International Journal of Business Analytics and Intelligence
7 (2) 2019, 06-10
http://publishingindia.com/ijbai/

Submitted: 22 July, 2019
Accepted: 18 August, 2019

On Data and Science

Favio Vázquez*

If we fail to predict the future, they call us a failure; if we do it too well, they call us a sorcerer. But as Poincaré said, "It is far better to foresee even without certainty than not to foresee at all."

We have been doing science for a while now. I am going to put beginning in 1637 when Descartes published "Discourse on the Method." The main result of that book is the distinction between knowledge and truth, and that the discourse of the scientist is related to knowledge (that we later discover to be always incomplete), not to look for the truth.

That is a very important point, because it gives us a focus; science wants to know stuff that are not "undisputed truths."

Again reframing what Poincaré said about Mathematics, if we wish to foresee the future of data science, our proper course is to study the history and present condition of the science.

Let us start with some history. Instead of boring you with paragraphs of text, I build this timeline about data science that will help understand where we are coming from and where we are going.

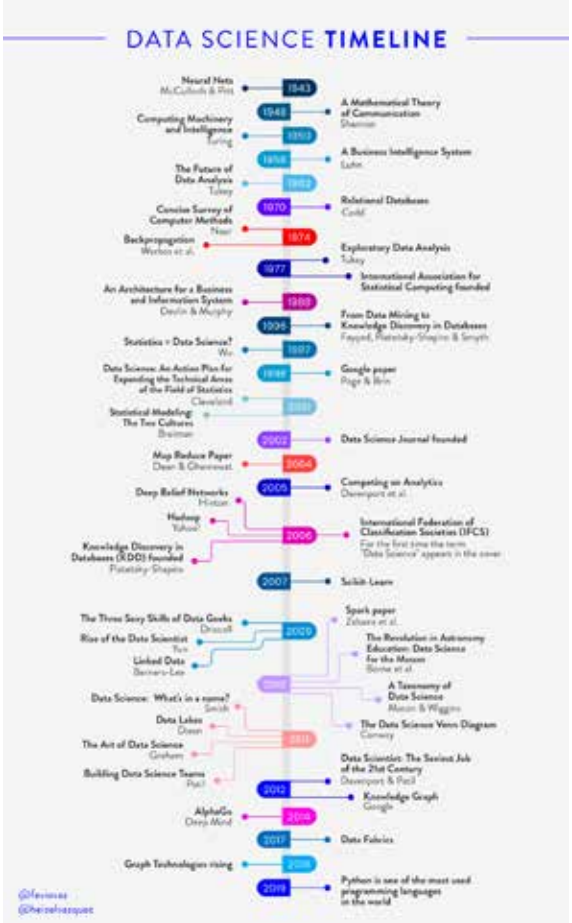

Fig. 1

* Data Science Course Instructor, Business Science University, Pennsylvania Area.

Not a new field at all, but with amazing developments in the past decades. I had to cut some parts and moments because of space, but the most important things are hopefully there.

This timeline will give you an idea of the history of the science of studying data, which we call data science right now, but that may not be the final term we have. So, always be prepared for a change.

The Role of Data in Science

Data have been close to science in almost all of its history. Sometimes, theory gets us to the data, but for the most of what we are going to discuss here and other upcoming articles, it's going to be in the model data -> theory.

I want to tell you a little story that will make you understand better why data are important, but not the most important part of science.

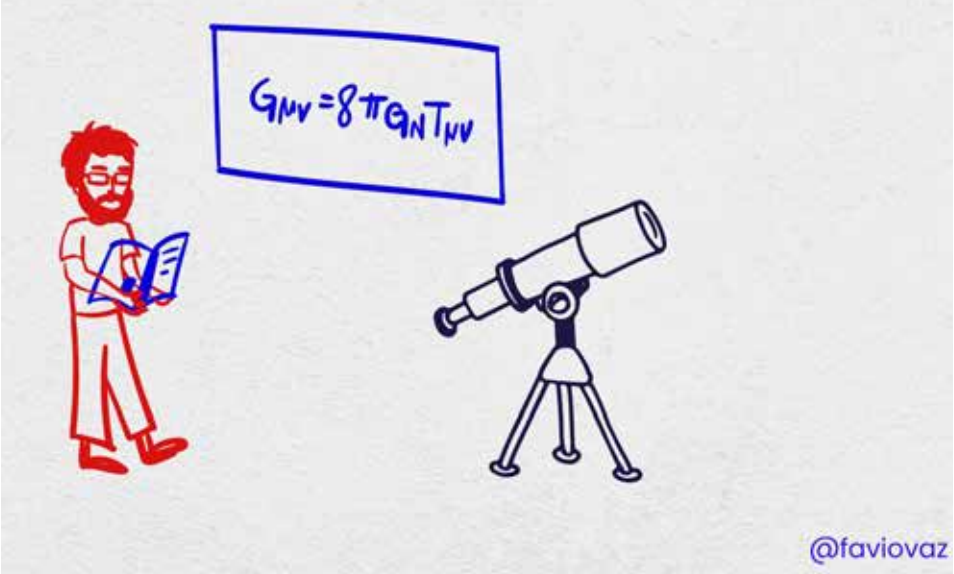

Fig. 2

The story is about a man called Tycho Brahe. He spent almost all his life measuring the positions of the stars, planets, the Moon, and Sun. For what? He wanted to learn how to predict eclipses. Also, he was unhappy with the Ptolemaic system and the Copernican theory was not enough for him either. So, he wanted to find the best way to describe the skies and its moving parts.

Sadly, he was not sure how, but he kept on measuring things until his final days. He died in 1601, and someone named Johannes Kepler, who became his assistant a year before, who was a great mathematician, had access to almost all his data. With that data, Kepler improved the Copernicus theory of the Universe and developed three laws that described the motion of the planets. Kepler's work served as the basis for the later studies of Isaac Newton about the theory of gravity and the motion of bodies.

The story is much more long and fun, if you want to know more about it, please take a look online. But you could be asking yourself at this point, what does this story has to do with data science?

The biggest learning we have from that story is that having data, and sometimes a lot of data, is worthless unless you have a good question to answer. This is still true nowadays, and the start of the modern love of data began with statistics.

The Role of Statistics

Fig. 3

I am not going to write a lot about statistics here, but I will point to two specific things that changed the world forever. First is an article called "The Future of Data Analysis" by John Tukey, published in 1962, and the other one is a presentation by Professor Jeff Wu titled "Statistics = Data Science" given in 1997.

These are pretty old references I know, but they are very important. Believe me.

In the article by Tukey, 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…"

This is 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.

Saying that this "new science" is defined by a ubiquitous problem rather than a concrete subject. He then goes on and talks about how to learn and get started with data analysis, and how to become a data analyst and also how to teach it. It is an amazing article that we all should read if we want to understand the beginnings of our field.

In the second piece, 35 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 now that statistics is a part of data science, but why? Because we say that we also need programming, business understanding, machine learning, and more. Maybe, it is just that statistics evolved and now some statisticians became data scientists. But, only some of them.

To understand the portion of statistics and statisticians that became data science and data scientists, we need to read the article "Statistical Modeling: The Two Cultures" by Leo Breiman published in 2001.

Here, he mentions that there is 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 ones consider 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 on 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, i.e., the algorithmic modeling culture, giving examples from his own work and others on how it can solve hard and complex problems.

The Role of Data Science

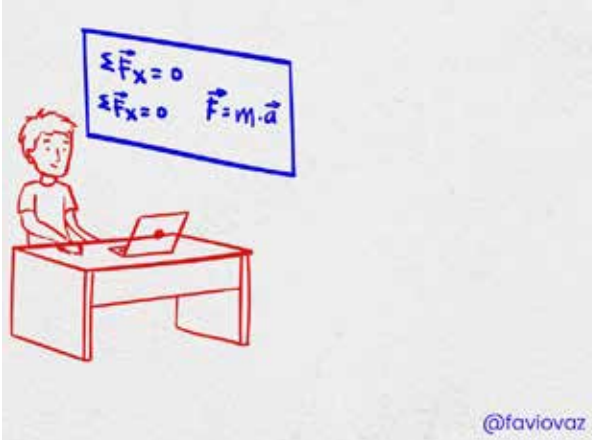

Fig. 4

Data science is the main focus of most sciences and studies right now. It needs a lot of things like AI, programming, statistics, business understanding, effective presentations skills, and much more. That is why, it is not easy to understand or study. But we can do it, we are doing it.

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

Where Are We Going?

A while ago, I published this chart about the interest on semantic technologies over the years.

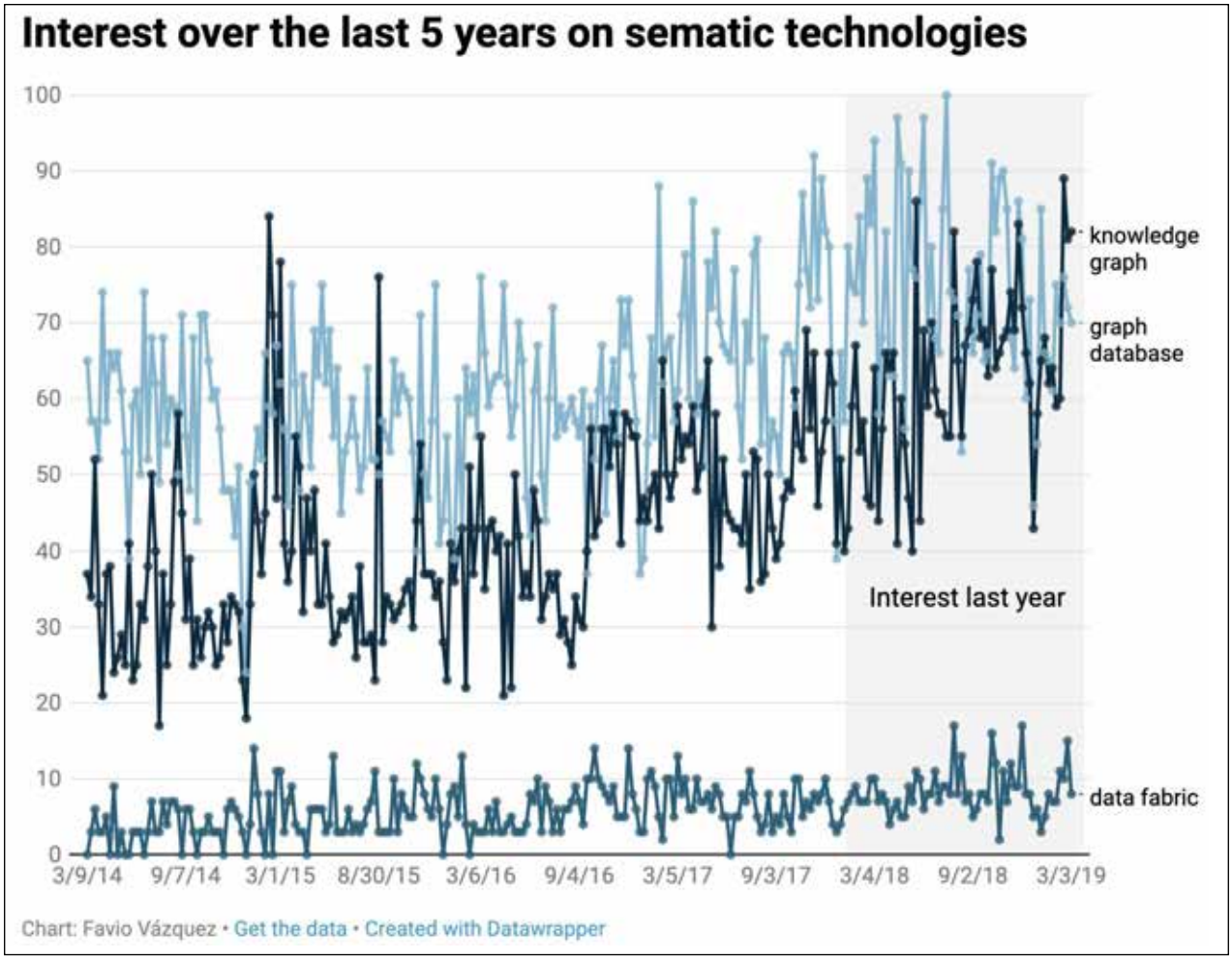

Fig. 5

We can easily see that it is increasing over time. Semantics in this context means the use of formal semantics to give meaning to the disparate and raw data that surrounds us, and also the relationship between signifiers and what they stand for in reality, their denotation.

When we talk about semantics in data, we normally mean a combination of ontology, linked data, graphs and knowledge-graphs, the data fabric, and more. You can read about all of that in the links at the beginning of the article.

But why? Why the shift? The thing is that all data modeling statements (along with everything else) in ontological languages for data are incremental, by their very nature. Enhancing or modifying a data model after the fact can be easily accomplished by modifying the concept.

We normally store data in graphs in these technologies. Whereas relational databases store highly-structured data in tables with predetermined columns and rows, graph databases can map multiple types of relational and complex data. And, it is better for what we have right now.

I have been in countless projects right now, and the common thing is we spend a lot of time trying to make sense of the data we have, and one of the reasons may be that we are not storing the data and its relationship in a good format. The promise of the data fabric is just to support all the data in the company, i.e., how it is managed, described, combined, and universally accessed.

Remember, data and context come first, this new paradigm integrates and harmonizes all relevant data sources — structured and unstructured data alike — using a built-in graph database and semantic data layer. The data fabric conveys the business context and meaning of your data, making it easier for business users to understand and properly utilize.

For me, that is the future of data science. We are moving in a direction where semantic technologies are going to be the standard in every company. But, we would not stop there. All the advances in augmented reality, virtual reality, and more will companion these shift.