# Big Data for the Social Sciences

## Introduction to Big Data for Social Science

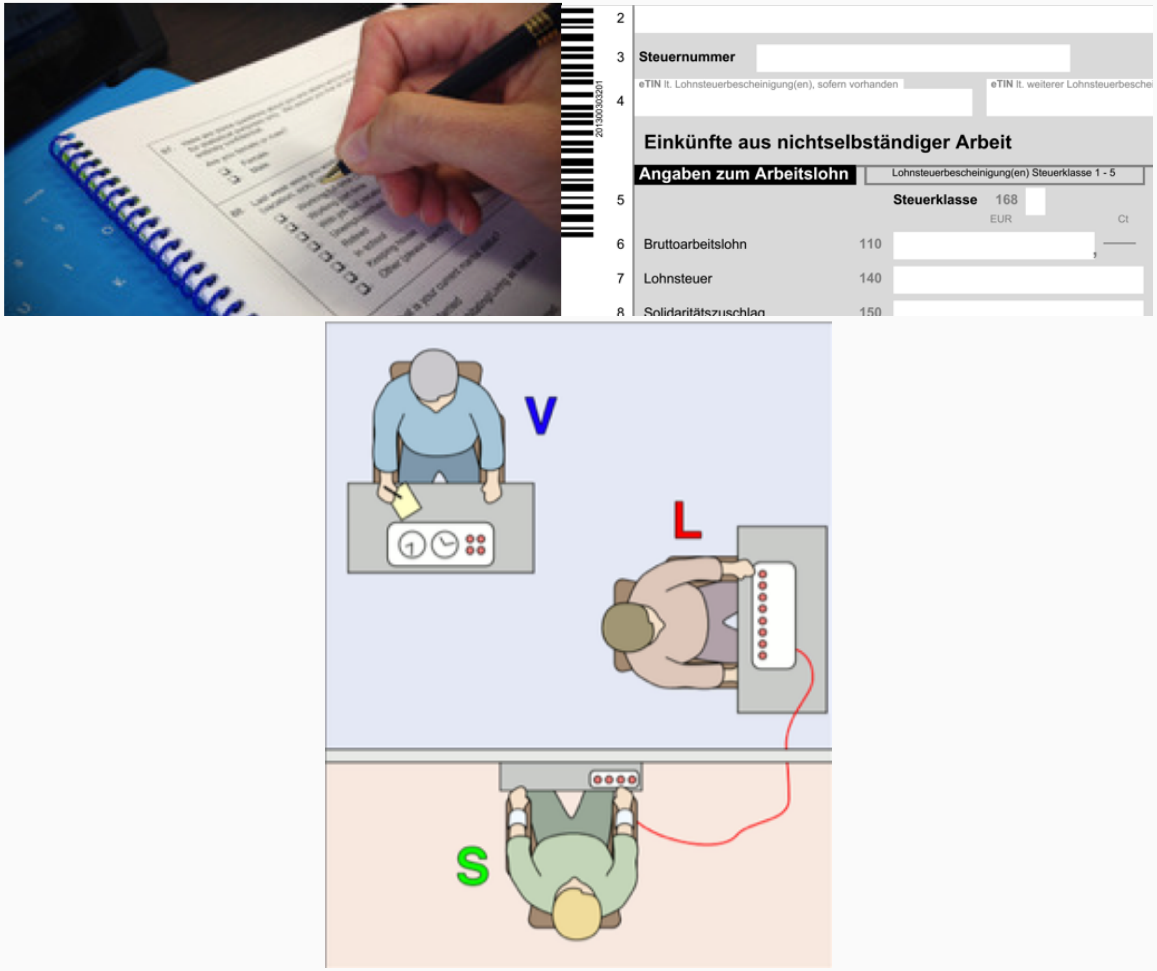
Frauke Kreuter[1] ...

June 3–4, 2019

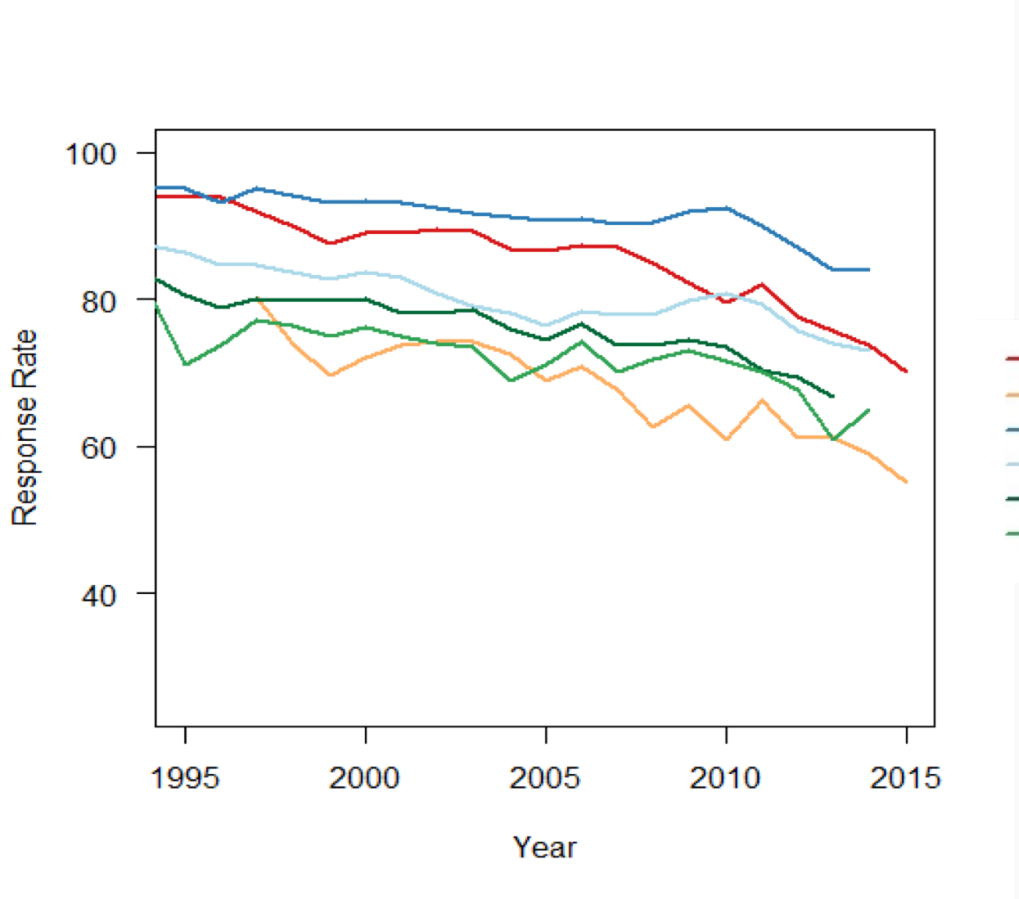
[1]fkreuter@umd.edu

# Until recently...

# Three main data sources

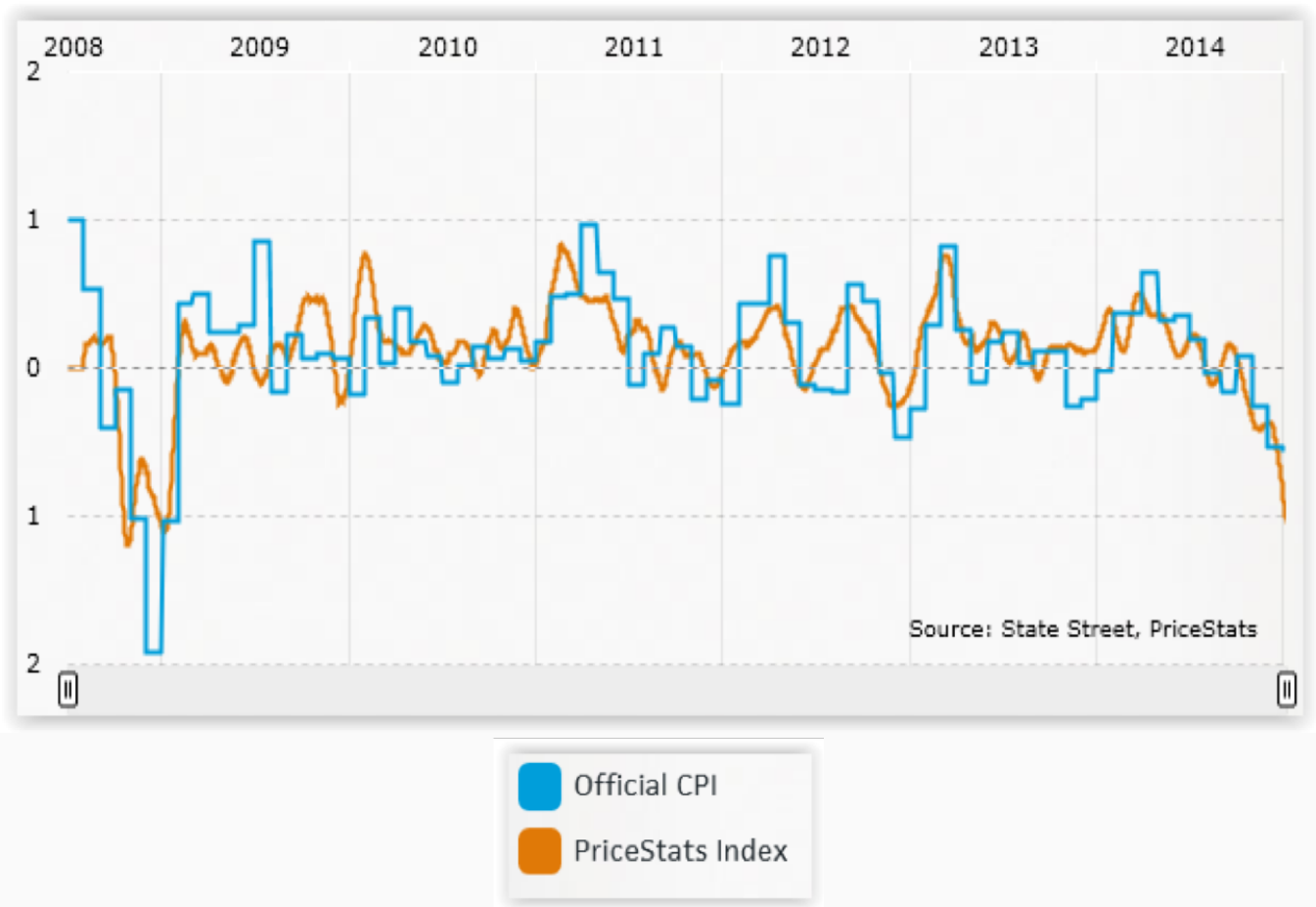


# U.S. Surveys RR



# Now…

# US Aggregated Inflation Series

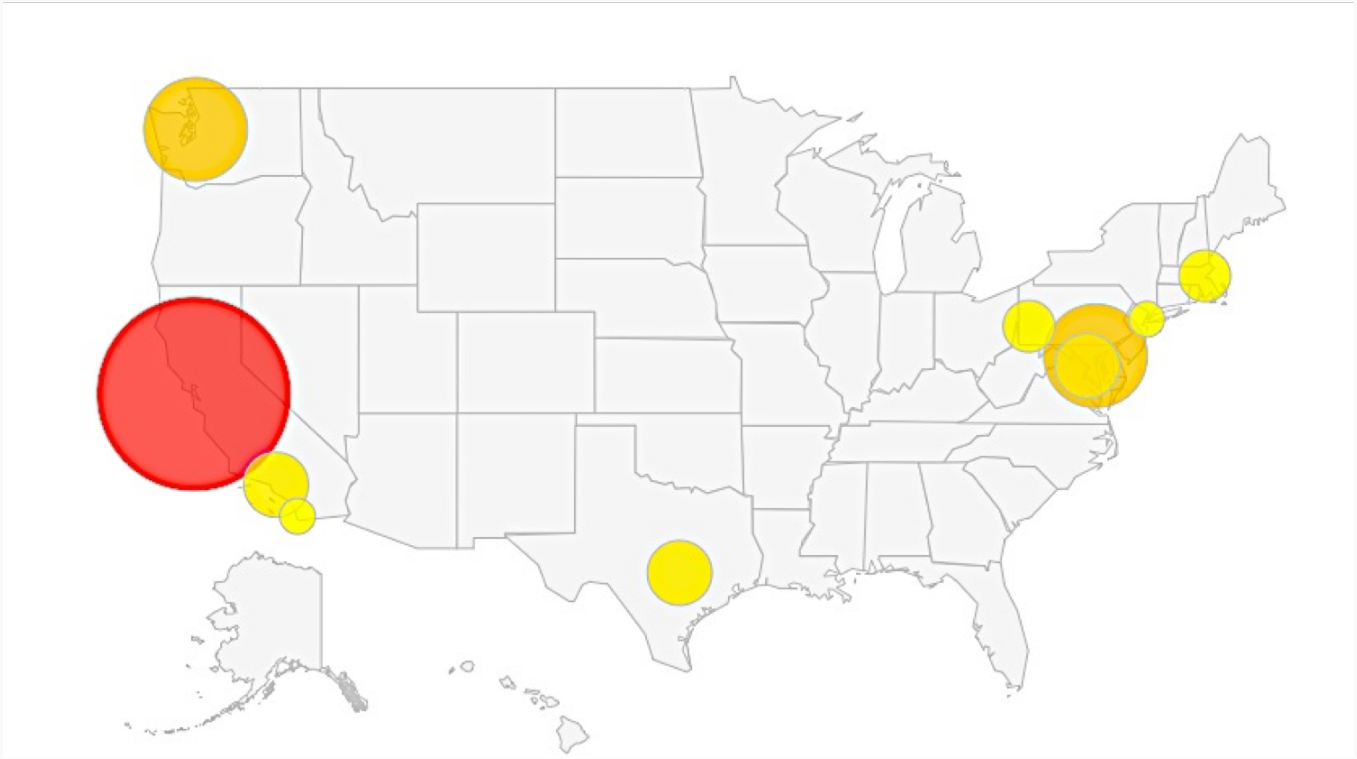



US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI. Accessed January 18, 2015 from the PriceStats website.

# Hadoop skill intensity

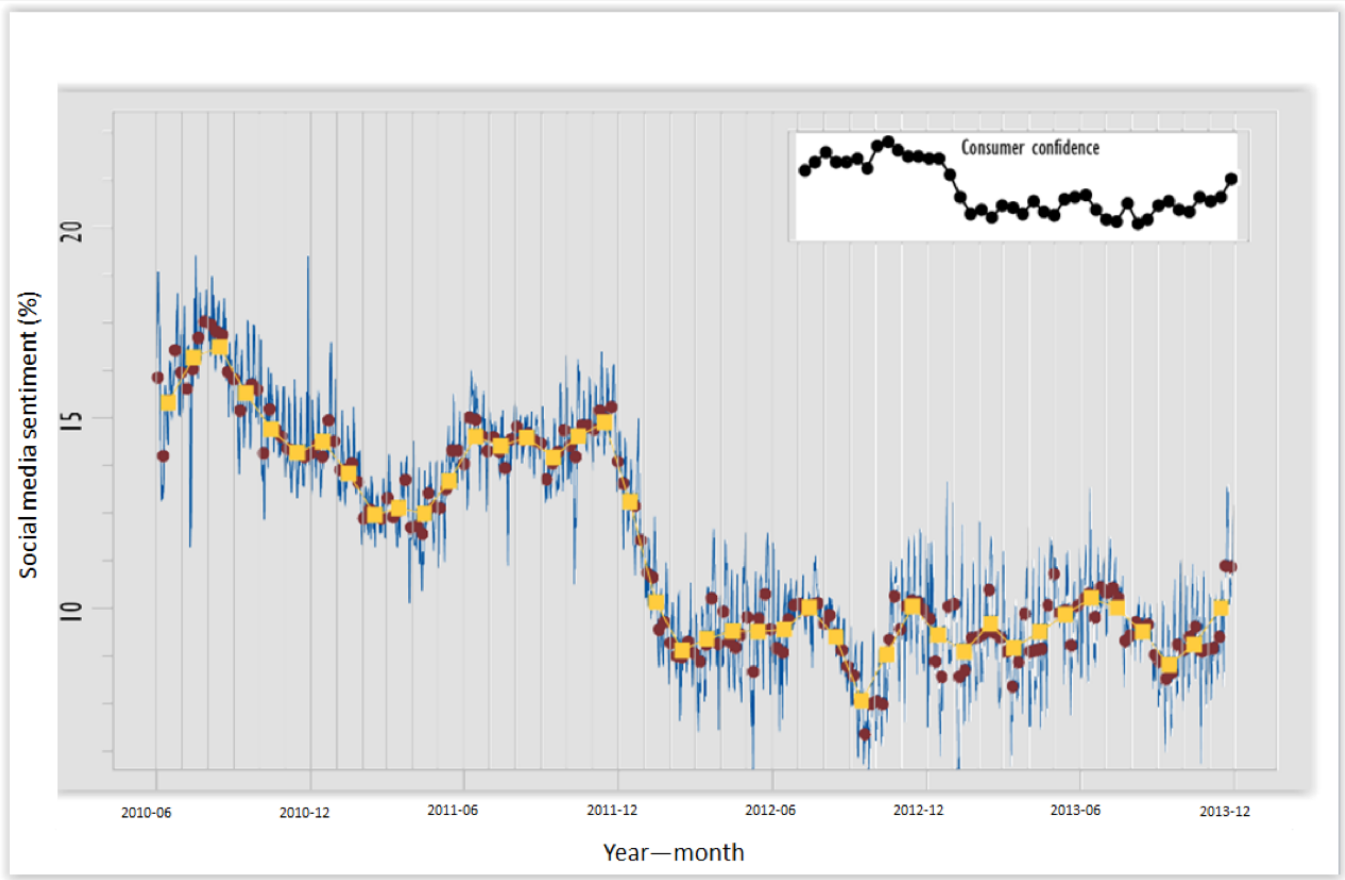
Hadoop skill intensity is highly concentrated

(normalized by total IT labor force size)



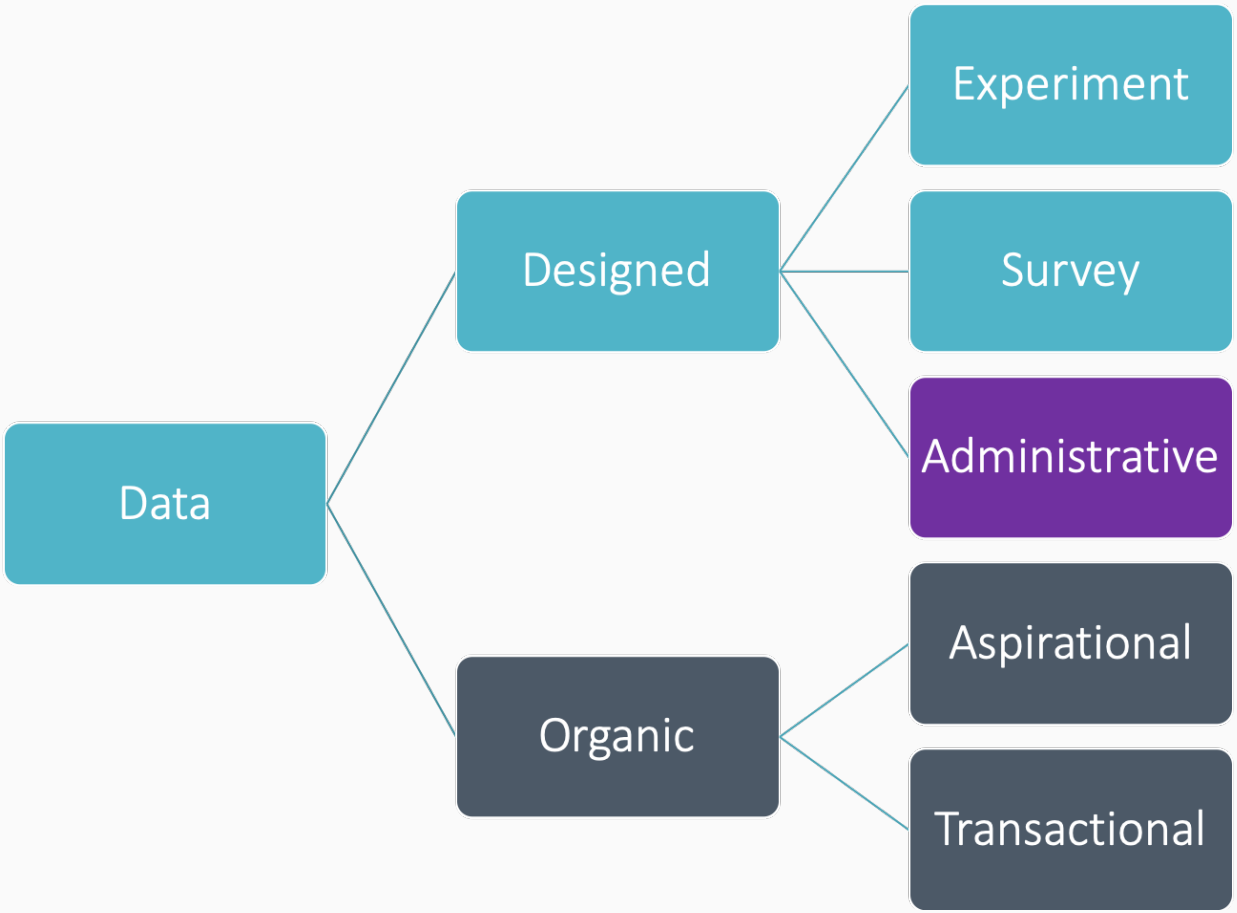
Source: Sonny Tambe – LinkedIn Data

# Social media sentiment




Social media sentiment (daily, weekly and monthly) in the Netherlands, June 2010 - November 2013. The development of consumer confidence for the same period is shown in the insert (Daas and Puts 2014).

# Data sources




Source: Roberto Rigobon

# Hope that found/organic data...

- Can replace or augment expensive data collections
- More (= better) data for decision making
- Information available in (nearly) real time

# New paradigm

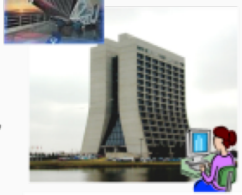
- New business model
- New analytical model
  - Outliers
  - Finegrained analysis
  - New units of analysis
- New sets of skills
  - Computer scientists
  - Citizen scientists
- Different cost structure

Source: Julia Lane

## Big Data

Jim Gray's paradigm

- **Observational Science**
  - Scientist gathers data by direct observation
  - Scientist analyzes data
- **Analytical Science**
  - Scientist builds analytical model
  - Makes predictions.
- **Computational Science**
  - Simulate analytical model
  - Validate model and makes predictions
- **Data Exploration Science**
  - **Data-driven science**
    Data captured by instruments or from the web, or data generated by simulation
  - Information extraction
  - Processed by software
  - Placed in a database / files
  - Scientist(s)/scholar(s) analyze(s) database / files
  - *Access crucial*



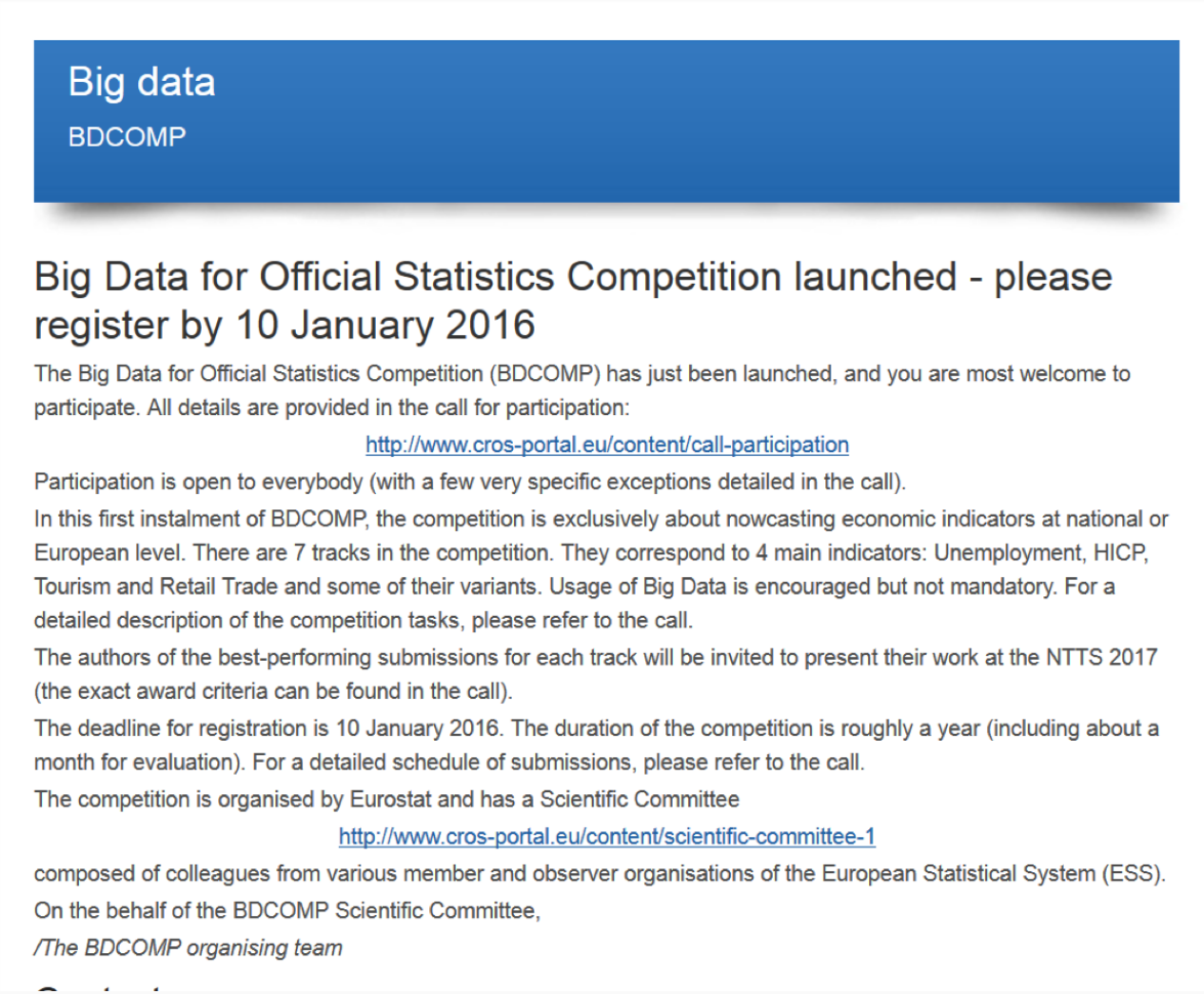
Training to Climb an Everest of Digital Data

By ASHLEE VANCE
Published: October 11, 2009

MOUNTAIN VIEW, Calif. — It is a rare criticism of elite American university students that they do not think big enough. But that is exactly the complaint from some of the largest technology companies and the federal government.

Source: Lee Giles

**Eurostat**



Big data

BDCOMP

## Big Data for Official Statistics Competition launched - please register by 10 January 2016

The Big Data for Official Statistics Competition (BDCOMP) has just been launched, and you are most welcome to participate. All details are provided in the call for participation:

http://www.cros-portal.eu/content/call-participation

Participation is open to everybody (with a few very specific exceptions detailed in the call).

In this first instalment of BDCOMP, the competition is exclusively about nowcasting economic indicators at national or European level. There are 7 tracks in the competition. They correspond to 4 main indicators: Unemployment, HICP, Tourism and Retail Trade and some of their variants. Usage of Big Data is encouraged but not mandatory. For a detailed description of the competition tasks, please refer to the call.

The authors of the best-performing submissions for each track will be invited to present their work at the NTTS 2017 (the exact award criteria can be found in the call).

The deadline for registration is 10 January 2016. The duration of the competition is roughly a year (including about a month for evaluation). For a detailed schedule of submissions, please refer to the call.

The competition is organised by Eurostat and has a Scientific Committee

http://www.cros-portal.eu/content/scientific-committee-1

composed of colleagues from various member and observer organisations of the European Statistical System (ESS).

On the behalf of the BDCOMP Scientific Committee,

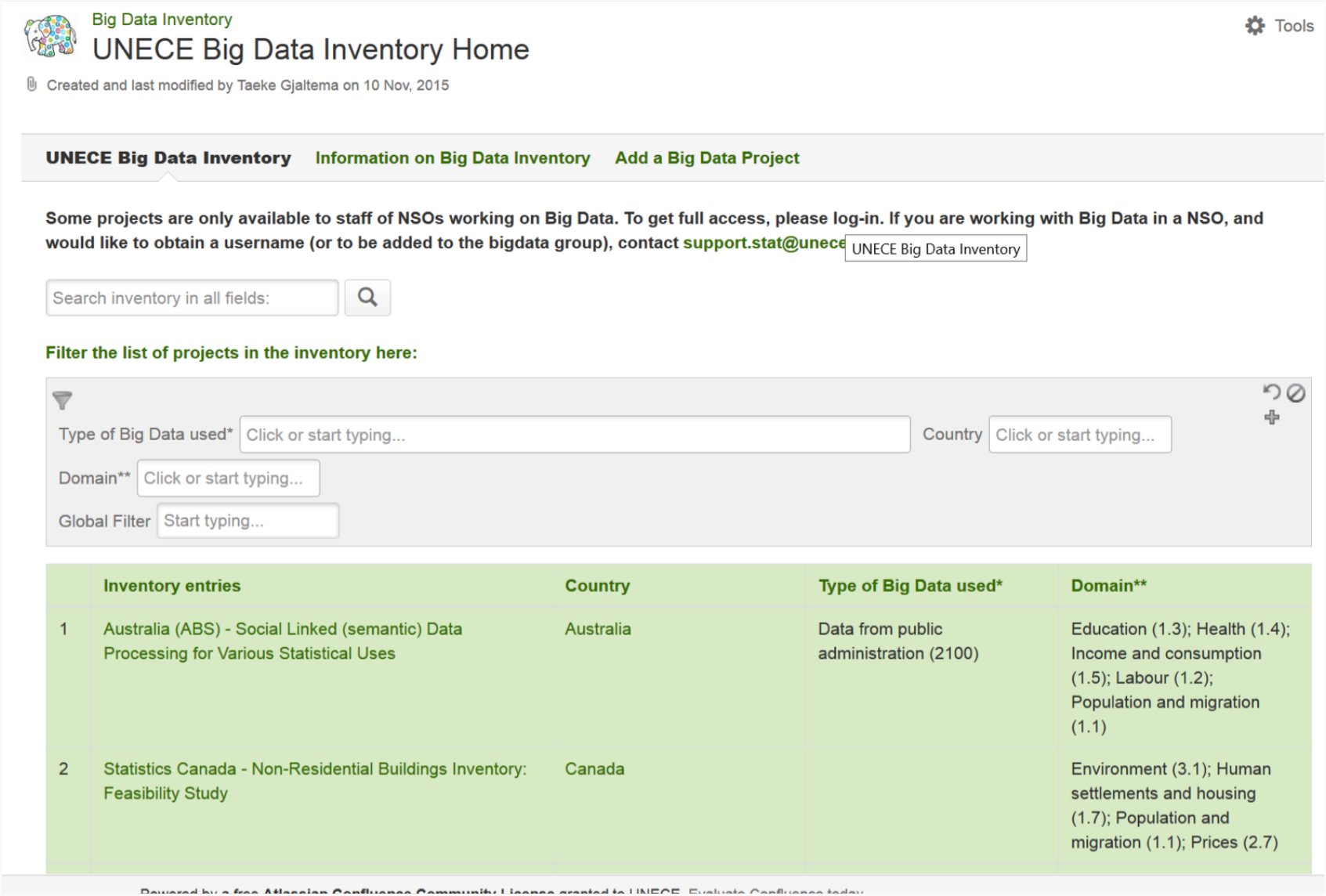
*/The BDCOMP organising team*

# UNECE and Big Data

The “Sandbox” provides a computing environment to load Big Data sets and tools

**Consumer price indices** – experimenting with the computation of price indexes
**Mobile telephone data** – statistics on tourism and daily commuting
**Smart meters** – statistics on power consumption
**Traffic loops** – traffic statistics using data from traffic loops
**Social media** – using Twitter data to analyze sentiment and to tourism flows
**Job portals** – computing statistics on job vacancies
**Web scraping** – automated data collection from web sources.

# UNECE



Big Data Inventory

UNECE Big Data Inventory Home

Tools

Created and last modified by Taeke Gjaltema on 10 Nov, 2015

**UNECE Big Data Inventory** | Information on Big Data Inventory | Add a Big Data Project

**Some projects are only available to staff of NSOs working on Big Data. To get full access, please log-in. If you are working with Big Data in a NSO, and would like to obtain a username (or to be added to the bigdata group), contact support.stat@unece**

UNECE Big Data Inventory

Search inventory in all fields:

**Filter the list of projects in the inventory here:**

Type of Big Data used* Click or start typing...

Country Click or start typing...

Domain** Click or start typing...

Global Filter Start typing...

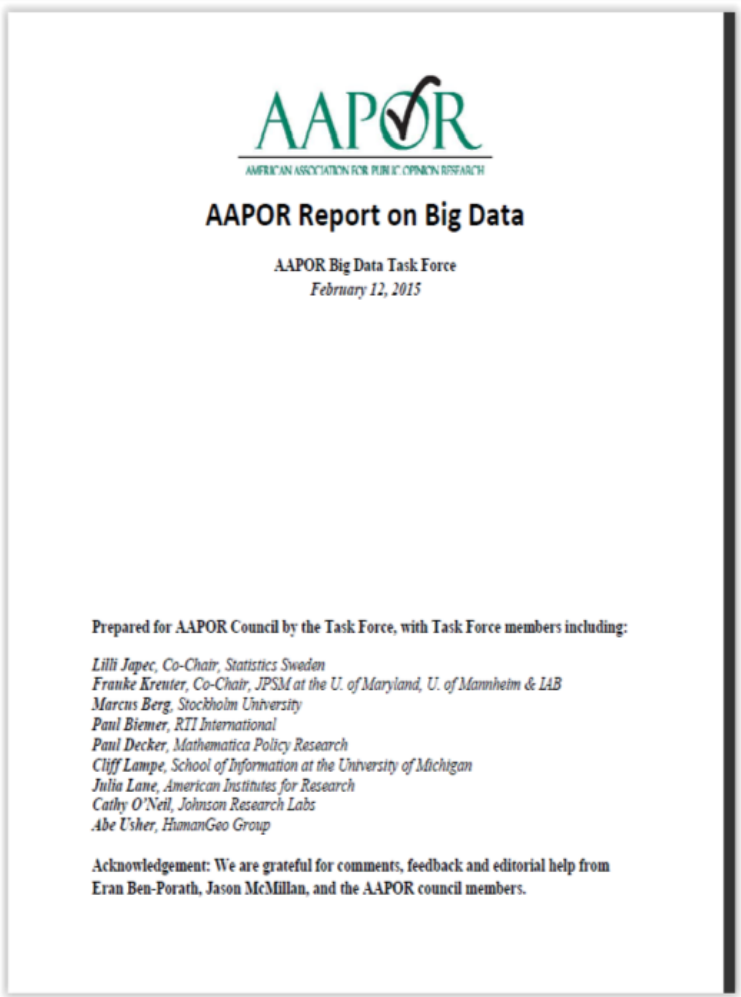
| | Inventory entries | Country | Type of Big Data used* | Domain** |
|---|---|---|---|---|
| 1 | Australia (ABS) - Social Linked (semantic) Data Processing for Various Statistical Uses | Australia | Data from public administration (2100) | Education (1.3); Health (1.4); Income and consumption (1.5); Labour (1.2); Population and migration (1.1) |
| 2 | Statistics Canada - Non-Residential Buildings Inventory: Feasibility Study | Canada | | Environment (3.1); Human settlements and housing (1.7); Population and migration (1.1); Prices (2.7) |

Powered by a free Atlassian Confluence Community License granted to UNECE. Evaluate Confluence today

# One day data


Source: Mostroem and Justesen, Statistics Sweden

# AAPOR Report on Big Data



AAPOR

AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

**AAPOR Report on Big Data**

AAPOR Big Data Task Force

*February 12, 2015*

**Prepared for AAPOR Council by the Task Force, with Task Force members including:**

*Lilli Japec, Co-Chair, Statistics Sweden*
*Franke Kreuter, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB*
*Marcus Berg, Stockholm University*
*Paul Biemer, RTI International*
*Paul Decker, Mathematica Policy Research*
*Cliff Lampe, School of Information at the University of Michigan*
*Julia Lane, American Institutes for Research*
*Cathy O'Neil, Johnson Research Labs*
*Abe Usher, HumanGeo Group*

**Acknowledgement: We are grateful for comments, feedback and editorial help from Eran Ben-Porath, Jason McMillan, and the AAPOR council members.**

**Paul Biemer**, RTI International
**Paul Decker**, Mathematica Policy Research
**Cliff Lampe**, University of Michigan
**Julia Lane**, New York University, CUSP
**Cathy OffNeil**, Author
**Abe Usher**, CTO Radiant Group and Digital Globe

# Committee on National Statistics

Panel 2015-2017: Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-Art Estimation Methods

**Robert Graves (chair),** *Georgetown University*

Committee on National Statistics

The National Academies of SCIENCES · ENGINEERING · MEDICINE

**Michael E. Chernew**, Harvard
**Piet Daas**, Statistics Netherlands
**Cynthia Dwork**, Microsoft Research
**Ophir Frieder**, Georgetown U.
**Hosagrahar Jagadish**, UMich
**Frauke Kreuter**, UMD & UMannheim
**Sharon Lohr**, Westat
**James P. Lynch**, UMD
**Colm O'Muircheartaigh**, U. Chicago
**Trivellore Raghunathan**, UMich
**Roberto Rigobon**, MIT
**Marc Rotenberg**, EPIC

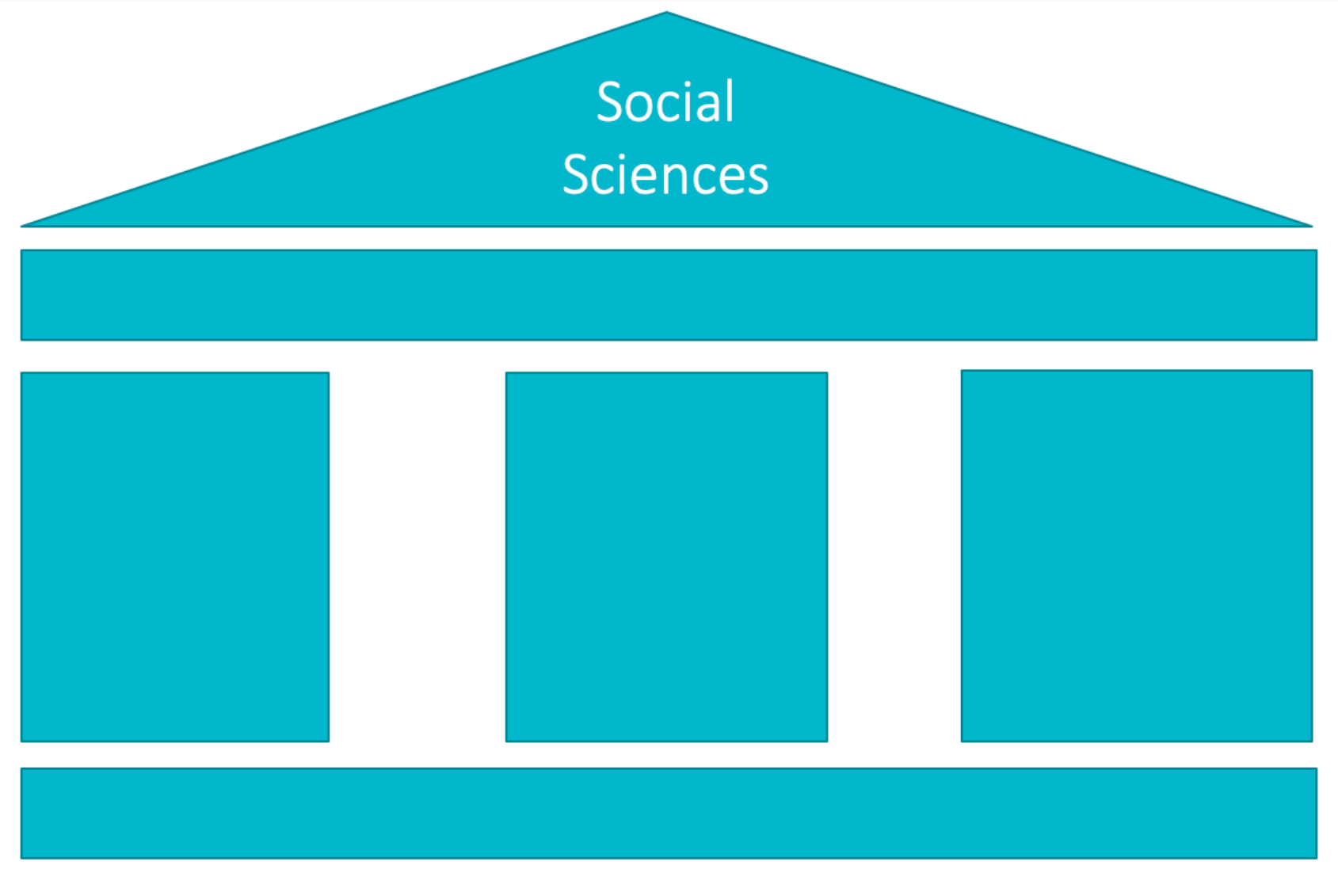
# Excitement over advantages

1. New research questions can be asked
    - spatial and temporal granularity
    - small parts of the populations
    - other form of data (text, visuals)
2. Reduced data collection costs
3. 'Instant' more timely availability

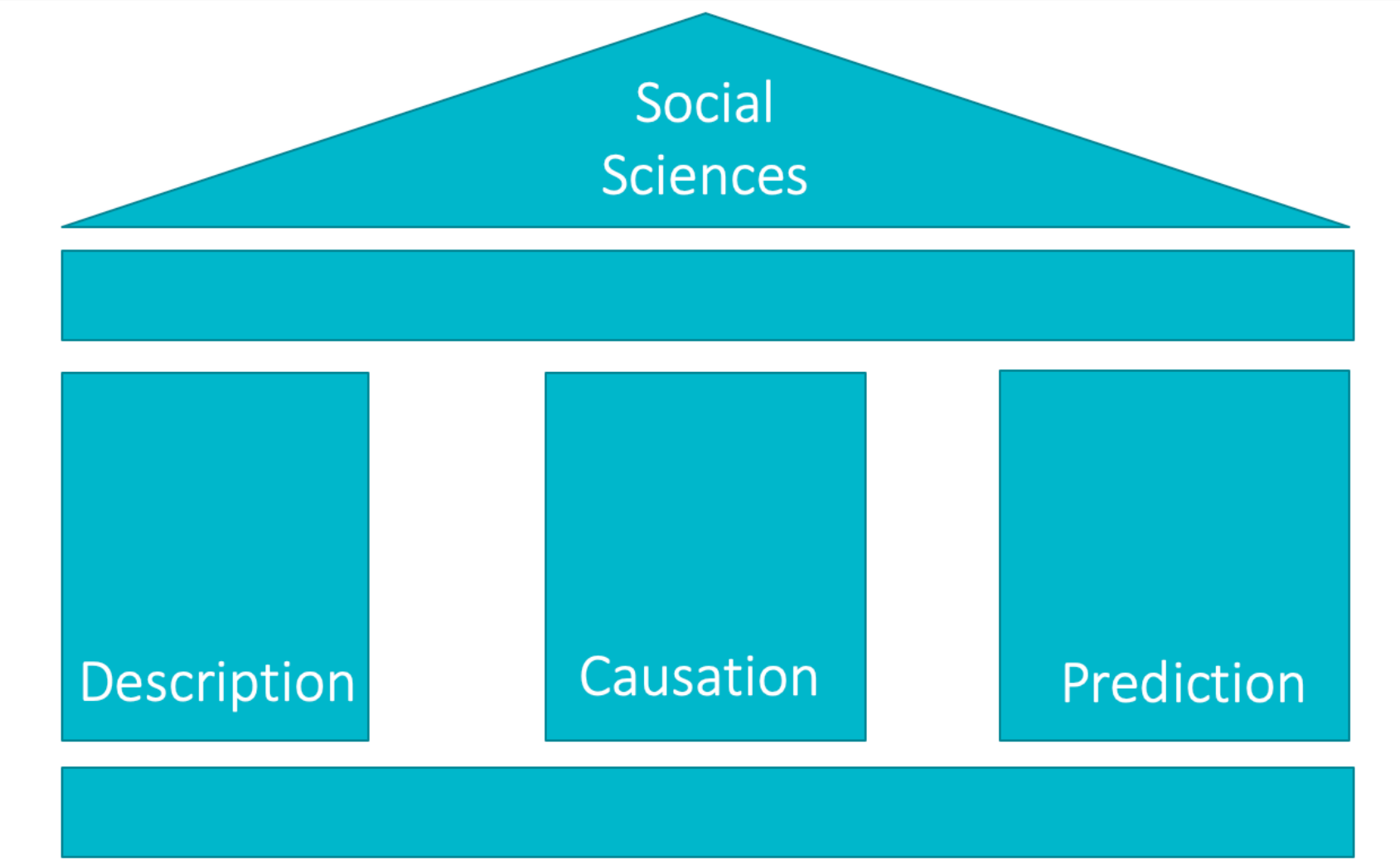
# Concerns over scientific value

1. Measurement
   Proxies and variable poor (Couper 2013)
2. Inference

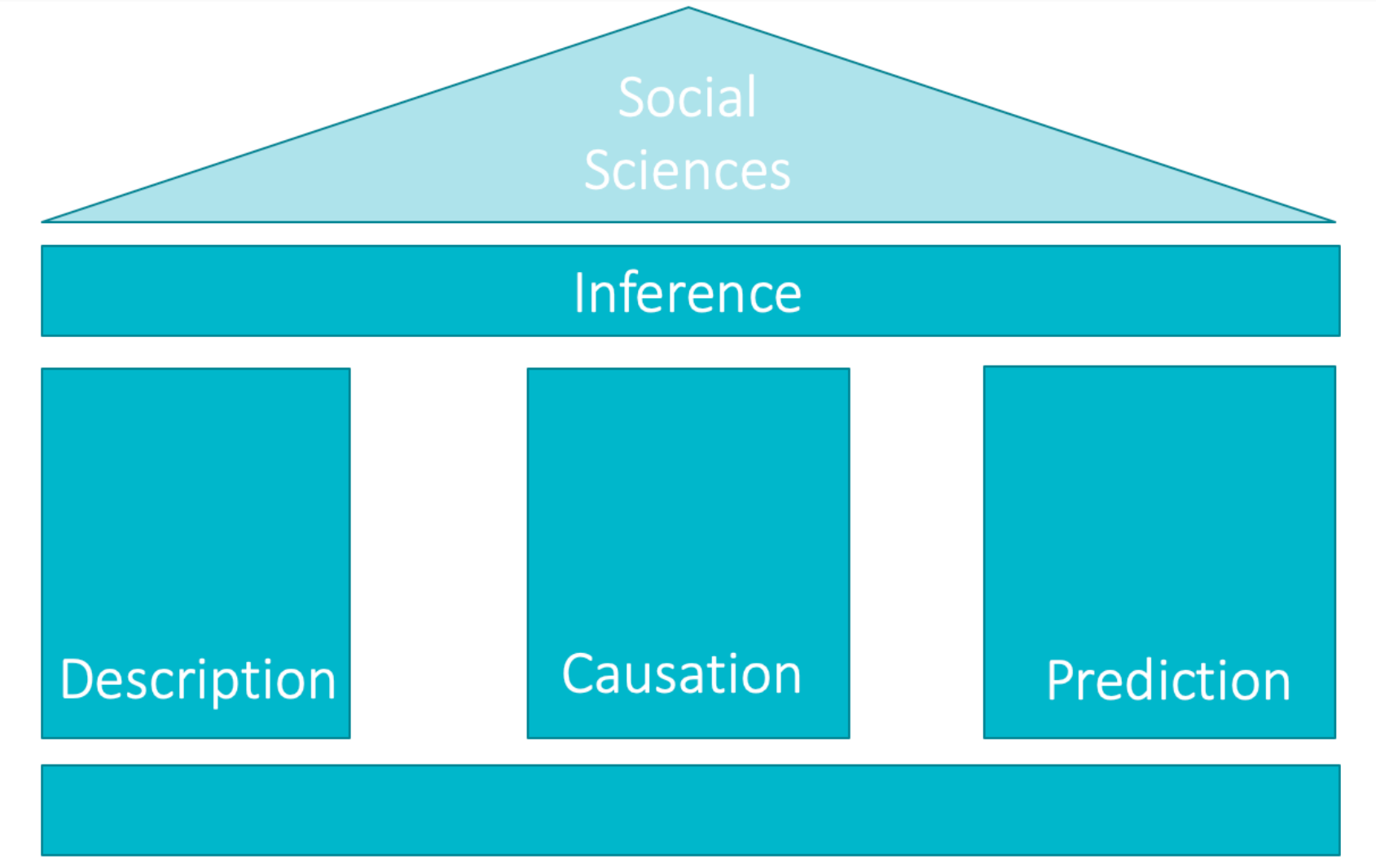
# The Social Science-House



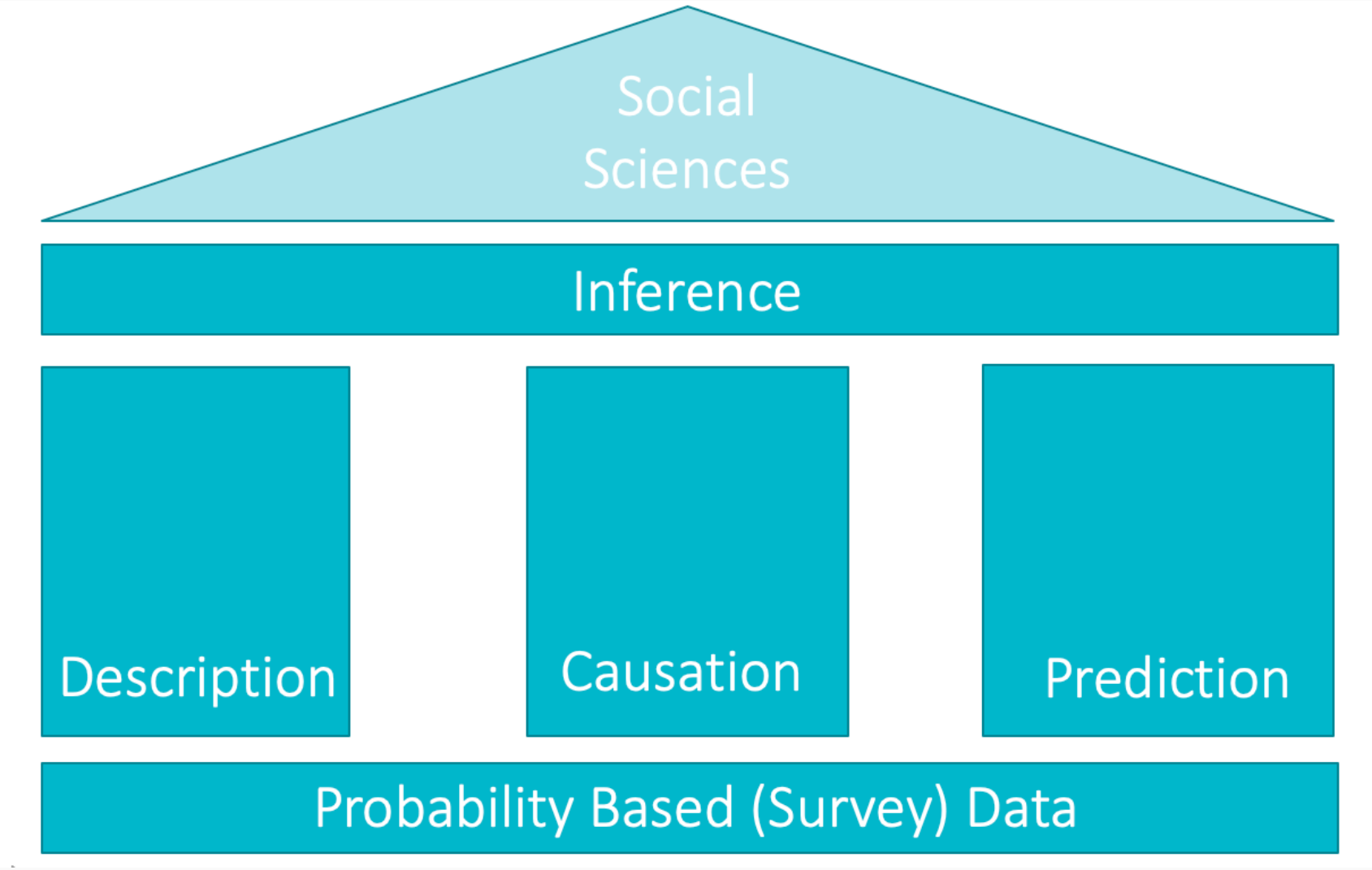
# The Social Science-House



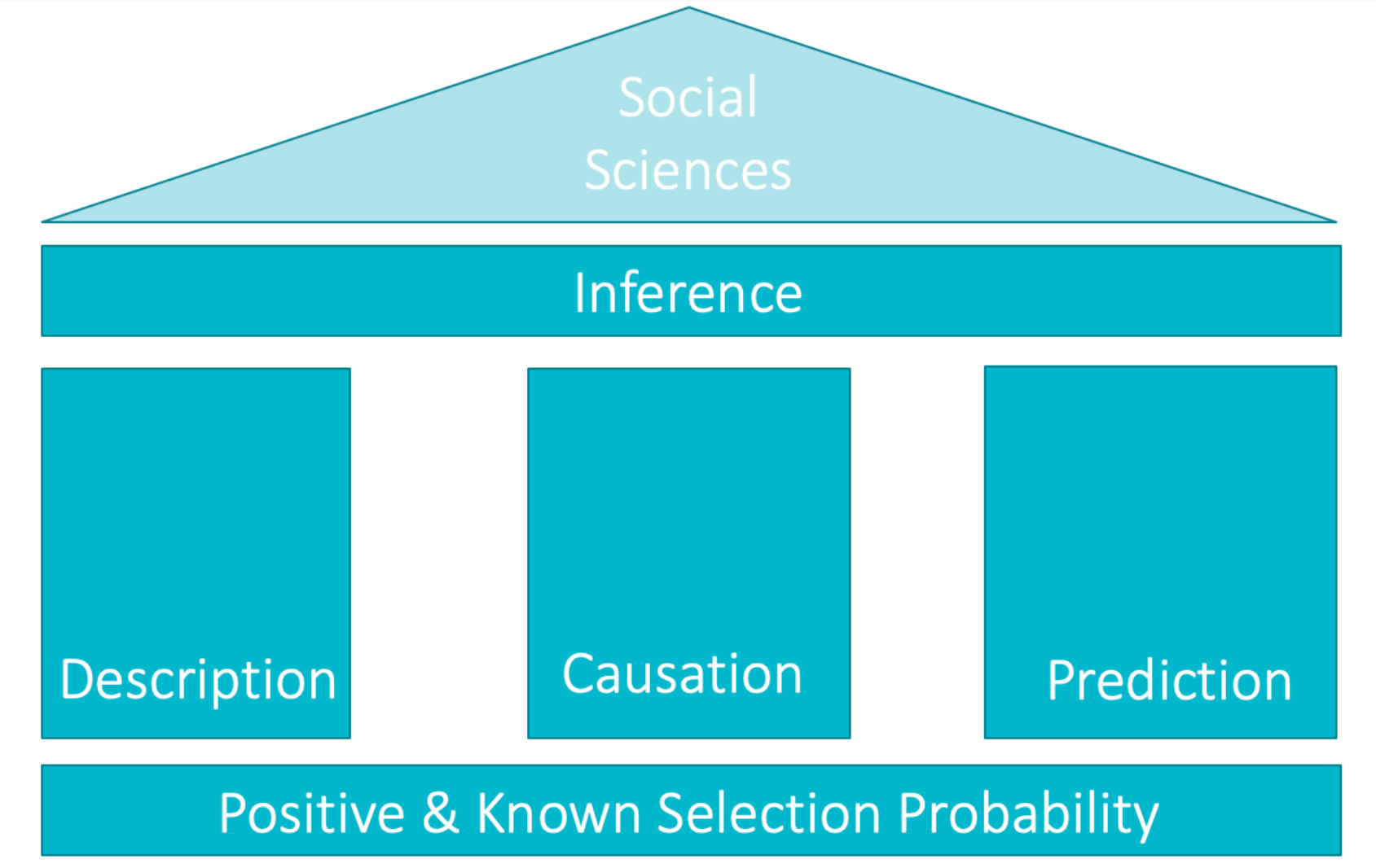
# The Social Science-House



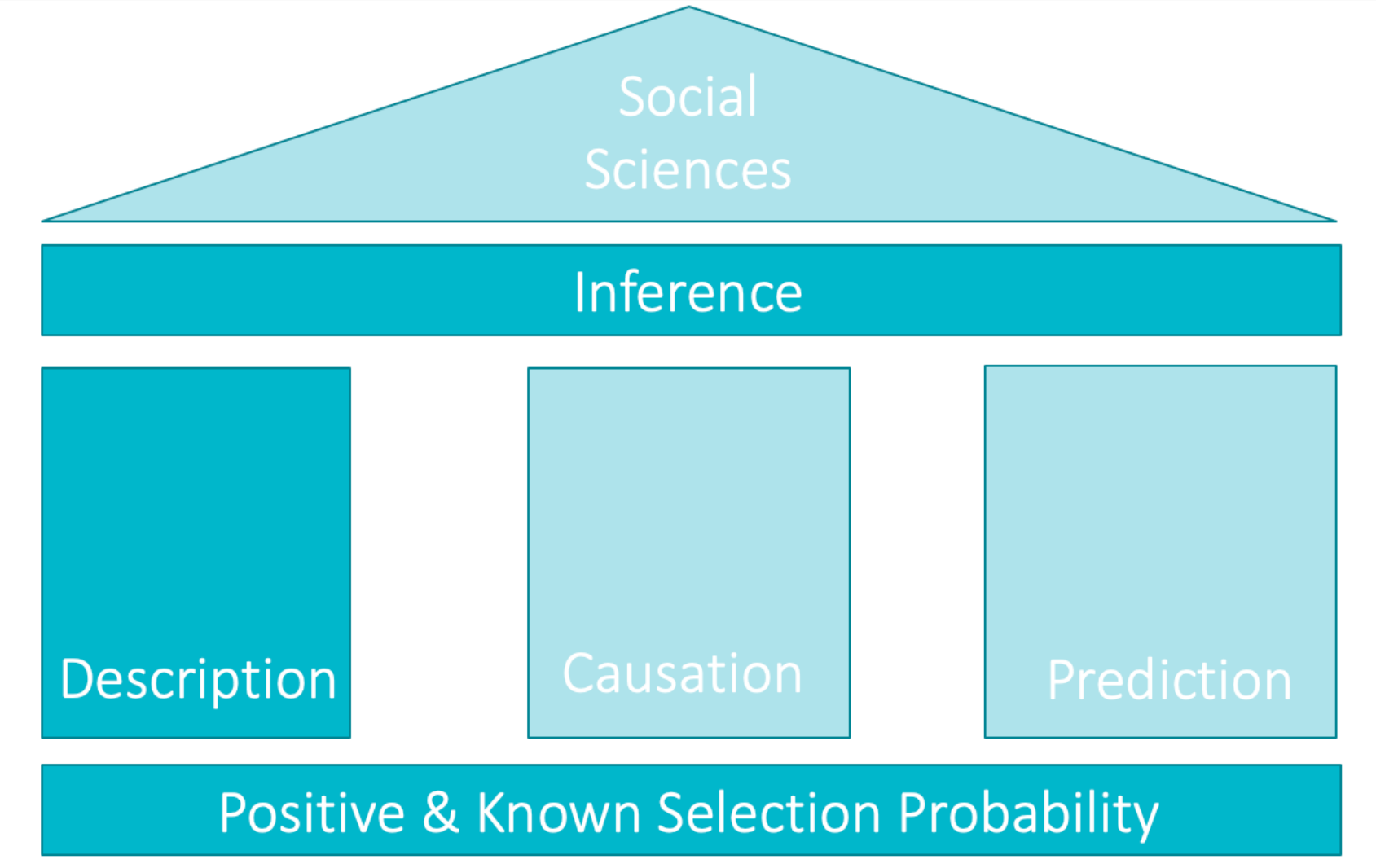
# The Social Science-House



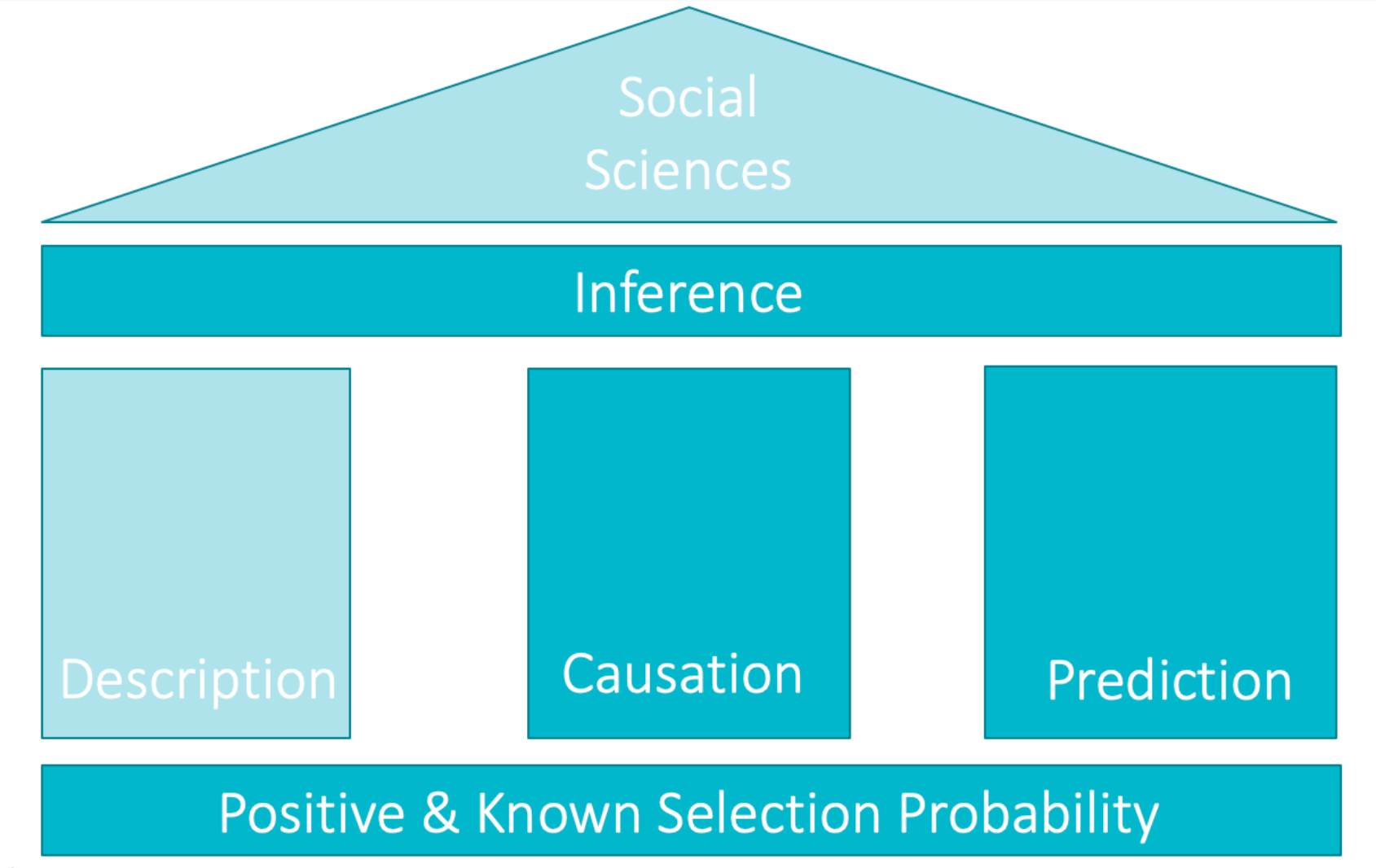
# The Social Science-House



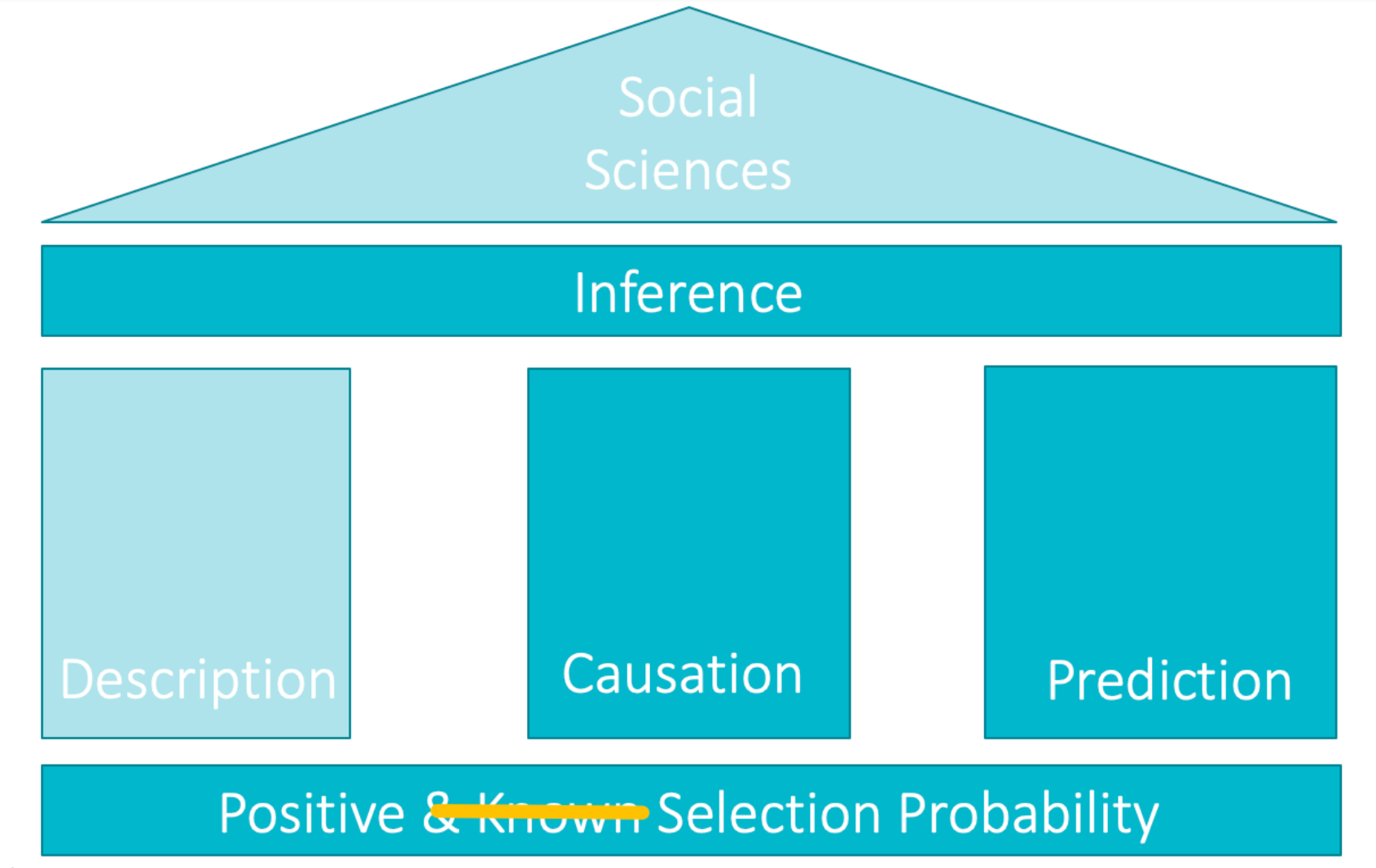
# The Social Science-House



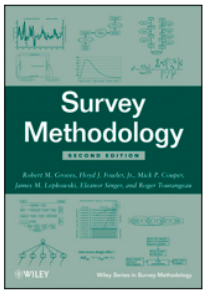
# The Social Science-House

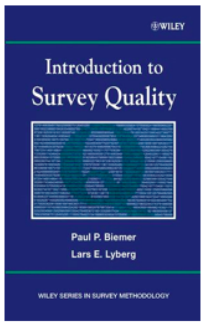


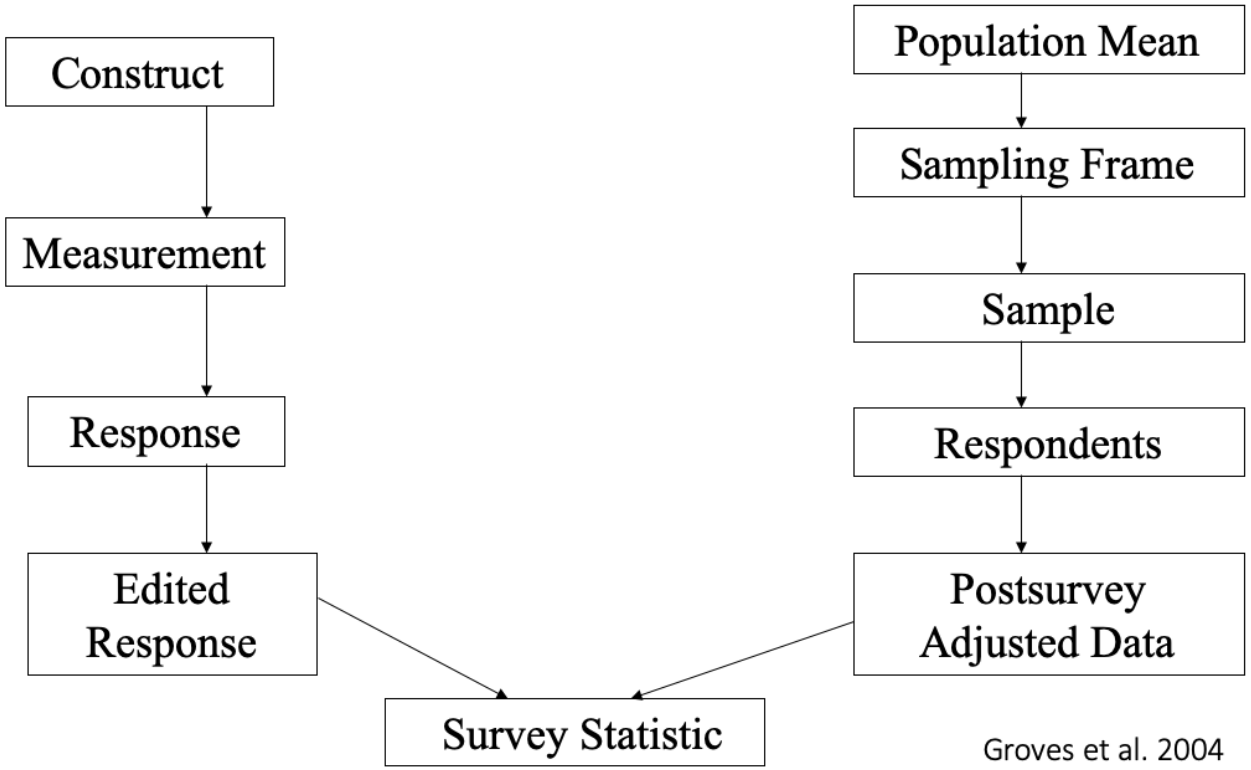
# The Social Science-House



# Data Generating Process

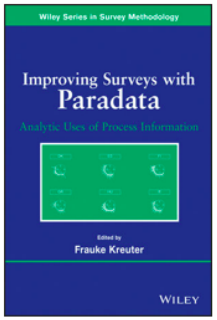


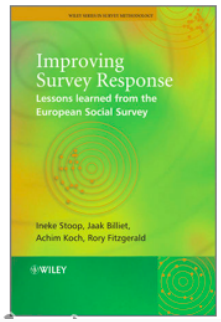



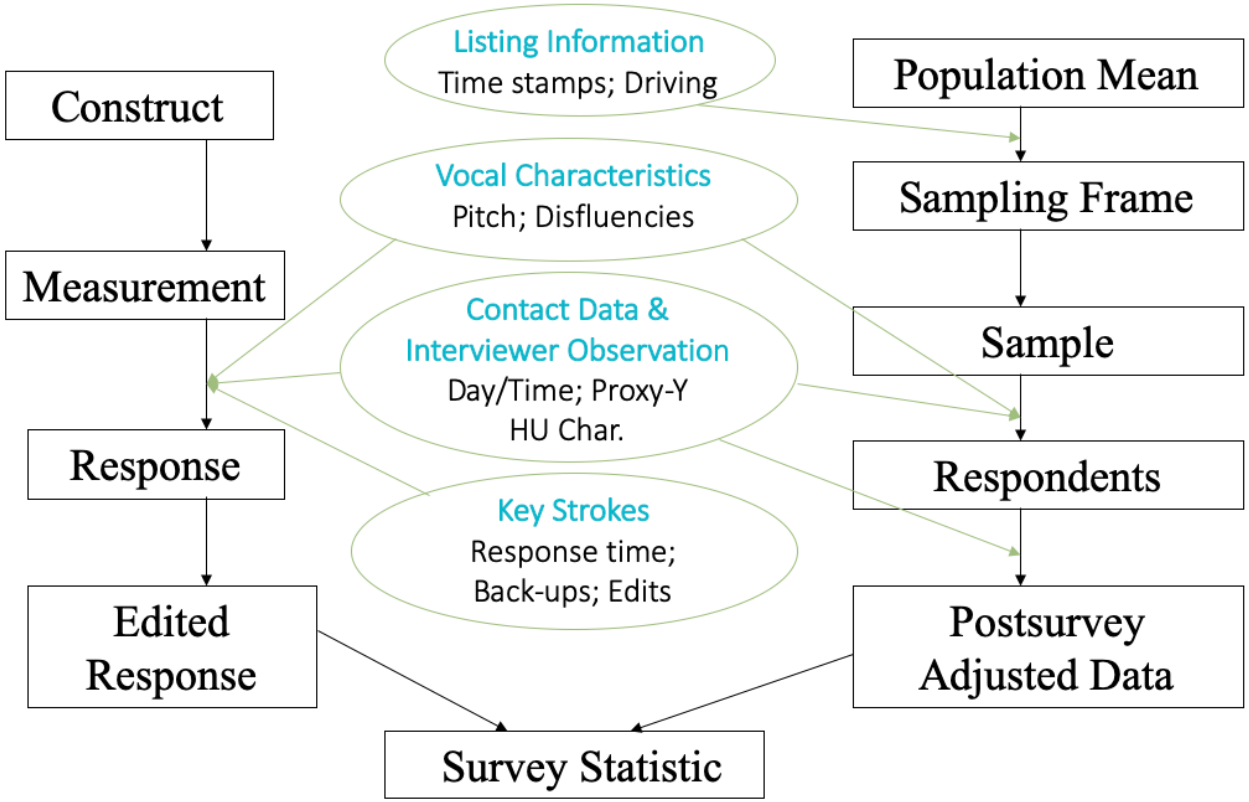




Groves et al. 2004

# Data Generating Process



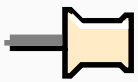



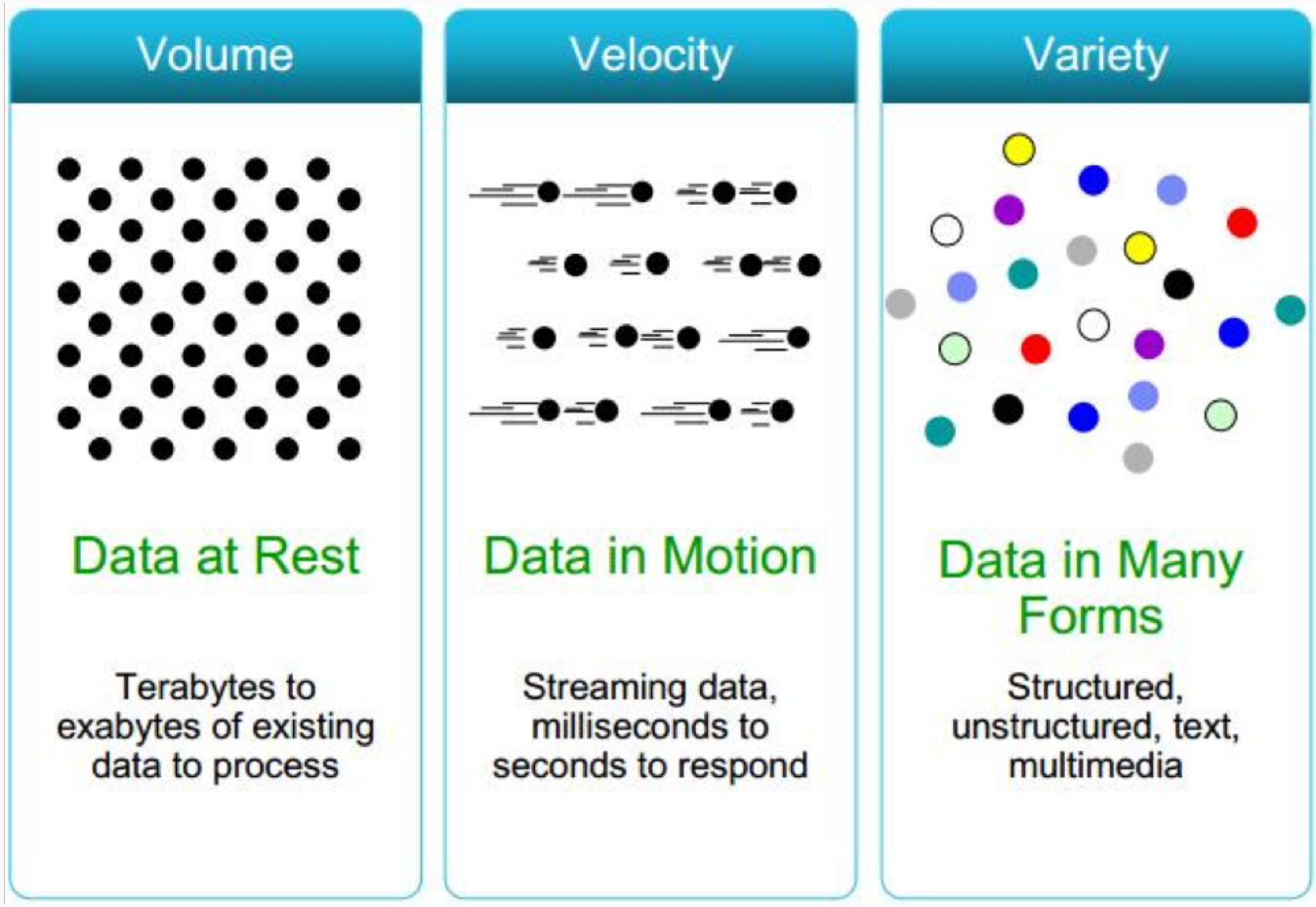
# Key Ingredients for Valid Inference

1. Data generating process needs to be known
2. Framework as tool to identify errors
3. Model or break confounders
4. Know your inferential goal

# “Found” – Boston Street Bumps

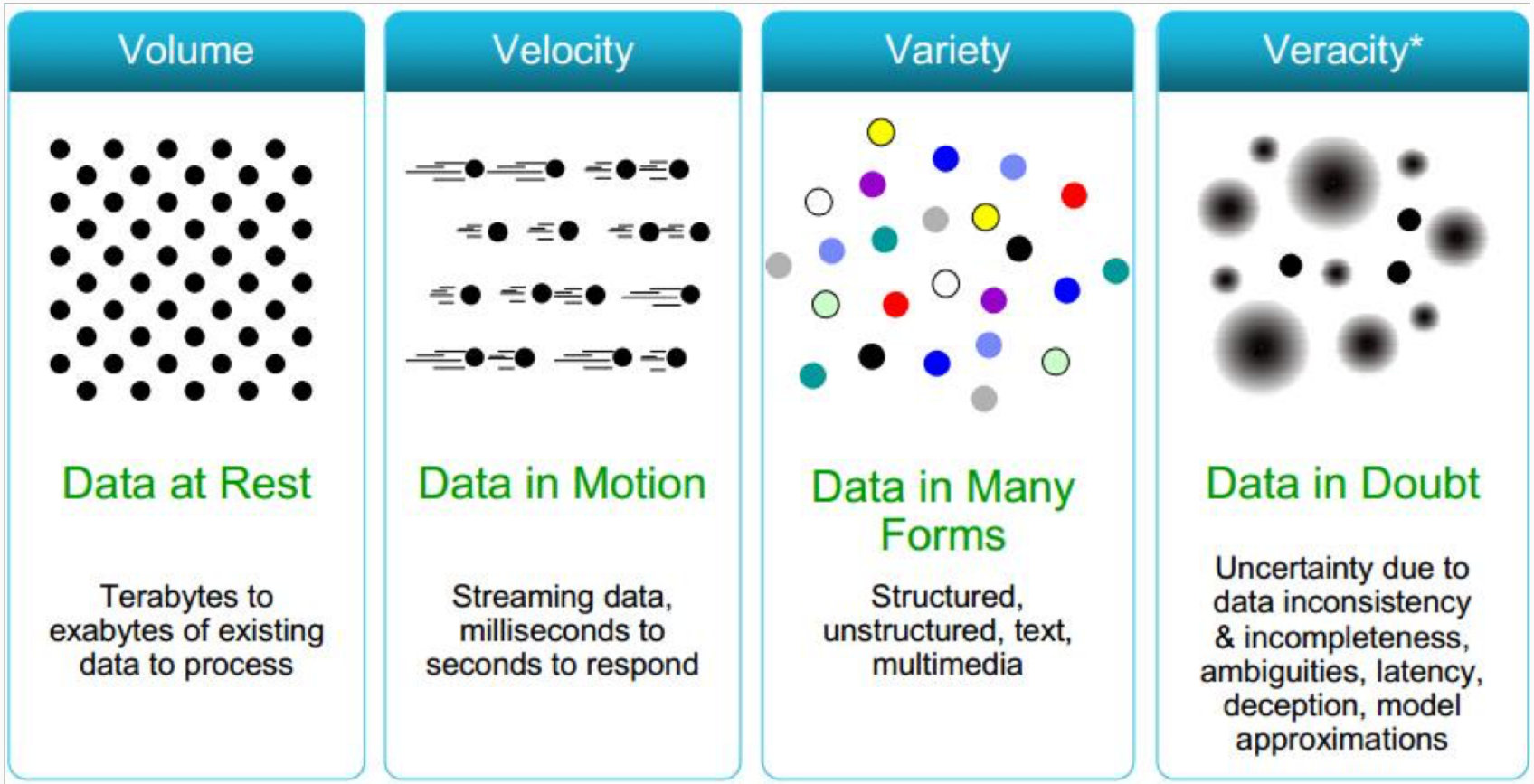


# Found/Organic Data



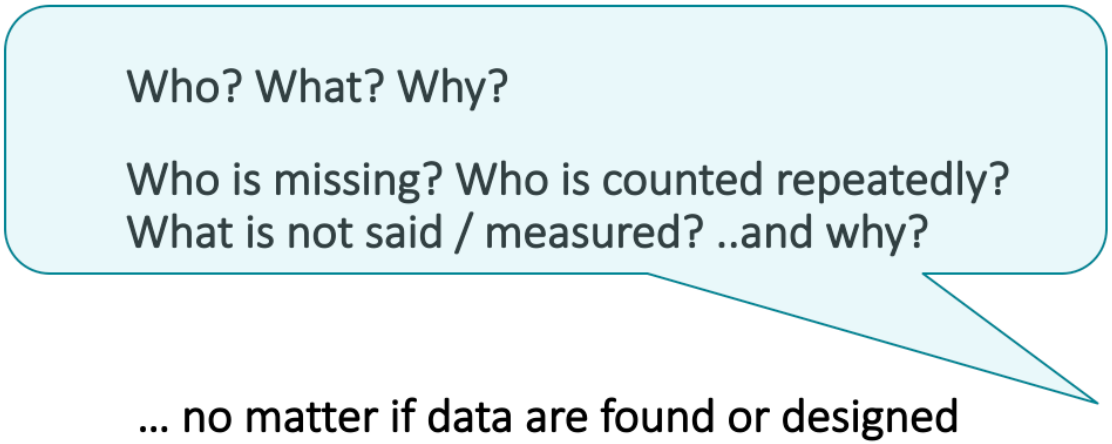

Source: http://www.rosebt.com/blog/data-veracity

# Found/Organic Data




Source: http://www.rosebt.com/blog/data-veracity

# Data Generating Process



# Data Generating Process

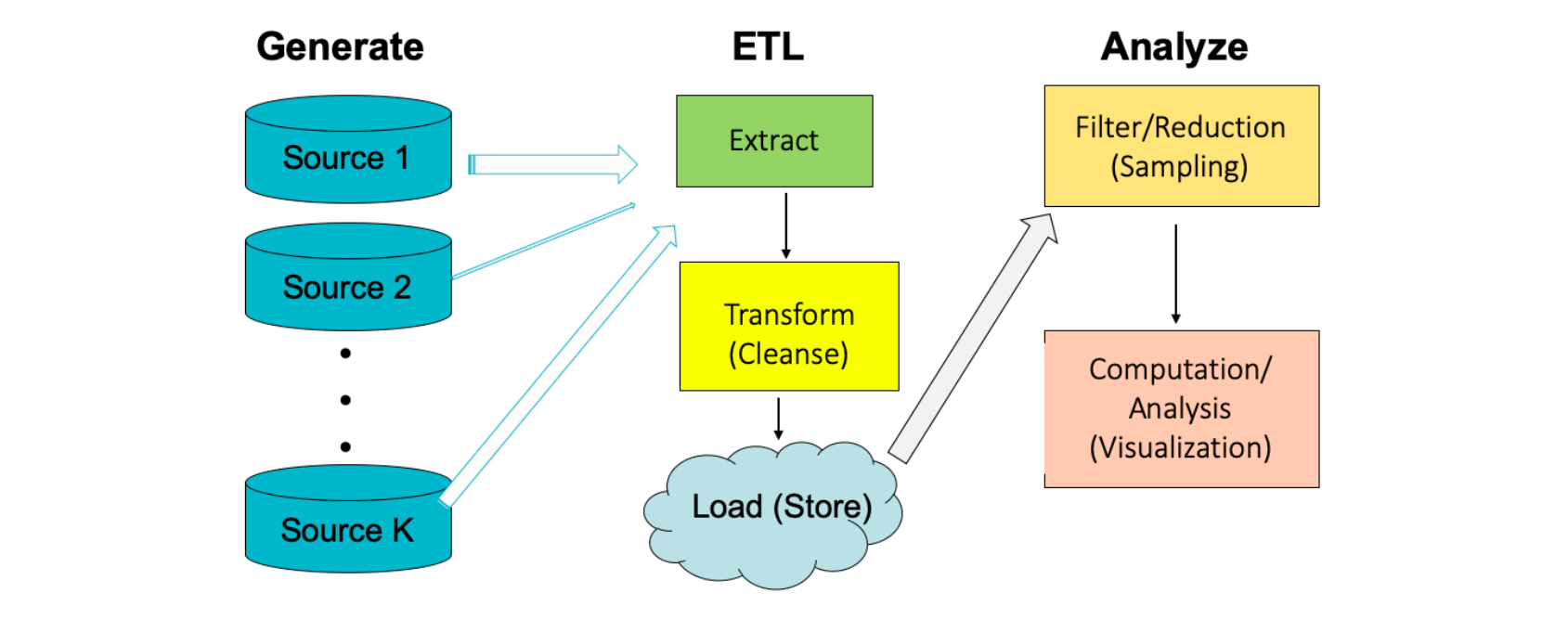
Who? What? Why?

Who is missing? Who is counted repeatedly?
What is not said / measured? ..and why?

... no matter if data are found or designed

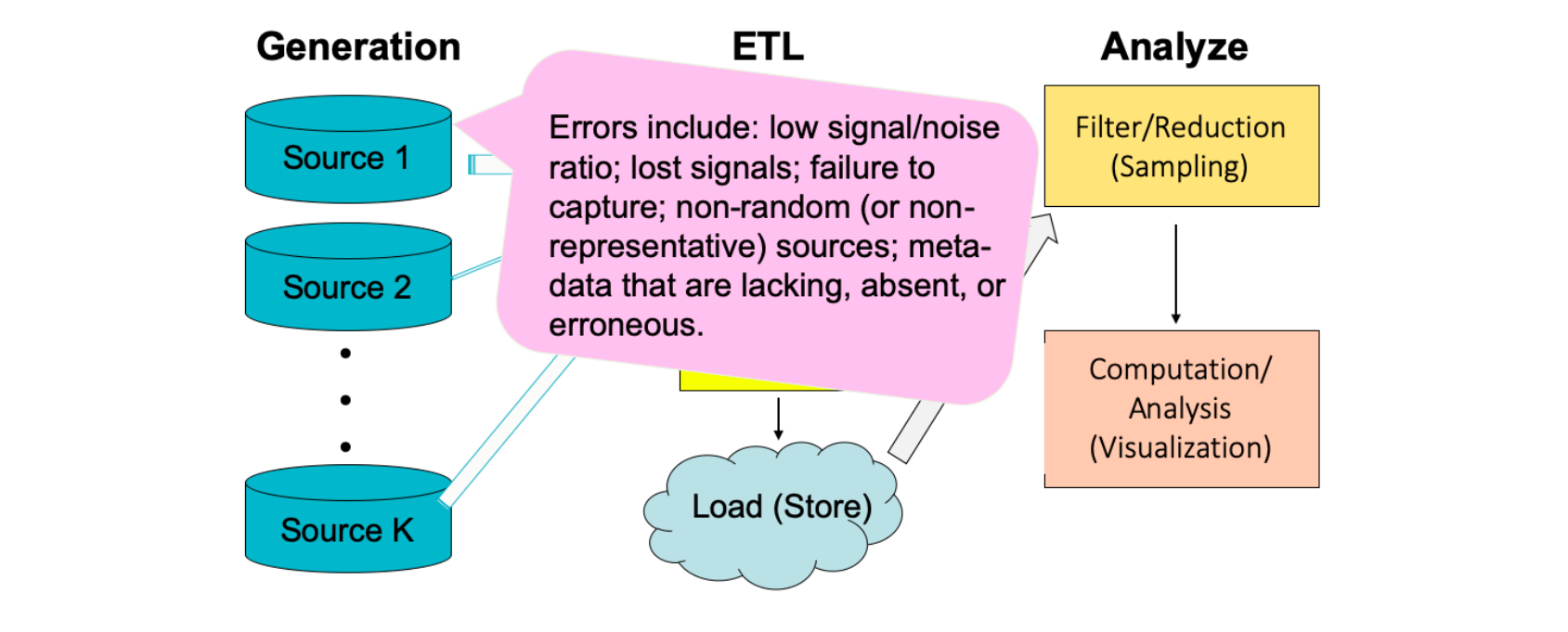
AAPOR Transparency Initiative
AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

# Big Data Process Map



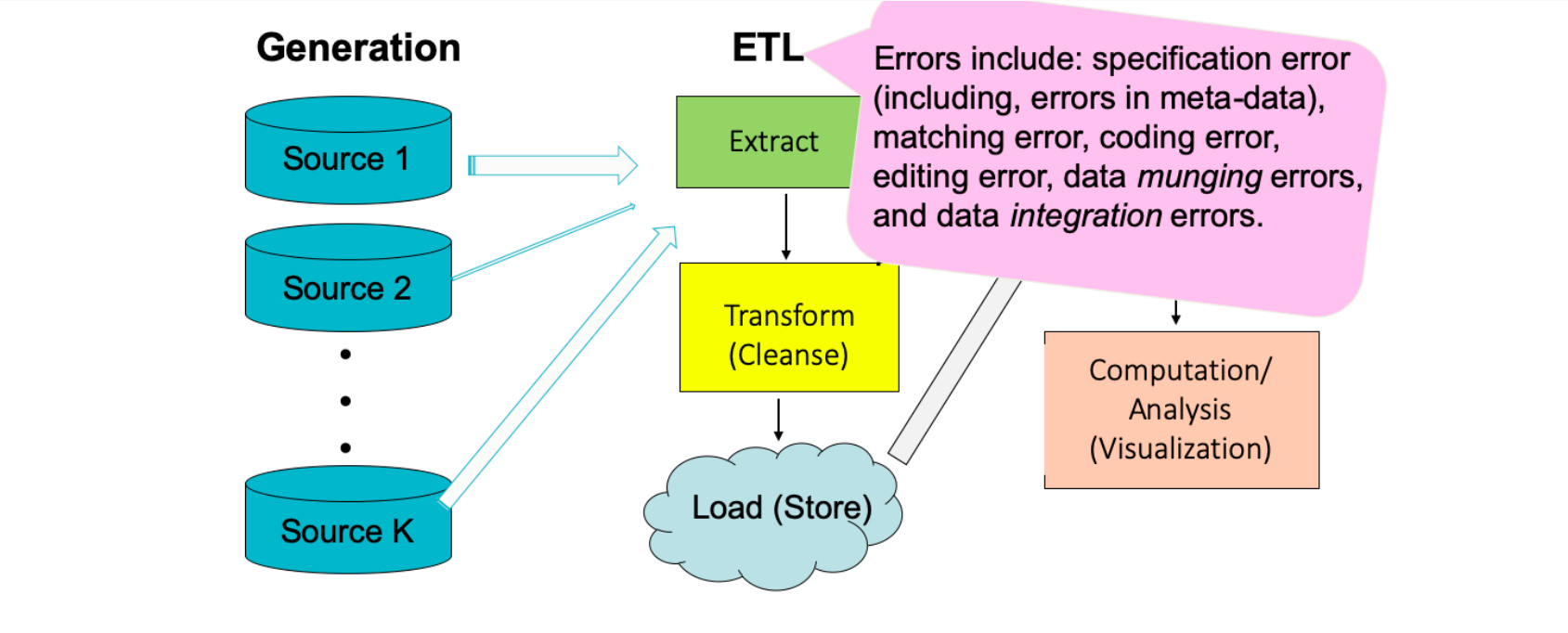

Source: Paul Biemer

# Big Data Process Map



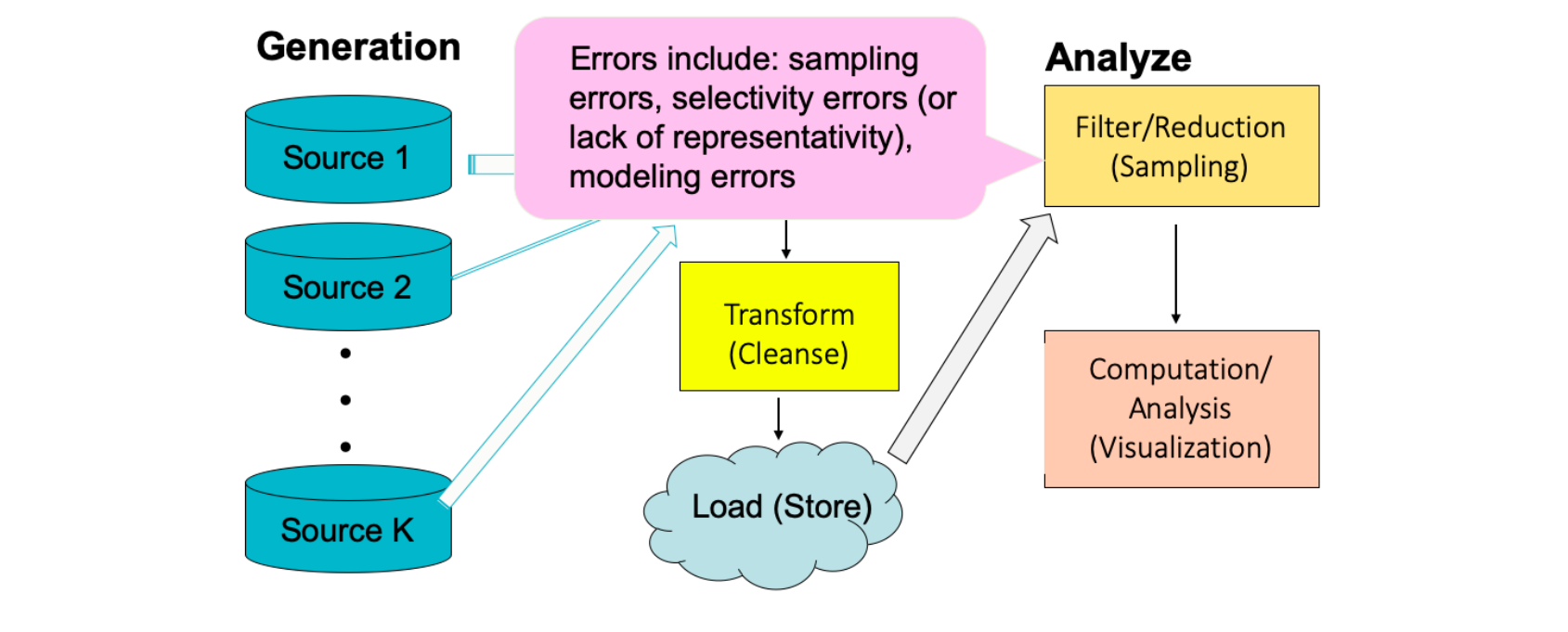

Source: Paul Biemer

# Big Data Process Map



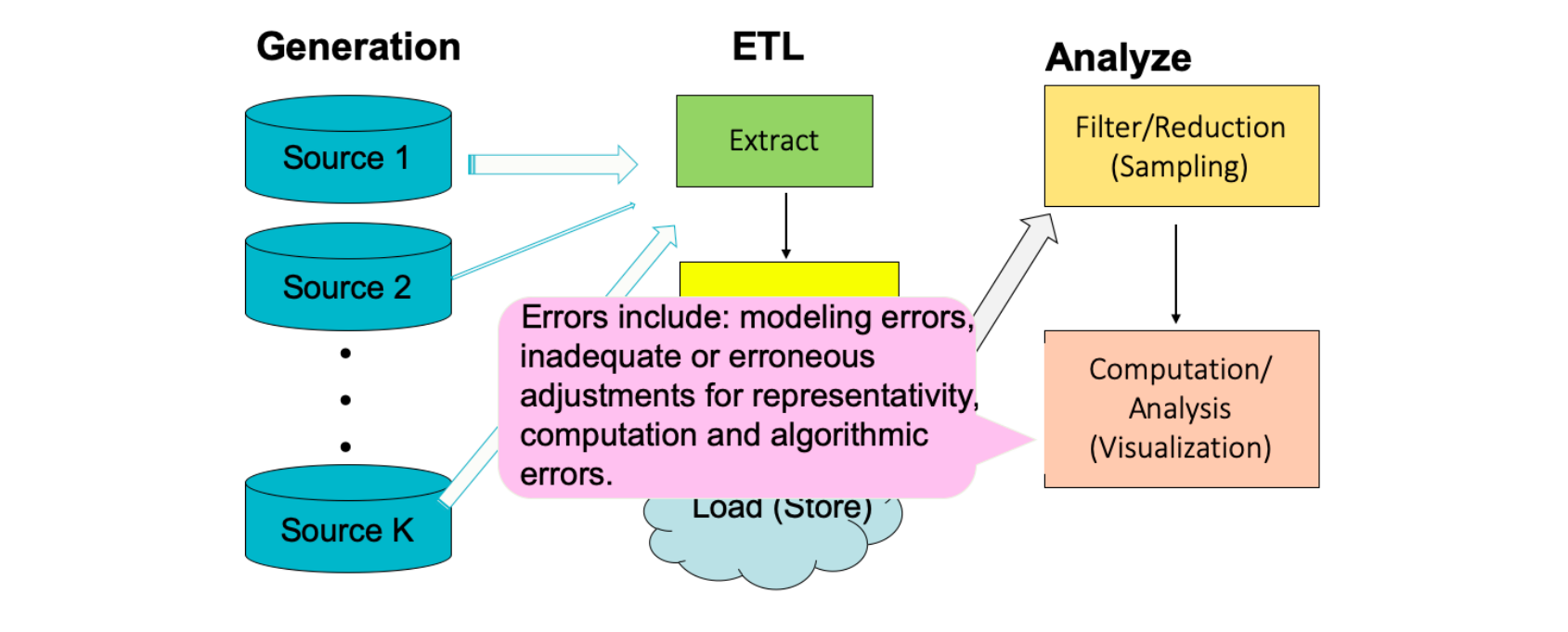

Source: Paul Biemer

# Big Data Process Map




Source: Paul Biemer

# Big Data Process Map

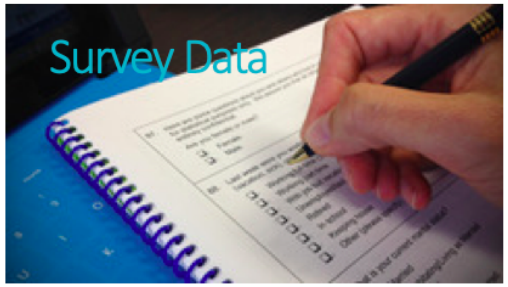



Source: Paul Biemer

# Scientific value can be high

1. If data generating process is known
2. If framework available to identify errors
3. If efforts are made to model or break confounders
4. If inferential goal is carefully kept in mind
5. Note: Cost often not lower (after processing)

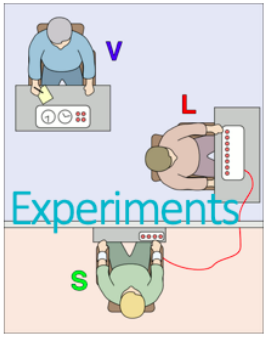
# Combining Data





*Real value is in combined data products.*

Problems:
Access
Consent



# Conclusion

1. Need vibrant and collaborative research programs to investigate **possibilities for combining** different public sector data as well as private data sources **while protecting privacy**
2. Research programs needed to **investigate the quality** of the data sources and **shed light on data collection processes**. This will inform issues of inference (presence of confounders and the possibility to take them into account through modeling)

# Conclusion

3. Need to **investigate alternative approaches** to informed (linkage) consent
4. More than ever will we need to form **interdisciplinary teams** to understand and model the data collection processes
5. We need to **build capacity among current employees** and researchers both in public and private sector