

Alternative Data Sources in Official Statistics

Frauke Kreuter

JPSM - Uni Maryland
University of Munich

@fraukolos

February 2022

Survey
Statistics
Perspective

Research
Examples

Privacy





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 Japac, Co-Chair, Statistics Sweden
Frauke 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.

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

BIG DATA AND SOCIAL SCIENCE

A Practical Guide to Methods and Tools



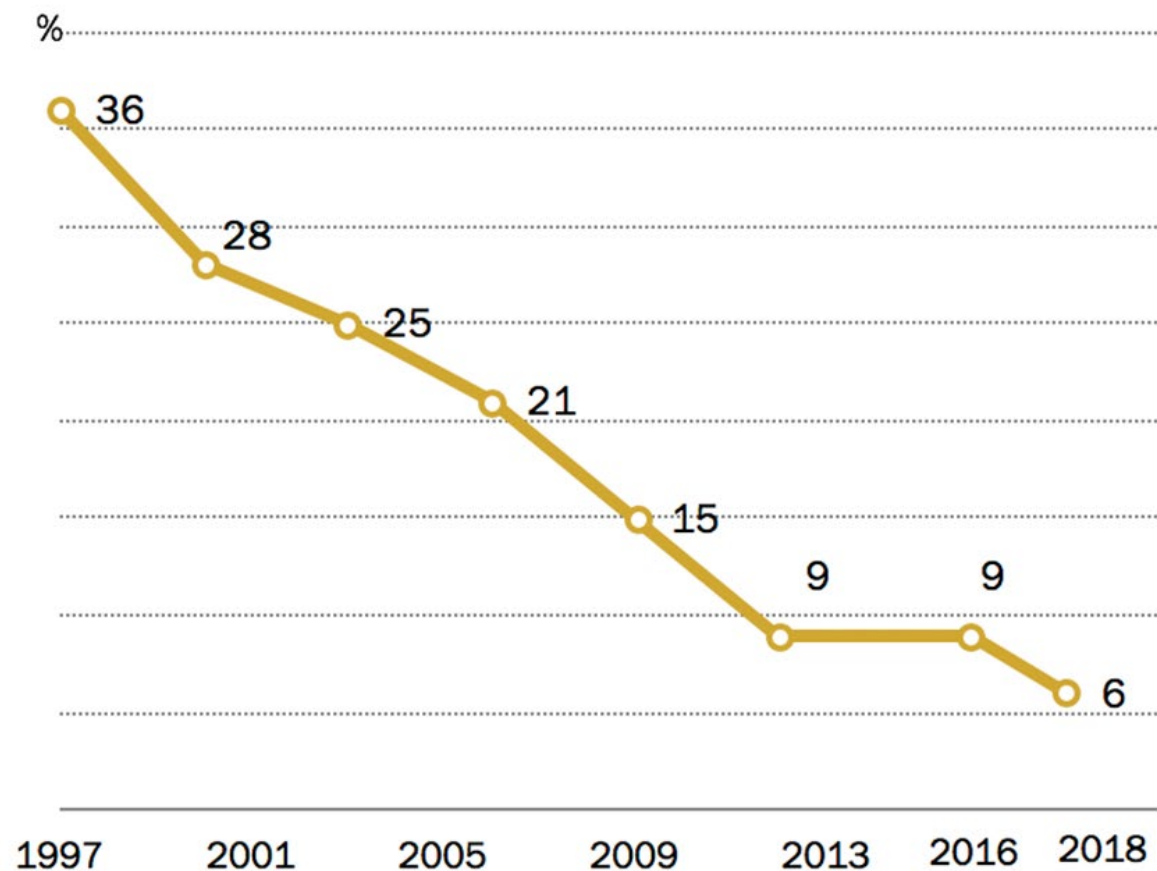
Edited by
**Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane**

 CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

<https://textbook.coleridgeinitiative.org/>

After brief plateau, telephone survey response rates have fallen again

Response rate by year (%)



Note: Response rate is AAPOR RR3. Only landlines sampled 1997-2006. Rates are typical for surveys conducted in each year.

Source: Pew Research Center telephone surveys conducted 1997-2018.

PEW RESEARCH CENTER

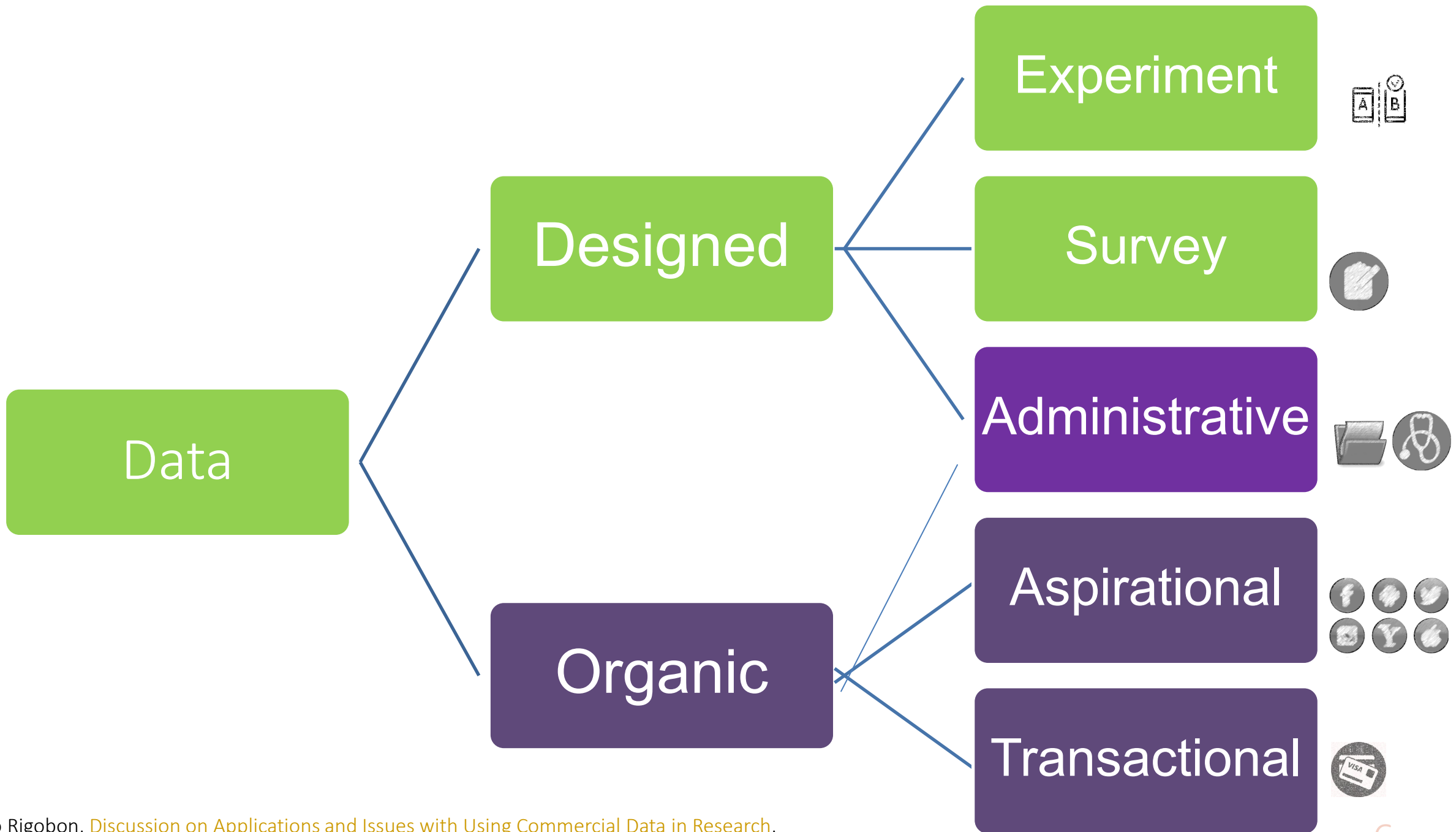
FEDERAL STATISTICS, MULTIPLE DATA SOURCES, AND PRIVACY PROTECTION

Next Steps



Survey-Statistician Perspective

1. Alternative data sources can **enhance our measurements**
2. **Purposeful design** is needed for success
3. **Data generating processes** need to be understood



Prediction of Initial Claims for Unemployment Insurance

The chart presents a prediction of Initial Claims for Unemployment Insurance using the University of Michigan Social Media Job Loss Index. The prediction is based on a factor analysis of social media messages mentioning job loss and related outcomes. See [Using Social Media to Measure Labor Market Flows](#) for details.

This research is a collaboration of University of Michigan's Institute for Social Research, Department of Economics, and Department of Electrical Engineering and Computer Science and Stanford University's Department of Computer Science. The Economic Indicators from Social Media project is part of the Michigan Node of the NSF-Census Research Data Network (NSF SES 1131500). You can find relevant academic papers about this work [here](#).

About this website: The computational and data infrastructure that powers this website is described [here](#).

For more information:

Matthew Shapiro, shapiro at umich.edu, (734) 764-5419 (Economics)
 Michael Cafarella, michjc at umich.edu, (734) 764-9418 (Computer Science)

Update (June, 2015)

We are currently in the process of revisiting our original model, which began to deviate in its estimates around mid-2014. We will be updating this site soon with our new model, along with details on our new model.

If you would like to view the original model's results, [click here](#).

Sources: *Initial Claims for Unemployment Insurance (seasonally adjusted)*, U.S. Department of Labor; *Prediction*, University of Michigan Social Media Job Loss Index.

Latest Estimate

 [download estimates](#)

Date	Initial Claims (Preliminary)	Initial Claims (Revised)	Prediction
July 15, 2017	233	n/a	296

Job Vacancy Prediction

Big Data ESSNet
presented in Sofia. 24-25 February 2017

- United Kingdom (lead)
- Germany
- Sweden
- Slovenia
- Italy
- Greece



Job Portal Ads



Pre-processed

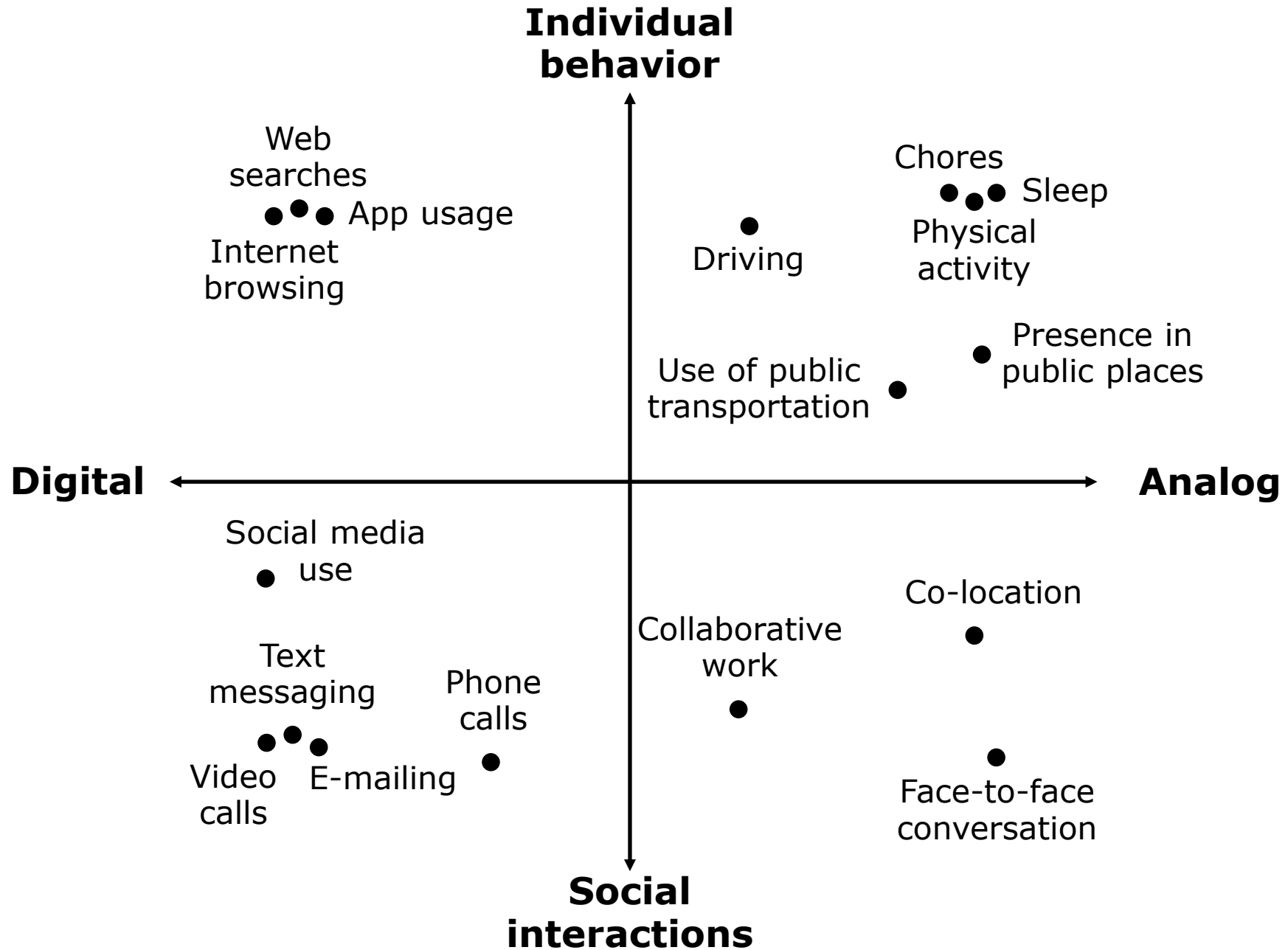


Data Processing –
Deduplication

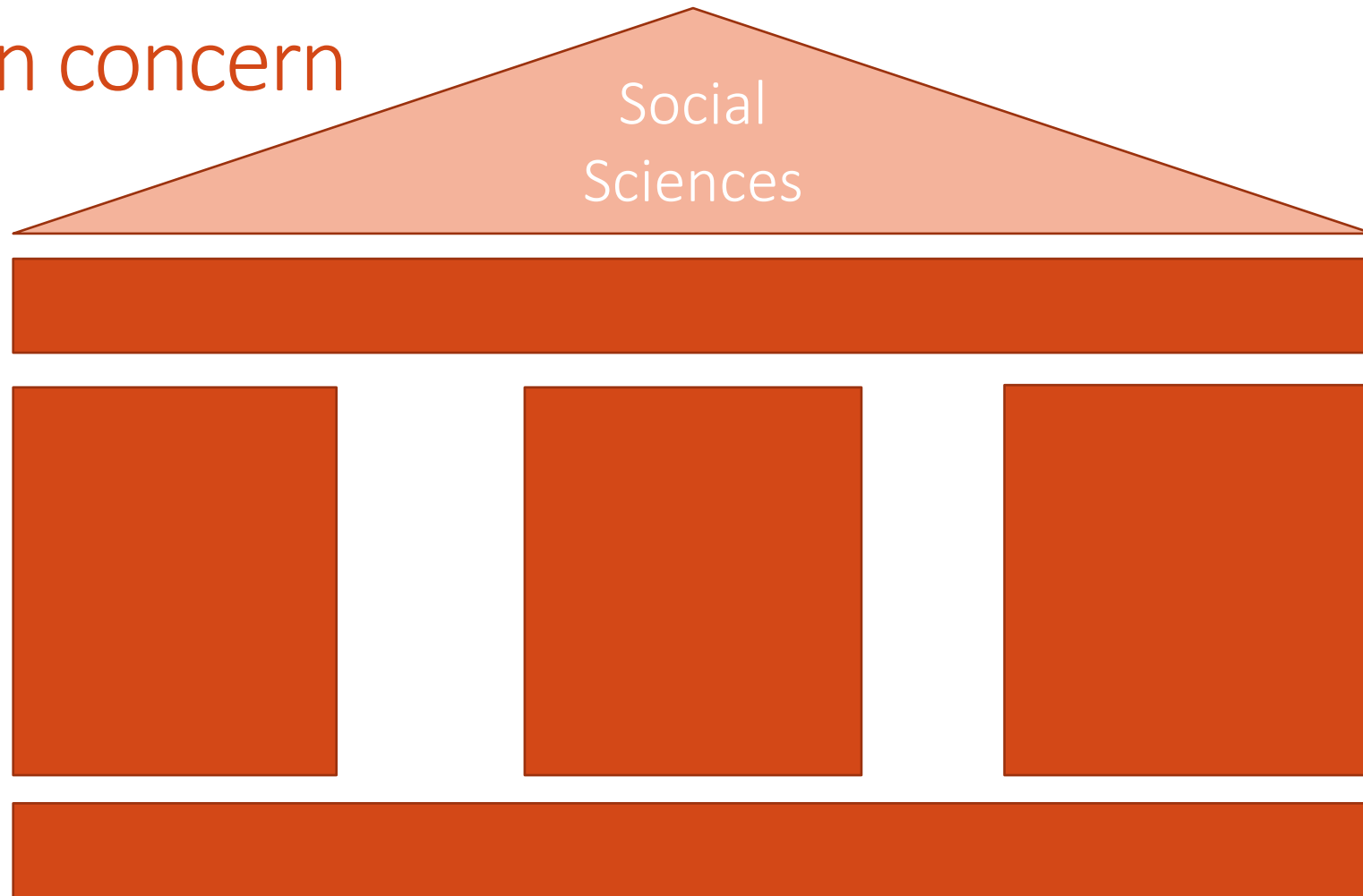


Data Analysis

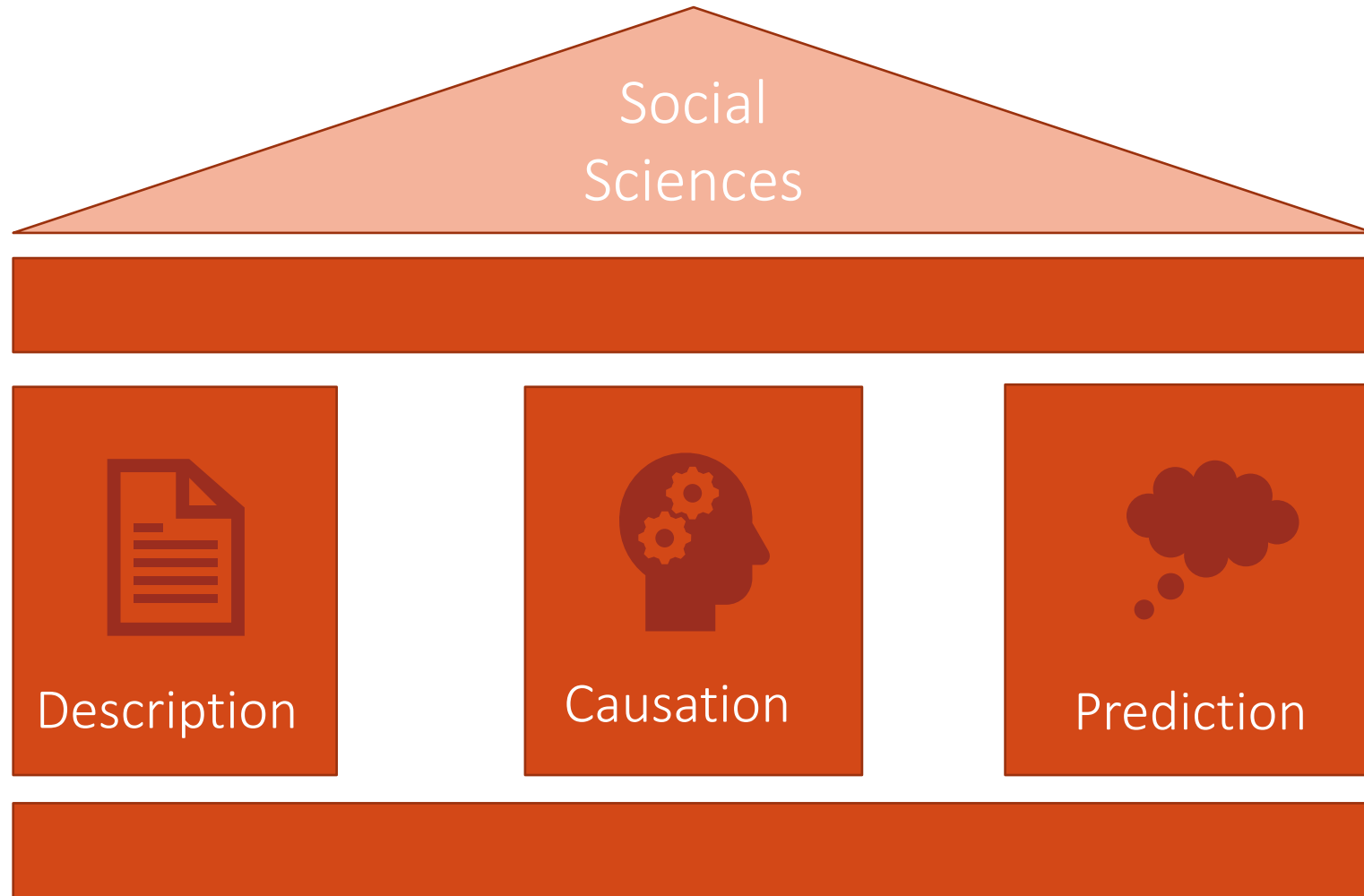
Organic – Behavioral



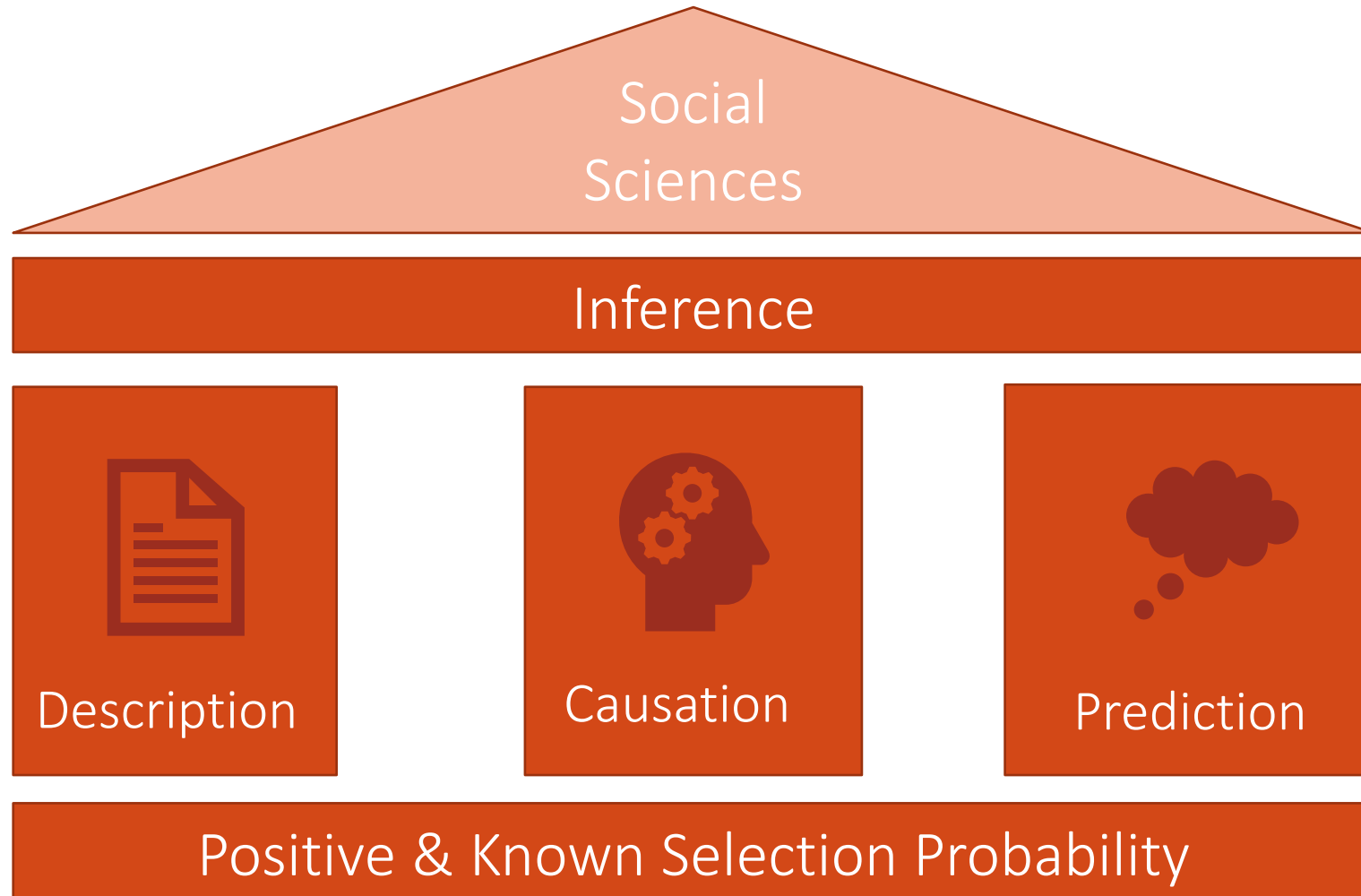
Common concern



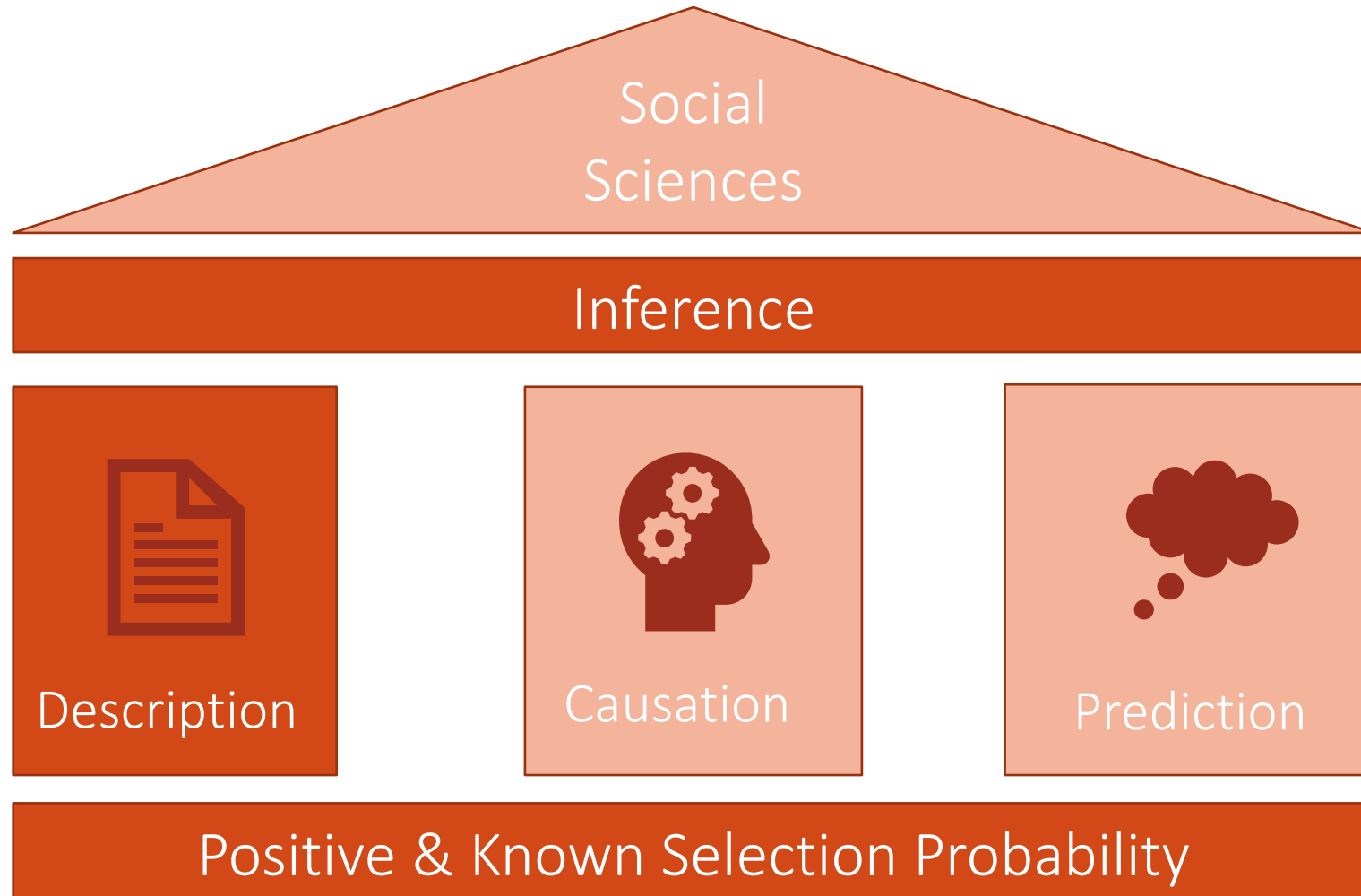
Kohler, Kreuter, Stuart 2019, Annual Review of Statistics and its Application, 6, 149-172



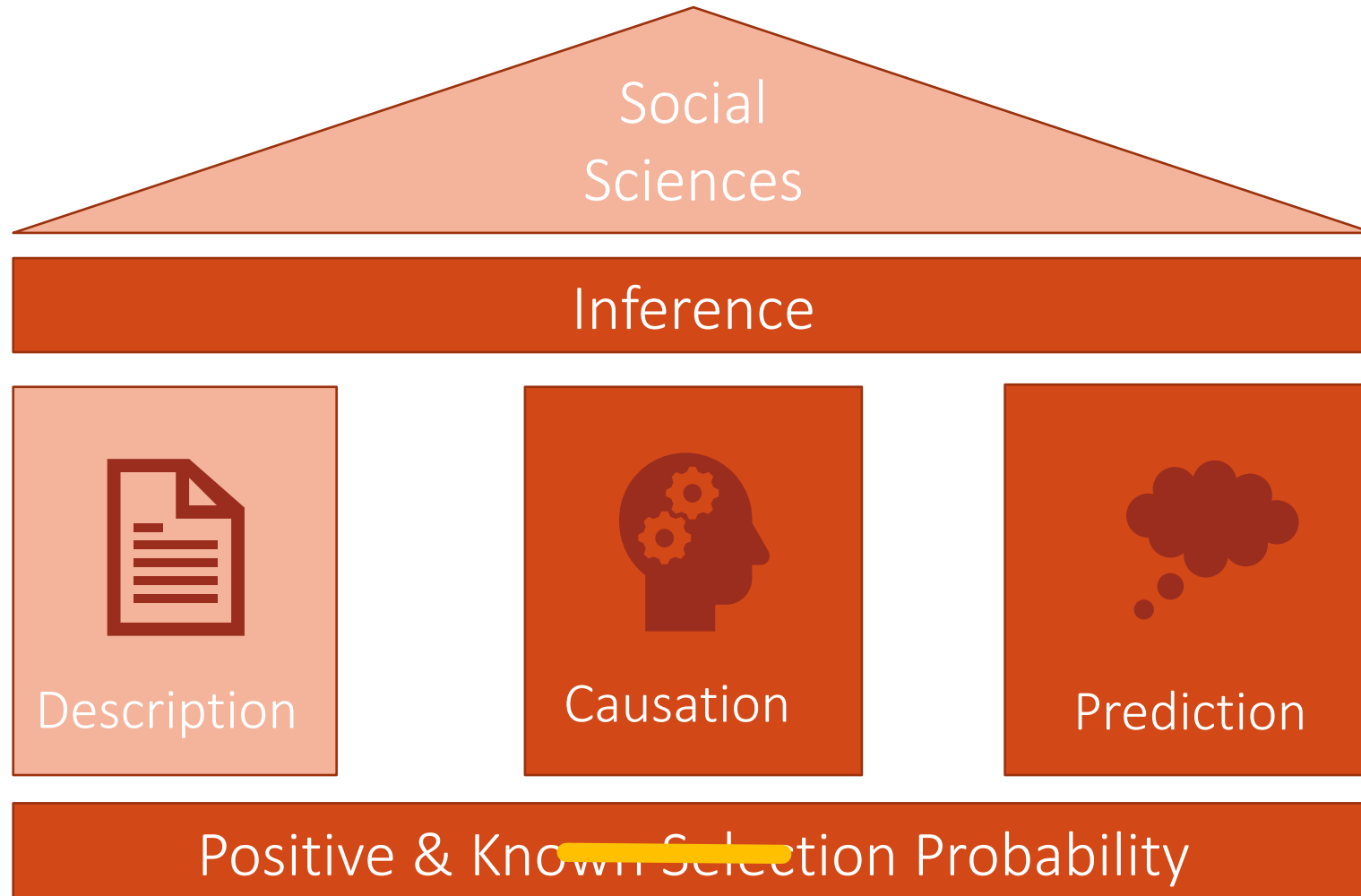
Kohler, Kreuter, Stuart 2019, Annual Review of Statistics and its Application, 6, 149-172



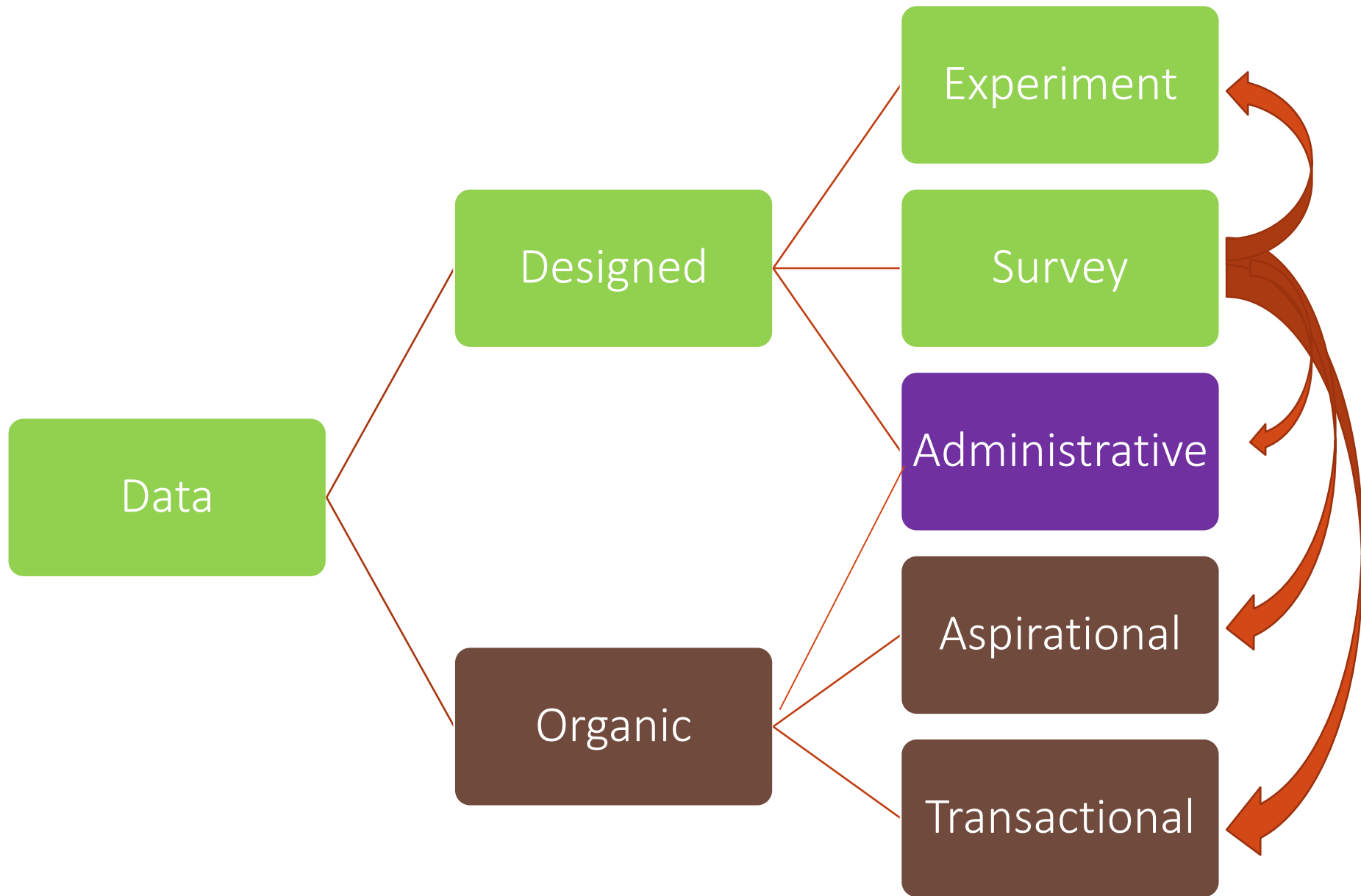
Kohler, Kreuter, Stuart 2019, Annual Review of Statistics and its Application, 6, 149-172



Kohler, Kreuter, Stuart 2019, Annual Review of Statistics and its Application, 6, 149-172



Kohler, Kreuter, Stuart 2019, Annual Review of Statistics and its Application, 6, 149-172



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



VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights



VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights

Behavioral (Big) Data



VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights

Surveys



VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights

Designed Product

collection

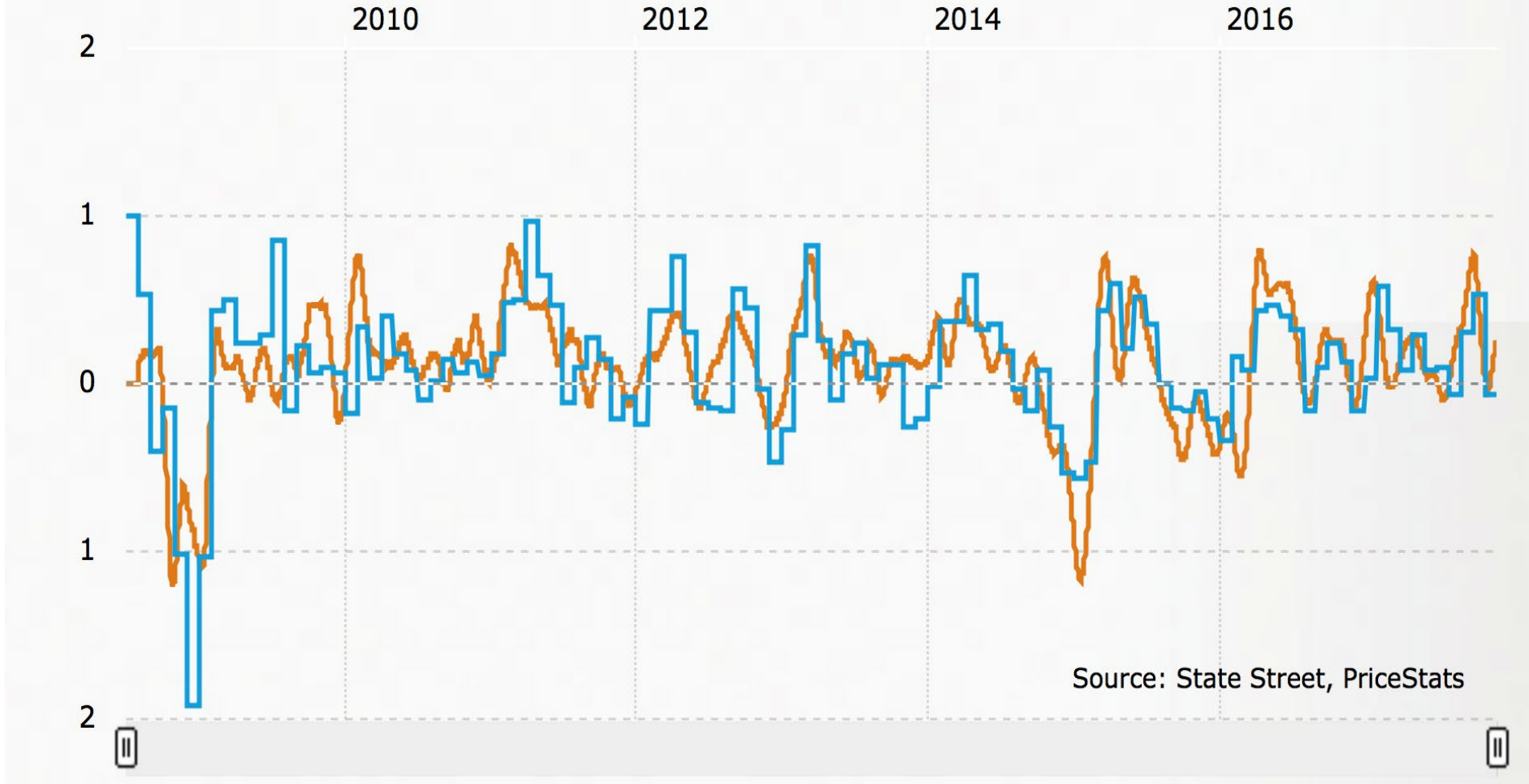
One way to think about a data ~~analysis~~ is to think of it as a product to be designed. [...] Producing a useful product requires careful consideration of who will be using it.

Roger Peng, 2018

US Aggregate Inflation Series

(Monthly Rate, 2008 - Present)

- Official CPI
- PriceStats Index



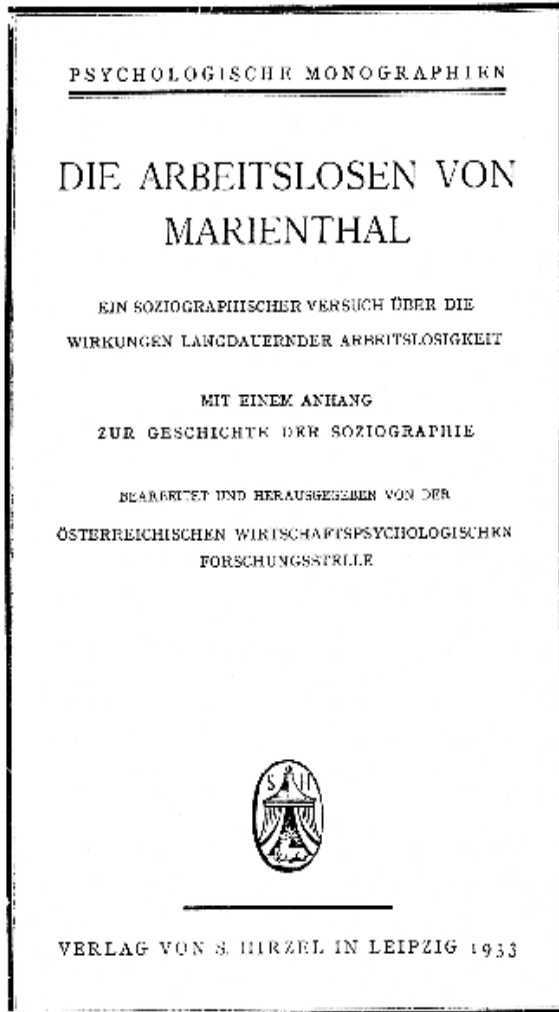




Example 1 – Economic Research

1. Old measurements possible at scale with new devices
2. Coverage error and non-participation error detection requires careful design and combined data
3. Measurement error detection will keep us busy for a while

Effects of Unemployment?



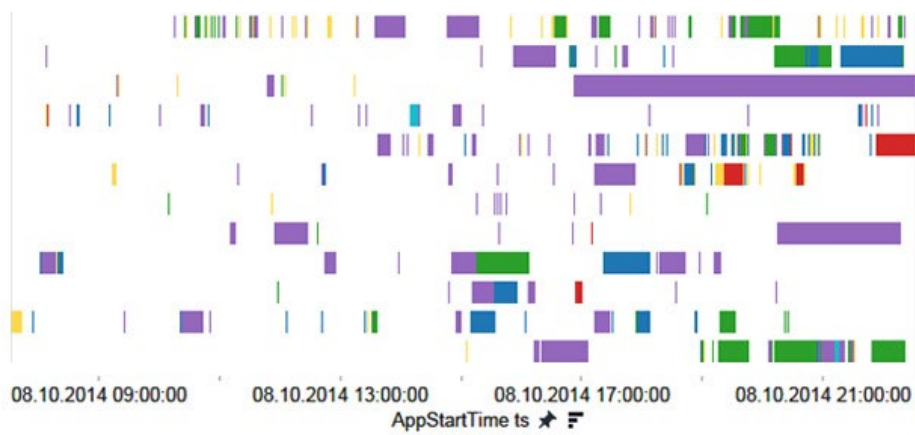
**The Sociography of an
Unemployed Community**

Marie Jahoda, Paul F. Lazarsfeld,
and Hans Zeisel

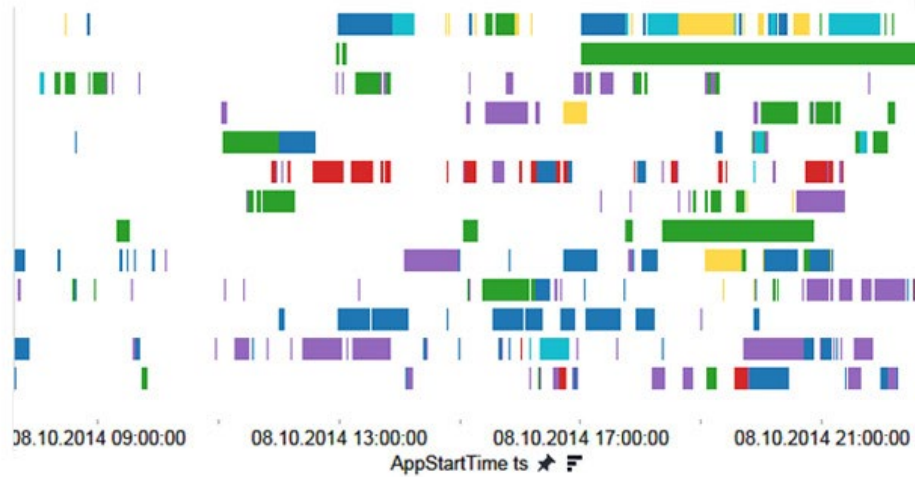


Source: Archives for the History of Sociology in Austria (Graz), »Marienthal« Virtual Archives

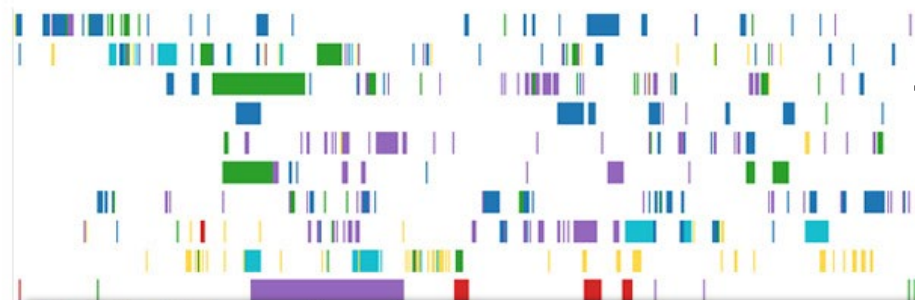




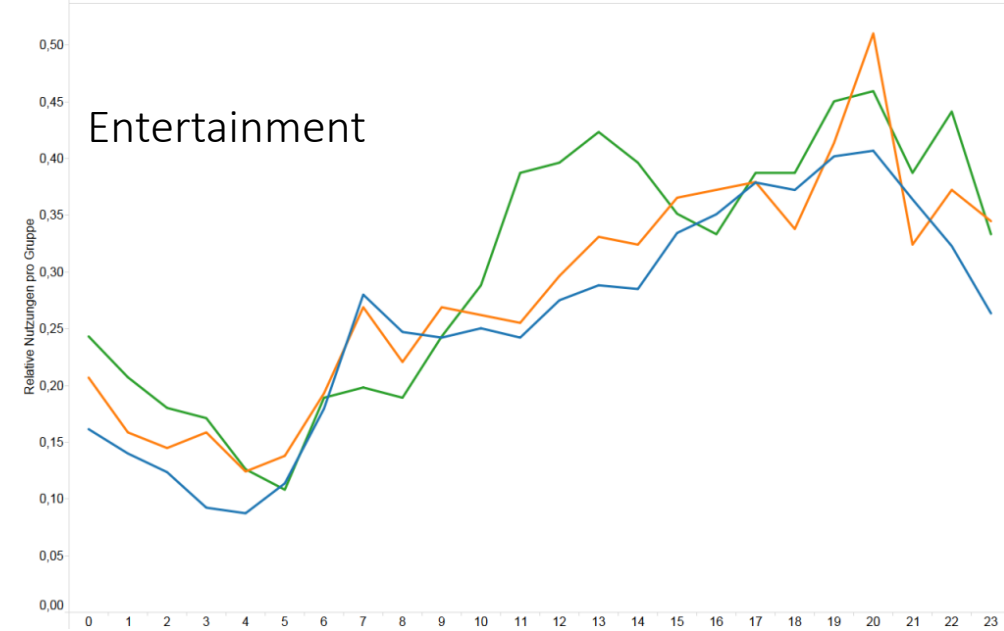
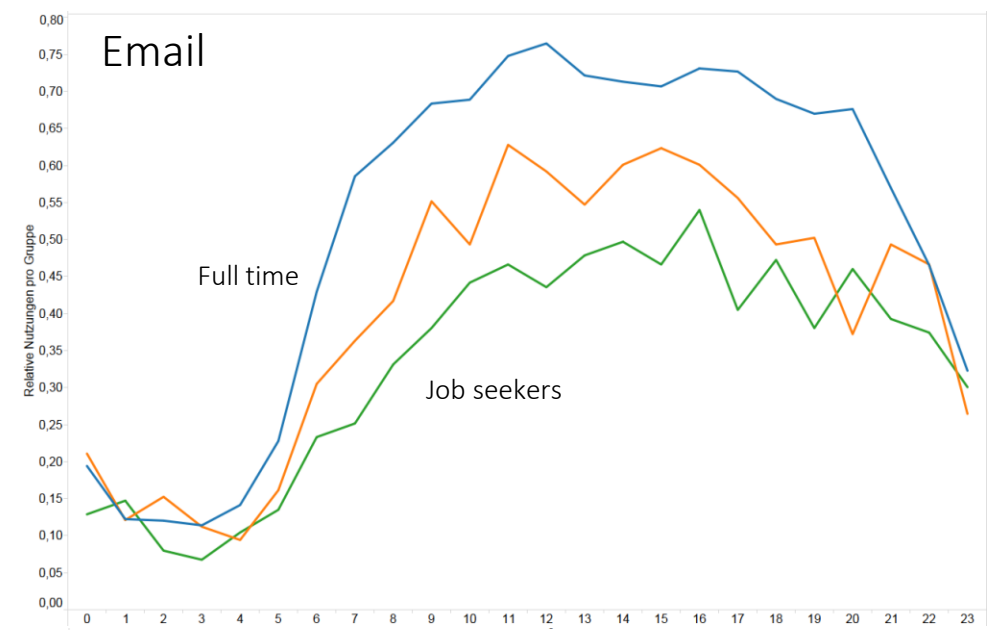
Full-time employed
→ App use past 5pm



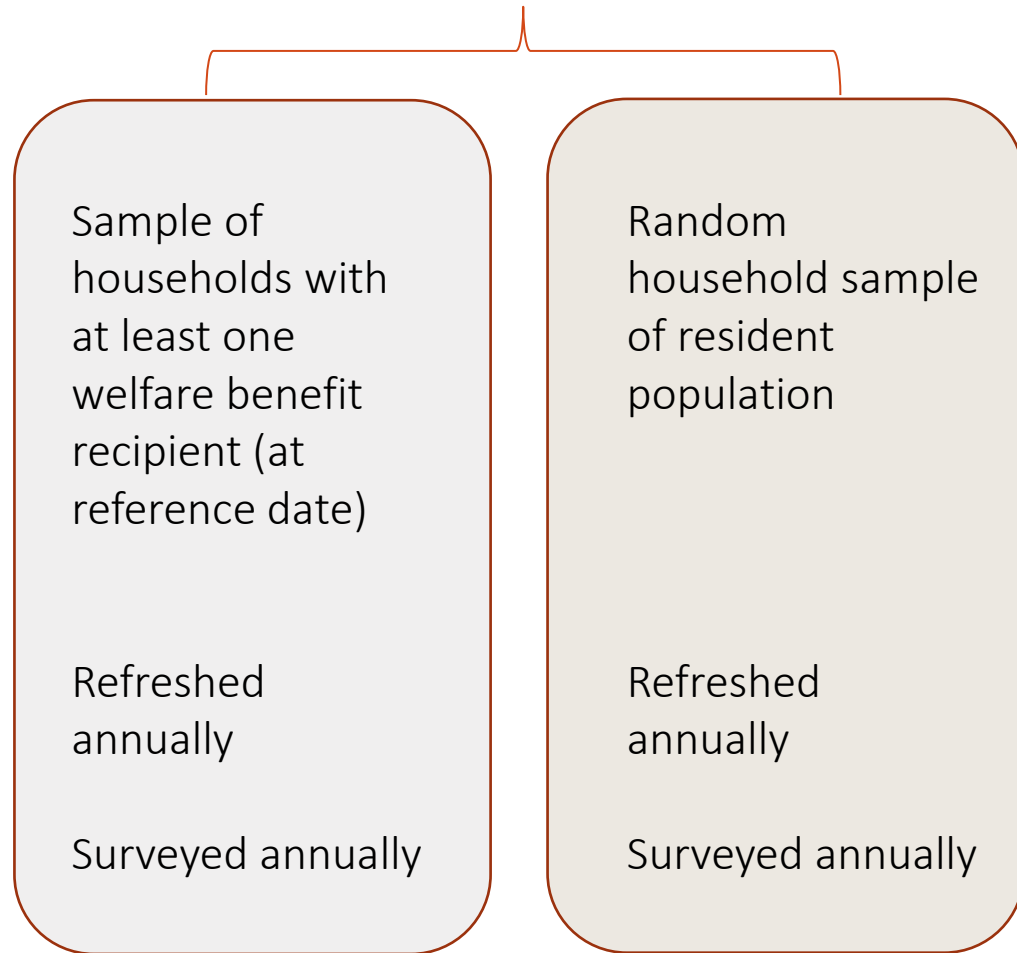
Part-time employed
→ App use at noon



Job seekers
→ Continuous app use



PASS – Panel (10 years) + Administrative Data



Meldung zur Sozialversicherung

Personalauswahl

Versicherungsnummer Personalnummer (freiwillige Angabe)

Name Vorsatz Zusatz Titel

Vorname

Straße und Hausnummer (Anschrift nur bei Anmeldung und Anschriftenänderung)

(Land) Postleitzahl Wohnort

Grund der Abgabe Entgelt in Gleitzone Namensänderung

Beschäftigungszeit

von bis Betriebsnummer des Arbeitgebers Personengruppe Mehrfachbeschäftigung Betriebsstätte Ost West

Beitragsgruppen KV RV ALV PV Angaben zur Tätigkeit Aktuelle Staatsangehörigkeit

Beitragspflichtiges Bruttoarbeitsentgelt (in DM ohne Pfennige / Euro ohne Cent) DM Euro Statuskennzeichen

Wenn keine Versicherungsnummer angegeben werden kann:

Geburtsname Vorsatz Zusatz Geburtsort

Geburtsdatum Geschlecht männlich weiblich

Inference to Population

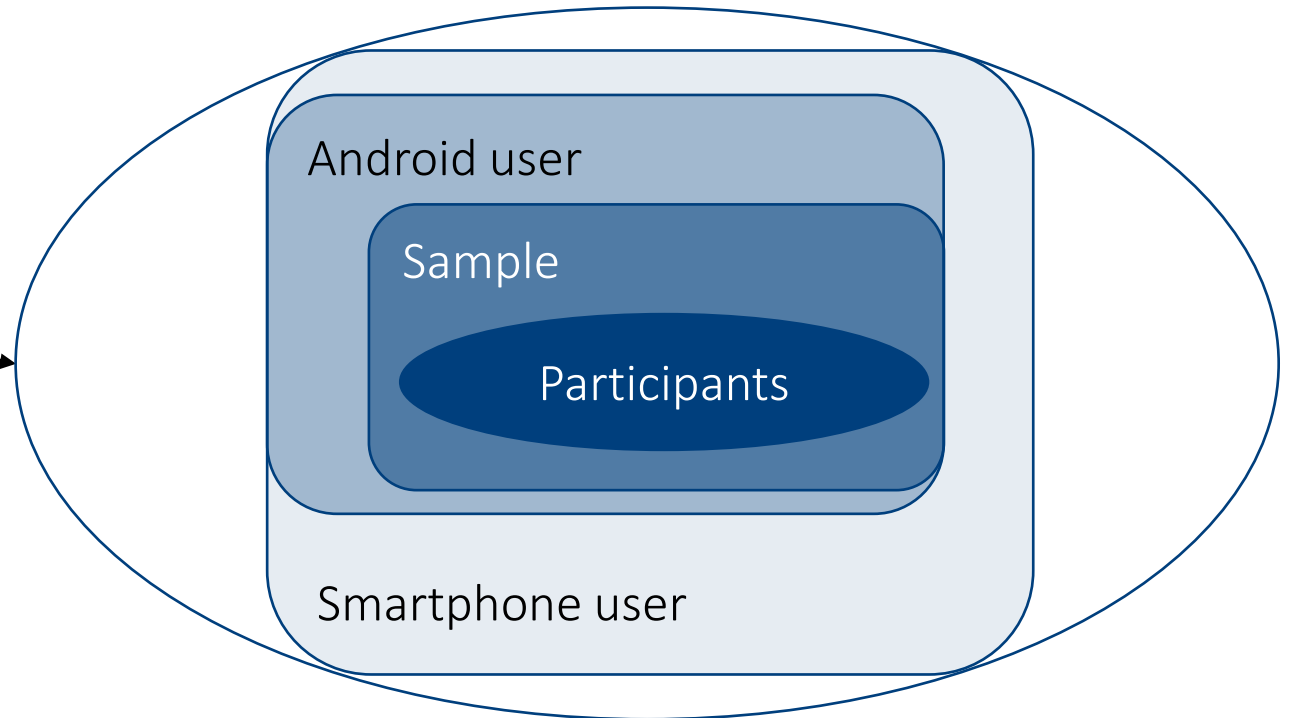
...owning a (specific) smartphone

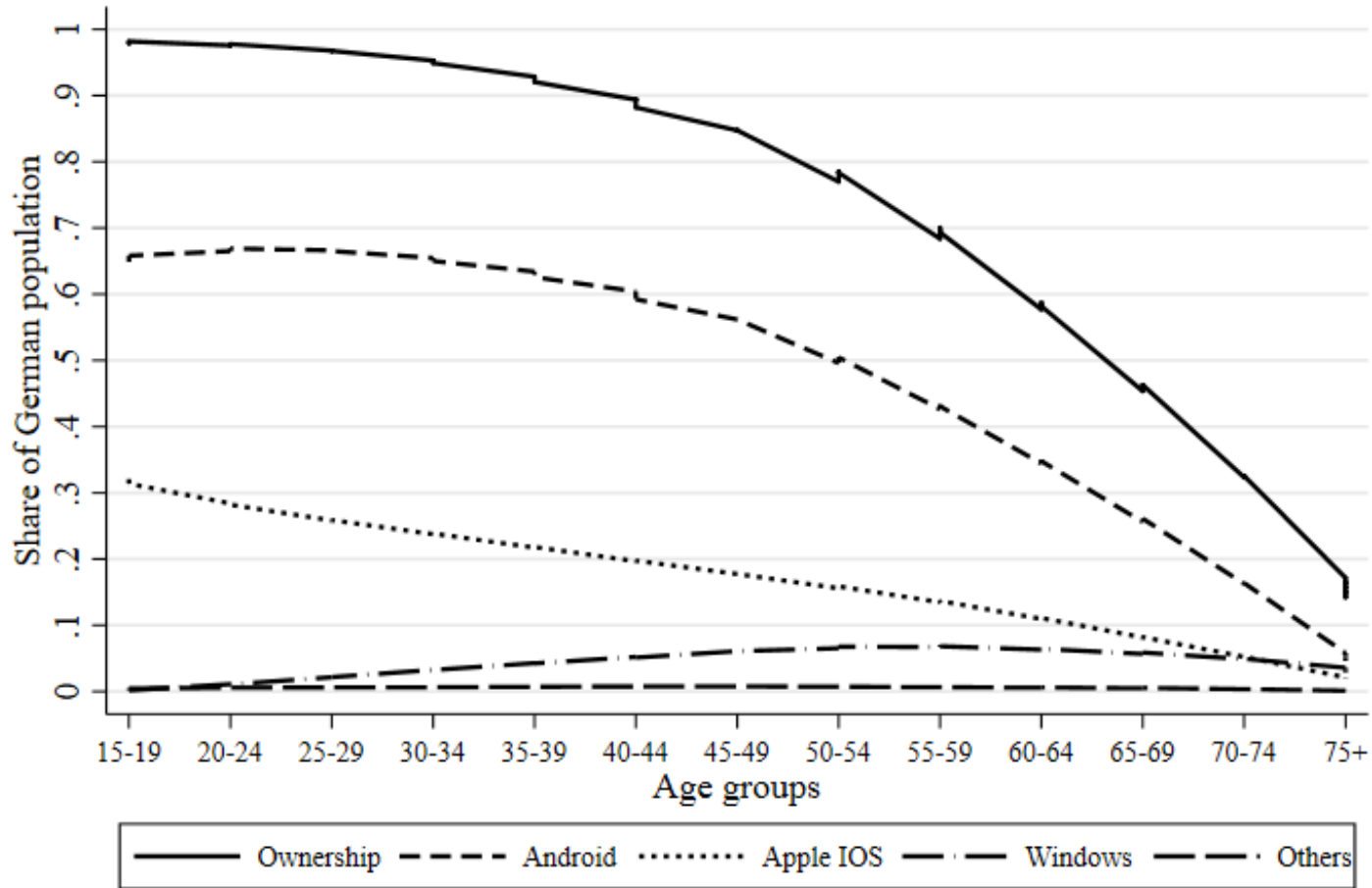
...being able to download an app

...being willing to download an app

} Nonparticipation error

Population
German Residents
PASS Panel at IAB
Wave 11 question on
smart phone use & OS



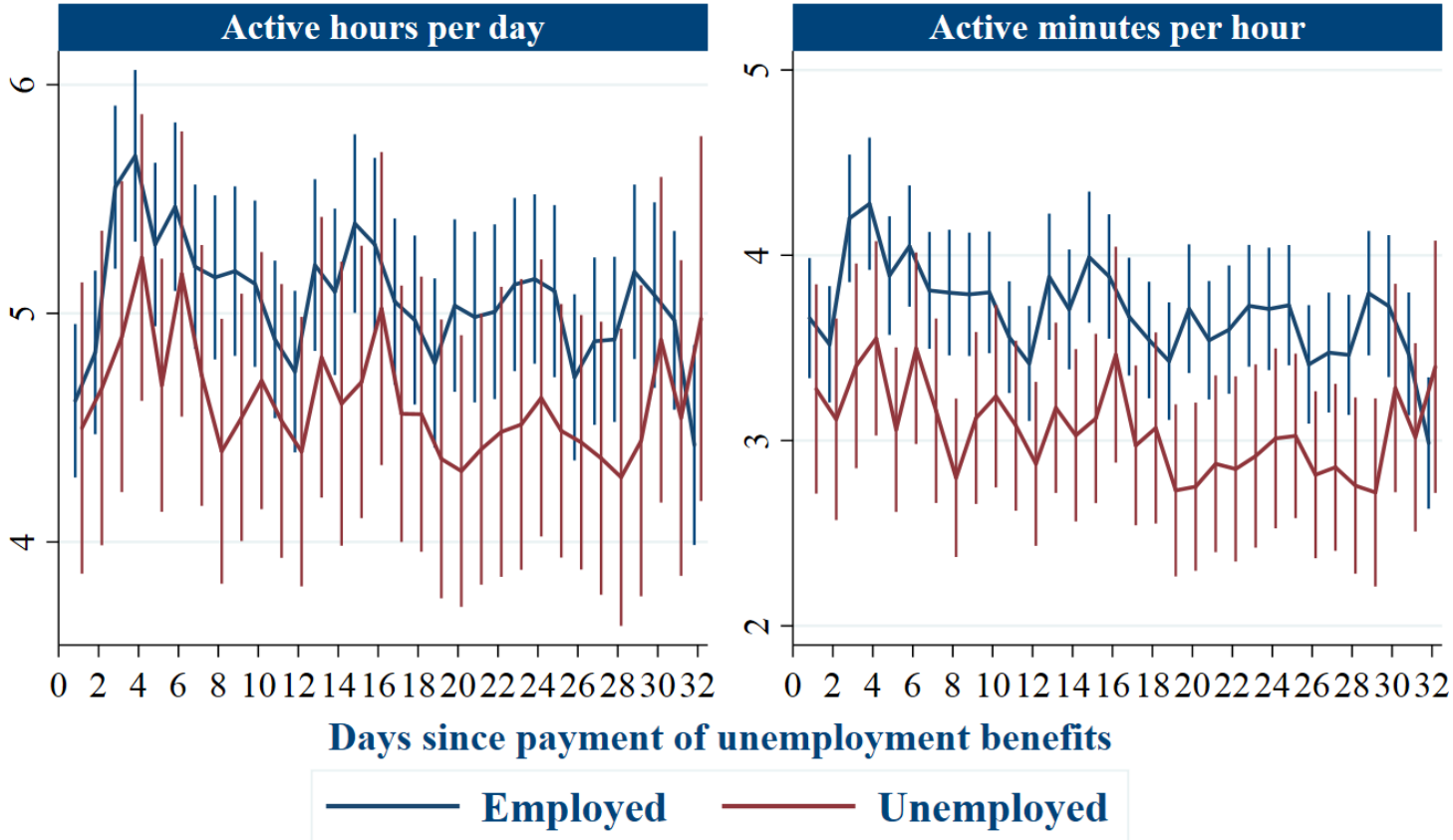


Smartphone ownership also correlates with...

- Educational attainment (higher)
- Immigrant (less likely)
- Region (less in East)
- Community size (smaller less)

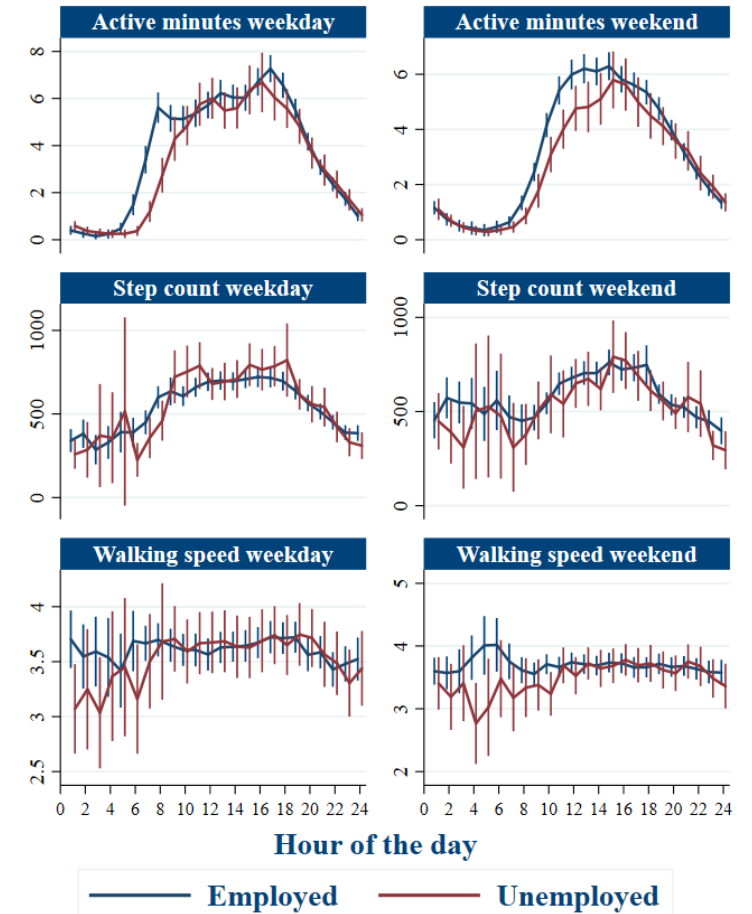
Sneak Peak

Rhythm of the payment of unemployment benefits



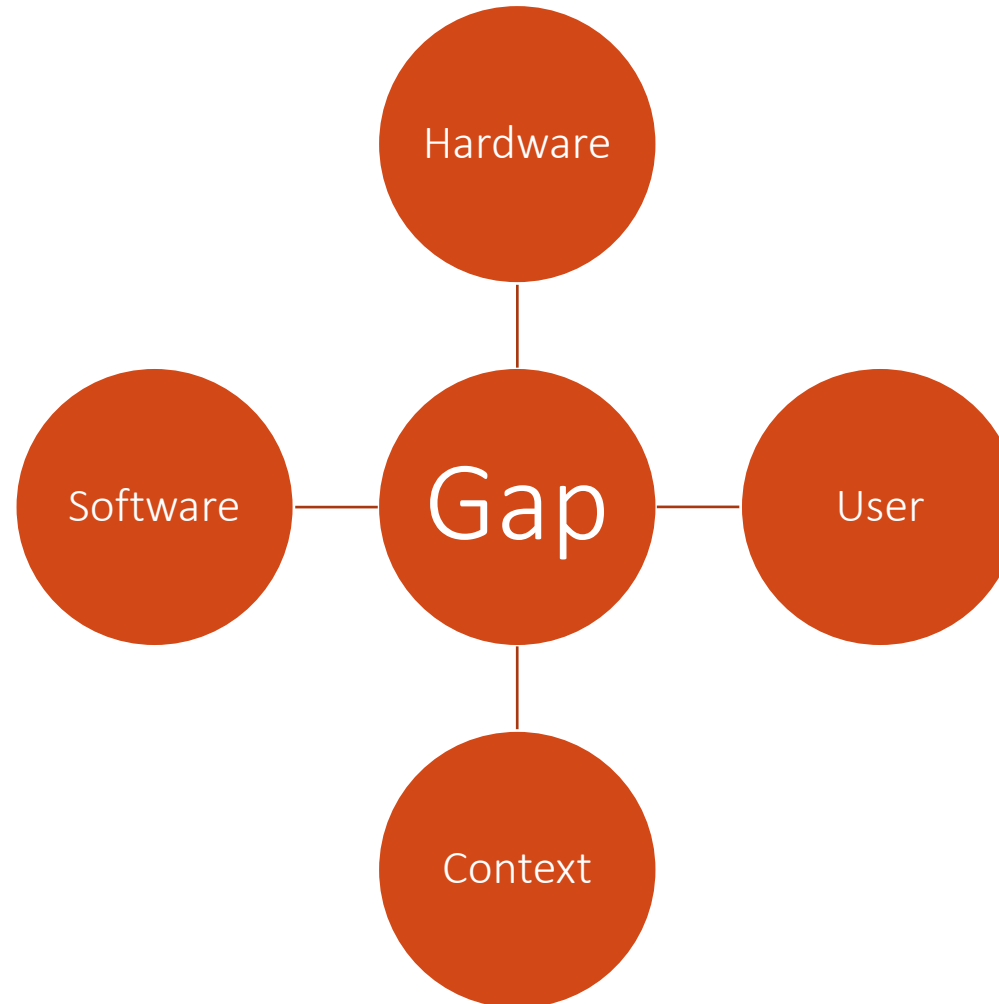
Predictive Margins with 95% confidence intervals.
 Controls: Gender, age, weekday, hours smartphone is kept nearby.

Loss of day structure / resignation

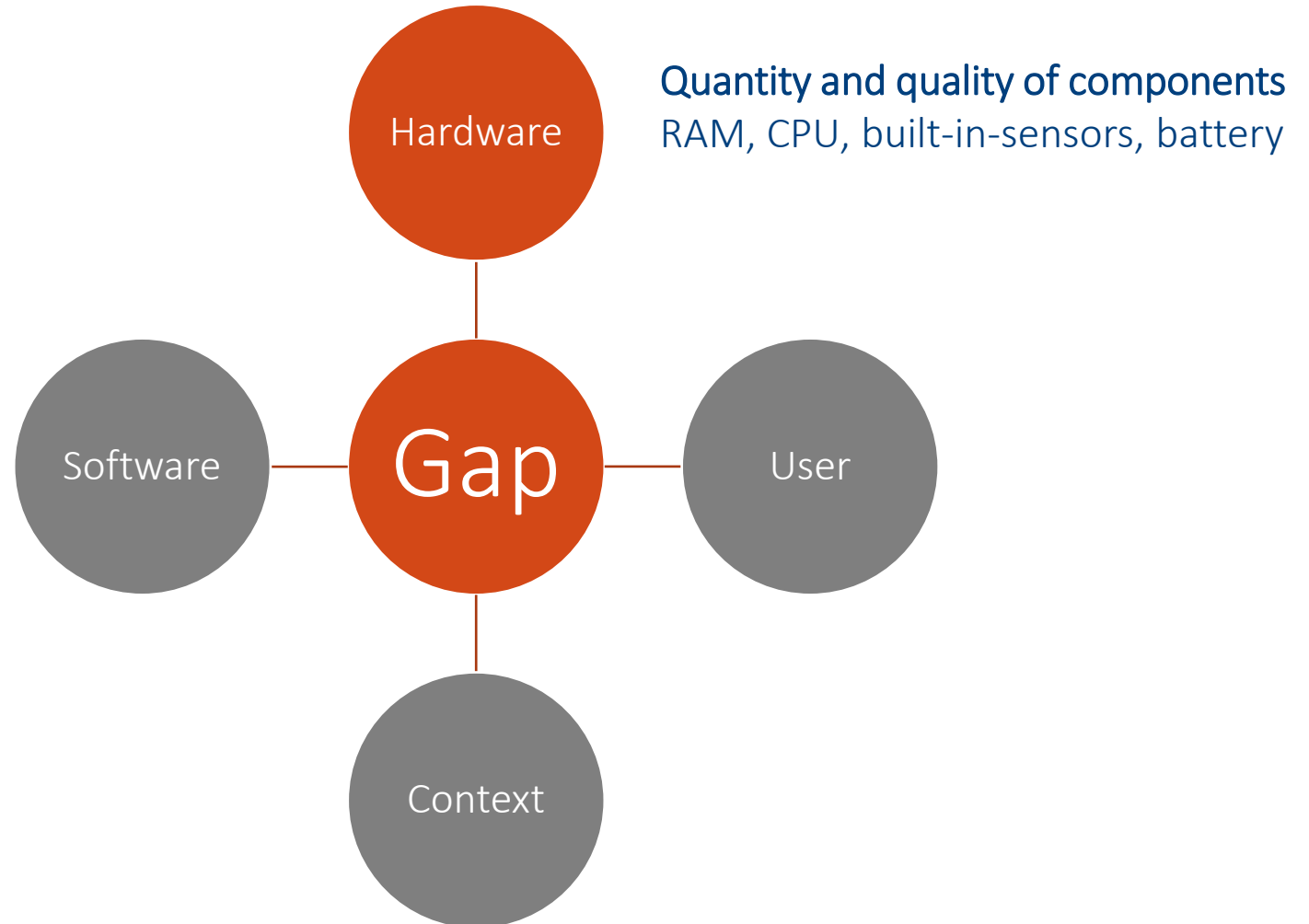


Predictive Margins with 95% confidence intervals.
 Controls: Gender, age, hours smartphone is kept nearby.

Error sources

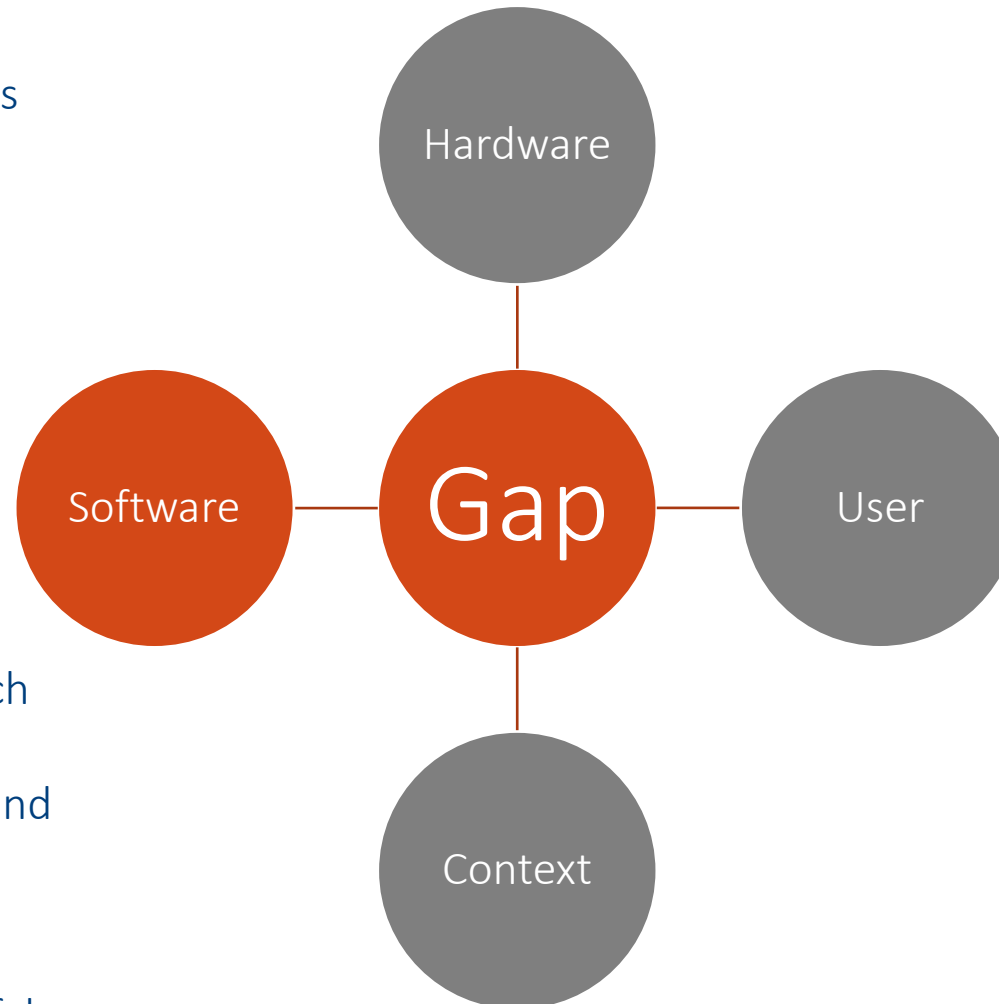


Error sources

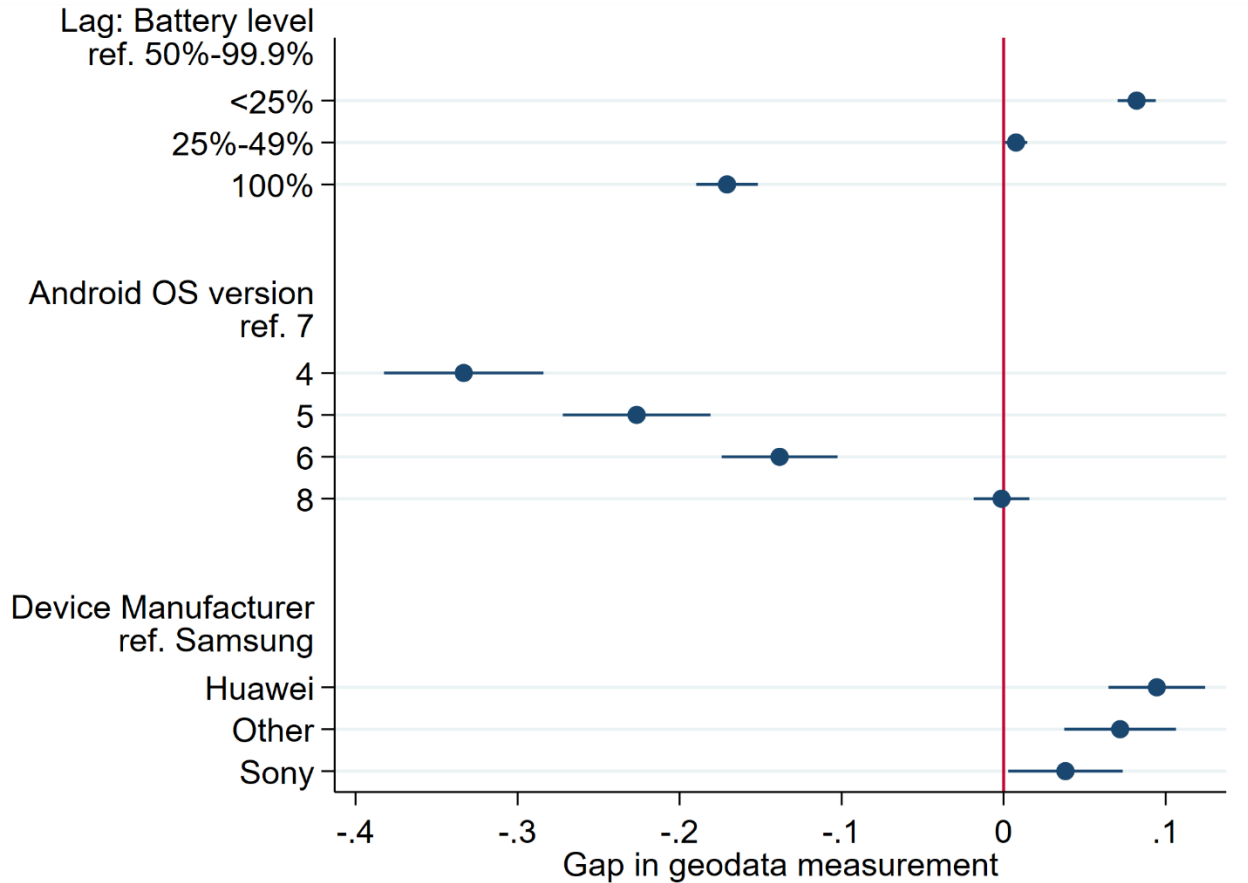


Error sources

- **Manufacturer Settings**
Device specific doze-/battery saving modes inhibit data collection
- **Operating System Settings**
Data collection may be inhibited by the Operating System (OS)
OS versions may vary in their rights management
- **Research App Settings**
How the research app collects the data (what, when, where, for how long, at which interval, from whom)
Interacts with device / OS / user: battery and RAM/CPU drain
- **Third Party Apps**
Battery saving apps, Task-killer apps, GPS faker apps



Device-related error sources



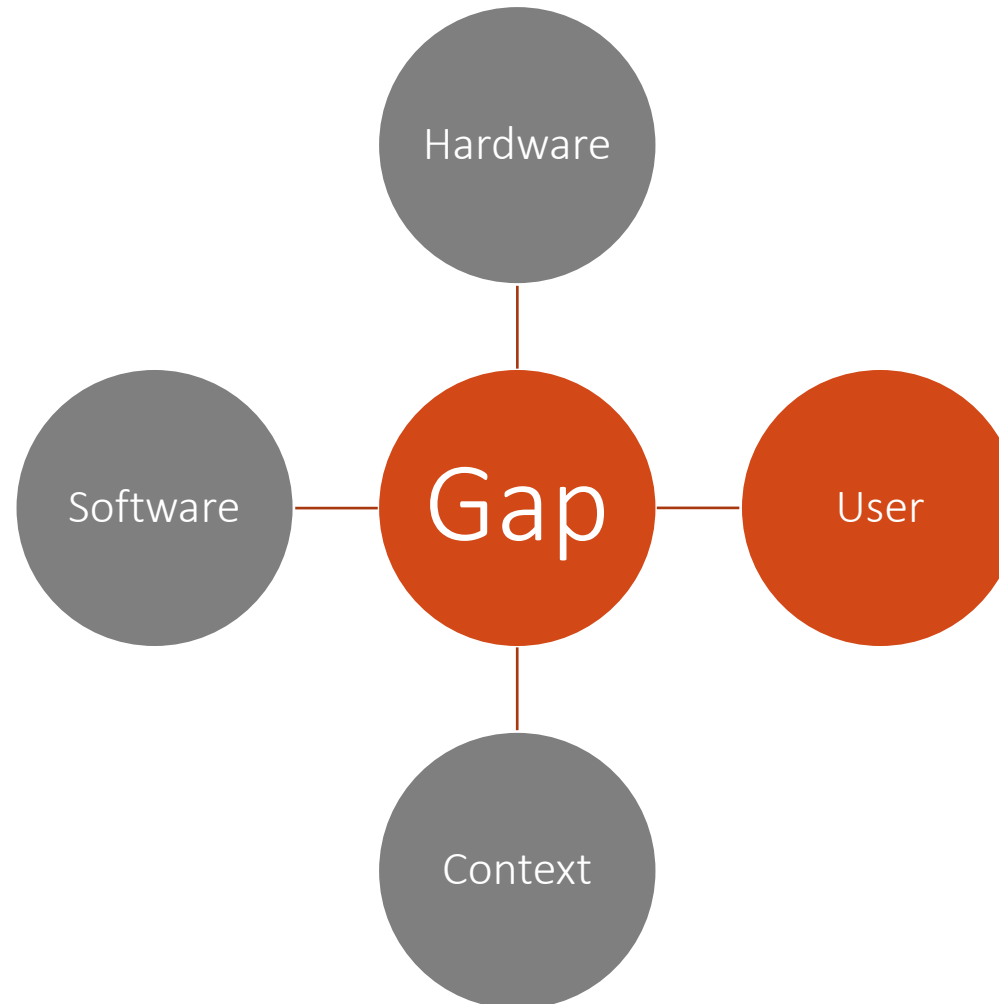
AME (with 95% CI) based on binomial probit regression with robust standard errors.

Low battery endangers data-collection

Older OS versions seem to be less prone to gaps

Device specific effects indicate hardware and software issues

Error sources



Participant characteristics

- Technical Competence

Participant behavior

- Fake data, kill / de-install battery-draining apps
- selectively turn off data collection

User-related error sources



codestring	timestamp	latitude	longitude	country
dfeh7r4v2v	05aug2018 10:28:48	52.2	8.6	Germany
dfeh7r4v2v	05aug2018 11:43:38	52.2	8.6	Germany
dfeh7r4v2v	05aug2018 12:22:50	8.6	52.2	
dfeh7r4v2v	05aug2018 12:52:49	8.6	52.2	

Apps falsify geolocation

Aim: Privacy, access location-specific content

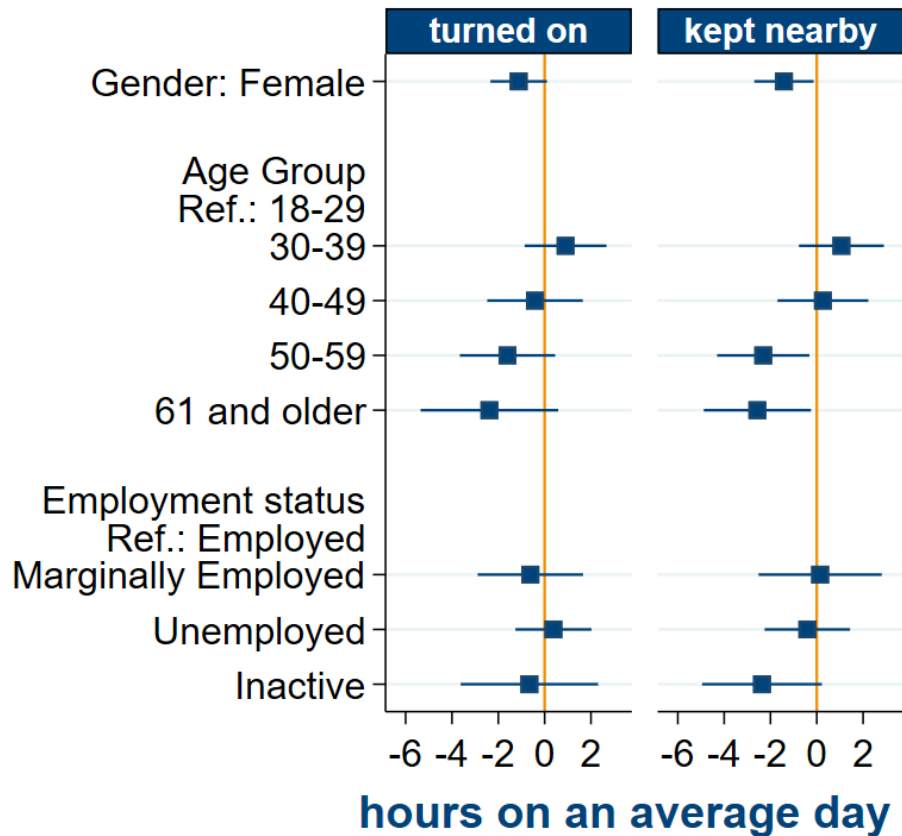
Validation with app usage data

4 / 621 participants had such apps installed

→ Replace false geo-positions with data from immediately before the app use

codestring	AppName	timestamp_start	timestamp_end
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:11:21	05aug2018 12:11:32
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:12:31	05aug2018 12:16:11
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:18:31	05aug2018 12:18:40
dfeh7r4v2v	Fake GPS with Joystick	05aug2018 12:19:00	05aug2018 12:19:03

Quality assessment from In-App surveys



- End of study survey includes rating questions

	Hours	Obs	Mean	Std. Dev.	Min	Max
turned on		462	20.9	5.8	1	24
kept nearby		462	11.3	6.2	0	24

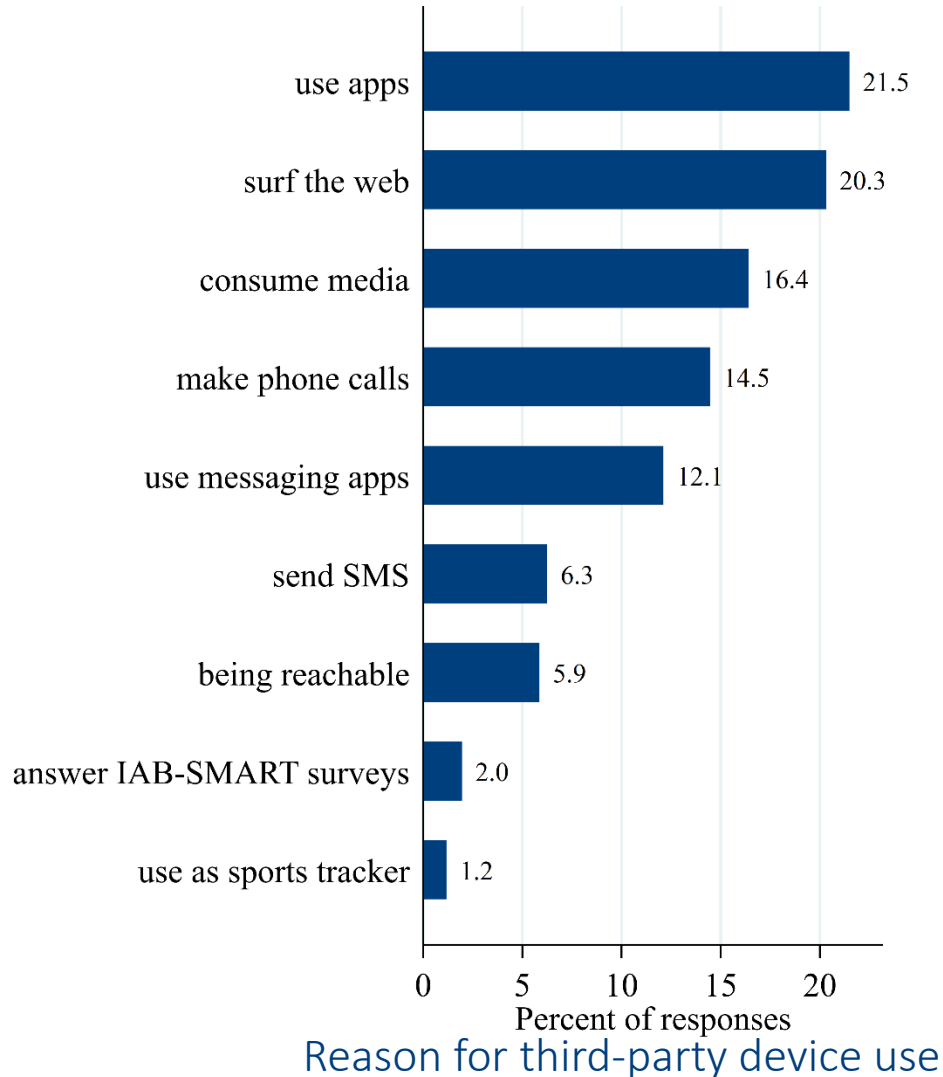
- Women tend to use their smartphone less than men
- Smartphone use drops at about 50 years of age
- There is no difference in use between employed and unemployed persons
- These characteristics and the usage information itself can be controlled in the models

389 participants, AMEs with 95% confidence intervals.

Turned on - On average, how many hours per day is your smartphone turned on?

Kept nearby - How many hours is the smartphone in your immediate vicinity (i.e. on your body, in the same building / car)?

Quality assessment from In-App surveys



- End of study survey includes questions about **third-party device use (3pdu)**

	Obs	Mean	Std. Dev.	Min	Max
Any 3pdu	465	0.16	0.4	0	1
Days with 3pdu	71	11.03	27.3	0	180
3pdu >10 days	471	0.03	0.2	0	1

- Reason for and extent of 3pdu determine scope of problem
- Depends on specific research questions



Example 2 – COVID-19 global survey

1. Scaling reach of surveys through public-private partnership
2. Daily monitoring and trend detection emphasized over full population coverage

Responding to the Need for Syndromic Surveillance

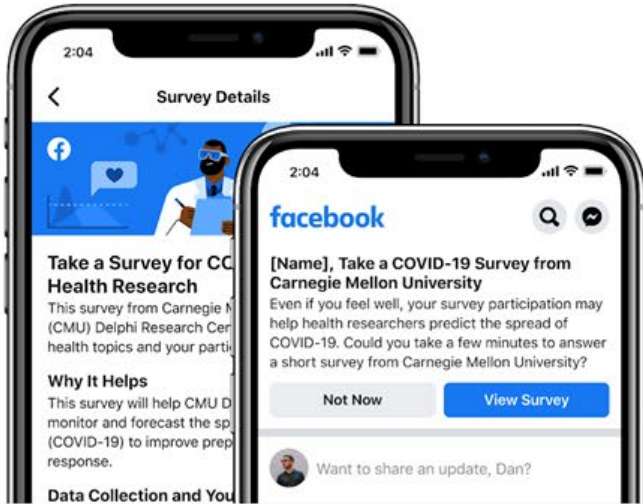
Syndromic surveillance enables policymakers and public health systems to make decisions before diagnosis data are available, especially in low resource areas with limited testing capabilities.

Facebook can reach large segments of the target population daily with the technical infrastructure to provide bias correction. And, the speed and scale of the symptom surveys allow them to act as early warning systems.

For more information see: covidmap.umd.edu



1 Who's Taking the Survey



Facebook invites a new, random sample of users to participate each day.

By now > 100 million responses since April 2020

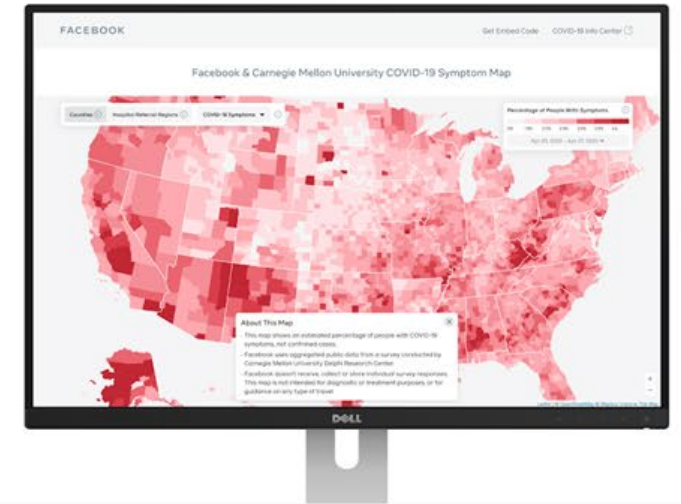
2 How the Survey Works



Users are sent to the survey hosted by UMD or CMU using Qualtrics.

Facebook does not receive responses, but does calculate weights to correct for non-response bias and sampling frame coverage bias using internal Facebook data for 115 countries or territories.

3 Using the Survey Data



Using the aggregated data, Facebook created a map visualization to help policymakers and public health systems make decisions.

The non-aggregate data are available to eligible academic and nonprofit researchers by request.

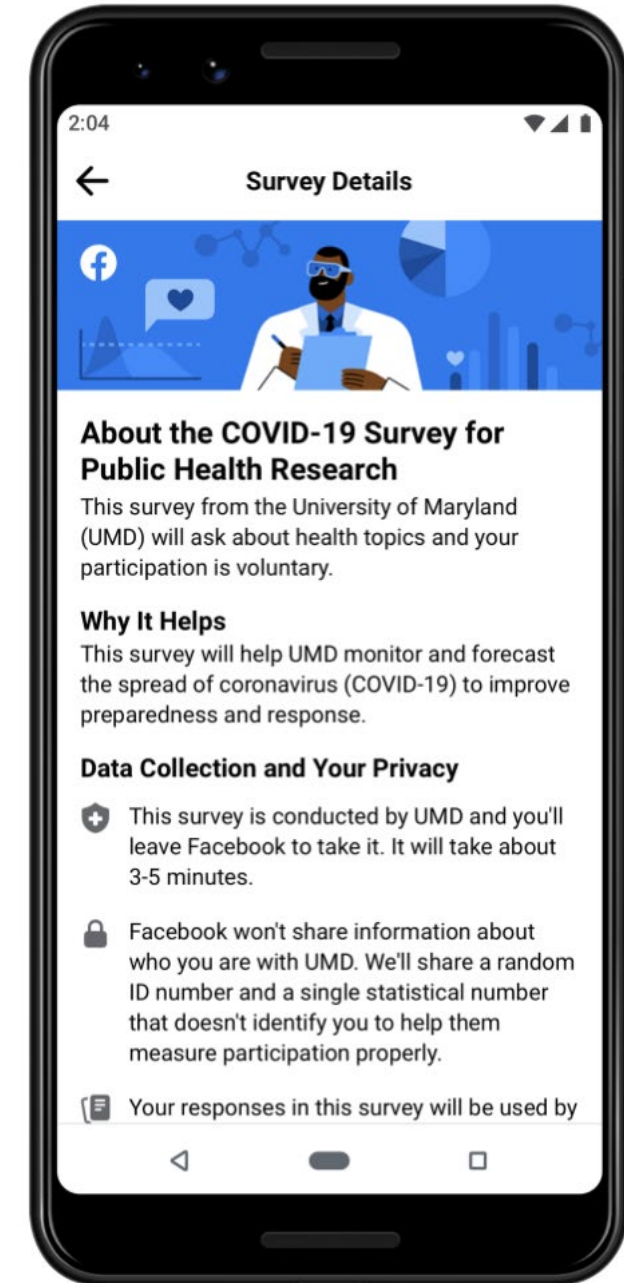
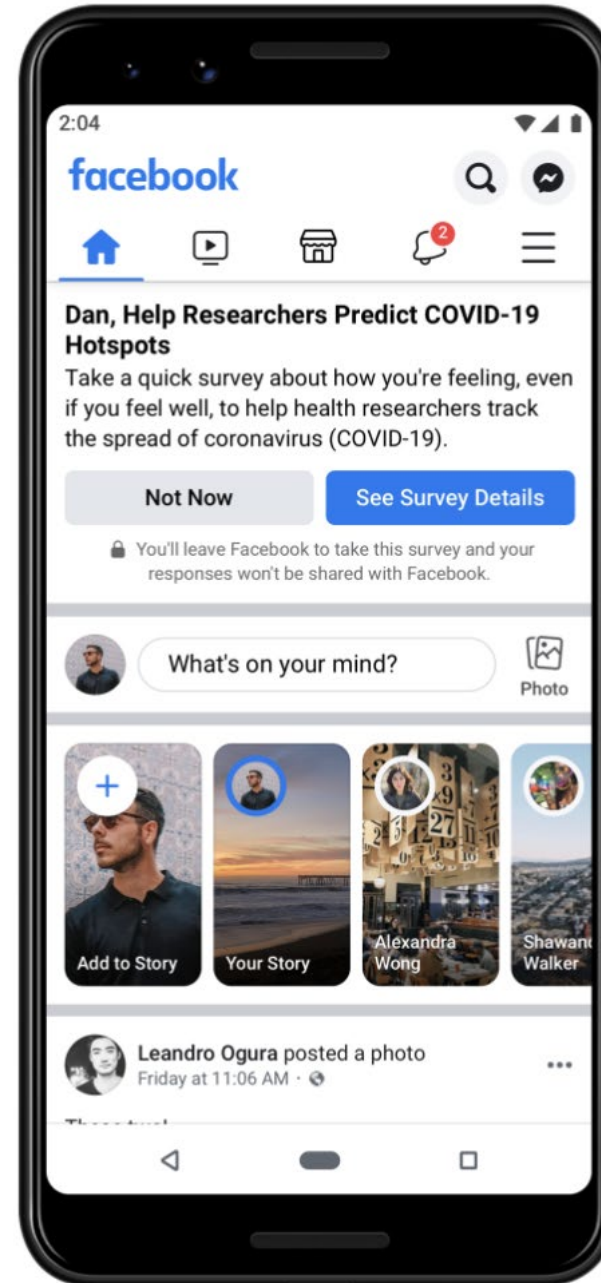
UMD Global Survey Instrument

Available in 50+ languages

Survey Instrument has 5 Sections:

- Consent
- Health symptoms
- Contacts with others
- Mental health and economic security
- Demographic characteristics

For more information see: covidmap.umd.edu



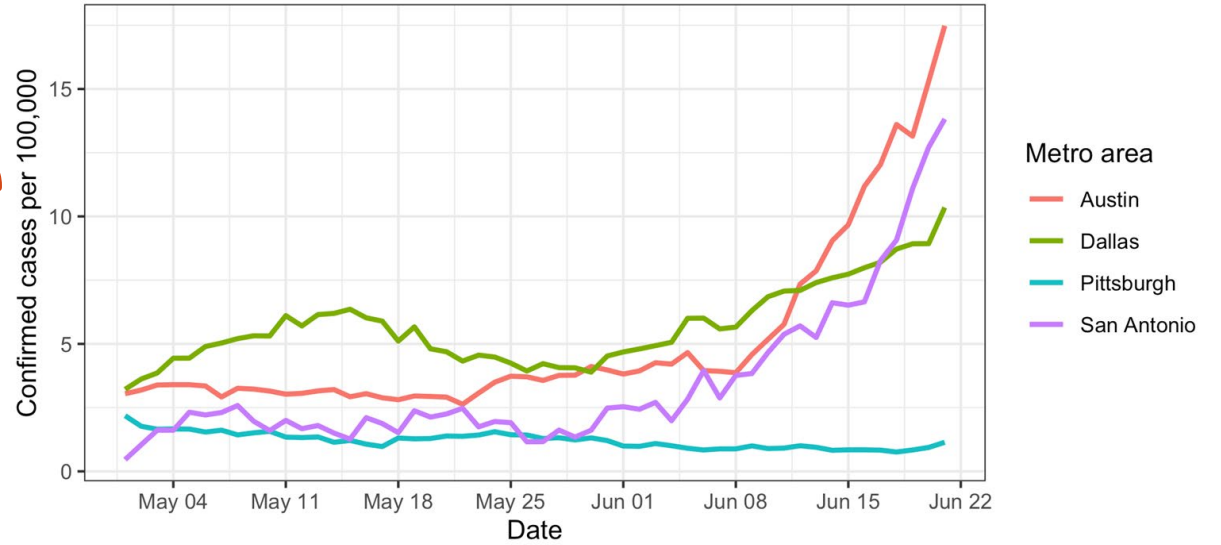
Early Insights for Fore

CMU Delphi Research Center is developing short term hospitalization forecasts in the US and deepening its partnerships with public health agencies.

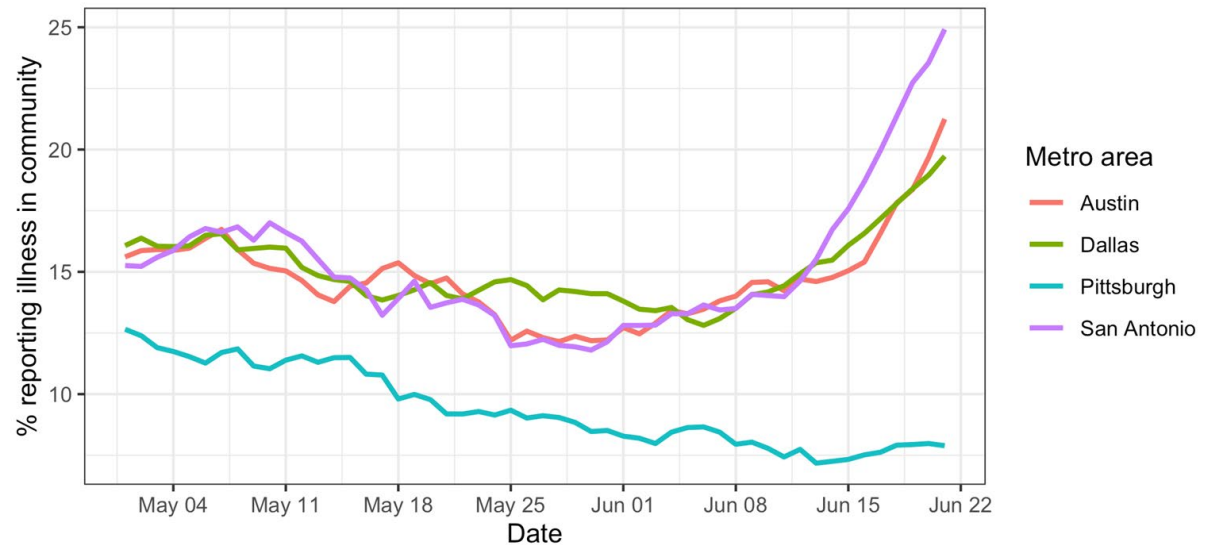
The symptom survey also shows noticeable correlation with confirmed case numbers, though the correlation varies across geographies.

For more information see: covidmap.umd.edu

Confirmed cases



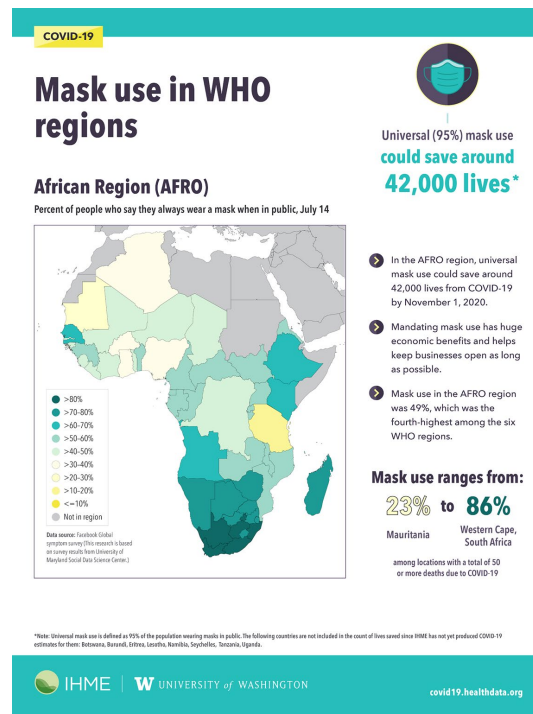
Symptom survey



Early Research Insights

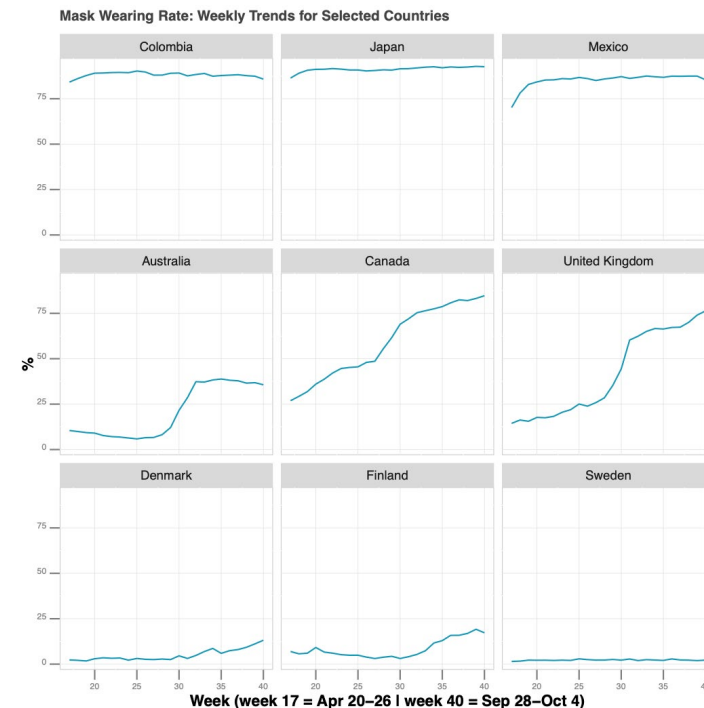
15 institutions are working with the non-aggregate data from at least one of the surveys.

IHME is mapping the prevalence of regular mask wearing, using the global Symptom Survey in conjunction with data from Premise.



SoDa has produced an interactive dashboard of mask-wearing behavior.

From April 2020 to present, we asked, “In the last 7 days, how often did you wear a mask when in public?”



For more information see: covidmap.umd.edu

How to Access Symptom Survey Data

Publicly Available, Aggregate Data

Global Survey Data:

<https://covidmap.umd.edu/api.html>

US Survey Data:

<https://cmu-delphi.github.io/delphi-epidata/api/covidcast.html>

Non-Aggregate Data for Research

Researchers from academic and non-profit institutions can request access.

Signed Data Use Agreements are required.

Central portal for project documentation and data access requests is on Facebook's Data for Good website: dataforgood.fb.com.

Additional Resources

Other Complimentary Data Sources Through Data for Good

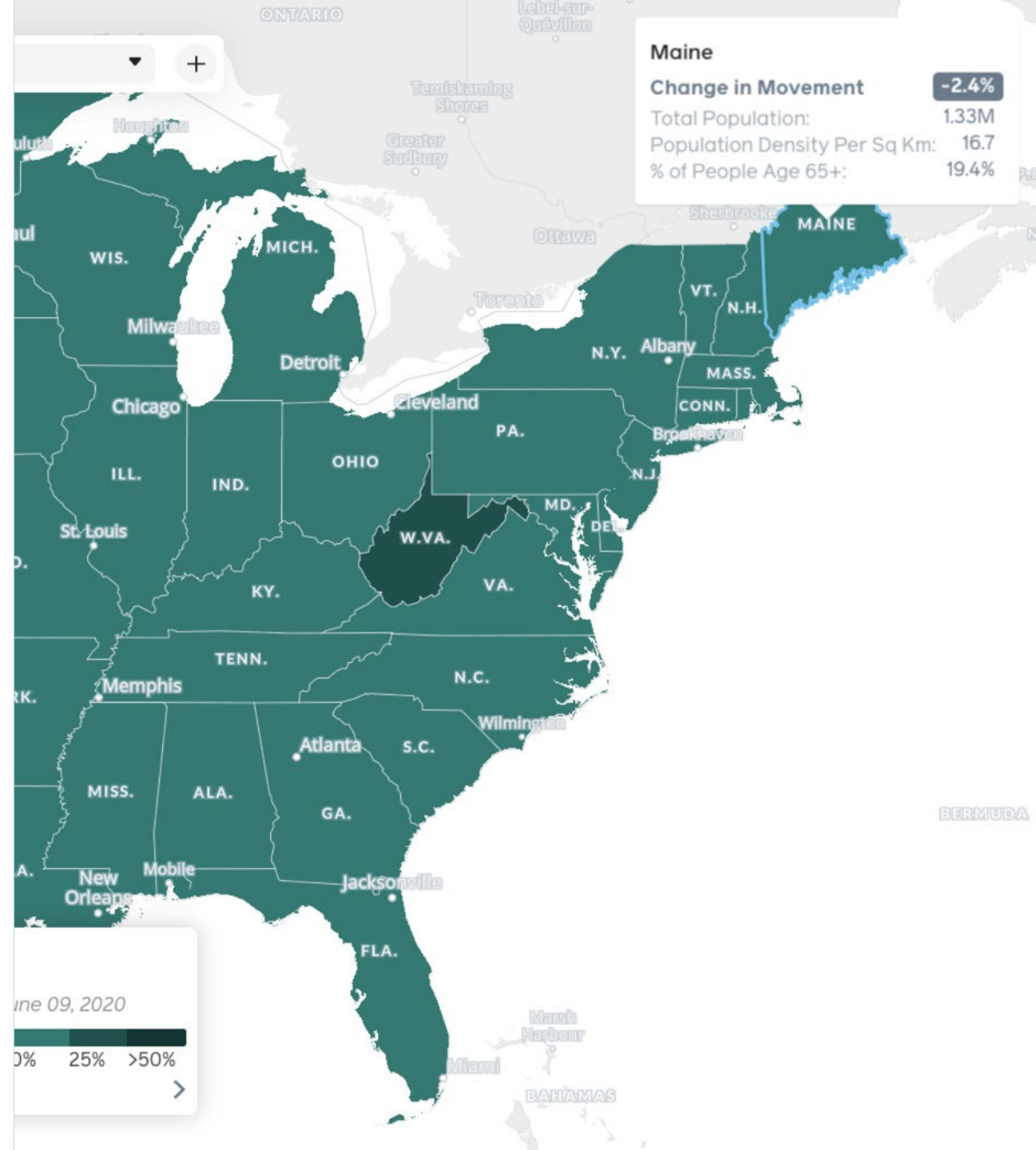
Population Density Maps

Social Connectedness Index

Movement Range Maps

More information on Facebook's Data for Good website: dataforgood.fb.com.

COVID-19 Symptom Data Challenge: symptomchallenge.org/.





Privacy

1. We can quickly face **higher privacy risks**
2. Researchers need to value **appropriate flow**
3. **Infrastructure** needed to support privacy efforts

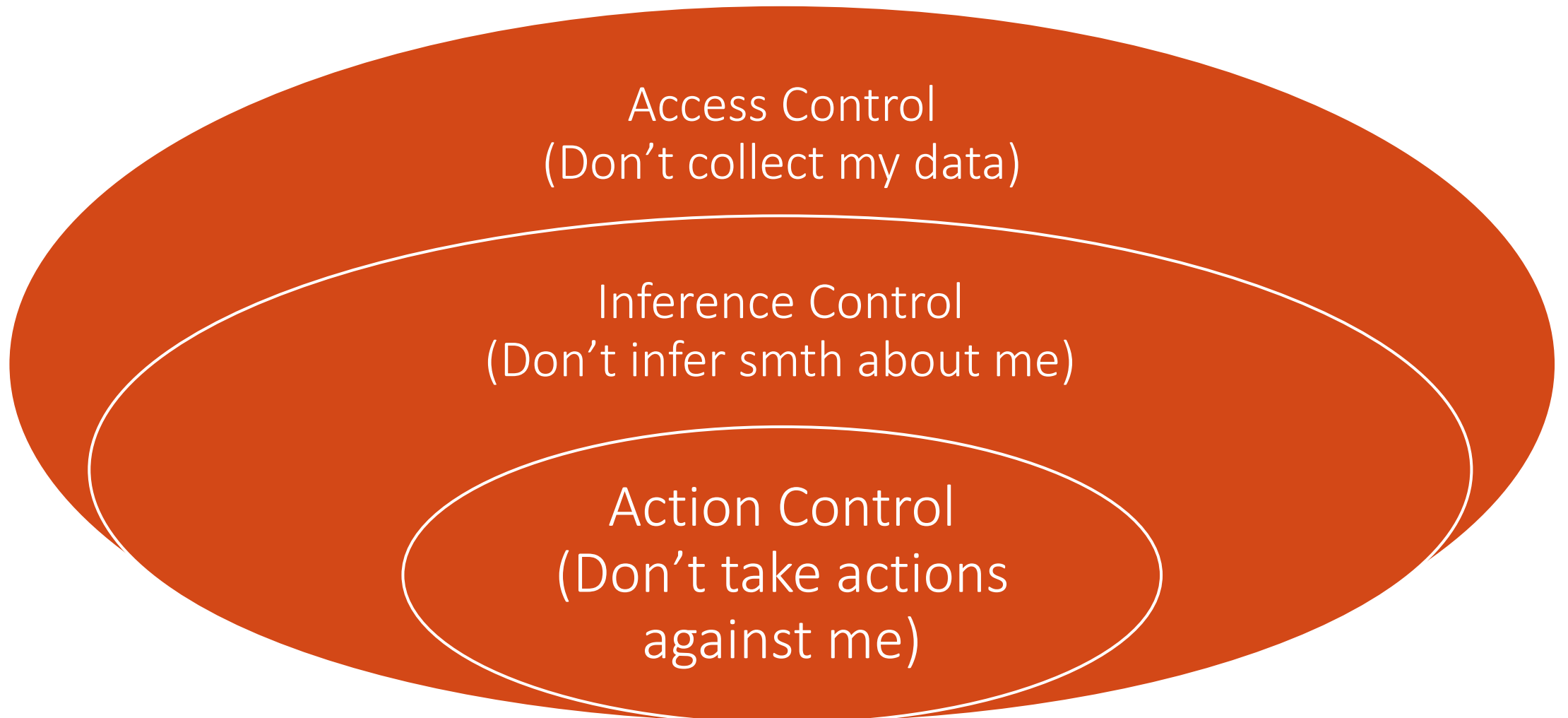
Microdata Releases

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.

Consent to give up control



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

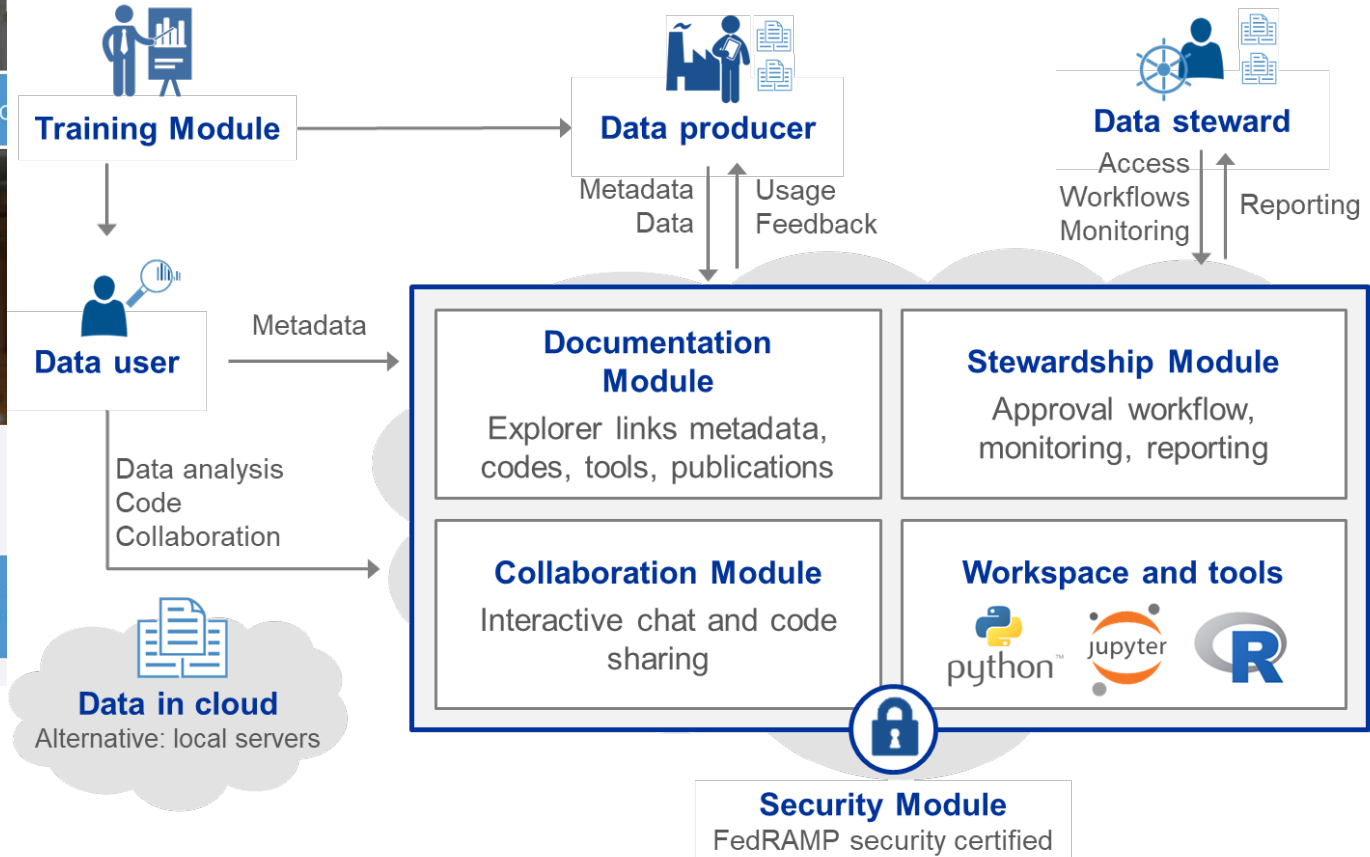
Web	Front	Back	Total
% agree	82.6	62.4	520

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?

Democratizing our Data: A Challenge to Invest in Data and Evidence-based Policy

Learn More

WHO WE



Who makes it happen?

The screenshot shows the top navigation bar of the IPSDS website with links for Home, Program, Project, Contact, FAQs, Imprint, Search, and an APPLY NOW button. Below the navigation is a large banner image of three people in conversation. The text on the banner reads: "INTERNATIONAL PROGRAM IN SURVEY AND DATA SCIENCE" and "offered through the University of Mannheim and the Joint Program in Survey Methodology (Universities of Maryland and Michigan West)". A "BE PART OF IT" button is visible at the bottom of the banner.

We are pleased to announce the launch of the International Program in Survey and Data Science. The program focuses on the nature of data, their availability, the way in which they are collected, integrated, and disseminated. It involves working with designed data from surveys as well as organic data. IPSDS was developed in response to the needs of researchers and practitioners for the appropriate methods and right tools to face these changes. The program offers a curriculum, world-class faculty, and a web-based learning environment that allows you to take courses



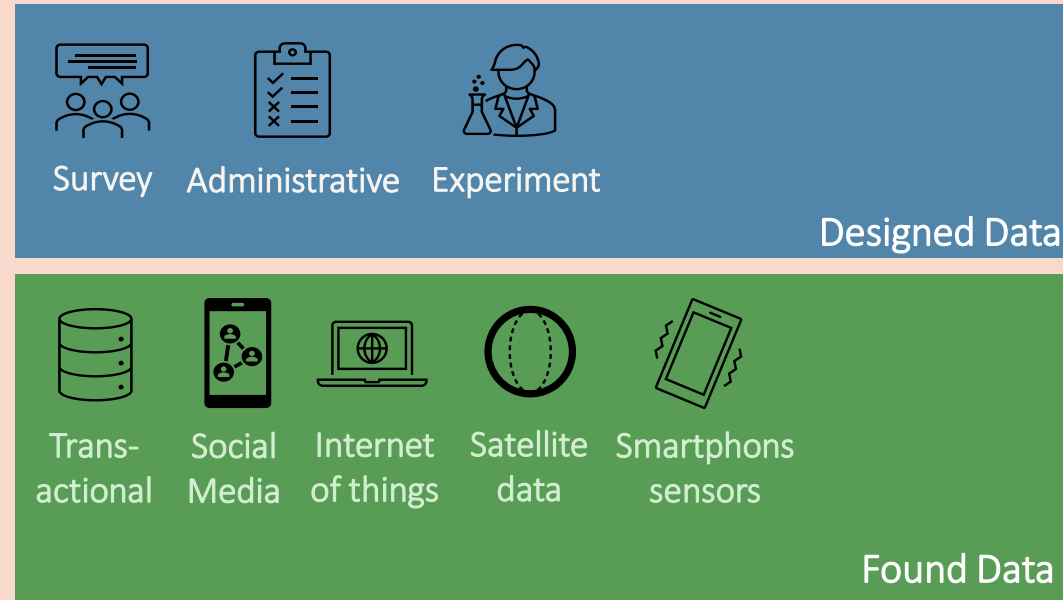
Federal Ministry of Education and Research

The screenshot shows the ASA DataFest website. The header features the ASA DataFest logo. Below the header are three blue boxes with white text: "What Is ASA DataFest?", "Hosting an Official ASA DataFest", and "Supporting ASA". To the right of these boxes is a section titled "What Is ASA DataFest?" with text describing the event as a gathering for 48 undergraduate students to work on a data set from the Los Angeles Police Department. Below the text are several small images showing students working at computers and in a lecture hall.



Mastering the New Data World

- The recent years have shown, that official statistics will not only rely on **survey data but also on new data sources** (so-called Big Data).
- **Each data source has its benefits and pitfalls** and could be used in different situations as well as in combination.
- These developments also emphasized the **need for well-trained professionals** to competently work with those new data sources and their combination.
- This holds especially true for **governmental agencies**, which compete with the private sector to hire data analysts. **Continuing education** in applied data science can help **filling the skill gap** to benefit from new and linked data sources.



We **designed a study program for working professionals** who want to enhance their data collection and analysis skills for **both designed and found data** while pursuing their careers.

International Partnership

- Funded by the German **BMBF** the **University of Mannheim** built together with the **University of Maryland** the **International Program in Survey and Data Science**.
- **Students** at JPSM, which offers a full- and part-time program come primarily from **federal statistical agencies** in the Washington, DC area and its related contractors.
- **Alumni** from JPSM and its sister-program in Michigan work at large **tech companies** such as Facebook, Google, Survey Monkey, and Qualtrics.
- The close ties to UMD through this program offer opportunities for other **US-based activities** of the MBS.



Designed for Working Professionals

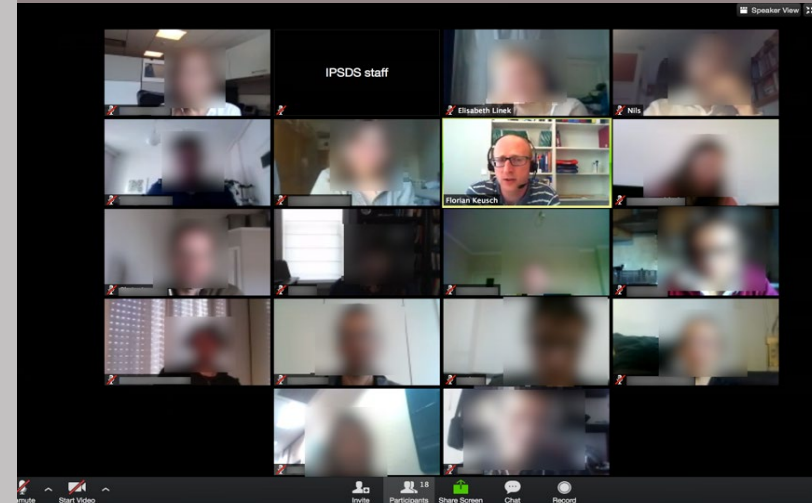
Asynchronous

- Pre-recorded video lectures
- Required readings and assignments
- Discussion forums

The image shows a collage of asynchronous learning materials. On the left, there is a video player with a paused video of a man speaking. Above it, a browser window displays a document titled 'Introduction - How to do survey research and data science'. To the right, a slide from the University of Mannheim features Prof. Dr. Thomas Fetzter, LL.M. (Vanderbilt) and the title 'Privacy Law'. The slide includes the Mannheim Centre for Competition and Innovation logo and a photograph of a large, historic building.

Synchronous

- Small virtual classrooms
- Weekly one-hour discussions led by the instructor



- 4 products: (1) Master; (2) Certificates; (3) Open Courses; (4) Tailored Courses

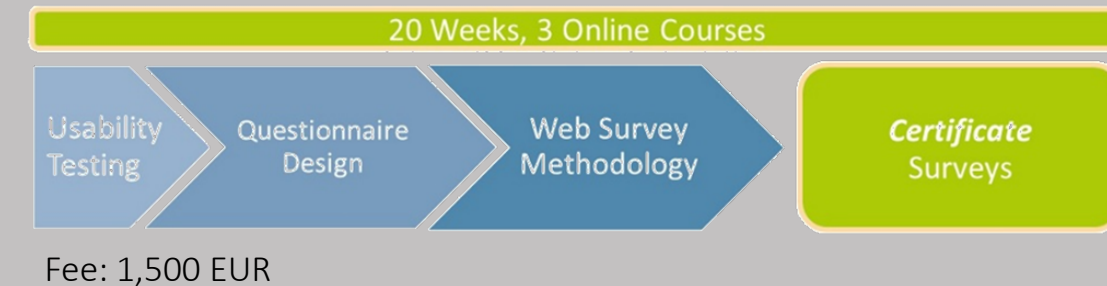
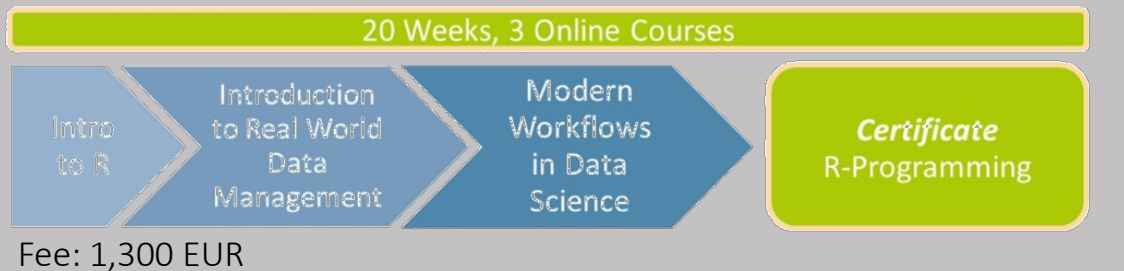
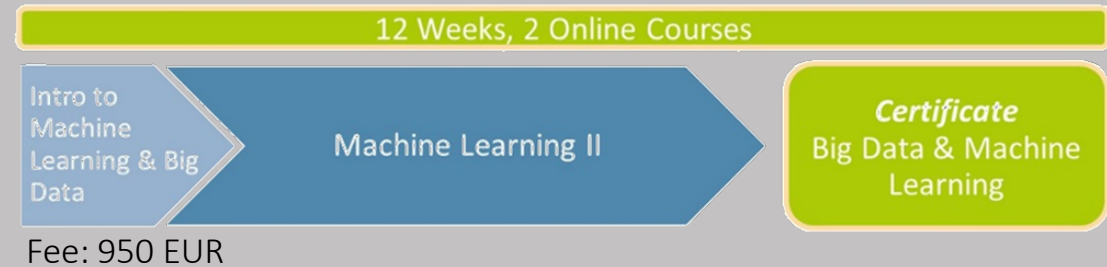
Open Courses – Summer 2022

Course	Instructor(s)	Dates	Fee
Introduction to Record Linkage with Big Data Application	Manfred Antoni (IAB) Stefan Bender (Deutsche Bundesbank) Christian Borgs (Uni Duisburg-Essen) Joe Sakshaug (IAB)	June 1 – July 20, 2022	800.00 EUR
Ethical Considerations for Data Science Research	Jessica Vitak (University of Maryland)	July 18 – August 8, 2022	400.00 EUR
Introduction to Python and SQL	Diego Fregolent Mendes des Oliveira (Rensselaer Polytechnic Institute)	June 9 – 30, 2022	400.00 EUR
Item Nonresponse and Imputation	Jörg Drechsler (IAB)	June 20 – July 11, 2022	400.00 EUR
Web Survey Methodology	Jernej Berzelak (National Institute of Public Health, Slovenia)	June 7 – July 26, 2022	800.00 EUR
Web Scraping and APIs	Sascha Göbel (Goethe University Frankfurt)	July 15 – August 5, 2022	400.00 EUR
Review of Statistical Concepts	Anna-Carolina Haensch (Ludwig-Maximilians-University of Munich)	June 1 – August 17, 2022	1,200.00 EUR
Privacy Law	Thomas Fetzer (University of Mannheim)	July 7 – 28, 2022	400.00 EUR

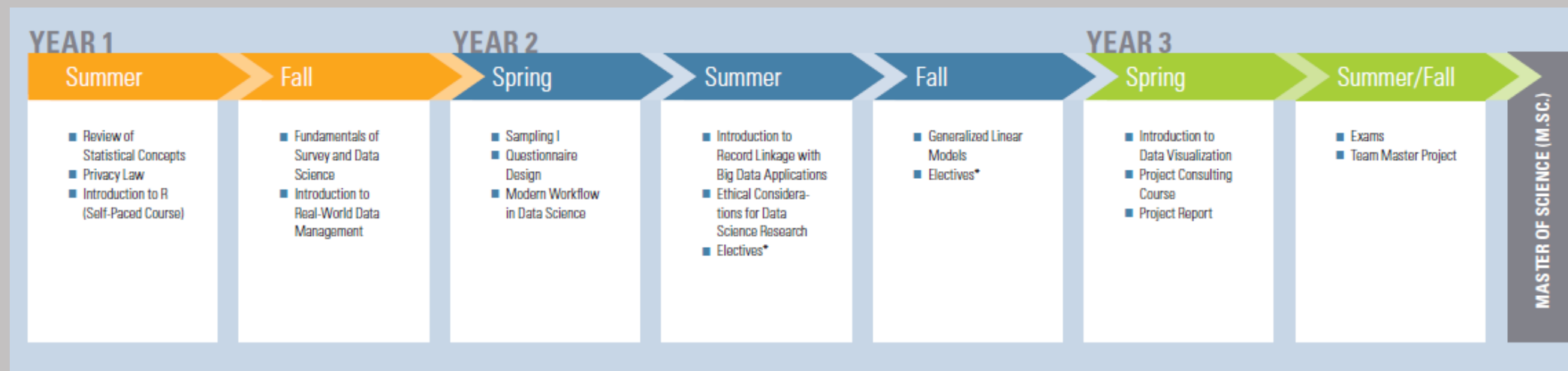
Full list of current and upcoming courses: <https://survey-data-science.net/program/courses/all>

February 7, 2022

Six Certificates



Master of Applied Data Science and Measurement



- The 30 months, 90 ECTS credits MDM curriculum is **oriented on the data-life-cycle** and covers the skills needed for working successfully with data.
- A **wide range of Electives** for example in **Big Data & Machine Learning, Python/SQL, APIs and Advanced Topics in Official Statistics** allow the students to tailor the study contents to their needs.
- The courses are strongly **practice-oriented** to transfer the acquired competences directly to the student's place of work.
- In the Project Consulting Course, students work in groups on a **data project from real clients**, for example from Bundesbank.
- The **Master-Project** is conducted in **group work** to mirror real-world data projects, which are usually completed in teams.
- Full price: EUR 22,500. Official Statistics Scholarship price: EUR 9,900.

More information: <https://www.mannheim-business-school.com/en/mba-master/mannheim-master-of-applied-data-science-and-measurement/>

February 7, 2022



Summary

1. Great potential: **New questions** can be asked
2. **Inference issues** and **data quality** questions do not go away
3. **Privacy** needs to be considered at the design stage
4. It is important to **empower** oneself and those around us



fkreuter@umd.edu

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

<http://socialdatascience.umd.edu/>