Between Data Lakes and Research Data Management –

Data Engineering Tasks for the Next Decade

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Universität Regensburg

My main research interests

Abstraction over data

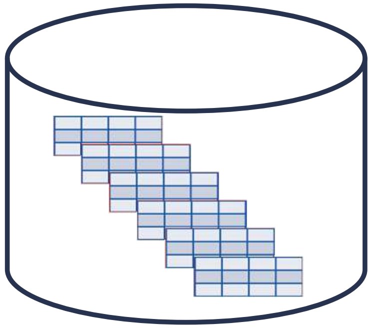
Evolution of data structures and processes

Automation of data engineering processes

… and abstraction you can also see in my paintings …

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

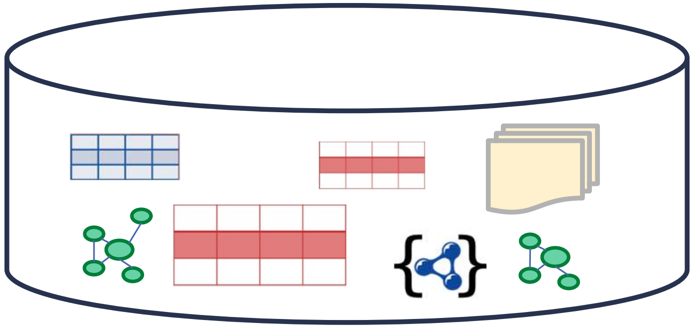
Motivation: The Academic Point of View onto Research Data



Structured datasets

- Complete, consistent, regular and relational-like

.. and in Reality (Data Lake)



Research Data

- Heterogeneous data formats, systems, and schemas
- Sometimes noisy, error-prone, and incomplete

NFDI: Definition of Standards (to enable Data Integration) and Processes (to guarantee Data Quality)

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

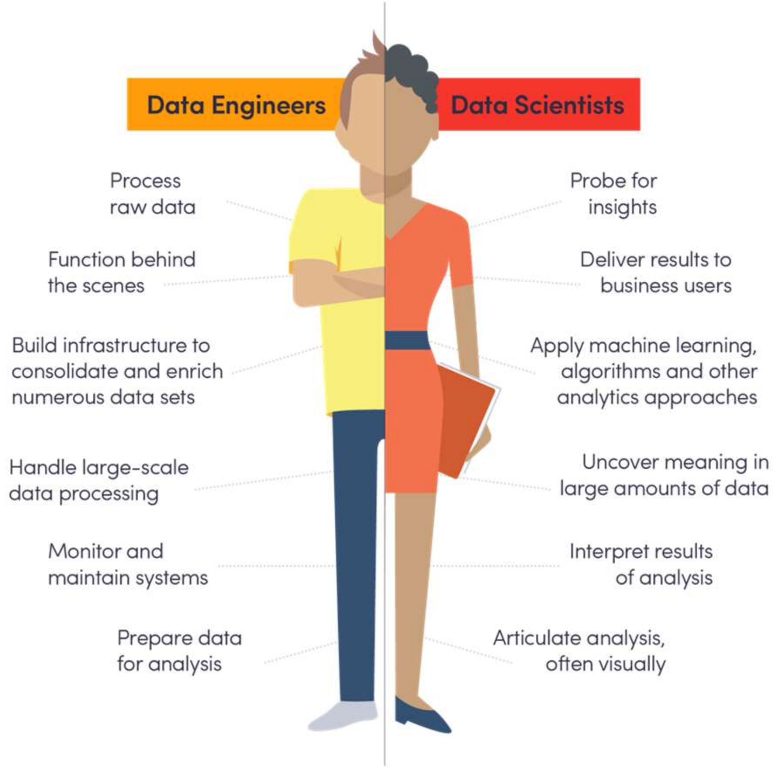
What are the next Data Engineering Tasks? – Structure of the Talk

Presence and Future (ongoing Research Tasks):

- First Generation: **Data Preprocessing**
- Second Generation: Data Engineering **Pipelines**

Future:

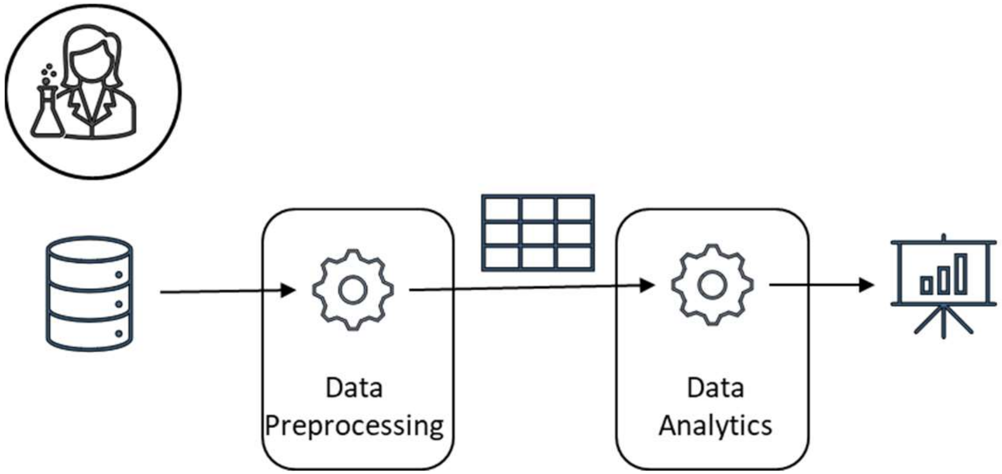
- Third Generation: **Adaptive** Data Engineering Pipelines
- Fourth Generation: **Automatic Data Curation** based on Recommender Technologies



First Generation: Data Preprocessing Algorithms

Data Engineering subtasks:

1. Data Selection (Sampling)
2. Data Understanding
3. Cleaning and Data Correction
4. Data Transformation



Data Engineering is **time-consuming** (**80% of the overall effort**), error-prone and expensive

- Choice, parametrization and application of algorithms → manual task
- Skills in computer science are needed for this data preprocessing

First Generation: Data Preprocessing Algorithms

Overview on data engineering subtasks

1. **Data Understanding**
   - Schema Extraction
   - Column Type Inference
   - Inference of Integrity Constraints/Pattern
   - Data Exploration

2. **Cleaning and Data Correction**
   - Concept Shift Detection
   - Bias Detection
   - Outlier Detection and Correction
   - Duplicate Elimination
   - Missing Value Imputation

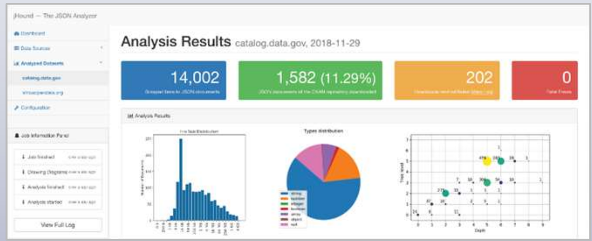
3. **Data Transformation**
   - Matching and Mapping
   - Data Integration
   - Datatype Transformation
   - Transformation between different Data Models

Our own work in this field

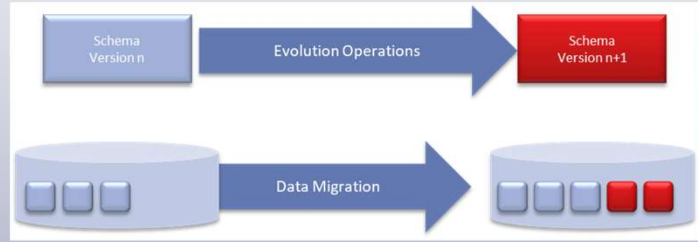
Abstraction, Evolution

Data Understanding – Exploring characteristics of JSON data

M. Möller, N. Berton, M. Klettke, S. Scherzinger, U. Störl: *jHound: Large-Scale Profiling of Open JSON Data*. BTW 2019

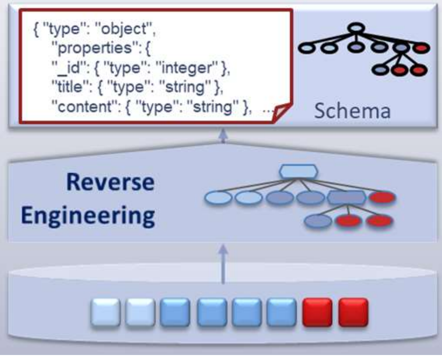


Schema Evolution and **Data Migration***

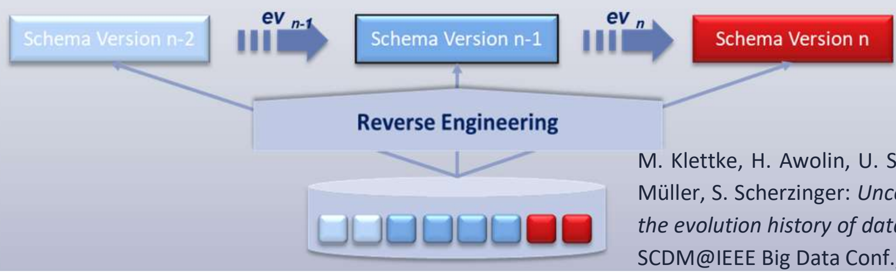


Reverse Engineering:
Schema Extraction*

M. Klettke, U. Störl, S. Scherzinger: *Schema Extraction and Structural Outlier Detection for NoSQL Data Stores*. BTW 2015

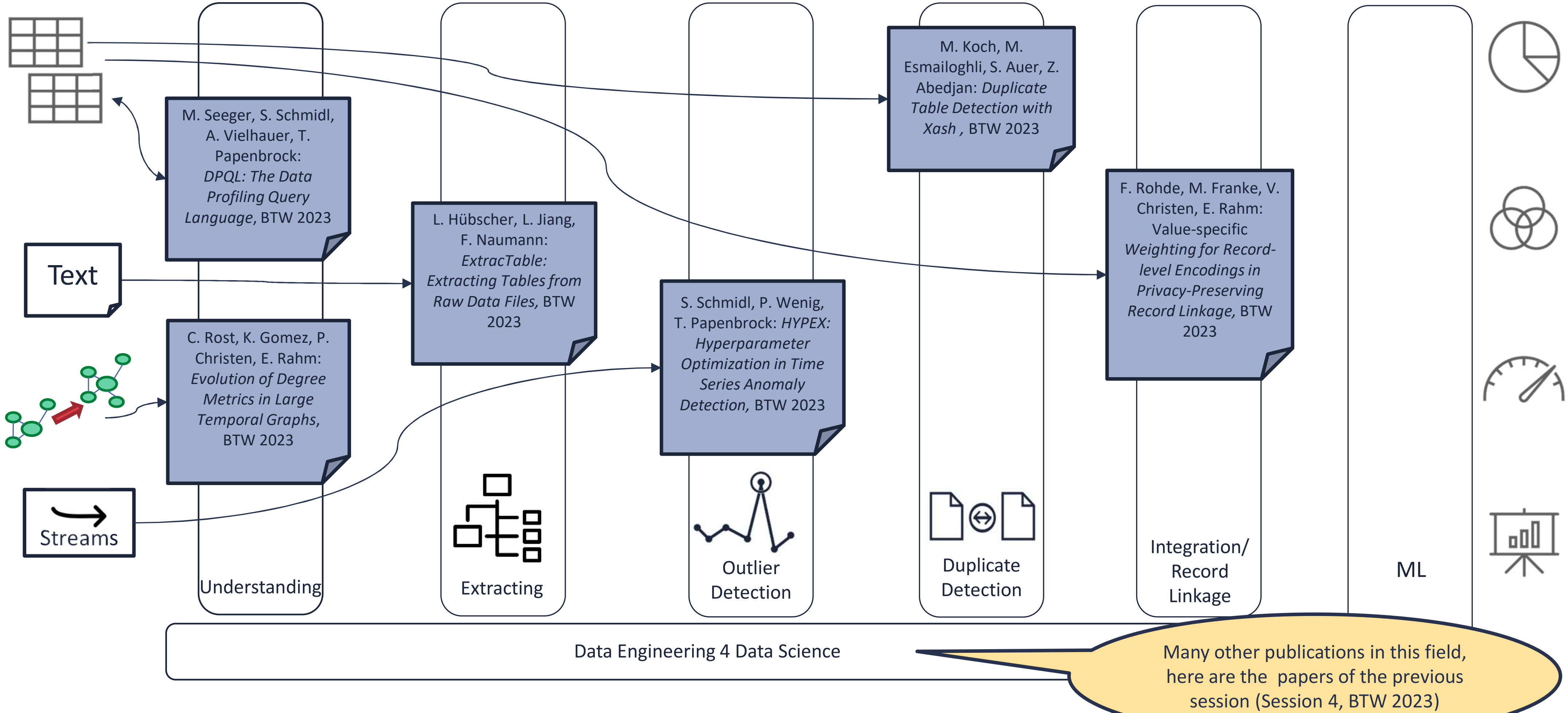


Reverse Engineering: Schema Version Extraction*



M. Klettke, H. Awolin, U. Störl, D. Müller, S. Scherzinger: *Uncovering the evolution history of data lakes*. SCDM@IEEE Big Data Conf., 2017

* together with Stefanie Scherzinger and Uta Störl

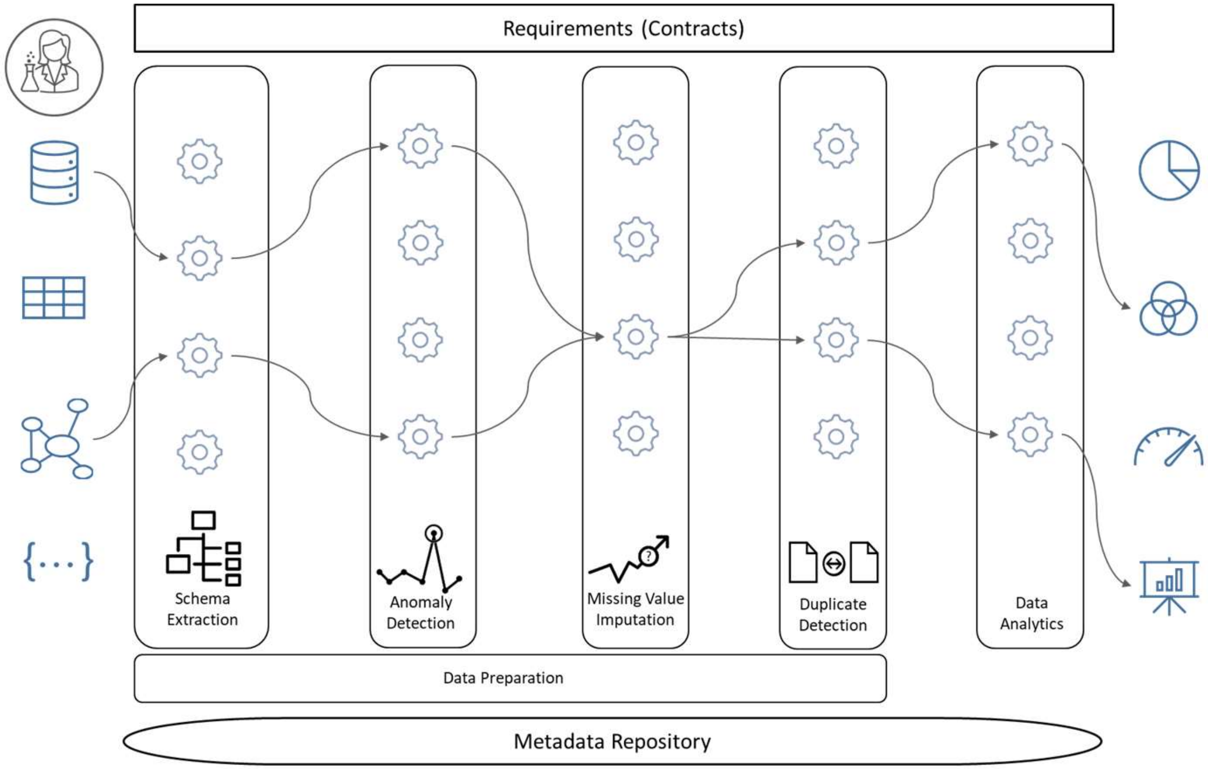

M. Seeger, S. Schmidl, A. Vielhauer, T. Papenbrock: *DPQL: The Data Profiling Query Language*, BTW 2023
Text
C. Rost, K. Gomez, P. Christen, E. Rahm: *Evolution of Degree Metrics in Large Temporal Graphs*, BTW 2023
Streams
Understanding
L. Hübscher, L. Jiang, F. Naumann: *ExtracTable: Extracting Tables from Raw Data Files*, BTW 2023
Extracting
S. Schmidl, P. Wenig, T. Papenbrock: *HYPEX: Hyperparameter Optimization in Time Series Anomaly Detection*, BTW 2023
Outlier Detection
M. Koch, M. Esmailoghli, S. Auer, Z. Abedjan: *Duplicate Table Detection with Xash* , BTW 2023
Duplicate Detection
F. Rohde, M. Franke, V. Christen, E. Rahm: Value-specific *Weighting for Record-level Encodings in Privacy-Preserving Record Linkage*, BTW 2023
Integration/ Record Linkage
ML
Data Engineering 4 Data Science
Many other publications in this field, here are the papers of the previous session (Session 4, BTW 2023)

Second Generation: Data Engineering Pipelines

Data Engineering Tools:

- For each data engineering task **many implementations** are available
  - for different data models/
  - for different data characteristics
  - applying different methods
- Toolsets are providing a **variety of algorithms**

Manual task: **select and compose** the **algorithms** in pipelines



Second Generation: Data Engineering Pipelines

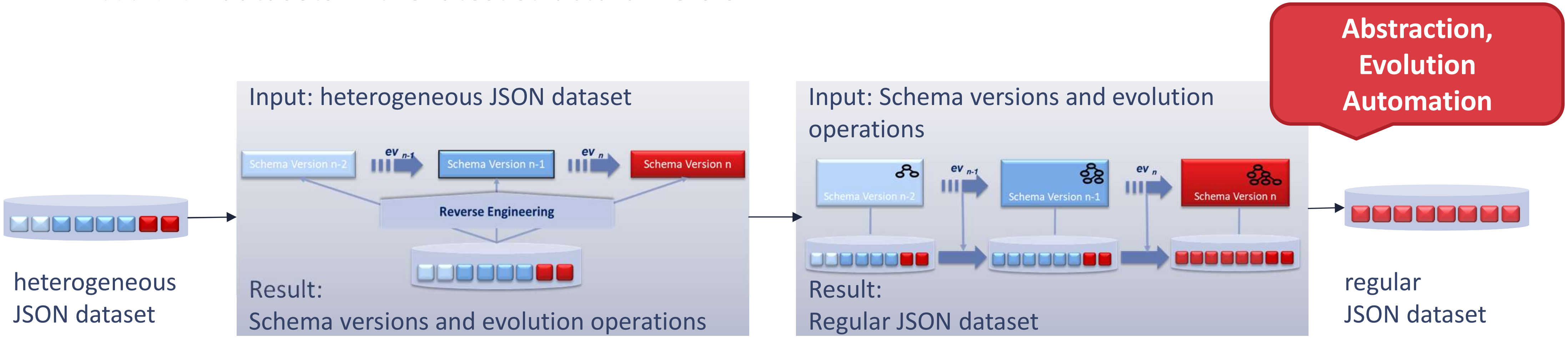
- **Pipelining idea** in:
  - **ETL** (processes), **Machine learning** (pipelines), **Data science** (pipelines)
- Some of these available **toolsets** are:
  - ETL tools for Data warehouses and BI tools (e.g. Talend, Tableau Prep, Qlik, …)
  - Python and data science libraries (NumPy, pandas, SciPy, scikit-learn, feature-engineering)
  - Data preparation parts in data mining tools (Weka, RapidMiner)
  - Data wrangling/ Data Lake processing (Snowflake, IBM InfoSphere DataStage, luigi)
- Ongoing Research:
  - Sebastian Baunsgaard, Matthias Boehm, Ankit Chaudhary, Behrouz Derakhshan, Stefan Geißelsöder, Philipp M. Grulich, Michael Hildebrand, Kevin Innerebner, Volker Markl, Claus Neubauer, Sarah Osterburg, Olga Ovcharenko, Sergey Redyuk, Tobias Rieger, Alireza Rezaei Mahdiraji, Sebastian Benjamin Wrede, Steffen Zeuch***: ExDRa: Exploratory Data Science on Federated Raw Data.*** SIGMOD Conference 2021
  - Patrick Damme, Marius Birkenbach, Constantinos Bitsakos, Matthias Boehm, Philippe Bonnet, Florina M. Ciorba, Mark Dokter, Pawel Dowgiallo, Ahmed Eleliemy, Christian Faerber, Georgios I. Goumas, Dirk Habich, Niclas Hedam, Marlies Hofer, Wenjun Huang, Kevin Innerebner, Vasileios Karakostas, Roman Kern, Tomaz Kosar, Alexander Krause, Daniel Krems, Andreas Laber, Wolfgang Lehner, Eric Mier, Marcus Paradies, Bernhard Peischl, Gabrielle Poerwawinata, Stratos Psomadakis, Tilmann Rabl, Piotr Ratuszniak, Pedro Silva, Nikolai Skuppin, Andreas Starzacher, Benjamin Steinwender, Ilin Tolovski, Pinar Tözün, Wojciech Ulatowski, Yuanyuan Wang, Izajasz P. Wrosz, Ales Zamuda, Ce Zhang, Xiaoxiang Zhu: ***DAPHNE: An Open and Extensible System Infrastructure for Integrated Data Analysis Pipelines.*** CIDR 2022

Our own work: Combine Schema Version Extraction and Data Migration

Input: Set of JSON documents (in different structural versions)

Pipeline: Combining **Inference of Schema Versions and Evolution Operations** and **Data Migration**

Result: all **datasets in the latest structural version**

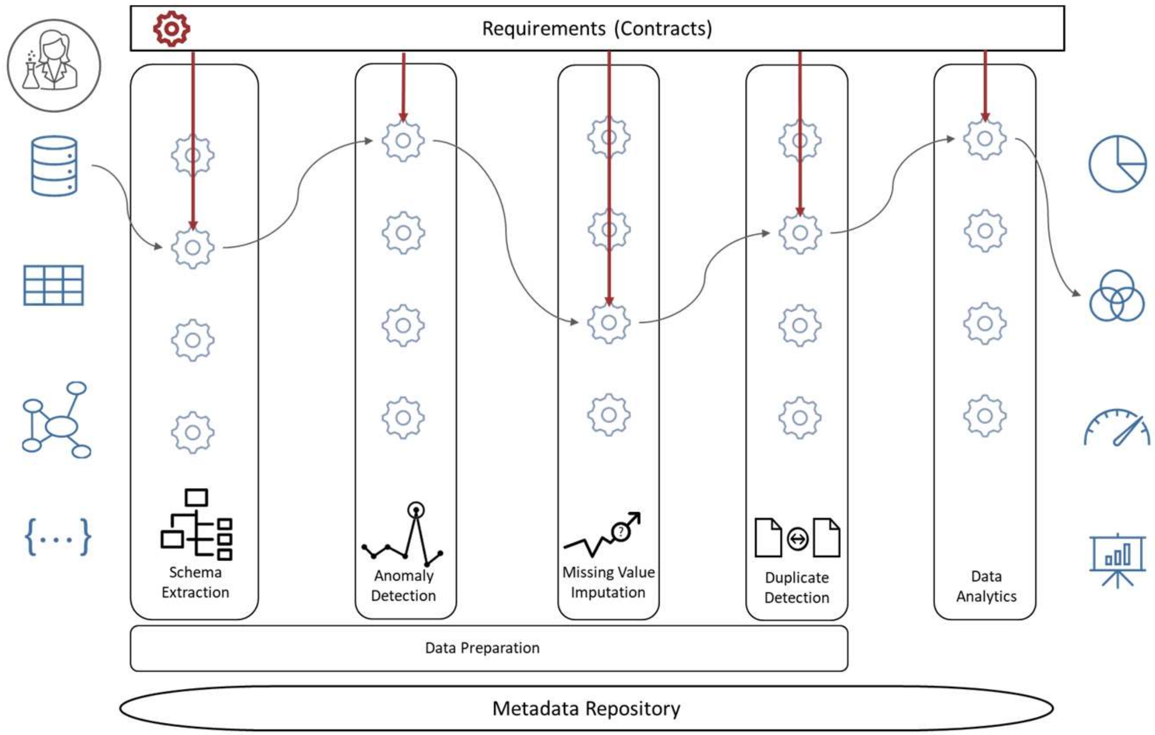


Third Generation: Intelligent Data Engineering Pipelines

I. Automatic orchestration
- **choice of algorithms**
- **composition** of data engineering algorithms to a workflow

II. Monitoring of the data
- **Detection of data changes**, changes of distributions, and **data bias**
- Monitoring of **evolving workflows**

Other terms for similar idea:
- **data democratization, AutoML**



Third Generation: Automatic workflow orchestration

Building blocks needed for this task:

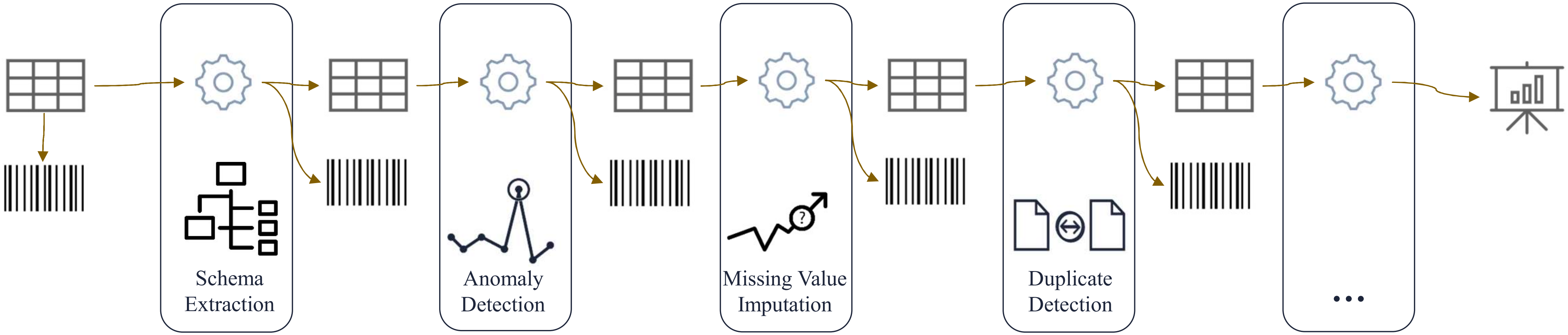
1. **Formal specification** of the **requirements**
2. **Formal metrics** (e.g. schema, data-types, pattern, constraints, data quality measures) of the datasets
3. **Formal characteristics for each data engineering algorithm (e.g. with pre- and postconditions)**
4. **Matches** between **requirements and algorithm characteristics**
5. **Opportunity to evaluate** workflows

Ongoing work:

- Patrick Damme, Marius Birkenbach, Constantinos Bitsakos, Matthias Boehm, Philippe Bonnet, Florina M. Ciorba, Mark Dokter, Pawel Dowgiallo, Ahmed Eleliemy, Christian Faerber, Georgios I. Goumas, Dirk Habich, Niclas Hedam, Marlies Hofer, Wenjun Huang, Kevin Innerebner, Vasileios Karakostas, Roman Kern, Tomaz Kosar, Alexander Krause, Daniel Krems, Andreas Laber, Wolfgang Lehner, Eric Mier, Marcus Paradies, Bernhard Peischl, Gabrielle Poerwawinata, Stratos Psomadakis, Tilmann Rabl, Piotr Ratuszniak, Pedro Silva, Nikolai Skuppin, Andreas Starzacher, Benjamin Steinwender, Ilin Tolovski, Pinar Tözün, Wojciech Ulatowski, Yuanyuan Wang, Izajasz P. Wrosz, Ales Zamuda, Ce Zhang, Xiaoxiang Zhu*: **DAPHNE: An Open and Extensible System Infrastructure for Integrated Data Analysis Pipelines**.* CIDR 2022
- Valerie Restat, Meike Klettke, Uta Störl: ***Towards a Holistic Data Preparation Tool***. EDBT/ICDT Workshops 2022
- Valerie Restat, Meike Klettke, Uta Störl: ***"FAIR" is not enough – A Metrics Framework to ensure Data Quality through Data Preparation.*** Workshop Data Engineering for Data Science (DE4DS)@BTW, 2023

Third Generation: Monitoring of the Data in Data Engineering Pipelines

- Monitoring of **data changes, changes of distributions, data bias** in Data Engineering workflows



- Definition of additional metadata for the data (barcode, data passports) and monitoring of these
  - Stefan Grafberger, P Groth, J Stoyanovich, S Schelter: *Data distribution debugging in machine learning pipelines,* VLDB Journal 31 (5), 2022
  - Meike Klettke, Adrian Lutsch, Uta Störl: *Kurz erklärt: Measuring Data Changes in Data Engineering and their Impact on Explainability and Algorithm Fairness*, Datenbank-Spektrum, 21(3): S. 245-249 (2021)
  - Erik Kleinsteuber, Samira Babalou, Birgitta König-Ries: *A Provenance Management Framework for Knowledge Graph Generation in a Web Portal,* DE4DS@BTW 2023

Third Generation: Monitoring of the Data in Evolving Data Engineering Pipelines

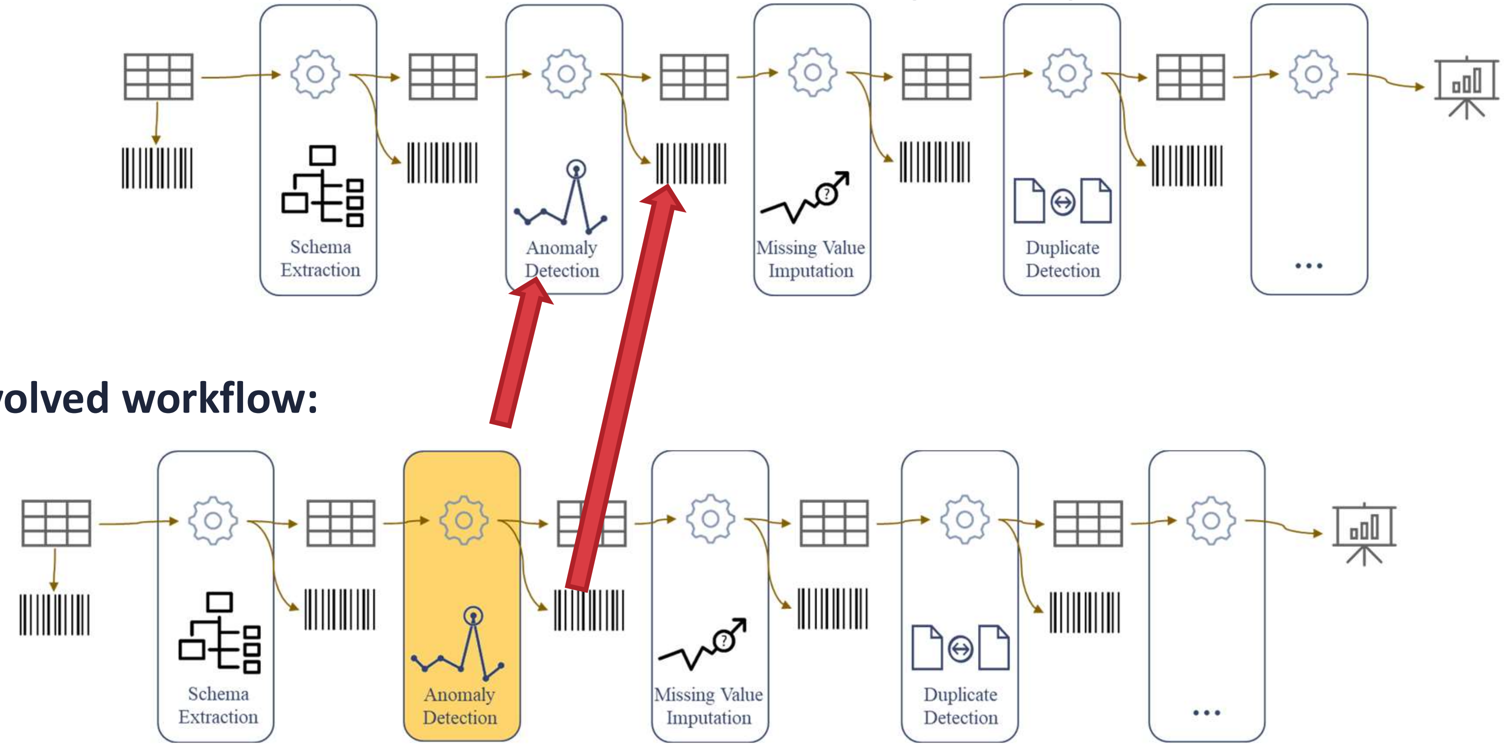
Monitoring of **evolving data engineering workflows**

- Reasons for evolution:
  - Replacement of algorithms
  - Correction of errors
  - Adding new algorithms
  - ...
- Monitoring changes
  - Applying the comparison of data **between** different workflows

Sihem Amer-Yahia, *Commodifying Data Exploration*, Keynote BTW 2023: **reuse vs. re-train**

Reference point: old version of a data engineering workflow

Evolved workflow:



Our own (future) plans in Data Engineering Pipelines

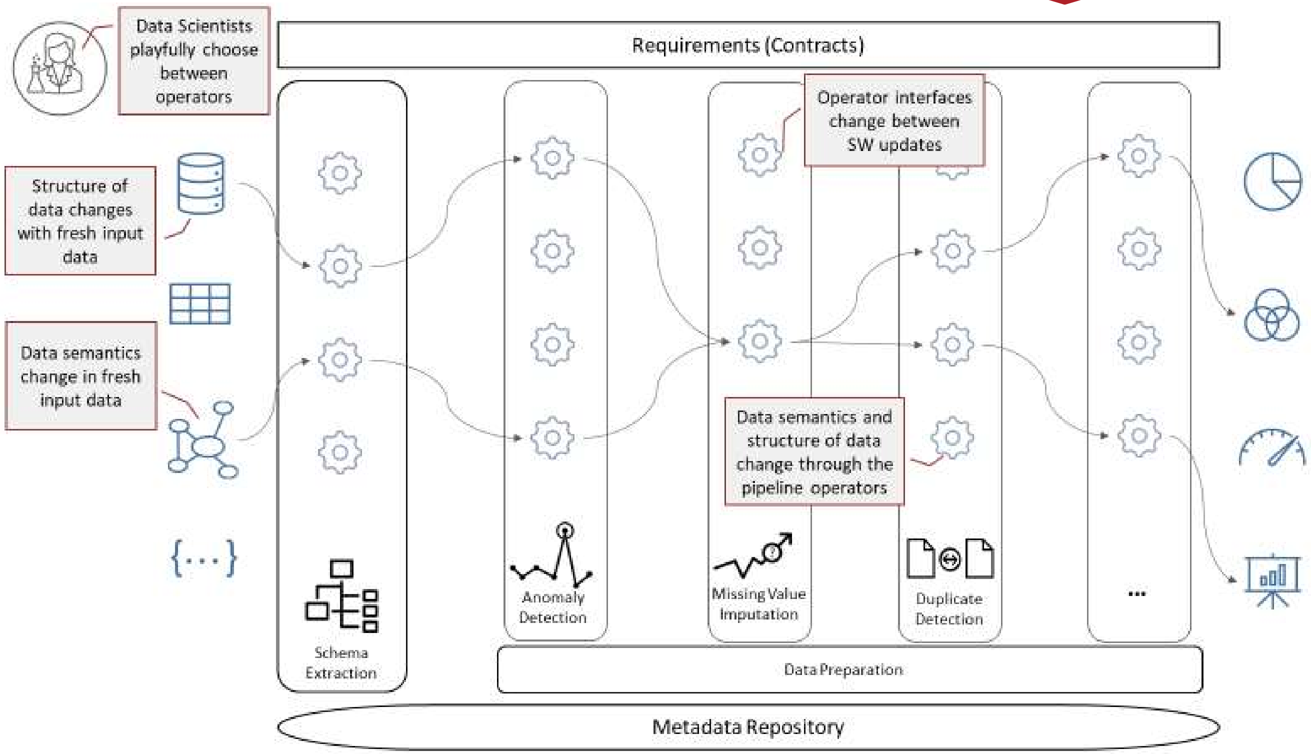
Evolution, Automation

Focussed on

- evolving data
- evolving operators
- evolving pipelines

Development of tools to

- **monitor data engineering workflows** (datasets, data characteristics, distributions)
- **monitor and evolution of evolving workflows**

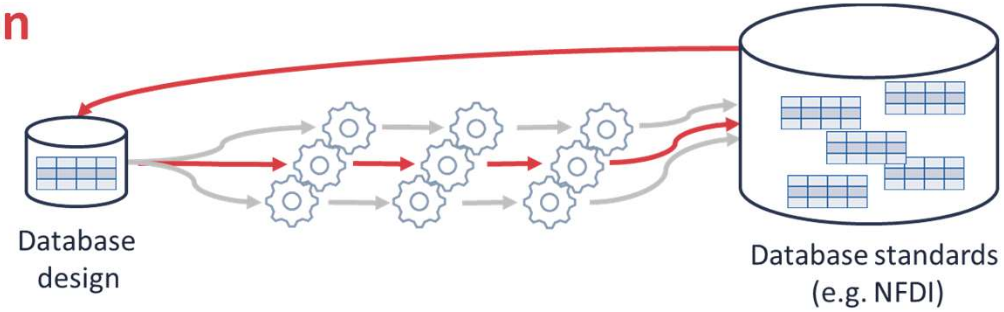


Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Fourth Generation: Automatic Data Curation

Developing tools for supporting **data curation**

- Let’s imagine: **NFDI** has defined the **standards** for research data for the different scientific fields
- Tools which guide design tasks would be nice



How can it be done?

- Data curations needs an understanding of the semantics of data
- Algorithms cannot understand attribute names, abbreviations, ect.

→ **workaround** is necessary

Usage of

- **recommender systems** combined with
- **similarity matching and mapping**

Aim:

…	C	E	F	…

during database design
- suggestion of further attributes,
- data types,
- constraints

(like an **autocomplete**)

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Recommender Systems for Database Design

Input: Set of Database schemas

Method:

- **Step 1 (optional) Preprocessing:** combining similar attributes
- **Step 2:** Creation of a **user/item matrix**, here **database/attribute matrix**
- **Step 3:** Calculation of **association rules**
- **Step 4 (optional) Postprocessing: combining similar attributes**
- **Step 5:** calculation of **support and confidence** values for the combined association rules
- **Step 6:** usage of the association rules for making design suggestions

A	B1	C	F1	…

B2	C	E	F2	…

C	F3	…

C	D	E	F2	…

A	C	E	F1	…

	A	B1	B2	C	D	E	F1	F2	F3
D1	X	X		X			X		
D2			X	X		X		X	
D3				X					X
D4				X	X	X		X	
D5	X			X		X	X		

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Recommender Systems for Database Design: Deriving Association Rules

	A	B1	B2	C	D	E	F1	F2	F3
D1	X	X		X			X		
D2			X	X		X		X	
D3				X					X
D4				X	X	X		X	
D5	X			X		X	X		

Let's assume:

min-support: 3, min-confidence: 60%

Association Rules:

- **C → E**, s:3(60%), c: 60%
- **E → C**, s:3(60%), c:100%

(s: support, c: confidence)

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Recommender Systems for Database Design: Deriving Association Rules

	A	B1	B2	C	D	E	F1	F2	F3
D1	X	X		X			X		
D2			X	X		X		X	
D3				X					X
D4				X	X	X		X	
D5	X			X		X	X		

Let's assume:
min-support: 3, min-confidence: 60%

- C → F1, s:2(40%), c: 40%
- C → F2, s:2(40%), c: 40%
- C → F3, s:1(20%), c: 20%

- F1 → C, s:2(40%), c: 100%
- F2 → C, s:2(40%), c: 100%
- F3 → C, s:1(20%), c: 100%

- E → F1, s:1(20%), c: 33%
- E → F2, s:2(40%), c: 66%

- F1 → E, s:1(20%), c: 50%
- F2 → E, s:2(40%), c: 100%

(s: support, c: confidence)

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Recommender Systems for Database Design: Deriving Association Rules

	A	B1	B2	C	D	E	F1	F2	F3
D1	X	X		X			X		
D2			X	X		X		X	
D3				X					X
D4				X	X	X		X	
D5	X			X		X	X		

Similar attributes (B1, B2)

Similar attributes (F1, F2, F3)

Let's assume:
min-support: 3, min-confidence: 60%

- C → F1
- C → F2
- C → F3

→ **C → F (F1,F2,F3)** s:5(100%), c: 100%

- F1 → C
- F2 → C
- F3 → C

→ **F (F1,F2,F3) → C** s:5(100%), c: 100%

- E → F1
- E → F2

→ **E → F (F1,F2,F3)** s:3(60%), c: 100%

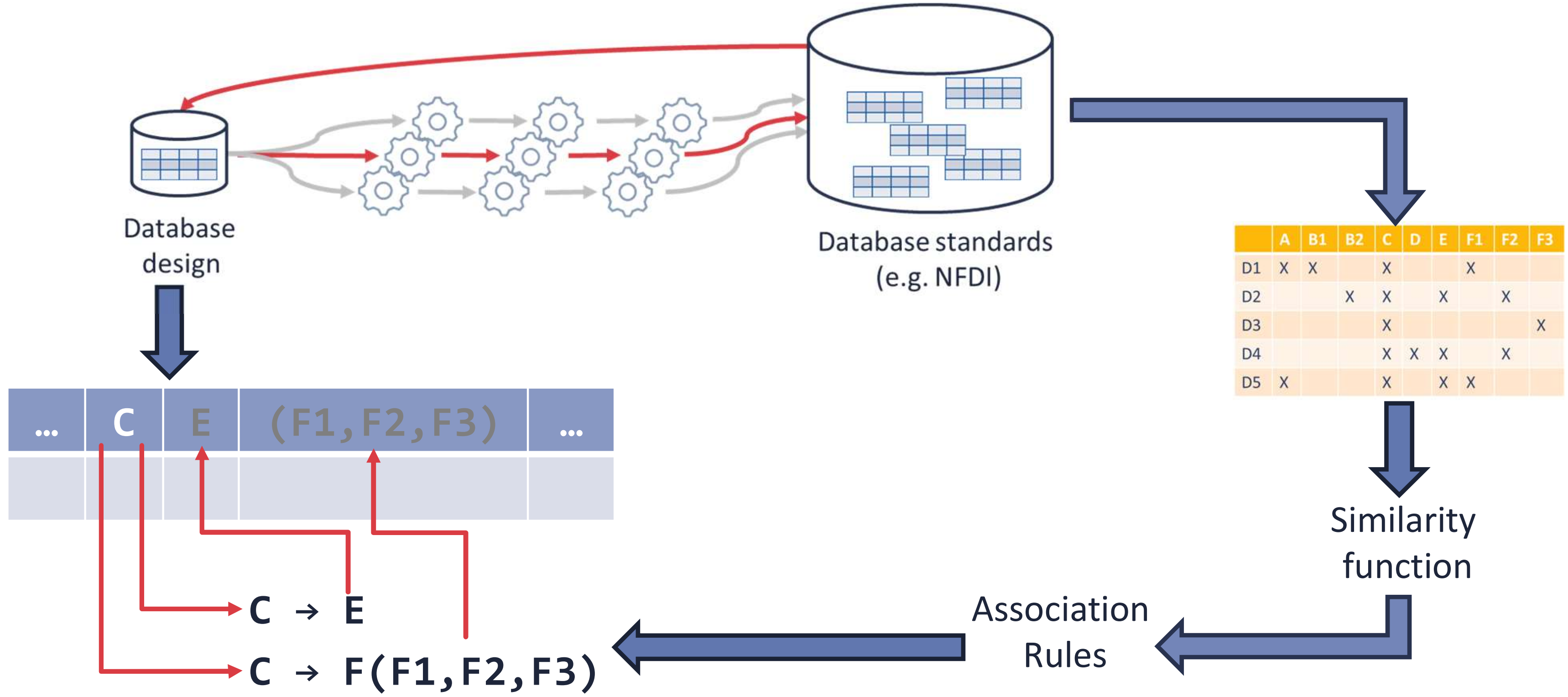
- F1 → E
- F2 → E

→ **F (F1,F2,F3) → E** s:3(60%), c: 60%

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Usage of the Association Rules for Database Designs



	A	B1	B2	C	D	E	F1	F2	F3
D1	X	X		X			X		
D2			X	X		X		X	
D3				X					X
D4				X	X	X		X	
D5	X			X		X	X		

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Overview of the Method and Differences to the Standard Method

Method:

- Using **available database standards** as well as **available database designs** for collaborative filtering
- Usage of **different data models** is possible
- Generating a user/item matrix, here database/attribute matrix
- Using a **similarity function** for finding similar items (attributes)
- Calculating association rules with **support and confidence**, (considering min-support and min-confidence)

Another modification:

- Different weights for different rows in the user/item matrix (also known as database/attribute matrix)
- **Standards are more reliable then other database schemas**

Characteristics:

- Like each collaborative filtering: learning system

Universität Regensburg

Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

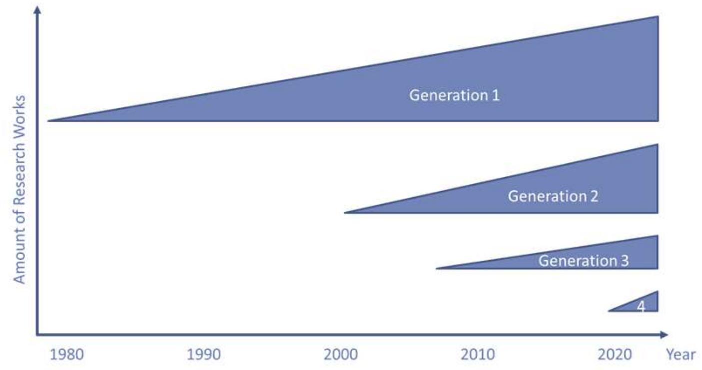
Conclusion and Future Work

Data Engineering Research:

- Generation 1: **Data Engineering algorithms** for each subtask
- Generation 2: **Pipelines** for defining whole data engineering processes
- Generation 3: **Advisor components** in these data engineering pipelines
- Generation 4: **Intelligent data curation**

In each of these subparts, **many open research tasks**, mainly

- **Monitoring data engineering workflows, esp. evolving data engineering workflows**
- **Database design tools based on recommender systems** combined with **similarity matching and mapping**



Estimation which bases on a very shallow keyword search in dblp

References (1/3)

- Abedjan, Z., Golab, L., Naumann, F., Papenbrock, T.: *Data Profiling*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers (2018)
- Baazizi, M.A., Colazzo, D., Ghelli, G., Sartiani, C.: *Parametric schema inference for massive JSON datasets*. VLDB J.28(4) (2019)
- Sebastian Baunsgaard, Matthias Boehm, Ankit Chaudhary, Behrouz Derakhshan, Stefan Geißelsöder, Philipp M. Grulich, Michael Hildebrand, Kevin Innerebner, Volker Markl, Claus Neubauer, Sarah Osterburg, Olga Ovcharenko, Sergey Redyuk, Tobias Rieger, Alireza Rezaei Mahdiraji, Sebastian Benjamin Wrede, Steffen Zeuch: *ExDRa: Exploratory Data Science on Federated Raw Data*. SIGMOD Conference 2021
- Boehm, M., Kumar, A., Yang, J.: *Data Management in Machine Learning Systems*. Synthesis Lectures on DataManagement. Morgan & Claypool Publishers (2019)
- Chandola, V., Banerjee, A., Kumar, V.: *Outlier detection: A survey*. ACM Computing Surveys14 (2007)
- Dimitriadou, K., Papaemmanouil, O., Diao, Y.: *Explore-by-example: an automatic query steering framework for interactive data exploration*. In: SIGMOD (2014)
- Dong, X.L., Halevy, A., Yu, C.: *Data integration with uncertainty*. VLDB J.18(2) (2009)
- Dong, X.L., Srivastava, D.: *Big data integration*. In: Proc.ICDE. IEEE (2013)
- Furche, T., Gottlob, G., Libkin, L., Orsi, G., Paton,N.W.: *Data Wrangling for Big Data: Challenges and Opportunities*. In: Proc. EDBT, vol. 16 (2016)
- Garcia, S., Luengo, J., Herrera, F.: *Data Preprocessing in Data Mining*, Intelligent Systems Reference Library, vol. 72. Springer (2015)

References (2/3)

- Golshan, B., Halevy, A.Y., Mihaila, G.A., Tan, W.: *Data Integration: After the Teenage Years*. In: Proc. PODS.ACM (2017)
- Grafberger, S., Stoyanovich, J., Schelter, S.: *Lightweight Inspection of Data Preprocessing in Native Machine Learning Pipelines.* In: Proc. CIDR (2021)
- Halevy, A., Rajaraman, A., Ordille, J.: *Data Integration: The Teenage Years*. In: Proc. VLDB (2006)
- Hodge, V.J., Austin, J.: *A Survey of Outlier Detection Methodologies*. Artif. Intell. Rev.22(2) (2004)
- Idreos, S., Papaemmanouil, O., Chaudhuri, S.: *Overview of Data Exploration Techniques*. In: SIGMOD (2015
- Ilyas, I.F., Chu, X.: *Trends in Cleaning Relational Data: Consistency and Deduplication*. Foundations and Trends in Databases 5(4) (2015)
- Inmon, W.H.: *Building the Data Warehouse*. Wiley; 4.edition (2005)
- Kim, W., Seo, J.: *Classifying schematic and data heterogeneity in multidatabase systems*. Computer24(12)(1991)
- Klettke, M., Awolin, H., Störl, U., Müller, D., Scherzinger, S.*: Uncovering the evolution history of data lakes*. In: Scalable Cloud Data Management@BigData (2017)
- Klettke, M., Störl, U., Scherzinger, S.: *Schema Extraction and Structural Outlier Detection for JSON-based NoSQL Data Stores*. In: Proc. BTW (2015)
- Lenzerini, M.: *Data Integration: A theoretical perspective*. In: Proc. PODS (2002)
- Moh, C., Lim, E., Ng, W.K.: *DTD-miner: A tool for mining DTD from XML documents.* In: Proc. WECWIS (2000)
- Möller, M.L., Berton, N., Klettke, M., Scherzinger, S., Störl, U.: *jHound: Large-Scale Profiling of Open JSON Data*. In: Proc. BTW (2019)

References (3/3)

- Morik, K., Kotthaus, H., Heppe, L., Heinrich, D., Fischer, R., Pauly, A., Piatkowski, N.: *The Care Label Concept: A Certification Suite for Trustworthy and Resource-Aware Machine Learning.* CoRR (2021)
- Nargesian, F., Zhu, E., Miller, R.J., Pu, K.Q., Arocena, P.C.: *Data lake management: Challenges and opportunities*. Proc. VLDB Endow. (2019)
- Naumann, F., Herschel, M.: *An introduction to duplicate detection*. Synthesis Lectures on Data Management2(1)(2010)
- Panse, F.: *Duplicate detection in probabilistic relational databases*. Ph.D. thesis, Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky (2014)
- Rahm, E., Bernstein, P.A.: *A survey of approaches to automatic schema matching*. VLDB J.10(4) (2001)
- Valerie Restat, Meike Klettke, Uta Störl: *"FAIR" is not enough – A Metrics Framework to ensure Data Quality through Data Preparation.* Workshop Data Engineering for Data Science (DE4DS)@BTW, 2023
- Ruiz, D.S., Morales, S.F., Molina, J.G.: *Inferring Versioned Schemas from NoSQL Databases and Its Applications.* In: Proc. ER, vol. 9381. Springer (2015)
- Seifert, C., Scherzinger, S., Wiese, L.: *Towards Generating Consumer Labels for Machine Learning Models*. In: Proc. CogMI. IEEE (2019)
- Terrizzano, I.G., Schwarz, P.M., Roth, M., Colino, J.E.: *Data Wrangling: The Challenging Journey from the Wild to the Lake.* In: Proc. CIDR (2015)
- Wang, H., Bah, M.J., Hammad, M: *Progress in outlier detection techniques*: A survey. IEEE Access7 (2019)