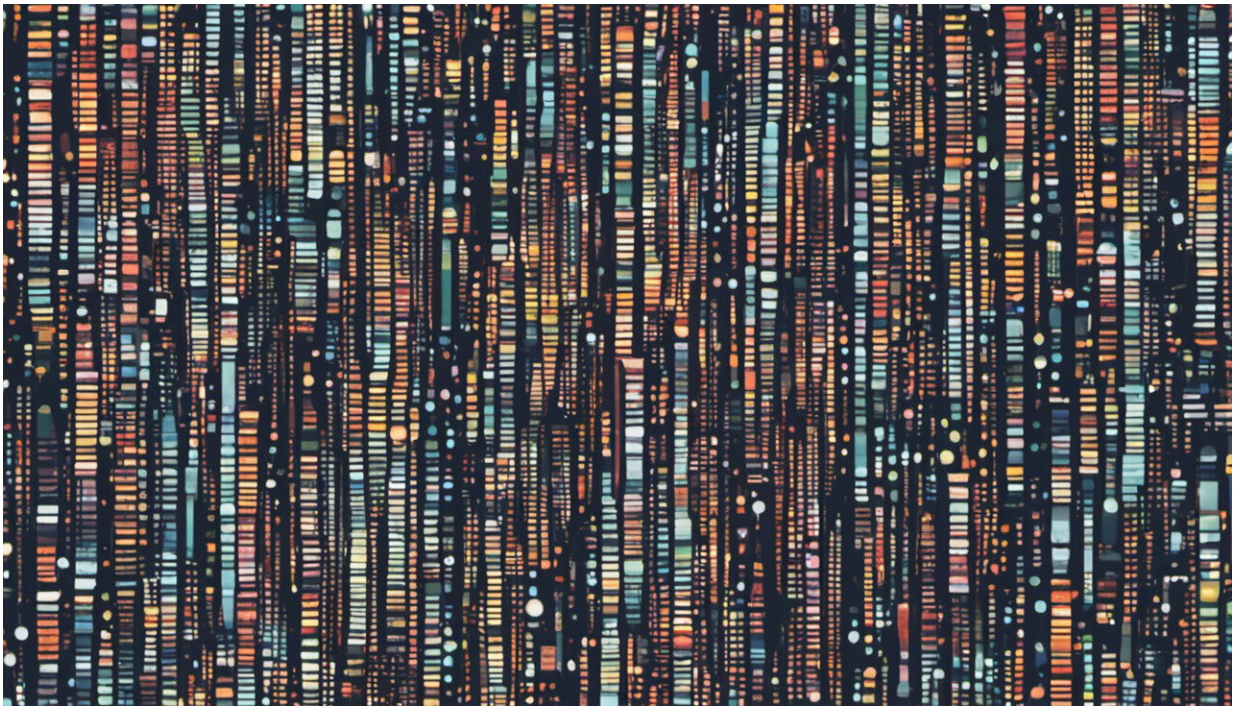
Researchers propose a better way to make sense of 'Big Data'

February 18 2014



Credit: AI-generated image (disclaimer)

Big Data is everywhere, and we are constantly told that it holds the answers to almost any problem we want to solve. Companies collect information on how we shop, doctors and insurance companies gather our medical test results, and governments compile logs of our phone calls and emails. In each instance, the hope is that critical insights are hidden

deep within massive amounts of information, just waiting to be discovered.

But simply having lots of data is not the same as understanding it. Increasingly, new mathematical tools are needed to extract meaning from enormous data sets. In work published online today, two researchers at Cold Spring Harbor Laboratory (CSHL) now challenge the most recent advances in this field, using a classic mathematical concept to tackle the outstanding problems in Big Data analysis.

What does it mean to analyze Big Data? A major goal is to find patterns between seemingly unrelated quantities, such as income and cancer rates. Many of the most common statistical tools are only able to detect patterns if the researcher has some expectation about the relationship between the quantities. Part of the lure of Big Data is that it may reveal entirely new, unexpected patterns. Therefore, scientists and researchers have worked to develop statistical methods that will uncover these novel relationships.

In 2011, a distinguished group of researchers from Harvard University published a highly influential paper in the journal Science that advanced just such a tool. But in a paper published today in *Proceedings of the National Academy of Sciences*, CSHL Quantitative Biology Fellow Justin Kinney and CSHL Assistant Professor Gurinder "Mickey" Atwal demonstrate that this new tool is critically flawed. "Their statistical tool does not have the mathematical properties that were claimed," says Kinney.

Kinney and Atwal show that the correct tool was hiding in plain sight all along. The solution, they say, is a well known mathematical measure called "mutual information," first described in 1948. It was initially used to quantify the amount of information that could be transmitted electronically through a telephone cable; the concept now underlies the

design of the world's telecommunications infrastructure. "What we've found in our work is that this same concept can also be used to find patterns in data," Kinney explains.

Applied to Big Data, mutual information is able to reveal patterns in large lists of numbers. For instance, it can be used to analyze patterns in data sets on the numerous bacterial species that help us digest food. "This particular tool is perfect for finding patterns in studies of the human microbiome, among many other things," Kinney says.

Importantly, mutual information provides a way of identifying all types of patterns within the data without reliance upon any prior assumptions. "Our work shows that mutual information very naturally solves this critical problem in statistics," Kinney says. "This beautiful mathematical concept has the potential to greatly benefit modern data analysis, in biology and in biology and many other important fields.

More information: "Equitability, mutual information, and the maximal information coefficient" appears online in PNAS on February 17, 2014. The authors are: Justin Block Kinney and Gurinder Singh Atwal. The paper can be obtained online at: www.pnas.org/content/early/201 … /1309933111.abstract

Provided by Cold Spring Harbor Laboratory

Citation: Researchers propose a better way to make sense of 'Big Data' (2014, February 18) retrieved 19 September 2024 from https://phys.org/news/2014-02-big.html