# BERD
## @NFDI

# BERD@NFDI in a Nutshell

By F. Stahl – B. Bischl – S. Gehrlein – F. Kreuter – K. Tochtermann

July 2021

UNIVERSITÄT MANNHEIM
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM
Universität zu Köln
UHH Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
IAB INSTITUT FÜR ARBEITSMARKT- UND BERUFSFORSCHUNG Die Forschungseinrichtung der Bundesagentur für Arbeit
ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics
gesis Leibniz Institute for the Social Sciences

# Why BERD@NFDI?

**Our understanding of individual and social behavior**

**is currently significantly expanded**

**due to the availability of new data types**

# Use Case: Unemployment Research

| 1930's | 1980's | Since 2010's |
|--------|--------|--------------|



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



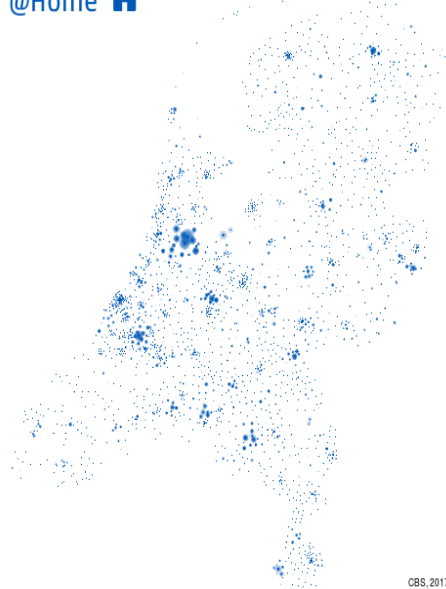Source: ISR Archive



Source: IAB SMART Study, Kreuter et al.

| 1930's | 1980's | Since 2010's |
|--------|--------|--------------|
| **+** detailed | **+** standardized | **+** standardized |
| **-** observer error | **+** large scale | **+** large scale |
| **-** small scale | **+** inference | **+** inference |
| **-** no inference | **-** expensive | **+** cheap |
| | **-** high burden | **+** low burden |
| | **-** misreports | **-** complex post-processing |
| | | **-** tools and infrastructure lacking |

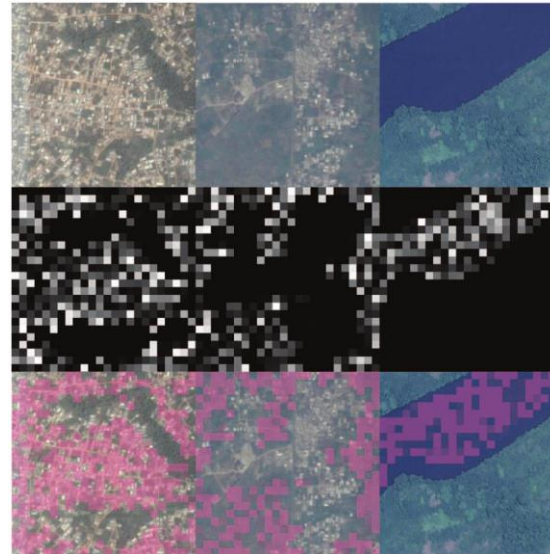# UN Sustainable Development Goals

## Sustainable Communities

@Home 🏠



CBS, 2017

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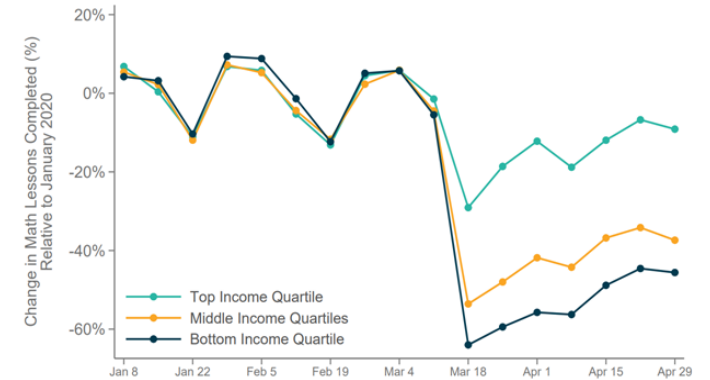towards-motives-behind-mobility

## No Poverty



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

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

## Reduced Inequalities



Chetty, R. et al. (2020): „The Economic Impacts of COVID-19: Evidence from
a New Public Database Built Using Private Sector Data.", *NBER Working Paper*,
No. 27431, 10.3386/w27431.

# Use Cases in all Disciplines

**Business**

- "From words to pixels: text and image mining methods for service research.", *Journal of Service Management,* 2019, 30(5), 593-620.
- "Understanding videos at scale: How to extract insights for business research.", *Journal of Business Research*, 2020, 123, 367-379.

**Economics**

- "Big Data as a Governance Mechanism." *The Review of Financial Studies*, 2019, 32(5), 2021-2061.
- "Seeing is be... auditor tenure, and audit fees.", *Journal of Accounting and Economics*, 2020, 69(1), 1-23.
- "Peers at Work.", *American Economic Review*, 2009, 99(4), 112-145.

**Psychology**

- "Predicting personality from patterns of behavior collected with smartphones.", *Proceedings of the National Academy of Sciences*, 2020, 1-...
- "A (More) Behavioural Science of Personality in the Age of Multi-Modal Sensing, Big Data, Machine Learning, and Artificial Intelligence.", *European Journal of Personality*, 2020, 34(5), 593-598.
- "After Aylan Kurdi: How tweeting about death, threat, and harm predict increased expressions of solidarity with refugees over time.", *Psychological Science,* 2018, 29(4), 623-634.
- "# Sad: Twitter Content Predicts Changes in Cognitive Vulnerability and Depressive Symptoms." Cognitive Therapy and Research, 2019, 43(4), 657-665.

**Communications**

- "Video content marketing: The making of clips.", *Journal of Marketing*, 2018, 82(4), 86-101.
- "Machine learning approaches to facial and text analysis: Discovering CEO oral communication styles.", *Strategic Management Journal*, 2019, 40(11), 1705-1732.
- "Video mining: Measuring visual information using automatic methods.", *International Journal of Research in Marketing*, 2019, 36(2), 216-231.

**Political Science**

- ... asure Protest Size Variation.", *American Political Science Review*, 2020, 114(4), 1343-1351.
- "Facing the electorate: Computational approaches to the study of nonverbal communication and voter impression formation." *Political Communication*, 2020, 1-23.
- ... ... of Legislators from Social Media Images." In: *Proceedings of the International AAAI Conference on Web and Social Media*, 2020, 14 (1), 726-737.
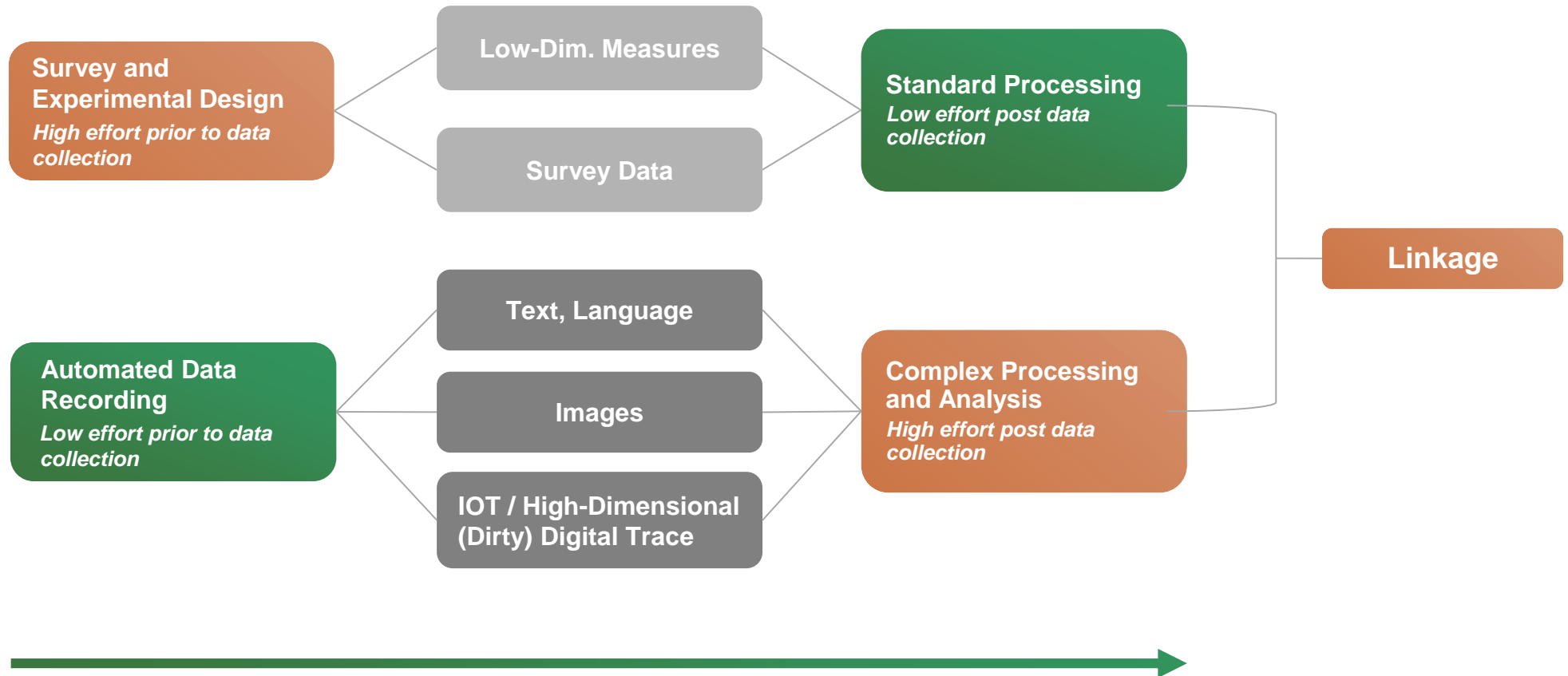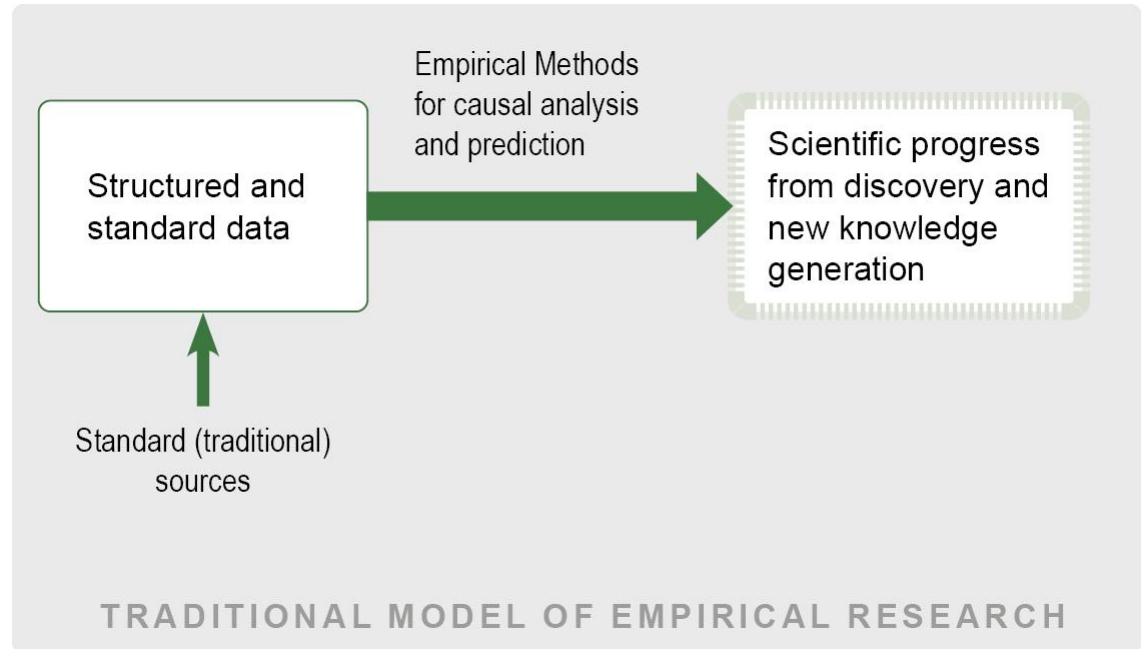
**Sociology**

- "Visual cross-platform analysis: digital methods to research social media images.", *Information, Communication & Society*, 2018, 23(2), 161-180.
- "Daytime Locations in Spatial Mismatch: Job Accessibility and Employment at Reentry From Prison", *Demography*, 2017, 54(2), 775-800.
- "Using deep learning and Google Street View to estimate the demographic makeup of neighborhoods across the United States", *PNAS*, 2017, 114(50), 13108-13113.
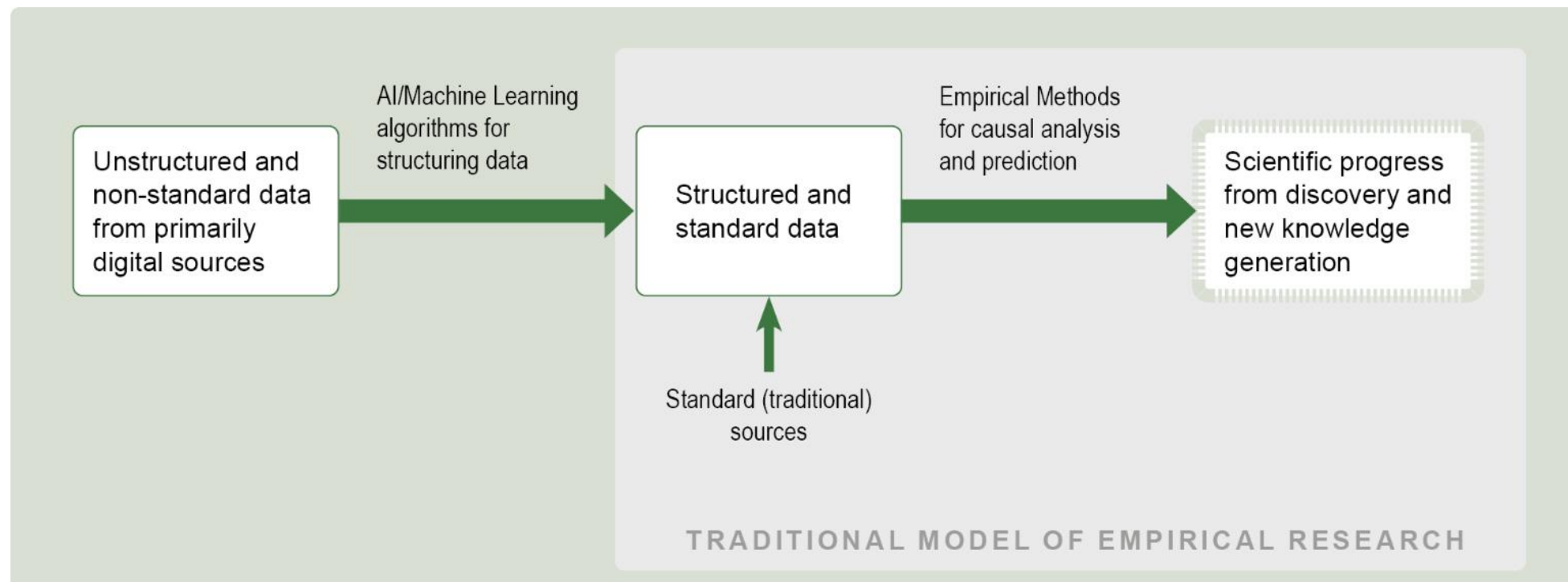
*Source: Destatis

**BERD provides infrastructure for the new enriched model of research to all social science:**

- **30% of all faculty and researchers in Germany***
- **33% of all students in Germany***

# Challenge of the New Analytical Paradigm

**Survey and Experimental Design**
*High effort prior to data collection*

Low-Dim. Measures

Survey Data

**Standard Processing**
*Low effort post data collection*

**Linkage**

**Automated Data Recording**
*Low effort prior to data collection*

Text, Language

Images

IOT / High-Dimensional (Dirty) Digital Trace

**Complex Processing and Analysis**
*High effort post data collection*

Structured and standard data

Empirical Methods for causal analysis and prediction

Scientific progress from discovery and new knowledge generation

Standard (traditional) sources

**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

- Abundant complex data and data types: Huge potential for exciting discoveries and social gains
- "Methodological" costs much higher in analysis
- Interwoven with technical burden
- Risk of misleading and irreproducible results

# Why BERD@NFDI?

**BERD@NFDI offers infrastructures to challenges**

**of the expanded empirical research**

# Hidden Technical Depth of Machine Learning
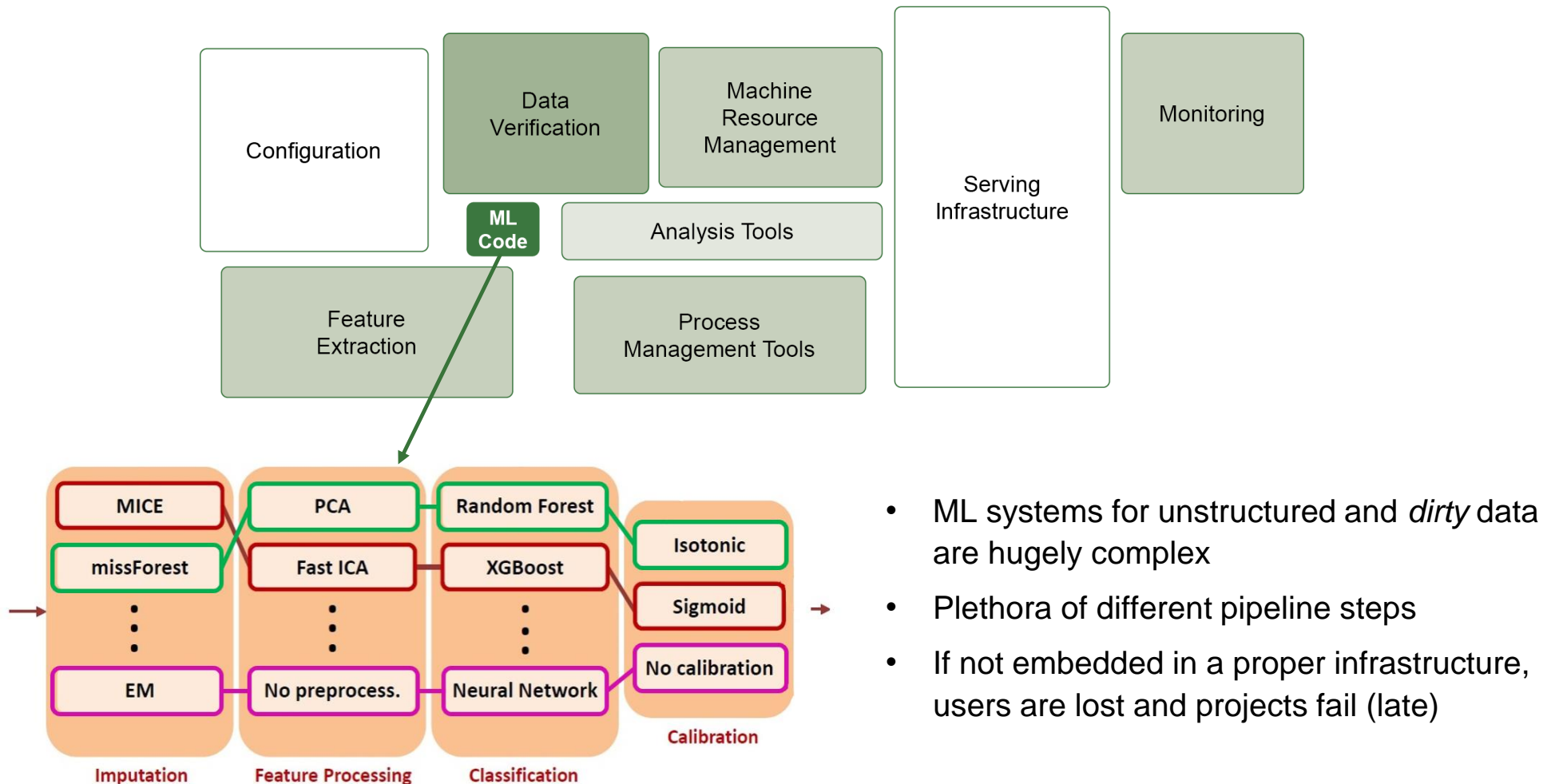
- ML systems for unstructured and *dirty* data are hugely complex

- Plethora of different pipeline steps

- If not embedded in a proper infrastructure, users are lost and projects fail (late)

Source: Sculley, D. et al. (2015): "Hidden technical debt in Machine learning systems", in: NIPS'15: Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 2, 2503-2511, https://dl.acm.org/doi/10.5555/2969442.2969519.

# BERD Building on OpenML
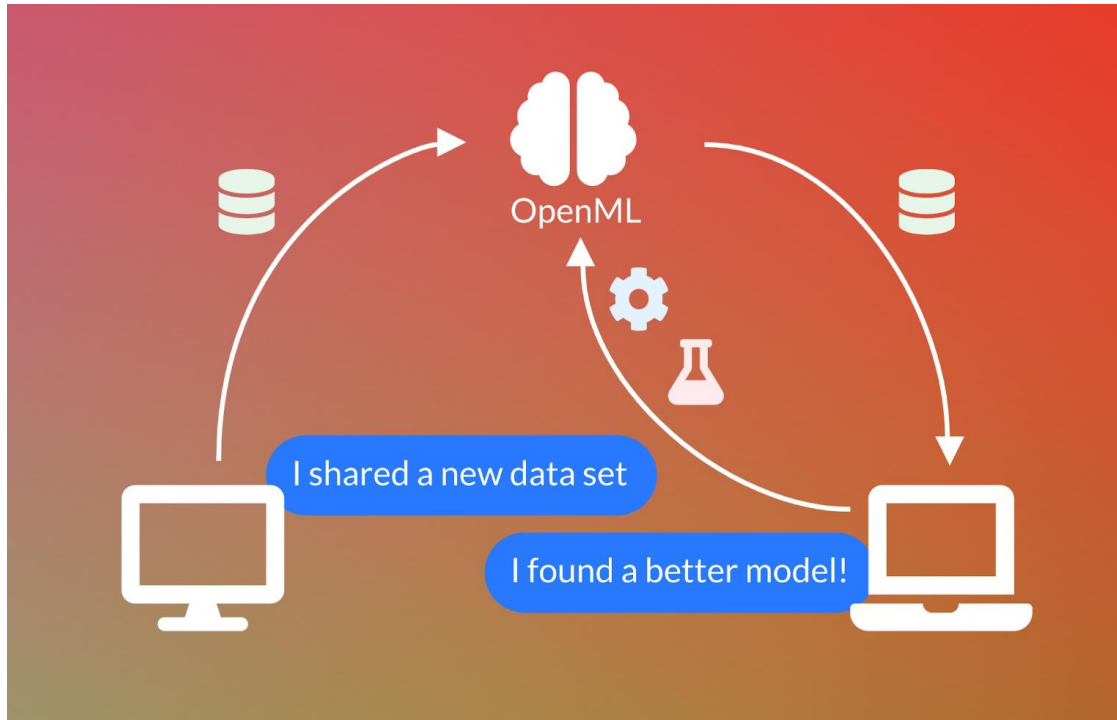
- All objects linked and searchable: data, algorithms, scripts, results
- Many major ML toolkits integrated
- Programming language agnostic
- Fully reproducible

# BERD as an Open Platform for Analysis
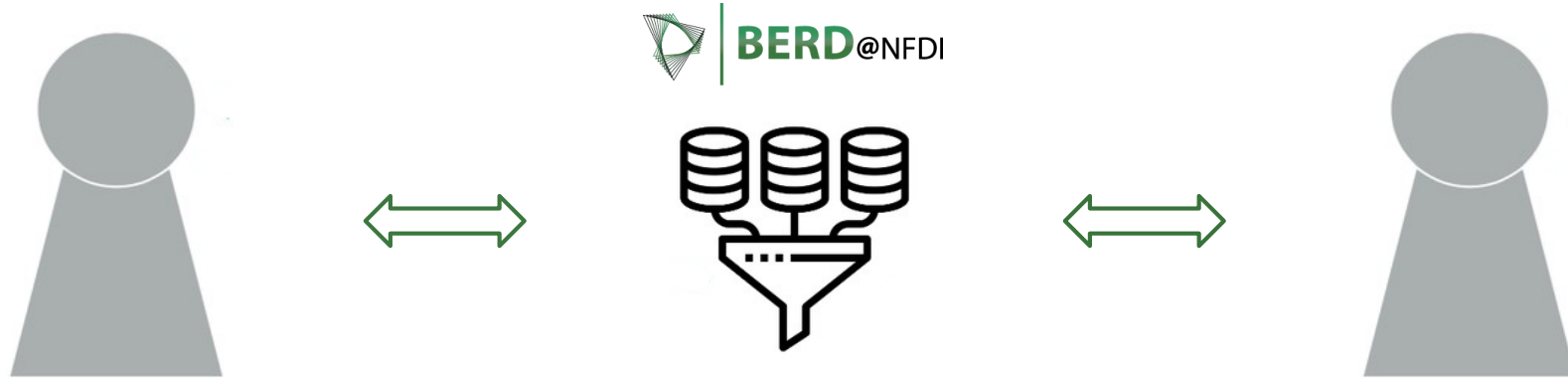
**Domain Specialist**

- Define task in accordance with theory
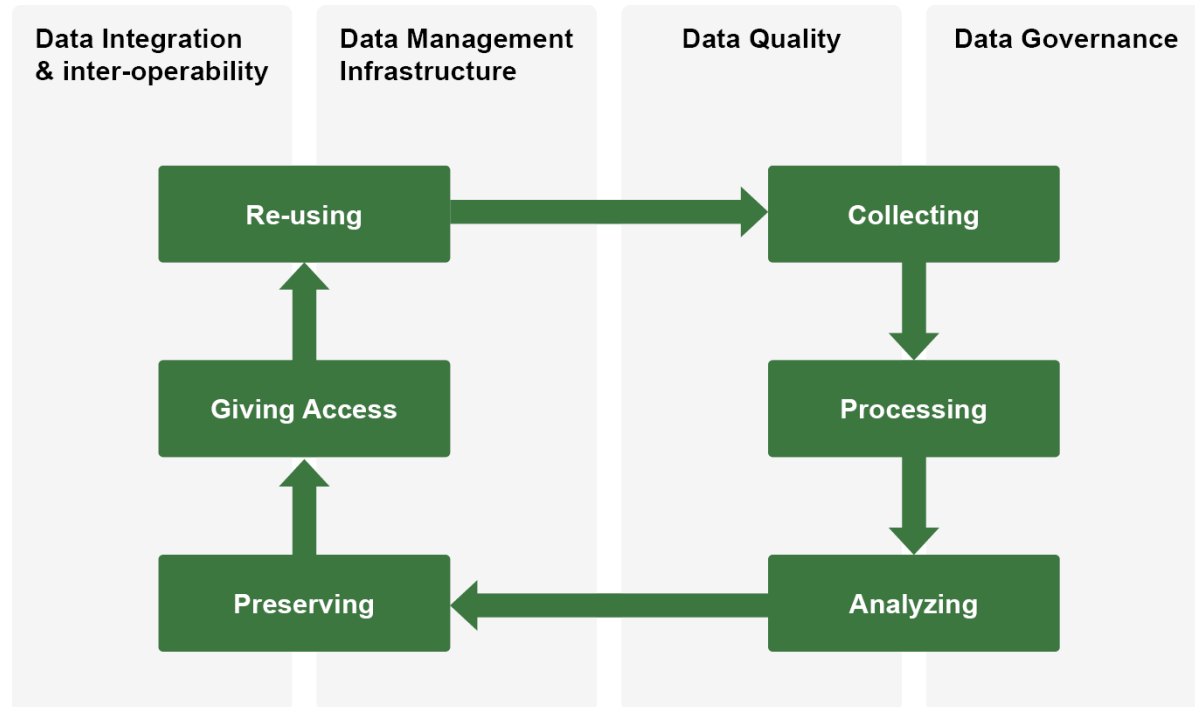- Refine theory based on results

**BERD**

- Interactive notebooks
- Similarity search on studies
- AutoML removes drudge work
- ...

**Data Scientist**

- Map task to analysis
- Refine and optimize analysis pipelines

BERD facilitates optimal collaboration between domain specialists and data scientists

# BERD is an Important Missing Piece of the NFDI

| Data Integration & inter-operability | Data Management Infrastructure | Data Quality | Data Governance |
| --- | --- | --- | --- |

Re-using → Collecting

Collecting ↓ Processing

Processing ↓ Analyzing

Analyzing → Preserving

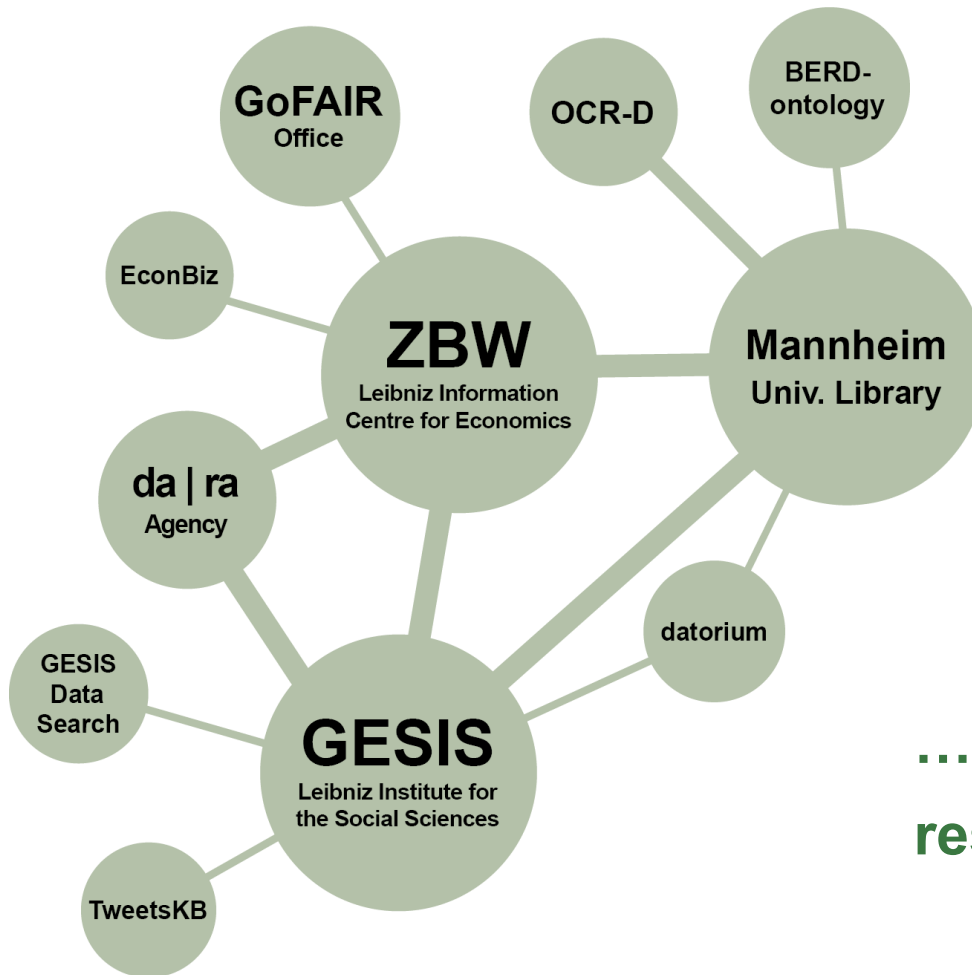Preserving ↑ Giving Access

Giving Access ↑ Re-using

- **Open**
  Linked unstructured and structured data

- **Fast and accessible computation**
  By cloud-based HPC solution

- **Best practices in ML**
  Platform provides guidance on methods

- **Reproducible and Transparent**
  Documented used data and methods

- **Management of the entire data life cycle**
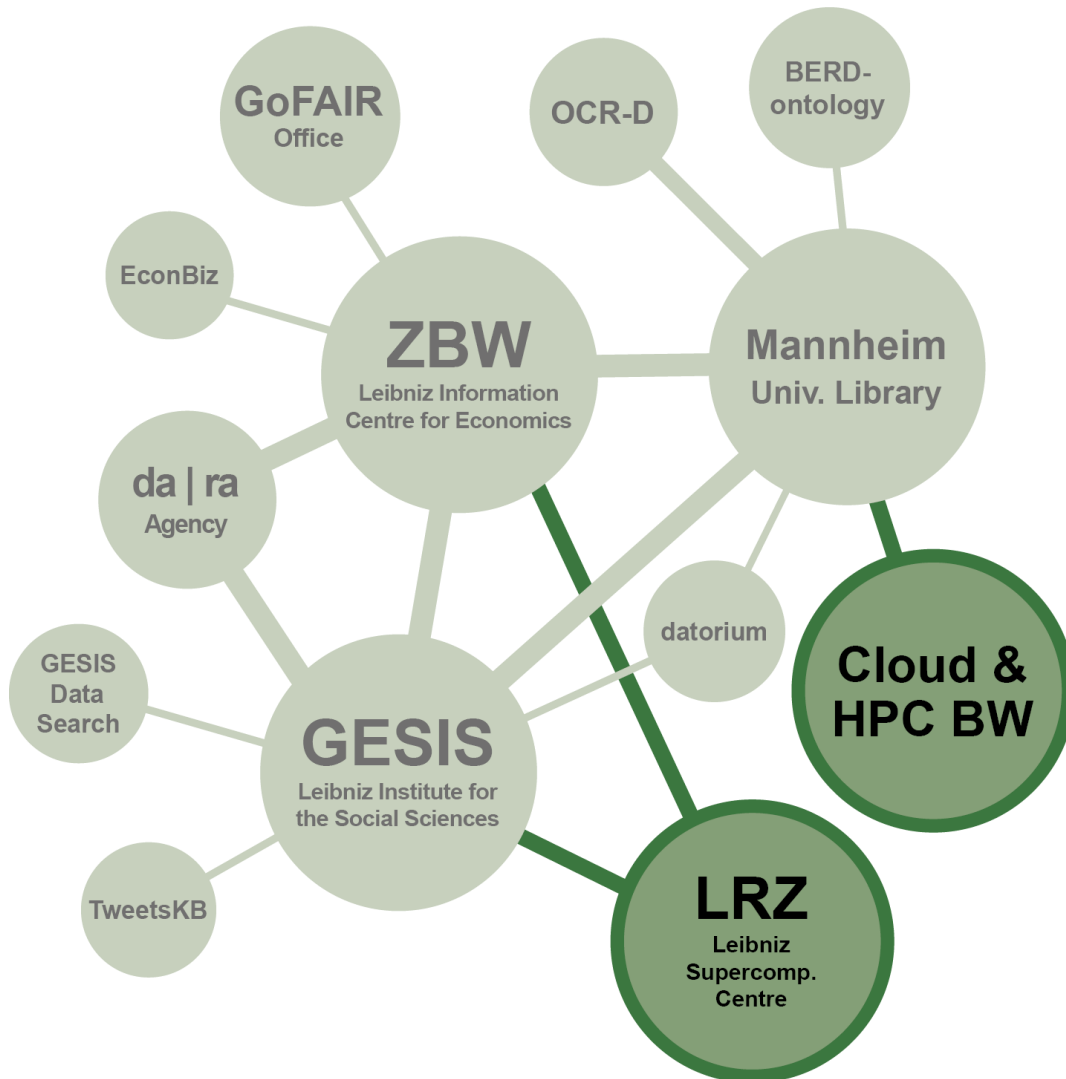
## Paradigm shift

from individual analysis and data silos to data and ML on one integrated platform

**BERD@NFDI builds on strong infrastructure**

**and long-standing partnerships**

GoFAIR
Office

OCR-D

BERD-
ontology

EconBiz

**ZBW**
Leibniz Information
Centre for Economics

**Mannheim**
Univ. Library

**da | ra**
Agency

datorium

GESIS
Data
Search

**GESIS**
Leibniz Institute for
the Social Sciences

TweetsKB

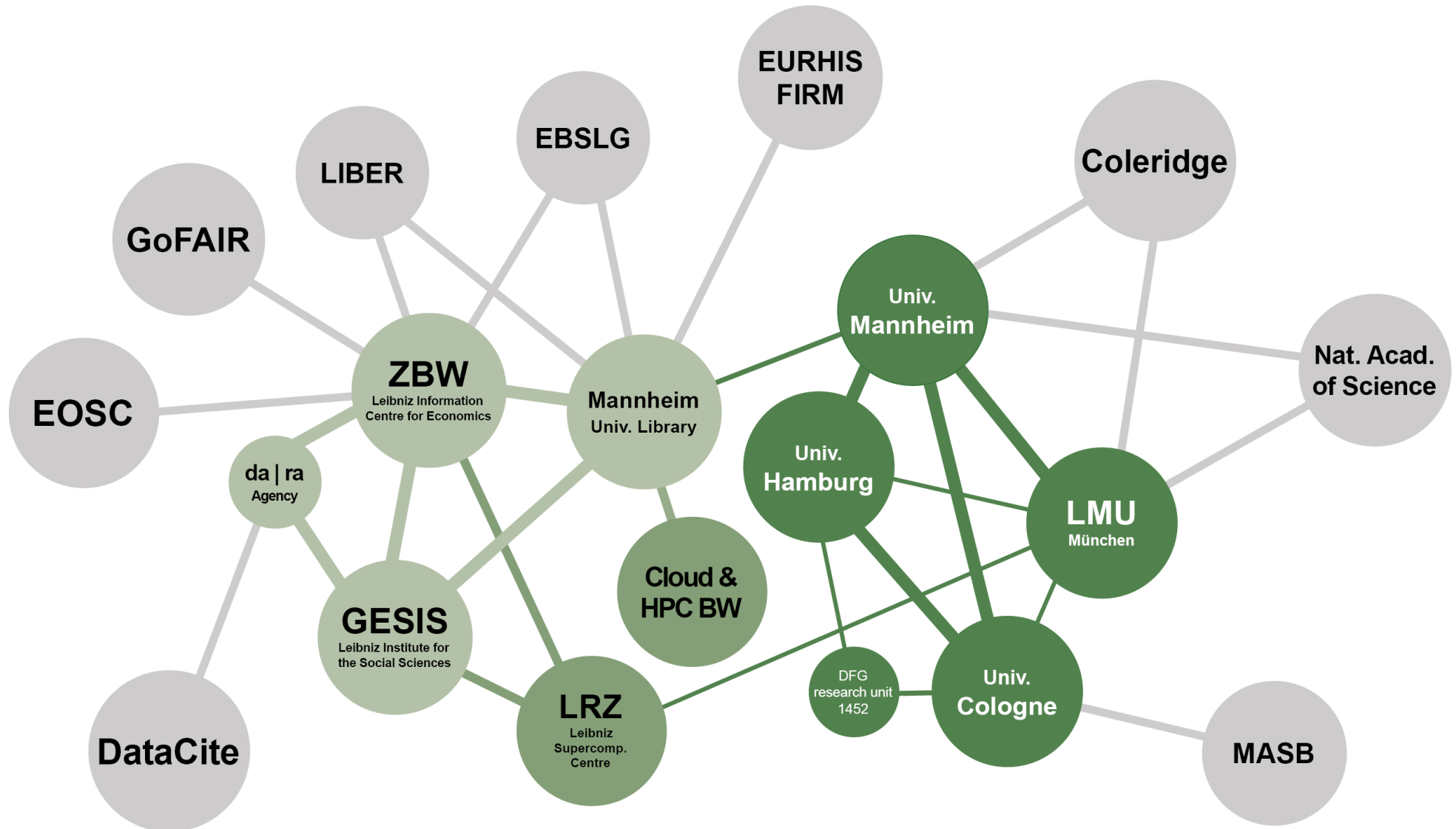**… develop and maintain nationwide research data services permanently!**

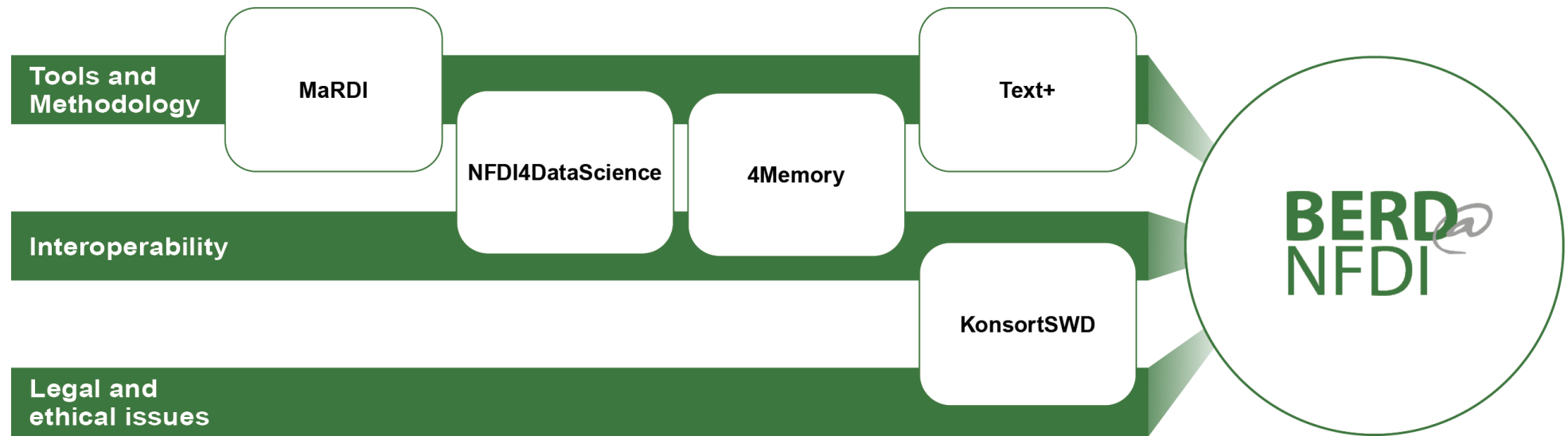… we provide the computing and storage power to deliver the BERD services!
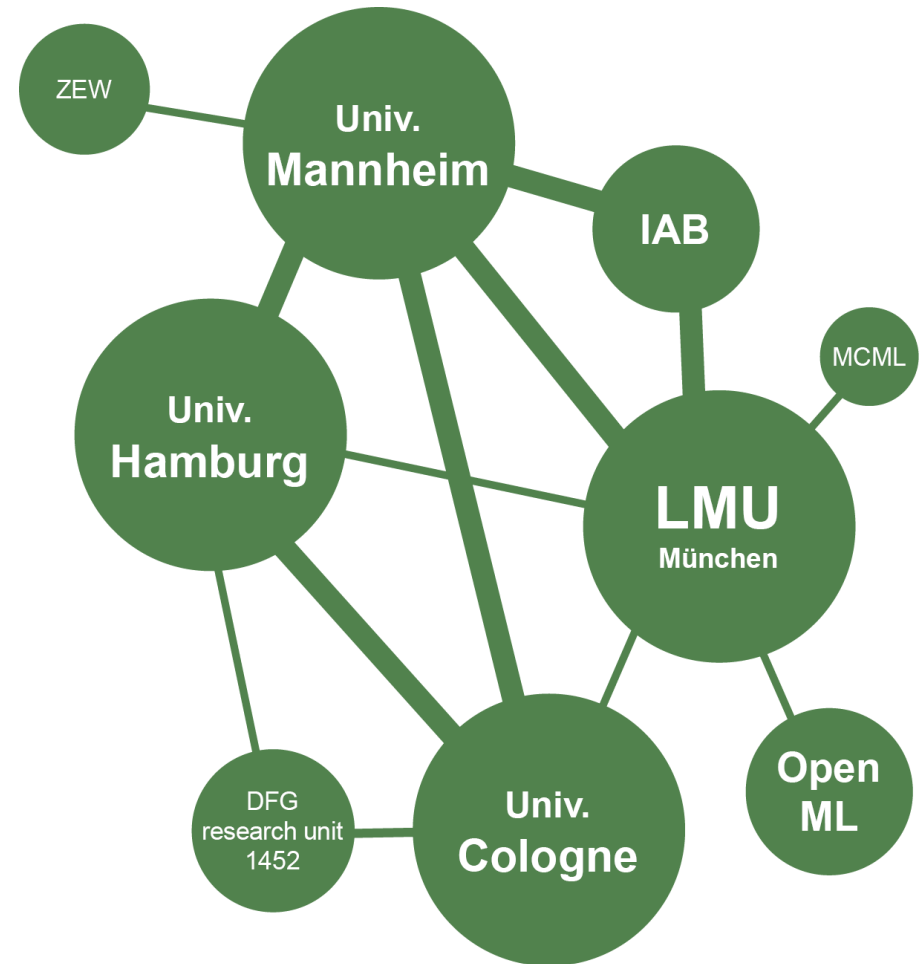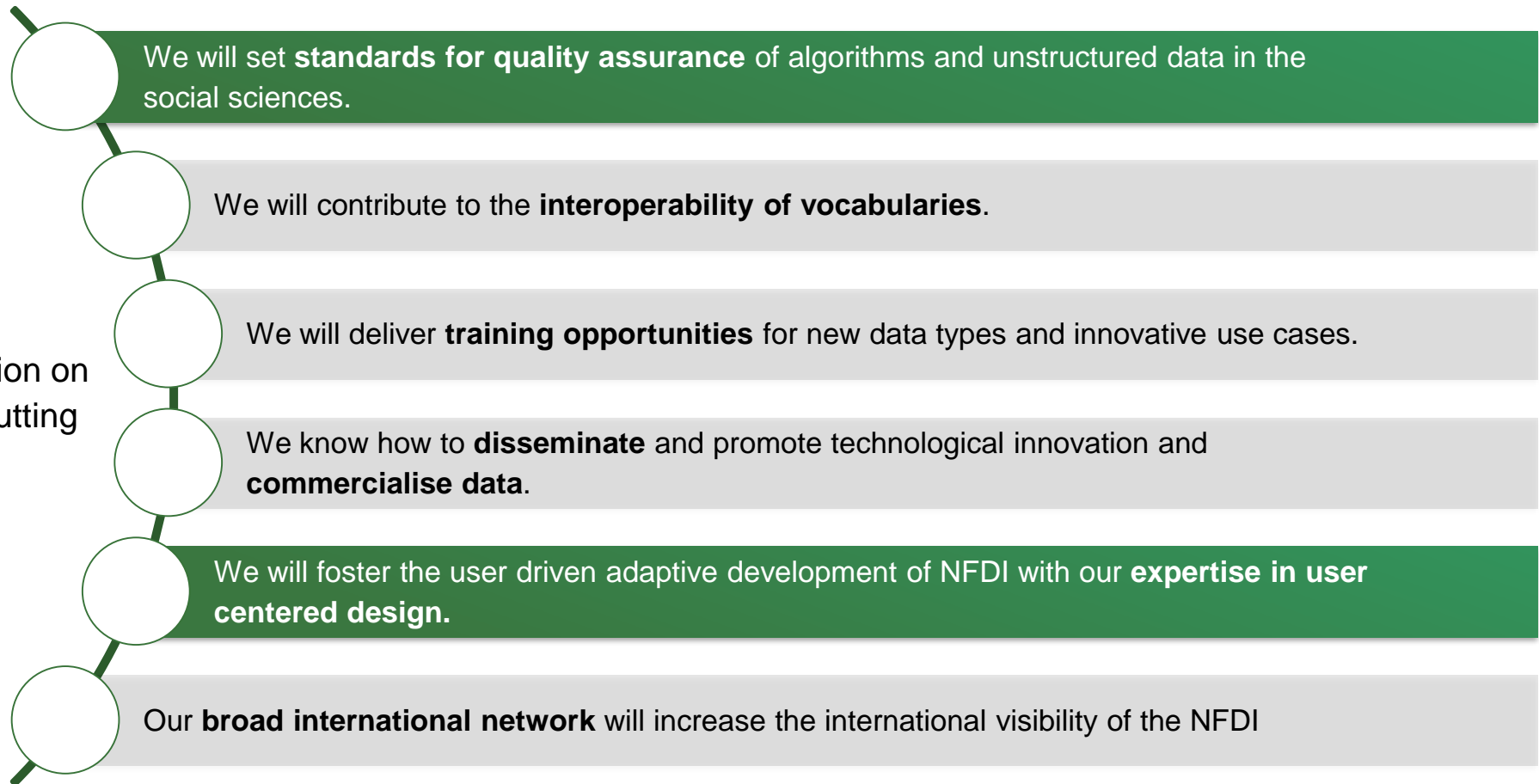
# The BERD Consortium - a Unique Combination
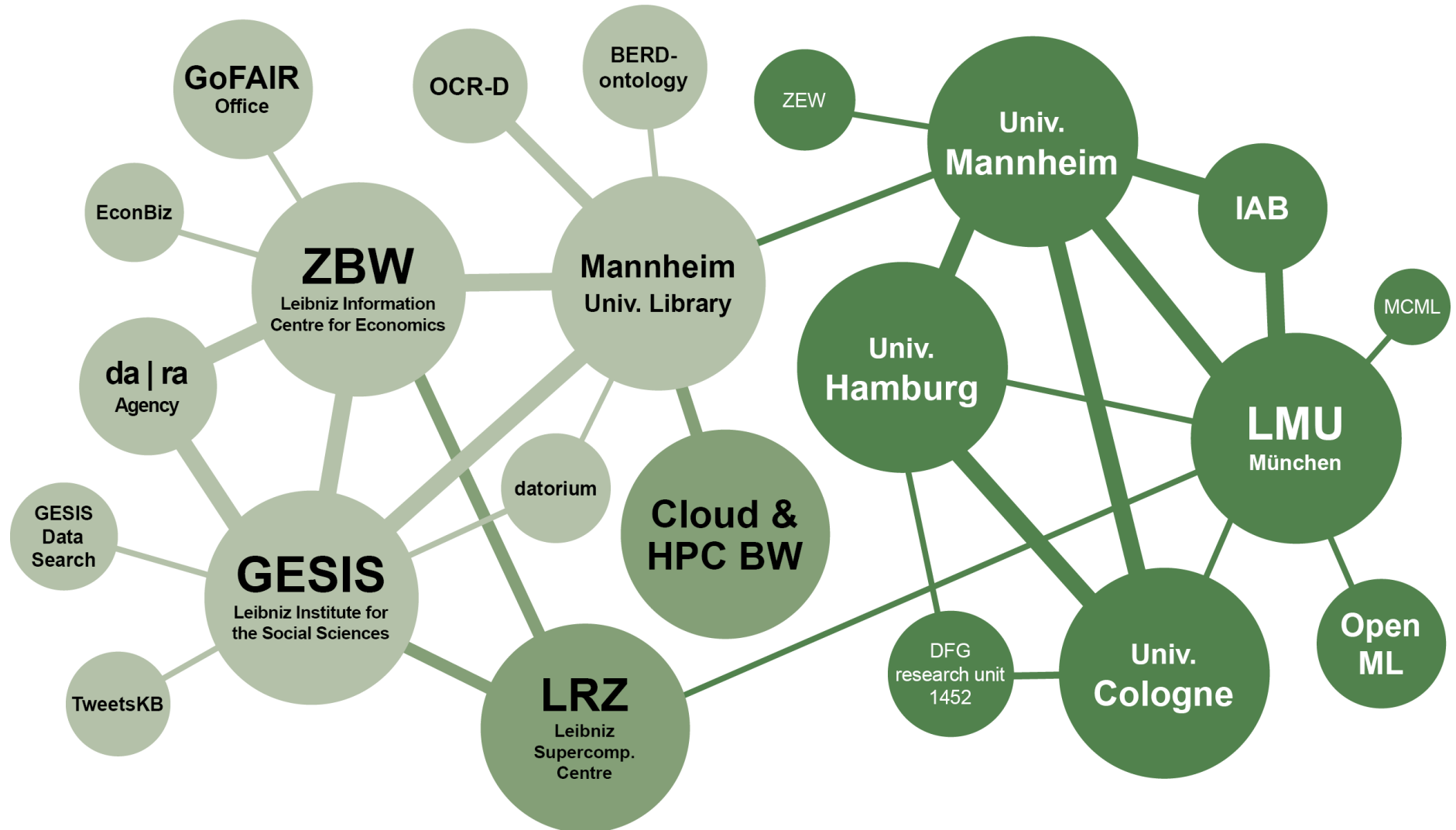
# Why BERD@NFDI?

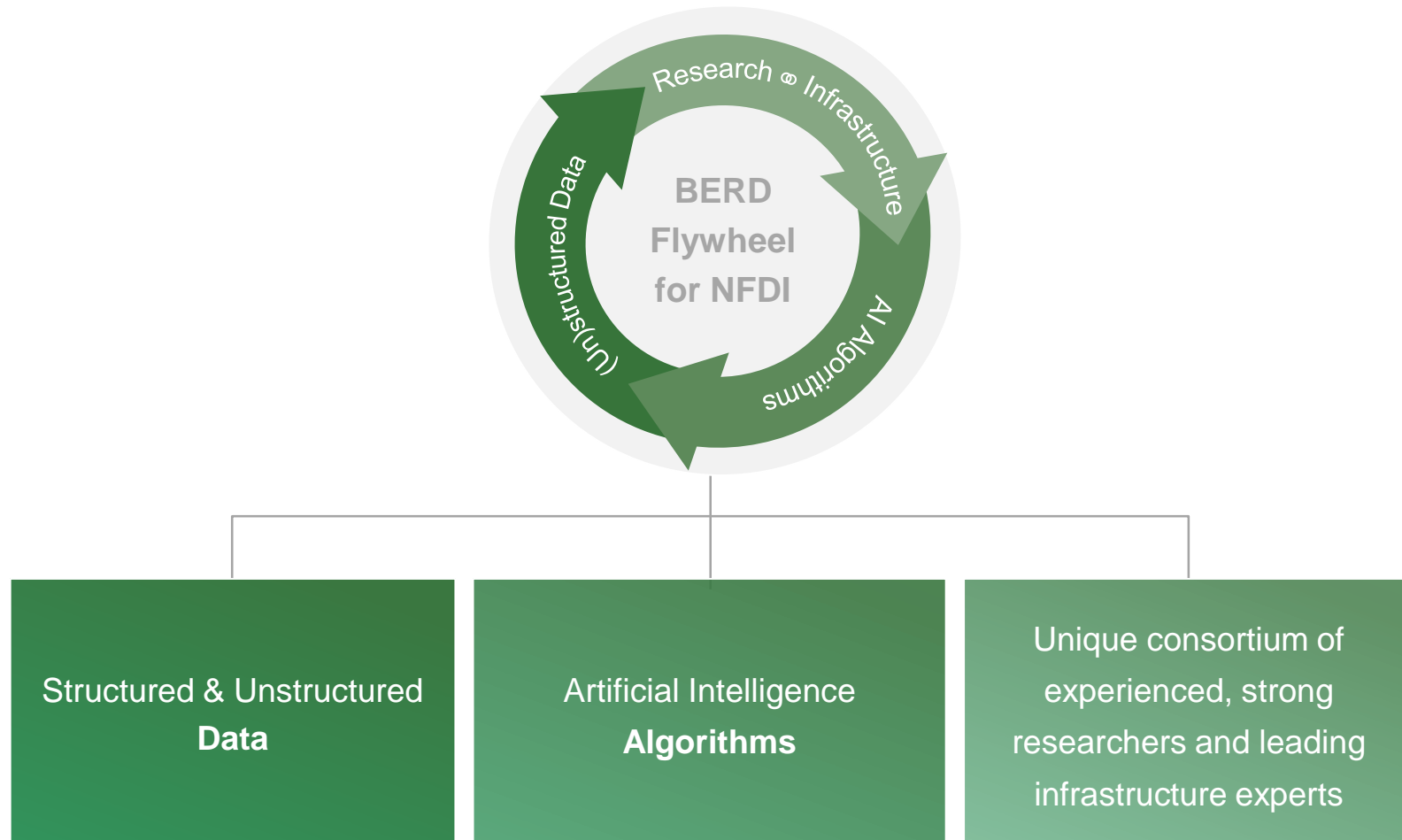BERD@NFDI makes an important and
unique contribution to the NFDI

# BERD in the NFDI

## We will cooperate closely with related consortia

# BERD in the NFDI

## We make important contributions to the cross-cutting topics of the NFDI

**Berlin Declaration on Cross-Cutting Topics**

- We will set **standards for quality assurance** of algorithms and unstructured data in the social sciences.

- We will contribute to the **interoperability of vocabularies**.

- We will deliver **training opportunities** for new data types and innovative use cases.

- We know how to **disseminate** and promote technological innovation and **commercialise data**.

- We will foster the user driven adaptive development of NFDI with our **expertise in user centered design.**

- Our **broad international network** will increase the international visibility of the NFDI

# BERD@NFDI in a Nutshell



BERD Flywheel for NFDI

Research ⚭ Infrastructure
AI Algorithms
(Un)structured Data

Structured & Unstructured **Data**

Artificial Intelligence **Algorithms**

Unique consortium of experienced, strong researchers and leading infrastructure experts