

# Brothers create computer data sifter

By Carolyn Y. Johnson | GLOBE STAFF DECEMBER 15, 2011



CHIEYU LIN, COURTESY OF PARDIS SABETI

**Brothers David, right, and Yakir, left, Reshef developed the new statistical tools under the guidance of professors from Harvard University and the Broad Institute.**

It is an unusual starting point for a high-profile paper in a leading science journal: Two brothers, students a year apart at universities down the Charles River from one another, decide to work together on a summer project. The research unfolds through ideas scribbled on the walls of a laboratory,

insights gained during downtime working as an emergency medical technician, and brainstorming shared at a fraternity house in Boston.

Yesterday, the influential journal *Science* published the fruits of that labor: the creation of a powerful computer program that rapidly flags patterns and identifies correlations in huge databases, from sports statistics to online social networks to the genomes being churned out by science laboratories.

While it is rare for two brothers in their mid-20s to share credit as the lead authors of a paper, the achievement demonstrates how creativity often arises from the back-and-forth of a team, in this case, David and Yakir Reshef, who have been collaborating since childhood.

“I think, in some sense, David and I have been roping each other into things for our entire lives,” said Yakir, 24 and a Fulbright scholar at the Weizmann Institute of Science in Israel.

The summer after his senior year at MIT, David began working with Pardis Sabeti, a biologist at the Broad Institute who had an interest in global health. David was developing an approach to sift through large, international health data sets, highlighting potential relationships between demographic information and the incidence of infectious diseases, such as cholera or HIV.

“We just wanted a simple way to figure out what was in the data sets,” said David, 25, who is pursuing a dual degree in the Harvard-MIT Division of Health Sciences and Technology. “At first we thought we would go find some methods that existed. It turned out to be a much more complicated question to answer.”

As the research advanced, David began to get excited. He saw a potentially big opportunity to develop tools that could rapidly and effectively identify all sorts of complex patterns hidden in data, ranging from the rise and fall of flu cases depending on the season to the swooping curve of female obesity when

graphed against average income.

He turned to his longtime collaborator, Yakir, a Harvard undergraduate who was working that summer on an ambulance crew based in Arlington, a job that he hoped would help prepare him for medical school.

The brothers had always been close. They lived in Israel and Kenya before their family moved to the United States, and in elementary school after coming to this country, they would sometimes speak to each other in Hebrew, using it as their private language.

When one was interested in something, he would get his brother interested. In his middle school years, Yakir became engrossed in computer programming from hanging out at his uncle's software startup during summer visits to Israel. He told his brother how much fun it was, and when David got to high school, he took a programming class and got hooked, too.

In high school, they spent as much time as possible in the computer lab, working together. But even the short commute between Harvard and MIT had allowed them to grow apart a bit.

“In some ways, this was a return to the good old days,” Yakir said. “This is a pattern with us. I was a little bit skeptical at the beginning.

“I credit him with having the thinking, ‘Even though it's crazy, we should try it.’ . . . He hatches harebrained schemes.”

So Yakir began bringing scientific articles to his summer job. Between calls, he would sit in the ambulance and think about the statistics and large data sets, a birds-eye view of broad trends in health data mixed with medicine in action.

They found that to solve the problem, they had to draw from all corners, expanding beyond global health. They worked with computer scientist

Michael Mitzenmacher at Harvard. A key insight came one night when the brothers were talking over Skype from David's fraternity to Mitzenmacher.

Over the years, academic scholarships put the key team members temporarily on different continents. The collaboration intensified over Skype, eventually resulting in a new statistical tool presented yesterday that rapidly and effectively mines data for relationships and ranks the strongest ones, without any preconceived notion of what that relationship might be.

The computer program they built cannot answer the question of whether one thing caused another, but by finding the strongest correlations, it can help scientists generate new hypotheses and questions to explore.

The brothers - both baseball fans, although Yakir says David is the better athlete - tried their tool on statistics and salaries from Major League Baseball. They found that hits, total bases, and a statistical measure of offensive performance were most strongly correlated with salary.

They also tested their tool on data from the World Health Organization, yeast gene activity, and genomic data describing the bacteria present in the human gut and found relationships not picked up with older methods.

Eli Upfal, a professor of computer science at Brown University who was not involved in the study, said the new tool has a solid mathematical foundation and worked well on real data, but its ultimate impact will be seen over time.

"It's not like someone solved an open problem in mathematics and you can check the proof," Upfal said. "This is more: Here is a tool and here's some very good mathematical justification for the tool, but the proof eventually is in it being adopted and shown to be practical. . . . This is the first step."

Sabeti said the team plans to extend and build on the tool, for example finding ways to look for complex relationships between more than two pieces of data.

“Every field is ripe for a tool like this, with the data deluge,” she said.

“Everywhere we go and give talks about it, somebody says, ‘I have a data set for you to look at,’ finance, sports, statistics.”

The researchers are making the tool available on a website. The brothers said they hope to work together again soon.

Introducing BostonGlobe.com digital subscriptions, just 99¢ for your first 4 weeks.

*Carolyn Y. Johnson can be reached at [cjohnson@globe.com](mailto:cjohnson@globe.com). Follow her on Twitter [@carolynjohnson](https://twitter.com/carolynjohnson).*