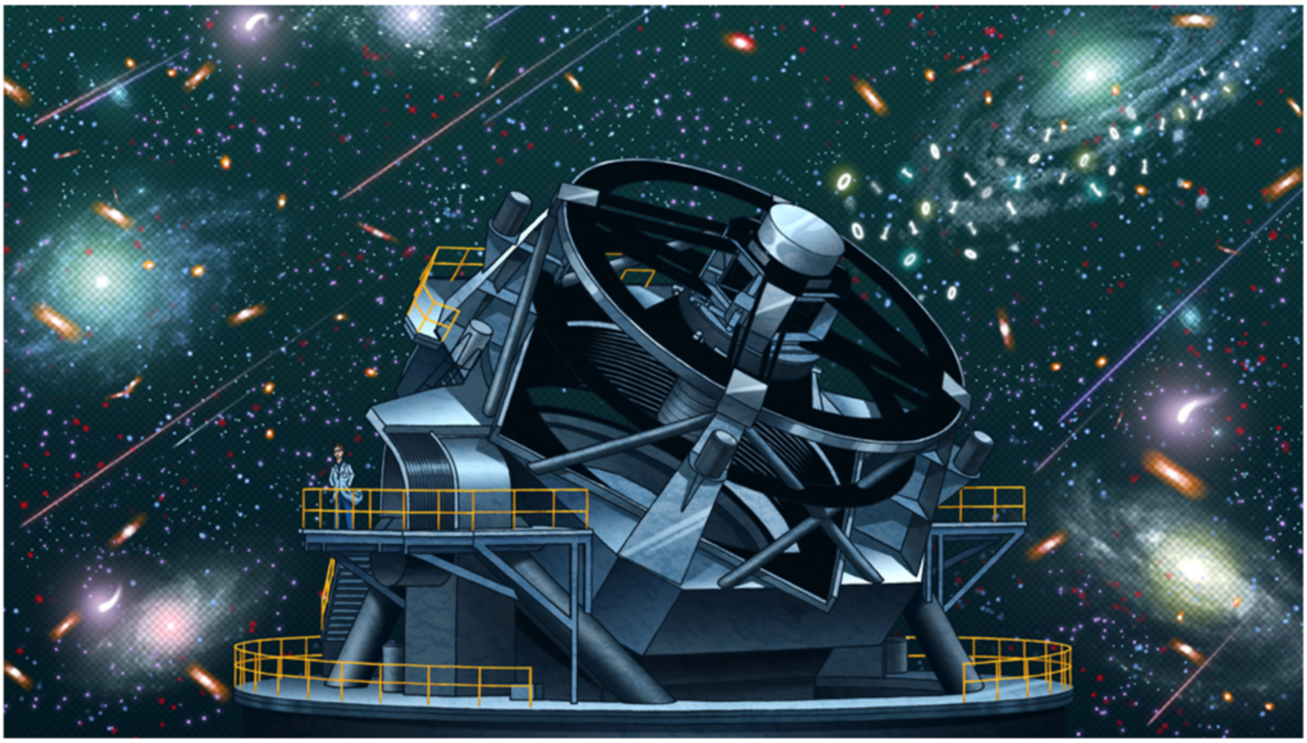
An editorially independent division of the SIMONS FOUNDATION

QUANTA MAGAZINE illuminating science

Quanta Magazine SimonsFoundation.org

Data Driven: The New Big Science

Who's Driving? Digits in the Sky *Revolutionary Algorithms* *The Facts of Life* *On Quantum Memory*



CHAPTER 2: DIGITS IN THE SKY

Yuta Onoda

A Digital Copy of the Universe, Encrypted

As physics prepares for ambitious projects like the Large Synoptic Survey Telescope, the field is seeking new methods of data-driven discovery.

By: Natalie Wolchover

October 2, 2013

Even as he installed the landmark camera that would capture the first convincing evidence of dark energy in the 1990s, Tony Tyson, an experimental cosmologist now at the University of California, Davis, knew it could be better. The camera's power lay in its ability to collect more data than any other. But digital image sensors and computer processors were progressing so rapidly that the amount of data they could collect and store would soon be limited only by the size of the telescopes delivering light to them, and those were growing too. Confident that engineering trends would hold, Tyson envisioned a telescope project on a truly grand scale, one that could survey hundreds of attributes of billions of cosmological objects as they changed over time.

It would record, Tyson said, "a digital, color movie of the universe."

Tyson's vision has come to life as the Large Synoptic Survey Telescope (LSST) project, a joint endeavor of more than 40 research institutions and national laboratories that has been ranked by the National Academy of Sciences as its top priority for the next ground-based astronomical facility. Set on a Chilean mountaintop, and slated for completion by the early 2020s, the 8.4-meter LSST will be equipped with a 3.2-billion-pixel digital camera that will scan 20 billion cosmological objects 800 times apiece over the course of a decade. That will generate well over 100 petabytes of data that anyone in the United States or Chile will be able to peruse at will. Displaying just one of the LSST's full-sky images would require 1,500 high-definition TV screens.

The LSST epitomizes the new era of big data in physics and astronomy. Less than 20 years ago, Tyson's cutting-edge digital camera filled 5 gigabytes of disk space per night

NEXT IN THE SERIES

Chapter 3: The Mathematical Shape of Things to Come

Friday Oct. 4

Quanta Magazine

Quanta Magazine is an online publication whose mission is to enhance public understanding of research developments in mathematics and the physical and life sciences. Quanta articles do not necessarily represent the views of the Simons Foundation.

About Quanta Magazine

Contact

Quanta@SimonsFoundation.org

Stay Connected & Informed

Sign Up for the Simons Foundation Newsletter

Email

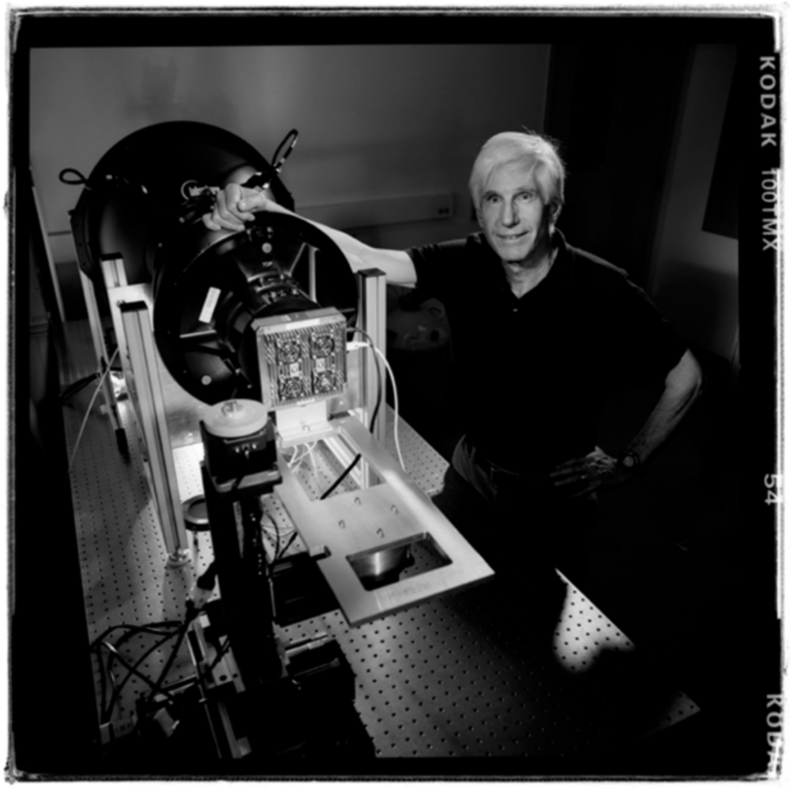
Subscribe

with revelatory information about the cosmos. When the LSST begins its work, it will collect that amount every few seconds — literally more data than scientists know what to do with.

"The data volumes we [will get] out of LSST are so large that the limitation on our ability to do science isn't the ability to collect the data, it's the ability to understand the systematic uncertainties in the data," said Andrew Connolly, an astronomer at the University of Washington.

Typical of today's costly scientific endeavors, hundreds of scientists from different fields are involved in designing and developing the LSST, with Tyson as chief scientist. "It's sort of like a federation," said Kirk Borne, an astrophysicist and data scientist at George Mason University. The group is comprised of nearly 700 astronomers, cosmologists, physicists, engineers and data scientists.

Much of the scientists' time and about one-half of the $1 billion cost of the project are being spent on developing software rather than hardware, reflecting the exponential growth of data since the astronomy projects of the 1990s. For the telescope to be useful, the scientists must answer a single question. As Borne put it: "How do you turn petabytes of data into scientific knowledge?"



Peter DaSilva for Quanta Magazine

Tony Tyson, an experimental cosmologist at the University of California, Davis, with a small test camera for the Large Synoptic Survey Telescope project, which he is helping to launch.

Physics has been grappling with huge databases longer than any other field of science because of its reliance on high-energy machines and enormous telescopes to probe beyond the known laws of nature. This has given researchers a steady succession of models upon which to structure and organize each next big project, in addition to providing a starter kit of computational tools that must be modified for use with ever larger and more complex data sets.

"How do you turn petabytes of data into scientific knowledge?"

Even backed by this tradition, the LSST tests the limits of scientists' data-handling abilities. It will be capable of tracking the effects of dark energy, which is thought to make up a whopping 68 percent of the total contents of the universe, and mapping the distribution of dark matter, an invisible substance that accounts for an additional 27 percent. And the telescope will cast such a wide and deep net that scientists say it is bound to snag unforeseen objects and phenomena too. But many of the tools for disentangling them from the rest of the data don't yet exist.

New Dimensions

Particle physics is the elder statesman of big data science. For decades, high-energy accelerators have been bashing particles together millions of times per second in hopes of generating exotic, never-before-seen particles. These facilities, such as the Large Hadron Collider (LHC) at CERN laboratory in Switzerland, generate so much data that only a tiny fraction (deemed interesting by an automatic selection process) can be kept. A network of hundreds of thousands of computers spread across 36 countries called the Worldwide LHC Computing Grid stores and processes the 25 petabytes of LHC data that were archived in a year's worth of collisions. The work of thousands of physicists went into finding the bump in that data that last summer was deemed representative of a new subatomic particle, the Higgs boson.

CERN, the organization that operates the LHC, is sharing its wisdom by working with other research organizations "so they can benefit from the knowledge and experience that has been gathered in data acquisition, processing and storage," said Bob Jones, head of CERN openlab, which develops new IT technologies and techniques for the LHC. Scientists at the European Space Agency, the European Molecular Biology

Laboratory, other physics facilities and even collaborations in the social sciences and humanities have taken cues from the LHC on data handling, Jones said.

When the LHC turns back on in 2014 or 2015 after an upgrade, higher energies will mean more interesting collisions, and the amount of data collected will grow by a significant factor. But even though the LHC will continue to possess the biggest data set in physics, its data is much simpler than those obtained from astronomical surveys such as the Sloan Digital Sky Survey and Dark Energy Survey and — to an even greater extent — those that will be obtained from future sky surveys such as the Square Kilometer Array, a radio telescope project set to begin construction in 2016, and the LSST.

Big Data in Physics

Some past, present and future large-scale physics experiments and the number of petabytes (10^{15} bytes) of data stored, or projected to be stored, per year.

Tevatron: 0.8
Particle accelerator at Fermilab in Illinois that discovered the top quark. (1983-2011)

Sloan Digital Sky Survey: 0.01
Largest sky survey to date, located at Apache Point Observatory in New Mexico. (2000-2014)

Large Hadron Collider: 25
World's largest particle accelerator at CERN in Switzerland and discoverer of the Higgs boson. (2010-?)

Large Synoptic Survey Telescope: 10
Optical astronomy survey telescope to be built in Chile. (2020-2030)

Square Kilometer Array: 365
Planned radio telescope with antenna arrays in South Africa and Australia. (2024-?)

"The LHC generates a lot more data right at the beginning, but they're only looking for certain events in that data and there's no correlation between events in that data," said Jeff Kantor, the LSST data management project manager. "Over time, they still build up large sets, but each one can be individually analyzed."

In combining repeat exposures of the same cosmological objects and logging hundreds rather than a handful of attributes of each one, the LSST will have a whole new set of problems to solve. "It's the complexity of the LSST data that's a challenge," Tyson said. "You're swimming around in this 500-dimensional space."

From color to shape, roughly 500 attributes will be recorded for every one of the 20 billion objects surveyed, and each attribute is treated as a separate dimension in the database. Merely cataloguing these attributes consistently from one exposure of a patch of the sky to the next poses a huge challenge. "In one exposure, the scene might be clear enough that you could resolve two different galaxies in the same spot, but in another one, they might be blurred together," Kantor said. "You have to figure out if it's one galaxy or two or N."

Beyond N-Squared

To tease scientific discoveries out of the vast trove of data gathered by the LSST and other sky surveys, scientists will need to pinpoint unexpected relationships between attributes, which is extremely difficult in 500 dimensions. Finding correlations is easy with a two-dimensional data set: If two attributes are correlated, then there will be a one-dimensional curve connecting the data points on a two-dimensional plot of one attribute versus the other. But additional attributes plotted as extra dimensions obscure such curves. "Finding the unexpected in a higher-dimensional space is impossible using the human brain," Tyson said. "We have to design future computers that can in some sense think for themselves."

Algorithms exist for "reducing the dimensionality" of data, or finding surfaces on which the data points lie (like that 1-D curve in the 2-D plot), in order to find correlated dimensions and eliminate "nuisance" ones. For example, an algorithm might identify a 3-D surface of data points coursing through a database, indicating that three attributes, such as the type, size and rotation speed of galaxies, are related. But when swamped with petabytes of data, the algorithms take practically forever to run.

Identifying correlated dimensions is exponentially more difficult than looking for a needle in a haystack. "That's a linear problem," said Alex Szalay, a professor of astronomy and computer science at Johns Hopkins University. "You search through the haystack and whatever looks like a needle you throw in one bucket and you throw everything else away." When you don't know what correlations you're looking for, however, you must compare each of the N pieces of hay with every other piece, which takes N-squared operations.

Adding to the challenge is the fact that the amount of data is doubling every year. "Imagine we are working with an algorithm that if my data doubles, I have to do four times as much computing and then the following year, I have to do 16 times as much computing," Szalay said. "But by next year, my computers will only be twice as fast, and in two years from today, my computers will only be four times as fast, so I'm falling farther and farther behind in my ability to do this."

A huge amount of research has gone into developing scalable algorithms, with techniques such as compressed sensing, topological analysis and the maximal

"Finding the unexpected in a

information coefficient emerging as especially promising tools of big data science. But more work remains to be done before astronomers, cosmologists and physicists will be ready to fully exploit the multi-petabyte digital movie of the universe that premiers next decade. Progress is hampered by the fact that researchers in the physical sciences get scant academic credit for developing algorithms — a problem that the community widely recognizes but has yet to solve.

higher-dimensional space is impossible using the human brain.”

“It’s always been the case that the people who build the instrumentation don’t get as much credit as the people who use the instruments to do the cutting-edge science,” Connolly said. “Ten years ago, it was people who built physical instruments — the cameras that observe the sky — and today, it’s the people who build the computational instruments who don’t get enough credit. There has to be a career path for someone who wants to work on the software — because they can go get jobs at Google. So if we lose these people, it’s the science that loses.”

Coffee and Kudos

In December 2010, in an effort to encourage the development of better algorithms, an international group of astronomers issued a challenge to computer geeks everywhere: What is the best way to measure gravitational lensing, or the distorting effect that dark matter has on the light from distant galaxies? David Kirkby read about the GREAT10 (GRavitational lEnsing Accuracy Testing 2010) Challenge on Wired.com and decided to give it a go.

Kirkby, a physicist at the University of California, Irvine, and his graduate student won the contest using a modified version of a neural network algorithm that he had previously developed for the BABAR experiment, a large physics collaboration investigating the asymmetry of matter and antimatter. The victory earned Kirkby a co-author credit on the recent paper detailing the contest, easing his switch from the field of particle physics to astrophysics. Also, with the prize money, “we bought a top of the line espresso machine for the lab,” he said.

Peter DaSilva for Quanta Magazine

David Kirkby, a physicist at the University of California, Irvine, holds an observing plate designed to capture data for a specific circular patch of the sky.

GREAT10 was one of a growing number of “data challenges” designed to find solutions to specific problems faced in creating and analyzing large physics and astronomy databases, such as the best way to reconstruct the shapes of two galaxies that are aligned relative to Earth and so appear blended together.

“One group produces a set of data — it could be blended galaxies — and then anybody can go out and try and estimate the shape of the galaxies using their best algorithm,” explained Connolly, who is involved in generating simulations of future LSST images that are used to test the performance of algorithms. “It’s quite a lot of kudos to the person who comes out on top.”

Many of the data challenges, including the GREAT series, focus on teasing out the effects of dark matter. When light from a distant galaxy travels to Earth, it is bent, or “lensed,” by the gravity of the dark matter it passes through. “It’s a bit like looking at wallpaper through a bathroom window with a rough surface,” Kirkby said. “You determine what the wallpaper would look like if you were looking at it directly, and you use that information to figure out what the shape of the glass is.”

Each new data challenge in a series includes an extra complication — additional distortions caused by atmospheric turbulence or a faulty amplifier in one of the detectors, for example — moving the goal posts of the challenge closer and closer to reality.

Data challenges are “a great way of crowd-sourcing problems in data science, but I think it would be good if software development was just recognized as part of your productivity as an academic,” Kirkby said. “At career reviews, you measure people based on their scientific contributions even though software packages could have a much broader impact.”

The culture is slowly changing, the scientists said, as the ability to analyze data becomes an ever-tightening bottleneck in research. “In the past, it was usually some post-doc or grad student poring over data who would find something interesting or something that doesn’t seem to work and stumble across some new effect,” Tyson said. “But increasingly, the amount of data is so large that you have to have machines with algorithms to do this.”

Dark Side of the Universe

Assuming that physicists can solve the computing problems they face with the LSST, the results could be transformative. There are many reasons to want a 100-petabyte digital copy of the universe. For one, it would help map the expansion of space and time caused by the still-mysterious dark energy, discovered with the help of the LSST’s predecessor, the Big Throughput Camera, which Tyson and a collaborator built in 1996.

When that camera, which could cover a patch of the sky the size of a full moon in a single exposure, was installed on the Blanco Telescope in Chile, astrophysicists immediately discovered dozens of exploding stars called Type IA supernovae strewn across the sky that revealed that most stuff in the universe is unknown. Light from nearby supernovae appeared to have stretched more than it should have during its journey through the expanding cosmos compared with light from faraway ones. This suggested that the expansion of the universe had recently sped up, driven by dark energy.

With the LSST, scientists hope to precisely track the accelerating expansion of the universe and thus to better define the nature of dark energy. They aim to do this by mapping a sort of cosmic yardstick called baryon acoustic oscillations. The yardstick was created from sound waves that rippled through the universe when it was young and hot and became imprinted in the distribution of galaxies as it cooled and expanded. The oscillations indicate the size of space at every distance away from Earth — and thus at any point back in time.

Baryon acoustic oscillations are so enormous that a truly vast astronomical survey is needed to make them a convenient measuring tool. By cataloguing billions of galaxies, the LSST promises to measure the size of these resonances more accurately than any other existing or planned astronomical survey. “The idea is that with the LSST, we will have onion shells of galaxies at different distances and we can look for this pattern and trace the size of the resonant patterns as a function of time,” Szalay said. “This will be beautiful.”

But, Szalay added, “it will be a nontrivial task to actually milk the information out of the data.”

print

Previous Chapter:
Who's Driving?

Next Chapter:
Chapter 3: The Mathematical Shape of Things to Come

Leave a Comment

Your email address will not be published. Your name will appear near your comment. Required *

Name *

Email *

Website

Comment

+ Post your comment