltech's Allen E. Puckett Professor of Electrical Engineering. Such images are like the mysterious dark matter that pervades the universe—they're everywhere, yet invisible. "Images account for the largest portion of the web's data, and we don't know how to treat them," he says. "This is big data to the tenth power."

To address that problem, Perona and his colleagues are working with a group at UC San Diego to develop a visual encyclopedia that combines image-processing and machine-learning algorithms with expert crowdsourcing. They've dubbed it Visipedia because, in the same way that Wikipedia relies on the public for content, it relies on both experts and regular users to submit, label, and annotate images.

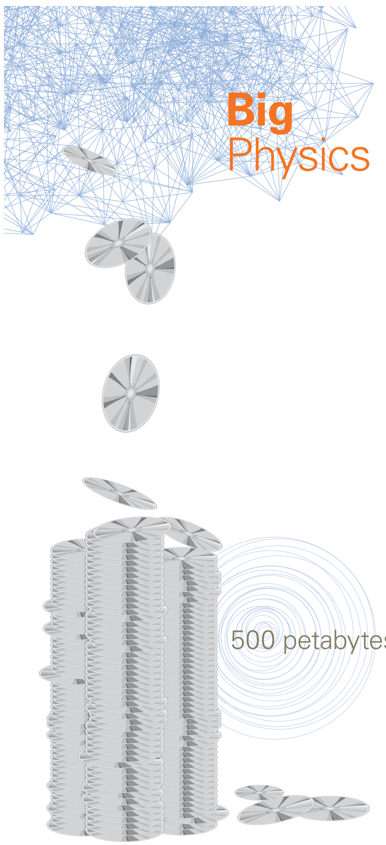
The researchers are starting relatively small, building Visipedia around bird images so as to take advantage of the enthusiasm and dedication of bird-watchers. The idea of Visipedia is for you to be able to upload a picture of a bird you've never seen before and get back more pictures of the same species—as well as a Wikipedia-like entry that identifies and describes it.

Initially, humans will do most of the image annotation. But as the annotators develop a systematic way to describe the images and identify important features—like the shape of a beak—the researchers hope that their computer-vision and machine-learning software will be able to learn enough from the human annotators to eventually do the same job itself. "You can think of this as a network of people and machines who are collaborating to achieve a certain goal," Perona says.

Ultimately, the Visipedia the scientists envision will be almost entirely automated—and it won't be aimed only at ornithologists and bird-watchers. This type of image recognition and searchability would be great for online shopping, Perona notes, or as a powerful tool for use in science, which is becoming increasingly dominated by visual data, organizing and automatically searching images ranging from stars and galaxies to cells and tissues and even atoms. Such a visual encyclopedia could help scientists hunt for images of a specific kind of rock or geological feature, for instance, from amid the thousands of pictures taken of the martian surface. "Once we get it just right for birds," Perona says, "you can imagine an enormous number of situations in which you would want something of this sort."



Big Physics



The Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) in Geneva—where last summer physicists discovered the Higgs boson—is a machine capable of slamming trillions of protons together at up to 99.9999991 percent of the speed of light, so fast that a proton careens around the accelerator's 27-kilometer circular track 11,000 times in one second.

The accelerator creates 600 million collisions per second, generating a flurry of other particles, which then decay into yet more particles. Detectors like the Compact Muon Solenoid—used by Caltech physicists at CERN—measure the velocity, position, electric charge, mass, and energy of every particle. That's a lot of data.

Indeed, there's so much data that sharing it with the thousands of physicists worldwide poses quite a challenge. In the 1990s, when construction of the LHC began, Caltech professor of physics Harvey Newman and computational scientist Julian Bunn came up with what turned out to be the best solution: a tiered system through which different types of data trickle down from CERN to institutions around the world for storage and sharing. With this distributed system, the data can be accessed by all the LHC partners without everything needing to be copied and sent to everyone.

CERN is the sole Tier 0 institution that creates and stores all of the raw data. There are 13 Tier 1 institutions that store the data for different regions of the world; that information is then distributed to and analyzed at hundreds of Tier 2 and Tier 3 institutions. About 500 petabytes of data are stored among all the institutions worldwide.

But that's not the whole of it. There's also computer-simulation data, the quantity of which is 10 times greater than that of experimental data, Bunn says. Comparing the simulation data with the experimental data allows physicists to search for unexpected phenomena—any discrepancies that might herald a new particle or even a new kind of physics. This simulation data, too, is distributed via the LHC grid, but because anyone can generate such simulations, it can be sent as easily from lower to higher tiers as from higher to lower.

The group led by Newman and Caltech physics professor Maria Spiropulu has also developed ever-better data-transfer methods; thanks to their work, more than 250 petabytes were transmitted through the LHC computing grid in 2012 alone. As the LHC continues to crank up its collision rates, Newman says, the flood of data will reach the exabyte range (an exabyte is a billion gigabytes). Over the next 10 years, data volumes and transfer rates are expected to grow a hundredfold.

A Faster Internet

A decade ago, it was impossible to transfer the large data sets that are now common at the Large Hadron Collider (LHC). The problem lay with the computers' so-called protocols—the systems of rules that dictate how data is transferred throughout a network like the Internet, setting up connections and automatically resending information that gets lost or delayed.

While the protocols of the early 2000s were satisfactory for the Internet and most people's needs, physicists like Caltech's Harvey Newman knew they would not be able to handle the oncoming deluge of LHC data. To solve the problem, Newman teamed up with Steven Low, professor of computer science and electrical engineering, and one of Caltech's experts on information networks.

At the time, Low says, there was no systematic way to design a protocol that would work for the huge networks required by physicists. "So what we did was try to really understand the problem by stepping back and developing a mathematical model of such networks," he explains. Working with professor John Doyle and a group of electrical engineers and computer scientists at Caltech, Low developed a deeper, structural understanding of these networks that allowed the team to build a protocol that could be as big and complex as needed. "This was not possible before," he says.

Low's ideas led to a new protocol called Fast TCP (for Transmission Control Protocol), which Newman used to set a new data-transfer record each year from 2003 to 2008. Newman and his group have since developed a sophisticated application called Fast Data Transfer—whose principal author is Caltech computational scientist Iosif Legrand—which doesn't just establish a protocol but optimizes the way huge data sets are transferred across the world, which is essential for doing LHC physics. This has allowed the team to continue breaking records: last fall, they hit a record-setting 339 gigabits per second, which is equivalent to sending one million full-length movies in one day.

Demands for high-speed data transfer have continued to grow, even among nonphysicists. And so, in 2006, Low and his colleagues started a company called FastSoft to commercialize Fast TCP. Last year, FastSoft was acquired by Akamai Technologies, which helps everyone from NBC to NASA deliver their online content. So that Grumpy Cat video you watched the other day? There's a good chance it was brought to you by Fast TCP.

Biology Gains Perspective

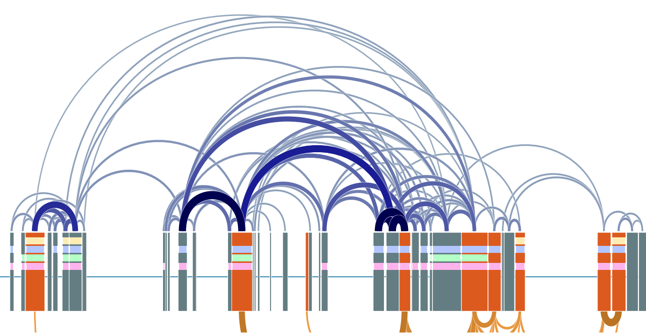
Barbara Wold, Bren Professor of Molecular Biology at Caltech, wants to untangle the complex webs of interacting genes—called gene regulatory networks—that determine whether a cell will ultimately become a muscle cell, a bone cell, or some other part of the adult organism. These regulatory networks consist of genes that turn one another on and off; how those interactions play out determines the cell's fate.

To understand this process in full genetic and biochemical detail, Wold and her colleagues use genomics, a field that focuses on the genome, with its 20,000 genes and hundreds of thousands of newly mapped regulatory elements. She says that the availability of "genome-wide" data for humans and key model organisms is transforming how scientists approach many problems in modern biology.

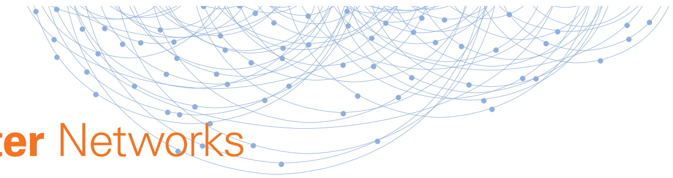
Of course, the amount of information is large, because each human or mouse genome consists of 6 billion DNA bases. This calls for new data-mining tools and ways to visualize data. Currently, integrating thousands of genomic datasets to extract new relationships among genes and their regulatory elements is a major challenge.

In June, Wold coauthored a white paper that announced a new global alliance for sharing genomic and clinical data. The ultimate goal of the alliance—of which Caltech is one of 70 founding institutional partners—is to pool data, including genome data and information about treatment received and outcomes. The resulting vast reservoir of data will help authorized researchers discover new causal relationships, doctors make better diagnoses, and scientists formulate new hypotheses. This is expected to be especially powerful for genomic diseases of high complexity such as cancer and autism.

"Creating and using large public databases that draw on data is a style of basic biology research that began with the first genome sequences," Wold says. "It is very exciting to see it fusing with clinical medicine to the benefit of both."



Left: Within a genome are sequences of DNA that regulate other genes, dictating when they turn on or off. Sometimes, those regulatory sequences and the genes they control are in completely different sections of the DNA strand—separated by thousands or even millions of DNA base pairs. This diagram illustrates the physical interactions between the genes in the mouse genome that are needed to turn a precursor cell called a myoblast into a myocyte, a type of skeletal muscle cell.



Better Networks

Mobile-phone base stations, fiber-optic cables, and satellites allow us to reach almost anyone anywhere on the planet. Our communication networks are huge, linking together millions of components—such as cell-phone towers and internet servers. In our data-driven world, the size and complexity of these networks will only increase.

The problem is, engineers don't have a systematic way of designing such intricate networks to function in the most efficient manner possible. "Network design is more of an art form than a science," says Michelle Effros, the George Van Osdol Professor of Electrical Engineering at Caltech and an expert on network and information theory. "People get good at it through experience and intuition."

Network components work differently linked together than they do as individual devices, researchers have discovered in recent years. But without a way to rigorously predict how a network will behave as a whole, engineers have to resort to trial and error, resulting in inefficient networks that can slow traffic and decrease reliability.

Now, Effros and her colleagues have devised some new mathematical models of generic network components that *do* predict how they would work when pieced together in a network. "Proving that such modeling is even possible is a surprising result," she says.

The researchers have so far developed models for the five most fundamental network components—which can be used to analyze all networks built from these components. Effros's ultimate goal is to keep enlarging this library, integrating the models into a piece of software that others can then use to design any kind of network they want. Using this tool, network engineers could, for instance, compare and optimize designs on a computer before they actually begin construction.

"The same ideas apply no matter what your network is," Effros says—whether you're talking about wireless networks, the Internet, or the sensors some grocery stores have installed on their shelves to monitor the freshness of foods. Some researchers are even exploring the possibility that such models can be used to understand the genetic networks that govern the development of embryos, she says.

"If we don't figure out how to use our networks properly and design them better, the path we're on will be limited," she says. "To keep expanding our communication capabilities requires real advances."

A Numbers Game

Analytics—also known as advanced statistics—has changed the way pros play baseball and basketball. And now it's changing Caltech basketball as well.

What is advanced statistics? "It's information beyond the box score," explains Oliver Eslinger, the head coach of the Caltech men's basketball team. Knowing how many points someone scored in a game isn't as informative as *how* that player got those points: when and where on the court he made his shots; whether those shots were contested jumpers or easy layups; whether they were the result of set plays or were assisted and, if so, which teammate passed the ball.

Eslinger and his coaching staff record and annotate every detail of every game and practice, developing formulas to better quantify the performance of each player and the team as a whole. Every play provides a plethora of data that can be used by coaches during practices, when determining game-day strategies and lineups, and when instructing players on how they can improve during the off-season.

The use of advanced statistics is still rare in Division III, Eslinger says. "I'd be surprised if any other DIII coaches are doing what we're doing," he notes. "I'd like us to be at the forefront of analytics in college sports." And for a place like Caltech, that's certainly fitting.

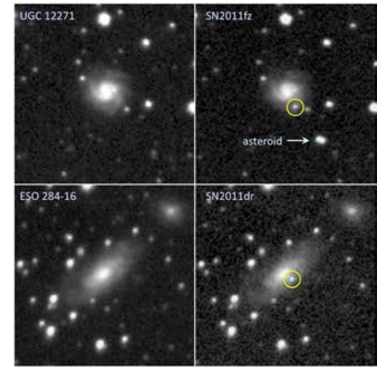
Flashes in the Night

Most stars remain static over the course of a human lifetime, but some rapidly brighten, dim, flare, or even explode. Objects whose brightness varies significantly are called transients, and they help astronomers understand how stars live and die, and how black holes form.

Caltech astronomers using Caltech's Palomar Transient Factory (PTF) and the Catalina Real-Time Transient Survey (CRTS) have trained automated telescopes on the heavens to search for such objects—and have brought in quite a haul. They have discovered thousands of variable stars and bright, black-hole-powered galactic nuclei. But the surveys may have made their biggest mark by recording thousands of new stellar explosions called supernovae.

Together, PTF and CRTS have discovered almost 4,000 of the more than 6,300 supernovae known so far. And the surveys' numbers rise daily. By collecting so much data, astronomers have been able to discover entirely new kinds of astronomical objects—such as new classes of supernovae—prompting new scientific inquiries, says astronomy professor George Djorgovski. And PTF and CRTS are only the beginning.

The Zwicky Transient Facility, also led by Caltech, is a PTF upgrade that's set to begin its survey in 2015. And, within the next decade, the Large Synoptic Survey Telescope (LSST) will begin a constant watch of the night sky with greater sensitivity and resolution than ever before.



Above: Examples of discoveries from the CRTS survey show, on the left, galaxies before supernova explosions. The images on the right show the galaxies with the supernovae circled. When the top right photograph was taken, there was an asteroid passing through the field at the same time; the software identifies asteroids and separates them from astrophysical transients like supernovae.

While PTF and CRTS might detect a few tens of transients per night, Djorgovski says, LSST should be able to find as many as 10 *million*.

In addition, this fall Caltech's Owens Valley Radio Observatory Long-Wavelength Array is set to begin imaging the entire viewable sky every second to search for transient signals at radio frequencies—in particular, signals from nearby exoplanets. These signals arise when particles spewing from the planets' stars interact with the planets' magnetic fields. During its hunt for transient signals, the array will generate 2.5 gigabytes of raw data every second, a rate similar in scale to that of the Large Hadron Collider, says astronomer Gregg Hallinan, who's leading the radio transient search.

The impending explosion of data makes the development of tools capable of analyzing all of it increasingly important, Djorgovski says. That's why he and his colleagues at Caltech and JPL are developing basic pattern-recognition and machine-learning algorithms to identify the information that's worth keeping.

Still, no matter how good the tools are, there is simply too much data for professional astronomers to analyze. The solution, Djorgovski says, is to make that information accessible to all. "When you have little data, data is precious," he says. "But when there's so much data that you can't possibly do it yourself, it's actually irresponsible *not* to let others do it."

The CRTS already makes its entire data set publicly available; the hope is that amateur enthusiasts will dive in and make their own discoveries. According to Djorgovski, it's this democratization of science that will be the most important consequence of the data explosion in astronomy.

"Anybody with an Internet connection has the same opportunity as astronomers at Caltech," he says. "And that's great." E&S