Data Handling: Import, Cleaning and Visualisation

Lecture 1: Introduction

Prof. Dr. Ulrich Matter
(University of St. Gallen)

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 of 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).

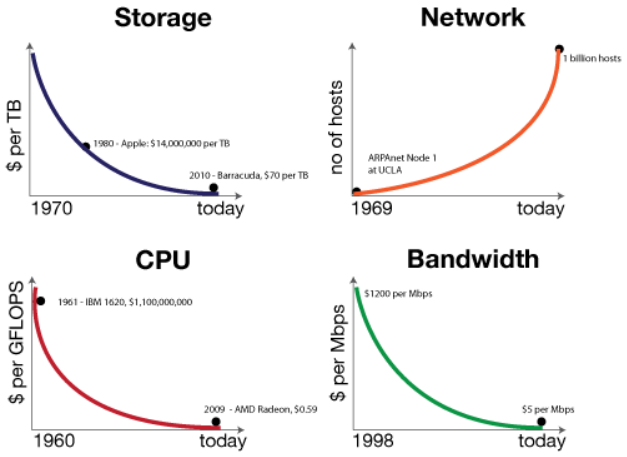

Figure 1: Source: http://radar.oreilly.com/2011/08/building-data-startups.html.

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 more practical and broader understanding of the matter at hand (i.e., data) than traditional Computer Science and Statistics from which Data Science borrows its methods. The various facets of this new craft are 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.

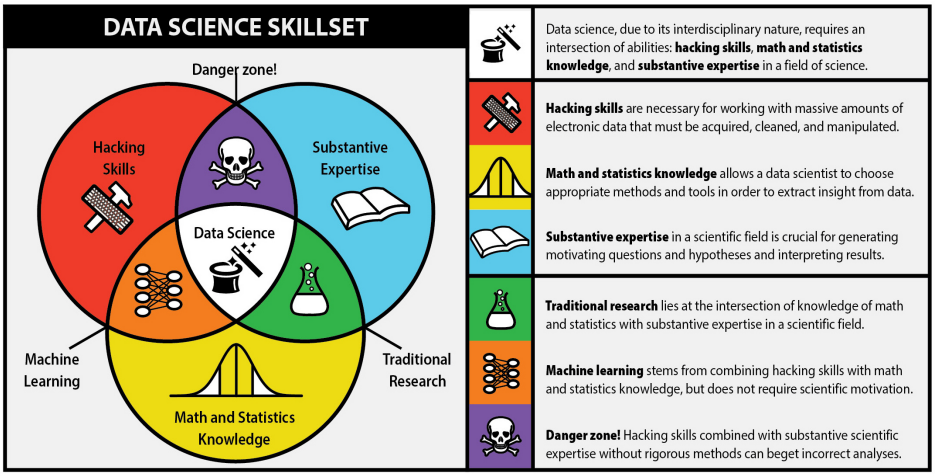

Figure 2: The 'Data Science Venn Diagram'. Source: http://berkeleysciencereview.com/how-to-become-a-data-scientist-before-you-graduate/.

Moreover, this course will revisit and apply/integrate concepts learned in the introductory Statistics course (3,222), and will generally presuppose 'substantive expertise' in undergraduate economics. Finally, the aim is to give you first practical insights into each part of the data science pipeline.

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.