What versus Why
Towards Computing Reality

Michael L. Brodie, Jennie Duggan
Computer Science and Artificial Intelligence Laboratory (CSAIL) | MIT
April 17, 2014

In late 2013, a colleague, Herb Lin, Chief Scientist of the US National Academies of Science, expressed his concern about Big Data. Herb is responsible for overseeing Academy studies for scientific rigor. He was concerned about the growing use of Big Data since Big Data analytics can only suggest *What* has occurred within a specific probability or confidence level. Big Data analytics does not and cannot say anything about *causation*, or *Why* a phenomenon occurred. This observation has led us over the course of six months to some fairly profound observations. First, with Big Data we are moving from discrete, top-down, well understood models (data schemas and computational models) to a new world of computing that is vague, bottom-up, and model-less. In fact *if we impose models on Big Data analytics, we may obscure or prevent its greatest value.* These concerns are shared with DARPA's Big Mechanism program.

The shift is not just from small data to Big Data, it is a profound shift from a bounded computational world based on known models to an unbounded, model-less computational world. For the moment we call this vision, *Computing Reality*. Reality is continuous and unbounded in the atomic and astronomic directions. We inhabit a small part of Reality. Most of our known world (life, nature) is vague and ill defined; otherwise science would be at an end. We in computing (databases) inhabit a well-ordered subset of our known world, in which everything is discrete, conforms to known, well-defined models, e.g., a Telecom billing system and its schema. Database systems are the backbone of the well-defined worlds of business and enterprises. They require a *single version of truth* so that a telephone bill has a justifiable charge. How many parts of your life are governed by *a single version of truth*? There is no absolute truth. Truth is relative to a model or system, e.g., US law, Catholicism, Islam, first order logic, or more loosely to the social and other rules – an ensemble of rules - of an individual or a family. While this is a much larger philosophical discussion, I conclude with the observation that using Big Data introduces us to a profound change – a paradigm shift from well-defined, limited worlds, to worlds open to interpretation, like life, closer to Reality. We have almost no computing models for this. We have known this for over 2,000 years since Plato gave us his *Allegory of the Cave*. As exciting and difficult as this may be – dealing with reality – there is an even more profound shift under way.

The grander shift concerns Scientific Discovery. The history of Western Thought has been dominated, since the 3rd Century BC when Aristotle (384-322 BC) invented the origins of the modern scientific method with a logical, objective approach to reasoning about phenomena. It took almost 1,000 years for Roger Bacon and then another four hundred years for Francis Bacon and René Descartes to develop what we know today as the *Scientific Method*, a method for determining (explaining, predicting) *causation* or *why* phenomena occur. Since the 17th Century and increasingly in the 21st century most human thought subscribes to some form of scientific thinking.

Scientific Discovery involves *What* and *Why*. A scientist makes observations about some (e.g., physical, social, business, medical) phenomenon. Using in depth domain knowledge scientists make hypotheses as to *why* the phenomenon occurred, e.g., the interaction of various key parameters. They create a testable model to reflect the hypotheses. Since empirical studies are extremely costly and time consuming, the scientist develops a computational model for the hypotheses, finds the relevant data and runs the data through the model. The scientist may repeat the process with different model settings, with different data, until the results converge within a confidence level on the hypothesized result, e.g., some measure of efficacy. If it does converge within the norms of the domain, e.g., p-values in medical and clinical experiments, the scientist uses this evidence to assert, typically in a published paper, that the model satisfactorily explains *what* – that there is a significant correlation amongst the hypothesized

variables to warrant empirical investigation. However, the scientist cannot make any assertions about causality – the ultimate goal – the *holy grail of science*. The model-driven, or theory-driven results establish an adequate basis for proceeding with the empirical studies to establish *why* – causality.

Since the origins of the Scientific Method, *What* typically precedes *Why* as the smaller part of the scientific discovery process to validate the direction prior to the main course – the long-duration, costly empirical study or clinical trial. Until 2008, *Why*, the search for causation, dominated *What* (what correlations exist in the data). With the emergence of Big Data most recently led by Machine Learning there appears to be a shift in which *What* is beginning to dominates *Why* marking a potential shift from Theory-Driven to Data-Driven.

The shift from *Why* to *What* is disturbing due to the inherent weakness in *What* and the lack of maturity of some models such as Machine Learning. One concern is philosophical, a shift to a new way of thinking. On the one hand, we are concerned about the misuse of Big Data analytics, e.g., see [2][3] when applied to significant phenomena, like the stock market, earthquakes, and medicines. On the other hand we are astounded by the potential of exploring far beyond traditional realms to the *unknown unknowns*, e.g., automatic hypothesis generation.

Another concern is for the potential imprecision of models in areas where precision matters. The renowned statistician George P. Box said in 1987 "All models are wrong, but some are useful" meaning that all models are abstractions of reality for a particular purpose. That purpose is met if the essential characteristics of the phenomena are adequately captured and modeled. In which case the model is useful, to some degree, for that purpose, and possibly for no other purpose. *The best model of a cat is a cat. By that token, the best model of the universe is the universe.* Yet it is generally impractical to impossible, let alone too specific, to use a cat for a model of a cat, let alone the universe. This research concerns modelling, analyzing, or better yet listening to, Big Data, and the role that data plays in the process. Creating abstractions to support our limited cognition helps in asking reasonable questions. Finding the right questions to ask is often the first step to solving a problem.

Our research interest is the former aspect of *what* in which Big Data Analytics and Machine Learning has become one of the celebrated movements of the decade. However, all models can be used in pursuit of *what* and have been for 1,000 years – statistics, economics, Fourier transforms, even visualizations of data are models for understanding data. The eight equations of fluid dynamics are particularly wonderful as they are as close to perfect as any model of the physical world gets for practical purposes and have been since they were defined in the 19th century. What is emerging in many areas of science and the humanities is the value of multiple models, called *ensemble models,* to understand or explain a phenomenon from multiple points of view.

Our first objective is to understand an appropriate framework for models including their veracity. Just as at the beginning of the 19th century when the *Generally accepted accounting principles* (GAAP) were established to ensure veracity in business dealings, we need a framework for modelling and analysis of Big Data, for example methods of verifying and validating Machine Learning models and their results.

A longer-term objective is in the latter – the potential of exploring unknown universes, of *Computing Reality*. In this world the data speaks for itself. Models are not imposed top-down but emerge from the data through the correlations deduced by methods such as machine learning with confidence intervals for each result.

Why is this so profound, so unusual, so exciting? Human thought is limited by the human mind. According to Miller's Law[3], the human mind (short term, working memory) is capable of less than ten (7 +/- 2) objects. Hence, humans have difficulty understanding complex models, e.g., more than ten variables. The conventional process is to imagine a small number of variables[1] then abstract or encapsulate that knowledge into a model that can subsequently expanded with a few more variables. Thus most scientific theories develop slowly over time into complex models. For example, Newton's model of physics was extended for 350 years through Bohr, Einstein, Heisenberg, and many more, up to Higgs, to form The Standard Model of Particle Physics. Scientific Discovery in particle physics is wonderful but has taken over 350 years. However due to its complexity no physicist has understood the Standard Model for decades, rather it is represented in complex, computational models. It has become a *what*.

[1] Most PhDs in the physical sciences involve less than 5 variables.

In comparison with a capacity of the human mind for fewer than ten variables let alone their potential correlations, Machine Learning can identify correlations amongst billions of variables. We envisage an acceleration of Scientific Discovery in which *What* (Big Data and applicable models) is used collaboratively with *Why* to iterate at a far greater rate by the automatic generation of potential models with means to estimate their veracity in cooperation with *Why* – empirical studies of highly probable models to establish causality.

The vision of *Computing Reality* is so compelling since we live immersed in a digital world, yet the computing on which it is based is restricted to such a small part of the real world – discrete, top-down, well-defined models and computations, compared to the many perspectives and vagueness of nature and day to day life. Steps towards *Computing Reality* include what is called the Open World Assumption in Artificial Intelligence, Probabilistic Databases[4], Probabilistic Computing, Approximate Computing, Social Computing, Social Networks, Web Science, Crowd Computing, among others. Since Data Science deals with Big Data, Data Science should investigate Computing Reality - the world of possible worlds, a multiverse in physics. Currently we have no mathematical models for unknown unknowns.

[1] G. Marcus and E. Davis, Eight (No, Nine!) Problems With Big Data, Opinion Page, The New York Times, April 6, 2014

[2] David Leinweber, Stupid Data Miner Tricks: How Quants Fool Themselves And The Economic Indicator In Your Pants, Forbes, July 24, 2012

[3] Miller, G. A. (1956). "The magical number seven, plus or minus two: Some limits on our capacity for processing information". Psychological Review 63 (2): 81–97. doi:10.1037/h0043158. PMID 13310704.

[4] Nilesh Dalvi, Christopher Ré, Dan Suciu, Probabilistic Databases: Diamonds in the Dirt, Communications of the ACM, Vol. 52 No. 7, Pages 86-94 10.1145/1538788.1538810