

Data science

Analyse prédictive et machine learning

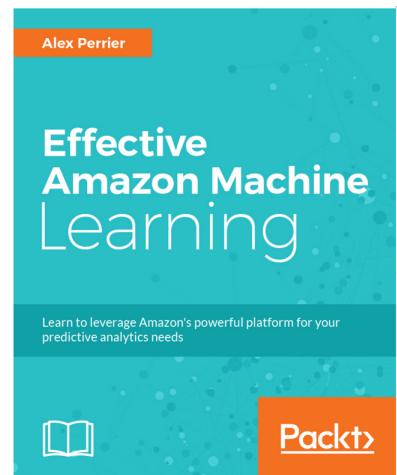
Prése

Alexis Perrier

- Data Scientist
- 20+ ans de software
- PhD TelecomParisTech 95' – E. Moulines

Connect:

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



PROGRAMME

- Machine learning avec scikit-learn
 - analyse prédictive
 - classification et régression
- Approches statistiques classiques:
 - régression linéaire,
 - régression logistique
- Modélisation machine learning
 - Random Forests, XGBoost
 - Support vector machines
 - Gradient stochastique
 - Adaboost, perceptron
 - Naive Bayes

- Concepts et Méthodes
 - biais, variance et overfitting
 - transformations de données
 - feature engineering
 - métriques et techniques d'évaluation
- datasets
 - iris, titanic, housing, ...
 - caravan, arbres, ...
- Python
 - notebook jupyter, anaconda
 - pandas, numpy
 - statsmodel et surtome

DÉROULEMENT

- Matin: théories, méthodes et démos
- Après-midi: Lab, workshop => notebooks jupyter
- Quizzes
- Projet final: Kaggle
 - <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/>

data science - machine learning - predictive
analytics - intelligence artificielle - deep
learning



🦊 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



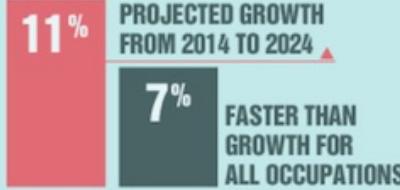
Tweet your reply

“Data scientist is
the sexiest job
of the 21st century.”

Harvard Business Review



JOB GROWTH AND DEMAND



DATA SCIENCE - MACHINE LEARNING - PREDICTIVE INTELLIGENCE ARTIFICIELLE - DEEP LEARNING

- **Data Analysis, Data Mining** : Exploration, trouver les tendances, les evolutions, les anomalies, etudier
- **Statistiques** : Trouver le modèle qui explique au mieux les données
- **Machine learning** : Le modèle apprend automatiquement à partir des données. Dimension importante : training
- **Analyse prédictive**: Construire ou entraîner des modèles qui peuvent "*prédirer*" à partir de données
- **Deep Learning** : Analyse prédictive supervisée avec des réseaux de neurones

[What is the difference between Data Analytics, Data Analysis, Data Mining, Data Science, Machine Learning?](#)

STATS VS MACHINE LEARNING

A TINY DROP OF HISTORY

Great article [Forbes: A Very Short History Of Data Science](#)

2001 Leo Breiman, Berkeley, publishes “[Statistical Modeling: The Two Cultures](#)”:

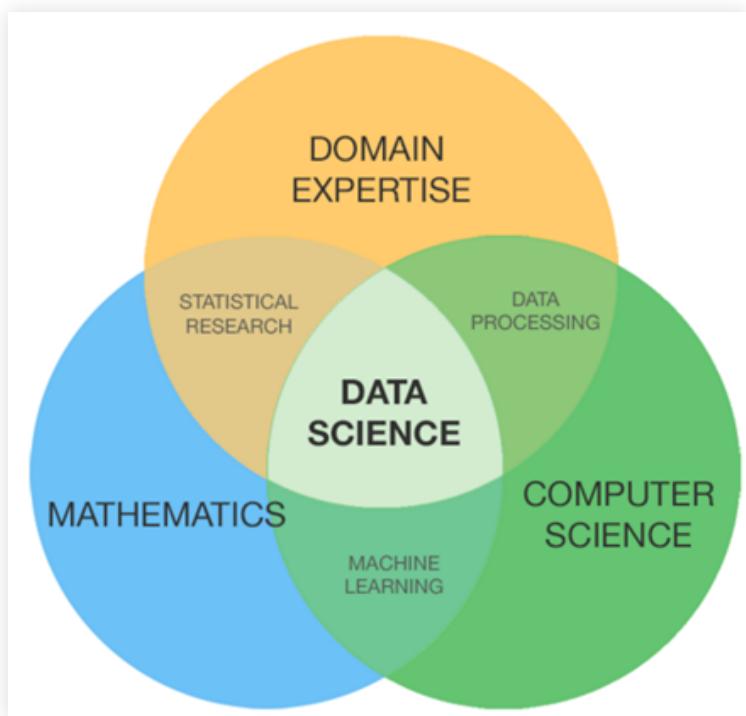
“There are two cultures in the use of statistical modeling to reach conclusions from data.

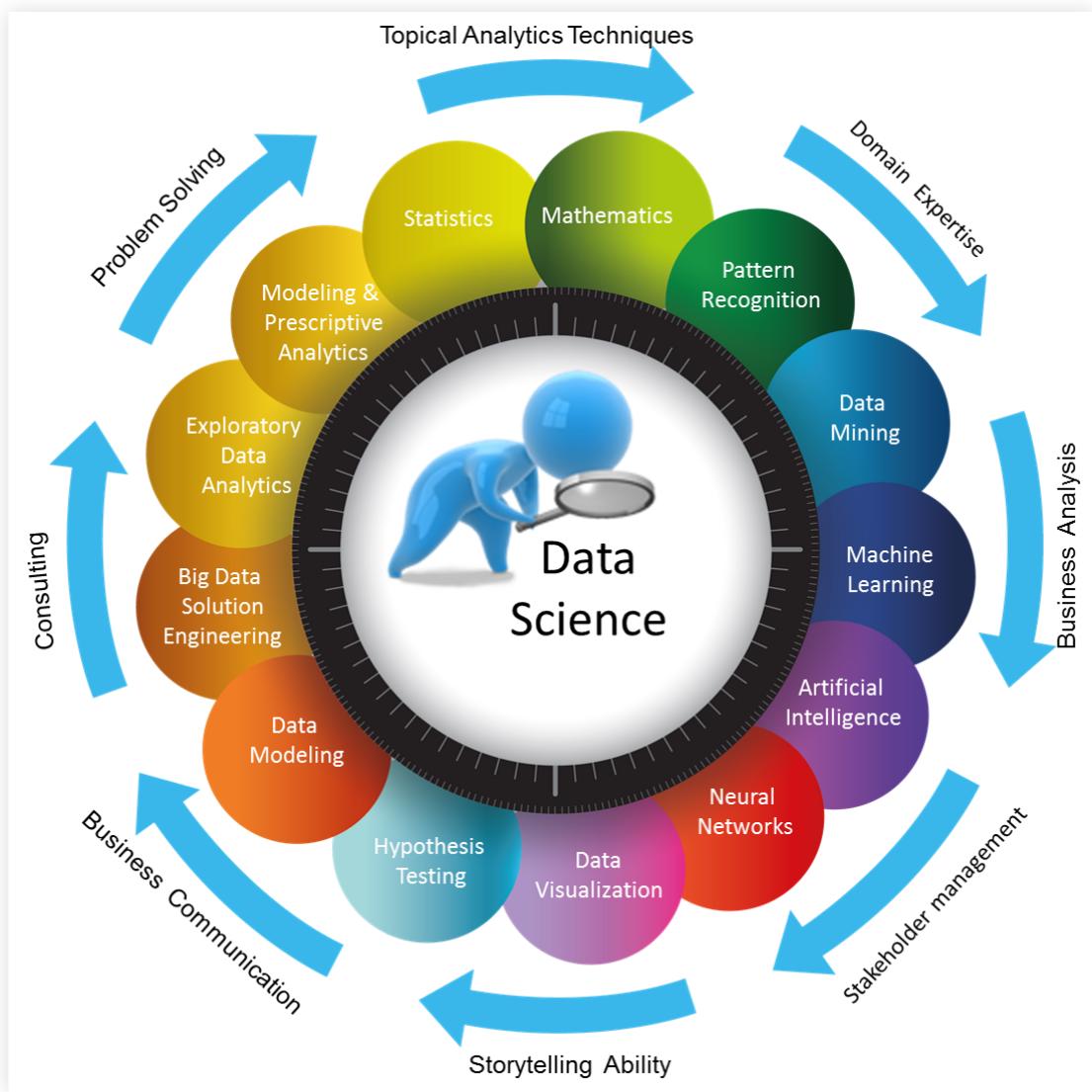
One assumes that the data are generated by a given stochastic data model. The other uses algorithmic data mechanism as unknown.

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 interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in recent years and can be used both on large complex data sets and as a more accurate and informative alternative to data models on small 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 statistics and adopt a more diverse set of tools.”

DATA SCIENCE: SKILLS

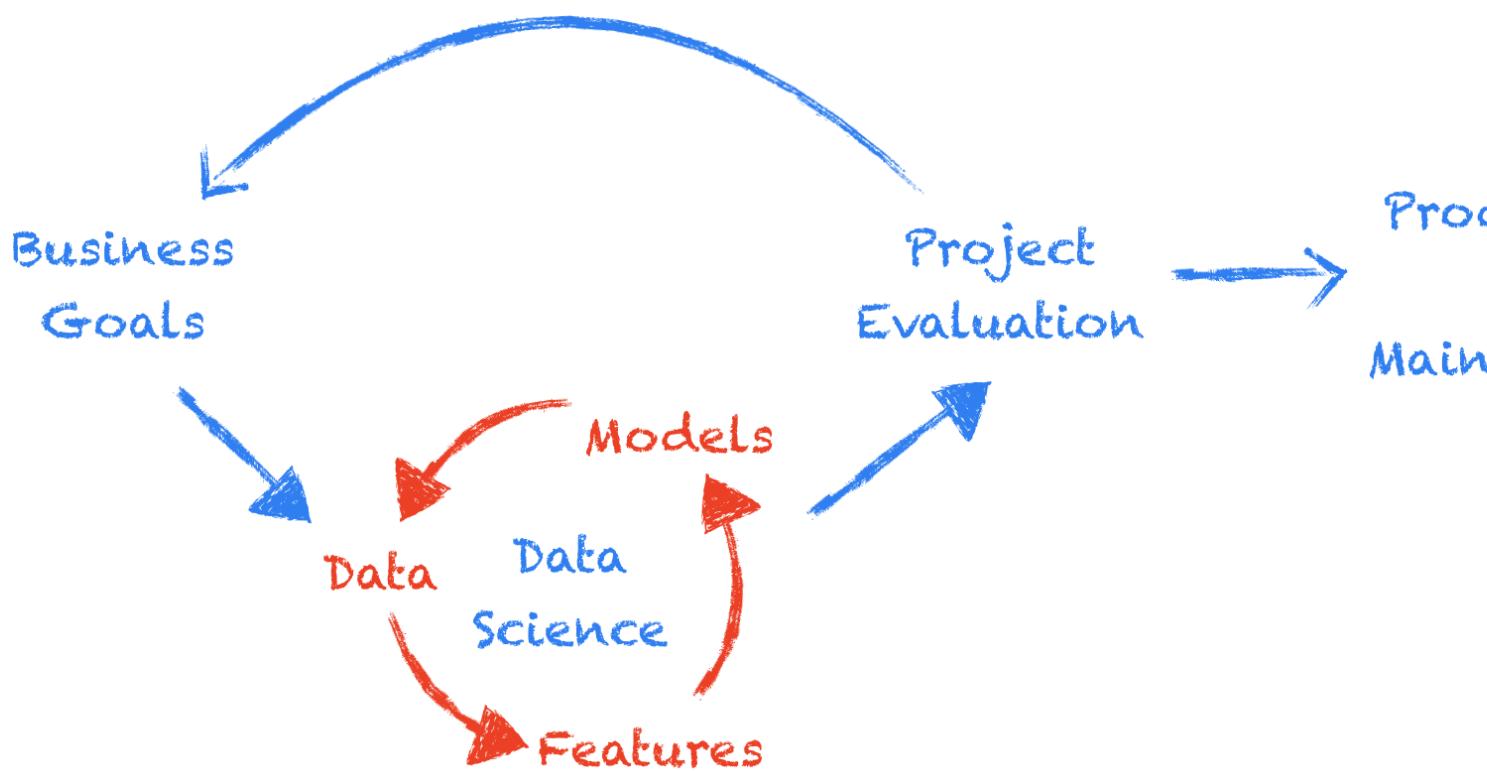




CHAMPS D'APPLICATIONS

- **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 checker
- **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
- **Image / Video:** automatic captionning, face and object recognition, ...

Data science workflow



A) LES DONNÉES

1. Définir le problème

- De quelles données disposent-on ?
- Sont-elles accessibles ?
- Que veut-on améliorer ?
- Comment mesurer l'efficacité de la solution, du modèle ?
- Choix des métriques

2. ETL: Extraction Transform Load

- Constituer le dataset
- Explorer et comprendre

LÀ COMMENCE LE TRAVAIL DE MODÉLISATION

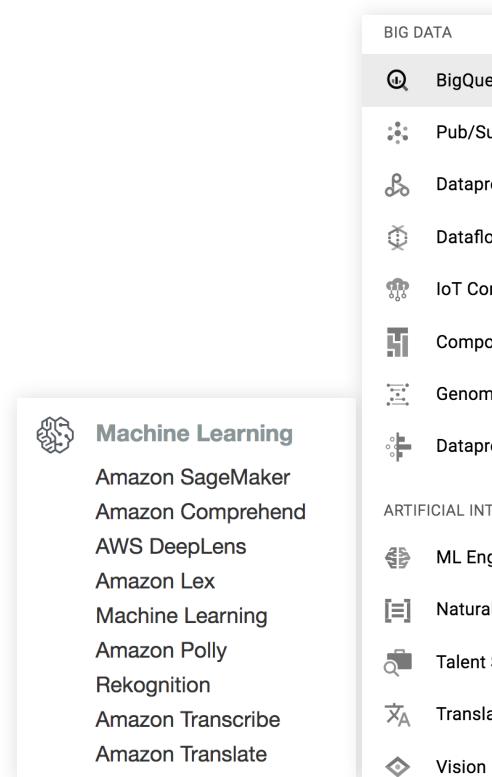
1. Travailler sur les variables

- Nettoyer et transformer : outliers, missing values, distributions, correlations, ...
- feature engineering
- feature selection

B) MACHINE LEARNING

4) Outils et plateforme

- Cloud (AWS, Google Cloud, Azure) ou local
- python (scikit-learn) ou R ou ...
- Modèles classiques
- Deep learning (TF, Keras, pytorch, ...)



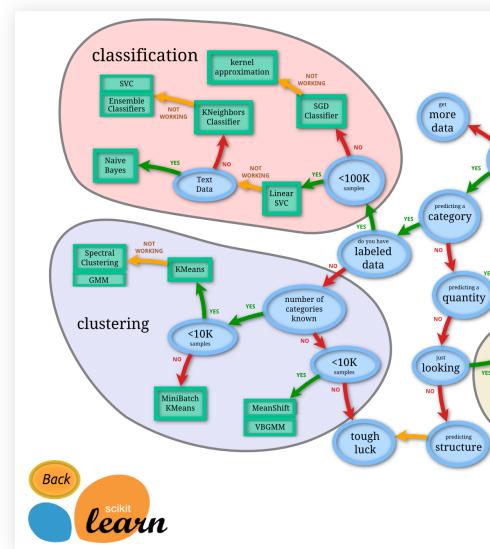
B) MACHINE LEARNING

5) Modélisation

- Choisir la bonne approche, le bon type de modèle
- *Train* le modèle
- Evaluer le modèle, scoring,
- Sélectionner les meilleurs paramètres du modèle

6) Appliquer sur de nouvelles données

- on quitte un environnement contrôlé (laptop / labo) pour le monde réel
- le modèle généralise t il bien ?
- les données ont-elles changées ?



C) NOUVELLE ITÉRATION

7) Présentation des résultats

- cycle itératif court
- communication
- data visualization

8) reprendre le problème au niveau des données

- il en faut plus
- il faut de nouvelles variables
-

ou au niveau de la définition du problème

- qu'est ce qu'on veut optimiser
- comment le mesurer
- accessibilité des données

D) MISE EN PRODUCTION

- ingénierie logicielle
- API
- streaming

A) LES DONNÉES

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

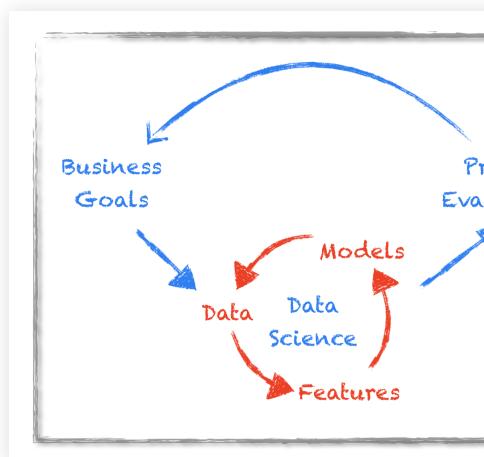
B) MACHINE LEARNING

- 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



DATA SCIENCE - MACHINE LEARNING - PREDICTIVE

[Can I learn Machine Learning completely with Kaggle?](#)

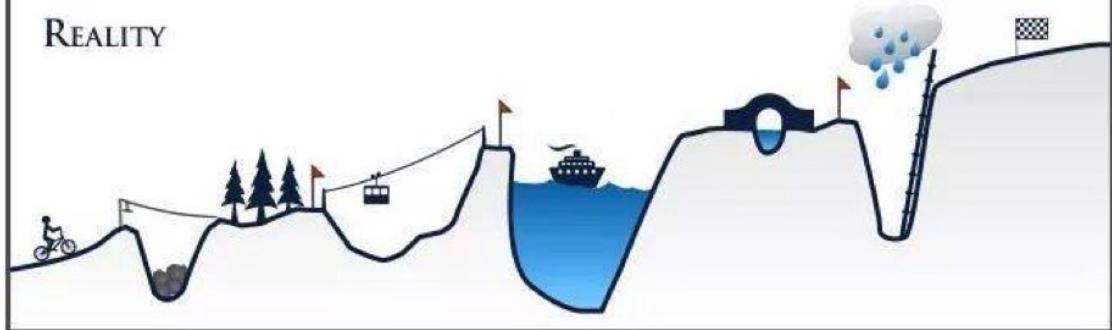
While modeling is the sexy part of any machine learning project, it is also one of the parts that you will actually spend most time on.

*In a business environment 80–90% of the time will be spent on defining problems worthwhile solving, defining the scope, procuring access to the **raw data**, **understanding** the data, generating **features**, presenting findings, and finally **deploy** the model to production via API or other automated approaches.*

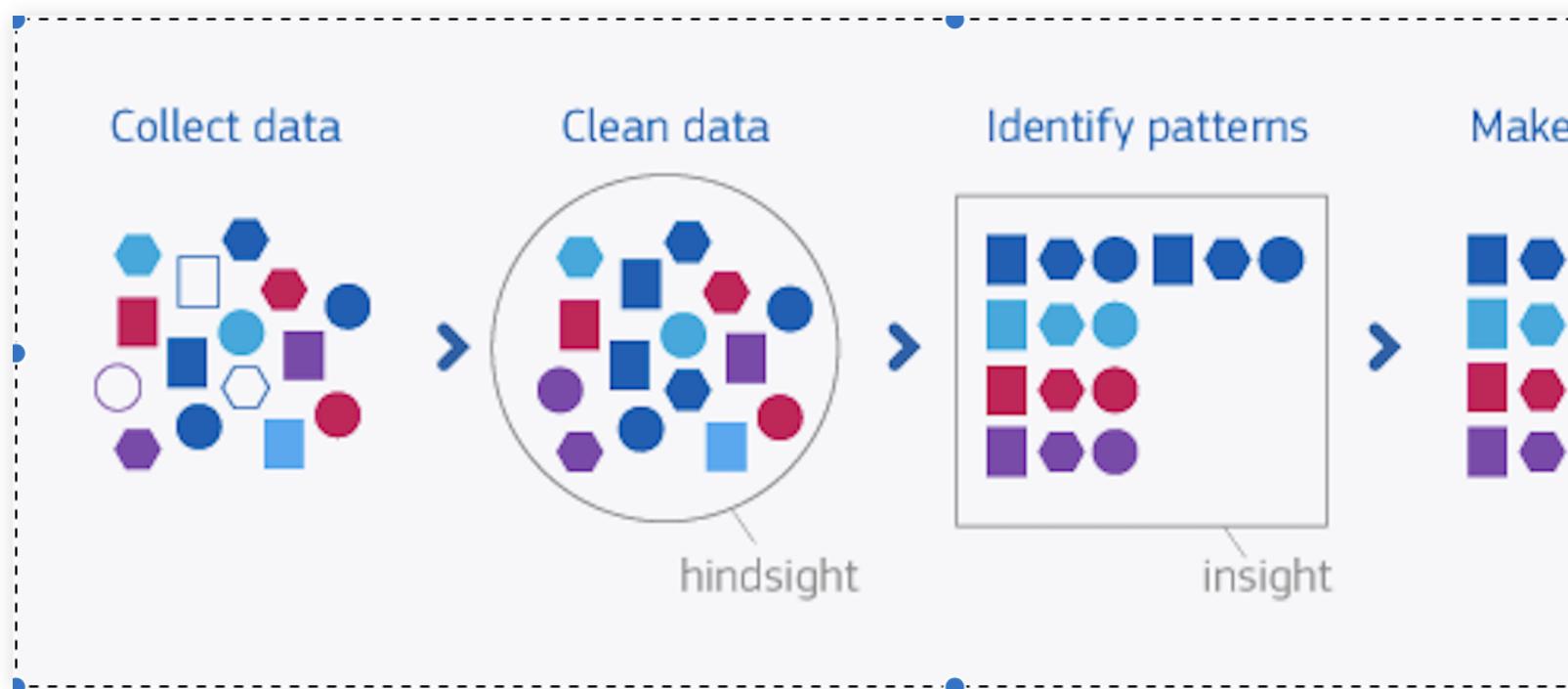
YOUR PLAN



REALITY



ANALYSE PRÉdictive

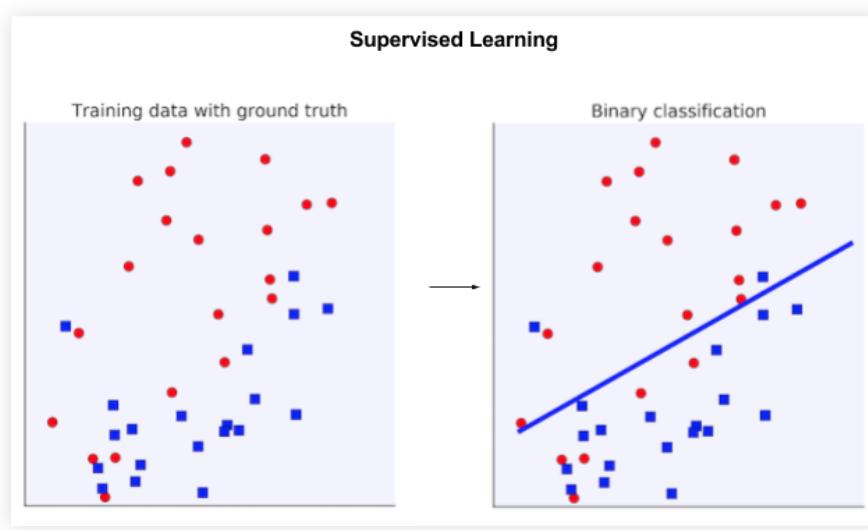


Predictive analytics : analyze current and historical facts to make predictions about **future or otherwise**.
Predictive analytics provides a predictive score (probability).

SUPERVISÉE

Le dataset d'apprentissage inclut la variable à prédire [cible]. On a un certain nombres d'exemples sur lesquels on peut entraîner un modèle

- logique de scoring, de classification et de prediction
- Random forest, Regression linéaire ou logistique, SVM, ...
- Classification: On connaît le nombre de classes



NON SUPERVISÉE

Le dataset d'apprentissage n'inclut pas de **ground truth**

- logique de clustering, de classification des échantillons sans connaître les classes
- notion de similarité et de distance entre échantillons
- K-means, K-NN, ...



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 discrète, une ou plusieurs catégories

CAS BINAIRE

- Achat, résiliation, click
- Survie, maladie, succès exercice
- Positif ou négatif
- Spam, fraude

MULTI CLASS - MULTINOMIALE

- Catégories, types (A,B,C),
- Positif, neutre ou négatif
- Espèces de plantes d'animaux
- Pays, planètes

ORDINALE

- Notes, satisfaction, ranking

REGRESSION À CLASSIFICATION

- discréteriser la variable

Age =>

- 0 - 12
- 12 - 24
- 25 - 49
- 50 - 65
- plus de 65

CLASSIFICATION REGRESSION

- Prédire une probabilité au

$0 < P(x \in$

Environnement

Python, anaconda et jupyter

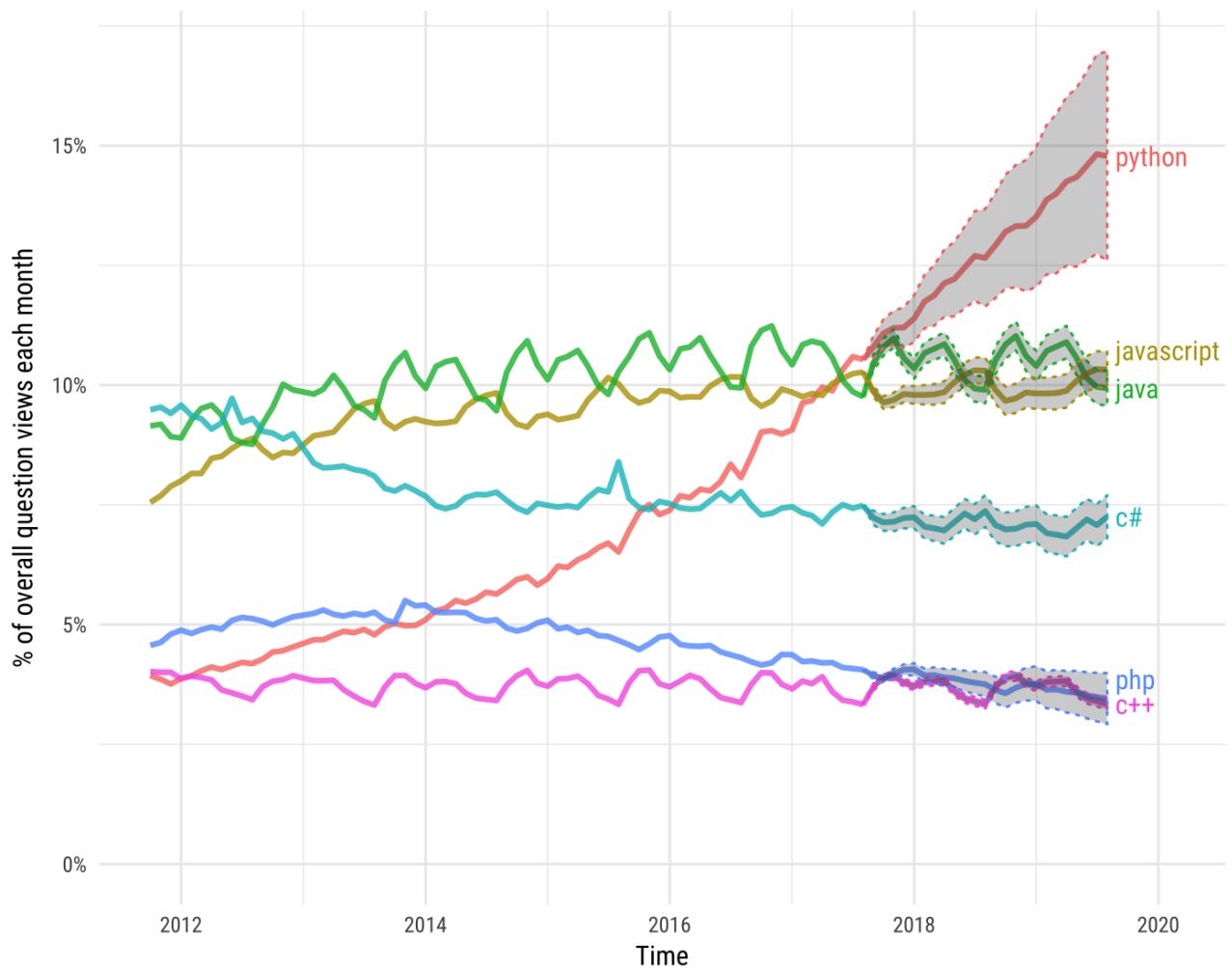
PYTHON

- 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 incohérences, 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.



PYTHON

- list comprehension

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

- pandas dataframe

```
df = pd.read_csv(filename)
```

```
df = df.groupby(by = 'age' ).reset_index(inplace = True)
```

```
df = df.age.apply(lambda a : une_fonction(a) )
```

- notebook jupyter

```
> jupyter notebook
```

- exemple

[Python_Pandas_Demo.ipynb](#)

QUEL PYTHON AVEZ-VOUS?

Dans un terminal

```
> python --version
```

```
[@:~]$ python --version
Python 3.6.2 :: Anaconda custom (x86_64)
[@:~]$ █
```

ANACONDA

- Distribution Anaconda et package manager conda

```
conda install package_name
```

- Data science en python:

- Dataframe: pandas, dask
- Math, science: numpy, scipy, statsmodel,
- Dataviz: matplotlib, plot.ly, bokeh
- Deep learning: Tensorflow, Keras, Mxnet, ...
- Text: Gensim, NLTK, Spacy.io
- scikit-learn: <http://scikit-learn.org/>

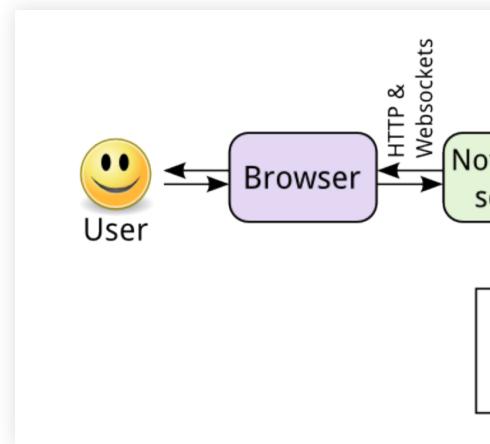


<https://www.anaconda.com/>

JUPYTER NOTEBOOK

- Executer du code dans le navigateur
- Partage et reproductibilité
- 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

HTTP://JUPYTER



EDITEURS DE TEXTE



Récapitulatif

RÉCAPITULATIF

- 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

LAB

2 datasets

- 200.000 arbres de Paris
 - Espèces, genres, famille
 - Adresse, geolocalisation
 - Environnement: rue, jardin, ..
 - hauteur et circonférence
- Arrondissement de Paris
 - Superficie

NOTEBOOK D'E ET PANDAS

- load dataset dans une dat
- visualisation des variables
- statistiques des variables des catégories
- trouver les outliers et les e

Arrondissements:

- quels arrondissements ont
 - le plus d'arbres
 - le plus de variétés d'arbres
 - les arbres les plus hauts
- hauteur et circonférence en fonction d'arbres
- Comment sont définis les arrondissements?
 - comment traiter les arrondissements avec cette colonne

Domanialité

* memes questions que pour les arrondissements

En joignant le dataset arrondissements au dataset superficie des arrondissements

Creer une variable code_postal
qui permette de joindre les 2 fichiers
75102

Joindre les 2 fichiers

Calculer le nombre d'arbres par quartier
utilisant groupby et count() en excluant
ceux qui ne sont pas dans Paris

