

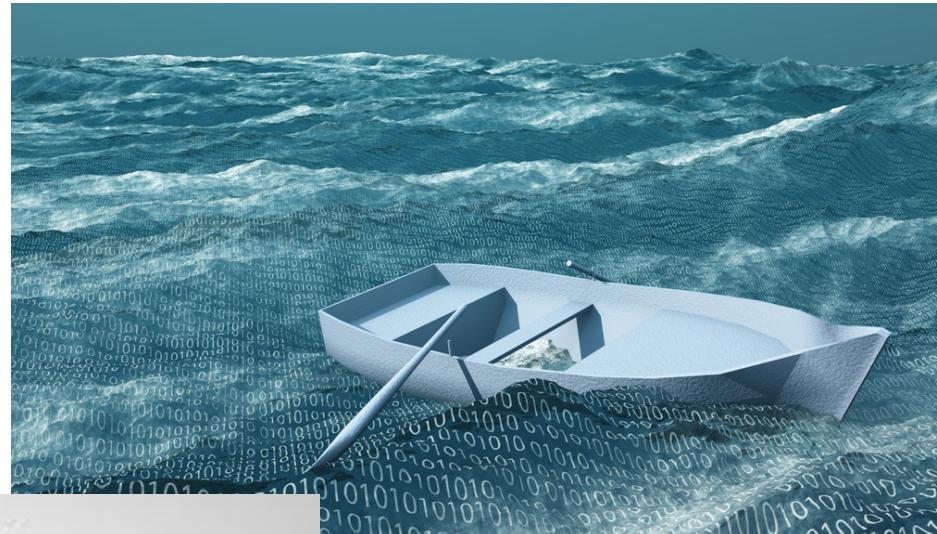
TECHNO DU BIG-DATA

Stockage, Exploitation et Valorisation des Données Massives

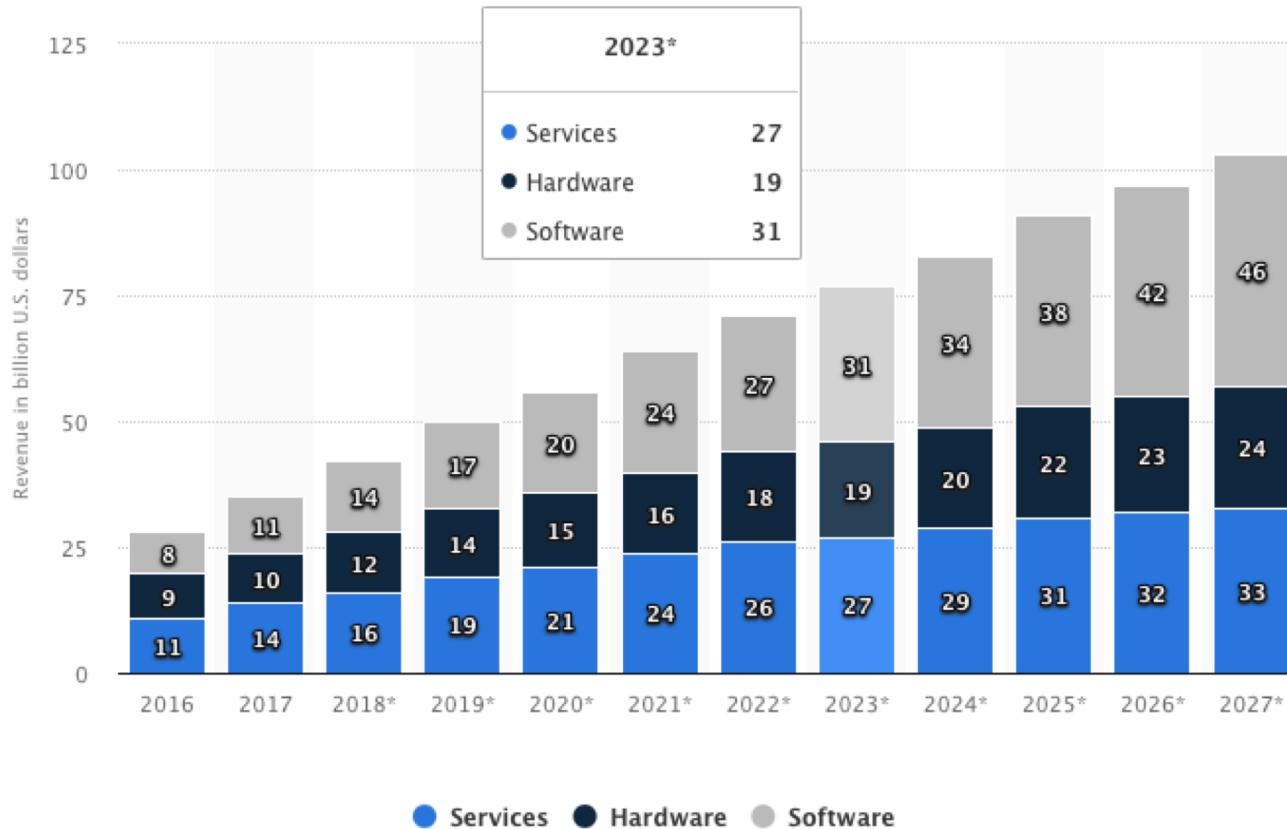
Stéphane Derrode, Dpt MI - stephane.derrode@ec-lyon.fr



Business Intelligence – « La Data »



Business Intelligence



<https://www.statista.com/statistics/301566/big-data-factory-revenue-by-type/>

Harvard Business Review :

<https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>

Risques



Intelligence artificielle : les géants du Web lancent un partenariat sur l'éthique

http://www.lemonde.fr/pixels/article/2016/09/28/intelligence-artificielle-les-geants-du-web-lancent-un-partenariat-sur-l-ethique_5005123_4408996.html

Déluge de données



Définition Big Data (Gardner, 2001)

Volume : Analyser de larges volumes de données par des techniques d'apprentissages (*machine learning*), de la fouille de données (*data mining*), de l'intelligence économique (*business intelligence*) et du calcul haute performance.

A titre d'exemple, *Twitter* généré en janvier 7 teraoctets de données chaque jour en janvier 2013, et *Facebook* 10!

Infographie sur les volumes : <http://future.arte.tv/fr/big-data-v-le-monde-ce-disque-dur/infographie-du-bit-au-yottaoctet>

Vélocité : La vélocité représente à la fois la fréquence à laquelle les données sont générées, capturées et partagées et mises à jour (bourse et *trading* haute fréquence et robots).

Variété : Il ne s'agit pas de données relationnelles traditionnelles, ces données sont brutes, semi-structurées voire non structurées. Ce sont des données complexes provenant du web (Web mining), au format texte (Text Mining) et images (Image Mining). Ce qui les rend difficilement utilisables avec les outils traditionnels.

Définition Big Data +

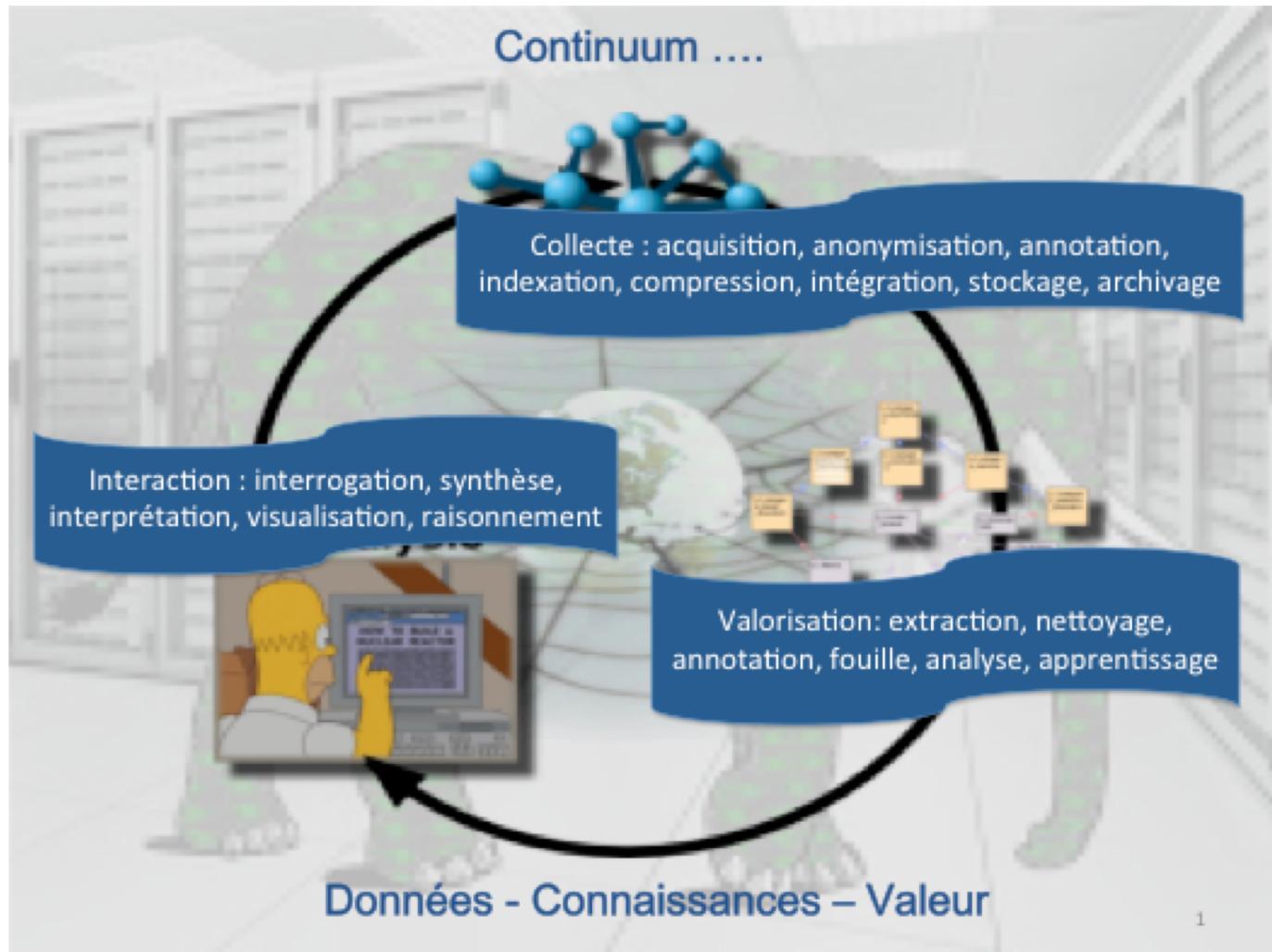
- + **Valeur** : Valeur ajoutée économique, sociétale, sanitaire, scientifique des données.
- + **Variabilité** : représente l'inconsistance des données dans le temps, ce qui entrave la manipulation et la gestion des données
- + **Véracité**: La qualité des données capturées, qui peuvent varier fortement (on parle de véracité des sources).

Les 10V

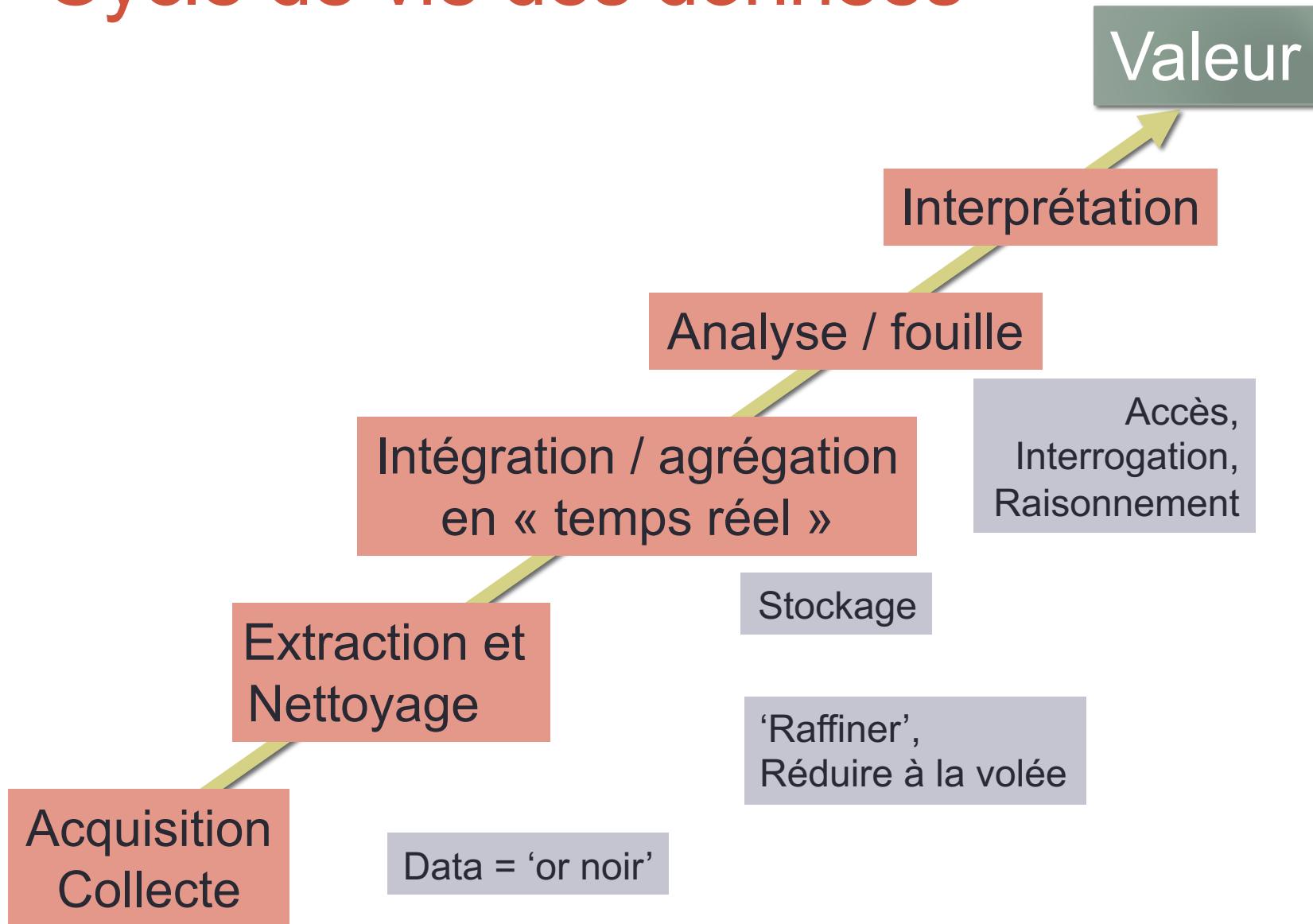
BIG | the Vs | 3v, 5v, 7v, 10v,

- **Volume** (length of a records, # of records) (entity-relationship databases)(datasets)
 - **Variety** (types: strings, pictures, voice, etc.) (structured, non-structured)
 - **Veracity** (**precision and accuracy of data**)
 - **Velocity** (of change)
 - **Value** (as a business/service)
 - **Volatility** (temporary; quick action)
 - **Vasting resources**
(storage, computation, transfer)
 - **Viability** (are data still useful?)
 - **Visibility** (open, hidden, ..)
 - **Validity**
(are there still valid/updated data?)
(in context validity)
(e-government datasets)
-
- The diagram illustrates the concept of Veracity. A vertical dotted line labeled "quality of data" is positioned to the right of the "Veracity" entry in the list. A pink dashed arrow points downwards from the "Veracity" entry towards this vertical line. To the left of the vertical line, four pink bullet points are listed: "- incomplete", "- redundant", "- inconsistent", and "- noisy". Below the vertical line, a blue dashed arrow points downwards, leading to a blue box containing the text: "filling missing values with estimated values calculated for complete records of the same dataset".

Cycle de vie des données



Cycle de vie des données



Stockage : Data-centers



société QTS à Atlanta

Le stockage de ces données massives est réalisé dans des entrepôts de données, ou data-centers, contenant des dizaines de pétaoctets (Po). Véritables usines confrontés à des problématiques industrielles (énergie, distribution, évacuation de la chaleur) et des contraintes de connexion, ils engloutissant 1% à 2% de la consommation électrique mondiale.

Traitements: Super-ordinateurs



Le superordinateur Blue Gene/P du Argonne National Lab utilise 250 000 processeurs en parallèle

MapReduce (Google, 2004) et son implémentation la plus connue, **Hadoop** (Yahoo, 2005) illustrent le paradigme actuel du calcul distribué : séparer le traitement de l'infrastructure logicielle permettant la répartition du calcul, la gestion des données distribuées, mais aussi l'hétérogénéité des ordinateurs et connections.

Libertés individuelles – « Big Brothers »

1. « **Big Data : les nouveaux devins** » (spécial investigation) 91,80% des données personnelles mondiales seraient détenues par 4 grands acteurs qui sont GAFA – 50 min.

<https://www.youtube.com/watch?v=5mmQeb8mXVk>

2. « **GAFA : une domination abusive** » – 15 min.

<http://www.rts.ch/emissions/geopolitis/6648267-google-et-les-autres-une-domination-abusive.html>

3. « **Données personnelles, protection des données** » : TV5monde – 15 min.

<http://enseigner.tv5monde.com/fle/big-data-que-fait-on-de-nos-donnees>

4. « **le Turbo Capitalisme, Nouveaux Loups de WallStreet** », 90 min. réalisé par Ivan Macaux avec Ali Baddou. CANAL+, 2015.

<https://www.youtube.com/watch?v=vOzH7Aj2MOA>

On parle maintenant de GAFAM = GAFA + Microsoft,

NATU : Netflix, Airbnb, Tesla, Uber. <http://tempsreel.nouvelobs.com/rue89/20150802.RUE3739/apres-les-gafa-les-nouveaux-maitres-du-monde-sont-les-natu.html>

Reportages ARTE Future

ARTE futur : dossier <http://future.arte.tv/fr/sujet/bigdata> (adresse obsolète)

1. Kenny Polcari : « J'ai peur » - 7 min.
2. Islande, le pays du stockage numérique écologique – 8 min.
3. La course à l'exploitation des données client – 6 min.
4. Le corps à la mesure de ses capacités (*quantified-self*) – 10 min.
5. Le nouveau système d'alerte pour les prématurés – 7 min.
6. La maîtrise des flux de données – 5 min.
7. Big Data et la protection des océans – 5 min.
8. L'utilisation du big data au CERN – 12 min.

<https://www.arte.tv/fr/videos/068833-002-A/l-utilisation-du-big-data-au-cern/>

Synthèse bibliographique

- Groupe de 5 élèves.
- Rapport de 5-10 pages (avec. Réf. Bib.). Pas de restitution orale.
- L'originalité de la présentation (forme / fond) sera prise en compte dans la note.
- Liste de sujets ci-après. Merci de me soumettre votre sujet originaux pour approbation

Synthèse bibliographique

Exemple de sujets

- Big Data et journalisme (data journalisme)
- Big data et jeux
- Big data et objets connectés (Internet des objets)
- Big Data et GAFAM
- Big-Data et start-up
- Open entreprise (cf C-Radar)
- La France et le Big-data
- Big Data et projet scientifique (astronomie, génome...)
- Le rôle du *data-scientist* dans l'entreprise
- Les *Data-Centers* dans le monde
- Enjeux de l'analyse prédictive (ex. Kxen)
- Big Data et libertés individuelles
- Big Data et plateforme de crowdfunding
- Big data et médecine personnalisée
- Big Data et Lean
- Big Data et politique
- Big Data et écologie (exemple d'un grand programme basé données)
- ...