Recent Developments in Federal Statistics – Surveys and Beyond

Frauke Kreuter

JPSM – Uni Mannheim – IAB

Melbourne June 26 ‘18

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.

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

REPORT

INNOVATIONS IN FEDERAL STATISTICS

Combining Data Sources While Protecting Privacy

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

CONSENSUS STUDY REPORT

FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION

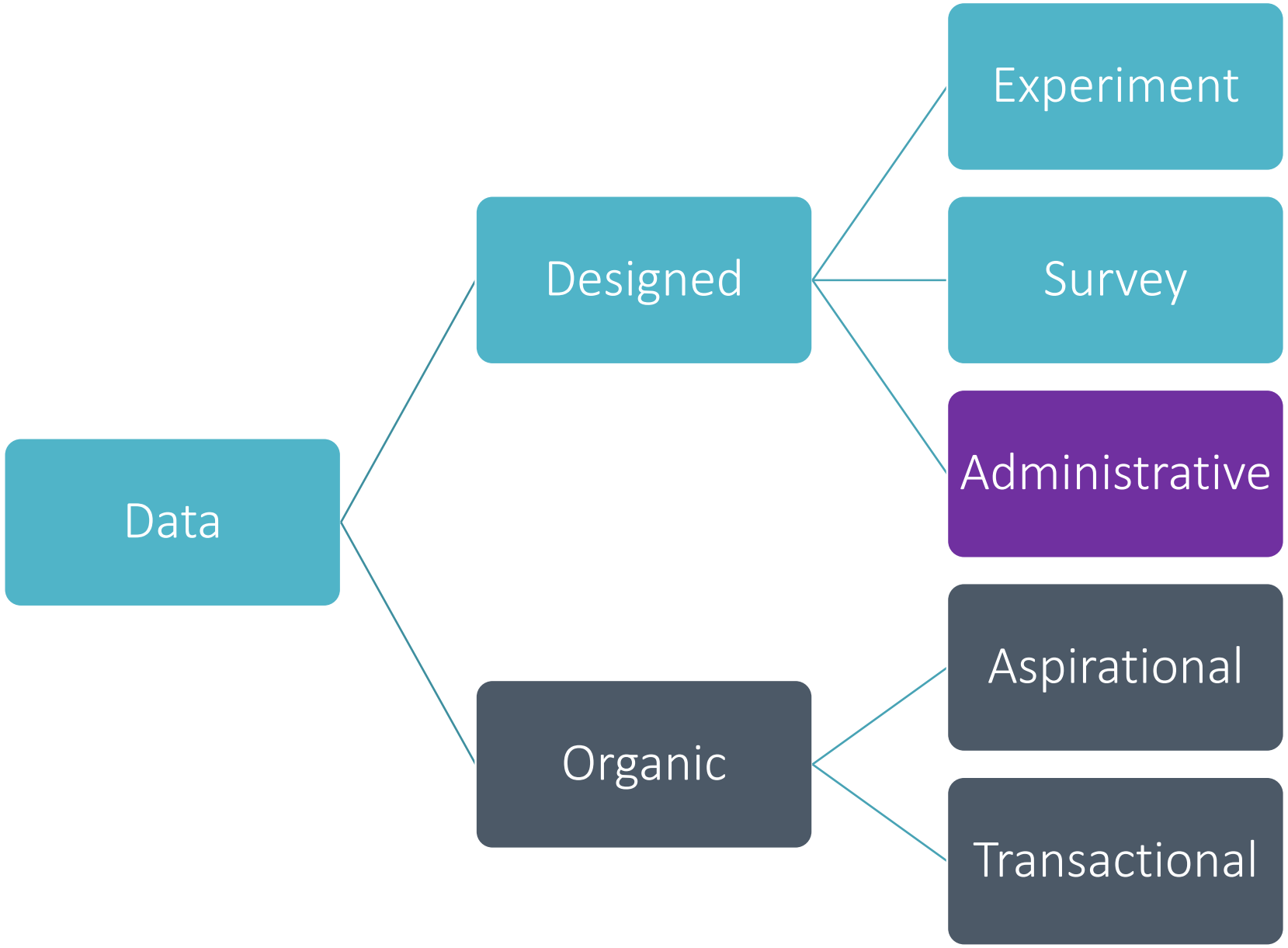
Next Steps

SPONSORED BY THE

Federal Ministry of Education and Research

AUFSTIEG DURCH BILDUNG
OFFENE HOCHSCHULEN

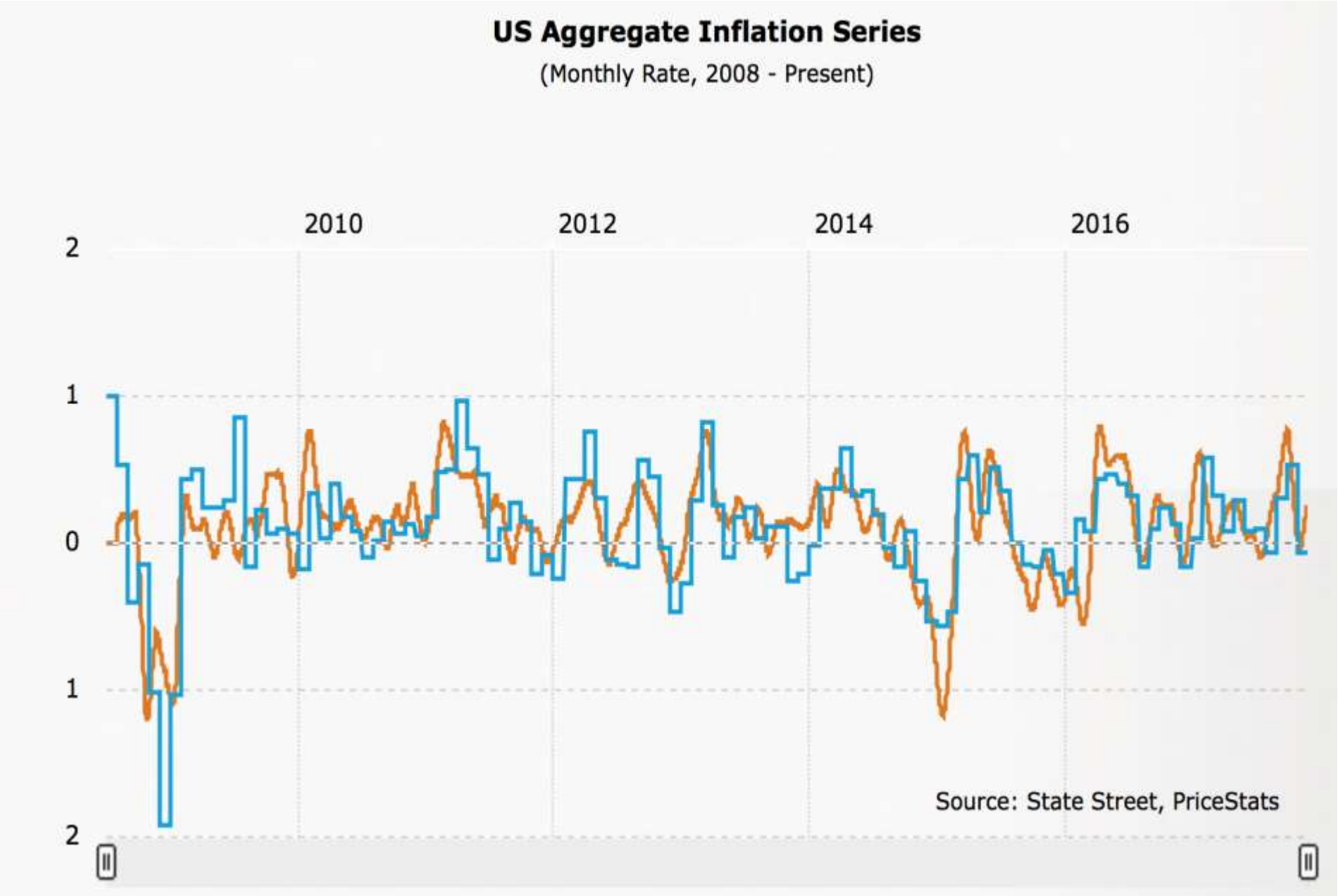
The project on which this report is based was funded by the Federal Ministry of Education and Research under the number [16OH22064]. Responsibility for the contents of this publication lies with the author.



Source: Roberto Rigobon, [Discussion on Applications and Issues with Using Commercial Data in Research](), BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics November 19, 2015

Economic Indicators

Examples for online data collection (and analysis)

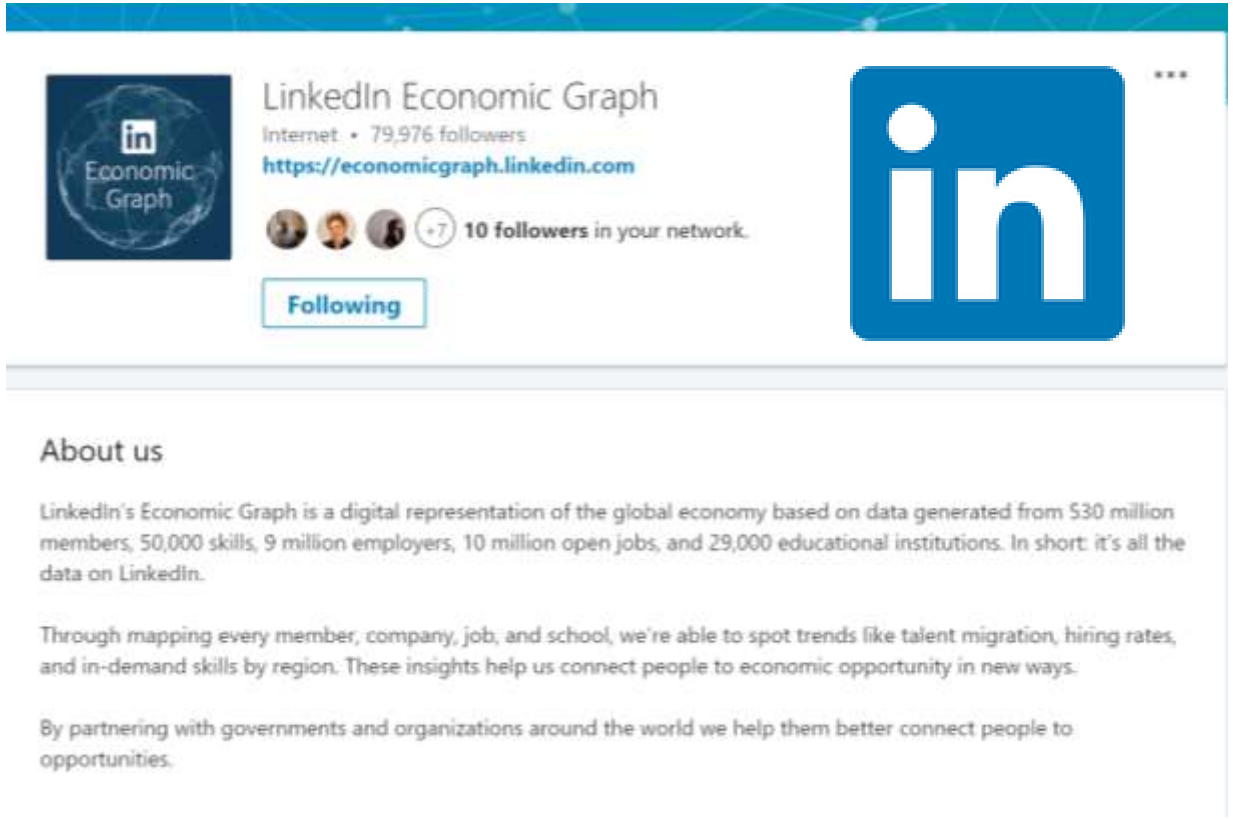


Official CPI

PriceStats Index

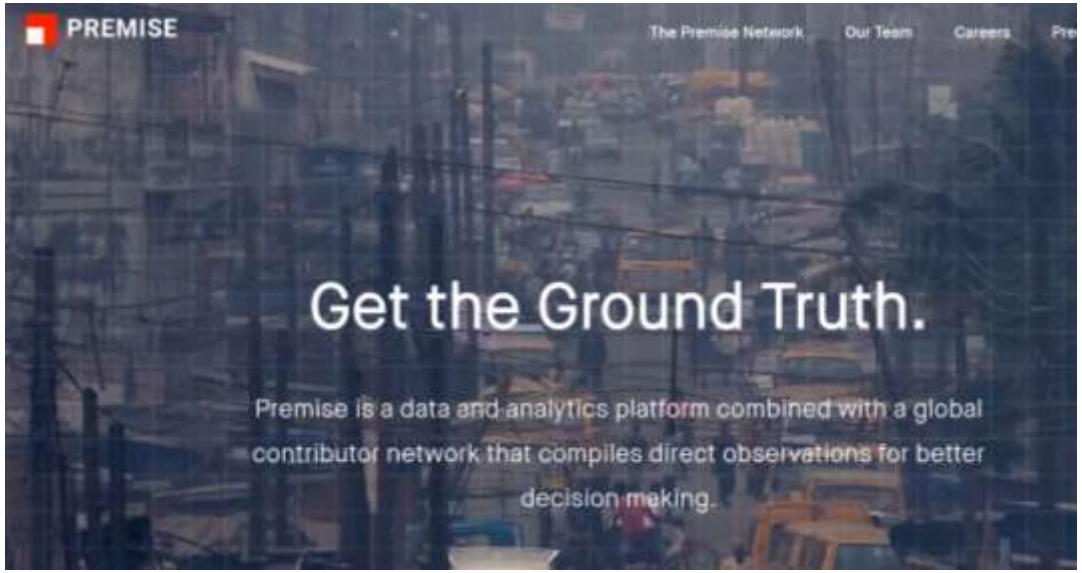
US Aggregated Inflation Series, Monthly Rate, PriceStats Index vs. Official CPI the PriceStats website 1/28/2018

Employment Websites + Aggregators



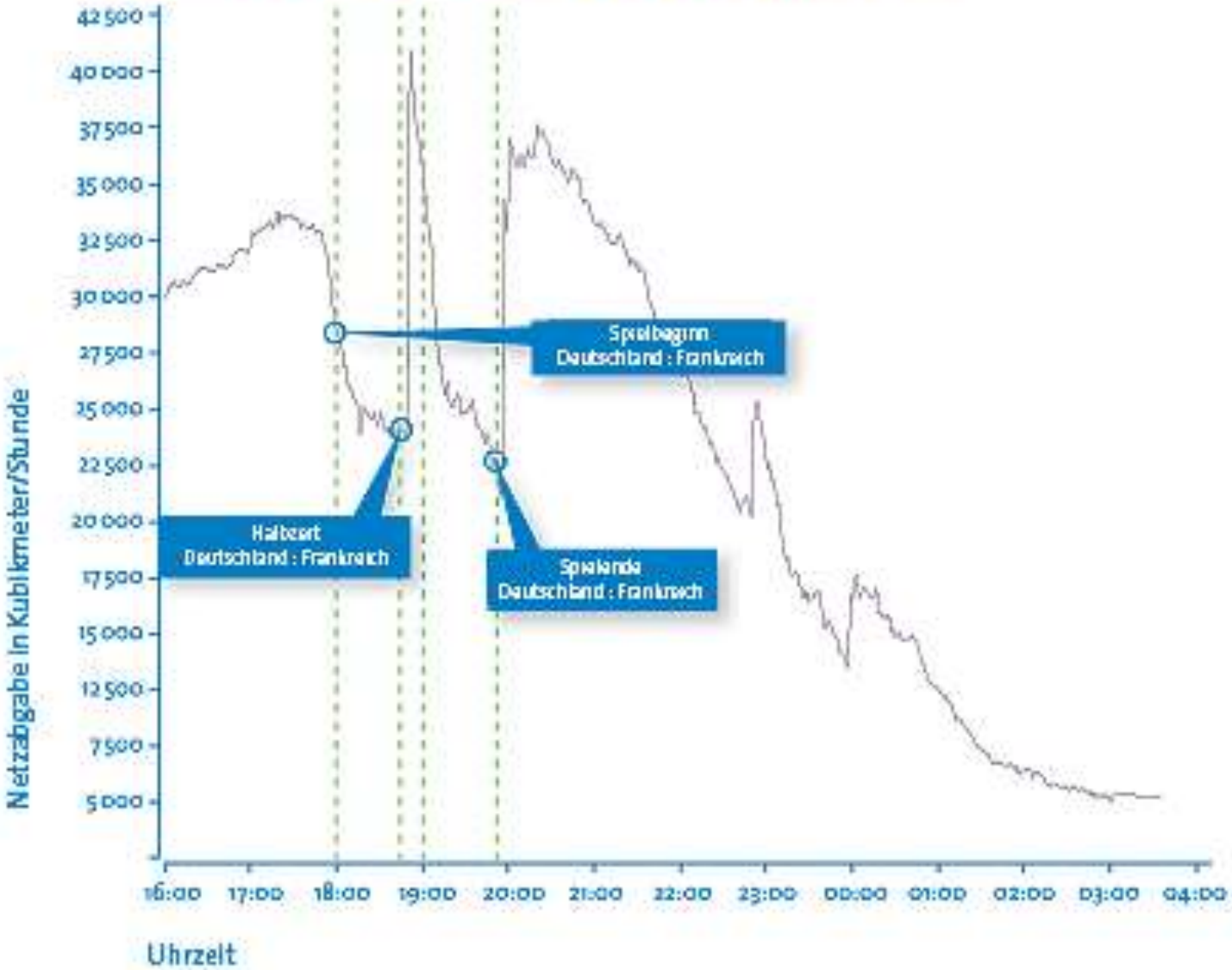
WANTEDANALYTICS
a CEB Company

Platforms - Crowdsourcing



Observations

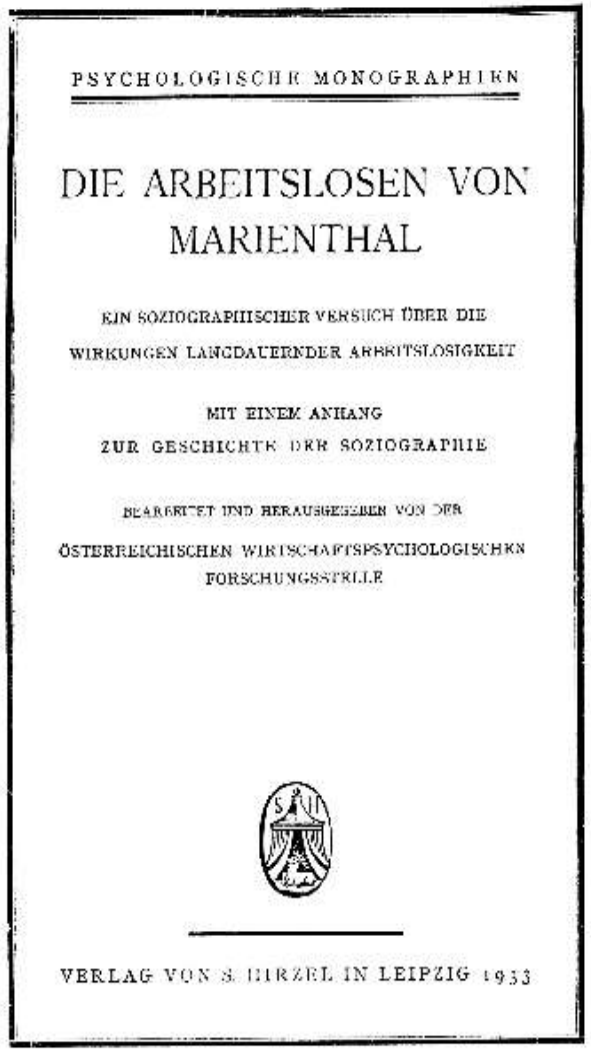
Netzabgabe für den 4. Juli 2014



1. IAB – Research Examples

1st Refugees

2nd Unemployment



PSYCHOLOGISCHE MONOGRAPHIEN

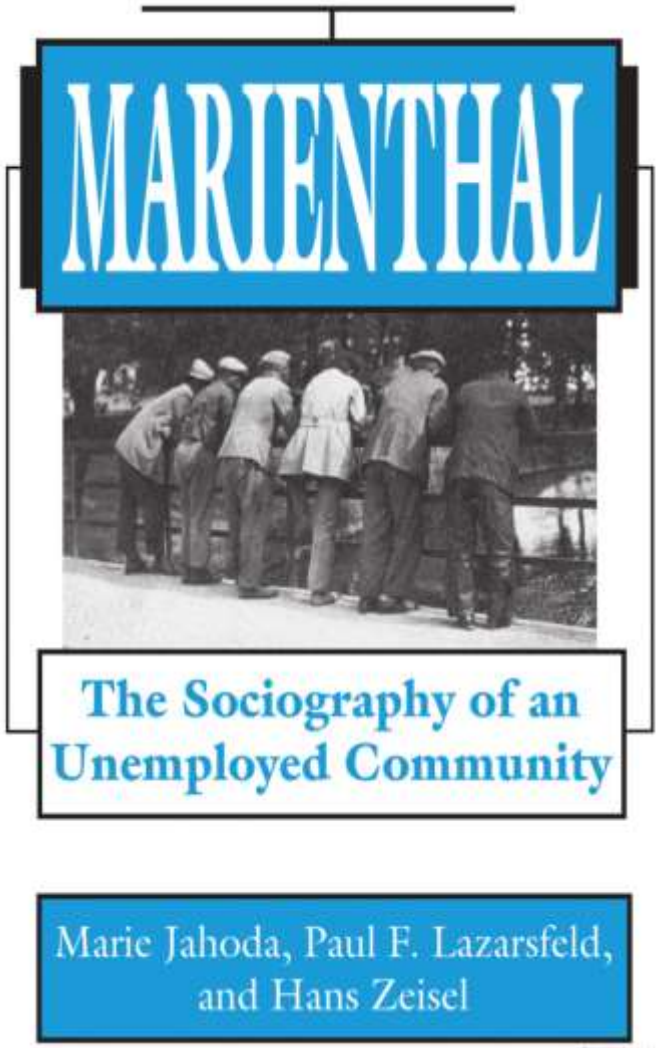
DIE ARBEITSLOSEN VON MARIENTHAL

EIN SOZIOGRAPHISCHER VERSUCH ÜBER DIE WIRKUNGEN LANGDAUERNDER ARBEITSLOSIGKEIT

MIT EINEM ANHANG ZUR GESCHICHTE DER SOZIOGRAPHIE

BEARBEITET UND HERAUSGEGEBEN VON DER ÖSTERREICHISCHEN WIRTSCHAFTSPSYCHOLOGISCHEN FORSCHUNGSSTELLE

VERLAG VON S. HIRZEL IN LEIPZIG 1933



IAB·Forum
Das Magazin des Instituts für Arbeitsmarkt- und Berufsforschung

9. Januar 2018 | Projekte

IAB-SMART-Studie: Mit dem Smartphone den Arbeitsmarkt erforschen

Autoren

Sebastian Bähr

Georg-Christoph Haas

Florian Keusch

Frauke Kreuter

Mark Trappmann

PASS – Panel (10 years) + Administrative Data

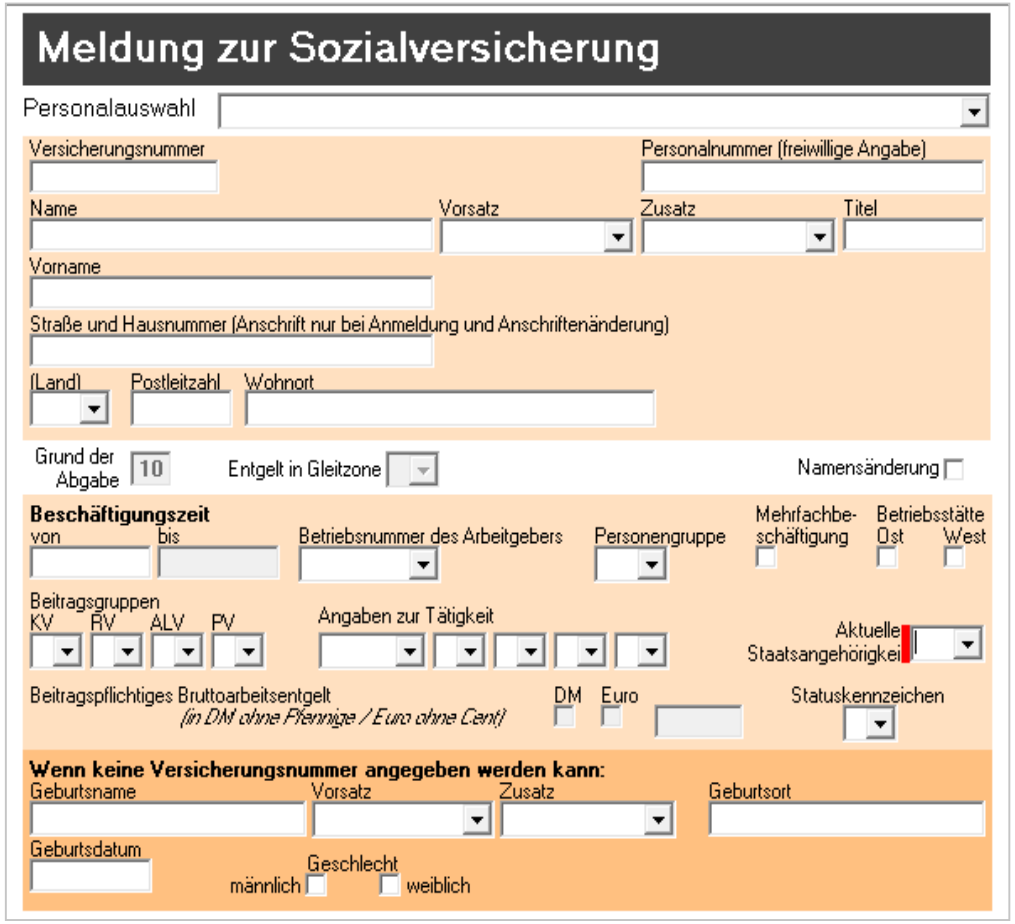
Sample of households with at least one welfare benefit recipient (at reference date)

Refreshed annually
Surveyed annually

Random household sample of resident population

Refreshed annually

Surveyed annually



Trappmann M., Christoph B., Achatz J., Wenzig C. (2009) PASS: a new panel study for labour market research, Int. J. of Manpower , 30, 7, pp.765-770

General Data Processing Regulation (GDPR)

Opt-In

Consent in GDPR

Consent needs to be freely given.

Consent needs to be specific, per purpose.

Consent needs to be informed.

Consent needs to be an unambiguous indication.

Consent is an act: it needs to be given by a statement or by a clear act.

Consent needs to be distinguishable from other matters.

Consent request needs to be in clear, plain language; intelligible and easily accessible.

Informed Consent – Privacy - Technology

Instruction booklet with IAB-Smart app screen shots – separate and active opt-in required for all use cases

Consent to Linkage by Framing and Mode in %

Phone	Front	Back	Total n
Gain	90.8	78.7	598
Loss	90.5	81.2	610
Total n	613	595	1208

Web	Front	Back	Total
Gain	82.6	62.4	520
Loss	86.3	75.4	489
Total	511	498	1009

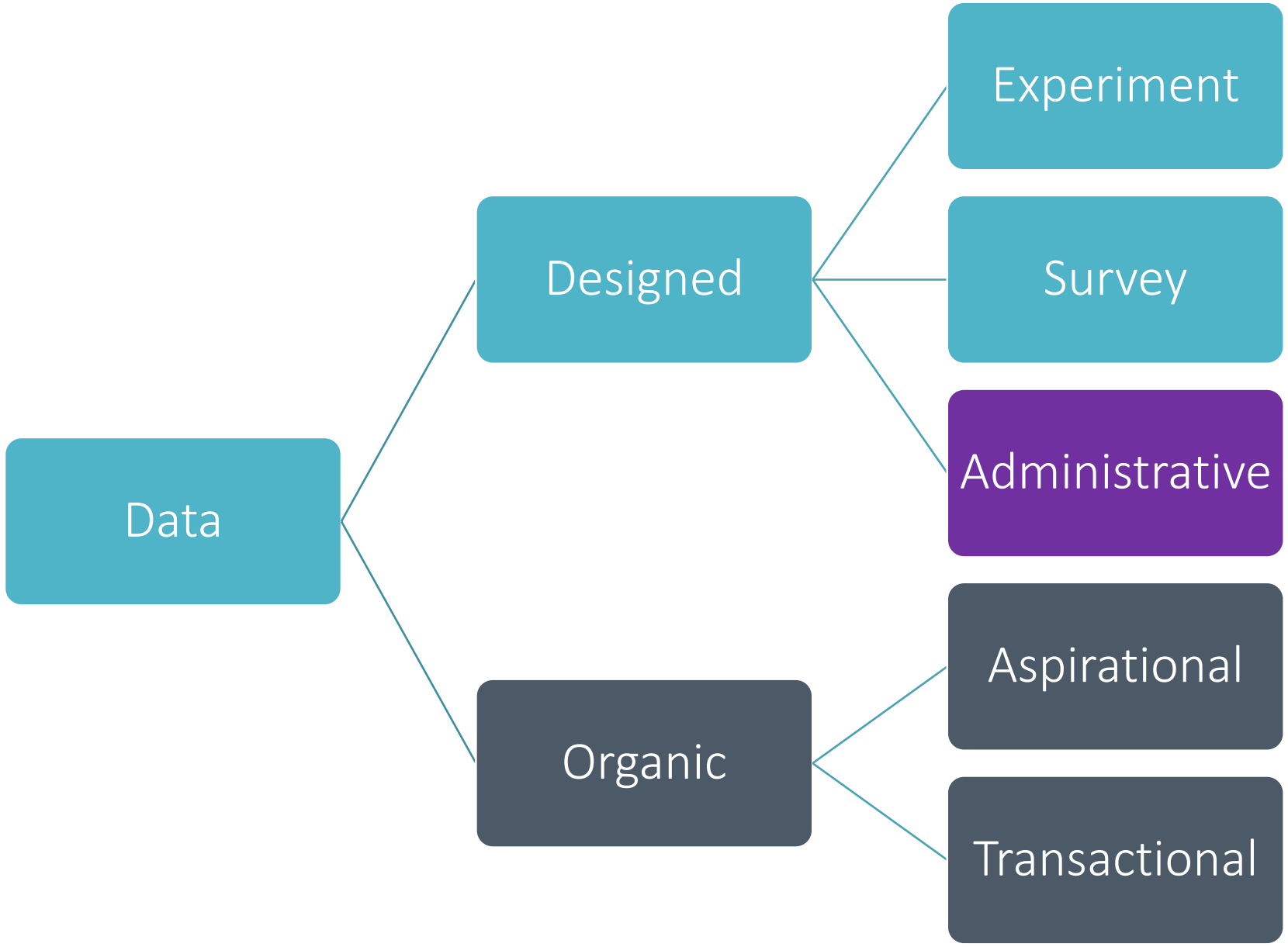
Lack of understanding

Phone	Consenters %correct	Non-consenters %correct
Answers sent to IAB	88.3	57.8
Merged with IAB	93.3	36.7
Name/Adress saved	68.3	38.8
Result lead to you	63.4	--
IAB only access	85.6	--
Public access to identifiable data	87.5	--

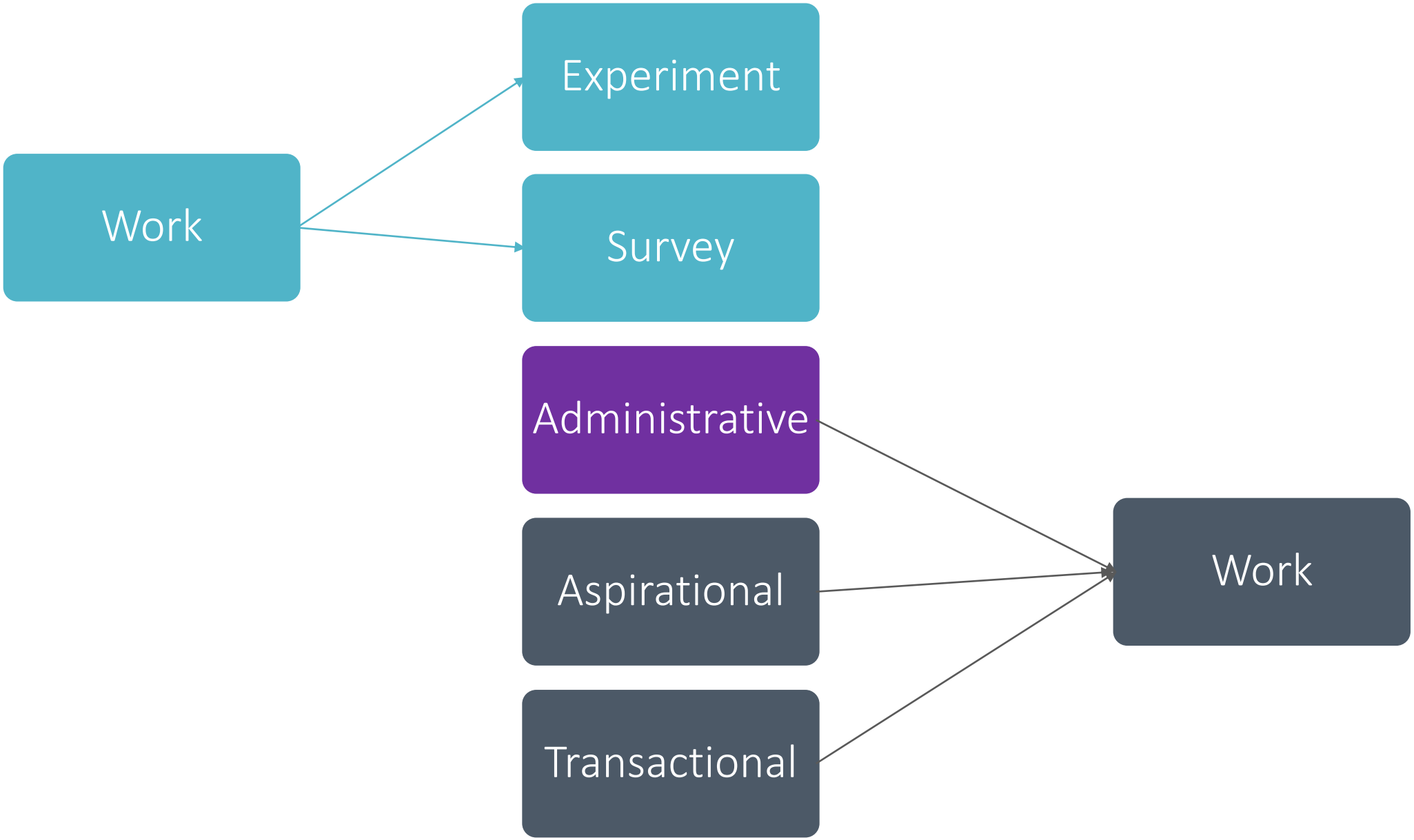
2. Role of methodologists

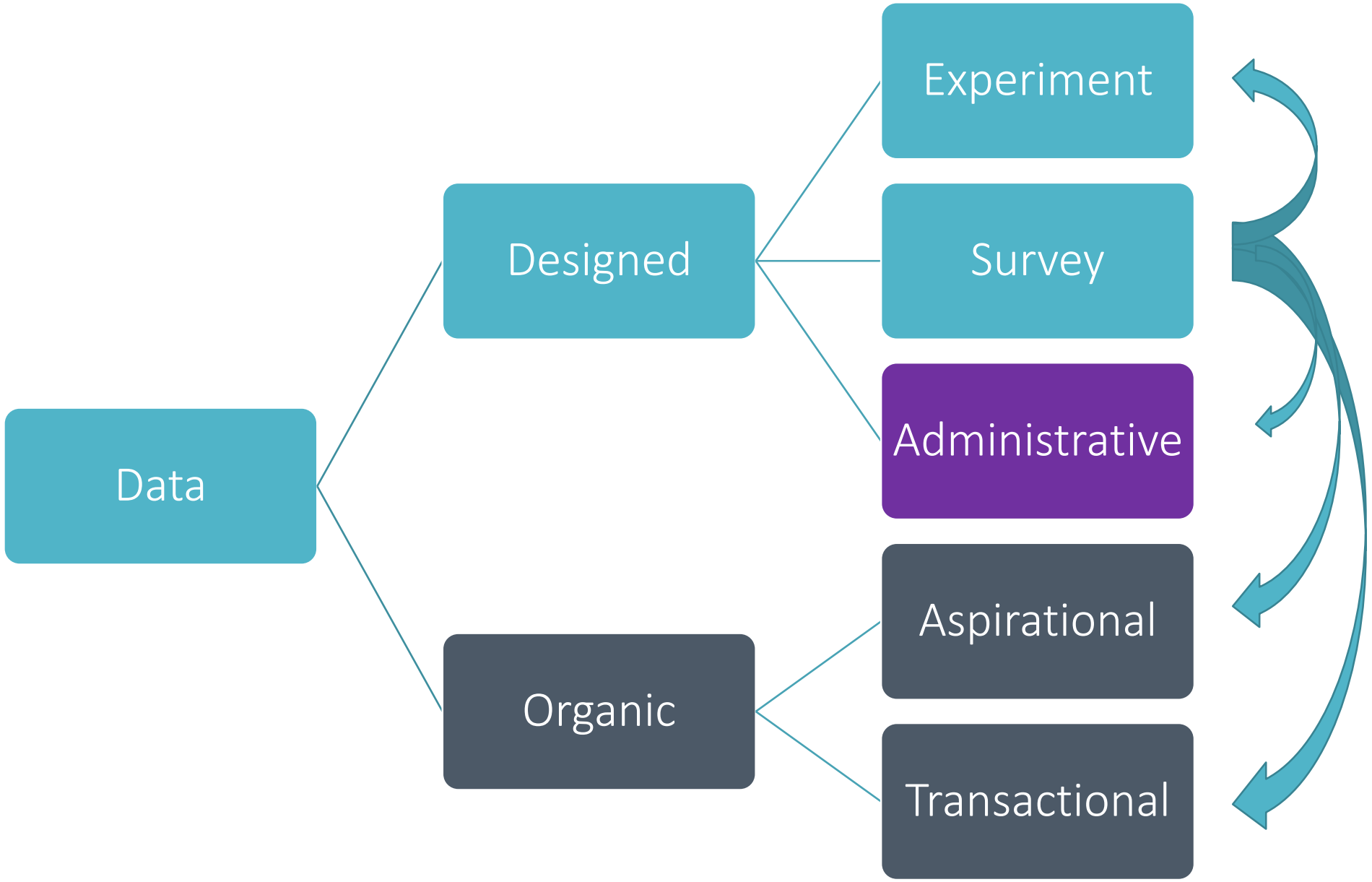
Data Output/Access	Communicate results and distribute data
Data Analysis	Variety of analysis methods suited for different data types
Data Curation/Storage	Curate and manage data
Data Generating Process	Collect data and understand how data are generated through administrative and other processes.
Research Question	Formulate research goal and know which data are best suited to achieve this goal.

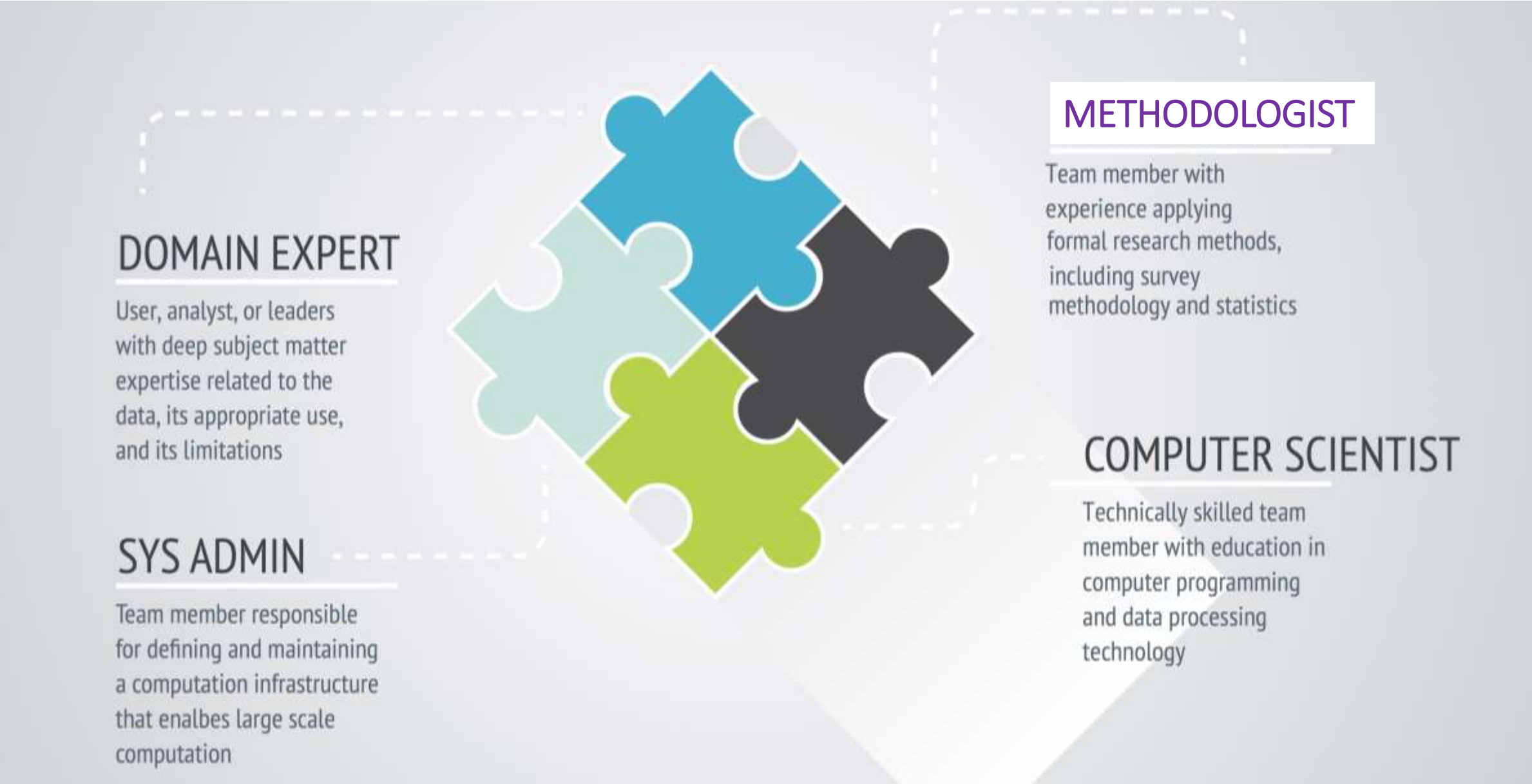
Big Data in survey research: AAPOR Task Force report. Japec, L.; Kreuter, F. ; Berg, M. ; Biemer, P. ; Decker, P.; Lampe, C.; Lane, J. ; O'Neil, C. ; Usher, A.. DOI: 10.1093/poq/nfv039. URL: http://poq.oxfordjournals.org/content/79/4/839.



Source: Roberto Rigobon, [Discussion on Applications and Issues with Using Commercial Data in Research](), BEA Expert Meeting on Exploiting Commercial Data for Official Economic Statistics November 19, 2015


Work
Experiment
Survey
Administrative
Aspirational
Transactional
Work


Data
Designed
Organic
Experiment
Survey
Administrative
Aspirational
Transactional



Big Data in survey research: AAPOR Task Force report. Japec, L.; Kreuter, F. ; Berg, M. ; Biemer, P. ; Decker, P.; Lampe, C.; Lane, J. ; O'Neil, C. ; Usher, A.. DOI: 10.1093/poq/nfv039. URL: http://poq.oxfordjournals.org/content/79/4/839.

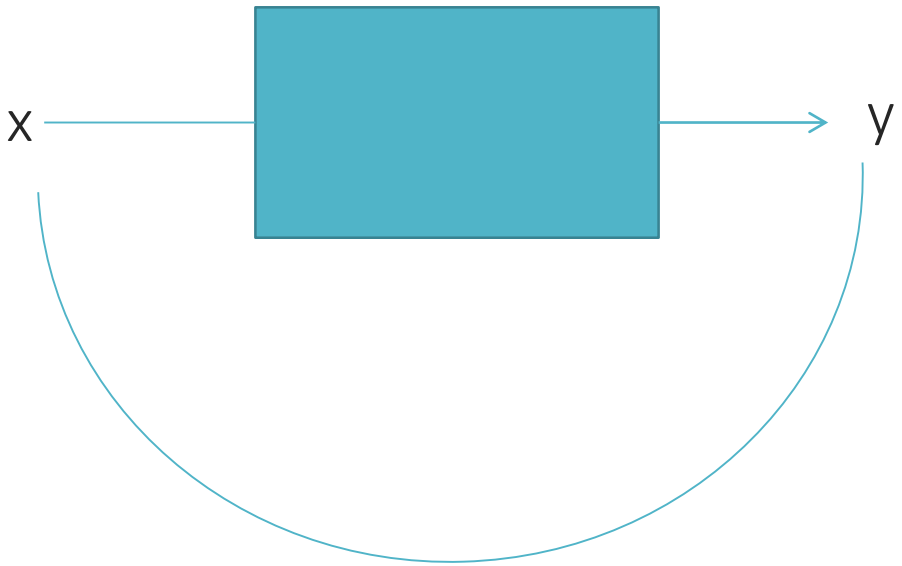
3. Some new vocabulary (skills)

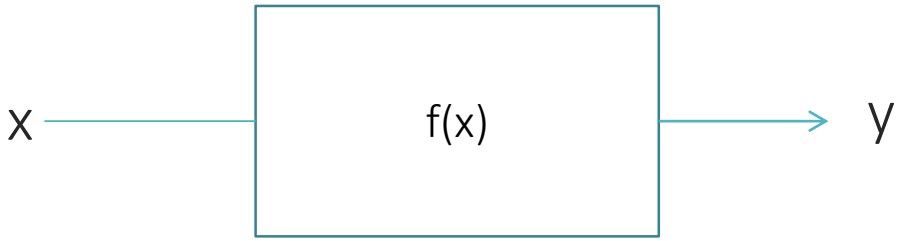
Data Output/Access	Example: map visualization / privacy / GDPR
Data Analysis	Example: Hadoop MapReduce; High Frequency Data; Machine Learning
Data Curation/Storage	Example: Record Linkage; Database Hadoop Distributed File System
Data Generating Process	Examples: geolocated social media + survey + administrative data
Research Questions	Examples: Behavior of interest (political participation/job searches)

Big Data in survey research: AAPOR Task Force report. Japec, L.; Kreuter, F. ; Berg, M. ; Biemer, P. ; Decker, P.; Lampe, C.; Lane, J. ; O'Neil, C. ; Usher, A.. DOI: 10.1093/poq/nfv039. URL: http://poq.oxfordjournals.org/content/79/4/839.

Usher 2015

Machine Learning





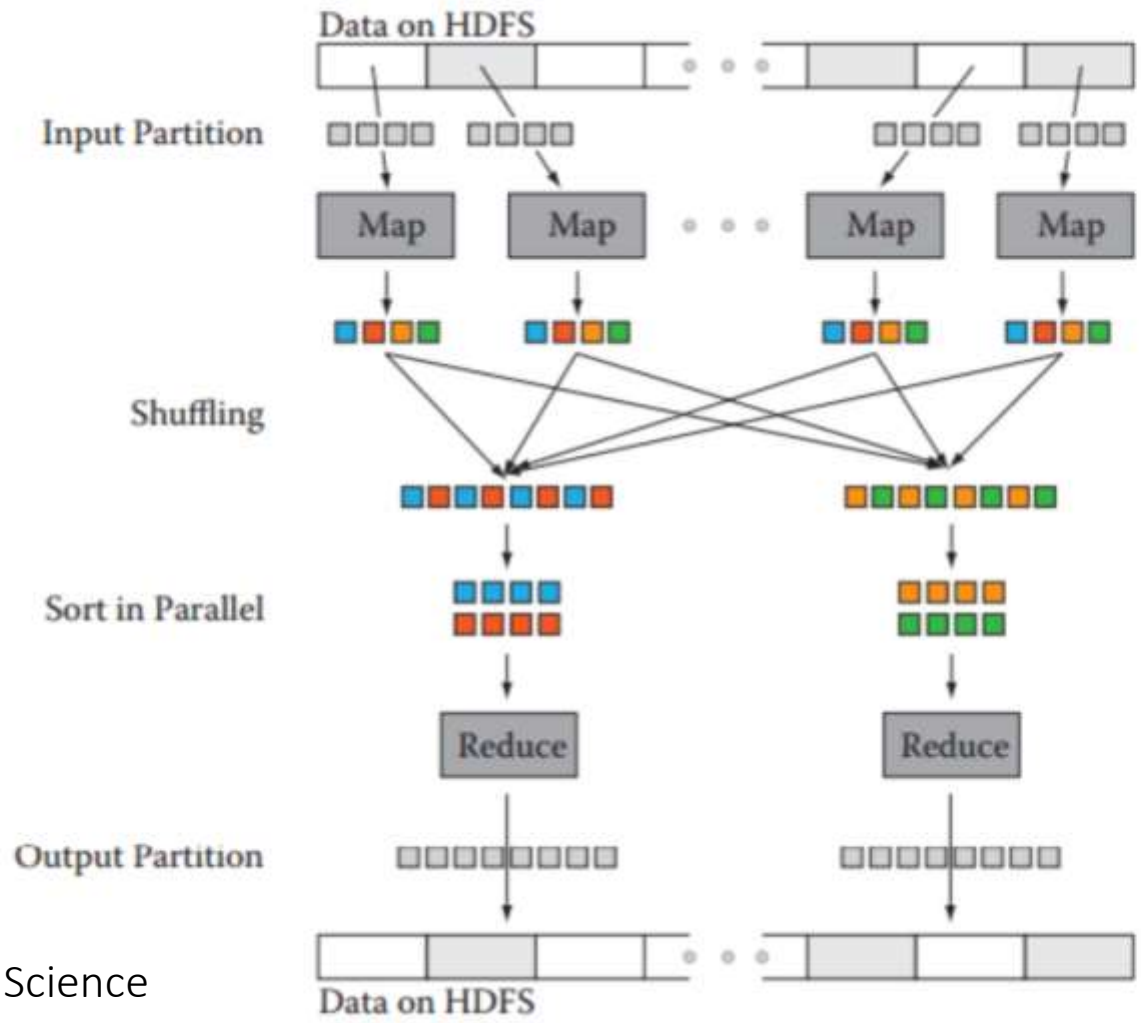
https://github.com/DataScienceSpecialization/courses from Roger Peng, Jeff Leek, Brian Caffo

Database Management

Text files and scripting language	• Your data is small • Your analysis is simple • You do not expect to repeat analyses over time
Statistical packages	• Your data is modest in size • Your analysis maps well to your chosen statistical package
Relational database	• Your data is structured • You will be analyzing data repeatedly over time
NoSQL database	• Your data is unstructured • Your data is extremely large

Foster et al. Big Data and Social Science: Practical Guide . 2016

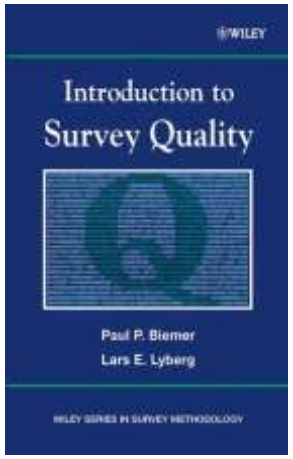
BD Programming – MapReduce

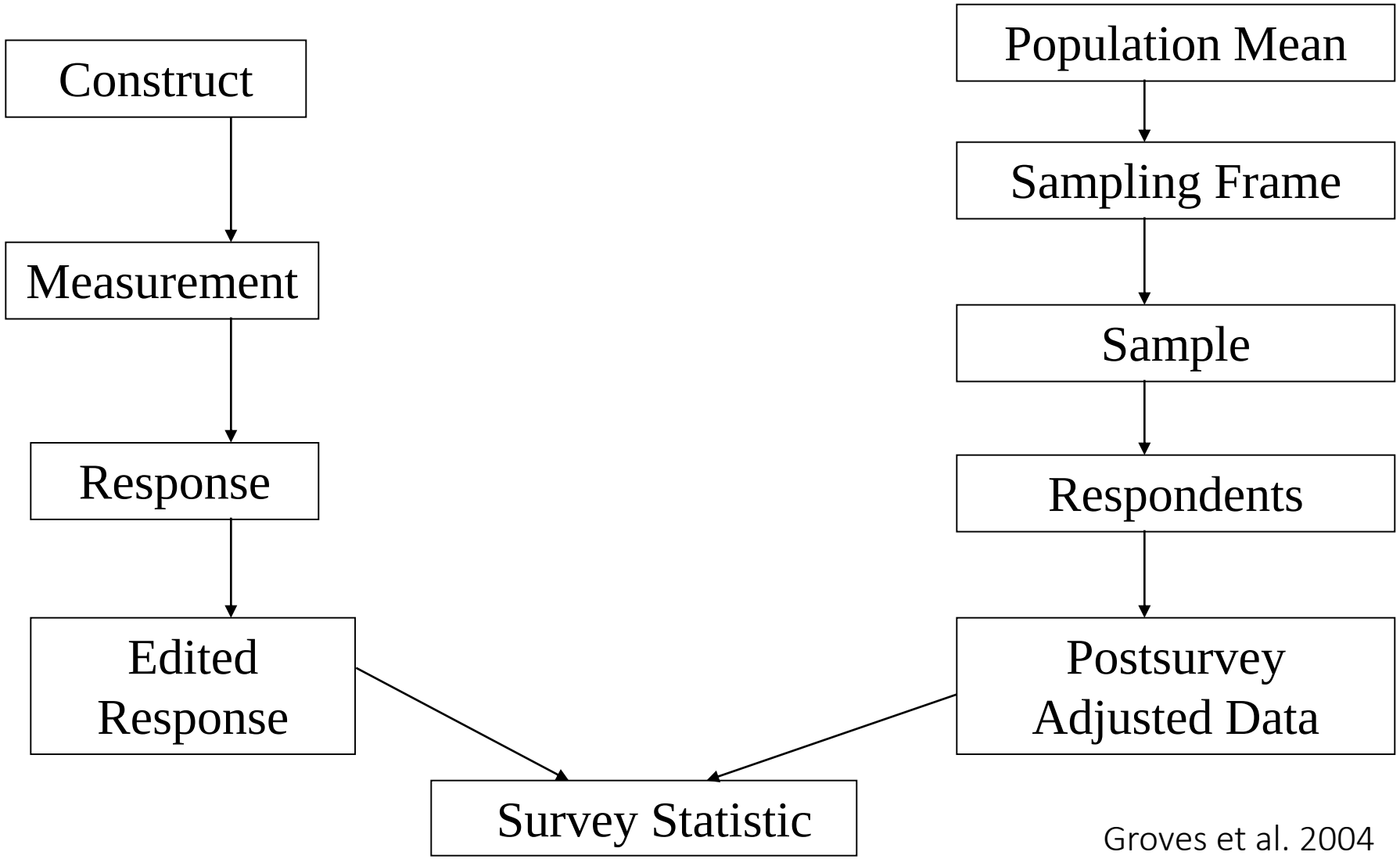


Foster et al. Big Data and Social Science

4. What we bring to the table

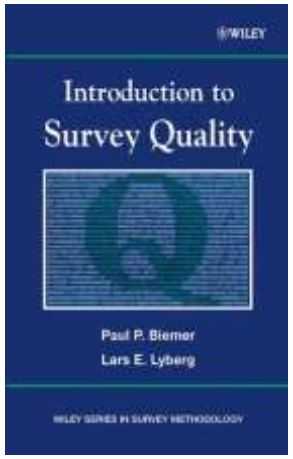
Data Generating Process

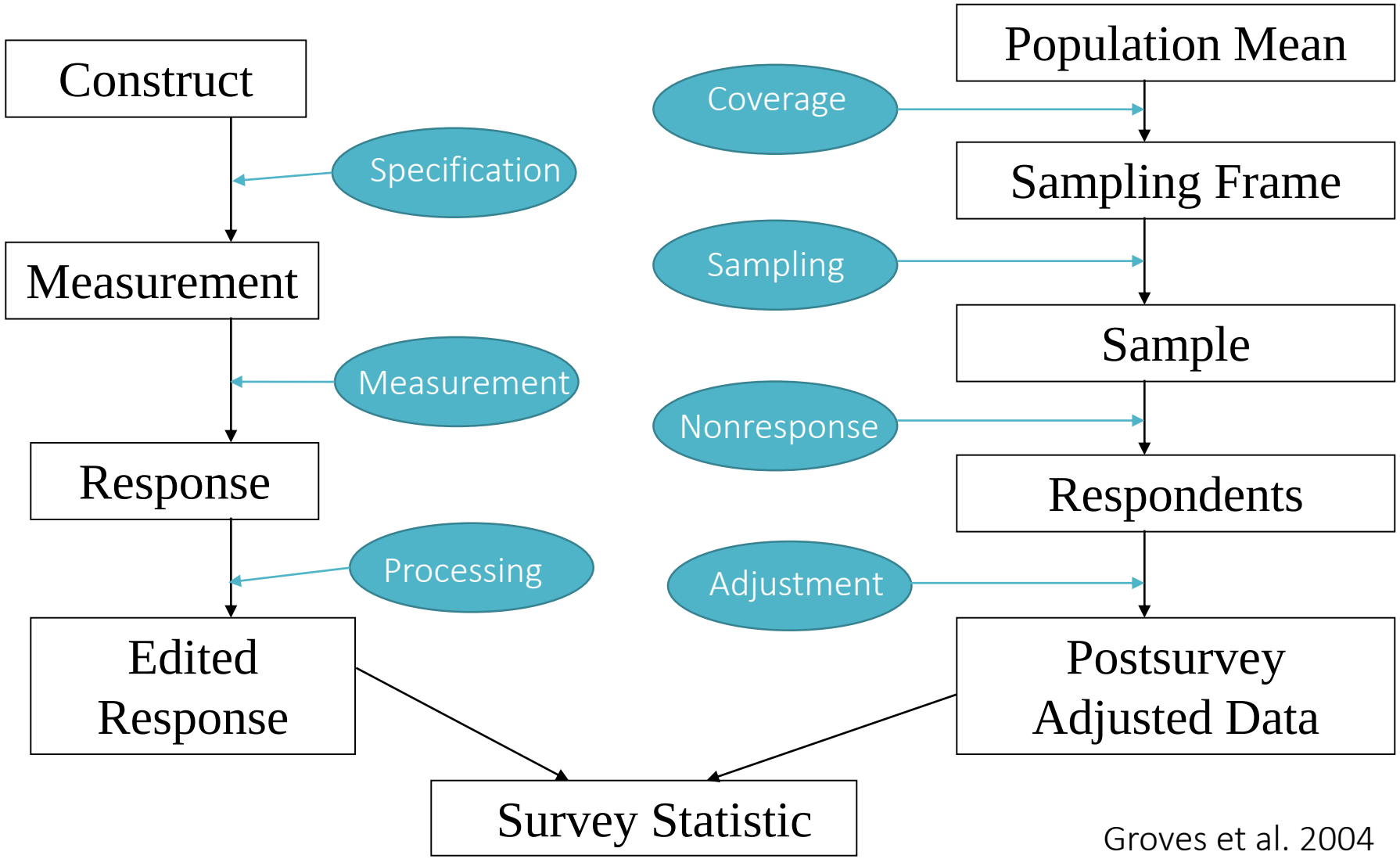




Groves et al. 2004

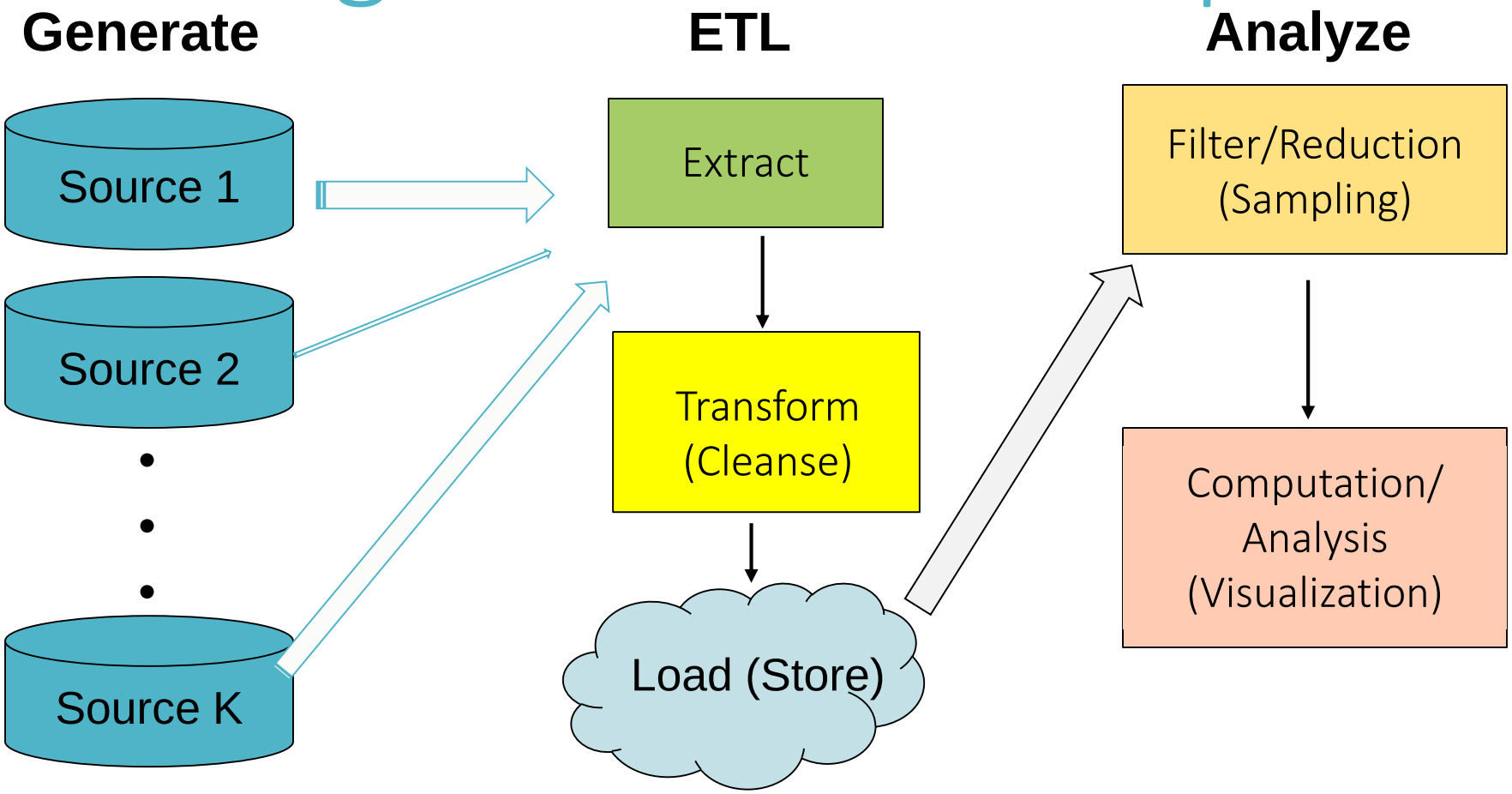
Data Generating Process





Groves et al. 2004

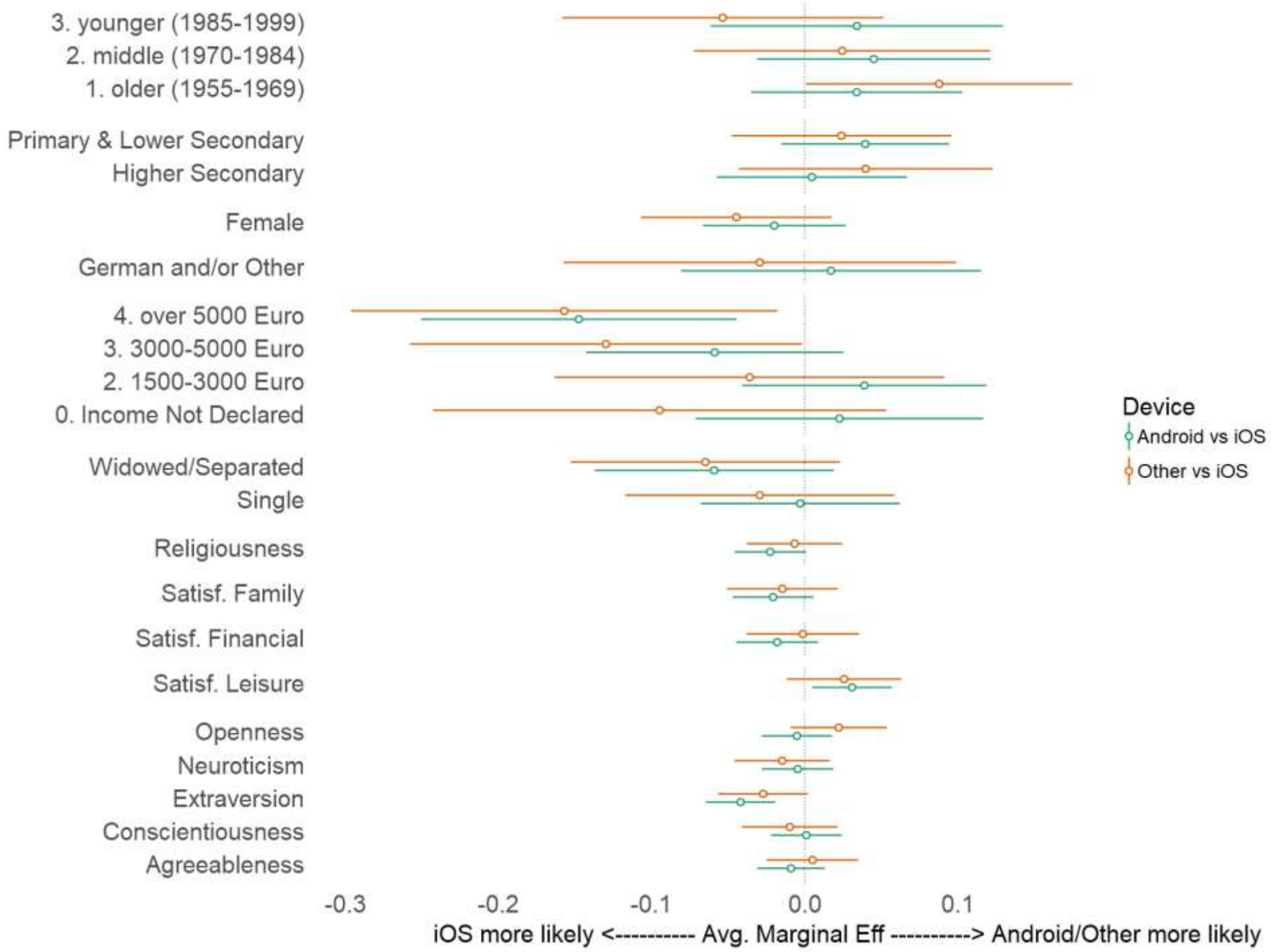
Big Data Process Map



Source: Paul Biemer in Japec, Kreuter et al. 2015 – AAPOR Task Force Report

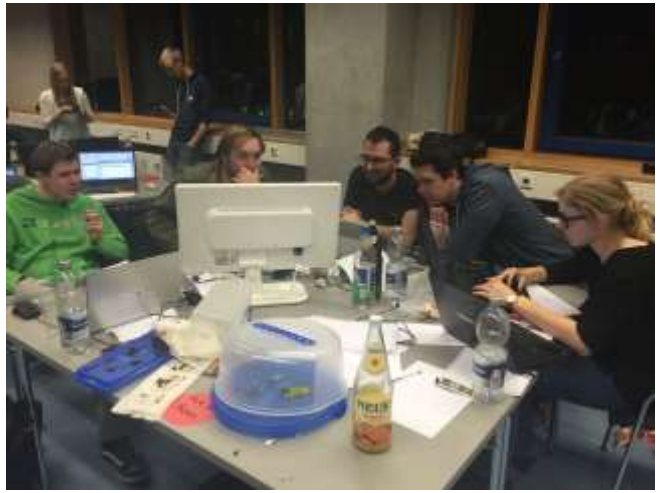
Boston Street Bumps

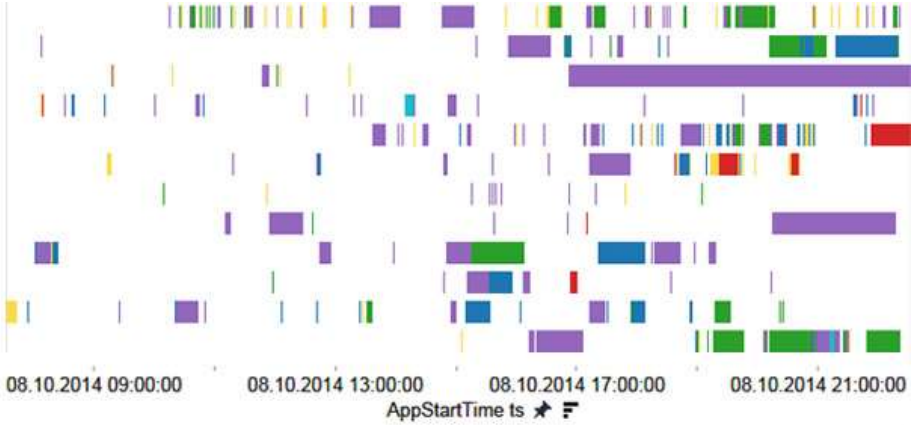
5. Back to the IAB - Example


3. younger (1985-1999)
2. middle (1970-1984)
1. older (1955-1969)
Primary & Lower Secondary
Higher Secondary
Female
German and/or Other
4. over 5000 Euro
3. 3000-5000 Euro
2. 1500-3000 Euro
0. Income Not Declared
Widowed/Separated
Single
Religiousness
Satisf. Family
Satisf. Financial
Satisf. Leisure
Openness
Neuroticism
Extraversion
Conscientiousness
Agreeableness
-0.3
-0.2
-0.1
0.0
0.1
iOS more likely <---------- Avg. Marginal Eff ----------> Android/Other more likely
Device
Android vs iOS
Other vs iOS

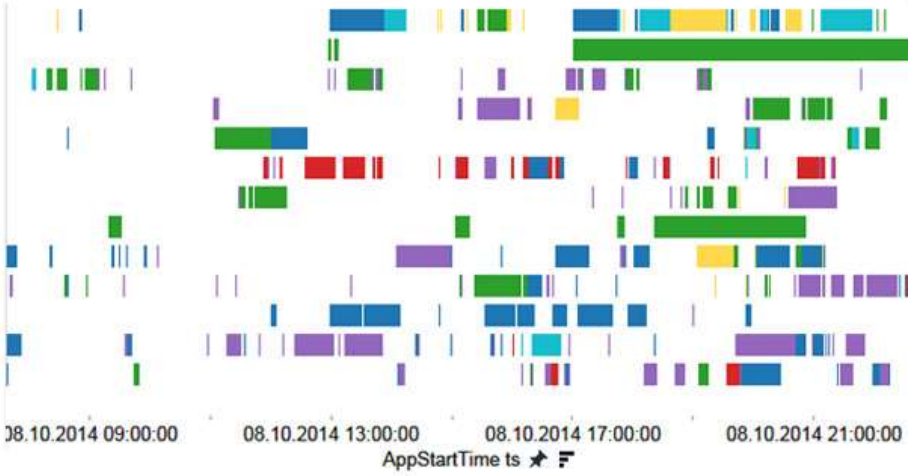
… typical attempts to find analysis help

DataFest

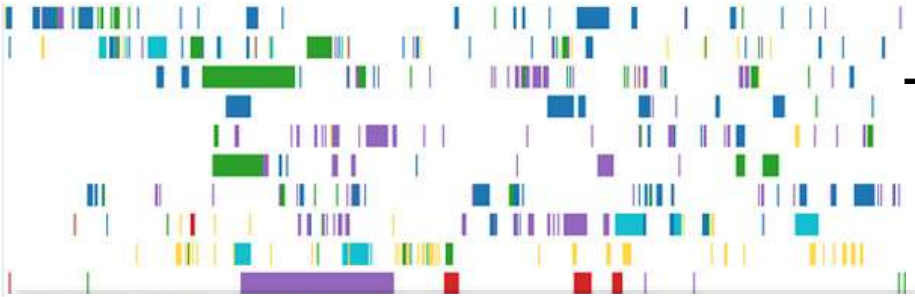




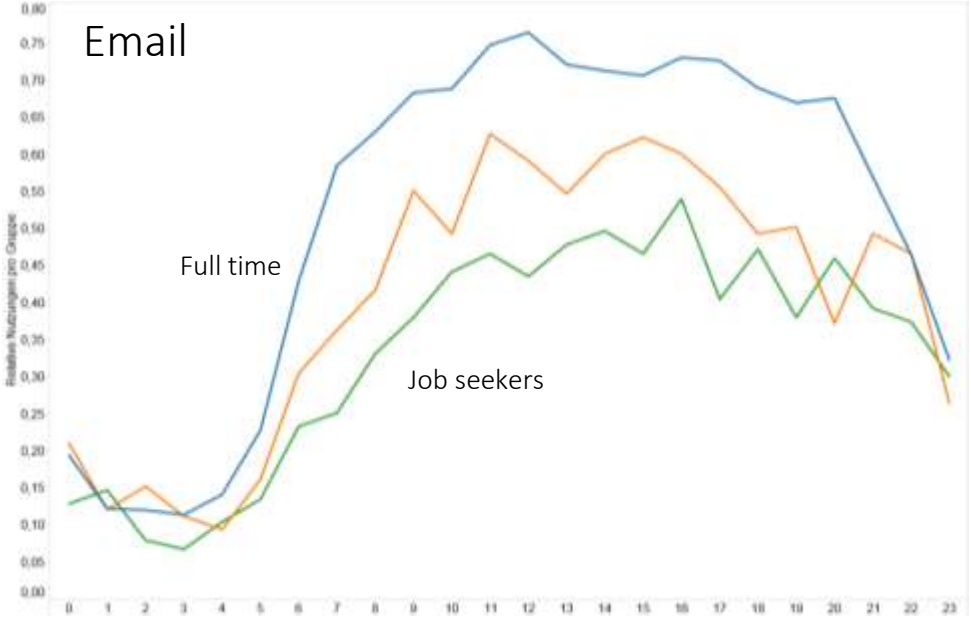
Full-time employed
→ App use past 5pm

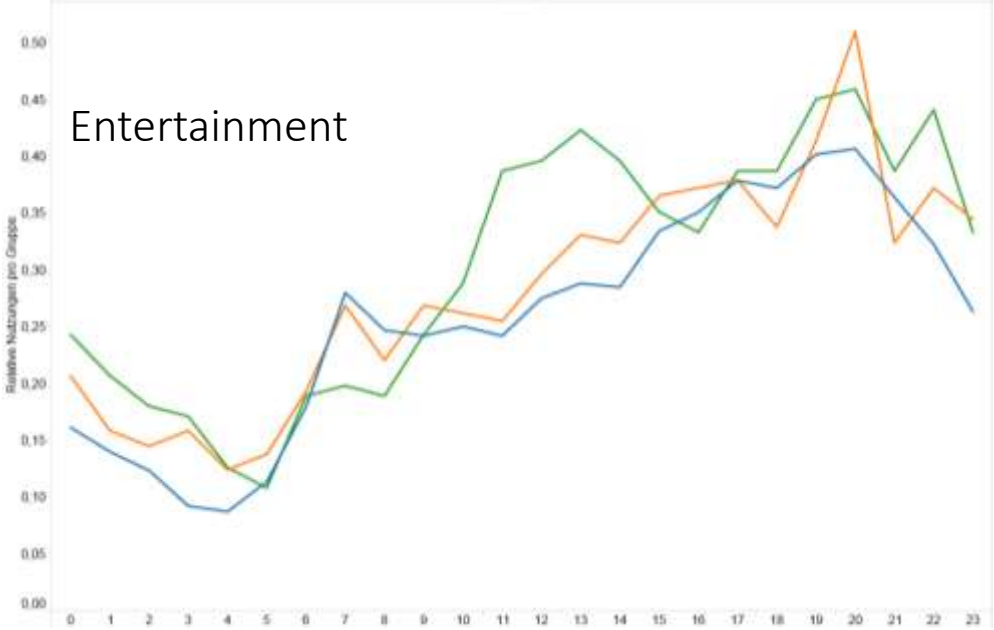


Part-time employed
→ App use at noon



Job seekers
→ Continuous app use





Schlickman et.al. (DataDiggers) App Nutzerverläufe statt Surveys? Eine Machbarkeitsstudie. DataFest Germany, Mannheim 2015, http://sswml.uni-mannheim.de/Teaching/DataFest%20Germany/DataFest%20Germany%202015/

Summary

1. Methodologists need to help teams find or create the right data for a specific purpose
2. It is easy to overlook what is missing. Privacy and confidentiality even more important
3. Methodologists should make use of all available data, and they can (with a few new skills)
4. Frameworks can help assess quality of single and combined data sources
5. Working in teams and with unlikely partners can accelerate

THANK YOU!

fkreuter@umd.edu

Twitter: @fraukolos