



CS Bits & Bytes is a bi-weekly newsletter highlighting innovative computer science research. It is our hope that you will use CS Bits & Bytes to engage in the multi-faceted world of computer science to become not just a user, but a creator of technology. Please visit our website at: <http://www.nsf.gov/cise/csbytes>.

May 7, 2012

Volume 1, Issue 11

Big Data

What's the Big Deal about Big Data? Last year, people around the world stored enough data to fill 60,000 Libraries of Congress. YouTube claims they receive 24 hours of video every minute. And, when the Sloan Digital Sky Survey came online in 2000, it collected more data in its first few weeks than had been amassed in the entire history of astronomy!

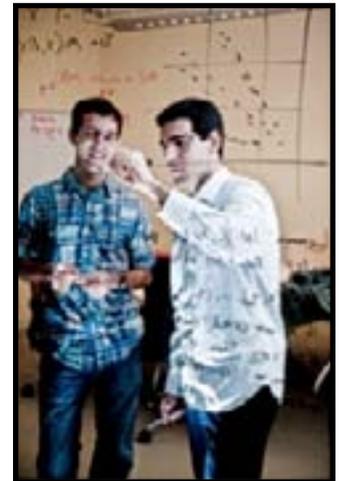


If researchers printed on paper each potential relationship in a recent data set containing abundance levels of bacteria in the human gut, the stack of paper would reach to a height of 1.4 miles, 6 times the height of the Empire State Building! Credit: Sigrid Knemeyer

Data are being accumulated at unprecedented rates and complexity. They are collected by sensors monitoring all around us (the environment, critical infrastructure such as bridges and smart grids, and even our homes), our pervasive use of the Internet (emails, images, videos, etc.), and modern experimental and observational studies.

"Big Data" is characterized not only by the enormous **volume** of data but also by the **variety** of those data and the **velocity** of its generation.

Computer science provides the tools to **collect, store, manage, analyze, and visualize** these large-scale and complex data sets to gain new insights, recognize relationships, and make increasingly accurate predictions. A new tool recently developed by researchers from the Broad Institute and Harvard University, with support from the National Science Foundation, can uncover patterns in large data sets in a way that no other software program can. Sophisticated computer programs search data sets with great speed and work well for searching for a specific pattern in a large data set, but have difficulty in detecting, scoring, and comparing different kinds of possible relationships in large data collections. Yet, this new tool, called



Brothers David Reshef and Yakir Reshef developed MIC under the guidance of professors from Harvard University and the Broad Institute. Credit: ChieYu Lin

Maximal Information Coefficient or MIC, can tease out multiple patterns from various data sets - health information from around the globe, the changing bacterial landscape of the gut, and even statistics amassed from a season of competitive sports.

The ability to gain new insights - **to move from data to knowledge to action** - has tremendous potential to transform how we live. It will drive discovery and decision-making in almost all aspects of our lives. Knowledge bases that enable biomedical discovery, more accurate diagnoses, and patient-centered therapy will revolutionize healthcare. Access to data is already transforming traditional business models - improving efficiency of operations and creating new opportunities to maintain the U.S. as a global leader. Forecasting and high-resolution models will improve our ability to manage and protect our increasingly stressed ecosystems and environment. And, new dynamic tools and analytics in educational settings will help to improve student performance and learning.



Image of Professor Daphne Koller.

Spot Light! Professor Daphne Koller is a Professor of Computer Science at Stanford University using data to observe and understand human learning in order to understand what learning strategies are more effective, and for whom. She is the co-founder of Coursera, a social entrepreneurship company that uses technology to bring quality higher education from top institutions to everyone around the world, for free. Dr. Koller was born in Jerusalem, Israel and came to the U.S. for her PhD at Stanford University. She likes to travel with her family and has visited over 50 countries.

Links:

Read more about MIC in the news from NSF (http://www.nsf.gov/news/news_summ.jsp?cntn_id=122597&org=CISE&from=news), the Broad Institute (<http://www.broadinstitute.org/news/3784>), Scientific American (<http://blogs.scientificamerican.com/observations/2011/12/16/how-to-find-meaning-in-a-maelstrom-of-data/>), and MIT news (<http://web.mit.edu/newsoffice/2011/large-data-sets-algorithm-1216.html>).

Read more about Dr. Koller at: <http://ai.stanford.edu/~koller/index.html>.

To learn more about Coursera, visit: <https://www.coursera.org/>.

Activity:

The short paragraphs below provide scenarios to generate discussions about data relationships and use of data in decision-making.

Discussion Topic 1: Scientists use big data to find the strength of relationships among data points. For example, if you look at measurements from physicals, you might find that there is a relationship between height and weight. As height increases, weight also tends to increase. This relationship may not hold for all people (e.g., there can be really light tall people and vice versa), yet a positive correlation between the two variables prevails. Researchers previously believed that there was a negative correlation between age and IQ. Recently, this correlation has turned out to be much weaker than we originally thought. Describe what is meant by a negative correlation between age and IQ. (from http://www.mhhe.com/socscience/sociology/statistics/stat_cor.htm).

Discussion Topic 2: The most commonly used measure of dependence between two variables is the Pearson correlation coefficient. Pearson can detect a linear relationship between two variables. What does it mean for a function to be linear? What would that imply for a relationship between two variables?

Discussion Topic 3: Brothers David and Yakir Reshef developed the Maximal Information Coefficient (MIC), which can efficiently detect non-linear relationships in millions of pairs of variables. The researchers rank their program on statistics from the 2008 baseball season to detect possible relationships among the data. Using MIC, the statistics most related to a hitter's salary were, in order, RPMLV: a measure of how many runs a player produced for his team (<http://www.baseballprospectus.com/glossary/index.php?mode=viewstat&stat=148>), hits, and total bases. Pearson ranked those three statistics, respectively, as the 14th, 37th, and 20th most closely related to a player's salary. Instead, Pearson ranked walks as the most related to a hitter's salary. If you were the owner, would you choose to pay the player with the most walks or hits? Why?

Further Exploration: If you have some experience running applications from the command line, you can experiment with MINE with datasets from baseball, social indicators, and genetics at <http://www.exploredata.net>.

CS BITS & BYTES

<http://www.nsf.gov/cise/csbytes/>

Please direct all inquiries to: csbitsandbytes@nsf.gov

National Science Foundation
Computer & Information Science & Engineering Directorate

4201 Wilson Blvd Suite 1105

Arlington VA, 22230

