Introduction

1 What is data analysis?

Data analysis is a set of methods for extracting information from a data set. It's also known as statistical learning. The idea is to use statistical models to understand how data is structured and how it interacts with each other.

> Example
>
> Let's imagine you work for the United Nations (UN). Your mission is to analyze life expectancy around the world. To do this, you'll have a measure of life expectancy in each UN member country, of course, but also GDP per capita, health expenditure, fertility rate, urbanization rate, education level of the country, and so on. The aim of data analysis is to find links between these different variables and the variable of interest, life expectancy, to visualize these data, and eventually to predict life expectancy from the other variables.

2 Course objectives

In this course, we aim to introduce methods that allow us to study a "high-dimensional" dataset (in the sense that we can't simply graph all the variables for each observation) without having to resort to a probabilistic model. The various techniques we'll be looking at can be used to:

- visualize data
- reduce data size;
- identify certain relationships between variables;
- divide the dataset into groups/classes.

This course is not intended to be exhaustive, in the sense of presenting all possible methods. Nor is it intended to be state-of-the-art, in the sense that it will not cover the latest developments in machine learning. Nor is it a programming course.

To finish this introduction, here is a quote from *Statistical Rethinking* by Richard McElreath (McElreath 2020) particularly accurate in this course.

> Statistics courses [...] tend to resemble horoscopes. There are two senses to this resemblance. First, in order to remain plausibly correct, they must remain tremendously vague. This is because the targets of the advice, for both horoscopes and statistical advices, are diverse. But only the most general advice applies to all cases. A horoscope uses only the basic facts of birth to forecast life events, and a [...] statistical guide uses only the basic facts of measurement and design to dictate a model. It is easy to do better, once more detail is available. In the case of statistical analysis, it is typically only the scientist who can provide that detail, not the statistician. Second, there are strong incentives for both astrologers and statisticians to exaggerate the power and importance of their advice. No one likes an astrologer who forecasts doom, and few want a statistician who admits the answers as desired are not in the data as collected. Scientists desire results, and they will buy and attend to statisticians and statistical procedures that promise them. What we end up with is too often *horoscopic*: vague and optimistic, but still claiming critical importance.

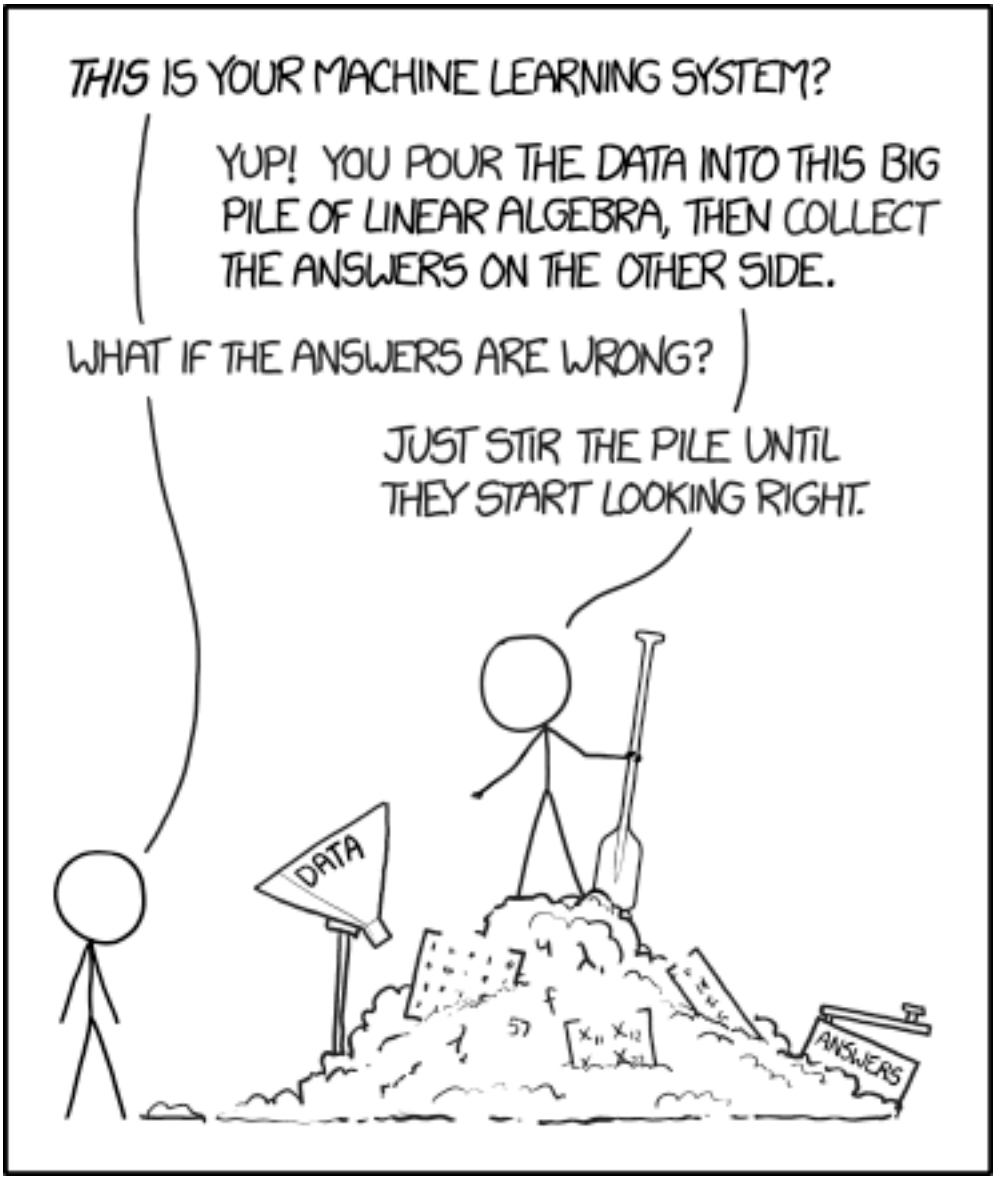

Figure 1: Machine learning (xkcd:1838).

References

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. 2nd ed. New York: Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608.