

Big data, data science & intelligence artificielle

Vincent Lefieux

Big data
Data science
Intelligence artificielle
Quelques jeux de données pour pratiquer
Références



Plan

Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Plan

Big data

Big data

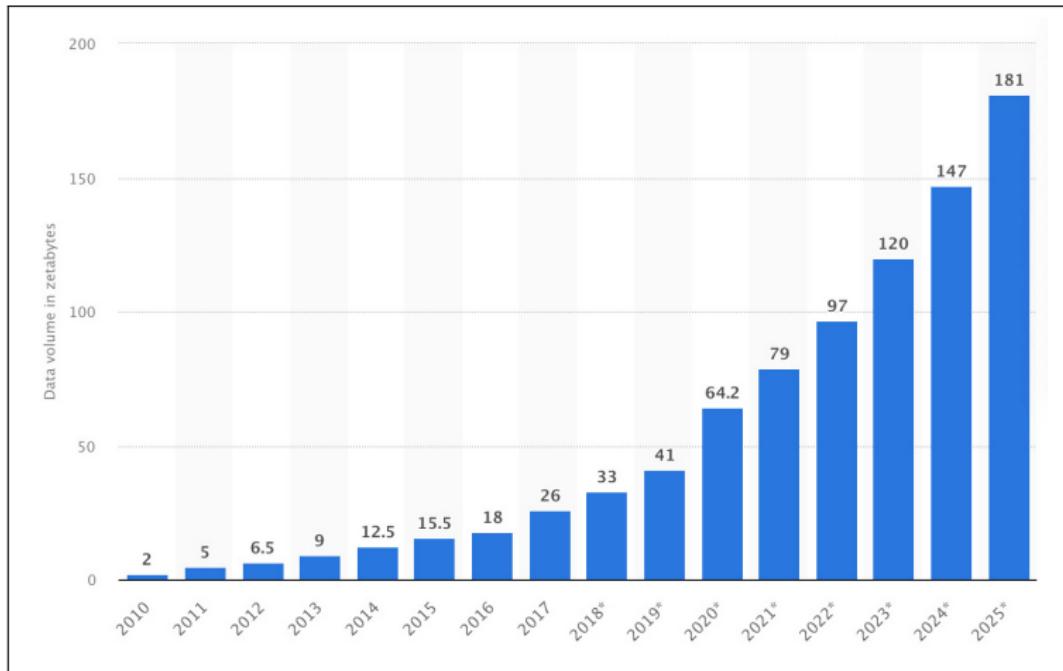
Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Des données toujours plus nombreuses II



Volume de données/informations créées, capturées, copiées et consommées dans le monde de 2010 à 2020, avec des prévisions de 2021 à 2025 (source : [Statista](#))

Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

Evolutions

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

► Avant :

- ▶ des données structurées,
- ▶ produites par des entreprises et des organisations,
- ▶ régulières mais peu fréquentes.

► Maintenant :

- ▶ des données non-structurées,
- ▶ produites par des individus,
- ▶ en temps réel.

Données structurées ou non

Structured Data vs Unstructured Data

Can be displayed
in rows, columns and
relational databases



Numbers, dates and strings



Estimated 20% of enterprise data (Gartner)



Requires less storage



Easier to manage
and protect with
legacy solutions



Cannot be displayed
in rows, columns and
relational databases



**Images, audio, video,
word processing files,
e-mails, spreadsheets**



Estimated 80% of enterprise data (Gartner)



Requires more storage



More difficult to manage and protect with legacy solutions



Big data

<https://lawtomated.com/>

[structured-data-vs-unstructured-data-what-are-they-and-why-care/](#)

Données produites par des entreprises et des organisations

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références



Mediametrie



Le Monde

Données produites par des individuels

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références



Big Data:
Expanding on 3 fronts
at an increasing rate.



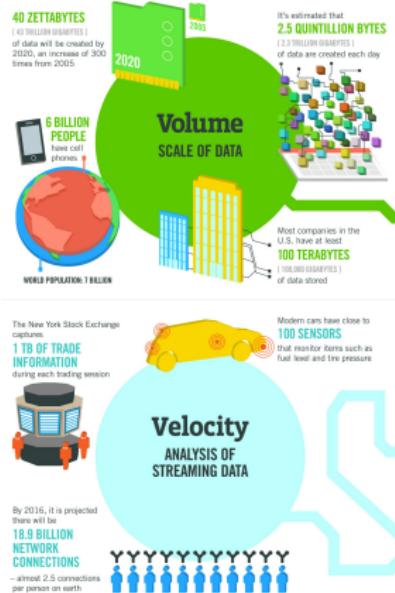
Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références



The FOUR V's of Big Data

From traffic patterns and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume**, **Velocity**, **Variety**, and **Veracity**.

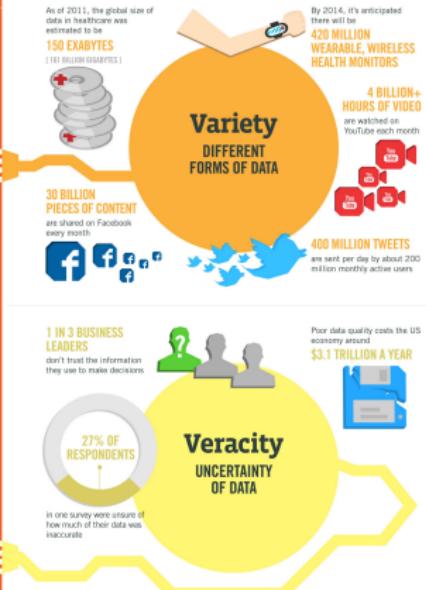
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, sensors, and mobile devices, as well as mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015, 4.4 MILLION IT JOBS will be created globally to support big data, with 1.9 million in the United States.



By 2016, it is projected there will be 18.9 BILLION NETWORK CONNECTIONS – almost 2.5 connections per person on earth.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MPTEC, Gartner



IBM

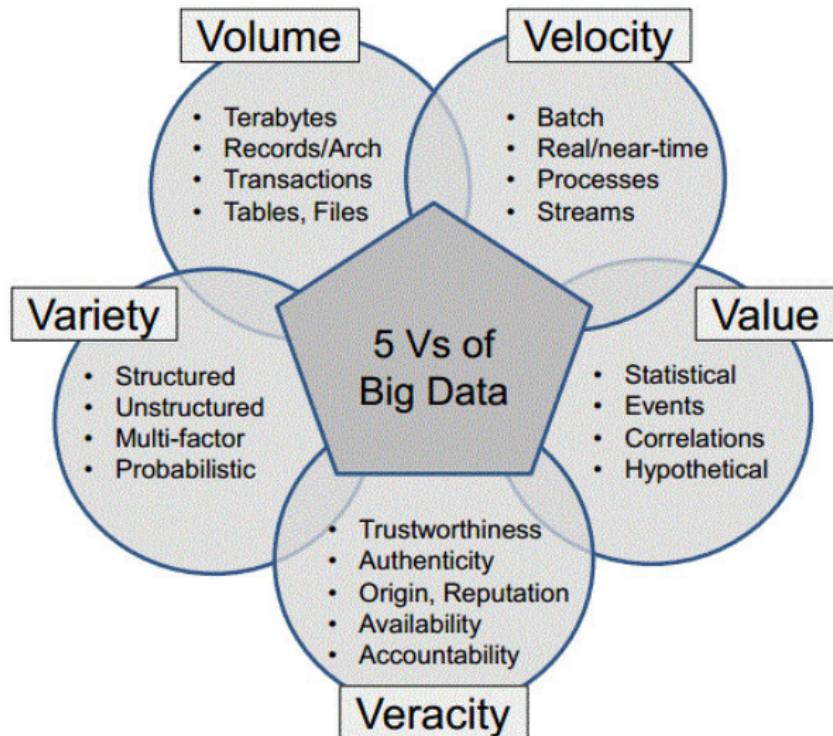
Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références



Big data

Data science

Intelligence
artificielleQuelques jeux de
données pour
pratiquer

Références

Plan

Data science

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Une nouvelle idée ?

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Four major influences act on data analysis today :

- ▶ *The formal theories of statistics.*
- ▶ *Accelerating developments in computers and display devices.*
- ▶ *The challenge, in many fields, of more and ever larger bodies of data.*
- ▶ *The emphasis on quantification in an ever wider variety of disciplines.*

Pas vraiment : (Tukey et Wilk, 1966)

Big data

Data science

Intelligence
artificielle

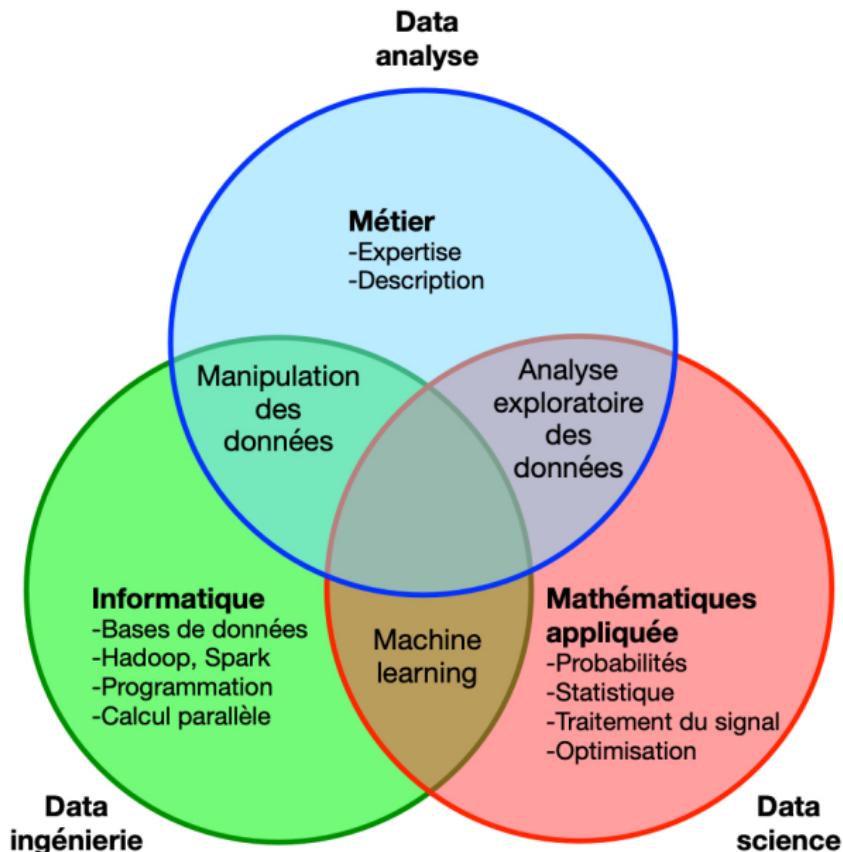
Quelques jeux de
données pour
pratiquer

Références

Four major influences act on data analysis today :

- ▶ *The formal theories of statistics.*
- ▶ *Accelerating developments in computers and display devices.*
- ▶ *The challenge, in many fields, of more and ever larger bodies of data.*
- ▶ *The emphasis on quantification in an ever wider variety of disciplines.*

Data - science, analyse et ingénierie



Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

Apprentissage supervisé ou non

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

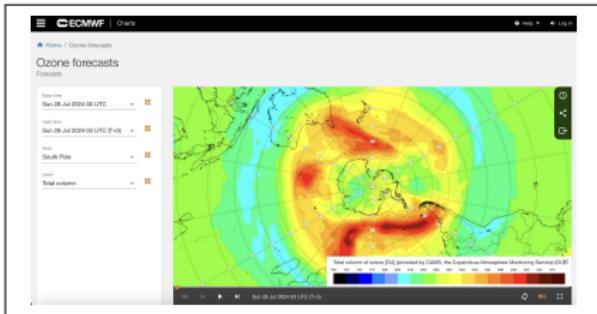
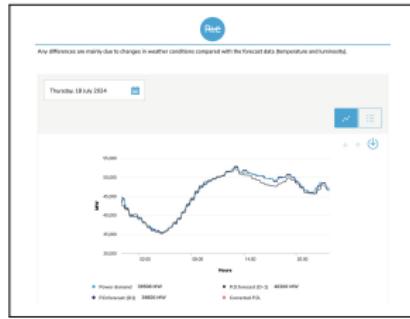
► Apprentissage supervisé :

Inférer (prévoir) une fonction/relation à partir de données d'apprentissages labellisées (ex : régression, classification supervisée).

► Apprentissage non-supervisé :

Trouver une “structure” dans des données non-labellisées (ex : analyse factorielle, classification non-supervisée ou *clustering*).

Exemples d'apprentissage supervisé : régression



Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

Exemples d'apprentissage supervisé : classification supervisée

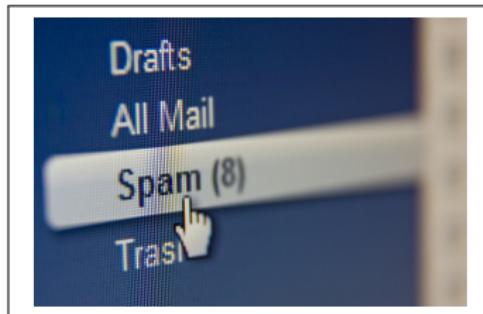
Big data

Data science

Intelligence artificielle

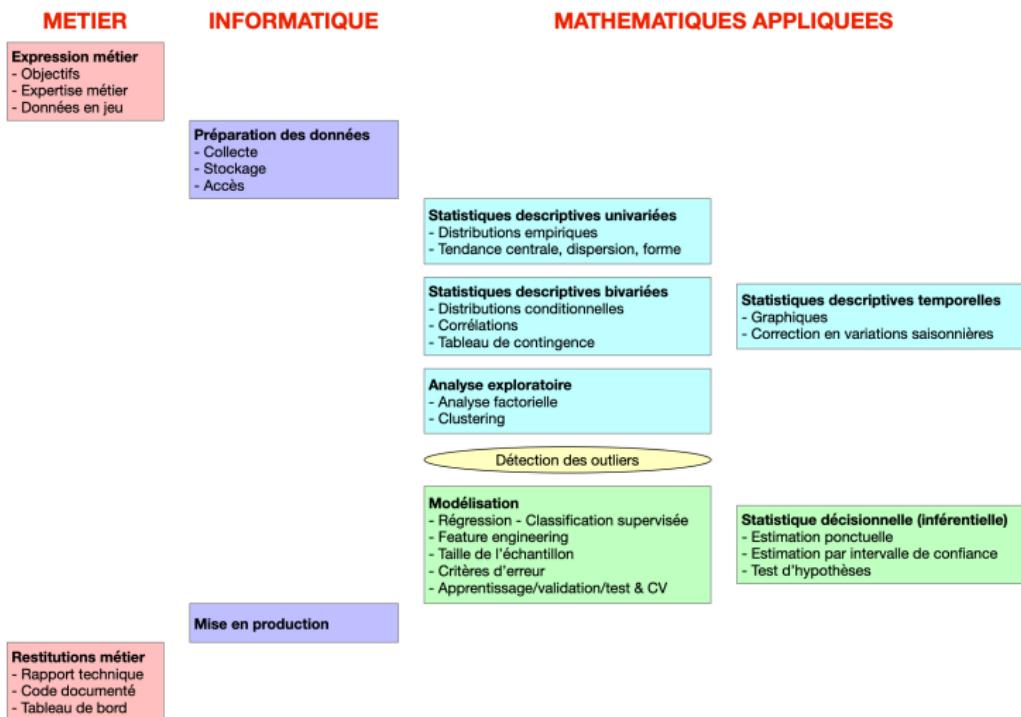
Quelques jeux de données pour pratiquer

Références



Label: 5	Label: 5	Label: 8	Label: 1	Label: 6	Label: 7
8	9	7	2	8	7
6	8	5	2	8	1
1	6	1	8	5	8
0	6	5	6	0	8

Un parcours « data » classique



Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

Python / R



- ▶ dispose d'une vaste communauté de développeurs et de data scientists,
- ▶ est utilisé au-delà de la data science (développement web, automatisation, etc.),
- ▶ dispose de nombreux packages très populaires en machine learning et en particulier en deep learning.



- ▶ dispose d'une vaste communauté de data analysts et de chercheurs académiques, en statistique, épidémiologie et sciences sociales, moindre que Python en data scientists,
- ▶ est circonscrit à data analyse et la data science,
- ▶ dispose de nombreux packages très populaires en visualisation et en analyse statistique mais est limité pour le deep learning.

Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

Plan

Intelligence artificielle

Big data

Data science

**Intelligence
artificielle**

Quelques jeux de
données pour
pratiquer

Références

Un peu de vocabulaire

- ▶ Selon l'OCDE (2023) : « An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment. ».
- ▶ 2 principales voies :
 - ▶ Les systèmes experts.
 - ▶ Le machine learning (incluant le deep learning).
- ▶ Des sous-domaines :
 - ▶ Computer vision.
 - ▶ NLP (Natural Language Processing) (TALN : Traitement Automatique du Langage Naturel) dédié au traitement du langage.

Big data

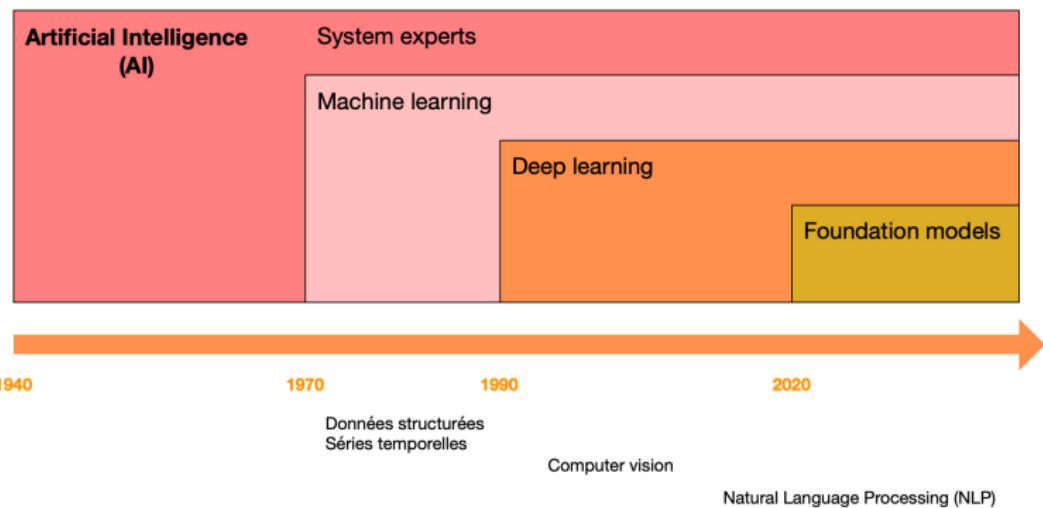
Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

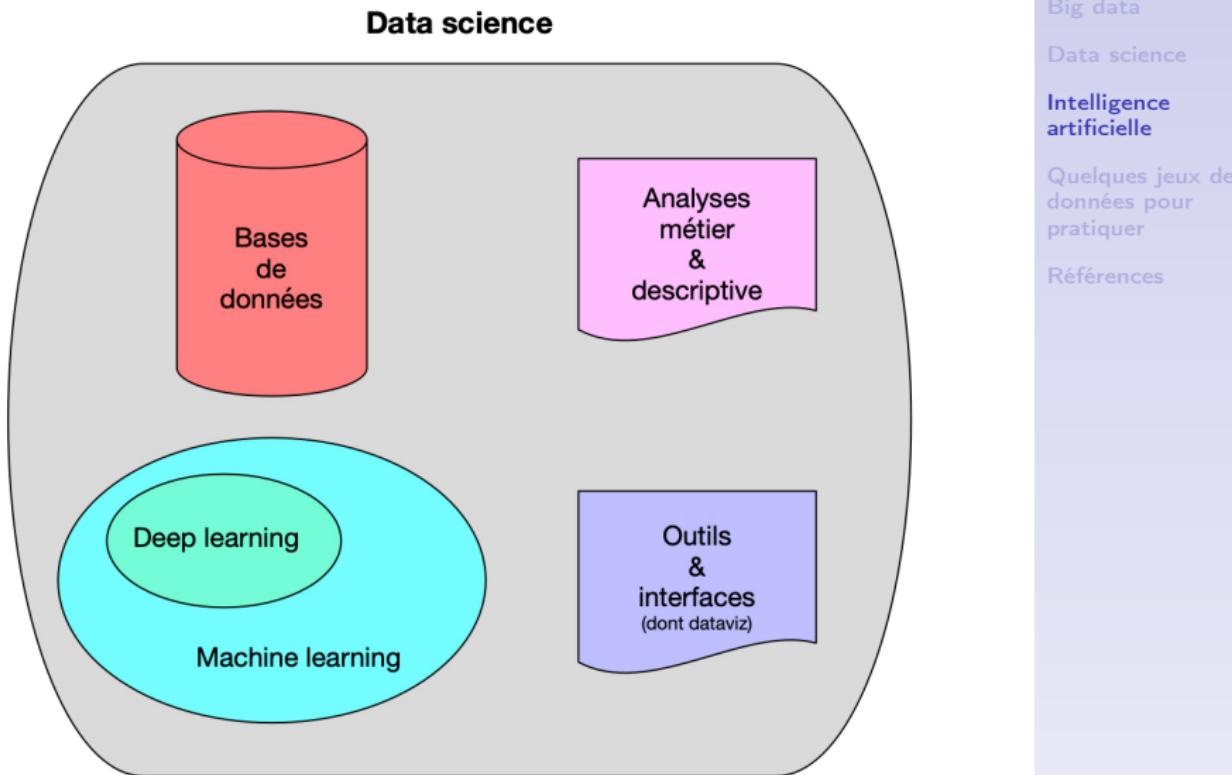
Références

Cartographie



Big data
Data science
Intelligence artificielle
Quelques jeux de données pour pratiquer
Références

IA et data science



Cognitivisme et connexionnisme

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

- ▶ L'IA a été influencée par des courants de pensée, issus de la cybernétique, parmi lesquels :
 - ▶ Le **cognitivisme** : il fait l'analogie entre la pensée et un processus de traitement de l'information. Il s'agit d'une approche par réduction (« top down »). Ce paradigme domine les sciences cognitives du milieu des années 1950 aux années 1980.
 - ▶ Le **connexionnisme** : il modélise la pensée ou le comportement comme un processus issu de réseaux d'unités simples interconnectées. Il s'agit d'une approche systémique (« bottom up »). Ce paradigme supplante le cognitivisme dans les années 1980.

Le terme « intelligence artificielle »

- Trenchard More, John McCarthy, Marvin Minsky, Oliver Selfridge et Ray Solomonoff, en 2006 présents lors de la **conférence de Dartmouth** sur l'IA en **1956** durant laquelle le terme *artificial intelligence* apparut.



- Le terme *intelligence* a un sens plus large en anglais (ex : CIA - Central Intelligence Agency).
- On lui préfère parfois le terme d'**intelligence augmentée** (*augmented intelligence*).

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Eléments chronologiques



Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

L'avènement des réseaux de neurones

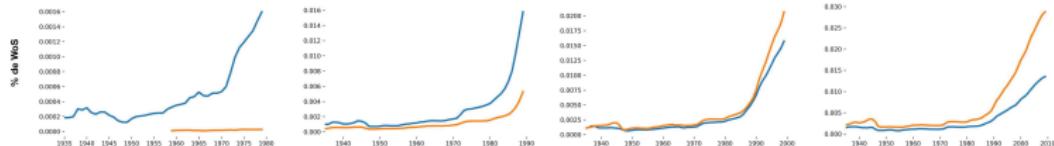
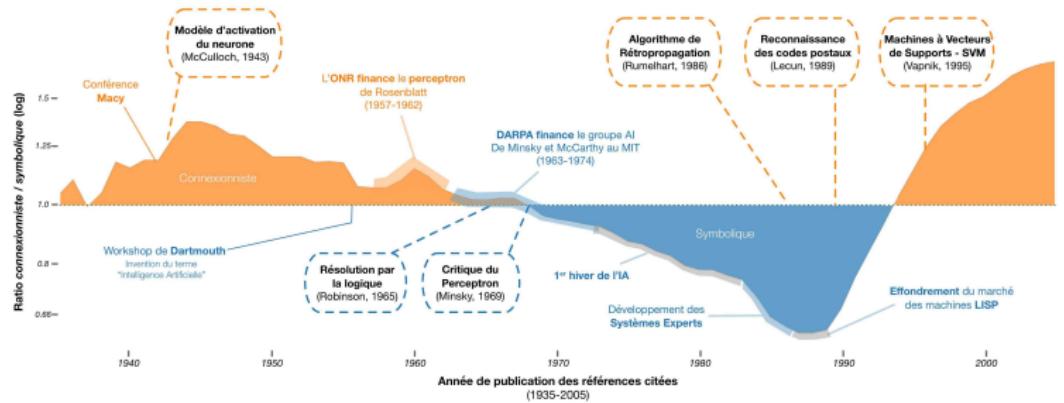
Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références



(Cardon et collab., 2018)

Les soubresauts de l'Histoire

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

- ▶ Une histoire avec des « **hivers** »...
- ▶ ... et des printemps grâce à :
 - ▶ de gros volumes de **données** (*big data*) labellisées,
 - ▶ des **moyens de calculs** puissants (ex : GPU),
 - ▶ des **algorithmes** plus performants.

Le grand jeu de l'IA



Big data

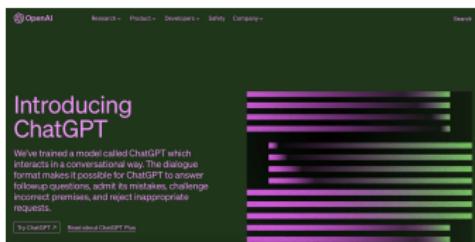
Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Des applications plus opérationnelles



Big data

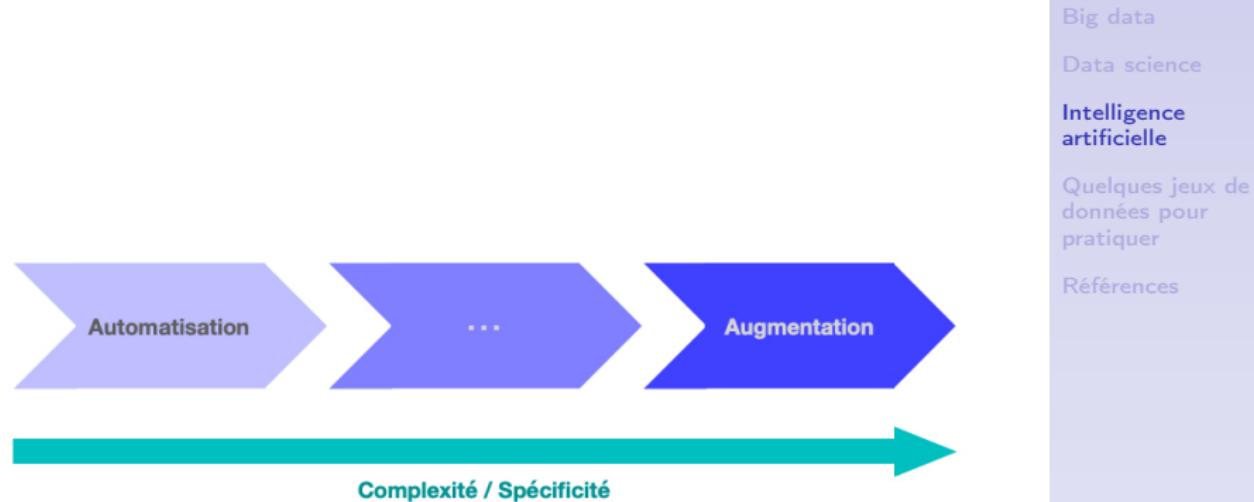
Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

Automatisation vs. Augmentation



Mythe et réalité



Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Des choix de société



Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références



Éthique de l'intelligence artificielle

Des points de vue différents parmi les récipiendaires du prix Turing 2019

ENTRETIEN

Yoshua Bengio, chercheur : « Aujourd'hui, l'intelligence artificielle, c'est le Far West ! Nous devons ralentir et réguler »

Le chercheur canadien, précurseur des réseaux de neurones artificiels, a signé l'appel à un moratoire sur le développement des applications d'intelligence artificielle comme ChatGPT. Il en explique les raisons dans un entretien au « Monde ».

Publié le 28 avril 2023 à 07h55, modifié le 29 avril 2023 à 17h04 · Claire Legros

Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

IA - INTELLIGENCE ARTIFICIELLE

Intelligence artificielle : l'un de ses pionniers quitte Google pour alerter sur ses dangers

Par Julien Dantchenko
Publié le mardi 2 mai 2023 à 17h44 · 0 min · 4 partages

Geoffrey Hinton a annoncé son départ de Google, pour qui il travaillait sur l'intelligence artificielle depuis une dizaine d'années. Le chercheur, primé en 2018 pour ses recherches, alerte sur les dangers de l'intelligence artificielle générative, comme celle utilisée par ChatGPT.



À l'origine, Geoffrey Hinton est considéré comme le pionnier de l'IA. AFP - Mladen Antonov / Getty / The Kinsey Wilson

ENTRETIEN

Yann Le Cun, directeur à Meta : « L'idée même de vouloir ralentir la recherche sur l'IA s'apparente à un nouvel obscurantisme »

Le chercheur français, qui est l'un des pères de l'intelligence artificielle et dirige un laboratoire de Facebook sur cette discipline, appelle, dans un entretien au « Monde », à accélérer les recherches pour améliorer la fiabilité des systèmes et conduire à un « nouveau siècle des Lumières ».

Publié le 28 avril 2023 à 08h00, modifié le 29 avril 2023 à 06h07 · Claire Legros

Des questions écologiques

- ▶ Luc Julia, Journal du Geek, 24 janvier 2019

« Il suffit d'ailleurs de regarder les chiffres : DeepMind, c'est 1500 CPU, environ 300 GPU, quelques TPU et 440 kWh. L'humain en face, c'est 20 Wh. Et lui sait faire bien d'autres choses que de jouer au go ! »



- ▶ Plus globalement, un bilan carbone important.
(Dhar, 2020)

<https://www.nature.com/articles/s42256-020-0219-9>

Big data

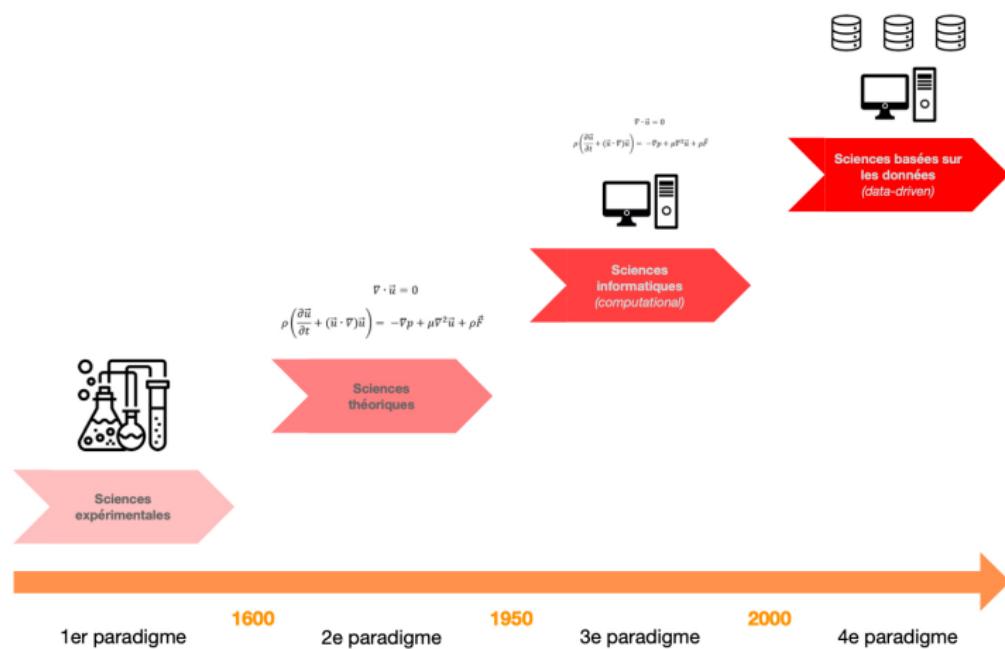
Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Changement de paradigme ?



Big data
Data science
Intelligence artificielle
Quelques jeux de données pour pratiquer
Références

Les modèles « boîte noire »

- ▶ nécessitent beaucoup de données,
- ▶ peuvent générer des résultats physiquement invraisemblables,
- ▶ sont incapables d'anticiper des situations inédites,
- ▶ génèrent uniquement des prévisions et ne permettent pas d'accroître les connaissances.



(Rudin et Radin, 2019)

<https://hdsr.mitpress.mit.edu/pub/f9kuryi8/release/8>

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Opportunités, conditions de réussite...

- ▶ Des opportunités :
 - ▶ La plasticité de l'IA en fait un levier de performance incontournable pour l'aide à la décision.
Ignorer les solutions IA peut limiter la performance opérationnelle (avec un risque image en sus)
 - ▶ L'IA est particulièrement adaptée aux données non structurées.
- ▶ Des conditions de réussite :
 - ▶ Compétences métier / SI / IA.
 - ▶ Données de qualité en quantité.
 - ▶ Moyens de calculs (GPU).
 - ▶ Focalisation sur les sujets pour lesquels l'IA est vraiment opportune.
 - ▶ Partenariats.

Big data

Data science

Intelligence artificielle

Quelques jeux de données pour pratiquer

Références

... et enjeux

► Des enjeux :

- ▶ **Sobriété** des algorithmes (enjeux économiques et développement durable).
- ▶ **Biais** algorithmiques.
- ▶ **Cybersécurité**.
- ▶ Questions **règlementaires et éthiques** (ex : RGPD, IA Act).
- ▶ **Propriété intellectuelle**.
- ▶ **RH** : conduite du changement (postes, compétences, responsabilité).

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Plan

Quelques jeux de données pour pratiquer

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Ressources

- ▶ Larry Winner (University of Florida, Department of Statistics) :
<http://www.stat.ufl.edu/~winner/datasets.html>
- ▶ James R. Eagan (Telecom ParisTech) :
<https://perso.telecom-paristech.fr/eagan/class/igr204/datasets>
- ▶ StatSci :
<http://www.statsci.org/datasets.html>
- ▶ University of California Irvine (UCI), Center for Machine Learning and Intelligent Systems :
<http://archive.ics.uci.edu/ml/>
- ▶ StatSci :
<http://www.statsci.org/datasets.html>
- ▶ Kaggle (data science competitions) :
<https://www.kaggle.com>

Big data

Data science

Intelligence
artificielle

Quelques jeux de
données pour
pratiquer

Références

Big data
Data science
Intelligence artificielle
Quelques jeux de données pour pratiquer
Références

Références

- Cardon, D., J.-P. Cointet et A. Mazieres. 2018, «La revanche des neurones : l'invention des machines inductives et la controverse de l'intelligence artificielle», *Réseaux*, vol. 5, n° 211, p. 173–220.
- Dhar, P. 2020, «The carbon impact of artificial intelligence», *Nature Machine Intelligence*, vol. 2, p. 423–425.
- Jani, K. 2016, «The promise and prejudice of big data in intelligence community», .
- Rudin, C. et J. Radin. 2019, «Why are we using black box models in ai when we don't need to ? a lesson from an explainable ai competition», *Harvard Data Science Review*, vol. 1, n° 2.
- Tukey, J. W. et M. B. Wilk. 1966, «Data analysis and statistics, an expository overview», dans *International workshop on managing requirements knowledge*, IEEE Computer Society, p. 695–709.