MOD 2.1 - « DÉFIS INFORMATIQUE DU BIG-DATA »

Stockage, Exploitation et Valorisation des Données Massives

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

Les données sont partout

The Economist

Identifiez 3 appareils qui collectent des données

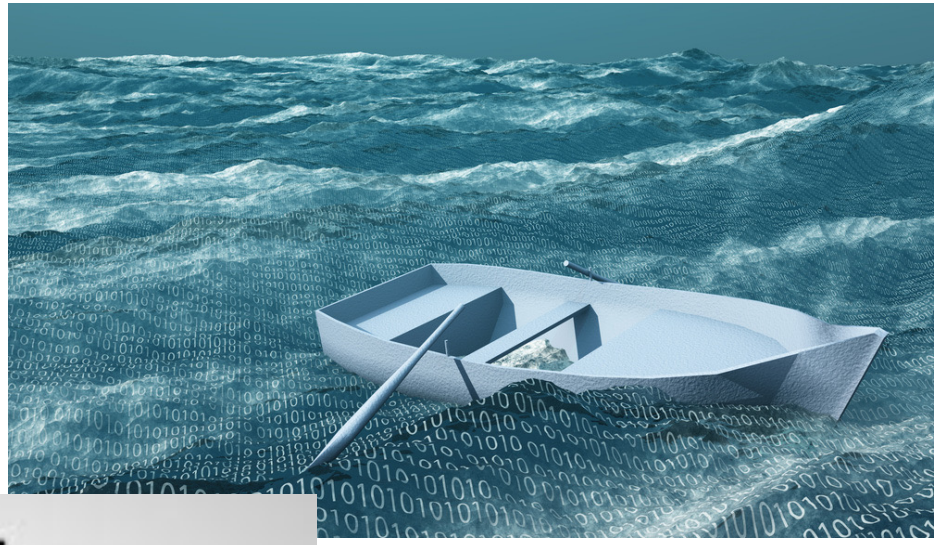
Les données sont partout

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**.

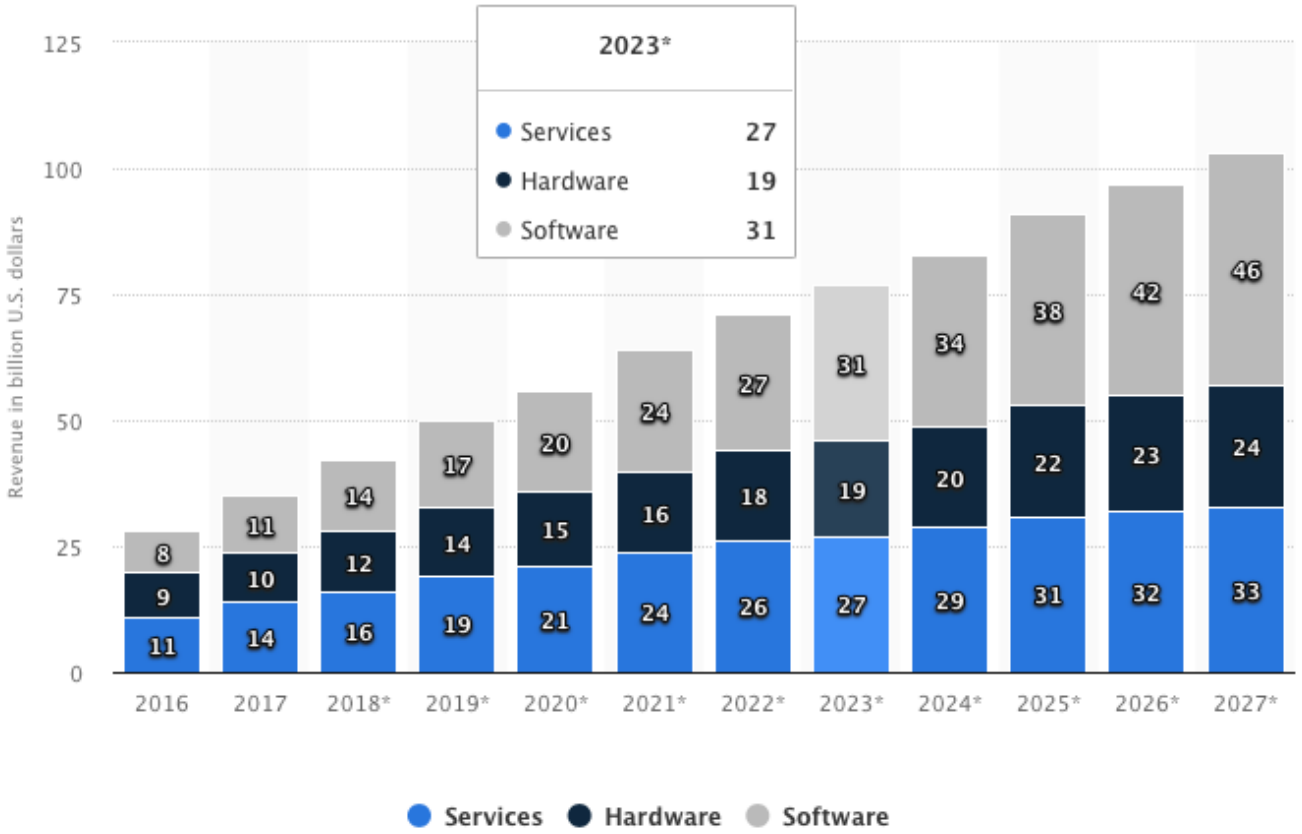
Business Intelligence – « La Data »



Les "**données massives**"

Marchés du big data

sources



Harvard Businesss Review :

Data Scientist: The Sexiest Job of the 21st Century

Risques

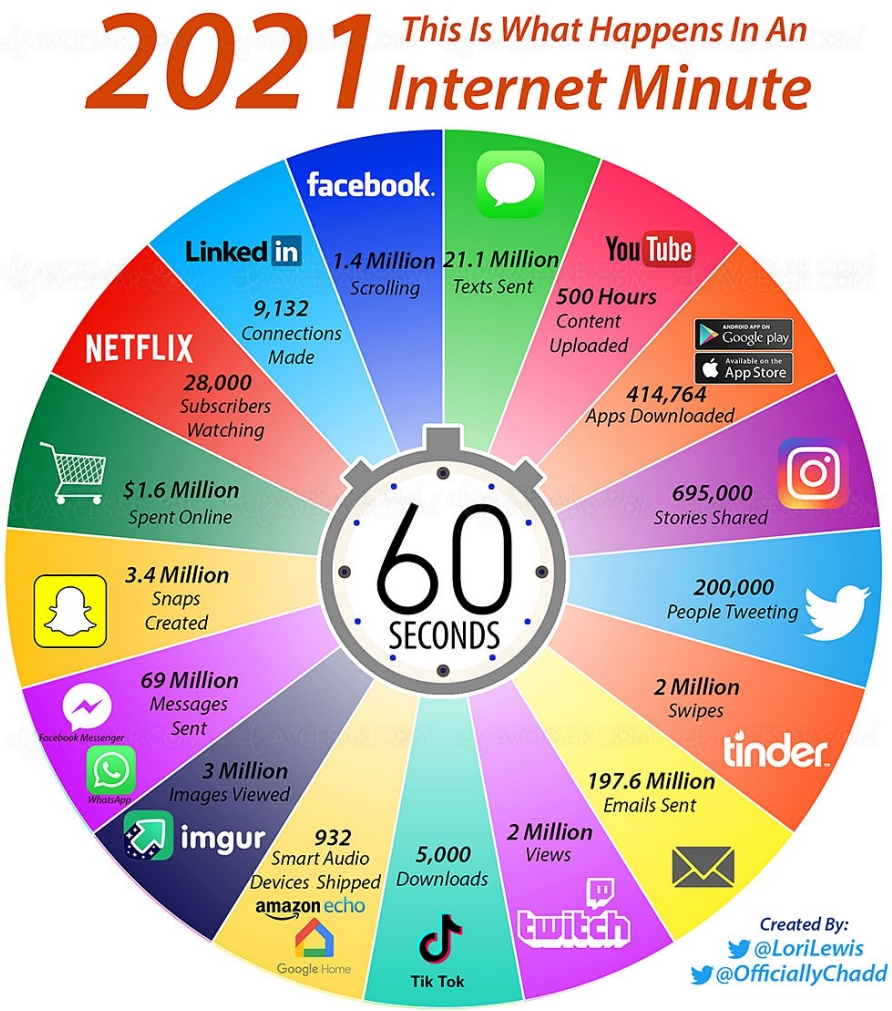
https://www.greenberg-art.com/.Toons/.Toons,%20social/Medprivacy.htm, 1999l

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

Risques

https://www.cartoonstock.com/directory/m/medical_record.asp

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.

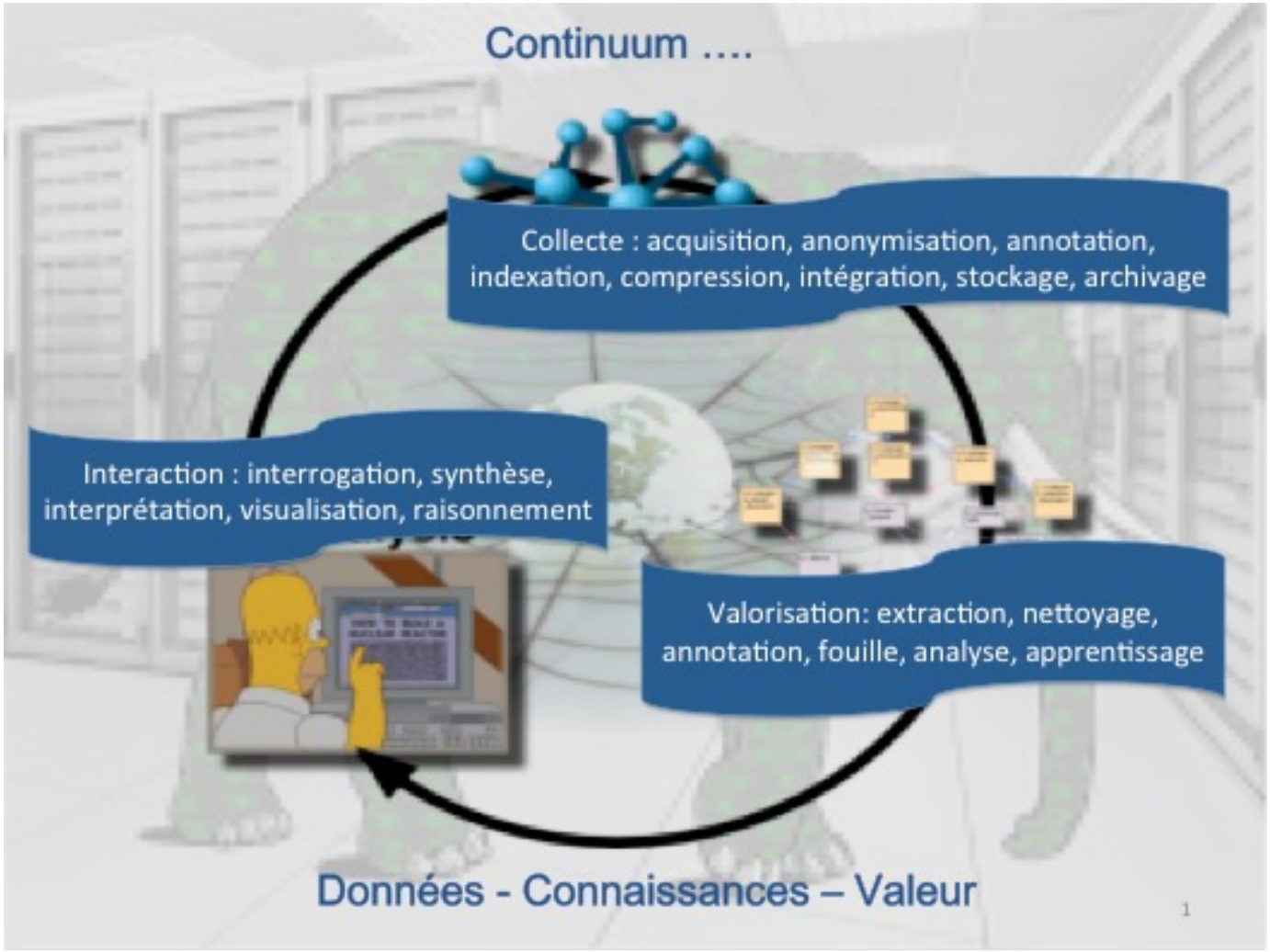
Définition Big Data +

+ **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

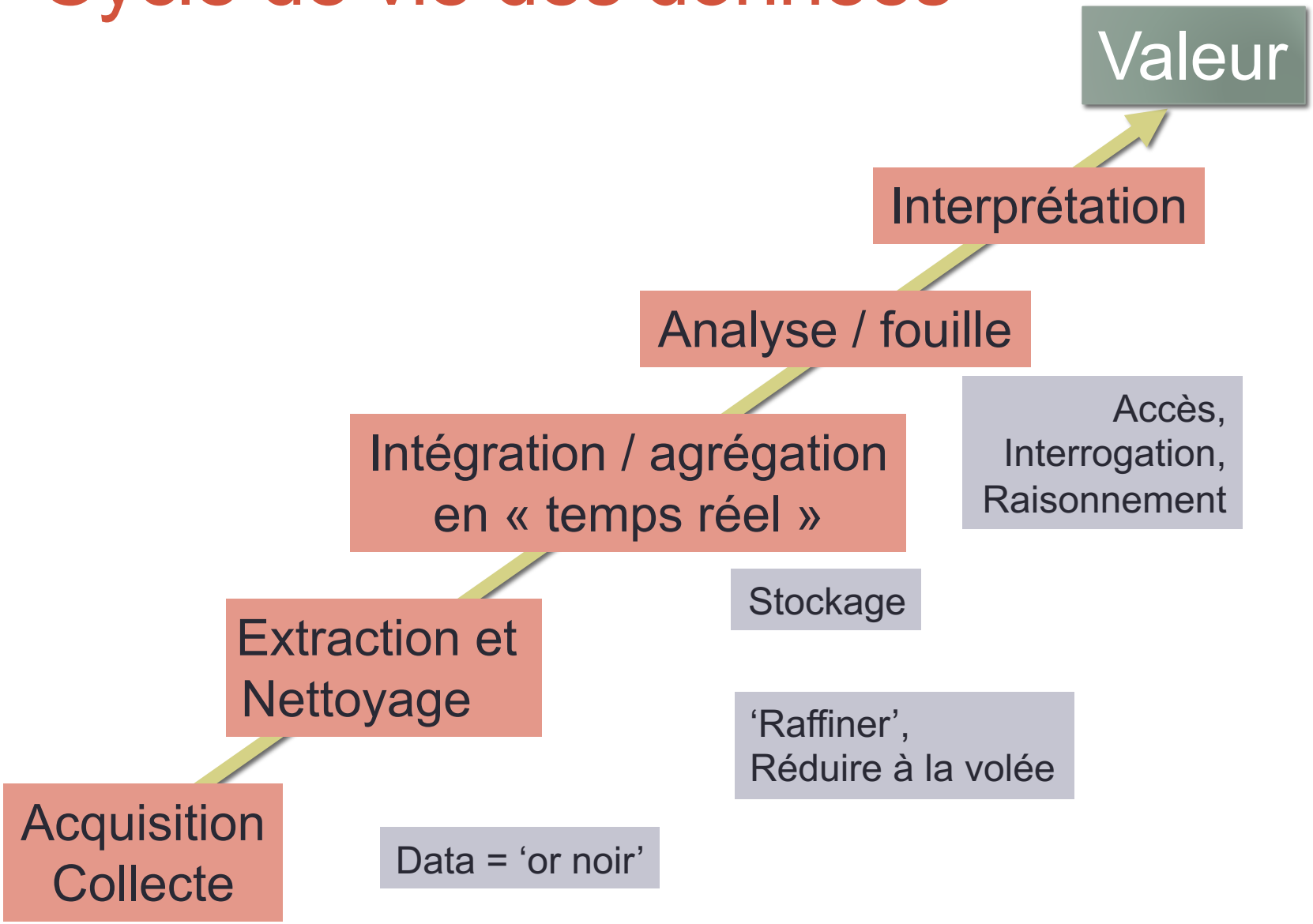
+ **Véracité:** La qualité des données capturées, qui peuvent varier fortement (on parle de véracité des sources).

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.

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 fin décembre
- Informations
  - Page principale (poly, énoncés TP, …) :
    http://perso.ec-lyon.fr/derrode.stephane/Teaching/ECL2A3A/ECL3A_MOD21/
  - Pedagogie3 (Dépôt de vos synthèses bibliographiques):
    https://pedagogie3.ec-lyon.fr/course/view.php?id=2304

Organisation du MOD

Num sÈa	Groupe(s	Intervenant	Jour	Date	Heure
1	MOD 2.1	Stéphane DERRODE	lun	03/10/2022	08:00-10:00
BE #1	MOD 2.1_3	Stéphane DERRODE	mer	05/10/2022	14:00-16:00
BE #1	MOD 2.1_3	Stéphane DERRODE	mer	05/10/2022	16:15-18:15
2	MOD 2.1	Stéphane DERRODE	lun	10/10/2022	08:00-10:00
BE #1	MOD 2.1_1	Stéphane DERRODE	mer	12/10/2022	14:00-16:00
BE #1	MOD 2.1_1	Stéphane DERRODE	mer	12/10/2022	16:15-18:15
3	MOD 2.1	Aurélien FARAVELON	lun	17/10/2022	08:00-10:00
BE #1	MOD 2.1_2	Stéphane DERRODE	mer	19/10/2022	14:00-16:00
BE #1	MOD 2.1_2	Stéphane DERRODE	mer	19/10/2022	16:15-18:15
BE #2	MOD 2.1_1	René CHALON	mer	19/10/2022	14:00-16:00
BE #2	MOD 2.1_1	René CHALON	mer	19/10/2022	16:15-18:15
4	MOD 2.1	Alessandro CERIONI	lun	24/10/2022	08:00-10:00
BE #3	MOD 2.1_3	Stéphane DERRODE	mer	26/10/2022	14:00-16:00
BE #3	MOD 2.1_3	Stéphane DERRODE	mer	26/10/2022	16:15-18:15
5	MOD 2.1	Guillaume LAVOUE	lun	07/11/2022	08:00-10:00
6	MOD 2.1	Xavier CALLENS	lun	14/11/2022	08:00-10:00
BE #2	MOD 2.1_2	René CHALON	mer	16/11/2022	14:00-16:00
BE #2	MOD 2.1_2	René CHALON	mer	16/11/2022	16:15-18:15
7	MOD 2.1	Aurélien FARAVELON	lun	21/11/2022	08:00-10:00
8	MOD 2.1	Emmanuel DELLANDREA	lun	28/11/2022	08:00-10:00
BE #3	MOD 2.1_1	Stéphane DERRODE	mer	30/11/2022	14:00-16:00
BE #3	MOD 2.1_1	Stéphane DERRODE	mer	30/11/2022	16:15-18:15
BE #2	MOD 2.1_3	René CHALON	mer	30/11/2022	14:00-16:00
BE #2	MOD 2.1_3	René CHALON	mer	30/11/2022	16:15-18:15
BE #3	MOD 2.1_2	Stéphane DERRODE	mer	07/12/2022	14:00-16:00
BE #3	MOD 2.1_2	Stéphane DERRODE	mer	07/12/2022	16:15-18:15

Synthèse bibliographique

- Groupe de 4-6 élèves.

  Merci de m'envoyer vos groupes constitués par mail (stephane.derrode@ec-lyon.fr)
- 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.
- Poids : 33% (contre 67% pour l'exam)
- Date boutoir : le vendredi 16 décembre (23h55)

- Sujet et groupe par mail pour le 11 octobre.
- Liste de sujets possible ci-dessous.

  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)
- …