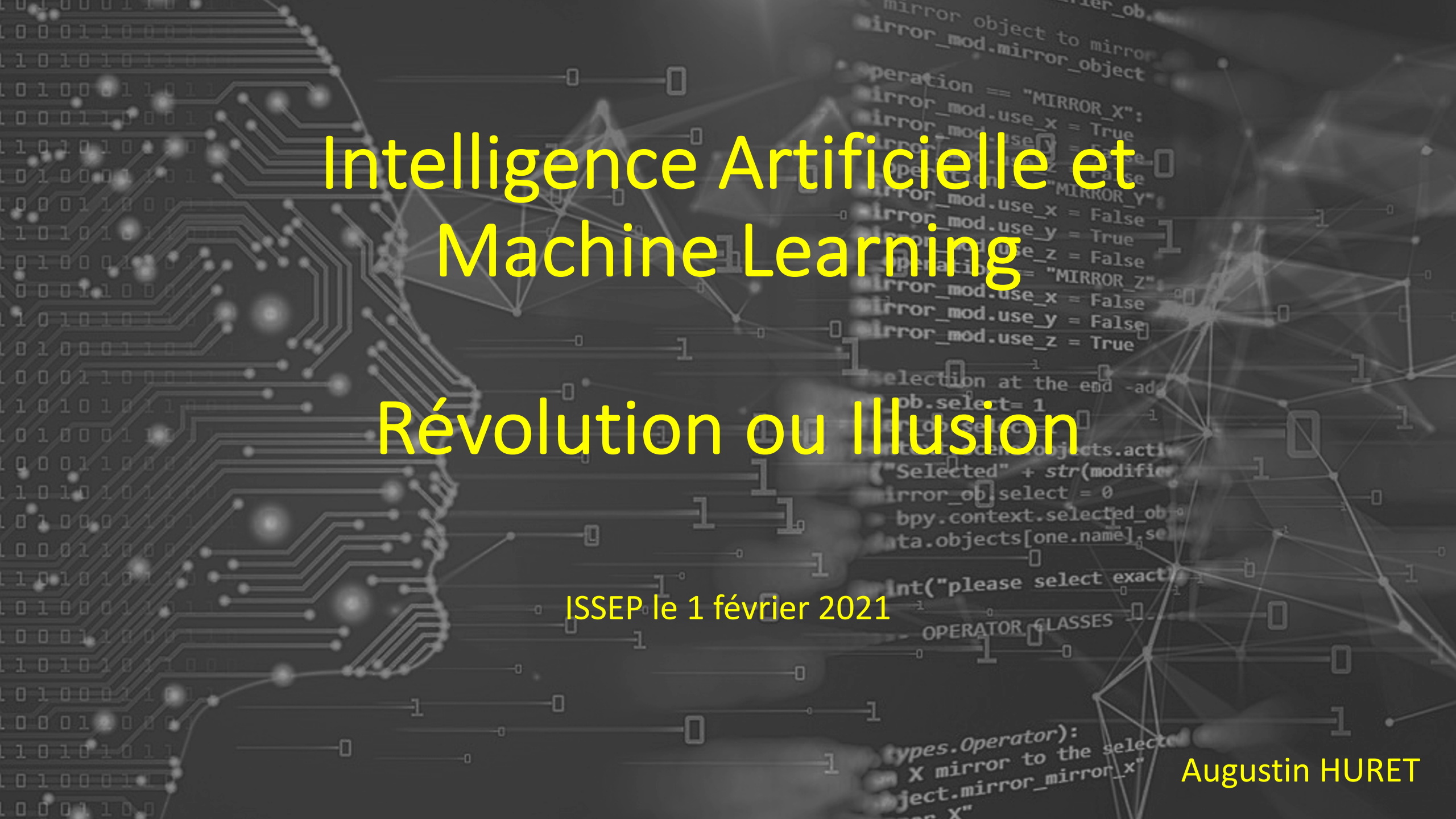


Intelligence Artificielle et Machine Learning

Révolution ou Illusion

ISSEP le 1 février 2021

Augustin HURET

Qu’est ce que l’Intelligence Artificielle

L'intelligence artificielle (IA) est « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence ».

Elle correspond donc à un ensemble de concepts et de technologies plus qu'à une discipline autonome constituée. D'autres, remarquant la définition peu précise de l'IA, notamment la CNIL, introduisent ce sujet comme « le grand mythe de notre temps ».

Souvent classée dans le groupe des sciences cognitives, elle fait appel à la neurobiologie computationnelle (particulièrement aux réseaux neuronaux), à la logique mathématique (partie des mathématiques et de la philosophie) et à l'informatique. Elle recherche des méthodes de résolution de problèmes à forte complexité logique ou algorithmique. Par extension elle désigne, dans le langage courant, les dispositifs imitant ou remplaçant l'homme dans certaines mises en œuvre de ses fonctions cognitives.

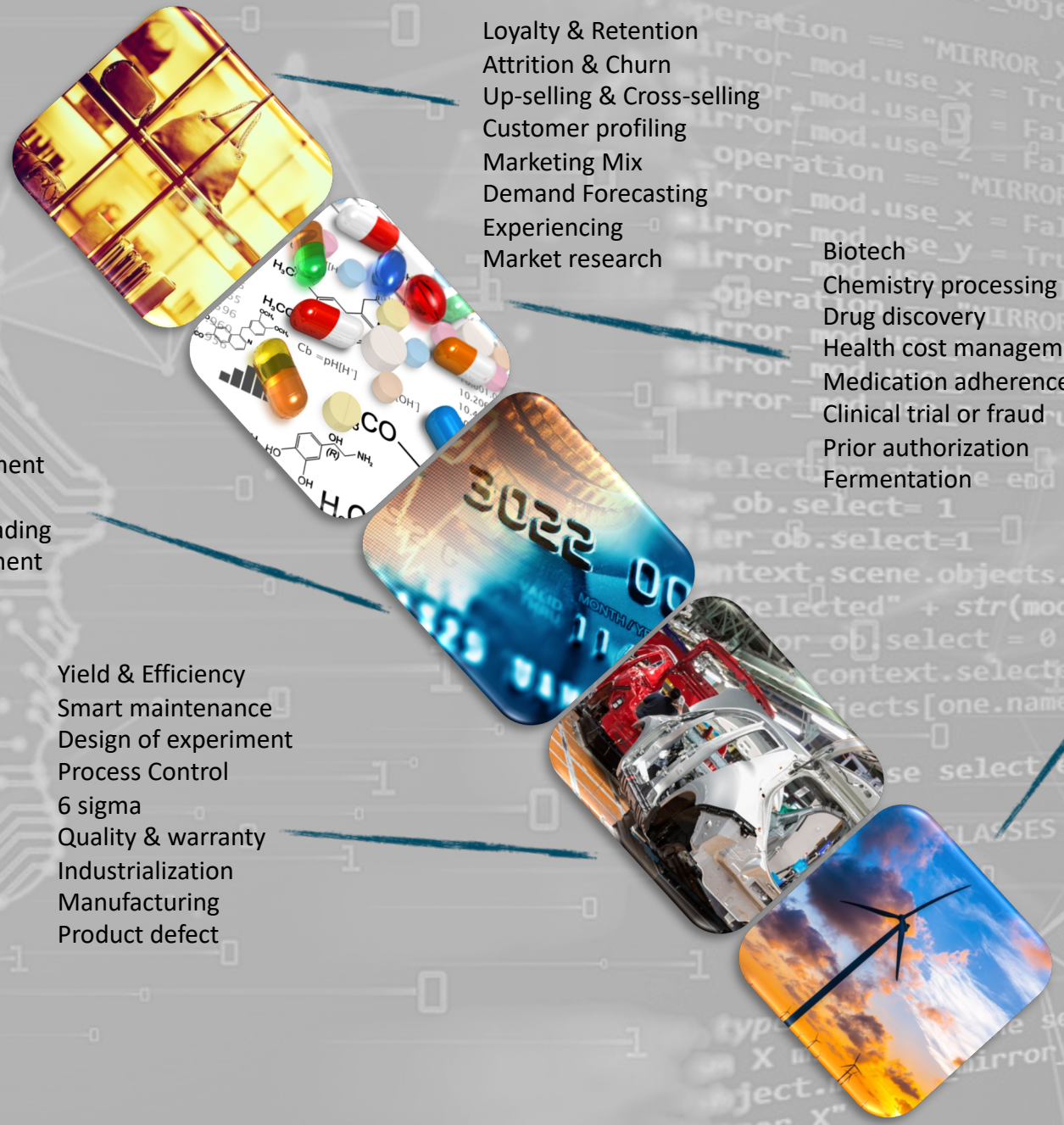
Qu’est ce que le Machine Learning

L'apprentissage automatique (en anglais : machine learning, litt. « apprentissage machine »), apprentissage artificiel ou apprentissage statistique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour donner aux ordinateurs la capacité d'« apprendre » à partir de données, c'est-à-dire d'améliorer leurs performances à résoudre des tâches sans être explicitement programmés pour chacune. Plus largement, il concerne la conception, l'analyse, l'optimisation, le développement et l'implémentation de telles méthodes.

L'apprentissage automatique comporte généralement deux phases. La première consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. L'estimation du modèle consiste à résoudre une tâche pratique, telle que traduire un discours, estimer une densité de probabilité, reconnaître la présence d'un chat dans une photographie ou participer à la conduite d'un véhicule autonome. Cette phase dite « d'apprentissage » ou « d'entraînement » est généralement réalisée préalablement à l'utilisation pratique du modèle. La seconde phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises afin d'obtenir le résultat correspondant à la tâche souhaitée. En pratique, certains systèmes peuvent poursuivre leur apprentissage une fois en production, pour peu qu'ils aient un moyen d'obtenir un retour sur la qualité des résultats produits.

Selon les informations disponibles durant la phase d'apprentissage, l'apprentissage est qualifié de différentes manières. Si les données sont étiquetées (c'est-à-dire que la réponse à la tâche est connue pour ces données), il s'agit d'un apprentissage supervisé. On parle de classification ou de classement3 si les étiquettes sont discrètes, ou de régression si elles sont continues. Si le modèle est appris de manière incrémentale en fonction d'une récompense reçue par le programme pour chacune des actions entreprises, on parle d'apprentissage par renforcement. Dans le cas le plus général, sans étiquette, on cherche à déterminer la structure sous-jacente des données (qui peuvent être une densité de probabilité) et il s'agit alors d'apprentissage non supervisé. L'apprentissage automatique peut être appliqué à différents types de données, tels des graphes, des arbres, des courbes, ou plus simplement des vecteurs de caractéristiques, qui peuvent être continues ou discrètes.

Dans tous les secteurs de l’économie



Early Intervention

Recruitment Fundraising

Development targeting

Student & employee retention

Risk profiling

Cyber security

Safety

Curriculum & program effectiveness

Fraud

Operational Risk

Credit scoring
Underwriting
Claim management
Anti-fraud
Quantitative trading
Asset management
Leasing

Yield & Efficiency
Smart maintenance
Design of experiment
Process Control
6 sigma
Quality & warranty
Industrialization
Manufacturing
Product defect

Loyalty & Retention
Attrition & Churn
Up-selling & Cross-selling
Customer profiling
Marketing Mix
Demand Forecasting
Experiencing
Market research

Biotech
Chemistry processing
Drug discovery
Health cost management
Medication adherence
Clinical trial or fraud
Prior authorization
Fermentation

Exploration
Efficiency
Performance
Emission
Hazard assessment
Targeting

Dans toutes les fonctions de l’entreprise

R&D
Manufacturing
Plant Management
Safety

Operations
Sales
Merchandising & Price strategy
Supply Chain

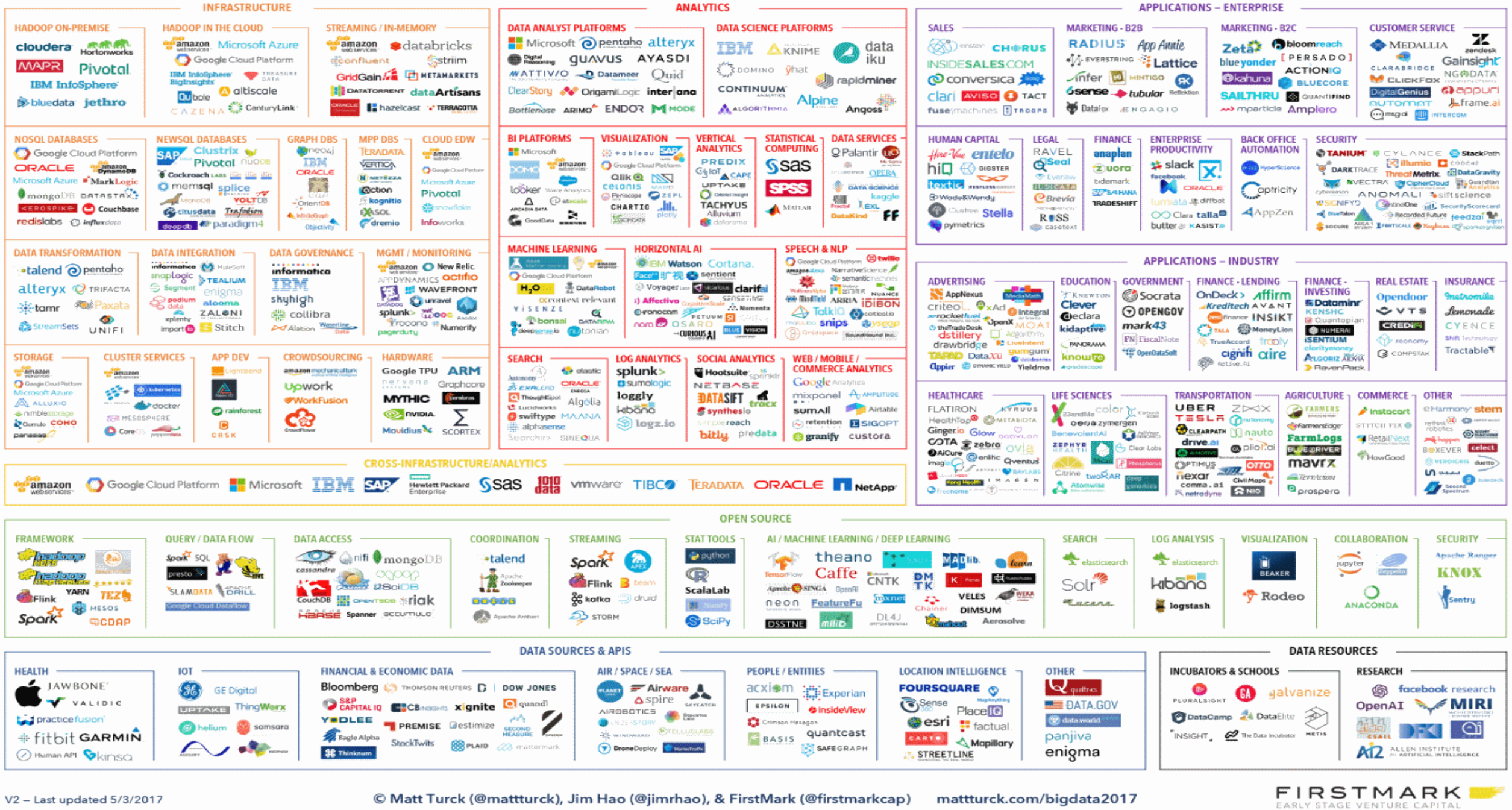
Digital Strategy
Marketing
E-commerce
Customer Experience

Finance
M&A
Risk Assessment
Fraud

Employee retention
Curriculum & program effectiveness
Recruitment

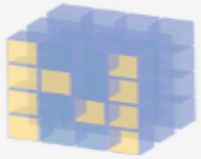
Sustainability
Efficiency

BIG DATA LANDSCAPE 2017



Data Science Tools

Basics


NumPy

pandas

SciPy.org

IA

PyBrain

scikit learn

TensorFlow

theano

PyTorch

Keras

fast.ai

Visualization

matplotlib

Seaborn

plotly

Bokeh

Statistiques

statsmodels

Environment

IP[y]:

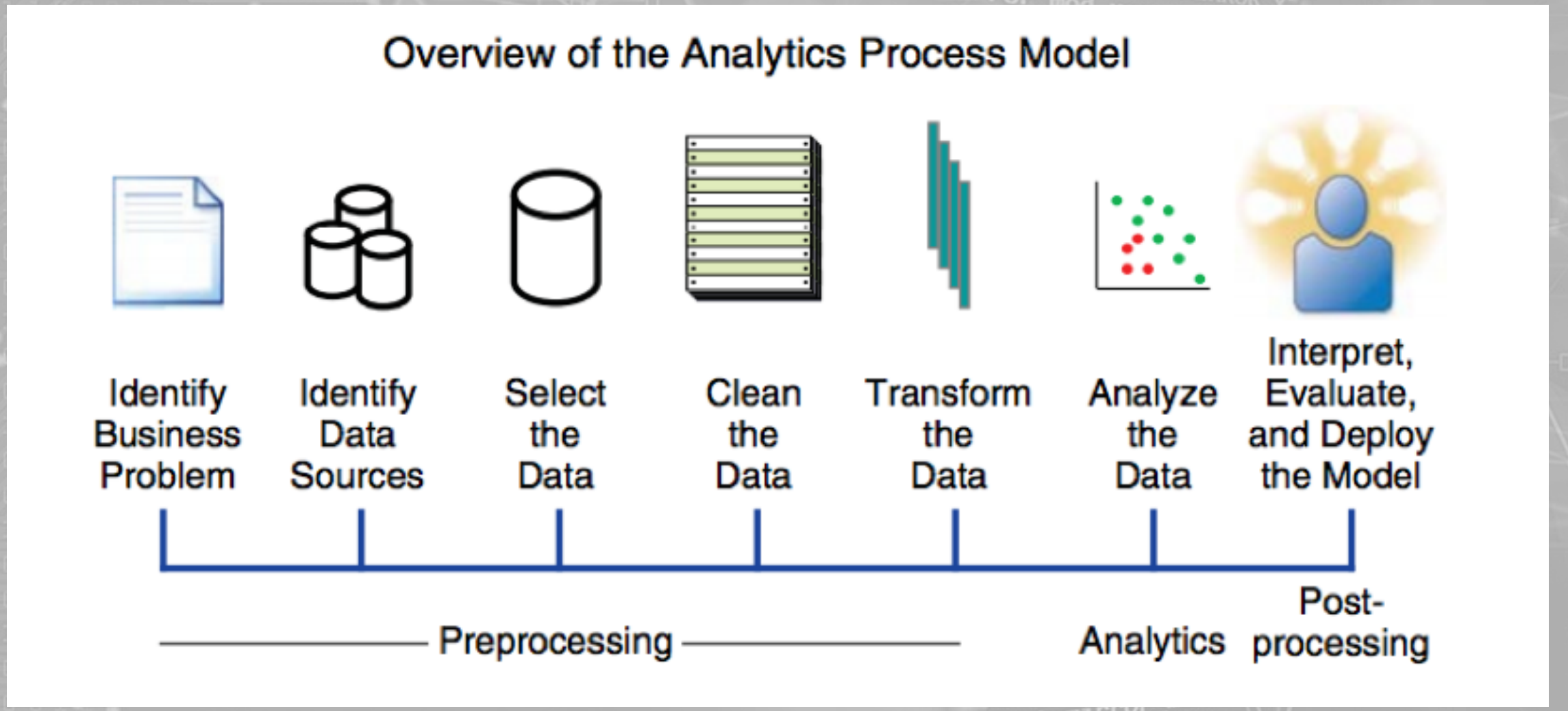
jupyter

Natural Language Processing

NLTK

L’intelligence *(non artificielle),* c’est de réussir à rendre simples les choses compliquées…

Le processus d'analyse de données et de construction de modèles



Les différentes familles d'algorithmes

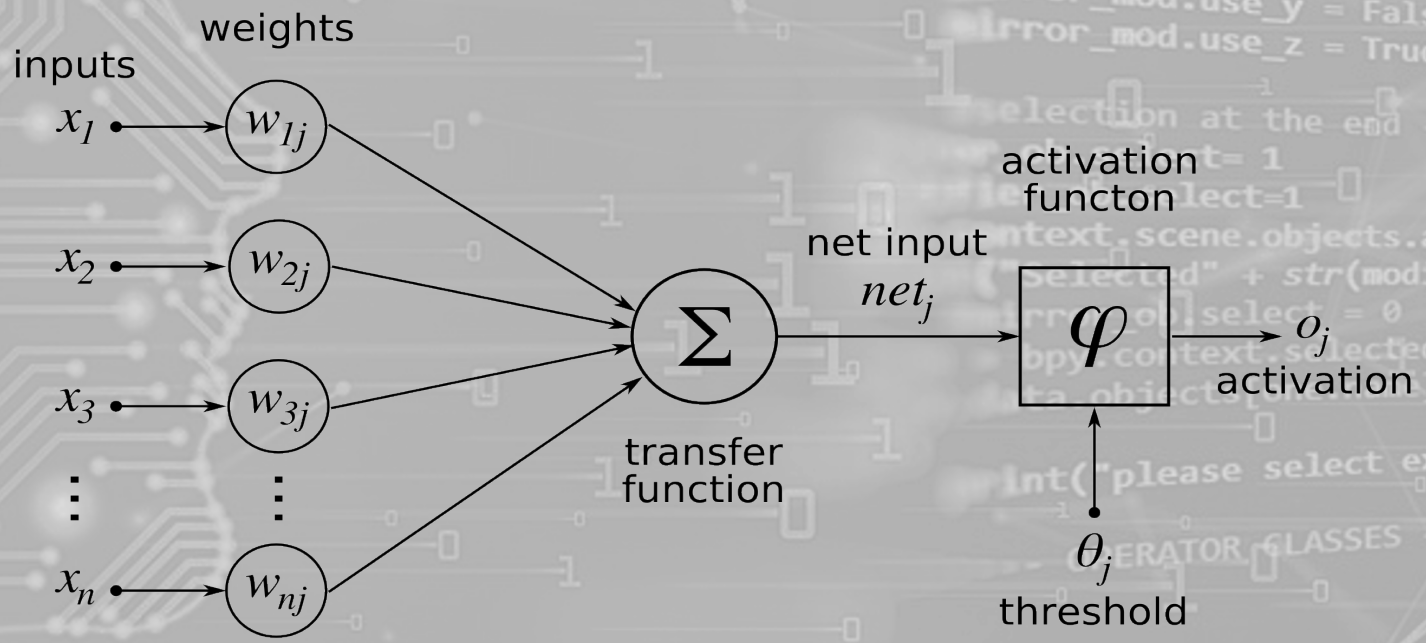
- Les réseaux de neurones – Neural Network
- Les plus proches voisins – K-means
- Les machines à supports vecteurs – SVM
- Les arbres de décision – Decision Tree
- Les forêts aléatoires – Random Forest
- Les régressions linéaires et logistiques

Réseau de neurones

•**Definition** :

An artificial neuron is a mathematical function from a set of inputs to a scalar output.

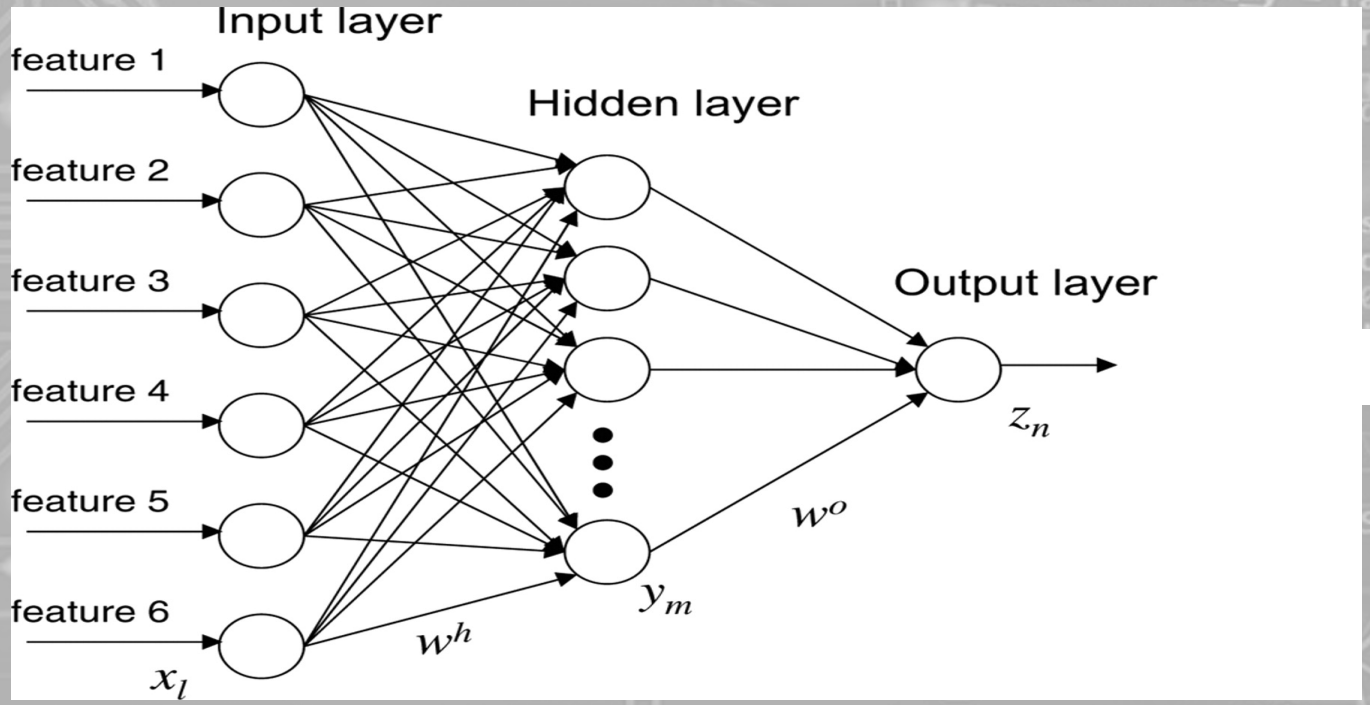
•**Structure**:



Réseau de neurones multi-couches

- **Definition** :
  A multi layer network based on scalar products

- **Structure**:



Réseau de neurones

- **Advantages:**
  - it can theoretically classify any class with enough hidden neurons
  - It can solve problems with several output classes

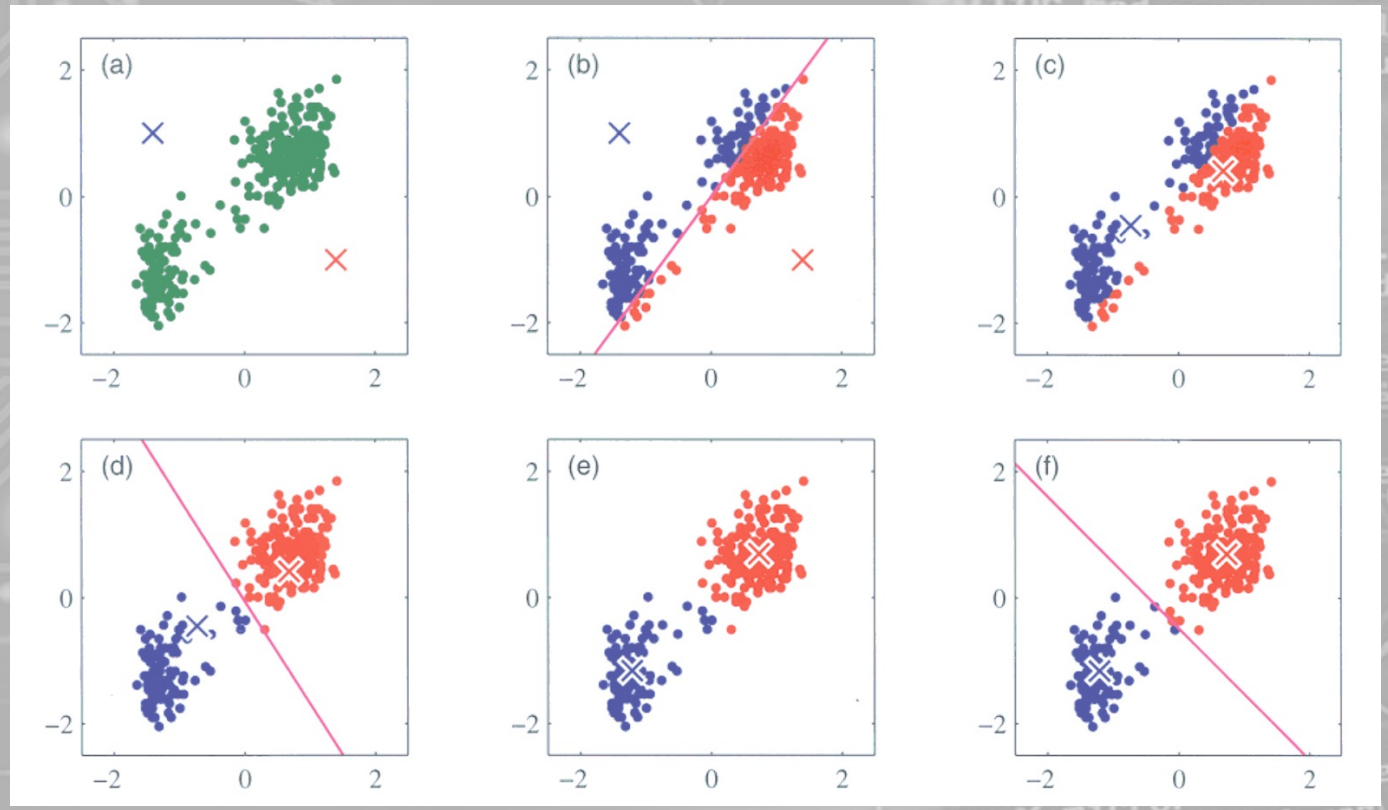
- **Disadvantages:**
  - Black box
  - A lot of parameters to tune (number of neurons, number of layers)
  - The cost function minimization can be very difficult and has often many local minima
  - Difficult to deal with discrete data

Plus proches voisins

•**Definition** :

Algorithm that partitions n observations into k clusters in which each observation belongs to the nearest mean

•**Structure**:



Plus proches voisins

- **Advantages:**
  - Often a good way to find large clusters
  - No overlap between clusters

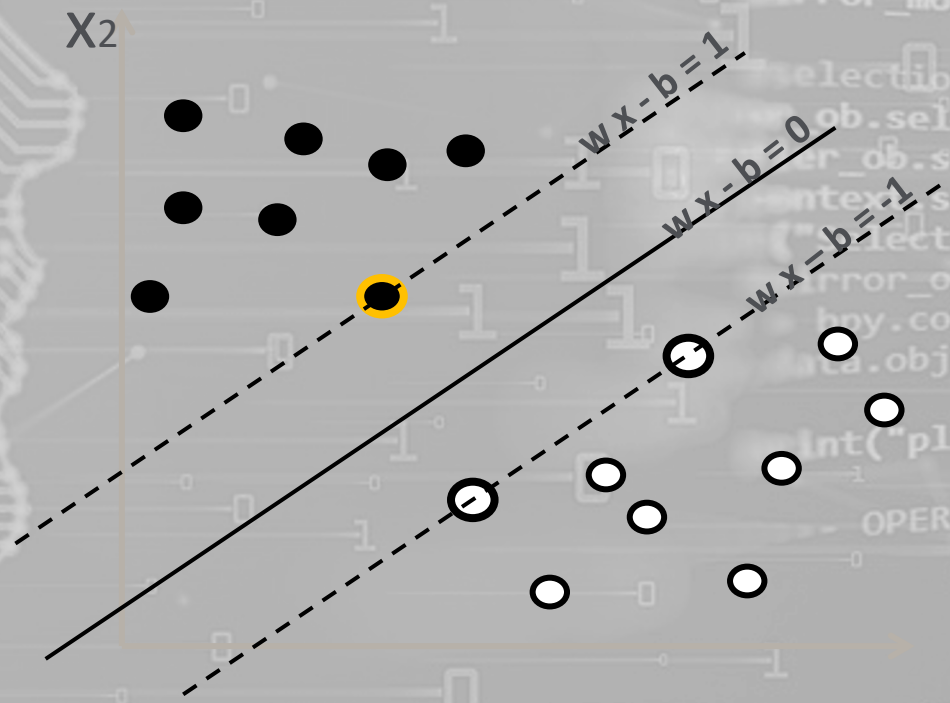
- **Disadvantages:**
  - We need to know the number K of clusters in advance
  - Heavily depends on the distance measure
  - Can fall in local minima

Machine à support vecteur

•**Definition** :

Algorithme that finds the maximum-margin hyperplane that divides the two classes of datapoints.

•**Structure**:

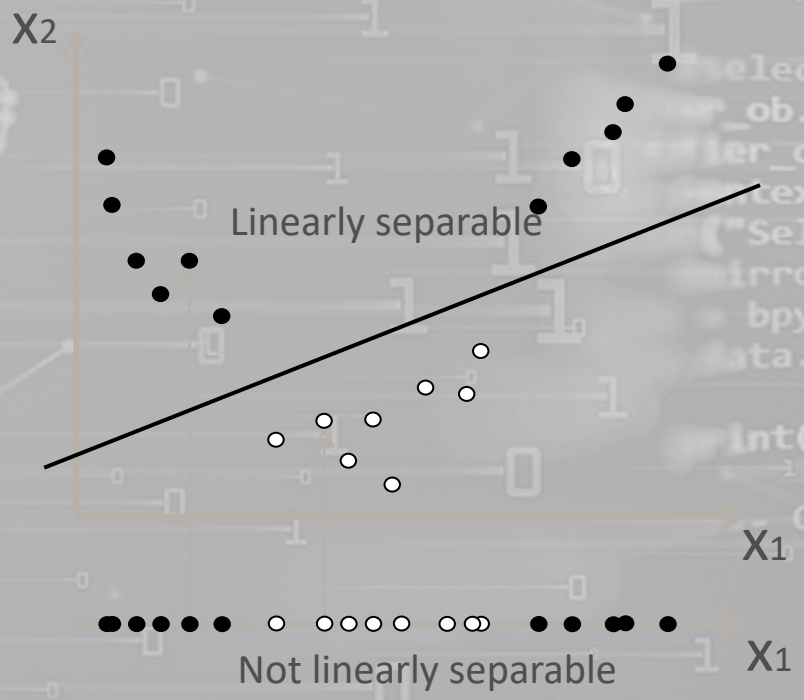


Machine à support vecteur

•**The Kernel trick**:

For databases that are not linearly separable, the Kernel trick can be used to find a higher dimension space where the two classes can be separated.

•**Structure**:



Machine à support vecteur

- **Advantages:**
  - It is very robust
  - It can solve problems with several output classes
  - It is fast
  - It provides the main drivers

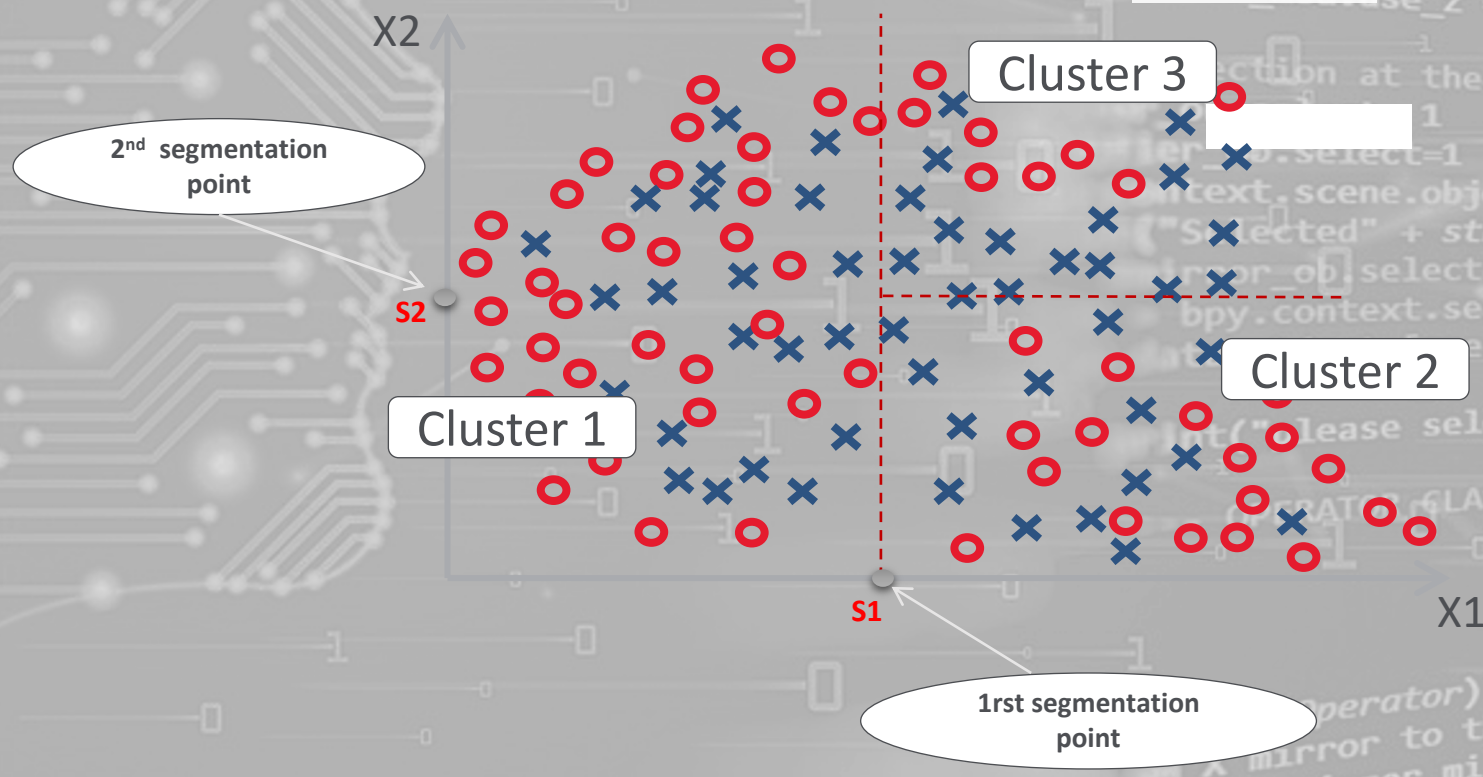
- **Disadvantages:**
  - Black box
  - Does not deal with missing information

Arbre de décision

•**Definition** :

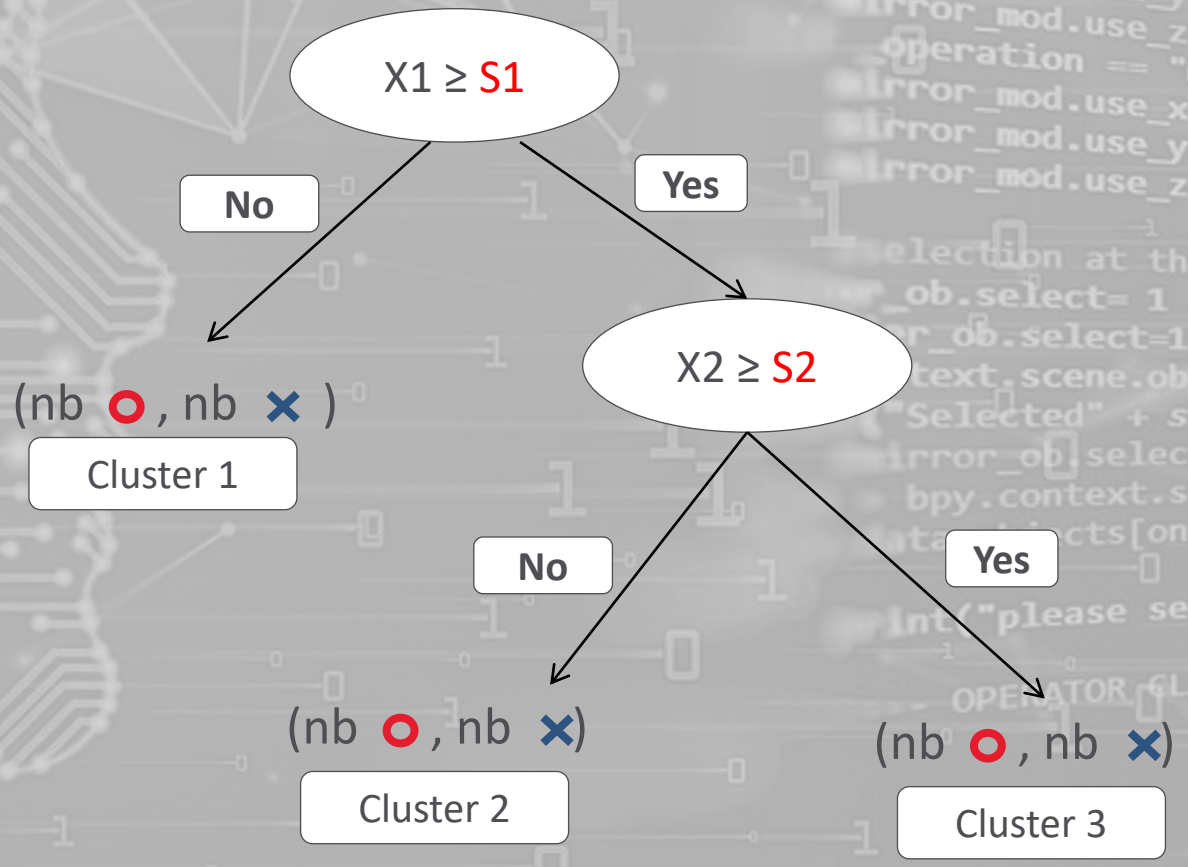
Algorithm that divides the space along each variable according to the best separation power. Build a tree od successive decisions based on majority votes.

•**Space structure**:



Arbre de décision

- **Tree Structure**



Arbre de décision

- **Advantages:**
  - Tree structure is easy to understand
  - Explanatory power
  - No overlap between clusters

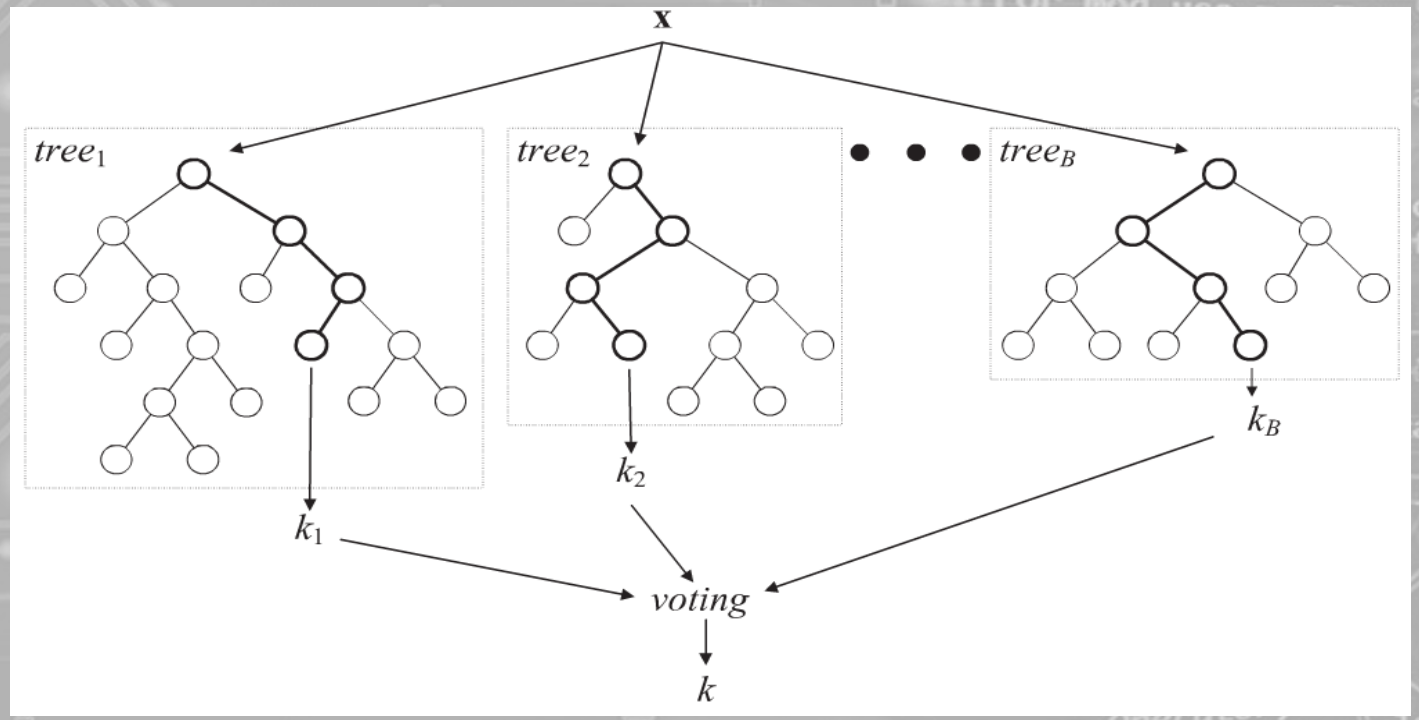
- **Disadvantages:**
  - Rules can have many variables
  - One segmentation point par axis
  - Difficulty to treat continuous variables without loosing information
  - Can hardly learn on data with missing information

Forêts aléatoires

- **Definition** :

  Combination of Decision Trees on various sub samples to improve the performance

- **Structure**:



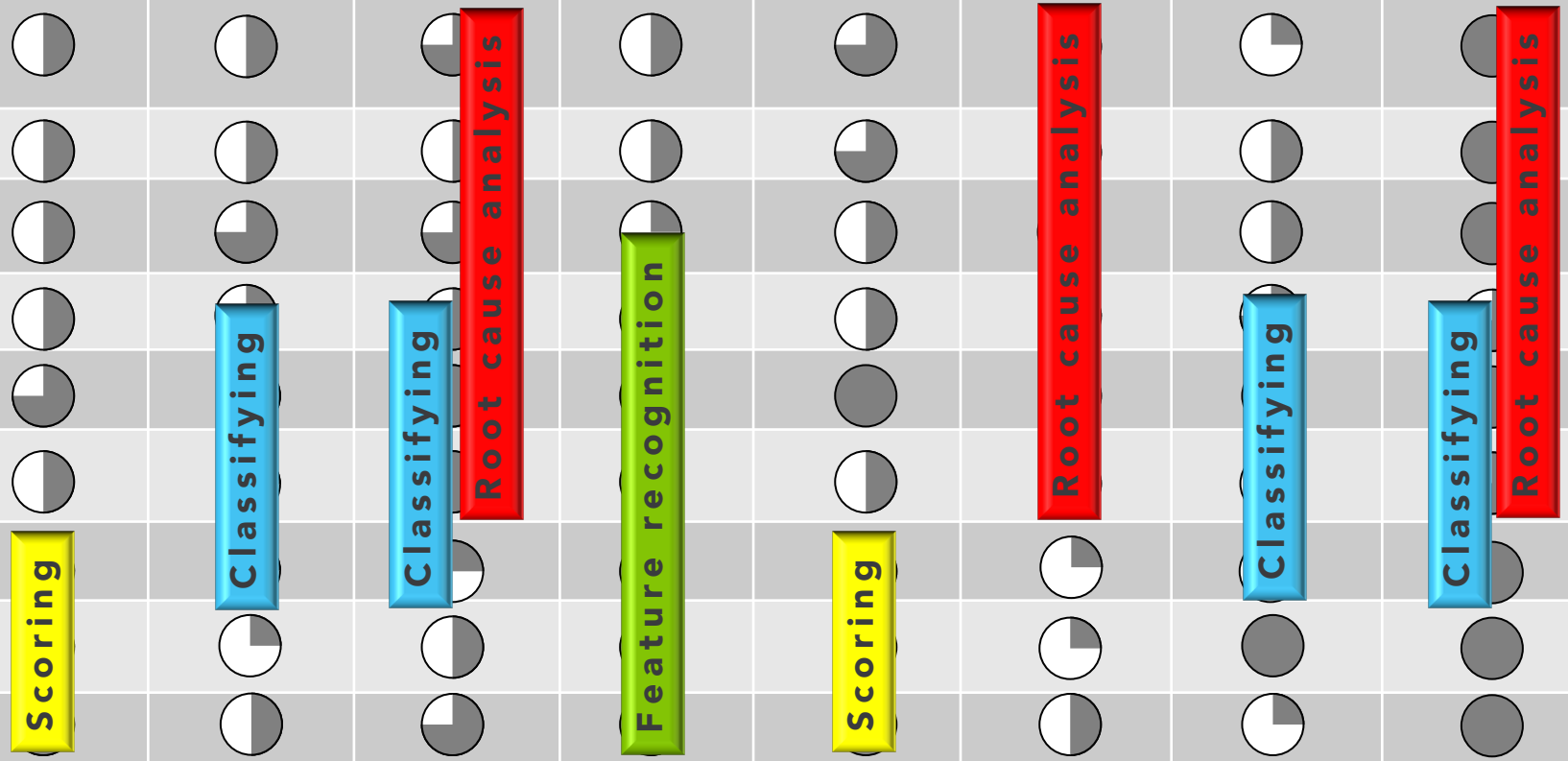
Forêts aléatoires

- **Advantages:**
  - Better predictive Power

- **Disadvantages:**
  - Loses the ability to interpret the rules
  - Difficulty to treat continuous variables without loosing information
  - Can hardly learn on data with missing information

Table de comparaison

Criteria	Logistic regression	Bayesian classifiers	Decision trees	Neural network	SVM	Assocation Rules	K-means	HyperCube
Can handle missing, discrete or continuous data								
Exhaustiveness								
Explanatory power (Causality)								
Noise resistance								
Predictive power								
Local phenomena detection								
Error rate advantage								
Ease of implementation								
Multi-variate scalability								



Distinctive Superior Average Inferior Remedial

Ethique et Impact social

- Remplacer les humains assurant des travaux à faible valeur ajoutée ?
- Remplacer les humains assurant des travaux à forte valeur ajoutée ?
- Déplacement de la valeur : révolution informatique, puis révolution numériques / AI
- Très volatile et transmissible
- AI contre AI – ex des nuages de drones
- Que reste-t il à l'homme ?
    - AI forts en prédiction, et l'explication ?
    - Intuition, imagination ?
    - Limites morales et éthiques, culture