An introduction to applied data science

A 5-day workshop towards practical big-data applications and analytics

Eduardo Barbaro and Coen Jonker

Data Scientistists at Mobiquity Inc

About the workshop and this manual:

Welcome to "An introduction to applied data science" a 5-day original workshop thought and written by Eduardo Barbaro and Coen Jonker. At the end of this week you will be able to (i) understand the basics of data science, (ii) mathematically describe data, and (iii) munge[1] it into an easy and communicable form. As the course develops, we will teach you how to access and look at data in innovative ways, as well as to extract value from big data. We aim a significant part of this training to help you to create/interpret a diverse set of numerical models, as well as to calculate descriptive analytics while fully understanding their meaning. We also focus on presentation, and how to show your results in beautiful smart charts and tables.

Although all the computations will be performed in Mobiquity's AWS cloud, we will guide you through all the necessary steps to install any software necessary for this course in your local machines. We use Mobiquity's AWS cloud to get the benefit of running our code in a fast, secure, and controlled environment.

This manual is divided in 11 independent short chapters. In Chapter 1 we set the stage with a short introduction to then answer together a bold question: "But what is data science?". In Chapter 2 we will spend some time to recap some basic mathematical concepts needed throughout the entire workshop. Chapters 3 and 4 are devoted to (important) practicalities, such as installing software/libraries and getting familiar with the AWS cloud environment. We also teach you some basic programming skills, so we are all on the same page. We move forward to our first data harvest, in Chapter 5. We teach you how to manipulate data, as well as to properly describe and clean it. Chapter 6 covers the basics of designing a smart numerical experiment.

Chapters 7 to 10 cover more advanced topics such as machine learning tools, databases (SQL and no-SQL) as well as a mapreduce implementation (Hadoop). In particular, in Chapter 10 we show in detail how to load, transform, and extract value from real big data. Finally, in Chapter 11 we come back to basics to show you rudiments of plotting and how to communicate your results.

We wish you a pleasant learning!

[1] describes the constructive operation of tying together systems and interfaces that were not specifically designed to interoperate

Contents

1

A historical view on data science

In this Chapter we briefly try to shed some light on the question "what is data science?". First of all, it is important to realize that data science is not a new concept. There are many complex definitions out there, but in our view, data science is simply the coupling of very well-established disciplines, such as mathematics and statistics, with a relatively young discipline, computer sciences.

Have a look at Fig. 1.1. It shows the evolution of 5 fundamental disciplines related to data science (Mathematics, Statistics, visualization, technology, computer science). Note that this cartoon depicts events as old as the "invention" of modern calculus by Newton/Leibniz (in the 17^{th} century) or foundation of probability theory by Cardano (in the 16^{th} century). You can see that back in those times all the disciplines were really self-contained areas without much interconnection. As we enter the 20^{th} century things start to get more interconnected. Statistics and Mathematics cannot be separated so clearly any more (e.g. stochastic models, survival models), and computer sciences and technology started to gain more terrain.

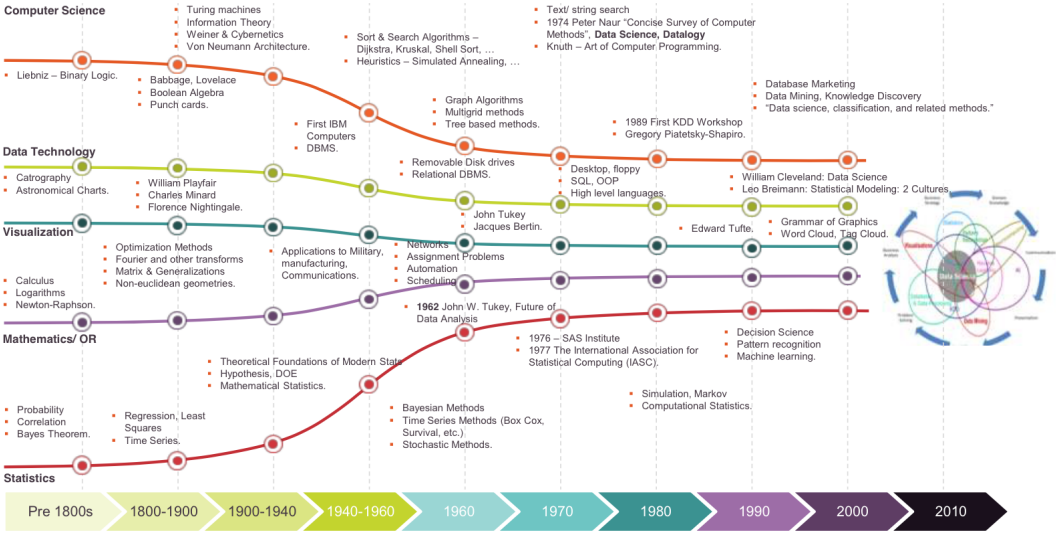

Figure 1.1: Most important events in the history of data science. *Credit: Mamatha Upadhyaya*

As early as 1962, John Wilder Tukey, an American Mathematician, wrote in his "The future of data":

> *For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt. I have come to feel that my central interest is in data analysis. Data analysis, and the parts of statistics which adhere to it, must take on the characteristics of science rather than those of mathematics data analysis is intrinsically an empirical science.*

That served as inspiration to Tukey who, in 1977, published his most well-known book "Exploratory Data Analysis". In that book, he argued that exploratory and confirmatory data analyses must proceed alongside. Also in the 1970's, Peter Naur, Danish computer science pioneer, published "Concise Survey of Computer Methods". In this book, the term data science was used for the first time. According to him, data science can be defined as:

> *The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.*

In the beginning of the 21^{st} century all the disciplines related to data science merged. In 2008, Jeff Hammerbacher (Facebook) and DJ Patil (LinkedIn) used the terminology "Data scientist" to define their work and teams. Since then, the use of this terminology has fully infiltrated the vernacular, and did not stop growing.

Today, data science and computer sciences (through machine learning) have been put together as almost synonyms. There are new terminologies appearing every year, such as Deep Learning, Big Data, Data Mining. Today, we understand that data science and big data do not mean (just) "lots of data". Instead, it means the creation of a new paradigm in how we do analysis and combine the use of our traditional tools (Mathematics and Statistics) with the technology available nowadays.

2

Some basics and beyond

3

Practicalities: Software Installations and AWS cloud

4

Practicalities: Programming language and libraries

5

Harvesting data

5.1 Data Manipulation

5.2 Analytics 1: Describing your data

5.3 Analytics 2: Cleaning your data

6 Numerical Modeling

6.1 Basics

6.2 Designing smart numerical experiments

7

Machine learning tools

7.1 Supervised methods

7.2 unsupervised methods

8 Databases

8.1 SQL

8.2 NoSQL

8.2.1 MongoDB

9 Map Reduce

9.1 Hadoop

10
Big Data

11 Communicating Results

11.1 Visuals