worms. Plagiarism shouldn't be tolerated, but you need a professional organization to handle the heat."

The arXiv's automated scanning for overlapping text is a refinement of an algorithm used last year by Cornell computer science graduate student Daria Sorokina to look at the server's then nearly 300 000 documents. The algorithm assigns unique numbers to word sequences and then compares those numbers across documents. Common phrases such as "this work was supported in part by" are excluded. "There is nothing new about document fingerprinting," says Cornell computer scientist Johannes Gehrke, an adviser on the project. "The novelty here was the application to the arXiv."

In the study, about 10% of arXiv manuscripts had text blocks that overlapped with other documents. After removing instances of authors reusing parts of their own text, different collaborators on a single project using the same text in separate conference abstracts, and other apparent false positives, less than 1% of manuscripts were still suspect, says Sorokina.

Close examination of 20 pairs of documents with among the highest levels of overlap exposed 16 as plagiarism. "In one case, an author copied descriptions of five or six methods that he was comparing," says Sorokina. "He didn't cite the sources. But the work of comparing was his own." One of the most common types of plagiarism found was the lifting of introductory or background material, especially in PhD theses, says Ginsparg. "The surprising thing is that people submit to the same database where they found [what they copied]. It's mind boggling, given the existence of Google, given the existence of searching on full text, that people wouldn't have an intuition that they would be caught."

"Some of it is different ethical norms," Ginsparg adds. "People in different countries, with different intellectual backgrounds, will sometimes argue that what they are doing is completely correct." The reassuring thing, he adds, "is that the most creative people, who are generating the ideas, don't have to start from someone else's article as a template. We'd be very surprised if authors of prominence showed up as perpetrators as opposed to victims."

Document fingerprinting catches only word-for-word plagiarism. But work is under way in the data-mining community on author identification and detection of the flow of ideas, says Gehrke. "Detecting content-based similarities with more sophisticated methods on a macroscale will be the next step."

In addition to implementing a check on new submissions to the arXiv, Ginsparg is talking to the editors of *Physical Review Letters* about applying the method to it and other American Physical Society publications. "More work needs to be done to include papers outside of the arXiv, and to go across journals," says Marty Blume, the recently retired APS editor-in-chief. "We have 30 000 submissions a year. We'll have to see how much [of the editors'] time it takes to run. And if we do it, what do we do with the results?" **Toni Feder**

Google to handle telescope data

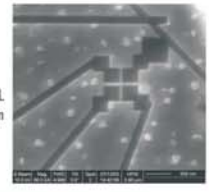
Nearly a decade ago, Anthony Tyson and some colleagues at Bell Labs were mulling how dark matter and dark energy might be studied, a notion that later led to the idea of designing a huge telescope. Meanwhile, another group at Bell wondered how to manage gargantuan amounts of data. Over the years, members of the two groups worked together on a telescope concept even as many left

MICHAEL MULLEN DESIGN, LSST CORPORATION

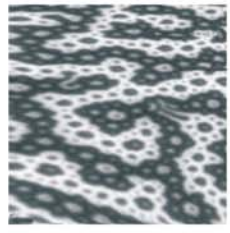
An artist's rendering shows the proposed design of the Large Synoptic Survey Telescope dome and local facilities. The telescope, which will see first light in 2014, will open a window on objects that change or move rapidly, including supernovae and near-Earth asteroids.

RT-SHPM

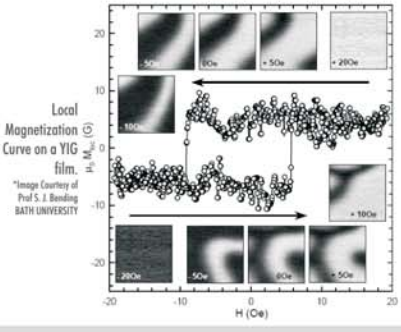
ROOM TEMPERATURE SCANNING HALL PROBE MICROSCOPE



WORLD'S SMALLEST HALL PROBE: 50nm



NdFeB DEMAGNETISED MAGNET



Local Magnetization Curve on a YIG film.
*Image Courtesy of Prof S. J. Bending BATH UNIVERSITY

- Scanning Hall Probe Microscopy
  - 50 nm spatial resolution!!
  - Real time scanning with SHPM!!
  - Unprecedented sensitivity: Up to 7mG/Hz$^{1/2}$
  - AFM or STM Tracking SHPM
- Multi-Mode Operation:
  - MFM, AFM, STM, EFM...

QUANTITATIVE & NON-INVASIVE MAGNETIC MEASUREMENTS AT NANOMETER SCALE

NanoMagnetics Instruments Ltd.

www.nanomagnetics-inst.com
info@nanomagnetics-inst.com

See www.pt.ims.ca/12304-20

Bell for other pastures. Today, Tyson, now a physics professor at the University of California, Davis, is director of the Large Synoptic Survey Telescope, and some of his former Bell colleagues are at Google Inc, which recently announced it is joining the project.

The 8.4-meter telescope, slated to see first light in Chile in 2014, will survey the entire visible sky every three nights with its 3-billion-pixel digital camera. A new image will emerge every 15 seconds—adding up to a staggering 30 to 50 terabytes a night. That's where Google comes in. The company, known principally for creating and operating one of the world's most popular Internet search engines, expects to apply its data management expertise to processing, organizing, and storing the reams of information the LSST will produce. "There's no way individuals can look at this data and discover things," Tyson said.

Don Sweeney, LSST project manager and a physicist at Lawrence Livermore National Laboratory, pointed out that often, privately owned telescopes have proprietary data that are not publicly available. "These [LSST] data will be available to everybody—an astronomer at Harvard or a high-school student in Iowa," Sweeney said. Google's involvement also means that all the information generated will be available to the public in a far more organized and ordered fashion than in the past.

The project's price tag is $380 million in 2006 dollars, and the money is expected to come from private foundations and US government agencies, including NSF and the Department of Energy. As of January, 20 organizations, including universities and other research institutions, had joined LSST. Google is the only private-sector partner.

Google spokeswoman Amanda Angelotti wouldn't provide specific information on exactly how Google will process and store the LSST data, saying technological details haven't been worked out yet. But company engineers have already begun wrestling with some data-management issues. Tyson said Google is also working to develop the most efficient way of processing LSST images. "We actually have to detect things that change in real time," he said.

Rob Pike, a former Bell scientist who is now Google's lead engineer on the LSST, said one of the project's challenges will be to continue to manage the constant influx of data in the face of inevitable technological breakdowns.

"[The LSST] will build on Google's expertise in providing reliable, large-scale distributed computing platforms. The flow of data is large and unrelenting, so computational downtime can cause logistical problems for data that pile up waiting to be processed," Pike said. "Reliability is important for throughput, and with the quantity of data involved, many machines must process it in parallel. Pieces of the system are sure to fail at some point—computers break, and the more you have, the more likely one will break—so the key is to keep the system running even when some of its machines are down."

Karen H. Kaplan

More physics in US high schools

US high-school students are taking physics in unprecedented numbers. And the upward trend in enrollments is likely to continue. These and other findings from a new statistics study by the American Institute of Physics were presented in January at a joint meeting of the American Association of Physics Teachers and the American Astronomical Society in Seattle.

In 2005 33% of all graduating high-school seniors had taken a physics class, up from 20% in 1987. "Now more than half of four-year college- and university-bound students take it," says Michael Neuschatz, coauthor of the AIP study. "When we look at the forces that are propelling the rise, it looks like the rise will continue."

Those forces include states' raising the science requirement for graduation, colleges and universities increasingly using tough courses such as physics in deciding who to admit, and tight competition compelling college-bound students to enhance their transcripts by taking physics, says Neuschatz.

The rise in enrollments is largely in nontraditional courses. The broadening of course offerings, Neuschatz says, "is helping to foster expansion beyond the usual reservoir of students. For example, girls have been brought in to the point that they are almost at parity." Girls make up 47% of high-school physics students, a percentage that's held for the past decade. The percentages of Hispanic and African American students that take physics in high school is growing, although these groups are still underrepresented compared to white and Asian American students.

Conceptual physics courses have opened the door to students who may not want to pursue math, science, or engineering. Indeed, says Neuschatz, the greater appeal of physics represents a "cultural change. It makes physics more appreciated by and accessible to a broader community." He adds that the growth in conceptual physics has not been at the expense of traditional classes. "It's a broadening, not a leaching." Enrollments in advanced placement courses grew about fivefold from 1987 to 2005.

More students means that more teachers can specialize in physics. The percentage that consider themselves physics specialists was 57% in 2005, up from 40% in 1993. However, the fraction who have a bachelor's degree in physics or physics education remained stable at one-third. Some 2300 teachers participated in the survey.

Reaching the Critical Mass: Findings from the 2005 Nationwide Survey of High School Physics Teachers is available online at http://www.aip.org/statistics/trends/hstrends.html. Single copies may be obtained free of charge from AIP, Statistical Research Center, One Physics Ellipse, College Park, MD 20740; e-mail stats@aip.org.

Toni Feder