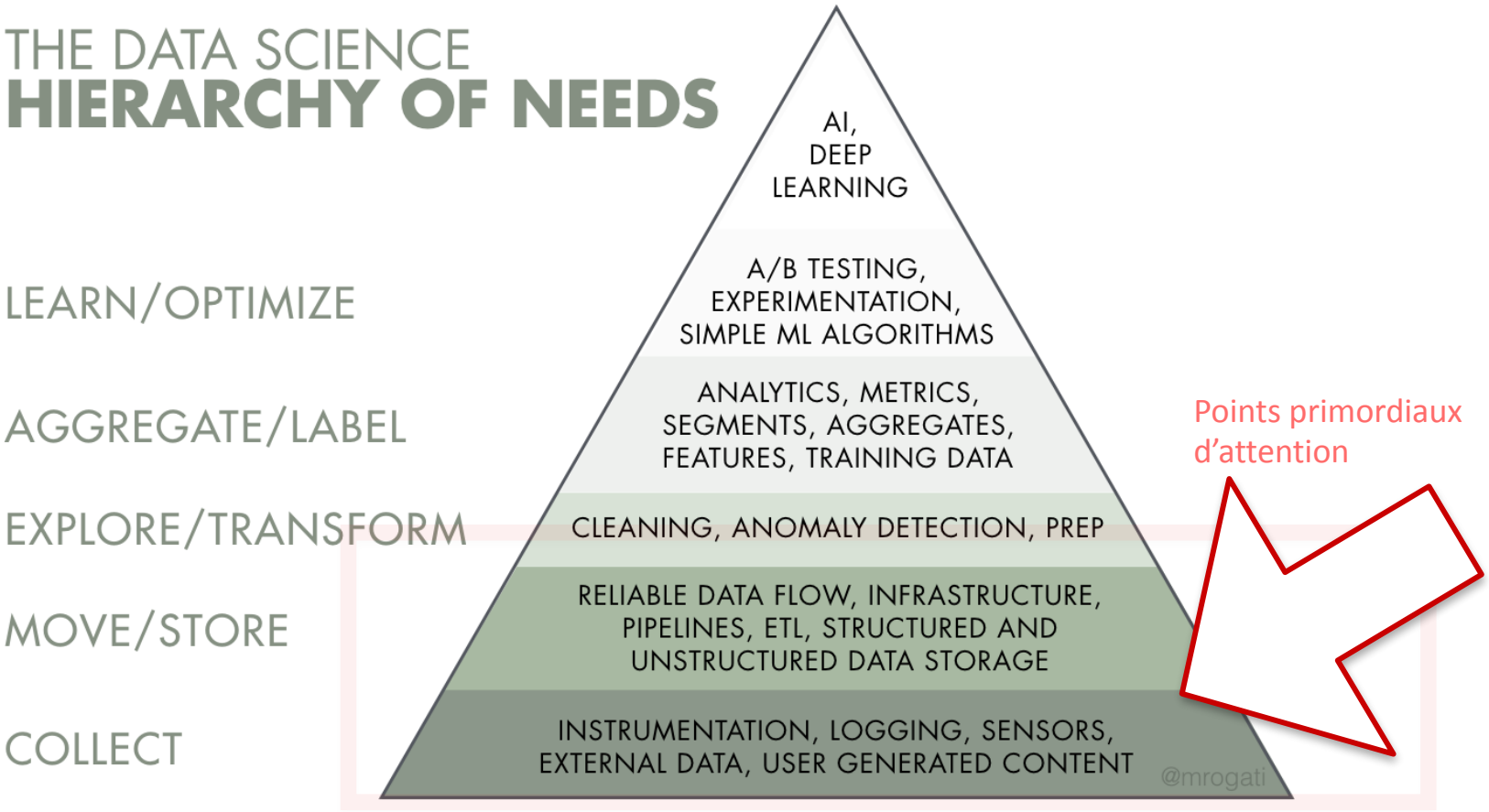
Session 2022

La qualité des données

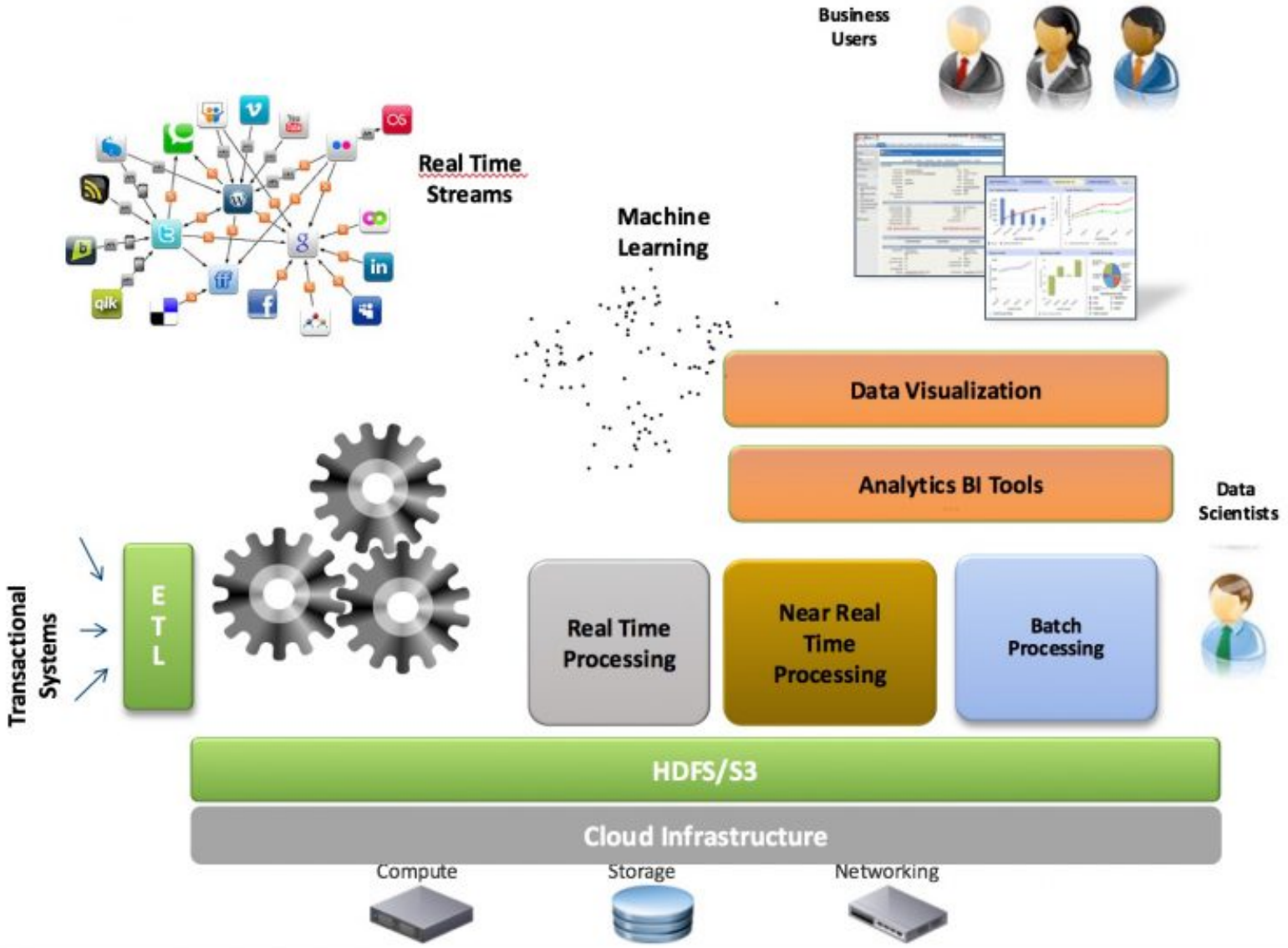
Mise en garde préliminaire

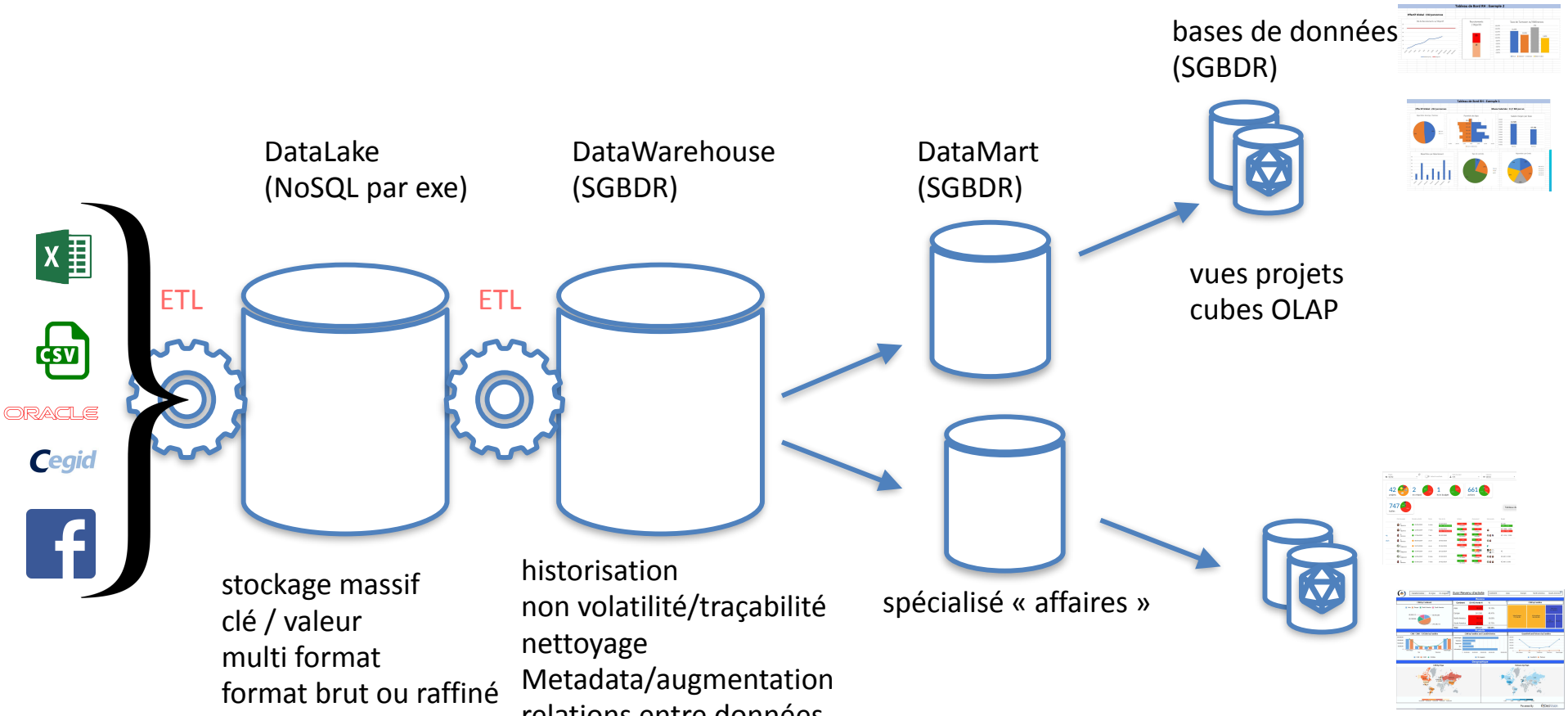
- Ce support est destiné à servir de base aux explications et exemples décrits en cours
- Il ne se substitue pas aux ouvrages plus complets relatifs à cette notion

le cheminement des données



Exemple de chaine de traitement




DataLake
(NoSQL par exe)
DataWarehouse
(SGBDR)
DataMart
(SGBDR)
bases de données
(SGBDR)
ETL
ETL
vues projets
cubes OLAP
stockage massif
clé / valeur
multi format
format brut ou raffiné
mise à l'échelle
historisation
non volatilité/traçabilité
nettoyage
Metadata/augmentation
relations entre données
spécialisé « affaires »

Quels enjeux?

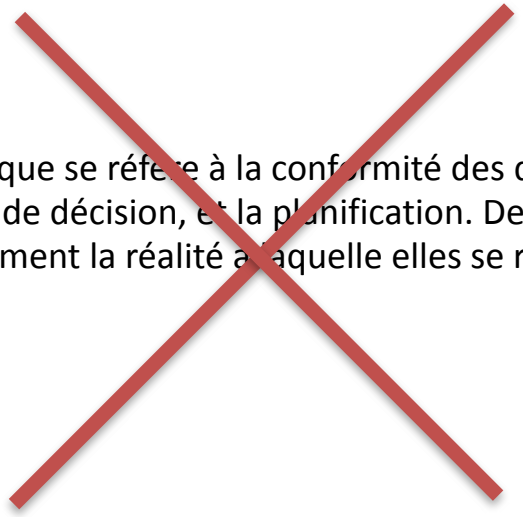
Plus de 55% des décisionnaires disent utiliser les données dans la prise de décisions. Mais seuls 33% les utilisent en toute confiance (Source études SAP & KPMG).

Les données sont de plus en plus stratégiques, mais leur qualité reste un enjeu majeur pour une très large majorité d'entreprises. De mauvaises décisions, fondées sur des données erronées, ont un impact direct sur le résultat des entreprises.

https://blog.atinternet.com/fr/infographie-la-qualite-des-donnees-en-digital-analytics/

Video : la qualité des données un enjeu majeur pour les entreprises
https://www.youtube.com/watch?v=YGpeO5dxVKs

La qualité des données, c’est quoi ?



La qualité des données, en informatique se réfère à la conformité des données aux usages prévus, dans les modes opératoires, les processus, les prises de décision, et la planification. De même, les données sont jugées de grande qualité si elles représentent correctement la réalité à laquelle elles se réfèrent.

Wikipedia

La qualité de la donnée, c’est quoi ?

- Une donnée est dite de (bonne) qualité si elle répond aux trois conditions suivantes :
  - Être unique : unicité de la réponse
  - Être intelligible : la réponse est cohérente avec la notion qu’elle renseigne
  - Être correcte : la réponse correspond à l’état de la donnée dans le contexte de la requête à l’instant de celle-ci.

La qualité de la donnée, c’est quoi ?

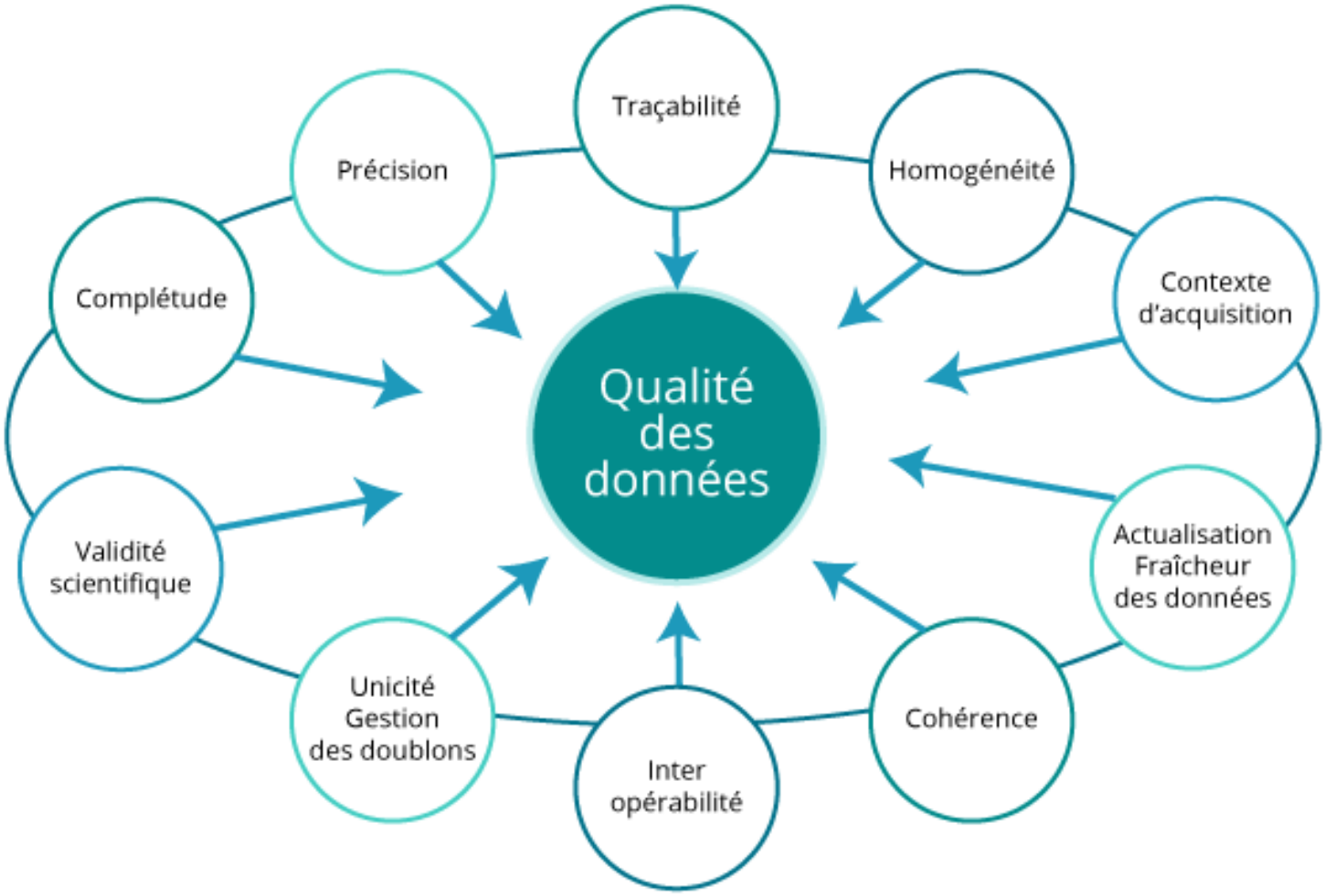
- Attention aux deux points suivants :
  - La définition de la donnée est propre au système d’information qui la traite : les unités seront choisies judicieusement et propre à un contexte (système international vs impérial par exemple).
  - Une même requête peut fournir deux résultats différents à deux instants distincts : ex l’âge d’une personne ou son adresse.

Faire perdurer la qualité

- La donnée de qualité doit la rester :
  - La donnée doit être rafraîchie selon des processus maitrisés à intervalles de temps cohérents avec son cycle de vie
  - Les procédures internes doivent être identifiées, documentées et suivies
  - L'obsolescence des données doit être évaluée et confrontée au coût de la non qualité

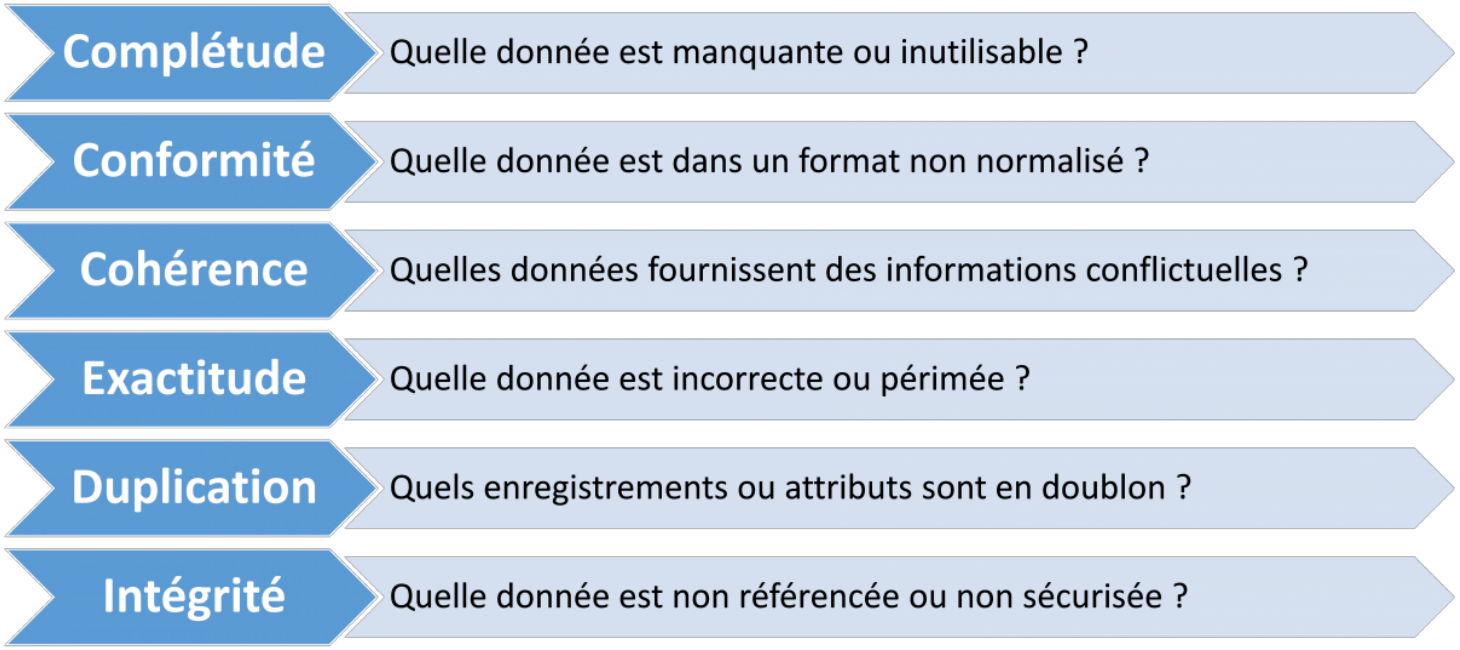
Les facteurs de qualité des données

De nombreux facteurs impactent la qualité des données



source INPN

Les principaux critères d'évaluation



cf : Source :https://blog.atinternet.com/fr/infographie-la-qualite-des-donnees-en-digital-analytics/

Les bonnes pratiques

- Créer de la valeur
- Mesurer la santé de ses données
- Distribuer les responsabilités
- Utiliser les outils à disposition

Créer de la valeur

- Identifier les données clés de l'entreprise, celles pour lesquelles l'activité peut courir un risque
- Confronter ces données aux processus de l'entreprise et s'assurer de leur pertinence
- Identifier et isoler les éléments qui pourraient contaminer la donnée et affecter leur exploitation
- Faire vivre ces données en les remettant en cause régulièrement
- Evaluer les pertes et profits engendrés par l'exploitation de ces données, documenter les éléments pour tracer les résultats

NF CONSULTING mutatis mutandis

Mesurer la santé des données

- Mettre en place des indicateurs de traitements sur les processus. i.e le taux de processus en succès/échec, le taux d'enregistrements qui ont dégradé les traitements…
- Identifier les seuils de qualité tangibles à partir desquels les traitements deviennent à risque ou erronés.
- Organiser les mesures préventives et curatives
- identifier les responsables de la production et du contrôle des données.

Distribuer les responsabilités

- A chaque niveau de production et de traitement un ou plusieurs responsable des données sont affectés à l'identification et au suivi des seuils. Leur vision métier est déterminante.
- Selon leur niveau de responsabilité fonctionnelle ou hiérarchique, l'étendue des données en responsabilité varie.
- Chacun doit pourvoir activer ou signaler une mesure à entreprendre pour consolider la qualité des données.
- exemple , pour une organisation , on définit les rôles suivants:
  - responsable d'une source de données et des règles/seuils correspondants
  - référent des données et des règles pour un domaine métier
  - responsable des données pour toute ou partie du SI

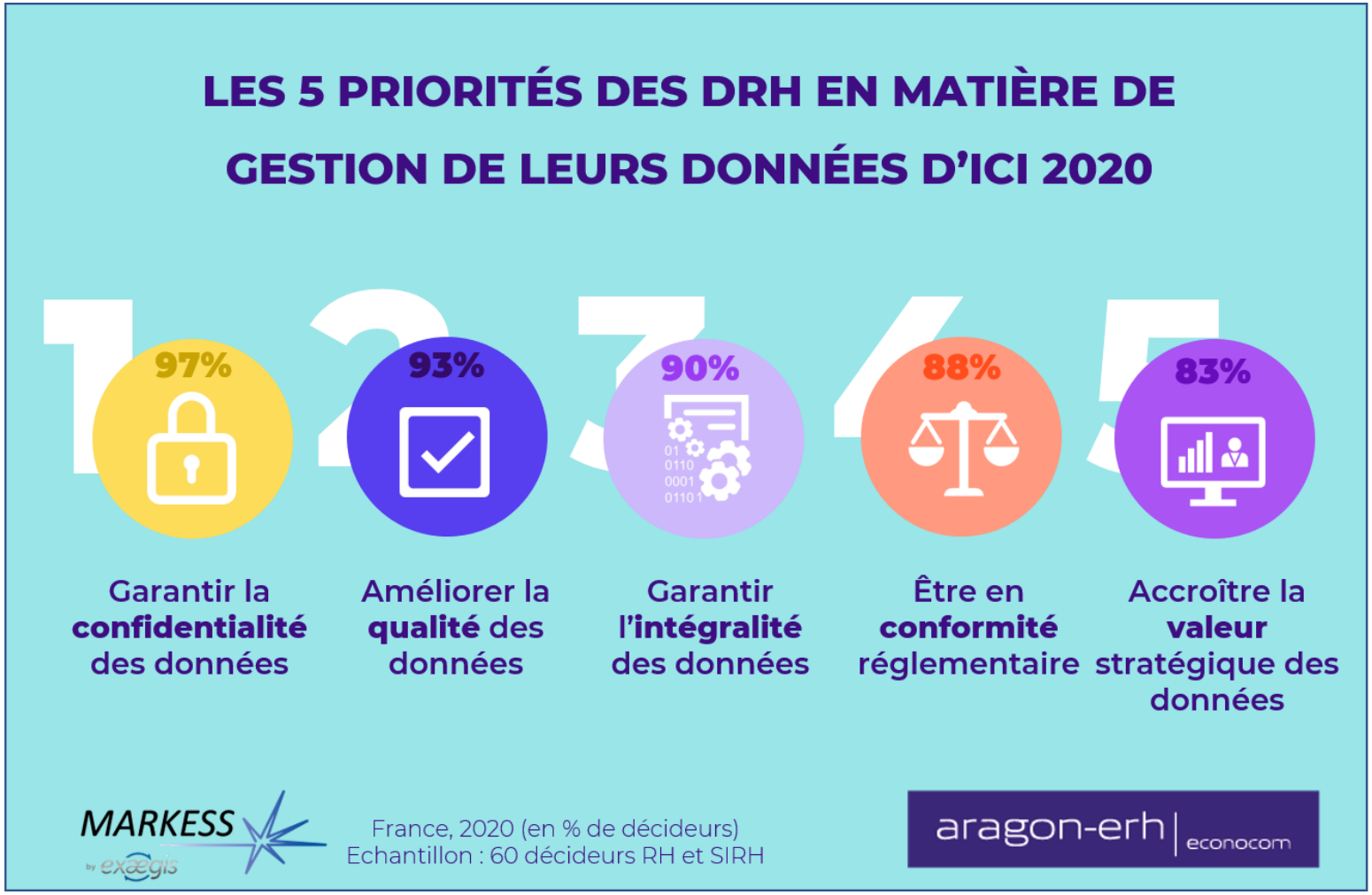
utiliser les outils à disposition

- Pour initier une démarche de qualité des données, des outils très variés existent. Il n'est pas nécessaire d'avoir un système qualité complet pour la démarrer.

- Une démarche MDM (Master Data Management) aboutie n'est pas nécessaire au départ. Elle viendra s'installer au fur et à mesure de la progression de l'organisation.

- Une organisation technique à minima peut être constituée comme suit:
  - un ETL pour l'intégration des données
  - une base de données pour stocker les informations
  - un environnement de traitement et de datavision pour exploiter les données, y compris qualifier leur santé

Cas d'application dans les ressources humaines



Renforcement du modèle

Le DQM

La qualité de données peut être altérée par différents facteurs que sont la quantité grandissante des données à gérer, l'hétérogénéité des sources, les différents silos applicatifs, les erreurs humaines de saisie, …

Pour remédier à ces effets, le Data Quality Management (DQM) offre un ensemble d'outils et de bonnes pratiques agissant tout au long du cycle de vie des données, afin d'en d'assurer leur qualité.

L'approche DQM est basée sur un processus pérenne en proposant un ensemble d'initiatives allant de la mesure, la correction et le contrôle de la qualité des données, à la mise en place de bonnes pratiques de gouvernance de la donnée

https://www.micropole.com/fr-fr/offres/data-governance-architecture/qualite-des-donnees

Le DQM

- Le DQM (Data Quality Management), associé à un flot ou plusieurs de travail (i.e Pipline) , est un processus qui vise à s'assurer que :
  - les valeurs obligatoires sont présentes
  - le types de données sont valides
  - les codes utilisés sont valides

- on veillera notamment à uniformiser les nommages et les traitements pour favoriser une qualité mesurée et reproductible.
- Le DQM permet de préparer les données (nettoyage, interpolation, dédoublonnage, standardisation, coquilles, format des dates et des numéros, données volontairement modifiées ou malveillantes…)
- Le DQM fonctionne par des règles programmées.
  - EX : définir une règle pour convertir toutes les dates MM/DD/YY en JJ/MM/AAAAA
- ex Outil IBM DQM

Principaux avantages

Une bonne qualité des données permet :

- d'améliorer la pertinence des data et des processus
- d'assurer la visibilité des informations
- d'assurer la fiabilité de l'information

https://www.capgemini.com/fr-fr/service/digital-services/insights-data-2/data-quality/#

Avantages pour les entreprises

- La Data aide à identifier de nouvelles opportunités et à améliorer les résultats de l'entreprise
- La qualité des données facilite la migration des données
- Garantir la qualité des données réduit le temps et les coûts de traitement des données

https://www.astera.com/fr/type/Blog/gestion-de-la-qualité-des-données/

Les risques
d'une mauvaise qualité des données

Des données mal organisées et obsolètes peuvent entrainer :

- une mauvaise prise de décisions
- des pertes d'opportunités
- une augmentation des coûts

Le bon outil de gestion : stratégie de DQM centralisée

- Data Quality Management:
  - Définir les principaux objectifs de réussite du programme de qualité des données
  - Communiquer le plan de gestion de la qualité à l'échelle de l'organisation
  - Évaluez les données entrantes par rapport aux métriques de qualité des données définies.
  - Analyser les résultats de la qualité des données et identifier les causes profondes des mauvaises données
  - Surveillez et ajustez les flux de travail relatifs à la qualité des données en fonction de l'évolution des besoins en données

Le MDM (Master Data Management)

Dans le cadre d'un DQM, gestion de la donnée de référence.

- « C'est une technologie de gestion des données de référence, aussi appelées données maîtres, qui vise à obtenir une information unique et partagée dans l'entreprise. » Dominique Mariko , MDM &DQM : synergies digitales et collaboratives 2016
- L'objectif est d'améliorer la qualité des données, notamment celles essentielles à l'activité et à la performance de l'entreprise (données de production, client, fournisseur, traçabilité réglementaire, reporting …)
- La finalité est notamment de synchroniser toutes les bases du SI en s'assurant de l'unicité, de l'intégrité, de l'exactitude et de la fraîcheur des données tout en s'assurant notamment d'un cycle fiable et adapté de la mise à jour pour conserver ou augmenter la qualité des données (démarche PDCA)
- Cette démarche permet de favoriser la numérisation de l'entreprise qui requiert une donnée fraîche, juste et porteuse de sens.
- Cette performance maitrisée permet d'ajuster les données utiles et optimiser les investissements autour de la data. Une diminution peut être recherchée ou une efficience renforcée à budget égal.

Le MDM

- Principales fonctionnalités recherchées:
  - Création de modèles de données maintenables et contrôlables pour assurer la pérennité de la qualité (données, meta données hierarchies, liens).
  - La gestion de l'unicité de la donnée (sources concurrentes) et la disponibilité des fonctionnalités d'enregistrement (gestion des erreurs et des conflits)
  - La gestion des connecteurs possibles et disponibles compatibles avec l'architecture et la dynamique du système d'information.
  - Intégration des caractéristiques métier du cycle de vie des données.

Les ETL

- Parmi les plus connus pour le nettoyage et l'exploitation des données:
  - Talend : https://fr.talend.com
  - Pentaho (Hitachi) : https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html

  - Les ETL (Export Transform Load) permettent notamment de modifier une donnée d'origine dans un format quelconque vers un format standardisé et nettoyé pour augmenter sa valeur métier.
  - Les ETL s'insèrent dans des suites logicielles de l'éditeur, en communicant par API voire par fichiers avec les flux du SI

  - => Il faut savoir quoi chercher et quelles sont les valeurs typiques attendues ou acceptables : la compétence métier est indispensable.

Les outils mathématiques

- Plusieurs ateliers mathématiques permettent d'optimiser le traitement des données:
- Matlab https://fr.mathworks.com/products/matlab.html
- Mathematica https://www.wolfram.com/mathematica/
- Scilab http://www.scilab.org
- Sans secret, les outils de qualité sont payants. Scilab est cependant un bon candidat open source.
- Ils nécessitent de très bonnes bases en mathématiques pour être utilisés efficacement et à bon escient. C'est un travail de spécialistes.

outils mathématiques

- L'outil mathématique est très puissant à partir du moment où ses hypothèses d'utilisation sont solides et le paramétrage pertinent. L'analyse du problème devra être fine et suffisamment complète.

- Les méthodes mathématiques utilisables classiquement:
  - probabilité , statistiques (moyenne, écart type, variance)
  - corrélation , convolution
  - interpolation , extrapolation, régression

- Les algorithmes utilisés seront sélectionnés en fonction du contexte et de la pertinence de chacun.

- Une approche critique basée sur le contexte et l'expérience sont nécessaires pour automatiser efficacement les traitements.

- Cf : ouvrages de mathématiques ex https://www.editions-ellipses.fr/index.php?controller=attachment&id_attachment=43228

Exemples

Sécurité et qualité des données : freins ou opportunités du big data?

Une grande majorité des entreprises interrogées (55 %) reconnaît l'incapacité des procédés en place à garantir la qualité des données.
Même les "datarati" (organisation ayant une démarche mature de big data) souffrent de cette insuffisance pour 45 % d'entre eux.

La situation semble moins mauvaise au Royaume-Uni et en France, où respectivement 53 % et 48 % des entreprises questionnées estiment disposer de processus de collecte garantissant la qualité des données.

Les erreurs liées à la saisie manuelle constituent dans 68 % des cas la cause du problème.

On comprend alors pourquoi seulement 27 % des entreprises interrogées font confiance aux données pour la prise des décisions les plus importantes, alors qu'elles sont 32 % à faire confiance à l'expérience et 33 % à faire confiance à l'instinct.

cf référence usine digitale en annexe.

Les idées neuves pour améliorer la qualité des données pour votre SI

Gouvernance des données

Grâce à des solutions dédiés (par ex: Gathering tools) , il est possible de remplacer des processus réalisés sous Excel par une application légère tout en préservant l'apparence et les fonctionnalités du processus existant

cf lien Gathering tools en annexe

Cas d'applications : ADIDAS

Frédéric Gaudin, Responsable domaines Fonctionnels Sales & Marketing chez Adidas France, revient sur la mise en place d'une solution simple et efficiente pour **faciliter et sécuriser la construction des forecasts commerciaux**.

Bénéfices de cette collaboration avec Gathering Tools **:**

- 1 jour au lieu de 5 pour la **collecte et la consolidation** des informations
- **Aucun changement visible pour les utilisateurs** : l'apparence et les fonctionnalités du classeur Excel sont conservées
- Une technologie qui vient renforcer la **proximité entre IT et métiers**

TP1

Sujet TP1:

Comparer deux jeux de données de climat pour déterminer la capitale européenne dont les données de température sont fournies dans le fichier Climat.xlsx
On se servira du fichier Savukoskikirkonkyla.xlsx issu de l'open data pour servir de référence.

Objectifs :

- Mettre en oeuvre un environnement de traitement graphique de données issues de sources plus ou moins fiables.
- Corriger un jeu de données mal formé
- proposer un candidat potentiel pour l'origine des données.

Déroulement:

- Pour l' échantillon SI, calculez :
  - moyenne par mois
  - écart type par mois
  - min /max par mois et par année
- utiliser par Python Scipy pour les parties mathématiques.
- tracer les courbes de chaque mois avec une bibliothèque graphique python Matplotlib, 12 vues mensuelles
- Assembler les courbes sur un seul graphique (J1 -> J365) : vue annuelle
- Présenter la valeur lue en parcourant la courbe à l'aide du pointeur,
- Présenter les valeurs précédentes par mois glissant de 30 jours centré sur la valeur lue
- Recommencez avec le jeu SI-erreur après avoir corrigé les valeurs en erreur. Précisez vos méthodes.
- Les données corrigées sont elles proches des valeurs sans erreur ?
- A partir de données opendata du second fichier, retrouver le type de climat
  - reprendre les données typiques de la localisation proche fournies en complément , comparer les écarts.
  - Qu'en concluez vous ?
  - De quelle la capitale européenne avez vous eu les données .

Outils : à utiliser Python + matplotlib, Jupyter éventuellement. Pas de R ni d'autre langage autorisés
Evaluation:
Démonstration des solutions techniques et argumentation sur les méthodes utilisées

Références

https://www.astera.com/fr/type/Blog/gestion-de-la-qualité-des-données/

https://www.usine-digitale.fr/article/securite-et-qualite-des-donnees-freins-ou-opportunites-du-big-data.N328082

https://www.sparklane-group.com/fr/blog/comment-resoudre-le-probleme-de-la-qualite-des-donnees-clients/

https://www.gathering-tools.com/2017/08/31/idees-ameliorer-qualite-donnees-dans-si/

7-success-story-adidas-gathering-toolspdf.pdf

Merci

Session 2019

Nicot Francois Consulting
SARL au capital de 2500 euros
SIRET 809528250 00016 R.C.S Rennes