International Journal of Business Analytics and Intelligence
8 (1) 2020, 04-08
http://publishingindia.com/ijbai/

The Roots of Data Science

Favio Vazquez*

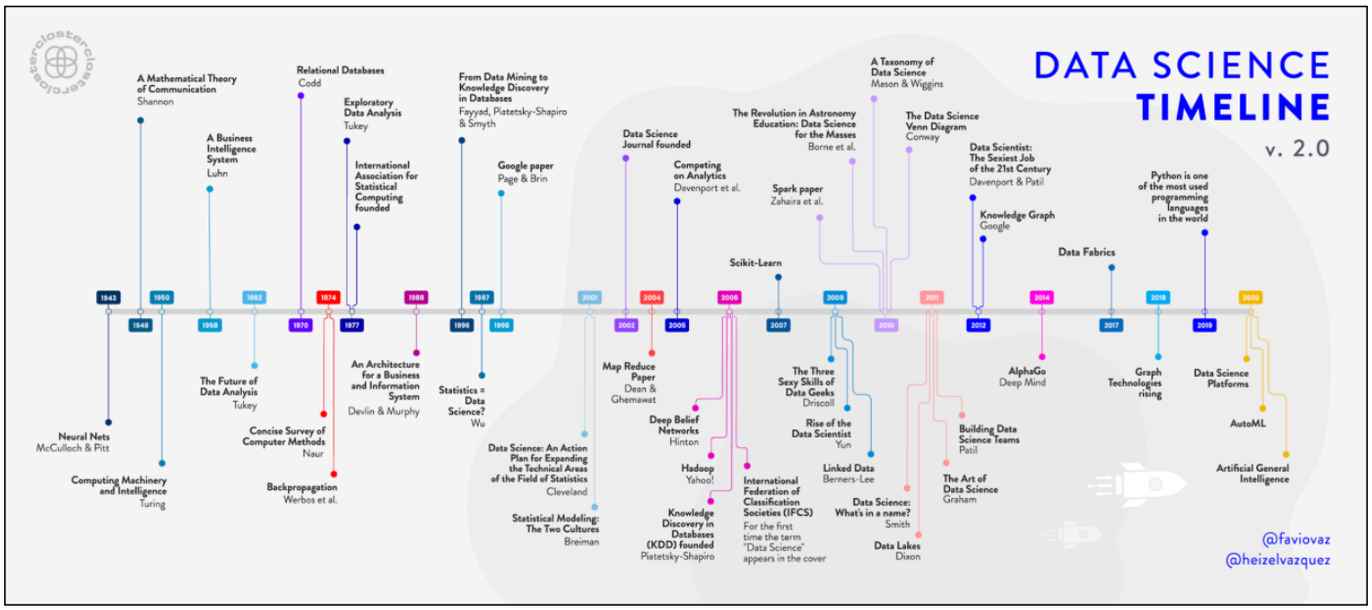
We have a name for the scientific exploration of data to find valuable information and solve business problems: Data Science (DS). This statement is not a definition, just a basic description.

DS has been there for a while now. I have covered my thoughts on the field extensively in the past (check the bottom for resources), but I want to talk about something different in this article.

DS is not coming out of anywhere. It comes from an iteration of statistics, specifically the algorithmic tradition of statistics and data analysis. Three great articles are the founders of our field (as we know it right now):

- The Future of Data Analysis by John Tukey.
- Statistical Modeling: The Two Cultures by Leo Breiman.
- Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics by William Cleveland.

But also those papers depend on a bigger picture. I tried to show the full development of the field of DS in the timeline I created with my sister Héizel Vázquez:



Please make sure to check the sources as well.

As we can see, the oldest of the three papers I mention is the one by Tukey. I will tell you why I'm proposing that this is the beginning of our field (even though there are papers before it in the timeline). Still, before that, I want to discuss some notions from the article: A Business Intelligence System by Luhn.

Before the Beginning

In the paper "A Business Intelligence System," Luhn proposed an automatic system to disseminate information to the various sections of any industrial, scientific, or government organization.

But I want to bring two definitions from this article. First, the definition of a business:

* CEO at Closter, Mexico City Area, Mexico.

> "Business is a collection of activities carried on for whatever purpose, be it science, technology, commerce, industry, law, government, defence, et cetera."

And the definition of intelligence:

> "The notion of intelligence is […] the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal".

I like these definitions. And as we will see, the notion of a "business intelligence system" will guide later authors in their interpretation of data science.

Luhn also said:

> "Efficient communication is key to progress in all fields of human endeavor. […] Information is now being generated and utilized at an ever-increasing rate because of the accelerated pace and scope of human activities and the steady rise in the average level of education. […] There is also a growing need for more prompt decisions at levels of responsibility far below those customary in the past. […] In view of the present growth trends, automation appears to offer the most efficient methods for retrieval and dissemination of this information."

Let's talk about this. The author is talking about communication and information. Also, companies are utilizing this information to make decisions that are much more advanced than what we could do in the past. Please remember that this article is from 1958.

Luhn also talks about the importance of automation for disseminating this information, something that was one of the foundations of data mining (one of the fathers of data science).

So far, we have four significant components that will be important soon:

- Information
- Intelligence
- Business
- Automation

From Statistics to Data Analysis

John Tukey is one of the most important statisticians in history. In the fantastic article "The Future of Data Analysis" he said this:

> For a long time, I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. […] All in all, I have come to feel that my central interest is in data analysis.

A huge statement to make by a statistician. In this time, the words "data science" did not exist as today, but the way Tukey described data analysis is very close to what we call now data science. He even called data analysis a science, because it passes these three tests:

- Intellectual content.
- Organization into an understandable form.
- Reliance upon the test of experience as the ultimate standard of validity.

There's also an important thing he says in the article:

> "If data analysis is to be helpful and useful, it must be practiced."

This seems obvious, but it guided how people then described DS. What do we understand from this article? There's an evolution of statistics that will create what Tukey called a new data analysis, defined as:

> "[The] Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

Something crucial in this definition is the idea that data analysis also includes the planning for gathering data to analyze, and also the machinery of statistics for analyzing and also interpreting data.

Aside than saying that data analysis should be practised, Tukey says that:

> "We need to face up to more realistic problems."

He was talking about data analysis, of course. Even though the article has the idea of explaining how to teach and use data analysis in a formal and academic realm, Tukey understands that we need to go beyond textbooks and idealistic scenarios. Something important to understand the connection between Luhn's article and this one.

From Data Analysis to Data Science

Thirty-five years later after Tukey's publication, Jeff Wu said this:

Statistics = Data Science?

Where he proposed that statistics should be renamed "data science" and statisticians should be named "data scientists". In today's standards, we know that statistics alone is not of data science, but why? Because we also need programming, business understanding, machine learning, and more (but more on that soon).

In a conversation with Jeff Wu, he mentioned that:

> "My lecture was entitled Statistics = Data Science?. There I characterized statistics as a trilogy of data collection, data analysis and decision making. I was talking about analytic statistics, rather than descriptive statistics. I suggested a change of our name from "statistics" to "data science" and "statistician" to "data scientist." I remember I even jokingly said in the lecture that by merely changing the name to data scientist, the salary will go higher. This is true nowadays. It's interesting."

Something interesting about Wu's definition of statistics is that data analysis is a part of it. I'm not entirely sure if Tukey will agree with Wu, but the idea is clear:

> Data science depends on data collection, data analysis, and decision making.

Finally, we start talking about something else: decision making. This is one of the connections between Tukey's views on data analysis and statistics, and Luhn's views on business intelligence.

Please check the timeline to remember the dates of the articles and presentations that I'm talking about.

Four years after Wu's presentations (2001), two papers put everything together. In April of 2001, Cleveland proposed an action plan to enlarge the technical areas of the field of statistics, and he called it Data Science. And then, in August of the same year, Breiman proposed that the use of the algorithmic modeling (as a different statistical culture) will be better to solve problems with data, rather than the classical statistical modeling.

The two articles are relevant in different ways, Cleveland's article aimed to create an academic plan to teach data science (similar to what Tukey did for data analysis) and Breiman's article had the idea to talk about the practical implications of data science and its relation to business (close to what Luhn wanted to explain with an application).

Even though Cleveland's article was directed to universities and educational institutes, he mentioned:

> Universities have been chosen as the setting for implementation because they have been our traditional institutions for innovation […]. But a similar plan would apply to government research labs and corporate research organizations.

So he's recognizing the importance of the government and also organizations in the process of institutionalizing data science as a serious field.

In the article, Cleveland states that data science depends on four big things (he talks about six things, but taking out the parts related to teaching DS):

Multidisciplinary Projects: Here he mentions:

> The single biggest stimulus of new tools and theories of data science is the analysis of data to solve problems posed in terms of the subject matter under investigation. Creative researchers, faced with problems posed by data, will respond with a wealth of new ideas that often apply much more widely than the particular data sets that gave rise to the ideas.

Important things to highlight here:

- Data analysis and data science have the primary goal of solving problems (that will be important when we talk about Breiman's article).
- The practitioner of data science needs to work on different issues and fields to be able to have a bigger picture, to exploit creativity, and to understand different types of data and problems posed by data.

Models and Methods: Here he mentions:

> The data analyst faces two critical tasks that employ statistical models and methods: (1) specification-

> the building of a model for the data; (2) estimation and distribution-formal, mathematical probabilistic inferences, conditional on the model, in which quantities of a model are estimated, and uncertainty is characterized by probability distributions.

Important to notice that he talks about the practitioner of data science as the data analyst, but we will refer to them as data scientists (something to think about).

In here we have to highlight that:

- Modeling is at the core of data science. This is the process of understanding the "reality", the world around us, but creating a higher level prototype that will describe the things we are seeing, hearing, and feeling. Still, it's a representative thing, not the "actual" or "real" thing. Tukey also talks about this in his articles.
- Data science needs a method (and a methodology).
- The data scientist creates models for the data and uses statistical techniques and methods to develop these methods. As we will see in Breiman's article, he emphasizes algorithms instead of formal mathematical methods.

Computing with Data: Here he mentions:

> Data analysis projects today rely on databases, computer and network hardware, and computer and network software. […] Along with computational methods, computing with data includes database management systems for data analysis, software systems for data analysis, and hardware systems for data analysis.

He also talks about the gap between statisticians and computer scientists:

> […] One current of work is data mining. But the benefit to the data analyst has been limited, because the knowledge among computer scientists about how to think of and approach the analysis of data is limited, just as the knowledge of computing environments by statisticians is limited.

And one of his ideas is that a "merger of the knowledge [Statistics and Computer Science] bases would produce a powerful force for innovation".

Some other things to highlight:

- The data scientists need an understanding of databases and computational software. Programming is there as well. He also talks about statistical packages and related software. But now we know that the path for data science nowadays depends on the understanding of some programming languages (mostly Python and R right now).
- Data science also depends on technological advances. This was true in 2001 and is true today as well. The methods that the data scientists use are shaped by the theoretical developments (check the timeline) but also on the fact that today we have powerful computers, cheaper and faster memory, high-speed internet, and also GPUs and TPUs.
- We need statisticians to learn computer science and computer scientists to learn statistics. This gap is filled right now by data scientists, but we can't forget that moving between these fields is becoming more usual, and we need experts in statistics to learn computer science and experts in computer science to learn statistics, not only people that are proficient in both.

Theory: Here, he mentions:

> Theory, both mathematical and non-mathematical theory, is vital to data science. Theoretical work needs to have a clearly delineated outcome for the data analyst, albeit indirect in many cases. Tools of data science-models and methods together with computational methods and computing systems-link data and theory. New data create the need for new tools. New tools need a new theory to guide their development.

Data science is a practical field, but it needs theory to understand and explain their methods and models. Today we know that if you want to understand machine learning, you will need an understanding of linear algebra, differential calculus, statistics, and probability (to mention some of the most important).

Important things to highlight:

- The tools of data science and its models link the data and the theory. We need to understand the theory to create better models, and when we build models, we use all the theoretical tools.
- Different datasets need different theoretical backgrounds. This is clear in Tukey's paper, where he

mentions some of the most important pieces of mathematics and statistics to work with different datasets. We saw this when big data exploded, and we had to analyze disparate sources of data.

- The theoretical advancements guide the creation of new tools and models. This reminds the history of science, where not only data and experiments led to the creation of new theories, but also, the new theories developed guided experiments, models, and tools.

The Algorithmic Modeling Culture in Data Science

As I mention, in August of 2001, Leo Breiman published a paper on the two cultures of statistics: The data modeling and the algorithmic modeling one. One of the remarks he makes in his article is:

> The roots of statistics, as in science, lie in working with data and checking theory against data. I hope in this century our field will return to its roots.

Here he mentions that there are some people in the statistical culture that are driven by data modeling and some by algorithmic modeling. Where the first ones assume that we have a stochastic data model that maps input variables x to response variables y. And the second one considers that the mapping process is both complex and unknown, and their approach is to find a function $f(x)$ that operates on x to predict the responses y.

He then goes to discuss why the data modeling culture has been bad for statistics for so long, leading to irrelevant theories and questionable scientific conclusions keeping statisticians from using more suitable algorithmic models and working on exciting new problems. Also, he talks about the wonders of the other part of the spectrum, the algorithmic modeling culture giving examples from his works, and others on how it can solve hard and complex problems.

He states that the algorithmic culture:

> […] Shifts focus from data models to the properties of algorithms. It characterizes their "strength" as predictors, convergence if they are iterative, and what gives them good predictive accuracy. The one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution.

He's not saying that the "old" statistical culture is useless. Instead, he means that the algorithmic culture is better suited for the current (back in 2001, of course) problems posed by data.

One of the guiding principles in this new culture, more close to what we do in data science, accordingly to Breiman is:

> The goal is not interpretability, but accurate information.

If we see what happened after his paper, that's exactly what happened. The advancements in algorithms, methods, and models were to improve accuracy, sacrificing interpretability. Luckily in the past years, there has been a tremendous advancement in the explainability and interpretability of "black boxes"; we now have tools to explain how a random forest, support vector machine, or a deep neural network works.

I love the way that he explains the usage of the algorithmic culture in several of these consulting works, and we can see how, changing some fundamentals aspects of the usage of data and mathematics, we can improve accuracy and solve much more complex problems.

Conclusions

Combining the work of the authors mentioned in this article, and adding to that the theoretical, computational, and scientific advancements proposed in the timeline, we can understand the historical development of data science.

The roots of data science are statistics, but also the idea of using data to solve business problems. Data science should be practical, but it relies upon the usage of the theory of mathematics and computer science to function. The practice and study of data science should be a part of every university, government, and organization that wants to use data to solve complex problems.

Data science has become the standard solving problem framework for academia and the industry, and it's going to be like that for a while. But we need to understand where we are coming from, who we are and where we are going.

If we channel the resources we have right now to make this area of knowledge work together for a greater good, we can make a tremendous positive impact in the world and our lives. It's our time.