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 isn't *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 something is true for 0, and can prove that something is true for $n+1$, we can deduce that something is true for all numbers.

Notice that both memorization and deduction doesn't 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 isn't 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.

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 isn't science.

2 What is data science?

The short answer is: All science is data science.

A generation before Newton, an astronomer named Kepler "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, and upon his death, Kepler stole that dataset. After about nine years of trying to fit different types of curves 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.

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.

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 be build a model of this interaction? Can we optimize that interaction to encourage more purchases and less support phone calls from customers?

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.

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

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.).