

Workshop

The world of blended data

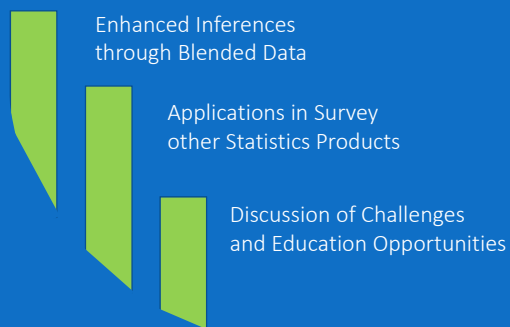


Frauke Kreuter @frauolos
JPSM - Uni Maryland
SOWI - Uni Mannheim – IAB

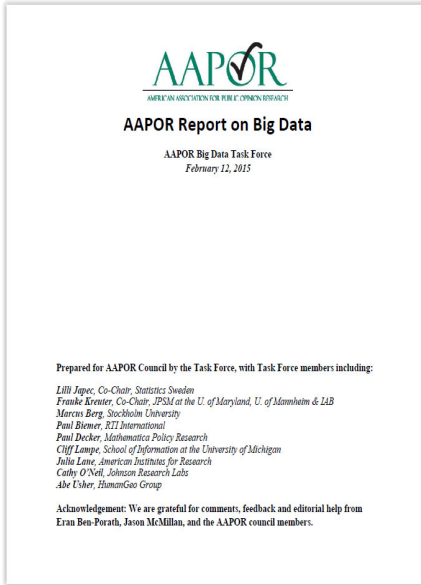
1

Overview

The world of blended data



2

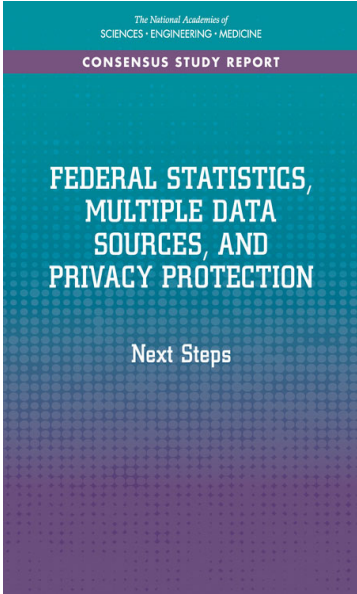


AAPOR Report on Big Data
AAPOR Big Data Task Force
February 12, 2015

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

Lilli Jager, Co-Chair, Statistics Sweden
Frankie Kreuter, Co-Chair, JPSM at the U. of Maryland, U. of Mannheim & IAB
Marcus Berg, Stockholm University
Pauli Biemer, ZITI International
Pauli Decker, Mathematics Policy Research
Cliff Lampe, School of Information at the University of Michigan
Julia Lane, American Institutes for Research
Cathy O'Neill, Johnson Research Labs
Abe Usher, HumanGen Group

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

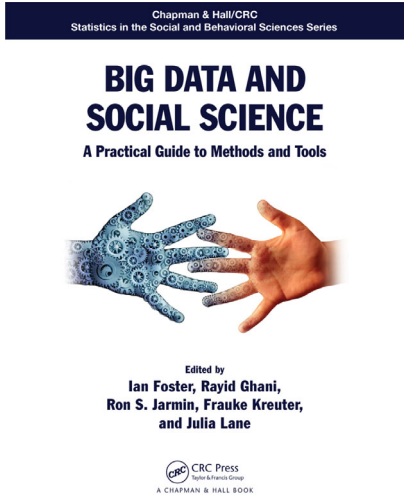


The National Academies of
SCIENCES • ENGINEERING • MEDICINE

CONSENSUS STUDY REPORT

**FEDERAL STATISTICS,
MULTIPLE DATA
SOURCES, AND
PRIVACY PROTECTION**

Next Steps



Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series


**BIG DATA AND
SOCIAL SCIENCE**
A Practical Guide to Methods and Tools

Edited by
**Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane**




CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

<https://coleridge-initiative.github.io/big-data-and-social-science/>

3




Training Computing Connecting Mailing List


A collaboration presented by   

<http://coleridgeinitiative.org>

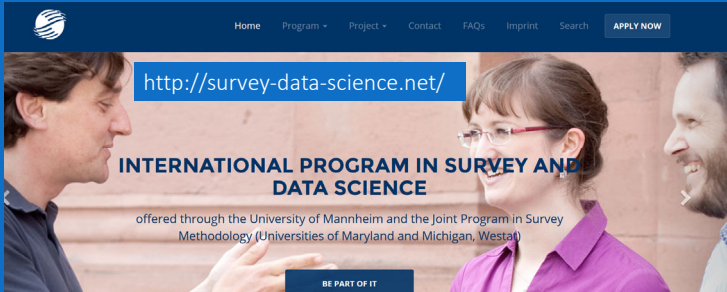
**COLERIDGE
INITIATIVE**



SPONSORED BY THE

 **Federal Ministry
of Education
and Research**

**AUFSTIEG DURCH
BILDUNG >>**
OFFENE HOCHSCHULEN



Home Program Project Contact FAQs Imprint Search **APPLY NOW**

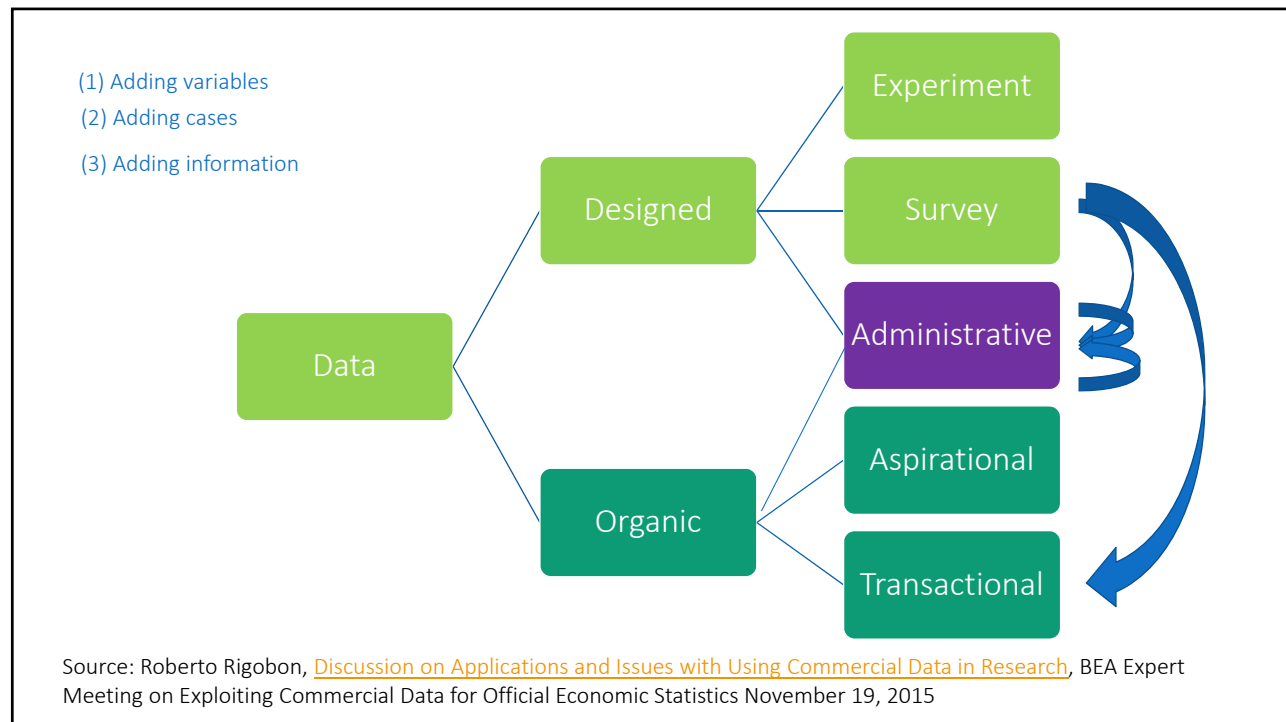
<http://survey-data-science.net/>

**INTERNATIONAL PROGRAM IN SURVEY
AND DATA SCIENCE**

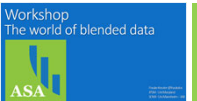
offered through the University of Mannheim and the Joint Program in Survey
Methodology (Universities of Maryland and Michigan, West)

BE PART OF IT

4



5



Additional Resources

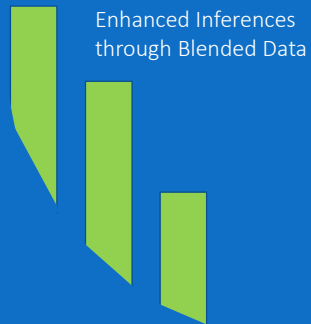
1. Lohr, S. L. and Raghunathan, T. E. (2017). Combining survey data with other data sources. *Statistical Science*, **32**, 293-312
2. Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics*, 12, 685-726
3. Rao, J.N.K. (under review). On Making Valid Inferences by Combining Data from Surveys and Other Sources. Carleton University, Ottawa, Canada
4. Thompson, M. E. (2019). Combining data from new and traditional sources in population surveys. *International Statistical Review*, **87**, S79-S89

1. Christen, P (2012). Data Matching. Concepts and techniques for record linkage, entity resolution, and Duplicate Detection. Springer
2. Chambers, R. L., Fabrizi, E. and Salvati, N. (2019). Small area estimation with linked data. Technical report appeared as arXiv: 1904.00364v1
3. Drechsler, J. (2011). Synthetic datasets for statistical disclosure control. Springer

1. Valliant, R. and Dever, J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, **40**, 105-137
2. Yang, S., Kim, J. K. and Song, R. (2019). Doubly robust inference when combining probability and non-probability samples with high-dimensional data. Technical Report: arXiv: 1903.05212v1

6

Enhanced Inferences



7



Enhanced Inferences

1. Combined data can **enhance our measurements**
2. **Purposeful design** is needed for success
3. Knowledge on **data generating processes** is key

8



Enhanced Inferences

1. Combined data can enhance our measurements
2. **Purposeful design** is needed for success
3. Knowledge on data generating processes is key

9

9

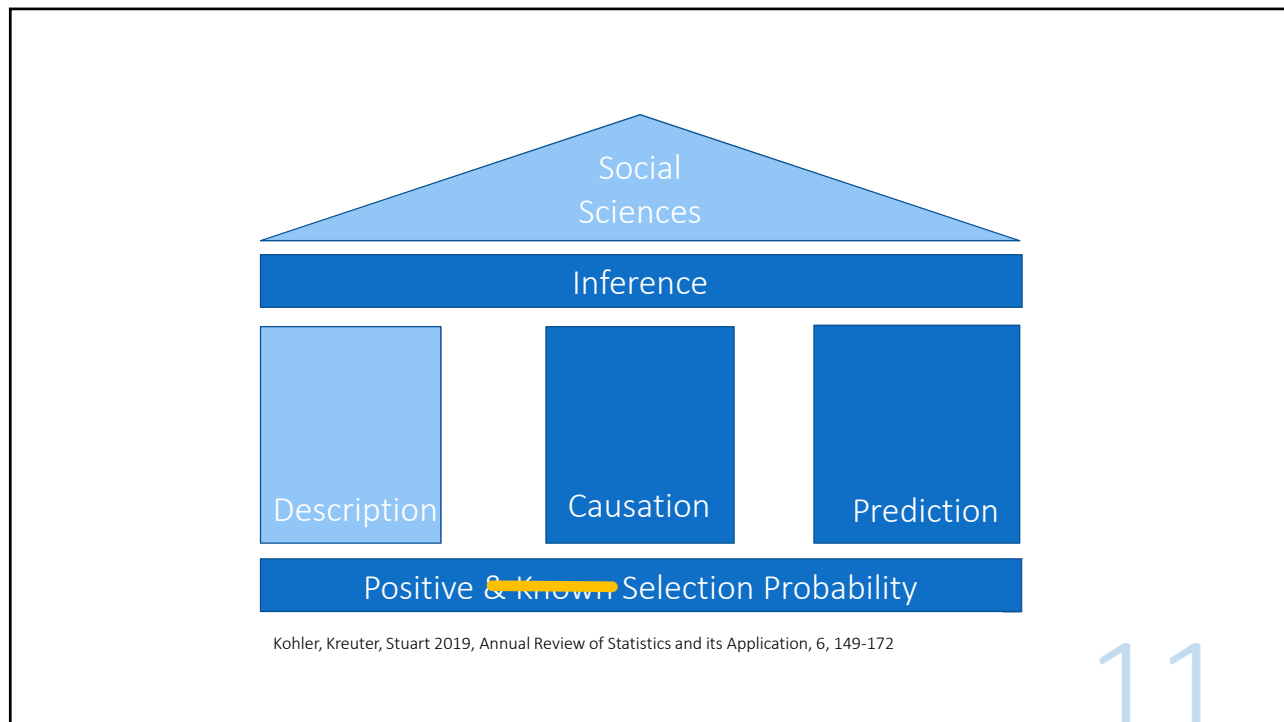
collection

One way to think about a data *analysis* is to think of it as a product to be designed. [...] Producing a useful product requires careful consideration of who will be using it.

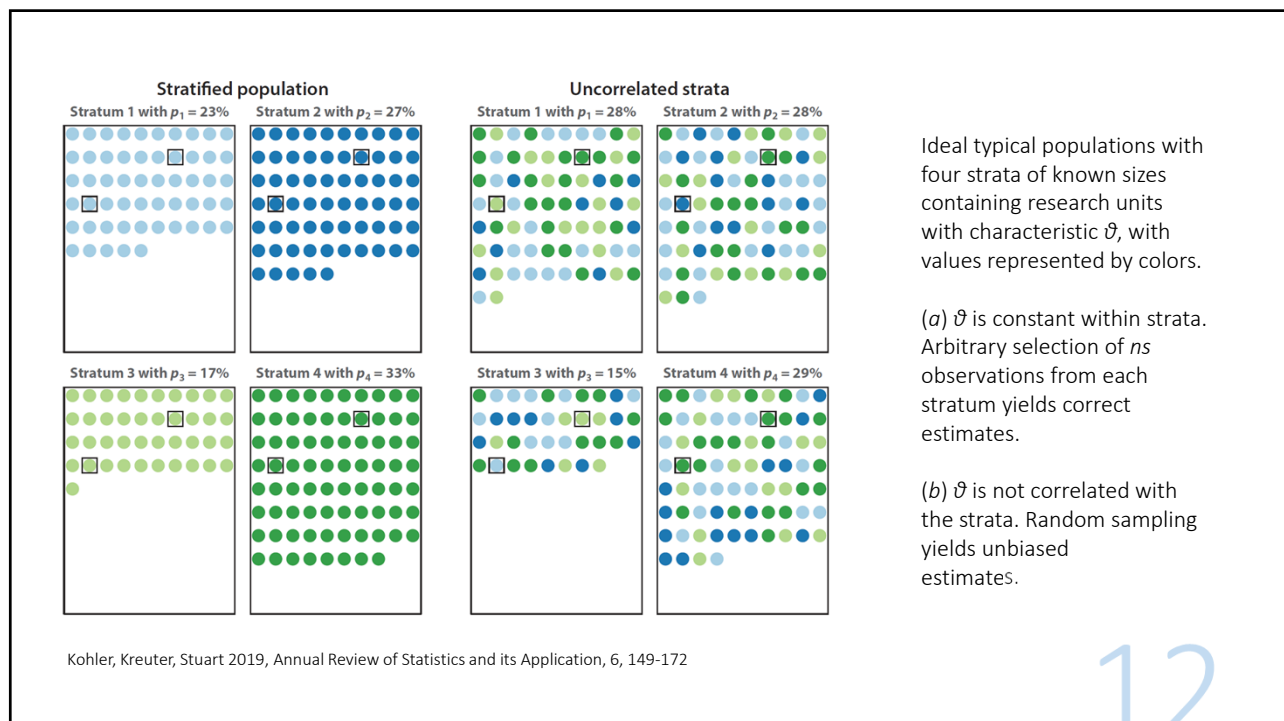
Roger Peng, 2018

10

10



11



12



Enhanced Inferences

1. Combined data can enhance our measurements
2. Purposeful design is needed for success
3. Knowledge on **data generating processes** is key

13

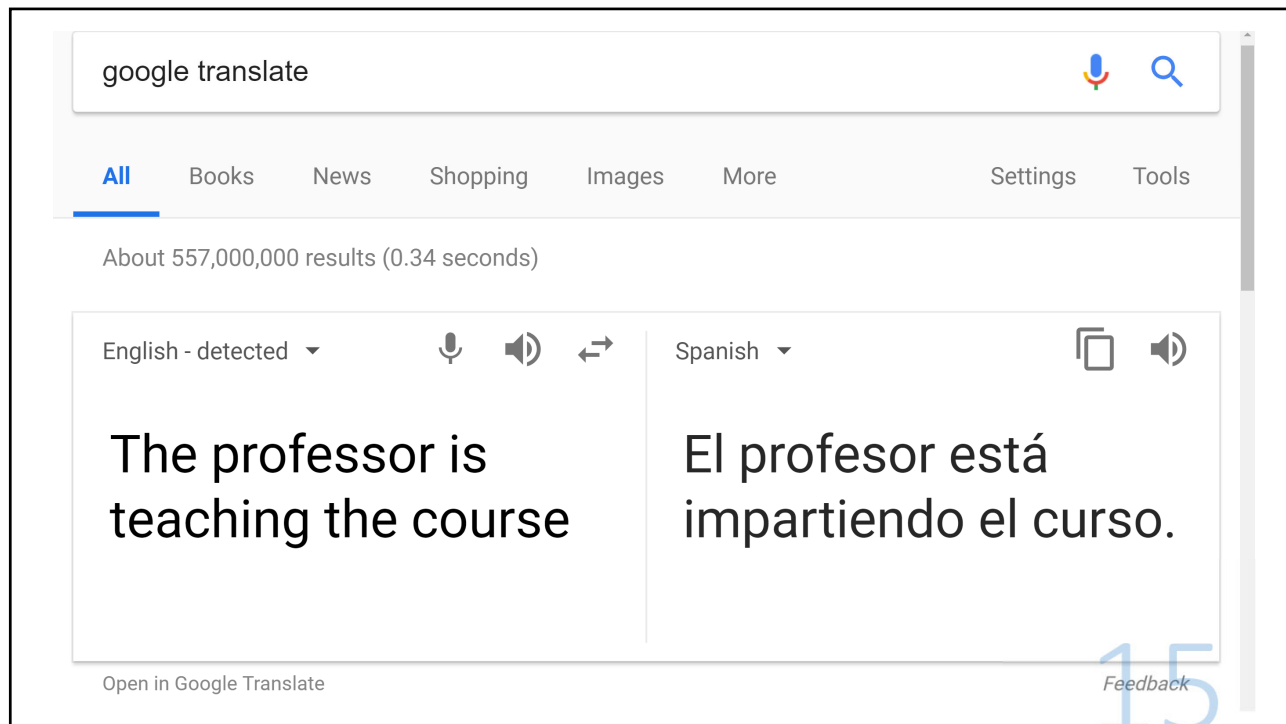
13

By far the worst approach is to wait for data quality problems to surface on their own.

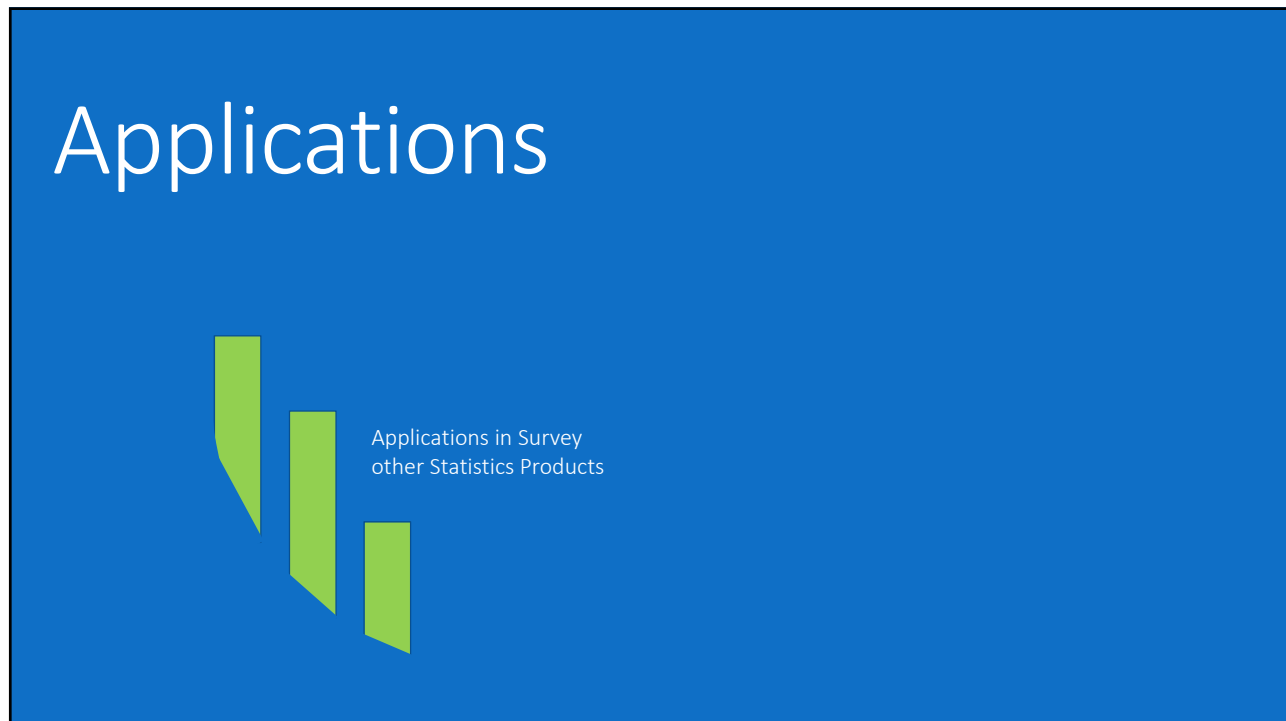
T. Herzog, F. Scheuren, W. Winkler, 2007

14

14



15



16

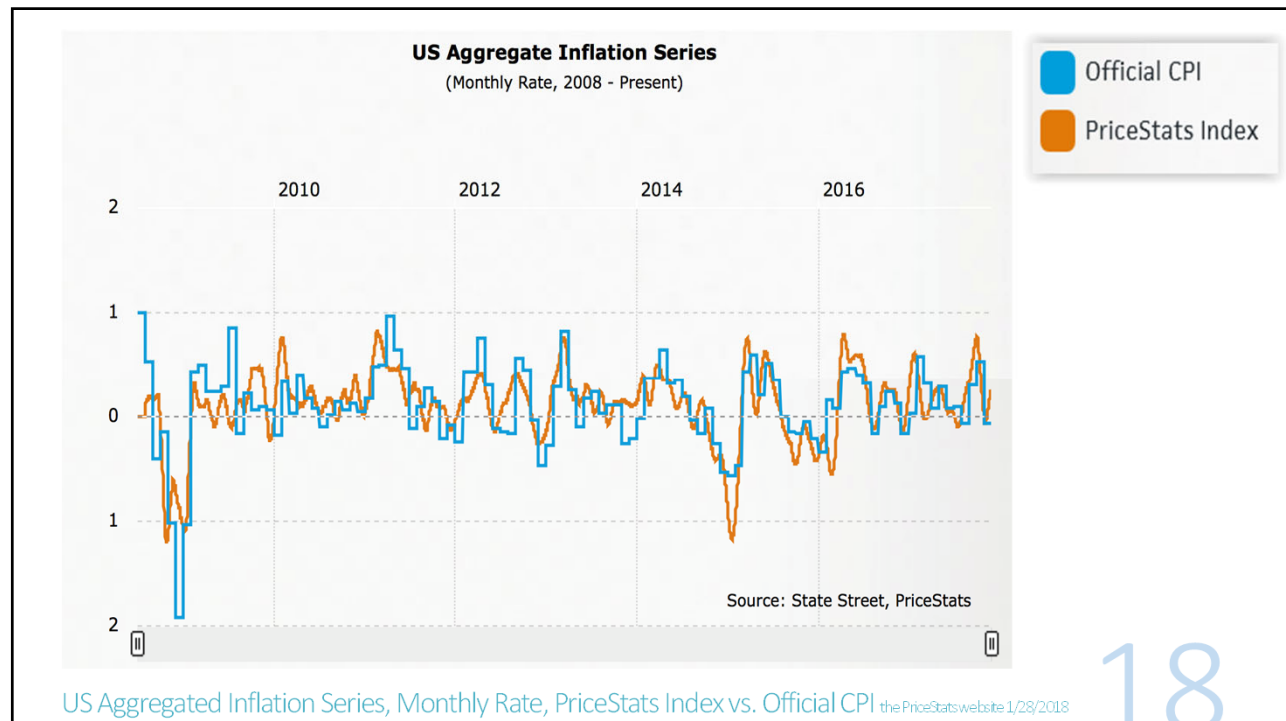


Applications

1. Data from different sources
and
2. Blended collection on same source

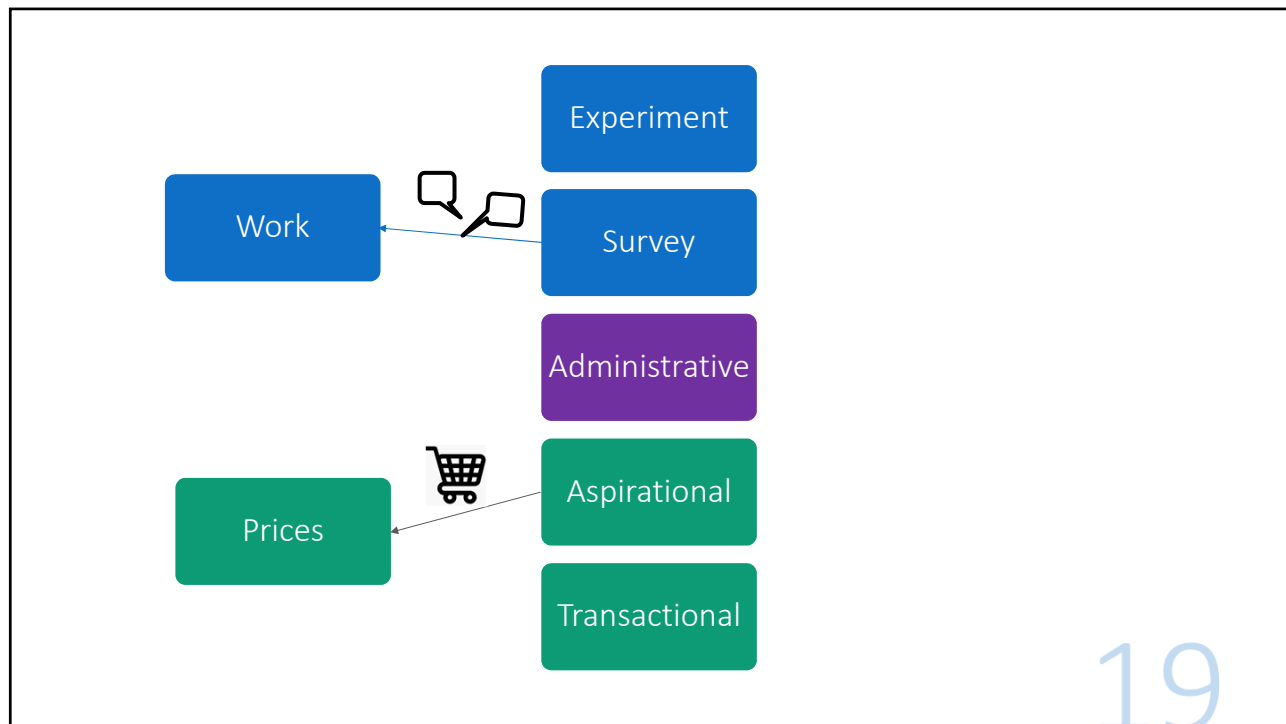
17

17

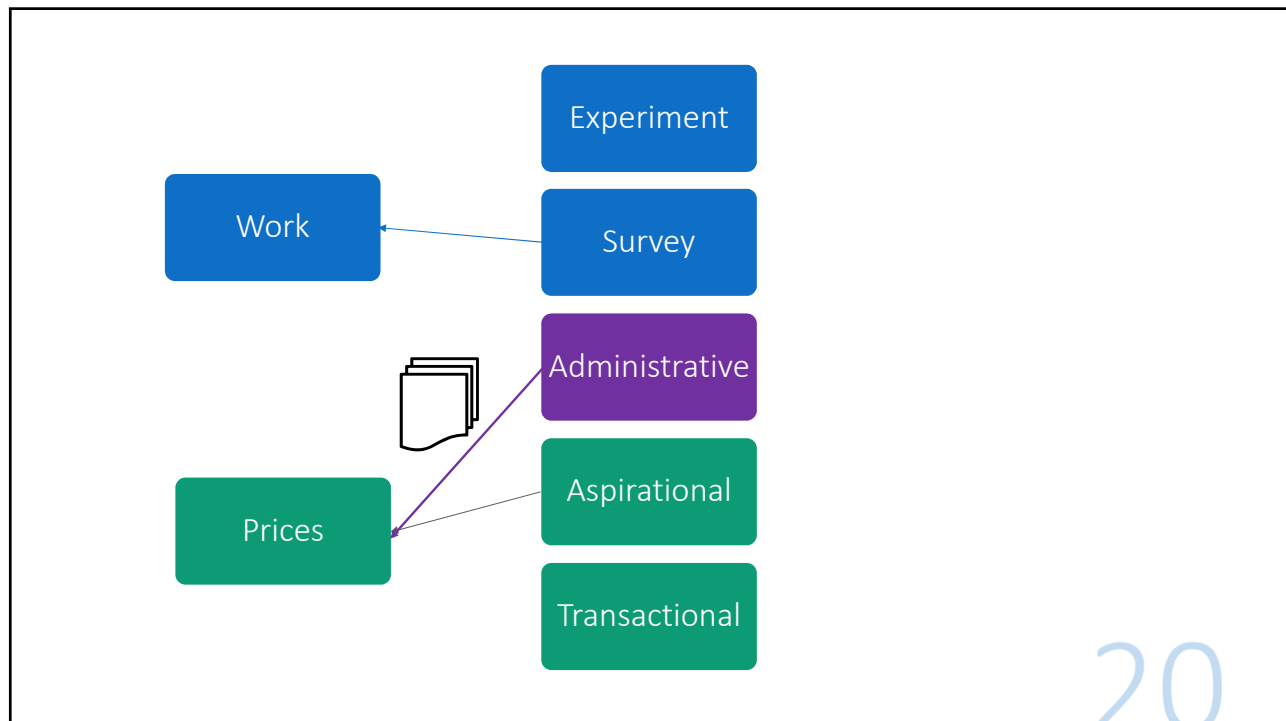


18

18



19



20



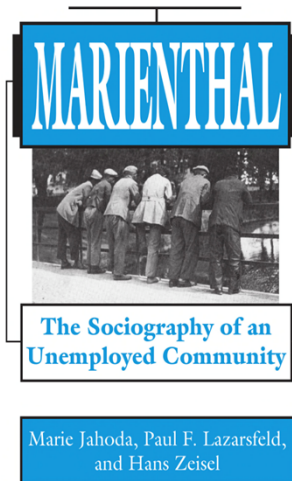
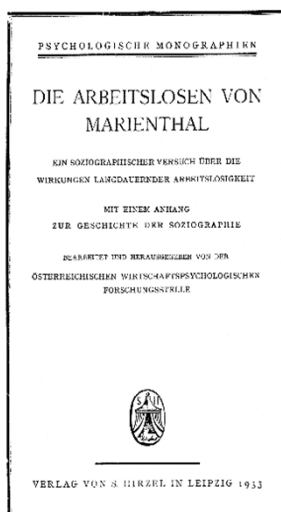
Applications

1. Data from different sources
and
2. Blended collection on same source

21

21

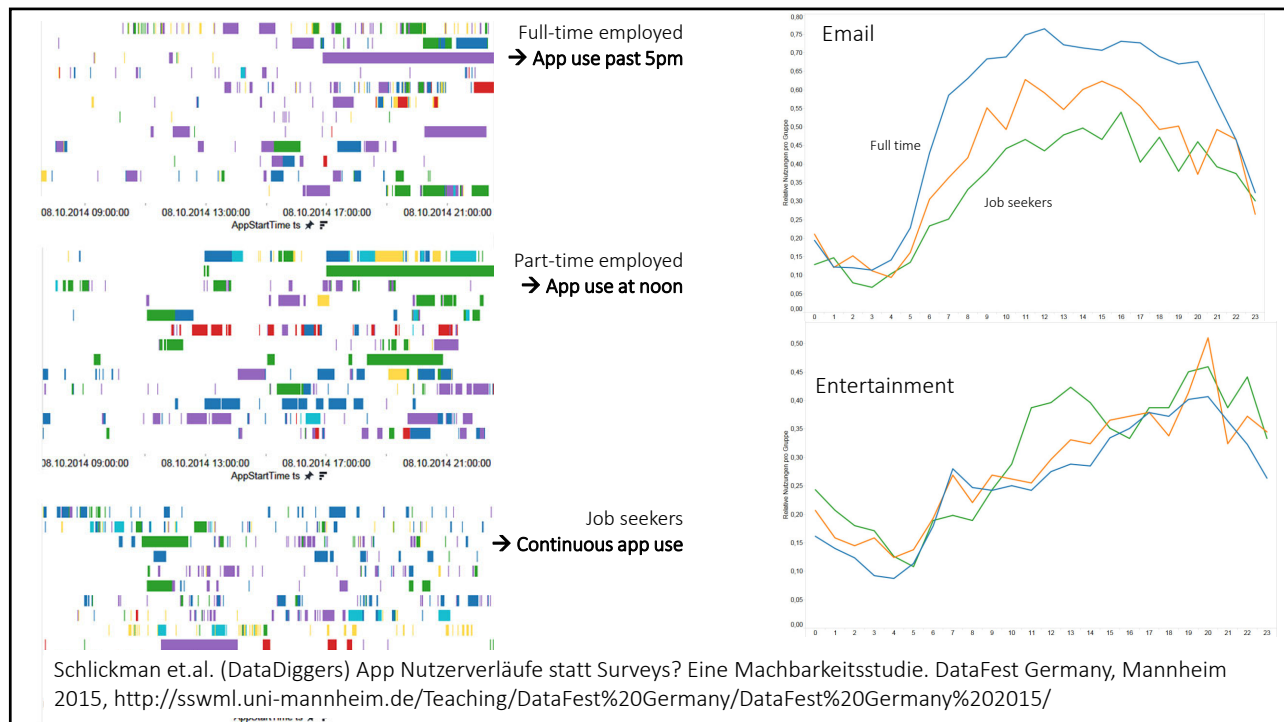
Effects of Unemployment?



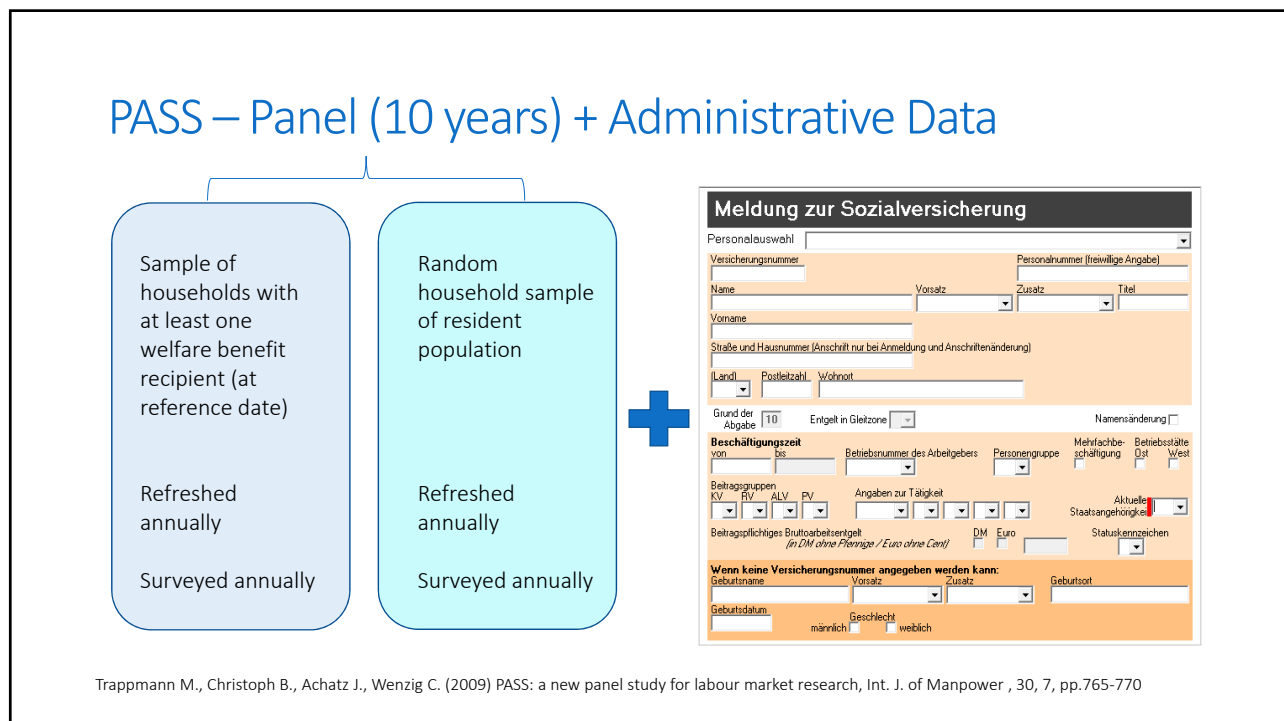
Source: Archives for the History of Sociology in Austria (Graz), »Marienthal« Virtual Archives



22



23



24

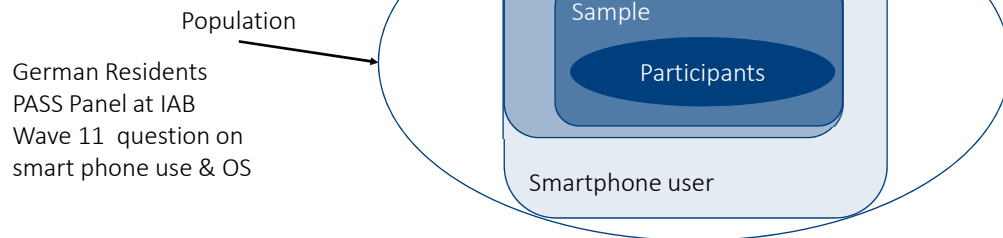
Inference to Population

...owning a (specific) smartphone

...being able to download an app

...being willing to download an app

} Nonparticipation error



// Page 25

25

Challenges and Education Opportunities



Discussion of Challenges
and Education Opportunities

26



Challenges and Opportunities

1. We can quickly face **high privacy risks**
2. Researchers need to value **appropriate flow**
3. **Infrastructure** needed – physical and mental

27

27

Microdata Releases

H1B Salary Database						
Companies ▾ Job Titles ▾ Cities Highest Paid ▾ Refinance Student Loan & Save! Subscribe						
EMPLOYER	JOB TITLE	BASE SALARY	LOCATION	SUBMIT DATE	START DATE	
FEDERAL RESERVE BANK OF SAN FRANCISCO	APPLICATION DEVELOPER IV	160,000	SAN FRANCISCO, CA	05/27/2016	07/05/2016	
FEDERAL RESERVE BANK OF SAN FRANCISCO	APPLICATION DEVELOPER IV	160,000	SAN FRANCISCO, CA	06/08/2016	07/18/2016	
FEDERAL RESERVE BANK OF SAN FRANCISCO	APPLICATION DEVELOPER IV	162,500	SAN FRANCISCO, CA	05/27/2016	07/05/2016	
FEDERAL RESERVE BANK OF SAN FRANCISCO	APPLICATIONS INTEGRATION ENGINEER	123,000	SAN FRANCISCO, CA	12/21/2017	03/15/2018	
FEDERAL RESERVE BANK OF SAN FRANCISCO	AUTOMATION TEST & INTEGRATION ENGINEER	110,000	RICHMOND, VA	12/13/2017	12/13/2017	
FEDERAL RESERVE BANK OF SAN FRANCISCO	AUTOMATION TEST & INTEGRATION ENGINEER	110,000	RICHMOND, VA	12/13/2017	12/13/2017	
FEDERAL RESERVE BANK OF SAN FRANCISCO	DOCUMENT MANAGEMENT SENIOR ANALYST	110,860	SAN FRANCISCO, CA	02/07/2018	03/01/2018	
FEDERAL RESERVE BANK OF SAN FRANCISCO	DOCUMENT MANAGEMENT SENIOR ANALYST	110,860	SAN FRANCISCO, CA	03/02/2018	07/31/2018	
FEDERAL RESERVE BANK OF CLEVELAND	ECONOMIC ANALYST INTERN	56,000	CLEVELAND, OH	09/04/2015	10/01/2015	
FEDERAL RESERVE BANK OF SAN FRANCISCO	ECONOMIC RESEARCH ADVISOR	205,000	SAN FRANCISCO, CA	04/21/2016	08/15/2016	

28

Microdata Releases

AOL

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.

29

Microdata Releases

Netflix

Those fears were highlighted in December, when an in-the-closet lesbian mother sued Netflix for privacy invasion, alleging the movie-rental company made it possible for her to be outed when it disclosed insufficiently anonymous information about nearly half-a-million customers as part of its \$1 million contest.

The federal suit claimed Netflix violated fair-trade laws and a federal privacy law designed to protect video rental records when the Los Gatos, California, company launched the popular contest in 2006. The FTC also contacted Netflix about the first contest, which lasted three years, according to a Netflix blog post Friday.

30



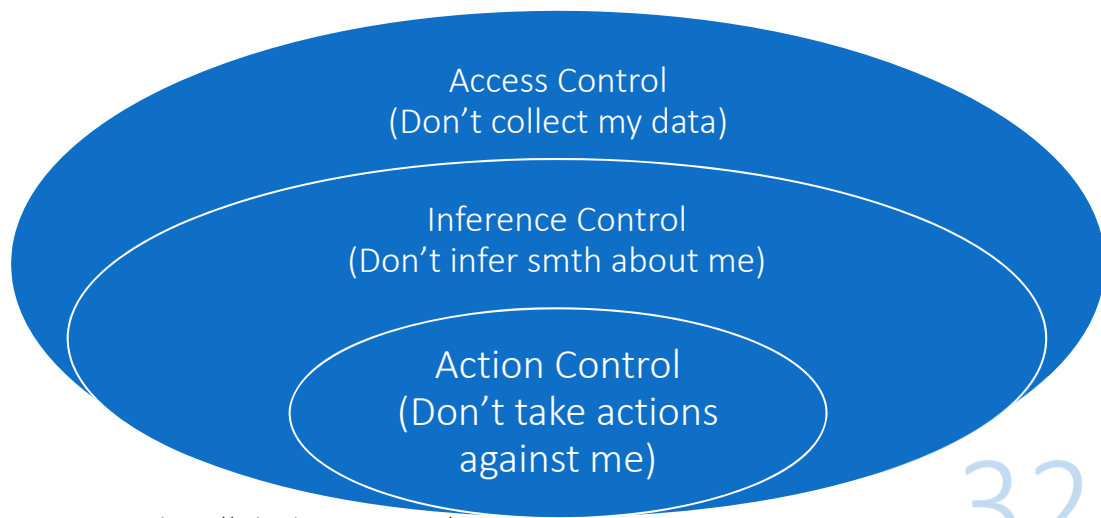
Challenges and Opportunities

1. We can quickly face high privacy risks
2. Researchers need to value **appropriate flow**
3. **Infrastructure** needed – physical and mental

31

31

Consent to give up control



Ghani 2018: Presentation in <https://coleridgeinitiative.org/>

32

32

The data you *already provided* to us would be **much more (gain frame) / much less (loss frame)** valuable if you would allow us to link them with Do you agree?

Web	Back	Total
% agree: gain	62.4	520
% agree: loss	75.4	489
Total	498	1009

Phone	Front	Back	Total n
% agree	90.8	78.7	598

Web	Front	Back	Total
% agree	82.6	62.4	520

The data you are **about to provide (front) / already provided (back)** to us would be **much more** valuable if you would allow us to link them with Do you agree?

Sakshaug et al. 2018

33

33

Onboarding

Linkage to PASS

Thank you for downloading the IAB-SMART-App and participating in our study.

An important component of our study will be the evaluation of the app together with the results from the "Quality of Life and Social Welfare" survey. Therefore, we need your consent. You are free to revoke your consent at any time.

I agree that the data from this app will be evaluated together with the data from the Quality of Life and Social Welfare survey. ☐

Quit **Continue**

Registration

Thank you for downloading the IAB-SMART-App and participating in our study.

An important component of our study will be the evaluation of the app together with the results from the "Quality of Life and Social Welfare" survey. Therefore, we need your consent. You are free to revoke your consent at any time.

Registration

Please enter your registration code.

Registration code

CANCEL **OK**

Quit **Continue**

Consent to data processing

Terms of Use and Privacy Policy

Names and addresses will be strictly separated from the collected app data. Data will be analysed in such a way that no conclusions about your identity are possible.

Please read the [data privacy policy](#) and the [terms of use](#) carefully and agree to the data processing.

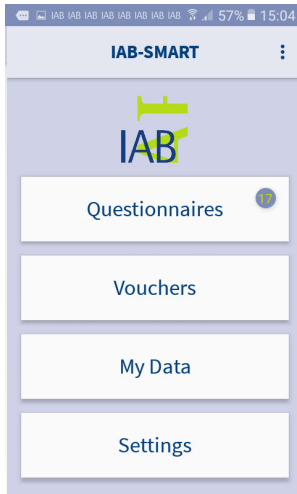
Hereby, I agree to the data processing and accept the terms of use. ☐

Back **Continue**

// Page 34

34

Withdrawing Consent

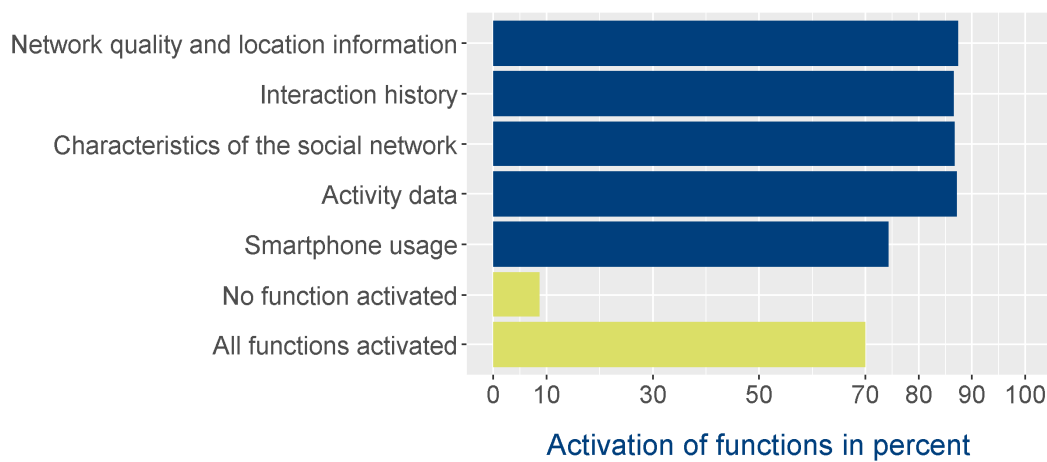


- Network quality and location information (every half hour)
- Interaction history
- Characteristics of the social network
- Activity data (every two minutes)
- Smartphone usage

// Page 35

35

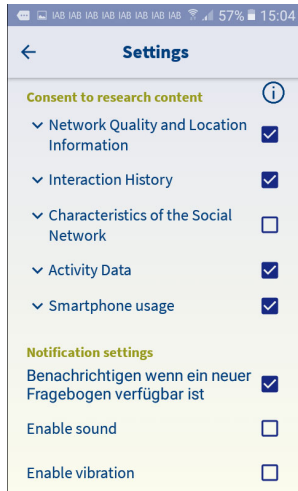
Which functions get activated?



// Page 36

36

Withdrawing consent



Overall, 129 (18.8%) individuals have made 590 changes

- 201 deactivations
- 389 activations

// Page 37

37



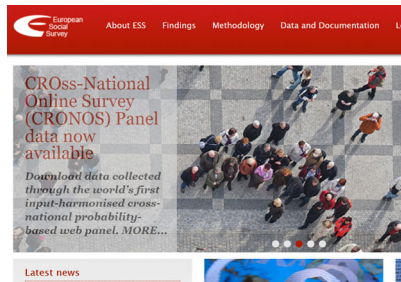
Challenges and Opportunities

1. We can quickly face high privacy risks
2. Researchers need to value appropriate flow
3. **Infrastructure** needed – physical and mental

38

38

Tiered Access Solutions



For this type of data no evidence of privacy risks and therefore no known benefit from implementing privacy guarantees.

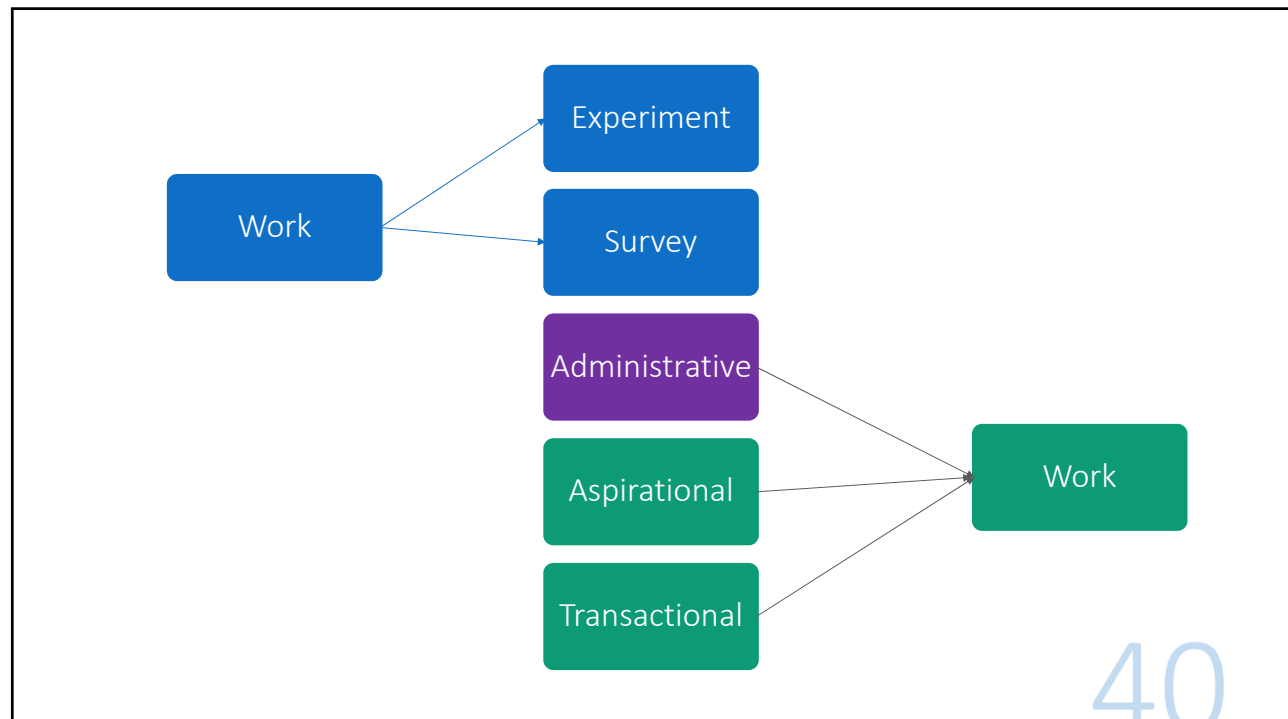


Safe data, safe people, in safe environment. Research needed to automatize agreements, input and output control.

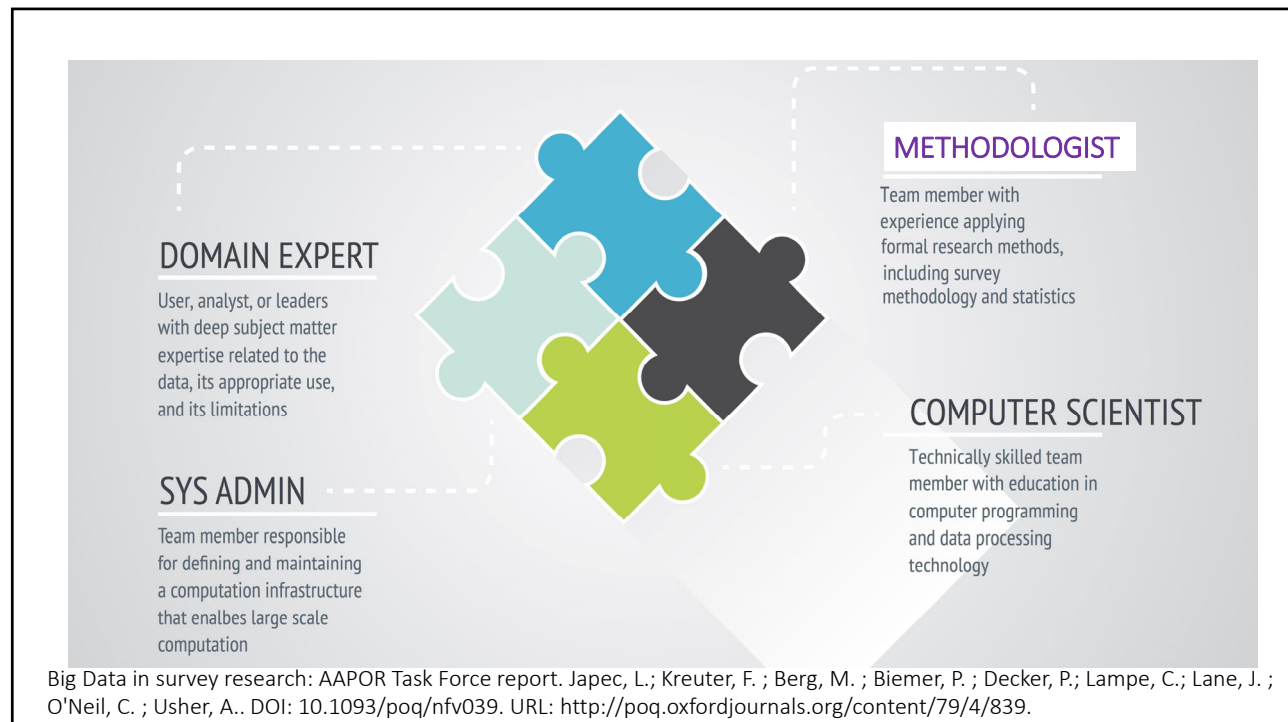


Raw data - out of researchers' reach. Federated learning. Differential privacy are being explored.

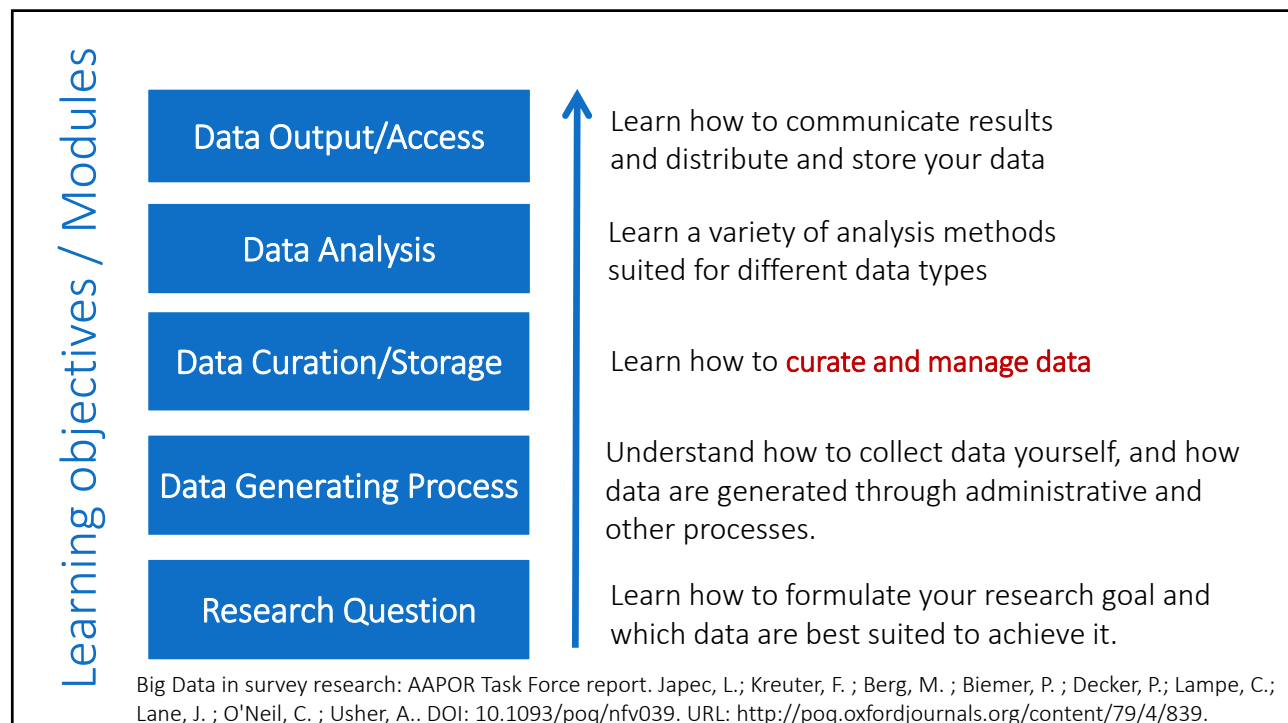
39




40



41



42



http://coleridgeinitiative.org

COLERIDGE INITIATIVE

Training Computing Connecting Mailing List

A collaboration presented by CHICAGO NYU MARYLAND


Home Program Project Contact FAQs Imprint Search **APPLY NOW**

http://survey-data-science.net/


INTERNATIONAL PROGRAM IN SURVEY AND DATA SCIENCE

offered through the University of Mannheim and the Joint Program in Survey Methodology (Universities of Maryland and Michigan, West)


BE PART OF IT



SPONSORED BY THE



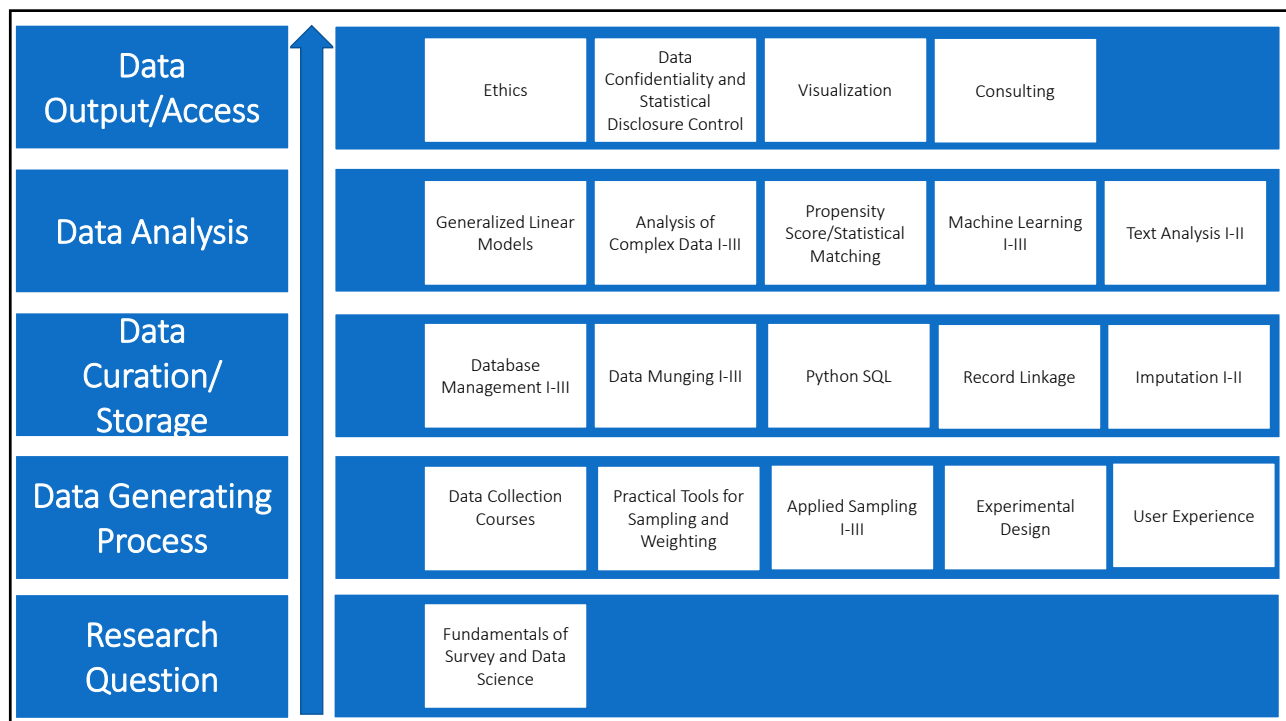
Federal Ministry of Education and Research



AUFSTIEG DURCH BILDUNG >>
OFFENE HOCHSCHULEN

43

43



44



Summary

1. Great potential: **New questions** can be asked
2. **Inference issues** and **data quality** questions remain
3. **Privacy** needs to be considered at the **design stage**
4. It is important to **empower** oneself and those around us
5. Thanks to the **ASA for organizing this workshop!**

45

Thank You



Frauke Kreuter @fraukolos
JPSM - Uni Maryland
SOWI - Uni Mannheim – IAB

46