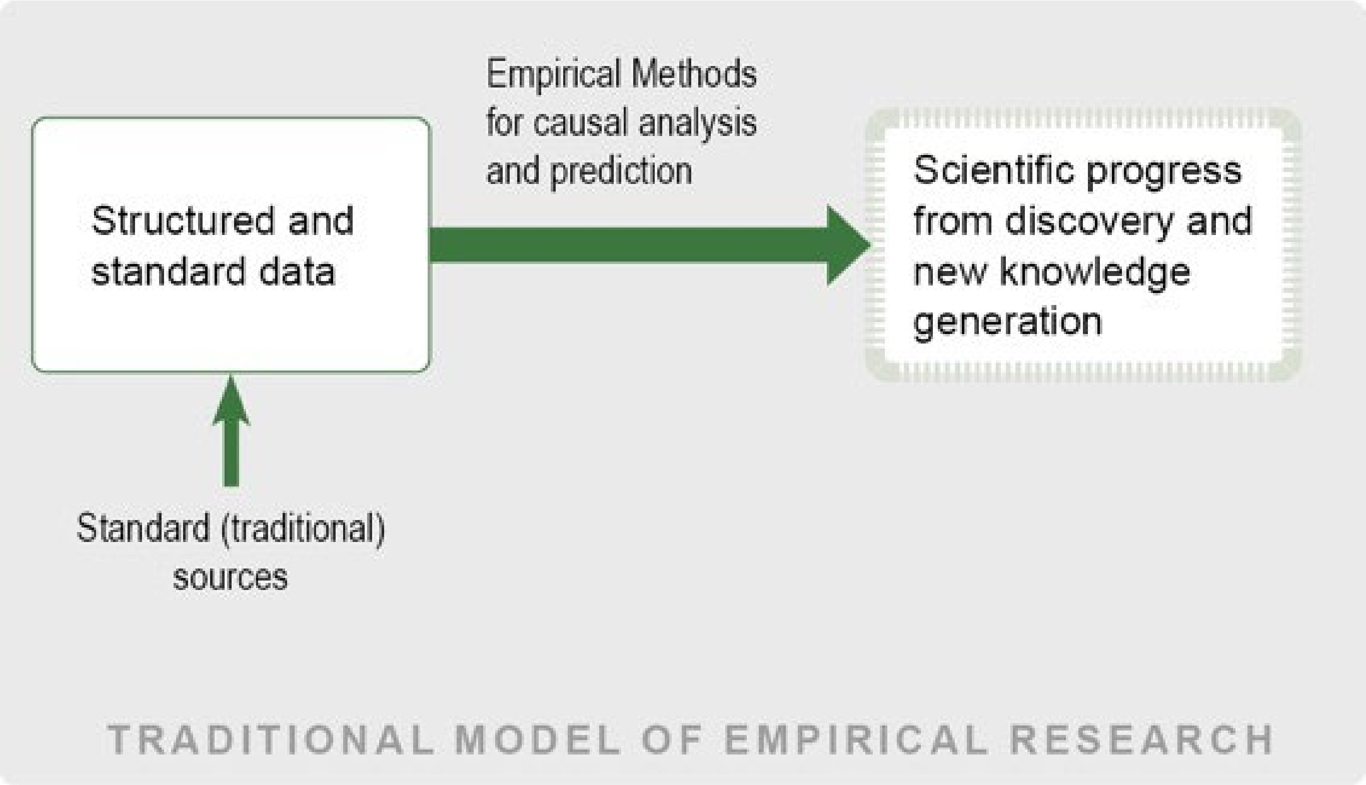
New Measurements in New Data Infrastructures.

The long road from vision to practice.

Frauke Kreuter
Joint Program in Survey Methodology, University of Maryland
LMU Munich Statistics and Data Science



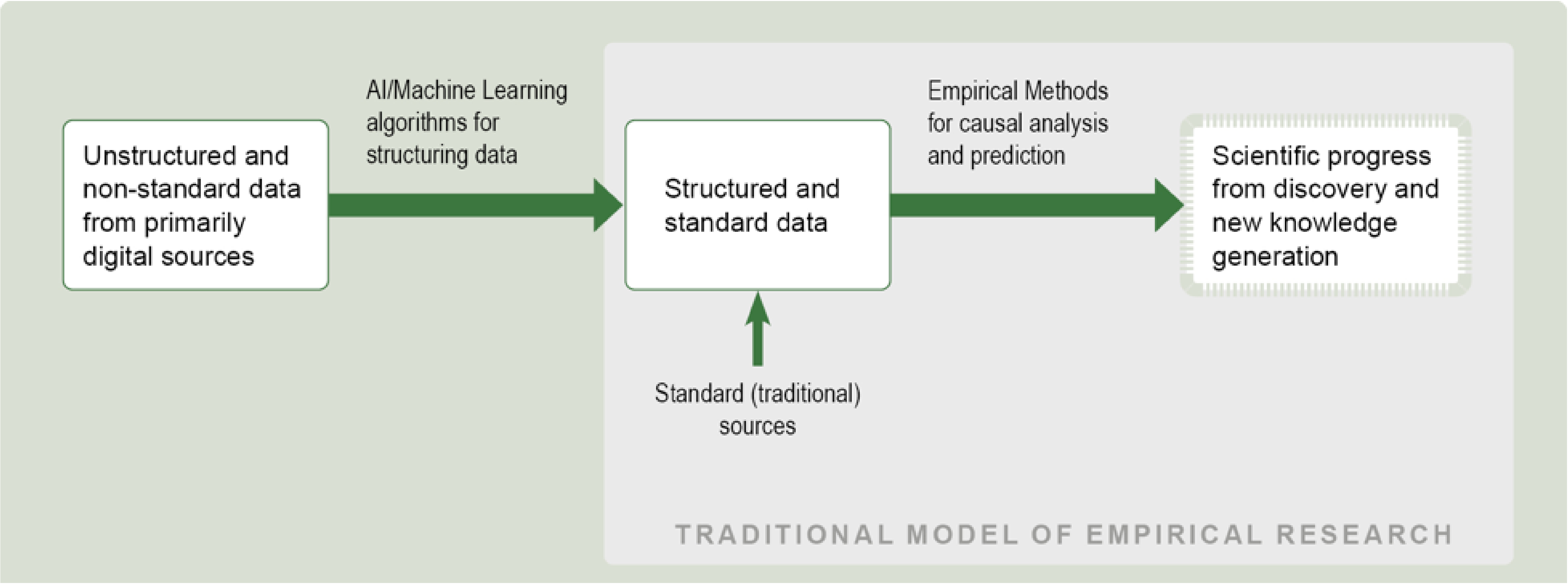
Structured and standard data

Standard (traditional) sources

Empirical Methods for causal analysis and prediction

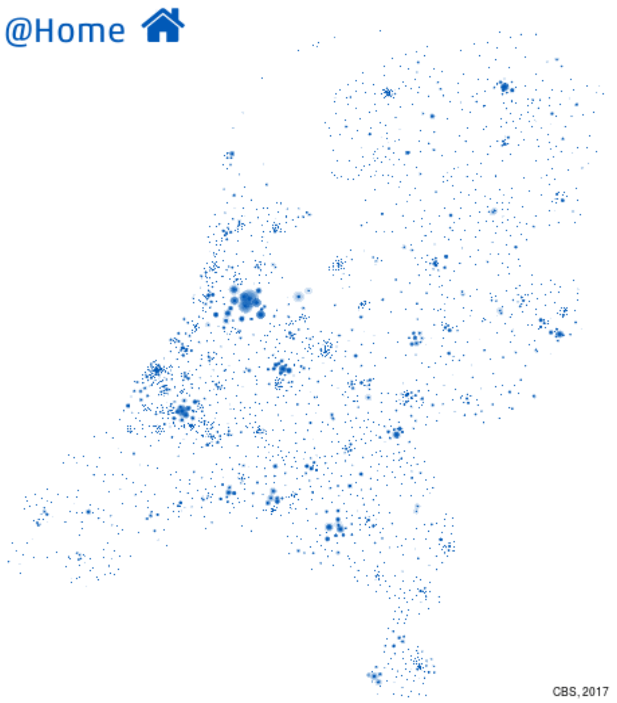
Scientific progress from discovery and new knowledge generation

TRADITIONAL MODEL OF EMPIRICAL RESEARCH


Unstructured and non-standard data from primarily digital sources
AI/Machine Learning algorithms for structuring data
Structured and standard data
Standard (traditional) sources
Empirical Methods for causal analysis and prediction
Scientific progress from discovery and new knowledge generation
TRADITIONAL MODEL OF EMPIRICAL RESEARCH

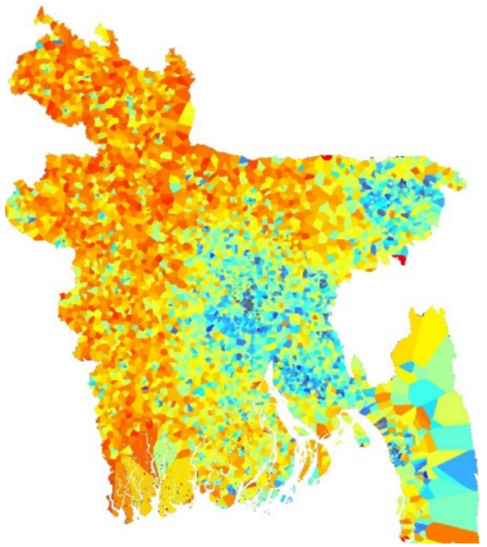
New Linked Measurements

Sustainable Communities



Source: https://www.cbs.nl/en-gb/our-services/innovation/project/towards-motives-behind-mobility

No Poverty

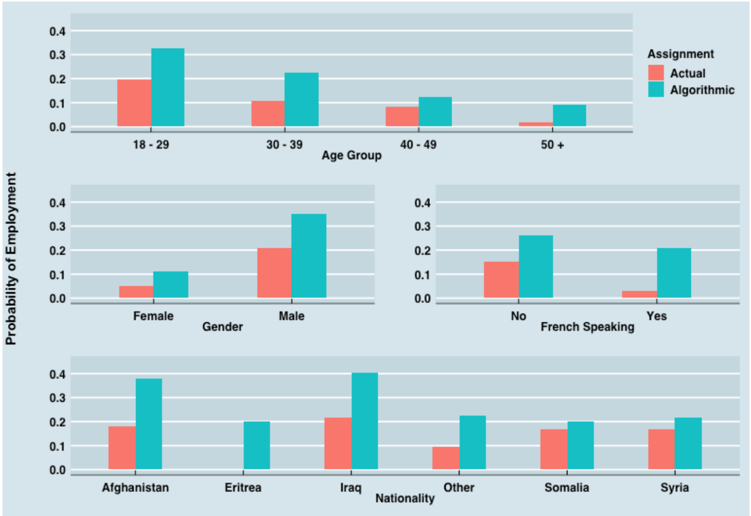


https://doi.org/10.6084/m9.figshare.c.3662800.v1

See also Steele, J. et al. (2017): „Mapping poverty using mobile phone and satellite data.", *Journal of the Royal Society Interface*, 14.

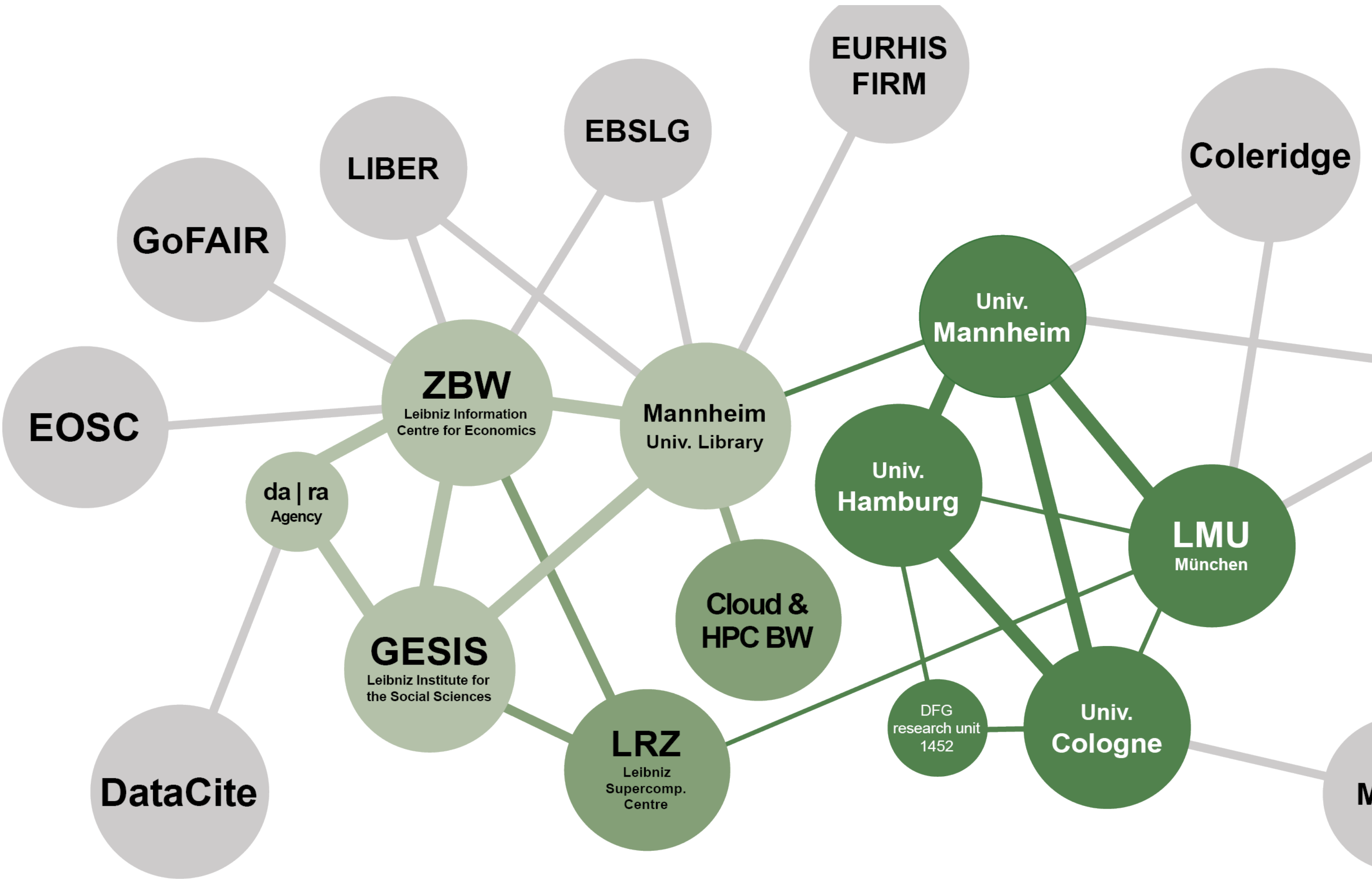
See also Jean, N. et al. (2016): „Use of satellite imagery and machine learning to predict poverty.", *Science*, 353(6301), 790-794.

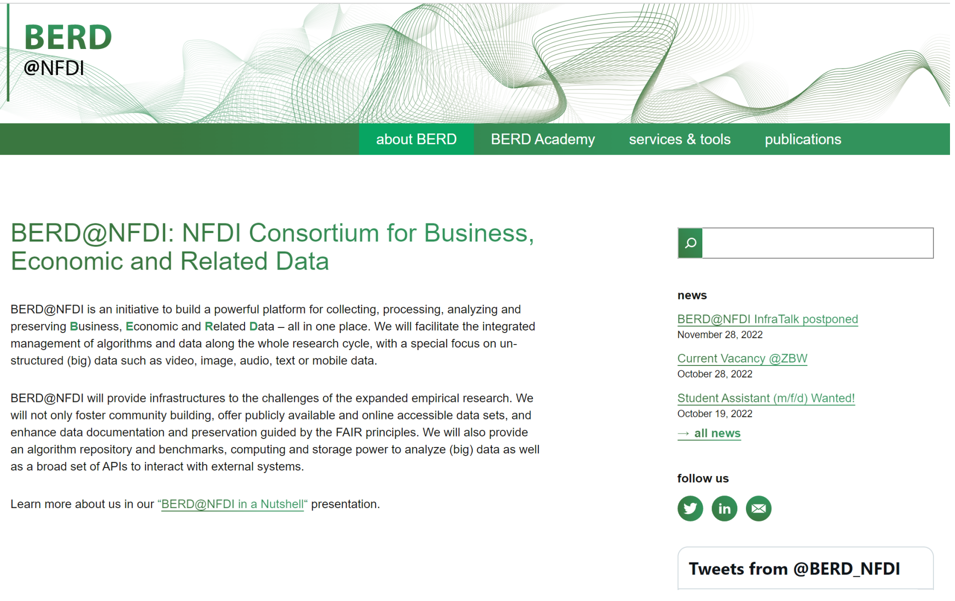
Decent Work and Economic Growth



See also Bansak, K. et al. (2018) Improving refugee integration through data-driven algorithmic assignment. Science, 359, 6373, 325-329.

Context





BERD@NFDI: NFDI Consortium for Business, Economic and Related Data

BERD@NFDI is an initiative to build a powerful platform for collecting, processing, analyzing and preserving **B**usiness, **E**conomic and **R**elated **D**ata – all in one place. We will facilitate the integrated management of algorithms and data along the whole research cycle, with a special focus on unstructured (big) data such as video, image, audio, text or mobile data.

BERD@NFDI will provide infrastructures to the challenges of the expanded empirical research. We will not only foster community building, offer publicly available and online accessible data sets, and enhance data documentation and preservation guided by the FAIR principles. We will also provide an algorithm repository and benchmarks, computing and storage power to analyze (big) data as well as a broad set of APIs to interact with external systems.

Learn more about us in our "BERD@NFDI in a Nutshell" presentation.

news

BERD@NFDI InfraTalk postponed
November 28, 2022

Current Vacancy @ZBW
October 28, 2022

Student Assistant (m/f/d) Wanted!
October 19, 2022

→ **all news**

follow us

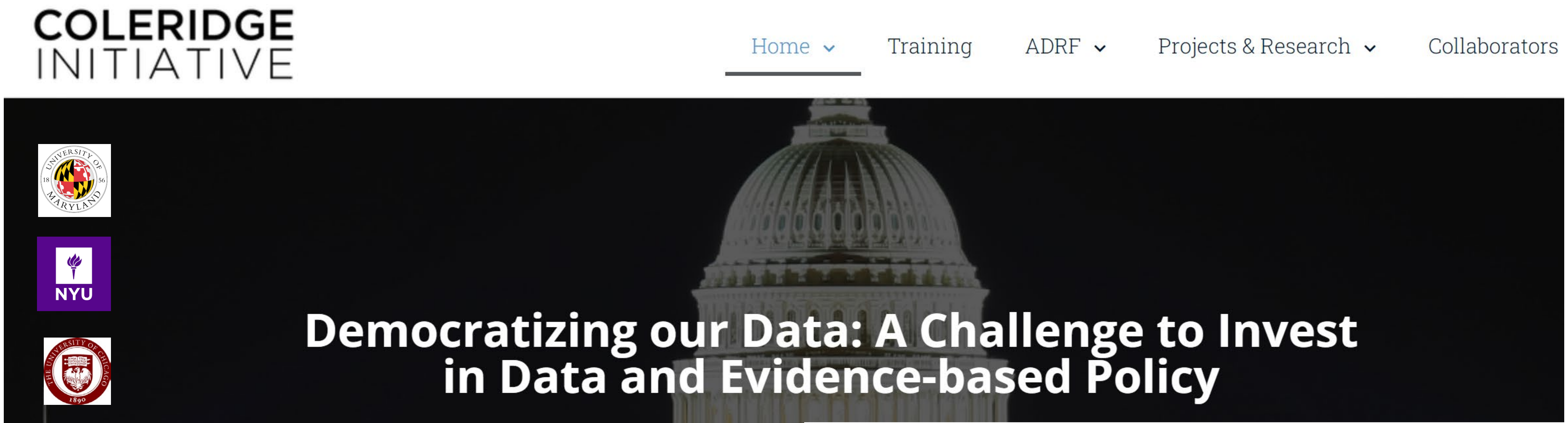
Tweets from @BERD_NFDI

Lessons Learned

1. Creating new measures out of linked data is not enough. The effort has to be tied to a product.
2. Research questions need to guide decisions on measurements and data sources.
3. Data Science is a “Team Sport” … and needs to be treated as such.

Lessons Learned

1. Creating new measures out of linked data is not enough. The effort has to be tied to a product.



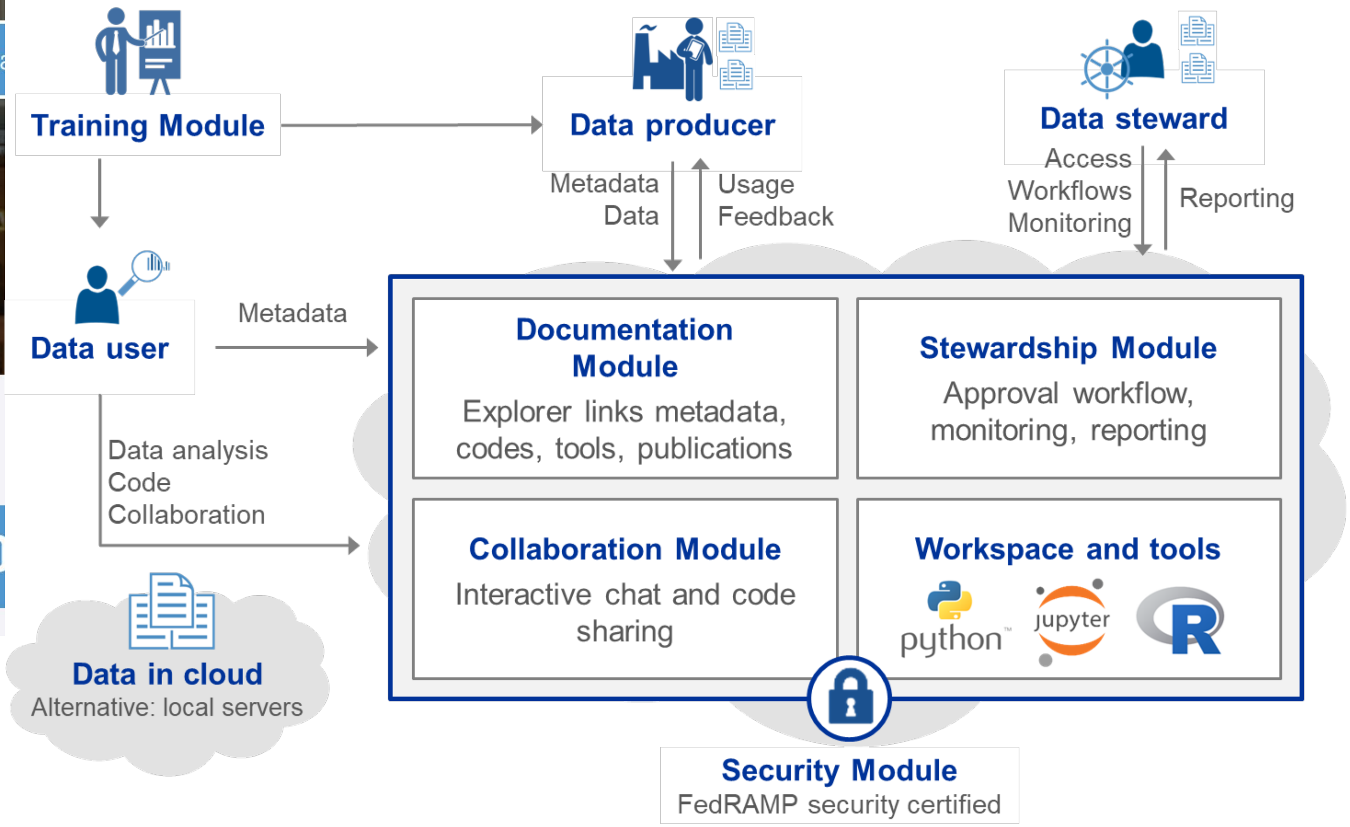
COLERIDGE INITIATIVE

Home ⌄ Training ADRF ⌄ Projects & Research ⌄ Collaborators

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

Kreuter, F., Ghani, R., & Lane, J. (2019). Change Through Data: A Data Analytics Training Program for Government Employees. *Harvard Data Science Review*, *1*(2). https://doi.org/10.1162/99608f92.ed353ae3

https://coleridgeinitiative.org/



Training Module

Data producer

Metadata Data

Usage Feedback

Data steward

Access Workflows Monitoring

Reporting

Data user

Metadata

Data analysis Code Collaboration

Documentation Module

Explorer links metadata, codes, tools, publications

Stewardship Module

Approval workflow, monitoring, reporting

Collaboration Module

Interactive chat and code sharing

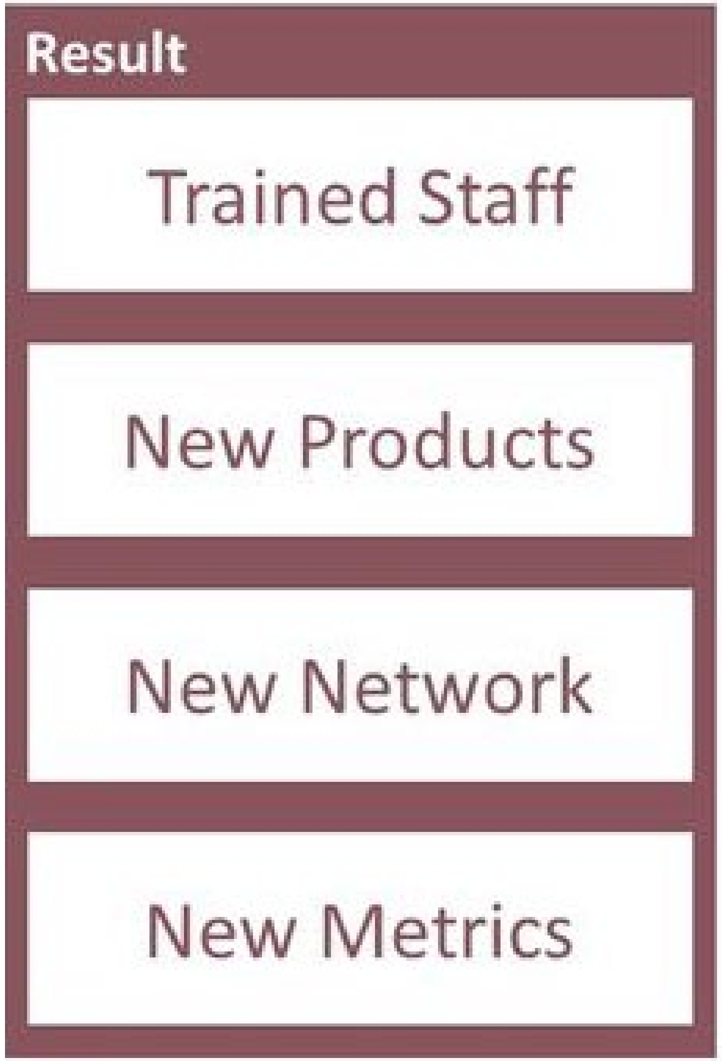
Workspace and tools

Data in cloud

Alternative: local servers

Security Module

FedRAMP security certified



Result

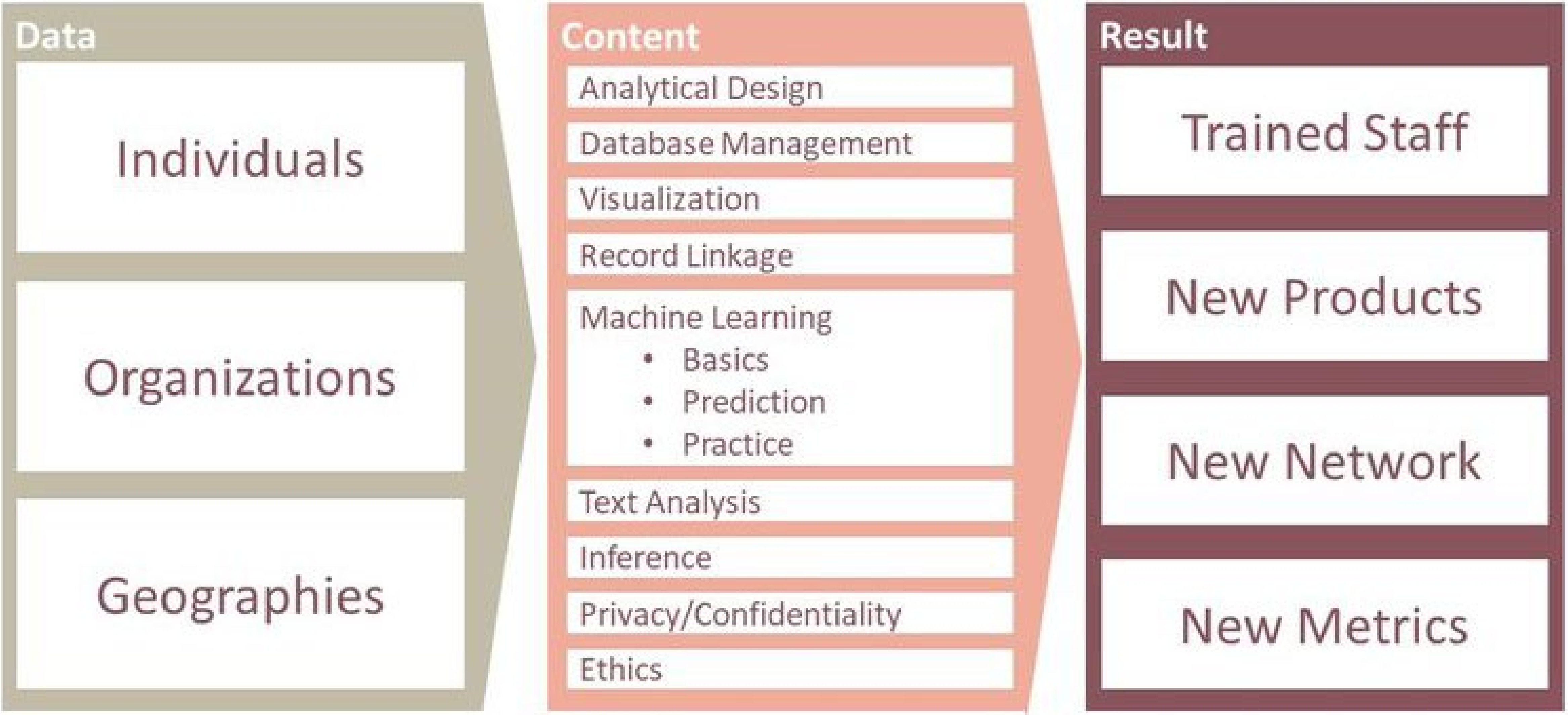
Trained Staff

New Products

New Network

New Metrics

Kreuter, Ghani, Lane (2019): Change through data. A Data Analytics Training Program for Government Employees.
Harvard Data Science Review,1.2



Data

Individuals

Organizations

Geographies

Content

Analytical Design

Database Management

Visualization

Record Linkage

Machine Learning

- Basics
- Prediction
- Practice

Text Analysis

Inference

Privacy/Confidentiality

Ethics

Result

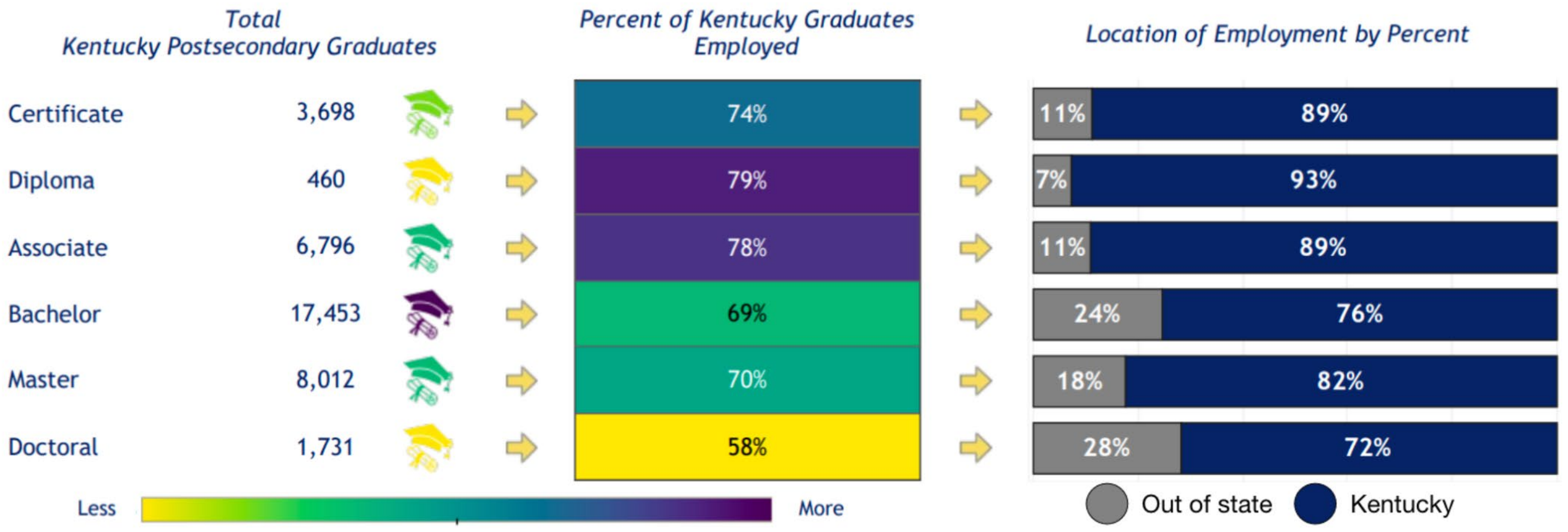
Trained Staff

New Products

New Network

New Metrics

Kreuter, Ghani, Lane (2019): Change through data. A Data Analytics Training Program for Government Employees.
Harvard Data Science Review,1.2

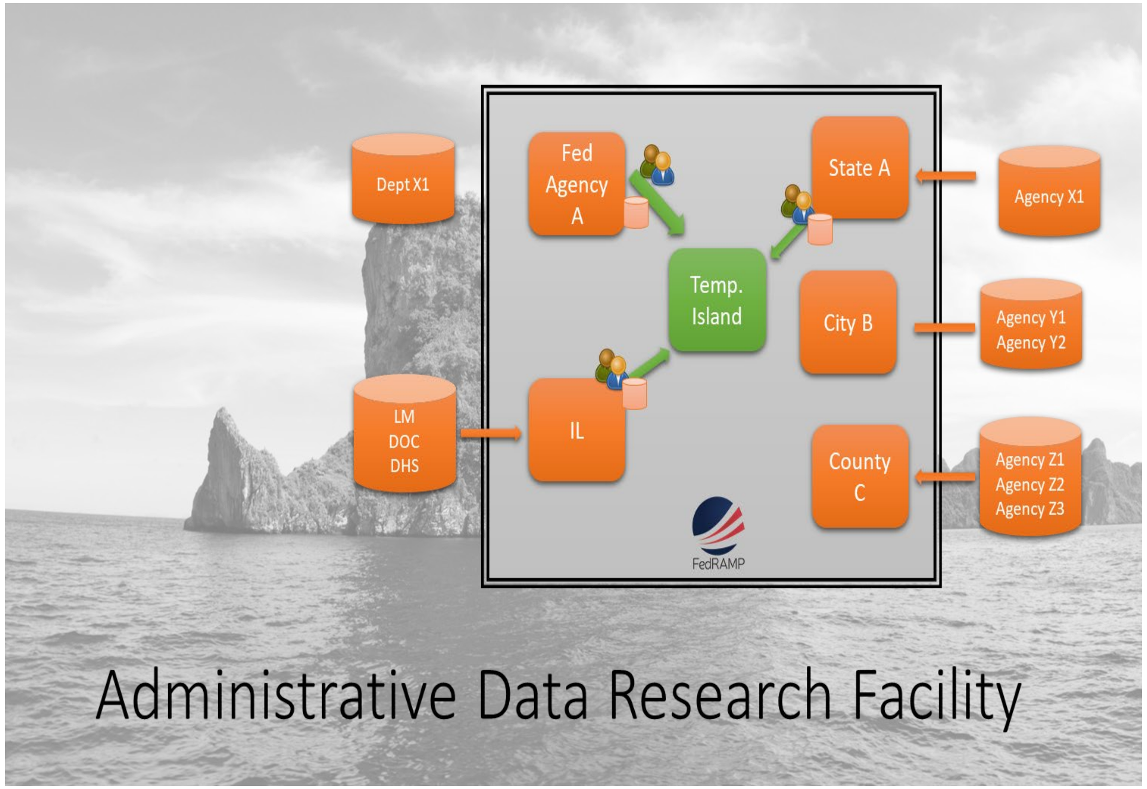


	Total Kentucky Postsecondary Graduates	Percent of Kentucky Graduates Employed	Location of Employment by Percent: Out of state	Location of Employment by Percent: Kentucky
Certificate	3,698	74%	11%	89%
Diploma	460	79%	7%	93%
Associate	6,796	78%	11%	89%
Bachelor	17,453	69%	24%	76%
Master	8,012	70%	18%	82%
Doctoral	1,731	58%	28%	72%

Enlarge Image

Kentucky Postsecondary Graduate Outcomes by Credential Level: Five years post-graduation, includes all Kentucky graduates that were employed within Indiana, Kentucky, Ohio, or Tennessee from academic year 2013

Expansion

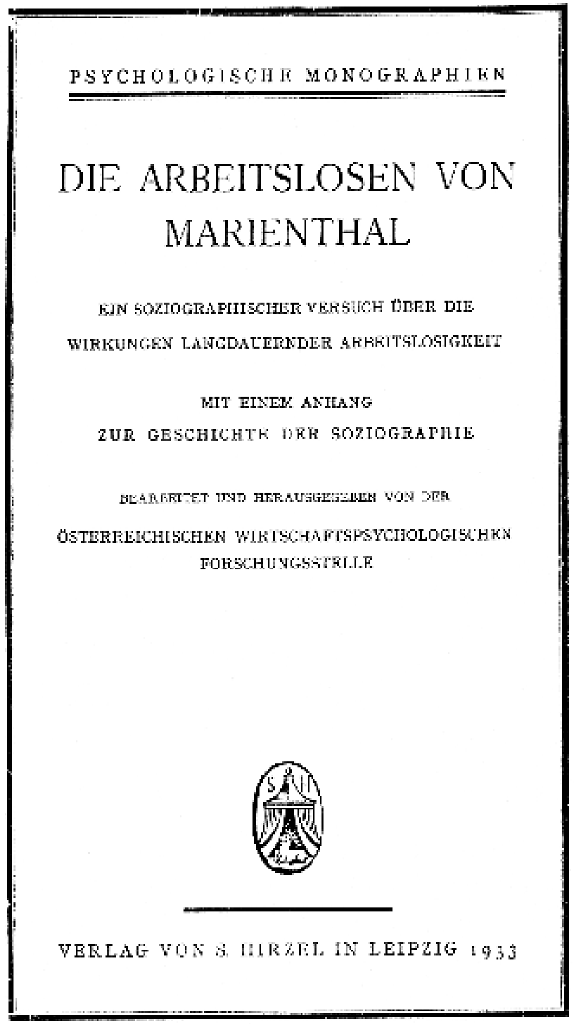

Dept X1
Fed Agency A
State A
Agency X1
Temp. Island
City B
Agency Y1
Agency Y2
LM
DOC
DHS
IL
County C
Agency Z1
Agency Z2
Agency Z3
FedRAMP
Administrative Data Research Facility

Quarter: 2016-Q4
Total Organizations: 0
Total Participants: 0
Number of Participants ● 5 ● 10 ● 15 ● 20
Number of Organizations □ 0 □ 1-5 □ 5-10 □ 10+

Lessons Learned

2. Research questions need to guide decisions on measurements and data sources.

Effects of Unemployment?



PSYCHOLOGISCHE MONOGRAPHIEN

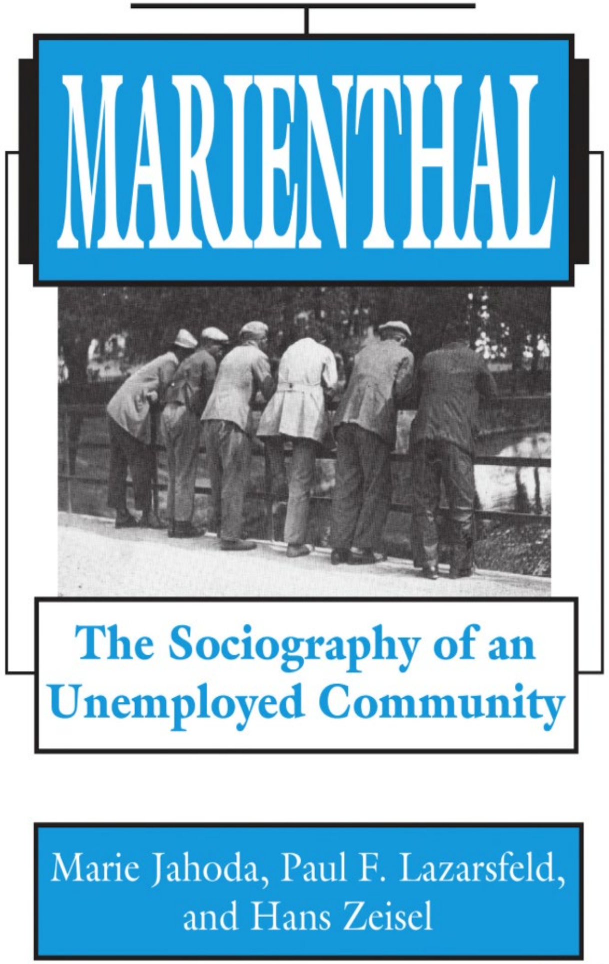
DIE ARBEITSLOSEN VON MARIENTHAL

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



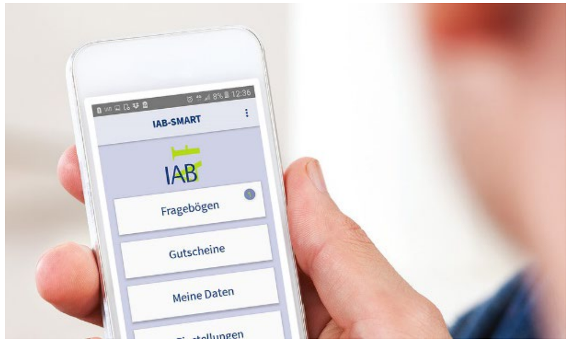
MARIENTHAL

The Sociography of an Unemployed Community

Marie Jahoda, Paul F. Lazarsfeld, and Hans Zeisel

Kreuter, Haas, Keusch, Bähr, Trappmann. 2018. “Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent.” Social Science Computer Review

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



Panel + Administrative Data as Frame

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

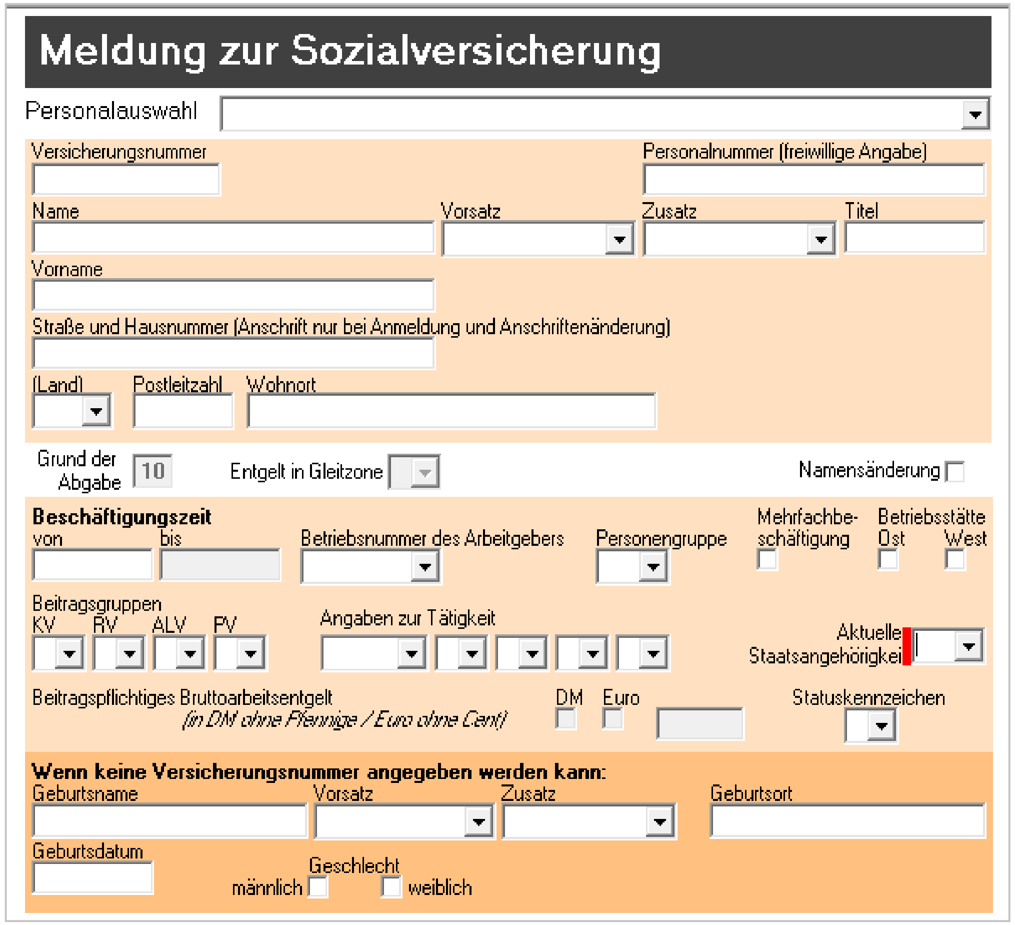
Refreshed annually

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

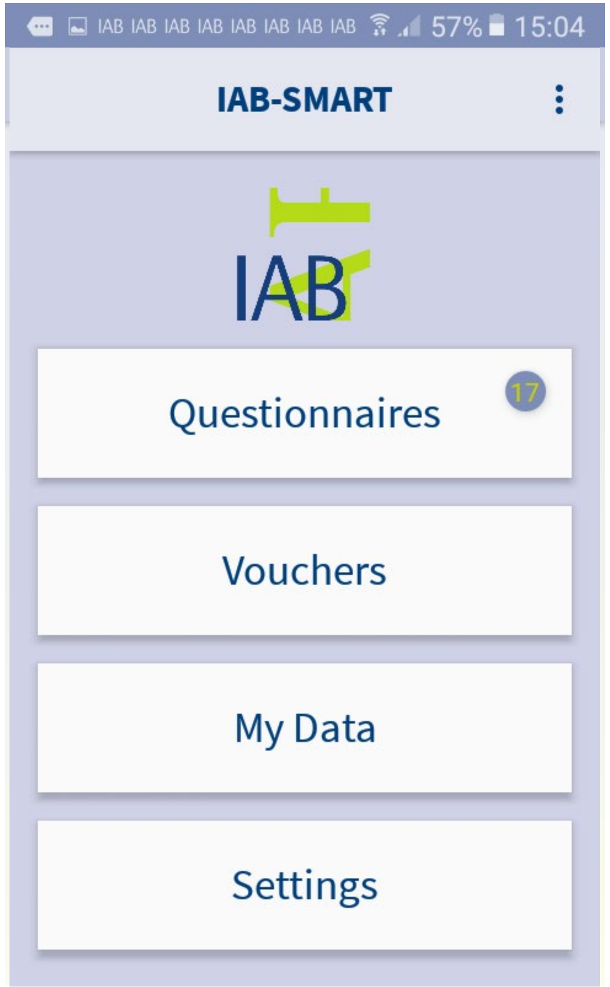
Geburtsort

Geburtsdatum

Geschlecht männlich □ □ weiblich

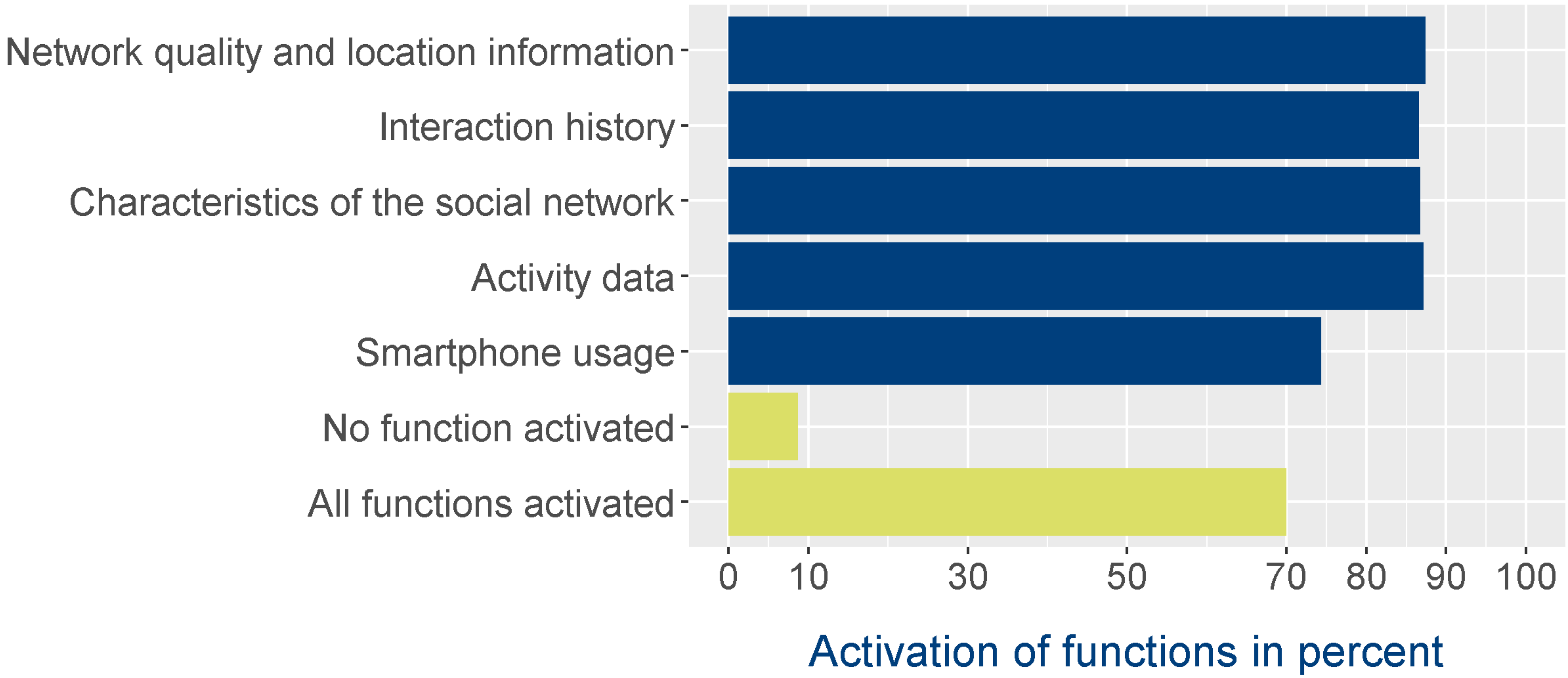
Trappmann M., Christoph B., Achatz J., Wenzig C. (2009) PASS: a new panel study for labour market research, Int. J. of Manpower , 30, 7, pp.765-770

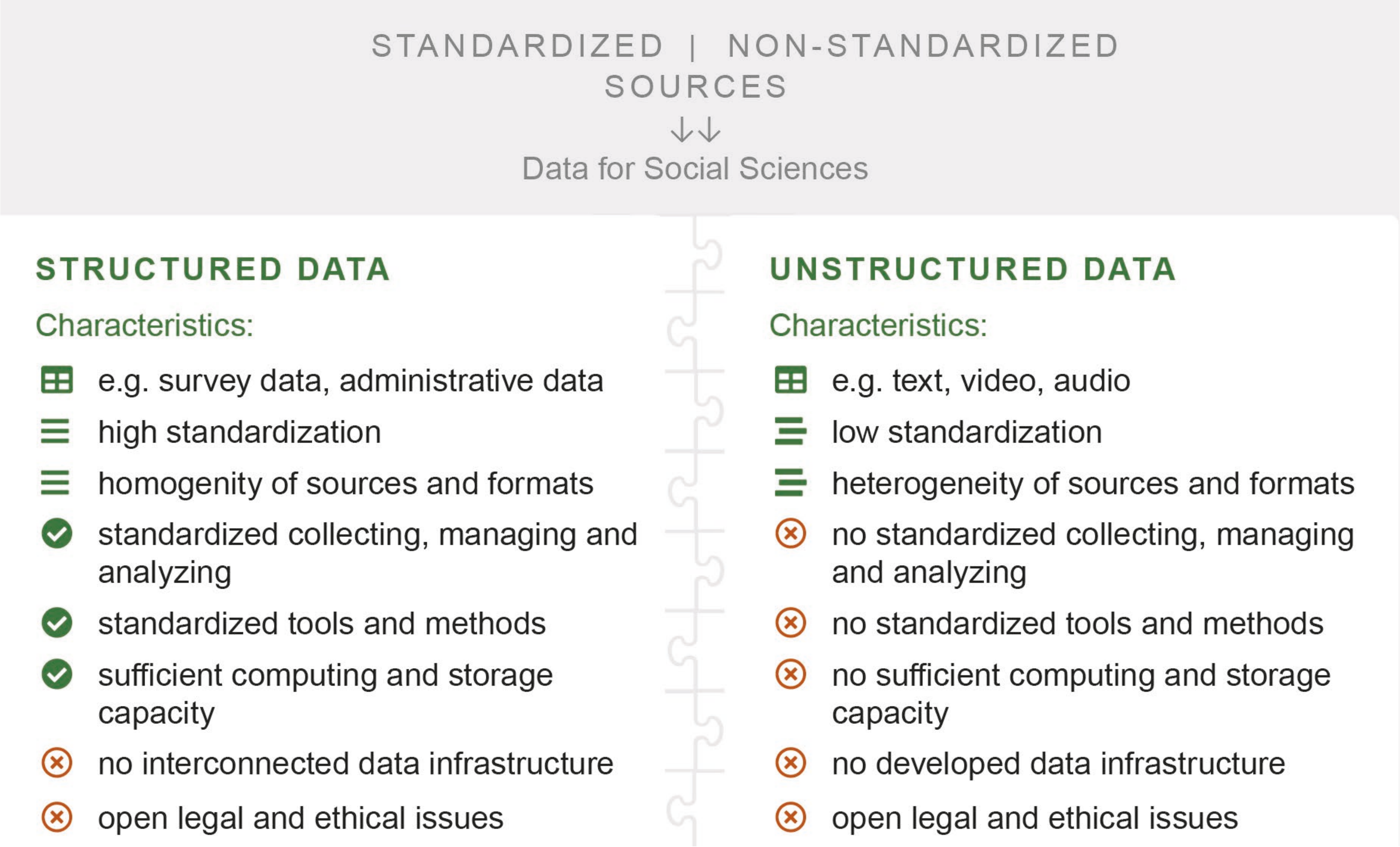
Data from Smart Phone Sensors



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

Great Uptake





STANDARDIZED | NON-STANDARDIZED SOURCES

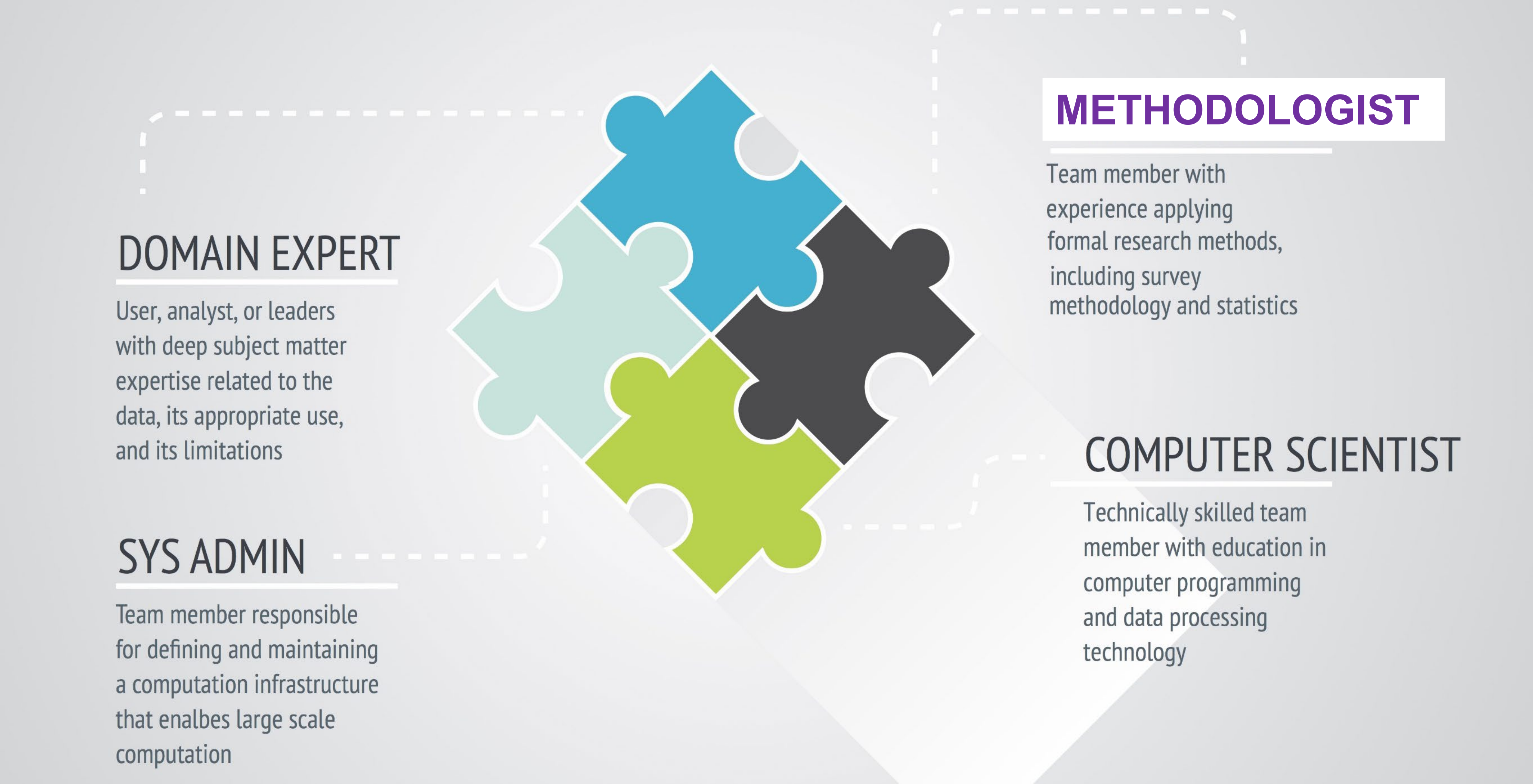
↓↓

Data for Social Sciences

STRUCTURED DATA	UNSTRUCTURED DATA
Characteristics:	Characteristics:
⊞ e.g. survey data, administrative data	⊞ e.g. text, video, audio
≡ high standardization	≡ low standardization
≡ homogenity of sources and formats	≡ heterogeneity of sources and formats
✓ standardized collecting, managing and analyzing	⊗ no standardized collecting, managing and analyzing
✓ standardized tools and methods	⊗ no standardized tools and methods
✓ sufficient computing and storage capacity	⊗ no sufficient computing and storage capacity
⊗ no interconnected data infrastructure	⊗ no developed data infrastructure
⊗ open legal and ethical issues	⊗ open legal and ethical issues

Lessons Learned

3. Data Science is a “Team Sport” … and needs to be treated as such.



METHODOLOGIST

Team member with experience applying formal research methods, including survey methodology and statistics

DOMAIN EXPERT

User, analyst, or leaders with deep subject matter expertise related to the data, its appropriate use, and its limitations

COMPUTER SCIENTIST

Technically skilled team member with education in computer programming and data processing technology

SYS ADMIN

Team member responsible for defining and maintaining a computation infrastructure that enalbes large scale computation

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.

Work in Teams

Lessons Learned

1. Creating new measures out of linked data is not enough. The effort has to be tied to a product.
2. Research questions need to guide decisions on measurements and data sources.
3. Data Science is a “Team Sport” … and needs to be treated as such.