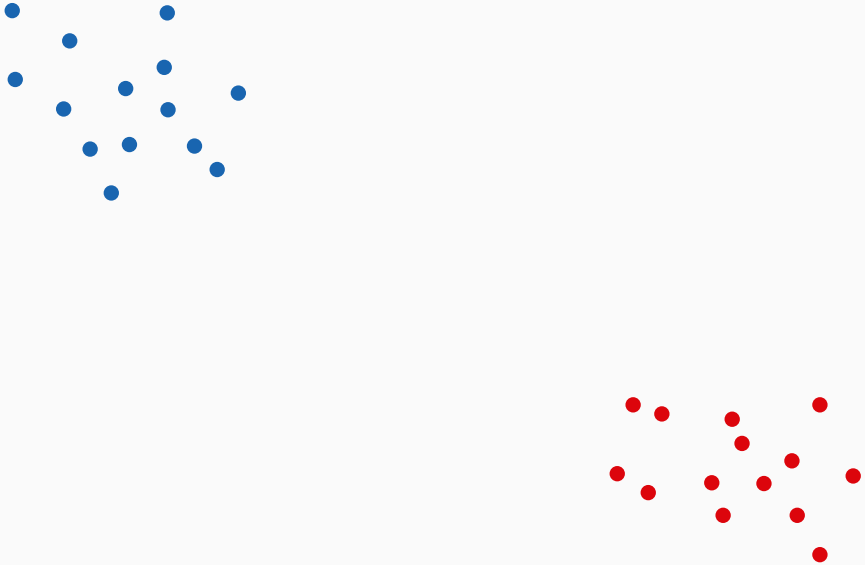
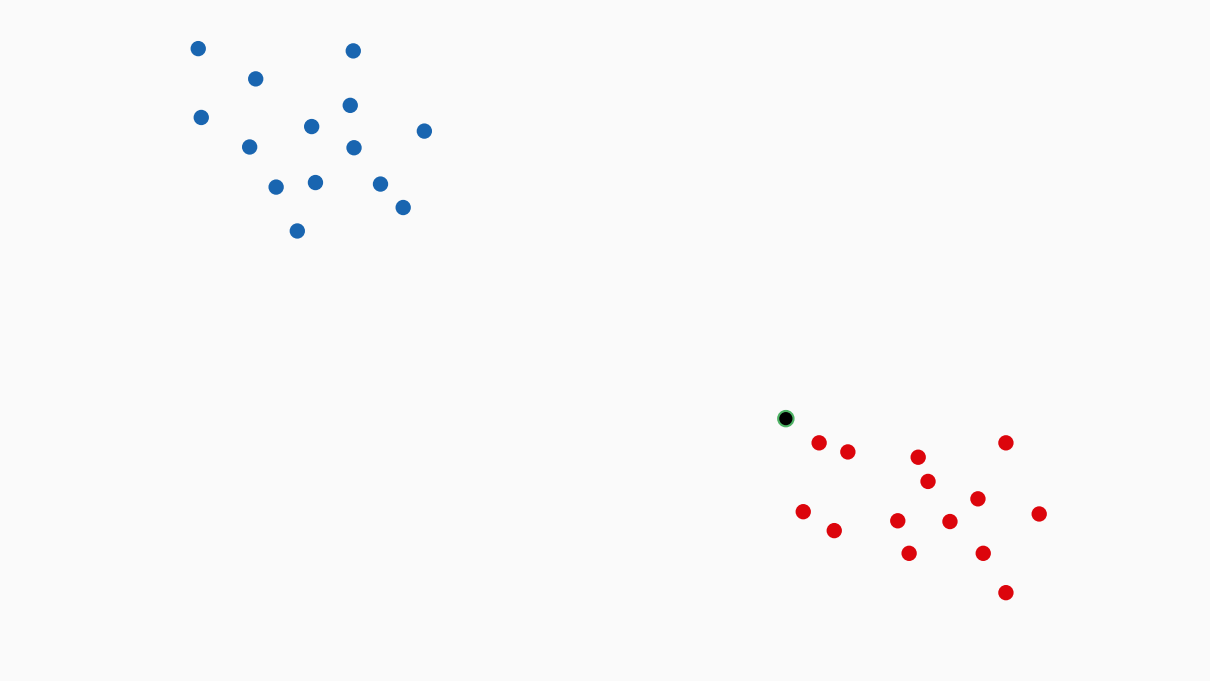
Apprentissage Automatique : Introduction

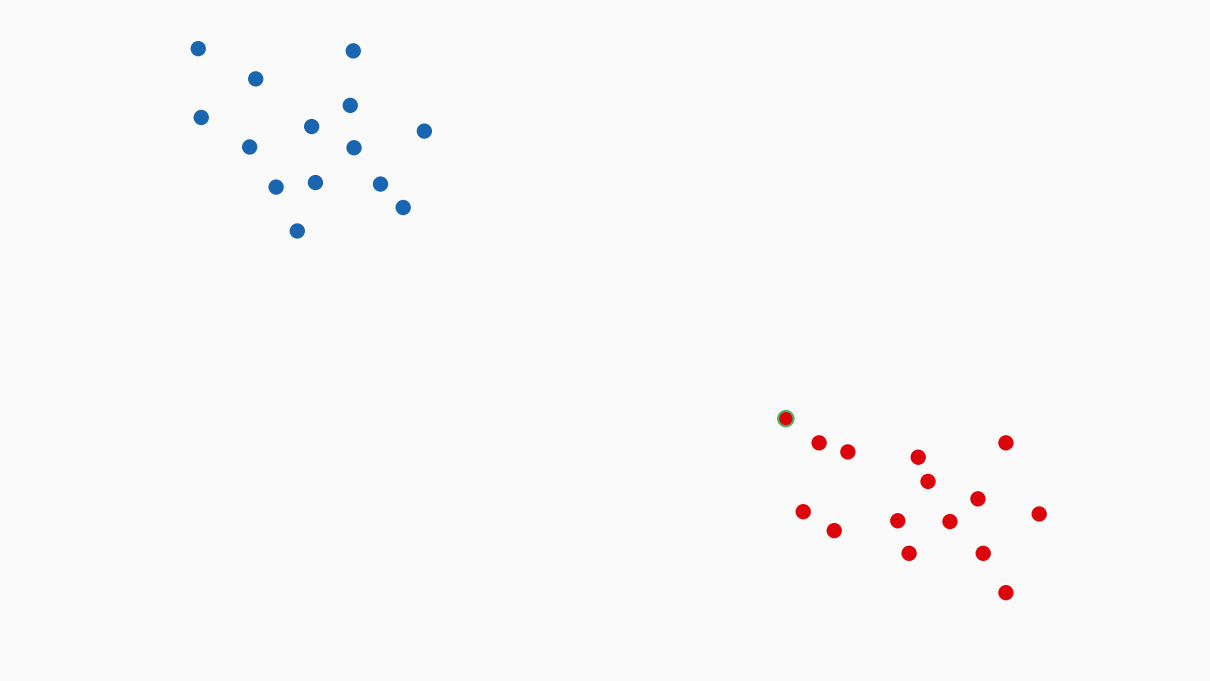
L. Grobol (MoDyCo, Université Paris Nanterre)

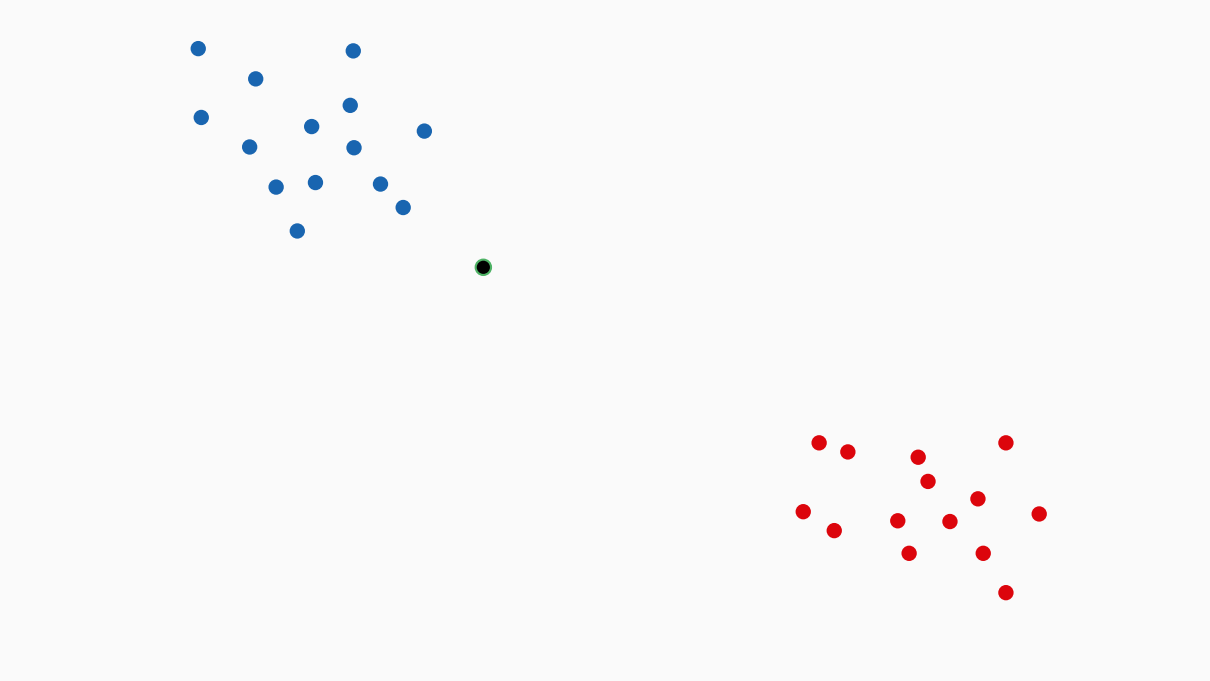
M2 Plurital

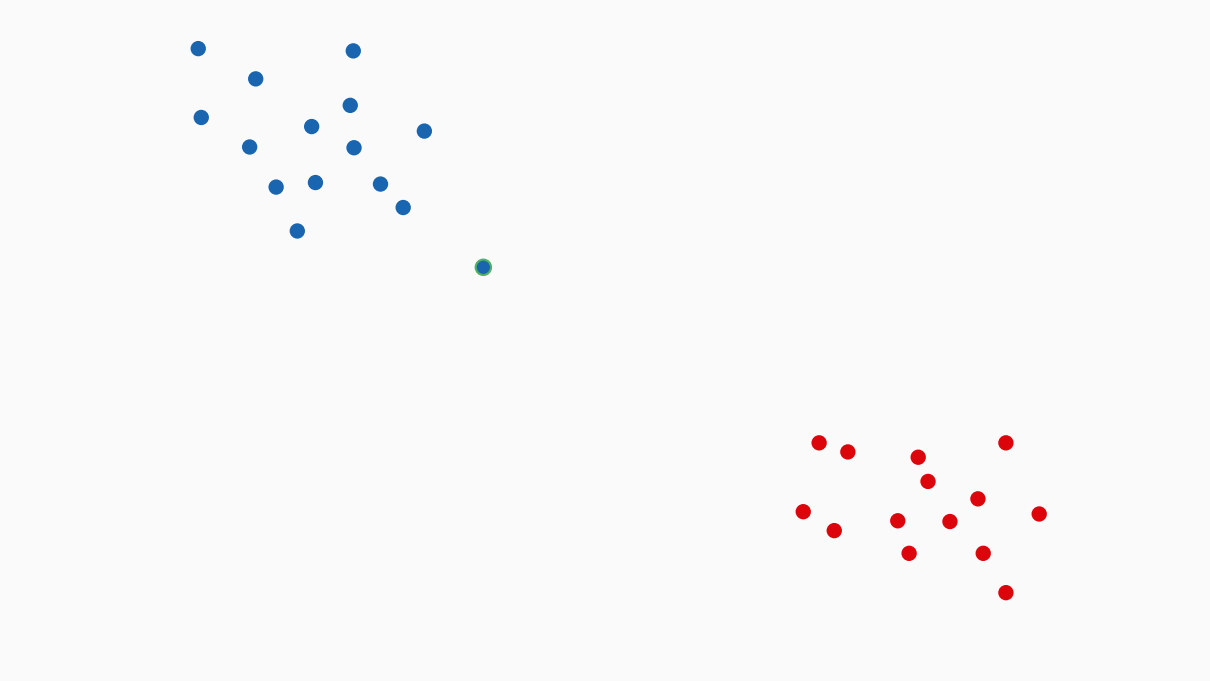
Nanterre, France, 2024-09-24

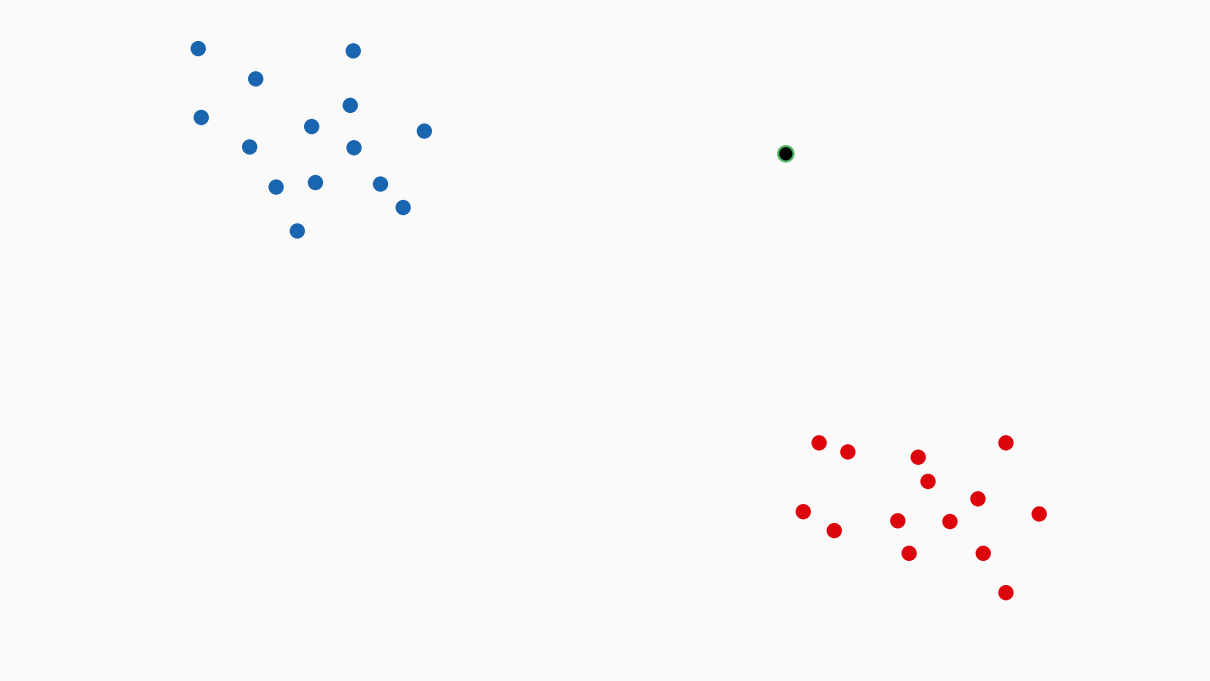


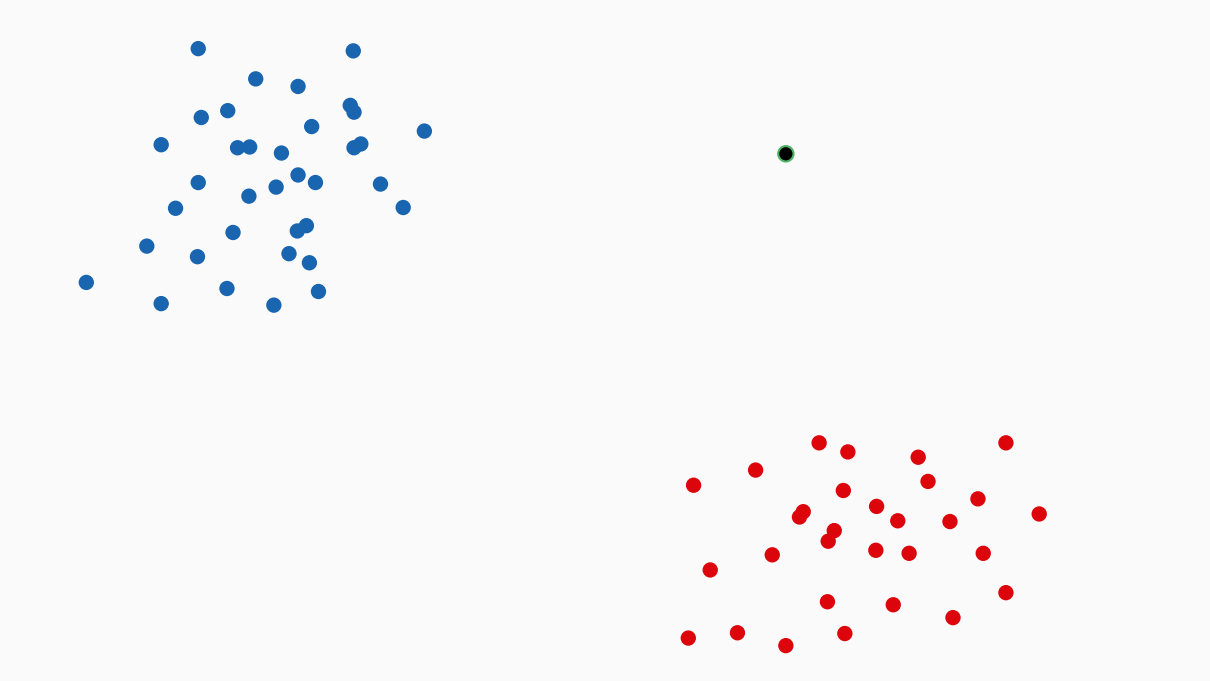


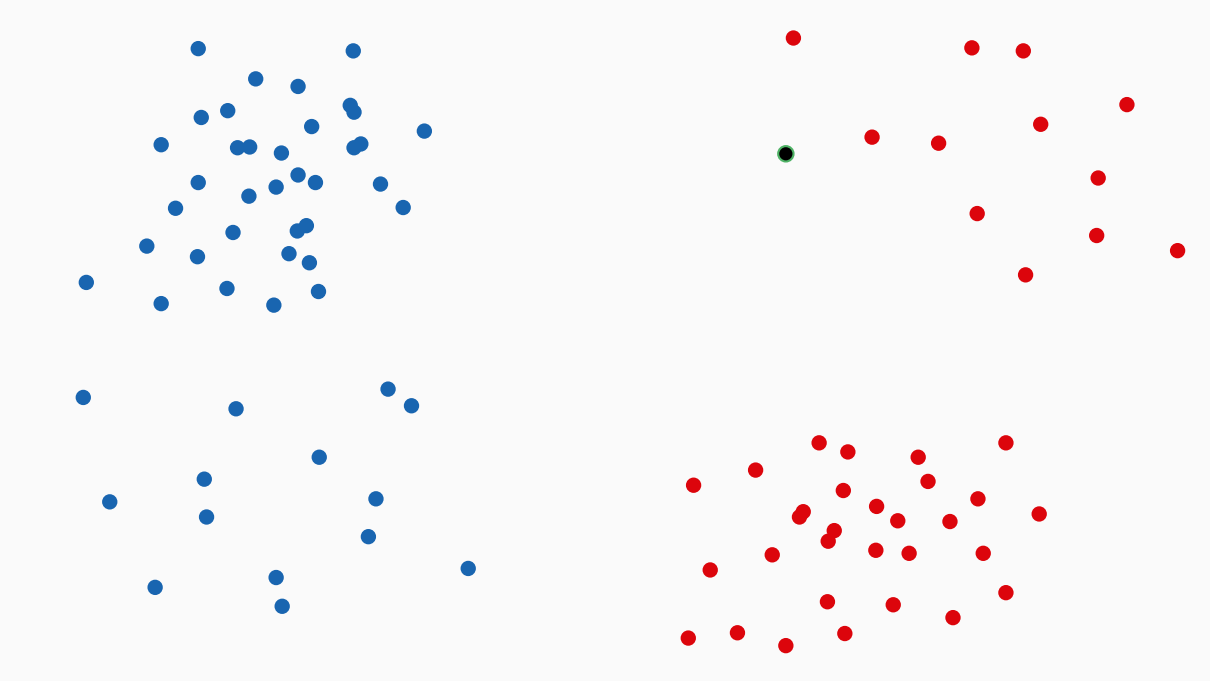


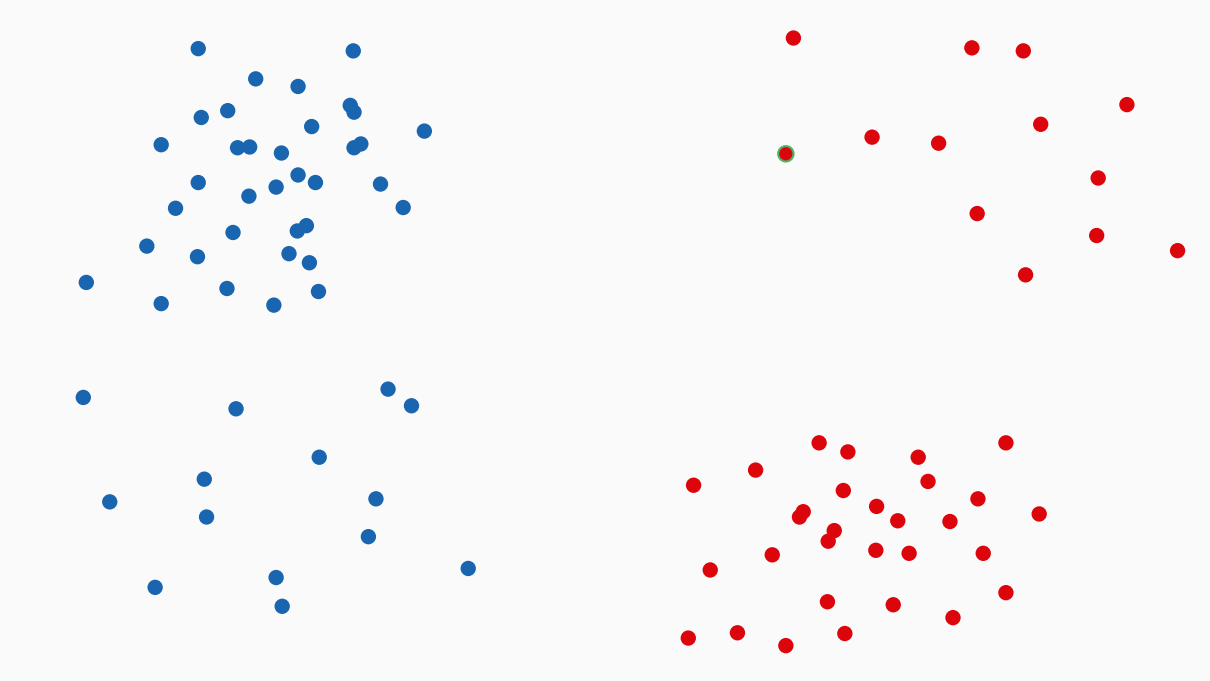












Apprendre

Dans tout ce qui suit, *apprendre* a un sens bien particulier :

Apprendre

Dans tout ce qui suit, *apprendre* a un sens bien particulier :

Identifier des régularités dans des données.

Apprendre

Dans tout ce qui suit, *apprendre* a un sens bien particulier :

Identifier des régularités dans des données.

→ Pour mieux les comprendre ou en générer des similaires

Apprendre

Dans tout ce qui suit, *apprendre* a un sens bien particulier :

Identifier des régularités dans des données.

→ Pour mieux les comprendre ou en générer des similaires

→ Pour inférer les propriétés d'objets inconnus

Apprendre

Dans tout ce qui suit, *apprendre* a un sens bien particulier :

Identifier des régularités dans des données.

- → Pour mieux les comprendre ou en générer des similaires : non-supervisé
- → Pour inférer les propriétés d'objets inconnus

Apprendre

Dans tout ce qui suit, *apprendre* a un sens bien particulier :

Identifier des régularités dans des données.

- → Pour mieux les comprendre ou en générer des similaires : non-supervisé
- → Pour inférer les propriétés d'objets inconnus : supervisé

Anthropomorphisme

La question de l'apprentissage chez les êtres vivants est encore très ouverte, mais il y a de bonnes raisons de penser que ce qu'on fait en apprentissage automatique n'en est pas un modèle raisonnable.

Anthropomorphisme

La question de l'apprentissage chez les êtres vivants est encore très ouverte, mais il y a de bonnes raisons de penser que ce qu'on fait en apprentissage automatique n'en est pas un modèle raisonnable.

C'est très clair en TAL, où nos modèles n'apprennent ni les mêmes choses ni de la même façon que les locuteurices humain·es.

Anthropomorphisme

La question de l'apprentissage chez les êtres vivants est encore très ouverte, mais il y a de bonnes raisons de penser que ce qu'on fait en apprentissage automatique n'en est pas un modèle raisonnable.

C'est très clair en TAL, où nos modèles n'apprennent ni les mêmes choses ni de la même façon que les locuteurices humain·es.

En revanche ça peut fournir des approximations intéressantes des conditions d'accès aux données.

Anthropomorphisme

On va quand même faire un certain nombre d'abus de langage parce que c'est pratique et qu'on y est assez prédisposé·es

Anthropomorphisme

On va quand même faire un certain nombre d'abus de langage parce que c'est pratique et qu'on y est assez prédisposé·es

Mais c'est important de garder dans un coin de son esprit que ce sont effectivement des abus de langage.

Toutefois : en soi, ce qu’on fait ici peut être implémenté par des humain·es.

Toutefois : en soi, ce qu'on fait ici peut être implémenté par des humain·es.

Mais en pratique, on le fait plutôt faire par des machines :

- Les humain·es sont mal équipées pour repérer des patterns dans des données trop vastes.

Toutefois : en soi, ce qu'on fait ici peut être implémenté par des humain·es.

Mais en pratique, on le fait plutôt faire par des machines :

- Les humain·es sont mal équipées pour repérer des patterns dans des données trop vastes.
- Les humain·es et les algorithmes d'apprentissage ont des biais inductifs très différents

Toutefois : en soi, ce qu'on fait ici peut être implémenté par des humain·es.

Mais en pratique, on le fait plutôt faire par des machines :

- Les humain·es sont mal équipées pour repérer des patterns dans des données trop vastes.
- Les humain·es et les algorithmes d'apprentissage ont des biais inductifs très différents
  → Espèrément complémentaires.

En linguistique

Non-supervisé :

- Identifier des classes de mots ayant des propriétés communes (morpho, syntaxe...)
- Formuler des hypothèses phylogénétiques/typologiques sur les langues (un peu dangereux)
- ...

Supervisé :

- Modéliser des phénomènes linguistiques pour l'annotation automatique de corpus.
- Modèles cognitifs computationnels, modèles de langues (normes)
- ...

Données

Fondamentalement, l'apprentissage automatique, c'est donc une question de données.

Données

Fondamentalement, l'apprentissage automatique, c'est donc une question de données.

Les résultats, les applications vont dépendre primairement de la quantité de données à laquelle on a accès et de leur distribution.

Données

Fondamentalement, l'apprentissage automatique, c'est donc une question de données.

Les résultats, les applications vont dépendre primairement de la quantité de données à laquelle on a accès et de leur distribution.

C'est une approximation mathématique qui plaît bien en informatique : on suppose qu'il existe un ensemble de données, on a accès à une partie, un échantillon de cet ensemble dont on se sert comme une approximation du tout.

Évidemment, pour qu'on puisse repérer dans cet échantillon des régularités intéressantes, il faut qu'il soit suffisamment couvrant pour que ces régularités soient observables.

Données

Évidemment, pour qu'on puisse repérer dans cet échantillon des régularités intéressantes, il faut qu'il soit suffisamment couvrant pour que ces régularités soient observables.

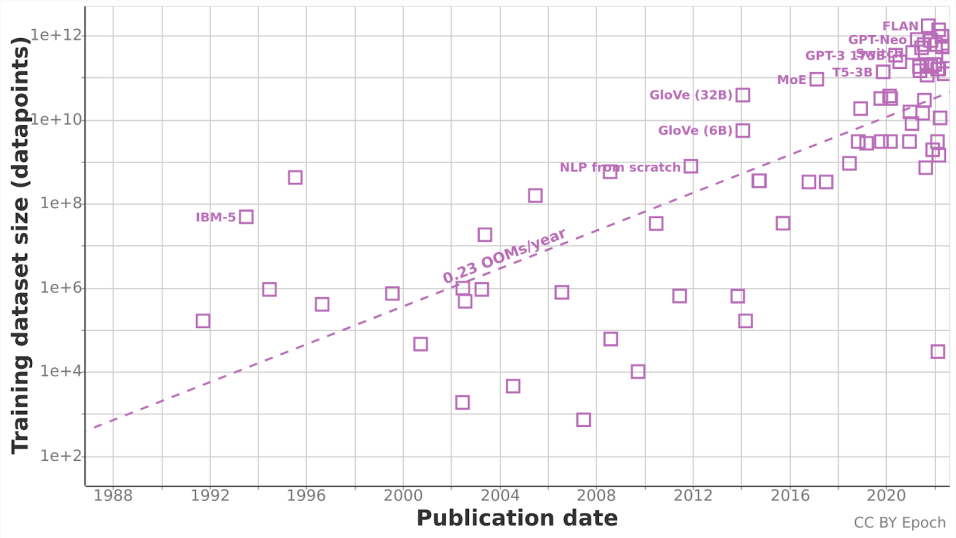
Une stratégie pour ça, c'est de collecter un échantillon le plus grand possible en espérant qu'il va couvrir toute la diversité de données.

Données

Évidemment, pour qu'on puisse repérer dans cet échantillon des régularités intéressantes, il faut qu'il soit suffisamment couvrant pour que ces régularités soient observables.

Une stratégie pour ça, c'est de collecter un échantillon le plus grand possible en espérant qu'il va couvrir toute la diversité de données.

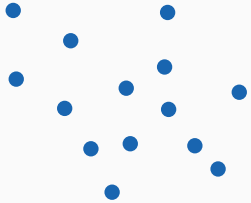
C'est *possible*, mais pas forcément idéal en pratique : en linguistique par exemple, le lexique et les constructions ne sont pas uniformément présents, et on court le risque de noyer des points de données rares mais intéressants.

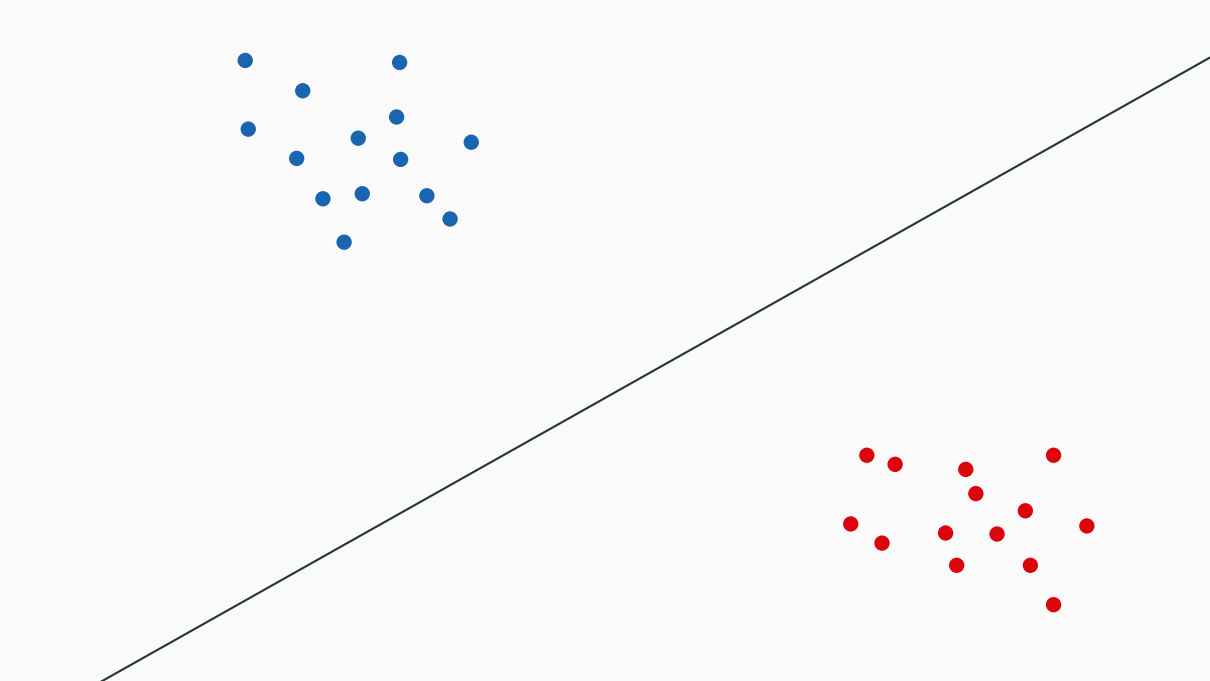


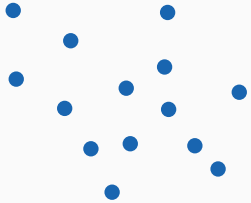
(Villalobos 2022)

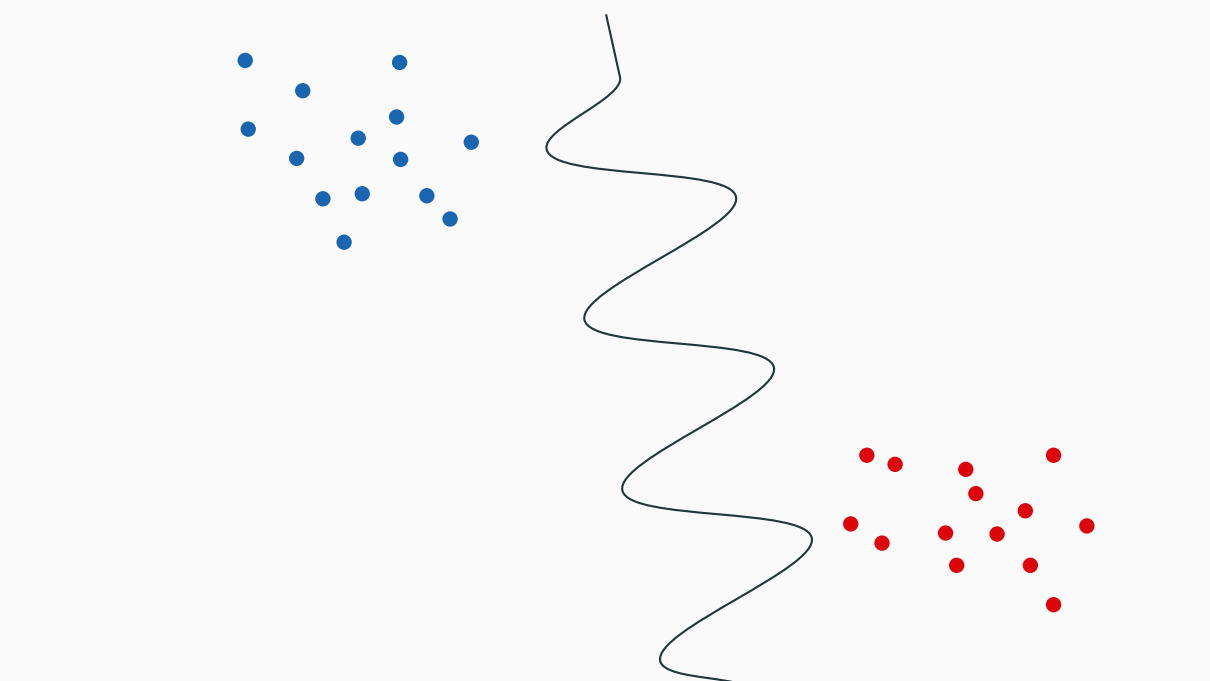
Modèles

Même dans le cas les données seraient générées par une règle déterministe, il n'est pas en général possible de la retrouver avec certitude, comme on ne dispose que d'un échantillon.







Modèles

Même dans le cas les données seraient générées par une règle déterministe, il n'est pas en général possible de la retrouver avec certitude, comme on ne dispose que d'un échantillon.

Modèles

Même dans le cas les données seraient générées par une règle déterministe, il n'est pas en général possible de la retrouver avec certitude, comme on ne dispose que d'un échantillon.

Mais on peut construire différents modèles des données avec différentes propriétés.

Modèles

Même dans le cas les données seraient générées par une règle déterministe, il n'est pas en général possible de la retrouver avec certitude, comme on ne dispose que d'un échantillon.

Mais on peut construire différents modèles des données avec différentes propriétés.

Le choix d'un modèle est un arbitrage (« *there is no free lunch* ») qui dépend des exigences des applications.

Des idées de propriétés désirables ?

Propriétés des modèles

- Capacités de mémorisation.

Propriétés des modèles

- Capacités de mémorisation.
- Capacités d'extrapolation.

Propriétés des modèles

- Capacités de mémorisation.
- Capacités d'extrapolation.
- Coût computationnel :

Propriétés des modèles

- Capacités de mémorisation.
- Capacités d'extrapolation.
- Coût computationnel :
  - → Pour trouver le modèle.

Propriétés des modèles

- Capacités de mémorisation.
- Capacités d'extrapolation.
- Coût computationnel :
  - → Pour trouver le modèle.
  - → Pour utiliser le modèle une fois trouvé.

Propriétés des modèles

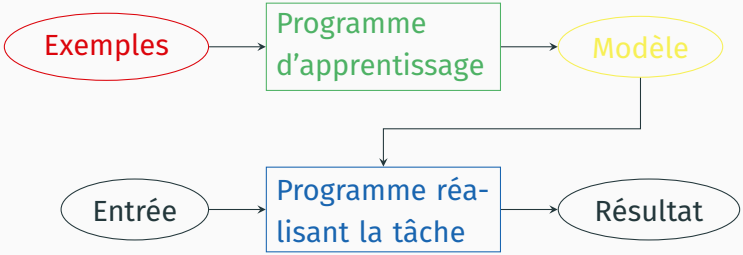
- Capacités de mémorisation.
- Capacités d'extrapolation.
- Coût computationnel :
  - → Pour trouver le modèle.
  - → Pour utiliser le modèle une fois trouvé.
- N'importe quelle autre propriété spécifique à l'application.

Propriétés des modèles

- Capacités de mémorisation.
- Capacités d'extrapolation.
- Coût computationnel :
  - → Pour trouver le modèle.
  - → Pour utiliser le modèle une fois trouvé.
- N'importe quelle autre propriété spécifique à l'application.
- La simplicité

Propriétés des modèles

- Capacités de mémorisation.
- Capacités d'extrapolation.
- Coût computationnel :
  - → Pour trouver le modèle.
  - → Pour utiliser le modèle une fois trouvé.
- N'importe quelle autre propriété spécifique à l'application.
- La simplicité (rasoir d'Occam).


Exemples
Programme d'apprentissage
Modèle
Entrée
Programme réalisant la tâche
Résultat

Paramètres et hyperparamètres

Un modèle généré, « appris », par un algorithme d'apprentissage peut être vu comme un ensemble de règles.

Ces règles (comme tout ce que manipule un ordinateur) sont matérialisées par des valeurs logiques ou numériques qu'on appelle paramètres. Les paramètres d'un modèle sont une représentation (d'une approximation) des données d'entraînement et sont obtenues *a posteriori* dans le processus d'apprentissage.

Paramètres et hyperparamètres

Un modèle généré, « appris », par un algorithme d'apprentissage peut être vu comme un ensemble de règles.

Ces règles (comme tout ce que manipule un ordinateur) sont matérialisées par des valeurs logiques ou numériques qu'on appelle paramètres. Les paramètres d'un modèle sont une représentation (d'une approximation) des données d'entraînement et sont obtenues *a posteriori* dans le processus d'apprentissage.

Les choix *a priori* de l'utilisateurice d'un algorithme d'apprentissage (taille, architecture, variante particulière de l'algorithme, transformations des données…), qui influencent le modèle final, mais ne sont pas *appris*, sont appelés hyperparamètres.

Paramètres et hyperparamètres

Un modèle généré, « appris », par un algorithme d'apprentissage peut être vu comme un ensemble de règles.

Ces règles (comme tout ce que manipule un ordinateur) sont matérialisées par des valeurs logiques ou numériques qu'on appelle paramètres. Les paramètres d'un modèle sont une représentation (d'une approximation) des données d'entraînement et sont obtenues *a posteriori* dans le processus d'apprentissage.

Les choix *a priori* de l'utilisateurice d'un algorithme d'apprentissage (taille, architecture, variante particulière de l'algorithme, transformations des données…), qui influencent le modèle final, mais ne sont pas *appris*, sont appelés hyperparamètres.

Là encore il est très rare qu'on échappe à des arbitrages : « *there is no free lunch* », on ne rase pas gratis.

Taille des modèles

Comme pour les données, au-delà de leur expressivité, la taille des modèles (en termes de paramètres) est une question omniprésente.

Taille des modèles

Comme pour les données, au-delà de leur expressivité, la taille des modèles (en termes de paramètres) est une question omniprésente.

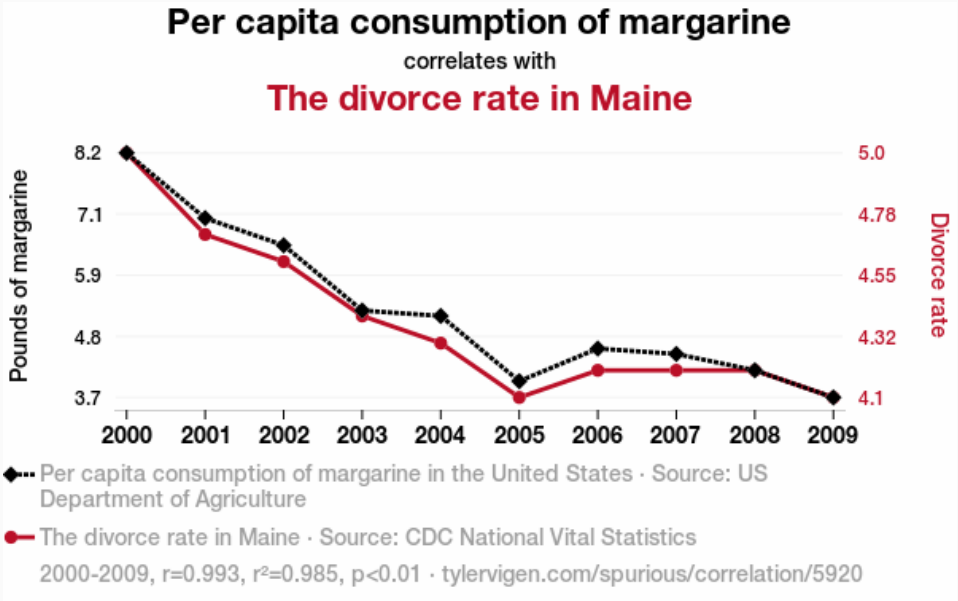
En général : un modèle avec plus de paramètres peut stocker plus d'information, donc mémoriser des régularités plus complexes, ce que ne peut pas forcément faire un modèle trop petit (qui sous-apprendrait).

Taille des modèles

Comme pour les données, au-delà de leur expressivité, la taille des modèles (en termes de paramètres) est une question omniprésente.

En général : un modèle avec plus de paramètres peut stocker plus d'information, donc mémoriser des régularités plus complexes, ce que ne peut pas forcément faire un modèle trop petit (qui sous-apprendrait).

Mais un modèle plus gros risque aussi de mémoriser des corrélations fallacieuses ou des biais dans les données au lieu de régularités intéressantes.


Per capita consumption of margarine
correlates with
The divorce rate in Maine
Pounds of margarine
8.2
7.1
5.9
4.8
3.7
Divorce rate
5.0
4.78
4.55
4.32
4.1
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
Per capita consumption of margarine in the United States · Source: US Department of Agriculture
The divorce rate in Maine · Source: CDC National Vital Statistics
2000-2009, r=0.993, r²=0.985, p<0.01 · tylervigen.com/spurious/correlation/5920

(Vigen 2024)

Taille des modèles

Comme pour les données, au-delà de leur expressivité, la taille des modèles (en termes de paramètres) est une question omniprésente.

En général : un modèle avec plus de paramètres peut stocker plus d'information, donc mémoriser des régularités plus complexes, ce que ne peut pas forcément faire un modèle trop petit (qui sous-apprendrait).

Mais un modèle plus gros risque aussi de mémoriser des corrélations fallacieuses ou des biais dans les données au lieu de régularités intéressantes.

Taille des modèles

Comme pour les données, au-delà de leur expressivité, la taille des modèles (en termes de paramètres) est une question omniprésente.

En général : un modèle avec plus de paramètres peut stocker plus d'information, donc mémoriser des régularités plus complexes, ce que ne peut pas forcément faire un modèle trop petit (qui sous-apprendrait).

Mais un modèle plus gros risque aussi de mémoriser des corrélations fallacieuses ou des biais dans les données au lieu de régularités intéressantes. On parle de sur-apprentissage.

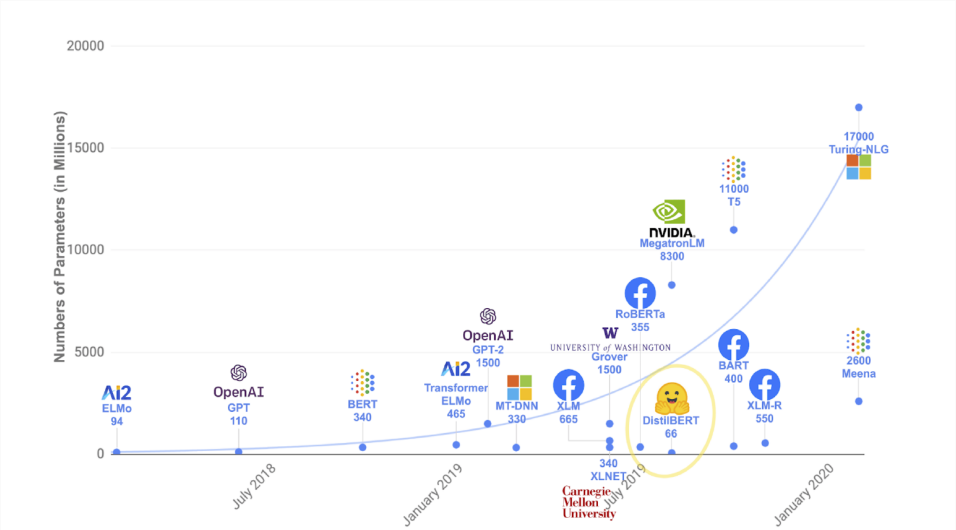
Taille des modèles

Comme pour les données, au-delà de leur expressivité, la taille des modèles (en termes de paramètres) est une question omniprésente.

En général : un modèle avec plus de paramètres peut stocker plus d'information, donc mémoriser des régularités plus complexes, ce que ne peut pas forcément faire un modèle trop petit (qui sous-apprendrait).

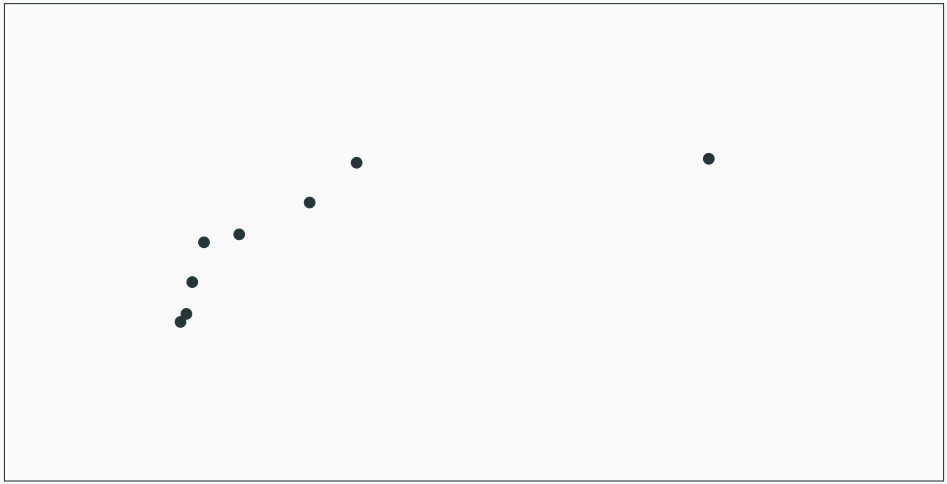
Mais un modèle plus gros risque aussi de mémoriser des corrélations fallacieuses ou des biais dans les données au lieu de régularités intéressantes. On parle de sur-apprentissage.

Débattez de ce dernier mot (3 rounds)



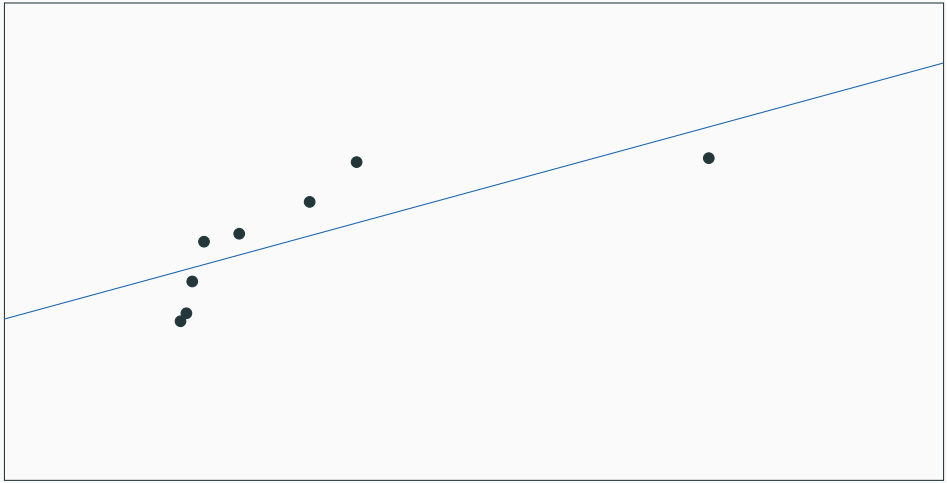
(Mehta et al. 2023)

Population de points

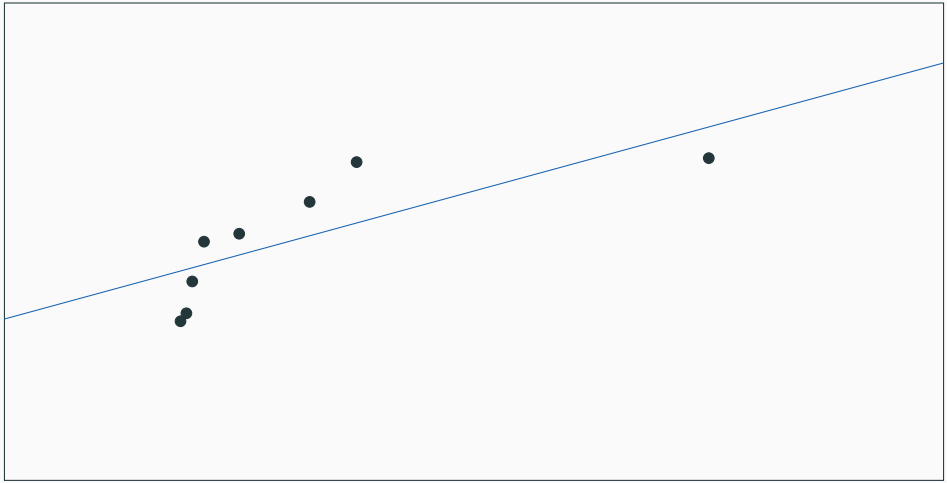


Voir aussi `ggbm.at/TagmZdxN`

Régression linéaire : $y = ax + b$

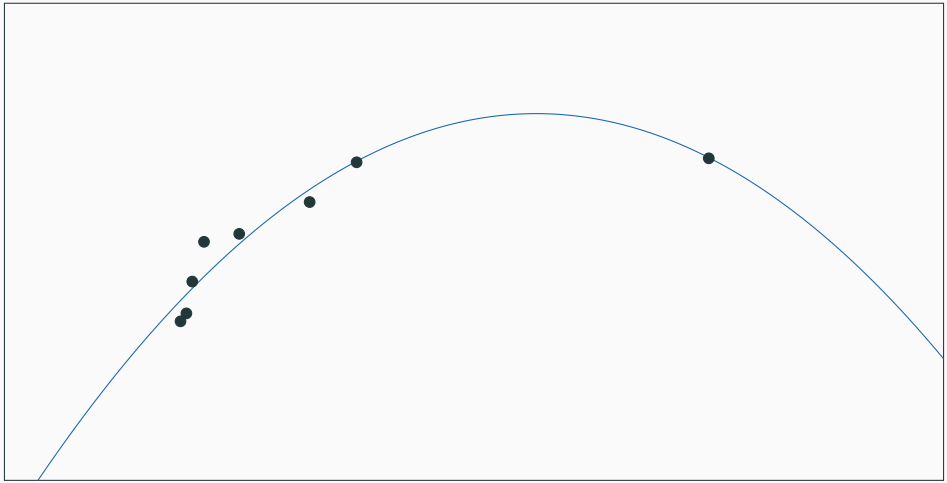


Régression linéaire : $y = ax + b$

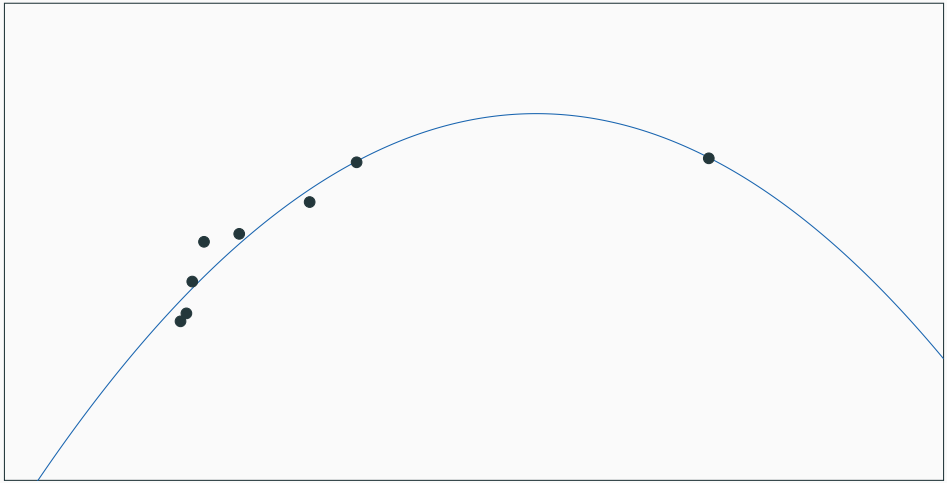


Léger sous-apprentissage (« *underfit* »)

Régression quadratique : $y = ax^2 + bx + c$

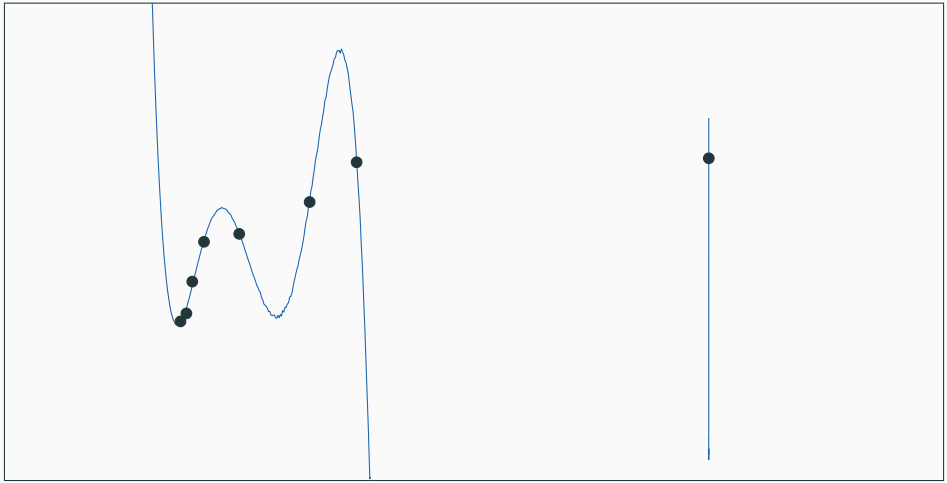


Régression quadratique : $y = ax^2 + bx + c$

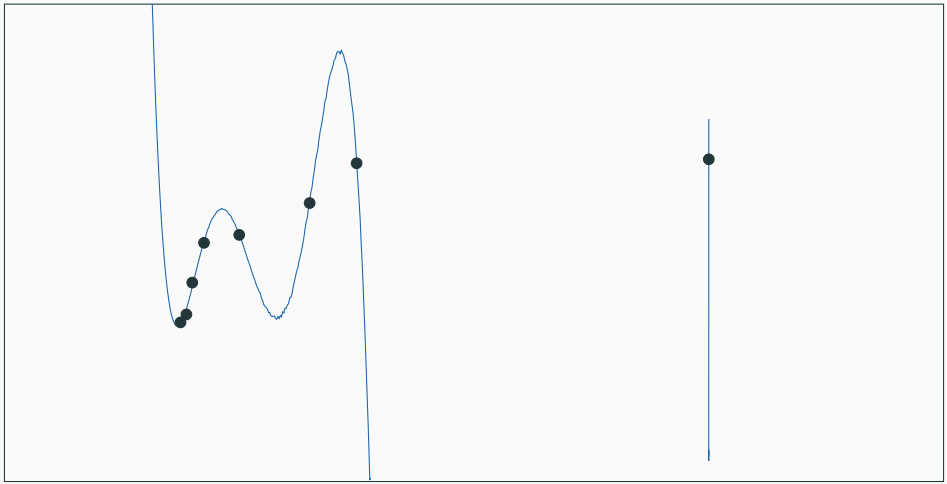


Sur-apprentissage (« *overfit* »)

Polynôme d'interpolation de Lagrange : $y = a_0 + a_1 x + a_2 x^2 + ... + a_6 x^6$



Polynôme d'interpolation de Lagrange : $y = a_0 + a_1 x + a_2 x^2 + \ldots + a_6 x^6$



Sur-apprentissage catastrophique

Appendix

References i

Mehta, Sandeep, Darpan Shah, Ravindra Kulkarni et Cornelia Caragea (24 avr. 2023). ***Semantic Tokenizer for Enhanced Natural Language Processing.*** arXiv : 2304.12404 [cs].
URL : http://arxiv.org/abs/2304.12404 (visité le 03/06/2024). preprint.

Vigen, Tyler (2024). ***Per Capita Consumption of Margarine Correlates with The Divorce Rate in Maine (R=0.993).***
URL : https://tylervigen.com/spurious/correlation/5920_per-capita-consumption-of-margarine_correlates-with_the-divorce-rate-in-maine (visité le 03/06/2024).

Villalobos, Pablo (20 sept. 2022). ***Trends in Training Dataset Sizes.*** Epoch AI.
URL : https://epochai.org/blog/trends-in-training-dataset-sizes (visité le 03/06/2024).

Licence