Combined Data Products COVID-19 context

Frauke Kreuter

JPSM - Uni Maryland
LMU - IAB

@fraukolos

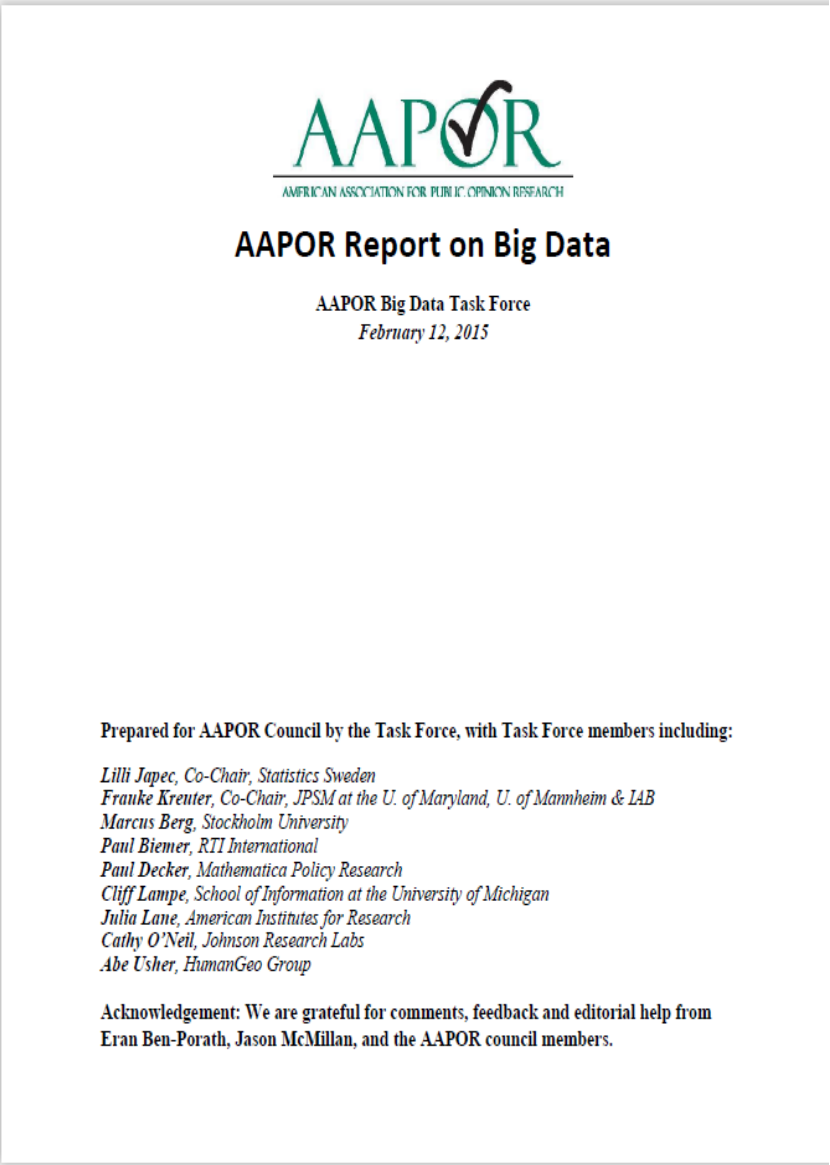
octocber 2020

Photo by ELVIS lawrence on Unsplash

Survey Statistics Perspective

Research Examples

Privacy



AAPOR
AMERICAN ASSOCIATION FOR PUBLIC OPINION RESEARCH

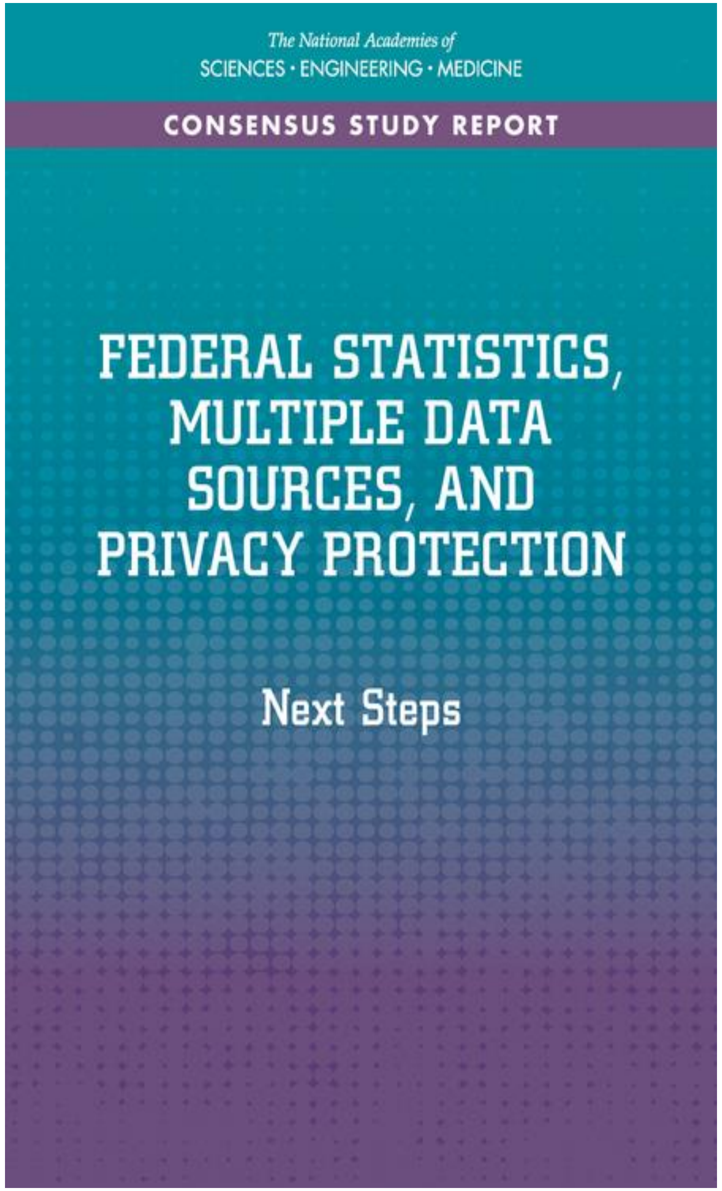
AAPOR Report on Big Data

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
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.

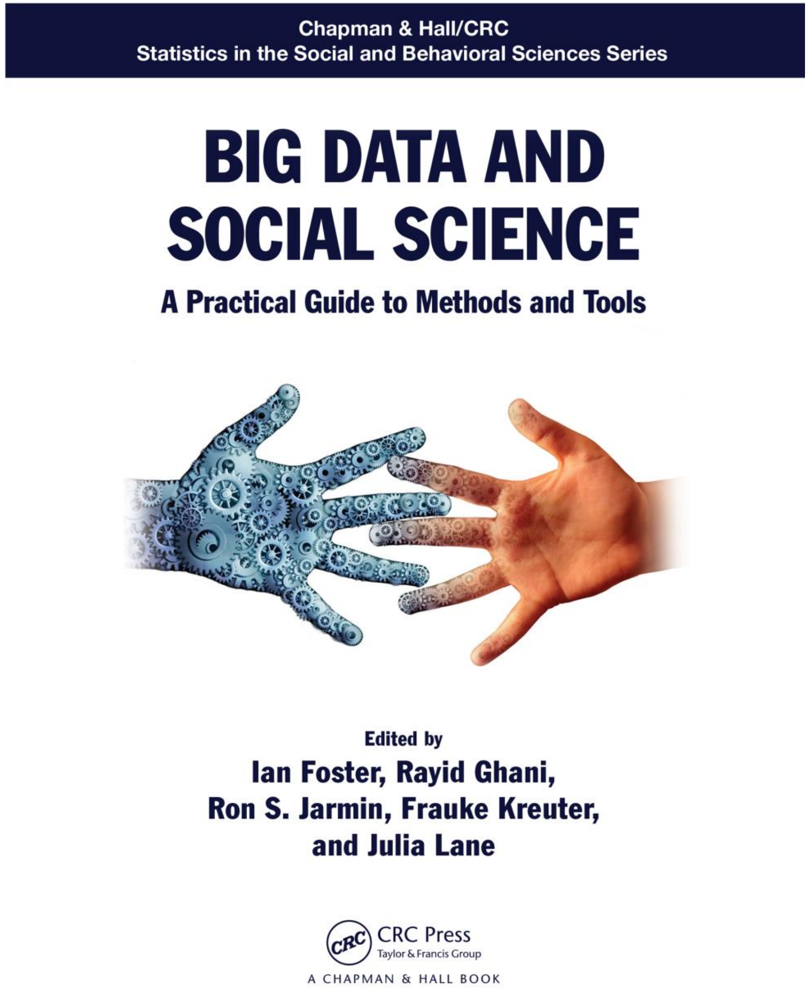


The National Academies of
SCIENCES · ENGINEERING · MEDICINE

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

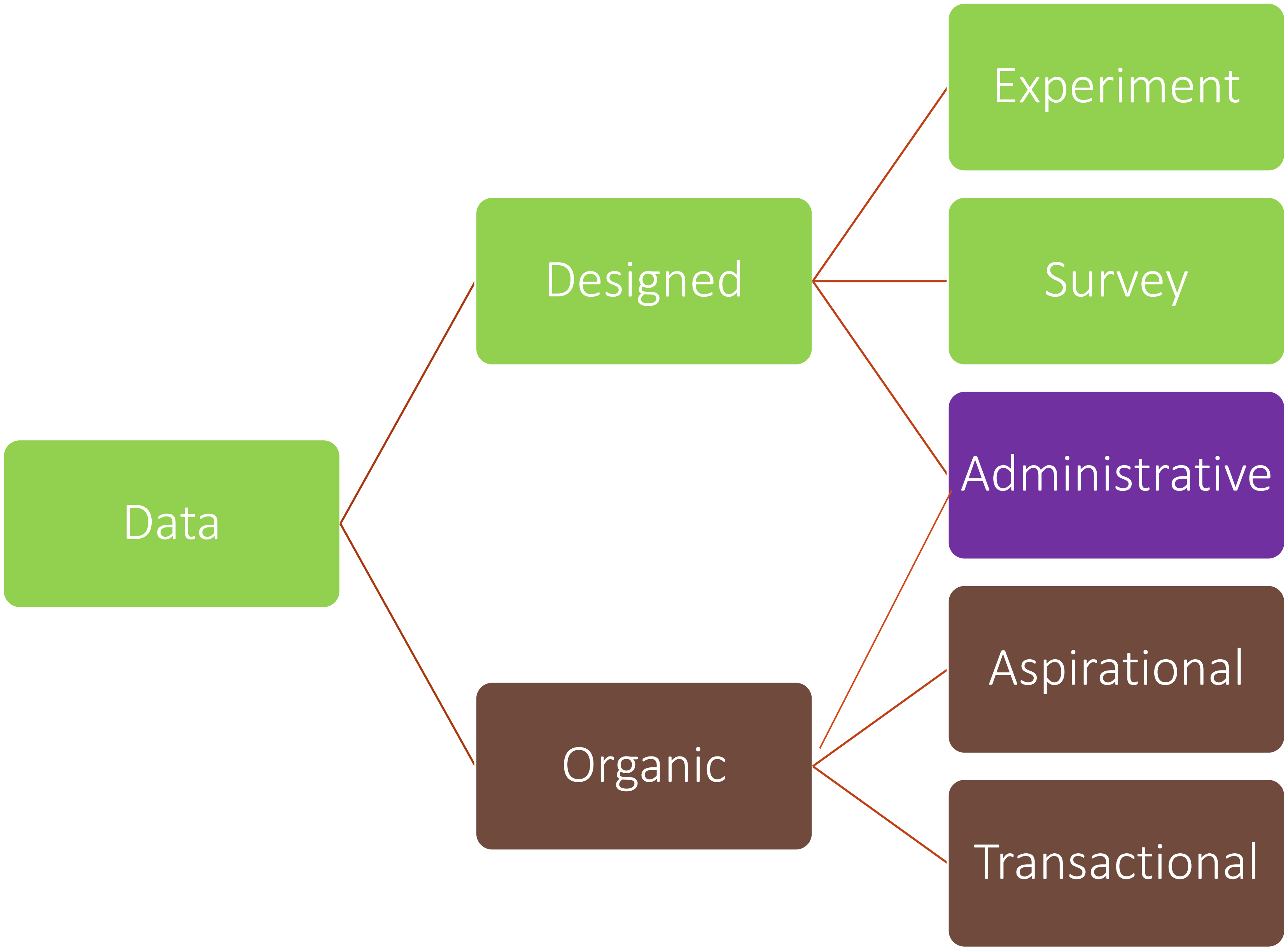
Edited by
Ian Foster, Rayid Ghani,
Ron S. Jarmin, Frauke Kreuter,
and Julia Lane

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

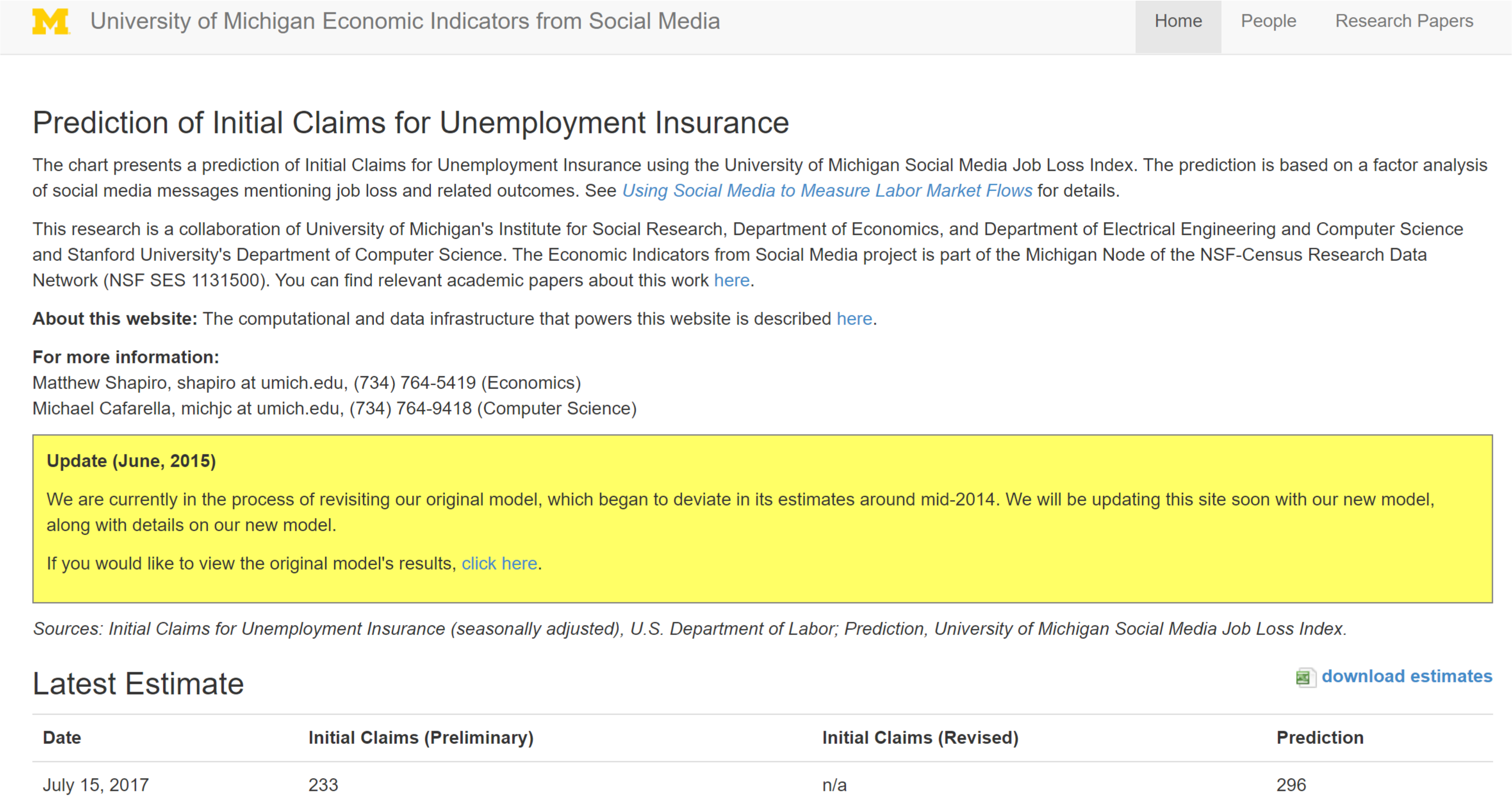
https://textbook.coleridgeinitiative.org/

Survey-Statistician Perspective

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



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


University of Michigan Economic Indicators from Social Media
Home
People
Research Papers

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

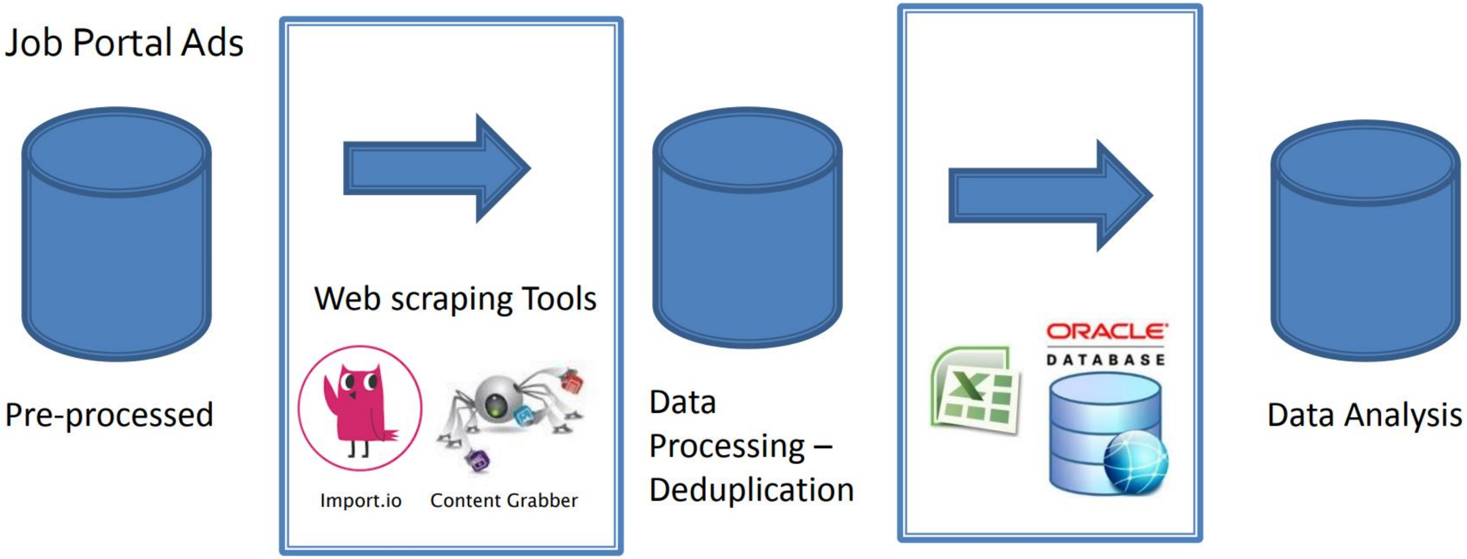
download estimates

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

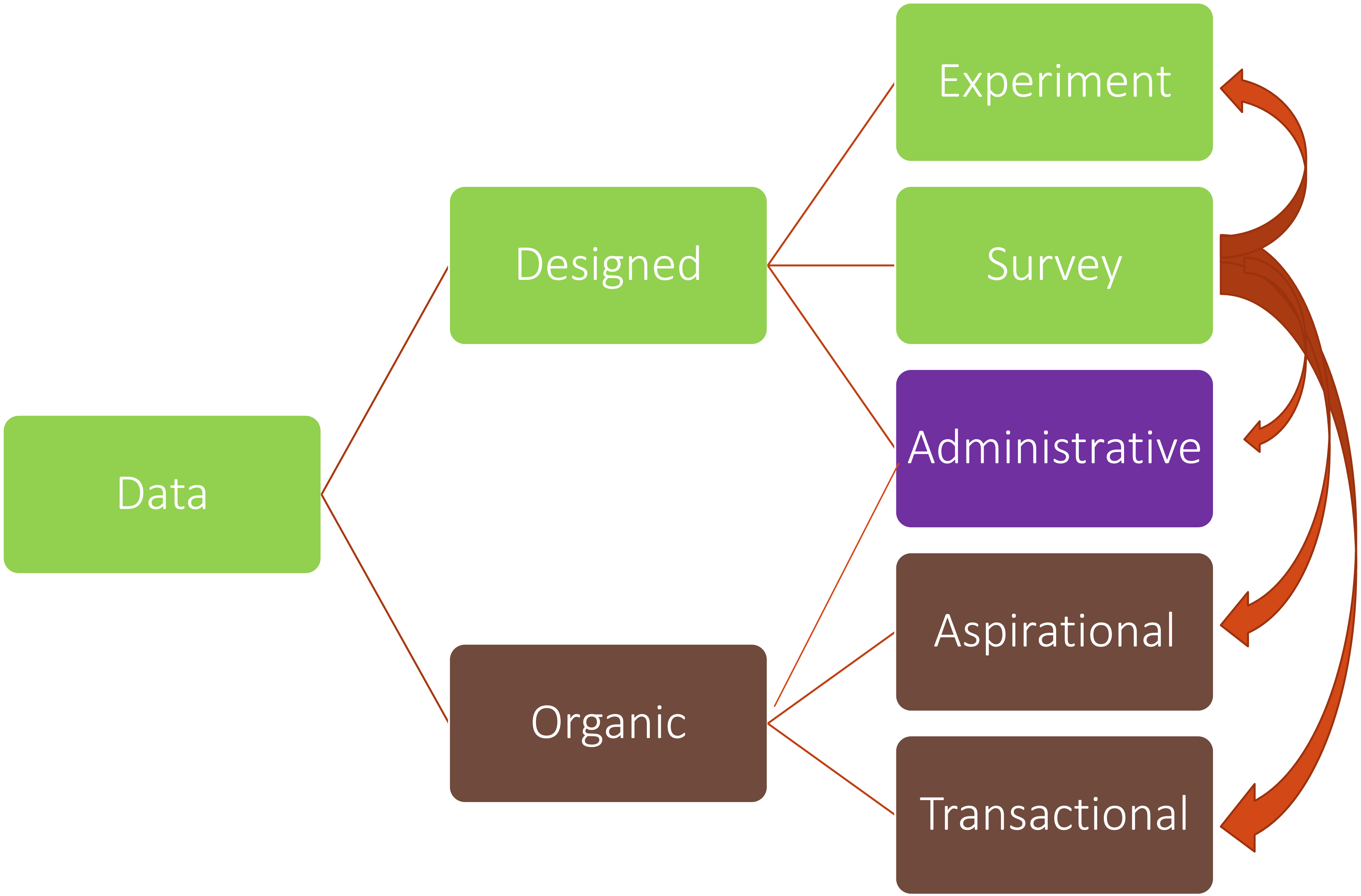
Web scraping Tools

Import.io

Content Grabber

Data Processing – Deduplication

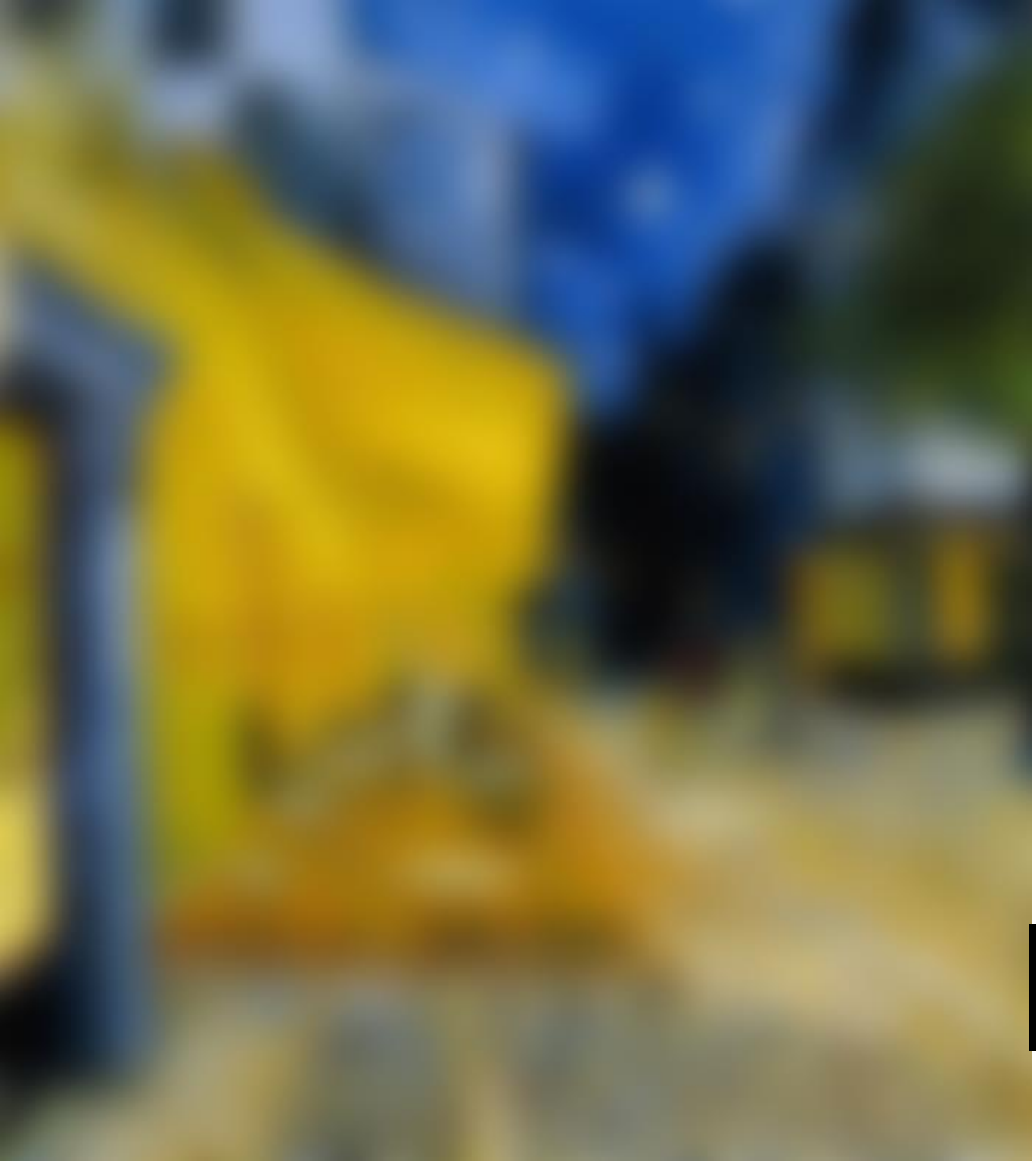
Data Analysis



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

Big Data

VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights

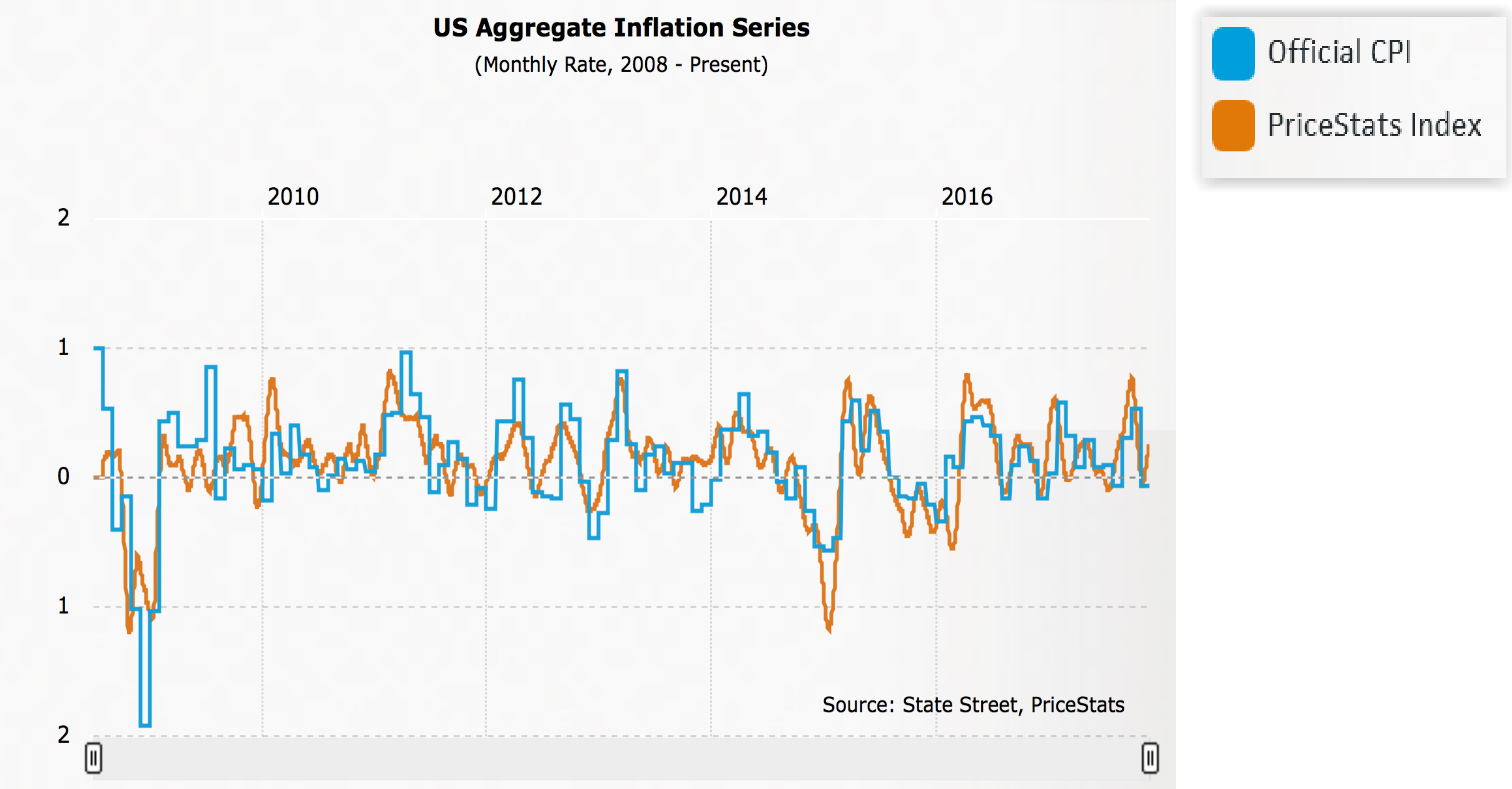
Surveys

VINCENT VAN GOGH

Credit: Ralph Klüber, p3 Insights

Designed Product

> One way to think about a data ~~analysis~~ collection 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



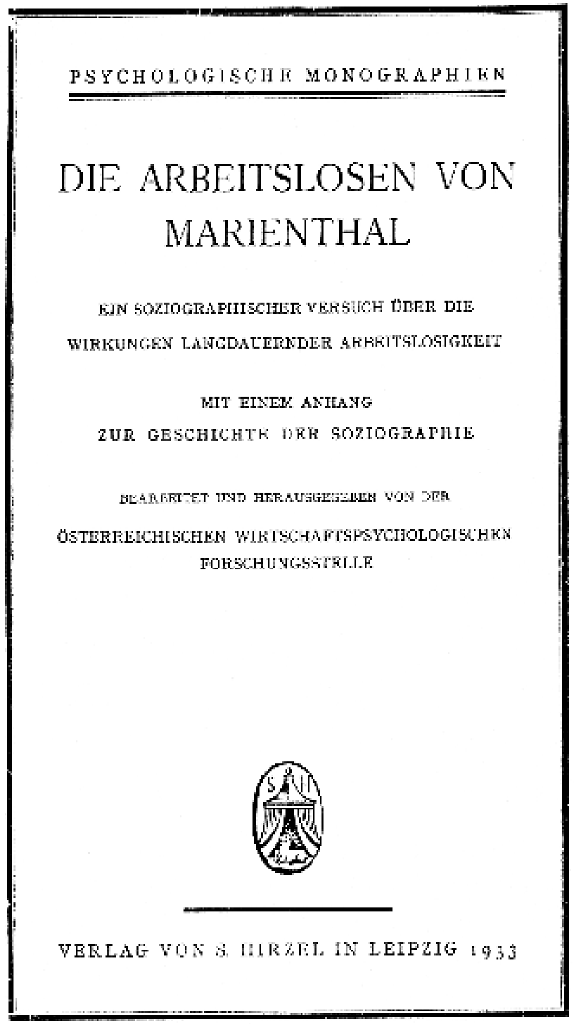
Source: State Street, PriceStats

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

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?



PSYCHOLOGISCHE MONOGRAPHIEN

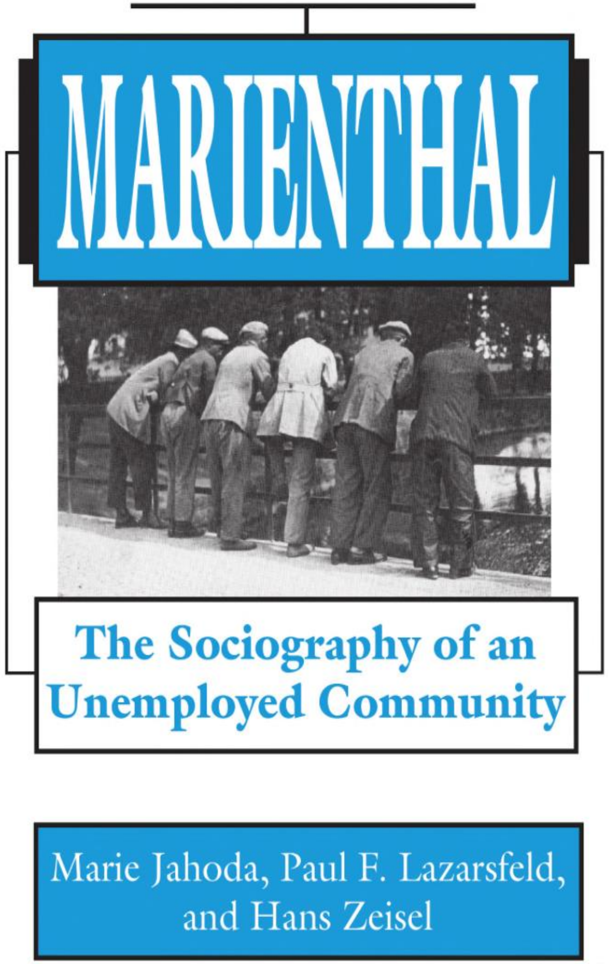
DIE ARBEITSLOSEN VON MARIENTHAL

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

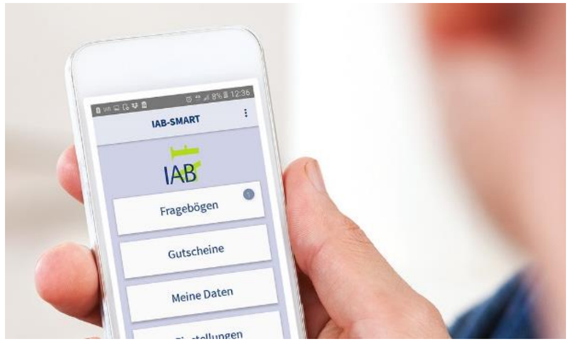


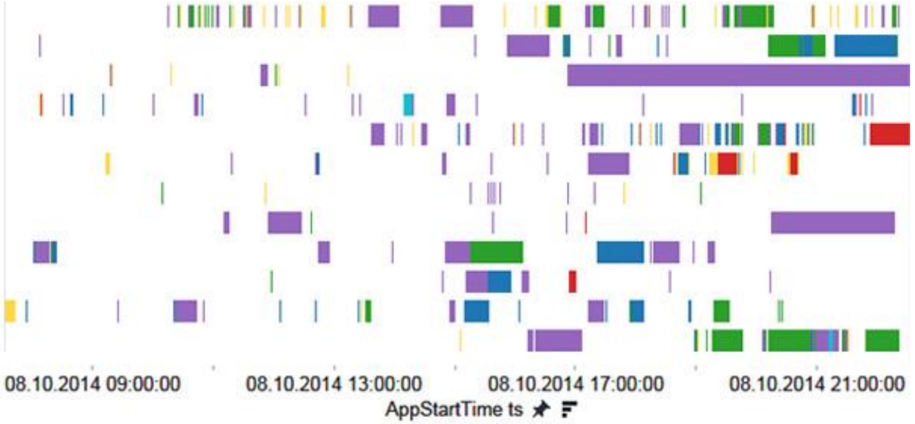
MARIENTHAL

The Sociography of an Unemployed Community

Marie Jahoda, Paul F. Lazarsfeld, and Hans Zeisel

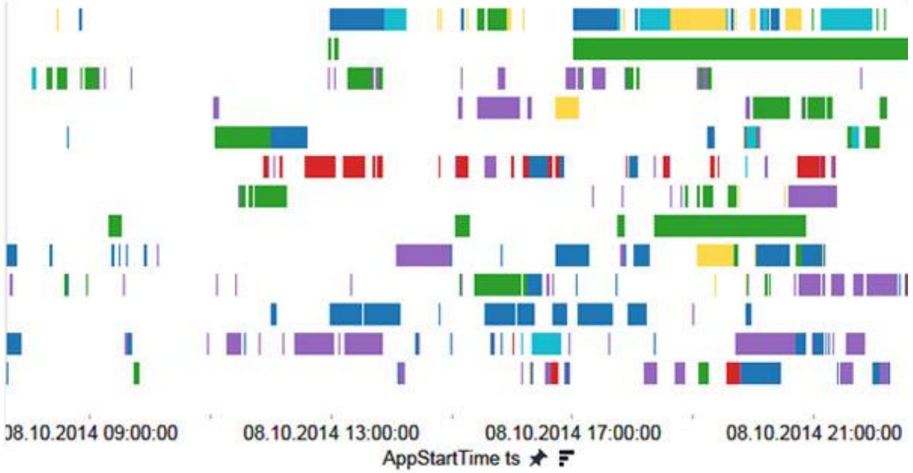
Source: Archives for the History of Sociology in Austria (Graz), »Marienthal« Virtual Archives





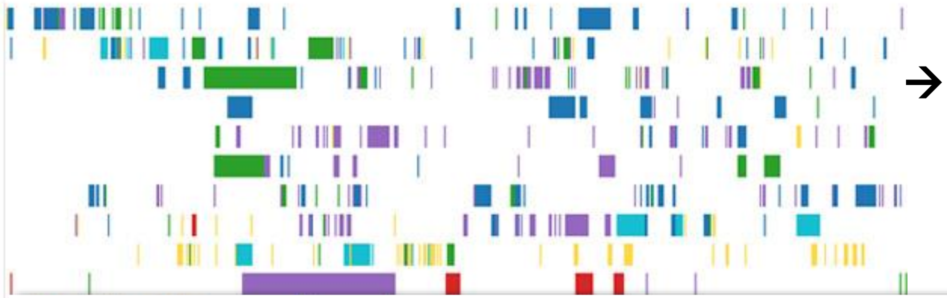
Full-time employed

→ **App use past 5pm**



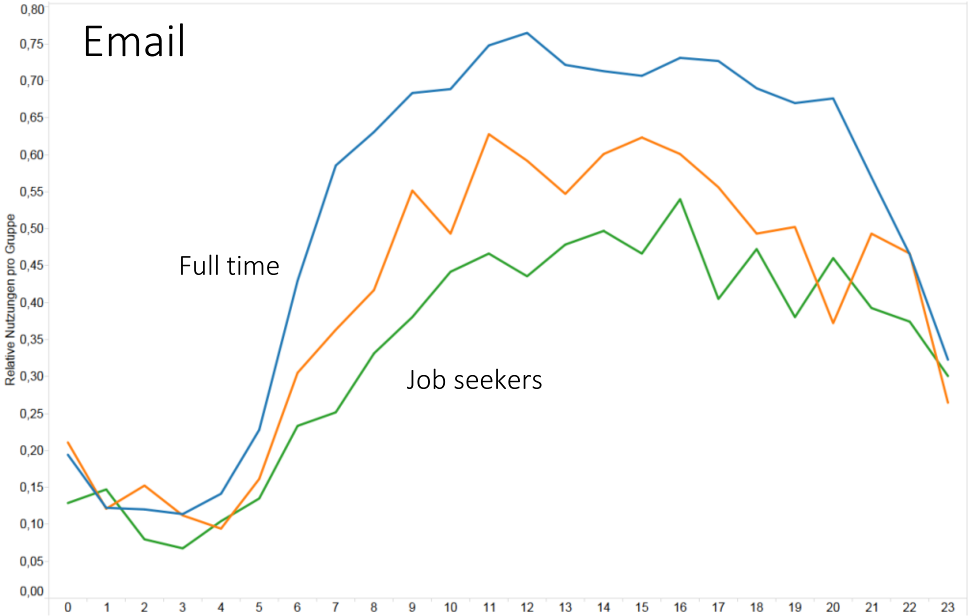
Part-time employed

→ **App use at noon**

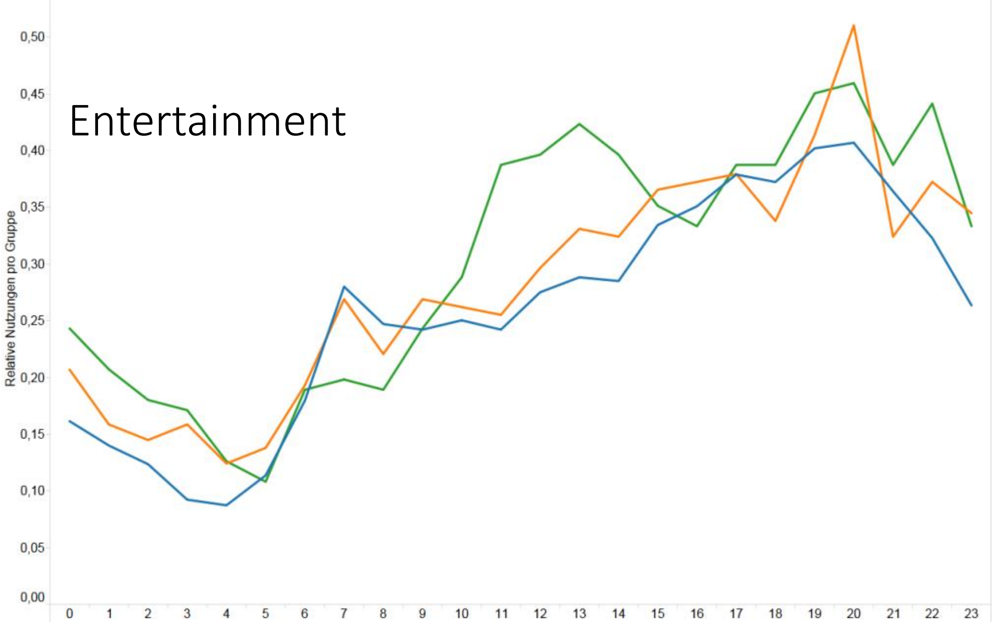


Job seekers

→ **Continuous app use**



Email



Entertainment

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/

PASS – Panel (10 years) + Administrative Data

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

Refreshed annually

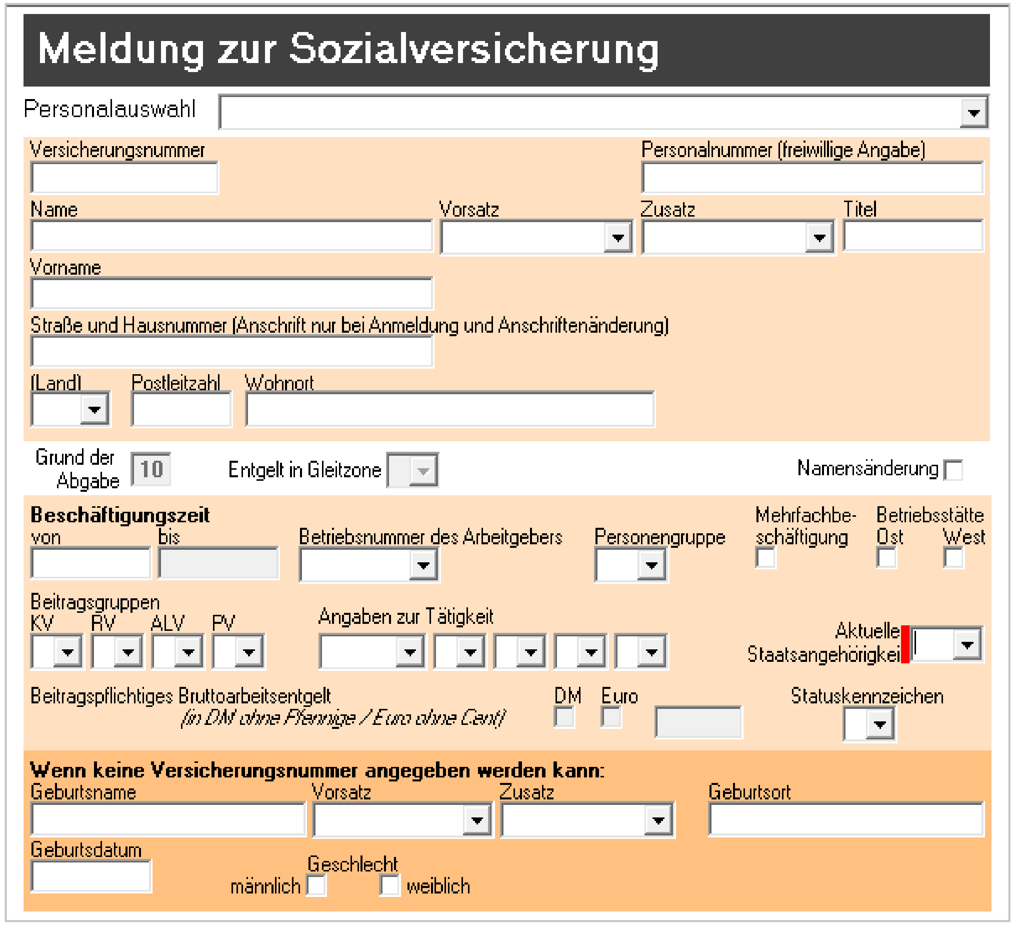
Surveyed annually

Random household sample of resident population

Refreshed annually

Surveyed annually

+



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 10

Entgelt in Gleitzone

Namensänderung □

Beschäftigungszeit

von

bis

Betriebsnummer des Arbeitgebers

Personengruppe

Mehrfachbeschäftigung □

Betriebsstätte Ost □ West □

Beitragsgruppen KV RV ALV PV

Angaben zur Tätigkeit

Aktuelle Staatsangehörigkei

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

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

Inference to Population

...owning a (specific) smartphoneCoverage error

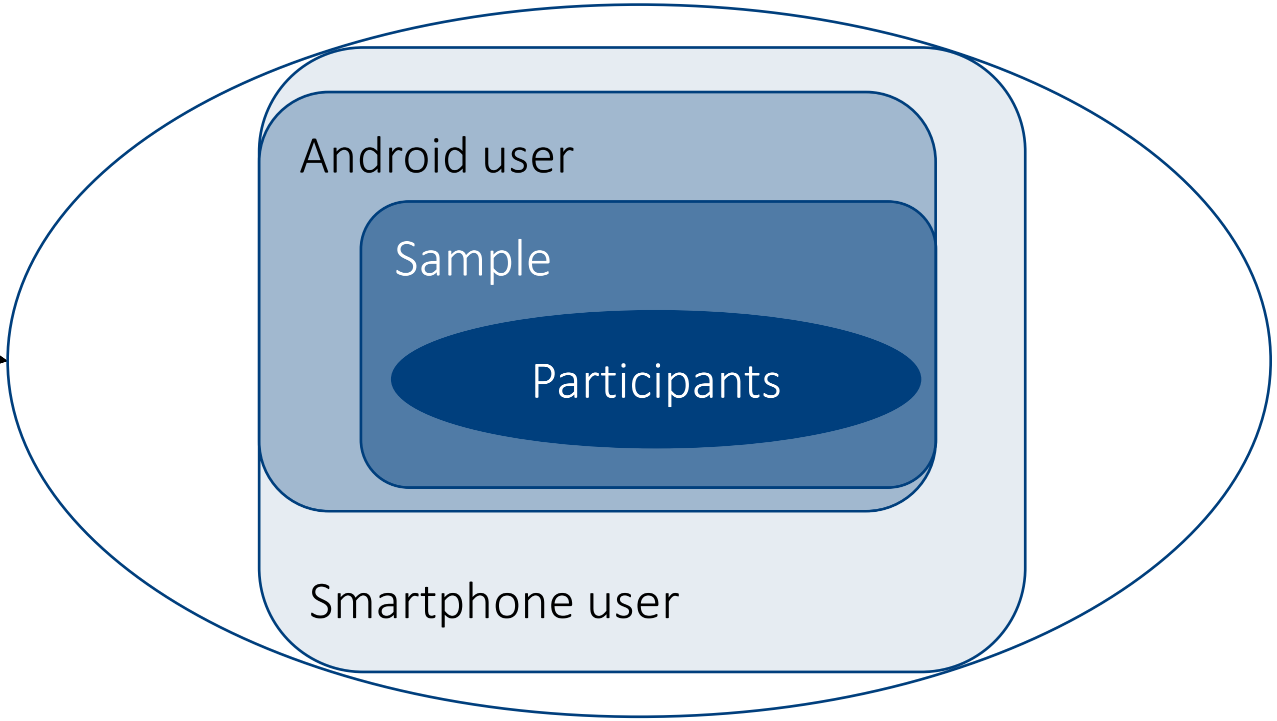
...being able to download an app

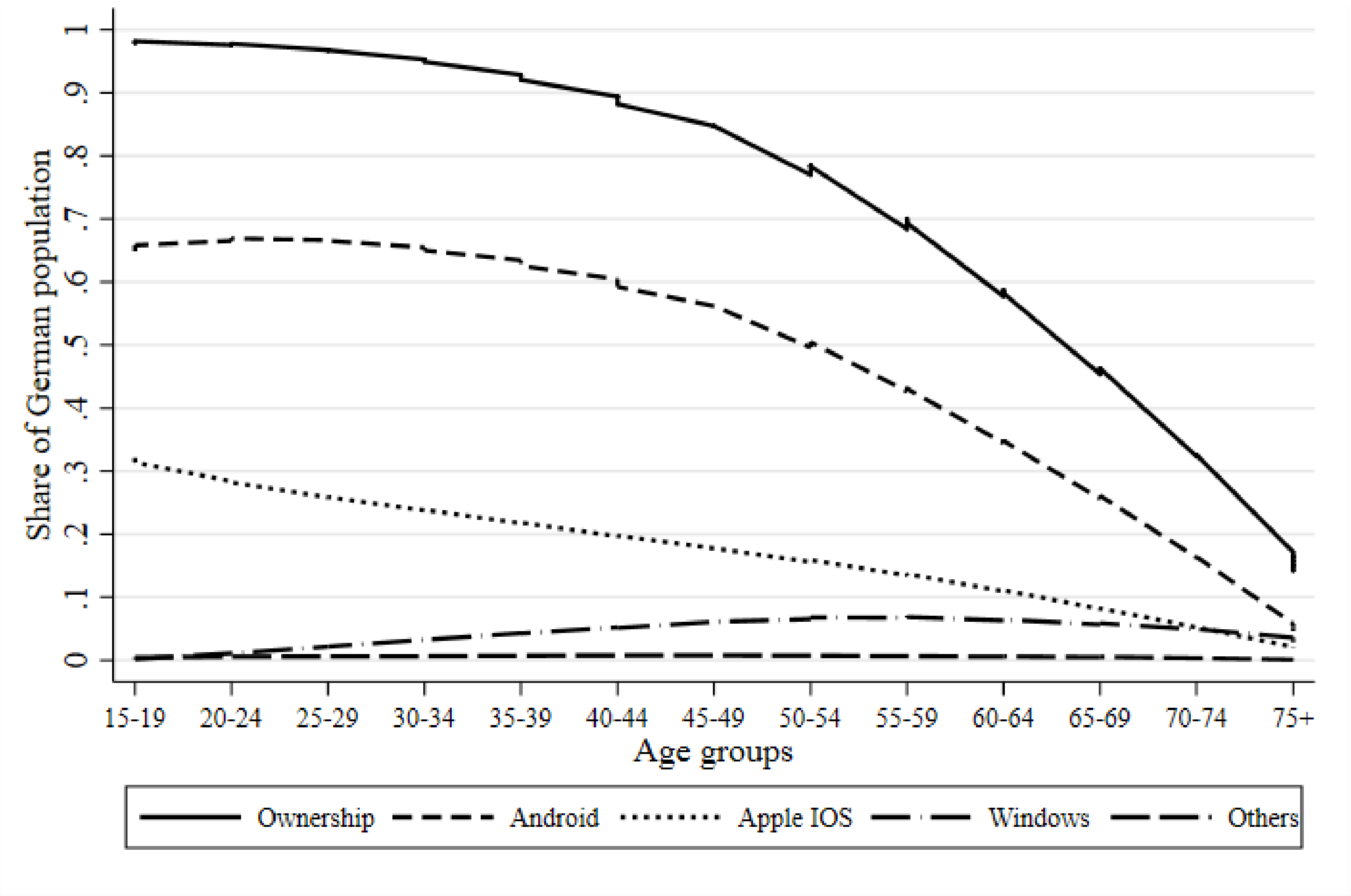
...being willing to download an app

} Nonparticipation error

Population

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



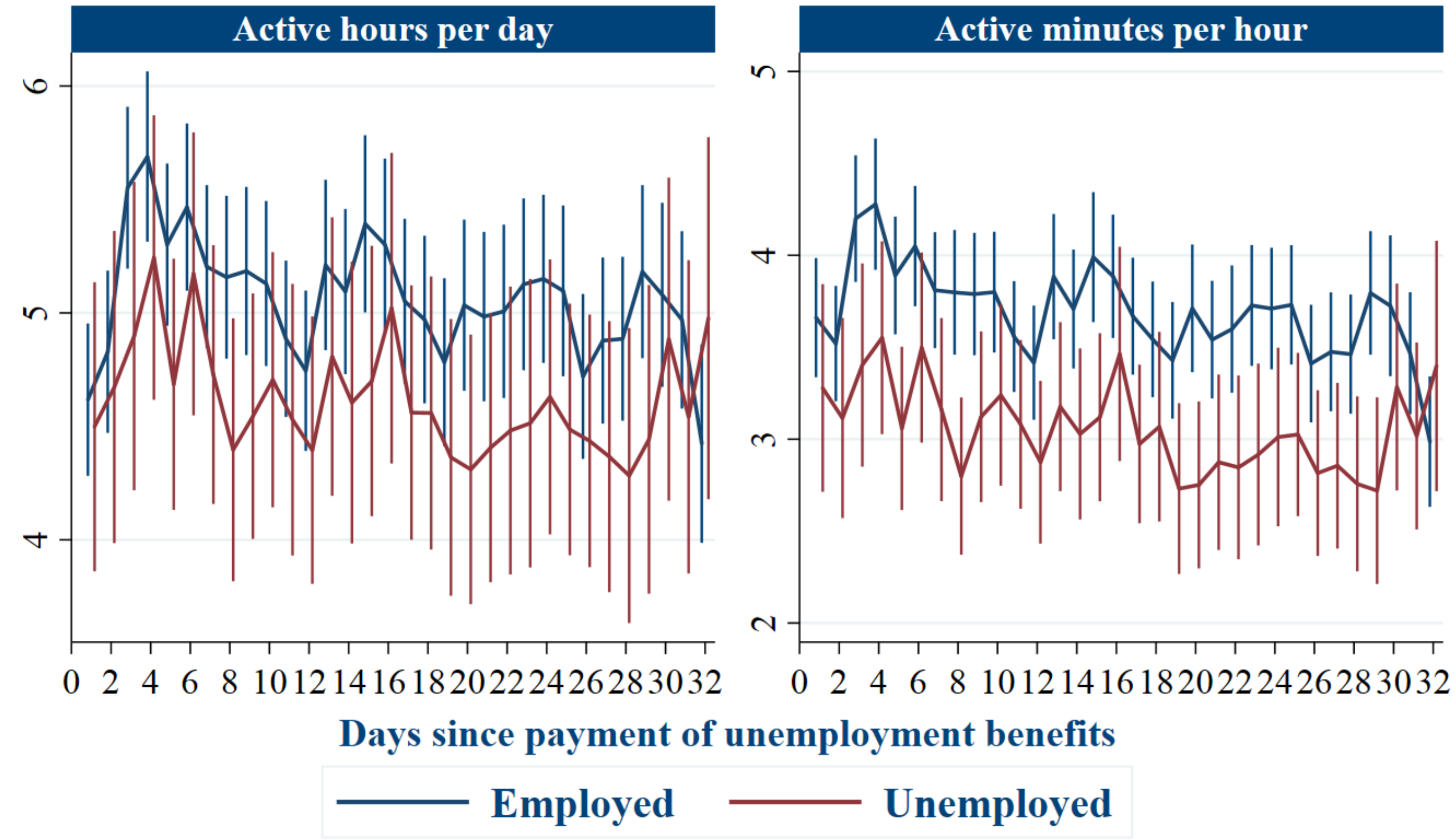


Smartphone ownership also correlates with…

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

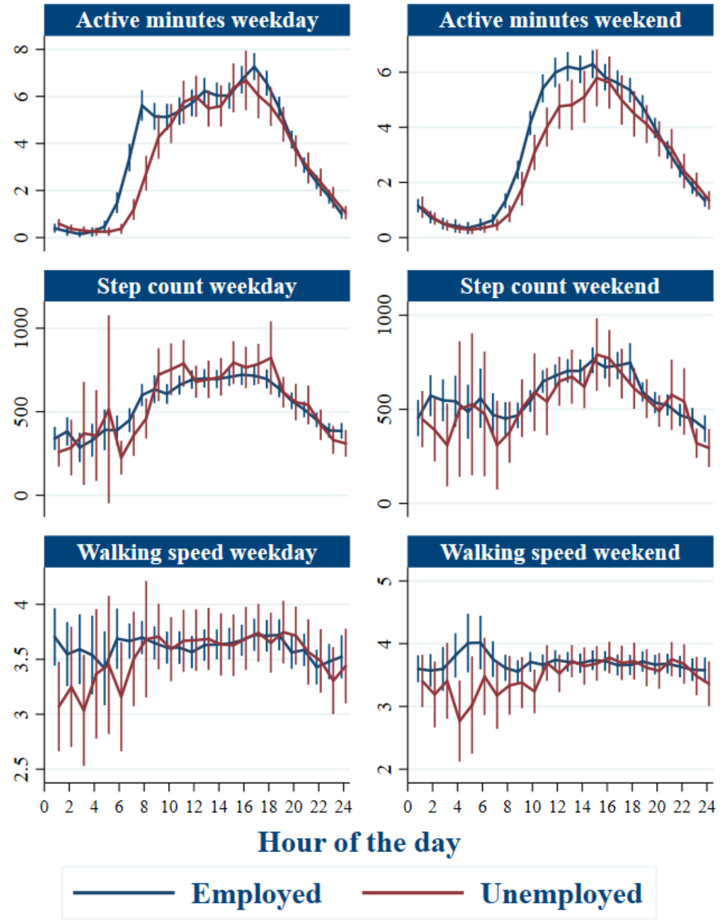
Sneak Peak

Rhythm of the payment of unemployment benefits



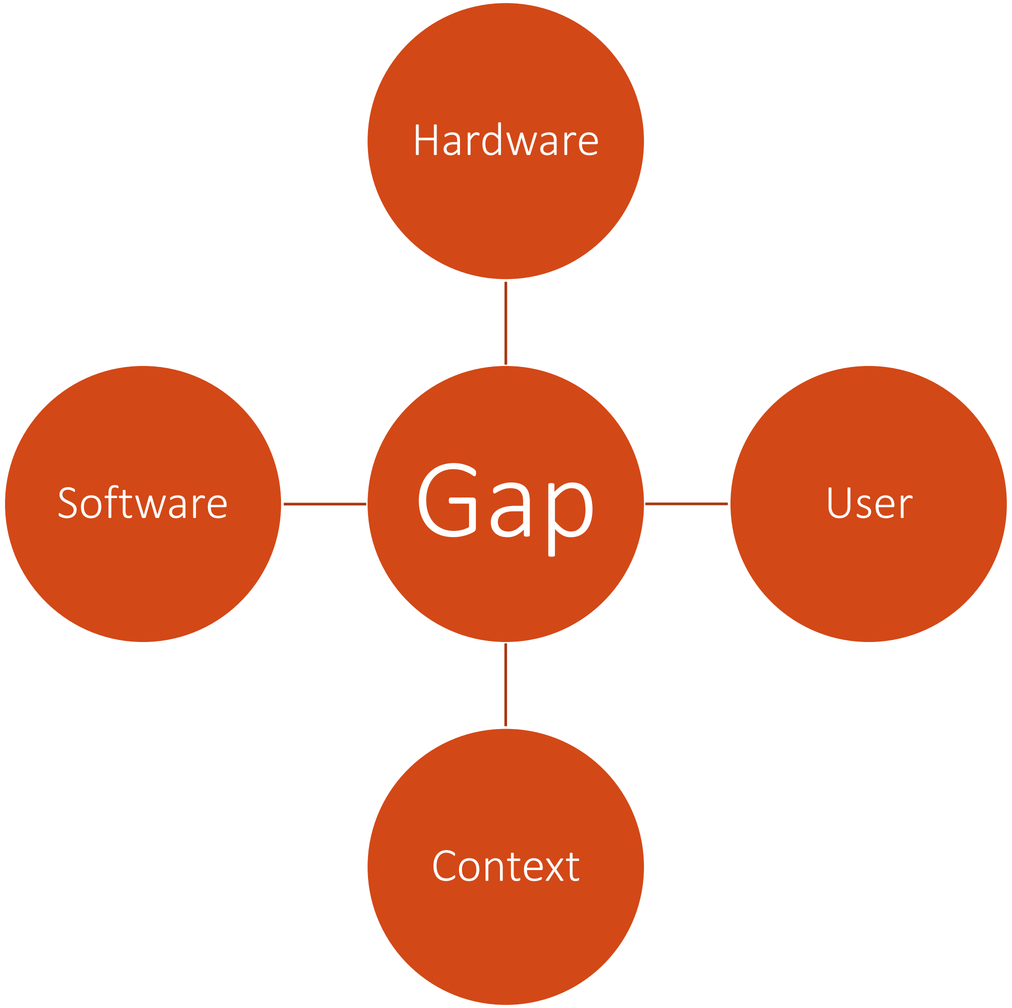
Predictive Margins with 95% confidence intervals.
Controls: Gender, age, weekday, hours smartphone is kept nearby.

Loss of day structure / resignation

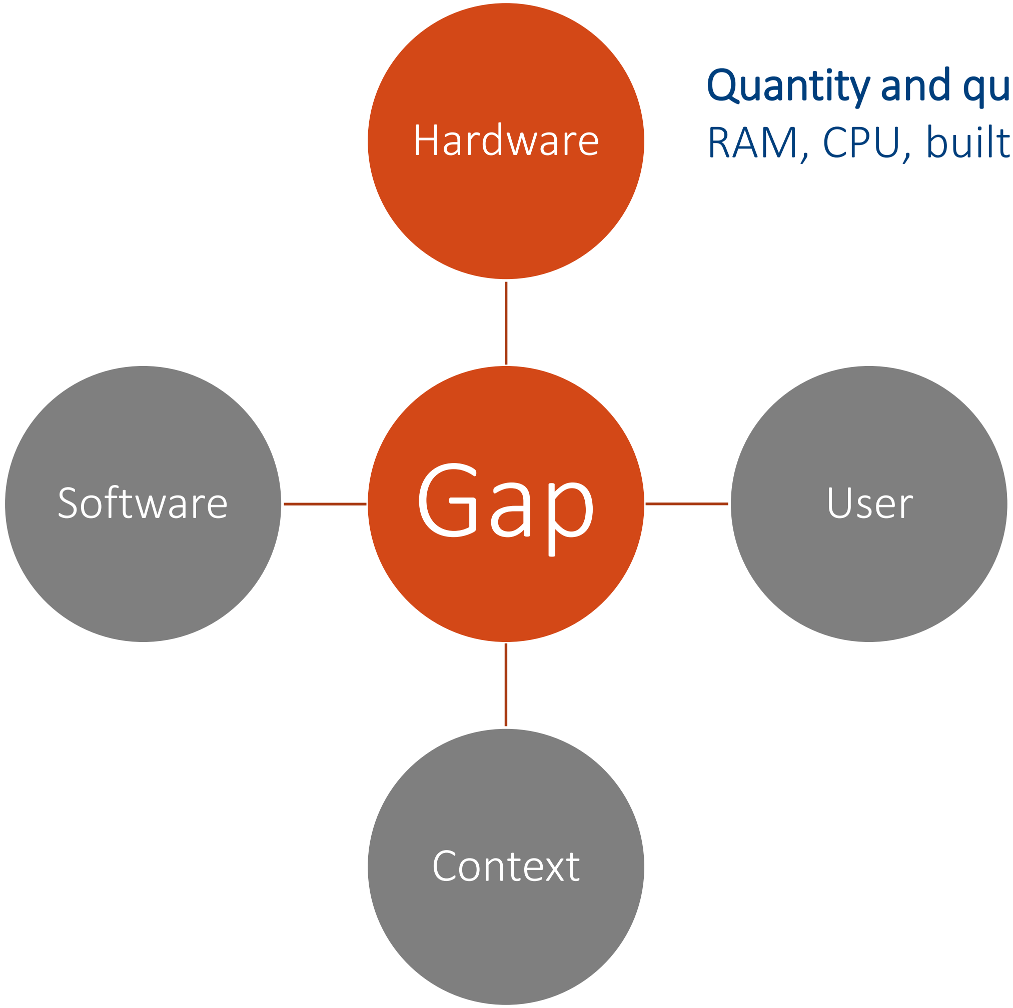


Predictive Margins with 95% confidence intervals.
Controls: Gender, age, hours smartphone is kept nearby.

Error sources



Error sources



Quantity and quality of components

RAM, CPU, built-in-sensors, battery

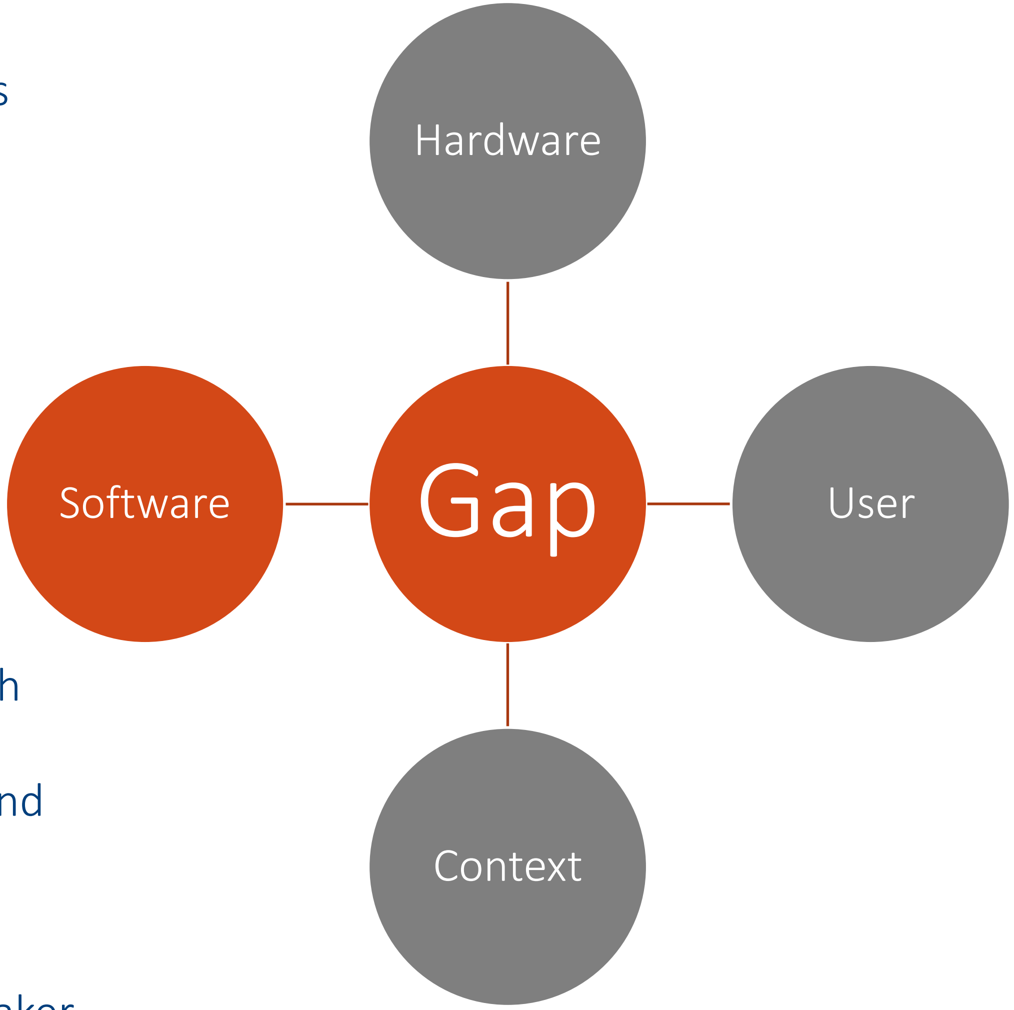
Error sources

- **Manufacturer Settings**
  Device specific doze-/battery saving modes inhibit data collection

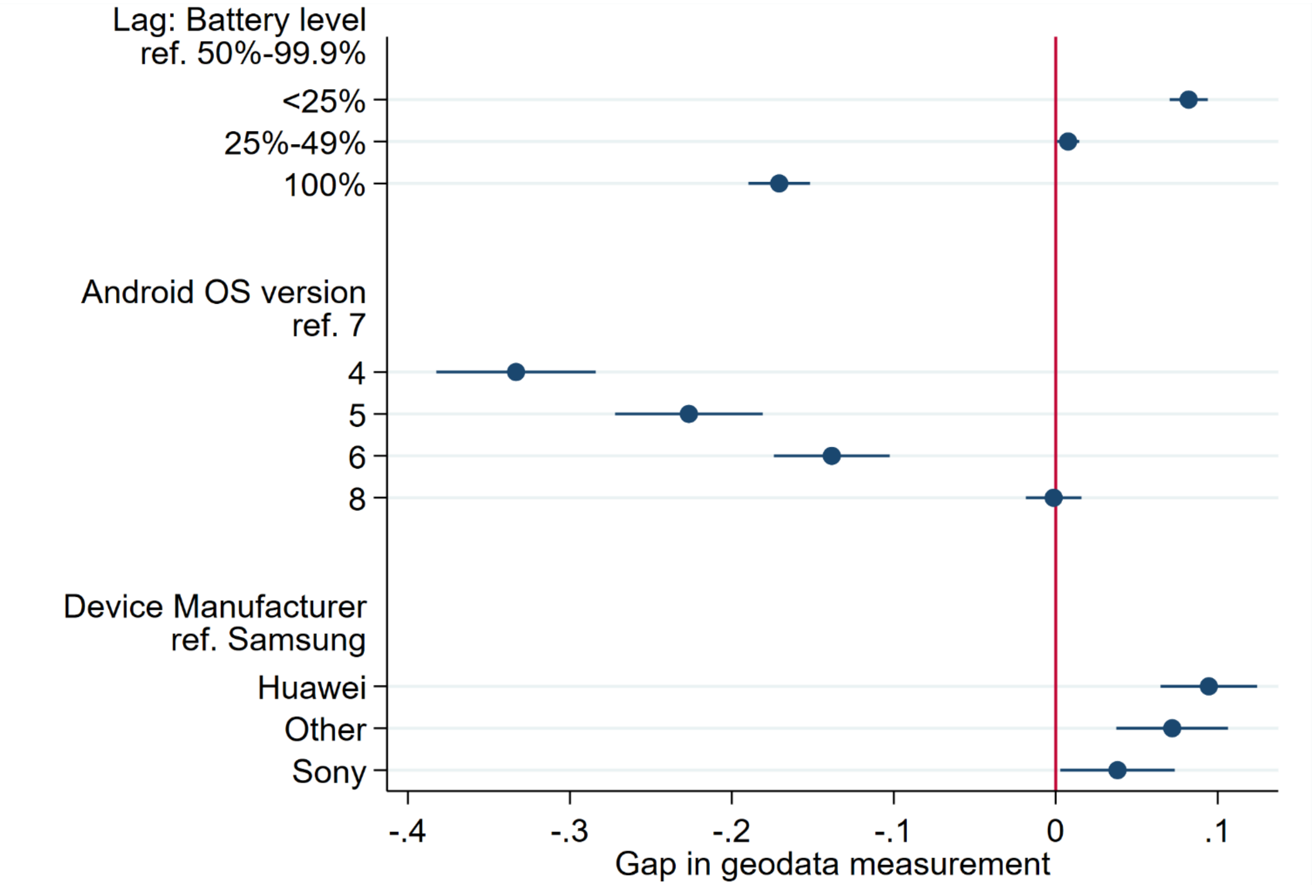
- **Operating System Settings**
  Data collection may be inhibited by the Operating System (OS)
  OS versions may vary in their rights management

- **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



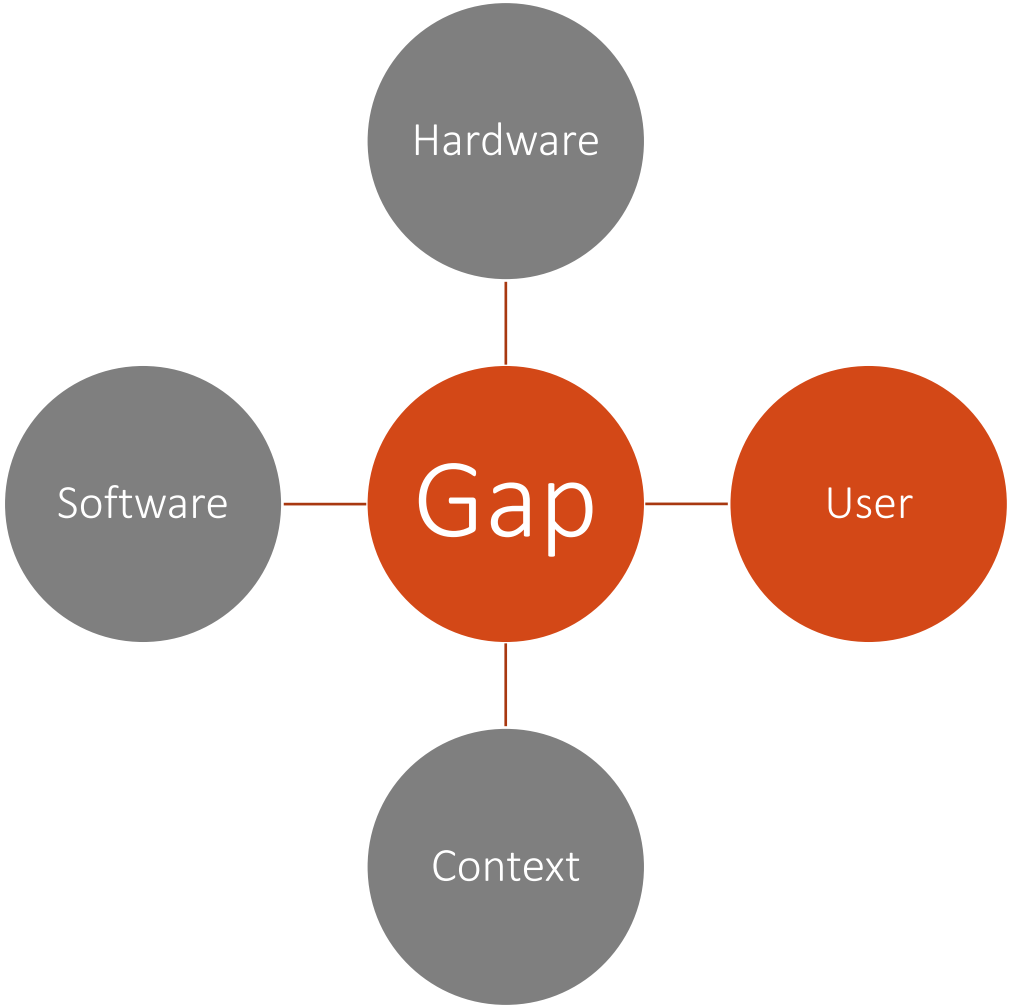
AME (with 95% CI) based on binomial probit regression with robust standard errors.

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



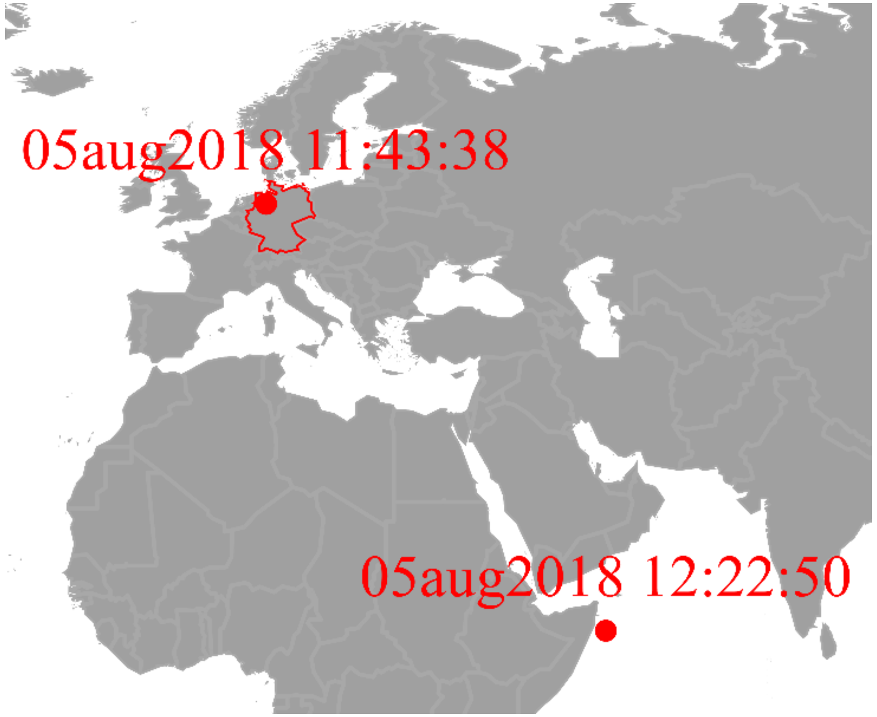
Participant characteristics

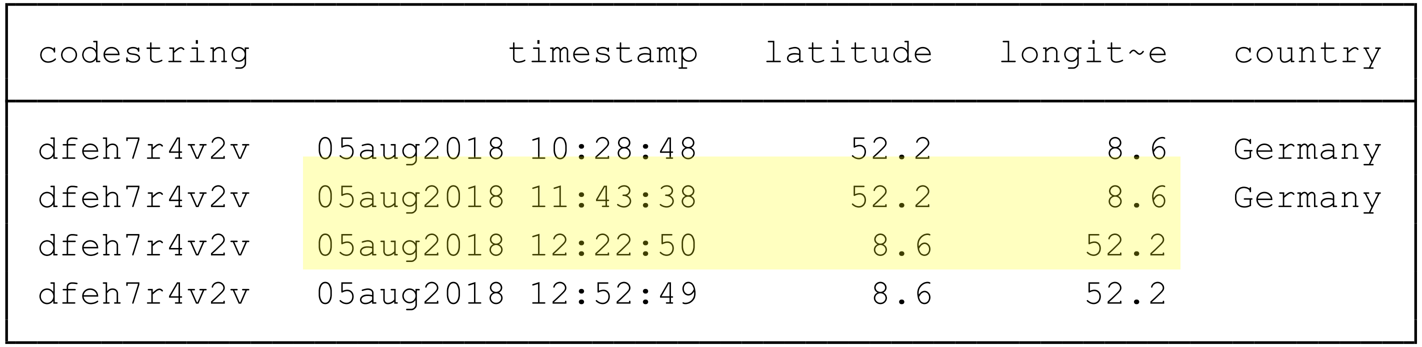
- Technical Competence

Participant behavior

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

User-related error sources





codestring	timestamp	latitude	longit~e	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	

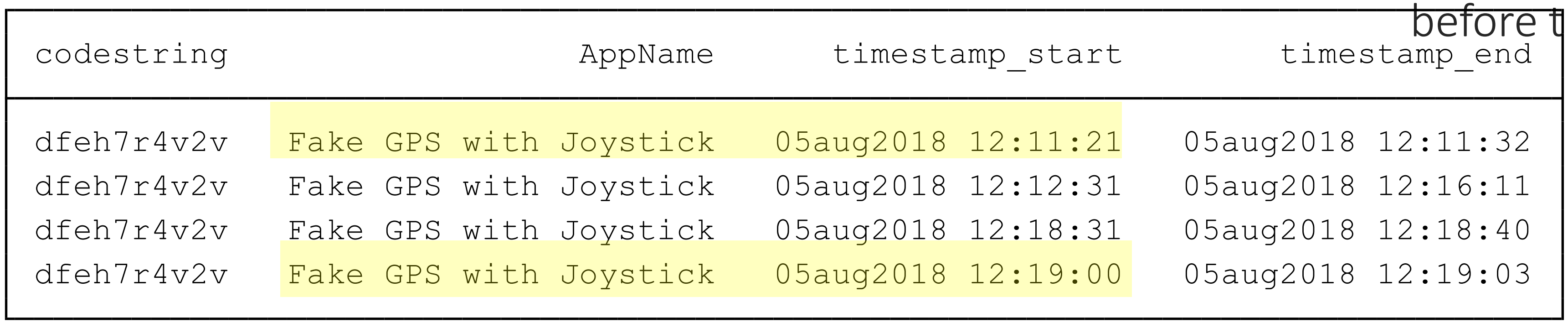
Apps falsify geolocation

Aim: Privacy, access location-specific content

Validation with app usage data

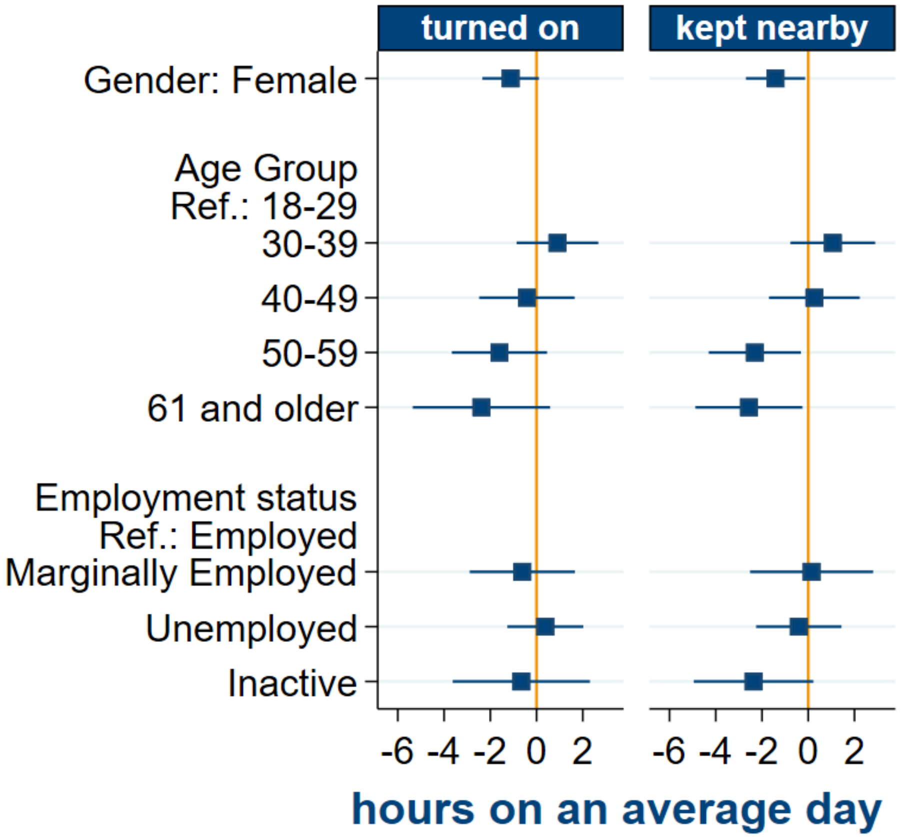
4 / 621 participants had such apps installed

→ 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



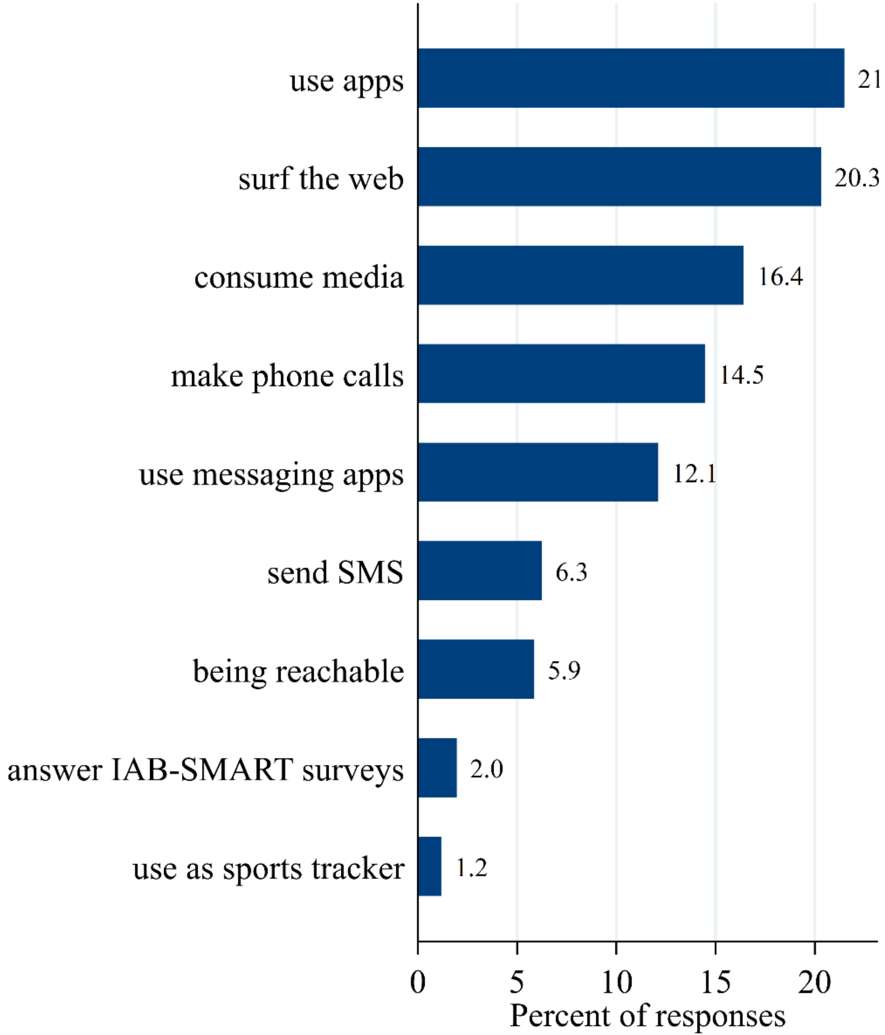
389 participants, AMEs with 95% confidence intervals.
Turned on - On average, how many hours per day is your smartphone turned on?
Kept nearby - How many hours is the smartphone in your immediate vicinity (i.e. on your body, in the same building / car)?

- 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

Quality assessment from In-App surveys



Reason for third-party device use

- 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

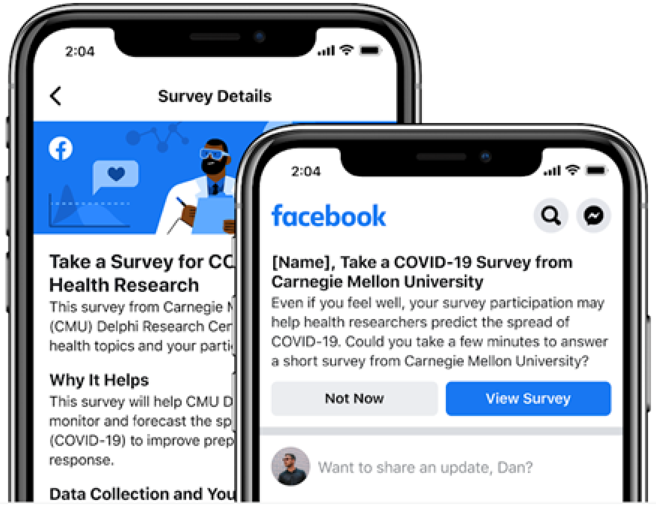
Responding to the Need for Syndromic Surveillance

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.

Project Overview

1 Who's Taking the Survey



Take a Survey for CC
Health Research
This survey from Carnegie M
(CMU) Delphi Research Cer
health topics and your parti

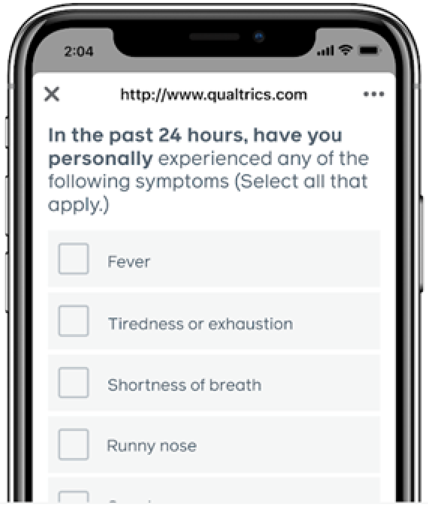
Why It Helps
This survey will help CMU D
monitor and forecast the sp
(COVID-19) to improve prep
response.

Data Collection and You

[Name], Take a COVID-19 Survey from Carnegie Mellon University
Even if you feel well, your survey participation may help health researchers predict the spread of COVID-19. Could you take a few minutes to answer a short survey from Carnegie Mellon University?

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

2 How the Survey Works



In the past 24 hours, have you **personally** experienced any of the following symptoms (Select all that apply.)

□ Fever
□ Tiredness or exhaustion
□ Shortness of breath
□ Runny nose

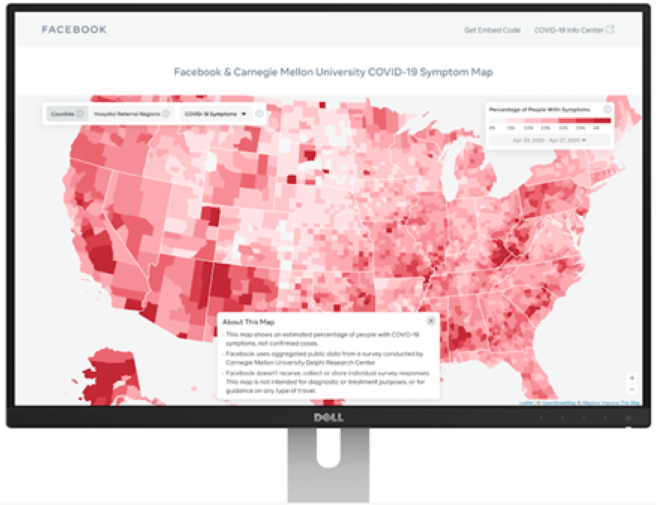
Carnegie Mellon University

University of Maryland 1856

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


FACEBOOK
Get Embed Code
COVID-19 Info Center
Facebook & Carnegie Mellon University COVID-19 Symptom Map
Counties
Hospital Referral Regions
COVID-19 Symptoms
Percentage of People With Symptoms
DELL

About This Map
- This map shows an estimated percentage of people with COVID-19 symptoms, not confirmed cases.
- Facebook uses aggregated public data from a survey conducted by Carnegie Mellon University Delphi Research Center.
- Facebook doesn't receive, collect or store individual survey responses. This map is not intended for diagnostic or treatment purposes, or for guidance on any type of travel.

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

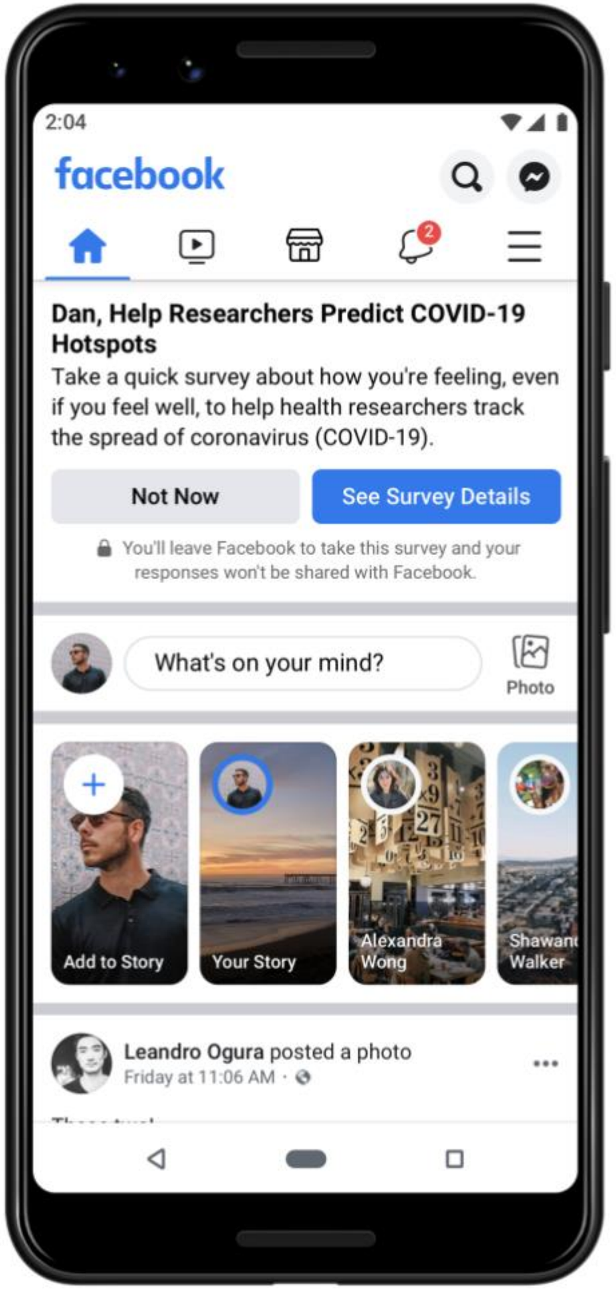
The non-aggregate data are available to eligible academic and nonprofit researchers by request.

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



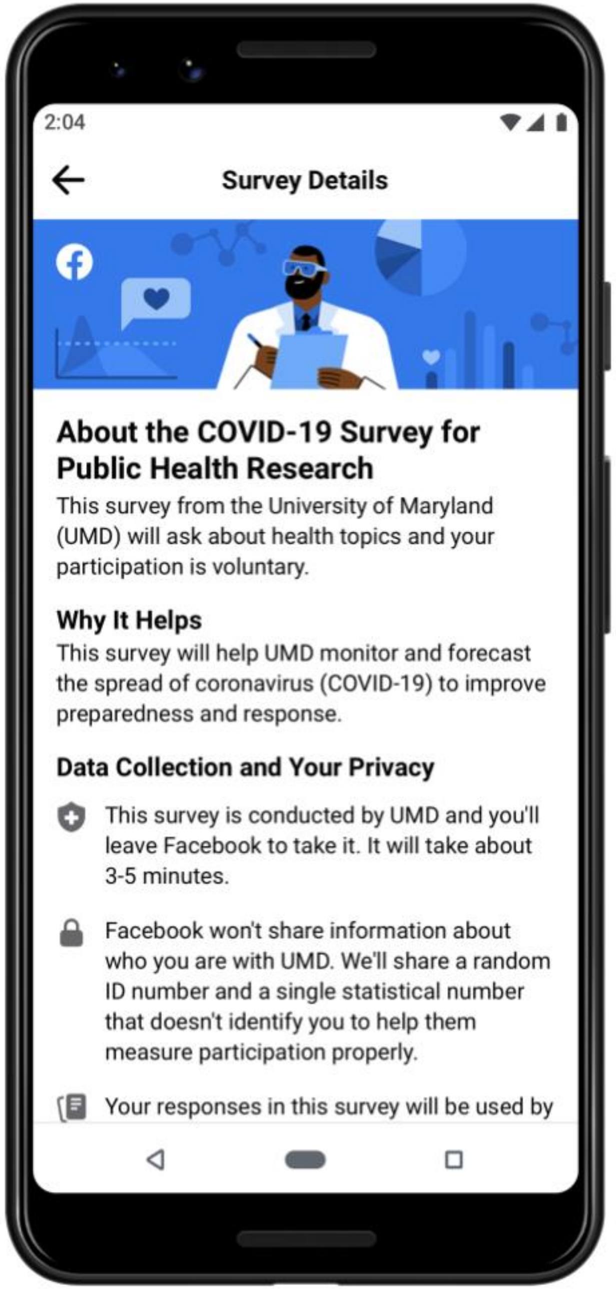
Dan, Help Researchers Predict COVID-19 Hotspots

Take a quick survey about how you're feeling, even if you feel well, to help health researchers track the spread of coronavirus (COVID-19).

Not Now | See Survey Details

You'll leave Facebook to take this survey and your responses won't be shared with Facebook.

Leandro Ogura posted a photo
Friday at 11:06 AM ·



Survey Details

About the COVID-19 Survey for Public Health Research

This survey from the University of Maryland (UMD) will ask about health topics and your participation is voluntary.

Why It Helps

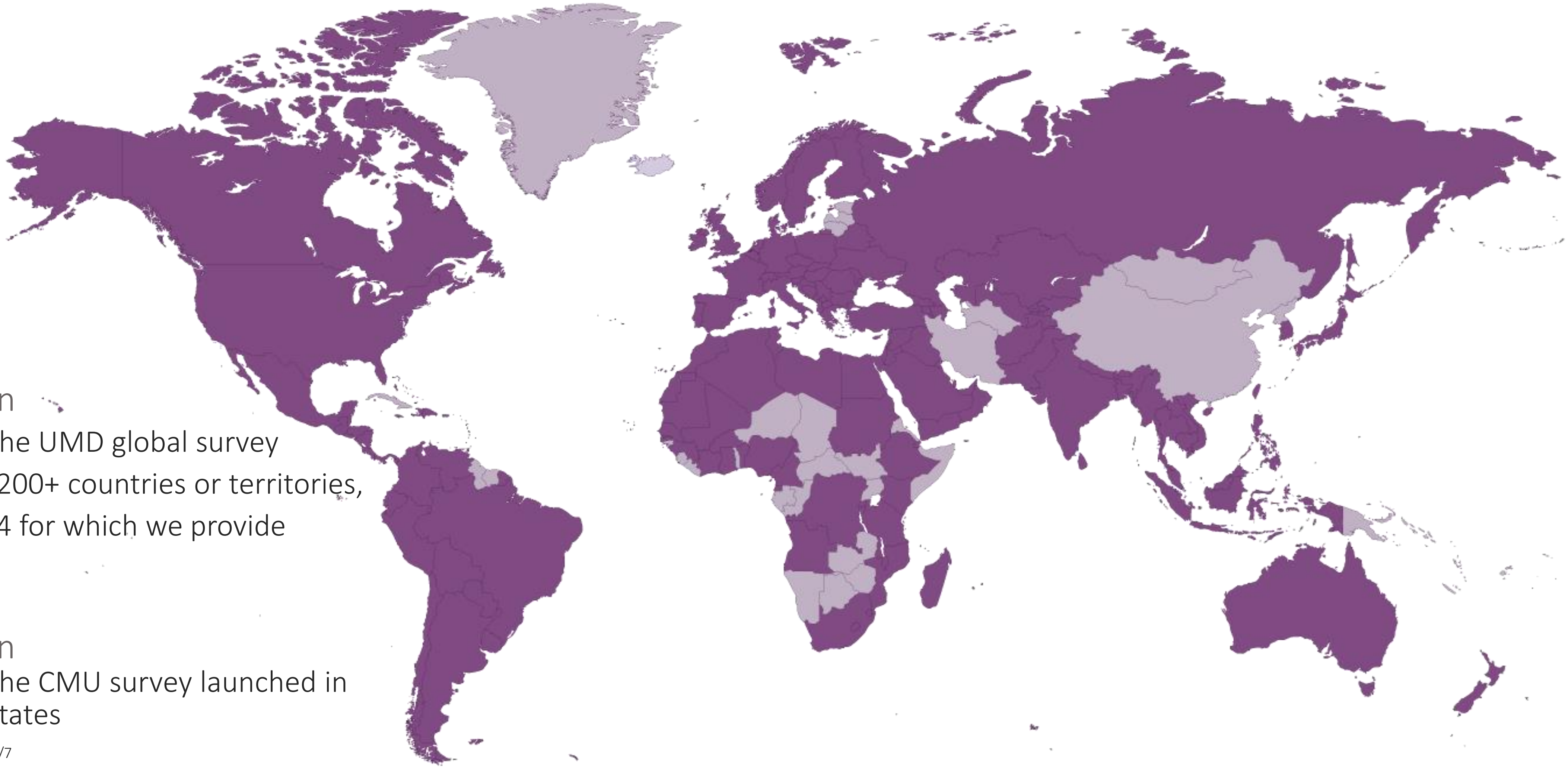
This survey will help UMD monitor and forecast the spread of coronavirus (COVID-19) to improve preparedness and response.

Data Collection and Your Privacy

This survey is conducted by UMD and you'll leave Facebook to take it. It will take about 3-5 minutes.

Facebook won't share information about who you are with UMD. We'll share a random ID number and a single statistical number that doesn't identify you to help them measure participation properly.

Your responses in this survey will be used by



20.3 million

Completed the UMD global survey launched in 200+ countries or territories, including 114 for which we provide weights

10.6 million

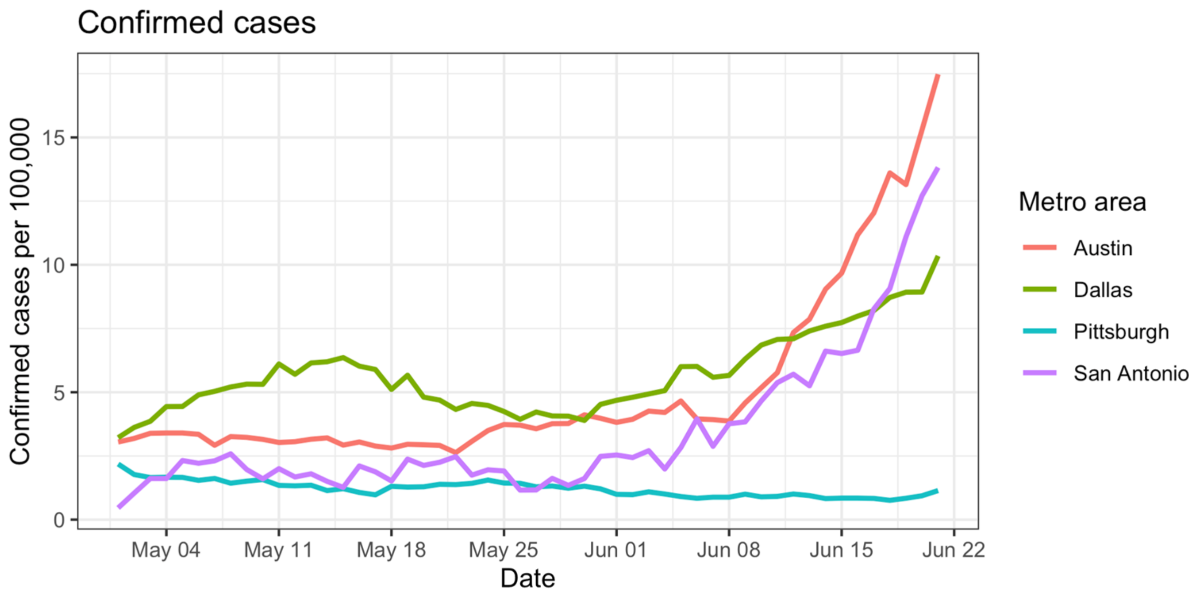
Completed the CMU survey launched in the United States

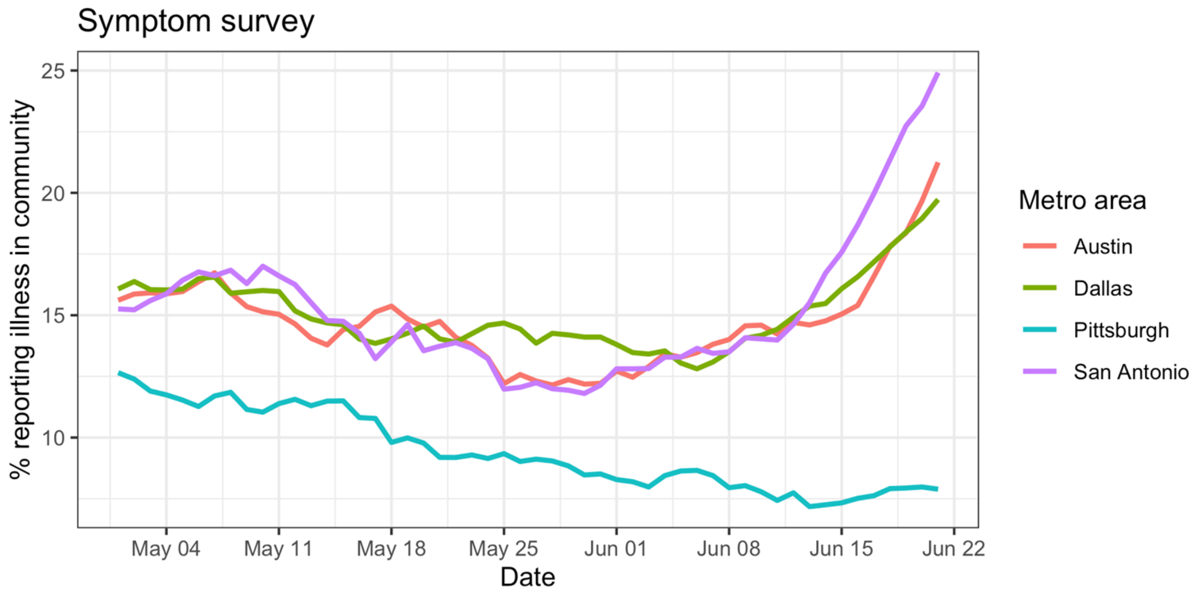
Note: sample size as of 9/7

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.

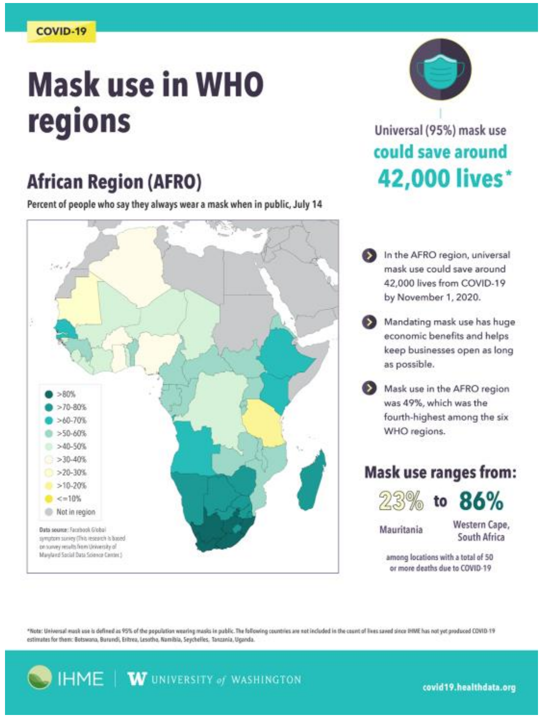




Early Research Insights

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.



COVID-19

Mask use in WHO regions

African Region (AFRO)

Percent of people who say they always wear a mask when in public, July 14

- ● >80%
- ● >70-80%
- ● >60-70%
- ● >50-60%
- ● >40-50%
- ○ >30-40%
- ○ >20-30%
- ● >10-20%
- ● <=10%
- ● Not in region

Data source: Facebook Global symptom survey (This research is based on survey results from University of Maryland Social Data Science Center.)

Universal (95%) mask use could save around 42,000 lives*

- In the AFRO region, universal mask use could save around 42,000 lives from COVID-19 by November 1, 2020.
- Mandating mask use has huge economic benefits and helps keep businesses open as long as possible.
- Mask use in the AFRO region was 49%, which was the fourth-highest among the six WHO regions.

Mask use ranges from:

23% to 86%

Mauritania — Western Cape, South Africa

among locations with a total of 50 or more deaths due to COVID-19

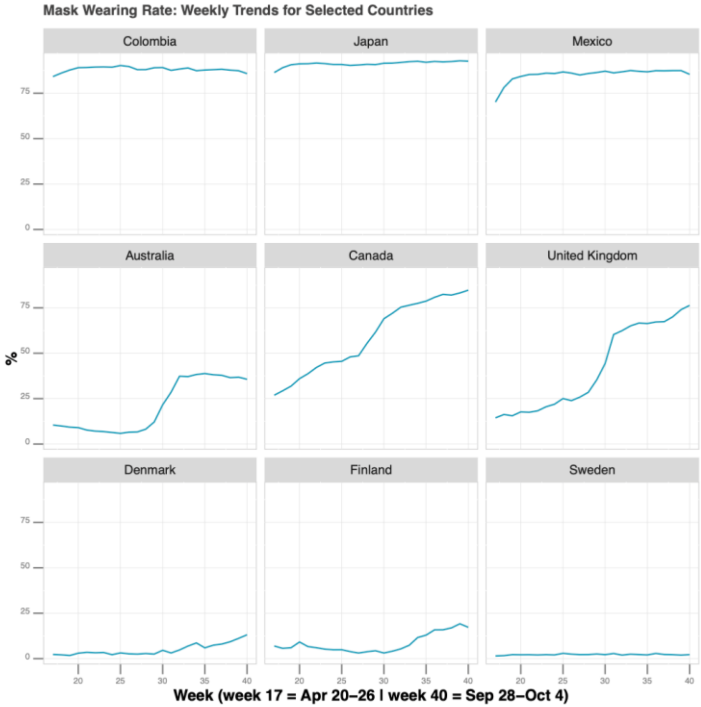
*Note: Universal mask use is defined as 95% of the population wearing masks in public. The following countries are not included in the count of lives saved since IHME has not yet produced COVID-19 estimates for them: Botswana, Burundi, Eritrea, Lesotho, Namibia, Seychelles, Tanzania, Uganda.

IHME | UNIVERSITY of WASHINGTON

covid19.healthdata.org

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?”



Mask Wearing Rate: Weekly Trends for Selected Countries

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

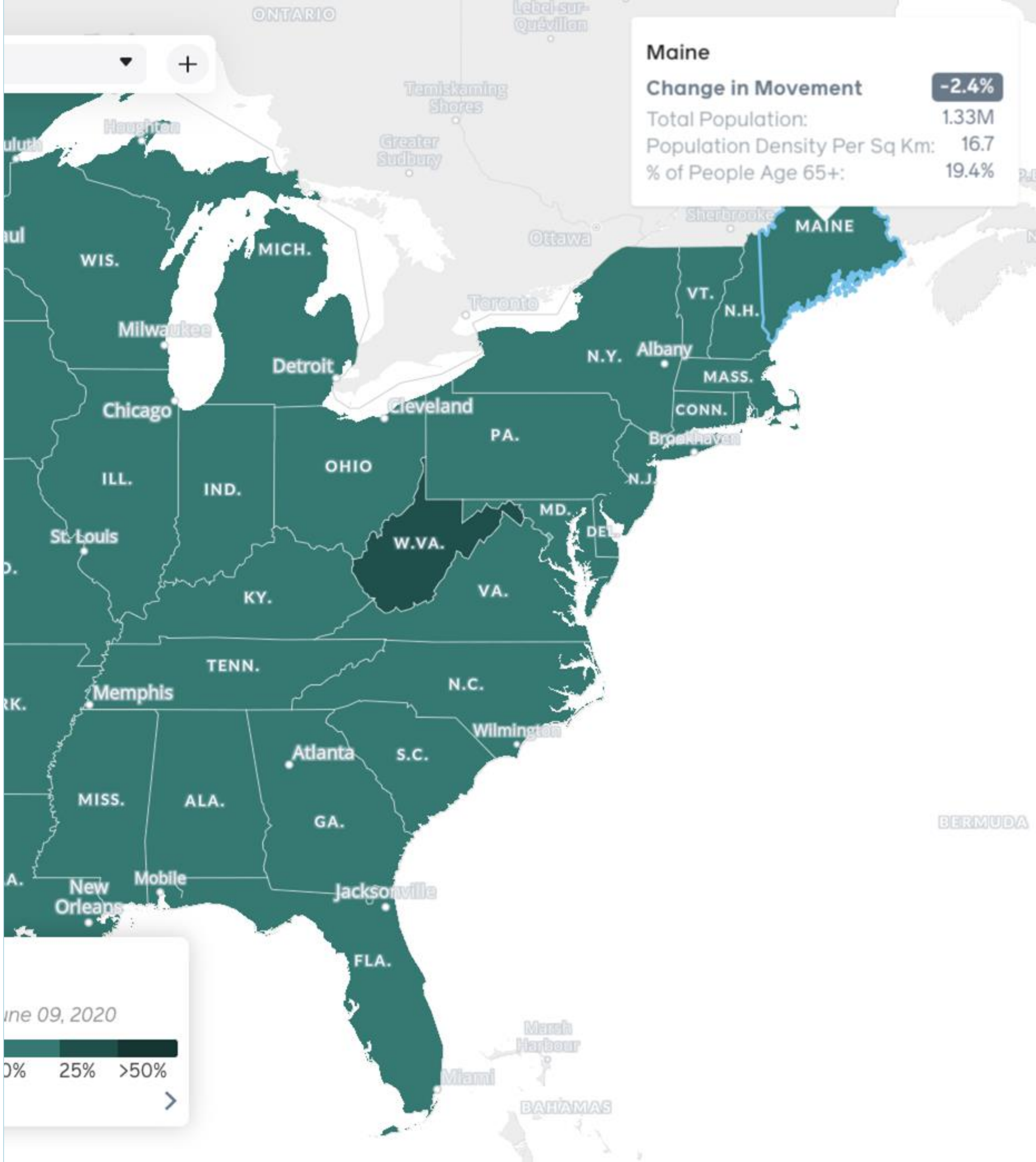
Population Density Maps

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/.



Maine

Change in Movement	-2.4%
Total Population:	1.33M
Population Density Per Sq Km:	16.7
% of People Age 65+:	19.4%

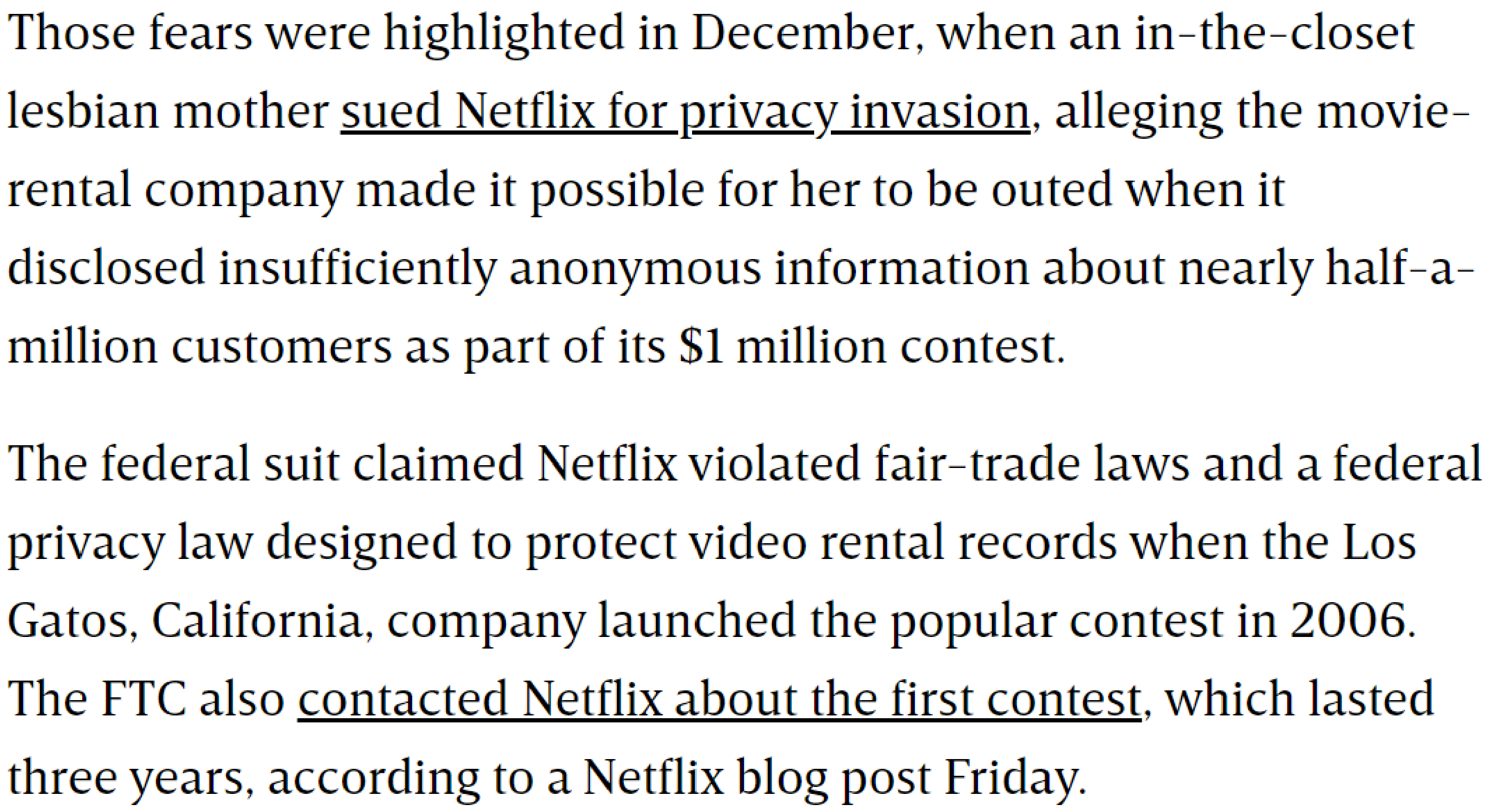
une 09, 2020

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

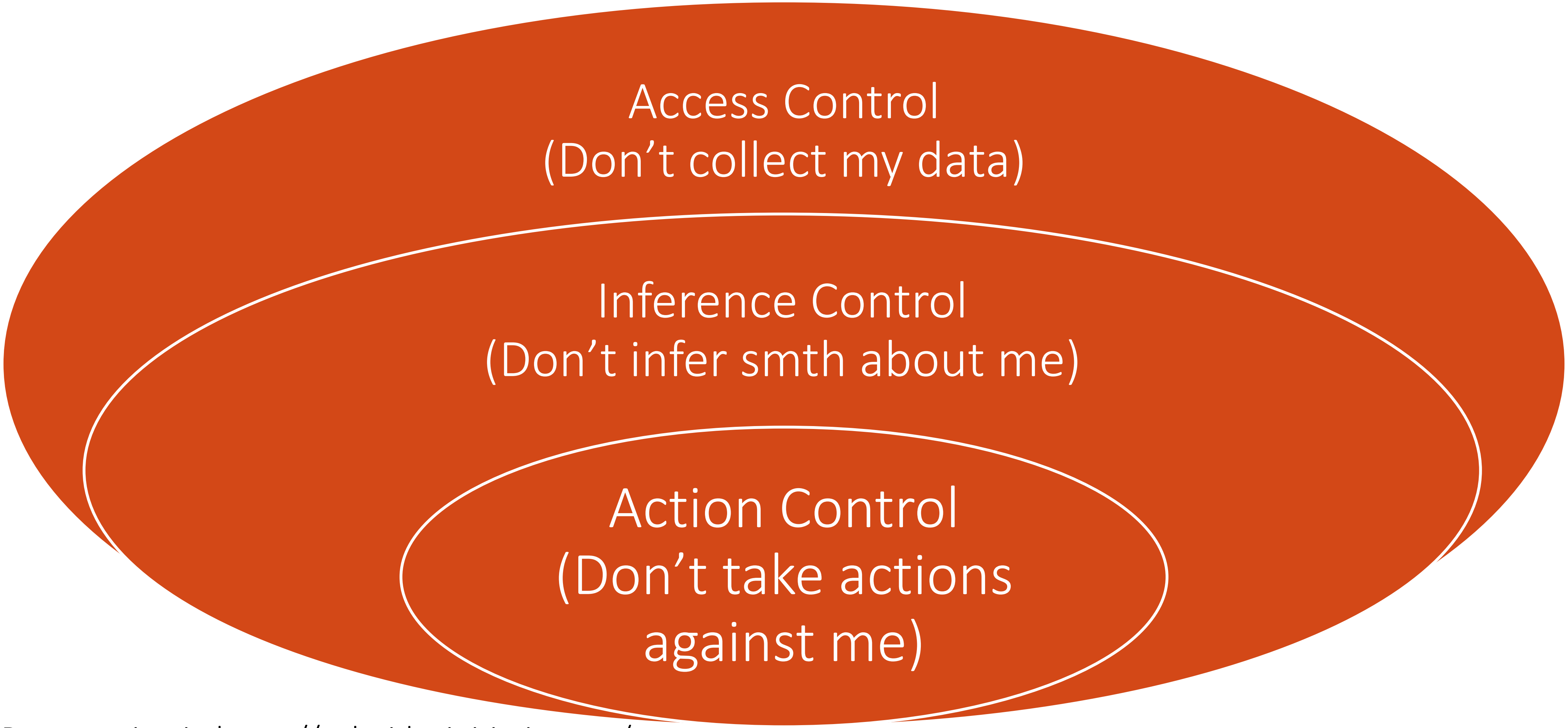
Netflix



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



Access Control
(Don’t collect my data)

Inference Control
(Don’t infer smth about me)

Action Control
(Don’t take actions against me)

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

*The data you **already provided** to us whould 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

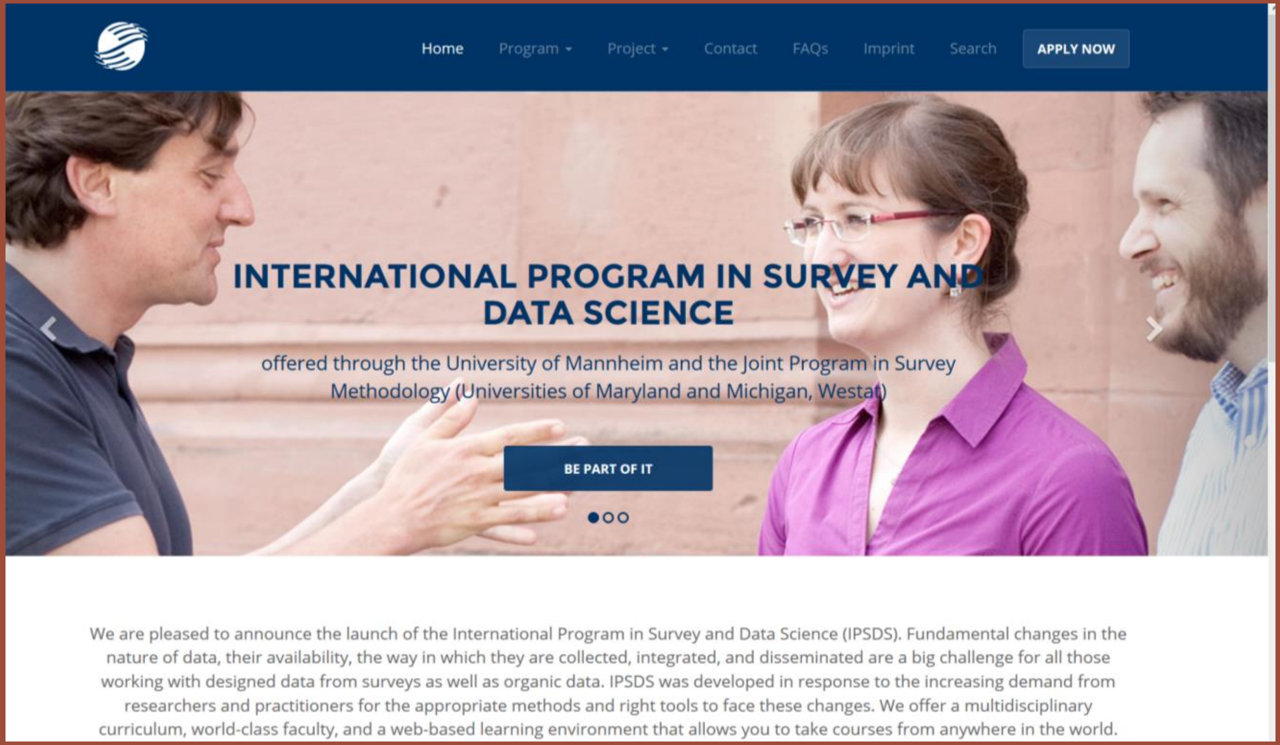
Web	Front	Back	Total
% agree	82.6	62.4	520

*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

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



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, Westat)

We are pleased to announce the launch of the International Program in Survey and Data Science (IPSDS). Fundamental changes in the nature of data, their availability, the way in which they are collected, integrated, and disseminated are a big challenge for all those working with designed data from surveys as well as organic data. IPSDS was developed in response to the increasing demand from researchers and practitioners for the appropriate methods and right tools to face these changes. We offer a multidisciplinary curriculum, world-class faculty, and a web-based learning environment that allows you to take courses from anywhere in the world.

THANK YOU!

fkreuter@umd.edu https://survey-data-science.net/ http://socialdatascience.umd.edu/