Inf2 – Foundations of Data Science 2022

Lecture: Introduction and Logistics

22nd September 2022

1 What is data science?

Data science is all about exploring, explaining, and inferring insights from vast amounts of data. The end-goal of this course is to provide students with tools they require to be great data scientists, the lucky practitioners that develop and/or use data-scientific technologies! So let's first define what a data scientist does.

What is a data scientist? Data scientists are able to turn raw data into understanding, insight and knowledge. Some of the activities that data scientists do on a daily basis include the following:

- Find, collect, check and clean data.
- Explore data and infer knowledge and insights.
- Explain and interpret these inferences.
- Communicate inferences about the data to others.
- Make prediction about future trends.

This is why data scientists need to employ interdisciplinary tools, from statistics, artificial intelligence and machine learning. We will be addressing all of these different aspects in the course, and have closely aligned our learning goals with the varied skill set that data scientists need to be successful in the field.

This process is broken up into steps and shown in Figure 1. We will expand on all of the steps of the data science cycle in this course.

The data science process. We can see data science as a process, comprising interconnected steps (Figure 1).

The data explosion: volume, variety and velocity Data is always around us, and it always has been. But the development of technology has made it possible to collect and store vast amounts of data. Just to get some perspective, each day on Earth we generate over 500 million tweets (how many of these are generated by bots?), 294 billion emails, 4 million gigabytes of

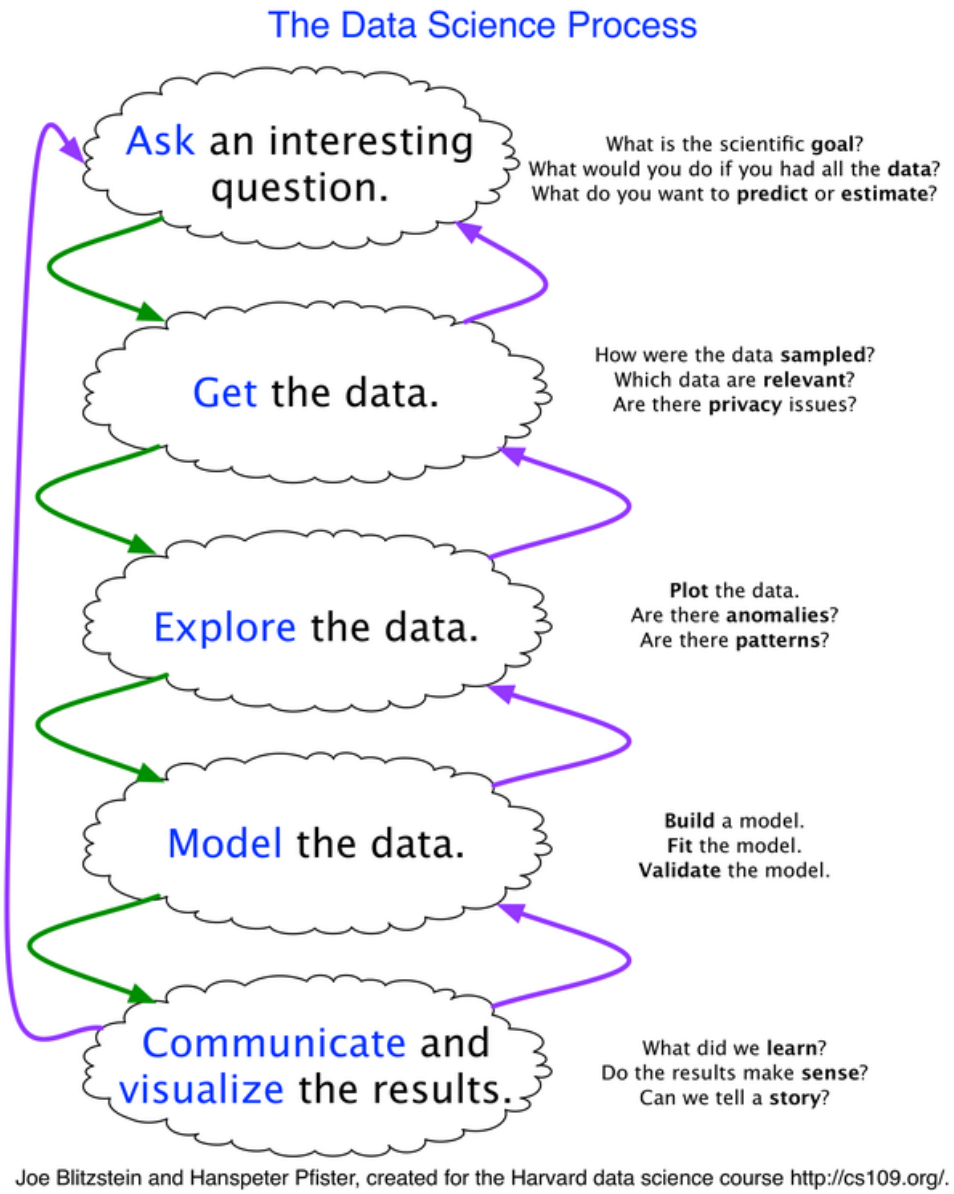

Figure 1: The data science process. Credit: Joe Blitzstein and Hanspeter Pfister, created for the Harvard data science course CS109

Facebook data, 65 billion WhatsApp messages and 720,000 hours of new content added daily on YouTube.

In 2018, the total amount of data created, captured, copied and consumed in the world was 33 zettabytes (ZB) – the equivalent of 33 trillion gigabytes. This grew to 59ZB in 2020 and is predicted to reach a mind-boggling 175ZB by 2025. One zettabyte is 8,000,000,000,000,000,000,000 bits (Vopson, 2021).

Parallel to this explosion of data generation, there is also consistent improvement in computational methods; we are now able to mechanise some aspects of data analysis that data scientists care about. So, while the amount of digital data in the world doubles every two years, so does the computational power that lies at our disposal to analyse it. One should state that the speed of computation, which was commonly considered to double every two years, may have plateaued.

How does data science relate to other fields? Data Science, Statistics, Machine Learning. These are some of the terms used in the scientific literature and popular media. Let's try to sort through the confusion.

Data science is a multidisciplinary field which uses scientific methods, processes, and systems in a range of forms. Both statisticians and data scientists care about analysing and explaining data. Indeed, some of the methods used by data scientists are taken from statistics, and we shall learn them in this course.

Statisticians focus on using structured models and parameter estimation to quantify uncertainty in data. Data science includes statistical methods, but in addition, data scientists need to do a lot of work on processing the data itself and check that the data makes sense. Data scientists often deal with huge databases, which is why they rely heavily on computational methods for their analysis. Historically, statisticians have focused on small quantities of data.

Machine learning is all about algorithms that are able to learn from data. Data science uses many tools from machine learning, and we will study some of these in the course. But data scientists begin by exploring the data and formalise questions that can be asked on the data. They care not only about making predictions about trends in the data, but in explaining why these trends arise. They need to communicate quantitative and qualitative arguments to the general public. Often, they also care about supporting practitioners in the field (e.g., alerting teachers to struggling students based on their interactions with educational software). The value of a machine learning algorithm is measured by its performance on unseen data. But the value of a data science analysis also needs to consider the human in the loop. We're (happily) not at the stage when these complex tasks can be replaced with computer algorithms. There is a lot of room for skill and creativity that cannot be automated.

In all these fields, it's important to know where the data comes from, how it's been collected, and to be able to reason about whether a dataset has been collected ethically, or if the project we are undertaking is ethical.

Examples of data science at work Let's consider two examples of data science at work. Both of these examples are meant to highlight the inherent biases that unfortunately exist in

society, and that also arise in the data sets that we generate.

- Example of gender tropes in films. In this example we will see the gender bias reflected in script writing in movies.
- Case study: COMPAS. In this example we will describe racial bias in the criminal justice system.

Data science career prospects Data science is a lucrative field. Although the study and analysis of data started off in academia, industry is increasingly taking a leading role and applying and developing data science methods. Demand is high for data professionals—data scientists and mathematical science occupations are growing at a significant higher rate than in other high tech fields.

2 Course logistics

- Learning outcomes
- Course activities
- Assessment

3 Telling stories with data

References

Vopson, M. M. (2021). 'The world's data explained: how much we're producing and where it's stored'. *The Conversation* URL `https://theconversation.com/the-worlds-data-explained-how-much-were-producing-and-where-its-all-stored-159964`