

> Comprendre l'IA et ses impacts

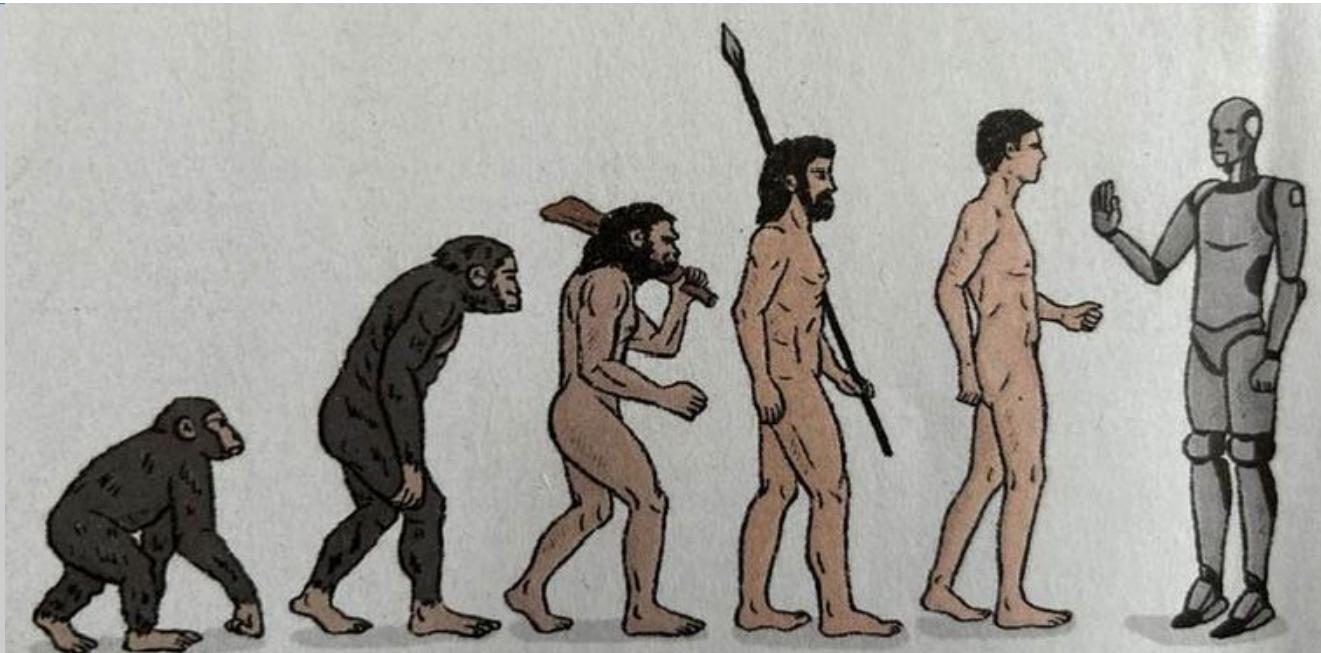
Université du Temps Libre

```
/* question 5 */
function TraiterEntrees() {
    if (isset($_POST) && count($_POST)==1)
    {
        /* on devrait valider la syntaxe de nom ici, pour le moment
         * on fait simple, cf. cours et TP11 */
        $nom = $_POST['nom'];
        CreationPersonne($nom);
    }
    if (isset($_GET) && count($_GET)==2)
    {
        if ($_GET['action']=="DEL")
        {
            /* récupérer et valider l'ID avant de le jouer (à faire) ...
             * $_GET['id']; ... */
        }
    }
}
```

5 – IA & Données

Les formes primitives d'intelligence artificielle que nous avons déjà se sont montrées très utiles. Mais je pense que le développement d'une intelligence artificielle complète pourrait mettre fin à l'humanité.

(Stephen Hawking) [46]



- À quoi servent les données en IA ?
- Où trouver / collecter ces données ?
- Qualité & Représentation
- Bonnes pratiques
- Exemples d'échecs

À quoi servent les données ?

Les données décrivent le monde !

- Nous percevons le monde *via* des observations, et aussi au moyen d'outils de mesure
- Pour comprendre (et prévoir) on a bâti des modèles qui *expliquent* ce qu'on observe



L'IA cherche à proposer des modèles de simulation du monde impliquant un raisonnement

L'apprentissage automatique vise à construire ces modèles d'après des données

Exploitation des données

- Un *exemple* est une **liste de mesures** qualifiant/quantifiant les observations = **numérisation**
- Ces mesures *font sens* dans un **espace de représentation**
exemples = **points** dans cet espace

L'apprentissage automatique peut aussi **trouver comment structurer l'espace de représentation** pour permettre aux modèles d'atteindre un certain objectif

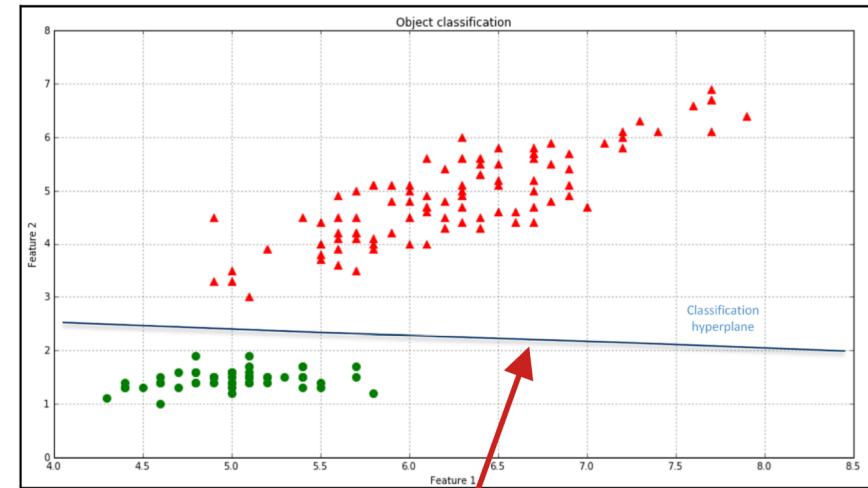
Des mesures, pour faire quoi ?

- L'ensemble des tâches réalisables par les IA actuelles se classent en 3 familles :
 - Classification supervisée
 - Classification non supervisée
 - Régression
- Toutes s'appuient sur un passage des données au modèle, très souvent au moyen de l'apprentissage automatique

Classification Supervisée

Affecter les exemples à un groupe connu (*labélisé*)

- Reconnaissance d'images (visage, posture, caractères, objets...), de la parole, de textes, d'émotions, de gestes (notion de dynamique)...



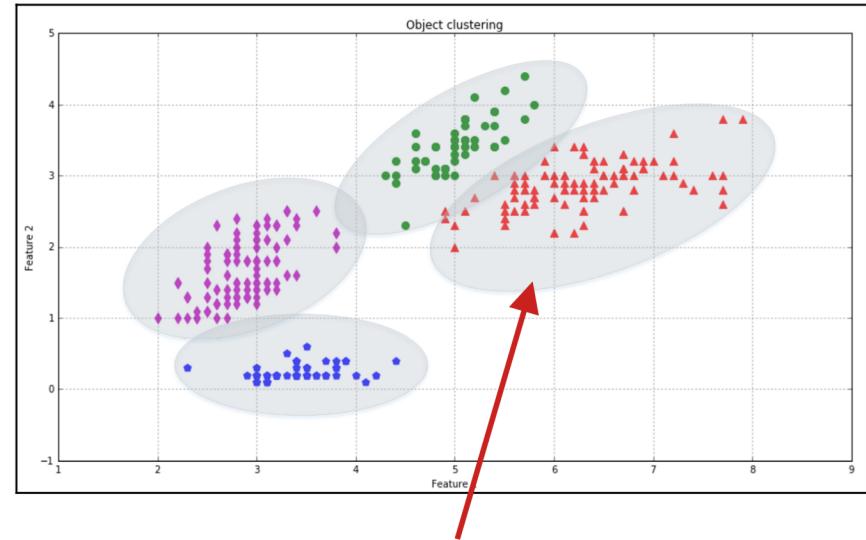
Source : [47]

Dans l'espace de représentation, il faut séparer les groupes de points

Classification non supervisée

Regrouper les exemples

- Structuration d'ensembles
- Découverte de motifs
- Réduction de variété
- Production d'arbres de décision...

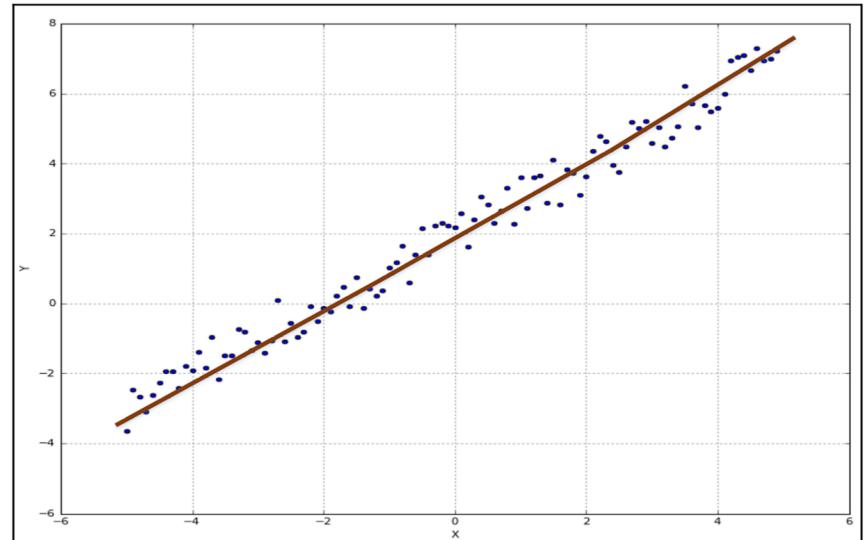


Dans l'espace de représentation, il faut **regrouper les points similaires**

Régression

Prédire l'évolution

- Evolution d'un processus (trajectoires, épidémies, séries financières, etc.)
- Prochaine forme d'une série (ex. mot d'un texte)



Source : [47]

Dans l'espace de représentation, il faut *relier* les points

Apprentissage automatique

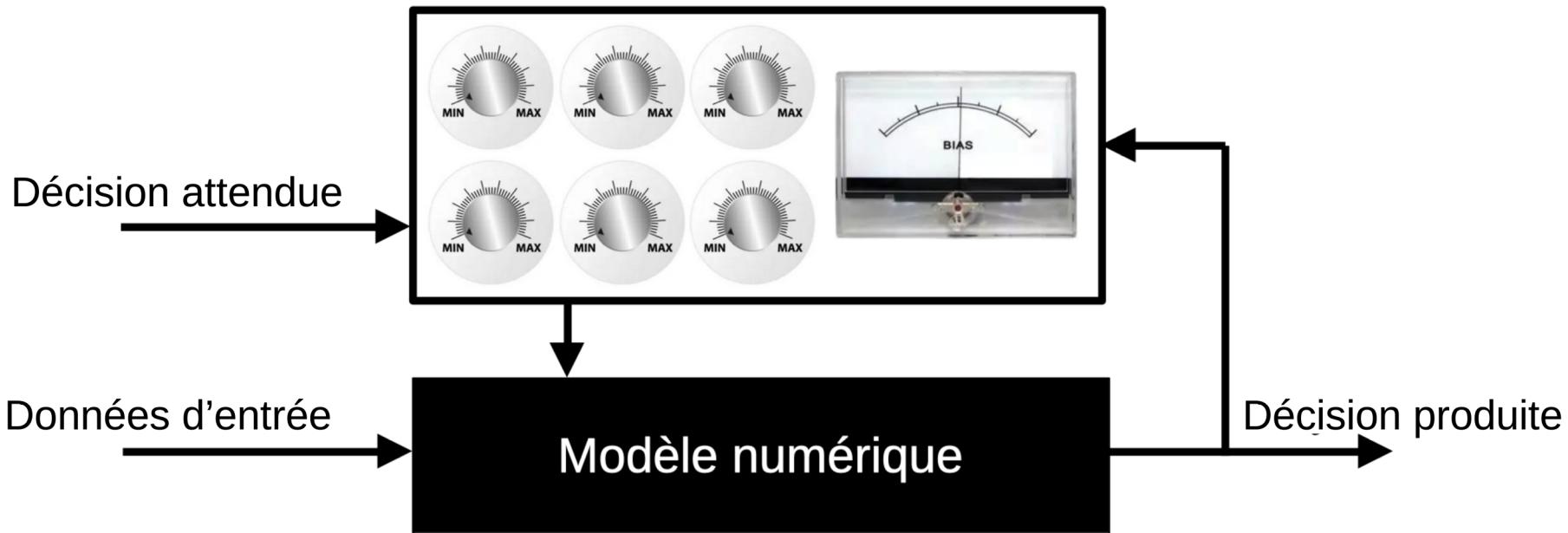
- Aussi appelé *machine learning (ML)*
- Principe énoncé par A. Turing [48] en 1948
- Appellation attribuée A.L. Samuel [49] en 1959
- Sous-domaine de l'intelligence artificielle
- Algorithmes de construction et/ou de paramétrage de modèle d'IA, d'après les données
- Sous-ensemble célèbre : *Deep Learning* (≈ 2010)



Image: Science Photo Library

Image : IEEE Society

Apprentissage (supervisé)



Où trouver / collecter les données ?

Les données sont partout !

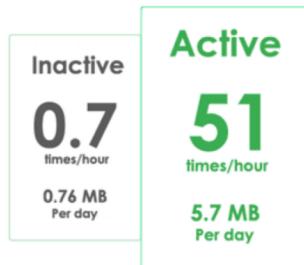
- Capteurs, smartphones, véhicules connectés, électroménager...
- Documents produits
- Interactions, connexions, transactions...
- La moindre information a une source, une date, un lieu, un contexte, une intensité, un média...

Tout est numérisable ...



[Matrix 4 - 2021]

Sans oublier....

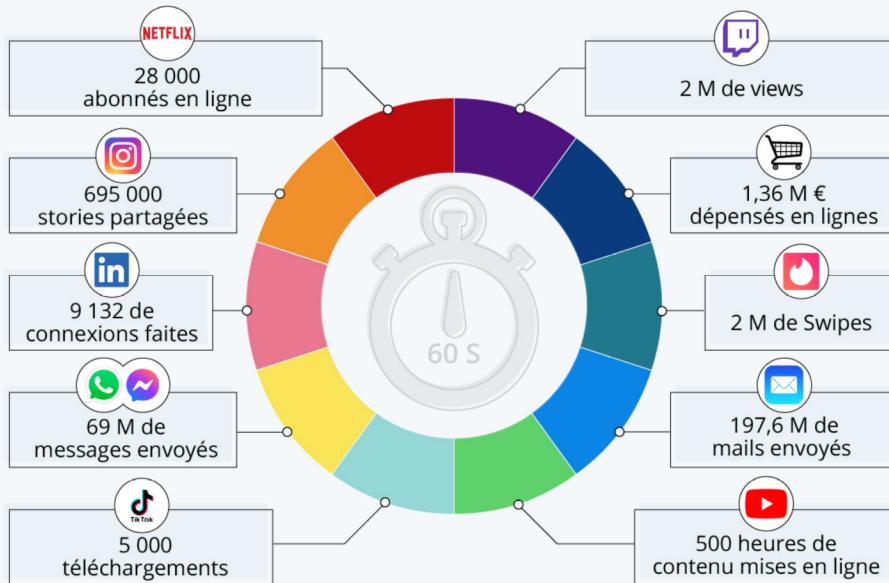


Sources : [50,51,52]



Une minute sur Internet en 2021

Estimation de l'activité et des données générées sur Internet en l'espace d'une minute



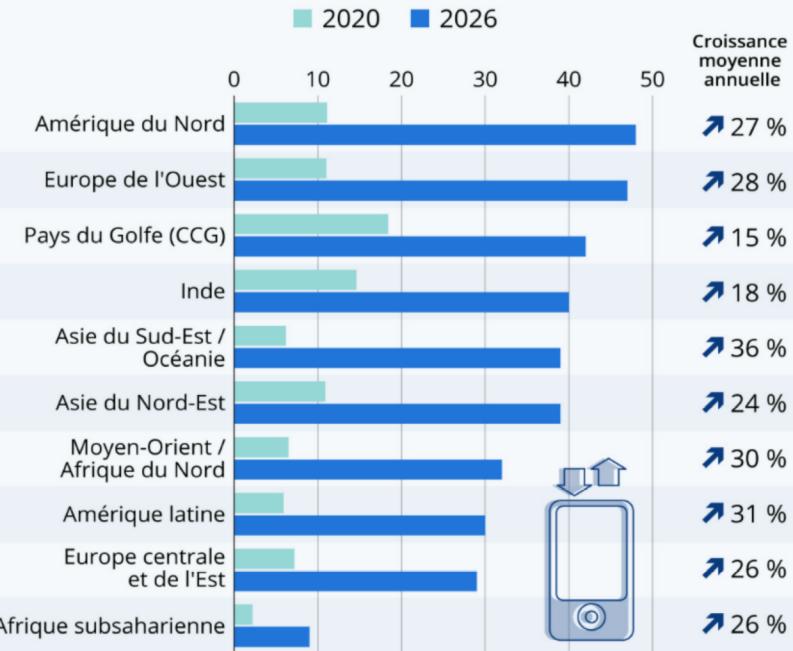
Source : Lori Lewis via AllAccess



statista

Données mobiles : comment le trafic va exploser

Prévision de l'évolution du trafic moyen de données mobiles par smartphone et par région, en Go par mois



Source : Ericsson Mobility Report (juin 2021)

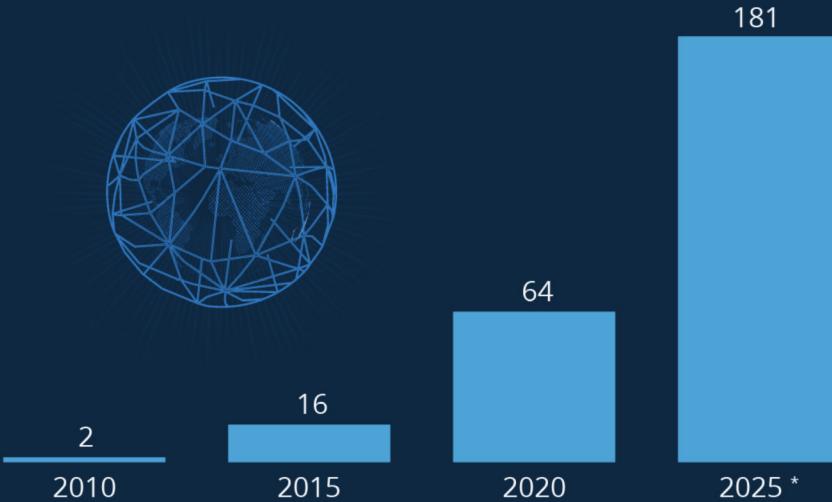
statista

Mégadonnées

- Aussi appelé *Big Data*
- Appellation issue de [53] en 1997
- Phénomène ancien (50'), en constante progression, exponentielle en volume de données
- Concepts liés : data science, cloud computing, DFS, HPC, data lake...
- Problèmes techniques et éthiques

Le Big Bang du Big Data

Estimation du volume de données numériques créées ou répliquées par an dans le monde, en zettaoctets



Un zettaoctet équivaut à mille milliards de gigaoctets.

* Prévision en date de mars 2021.

Sources : IDC, Seagate, Statista

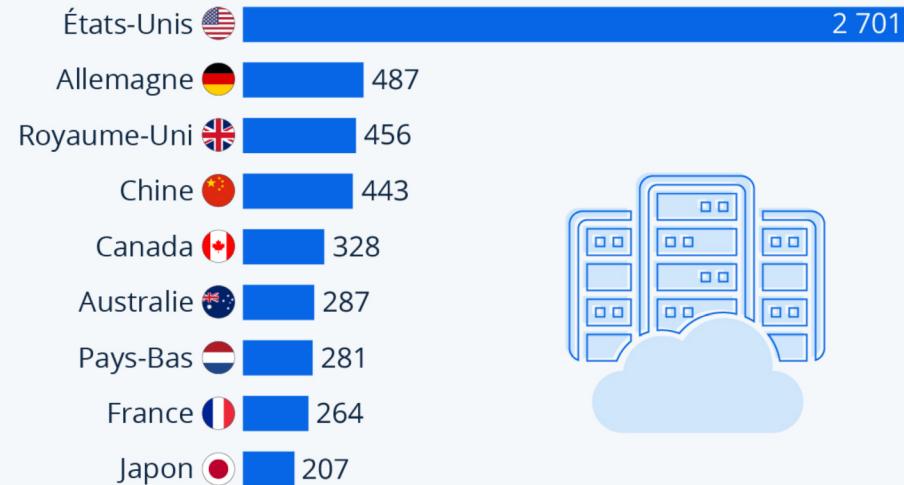


statista



Les pays qui hébergent le plus de data centers

Nombre de centres de données recensés par pays en septembre 2022 *



* Sélection des pays avec plus de 200 data centers répertoriés.

Source : Cloudscene



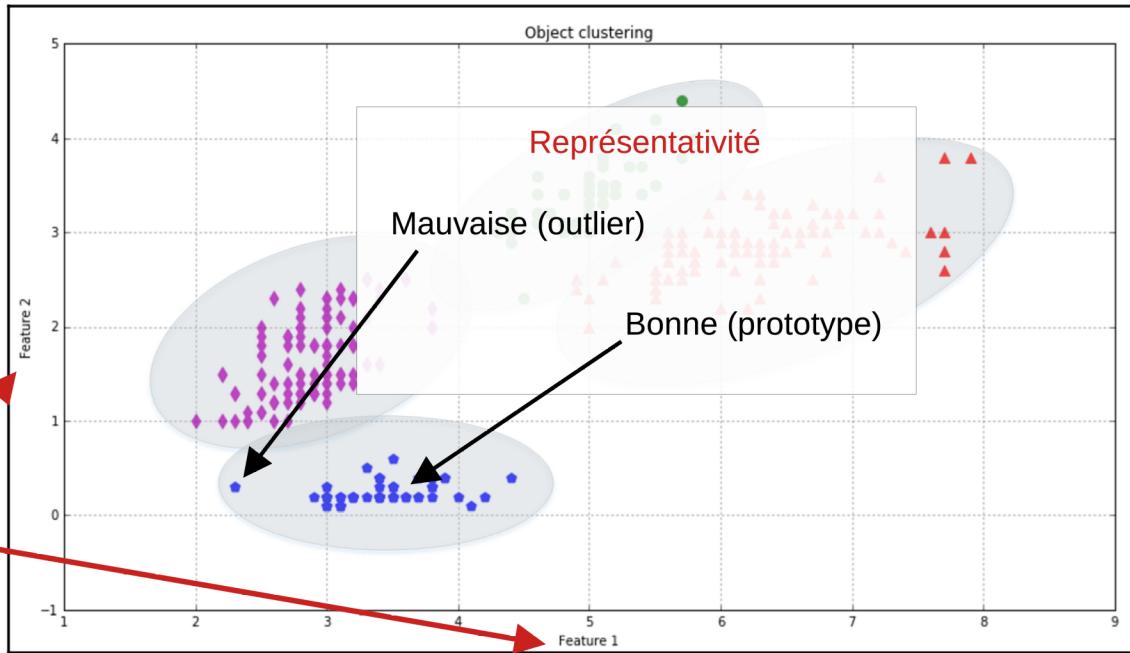
statista

Qualité & Représentation

De quoi parle-t-on ?

- **Représentation** : lié à *l'encodage*, *i.e.* en machine, à la manière de décrire l'information
 - Codage brut pour les nombres
 - Représentation scalaire ou vectorielle pour le reste
- **Représentativité** : est-ce qu'une donnée est « emblématique » (on la dit *prototype*) ou au contraire peu représentative (presque *outlier*)
 - Proche ou éloignée de la donnée caractéristique *moyenne*

De quoi parle-t-on ?

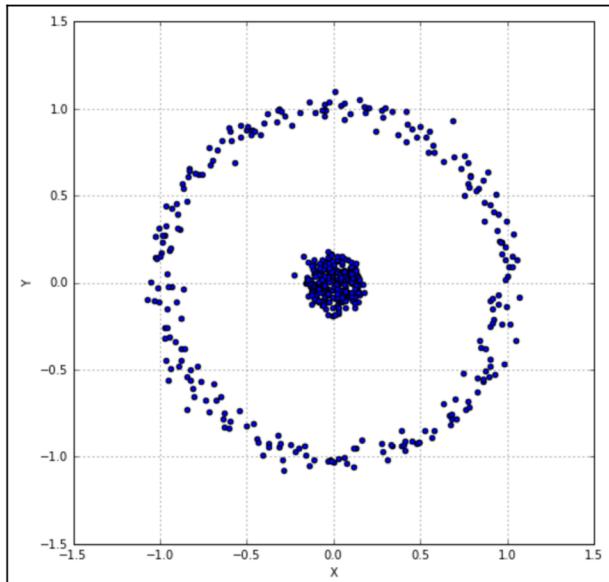


Source schéma : [47]

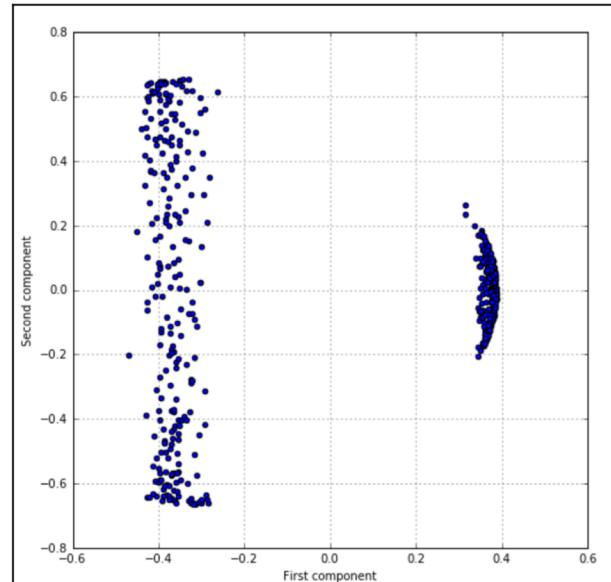
Choix de représentation

- Certains espaces de représentation sont plus propices pour atteindre l'objectif attendu
- L'expert peut définir lui même cet espace...
- L'espace peut être construit **automatiquement** par le ML afin d'optimiser l'objectif
 - fonctions noyau (ex. dans SVM)
 - *feature extractor* dans architecture Deep
 - encodeur/décodeur dans un transformer...

Choix de représentation



Classification non supervisée très difficile



Situation plus favorable en classification

Source schéma : [47]

Hypothèses de travail

- La dispersion des mesures est connue
- La distribution des données est connue
- La **stationnarité** des distributions : Les lois qui caractérisent les exemples ne changent pas entre 2 échantillons
 - Problème d'échantillonnage et de variabilité
 - Données non stable par nature, ex. empreintes digitales
 - Mesures forgées pour tromper le système...

Nettoyage des données

- Données acquises peuvent présenter ...
 - duplication : copies multiples d'une même info
 - des valeurs manquantes dans des champs importants
 - des valeurs erronées ou obsolètes
 - des problèmes de formatage (dont orthographe, conventions de codage...)

Suppression/correction des données erronées,
corrompues, mal formatées, dupliquées ou incomplètes

Des données aux datasets

- Nettoyage est une tâche fastidieuse, indispensable et pourtant mal considérée [54,55]
- Approche récente utilise l'IA pour nettoyer les données, avançant un gain de 50 % en temps [56]
- Nombreux datasets [57] dans tous les domaines
- Indispensables pour comparer les résultats au niveau scientifique

WE REALIZED
ALL OUR DATA
IS FLAWED.



GOOD

...SO WE'RE NOT
SURE ABOUT OUR
CONCLUSIONS.



BAD

...SO WE DID LOTS
OF MATH AND THEN
DECIDED OUR DATA
WAS ACTUALLY FINE.



VERY BAD

...SO WE TRAINED
AN AI TO GENERATE
BETTER DATA.



MCAI vs DCAI

- **model-centric AI** : focus sur le modèle
 - L'attention est portée sur le modèle, et les algorithmes de ML
 - Le modèle «rattrape les erreurs »
- **data-centric AI** : focus sur les données et la méthode pour créer les datasets [58,59,60]

Données = cœur du problème

- Il arrive (souvent) que le modèle ne soit pas performant, on remet alors en cause :
 - Les *features* (espace de représentation)
 - Les données (qualité, stabilité, représentativité...)
 - Le modèle d'IA

IA et biais

Bug ou Biais ?

- Bug : le système ne fait pas ce pour quoi il est programmé :
pb technique, d'ingénierie
- Biais : le système s'applique mal à la réalité
 - Hypothèses préjugées dans les algorithmes
 - Mauvaise représentativité des données

Problème de conception

Mais alors, le machine learning...

est un mécanisme qui reproduit avant tout les biais présents dans les données !

- Il est alors important
 - De comprendre les biais
 - De trouver d'où ils viennent pour les limiter
 - D'en tenir compte dans les évaluations

Catégories de biais en IA

- Cognitif :
 - Erreur de raisonnement, lors de modélisation / simplification d'une situation : **modèle**
 - Longue liste de biais de ce type (180...[61])
- **Données incomplètes :**
 - Pas assez représentatives, déséquilibrées
 - Pouvant aussi inclure un ou plusieurs biais

Real world patterns of health inequality and discrimination



Unequal access and resource allocation



Discriminatory healthcare processes



Biased clinical decision making



World → Data

Use ← Design



Sampling biases and lack of representative datasets

Discriminatory data



Patterns of bias and discrimination baked into data distributions

Application injustices



Disregarding and deepening digital divides



Exacerbating global health inequality and rich-poor treatment gaps



Hazardous and discriminatory repurposing of biased AI systems

Biased AI design and deployment practices



Power imbalances in agenda setting and problem formulation



Biased and exclusionary design, model building and testing practices



Biased deployment, explanation and system monitoring practices

Bonnes pratiques

Bonnes pratiques indispensables

- Méthodologie indispensable pour atteindre un certain niveau de qualité et de reproductibilité
- Besoin aussi bien scientifique que sociétal
 - Sci : progresser dans la connaissance
 - Soc : l'IA n'est pas neutre en terme d'impact
- Encore trop souvent manque de transparence
 - Mais faut il être totalement transparent ?
- Quelques initiatives : [63,64,65]

4 PRINCIPLES OF RESPONSIBLE AI & BEST PRACTICES TO ADOPT THEM



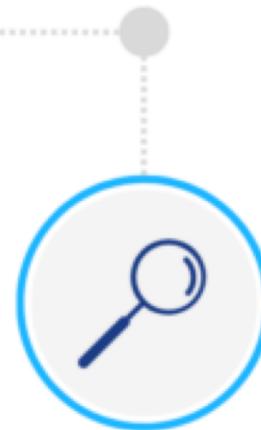
Fairness



Privacy



Security



Transparency

Fairness (équité)

- Système qui n'engendre pas de discrimination
- Analyser des données pour vérifier l'équité de représentation [66]
- Analyser les sous-populations pour voir si la performance du modèle est sélective [67]
- Concevoir des modèles en tenant compte de l'équité → avec des experts
- Monitorer le système dans le temps, car il dévie forcément à l'usage [68]

Privacy (confidentialité)

Datasets peuvent contenir des données sensibles, dont la violation ou l'usage détourné peut constituer un risque pour la population en général ou des sous-groupes

- Trier les données par degré de sensibilité
- Politiques du moindre privilège [69]
- Protéger les données ET le modèle (cryptographie, mais aussi modèles IA via des données synthétiques)

Security (sécurité)

L'attaque du modèle peut amener celui-ci à rendre une décision incorrecte dans le cadre d'un fonctionnement normal

- Evaluer l'intérêt d'une attaque sur l'IA ou ses données : qui et pourquoi ?
- Attaque/défense : missionner des équipes pour attaquer le système et évaluer ses faiblesses
- Se documenter auprès des sources scientifiques pour voir les avancées

Transparency (transparence)

Comprendre les décisions de l'IA et les expliquer pour les contrôler

- Responsabilité en cas de problème
- Simuler pour améliorer
- Réfléchir à, et fixer, un niveau d'explicabilité minimum
- Privilégier les IA explicables [44]
- Travailler avec de faibles volumes de données pour relier les décisions aux données d'entrée et faciliter la compréhension

Exemples d'échecs

Sécurité / Confidentialité

- [70] estimation de caract. génomiques de population par inversion de modèle
- [71,72] désanonymisation de datasets
- [73,74,75] dysfonctionnement de véhicules autonomes
- [76,77] modifications invisibles pour un humain qui mettent l'IA en échec

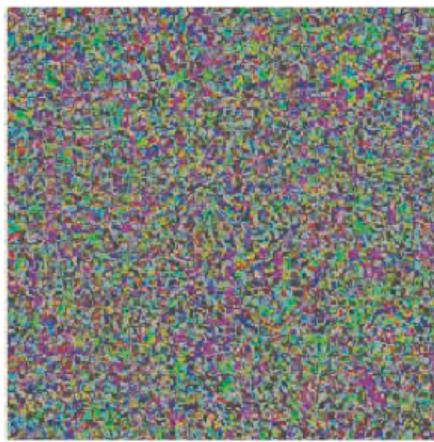
Original image



Dermatoscopic image of a benign melanocytic nevus, along with the diagnostic probability computed by a deep neural network.



Adversarial noise



+ 0.04 ×

Perturbation computed by a common adversarial attack technique.
See (7) for details.

Adversarial example



=

Combined image of nevus and attack perturbation and the diagnostic probabilities from the same deep neural network.



Source : [76]



Cup(16.48%)
Soup Bowl(16.74%)



Bassinet(16.59%)
Paper Towel(16.21%)



Teapot(24.99%)
Joystick(37.39%)



Hamster(35.79%)
Nipple(42.36%)

Source : [75]

Original



Adversarial



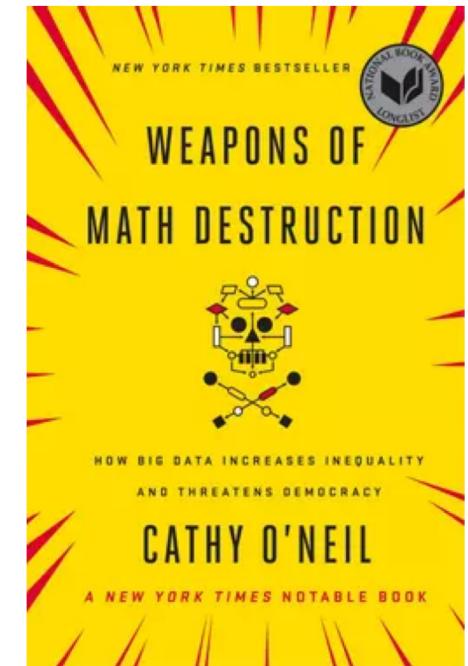
Classified as: Stop Speed limit (30)

Source : [77]

Boucle de rétroaction

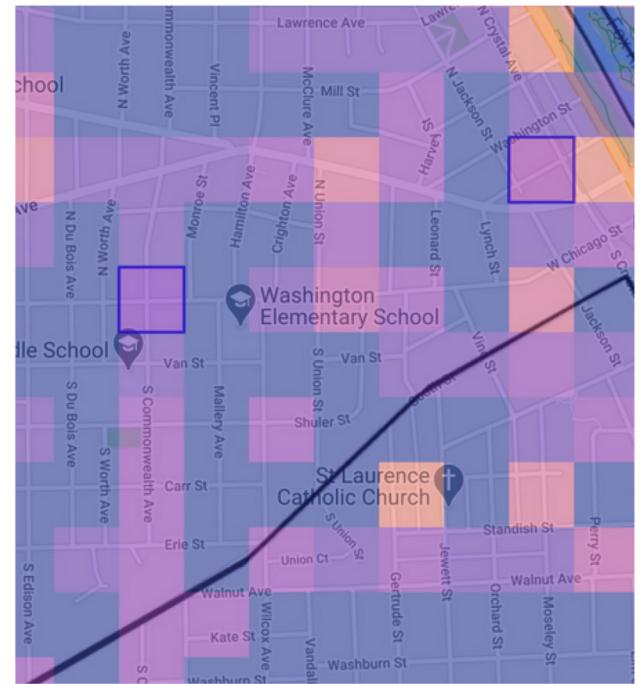
Application *aveugle* ‘un modèle biaisé entraîne une mauvaise décision

- Son application renforce la situation qui va dans le sens du biais
- [78,68] : enseignement, prévention de la violence, accord de prêts...
- Algorithme *obscure* de la CAF qui accroît la précarité [79]



Police prédictive !

- PredPol [80] : Echec dans 99 % des cas...
- Les technologies dites de **police prédictive** ont été déclarées comme présentant un **niveau de risque « inacceptable »** par le Parlement européen, lors de l'adoption de l'AI Act [81,82] en juin 2023, le règlement qui légifère l'utilisation de l'intelligence artificielle **dans l'Union Européenne**

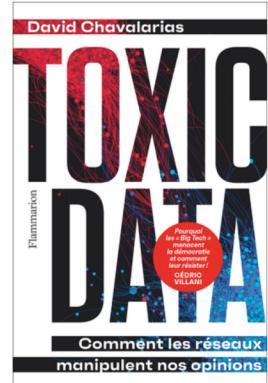


Proactively patrol to reduce crime rates and victimization

Bulle de filtres

Les algorithmes de recommandation peuvent mener à des « bulles » [83] décrivant une réalité différente

- Montée de la radicalisation et du populisme, irrégularités électorales [84, 85]
- Génocide [86]



THE NEW YORK TIMES BESTSELLER

THE FILTER BUBBLE

What the Internet is Hiding from You



'Astounding'
Andrew Marr

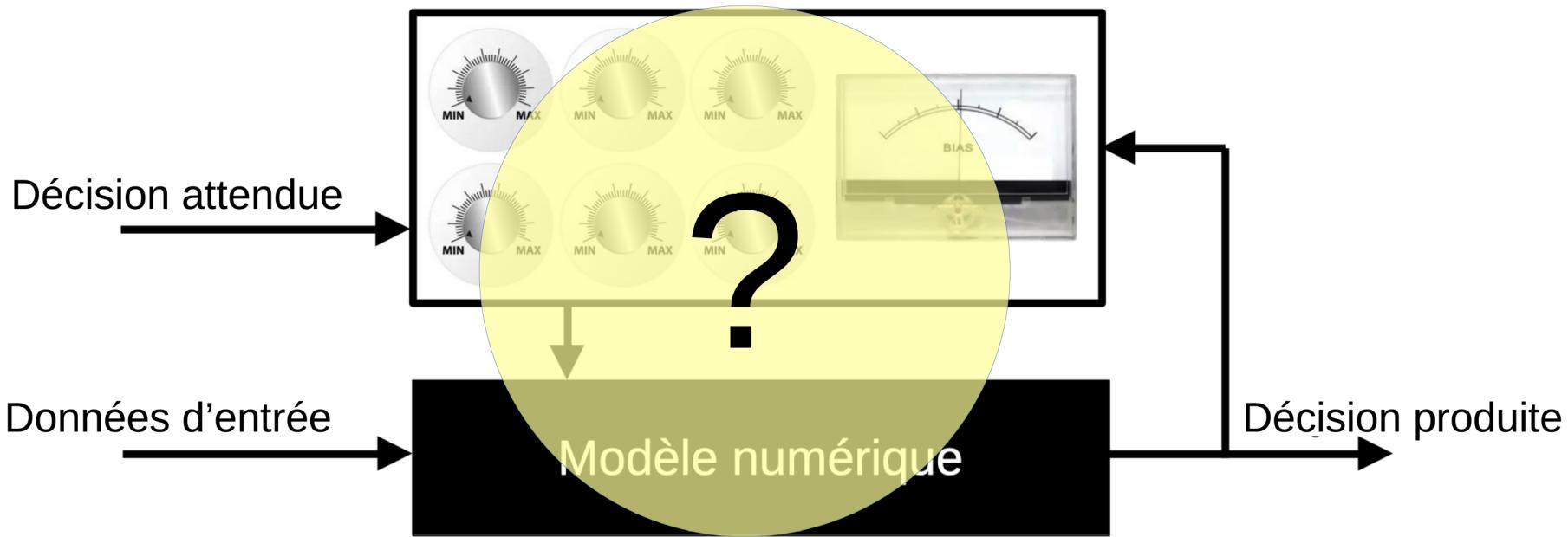
'Explosive'
Chris Anderson

ELI PARISER

Datasets biaisés

- Personnes de couleur non reconnues par les systèmes de reconnaissance faciale [87,88]
- 99% des Fortune'500 et entreprises de plus de 50 empl. recrutent via des IA qui écarteraient 27 millions de personnes des viviers de recrutement [89]
- Incitation au suicide [90,91]
- Faux cas de maltraitance [92]

Prochaine séance...



Données

UTL :: Comprendre l'IA et ses impacts



Av. Monge, 37200 Tours



brouard@univ-tours.fr



<https://www.univ-tours.fr/>



+33 247 367 019 (T. BROUARD)