What is Data Science?

Alex Sverdlov
alex@theparticle.com

1 What is science?

One way to gain knowledge is to be told things, and to memorize. While this certainly gets us knowledge, the knowledge itself is not *new*. Someone must have it for us to be able to memorize it.

Another way to gain knowledge is to use deduction. By applying logic on things we already know, we can deduce things that weren't previously obvious. For example, if we know statement X is true for 0, and can prove that it is also true for $n + 1$, we can deduce that X is true for all natural numbers.

Notice that both memorization and deduction do not actually create *new* knowledge. It is just recalling or recombining existing facts. Deduction is very powerful, but the knowledge it creates was already there—just not obvious. To gain truly new knowledge, stuff that is not just a combination of existing knowledge, we need something more powerful: we need *inference*.

Inference is our ability to form beliefs and to refine said beliefs based on observations. Implicit in this process is our inability to prove most things we infer—otherwise we would just use deduction to get to that knowledge.

If observations confirm our beliefs, then our beliefs are strengthened. If we craft experiments/observations to try to falsify our beliefs, then we call this process *science*. Said another way: if a belief cannot be falsified via an experiment or observation, then it is not science.

2 What is *data* science?

The short answer is: All science is data science.

A generation before Isaac Newton, an astronomer named Kepler[1] "discovered" the laws of planetary motion: *planets move in elliptical orbits with the sun at the center*.

What makes Kepler's discovery stand out is that it was perhaps the first to be made from data. The story is that Tycho Brahe (another astronomer) made very precise measurements of planetary positions in the sky, and upon his death, Kepler stole that dataset. After about

[1] Johannes Kepler was a German astronomer, mathematician, and astrologer.

nine years of trying to fit different types of circles (and nested Platonic solids) to the data, Kepler stumbled onto an *ellipse*, and the rest is history.

Ellipses are not as elegant as circles. But circles didn't fit the data—ellipses did. After Kepler, data and observations played a key role in all science.[2].

3 Synthetic Domains

When we say science, we often refer to physics, chemistry, biology, etc. These sciences apply the scientific method to understanding the natural world. The data/observations generally comes from observing the real world. e.g. Newton comes up with $F = ma$ and can confirm it against the motion of real world objects.[3]

The term Data Science often has a grander meaning: applying the scientific method on data, which is often not observed from (or generated by) the natural world.

For example, a business interacts with customers. Each transaction is a data point. An observation. Using said observations can we be build a model of this interaction? Can we optimize that interaction to encourage more purchases and less support phone calls from customers?

Domain knowledge plays a very important role in data science. While everyone can relate to the natural world, deep knowledge of various domains is hard to acquire and accurately wield. Data scientists often have to learn more about the domain than about technical aspects of working with data—which translates into variation in primary skill-set: some folks are stronger in business than tech, and vice versa.

The other "problem" that synthetic domains present is high dimensionality. Real world is just 4 dimensions (x,y,z,time); some forms of string theory have 11-dimensions. Data scientists often work with hundreds or thousands of dimensions—which presents its own problems, such as sparsity, etc. A 1-megapixel photo is a point in million dimensional space. Most points in million-dimensional space are not valid photos—no two photos will ever perfectly match each other unless one of them is a copy. If our task is to find similar photos, we cannot just compare pixels.

4 Reductionism

Another key idea in acquiring knowledge is that generally things can be understood by examining their component parts. For example, cakes may be understood by examining the ingredients and cooking instructions.

Applied recursively, cake ingredients may be understood by examining their chemical makeup, which subsequently can be understood by examining the atomic structure of said chemicals, which subsequently can be understood by examining the types of quarks involved, etc.

[2]This is an oversimplification.

[3]How does computer science fit into this?

This reductionist view creates a hierarchy of knowledge and sciences. Physics, chemistry, biology, etc., they're just different levels in the hierarchy.

Not every hierarchy is useful for the problem at hand. We would be hard-pressed to decide if the cake tastes good by examining the underlying quarks.

There is a similar argument regarding books: To understand a book, we need to understand chapters, then paragraphs, then sentences, then words, then letters. So once we know the alphabet, all the books are the same: they are all just different arrangements of the same letters. This is clearly silly.

What this means for data scientists is that things are modeled at a certain level, that is hopefully appropriate for the problems we would like to solve. For example, we could model customers as individuals (almost impossible) or as groups (by age, gender, education, location, etc.).

5 Correlation is not causation

Before we get too excited about data and data science, it is worth looking at problems that data alone cannot answer (no matter how much data we have). The most famous being that correlation is not causation.

Causal relationships are beyond statistics. Even with a lot of observations, the best we can do is say that two variables appear to be correlated or not. We may even be able to fit a function $f(x) = y$ which allows us to calculate one variable from the other.

None of this means that x causes y or that y is somehow derived from x. Even if we first observe x and then at a later time observe y, we cannot say that x causes y.

What this means is that statisticians can look at a million fire & fire-alarm data points, and still not be able to express that fire causes fire-alarm or vice versa.

5.1 Coincidence

For the fire & fire-alarm scenario, our common sense tells us that fire causes fire-alarm to trigger. But what about nameless variables x and y. Suppose we observe x and y together a lot—they appear to be highly correlated. Can we claim that they are somehow connected?

The ugly truth is that with enough data points in many dimensions, there will almost always be pairs of points that are highly correlated by pure chance. Massive studies that attempt to find links between pairs of variables often find such correlations.

6 Counterfactual

Data points are facts. Anything counterfactual is by definition *not in the data*.

Human beings reason about such things all the time: We know a customer bought XYZ, but what would they do if they didn't buy XYZ? Would they buy a different model? A

competitor's ZYX? Wait for XYZ to go on sale?, etc. Getting computers to do that from any amount of data is proving elusive.

7 Past is not the future

There is a famous saying: "past performance is not indicative of future results."

Data is recordings/observations of the past. Using it to reason about the future often works (that is what living creatures do all the time). This predictability depends on the specific characteristics of a particular problem.

7.1 Random vs Chaotic

Random and chaotic are kind of opposites. Random things cannot be predicted short term, but often follow a distribution which allows for some accurate long term predictions (averages, etc.)

Chaotic processes can be predicted short term, but long term predictions depend on accuracy of variables (and since we do not have infinite-precision variables, long term prediction is impossible).

We can imagine processes being both random *and* chaotic.