

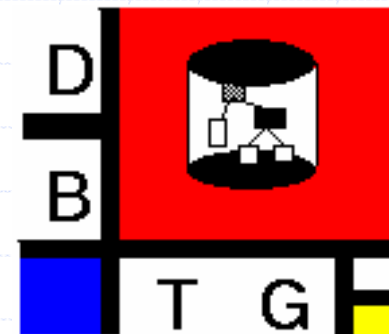
# Data Provenance: A Categorization of Existing Approaches

Boris Glavic  
Klaus Dittrich

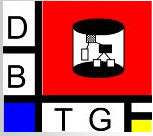


Institut für Informatik  
Universität Zürich

Database Technology  
Research Group



Binzmühlestrasse 14, CH-8050 Zürich  
e-mail: [glavic@ifi.unizh.ch](mailto:glavic@ifi.unizh.ch), <http://www.ifi.unizh.chh>  
Tel.: +41-44-635 4329, Fax: +41-44-635 6809



# Gliederung

---



**Einführung**



**Klassifikationsschema für Data Provenance**



**Zusammenfassung und Ausblick**



# Einführung - Data Provenance

---

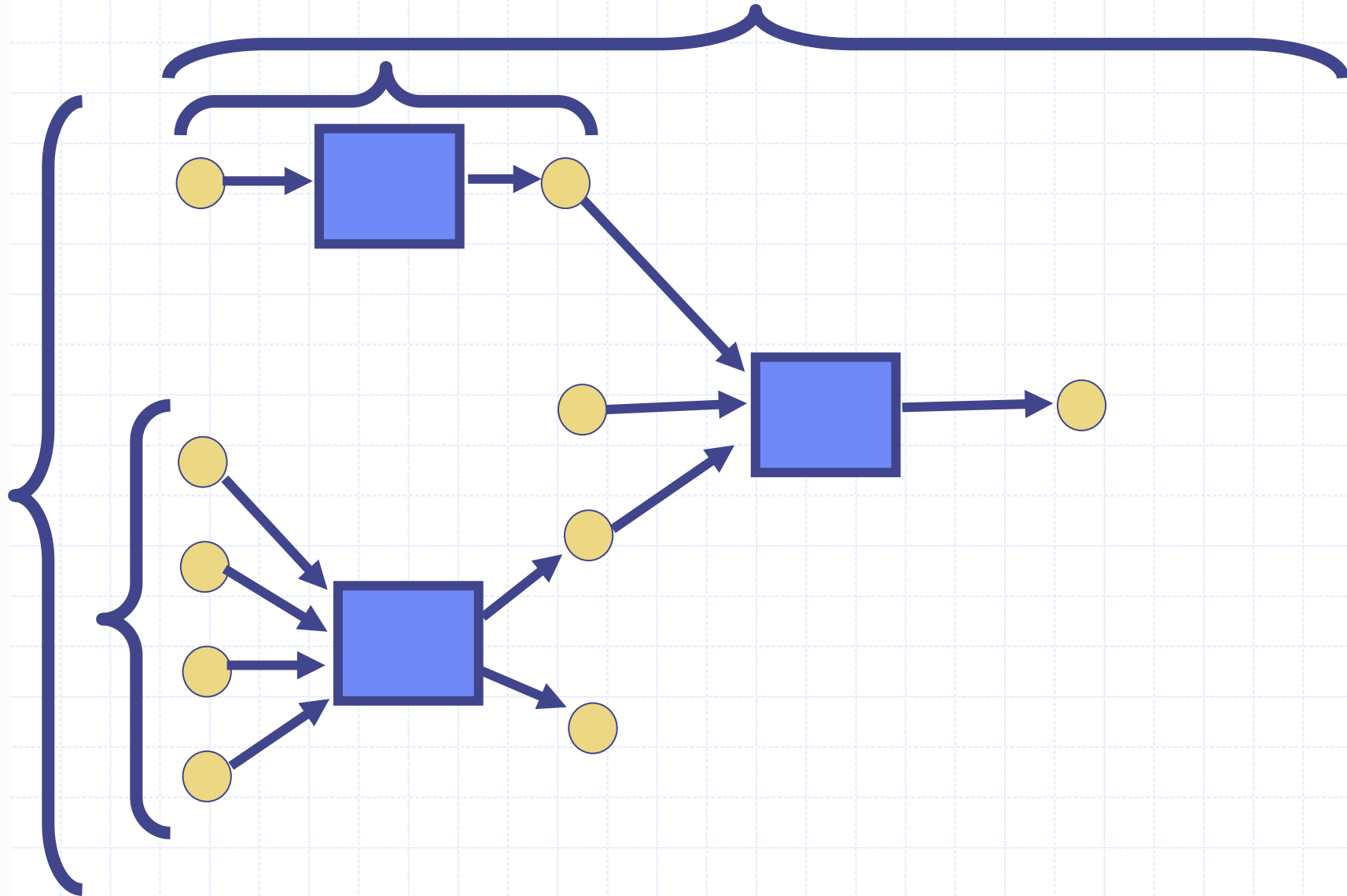
- Data Provenance (Data Lineage, Data Pedigree)
- Herkunft von Daten
  - Informationen über Entstehungsprozesse (Transformationen)
  - Informationen über Ursprungsdaten

# Einführung - Anwendungsgebiete

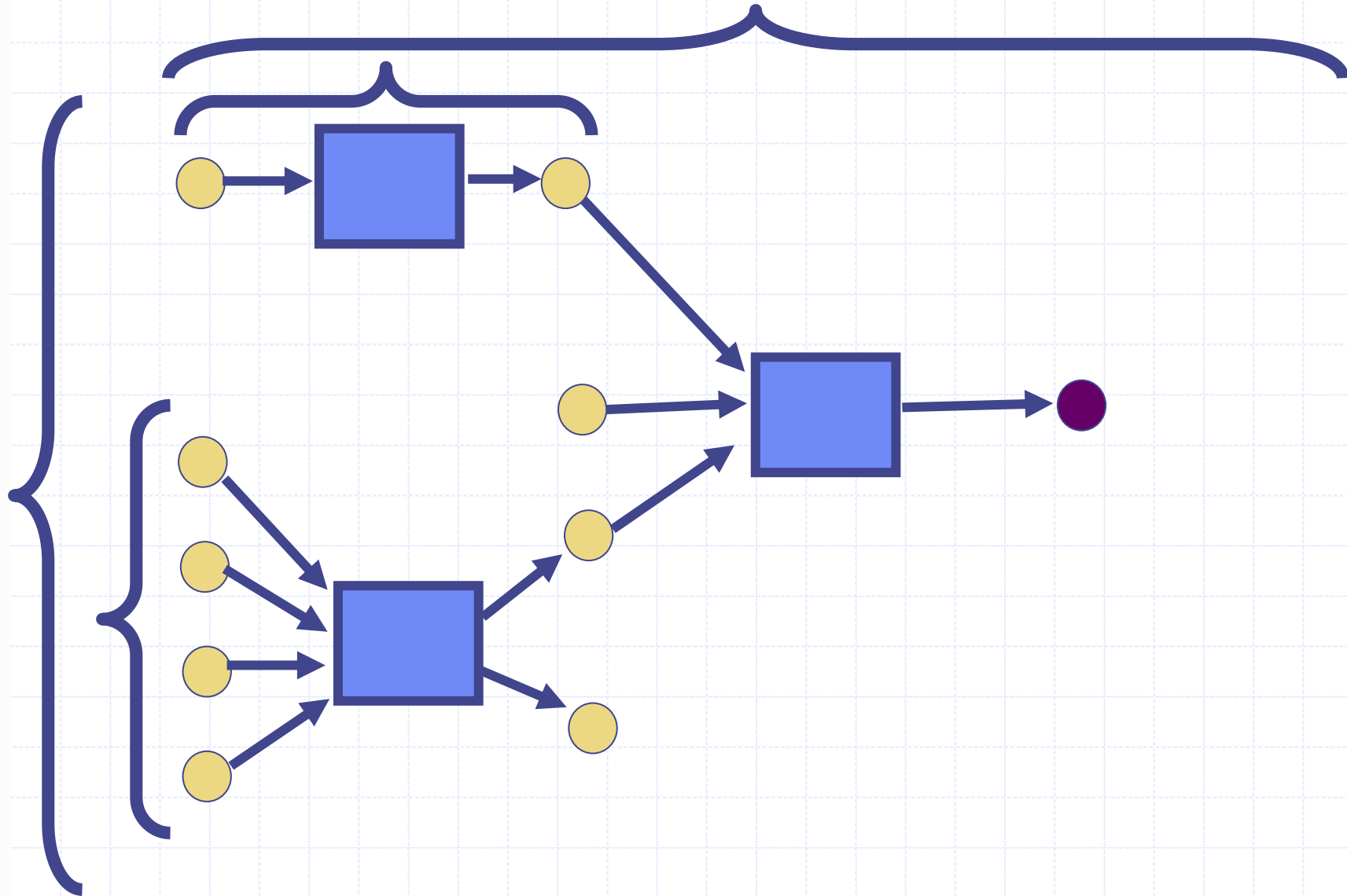
---

- "gepflegte Datenbanken" (curated databases)
- Data warehouses
- "E-science"
- Workflow management
  
- Datenqualität / Erkenntnisgewinn
- Wiederholbarkeit von Prozessen / "Virtuelle Daten"
- Urheberschaft / Zuständigkeit
- Fehlerursachen erkennen

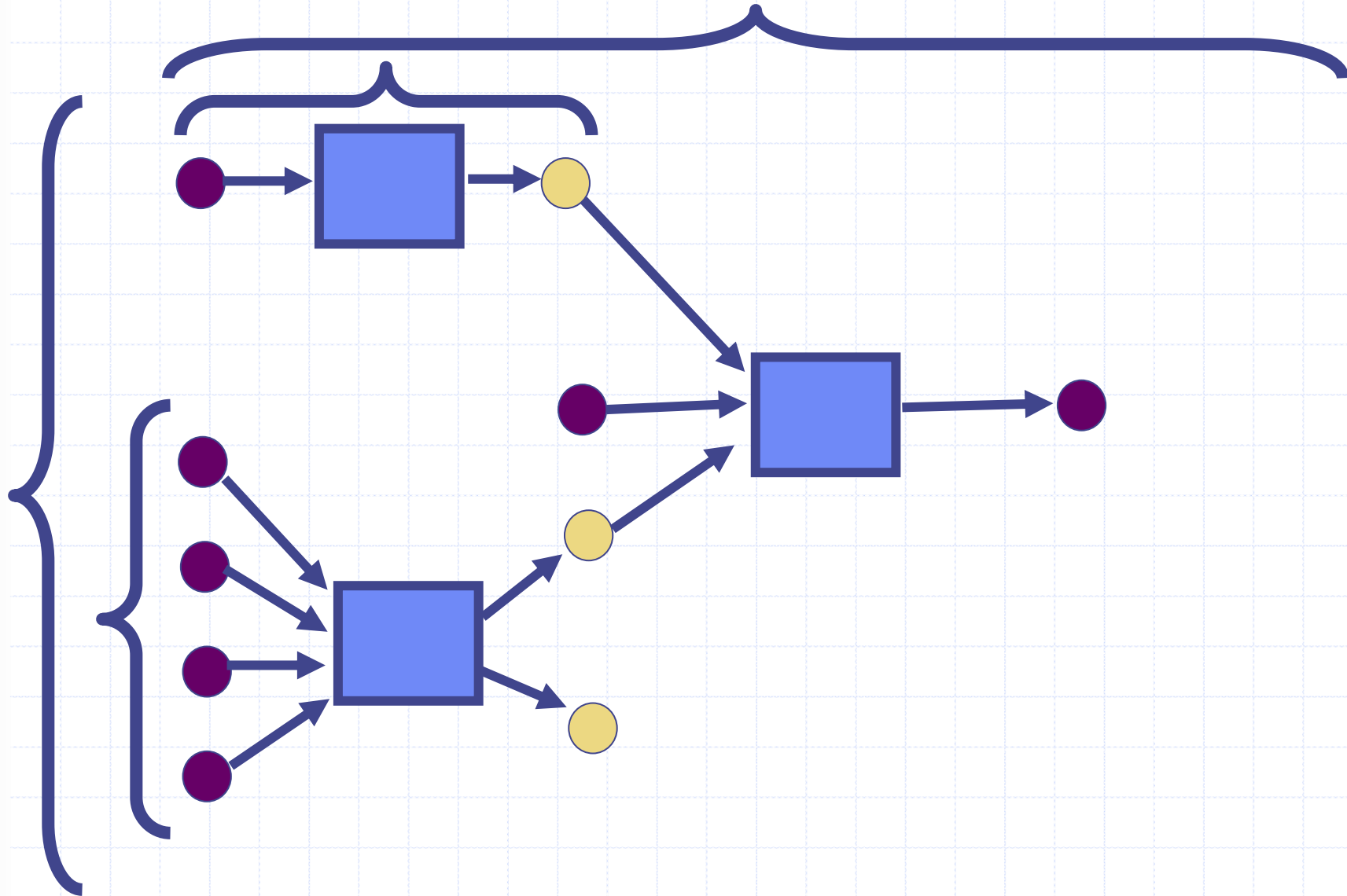
# Einführung



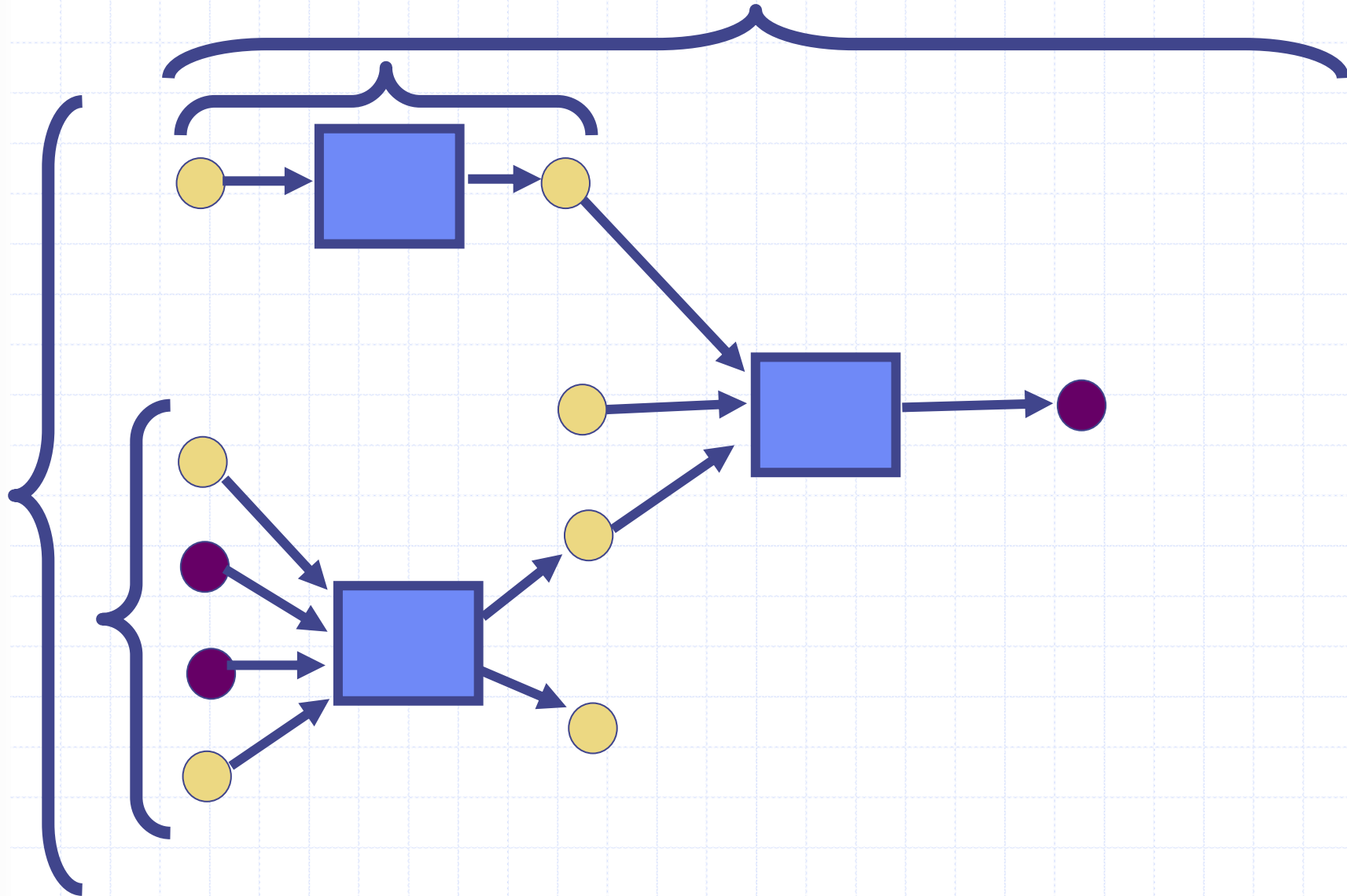
# Einführung

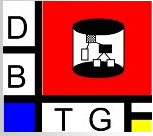


# Einführung



# Einführung



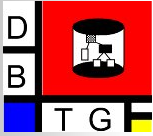


# Gliederung

---

- Einführung
- Klassifikationsschema für Data Provenance
- Zusammenfassung und Ausblick



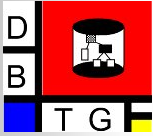


# Klassifikationsschema für Data Provenance - Begriffe

---

- Provenance model
- Provenance management system (PMS)
- Data Item
- Transformation (Source / Result)
- Level of detail
- Data repository
- Source provenance <-> Transformation provenance



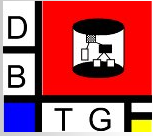


# Klassifikationsschema für Data Provenance

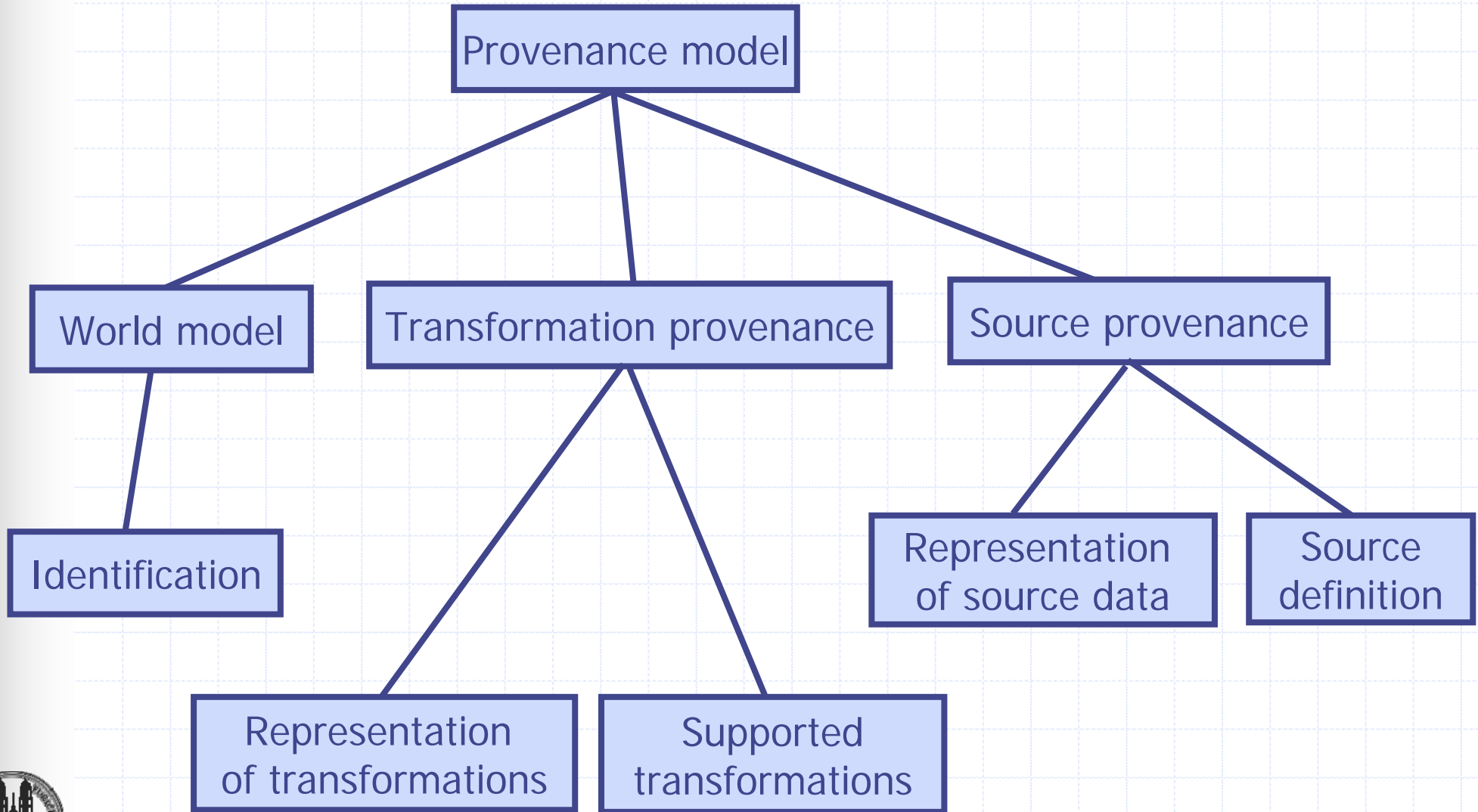
---

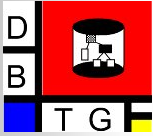
- Drei Hauptaspekte
  - Provenance model
  - Query and Manipulation Functionalities
  - Storage and recording



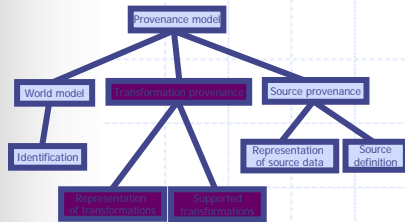


# Provenance model





# Provenance model



Transformation provenance

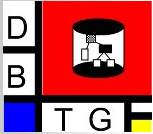
Representation of transformations

Supported transformations

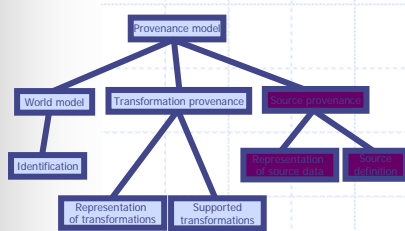
- Transformation class
- Transformation
- Meta data
- handle different detail levels

- Automatic transformations
- Manuel transformations
- semi-automatic transformations





# Provenance model



Source provenance

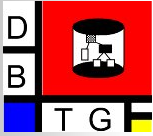
Representation of source data

Source definition

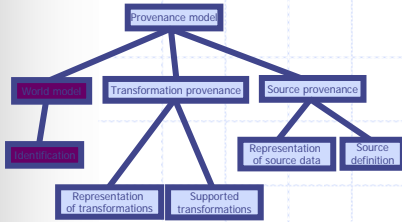
- Original source data
- Source hierarchy
- Meta data
- Handle different detail levels

- Input Source
- Contributing Source
- Original Source





# Provenance model



Identification

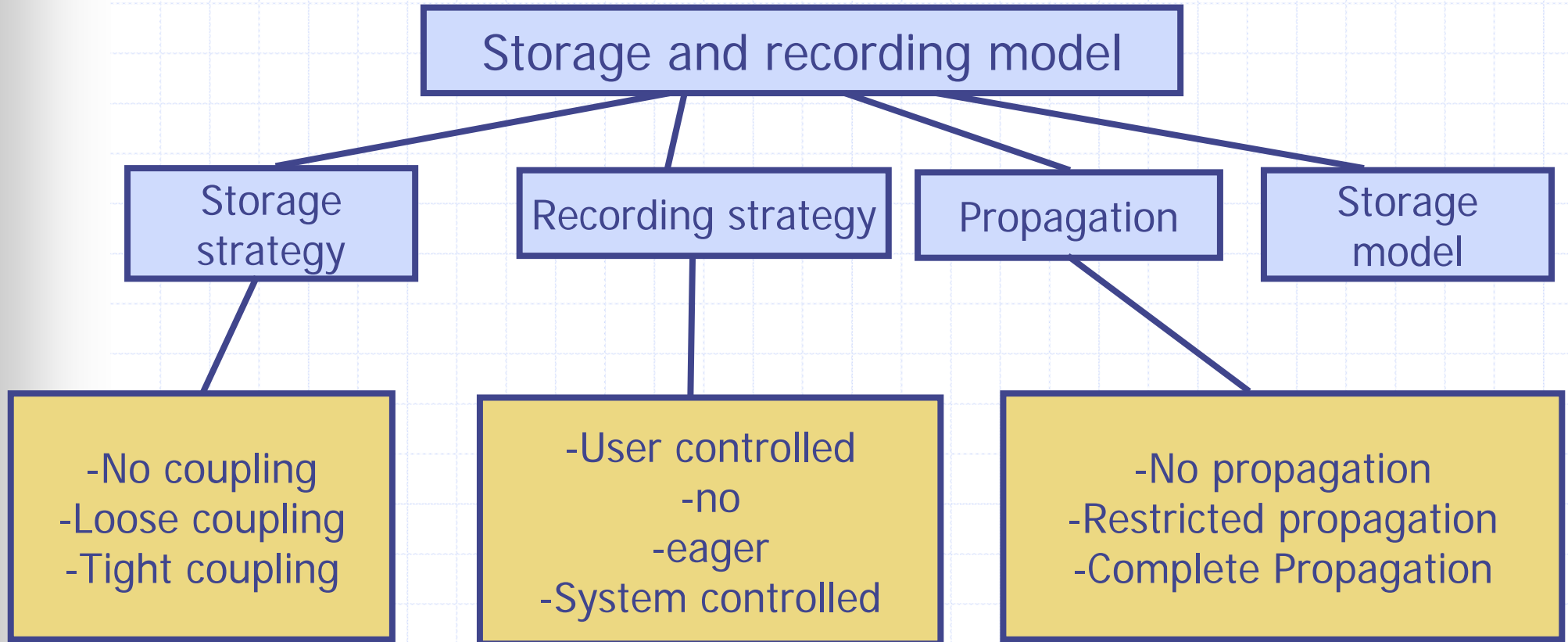
- Recognise item duplicates
- Recognise item versions
- Recognise transformation duplicates
- Recognise transformation versions

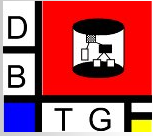
World model

- Closed world model
- Open world model

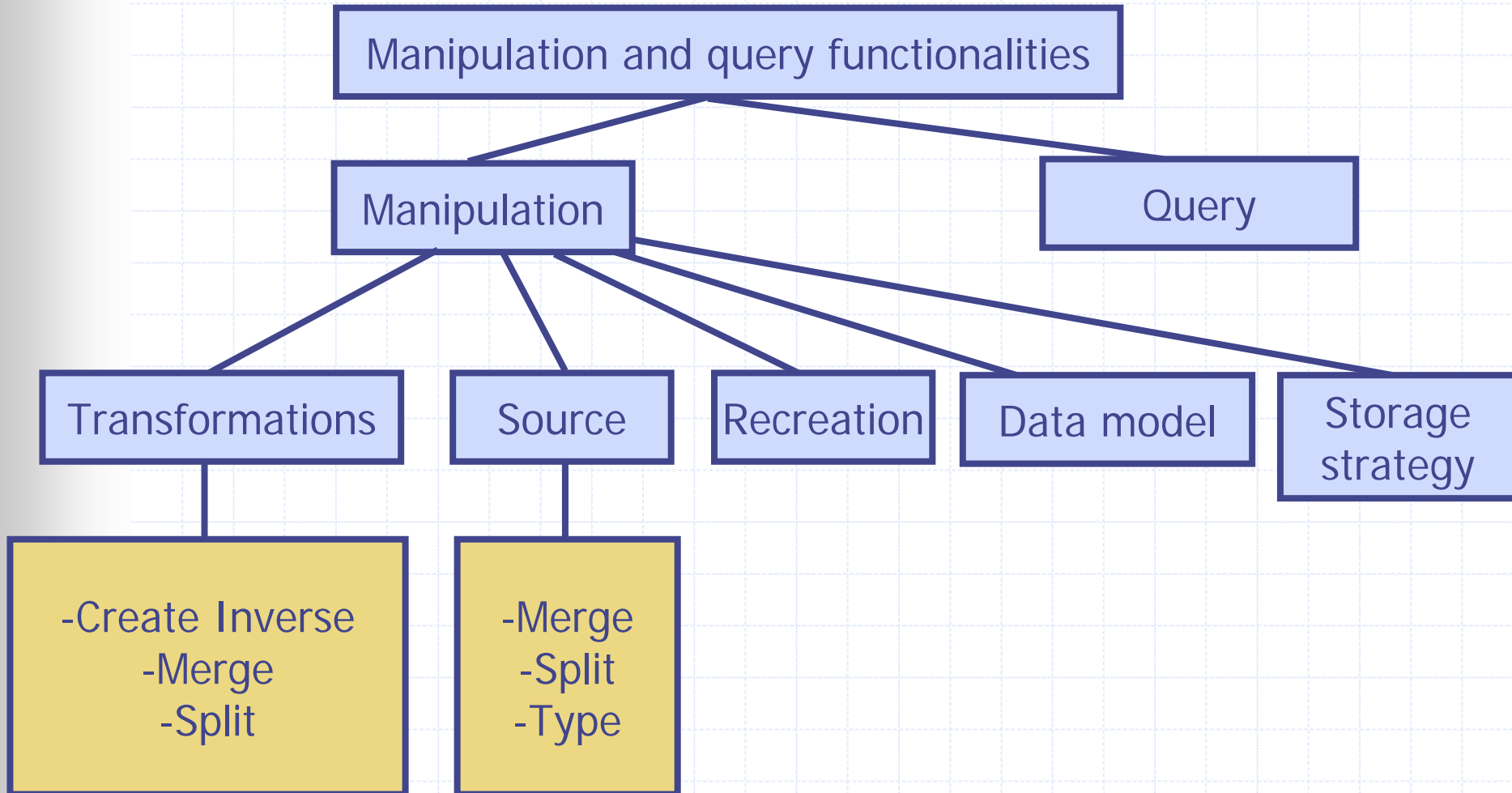


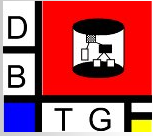
# Storage and Recording Model





# Manipulation and Query





# Gliederung

---

- Einführung
- Klassifikationsschema für Data Provenance
- Zusammenfassung und Ausblick



# Zusammenfassung

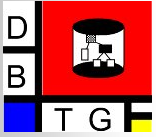
---

- Data Provenance
  - Relativ neues Forschungsgebiet
  - Verknüpft mit anderen Themengebieten aus der Datenbankforschung
    - Datenintegration
    - Temporale Datenbanken / Versionsverwaltung
    - Data Mining / Machine Learning
    - Datenqualität
- Klassifikationsschema für Provenance Management Systeme und Provenance Modelle
- Anwendung des Klassifikationsschemas im Artikel

# Ausblick

---

- Prozessorientiertes Datenmodell / Prozessorientierte Erweiterung konventioneller Datenmodelle
- Formale Beschreibung von Prozessklassen
- Provenance für Data Mining <-> Data Mining für Provenance
- Identifizierung von Versionen und Duplikaten in "open world models"
- Unterstützung unterschiedlicher Granularitäten
- Effiziente Speicherung von Provenance Informationen
- Spezialisierte Anfragesprachen / Anfrageoptimierung
- ...



# Fragen

