

Master in Data Science – Introduction

Israel Herraiz, Igor Arambasic

October 17, 2018

Contents

- 1** The Data Scientists
- 2** The Data Science Process
- 3** Program of the Master in Data Science
- 4** Recommendations to follow the master
- 5** This week's session

1 The Data Scientists

2 The Data Science Process

3 Program of the Master in Data Science

4 Recommendations to follow the master

5 This week's session

Why data science?

The intelligente use of data:

- has become a source of competitive advantage
 - Better knowledge about the market
 - Better knowledge about your customers

Why data science?

The intelligente use of data:

- has become a source of competitive advantage
 - Better knowledge about the market
 - Better knowledge about your customers

The data science process

Data → Information → Decision

Why data science?

The intelligente use of data:

- has become a source of competitive advantage
 - Better knowledge about the market
 - Better knowledge about your customers

The data science process

Data → Information → Decision

The goal

Data-Driven Decision Making

What is a data scientist?

Key skills

- Fitting in an **organization**, leading projects in a heterogeneous environment, aligning with strategy
- Data-Analytic **Thinking**
- How to extract **knowledge** from data

The three facets of a data scientist

- Functional (domain knowledge)
- Analytical (how to extract knowledge from data)
- Technical (how to implement the data science process)

What is a data scientist?

Key skills

- Fitting in an **organization**, leading projects in a heterogeneous environment, aligning with strategy
- Data-Analytic **Thinking**
- How to extract **knowledge** from data

The three facets of a data scientist

- Functional (domain knowledge)
- Analytical (how to extract knowledge from data)
- Technical (how to implement the data science process)

Question

- Are the three facets equally important?
- Which facet is the most important?

Skills demanded in job postings

Analytical skills

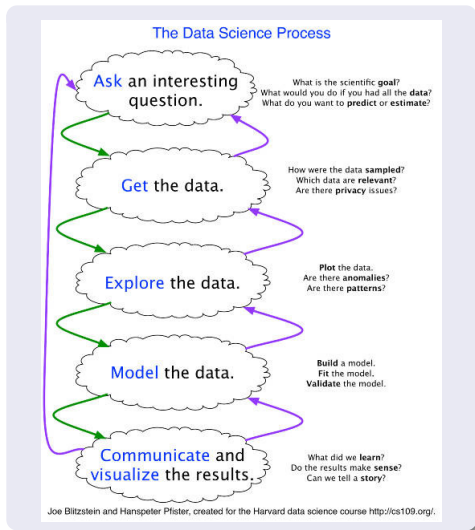
- **Machine Learning**
- **Statistics**
- Bias towards engineering and scientific backgrounds

Technical skills

- **R** or SAS for statistical programming
- **Python**
- **SQL**
- **Hadoop, Spark and Big Data** (covered only partially)
- Visualization technologies (**Tableau**, Qlik, **D3.js**)
- Excel

- 1 The Data Scientists
- 2 The Data Science Process**
- 3 Program of the Master in Data Science
- 4 Recommendations to follow the master
- 5 This week's session

The Data Science process



- 1 Ask a question
 - Domain expertise
- 2 Get and prepare the data
 - Python, Pandas
- 3 Explore the data
 - Pandas, Matplotlib, R, Spark, Tableau
- 4 Model the data
 - R, Spark
- 5 Communicate the results
 - D3.js, Tableau

<http://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html>

But it is always iterative...

Very rarely (never?) you will do it in just one pass



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

What kind of questions can we answer to?

Machine Learning & Statistics Applications

Supervised Learning

- Classification
- Regression
- Ranking

Unsupervised Learning

- Clustering
- Association Mining
- Segmentation
- Dimension Reduction

Reinforcement Learning

- Decision Process
- Reward System
- Recommendation Systems

<http://www.kdnuggets.com/2015/09/questions-data-science-can-answer.html>

What kind of questions can we answer to?

Classifications of questions

- Is this A or B ?
 - Binary classification
- Is this A , B , C or D ?
 - Multi-class classification
- Is this normal or weird?
 - Anomalies detection
- How much or how many?
 - Regression
- How is this data organized?
 - Unsupervised learning
 - Dimensionality reduction
- What should I do now?
 - Recommendation systems

- 1 The Data Scientists
- 2 The Data Science Process
- 3 Program of the Master in Data Science**
- 4 Recommendations to follow the master
- 5 This week's session

The (main) program

■ Intro to data science

- Setup environment, working with the command line, intro to Git
- Intro to SQL

■ Data analysis

- Python for Data Analysis
- Intro to Statistical Programming (R)

■ Machine Learning and Statistics (R, Python)

- Including a brief introduction to **Deep Learning**

■ Big Data

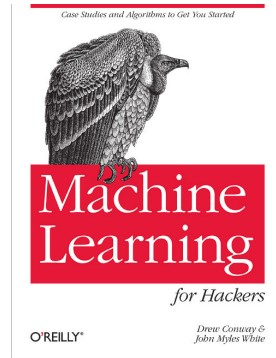
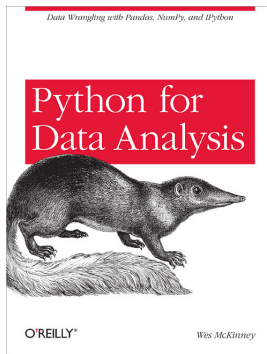
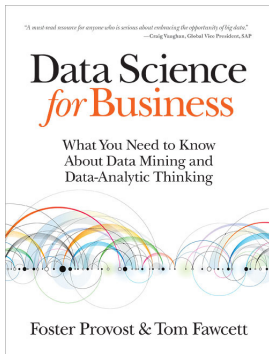
- Very brief intro to Hadoop and Spark, using Python

■ Visualization and Business Intelligence

- Dashboards development with Tableau
- Brief intro to D3.js

- 1 The Data Scientists
- 2 The Data Science Process
- 3 Program of the Master in Data Science
- 4 Recommendations to follow the master**
- 5 This week's session

Recommended books



Share promiscuously

You need visibility

- Use the social networks to position yourself as a data scientist
 - **Crucial:** Profile in [LinkedIn](#) and share, share, share!
 - Share frequently about your progress in the master
- Share your code in [Github](#) or Bitbucket
 - Try to use it frequently, so it always shows recent activity
 - Don't worry if you don't know Git, we will see in the first session
- Don't mix (too much) your personal postings with your data scientist postings
 - For instance, use Facebook for your personal network, and Twitter for your professional postings

Recommended readings

Amadeus Data Scientists series

<http://www.amadeus.com/blog/tag/data-scientist/>

Guía de las profesiones de Internet

<http://www.avanzaentucarrera.com/llegaraser/profesiones-y-profesionales-de-internet>

Take away

Key ideas to remember during the master

Take away

Key ideas to remember during the master

- Extracting useful **knowledge from data** to solve **business problems** can be treated systematically by following a **process** with reasonably **well-defined stages**.

Take away

Key ideas to remember during the master

- Extracting useful **knowledge from data** to solve **business problems** can be treated systematically by following a **process** with reasonably **well-defined stages**.
- From a **large mass of data**, **information technology** can be used to **find informative descriptive attributes** of entities of interest.

Take away

Key ideas to remember during the master

- Extracting useful **knowledge from data** to solve **business problems** can be treated systematically by following a **process** with reasonably **well-defined stages**.
- From a **large mass of data**, **information technology** can be used to **find informative descriptive attributes** of entities of interest.
- If you look too hard at a set of data, you will find something – but it **might not generalize beyond** the data you are looking at.

Take away

Key ideas to remember during the master

- Extracting useful **knowledge from data** to solve **business problems** can be treated systematically by following a **process** with reasonably **well-defined stages**.
- From a **large mass of data**, **information technology** can be used to **find informative descriptive attributes** of entities of interest.
- If you look too hard at a set of data, you will find something – but it **might not generalize beyond** the data you are looking at.
- Formulating **data science solutions** and evaluating the results involve thinking carefully about the **context in which they will be used**.

Take away

Key ideas to remember during the master

- Extracting useful **knowledge from data** to solve **business problems** can be treated systematically by following a **process** with reasonably **well-defined stages**.
- From a **large mass of data**, **information technology** can be used to **find informative descriptive attributes** of entities of interest.
- If you look too hard at a set of data, you will find something – but it **might not generalize beyond** the data you are looking at.
- Formulating **data science solutions** and evaluating the results involve thinking carefully about the **context in which they will be used**.
- It's not about the technologies. Technology will always change very fast. **Learn the concepts**, apply them with technology. **Be open to learning** new technologies (and sometimes it will also imply learning new concepts). **The only constant is change**.

- 1 The Data Scientists
- 2 The Data Science Process
- 3 Program of the Master in Data Science
- 4 Recommendations to follow the master
- 5 This week's session**

Program for this week

Intro to data science

We have just completed this part.

Program for this week

Intro to data science

We have just completed this part.

The environment

Setting up the environment with VirtualBox and Ubuntu

Program for this week

Intro to data science

We have just completed this part.

The environment

Setting up the environment with VirtualBox and Ubuntu

Starting with Git

Intro to Git, and the importance of sharing our source code

Program for this week

Intro to data science

We have just completed this part.

The environment

Setting up the environment with VirtualBox and Ubuntu

Starting with Git

Intro to Git, and the importance of sharing our source code

Getting familiar with the command line

Many times the shell command line is enough to answer to a lot of questions