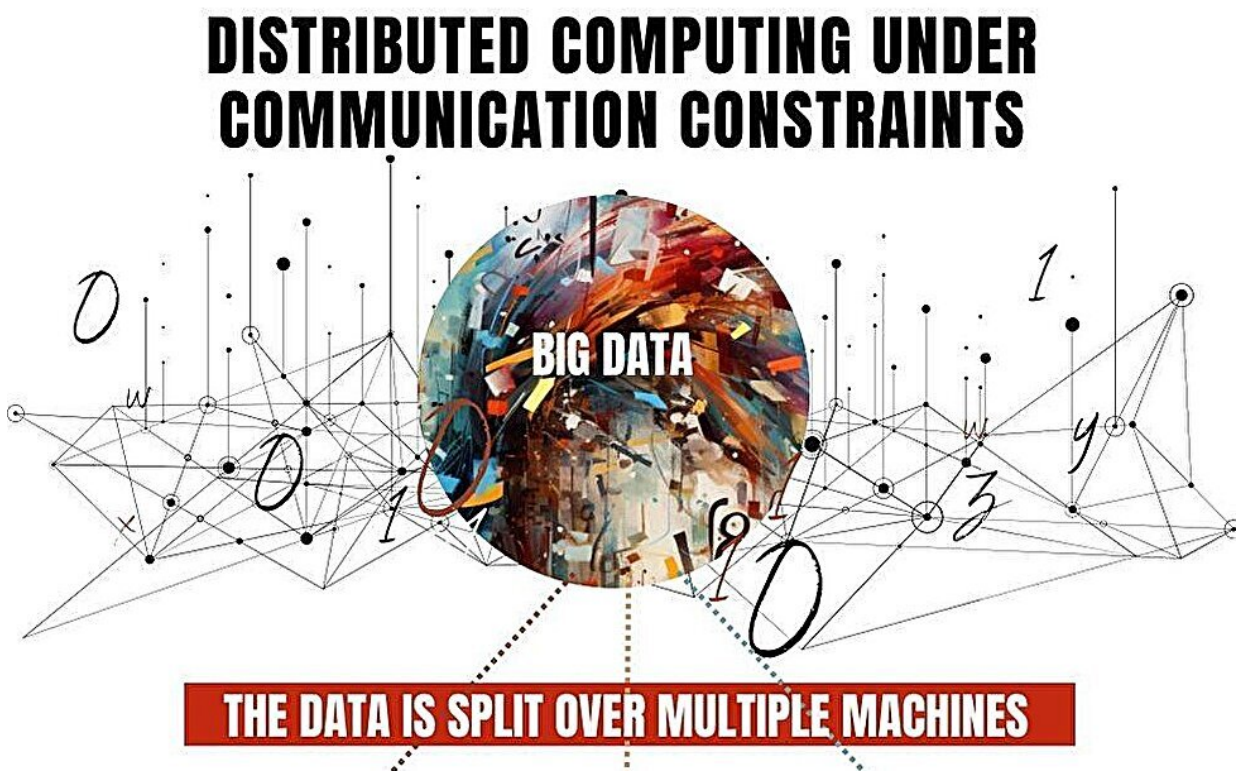
Improving accuracy, reliability and interpretability of distributed computing

October 4 2023, by Fabio Todesco



Credit: Weiwei Chen/Bocconi University

A new study by Botond Szabo (Bocconi Department of Decision Sciences) published in *The Annals of Statistics* lays the cornerstone for more accurate, reliable and interpretable distributed computing methods.

In the world of big data, when the need arises to estimate many parameters in very complex statistical models that make use of large amounts of available information, computation time becomes unsustainable even with the fastest supercomputers. One of the strategies developed to cope with the issue is distributed (or parallel) computing.

Data (or tasks, in some cases) are divided among many machines and only summary information (the results of computations) is sent to a central location, say a meteorological station, an astronomy observatory, or a traffic control system. This method also mitigates privacy concerns since most data don't have to be moved around.

In any case, even communicating only summary information between servers can be costly, so statisticians have borrowed from electric engineers the idea of bandwidth limitation. "The goal," says Professor Szabo, "is to minimize the flow of data, losing as little information as possible.

"Furthermore, parallel computing is often a black-box procedure, i.e., a procedure which transforms inputs into outputs in not-well-understood ways, and this makes results neither completely interpretable, nor reliable. Finding mathematical models which give theoretical underpinnings to such procedures would be desirable."

In his paper with Lasse Vuursteen (Delft University of Technology) and Harry van Zanten (Vrije Universiteit Amsterdam), Prof. Szabo derives the best tests to minimize the loss of information in a distributed framework where the data is split over multiple machines and their communication to a central machine is limited to a given quantity of bits.

In statistics, a test is a procedure that determines whether a hypothesis about a parameter is true and how much you can rely on this result. In other words, it quantifies uncertainty. When we read that a hypothesis is

"not statistically significant," it means that no evidence was found in the data to support the hypothesis.

"The tests we develop in the paper allow us to achieve the highest accuracy for a given amount of transmitted information or the minimum amount of information to be transmitted for a needed level of accuracy," explains Prof. Szabo.

The paper is a foundational work, using an idealized mathematical case, but Prof. Szabo is already working on more complex settings. "In the long-term," he says, "we could hopefully obtain more efficient communication algorithms, underpinned by theoretical guarantees."

More information: Botond Szabó et al, Optimal high-dimensional and nonparametric distributed testing under communication constraints, *The Annals of Statistics* (2023). DOI: 10.1214/23-AOS2269

Provided by Bocconi University