
CHIPS

ANOTHER WAY OF DOING SCIENCE

BY SHARON TREGASKIS

# CAUSE AND EFFECT

In 1979, multiple scientific teams working independently in London, New York, Paris, and Princeton began zeroing in on a mysterious protein that seemed to speed cancer growth. They later named their quarry p53, a nod to its apparent 53-kilodalton weight.

Early studies suggested that the code responsible for the protein—the gene *TP53*, since found on the short arm of human chromosome 17 and in the DNA of most mammals whose genomes have been sequenced—was an oncogene, the cancer biology equivalent of your car's accelerator.

Scientists would only discover years later that those first experimental cell lines had featured rogue forms of p53. As it turns out, unmutated p53 is actually more akin to a brake. It regulates the activity of other genes, monitoring the fidelity of the cell-division process, stopping erratic cell replication, and spurring repairs.

In cases of irreparable mutations, p53—dubbed "molecule of the year" by *Science* in 1993 and now widely known among cancer biologists as the "guardian of the genome"—induces senescence and even programmed cell death to stop precancerous cells in their tracks, before things get ugly.

Data can deceive and confuse. Getting it to reveal meaningful relationships is the charge of a new team of big thinkers from Pitt, Carnegie Mellon, and the Pittsburgh Supercomputing Center.

ILLUSTRATIONS | STACY INNERST

But like Longfellow's little girl with a curl, when *TP53* goes bad, it goes horrid—hence that early mistaken oncogene hypothesis. We now know that more than 50 percent of cancers feature mutated *TP53*; it's the most commonly mutated gene found in human malignancies. It's also among the more potent; tumors in which biopsies reveal *TP53* mutations behave more aggressively and are associated with worse patient outcomes.

"The excitement generated by [p53] and its fellow tumor suppressors is reaching a crescendo," wrote *Science*'s then editor-in-chief Daniel Koshland Jr. 23 years ago when the journal made its molecule of the year announcement, "with exhilarating possibilities for prevention and cure of cancer."

The possibilities of p53 remain tantalizing, yet the quest to realize them in the past quarter century has turned into a long, hard slog. Part of the issue is the mind-boggling volume of biomedical data now available to investigators. Advances in biotechnology have fueled the exponential growth of biomedical datasets.

The prospect of coaxing gold from the dross has become epic. "The good news is that we have so much data. But the bad news is that we have so much data," says Jeremy Berg, University of Pittsburgh PhD professor of computational and systems biology and Pittsburgh Foundation Professor.

The challenge? Even more than having the means to churn it all is the problem of finding real meaning. Or, for Berg and his colleagues: "meaningful relationships that lead us to new insights in health and disease."

Complicating it all, notes Pitt's Gregory Cooper, MD/PhD director of the new Center for Causal Discovery (CCD), a group trying to figure all this out, is not just the many data points, but the enormous number of variables —"even more than the number of data points at times"—that can arise in the medical world.

It's not surprising that researchers would turn to computers to help them discover connections within data. But the CCD team is actually asking computers to, effectively, play a collaborative role in the process of scientific inquiry. This group—and the few others like it elsewhere—is overhauling how modern science is done.

"You want to know what experiments are worth doing out of the millions that could be done," says Clark Glymour, a PhD and founding chair of Carnegie Mellon's Department of Philosophy. (That's right, philosophy. Glymour's realm is the study of knowledge, a.k.a. epistemology, how we know what we know.) Glymour is the CMU lead for the CCD, a partnership of Pitt, CMU, and the Pittsburgh Supercomputing Center.

Glymour (pronounced *glee-more*) says he and his colleagues can put together "search procedures that suggest what experiments are most worth doing because the computer returns the most likely results."

"The possibility space is vast," says philosopher of science Richard Scheines, a PhD and former graduate student of Glymour's who is now dean of CMU's Dietrich College of Humanities & Social Sciences. "If you can't sift through all those possibilities intelligibly, it's impossible to expect the community to run through the experiments. We're hoping we can give scientists a much narrower focus on what to confirm."

## Meaning Amid the Morass

Before big data came along, scientific discovery was a granular process. Well-read investigators with a keen eye could (and still do, of course) integrate the latest published articles with their own observations and experimental manipulations to spark new hypotheses.

Naturally, as more opportunities for collecting and distilling data have become available throughout the past few decades, many scientists have looked in those corners for answers. Yet, "20 years ago, if you wrote a grant to collect a large volume of data saying, *I don't know what I'll find, but I'll first collect data and find out later*, the grant would be doomed to fail," says Xinghua Lu, an MD/PhD who codirects Pitt's Center for Translational Bioinformatics and heads the CCD's cancer team. "The reviewers would certainly criticize the application as a fishing expedition with no hypothesis. Nowadays, biology is becoming more of a data science, and data-driven hypothesis generation is more accepted."

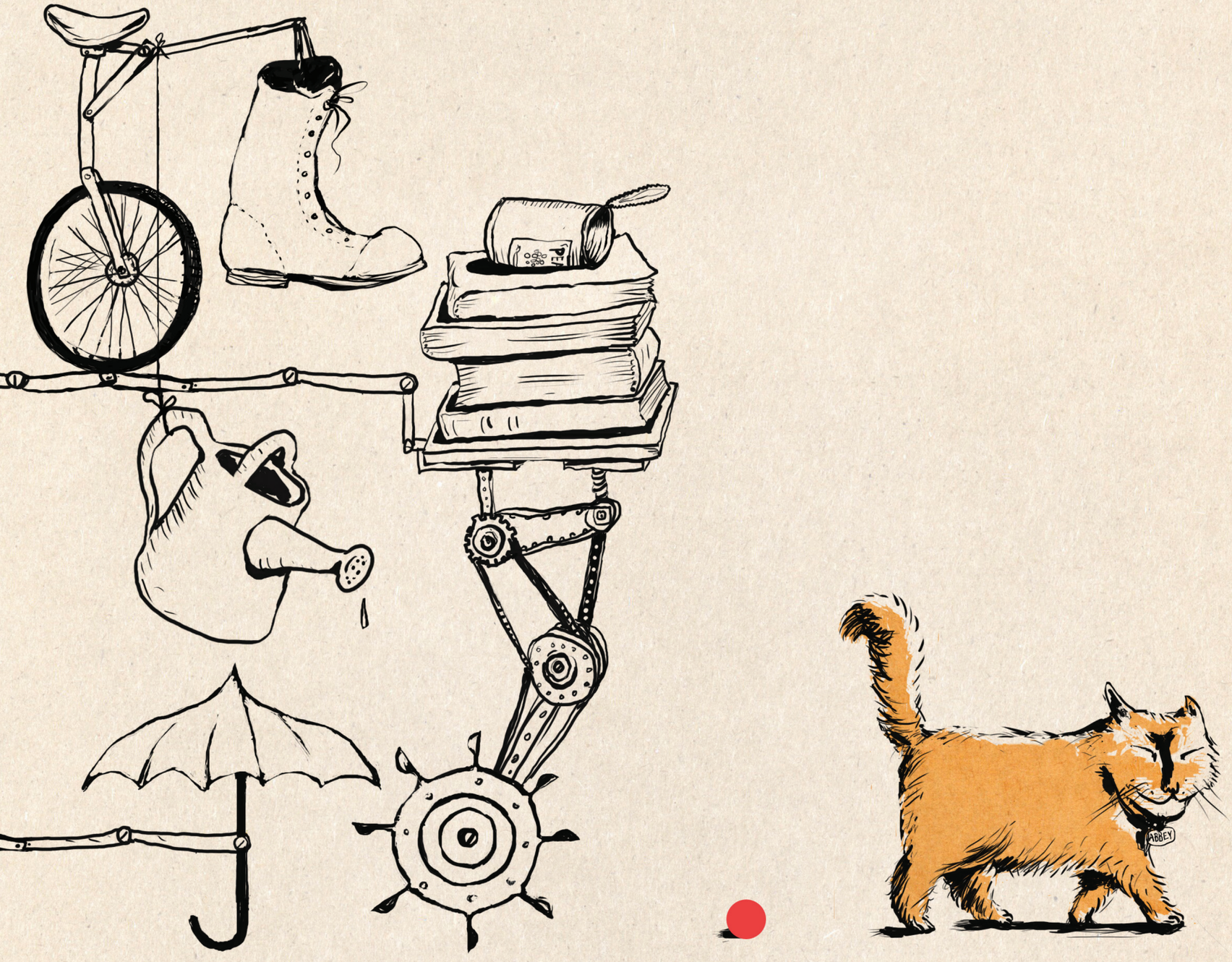
With this approach, you might hit the jackpot. But it's also not particularly difficult to find patterns and associations that mean nothing.

Consider what some jokesters revealed using data compiled by the U.S. Department of Agriculture, the Centers for Disease Control and Prevention, and the like: The per capita consumption of cheese has a 95 percent correlation with the number of people who died by becoming tangled in their bedding. The divorce rate in Maine has a staggering 99 percent correlation with per capita consumption of margarine. Such spurious associations aren't fabrications. (See an array of head-scratching graphs compiled using publicly available datasets on tylervigen.com.)

The trouble, as anyone who's taken an introductory statistics class knows, is that correlation does not equal causation.

To find meaning amid the morass, the CCD team is building computational tools to reveal networks of causal relationships from big biological data. As with any process of innovation, they'll have to jump their share of hurdles, from protecting the privacy of patients whose cases appear in datasets to developing new techniques for optimizing computation speed and efficiency. They're also working out theories to guide the crafting of algorithms that can sift through more data points and variables than mere mortals can keep in mind. What's really unique about their enterprise is how they are deploying

**It's not surprising that researchers would turn to computers to help them discover connections within data. But this team is actually asking computers to, effectively, play a collaborative role in the process of scientific inquiry.**




**What's disrupting that pathway? The huge number of variables in health can make the work of finding meaningful relationships in biomedical data astonishingly complex.**

artificial intelligence and machine learning techniques to generate likely hypotheses for scientists to test.

We can imagine that science will always rely on clever, well-read researchers with keen instincts, yet with millions of data points accumulating and overall funding for basic research plummeting, investigators have little time to waste. "We could be a lot more efficient in terms of how we store and share data and how we analyze it," says Cooper, a Pitt professor of biomedical informatics and intelligent systems. "This is a very focused effort on trying to make that whole effort of going from datasets to insight and knowledge much more efficient."

Founded with an $11 million, four-year grant, the CCD is one of 11 centers nationwide established as part of the National Institutes of Health's Big Data to Knowledge (BD2K) initiative. Each enterprise has a unique agenda. Pitt and CMU are committed to causality, Stanford to data annotation and retrieval efforts, Harvard to patient-centered data collation, and so on.

When Cooper decided to apply for BD2K funds in 2013, he had a deep well of talent from which to draw. Cooper arrived at Pitt in 1990, when he joined an early group of biomedical informatics researchers. He came, in part, for the chance to work with Glymour, a scholar of probability, machine learning, and causal discovery (who then had an adjunct appointment in Pitt's Department of History and Philosophy of Science). Pitt's Ivet Bahar, a PhD Distinguished Professor, who holds the JK Vries Chair of Computational and

Systems Biology, and Berg are principal investigators for the center.

In the realm of cause and effect in biomedicine, the CCD cast is a dream team.

Bahar founded what's now called Pitt's Department of Computational and Systems Biology in 2004. She's among the leaders of a new field that's starting to make sense of dynamic, complex interactions within the human body—events obscured when investigators focus on a single variable. Berg is founding director of UPMC's Institute for Personalized Medicine and the University's associate vice chancellor for science strategy and planning for the health sciences. (He is also a former director of the National Institute of General Medical Sciences at the NIH.)

For his part, Glymour has devoted more than three decades to pondering theories of knowledge—including what constitutes compelling evidence (both Freudian psychology and Einstein's theory of general relativity have captured his gaze). In the '80s, he and CMU colleagues Peter Spirtes and Scheines developed TETRAD, a software program that can generate causal inferences (now in its fifth iteration).

So how do these big thinkers think science should work?

## Exhibit A: Cancer

Let's go back to p53 and use it as a case study. PubMed lists more than 79,000 papers in the scientific literature that mention the protein. Libraries of biological samples, genomic sequences, and tumor imaging are widely available. Electronic medical records document the clinical trajectory of individual patients with and without mutated forms of p53. Wet-lab analyses reveal ever greater detail about the protein's (dys)function. "In the old days, the data just wasn't available," says Adrian Lee, a Pitt PhD professor of pharmacology and chemical biology. "Now people have to share the data. There are thousands and thousands of [experimental results available]. It's like a gold mine for computational biologists."

But despite decades of efforts, notes Lu, molecular therapy directly targeting mutant p53 has remained elusive.

And that's just one gene. Any given tumor might have as many as 10,000 glitches scattered throughout hundreds of genes within its genome.

The majority of those coding errors are irrelevant—passengers, if you will—whereas a few key mutations drive a tumor's growth.

"There is an urgent need," says Lu, to develop methods to find cancer drivers.

Yet investigators tend to simplify their experimental models by considering each mutation in isolation. That's like focusing on a single vehicle for insights into what's happening during a Los Angeles rush hour. The reality is vastly more complicated on the roadways and in the human body. Rogue genes frequently exhibit a diabolical synergy, amplifying one another's effects and silencing healthy genes that might lessen the damage. Compound the layers of variability embodied within a single patient by the millions of people with cancer, and the problem of identifying relevant targets and developing tailored treatments fast becomes a conceptual and computational quagmire.

When cancer biologist Lee arrived in Pittsburgh in 2010 to direct the new Women's Cancer Research Center at the University of Pittsburgh Cancer Institute, he'd never heard of causal modeling. But for more than a decade, he'd been compiling huge datasets generated by the genomic sequencing of breast tumors; he knew there had to be a better way to make sense of the relationships between genetic codes and disease states in the quest for targeted therapies.

To find that method, he partnered with Lu. The work of rethinking science might seem pretty darn abstract, yet Lee and Lu share a decidedly practical bent. A native of Shanghai, Lu trained as an emergency physician and finished a master's degree in cardiology before traveling to the United States to pursue a PhD in pharmacology, advanced study in artificial intelligence and computer science, and a Pitt certificate in bioinformatics. Says Lu: "I have 10 years' experience seeing patients, and I really want whatever I work on to be as close to the clinical setting as possible."

Lu and Lee were determined to craft an algorithm to help them detect relevant trends. Their first step, says Lee, was bringing together collaborators with expertise in fields that have traditionally worked independently. Algorithm development took months of fine-tuning before the team even thought of designing a wet-lab experiment to test any of the hypotheses emerging.

"We have these meetings, and Xinghua [Lu] says, *Hey, we did this.* We say, *Did you filter for this weird feature of this gene?*" says Lee. "And after they generate the algorithm, we test it; and it's right, or wrong, or half-right; and then they refine it." Once everyone agrees that the algorithm might be getting close, they apply it to additional datasets or design an experiment for the wet lab to test the suppositions. Says Lee: "It's an iterative process."

(He makes it all sound so polite. Glymour sees the approach of his separate research group like this: "We yell at each other, work out counterexamples, try simulations.")

Lu notes that the multidisciplinary environment of the CCD has allowed the team to reveal insights into p53 derived from a computer-generated hypothesis. By slicing and dicing details from two national datasets—one containing genomic information, the other with clinical details—they were able to identify perturbed signaling pathways that affect patient outcomes and zero in on a particular signaling complex triggered by the mutual activation of multiple mutated genes (including *TP53*).

After their algorithm generated its hypothesis about the signaling complex, the team designed a series of experiments in the wet lab to probe the computer's theory. By disrupting the signaling complex in the lab, they inhibited the cancer from growing.

It seems the computer was right—and because the same signaling complex becomes muddled in dozens of other types of cancer, including ovarian, lung, and esophageal, the team has a head start on identifying similar targets in a significant proportion of cancers afflicting other parts of the body.

## Test Drives

This is the name of the new game—the computer as a collaborator.

And with this productive colleague, CCD investigators have identified a few biomedical challenges on which to test-drive their formulations. There's the cancer signaling pathways group, which generated the p53 study; headed by Lu, it's seeking targets for treatment. An

fMRI group, led by Glymour, is identifying causal influences among brain regions by digging into functional magnetic resonance imaging data. A lung group, led by Pitt's Takis Benos, a PhD professor of computational and systems biology with joint appointments in biomedical informatics and computer science, means to detect the cellular factors that lead to chronic obstructive pulmonary disease and idiopathic pulmonary fibrosis.

"One of the things that we've found, perhaps not surprisingly," says Cooper, "is that when you start applying an algorithm to a particular domain—understanding cancer signaling pathways, for example—there are details that you need to attend to that are special to that area. The algorithms are modified to adapt to that area and the kind of data that's common in that area."

A fourth effort, led by Pitt's Jeremy Espino, an MD and director of information technology and open source software development for the Department of Biomedical Informatics, is slated to run for one year in partnership with scholars at Harvard. The team is establishing a cloud-based tool with which investigators anywhere in the world will be able to manipulate a massive autism dataset (on a server farm run by Amazon).

"This is a relatively new thing that's happening in research," says Espino. "Datasets are so large that it's difficult to even have enough computational resources to digest the information."

Like the other three projects, the Pitt-Harvard autism team will develop a proof of concept, then share their techniques so similarly complex datasets can be made more broadly accessible.

## Good Old-Fashioned Good Assumptions

CCD investigators constantly balance their pursuit of clinically germane insights with the imperative to design computational processes elegant enough to crunch through relevant data in a timely fashion.

If the balance shifts too far in favor of computational elegance, biomedical relevance plummets.

At a conference this fall, says Lee, a computer scientist presented a compelling technique for mining biomedical data, but a basic biological assumption he'd embedded within his algorithm was just plain wrong.

"The assumptions are often really, really naive," says Lee. "In that case, the first assumption was so utterly flawed, the outcome would have to be wrong. But without the assumption, you can't run the algorithm."

Without the input of biomedical colleagues, says Glymour, who's tackled the challenge of modeling everything from climate science to long-range prediction of forest fires, even mineral identification for a NASA robot, he's sure the algorithms he helps the fMRI group craft would fast collapse under their own weight. "My biological collaborators keep me from doing incredibly stupid things," he says, tongue only slightly in cheek.

The CCD's approach has investigators cross-checking their assumptions throughout. "Xinghua knows the underlying statistical and mathematical models," says Lee, "and we can give input to make sure we don't generate spurious findings."

In their ongoing work, Lu, Cooper, and Lee are concentrating further on algorithms related to driver mutations.

"Say you have 100 genes, each in the same pathway, and in each of 100 patients, that pathway is differently mutated," says Lee. "But if they all affect the same output, you'd never notice."

Imagine, again, that LA traffic metaphor. There are many reasons why a midtown intersection may have recently started experiencing frequent pileups: nearby road construction or a recently opened office building, perhaps.

In effect, Lu's algorithm detects the vexed midtown intersection—the cellular signaling pathways where those mutations' effects converge. So instead of focusing on the many different causes of traffic jams that occur at the crossroads, engineers can modify the intersection to bring traffic flow back to normal.

"The majority of the mutated genes—you have no way to treat them," says Lu. "There are not so many drugs to target particular mutated genes. Even if you had those drugs, you could only help 1 or 2 percent of patients. Target a particular pathway, and you might help 20 percent of patients."

Says Lee: "Xinghua's network view is more of a systems approach, and it's a huge advance."

## Sharing the Gospel

To speed discovery, the CCD is bringing more researchers into the fold. Last June, the center offered its first in-person training. The free, four-day course introduced 75 investigators from around the world to the concepts and software necessary to pursue causal discovery in their own research. Throughout the past year, the CCD has expanded its online offerings with recorded lectures, papers, and software tutorials.

"For this to be useful to a wider population of biomedical researchers," says Scheines, who leads the training effort, "we need to develop training materials for biomedical scientists who want to use data science approaches like ours and for data scientists who want to apply our techniques to biomedicine. We need to train both communities fairly broadly."

**"You want to know what experiments are worth doing out of the millions that could be done."**

And, Scheines says, "We really want to get a community set up where people can work together."

Consequently, software developers throughout the CCD are making accessibility a top priority.

"Everyone has their own workflow, statistical packages, computational packages they use," says Scheines. "We need to make our materials easy to fold into their processes if they're going to be widely adopted."

"Causal discovery of biological networks from data—already nearly three decades in the making—has the capacity to advance science now," says Cooper. Glymour assumes a curmudgeonly tone. "I began this work in 1980," he says. "It's taken 35 years of just repeating this strategy, publishing results, for the NIH to pay attention. The turnover is not fast.

"It takes a while for people to absorb radical new methods, even with proof that they work." ■