Advanced Data Management (CSCI 640/490)

Data Integration

Dr. David Koop

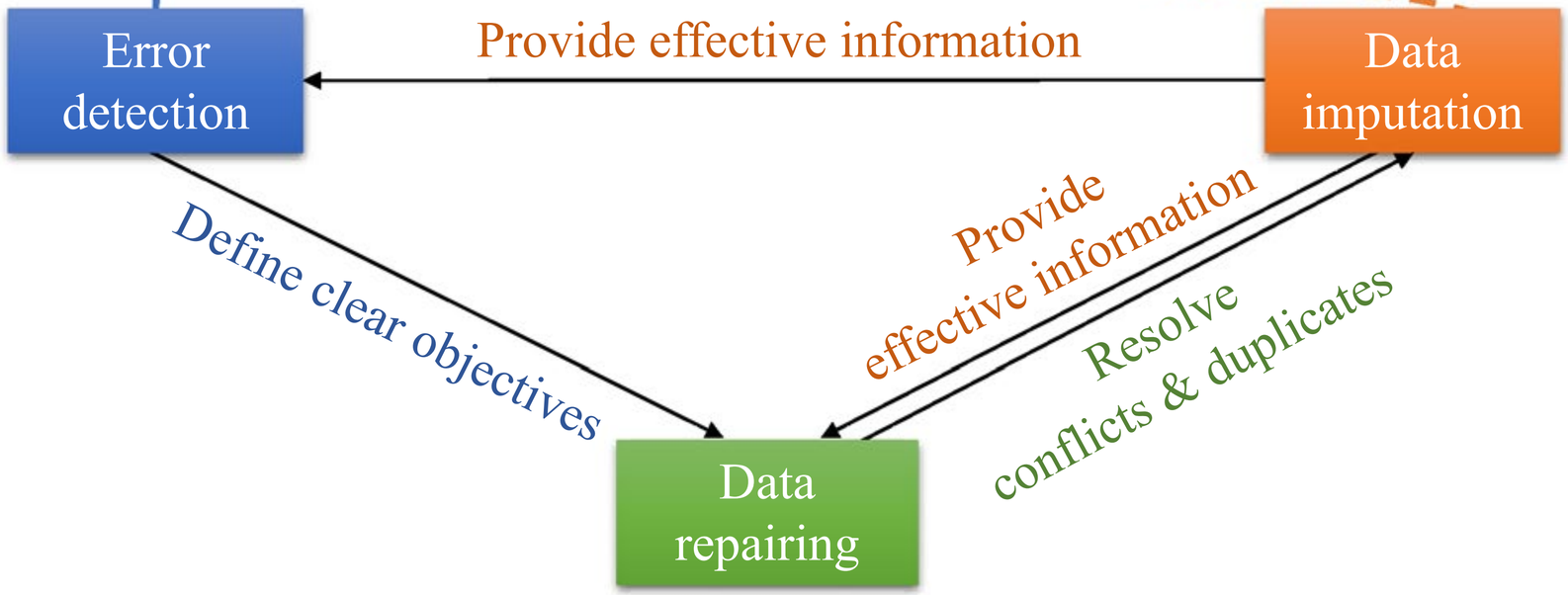
Data Cleaning Types

- How can statistical techniques improve efficiency or reliability of data cleaning? (Data Cleaning **with** Statistics)
  - Example: Wrangler, polars
- How how can we improve the reliability of statistical analytics with data cleaning? (Data Cleaning **for** Statistics)
  - Example: SampleClean

[D. Haas et al., 2016]

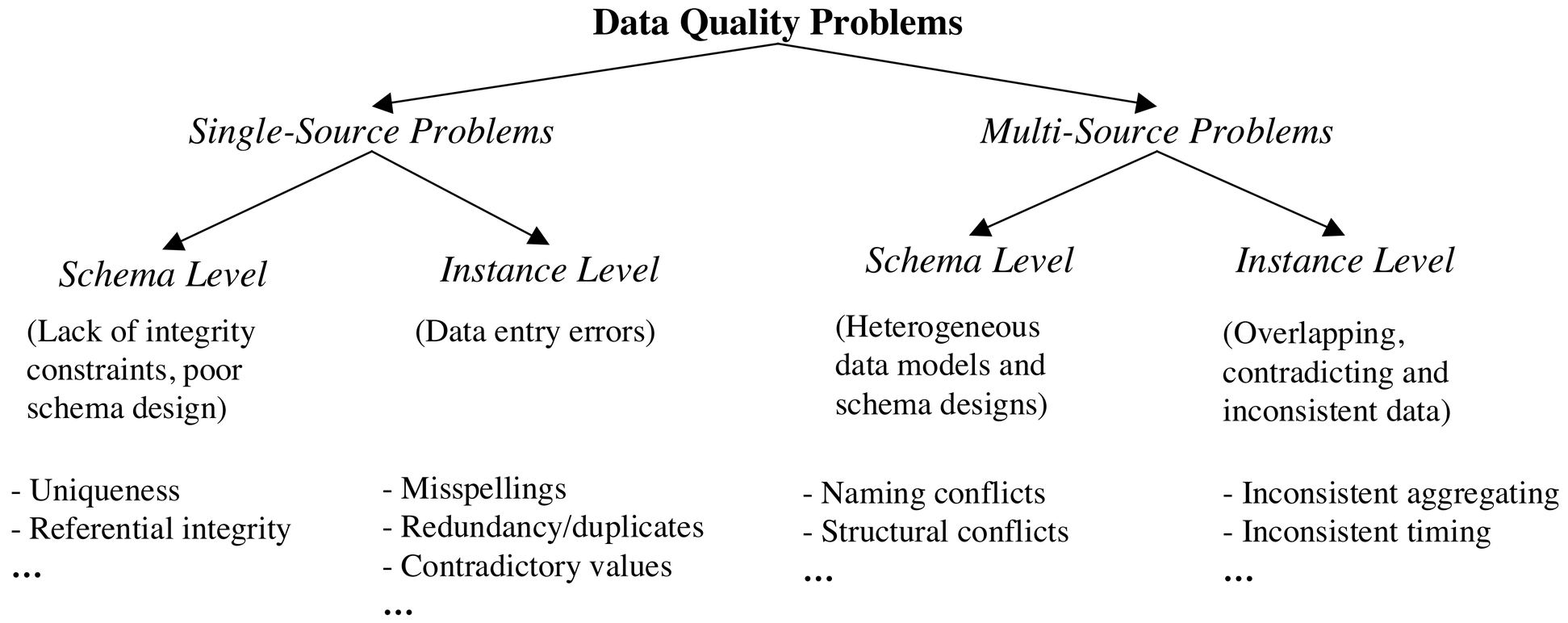
Data Cleaning Overview

	Length (cm)	Size	Material	Price ($)
Pants 1	95	XL	Silk	70
Pants 2	0.9	M	Cotton	30
Pants 3	100	L	?	35



[J. Zhu et al., 2024]

Classifying Data Quality Problems

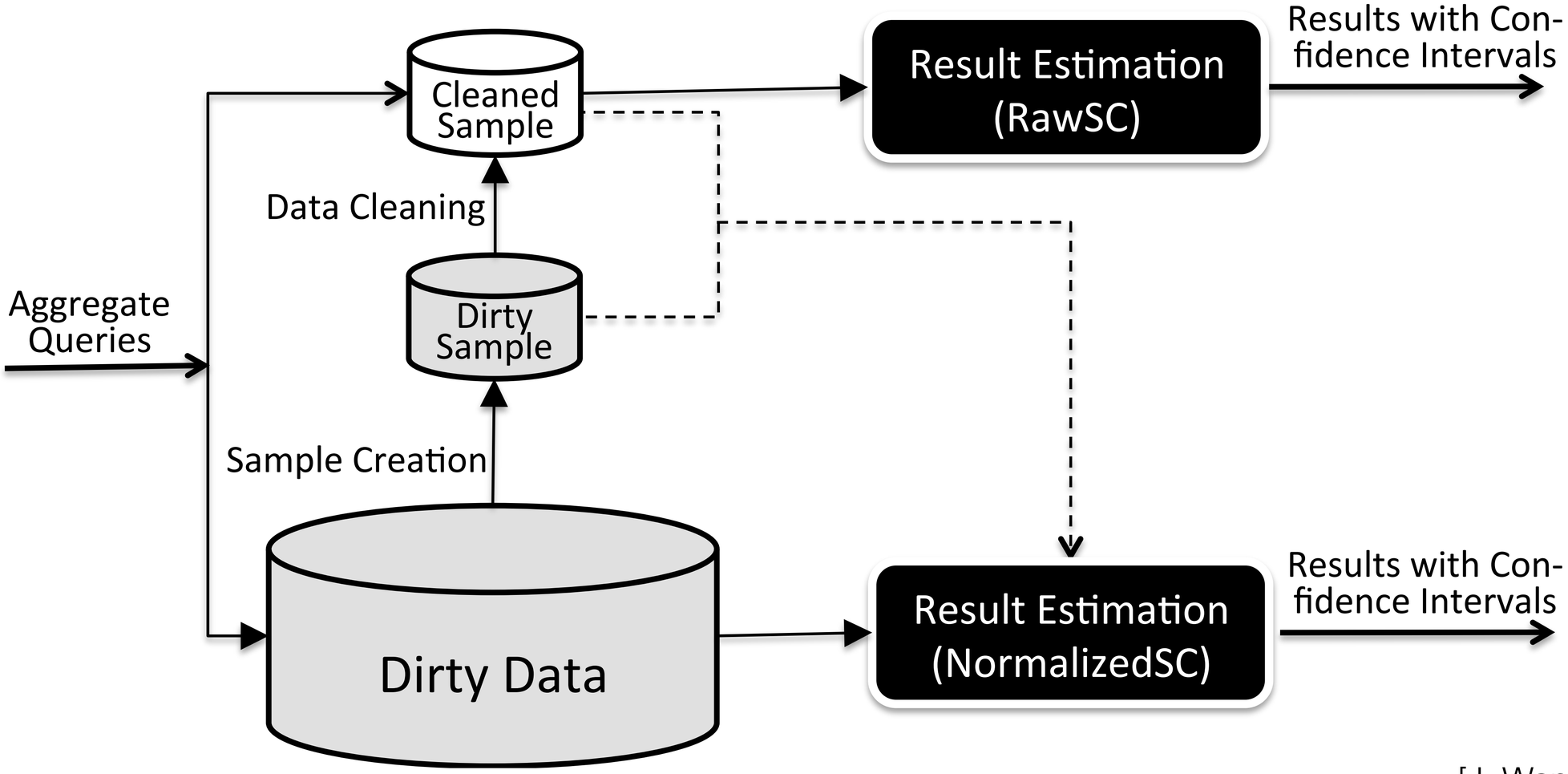


[E. Rahm & H. H. Do, 2000]

SampleClean (and Variants)

- Dirty Data?
  - Missing Values
  - Duplicate Values
  - Incorrect Values
  - Inconsistent Values
- Estimate query results using a sample of the data
- Two ideas:
  - Direct Estimate
  - Correction

SampleClean Framework



[J. Wang et al., 2014]

HoloClean

- A holistic data cleaning framework that combines qualitative methods with quantitative methods:
  - Qualitative: use integrity constraints or external data sources
  - Quantitative: use statistics of the data
- Driven by probabilistic inference. Users only need to provide a dataset to be cleaned and describe high-level domain specific signals.
- Can scale to large real-world dirty datasets and perform automatic repairs with high accuracy

[T. Rekatsinas et al., 2017]

HoloClean

Input

Dataset to be cleaned

	DBAName	Address	City	State	Zip
t1	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60609
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60609
t4	Johnnyo's	3465 S Morgan ST	Cicago	IL	60608

Denial Constraints

c1: DBAName → Zip
c2: Zip → City, State
c3: City, State, Address → Zip

Matching Dependencies

m1: Zip = Ext_Zip → City = Ext_City
m2: Zip = Ext_Zip → State = Ext_State
m3: City = Ext_City ∧ State = Ext_State∧
∧ Address = Ext_Address → Zip = Ext_Zip

External Information

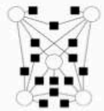
Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608
1208 N Wells ST	Chicago	IL	60610
259 E Erie ST	Chicago	IL	60611
2806 W Cermak Rd	Chicago	IL	60623

The HoloClean Framework

1. Error detection module

2. Automatic compilation to a probabilistic graphical model

3. Repair via statistical learning and infernece



Output

Proposed Cleaned Dataset

	DBAName	Address	City	State	Zip
t1	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	60608
t2	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t3	John Veliotis Sr.	3465 S Morgan ST	Chicago	IL	**60608**
t4	**John Veliotis Sr.**	3465 S Morgan ST	**Chicago**	IL	60608

Marginal Distribution of Cell Assignments

Cell	Possible Values	Probability
t2.Zip	60608	0.84
	60609	0.16
t4.City	Chicago	0.95
	Cicago	0.05
t4.DBAName	John Veliotis Sr.	0.99
	Johnnyo's	0.01

[T. Rekatsinas et al., 2017]

Data Cleaning and AI

- Traditional Methods are often efficient and interpretable
- Deep Learning is expensive and hard to understand but can be more effective
- Machine Learning provides a balance?

		Cost	Generalization	Interpretability	Efficiency	Effectiveness
	Traditional	¥	⚛	💬💬💬	⚡⚡⚡	🌷🌷
AI	ML	¥¥	⚛⚛	💬💬	⚡	🌷🌷
AI	DL	¥¥¥	⚛⚛⚛	💬	⚡	🌷🌷🌷

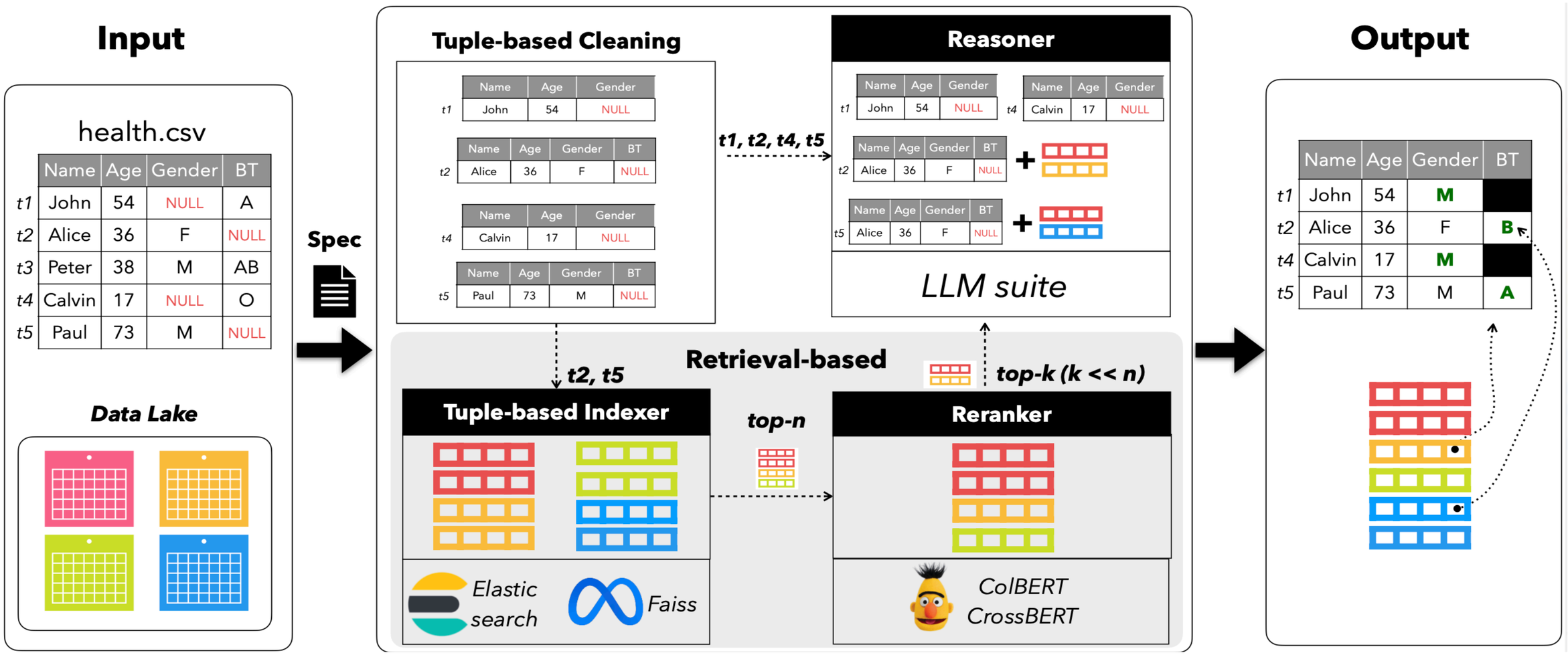
[J. Zhu et al., 2024]

Data Repair using LLMs (RetClean)

- Non-retrieval based: Send tuple to LLMs and identify tuple(s) and column(s) to be fixed
- Retrieval-based:
  - Indexer: Get top-k relevant tuples from a database/data lake
  - Reranker: Rank relevance using ColBERT/CrossBERT
  - Reasoner: Determine, using LLM, which tuple and value to use for fix
  - Reasoner keeps track of lineage

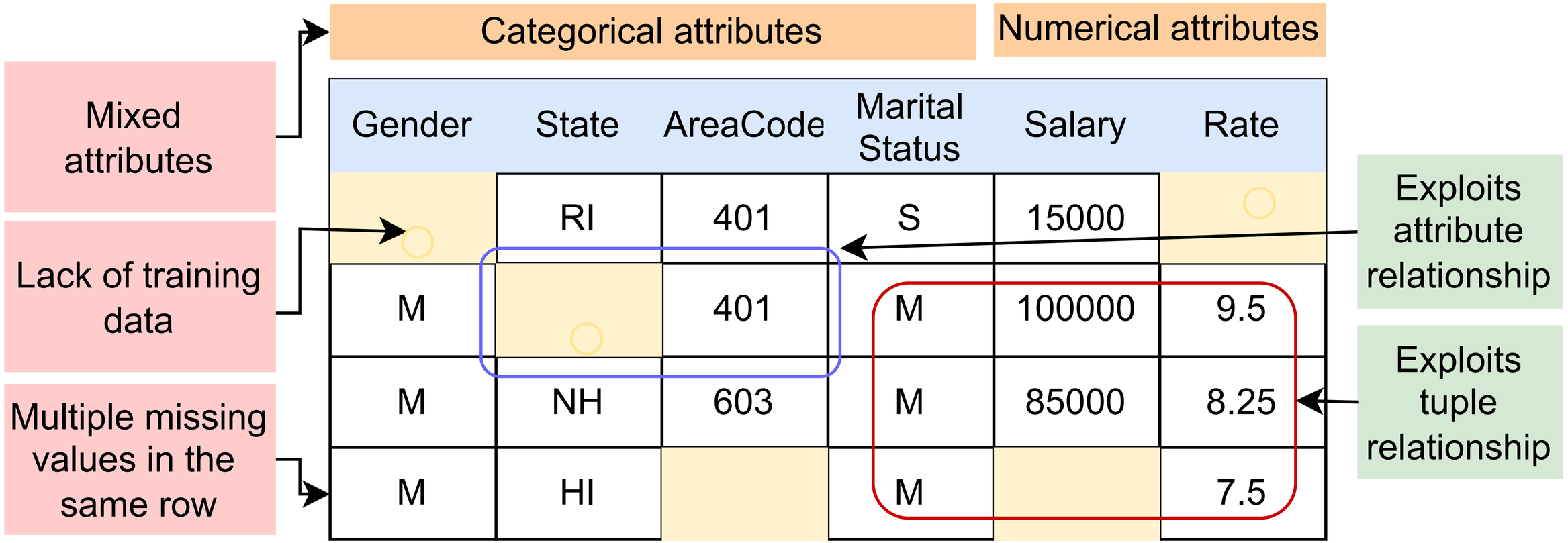
[Naeem et al., 2024]

Data Repair using LLMs (RetClean)



[Naeem et al., 2024]

Data Imputation Challenges



[R. Cappuzzo et al., 2024]

CSAN Panel: Real Jobs in the Real World

- Tuesday, Oct. 7, 5:30–7:30pm
- Provides an insight into jobs from NIU alumni
- Food is Provided
- Sponsored by the Computer Science Alumni Network and the NIU Alumni Association

Test 1

- This Wednesday, October 8, 12:30-1:45pm in PM 103
- In-Class, paper/pen & pencil
- Covers material through this week
- Format:
  - Multiple Choice
  - Free Response
  - One extra 2-sided page for CSCI 640 Students
- Info will be on the course webpage

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

Outline

- Data Integration
- Data Matching (Entity Resolution)
- Data Fusion: Monday
- 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

Introduction to Data Integration

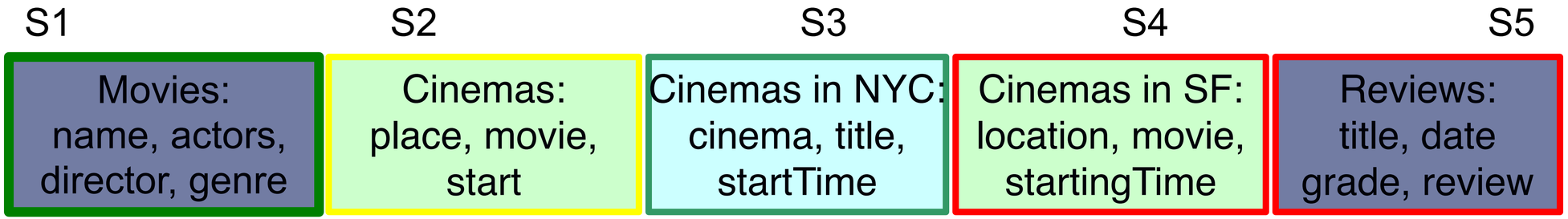
A. Doan, A. Halevy, and Z. Ives

Data Integration

```
select title, startTime
from Movie, Plays
where Movie.title=Plays.movie AND
            location="New York"  AND
            director="Woody Allen"
```

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.



[AH Doan et al., 2012]

Data Integration & Data Matching

- Data Integration: focus on integrating data from different sources
- Data Matching (aka Entity Resolution aka Record Linkage):
  want to know that two entities (often in different sources) are the same "real" entity

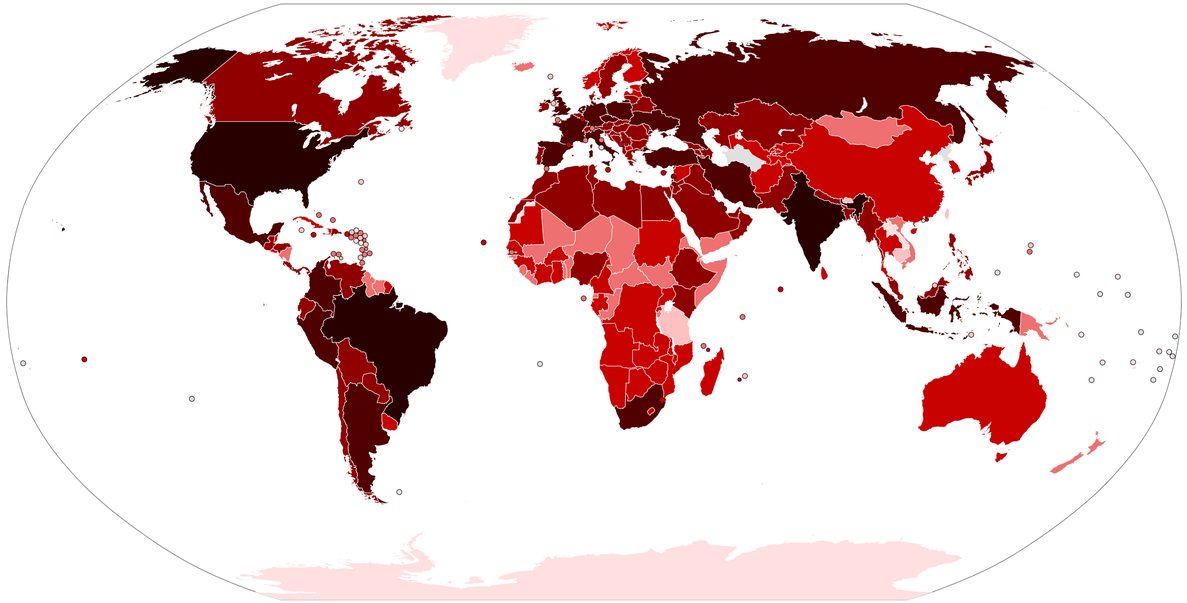
Record Linkage Motivation

- Often data from different sources need to be integrated and linked
  - To allow data analyses that are impossible on individual databases
  - To improve data quality
  - To enrich data with additional information
- **Lack of unique entity identifiers** means that linking is often based on personal information
- When databases are linked across organisations, maintaining privacy and confidentiality is vital
- The linking of databases is challenged by **data quality**, **database size**, and **privacy concerns**

[P. Christen , 2019]

Motivating Example

- Preventing the outbreak of epidemics requires monitoring of occurrences of unusual patterns of symptoms, ideally in real time
- Data from many different sources will need to be collected (including travel and immigration records; doctors, emergency and hospital admissions; drug purchases; social network and location data; and possibly even animal health data)



[P. Christen , 2019], image: [Pharexia, Wikipedia]

Record Linkage

P. Christen