Introduction à l'intelligence artificielle

Anna Bonnet

Laboratoire de Probabilités, Statistique et Modélisation, Sorbonne Université

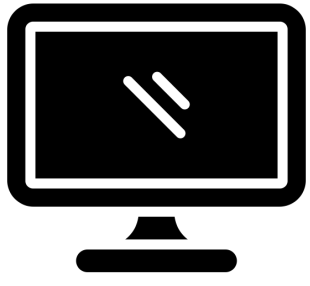
Ecole d'été en IA, SCAI, Sorbonne Université
3 juillet 2023

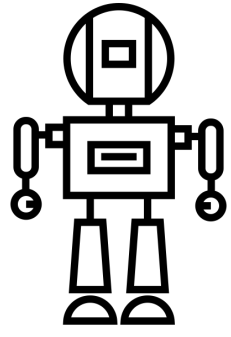
Qu'est-ce que l'intelligence artificielle ?

"Ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine" (Larousse, version actuelle)

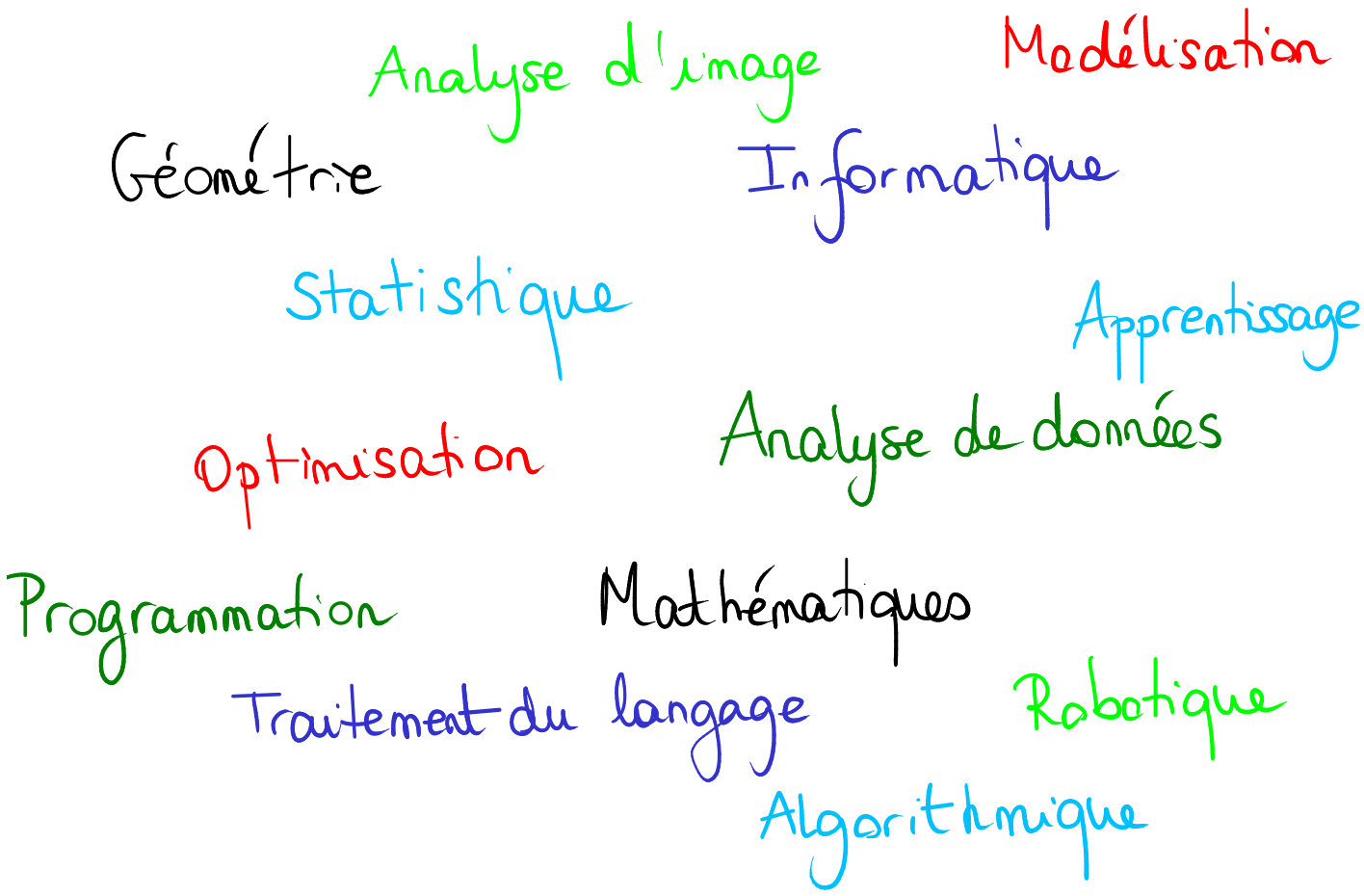
"Discipline étudiant la possibilité de faire exécuter par l'ordinateur des tâches pour lesquelles l'homme est aujourd'hui meilleur que la machine" (Rich et Knight, 1990)

"Automatisation des activités associées au raisonnement humain, telles que la décision, la résolution de problèmes, l'apprentissage" (Bellman, 1978)

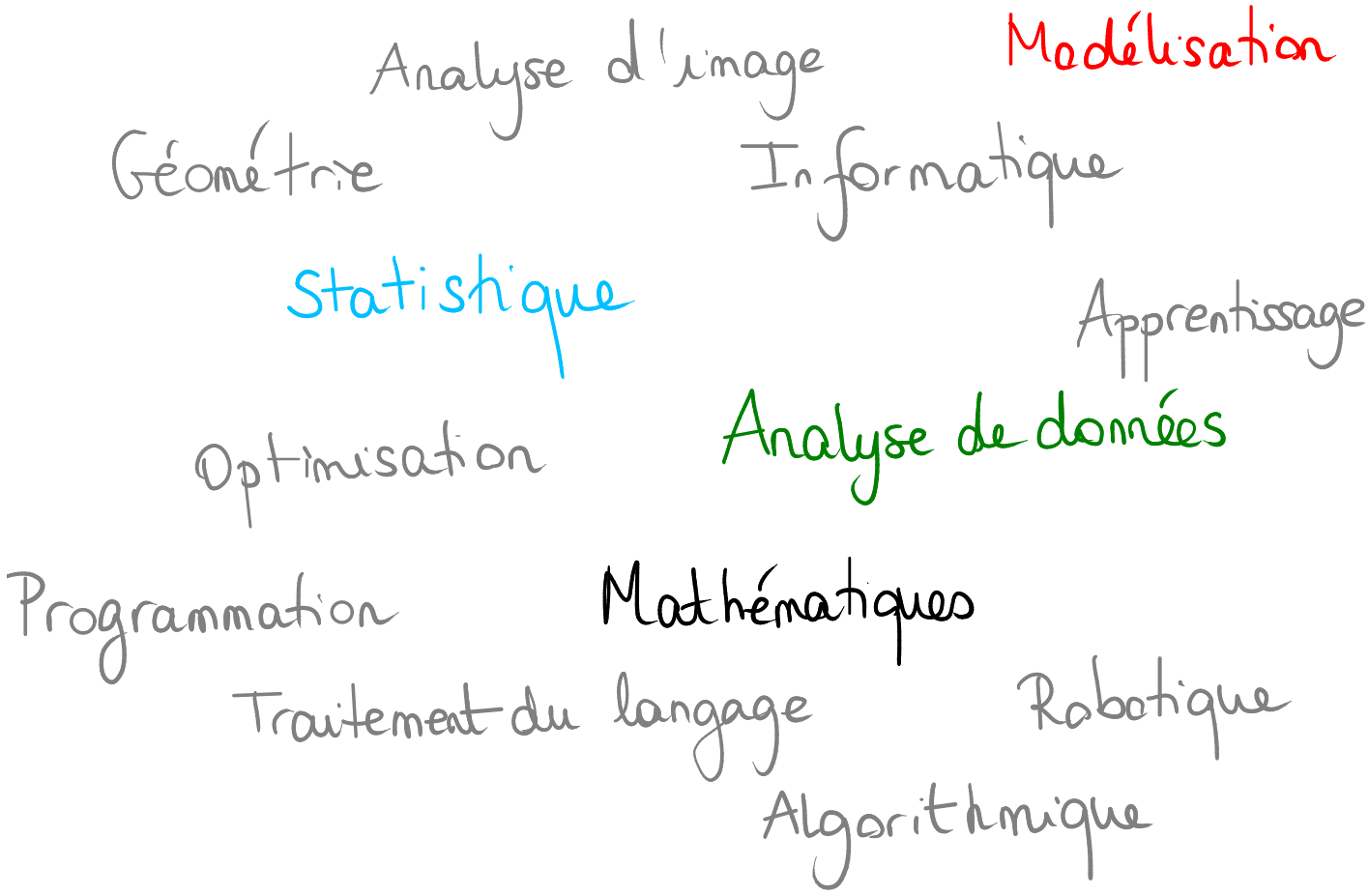




Domaines de l'IA



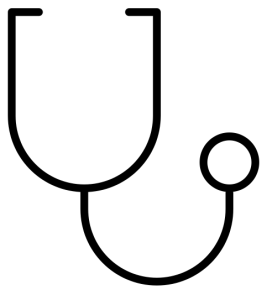
Domaines de l'IA



Etapes du développement d'une méthode d'IA

- Récolte des données
- Mise en forme des données
- Etape de modélisation
- Développement d'une nouvelle méthode
  - Implémentation de la méthode
  - Analyse théorique
  - Optimisation numérique
- Interprétation des résultats



1. Exemple de tâches qu'on peut réaliser avec une IA

2. Histoire et grandes avancées

3. Mise en pratique : comment modéliser un problème en biologie ?

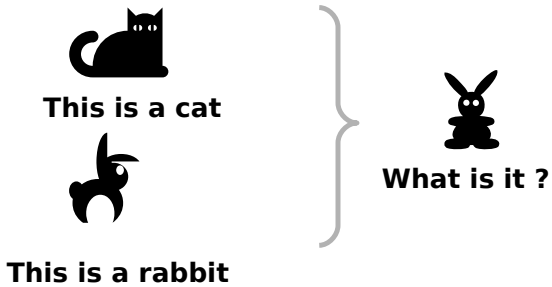
Plan

Apprentissage supervisé : classification

Idée

Apprendre à réaliser une tâche grâce à des exemples, par exemple prédire la valeur d'une observation Y en fonction de caractéristiques X

- Si Y est qualitatif, on parle de classification

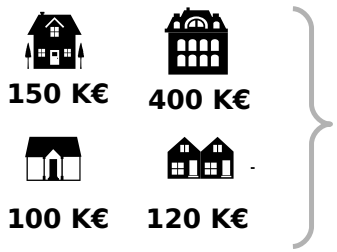


Y=chat/lapin, X=photo

- Deux étapes : apprentissage sur des données et test sur de nouvelles observations
- Objectif : maximiser la probabilité de prédire le bon label

Apprentissage supervisé : régression

- Si Y est quantitatif, on parle de régression



Y=prix, X=localisation, surface...

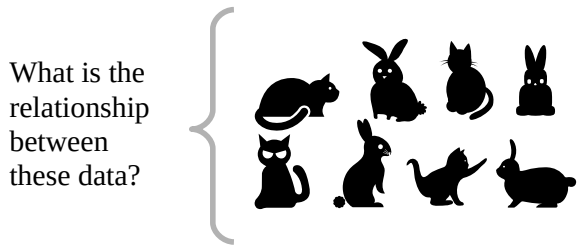
- Deux étapes : apprentissage sur des données et test sur de nouvelles observations
- Objectif : minimiser l'écart entre la valeur prédite et la vraie

Apprentissage non supervisé

Idée

Apprendre des caractéristiques sur des données non labellisées

- Clustering : identifier des caractéristiques communes



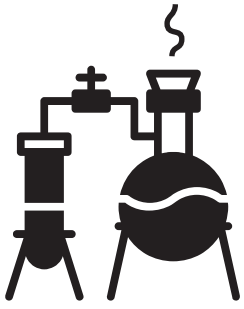
- Réduction de dimension : réduire la taille des données en conservant les principales caractéristiques

Can we simplify the data while
keeping its meaning?

Plan

Histoire (courte) : statistique et IA

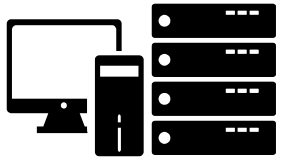
Experimental science



Theoretical science

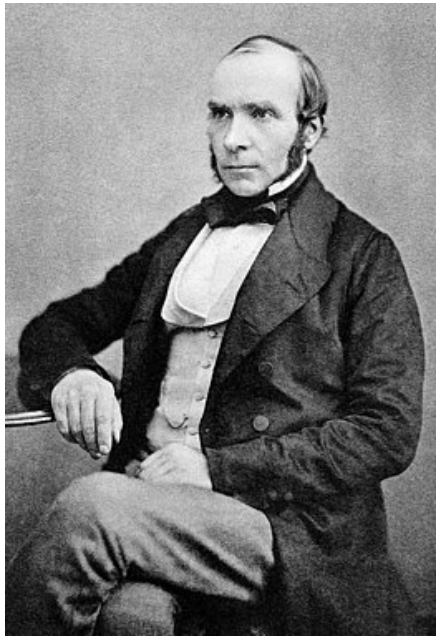
$$\nabla \times H = J + \frac{\partial D}{\partial t}$$

Computational science



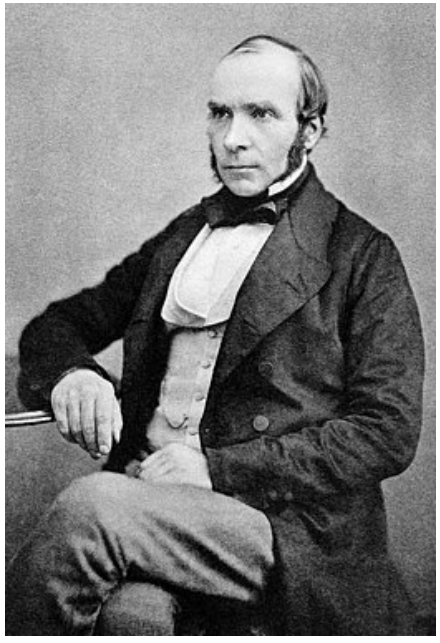
⋆ L'IA moderne est le produit de grandes évolutions en sciences expérimentales, théoriques et numériques.

Jusqu'au XIXe siècle : Statistique descriptive



John Snow (1813-1858)

Jusqu'au XIXe siècle : Statistique descriptive



John Snow (1813-1858)

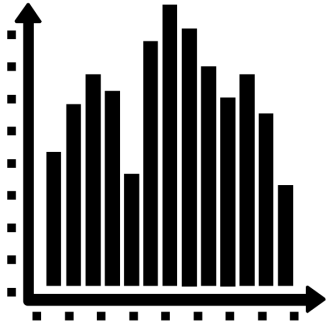
Jon Snow (283- ?)

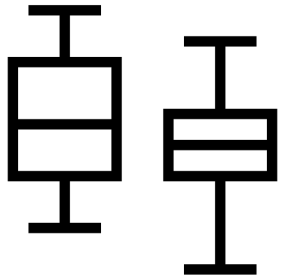
- Collecte des cas de choléra (1849) pendant une épidémie
- Représentation du nombre de cas sur une carte
- Identification de la propagation de la bactérie dans les eaux usées
- Pionnier en épidémiologie et statistique spatiale

Visualisation des données

Florence Nightingale (1820-1910)

- Infirmière et pionnière dans l'utilisation des statistique en santé
- Développement d'outils visuels pour représenter les données



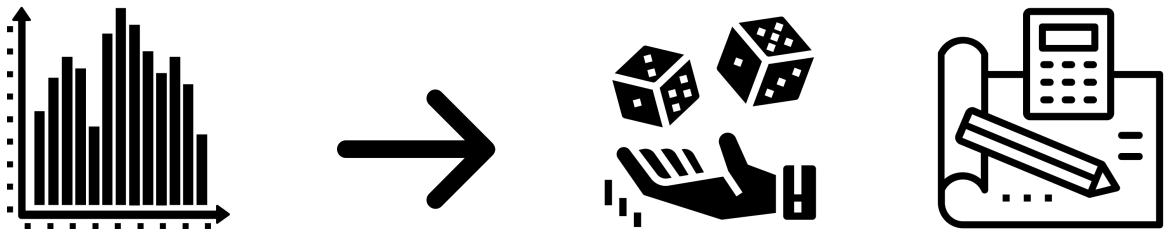


XIXe siècle - Début XXe : Modélisation statistique

Plusieurs grands noms : Bayes, Laplace, Gauss, Pearson, Fisher...

Idée

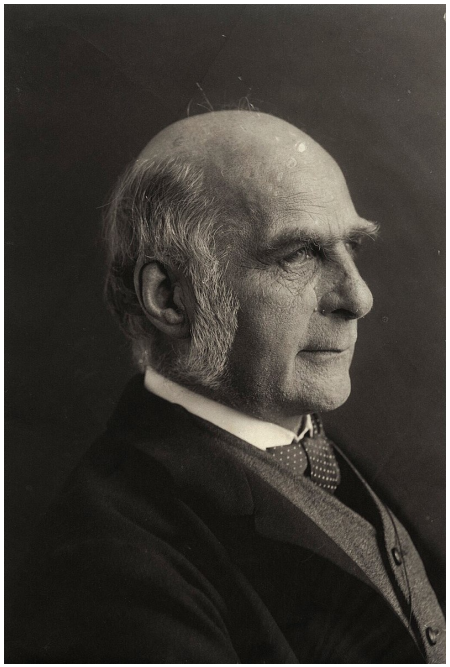
- Observation de données
- Proposer un modèle mathématique qui explique d'où viennent les données
- Utilisation de la théorie des probabilités pour comprendre les observations



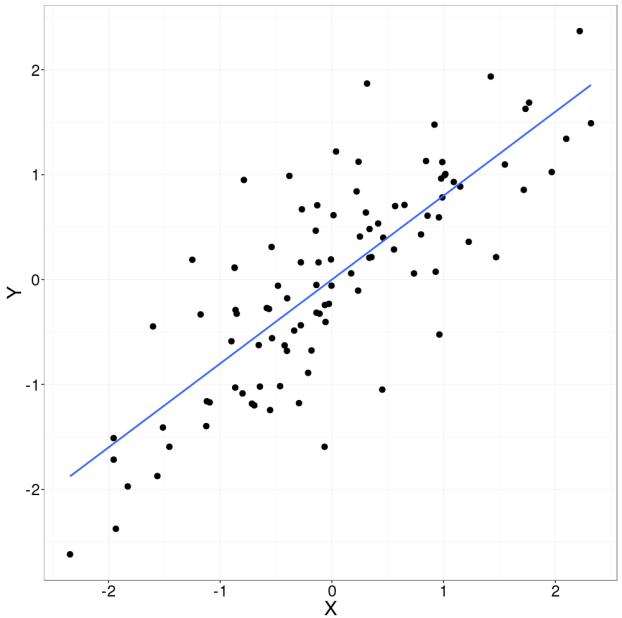
Travaux de Galton sur l'hérédité

Idée

Etudier le lien entre la taille des pères X et la taille de leurs fils Y



Francis Galton (1822-1911)



Objectif

Trouver les "meilleures" valeurs de a et b telles que $Y \simeq aX + b$

Début XXe : Fisher et la significativité

Expérience : The Lady testing tea

L'expérience fournit à un sujet 8 tasses de thé classées au hasard : 4 préparées en versant d'abord le thé, puis en ajoutant du lait, 4 préparées en versant d'abord le lait, puis en ajoutant le thé. Le sujet doit déterminer dans quelles tasses on a versé le lait d'abord.

Si on désigne au hasard 4 tasses :

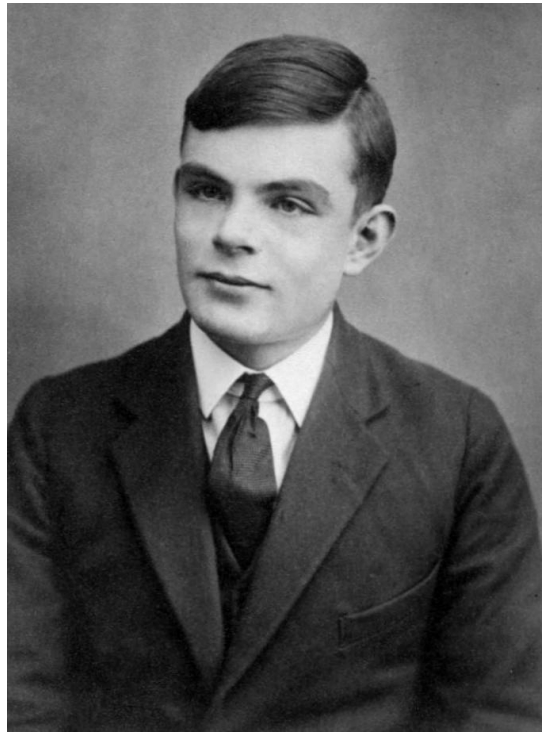
Nb de succès	Combinaisons de sélection	Nb de combinaisons
0	oooo	$1 \times 1 = 1$
1	ooox, ooxo, oxoo, xooo	$4 \times 4 = 16$
2	ooxx, oxox, oxxo, xoxo, xxoo, xoox	$6 \times 6 = 36$
3	oxxx, xoxx, xxox, xxxo	$4 \times 4 = 16$
4	xxxx	$1 \times 1 = 1$

Conclusion

Au total il y a 70 combinaisons, donc 1 chance sur 70 $\simeq 1.4\%$ de chances de trouver la bonne au hasard !

Les début de l'informatique

- 2nde guerre mondiale : décryptage d'Enigma
- 1945-1950 : travail sur les premiers ordinateurs
- Années 1950 : Test de Turing
- 1952 : Programme jeu d'échecs



Alan Turing (1912-1954)

"D'ici cinquante ans, il n'y aura plus moyen de distinguer les réponses données par un homme ou un ordinateur, et ce sur n'importe quel sujet."

Les débuts de l'informatique

- Projet Manhattan (bombe atomique)
- Architecture des ordinateurs
- Contributions dans de nombreux domaines (mécanique quantique, logique, théorie des jeux...)

John Von Neumann
(1903-1957)

Naissance de l'IA : connexionnisme vs symbolisme

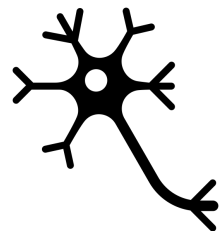
Années 50

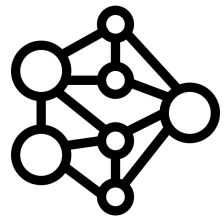
Naissance de l'IA (le terme apparaît en 1956) avec une opposition de 2 courants

- **Symbolisme** : recréer la pensée grâce à des liens logiques
- "Penser, c'est calculer des symboles qui ont à la fois une réalité matérielle et une valeur sémantique de représentation"

Tout homme est mortel
Socrate est un homme
Donc Socrate est mortel.

- **Connexionnisme** : modéliser le fonctionnement du cerveau
- Inspiré des travaux de McCulloch et Pitt (1943) : premier concept de réseau de neurones artificiels
- "Penser s'apparente à un calcul massivement parallèle de fonctions élémentaires"





Premier réseau de neurones (Rosenblatt, 1957)

$$X = (X_1, X_2)$$
$$= (\text{heures sommeil, heures travail})$$
$$Y = (\text{succès, échec})$$
$$= (1, 0)$$

$$(X_1, X_2) \rightarrow \text{Calcul de } \hat{Y} = w_1 X_1 + w_2 X_2 \begin{cases} \nearrow \text{ si } \hat{Y} \geq 0.5 \text{ alors } \tilde{Y} = 1 \\ \searrow \text{ si } \hat{Y} < 0.5 \text{ alors } \tilde{Y} = 0 \end{cases}$$

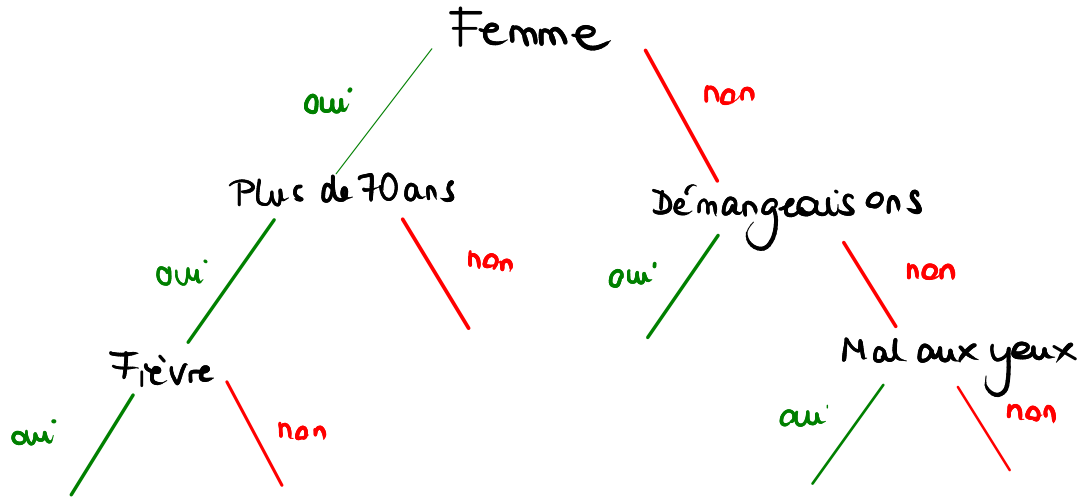
Objectif

Calculer les poids w_1 et w_2 tels que $\tilde{Y}$ soit le plus proche possible du vrai Y.

On vient de construire un réseau à 1 neurone : le perceptron !

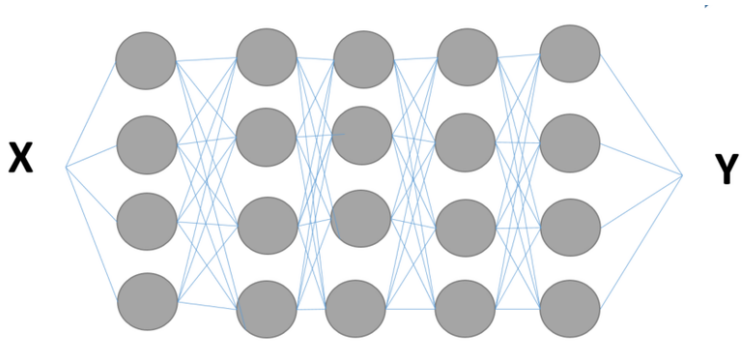
Années 60-70

- Années 60 : premières déceptions
  - Côté symbolisme : malgré des financements très importants, les progrès se limitent à la résolution de jeux assez simples (puzzles, dames...)
  - Côté connexionnisme : limites dues à la faible puissance de calcul des ordinateurs
- Années 70-80 : Succès des symbolistes avec les **systèmes experts**, arbre de décision pour l'assistance médicale.



IA moderne : le retour des réseaux de neurones

- Années 2000 : retour en force des réseaux de neurones, avec les mêmes concepts mais avec des machines beaucoup plus puissantes
- Beaucoup plus de neurones, fonctions beaucoup plus compliquées

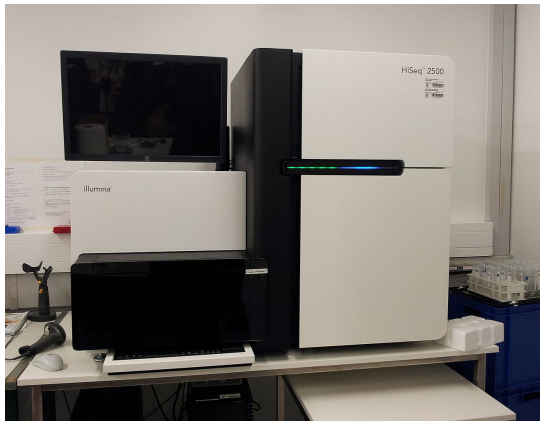


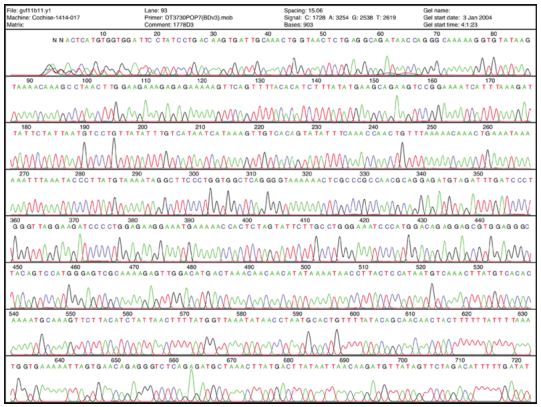
- Gros moyens de calcul, données massives accessibles (big data)
- Temps de calcul, coût écologique
- Décalage entre les performances numériques et la compréhension des modèles, question de l'**interprétabilité** des résultats

Plan

Pourquoi a t-on besoin de modèles mathématiques en biologie ?

- Grands progrès techniques en biologie, qui donnent accès à une masse de données



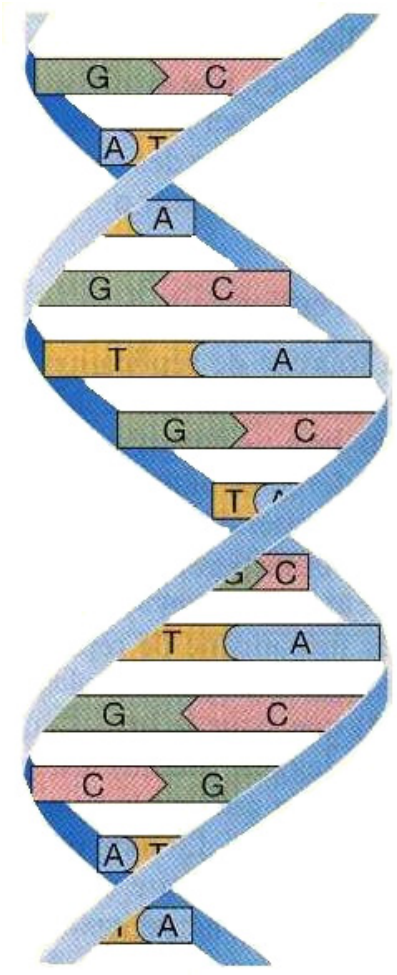


- Que faire de ces données ? Comment chercher (et trouver) l'information pertinente ?

- Mise en place de modèles mathématiques, descriptifs ou prédictifs

Qu'est-ce que l'ADN ? Quelle information contient-il ?

- Présent dans toutes nos cellules
- Contient l'information génétique (génome) permettant le développement, le fonctionnement et la reproduction des êtres vivants
- Molécule en forme de double hélice, composée de 2 brins **complémentaires**
- Chaque brin est composé d'un enchaînement de bases A, T, G, C
- L'ordre dans lequel se succèdent les bases définit une **séquence**
- **Séquençage** : technique qui permet de déterminer une séquence d'ADN



Coût du séquençage

- **2003** : Premier génome humain séquencé (3 milliards de paires de bases, 10 ans + 2.7 milliards de dollars !)
- **Aujourd'hui** : 100 dollars en 24h !!

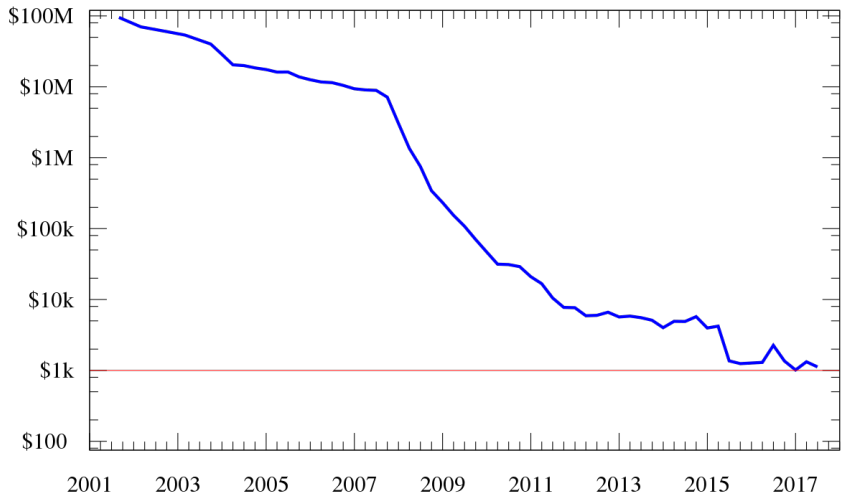


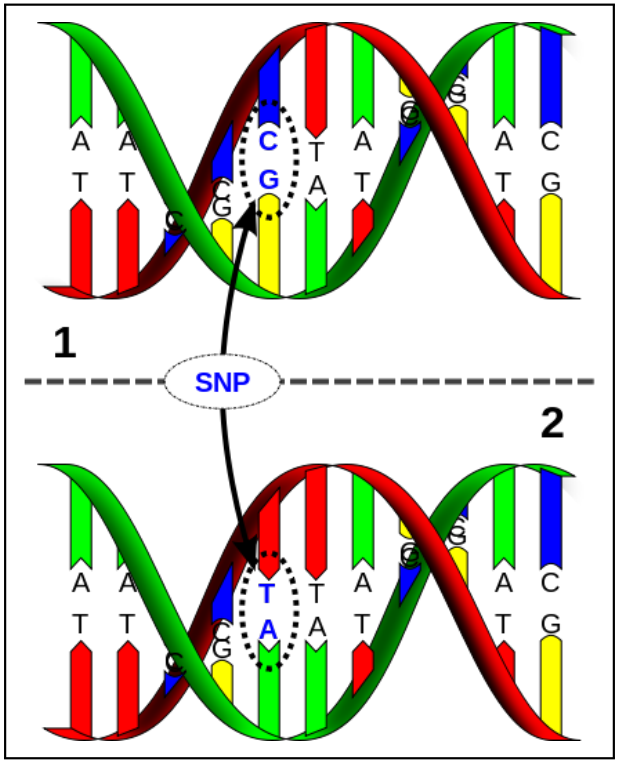
Figure: Coût (en dollars) du séquençage d'un génome humain (source : National Human Genome Research Institute)

Question

Que fait-on de tous ces génomes séquencés ?

Variations de séquence entre les individus

- 2 séquences humaines se ressemblent à plus de 99%
- **Single Nucleotide Polymorphism (SNP)** : changement d'une base
- A l'origine des différences entre individus d'une même espèce
- Chez l'humain : quelques millions de SNPs



Quels sont les effets de ces variations de séquence entre les individus ?
Comment les étudie-t-on ?

Etudes d'association

	malade/sain	snp_1	snp_2	...	snp_p	Age	Fumeur
$i = 1$	0	A	G	...	A	38	oui
$i = 2$	1	A	A	...	C	15	non
$\vdots$		$\vdots$	$\vdots$		$\vdots$		
$i = N$	1	T	G		G	90	non

Pour chaque individu :

- Malade (1) ou sain (0)
- Variants génétiques mesurés sur p SNPs
- Données cliniques (non génétiques)

Question

Est-ce que certains variants (lesquels) sont souvent (tout le temps ?) présents chez les individus malades ?

Stratégie univariée : SNP par SNP

Principe

Pour chaque SNP, on compte la proportion de A/T et de G/C chez les individus sains et chez les individus malades, et on les compare.

- Exemple 1 :

	A/T	C/G
Malades	70%	30%
Sains	70%	30%

- Exemple 2 :

	A/T	C/G
Malades	90%	10%
Sains	20%	80%

- Exemple 3 :

	A/T	C/G
Malades	70%	30%
Sains	60%	40%

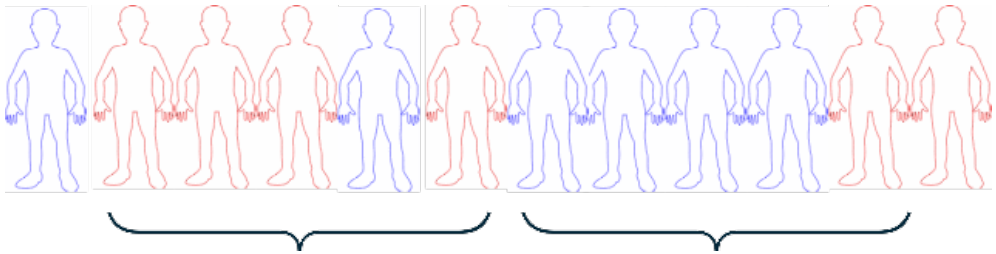
Question

Quand est-ce que la différence de proportions est assez grande pour qu'on déclare que le variant a un effet sur la maladie ?

Test de significativité (Travaux de Fisher)

Question

Dans l'exemple 3, est-ce que la différence qu'on observe est réellement due à l'effet du variant 1 ou bien **au hasard** ?



- Grâce à la théorie des probabilités, on sait déterminer un intervalle dans lequel devrait se trouver la différence entre les proportions **si elle est uniquement due au hasard**.

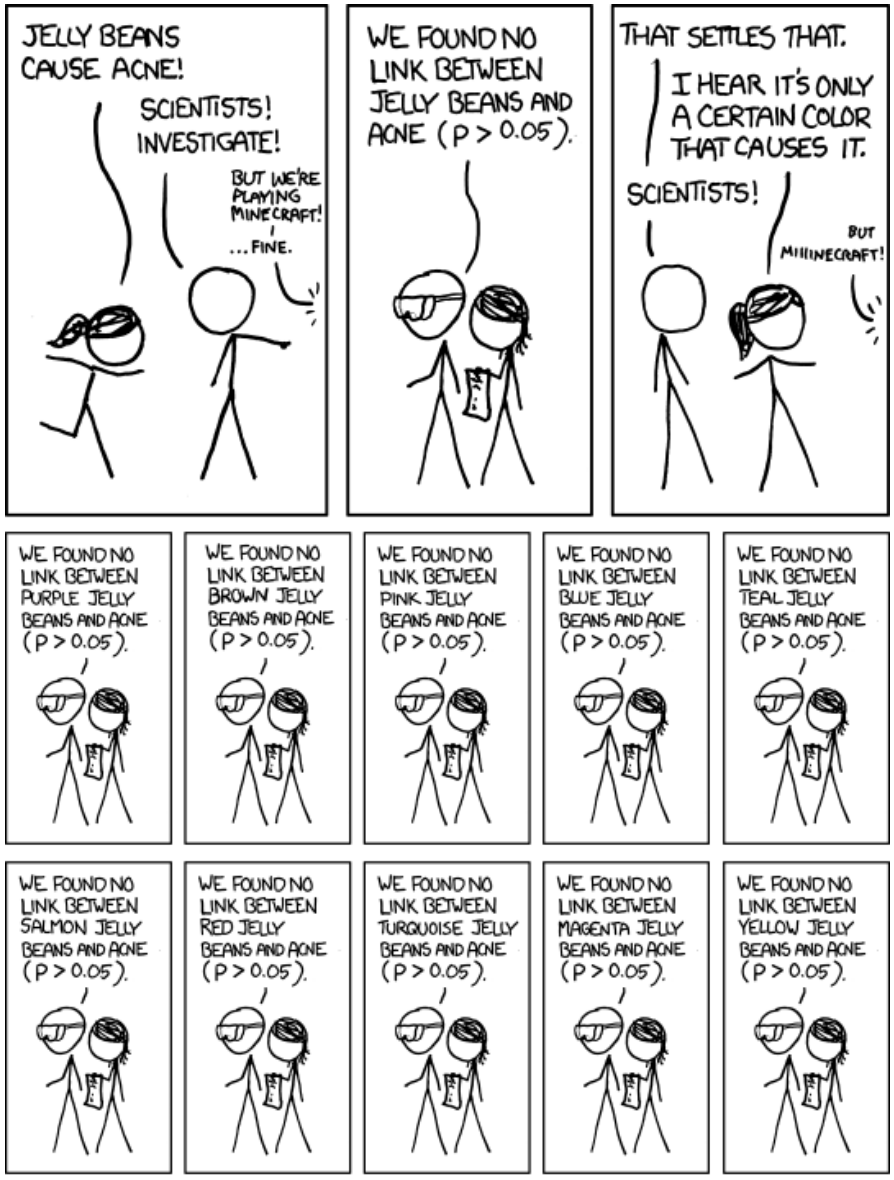
Analogie

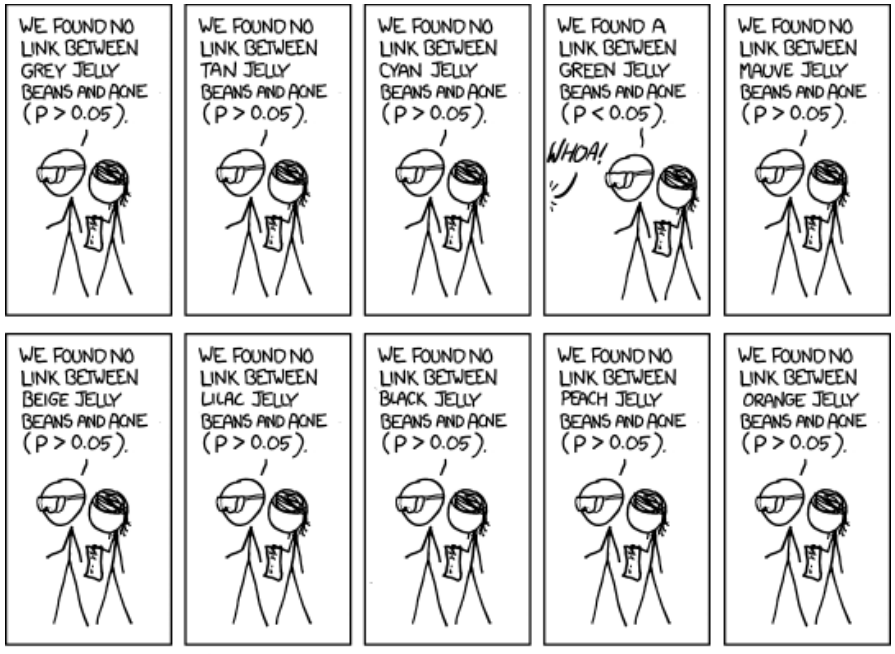
Si on lance 10 fois une pièce équilibrée, parfois on obtient 5 Pile et 5 Face, mais il est aussi très probable de tomber sur 4 Pile et 6 Face, ou l'inverse. Ces différences ne sont pas dues à un défaut de la pièce mais juste au hasard de l'expérience.

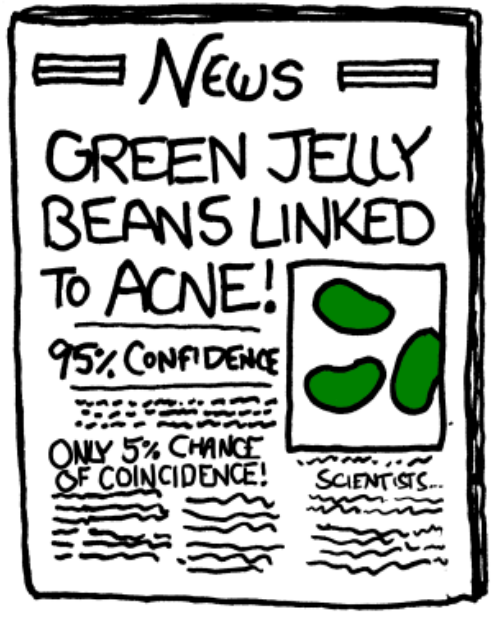
Limites de cette première stratégie

- Un test par SNP (1 million de tests à faire !)
- On va détecter uniquement les SNPs qui ont un effet très fort
- On va rater les SNPs qui doivent être associés à d'autres pour avoir un effet
- En faisant 1 million de tests on va avoir des **faux positifs** (si on lance une pièce longtemps, on va finir par observer 10 fois Pile d'affilée, ce qui est très rare si on la lance 10 fois).

Attention aux tests multiples !







Modélisation de tous les SNPs ensemble

On peut regarder l'effet joint **de tous les SNPs en même temps**, et même d'autres variables comme les donnée cliniques.

$$(X_1, \ldots, X_n) \to Z = w_1 X_1 + \cdots + w_n X_n + \epsilon \begin{cases} \nearrow \text{ si } Z \geq 0.5 \text{ alors } Y = 1 \\ \searrow \text{ si } Z < 0.5 \text{ alors } Y = 0 \end{cases}$$

- $Y =$ sain/malade, $0/1$
- $X = (X_1, \ldots, X_n)$ contient toute l'information génétique et les variables cliniques
  - ϵ correspond à tous les autres effets, on l'appelle le terme d'erreur

Objectif

Estimer $w_1, \ldots, w_n$ à partir de beaucoup d'observations (X, Y)

- Avantages : il existe une résolution simple du problème + on sait interpréter w
- Inconvénient : la modélisation est peut-être trop simplifiée (effets additifs)

2e stratégie

Nouveau modèle plus général

$$Y = f(X) + \epsilon$$

- Y = sain/malade, 0/1
- X contient toute l'information génétique et les variables cliniques
- f décrit le lien entre les deux
- ϵ le terme d'erreur

Questions

- Est-ce que $f = 0$ (pas de lien entre X et Y) ? $\rightarrow$ **test statistique**
- Sinon, que vaut f ? $\rightarrow$ **estimation**
- Comment savoir quelles variables (génétiques ou cliniques) ont un effet sur Y ? $\rightarrow$ **sélection**

Comparaison des deux stratégies

- On peut faire des **hypothèses** qui permettent de simplifier le modèle mathématique
- Compromis entre la complexité du modèle, sa pertinence biologique et ce que l'on sait faire mathématiquement
- L'**interprétabilité** d'une méthode est très importante
- Quelle stratégie est meilleure ?

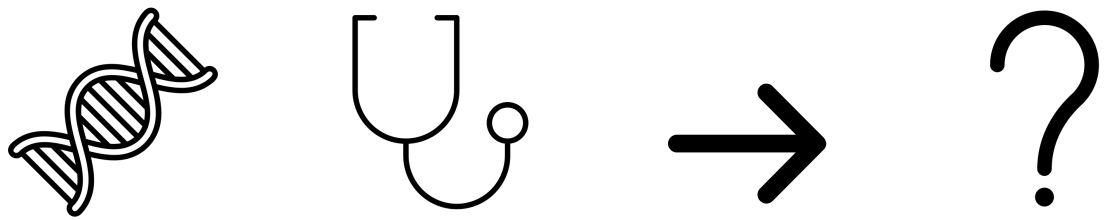
Comparaison des deux stratégies

- On peut faire des **hypothèses** qui permettent de simplifier le modèle mathématique
- Compromis entre la complexité du modèle, sa pertinence biologique et ce que l'on sait faire mathématiquement
- L'**interprétabilité** d'une méthode est très importante
- Quelle stratégie est meilleure ?

Ça dépend du domaine et de ce que l'on cherche à faire !

Autre question : prédire le résultat sur un nouveau patient

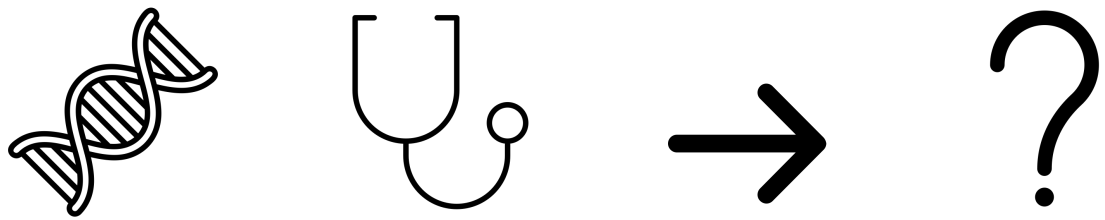
- On va mesurer X pour un nouveau patient, on cherche à prédire Y (malade/non malade) $\rightarrow$ problème de prédiction



- Ici, il est très important de comprendre exactement comment a été obtenu le résultat de notre méthode ! On préférera sans doute la première stratégie.

Autre question : prédire le résultat sur un nouveau patient

- On va mesurer X pour un nouveau patient, on cherche à prédire Y (malade/non malade) $\rightarrow$ problème de prédiction



- Ici, il est très important de comprendre exactement comment a été obtenu le résultat de notre méthode ! On préférera sans doute la première stratégie.
- On peut imaginer le même modèle pour prédire si une image est un chat ou un lapin : ici, on peut vouloir maximiser le taux de bonne réponse. On choisira plutôt la deuxième stratégie.

La recherche en IA, ça consiste en quoi exactement ?

- Poser les problèmes qui nécessitent d'utiliser l'IA (énergie, écologie, transport...)
- Proposer une modélisation mathématique adaptée (ni trop simple, ni trop complexe)
- Développer des méthodes numériques rapides et efficaces
- Analyser théoriquement les modèles et les méthodes proposées
- Interpréter les résultats en accord avec la personne qui les a produits et l'experte du domaine d'application

En bref

La recherche en IA fait intervenir des spécialistes des domaines d'application, des modélisatrices, des mathématiciennes théoriques et appliquées, des ingénieures, des informaticiennes, des gens à l'interface entre tous ces domaines !

Bilan

- La recherche en IA est très excitante, elle évolue constamment et elle s'applique dans tous les domaines
- Il y a encore un décalage entre les résultats impressionnants de l'IA et la compréhension mathématique de ses modèles
- Il faut favoriser la collaboration et la discussion entre les spécialistes des différents domaines
- Il ne faut pas négliger la recherche dans les domaines qui sont moins à la mode !!

Quelques perspectives

Article

"Tackling climate change with machine learning" (D. Rolnick et al, 2022)

- Une étude pluridisciplinaire publiée par des expert.e.s de plusieurs secteurs (système électrique, transport, industrie, agriculture...)
- S'adresse à un public large (chercheur.e.s, ingénieur.e.s, politiques, entrepreneur.e.s...)
- Chaque proposition est classée parmi 3 catégories
  - High-leverage : les outils d'IA sont particulièrement adaptés
  - Long-term : effets à prévoir d'ici 2040
  - Uncertain impact : impact difficile à mesurer, notamment avec la prise en compte de l'effet rebond
- Quelques idées à retenir :
  - L'IA peut apporter des solutions mais également faire partie du problème
  - L'IA seule ne peut pas résoudre la crise climatique
  - Nécessité du partage des connaissances entre les domaines

Références

- Formation Deep Learning Fidle CNRS `https://fidle.cnrs.fr`
- Illustrations : The Noun Project `https://thenounproject.com/`, Fidle CNRS et Futura Science, BD de xkcd
- IA et changement climatique : Tackling Climate Change with Machine Learning, D. Rolnick et al (2022), `https://dl.acm.org/doi/10.1145/3485128`
- Histoire de l'Intelligence Artificielle
  - Dominique Cardon, Jean-Philippe Cointet, Antoine Mazières (2018). "La revanche des neurones "
  - Polycopié "Introduction à l'intelligence Artificielle" de M. Bienvenu `https://www.labri.fr/perso/meghyn/papers/cours_IA.pdf`