Modulo 5

Data Analytics

Lesson 4: Analytics over Big Data sources
Data Science fundamentals

Alejandro Maté
Juan Carlos Trujillo

Table of contents

INTRODUCTION TO DATA SCIENCE

- What is Data Science?
  - Using together **algorithms**, **data**, and **software** to infer new knowledge (data insights) from heterogeneous data
  - Applying Data Science requires **domain knowledge**
    - Data insights are **only as relevant** as how much they **impact the business**
  - Applying Data Science requires knowledge of **data manipulation and interpretation**
    - You **do not** need to be a Computer Scientist
    - You need to know **how to use the tools**
  - Applying Data Science requires **communicating results**
    - Visualizations can be key

INTRODUCTION TO DATA SCIENCE

- What is a Data Scientist? What can he or she does? What questions can he or she **answer** us?
  - "How do I build a pipeline that can handle 10 000 data requests per minute?"
  - "What's driving our user growth numbers?"
  - "How many different user types do we really have?"

INTRODUCTION TO DATA SCIENCE

- What is a Data Scientist? What can he or she does? What questions can he or she **answer** us?
  - "How do I build a pipeline that can handle 10 000 data requests per minute?"
  - "What's driving our user growth numbers?"
  - "How many different user types do we really have?"

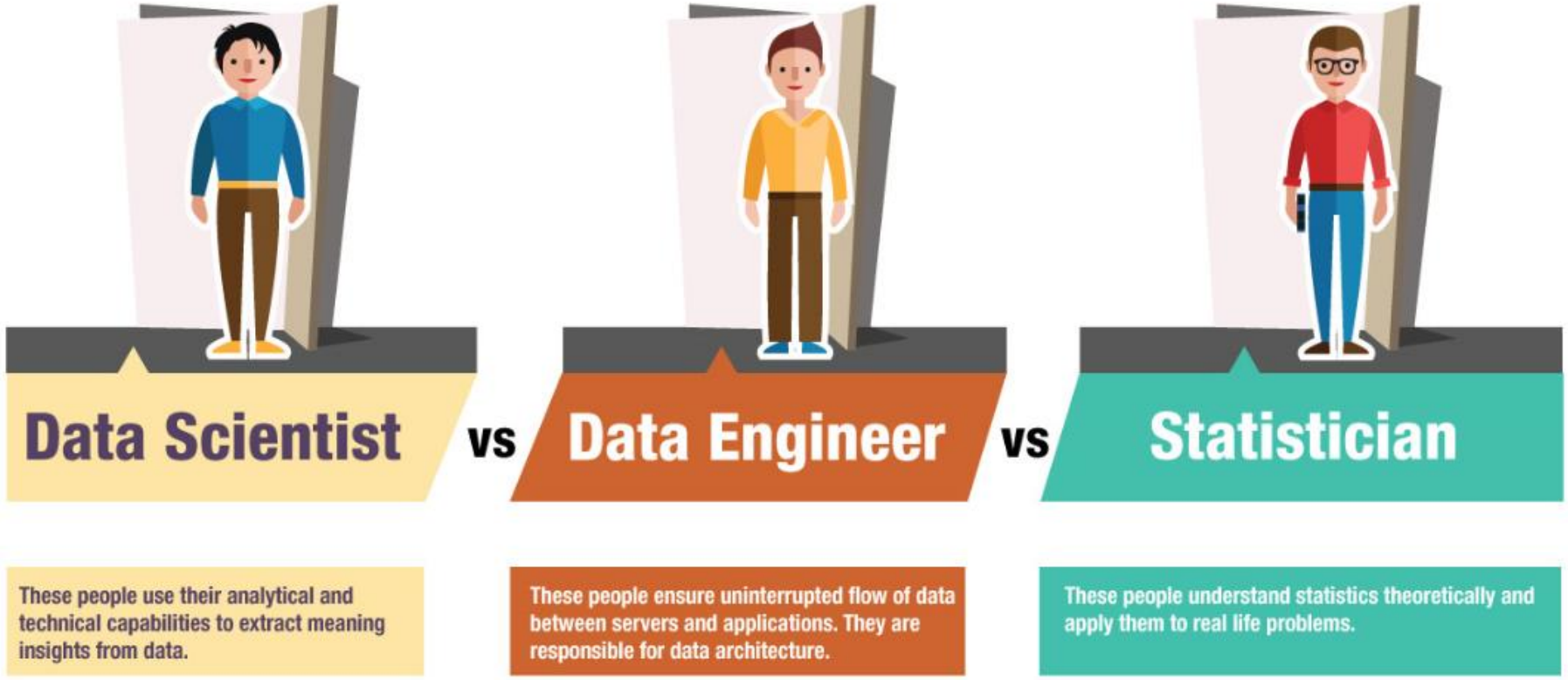
  - **ANSWER:** ~~You can be a Jack-of-all-trades~~

INTRODUCTION TO DATA SCIENCE

- NOWADAYS:
  - Data science has been around for some time now, and the Data Scientist has split in several jobs
  - Most companies **know who** should work on their tasks but **rarely know what they need** (i.e. we want ***value*** out of our data)
  - In SMEs, especially Spain, the Data Scientist needs to do ***everything***

- FUTURE YEARS:
  - **New skills** will be constantly required (e.g. new GenAI tools)
  - New generation of **technologies** in multiple fields (data engineering, AI, visualizations)
  - It is **recommended** to focus on one career

PROFILES


Data Scientist
vs
Data Engineer
vs
Statistician
These people use their analytical and technical capabilities to extract meaning insights from data.
These people ensure uninterrupted flow of data between servers and applications. They are responsible for data architecture.
These people understand statistics theoretically and apply them to real life problems.

SAMPLE JOB OFFERS

Técnico/a CRM Data Scientist

[redacted]

Las **funciones principales** a desempeñar por la persona seleccionada serán entre otras:

- Extraer y depurar datos de diferentes fuentes para poder procesarlos y crear atributos adaptados a cada modelo e implementarlos para su análisis posterior.
- Una vez se cuenta con los datos preparados, procesarlos aplicando estadística, softwares analíticos y modelos predictivos y representarlos de forma que sean comprensibles.
- Diseñar, desarrollar e implantar Custom Activity y Custom Split para integrarlos con Journey Builder.
- Desarrollar aplicaciones en Heroku para su integración en MKT Cloud.
- Desarrollar nuevas vías de comunicaciones entre Salesforce y tecnología propia.

Los **requisitos** principales son:

- Experiencia en Salesforce Marketing Cloud
- Dominio avanzado de Excel (funciones, tablas dinámicas, análisis de datos), power point y herramientas de análisis de datos (SQL, Python).
- Conocimiento de software de visualización (Power BI, SAP). Lenguaje de Programación DAX
- Conocimiento en Custom Activity
- Nivel de inglés avanzado.

SAMPLE JOB OFFERS

Técnico/a CRM Data Scientist

[redacted]

Las fu

- E
  a
- U
  a
- D
  B
- D
- D

Los re

- E
- D
  h
- Conocimiento de software de visualización (Power BI, SAP). Lenguaje de Programación DAX
- Conocimiento en Custom Activity
- Nivel de inglés avanzado.

The impact you'll make

- Drive strategic business decisions with insights obtained by manipulating large volumes of data from a variety of different sources: system logs, satellite imagery, vertical imagery, street level imagery, GPS traces, etc. individually and by fusing several combinations of these sources
- Design, develop and maintain data processing pipelines to use these data sources individually or combined and to predict various map features and attributes by applying statistical models/ ML / AI models
- State the quality level of our maps via statistical experiments that help us understand where we are and why, and deliver effective insights that help drive our map editing strategy
- Develop processes and tools to monitor and analyze model and system performance and data accuracy
- Collaborate with teams from diverse backgrounds (e.g., data scientists, software engineers, product managers, map and GIS experts …) across the company and with our customers on projects and knowledge sharing

PROFILES

• RESPONSIBILITIES (1/3)

DATA SCIENTIST:

- Develop and plan required analytic projects in response to **business needs**
- Contribute to data mining architectures, modeling standards, reporting, and **data analysis methodologies**
- **Collaborate with stakeholders** to integrate data mining results with existing systems
- Communicate and **present results**

PROFILES

• RESPONSIBILITIES (2/3)

DATA ENGINEER:

- **Design, construct, install, test** and maintain highly scalable data management systems
- **Improve data** foundational procedures guidelines and standards
- **Integrate new** data management **technologies** and software engineering tools into existing structures
- Create **custom software components** (e.g. specialized UDF's) and analytic applications

PROFILES

• RESPONSIBILITIES (3/3)

STATISTICIAN:

- Apply **statistical theories** and methods to solve practical problems of various industries
- Determine **methods** for finding or collecting data
- **Design surveys** or experiments or opinion polls to collect data
- Analyze, interpret & undertake **data analysis**
- **Report conclusions** from their analyses

PROFILES

- SKILLS

DATA SCIENTIST

- Programming
- Mathematics
- Business Understanding
- Statistics
- Data Visualization
- Machine Learning
- Attention to detail
- **Domain knowledge**

DATA ENGINEER

- Database Design
- Production coding
- Data collection
- Data Warehousing
- Data transformation
- Work diligently with data

STATISTICIAN

- Technical and Analytics Skills
- Mathematics
- Operational Research Writing Skills
- Ability to analyze
- Model and interpret data
- Flair of explaining difficult concepts in simple manner

DATA SCIENTIST

- TOOLS

PROFILES

- SALARY (AVG)

	DATA SCIENTIST	DATA ENGINEER	STATISTICIAN
SPAIN*	50.658 €	43.094 €	? €
EEUU*	$123,111	$125,313	$89,517

* INDEED 2025

PROFILES

- CHANGELLING TASKS YOU MUST BE ABLE TO COPE WITH

DATA SCIENTIST

- Design a Recommender System
- Dirty data cleaning
- Lack of clear questions
- Results not used for decision makers

DATA ENGINEER

- Data warehouse design
- Writing queries efficiently
- Dealing with data grow
- Generating insights in a timely manner
- Integrate disparate data sources

STATISTICIAN

- Risk Pricing Model
- Analyze trends
- Presenting information in a variety of formats
- Conveying complex information to people who may not be specialists

PROFILES

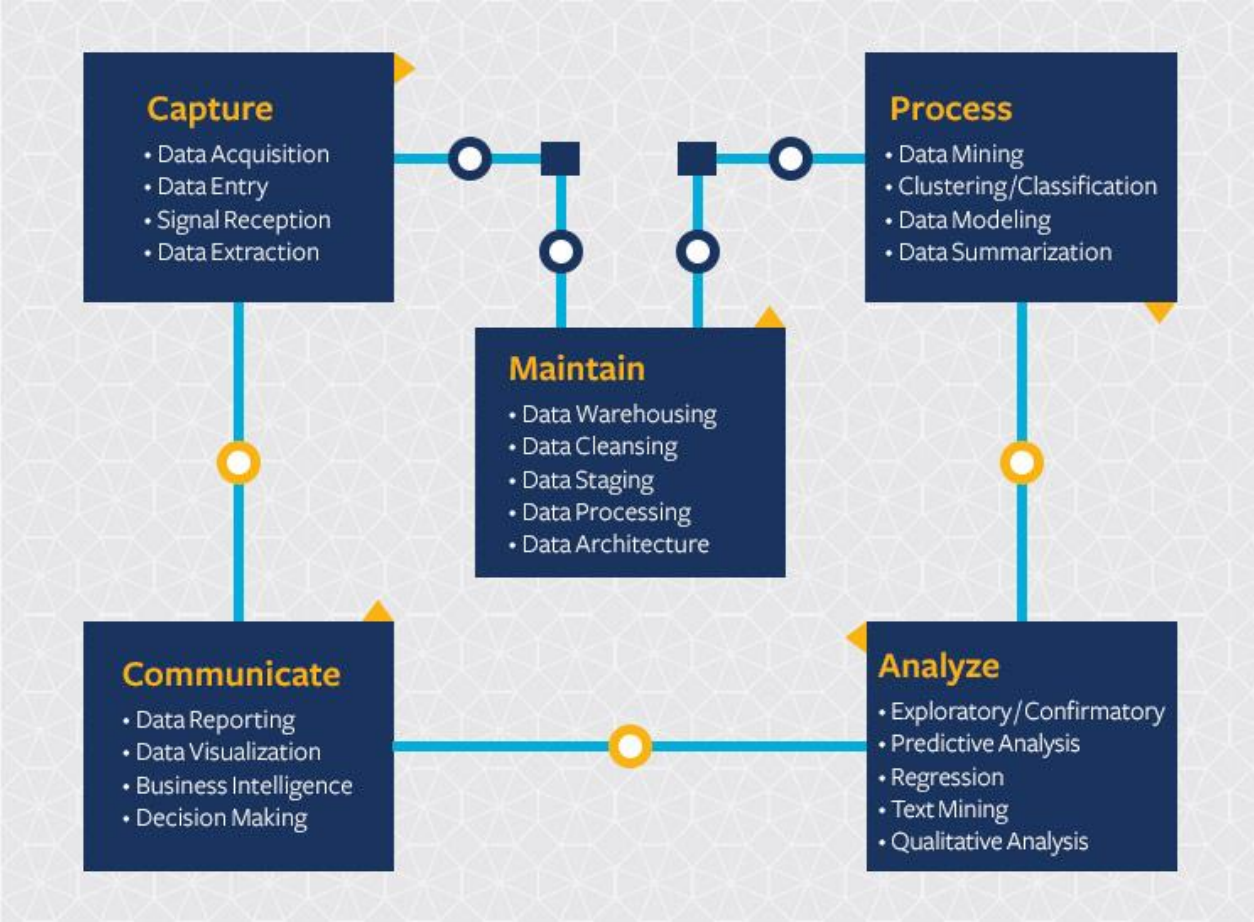
- WHERE CAN I FIND A JOB?

EVERYWHERE!!

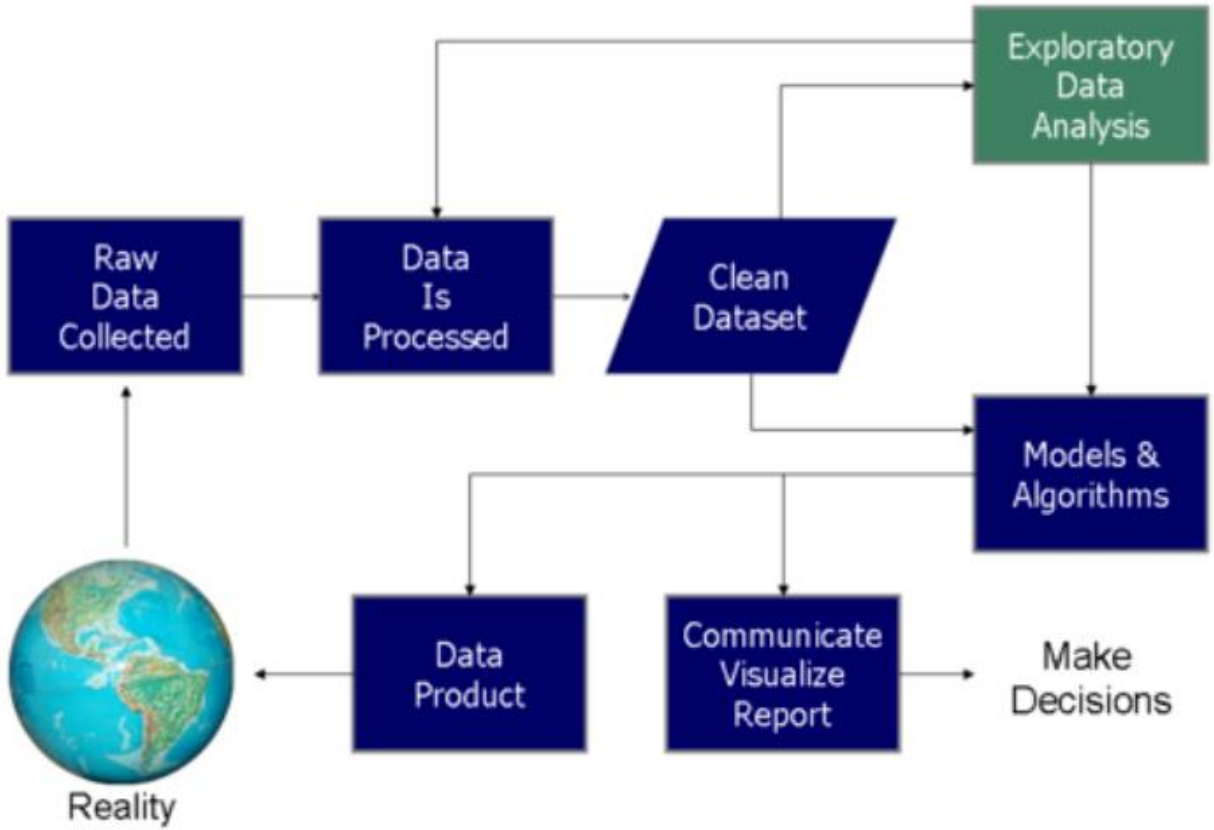
THE DATA SCIENCE PROCESS

- The Data Science Lifecycle

Source: https://datascience.berkeley.edu/about/what-is-data-science/



THE DATA SCIENCE PROCESS



THE DATA SCIENCE PROCESS

1. FRAME THE PROBLEM (1/2)

DATA SCIENTIST

STATISTICIAN

- We must know what we need to answer
- Our objective is to transform data questions in actionable insights
- Sometimes the customer does not know the question, or she can not explain it properly
- We must help the customer refine the initial objective into **concrete questions**

THE DATA SCIENCE PROCESS

1. FRAME THE PROBLEM (2/2)

- E.g. Management wants to improve sales
  - *So far* we cannot kidnap customers and force them to buy
- We could **refine** the request into several questions:
  - Do we wish to identify the characteristics of a customer that will buy our product?
  - Do we wish to analyze sales opportunities across the year?
  - Are we trying to locate shops that could increase their sales?

DATA SCIENTIST

STATISTICIAN

- With those questions, Management might reveal that their true **objective** is to understand why certain segments of customers have bought less than expected (more specific question)

THE DATA SCIENCE PROCESS

2. COLLECT RAW DATA FOR YOUR PROBLEM

- What data do we need?
- How can we get it?
- Which queries can give us that data?
- Do we need to buy external datasets?

DATA ENGINEER

THE DATA SCIENCE PROCESS

3. PROCESS THE DATA FOR ANALYSIS

- We require basic understanding and to check errors prior to the analysis – Data profiling
  - Duplicated values
  - Empty entries
  - Corrupted values
  - Time zone differences
  - Data range errors
  - Does the aggregated data make sense?

DATA ENGINEER

DATA SCIENTIST

THE DATA SCIENCE PROCESS

3. PROCESS THE DATA FOR ANALYSIS

DATA ENGINEER

DATA SCIENTIST

- Using jupyter:
  - Describe() for automatic calculus over numerical columns
  - Describe(include='all') for automatic calculus over all columns
  - Number of rows – len(DF.index)
  - Number of blanks/nulls on a column - .isna().sum()
  - Number of distinct values on a column - .nunique()
  - Maximum and minimum (non-null) values - .max()/.min()

THE DATA SCIENCE PROCESS

4. EDA (EXPLORATORY DATA ANALYSIS)

DATA SCIENTIST

- Once our data is ready
- Provides deeper understanding of the data
- Prioritize our questions (we will have a deadline)

LETS PLAY WITH THE DATA!!

THE DATA SCIENCE PROCESS

5. PERFORM IN-DEPTH ANALYSIS

- Crunch the data and find every insight that we can
- Often aided by Machine Learning and Statistical models

- Our tools:
  - Mathematics
  - Statistical
  - Technological knowledge
  - Domain knowledge

DATA SCIENTIST

STATISTICIAN

THE DATA SCIENCE PROCESS

6. COMMUNICATE RESULTS OF THE ANALYSIS

- Use suitable tools and schemas for visualization
- Propose actions for solve the problem
- Explain effects of inaction to those problems

STATISTICIAN

DATA SCIENCE – Classroom exercise

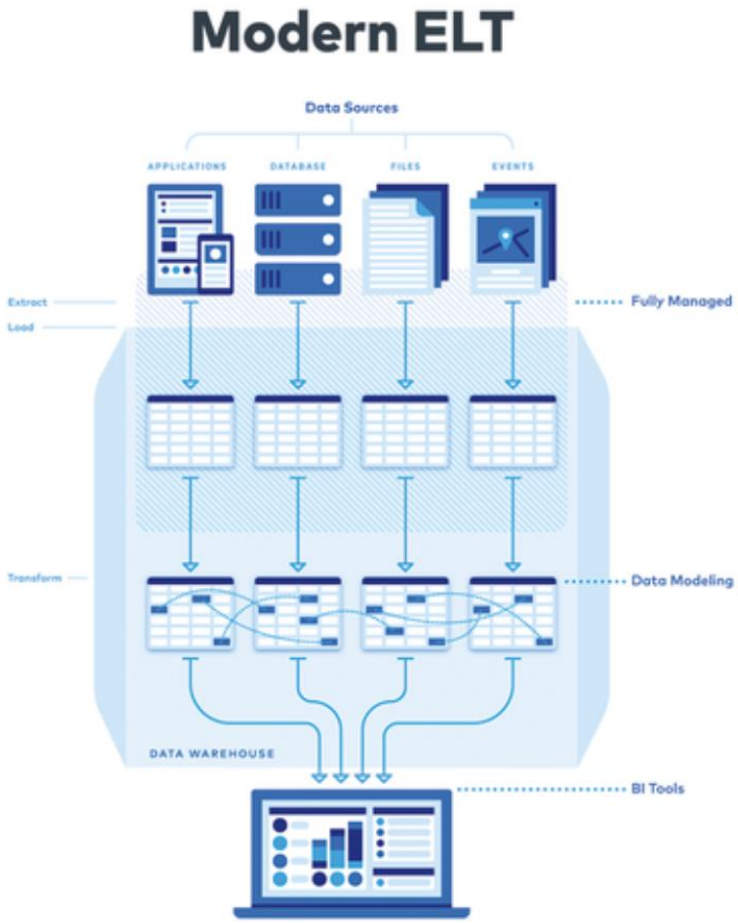
- Connect to San Francisco Open Data portal
  - The link is on the main tab on the virtual campus
  - Click on “view data”
  - Answer the following:
    - What kind of domain knowledge would you need to interpret the data?
    - What can this data be useful for?

DATA SCIENCE – Classroom exercise

- Connect to San Francisco Open Data portal
  - The link is on the main tab on the virtual campus
  - Click on “view data”
  - Answer the following:
    - If we wished to build a data warehouse with this data, which profile would be most adequate?
    - What if we wished to try to predict what kind of unit should be dispatched before grabbing the call?
    - And if we wished to know the likelihood of a call being a fake call?

THE DATA SCIENCE PROCESS

- Compared to traditional processes:
  - Data is **transformed continuously** during the analysis
  - **Data modeling and integration** is performed "**on the fly**" as more information is needed
  - **Transformed data may be discarded**, raw data remains in the data lake
  - **Results** are most often **reports or AI models**



Big Data sources

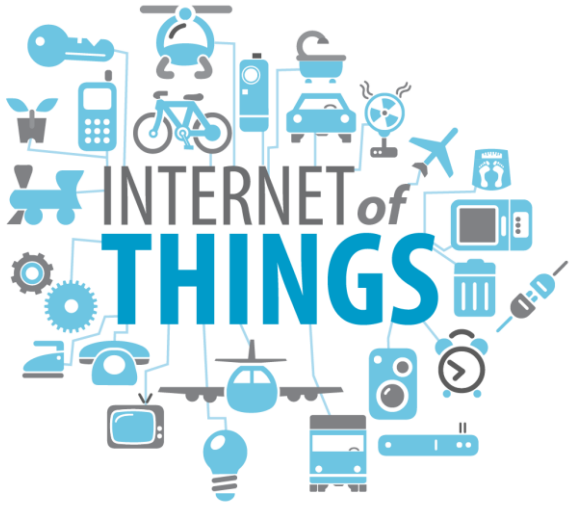
- What is Big Data?
  - Most often referred as data that is "**too big**" "**too heterogeneous**" and "**too fast**" to be processed with traditional algorithms and techniques
  - Big Data comes from a **variety of sources** and is interesting for organizations thanks to the **data insights** they can obtain
    - What do our customers like most?
    - What is the optimal path for our taxis at each point in time?
    - How do our citizens behave? Where do they go, when?

Types of Big Data sources

- We can consider Big Data as data coming from several types of data sources and in different formats:
  - Social networks
    - Register interactions between users
    - Heterogeneous format and media
    - The kind of interaction depends on the network
    - The information is often represented as a graph

Types of Big Data sources

- We can consider Big Data as data coming from several types of data sources and in different formats:
  - Sensors and general IoT
    - Small pieces of constant information from devices
    - Information must be processed or stored in real time
      - Otherwise it is lost
    - Typically some sort of Json format:
      - [{temperature: 23, humidity: 44, air: 2}]



Types of Big Data sources

- We can consider Big Data as data coming from several types of data sources and in different formats:
  - Clickstream and Logs:
    - Interactions between users and webpages
    - Traditional datasource that has evolved due to the increase in the number of interactions on the web

Types of Big Data sources

- We can consider Big Data as data coming from several types of data sources and in different formats:
  - Medical records:
    - Set of files in multiple formats related to patients
    - Diseases, health information, visits, habits
    - Difficult access due to privacy issues and sensitive information

Types of Big Data sources

- We can consider Big Data as data coming from several types of data sources and in different formats:
  - General unstructured data:
    - Text corresponding to news, product comments, etc.
    - Difficult to interpret due to the lack of metadata

Google News


The New York Times.

ARMISTICE SIGNED, END OF THE WAR!
BERLIN SEIZED BY REVOLUTIONISTS;
NEW CHANCELLOR BEGS FOR ORDER;
OUSTED KAISER FLEES TO HOLLAND

Types of Big Data sources

- We can consider Big Data as data coming from several types of data sources and in different formats:
  - Curated data:
    - Data prepared to be reused by third parties
    - Sold by companies or offered by public entities
    - Statistics agencies, town halls, and governments offer data for free
    - When is offered for free it is often called “Open Data”

OPEN DATA

OPEN DATA

•QUICK RECALL – OPEN DATA

- "Data that can be freely used, shared and built-on by anyone, anywhere, for any purpose"[1]

- "A piece of data that anyone is free to use, reuse, and redistribute it subject only, at most, to the requirement to attribute and/or share-alike"[2]

(1): Laura James, CEO in the Open Knowledge Foundation
(2): opendatahandbook.org

OPEN DATA

- ADVANTAGES

PERFORMANCES	Improving efficiency of public services
	Improving quality
ECONOMIC	Developing innovative services
	Creating new business models
SOCIAL	Enhance participation
	Improving transparency & accountability

OPEN DATA

- DISADVANTAGES

DARK SIDE OF OPEN DATA	Risk of violating legislation by opening data
	Difficulties with data ownership
	Privacy can be violated unintentionally
	Misinterpretation and misuse
DARK SIDE TO THE IMPLEMENTATION OPEN DATA	Little attention for public value and solving societal problems
	Unclear responsibility & accountability
	Not citizens but others profit from open data

OPEN DATA

- WIKIDATA
  - Collaboratively edited knowledge base.
  - Hosted by the Wikimedia Foundation

  - Document oriented DB, focused on items
  - It offers a query services with SPARQL
  - Public domain license

WIKIDATA

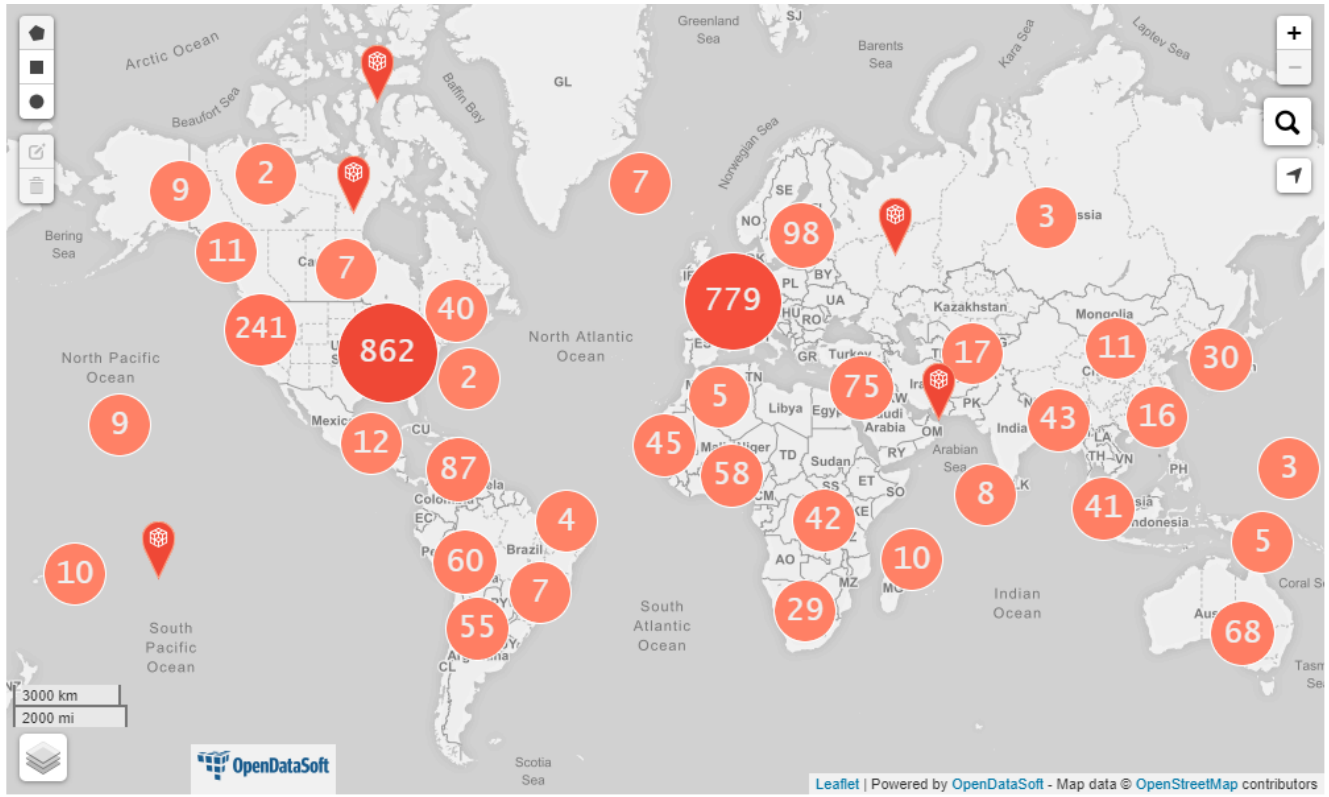
OPEN DATA

- LINKING OPEN DATA PROJECT
  - Aims at making data freely available for everyone
  - Under Creative Commons or Talis Licenses
  - Examples:
    - Wikipedia, Wikibooks, Geonames, MusicBrainz, Wordnet
  - The OD cloud

LINKINGOPENDATA
W3C SWEO Community Project

OPEN DATA

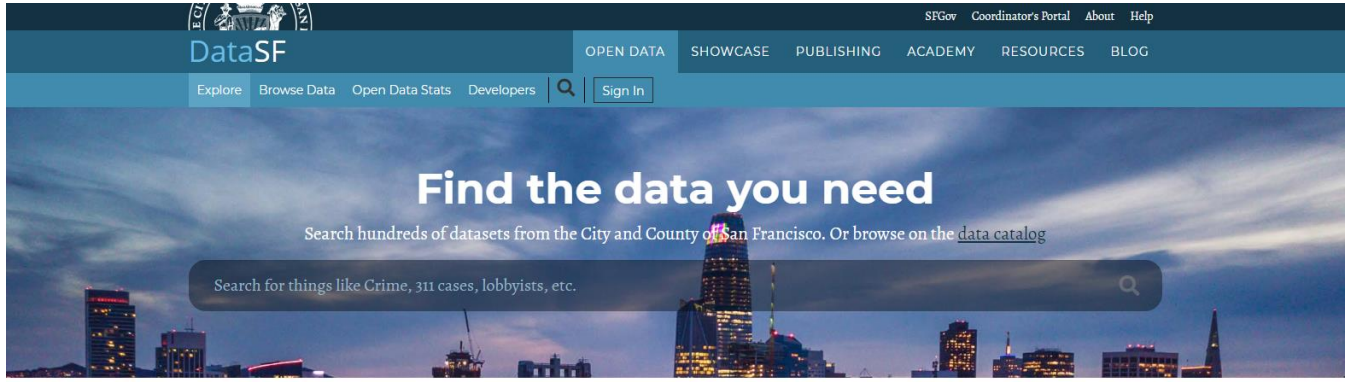
- IS OPEN DATA SPREAD OUT?



OPEN DATA

•FOREIGN OPEN DATA PORTALS (1/3)

- San Francisco Open Data



OPEN DATA

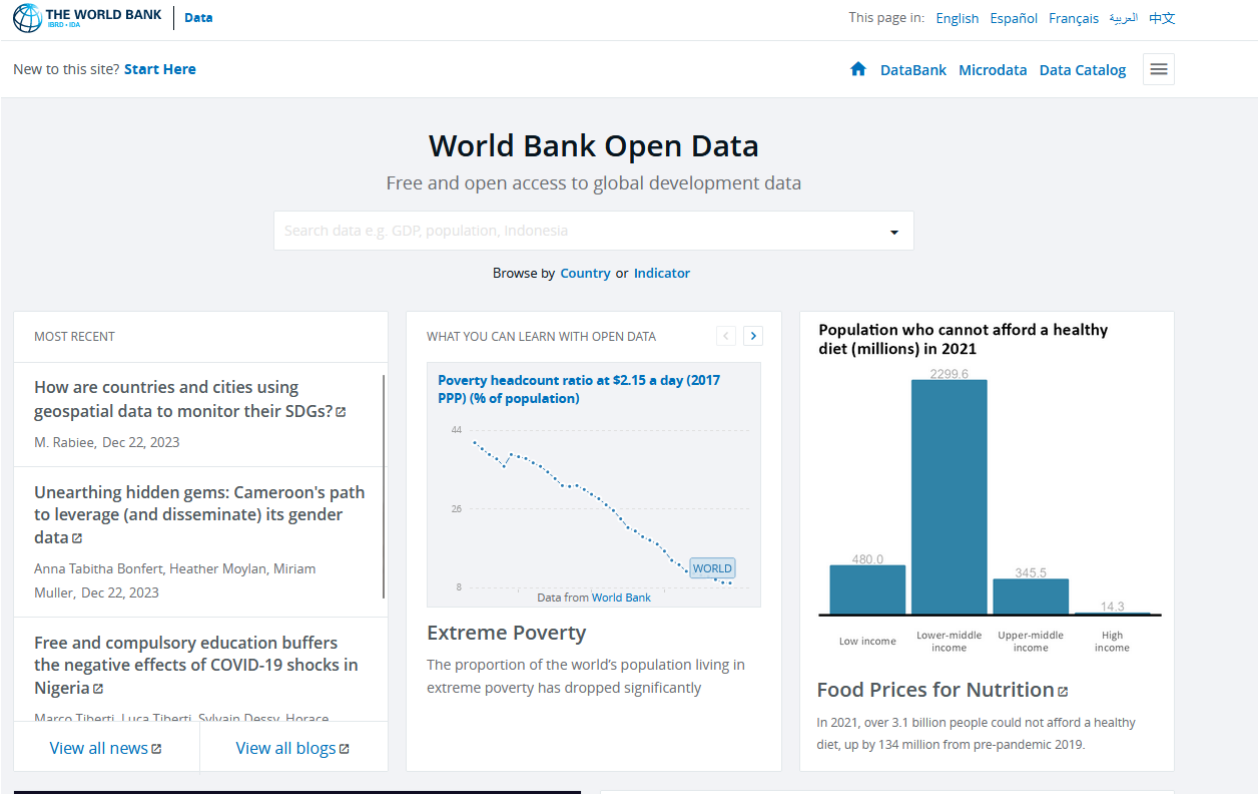
•FOREIGN OPEN DATA PORTALS (2/3)

- United States Open Data

OPEN DATA

•FOREIGN OPEN DATA PORTALS (3/3)

- World Bank Open Data



OPEN DATA

•SPANISH OPEN DATA PORTALS (1/2)

- Government of Spain Open Data

Impacto.

OPEN DATA

•SPANISH OPEN DATA PORTALS (2/2)

- Townhall of Malaga Open Data

OPEN DATA

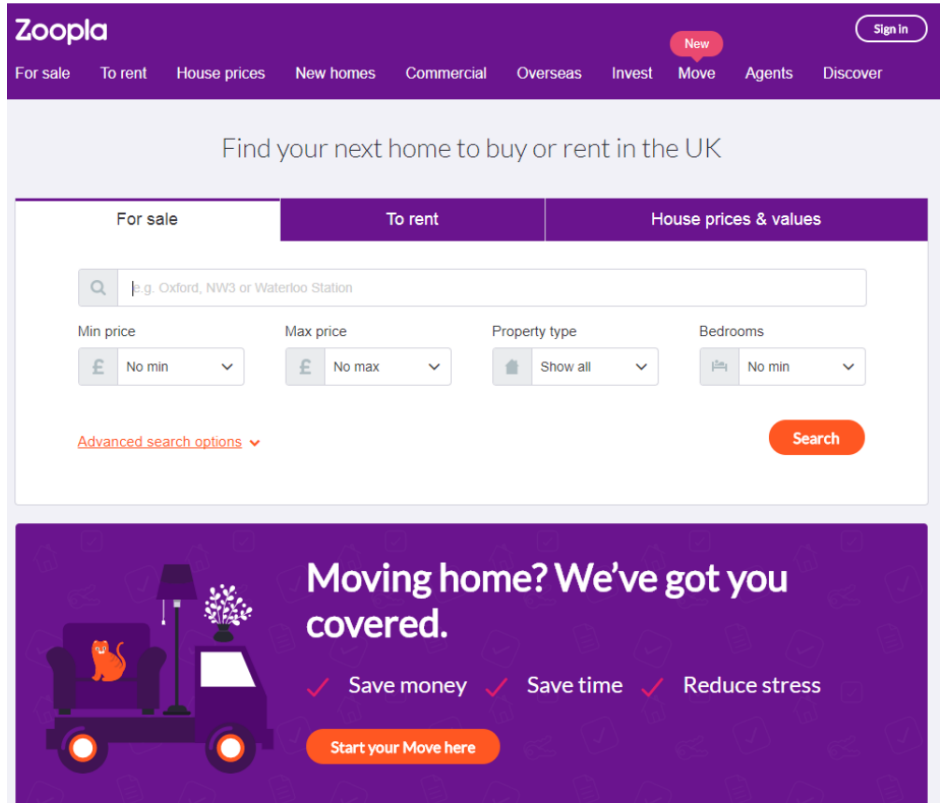
•DIFFERENT USES OF OPEN DATA (1/3)

- Plague Inc.

OPEN DATA

•DIFFERENT USES OF OPEN DATA (2/3)

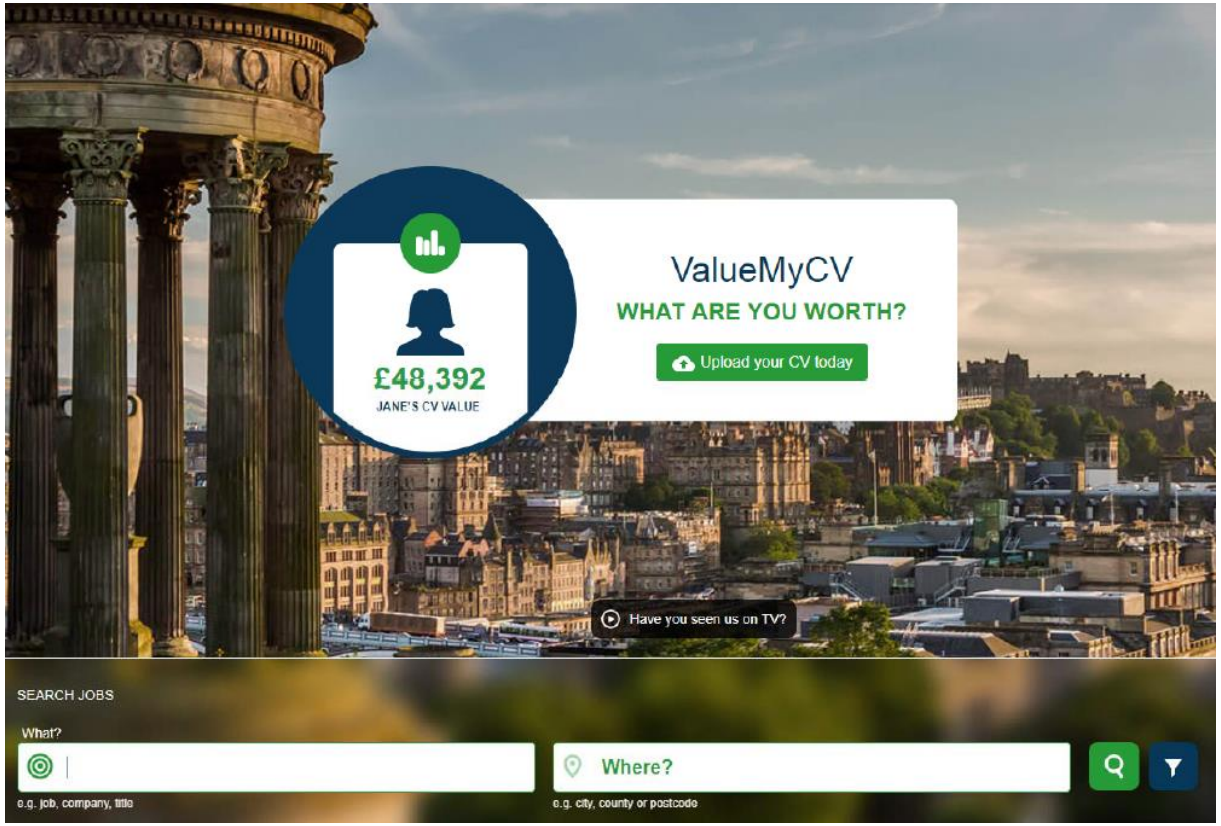
- Zoopla



OPEN DATA

•DIFFERENT USES OF OPEN DATA (3/3)

- Adzuna



OPEN DATA

- COMPANIES WITH BUSINESS MODELS BASED DATA SCIENCE OVER OPEN DATA
  - **T4 MEDIA**: Advertising company that uses Open Data to automate the process of identifying more appropriate outdoor advertising spaces
  - **RUBICON HERITAGE SERVICES LTD:** An archeology company that uses Open Data to enable fund allocation on its prospecting equipment and offer real-time verification
  - **MIME CONSULTING:** Consulting company tracking the status of the education system in UK to track learning success and recommend careers for teenagers

Analytics over Big Data sources: Data Science fundamentals

Alejandro Maté
Juan Carlos Trujillo