Towards a Quality Framework for Blends of Designed and Organic Data

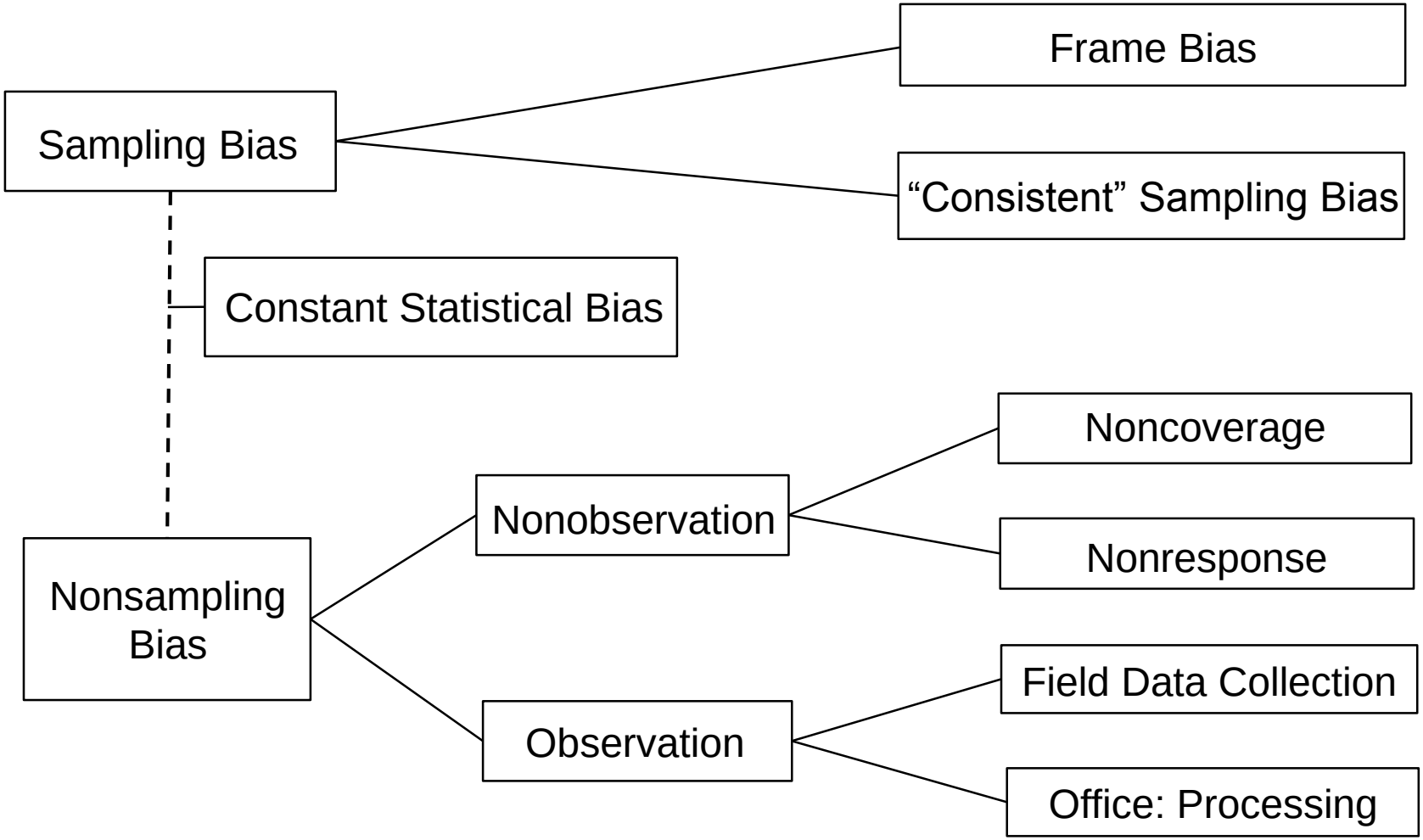
Robert M. Groves

GEORGETOWN UNIVERSITY

Outline

1. Evolution of concepts of statistical quality
2. Alternative taxonomies of quality
3. The rise of organic data, the demise of designed data
4. Blending organic and designed data
5. The need for a new quality framework for organic data
6. Blending models

Sampling and Nonsampling Bias



Adapted from Deming (1943) and Kish (1965)

Total Survey Error

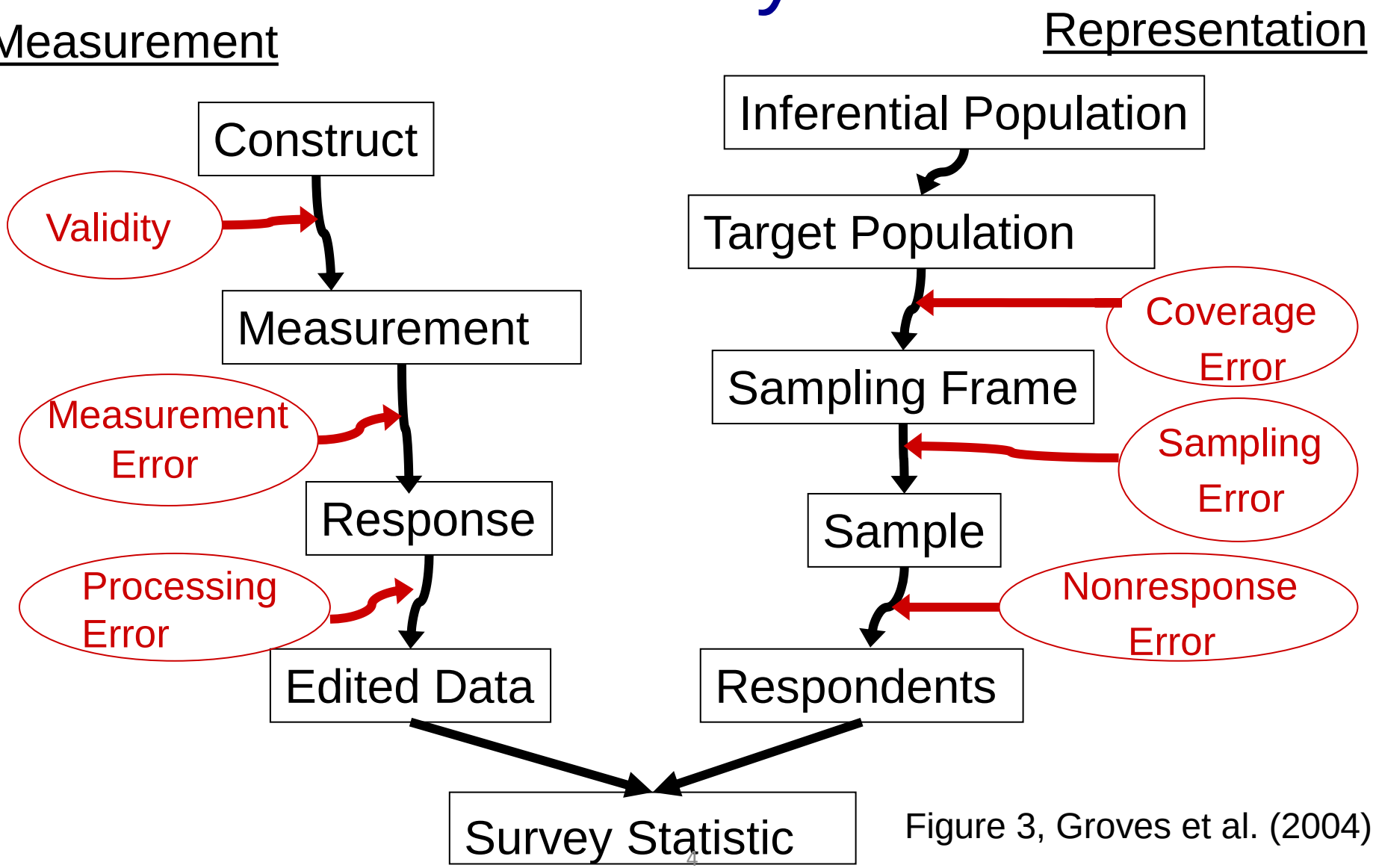


Figure 3, Groves et al. (2004)

Fitness for Use

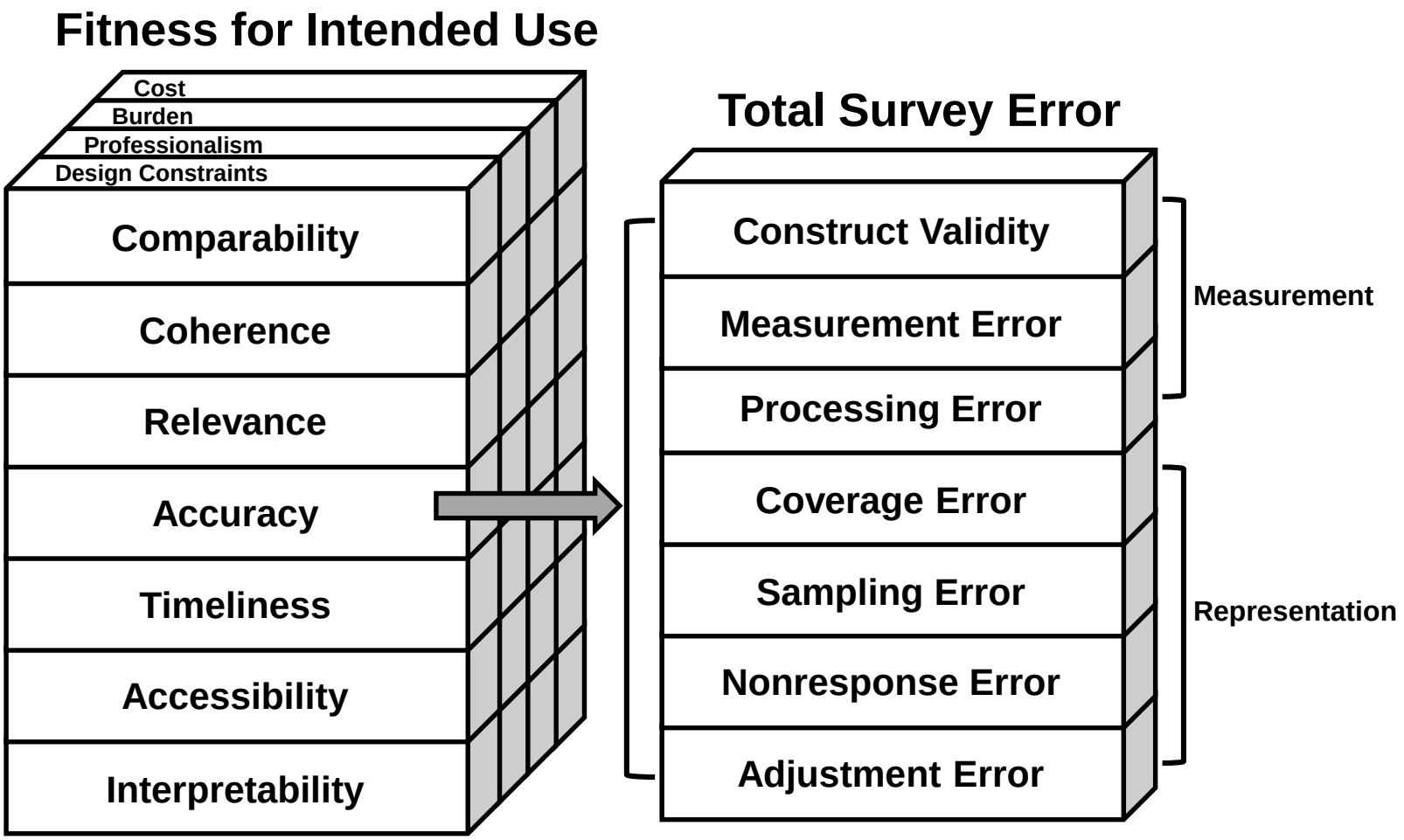
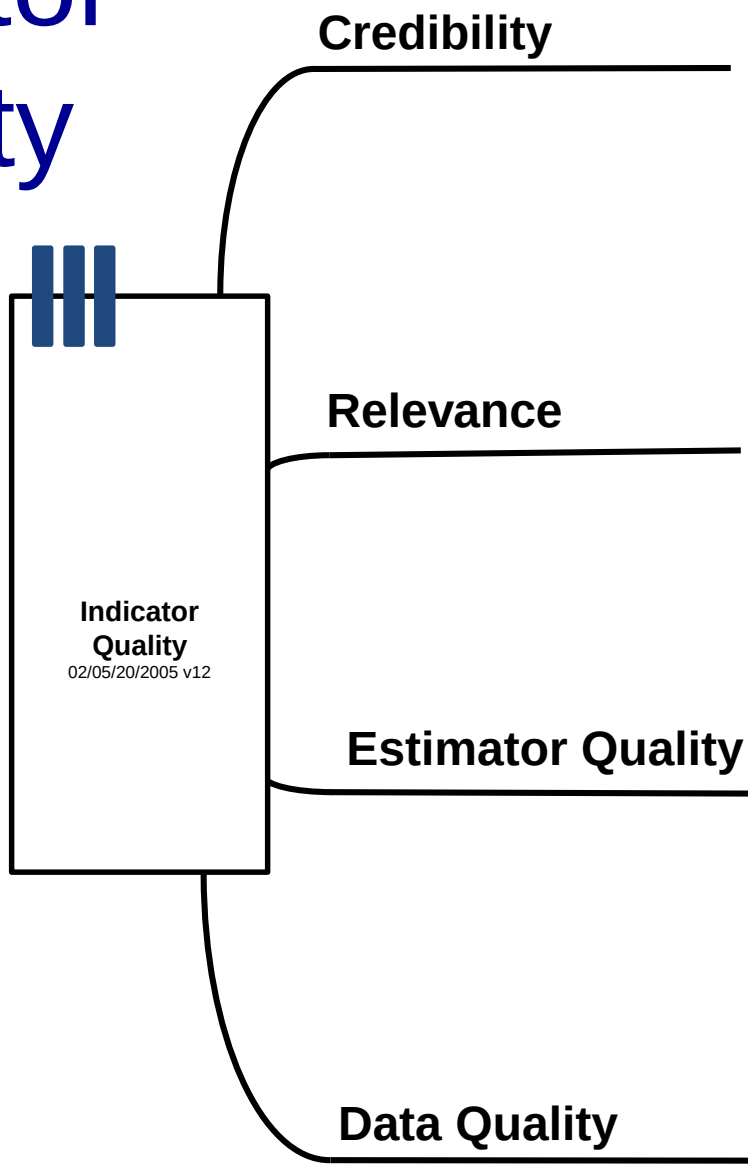


Figure 2, Hansen et al. (2010)

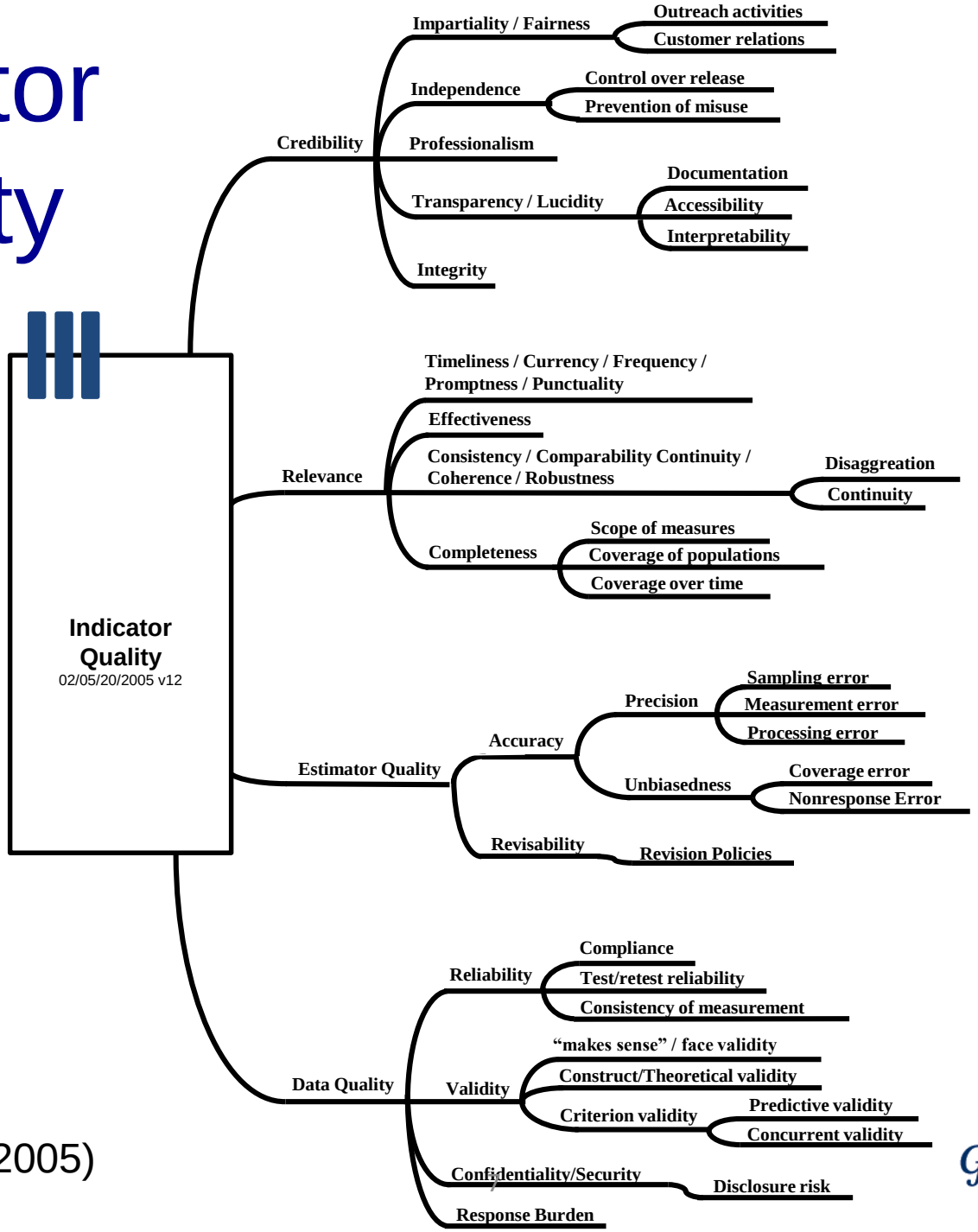
GEORGETOWN UNIVERSITY

Indicator Quality



Groves (2005)

Indicator Quality



Groves (2005)

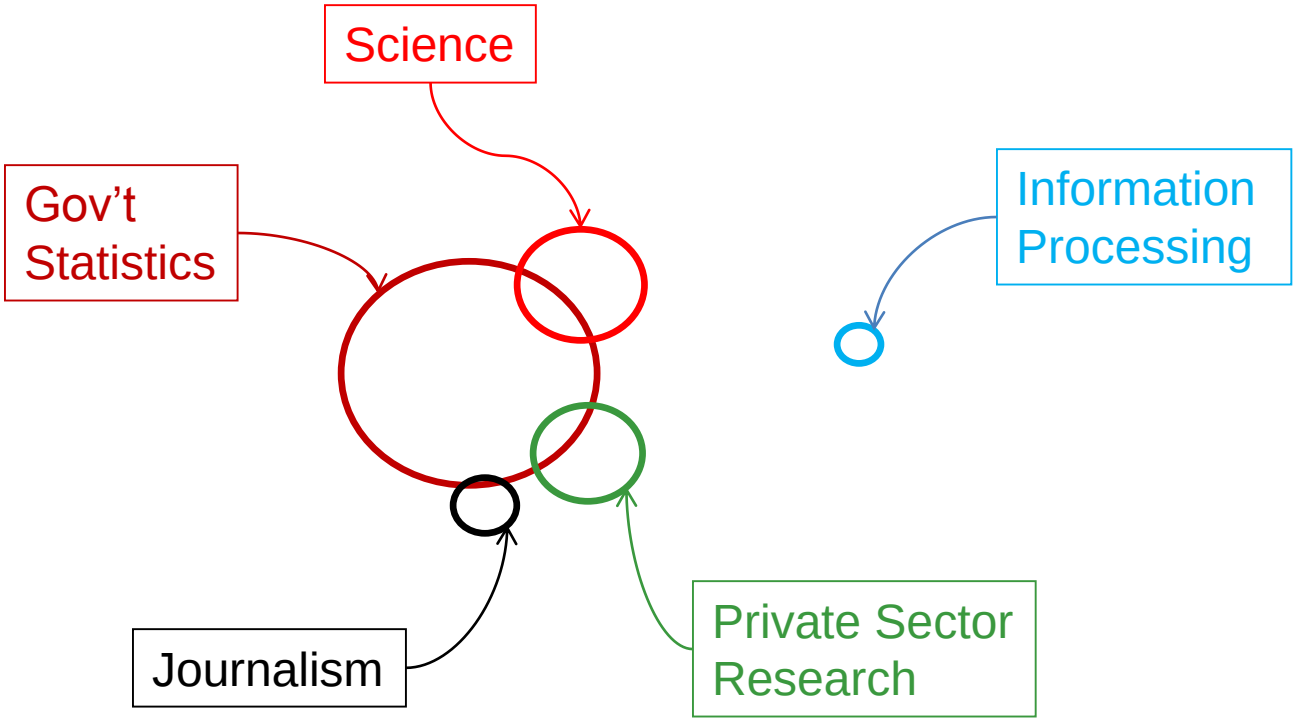
GEORGETOWN UNIVERSITY

Summary of Quality Frameworks

- They offer a language
- They offer a conceptual framework, relating components to one another
- They permit focused design of data evaluation
- But, they have not in general yielded measureable error components

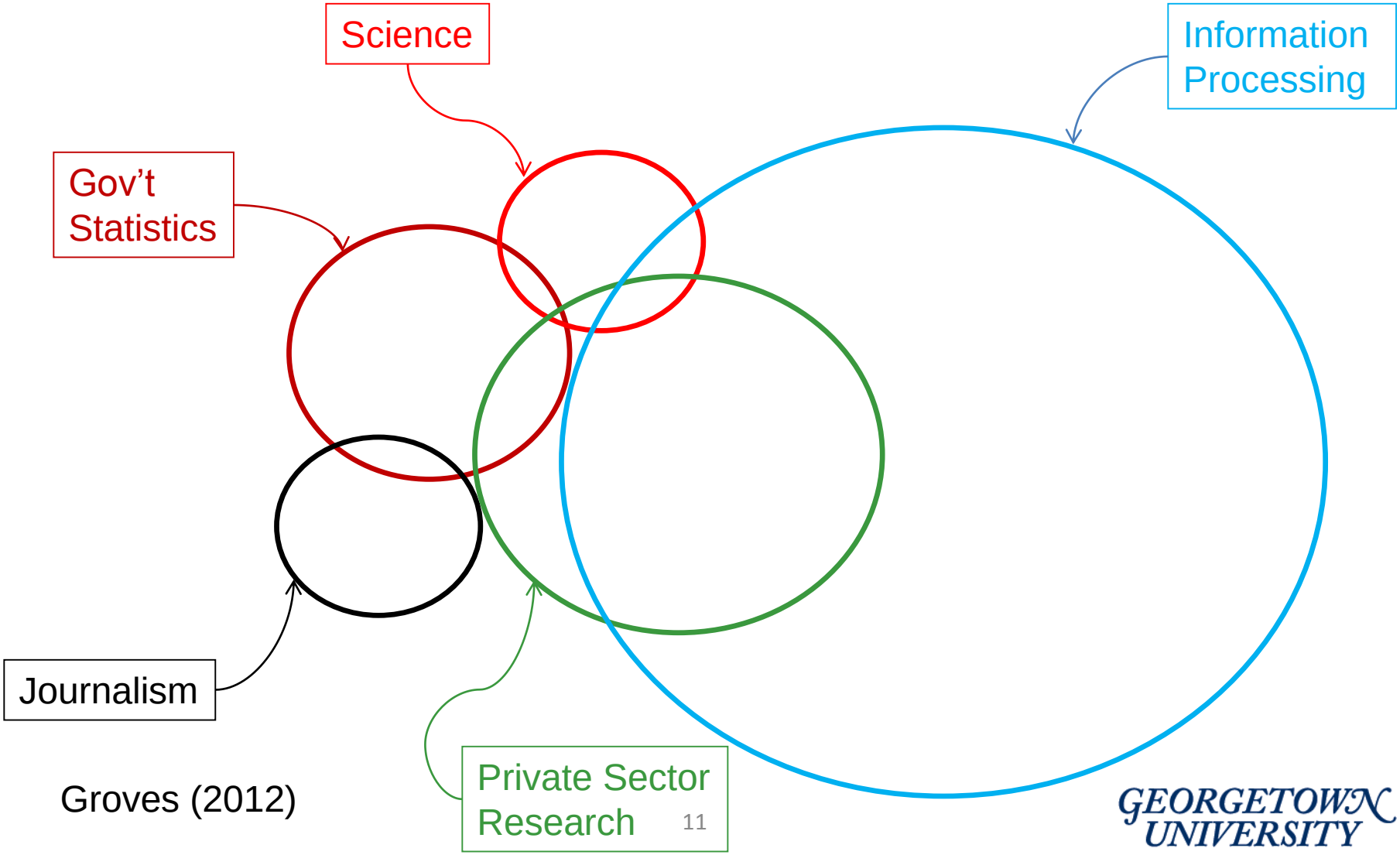
THE RISE OF ORGANIC DATA

Relative Sizes of Digital Data Production, c.1960



Groves (2012)

Relative Sizes of Digital Data Production, c.2010



Groves (2012)

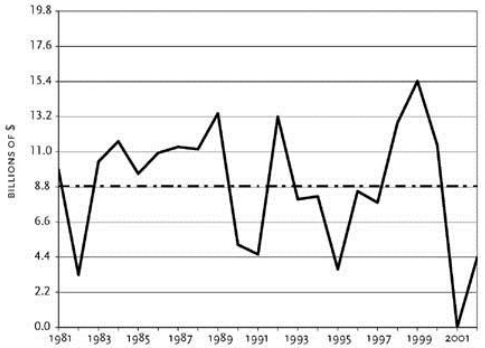
GEORGETOWN UNIVERSITY

Types of Organic Data

- Transaction data (scanner, credit card, utility usage, health service, tweets)
- Social network communication data (Twitter, instagram)
- Data from software agents in mobile devices (GPS-associated data)
- Data from the internet of things (weather sensors)
- Biometric data
- Communication data (blogs, texts, emails)
- Internet search data
- CCTV digital data
- Data scraped from websites

Credit Card and Nielsen Scanner Data

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



GEORGETOWN UNIVERSITY

Organic Data from Mobile Devices

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

Bus Route Image from: Wakefield (2013)
http://www.bbc.com/news/technology-22357748

GEORGETOWN UNIVERSITY

Picture

Following

Followers

TWEETS 268 | FOLLOWING 582 | FOLLOWERS 1,307 | LIKES 11

Message Content

Follow

Robert M. Groves
@RobertMGroves

Name

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

Bio

Georgetown University
provost.georgetown.edu
Joined September 2012

Tweet to Robert M. Groves

Tweets | Tweets & replies

Retweets

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…

1 1

Tweets

Robert M. Groves @RobertMGroves · Jan 13
New post: "Seeking Faculty Input on What They Think about Georgetown" at blog.provost.georgetown.edu

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/…

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

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

Adapted from Daas, et al. (2015)

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

Following

TWEETS 268 | FOLLOWING 582 | FOLLOWERS 1,307 | LIKES 11

Follow

Who to follow · Refresh · View all

Capital Weather Gang ... Follow

POLITICO @politico Follow

Washington Post @was... Follow

Find friends

Trends · Change

#twitterdown

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 Twitterverse 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

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

GUWLI GEORGETOWN UNIVERSITY WOMEN'S LEADERSHIP INSTITUTE

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

Followers

TWEETS 268 | FOLLOWING 582 | FOLLOWERS 1,307 | LIKES 11

Follow

Follow

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

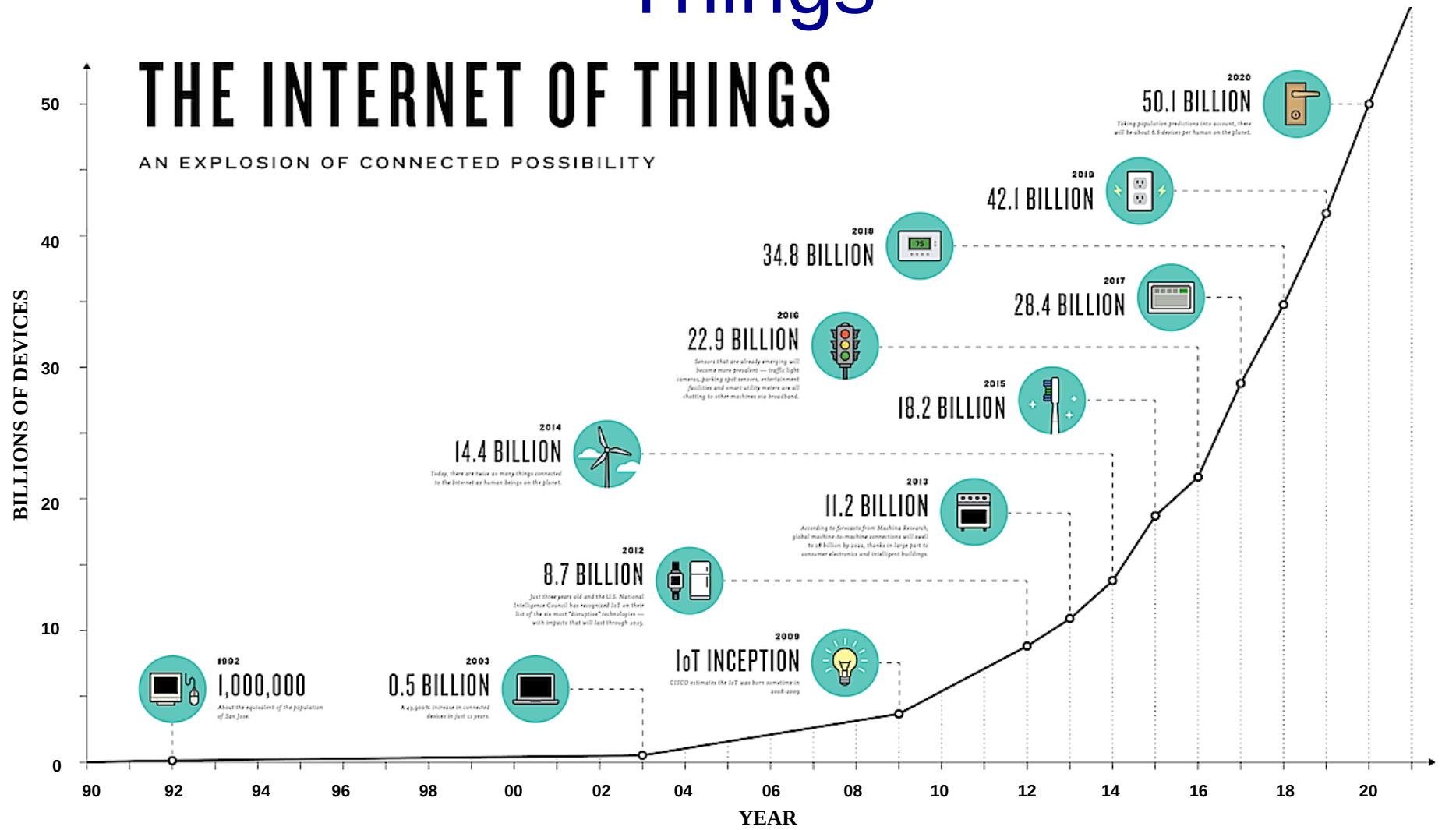
UM SurveyMethodology
@MichPSM
The University of Michigan's Program in Survey Methodology is a program where students learn the science of surveys.

Follow

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

GEORGETOWN UNIVERSITY

Exponential Growth of Internet of Things



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

GEORGETOWN UNIVERSITY

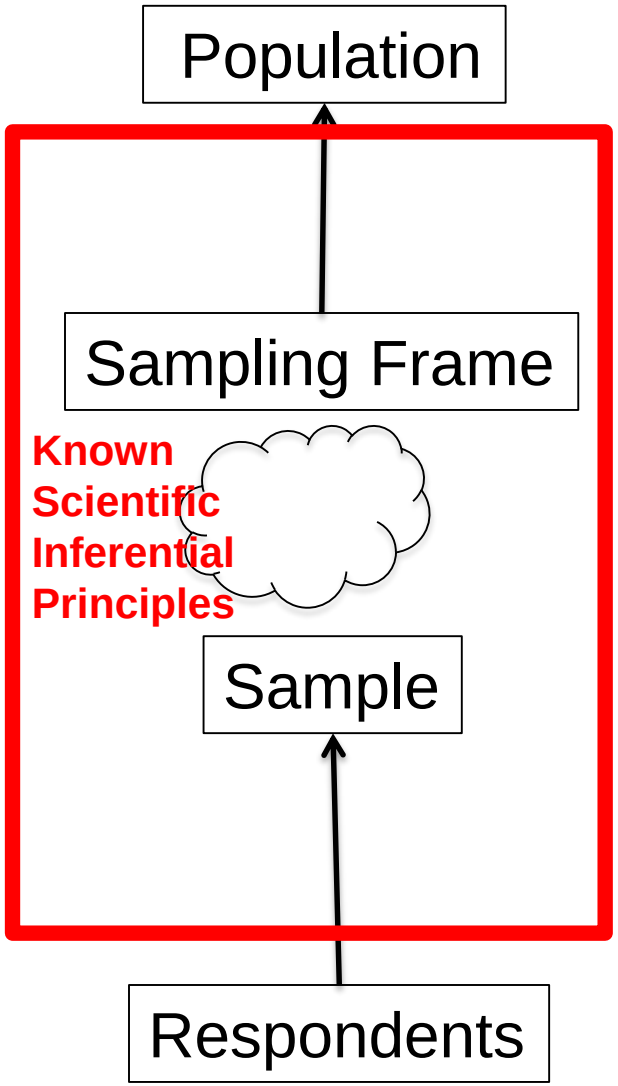
Organic data strengths

- They measure novel behaviors
- Frequency is near real-time
- High spatial granularity
- Ability to measure networks

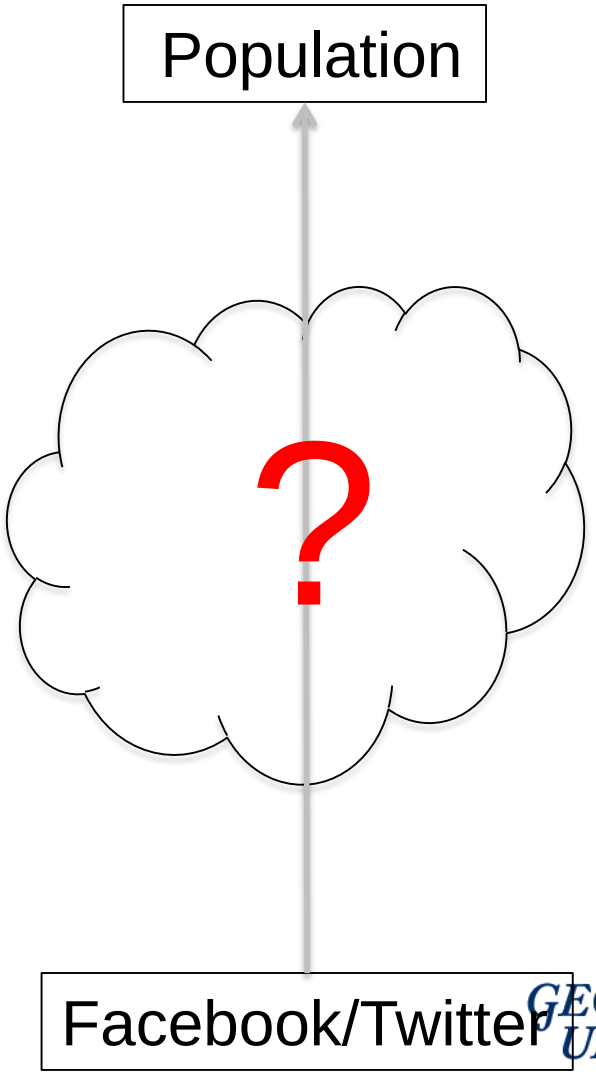
Organic data weaknesses

- Not all members of the population are covered
- Not multivariate; lean in variables
- Identifying who supplied the data is difficult
- The structure of the data is hostile to analysis
- Access and content are controlled by owners of the data

The Sample Survey Paradigm



The "Big Data" Paradigm



THE NEED FOR A QUALITY FRAMEWORK FOR ORGANIC DATA

Four Examples of Organic Data Problems Needing a Quality Framework

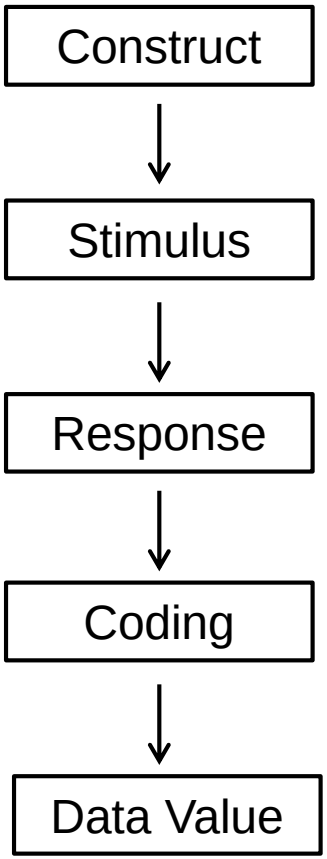
1. Organic data generated without a known external stimulus
2. Organic data stimulated by a single process that is not known by the statistician
3. Organic data that are products of models themselves
4. Organic data from text

The Absence of a Measurement Stimulus

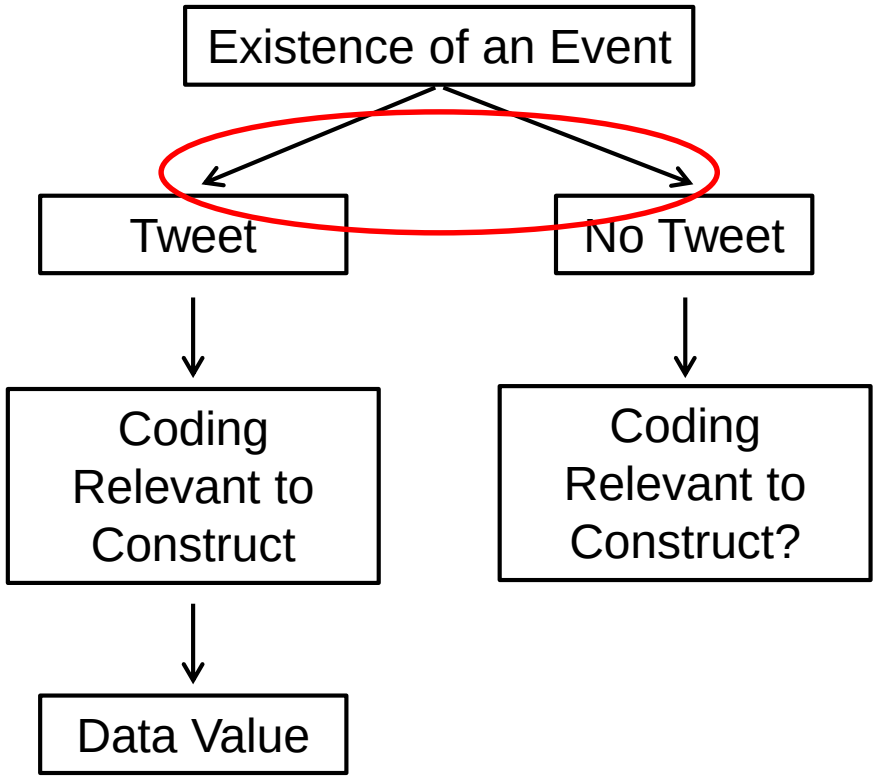
- Many sources of organic data do not arise in reaction to a uniform stimulus applied to all measurement units
- Examples:
  - Tweets
  - Blogs

The Absence of a Measurement Stimulus

Designed Data



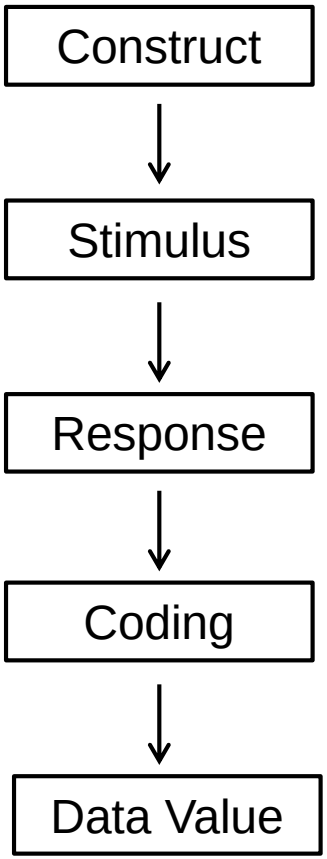
Organic Data



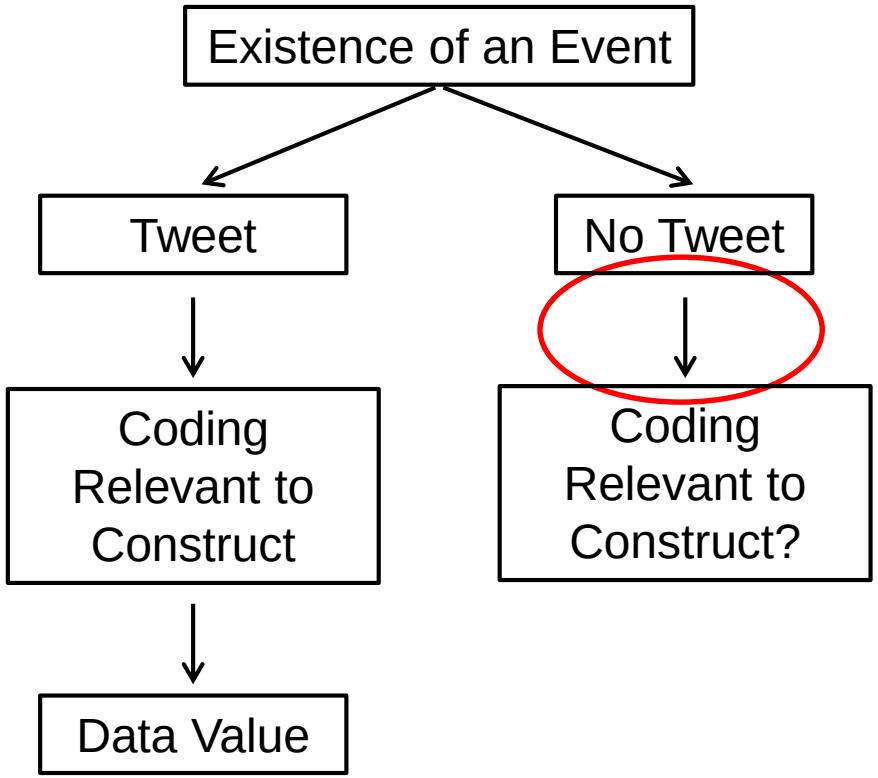
GEORGETOWN UNIVERSITY

A Measurement Stimulus that is a Known Unknown

Designed Data



Organic Data

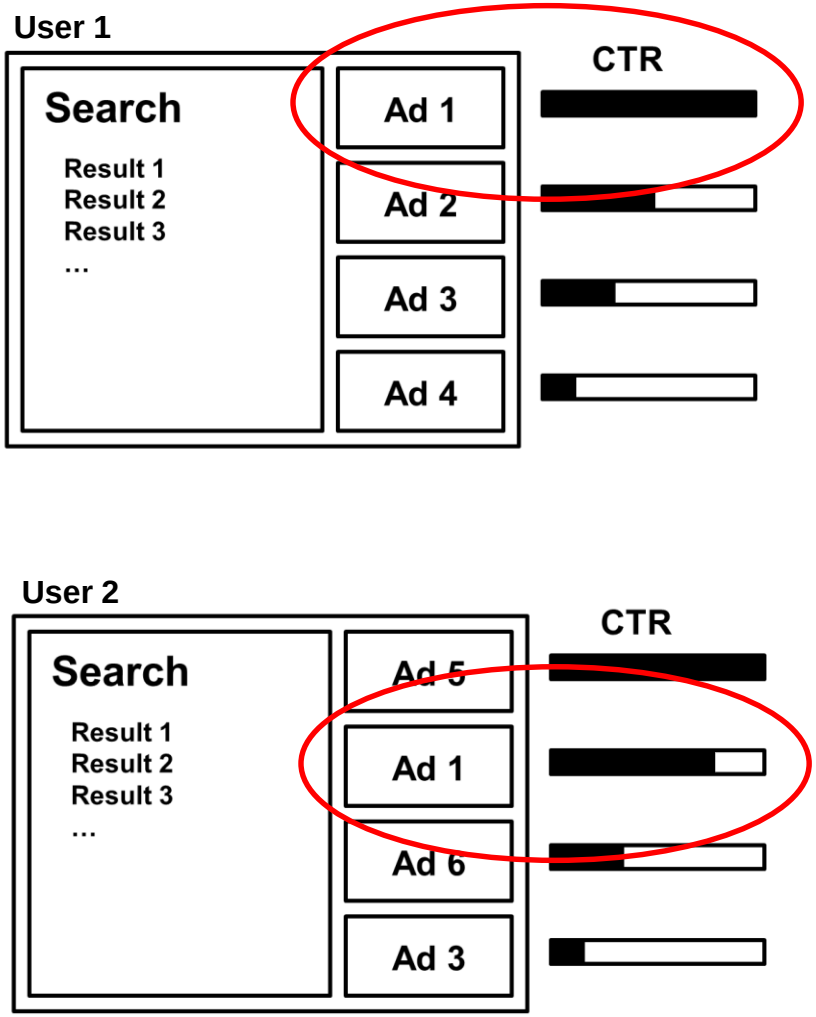


GEORGETOWN UNIVERSITY

Organic Data With Intermediate Measurement Error Features

- Organic data created by multi-step processes affecting measurement error properties
- Example:
  - Click data on web-page advertisements that were displayed based on predictive models of user interest

Advertisement Location Affects Click Through Rate



Adapted from Figure 1, Richardson et al. (2007)

GEORGETOWN UNIVERSITY

Data Generated from Models using Unstructured Data

- Some organic data come from model-based coding
- Example:
  - Facial recognition
  - Traffic data from CCTV
  - GPS

Measurement Error Estimates Included with Data Items

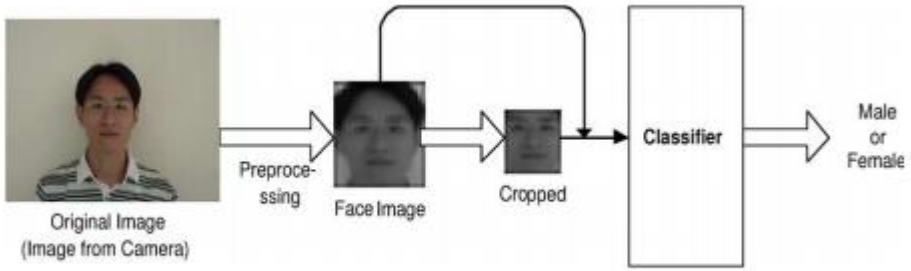


Fig. 1. The process of appearance-based gender classification.

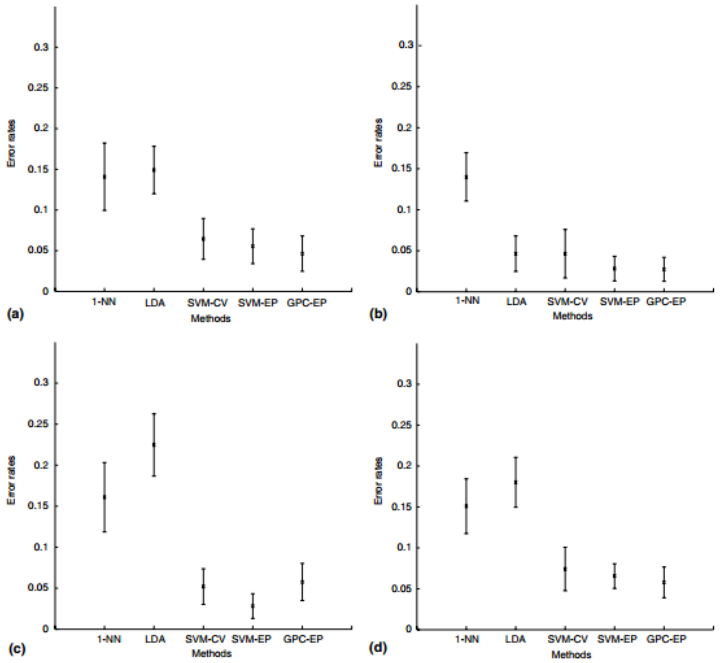


Fig. 5. Classification error rates of various methods for four kinds of gender classification data sets (PF'01 DB): (a) data set P-I, (b) data set P-II, (c) data set P-III, (d) data set P-IV.

Figure 1 and Figure 5
Kim et al. (2006)

GEORGETOWN UNIVERSITY

Text to Number Translation

- Organic data based on coding from text data
- Example:
  - Twitter text coding
  - Blog coding
  - LinkedIn coding

Coding in Sample Survey

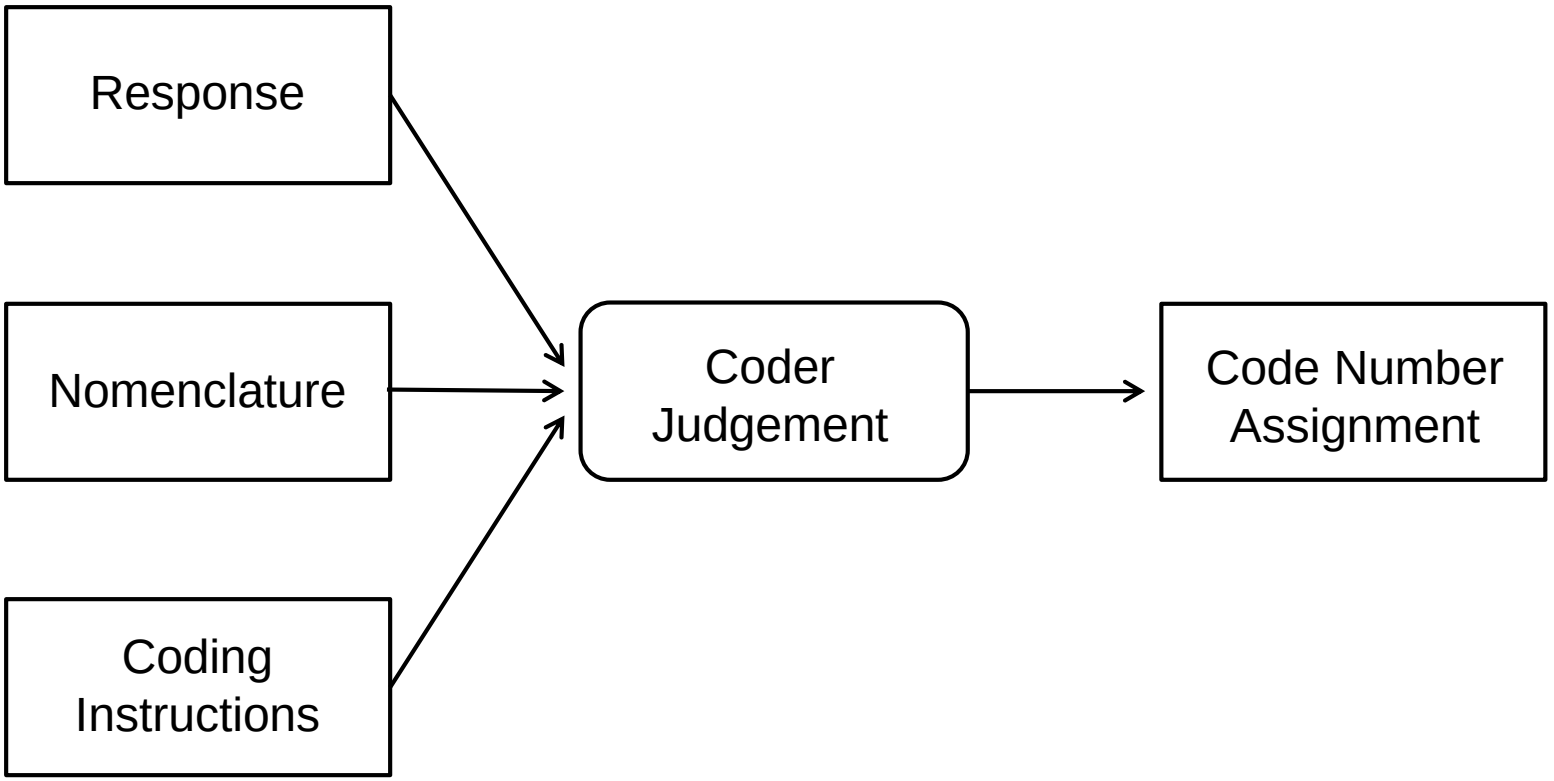


Figure 7.1, Biemer & Lyberg (2003)

GEORGETOWN UNIVERSITY

Sources of Coding Errors from Lexical Polysemy

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.

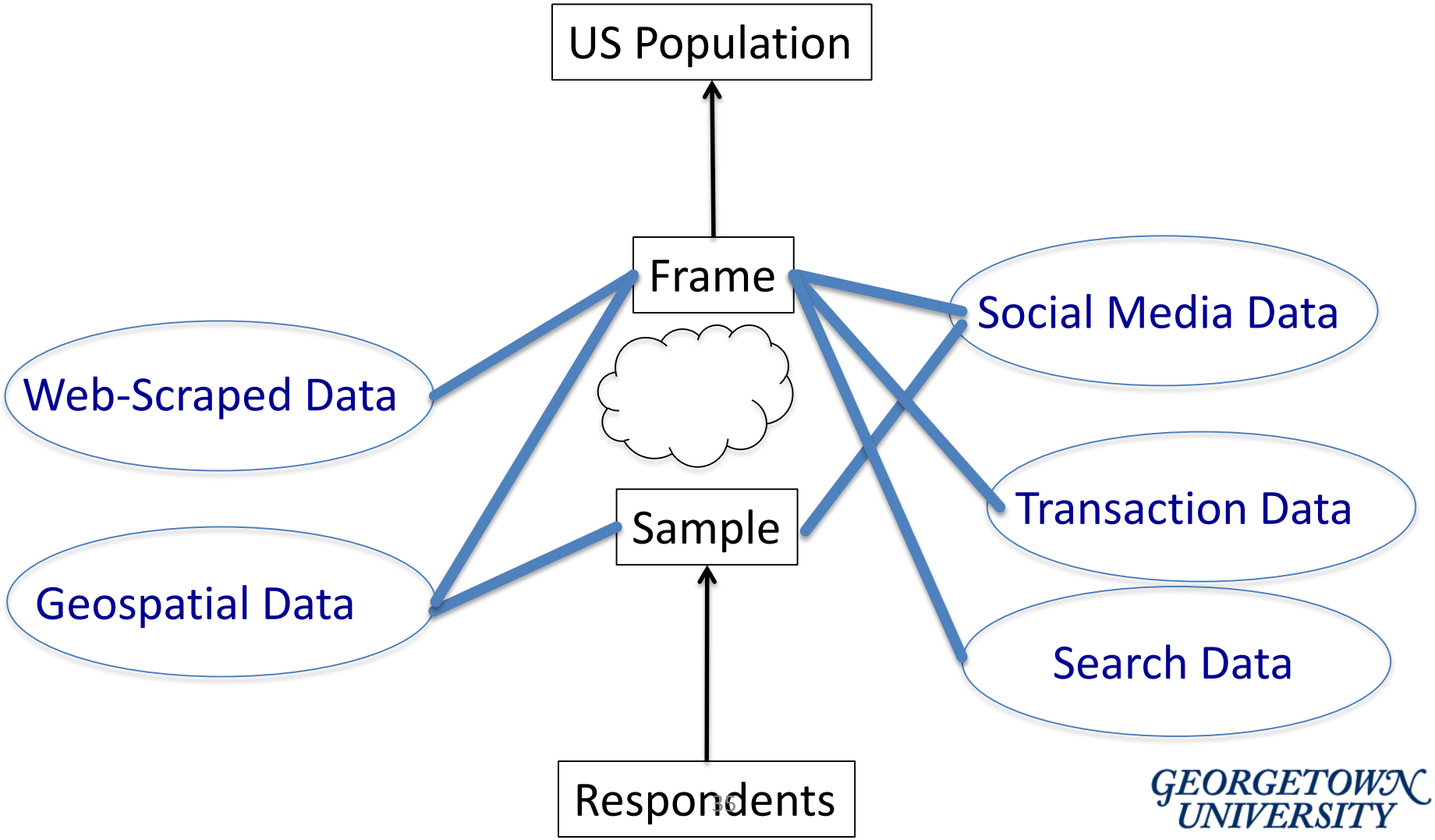
Adapted from Frisbie & Sudman (1968)

BLENDING ORGANIC AND DESIGNED DATA

Blending organic and survey data combines the temporal and spatial granularity of organic data with the inferential power of survey data

- Statistical models are key
- Shared covariates across data sets are desirable
- Focusing on coverage and measurement errors is critical

The Blended Paradigm

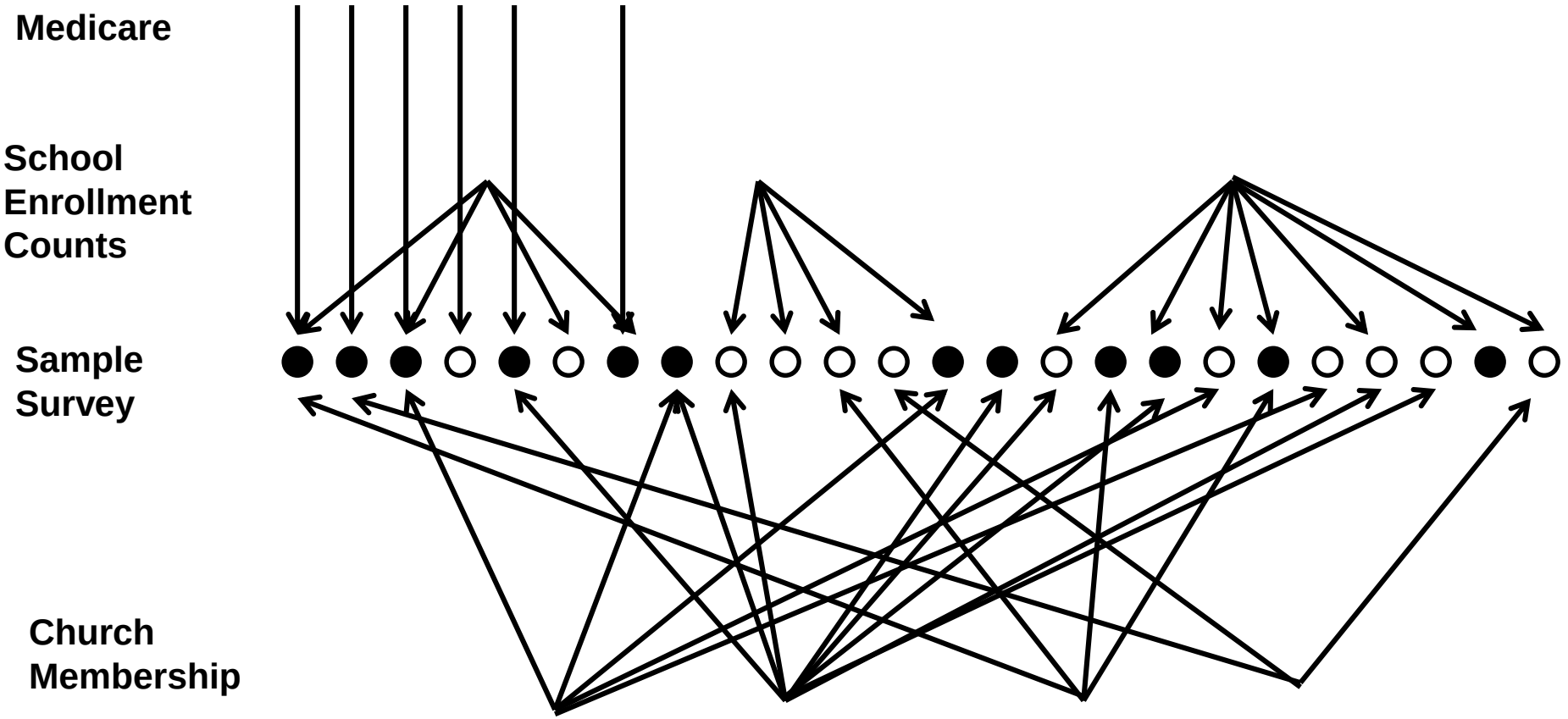


GEORGETOWN UNIVERSITY

Organic Data as Covariate Source

- Problem: estimation of statistic for many small domains
- Data resources:
  - National survey with designed data on sample of PSUs
  - Most recent census data
  - Organic data as a covariate of the key quantity to be estimated
- Example: Marchetti *et al.*(2015) for income using mobile phone GPS traffic as covariate

Small Area Estimation

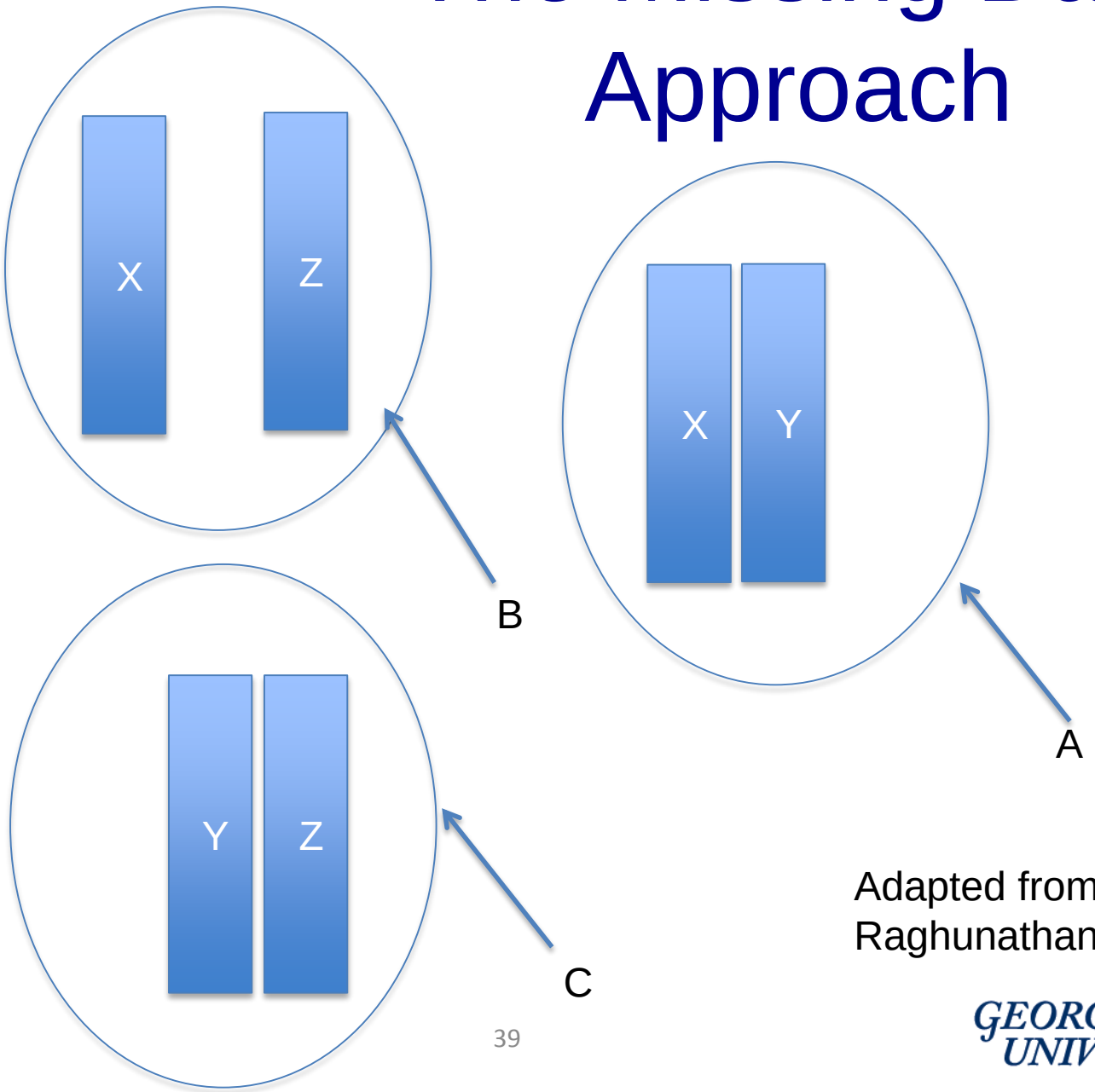


GEORGETOWN UNIVERSITY

Blending in a Missing Data Framework

- Problem: Improved estimation borrowing strength from multiple data sources
- Data resources:
  - National survey with designed data on some measures (X,Y)
  - Organic data sources with other variables (X,Z), (Y,Z)
- Example: Schenker and Raghunathan *et al.* (2006, 2007, 2010)

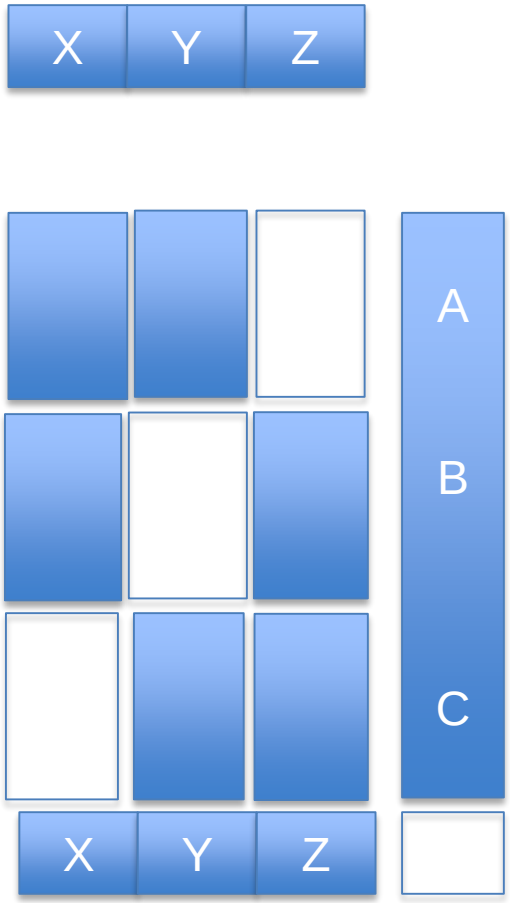
The Missing Data Approach



Adapted from:
Raghunathan (2015)

GEORGETOWN UNIVERSITY

The Missing Data Approach

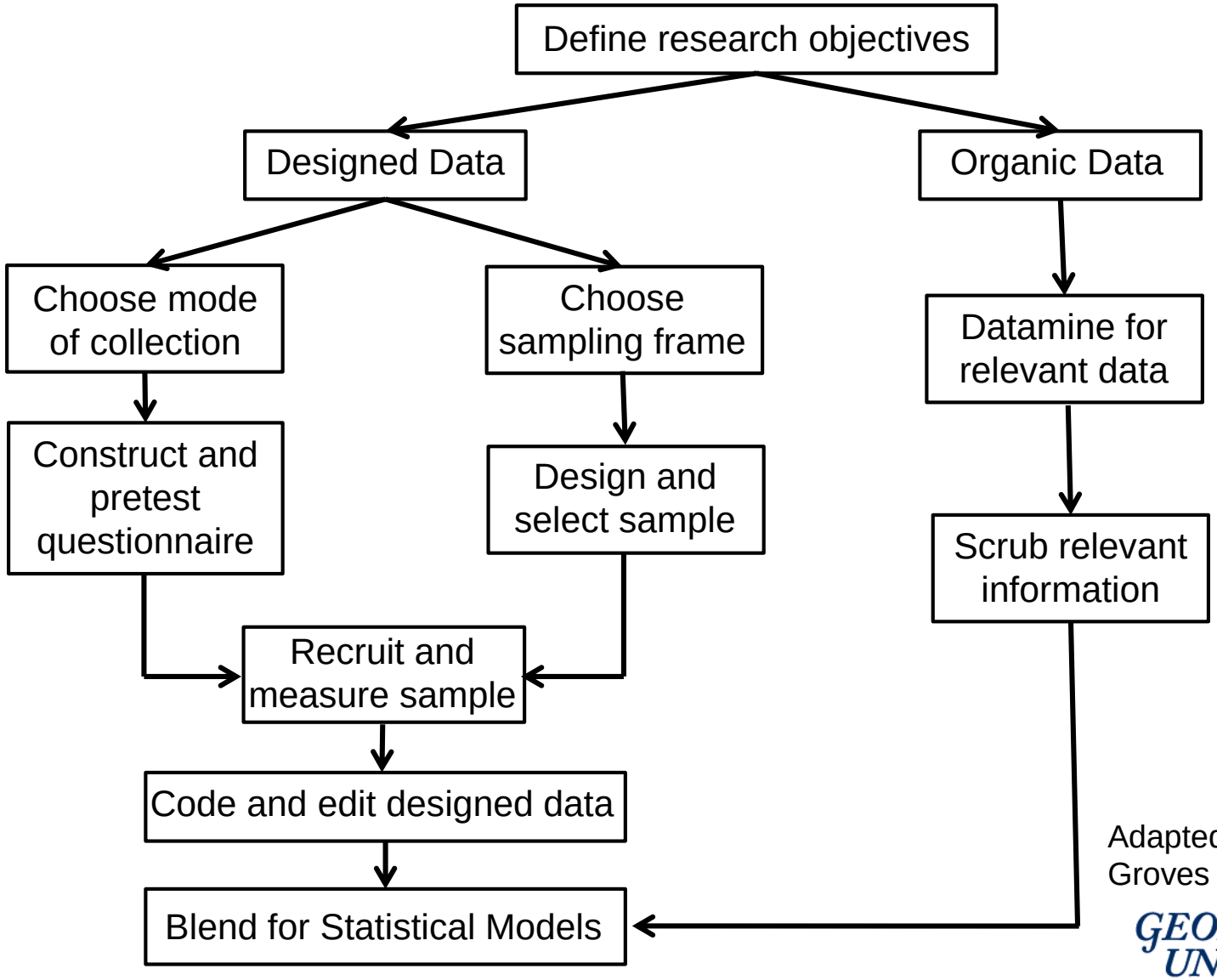


Adapted from:
Raghunathan (2015)

GEORGETOWN UNIVERSITY

MOVING FORWARD

A New World for Official Statistics



Adapted from Figure 2.4, Groves et al. (2009)

GEORGETOWN UNIVERSITY

Model Based World

In order to move to a fully model based world, we must:

- Develop standards of transparency
- Invent simple statistics understandable by large groups
- Faithfully display uncertainty of estimates
- Develop standards of sensitivity analysis.