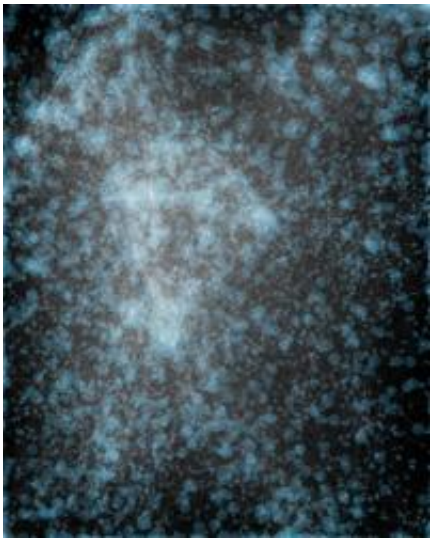
Tool detects patterns hidden in vast data sets

December 15 2011



This graphic depicts the top 0.25 percent of the relationships that the researchers' techniques found in data on the concentration of microbes in the human gut. Image courtesy of David Reshef

Researchers from the Broad Institute and Harvard University have developed a tool that can tackle large data sets in a way that no other software program can. Part of a suite of statistical tools called MINE, it can tease out multiple patterns hidden in health information from around the globe, statistics amassed from a season of major league baseball, data on the changing bacterial landscape of the gut, and much more. The researchers report their findings in a paper appearing in the December 16 issue of the journal *Science*.

From Facebook to physics to the global economy, the world is filled with data sets that could take a person hundreds of years to analyze by eye. Sophisticated computer programs can search these data sets with great speed, but fall short when researchers attempt to even-handedly detect different kinds of patterns in large data collections.

"There are massive data sets that we want to explore, and within them, there may be many relationships that we want to understand," said Broad Institute associate member Pardis Sabeti, senior author of the paper and an assistant professor at the Center for Systems Biology at Harvard University. "The human eye is the best way to find these relationships, but these data sets are so vast that we can't do that. This toolkit gives us a way of mining the data to look for relationships."

The researchers tested their analytical toolkit on several large data sets, including one provided by Harvard colleague Peter Turnbaugh who is interested in the trillions of microorganisms that live in the gut. Working with Turnbaugh, the research team harnessed MINE to make more than 22 million comparisons and narrowed in on a few hundred patterns of interest that had not been observed before.

"The goal of this statistic is to take data with a lot of different dimensions and many possible correlations and pick out the top ones," said Michael Mitzenmacher, a senior author of the paper and professor of computer science at Harvard University. "We view this as an exploration tool – it can find patterns and rank them in an equitable way."

One of the tool's greatest strengths is that it can detect a wide range of patterns and characterize them according to a number of different parameters a researcher might be interested in. Other statistical tools work well for searching for a specific pattern in a large data set, but cannot score and compare different kinds of possible relationships.

MINE, which stands for Maximal Information-based Nonparametric Exploration, is able to analyze a broad spectrum of patterns.

"Standard methods will see one pattern as signal and others as noise," said David Reshef, a co-first author of the paper who is currently a graduate student in the Harvard-MIT Health Sciences and Technology program and also worked on this project as a graduate student in the department of statistics at the University of Oxford. "There can potentially be a variety of different types of relationships in a given data set. What's exciting about our method is that it looks for any type of clear structure within the data, attempting to find all of them."

Not only does MINE attempt to identify any pattern within the data, but it also attempts to do so with an eye toward capturing different types of patterns equally well. "This ability to search for patterns in an equitable way offers tremendous exploratory potential in terms of searching for patterns without having to know ahead of time what to search for," said David Reshef.

MINE is especially powerful in exploring data sets with relationships that may harbor more than one important pattern. As a proof of concept, the researchers applied MINE to social, economic, health, and political data from the World Health Organization (WHO) and its partners. When they compared the relationship between household income and female obesity, they found two contrasting trends in the data. Many countries follow a parabolic rate, with obesity rates rising with income but peaking and tapering off after income reaches a certain level. But in the Pacific Islands, where female obesity is a sign of status, countries follow a steep trend, with the rate of obesity climbing as income increases.

"Many data sets will contain these types of complicated relationships that are guided by multiple drivers," said Sabeti. MINE is able to identify these. "This greatly extends our capability to find interesting

relationships in data."

Researchers can use MINE to generate new ideas and connections that no one has thought to look for before.

"Our tool is a hypothesis generator," said Yakir Reshef, a co-first author of the paper and a Fulbright scholar at the Weizmann Institute of Science. "The standard paradigm is hypothesis-driven science, where you come up with a hypothesis based on your personal observations. But by exploring the data, you get ideas for hypotheses that would never have occurred to you otherwise."

In addition to testing the ability of the suite of tools to detect patterns in biological and health data, the researchers examined data collected from the 2008 baseball season.

"One question that we thought would be particularly interesting would be to see what things were most strongly associated with salary," said David Reshef. The researchers generated a list of relationships, finding that the strongest associations with salary were hits, total bases, and an aggregate statistic that reflects how many runs a player generated for a team. "Given the stakes, baseball is so well documented. We're curious to see what can be done in this realm with tools like MINE."

Researchers from many different fields, including systems biology, computer science, statistics, and mathematics, all contributed to this project. "People are getting better at combining data from different sources, and in some ways, this project is in the spirit of that," said Yakir Reshef. "The project brought together authors from many disciplines. It symbolizes the kind of collaborations that we hope people will use this for in the future."

More information: Reshef, DN et al., Detecting novel associations in

large data sets, *Science*, DOI: 10.1126/Science1205438