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.

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, Chapter 11 covers the basics of plotting and communicating results. 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.

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

But what is data science?

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