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 données (Big Data)

Machine Learning

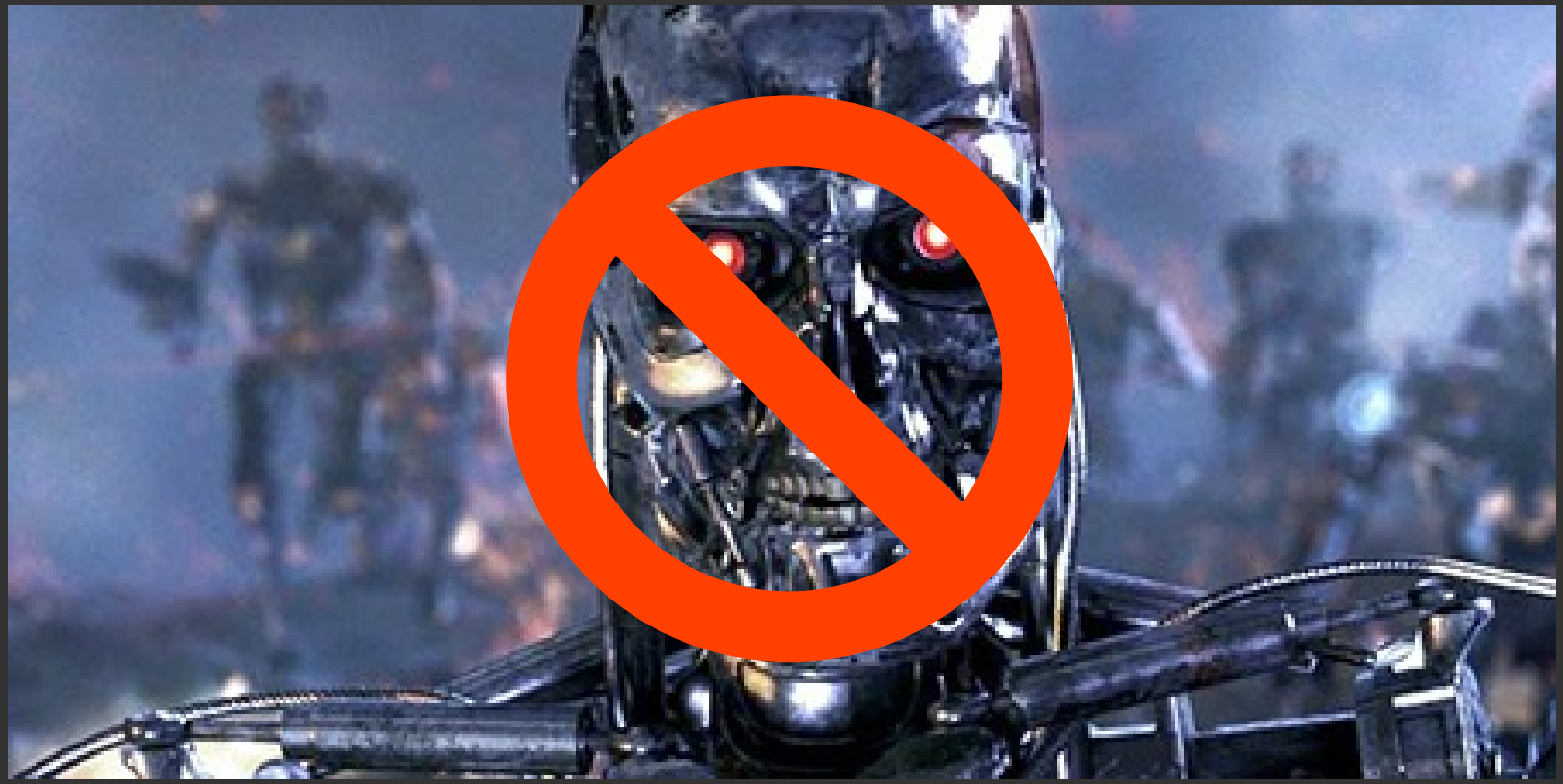
- Branche appliquée des statistiques
- Deep Learning
  - Petit frère du ML
- NLP = Natural Langage Processing
  - Branche à part du Machine Learning
- Intervient quand l'être humain montre ses limites

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
- ...

Machine Learning – But

Prévoir les valeurs de sorties à partir d'attributs (features/colonnes) grâce à l'application d'un modèle choisi

Machine Learning - Etapes

1. Définition du problème à résoudre
2. Acquisition des données d’apprentissages et de tests
3. Analyser, explorer les données
4. Préparer et nettoyer les données
5. Choisir un modèle d’apprentissage
   - Savoir quel problème on cherche à résoudre
6. Visualiser les résultats, et ajuster ou modifier le modèle d’apprentissage
7. Tester en production

Modèle

- Représentation simplifiée de la réalité
- Représentation mathématique de relation entre des données
- équation mathématique en résumé
- Performance par un score
  - Pouvant être de plusieurs types
  - Sujet à une interprétation qui lui est propre dépendamment des cas

Variables

- Influencent l'algorithme qui va être utilisé
- Peuvent être de plusieurs types
  - Quantitatives (nombre)
  - Qualitative (pas un nombre)

Sources :

- https://openclassrooms.com/fr/courses/4525266-decrivez-et-nettoyez-votre-jeu-de-donnees/4725615-decouvrez-les-4-types-de-variables
- https://www.stat.berkeley.edu/~stark/SticiGui/Text/histograms.htm - anglais

Variables

- Quantitatives (nombre) :
  - Discrète : la valeur de la variable est finie. Ex : âge
  - Continue : la variable peut prendre une infinité de valeurs. Ex : temps

Variables

- Qualitatives (ou catégorielles) (pas un nombre) :
  - Ordinale : La variable peut-être ordonnée. Ex : 1er, 2ème...
  - Nominale : La variable ne peut pas être ordonnée. Ex : les couleurs
  - Dichotomiques : La valeur ne peut avoir que deux états. Ex : Vrai / Faux

Variables – Quantitative ou Qualitative ?

(nombre ou texte)

Modèle d'une voiture	**Code postal**	**État d'un interrupteur**
Temps d'une course	**Nombre de questions dans le prochain test**	**Nombre de personnes dans ce TP**
Vitesse d'un véhicule	**Votre heure de réveil**	**Température du jour**

Variables – Quantitative ou Qualitative ?

(nombre ou texte)

Modèle d'une voiture
Qualitative

Code postal
Qualitative

État d'un interrupteur
Qualitative

Temps d'une course
Quantitative

Nombre de questions dans le prochain test
Quantitative

Nombre de personnes dans ce TP
Quantitative

Vitesse d'un véhicule
Quantitative

Votre heure de réveil
Quantitative

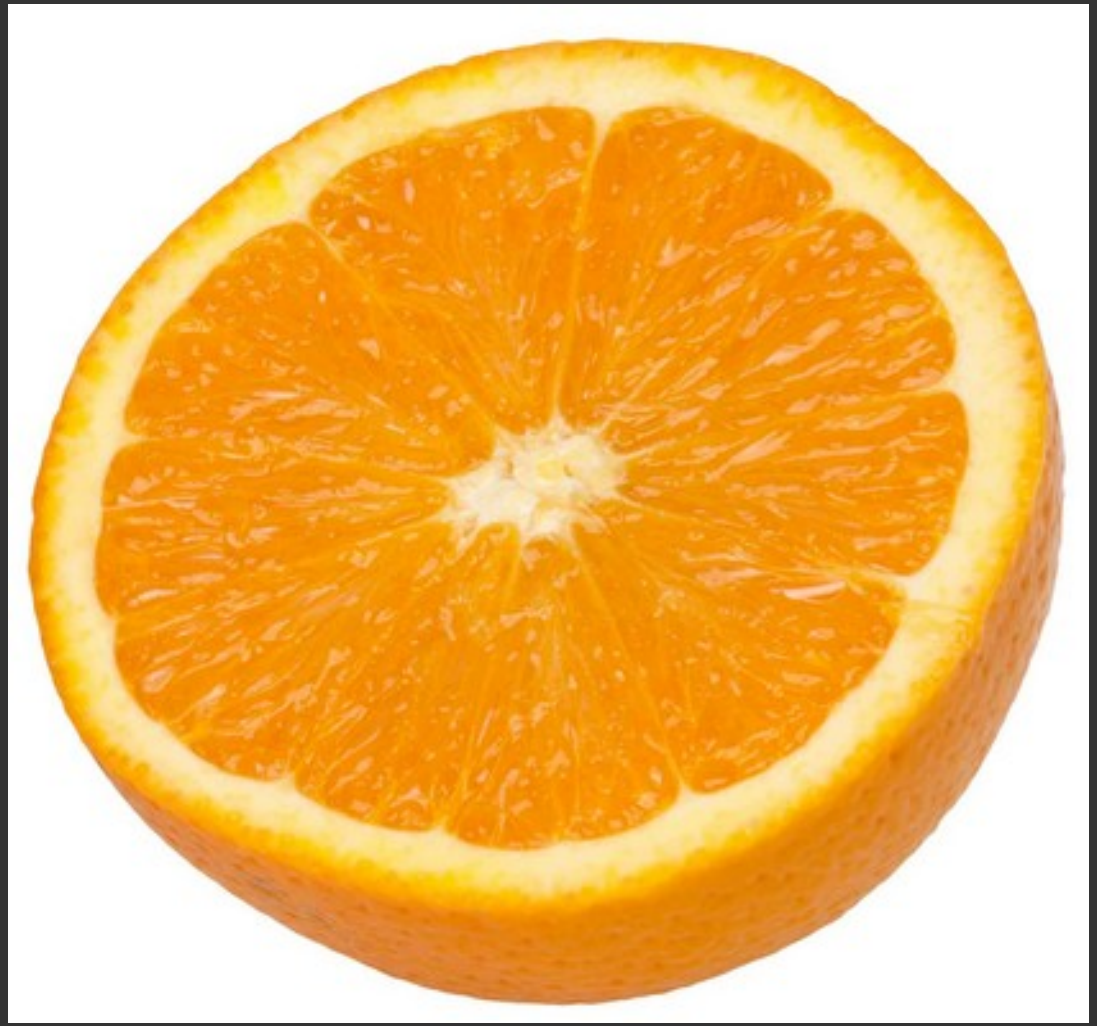
Température du jour
Quantitative

Code postaux et numéro de téléphone

- Variables qualitatives malgré la présence de numéros
- Faire des opérations sur ces données n'a aucun sens

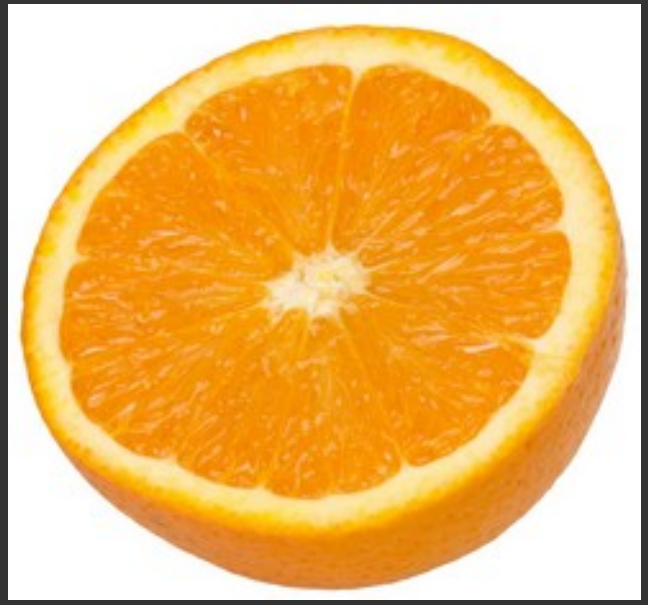
Variables – Quantitative ou Qualitative ?

- Qualitative :
  - Je suis rapide et je suis grand
- Quantitative :
  - Je cours 100 m en 13.42s et je mesure 180 cm



Apprentissage supervisé

- Les données sont libellées



(Une) Orange

Apprentissage supervisé

- Deux types :
  - Classification
  - Régression
- A besoin d'exemple**s** pour s'entraîner
  - Un être humain doit intervenir au début

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

Apprentissage supervisé - Dataset

Year	Liquid fuel	Solid fuel	Gas fuel	Cement production	Gas flaring
2010	3,107	3,812	1,696	446	67
2011	3,134	4,055	1,756	494	64
2012	3,200	4,106	1,783	519	65
2013	3,220	4,126	1,806	554	68
2014	3,280	4,117	1,823	568	68

Entrées

Sorties

Sources :

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

Apprentissage supervisé – Préparation

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

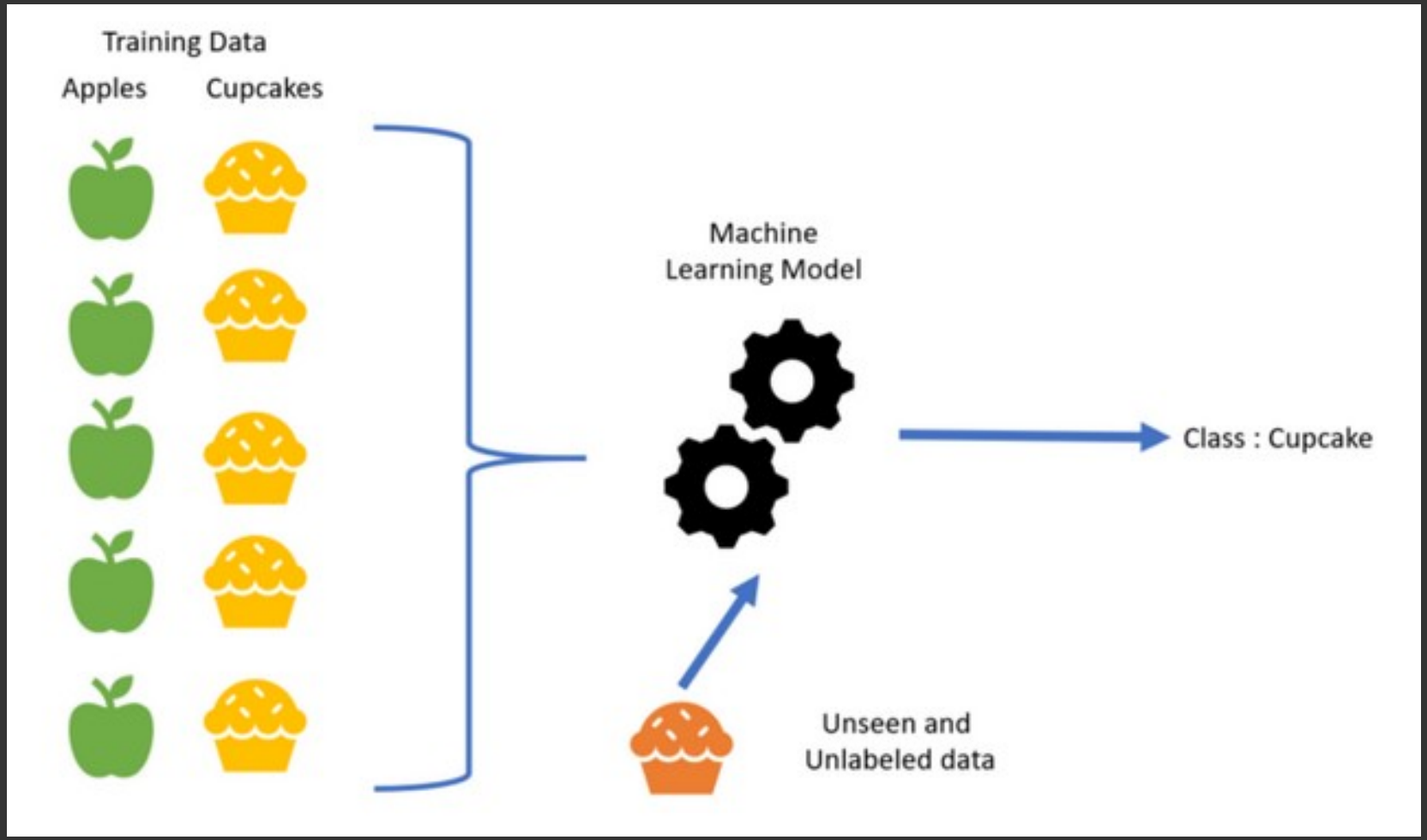
Apprentissage supervisé - Classification

- S'utilise pour les valeurs qualitatives ou quantitatives
  - Exemple : classement d'images
- Classes binaires ou multiples
- Exemple d'algorithmes (liste non-exhaustive) :
  - Régression logistique (Logistic Regression)
  - k plus proches voisins (K-Nearest Neighbor)
  - Forêt d’arbres décisionnels (Random Forest)
  - Boosting de gradient (Gradient boosting)

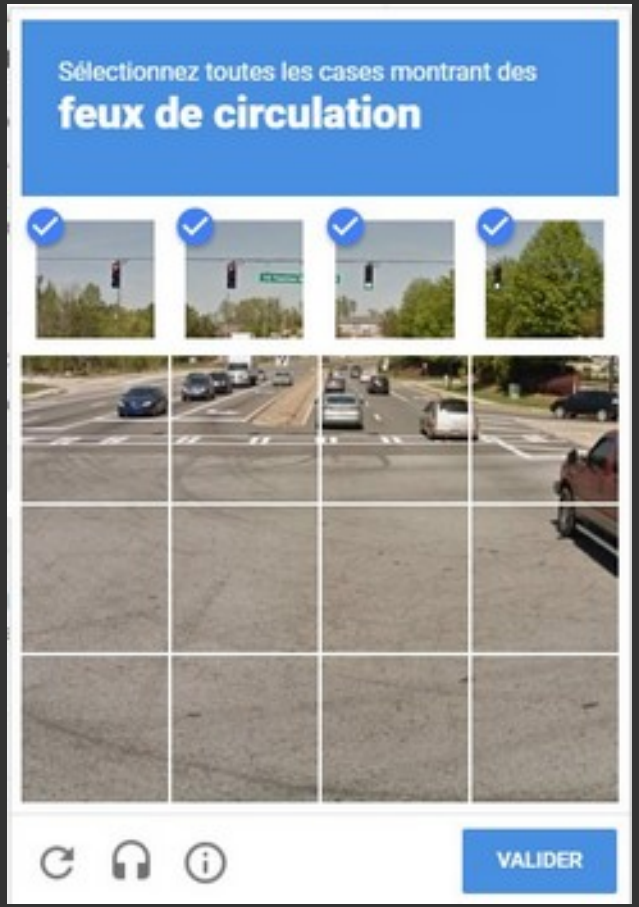
Sources :

- https://moncoachdata.com/blog/modeles-de-machine-learning-expliques
- https://larevueia.fr/algorithme-du-plus-proche-voisin/
- https://datascience.eu/fr/apprentissage-automatique/gradient-boosting-ce-que-vous-devez-savoir/

Apprentissage supervisé - Classification



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

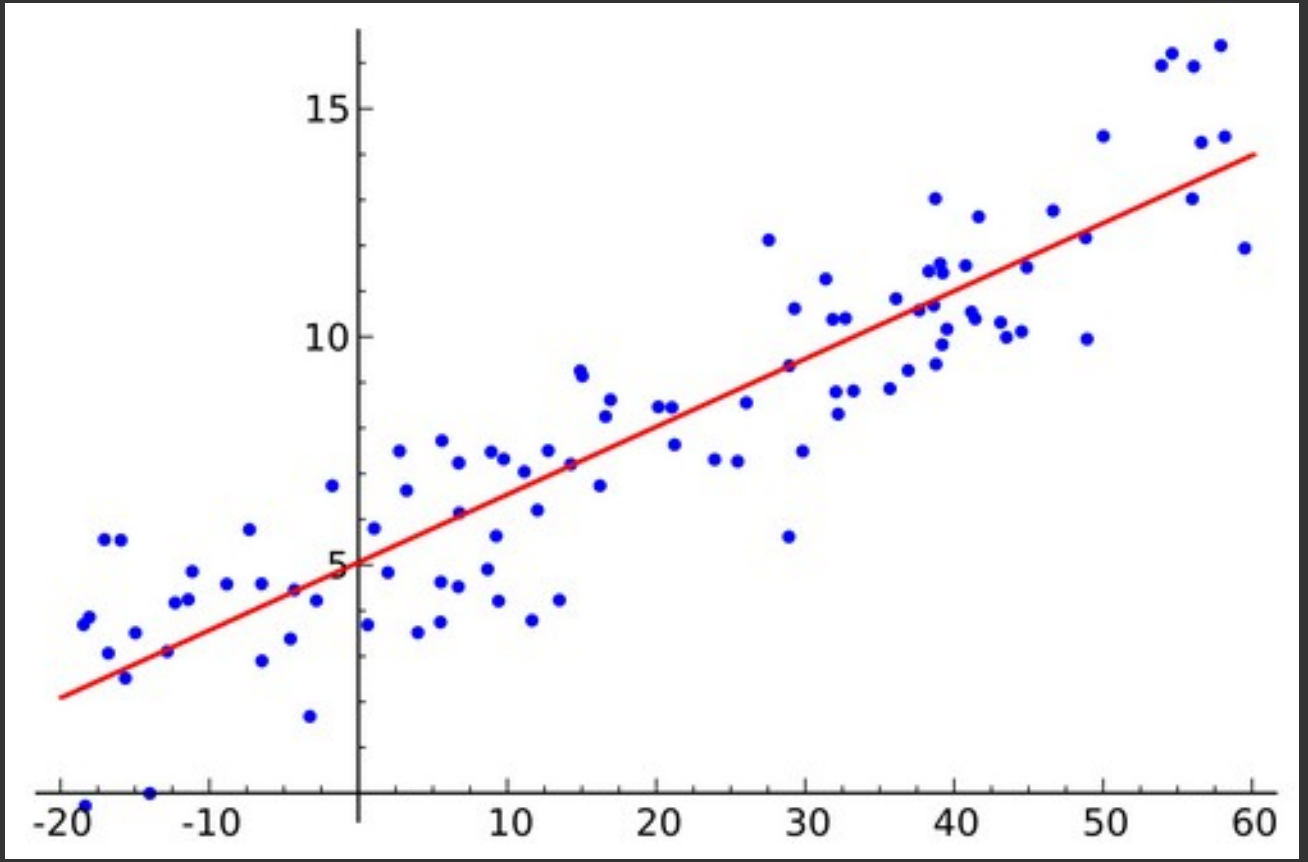
- S'utilise pour les valeurs qualitatives
- Prédit les valeurs de sorties à partir des valeurs d'entrées (features)
  - Exemple : prix de l'immobilier
- Exemple d'algorithmes (liste non-exhaustive) :
  - Régression linéaire (Simple / Multiple Linear Regression)
  - Lasso Régression
  - Boosting de gradient (Gradient boosting)

Sources :

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

Apprentissage supervisé - Régression

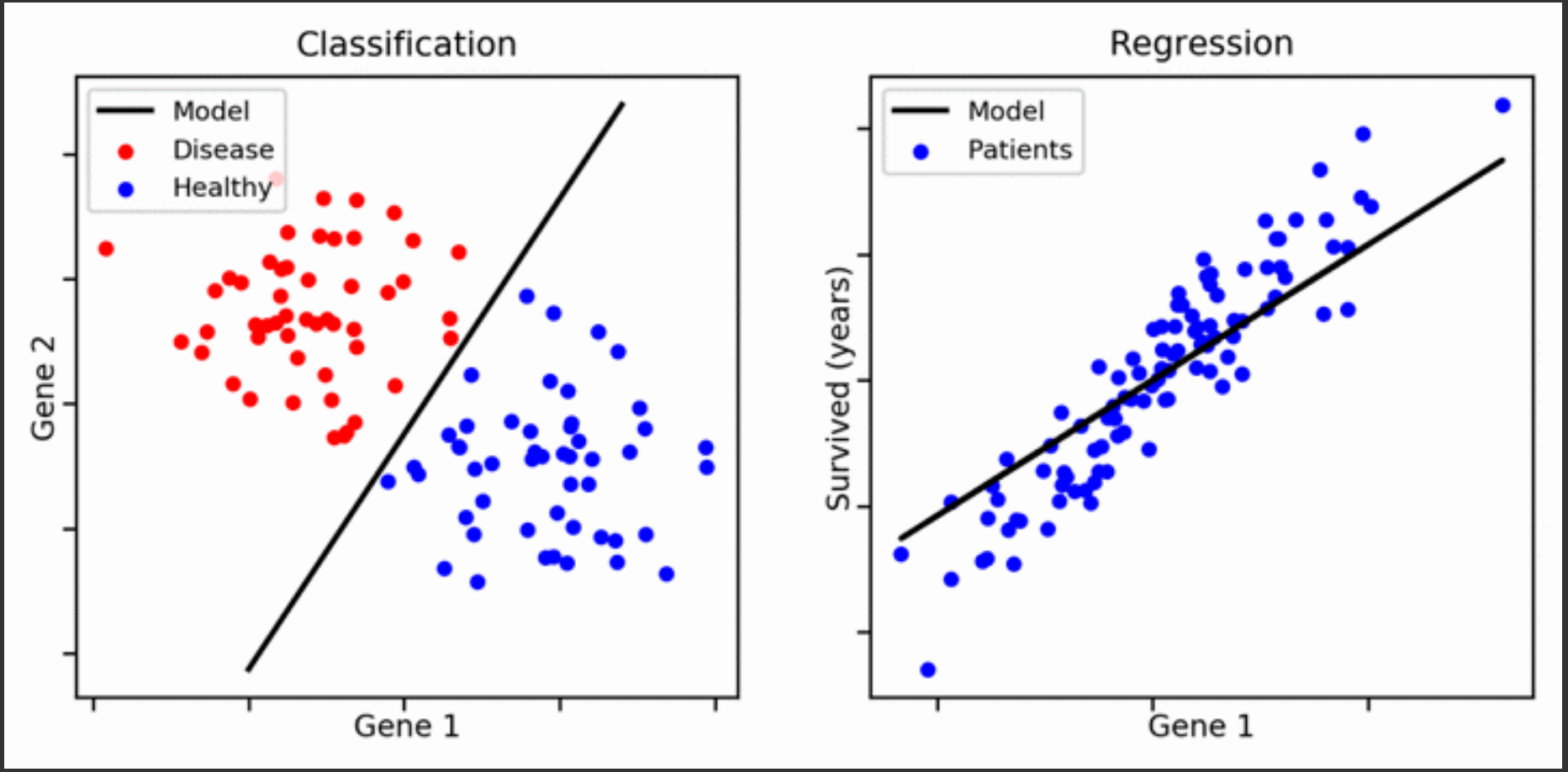
Prix habitation (€)	Surface (m²)
150 000	68
178 000	75
…	…



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 – Loi de Pareto
  - ~ 80 % des données servent à l'entraînement du modèle
  - ~ 20 % servent de test
- Test de notre modèle sur des données jamais vues

Sources :

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

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

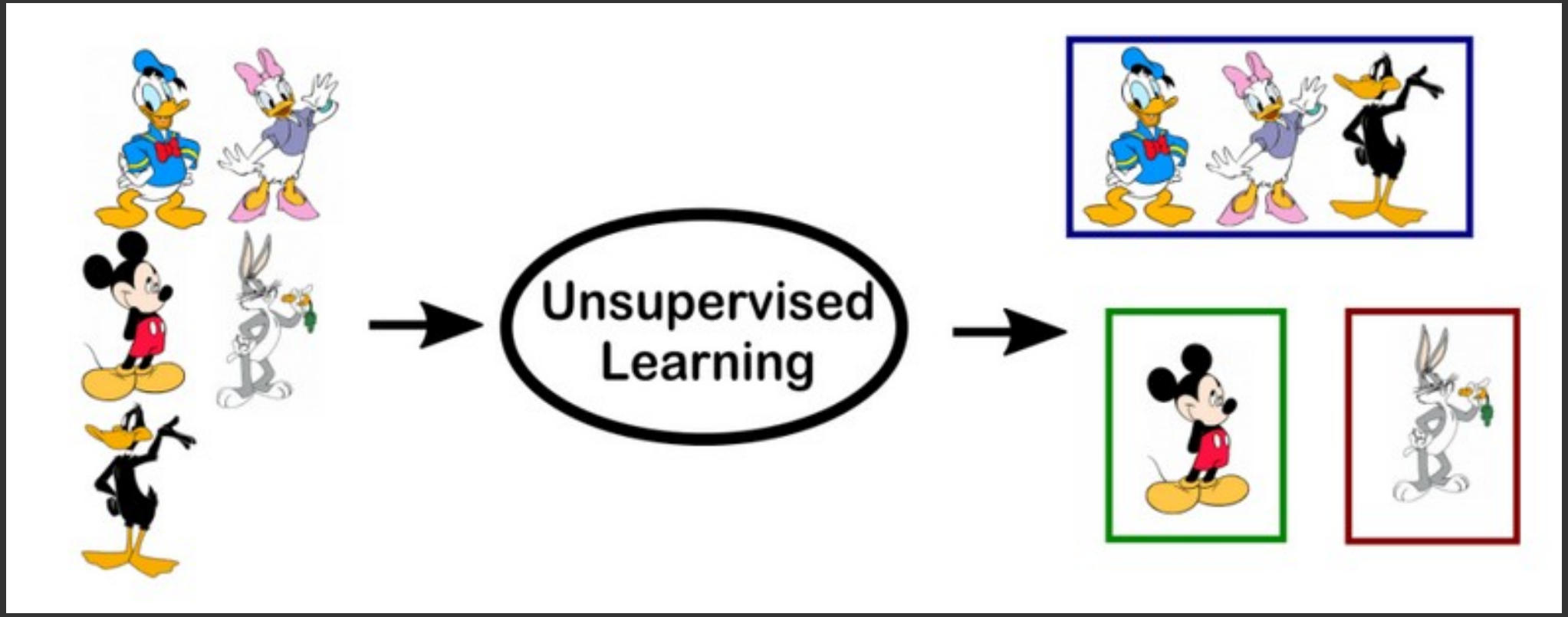
Sources :

- 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 :

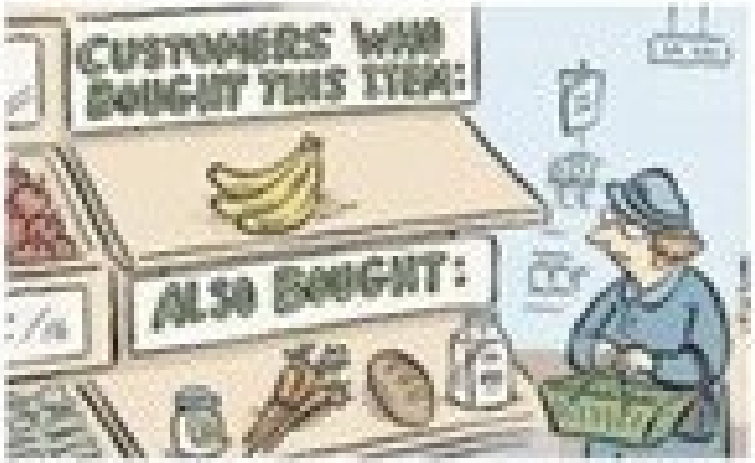
- https://fr.wikipedia.org/wiki/Algorithme_APriori

Apprentissage non-supervisé - Association

Association

People that buy X tend to buy Y

People that buy A+B tend to buy C

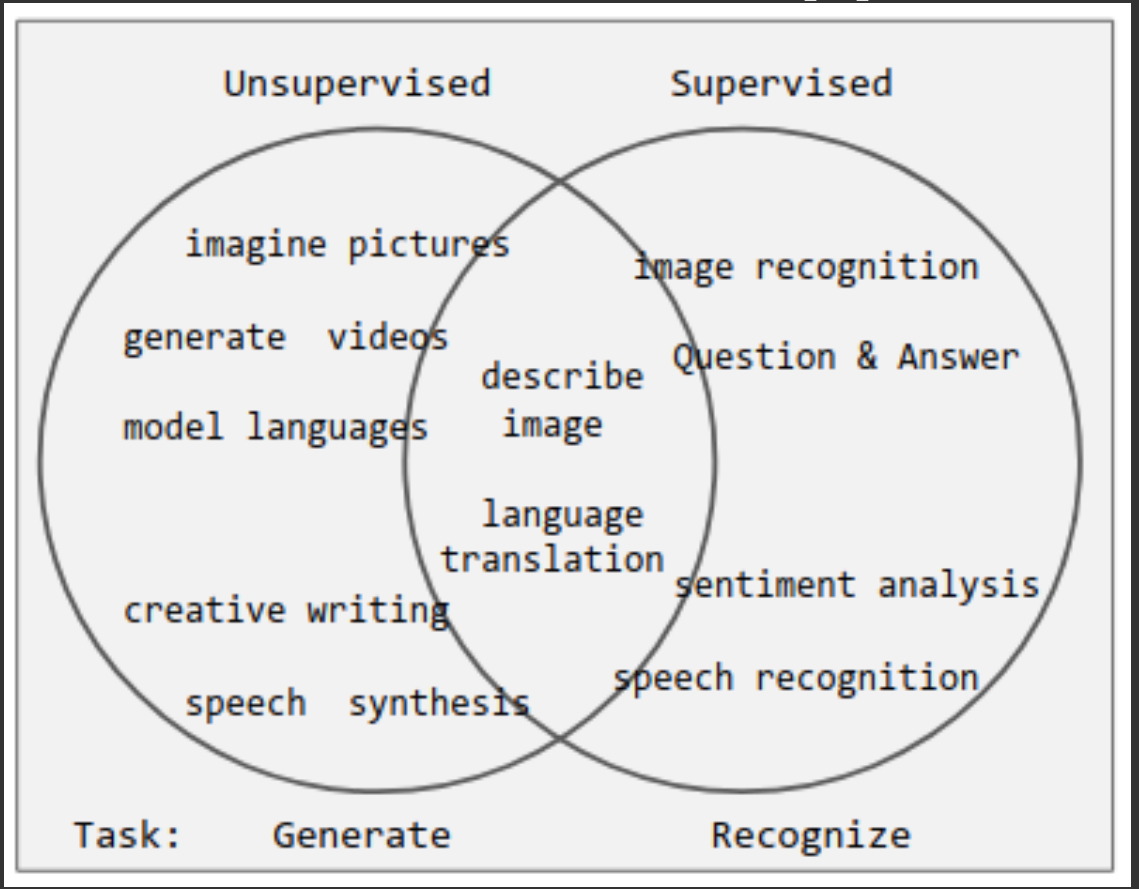


Apprentissage non-supervisé – Association
Corrélation != Causalité

- 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 :

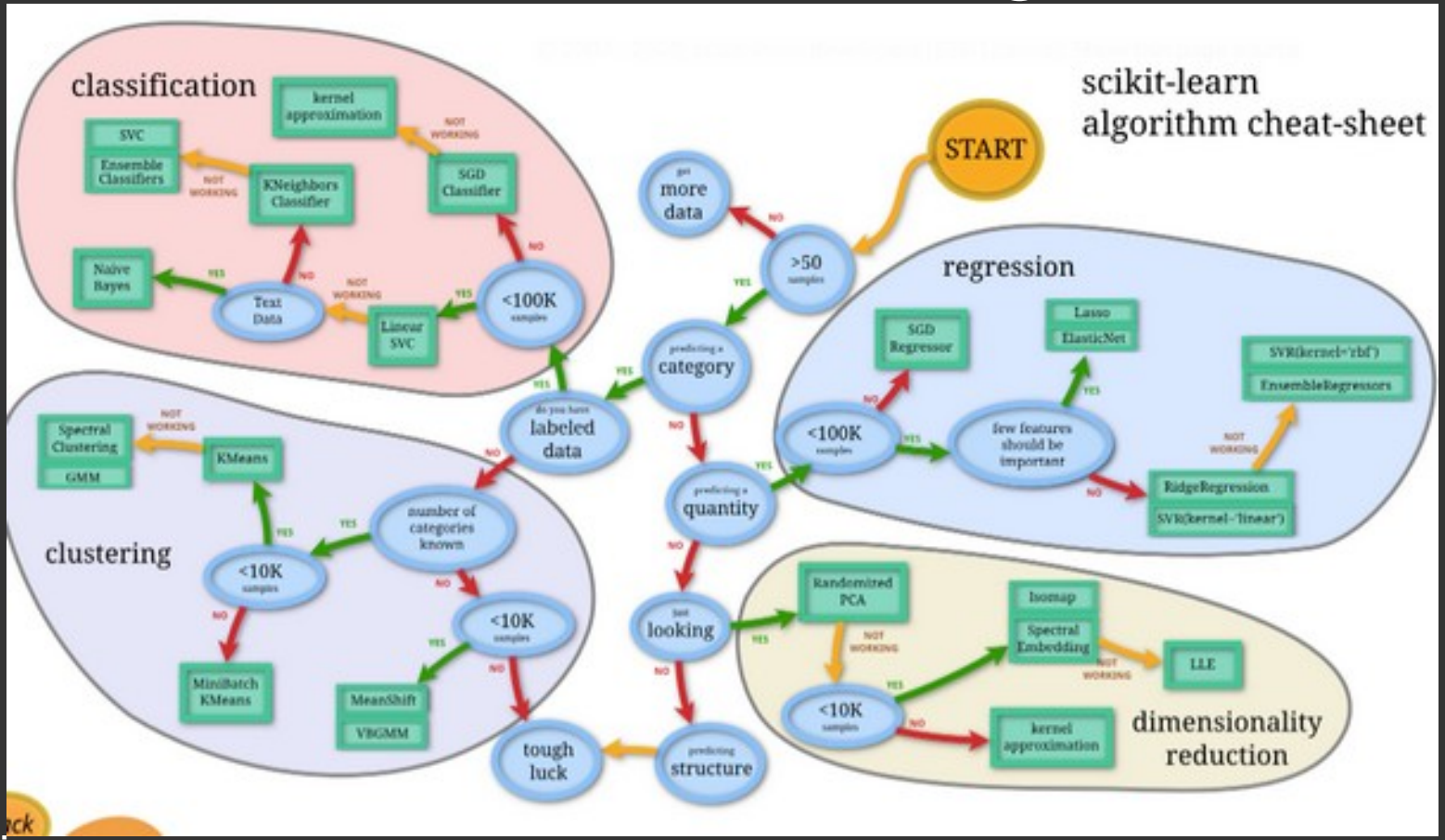
- https://en.wikipedia.org/wiki/Data_mining

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
- https://www.datacorner.fr/ml-memento/

Reconnaissance d'images

Grouper des clients

Google Translate

Kinect

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 !

Kinect

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

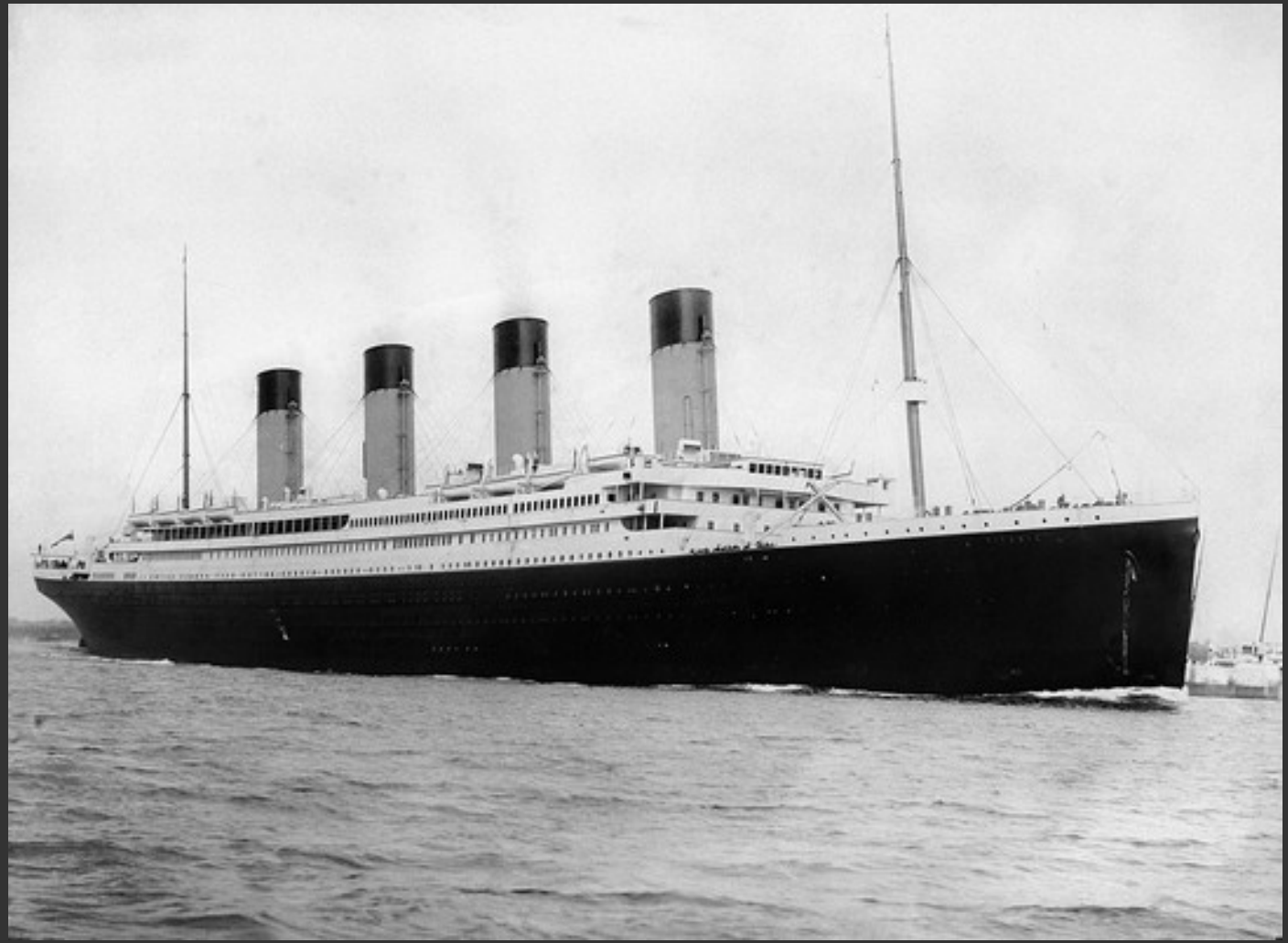
Les gauchers achètent des ciseaux pour gauchers

Non supervisé

Mes chances de survivre le 14 avril 1912

Sources :

- https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/BodyPartRecognition.pdf



Le Royal Mail Ship Titanic

Titanic

- Mis en service le 10 avril 1912
- 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 jupyter 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 data-scientifiques
- Propose un nombre conséquent de jeux de données réalisés par la communauté
- Propose une interface proche de google colab pour expérimenter

Sources :

- https://www.kaggle.com/

matplotlib

- Bibliothèque Python permettant la gestion de graphiques
- Intégrée à Jupyter et Google Colab
  - Pas besoin de pip donc
- Pas très élégant à la base → utilisation de seaborn

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

Questions ?