

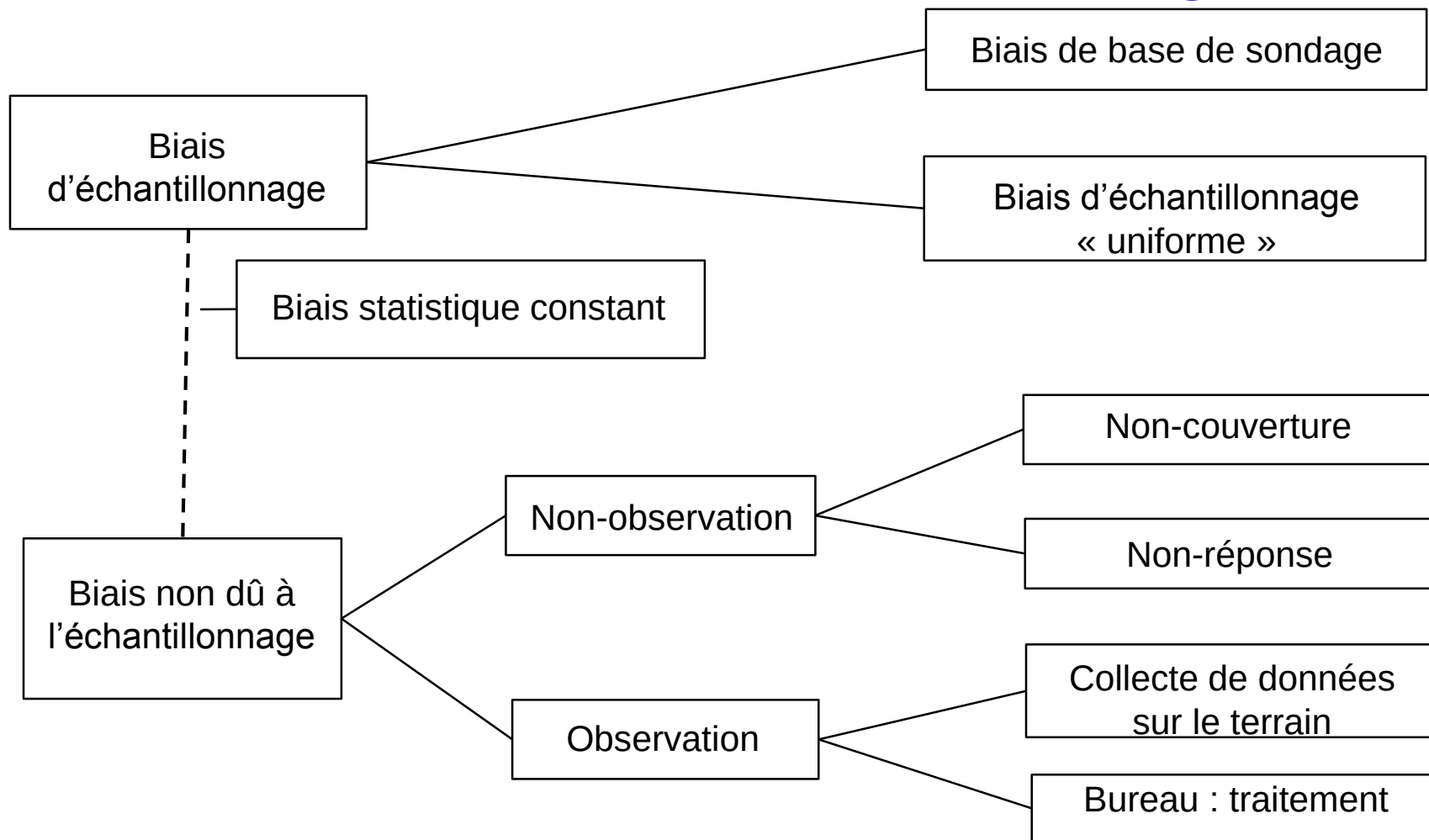
Vers un cadre de qualité pour les mélanges de données conçues et de données organiques

Robert M. Groves

Aperçu

1. Évolution des concepts de qualité statistique
2. Autres taxonomies de la qualité
3. Montée des données organiques et fin imminente des données conçues
4. Mélange de données organiques et de données conçues
5. Nécessité d'un nouveau cadre de qualité pour les données organiques
6. Modèles de mélange

Biais d'échantillonnage et biais non dû à l'échantillonnage



Erreur d'enquête totale

Mesure

Représentation

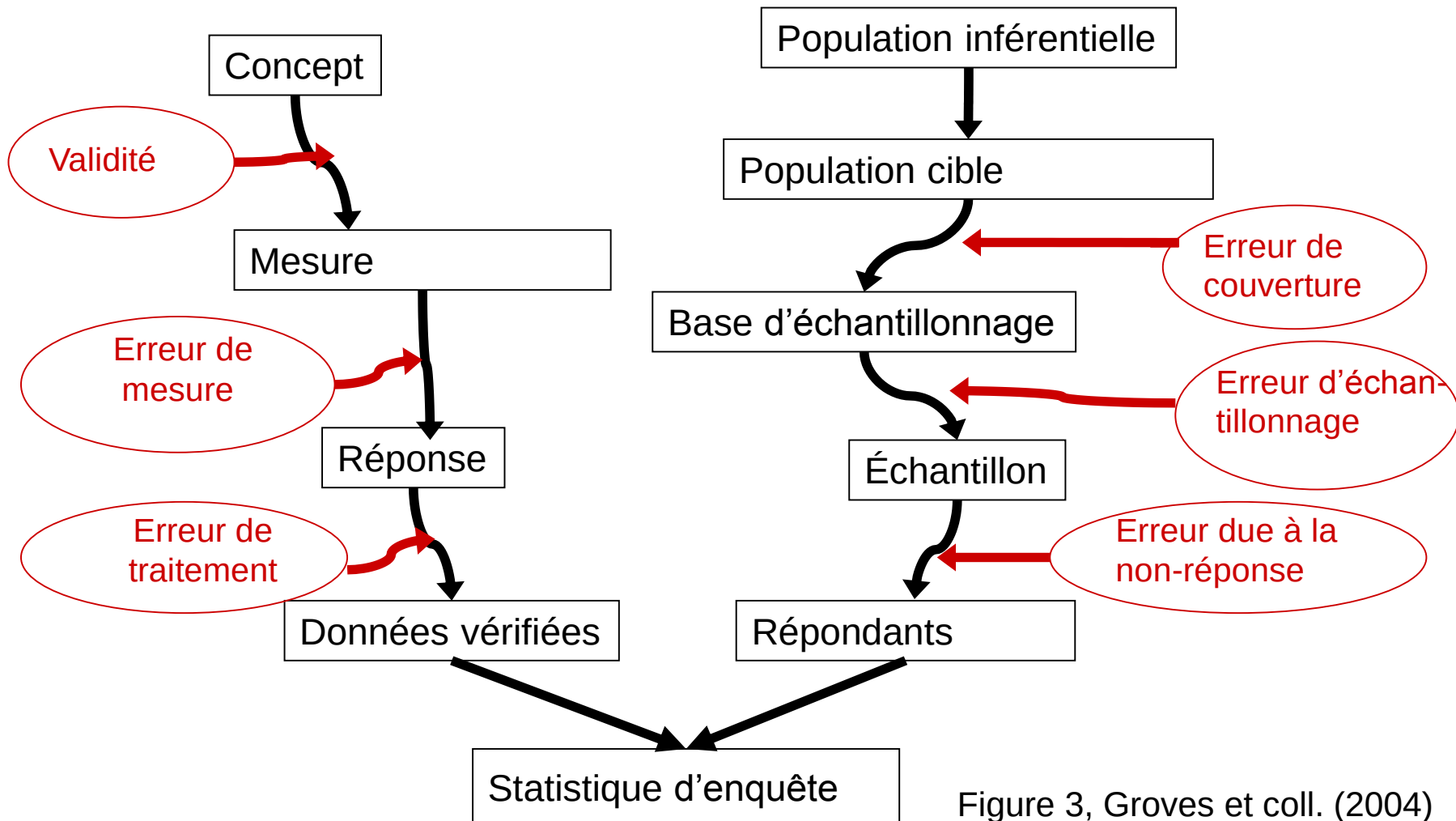


Figure 3, Groves et coll. (2004)

Adaptation des données à leur utilisation

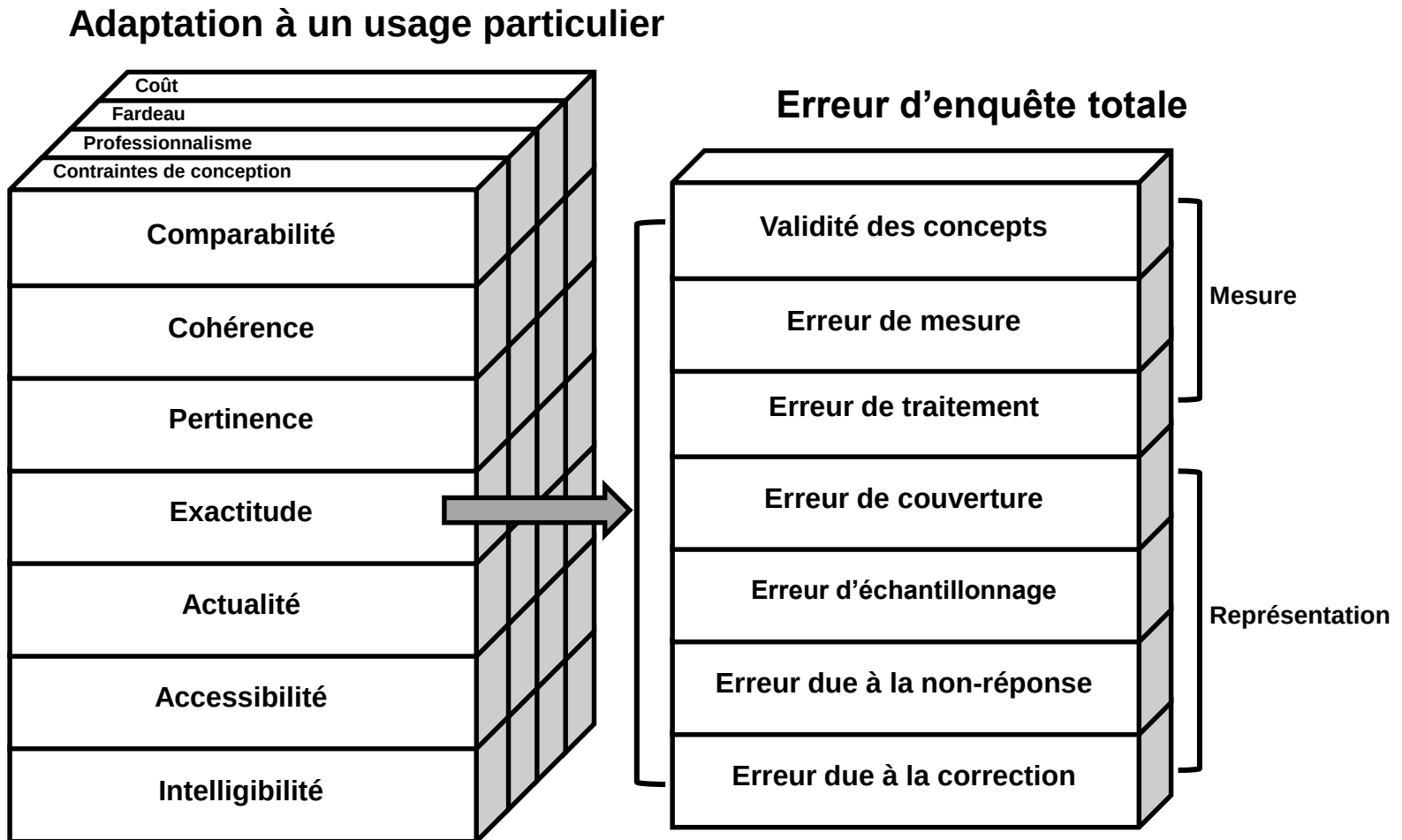
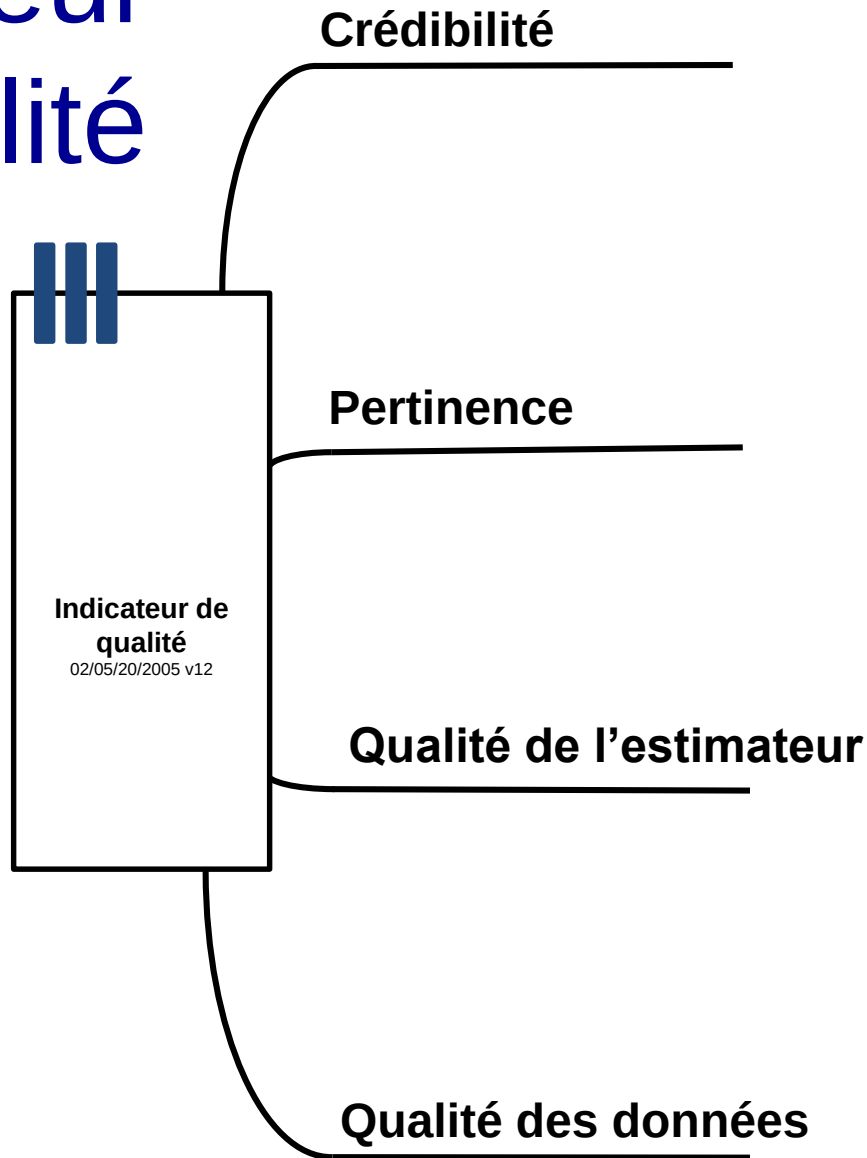
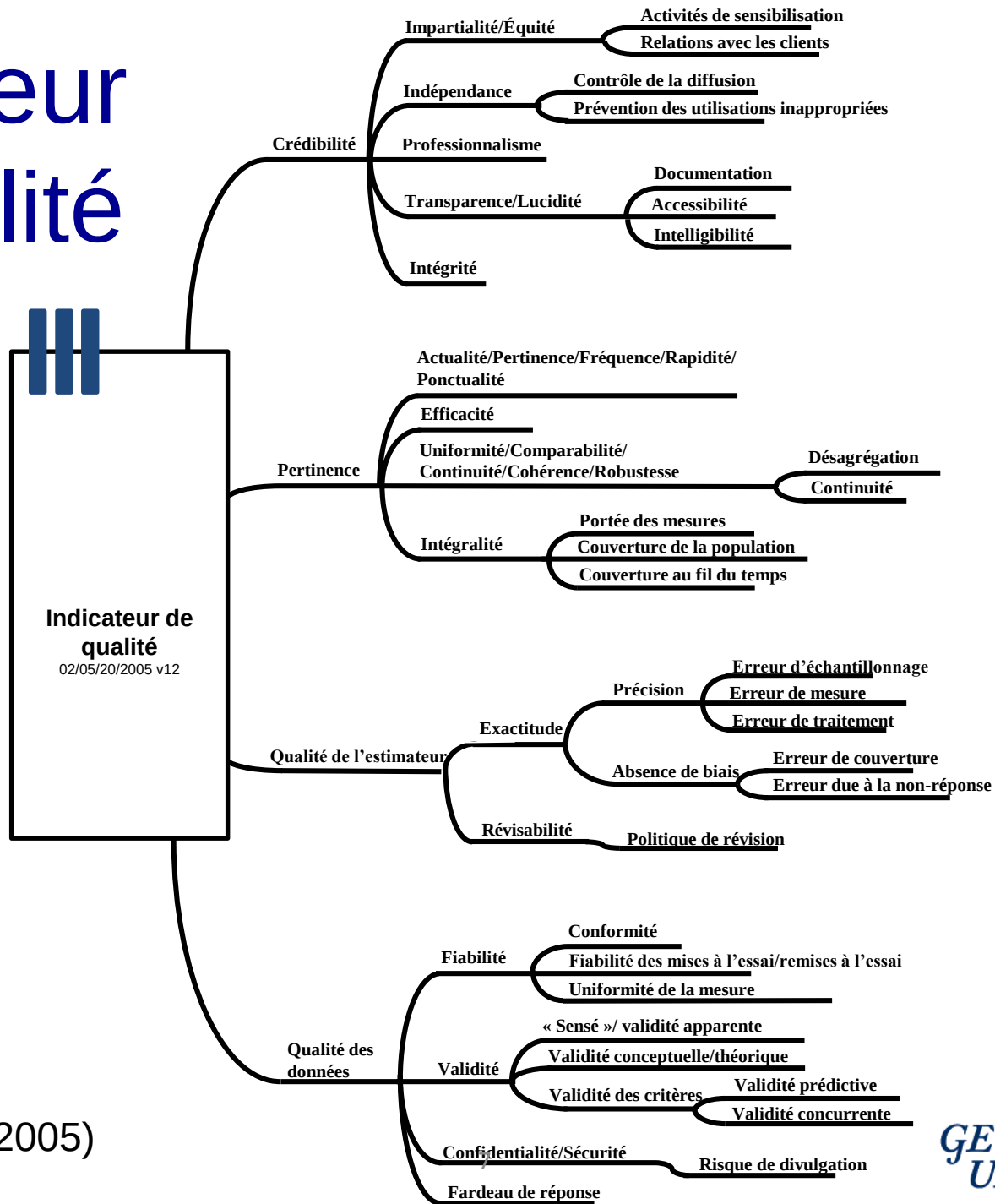


Figure 2, Hansen et coll. (2010)

Indicateur de qualité



Indicateur de qualité



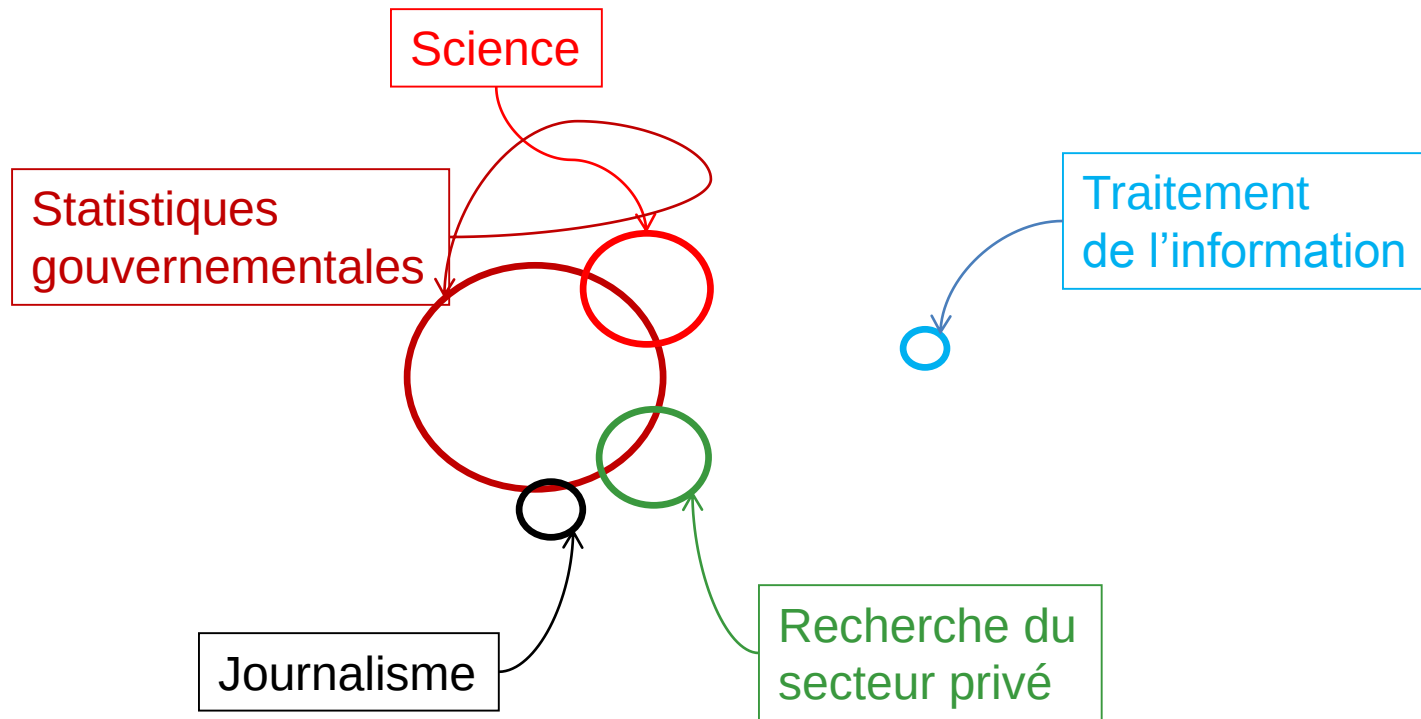
Groves (2005)

Sommaire des cadres de qualité

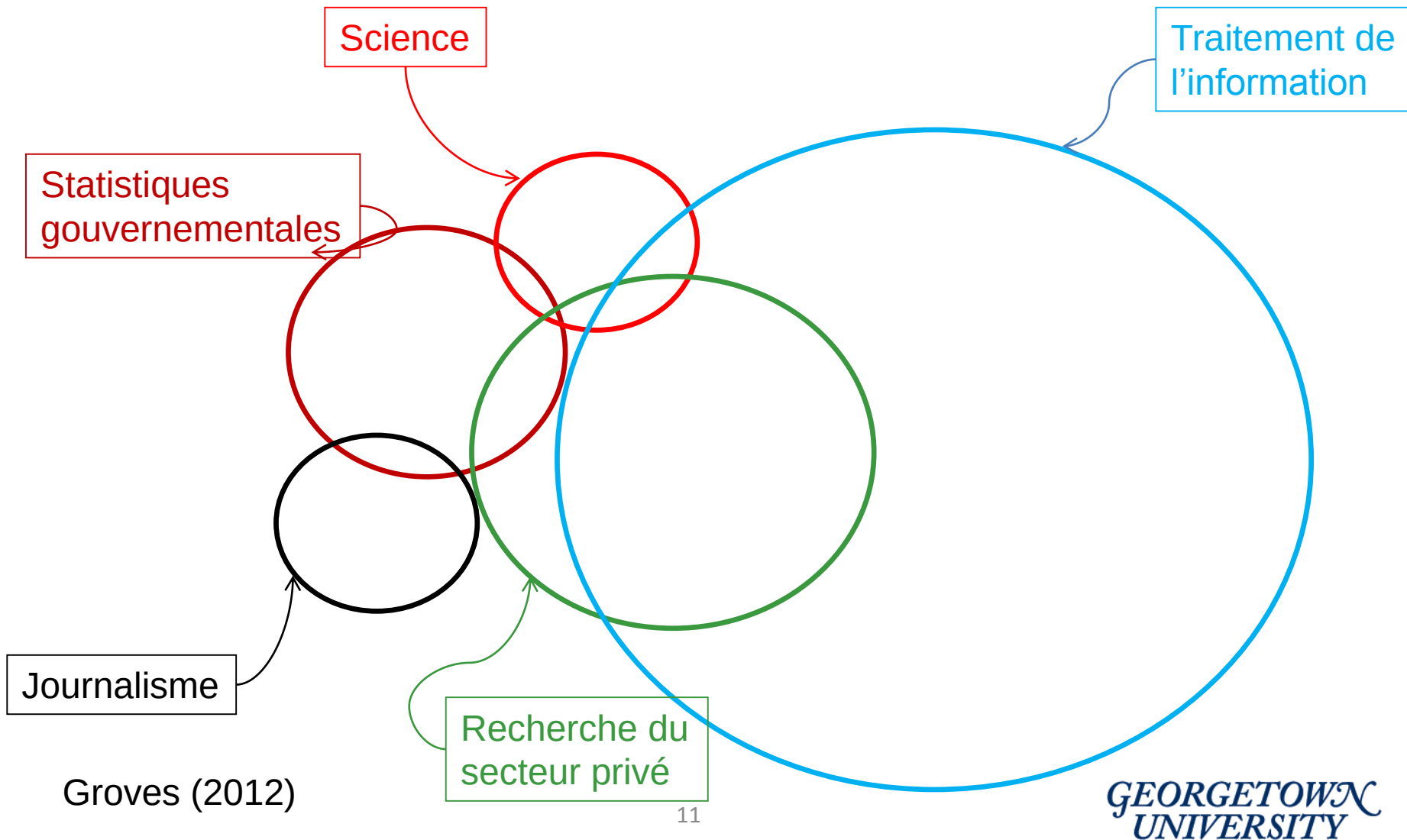
- Ils offrent un langage
- Ils offrent un cadre conceptuel, qui relie les composantes entre elles
- Ils permettent une évaluation ciblée de la conception des données
- Toutefois, ils n'ont généralement pas produit des composantes d'erreur mesurables

MONTÉE DES DONNÉES ORGANIQUES

Importance relative de la production de données numériques, autour des années 60



Importance relative de la production de données numériques, autour de 2010



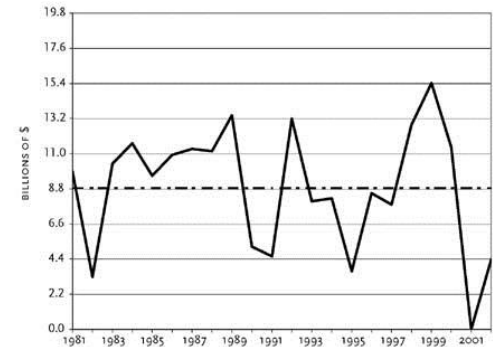
Types de données organiques

- Données de transactions (lecteur optique, carte de crédit, utilisation des services publics, service de santé, gazouillis)
- Données de communication des réseaux sociaux (Twitter, Instagram)
- Données d'agents logiciels d'appareils mobiles (données associées aux GPS)
- Données de l'Internet des objets (capteurs atmosphériques)
- Données biométriques
- Données de communication (blogues, textes, courriels)
- Données de recherche sur Internet
- Données numériques de télévision en circuit fermé
- Données tirées de sites Web

Données de carte de crédit et de scanner Nielsen



12/5/2016	8:13 AM	Giant	\$2.07
12/5/2016	8:46 AM	Giant	\$4.35
12/5/2016	9:16 AM	Safeway	\$2.31
12/5/2016	9:31 AM	Giant	\$4.92
12/5/2016	11:15 AM	Costco	\$1.31
12/5/2016	12:47 PM	Costco	\$3.11
12/5/2016	12:58 PM	Safeway	\$3.01
12/5/2016	1:38 PM	Costco	\$2.17
12/5/2016	2:24 PM	Safeway	\$1.32
12/5/2016	2:51 PM	Costco	\$3.05
12/5/2016	4:05 PM	Giant	\$4.78
12/6/2016	9:38 AM	Safeway	\$3.03
12/6/2016	10:35 AM	Giant	\$1.57
12/6/2016	1:15 PM	Giant	\$2.70
12/6/2016	2:42 PM	Safeway	\$1.57
12/7/2016	11:11 AM	Costco	\$2.80
12/7/2016	1:45 PM	Safeway	\$4.41
12/7/2016	3:41 PM	Safeway	\$2.76



Données organiques d'appareils mobiles



12/5/2016	8:13 AM	Airtel	23292 miles	24576 miles	22045 miles	8°30'26"N 13°15'01"W
12/5/2016	8:19 AM	Airtel	23291 miles	24573 miles	22045 miles	8°29'71"N 13°14'31"W
12/5/2016	8:26 AM	Airtel	23290 miles	24574 miles	22046 miles	8°29'97"N 13°15'37"W
12/5/2016	8:37 AM	Airtel	23289 miles	24575 miles	22045 miles	8°28'79"N 13°15'57"W
12/5/2016	8:41 AM	Airtel	23296 miles	24570 miles	22047 miles	8°30'25"N 13°15'48"W
12/5/2016	8:46 AM	Airtel	23294 miles	24579 miles	22043 miles	8°28'38"N 13°14'46"W
12/5/2016	8:55 AM	Airtel	23292 miles	24581 miles	22045 miles	8°28'88"N 13°14'98"W
12/5/2016	9:03 AM	Airtel	23289 miles	24569 miles	22043 miles	8°29'26"N 13°14'75"W
12/5/2016	9:12 AM	Airtel	23297 miles	24575 miles	22045 miles	8°29'53"N 13°14'75"W
12/5/2016	9:18 AM	Airtel	23288 miles	24579 miles	22043 miles	8°29'57"N 13°14'84"W
12/5/2016	9:31 AM	Airtel	23291 miles	24569 miles	22046 miles	8°30'09"N 13°14'91"W
12/5/2016	9:36 AM	Airtel	23298 miles	24573 miles	22044 miles	8°28'48"N 13°14'37"W
12/5/2016	9:40 AM	Airtel	23297 miles	24576 miles	22045 miles	8°28'18"N 13°14'88"W
12/5/2016	9:47 AM	Airtel	23297 miles	24574 miles	22042 miles	8°29'99"N 13°15'13"W
12/5/2016	9:59 AM	Airtel	23290 miles	24571 miles	22046 miles	8°29'52"N 13°14'38"W
12/5/2016	10:02 AM	Airtel	23288 miles	24570 miles	22047 miles	8°29'25"N 13°14'50"W
12/5/2016	10:09 AM	Airtel	23294 miles	24571 miles	22043 miles	8°29'88"N 13°15'68"W
12/5/2016	10:16 AM	Airtel	23293 miles	24576 miles	22042 miles	8°28'49"N 13°15'99"W

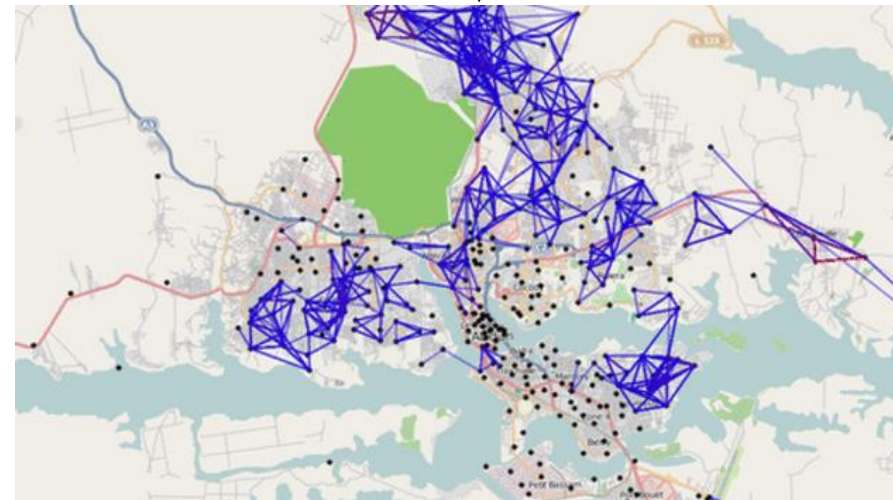


Image de trajets d'autobus de : Wakefield (2013)
<http://www.bbc.com/news/technology-22357748>



Photo

Abonnements

Abonnés

TWEETS
268

FOLLOWING
582

FOLLOWERS
1,307

LIKES
11

Contenu du message



+ Follow

Robert M. Groves

@RobertMGroves

Nom

Provost of Georgetown's main campus, professor in mathematics and statistics, and sociology. His blog can be found at blog.provost.georgetown.edu

📍 Georgetown University

🌐 provost.georgetown.edu

📅 Joined September 2012

Biographie

✉ Tweet to Robert M. Groves

Tweets

Tweets & replies

Gazouillis partagés



Robert M. Groves Retweeted
Georgetown College @GeorgetownColl · Jan 14
Way to go @dvfernandez1! #HoyaSaxa

Georgetown Univ. @Georgetown

Congratulations Daniela Fernandez (C'16) the recipient of the Christopher Benchley Ocean Youth Award! [twitter.com/StateDeptOES/s...](https://twitter.com/StateDeptOES/status/721111111111111111)

🔄 1 ❤️ 1 ⋮



Robert M. Groves @RobertMGroves · Jan 13

New post: "Seeking Faculty Input on What They Think about Georgetown" at blog.provost.georgetown.edu

🔄 1 ❤️ 1 ⋮



Robert M. Groves Retweeted

GU Security Studies @GeorgetownCSS · Jan 13

Dr. @hoffman_bruce's book Anonymous Soldiers was named the Book of the Year in the 2015 National Jewish Book Awards! [tabletmag.com/scroll/196601/...](http://tabletmag.com/scroll/196601/)

🔄 3 ❤️ 1 ⋮



Robert M. Groves Retweeted

GU Medical Center @gumedcenter · Jan 11

Learn how a @GeorgetownLombardi researcher contributed to the science behind #breastcancer screening recommendations bit.ly/1Q0VMDD

🔄 3 ❤️ 1 ⋮

Gazouillis

Who to follow · Refresh · View all



DCist @DCist

+ Follow

Trends · Change

#twitterdown
106K Tweets

#HowIStaySaneInMyWorkspace
Just started trending

#NationalPopcornDay
Trending for 5 hours now

Ice Prince
12K Tweets

#TravelTuesday
Trending for 5 hours now

#GrantsNotDebt
Trending for 2 hours now

Edgar Allan Poe
10.1K Tweets

Pete Rose
Started trending in the last hour

Adapté de
Daas et coll.
(2015)

Abonnements



Robert M. Groves

@RobertMGroves

Provost of Georgetown's main campus, professor in mathematics and statistics, and sociology. His blog can be found at blog.provost.georgetown.edu

📍 Georgetown University

🌐 provost.georgetown.edu

📅 Joined September 2012

✉ Tweet to Robert M. Groves

TWEETS
268

FOLLOWING
582

FOLLOWERS
1,307

LIKES
11



+ Follow



The Caravel
@TheCaravelGU

Georgetown University's one and only news source dedicated to international affairs. Set sail with us and get a uniquely student-driven perspective...

⚙️ + Follow



GUClassof2002
@GUClassof2002

The official Twitverse home for Georgetown University's Class of 2002 alumni. Gearing up for 10 year reunion May 31, 2012-June 2, 2012. Hoya Saxa!

⚙️ + Follow



GUclubStLouis
@GUclubStLouis

+ Follow

Who to follow · Refresh · View all



Capital Weather Gang

+ Follow



POLITICO

+ Follow



Washington Post

+ Follow

Find friends

Trends · Change

#twitterdown



GU Art Gallery
@GUartgallery

We exhibit pro artists, students, and faculty across the Georgetown University campus. Visit our HQ, the Spagnuolo Art Gallery, at 36th and...

⚙️ + Follow



Biology @ Georgetown
@GUBiology

We are the Biology department @Georgetown University. Our programs are designed to provide the finest education in biological principles...

⚙️ + Follow



GUOCAF
@guocaf

The Office of Campus Activity Facilities strives to provide the GU community with facilities & services for enlightenment & entertainment. ...

⚙️ + Follow



2015 Black Alumni Summit
Oct. 23-25, 2015 in Washington, DC

The Georgetown University Black Alumni Summit is a gathering planned by and for the black undergraduate students of Georgetown University. In collaboration with the Office of Advancement, our intention is to bring several hundred black alumni back to the Hilltop to...



GUWLI
@GUWLI

GEORGETOWN UNIVERSITY
WOMEN'S LEADERSHIP INSTITUTE



GUWLI
@GUWLI

GEORGETOWN UNIVERSITY
WOMEN'S LEADERSHIP INSTITUTE



Abonnés

TWEETS
268

FOLLOWING
582

FOLLOWERS
1,307

LIKES
11



Follow

Robert M. Groves

@RobertMGroves

Provost of Georgetown's main campus, professor in mathematics and statistics, and sociology. His blog can be found at blog.provost.georgetown.edu

📍 Georgetown University

🌐 provost.georgetown.edu

📅 Joined September 2012

✉ Tweet to Robert M. Groves

Who to follow · Refresh · View all

Common \$ense
@GUcommonsense
Georgetown University's premier financial literacy program created and led by students. Visit us in the Financial Aid Office, G-19 Healy Hall.

Follow

Georgetown Ministry
@gmcgt
Dedicated to ending homelessness one person at a time. Like us on Facebook @ Georgetown Ministry Center

Follow

HoyasLead
@HoyasLead
Hoya Athletics office of Student-Athlete Leadership and development

Follow

Georgetown Physics
@GUPhysics

Follow

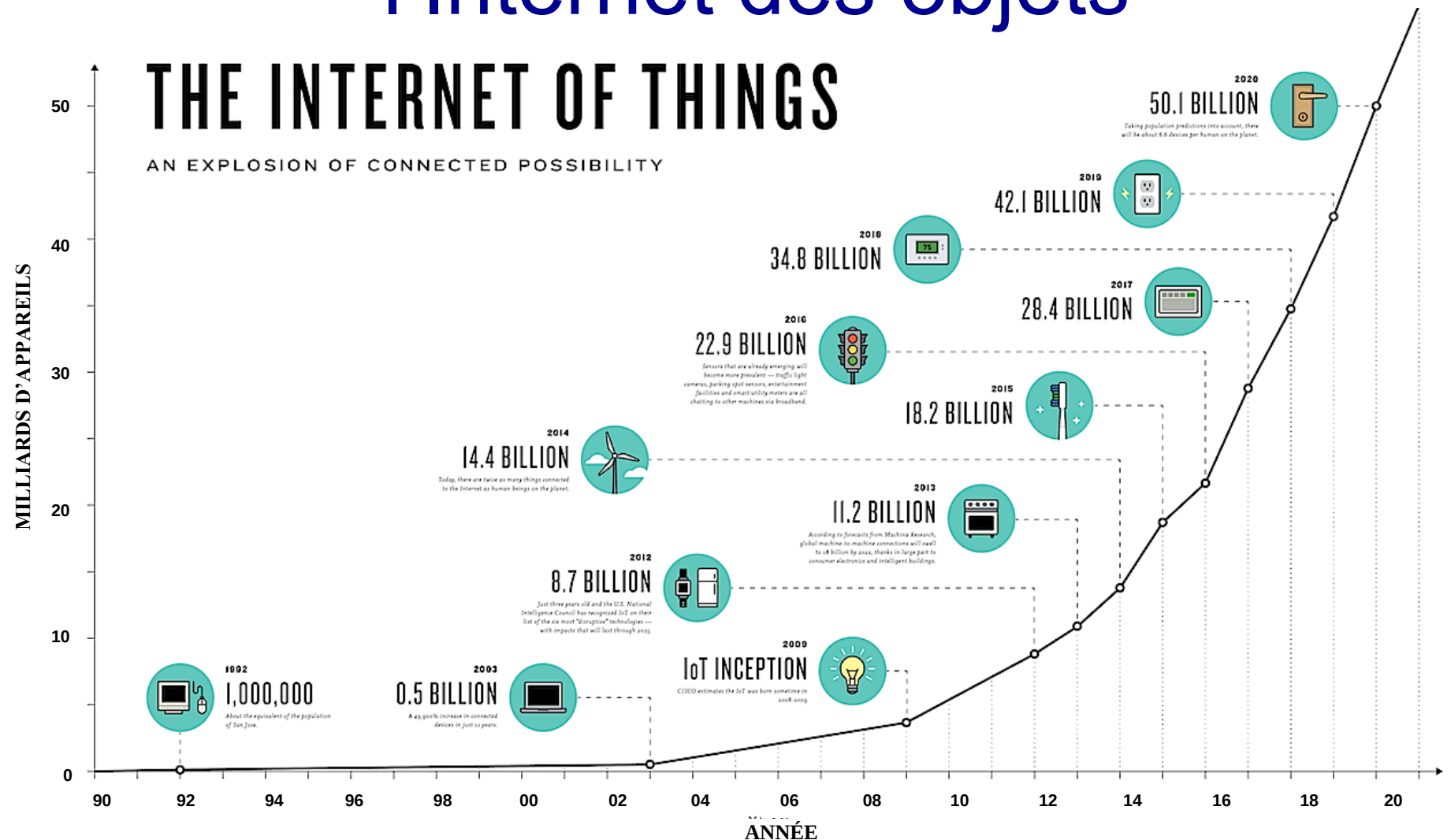
UM SurveyMethodology
@MichPSM
The University of Michigan's Program in Survey Methodology is a program where students learn the science of surveys.

Follow

McCourt E&E
@McCourtEE
The Georgetown McCourt Energy & Environmental Policy group engages students, academics, and practitioners on policy issues related to energy &...

Follow

Croissance exponentielle de l'Internet des objets



De Clerck (2015), <http://www.i-scoop.eu/internet-of-things/>

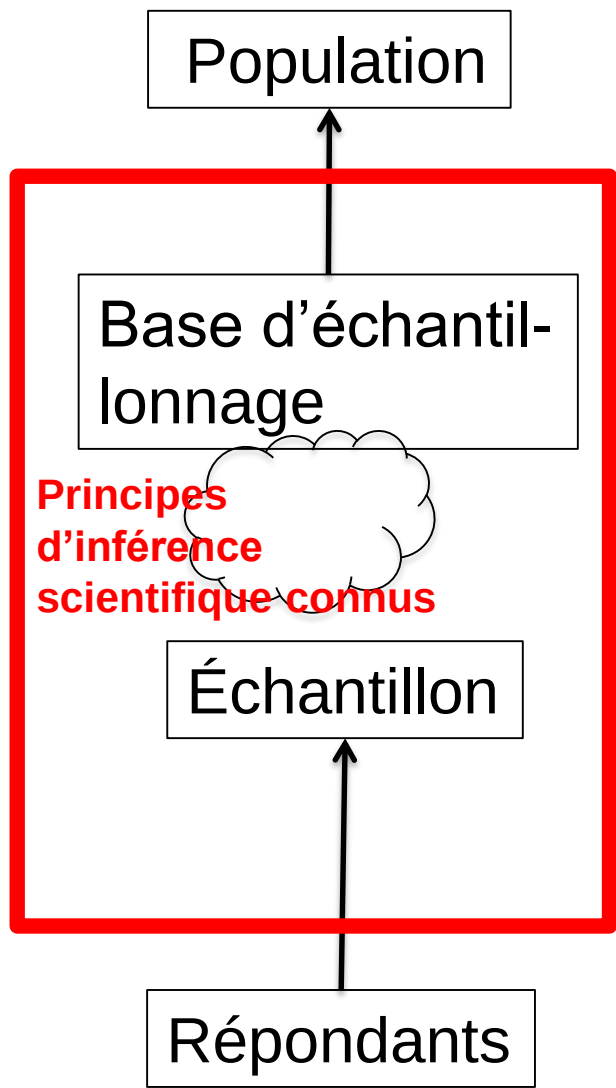
Points forts des données organiques

- Elles mesurent les comportements nouveaux
- La fréquence est presque en temps réel
- Grande granularité spatiale
- Capacité de mesurer des réseaux

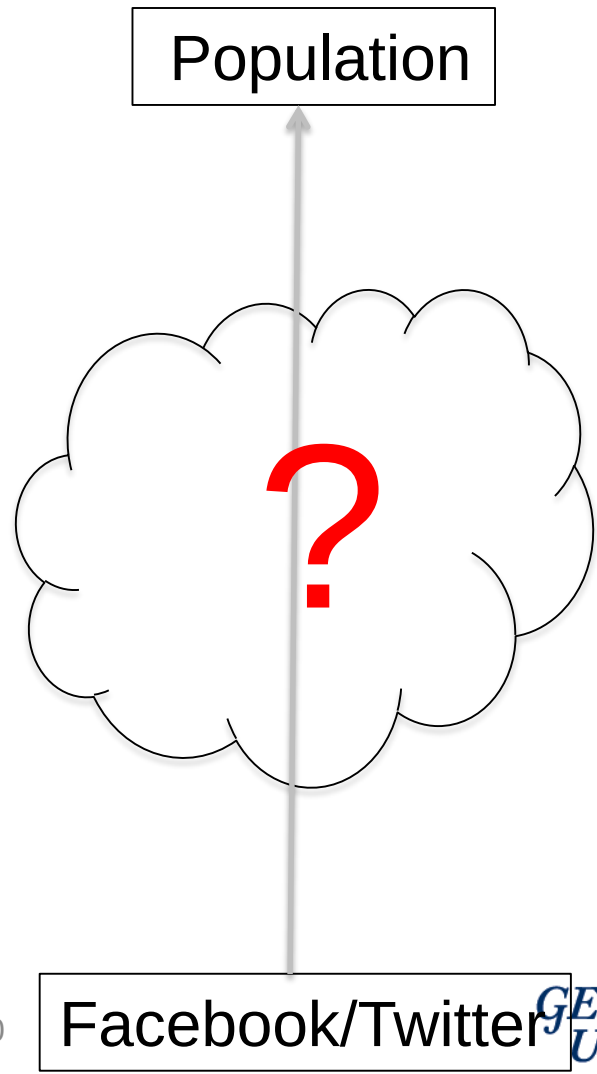
Faiblesses des données organiques

- Ce ne sont pas tous les membres de la population qui sont couverts
- Non multidimensionnelles; peu de variables
- L'identification des fournisseurs de données est difficile
- La structure des données ne favorise pas l'analyse
- L'accès et le contenu sont contrôlés par les détenteurs des données

Le paradigme de l'enquête sur échantillon



Le paradigme des « mégadonnées »



NÉCESSITÉ D'UN CADRE DE QUALITÉ POUR LES DONNÉES ORGANIQUES

Quatre exemples de problèmes de données organiques nécessitant un cadre de qualité

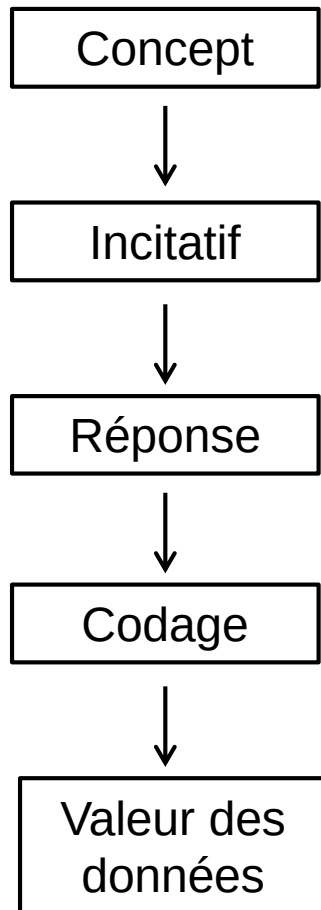
1. Données organiques générées sans stimulus externe connu
2. Données organiques stimulées par un processus unique qui n'est pas connu du statisticien
3. Données organiques qui sont elles-mêmes des produits de modèles
4. Données organiques tirées de textes

Absence d'un incitatif de mesure

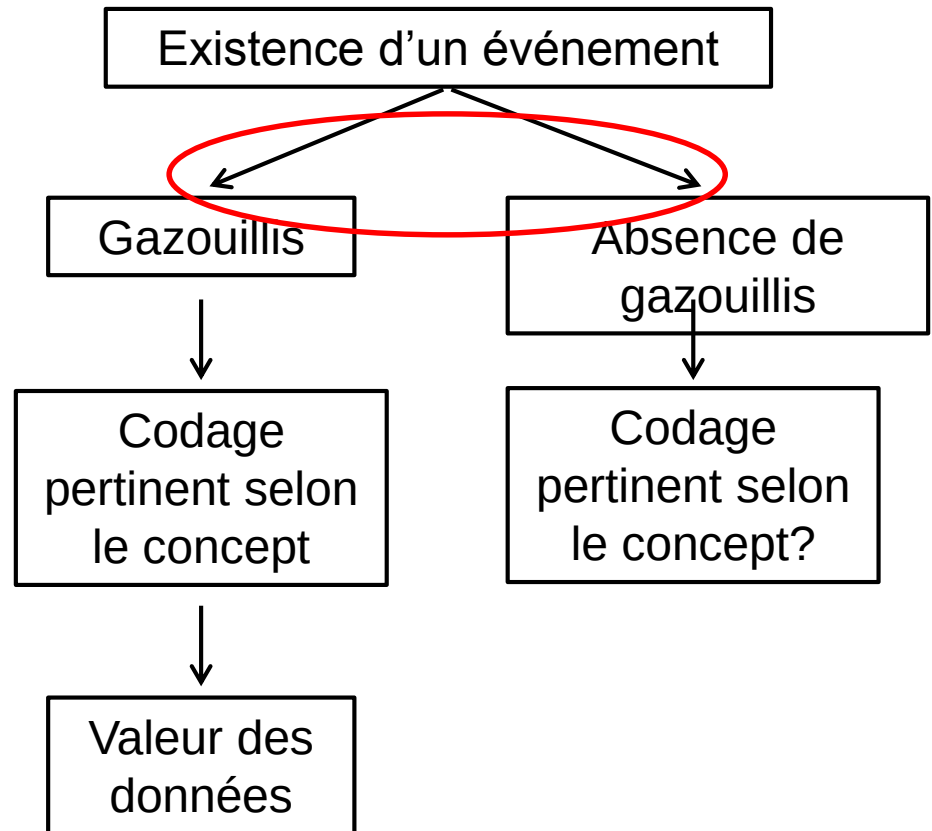
- De nombreuses sources de données organiques ne découlent pas d'un incitatif uniforme appliqué à toutes les unités de mesure
- Exemples :
 - Gazouillis
 - Blogues

Absence d'un incitatif de mesure

Données conçues

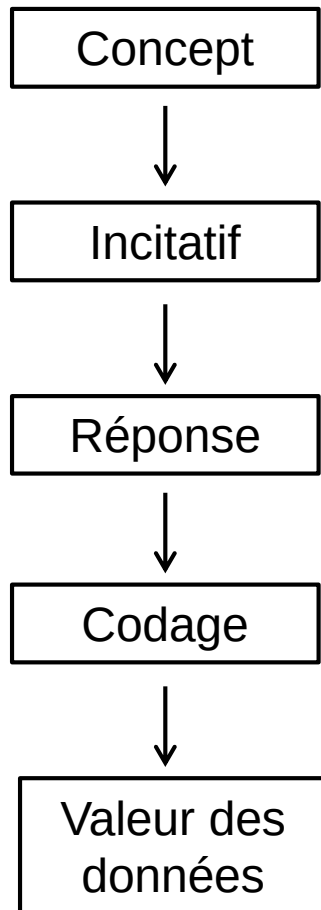


Données organiques

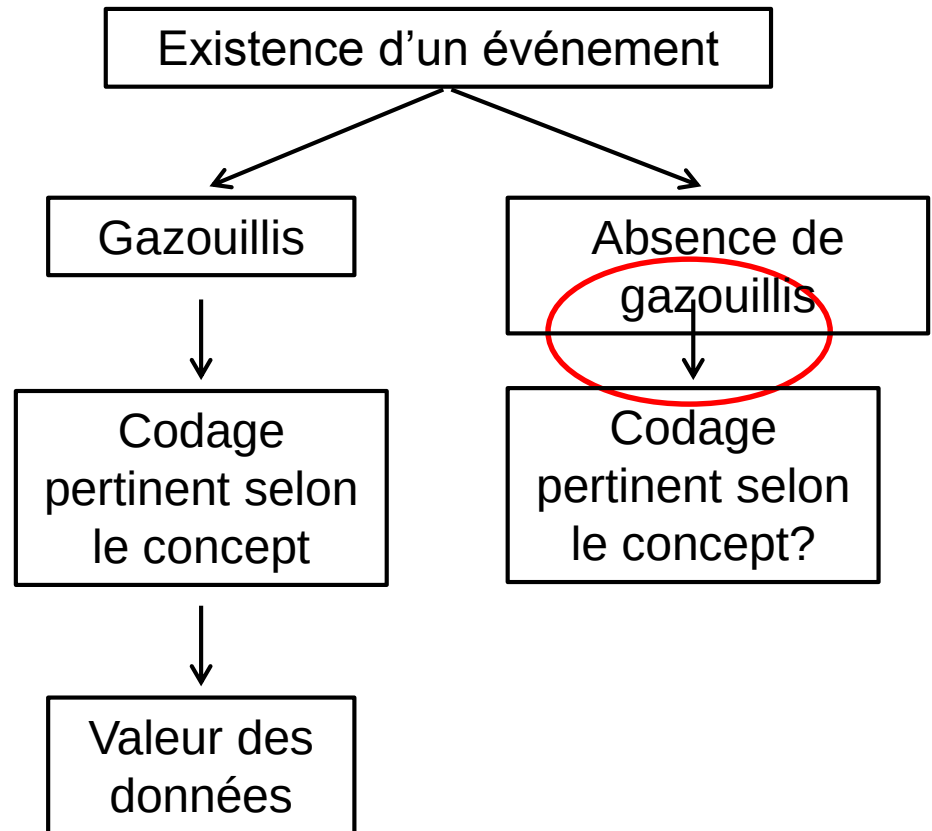


Incitatif de mesure connu/inconnu

Données conçues



Données organiques

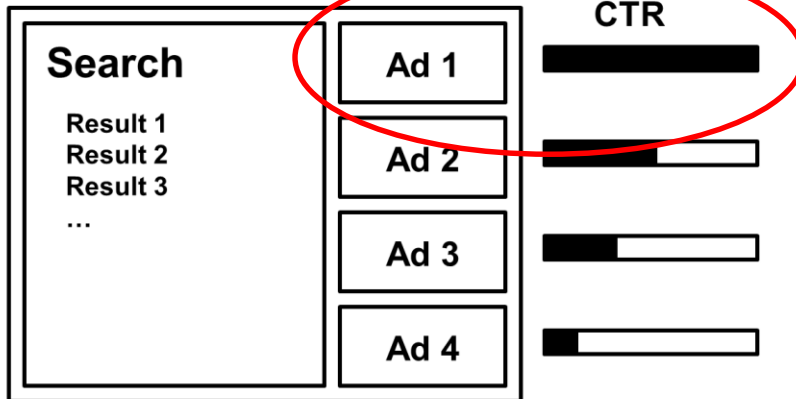


Données organiques avec caractéristiques d'erreur de mesure intermédiaires

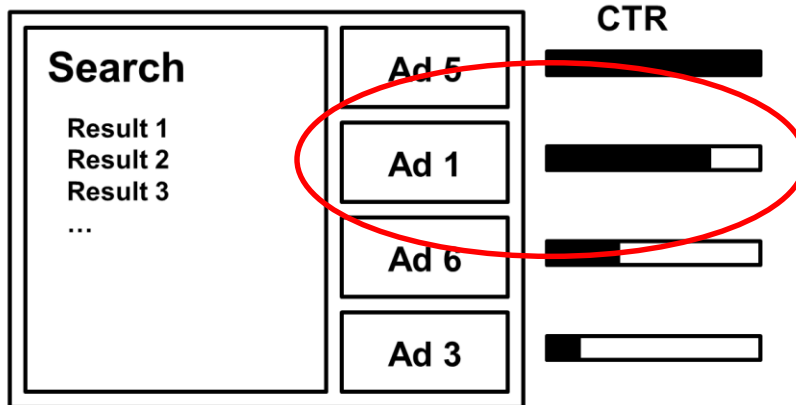
- Les données organiques qui sont créées au moyen de processus comportant plusieurs étapes ont des répercussions sur les propriétés de l'erreur de mesure
- Exemple :
 - Données à cliquer dans des annonces de pages Web qui ont été affichées en fonction de modèles prédictifs de l'intérêt des utilisateurs

L'endroit où se trouve l'annonce a des répercussions sur le taux de clics

User 1



User 2



Adapté de la figure 1,
Richardson et coll.
(2007)

Données générées à partir de modèles reposant sur des données non structurées

- Certaines données organiques proviennent d'un codage fondé sur un modèle
- Exemple :
 - Reconnaissance faciale
 - Données sur la circulation provenant d'un système de télévision en circuit fermé
 - GPS

Estimations de l'erreur de mesure inclues dans les éléments de données

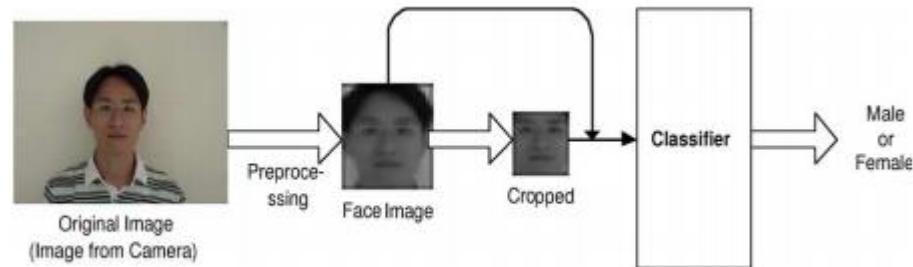


Fig. 1. The process of appearance-based gender classification.

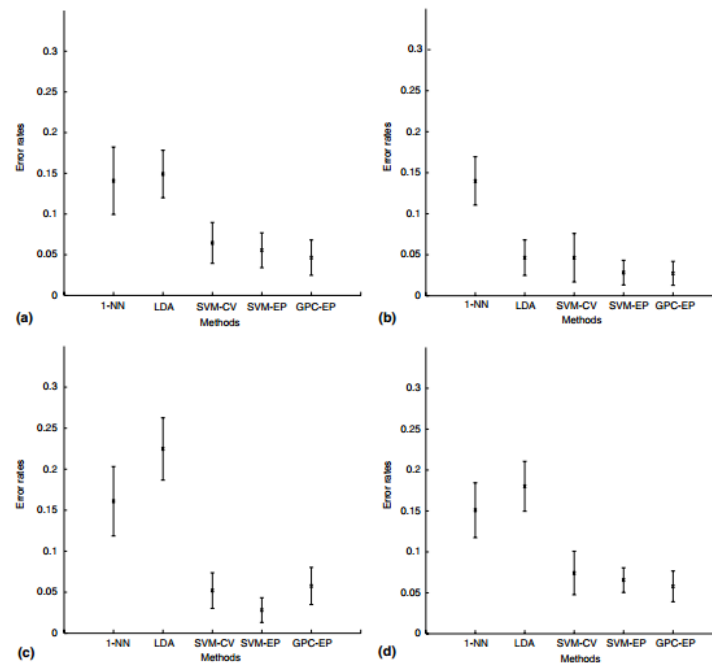


Fig. 5. Classification error rates of various methods for four kinds of gender classification data sets (PF01 DB): (a) data set P-I, (b) data set P-II, (c) data set P-III, (d) data set P-IV.

Figure 1 et Figure 5,
Kim et coll. (2006)

Transposition de textes en chiffres

- Données organiques fondées sur le codage de données textuelles
- Exemple :
 - Codage de textes de Twitter
 - Codage de blogues
 - Codage de LinkedIn

Codage dans une enquête-échantillon

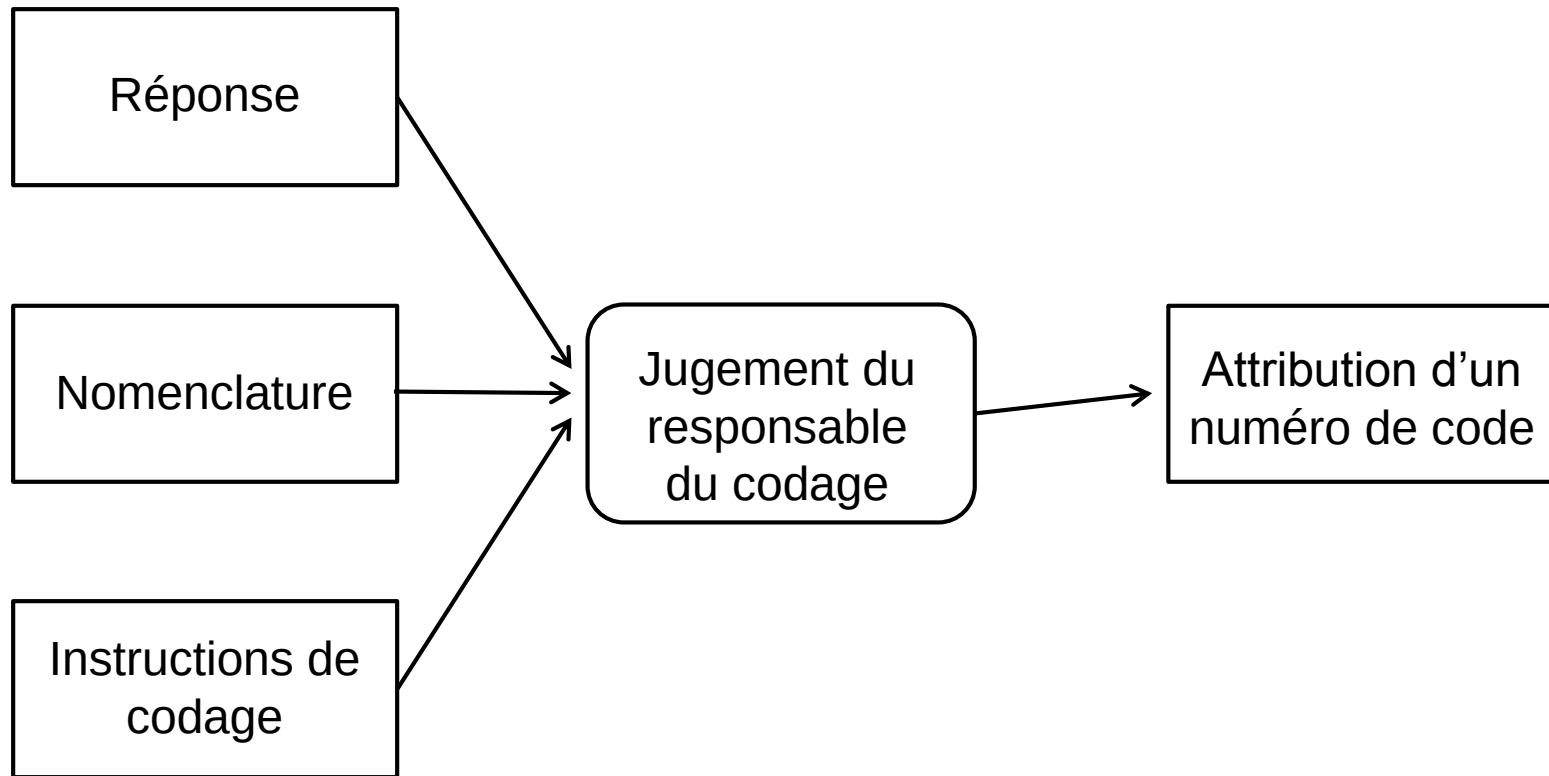


Figure 7.1, Biemer et Lyberg (2003)

Erreurs de codage découlant de la polysémie lexicale

Voici un exemple de la polysémie du mot *back* en anglais.

I don't want to get back at him for it though.

We are so backed up at work I get overtime.

I wish my baby would get back to normal.

I have to get back to work one of these days.

My back hurts all the time.

I have a bad back.

The ceiling in the back of the house is falling in.

I called her back and said I was sorry.

I have to go back to the store and get food.

This morning the cars were so backed up for miles and I was late to work.

I never had these problems back in Alabama.

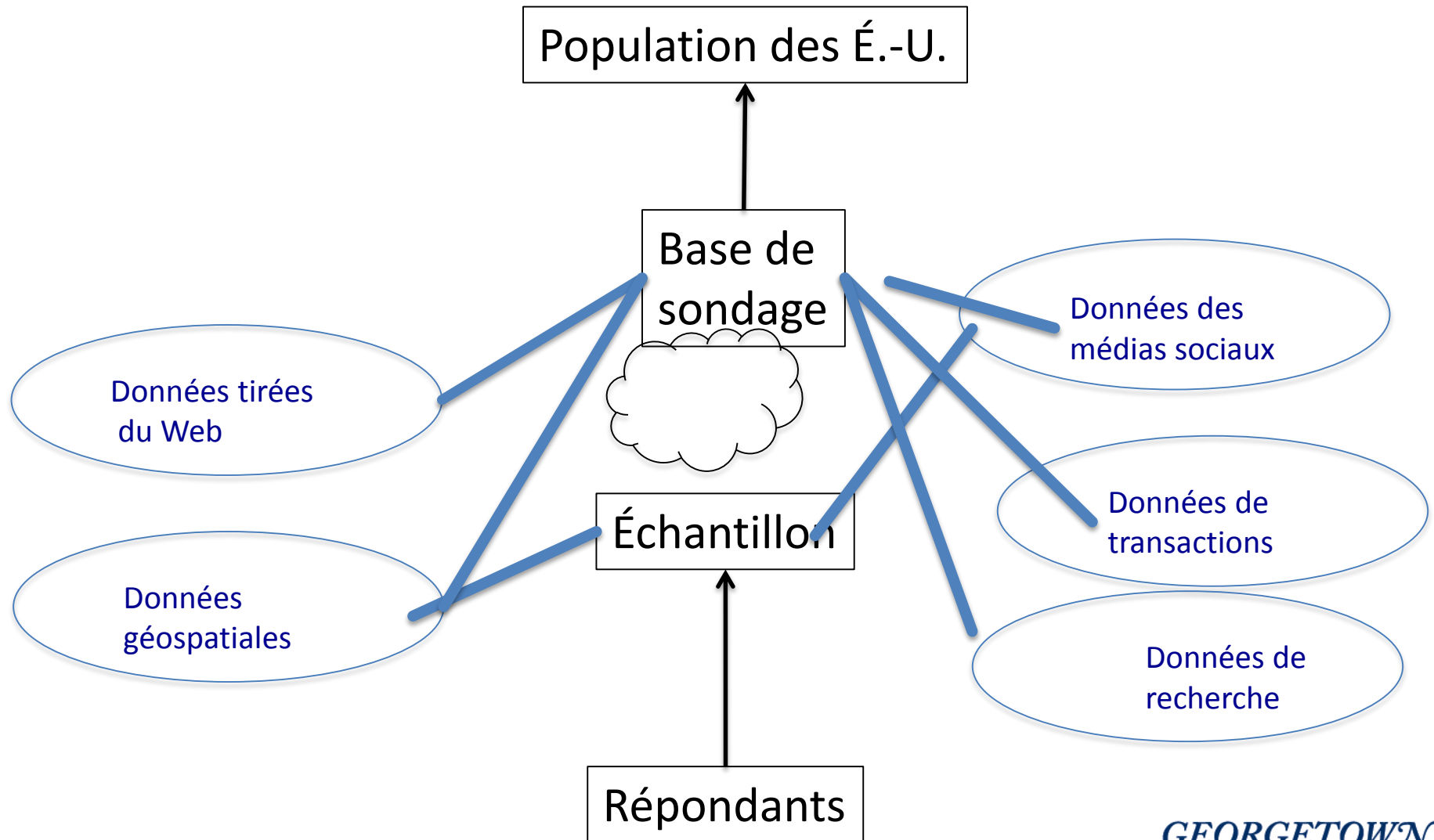
Adapté de Frisbie et Sudman (1968)

MÉLANGE DE DONNÉES ORGANIQUES ET DE DONNÉES CONÇUES

Le mélange de données organiques et de données d'enquête permet de combiner la granularité temporelle et spatiale des données organiques et le pouvoir d'inférence des données d'enquête

- Les modèles statistiques jouent un rôle clé
- Des covariables partagées entre les ensembles de données sont souhaitables
- Il est essentiel de mettre l'accent sur les erreurs de couverture et de mesure

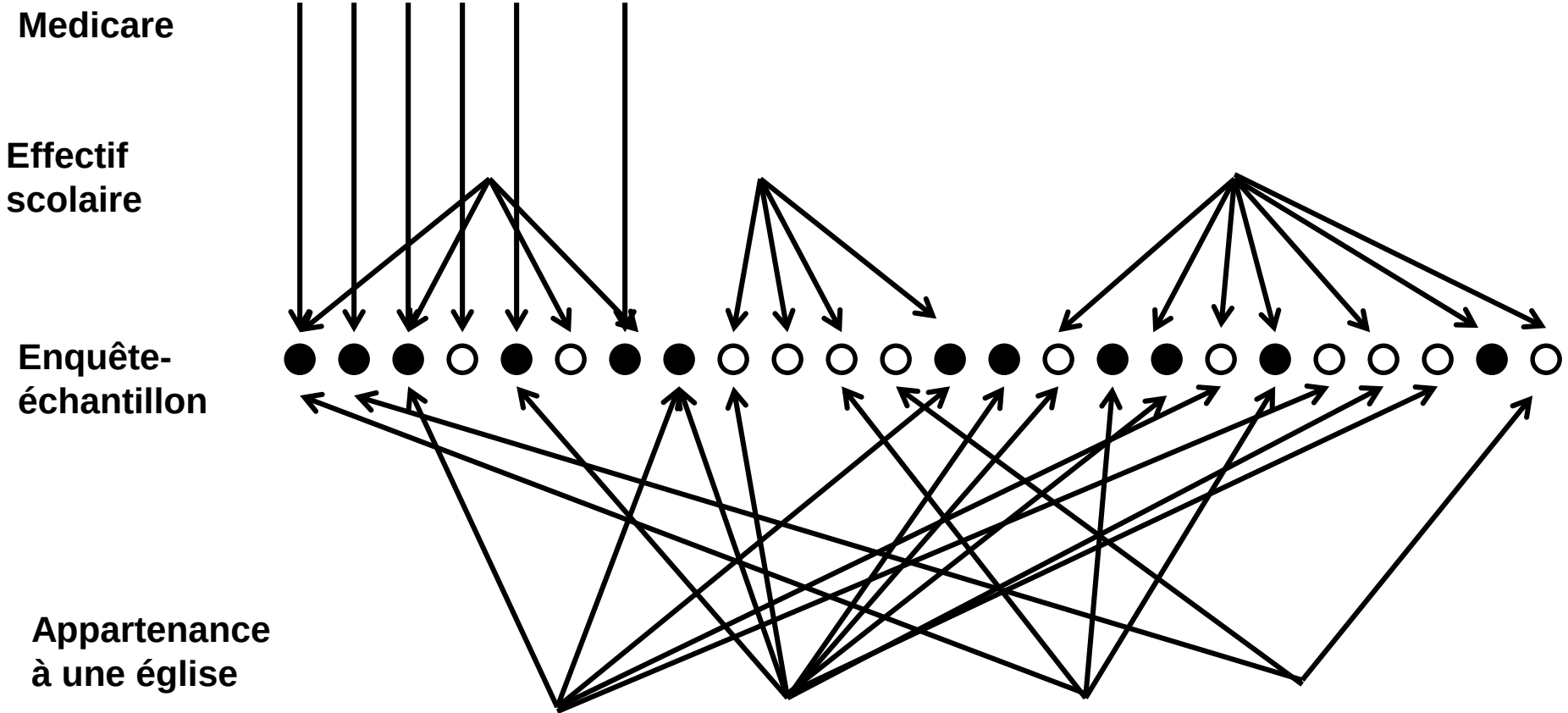
Le paradigme regroupé



Données organiques comme source de covariables

- Problème : estimation de statistiques pour de nombreux petits domaines
- Ressources de données :
 - Enquête nationale avec données conçues à partir d'un échantillon d'UPE
 - Données de recensement les plus récentes
 - Données organiques comme covariables de la principale quantité à estimer
- Exemple : Marchetti et coll. (2015) pour le revenu à partir de l'utilisation du GPS au moyen de téléphones mobiles comme covariable

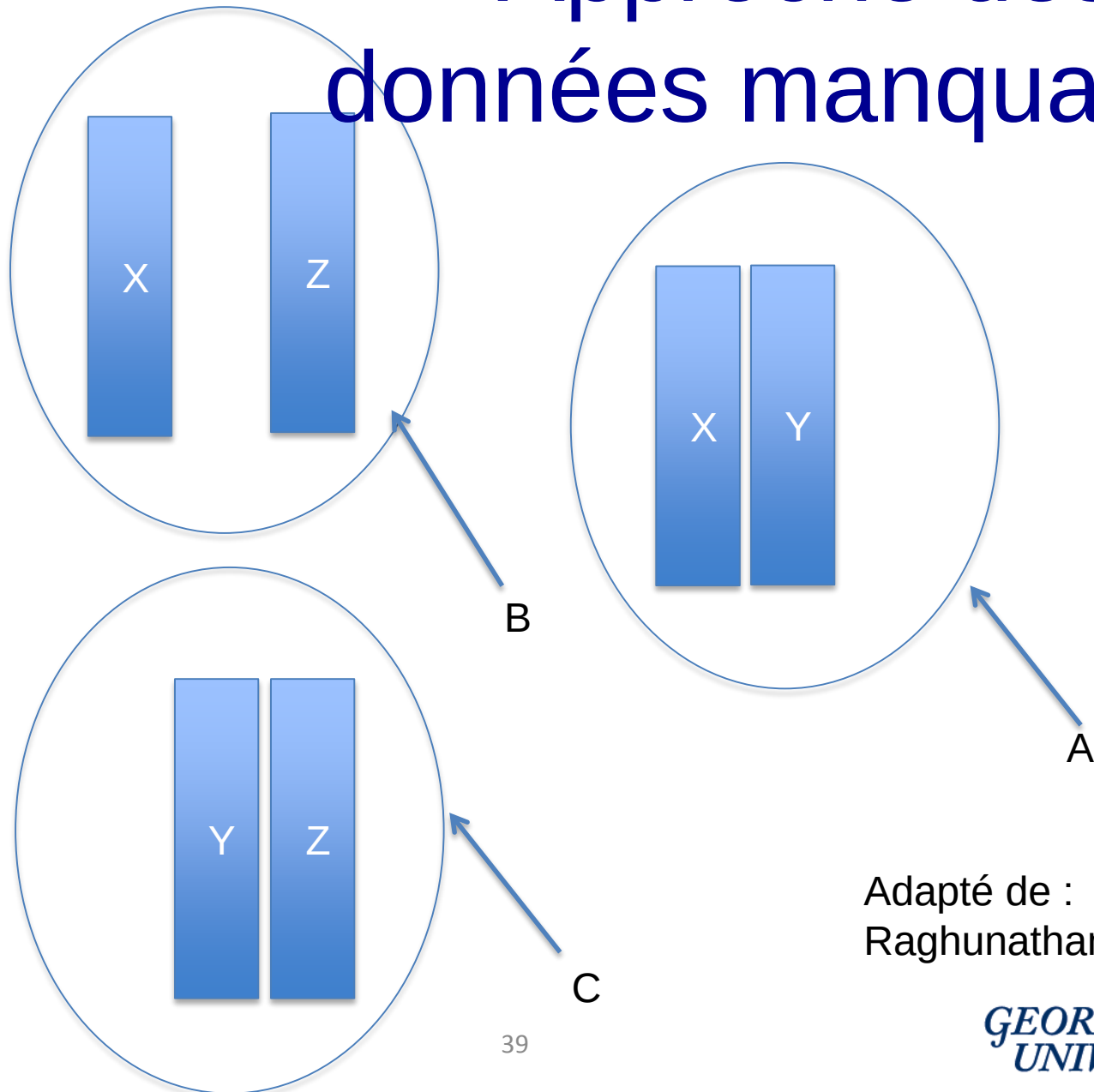
Estimation sur petits domaines



Mélange dans un cadre de données manquantes

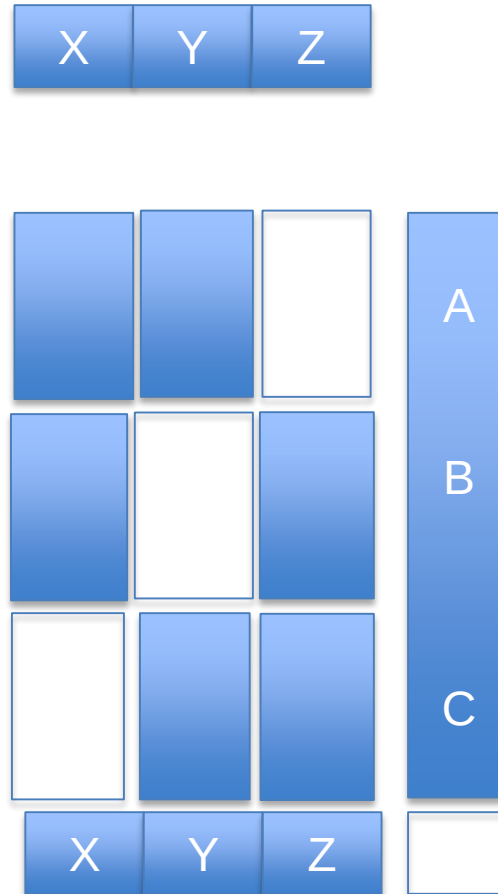
- Problème : estimation améliorée empruntant de l'information de plusieurs sources de données
- Ressources de données :
 - Enquête nationale avec données conçues pour certaines mesures (X,Y)
 - Sources de données organiques avec d'autres variables (X,Z), (Y,Z)
- Exemple : Schenker et Raghunathan et coll. (2006, 2007, 2010)

Approche des données manquantes



Adapté de :
Raghunathan (2015)

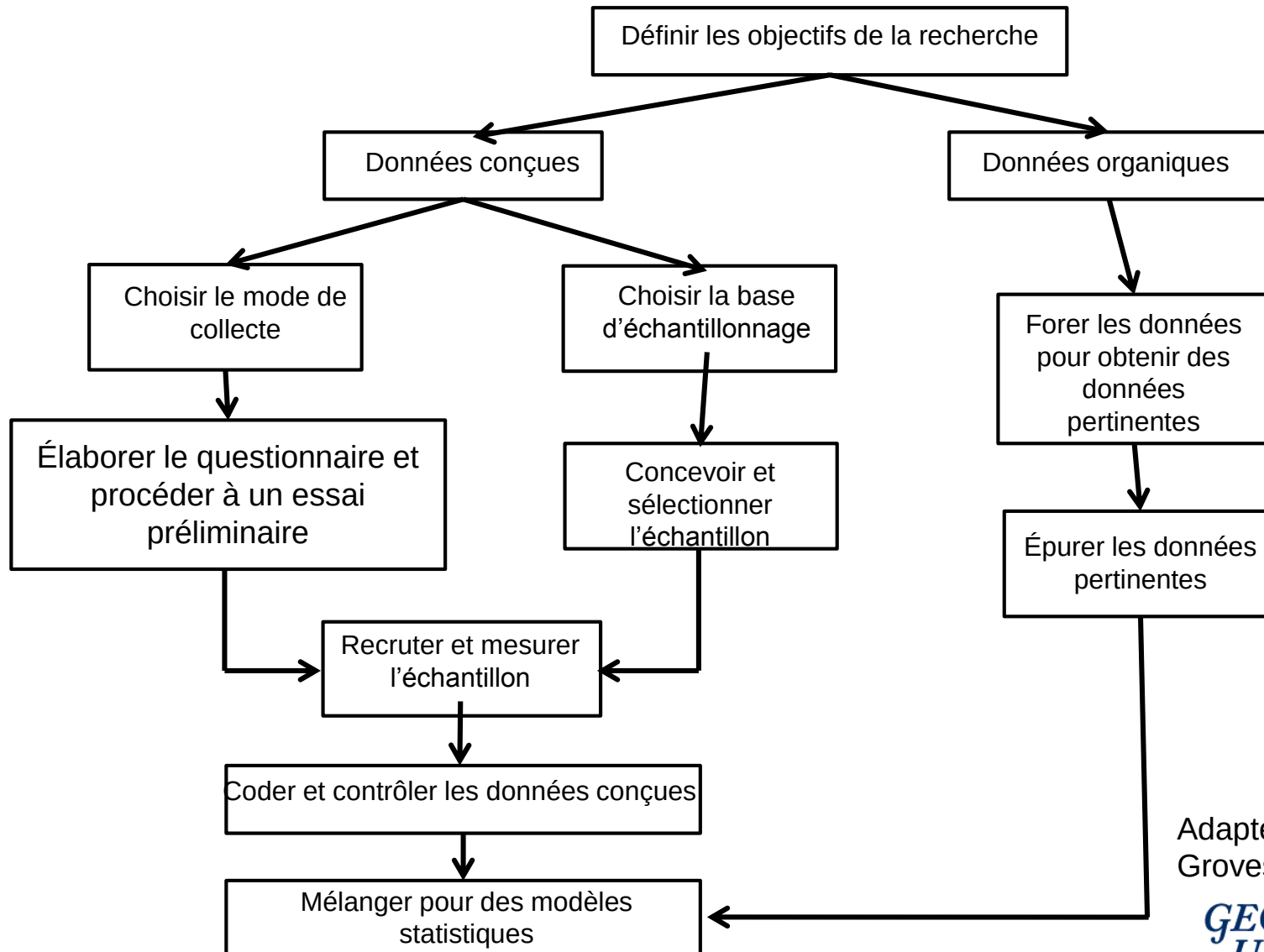
Approche des données manquantes



Adapté de :
Raghunathan (2015)

À VENIR

Un nouvel environnement pour les statistiques officielles



Adapté de la figure 2.4,
Groves et coll. (2009)

Environnement fondé sur un modèle

Pour passer à un environnement entièrement fondé sur un modèle, nous devons :

- Élaborer des normes de transparence
- Inventer des statistiques simples pouvant être comprises par de grands groupes
- Montrer fidèlement l'incertitude des estimations
- Élaborer des normes pour l'analyse de sensibilité.