



Supporting Research Data Management with Generative AI – a Modern FAIRytale?

Prof. Dr. Sandra Geisler
Lecture Series NFDI4DS
27.06.2025



INTERNET OF
PRODUCTION



Juniorprofessur
für Datenstrom-
Management
und -Analyse

RWTH AACHEN
UNIVERSITY

Data Sharing in Research

SPRINGER NATURE Link

Find a journal Publish with us

Home > Orphanet Journal of Rare Diseases > Volume 16, article number 161, (2021) > FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare

Data silos are undermining development and care for patients

Position statement | Open access | Published: 16 November 2021

Volume 16, article number 161, (2021) | DOI: <https://doi.org/10.1136/bmj-2024-081554>

Download PDF  You have access

Nathan Denton , Monique Molloy , Daniel Mulberg, Paul Howard  & Eric D. Mandel 

5413 Accesses  32 Altmetrics

Abstract

Data silos are proliferating while genetic and immunological advances are uncovering previously unknown etiologies. This fragmentation hinders clinical, and research communities that disease-specific treatments are on the way. However, we fear the tendency of various stakeholders to balkanize databases in proprietary formats, driven by current economic and academic incentives, will inevitably fragment the expanding knowledge base and undermine current and future research efforts to develop much-needed treatments. The proliferation of proprietary databases, compounded by a paucity of meaningful outcome measures and/or good natural history management insights. From 60 percent to 80 percent of this information is unstructured, often locked away in incompatible systems.

Lekadir et al., 2025. FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ* 2025; 388; <https://doi.org/10.1136/bmj-2024-081554>

AI tools in healthcare should be:

F U T U R E

FAIR **UNIVERSAL** **TRACEABLE** **USABLE** **ROBUST** **EXPLAINABLE**

HealthData Management

DEMOCRATIZATION INTEROPERABILITY TECHNOLOGY

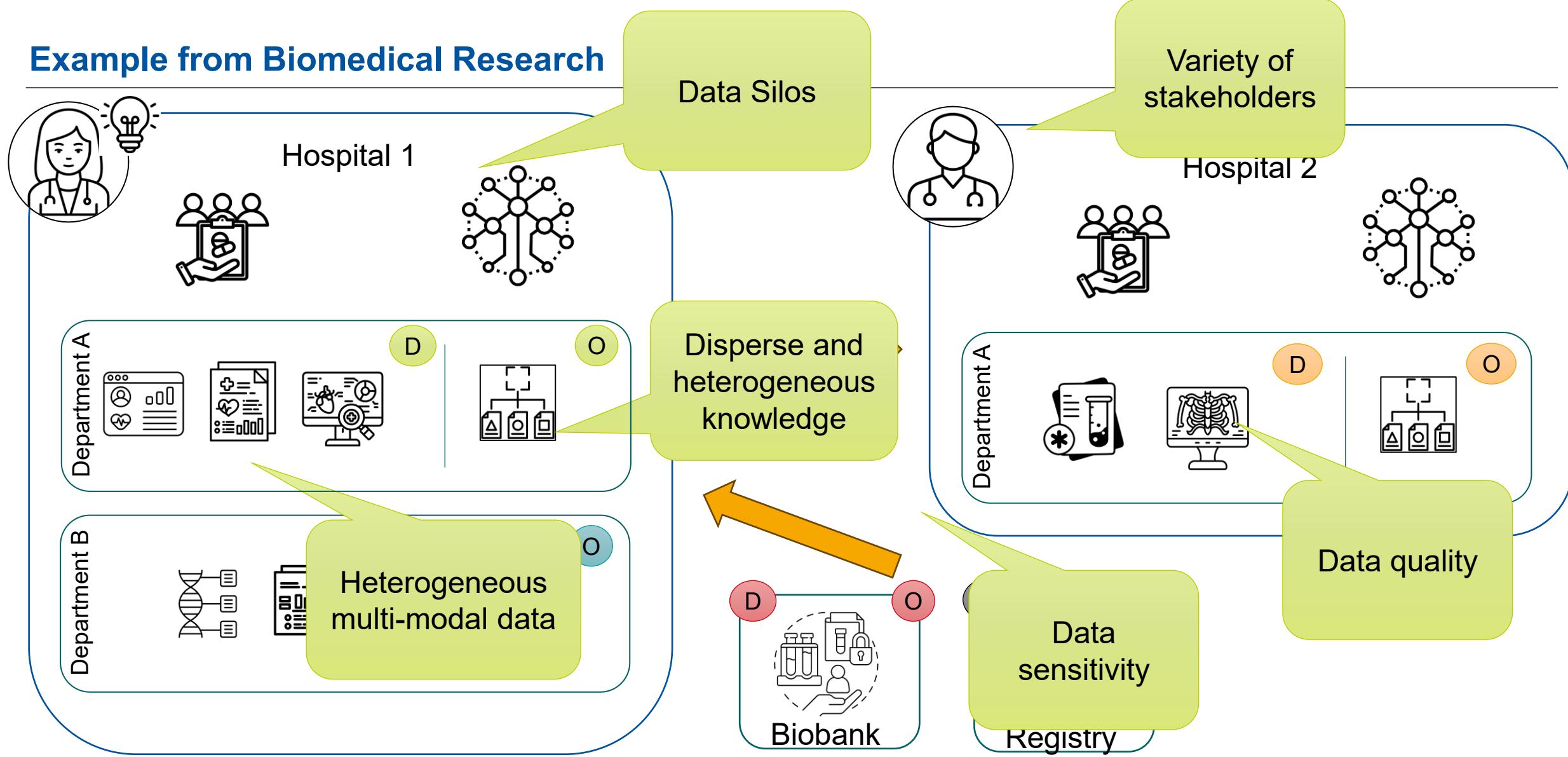
Democratization can solve healthcare's problems

Technology can play a significant role in enhancing

SHARE:   



Example from Biomedical Research



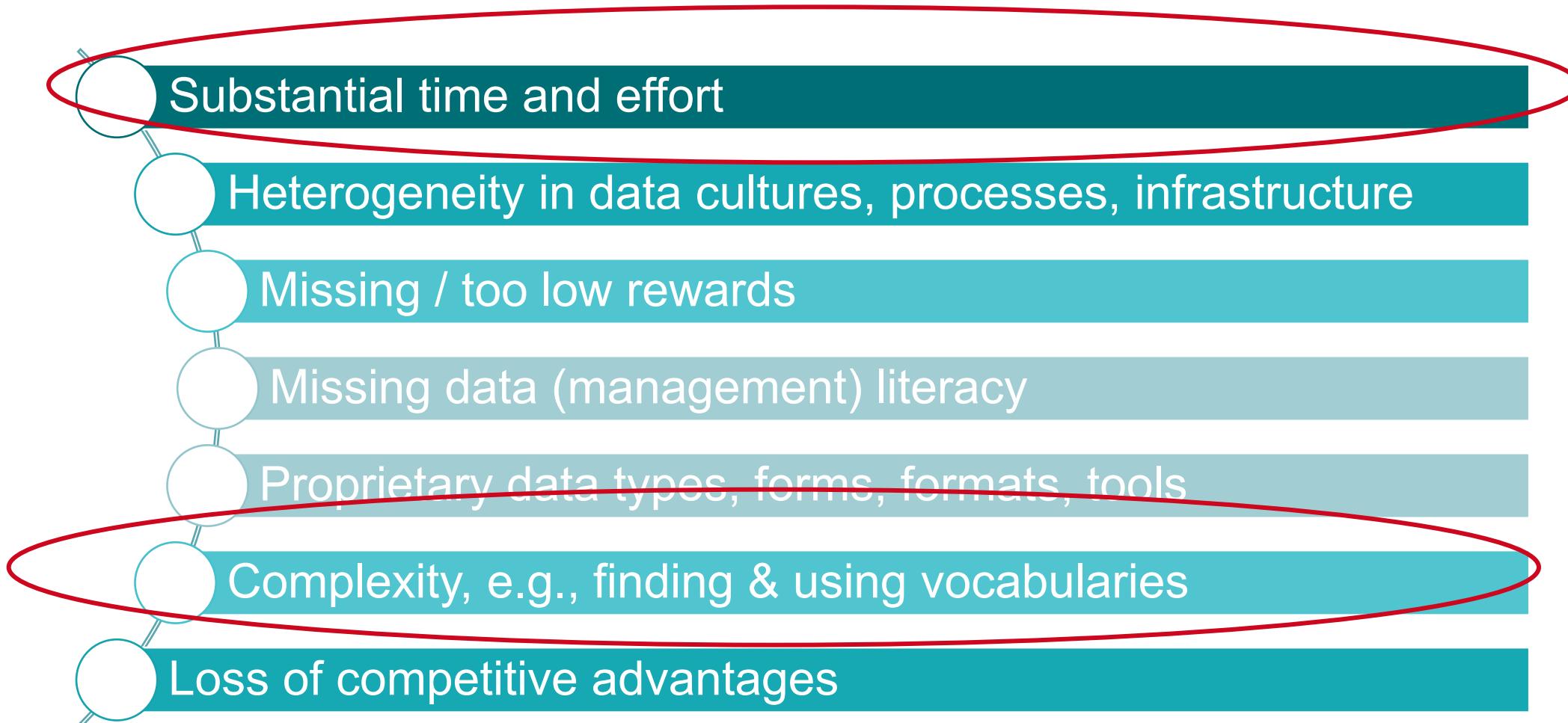
EXC "Internet of Production" - Smart, Connected Production



“The vision of the Internet of Production (IoP) is to enable a new level of **cross-domain** collaboration by providing **semantically adequate** and **contextual data** from production, development and usage **in real time** at an appropriate granularity.”

<https://www.iop.rwth-aachen.de/>

Barriers for FAIR Research Data Management



[Feger et al., 2019; Kim et al., 2023; Jussen et al., 2023, Stuart et al., 2018]

LLMs – Typical tasks

- Complete and create text
- Summarizing, answering questions, translating, ...
- Intelligent assistants
 - Office applications
 - Teaching & learning
 - Programming
 - Robot control
 - Science communication



From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles

Sandra Geisler
RWTH Aachen University
Aachen, Germany
geisler@cs.rwth-aachen.de

David Chaves-Fraga
Universidade de Santiago de Compostela
Santiago de Compostela, Spain
david.chaves@usc.es

Maurizio Lenzerini
Sapienza University of Rome
Rome, Italy
lenzerini@diag.uniroma1.it

Anastasia Dimou
KU Leuven
Leuven, Belgium
anastasia.dimou@kuleuven.be

Anisa Rula
University of Brescia
Brescia, Italy
anisa.rula@unibs.it

Sascha Welten
RWTH Aachen University
Aachen, Germany
welten@dbis.rwth-aachen.de

Cinzia Cappiello
Politecnico di Milano
Milan, Italy
cinzia.cappiello@polimi.it

Anastasia Dimou
KU Leuven
Leuven, Belgium
anastasia.dimou@kuleuven.be

Anisa Rula
University of Brescia
Brescia, Italy
anisa.rula@unibs.it

Maria-Ester Vidal
TIB Hannover
L3S Research Center
Leibniz University Hannover
Hannover, Germany
vidal@l3s.de

ABSTRACT

Knowledge graphs (KGs) play a crucial role in the integration and organization of heterogeneous data and knowledge, enabling advanced data analytics and decision-making across various industries. This vision paper addresses critical challenges in managing KGs, emphasizing their relevance in integrating information from disparate sources. We propose the concept of knowledge graph ecosystems and life cycles to systematically manage tasks, e.g., data integration, standardization, continuous updates, efficient querying, and provenance tracking. By adopting our approach, organizations can enhance the accuracy, consistency, and reliability of KGs, thus improving knowledge management, enabling the extraction of valuable insights, and ensuring transparency and accountability.

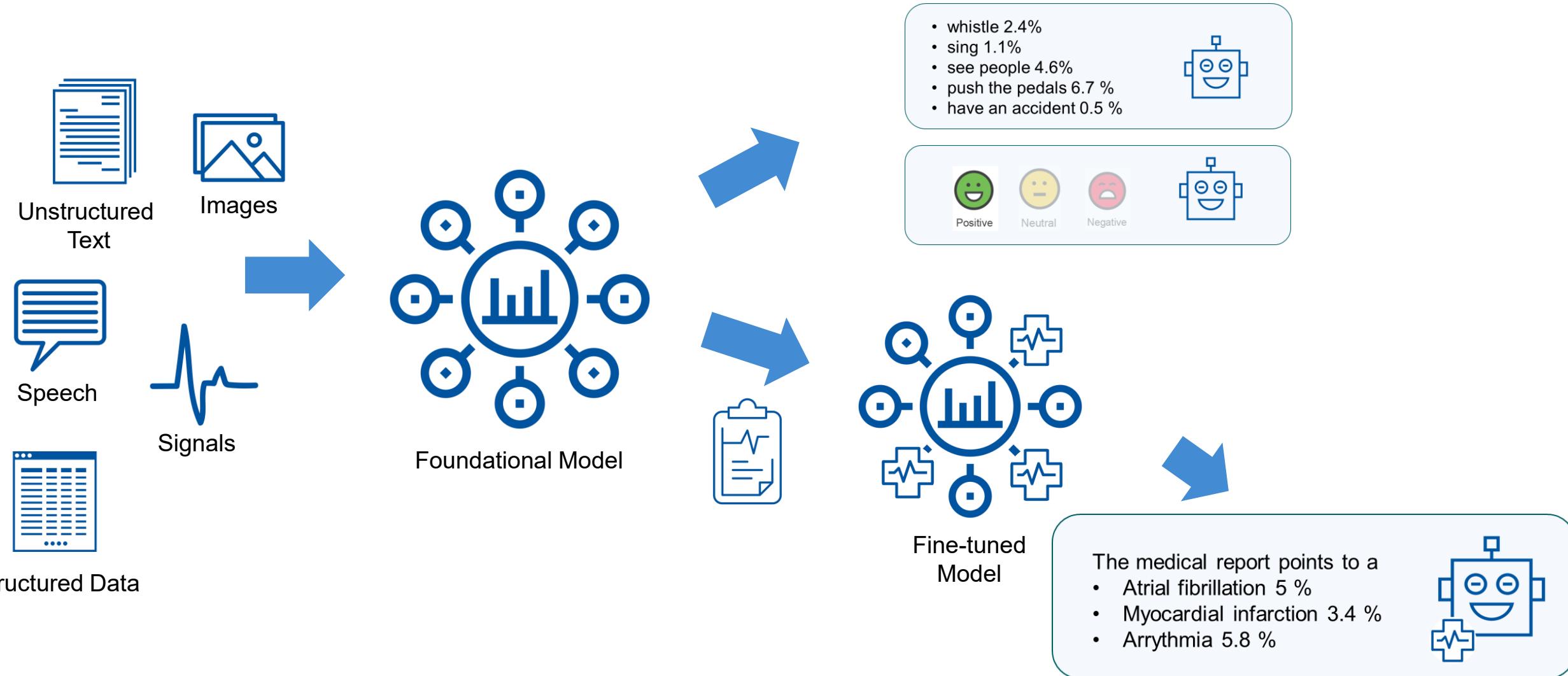
PVLDB Reference Format:

Sandra Geisler, Cinzia Cappiello, Irene Celino, David Chaves-Fraga, Anastasia Dimou, Ana Iglesias-Molina, Maurizio Lenzerini, Anisa Rula, Dylan Van Assche, Sascha Welten, and Maria-Ester Vidal. From Genesis to Maturity: Managing Knowledge Graph Ecosystems Through Life Cycles. PVLDB, 18(5): 1390 - 1397, 2025.
doi:10.14778/3718057.3718067

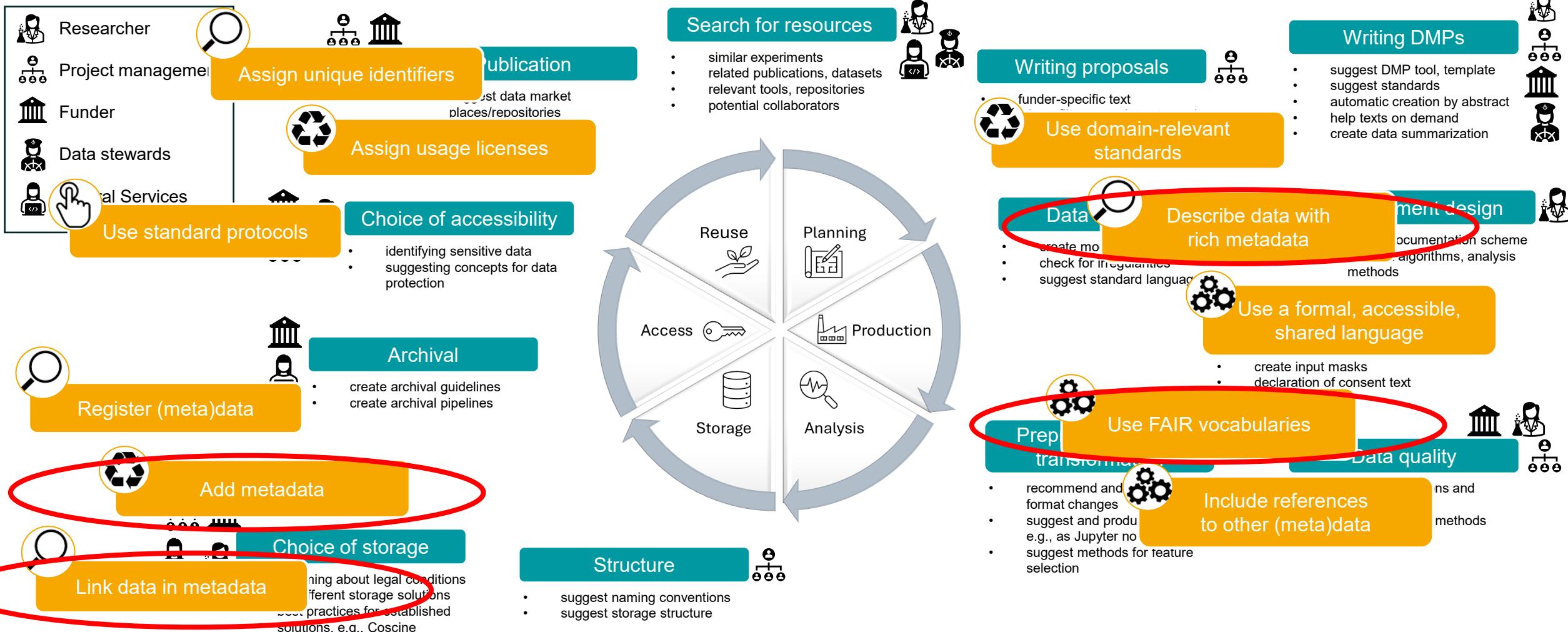
https://www.youtube.com/watch?v=v0Mrhyvme_EQ
<https://www.microsoft.com/de-DE/microsoft-365/copilot?>



Fine-tuning & Retrieval Augmented Generation (RAG)



Research Data Management Cycle



Geisler & Kim, Unlocking the potential: LLMs Transforming Research Data Management, Tag der Forschungsdaten NRW, 2023

Metadata Annotation of Data Sets and Data – Use Case Fineblanking

Material Forming - ESAFORM 2024
Materials Research Proceedings 41 (2024) 1436-1445

Materials Research Forum LLC
<https://doi.org/10.21741/9781644903131-159>

Predicting fine blanking process signals from sheet metal thickness

MOON Jiyoung^{1,a*}, GELBICH Daria^{1,b}, BECKER Marco^{1,c}, NIEMIETZ Philipp^{1,d}
and BERGS Thomas^{1,2,e}

¹Manufacturing Technology Institute MTI of RWTH Aachen University, Campus-Boulevard 30,
52074 Aachen, Germany

²Fraunhofer Institute for Production Technology IPT, Steinbachstr. 17, 52074 Aachen, Germany

^aj.moon@mti.rwth-aachen.de, ^bd.gelbich@mti.rwth-aachen.de, ^cm.becker@mti.rwth-aachen.de,
^dp.niemietz@mti.rwth-aachen.de, ^et.bergs@mti.rwth-aachen.de

Keywords: Sheet M

Abstract. In sheet metal forming processes, it is often difficult to monitor process conditions and product quality in real time due to the inaccessibility of the tooling. This lack of information, such as force or strain measurements, makes it challenging to obtain detailed information about the process. In addition, the complexity of the process signals can depend on different internal factors, such as sheet metal properties, tool geometry, and process parameters, which can lead to complex and non-linear relationships between the measured signals.

Name	Meaning
ground_truth	Actual values of process signals, obtained through sensors or manual measurement. It serves as the reference for training the model.
label_data	Labels corresponding to the ground truth data, indicating specific process states or categories.
scaler_model	A trained machine learning model that scales the input features to a standard range, typically between 0 and 1.
training_data	Data used to train the machine learning model, consisting of input features and their corresponding labels.

M2,M4,M6

-7050.158868533855, -6545.197609458095, -7069.382093213849
-6974.289952503707, -6450.2257491644905, -7004.105439762875
-6888.948591670646, -6305.549122272035, -6907.7276024169705
-6751.457746721394, -6118.219682318749, -6766.739859775819
-6565.683333005749, -5939.763193920102, -6578.869012704587
-6378.316085990314, -5778.07016807039, -6404.416245037385
-6205.615321967468, -5646.562328180286, -6231.639715977421
-6064.323691867901, -5552.310844229093, -6085.741985763209
-5960.395467087033, -5498.973085403063, -5982.132704421241
-5893.883970595199, -5485.416242488852, -5921.408832479124
-5870.69053169583, -5517.542405219507, -5894.467360214813
-5897.609639445589, -5579.899152752291, -5913.902163190129
-5952.995595790824, -5634.899606358309, -5971.307026071324
-6014.17204583685, -5665.853566171, -6020.791055125605
-6046.391837716967, -5674.289459888521, -6048.577688236898

[Moon et al., 2024]



Add dataset metadata

No thanks,
but please create a metadata schema for the data set

 Creating a metadata schema for your dataset will involve defining the essential attributes that describe the dataset, including its content, structure, provenance, and relevant contextual information. This schema will help in organizing and documenting the dataset to facilitate understanding, sharing, and future use.

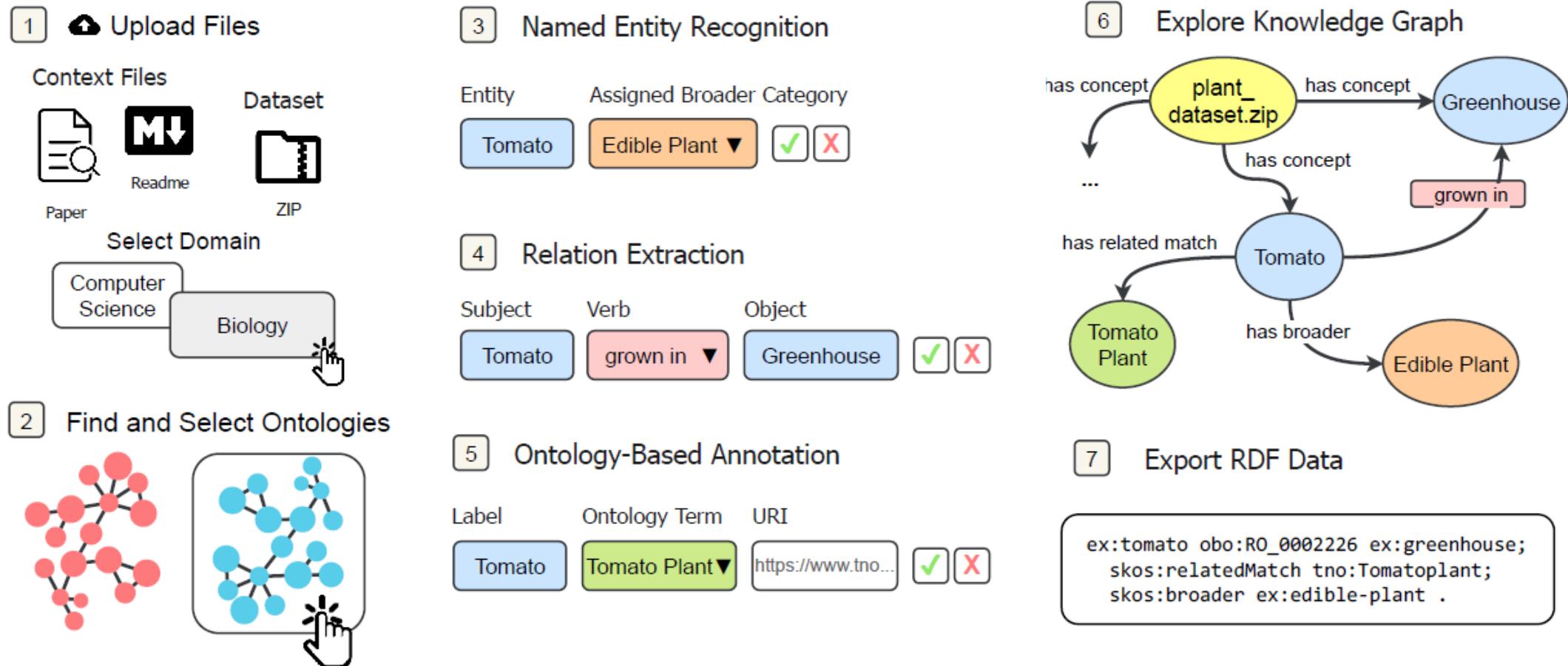
Metadata Schema for the Fine Blanking Process Signals Dataset:

- 1. Dataset Overview**
 - **Title:** Fine Blanking Process Signals with Synthetic Data from C-DCGAN Model
 - **Description:** A dataset containing time-series process signals from force sensors in a fine blanking process, with synthetic data generated using a Conditional Deep Convolutional GAN (C-DCGAN) for various sheet metal thicknesses.
 - **Creator(s):**
 - MOON Jiyoung, GELBICH Daria, BECKER Marco, NIEMIETZ Philipp, BERGS Thomas
 - **Publisher:** Materials Research Forum LLC
 - **Date of Creation:** [Date]
 - **Licensing:** Creative Commons Attribution 3.0 license
 - **Version:** 1.0

- 2. Data Provenance**
 - **Source:** RWTH Aachen University, Manufacturing Technology Institute (MTI)
 - **Collection Method:** Force signals were measured using piezoelectric sensors during a fine blanking process at 30 strokes per minute.
 - **Synthetic Data Generation:** Synthetic force signals were generated using a Conditional Deep Convolutional GAN (C-DCGAN), conditioned on sheet metal thickness.
- 3. Content Description**
 - **Type of Data:** Time-series data from force sensors (M2, M4, M6), synthetic signals, sheet thickness metadata
 - **Variables:**
 - **ForceSignal_M2:** Force measurement from sensor M2 (N)
 - **ForceSignal_M4:** Force measurement from sensor M4 (N)
 - **ForceSignal_M6:** Force measurement from sensor M6 (N)
 - **SheetThickness:** Thickness of the sheet metal (mm) [3.95 mm, 4.00 mm, 4.05 mm]
 - **ProcessStage:** The stage of the fine blanking process (Shearing or Stripping)
 - **SyntheticSignal:** Synthetic force signal generated by C-DCGAN (Y/N)
 - **File Formats:** CSV files containing the time-series data
- 4. Structure**
 - **Ground Truth Data:**
 - **Directory:** `ground_truth/` 
 - **File Naming Convention:** `DiCo-R_M2_M4_M6_<stage>_segX.csv`

Slide Credits: Martin Görz

KONDA – Semantification of Datasets



Kim, Görz & Geisler, KONDA: An LLM-based Tool for Semantic Annotation and Knowledge Graph Creation Using Ontologies for Research Data,
Sci-K Workshop@ISWC'25 & CoRDI ,2025

KONDA – The User Interface

The screenshot shows the KONDA user interface. At the top, there is a header bar with a search input field and some icons. Below the header, the main area is divided into two sections: "Ontology" and "Export".

Ontology: This section displays an ontology in RDF triples. The triples include prefixes for owl and rdfs, and definitions for properties like RO_0002021 and external, and concepts like OEO_00000509.

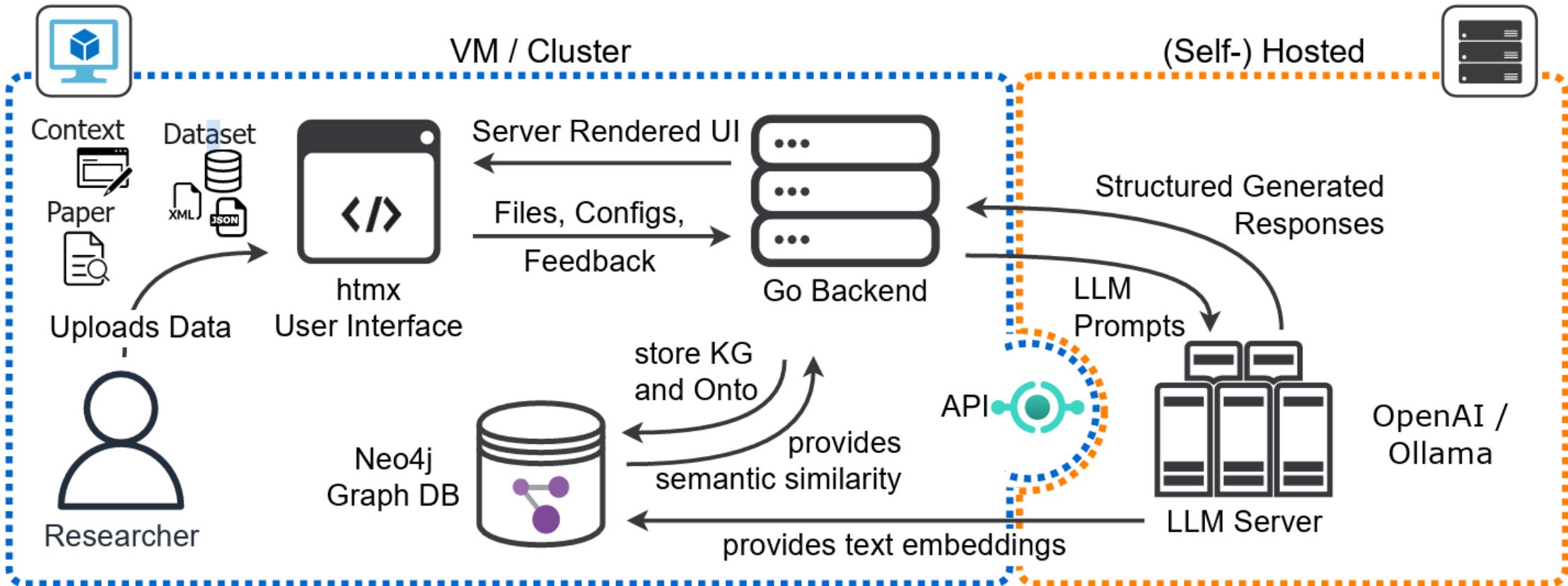
```
@prefix owl: <http://www.w3.org/2002/07/owl#> .  
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .  
  
<http://purl.obolibrary.org/obo/RO_0002021> a owl:ObjectProperty;  
    rdfs:label "RO_0002021";  
    rdfs:domain <http://example.org/ontology#Concept>;  
    rdfs:range <http://example.org/ontology#Concept> .  
  
<neo4j://graph.schema#external> a owl:DatatypeProperty;  
    rdfs:label "external";  
    rdfs:domain <http://example.org/ontology#Concept>, owl:ObjectProperty;  
    rdfs:range <http://www.w3.org/2001/XMLSchema#boolean> .  
  
<http://openenergy-platform.org/ontology/oeo/OEO_00000509> a owl:ObjectProperty;  
    rdfs:range <http://example.org/ontology#Concept>;  
    rdfs:domain <http://example.org/ontology#Concept>;  
    rdfs:label "OEO 00000509" .
```

Export: This section has a "Format" dropdown menu labeled "Pick a format" and an "Export" button.

At the bottom of the interface, there are navigation buttons: "Back", "Rerun", and a "Help" button.

Slide Credits: Martin Görz, Soo-Yon Kim

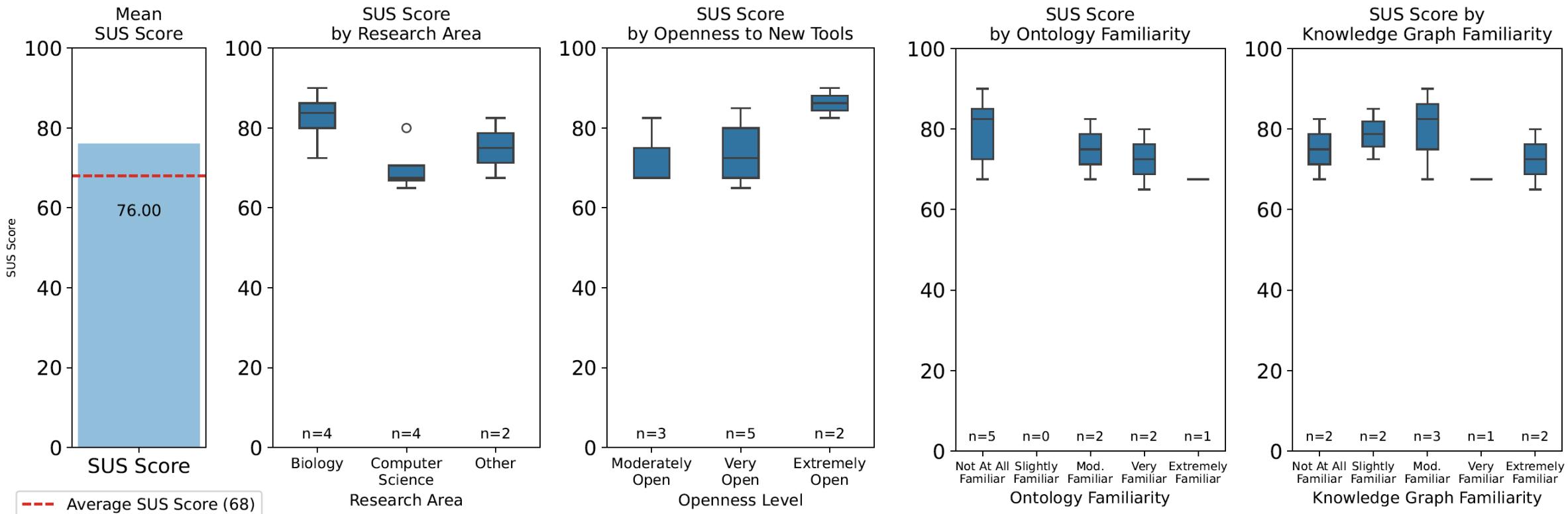
KONDA - Architecture



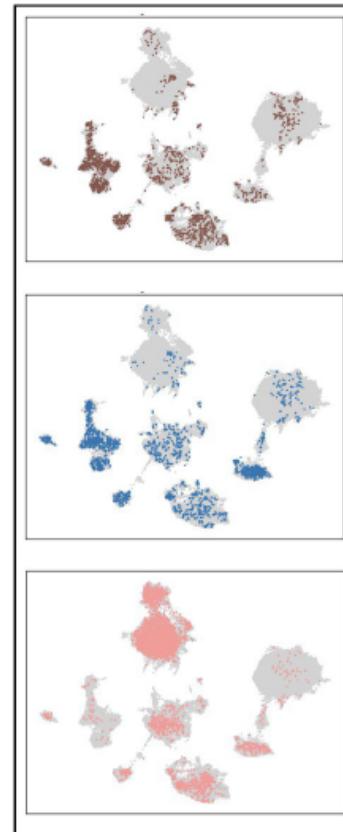
Kim, Görz & Geisler, KONDA: An LLM-based Tool for Semantic Annotation and Knowledge Graph Creation Using Ontologies for Research Data,
Sci-K Workshop@ISWC'25 & CoRDI ,2025

KONDA - Qualitative Evaluation

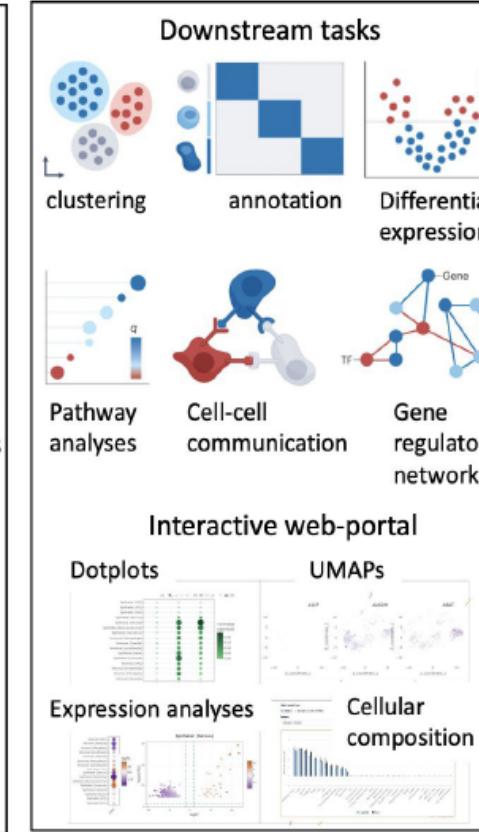
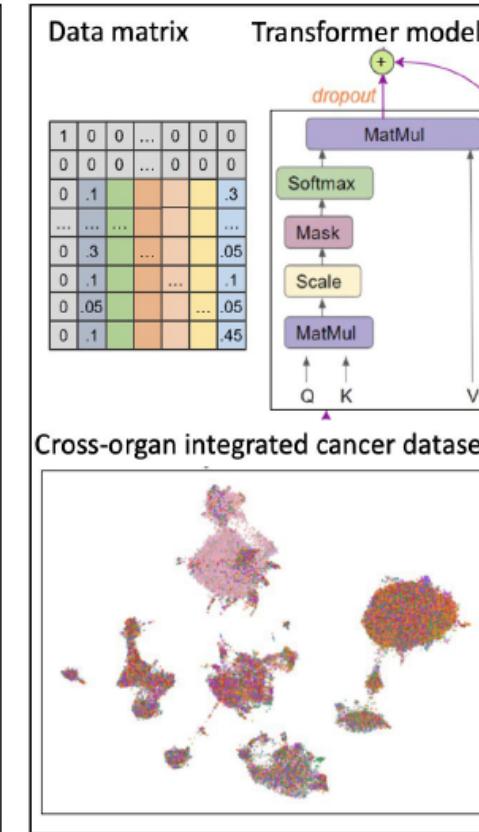
Study setup: 10 researchers from RWTH Aachen (computer science, biology, economics, materials science)



Kim, Görz & Geisler, KONDA: An LLM-based Tool for Semantic Annotation and Knowledge Graph Creation Using Ontologies for Research Data, Sci-K Workshop@ISWC'25 & CoRDI ,2025



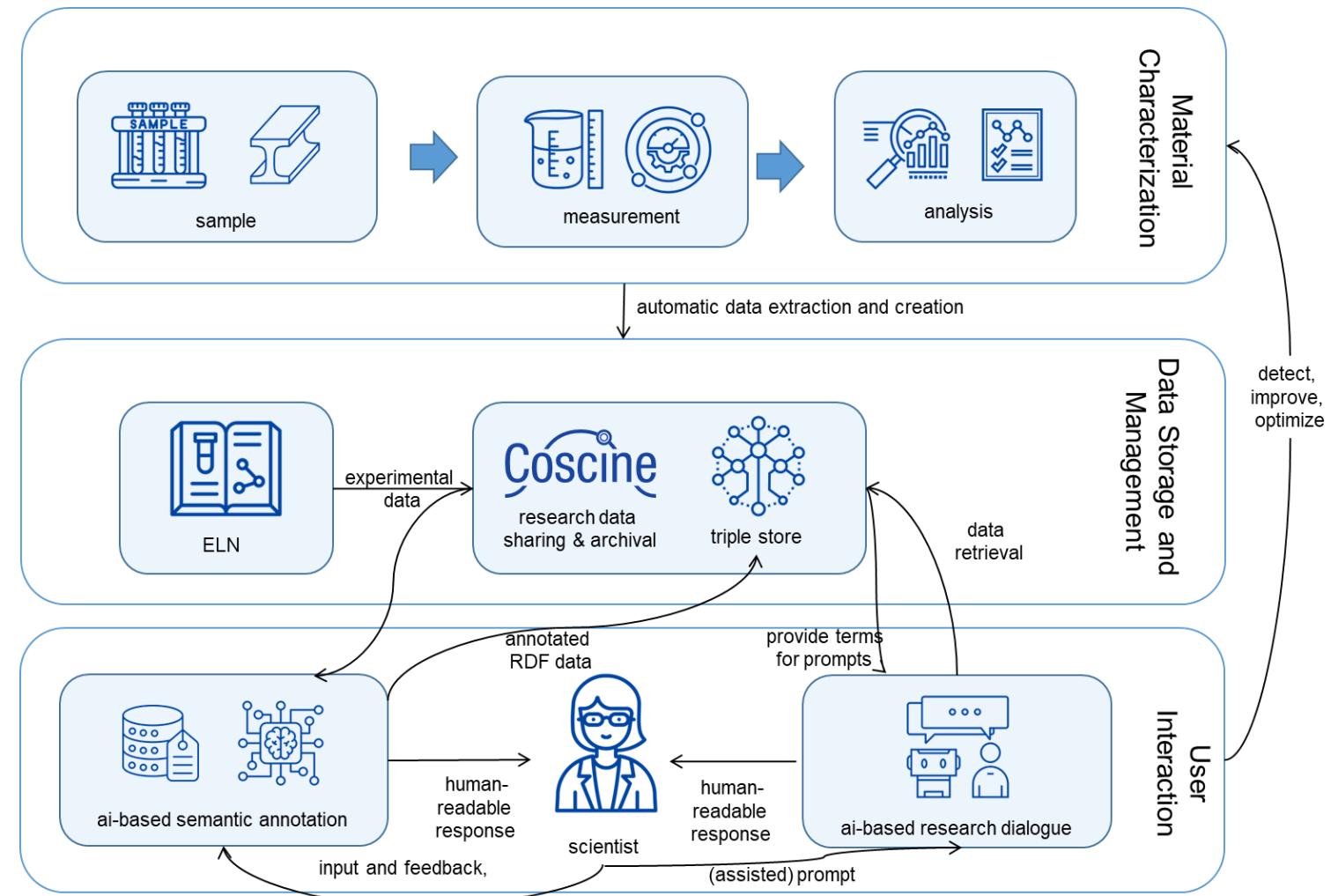
Identify entities (disease, cell-type, genes)
Assign Identifiers (ontologies)
Harmonize Identifiers (Standardization)
Common Metadata Exchange Schema
Overall FAIR data assembly



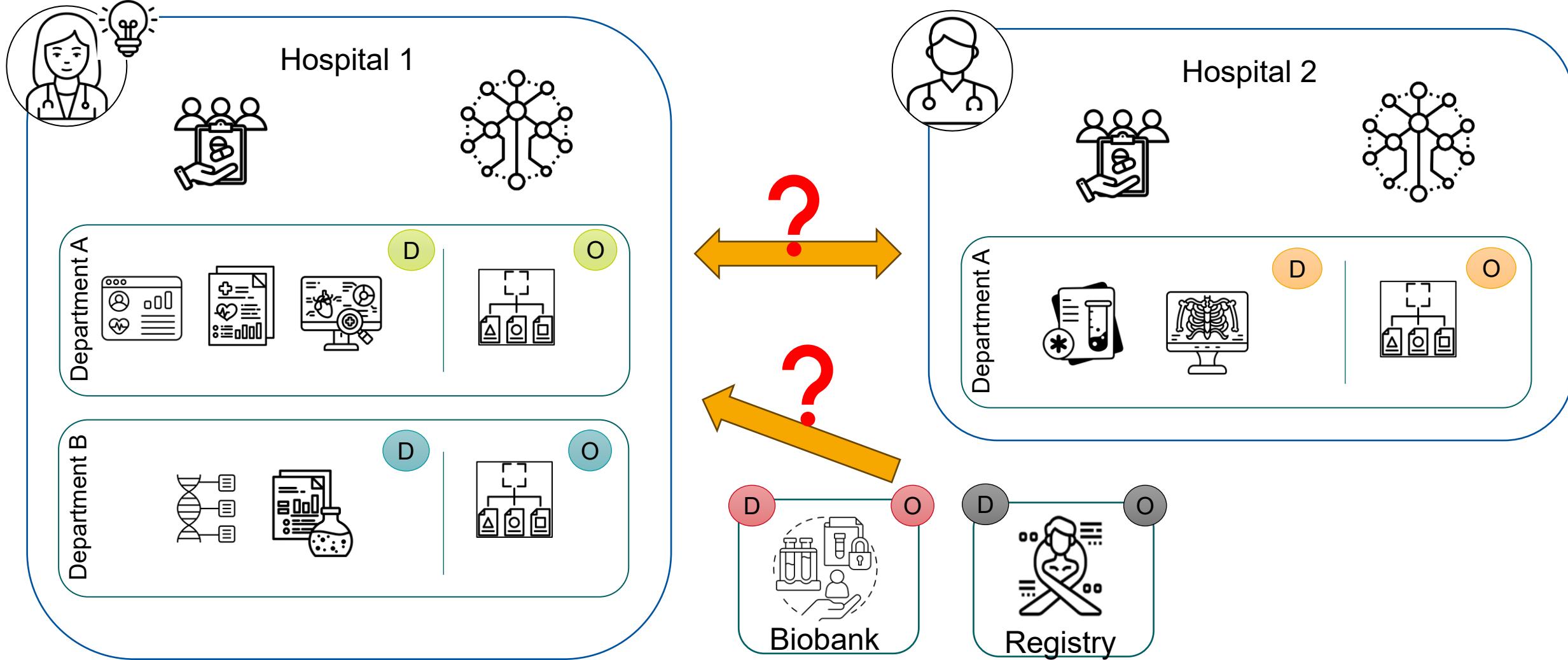
Downstream tasks and interactive user-portal

Slide credits: Sikander Hayat, Yongli Mou

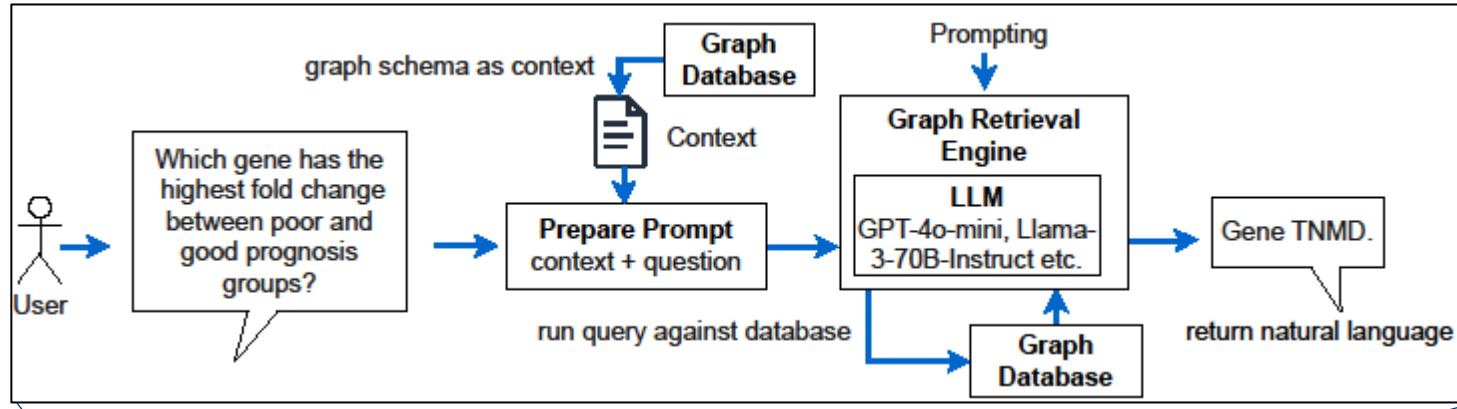
Semantic Annotation in Material Science



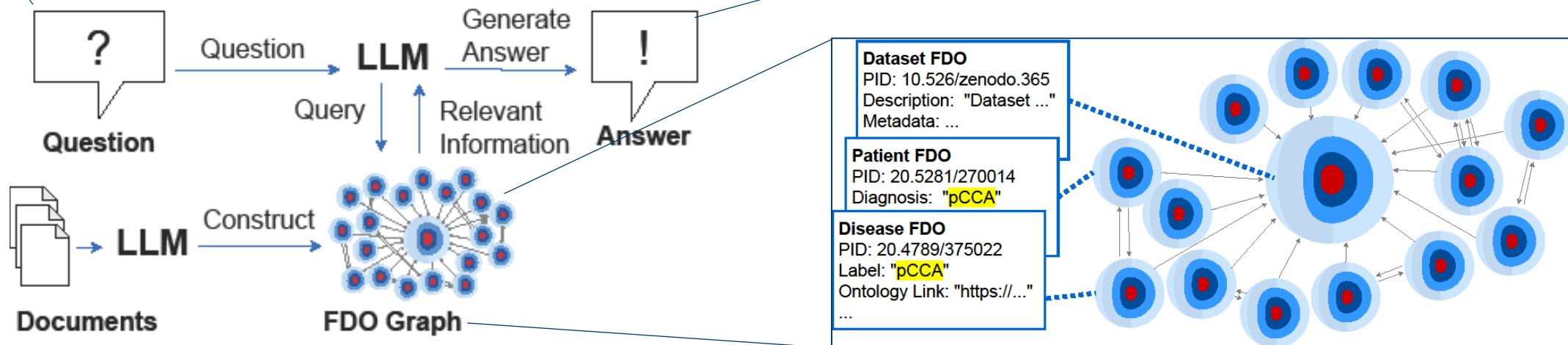
Where do we get similar / useful data?



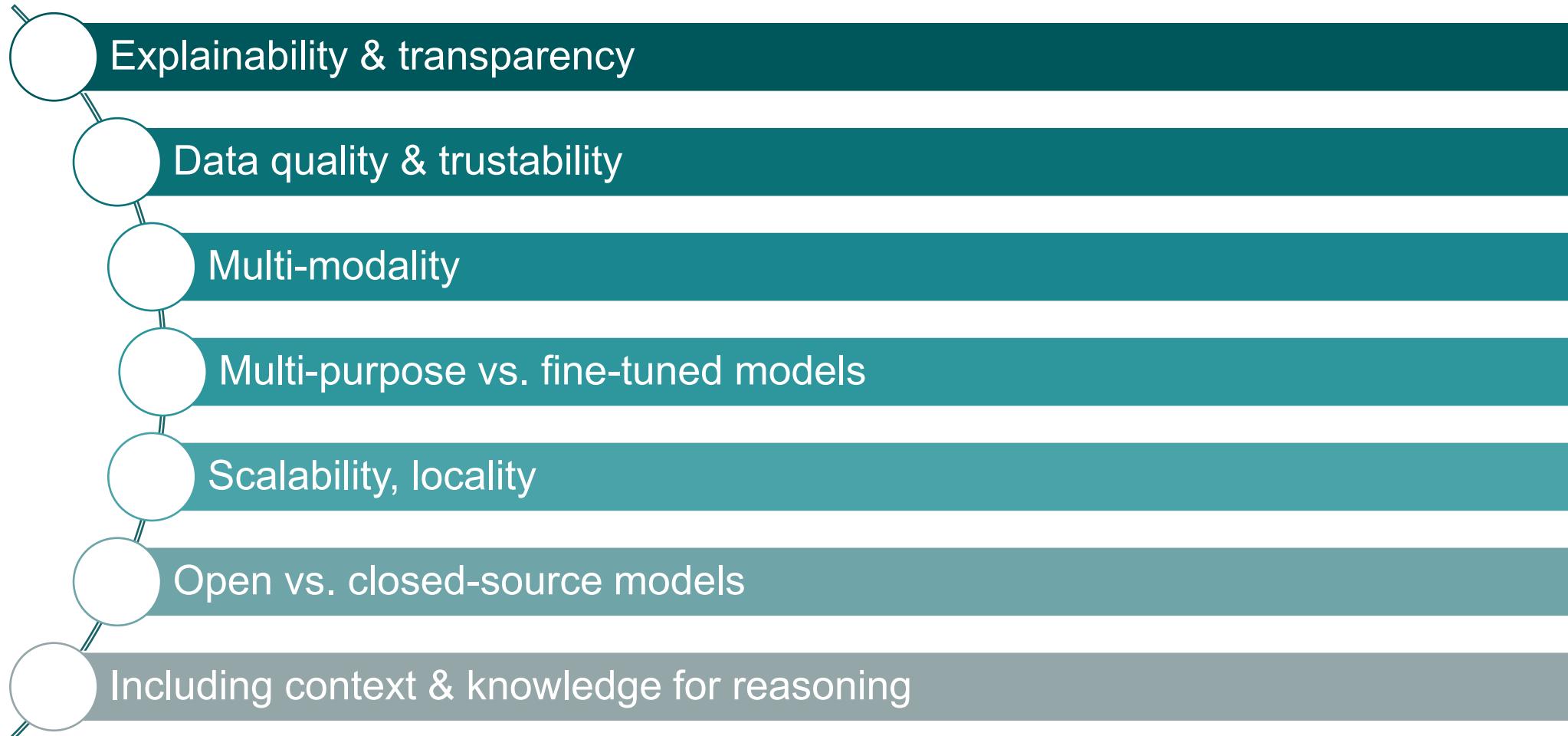
FAIR GraphRAG: A RAG Approach for Semantic Data Analysis



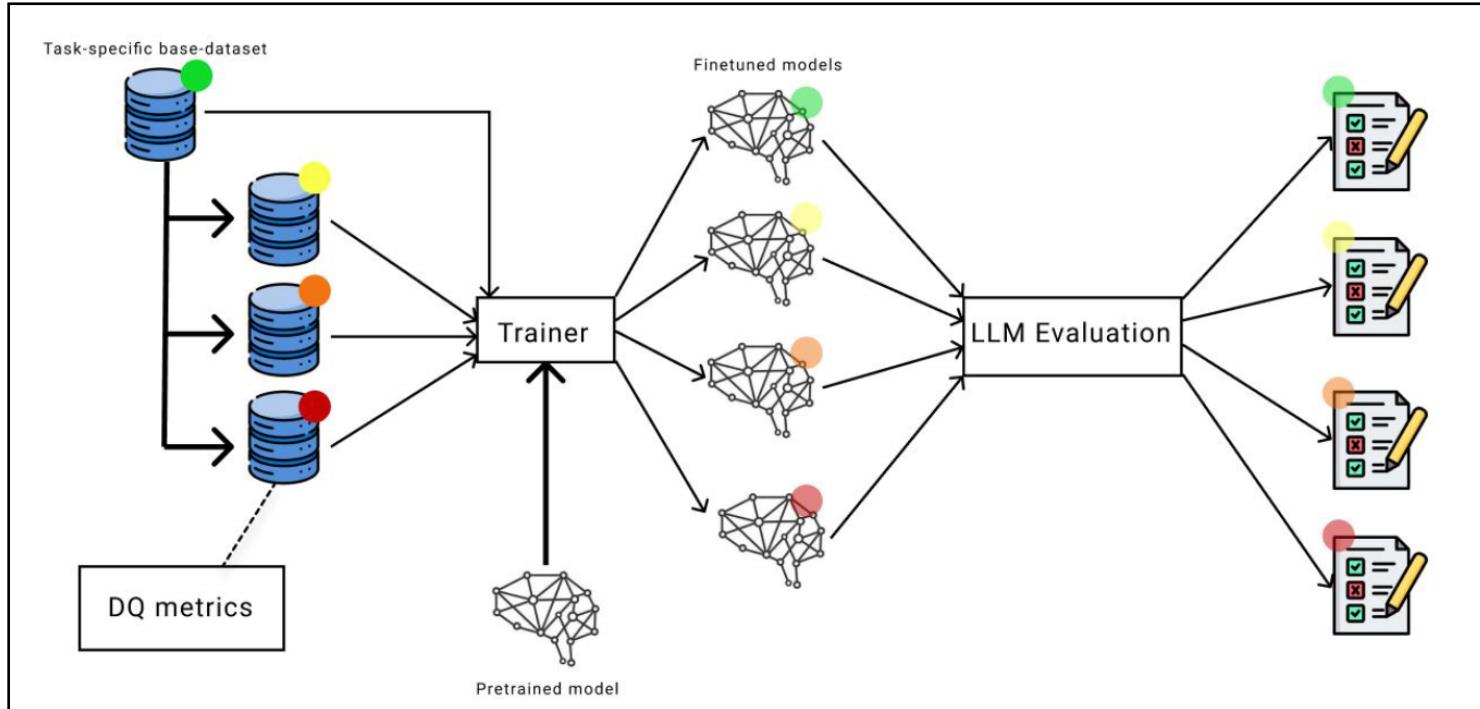
Flueh, Kim, Schneider & Geisler, 2025. FAIR GraphRAG: A Retrieval-Augmented Generation Approach for Semantic Data Analysis, ICKG 2025 (to appear)



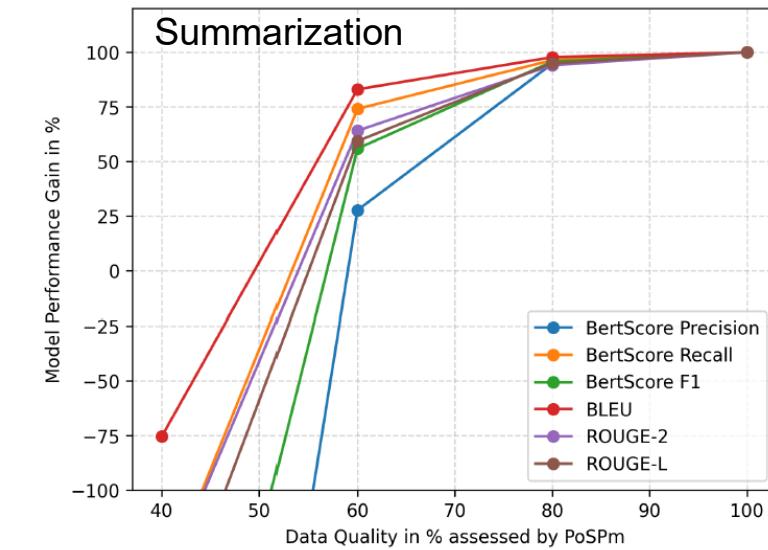
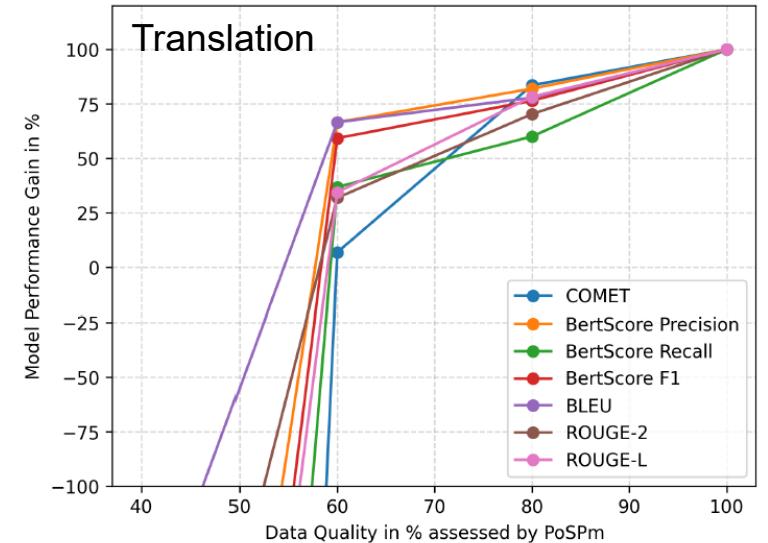
Adjusting screws



Quality of Training Data



Slide credits: Jan Dirks



Prompting Techniques – Improve Response Quality

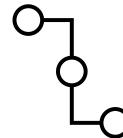
Inner Monologue*



Have an inner monologue on how to solve the task first. Only then give a final answer.

Chain-of-Thought

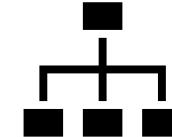
[Wei et al., 2022]



Solve the task by solving subtask 1, then subtask 2, then give the final answer.

Tree-of-Thought

[Yao et al., 2023]



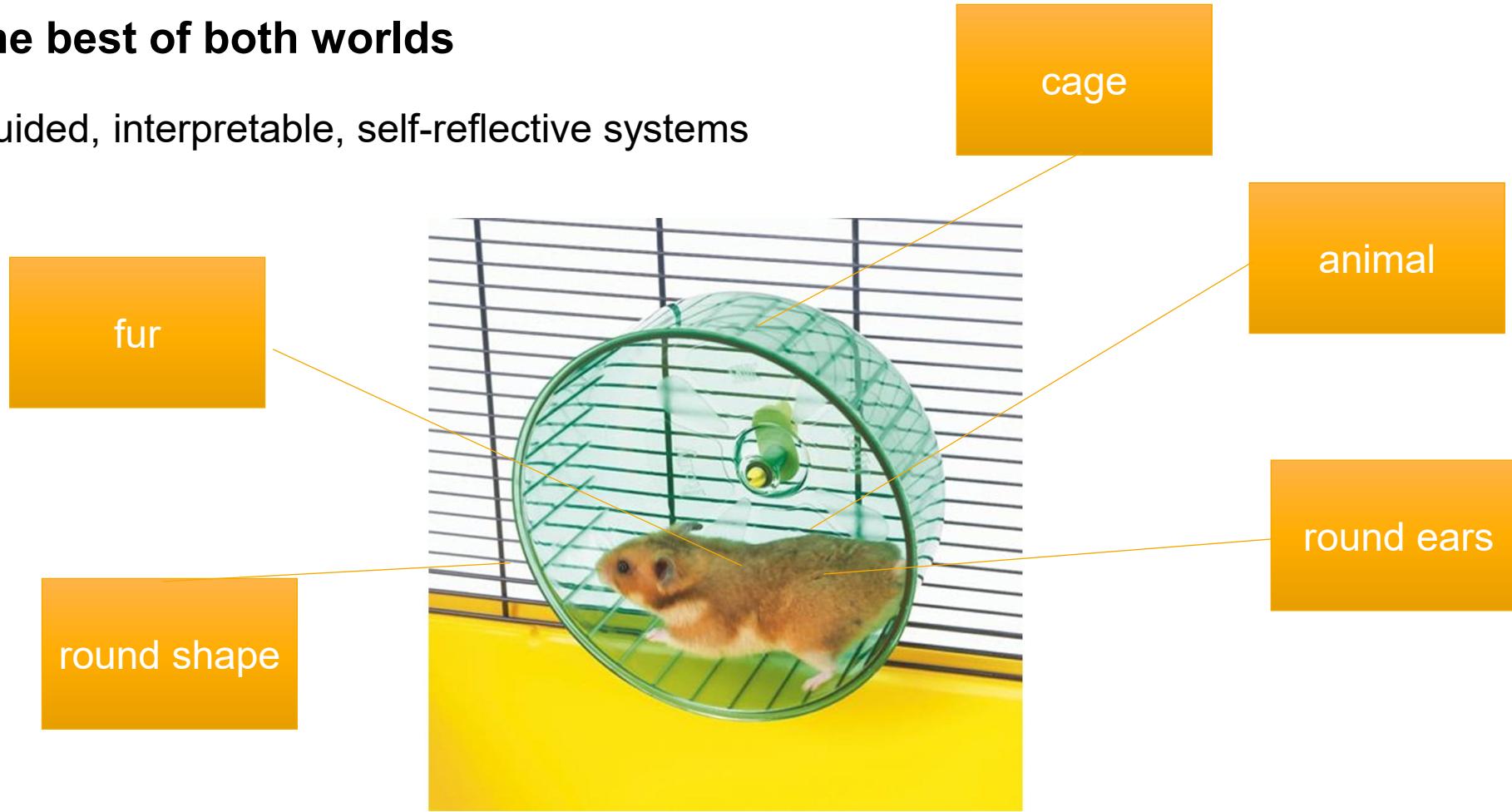
Propose multiple solutions for subtask 1. Evaluate the suggestions. Proceed with subtask 2 with the knowledge from the prior step. Derive the final answer.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... & Resnik, P. (2024). The prompt report: a systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.

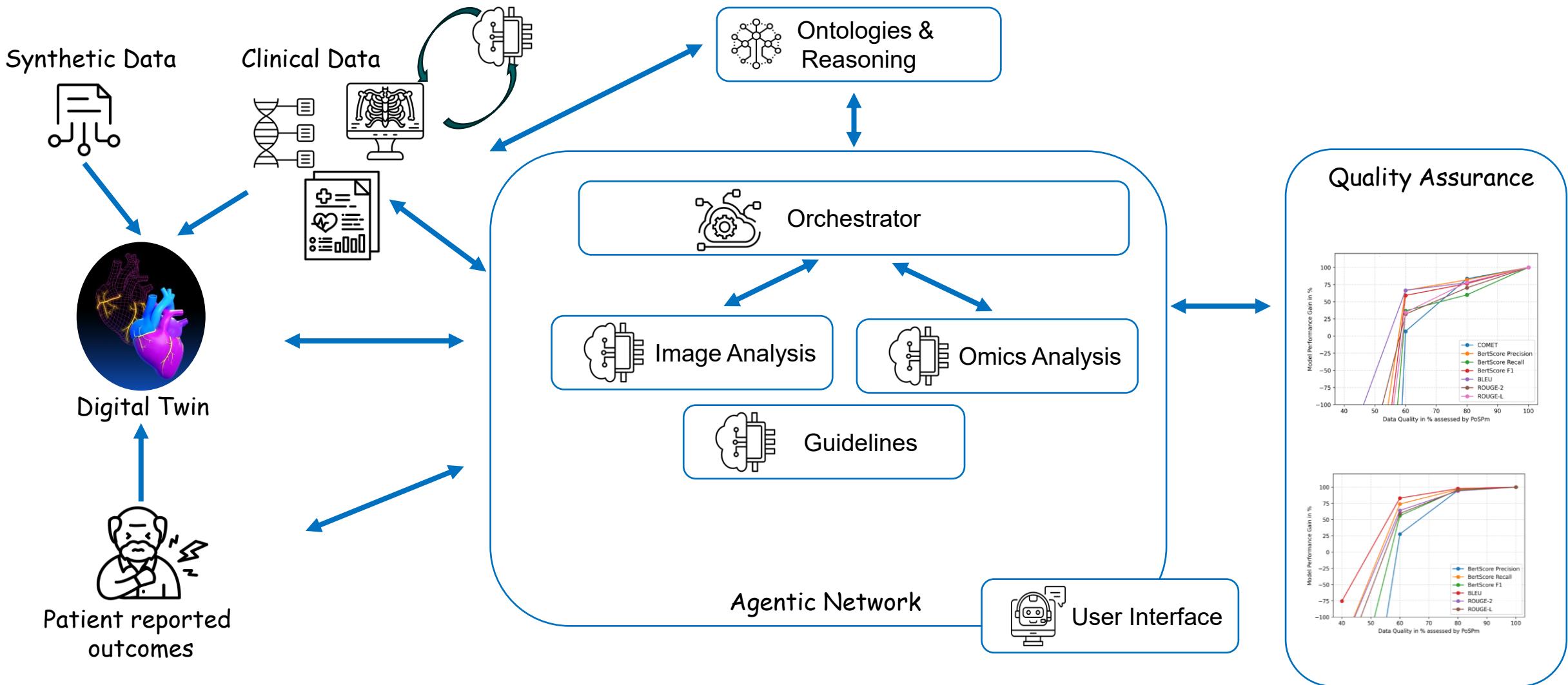
* <https://medium.com/contact-research/think-before-you-speak-the-inherent-statelessness-of-large-language-models-ceff8ab20ff1>, last accessed 3 June, 2025

Combining the best of both worlds

- Knowledge-guided, interpretable, self-reflective systems



Agentic AI



Conclusion & next steps



LLMs helpful to support FAIR RDM

- Understand and annotate data sets
- Search for existing ontologies, combination and generation of new ontologies
- Leverage discovery, interoperability, and reuse



But...

- Manual checks and some knowledge about ontologies and KGs still necessary
- Understanding the data is limited, but improvements, such as prompt-engineering, fine-tuning can help



Next steps

- Intensive analysis in various domains
- Integration of methodologies for ontology generation and quality checks
- Knowledge Graph reuse, versioning, logging
- Integration of symbolic knowledge & reasoning

Contact us



JProf. Dr. Sandra Geisler

Principal Investigator IoP
Data Stream Management and Analysis (DSMA)
Phone: +49 241 80-21508
E-Mail: geisler@cs.rwth-aachen.de



Soo-Yon Kim, M.Sc.

Data Steward IoP
Data Stream Management and Analysis
(DSMA)
Phone: +49 241 80-21514
E-Mail: soo-yon.kim@cs.rwth-aachen.de



Prof. Dr. Stefan Decker

Principal Investigator IoP
Databases and Information Systems (DBIS)
Phone: +49 1590 4338009
E-Mail: decker@dbis.rwth-aachen.de



Acknowledgment

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612.

References

- Feger, S. S., Dallmeier-Tiessen, S., Schmidt, A., & Woźniak, P. W. (2019, May). Designing for reproducibility: A qualitative study of challenges and opportunities in high energy physics. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14). <https://doi.org/10.1145/3290605.3300685>
- Jussen, I., Möller, F., Schweihoff, J., Gieß, A., Giussani, G., & Otto, B. (2024). Issues in inter-organizational data sharing: Findings from practice and research challenges. *Data & Knowledge Engineering*, 102280. <https://doi.org/10.1016/j.datak.2024.102280>
- Kim, S. Y., Hillemacher, S., Decker, S., Rumpe, B., & Geisler, S. (2023). Designing and Implementing Practicable Data Management Plans in Large-Scale Projects. *Bausteine Forschungsdatenmanagement*, (3), 1-12. <https://doi.org/10.17192/bfdm.2023.3.8571>
- Geisler, S. and Kim, S.-Y., "Unlocking the Potential: LLMs Transforming Research Data Management," in From Data to Diamonds - Empowering Research with AI and RDM, Aachen, Germany, Nov. 2023, pp. 1–35. <https://doi.org/10.18154/RWTH-2023-10498>
- Moon, J., Gelbich, D., Becker, M., Niemitz, P., & Bergs, T. Predicting fine blanking process signals from sheet metal thickness. *Materials Research Proceedings*, Vol. 41, pp 1436-1445. <https://doi.org/10.21741/9781644903131-159>
- Stuart, D., Baynes, G., Hrynaszkiewicz, I., Allin, K., Penny, D., Lucraft, M., et al. (2018). Whitepaper: Practical challenges for researchers in data sharing. Springer Nature, <https://doi.org/10.6084/m9.figshare.5975011.v1>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36, 11809-11822.