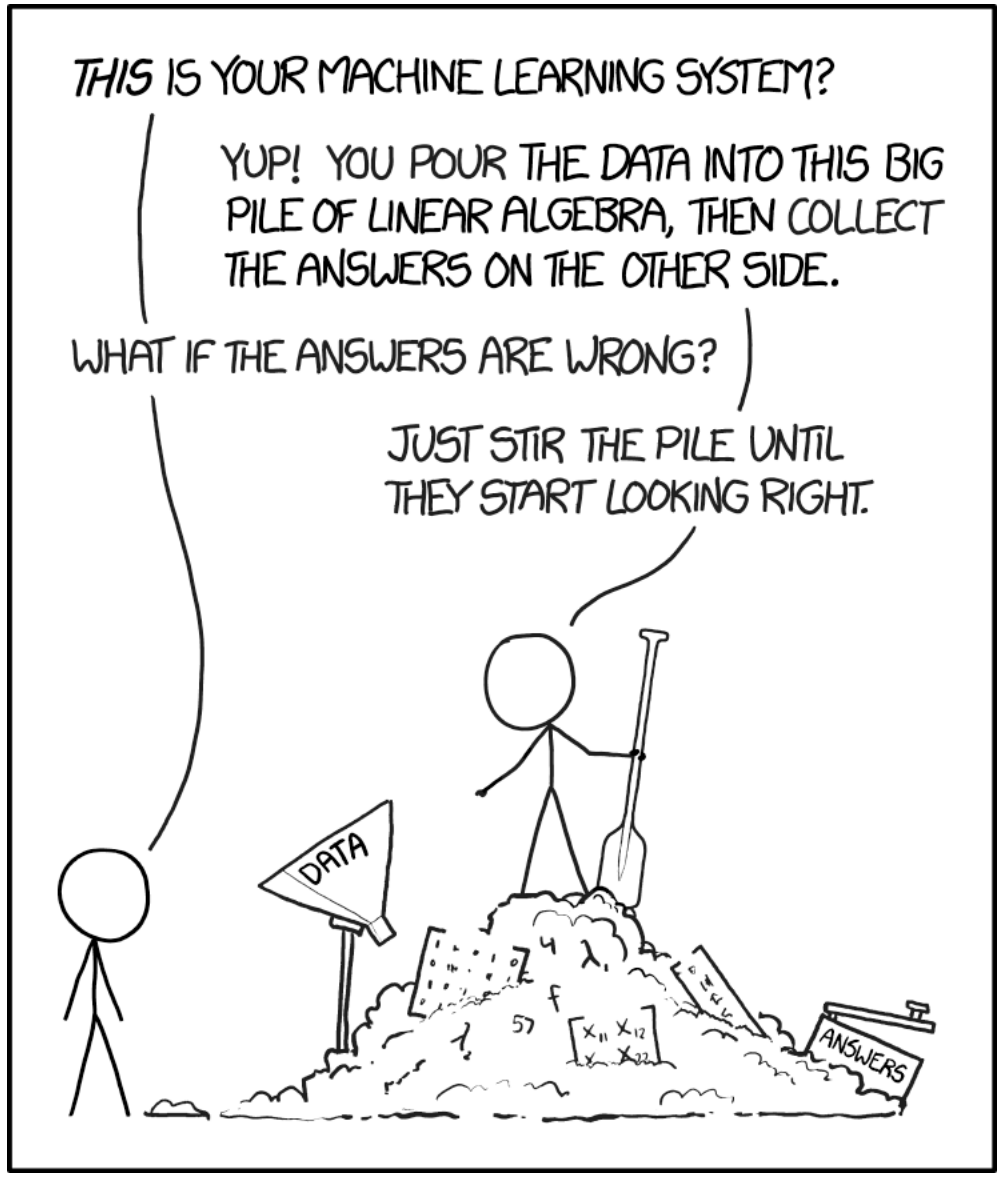


https://imgs.xkcd.com/comics/machine_learning.png

Présentations

Alexis Perrier

- Data Scientist
- PhD telecomParis 95' – Eric Moulines
- Paris - Boston (Berklee) – DC (freelance)
- Openclassrooms, UPEM, PCC @Polytechnique

alexis.perrier@pm.me
linkedin.com/in/alexisperrier
twitter: @alexip

Programme Data Science 2019 -2020

Semaine	quoi	qui
23/09 au 27/09	stats et regression	AP
30/09 au 04/10	analyse predictive supervisée	AP
07/10 au 11/10	non supervisé et intro au deep learning	Laurent Risser
14/10 au 18/10	SVM – machines a vecteurs de support	Alice Martin
25/11 au 29/11	NLP	AP
16/12 au 20/12	Deep learning et datacamp	Chafik Samir et Eric Moulines
27/01 au 31/01	Méthodes statistiques pour les expériences simulées	Fabrice Gamboa
24/02 au 28/02	Analyse bayésienne et la programmation probabiliste	AP

Des stats au machine learning

L'approche statistique – 3 jours

Python
- jupyter notebook, colab, kaggle kernel, anaconda
- librairies: numpy, scipy, statsmodel, scikit-learn, pandas

- Linéarité et corrélation
- tests statistiques
- régression lineaire univariée, multivariée
- classification, régression logistique
- interpretation, diagnostique
- hypothèses de la régression linéaire
- bases mathématiques

L'approche machine learning – 5 jours

Analyse prédictive et performance
- GLM & régression polynomiale
- décomposition biais-variance
- overfitting: détection et solutions
    - Gradient stochastique
    - regularization
    - cross validation

- Arbres de decisions
- ensembling / bagging
- random forests
- feature engineering
- imbalanced dataset

Boosting => Kaggle
- XGBoost
- LightGBM
- CatBoost

Questions ?

Tour de table

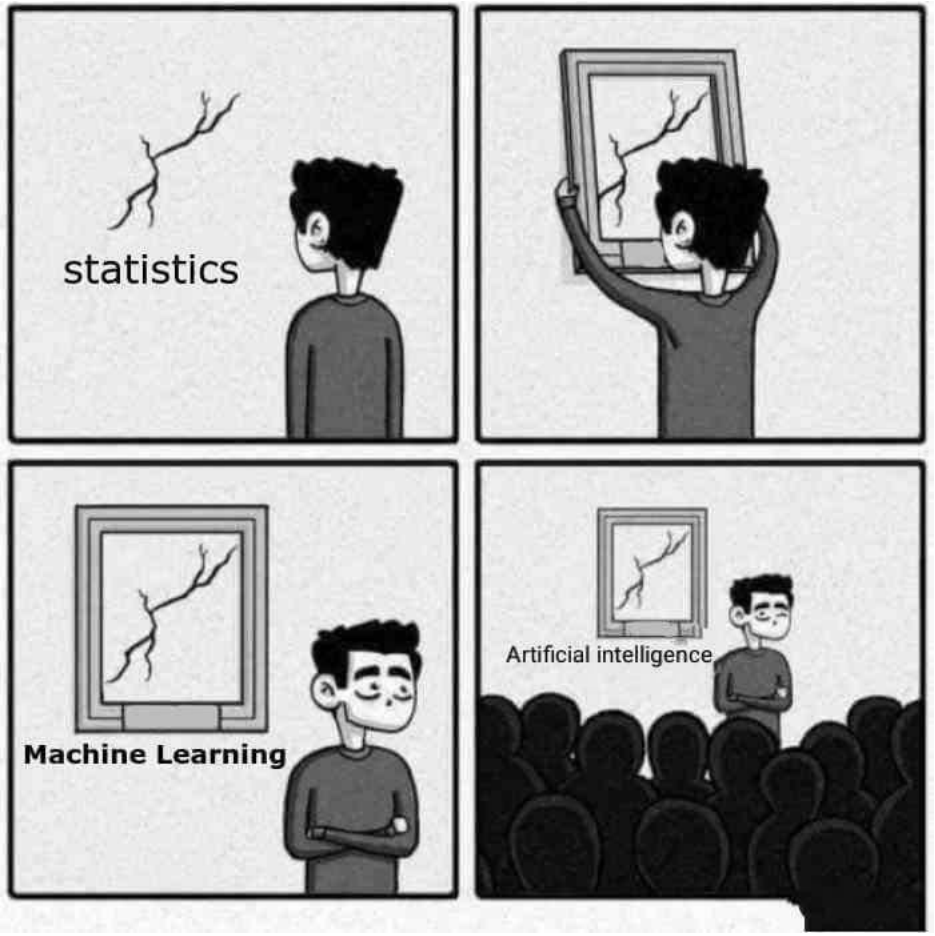
- C'est quoi le data science ?
- Pourquoi avoir choisi cette option ?
- Niveau en stats ?
- Niveau en python ?
- Github ?
- Manque t il des choses dans ce programme ?
- twitter ? kaggle ? insta ? linkedin ?
- laptop ?
- **Votre question**

alexis.perrier@pm.me
linkedin.com/in/alexisperrier
twitter: @alexip

Outils

- Moodle
- github
- Slack ?

data science - machine learning - predictive analytics - intelligence artificielle - deep learning – statistiques ?



Baron Schwartz

@xaprb

Follow

When you’re fundraising, it’s AI
When you’re hiring, it’s ML
When you’re implementing, it’s linear regression
When you’re debugging, it’s printf()

12:52 AM - 15 Nov 2017

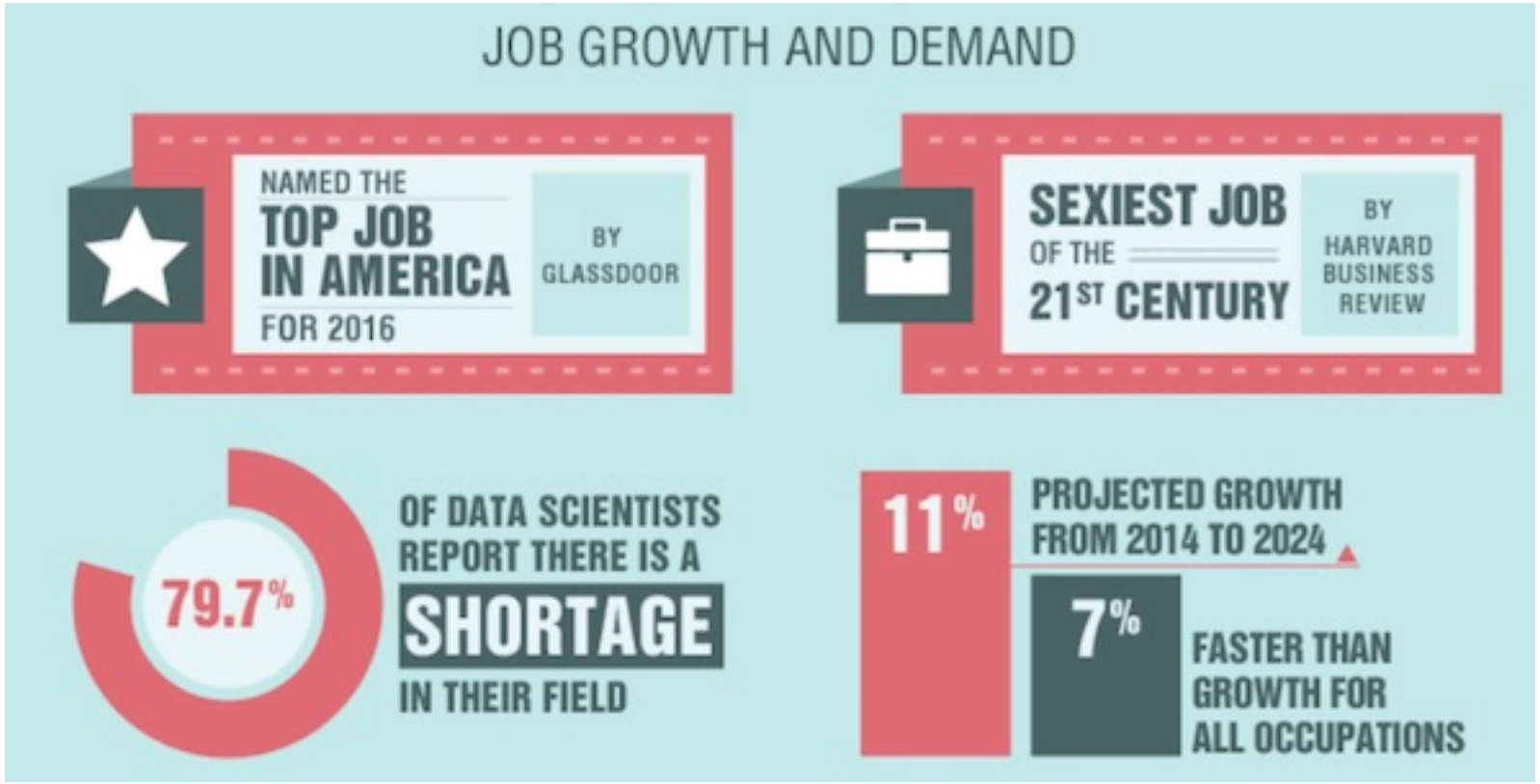
5,595 Retweets **12,717** Likes

93 5.6K 13K

Tweet your reply

Don't believe the hype?

"Data scientist is the sexiest job of the 21st century."
Harvard Business Review



Application sur des données existantes

- **Data Analysis, Data Mining** : Exploration, trouver les tendances, les evolutions, les anomalies, etudier les corrélations => data visualization, analyse non supervisée

- **Statistiques** : Trouver le modèle qui explique au mieux toutes les données

Application a des données nouvelles (unseen data)

- **Machine learning** : Le modèle apprend automatiquement à partir des données. apprentissage automatique
- **Analyse prédictive**: Construire ou entrainer des modèles qui peuvent *"prédire"* sur des nouvelles données à partir de données existantes.
- **Deep Learning** : Analyse prédictive supervisée avec des réseaux de neurones
- Autres variantes: Transfer learning, reinforcement learning, …

Data Science:

- Comprend toutes les techniques connues pour extraire de l'info des données

What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?

2001 Leo Breiman, Berkeley, publishes "Statistical Modeling: The Two Cultures":

"There are two cultures in the use of statistical modeling to reach conclusions from data.

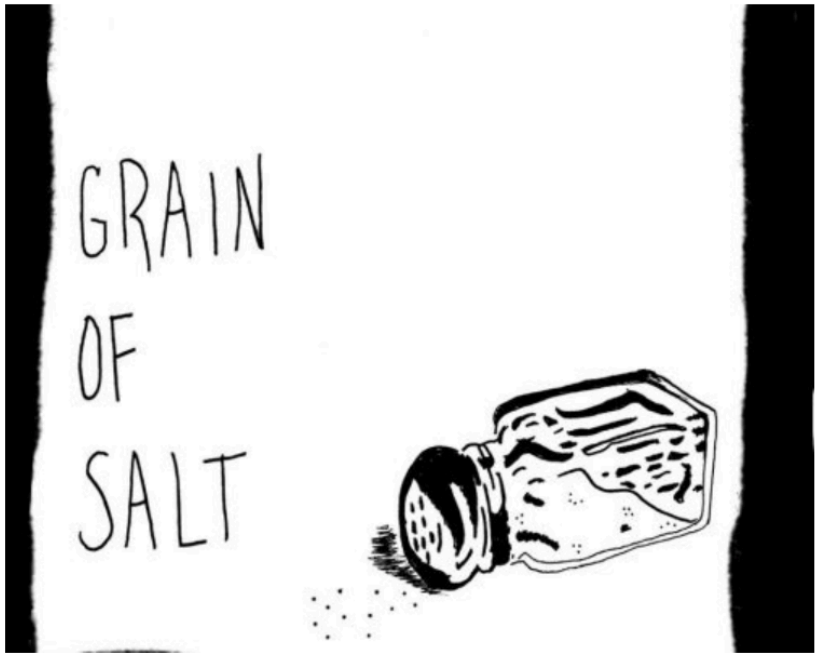
One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.

The statistical community has been committed to the almost exclusive use of data models.

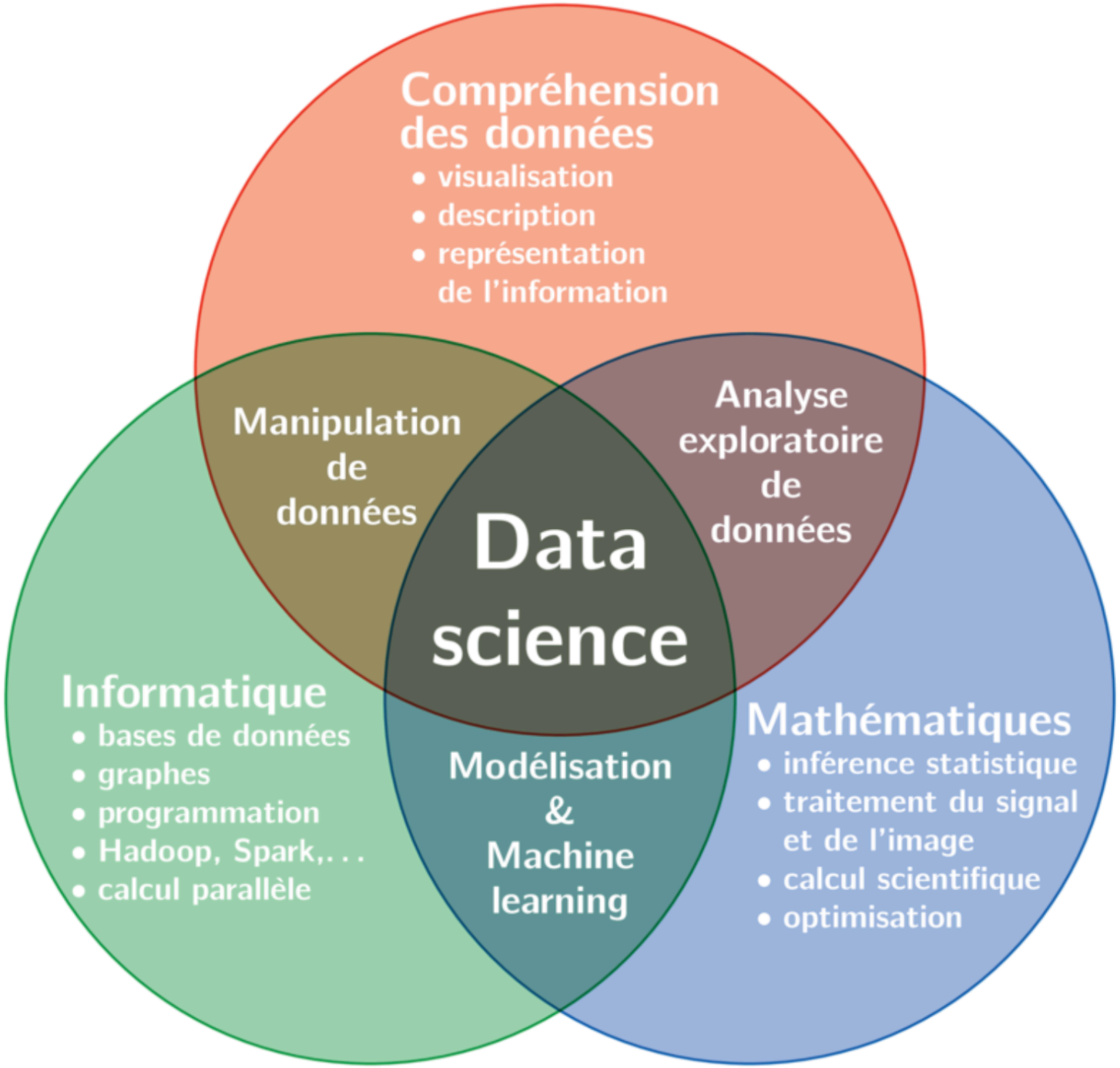
This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

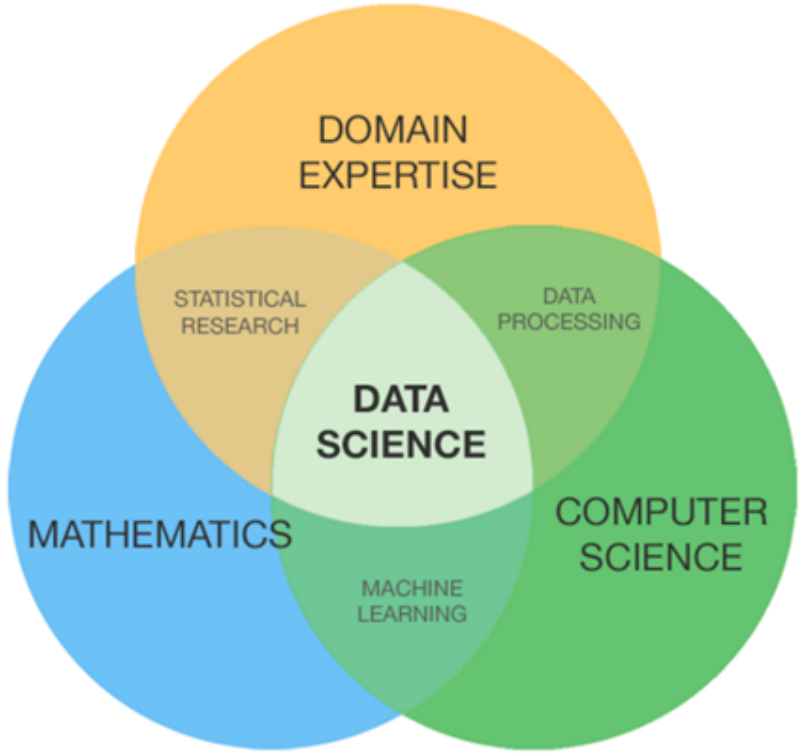
Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets.

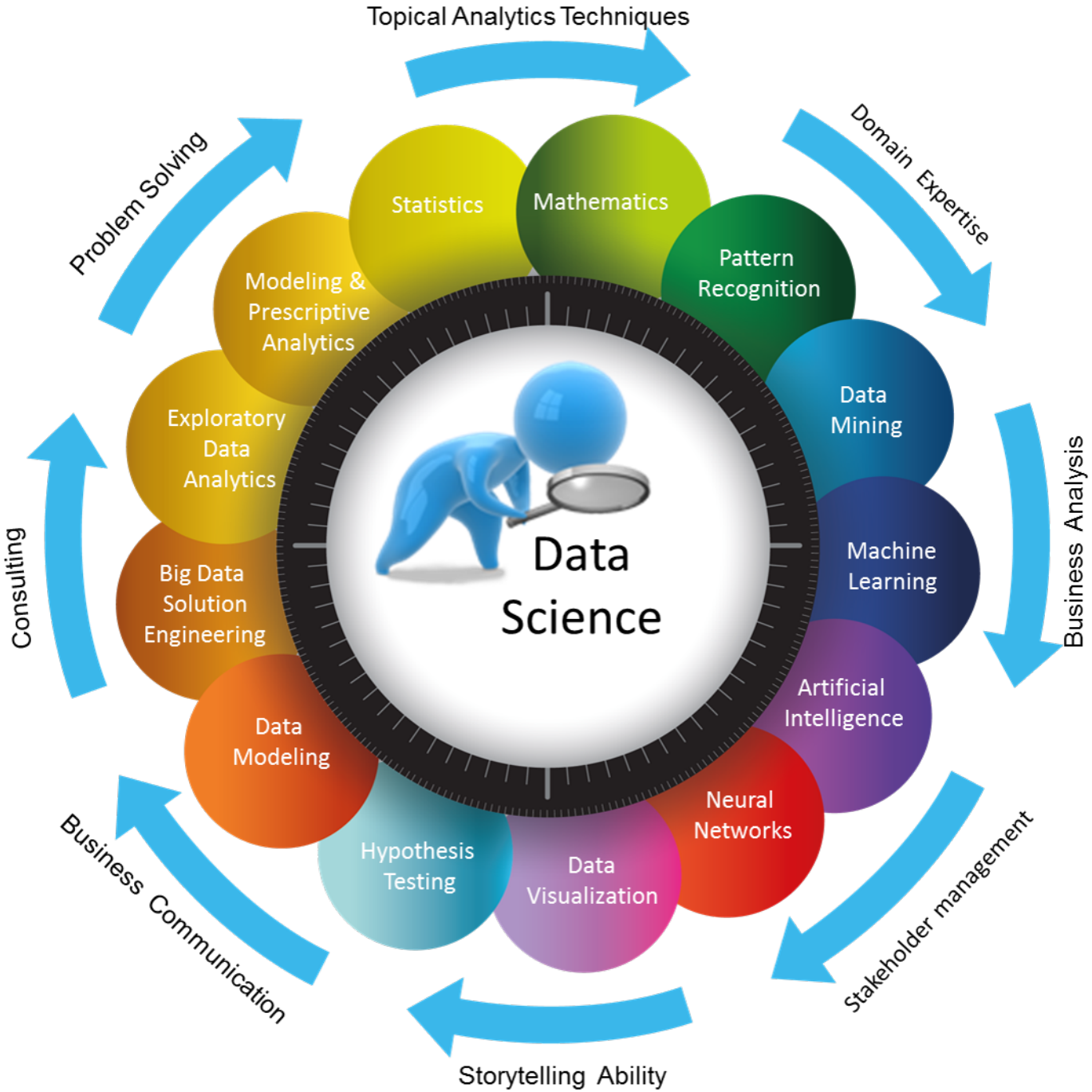
If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools."



la Data Science c'est quoi finallement?


Compréhension des données
• visualisation
• description
• représentation de l'information
Manipulation de données
Analyse exploratoire de données
Data science
Informatique
• bases de données
• graphes
• programmation
• Hadoop, Spark,...
• calcul parallèle
Modélisation & Machine learning
Mathématiques
• inférence statistique
• traitement du signal et de l'image
• calcul scientifique
• optimisation


DOMAIN EXPERTISE
STATISTICAL RESEARCH
DATA PROCESSING
DATA SCIENCE
MATHEMATICS
COMPUTER SCIENCE
MACHINE LEARNING


Topical Analytics Techniques
Domain Expertise
Business Analysis
Stakeholder management
Storytelling Ability
Business Communication
Consulting
Problem Solving
Statistics
Mathematics
Pattern Recognition
Data Mining
Machine Learning
Artificial Intelligence
Neural Networks
Data Visualization
Hypothesis Testing
Data Modeling
Big Data Solution Engineering
Exploratory Data Analytics
Modeling & Prescriptive Analytics
Data Science

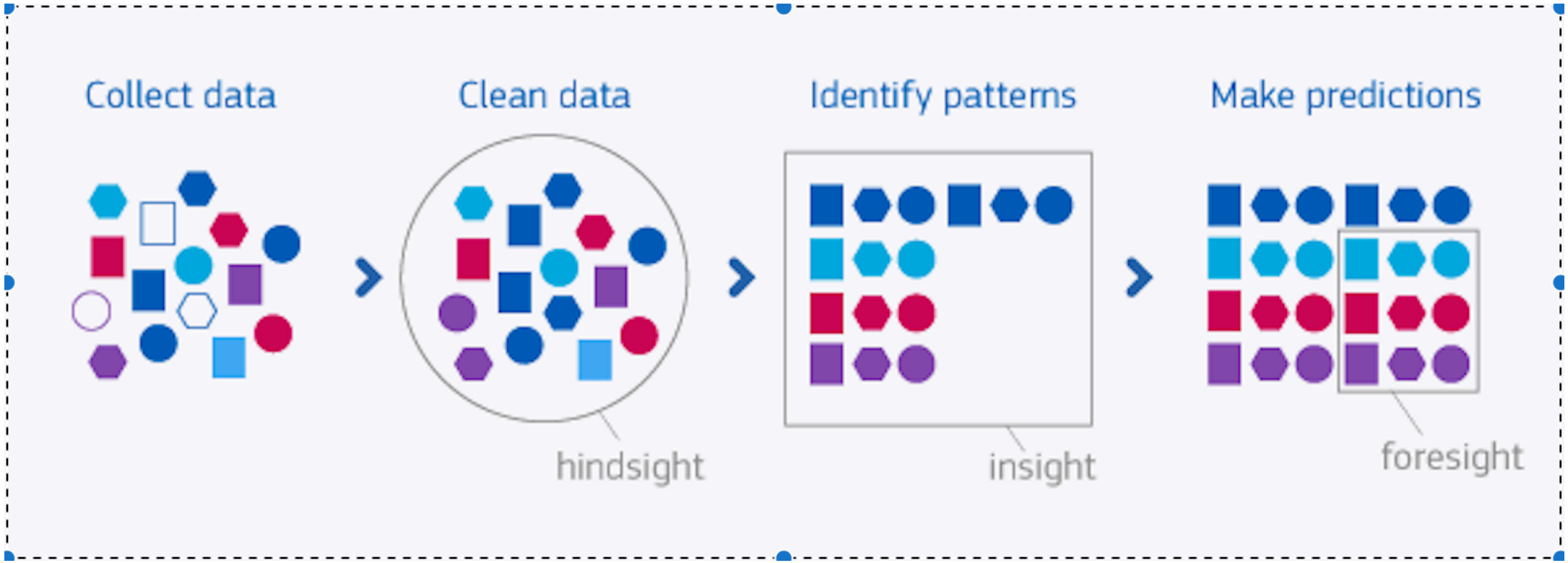
CHAMPS D'APPLICATIONS

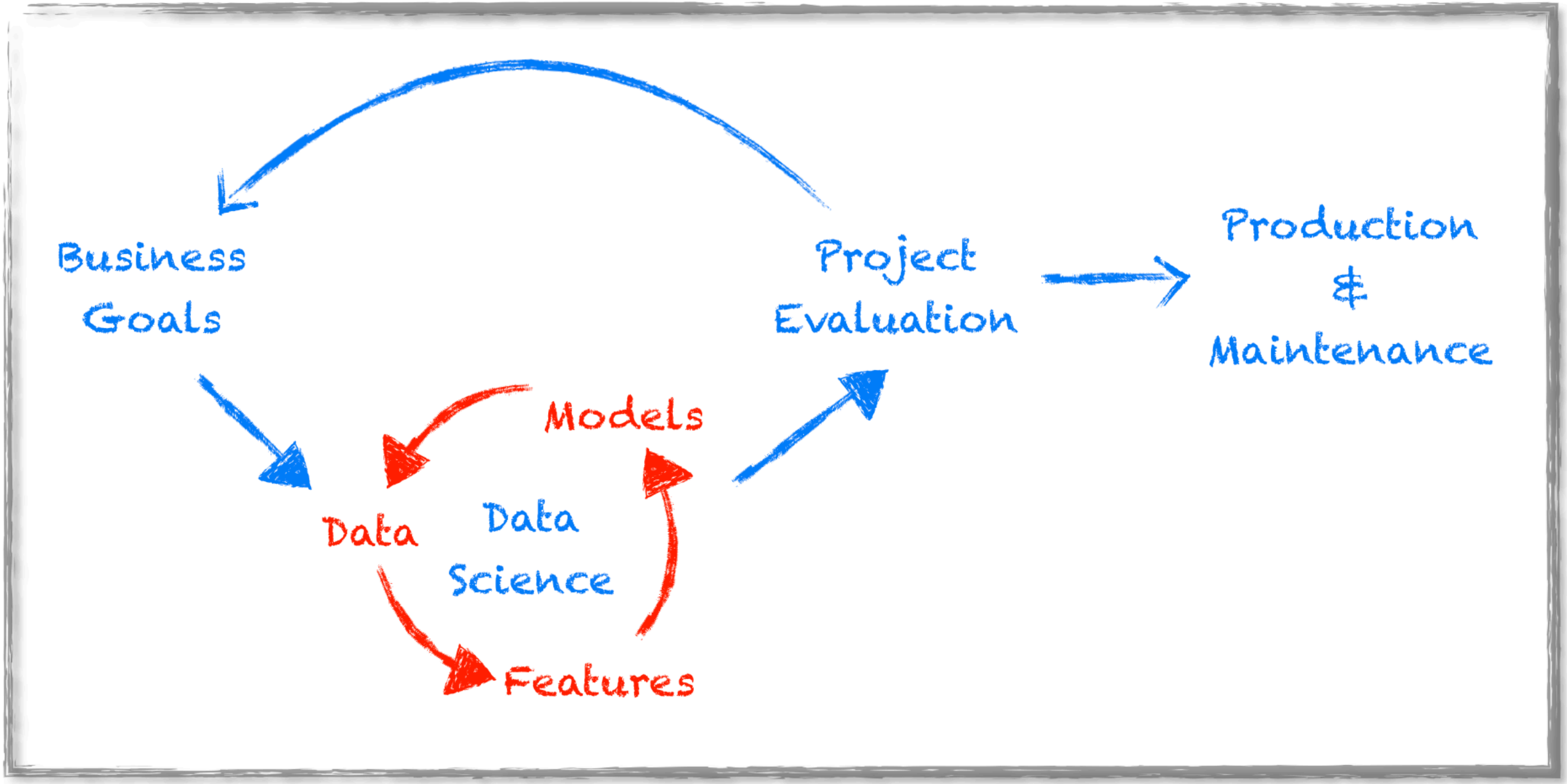
- **Predictions**: market, demand, supply prices, population, weather, earthquakes, ...
- **Patterns**: customer behavior patterns
- **Detection**: Spam, Fraud, Failures, Cyber attacks
- **Extracting meaning** from large sets of data: handwritten health records, exoplanets
- **NLP**: translation, speech to text, speech recognition, sentiment analysis, topic modeling, spell checking
- **Recommender systems**: Netflix, Spotify, Amazon
- **Ranking systems**: search results
- **Autonomous systems** (reinforcement learning / AI): playing games, self driving cars, drones
- **Time series**: algorithmic trading, signal processing, IoT, sales forecast
- **Image / Video**: automatic captionning, face and object recognition, ...

deepindex

Questions?

Data science workflow


Collect data
Clean data
Identify patterns
Make predictions
hindsight
insight
foresight


Business Goals
Project Evaluation
Production & Maintenance
Models
Data
Data Science
Features

Les etapes d'un projet de Data Science

A) LES DONNÉES

1) Définir le problème
2) ETL: Extraction Transform Load
3) Travailler sur les variables

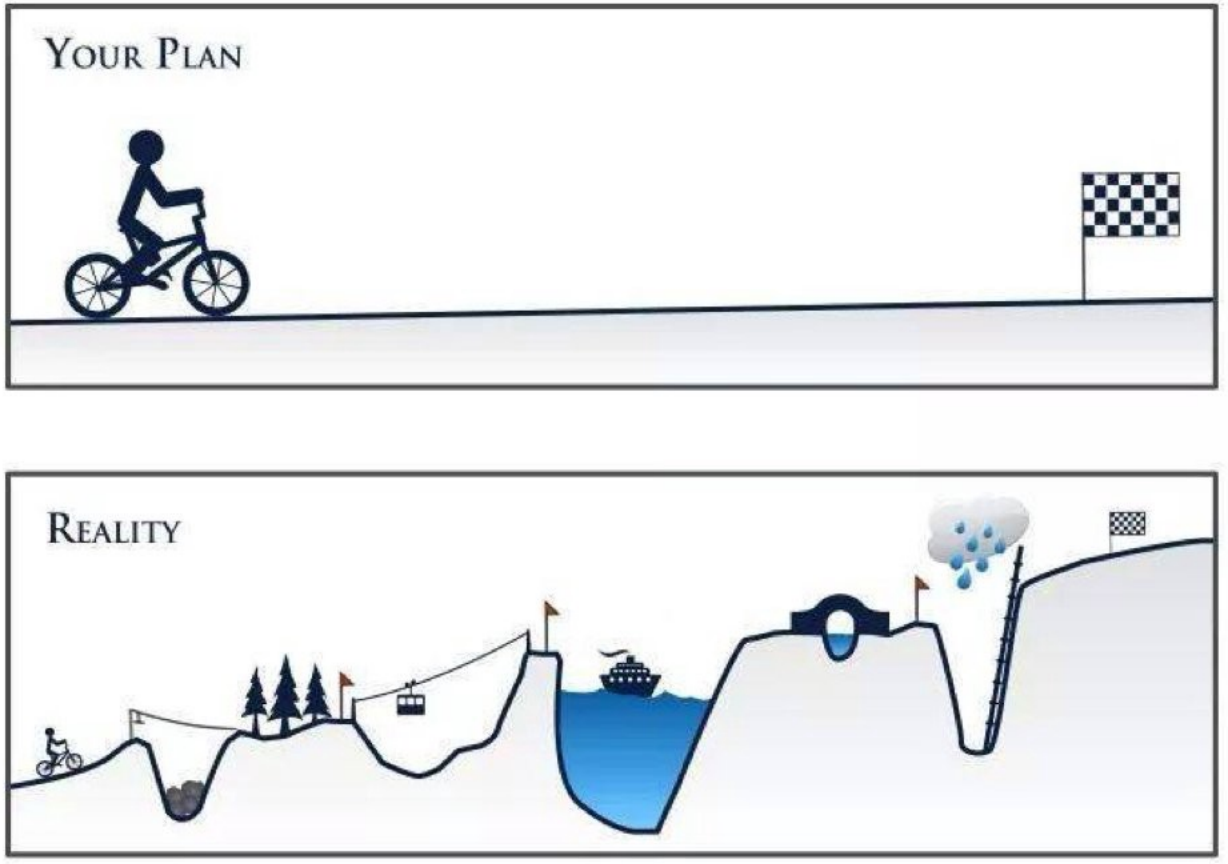
B) MACHINE LEARNING

4) Outils et plateforme
5) Modélisation
6) Le test des nouvelles données

C) NOUVELLE ITÉRATION

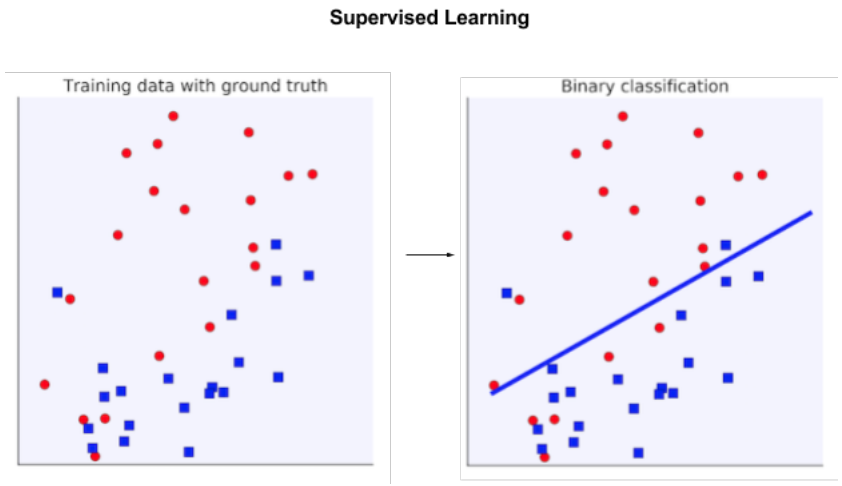
7) Présentation des résultats
8) reprendre le problème

D) MISE EN PRODUCTION



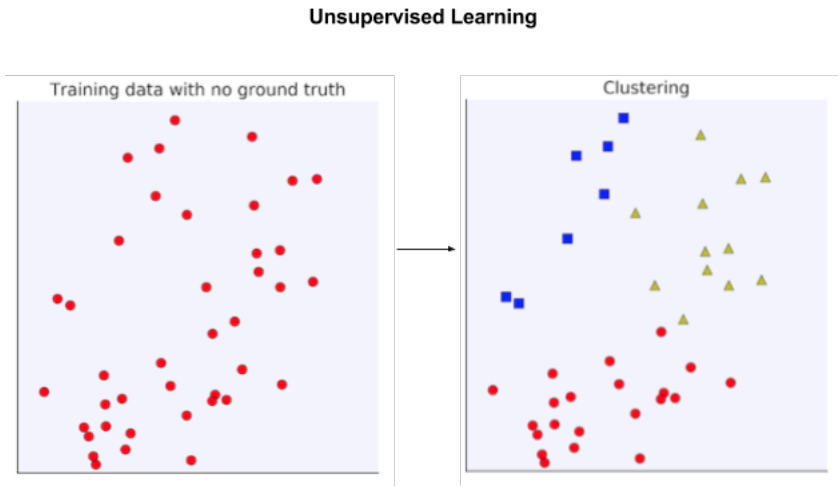
SUPERVISÉE

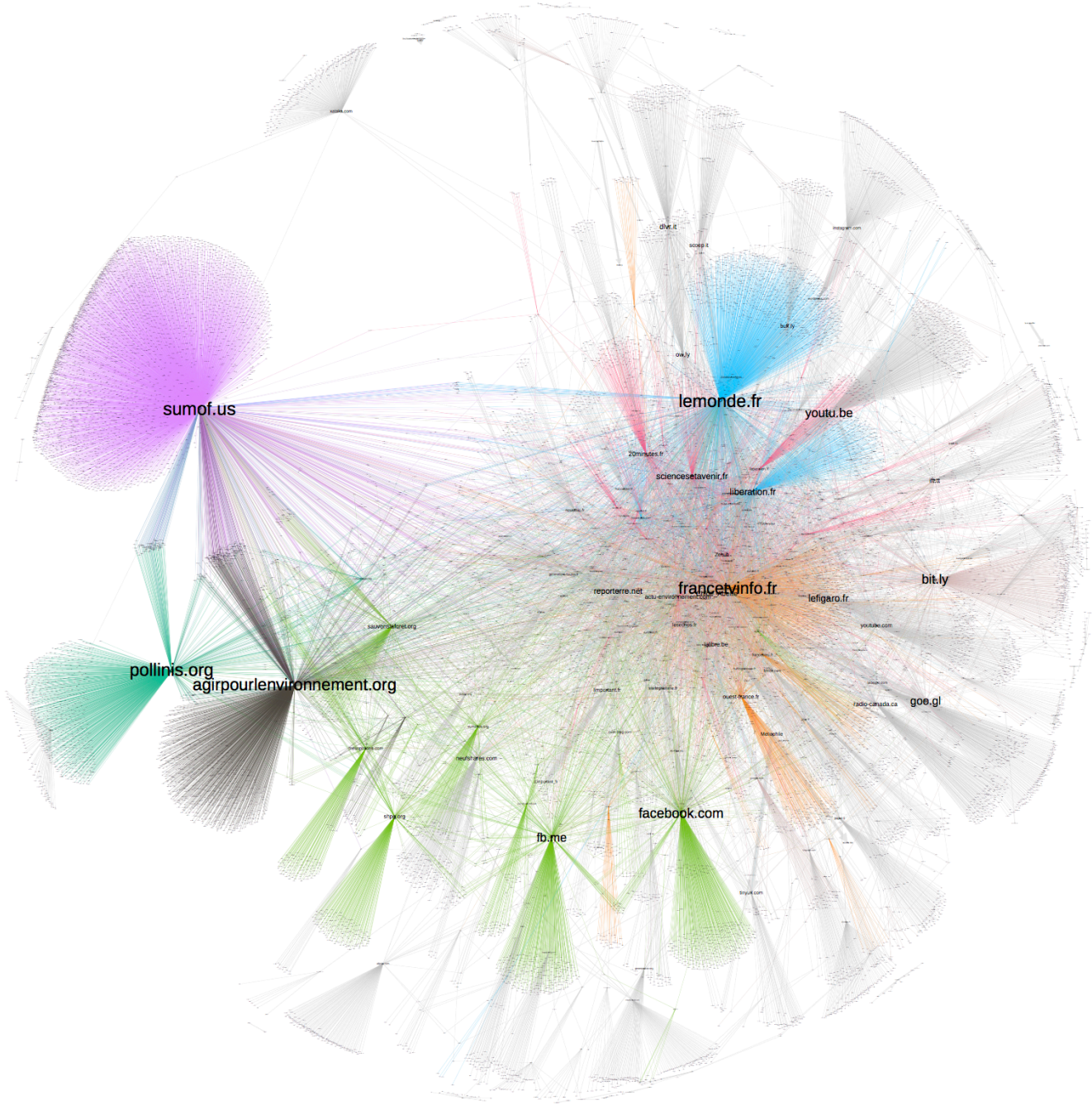
Le dataset d'apprentissage inclut la variable à prédire [cible].
On a un certain nombres d'exemples sur lesquels on peut entrainer un modèsle
logique de scoring, de classification et de prediction Random forest, Regression lineaire ou logistique, SVM, ...
Classification: On connait le nombre de classes



Non SUPERVISÉE

Le dataset d'apprentissage n'inclut pas de variable cible. Il n'y a pas de **ground truth**
logique de clustering, de classification automatique des échantillons sans connaitre a priori le nombre de classes
notion de similarité et de distance entre les échantillons
K-means, K-NN, ..




sumof.us
lemonde.fr
youtu.be
20minutes.fr
sciencesetavenir.fr
liberation.fr
reporterre.net
francetvinfo.fr
lefigaro.fr
bit.ly
youtube.com
pollinis.org
agirpourlenvironnement.org
goo.gl
radio-canada.ca
facebook.com
fb.me

REGRESSION La variable cible est continue

Age, taille, poids,
nombre d'appels, de clicks, volume de vente, consommation
Température, Salaire, ... Probabilité d'une action Temps, délai, retard

CLASSIFICATION

La variable cible est discrete, une catégorie, une classe

CAS BINAIRE

Achat, resiliation, click
Survie, maladie, succes examen, admission, Positif ou negatif
Spam, fraude

MULTI CLASS - MULTINOMIALE

Catégories, types (A,B,C),
Positif, neutre ou negatif Espèces de plantes d'animaux, ... Pays, planetes

ORDINALE

Notes, satisfaction, ranking

REGRESSION À CLASSIFICATION

discrétiser la variable
Age =>
0-12 12-24 25-49 50-65 plus de 65

CLASSIFICATION À REGRESSION

Prédire une probabilité au lieu d'une classe

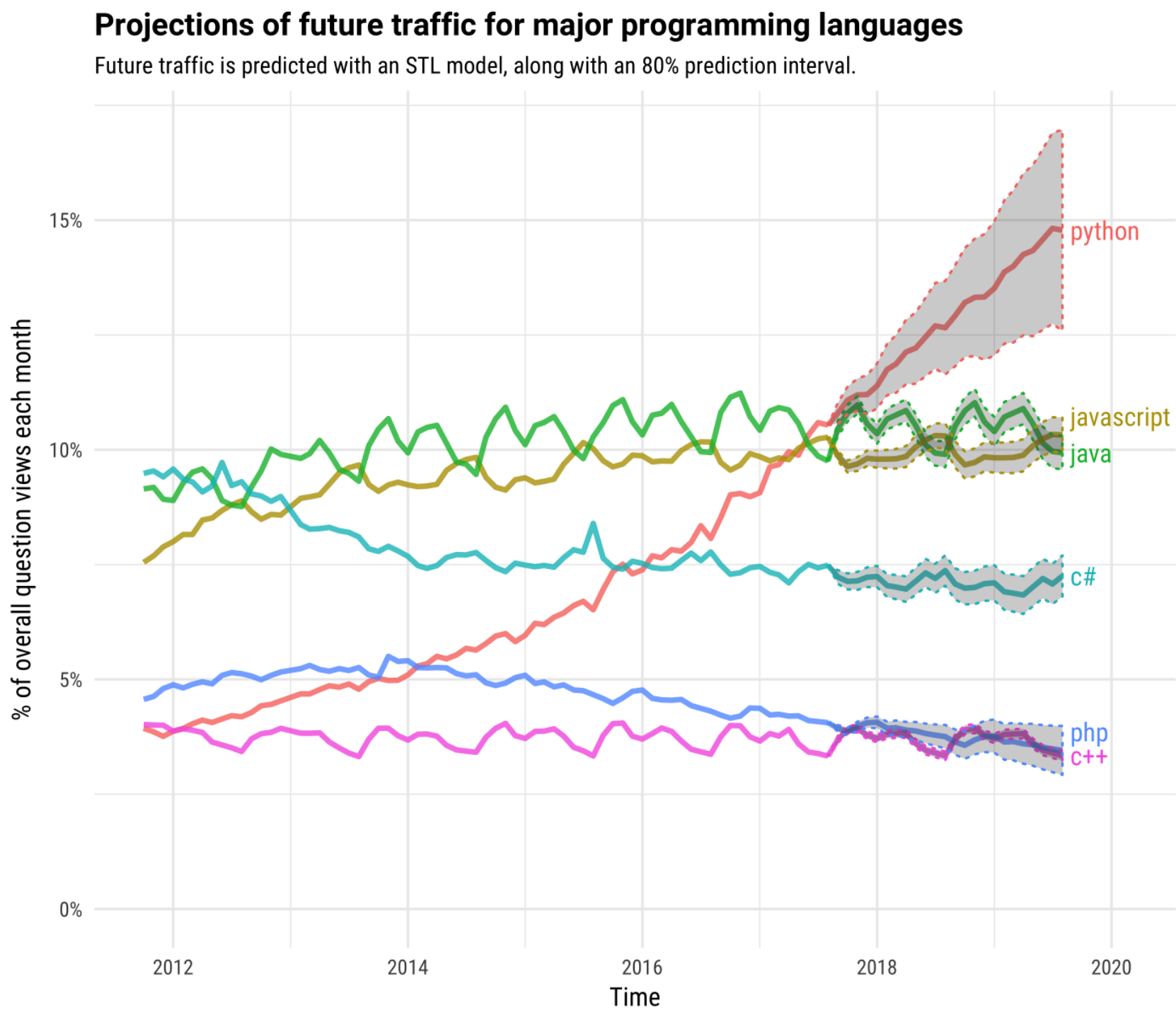
0< p(y \in class) < 1

Les outils

python anaconda jupyter notebooks

PYtHON

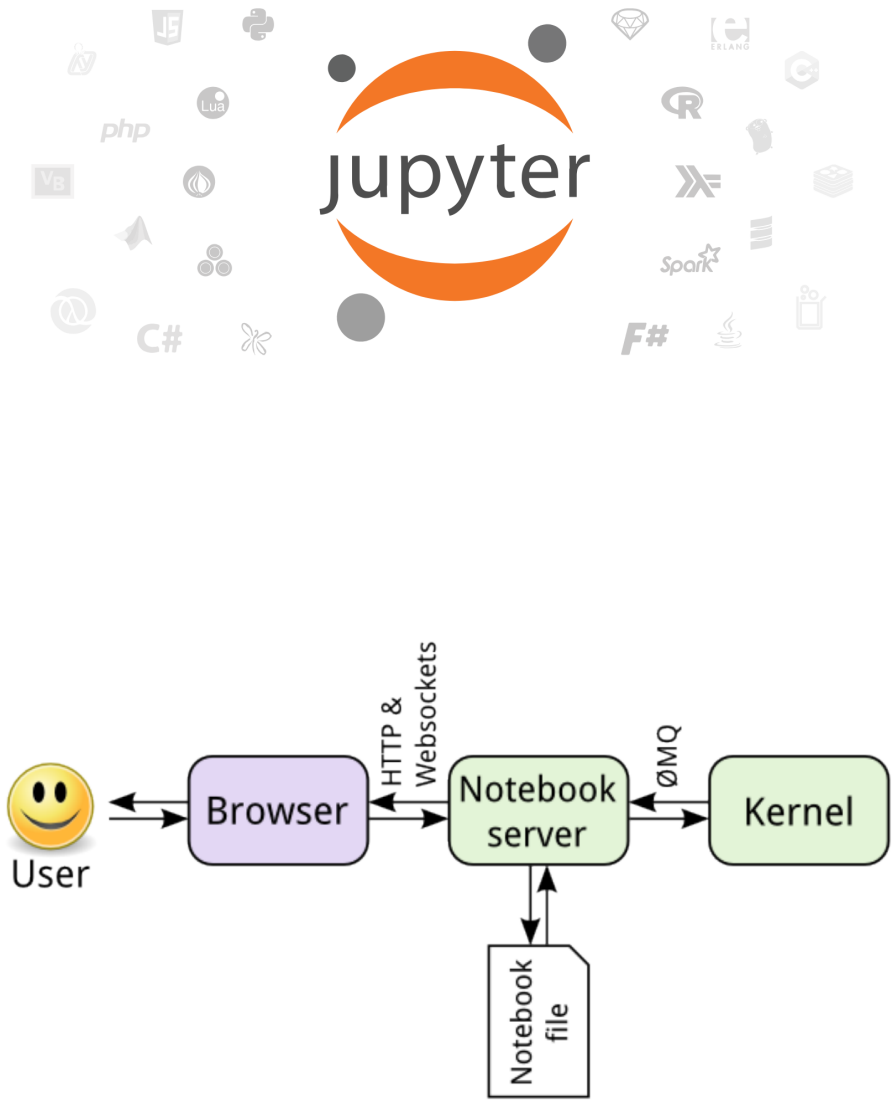
- Beaucoup d'applications: web, data science,
- scientific, ...
- Créé en 1991 par Guido von Rossum! 30 ans déjà!
- 130.000 packages et librairies
- Duck typing, pas de compilation, pas de ; ou de {}
- Indentation => le code est lisible
- Performances
- Mais il y a des surprises, des incoherences, des idioms, ...
- Python 2.7 ou python 3.6


Projections of future traffic for major programming languages
Future traffic is predicted with an STL model, along with an 80% prediction interval.
% of overall question views each month
15%
10%
5%
0%
2012
2014
2016
2018
2020
Time
python
javascript
java
c#
php
c++

```
liste_a = [n for n in range(100) if n % 2 ==0 ]
```

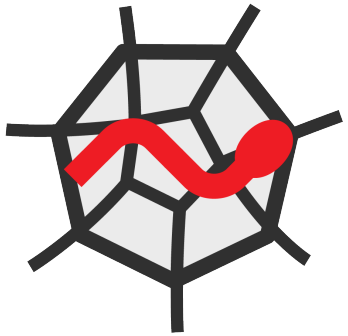
JUPYTER NOTEBOOK Executer du

code dans le navigateur
Partage et reproductabilité Calcul et visualisation
Multilingue: R, python, ...
Local ou cloud
$ Jupyter notebook
AWS Sagemaker, Google datalab, Kaggle kernels
A base de cellules
Documentation: markdown et latex Kernels: Python, R, Julia, Scala, ... Shell terminal
Alt: Beaker, Apache zeppelin
Mais: Le code est séquentiel + State problems



- github du cours
- https://github.com/alexisperrier/XEmines_DataScience_2020
- Notebooks
- https://github.com/alexisperrier/XEmines_DataScience_2020/blob/master/notebooks/Python_Pandas_Demo.ipynb
- Colab
- https://colab.research.google.com/
- Colab + github
- https://colab.research.google.com/github/alexisperrier/XEmines_DataScience_2020/blob/master/notebooks/Python_Pandas_Demo.ipynb

Editeurs de textes



SPYDER

The Scientific Python Development Environment

RÉCAPITULATIF

https://www.anaconda.com/

HTTP://JUPYTER.ORG/

Programme des 2 semaines
Révisions de python
Différence entre Data Science, Machine Learning et analyse prédictive Approche statistique vs approche machine learning
Déroulement d'un projet de Data science
Supervisée vs non-supervisée
Regression vs Classification
Anaconda, Python et Jupyter

Demo – les arbres de Paris

Votre tour

COLAB

- Copier le notebook les-arbres sur votre google drive
- download le dataset sur votre google drive
- Faites tourner le notebook et repondez aux questions

En local

- downloadez le notebook et le dataset
- > jupyter notebook
- Faites tourner le notebook et repondez aux questions

Liens

- What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning, and Big Data?
- Statistical Modeling: The Two Cultures