INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Course Presentation

IA Frameworks

Brendan Guillouet - 02/11/2020

Institut National des Sciences Appliquées de Toulouse

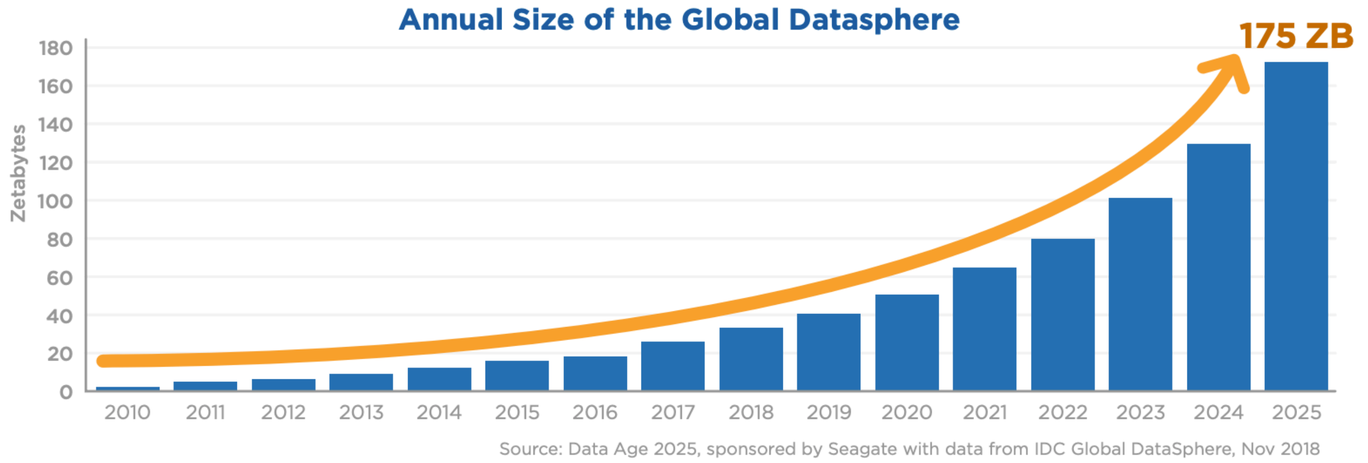
Table of contents

Definition & Context

IA & Big Data

Big Data and its 3 Vs:

- **Volume:** Exponential Growth
  - According to IDC: from 33 zeta bytes in 2018 to 175 in 2025



- **Volume:** numeric, signal, images, time series, graph, text etc...
- **Velocity:** real time decision
- (**Valorisation**)

Now **ARTIFICIAL INTELLIGENCE** (again)

- **BIG DATA**
- **COMPUTATION POWER AVAILABILITY**: GPU, cloud computing
- **'NEW ALGORITHMS'**: Deep Learning, Reinforcement learning,
- That produce impressive results, (Alpha Go, Atari Game defeated, write poetry etc..)

- … and benefits some good marketing

The definition is still very unclear, but IA is still NOT here.

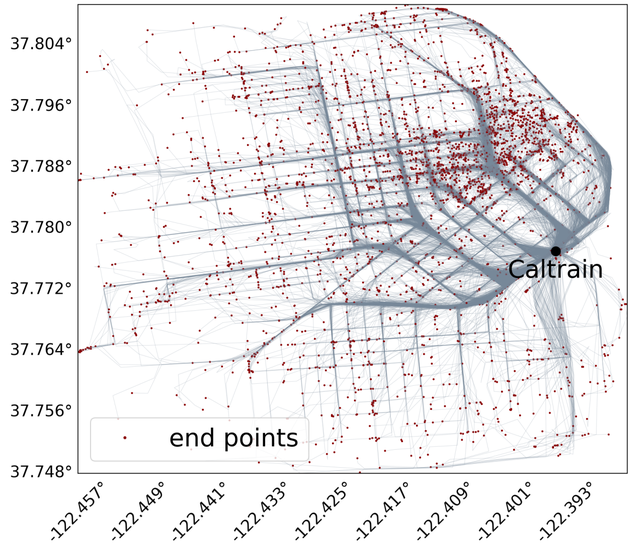
A wide range of application fields

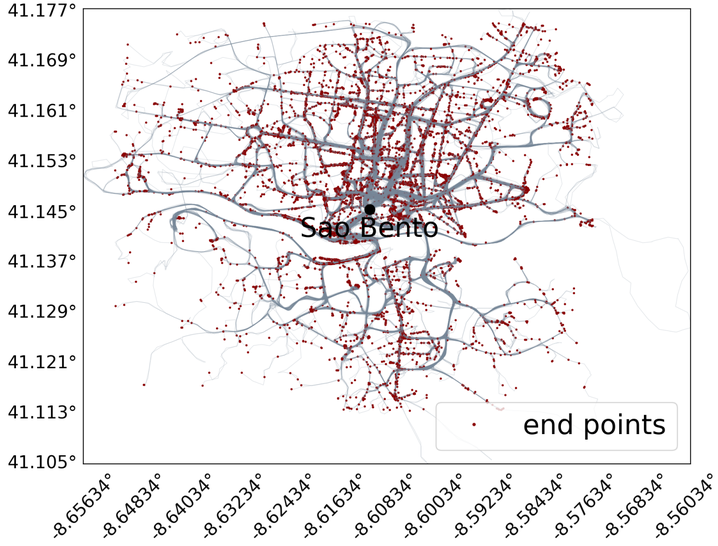
… at different levels of advancement.

- **E-commerce**: Recommendation system, advertising, targeting.
- **Text translation**: DeepL, google translate.
- **Image & video**: Classification, videosurveillance etc…
- **Game**: Chess, go, Atari game, Starcraft.

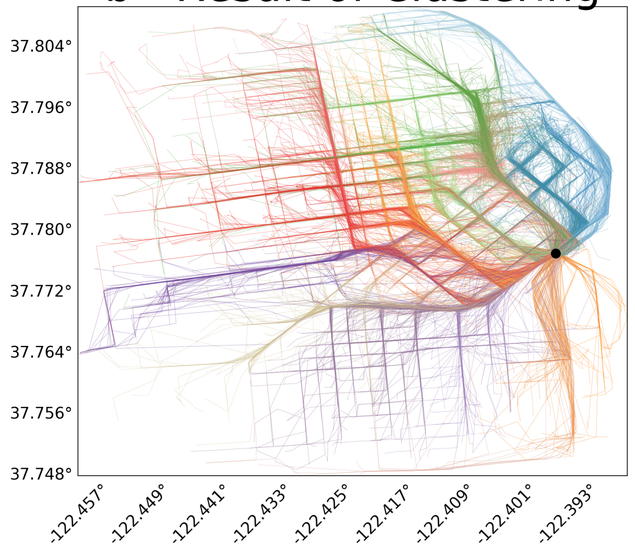
- **Public**: administrations, e-health (*open data*).
- **Industry**: 4.0 industry, autonomous vehicle.
- **Meteorology**: Hybridation of learning and physical model.
- **Certification.**
- **Internet of things.**

Variety: Trajectory clustering from GPS data

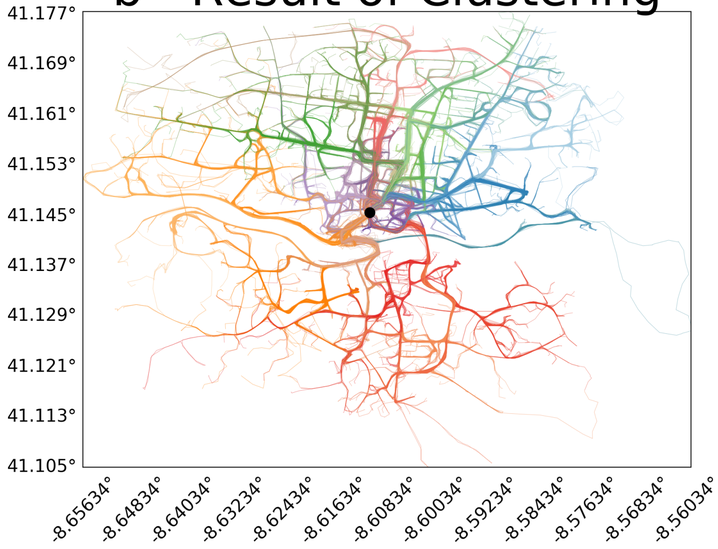




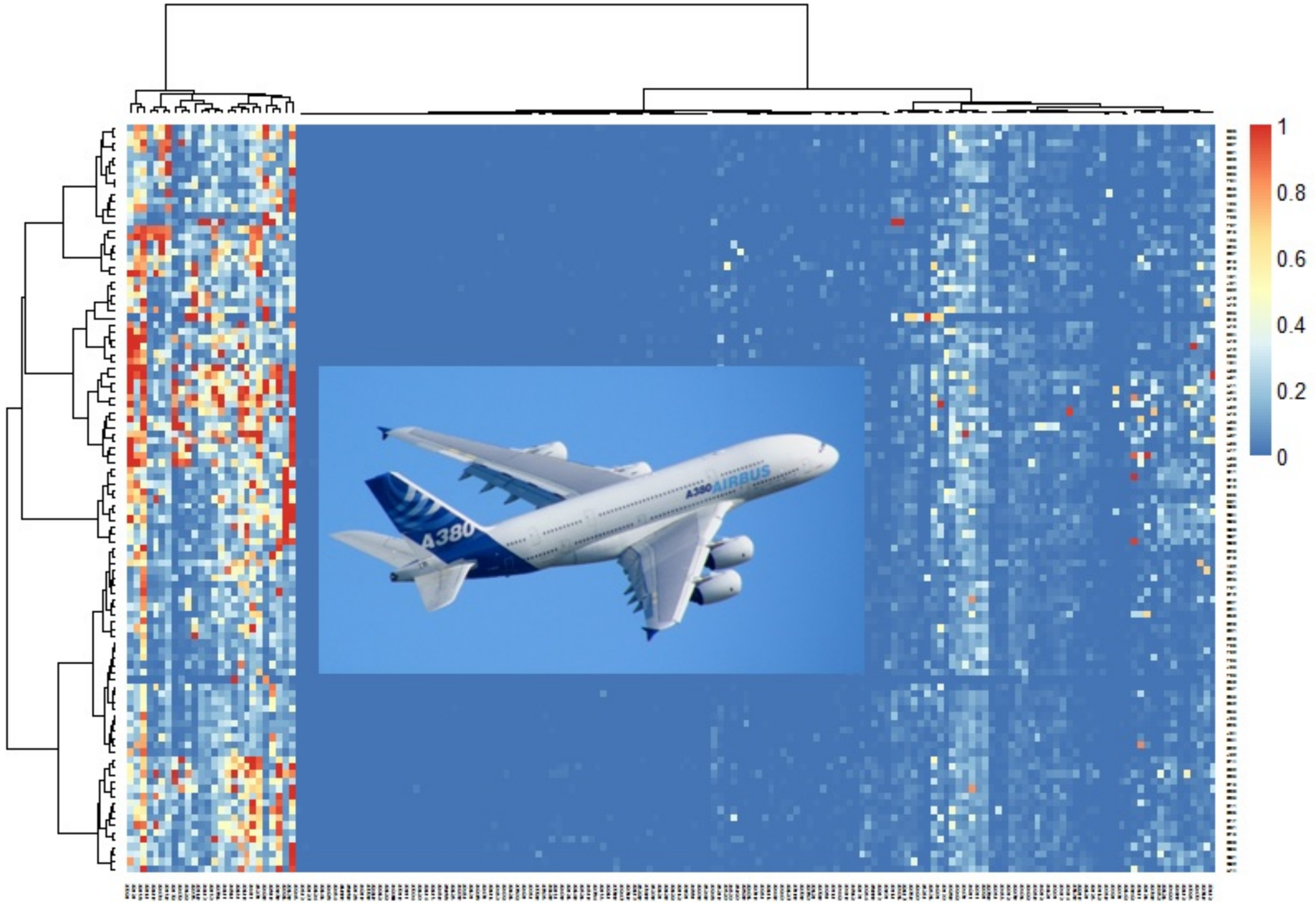
b - Result of Clustering



b - Result of Clustering



Analysis of flight incident messages



2019 INTERNSHIP OF INSA STUDENT

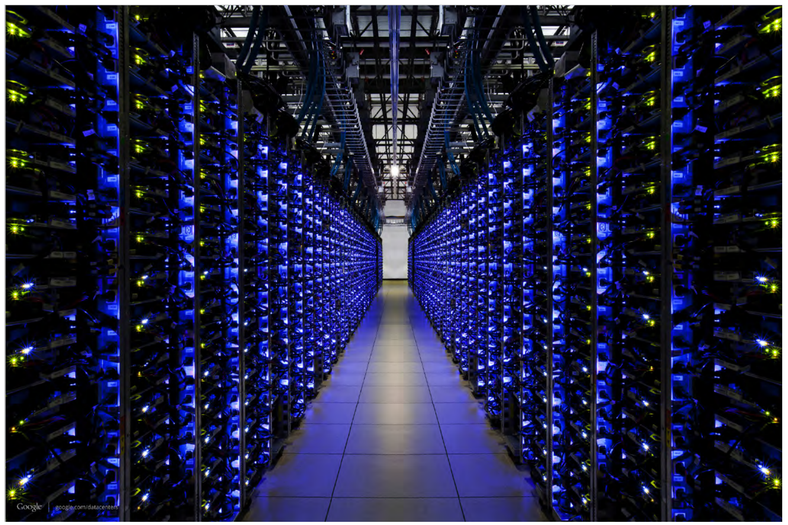
- Detection of field (agriculture) and building (military) from satellite images. *SSII & Startup*
- Anonymisation of name on text documents*. SSII*
- Fall detection of elder people*. PME*
- Autonomous drone flight*. SSII*
- Rain prediction in Ivory Coast*. Startup*
- Detection of diabetic retinopathy from x-ray photo*. Startup (Norway)*
- Estimation of fish results. *Research Center.*

2020 internship of INSA student.

TODO

What do we need to start?

- Local or cloud computing? (GCP, AWS, AZURE, Custom solution?)
- Size of disk ? Size of RAM? CPU or GPU? How many of its ?
- From where and how di we get the data: local/bucket? Security?
- How are they stored (csv, json, database..) ?
- Distributed architecture?
- In which format do we produce the results?
- How to automate this?



Job specification

Task	Data Engineer	Data Analyst	Data Scientist
Managing data warehousing.	✓✓	✗	✗
Ensure data accessibility & consumption via API.	✓✓	✗	✗✗
Continuously monitoring and testing the system to ensure optimized performance.	✓✓	✗	✗
Extraction Transformation & Loa.	✓✓	✓	✓
Cleaning and organizing raw data.	✓	✓	✓
Programming skills.	✓	✓	✓
Using descriptive statistics to get a big-picture view of their data.	✗	✓✓	✓
Analyzing interesting trends found in the data.	✗	✓✓	✓
Visualisations and dashboards to help the company interpret and make decisions from data.	✗	✓✓	✓
Presenting the results of a technical analysis to business clients or internal teams.	✗	✓✓	✓
Evaluating statistical models to determine the validity of analyses.	✗	✓	✓✓
Using machine learning to build better predictive algorithms.	✗	✓	✓✓
Testing and continuously improving the metrics of ML models.	✗	✗	✓✓
Continuously performs scientific Watch	✗	✗	✓✓
Building data visualisation to summarise the conclusion of an advanced analysis.	✗	✗	✓✓

Job specification

You are here

Task	Data Engineer	Data Analyst	Data Scientist
Managing data warehousing.	✓✓	✗	✗
Ensure data accessibility & consumption via API.	✓✓	✗	✗✗
Continuously monitoring and testing the system to ensure optimized performance.	✓✓	✗	✗
Extraction Transformation & Loa.	✓✓	✓	✓
Cleaning and organizing raw data.	✓	✓	✓
Programming skills.	✓	✓	✓
Using descriptive statistics to get a big-picture view of their data.	✗	✓✓	✓
Analyzing interesting trends found in the data.	✗	✓✓	✓
Visualisations and dashboards to help the company interpret and make decisions from data.	✗	✓✓	✓
Presenting the results of a technical analysis to business clients or internal teams.	✗	✓✓	✓
Evaluating statistical models to determine the validity of analyses.	✗	✓	✓✓
Using machine learning to build better predictive algorithms.	✗	✓	✓✓
Testing and continuously improving the metrics of ML models.	✗	✗	✓✓
Continuously performs scientific Watch	✗	✗	✓✓
Building data visualisation to summarise the conclusion of an advanced analysis.	✗	✗	✓✓

A new economic model

Situation :

- Algorithm are available and free (python library, github, etc..)
- Most used software (Tensorflow, spark) are under free licences (GNU, MIT, Apache)

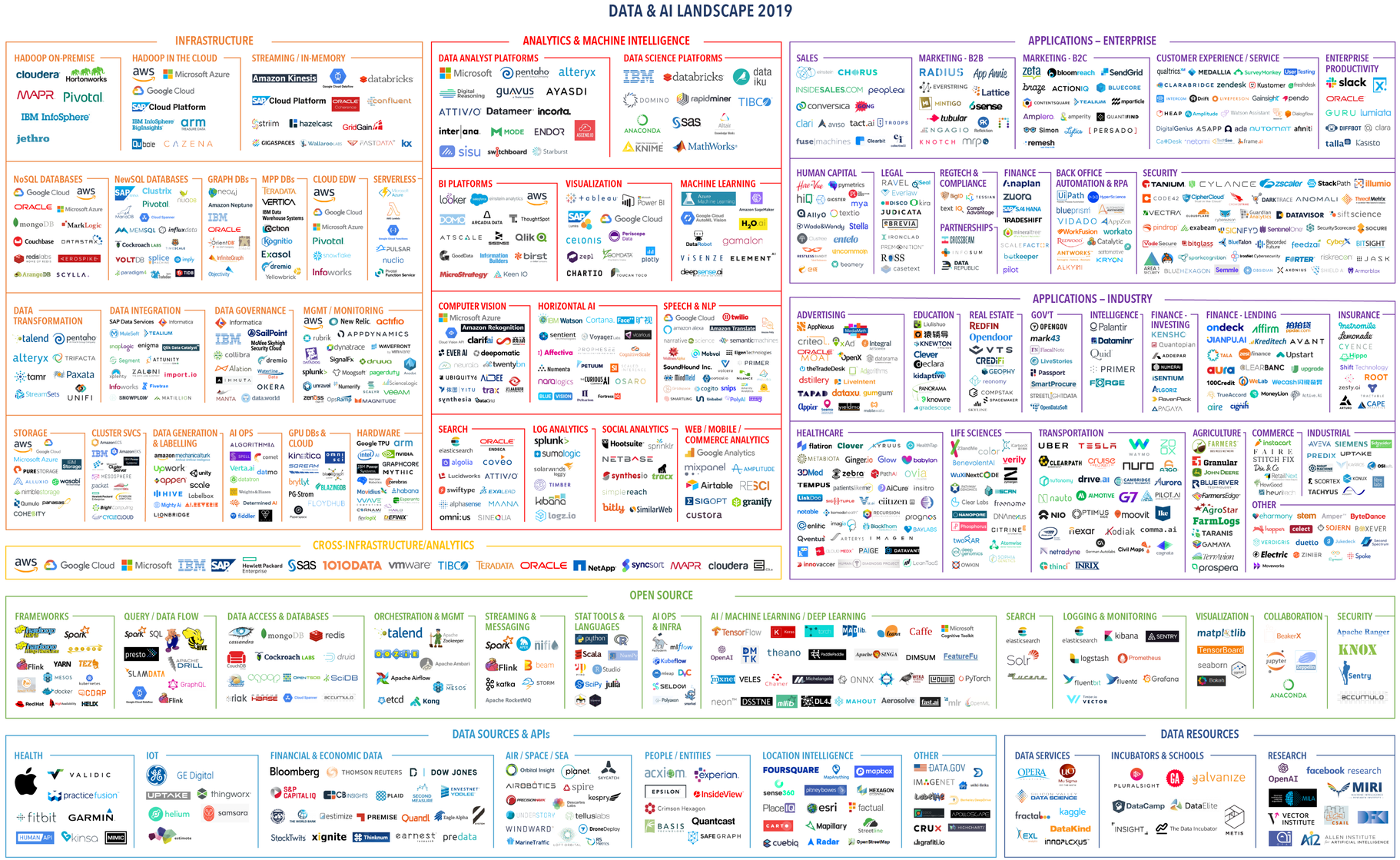
New strategy: **Sell services.**

- Enthought **(Canopy),** continuum analytics **(Anaconda),** Horton works, Cloudera **(Hadoop, Spark),** databricks **(Spark),** Oxdata **(H20)**, Microsoft, etc..

New market: **Cloud computing** (pick-and-shovel play)

- Actors: Amazon web service platform, Microsoft Azure, Google cloud computing.
- A Hughe variety of services: Storange, ML instance, ML pipeline, dataflow, etc…

BIG DATA & AI LANDSCAPE (TURCK, 2019)



(New ?) Data science

As a **Statistician :**

- Deep comprehension of algorithms.
- Scientific watch.

What's really new ?

- Programming skills
- Better understanding of all data-pipeline technology
- Manage bigger data and associate tools.
- New question about ethics.

(New ?) Data science

As a **Statistician :**

- Deep comprehension of algorithms.
- Scientific watch.

What's really new ?

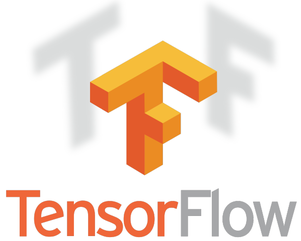
- Programming skills
- Better understanding of all data-pipeline technology
- Manage bigger data and associate tools.
- New question about ethics.

Learn to "self-learn"

Organisation & Evaluation

Objective - Tools

ML Python Libraries



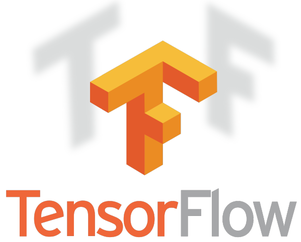
Python Environment

Viz' Python Libraries

Framework & Tool

Objective - Tools

ML Python Libraries



Python Environment

Viz' Python Libraries

Framework & Tool

Objectives - Data & Algorithm

Machine learning & High dimensional Statistics

Data

- Structured data
- Time series & Signal
- Image

Algorithm

- Machine learning: *SVM, RF, Regression*
- Deep learning: *MLP, CNN*

Objectives - Data & Algorithm

Machine learning & High dimensional Statistics

Data

- Structured data
- Time series & Signal
- Image

Algorithm

- Machine learning: *SVM, RF, Regression*
- Deep learning: *MLP, CNN*

Artificial Intelligence Frameworks

Data

- Text *(Cdiscount)*
- Video Game *(pacman like)*
- Note data *(movielens)*

Algorithm

- Text cleaning : TFIDF, steming
- More Deep learning: *RNN, Words embedding*
- *Reinforcement learning: PG, Q-learning*
- *Recommendation system: NMF*

Six labs

Session 1 - 02/11/20

- Natural language processing: Text Cleaning + Text Vectorisation.
- Natural language processing: Words Embedding.

Session 2 - 16/11/20

- Natural language processing: Recurrent Network.
- Python environment + Github Repo + Python Script.

Session 3 - 30/11/20

- Introduction to google cloud computing.
- Docker.

Session 4 - 07/12/20

- Reinforcement learning: PG Gradient.
- Reinforcement learning: Deep Q-learning.

Session 5 - 14/12/20

- Recommendation system.
- ???

Session 6 - 04/01/21 - Free time

Evaluation- Objective

Evaluation is linked to defi IA's 2021 challenge.

You will be evaluated on your capacity of acting like a Data Scientist, ie :

- Handle a new dataset and explore it.
- Find a solution to address the challenge's problem with a high score (above baseline).
- Explain the chosen algorithm.
- Write a complete pipeline to easily reproduce the results.
- Justify the choice of the algorithm and the environment (CPU/GPU, Cloud etc..).
- Share it and make your results easily reproducible (git, docker/conda environment.).

Evaluation

Oral Presentation or Rapport (40%)

- Quality of the presentation - 25%.
- In-deep explanation of the chosen algorithm - 25%.
- Choice of the tools-infrastructure used - 25%.
- Results obtained - 25%.

Evaluation

Oral Presentation or Rapport (40%)

- Quality of the presentation - 25%.
- In-deep explanation of the chosen algorithm - 25%.
- Choice of the tools-infrastructure used - 25%.
- Results obtained - 25%.

Date/deadline: ???

Evaluation

Project (60%)

- The git contains a clear markdown readme, which describes : -33%
  - Which result you achieved? In which computation time? On which engine?
  - What do I have to install to be able to reproduce the code?
  - Which command do I have run to reproduce the results ?
- The code is easily reproducible : -33%
  - Packages required has to be well described (a *requirements.txt* files is the best)
  - Python / Conda / docker are furnished.
- The code should be clear en easily readable : -33%
  - Final results can be run in a script and not a notebook.
  - Only final code can be found in the script

Evaluation

Project (60%)

- The git contains a clear markdown readme, which describes : -33%
  - Which result you achieved? In which computation time? On which engine?
  - What do I have to install to be able to reproduce the code?
  - Which command do I have run to reproduce the results ?
- The code is easily reproducible : -33%
  - Packages required has to be well described (a *requirements.txt* files is the best)
  - Python / Conda / docker are furnished.
- The code should be clear en easily readable : -33%
  - Final results can be run in a script and not a notebook.
  - Only final code can be found in the script

Date/deadline: ???