INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

IA Frameworks

Course Presentation

Brendan Guillouet

November ??, 2020

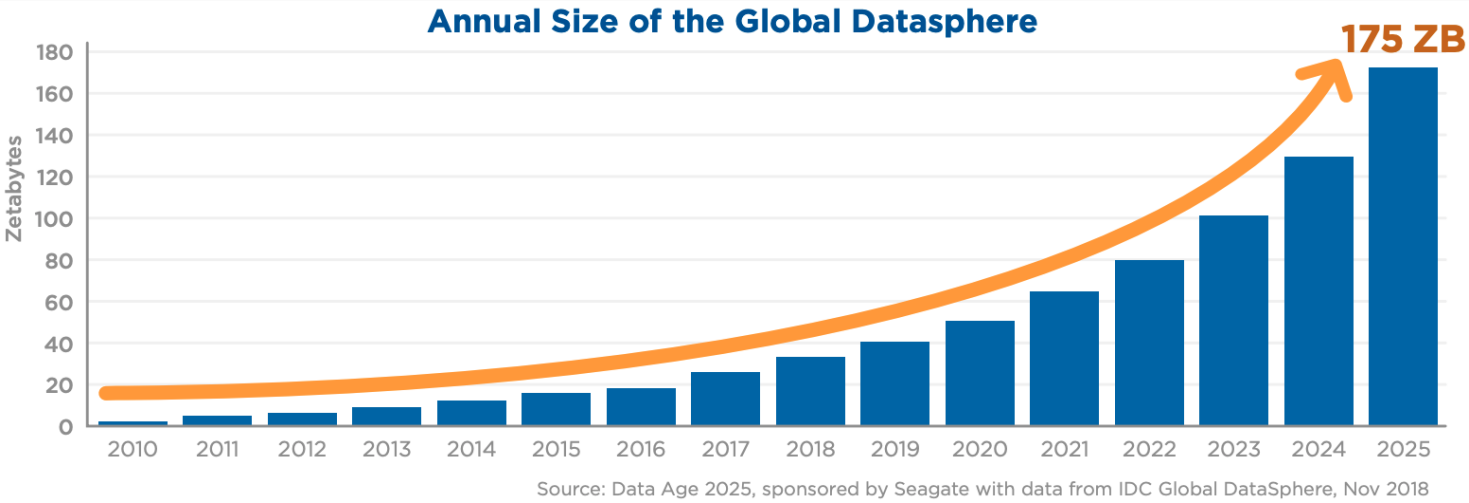
Institut National des Sciences Appliquées de Toulouse

Table of contents

Definition & Context

Big Data and its 3 Vs :

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

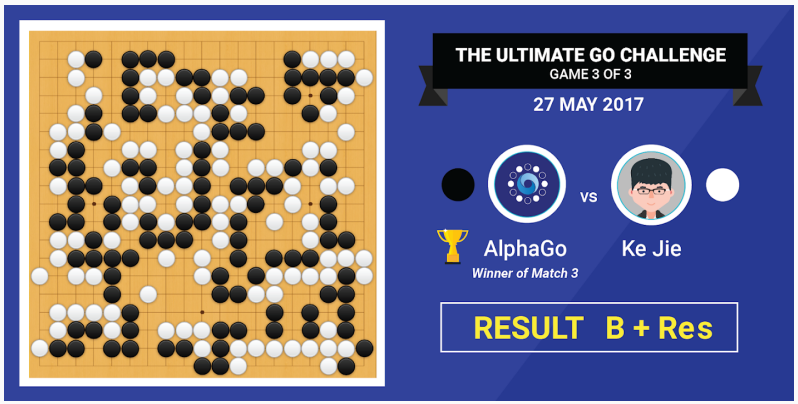


- **Variety** : numeric, signal, images, time series, graph, text etc...
- **Velocity** : real time decision.
- (**Valorisation**).

IA & Big Data

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..).

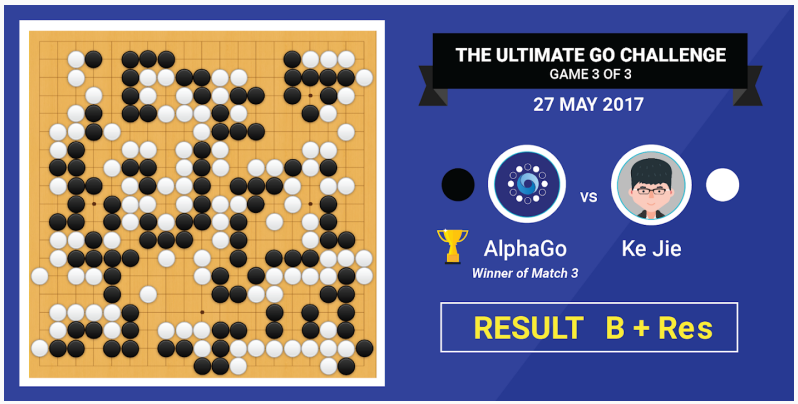


- .. and benefits some good marketing.

IA & Big Data

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.

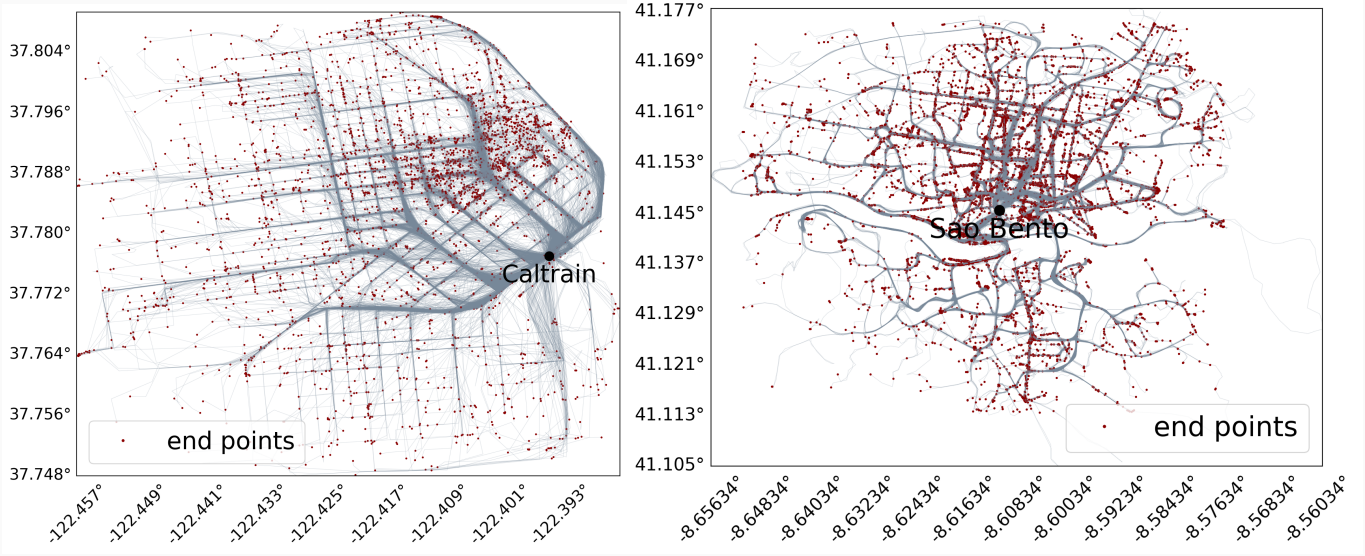
A wide range of application fields...

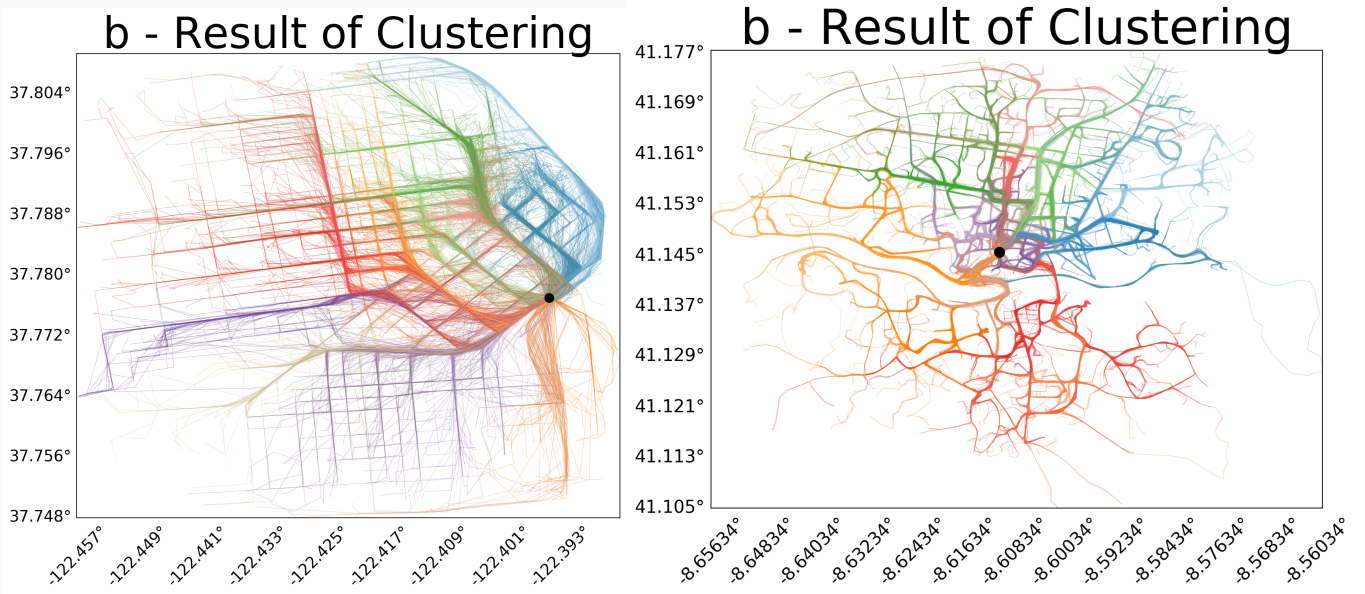
..at different levels of advancement.

- **E-commerce** : Recommendation system, advertising, targeting.
- **Text translation** : DeepL , Google Translate.
- **Image & Video** : Classification, videosurveillance etc..
- **Game** : Chess, Go, Atari Game, StarCraft.

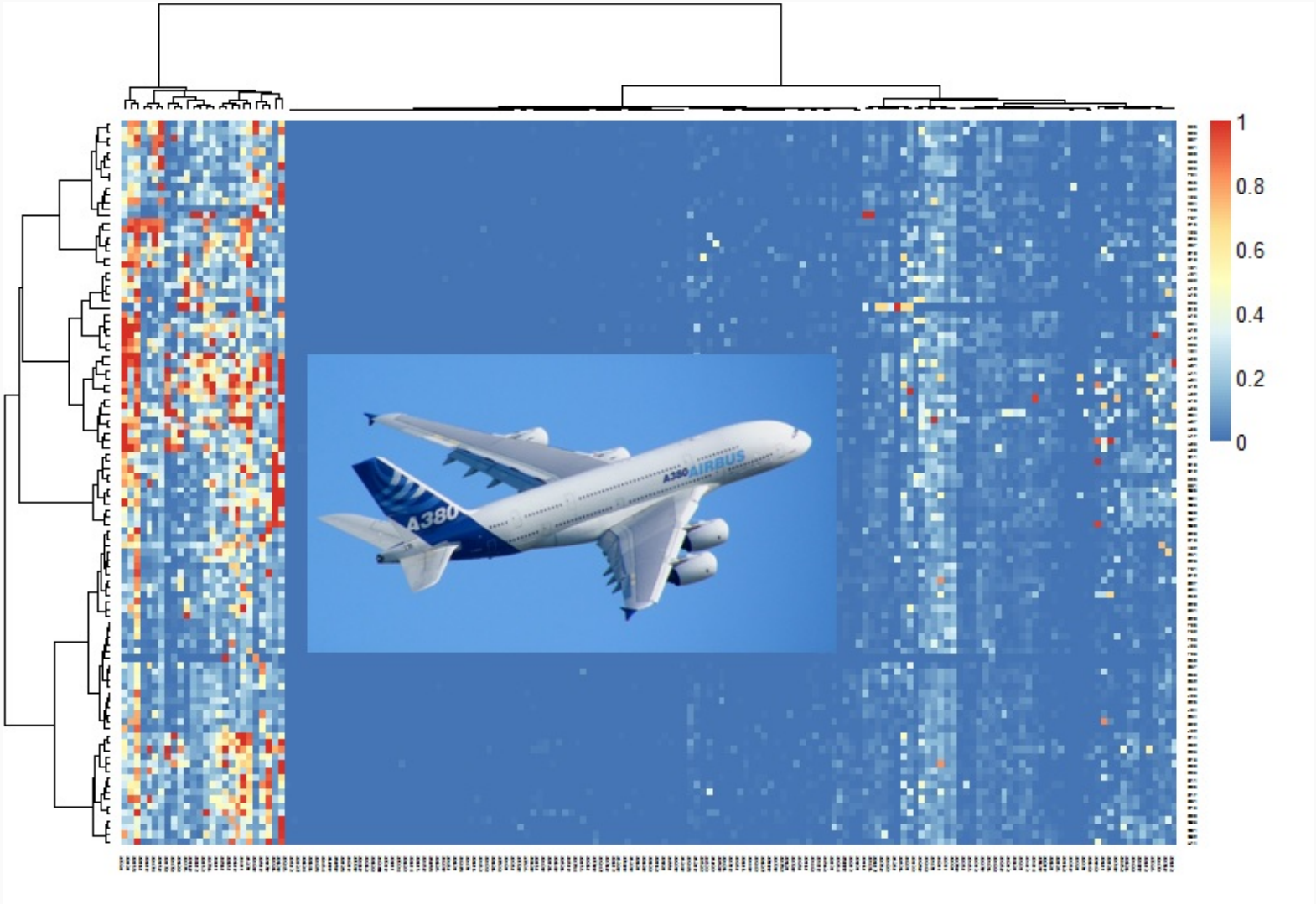
- **Public** : administrations, e-health (*open data*)
- **Industry** : 4.0 industry, Autonomous vehicle.
- **Meteorology** : Hybridation of learning and physical model
- **Certification**.
- **Internet of things**.

Variety : Trajectory clustering from GPS data.





Analysis of flight incident messages



2020 internship of INSA student

- Detection of field (agriculture) and batiment (military) from satellite images. *SSII & Startup*
- Anonymization 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*

Job Specification

Task

- Managing data Wharehousing
- Ensure Data accessability & consumption via API.
- Continuously monitoring and testing the system to ensure optimized performance.
- Extraction Transformation & Load
- 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.
- Creating visualizations and dashboards to help the company interpret and make decisions with the 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 accuracy of machine learning models.
- Continuously Perform scientific Watch.
- Building data visualizations to summarize the conclusion of an advanced analysis.

Job Specification

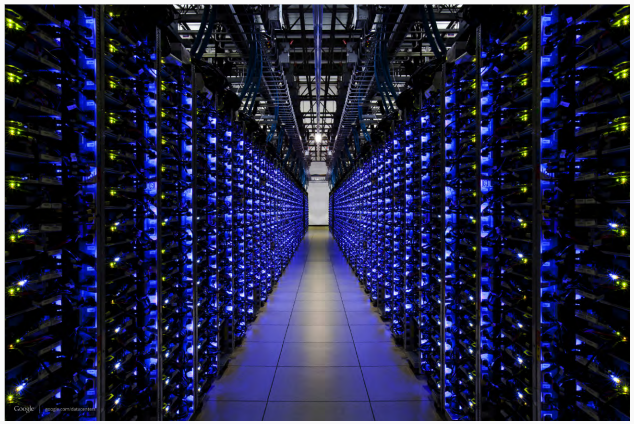
Task	Data Engineer	Data Analyst	Data Scientist
• Managing data Wharehousing	✔✔	✗	✗
• Ensure Data accessability & consumption via API.	✔✔	✗	✗
• Continuously monitoring and testing the system to ensure optimized performance.	✔✔	✗	✗
• Extraction Transformation & Load	✔✔	✔	✔
• Cleaning and organizing raw data.	✔	✔	✔
• Programming skills.	✔	✔	✔
• Using descriptive statistics to get a big-picture view of their data.	✗	✔✔	✔
• Analyzing interesting trends found in the data.	✗	✔✔	✔
• Creating visualizations and dashboards to help the company interpret and make decisions with the 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 accuracy of machine learning models.	✗	✗	✔✔
• Continuously Perform scientific Watch.	✗	✗	✔✔
• Building data visualizations to summarize the conclusion of an advanced analysis.	✗	✗	✔✔

Job Specification

Task	Data Engineer	Data Analyst	You're Here ↓ Data Scientist
• Managing data Wharehousing	✔✔	✗	✗
• Ensure Data accessability & consumption via API.	✔✔	✗	✗
• Continuously monitoring and testing the system to ensure optimized performance.	✔✔	✗	✗
• Extraction Transformation & Load	✔✔	✔	✔
• 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.	✗	✔✔	✔
• Creating visualizations and dashboards to help the company interpret and make decisions with the 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 accuracy of machine learning models.	✗	✗	✔✔
• Continuously Perform scientific Watch.	✗	✗	✔✔
• Building data visualizations to summarize the conclusion of an advanced analysis.	✗	✗	✔✔

What do we need to start ?

- Local or cloud computing ? *(GCP, AWS, AZURE)*
- Size of disk ? Size of RAM ? CPU or GPU ? Hom many of its ?
- From where and how did you get the data : local/bucket ?
- How are they stored ? Csv, Data base ? Which one ?
- Distributed architecture ? *(Hadoop, Spark)*
- In which format do you expect the results ?
- How it is automate ?



DataFarming

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)
- **Margin** on hardware are not really valuable

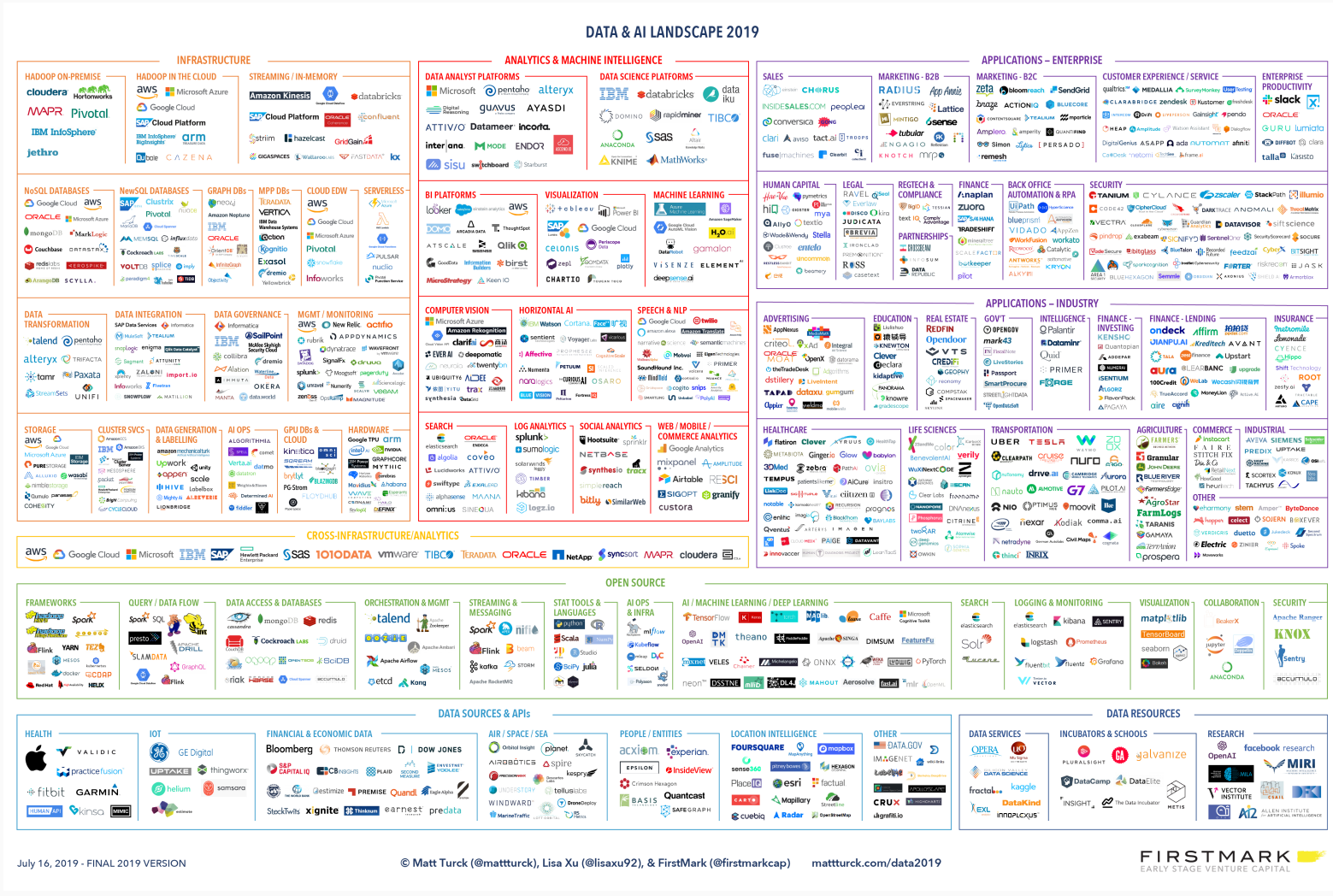
New Strategy : **Sell services**

- Enthought (**Canopy**), Continuum analytics (**Anaconda**), Horthon Works, Cloudera (**Hadoop, Spark...**), Databricks (**Spark**), Oxdata (**H20**), Revolution Analytics (**RHadoop**) – Microsoft

New market : **Cloud computing** (Pick-And-Shovel Play)

- Actors : Amazon Web Service Platform, Microsoft Azure, Google Cloud Computing **aaS**, Software **aaS**, Service **aaS**...
- A huge variety of services, Storage, Instances, 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 for a **Data Scientist**.

- Programming skills.
- Better understanding of all data-pipeline technology.
- Manage bigger data.
- New questions about ethics.

$\Longrightarrow$ Learn to "**self-learn**".

Organization & Evaluation

Objectives - Tools

Tools you know

ML Python Libraries

Python Environment

Viz' Python Libraries

Framework & Tool

Objectives - Tools

Tools you'll discover

Objectives - Tools

Tools you'll discover

Question :

- What are the best tools to used ? In which circumstances ?
- Is it valuable to use CPU/GPU ? Own laptop/cloud resources ? Distributed architecture ?

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 (*MovieLense*)

Algorithm

- Text cleaning : *TFIDF, steming.*
- More Deep Learning : *RNN, Words Embedding.*
- Reinforcement learning : *Policy Gradient / Q-learning.*
- Recommendation System : *NMF.*

Six - Labs

- Session 1 - ??- ??-20
  - Github Reminder, Python environment and Python Script.
  - Introduction to Google Cloud Computing.
- Session 2 - ??- ??-20
  - Containerization with Docker.
  - Natural language processing : Text Cleaning + Text Vectorization.
- Session 3 - ??- ??-20
  - Natural language processing : Words Embedding.
  - Natural language processing : Recurrent Network.
- Session 4 - ??- ??-20
  - Reinforcement learning : PG Gradient.
  - Reinforcement learning : Deep Q-learning.
- Session 5 - ??- ??-20
  - Recommendation System.
  - Data Visualization.
- Session 6 - ??- ??-20
  - ”Free Time”.

Evaluation - Objective

You will be evaluated on your capacity of acting like a Data Scientist, i.e.

1. Understand an algorithm you haven't seen during course.
2. Being able to explain it clearly.
3. Make it run on an dataset to evaluate its performances.
4. Make it run on the appropriate tools (SPark ? Cloud ? GPu ?)
5. Share it and make your results easily reproducible (Git - docker, conda environment.).

Notation

- Oral presentation - (40%)
- Project - (60%)

Evaluation

Oral presentation - (40%)

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

Date : January ??, 2020.

Evaluation

Project - (60%) :

- The git should contain 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 to run to reproduce the results ?
- The code has to be easily reproducible. (33%)
  - Packages required has to be well described. (a *requirements.txt* files is the best)
  - Conda command or docker command can be furnish
- The code should be clear and easily readable. (33%)
  - Final results can be run in a script and not a notebook.
  - Only final code can be found in this script.

Deadline : January ?? 2020.

Evaluation

$\Longrightarrow$ Some proposition of subjects

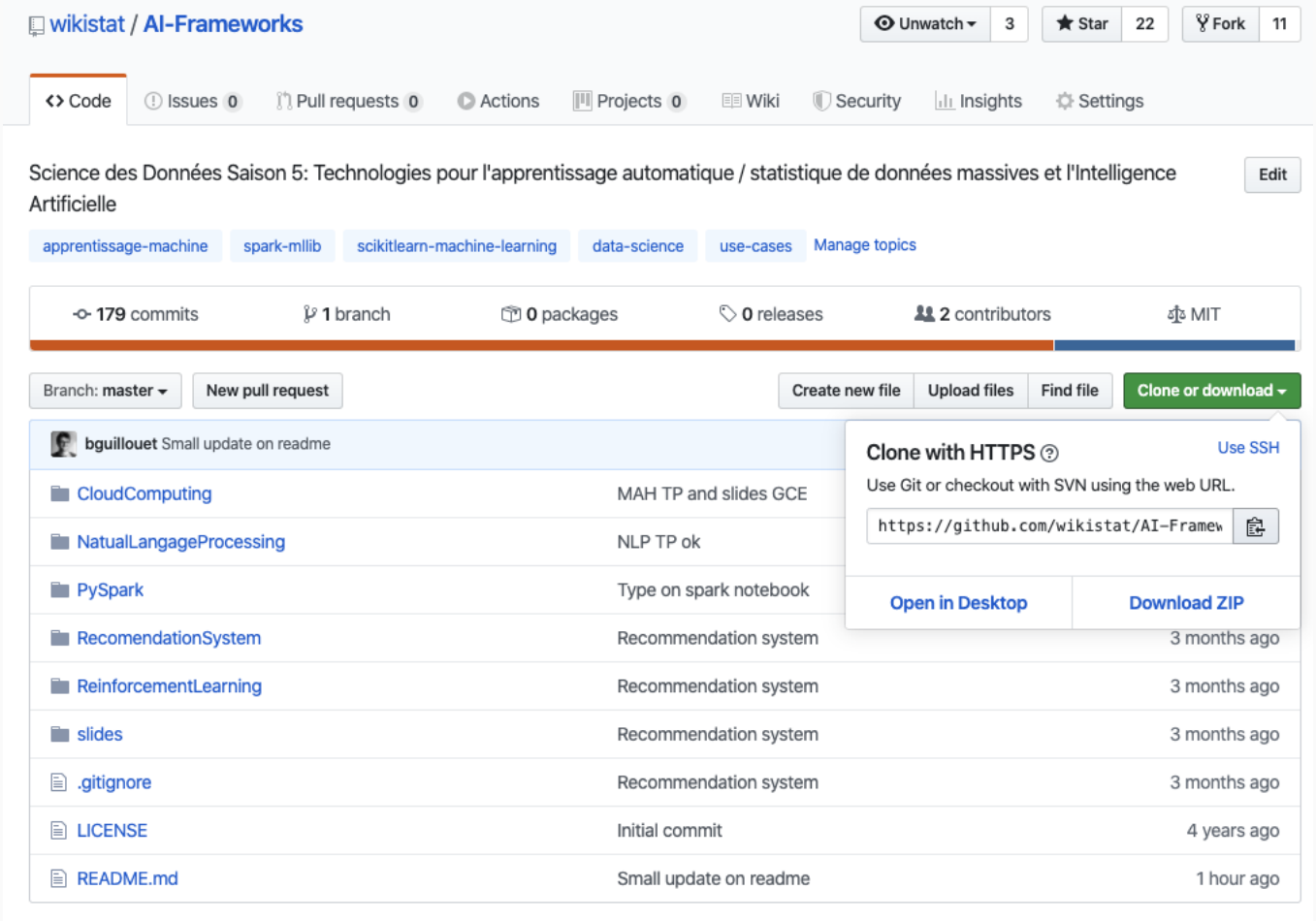
- Image : Generative Adversarial Network.
- Image : Object segmentation & Localization
- Image : Siamese Network.
- One-Shot Learning.
- NLP :Recurrent Neural Network (Text generation)
- NLP : Words Embedding Comparison.
  (Glove/Fasttestx/Bert/Elmo)
- RL : Actor Critic.
- RL : Application on real data (Pacman).
- Google CLoud : Comparizon of AutoML with other services.

$\Longrightarrow$ You're free to choose the infrastructure you want.

$\Longrightarrow$ Group of 4 to 5.

Github Reminder

Github - clone



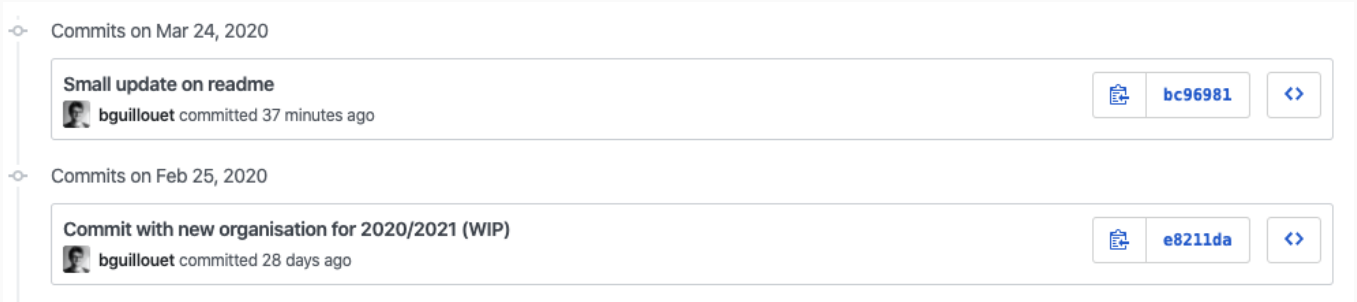
Github - get new commit

Local last commit

```
~/INSA/AI-Frameworks master* ↓
INSA ❯ git log
commit e8211da86b578d69641c8892dc969d1a44013d7d (HEAD -> master)
Author: bguillouet <brendan.guillouet@gmail.com>
Date:   Tue Feb 25 15:46:47 2020 +0100

    Commit with new organisation for 2020/2021 (WIP)
```

Github



Github - pull

```
~/INSA/AI-Frameworks master* ↓ 10s
INSA ❯ git status
On branch master
Your branch is behind 'origin/master' by 1 commit, and can be fast-forwarded.
  (use "git pull" to update your local branch)

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git checkout -- <file>..." to discard changes in working directory)

        modified:   README.md

Untracked files:
  (use "git add <file>..." to include in what will be committed)

        new_filte.txt

no changes added to commit (use "git add" and/or "git commit -a")

~/INSA/AI-Frameworks master* ↓
INSA ❯ git pull origin master
From github.com:wikistat/AI-Frameworks
 * branch            master     -> FETCH_HEAD
Updating e8211da..bc96981
error: Your local changes to the following files would be overwritten by merge:
        README.md
Please commit your changes or stash them before you merge.
Aborting
```

Github - commit & pull

```
~/INSA/AI-Frameworks master* ↓
INSA ❯ git add README.md

~/INSA/AI-Frameworks master* ↓
INSA ❯ git commit -m "my commit"
[master 7735af0] my commit
 1 file changed, 31 insertions(+), 10 deletions(-)

~/INSA/AI-Frameworks master* ↓↑
INSA ❯ git pull origin master
From github.com:wikistat/AI-Frameworks
 * branch            master     -> FETCH_HEAD
Auto-merging README.md
CONFLICT (content): Merge conflict in README.md
Automatic merge failed; fix conflicts and then commit the result.
```

Github - create new branch

```
~/INSA/AI-Frameworks master* ↓
INSA ❯ git status
On branch master
Your branch is behind 'origin/master' by 1 commit, and can be fast-forwarded.
  (use "git pull" to update your local branch)

Changes not staged for commit:
  (use "git add <file>..." to update what will be committed)
  (use "git checkout -- <file>..." to discard changes in working directory)

        modified:   README.md

Untracked files:
  (use "git add <file>..." to include in what will be committed)

        new_filte.txt

no changes added to commit (use "git add" and/or "git commit -a")

~/INSA/AI-Frameworks master* ↓
INSA ❯ git checkout -b TP1
M       README.md
Switched to a new branch 'TP1'

~/INSA/AI-Frameworks TP1*
INSA ❯ git add README.md

~/INSA/AI-Frameworks TP1*
INSA ❯ git commit -m "my commit"
[TP1 6a37c81] my commit
```

Github - pull on master

```
~/INSA/AI-Frameworks TP1*
INSA ❯ git checkout master
Switched to branch 'master'
Your branch is behind 'origin/master' by 1 commit, and can be fast-forwarded.
  (use "git pull" to update your local branch)

~/INSA/AI-Frameworks master* ↓
INSA ❯ git pull origin master
From github.com:wikistat/AI-Frameworks
 * branch            master     -> FETCH_HEAD
Updating e8211da..bc96981
Fast-forward
 README.md | 11 ++++++++---
 1 file changed, 8 insertions(+), 3 deletions(-)

~/INSA/AI-Frameworks master*
INSA ❯ git log
commit bc96981b978d6f12693d64aa2e4cbdf8da945160 (HEAD -> master, origin/master, origin/HEAD)
Author: bguillouet <brendan.guillouet@gmail.com>
Date:   Tue Mar 24 16:05:06 2020 +0100

    Small update on readme

commit e8211da86b578d69641c8892dc969d1a44013d7d
Author: bguillouet <brendan.guillouet@gmail.com>
Date:   Tue Feb 25 15:46:47 2020 +0100

    Commit with new organisation for 2020/2021 (WIP)
```