Data Engineering: Evolution von Daten und Evolution von Datenbanksystemen

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

Universität Regensburg

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

Diplom im Fach Informatik, Universität Rostock

1992

Laura

1995

Promotion
Thema: Akquisition von Integritätsbedingungen in Datenbanken
Universität Rostock

1997 1998

Helena

2003

Jakob

Habilitation
Thema: Modellierung, Bewertung und Evolution von XML-Dokumentkollektionen
Universität Rostock

2007 2009

Hanna

Prorektorin für Internationales, Gleichstellung und Vielfaltsmanagement
Universität Rostock

2018 2020

Moritz

W3 Professur Data Engineering Universität Regensburg

2022 2024

Freja

Wer sind wir?

Einige Forschungsschwerpunkte:

- Evolution von NoSQL-Datenbanken und Graph-Datenbanken
- Abstraktionen über Daten, Profiling in nicht-relationalen Datenbanken
- Evolution in Data-Engineering-Pipelines
- Selbst-adaptierende Data-Engineering-Methoden und –Prozesse

Und Themen in der Lehre:

- Datenbanken 1
- Data Engineering
- Datenbanken 2 (ab nächstem Semester)
- Softwareprojekte (ab nächstem Semester)
- Programmieren 1

Inhalte der heutigen Antrittsvorlesung

ausgehend von **Evolution** als einem der Hauptthemen am Lehrstuhl Data Engineering

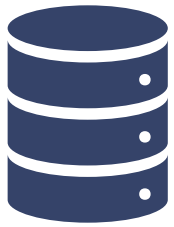
- klassische Kernthemen des Faches Data Engineering
- Erweiterung und Veränderung dieser Themen in den letzten Jahren
  - Einordnung der aktuellen Forschungsthemen in die Themen des Faches
  - Zuordnung der Lehrveranstaltungen
  - zukünftige Pläne

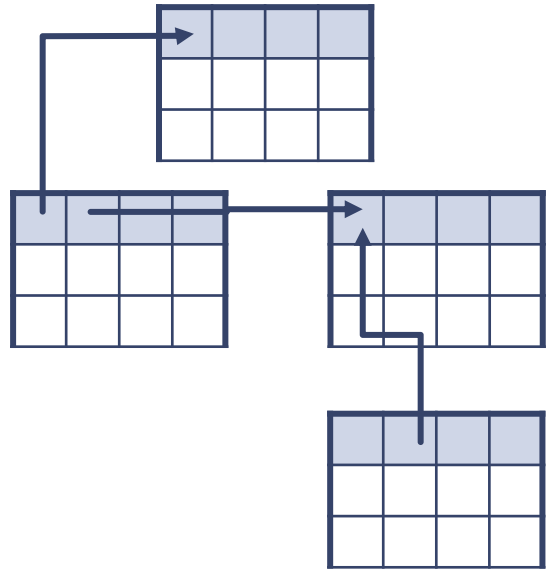
Das Fach Data Engineering

Datenbanken 1:
Vorlesungen und Übungen

- stammt aus dem Fach **Datenbanken und Informationssysteme**
- Methoden und Tools für die
  - dauerhafte, zuverlässige und verlässliche Datenspeicherung
  - Garantie von Strukturierung und Datenkonsistenz
  - Realisieren paralleler Zugriffe auf Datenbestände

  - basierend auf etlichen formalen Grundlagen wie
    - relationales Datenmodell, relationale Algebra
    - Modellierungsmethode für Datenbanken (ERM)
    - Indexverfahren
    - Anfrageoptimierung
      …
    - zahlreiche Systeme: Oracle, IBM DB2, PostGres, mysql





zentrale Komponente jeder Software-Architektur (Backend)

Universität Regensburg

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

Data Engineering – Anforderungen und Aktuelles

Neue Datenformate, insbesondere für Objekte:

- objektorientierte Datenbanken
- XML-Datenbanken
- JSON-Datenbanken (NoSQL)

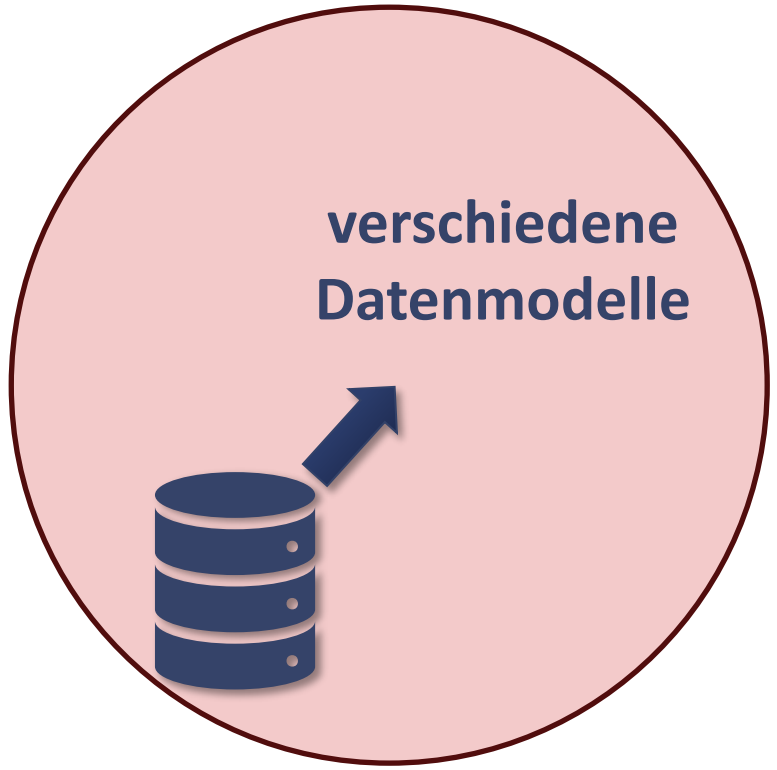
für komplexe vernetzte Strukturen:

- Graphdatenbanken

und damit: häufig kommen in Anwendungen etliche Datenmodelle zusammen

Datenbankaufgaben dabei:

- Datenintegration,
- Verteilung von Anfragen
- Transformation zwischen Datenmodellen



Universität Regensburg

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

Unsere Arbeiten/ Themen dazu:

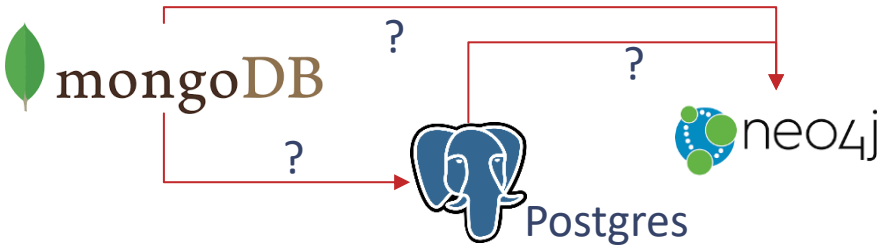
Große Tradition von Verfahren für nicht-relationale Daten:

- XML-Dokumente
- NoSQL
- Graphen

insb. Evolution, Transformation

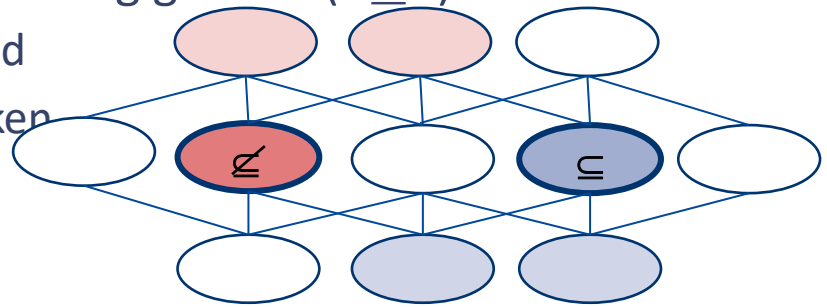
Datenbanken 2: Vorlesungen und Übungen

Verbindung verschiedener DBMS zu Multi-Modell-Datenbanken



Ableitung von Inklusionsabhängigkeiten (A⊆B) für

- NoSQL- Datenbanken und
- Multi-Modell-Datenbanken



(zur Datenintegration und Evolution)

Universität Regensburg

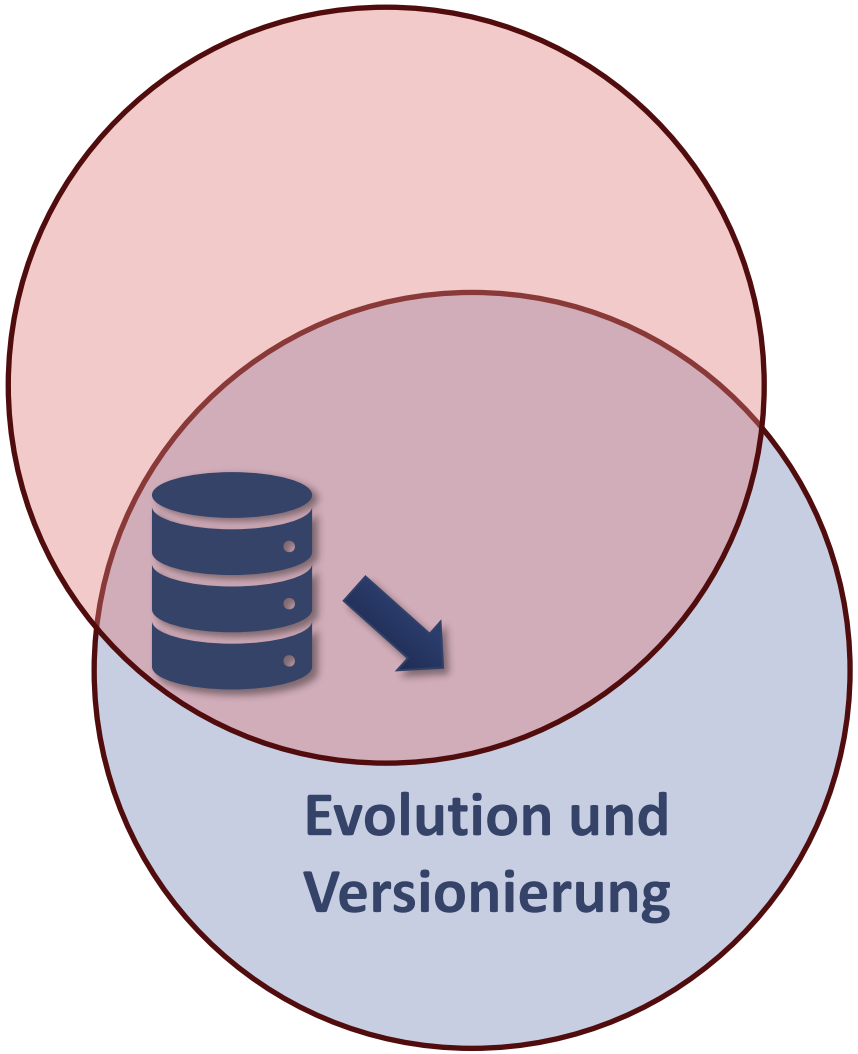
Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Data Engineering – Anforderungen und Aktuelles

Evolution von Daten

= Weiterentwicklung über die Zeit

- **Daten sind länger im Einsatz als Software**
- Methodenentwicklung für Design verfügbar
- für Evolution (Re-Design) vorhandener Datenbanken notwendig



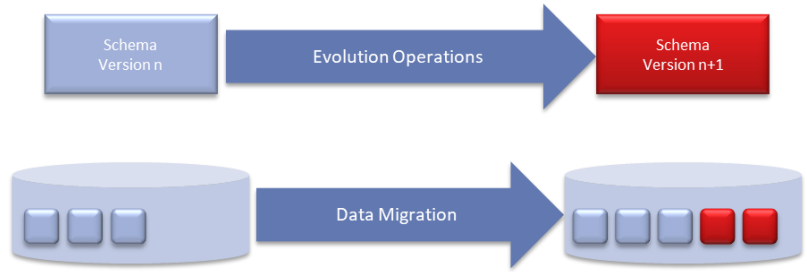
Universität Regensburg

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

Unsere Arbeiten/ Themen dazu:

Schemaevolution und Datenmigration für NoSQL-Datenbanken

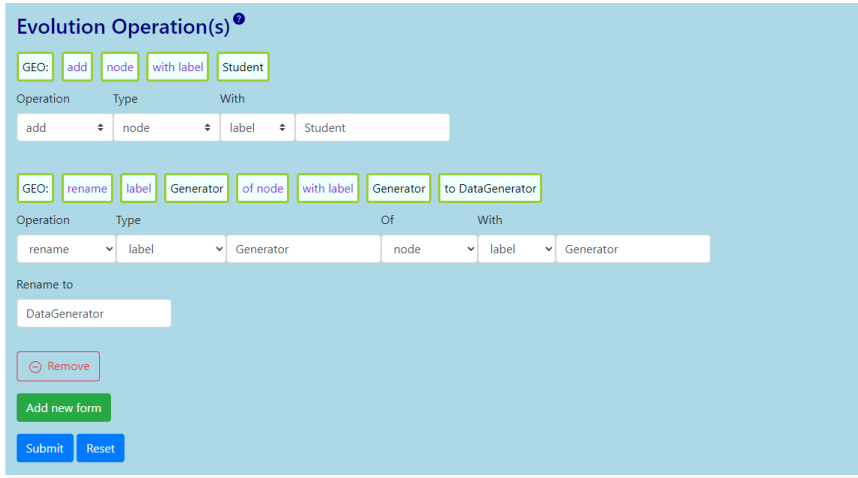
- für JSON-Datenbanken
- einfache Operationen (`add, delete, rename`) und
- komplexe Operationen (`move, copy`)



- **Query Rewriting** zur Verwendung der Daten in verschiedenen Versionen

Evolution von Graphdaten

- Ableitung **struktureller Informationen** und **Statistiken** zu Dateneigenschaften
- **Human-in-the-loop** Ansatz zur Evolution

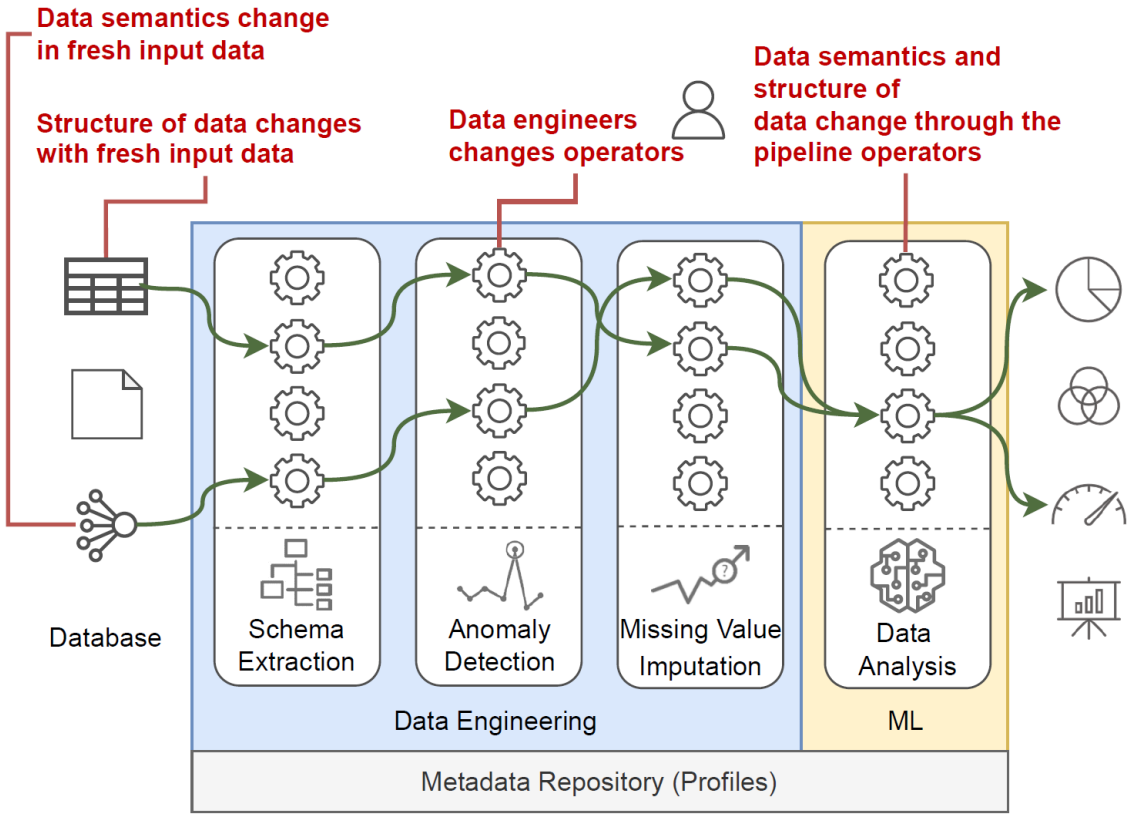


basierend auf Neo4j

Und unsere gegenwärtigen/zukünftigen Pläne dazu:

Evolution in Data Engineering Workflows

- verschiedene Gründe für Evolution von Pipelines
- **Daten-Eigenschaften**:
  - Änderungen der Datenformate, Schemata, Datentypen, …
- **Pipeline-Eigenschaften**:
  - Änderung der Algorithmen in den Workflows
  - Veränderung der Pipeline
- In allen Fällen:
  - **re-use oder re-train**?



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

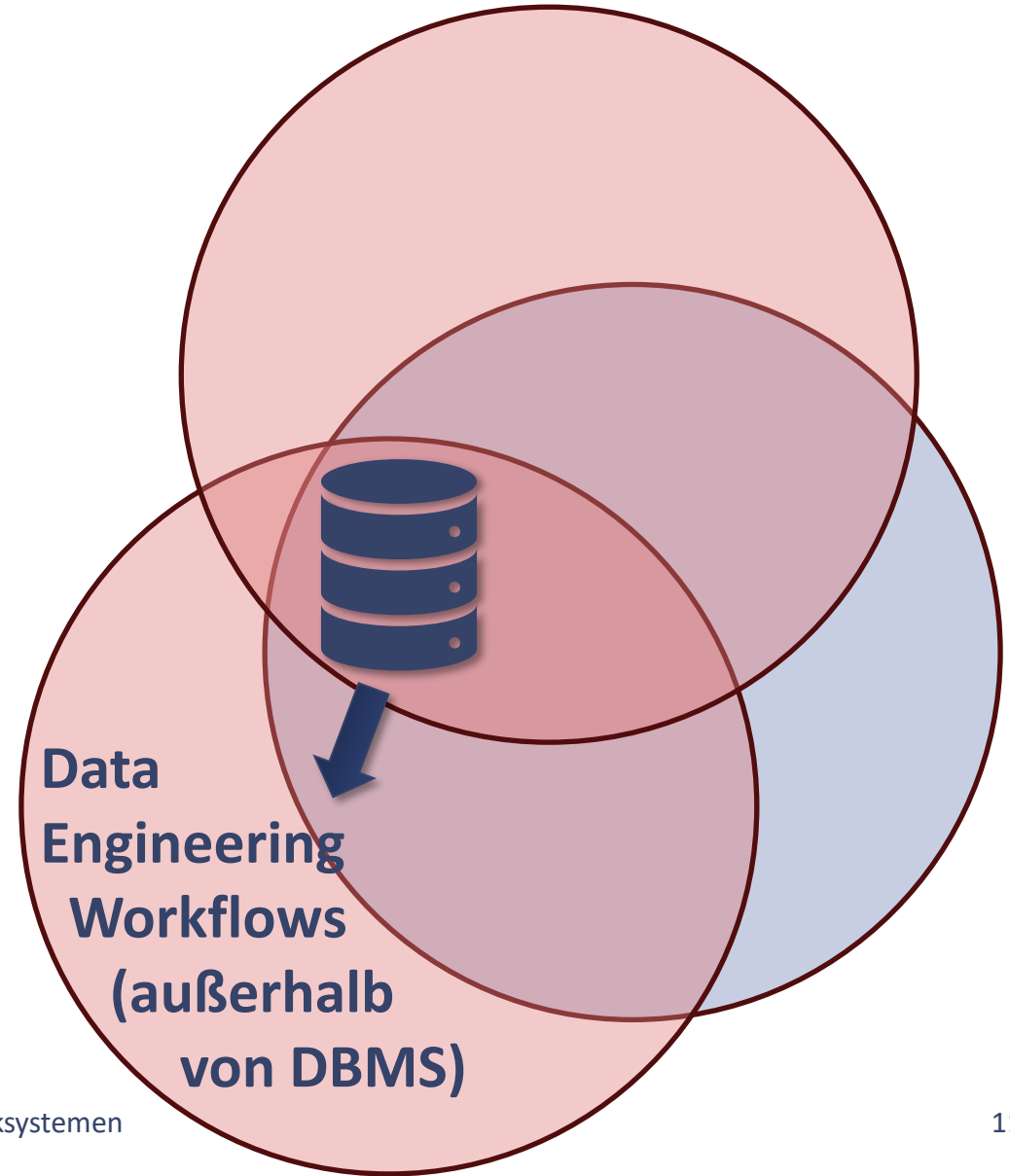
Data Engineering – Anforderungen und Aktuelles

Daten außerhalb von Datenbank-Managementsystemen

- Data Science Workflows beinhalten zahlreiche Data Engineering-Aufgaben

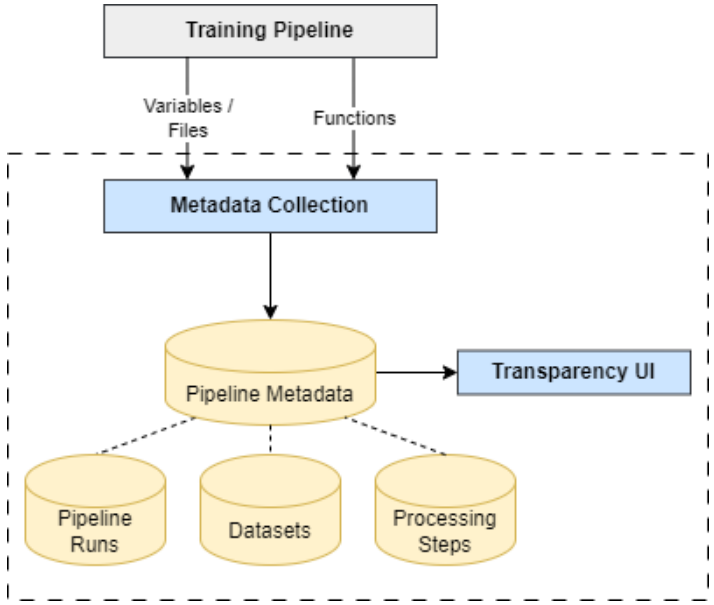
Aufgabe hierbei:

- Dankbankfunktionalität außerhalb von Datenbankmanagementsystemen zu entwickeln und auszuführen



Unsere Arbeiten/ Themen dazu:

Monitoring
von Data Engineering Pipelines



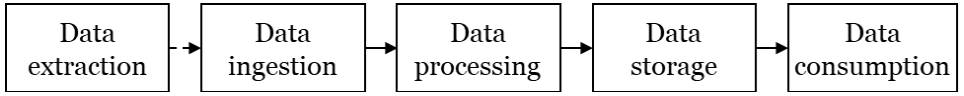
Provenance in und für **Data Engineering Workflows**:

- Data Provenance basiert auf Operationen der relationalen Algebra

Zwei Entwicklungsrichtungen:

- Data Preprocessing außerhalb von Datenbanken (Protokollierung und Formalisierung der **Workflow-Algorithmen**)
- Einbeziehung anderer Datenmodelle (**Graphdaten**)

Cloud-Data-Engineering-Workflows
mit den Teilaufgaben



Data Engineering: Vorlesungen und Übungen

Und unsere zukünftigen Pläne dazu:

CENTURY – Geschichte Mittel- und Osteuropas im langen 20. Jahrhundert (1918 bis heute)

Informatikanteile dabei:

- Workflows für natürlichsprachige historische Texte
- Generieren von Informationsgraphen
- Entwickeln von Portalen zur Exploration und Suche

Herausforderung:

- low-ressouce languages
- Evolution der Quellen über die Zeit

zusammen mit Udo Kruschwitz (Kompetenz NLP)

Nutzung der Kompetenz in den Bereichen

- Graphdatenbanken
- Ableitung von Informationen aus Daten/ Text
- Evolution und Transformation von Daten
- Eye-Tracking-Daten

für weitere Anwendungen

im Sommersemester 2025

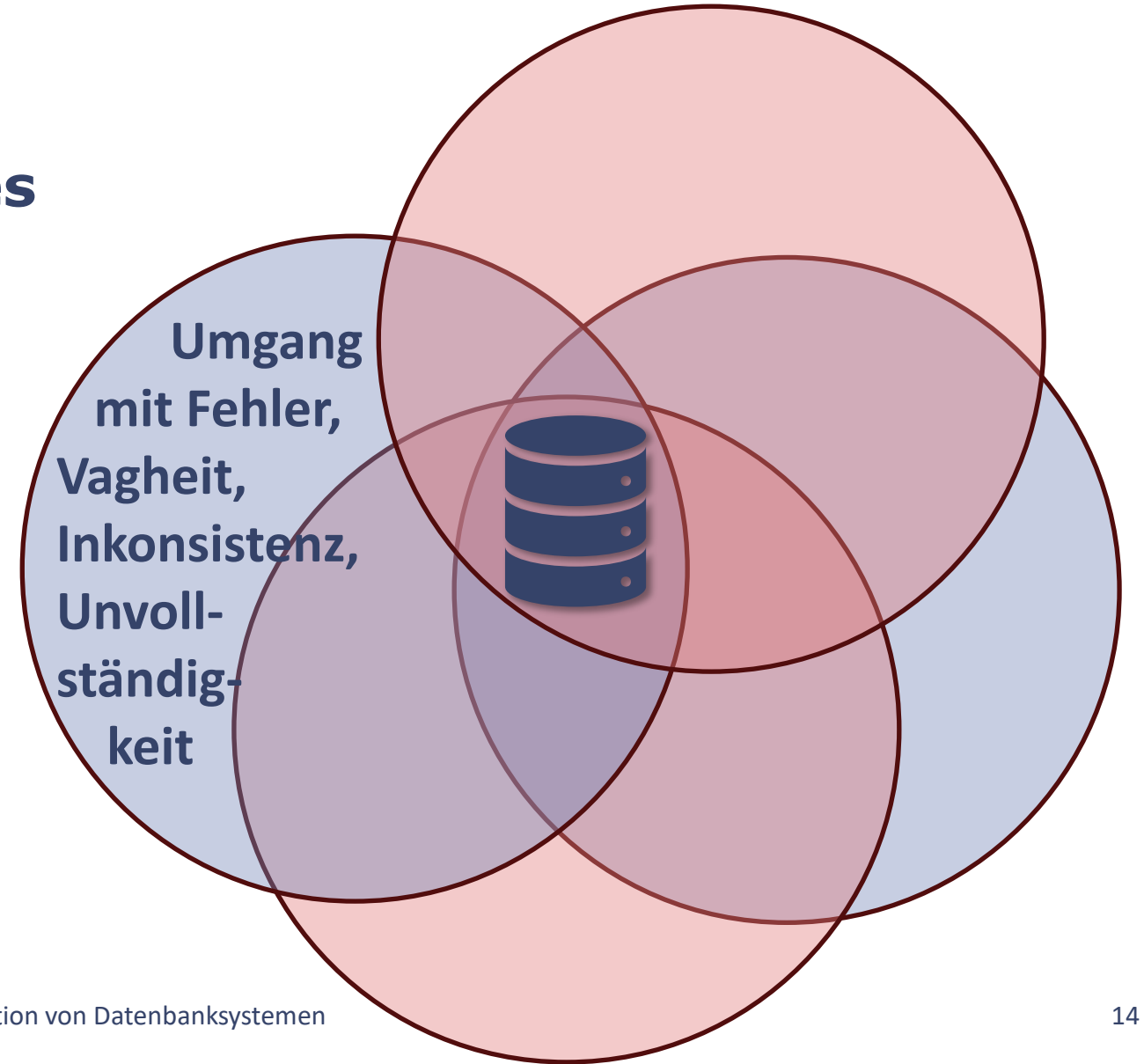
Softwareprojekt: Entwicklungs- und Teamarbeit

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

Data Engineering – Anforderungen und Aktuelles

Umgang mit Unvollständigkeit, Vagheit, Unvollständigkeit, Inkonsistenz

- insbesondere, wenn Daten in "nichtrelationalen" Datenbanken (JSON – NoSQL, Graph) liegen sowie
- in Data Engineering-Workflows außerhalb der Datenbanken verarbeitet werden

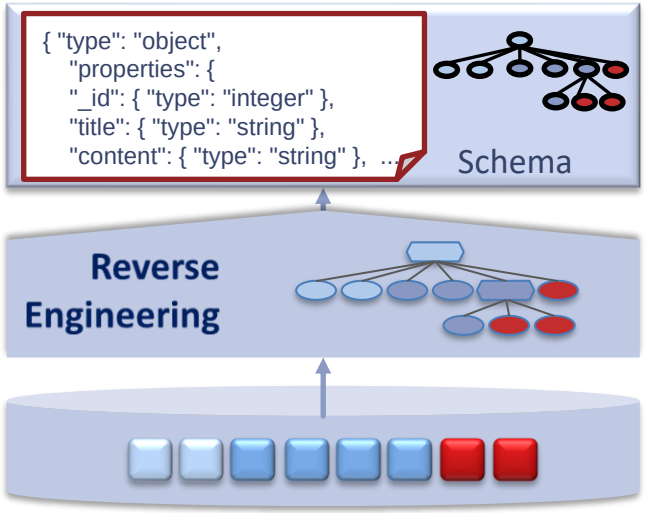


Universität Regensburg

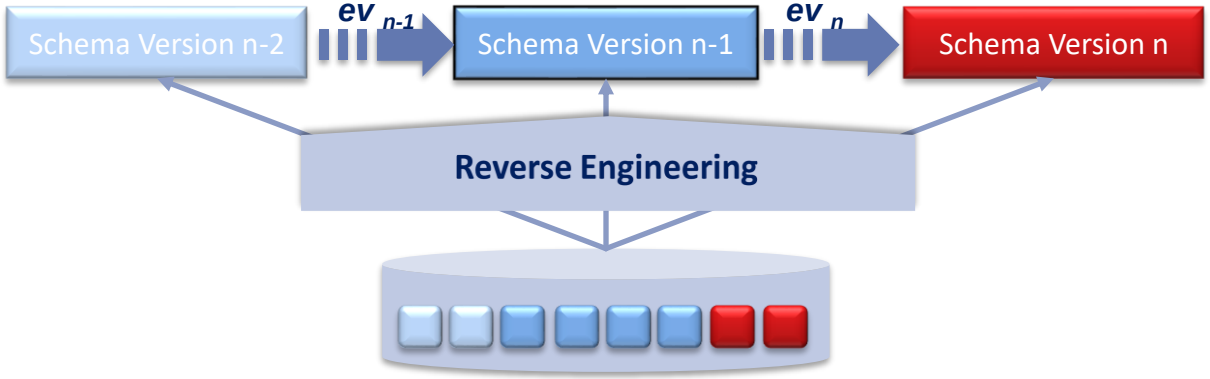
Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Unsere Arbeiten/ Themen dazu:

Schema-Ableitung aus Daten (Reverse-Engineering)



Ableitung von Schema-Versionen und Evolutionsoperationen



Tritt in **jedem Thema** auf:

– Ableiten von Graphschemata und -mustern, Monitoren von Datenänderungen, Auswertung von Texten, Erkennen von Integritätsbedingungen, …

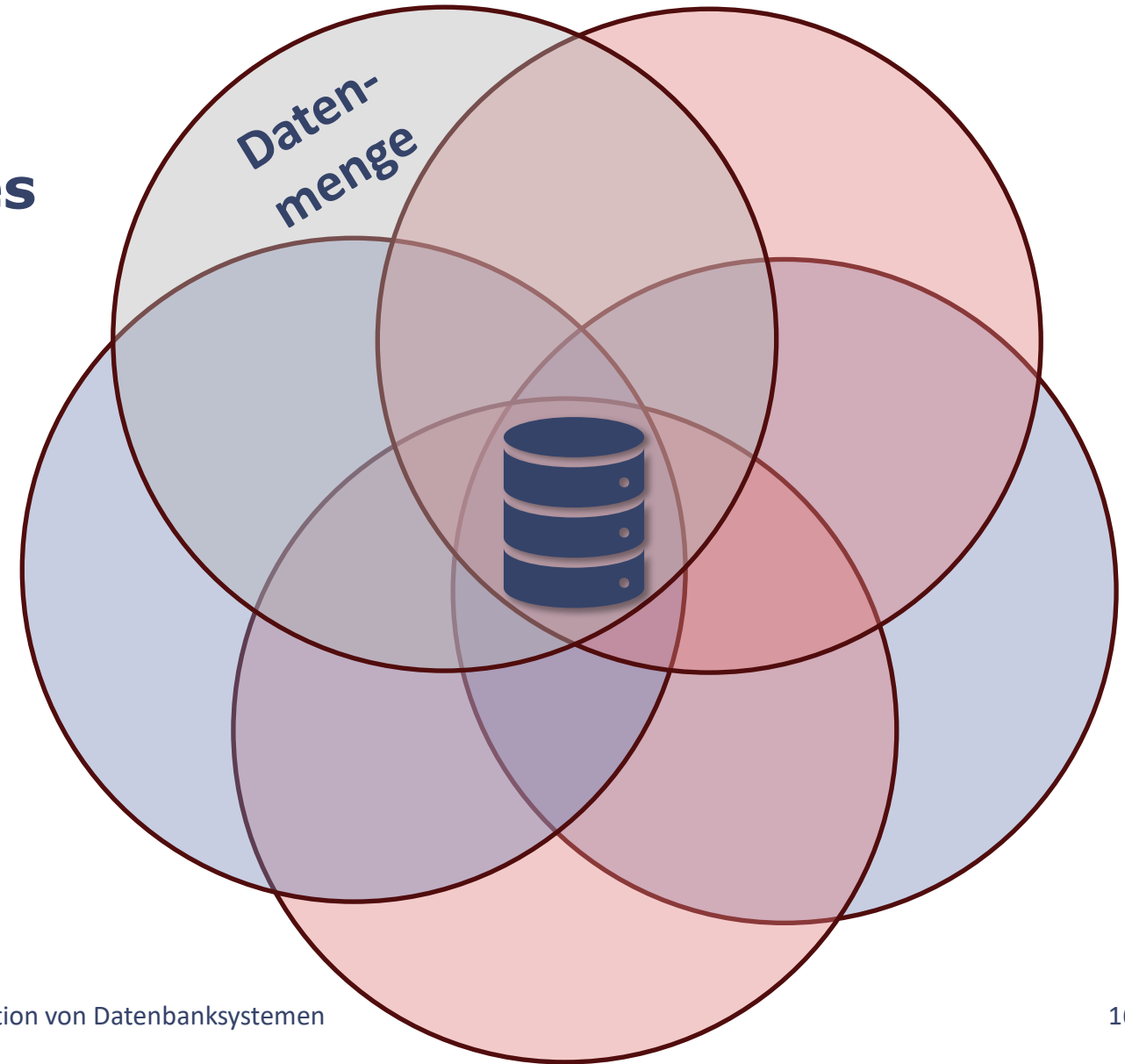
Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Data Engineering – Anforderungen und Aktuelles

und zuletzt: die Datenmenge steigt ständig und überall

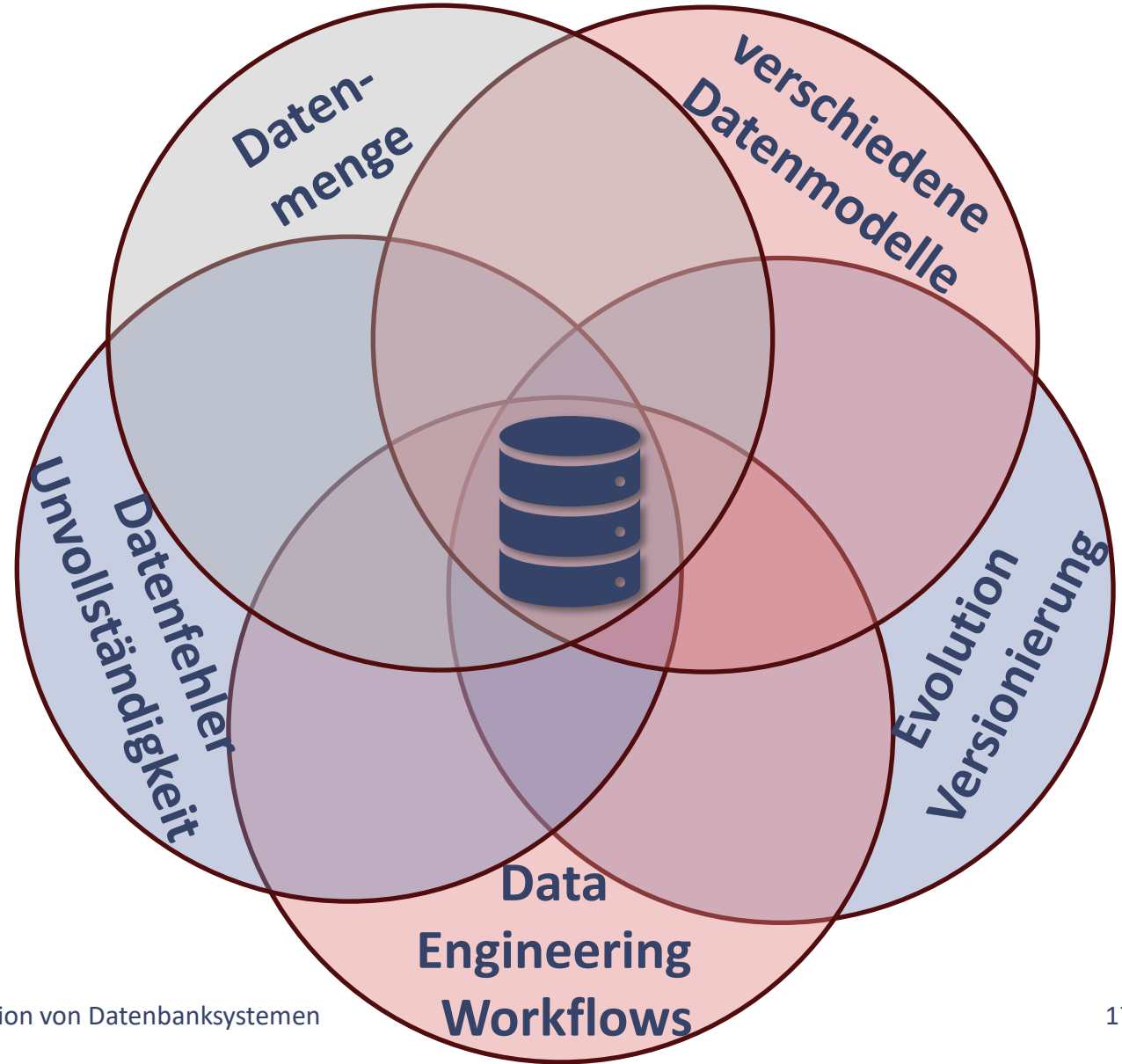
Skalierbarkeit –

- bei uns kein separates Thema
- Anforderung, die bei jedem Verfahren besteht



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

Zwischenstand



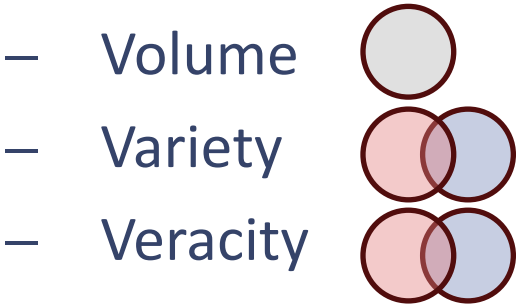
Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Big Data Movement

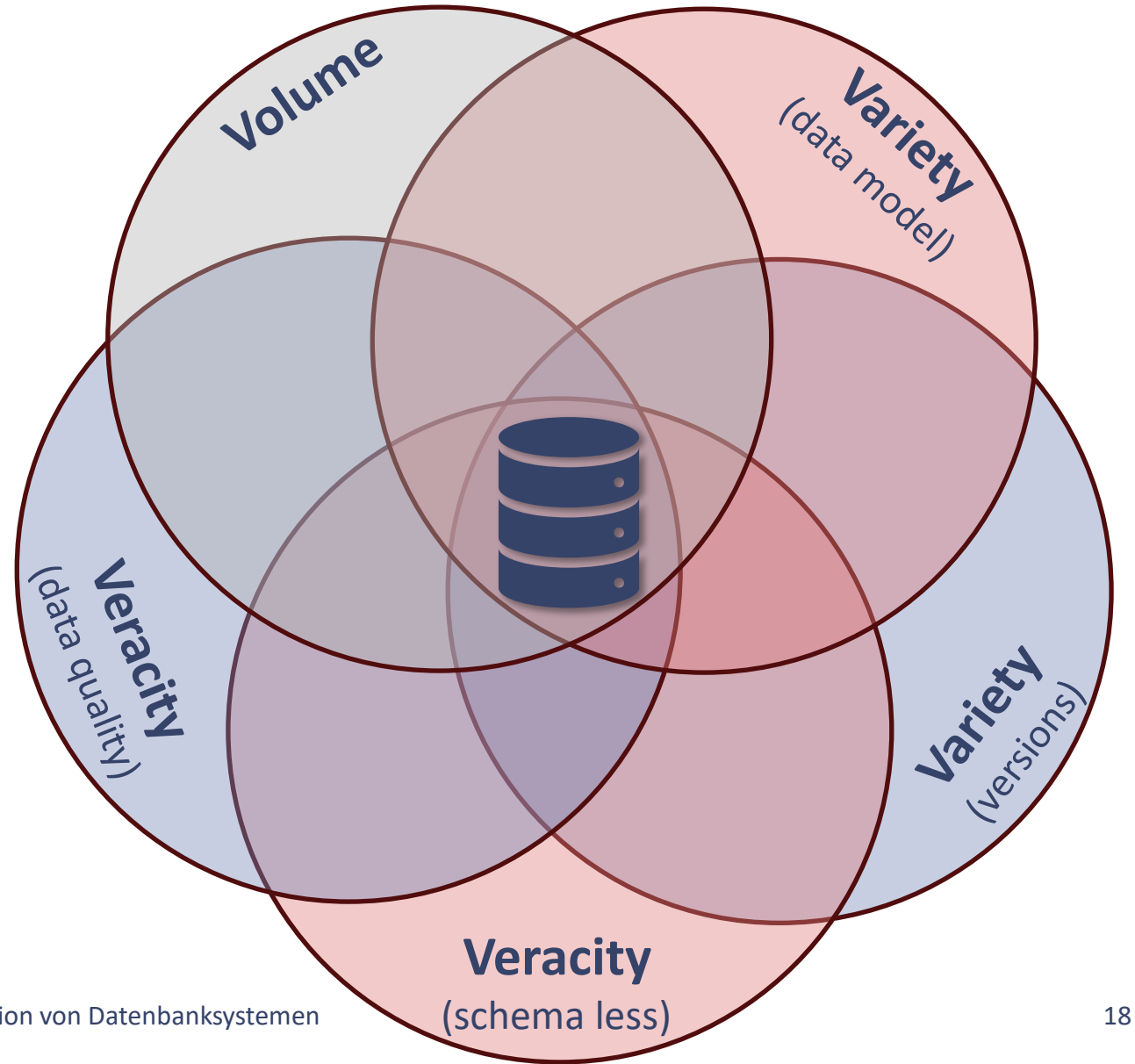
wurde in den letzten Jahren ausführlich diskutiert

Big Data:

- Volume
- Variety
- Veracity
- Velocity
- Value



➡ Dimensionen reflektieren die "großen Linien" des Faches



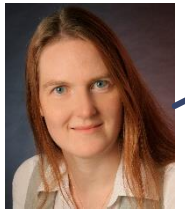
Themen am Lehrstuhl Data Engineering

NoSQL-Datenbanken
Query Rewriting

Cloud Data Engineering und LLMs für Datenbank-aufgaben

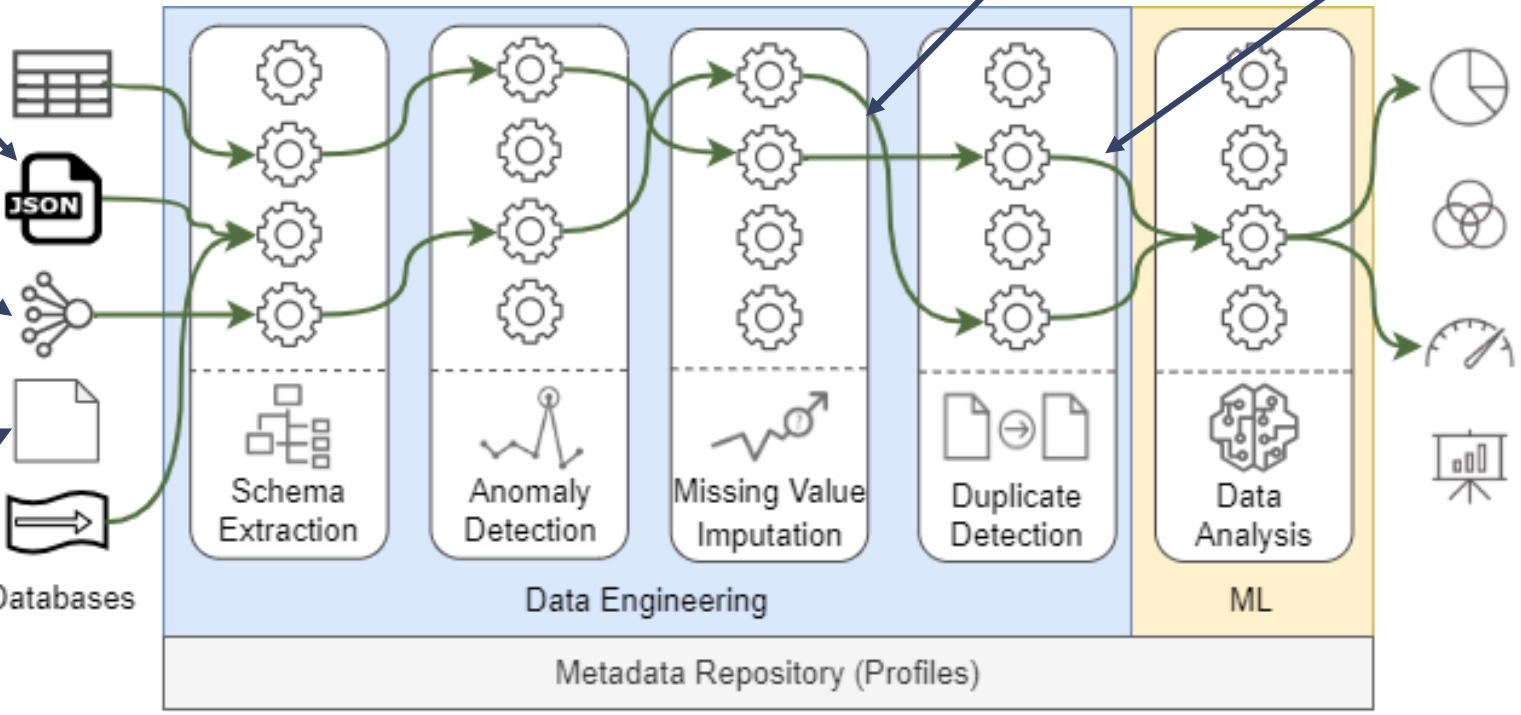
Verbinden von Data- und Workflow-Provenance

Graphdatenbank-Evolution und -Transformation



DH-Projekte

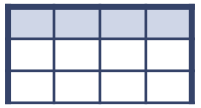
Data Engineering für Eye-Tracking Daten



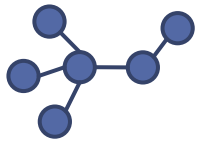
Prof. Dr.-Ing. habil. Meike Klettke
Lehrstuhl für Data Engineering
FAKULTÄT FÜR INFORMATIK UND DATA SCIENCE

Take-Away-Messages

Evolution (von Daten) hat eine ebenso große Bedeutung wie Design, muss kontinuierlich erfolgen



Daten so **strukturiert** wie möglich speichern, in Systemen, die diese Strukturierung kontrollieren (relationale Datenbanken, Schema first)



Andernfalls: andere Datenbankmanagementsysteme, zumindest **partielle Schemakontrolle**

wenn Daten außerhalb von Datenbanksystemen primär gespeichert sind, dann so viele **Data Engineering Techniken** wie möglich anwenden

Primäre Datenquelle in anderem Format, z.B. (historische) Texte, dann **Erschließung** mit Information Retrieval (NLP) und Data Engineering Technologien

Vielen Dank für die Aufmerksamkeit und

ich freue mich sehr auf die zukünftige (weitere) Zusammenarbeit!