Advanced Data Management (CSCI 640/490)

Data Integration

Dr. David Koop

Outline

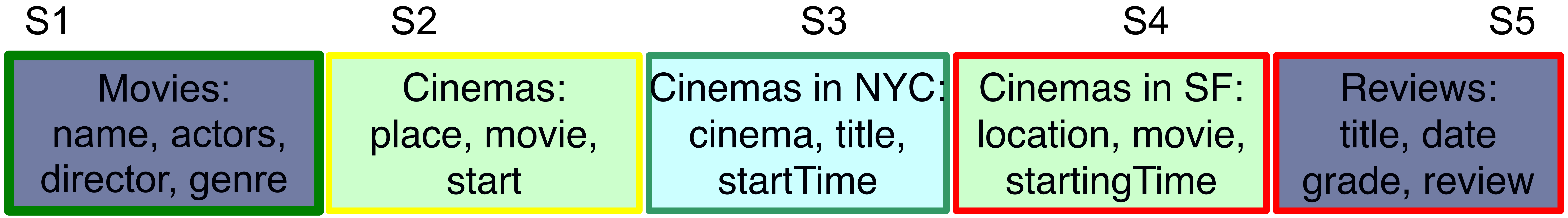
- Data Integration: Last Week
- Data Matching (Entity Resolution): Today
- Data Fusion: Today
- Data Fusion Techniques: Wednesday
    - Integrating Conflicting Data: The Role of Source Dependence, X. L. Dong et al., 2009
    - **Quiz** at the beginning of class on Wednesday, Oct. 15

Data Integration

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
           location=“New York”  AND
           director=“Ava DuVernay”
```

Movie: Title, director, year, genre
Actors: title, actor
Plays: movie, location, startTime
Reviews: title, rating, description

Sources S1 and S3 are relevant, sources S4 and S5 are irrelevant, and source S2 is relevant but possibly redundant.



S1
Movies:
name, actors,
director, genre

S2
Cinemas:
place, movie,
start

S3
Cinemas in NYC:
cinema, title,
startTime

S4
Cinemas in SF:
location, movie,
startingTime

S5
Reviews:
title, date
grade, review

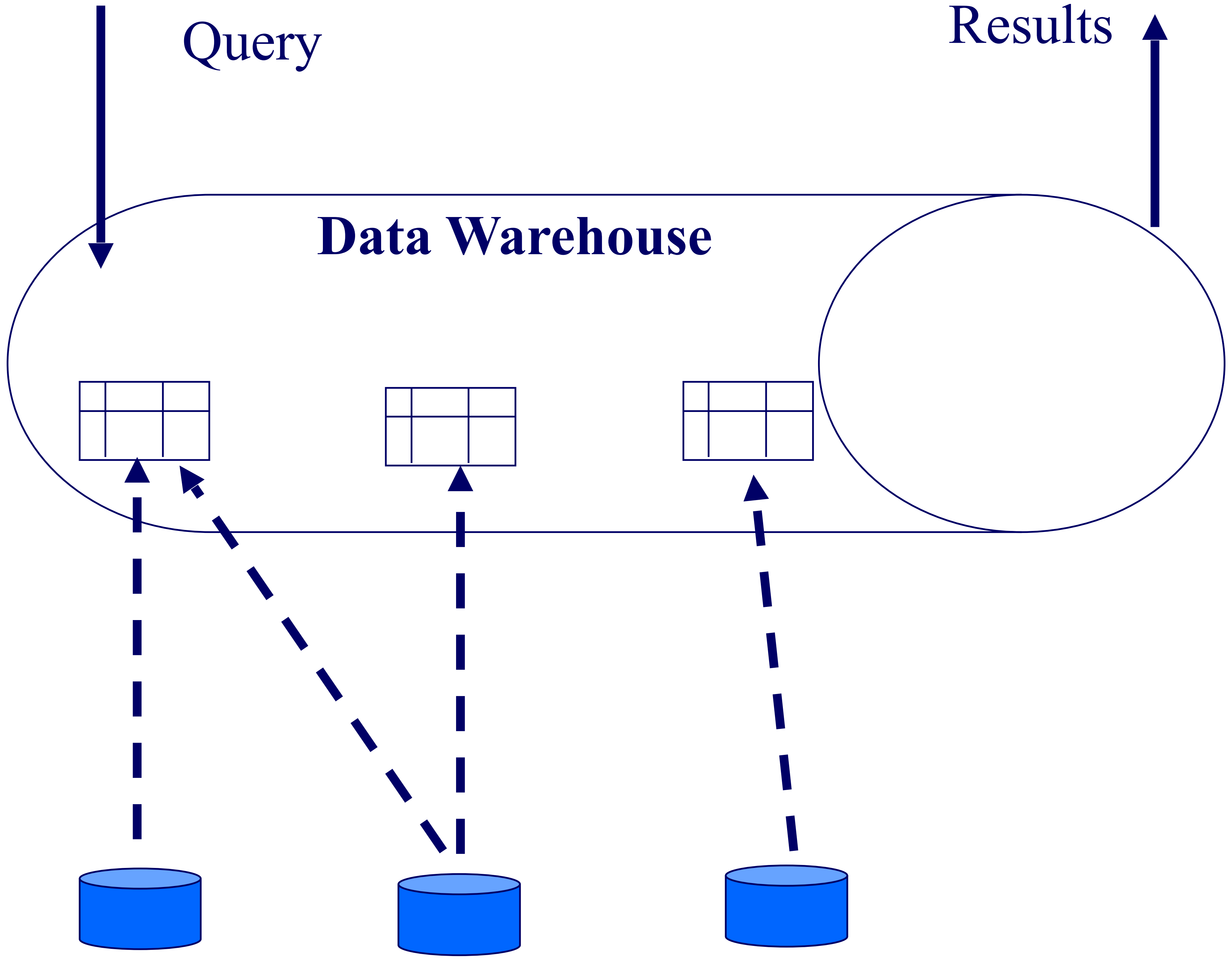
[AH Doan et al., 2012]

Data Integration

- Lots of data sources, how do we answer questions where we need to access data from more than one?
- Schema matching
- Problem of heterogeneity
- AI-Complete problem: difficulty is the same as making computers as intelligent as people
- Two techniques:
  - Mediation
  - Data Warehouses

Data Warehouses: Offline Replication

- Determine physical schema
- Define a database with this schema
- Define procedural mappings in an "ETL tool" to import the data and clean it.
- Periodically copy all of the data from the data sources
  - Note that the sources and the warehouse are basically independent at this point

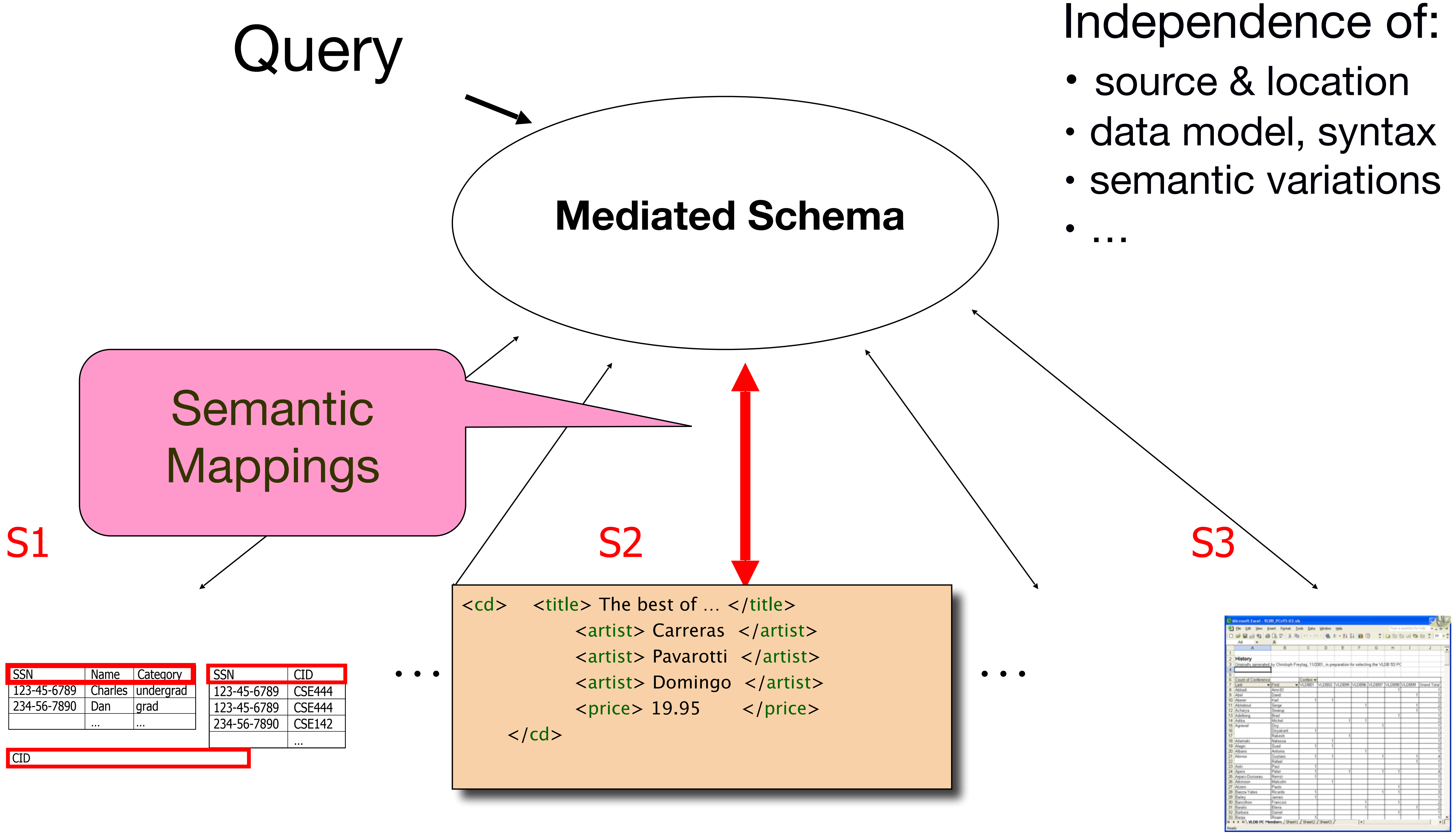


[A. Doan et al., 2012]

Virtual Data Warehouses

Independence of:

- source & location
- data model, syntax
- semantic variations
- …



SSN	Name	Category
123-45-6789	Charles	undergrad
234-56-7890	Dan	grad
	…	…

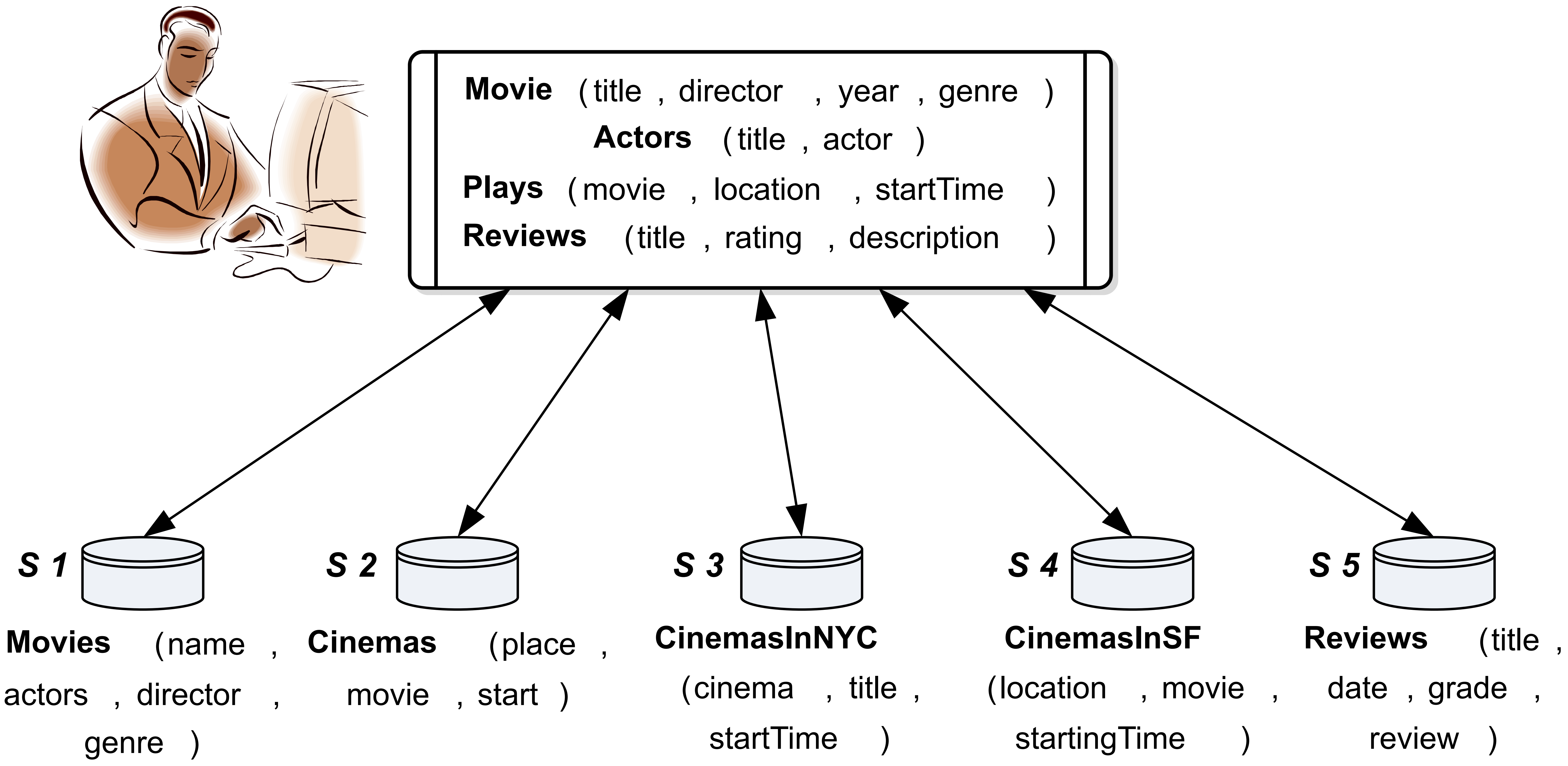
SSN	CID
123-45-6789	CSE444
123-45-6789	CSE444
234-56-7890	CSE142
	…

CID

```
<cd>   <title> The best of … </title>
           <artist> Carreras  </artist>
           <artist> Pavarotti  </artist>
           <artist> Domingo  </artist>
           <price> 19.95       </price>
    </cd>
```

[A. Doan et al., 2012]

Integrated Schema Example



Movie (title, director, year, genre)
Actors (title, actor)
Plays (movie, location, startTime)
Reviews (title, rating, description)

Movies (name, actors, director, genre)

Cinemas (place, movie, start)

CinemasInNYC (cinema, title, startTime)

CinemasInSF (location, movie, startingTime)

Reviews (title, date, grade, review)

[A. Doan et al., 2012]

Why is Data Integration Hard?

- Systems-level reasons:
  - Managing different platforms
  - SQL across multiple systems is not so simple
  - Distributed query processing
- Logical reasons:
  - Schema (and data) heterogeneity
- ‘Social’ reasons:
  - Locating and capturing relevant data in the enterprise.
  - Convincing people to share (data fiefdoms)
    - Security, privacy and performance implications

[A. Doan et al., 2012]

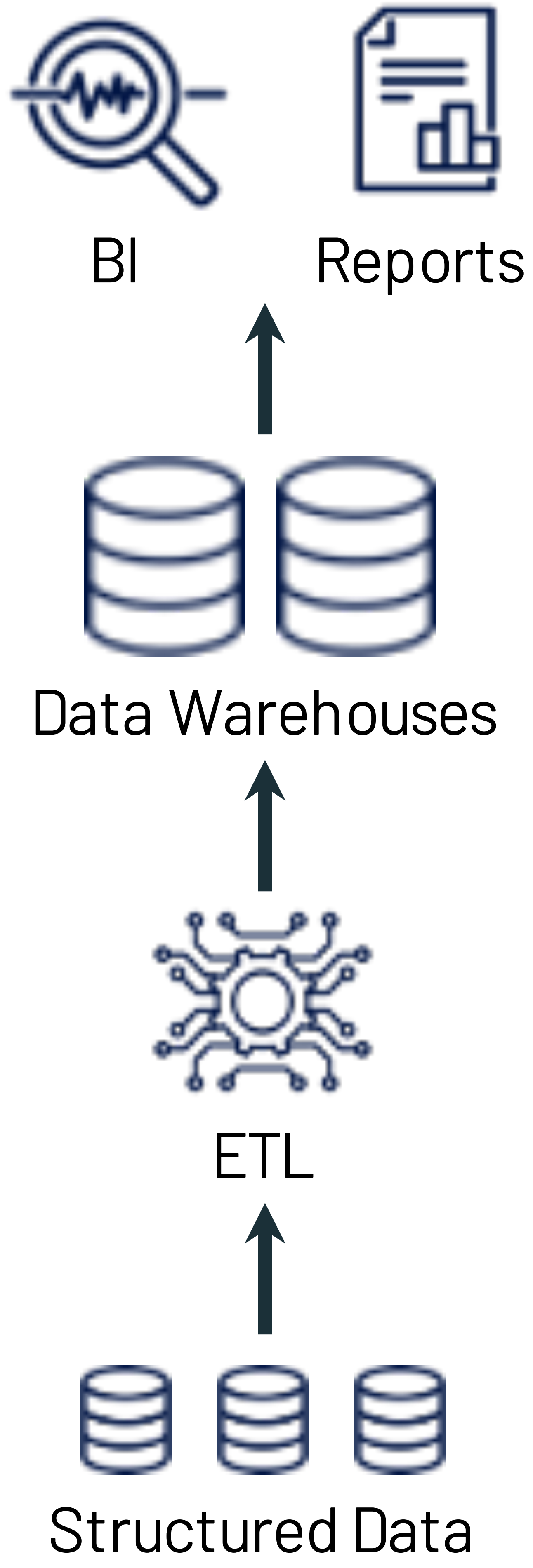
Assignment 3

- Clean the Ask a Manager Salary Survey Data
- Use polars to clean and transform data
- Will add a few more tasks or tasks using another tool

Data Lakes & Data Lakehouses

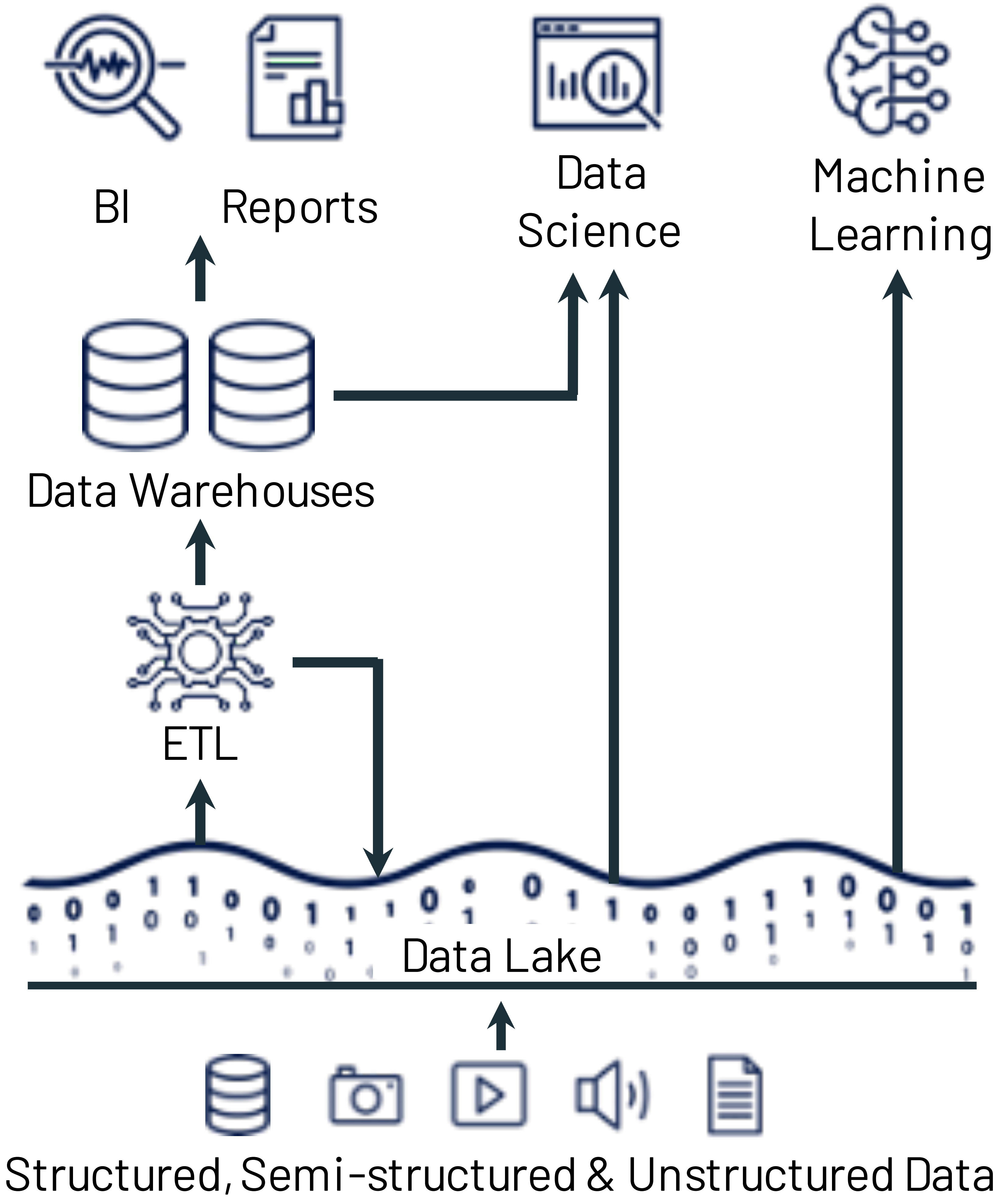
Data Warehouse

- Problem: Data stored in different files/ locations, want to run reports on that data
- Solution: load it into one big database with a set schema
- Problems:
  - Outdated data
  - Work for unknown usage
  - Schema



[M. Armbrust et al., 2021]

Data Lake

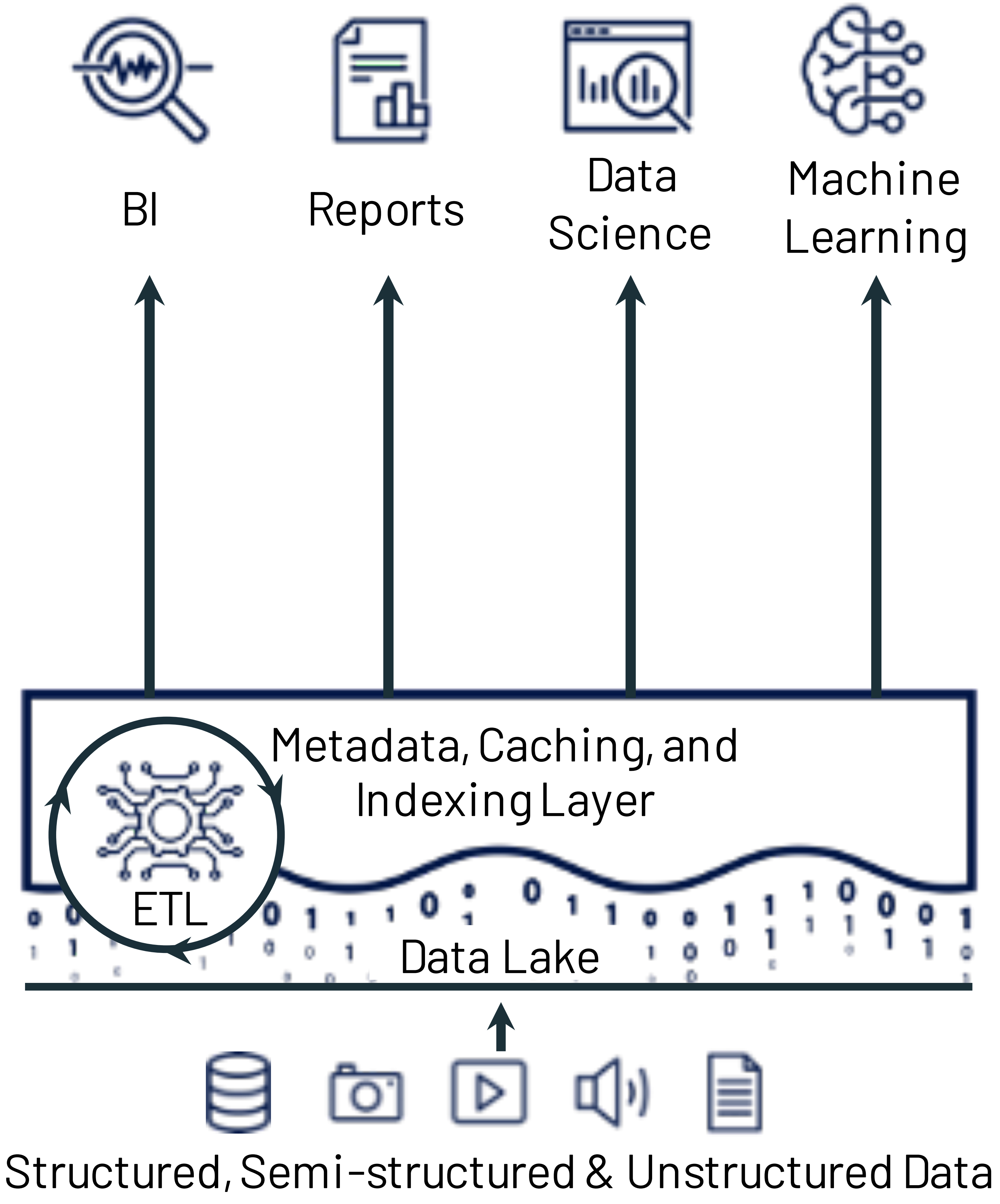


- New types of data, hard to cram into a database
- Distributed data
- Some data already used as files: data science, machine learning
- Basically, was HDFS, now usually cloud object stores like S3, Azure
- May not use some of the data
- Sometimes known as "data swamp"

[M. Armbrust et al., 2021]

Data Lakehouse

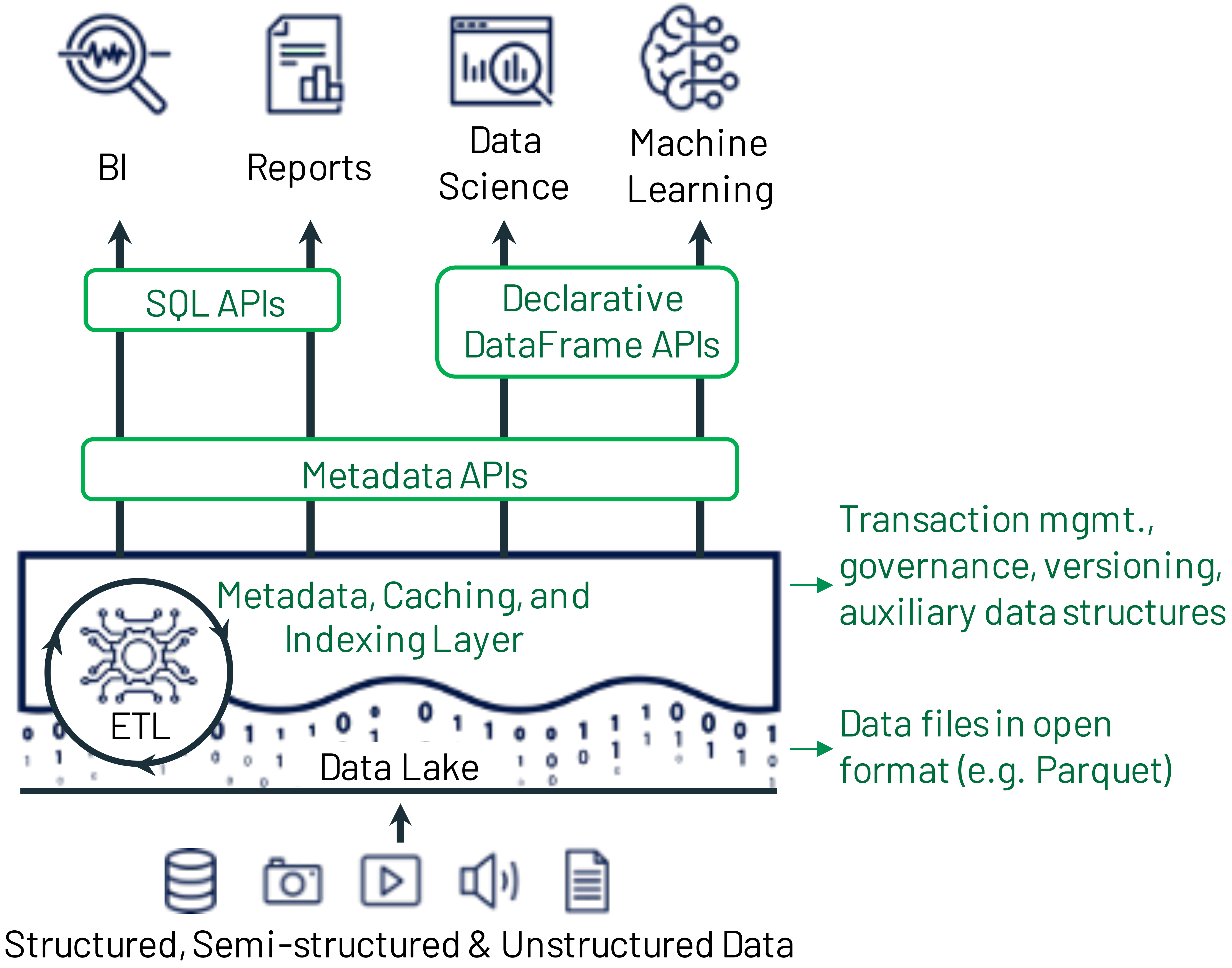
- Problems with data lakes:
  - Reliability: hard to keep lake and warehouse consistent
  - Data staleness: analysts often use out-of-date data
  - Limited support for adv. analytics: want to use with machine learning
  - Cost: storing data twice!


BI
Reports
Data Science
Machine Learning
Metadata, Caching, and Indexing Layer
ETL
Data Lake
Structured, Semi-structured & Unstructured Data

[M. Armbrust et al., 2021]

Data Lakehouse

- Solutions:
  - Reliable data management on data lakes: add metadata APIs
  - Support for machine learning and data science: allow use of declarative dataframe APIs
  - SQL performance: allow use of SQL APIs


BI
Reports
Data Science
Machine Learning
SQL APIs
Declarative DataFrame APIs
Metadata APIs
Metadata, Caching, and Indexing Layer
ETL
Data Lake
Transaction mgmt., governance, versioning, auxiliary data structures
Data files in open format (e.g. Parquet)
Structured, Semi-structured & Unstructured Data

[M. Armbrust et al., 2021]

Apache Iceberg: An Architectural Look Under the Covers

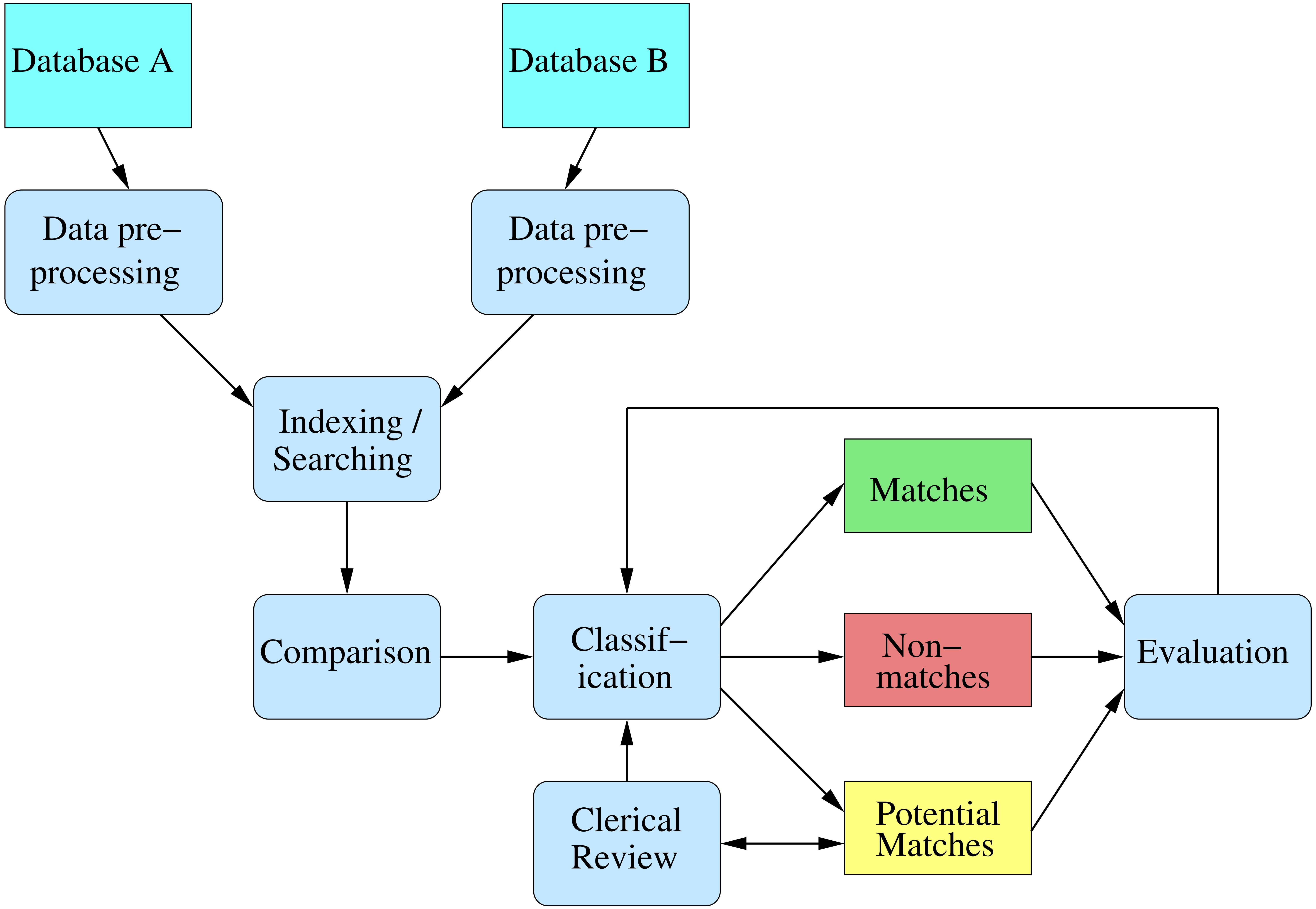
Apache Iceberg: An Architectural Look Under the Covers

A. Merced

Record Linkage

P. Christen

Record Linkage Process



[P. Christen , 2019]

Record Linkage Techniques

- Deterministic matching
  - Rule-based matching (complex to build and maintain)
- Probabilistic record linkage [Fellegi and Sunter, 1969]
  - Use available attributes for linking (often personal information, like names, addresses, dates of birth, etc.)
  - Calculate match weights for attributes
- “Computer science” approaches
  - Based on machine learning, data mining, database, or information retrieval techniques
  - Supervised classification: Requires training data (true matches)
  - Unsupervised: Clustering, collective, and graph based

[P. Christen , 2019]

Record Linkage/Entity Resolution Recipe

- Problem: Link references to the same entity
- Short Answers:
  - Random Forest with attribute similarity features
  - Deep Learning to handle text and noise
  - End-to-end solutions still being worked on

[X. L. Dong and T. Rekatsinas, 2018]

Data Integration and Data Fusion

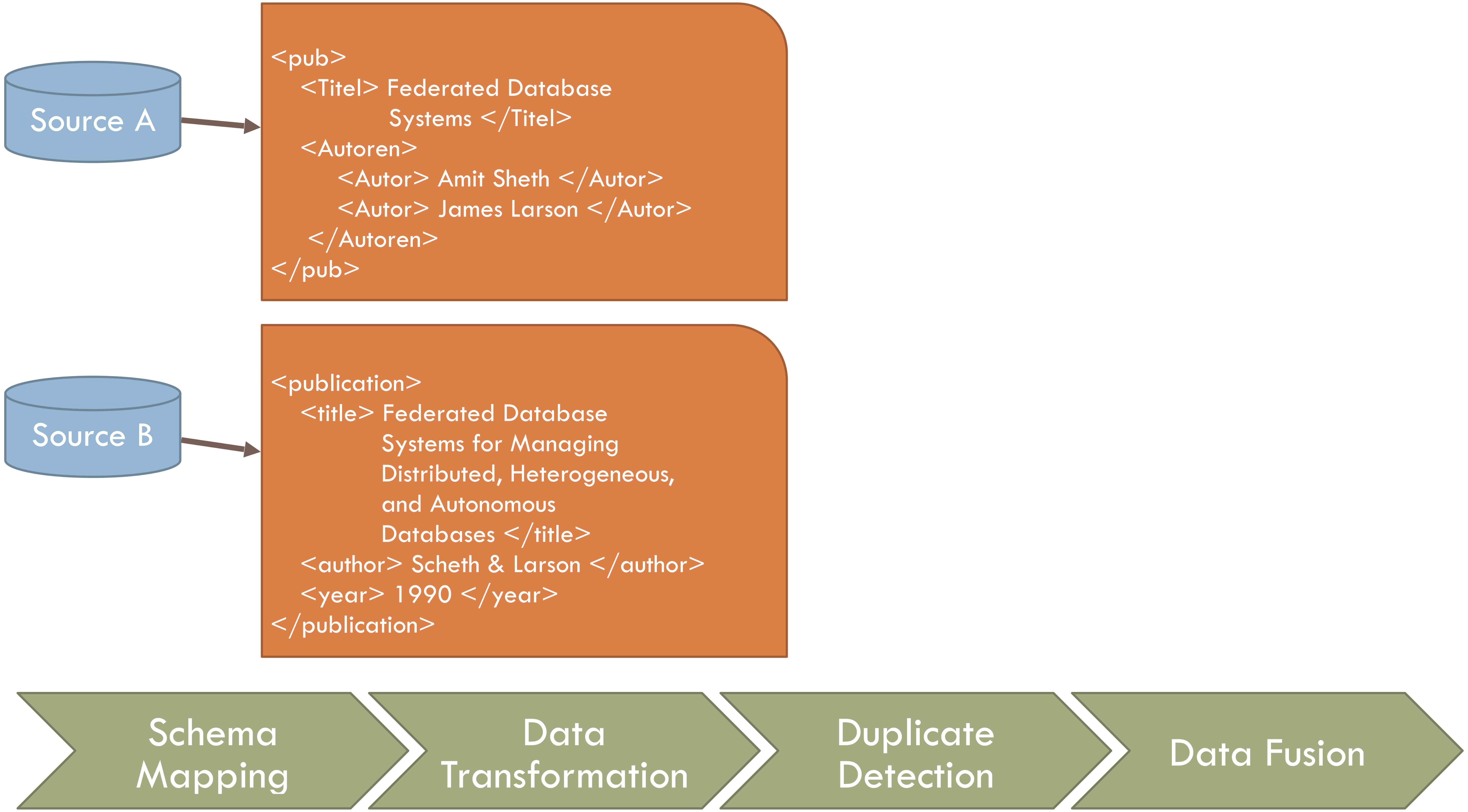
- Data Integration: focus on integrating data from different sources
- When sources are orthogonal, no problems
- What happens when two sources provide the same type of information and they **conflict**?
- Data Fusion: create a single object while resolving conflicting values

Data Fusion

Data Fusion— Resolving Data Conflicts in Integration

X. L. Dong and F. Naumann

Information Integration

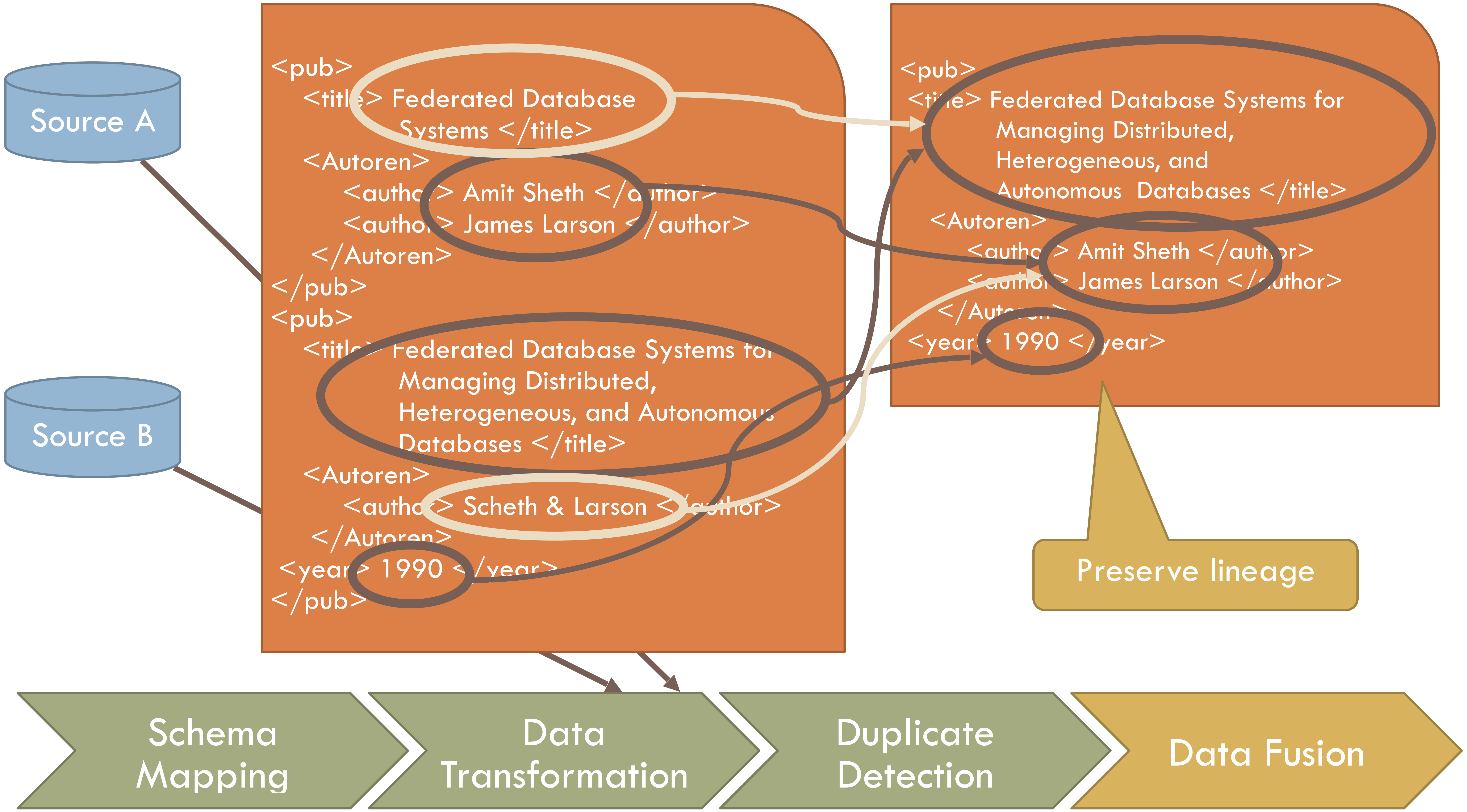


```
<pub>
   <Titel> Federated Database
            Systems </Titel>
   <Autoren>
       <Autor> Amit Sheth </Autor>
       <Autor> James Larson </Autor>
    </Autoren>
</pub>
```

```
<publication>
   <title> Federated Database
           Systems for Managing
           Distributed, Heterogeneous,
           and Autonomous
           Databases </title>
   <author> Scheth & Larson </author>
   <year> 1990 </year>
</publication>
```

[L. Dong and F. Naumann, 2009]

Information Integration


Source A
Source B
<pub>
<title> Federated Database Systems </title>
<Autoren>
<author> Amit Sheth </author>
<author> James Larson </author>
</Autoren>
</pub>
<pub>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<Autoren>
<author> Scheth & Larson </author>
</Autoren>
<year> 1990 </year>
</pub>
<pub>
<title> Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases </title>
<Autoren>
<author> Amit Sheth </author>
<author> James Larson </author>
</Autoren>
<year> 1990 </year>
Preserve lineage
Schema Mapping
Data Transformation
Duplicate Detection
Data Fusion

[L. Dong and F. Naumann, 2009]

Outline

- ~~Combining Data~~
- ~~Data Integration~~
- ~~Data Matching (Entity Resolution)~~
- ~~Data Fusion~~
- Data Fusion Techniques: Wednesday
    - Integrating Conflicting Data: The Role of Source Dependence, X. L. Dong et al., 2009
    - **Quiz** at the beginning of class