Data Handling: Import, Cleaning and Visualisation

Lecture 1: Introduction

Ulrich Matter

20/09/2018

1 The recent rise of big data and Data Science

Lower computing costs, a stark decrease in storage costs for digital data, as well as the diffusion of the Internet have over the last few decades led to the development of new products (e.g., smartphones) and services (e.g., web search engines, cloud computing). Figure 1 quantifies some of the key characteristics of this technological change. A side product of these developments is a stark increase in the availability of digital data describing all kind of every-day human activities (Einav and Levin 2014; Matter and Stutzer 2015). As a consequence, new business models and economic structures are emerging with data as their core commodity (i.e., AI-related technological and economic change). For example, the current hype surrounding 'Artificial Intelligence' (AI), largely fueled by the broad application of machine-learning techniques such as 'deep learning' (a form of neural networks), would not be conceivable without the increasing abundance in large amounts of digital data on all kind of socio-economic entities and activities. In short, without understanding and handling the underlying data streams properly, the AI-driven economy cannot function. The same rationale applies, of course, to other ways of making use of digital data, be it traditional big data analytics or scientific research (e.g., applied econometrics).

The need for proper handling of large amounts of digital data has given rise to the interdisciplinary field of 'Data Science' as well as an increasing demand for 'Data Scientists'. While nothing within Data Science is particularly new on its own, it is the combination of skills and insights from different fields (particularly Computer Science and Statistics) that has proven to be very productive in meeting new challenges posed by a data-driven economy. In that sense, Data Science is rather a craft than a scientific field. As such, it presupposes a deeper and more practical understanding of the matter at hand (data) than the scientific disciplines Computer Science and Statistics from which it borrows its methods. This is often illustrated in the 'Data Science' Venn-Diagram (see Figure 2), reflecting the combination of knowledge and skills from Mathematics/Statistics, substantive expertise in the particular scientific field in which Data Science is applied, and 'hacking skills', that is, the skills necessary for *acquiring, cleaning, and manipulating* massive amounts of electronic data. It is exactly this skill-set our course will focus on.

Moreover, this course will revisit and apply/integrate concepts learned in the introductory Statistics course (3,222), and will generally presuppose 'substantive expertise' in economics.

2 A brief introduction to data

In order to better understand the role of data in today's economy and society, we have a look at the usage forms and purposes of data records in human history. In a second step, we look at how a computer processes data. Finally, we transfer the concept of analogue data storage to the digital realm.

2.1 Data in human history

Throughout human history, the recording and storage of data has primarily been motivated by measuring, quantifying, and keeping record of both our social and natural environments. Early on, the recording of data has been related to economic activity and scientific endeavor. The neolithic transition from hunter-gatherer societies to agriculture and settlements (the economic development sometimes referred to as the 'first industrial

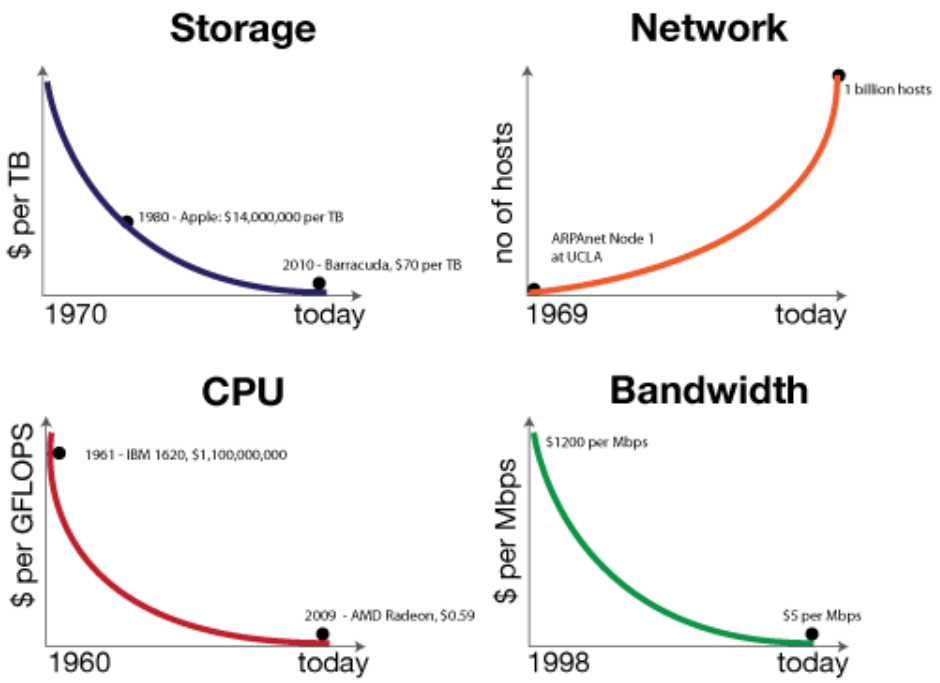

Figure 1: Source: http://radar.oreilly.com/2011/08/building-data-startups.html.

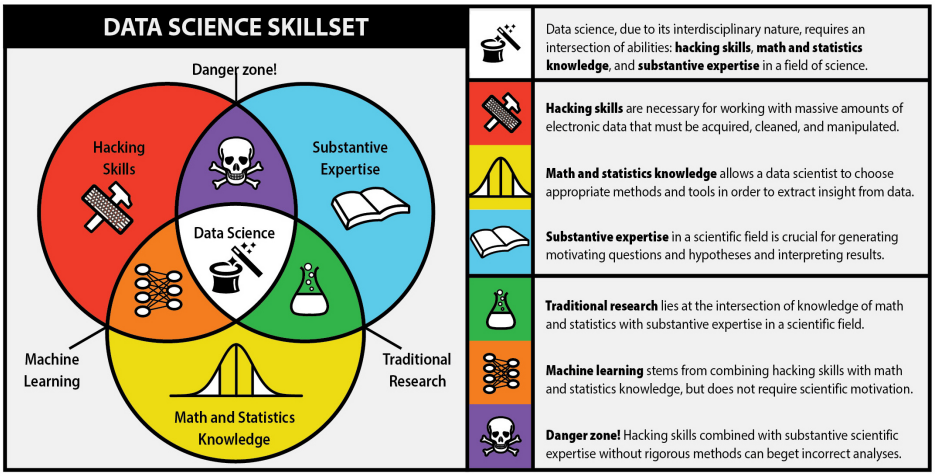

Figure 2: Source: http://berkeleysciencereview.com/how-to-become-a-data-scientist-before-you-graduate/.

Figure 3: YBC 7289. Photo by Bill Casselman.

revolution'), came along with a division of labor and more complex organizational structures of society. The change to agricultural food production had on the one hand the effect that more mouths could be fed, but on the other hand also that food production would need to follow a careful planning (the right time to seed and harvest) and that the produced food (e.g. grains) would partly be stored and not consumed entirely on the spot. It is believed that partly due to these two practical problems, keeping track of time and keeping record of produced quantities, neolithic societies started to use signs (numbers/letters) carved in stone or wood. Keeping record of the time and later measuring and keeping record of produce quantities in order to store and trade, likely led to the first 'data sets'. Simultaneously, the development of mathematics, particularly geometry took shape.

References

Einav, Liran, and Jonathan Levin. 2014. "Economics in the Age of Big Data." *Science* 346 (6210): 1243089–1–1243089–6. doi:10.1126/science.1243089.

Matter, Ulrich, and Alois Stutzer. 2015. "pvsR: An Open Source Interface to Big Data on the American Political Sphere." *PLOS ONE* 10 (7). Public Library of Science: 1–21. doi:10.1371/journal.pone.0130501.