

# Introduction au Machine Learning

Laure Delisle

Element AI

21 mars 2019

# Background

Laure Delisle – Research engineer

- | Element AI
  - | Lastline
  - | CEA
  - └ Airbus defence
- ⋮
- └ NDH Kids



# Machine Learning 101 - plan

## Machine learning

- définition
- ML vs Intelligence Artificielle / Deep Learning / Data Science
- taxonomie

## Process, données, vocabulaire

- données, data preparation
- training
- mesure de performance

## En pratique

- dataset NSL-KDD (classification, clustering)

# Objectifs pédagogiques

Machine learning

- └ Définition, taxonomie
- └ Algorithmes
- └ Techniques
- └ Mise en application

# Machine Learning (apprentissage machine)

Ensemble de techniques, reposant sur des **statistiques** et **algorithmes**, par lesquelles un **programme** informatique est capable d'**apprendre par expérience** à réaliser un ensemble de **tâches** sous contrainte d'une **mesure de performance**.

```
Offset(h) 00 01 02 03 04 05 06 07 08 09 0A 0B 0C 0D 0E 0F
00000000 4D 5A 90 00 03 00 00 00 04 00 00 00 FF FF 00 00
00000010 B8 00 00 00 00 00 00 00 40 00 00 00 00 00 00 00
00000020 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00000030 00 00 00 00 00 00 00 00 00 00 00 00 80 00 00 00
00000040 0E 1F BA 0E 00 B4 09 CD 21 B8 01 4C CD 21 54 68
00000050 69 73 20 70 72 6F 67 72 61 6D 20 63 61 6E 6E 6F
00000060 74 20 62 65 20 72 75 6E 20 69 6E 20 44 4F 53 20
00000070 6D 6F 64 65 2E 0D 0D 0A 24 00 00 00 00 00 00 00
00000080 50 45 00 00 4C 01 03 00 8D FA 81 4D 00 00 00 00
00000090 00 00 00 00 E0 00 02 01 0B 01 08 00 00 0A 00 00
000000A0 00 08 00 00 00 00 00 00 9E 28 00 00 00 20 00 00
000000B0 00 40 00 00 00 00 40 00 00 20 00 00 00 02 00 00
000000C0 04 00 00 00 00 00 00 00 04 00 00 00 00 00 00 00
000000D0 00 80 00 00 00 02 00 00 01 82 00 00 03 00 40 85
000000E0 00 00 10 00 00 10 00 00 00 10 00 00 10 00 00
000000F0 00 00 00 00 10 00 00 00 00 00 00 00 00 00 00 00
```

$$S(x) = \frac{1}{1 + e^{-x}}$$



# AI, DL, DS ?

## Intelligence artificielle

- ↳ Système capable de réaliser des tâches qui nécessite normalement une intelligence organique (prise de décision, perception visuelle, compréhension du langage...). <sup>[1]</sup>

## Deep Learning

- ↳ Sous-famille du Machine Learning, regroupant des algorithmes de **réseaux de neurones**, inspirés de la structure neuronale du cerveau.

## Data Science

- ↳ Ensemble de techniques pour **préparer, visualiser, analyser des données** pour en extraire des informations ou prendre des décisions.

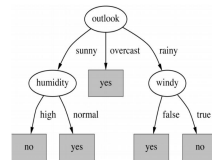
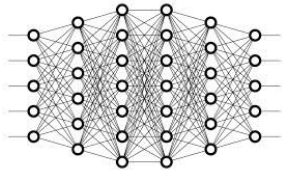
<sup>[1]</sup> From the Oxford Dictionary

# Interconnexion - ML, AI, DL, DS

Intelligence Artificielle

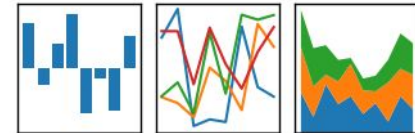
Machine Learning

Deep Learning



Data Science

data ingestion  
data preparation



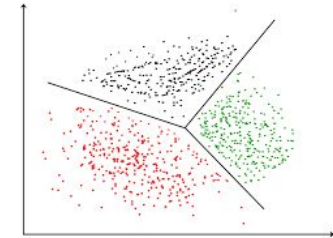
operational research  
classic computer vision  
classic pattern recognition  
...

# Taxonomie - Machine Learning

## Apprentissage supervisé

- └ données '**labellées**' (étiquetées)
- └ régression
- └ classification

var 1	...	var n	target
'blanc'	...	83.2	A
'vert'	...	47.5	B
' <b>blanc</b> '	...	<b>75.7</b>	<b>?</b>

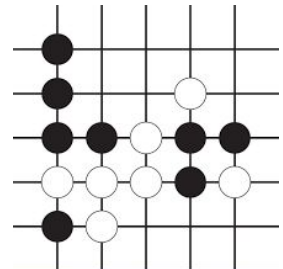


## Apprentissage non supervisé

- └ données **non 'labellées'**
- └ clustering

## Apprentissage par renforcement

- └ environnement + actions + **récompenses** + **agent autonome**
- └ end-to-end
- └ inverse
- └ par démonstration





# Catalystes et limites

## Historique

- régression: 1805 (Legendre), 1809 (Gauss)
- classification: 1955 (KMeans)
- clustering: 1990's (Kernel machines, Graphical models)
- deep learning: 1958 (Perceptron), 1986 (Backpropagation)

## Catalystes

- plus de **données**, dataset plus larges
- **modèles** plus complexes
- puissance de **calcul** (CPU, GPU, TPU)

## Limites

- données adéquates et préparées
- overfitting
- connaissance du domaine
- déploiement en production

# Process, méthodes, vocabulaire

## Données

- types
- manquantes
- déséquilibrées
- aberrantes

## Variables

- sélection (filtre, wrapper, embedded)
- engineering
- réduction de dimension

## Phases

- training
- validation
- testing

# Process, méthodes, vocabulaire

## Training

- loss
- descente de gradient
- vanishing/exploding gradient
- optimization

## Validation / testing

- cross-validation, validation/test, out-of-time
- mesures de performance
- compromis variance/biais
- overfitting, underfitting

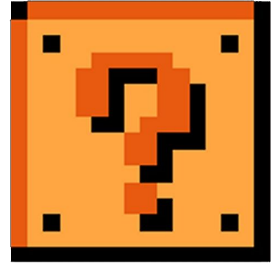
# Données - types

## Tabulaires

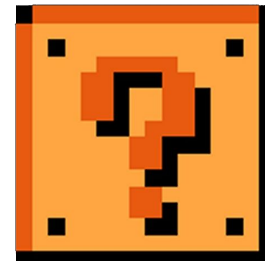
- | numériques
- | catégoriques
- | ordinales

## Séquentielles

- | temporelles
- | texte



# Données manquantes (missing)

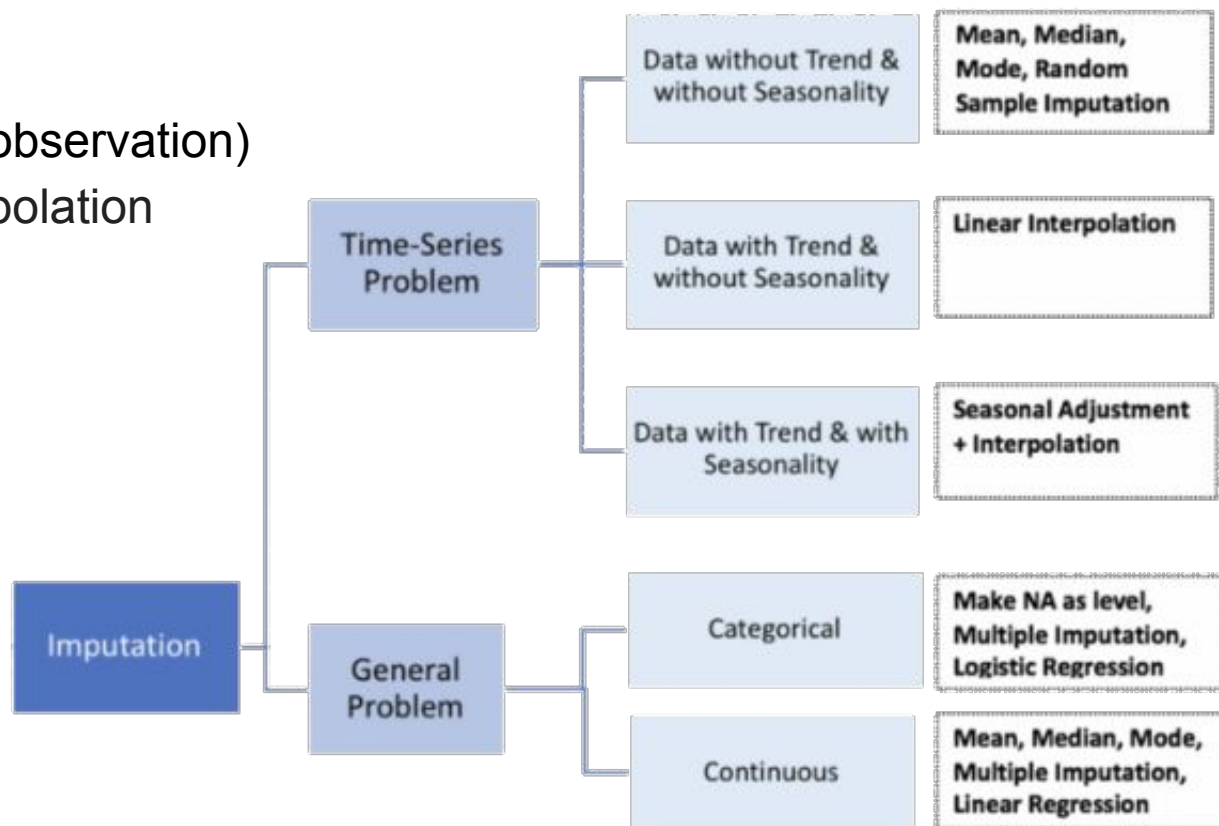


Pourquoi ?

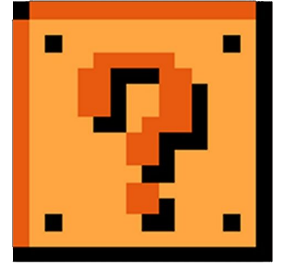
- └ au hasard
- └ conditionnellement à une autre variable

Stratégie

- └ suppression (col, observation)
- └ 'imputation' : interpolation



# Données aberrantes (outliers)



Pourquoi ?

- au hasard
- erreur de mesure
- conditionnellement à une autre variable

Stratégie

- suppression (col, observation)
- choix d'une méthode plus robuste

# Données déséquilibrées (unbalanced)

## Biais

- └ sous-représentation
- └ le modèle apprend davantage depuis la classe majoritaire

## Stratégie

- └ up-sampling la classe minoritaire
- └ down-sampling la classe majoritaire
- └ collecter davantage de données, changer de dataset

## Risk

- └ study by Joy Buolamwini, M.I.T. <sup>[1]</sup>

***Facial Recognition Is Accurate,  
if You're a White Guy***

[1] <https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html>

# Training, testing

## Apprentissage

- calcul d'erreur
- ajustement par descente du gradient

## Test - mesure de performance

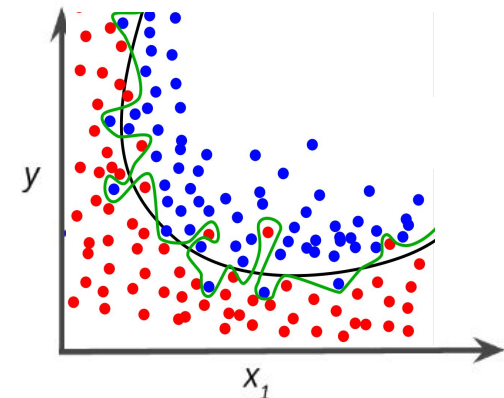
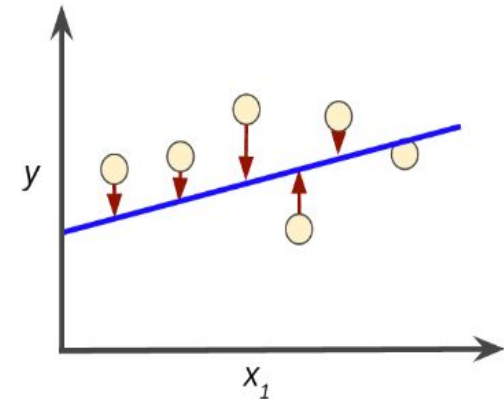
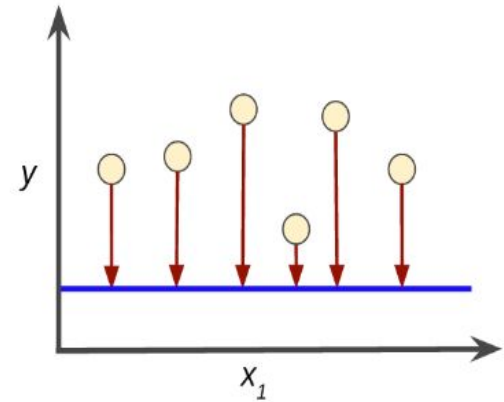
- données non utilisées pour l'apprentissage
- matrice de confusion (accuracy/precision/recall/F1)
- SSE,  $R^2$

## Validation

- 3eme jeu de données
- pour ajuster des hyperparamètres

## Risques

- overfitting, underfitting
- exploding gradient





# Régression

## Contexte

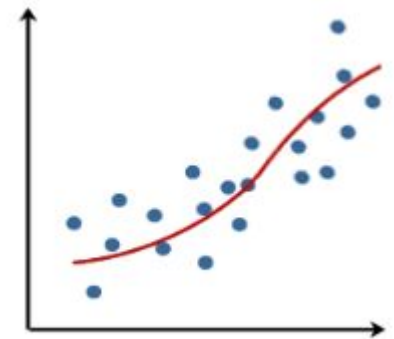
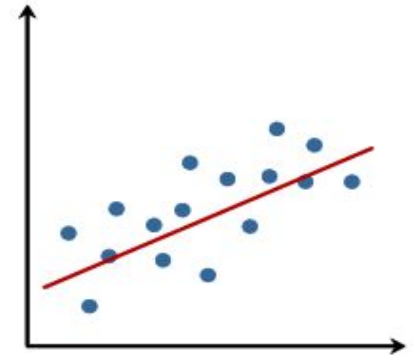
- └ données labellées (apprentissage **supervisé**)
- └ label **numérique**

## Modèle

- └  $Y = f(\mathbf{X}, \boldsymbol{\beta})$
- └  $\mathbf{X}$  : variables
- └  $\boldsymbol{\beta}$  : paramètres du modèles
- └  $Y$  : variables

## Mesure de performance

- └ SSE,  $R^2$



# Classification

## Contexte

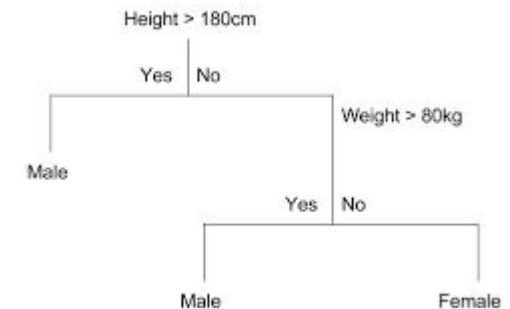
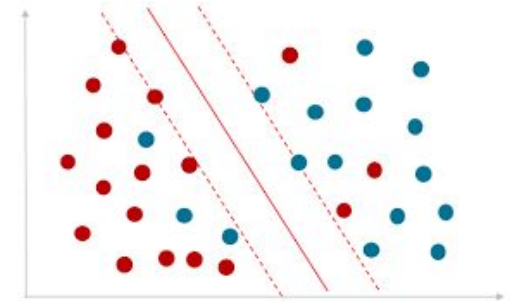
- données labellées (apprentissage **supervisé**)
- label **catégorique**

## Modèles

- régression logistique, Naive Bayes
- arbres de décision, random forest
- réseaux de neurones

## Mesure de performance

- matrice de confusion
- accuracy, precision, recall, F1



# TP - classification

## Objectifs

- load le dataset NSL-KDD (intrusion detection)
- préparation des données
- train / test split
- fit du modèle
- prédiction et évaluation

## Méthodes mise en oeuvre

- type de ML : supervisé > classification
- algorithme : Decision Tree / XGBoost
- techniques : one-hot encoding, standardization
- metric : confusion matrix

## Dataset

- <https://plg.uwaterloo.ca/~gvcormac/treccorpus07/>

# Clustering

## Contexte

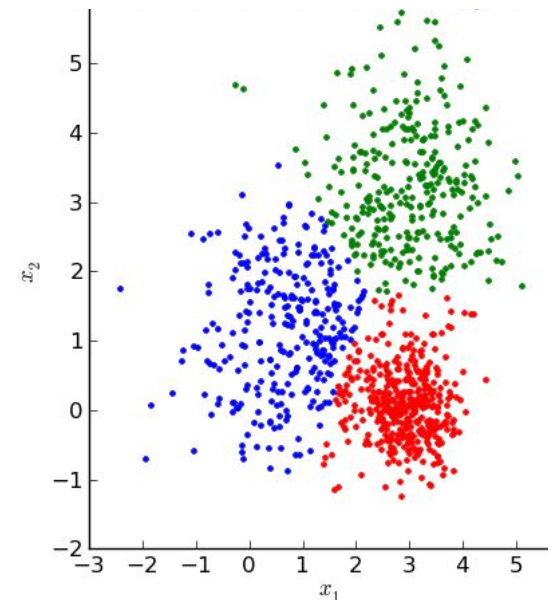
- └ données non labellées (apprentissage **non-supervisé**)
- └ label **catégorique**

## Modèles

- └ kNN, k-means
- └ hiérarchique, locally sensitive hashing
- └ density-based

## Mesure de performance

- └ homogénéité, completeness
- └ silhouette
- └ critère de variance ratio (Calinski-Harabaz)



# TP - clustering

## Objectifs

- load le dataset NSL-KDD (intrusion detection)
- préparation des données
- expérimentation avec différents modèles
- évaluation de la qualité des clusters

## Méthodes mise en oeuvre

- type de ML : non-supervisé > clustering
- algorithme : K-means, DBSCAN
- metric : silhouette, homogénéité

# General resources

## Datasets:

- <http://archive.ics.uci.edu/ml/index.php> (many open datasets to practice)

## Apprentissage par renforcement

- <http://incompleteideas.net/book/RLbook2018.pdf> (bible du domaine)

## Deep Learning

- <https://www.deeplearningbook.org/> (bible, un peu datée (2016))

# Régression (2)

## Hypothèses

- Y est une variable continue
- les variables X sont linéairement indépendantes
- les observations sont indépendantes (erreurs non corrélées)

## Régression linéaire

- polynome (2D)  $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n$
- multivariate  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n = \boldsymbol{\beta} \mathbf{x}$  (with  $x_0 = 1$ )