CIO JOURNAL.

Why Do We Need Data Science When We've Had Statistics for Centuries?

By **IRVING WLADAWSKY-BERGER**

May 2, 2014 11:46 am ET

Data Science is emerging as one of the hottest new professions and academic disciplines in these early years of the 21st century. A number of articles have noted that the demand for data scientists is racing ahead of supply. People with the necessary skills are scarce, primarily because the discipline is so new. But, the situation is rapidly changing, as universities around the world have started to offer different kinds of graduate programs in data science. This year, for example, New York University is offering two new degrees–a general Master in Data Science, and a more domain-specific Master in Applied Urban Science and Informatics.

It's very exciting to contemplate the emergence of a major new discipline. It reminds me of the advent of computer science in the 1960s and 1970s. Like data science, computer science had its roots in a number of related areas, including math, engineering and management. In its early years, the field attracted people from a variety of other disciplines who started out using computers in their work or studies, and eventually switched to computer science from their original field.

This was the case with me. I used computers extensively while a student at the University of Chicago, where I worked closely with Prof. Clemens Roothaan, one of the pioneers in the use of computers in physics and chemistry. As an undergraduate, I worked part-time at the university's supercomputing center which he founded. Later he was my thesis advisor as a graduate student in physics. When the time came to look for a job, I realized that I enjoyed the computing side of my work more than the physics. I decided to switch fields and

in 1970 joined the computer science department at IBM's Watson Research Center.

Not unlike data science today, computing had to overcome the initial resistance of some prominent academics. I still remember a meeting in 1965 with a very eminent physicist from whom I was taking a graduate course. He asked me what I planned to do research on for my degree, and I told him that I was already working with Prof. Roothaan on atomic and molecular calculations. He just said that good theoretical physics should require no more than pencil and paper, rather than these elaborate new computers. In his mind, this wasn't real physics. A number of the physics faculty felt the same way. Change does not come easy, even for brilliant physicists.

Computer science has since become a well respected academic discipline. It has grown extensively since its early days and expanded in many new directions. It's quite possible that being around in the early days of computer science and computing in general is part of the reason I'm so interested in the evolution of data science today. So, what is data science all about? One of the best papers on the subject is Data Science and Prediction by Vasant Dhar, professor in NYU's Stern School of Business and Director of NYU's Center for Business Analytics. The paper was published in the *Communications of the ACM* in December 2013. "Use of the term *data science* is increasingly common, as is *big data,*" Mr. Dhar writes in the opening paragraph. "But what does it mean? Is there something unique about it? What skills do *data scientists* need to be productive in a world deluged by data? What are the implications for scientific inquiry?"

He defines data science as being essentially the systematic study of the extraction of knowledge from data. But analyzing data is something people have been doing with statistics and related methods for a while. "Why then do we need a new term like data science when we have had statistics for centuries? The fact that we now have huge amounts of data should not in and of itself justify the need for a new term."

In short, it's all about the difference between *explaining* and *predicting*. Data analysis has been generally used as a way of explaining some phenomenon by extracting interesting patterns from individual data sets with well-formulated queries. Data science, on the other hand, aims to discover and extract *actionable* knowledge from the data, that is, knowledge that can be used to make decisions and predictions, not just to *explain* what's going on.

The raw materials of data science are not independent data sets, no matter how large they are, but heterogeneous, unstructured data set of all kinds – text, images, video. The data scientist will not simply analyze the data, but will look at it from many angles, with the hope of discovering new insights.

One of the problems with conducting such an in-depth, exploratory analysis is that the multiple data sets that are typically required to do so are often found within organizational *silos;* be they different lines of business in a company, different companies in an industry or different institutions across society at large. Data science platforms and tools aim to address this problem by working with, linking together and analyzing data sets previously locked away in disparate *silos*.

"Unlike database querying, which asks *What data satisfies this pattern (query)?* discovery asks *What patterns satisfy this data?*," notes Mr. Dhar. "Specifically, our concern is finding interesting and robust patterns that satisfy the data, where *interesting* is usually something unexpected and actionable and *robust* is a pattern expected to occur in the future."

The article discusses the key skills data scientists should have, starting with machine learning, a complex concept which Mr. Dhar explains in a particularly simple way.

"Most of us are trained to believe theory must originate in the human mind based on prior theory, with data then gathered to demonstrate the validity of the theory. Machine learning turns this process around. Given a large trove of data, the computer taunts us by saying, *If only you knew what question to ask me, I would give you some very interesting answers based on the data.* Such a capability is powerful since we often do not know what question to ask. . ."

> "Suitably designed machine learning algorithms help find such patterns for us. To be useful both practically and scientifically, the patterns must be predictive. The emphasis on predictability typically favors Occam's razor, or succinctness, since simpler models are more likely to hold up on future observations than more complex ones, all else being equal. . ."

Data scientists should also have good computer science skills–including data structures, algorithms, systems and scripting languages–as well as a good understanding of correlation, causation and related concepts which are central to modeling exercises involving data.

> "The final skill set is the least standardized and somewhat elusive and to some extent a craft but also a key differentiator to be an effective data scientist – the ability to formulate problems in a way that results in effective solutions. . . formulation expertise involves the ability to see commonalities across very different problems . . ."

Like computing, one of the most exciting part of data science is that it can be applied to many domains of knowledge. But doing so effectively requires domain expertise to identify the important problems to solve in a given area, the kinds of questions we should be asking and the kinds of answers we should be looking for, as well as how to best present whatever insights are discovered so they can be understood by domain practitioners in their own terms. *Garbage-in, garbage-out*, a phrase I often heard in the early days of computing, is just as applicable to data science today.

Physics, chemistry, biology and other natural science disciplines have long been practicing their own version of data science. In physics, for example, "a theory is expected to be *complete* in the sense a relationship among certain variables is intended to explain the phenomenon completely, with no exceptions. . . In such domains, the explanatory and predictive models are synonymous."

But given our newfound ability to gather valuable data on almost any topic, prediction can now apply to *softer* disciplines like the health and social sciences. Mr. Dhar points out that while these fields generally lack solid theories, "large amounts of data can result in accurate predictive models, even though no causal insights are immediately apparent. As long as their prediction errors are small, they could still point us in the right direction for theory development."

Finally, beyond access to the appropriate skills, are there cultural and management implications in embracing data science in the business world?

> "Besides recognizing and nurturing the appropriate skill sets, it requires a shift in managers' mind-sets toward data-driven decision making to replace or augment intuition and past practices. A famous quote by 20th-century American statistician W. Edwards Demming - *In God we trust, everyone else please bring data* - has come to characterize the new orientation, from intuition-based decision making to fact-based decision making. . . It is suddenly possible to test many of their established intuitions, experiment cheaply and accurately, and base decisions on data. This opportunity

requires a fundamental shift in organizational culture, one seen in organizations that have embraced the emerging world of data for decision making."

Irving Wladawsky-Berger is a former vice-president of technical strategy and innovation at IBM. He is a strategic advisor to Citigroup and is a regular contributor to CIO Journal.

Share this: http://on.wsj.com/1iRCO1w

DATA SCIENCE (HTTP://BLOGS.WSJ.COM/CIO/TAG/DATA-SCIENCE/)

DATA SCIENTIST (HTTP://BLOGS.WSJ.COM/CIO/TAG/DATA-SCIENTIST/)