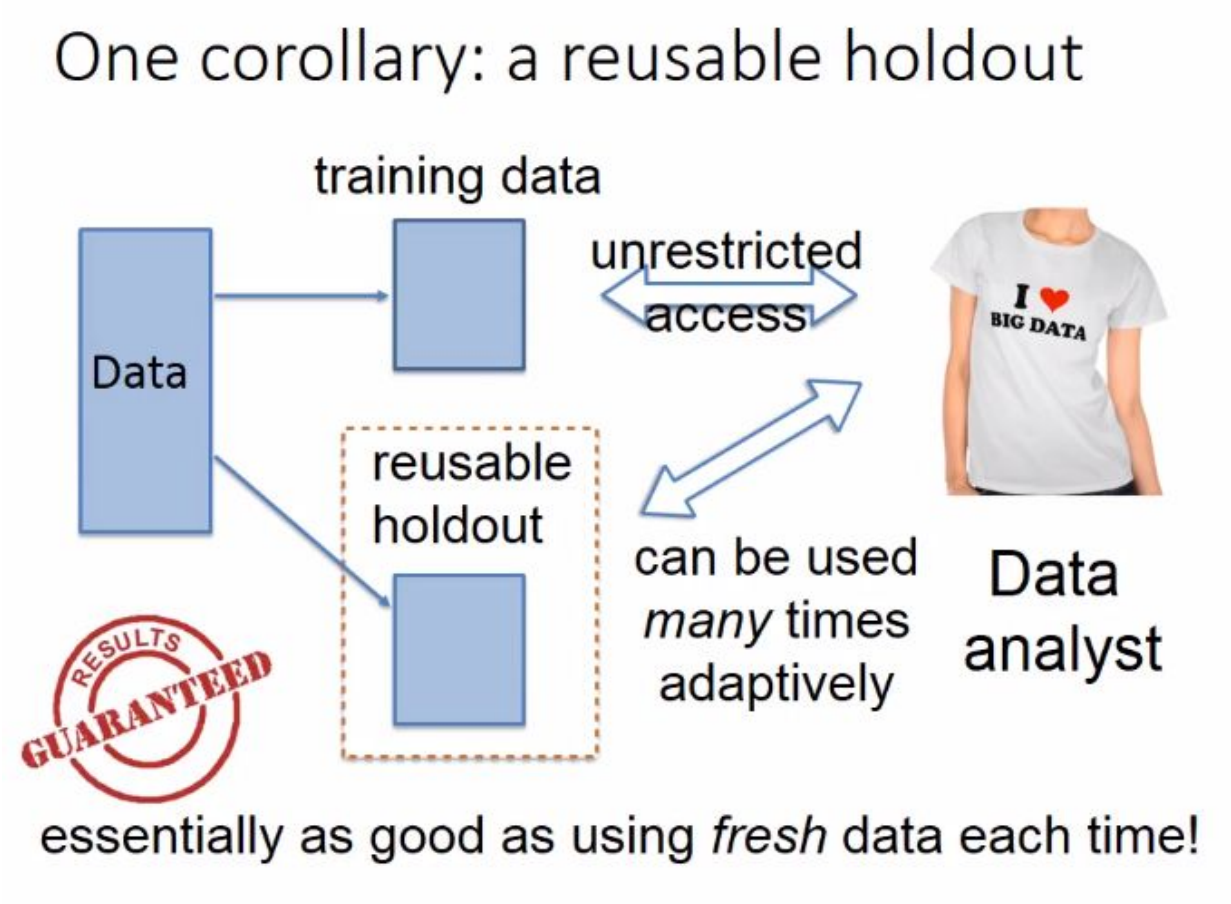
New algorithm aimed at combating science's reproducibility problem

August 6 2015



Big data sets are important tools of modern science. Mining for

correlations between millions of pieces of information can reveal vital relationships or predict future outcomes, such as risk factors for a disease or structures of new chemical compounds.

These mining operations are not without risk, however. Researchers can have a tough time telling when they have unearthed a nugget of truth, or what amounts to fool's gold: a correlation that seems to have predictive value but actually does not, as it results just from random chance.

A research team that bridges academia and industry has developed a new mining tool that can help tell these nuggets apart. In a study published in *Science*, they have outlined a method for successively testing hypotheses on the same data set without compromising statistical assurances that their conclusions are valid.

Existing checks on this kind of "adaptive analysis," where new hypotheses based on the results of previous ones are repeatedly tested on the same data, can only be applied to very large datasets. Acquiring enough data to run such checks can be logistically challenging or cost prohibitive.

The researchers' method could increase the power of analysis done on smaller datasets, by flagging ways researchers can come to a "false discovery," where a finding appears to be statistically significant but can't be reproduced in new data.

For each hypothesis that needs testing, it could act as a check against "overfitting", where predictive trends only apply to a given dataset and can't be generalized.

The study was conducted by Cynthia Dwork, distinguished scientist at Microsoft Research, Vitaly Feldman, research scientist at IBM's Almaden Research Center, Moritz Hardt, research scientist at Google,

Toniann Pitassi, professor in the Department of Computer Science at the University of Toronto, Omer Reingold, principle researcher at Samsung Research America, and Aaron Roth, assistant professor in the Department of Computer and Information Science in the University of Pennsylvania's School of Engineering and Applied Science.

Adaptive analysis, where multiple tests on a dataset are combined to increase their predictive power, is an increasingly common technique. It also has the ability to deceive.

Imagine receiving an anonymous tip via email one morning saying the price of a certain stock will rise by the end of the day. At the closing bell, the tipster's prediction is borne out, and another prediction is made. After a week of unbroken success, the tipster begins charging for his proven prognostication skills.

Many would be inclined to take up the tipster's offer and fall for this scam. Unbeknownst to his victims, the tipster started by sending random predictions to thousands of people, and only repeated the process with the ones that ended up being correct by chance. While only a handful of people might be left by the end of the week, each sees what appears to be a powerfully predictive correlation that is actually nothing more than a series of lucky coin-flips.

In the same way, "adaptively" testing many hypotheses on the same data, each new one influenced by the last, can make random noise seem like a signal: what is known as a false discovery. Because the correlations of these false discoveries are idiosyncratic to the dataset in which they were generated, they can't be reproduced when other researchers try to replicate them with new data.

The traditional way to check that a purported signal is not just coincidental noise is to use a "holdout." This is a data set that is kept

separate while the bulk of the data is analyzed. Hypotheses generated about correlations between items in the bulk data can be tested on the holdout; real relationships would exist in both sets, while false ones would fail to be replicated.

The problem with using holdouts in that way is that, by nature, they can only be reused if each hypothesis is independent of each other. Even a few additional hypotheses chained off one another could quickly lead to false discovery.

To this end, the researchers developed a tool known as a "reusable holdout." Instead of testing hypothesis on the holdout set directly, scientists would query it through a "differentially private" algorithm.

The "different" in its name is a reference to the guarantee that a differentially private algorithm makes. Its analyses should remain functionally identical when applied to two different datasets: one with and one without the data from any single individual. This means that any findings that would rely on idiosyncratic outliers of a given set would disappear when looking at data through a differentially private lens.

To test their algorithm, the researchers performed adaptive data analysis on a set rigged so that it contained nothing but random noise. The set was abstract, but could be thought of as one that tested 20,000 patients on 10,000 variables, such as variants in their genomes, for ones that were predictive of lung cancer.

Though, by design, none of the variables in the set were predictive of cancer, reuse of a holdout set in the standard way showed that 500 of them had significant predictive power. Performing the same analysis with the researchers' reusable holdout tool, however, correctly showed the lack of meaningful correlations.

An experiment with a second rigged dataset depicted a more realistic scenario. There, some of the variables did have predictive power, but traditional holdout use created a combination of variables with wildly overestimated this power. The reusable holdout tool correctly identified the 20 that had true statistical significance.

Beyond pointing out the dangers of accidental overfitting, the reusable holdout algorithm could warn users when they were exhausting the validity of a dataset. This is a red flag for what is known as "p-hacking," or intentionally gaming the data to get a publishable level of significance.

Implementing the reusable holdout algorithm will allow scientists to generate stronger, more generalizable findings from smaller amounts of data.

More information: "The reusable holdout: Preserving validity in adaptive data analysis," by C. Dwork et al. *Science*, www.sciencemag.org/lookup/doi/ … 1126/science.aaa9375

Provided by University of Pennsylvania

Citation: New algorithm aimed at combating science's reproducibility problem (2015, August 6) retrieved 3 May 2024 from https://phys.org/news/2015-08-algorithm-aimed-combating-science-problem.html