PHYS.ORG

# Bombarded by explosive waves of information, scientists review new ways to process and analyze Big Data

August 26 2014

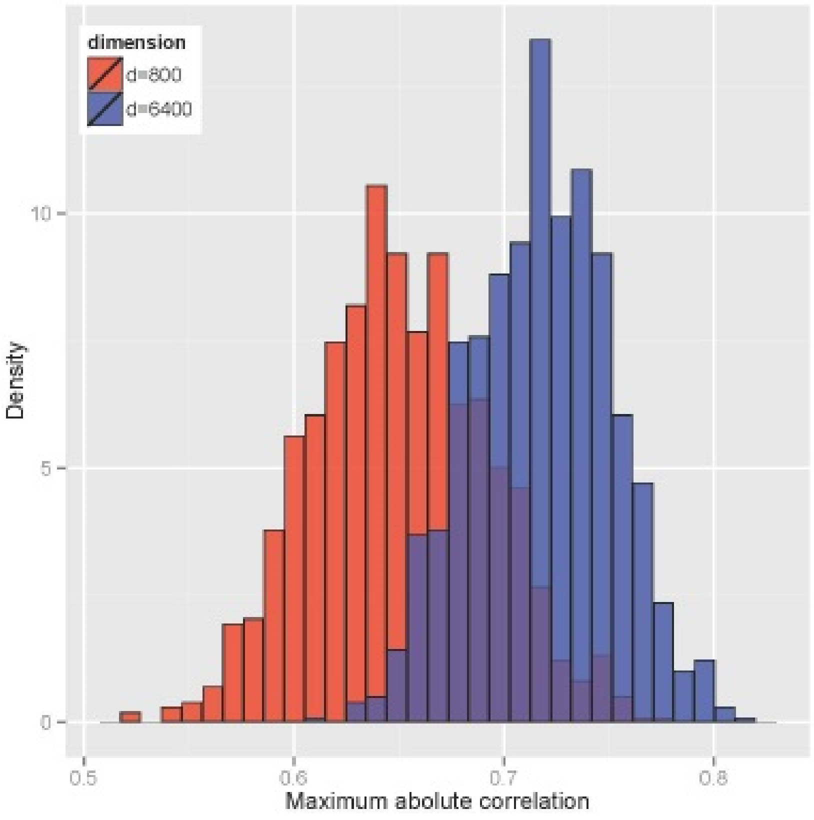

Illustration of spurious correlation by showing the distribution of the maximum absolute sample correlation coefficients between the first and the four of the rest of 800 (in red) and 6,400 (in blue) independently drawn standard Gaussian random variables with sample size n = 60. It can be seen that the maximum spurious correlation coefficient is very high. Credit: Science China Press

Big Data presents scientists with unfolding opportunities, including, for instance, the possibility of discovering heterogeneous characteristics in the population leading to the development of personalized treatments and highly individualized services. But ever-expanding data sets introduce new challenges in terms of statistical analysis, bias sampling, computational costs, noise accumulation, spurious correlations, and measurement errors.

The era of Big Data – marked by a Big Bang-like explosion of information about everything from patterns of use of the World Wide Web to individual genomes – is being propelled by massive amounts of very high-dimensional or unstructured data, continuously produced and stored at a decreasing cost.

"In genomics we have seen a dramatic drop in price for whole genome sequencing," state Jianqing Fan and Han Liu, scientists at Princeton University, and Fang Han at Johns Hopkins. "This is also true in other areas such as social media analysis, biomedical imaging, high-frequency finance, analysis of surveillance videos and retail sales," they point out in a paper titled "Challenges of Big Data analysis" published in the Beijing-based journal *National Science Review*.

With the quickening pace of data collection and analysis, they add, "scientific advances are becoming more and more data-driven and researchers will more and more think of themselves as consumers of data."

Increasingly complex data sets are emerging across the sciences. In the field of genomics, more than 500 000 microarrays are now publicly available, with each array containing tens of thousands of expression values of molecules; in biomedical engineering, tens of thousands of terabytes of functional magnetic resonance images have been produced, with each image containing more than 50 000 voxel values. Massive and high-dimensional data is also being gathered from social media, e-commerce, and surveillance videos.

Expanding streams of social network data are being channeled and collected by Twitter, Facebook, LinkedIn and YouTube. This data, in turn, is being used to predict influenza epidemics, stock market trends, and box-office revenues for particular movies.

The social media and Internet contain burgeoning information on consumer preferences, leading economic indicators, business cycles, and the economic and social states of a society.

"It is anticipated that social network data will continue to explode and be exploited for many new applications," predict the co-authors of the study. New applications include ultra-individualized services.

And in the area of Internet security, they add, "When a network-based attack takes place, historical data on network traffic may allow us to efficiently identify the source and targets of the attack."

With Big Data emerging from many frontiers of scientific research and technological advances, researchers have focused on the development of new computational infrastructure and data-storage methods, of fast algorithms that are scalable to massive data with high dimensionality.

"This forges cross-fertilization among different fields including statistics, optimization and applied mathematics," the scientists add.

The massive sample sizes giving rise to Big Data fundamentally challenge the traditional computing infrastructure.

"In many applications, we need to analyze Internet-scale data containing billions or even trillions of data points, which makes even a linear pass of the whole dataset unaffordable," the researchers point out.

The basic approach to store and process such data is to divide and conquer. The idea is to partition a large problem into more tractable and independent sub-problems. Each sub- problem is tackled in parallel by different processing units. On a small scale, this divide-and-conquer strategy can be implemented either by multi-core computing or grid computing.

On a larger scale, handling enormous arrays of data requires a new computing infrastructure that supports massively parallel data storage and processing.

The researchers present Hadoop as an example of a basic software and programming infrastructure for Big Data processing. Alongside Hadoop's distributed file system, they review MapReduce, a programming model for processing large datasets in a parallel fashion, cloud computing, convex optimization, and random projection algorithms, which are specifically designed to meet Big Data's computational challenges.

Hadoop is a Java-based software framework for distributed data management and processing. It contains a set of open source libraries for distributed computing using the MapReduce programming model and its own distributed file system called HDFS. Hadoop automatically facilitates scalability and takes cares of detecting and handling failures.

HDFS is designed to host and provide high-throughput access to large datasets that are redundantly stored across multiple machines. It ensures Big Data's survivability and high availability for parallel applications.

In terms of statistical analysis, Big Data presents another set of new challenges. Researchers tend to collect as many features of the samples as possible; as a result, these samples are commonly heterogeneous and high dimensional.

High dimensionality brings new problems, including noise accumulation, spurious correlation, and incidental endogeneity. For instance, high dimensionality gives rise to spurious correlation. In studying the association between cancers and certain genomic and clinical factors, it might be possible that prostate cancer is highly correlated to an unrelated gene. However, such a high correlation could be explained by high dimensionality: In studies that include so many features, ranging from genomic information to height, weight and gender to favorite foods and sports, some high correlations emerge merely by chance.

**More information:** Jianqing Fan, Fang Han, and Han Liu. "Challenges of Big Data analysis." *Natl Sci Rev* (June 2014) 1 (2): 293-314 nsr.oxfordjournals.org/content/1/2/293.full

Provided by Science China Press

Citation: Bombarded by explosive waves of information, scientists review new ways to process and analyze Big Data (2014, August 26) retrieved 27 April 2024 from https://phys.org/news/2014-08-bombarded-explosive-scientists-ways-big.html