Big Data Analytics

MOD 2.1 : Défis informatiques du Big Data

Emmanuel Dellandréa
emmanuel.dellandrea@ec-lyon.fr

ÉCOLE CENTRALELYON

ÉCOLE CENTRALELYON

Big Data Analytics

- De quoi s'agit-il ?
  - Développement d'applications à visée analytique qui traitent les données massives (Big Data) pour en tirer du sens
- Un champs immense d'applications :
  - Analyser les données en mouvement
  - Analyser une grande variété de données
  - Traiter un volume conséquent de données
  - Découvrir et expérimenter

ÉCOLE CENTRALELYON

Big Data – un immense champs d'applications

- Analyser les données en mouvement
  - La surveillance d'un grand nombre de nœuds de données (cloud, hébergement)
  - La sécurité électronique et la détection de fraude dans un contexte banque assurance
  - Le suivi en temps réel et à forte réactivité de clients (commerce de détail, téléassistance)
  - Le pilotage avancé de systèmes complexes (recherche en astronomie et en physique des particules, domaine spatial, armée)
  - L'analyse des logs
  - Le tracking des identifiants, connexions et pistage des transports

ÉCOLE CENTRALELYON

Big Data – un immense champs d'applications

- Analyser une grande variété de données
  - Le décodage et l'analyse des humeurs sur les réseaux sociaux
  - L'accostage avec la stratégie d'entreprise
  - Le suivi de phénomènes propagés (épidémies, intrusions, terrorisme)
  - L'analyse multimédia à partir de formats vidéo, audio, texte, photos (sécurité, traductions, médiamétrie)

ÉCOLE CENTRALELYON

Big Data – un immense champs d'applications

- Traiter un volume conséquent de données
  - Rapprochement d'objets métiers (produits, clients, déploiement, stocks, ventes, fournisseurs, lieux promotionnels)
  - Détection de fraudes, repérage de clients indélicats ou manipulateurs
  - Management du risque, prise de décision sensible appuyée de modèles prévisionnels
  - Analyse de contexte, d'environnement

ÉCOLE CENTRALELYON

Big Data – un immense champs d'applications

- Découvrir et expérimenter
  - Approcher la notion de désir ou de sentiment déclencheur d'action
  - Cerner l'entourage social d'un client, les conditions favorables
  - Expérimenter l'impact d'un nouveau produit, son ressenti
  - Mesurer l'efficacité d'une stratégie de conquête, mesurer des écarts ou des erreurs de positionnement d'image
  - Profilage de nouveaux comportements
  - Identification de nouveaux facteurs d'influence

ÉCOLE CENTRALELYON

Plan

- MapReduce et Hadoop
- Exploration et préparation des données
- Le Machine Learning
- Quelques outils de Machine Learning pour le Big Data
- Un exemple d'application d'une méthode de Machine Learning

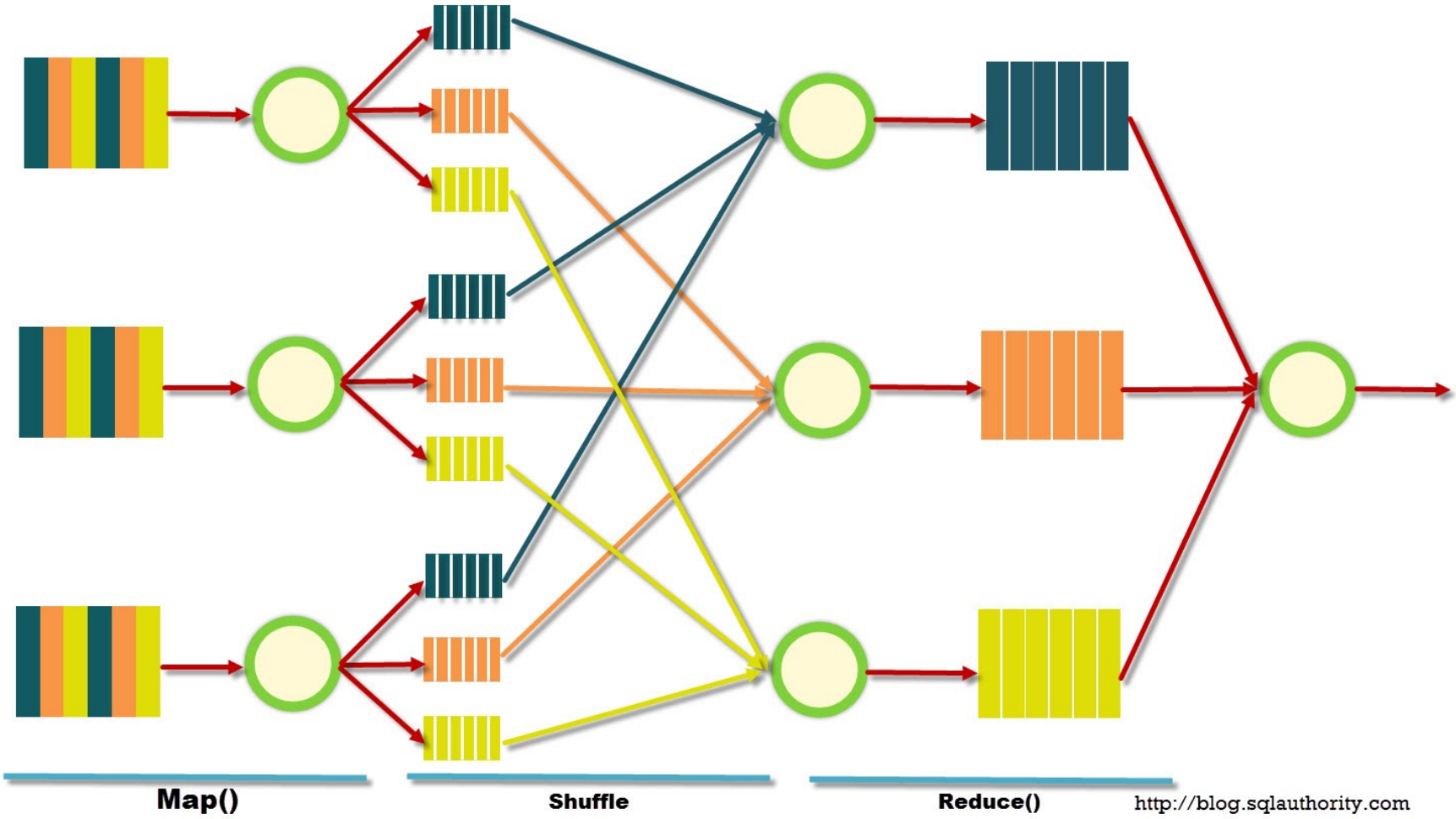
ÉCOLE CENTRALELYON

MAPREDUCE ET HADOOP

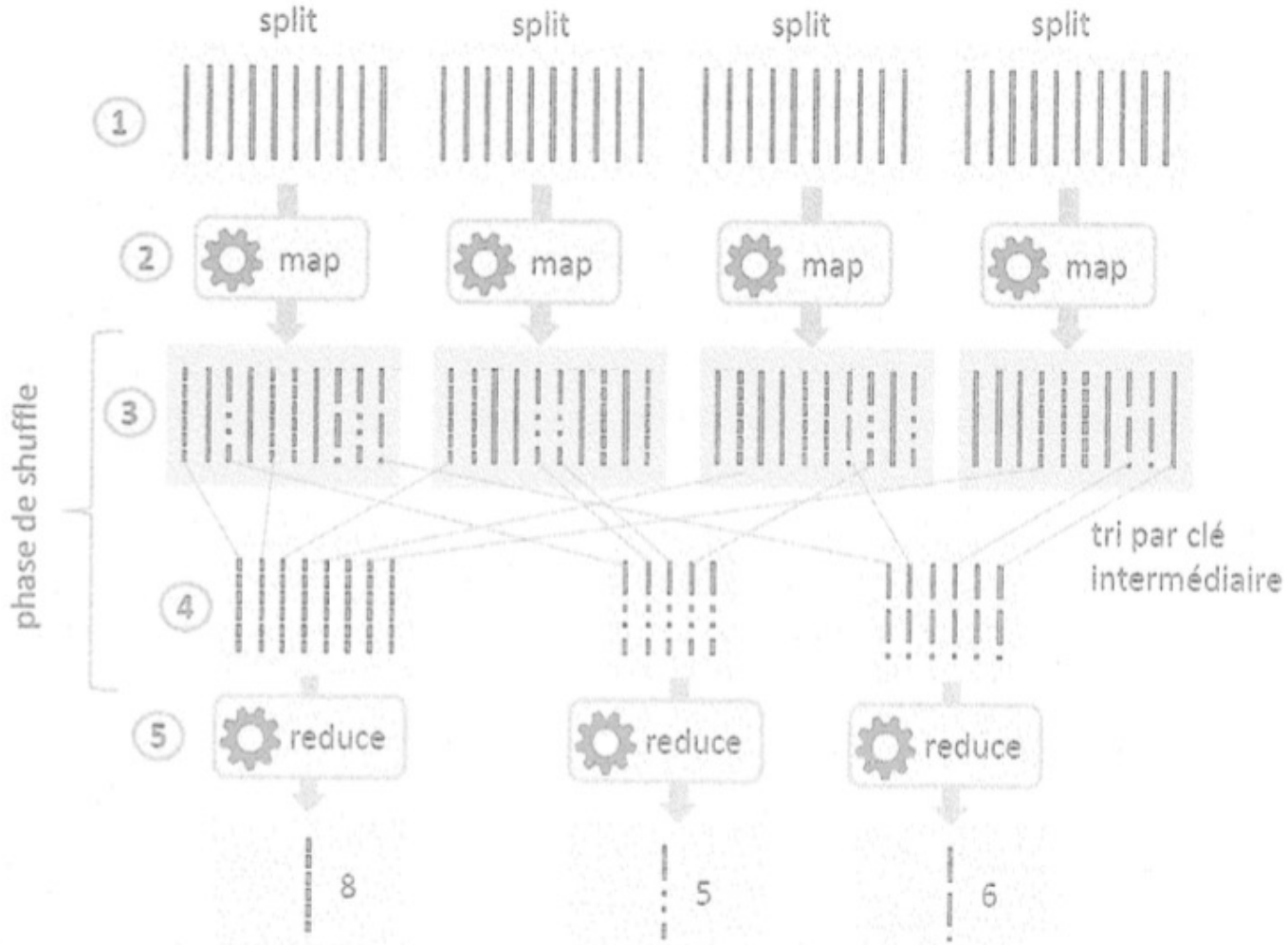
MapReduce et Hadoop

- Objectif : automatiser le déploiement de traitements massivement parallèles sur des clusters de centaines ou de milliers de serveurs
- Le pattern MapReduce :
  - Permet de paralléliser le traitement de grands volumes de données
  - Le calcul doit être formulé avec deux fonctions map et reduce qui opèrent sur des listes de couples clés-valeurs
- Le framework Hadoop :
  - Plateforme d'exécution de MapReduce
  - Fonctionne sur des clusters de machines

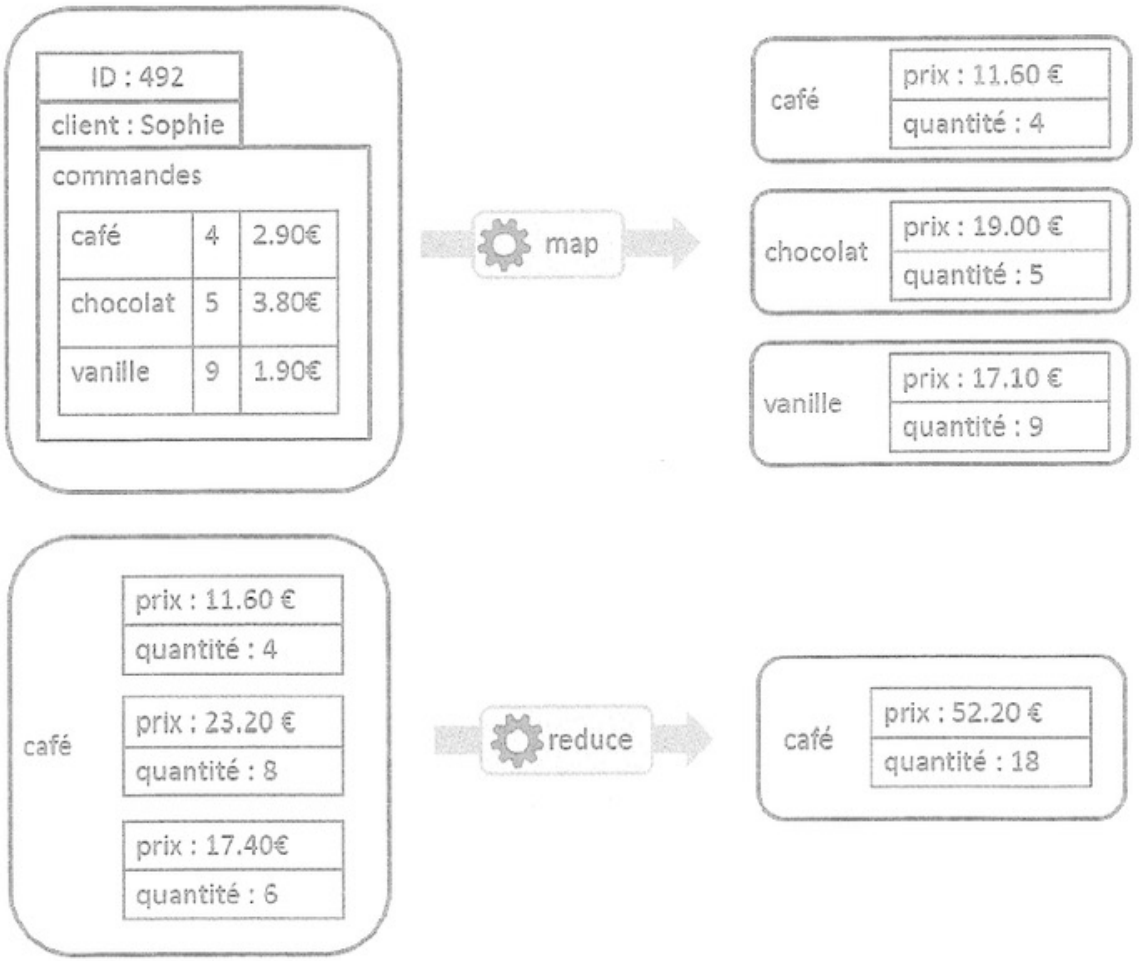
Le pattern MapReduce



Le pattern MapReduce



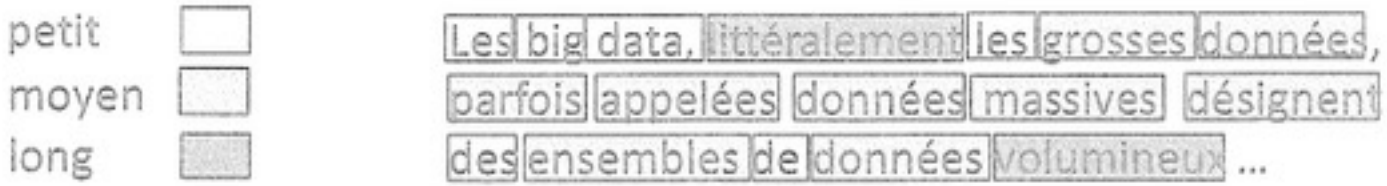
Le pattern MapReduce



Un exemple de fonctions *map* et *reduce* adaptées au calcul d'un prix total et d'une quantité totale par ventes à partir d'un historique de ventes.

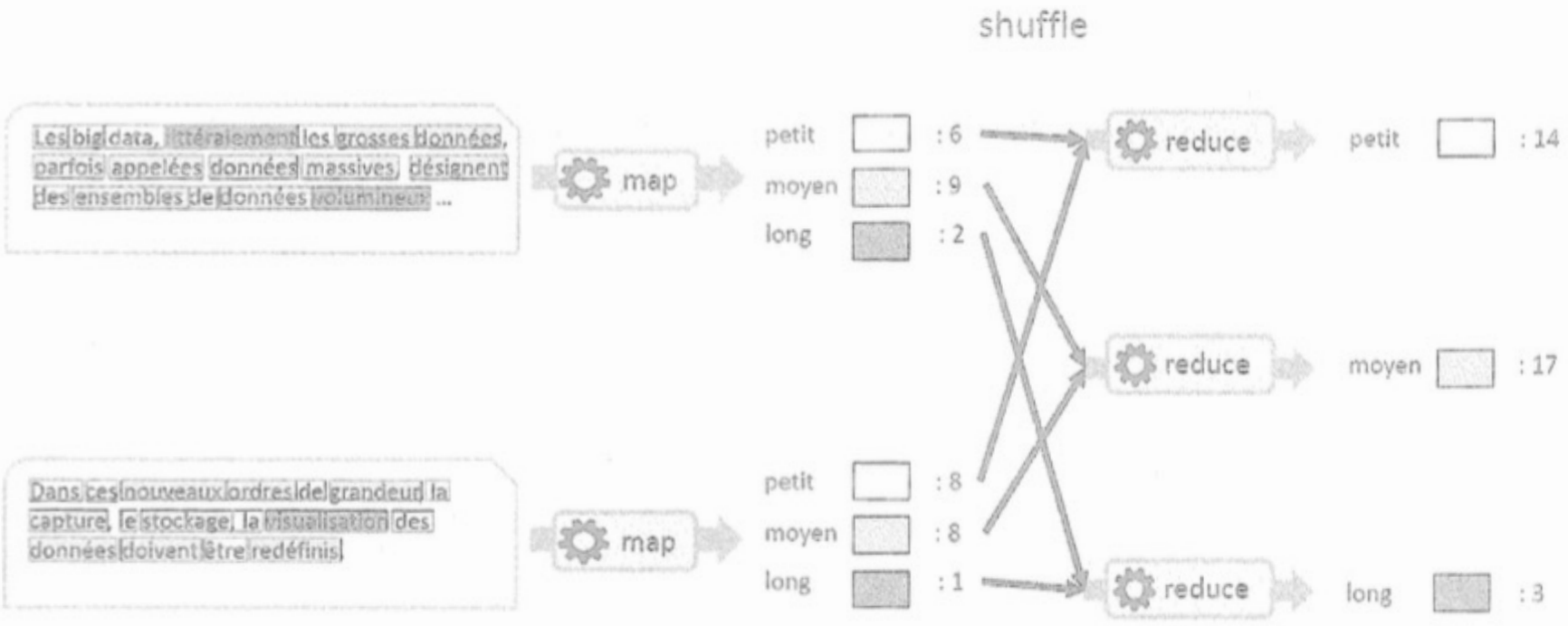
Le pattern MapReduce

- Exemple d'analyse statistique d'un texte :
  - Analyser plusieurs millions de documents et construire l'histogramme de la longueur des mots qu'ils contiennent



Un texte dont les mots ont été triés en trois catégories.
Les mots longs, de plus de 10 lettres, les mots de longueur moyenne de 5 à 9 lettres,
les petits mots de moins de cinq lettres.

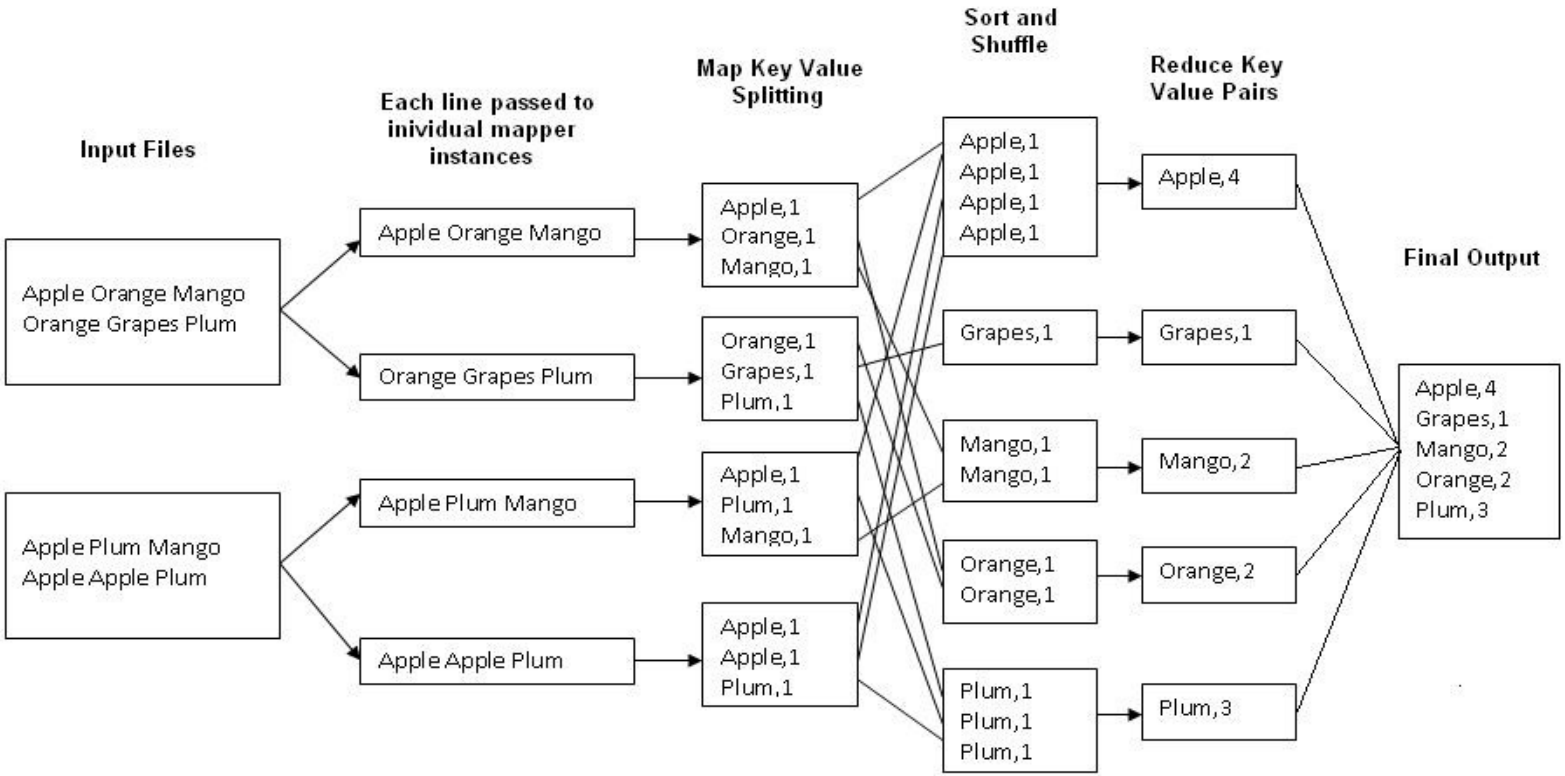
Le pattern MapReduce



La succession des tâches *map* et *reduce* utilisées pour le calcul de l'histogramme de fréquence de mots d'une certaine longueur.

ÉCOLE CENTRALELYON

Le pattern MapReduce



Le framework Hadoop

- De quoi s'agit-il ?
  - Hadoop est un framework MapReduce open source écrit en Java et développé sous l'égide de la fondation Apache depuis 2005
- Structure :
  - Planning des exécutions
  - Tolérance aux pannes
  - Découpage des données en lots
  - Fusion et tri des listes intermédiaires
  - Monitoring des processus

ÉCOLE CENTRALELYON

Exploration et préparation des données

ÉCOLE CENTRALELYON

Exploration et préparation des données

- Objectifs :
  - Transformer les données brutes
    - éparses
    - hétérogènes
    - de qualité variable
  - en un jeu de données
    - de qualité homogène
    - exploitable pour une analyse statistique où l'apprentissage d'un modèle de Machine Learning

ÉCOLE CENTRALELYON

Diversité des sources

- Les SI transactionnels d'une entreprise
- Les entrepôts de données comportementales
- Les données géographiques
- Les données de type Open Data
- Les bases de données commerciales
- Les données obtenues par crawling
- …

ÉCOLE CENTRALELYON

Diversité des formats

- Fichiers classiques
- Bases de données relationnelles
- Bases NoSQL
- …

Diversité de la qualité

- L'exhaustivité
  - plus le jeu de données sera complet, meilleure sera la qualité
- La granularité
  - degré de finesse des données (spatiales, temporelles, sociales)
- L'exactitude
  - fiabilité des informations ou valeurs enregistrées
- La fraîcheur
  - fréquence de rafraîchissement et date de mise à jour
- …

ÉCOLE CENTRALELYON

Exploration des données

- La visualisation des données
  - Elle a pour but d'aider à comprendre les données en s'appuyant sur diverses représentations
- Les statistiques descriptives
  - Utiles en complément de la visualisation, pour résumer certains aspects d'un ensemble de données
  - Moyenne, écart-type, médiane, quantiles
  - Histogrammes
- Les tableaux croisés
- …

Préparation des données

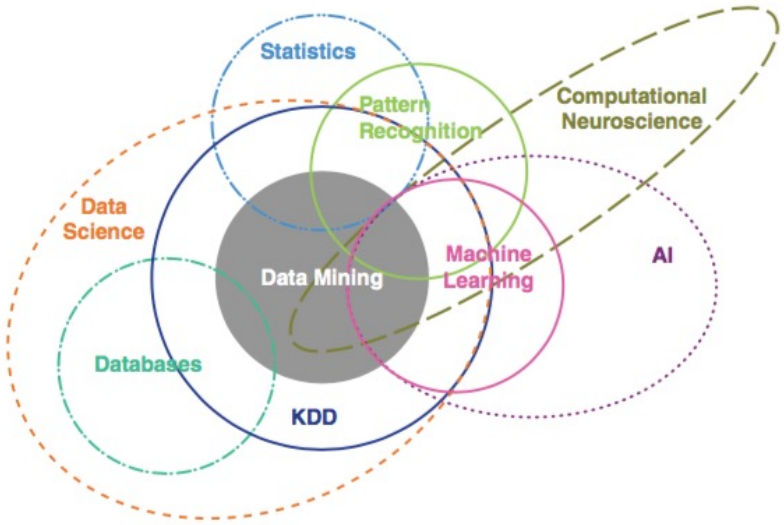
- Objectif : constituer un jeu de données de qualité homogène
- Nettoyer les données
  - éliminer toutes les informations que l'on ne souhaite pas conserver (données erronées, inexactes, …)
- Transformer les données
  - manipuler, modifier, créer de nouvelles informations à partir des informations disponibles
- Enrichir les données
  - utilisation de nouvelles informations que l'on croise avec les données existantes pour identifier de nouvelles corrélations significatives

ÉCOLE CENTRALELYON

LE MACHINE LEARNING

Le Machine Learning

- De quoi s'agit-il ?
  - Un ensemble d'outils statistiques ou géométriques et d'algorithmes permettant d'automatiser la construction d'une fonction de prédiction à partir d'un ensemble d'apprentissage

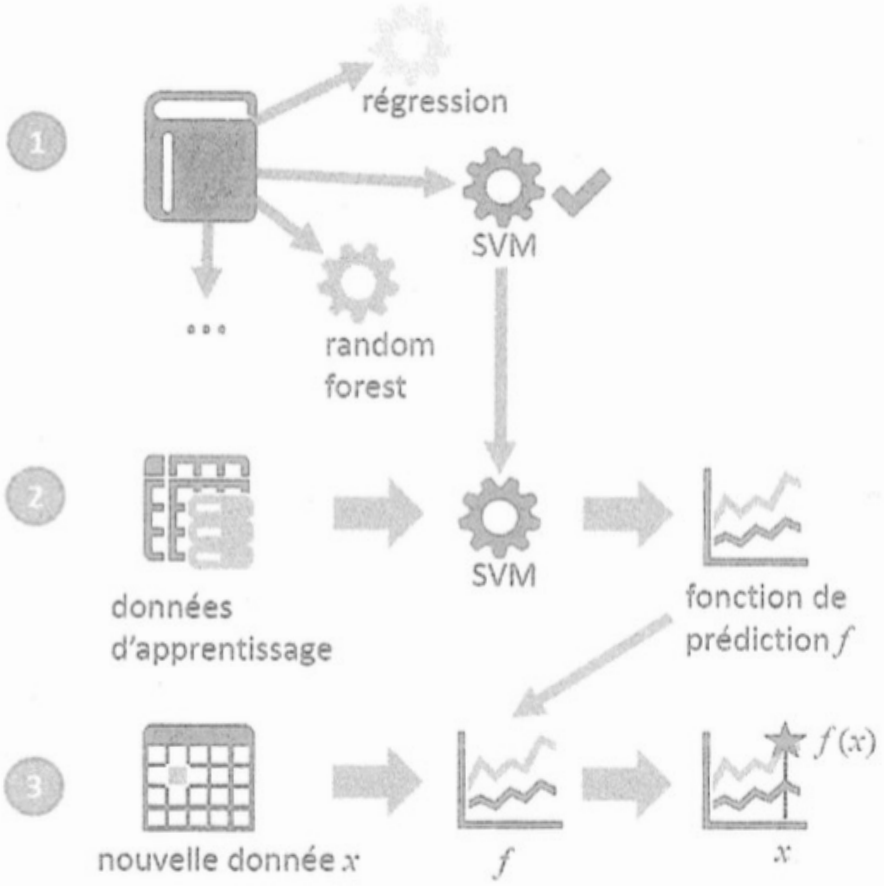


Source: SAS Enterprise Miner Training material from 1998.

Le Machine Learning

- Quelques exemples d'utilisation :
  - Détecter des comportements frauduleux lors de transactions financières en ligne
  - Estimer un taux de transformation sur un site marchand en fonction du nombre de clics sur certaines pages
  - Prédire les risques de non-solvabilité d'un client en fonction de ses ressources et de son profil socioprofessionnel
  - Anticiper les intentions de résiliation d'un service en fonction des activités d'un souscripteur
  - Découvrir les préférences d'un client que l'on souhaite retenir pour lui suggérer des produits et des services adaptés à ses goûts et ses besoins

Choix d'un algorithme de Machine Learning



Sélection d'un algorithme de machine learning dans une librairie de ML

L'entrainement de l'algorithme retenu à partir des données produit une fonction de prédiction f

Prédiction pour une nouvelle observation

(1) Choix d'un modèle de ML, (2) apprentissage (ou entraînement) du modèle et enfin (3) prédiction pour une nouvelle observation.

ÉCOLE CENTRALELYON

Propriétés d'un bon algorithme de ML

- Déployabilité : possibilité de passage à l'échelle sur un framework comme Hadoop
- Robustesse : ne pas être sensible à des données incohérentes et incomplètes
- Transparence : possibilité de prévenir les dégradations des performances pendant la phase d'apprentissage
- Adéquation aux compétences disponibles : ne nécessite pas d'expertise trop poussée
- Proportionnalité : adéquation entre le temps investi dans l'amélioration d'un algorithme avec le gain apporté

ÉCOLE CENTRALELYON

Evaluation de la performance d'un modèle

- Matrice de confusion

		Classe prédite	
		Classe	Non classe
Classe réelle	Classe	VP	FN
	Non classe	FP	VN

- VP / VN : vrais positifs / vrais négatifs
- FP / FN : faux positifs / faux négatifs

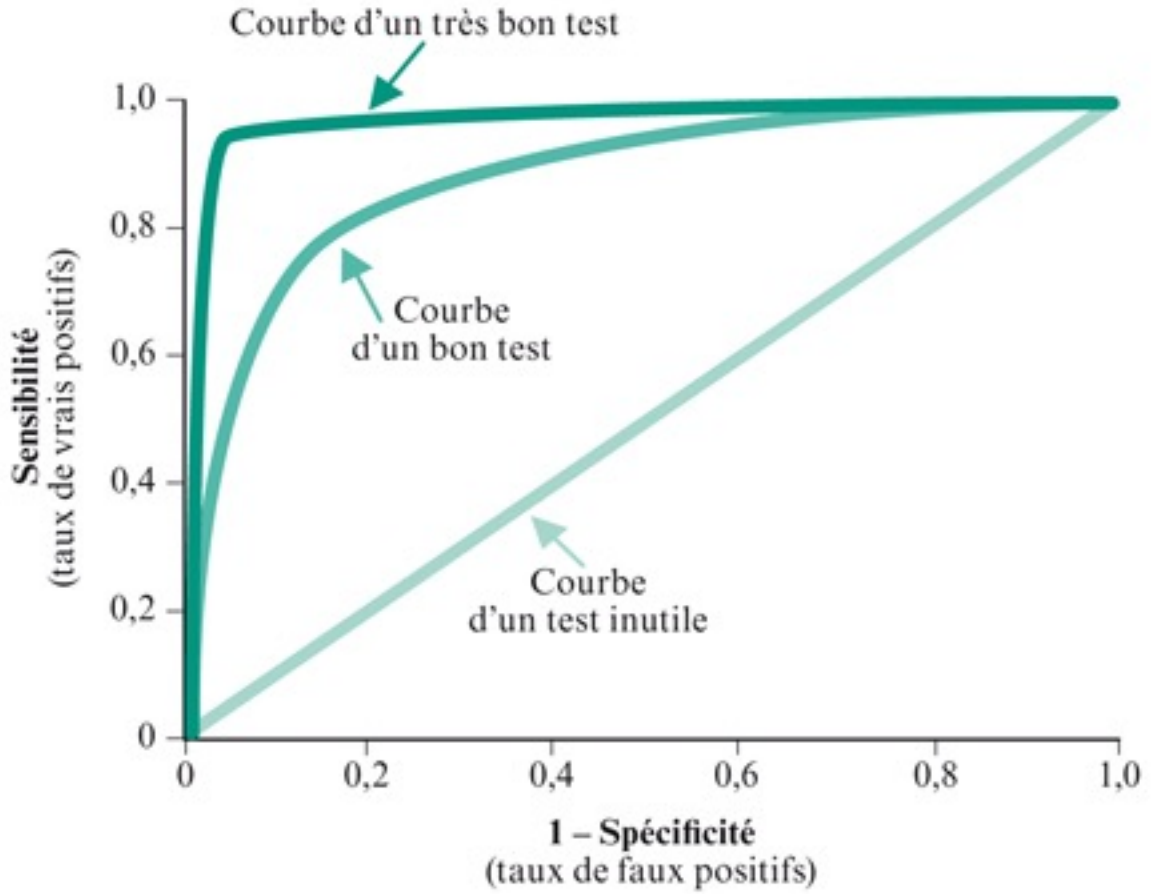
Evaluation de la performance d'un modèle

- Taux de classification : proportion d'éléments bien classés
  -> (VP + VN) / (VP + VN + FP + FN)
- Rappel (sensibilité) : proportion d'éléments bien classés par rapport au nombre d'éléments de la classe à prédire
  -> VP / (VP + FN)
- Précision : proportion d'éléments bien classés pour une classe donnée
  -> VP / (VP + FP)
- F-mesure : mesure de compromis entre précision et rappel
  -> 2*(Rappel*Précision) / (Rappel + Précision)

		Classe prédite	
		Classe	Non classe
Classe réelle	Classe	VP	FN
	Non classe	FP	VN

Evaluation de la performance d'un modèle

- Courbe ROC



Evaluation de la performance d'un modèle

- Pour un modèle de régression :
  - Erreur quadratique moyenne

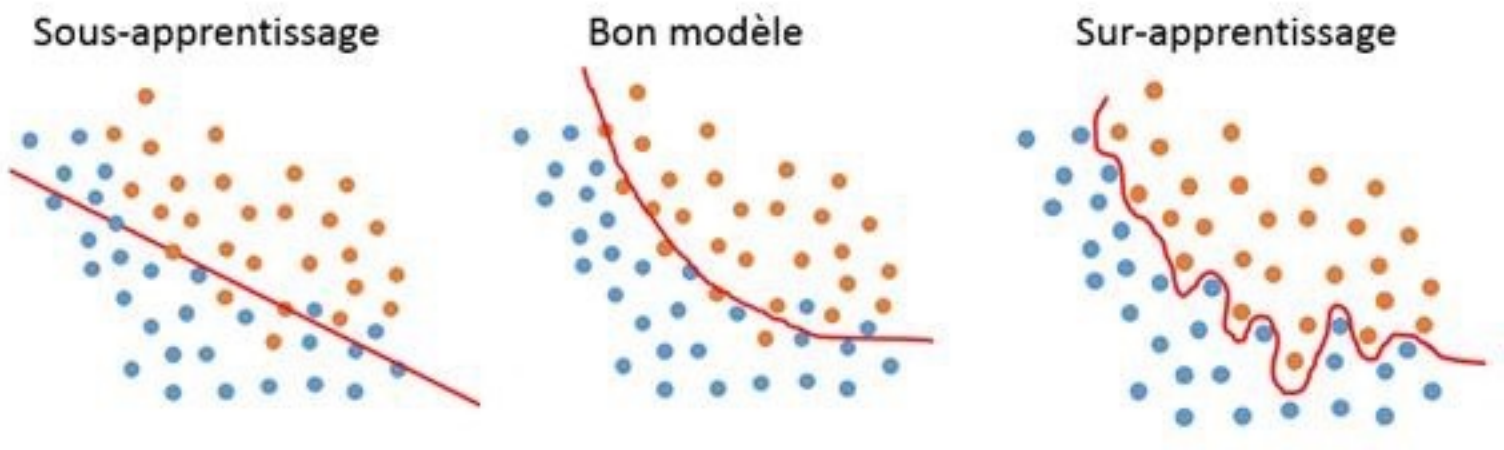
    -> Moyenne des différences au carré entre les vraies valeurs et les valeurs prédites

  - Coefficient de corrélation de Pearson

    -> Mesure la relation linéaire ("proportionnalité") entre les valeurs réelles et les valeurs prédites

Problème du surapprentissage

- Situation dans laquelle un modèle de ML reproduit avec une grande précision les données d'apprentissage tout en étant incapable de faire des extrapolations précises



ÉCOLE CENTRALELYON

Machine Learning et Big Data

- Parallélisation des algorithmes :
  - De nombreux algorithmes ont été conçus à une époque où les données comportaient au plus quelques milliers d'observations
  - Il faut donc que ces algorithmes soient parallélisables pour être utiles dans un contexte Big Data

ÉCOLE CENTRALELYON

Les différents types de Machine Learning

- Apprentissage supervisé vs non supervisé
  - Supervisé :
    - Suppose que l'on dispose d'un ensemble d'exemples caractérisés par des variables prédictives, et dont on connaît les valeurs pour la variable cible
    - L'objectif de l'apprentissage est de construire une fonction de prédiction permettant l'association entre les variables prédictives et la variable cible
  - Non supervisé :
    - On ne dispose pas de valeurs pour la variable cible
    - L'objectif est de réaliser le regroupement en catégories des exemples fournis en exemple

ÉCOLE CENTRALELYON

Les différents types de Machine Learning

- Régression vs classification
  - Régression :
    - La variable cible est quantitative
  - Classification :
    - La variable cible est qualitative

ÉCOLE CENTRALELYON

Les différents types de Machine Learning

- Algorithmes linéaires vs non linéaires
  - Linéaires :
    - La fonction de prédiction *f* est une combinaison linéaire des variables prédictives : $f(x) = a_1x_1+a_2x_2+\ldots+a_nX_n$
  - Non linéaires :
    - La fonction *f* est non linéaire

ÉCOLE CENTRALELYON

Les différents types de Machine Learning

- Modèle paramétrique vs non paramétrique
  - Modèle paramétrique :
    - La fonction de prédiction f a une forme particulière, avec un nombre spécifié par avance de paramètres ajustables estimés à partir des données d'apprentissage
  - Modèle non paramétrique :
    - Aucune forme particulière n'est postulée par avance pour la fonction de prédiction

ÉCOLE CENTRALELYON

Les différents types de Machine Learning

- Modèle géométrique vs probabiliste
  - Modèle géométrique :
    - Repose sur la notion de proximité d'une observation avec d'autres exemplaires (métrique)
  - Modèle probabiliste :
    - Utiles lorsque les valeurs des variables prédictives et des variables cibles obéissent à certaines lois de probabilité

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- Objectif : Réduire le nombre de variables prédictives d'un ensemble d'observations
  - Le temps d'exécution d'un algorithme d'apprentissage dépend simultanément de la taille N de l'échantillon d'apprentissage et du nombre p de variables prédictives
  - Un modèle prédictif avec peu de variables est plus facile à interpréter et permet généralement d'obtenir de meilleurs résultats (apprentissage facilité)
- Quelques exemples :
  - L'analyse en composantes principales
  - Les approches de type Forward Selection
  - Les algorithmes génétiques

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- L'analyse en composantes principales
  - Consiste à chercher un nombre restreint de combinaisons linéaires des variables prédictives d'origine
  - Ces nouvelles variables sont calculées de manière à être non corrélées et à expliquer l'essentiel de la variabilité des variables d'origine

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- Les algorithmes génétiques
  - Principe : maintenir pendant plusieurs générations une population de taille constante d'individus caractérisés par leur chromosome
  - Chromosome : chaîne de gènes binaires
  - Gène : associé à un descripteur de l'espace d'origine
    - Valeur = 1 : inclusion de ce descripteur dans le nouvel espace
    - Valeur = 0 : exclusion
  - Evolution gouvernée par trois règles
    - Sélection
    - Croisement
    - Mutation

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- Les algorithmes génétiques
  - Sélection
    - Permet l'élimination des individus les moins performants par rapport à un objectif donné
  - Croisement
    - Permet la création de nouveaux individus par combinaison des chromosomes des deux parents
  - Mutation
    - Permet d'éviter une stagnation qui se produirait au cours du processus de recherche, et d'augmenter le domaine d'exploration
    - Principe : modifier de manière aléatoire les gènes d'un individu sélectionné

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- Les algorithmes génétiques
  - La performance d'un individu est donnée par une fonction d'évaluation
  - Dans le cas d'un problème de classification, la fonction d'évaluation est le taux de reconnaissance du classifieur associé à l'espace représenté par le chromosome

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- Les approches de type Forward Selection
  - Particulièrement adaptées pour trouver les quelques descripteurs les plus importants
  - Le critère à maximiser, comme pour l'algorithme génétique est dans ce cas le taux de reconnaissance du classifieur

ÉCOLE CENTRALELYON

Les techniques de réduction dimensionnelle

- Les approches de type Forward Selection
  - Calculer un critère J pour chaque descripteur et sélectionner le meilleur
  - Considérer toutes les combinaisons possibles à deux dimensions de ce descripteur avec les autres descripteurs et sélectionner le meilleur
  - Ajouter ainsi un descripteur à chaque étape en prenant toujours celui qui correspond à la valeur la plus élevée du critère J
  - Arrêter le processus lorsque le nombre de descripteurs désiré est atteint

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La régression linéaire
- Les k plus proches voisins
- La classification naïve bayésienne
- La régression logistique
- Les arbres de décision
- Les forêts aléatoires
- Les réseaux de neurones

- Et pour l'apprentissage profond :
  - Réseaux convolutifs
  - Transformers

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La régression linéaire
  - Modèle supervisé paramétrique
  - Suppose que la fonction de prédiction *f* a la forme :
  - $f(x) = a_1x_1+a_2x_2+...+a_nx_n +b$
  - L'apprentissage consiste à calculer les coefficients a_i et b qui minimisent les erreurs de prédiction sur les données d'apprentissage

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La régression linéaire
  - Avantages :
    - L'apprentissage se résume à l'inversion d'une matrice construite à partir des données d'apprentissage
    - Aucun calcul numérique complexe
    - Le modèle est généralement simple à interpréter
  - Inconvénients :
    - Pour être efficace, il faut que la relation soit effectivement linéaire
    - Est sensible aux valeurs aberrantes
    - Du fait de la linéarité, néglige les interactions entre les variables prédictives

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les k plus proches voisins
  - Modèle supervisé non paramétrique
  - Considère chaque donnée de l'ensemble d'entrainement comme un point dans l'espace des valeurs prédictives
  - Pour une donnée à classer, on détermine dans cet espace les k exemplaires les plus proches et on affecte à cette donnée la classe majoritaire parmi ses k voisins

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les k plus proches voisins
  - Avantages :
    - Aucune hypothèse sur la distribution des données
    - Modèle très simple
  - Inconvénients :
    - Modèle très sensible au bruit
    - Lorsque le nombre de variables est grand, le calcul de la distance peut être couteux. Il faut alors utiliser des techniques de réduction dimensionnelle

Les principaux algorithmes de Machine Learning

- La classification naïve bayésienne
  - Modèle supervisé souvent très performant malgré sa simplicité
  - Repose sur l'indépendance des probabilités conditionnelles des variables prédictives sachant qu'une catégorie est fixée
  - Considère le vecteur x des valeurs des variables prédictives comme une variable aléatoire dont la distribution dépend de la classe
  - Formule de Bayes :

$$p(\omega_i / x) = \frac{p(x / \omega_i) p(\omega_i)}{p(x)}$$

Les principaux algorithmes de Machine Learning

- La classification naïve bayésienne
  - On ne connaît pas $p(\omega_i / x)$ mais on peut estimer $p(x / \omega_i)$ ainsi que $p(\omega_i)$ à partir de l'ensemble d'apprentissage
  - De plus, p(x) étant le même pour toutes les classes, on peut réaliser la classification à partir du classifieur suivant :

$$x \in \omega_i \Leftrightarrow p(\omega_i / x) > p(\omega_j / x) \ \forall j$$

  - soit :

$$x \in \omega_i \Leftrightarrow p(x / \omega_i) p(\omega_i) > p(x / \omega_j) p(\omega_j) \ \forall j$$

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La classification naïve bayésienne
  - But de l'apprentissage pour le classifieur bayésien
    - Estimer la probabilité a priori des classes
    - Estimer la densité de probabilités des classes
  - Estimation de la probabilité a priori des classes
    - Rapport du nombre d'éléments de l'ensemble d'apprentissage appartenant à la classe considérée sur le nombre total d'éléments dans l'ensemble d'apprentissage

Les principaux algorithmes de Machine Learning

- La classification naïve bayésienne
  - Estimation de la densité de probabilités des classes dans le cas gaussien

$$p(x/\omega_i) = N(x;\mu_i,\Sigma_i) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_i)}} \exp\left(-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)\right)$$

  - Méthode du maximum de vraisemblance
    - Calculer le vecteur moyenne μ_i et la matrice de covariance Σ_i de chaque classe à partir de l'ensemble d'apprentissage

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La classification naïve bayésienne
  - Avantages :
    - Algorithme simple et efficace
    - Utile même lorsque l'indépendance des variables prédictives conditionnées sur les classes est difficile à justifier
  - Inconvénients :
    - Les prédictions de probabilités pour les différentes classes sont erronées lorsque l'hypothèse d'indépendance conditionnelle est invalide

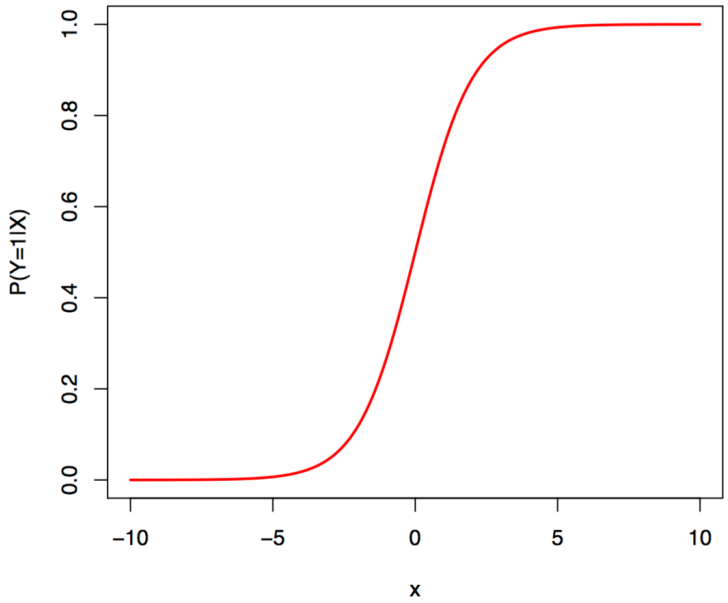
ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La régression logistique
  - Modèle de classification linéaire
  - Equivalent de la régression linéaire lorsque la variable cible ne peut prendre que deux valeurs (0 ou 1)
  - Utilisation d'une fonction de score S des variables prédictives :
  - $S(x) = a_1x_1+...+a_nx_n$
  - On recherche alors les coefficients a_i tels que le score S(x) soit positif lorsque les chances d'appartenir au groupe 1 sont grandes, et tel que S(x) soit négatif lorsque les chances d'appartenir au groupe 0 sont grandes
  - Entre les deux, on utilise une fonction d'interpolation pour calculer la probabilité $P_1(x)$ d'appartenance au groupe 1

Les principaux algorithmes de Machine Learning

- La régression logistique
  - Expression de la probabilité conditionnelle :
  - P(y=1/x) = logit(S(x))
  - avec logit(S) = 1 / (1+exp(-S))



ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- La régression logistique
  - Avantages :
    - La classification d'une nouvelle observation est très rapide puisqu'elle se résume à l'évaluation d'une fonction de score linéaire S
    - Algorithme simple peu sensible au surapprentissage
    - Interprétation aisée des coefficients a_i
  - Inconvénients :
    - L'hypothèse de linéarité du score ne permet pas de tenir compte des interactions entre variables
    - La phase d'apprentissage peut être longue car l'opération d'optimisation des coefficients a_i peut être complexe
    - L'algorithme est limité aux variables cibles binaires

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les arbres de décision :
  - Modèles supervisés non paramétriques, très flexibles
  - Ne reposent sur aucun modèle probabiliste
  - Consistent à classer une observation au moyen d'une succession de tests concernant les valeurs des variables prédictives
  - Chaque test est représenté par un nœud de l'arbre
  - Chaque branche correspond à une réponse possible à la question posée
  - La classe de la variable est déterminée par la feuille à laquelle parvient l'observation à l'issue de la suite de tests

ÉCOLE CENTRALELYON

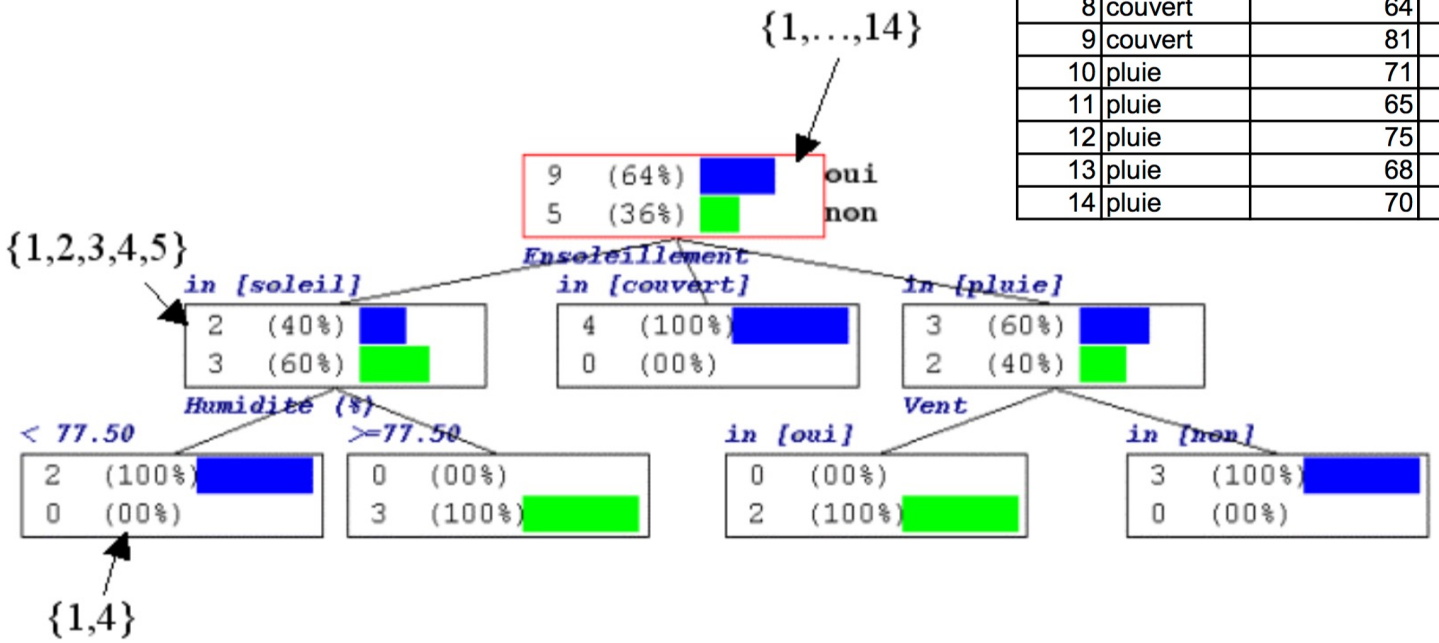
Les principaux algorithmes de Machine Learning

- Les arbres de décision :
  - La phase d'apprentissage consiste donc à trouver les bons tests pour classer correctement les observations par rapport à leur valeur pour la variable cible
  - Il existe différentes stratégies d'apprentissage pour déterminer les nœuds de l'arbre
  - L'objectif reste le même : les feuilles doivent être homogènes en ne contenant que les observations appartenant à une seule et même classe

Les principaux algorithmes de Machine Learning

- Les arbres de décision :

Numéro	Ensoleillement	Température (°F)	Humidité (%)	Vent	Jouer
1	soleil	75	70	oui	oui
2	soleil	80	90	oui	non
3	soleil	85	85	non	non
4	soleil	72	95	non	non
5	soleil	69	70	non	oui
6	couvert	72	90	oui	oui
7	couvert	83	78	non	oui
8	couvert	64	65	oui	oui
9	couvert	81	75	non	oui
10	pluie	71	80	oui	non
11	pluie	65	70	oui	non
12	pluie	75	80	non	oui
13	pluie	68	80	non	oui
14	pluie	70	96	non	oui



ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les arbres de décision :
  - Avantages :
    - Les variables prédictives peuvent être qualitatives ou quantitatives
    - La phase de préparation des données est réduite (ni normalisation, ni traitement des données manquantes)
    - On tient compte des interactions entre variables car pas d'hypothèse de linéarité
    - Permet de traiter des problèmes de classification à plusieurs classes
    - Il est souvent possible d'interpréter le processus de classification comme l'application d'un ensemble de règles intelligibles

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les arbres de décision :
  - Inconvénients :
    - Exiger que toutes les observations soient parfaitement classées peut mener au surapprentissage
    - Le critère affecté au premier nœud a une très grande influence sur le modèle de prédiction

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les forêts aléatoires :
  - Le but est de conserver les avantages des arbres de décision tout en éliminant leurs inconvénients
  - Algorithme de classification ou de régression non paramétrique, très flexible et très robuste

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les forêts aléatoires :
  - Repose sur trois idées principales :
    - A partir d'un échantillon initial de N observations dont chacune est décrite par p variables prédictives, on crée artificiellement B nouveaux échantillons de même taille M par tirage avec remise (bootstrap). On entraine alors B arbres de décision différents
    - Parmi les p variables prédictives, on n'en utilise qu'un nombre m<p choisies au hasard. Elles sont alors utilisées pour faire la meilleure segmentation possible
    - L'algorithme combine plusieurs algorithmes faibles (les B arbres de décision) pour en constituer un plus puissant en procédant par vote : pour classer une nouvelle observation, on la fait passer par les B arbres et on sélectionne la classe majoritaires parmi les B prédictions

ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

- Les forêts aléatoires :
  - Avantages :
    - Un des meilleurs algorithmes à l'heure actuelle
    - N'est pas sujet au surapprentissage
    - Utile pour se faire une idée du pouvoir prédictif des données
    - Peut traiter des données avec des milliers de variables prédictives
    - Conserve une bonne puissance prédictive même lorsqu'une grande partie des données est manquante
  - Inconvénients :
    - Algorithme complexe dont l'implémentation peut être délicate (utiliser de préférence une librairie existante)
    - On ne conserve pas le caractère intelligible des arbres de décision

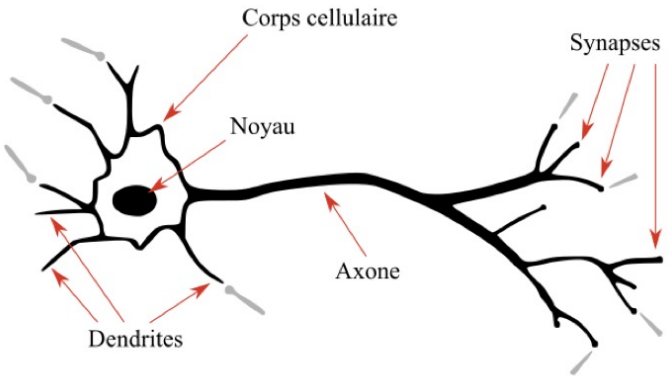
ÉCOLE CENTRALELYON

Les principaux algorithmes de Machine Learning

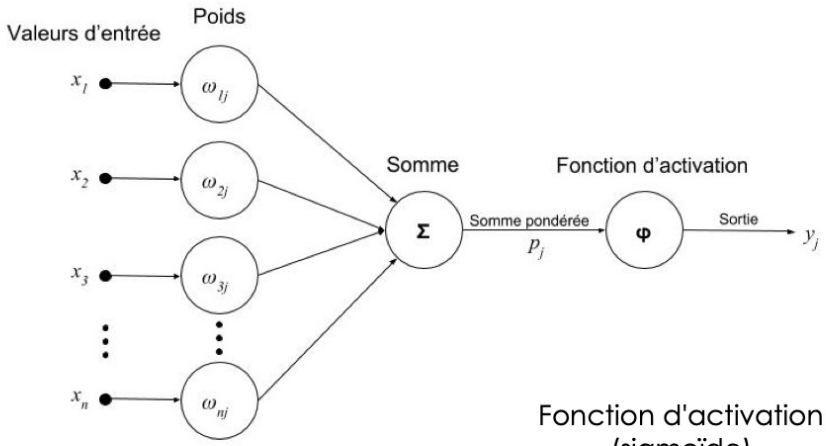
- Les réseaux de neurones
  - Les réseaux de neurones artificiels, et particulièrement les réseaux multicouches de type perceptron sont une technique de classification très puissante et largement répandue
  - Ils sont inspirés des réseaux de neurones biologiques
  - Ils consistent en un réseau orienté composé de neurones artificiels organisés en couches

Les principaux algorithmes de Machine Learning

Neurone biologique



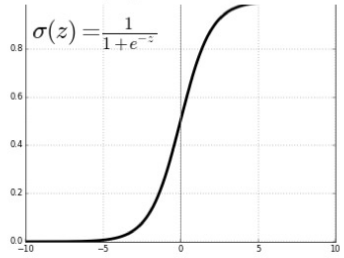
Neurone artificiel



Correspondances

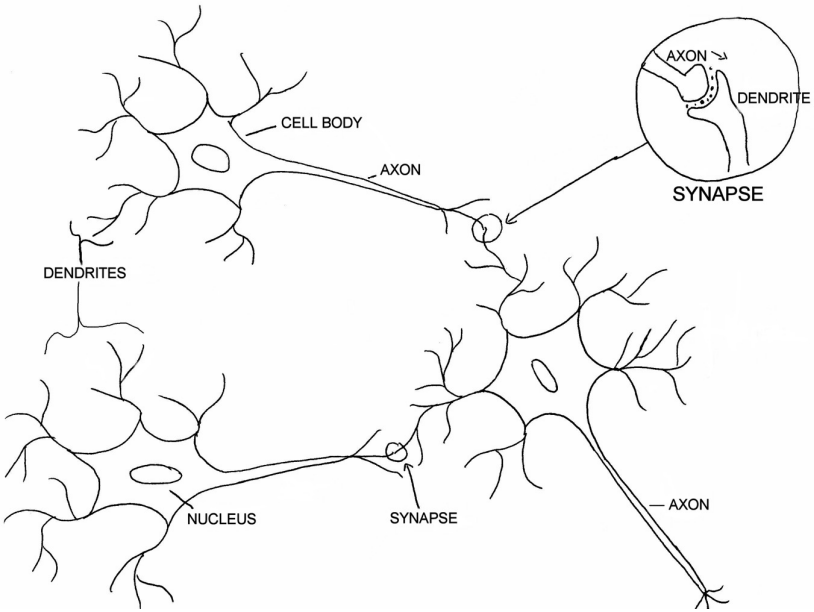
Neurone biologique	Neurone artificiel
Dentrite	Signal d'entrée
Synapse	Poids de la connexion
Axone	Signal de sortie

Fonction d'activation (sigmoïde)

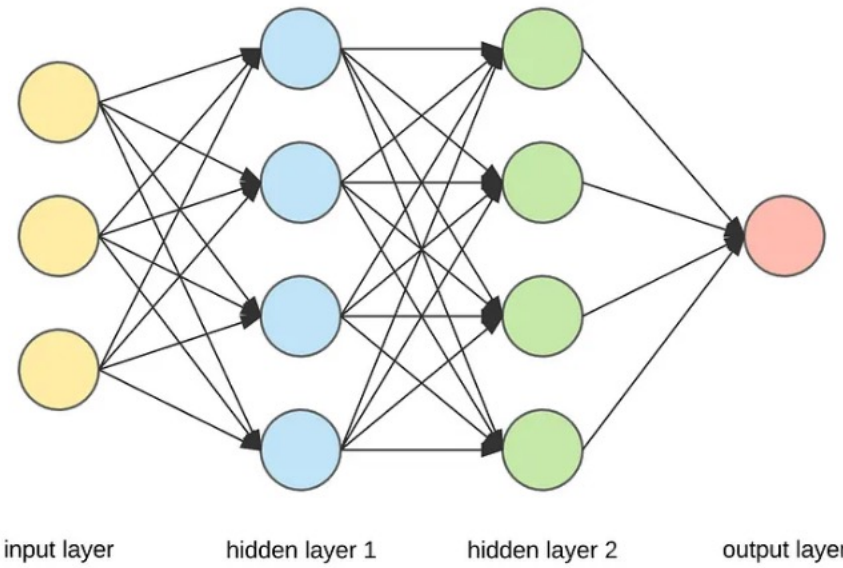


Les principaux algorithmes de Machine Learning

Réseau de neurones biologiques



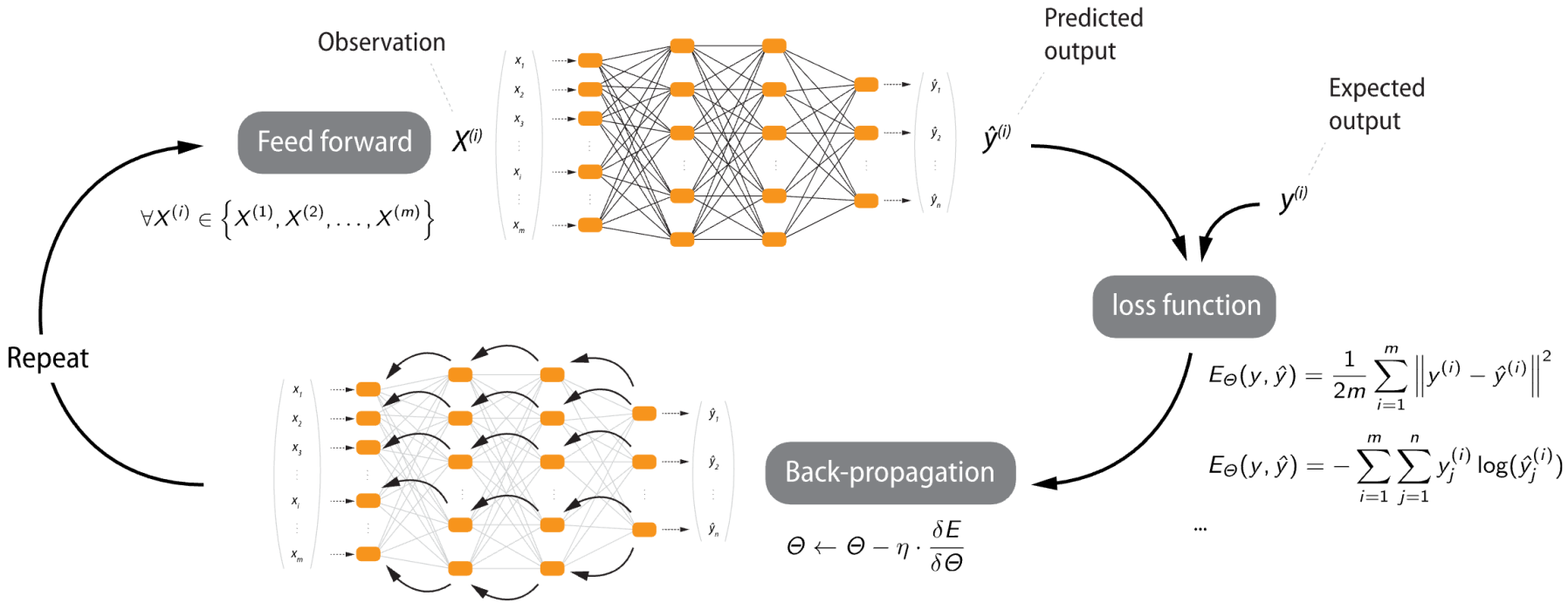
Réseau de neurones artificiels



Source : https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

Les principaux algorithmes de Machine Learning

- Apprentissage des réseaux de neurones



Source : FIDLE, section 1 History and Basic Concepts, https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/wikis/home

ÉCOLE CENTRALELYON

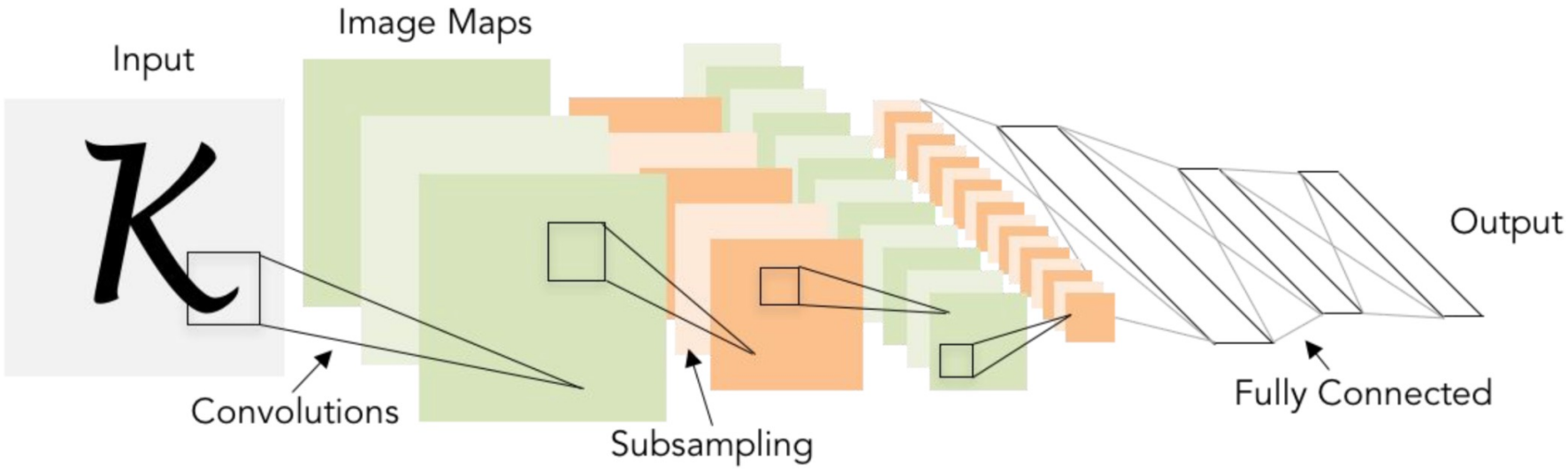
Les principaux algorithmes de Machine Learning

- Les réseaux de neurones
  - Avantages :
    - Permet de traiter des problèmes de classification non linéaires complexes
  - Inconvénients :
    - Choix de la structure délicat
    - Risque de tomber dans un minimum local lors de l'apprentissage

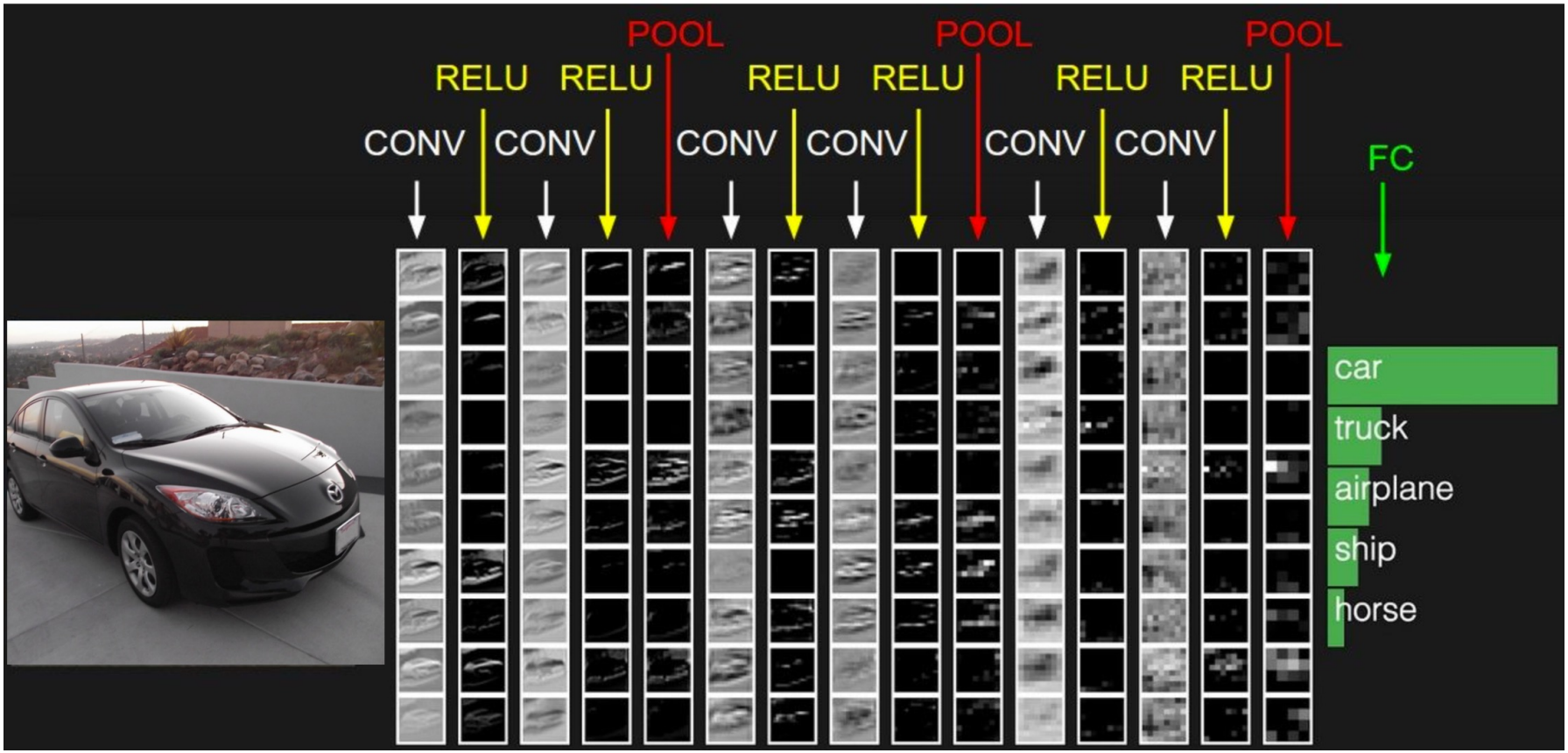
Deep Learning

- Les réseaux de neurones profonds (deep learning)
  - ➔ Réseaux de neurones convolutifs (CNN) : LeNet

[LeCun et al., 1998]



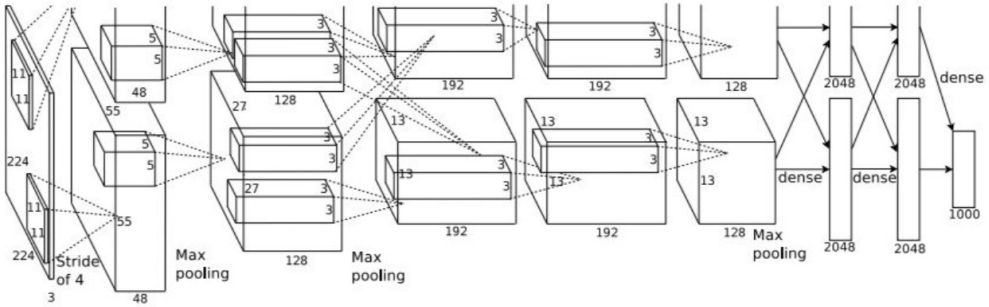
Deep Learning



Deep Learning

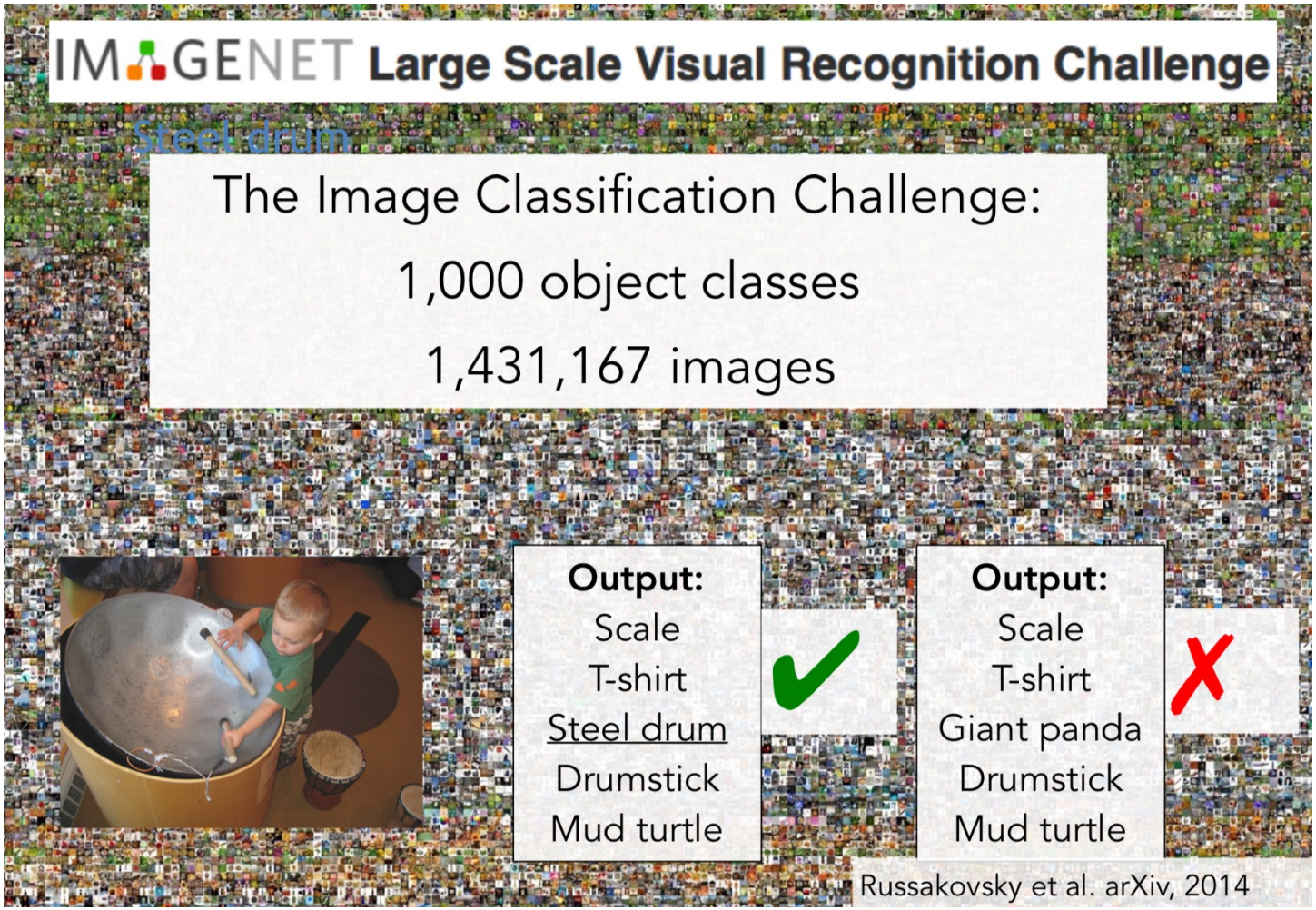
Case Study: AlexNet

[Krizhevsky et al. 2012]



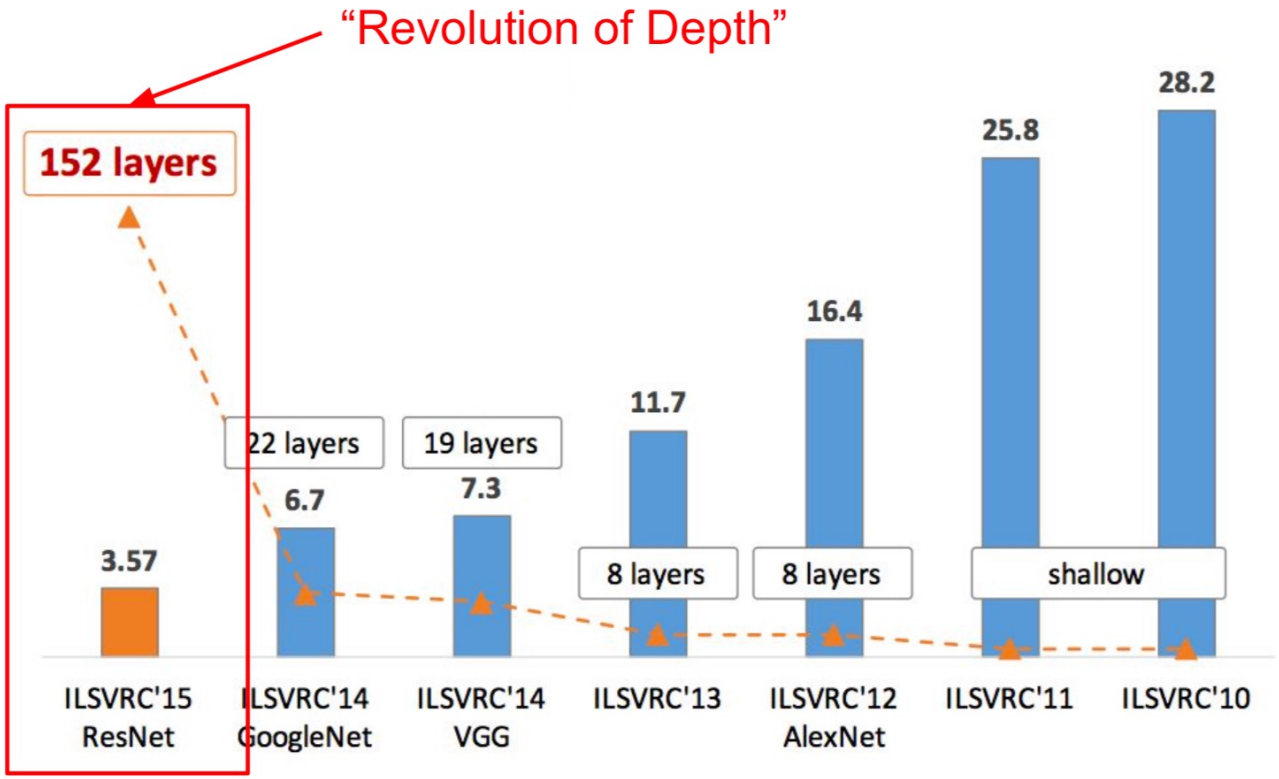
Full (simplified) AlexNet architecture:
[227x227x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

Deep Learning



Deep Learning

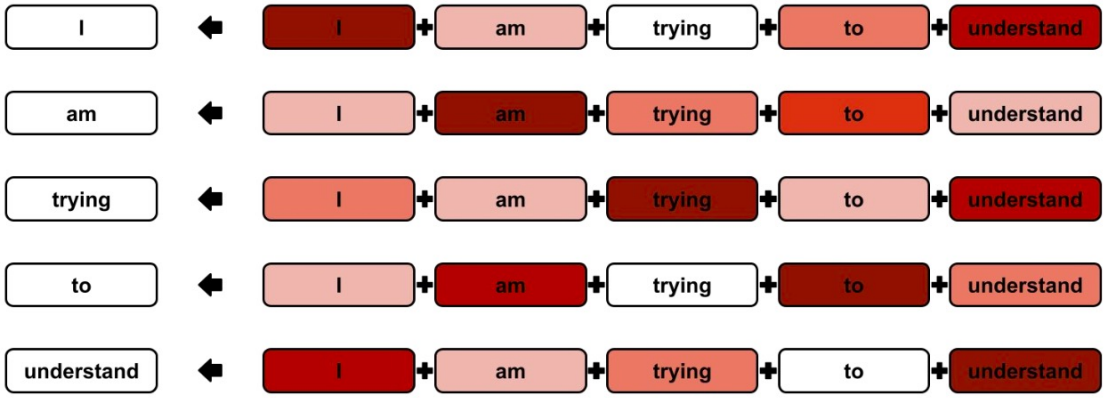
ImageNet Large Scale Visual Recognition Challenge (ILSVRC) winners



Transformers

- Principe fondamental des modèles Transformers* : l'auto-attention
  - Calculer une pondération pour chaque élément de l'entrée, en fonction de sa relation avec tous les autres éléments de l'entrée.
  - Utiliser ces pondérations d'attention pour générer des représentations de plus haut niveau pour la séquence, en combinant les informations de chaque élément de la séquence pondérées par leurs poids d'attention.

➔ Permet de capturer des relations complexes entre les différents éléments de la séquence

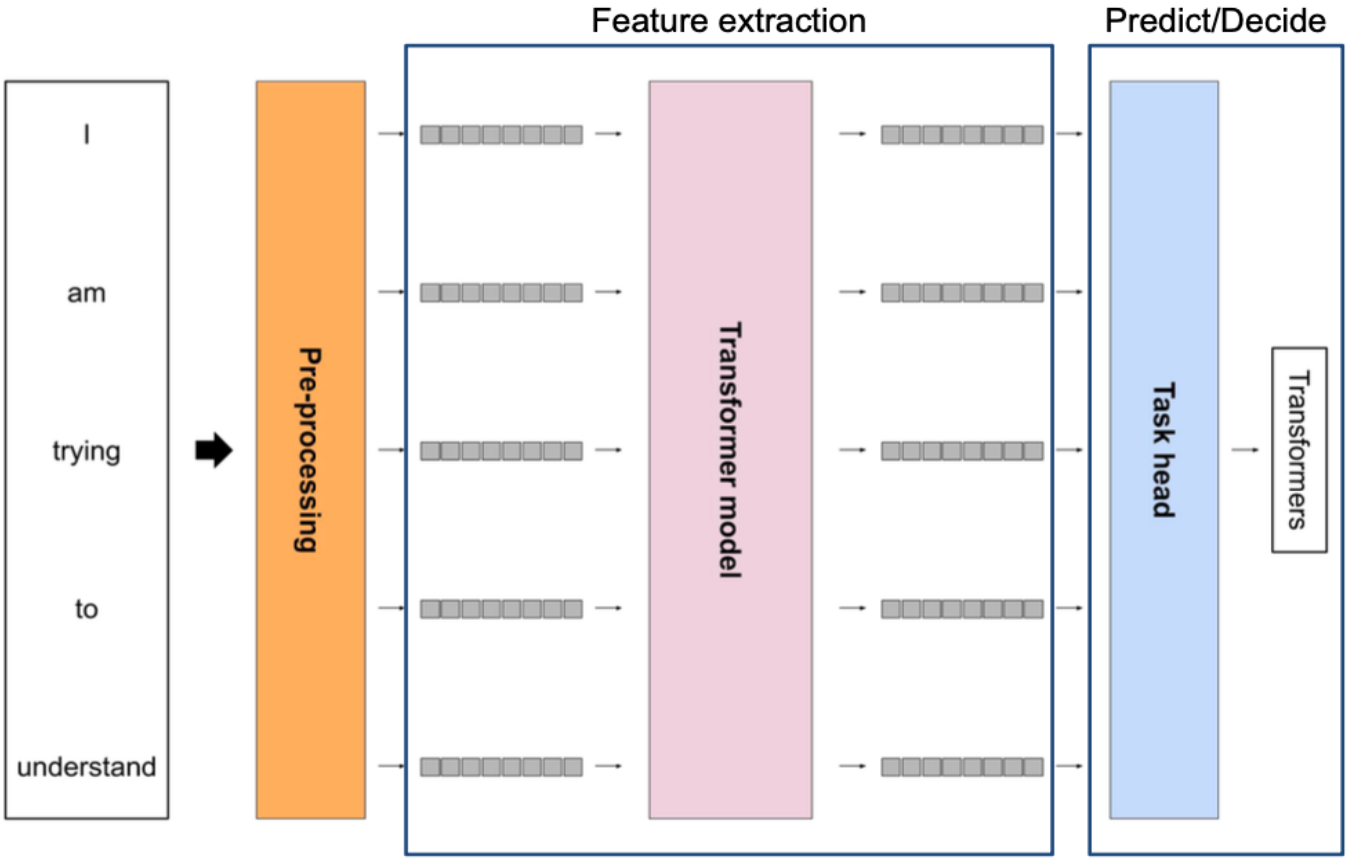


Source : FIDLE, section 8 Transformers, https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/wikis/home

* « Attention Is All You Need », Ashish Vaswani et al., NIPS, 2017.

Transformers

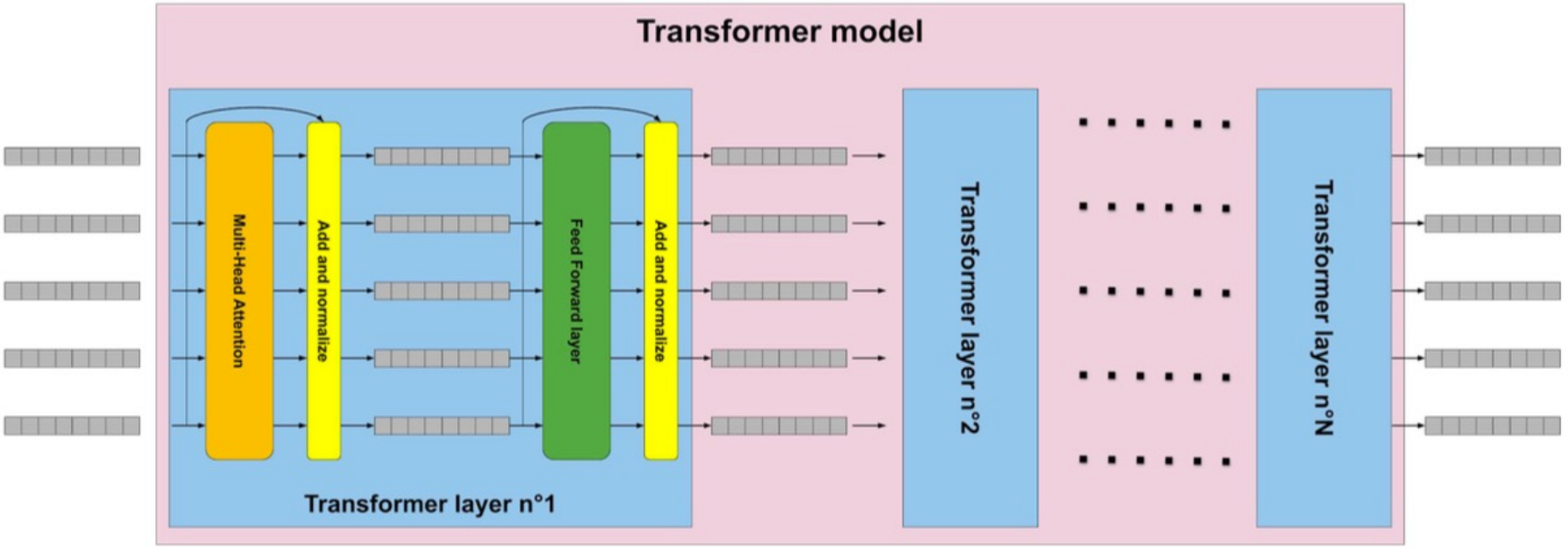
- Architecture globale du modèle



Source : FIDLE, section 8 Transformers, https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/wikis/home

Transformers

- Architecture du Transformers



Source : FIDLE, section 8 Transformers, https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/wikis/home

ChatGPT

- ChatGPT ?
  - Un agent conversationnel basé sur un modèle de langage développé par la société OpenAI
  - Système génératif de texte s'appuyant sur un prompt (texte d'entrée)
  - Son objectif : prédire à chaque étape le mot suivant d'une séquence de mots
  - Comment y avoir accès ? ➔ https://chat.openai.com

ChatGPT

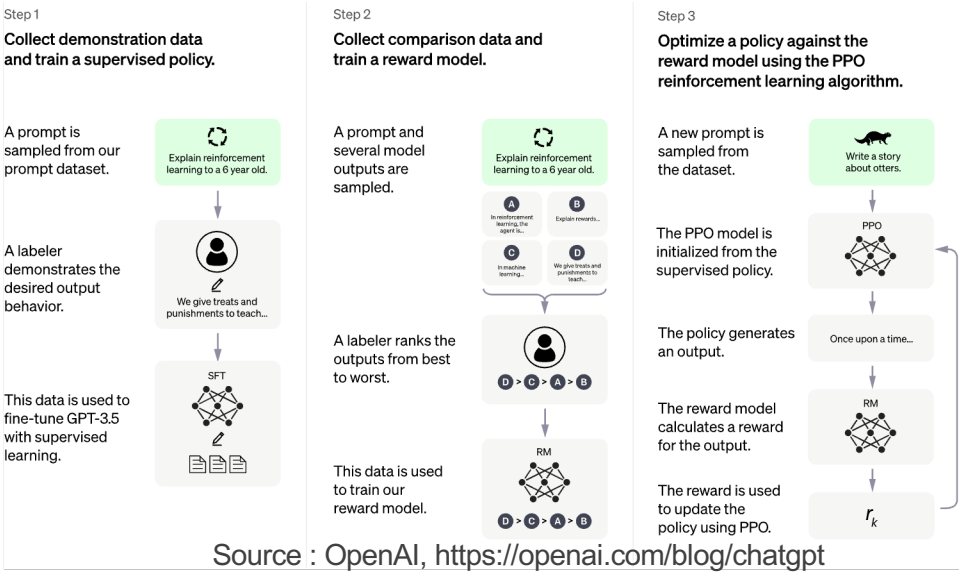
- Genèse de ChatGPT :
  - 2018 : GPT (Improving Language Understanding by Generative Pre-Training) ➔ Utilisation de la partie décodeur d'un transformer pré-entrainé et affiné pour réaliser différentes tâches (117 millions de paramètres)
  - 2019 : GPT-2 (Language Models are Unsupervised Multitask Learners) ➔ Q/A avec un prompt en langage naturel et des réponses proches de celles d'un humain (1,5 milliard de paramètres)
  - 2020 : GPT-3 (Language Models are Few-Shot Learners) ➔ amélioration du modèle précédent avec un réseau beaucoup plus grand (175 milliards de paramètres, 96 couches)
  - 2022 : ChatGPT & InstructGPT (Training language models to follow instructions with human feedback) ➔ amélioration du modèle en utilisant un apprentissage supervisé, et par renforcement
  - 2023 : GPT-4 ➔ modèle beaucoup plus grand que le précédent, multimodal (permet l'utilisation d'images et de textes en entrée) (1,6 trillions de paramètres)

ChatGPT

- Caractéristiques de chatGPT (mai 2023)
  - Basé sur l'architecture GPT-3.5 (GPT-4 pour la version payante)
  - Données d'apprentissage collectées jusqu'en septembre 2021
  - Modèle hors-ligne (n'a pas accès à internet)
  - Disponible en plusieurs langues (anglais, français, espagnol, chinois, …)
  - Jusqu'à 175 milliards de paramètres (en fonction du modèle GPT-3 utilisé)

ChatGPT

- Apprentissage de ChatGPT en 4 étapes principales
  - 0) Apprentissage auto-supervisé du modèle GPT-3.5 sous-jacent
  - 1) Apprentissage supervisé (Fine-Tuning) du modèle ChatGPT sur des exemples de questions/réponses rédigés par des humains
  - 2) Apprentissage supervisé d'un modèle d'évaluation des réponses
  - 3) Apprentissage par renforcement (Fine-Tuning) utilisant ce modèle d'évaluation pour améliorer les performances de ChatGPT



Source : OpenAI, https://openai.com/blog/chatgpt

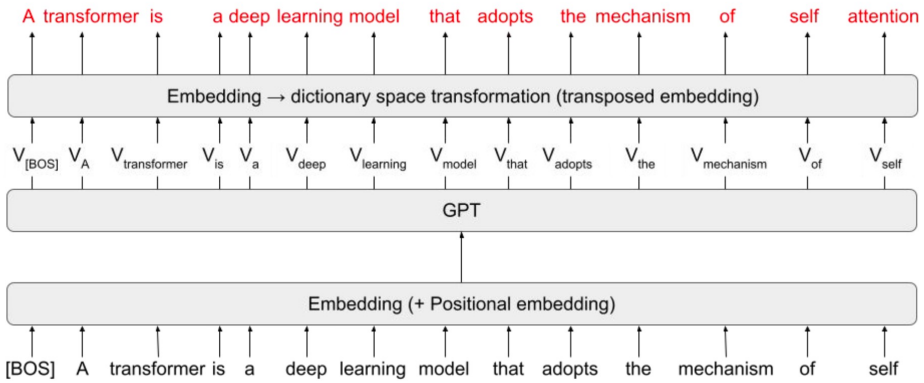
ChatGPT

- Etape 0 : Apprentissage non supervisé du modèle GPT-3.5
  - Objectif : apprendre les structures syntaxiques et grammaticales de la langue, ainsi que des connaissances générales sur de nombreux domaines
  - Données :

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Source : « Language Models are Few-Shot Learners», Brown et al., 2020

  - Stratégie : prédiction étape par étape du mot suivant dans une séquences de mot



Source : FIDLE, section 8 Transformers, https://gricad-gitlab.univ-grenoble-alpes.fr/talks/fidle/-/wikis/home

ChatGPT

- Etape 1 : Apprentissage supervisé du modèle ChatGPT (Fine Tuning)
  - Objectif : adapter le modèle à la tâche de conversation
  - Données : exemples de séquences de dialogues (questions/réponses) rédigés par des humains
  - Stratégie : l'apprentissage du modèle pré-entrainé est poursuivi avec ces données

- Etape 2 : Apprentissage supervisé d'un modèle d'évaluation des réponses
  - Objectif : obtenir un modèle qui sera capable d'évaluer les réponses fournies par le modèle génératif, nécessaire pour l'étape 3
  - Données : exemples de questions avec plusieurs réponses générées par le modèle. Un humain ordonne ces réponses par qualité
  - Stratégie : un réseau de neurones est appris à partir de des données

ChatGPT

- Etape 3 : Apprentissage par renforcement du modèle ChatGPT (Fine Tuning)
  - Objectif : optimiser le modèle
  - Données : exemples de questions
  - Stratégie :
    - le modèle ChatGPT génère une réponse pour chaque question choisie
    - Le modèle d'évaluation prédit la qualité de cette réponse
    - Cette valeur est utilisée comme signal de récompense d'un apprentissage par renforcement pour améliorer la qualité de génération du modèle

ÉCOLE CENTRALELYON

ChatGPT : Quelles limites ?

- Phénomène d'hallucination : ChatGPT peut générer des contre-vérités présentées de manière très cohérente et convaincante (et même de fausses sources)
- Apporte des réponses non sourcées (inadapté à la recherche d'informations), et non exhaustives
- Pose des problèmes de biais et de droits d'auteurs (les données utilisées pour l'apprentissage ne sont pas complètement connues)
- Pose des problèmes de confidentialité (les données utilisateurs sont utilisées par OpenAI)
- Pose des problèmes de fraude (appropriation de texte rédigé par autrui), non de plagiat (le texte généré par ChatGPT est généralement unique)

ChatGPT pour un développeur ?

- Obtenir des explications sur des concepts de programmation

  ➔ Ex de prompt : « Explique-moi le principe de l'héritage de la programmation objet »
- Générer des quizz pour tester ses connaissances

  ➔ Ex de prompt : « Génère-moi un quizz de 5 questions pour tester mes connaissances en programmation objet »
- Obtenir la documentation d'une fonction

  ➔ Ex de prompt : « Comment fonctionne la fonction plot de la librairie matplotlib ? »
- Obtenir des informations sur des éléments du langage

  ➔ Ex de prompt : « En Python, quelle est la différence entre les listes et les tuples ? »
- Générer des commentaires de code (dont docstring)

  ➔ Ex de prompt : « Réécris ce code Python en ajoutant les commentaires et les docstrings : [le code] »

ChatGPT pour un développeur ?

- Détecter et expliquer des erreurs de codage

  ➔ Ex de prompt : « Quelle est l'erreur dans ce code Python : [le code] »
- Générer du code pour faire un traitement particulier

  ➔ Ex de prompt : « Utilise la librairie matplotlib pour réaliser l'affichage d'une courbe en 3D »
- Améliorer l'écriture / l'efficacité d'un code (temps, mémoire)

  ➔ Ex de prompt : « Réécris ce code en utilisant les recommandations PEP8 : [le code] »
- Générer les tests unitaires d'un code

  ➔ Ex de prompt : « Ecris moi les tests unitaires pour cette classe : [le code] »
- …

ChatGPT : conseils

- Soigner le prompt :
  - Donner le contexte
  - Être précis dans les instructions données
  - Affiner itérativement les instructions si besoin
- **Faire preuve d'esprit critique**
  - Le code généré ne doit jamais être utilisé sans vérification car il peut contenir des erreurs et/ou ne pas être adapté à la problématique (problème d'hallucination)

➔ ChatGPT ne remplacera pas le savoir-faire et l'expertise du programmeur, mais peut l'aider à gagner en efficacité

ÉCOLE CENTRALELYON

Exemples d'outils pour l'analyse de données massives

Apache Mahout

- Librairie Java d'implémentations parallélisables des principaux algorithmes de Machine Learning
- De nombreuses implémentations reposent sur MapReduce
- Récemment, orientation vers YARN (gestion des ressources et de planification de tâches)
- Algorithmes proposés :
  - Classification (random forest, naïve Bayes, regression logistique, …)
  - Clustering (k-means, …)
  - Réduction dimensionnelle (analyse en composantes principales, …)

ÉCOLE CENTRALELYON

Data Science Studio

- Edité par la société Dataiku
- Plateforme logicielle basée sur Hadoop
- Combinaison de briques logicielles dont des outils de gestion de données, de statistiques, de visualisation et d'analyse prédictive
- Parmi les fonctionnalités : le nettoyage, l'enrichissement et la modélisation des données, des fonctions de Machine Learning et d'analyse prédictive

Weka

- Développé par l'université de Waikato en Nouvelle-Zélande
- Suite logicielle d'outils de Machine Learning
- Licence open source, écrit en Java (donc multi-plateformes)
- Utilisable par le biais d'une interface, ou par une API
- Dispose de modules pour être appliqué sur des données massives :
  - DistributedWekaHadoop (map et reduce specifique à Hadoop)
  - DistributedWekaSpark (map et reduce specifique à Spark)
  - DistributedWekaBase (map et reduce non spécifique)

PyTorch

- Librairie Python open source pour l'apprentissage profond
- Permet de paralléliser les calculs sur cartes GPU
- Mise en œuvre simplifiée d'apprentissages pour des modèles à base de réseaux de neurones :
  - Réseaux convolutifs
  - Transformers

ÉCOLE CENTRALELYON

Exemple d'Application d'une méthode de Machine Learning avec Weka

Weka

- Télécharger et installer Weka :

https://waikato.github.io/weka-wiki/downloading_weka/

- Réaliser différentes expérimentations
  - en considérant plusieurs modèles de classification (arbres de décision, forêts aléatoires, classification bayésien, réseau de neurones, …)
  - sur différents jeux de données :
    - breast-cancer : prédiction du cancer du sein
    - diabetes : prédiction du diabète
    - vote : prédiction du parti politique (démocrate, républicain)
    - ...

ÉCOLE CENTRALELYON

Quelques références

- Big Data et Machine Learning, P. Lemberger, M. Batty, M. Morel, J.-L. Raffaëli, Editions Dunod, 2015
- Big Data, Data Mining, and Machine Learning, J. Dean, Editions Wiley, 2014
- Apache Mahout, http://mahout.apache.org
- Data Science Studio, Dataiku, http://www.dataiku.com/dss/
- Weka, https://www.cs.waikato.ac.nz/ml/weka/
- PyTorch, https://pytorch.org/

ÉCOLE
CENTRALE LYON

36 av. Guy de Collongue
69134 Écully cedex
T + 33 (0)4 72 18 60 00
www.ec-lyon.fr