Big Data

Sergey Kirgizov

Plan de cours

8 CM, 4 séances TD, 7 séances TP

- Introduction, Histoire, Motivation
- Stockage et modélisation de données
- La complexité des algorithmes et des données
- Méthodes d'analyse : K-Means et PCA
- Architecture
- Visualisation
- Traitement éthique de données, sécurité

Le lien !

Cours, TP, TD, datasets

`https://kirgizov.link/teaching/esirem/bigdata`

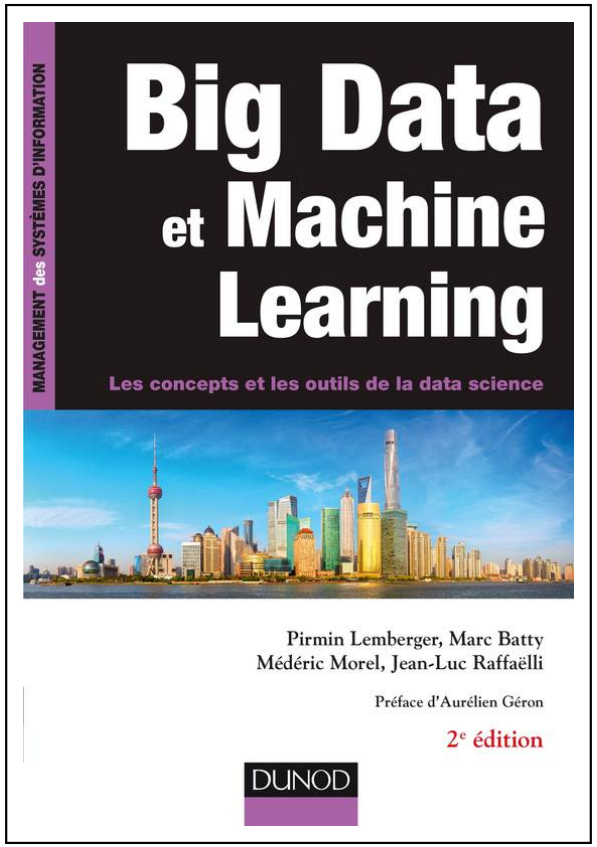
Les notes

- un contrôle intermédiaire
- un contrôle terminal
- un projet à faire pendent les TPs

Livres

Data et Machine Learning : Les concepts et les outils de la data science.

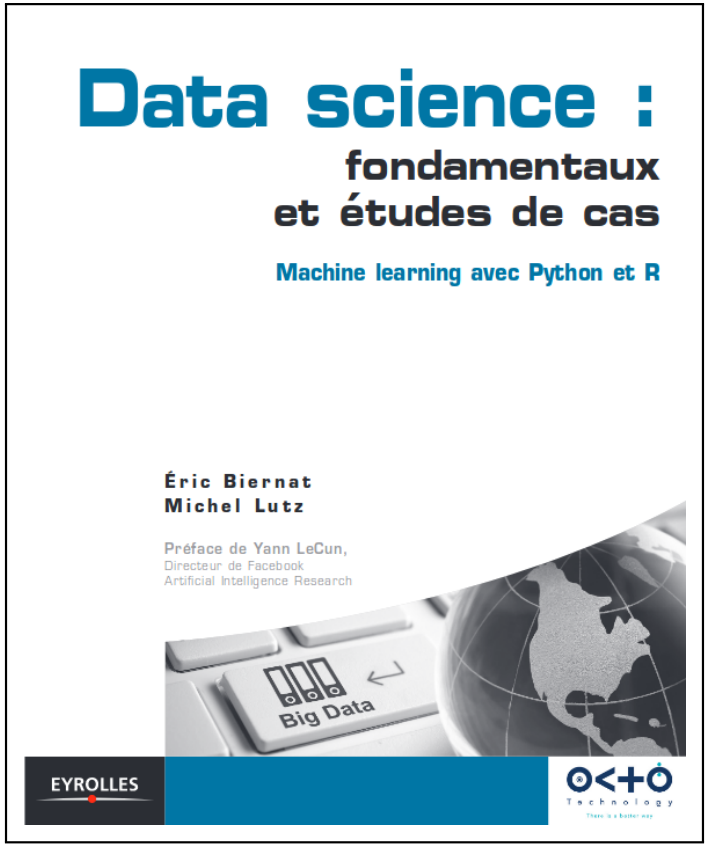
Jean-Luc Raffaëlli, Médéric Morel, Marc Batty, Pirmin Lemberger



Livres

Data science : fondamentaux et études de cas : Machine Learning avec Python et R

Michel Lutz, Eric Biernat



Autres liens

Big Data, Machine Learning : qu'est-ce que la science des données ?

Aurélien Garivier
Professeur à l'École Normale Supérieure de Lyon.

`http://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/mydocs/IREM201701.pdf`

`http://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/indexa069.html?q=node/113`

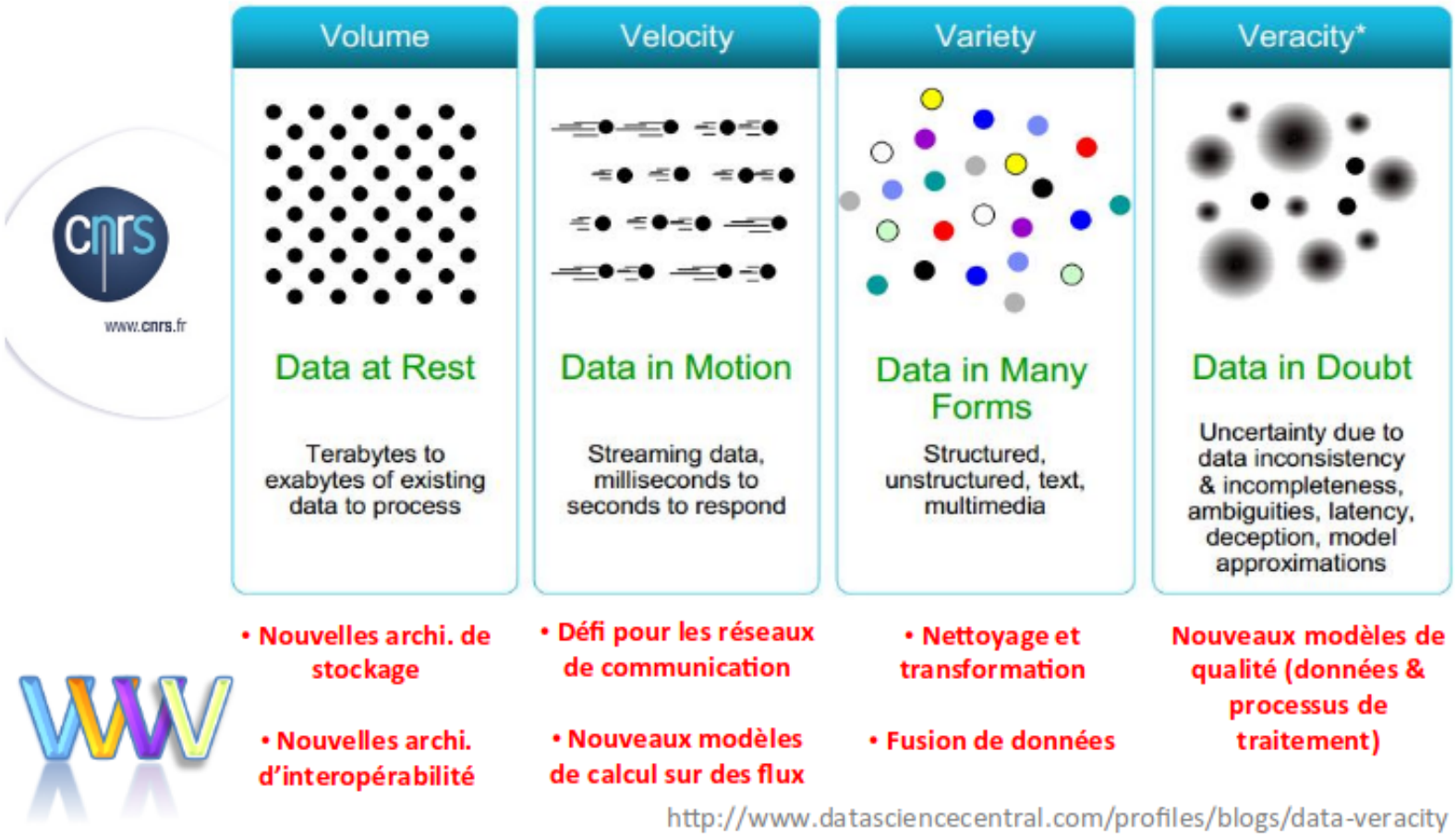
coursera

`https://www.coursera.org/courses?query=bigdata`

Big data...

Big data...
qu'est-ce que c'est ?

Complexité multidimensionnele des Big Data



[Bouzeghoub, Mastodons : Une approche interdisciplinaire des Big Data]

Big Data = VVVV?!

Combien de lettres “V” contient Big Data ?

- V V V = volume, vélocité et variété

  Un rapport de recherche par Doug Laney du META Group en 2001.

- V V V V volume, vélocité, variété, véracité

  IBM, 2011

- V V V V V = volume, vélocité, variété, véracité, valeur

  “The Missing V’s in Big Data : Viability and Value” by Neil Biehn. Wired 2013

- V V V V V V V = 5V + visualisation + variabilité

  “Understanding Big Data : The Seven V’s” by Eileen McNulty. Dataconomy 2014

Jeu de langage de Ludwig Wittgenstein ?

"... Le sens n'apparaît donc que dans un contexte concret. Ceci signifie que nous n'apprenons pas le sens des mots que nous utilisons en apprenant des concepts mais dans la pratique du langage ..."

`https://fr.wikipedia.org/wiki/Jeu_de_langage_(philosophie)`

Histoire d'informatique = histoire du big data

Il y 20 000 ans. Os d'Ishango, Congo

https://fr.wikipedia.org/wiki/Os_d'Ishango

Il y a 2700 ans. Bibliothèque d'Assurbanipal de l'Assyrie antique

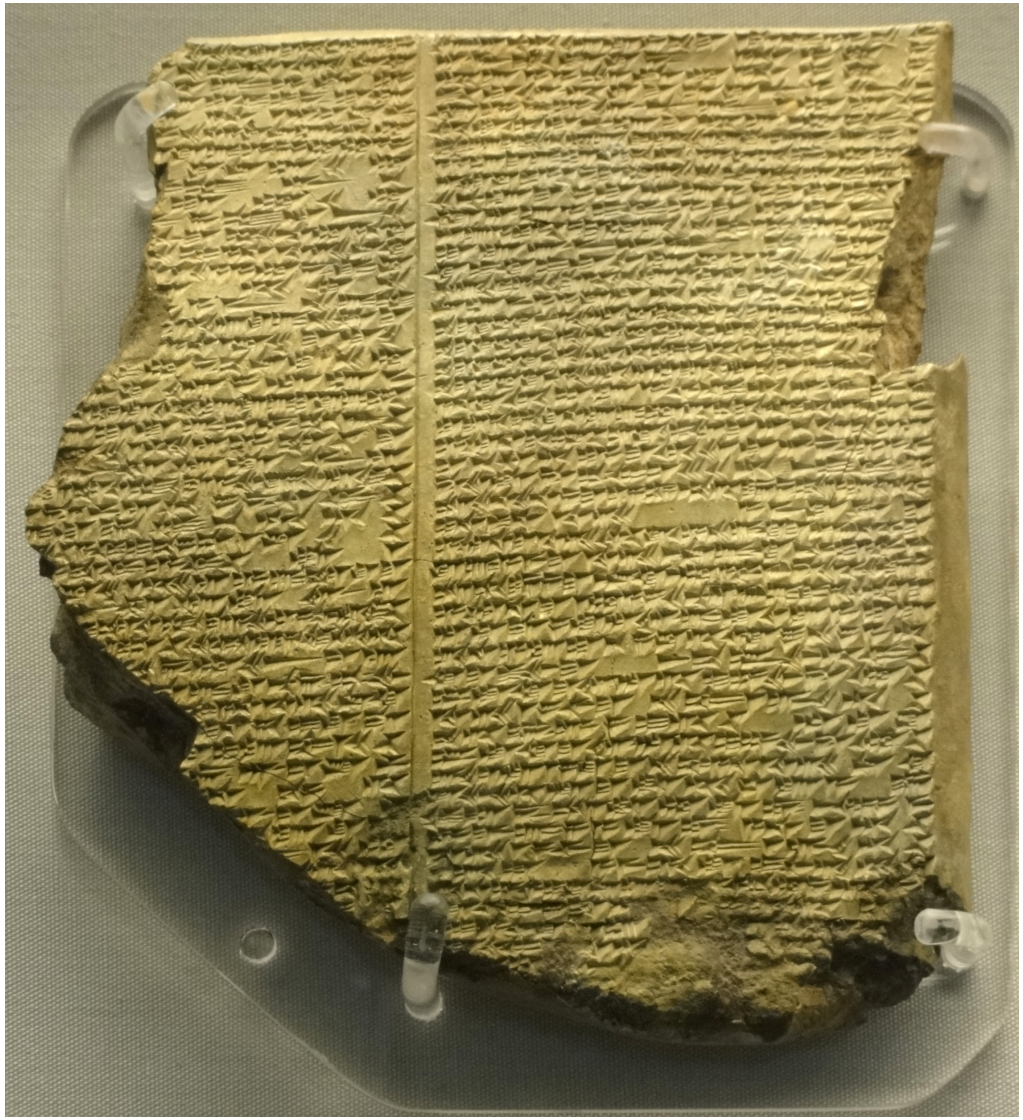
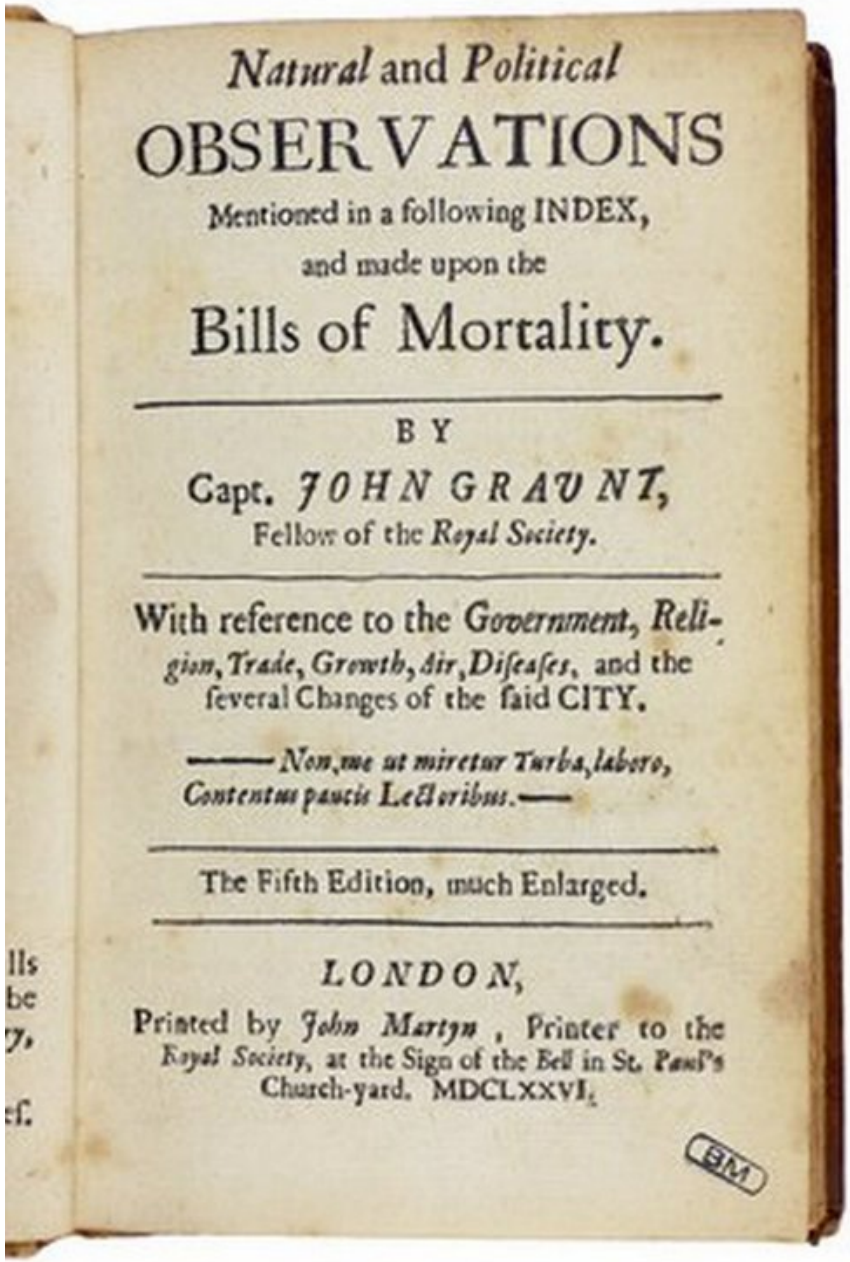


Photo par Fæ de Wikipedia

“The tablets were often organized according to shape : four-sided tablets were for financial transactions, while round tablets recorded agricultural information.(In this era, some written documents were also on wood and others on wax tablets.) Tablets were separated according to their contents and placed in different rooms : government, history, law, astronomy, geography, and so on. The contents were identified by colored marks or brief written descriptions, and sometimes by the "incipit," or the first few words that began the text”

https://en.wikipedia.org/wiki/Library_of_Ashurbanipal

Il y a 350 ans. John Graunt



Natural and *Political*

OBSERVATIONS

Mentioned in a following INDEX,
and made upon the

Bills of Mortality.

BY

Capt. *JOHN GRAUNT*,
Fellow of the *Royal Society*.

With reference to the *Government*, *Religion*, *Trade*, *Growth*, *Air*, *Diseases*, and the several Changes of the said CITY.

——*Non, me ut miretur Turba, laboro,*
Contentus paucis Lectoribus.——

The Fifth Edition, much Enlarged.

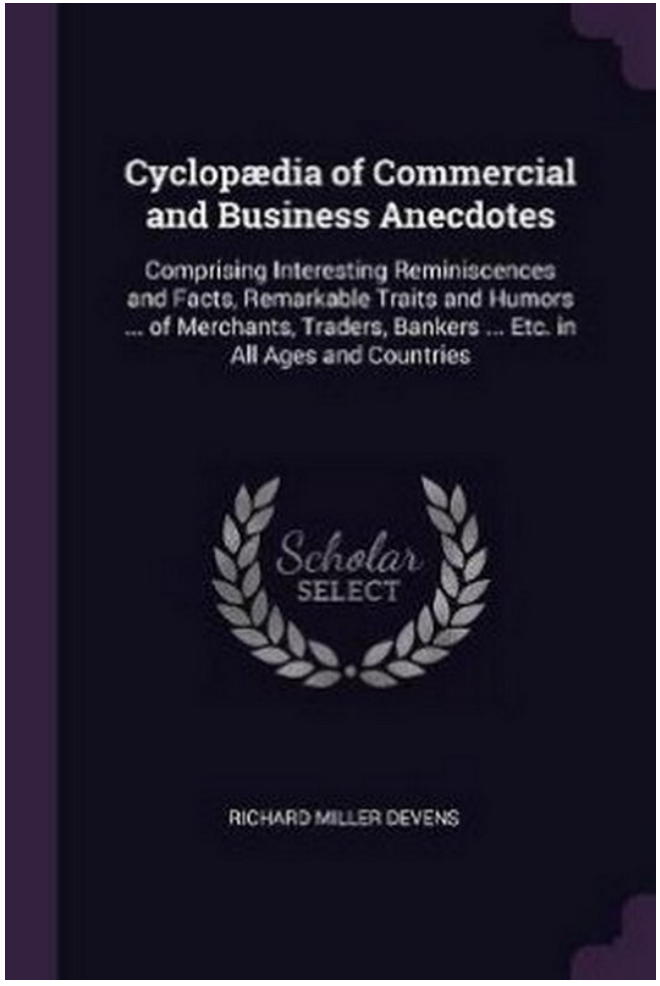
LONDON,
Printed by *John Martyn*, Printer to the *Royal Society*, at the Sign of the Bell in St. *Paul's* Church-yard. MDCLXXVI.

Afin d'essayer de mettre au point un système pour détecter l'apparition de la peste bubonique à Londres, il avait analysé les bulletins de décès publiés hebdomadairement dans la capitale anglaise sous le règne de Charles II.

– `https://fr.wikipedia.org/wiki/John_Graunt`

Il y a 155 ans. “Business Intelligence”

L’apparition du terme “business intelligence”, connu aujourd’hui comme “l’informatique décisionnelle”.



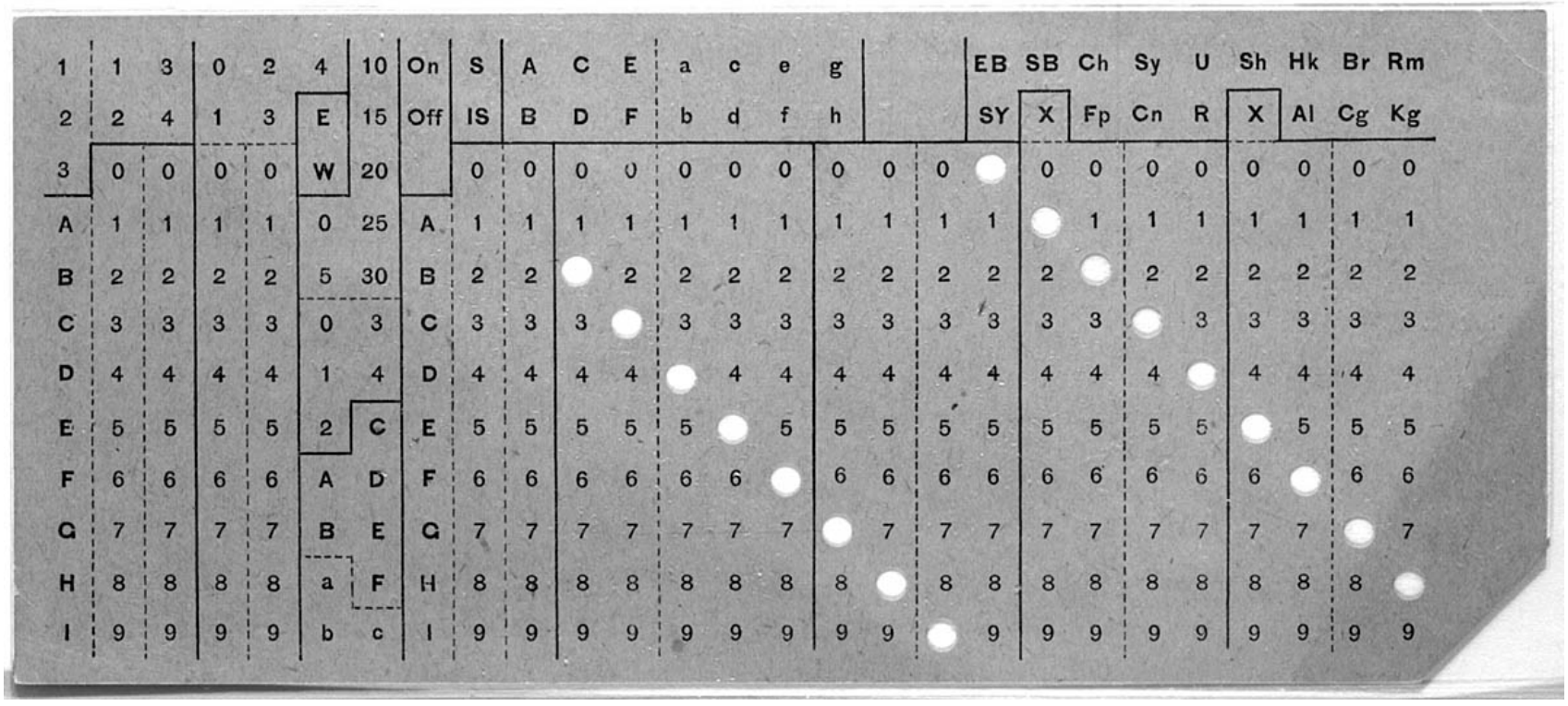
Richard Millar Devens a utilisé le terme “business intelligence” pour décrire comment le banquier Sir Henry Furnese a réalisé des bénéfices en recevant et en agissant sur la base d’informations sur son environnement, avant ses concurrents.

> *Throughout Holland, Flanders, France, and Germany, he maintained a complete and perfect train of business intelligence. The news of the many battles fought was thus received first by him [banker Sir Henry Furnese], and the fall of Namur added to his profits, owing to his early receipt of the news.*
>
> — R. M. Devens
> Cyclopædia of Commercial and Business Anecdotes (1865)

`http://en.wikipedia.org/wiki/Business_intelligence#History`

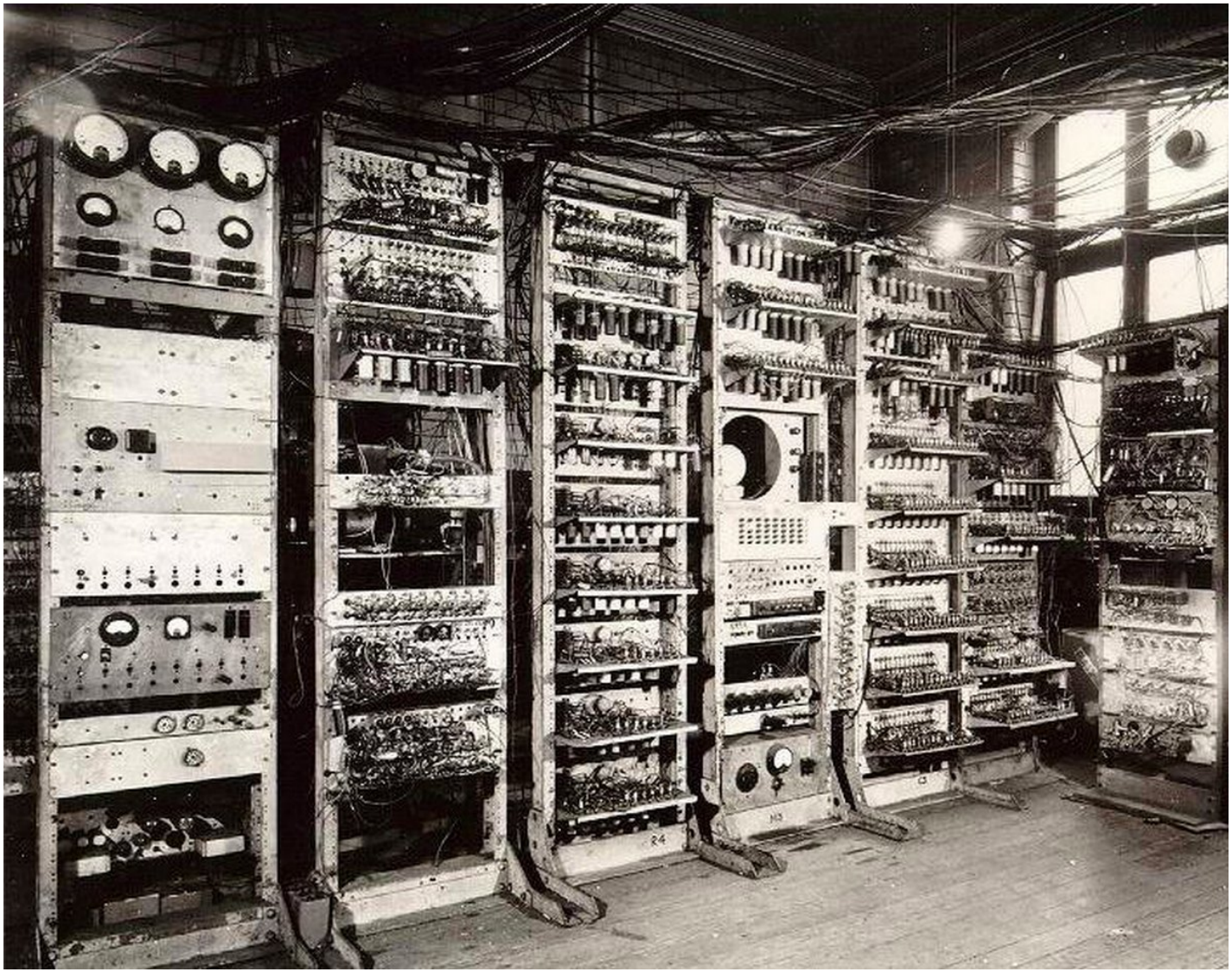
1890. Une tabulatrice de Herman Hollerith

Photo IBM d'Adam Schuster.



- Recensement des États-Unis d'Amérique de 1890 pourrait durer 8 ans...
- À l'aide de cartes perforées, la tabulatrice réduit la main-d'œuvre de 10 ans à 3 mois !
- Hollerith a crée une entreprise “TMC”, il a été achetée par C-T-R (devenu IBM).

Colossus britannique de 1943



“Colossus est une série de calculateurs électroniques fondé sur le système binaire. Le premier, Colossus Mark 1, est construit en l’espace de onze mois et opérationnel en décembre 1943, par une équipe dirigée par Thomas “Tommy” Flowers et installé près de Londres, à Bletchley Park : constitué de 1 500, puis 2 400 tubes à vide, il accomplissait 5 000 opérations par seconde. Il était utilisé pendant la Seconde Guerre mondiale pour la cryptanalyse du code Lorenz. Ce code était utilisé par les hauts dirigeants allemands pour communiquer entre eux alors qu’Enigma était utilisée au quotidien pour les autres types de communication. Plus rapide, le Colossus Mark II servit notamment pour le lancement surprise du Débarquement.”

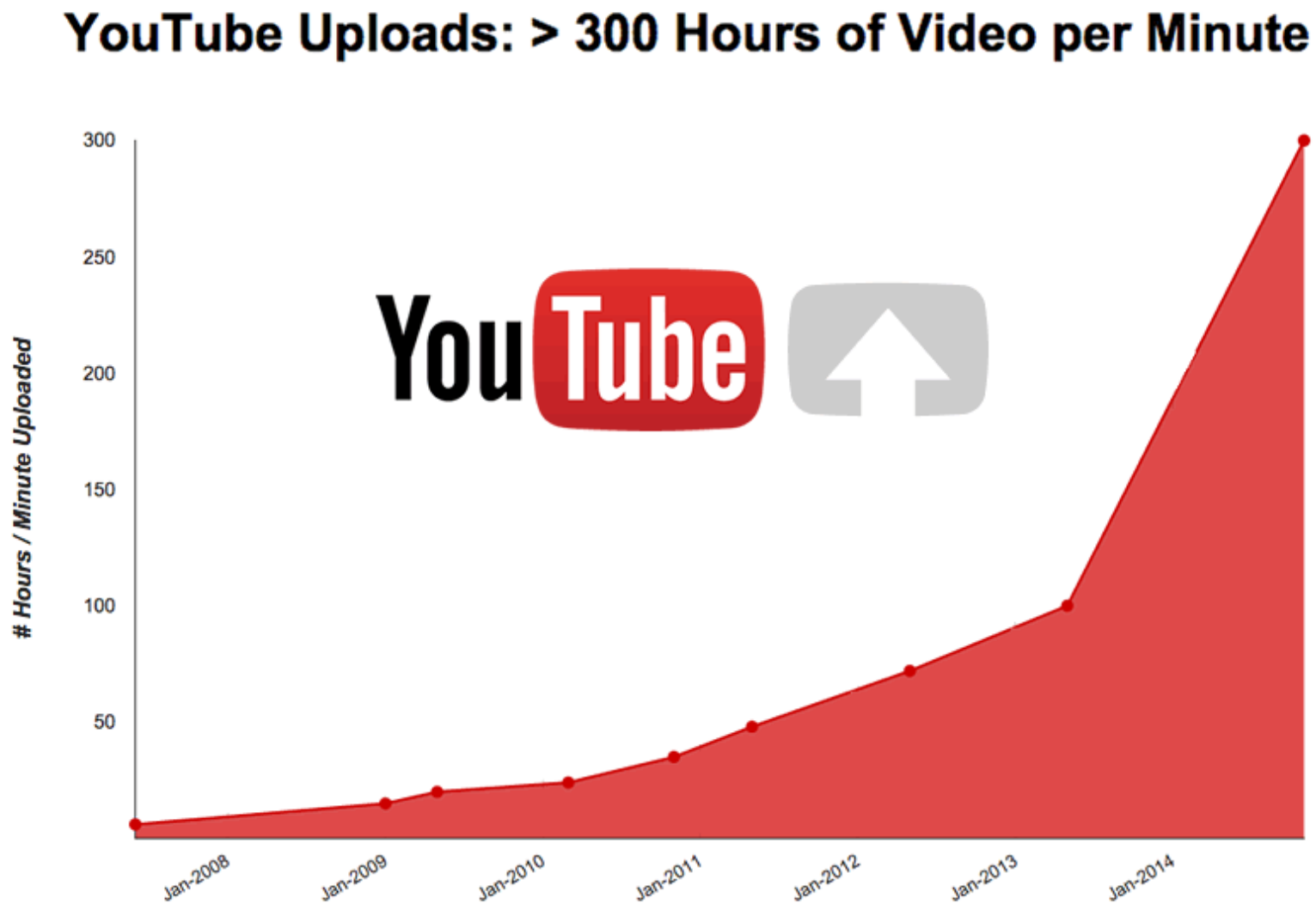
`https://fr.wikipedia.org/wiki/Colossus_(ordinateur)`

Ok pour les notes historiques

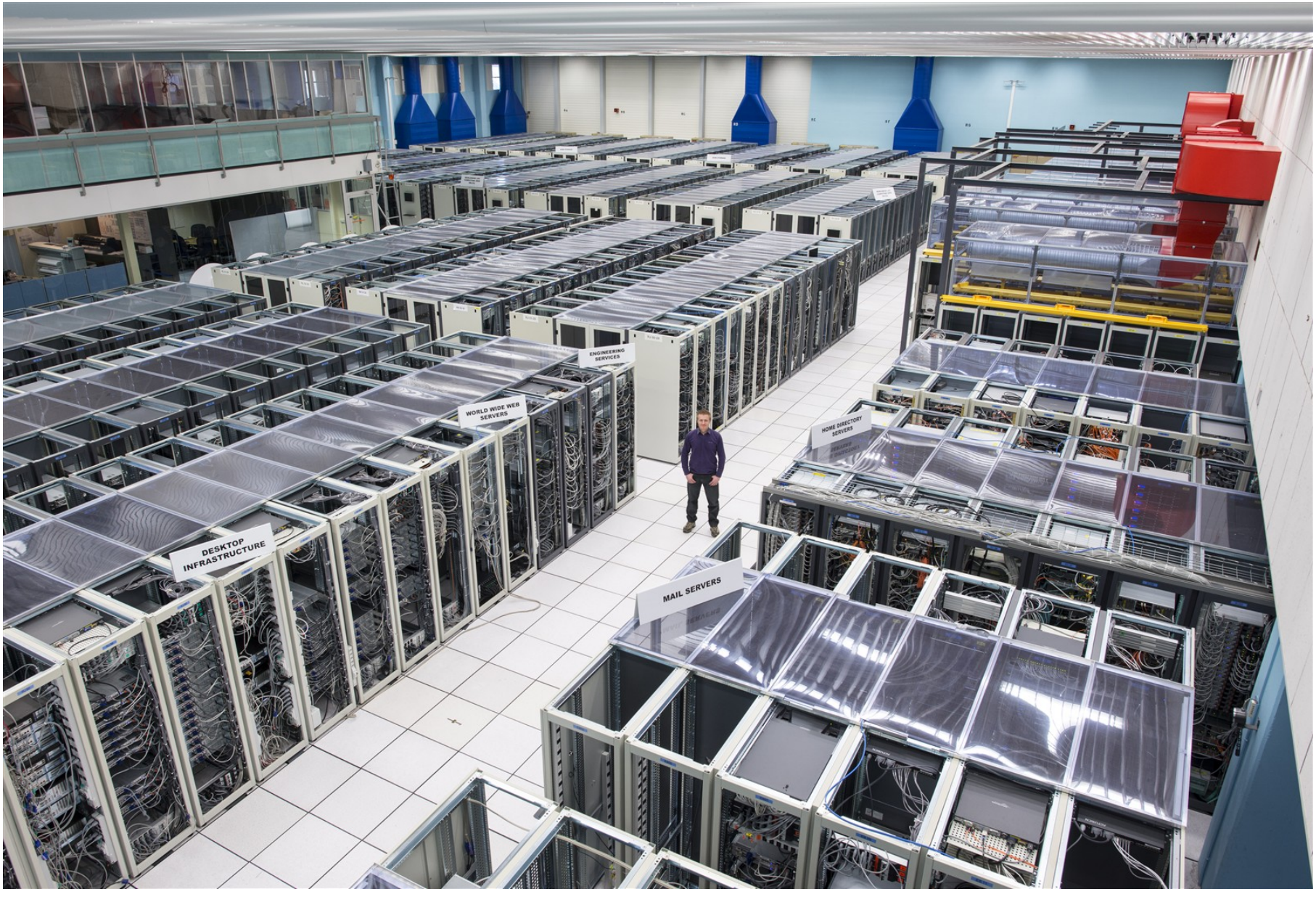
et aujourd'hui ?


BA-DUM-TSS!
The Oatmeal

Un exemple de notre temps



https://tubularinsights.com/youtube-300-hours/



Le système de stockage du CERN EOS Open Storage. `http://eos.web.cern.ch/`
Les expériences du LHC produisent de l'ordre de 90 pétaoctets de données par an.

Ce n'est pas évident...
Comment on va faire ?

Motivation

Big data $\Rightarrow$ Big Money ?

Big data $\Rightarrow$ Big Power ?

Big data $\Rightarrow$ Big Bonheur ?

Big data $\Rightarrow$ Beaucoup d'Intéressant ?

Entreprises

Ok, nous avons collecté beaucoup de données de nos clients :

- leurs mails
- leur historique d'achats
- leur consommation énergétique
- leurs transactions bancaires
- les jeux auxquels ils jouent
- leurs coordonnées GPS
- liste de leurs amis, historique de la navigation web
- etc

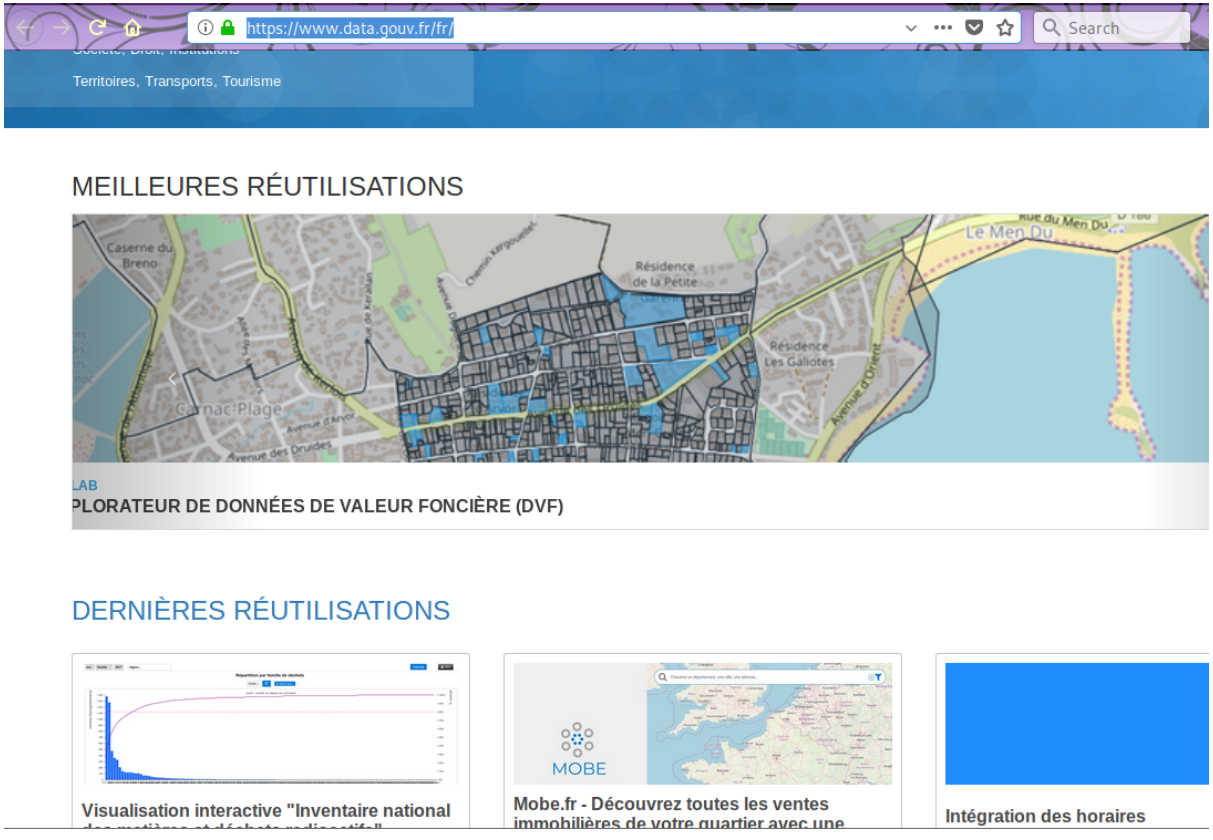
Que faire maintenant avec ça ?
Comment transformer ces données en argent ?

Science

- “Décoder le premier génome humain a nécessité 10 ans, mais prend aujourd’hui moins d’une semaine”
- “Le NASA Center for Climate Simulation (NCCS) stocke 32 Po de données d’observations et de simulations climatiques”
- CERN et son “Large Hadron Collider“
- SETI@home — recherche d’une intelligence extraterrestre.
- etc

Les méthodes statistiques et beaucoup des algorithmes “machine learning” ils marchent bien mieux quand on a beaucoup des données.

Mouvement “OPEN DATA”



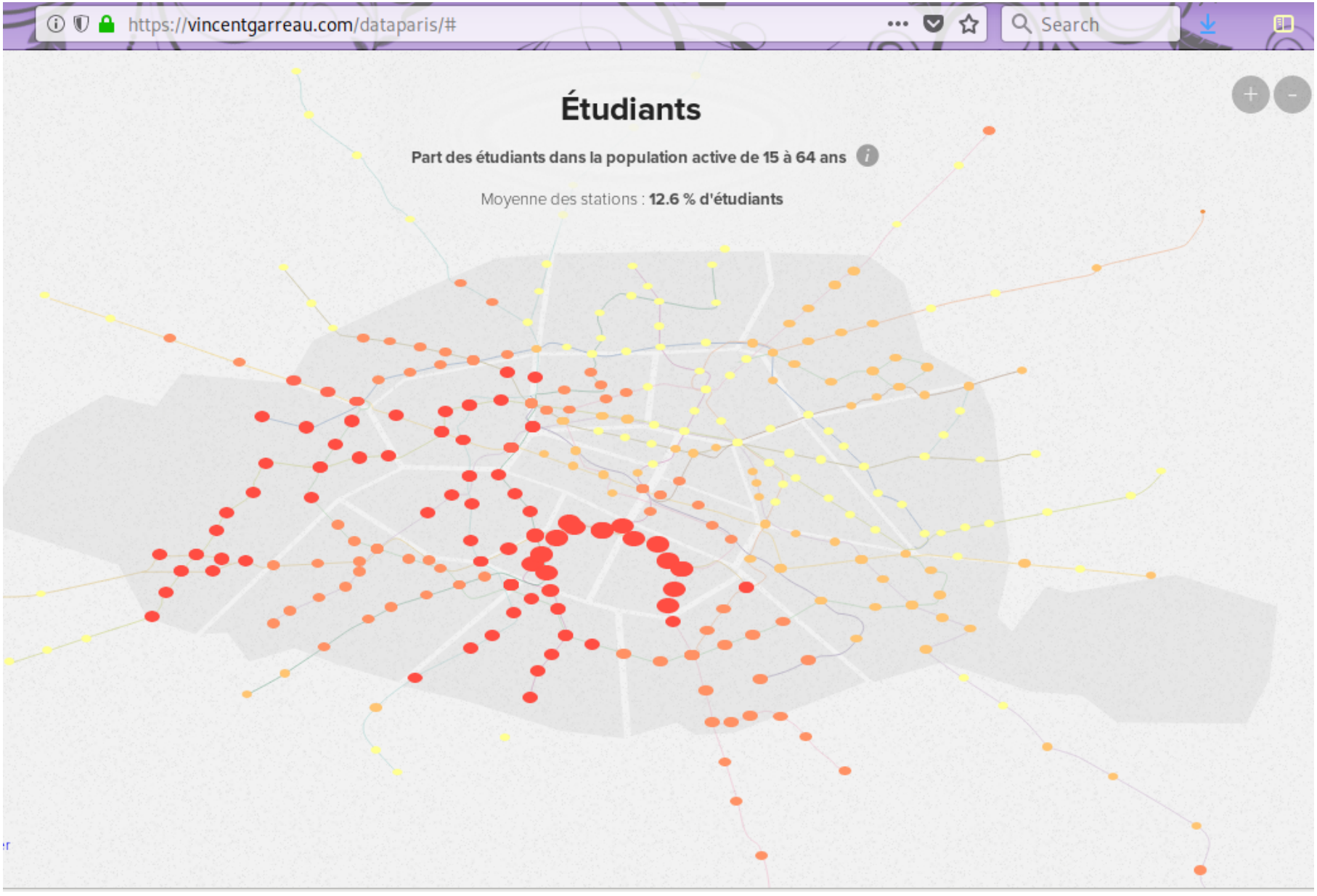
https://www.data.gouv.fr/fr/

Service public de la donnée

Des données sur lesquelles vous pouvez compter

“Le service public de la donnée créé par l’Article 14 de la loi pour une République numérique vise à mettre à disposition, en vue de faciliter leur réutilisation, les jeux de données de référence qui présentent le plus fort impact économique et social. Il s’adresse principalement aux entreprises et aux administrations pour qui la disponibilité d’une donnée de qualité est critique. Les producteurs et les diffuseurs prennent des engagements auprès de ces utilisateurs. La mission Etalab est chargée de la mise en oeuvre et de la gouvernance de ce nouveau service public. Elle référence l’ensemble des données concernées sur cette page.”

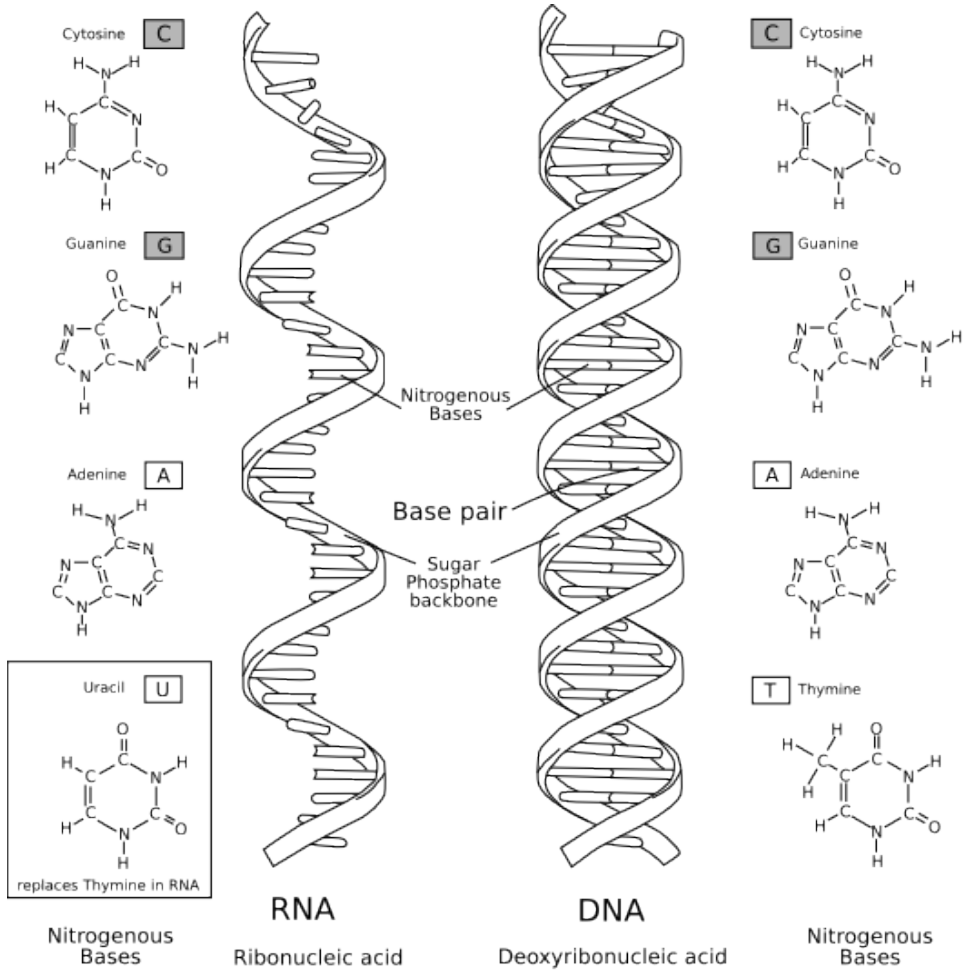
`https://www.data.gouv.fr/fr/reference`



https://vincentgarreau.com/dataparis/

Stockage et modélisation de données

Le stockage des données



by Antilived, Fabiolib, Turnstep, Westcairo @ Wikipedia
`https://fr.wikipedia.org/wiki/Big_data#Histoire`

Combien de bits pour représenter un ADN humain ?

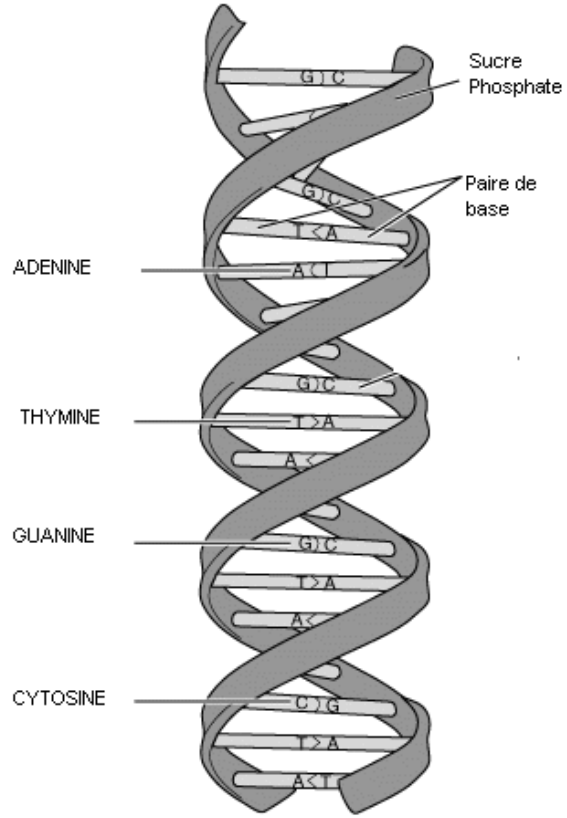
810 Mo

Combien de bits pour représenter l'ADN d'une amibe ?

157 Go

Paire de bases — unité de mesure d'information ADN

Une paire de bases (pb) est l'appariement de deux bases nucléiques situées sur deux brins complémentaires d'ADN.



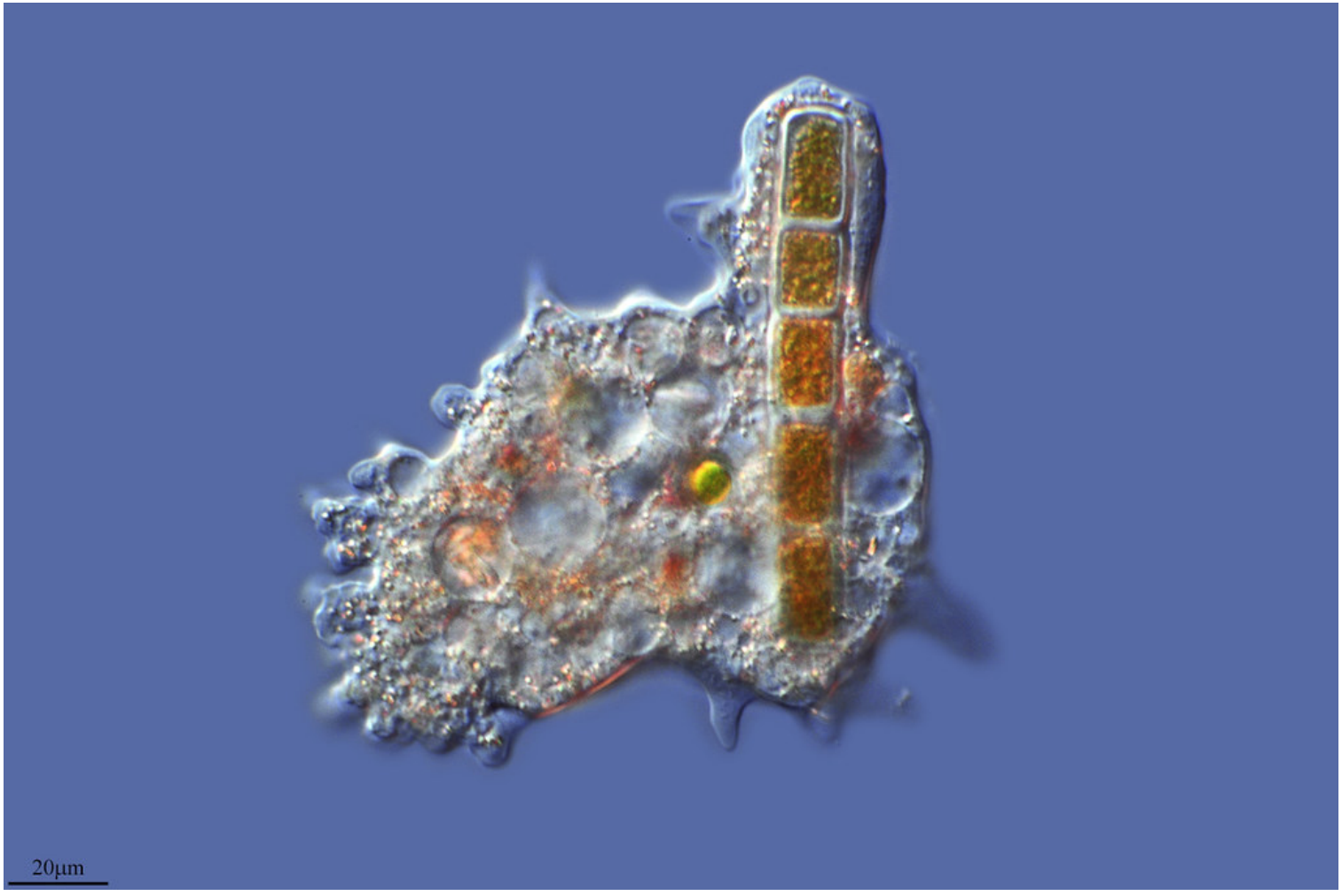
`https://fr.wikipedia.org/wiki/Paire_de_bases`

Différents nucléotides

- A — adénine
- G — guanine
- T — thymine
- C — cytosine

1 Mpb = 1 000 000 pb.
1 Mpb ≈ 244 Kilooctes

Amoeba dubia, est une amibe microscopique qui se distingue par la taille de son génome qui serait le plus grand du monde vivant connu avec 675 milliard de paires de bases.

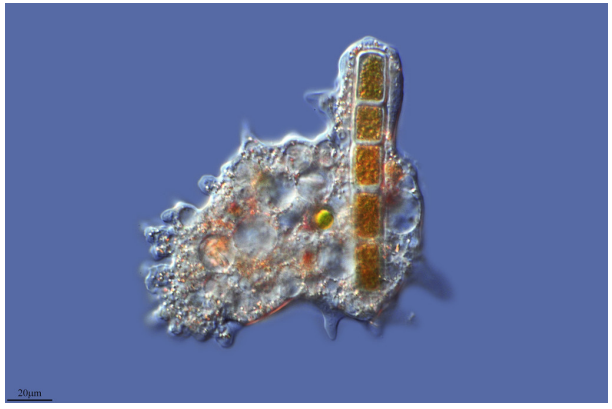


https://www.flickr.com/photos/microagua/34096401014

Homo sapiens

3 400 Mpb $\approx$ 810 Mo

Amoeba dubia



675 000 Mpb $\approx$ 157 Go

Stockage et modélisation de données

- Bit
  - 0 ou 1
  - Vrais ou Faux
  - Absence / présence

Stockage et modélisation de données

- Bit
  - 0 ou 1
  - Vrais ou Faux
  - Absence / présence
- Mots binaires, par exemple un octet, “byte”

Stockage et modélisation de données

- Bit
  - 0 ou 1
  - Vrais ou Faux
  - Absence / présence
- Mots binaires, par exemple un octet, “byte”
- Structures : tableau, arbres, listes, graphes, série temporelles

Stockage et modélisation de données

- Bit
  - 0 ou 1
  - Vrais ou Faux
  - Absence / présence
- Mots binaires, par exemple un octet, “byte”
- Structures : tableau, arbres, listes, graphes, série temporelles
- Relations entre les structures

Stockage et modélisation de données

- Bit
  - 0 ou 1
  - Vrais ou Faux
  - Absence / présence
- Mots binaires, par exemple un octet, “byte”
- Structures : tableau, arbres, listes, graphes, série temporelles
- Relations entre les structures
- Systèmes (de gestion) de structures :
  - Bases de données SQL
  - Bases de données NoSQL
- Système d’information autour de ces structures.

La complexité des algorithmes et des données

La complexité des données

Comment mesurer la complexité ?

Comment mesurer la complexité ?

Comment mesurer la complexité ?

Par la taille ? !

1. abababababababababababababababa
2. 12345678910111213141516171819 20
3. bwert2362223767465866253542514 5

1. `abababababababababababababababa`
2. `12345678910111213141516171819 20`
3. `bwert2362223767465866253542514 5`

La taille est la même, complexité différente !

Comment mesurer la complexité ?

Comment mesurer la complexité ?

Comment mesurer la complexité ?

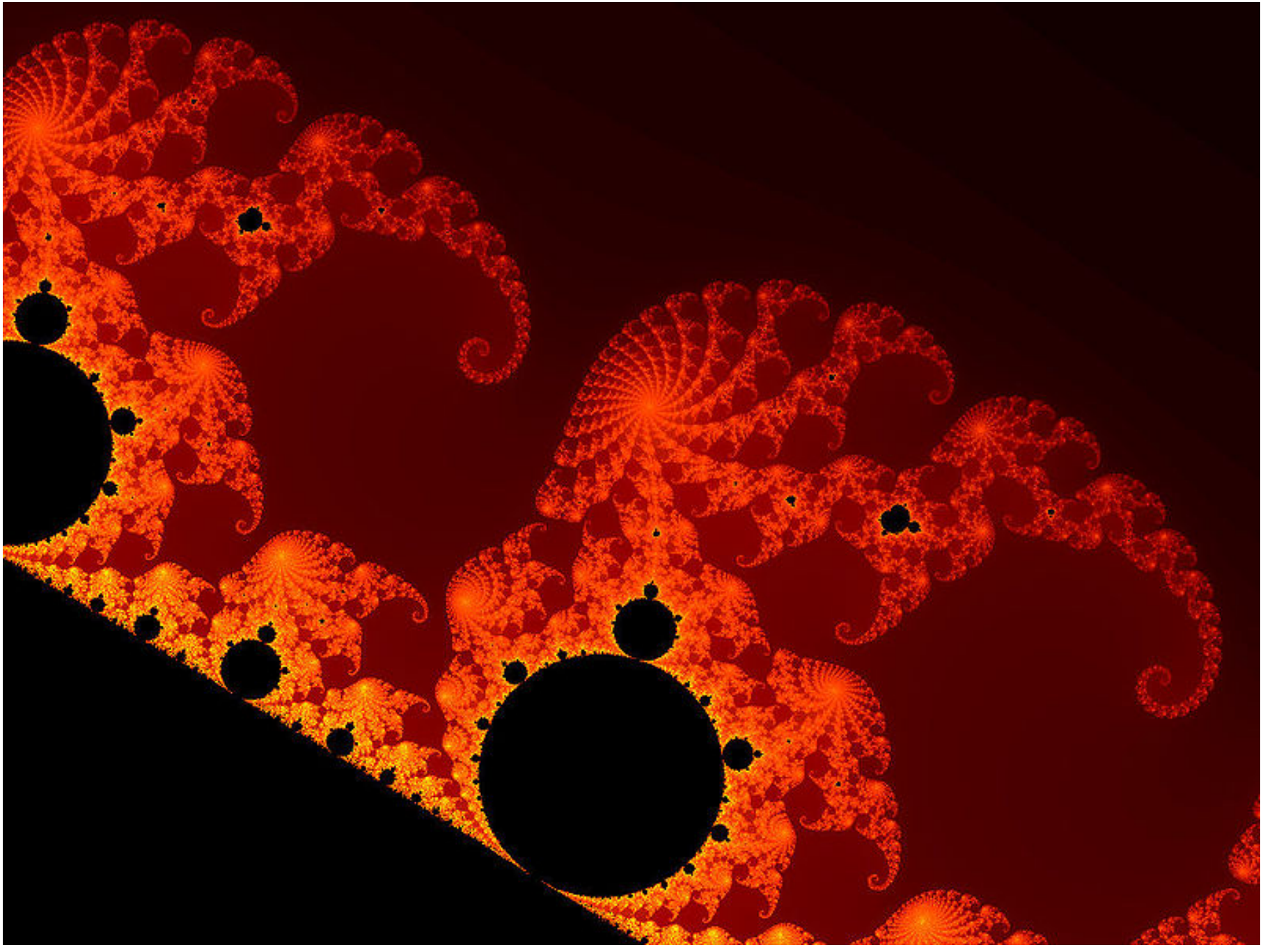
Par la compressibilité ? !

Comment mesurer la compressibilité ?

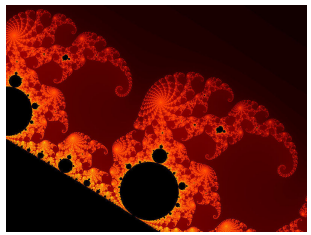
Comment mesurer la compressibilité ?

Comment mesurer la compressibilité ?

Par un “.zip” ? !



Ensemble de Mandelbrot



“L’ensemble de Mandelbrot est une fractale définie comme l’ensemble des points c du plan complexe pour lesquels la suite de nombres complexes définie par récurrence par :

$$\begin{cases} z_0 = 0 \\ z_{n+1} = z_n^2 + c \end{cases}$$

est bornée”

— Wikipedia

Complexité de Kolmogorov

Informalement, la complexité de Kolmogorov “La longueur du plus petit programme qui génère les données en question”.

`https://fr.wikipedia.org/wiki/ComplexitÃl'_de_Kolmogorov`

La complexité de Kolmogorov n'est pas calculable.

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

Démonstration.
Raisonnement par l'absurde. Supposons qu'une telle fonction $K(X)$ existe.

```
fonction K(n) {...}
```

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

Démonstration.
Raisonnement par l'absurde. Supposons qu'une telle fonction $K(X)$ existe.

```
fonction K(n) {...}
```

La taille de cette fonction est k charactères.

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

Démonstration.
Raisonnement par l'absurde. Supposons qu'une telle fonction $K(X)$ existe.

```
fonction K(n) {...}
```

La taille de cette fonction est k charactères.
La programme suivant donne le plus petit nombre à avoir une complexité de Kolmogorov supérieure à k+100.

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

Démonstration.
Raisonnement par l'absurde. Supposons qu'une telle fonction $K(X)$ existe.

```
fonction K(n) {...}
```

La taille de cette fonction est k charactères.
La programme suivant donne le plus petit nombre à avoir une complexité de Kolmogorov supérieure à k+100.

```
fonction K(n) {...}
n = 1
while ( K(n) < k + 100 ) {
  n = n + 1
}
print (n)
```

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

Démonstration.
Raisonnement par l'absurde. Supposons qu'une telle fonction $K(X)$ existe.

```
fonction K(n) {...}
```

La taille de cette fonction est k charactères.
La programme suivant donne le plus petit nombre à avoir une complexité de Kolmogorov supérieure à k+100.

```
fonction K(n) {...}
n = 1
while ( K(n) < k + 100 ) {
  n = n + 1
}
print (n)
```

Mais ! Sa taille est $k + 56$

La complexité de Kolmogorov n'est pas calculable.

Théorème.
Il est impossible d'écrire un programme qui calculera la valeur la complexité d'un nombre entier X.

Démonstration.
Raisonnement par l'absurde. Supposons qu'une telle fonction $K(X)$ existe.

```
fonction K(n) {...}
```

La taille de cette fonction est k charactères.
La programme suivant donne le plus petit nombre à avoir une complexité de Kolmogorov supérieure à k+100.

```
fonction K(n) {...}
n = 1
while ( K(n) < k + 100 ) {
  n = n + 1
}
print (n)
```

Mais ! Sa taille est $k + 56 < k + 100$
! ! !

Paradoxe de Berry

“Le plus petit entier non nommable en moins de vingt mots”

Paradoxe de Berry

“Le plus petit entier non nommable en moins de vingt mots”

Il y a ....

Paradoxe de Berry

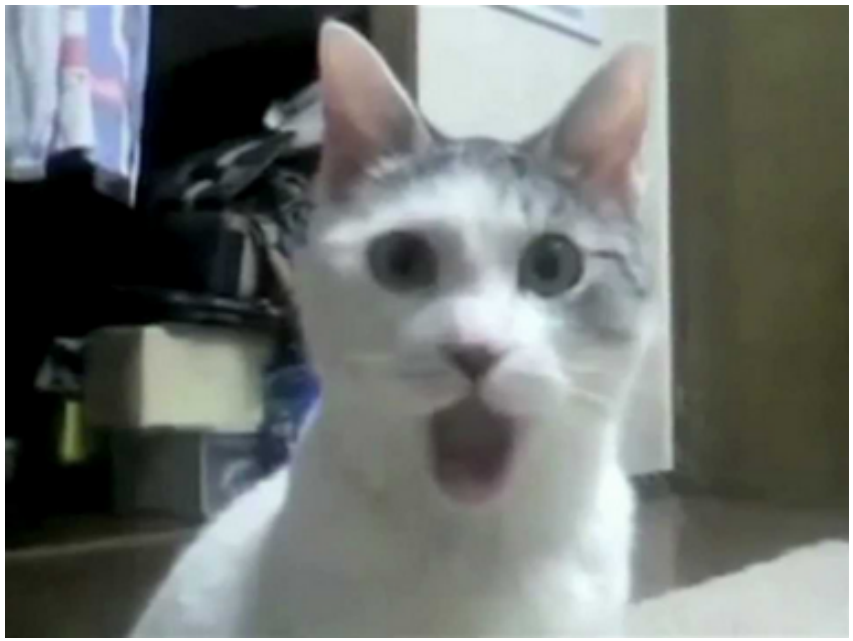
“Le plus petit entier non nommable en moins de vingt mots”

Il y a .... 11 mots !

Paradoxe de Berry

“Le plus petit entier non nommable en moins de vingt mots”

Il y a .... 11 mots !



Taille, complexité, richesse

En pratique on peut tenter juste de “zipper” les données pour réduire la taille tout en gardent leur complexité, leur richesse.

Taille, complexité, richesse

En pratique on peut tenter juste de “zipper” les données pour réduire la taille tout en gardent leur complexité, leur richesse.

C’est la compression sans pertes.

Concours de compression

- 500'000€ Prize for Compressing Human Knowledge.
  Compress the 1GB file enwik9 to less than the current record of about 114MB
  `http://prize.hutter1.net/`

- Large Text Compression Benchmark
  `http://www.mattmahoney.net/dc/text.html`
  Actuallement (Aug 16, 2023) c'est Fabrice Bellard avec son NNCP qui occupe la première place de cette compétition.
  `https://bellard.org/`

Fabrice Bellard

Un programmeur informatique français né en 1972 à Grenoble.
FFmpeg, QEMU, TCC, compression de texte, formules pour π.

Compression AVEC perte

Compression AVEC perte

Compression AVEC perte

Compression AVEC perte

Compression AVEC perte

C'est également utile pour bien anonymiser les données !

La complexité des algorithmes

La complexité des algorithmes

La notation grand O de Bachmann-Landau.

La complexité des algorithmes

La notation grand O de Bachmann-Landau.

On dit

$$f(t) \in O(g(t))$$

La complexité des algorithmes

La notation grand O de Bachmann-Landau.

On dit

$$f(t) \in O(g(t))$$

lorsqu'il existe des constantes $N > 0$ et $C > 0$ telles que pour tout $t > N$ on a

La complexité des algorithmes

La notation grand O de Bachmann-Landau.

On dit

$$f(t) \in O(g(t))$$

lorsqu'il existe des constantes $N > 0$ et $C > 0$ telles que pour tout $t > N$ on a $|f(t)| \leq C|g(t)|$.

Langages “Big Data”

- R
- Python
- C, Go, Java
- FPGA : VHDL, Verilog
- Cartes graphiques !

Chaque langage a une niche et sa propre culture !

Méthodes d'analyse

Méthodes d'analyse

1. Quoi faire ?
2. À partir de quoi ?
3. Comment faire ?

Méthodes d'analyse

1. Quoi faire ?
2. À partir de quoi ?
3. Comment faire ?

- Fonction de hachage
- PageRank de Google
- Map-Reduce
- Clustering / Partitionnement
- Résumé automatique...
- (Deep) Machine learning
- Autres méthodes statistiques

Architecture

Architecture

C'est trop, il faut partager !

Architecture

C'est trop, il faut partager !

- Stockage de données distribué : hdfs, blockchain, git, MySQL galera...

C'est trop, il faut partager !

- Stockage de données distribué : hdfs, blockchain, git, MySQL galera...
- Algorithmes distribués : map-reduce...

C'est trop, il faut partager !

- Stockage de données distribué : hdfs, blockchain, git, MySQL galera...
- Algorithmes distribués : map-reduce...
- Frameworks distribués : Hadoop, spark

Architecture

C'est trop, il faut partager !

- Stockage de données distribué : hdfs, blockchain, git, MySQL galera...
- Algorithmes distribués : map-reduce...
- Frameworks distribués : Hadoop, spark
- Architecture distribuée : Docker swarm, kubernetes, cloud, AWS, Microservices

Architecture

C'est trop, il faut partager !

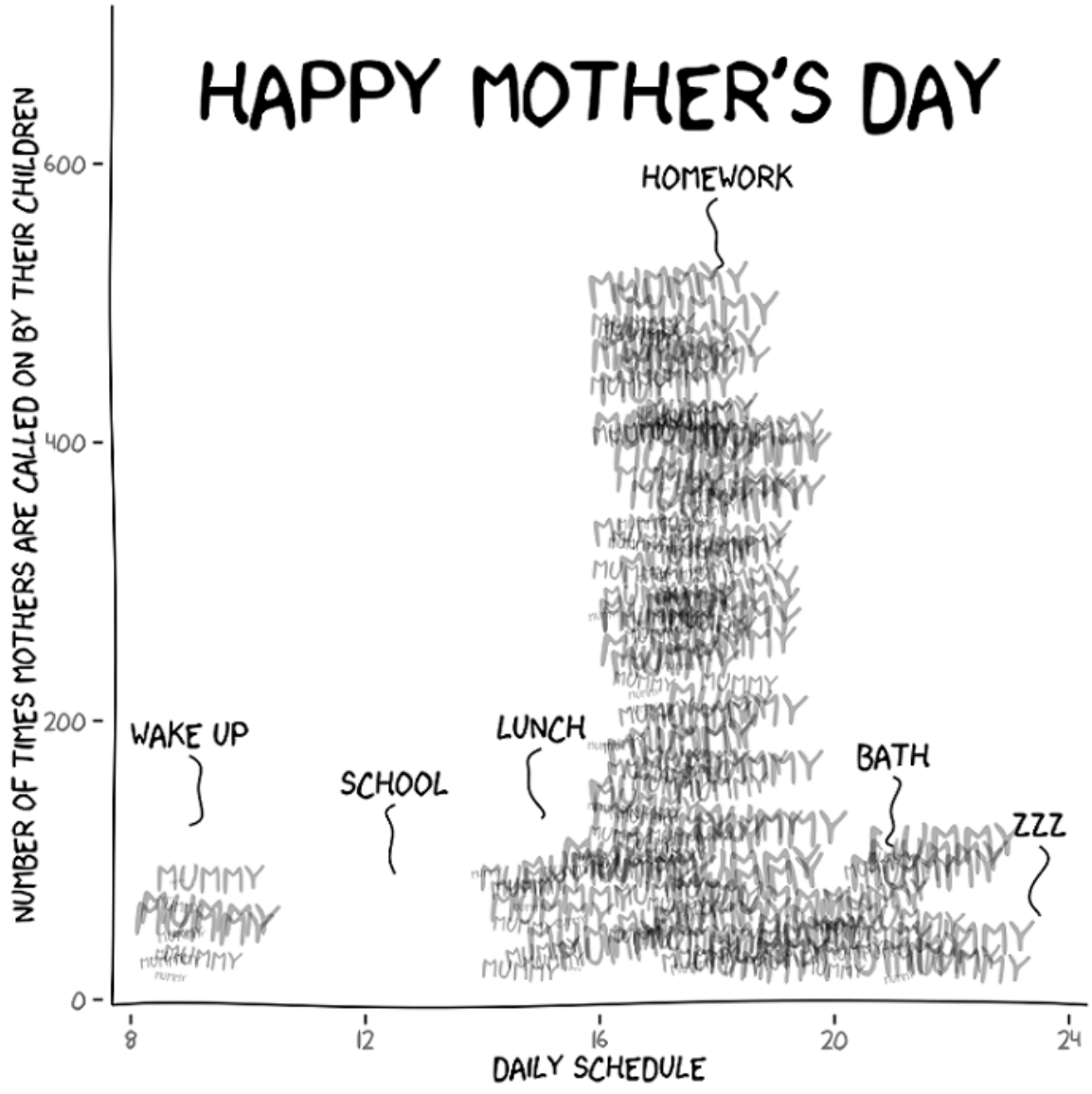
- Stockage de données distribué : hdfs, blockchain, git, MySQL galera...
- Algorithmes distribués : map-reduce...
- Frameworks distribués : Hadoop, spark
- Architecture distribuée : Docker swarm, kubernetes, cloud, AWS, Microservices
- Calcul parallèle : MPI et openMP

Architecture

C'est trop, il faut partager !

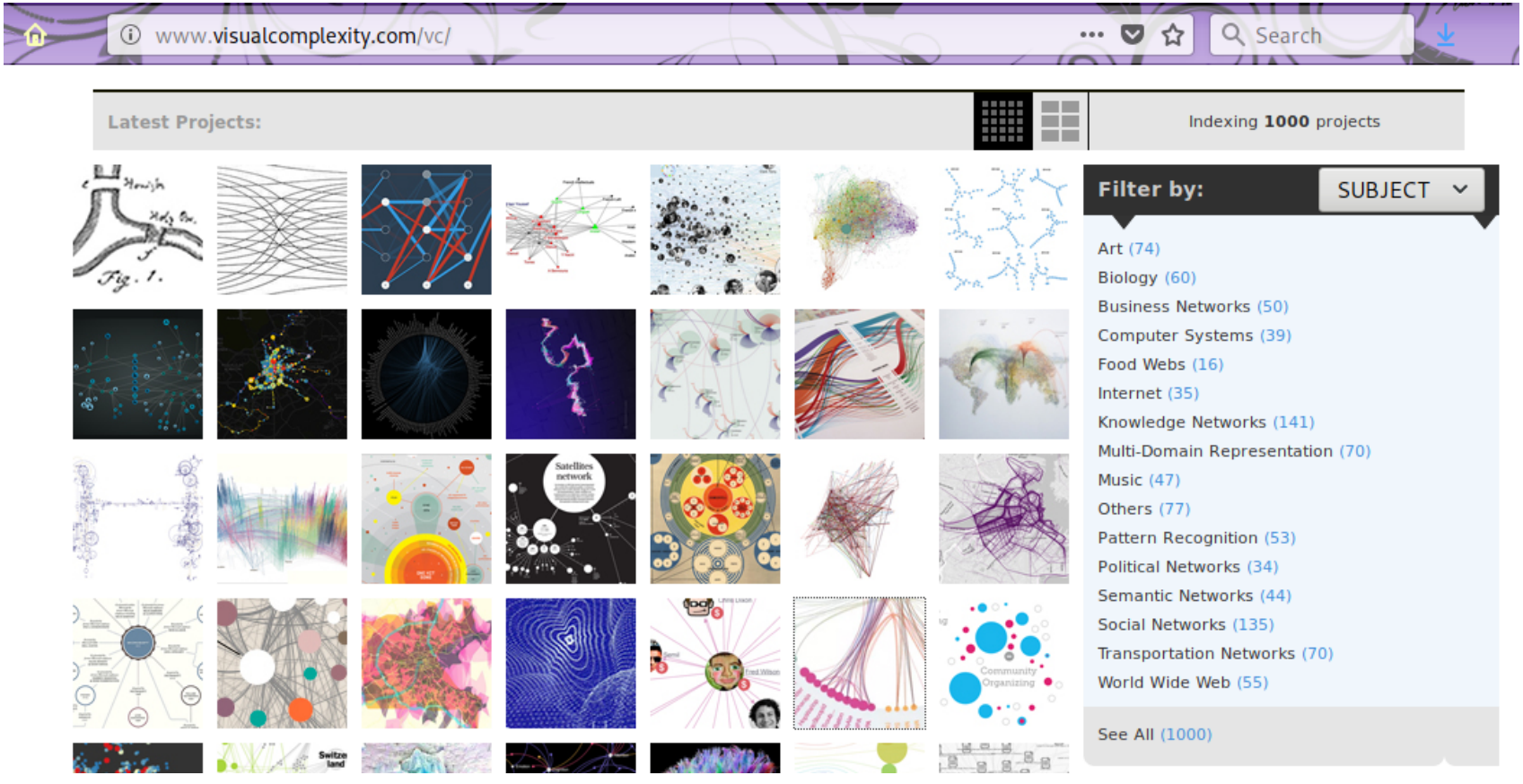
- Stockage de données distribué : hdfs, blockchain, git, MySQL galera...
- Algorithmes distribués : map-reduce...
- Frameworks distribués : Hadoop, spark
- Architecture distribuée : Docker swarm, kubernetes, cloud, AWS, Microservices
- Calcul parallèle : MPI et openMP
- GPU et FPGA

Visualisation



https://insidebigdata.com/2015/05/13/big-data-humor-a-mothers-day-plot/humor_mothersday/

Visualisation



http://www.visualcomplexity.com/vc/

Traitement éthique de données

Traitement éthique de données

Principes :

- principe de minimisation
  enlever les données qui ne sont pas utiles (contre logique big-data)

Traitement éthique de données

Principes :

- principe de minimisation
  enlever les données qui ne sont pas utiles (contre logique big-data)
- principe de finalité
  pas le droit de croiser des données de différents services
  collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités

Traitement éthique de données

Principes :

- principe de minimisation
  enlever les données qui ne sont pas utiles (contre logique big-data)
- principe de finalité
  pas le droit de croiser des données de différents services
  collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement de manière incompatible avec ces finalités

Contrainte :

- contrainte de compatibilité ne pas empêcher les progrès en France !

src : `http://perso.ens-lyon.fr/aurelien.garivier/www.math.univ-toulouse.fr/_agarivie/mydocs/IREM201701.pdf`

Explicabilité des décisions !

"En résumé, il semble bien qu'une décision administrative en France concernant une personne physique ne puisse pas se baser sur un algorithme d'apprentissage opaque."

1. `https://perso.math.univ-toulouse.fr/mllaw/home/statisticien/explicabilite-des-decisions-algorithmiques/`
2. `https://www.legifrance.gouv.fr/eli/decret/2017/3/14/PRMJ1632786D/jo/texte`

Quelques points

- Anonymisation des données

Quelques points

- Anonymisation des données
  Comment garantir la non-identifiabilité ?

Quelques points

- Anonymisation des données
  Comment garantir la non-identifiabilité ?
- Non-discrimination

Quelques points

- Anonymisation des données
  Comment garantir la non-identifiabilité ?
- Non-discrimination
- Distorsion de concurrence
  barrière à l'entrée du fait d'avoir les données

Quelques points

- Anonymisation des données
  Comment garantir la non-identifiabilité ?
- Non-discrimination
- Distorsion de concurrence
  barrière à l'entrée du fait d'avoir les données
- Ouverture et transparence versus protection du secret d'affaire

Questions ?