Introduction à l'intelligence artificielle

Les différents types d'apprentissage

Laetitia Chapel - Professeure en IA à l'institut Agro Rennes-Angers - laboratoire IRISA

janvier 2025

Objectifs pédagogiques

A la fin de la séance, vous serez capable de :

- distinguer apprentissage supervisé / non supervisé / par renforcement
- comprendre ce que le modèle apprend (et ce qu'il ne comprend pas)
- manipuler des cas jouets simples, proches de l'agronomie
- avoir une intuition des limites (données, biais, sur-apprentissage)

Facteurs clés des performances de l'IA

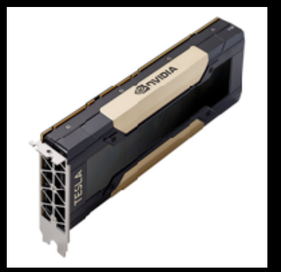
Une « nouvelle » puissance de calcul : les GPU

- Processeur ou unité de calcul
  - peut être composé de plusieurs coeurs (=mini-processeurs)
  - concept clé : les FLOPS (Floating Point Operations Per Second)
  - peut effectuer une même instruction en parallèle sur des données différentes

Intel Xeon E5 - 2680v4

28 coeurs, 268.8 GigaFlops

- CPU vs GPU
  - CPU : quelques coeurs pour des calculs généraux (calcul en ligne)
  - GPU : des centaines voire milliers de petits coeurs (calcul parallèle)

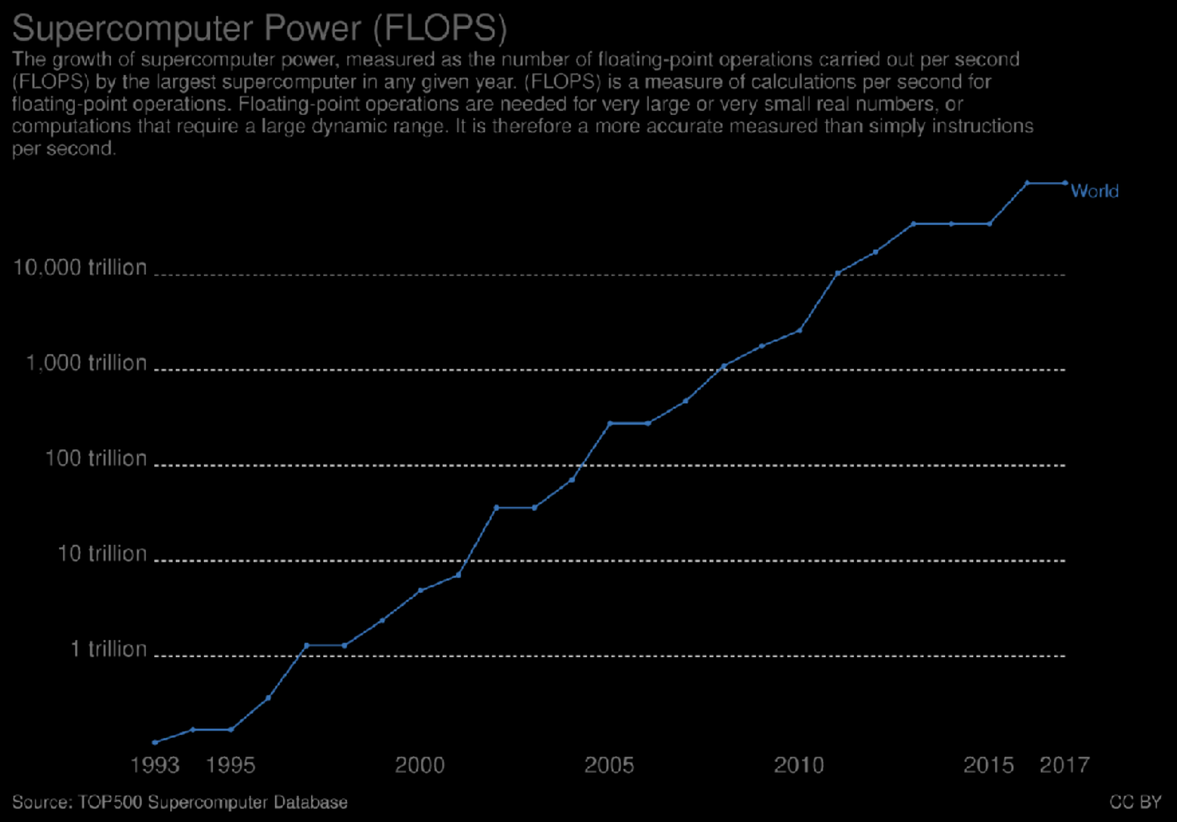


NVIDIA tesla P100

43000 coeurs = 149 TeraFlops

Facteurs clés

Une « nouvelle » puissance de calcul : les GPU



année	dollars par FLOP (corrigé)
1945	$2 quadrillion
1984	$100,000,000
1997	$57,000
Avril 2000	$2,000
Mai 2000	$1,000
Août 2003	$100
Août 2012	$1
Juin 2015	$0.1
Mai 2023	$0.01

source : wikipedia

Facteurs clés

Une « nouvelle » puissance de calcul : les GPU

...qui permet d'utiliser des modèles de plus en plus complexes et volumineux...

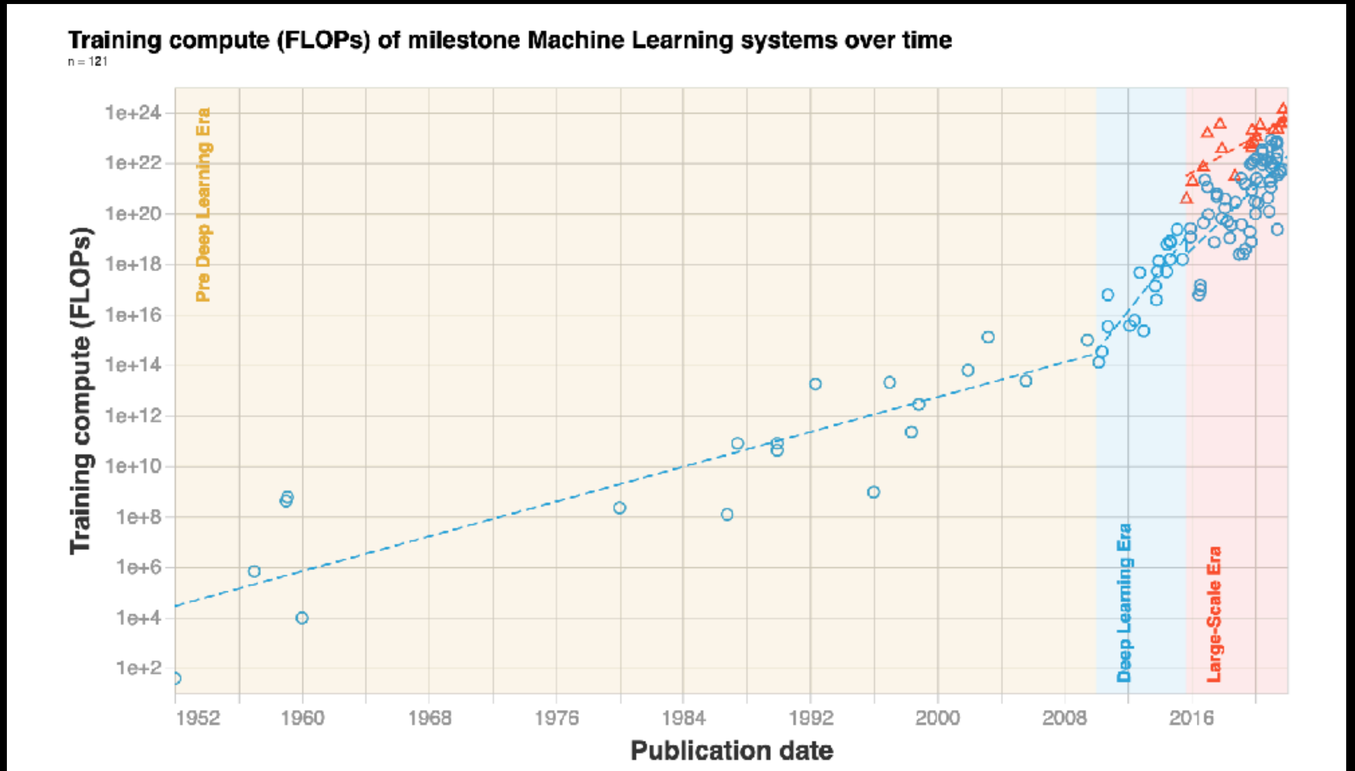


Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

source : COMPUTE TRENDS ACROSS THREE ERAS OF MaCHINE LEARNING

Facteurs clés

Une nouvelle profusion de données d'entraînement

... entrainés sur des données de plus en plus volumineuses

Computer vision

Data-set		Year	Nb. images	Size
MNIST	(classification)	1998	60K	12Mb
Caltech 101	(classification)	2003	9.1K	130Mb
Caltech 256	(classification)	2007	30K	1.2Gb
CIFAR10	(classification)	2009	60K	160Mb
ImageNet	(classification)	2012	1.2M	150Gb
MS-COCO	(segmentation)	2015	200K	32Gb
Cityscape	(segmentation)	2016	25K	60Gb
LAION-5B	(multi-modal)	2022	5.85B	240Tb

Natural Language Processing

Data-set		Year	Size
SST2	(sentiment analysis)	2013	20Mb
WMT-18	(translation)	2018	7Gb
OSCAR	(language model)	2020	6Tb

Quelle est la taille du training set de chatgpt ?

OpenAI n'a pas communiqué la **taille exacte** du jeu de données (training set) utilisé pour entraîner ChatGPT, y compris ses différentes versions comme GPT-3 ou GPT-4. Cependant, nous avons quelques informations générales sur la nature de ces données :

1. **Corpus varié et immense** : Le modèle est formé à partir de vastes ensembles de données textuelles disponibles publiquement, comprenant des livres, des articles scientifiques, des sites web, des forums, des articles de presse, et d'autres formes de contenu en ligne. Le volume total de ces données est énorme, mais OpenAI n'a pas fourni de chiffre spécifique.
2. **Données de sources diverses** : Les sources de données incluent des livres, Wikipédia, des archives de recherche, des discussions en ligne et bien d'autres contenus textuels disponibles publiquement avant la **fin de l'année 2021** (pour GPT-4).
3. **Ordre de grandeur** : Pour GPT-3, il a été révélé que le modèle a été formé sur environ **570 Go de données textuelles** après prétraitement, provenant d'environ **300 milliards de tokens** (morceaux de texte). Cependant, ces chiffres ne sont pas applicables directement à GPT-4, car il est probable que GPT-4 a été formé sur un ensemble de données encore plus grand et diversifié.

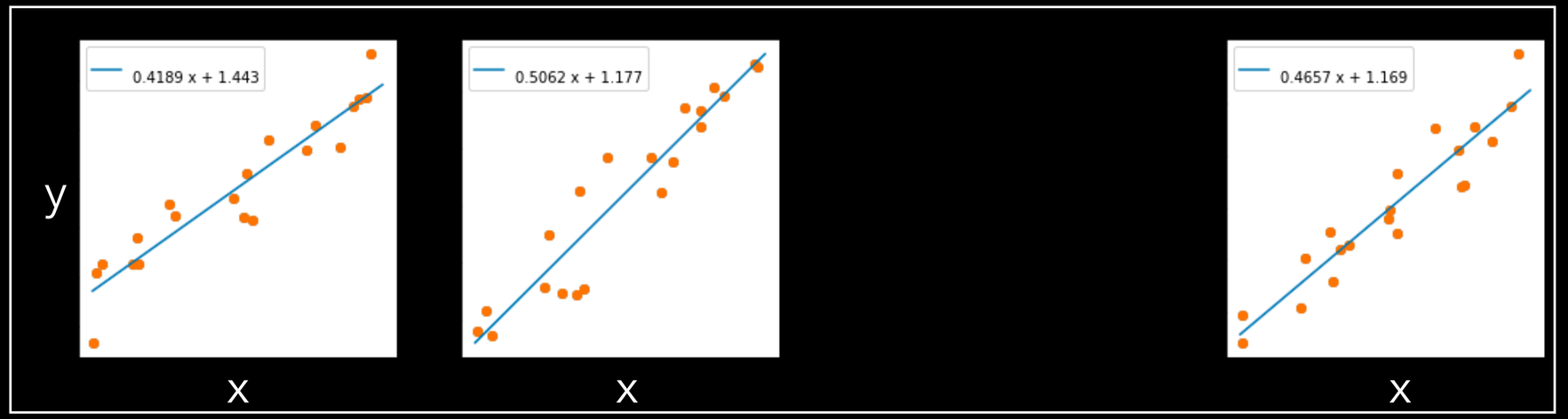
En résumé :

Bien que la taille exacte du jeu de données utilisé pour entraîner ChatGPT, notamment GPT-4, ne soit pas publique, elle est probablement énorme, englobant plusieurs centaines de gigaoctets, voire des téraoctets de texte, provenant de diverses sources accessibles publiquement.

Facteurs clés

Une nouvelle profusion de données d'entraînement

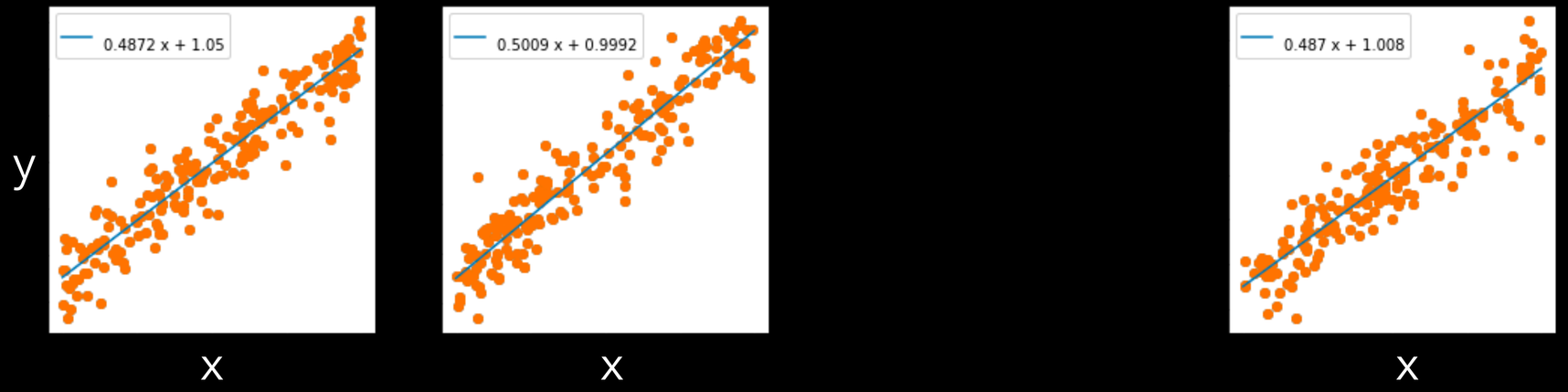
- Plus on a de points, mieux c'est ! (sous réserve que l'on ait des ressources mémoire/ calcul suffisantes…)
- Un exemple avec un modèle théorique $y = 1 + 0.5x + \epsilon$ avec $\epsilon \sim \mathcal{N}(0,1)$
- M tirages de $N = 20$ points, avec $x \in U[0,20]$



Facteurs clés

Une nouvelle profusion de données d'entraînement

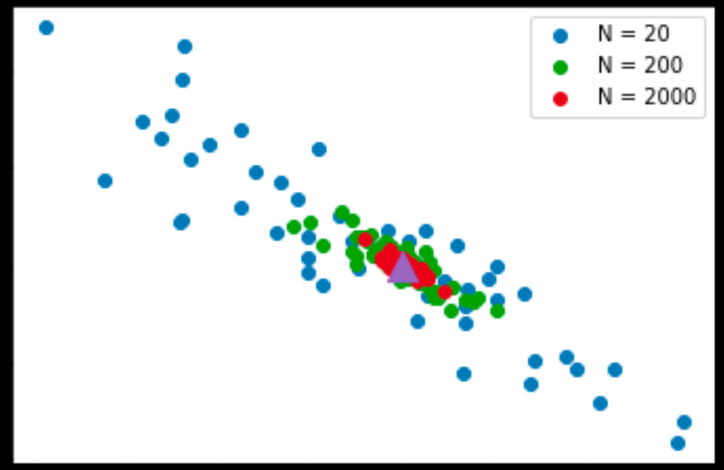
- Plus on a de points, mieux c'est ! (sous réserve que l'on est des ressources mémoire/ calcul suffisantes…)
- Un exemple avec un modèle théorique $y = 1 + 0.5x + \epsilon$ avec $\epsilon \sim \mathcal{N}(0,1)$
- M tirages de $N = 200$ points, avec $x \in U[0,20]$



Facteurs clés

Une nouvelle profusion de données d'entraînement

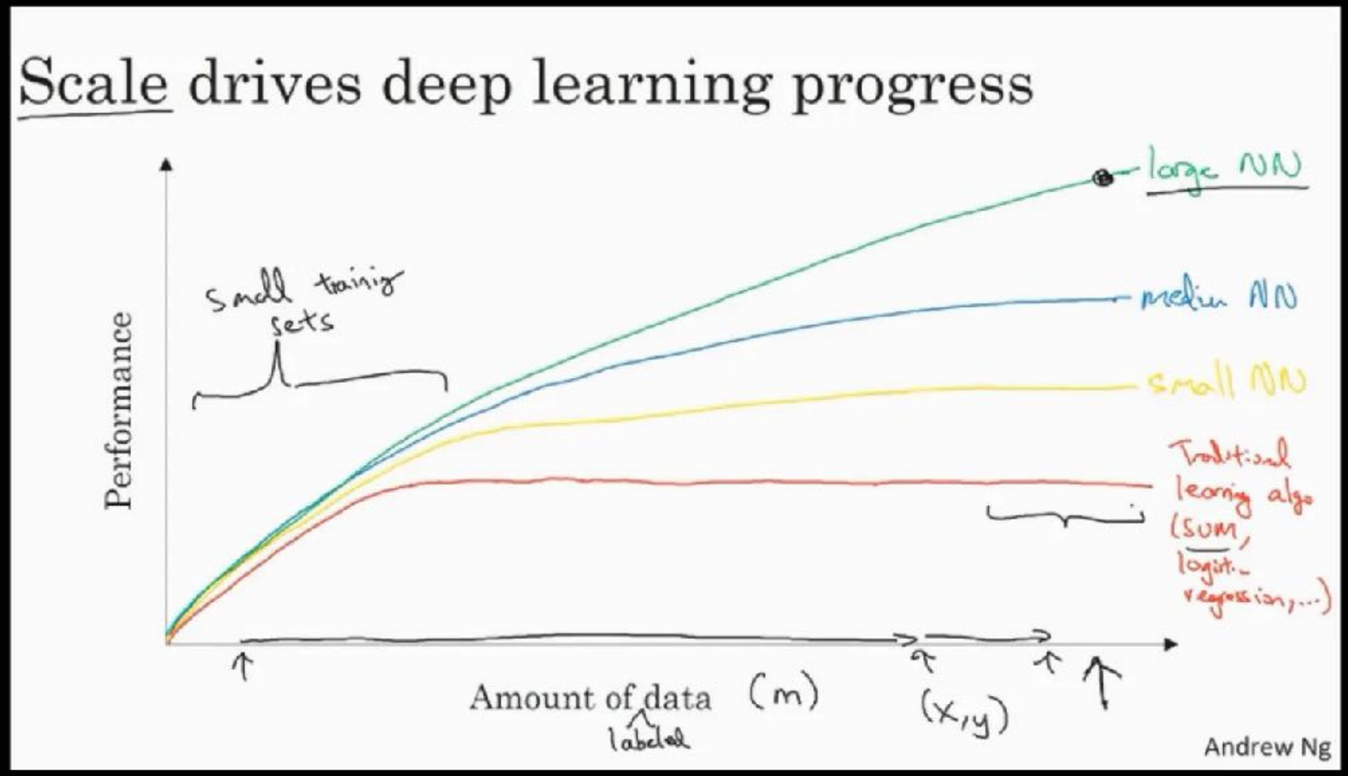
- Plus on a de points, mieux c'est ! (sous réserve que l'on est des ressources mémoire/ calcul suffisantes…)
- Un exemple avec un modèle théorique $y = 1 + 0.5x + \epsilon$ avec $\epsilon \sim \mathcal{N}(0,1)$
- Pour $N = 20,\ 200,\ 2000$ points, impact sur la variance (pour une complexité fixée)



Facteurs clés

Une nouvelle profusion de données d'entraînement

- Influence de la quantité de données sur la performance



Le mot-clé : la généralisation

Imiter l'intelligence humaine

cténophore

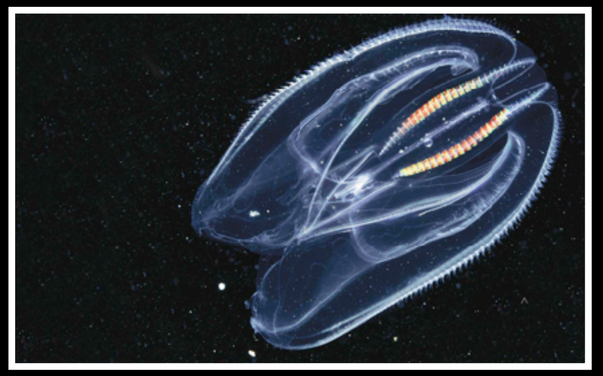
salpe

Le mot-clé : la généralisation

Imiter l’intelligence humaine

cténophore

salpe



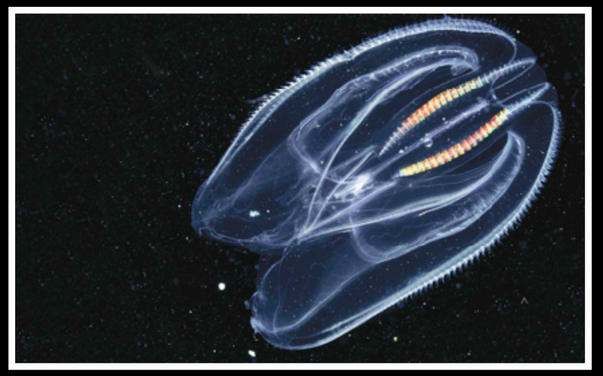
Le mot-clé : la généralisation

Imiter l'intelligence humaine

cténophore

salpe

sur de nouvelles données



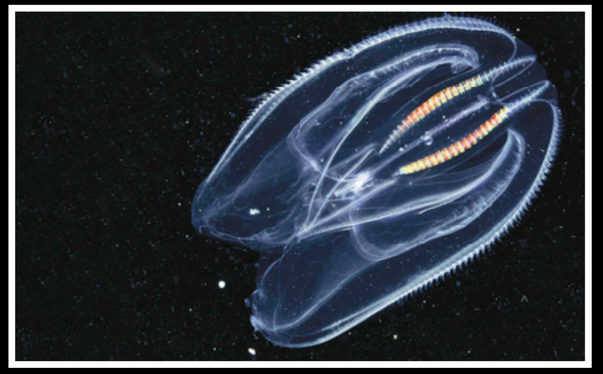
Le mot-clé : la généralisation

Imiter l'intelligence humaine

cténophore

salpe

sur de nouvelles données



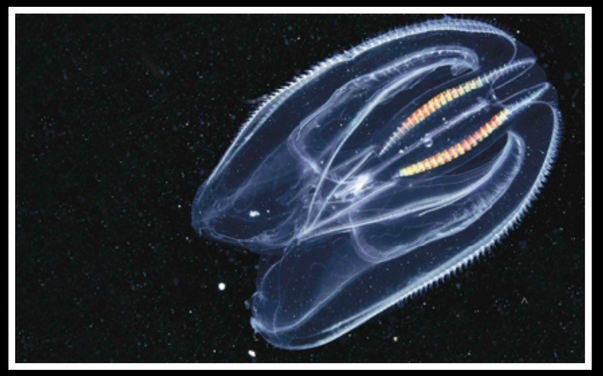
Le mot-clé : la généralisation

Imiter l'intelligence humaine

cténophore

salpe

sur de nouvelles données



être capable de transfert

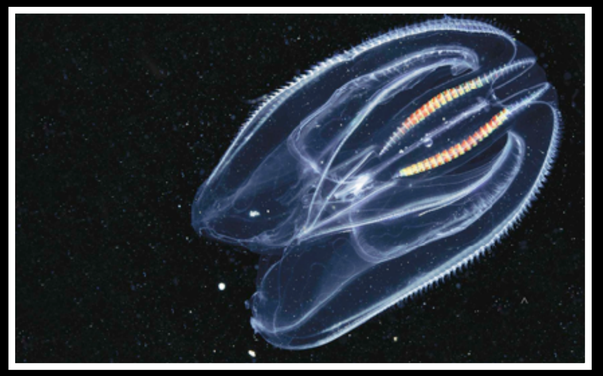
Le mot-clé : la généralisation

Imiter l'intelligence humaine

cténophore

salpe

sur de nouvelles données



être capable de transfert

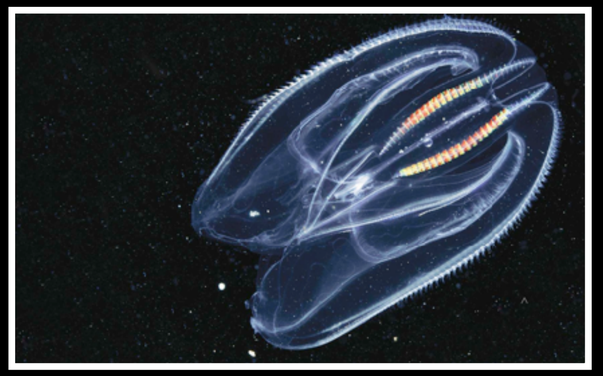
Le mot-clé : la généralisation

Imiter l'intelligence humaine

cténophore

salpe

sur de nouvelles données



être capable de transfert

en ligne

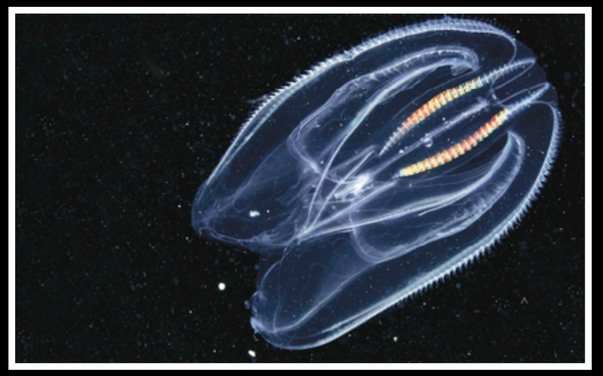
Le mot-clé : la généralisation

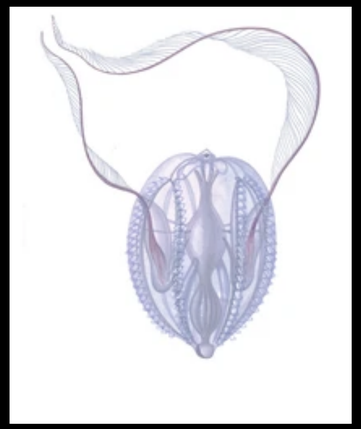
Imiter l'intelligence humaine

cténophore

salpe

sur de nouvelles données





être capable de transfert

en ligne

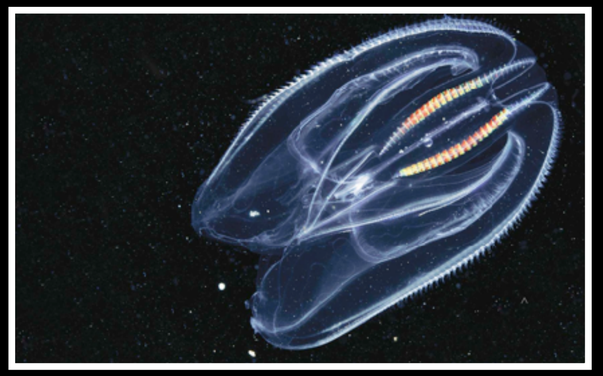
Le mot-clé : la généralisation

Imiter l'intelligence humaine

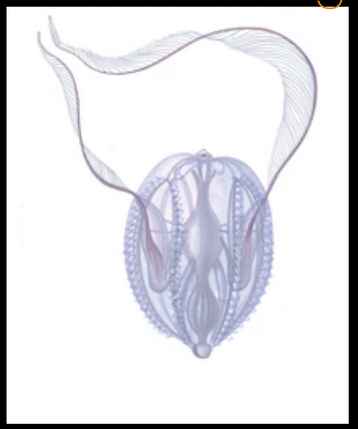
cténophore

salpe

sur de nouvelles données



données hétérogènes



être capable de transfert

en ligne

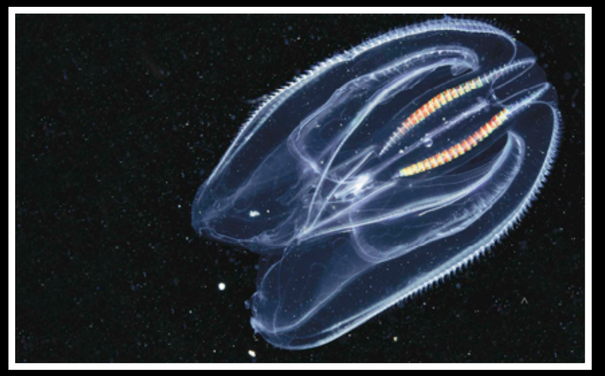
Le mot-clé : la généralisation

Imiter l'intelligence humaine

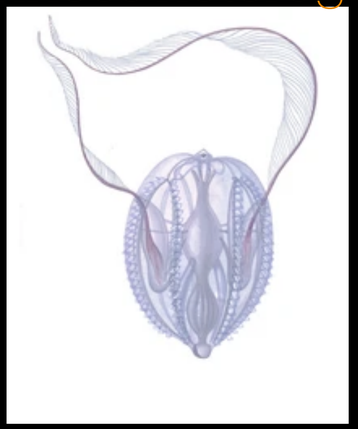
cténophore

salpe

sur de nouvelles données



données hétérogènes



généralisation ≠ apprentissage par coeur

être capable de transfert

en ligne

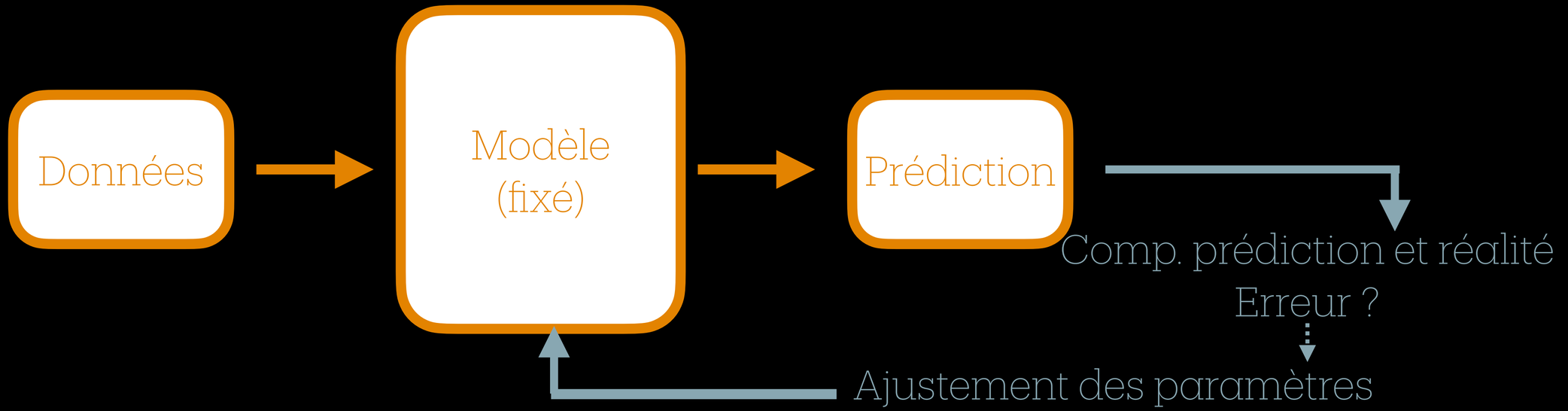
Où allons nous ?

- Comment une machine apprend ?
- À partir de quels types de données ?
- Avec quels objectifs ?

Apprendre = ajuster un modèle à partir de données

Apprentissage machine

- Un algorithme ***apprend*** quand il améliore ses performances sur une tâche à partir de données, sans être explicitement programmé pour chaque cas.



- Selon le type de données disponibles et l'objectif, on va apprendre de façons différentes

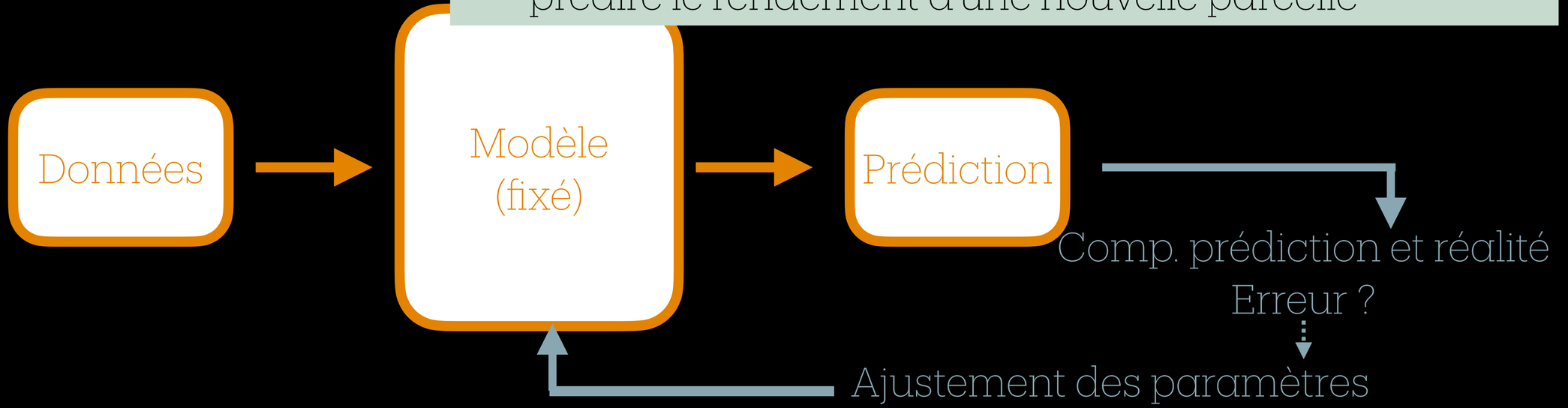
  ⇒ différents types d'apprentissage.

Apprentissage m

- Un algorithme ***apprend*** quand il amé
  données, sans être explicitement prog

Exemple (très simple)

- On donne à la machine :
  - des caractéristiques de parcelles (pluie, sol, azote)
  - le rendement observé
- Elle apprend à :
  - prédire le rendement d'une nouvelle parcelle



- Selon le type de données disponibles et l'objectif, on va apprendre de façons différentes

  ⇒ différents types d'apprentissage.

Trois grands types d'apprentissage

Type d'apprentissage	Données	Objectif
Supervisé	données + réponse	prédiction
Non supervisé	données seules	découverte de structure
Renforcement	interaction + récompense	décider

Apprentissage supervisé

Principe

- Prédire une relation entre les descripteurs des données et leurs étiquettes
- Étiquetage = associer une information / contexte / signification à des données brutes
- X : descripteurs, Y : étiquette $\Rightarrow$ trouver une fonction $f(X) \approx Y$
- Deux grandes sous-familles :
  - Classification $Y \in \{ \dots \}$ un ensemble fini de valeurs possibles
  - Régression $Y \in \mathbb{R}$ un ensemble continu de valeur

Apprentissage supervisé

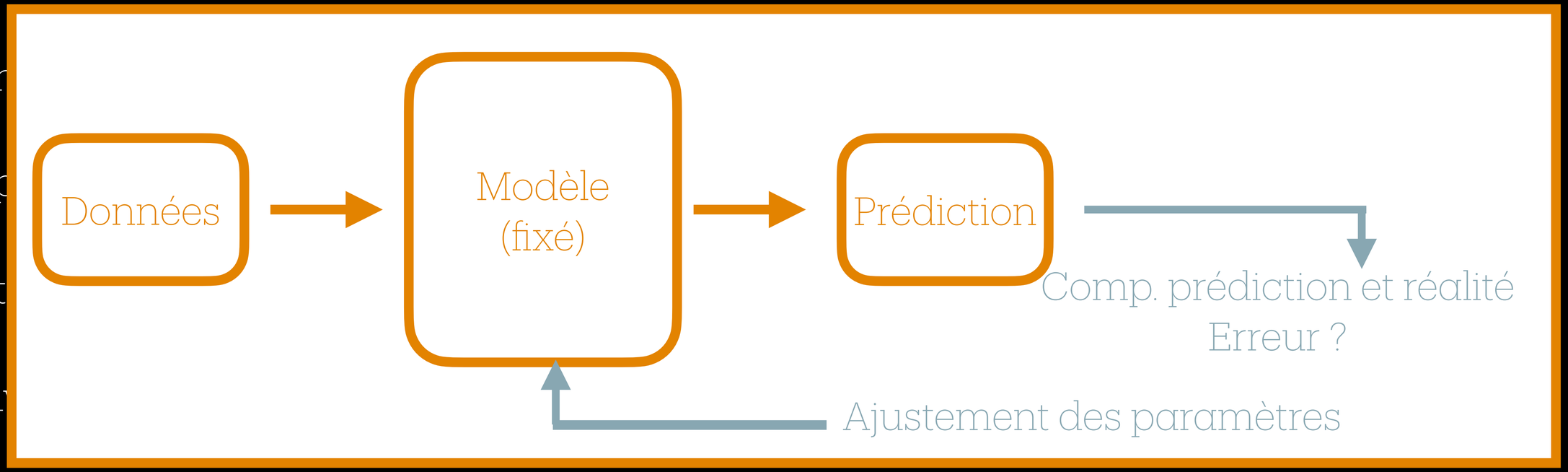
Cas jouet pour un problème de régression

- Peut-on prédire le rendement d'une parcelle à partir de données simples ?
- Descripteurs/variable X : pluie cumulée, apport azoté, température
- Etiquette Y : rendement
- $\Rightarrow$ trouver une fonction $f(X) \approx Y \Rightarrow f(pluie, azote, temp) \approx rendement$
- Modèle linéaire
  - Rendement ≈ a × Pluie + b × Azote + c × Température + d

Apprentissage supervisé

Cas jouet pour un problème de régression

- Peut-on
- Descrip
- Etiquet
- ⇒ trou
- Modèle linéaire
  - Rendement ≈ a × Pluie + b × Azote + c × Température + d



Apprentissage supervisé

Cas jouet pour un problème de régression

- Hypothèses :
  - les descripteurs sont utiles pour prédire l'étiquette/label
  - la fonction f que l'on a choisi permet de bien modéliser le phénomène
  - on est capables de trouver les bons paramètres
  - on est capable d'évaluer correctement le modèle f

Apprentissage supervisé

Cas jouet pour un problème de régression

- Hypothèses :
  - les descripteurs sont utiles pour prédire l'étiquette/label
  - la fonction f que l'on a choisi permet de bien modéliser le phénomène
  - on est capables de trouver les bons paramètres
  - on est capable d'évaluer correctement le modèle f

Comment dissocier généralisation
et apprentissage par coeur ?

Apprentissage supervisé

Cas jouet pour un problème de régression

- Hypothèses :
  - les descripteurs sont utiles pour prédire l'étiquette/label
  - la fonction f que l'on a choisi permet de bien modéliser le phénomène
  - on est capables de trouver les bons paramètres
  - on est capable d'évaluer correctement le modèle f

Beaucoup de données annotées, de bons descripteurs !

Comment dissocier généralisation et apprentissage par coeur ?

Apprentissage supervisé

Cas jouet pour un problème de régression

- Hypothèses :
  - les descripteurs sont utiles pour prédire l'étiquette/label
  - la fonction f que l'on a choisi permet de bien modéliser le phénomène
  - on est capables de trouver les bons paramètres
  - on est capable d'évaluer correctement le modèle f

Beaucoup de données annotées, de bons descripteurs !

Plus elle est compliquée, plus on a des chances que la prédiction sera bonne

Comment dissocier généralisation et apprentissage par coeur ?

Apprentissage supervisé

Cas jouet pour un problème de régression

- Hypothèses :
  - les descripteurs sont utiles pour prédire l'étiquette/label
  - la fonction f que l'on a choisi permet de bien modéliser le phénomène
  - on est capables de trouver les bons paramètres
  - on est capable d'évaluer correctement le modèle f

Beaucoup de données annotées, de bons descripteurs !

Plus elle est compliquée, plus on a des chances que la prédiction sera bonne

Bonnes procédures d'optim + puissance de calcul

Comment dissocier généralisation et apprentissage par coeur ?

Apprentissage supervisé

Jeu d'entrainement vs jeu de test

- Si le modèle f est très complexe, l'apprentissage par coeur est possible
- La généralisation mesure la capacité d'un modèle à « donner la bonne réponse » sur de nouvelles données

Apprentissage supervisé

Jeu d'entrainement vs jeu de test

- Si le modèle f est très complexe, l'apprentissage par coeur est possible
- La généralisation mesure la capacité d'un modèle à « donner la bonne réponse » sur de nouvelles données

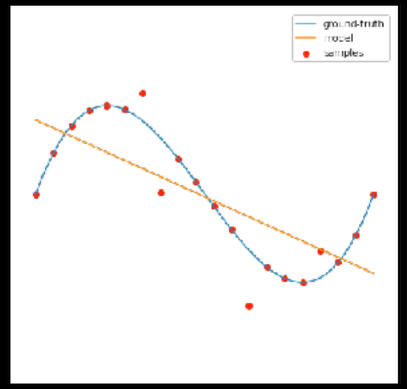
On va faire « semblant » d'avoir de nouvelles données
- Jeu d'apprentissage : pour « apprendre » le modèle
- Jeu de test : pour tester la capacité de généralisation

Sur-apprentissage : bonne précision, mauvaise généralisation

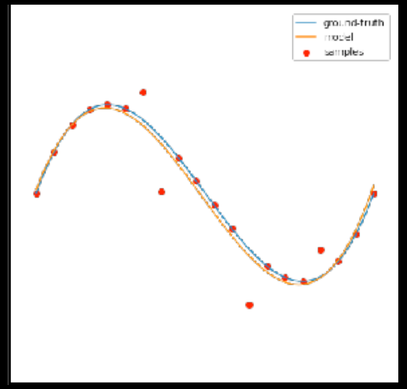
Apprentissage supervisé

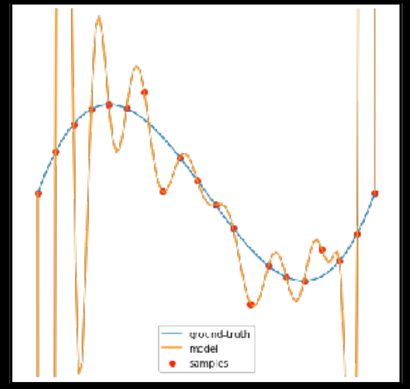
Sur-apprentissage

- Si le modèle f est très complexe, l'apprentissage par coeur est possible
- La généralisation mesure la capacité d'un modèle à « donner la bonne réponse » sur de nouvelles données



Sous-apprentissage



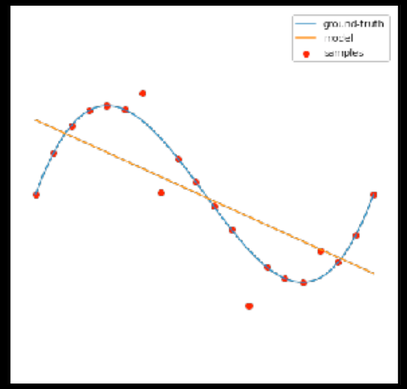


Sur-apprentissage

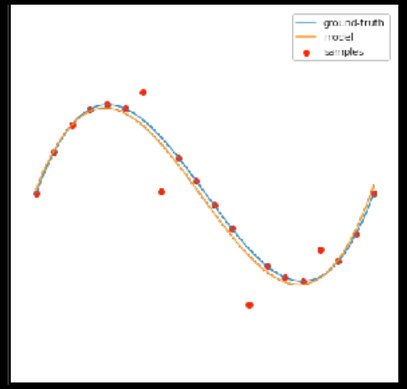
Apprentissage supervisé

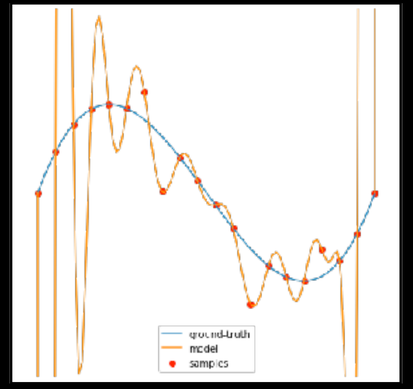
Sur-apprentissage

- Si le modèle f est très complexe, l'apprentissage par coeur est possible
- La généralisation mesure la capacité d'un modèle à « donner la bonne réponse » sur de nouvelles données



Sous-apprentissage





Sur-apprentissage

Moins « probable » quand on a beaucoup de données

Apprentissage supervisé

Cas jouet pour un problème de classification

- Prédiction de maladie de plants de tomate à partir de photos de feuilles

  - les descripteurs ? Il faut être capable de les « déduire » ou « apprendre » à partir des données (on verra ça plus tard....)

  - les étiquettes sont discrètes (feuille saine, feuille avec du mildiou) - potentiellement un grand nombre de classes

Apprentissage supervisé

Cas jouet pour un problème de classification

- Prédiction de maladie de plants de tomate à partir de photos de feuilles

mildiou précoce

  - les descripteurs ? Il faut être capable de les « déduire » ou « apprendre » à partir des données (on verra ça plus tard....)

  - les étiquettes sont discrètes (feuille saine, feuille avec du mildiou) - potentiellement un grand nombre de classes

Apprentissage supervisé

Cas jouet pour un problème de classification

- Prédiction de maladie de plants de tomate à partir de photos de feuilles

mildiou précoce

saine

  - les descripteurs ? Il faut être capable de les « déduire » ou « apprendre » à partir des données (on verra ça plus tard....)

  - les étiquettes sont discrètes (feuille saine, feuille avec du mildiou) - potentiellement un grand nombre de classes

Un exemple d'apprentissage

Teachable Machine

Reconnaissance de feuilles de tomates malades

- TEST 1 : une image d'apprentissage par classe

1. Importer les 2 images du répertoire apprentissage de TEST1, renommer les classes et faire la prédiction pour la feuille de tomate saine. Que remarquez-vous ?

Un exemple d'apprentissage

Teachable Machine

Reconnaissance de feuilles de tomates malades

- TEST 1 : une image d'apprentissage par classe

1. Importer les 2 images du répertoire apprentissage de TEST1, renommer les classes et faire la prédiction pour la feuille de tomate saine. Que remarquez-vous ?
    - le modèle prédit une probabilité + pas d'apprentissage par coeur !
    - la probabilité la plus forte pour l'image saine est mildiou : le modèle n'est pas très performant... Il faut dire que l'on a une seule donnée d'apprentissage !

Un exemple d'apprentissage

Teachable Machine

Reconnaissance de feuilles de tomates malades

- TEST 1 : une image d'apprentissage par classe

1. Importer les 2 images du répertoire apprentissage de TEST1, renommer les classes et faire la prédiction pour la feuille de tomate saine. Que remarquez-vous ?
    - le modèle prédit une probabilité + pas d'apprentissage par coeur !
    - la probabilité la plus forte pour l'image saine est mildiou : le modèle n'est pas très performant... Il faut dire que l'on a une seule donnée d'apprentissage !
2. Tester sur une image non vue pendant l'entraînement

Un exemple d'apprentissage

Teachable Machine

Reconnaissance de feuilles de tomates malades

- TEST 1 : une image d'apprentissage par classe

1. Importer les 2 images du répertoire apprentissage de TEST1, renommer les classes et faire la prédiction pour la feuille de tomate saine. Que remarquez-vous ?
    - le modèle prédit une probabilité + pas d'apprentissage par coeur !
    - la probabilité la plus forte pour l'image saine est mildiou : le modèle n'est pas très performant... Il faut dire que l'on a une seule donnée d'apprentissage !
2. Tester sur une image non vue pendant l'entraînement

    - On apprend sur des données, on évalue la capacité de généralisation sur le test

Un exemple d'apprentissage

Reconnaissance de feuilles de tomates malades

- TEST 2 : 20 images d'apprentissage par classe

1. Importer les images du répertoire apprentissage de TEST2, renommer les classes et faire la prédiction pour les données de test (les mêmes que pour le TEST1). Que remarquez-vous ?

Un exemple d'apprentissage

Reconnaissance de feuilles de tomates malades

- TEST 2 : 20 images d'apprentissage par classe

1. Importer les images du répertoire apprentissage de TEST2, renommer les classes et faire la prédiction pour les données de test (les mêmes que pour le TEST1). Que remarquez-vous ?
    - le modèle est beaucoup plus confiant !

Un exemple d'apprentissage

Reconnaissance de feuilles de tomates malades

- TEST 2 : 20 images d'apprentissage par classe

1. Importer les images du répertoire apprentissage de TEST2, renommer les classes et faire la prédiction pour les données de test (les mêmes que pour le TEST1). Que remarquez-vous ?
    - le modèle est beaucoup plus confiant !
2. On a maintenant des photos de moisissure de feuille de tomate. Tester le modèle dessus et discuter le résultat

Un exemple d'apprentissage

Reconnaissance de feuilles de tomates malades

- TEST 2 : 20 images d'apprentissage par classe

1. Importer les images du répertoire apprentissage de TEST2, renommer les classes et faire la prédiction pour les données de test (les mêmes que pour le TEST1). Que remarquez-vous ?
    - le modèle est beaucoup plus confiant !
2. On a maintenant des photos de moisissure de feuille de tomate. Tester le modèle dessus et discuter le résultat
    - Aucune chance qu'il donne la bonne réponse puisqu'il n'a pas été appris sur cette classe

Un exemple d'apprentissage

Détection de taches sur les feuilles de tomates

- Détection de taches sur des feuilles de tomates
- Tâche de **segmentation sémantique** : tâche de vision par ordinateur qui consiste à attribuer une étiquette à chaque pixel d'une image, afin d'identifier les différentes catégories présentes (ex. sol, plante, ciel).

Un exemple d'apprentissage

Détection de tâches sur les feuilles de tomates

- On va utiliser la librairie SAM (=segment anything) — « try the playground » — https://ai.meta.com/sam3/ qui permet de détecter les objets à partir d'un prompt

- Pas besoin d'apprentissage, c'est un **modèle de fondation** ≈ il n'a pas été appris sur nos données mais a tellement vu d'exemples lors de l'apprentissage qu'il généralise (très) bien

1. Reprendre les images d'apprentissage de TEST1 et détecter le contour de la feuille

2. Tester avec la feuille qui contient des moisissures et essayer de segmenter les taches (=spot)

Apprentissage non supervisé

Les classes ne sont pas disponibles

- Hypothèses :
  - on dispose de descripteurs X mais pas d'étiquette
  - Objectif : trouver de la structure dans les données
  - Par exemple : faire des groupes d'individus qui se ressemblent, segmenter des images en zones homogènes

Apprentissage par renforcement

On apprend des actions

- L'apprentissage par renforcement, c'est :
  - Un agent
  - Un environnement
  - Des actions
  - Une récompense

- pour résoudre des problèmes tels que :
  - définir une stratégie d'irrigation
  - prendre des décisions de fertilisation

- Un type d'apprentissage beaucoup plus compliqué, qui nécessite beaucoup de données et de grosses compétences dédiées