DATASCIENCE, LEARNING AND APPLICATIONS

DALAS - introduction

20 janvier 2024

Laure Soulier - Nicolas Baskiotis

SCIENCES SORBONNE UNIVERSITÉ

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

MLIA Machine Learning & Deep Learning for Information Access

Course organization

Course content

1. Introduction, data science
2. Web scrapping
3. Exploratory data analysis (EDA)
4. Data visualization (DataViz)
5. Storytelling and dashboard
6. Correlation analysis
7. Pipeline : regression
8. Pipeline : classification
9. Pipeline : unsupervised analysis
10. Deployment, docker and MLOps
11. Bias in ML

→ Lectures with basic notions, illustrative examples

→ TME including practical activities and a project

Organization

Horaires :

- Lectures : Monday 1.45-3.45pm
- TME : Thursday 1.45-6pm

Evaluation

- Continuous assessment - Project (50%)
  - Technical report (choices, algorithms, results, ...)
  - Defense considering a different audience : the client
- Final exam (50%)

Continuous assessment - Project (50%)

Project

- Identifying a topic, the data (crawled from the web, and possibly - in addition- from an open data portal)
- Cleaning, visualizing and analyzing data
- Performing different analyzes to answer to the initial topic
- Working group (2 people) : Git tutorial : https://github.com/baskiotisn/2IN013robot2023/blob/d979333fb80c9b6acd9515aaec040943d10d365c/docs/tutoriel_git.pdf

Objectives

Students ⇒ Data Engineer ⇒ Data Analyst ⇒ Data scientists

- Provide keys to understanding the role and management of data in companies
- Acquire data processing methodology for data science and machine learning
- Address data processing/analysis issues using concrete examples

But also...

- Develop creativity around data processing/analysis and its applications
- Teamwork

Context

Context : from Big Data to Data Science

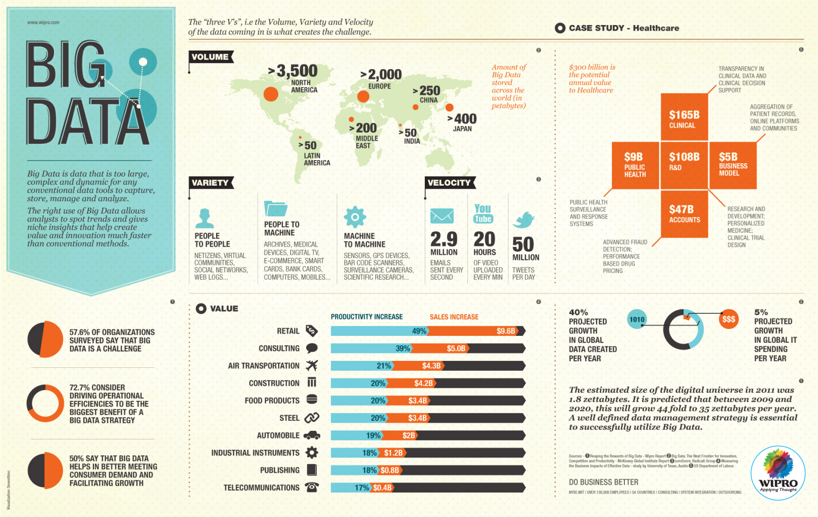
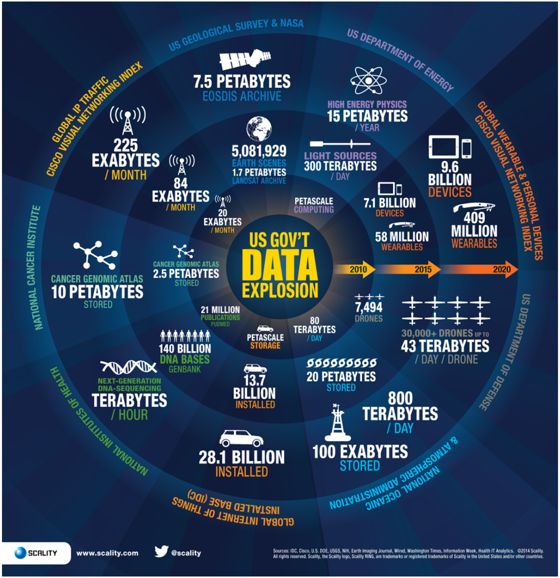


Figure 1 – Source : https://www.e-marketing.fr/Thematique/data-1091/Infographies/Saisir-big-data-infographie-196820.htm

Contexte

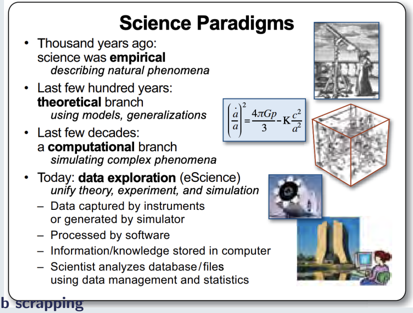


Sources: IDC, Cisco, U.S. DOE, USGS, NIH, Earth Imaging Journal, Wired, Washington Times, Information Week, Health IT Analytics. ©2014 Scality. All rights reserved. Scality, the Scality logo, Scality RING, are trademarks or registered trademarks of Scality in the United States and/or other countries.

Data driven science : The 4th paradigm

Extract from "The fourth paradigm " book

"I wanted to point out that almost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, ?data-intensive ? science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. Lots of new tools are needed to make this happen." (Jim Gray - Turing Price)



Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
    or generated by simulator
  – Processed by software
  – Information/knowledge stored in computer
  – Scientist analyzes database/files
    using data management and statistics

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$

The importance of data analytics for companies

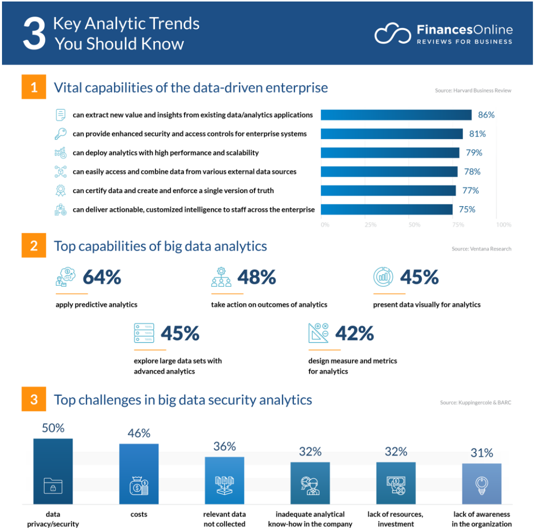


Figure 2 – Source : https://financesonline.com/

The importance of data analytics for companies

SCIENCES SORBONNE UNIVERSITÉ

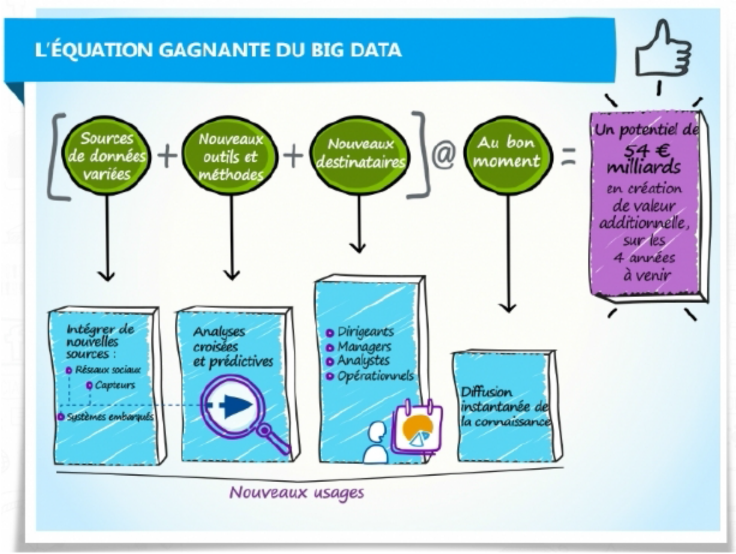
....

For those who succeed in optimizing its use, **data** becomes **information**, then, when properly shared within the company, it becomes and constitutes its **knowledge**. It can be a source of services and innovations, particularly when cross-referenced with other data from a variety of sources.
* *Enjeux Business des données - CIGREF 2014*

...

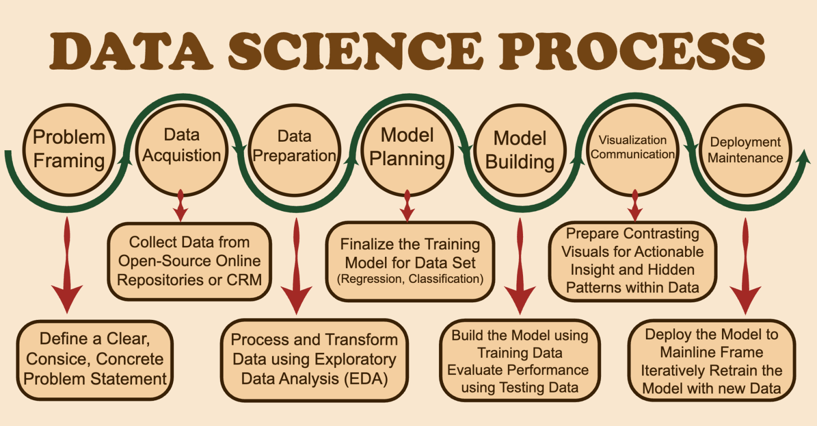
Data is therefore one of the **main intangible assets** of our organizations, and yet it is not yet managed with the same rigor and capital and human resources in particular. In a context where data has become critical to business activity, it is imperative to implement structured, industrial data management.
* *Enjeux Business des données - CIGREF 2014*

Etude IDC - Microsoft 2014


L'ÉQUATION GAGNANTE DU BIG DATA
Sources de données variées
+
Nouveaux outils et méthodes
+
Nouveaux destinataires
@
Au bon moment
=
Un potentiel de 54 € milliards en création de valeur additionnelle, sur les 4 années à venir
Intégrer de nouvelles sources :
Réseaux sociaux
Capteurs
Systèmes embarqués
Analyses croisées et prédictives
Dirigeants
Managers
Analystes
Opérationnels
Diffusion instantanée de la connaissance
Nouveaux usages

Data Science : process and jobs

Data Science process



DATA SCIENCE PROCESS

Problem Framing

Define a Clear, Consice, Concrete Problem Statement

Data Acquistion

Collect Data from Open-Source Online Repositories or CRM

Data Preparation

Process and Transform Data using Exploratory Data Analysis (EDA)

Model Planning

Finalize the Training Model for Data Set (Regression, Classification)

Model Building

Build the Model using Training Data Evaluate Performance using Testing Data

Visualization Communication

Prepare Contrasting Visuals for Actionable Insight and Hidden Patterns within Data

Deployment Maintenance

Deploy the Model to Mainline Frame Iteratively Retrain the Model with new Data

Figure 3 – Source https://www.embedded-robotics.com/data-science-process/

Data Science process : Problem framing

→ Identifying what to do : clear, concise, and concrete end-goal

- Discuss with your client & gather information about the problem
- Identify existing data. Warning GDRP, 1st party : collected by the company (CRM, analytics, logs...) - 2nd party : collected by a partner (e.g. media) - 3rd party : bought
- Identifying "marketing" use cases, evaluating the strategy, the difficulty, the delay, ...
- Identifying SMART objectives (Specific, Measurable, Reachable, Realistic, Temporal)

What is not Data Science

- Playing with data without objective
- Choose the analysis according to the results
- Technical difficulties takes precedence over the main objectives

Data Science process : Data Acquisition

→ Collecting data relevant for the objective

- Time to get hands dirty !
  - Seeking information sources : access modality/scrapping, local storage (database, datawarehouse, cloud, ...)
  - Might be dispersed, (un)structured, of variable quality, of heterogeneous format
- Different data sources
  - Online dataset repository
  - Web servers
  - Databases
  - Research labs
  - Social, administrative, public institutions
  - Survey repositories
  - API on website, on CRM, ...

Data Science process : Data Preparation

→ Organizing and cleaning data, resolving anomalies and unusual patterns

- Conditioning data to the same format, unifying labels
- Merging data. be careful of data format (date /hour format, postal code, phone number, ...)
- Identifying conflictual values (age cannot be negative, email must contain '@', ...)
- Identifying missing values, outliers, noise, bias. Deciding how to deal with them
- Remove duplicate values

Be aware of...

- GDRP
- data usage license
- privacy and protection of personal data

Data Science process : Model Planning and Building

→ The main skill of data scientist !

- Analyze and understand data from a visual perspective : Exploratory Data Analytics - EDA (variable distribution, correlation, ...)
- Identifying the suitable model/algorithm (regression, classification, generation, unsupervised methods, ...)
- Dividing data into training/testing set
- Sometime might be required to label/annotate data
- Building the model (training, parameter tuning, fine-tuning, ...)
- Evaluating its performances (VERY IMPORTANT ! ! ! !) : designing a protocol, metrics, tasks, qualitative analysis
- Validating with experts

Data Science process : Visualization - Communication

→ Presenting your results in a clear and pedagogical manner

- Data visualization, dashboards, ...
- Data Storytelling : be aware of your client background, understanding, skills.
- Match the result with the initial problem vocabulary

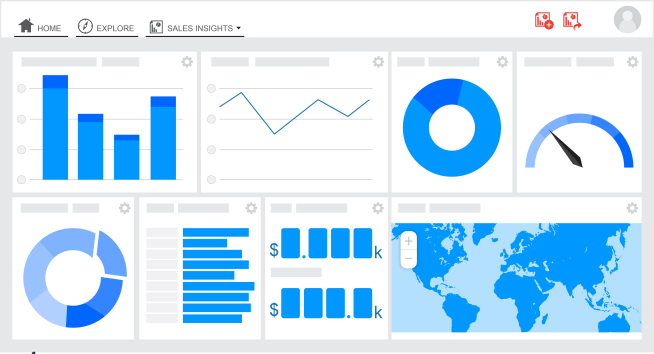


Figure 4 – https://commons.wikimedia.org/wiki/File:Infruid%27s_Self-Service_BI_Tool_Dashboard.jpg

Data Science process : Deployment - Maintenance

→ When your model is validated, must be used in real cases !

- Infrastructure/Docker/Server/Cloud
- MLOps : automatic monitoring of the data science pipeline
- Data science is an iterative process : new data acquisition, new feedback → new model

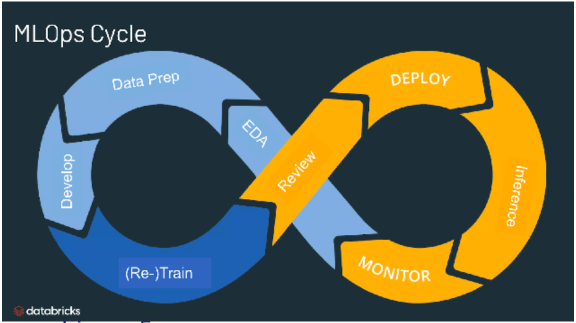


Figure 5 – https://www.databricks.com/fr/glossary/mlops

Use case

See example here : "Data Science Process : A Case-Study"
https://www.embedded-robotics.com/data-science-process/

Data Science process vs. jobs

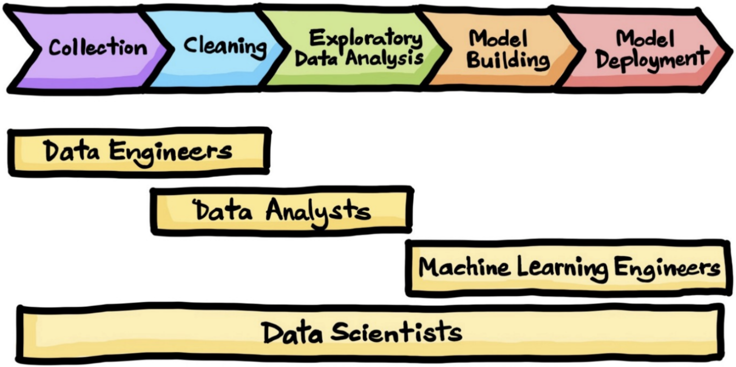
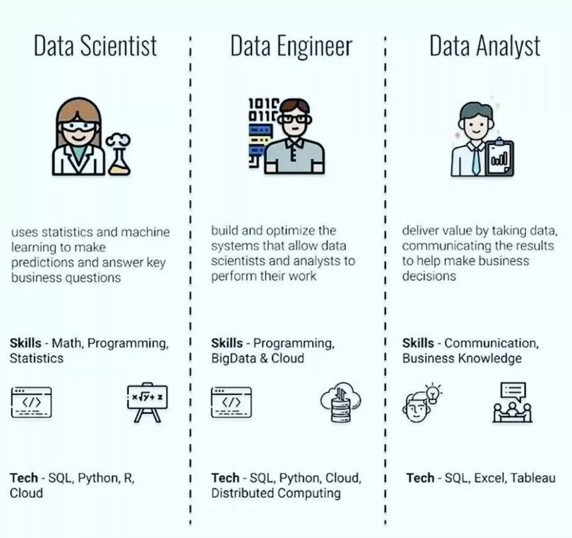


Figure 6 – Data science life cycle. (Drawn by Chanin Nantasenamat in collaboration with Ken Jee) - https://towardsdatascience.com/the-data-science-process-a19eb7ebc41b

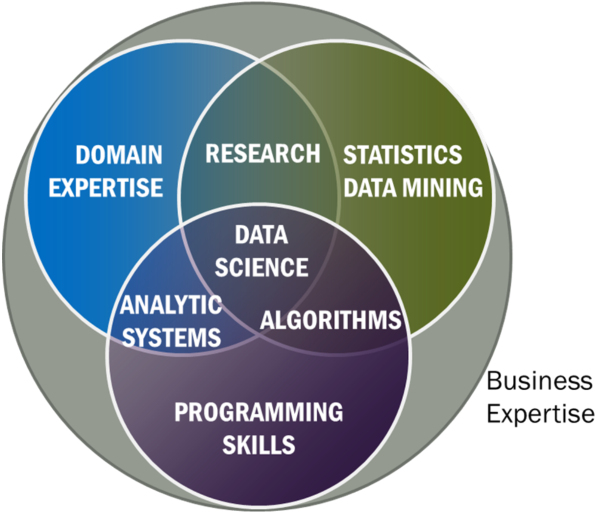
Data scientist skills



Data Scientist	Data Engineer	Data Analyst
uses statistics and machine learning to make predictions and answer key business questions	build and optimize the systems that allow data scientists and analysts to perform their work	deliver value by taking data, communicating the results to help make business decisions
Skills - Math, Programming, Statistics	**Skills** - Programming, BigData & Cloud	**Skills** - Communication, Business Knowledge
Tech - SQL, Python, R, Cloud	**Tech** - SQL, Python, Cloud, Distributed Computing	**Tech** - SQL, Excel, Tableau

Figure 7 – https://idreamcareer.com/blog/data-analyst-vs-data-scientist/

New challenges of data scientist in the big data area


DOMAIN EXPERTISE
RESEARCH
STATISTICS DATA MINING
DATA SCIENCE
ANALYTIC SYSTEMS
ALGORITHMS
PROGRAMMING SKILLS
Business Expertise