

Données ? IA ? J'en utilise déjà sans le savoir

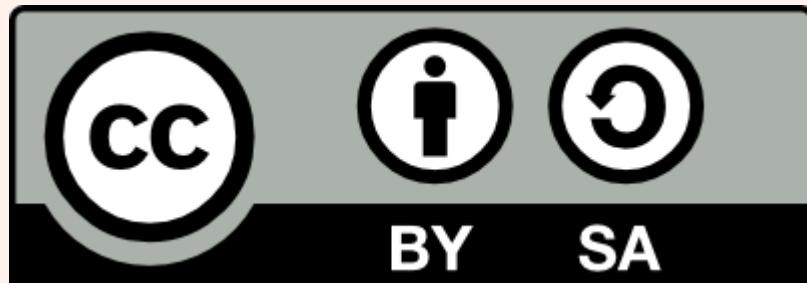
**Maëlle Fouquenet, Joël Combin et
Clément Mandron, Dataactivist**

Webinaire 1 - INTEFP - 2022

Ces slides en ligne : http://dataactivist.coop/webinaires_intefp/

Sources : https://github.com/dataactivist/webinaires_intefp/

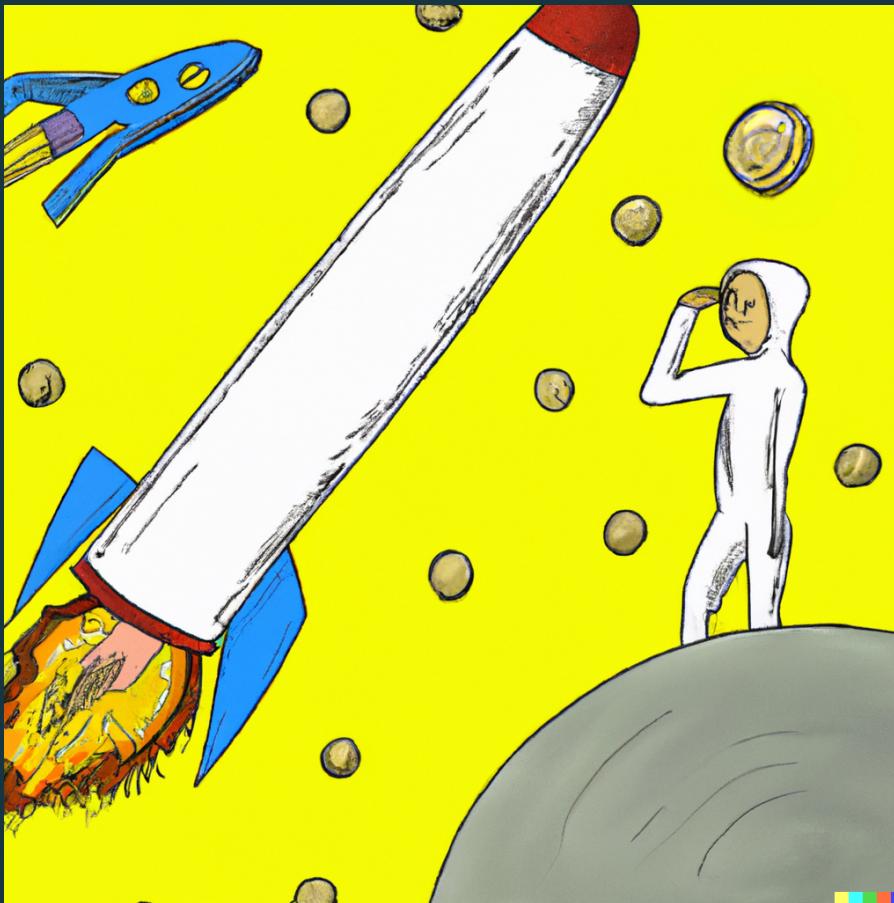
Les productions de Dataactivist sont librement réutilisables selon les termes de la licence [Creative Commons 4.0 BY-SA](#).



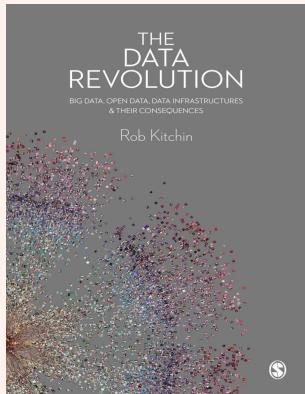
Au programme

1. Dessine moi une donnée
2. La mise en données du monde
3. Le cadre juridique des données
4. Ceux qui mangent les données : algorithmes, sciences de la données et intelligence artificielle

Dessine moi une donnée



Une définition des données



Les données sont généralement comprises comme étant la matière première produite dans l'abstraction du monde en catégories, mesures et autres formes de représentation - nombres, caractères, symboles, images, sons, ondes électromagnétiques, bits - qui constituent les fondations sur lesquelles l'information et le savoir sont créés.

Les données, la base de l'informatique

L'informatique consiste dans le traitement de l'information, ou de la donnée. La naissance de l'informatique est le point de départ d'un déluge de données. Le volume des données créées et traitées ne cesse de croître en même temps que les capacités de calcul des ordinateurs.

Data inflation

2

Unit	Size	What it means
Bit (b)	1 or 0	Short for “binary digit”, after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10} , bytes	From “thousand” in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From “large” in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From “giant” in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From “monster” in Greek. All the catalogued books in America’s Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America’s postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of <i>The Economist</i>
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

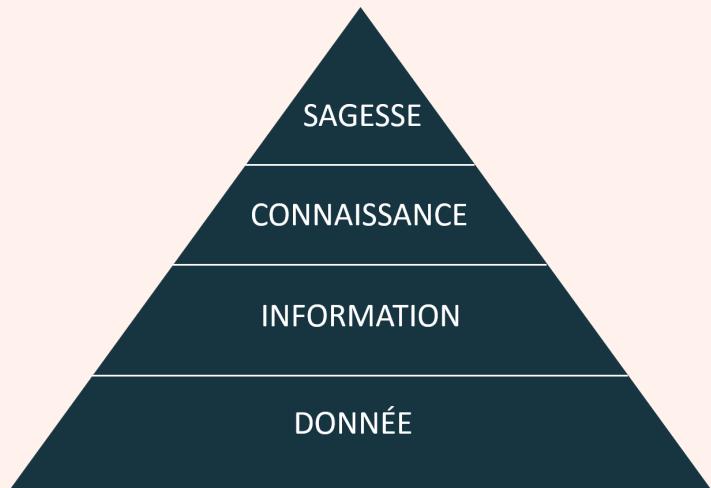
The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures.

Source: *The Economist*

Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

La pyramide DIK

Attribuée à [Russell Ackoff](#) en 1989, elle signifie que :



- la **donnée** est la matière "brute" de l'information conçue plutôt pour des machines
- **l'information** sont des données qui ont été interprétées pour dégager du sens pour des humains
- en donnant du sens à de l'information, on obtient de la **connaissance**
- en donnant du sens à la connaissance on obtient de la **sagesse** (ou compétence).

Un peu de sémantique

Latin : dare (donner) > datum (donné) > data (donnés)

Ce qui est évident, va de soi, est accepté sans discussion.

Data ou capta ?

Techniquement, ce que nous nous appelons "donnée" est en réalité "**capturé**" (du latin "capere", qui signifie "prendre") ; les *capta* sont les unités de données qui ont été sélectionnées et collectées parmi l'ensemble de toutes les données possibles.

Kitchin, *The Data Revolution*, 2014

La mise en données du monde

Le censeur à Rome, ancêtre de la statistique

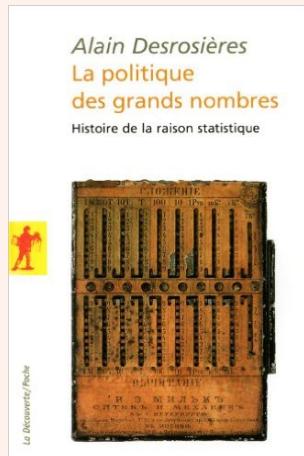


Source : *Asterix chez les pictes*, © Albert René 2013

La statistique : outil de gouvernement et de preuve

La statistique est à la fois :

- **outil de gouvernement** (*Statistik* - 18e siècle)
- **outil de preuve** (*statistics* - 19e siècle)



Comment comprendre qu'un même mot, statistique, évoque pour les uns la simple quantification (...), et pour d'autres l'idée de grands nombres et de régularités tendancielles appuyées sur le calcul des probabilités (...) ?

"Datafication" : la mise en données du monde

"L'immense gisement de données numériques découle de la capacité à paramétrier des aspects du monde et de la vie humaine qui n'avaient encore jamais été quantifiés. On peut qualifier ce processus de « **mise en données** » (**datafication**).

(...) La mise en données désigne autre chose que la numérisation, laquelle consiste à traduire un contenu analogique - texte, film, photographie - en une séquence de 1 et de 0 lisible par un ordinateur. Elle se réfère à une action bien plus vaste, et aux implications encore insoupçonnées : **numériser non plus des documents, mais tous les aspects de la vie.**"

Kenneth Cukier (2013), "Mise en données du monde, le déluge numérique", *Le Monde diplomatique*

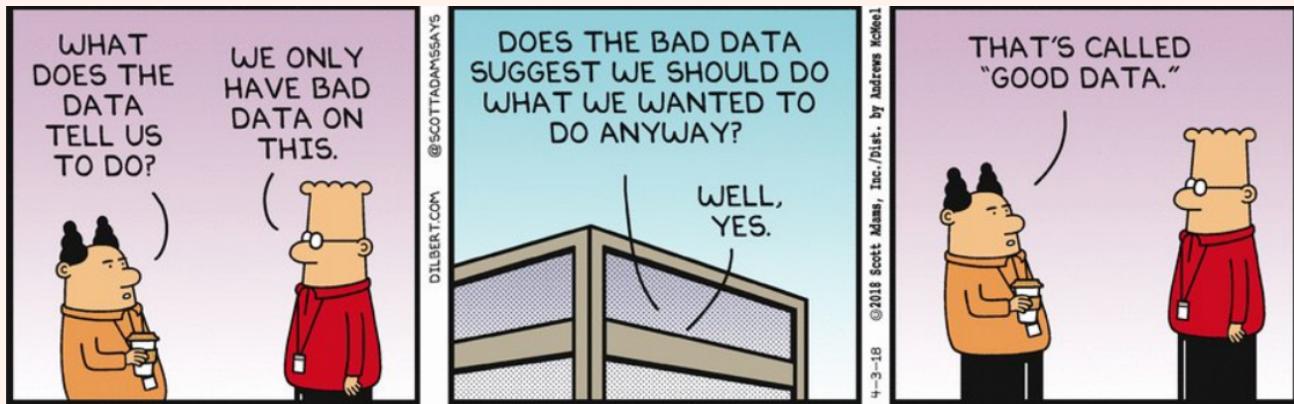
La mise en données du monde

- **Concrètement, aujourd'hui, quels aspects de votre vie sont mis en données ?**
- Recherches internet (**cookies**), pratiques sportives (**montres connectées**), consommation énergie (**compteurs connectés**), régime alimentaire (**appli type Yuka**), trajets dans les transports en commun (**Pass Navigo**)...
- Cette mise en donnée est rendue possible par le développement de **capteurs** qui viennent collecter et agréger ces données.
- La question devient peut-être, *quels aspects de votre vie ne sont pas (encore) mis en données ?*

Exemple : les données de bornage des téléphones

Le nouveau positivisme des données

- Attention, les données, même provenant de sources officielles, ne sont pas pour autant neutres, irréprochables ou porteuses de "LA" vérité



- Avez-vous des exemples de données officielles pouvant faire l'objet de critiques ?

Dilbert © Scott Adams

Open data et cadre juridique

1978 : La loi CADA, vers le "droit de savoir"

- Le fondement : la **Déclaration des Droits de l'Homme et du Citoyen** de 1789 dans son article 15, "la Société a le droit de demander compte à tout Agent public de son administration."
- Le droit d'accès des citoyens à l'information publique émerge en **1978 avec la loi dite CADA** du nom de la Commission d'Accès aux Documents Administratifs.
- La France était le **3e pays au monde** après la Suède et les Etats-Unis avec le Freedom of Information Act (FOIA) en 1966 à accorder un "droit de savoir" avec pour but d'améliorer les relations entre le public et l'administration.

Il faudrait maintenant désigner la loi comme le Code des relations entre le public et l'administration (CRPA) qui, dans son **livre 3**, codifie le droit d'accès et de réutilisation mais il est encore moins connu que la loi CADA...

Du droit d'accès à l'open data

Les "bases de données" : des documents administratifs ?

Le guide CADA-CNIL rappelle la définition des "bases de données"

On entend par base de données un recueil d'œuvres, de données ou d'autres éléments indépendants, disposés de manière systématique ou méthodique, et individuellement accessibles par des moyens électroniques ou par tout autre moyen (art L112-3 du code de la propriété intellectuelle)

Loi pour une République Numérique : l'ouverture des données par défaut

La loi pour une République Numérique impose un principe d'**ouverture des données par principe** qui ne fait pas l'objet de sanctions à toutes les administrations, les entreprises délégataires d'une mission de service public et les **collectivités locales de plus de 3500 habitants et 50 agents.**



Les 8 principes de l'open data

- 1/ **Des données complètes** : toutes les données publiques doivent être rendues disponibles dans les limites légales liées à la vie privée ou la sécurité.
- 2/ **Des données primaires** : les données ouvertes sont telles que collectées à la source, non-agrégées avec le plus haut niveau de granularité
- 3/ **Des données fraîches (*timely*)** : les données doivent être disponibles dès qu'elles sont produites
- 4/ **Des données accessibles** : les données doivent être utilisables par le plus grand nombre d'usagers potentiels

Les 8 principes de l'open data

5/ **Des données exploitables par les machines** : Les données peuvent être traitées automatiquement par les machines

6/ **Des données non discriminatoires** : Elles peuvent être utilisées par tous sans réclamer un enregistrement préalable

7/ **Des données dans un format ouvert** : Ce format ne doit pas être la propriété d'une organisation en particulier (.xls) et doit être gouverné par ses usagers (exemple : CSV)

8/ **Des données dans une licence ouverte** : Idéalement dans le domaine public sinon dans une licence conforme à l'[Open Definition](#) : Licence Ouverte (CC-BY) ou ODBL (CC-BY-SA)

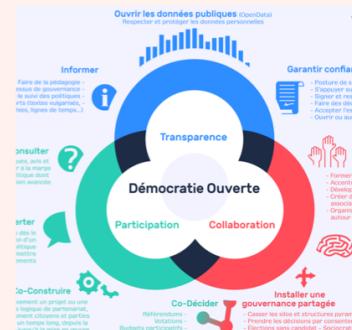
Les multiples facettes de l'open data



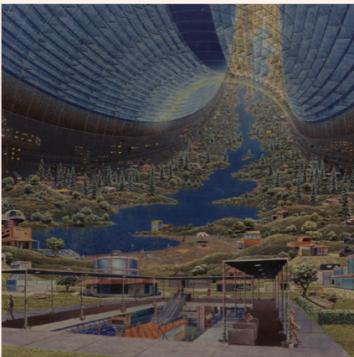
La transparence



Le partage des données en science



La transformation des administrations



La libre circulation de l'information



Les industries de la donnée

Ouverture par défaut... de toutes les données ?

Les données à caractère personnel sont évidemment exclues. Elles sont même encadrées par un règlement spécifique : Le règlement général sur la protection des données (RGPD).

Adopté à la mi-avril 2016 après 4 années de débat, 3 objectifs :

- Renforcer les droits des personnes.
- Responsabiliser les acteurs traitant des données personnelles des résidents européens.
- Crédibiliser la régulation.



Le RGPD : crédibiliser la régulation

-  Les autorités de protection peuvent notamment : limiter temporairement ou définitivement un traitement, suspendre les flux de données, ordonner la rectification, la limitation ou l'effacement des données
-  Amendes : jusqu'à 10 ou 20 millions d'euros, ou, dans le cas d'une entreprise, de 2% jusqu'à 4% du chiffre d'affaires annuel mondial, le montant le plus élevé étant retenu.
-  Sanction conjointement adoptée entre l'ensemble des autorités concernées, donc potentiellement pour le territoire de l'UE

Le RGPD : Renforcer les droits des personnes



Transparence et consentement

explicite : plus de lisibilité sur ce qui est fait de mes données (preuve de consentement explicite dans des termes clairs) et j'exerce mes droits plus facilement (droit d'accès, droit de rectification, droit à l'oubli).

Le RGPD : Renforcer les droits des personnes



Protection des mineurs : Les services en ligne doivent obtenir le consentement des parents des mineurs de moins de 16 ans avant leur inscription.

Source : [CNIL](#)

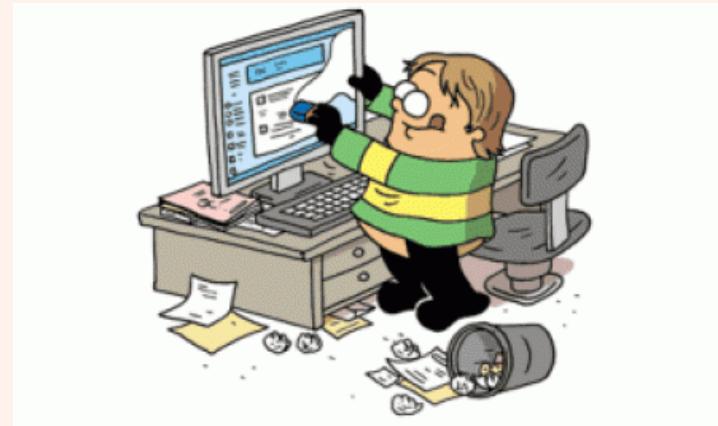


Guichet unique : En cas de problème, je m'adresse à l'autorité de protection des données de mon pays, quelque soit le lieu d'implantation de l'entreprise qui traite mes données.

Le RGPD : Renforcer les droits des personnes



Sanctions renforcées : En cas de violation de mes droits, l'entreprise responsable encourt une sanction pouvant s'élever à 4% de son chiffre d'affaires mondial.



Droit à l'oubli : Je peux demander à ce qu'un lien soit déréférencé d'un moteur de recherche ou qu'une information soit supprimée s'ils portent atteinte à ma vie privée.

Source : [CNIL](#)

Le RGPD : Responsabiliser les acteurs

- **Privacy by design** : protection des données personnelles dès la conception du produit et par défaut
- **Minimisation** : limiter la quantité de données personnelles dès le départ
- **Accountability** : mettre en place des mesures de protection des données et démontrer cette conformité à tout moment
- Fin des obligations déclaratives sauf si risque accru pour la vie privée

Ceux qui mangent les données



L'explosion des données ?

La conjonction de deux mouvements :

- la hausse exponentielle de la capacité à stocker et traiter des données numériques
- la hausse exponentielle de la volonté de collecter des données numériques

Il en résulte ce qui a été appelé le "big data":

- se définit par les trois V (selon la société Gartner) : "volume", "velocity", "variety" (+ "veracity" ?)
- Kitchin ajoute l'exhaustivité, la résolution, la *scalability*

L'explosion des données ?

Supercomputers and Smartphones

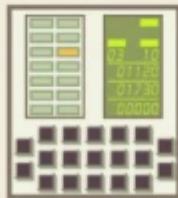
Take a look at your smartphone. The device you are holding is more powerful than the most advanced supercomputers of the early 1990s.
But that's not the only incredible statistic...

From A to B

Today's *TomTom Go GPS* computer runs at 500 Mhz. This is approximately 244 times faster than *NASA's Apollo Guidance Computer*, which navigated to the moon in 1966 at just 2.048 Mhz.



*TomTom Go
GPS computer*



*NASA's Apollo
Guidance Computer*

Gaming

Sony's PlayStation 4 will debut with 1.84 teraflops of raw computing power, 150 times the power of IBM's 1997 chess-grandmaster-beating *Deep Blue*.



*Sony
PlayStation 4*



*IBM
Deep Blue*

L'explosion des données ?

1996...

Alta Vista is a very large project, requiring the cooperation of at least 5 servers, configured for searching huge indices and handling a huge Internet traffic load. The initial hardware configuration for Alta Vista is as follows:

```
Alta Vista -- AlphaStation 250 4/266
    4 GB disk
    196 MB memory
    Primary web server for gotcha.com
    Queries directed to WebIndexer or NewsIndexer

NewsServer -- AlphaStation 400 4/233
    24 GB of RAID disks
    160 MB memory
    News spool from which news index is generated
    Serves articles (via http) to those without news server

NewsIndexer -- AlphaStation 250 4/266
    13 GB disk
    196 MB memory
    Builds news index using articles from NewsServer
    Answers news index queries from Alta Vista

Spider -- DEC 3000 Model 900 (replacement for Model 500)
    30 GB of RAID disk
    1GB memory
    Collects pages from the web for WebIndexer

WebIndexer -- Alpha Server 8400 5/300
    210 GB RAID disk (expandable)
    4 GB memory (expandable)
    4 processors (expandable)
    Builds the web index using pages sent by Spider.
    Answers web index queries from Alta Vista
```

Thank you,
Alta Vista Technical Support

L'explosion des données ?

2016...



De la statistique à la *data science*

- la statistique est une relativement vieille science (développement au 18e siècle), pour aider les États (*Statistik*) à compter (les contribuables, les soldats potentiels...) mais aussi des entreprises privées (au départ, les assureurs => actuariat)
- la statistique repose sur une branche des mathématiques, les probabilités, qui émerge au milieu du 17e siècle, avec Pascal et Fermat notamment.
- c'est pourquoi la statistique est une discipline pratiquée par des mathématiciens, avec une importante formalisation mathématique.
- la pratique de la statistique recouvre une forte dimension théorique : on part de problèmes théoriques, et de données d'illustrations, plutôt que de données et de problèmes réels.

Au commencement était la statistique

I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

Je dis tout le temps que le métier sexy dans les dix ans à venir sera celui de statisticien. Les gens pensent que je plaisante, mais qui aurait pu deviner que les ingénieurs informatiques auraient été le métier sexy des années 1990 ?

Hal Varian (économiste en chef, Google), *The McKinsey Quarterly*, January 2009

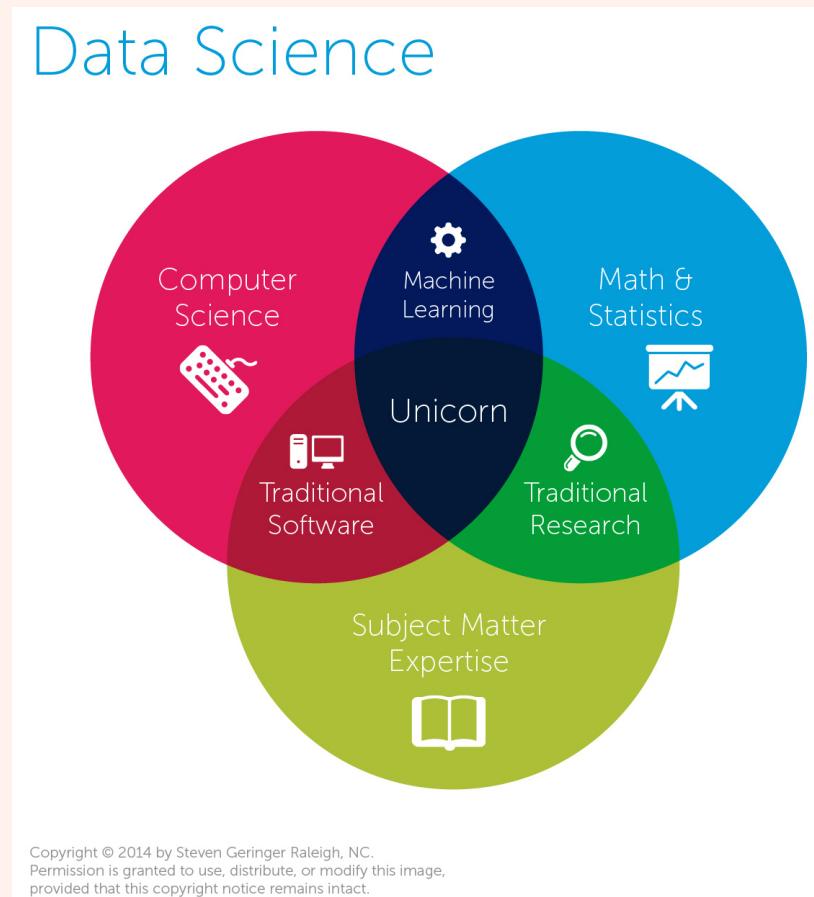
Data science is the new statistics?

I think data-scientist is a sexed up term for a statistician

Je pense que data scientist est un terme sexy pour dire statisticien

Nate Silver

Data science is the new statistics?



Data science is the new statistics?

La data science, comparativement à la statistique "traditionnelle", est un métier de praticien, presque de bidouilleur : elle nécessite des compétences mathématiques et statistiques, certes, mais aussi une compétence "métier" (compréhension du domaine d'application) et une solide maîtrise de l'informatique.

Changement de paradigme : le *machine learning*

- statistique classique : les problèmes doivent pouvoir être résolus de manière analytique, sans puissance de calcul particulière (d'où le succès du fréquentisme)
- le développement de la puissance de calcul permet de résoudre des problèmes statistiques par la simulation (**MCMC**) : on n'a pas besoin de connaître la solution mathématique, il "suffit" de faire de nombreuses simulations aléatoires.

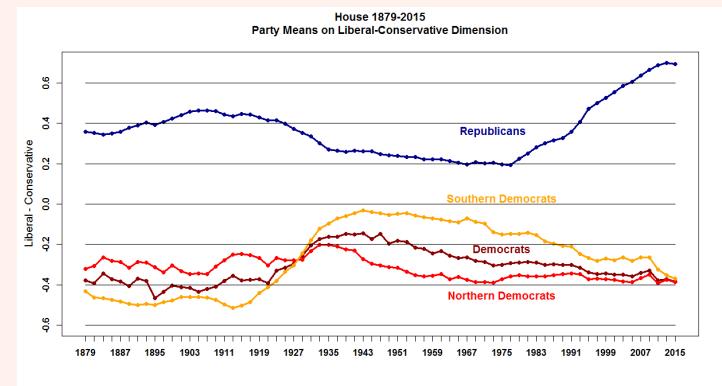
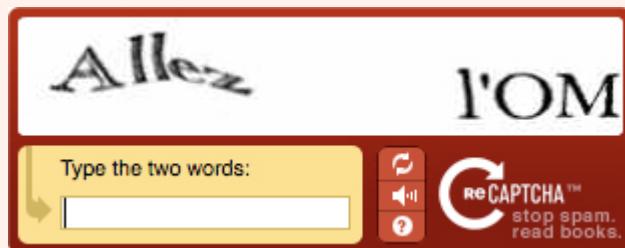
Changement de paradigme : le machine learning

- Fondamentalement, modélisation et machine learning ne sont pas différents, du point de vue d'un statisticien : modéliser un Y en fonction d'un vecteur de X_i
- Une des différences principales toutefois : veut-on *prévoir* ou *comprendre/analyser* ?
- Et donc : peut-on, veut-on interpréter les coefficients ?

En pratique : le machine learning porte sur des données plus complexes que la modélisation traditionnelle, avec souvent beaucoup de valeurs manquantes.

Changement de paradigme : le *machine learning*

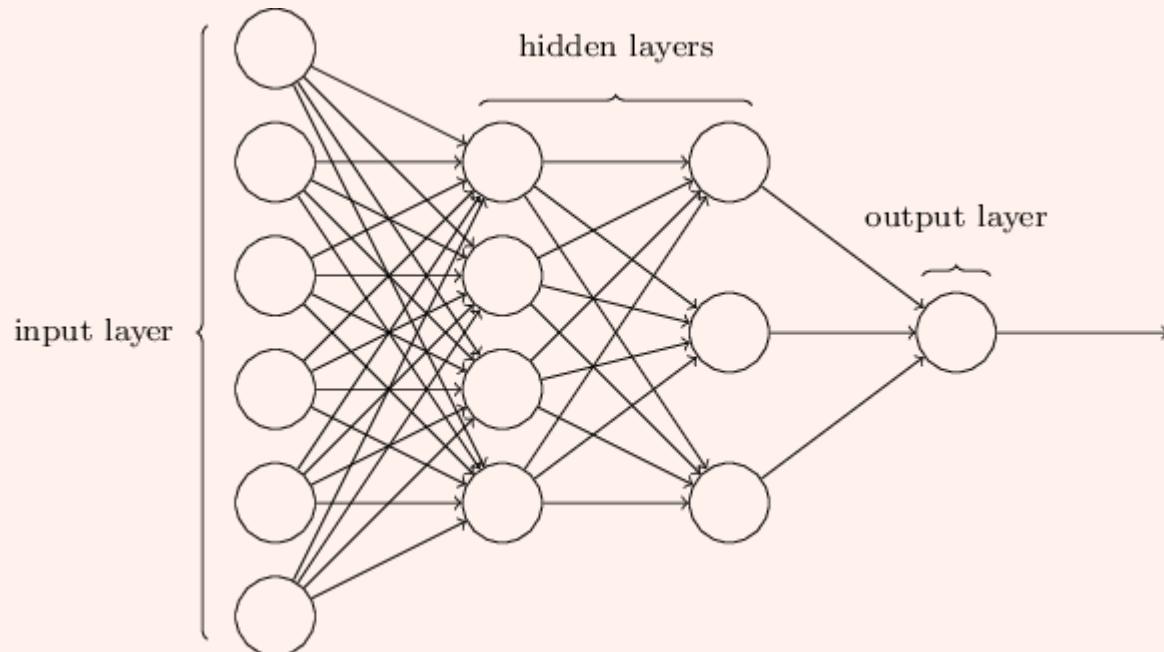
Apprentissage supervisé vs non-supervisé



Ce qu'on appelle aujourd'hui l'intelligence artificielle

Un terme ambigu et recouvrant des réalités socio-techniques très diverses, de la fonction Excel au modèle GPT-3 ou Dall-E en passant par des Mechanical Turks...

Une classe de modèles importante : le *deep learning*



Algorithmes, IA, code source

- Un **algorithme** "est la description d'une suite d'étapes permettant d'obtenir un résultat à partir d'éléments fournis en entrée" (**CNIL**).
- Cet algorithme est considéré comme un **algorithme public** (au sens de la loi pour République numérique) lorsqu'il est utilisé dans le cadre d'une mission de service public, en particulier pour prendre une décision administrative individuelle.
- Un algorithme peut aller de modèles procéduraux très simples à des modèles d'IA très complexes -- avec un rapport variable entre importance des règles (calcul de l'impôt) et importance des données d'entraînement (détection de la fraude fiscale) (modèles procéduraux vs modèles auto-apprenants).
- Le **code source** désigne la manière dont un algorithme est traduit dans une suite concrète d'instructions informatiques, dans un langage informatique donné.

S'agissant de modèles auto-apprenants/de *machine learning* supervisé, **l'ouverture du code source ne suffit pas à sa transparence** : il est très dépendant des données d'entraînement en amont, et se caractérise par ses *poids* (le modèle entraîné) en aval.

Temps de questions- réponses

**Posez vos questions dans le
chat !**

Merci !

Contact : joel@dataactivist.coop