CLAYTON CHRISTENSEN INSTITUTE for DISRUPTIVE INNOVATION

November 21, 2013

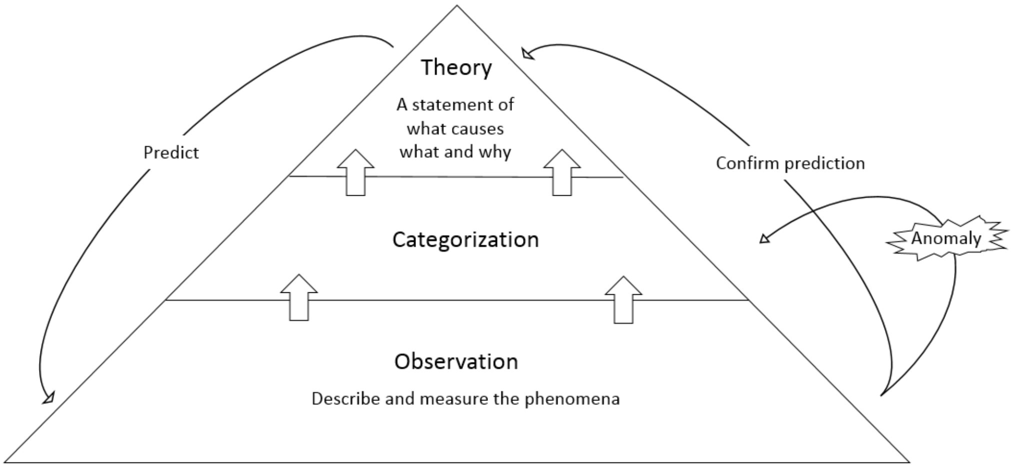
Big Data: The end of theory in healthcare?

by Ben Wanamaker and Devin Bean

Five years ago, Chris Anderson wrote an article in Wired titled, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." He claimed that, given enough information, "correlation is enough" to make robust and informative predictions. Since then, McKinsey and others have heralded a "Big Data Revolution in Healthcare," and the White House and others have poured millions of dollars in big data initiatives. Does this mark the end of theory relevance in healthcare?

We answer emphatically: No. By ignoring theory, big data exclusionists risk cementing themselves in a world where they can't cope with anomalies and changing circumstances, never able to progress toward conscious competence. At the same time, they distract from the very real and potentially disruptive ways that big data analytics could be revolutionary in healthcare and other industries.

The Classical Theory Development Model


Theory
A statement of what causes what and why
Predict
Confirm prediction
Categorization
Anomaly
Observation
Describe and measure the phenomena

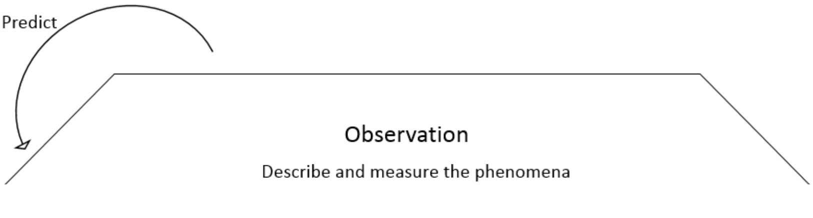
A theory is a statement of what causes what and why. Generally, researchers develop theory in three steps. First they observe, carefully describing and measuring the phenomena. Next, they group the observations into distinct categories, generally looking for similarities and differences among attributes. Finally, researchers develop a theory that explains how a certain set of attributes leads to a certain result.

After developing the initial theory (aka hypothesis), researchers seek to test the theory by using it to make predictions. In this stage, they typically encounter anomalies – things that the theory predicts would not happen do happen or vice versa. These anomalies, encountered because theories based primarily on attributes show correlation but do not yet explain causation, guide researchers to revisit the categorization scheme, continually refining the theory into a more powerful tool.

Eventually, what began as an attribute-based correlative theory evolves into a circumstance-based causal theory that generates useful predictions and robust explanations. These theories often spur dramatic advances – germ theory and genetic theory are just two examples.

Big Data and the Theory-less Model

Big data proponents, however, present a different model:

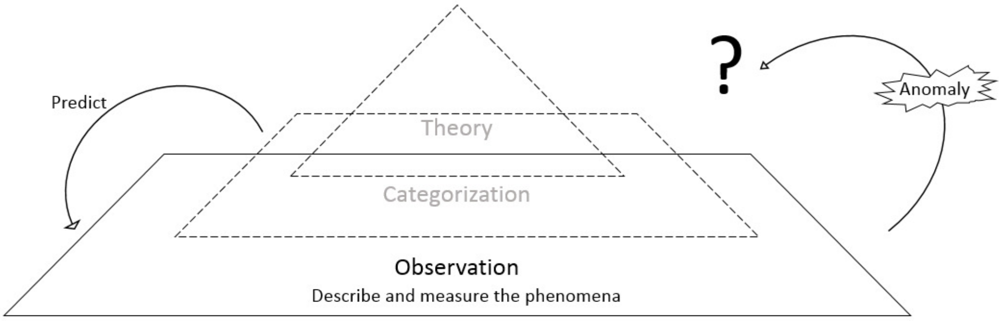


In the words of one biologist who questions the implications of big data on theory, "If you measure enough variables, it doesn't matter whether you understand the relationship between cause and effect; all you need is a relationship between one variable and another." Proponents suggest that data-driven prediction engines will generate predictions that will obviate the need for theory-drive analysis.

This model, however, presents two major problems.

1. **Theory, implicit or explicit, underlies all data analysis.** What we measure, how we measure it, and the predictive engines we apply all make implicit theory-based assumptions about the significance of certain forms of data and the meaning of the resultant outputs. Google, for instance, ranks pages via an algorithm that weighs inbound links, keywords, and other criteria. Each of these criteria are selected because, theoretically, they are relevant to a page's value.
2. **Implicit theory has no method of dealing with anomalies.** When an analytic engine fails, either because of false theoretical assumptions about attribute significance or because of external circumstance changes, there is no way to explicitly deal with the failure. Now that small molecule innovation is no longer delivering blockbuster drugs for big pharma, even the most advanced modeling engines that money can buy cannot generate predictions robust enough to guarantee drug development success.

Indeed, the theory-less model above should be re-drawn as follows:



Suppose that a successful big data driven algorithm encounters an unexpected situation. It has no method of dealing with the anomaly, and so crashes. Because theory is not explicitly examined and refined, not only will researchers have a difficult time learning from the anomaly, it could happen again. In 2010, for example, the Dow Jones plunged more than 600 points in minutes and then recovered just as abruptly, sending investors and businesses into a tailspin. The cause? Computer algorithms, at least in part. Because there is no good way to consciously learn from the anomaly, such volatility is still unpredictable and could happen again.

The Data Revolution Through the Lens of Disruption

All of this is not to say that big data is not important, only that current conceptualizations miss the real ways that big data is revolutionary! Fortunately, a theory exists that enables us to better understand the role of big data: the theory of disruptive innovation.

Disruptive innovations make technologies, products, and services that were once only accessible to the highly trained available to less skilled and less wealthy individuals. In the process, large companies fall and new ones rise from their ashes.

No big data proponent would disagree with the statement that "big data is potentially disruptive." But the *ways* in which it may be disruptive, as determined by the theory, are where we should direct conversation in order to cultivate significant innovation. Theory suggests that big data analysis technology, coupled with the proper business model (a topic for another blog), may be a disruption enabler in at least two ways:

1. **Enabling technologies make powerful data analytics affordable and accessible.** Inexpensive data storage, massive computing power, and widespread information transfer infrastructures make predictive analytics available to small players than ever before. Patient population analytics, for example, used to be available only to the largest and wealthiest of health systems who could hire talented teams of statisticians. Now, solutions exist that are affordable even for small providers.
2. **Accessible and affordable analytics drive operating efficiencies.** One third to one half of U.S. health care spend is on ineffective or inefficient care – hundreds of billions of dollars each year. Theory-based big data solutions can help change this. For example, theory-based (rather than simply correlative) population health management analytics updated with the latest best clinical practices from medical journals could help test, validate, and adopt new standards of care much more quickly than the average adoption timeline of 17 years and in so doing help obsolete outdated, wasteful, and expensive treatment methods While perhaps not as headline-making as "The End of Theory," "Save Hundreds of Billions of Dollars With Better Health Outcomes" is pretty close.

The final implication, then, is that big data used incorrectly could lead to incorrect conclusions and missed opportunity. Use correctly, it does have tremendous potential to revolutionize health care by making analytics more affordable and accessible and by driving efficiencies in care delivery – and these are the implications that we should be talking about, funding, and developing. So don't fire the research team just yet – instead, use them to develop robust theory that can drive health care innovation.

How much more data do we need to collect on HIV before a vaccine is available?

It's not just that theory underlies all data *analysis* but also all data *collection*.

–bks

by Bradley K. Sherman on Nov 25, 2013 6:54 PM PDT REPLY