Master in Data Science
ed 23

Guillermo Ribeiro, Juan Arevalo, Daniel Mateos, Igor Arambasic

About Igor…

Overview

• Data Scientist (10+ years)
• Data Science Manager (10+ years)
• Big Data (5+ years)

2002-2014@GAPS
Grupo de Aplicaciones de Procesado de Señal

2014-@AMADEUS

• PhD in Telecommunications Science, Universidad Politécnica de Madrid, 2008

Amadeus

What do I do?

Head of Airlines Data Science Unit
Responsible of DE, DS and Devops

What I used to in my daily life?

Scala, **Spark**, Scoobi, **Python**, **Pandas**, **Shell**
SQL, Impala, Hive, Cassandra,
Hadoop, HDFS, Hue, Ssh
Tableau, Qlik
Jenkins
Git

About Dani…

Overview

• Data scientist (6+ years)
• Big Data (6+ years)

2008-2013 Molecular Biologist and Bioinformatician @ CNIC (Madrid)

• PhD in Molecular Biology, Universidad Autónoma de Madrid, 2013

Current

What do I do?

Lecturer @ IE, KSchool
Independent Data Scientist

Skills

Python, Spark, Scala,
ML, Dataviz
AWS

About YOU!

- Quick introduction
  - What is your background
  - Why do you want to do data science?
  - What do you expect from this Master?
  - How did you find us?
  - Why did you choose us and not some other DS master?

Before we start…

- Do we **ALL** have access to basecamp?
- Do we **ALL** have the VM, VirtualBox, VirtualBox Extensions downloaded?
- Do we **ALL** have the Github account?

Agenda

Introduction

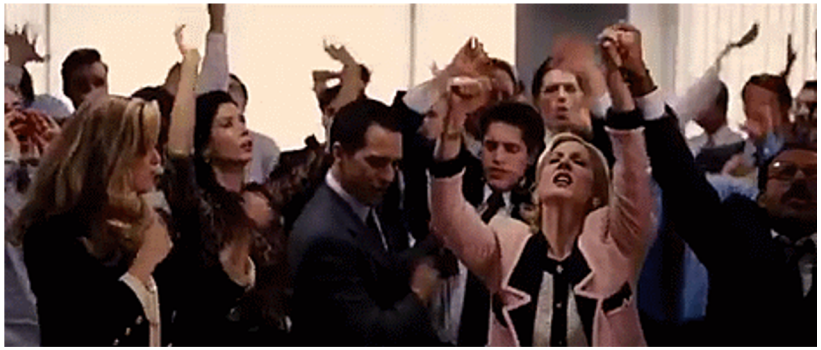
What is Data Science? Where do we fit? What do we do? What does Data science project life cycle look like?

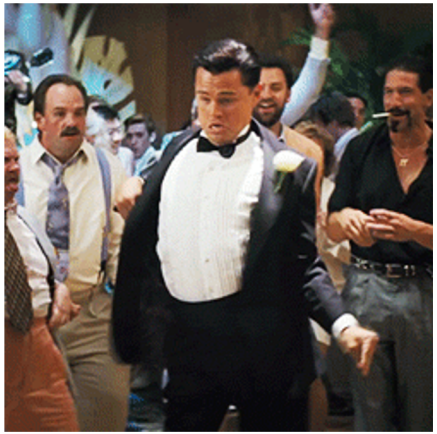
Program of the Master in Data Science

What job position can you apply for after the master?

Takeaways

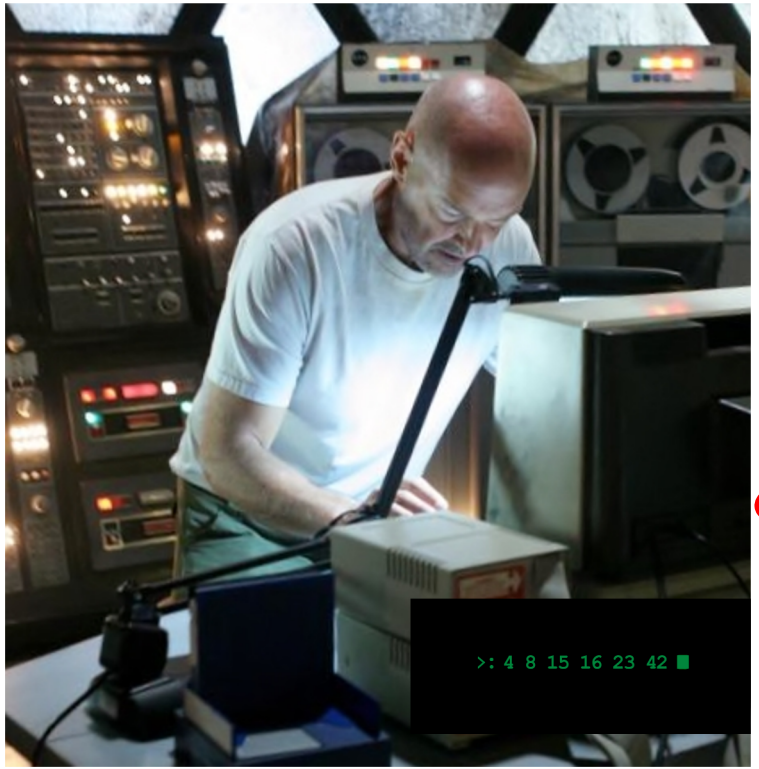
This is what people think we do…





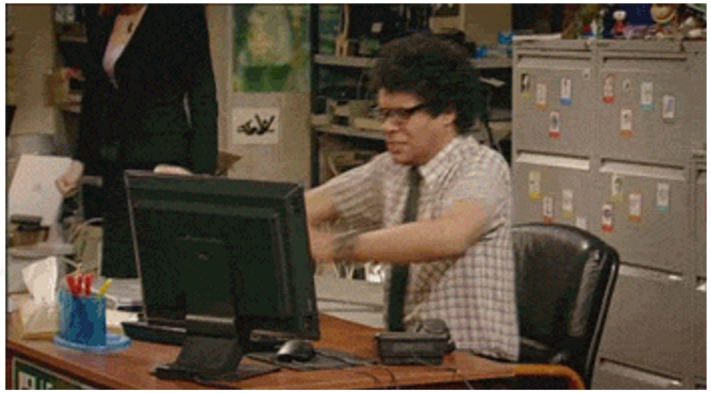
FIESTA!!!

... What WE think we do…



Saving the WORLD!!!!!

Our daily job consist of resolving the problems with DATA.



Why are companies interested in data science?

- All companies have data!!! ALL !!!! The data allows you to get to know (better) your customers and market and optimize their business!

    - IKEA, MAPFRE, AMADEUS, BBVA, CEPSA, ZARA, RENFE etc

- Why are some companies exploit more data than others?

    - Renfe vs Iberia/Air Europa/Vueling?

        - Which origin/destination?
        - At which price? At which time?
        - How much of each service (Business/Premium/Economy)
        - Through which channel to make the offer?
        - Passengers data!

            when you buy, how do you pay, through which channel, how often, nationality, age …

ALL companies should be DATA driven!!!

- The intelligent use of data has become a source of competitive advantage
- **When a company starts hiring DS their goal is Data-Driven Decision Making**
- Data → Information → Decision

“But we have business analysts and experts to drive our business!”

However…

- these kind of decision are based on experience, sensations, sentiment.
- Human bias lead to imperfect decision!

Data Science:

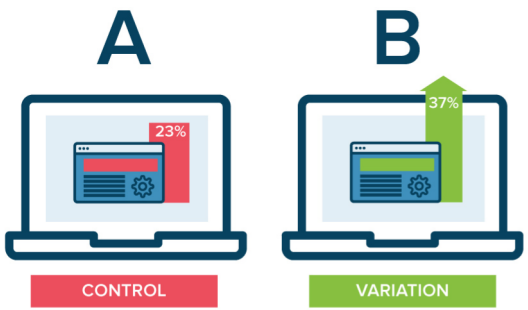
- The problem solved with the same approach over the same data set gives always the same answer.

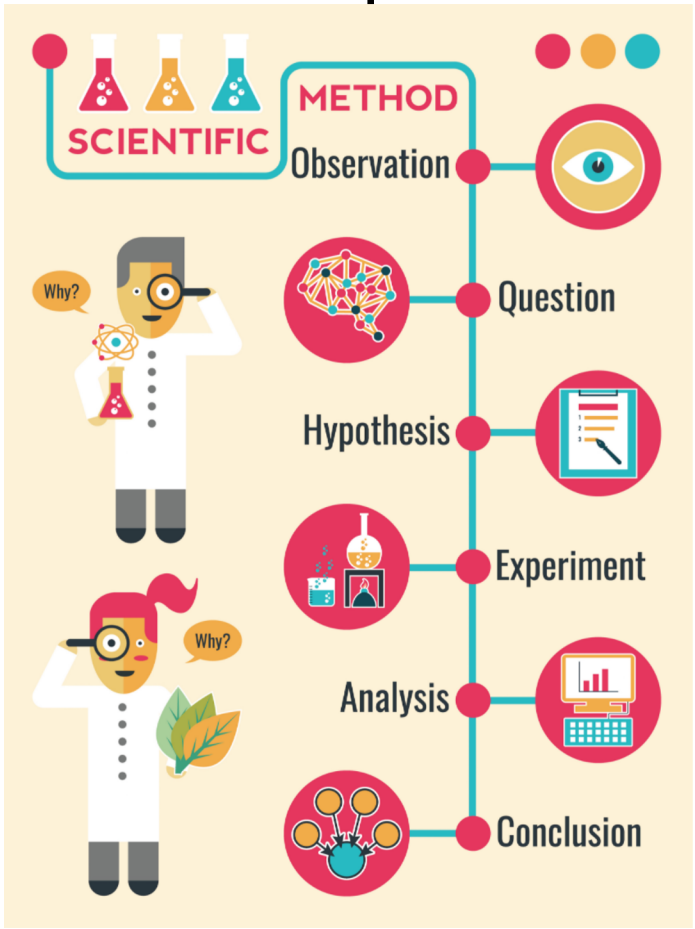
Why do we call it a science if we don’t publish scientific papers?

Creative process with lot of thinking, questioning, experimenting and debating.

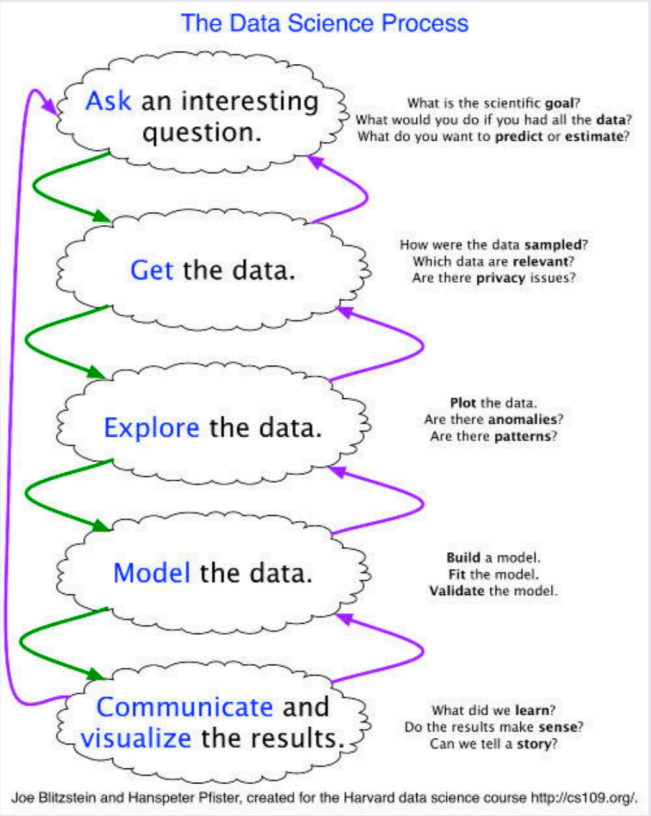
We do a study, which is based on data, and reach a conclusion.

As in the SCIENCE the process should be repeatable, verifiable and should lead to the same conclusions!





What does the Data Science project look like?



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

https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html

What does the Data Science project look like?

- Data science Practical definition by Mason & Wiggins (2010) in five steps:

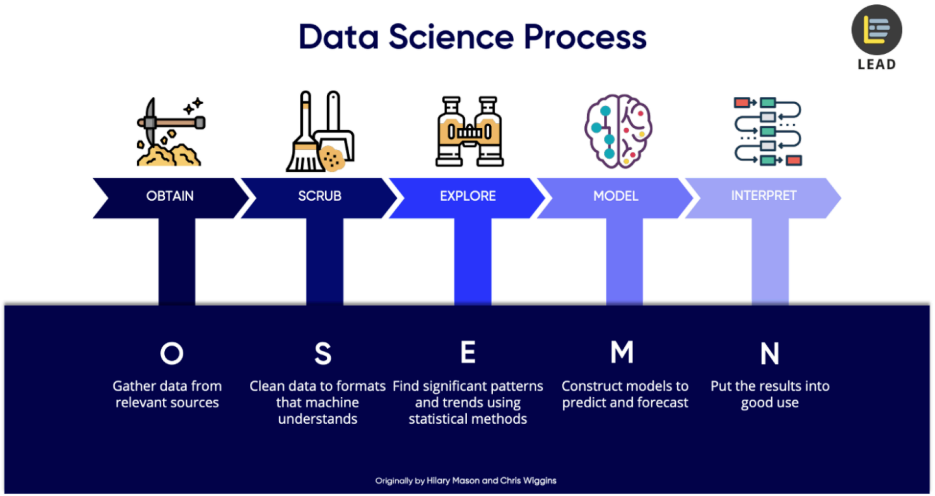
(1) **o**btaining data,

(2) **s**crubbing data,

(3) **e**xploring data,

(4) **m**odeling data

(5) i**n**terpreting data.

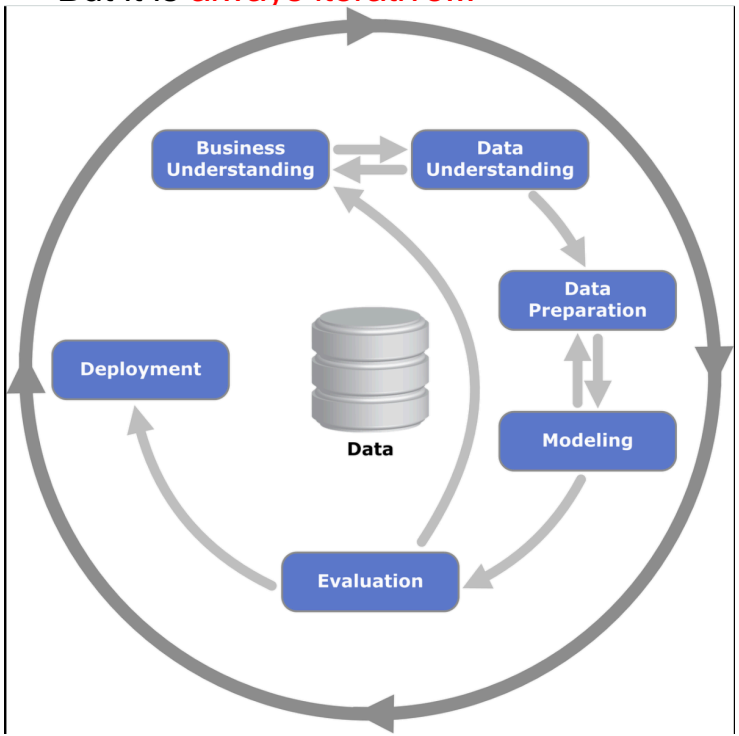


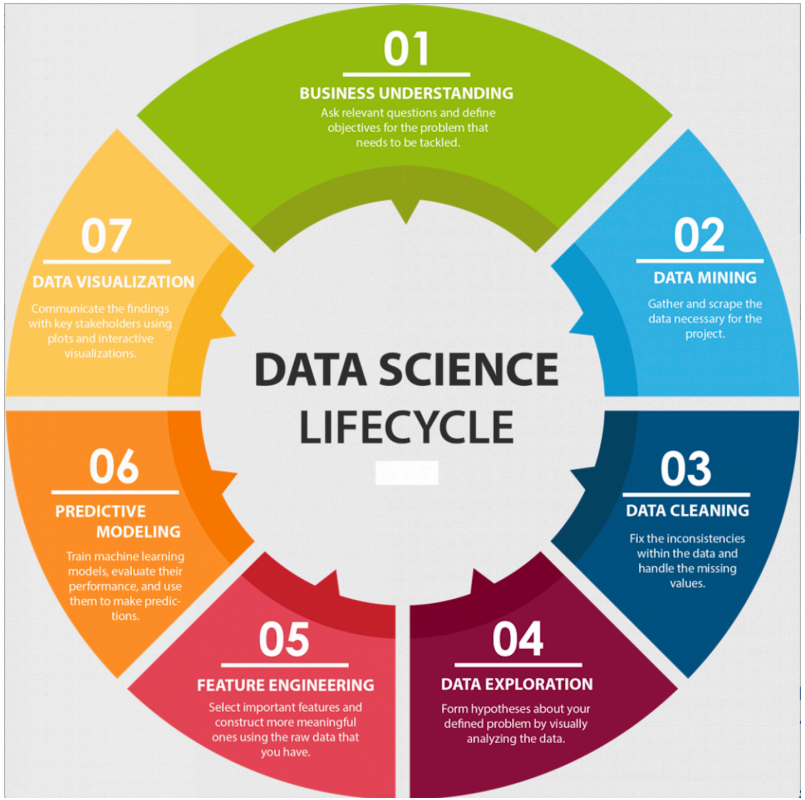
- Together, these steps form the OSEMN model (which is pronounced as **awesome**).

https://towardsdatascience.com/5-steps-of-a-data-science-project-lifecycle-26c50372b492

It can be presented in many diff ways

But it is always iterative...





https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

How do we spend our time?

80% on preparing the data!

And the rest?

Do ALL DS spend 80% on preparing the data????

- Actually, no…
  this is a typical time split of a common DS project

- **There are many different DS job definitions/descriptions**
- **And many more flavors of data professionals**
  - Data Scientist
  - Data Science Research
  - Data Toolmaking / Platform Engineering
  - Data architect
  - Data Analyst
  - ML/AI engineer
  - Data engineer
  - Statistician
  - Decision scientist
  - Big Data DevOps
  - Big Data Engineer
  - ML/AI Specialist

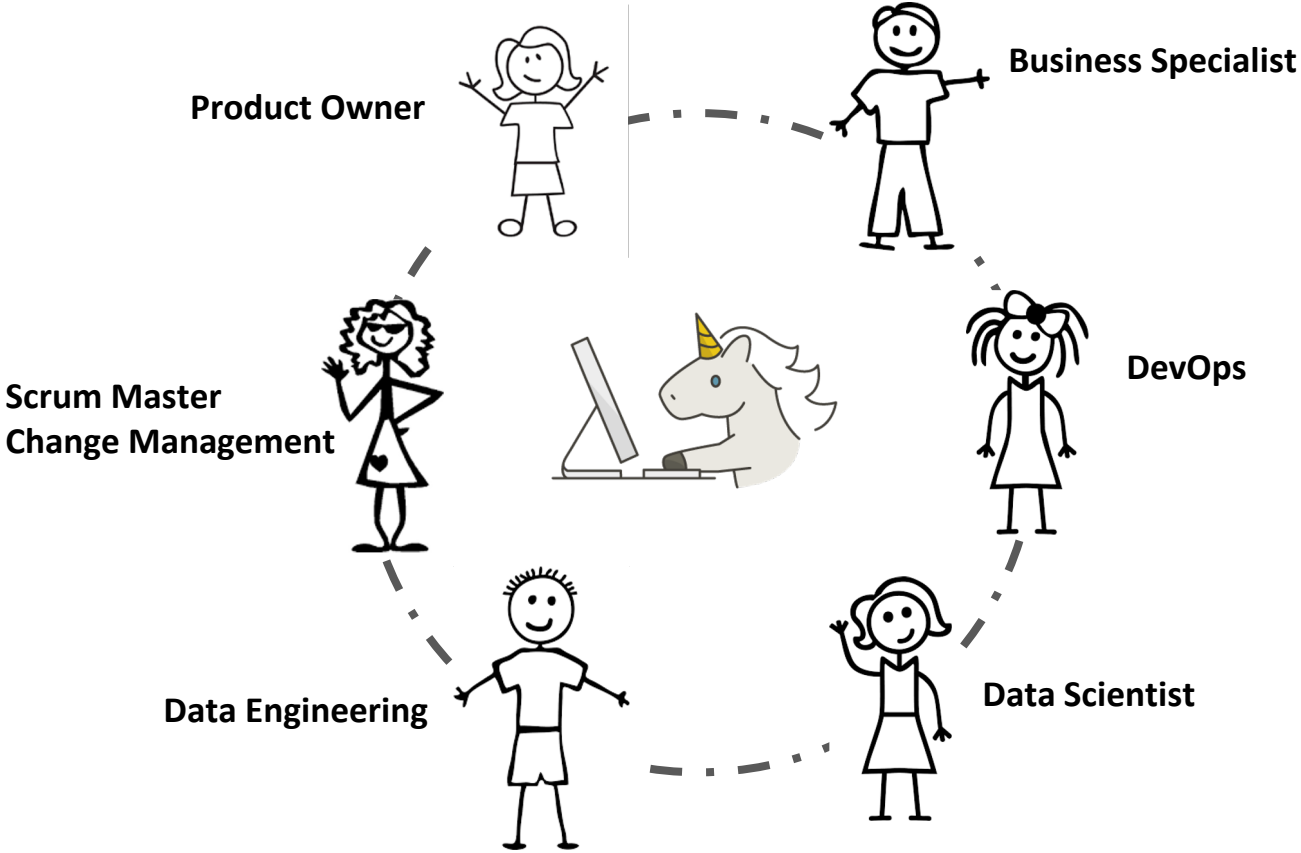
https://towardsdatascience.com/which-flavor-of-data-professional-are-you-5e01375584ce

In KSchool we educate the DS profile as hybrid of DE and DS roles

- **If you want to have an agile team your people should NOT depend on others for simple things**
- **If you want to do a DS project from end to end you need to know data engineering also!**

- Our background motivation:
  - I don't do bash!
  - I don't do git!
  - I don't know what this column means. I receive it like this and I use it… talk to someone else…
  - Why should I change the date format? DE should do it!
  - I only use sci-kit learn. This is enough
  - R is much better than Python. You don't need anything else…

How do we fit in business organization?



There is no such thing as unicorns!

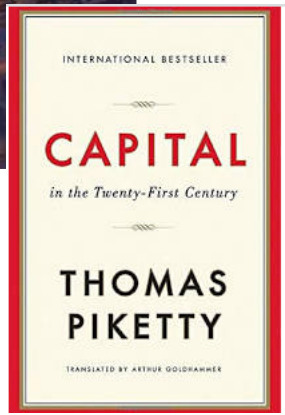
The three facets of a data scientist

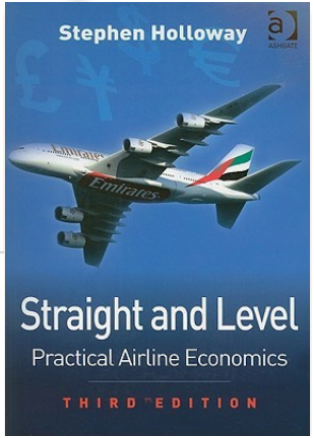
- Functional (domain knowledge)
- Analytical (how to extract knowledge from data)
- Technical (how to implement the data science process)

Question

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

There is a reason why we don’t teach functional knowledge at Master !!!





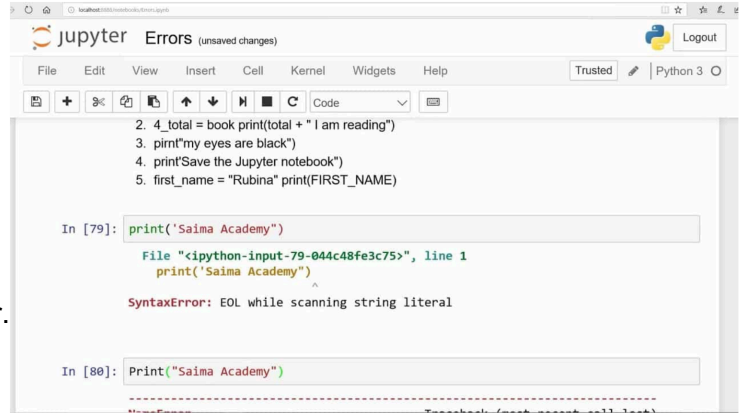
Program of the Master in Data Science

- 01. Introducción
  - ¿Qué es el Data Science? ¿Por qué es importante? ¿Quiénes son los profesionales que se dedican a ello?
  - Preparación de entorno de trabajo: Linux y GIT (en máquina virtual).
  - Manipulando las herramientas imprescindibles de linea de comandos de Linux (sort, unique, cut, paste, grep, sed…).

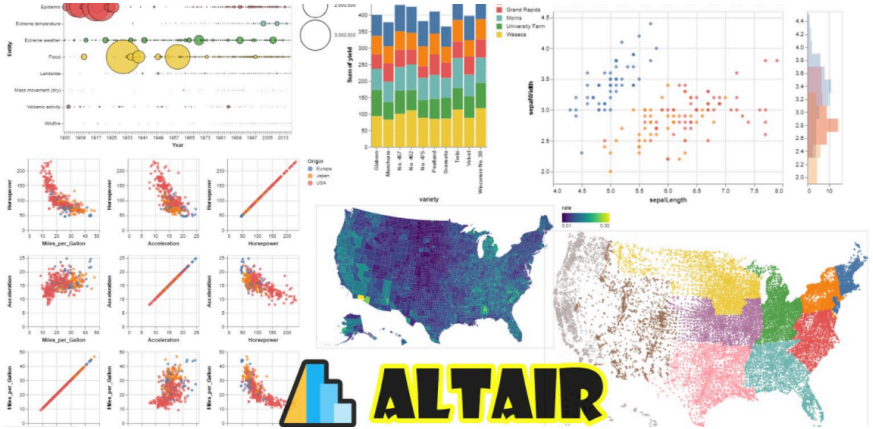
- Why do we need the Virtual Machine (VM)?.
  - VM is a machine inside a machine. As such it works slower than native system, but it assures **we ALL have the same working environment!**
  - Linux environment (and Command line tools) is a must for a serious DS/DE
  - Some classes will not be based on VM (Tableau only works on Mac & Win). You will know this before the class!
  - VM includes tools, data set and books used throughout the master
    - Python 3
    - Jupyter Notebook
    - R studio
    - Spark
  - The VM is not an absolute must but **we do not give support for any other machine** if your tools are not working properly.

Program of the Master in Data Science

- 02. Data Hacking con Python
  - Introducción a Python (Jupiter Notebook).
  - Álgebra y métodos numéricos en Python (Numpy).
  - Estadística con Python.
  - Data Science con Python (Pandas, Dataframes, APIs y web scrapping).
  - Visualización de datos con Python: Matplotlib, Seaborn, Vegalite y Altair.
  - Caso práctico: Data Science Challenge.



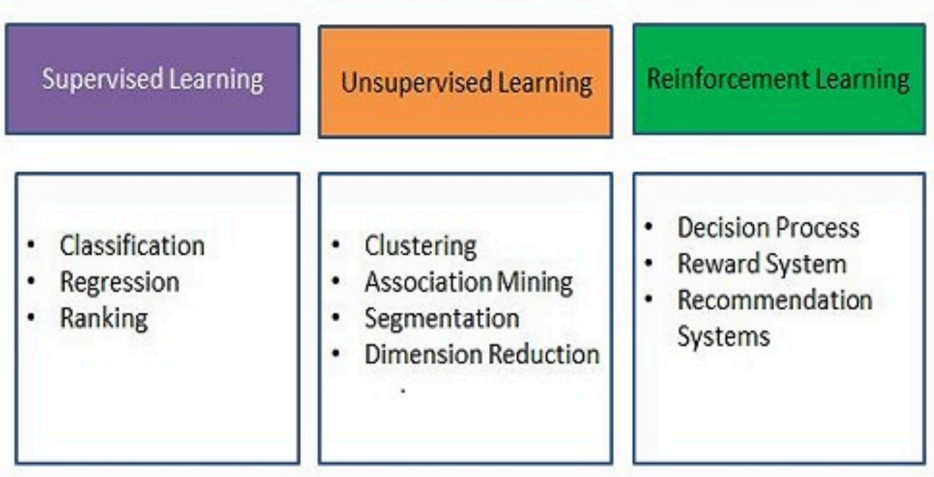
	Series			Series			DataFrame	
	apples			oranges			apples	oranges
0	3	+	0	0	=	0	3	0
1	2		1	3		1	2	3
2	0		2	7		2	0	7
3	1		3	2		3	1	2



Program of the Master in Data Science

- 03. Machine Learning con Python
  - Feature engineering con Python.
  - Aprendizaje supervisado con Python. Máquinas de vector soporte (SVM). Gradient boosting machines. Árboles y bosques. K-vecinos.
  - Aprendizaje no supervisado con Python.
  - Introducción a sistemas de recomendación.

What kind of questions can we answer to?



https://www.kdnuggets.com/2015/09/questions-data-science-can-answer.html

Program of the Master in Data Science

- 03. Machine Learning con Python
  - Feature engineering con Python.
  - Aprendizaje supervisado con Python. Máquinas de vector soporte (SVM). Gradient boosting machines. Árboles y bosques. K-vecinos.
  - Aprendizaje no supervisado con Python.
  - Introducción a sistemas de recomendación.

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

Program of the Master in Data Science

- 04. Diseño de Producto de Datos

Tutoria 1

- 05. Deep Learning
  - Introducción a Deep Learning: Python, Keras y Tensorflow.
  - Construcción de modelos predictivos basados en redes neuronales.

- 06. Procesamiento del Lenguaje Natural (NLP)
  - Procesamiento del Lenguaje Natural.
  - Acceso a recursos lingüísticos y colecciones de datos.
  - Librerías de PLN en Python: NLTK, TextBlob y spaCy.

- 07. Machine Learning con Google Cloud Platform
  - Google Cloud Solutions for Big Data & Machine Learning.
  - Análisis del marketing digital, Geo Analytics y Machine Learning con BigQuery (SQL).

Program of the Master in Data Science

- 07. Machine Learning con Google Cloud Platform
  - Google Cloud Solutions for Big Data & Machine Learning.
  - Análisis del marketing digital, Geo Analytics y Machine Learning con BigQuery (SQL).

- 08. Data Hacking con R
  - Introducción a R.
  - Manejo y limpieza de datos con R.
  - Visualización de datos con R.
  - Aprendizaje supervisado con R.
  - Aprendizaje no supervisado con R.
  - Series temporales con R.

Tutoria 2

Program of the Master in Data Science

- 09. Big Data
  - Big Data con Spark.
  - Herramientas para el trabajo remoto: ssh, autenticación.
  - Spark sobre Amazon Web Services.
  - Spark sobre Google Cloud Platform.
- 10. Visualización
  - Introducción a Tableau y visualización.
  - Visualizaciones en Tableau.
- 11. Resumen práctico de todo el máster
  - Claves del éxito de un proyecto de Data Science.
  - Kaggle Hackathon.
- TFM

Trabajo Fin de Master (TFM)

It is normal that this question is refined after iterations with the data

Goals of the TFM

- Students will show their capacity to work as data scientists
- Starting from raw data, all the way up to solving a research/business question
- TFM will include the following phases:
  - Data acquisition
  - Data cleansing and preparation ○ Analysis
  - Frontend / visualization

Evaluation criteria

- Clarity in the document, easiness to replicate work
- Complexity of the data and the analysis methodologies
- Clarity and correctness of source code
- Relevance and fit-to-purpose of the chosen analytical methods and technologies
- UX and usability of the frontend

We will evaluate as the “consumer” of the report, with zero knowledge of data science!

- (0-10 points, 0.1 points less per hour of delay)

The delivery is done through github!

You can do it in pairs!

What job position can you apply for after the master?

Harvard Business Review

Data | Data Scientist: The Sexiest Job of the 21st Century

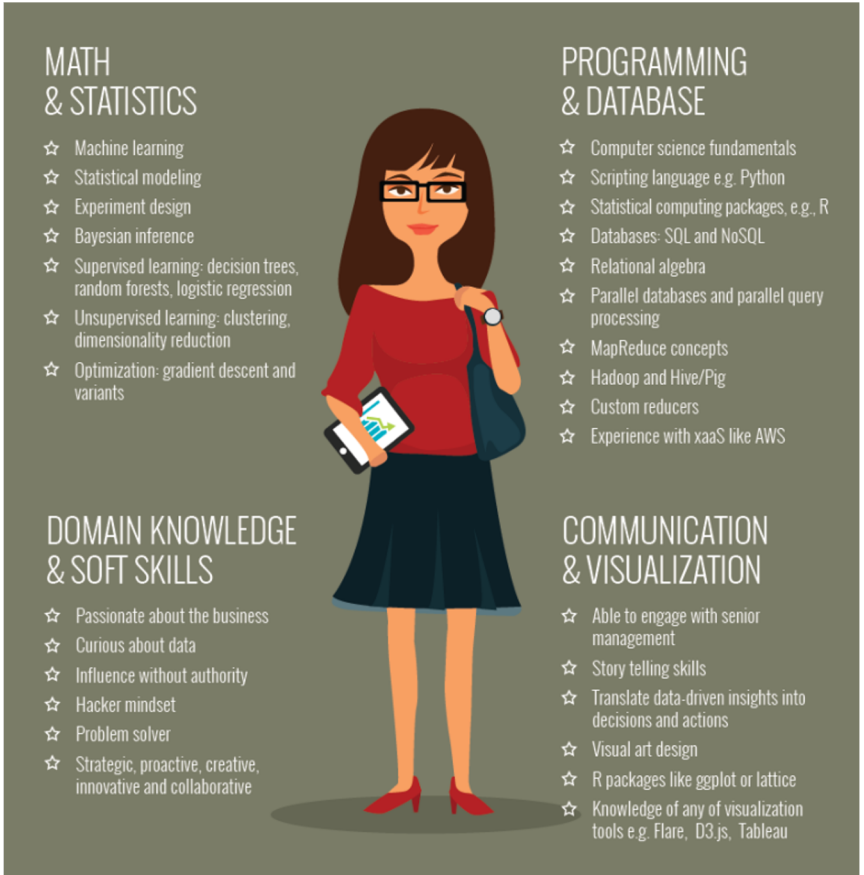
DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

From the October 2012 Issue

Recruiters look for unicorns



What job position can you apply for after the master?

Skills demanded in Data Engineering or Data Scientist job postings

Analytical skills

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

Technical skills

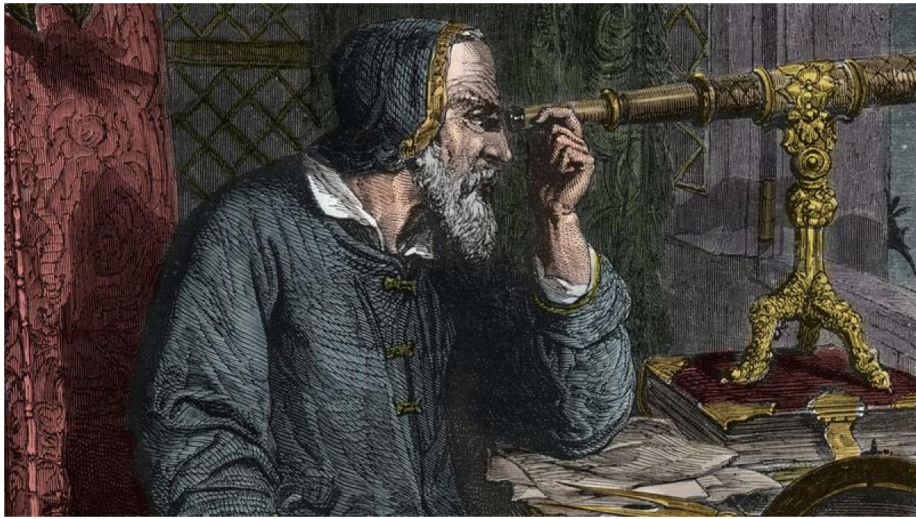
- R
- Python
- SQL (BigQuery)
- Hadoop, Spark and Big Data (covered only partially)
- Visualization technologies (Tableau, Qlik)
- Excel

Who can be a Data Scientist?

No, **you don’t need a PhD!!!!**

… but Scientific Mindset is a plus ☺

- Don’t believe anything
- Measure everything
- Question everything



Additional key skills

- Fitting in an **organization**, leading projects in a heterogeneous environment, aligning with strategy
- Data-Analytic **Thinking**
- Communication**, communication,** communication

Why would someone choose you if you have no DE/DS experience????

You need visibility!!! → Share promiscuously

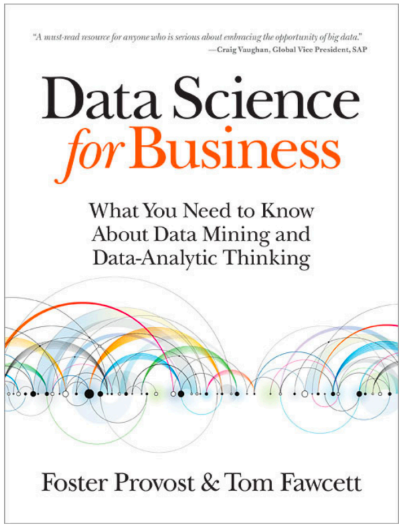
Use the social networks to position yourself as a data scientist

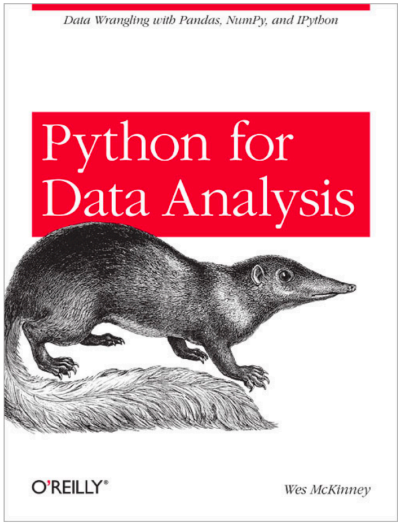
- Crucial: Profile in LinkedIn and share, share, share!
- **Share frequently about your progress in the master**
- Share your code in Github or Bitbucket
- Share and update frequently, so it always shows recent activity
- **Your TFM and your github account are your DS portfolio!!!**

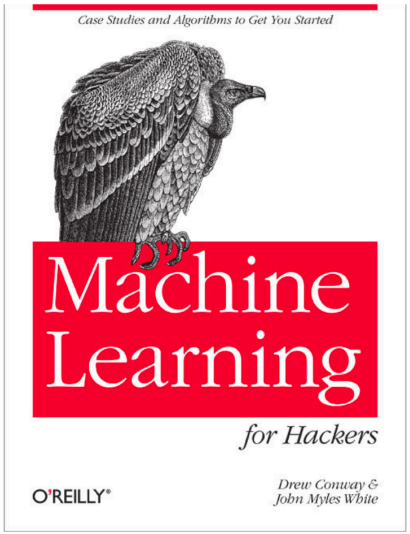
- Don’t mix (too much) your personal postings with your data scientist postings

Recommendations to follow the master

- Recommended books







- **For every hour of class with professor you need to dedicate at least 1 hour on your own**

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.

- 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.

- Practice, practice, practice…
- Share, share, share

- There are no unicorns!!!!

Program for this week

- Intro to data science
  - We have just completed this part.
- The environment
  - Setting up the environment with VirtualBox and Ubuntu
  - **The only class with 2 professors**
- Starting with Git
  - Git and the importance of sharing our source code
  - Working with git is spread over first 3 days
- Getting familiar with the command line
  - Many times the shell command line is enough to answer to a lot of questions

Any Questions or we start?