

Dr Aude Sportisse
Chercheuse dans l'équipe Maasai, Inria Sophia Antipolis
Enseignante à EFELIA

aude.sportisse@inria.fr

Introduction à l'Intelligence Artificielle

Appliquée à la biologie

Septembre-Décembre 2023

Déroulement de l'UE

- 16h de cours en 4 séances: **28/09, 5/10, 19/10, 9/11 de 8h à 12h**
 - Introduction
 - Apprentissage statistique
 - TD spécifique: IA appliquée à la biologie avec **Marc Bailly-Bechet**
 - Limites de l'IA et ouvertures
- Déroulé des séances focus IA:
 - ~1h20 de cours + ~10min de pause
 - ~1h20 de cours + ~10min de pause
 - ~1h de TP (2 choses: **apporter son ordinateur + avoir accès à internet**, pas besoin de télécharger de logiciel de code)
 - Dernière séance: un temps pour exposés+discussion

Evaluation

- 2 QCM (20%)
- Exposé (80%) de ~5 minutes sur un des thèmes suivants: présentation + résumé
 - **Génomique et IA:** lecture de l'ADN n'a plus de secrets pour les algorithmes, Jean-Philippe Vert, La Recherche, 2019
 - **Images médicales et IA:** *Deep Clustering for Abdominal Organ Classification in Ultrasounds Imaging*, Hind Dadoun, Hervé Delingette, Anne-Laure Rousseau, Eric de Kerviler et Nicholas Ayache, 2023 (En anglais, difficile, article de recherche)
 - **Prédiction structurelle de protéine et IA:** *Highly accurate protein structure prediction with AlphaFold*, John Jumper et al., 2021 (En anglais, difficile, article de recherche)
 - **Protection des données en médecine:** *Machine learning and genomics: precision medicine vs. patient privacy*, Chloe-Agathe Azencott (En anglais, difficile)
 - **Limites de l'IA pour la démonstration de théorèmes:** Quand les machines font des mathématiques, Jean-Paul Delaye, Pour la Science, 2020
 - **Légitimité de l'IA:** lecture de *When is automated decision making legitimate?* (En anglais), Solon Barocas, Moritz Hardt, Arvind Narayanan, novembre 2022
 - **ChatGPT:** lecture de *Les élucubrations mathématiques de ChatGPT*, Jean-François Colonna, Pour la Science, juillet 2023
 - Présenter un **algorithme** parmi, me voir pour références: algorithme du gradient stochastique (SGD, algorithme central en apprentissage statistique) ou k-moyennes

Objectifs du cours

- Expliquer les concepts et les méthodes clés de l'IA,
- Appréhender les capacités, limites et enjeux des nouveaux systèmes d'IA,
- Débuter une intégration des progrès apportés par l'IA dans l'exploration des problématiques de recherche liées aux sciences de la vie et aux neurosciences.

Ce cours ne permet pas de devenir spécialiste en IA, mais peut permettre d'être plus à l'aise au cours de discussions entre experts (biologistes) et chercheurs en IA, pendant sa thèse par exemple.

Programme du jour

1. Historique de l'IA
2. Définition de l'IA (?)
3. Applications récentes en biologie
4. Introduction à l'apprentissage statistique
5. Zoom sur les modèles linéaires
6. TP: modèles linéaires



1. Historique de l'IA

Le rêve d'une machine intelligente



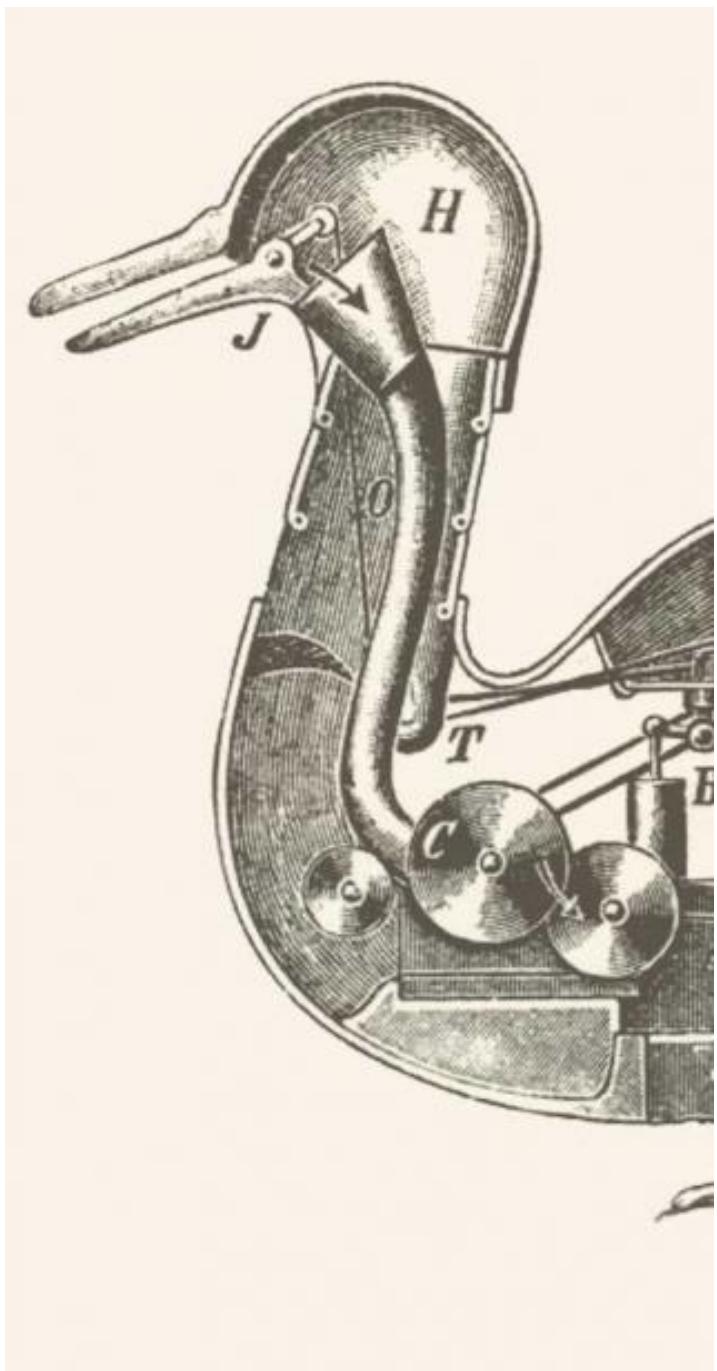
Héphaïstos remet à Thétis les armes forgées pour Achille vers IV av J.-C.; fresque du Ier siècle ap J.-C.

Dès l'Antiquité, le rêve d'une machine intelligente (avec intervention divine)

- **Mythologie juive:** le Golem est un automate à forme humaine en bois ou argile, un rabbin inscrit un mot magique sur son front pour qu'il le serve.
- **Grèce antique (IVe siècle av J.-C.), l'Iliade:** le Dieu du feu Héphaïstos construit des tables à trois pieds pour le servir; les portes de l'Olympe s'ouvrent d'elles-mêmes

Systématiser la pensée, automates

- **XVIIe siècle:** Descartes introduit l'idée de l'animal machine, qui aurait certaines activités humaines, mais pas toutes l'intelligence.
- **XVIIIe siècle:** Vaucanson invente des automates: joueurs de flûte, canard
1747: publication de « L'homme machine » par Julien Offray de La Mettrie



Canard de Vaucanson

Inventeur et mécanicien français, Jacques Vaucanson (1709-1782) est célèbre pour ses automates qui reproduisent les êtres vivants.

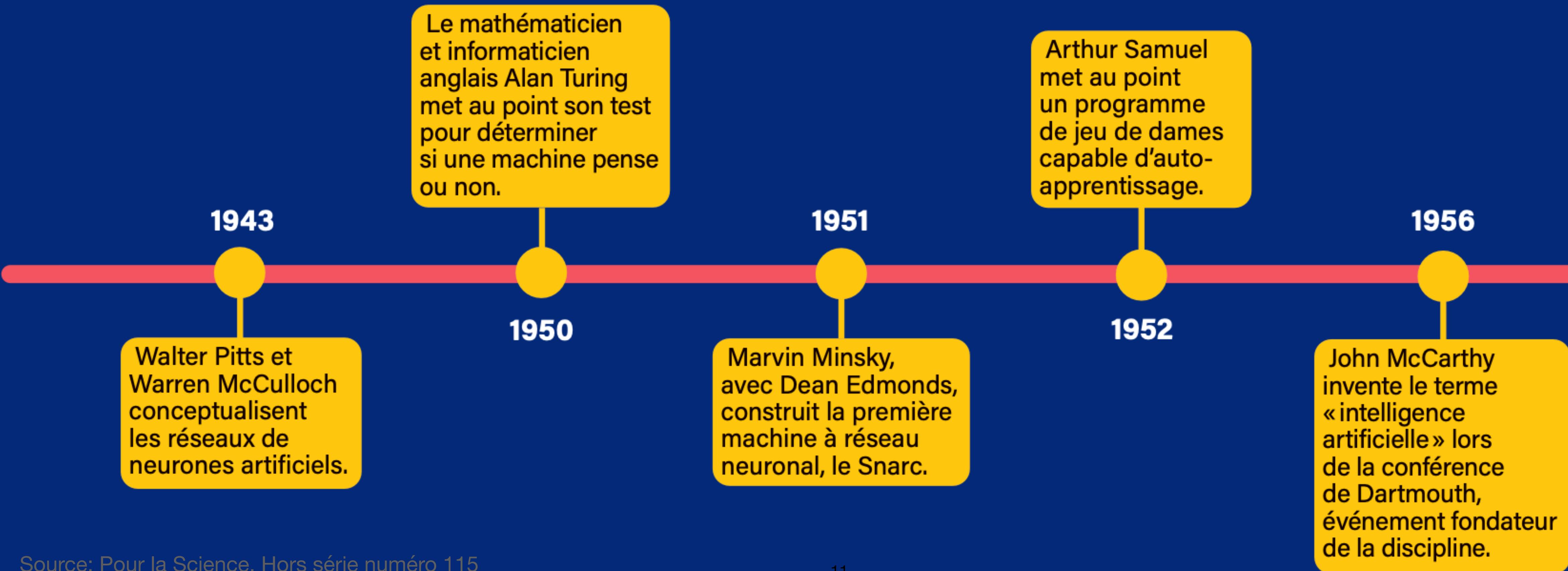
En 1738, il présente un "**Joueur de flûte**" de **1,50 m qui joue de la musique comme un humain**, avec des mouvements de lèvres, de doigts, et le contrôle du souffle.

Plus sophistiqué, le "**Canard digérateur**" bouge les ailes, cancane, ingère des graines et les "digère" aussi, avec un réalisme caractéristique du courant pédagogique de l'époque. Le mécanisme, placé dans l'imposant piédestal, est visible par tous, dans le but de montrer la complexité du travail accompli. La **digestion relève de l'exploit** : le canard semble rendre ce qu'il a avalé après une véritable digestion ! Au-delà de l'effet de salon, la **conception d'automates marque une étape importante dans la mécanisation**, qui ouvre des perspectives pour l'industrie naissante au XVIII^e siècle.

Source: Bibliothèque Nationale de France, <https://gallica.bnf.fr/essentiels/evenement/automates-vaucanson>

« Une histoire à rebondissement »

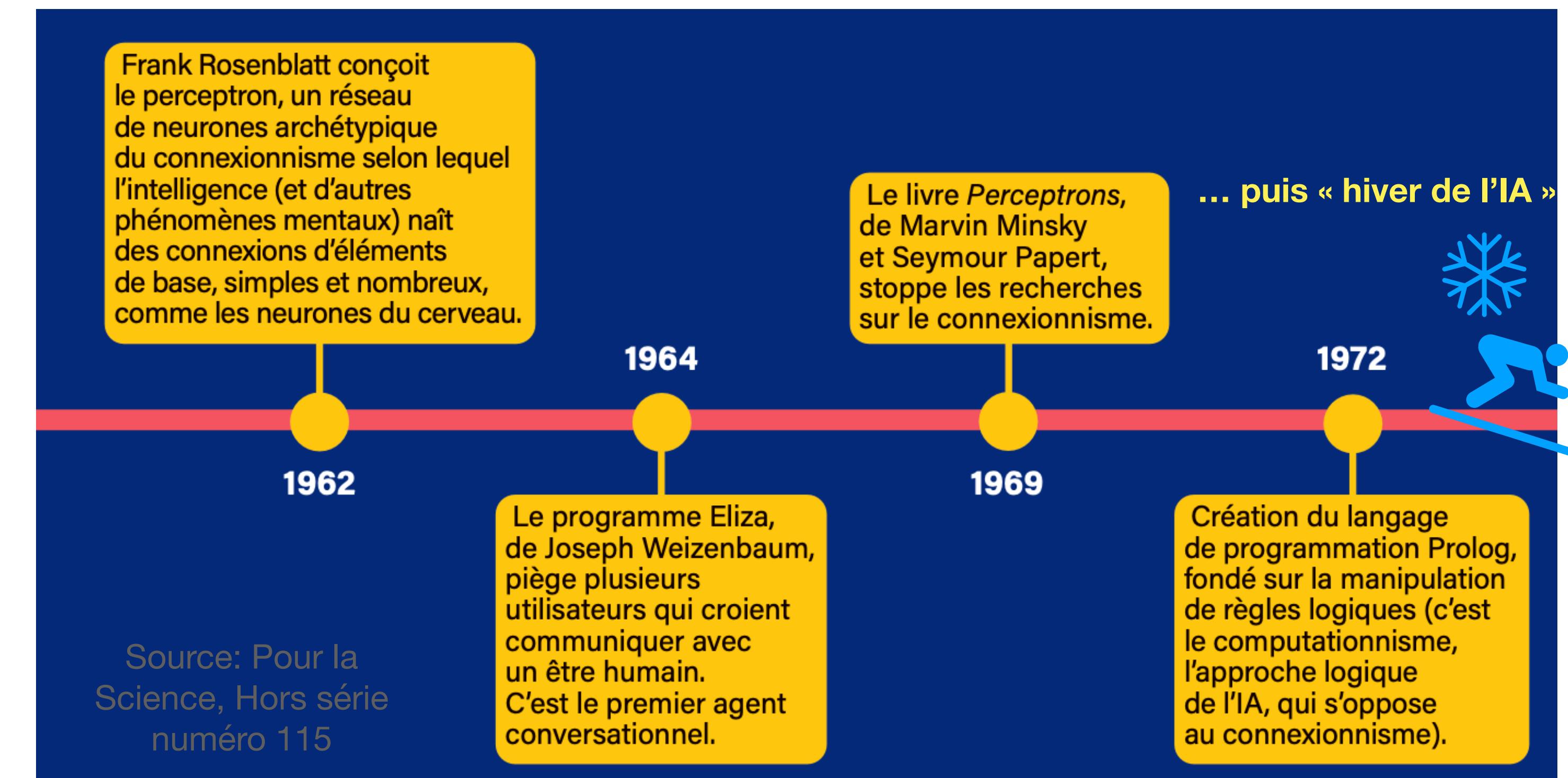
- 1943-1955: Premiers ordinateurs (en 43), Premiers travaux en IA
- 1956: Fondements de l'IA et invention du terme lors d'une longue conférence (2 mois)...



« Une histoire à rebondissement »

Premiers programmes de l'IA; des travaux sur les réseaux de neurones

- **1954:** Allen Newell a un projet de programme d'échecs, il travaille avec des **psychologues**, notamment Herbert Simon et Cliff Shaw.
- **Logic Theorist (1956):** premier programme de démonstration en Logique des propositions (but: prouver des théorèmes existants en imitant le raisonnement humain).
- Newell et Simon pensent qu'avant 68, un programme sera champion d'échecs et démontrera un important théorème mathématique...
- 1967: Richard Greenblatt crée le premier programme d'échecs capable de battre un joueur normal mais pas le champion du monde...



Deux approches s'opposent... l'IA symbolique et le connexionnisme

Mais les deux approches sont décevantes, l'IA est délaissée jusque dans les années 80 (premier hiver de l'IA)

Deux approches se dégagent (et s'opposaient)

John McCarthy: « *on a besoin de machines qui puissent résoudront des problèmes, pas qui puissent penser comme des gens* »

Quel concept décrit un canard et permet d'exclure les manchots?

- **IA symbolique (approche la plus ancienne, presque délaissée à partir des années 80-90, mais regain?):** automatiser en représentant les situations en **symboles compréhensibles** par nous humains et en les traitant par des algorithmes.

Techniques symboliques appelés GOFAI « Good Old-Fashion Artificial Intelligence »

- **IA sub-symbolique (approche récente, véritable essor dans les années 90):** pas de représentation directement compréhensible, **reproduction de phénomènes naturels** avec des méthodes **statistiques**

- Par exemple: le **connexionnisme**, qui utilise les réseaux de neurones (s'inspire du fonctionnement de notre système nerveux)

	Aplatir	Taille	Envergure	Couleur	Classe
$e_1 =$	VRAI	30	49	Roux	+ (canard)
$e_2 =$	FAUX	70	32	Gris	- (manchot)
$e_3 =$	VRAI	40	46	Orange	+ (canard)
$e_4 =$	FAUX	60	33	Orange	- (manchot)



Source: A. Cornuéjols et L. Millet. Apprentissage artificiel, concepts et algorithmes

Critiques de l'IA symbolique: programme *incapable de comprendre les symboles* qu'il utilise (John Searle, **philosophe**); peu de traitement symbolique dans le raisonnement humain (Hubert Dreyfus, **philosophe**); **besoin comprendre la machinerie physique** du problème avant de pouvoir comprendre des symboles (David Marr; **neuroscientifique** et **physiologiste**)

Pourquoi le premier hiver de l'IA?

- chercheurs **trop optimistes**, plan trop ambitieux pour réaliser un traducteur automatique (5 ans) mais il faudrait **énorme quantité d'information**, impossible à gérer à l'époque -> *annulation des financements aux USA (1966)*
- algorithmes non adaptés aux problèmes de grande taille: **manque de mémoire, pas assez de puissance de calcul** -> *rapport Lighthill (1973)* et arrêt des financements en UK.
- Livre *Perceptrons*, de Minsky et Papert: réseaux de neurones incapables de calculer des fonctions très simples, aucun réseau de neurones capable de distinguer deux nombres écrits en binaires.

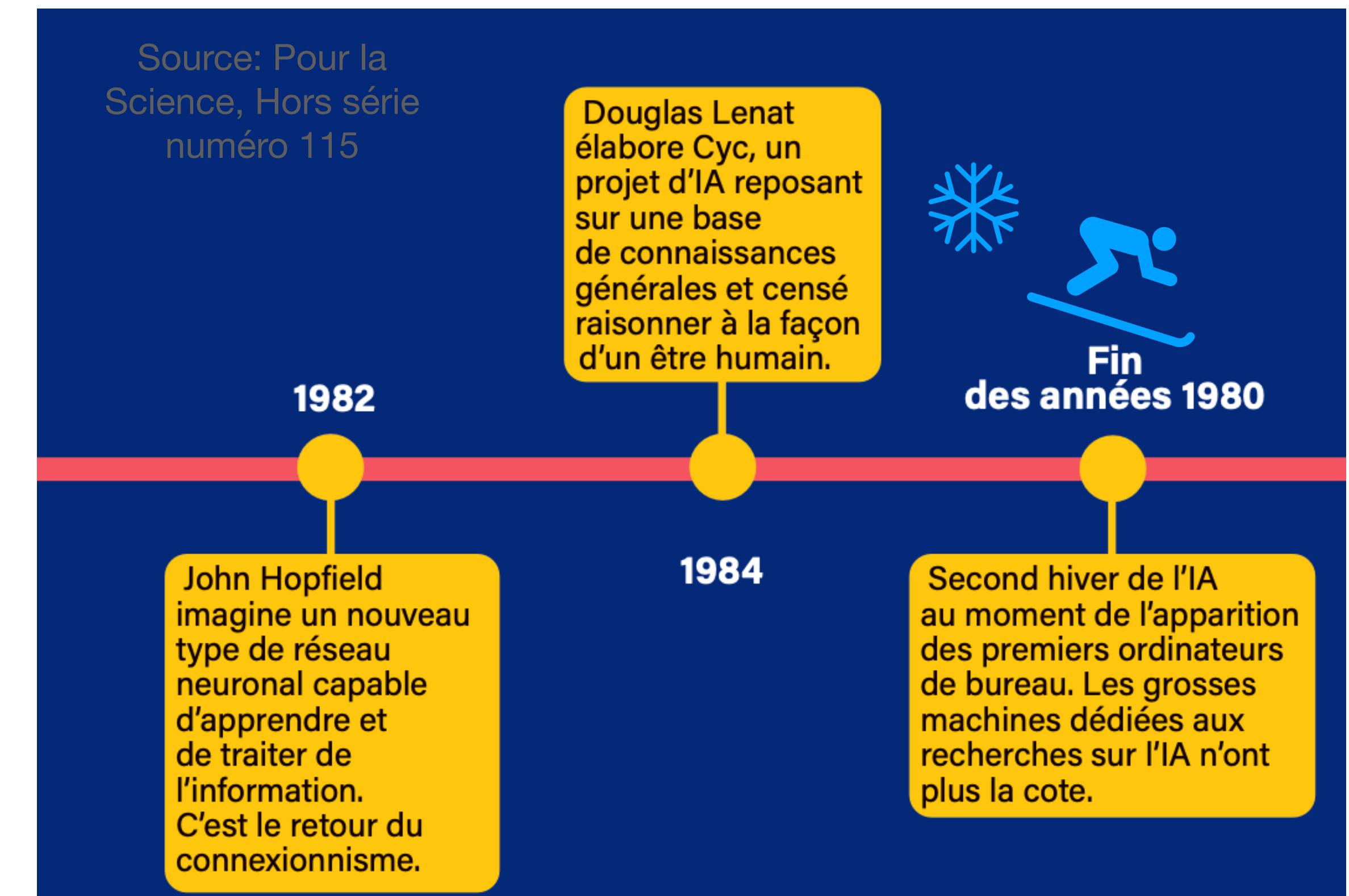


Un homme ajuste le réseau de câblage aléatoire entre les capteurs de lumière et l'unité d'association du Perceptron du scientifique Frank Rosenblatt (1962)

« Une histoire à rebondissement »

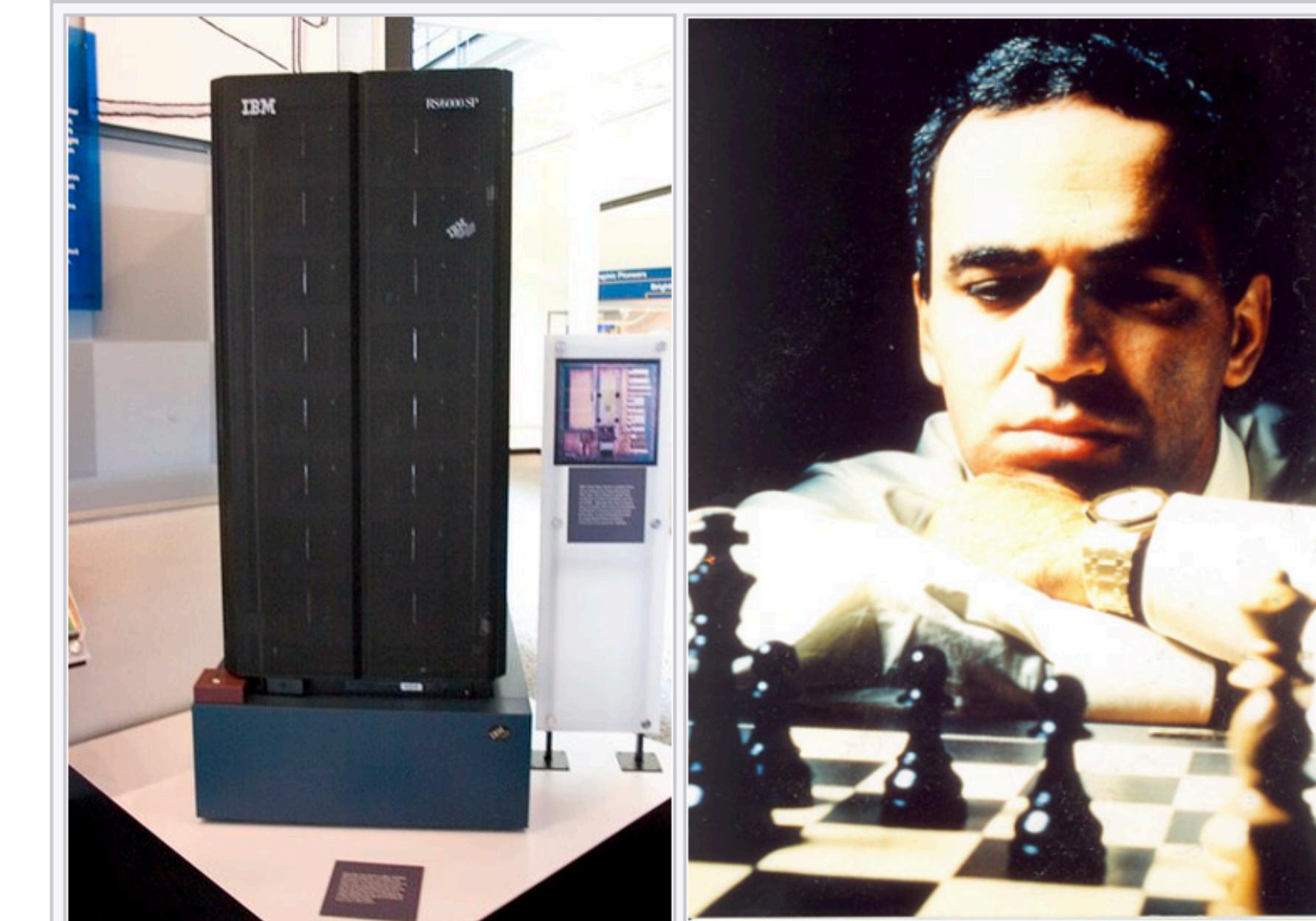
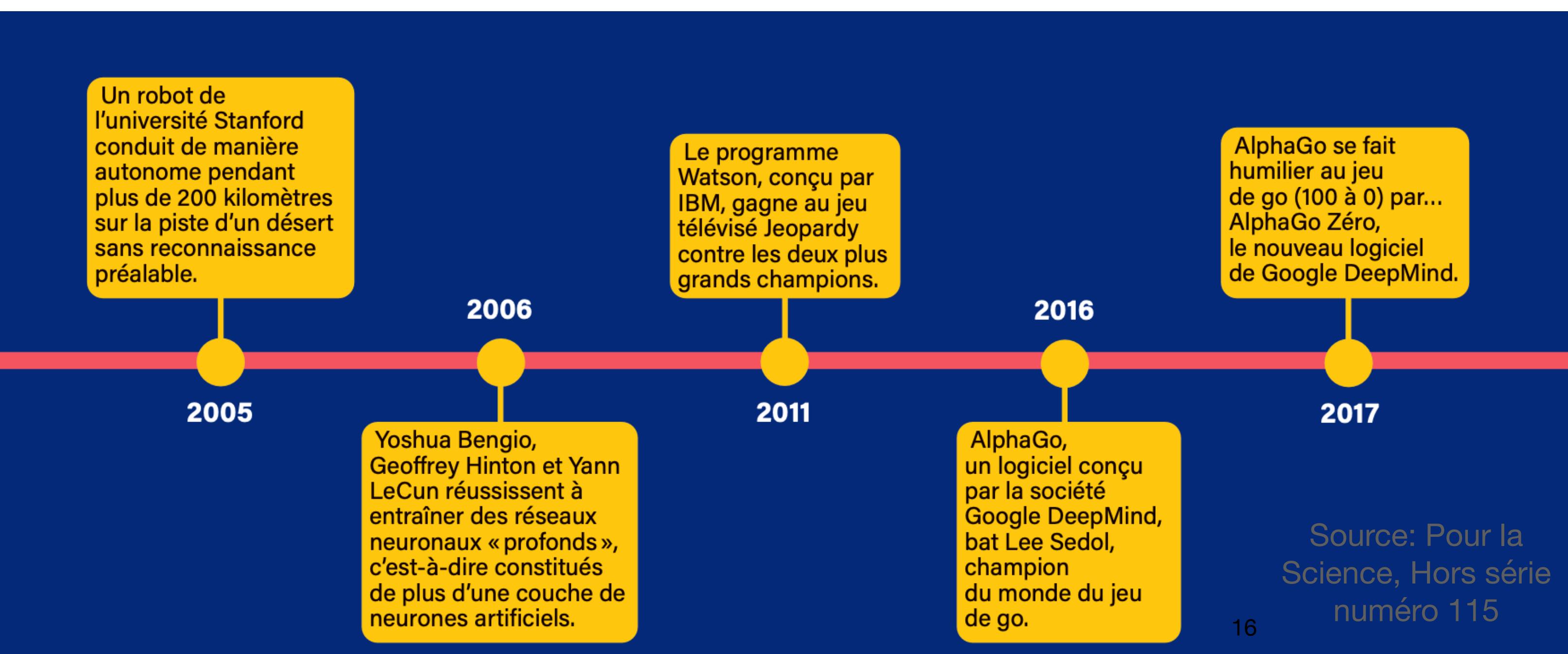
Nouveau (court) regain, notamment avec les systèmes basés sur la connaissance

- P. McCorduck: « *AI researchers were beginning to suspect - reluctantly, for it violated the scientific canon of parsimony- that intelligence might very well be based on ability to use large amounts of diverse knowledge in different ways* »
- Systèmes Experts: tâche spécialisée, basé sur un grand nombre de règles heuristiques élaborées par des experts humains
 - DENDRAL (1969, mais aucun impact à cette date): déterminer la structure moléculaire d'une molécule étant donné sa formule et les résultats de sa spectrométrie de masse



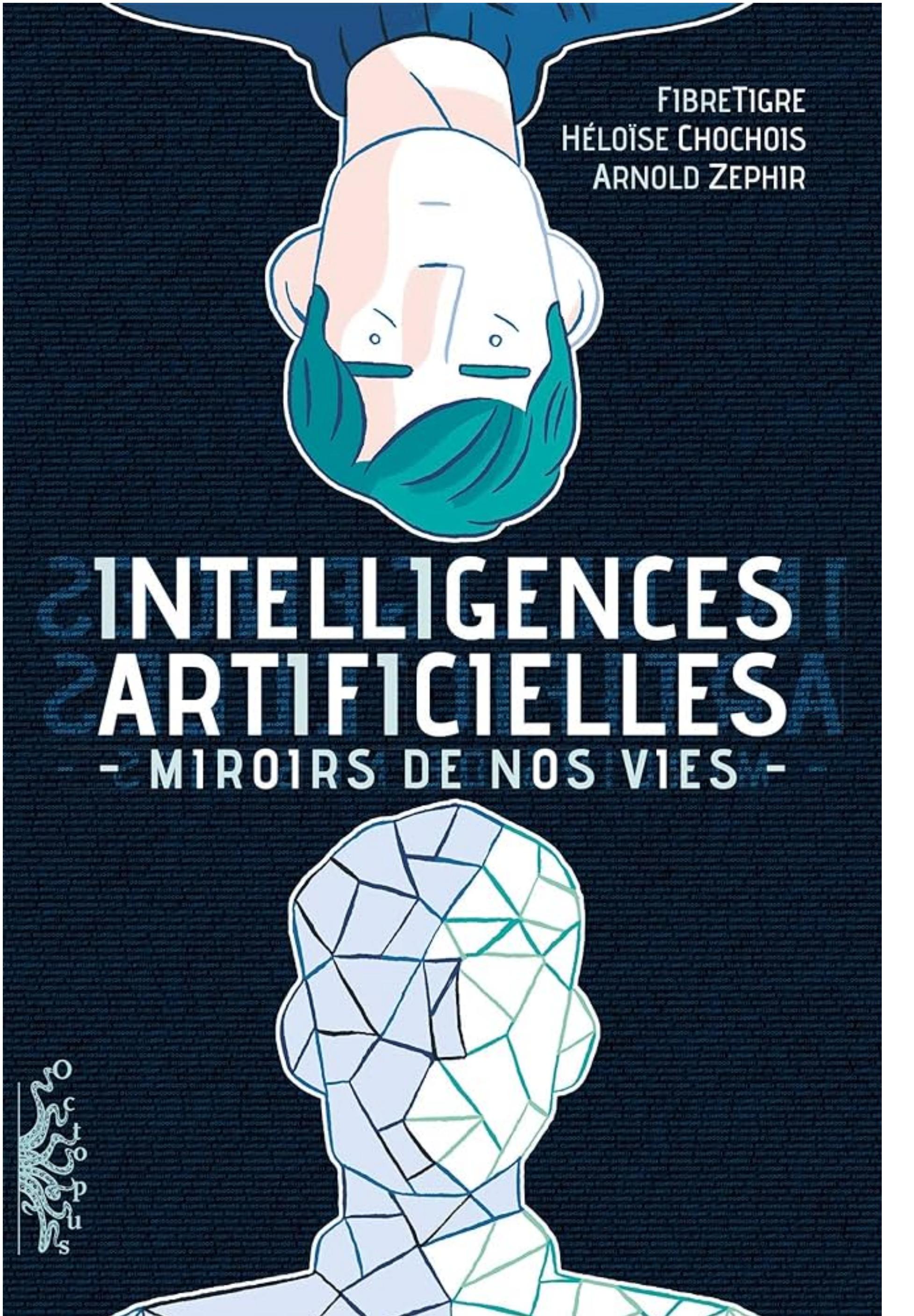
« Une histoire à rebondissement »

- **11 mai 1997:** Deep Blue, d'IBM, bat aux échecs le champion du monde en titre, Garry Kasparov. (En 58, les chercheurs pensaient que ça allait avoir lieu avant 68)
- **Début des années 2000:** la puissance des ordinateurs permet à l'IA de faire des progrès spectaculaires



Un ordinateur IBM similaire à « Deep Blue »¹ exposé au musée de l'histoire de l'ordinateur de Mountain View.

Le grand maître russe et champion du monde Garry Kasparov, ici en 1996, qui affronta Deep Blue lors de deux matchs. Source: Wikipédia



2. Définition de l'IA (?)



Comment définir l'intelligence
artificielle?
À vous de jouer !

Vos définitions:

- (...)

Définition(s) de l'IA

- L'objectif de l'IA est de créer des systèmes **intelligents**, mais de quelle intelligence parle-t-on ? **Nécessité de définir des tâches**
 - Veut-on une machine qui *raisonne* ou *agit* intelligemment ?
 - Définit-on l'IA à partir de l'**être humain** ou d'un **standard de rationalité** ?

Définitions de l'IA

Déjà à l'intersection entre plusieurs disciplines

- **Machine qui pense comme les humains:** le but est de reproduire le raisonnement humain dans les machines (imiter les méthodes de raisonnement - sciences cognitives, psychologie - ou reproduire le fonctionnement du cerveau - neurobiologie)
- **Machine qui agit comme les humains:** le but est de ne plus distinguer la machine et l'humain (propriété définie par Alan Turing)
- **Machine qui pense rationnellement:** le but est d'obtenir un raisonnement rationnel, défini par les mathématiques et surtout la logique.
- **Machine qui agit rationnellement:** utiliser la définition d'action rationnelle utilisée en économie pour agir dans le but d'obtenir le meilleur résultat en sachant prendre en compte des éléments d'imprévisibilité.

Le test de Turing: une machine indiscernable de l'humain

Test de Turing « total » où le juge a une « vraie » interaction avec la machine:

Vision artificielle (but: reconnaître les objets)

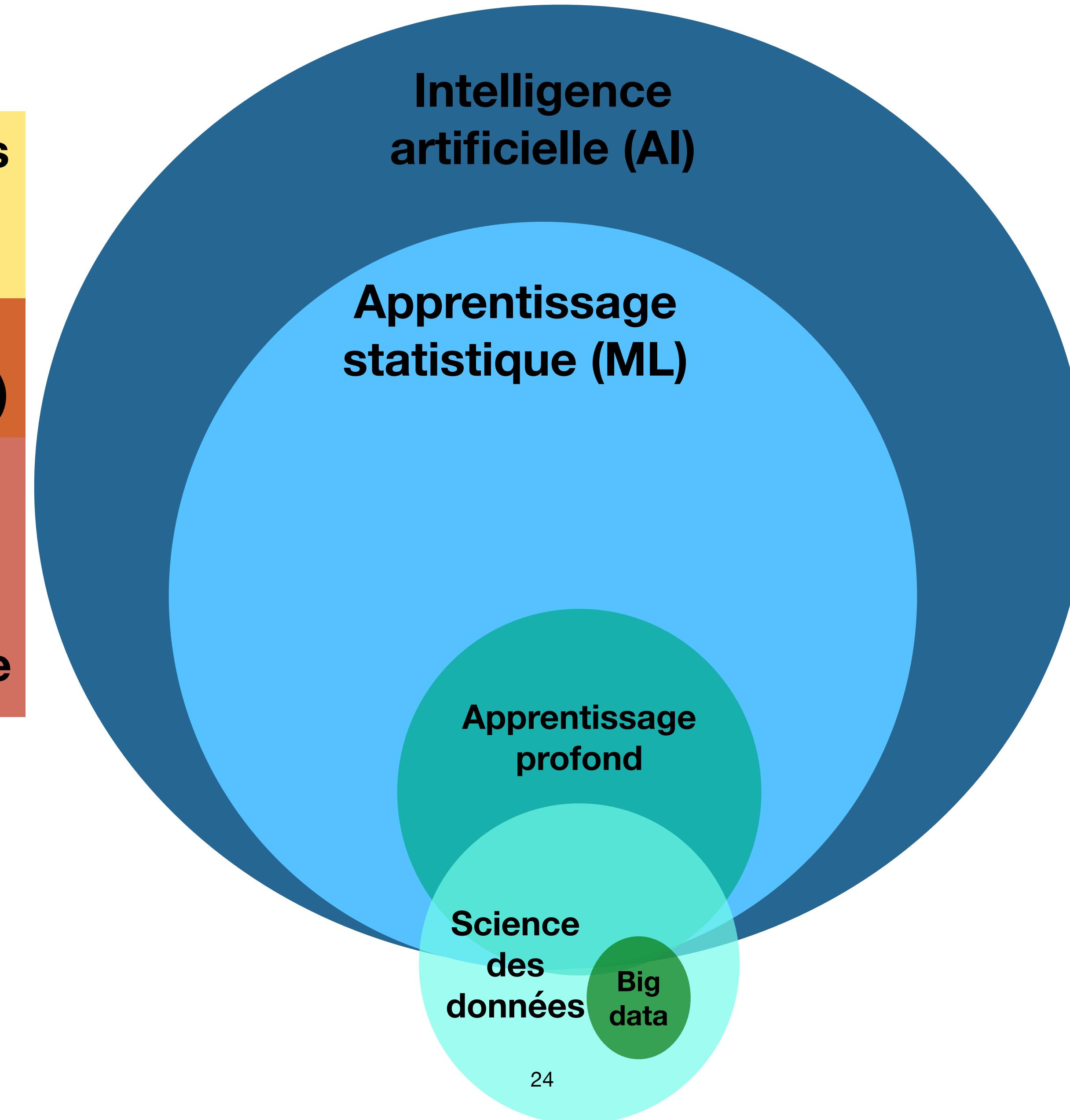
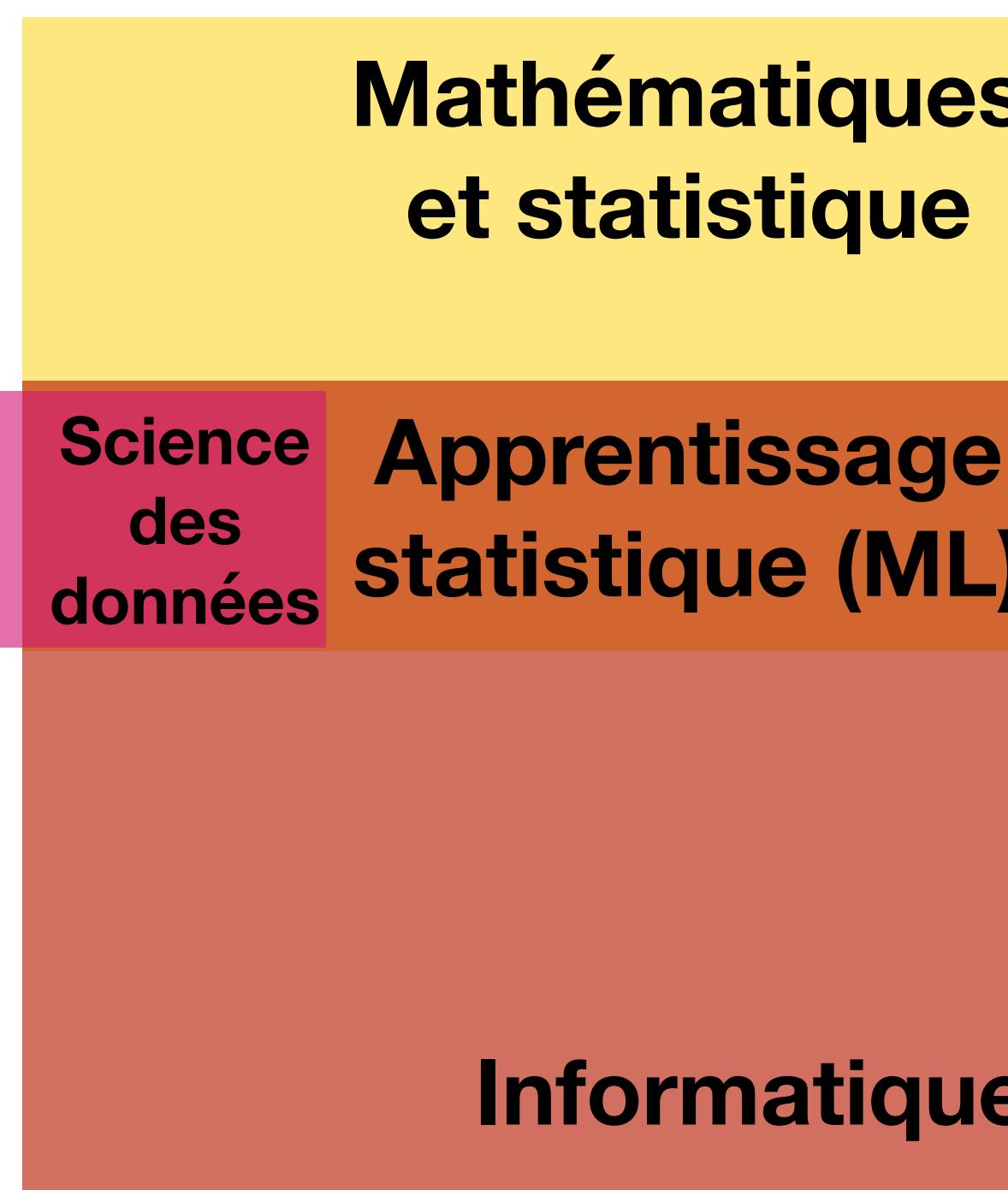
Robotique (but: manipuler les objets)

Test de Turing (1950) évalue les capacités suivantes:

- Traitement du langage naturel (but: communiquer)
- Représentation des connaissances (but: mémoriser)
- Raisonnement automatique (but: produire des réponses à partir de ses connaissances)
- Apprentissage automatique (but: élargir sa base de connaissance)

IA, ML, data science ?

Domaine d'expertise



Apprentissage statistique (ML):
Herbet Alexander Simon, économiste et sociologue (1916-2001): « *Learning is any process by which a system improves performance from experience.* »

Science des données:
étude/science de l'extraction généralisable de connaissances à partir de données

IA: un mot pour le buzz ?

Serge Abiteboul, directeur de recherche Inria, ENS Paris:

« En fait, j'ai un problème : je vous en parle, mais je ne sais pas trop ce qu'est l'intelligence artificielle ! L'expression fait fantasmer. Mais qu'est-ce qu'elle signifie ? Depuis ma thèse, je travaille sur des systèmes de gestion de base de données, qui répondent aux questions des humains. C'est quand même intelligent de répondre à des questions ! J'ai travaillé sur des bases de connaissances qui font de la déduction. Là encore, c'est intelligent de raisonner. Plus récemment, l'apprentissage automatique m'a permis d'introduire de nouvelles fonctionnalités dans des systèmes sur lesquels nous travaillons avec des étudiants. Distinguer ce qui en informatique tient de l'intelligence artificielle ou pas, ça n'aide en rien. Pour moi, c'est avant tout un buzzword, surtout utile pour récupérer des financements ou impressionner des amis. Le truc cool, aujourd'hui, n'est pas l'intelligence artificielle, mais l'apprentissage automatique qui vient compléter d'autres techniques essentielles de l'informatique. »

Source: Pour la Science, Hors série numéro 115

SPOTLIGHT ON BIG DATA

Spotlight

ARTWORK Tamar Cohen, Andrew J Buboltz
2011, silk screen on a page from a high school
yearbook, 8.5" x 12"

Data Scientist: The Sexiest Job of the 21st Century

Meet the people who
can coax treasure out of
messy, unstructured data.
by Thomas H. Davenport
and D.J. Patil

70 Harvard Business Review October 2012



Havard Business Review, Octobre 2012

Où en est l'IA ?

- **Ce que l'IA peut faire actuellement** (non exhaustif) :
 - **Apprendre** (acquérir des connaissances et des compétences)
 - **Optimiser** (ex: afficher des résultats de recherche rapidement)
 - **Percevoir** (ex: utiliser des capteurs pour déterminer un état)
 - **Imiter** (ex: exploiter des corrélations pour reproduire des motifs complexes, notamment avec des images ou du texte)
- Actuellement l'IA est dite « faible », **sans autonomie ni conscience, spécialisée dans une tâche spécifique.**
- **L'IA dit « forte » n'est que de la science fiction !** (Samatha dans Her, HAL dans l'Odyssée de l'espace, les androïdes dans Blade Runner,...)

Francis Bach, directeur de recherche Inria, ENS Paris:

« *Par nature, les méthodes actuelles de l'IA ne peuvent se mettre en œuvre que sur des tâches pré-définies (comme reconnaître un piéton dans une image ou jouer au jeu de Go): si les méthodes d'apprentissage utilisées sont génériques, les algorithmes finaux et les données sont spécifiques à chacune des tâches. Nous sommes encore loin d'une intelligence "multi-usage" qui permet de s'adapter à des situations qui n'ont jamais été observées auparavant (l'algorithme qui a battu le champion du monde de Go ne pourra pas être utilisé directement pour jouer aux échecs). De plus, ces tâches doivent souvent être déjà sous forme numérique, même si les progrès constants de la robotique laissent entrevoir l'utilisation de robots dans des situations plus incertaines et moins contrôlées que des ateliers d'usine. Ainsi, les technologies actuelles permettent d'aider les humains dans certaines tâches (par exemple, la recherche d'information, ou la traduction automatique d'une langue à une autre), sans pour autant pouvoir les remplacer. »*



Source: IA et emploi : une menace artificielle
Philippe Askenazy, Francis Bach, 2019

3. Applications récentes en biologie

L'IA pour la biologie en France

Deux exemples parmi beaucoup d'autres

- **CBIO - Centre for Computational Biology, Mines Paristech**
 - Chloé-Agathe Azencott, enseignante-chercheuse (titre de thèse, 2010: Statistical machine learning and data mining for chemoinformatics and drug discovery)
- **Owkin: *Understanding complex biology through AI, full-stack AI biotech***
 - Jean-Philippe Vert, chercheur à Owkin, directeur R&D (titre de thèse, 2004: Kernel methods for computational biology)



L'IA pour la biologie au cœur de l'écosystème azuréen

- **3iA Côte d'Azur:** un axe entièrement dédié « AI for computational biology and bio-inspired AI »
- **Equipes INRIA:** Compo (Méthodes computationnelles pour la prise en charge thérapeutique en oncologie), Biocore (Biological control of articula ecosystems), ABS (Algorithmes et Biologie Structurale, biologie moléculaire), Epione (E-Patient : Images, données & mOdèles pour la médecine numériquE),...
- **Neuromod:** biologically inspired Machine Learning
- **Des start-ups, par exemple:**
 - Bioceanor (IA pour la biologie océanique, surveillance de qualité de l'eau, aquaculture)
 - Therapixel (IA pour l'imagerie médicale)

COMPUTATIONAL BIOLOGY

- Molecules: mining conformational spaces of huge dimensions to reveal biological functions
- Networks: combining single cell atlases and interaction networks (protein, metabolic, genetic, signaling, etc.) to reveal molecular pathways
- Cells/tissues: 3D+t super-resolution /multispectral microscopy to reveal differentiation/development complexity
- Brain: neuron-to-brain integration to model brain activity & computational neuroscience

BIO-INSPIRED AI

- Neuronal level: spiking models to better understand neuronal dynamics
- Cognition: neuronal dynamics for the analysis of learning/perception/action sequences
- Simulation/electronics: brain models to provide new neuromorphic-biomimetic algorithms/architectures

Source: <https://3ia.univ-cotedazur.eu/>

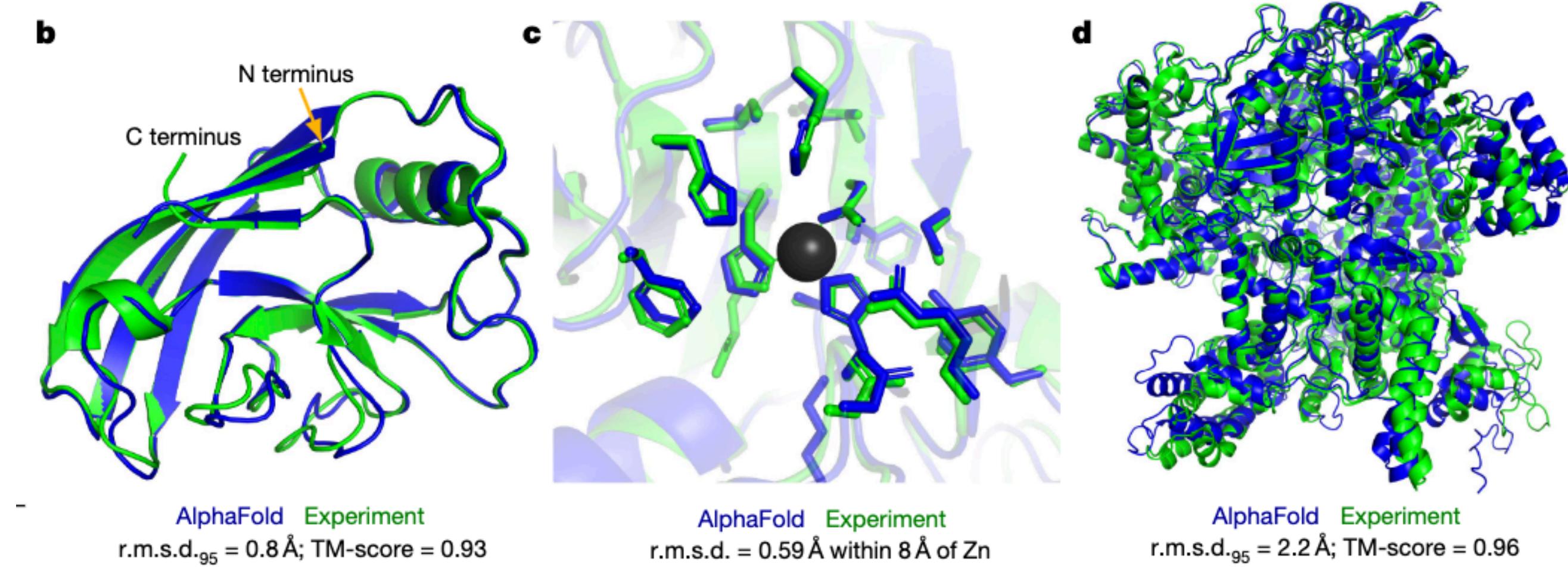
Ellen Van Obberghen-Schilling (Inserm)
IA appliquée à l'imagerie pour la détection du cancer

*AI-powered analysis of the tumor microenvironment
« Our project will integrate tissue imaging modalities and artificial intelligence-based analysis tools for a deeper understanding and control of cancer, targeting tumor microenvironment and on the role of the extracellular matrix (ECM) in carcinoma progression, spread and response to therapy. »*



Application 1: AlphaFold2 (novembre 2020)

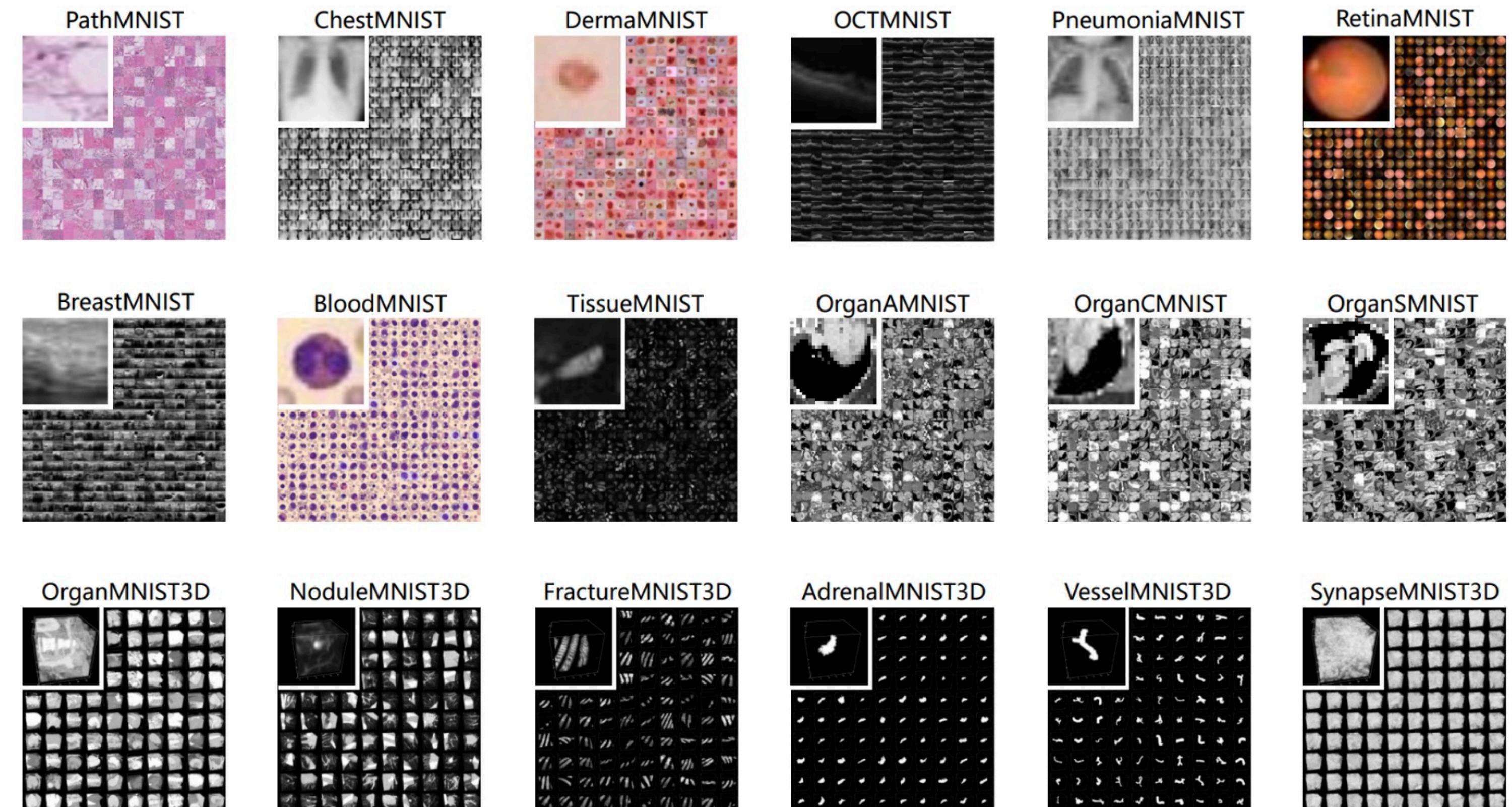
- But: **prédire la structure tridimensionnelle d'une protéine** à partir de sa séquence linéaire en acides aminés.
- Avant AlphaFold2 (novembre 2020, gagnant de la compétition CASP, *Critical assistant of structure prediction*), il manquait les informations structurelles pour 48000 protéines humaines; maintenant il n'en manque que pour 29.
- Modèle d'apprentissage profond (avec des **réseaux de neurones**).



Source: Highly accurate protein structure prediction with AlphaFold, 2021, Nature, Jumper et al.

Application 2: classification d'images médicales (2D, 3D)

- 12 datasets 2D, 6 datasets 3D
- Exemple pour dermaMNIST: le but est de **prédir si une lésion est bénigne ou non.**
- Modèle d'apprentissage profond (avec des **réseaux de neurones**).



Source: [medmnist](#)

- Prédiction de structure de protéines
- Diagnostic du cancer
- Immunologie, prédiction de peptides immunogéniques
- Identification d'expressions aberrantes par des données de séquençage ARN
- Support à la prise de décision des experts en pharmacovigilances
- Simulation de la signalisation intra-cellulaire au niveau du genome entier
- Interprétation des EEG
- ...

4. Introduction à l'apprentissage statistique

Jeu de données

- Jeu de données (dataset): regroupe plusieurs **valeurs/entrées** ayant un lien cohérent entre elles.
- Il peut généralement se présenter sous la forme d'un **tableau (= matrice)**.
- Ces valeurs/entrées peuvent être numériques, mais aussi des images, du texte,...

Une ligne = une observation

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
	82	17.0	18.4	17.7	5	5	7	-4.3301	-4.0000	-3.0000	87	Nord	Sec
	92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
	114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
	94	17.4	20.5	20.4	8	8	7	-0.5000	-2.9544	-4.3301	114	Ouest	Sec
	80	17.7	19.8	18.3	6	6	7	-5.6382	-5.0000	-6.0000	94	Ouest	Pluie

Une colonne = une variable

Jeu de données

maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
82	17.0	18.4	17.7	5	5	7	-4.3301	-4.0000	-3.0000	87	Nord	Sec
92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
94	17.4	20.5	20.4	8	8	7	-0.5000	-2.9544	-4.3301	114	Ouest	Sec
80	17.7	19.8	18.3	6	6	7	-5.6382	-5.0000	-6.0000	94	Ouest	Pluie

Extrait du jeu de données ozone

- En mathématiques, on *représente* ce jeu de données par une matrice:

$$X = \begin{pmatrix} 82 & 17.0 & 18.4 & \dots \\ 92 & 15.3 & 17.6 & \dots \\ \vdots & \vdots & \vdots & \end{pmatrix}.$$

- Une **matrice** est seulement un tableau à **2 dimensions** de nombres.
- Un **vecteur** est un tableau à **1 dimension** peut par exemple être l'ensemble des valeurs qui constitue la première colonne de X (ici: 82, 92, ...) ou l'ensemble des valeurs qui constitue la première ligne de X (ici: 82, 17.0, ...)

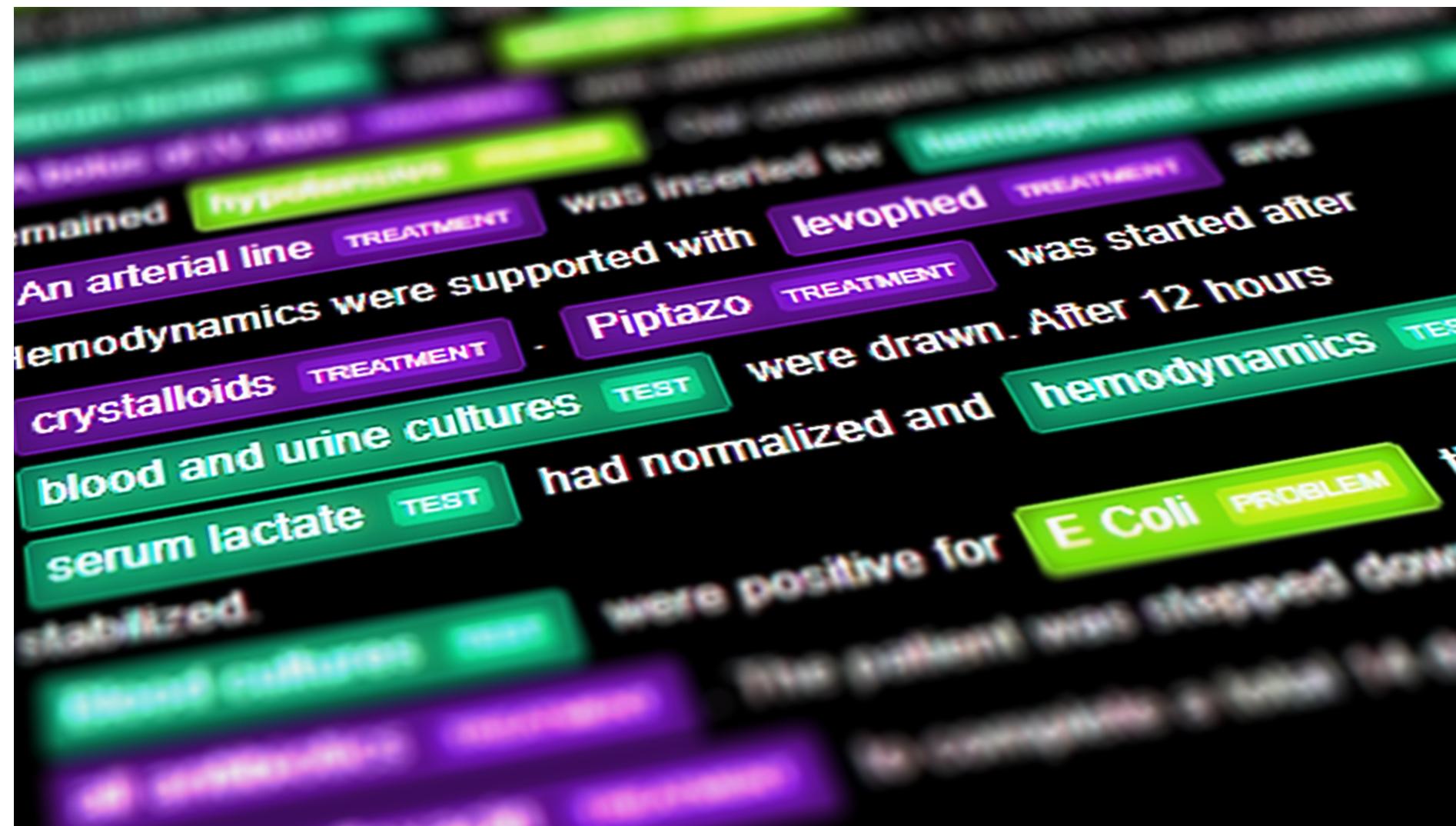
Différents types de variables

- **Variable *continue***: ses valeurs sont **numériques** et sont comprises dans un ensemble fini ou infini, elles peuvent prendre toutes les valeurs de cet ensemble.
- **Variable *catégorielle***: se réfère à une caractéristique **non quantifiable**, cette variable peut prendre des modalités/catégories.
 - Exemple: dans le jeu de données ozone, la variable vent donne l'orientation du vent à 12 h et prend les catégories « Est », « Ouest », « Nord », « Sud ».
 - Une variable catégorielle peut être ordinaire, si les catégories peuvent être classées, par exemple si une variable donne la mention au bac.

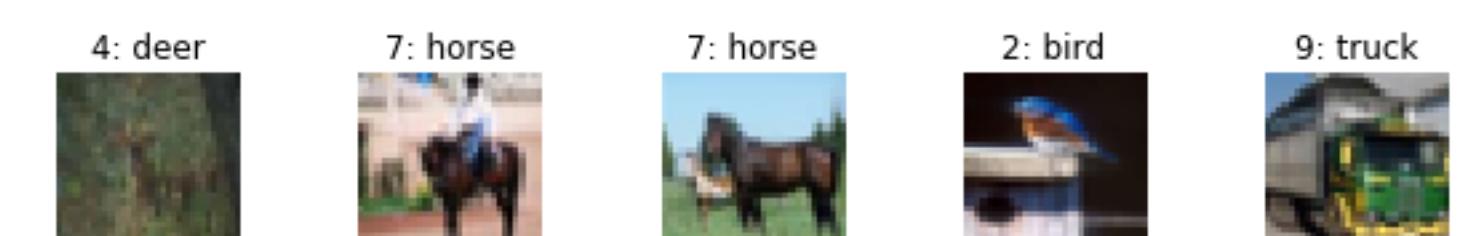
Différents types de jeu de données

maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
82	17.0	18.4	17.7	5	5	7	-4.3301	-4.0000	-3.0000	87	Nord	Sec
92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
94	17.4	20.5	20.4	8	8	7	-0.5000	-2.9544	-4.3301	114	Ouest	Sec
80	17.7	19.8	18.3	6	6	7	-5.6382	-5.0000	-6.0000	94	Ouest	Pluie

Jeu de données tabulaire, ozone



Jeu de données textuel

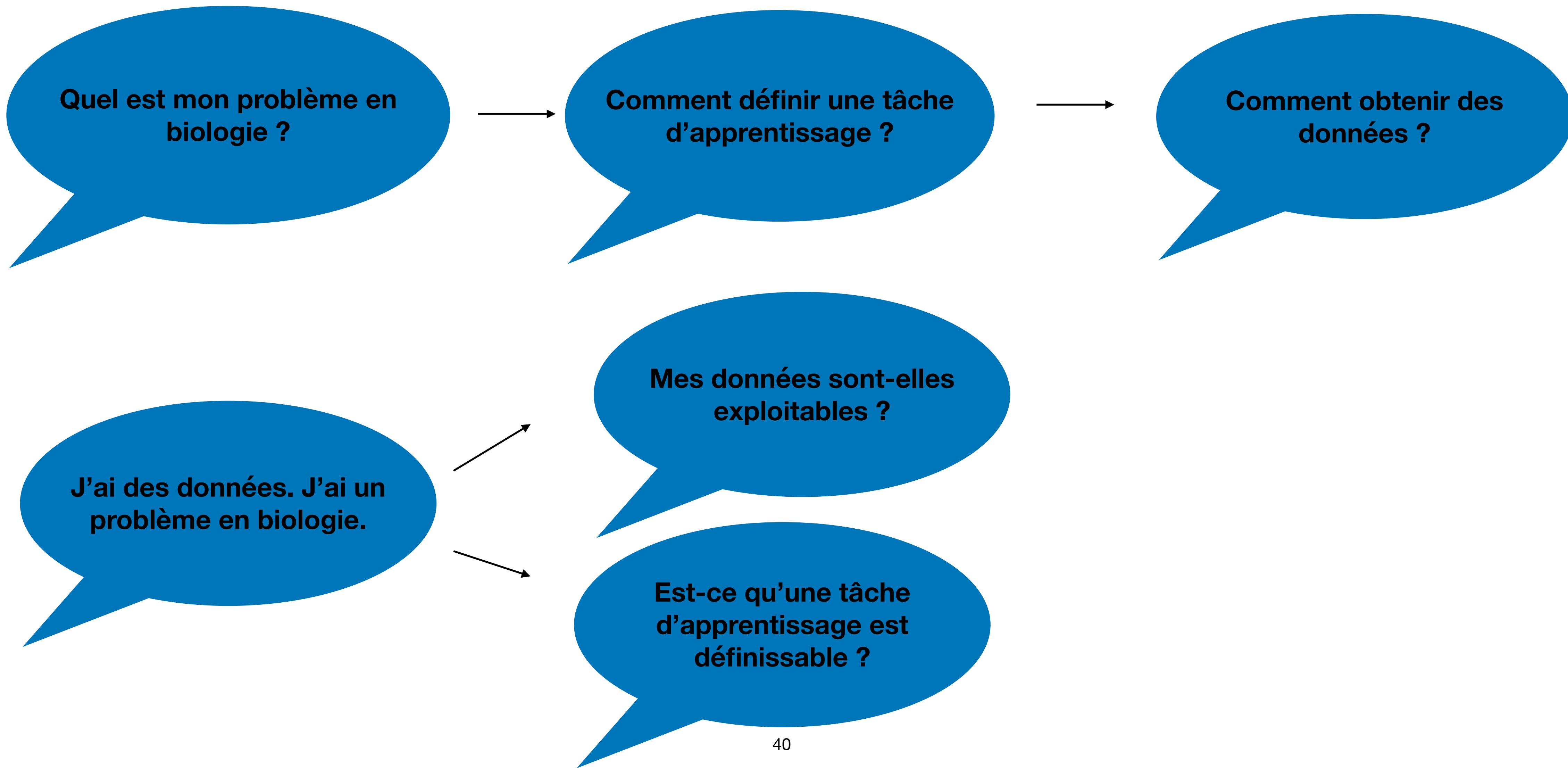


Jeu de données d'images, CIFAR10

À vous de répondre...

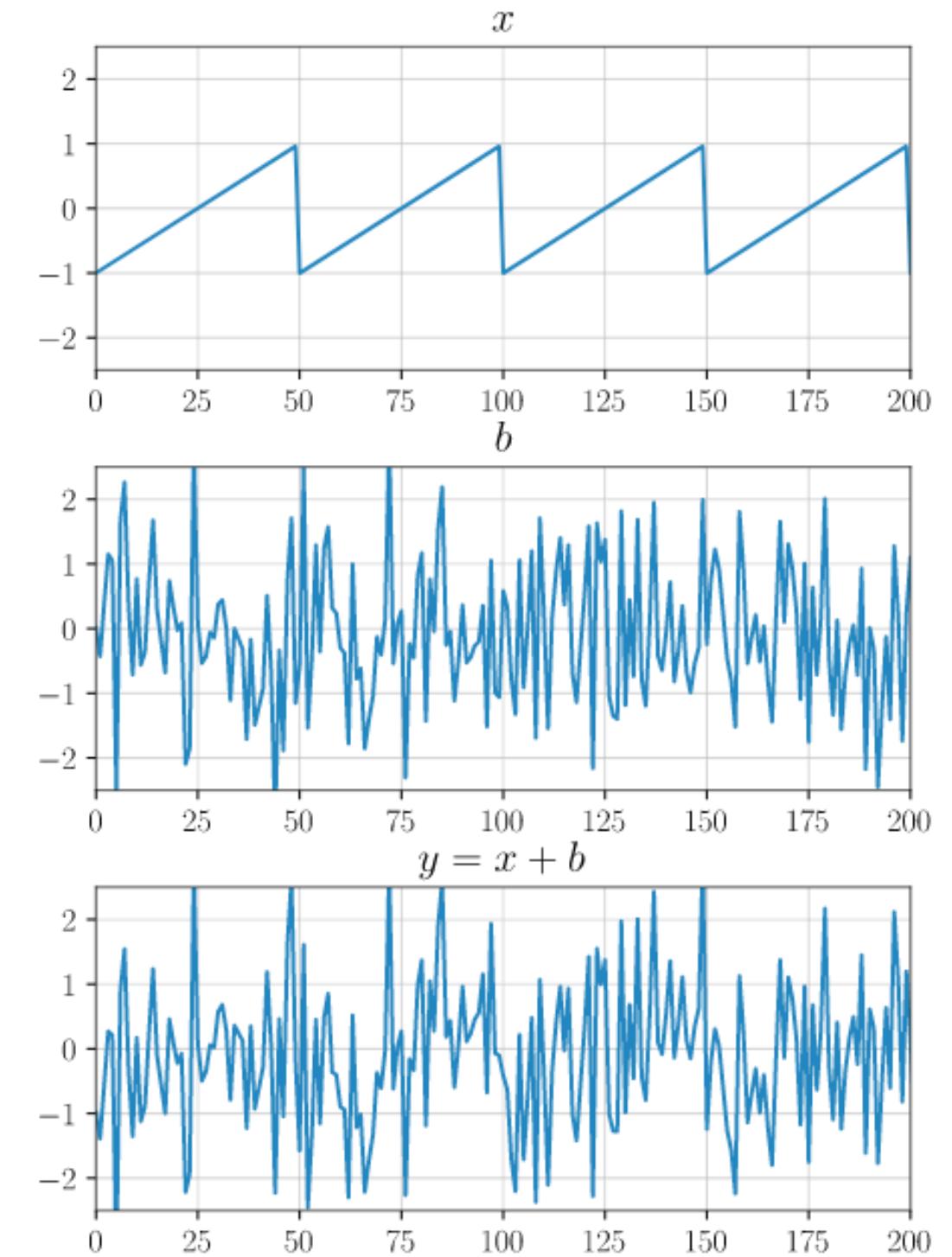
- Quelles tâches peut-on définir à partir d'un jeu de données ?
 - Exemple avec le jeu ozone: (...)
- Quelles questions doit-on se poser lorsque l'on dispose d'un jeu de données ?
 - (...)

En tant qu'experts (biologistes, médecins, ...), vous allez plus souvent être confrontés aux chemins suivants:



Qualité des données

- Les données peuvent être **bruitées**: il y a un aspect qui n'est pas caractérisé par une variable dans le jeu de données mais qui influe sur le résultat. Par exemple, pour la concentration d'ozone, l'altitude où elle a été mesurée.
- Certaines valeurs peuvent être **manquantes**: très souvent, il peut arriver qu'une caractéristique ne soit pas mesurée par manque de temps, parce que le capteur est cassé, ou dans un sondage parce que les personnes n'ont pas envie de répondre à une question intime (exemple: quel salaire gagnez-vous?). Cela crée des *trous* dans la matrice, et mathématiquement, on comprend bien que ce sera un problème pour faire des calculs même simple (addition).



Source: Vincent Mazet, traitement du signal

Heart rate	Death	Anticoagulant. therapy	Glasgow score	...
88	0	No	3	
103	0	NA	5	
NA	0	Yes	6	
NA	0	No	NA	
62	0	Yes	6	
NA	0	No	NA	
:	:	:	:	

Deux défis qui peuvent mettre les algorithmes en difficulté

- **Grande dimension**, quand on a beaucoup plus de variables explicatives que d'observations: c'est notamment le cas en génomique (plusieurs millions de gênes). Beaucoup d'algorithmes de statistique classique ne sont adaptés qu'au cas où il y a plus de d'observations que de variables explicatives.
 - En fait, dans un espace de grande dimension, il y a **beaucoup de vide**. Par exemple, une boule occupe 52,3% d'un « cube » en dimension 3; en dimension 8, elle n'occupe que 1,5% du « cube ». On veut synthétiser l'information contenue dans les données de grande dimension en plus petite dimension: c'est le but de la réduction de dimension.
- **Très peu d'observations**: c'est notamment le cas lors d'études cliniques invasives pour les patients.
 - Quand il n'y a vraiment pas assez d'observations, cela ne sert à rien d'entrainer un modèle d'apprentissage compliqué. Par exemple, il vaut mieux se servir de tests statistiques adaptés aux petits échantillons.

Beaucoup de jeux de données publics

- UCI Machine Learning Repository: <https://archive.ics.uci.edu/datasets>
- Kaggle: <https://www.kaggle.com/datasets>
- scikit-learn: <https://scikit-learn.org/stable/modules/classes.html#module-sklearn.datasets>

The screenshot shows the homepage of the UC Irvine Machine Learning Repository. At the top, there is a navigation bar with links for 'Datasets', 'Contribute Dataset', 'About Us', a search bar labeled 'Search datasets...', and a 'Login' button. Below the navigation bar, there is a section titled 'Filters' with dropdown menus for 'Keywords', 'Data Type', 'Subject Area', 'Task', '# Attributes', '# Instances', and 'Attribute Type'. To the right of the filters, there is a section titled 'Browse Datasets' with a 'SORT BY # VIEWS, DESC' button and an 'EXPAND ALL' button. The main content area displays four dataset entries: 'Iris', 'Heart Disease', 'Adult', and 'Dry Bean Dataset', each with a small icon, a brief description, and detailed statistics like 'Classification', 'Tabular/Multivariate', 'Instances', and 'Attributes'.

Dataset	Description	Type	Instances	Attributes	
Iris	A small classic dataset from Fisher, 1936. One of the earliest known datasets used for evaluation.	Classification	Tabular	150 Instances	4 Attributes
Heart Disease	4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach	Classification	Multivariate	303 Instances	13 Attributes
Adult	Predict whether income exceeds \$50K/yr based on census data. Also known as "Census Income".	Classification	Multivariate	48.84K Instances	14 Attributes
Dry Bean Dataset	Images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera.	Classification	Multivariate	13,611 Instances	10 Attributes

Beaucoup de jeux de données sont considérés comme *jouets*, ce n'est pas parce qu'une méthode d'apprentissage marche bien sur un jeu de données qu'elle sera performante sur des jeux de données réelles (plus de bruit, moins de qualité,...)

Pourquoi?
On pourrait demander à des experts d'analyser quelques données !

Apprentissage statistique

Oui mais ce ne serait pas automatisé, si on a beaucoup de données; si on a des données plus complexes, c'est impossible même pour un expert !

On peut classer en deux grandes catégories:

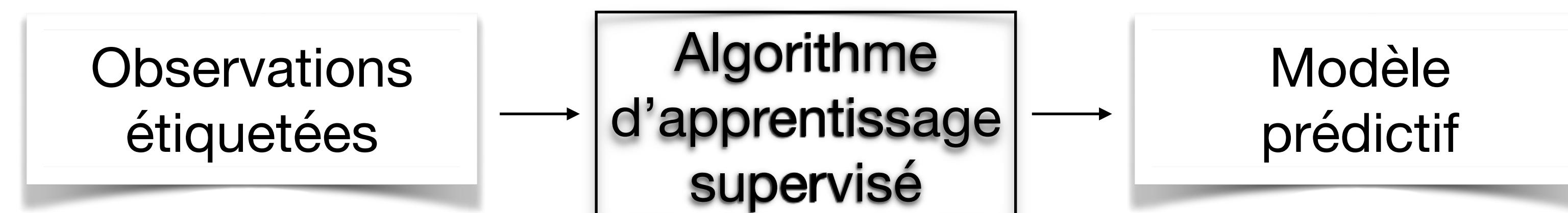
- **Apprentissage supervisé:** apprendre un modèle prédictif
 - Exemple: prédire la classe d'une tumeur à partir de l'expression génétique
- **Apprentissage non supervisé:** apprendre une structure sous-jacente
 - Exemple: classification cellulaire à partir de données de séquençage d'ADN

Mais aussi...

- **Apprentissage par renforcement:** apprendre une politique permettant de maximiser une récompense au cours du temps
- **Apprentissage semi-supervisé:** apprendre un modèle prédictif lorsque la variable à prédire est manquante pour certaines valeurs

Apprentissage supervisé

- **But:** apprendre un modèle prédictif



Nom des descripteurs (=variables explicatives)

maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
82	17.0	18.4	17.7	5	5	7	-4.3301	-4.0000	-3.0000	87	Nord	Sec
92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
94	17.4	20.5	20.4	8	8	7	-0.5000	-2.9544	-4.3301	114	Ouest	Sec
80	17.7	19.8	18.3	6	6	7	-5.6382	-5.0000	-6.0000	94	Ouest	Pluie

Quantité que l'on
veut prédire
(variable cible,
étiquette)

Extrait du jeu de données ozone

- **MaxO3:** valeur max d'ozone observée sur une journée ;
 - T9, T12 et T15: températures prises à 9 h, 12 h et 15 h ;
 - Ne9, Ne12, Ne15: nébulosités prises à 9 h, 12 h et 15 h ;
 - Vx9, Vx12 et Vx15: composantes est-ouest du vent mesurées à 9 h, 12 h et 15 h ;
- MaxO3V: teneur maximale en ozone observée la veille ;
 - vent: l'orientation du vent à 12 h ;
 - pluie: la présence ou non de pluie.

Apprentissage supervisé

maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
82	17.0	18.4	17.7	5	5	7	-4.3301	-4.0000	-3.0000	87	Nord	Sec
92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
94	17.4	20.5	20.4	8	8	7	-0.5000	-2.9544	-4.3301	114	Ouest	Sec
80	17.7	19.8	18.3	6	6	7	-5.6382	-5.0000	-6.0000	94	Ouest	Pluie

Extrait du jeu de données ozone

- Les variables explicatives sont représentées par une matrice X .
- La variable cible est représentée par un vecteur y .
- L'algorithme va apprendre une fonction mathématique f telle que

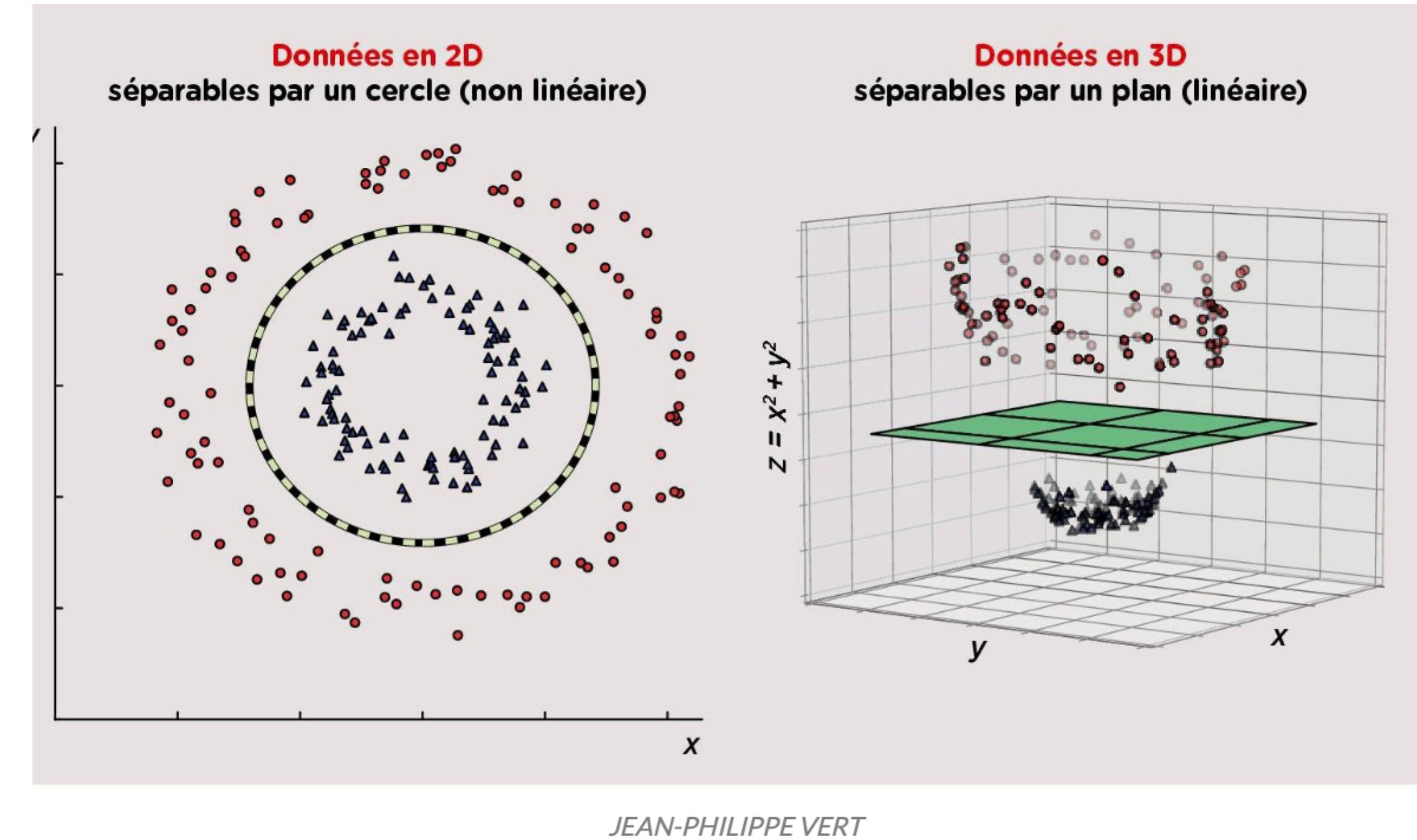
$$y = f(X)$$

- Avec comme objectif de prédire l'étiquette d'une nouvelle observation X_{new}

$$y_{new} = f(X_{new})$$

Apprentissage supervisé: des exemples

- **Régression linéaire, régression logistique:** focus dans la section suivante
- **Méthodes à noyau, Machines à vecteurs de supports** (Vapnik, Chervonenkis, 1960-1990): l'idée est d'appliquer une transformation (un noyau) qui envoie les données dans un autre espace (de plus grande dimension) pour mieux séparer les données, de manière linéaire.
- **Forêts aléatoires** (Breiman, 2001): méthode non linéaire qui repose sur un ensemble de décisions.



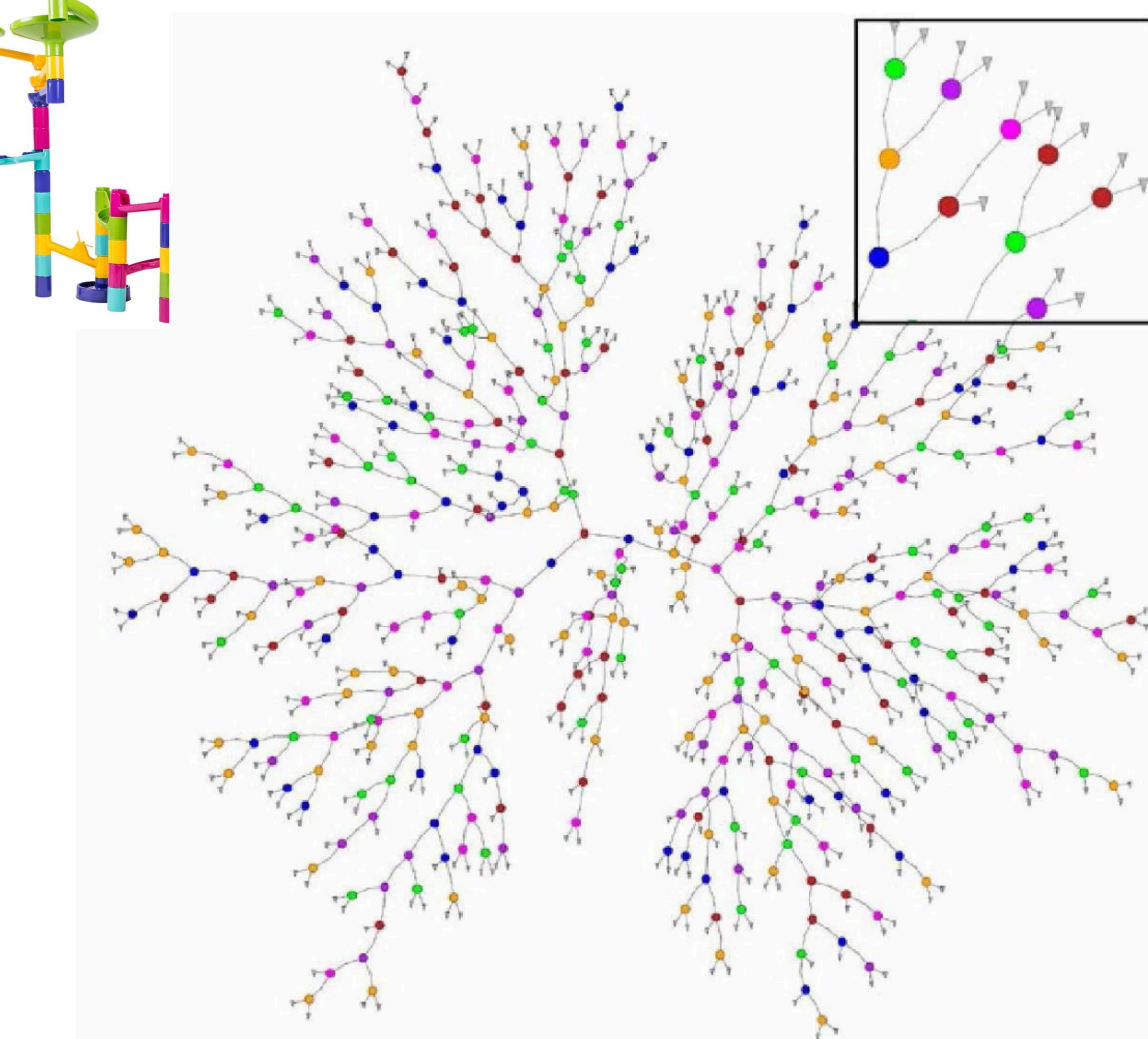
JEAN-PHILIPPE VERT

« En envoyant les données dans un espace de dimension infinie, les méthodes à noyaux agissent à contre-pied de l'intuition qui consiste à réduire la dimension. Grâce à une astuce mathématique, on résout le problème calculatoire, car il existe des opérations mathématiques qui permettent de calculer avec un nombre infini de paramètres, sans faire un calcul infini », explique Jean-Philippe Vert, chercheur en machine learning et bio-informatique chez Google Brain, à Paris (lire p. 70).

Apprentissage supervisé

Forêts aléatoires (ou forêts d'arbres décisionnels)

- Les données (billes) sont comme lancées dans un circuit à billes (**arbre**) avec plusieurs directions: au bout du circuit, chaque bille arrive dans une **feuille** en fonction des directions/ décisions qui ont été prises. Au sein de la feuille, on fait un *vote*, si les billes sont majoritairement **ROUGE**, alors si on lance une nouvelle bille qui arrive dans cette feuille, elle sera prédicté **ROUGE**.
- Gérard Biau, professeur à Sorbonne Université:
« *L'idée des forêts aléatoires ? Au lieu de prendre la décision sur un seul arbre, on en agrège n très grand nombre (...). Autrement dit, on fait voter tous les arbres de manière collective* »



Dans un arbre de décision, à chaque noeud (en couleur), une séparation a lieu en fonction de la réponse donnée à une question, jusqu'à atteindre une feuille (un triangle). -

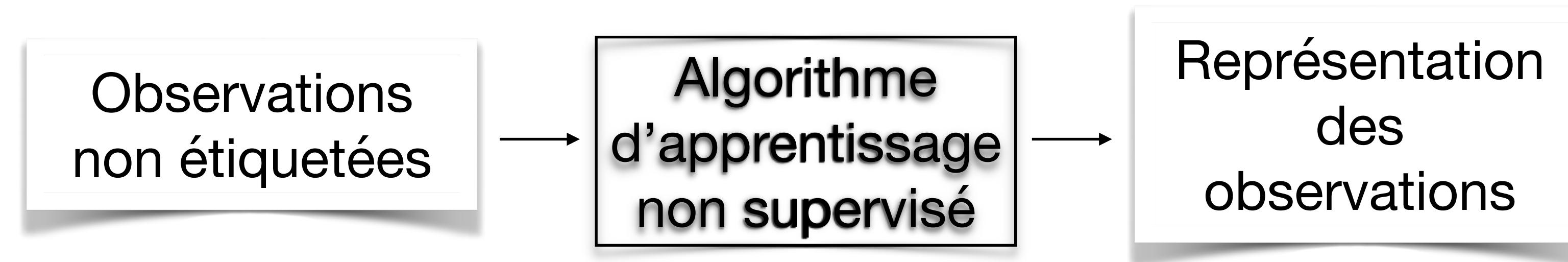
Plus de détails/formalisme à la prochaine séance !

Pour en savoir plus Gérard Biau et Erwan Scornet, « A Random Forest Guided Tour », *TEST*, 25, 197, 2016. Christophe Giraud, *Introduction to High-Dimensional Statistics*, CRC Press/Chapman and Hall, 2014.

Source: La Recherche, numéro spécial, dossier 31

Apprentissage non supervisé

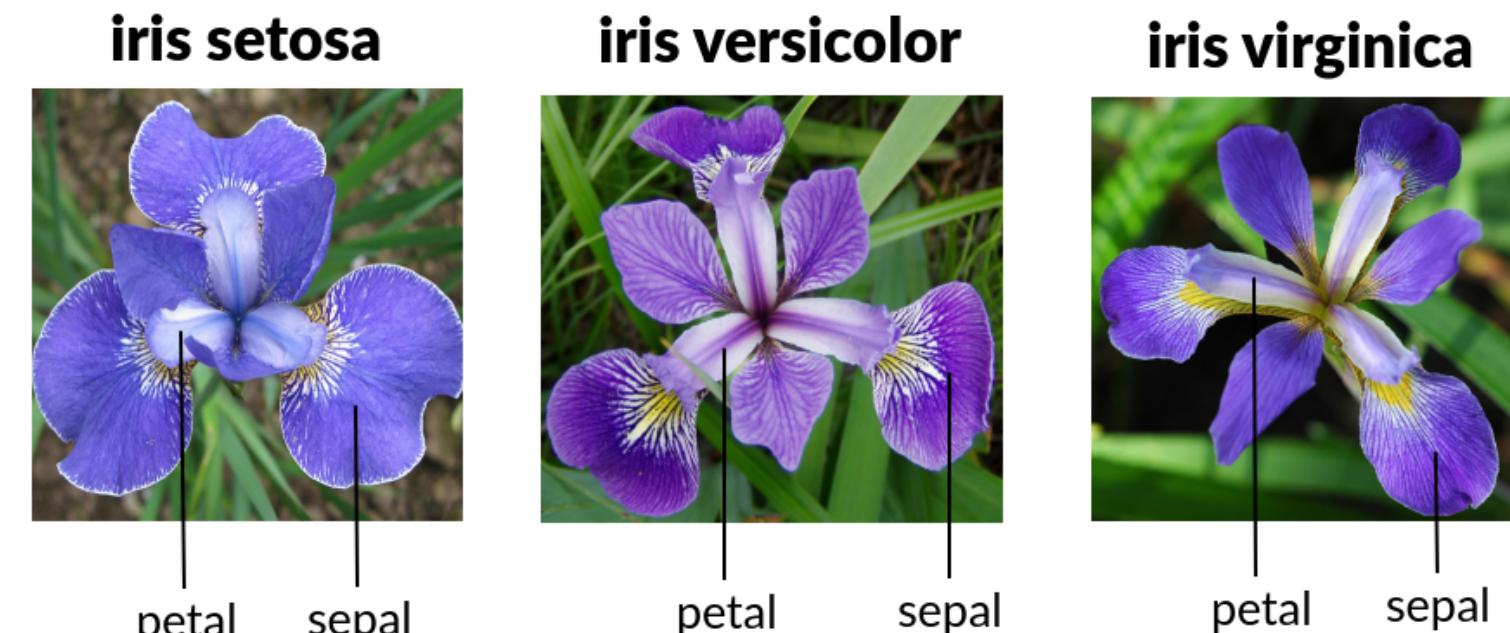
- **But:** apprendre une structure sous-jacente



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

Extrait du jeu de données iris

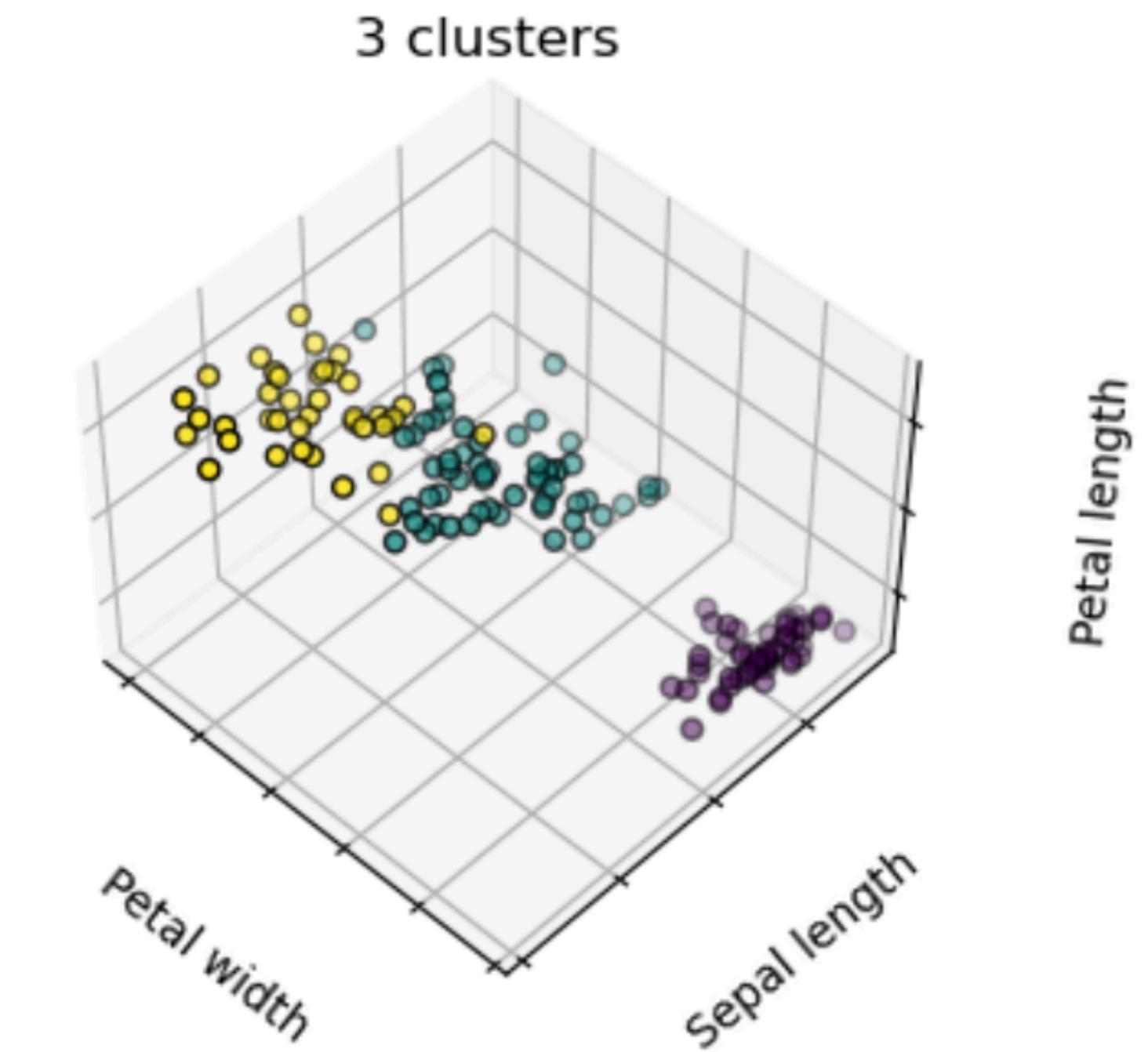
On n'a pas accès à l'espèce des iris (setosa, versicolor ou virginica)



Apprentissage non supervisé: des exemples

- **Clustering ou partitionnement de données** (Lloyd, 1957, algorithme des k-moyennes): l'algorithme va apprendre à **regrouper les données en individus semblables**. On aura différents groupes qui se dégagent en fonction de leurs caractéristiques.
- **Réduction de dimension** (Harold Hotelling, 1930, formalisme de l'ACP): l'algorithme va apprendre une **représentation informative des données**, cela peu permettre de réduire la taille des données en mémoire ou visualiser les données (en moins de 3 dimensions).

Algorithme des k-moyennes appliqué pour iris



Source: [scikit-learn](#)

Apprentissage non supervisé

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

Extrait du jeu de données iris

- Les variables explicatives sont représentées par une matrice X .
- ~~La variable cible est représentée par un vecteur y . PAS ACCES~~
- **Clustering:** l'algorithme va apprendre une partition de X en ensembles d'éléments semblables, appelés clusters.
- **Réduction de dimension:** l'algorithme va apprendre une fonction mathématique f telle que $f(X)$ soit une représentation informative de X .

Apprentissage non supervisé

Analyse à Composantes Principales (ACP)

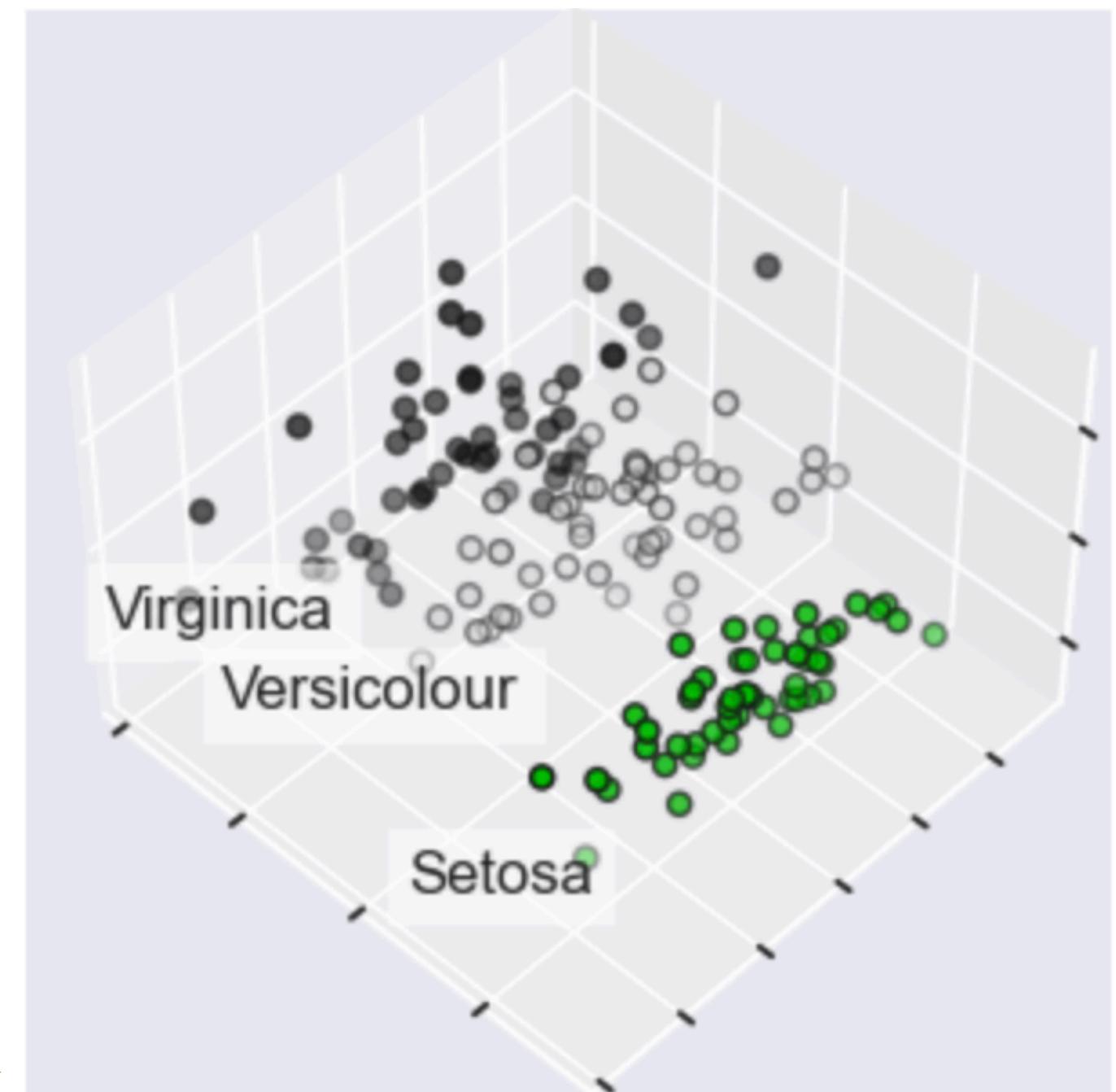
- Le but est de synthétiser l'information contenue dans les données. L'ACP va **trouver un espace où les données sont le mieux représentées possible**, notamment:

- en débruitant: on oublie les axes qui ne nous aident pas et ajoutent du bruit.
- en décolérant: les nouveaux axes donnent des informations distinctes sur les données, ils ne sont pas corrélés.
- L'ACP va donner un pourcentage de « la variance du jeu de donnée » expliquée par chacun des axes trouvés: cela donne une idée de combien d'axes il faut pour bien représenter les données.

Pour iris, les 3 premiers axes expliquent 99,5% de la variance du jeu de données.



Figure 1: Camel or dromedary? source J.P. Fénelon



Source: [scikit-learn](#)

Plus de détails/formalisme à la prochaine séance !

À vous de répondre...

- Quel serait un algorithme d'apprentissage parfait ?

Danger du sur-entraînement

- Le but d'un algorithme d'apprentissage est d'apprendre à partir d'un certain jeu de données pour ensuite être appliqué à de **nouvelles données**.
- On veut évaluer sa capacité à **généraliser**.
 - Exemple: pour le jeu de données ozone, on veut que l'algorithme soit performant pour prédire la valeur maximale de l'ozone observée sur une nouvelle observation.

J'ai une idée ! Ma nouvelle observation sera une observation contenue dans mon jeu de données initial !

Trop facile ! L'algorithme aura déjà « vu » cette observation, il peut l'apprendre par cœur !
Et surtout, ça ne sert à rien ! Tu connais le maximum d'ozone sur cette observation, aucun intérêt à la prédire !

J'ai une autre idée ! Et si j'entraînais mon algorithme pendant une semaine sur mes données ? Il serait très fort !

Surtout pas ! Tu risques d'avoir un algorithme qui aura simplement appris par cœur tes données (en une semaine, il les aura vu tellement de fois !).

Danger du sur-entraînement

- **Généralisation:** capacité de l'algorithme à être performant sur des données qu'il n'a jamais vu. C'est difficile d'être performant sur de nouvelles données, car elles peuvent être bruitées différemment, elles peuvent être *très différentes*.
- On parle de **sur-entraînement (overfitting)** lorsque la performance de l'algorithme est meilleure sur les données initiales que sur les nouvelles données.
- On appelle l'**erreur de généralisation** l'erreur faite par l'algorithme sur de nouvelles données. Le but est de minimiser cette erreur.

Danger du sous-entraînement

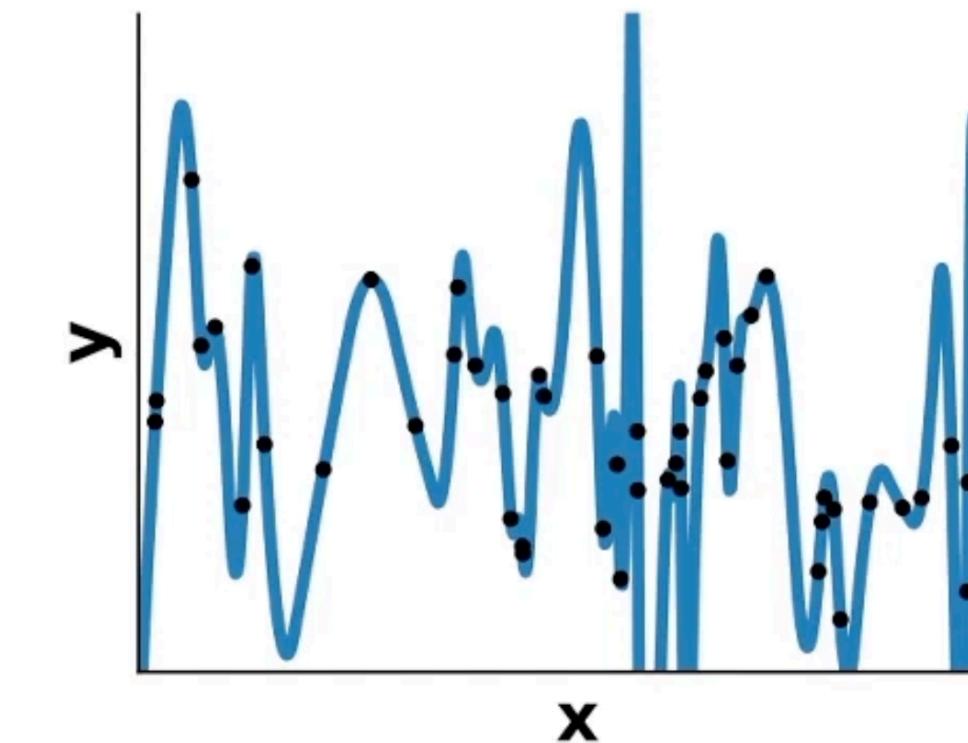
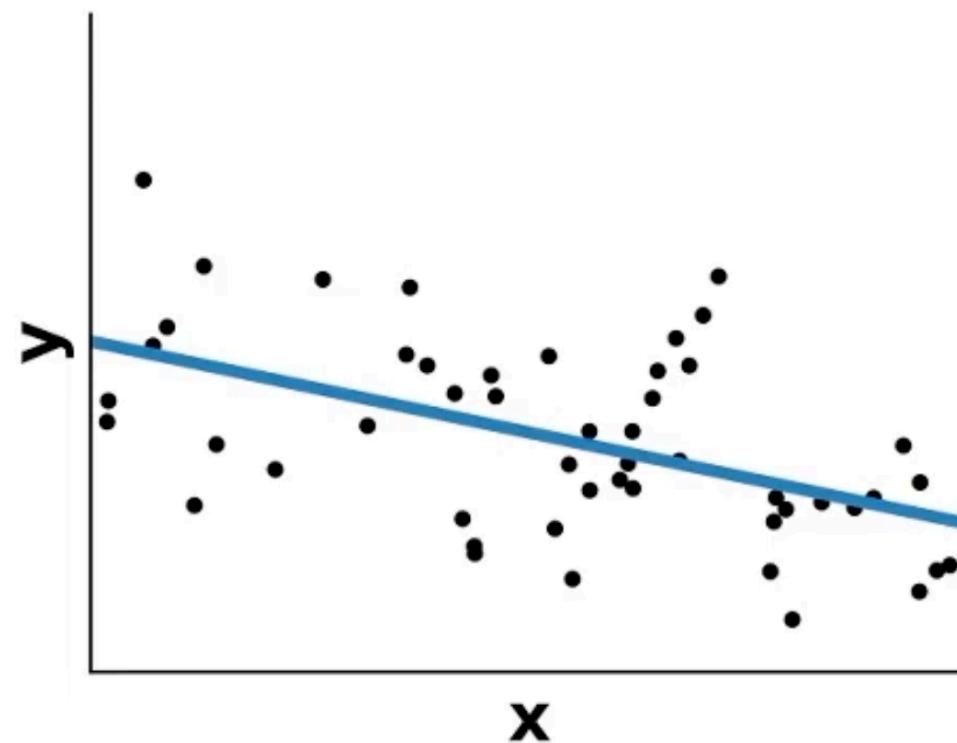
Parfait ! Et bien je vais très peu entraîner mon algorithme ! Quelques minutes, et hop, il sera performant !

Non ! Si tu ne l'entraînes pas assez, s'il n'est pas assez complexe, il ne sera bon ni sur les données initiales, ni les nouvelles données. Il faut quand même lui laisser le temps d'apprendre !

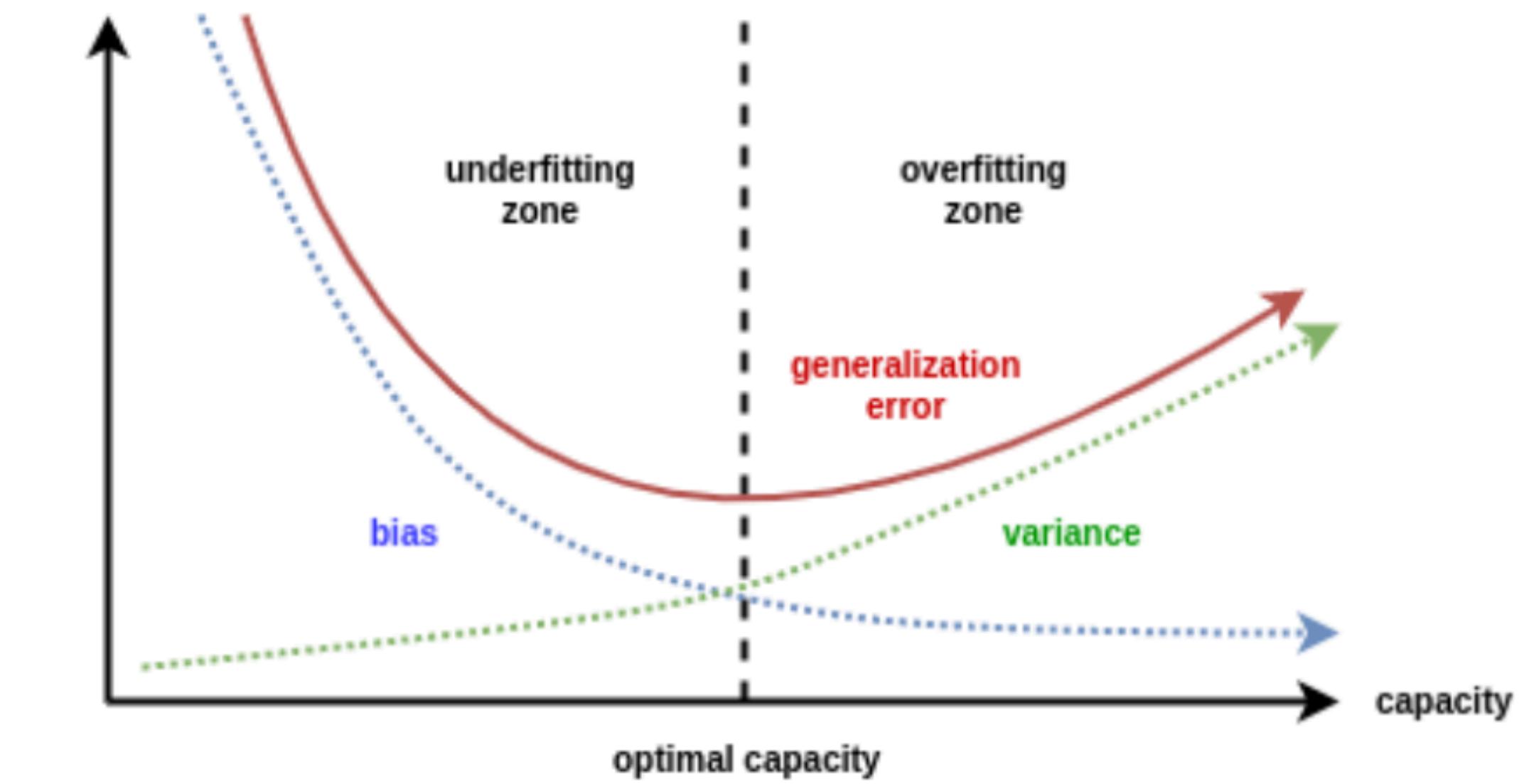
- C'est ce qu'on appelle le **sous-entraînement** (**underfitting**).

Trouver le bon compromis

Which data fit do you prefer?



Source: [scikit-learn](#)



Source: [blog de Daniel Saunders](#)

Comment évaluer la performance d'un algorithme ?

- On va créer de *fausses* nouvelles données, appelé le **jeu de test (test set)**.
- Les autres observations feront partie du **jeu d'entraînement (train set)**.
- Généralement, le découpage typique est 80% train / 20% test.

maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v	vent	pluie
82	17.0	18.4	17.7	5	5	7	-4.3301	-4.0000	-3.0000	87	Nord	Sec
92	15.3	17.6	19.5	2	5	4	2.9544	1.8794	0.5209	82	Est	Sec
114	16.2	19.7	22.5	1	1	0	0.9848	0.3473	-0.1736	92	Nord	Sec
94	17.4	20.5	20.4	8	8	7	-0.5000	-2.9544	-4.3301	114	Ouest	Sec
80	17.7	19.8	18.3	6	6	7	-5.6382	-5.0000	-6.0000	94	Ouest	Pluie

Jeu d'entraînement

Jeu de test

- L'algorithme apprend sur le jeu d'entraînement, et on calcule l'erreur de généralisation (évaluation de la performance) sur le jeu de test.

Quelques questions à se poser face à des données

- **Mes données sont-elles exploitables?** (Dimension: pas assez d'observations ? Grande dimension ? ; propreté du jeu de données: données manquantes / bruit ; ...)
- Quel **type d'algorithmes d'apprentissage** en IA est **applicable** ? Quelle tâche pourrais-je définir ?

5. Zoom sur les modèles linéaires

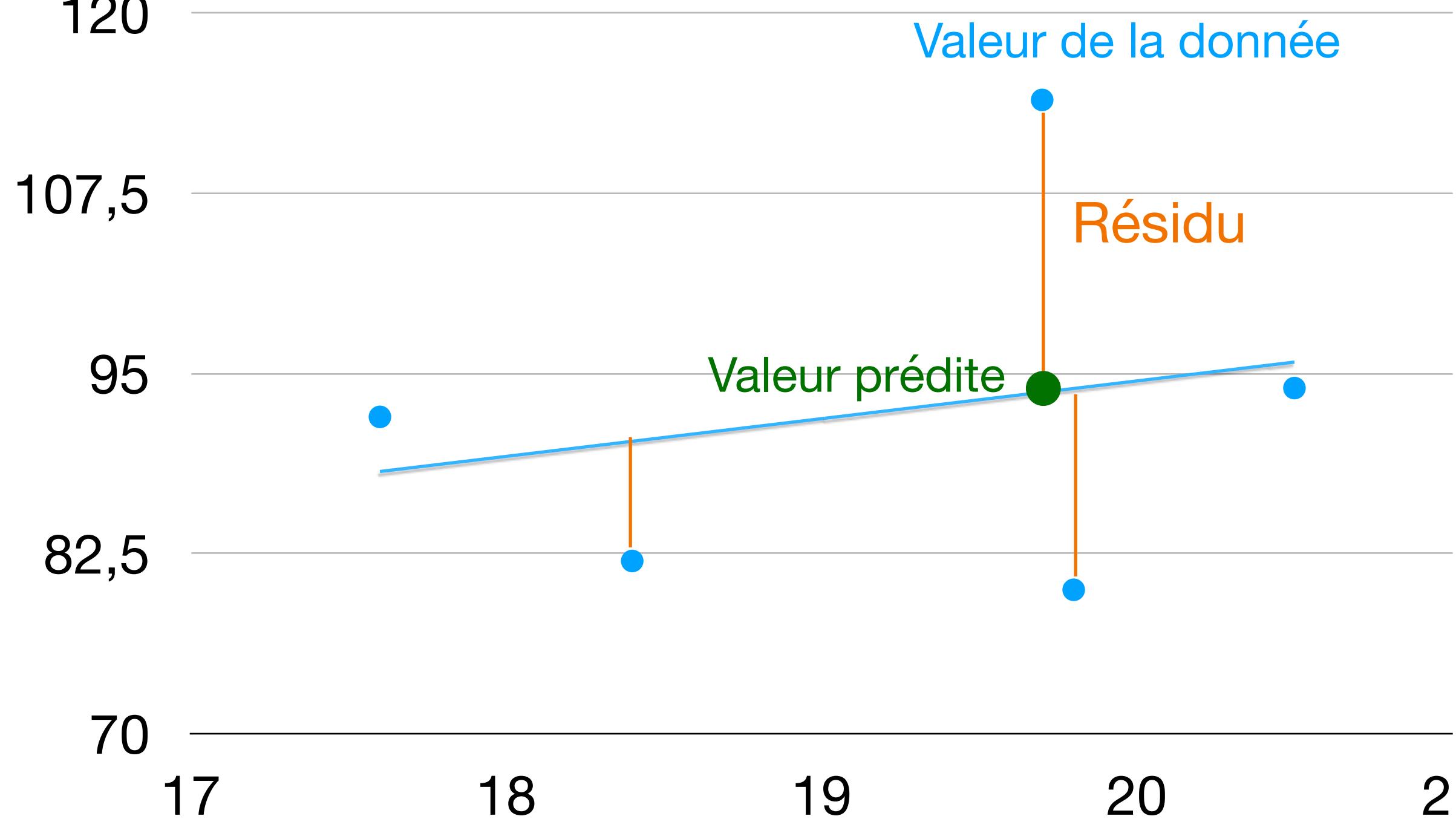
Modèles linéaires

- Avantages:
 - ✓ modèles très simples
 - ✓ facile à interpréter
 - ✓ facile à entraîner
- C'est une base de référence à laquelle comparer un algorithme en apprentissage statistique.
- C'est aussi une base de beaucoup d'autres algorithmes plus complexes, comme les réseaux de neurones. -> prochaine séance
- Contexte de l'**apprentissage supervisé**: on veut prédire une variable cible y à partir d'un jeu de données X .
- Deux exemples abordés:
 - Lorsque la variable à prédire est **continue**: c'est la **régression linéaire**.
 - Lorsque la variable à prédire est **catégorielle**: c'est la régression logistique, aussi appelée **classification**.

Un exemple de régression linéaire simple

- But: trouver une **relation linéaire** entre le maximum d'ozone (**maxO3**) et la température (**T12**). Cela revient à trouver **a** et **b** tels que:

$$\text{maxO3} = a * \text{T12} + b$$



$$82 = a * 18.4 + b$$

...

Objectif: On veut un algorithme qui **minimise les résidus**: la différence entre la courbe bleue (courbe paramétrée par les coefficients **a** et **b**) et les valeurs des données.

maxO3	T12
82	18.4
92	17.6
114	19.7
94	20.5
80	19.8

Un exemple de régression linéaire simple

- On veut trouver un algorithme qui **minimise les résidus**: la différence entre les valeurs prédites et les valeurs des données.
- On appelle l'erreur que l'on souhaite minimiser: **l'erreur quadratique moyenne**

Moyenne(Résidus sur les données)

Nombre de données →

$$\sum_{i=1}^n (y_i - (aX_i + b))^2$$

y_i c'est simplement toutes les valeurs que peut prendre maxO3

X_i c'est simplement toutes les valeurs que peut prendre T12

Régression linéaire multiple

- C'est simplement le cas général: au lieu d'avoir une seule variable *explicative* (comme la température avant), on en a plusieurs.
- Le principe est le même, on cherche à trouver une fonction linéaire f telle que

$$y = f(X).$$

Variables explicatives

maxO3	T12	Ne12	Vx12
82	18.4	5	-4.0000
92	17.6	5	1.8794
114	19.7	1	0.3473
94	20.5	8	-2.9544
80	19.8	6	-5.0000

Cela revient à trouver a_1 , a_2 , a_3 et b tels que:

$$\mathbf{maxO3} = a_1 * \mathbf{T12} + a_2 * \mathbf{Ne12} + a_3 * \mathbf{Vx12} + b$$

Algorithmes pour la régression

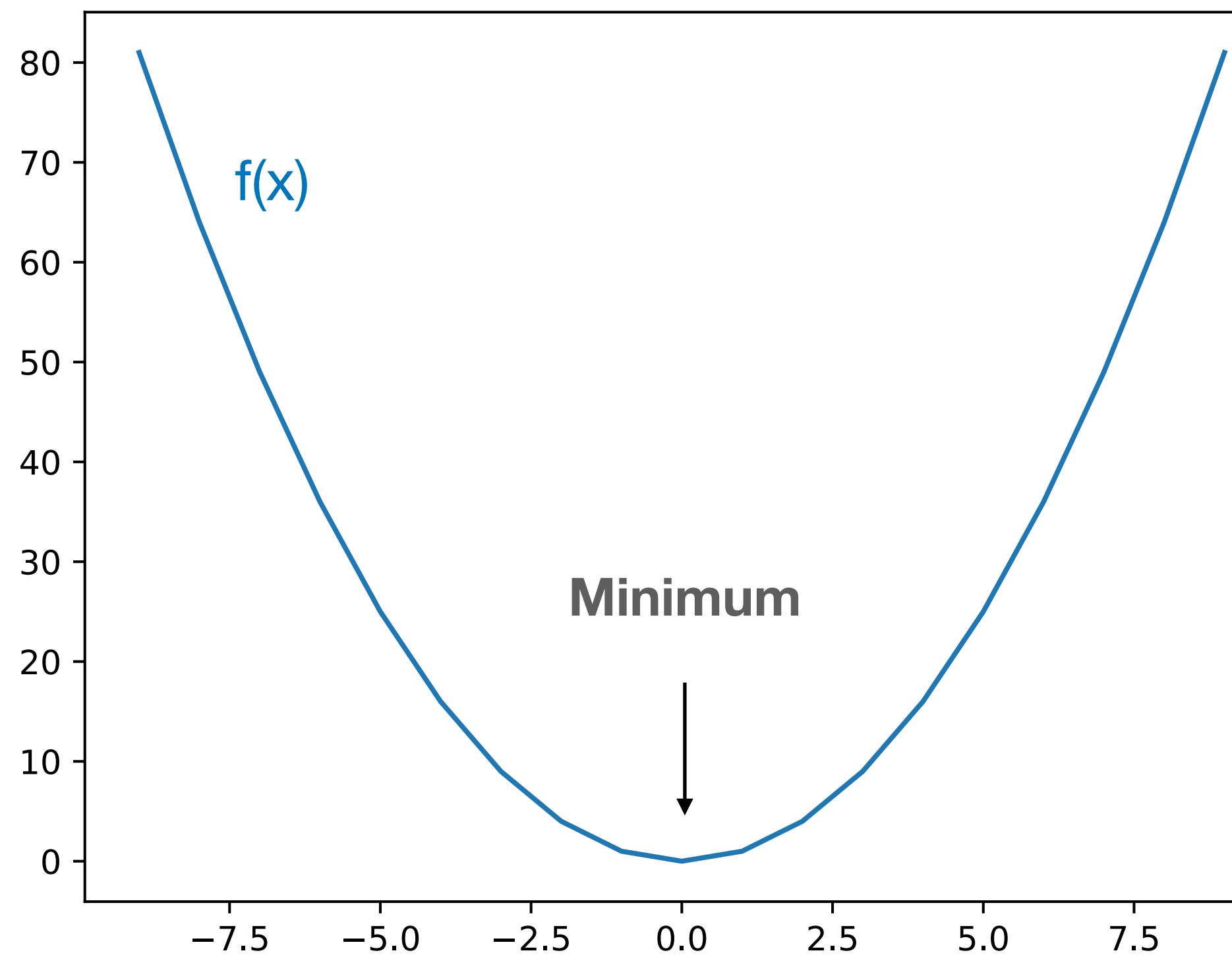
On veut trouver a et b tels que l'erreur quadratique est minimale

$$\operatorname{argmin}_{a \in \mathbb{R}, b \in \mathbb{R}} \sum_{i=1}^n (y_i - (aX_i + b))^2$$

- Sous certaines conditions sur la matrice X , il y a une solution explicite: c'est la **méthode des moindres carrés** (Legendre et Gauss, XIXe siècle).
- C'est assez calculatoire, on va se focaliser sur un algorithme central qui est la base de beaucoup d'algorithmes plus compliqué en apprentissage statistique, en IA... l'**algorithme du gradient**.

Algorithme du gradient

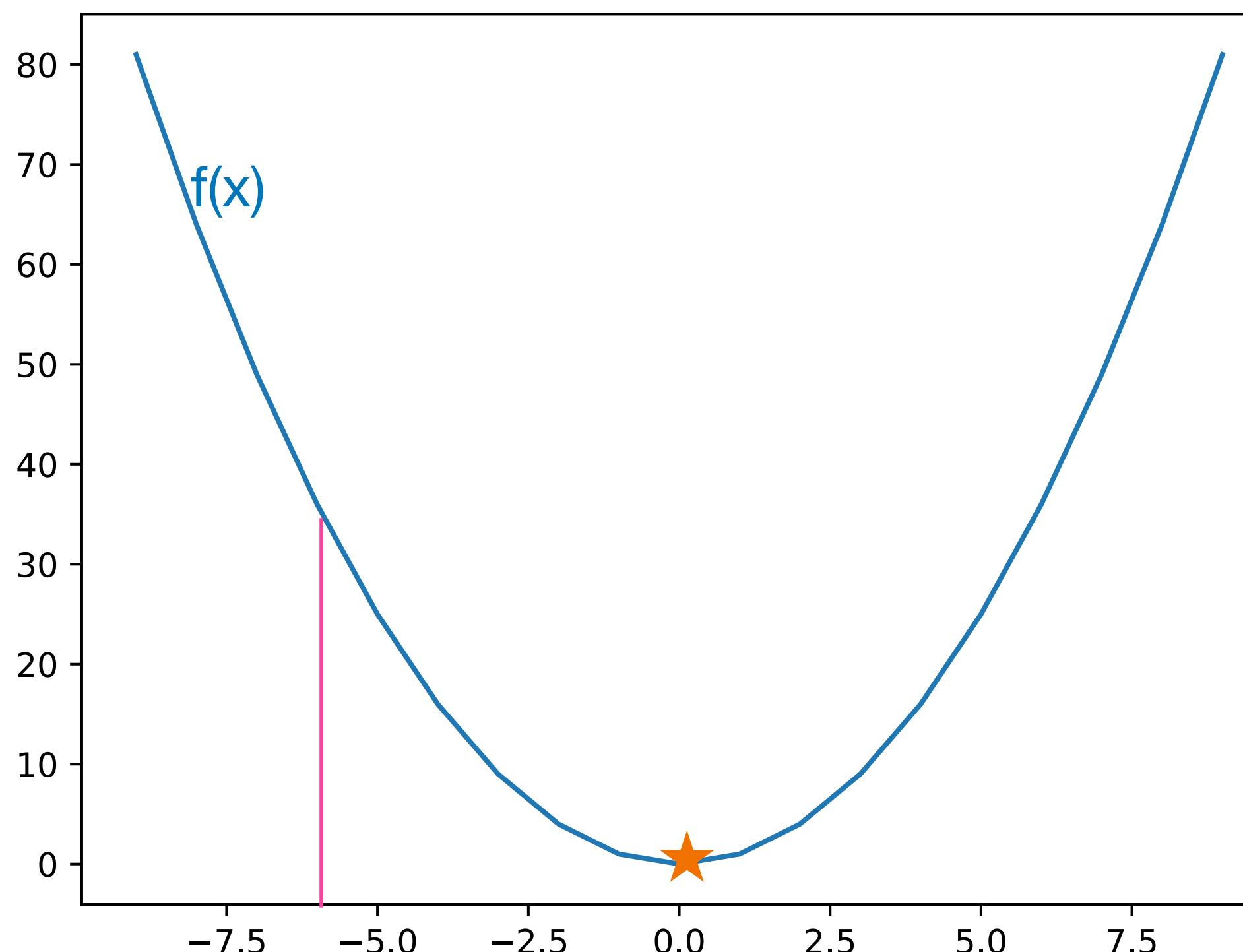
- Algorithme classique permettant de trouver le minimum d'une fonction $f(x)$



Cas facile: la fonction est ici dite convexe

Algorithme du gradient: détails

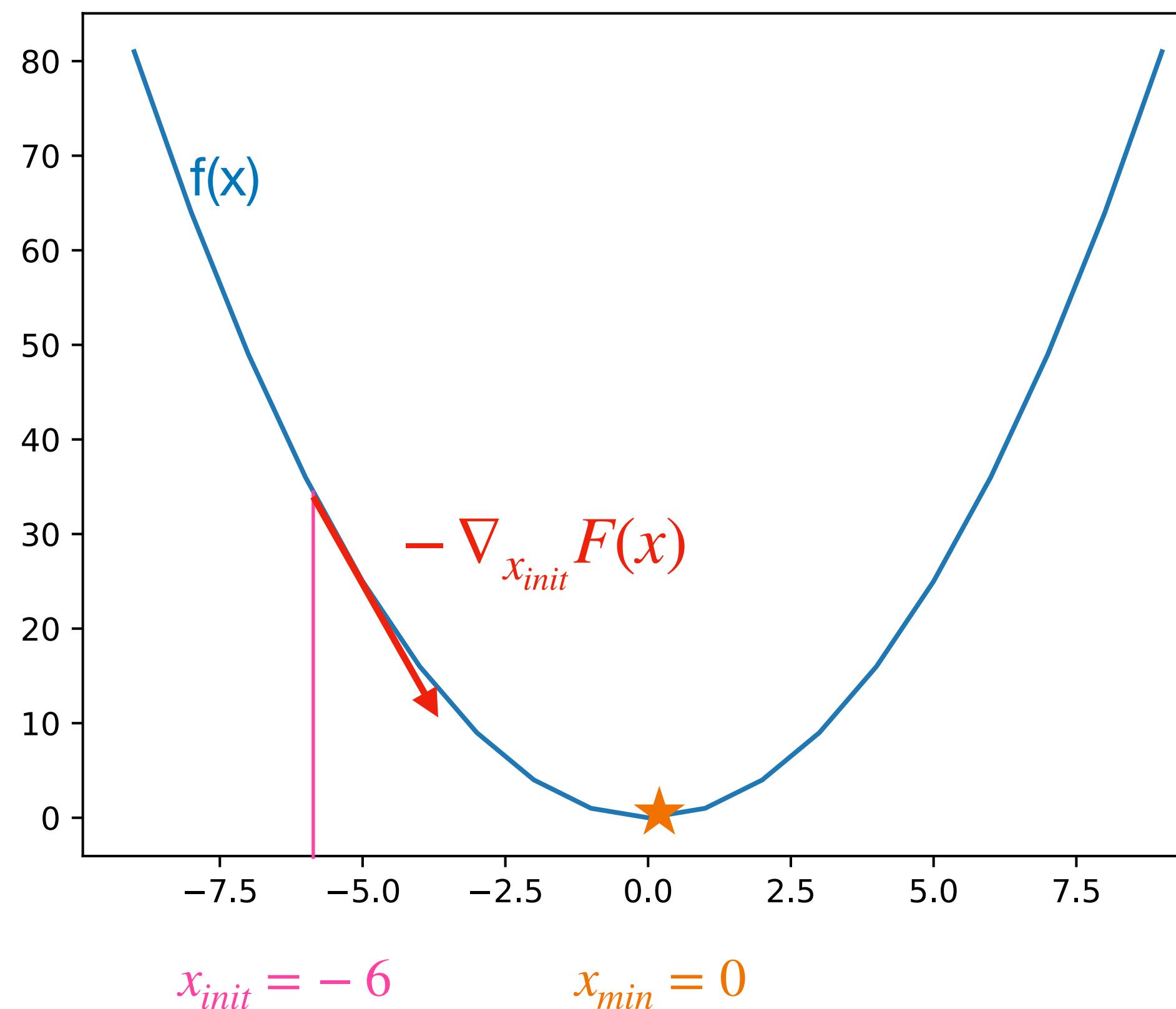
- **Initialisation:** on choisit une solution aléatoire $x_{init} = -6$



$$x_{init} = -6$$

$$x_{min} = 0$$

- Le **gradient d'une fonction** pointe vers dans la direction où la fonction croît le plus rapidement.
- Pour **minimiser** la fonction, on se déplace donc vers la **direction opposée du gradient**.



- **Initialisation:** on choisit une solution aléatoire $x_{init} = -6$
- **Mise à jour:** on se déplace dans la direction opposée du gradient

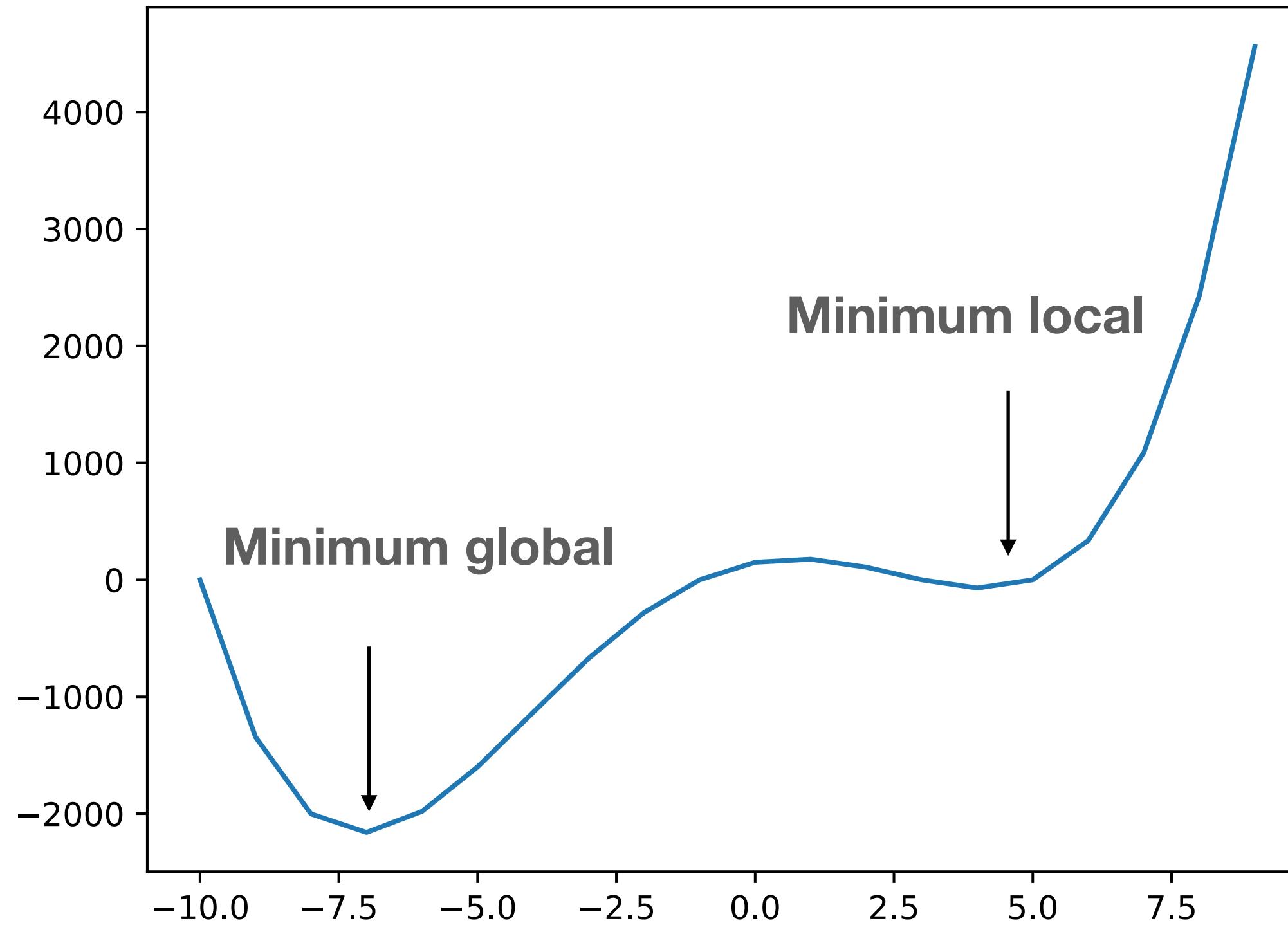
$$x_{suivant} = x_{init} - \eta \nabla F(x_{init})$$



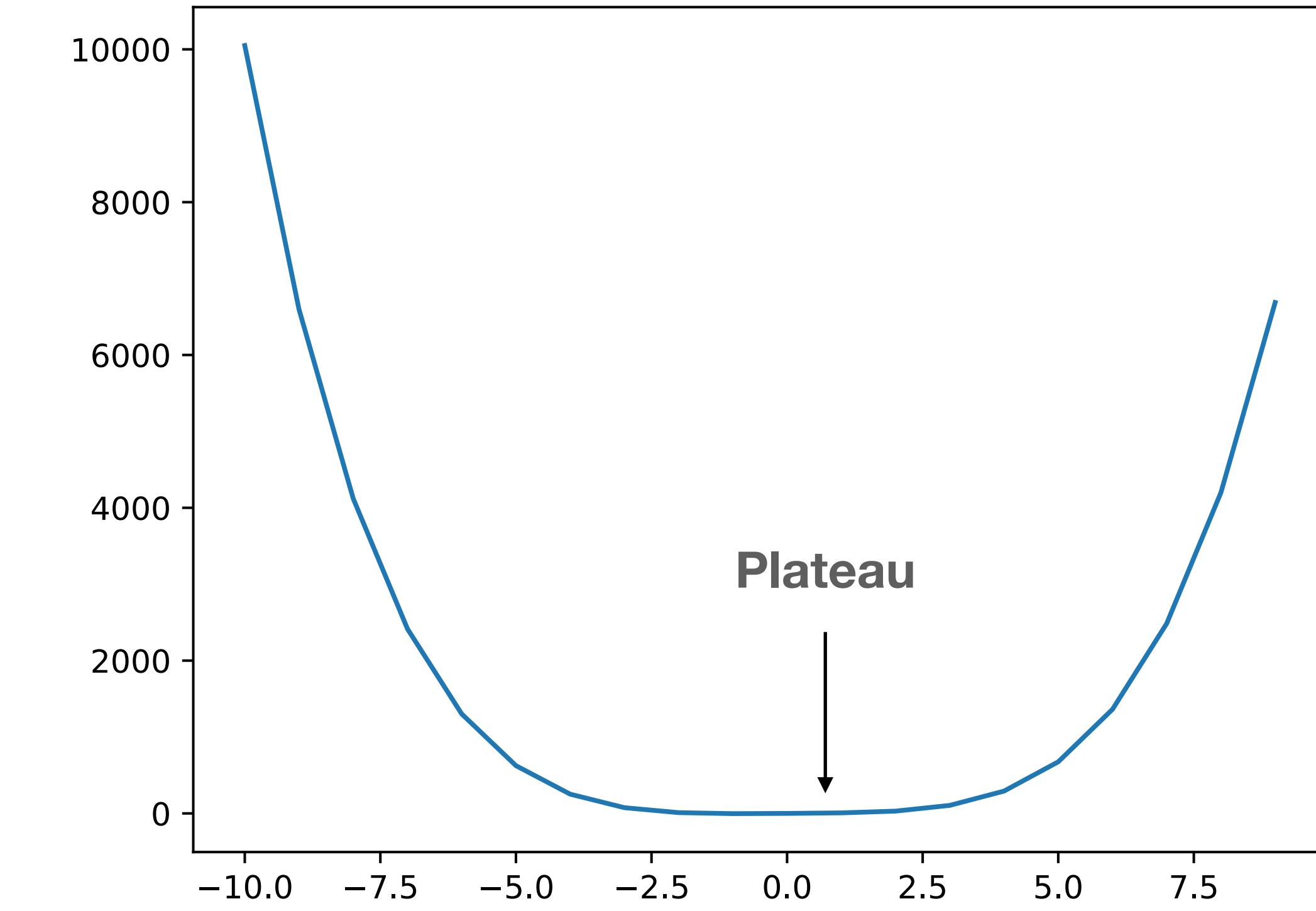
pas de l'algorithme du gradient

- On recommence jusqu'à ce qu'un nombre d'itérations soit atteint (d'autres critères d'arrêts moins naïfs sont possibles).
- **Pas de l'algorithme du gradient:** c'est un hyperparamètre (à choisir par l'utilisateur).
 - Si on le choisit trop grand, plus l'algorithme va aller vite vers le minimum (mais il va peut-être le rater).
 - Si on le choisit trop petit, l'algorithme ne va pas assez vite.

Deux exemples où c'est plus compliqué



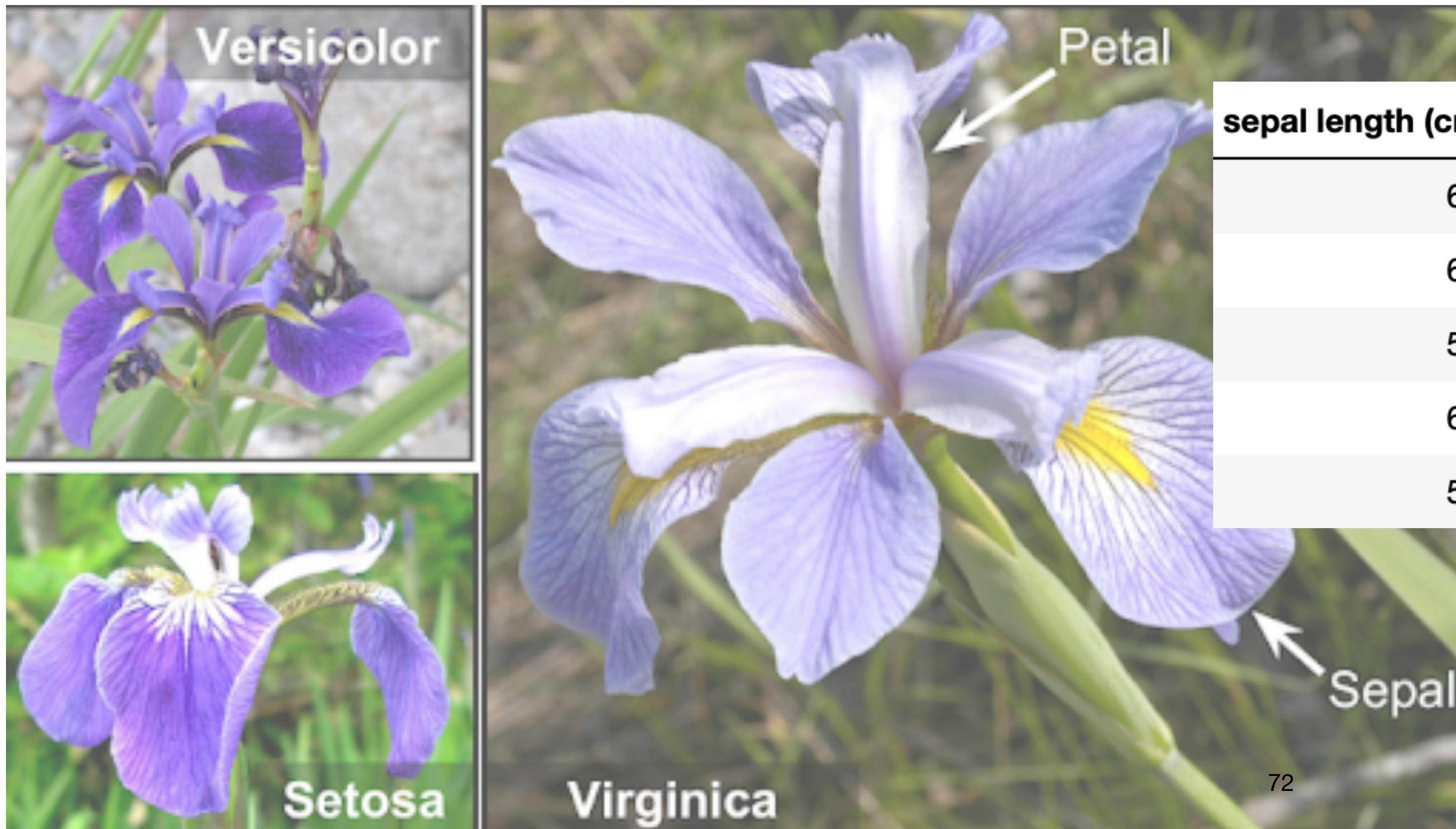
Attention au minimum local



Attention au plateau

Un exemple de classification

On veut prédire la classe « espèces » (species) qui est soit: versicolor, virginica ou setosa.



sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
6.1	2.8	4.7	1.2	versicolor
6.5	3.0	5.2	2.0	virginica
5.0	3.4	1.6	0.4	setosa
6.4	2.8	5.6	2.1	virginica
5.1	3.8	1.6	0.2	setosa

Un exemple de classification

- Comme la régression linéaire, on pourrait vouloir écrire:

species = a1*sepal_length+a2*sepal_width+a3*petal_length+a4*petal_width+b



 Le résultat ne sera pas catégoriel

- Mais on ne peut pas: la variable **species** est **catégorielle** et ne prend que 3 valeurs: **versicolor**, **virginica** ou **setosa**.
- Pour chaque observation, on va prédire une **probabilité d'appartenance à la classe**: **versicolor**, **virginica** ou **setosa**.

Un exemple de classification

- Pour chaque observation, on va prédire une **probabilité d'appartenance à la classe**: **versicolor**, **virginica** ou **setosa**.
 - La somme des probabilités doit sommer à 1, car l'observation est forcément dans une de ces classes.
 - On choisira la classe qui a la plus grande probabilité.
- **Comment transformer un nombre réel en probabilité ?**

$a1 * \text{sepal_length} + a2 * \text{sepal_width} + a3 * \text{petal_length} + a4 * \text{petal_width} + b$

Transformer un nombre en une probabilité: lien Logit

Cas de deux classes

On simplifie le problème: l'espèce peut être setosa ou virginica

- **z = combinaison linéaire des variables explicatives:**

$$z = a_1 * \text{sepal_length} + a_2 * \text{sepal_width} + a_3 * \text{petal_length} + a_4 * \text{petal_length} + b$$

$$\bullet \quad \mathbb{P}(y = \text{setosa} | X) = \frac{e^z}{1 + e^z}$$

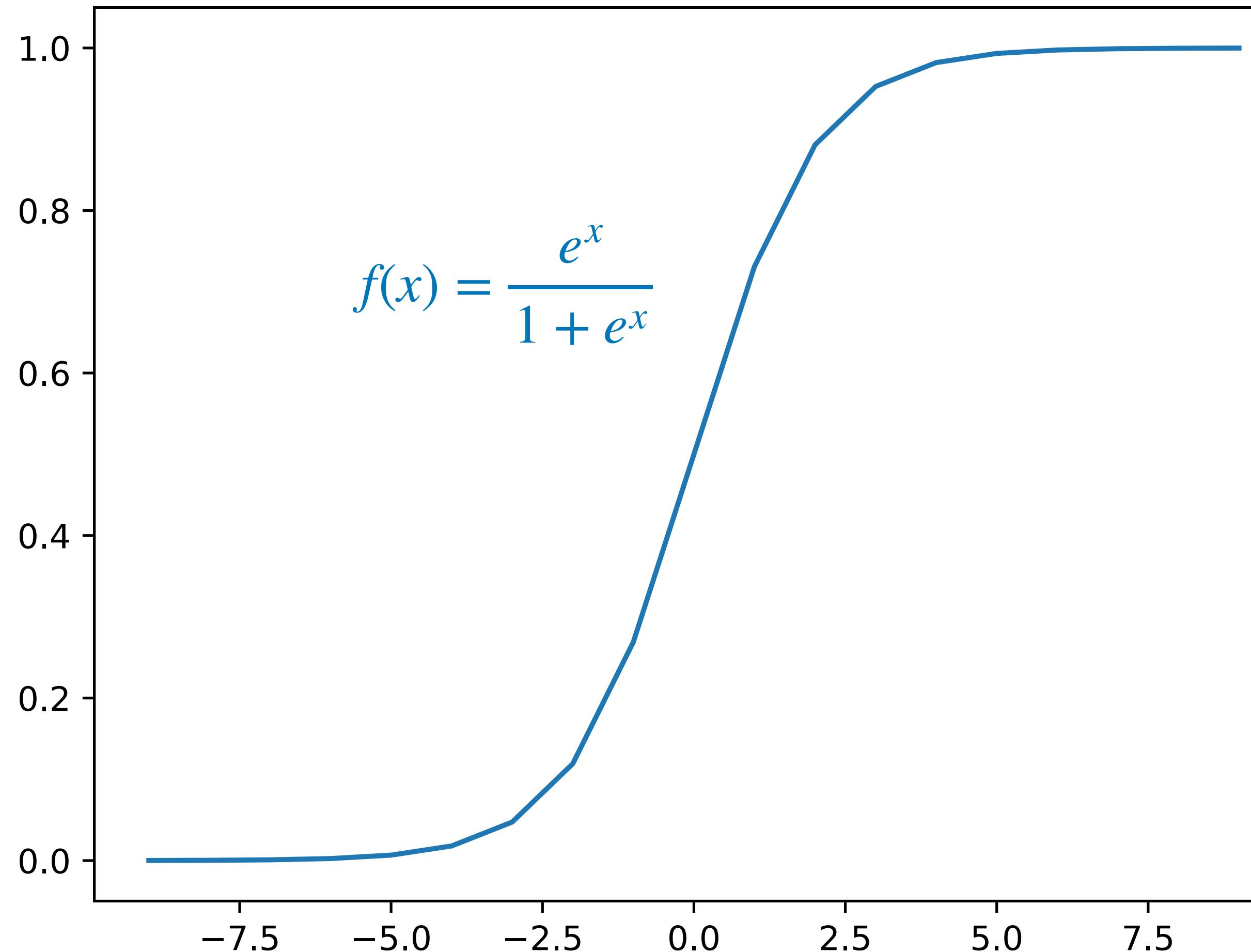
$$\bullet \quad \mathbb{P}(y = \text{virginica} | X) = 1 - \frac{e^z}{1 + e^z}$$

Rappels

- Une probabilité est un nombre compris entre 0 et 1.
- Une probabilité = 1: cela veut dire que l'évènement est sûr.

Même technique de passage à une *fonction de lien* pour le cas de plus de deux classes

Transformer un nombre en une probabilité: lien Logit



Exemple de mesure de performance: la matrice de confusion

Quelles observations / combien sont mal classées ?

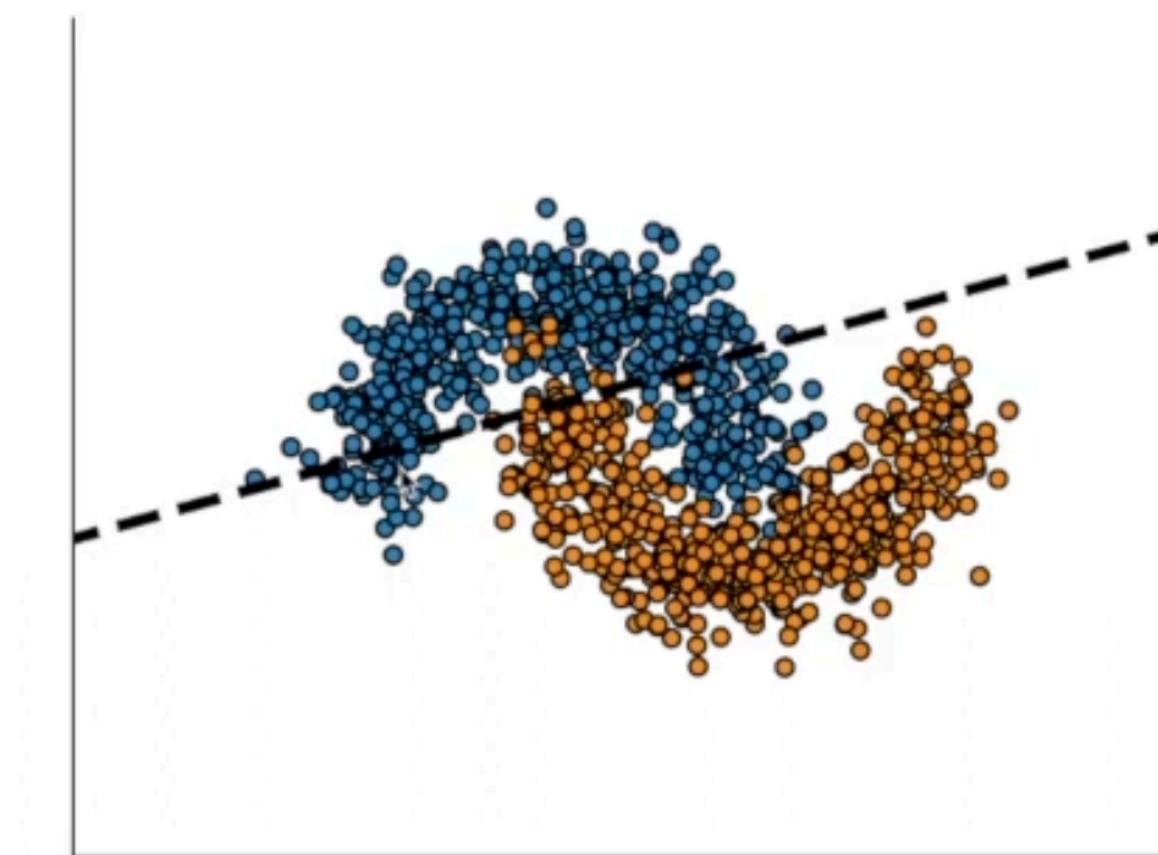
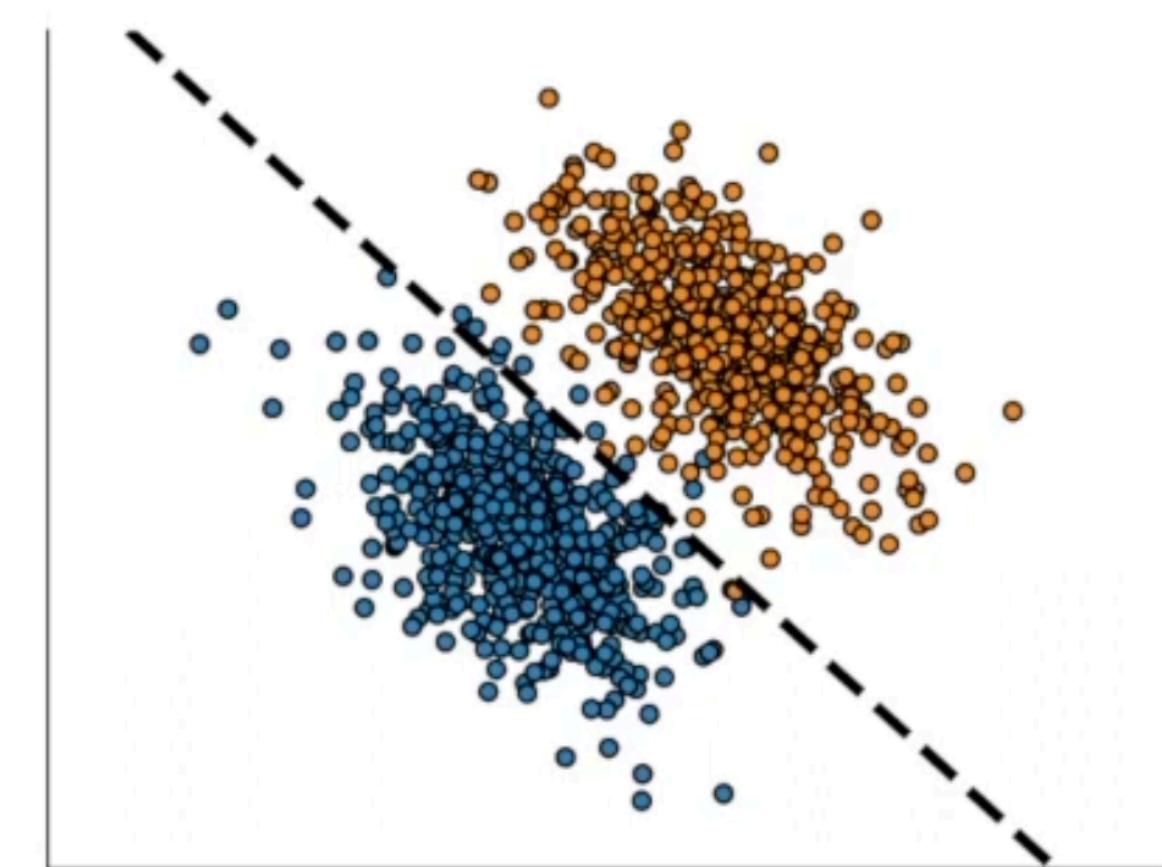
	Setosa	Versicolor	Virginica
Setosa	29	2	0
Versicolor	0	30	0
Virginica	0	0	23

29 observations ont été bien classées; 2 ont été classées en « versicolor » alors qu'elles sont des « setosa »



Limitation des modèles linéaires

- Modèles linéaires: ils utilisent des lignes droites (ou des généralisation de lignes droites en dimension supérieur) pour séparer les données.
- Limitation: **ces modèles ne peuvent pas séparer des données qui ne sont pas linéairement séparables**, ce qui arrive souvent en pratique ! En fait, si on utilise un modèle linéaire dans ce cas, on est dans un cas de sous-apprentissage, notre modèle n'est pas assez complexe pour les données.



Source: [scikit-learn](#)

6. TP: modèles linéaires

https://mybinder.org/v2/gh/AudeSportisse/Efelia-cours/HEAD?labpath=TP1_linear_models.ipynb

Prochaine séance

1. Programme prévisionnel:

- Forêts aléatoires
- Réduction de dimension
- Réseaux de neurones

2. Choisir son exposé

Bibliographie (liens cliquables)

- Mooc de [scikit-learn](#), développeurs scikit-learn / Inria, 2022
- [Introduction à l'intelligence artificielle et aux modèles génératifs](#), **Pierre-Alexandre Mattei et Serena Villata**, chercheur Inria, chercheuse CNRS, Université Côte d'Azur, 2022
- *Jusqu'où ira l'intelligence artificielle ?*, Pour la Science, Hors-série numéro 115, 2022
- Cours, [Intelligence Artificielle](#), **Chloé-Agathe Azencott**, enseignante-chercheuse au Center for Computational Biology (CBIO), 2021
- *Les maths et le réel, comment décoder le monde*, La Recherche hors-série n° 31, septembre-octobre 2019
- *Formation des médiateur-rites de la MIA*, **Lucile Sassatelli**, professeure des Universités en informatique, 2023
- Cours, [Chapitre 1: Introduction à l'intelligence artificielle](#), **Meghyn Bienvenu**, chercheuse CNRS, Université de Bordeaux, 2006-2008
- Cours, [Intelligence artificielle, définition, généralités, historique, domaines](#), **Dominique Pastre**, professeur d'informatique honoraire, ENS, Université Paris 5, 1999-2000