

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

A PPRENTISSAGE

sandra.bringay@univ-montp3.fr

INTRODUCTION

- Data Scientist**
 - Sa valeur repose sur sa capacité à **décrire le monde et à faire des prédictions**
- Apprentissage**
 - Faire des **prédictions** sur de **nouvelles données** à partir de **modèles appris** sur des **données existantes** :
 - Une tumeur donnée est-elle bénigne ou maligne ?
 - Lequel de vos clients va vous quitter ?
 - Est-ce que cet e-mail est un spam ?
 - Classer les articles Wikipedia par catégorie

Apprentissage 2

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

OBJECTIFS

- Panorama des méthodes d'apprentissage et premier élément de théorie
- Utiliser scikit-learn, une des bibliothèques Python populaires

Apprentissage 3

ORGANISATION

- Sandra Bringay
 - Méthodologie
 - Premières expérimentations
- Samuel Rochette
 - Cœur des algorithmes
 - Expérimentations plus détaillées
- Maximilien Servajean
 - Pré-requis théoriques

Apprentissage 4

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

EVALUATION

- 2 comptes rendus pour les différentes parties du cours
- Les plus détaillés possibles : **ce qui nous intéresse c'est les interprétations (et les résultats) mais pas le code !**
- En binôme
- Partiel écrit (question de cours)
- Note finale = Partiel écrit (/20) + Bonus de 2 points (sur les comptes rendus)

Apprentissage 5

RESSOURCES BIBLIOGRAPHIQUES

- <http://cours.zucker.fr/M2IFI/>
- <http://www.grappa.univ-lille3.fr/~torre/Enseignement/Cours/Apprentissage-Automatique/>
- <http://www.cril.univ-artois.fr/~koriche/Apprentissage2013-Partie1.pdf>
- <http://www.lifl.fr/~pietquin/teaching/FAACours1.pdf>
- <http://pageperso.lif.univ-mrs.fr/~francois.denis/IAAM1/expoM1.pdf>
- <http://scikit-learn.org/>
- <https://www.datacamp.com/>
- <https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-tutor3-python-scikit.pdf>

Apprentissage 6

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

APPRENTISSAGE ... UN BUZZ WORD

• machine learning
Terme de recherche

• data science
Terme de recherche

• artificial intelligence
Terme de recherche

+ Ajouter une comparaison

Dans tous les pays ▾ Cinq dernières années ▾ Toutes les catégories ▾ Recherche sur le Web ▾

Évolution de l'intérêt pour cette recherche

Apprentissage 7

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

Apprentissage : qu'est-ce que c'est pour vous ?

Apprentissage 8

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

TENTATIVES DE DÉFINITION

- Arthur Samuel, 1959 (auto-apprentissage – jeu échec)
 - Field of study that gives computers the ability to learn without being explicitly programmed
- Tom Mitchell, 1998
 - A computer program is said to learn from **experience** E with respect to some class of tasks T and **performance measure** P, if its performance at tasks in T, as measured by P, improves with experience E

Apprentissage 9

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

APPRENTISSAGE ET IA ?

IA Forte

- Machines reproduisant l'intelligence humaine
- Raisonnement à base de connaissances
- Prédicats, symboles, inférence logique
- Raisonnements de haut niveau

IA Faible

- Machines reproduisant le comportement humain
- Apprentissage à base d'exemples
- Hypothèses, calculs numériques
- Tâches spécifiques

Apprentissage 10

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

APPRENTISSAGE ET IA ?

IA Forte

- Machines reproduisant l'intelligence humaine
- Raisonnement à base de connaissances
- Prédicats, symboles, inférence logique
- Raisonnements de haut niveau

IA Faible

- Machines reproduisant le comportement humain
- Apprentissage à base d'exemples
- Hypothèses, calculs numériques
- Tâches spécifiques

Apprentissage 11

UNIVERSITÉ PAUL VALÉRY MONTPELLIERS

RECONNAISSANCE DE CARACTÈRES MANUSCRITS

Apprentissage 12

RECONNAISSANCE Vocale - ANALYSE DU LANGAGE

Apprentissage 13

DÉTECTION DE VISAGES

- vidéo-surveillance, biométrie, robotique, commande d'interface homme-machine, photographie, indexation d'images et de vidéos, recherche d'images par le contenu, etc.

Apprentissage 14

GOOGLE CAR

- Voiture sans conducteur
- Télédétection par laser (LiDAR), caméra, radars, GPS, capteurs...

Apprentissage 15

SYSTÈMES DE RECOMMANDATION

- Collecte passive des données (film ou série visionné, en entier ou non, pauses, utilisation du Replay, etc.),
- Informations saisies (recherche, notation, etc.)
- Caractéristiques des vidéos (genre, acteurs, récompenses, éléments particuliers, etc.)

Apprentissage 16

WATSON

- Super ordinateur de IBM
- Il répond à des questions formulées en langage naturel (Jeopardy)
- Il trouve une **signification** à la question posée par rapport à ce qu'il a **appris** et **emmagasiné** dans sa mémoire
- Il émet ensuite des hypothèses via des **algorithmes de réflexion** qu'il valide à partir de ce qu'il sait, avec un **score de confiance**
- Il peut ainsi **argumenter sa réponse**
- Il **apprendre de ses erreurs**
- Application des capacités analytiques : santé, finance...

Apprentissage 17

DEEPMIND

- Objectif : « **résoudre l'intelligence** » en combinant :
 - apprentissage automatique
 - neurosciences des systèmes
- pour construire de puissants **algorithmes d'apprentissage généraliste**
- Victoire de AlphaGo en 2016**
- supériorité sur le meilleur joueur de la planète

Apprentissage 18

 ET ENCORE ...

- Traduction automatique
- Détection de spam
- Imagerie médicale
- Bio-informatique
- Vidéo-surveillance
- Aide aux personnes
- E-Learning
- Economie
- Publicité
- ...

Apprentissage 19

 TENTATIVES DE DÉFINITION

Apprentissage : qu'est-ce que c'est pour vous ?

Apprentissage 20

 TENTATIVES DE DÉFINITION

Apprentissage : qu'est-ce que c'est pour vous ?

Apprentissage 21

 TENTATIVES DE DÉFINITION

Apprentissage : qu'est-ce que c'est pour vous ?

La machine parvient rapidement à reconnaître des régularités dans les données et à réagir

Apprentissage 22

 TENTATIVES DE DÉFINITION

Apprentissage : qu'est-ce que c'est pour vous ?

La machine parvient rapidement à reconnaître des régularités dans les données et à réagir

La machine crée de la connaissance de manière automatique à partir de données brutes

La machine exploite cette connaissance pour prendre des décisions

Apprentissage 23

 TENTATIVES DE DÉFINITION

Apprentissage : qu'est-ce que c'est pour vous ?

La machine parvient rapidement à reconnaître des régularités dans les données et à réagir

La machine crée de la connaissance de manière automatique à partir de données brutes

La machine exploite cette connaissance pour prendre des décisions

Stratégie pilotée par les données (data-driven strategy)

Apprentissage 24



UN CHANGEMENT DE PARADIGME

- Avec l'apprentissage, on cherche davantage à établir des **corrélations** entre 2 événements plutôt qu'une **causalité**
- Exemple:
 - on peut détecter une **corrération entre la consommation de sucre et les maladies cardiaques**, sans pour autant dire que l'une est la cause de l'autre
 - la corrélation est utile pour identifier les **personnes susceptibles d'avoir des maladies cardiaques**.
 - on ne cherche pas à expliquer POURQUOI il y a une corrélation

Apprentissage

25



TENTATIVES DE DÉFINITION



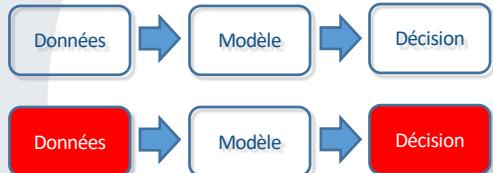
- Les **données** disponibles construisent le **modèle**
- Le modèle permet de prendre des **décisions**

Apprentissage

26



TENTATIVES DE DÉFINITION



- On se sert du modèle sur de **nouvelles données** afin de **prendre des décisions**

Apprentissage

27



LES DONNÉES



- Les données sont appelées **échantillons**.
 - Les échantillons sont souvent notés sous forme de vecteur :
- $$x = (x_1, x_2, \dots, x_p)$$
- où p est le nombre d'**attributs** (ou **coordonnées, dimensions**)

Apprentissage

28



LES DONNÉES



- **Les données labélisées**
 - les données sont accompagnées d'un **label** (ou **étiquette, classe**) qui identifie la décision à prendre pour chaque échantillon
 - Il est souvent très **coûteux** (en temps et en argent) d'avoir de grands volumes de données labélisées
 - Exemple : mail spams ou non spams
- **Les données non-labélisées**
 - Les données ne sont pas accompagnées de labels
 - Même si les données non labélisées sont plus difficiles à exploiter, elles sont beaucoup plus accessibles
 - Exemple : récupération de millions d'images sur le web pour faire de la reconnaissance de visages

Apprentissage

29



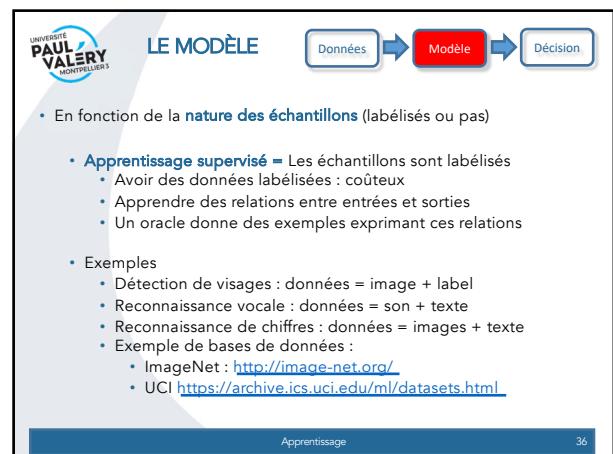
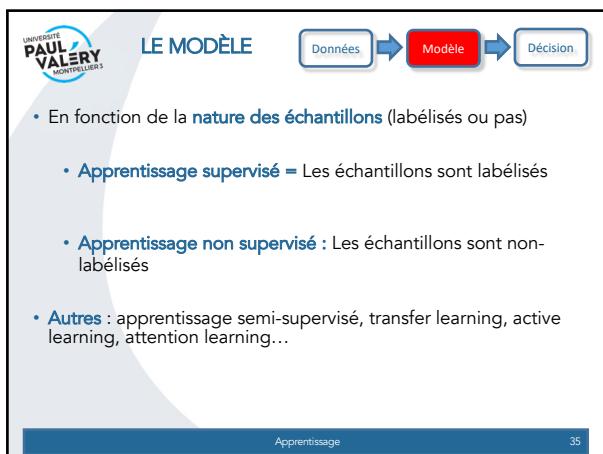
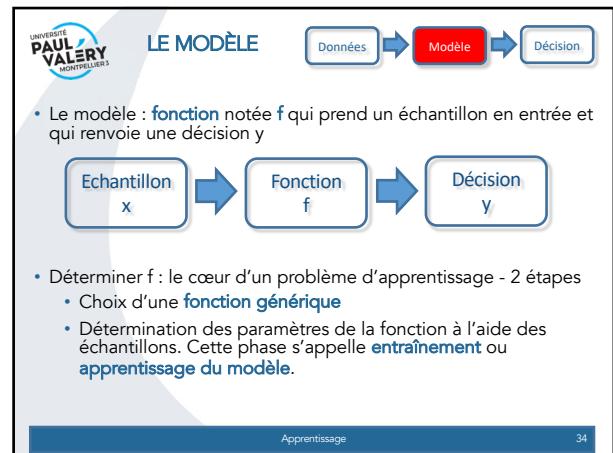
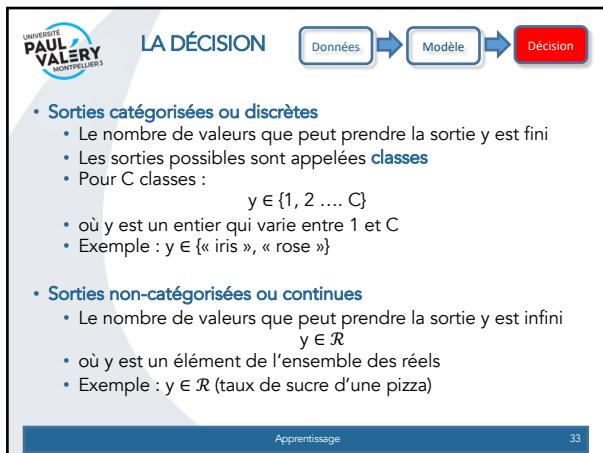
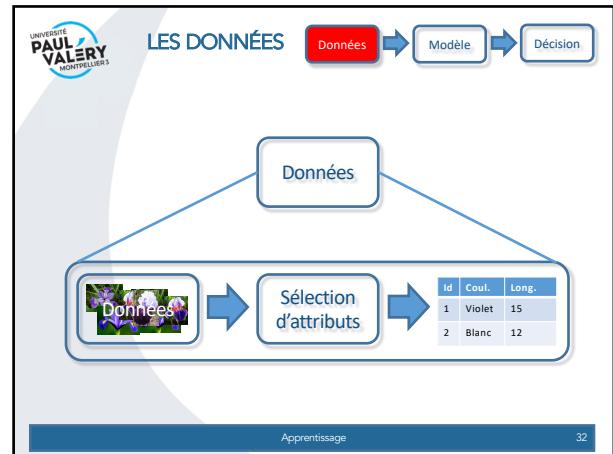
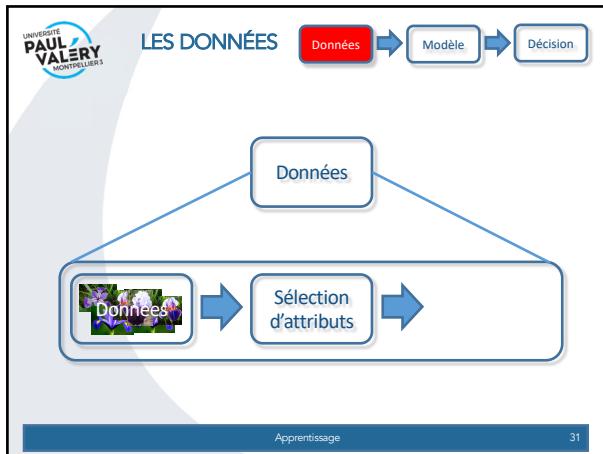
LES DONNÉES



- **Caractérisation des données**
 - Les données brutes sont souvent inexploitables
 - Généralement, on applique des **prétraitements** sur les données
 - pour extraire des **caractéristiques** pertinentes pour la prise de décision (**features**)
$$x = (x_1, x_2, \dots, x_p)$$
- Le **choix des caractéristiques (feature selection)** fait souvent appel
 - au bon sens
 - à des facteurs de corrélation statistiques
 - à des itérations successives (choix empirique)

Apprentissage

30



LE MODÈLE

- En fonction de la **nature des échantillons** (labeledés ou pas)
- Apprentissage non supervisé :** Les échantillons sont non-labélisés
 - Plus difficile à extraire mais possible (e.g. web)
 - Apprendre une structure dans un ensemble de données
 - Pas d'oracle
- Exemples
 - Clustering d'images
 - Détection de communautés (graphes d'amis Facebook, préférences Netflix etc.)

Apprentissage 37

SUPERVISÉ OU NON SUPERVISÉ ?

Choix de nouveaux clients 	Localisation de points de vente 
Popularité des pages 	Paramétrage d'une ligne de production 

Apprentissage 38

CHOIX DE NOUVEAUX CLIENTS

Banque : BNP, La poste, Caisse d'épargne...

Age
Profession
Santé
Solde
Crédits
...

Acceptation du prêt ?

Apprentissage 39

CHOIX DE NOUVEAUX CLIENTS

Banque : BNP, La poste, Caisse d'épargne...

Age
Profession
Santé
Solde
Crédits
...

Acceptation du prêt ?

Supervisé

Apprentissage 40

LOCALISATION DES POINTS DE LIVRAISON PIZZA HUT

Slide 29

www.colinlaury-data-science

Apprentissage 41

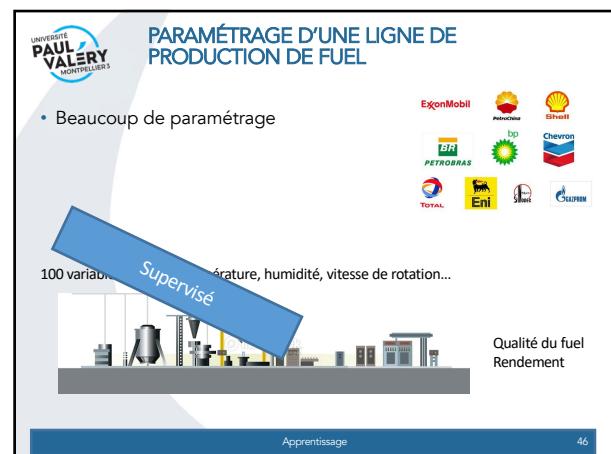
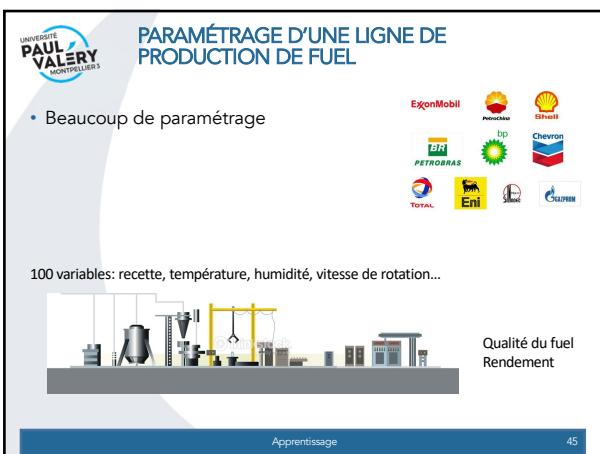
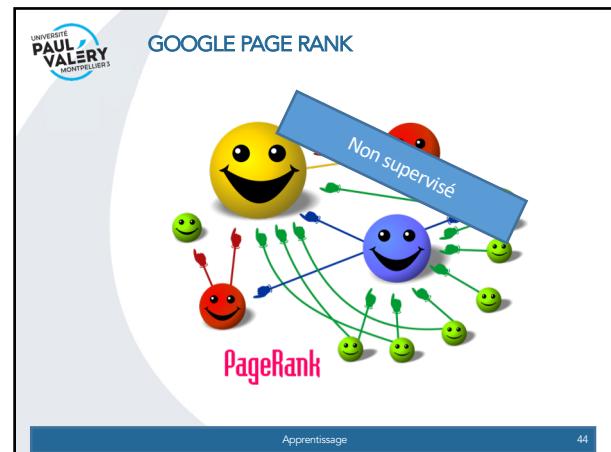
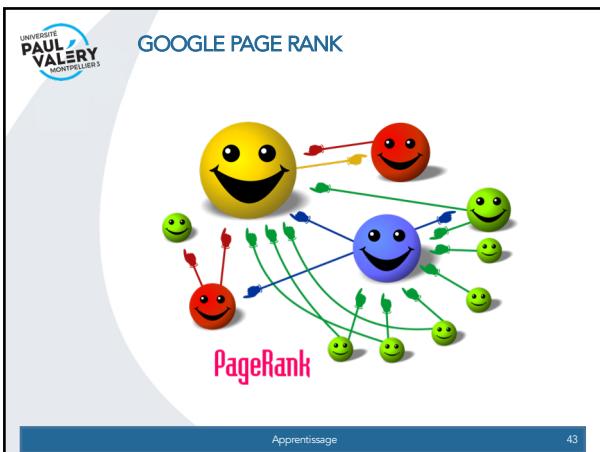
LOCALISATION DES POINTS DE LIVRAISON PIZZA HUT

Slide 29

www.colinlaury-data-science

Non supervisé

Apprentissage 42



- LES 5 ÉTAPES**
1. **Récupération des données à analyser.** Étape non triviale, souvent la plus longue... e.g. récolte de tweets, d'images sur le web...
 2. **Sélection des caractéristiques.** Souvent difficile de faire le choix des caractéristiques à utiliser pour décrire les données + Longs prétraitements
 3. **Choix du modèle.** Pas de méthodes automatiques. Dépend en grande partie des données à analyser et ... de l'expérience du data scientist
 4. **Entraînement du modèle.** Pour chaque modèle, il faut un d'entraînement. Étape longue en temps de calcul selon le modèle et la taille des données d'entrée
 5. **Evaluation du modèle.** Métrique différente selon les modèles + En général, validation croisée : découpage de l'ensemble de données en 2 parties, entraînement du modèle avec la première partie et test sur la seconde partie. On calcule ensuite des indicateurs pour évaluer le modèle
- Apprentissage 47



ENVIRONNEMENTS

• Python possède des librairies pour à peu près tout ce que vous pouvez imaginer :

- numpy et scipy pour les calculs
- Matplotlib et Seaborn pour la visualisation
- Scikit-learn pour les algorithmes
- Pandas pour les gérer les données (les charger, appliquer des opérations d'algèbre relationnelle, etc.)
- Tensorflow et PyTorch pour le deep learning
- etc.

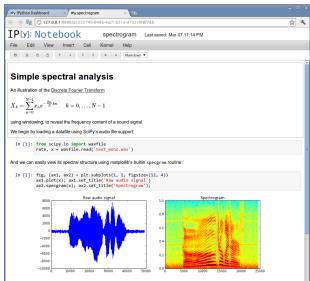


Apprentissage 49

NOTEBOOK

• Une interface web dans laquelle vous pouvez

- taper du code Python,
- l'exécuter
- voir directement les résultats, y compris sous forme de graphiques



Apprentissage 50

jupyter

Créez un notebook

Exécuter une cellule

Apprentissage 51

UN PEU DE PRATIQUE

- Moodle : W123M15

Apprentissage 52