Is Data Science a science?

Maxim Scherbak Follow

Mar 12, 2019 · 9 min read ★

What to expect from your first Data Science project? A guide for businesses who wish to dip their toes into Data Science and AI. Part 1

Any successful Data Science project for business is a result of successful collaboration between Data Scientists, Business stakeholders, business experts and many other parties. Recently, there was a handful of articles on Medium for Data Scientists about how to become more business oriented to solve business problems more effectively. But there were not many posts addressing the other side of collaboration, helping business to become more prepared for Data Science and AI projects, which I think is equally important.

This article is a non-technical conceptual introduction to Data Science and AI for for Business people — domain experts, project managers, and stakeholders. Unlike many other non-technical articles explaining some parts of Data Science, AI or machine learning and giving only a small part of a picture, this series of articles will give a helicopter (or even satellite) view of Data Science's place on Business Data Analytics landscape.

A big part of my current role is finding Data science and AI opportunities in businesses. The biggest problem I face when I start any conversation about Data Science and AI is a communication with business people who are not Data Scientists. Not because I use Data Scientific jargon, but mostly because my normal language is understood not in a way I wanted. It took me some time to realise that this happens because of assumptions

people make to fill in gaps in their understanding.

The hype only makes it worse: everybody has heard about Data Science or AI somewhere, most likely from an article describing a new cool thing that Google, Amazon or Facebook have invented. These articles pour buzzwords like Data Science, AI, Machine learning and all other "learnings" on readers not explaining what they are. In these circumstances people have to make their own assumptions to fill these gaps to build a picture in their minds.

> ***These assumptions turn any constructive conversation about Data Science or AI solution into a minefield for Data Scientists to walk on, who can never be sure that people understood him/her correctly and not in their own way.***

It could turn out quite late in the project that 5 people who attended a kick-off meeting had 5 different understandings of it. That will not only deem any project to fail but will also cause much frustration to both businesses and data scientists. Both will learn to stay clear of each other in the future to avoid further failures and, sadly, to miss opportunities that this fruitful collaboration may give.

This is why I start any conversation about Data Science opportunities with new business people from a quick education course, which is based on the system I have developed over the

past 2 years. It is designed to lay a solid foundation for future dialogue about AI and Data Science opportunities by filling typical understanding gaps and explaining common misconceptions. The system is based on numerous projects and conversations I had with business analysts, project managers and directors in years of consulting both in Business Intelligence and Data Science.

> **I found this approach very effective in practice — by just spending a couple of hours for an educational workshop I could achieve multiple goals: dramatically reduce risks from misunderstanding, reduce people's anxiety about security of their jobs caused by a perceived threat from AI, and it also helps to inspire their own thinking, sometimes to a level when they find AI opportunities for their businesses themselves!**

This article is one of the few I'm going to write to cover the system. Although anyone can find it useful, it is designed for people working in Business Data Analytics, who wish to explore Data Science and AI opportunities. It is based on comparing traditional Business Data Analytics and Data Science and explains the following questions: How and why they are different, What are the risks specific to any Data Science project, How to minimise those risks, and what makes a data science project successful.

I'm starting this series of articles from a slightly philosophical question:

Is Data Science a science?

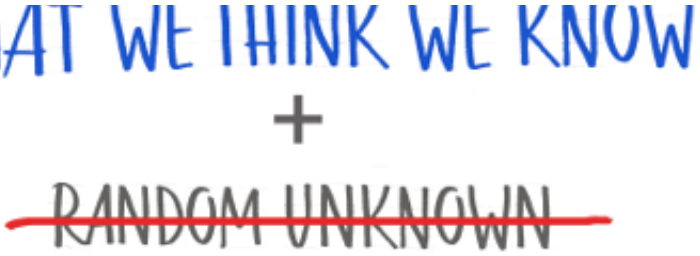
TLDR: Yes, it is! Read below why

According to Wikipedia, science is a is a systematic enterprise that builds and organises knowledge in the form of testable explanations and predictions about the universe. This definition is a bit dry and I don't use it in my talks, instead, I start from what I call the "Grand formula of knowledge":

Our knowledge is what allows us to understand life around us and make predictions. And we are full of this knowledge: we know

language that allows us to communicate, we have social skills that allow us to understand and influence other people, we know traffic rules that help us to drive safely and so on. Just imagine what our life would be without this knowledge — correct, a chaos, the random unknown from the above formula. We would not be able to communicate and predict other people, driving a car would be playing a Russian roulette. The above formula shows that the more we know, less random unknown we will be facing, and the other way around: the less we know the more room for the random unknown.

Living with the unknown is uncomfortable, that's why we naturally want explanations for anything, we want to know about anything around us. We want knowledge that is good enough to explain things and that doesn't have to be a scientific knowledge — anything that helps to explain and predict works, including religious or supernatural. Our collective efforts in looking for answers are paying off — most of the time we live a comfort zone, where our knowledge serves us relatively well and where we can safely ignore the random unknown from above equation and believe:

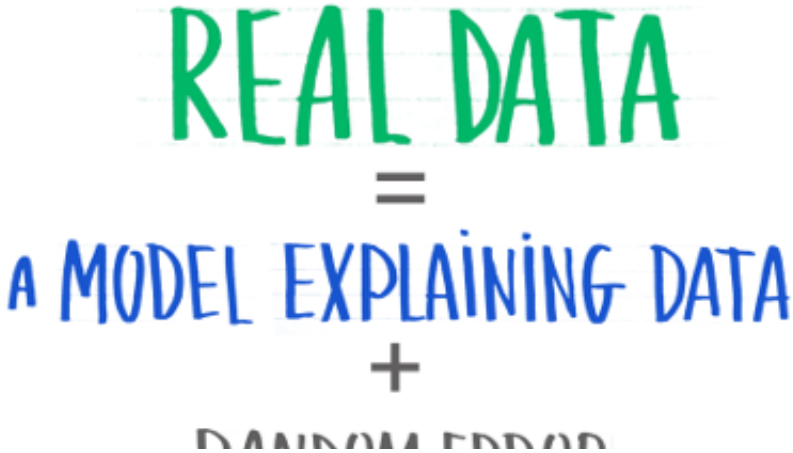


Not all of us prefer comfort, some spend more time outside of this comfort zone dealing more with the unknown — they are scientists, explorers, experimenters. Their goal is to make unknown known and make our lives more safe, comfortable and fun.

Back in the days, my PhD thesis was evaluated against a criteria called “scientific value” — a condition, confirming that the same work hasn’t been done before and possible outcome has value. In the terms of the above formula, this means that my thesis had to deal with some bothering unknown, making it a valuable known.

A Monument of the Discoveries (Padrão dos Descobrimentos) in Lisbon. The figure with the ship in his hands is Henry the Navigator, a Portuguese Prince who is regarded as the main initiator of what would be known as the Age of Discovery, which changed the world at a much bigger scale than the Romans before. The other figures are famous Portuguese navigators, artists, writers and scientists of that time. They all did not prefer comfort.

You probably know now where I'm going. Data science deals with real life data, its goal is building a data model that explains the data as good as possible and that is capable making accurate judgements about new data (predictions). Incorrect predictions, called prediction errors, are random and their amount indicate how good the model explains the real data. Having said that, I can apply the "Grand formula of knowledge" to Data Science beautifully:



Now I can answer the topic question: “Is Data Science a science?” Yes, it is! Like other sciences, it discovers previously unknown facts, it uses systematic approach, it creates reproducible results. In other words, it has all the attributes of a science.

Comparing to other sciences, like physics, Data Science does not explain real life processes directly, but through the data these processes produce. This is what makes data science special — it can explain processes of a different nature using the same methods, as long as these processes produce data. If you look at Data Science competition history on Kaggle, you can find any types of problems from social sciences, all the way to nuclear physics. All these problems can be solved with Data Science! Magic? No, it is just because of a unique approach that Data Science uses — it abstracts itself from the nature of the process, through the data this process produces. For example, for predicting result of a coin flip, Data Science will not care about shape or mass of the coin, force applied to it, earth gravity etc. Instead, it will make predictions based on a data that this coin produced. In order to obtain this data, one just needs to flip the coin multiple times and write down the results.

You may wonder how Data Science is different from Statistics, a

well-established sub-field of Mathematics? It's a very debated question that hasn't been answered and accepted by both Data Scientists and Statisticians yet. There are only opinions and I have an opinion too. Well, from above formula perspective there is no difference — they both do the same. The main difference to me is in how they do it.

Data Science was born in response to new data challenges of the modern digital age: we now have much more data than statisticians had in the past and the demand for practical data research has exploded exponentially in the past decades. All those petabytes that humanity now generates daily are not laying around ready to be taken and used to build models. Unfortunately, all this data is usually scattered across different databases, applications and systems each of which might have its own tricky interface. On the top of that, the industry now expects not only a result of data modelling with some conclusions, they now expect a data product — an application that would deliver benefits and make return of investment. All these new challenges require a very broad set of technical skills, that statisticians didn't have before.

On the other hand, with the amounts of data available, Data Science doesn't usually face a challenge typical for statistics: making judgements about big things from a tiny sample of data. This is a very difficult challenge that required high maths and

statistics skills. Nowadays, when Facebook covers 72% of population of North America and most Data Scientists have access to up to 100% of data that could possibly be available, it is much easier to validate a model on real data rather than spending time for developing perfect theoretical models. As a result, practical Data Science does require good expertise in Stats, but it usually does not demand a Data Scientist to be a top Stats expert.

I believe that Data Science is not just a new fancy name for statistics, it is deeply modernised Statistics that deserves its own name.

A problem of early adoption

Thanks to tech giants, Data Science and AI have already become a part of our personal life through social media, mobile and smart devices and IoT kicking in. In contrast, businesses in general are just starting to explore commercial potential of Data Science and AI. But is does not go very well, unfortunately — most of the projects do not go beyond proof of concept phase. There is no published figure on how many Data Science projects fail yet, but there is a similar measure for projects of previous technology hype — Big Data. According to Gartner, their failure rate can be as high as 85%. I'm sure that this figure is even worse for AI and Data Science projects.

There are many reasons for lack of success in the early commercial adoption of Data Science in big companies, but most of them stem from misunderstanding of Data Science concept: what it is good for, how it is different from things that are familiar and how to use it. For example, a typical misconception among Business Data Analytics people, that Data Science is just a "BI on steroids", that it is the same old Business Intelligence that works much more cleverly. This misconception alone will deem a Data Science project to fail, because it creates a false belief that the old established BI framework would work for Data Science just as well. All they need to do, is to add Data Scientist into the mix. If they only knew how wrong it is, they would not start a project at all!

In the next two parts I will explore Data Science concept and its use in a commercial environment. In the next part, "Is Data Science a BI on steroids?", will touch on the fundamental difference of Data Science from BI, a traditional Data Analytics that companies have been using for years. It will also explain concepts that will be used in the third part "How to benefit from data science projects?".

So watch out for the new parts very soon…

I've worked in Business data analytics for more than 13 years now and have made a journey from an ordinary BI developer to an AI

researcher, developer and evangelist.

You can find me on LinkedIn

Data Science Data Analytics Business Data Analytics Artificial Intelligence

Project Management