

IA, Deep Learning et Machine Learning

EFREI Paris - M2 DEV Groupe 2



Kévin Duranty
Formateur Data & Webmarketing



Planning de la formation

mercredi, 17 janvier

9h - 17h30

IA & Datasience :

- Introduction à l'intelligence artificielle
- Manipulation de donnée avec Numpy et Pandas.

lundi, 13 mai

9h - 17h30

Machine Learning :

- Machine Learning : Apprentissage Supervisé/Non supervisé
- Bibliothèque Scikit-Learn

mardi, 14 mai

9h - 17h30

DeepLearning :

- Conception d'un perceptron
- Création d'un réseau de neurones profond avec la bibliothèque Tensorflow.

lundi, 10 juin

9h - 17h30

Traitemet du langage Naturel (NLP) :

- Tokenization TF-IDF & WordEmbedding
- Réseau de neurone Récurrent (RNN, LSTM & GRU)
- Large modèle de langue (LLM) : Bert, Mistral et GPT.

mardi, 11 juin

9h - 17h30

Computer Vision : Vision par Ordinateur (CV) :

- Type d'application : Classification, détection, segmentation
- Réseau de neurone à convolution
- Architecture Resenet
- Modèles YOLO.

mercredi, 12 juin

9h - 17h30

Projet individuel.

- Entrainement d'un modèle de machine learning/deep learning à partir d'un jeu de données.

01

Introduction à l'IA

1. Introduction à l'Intelligence Artificielle

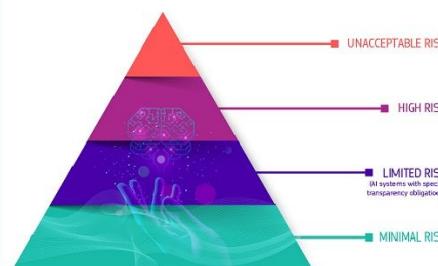
Définitions : Intelligence Artificielle



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 humaine ».



L'IA désigne la possibilité pour une machine de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité.

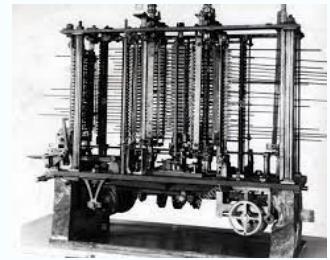


La Commission européenne présente un [cadre juridique pour l'IA](#), abordant les risques et positionnant l'Europe à l'avant-scène mondiale. Le texte réglemente l'utilisation de l'IA, garantit la sécurité, favorise l'innovation et traite des systèmes à haut risque, notamment l'identification biométrique.

Une mise en application est attendue pour 2024.

1. Introduction à l'Intelligence Artificielle

Evolution des technologies



Machine Analytique

1835

Charles Babbage & Ada Lovelace imaginent pour la première fois une machine à calculer programmable.

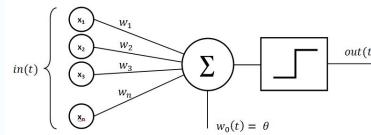


Machine de Turing



1936

Imaginée par Alan Turing, cette machine est un modèle abstrait du fonctionnement des appareils mécaniques de calcul, tel qu'un ordinateur.

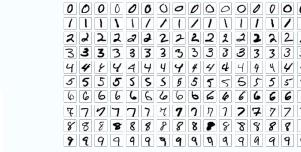


Perceptron
ELIZA (ChatBot)

1957, 1964

Le perceptron est un algorithme d'apprentissage supervisé de classificateurs binaires inventé par Frank Rosenblatt.

Le programme Eliza écrit par Joseph Weizenbau qui simule un psychothérapeute en reformulant la plupart des affirmations d'un "patient" en questions.



Deeper Blue
MNIST



1997 et 1998

Deep Blue est un superordinateur spécialisé dans le jeu d'échecs célèbre pour avoir battu le champion du monde Garry Kasparov.

La base de données MNIST pour Modified ou Mixed National Institute of Standards and Technology, est une base de données de chiffres écrits à la main.



2011 - Watson
2016 - AlphaGO

2011, 2014 et 2016

Watson est un programme informatique d'intelligence artificielle conçu par la société IBM dans le but de répondre à des questions formulées en langage naturel.



Quant à AlphaGO c'est une IA spécialisée dans le jeu de GO ayant battu en mai 2017 le champion du monde Ke Jie.

1. Introduction à 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 humaine » selon le [Larousse](#).

Informatique

Cybersécurité



Réseau



Webmarketing



Intelligence Artificielle



Apprentissage par
renforcement

Algorithme
évolutionniste

Machine Learning

DeepLearning

IA Générative

1. Introduction à l'Intelligence Artificielle

Les branches de l'IA



Vision par Ordinateur
Computer Vision



Les domaines
de l'IA

1. Introduction à l'Intelligence Artificielle

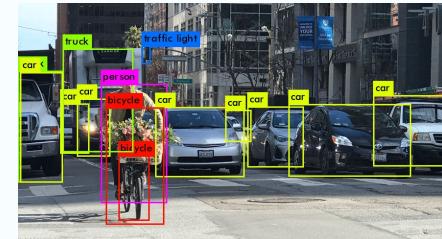
Computer Vision : Vision par ordinateur

La vision par ordinateur, également appelée "computer vision" en anglais, est un domaine de l'informatique qui s'intéresse à la réalisation de tâches visuelles par des ordinateurs. Cela inclut la capture, l'analyse et la compréhension d'images et de vidéos par des ordinateurs.

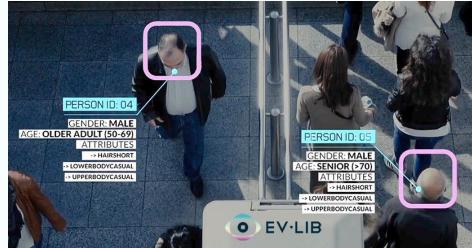
Classification



Détection



Identification



Segmentation



1. Introduction à l'Intelligence Artificielle

Les branches de l'IA



Vision par Ordinateur
Computer Vision



NLP: Natural Language Processing



Les branches
de l'IA



1. Introduction à l'Intelligence Artificielle

NLP : Natural Language Processing

*Traitement du langage Naturel



Analyse de sentiment



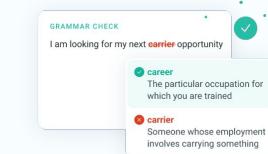
Traduction automatique



Chatbots



Classification de texte



Correction automatique



Résumé et générateur de texte

Ressources : [Chat GPT3 OpenIA](#) | [Scriben](#) | [Google Traduction](#) | [Tidio](#) | [Huggingface](#)

1. Introduction à l'Intelligence Artificielle

Les branches de l'IA



Analyse prédictive



Vision par Ordinateur
Computer Vision

Robotique
IA Incarnée



Les branches
de l'IA

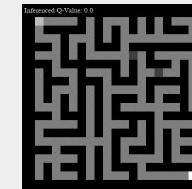
NLP: Natural Language Processing



IA Générative



Language Modeling



Agents autonomes



1. Introduction à l'Intelligence Artificielle

Importance de la donnée

Donnée : Représentation conventionnelle d'une information permettant d'en faire le traitement automatique. → **anglicisme data**.

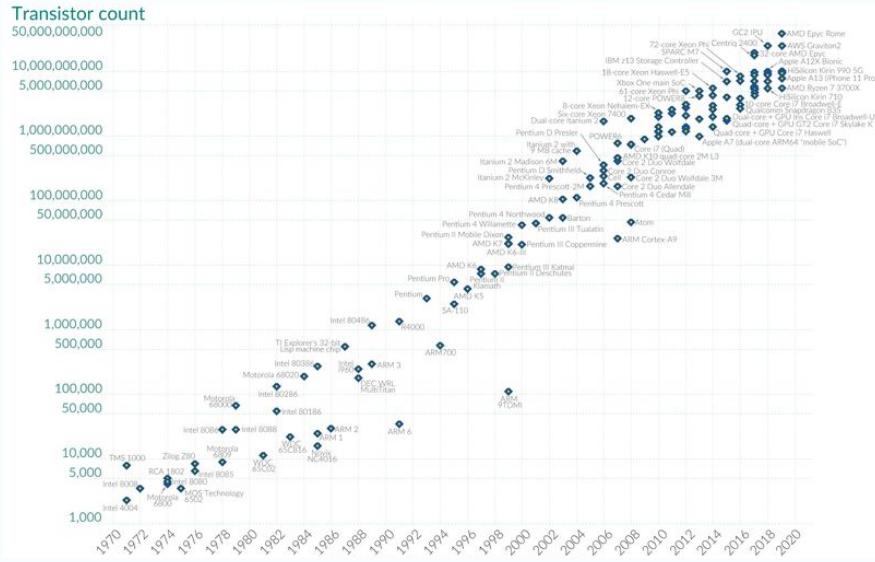
Les 3V qualifiant le BigData

Ces 3V permettent de mieux appréhender les enjeux et les fondements du Big Data :

- V comme Volume
- V comme Vitesse et Visualisation
- V comme Variété

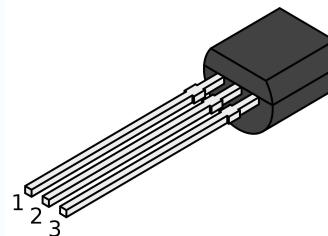
1. Introduction à l'Intelligence Artificielle

Les 3V → V comme Vitesse



La [Loi de Moor](#) décrit depuis 1965 un **doublement du nombre de transistors présents sur une puce** de microprocesseur tous les deux ans.

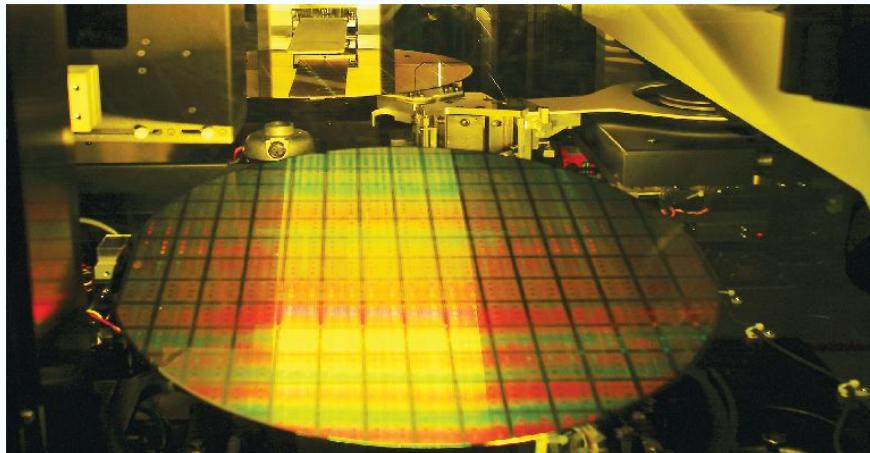
La puissance de calcul a été multipliée par 1 million en 40 ans.



Un transistor est un composant électronique qui peut être utilisé pour amplifier ou commuter des signaux électriques.

1. Introduction à l'Intelligence Artificielle

Les 3V → V comme Vitesse



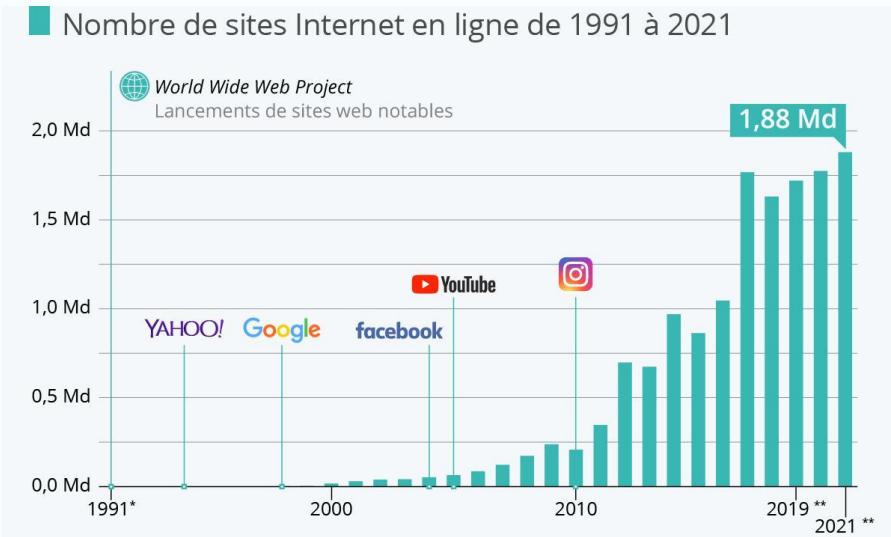
La finesse de gravure atteint aujourd’hui les 5 nanomètres pour les puces fabriquées par la fonderie taïwanaise **TSMC**.

Les processeurs sont fabriqués grâce à des lasers sur des plaques de silicium.

La limite physique étant la taille de l’atome de l’ordre de 0,1 nanomètre.

1. Introduction à l'Intelligence Artificielle

Les 3V → V comme Variété



Chercheur au CERN dans les années 90 Tim Berners-Lee est un informaticien britannique qui a développé le langage html et le protocole http.

Il met en ligne en 91 le premier site [internet](#).



1. Introduction à l'Intelligence Artificielle

Les 3V → V comme Volume

L'un des phénomènes expliquant l'explosion des données provient des coûts de stockage de l'information qui ont chuté drastiquement.

10M\$ pour stocker 1Go dans les années 50, aujourd'hui cela coûte moins de 1\$



RAMAC 305 d'IBM - 1956
5Mo, 8,8 ko/s → 50 000\$



Une multiplication du nombre de données créés par 45 de 2020 à 2035.

En une minute sur internet c'est : **2 Millions** de message sur Snapchat, **65K photos** publiées sur Instagram, **240K photos** sur Facebook, **575K posts** sur Twitter, **200 vidéos** ajouté sur Youtube.

1. Introduction à l'Intelligence Artificielle

Nous utiliserons l'outil [Google Colab](#) pour la manipulation des données avec les bibliothèques de data science Pandas et Numpy. Voici un panorama de l'environnement de développement data en Python :

TP1 - **Maîtrise de la bibliothèque Numpy**
→ Accédez au fichier suivant en utilisant Google Colab : [TP1 - Maîtrise de la bibliothèque Numpy](#)

IDE (environnement de dev)



Bibliothèques Data Science



02

Datascience

Correction du TP2 à 16h !

2. Datascience

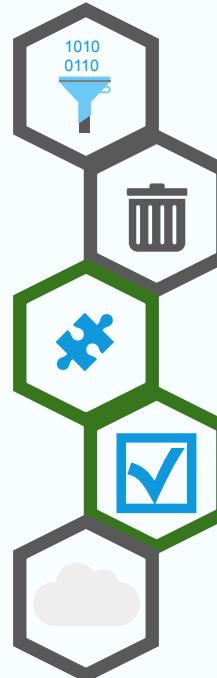
01 Collecte et Visualisation de la donnée

Nettoyage de la donnée **02**

03 Construction du modèle & Entraînement

Évaluation du modèle **04**

05 Déploiement du modèle



2. Datascience

Qu'est-ce qu'une donnée ?

Le machine learning utilisent essentiellement deux structures de données pour la conception et l'exploitation des algorithmes :

Les données structurées

Données organisées selon un modèle prédefini : tableau, base de données,

Code client	Nom	Prenom	Ville	Adresse e-mail	Code Postal	Mairie
X-100001	Mme	Elisabeth	ANGERS SUR MER	5, place de la Mairie	66790	ANGLES SUR MER
X-100002	Mme	Aline	ANGERS	5, place de la République	44000	ANGERS
X-100003	Mlle	Olivier	ANGERS	12, place des Fromages	79000	NANTES
X-100004	M.	Philippe	BONNEAU	31, rue des Colombes	15130	RENNES
X-100005	M.	Denis	BOSSEAU	23, avenue des Pins	44000	ANGERS
X-100006	Mme	Delphine	CORNETTE	5, impasse des Bergères	44000	ANGERS
X-100007	M.	Mathilde	COMBRET	36, rue des Boulangers	13120	ARCACHON
X-100008	M.	Dominique	DABAGIET	1, rue Claude Frérotte	64000	PERPIGNAN
X-100009	M.	Isabelle	DEBOIS	12, rue de la grande Armée	72000	LE MANS
X-100010	M.	Emmanuel	DECERF	36, rue des Oliviers	17200	ROCHEFORT
X-100011	Mme	Olivier	DEJOYEZ	45, rue du Calme	79000	NANTES
X-100012	Mme	Valérie	DE MASQUE	5, rue des Champs	44000	ANGERS
X-100013	M.	Denis	DEPERNET	46, rue des Hirondelles	13120	ARCACHON
X-100014	M.	Stéphane	DELON	75, rue des Maréchaux	66790	ANGLES SUR MER
X-100015	Mme	Hélia	DESKAMP	31, boulevard de l'Industrie	75016	PARIS
X-100016	M.	Hervé	DEVAMNY	5, rue de l'Amitié	69003	LYON
X-100017	M.	Ferdinand	DIVIU	29, rue Pouydebat	17000	LA ROCHELLE
X-100018	Mme	Isabelle	DUVAL	15, avenue des Facultés	17000	LA ROCHELLE
X-100019	Mme	Élise	EPINARD	25, rue des Acacias	17000	LA ROCHELLE
X-100020	Mme	Clémence	FOURNIER	15, boulevard de la Source	16000	ANGERS
X-100021	M.	Jean	GARNIER	5, place des Fromages	75008	PARIS
X-100022	M.	Agathe	GRANAU	96, impasse des Miches	66790	ANGLES SUR MER
X-100023	M.	Thierry	HARIS	44, rue des Tilleuls	72000	LE MANS
X-100024	M.	William	HESSEM	12, avenue du Casino	34130	MONTELÉGAR

les données non-structurées

Données aux formats brut tels que du texte, des images, des fichiers audio/vidéo.



Par la suite, nous utiliserons ce type de structure.

2. Datascience

Qu'est-ce qu'une donnée ?

Données structurées

On parle de jeu de données lorsqu'on exploite un fichier contenant l'ensemble de nos informations structurées sous forme de tableau.

Code Client	Nom	Pénomé	Ville	Titre	Prénom	Nom	Rue et numéros	Code Postal	Ville
0100001	Mme	Isabelle	ARVOVILLE	S, place de la Mairie	66790	ARVILLERS SUR			
0100001	Mme	Aline	ARVOVILLE	3, place de la République	66000	ARVILLERS SUR			
0100002	Mme	Elise	ACROIX	12, place des Frangaises	78000	NEUILLY			
0100002	M.	Philippe	BONHOMME	24, rue des Colombes	75136	RENNES			
0100003	M.	Yves	BOUTEILLES	23, avenue des Pins	44000	SAINT NAZAIRE			
0100003	Mme	Yvette	COUVELAIS	3, impasse des Vignevants	44000	AMERES BE			
0100004	M.	Marie	COURBET	36, rue des Boulangers	33120	APPEACHON			
0100004	M.	Dominique	DEBLAIS	1, rue Claude Trévoux	66000	PERPIGNAN			
0100005	M.	Robert	DESOL	12, rue de la Grange d'Arche	72000	LE MANS			
0100005	M.	Bernard	DECAMPS	26, rue des Oliviers	17300	ROQUEFORT			
0100006	Mme	Olivier	DEJOUR	33, rue du Calme	78930	NANTERRE			
0100006	Mme	Valérie	DE MAISOL	5, rue des Etangs	44000	NANTES			
0100007	M.	Dominique	DESERET	46, rue des Hêtres pendus	33120	APPEACHON			
0100007	M.	Magali	DEUFON	72, rue des Marathons	66790	ARVILLERS SUR			
0100008	Mme	Hélia	DEVALDAMP	54, boulevard de l'Évitement	75016	PARIS			
0100008	M.	Isabelle	DEVILLE	5, rue de l'Amitié	69003	LYON			
0100009	M.	François	DIVLU	29, rue Poujol	12000	LA ROCHELLE			
0100009	Mme	Isabelle	ELLET	15, avenue des Facultés	12000	LA ROCHELLE			
0100010	Mme	Datelle	EPINAILLARD	25, rue des Alouettes	12000	LA ROCHELLE			
0100010	Mme	Christine	FONTAINE	13, boulevard de la Source	14000	AMBOISE			
0100011	M.	Yannick	GONTHIER	5, place des Frangaises	75016	PARIS			
0100011	M.	Auguste	GRANAUDE	56, impasse des Mimosas	66790	ARVILLERS SUR			
0100012	M.	Thierry	HABHS	14, rue des Taillieurs	72000	LE MANS			
0100012	M.	William	HEDON	32, avenue du Closat	34130	MONTPELLIER			

En science des données, et plus précisément en les statistiques, afin de décrire nos données, nous utilisons un langage particulier.

⇒ Les **lignes** sont appelées des **entrées, individus..**

⇒ Les **colonnes** sont appelées des **variables ou features.**

⇒ L'ensemble des individus forme une population.

Les premiers éléments à observer lorsque l'on récolte un jeu de données sera donc **le nombre d'individus** du jeu de données (**sa population**) et **le nombre de variables**.

2. Datascience

Qu'est-ce qu'une donnée ?

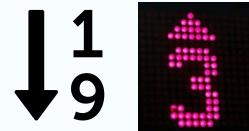
Données structurées

Il existe ensuite 2 types de variables : les variables **quantitatives** et les variables qualitatives.

Les **variables quantitatives** (ou numériques) regroupent les **informations quantifiables** (qui peuvent être comptées) et enregistrées sous forme de nombre, exemples : âge, taille, poids, prix.

Les variables quantitatives sont divisées en deux catégories :

- Variables quantitatives **discrètes**, qui prennent un nombre fini de valeurs réelles dans un intervalle donné (ex : nombre d'enfant, étage, chaîne TV).
- Variables quantitatives **continues**, qui peuvent prendre un nombre infini de valeurs réelles dans un intervalle donné (ex: salaire, température, vitesse).



2. Datascience

Qu'est-ce qu'une donnée ?

Données structurées

Il existe ensuite 2 types de variables : les variables quantitatives et les variables **qualitatives**.

Les **variables qualitatives** (ou catégorielles) regroupent les **informations non-quantifiables**, qui ne peuvent pas être comptées et sur lesquelles on ne peut réaliser des opérations arithmétiques (ex : nom de famille, genre, adresse).

Les variables qualitatives sont également divisées en deux catégories :

- Variables qualitatives **ordinales**, qui présentent des valeurs définies par une relation d'ordre entre les différentes catégories possibles exemple : mentions au BAC, le nutri-score, étiquette énergie.
- Variables qualitatives **nominales**, qui présentent des valeurs n'ayant pas de relation d'ordre entre elles : prénom, secteur d'activité, ville.



First Name *

2. Datascience

Où trouver les données ?

Base de données



En entreprise de nombreux moyens sont déployés pour collecter la donnée de manière automatisée. C'est la mission du Data architect d'organiser cette collecte. La RGPD est le cadre de loi européen qui définit de nouveaux droits relatifs aux données personnelles.



Like



Géolocalisation



Vidéo



Biométrique



Audio



Navigation



Médicale



Banquaire



Fichiers



Connexion

2. Datascience

Où trouver les données ?

Banque de données

Les banques de données sont des plateformes web où il est possible de récupérer des données sous format .csv, .xls ou encore .json.

Les institutions



data.gouv.fr



Les plateformes tiers

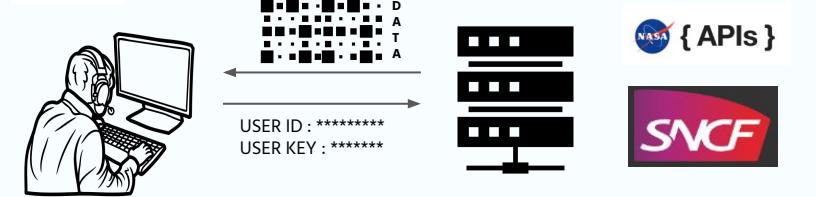


Les entreprises

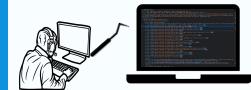


API

Les API (application programming interface) est un programme qui permet la connexion entre plusieurs services.



Scraping



Le scraping est une des méthodes permettant de collecter de la donnée. Le principe est d'extraire de l'information provenant d'un site web en récupérant son code.

2. Data préparation

Exercices corrigés

Lien du jeu de données



TP2 - Maîtrise de la bibliothèque Pandas

L'objectif de ce TP est de manipuler un jeu de données à l'aide de la bibliothèque [Pandas](#).

- Accédez au fichier suivant en utilisant Google Colab :
[**TP2 - Maîtrise de la bibliothèque Pandas**](#)
- Suivez les instructions présentes dans le notebook.

2. Data Science

Banque de données

Les institutions



data.gouv.fr



Les plateformes tiers



Les entreprises



RATP
DEMANDEZ
NOUS
LA VILLE



Analyse exploratoire de données

Utilisez l'une des plateformes proposées ci-contre pour récupérer un jeu de données au format CSV, puis analysez les données en 3 parties dans un notebook :

1. Exploration des données
2. Visualisation des données (au moins 2 graphiques)
3. Nettoyage des données

Déposez le TP2 et votre analyse exploratoire [ici](#).

03

Statistique et Visualisation

3. Statistique et Visualisation

Statistique

Une statistique est un indicateur numérique calculé à partir d'un échantillon.

- Une analyse univariée est une analyse effectuée sur une variable à la fois.
- Une analyse bivariée est une analyse menée entre deux variables qui vise à observer les éventuels liens de corrélation et contredire leur interdépendance.

Corrélation exemple : Les évolutions des ventes de glaces et des coups de soleil.

Il existe deux termes pour catégoriser les statistiques, qui reviennent beaucoup dans le lexique courant :

1. Un indice statistique, c'est une statistique construite à partir d'une certaine vision, à partir de connaissances d'un domaine.
2. Un indicateur qui est une statistique plus neutre, construite sans à-priori et sans intention derrière.

3. Statistique et Visualisation

La Moyenne

Pour toute liste d'éléments réels (x_1, \dots, x_n), on définit sa moyenne arithmétique par la formule :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Ecart-type

Pour toute liste d'éléments réels (x_1, \dots, x_n) et de moyenne \bar{x} on définit l'écart-type par la formule :

$$\sigma = \sqrt{V} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

3. Statistique et Visualisation

Les quantiles de distribution

En statistiques et en théorie des probabilités, les quantiles sont **les valeurs** qui divisent un jeu de données en **intervalles de même probabilité égale**.

- **Les quartiles :**
 - Le premier quartile est la statistique notée générale q1 ;
 - Le second quartile n'est autre que la médiane ;
 - Le troisième quartile est noté q3 et son écart au 1er quartile définit l'écart interquartile, qui est une des mesures classiques de la dispersion de l'échantillon de données, néanmoins plus robuste que l'écart-type.
- **Les déciles.** Ils sont d'usage fréquent en géologie minière, en hydrologie, ainsi que dans nombre de statistiques médicales ;
- **Les centiles,** ou percentiles selon un anglicisme fréquent. Ainsi, le 5e centile partage l'échantillon en 5 % des données sous lui, et les 95 % restant au-dessus de lui. Le dernier centile (le 99e) joue fréquemment un rôle de seuil d'alerte extrême pour des mesures qui traduisent l'intensité d'un phénomène sujet à des évolutions critiques et en permettent ainsi le suivi

3. Statistique et Visualisation

Histogramme

Un histogramme représente la distribution d'une variable numérique pour un ou plusieurs groupes. Les valeurs sont divisées en bacs, chaque bac est représenté sous forme de barre.

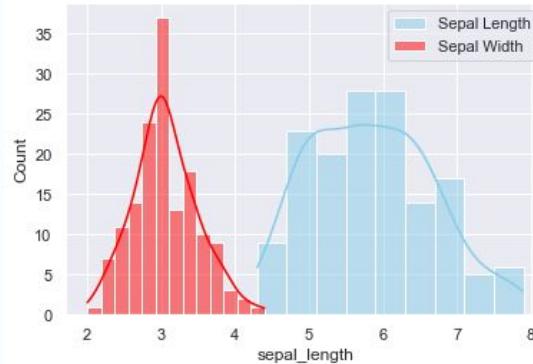
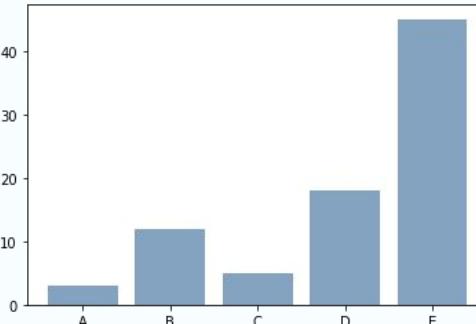


Diagramme à barres

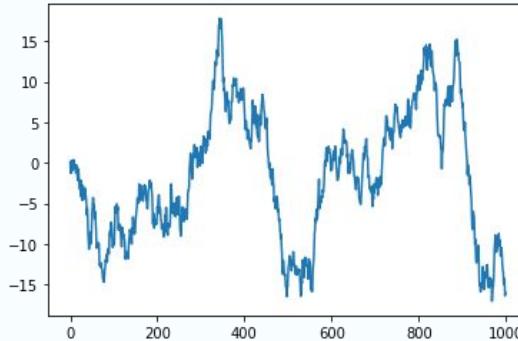
Un graphique à barres montre la relation entre une variable numérique et une variable catégorielle. Chaque entité de la variable catégorielle est représentée sous forme de barre. La taille de la barre représente sa valeur numérique.



3. Statistique et Visualisation

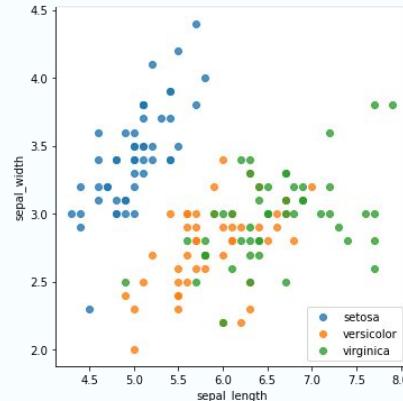
Courbes linéaires

Un graphique linéaire affiche l'évolution d'une ou plusieurs variables numériques. C'est l'un des types de graphique les plus courants.



Nuage de points

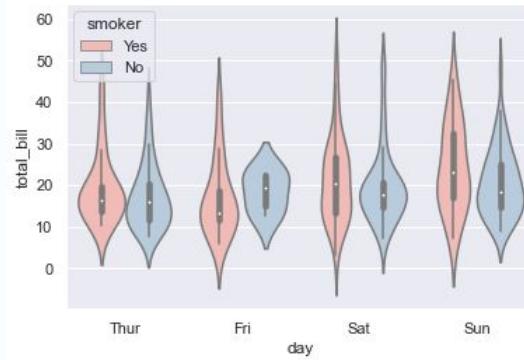
Un nuage de points affiche la relation entre 2 variables numériques. Chaque point de données est représenté par un cercle.



3. Statistique et Visualisation

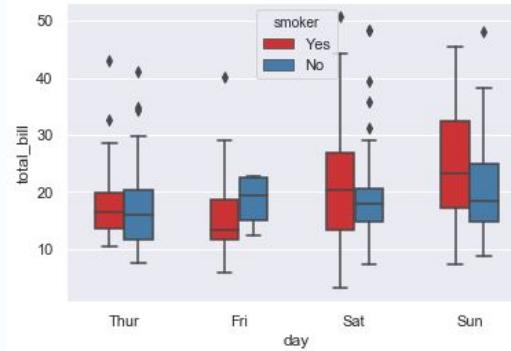
Violon

Un graphique en violon permet de visualiser la distribution d'une variable numérique pour un ou plusieurs groupes.



Boîte à moustaches

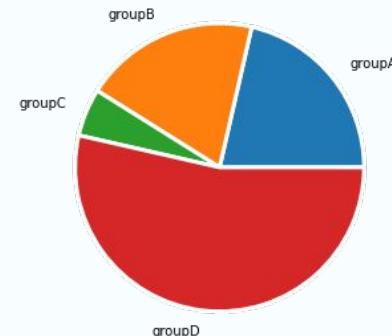
Une boîte à moustaches résume la distribution d'une variable numérique pour un ou plusieurs groupes. Il permet d'obtenir rapidement la médiane, les quartiles et les valeurs aberrantes, mais masque également les points de données individuels de l'ensemble de données.



3. Statistique et Visualisation

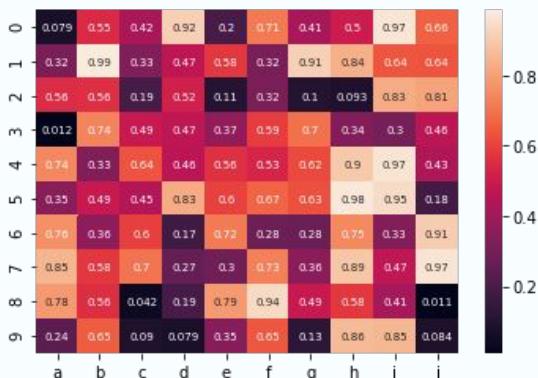
Diagramme circulaire

Un camembert est un cercle divisé en secteurs qui représentent chacun une proportion de l'ensemble. C'est l'un des types de visualisation les plus courants, mais aussi probablement le plus critiqué .



Carte de chaleur

Une carte thermique est une représentation graphique de données où chaque valeur d'une matrice est représentée par une couleur.



Exporter son notebook en html :

```
jupyter nbconvert --to html /PATH/TO/YOUR/NOTEBOOKFILE.ipynb
```

04

Machine Learning

Planning de la formation

mercredi, 17 janvier

9h - 17h30

IA & Data Science :

- Introduction à l'intelligence artificielle
- Manipulation de donnée avec Numpy et Pandas.

lundi, 13 mai

9h - 17h30

Machine Learning :

- Machine Learning : Apprentissage Supervisé/Non supervisé
- Bibliothèque Scikit-Learn

mardi, 14 mai

9h - 17h30

DeepLearning :

- Conception d'un perceptron
- Création d'un réseau de neurones profond avec la bibliothèque Tensorflow.

lundi, 10 juin

9h - 17h30

Traitemet du langage Naturel (NLP) :

- Tokenization TF-IDF & WordEmbedding
- Réseau de neurone Récurrent (RNN, LSTM & GRU)
- Large modèle de langue (LLM) : Bert, Mistral et GPT.

mardi, 11 juin

9h - 17h30

Computer Vision : Vision par Ordinateur (CV) :

- Type d'application : Classification, détection, segmentation
- Réseau de neurone à convolution
- Architecture Resenet
- Modèles YOLO.

mercredi, 12 juin

9h - 17h30

Projet individuel.

- Entrainement d'un modèle de machine learning/deep learning à partir d'un jeu de données.

2. Datascience

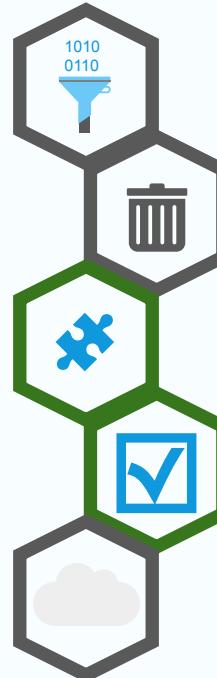
01 Collecte et Visualisation de la donnée

Nettoyage de la donnée **02**

03 Construction du modèle & Entraînement

Évaluation du modèle **04**

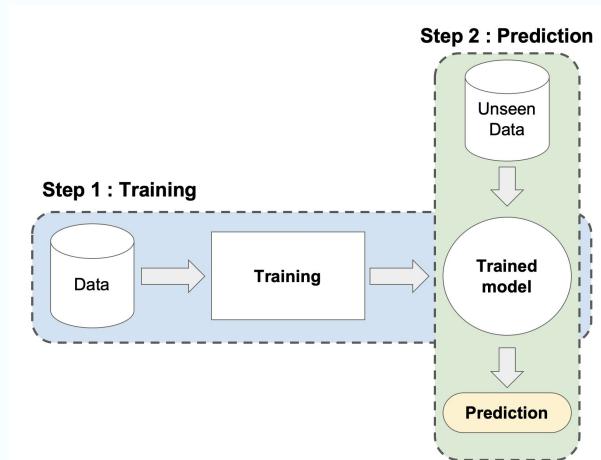
05 Déploiement du modèle



4. Machine Learning

Machine Learning

Le machine learning (apprentissage automatique) est une branche de l'intelligence artificielle qui consiste à entraîner des algorithmes sur un ensemble de données afin de prédire un résultat. L'élément à prédire est appelé **Label** ou **Target**, les algorithmes quant à eux sont appelés des **modèles**.



La première phase consiste à entraîner le modèle sur des données d'entraînement. Le modèle s'entraîne à prédire la modèle target.

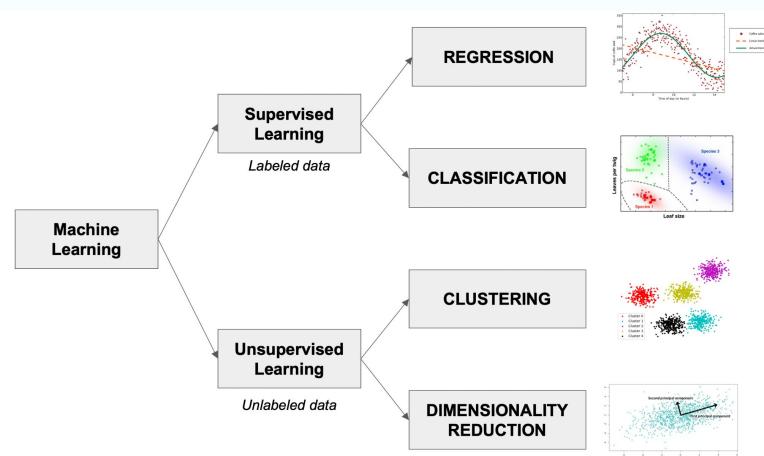
La deuxième phase consiste à évaluer le modèle sur des données qu'il n'a pas vu durant son entraînement. Cette phase sert à évaluer la performance du modèle.

→ Si les performances du modèle sont correctes alors il sera intégré aux applications métiers, sinon le modèle s'entraîne à nouveau avec des paramètres différents.

4. Machine Learning

Le machine learning est divisé en deux branches :

- L'apprentissage **Supervisée** : le modèle est entraîné à partir de données étiquetées (label/targette cible)
- L'apprentissage **Non-Supervisée** : le modèle est entraîné sur des données non étiquetées et doit découvrir des structures intrinsèques aux données. Ces modèles tentent généralement de regrouper les données similaires ou de trouver des relations entre les données.



Régression : La targette à prédire est une valeur continue.

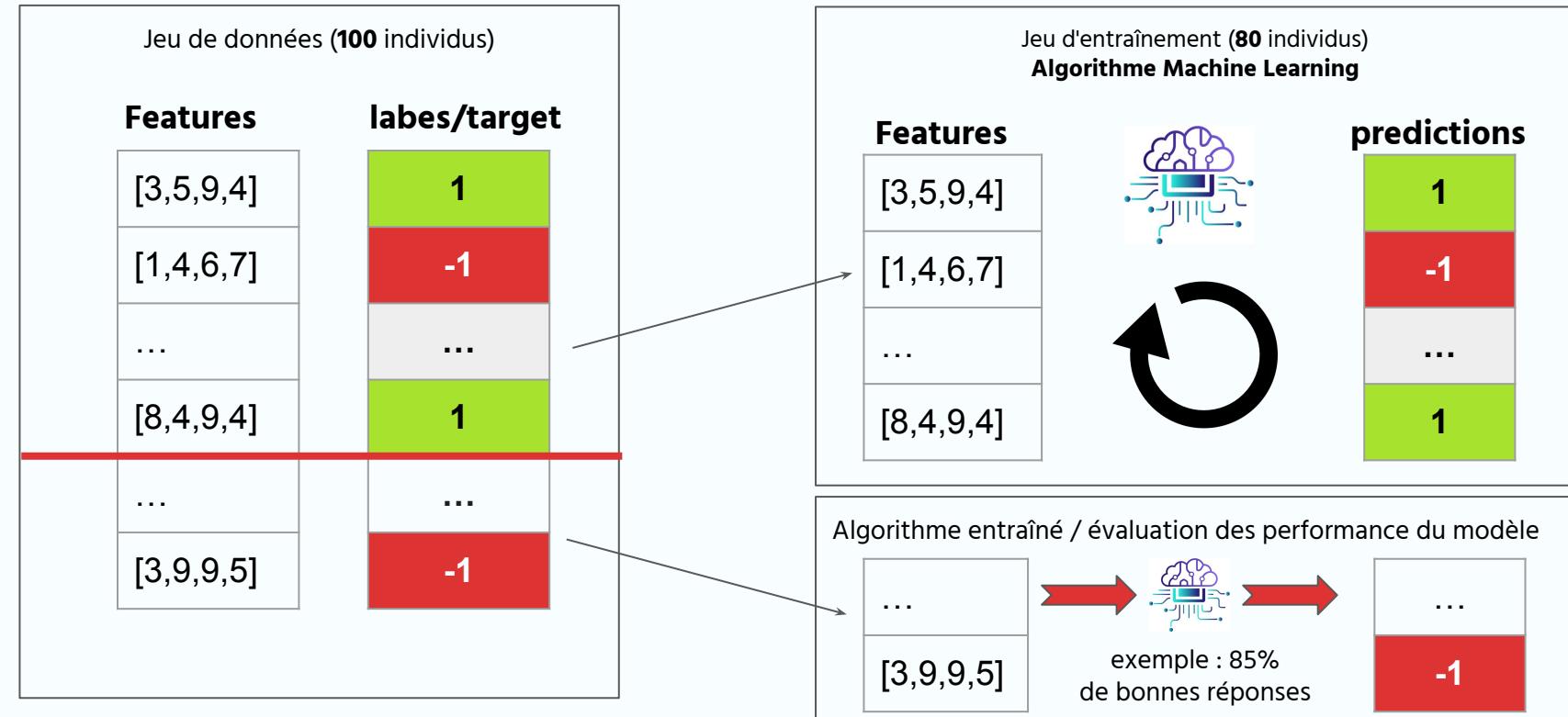
Classification : La targette à prédire est une classe

Clustering : Rassemblement de données en groupe homogène.

Réduction de dimension/Embedding : Changement des dimensions d'un jeu de données sans perte d'information.

4. Machine Learning

Le machine learning consiste à envoyer des données (features/labels) à un algorithme pour qu'il ajuste de lui-même ses coefficients afin d'améliorer les performances de ses prédictions.



[Lien du jeu de données](#)

Corrections à 17h

TP3 - Machine Learning avec Scikit-Learn

L'objectif de ce TP est d'entraîner un modèle de machine learning avec la bibliothèque Scikit-learn.

- Accédez au fichier suivant en utilisant Google Colab :
[TD3 - Machine Learning](#)
- Suivez les instructions présentes dans le notebook.

[TP Bonus - MINST Dataset](#)



2. Data Science

Correction à 17h

Les institutions



data.gouv.fr



Les plateformes tiers



Les entreprises



DEMANDEZ
NOUS
LA VILLE



Analyse exploratoire de données

Utilisez l'une des plateformes proposées ci-contre pour récupérer un jeu de données au format CSV, puis analysez les données en 3 parties dans un notebook :

1. **Exploration des données**
2. **Visualisation des données (au moins 2 graphiques)**
3. **Nettoyage des données**
4. **Entraîner un modèle de machine learning à prédire l'une de vos variables.**

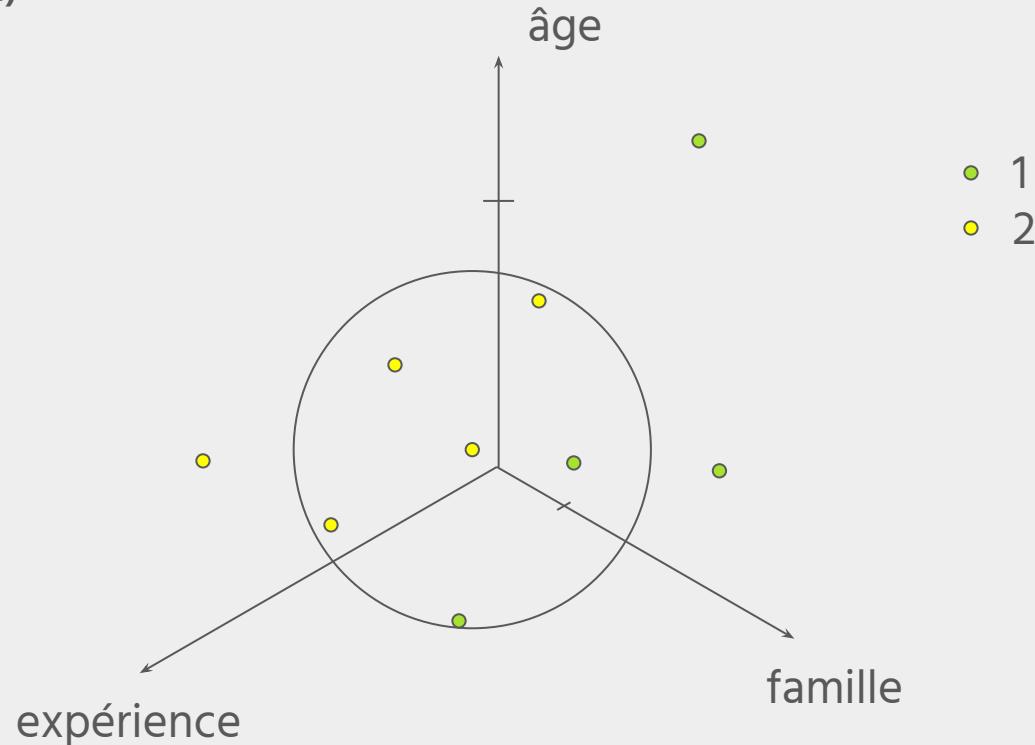
Déposez votre analyse exploratoire [ici](#).

4. Machine Learning

Exercices corrigés

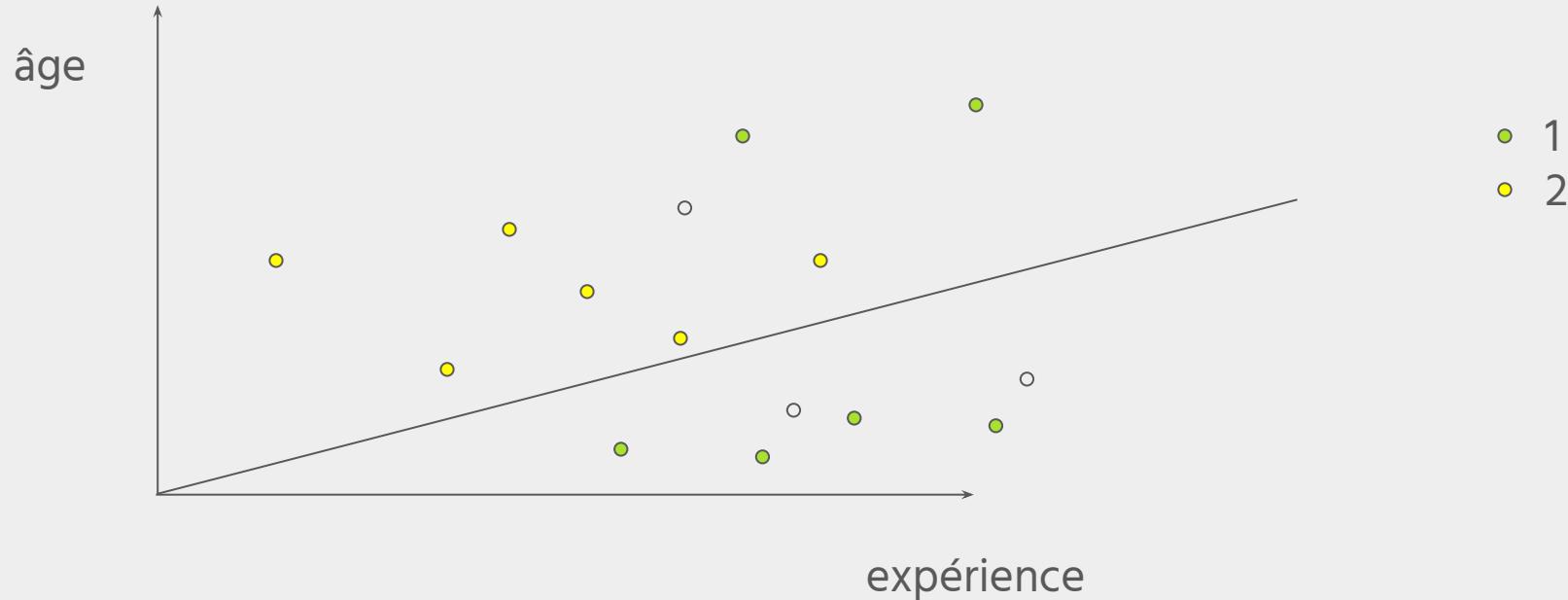
KNN - K Nearest Neighbors

(vote des k-voisins les plus proches)



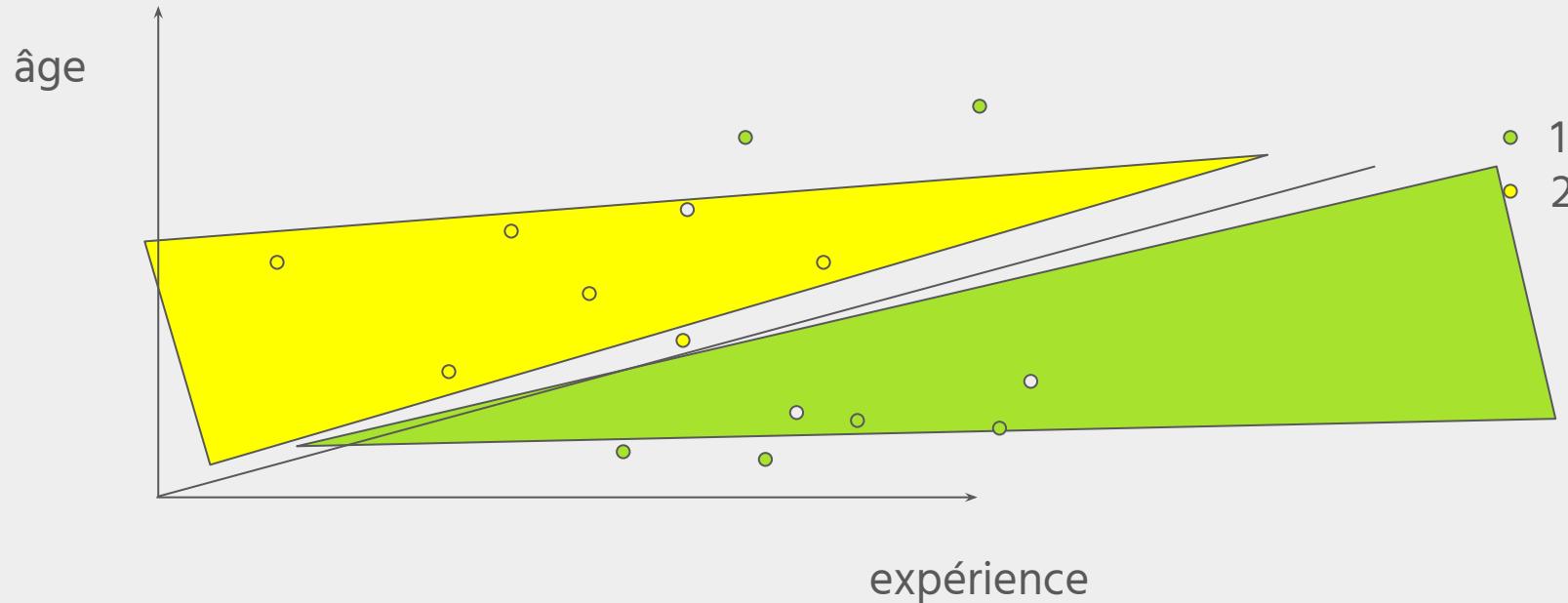
LogisticRegression

(Régression logistique)



LogisticRegression

(Régression logistique)

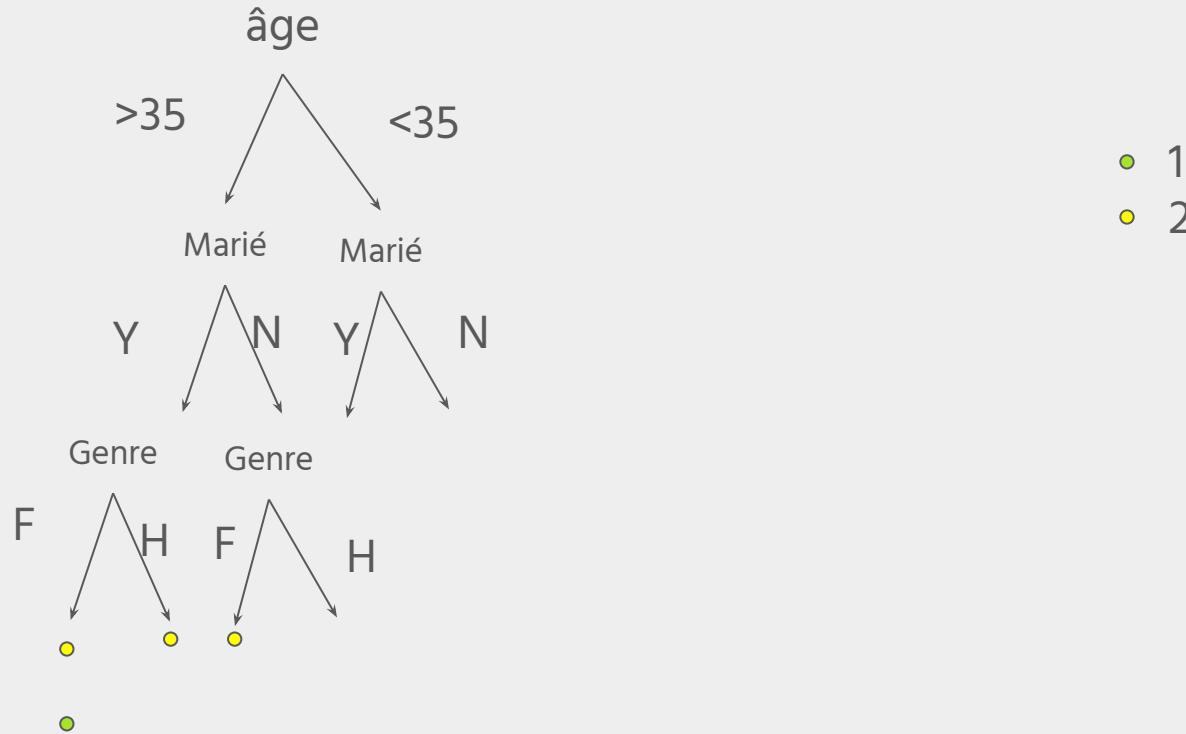


4. Machine Learning

Exercices corrigés

Decision Tree Classifier

(Arbre de décision)



05

Deep Learning

5. Deep Learning



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 humaine » selon le [Larousse](#).

Informatique

Cybersécurité



Réseau



Webmarketing



Intelligence Artificielle



Apprentissage par renforcement

Algorithme évolutionniste

Machine Learning ([lien](#))

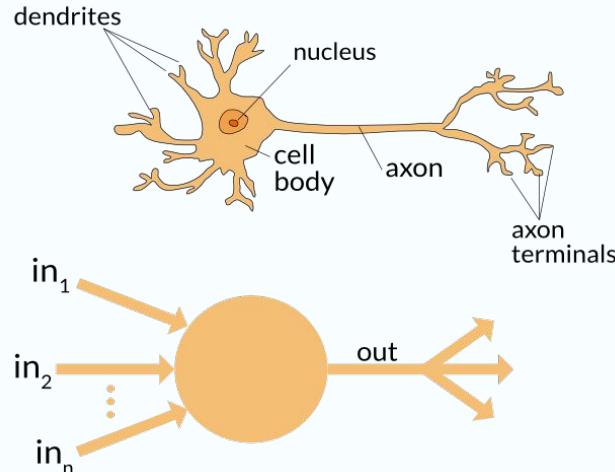
DeepLearning

IA Générative

5. Deep Learning

Le neurone artificiel : Le Perceptron

→ Accédez au fichier suivant en utilisant Google Colab :
[TD4 - Deep Learning](#)



Un perceptron est un algorithme d'apprentissage automatique supervisé, développé en 1957 par [Frank Rosenblatt](#), il est considéré comme la plus petite unité de base dans un réseau de neurones.

Un neurone est la cellule de base du système nerveux, responsable de la transmission de l'information électrique dans le corps. Un neurone est constitué de trois parties principales : le corps cellulaire, les dendrites et l'axone.

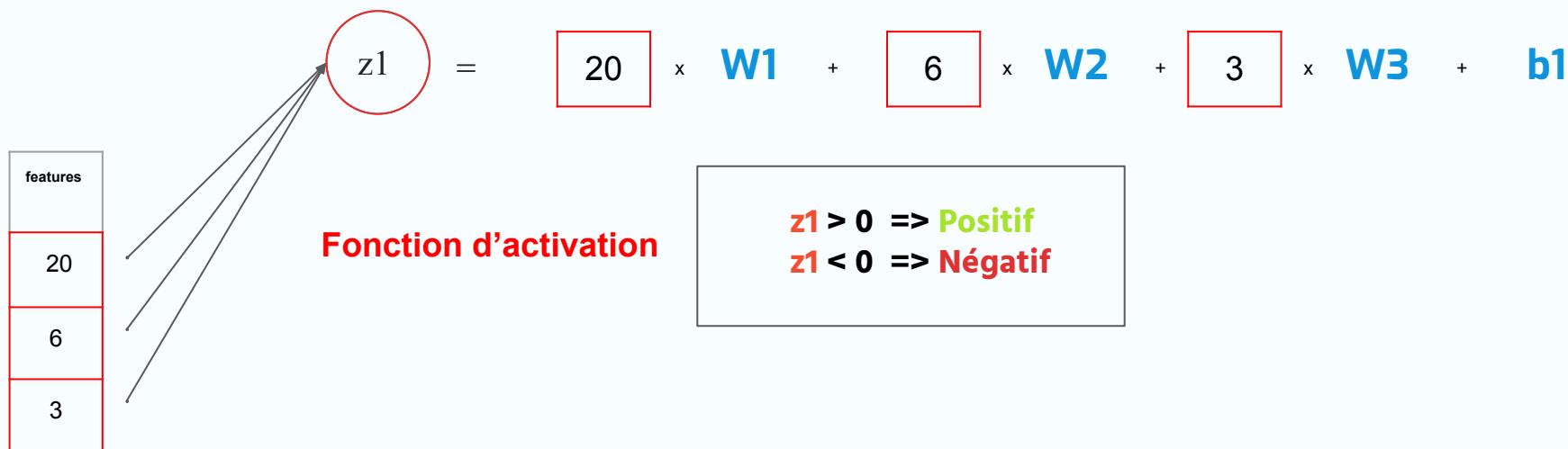
Les dendrites reçoivent les signaux électriques provenant d'autres neurones ou de l'environnement, et les transmettent au corps cellulaire. Le corps cellulaire intègre les signaux entrants et, si la somme de ces signaux dépasse un seuil de déclenchement, il envoie un signal électrique le long de l'axone, qui transmet l'information à d'autres neurones.

5. Deep Learning

Paramètres d'un modèle

Input

Un paramètre désigne une valeur ajustable dans un modèle qui est apprise à partir des données durant l'entraînement.



5. Deep Learning

Mise à jour des poids du perceptron.

$$W' = W + \alpha (Y_{\text{true}} - Y_{\text{pred}}) \times X$$

Où W' sont les **poids** actualisés, α est appelé le **learning rate** et Y le label prédit et réel.

$$W'_{(i)} = W_{(i)} + \alpha (Y_{(i)\text{true}} - Y_{(i)\text{pred}}) \times X_{(i)}$$

Les poids sont mis-à-jour à chaque prédition, ou après chaque lot de données (également appelé batch).

5. Deep Learning

Entrainement d'un algorithme pour le financement d'un prêt bancaire



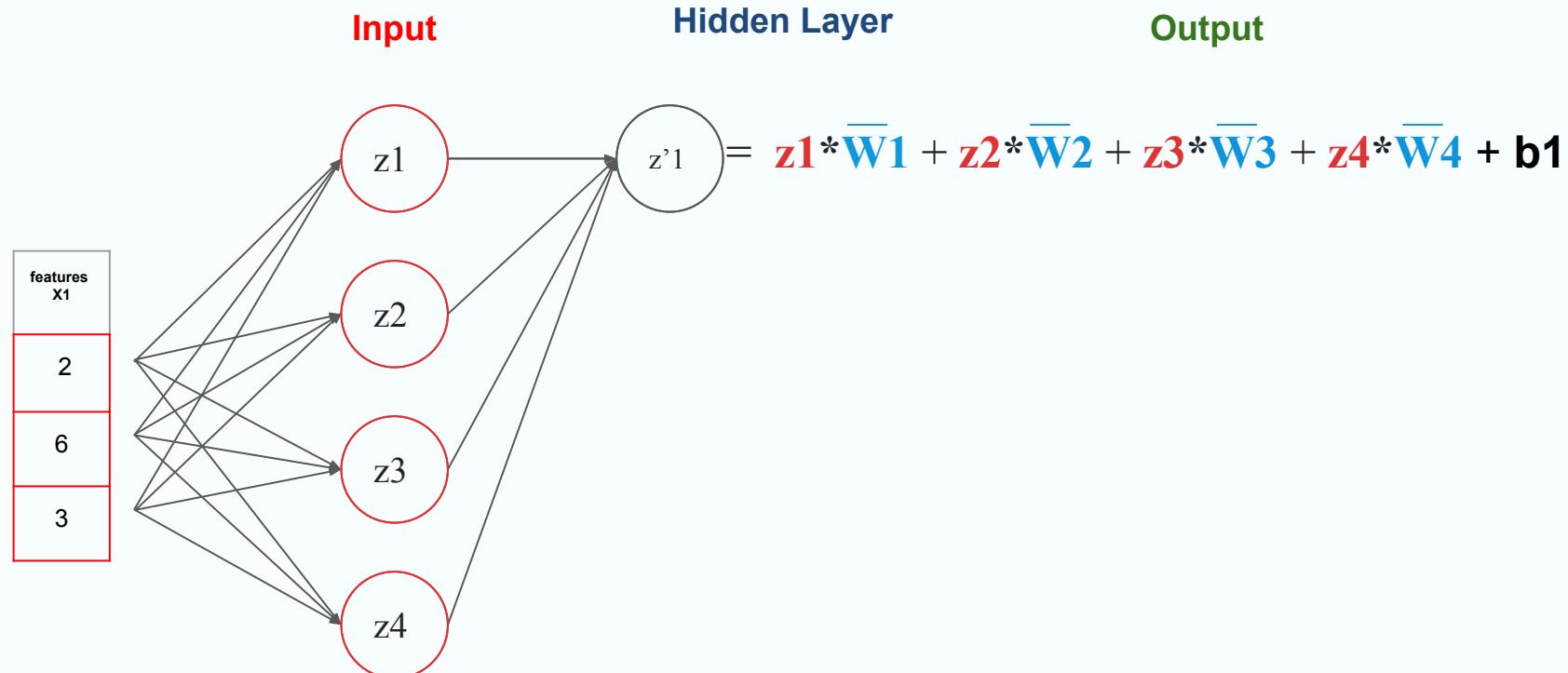
5. Deep Learning

Entrainement d'un algorithme pour le financement d'un prêt bancaire



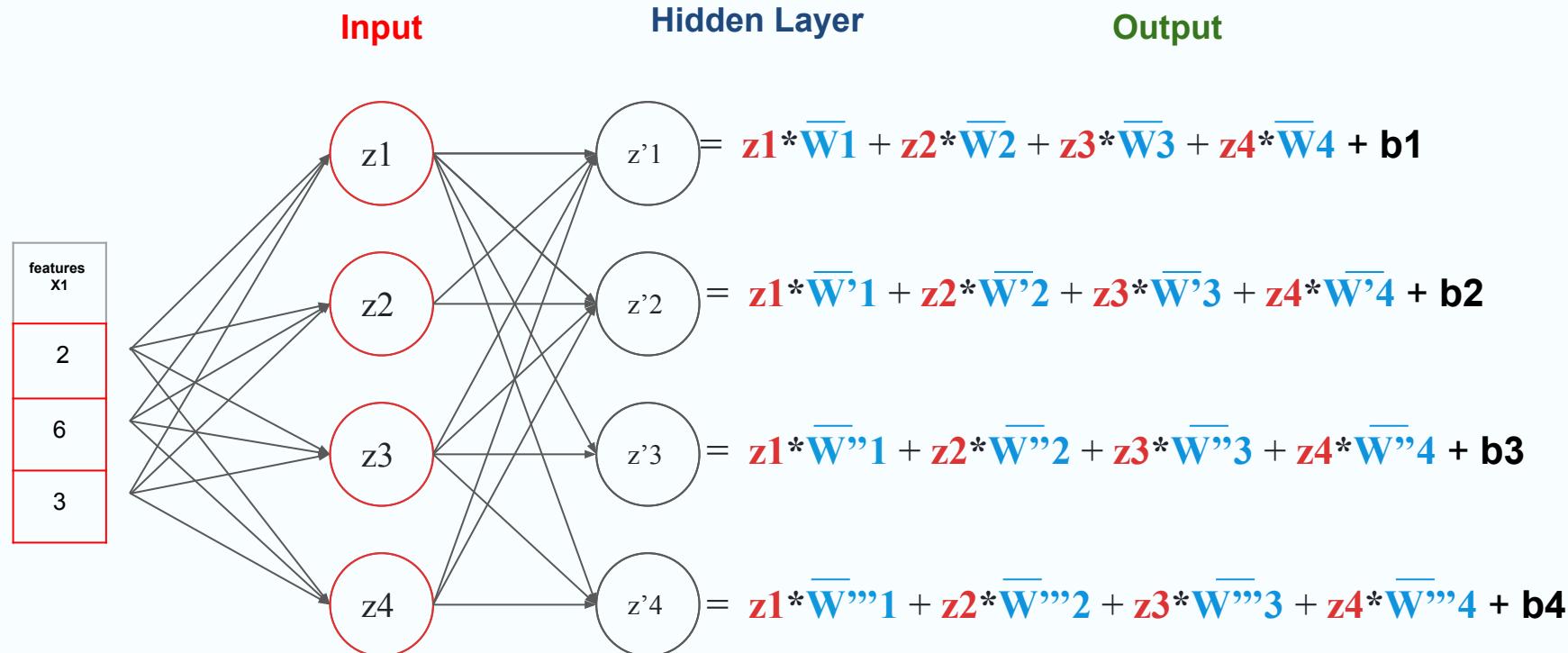
5. Deep Learning

Entrainement d'un algorithme pour le financement d'un prêt bancaire



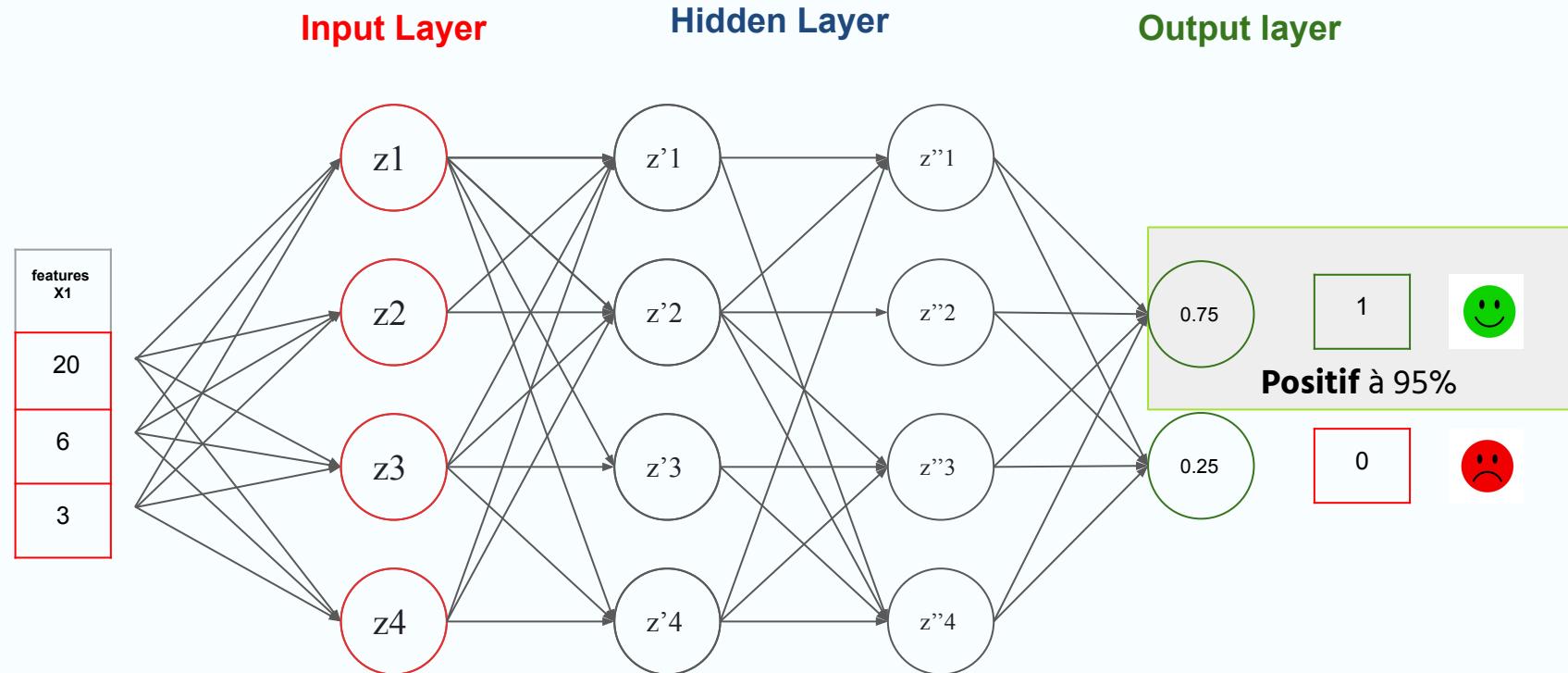
5. Deep Learning

Entrainement d'un algorithme pour le financement d'un prêt bancaire



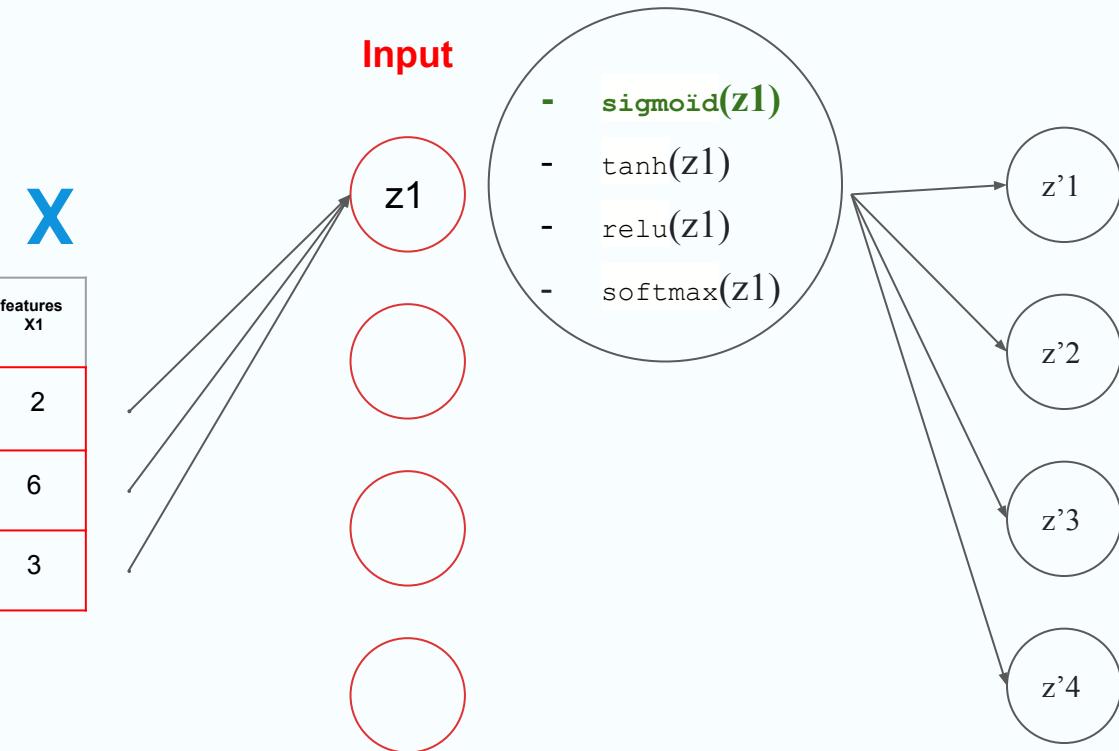
5. Deep Learning

Entrainement d'un algorithme pour le financement d'un prêt bancaire



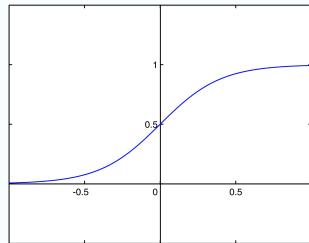
5. Deep Learning

Fonctions d'activation

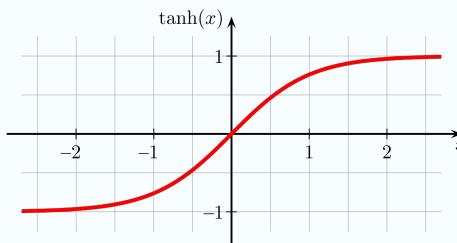


5. Deep Learning

Fonctions d'activation



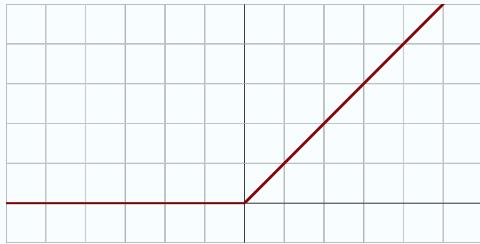
[Fonction sigmoïde](#) : une fonction en forme de S qui comprime les valeurs d'entrée entre 0 et 1, souvent utilisée pour la classification binaire ou la régression logistique.



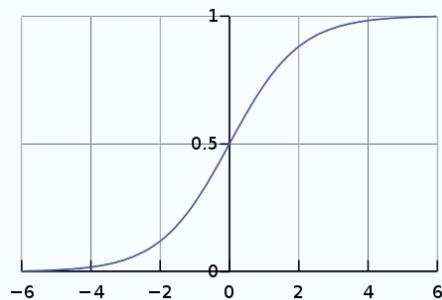
[Fonction tangente hyperbolique](#) (\tanh) : une fonction similaire à la fonction sigmoïde, mais qui comprime les valeurs d'entrée entre -1 et 1.

5. Deep Learning

Fonctions d'activation



Fonction ReLU (Rectified Linear Unit) : une fonction qui renvoie la valeur d'entrée si elle est positive, sinon renvoie 0. Cette fonction est souvent utilisée pour les réseaux de neurones profonds en raison de sa simplicité et de son efficacité.



Fonction softmax : une fonction qui renvoie un vecteur de probabilités normalisées, souvent utilisée pour la classification multi-classes.

5. Deep Learning

Descente de gradient

Une fonction de perte (ou fonction d'erreur) est une mesure utilisée pour évaluer la qualité des prédictions d'un modèle de machine learning.

Elle calcule l'écart entre les sorties prédites par le modèle et les sorties attendues (étiquettes de classe) pour un ensemble de données d'entraînement.

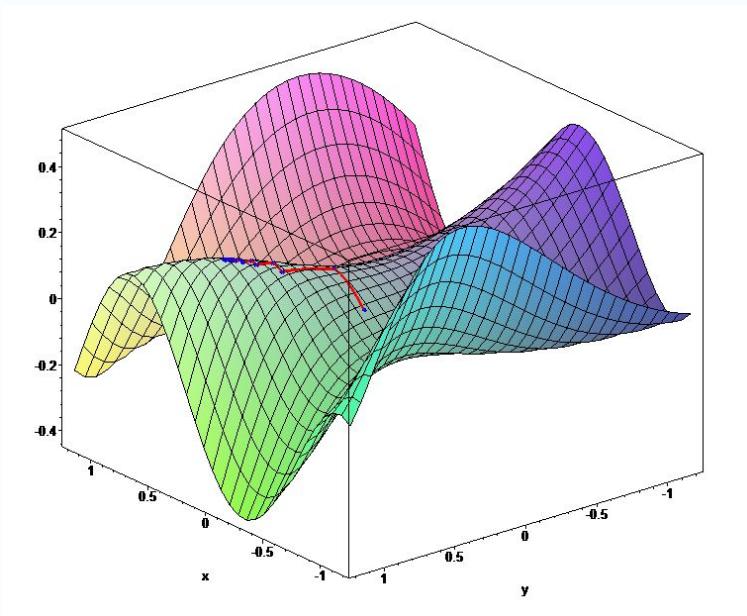
$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

L'objectif de l'apprentissage automatique est de minimiser cette fonction de perte, c'est-à-dire de trouver les paramètres du modèle qui minimisent l'écart entre les prédictions et les étiquettes de classe réelles.

$$W' = W + \alpha (Y_{\text{true}} - Y_{\text{pred}}) \times X$$

5. Deep Learning

Descente de gradient



$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

La descente de gradient est un algorithme d'optimisation qui agit sur les poids et les biais (ou "biais de neurones") du réseau de neurones.

L'objectif de la descente de gradient est de minimiser la fonction de perte du modèle en ajustant les poids et les biais du réseau de neurones.

Pour ce faire, elle calcule le gradient de la fonction de perte par rapport à chaque poids et biais du réseau de neurones et utilise ce gradient pour ajuster progressivement les valeurs des poids et des biais dans la direction de la pente descendante de la fonction de perte.

05

Introduction au NLP

Planning de la formation

mercredi, 17 janvier

9h - 17h30

IA & Datasience :

- Introduction à l'intelligence artificielle
- Manipulation de donnée avec Numpy et Pandas.

lundi, 13 mai

9h - 17h30

Machine Learning :

- Machine Learning : Apprentissage Supervisé/Non supervisé
- Bibliothèque Scikit-Learn

mardi, 14 mai

9h - 17h30

DeepLearning :

- Conception d'un perceptron
- Création d'un réseau de neurones profond avec la bibliothèque Tensorflow.

lundi, 10 juin

9h - 17h30

Traitemet du langage Naturel (NLP) :

- Tokenization TF-IDF & WordEmbedding
- Réseau de neurone Récurrent (RNN, LSTM & GRU)
- Large modèle de langue (LLM) : Bert, Mistral et GPT.

mardi, 11 juin

9h - 17h30

Computer Vision : Vision par Ordinateur (CV) :

- Type d'application : Classification, détection, segmentation
- Réseau de neurone à convolution
- Architecture Resenet
- Modèles YOLO.

mercredi, 12 juin

9h - 17h30

Projet individuel.

- Entrainement d'un modèle de machine learning/deep learning à partir d'un jeu de données.

5. Introduction au NLP

Les branches de l'IA



Vision par Ordinateur
Computer Vision

Analyse
prédictive

Robotique
IA Incarnée

NLP: Natural Language Processing



Les branches
de l'IA



IA Générative



Language Modeling

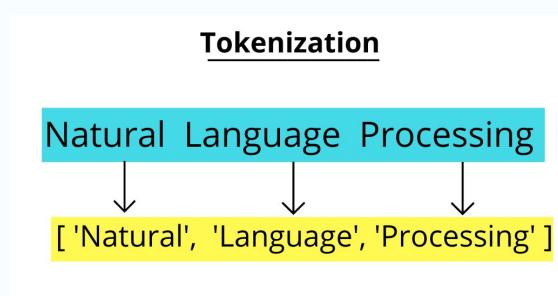


Agents autonomes



5. Introduction au NLP

Le traitement du langage naturel (**NLP** en anglais pour natural language processing) est un champ de l'informatique, de l'intelligence artificielle et de la linguistique qui traite des interactions entre les ordinateurs et les langues humaines (naturelles).



Le but du NLP est de permettre aux ordinateurs de comprendre, d'interpréter et de générer des langues humaines.

Dans ce cours nous verrons les notions de :

- Text Preprocessing (Token, Stop Word, Stemming/lemmatization, Bag of Words).
- Term Frequency, Inverse Document Frequency, Word-Embedding, Text Similarity.

05

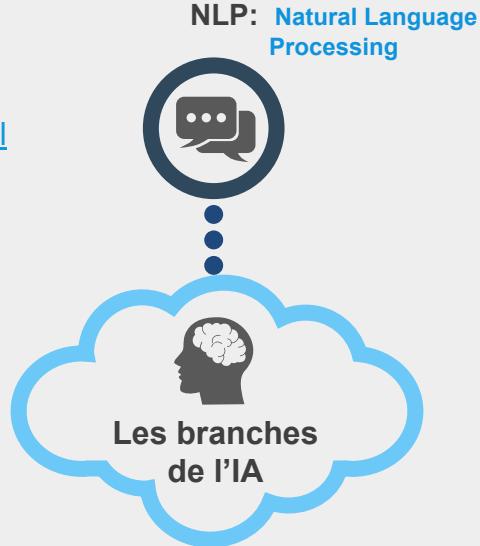
Correction 17h

5. Introduction au NLP

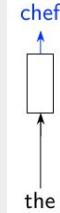
Les branches de l'IA

Support de cours :

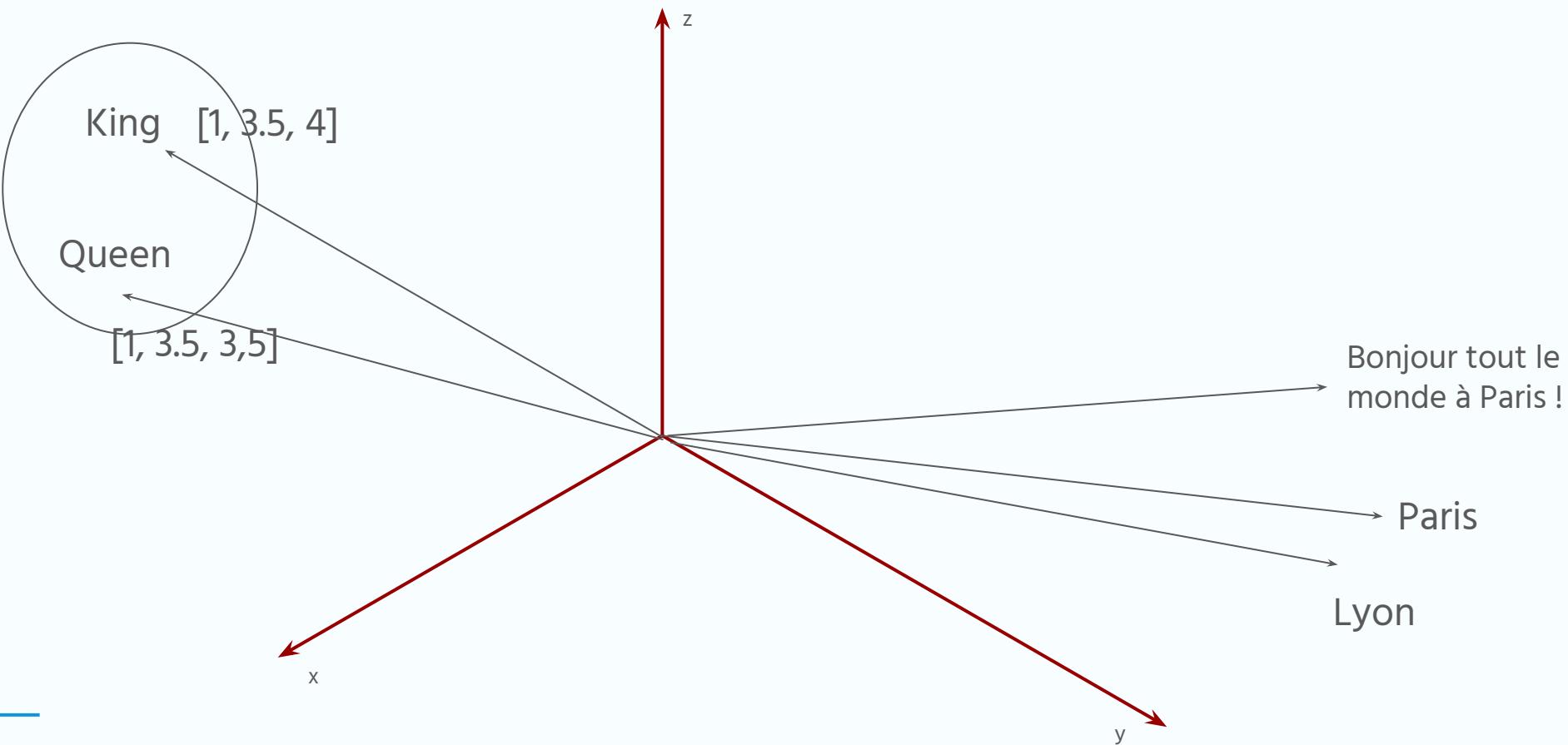
- [5. Traitement du langage naturel](#)
- [6. WordEmbedding](#)



Language Modeling



5. Introduction au NLP



06

Computer Vision

2. Computer Vision

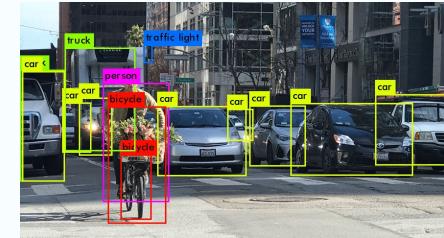
Computer Vision : Vision par ordinateur

La vision par ordinateur, également appelée "computer vision" en anglais, est un domaine de l'informatique qui s'intéresse à la réalisation de tâches visuelles par des ordinateurs. Cela inclut la capture, l'analyse et la compréhension d'images et de vidéos par des ordinateurs.

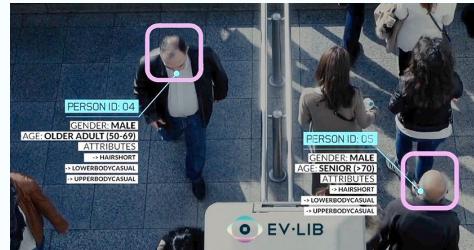
Classification



Détection



Identification



Segmentation



2. Computer Vision

Les branches de l'IA



Vision par Ordinateur
Computer Vision



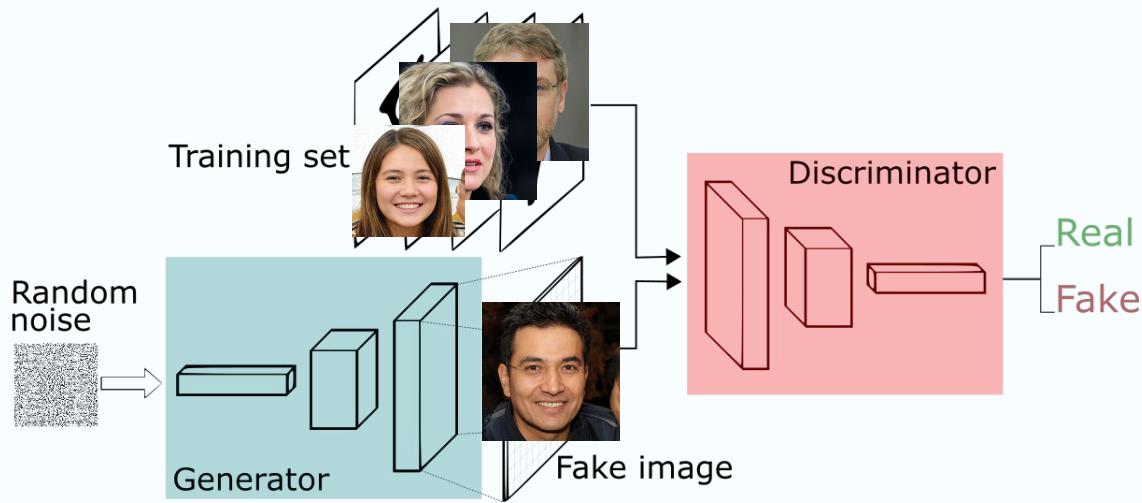
Support de cours :

- [7. Réseau de neurone à convolution](#)

2. Computer Vision

GAN : Generative Adversarial Network

*Réseaux antagonistes génératifs



Ressources : [Dall-E](#) : Open IA | [This person does not exist](#) | [Deepfakesweb](#) | [Generative.fm](#)

Planning de la formation

mercredi, 17 janvier	lundi, 13 mai	mardi, 14 mai
9h - 17h30	9h - 17h30	9h - 17h30
IA & Data Science : Numpy et Pandas	Machine Learning, Deep Learning : Sklearn et Tensorflow	Développement d'une API : FastAPI
<u>TD1. Maîtrise de Numpy</u> <u>TD2. Maîtrise de Pandas</u>	<u>TD3. Machine Learning</u> <u>TD4. Deep Learning</u>	<u>Projet FastAPI</u>
lundi, 10 juin	mardi, 11 juin	mercredi, 12 juin
9h - 17h30	9h - 17h30	9h - 17h30
Traitements du langage Naturel (NLP) : NLTK, HuggingFace, OpenAI	Computer Vision : Vision par Ordinateur (CV).	Projet de groupe.
<u>TD5. Traitement du langage naturel</u> <u>TD6. Word Embedding</u>	<u>TD7. Réseau de neurone à convolution</u>	

6. Evaluation - Machine Learning/Deep Learning

Objectif : Création d'une API & Entraînement d'un modèle

Projet à rendre [ici](#).
Au plus tard le **05/07/24**

Créez un dossier projet intitulé : Nom + Prénom	1pts
Créez un fichier api.py qui contiendra l'api développée avec la bibliothèque FASTAPI .	1pts
L'api possède une page de documentation complète : un titre, une présentation, des endpoints classés par thème et possédant une description.	2pts
L'api possède un endpoint POST training , qui permet de recevoir un jeu de données et d'entraîner un modèle sur les données envoyées.	
Le modèle entraîné peut être un modèle de Sklearn ou Tensorflow, le jeu choix du jeu de donnée est libre.	4pts
L'api possède un endpoint POST predict qui permet de faire une prédiction à partir du dernier modèle sauvegardé.	
Le point de terminaison possède dans sa documentation un exemple de requête avec la structure de donnée appropriée.	4pts
L'api possède un point de terminaison GET model qui fait appel soit à l'API de OpenAI , soit celle de HuggingFace .	4pts
L'api permet de gérer les erreurs en cas de données invalides, pour l'entraînement ainsi que pour la prédiction.	2pts
Le projet possède au maximum 6 fichiers, dont les fichiers suivants :	
<ul style="list-style-type: none">- api.py : le fichier de l'api.- requirements.txt : les dépendances nécessaires au projet.- function.py : le fichier contenant les fonctions principales.- app.py : l'application front développée avec la bibliothèque Streamlit permettant d'entrer en interaction avec l'API.- model : le modèle entraîné utilisé dans le point predict.- data.csv : le jeu de données utilisé pour l'entraînement.- Notebook.ipynb : le notebook contenant votre code brouillon.	2pts
Bonus : Le projet possède une application Streamlit permettant d'entrer en interaction avec l'API et de télécharger le modèle entraîné.	4pts