BIG DATA: la QUANTITÉ ne fait pas la QUALITÉ

Jef Wijsen

Département d'Informatique
Université de Mons

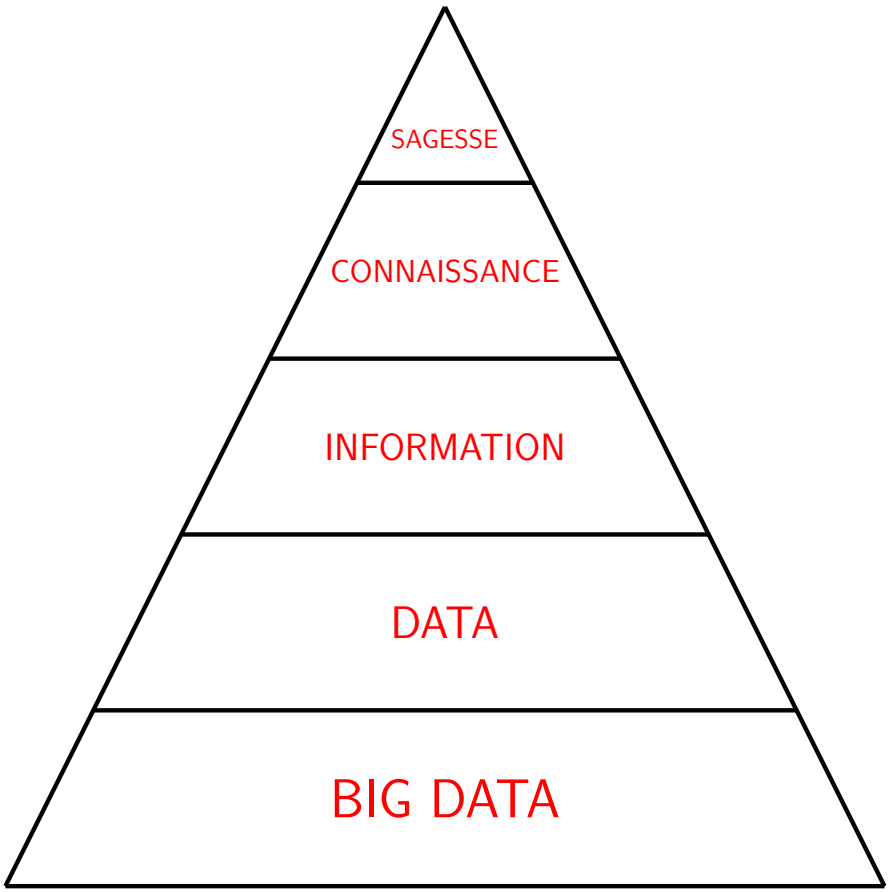
UMONS Faculté des Sciences

Troisième journée scientifique du Pôle hainuyer
Tournai, 25 avril 2017

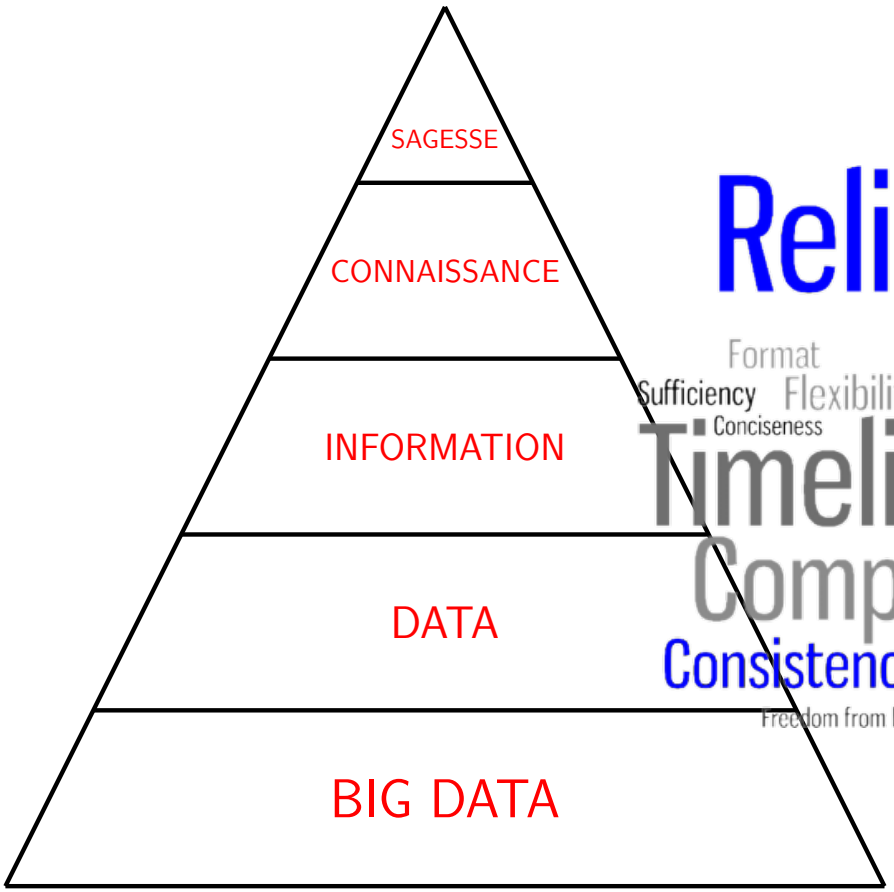
Expertise

DATA QUALITY

L'idéal



L'idéal


SAGESSE
CONNAISSANCE
INFORMATION
DATA
BIG DATA


Reliability
Accuracy
Format
Sufficiency
Flexibility
Conciseness
Timeliness
Currency
Comparability
Scope
Level-of-detail
Precision
Completeness
Effeciency
Relevance
Quantitativeness
Interpretability
Understandability
Usefulness
Usableness
Consistency
Informativeness
Clarity
Content
Importance
Freedom from bias

La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

La réalité

Souvent les données sont incohérentes, incomplètes, manquantes...

Que peut-on faire avec ces données ?

La réalité

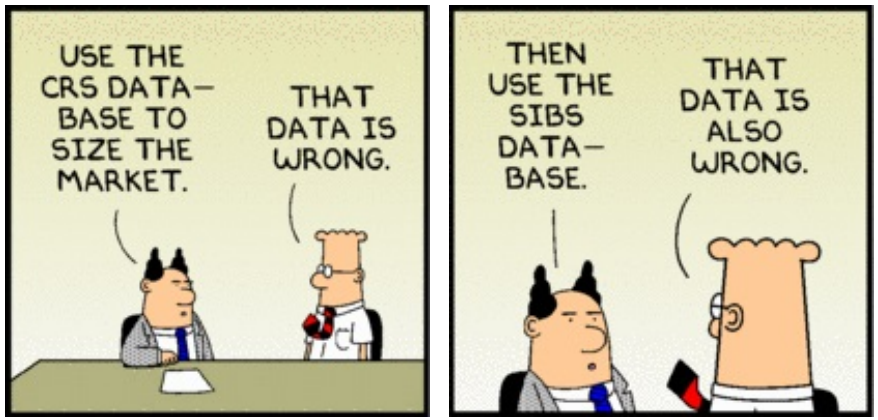
Souvent les données sont incohérentes, incomplètes, manquantes. . .

Que peut-on faire avec ces données ?

La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

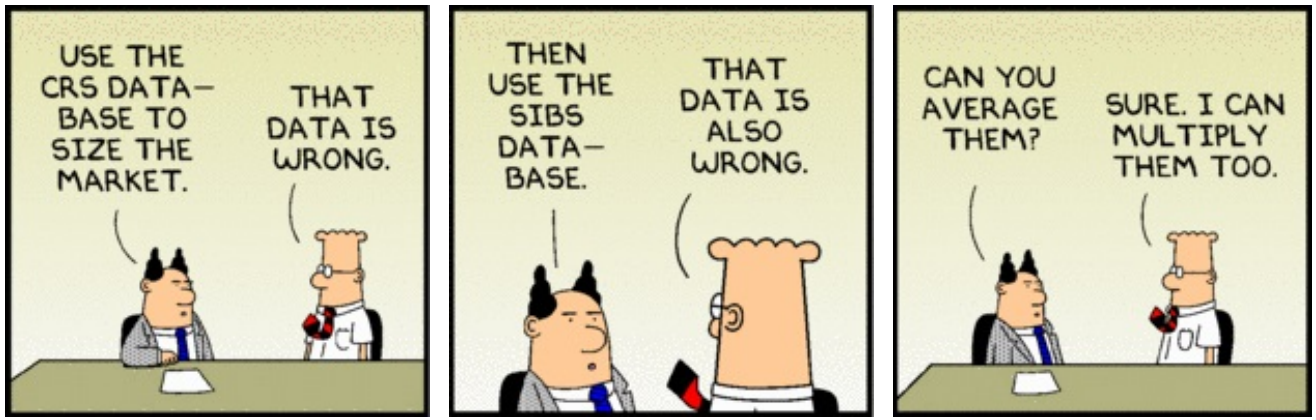
Que peut-on faire avec ces données ?



La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

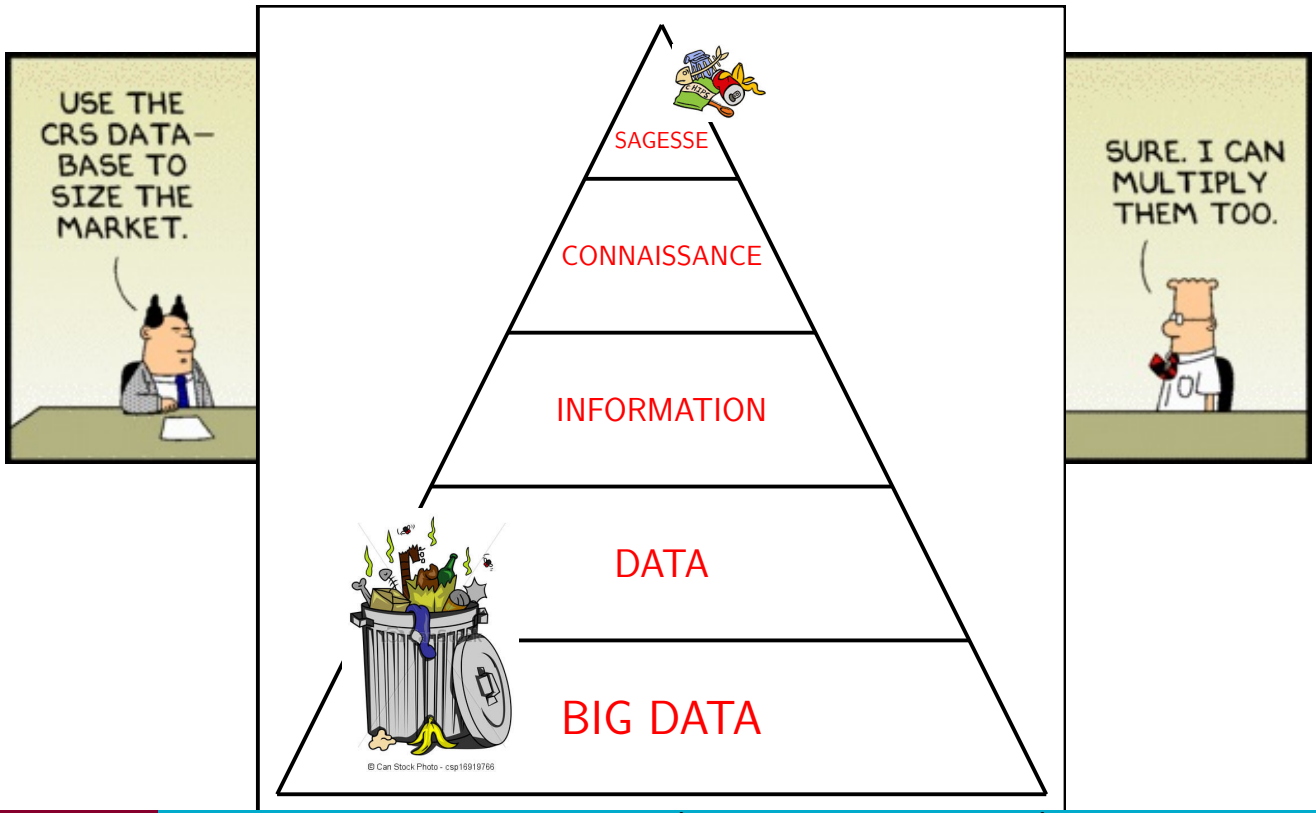
Que peut-on faire avec ces données ?



La réalité

Souvent les données sont incohérentes, incomplètes, manquantes. . .

Que peut-on faire avec ces données ?



Données imparfaites

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Données imparfaites

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Problèmes d'encodage

Données imparfaites

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Doublons

Données imparfaites

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Valeurs manquantes

Données imparfaites

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Valeurs impossibles

Notre expertise et domaine de recherche

Souvent les données sont incohérentes, incomplètes, manquantes...

Que peut-on faire avec ces données ?

Notre expertise et domaine de recherche

Souvent les données sont incohérentes, incomplètes, manquantes. . .

Que peut-on faire avec ces données ?

Notre expertise et domaine de recherche

Souvent les données sont incohérentes, incomplètes, manquantes. . .

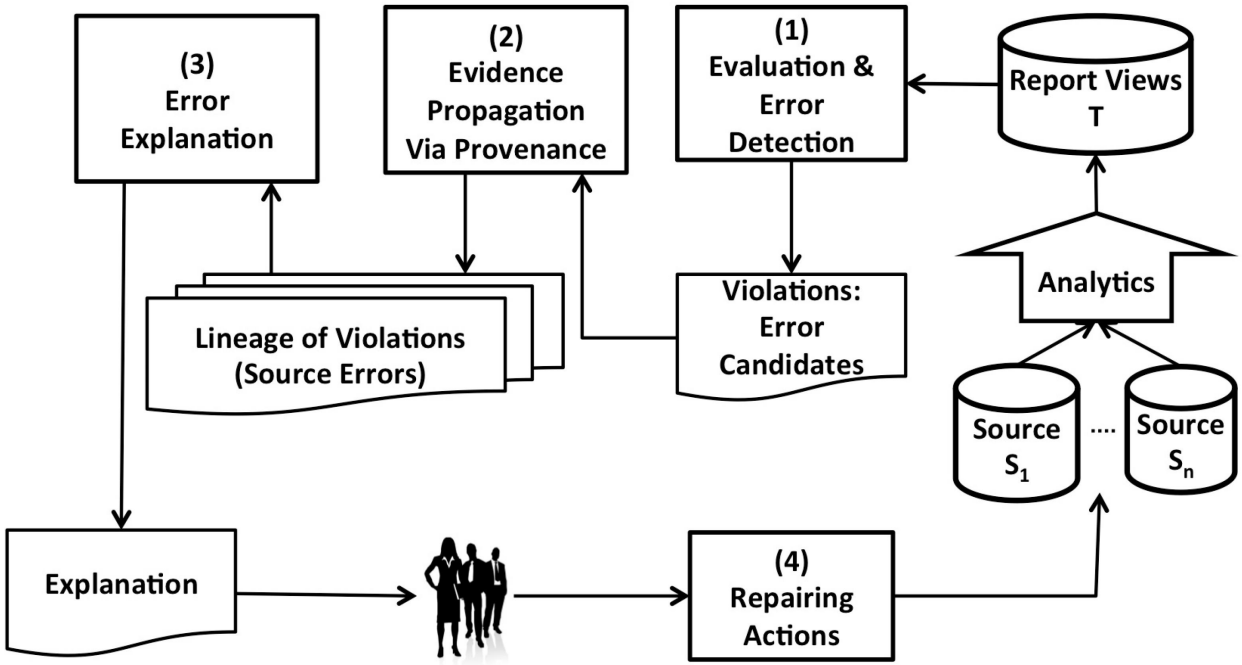
Que peut-on faire avec ces données ?

Clean Up

- Détection et suppression des doublons
- Détection et correction des erreurs
- Compléter des valeurs manquantes
- ...

Clean Up

- Détection et suppression des doublons
- Détection et correction des erreurs
- Compléter des valeurs manquantes
- ...



Source : Ihab F. Ilyas, *Effective Data Cleaning with Continuous Evaluation*

Embrace Imperfection

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes ?

1. *Combien de patients ont un groupe sanguin de A+ ?*
2. *Combien de patients ont un groupe sanguin autre que A+ ?*

Embrace Imperfection

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	♂	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaitre	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes ?

1. *Combien de patients ont un groupe sanguin de A+ ?*
2. *Combien de patients ont un groupe sanguin autre que A+ ?*

Embrace Imperfection

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	M	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaître	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes ?

1. *Combien de patients ont un groupe sanguin de A+ ?*
2. *Combien de patients ont un groupe sanguin autre que A+ ?*

Embrace Imperfection

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	M	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaître	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes ?

1. *Combien de patients ont un groupe sanguin de A+ ?*
2. *Combien de patients ont un groupe sanguin autre que A+ ?*

Embrace Imperfection

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	M	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaître	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes ?

1. *Combien de patients ont un groupe sanguin de A+ ?*
2. *Combien de patients ont un groupe sanguin autre que A+ ?*

Embrace Imperfection

Exemple

P	*PID*	*Prénom*	*Nom*	*GroupeSanguin*	*Genre*	···
	1	John	Adams	NULL	M	···
	2	Jan	Peeters	A+	M	···
	3	Jean	Lemaître	A+	M	···
	3	Jean	Lemaître	AB+	M	···

Attention

Si les données sont imparfaites et non nettoyables, nous devons repenser la façon de répondre aux requêtes.
Par exemple, comment répondre aux questions suivantes ?

1. *Combien de patients ont un groupe sanguin de A+ ?*
2. *Combien de patients ont un groupe sanguin autre que A+ ?*

Quelques réalisations

Publications

- Jef Wijsen : *Database repairing using updates. ACM Trans. Database Syst.* 30(3) : 722-768 (2005) ($>$ 170 citations)
- Paraschos Koutris, Jef Wijsen : *Consistent Query Answering for Primary Keys.* SIGMOD Record 45(1) : 15-22 (2016) (ACM SIGMOD Research Highlight)

Projets

- Développement d'outils de pilotage effectifs du réseau d'enseignement organisé par la Communauté française
- FEDER-IDEES L'Internet de demain pour développer les entreprises, l'économie et la société

Développement d'outils de pilotage effectifs du réseau d'enseignement organisé par la Communauté française

- Données opérationnelles dispersées, détaillées, de qualité variable :

COMPTAGE	Le comptage des élèves.
EDIFCf	Les infrastructures.
PERSONNEL	Le personnel de l'enseignement.
GESTELEV	Les grilles horaires et les attestations.
TEC	Les transports en commun, provenant de la Société Régionale Wallonne du Transport (SRWT) et de la Société de Transport Intercommunal de Bruxelles (STIB).
RESULTATS	Les résultats des évaluations externes certificatives (CEB et CE1D).

- Objectif décisionnel : améliorer le pilotage et faciliter la définition des actions pour améliorer la qualité de l'enseignement

FEDER-IDEES L'Internet de demain pour développer les entreprises, l'économie et la société

- À l'UMONS :
  - Data provenance, data quality, data curation
  - Cloud computing, big data
    (P. Manneback, expertise en calcul parallèle et intensif)
  - Internet of Things
    (B. Quoitin, expertise en réseaux et télécommunications)
- Collaboration étroite avec le CETIC

Merci de votre attention !