BERD
@NFDI

Structuring unstructured data for business, economic and related research

Anja Busch[1], Ulrich Krieger[2], Atif Latif[1], Fidan Limani[1], Irene Schumm[2], Ahmed Saleh[1]

1: ZBW – Leibniz Information Centre for Economics; 2: University Library, University of Mannheim

July 2021

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

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

- + detailed
- - observer error
- - small scale
- - no inference

1980’s

Source: ISR Archive

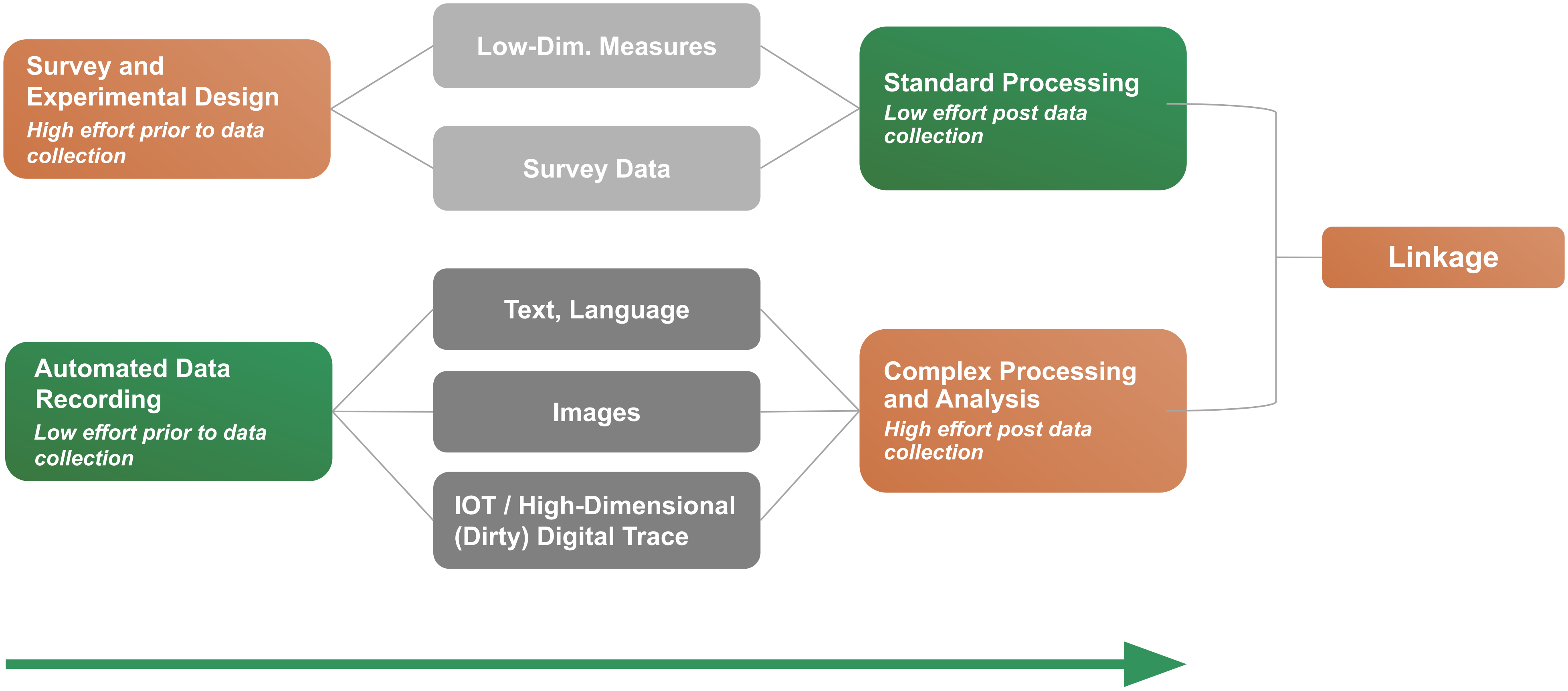
- + standardized
- + large scale
- + inference
- - expensive
- - high burden
- - misreports

Since 2010’s

Source: IAB SMART Study, Kreuter et al.

- + standardized
- + large scale
- + inference
- + cheap
- + low burden
- - complex post-processing
- - tools and infrastructure lacking

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

Text, Language

Images

IOT / High-Dimensional (Dirty) Digital Trace

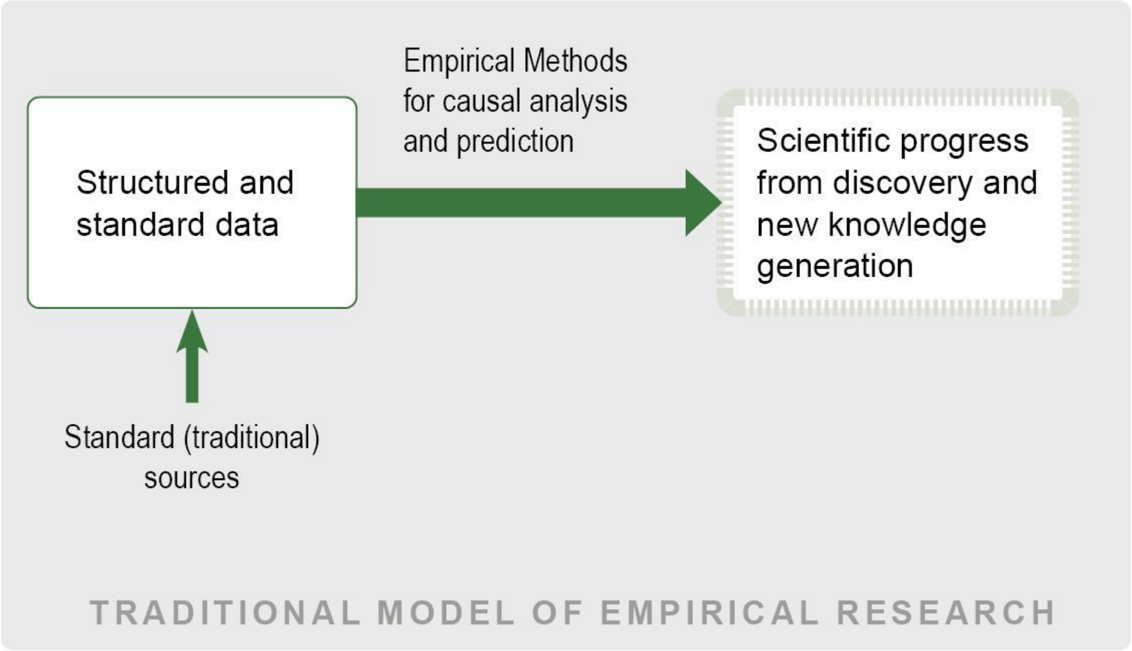
Automated Data Recording

Low effort prior to data collection

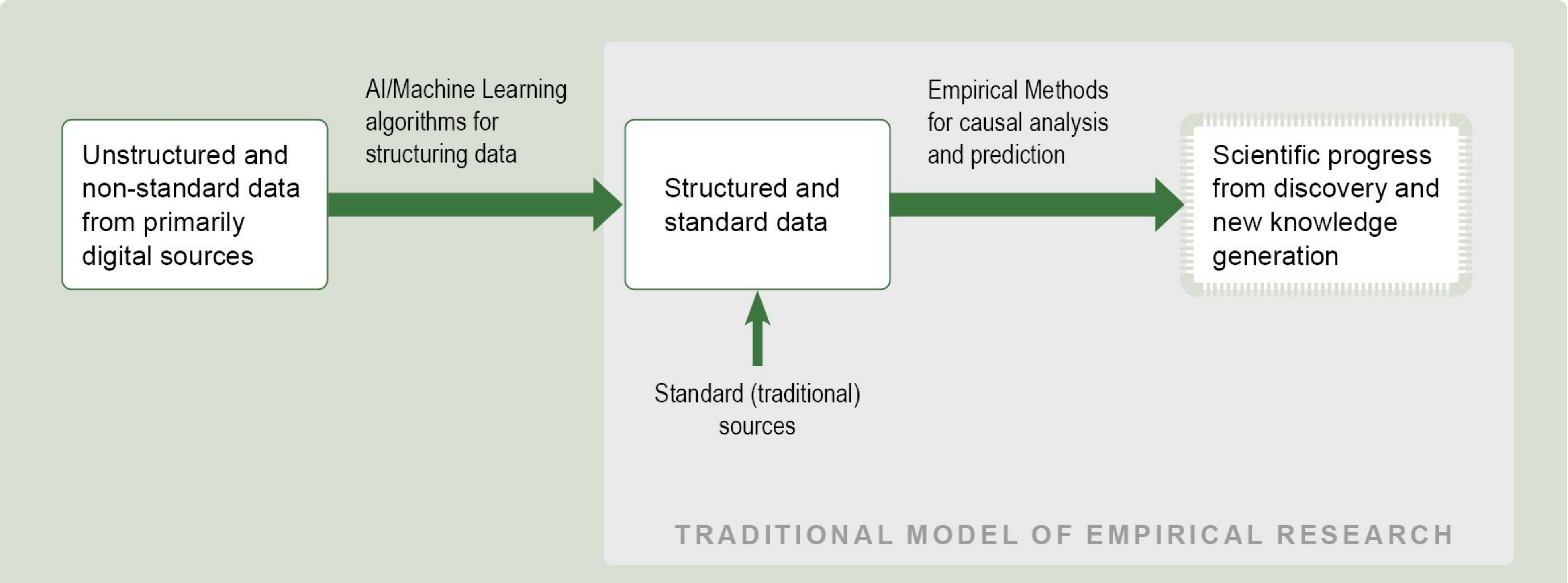
Complex Processing and Analysis

High effort post data collection

New Enriched Model of Social Science Research ...

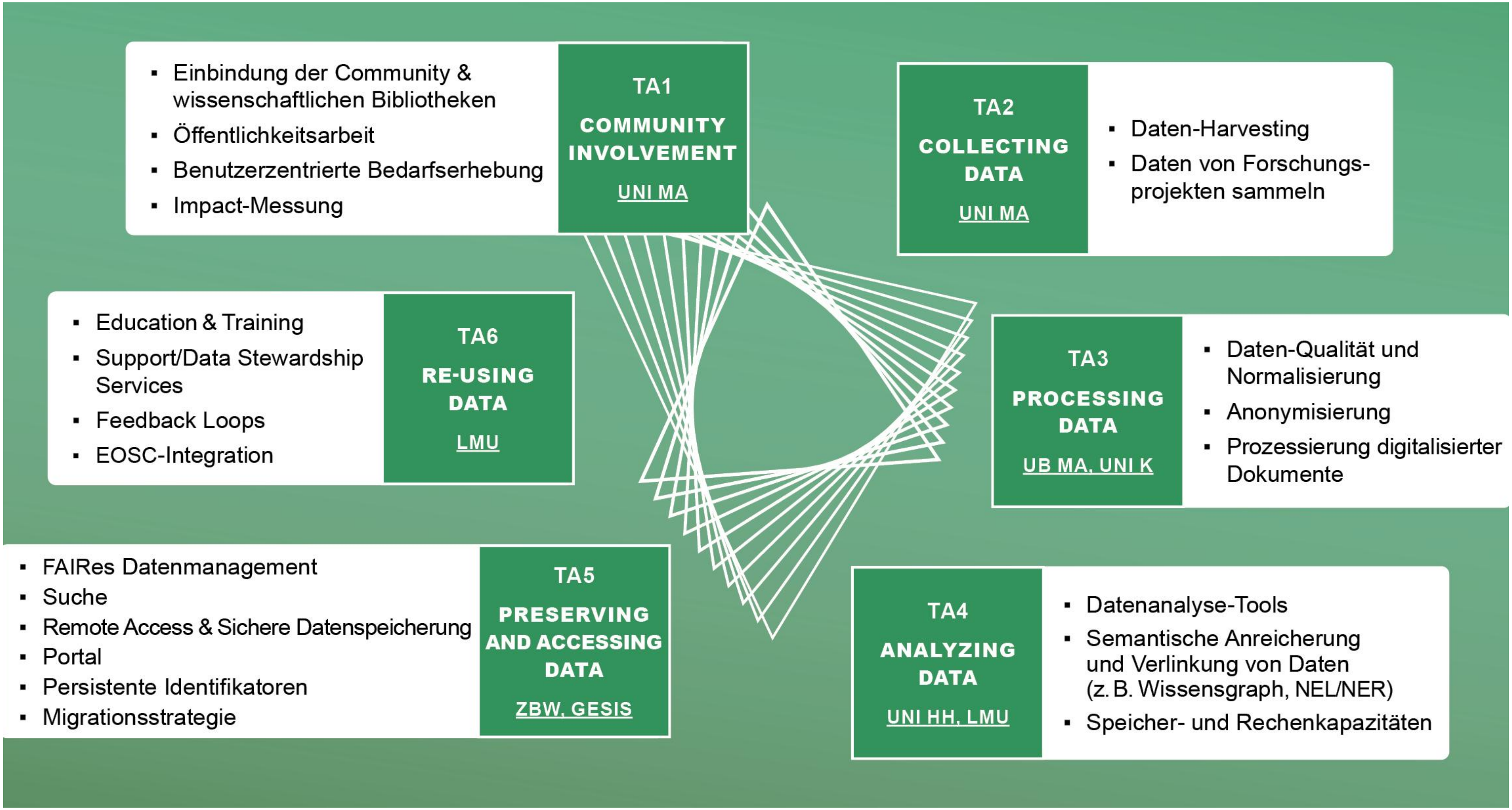

Empirical Methods for causal analysis and prediction
Structured and standard data
Scientific progress from discovery and new knowledge generation
Standard (traditional) sources
TRADITIONAL MODEL OF EMPIRICAL RESEARCH

New Enriched Model of Social Science Research ...


Unstructured and non-standard data from primarily digital sources
AI/Machine Learning algorithms for structuring data
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

- 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

BERD@NFDI Structure



TA1

COMMUNITY INVOLVEMENT

UNI MA

- Einbindung der Community & wissenschaftlichen Bibliotheken
- Öffentlichkeitsarbeit
- Benutzerzentrierte Bedarfserhebung
- Impact-Messung

TA2

COLLECTING DATA

UNI MA

- Daten-Harvesting
- Daten von Forschungsprojekten sammeln

TA3

PROCESSING DATA

UB MA, UNI K

- Daten-Qualität und Normalisierung
- Anonymisierung
- Prozessierung digitalisierter Dokumente

TA4

ANALYZING DATA

UNI HH, LMU

- Datenanalyse-Tools
- Semantische Anreicherung und Verlinkung von Daten (z. B. Wissensgraph, NEL/NER)
- Speicher- und Rechenkapazitäten

TA5

PRESERVING AND ACCESSING DATA

ZBW, GESIS

- FAIRes Datenmanagement
- Suche
- Remote Access & Sichere Datenspeicherung
- Portal
- Persistente Identifikatoren
- Migrationsstrategie

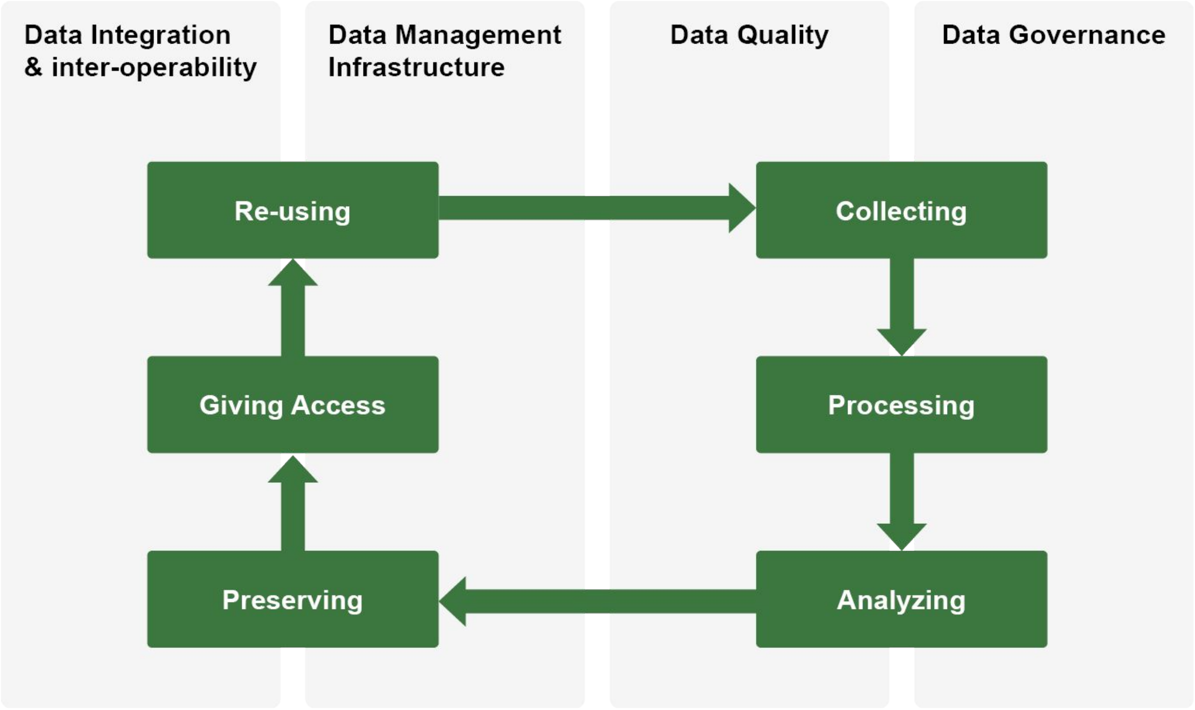
TA6

RE-USING DATA

LMU

- Education & Training
- Support/Data Stewardship Services
- Feedback Loops
- EOSC-Integration

BERD is an Important Missing Piece of the NFDI

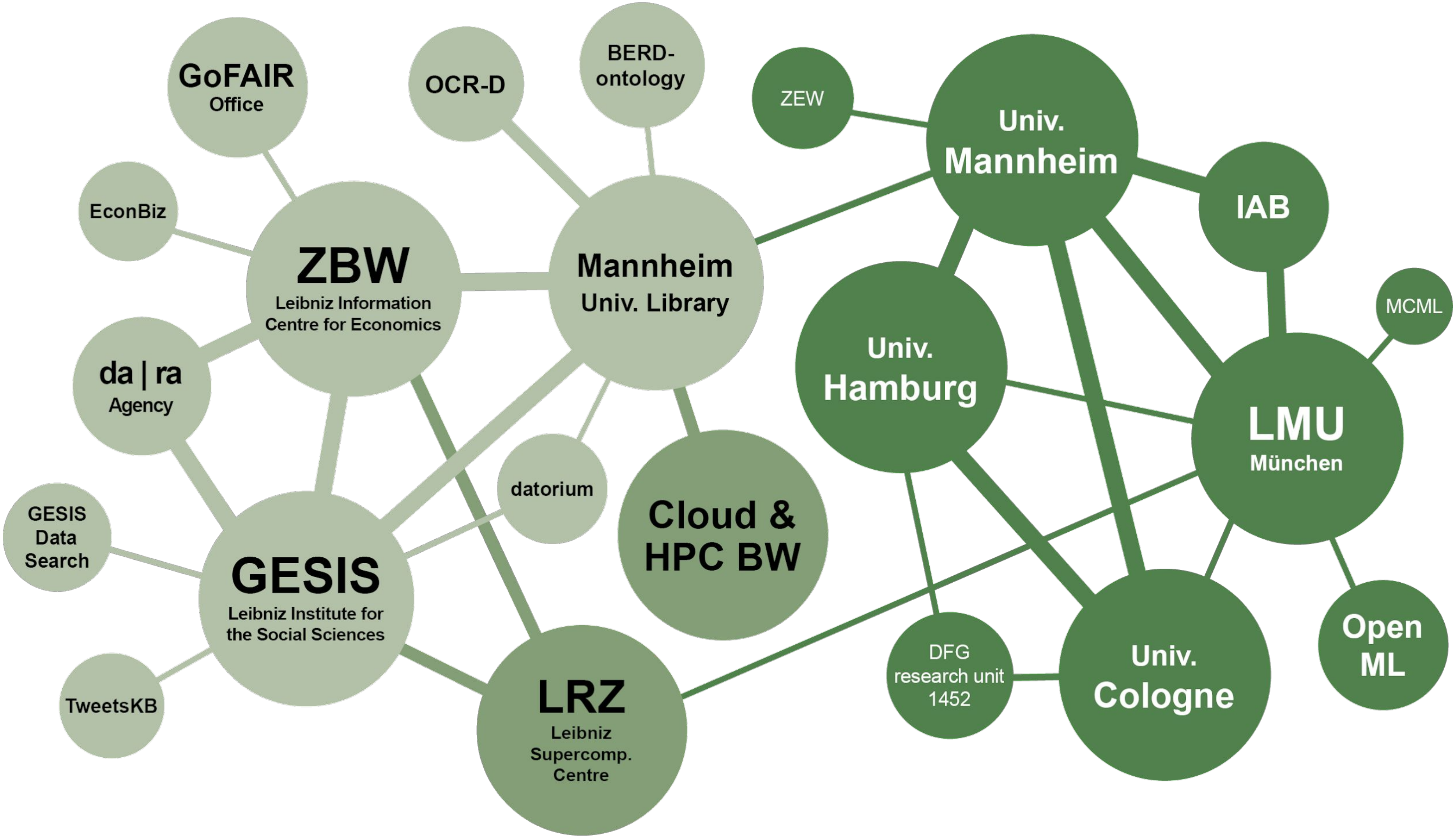


- **Open**
  Linked unstructured and structured data
- **Fast and accessible computation**
  By cloud-based HPC solution
- **Best practices in ML**
  Platform provides guidance on methods
- **Reproducible and Transparent**
  Documented used data and methods
- **Management of the entire data life cycle**

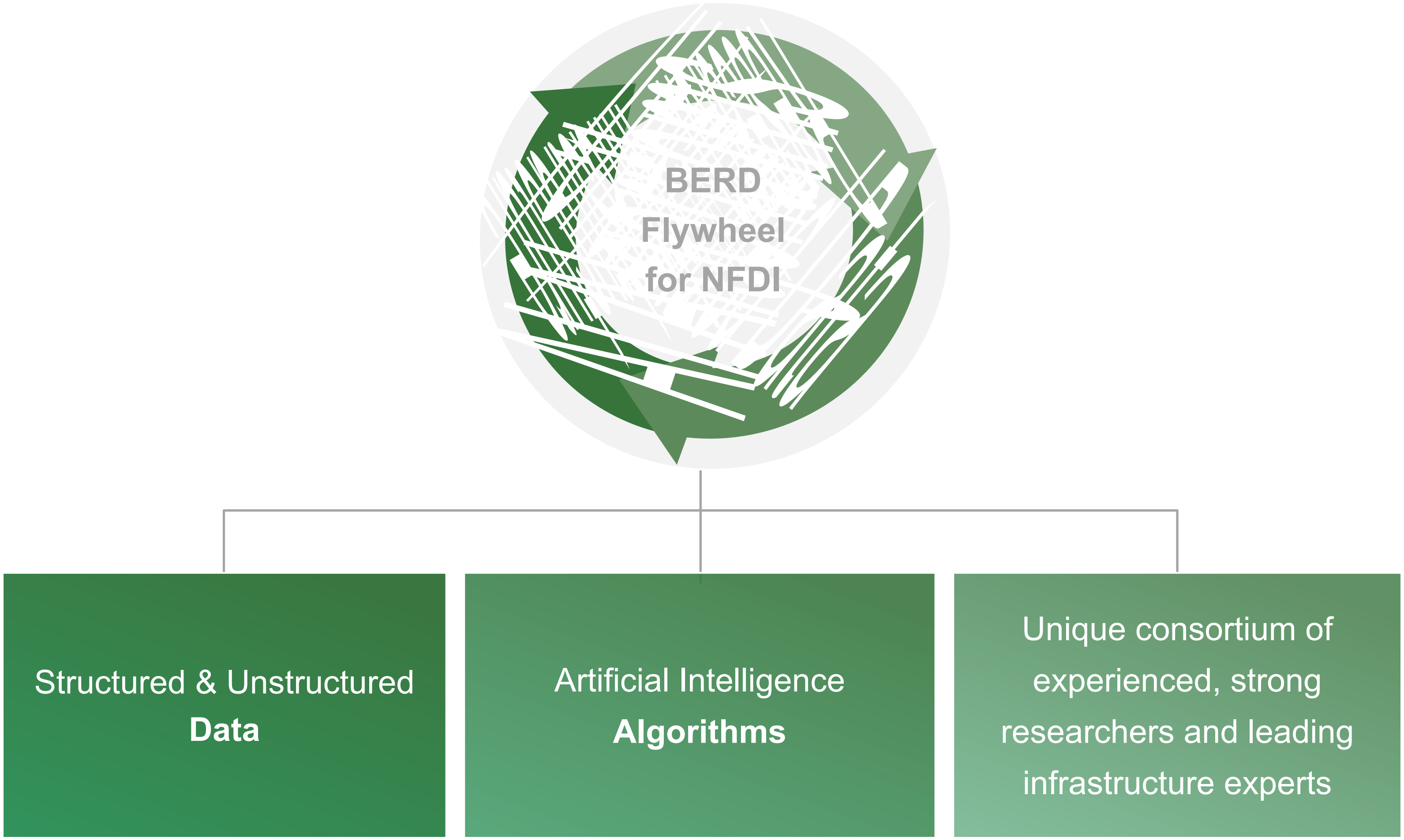
Paradigm shift

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

The BERD Consortium - a Unique Combination



BERD@NFDI in a Nutshell



BERD Flywheel for NFDI

Structured & Unstructured **Data**

Artificial Intelligence **Algorithms**

Unique consortium of experienced, strong researchers and leading infrastructure experts

Backup

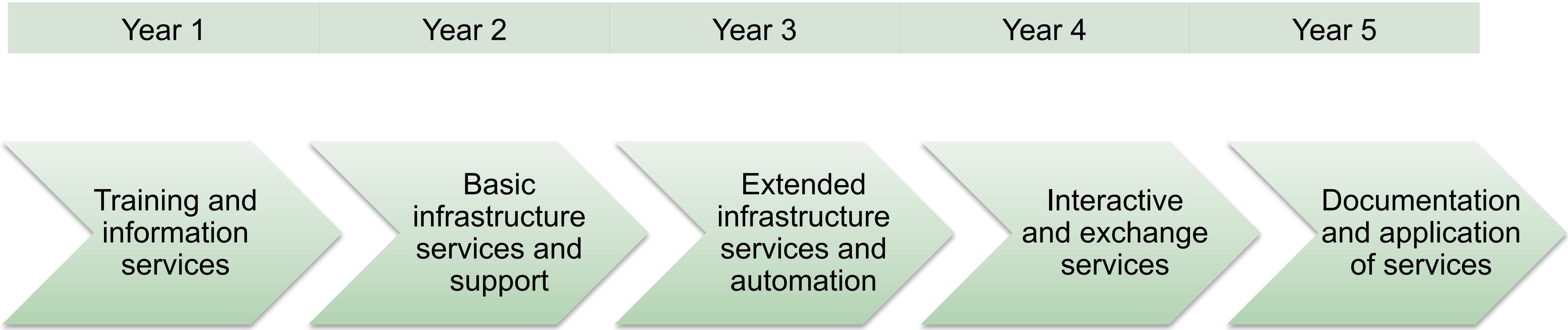
Services Roll-out Schedule

	Year 1	Year 2	Year 3	Year 4	Year 5
Task Area 1	Continuous feedback generation using methods of User-Centered Requirements Engineering; Web and social media presence; Scientific publication on BERD	Dissemination events			Evaluation report on impact and success
Task Area 2		Guidelines and legal consulting services	Focused crawlers and data harvesters	Upload feature	Ingestion services fully operable metadata normalization tool implemented
Task Area 3	Recommender service for OCR tools	Initial set of standards for data quality assessment and data normalization	Inventory of German Firm Data and Sources; Integrated OCR-D workflow	User interaction functionality for discussion of standards; Documentation of new anonymization techniques	Guidelines for data quality documentation and data normalization; New data sets
Task Area 4	Storage and computing capacity set up	Algorithm repositories connected; Initial reporting standards for performance of data analysis tools available		Continuous assessment of algorithms established; Graphical User Interface for BERD ontology	Information extraction from unstructured resources based on BERD ontology

Task Area 5		Metadata Schema specification;
		Prototype of search infrastructure;
		PID service technically integrated;
		Information portal
			Mapping of harvested metadata; Deep indexing for domain-specific searches;
			Single sign-on;
			Virtual BERD@NFDI environment;
			Migration service
				Metadata-based quality check for (incoming) harvested metadata	Continuous metadata normalization and preservation

Task Area 6	Self-learning modules for researchers; Training events for researchers and librarians; Consultancy service for research data management; Automated data stewardship services pilot	Support for BERD@NFDI infrastructure	Fully automated data stewardship services; Automated feedback loops	Export and exchange services	

Services Roll-out Schedule



Year 1

Year 2

Year 3

Year 4

Year 5

Training and information services

Basic infrastructure services and support

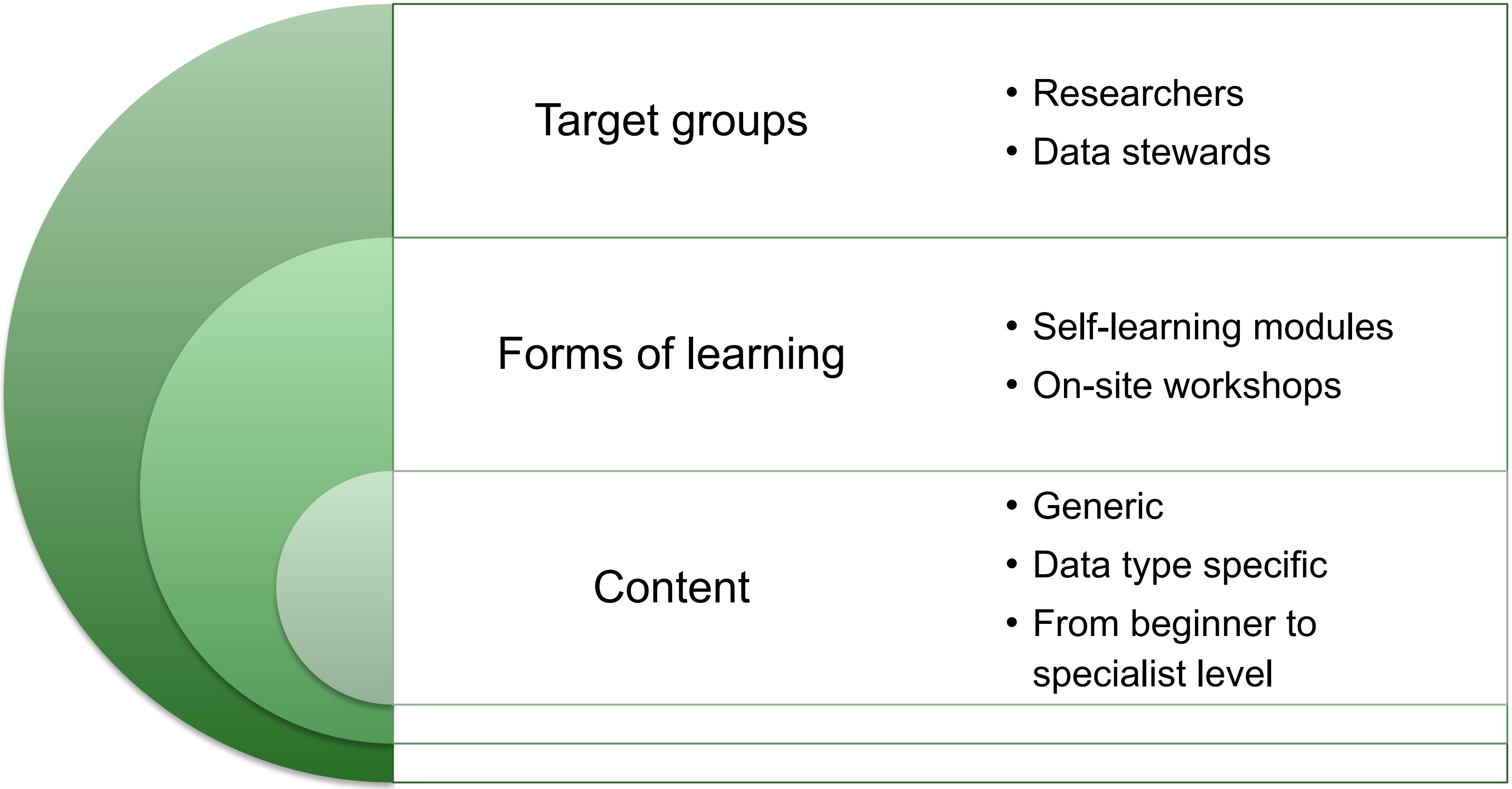
Extended infrastructure services and automation

Interactive and exchange services

Documentation and application of services

Apart from the implementation of new BERD@NFDI services, existing services (from BERD@BW, OpenML etc.) are continued and will be integrated into the BERD@NFDI information portal.

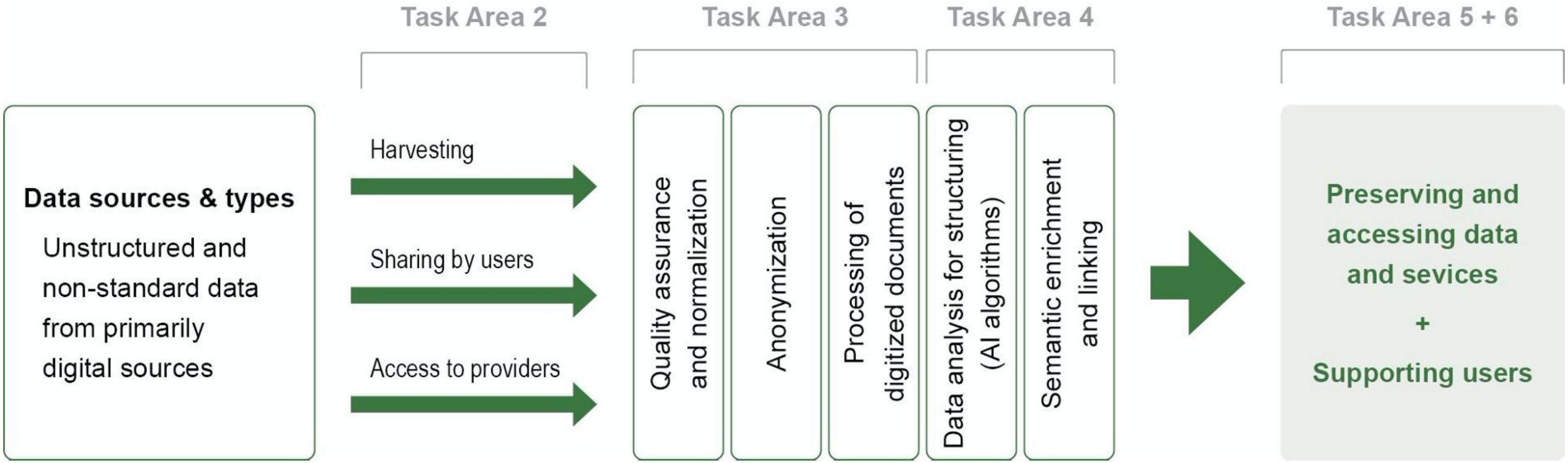
Training and Teaching



Target groups	• Researchers • Data stewards
Forms of learning	• Self-learning modules • On-site workshops
Content	• Generic • Data type specific • From beginner to specialist level

Based on vast experience at Mannheim University and LMU

BERD@NFDI takes the challenge


Task Area 2
Task Area 3
Task Area 4
Task Area 5 + 6
Data sources & types
Unstructured and non-standard data from primarily digital sources
Harvesting
Sharing by users
Access to providers
Quality assurance and normalization
Anonymization
Processing of digitized documents
Data analysis for structuring (AI algorithms)
Semantic enrichment and linking
Preserving and accessing data and sevices
+
Supporting users

BERD is indispensable for the Social Sciences NFDI

STRUCTURED DATA

Characteristics:

- ▦ e.g. survey data, administrative data
- ≡ high standardization
- ≡ homogenity of sources and formats
- ✓ standardized collecting, managing and analyzing
- ✓ standardized tools and methods
- ✓ sufficient computing and storage capacity
- ⊗ no interconnected data infrastructure
- ⊗ open legal and ethical issues

↓

KonsortSWD

UNSTRUCTURED DATA

Characteristics:

- ▦ e.g. text, video, audio
- ≡ low standardization
- ≡ heterogeneity of sources and formats
- ⊗ no standardized collecting, managing and analyzing
- ⊗ no standardized tools and methods
- ⊗ no sufficient computing and storage capacity
- ⊗ no developed data infrastructure
- ⊗ open legal and ethical issues

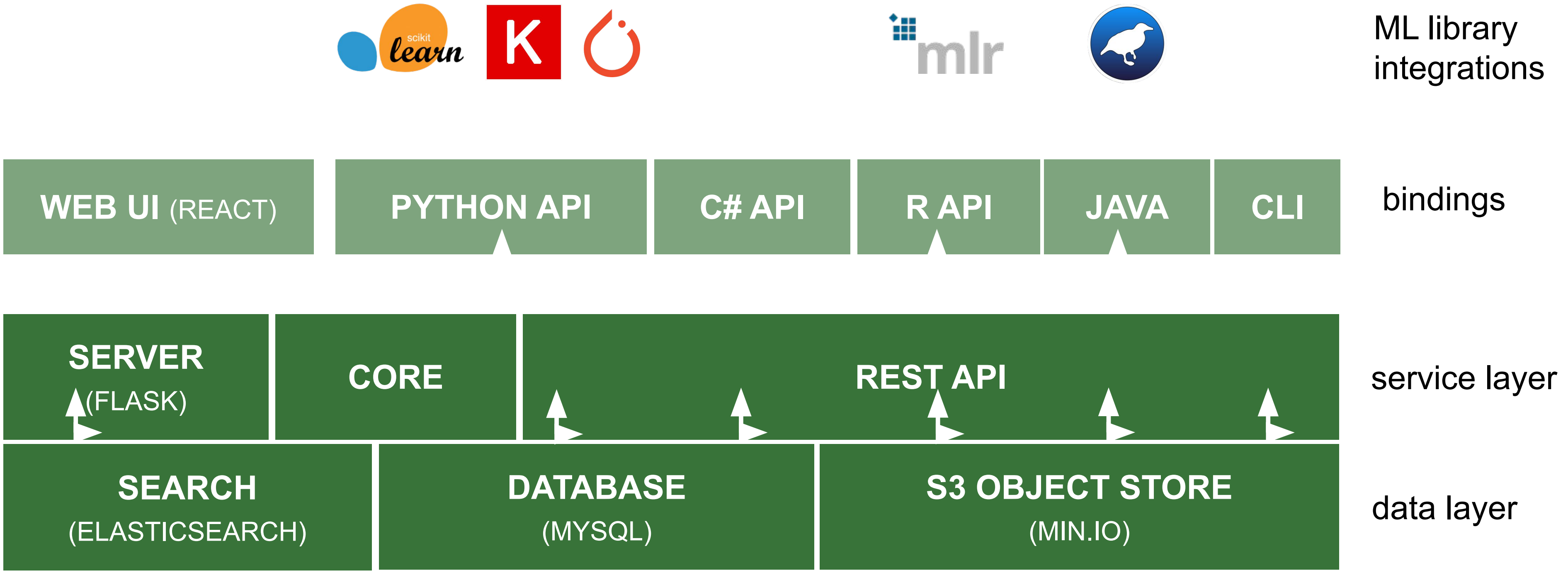
↓

BERD@NFDI

Common goal

Interconnected infrastructure for handling both structured and unstructured data

OpenML Technical Architecture



ML library integrations

WEB UI (REACT) | PYTHON API | C# API | R API | JAVA | CLI

bindings

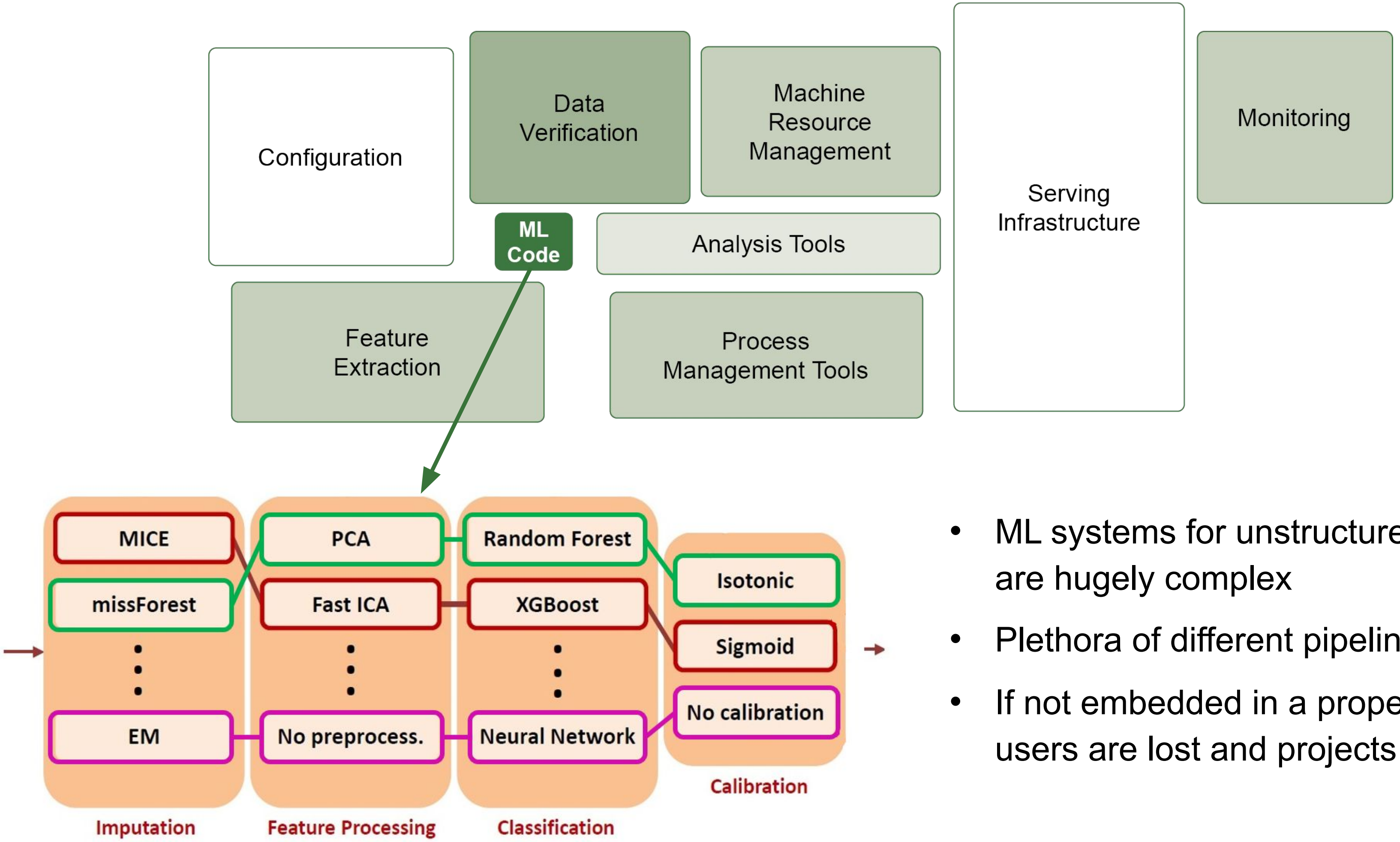
SERVER (FLASK) | CORE | REST API

service layer

SEARCH (ELASTICSEARCH) | DATABASE (MYSQL) | S3 OBJECT STORE (MIN.IO)

data layer

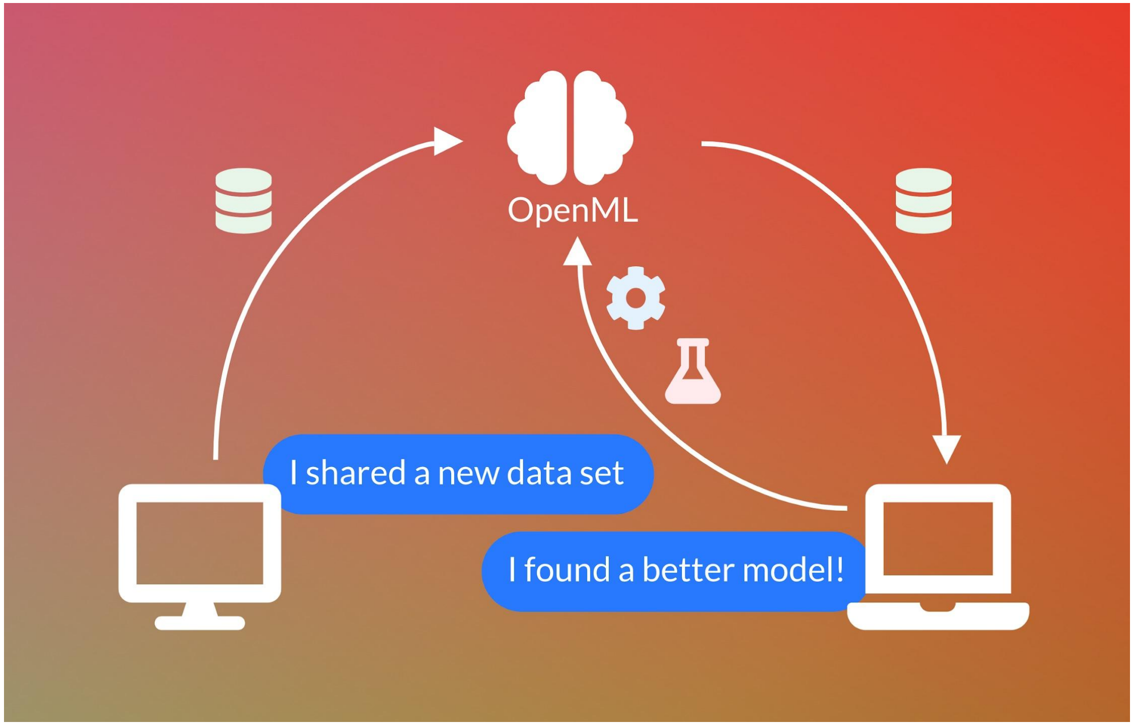
Hidden Technical Depth of Machine Learning



- 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)

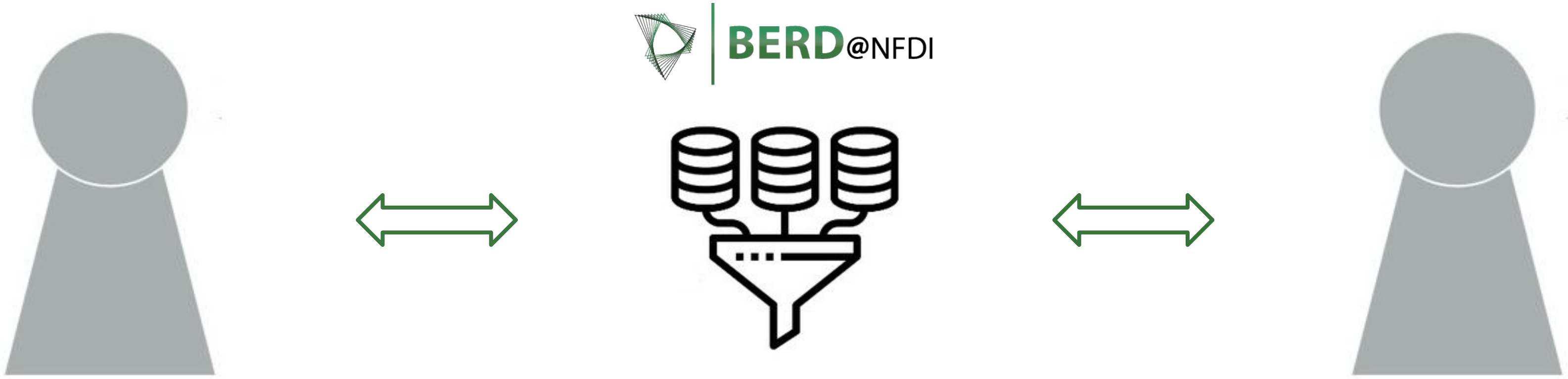
Source: Sculley, D. et al. (2015): “Hidden technical debt in Machine learning systems”, in: NIPS’15: Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 2, 2503-2511, https://dl.acm.org/doi/10.5555/2969442.2969519.

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