Data Science and Analytics

Department of Industrial and Systems Engineering

Lehigh University

Technical Report 15T-009

You can't manage what you don't measure.

Harvard Business Review

Data Science, Big Data, and Big Analytics

What is *Data Science*? The definition is still evolving and an Internet search for the term reveals dozens of variations. As a simple working definition, we define Data Science simply to be **the science of extracting knowledge from data**. From the recent attention Data Science has received in academic journals and the popular press, one gets the distinct impression that this is a new discipline. But is it really? Experts in data analysis, most notably statisticians, have been extracting knowledge from data for decades. In a recent article in *Forbes* entitled "A Very Short History of Data Science," [1] Gil Press traces the origins of Data Science as a discipline back to an article by John Tukey in 1962 called "The Future of Data Analysis" [2] in which he wrote

> Data analysis, and the parts of statistics which adhere to it, must...take on the characteristics of science rather than those of mathematics... data analysis is intrinsically an empirical science... How vital and how important... is the rise of the stored-program electronic computer? In many instances the answer may surprise many by being 'important but not vital,' although in others there is no doubt but what the computer has been 'vital.'

Given this early recognition by Tukey and others of the importance of Data Science as a field distinct from statistics, why has it taken so long for it to be recognized as a pronounced and important discipline? The most likely answer is that it was several more decades before the confluence of computational methods, computing technology, and mathematical techniques that allowed Tukey's vision to be realized would occur. Although it was possible to envision modern Data Science several decades ago, we simply did not have the means to generate, store, and share the volumes of data required for many of the applications that are driving modern needs and trends.

Big Data

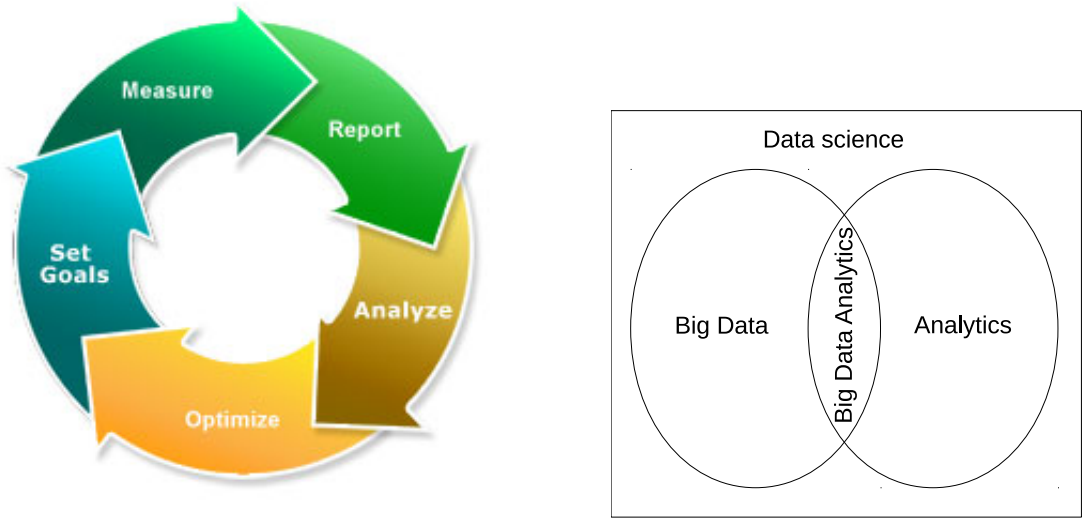
Another term that has recently gained traction and cachet in both the popular press and academic circles is *Big Data*. It is clear that we have now entered "the age of Big Data" and much of the recent emphasis on Data Science has been borne out of the explosion in the availability of Big Data, usually described as data having the following characteristics [3] :

- **Volume.** It is estimated that tens of exabytes of data are gathered worldwide each day and this amount is forecasted to double every 40 months. For example, it is estimated that Walmart collects more than 2 petabytes of data every hour from its customer transactions.
- **Velocity.** For many applications, the speed of data creation is even more important than its volume. Real-time information can help companies be more agile than their competitors.
- **Variety.** Big Data includes a wide variety of data types, including Facebook statuses, pictures on Google's Picasa or Flickr, articles in Wikipedia, Tweets on Twitter, readings from various sensors, YouTube movies, and much more. All of these are sources of unstructured data, not suitable to be stored in classical relational databases, which assume that data possess a certain structure.

[1] http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

[2] Annals of Mathematical Statistics, Volume 33, Number 1 (1962), pages 1-67

[3] http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf



(a) The Anaytics Process (source: https://www.informs.org/About-INFORMS/What-is-Analytics)

(b) Big Data Analytics Venn Diagram

Figure 1. The Analytics Process and Big Data Analytics

It would be a mistake, however, to equate Data Science with Big Data. Data does not have to be "big" in order for the extraction of knowledge from it to be challenging.

Analytics

Analytics is another term that has been variously defined and has recently increased in usage and popularity. The Institute for Operations Research and Management Science (INFORMS), the leading professional society of Analytics experts, defines it as the **scientific process of transforming data into insight for making better decisions.** [4] This definition differs from that of Data Science in that it makes explicit the end goal of having *the insight to make an informed decision*. The data is one input into a cyclic process, shown in Figure 1(a), in which the collection of data drives decisions, which in turn drive the collection of more data. Although it is easy to collect a large volume of data without first thinking about what decisions these data will be used to make, this indiscriminate approach collection is not likely to lead to meaningful results. The cyclic nature of the Analytics process is critical.

In "A Taxonomy of Data Science,"[5] Mason and Wiggins provide an alternative view of this process and state that there are five steps data scientists follow in analyzing data: Obtain, Scrub, Explore, Model, and Interpret. This describes a similar cycle, but explicitly includes the concept of developing a "model" following the exploration phase. Exploration can be seen as an informal and usually human-driven manual process of determining what modeling framework is most appropriate for the more rigorous analysis to follow. One cannot over-emphasize the importance of the underlying model that one adopts after the exploration phase, as it informs what conclusions one can and will eventually reach. In practice, the cyclic process in Figure 1(a) includes refinements in the model.

Finally, Analytics is often seen as consisting of three different types of analysis:

- **Descriptive.** Summarizing historical data and identifying patterns and trends.
- **Predictive.** Forecasting what future data will look like if current trends continue.
- **Prescriptive.** Determining what actions to take in order to change undesirable trends.

[4] https://www.informs.org/Sites/Getting-Started-With-Analytics/What-Analytics-Is
[5] http://www.dataists.com/2010/09/a-taxonomy-of-data-science/

Roughly speaking, with respect to the Analytics process in Figure 1(a), the first two of these types make up the *Analyze* step, while the third is the primary driver of the *Optimize* step. Most current Analytics research is focused on the second and third steps, each of which is challenging in its own right.

The close relationship between Data Science and Analytics should be evident from the above discussion. Although the Analytics process is data-driven and the focus is rightly on the data as the raw material, viewing Data Science through an Analytics lens highlights important steps in the overall process that can be challenging regardless of the size of the data. Prescriptive Analytics requires both the development of abstract modeling frameworks and the techniques from computational mathematics required to analyze these models once they are populated with specific data.

One of the core theoretical and methodological tools used in data-driven decision-making is that of Mathematical Optimization. From simple regression to the tuning of deep neural networks, mathematical optimization techniques are critical in the exploitation of data, though to use them efficiently can be extremely challenging. For example, it is not difficult to construct optimization models with small input size in terms of data, but for which the space of possible courses of action is astronomical and the search for the "optimal" course may be intractable. Typical examples include the straightforward problem of determining the routing of delivery vehicles to deliver goods to customers from a central warehouse. Even for a relatively small number of customers and vehicles, and with deliveries made from a single warehouse, this problem (known in academic papers as the *Vehicle Routing Problem*) is difficult to solve.

Big Data Analytics

When problems that are already computationally difficult at a small scale are made more realistic by including fine-grained data (e.g., demand forecasts) and the problem is scaled to the size faced by a company such as Amazon or Google, then we have entered the realm of *Big Data Analytics*. The techniques of Big Data Analytics encompass the computational challenges involved both in the *analysis* of the data and in the *exploitation* of it as part of a data-driven decision-making process. Simply put, Big Data Analytics (see Figure 1(b)) is the confluence of Big Data, Big Analytics, and Big Computation.

Machine learning, data mining, social network analysis, financial optimization, healthcare analytics, and computational biology are some of many prominent application domains where Big Data is available and mathematical optimization modeling is the natural framework for making decisions. In these applications, decision problems with millions or billions of variables are commonplace. Classical optimization algorithms are not designed to scale to instances of this size. There is a need to continually develop new approaches.

ISE and Big Data Analytics

Beginning around the year 2000, ISE instituted a departmental focus on Analytics. Since that time, Analytics has been our core strength and has since been identified in our strategic plan as both our primary strategic focus and our primary growth area. Throughout, we have grown our expertise in this area through new educational programs and initiatives, targeted hiring, and the development of new labs and research centers.

Among the educational initiatives that facilitated this growth was the development of a new undergraduate program called *Information and Systems Engineering* (I&SE). The goal of this program was to create a new curriculum which would integrate the use of information technologies into our traditional *Industrial Engineering* (IE) program. The degree was developed in response to trends similar to those that have driven the recent interest in Data Science.

The I&SE program drove the hiring of a large cohort of faculty who, in turn, contributed to the continued growth of the department's Analytics focus in its research program. This strategy began to take shape with a few initial hires in the area of Big Computation (with expertise in large-scale discrete optimization and parallel computing), continued with hires specialized in other areas of optimization (particularly nonlinear, robust, and stochastic optimization), and has grown further with more recent hires having advanced Data

Science credentials (expertise in Machine Learning, Data Mining, and Stochastic Processes).

Throughout this growth, ISE's research centers have advanced with a similar strategic focus on Analytics. The Laboratory for Computational Research at Lehigh (COR@L), home to the department's state-of-the-art compute cluster, powers the department's research efforts on large-scale computations and optimization. The Center for Value Chain Research, now moved outside of ISE, was originally an outgrowth of efforts by ISE faculty to study Analytics in a more practical context; the Enterprise Systems Center has a strategic partnership with SAS, one of the nation's largest Analytics companies; and finally, the Integrated Networks for Electricity Research Cluster, in which ISE is a partner, is in part an effort to exploit Analytics techniques in the design and operation of advanced electricity systems.