WIRED MAGAZINE: 16.07

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

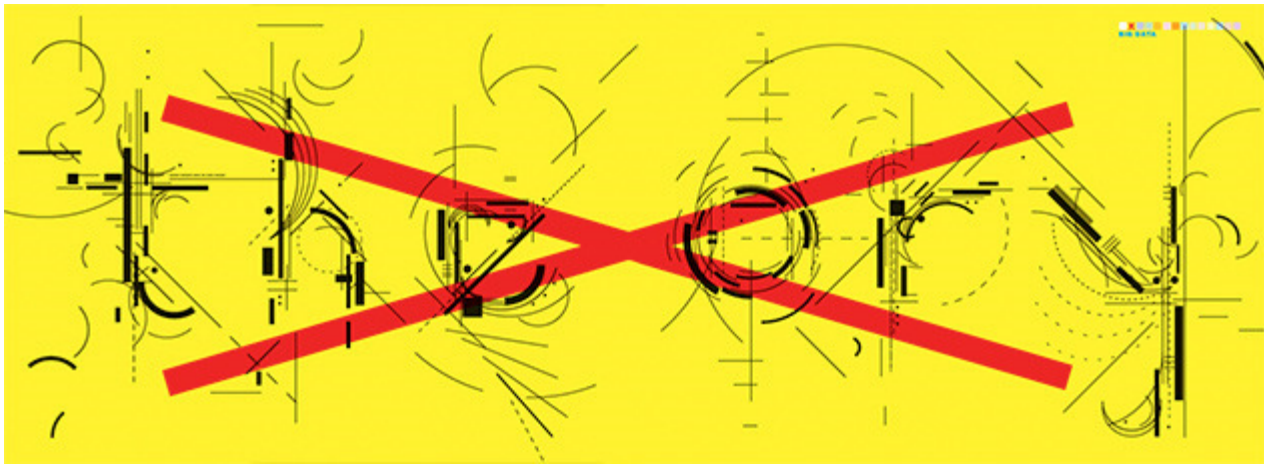

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our collection of facts and figures grows, so will the opportunity to find answers to fundamental questions. Because in the era of big data, more isn't just more. More is different.

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age.

The Petabyte Age is different because more is different. Kilobytes were stored on floppy disks. Megabytes were stored on hard disks. Terabytes were stored in disk arrays. Petabytes are stored in the cloud. As we moved along that progression, we went from the folder analogy to the file cabinet analogy to the library analogy to — well, at petabytes we ran out of organizational analogies.

At the petabyte scale, information is not a matter of simple three- and four-dimensional taxonomy and order but of dimensionally agnostic statistics. It calls for an entirely different approach, one that requires us to lose the tether of data as something that can be visualized in its totality. It forces us to view data mathematically first and establish a context for it later. For instance, Google conquered the advertising world with nothing more than applied mathematics. It didn't pretend to know anything about the culture and conventions of advertising — it just assumed that better data, with better analytical tools, would win the day. And Google was right.

Google's founding philosophy is that we don't know why this page is better than that one: If the statistics of incoming links say it is, that's good enough. No semantic or causal analysis is required. That's why Google can translate languages without actually "knowing" them (given equal corpus data, Google can translate Klingon into Farsi as easily as it can translate French into German). And why it can match ads to content without any knowledge or assumptions about the ads or the content.

Speaking at the O'Reilly Emerging Technology Conference this past March, Peter Norvig, Google's research director, offered an update to George Box's maxim: "All models are wrong, and increasingly you can succeed without them."

This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.

The big target here isn't advertising, though. It's science. The scientific method is built around testable hypotheses. These models, for the most part, are systems visualized in the minds of scientists. The models are then tested, and experiments confirm or falsify theoretical models of how the world works. This is the way science has worked for hundreds of years.

Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

But faced with massive data, this approach to science — hypothesize, model, test — is becoming obsolete. Consider physics: Newtonian models were crude approximations of the truth (wrong at the atomic level, but still useful). A hundred years ago, statistically based quantum mechanics offered a better picture — but quantum mechanics is yet another model, and as such it, too, is flawed, no doubt a caricature of a more complex underlying reality. The reason physics has drifted into theoretical speculation about *n*-dimensional grand unified models over the past few decades (the "beautiful story" phase of a discipline starved of data) is that we don't know how to run the experiments that would falsify the hypotheses — the energies are too high, the accelerators too expensive, and so on.

Now biology is heading in the same direction. The models we were taught in school about "dominant" and "recessive" genes steering a strictly Mendelian process have turned out to be an even greater simplification of reality than Newton's laws. The discovery of gene-protein interactions and other aspects of epigenetics has challenged the view of DNA as destiny and even introduced evidence that environment can influence inheritable traits, something once considered a genetic impossibility.

In short, the more we learn about biology, the further we find ourselves from a model that can explain it.

There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

The best practical example of this is the shotgun gene sequencing by J. Craig Venter. Enabled by high-speed sequencers and supercomputers that statistically analyze the data they produce, Venter went from sequencing individual organisms to sequencing entire ecosystems. In 2003, he started sequencing much of the ocean, retracing the voyage of Captain Cook. And in 2005 he started sequencing the air. In the process, he discovered thousands of previously unknown species of bacteria and other life-forms.

If the words "discover a new species" call to mind Darwin and drawings of finches, you may be stuck in the old way of doing science. Venter can tell you almost nothing about the species he found. He doesn't know what they look like, how they live, or much of anything else about their morphology. He doesn't even have their entire genome. All he has is a statistical blip — a unique sequence that, being unlike any other sequence in the database, must represent a new species.

This sequence may correlate with other sequences that resemble those of species we do know more about. In that case, Venter can make some guesses about the animals — that they convert sunlight into energy in a particular way, or that they descended from a common ancestor. But besides that, he has no better model of this species than Google has of your MySpace page. It's just data. By analyzing it with Google-quality computing resources, though, Venter has advanced biology more than anyone else of his generation.

This kind of thinking is poised to go mainstream. In February, the National Science Foundation announced the Cluster Exploratory, a program that funds research designed to run on a large-scale distributed computing platform developed by Google and IBM in conjunction with six pilot universities. The cluster will consist of 1,600 processors, several terabytes of memory, and hundreds of terabytes of storage, along with the software, including IBM's Tivoli and open source versions of Google File System and MapReduce.[1] Early CluE projects will include simulations of the brain and the nervous system and other biological research that lies somewhere between wetware and software.

Learning to use a "computer" of this scale may be challenging. But the opportunity is great: The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?