# Bases de données

## Introduction

BD000
v215b

2020-09-08

Département d'informatique
Faculté des sciences

UNIVERSITÉ DE SHERBROOKE

Christina.Khnaisser@USherbrooke.ca
http://info.USherbrooke.ca/ckhnaisser
Luc.Lavoie@USherbrooke.ca
http://info.USherbrooke.ca/llavoie

# Plan

- Contexte
- Modèles
- Futurologie
- Vocabulaire
- Références
- Les colles du prof !
- Bibliographie

# CONTEXTE

- Les domaines d'application
- Une caractérisation des problèmes
- Le complot des scientifiques
- Le complot des gestionnaires

- Besoin de traiter, de conserver et d'analyser de (très) grandes quantités d'informations
  - gouvernements (recensement, impôts, santé...)
  - banques et assurances
  - recherche scientifique (astronomie, chimie, génétique...)
  - télécommunications
  - secteurs industriels et manufacturiers
  - secteur énergétique
  - grande distribution
  - géomatique
  - agences de renseignement, de propagande et de marketing
  - ...

## Contexte
## Une caractérisation des problèmes (3 V)

- Comment caractériser les problèmes afin de déterminer les solutions les plus adéquates ?

- Spectre traditionnel des 3 V
  - **volume**
  - **variété**
  - **vélocité**

## Contexte
## Une caractérisation des problèmes (4 V)

- Comment caractériser les problèmes afin de déterminer les solutions les plus adéquates ?

- Spectre contemporain des 4 V, on ajoute :
  - **véracité**
    (afin de rendre compte de l'incertitude)
  - volume
  - variété
  - vélocité

## Contexte
## Une caractérisation des problèmes (5 V)

- Comment caractériser les problèmes afin de déterminer les solutions les plus adéquates ?

- Spectre « monétisé » 5 V, on ajoute
  - **valeur**
    (afin de rendre compte de la richesse analytique)
  - véracité
  - volume
  - variété
  - vélocité

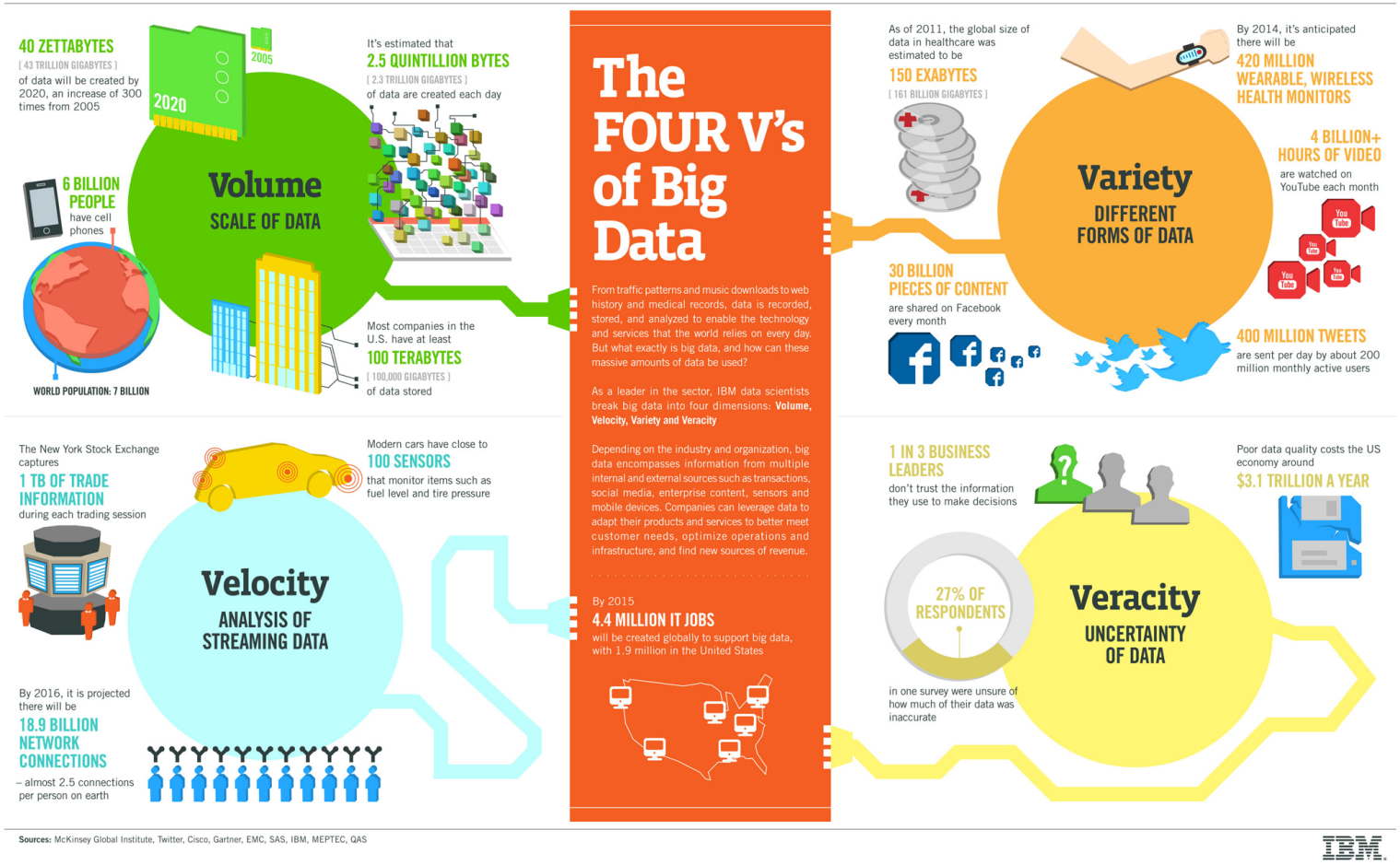
## Contexte
## Une caractérisation des problèmes (6 V)

- Comment caractériser les problèmes afin de déterminer les solutions les plus adéquates ?
- Spectre « géographique » 6 V, on ajoute
  - **virtualité**
    (afin de rendre compte de l'emplacement des données)
  - valeur
  - véracité
  - volume
  - variété
  - vélocité

## Contexte
## Une caractérisation des problèmes (7 V)

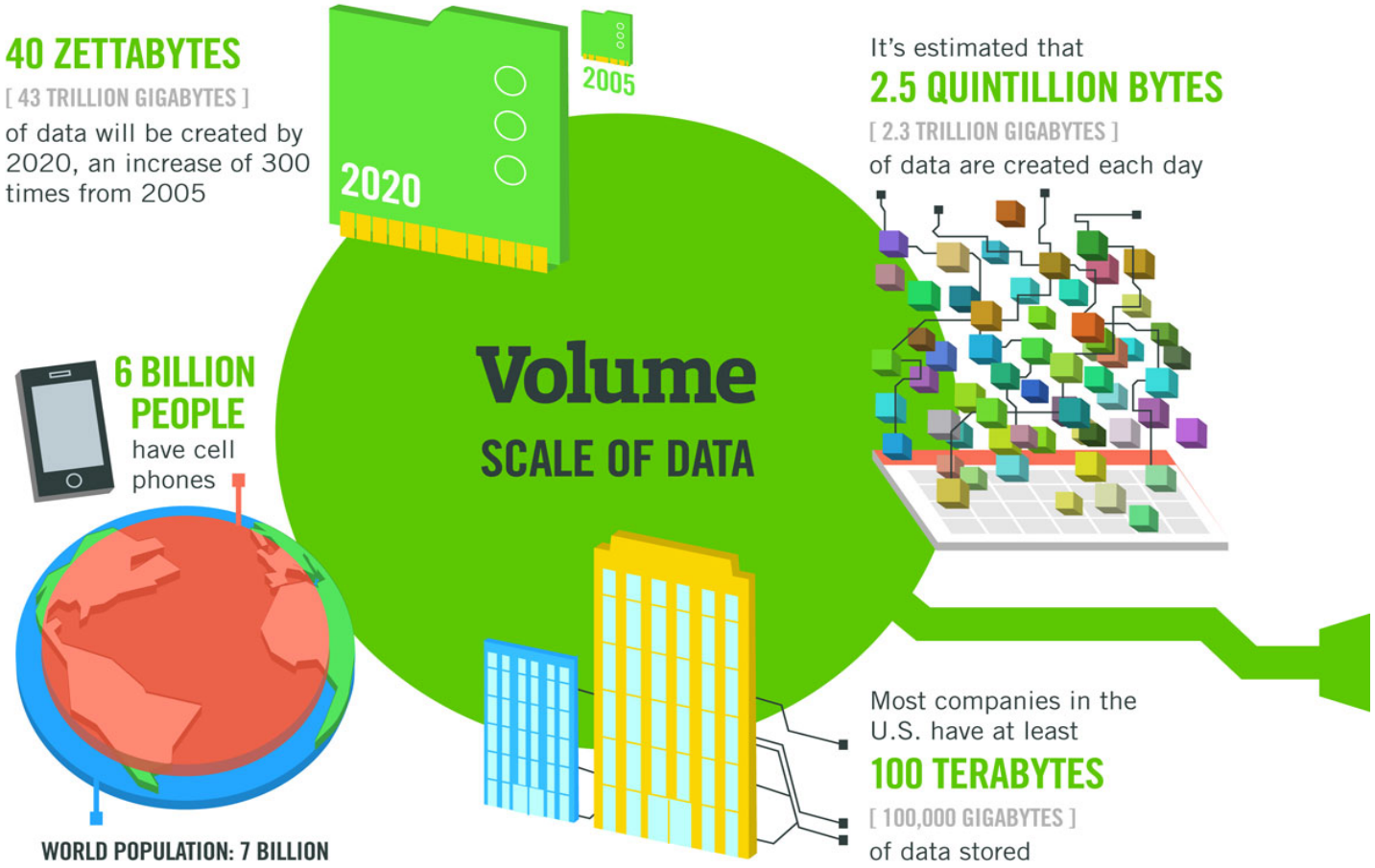
- Comment caractériser les problèmes en regard exigences découlant des lois et règlements quant à la protection et l'utilisation éthique des données ?
- Spectre « vertueux », Talend propose
  - **vertu (*virtue*)**
    (afin de rendre compte des lois et règlements quant à la protection et l'utilisation éthique des données)
  - virtualité
  - valeur
  - véracité
  - volume
  - variété
  - vélocité

# UNE ILLUSTRATION DES « V » PAR IBM



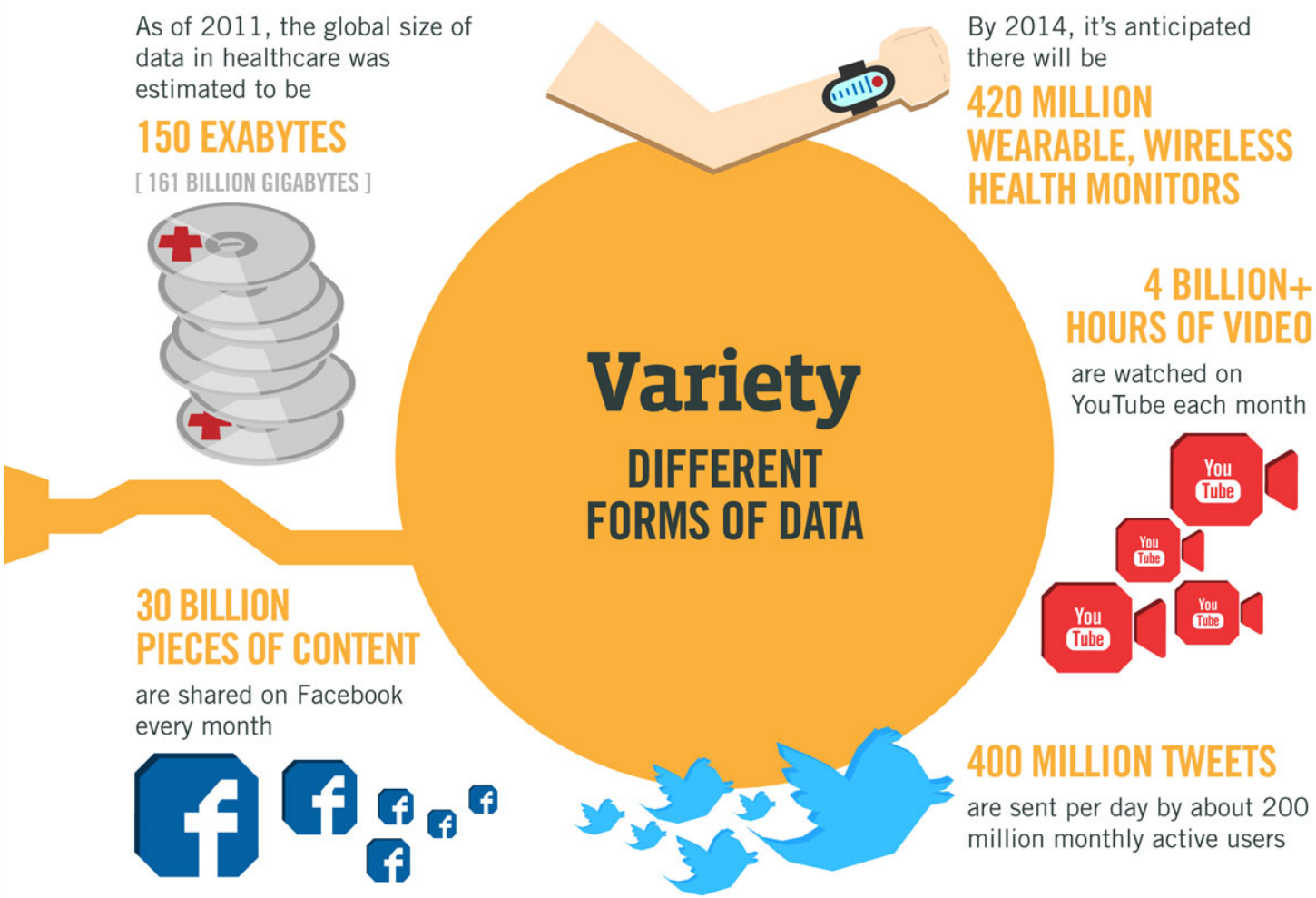

**source : http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg**

# IBM - Volume : Pour l'année de référence 2013, vraiment ?



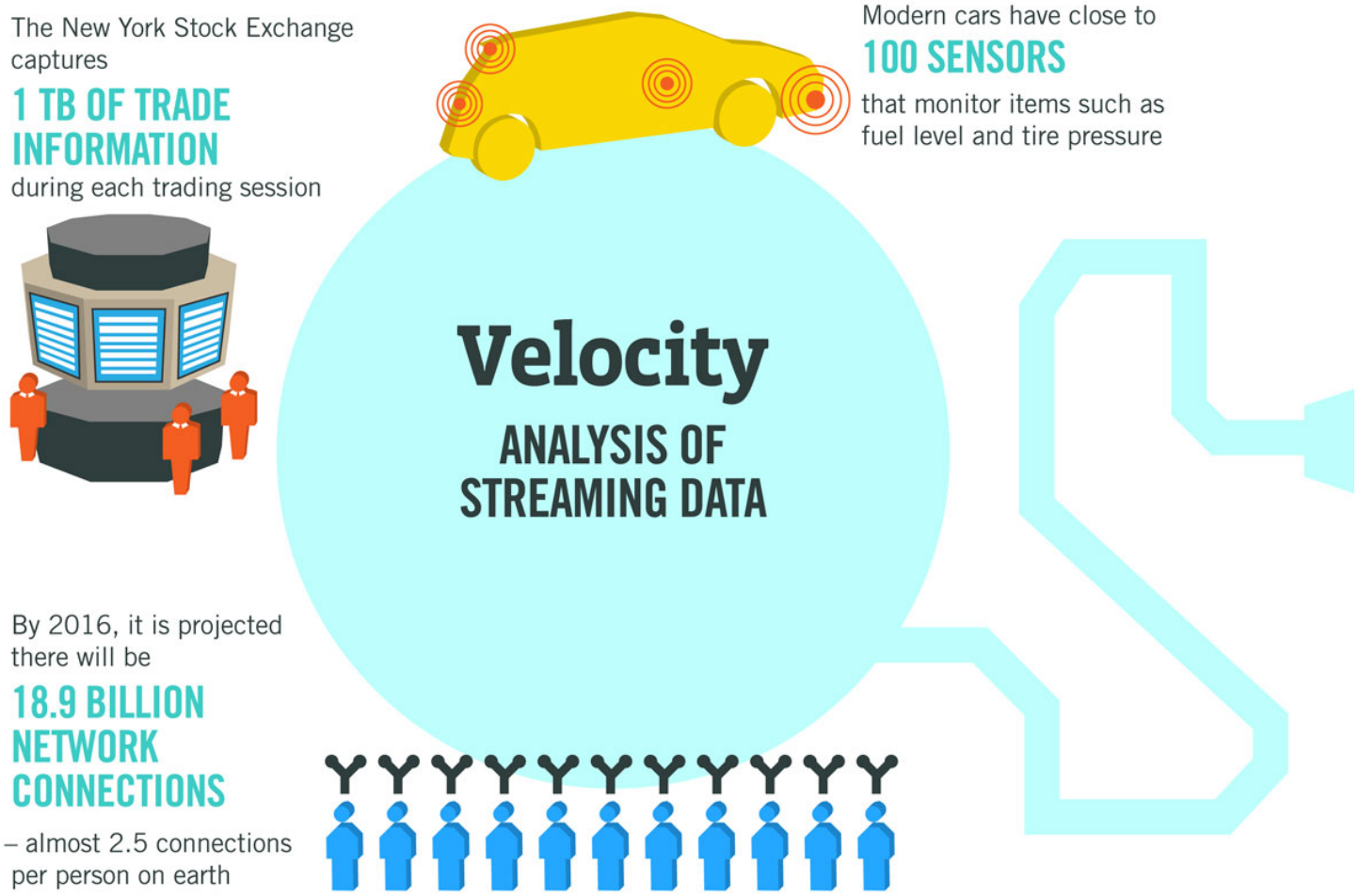

**source : http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg**

# *IBM – VARIÉTÉ : SOYEZ VIGILANTS ET CRITIQUES!*



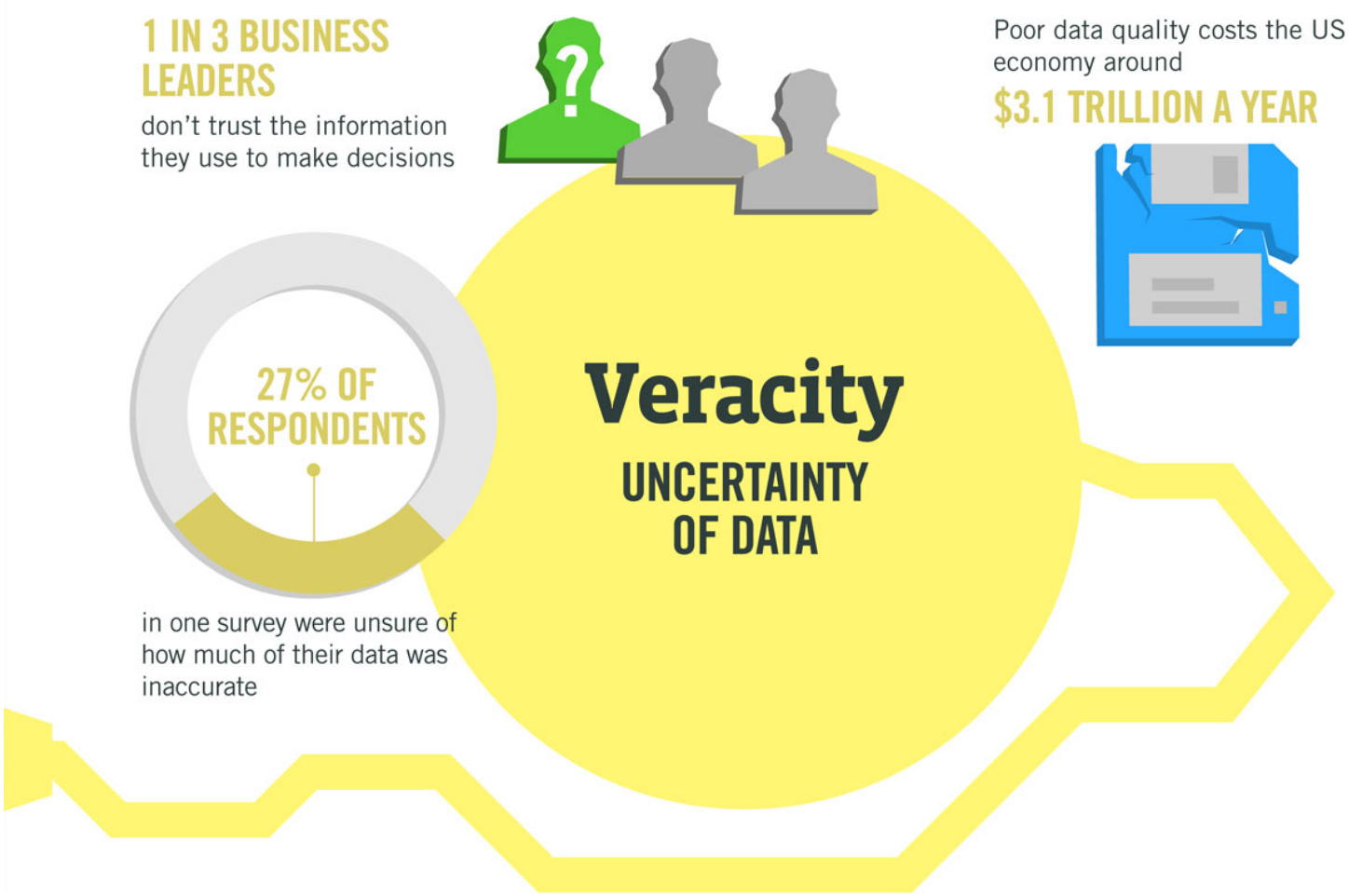

**source : http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg**

# *IBM – VÉLOCITÉ : SOYEZ VIGILANTS ET CRITIQUES!*



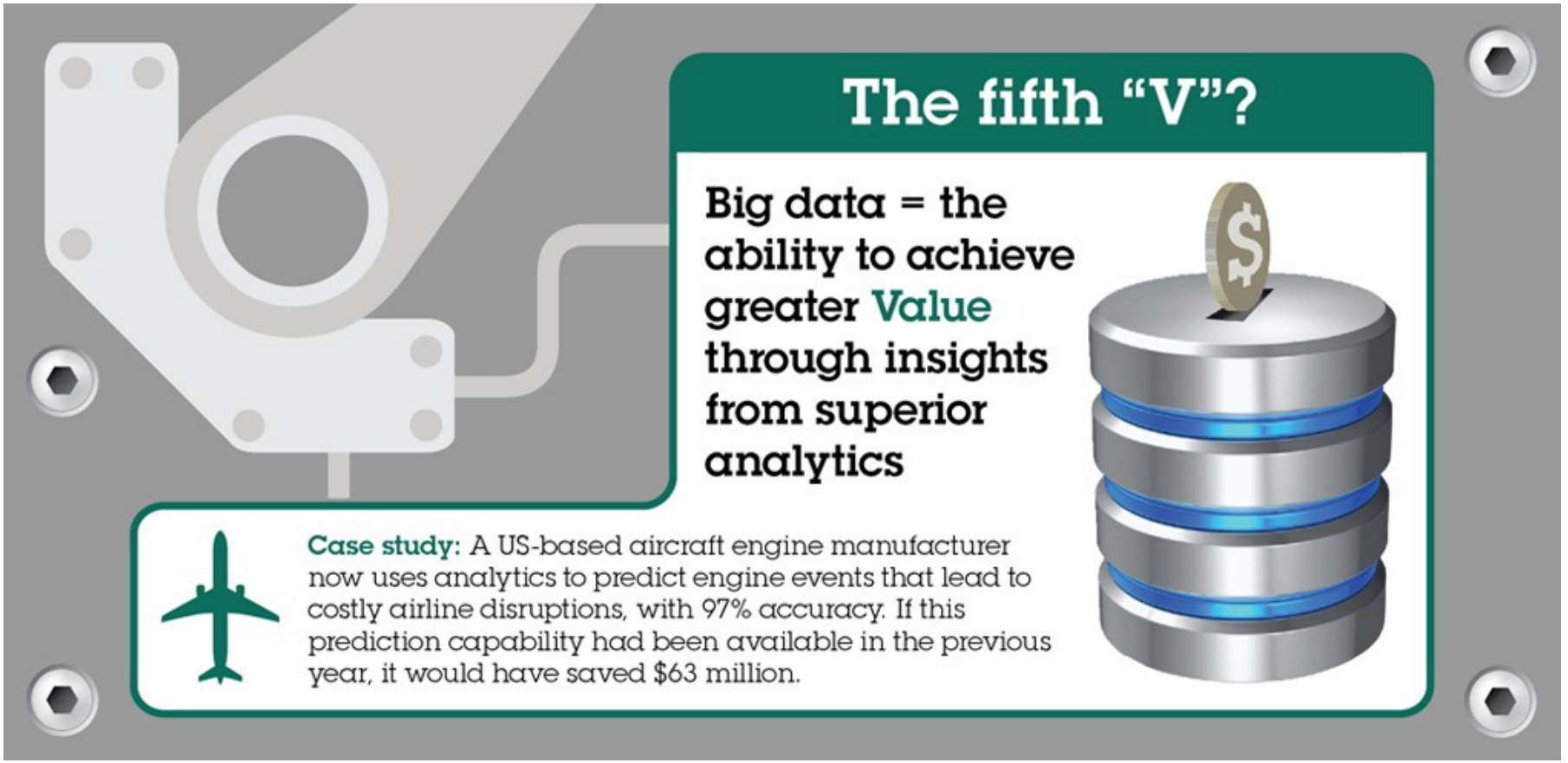

**source : http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg**

# *IBM – VERACITY : SOYEZ VIGILANTS ET CRITIQUES !*




**source : http://www-01.ibm.com/software/data/bigdata/images/4-Vs-of-big-data.jpg**

# IBM – VALEUR : LES PROMESSES DE *2015*...




**source :**
**http://www.ibmbigdatahub.com/infographic/extracting-business-value-4-vs-big-data**

## Illustration des autres V

- Suggestions ?

## Contexte
## Une caractérisation des problèmes (7 V + 1 T)

- Comment caractériser les problèmes afin de déterminer les solutions les plus adéquates sans parler de temporalité, voire d'historicisation ?

- Cet aspect est traité dans l'activité IGE 487.

- Intégrer différents types de données
  - Structurées
  - Semi-structurées
  - Non structurées
- Intégrer différentes sources de données
  - Internet des objets
  - Médias sociaux
  - ...

## Contexte
## Le complot des scientifiques

- Une volonté de comprendre « la réalité », puis
  - de documenter cette compréhension,
  - de la transmettre,
  - de l'opérationnaliser,
  - de l'appliquer.
- Une approche qui repose
  - sur le raisonnement axiomatique,
  - afin (notamment) de garantir la réfutabilité,
  - en se fondant sur la logique (du premier ordre)
- Cette approche a donné naissance à la **théorie relationnelle** (de modélisation des données).

- *Cette approche s'accorde très souvent avec la mission des organisations en ce sens qu'elle constitue un outil puissant pour en gérer optimalement les ressources.*
- *Le méta-complot des gestionnaires :*
  - *Tirer parti des avancées de la théorie relationnelle pour mieux gérer.*

# MODÈLES

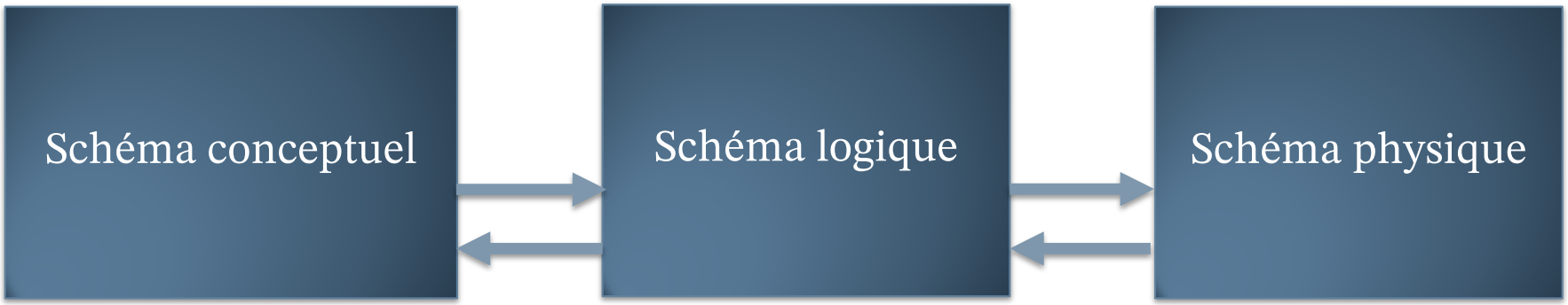
- Principes CMcm
- Trischématisation
- SGBDR
- Modèles et modèles de modèles

## Principe CMcm (Cohérence maximale pour couplage minimal)

- **Cohérence**
  rien ne se contredit.
- **Cohésion**
  tout se tient.
- **Couplage**
  liaison entre deux éléments; en particulier une liaison peut être une dépendance.
- *Si tous les éléments d'un ensemble sont couplés (transitivement) alors l'ensemble est (totalement) cohésif.*
- Il est (plus) facile de constater l'incohérence dans un ensemble cohésif.
- La cohésion rend très difficile la modification (l'évolution)
- On en déduit un principe :
  - Cohérence maximale pour un couplage minimal

## Modèles
## Trischématisation, une solution ?

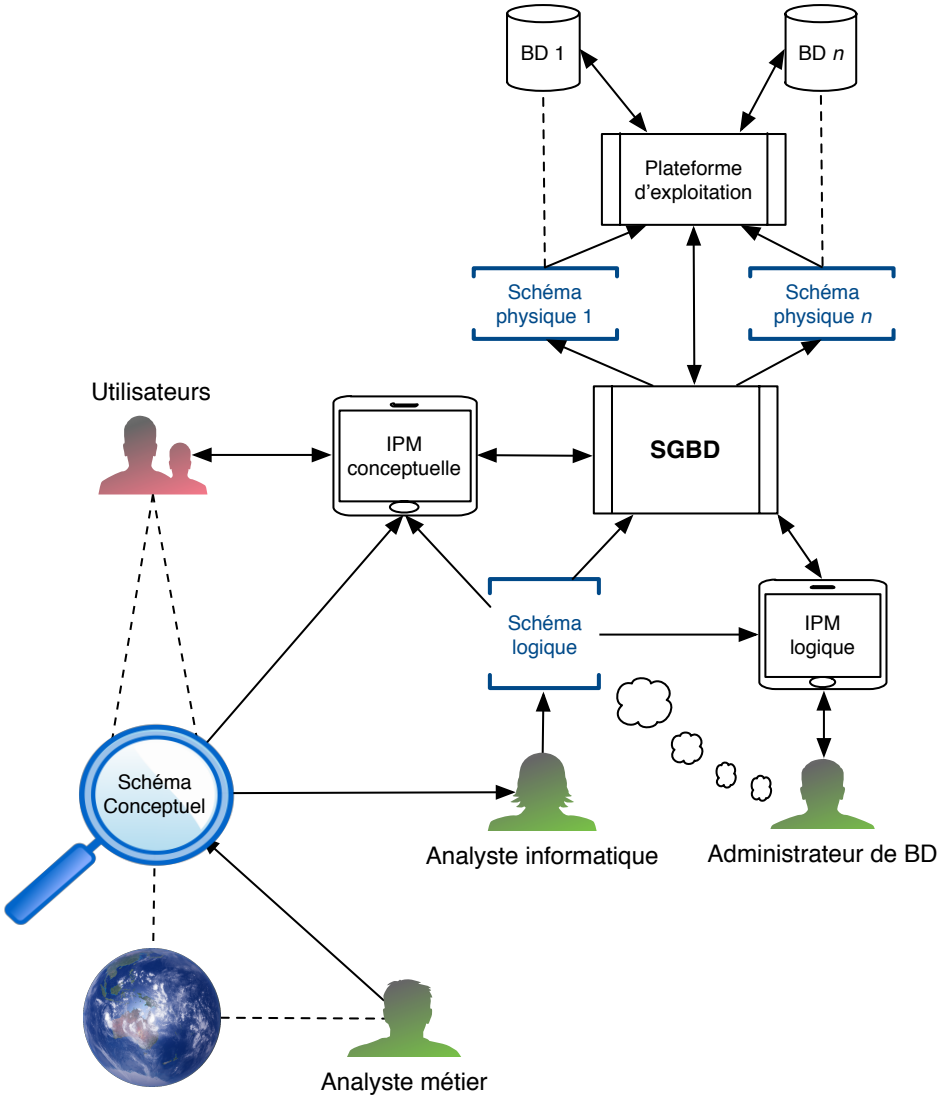
- L'application du principe CMcm au problème de l'intégrité a conduit à l'élaboration du modèle *trischématique*.
- Ce modèle a également permis de répondre à plusieurs préoccupations d'efficacité et d'efficience.
- D'où sa longévité.



# Trischématisation
# Un exemple : le SGBD

**Schémas :**

- **conceptuel**
- **logique**
- **physique**



## Théorie et modèles

- conceptuel
  - EA (entité-association)
    ER (*entity-relationship*)
  - EAE (EA étendu)
    ERE (*extended ER*)
  - ...
- logique
  - hiérarchique
  - réseau
  - relationnel
  - objet
  - ...
- physique
  - adressage dispersé (*hashing*), arbre de recherche (*B-Tree*)
  - horizontal, vertical, mixte, KV
  - ...

## TROIS NIVEAUX SONT-ILS SUFFISANTS ?

- Beaucoup de chercheurs estiment aujourd'hui qu'un quatrième niveau est nécessaire : celui des **modèles de connaissances**.

## Dans ce cours (1)

- Nous recherchons :
  - une fondation solide à la modélisation de données
  - une grande souplesse de modélisation
  - la capacité de dériver une « machine » performante à partir du modèle de données automatiquement (sans avoir à élaborer le modèle physique)

## Dans ce cours (2)

- Nous nous intéresserons à **une** théorie des schémas logiques :
  - la **théorie relationnelle**
  - et à certains de ses modèles.

- Nous nous intéresserons à **une** théorie des schémas conceptuels :
  - la **théorie entité-association**
  - et à certains de ses modèles.

## Modèles et schémas

- Un modèle de données (dérivé d'un modèle d'information s'appuyant lui-même sur un modèle de connaissances, chaque modèle étant décrit à l'aide d'une notation associée à un modèle de modélisation approprié) sera désormais désigné sous le nom de schéma…

- En fait, le mot modèle est
  - surutilisé
  - dans des contextes différents
  - pour décrire des choses différentes
  - … il faut donc toujours en préciser le sens

- Voilà qui est fait!

# FUTUROLOGIE

- Un exemple
- De l'esprit critique
- Exercices

## Futurologie
## Un exemple

- Préambule
  - De la difficulté de la prospective
  - De la nécessité de l'analyse critique
- Un exemple
  - Le taux de pénétration des appareils mobiles en 2013 avancée par IBM en 2010 (voir l'illustration du critère de volume proposée précédemment)

*Nombre d'abonnements au large bande mobile et taux de pénétration dans le monde d'ici à la fin 2013 (estimations) et taux composé de croissance annuelle (TCAC) pour 2010–2013*

**Amériques**
**460 millions** d'abonnements
Taux de pénétration de **48%**
TCAC de **28%** (2010–2013)

**Europe**
**422 millions** d'abonnements
Taux de pénétration de **68%**
TCAC de **33%** (2010–2013)

**CEI**
**129 millions** d'abonnements
Taux de pénétration de **46%**
TCAC de **27%** (2010–2013)

**Etats arabes**
**71 millions** d'abonnements
Taux de pénétration de **19%**
TCAC de **55%** (2010–2013)

**Afrique**
**93 millions** d'abonnements
Taux de pénétration de **11%**
TCAC de **82%** (2010–2013)

**Asie-Pacifique**
**895 millions** d'abonnements
Taux de pénétration de **22%**
TCAC de **45%** (2010–2013)

**https://itunews.itu.int/Fr/3855-Le-nombre-dabonnements-aumobile-frole-les-septmilliardsbrUn-telephone-pour-chacun-ou-presque.note.aspx**

## FUTUROLOGIE
### ESPRIT CRITIQUE - TÉLÉPHONE

- Une étude plus sérieuse estime à 3,4 milliards de personnes le nombre de possesseurs d'un appareil fonctionnel et connecté en 2013, soit moins de 50 % de la population mondiale. Quant au nombre d'appareils reliés effectivement à Internet, il était inférieur à 2 milliards... et plusieurs de ceux-ci étaient embarqués dans des équipements. De plus, les marges d'erreur sont considérables. Voir :
  - https://itunews.itu.int/Fr/3855-Le-nombre-dabonnements-aumobile-frole-les-septmilliardsbrUn-telephone-pour-chacun-ou-presque.note.aspx
- Voir également
  - http://www.journaldunet.com/ebusiness/internet-mobile/1009553-monde-le-nombre-d-abonnes-au-telephone-mobile/

## FUTUROLOGIE
### ESPRIT CRITIQUE - POPULATION

- Pour des statistiques relatives à la population mondiale, aux pyramides d'âges, etc., voir :
  - http://www.prb.org/pdf12/2012-population-data-sheet_french.pdf
  - https://www.prb.org/wp-content/uploads/2020/07/PRB2020WPDS-BOOKLET-FR.pdf
  - https://population.un.org/wpp/Publications/Files/WPP2019_DataBooklet.pdf
- et comparer les méthodologies et les conclusions.

## Futurologie
## Exercice 1

- Soumettez les illustrations de la caractérisation des « V » proposées par IBM à l'analyse critique.
- Faites de même pour l'illustration proposée par le GRIIS.

## Futurologie
## Exercice 2 et 3

- Quelles étaient les prétentions des grands acteurs des télécommunications en 2000 (Cisco, Nokia, Motorala, Ericsson, Nortel, etc.) ?
  - Qu'en est-il advenu ? Pourquoi ?
- Quelles sont les prétentions des grands acteurs du Web en 2020 (GAFAM et les autres) ?
  - Qu'en adviendra-t-il ? Pourquoi ?

## VOCABULAIRE
## CONCEPTS DE BASE

- informatique
- connaissance
- information
- donnée
- représentation
- valeur
- domaine
- type

- mesure d'encombrement

## VOCABULAIRE – DÉFINITIONS UTILES (1)

- informatique
  - 1a - Science du traitement de l'information.
  - 1b - Science du traitement automatique et rationnel de l'information.
  - 2 - Ensemble des techniques de la collecte, du tri, de la mise en mémoire, du stockage, de la transmission, et de l'utilisation des informations traitées automatiquement à l'aide de logiciels mis en oeuvre sur des ordinateurs.

## VOCABULAIRE – DÉFINITIONS UTILES (2)

- information
  - Élément de connaissance représentable par une donnée.
  - Élément de connaissance susceptible d'être transmis au moyen d'une suite de signes.

connaissance ?

donnée ?

signe ?

## VOCABULAIRE – DÉFINITIONS UTILES (3)

- connaissance
  - *Faculté mentale produisant une assimilation par l'esprit d'un contenu objectif préalablement traduit en signes et en idées.*
  - *Résultat de cette opération. La connaissance est une possession symbolique des choses. Elle comprend une infinité de degrés. La connaissance rationnelle, méthodique universelle a parfois été opposée au savoir empirique, chaotique, objectif.*

référence : Godin Christian, Dictionnaire de philosophie, Paris, Fayard, 2004, ISBN 978-2-213-62116-6. Cité par Wikipédia, https://fr.wikipedia.org/wiki/Connaissance [2020-08-16]

## VOCABULAIRE – DÉFINITIONS UTILES (4)

- donnée
  - valeur associée à une représentation (apte à être traitée par ordinateur).
- représentation
  - une suite de signaux
    (un signal est un phénomène physique mesurable, donc suffisamment stable pour être mesuré).
- valeur
  - élément d'un domaine.
- domaine
  - ensemble fini de valeurs propres
    (n'appartenant à aucun autre domaine).
- type
  - domaine muni d'une contrainte
    (qui restreint les valeurs acceptées).

## Remarques

- Remarquons la circularité entre la définition de valeur et celle de domaine. Leur existence relève des postulats de base de la théorie de l'information.
- Remarquons également la circularité entre connaissance, information, donnée et signe.
- Quelqu'un a-t-il une meilleure suggestion ? Il y a peut-être un Ph. D. à la clé!

## VOCABULAIRE – DÉFINITIONS UTILES (4)

- Exigences applicables aux définitions
  - Toute valeur doit avoir au moins une représentation (mais une même valeur peut avoir plusieurs représentations).
  - Lorsqu'une même représentation est associée à plus d'une valeur, le « contexte » doit permettre d'inférer la valeur que représente une donnée.
    - Par exemple, attendu un tuple, un attribut donne accès à une donnée dont la valeur appartient au type de l'attribut (l'attribut fournit donc le contexte par le type qui lui est associé).

## VOCABULAIRE — DÉFINITIONS UTILES

- schéma (d'une base de données) ≡
  - *modèle logique de données*
- base de données ≡
  une instance d'un schéma ≡
  une instance d'un modèle logique de données ≡
  - *un contenant regroupant les données représentant un état de la réalité telle que modélisée grâce au modèle conceptuel à l'origine du schéma.*

# VOCABULAIRE
## MESURE D'ENCOMBREMENT ET SYMBOLES ISO

**Unités internationales (indépendantes de la langue)**

- bit (b)
- octet (o)

**Préfixes (indépendantes de la langue)**

| Préfixe | Base 2 | Symbole | Base 10 | Symbole |
|---|---|---|---|---|
| kilo | $2^{10}$ | kio | $10^{3}$ | ko |
| mega | $2^{20}$ | Mio | $10^{6}$ | Mo |
| giga | $2^{30}$ | Gio | $10^{9}$ | Go |
| tera | $2^{40}$ | Tio | $10^{12}$ | To |
| exa | $2^{50}$ | Eio | $10^{15}$ | Eo |
| peta | $2^{60}$ | Pio | $10^{18}$ | Po |
| zetta | $2^{70}$ | Zio | $10^{21}$ | Zo |
| yotta | $2^{80}$ | Yio | $10^{24}$ | Yo |

# Conclusion

- Références
- Les colles du prof
- Quelques autres références....

## Références

- Bases de données (Databases) :
  - [Date 2004], chapitres 1 et 2
  - [Elmasri 2004], chapitres 1 et 2
  - [Elmasri 2016], chapitres 1 et 2
  - [Ullman 2008], chapitre 1
  - https://fr.wikipedia.org/wiki/Base_de_données
  - https://en.wikipedia.org/wiki/Database
- Mégadonnées (Big Data) :
  - https://fr.wikipedia.org/wiki/Big_data
  - https://en.wikipedia.org/wiki/Big_data

## Les colles du prof (2)

- À quels besoins fondamentaux répond une base de données?
- Quelles sont ses « qualités » essentielles?
- Quels sont ses fondements théoriques?
- Quelle est l'architecture d'un SGBD?
- Quelle différence y a-t-il entre un SGBD, une BD et un schéma?

## Les colles du prof (2)

http://en.wikipedia.org/wiki/Alan_Turing#mediaviewer/File:Alan_Turing_photo.jpg

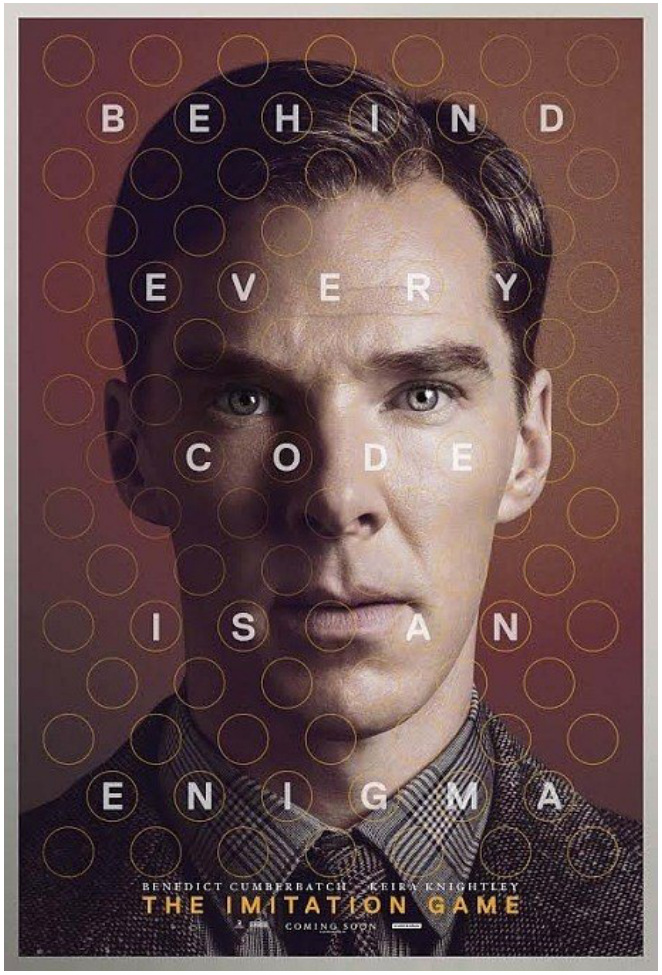
# Alan Turing (1912-1954)

Qui était Alan Turing?

Comment est-il mort?

Qu'est-ce que le prix Turing?

# Imitation Game (2014)




**https://www.imdb.com/title/tt2084970/reference**

Réalisateur :

Morten Tyldum

Scénaristes :

Graham Moore,

Andrew Hodges

Acteur principal :

Benedict Cumberbatch

Synopsis :

Une version romancée d'une partie déterminante de la vie de Turing : le décryptage d'Enigma.

## Biographies et autres sources


- https://fr.wikipedia.org/wiki/Alan_Turing
- Jean Lassègue, *Turing*, Paris, Les Belles lettres, 1998. (ISBN 978-2-2517-6014-8)
- Laurent Lemire, *Alan Turing : l'homme qui a croqué la pomme*, Paris, Hachette Littératures, 2004, 191 p. (ISBN 978-2-0123-5618-4)
- Jean Lassègue, *Les Génies de la science*, Pour la Science, nº 29 « Turing... et l'informatique future », nov. 2006 – janv. 2007. (ISBN 978-2-8424-5078-6)
- Andrew Hodges et Douglas Hofstadter, *Alan Turing : The Enigma*, Princeton University Press, 2012, 586 p. (ISBN 978-0-6911-5564-7)