Gestion de Projet Big Data & Développement d'applications Big Data

EDAH Kodjo
Consultant Systèmes d’Information, Big-Data

Objectifs

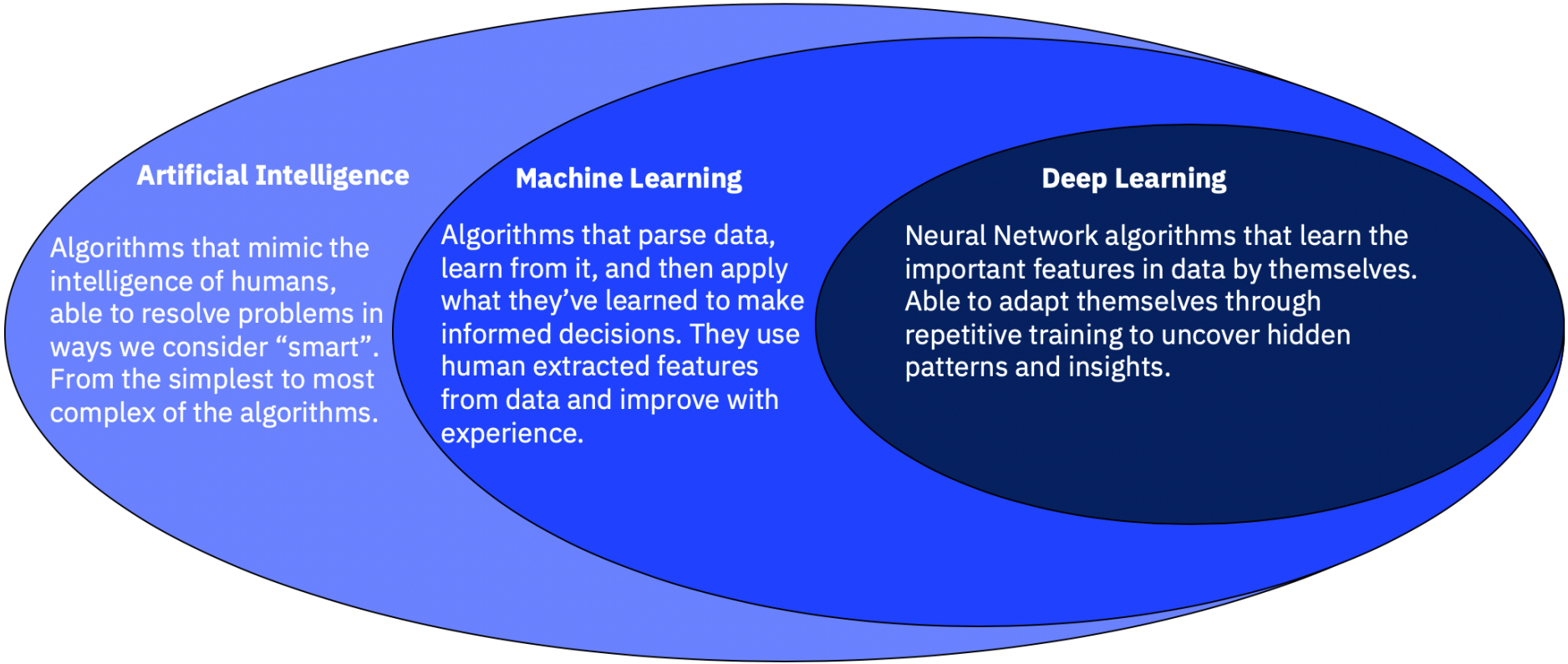
- Comprendre la notion et les spécificités du Big Data
- Connaître les technologies de l’écosystème Hadoop
- Connaître le langage python et utiliser les librairies de machine learning
- Savoir utiliser les outils de visualisation des données (Dataviz)

Partie 3 : Machine learning

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 »

Machine Learning et intelligence artificielle



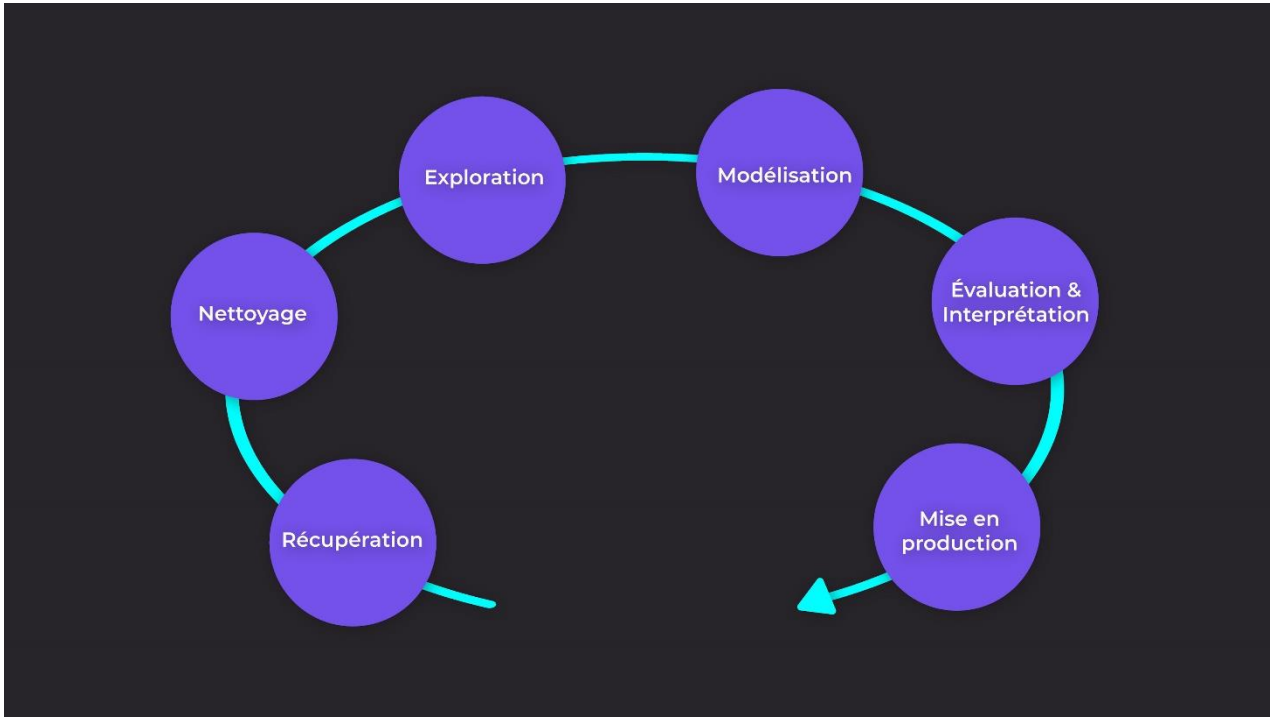
Intelligence artificielle

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 »

Exemple de problématique de machine learning

- Prédire les ventes
- Identification des objets (image, ……)
- Segmenter les utilisateurs d'un site en plusieurs groupes en fonction de leur comportement sur le site, catégoriser un produit
- Recommandation de produit

Le cycle de travail d'un data science

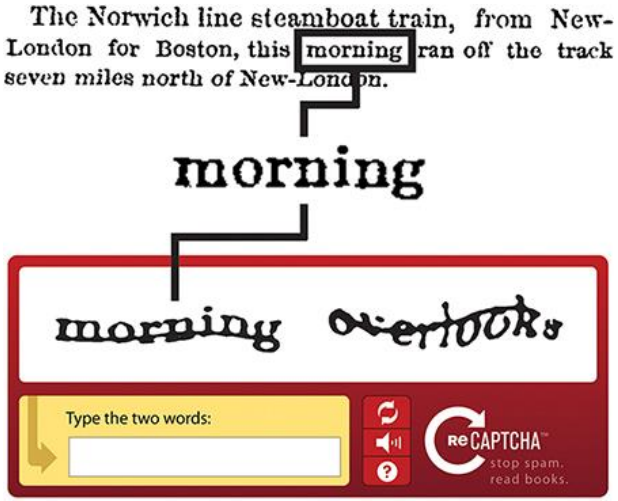


https://user.oc-static.com/upload/2019/04/23/15560369383553_2c2.jpg

Récupération des données

- Les bases de données existantes
- Les données brutes alternatives (image, son, document, pages web, etc.)
- Les réseaux sociaux
- Internet des Objets
- Création de nouveaux canaux d'acquisition de données

- Exemple : Les CAPTCHAs pour la digitalisation automatique de livres



https://user.oc-static.com/upload/2016/09/17/14741229513738_img-2.png

Nettoyage des données

- Suppression des données aberrantes et incohérentes
- Agrégation si nécessaire

- Les batchs ou job map-reduce, spark

Exploration des données

- Comprendre les différents comportements
- Détecter les schémas
- Tâche destinée au Data Analyst

- Résultats
  - Proposer plusieurs hypothèses sur les causes sous-jacentes à la génération du dataset
  - Proposer plusieurs pistes de modélisation statistique des données, qui vont permettre de résoudre la problématique de départ considérée.
  - Proposer si nécessaire de nouvelles sources de données qui aideraient à mieux comprendre le phénomène.

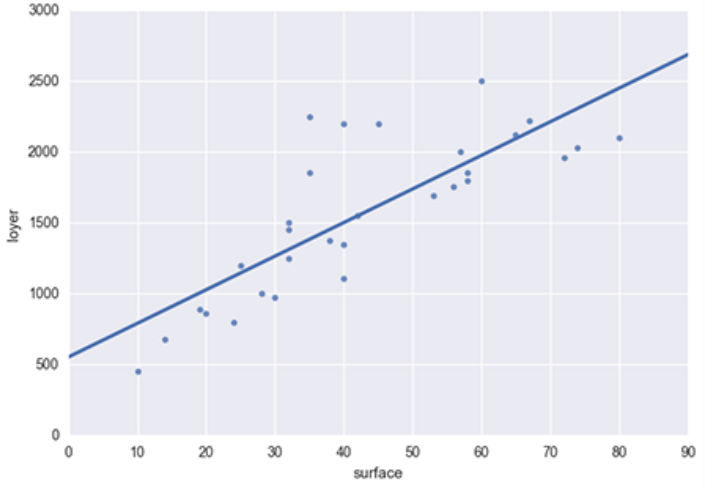
Modélisation

- C’est l’étape du machine learning ou apprentissage
- Application des algorithmes de d’apprentissage
  - la régression linéaire
  - K-nn
  - les Support Vector Machine (SVM)
  - les réseaux de neurones
  - les random forests.
  - Clustering
  - Collaborive filtering

PyTorch

scikit learn

TensorFlow

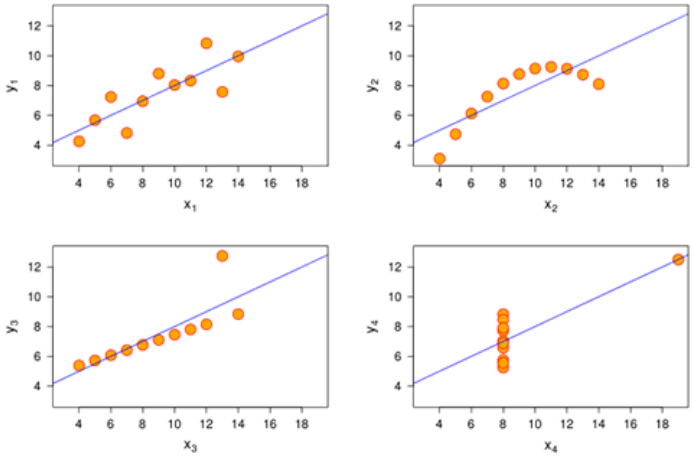


https://user.oc-static.com/upload/2016/09/17/14741406902223_download-2.png

Evaluation du modèle

- Le modèle représente t-il avec exactitude le phénomène ?
- Le modèle résout t-il le problème ?
- Quelle est la marge d'erreur ?
- Quelle est la performance du modèle

- Le quartet d'Anscombe



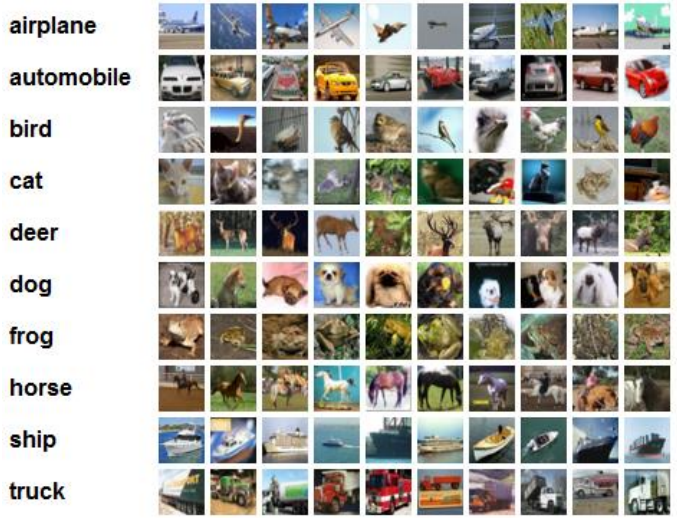
https://user.oc-static.com/upload/2016/09/17/14741418471714_640px-Anscombe.svg.png

Mise en production

- Déploiement du modèle en production
- Mise en place des supports de production
- Infrastrusture big data (Hadoop, AWS, Azure ……)

Familles d'algorithmes d'apprentissage existantes

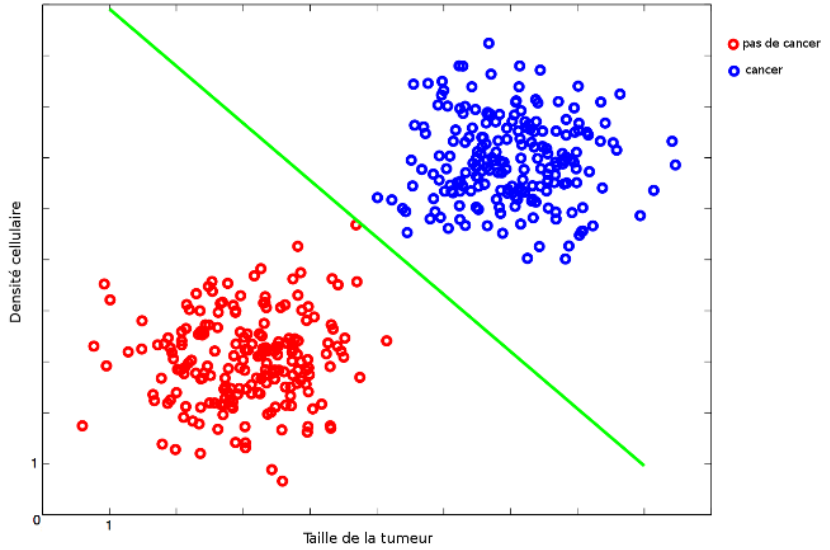
- Apprentissage « supervisé » : supervised learning
- Données sont annotées ou labélisées
- Problème : comment labéliser ?



https://user.oc-static.com/upload/2016/10/24/14773158929787_cifar_preview.png

Familles d'algorithmes d'apprentissage existantes

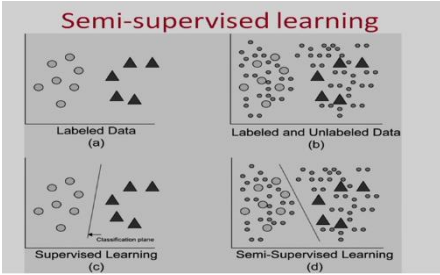
- Apprentissage « non supervisé » : unsupervised learning
- Les données ne sont pas annotées
- L'algorithme determine lui même les similarités dans le dataset



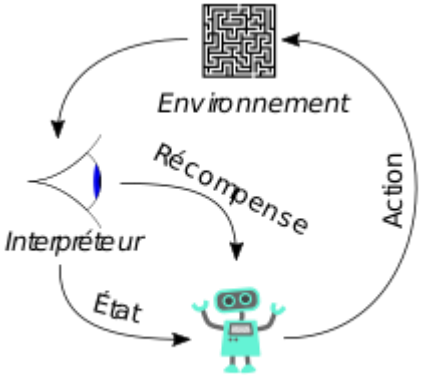
https://markdown.data-ensta.fr/uploads/upload_87ef9ad65f9163ff5a92e1691eb4d1bd.png

Familles d'algorithmes d'apprentissage existantes

- le semi-supervised learning : combine supervised et unsupervised



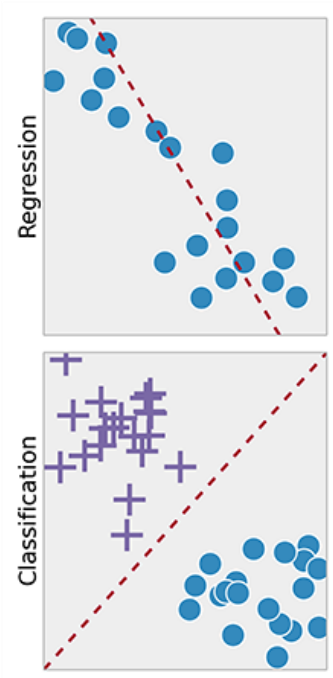
- le reinforcement learning : qui se base sur un cycle d'expérience / récompense et améliore les performances à chaque itération
  - Jeu de go, damier, échec



Source : https://www.digitalvidya.com/blog/semi-supervised-learning/ , https://upload.wikimedia.org/wikipedia/commons/thumb/1/1b/Reinforcement_learning_diagram.svg/250px-Reinforcement_learning_diagram.svg.png

Familles d'algorithmes d'apprentissage existantes

- Régression
  - Recherche t-on un nombre ?
  - Valeur continue
- Classification
  - Recherche t-on une catégorie ?
  - Valeur discrète



Source : https://user.oc-static.com/upload/2016/09/18/14742103795655_ml.png

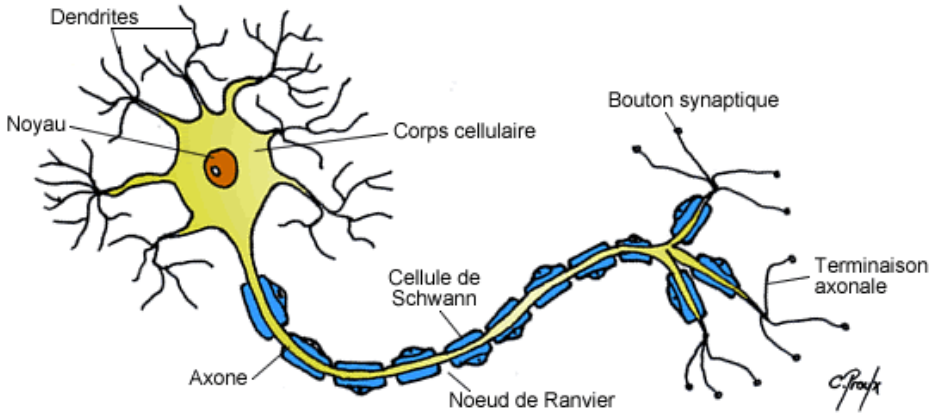
Segmentation des datasets

- Training set :
  - sous-ensemble destiné à l'apprentissage d'un modèle.
  - Exemple : Proportion 80% du dataset
- Validation set/ Test set
  - sous-ensemble destiné à l'évaluation du modèle.
  - Exemple : Proportion 20% du dataset

- N'effectuez jamais l'apprentissage sur des données d'évaluation

Les réseaux de neurones artificiels

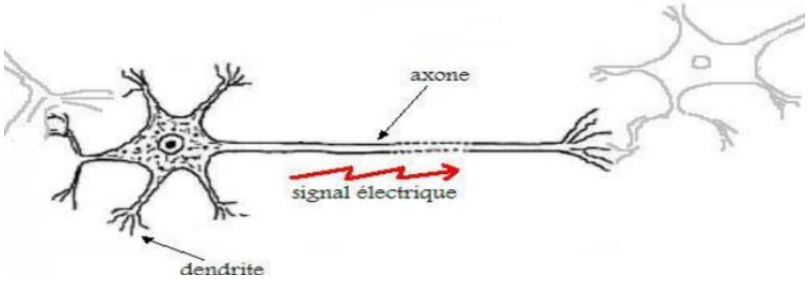
- Inspiré des réseaux de neurones humains



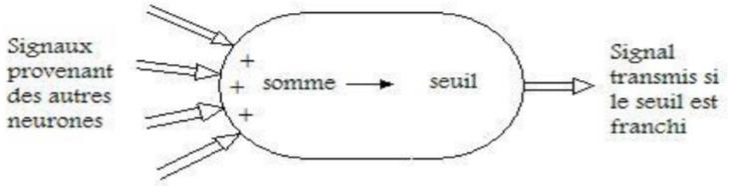
http://ressources.unisciel.fr/DAEU-biologie/P2/res/chap4_im05.png

Métaphore biologique

- Fonctionnement du cerveau Transmission de l'information et apprentissage

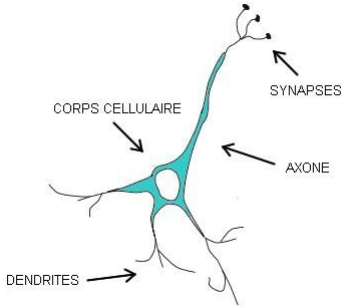


- Un perceptron !!!!

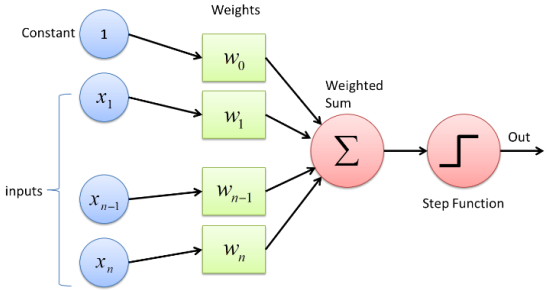


Un neurone : le perceptron

- Les neurones reçoivent des signaux (impulsions électriques) par les dendrites et envoient l ’information par les axones.
- Les contacts entre deux neurones (entre axone et dendrite) se font par l ’intermédiaire des synapses.

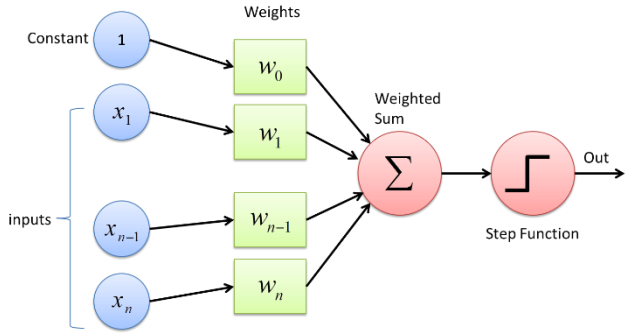


Neurone biologique	Neurone artificiel
Axones	Signal de sortie
Dendrites	Signal d'entrée
Synapses	Poids de la connexion



Un neurone : le perceptron

- Frank Rosenblatt en 1956
- Algorithme d'apprentissage supervisé de classifieurs binaires
- Inconvénient : linéarité



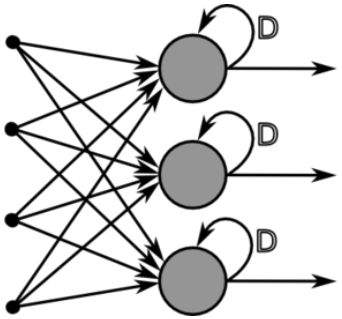
https://towardsdatascience.com/what-the-hell-is-perceptron-626217814f53

Un perceptron à *n* entrées $(x_1, \ldots, x_n)$ et à une seule sortie *o* est défini par la donnée de *n* poids (ou coefficients synaptiques) $(w_1, \ldots, w_n)$ et un biais (ou seuil) θ par[2] :

$$o = f(z) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^{n} w_i x_i > \theta \\ 0 & \text{sinon} \end{cases}$$

Un neurone : le perceptron

- La règle de Hebb
- Correction du modèle (loi de Widrow-Hoff)
- Inconvénient : linéarité



$$W_i' = W_i + \alpha(Y_t - Y)X_i$$

W_i' = le poids i corrigé

Y_t = sortie attendue

Y = sortie observée

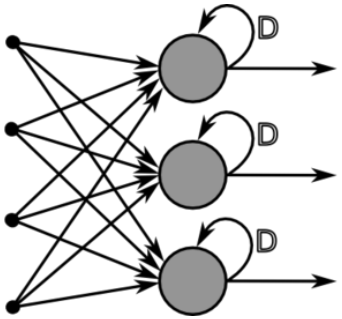
α = le taux d'apprentissage

X_i = l'entrée du poids i pour la sortie attendue Y_t

W_i = le poids i actuel

Un neurone : le perceptron

- La règle de Hebb
- Correction du modèle (loi de Widrow-Hoff)
- Inconvénient : linéarité



$$W_i' = W_i + \alpha(Y_t - Y)X_i$$

W_i' = le poids i corrigé

Y_t = sortie attendue

Y = sortie observée

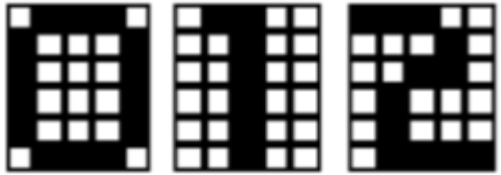
α = le taux d'apprentissage

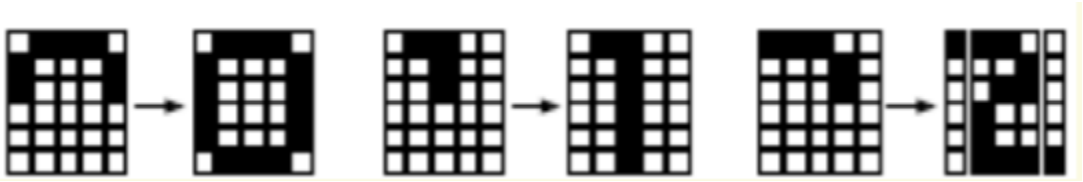
X_i = l'entrée du poids i pour la sortie attendue Y_t

W_i = le poids i actuel

Un neurone : le perceptron

- Exemple : reconstruction d'image
- Reconnaitre les chiffres 0, 1, 2

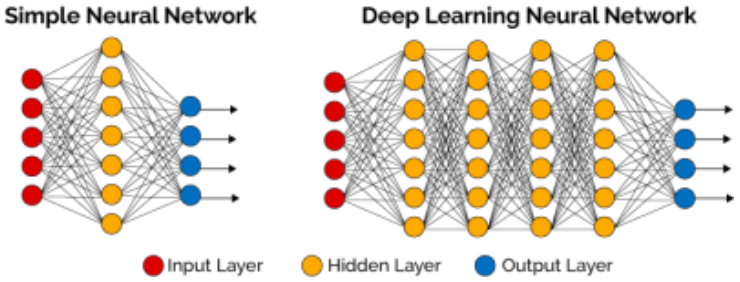


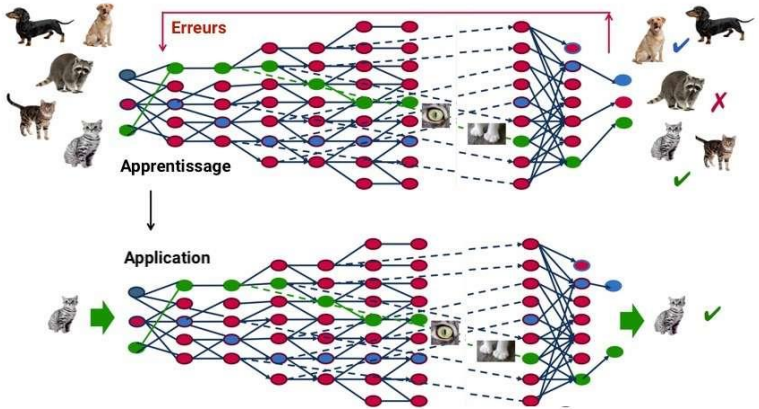


http://master-ivi.univ-lille1.fr/fichiers/Cours/rdf-semaine-8-neurones.pdf

Les réseaux de neurones : deep learning

- Réseau multi-couche

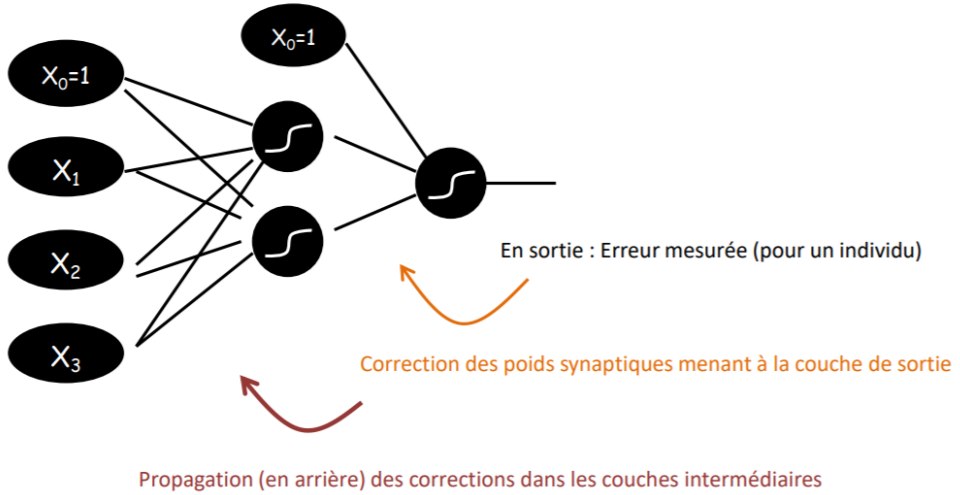




https://cdn.futura-sciences.com/buildsv6/images/mediumoriginal/d/c/d/dcdc8d74ca_125717_deep-learning.jpg

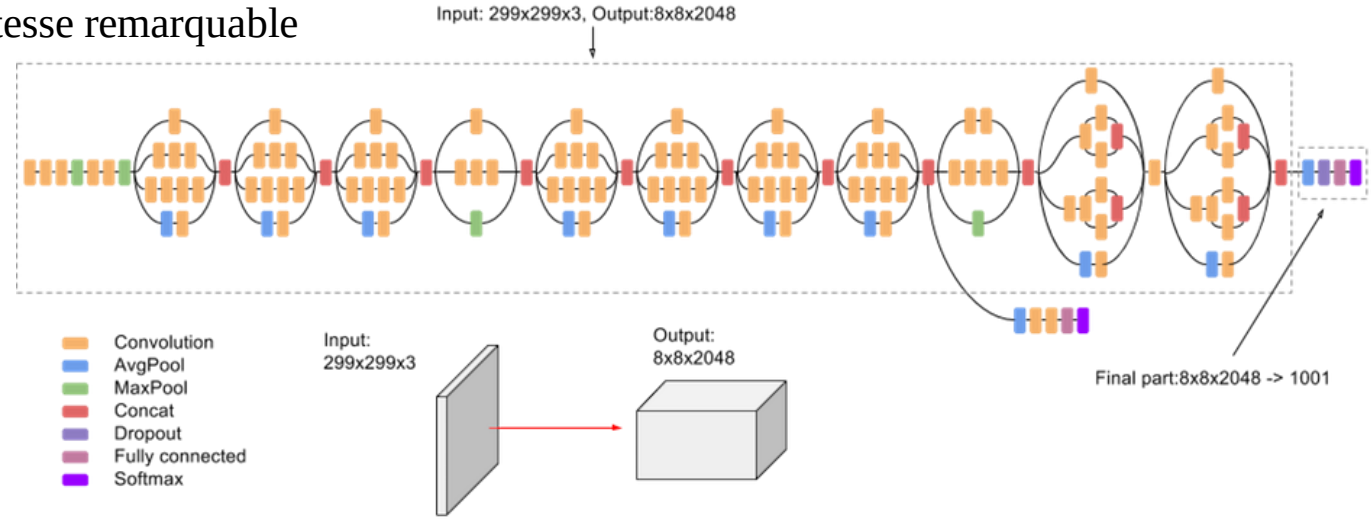
Les réseaux de neurones : deep learning

- La rétropropagation du gradient (backpropagation)
- Généraliser la règle de Widrow-Hoff – Rétropropagation



Les réseaux de neurones : deep learning

- Inception v3 sur Cloud TPU
- modèle de reconnaissance d'images
- Atteint une justesse remarquable



https://cloud.google.com/tpu/docs/images/inceptionv3onc--oview.png

TP : Python

Questions ?

Merci