Big Data
Open Data

MMI 2 – TP#6 S4

IUT CERGY-PONTOISE

Danielo **JEAN-LOUIS**
Développeur front-end

Le Machine Learning c'est quoi ?

Machine Learning

- Apprentissage automatique en français
- Abrégé ML
- Sous-branche de l'Intelligence Artificielle
- C'est la machine qui décide
  - A partir de données et d'algorithmes définis
- Nécessite un certain nombre de jeu de données

Machine Learning

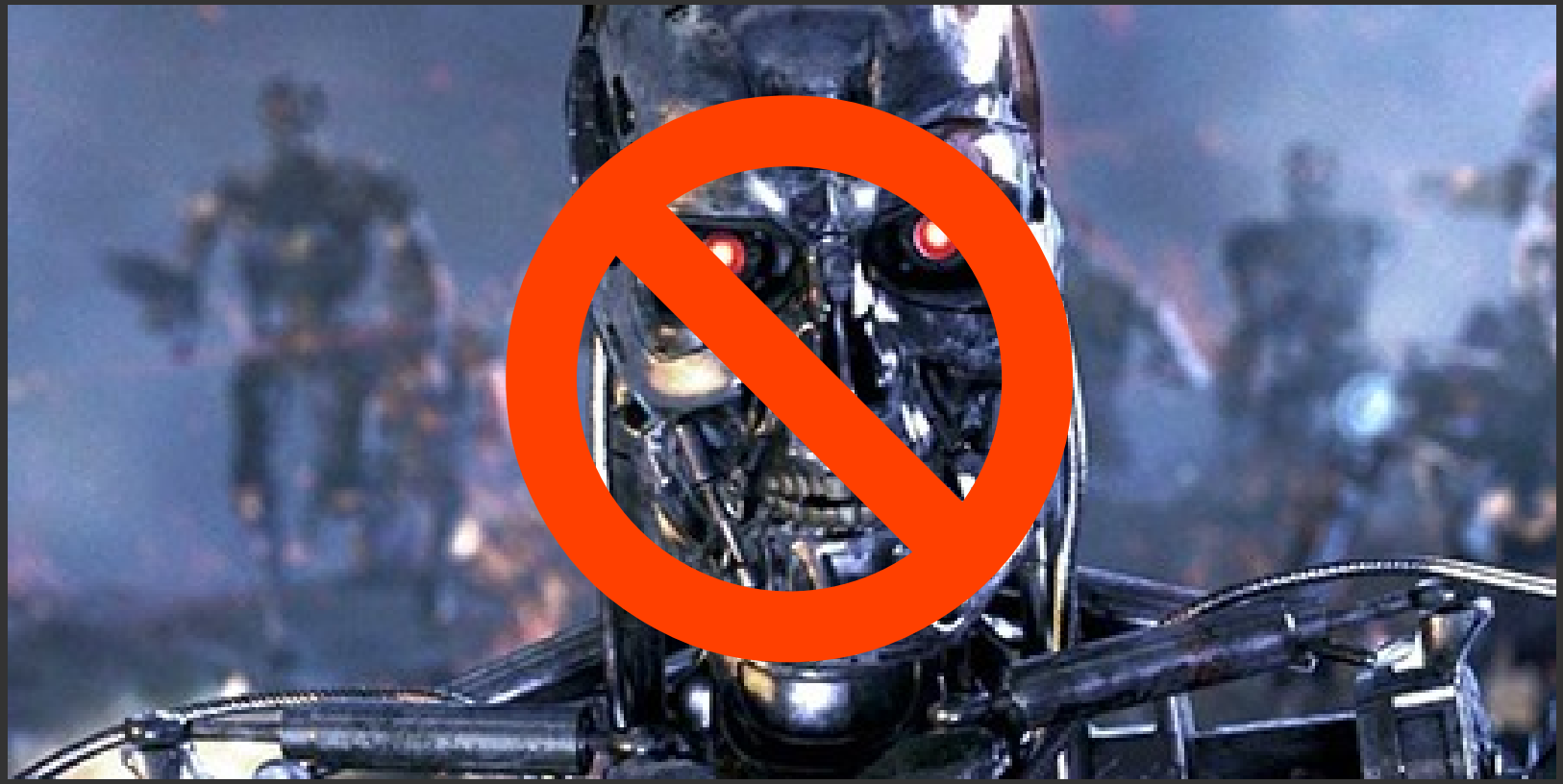
- Deep Learning
  - Petit frère du ML
- NLP = Natural Langage Processing
  - Branche à part du Machine Learning

Machine Learning

- Trois grandes catégories :
  - **Apprentissage supervisé**
  - **Apprentissage non-supervisé**
  - Apprentissage par renforcement
- Reste relativement limité
  - Puissance / Données

Sources :

- https://larevueia.fr/apprentissage-par-renforcement/



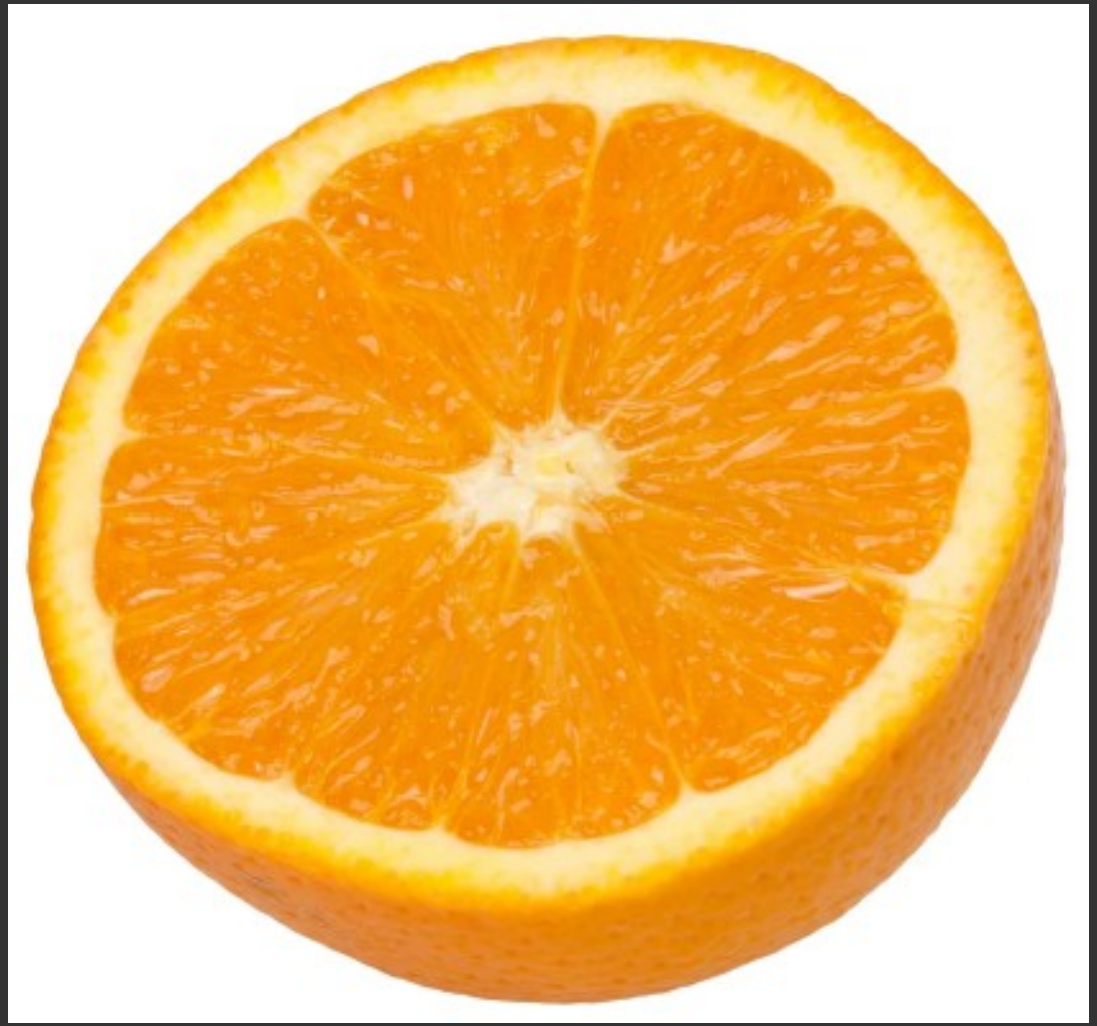
Le soulèvement des machines, ce n'est pas pour demain

Machine Learning - Applications

- Voitures autonomes
- Recommandations (netflix, amazon…)
- Détection de fraudes
- **Prédiction** des prix de l'immobilier
- ...

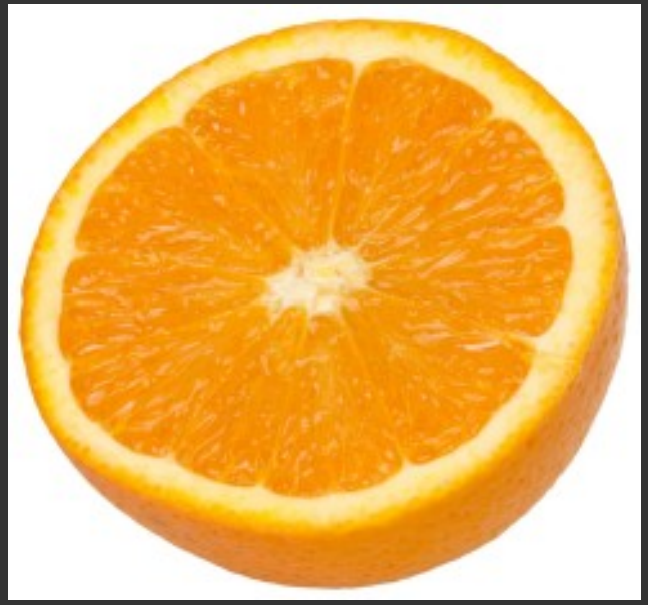
Modèle

- Représentation simplifiée de la réalité
- Représentation mathématique de relation entre des données
- "Fonction mathématique" en résumé



Apprentissage supervisé

- Les données sont libellées



(Une) Orange

Apprentissage supervisé

- Deux types :
  - Classification
  - Régression
- Nécessite une intervention humaine
- A besoin d'exemple**s** pour s'entraîner

Sources :

- http://www.vincentlemaire-labs.fr/cours/2.1-ApprentissageSupervise.pdf

Apprentissage supervisé – Préparation

- Phase permettant de dégager des features (caractéristiques/dimensions/paramètres)
- Se poser les bonnes questions
  - *Un problème bien posé est à moitié résolu*

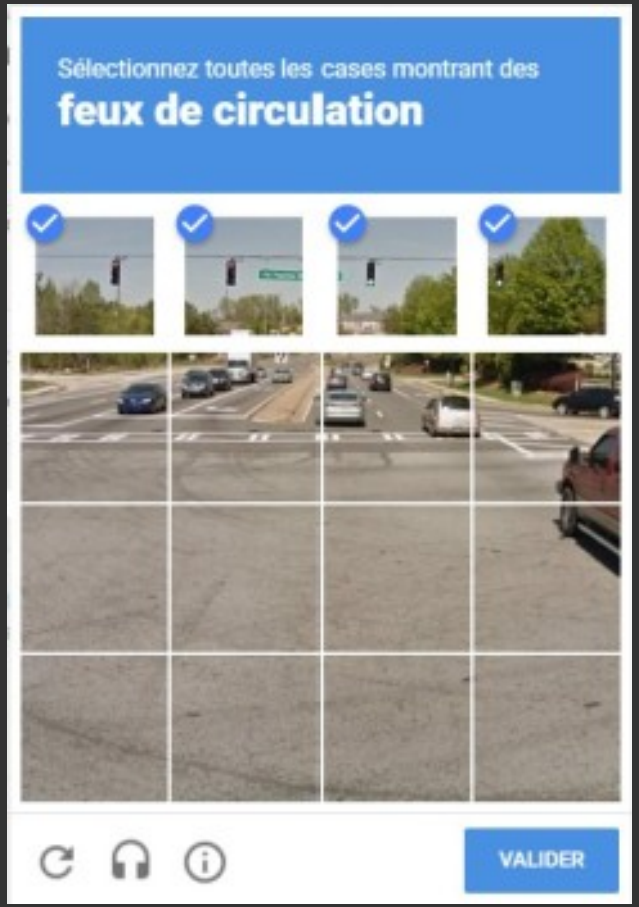
Apprentissage supervisé - Classification

- S'utilise pour les valeurs discrètes ou continues
  - Exemple : classement d'images
- Classes binaires ou multiples
- Exemple d'algorithmes :
  - Régression logistique (Logistic Regression)
  - k plus proches voisins (K-Nearest Neighbor)
  - Forêt d’arbres décisionnels (Random Forest)

Sources :

- https://moncoachdata.com/blog/modeles-de-machine-learning-expliques
- https://larevueia.fr/algorithme-du-plus-proche-voisin/

Apprentissage supervisé - Classification



Les captchas sont des moyens communautaires de classifier (et libeller) des données

Apprentissage supervisé - Classification

Chien ou bagel ?

Il est important de montrer plusieurs exemples pour entraîner le modèle

Sources :

- https://twitter.com/teenybiscuit/status/707004279324696577

Apprentissage supervisé - Régression

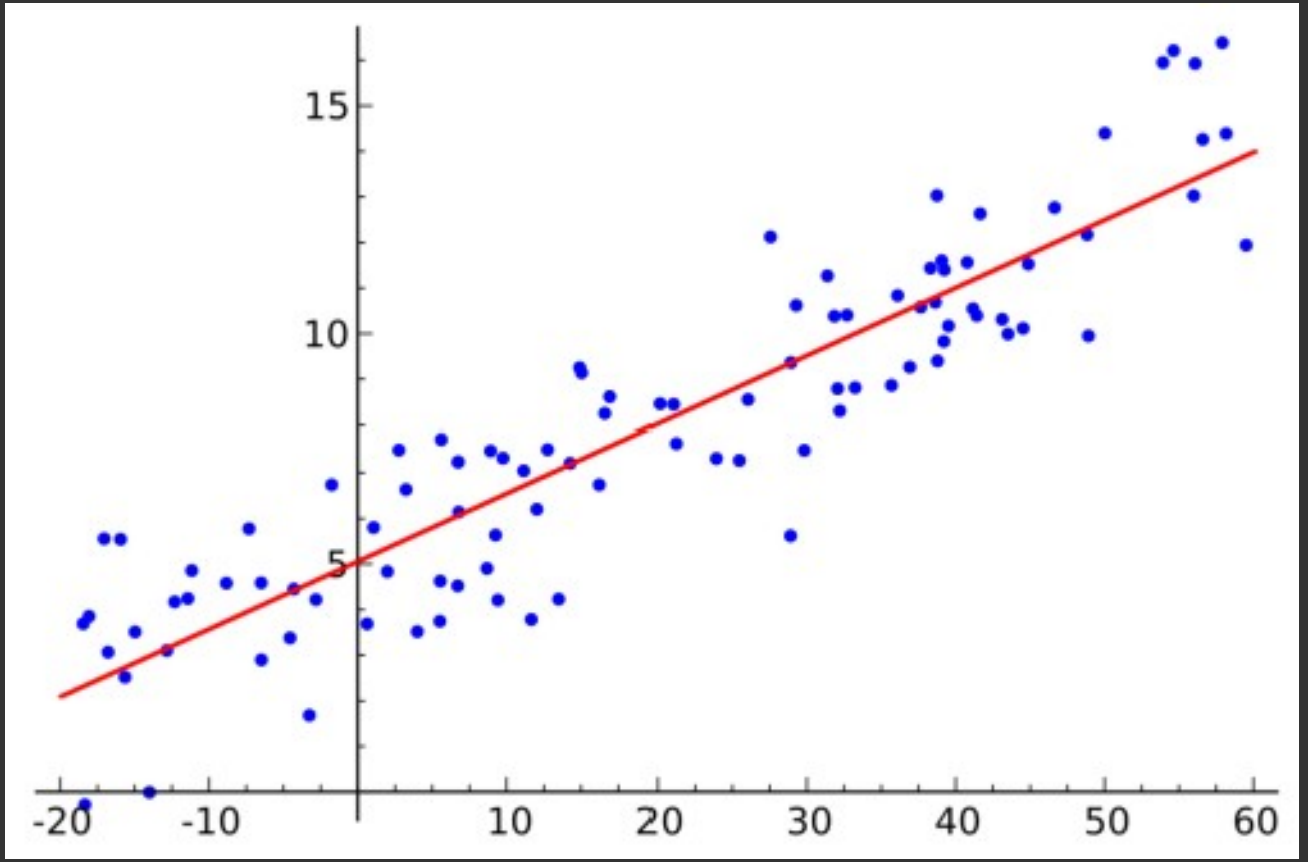
- S'utilise pour les valeurs continues (des nombres)
- Prédit la valeur de sorties à partir des valeurs d'entrées (features)
  - Exemple : prix de l'immobilier
- Exemple d'algorithmes :
  - Régression linéaire (Linear Regression)
  - Lasso Régression

Sources :

- https://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire

Apprentissage supervisé - Régression

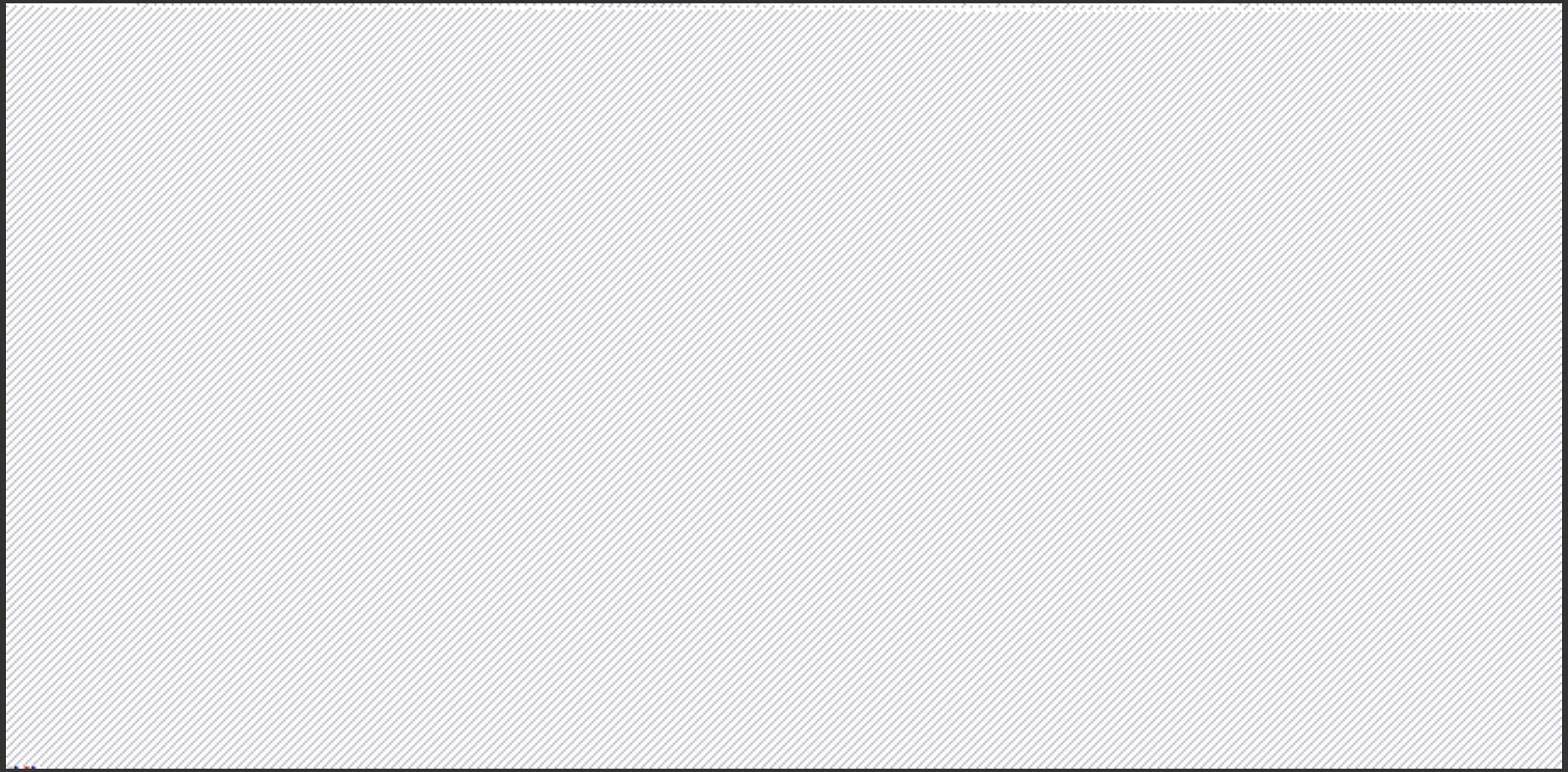
Prix habitation (€)	Surface (m²)
150 000	68
178 000	75
…	…



Sources :

- https://fr.wikipedia.org/wiki/R%C3%A9gression_lin%C3%A9aire

Apprentissage supervisé – Régression / Classification



Apprentissage supervisé – Attention au overfitting

- Surapprentissage en français
- Apparaît plus en Régression
- Important d'avoir des données disparates
- Multiplier les tests

Sources :

- https://fr.wikipedia.org/wiki/Surapprentissage
- https://larevueia.fr/7-methodes-pour-eviter-loverfitting/

Apprentissage supervisé - Validation

- Ne pas oublier de tester son modèle
- 80/20.
  - 80 % des données servent à l'entraînement du modèle
  - 20 % servent de test

Apprentissage non-supervisé

- Le contenu n'est pas libellé, l'algorithme va trouver lui-même les similarités, les liens
  - Exemple : https://cs.stanford.edu/people/karpathy/cnnembed/cnn_embed_6k.jpg
- Découvrir une tendance
  - Tendance peut changer en fonction de l'algorithme utilisé

Sources :

- https://datascientest.com/apprentissage-non-supervise
- https://dataanalyticspost.com/Lexique/apprentissage-non-supervise/
- http://www.vincentlemaire-labs.fr/cours/2.2-ApprentissageNonSupervise.pdf

Apprentissage non-supervisé

- Exemples d'applications :
  - Génération d'images
  - Détection de fraudes
  - Création d'articles

Sources :
- https://thispersondoesnotexist.com/

Apprentissage non-supervisé

- Exemple de types de familles d'algorithmes :
  - Groupement (clustering)
  - Association
    - Compréhension par contexte
- **Attention : méthode dangereuse**

Sources :

- https://datascientest.com/apprentissage-non-supervise
- https://dataanalyticspost.com/Lexique/apprentissage-non-supervise/
- http://www.vincentlemaire-labs.fr/cours/2.2-ApprentissageNonSupervise.pdf

Un ordinateur n'a pas de morale

- Elle exécute sans réfléchir aux conséquences
- Peut conduire à des dérives. Exemples :
  - Chambre d'écho (problème de fixation)
  - Le recrutement chez Amazon
- Nécessite une validation humaine (devrait)
- RGPD

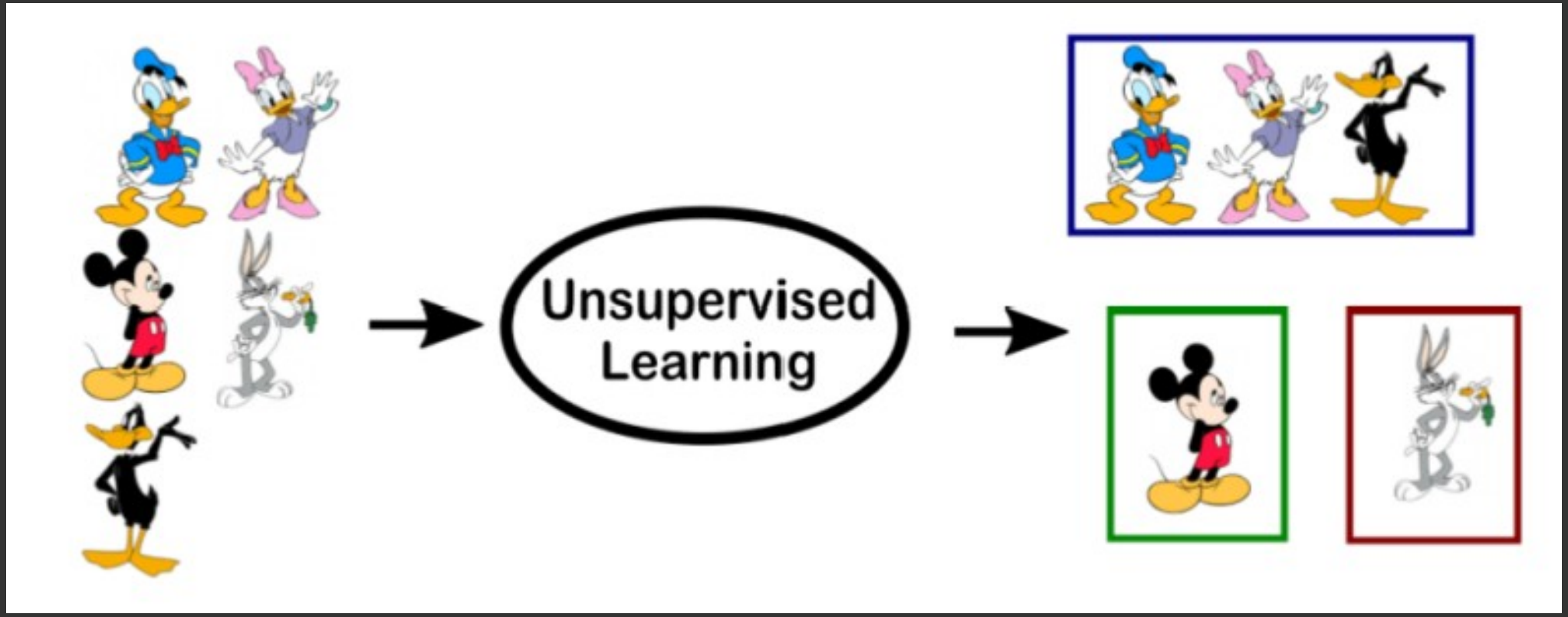
Sources :

- https://fr.wikipedia.org/wiki/R%C3%A8glement_g%C3%A9n%C3%A9ral_sur_la_protection_des_donn%C3%A9es
- https://larevueia.fr/les-5-plus-gros-fails-de-lintelligence-artificielle/
- https://larevueia.fr/le-machine-learning-pour-les-systemes-de-recommandations
- https://www.youtube.com/watch?v=DKvV1S3B4Uc

Apprentissage non-supervisé - Groupement

- Cherche à définir des groupes homogènes (infinis)
  - Exemple : Type d'acheteurs
- Nécessite une intervention humaine en aval
- Exemple algorithme :
  - K-moyennes (K-means)

Apprentissage non-supervisé - Groupement



L'algorithme a groupé les images tout seul

Apprentissage non-supervisé - Association

- Cherche à découvrir des associations entre éléments
  - Exemple : Contenu d'un caddie de supermarché
- Nécessite une intervention humaine en aval
  - Exemple : revoir l'agencement d'un magasin
- Exemple algorithme :
  - APriori (K-means)

Sources :

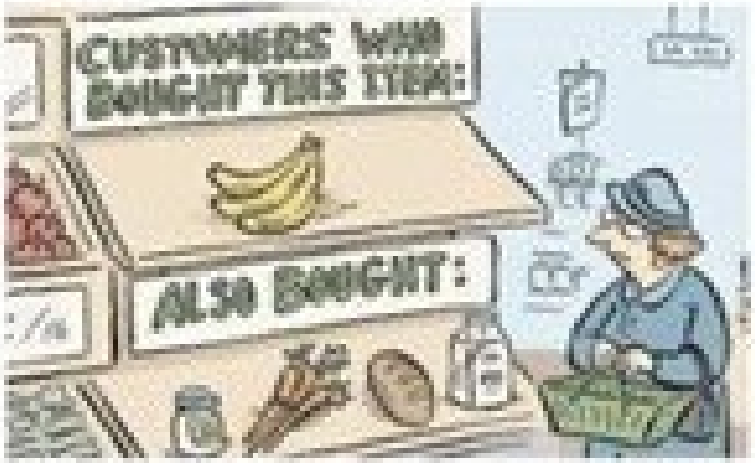
- https://fr.wikipedia.org/wiki/Algorithme_APriori

Apprentissage non-supervisé - Association

Association

People that buy X tend to buy Y

People that buy A+B tend to buy C

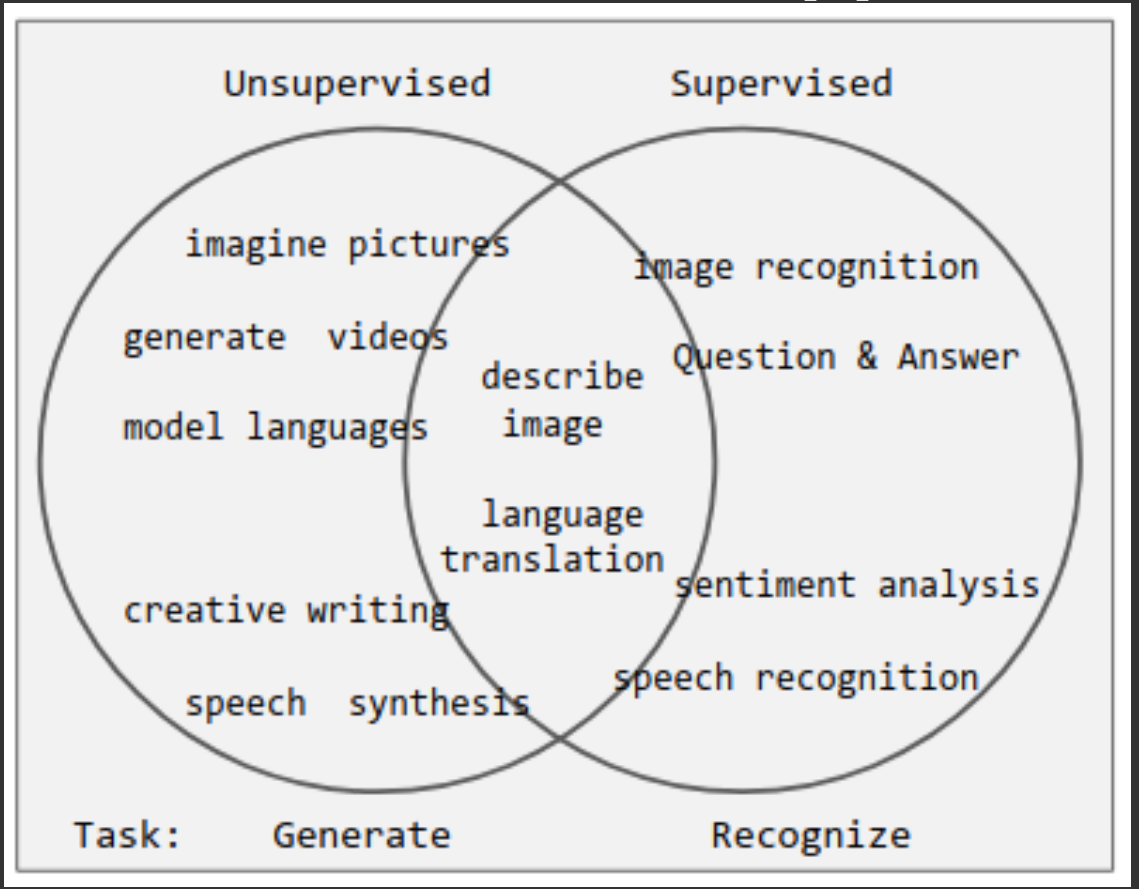


Apprentissage non-supervisé – Association
Corrélation != Causalité

- Ne pas se fier aveuglément aux résultats du modèle
- Une corrélation est trouvée pas forcément une causalité

Sources :
- http://www.tylervigen.com/spurious-correlations

En résumé – Panorama des applications



Sources :

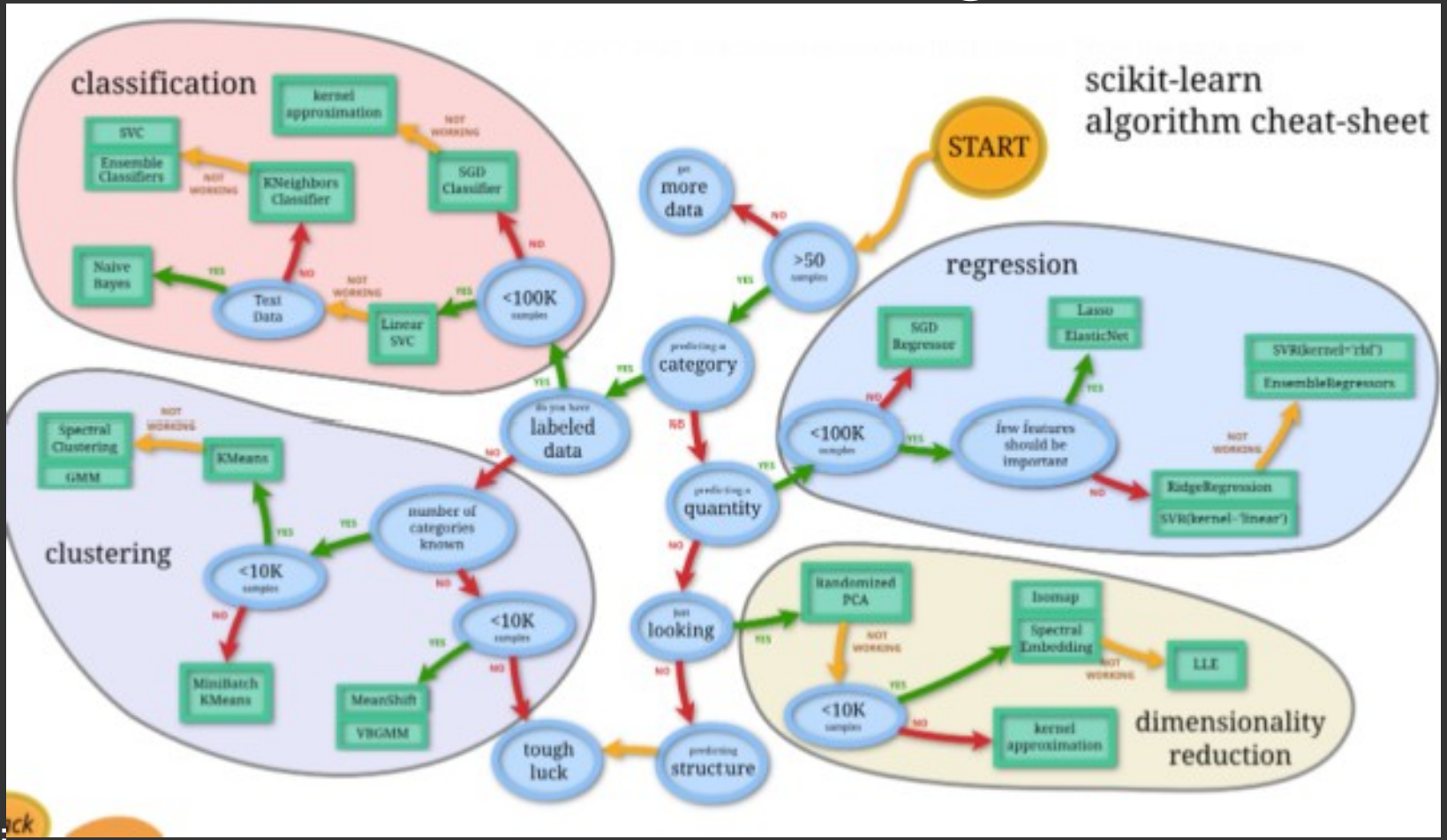
- https://en.wikipedia.org/wiki/Data_mining

En résumé – Types d'apprentissages

Supervisé → Effectuer une tâche

Non-supervisé → Découvrir quelque chose

En résumé – Panorama des algorithmes



Sources :

- https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html - anglais
- https://docs.microsoft.com/fr-fr/azure/machine-learning/algorithm-cheat-sheet - anglais

Reconnaissance d'images

Grouper des clients

Google Translate

Reconnaissance d'écriture

FaceID

Voitures autonomes

Alexa

Recommendation Netflix

Prédiction du salaire

Détection de SPAM

Suggestion d'achat

Les gauchers achètent des ciseaux pour gauchers

Mes chances de survivre le 14 avril 1912

Reconnaissance d'images

Supervisé

Grouper des clients

Non-Supervisé

Google Translate

Les deux !

Reconnaissance d'écriture

Supervisé

FaceID

Supervisé

Voitures autonomes

Les deux !

Alexa

Les deux !

Recommendation Netflix

Non supervisé

Prédiction du salaire

Supervisé

Détection de SPAM

Supervisé

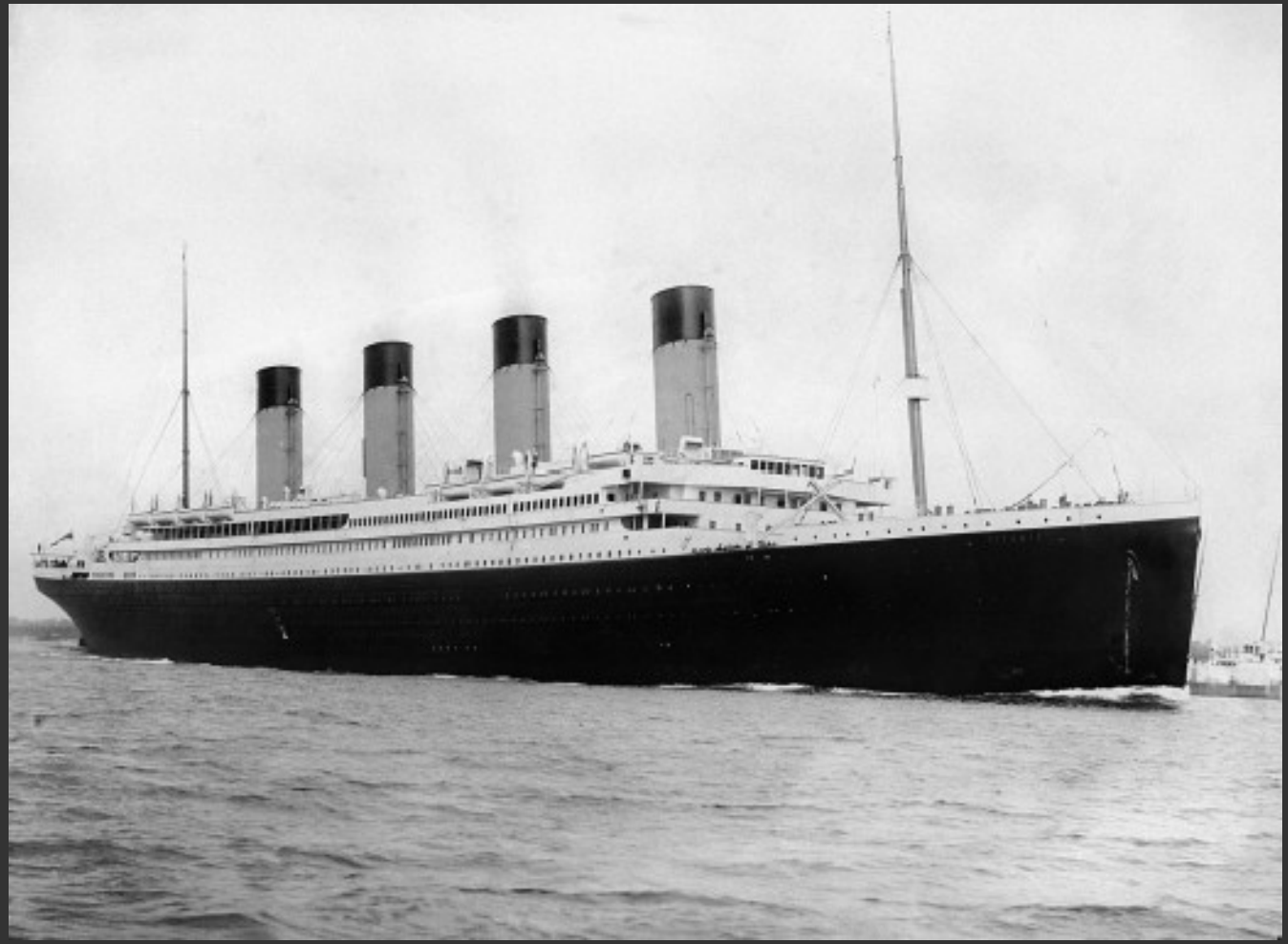
Suggestion d'achat

Non supervisé

Les gauchers achètent des ciseaux pour gauchers

Non supervisé

Mes chances de survivre le 14 avril 1912



Le Royal Mail Ship Titanic

Titanic

- Mis en service le 10 avril 1910
- Heurte un iceberg le 14 avril 1912
- 1 490 et 1 520 personnes trouvent la mort
- "Hello world" du Machine Learning… supervisé

Scikit learn

- Implémenté dans jupyer notebook et colaboratory
- Permet d'utiliser des algorithmes (supervisés ou non) de machine learning
- Contient des jeux de données par défaut pour tester
  - Fleur d'iris : autre classique de la datascience

Sources :
- https://scikit-learn.org

Kaggle

- Site gratuit nécessitant une inscription
- Concours de datascientifique
- Propose un nombre conséquent de jeux de données réalisés par la communauté
- Propose une interface proche de google colab

Sources :
- https://www.kaggle.com/

Pratiquons ! - Titanic

Pré-requis :

- Avoir la ressource ressources/titanic
- Lien : https://downgit.github.io/#/home?url=https://github.com/DanYellow/cours/tree/main/big-data-s4/travaux-pratiques/numero-6/ressources/titanic

Questions ?