# MOD 2.1 - « DÉFIS INFORMATIQUE DU BIG-DATA »

## Stockage, Exploitation et Valorisation des Données Massives

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

# Organisation du MOD

- **Organisation**
  - 3 TPs de 4 heures (des mercredis après-midi entre octobre et début décembre). CR à prévoir !
  - 8 interventions de 2h (lundis matin à 8h)
  - 1 fiche de synthèse à rédiger (description ci-après)
  - 1 examen de 1h30 programmé le 17 décembre à 10h.
- **Informations**
  - Page principale (slides, énoncés TP de S. Derrode, info pratiques) : http://perso.ec-lyon.fr/derrode.stephane/Teaching/ECL2A3A/ECL3A_MOD21/
  - Pedagogie3 :
    - Dépôt de vos synthèses bibliographiques
    - Dépôt de vos CR de TP
    - Enoncé TP de R. Chalon
- **Vous êtes >100 étudiants**
  - Présence obligatoire à tous les cours (présence systématique)
  - Les absences non justifiées seront sévèrement sanctionnées.
  - Des évaluations sensiblement plus difficiles que l'année dernière.

# BIG DATA

Définition, enjeux, illustrations

# Les données sont partout


The Economist

**Identifiez 3 appareils qui collectent des données**

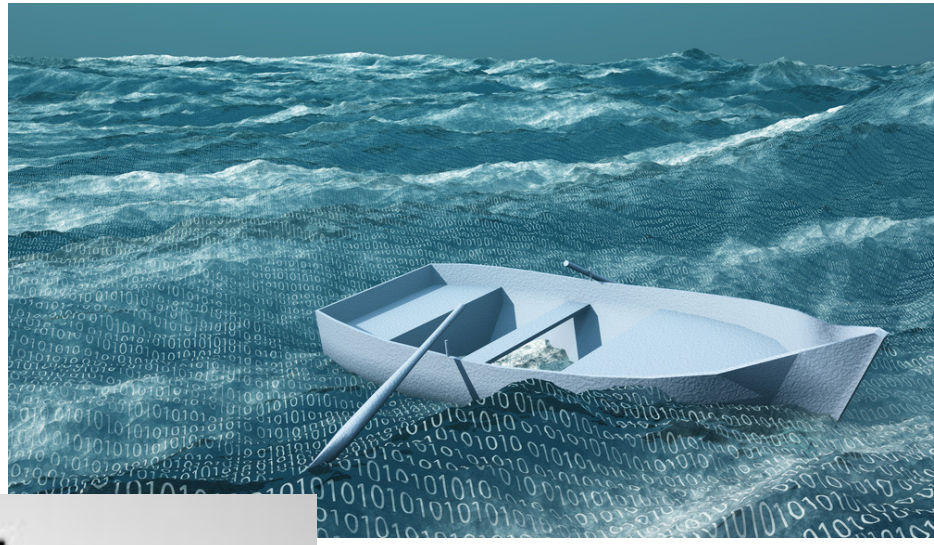
# Les données sont partout

**La pyramide Data-Information-Knowledge-Wisdom (ou compétences)**


Attribuée à Russell Ackoff en 1989, elle signifie que :

•Les **données** sont la matière "brute" de l'information conçues plutôt pour des machines.
•**L'information** pourrait être définie comme 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**.

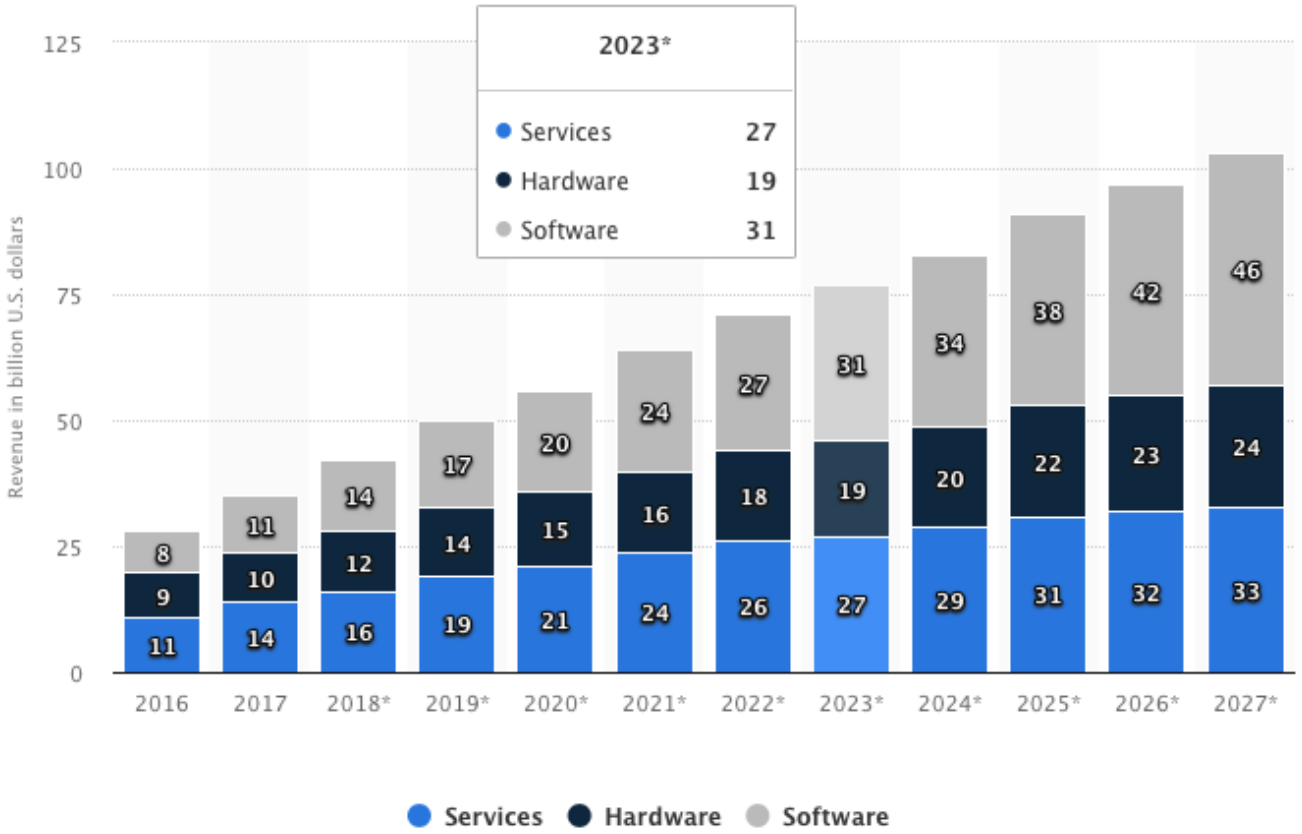
# Business Intelligence – « La Data »


Les "**données massives**"

# Marchés du big data

sources




**Harvard Businesss Review :**

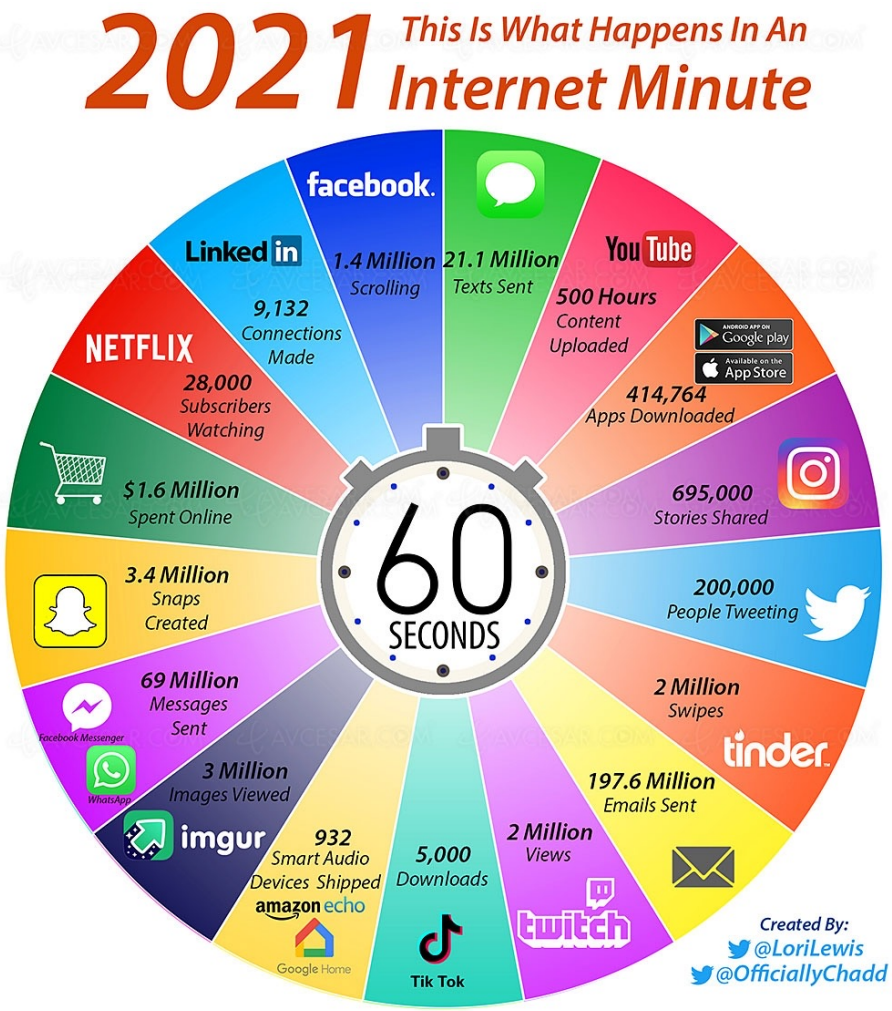
Data Scientist: The Sexiest Job of the 21st Century

# Risques


https://www.greenberg-art.com/.Toons/.Toons,%20social/Medprivacy.htm, 1999l

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

# Déluge de données




"Avec suffisamment de données, les chiffres parlent d'eux-mêmes."
Chris Anderson, journaliste *Wired Magazine*

Le déluge des données rend la méthode scientifique obsolète, l'analyse des motifs et des relations contenues dans les données massives produit intrinsèquement un savoir significatif et éclairé sur des phénomènes complexes. Il y a maintenant une meilleure manière de faire. Les petabytes nous permettent de dire que « la corrélation suffit ». Nous pouvons analyser les données sans hypothèses sur ce qu'elles peuvent montrer.
Anderson, C. (2008) "The end of theory: The data deluge makes the scientific method obsolete", *Wired*

# Déluge de données

**Big data** : “une nouvelle génération de technologies et architectures conçues pour extraire de la valeur économique ou scientifique à une grande variété de données en permettant leur capture à haute vitesse, leur découverte et/ou analyse”.

# 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 par Arte.

**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.

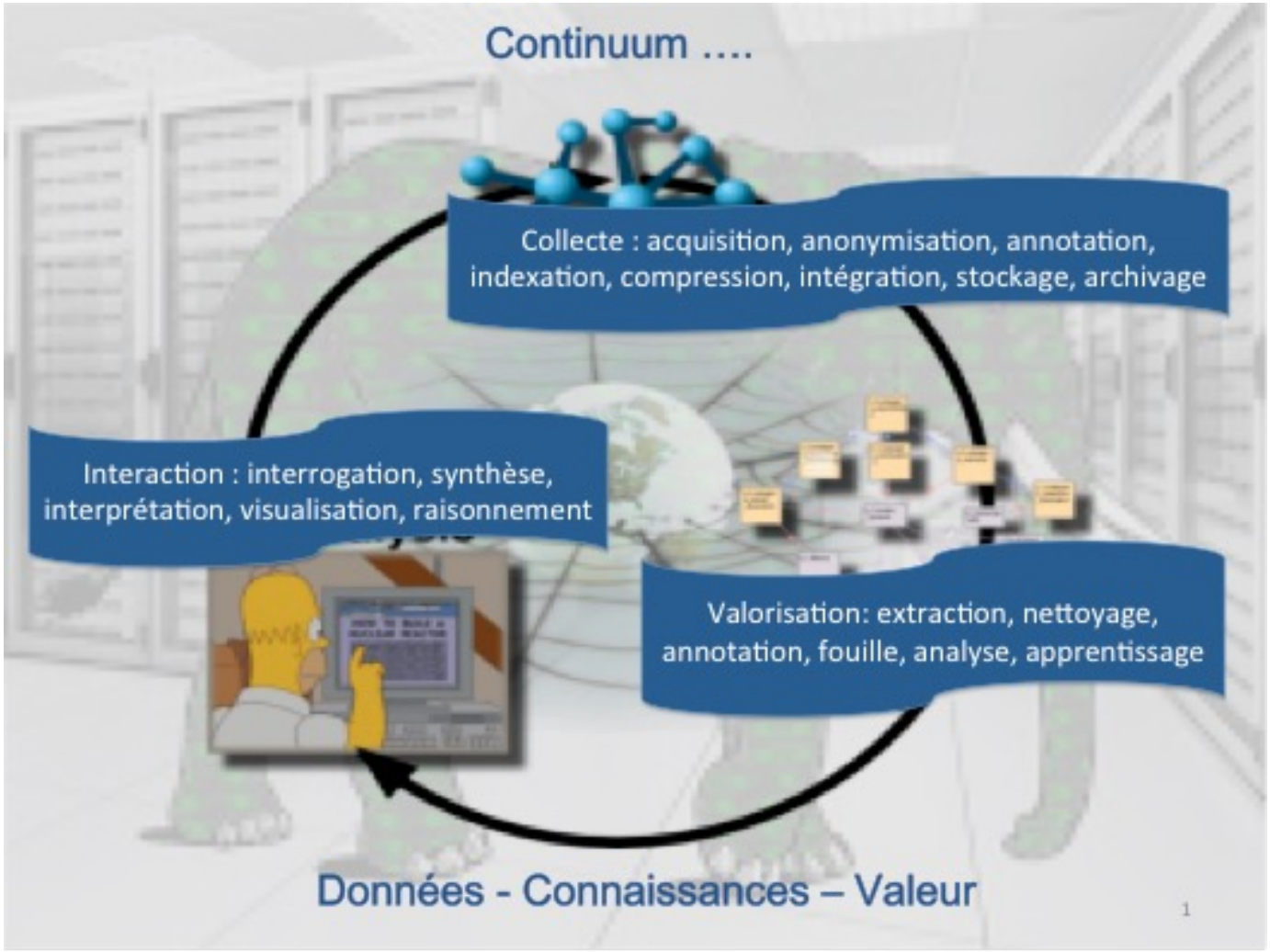
# Définition Big Data +

+ **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

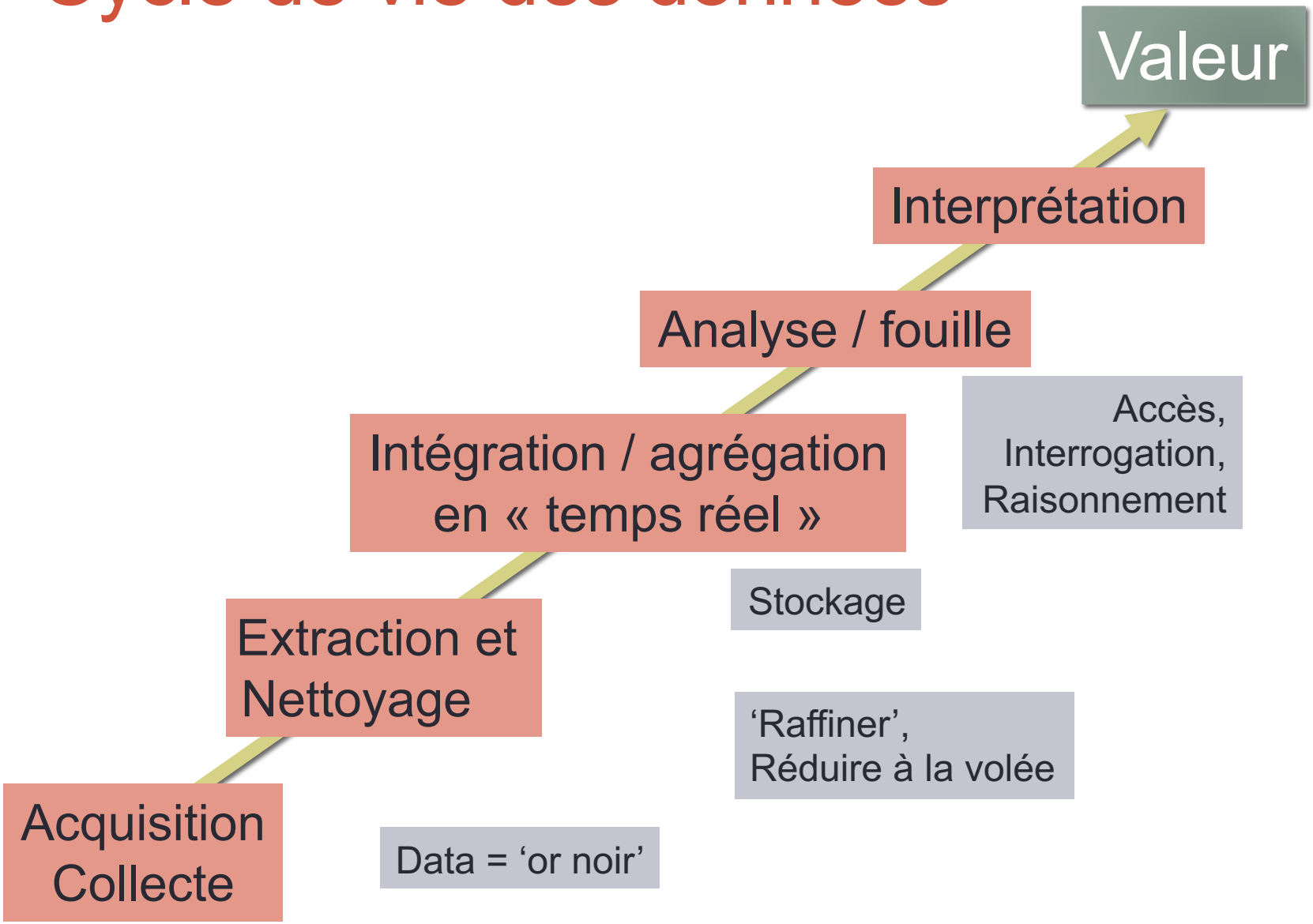
+ **Véracité:** La qualité des données capturées, qui peuvent varier fortement (on parle de véracité des sources).

# Cycle de vie des données




Source : site du GDR MADICS

# 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 Argone 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 2015) – 50 min.
   91,80% des données personnelles mondiales seraient détenues par 4 grands acteurs qui sont GAFA

2. **« GAFA : une domination abusive »** (RTS, Geopolitis) – 15 min.
   Regroupées sous l'acronyme GAFA (Google, Apple, Facebook, Amazon), ces entreprises totalisent plus de 300 milliards de dollars de chiffre d'affaires en 2014, soit approximativement le PIB de la Norvège.

3. **« Big Data : que fait-on de nos données ? »** (RTS/TV5monde) – 15 min.
   Toutes ces données sont-elles vraiment utilisées de manière bienveillante ? L'éthique ne doit-elle pas être remise au cœur de l'immense champ des possibles offert par les Big Data ?

4. **« Le Turbo Capitalisme, Nouveaux Loups de WallStreet (**CANAL+, 2015) – 1h30.
   Les Nouveaux Loups de Wall Street - la Bourse est truquée, le Pigeon, c'est vous !

- GAFAM = GAFA + Microsoft,
- Après les Gafa, les nouveaux maîtres du monde sont les Natu (Netflix, Airbnb, Tesla, Uber). (Le Nouvel Obs 2017)

# Reportages ARTE Future

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 maitrise des flux de données – 5 min.
7. Big Data et la protection des océans – 5 min.
8. L'utilisation du big data au CERN (Arte.tv 2017) – 12 min.

# RGPD ET AI ACT

Protection des données personnelles au niveau Européen. Réglementation des IA (en particulier génératives)

# Protection des données personnelles

**Loi de 1978 dite « Informatique et libertés ».** Le président Valéry Giscard d'Estaing décide la création d'un organisme de contrôle des données personnelles dans la société de l'information : **la CNIL**, et avec elle est promulguée une loi majeure, la **loi de 1978 dite « Informatique et libertés ».**

Les droits

- le droit d'information : chacun peut être informé des traitements dont ses données font l'objet, cet article est applicable en toutes circonstances, même aux cas relevant de la sécurité nationale ;
- le droit d'accès (et d'effacement, droit à l'oubli) : plus complet que le droit d'information, il permet à chacun d'accéder aux informations qui sont conservées sur lui, il est toutefois interdit d'en faire usage dans certains cas ;
- le droit de rectification : chacun peut demander à faire corriger les données stockées le concernant;
- le droit d'opposition : chacun peut s'opposer à faire l'objet d'un traitement, pour un motif légitime (le démarchage commercial est reconnu par la loi comme motif légitime).

**Ces droits se retrouveront, amplifiés et adjoints à d'autres, dans le RGPD.**

**Modification de 2004 :** La loi de 2004 met aussi en place un nouveau système relatif à la déclaration des fichiers et traitements ; toute structure souhaitant effectuer un traitement de données à caractère personnel ou stocker ces données doit effectuer une déclaration à la CNIL.

Source: Zeste de savoir

## RGPD (Loi Européenne, 2015) : Règlement Général de la Protection des Données

Concrètement, toute donnée concernant *un citoyen* européen, même traitée hors union, est dans le champ d'application du règlement ; c'est un cadre très large et protecteur pour les Européens.

**"Le présent règlement protège les libertés et droits fondamentaux des personnes physiques, et en particulier leur droit à la protection des données à caractère personnel."**

Pour les structures (y compris les administrations) traitant de la donnée personnelle, le principe de déclaration obligatoire à la CNIL est transformé en principe de responsabilité, permettant bien plus de souplesse, mais augmentant les sanctions (amendes dissuasives).

En ce qui concerne les utilisateurs, leurs droits sont renforcés, avec notamment l'arrivée du consentement « explicite » : le traitement des données personnelles est soumis à un consentement qui ne peut être forcé ; particulièrement, l'accès au service ne peut être conditionné à acceptation de traitement de données qui n'y seraient pas directement nécessaires.

Aux 4 droits de la lois Informatique et Liberté française, s'ajoute:

- le droit à la limitation du traitement (effacement partiel);
- le droit à la portabilité : permet à chaque citoyen de demander l'intégralité des données à caractère personnel les concernant.

**2 MIN POUR COMPRENDRE LE RGPD : GALÈRE OU OPPORTUNITÉ ?**

Source: Zeste de savoir

**AI Act (Européen) :** première réglementation (mondiale) de l'intelligence artificielle (texte final : fin 2023)

En avril 2021, la Commission européenne a proposé le premier cadre réglementaire de l'UE pour l'IA. Il propose que des systèmes d'IA qui peuvent être utilisés dans différentes applications soient **analysés et classés en fonction du risque qu'ils présentent pour les utilisateurs**. Les différents niveaux de risque impliqueront plus ou moins de réglementation. Une fois approuvées, ces règles seront les premières au monde sur l'IA.

La priorité du Parlement est de veiller à ce que les systèmes d'IA utilisés dans l'UE soient sûrs, transparents, traçables, non discriminatoires et respectueux de l'environnement.

**Risque inacceptable :** Les systèmes d'IA à risque inacceptable sont des systèmes considérés comme une menace pour les personnes et seront interdits. Ils comprennent:

- la manipulation cognitivo-comportementale de personnes ou de groupes vulnérables spécifiques : par exemple, des jouets activés par la voix qui encouragent les comportements dangereux chez les enfants
- un score social : classer les personnes en fonction de leur comportement, de leur statut socio-économique, de leurs caractéristiques personnelles
- des systèmes d'identification biométrique en temps réel et à distance, tels que la reconnaissance faciale

Certaines exceptions peuvent être autorisées : par exemple, les systèmes d'identification biométrique à distance "a posteriori", où l'identification se produit après un délai important, seront autorisés à poursuivre des crimes graves et seulement après l'approbation du tribunal

Source: Actualités du parlement Européen

## **AI Act (Européen) :** première réglementation (mondiale) de l'intelligence artificielle

**Risque élevé :** Les systèmes d'IA qui ont un impact négatif sur la sécurité ou les droits fondamentaux seront considérés comme à haut risque et seront divisés en deux catégories.

1. **Les systèmes d'IA** qui sont utilisés dans les produits relevant de la législation de l'UE sur la sécurité des produits. Cela comprend les jouets, l'aviation, les voitures, les dispositifs médicaux et les ascenseurs.
2. **Les systèmes d'IA** relevant de huit domaines spécifiques qui devront être enregistrés dans une base de données de l'UE :
   - l'identification biométrique et la catégorisation des personnes physiques
   - la gestion et l'exploitation des infrastructures critiques
   - l'éducation et la formation professionnelle
   - l'emploi, la gestion des travailleurs et l'accès au travail indépendant
   - l'accès et la jouissance des services privés essentiels et des services et avantages publics
   - les forces de l'ordre
   - la gestion de la migration, de l'asile et du contrôle des frontières
   - l'aide à l'interprétation juridique et à l'application de la loi.

Tous les systèmes d'IA à haut risque seront évalués avant leur mise sur le marché et au long de leur cycle de vie.

**L'IA générative**, comme ChatGPT, devrait se conformer aux exigences de transparence :

- indiquer que le contenu a été généré par l'IA
- concevoir le modèle pour l'empêcher de générer du contenu illégal
- publier des résumés des données protégées par le droit d'auteur utilisées pour la formation

**AI Act (Européen) :** première réglementation (mondiale) de l'intelligence artificielle (texte final : fin 2023)

**Risque limité :**

Les systèmes d'IA à risque limité doivent respecter des exigences de transparence minimales qui permettraient aux utilisateurs de prendre des décisions éclairées. Après avoir interagi avec les applications, l'utilisateur peut alors décider s'il souhaite continuer à l'utiliser. Les utilisateurs doivent être informés lorsqu'ils interagissent avec l'IA.

Cela inclut les systèmes d'IA qui génèrent ou manipulent du contenu image, audio ou vidéo (par exemple, les deepfakes, des contenus faux qui sont rendus crédibles par l'IA).

Source: Actualités du parlement Européen

# Synthèse bibliographique

- Groupe de 4-6 élèves.
- Rapport de 10 p. (avec. Réf. Bib.). Pas de restitution orale.
- L'originalité de la présentation (forme / fond) sera prise en compte dans la note.
- Date boutoir : le vendredi 20 décembre 2024 (23h55)

- Sujet et groupe pour le 7 octobre 2024 (soir), en remplissant le fichier partagé (l'adresse du fichier est sur la page du cours).

- Liste de sujets possibles ci-après.

Merci de me soumettre votre sujet originaux pour approbation préalable.

# 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 (ou en Asie, en Belgique…)
- 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)
- …