

COURSE PRESENTATION

AI FRAMEWORKS

TABLE OF CONTENTS

Definition & Context

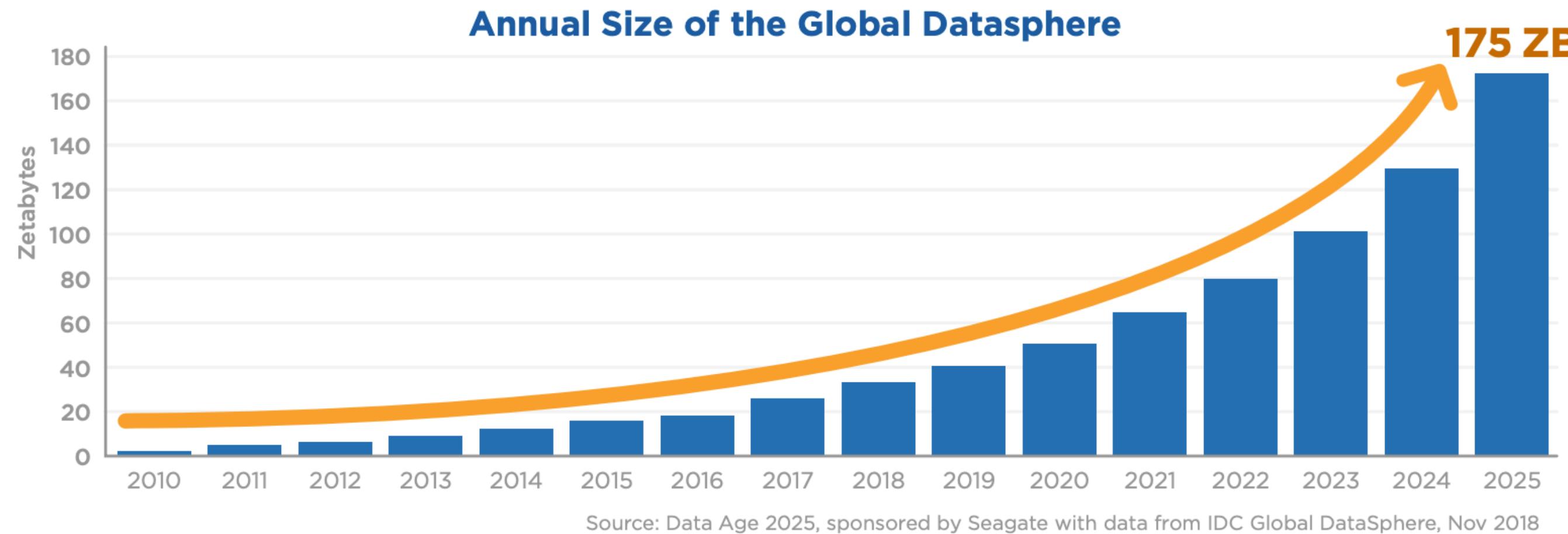
Organisation & Evaluation

DEFINITION & CONTEXT

AI & BIG DATA

Big Data and its 3 Vs:

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

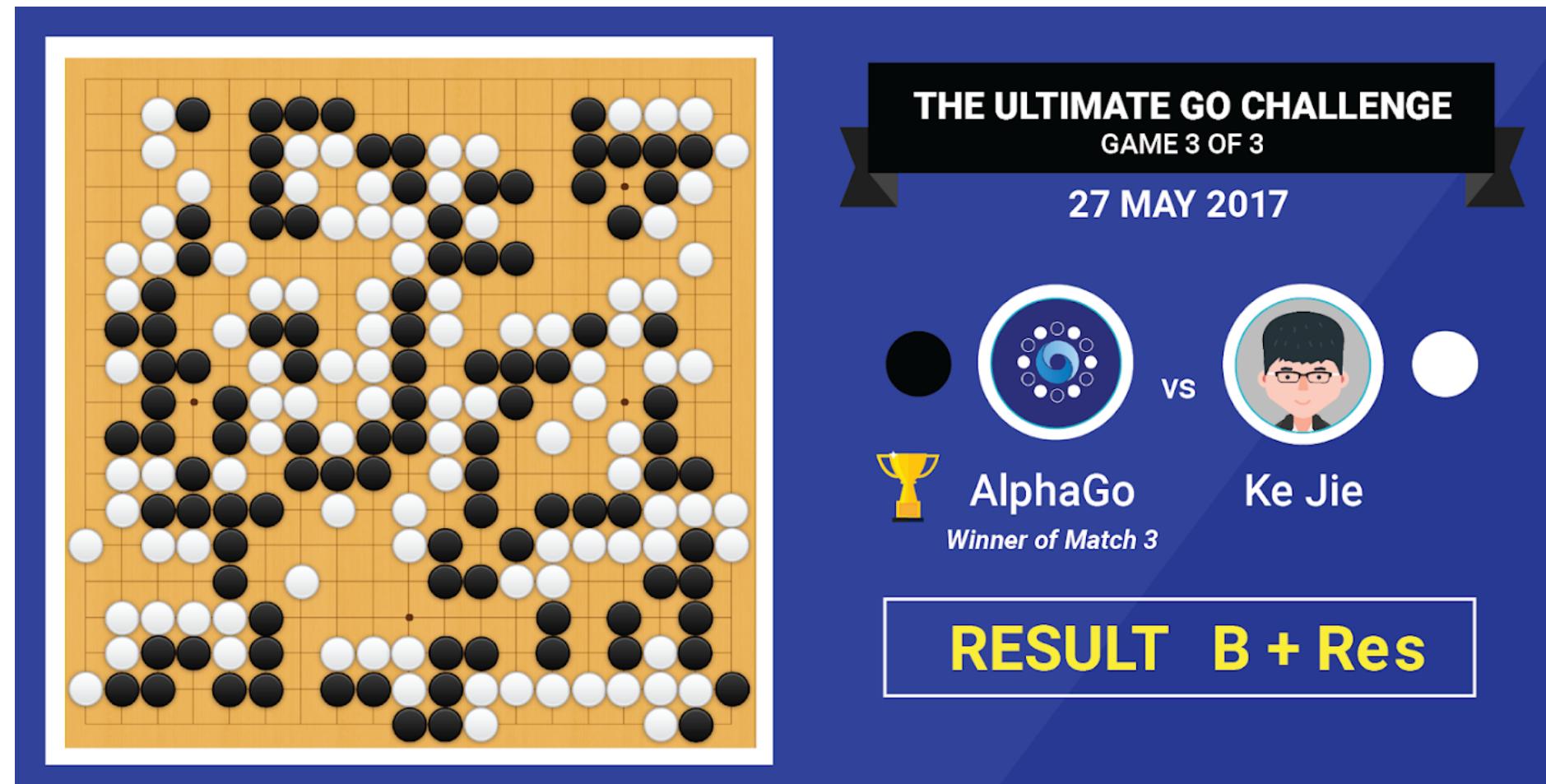


- **VARIETY:** numeric, signal, images, time series, graph, text etc...
- **Velocity:** real time decision
- **(VALORISATION)**

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

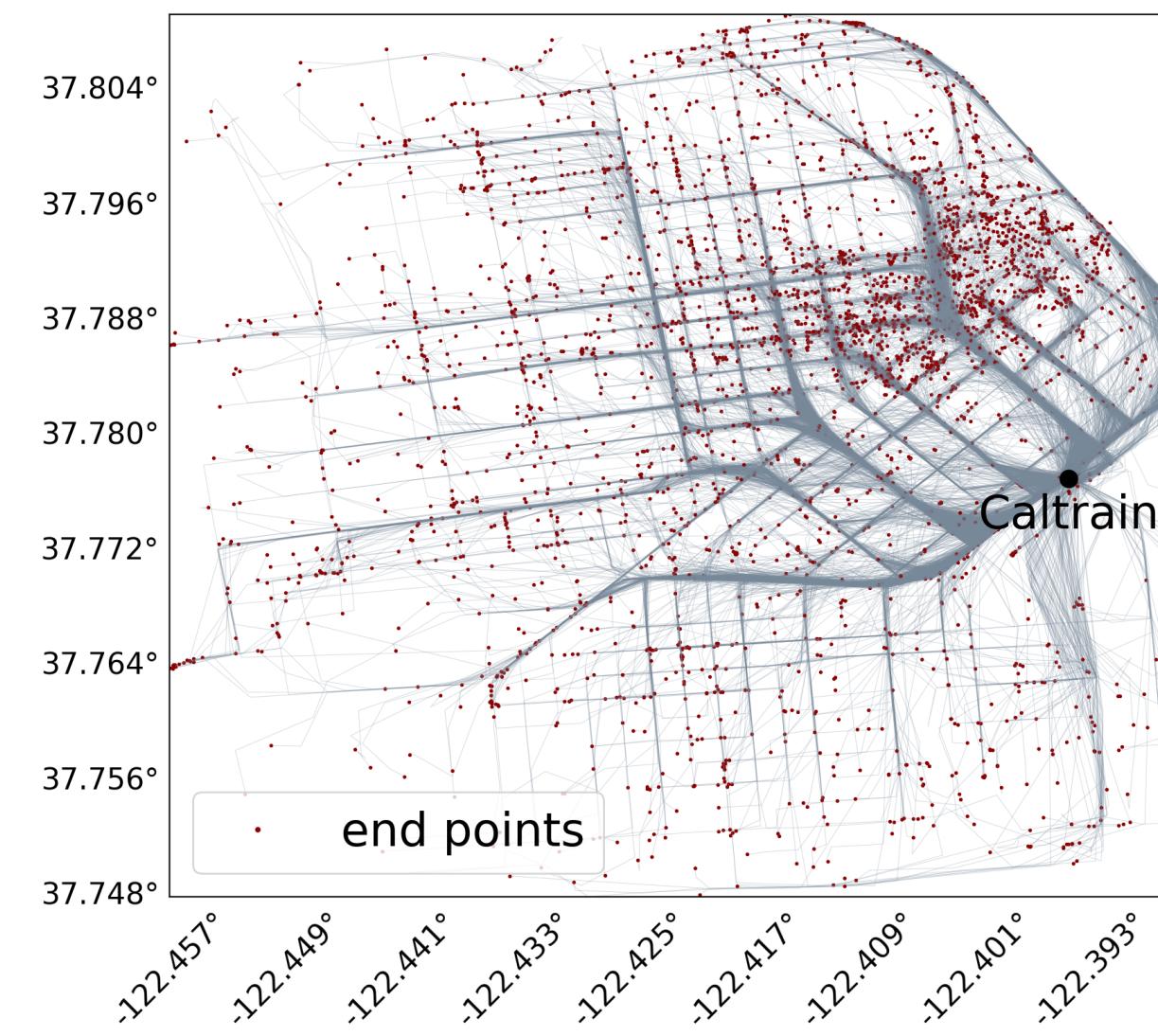
The definition is still very unclear, but IA is still NOT here.

A WIDE RANGE OF APPLICATION FIELDS

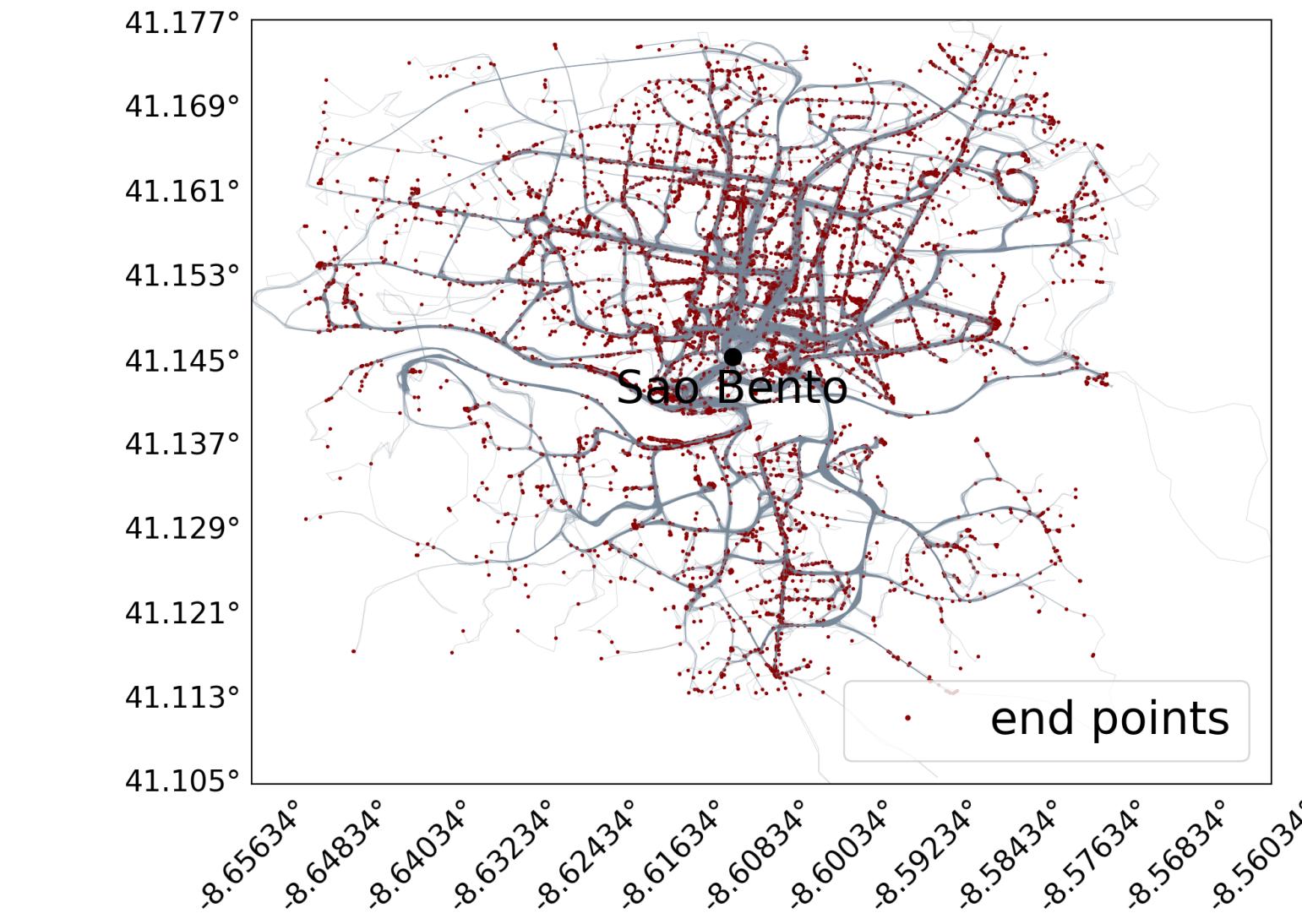
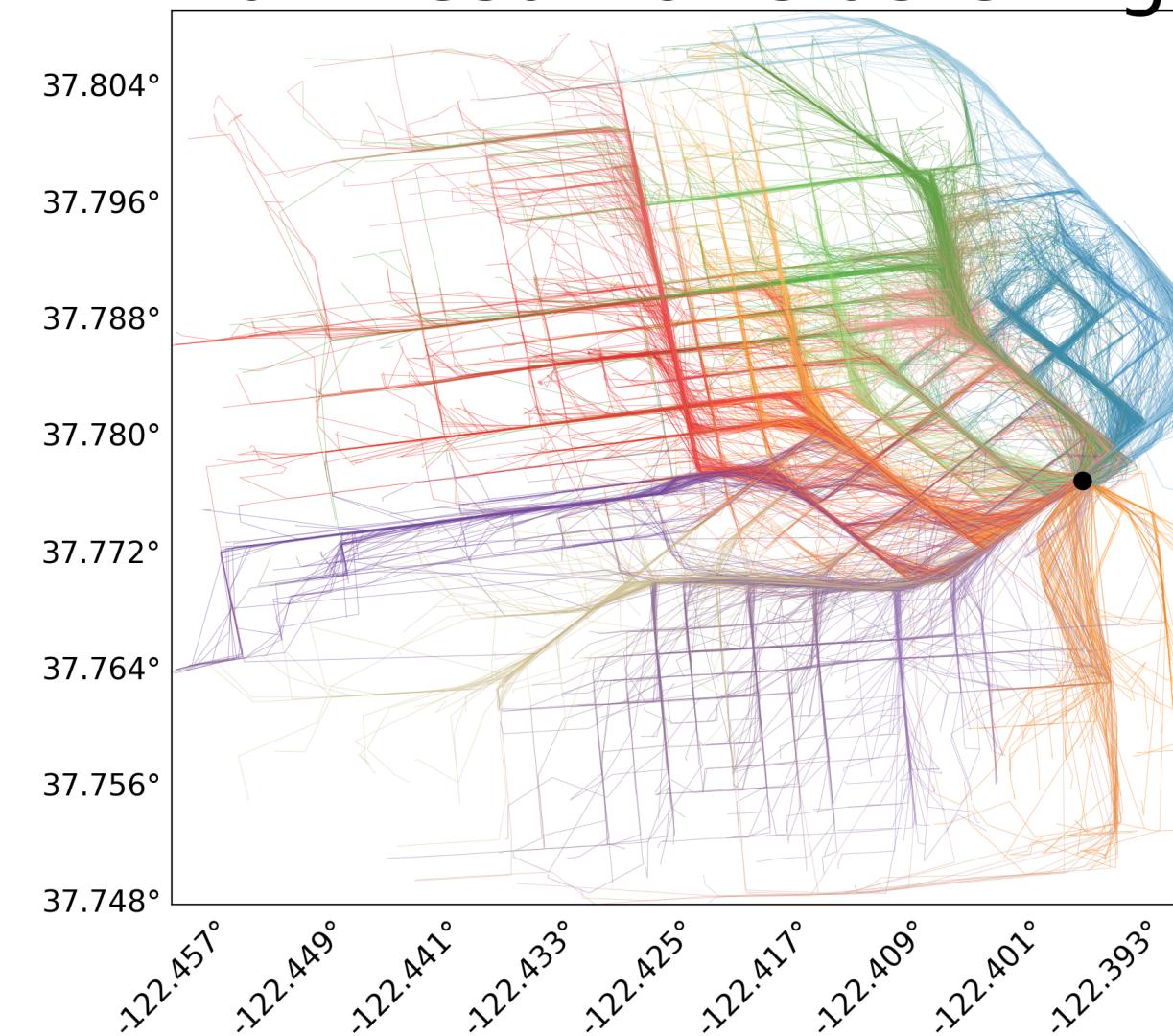
... at different levels of advancement.

- **E-COMMERCE:** Recommendation system, advertising, targeting.
 - **TEXT TRANSLATION:** DeepL, google translate, simplify, chatbot.
 - **IMAGE & VIDEO:** Classification, image generation, videosurveillance, video modification 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.**

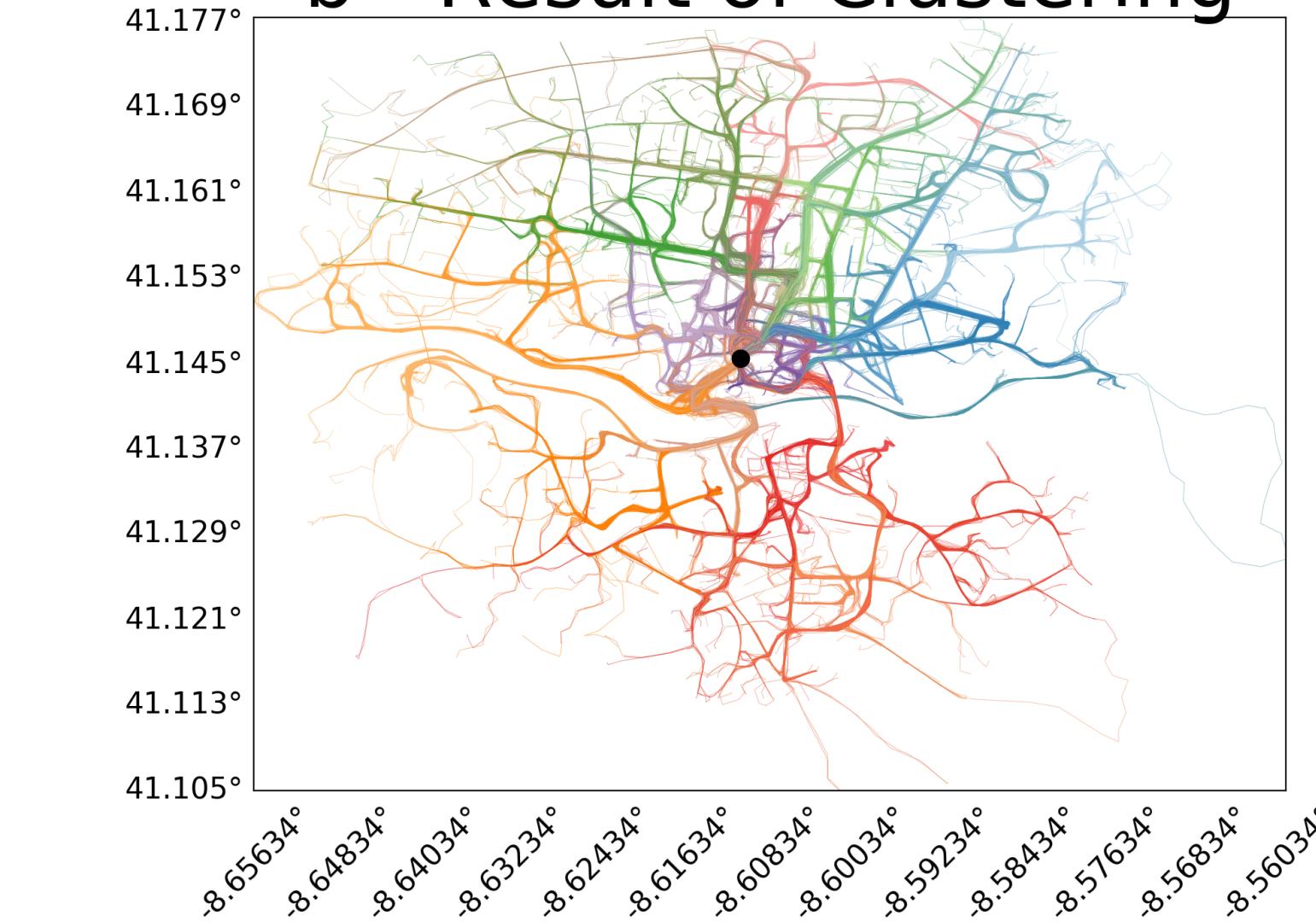
VARIETY: TRAJECTORY CLUSTERING FROM GPS DATA



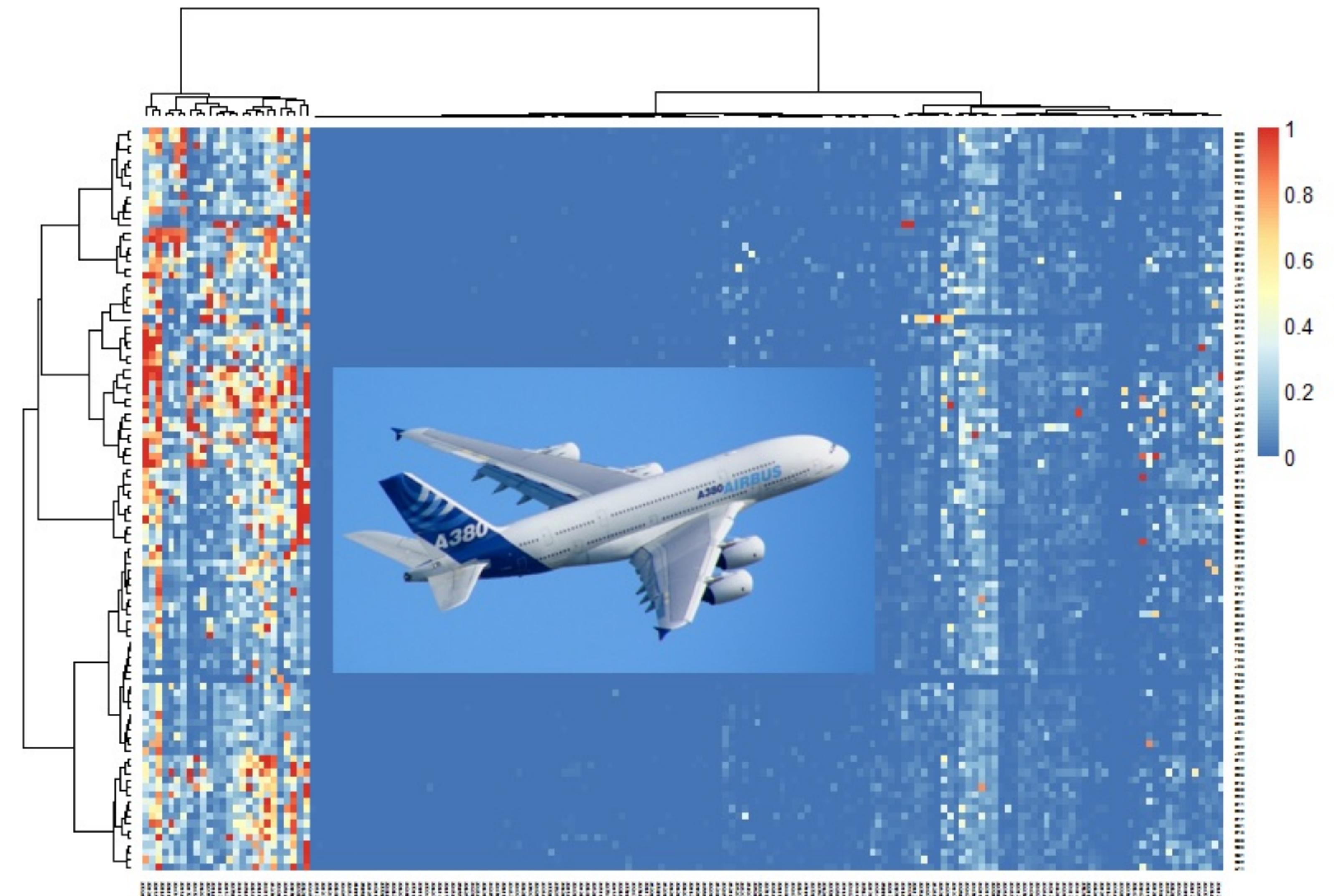
b - Result of Clustering



b - Result of Clustering



ANALYSIS OF FLIGHT INCIDENT MESSAGES



2019 INTERNSHIP OF INSA STUDENT

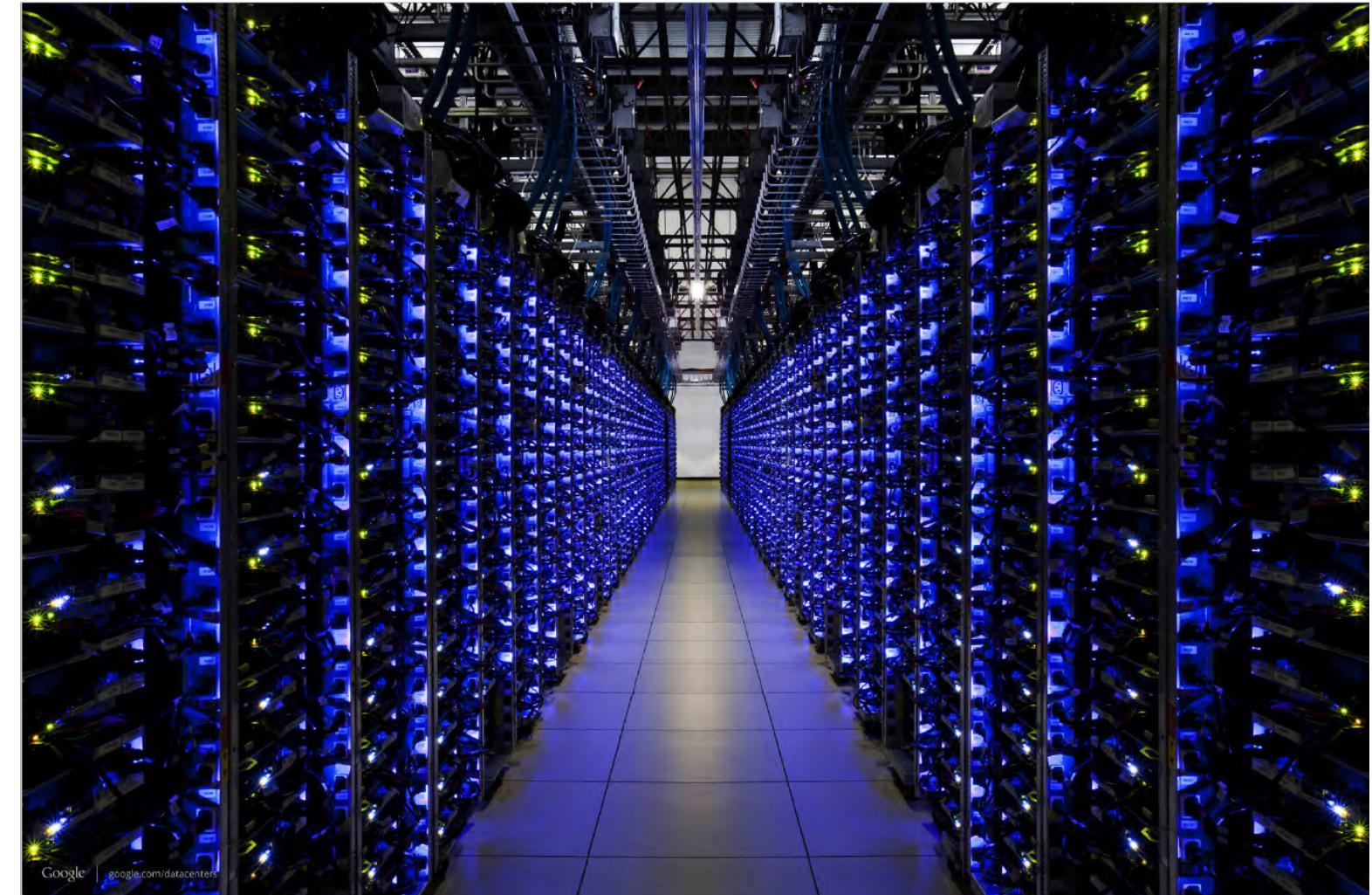
- Detection of field (agriculture) and building (military) from satellite images. *SSII & Startup*
- Anonymisation of name on text documents. *SSII*
- Fall detection of elder people. *PME*
- Autonomous drone flight. *SSII*
- Rain prediction in Ivory Coast. *Startup*
- Detection of diabetic retinopathy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

2020 INTERNSHIP OF INSA STUDENT.

- Image segmentation. *Big Company*
- Prediction road condition and traffic congestion. *Big Company*
- Change detection. *Research Center*
- Simulation of physics for plane conception. *Research Center*
- Object detection from radar data. *Big Company*
- Data anonymisation. *Startup*
- NLP on report. *Big Company*

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 do 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?

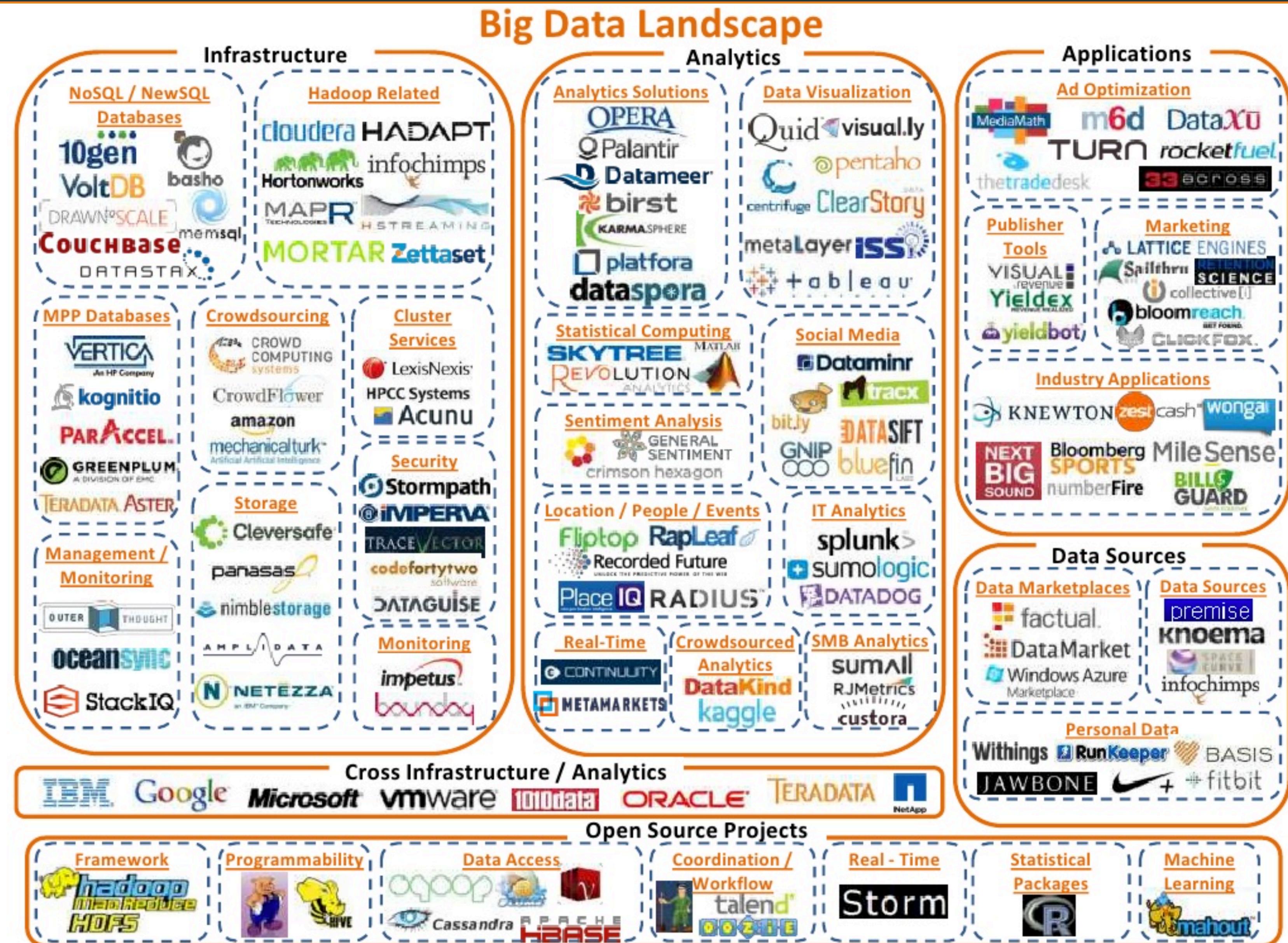


JOB SPECIFICATION

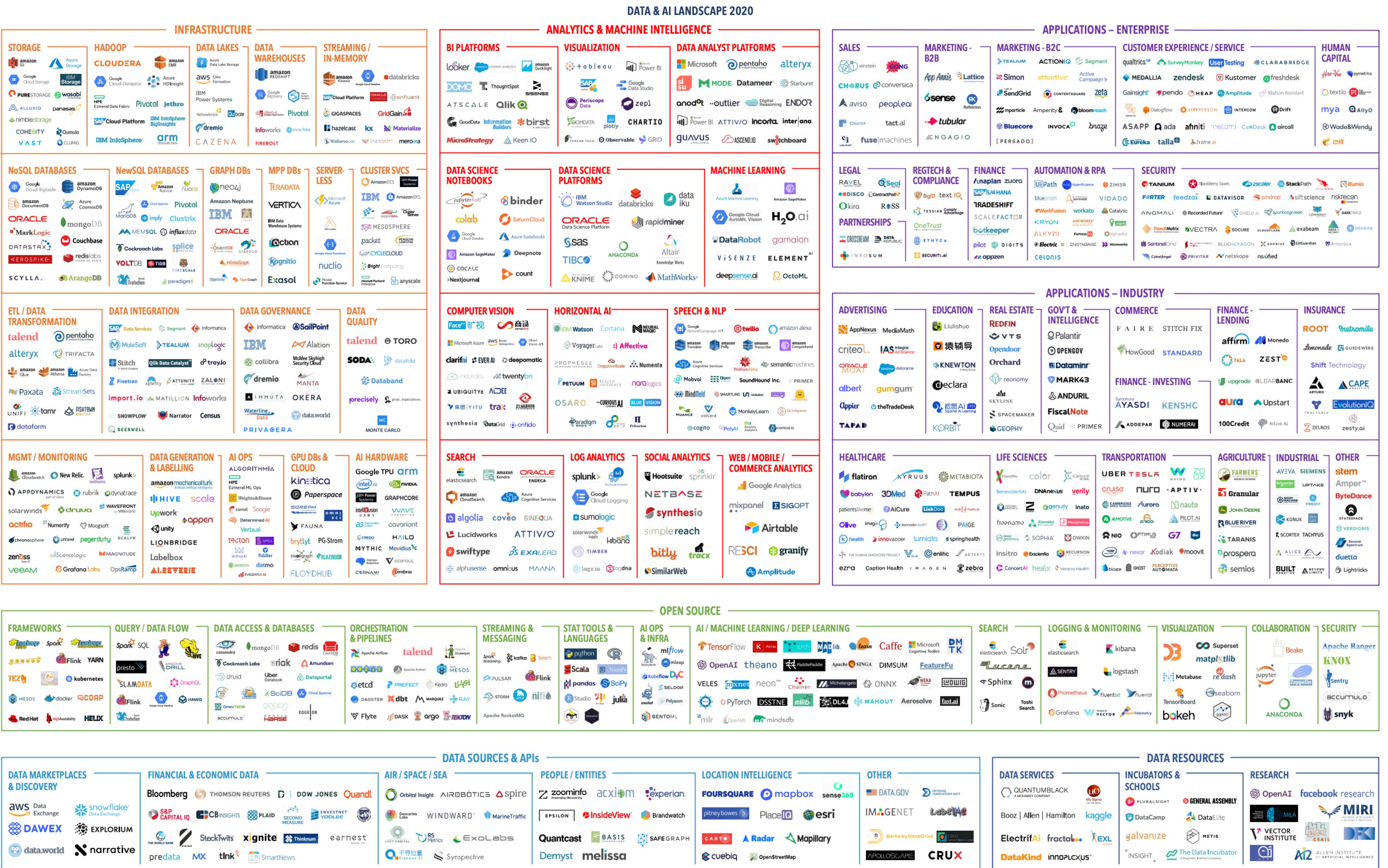
You are
here

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.	✗	✗	✓✓

BIG DATA & AI LANDSCAPE (TURCK, 2012)



BIG DATA & AI LANDSCAPE (TURCK, 2020)



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 (**H2O**), Microsoft, etc..

NEW MARKET: CLOUD COMPUTING (PICK-AND-SHOVEL PLAY)

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

(NEW ?) DATA SCIENCE

AS A **STATISTICIAN** :

- Deep comprehension of algorithms.
- Scientific watch.

WHAT'S REALLY NEW ?

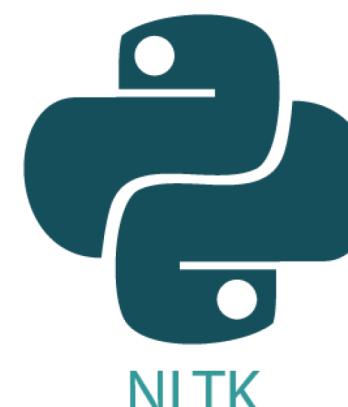
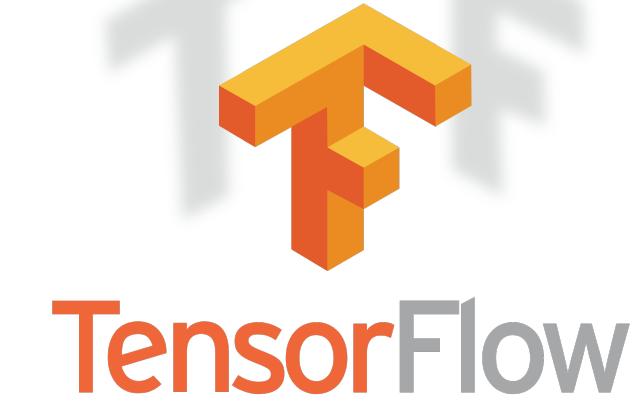
- More data, more algorithm.
- Programming skills
- Better understanding of all data-pipeline technology
- Manage bigger data and associate tools.
- New question about ethics.

Learn to “self-learn”

ORGANISATION & EVALUATION

OBJECTIVE - TOOLS

ML Python Libraries



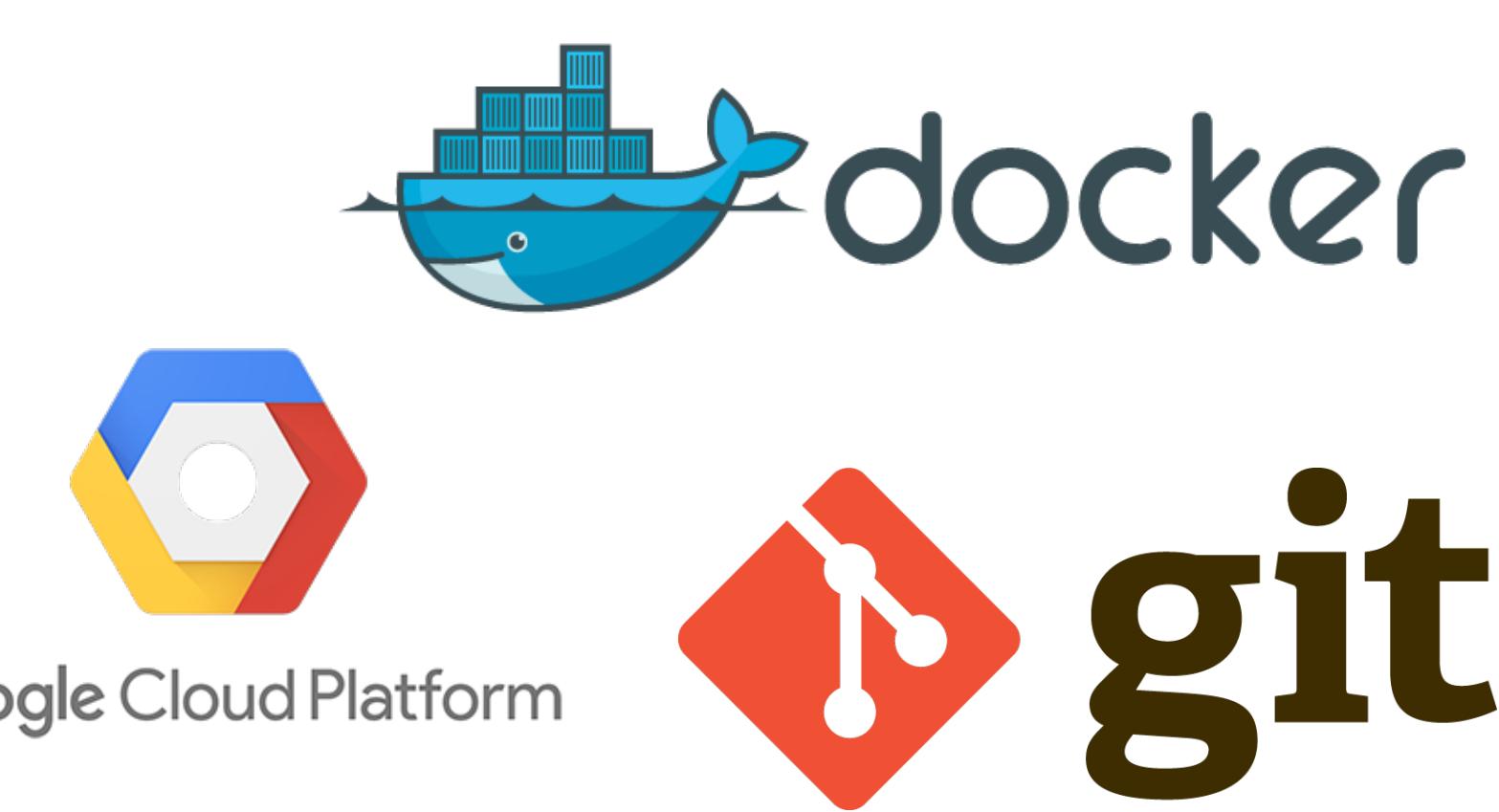
Python Environment



Viz' Python Libraries



Framework & Tool



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, GAN, VAE*

ARTIFICIAL INTELLIGENCE FRAMEWORKS

DATA

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

ALGORITHM

- Text cleaning : TFIDF, stemming
- 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.
- Code Development: Python environment + Python Script.

SESSION 3 - 30/11/20

- Code Development: Introduction to Google cloud computing.
- Code Development: Docker.

Video Course

TP : 9H-12H & 14H - 16H.

SESSION 4 - 07/12/20

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

SESSION 5 - 14/12/20

- Recommendation system.
- Introduction to DataVizualisation

SESSION 6 - 04/01/21 - Defi AI

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

RAPPORT (40%) - 10 PAGES MAXIMUM.

- Quality of the document - 25%.
- In-deep explanation of the chosen algorithm / features used, etc.. - 25%.
- Choice of the tools-infrastructure used - 25%.
- Results obtained - 25%.

Date/deadline: 2021-01-29

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: 2021-01-29