University of Amsterdam
Informatics Institute

INDE lab
INtelligent Data Engineering

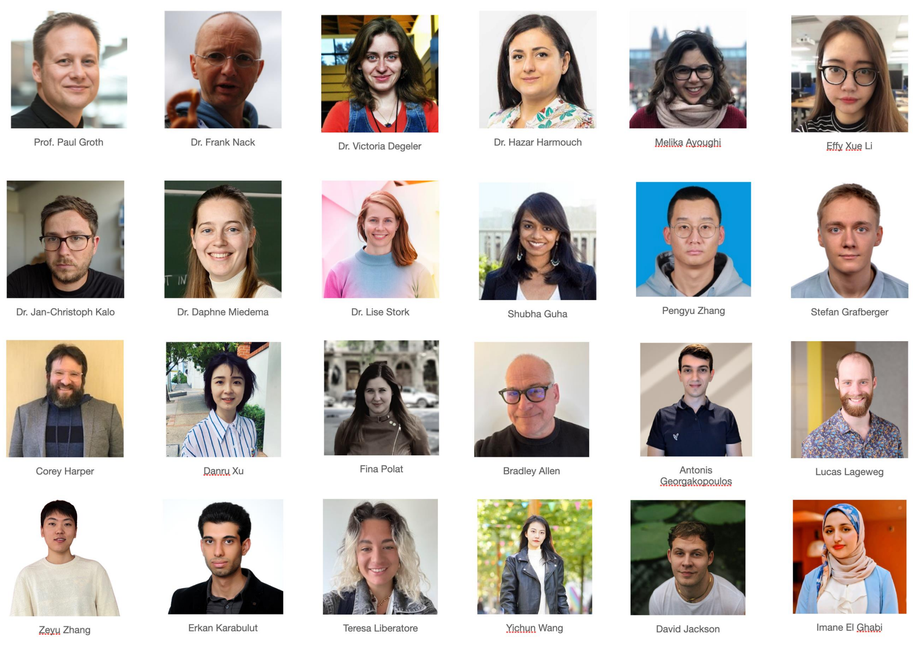
AI Systems for Data Work

Intelligent Data Engineering Lab (INDE lab)
indelab.org

Mission of INDE lab & The Team

We investigate intelligent systems that support people in their work with data and information from diverse sources.

In this area, we perform applied and fundamental research informed by empirical insights into data science practice.

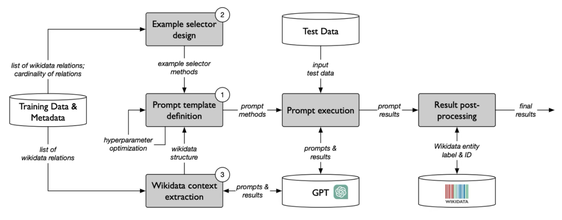


Research Topics

Automated Knowledge Graph Construction
(e.g. building knowledge graphs (KGs) from video and text; architectures for integrating KGs and LLMs; entity linking in temporal KGs)

Context Aware Data Systems
(e.g. rule learning & digital twins, understanding SQL usage in the wild, human - AI workflows)

Data Management for Machine Learning
(e.g. data quality assessment; using data provenance for ML debugging, data handling impact on ML models, data search)



Knowledge-centric Prompt Composition for Knowledge Base Construction from Pre-trained Language Models.
Xue Li et al.
LM-KBC 2024

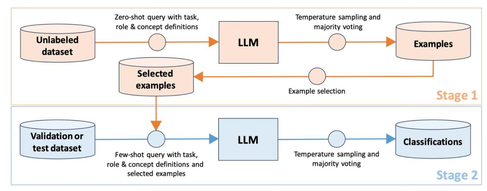
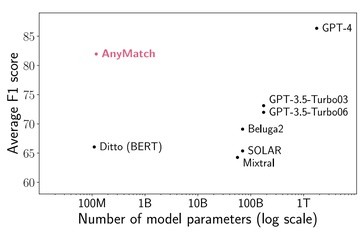


Figure 1: SHROOM-INDElab system workflow.

SHROOM-INDElab at SemEval-2024 Task 6: Zero- and Few-Shot LLM-Based Classification for Hallucination Detection.
Bradley Allen, Fina Polat, and Paul Groth. 2024.
SemEval 2024.

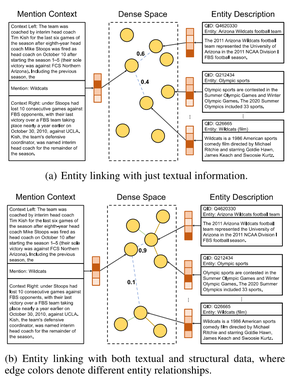
Examples of Recent Work

Efficient LLMs for Tabular Data



AnyMatch--Efficient Zero-Shot Entity Matching with a Small Language Model
Z Zhang, P Groth, I Calixto, S Schelter
arXiv preprint arXiv:2409.04073

Temporarly Robust Entity Linking

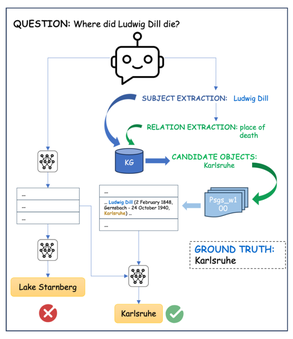


(a) Entity linking with just textual information.

(b) Entity linking with both textual and structural data, where edge colors denote different entity relationships.

TIGER: Temporally Improved Graph Entity Linker
P Zhang, C Cao, P Groth
ECAI 2024

Enhancing Neural QA with Knowledge Graphs



Retrieval-based Question Answering with Passage Expansion Using a Knowledge Graph.
B Kruit, Y Xu, JC Kalo
LREC-COLING 2024

Data Quality Assesment & ML

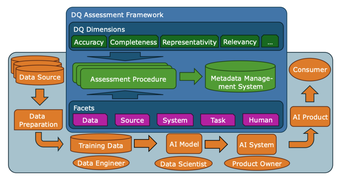
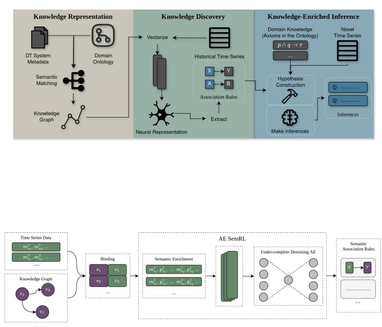


Figure 2: DQ assessment for an AI pipeline.

Data Quality Assessment: Challenges and Opportunities
S Mohammed, H Harmouch, F Naumann, D Srivastava
arXiv preprint arXiv:2403.00526, 2024

Digital Twins & Knowledge



3K: Knowledge-Enriched Digital Twin Framework.
E Karabulut, P Groth, V Degeler
LongeviIoT 2024

Knowledge Engineering & LLMs



Knowledge Engineering Using Large Language Models
BP. Allen, L Stork, P Groth.
Transactions on Graph Data & Knowledge

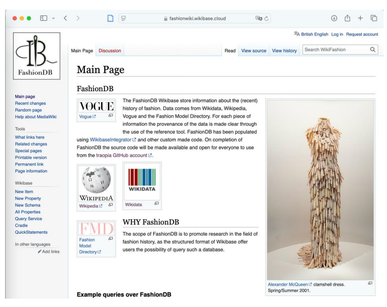
Challenge: Evaluation

- Empirical insights requires data, e.g.: observational data from user practice; survey response; evaluation environment; high quality evaluation / testing datasets, …
- Performing evaluations whether with users or computationally, e.g.:
  - LLMs makes out of distribution data harder to find;
  - LLMs generated answers appear right but need extra humana evaluation (i.e not in testsets);
  - LLM evals have potential issues;
  - comparing to proprietary models;
  - finding the right user population.
- **How are you tackling evaluation challenges?**

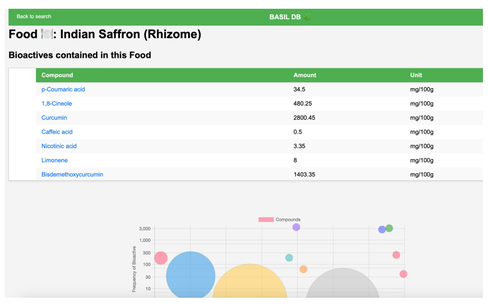
Data

Come talk to us about the databases and datasets we are building. We are also curious about the data you're making. What challenges do you have? Let us know

FashionDB



BasilDB



WDN Datasets



GitTables

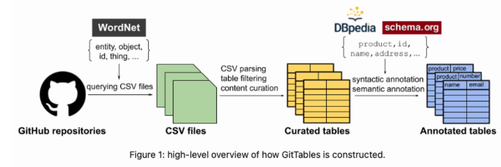


Figure 1: high-level overview of how GitTables is constructed.

GitTables: A Large-Scale Corpus of Relational Tables.
Madelon Hulsebos, Çağatay Demiralp, and Paul Groth. 2022.
DOI:https://doi.org/10.48550/arXiv.2106.07258