Data Cleaning and Integration

Felix Naumann, Fabian Panse, Matteo Paganelli
WS 2023/2024

Agenda

- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data Cleaning and Integration
- ❑ Seminar Topics

Agenda

- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data Cleaning and Integration
- ❑ Seminar Topics

Information Systems Team

Sebastian **Schmidl**

Phillip **Wenig**

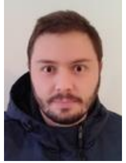
Diana **Stephan**

Prof. Felix **Naumann**

Dr. Matteo **Paganelli**

Dr. Fabian **Panse**

Tobias **Bleifuß**


Alejandro
Sierra-Múnera

Leon **Bornemann**

Duplicate Detection

Data Change **Data Fusion**

project **AKITA**

Entity Search

Data Profiling **Information Integration**

project **DataKnoller**

Web Science

Data Scrubbing

Data as a Service

project **AI4ART**

Information Quality

Data Cleansing

Text Mining

Linked Open Data

Dependency Detection

CSV parsing

Distributed Computing

Knowledge Management for the Arts

Web Data

project **Janus**

Entity Recognition

project **Metanome**

Data Preparation

Change Exploration

Sedir **Mohammed**

Gerardo **Vitagliano**

Mazhar **Hameed**

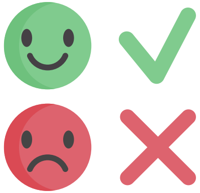
Daniel **Lindner**

Youri **Kaminsky**

Agenda

- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data Cleaning and Integration
- ❑ Seminar Topics

Course Information



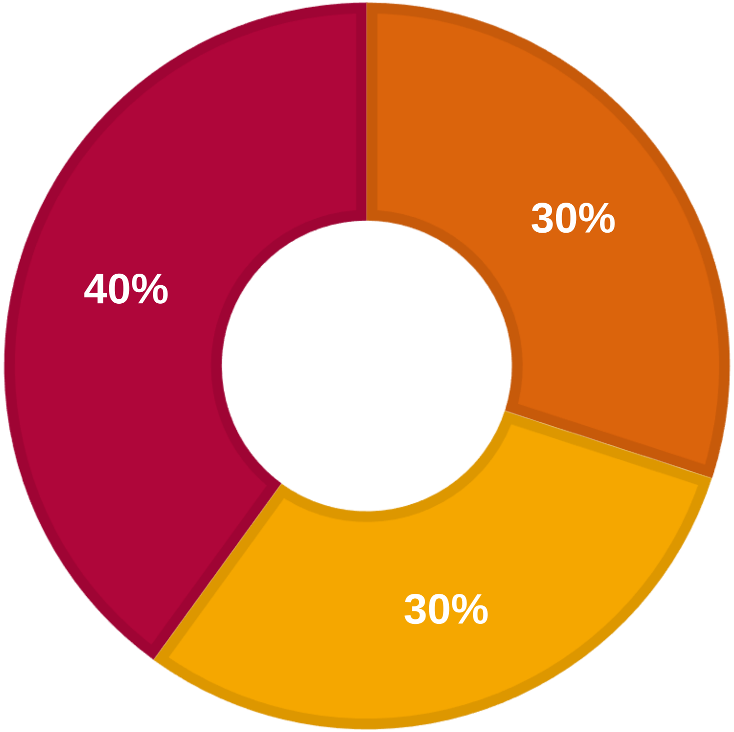
- **General Information**
  - Semester hours per week: 2
  - ECTS: 3
  - Total working time: 90 h (≃ 6 h per week)
  - Language: English
  - Maximum number of participants: 8
  - Enrollment period: 01.10.2023 - 31.10.2023
- **Tasks**
  - Writing a seminar report for a given topic
  - Giving a presentation for the same topic
  - Reviewing 2-3 other seminar reports

Grading

Writing Seminar Report

Giving Seminar Presentation

Reviewing other Seminar Reports



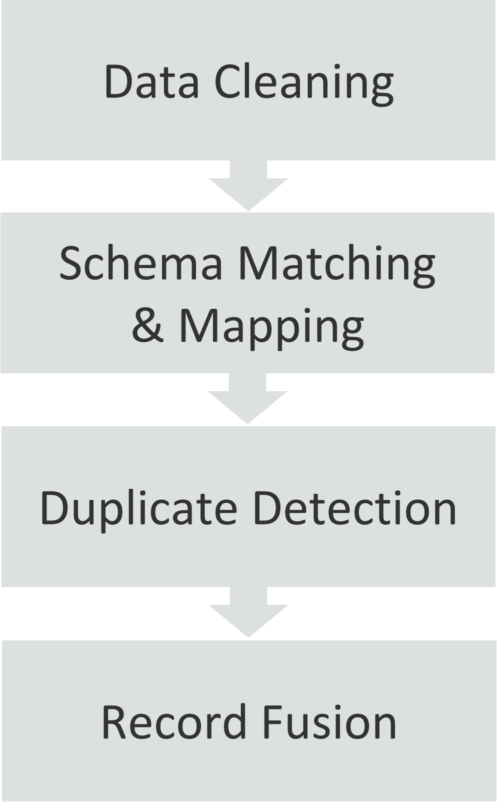
Schedule

Date	Topic
2023-10-16	**Seminar introduction**
2023-10-20 11:59 a.m.	Participation feedback and topic requests (online)
2023-10-20 6:00 p.m.	Notification of participation (online)
2023-10-23	Topic assignements and first discussions
2023-10-30 - 2023-12-18	Weekly meetings and progress reports
2023-12-25	Christmas break
2024-01-01	New Years break
2024-01-08	Submission of the seminar papers and review assignments
2024-01-15	Weekly meetings and progress reports
2024-01-22	Submission and discussion of paper reviews
2024-01-29 & 2024-02-05	Seminar presentations

Agenda

- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data Cleaning and Integration
- ❑ Seminar Topics

Data Integration Pipeline



Data Cleaning

Schema Matching & Mapping

Duplicate Detection

Record Fusion

id	name	size	city
1	Bob Lee	718	Rmo

→

id	name	size	city
1	Bob Lee	178	Rom

id	name	size	city
1	Bob Lee	178	Rom
2	Lilly Hall	169	Ulm

fname	lname	height	age
Tom	Britt	192	23
Bob	Lee	180	31

Matching:
{name} ↔ {fname,lname}
{size} ↔ {height}

id	fname	lname	size	city	age
1	Bob	Lee	178	Rom	-
2	Lilly	Hall	169	Ulm	-
					25

records 2 and 3 are duplicates!

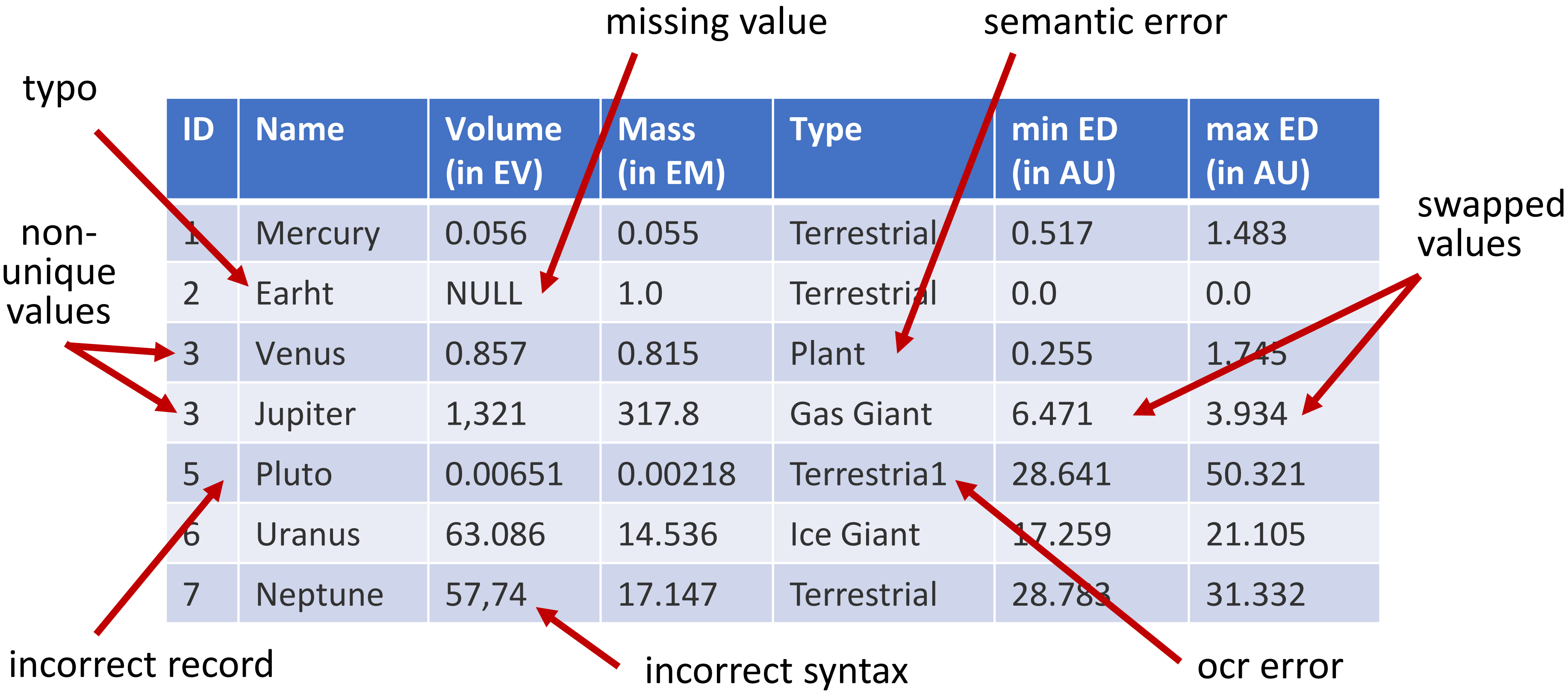
Duplicate Elimination
(Entity Resolution)

id	fname	lname	size	city	age
2	Lilly	Hall	169	Ulm	-
3	Lill	Hall	169	-	25

→

id	fname	lname	size	city	age
2	Lilly	Hall	169	Ulm	25

Data Cleaning



ID	Name	Volume (in EV)	Mass (in EM)	Type	min ED (in AU)	max ED (in AU)
1	Mercury	0.056	0.055	Terrestrial	0.517	1.483
2	Earht	NULL	1.0	Terrestrial	0.0	0.0
3	Venus	0.857	0.815	Plant	0.255	1.745
3	Jupiter	1,321	317.8	Gas Giant	6.471	3.934
5	Pluto	0.00651	0.00218	Terrestria1	28.641	50.321
6	Uranus	63.086	14.536	Ice Giant	17.259	21.105
7	Neptune	57,74	17.147	Terrestrial	28.783	31.332

typo

missing value

semantic error

non-unique values

swapped values

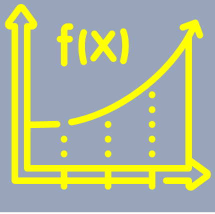
incorrect record

incorrect syntax

ocr error

Data Cleaning

Statistical techniques



Pattern-based techniques

[\w-]+@([\w-]+\.)+[\w-]+

Constraint-based techniques

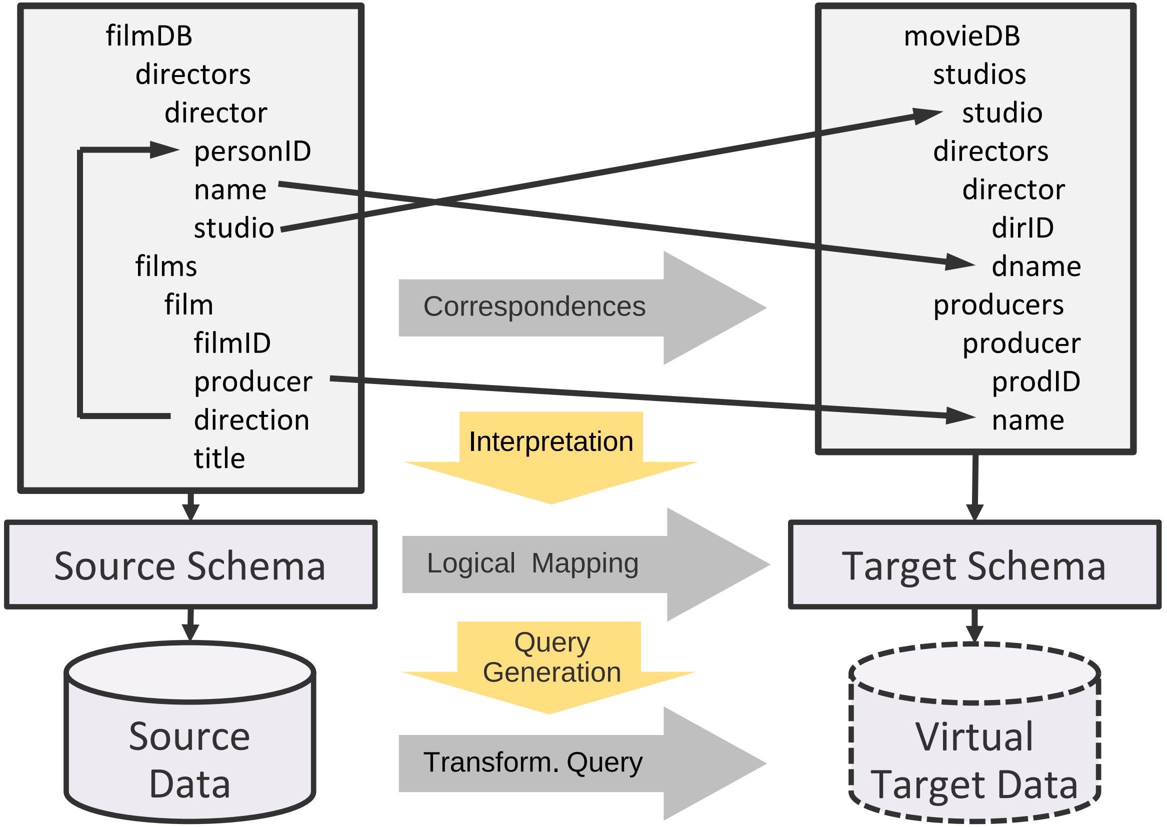
FDs: ZIP → City

ISBN is unique

Knowledge-based techniques

Eiffel Tower LocatedIn Paris

Schema Matching and Mapping



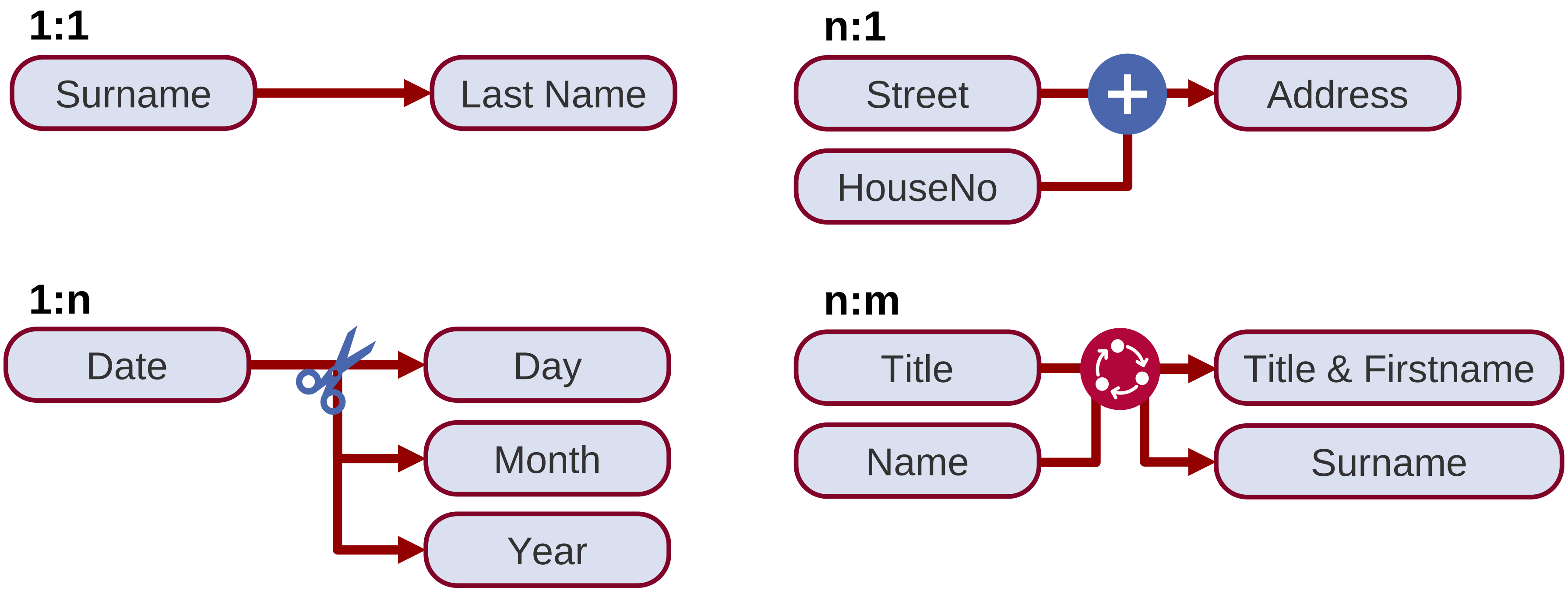
Schema Matching:

- Correspondencies between Schema Elements
- 1:1, 1:n, n:1, n:m

Schema Mapping:

- GaV, LaV, GLaV
- Tuple-Generating Dependencies (tgd)

Schema Matching - Correspondences



1:1

Surname → Last Name

n:1

Street, HouseNo → + → Address

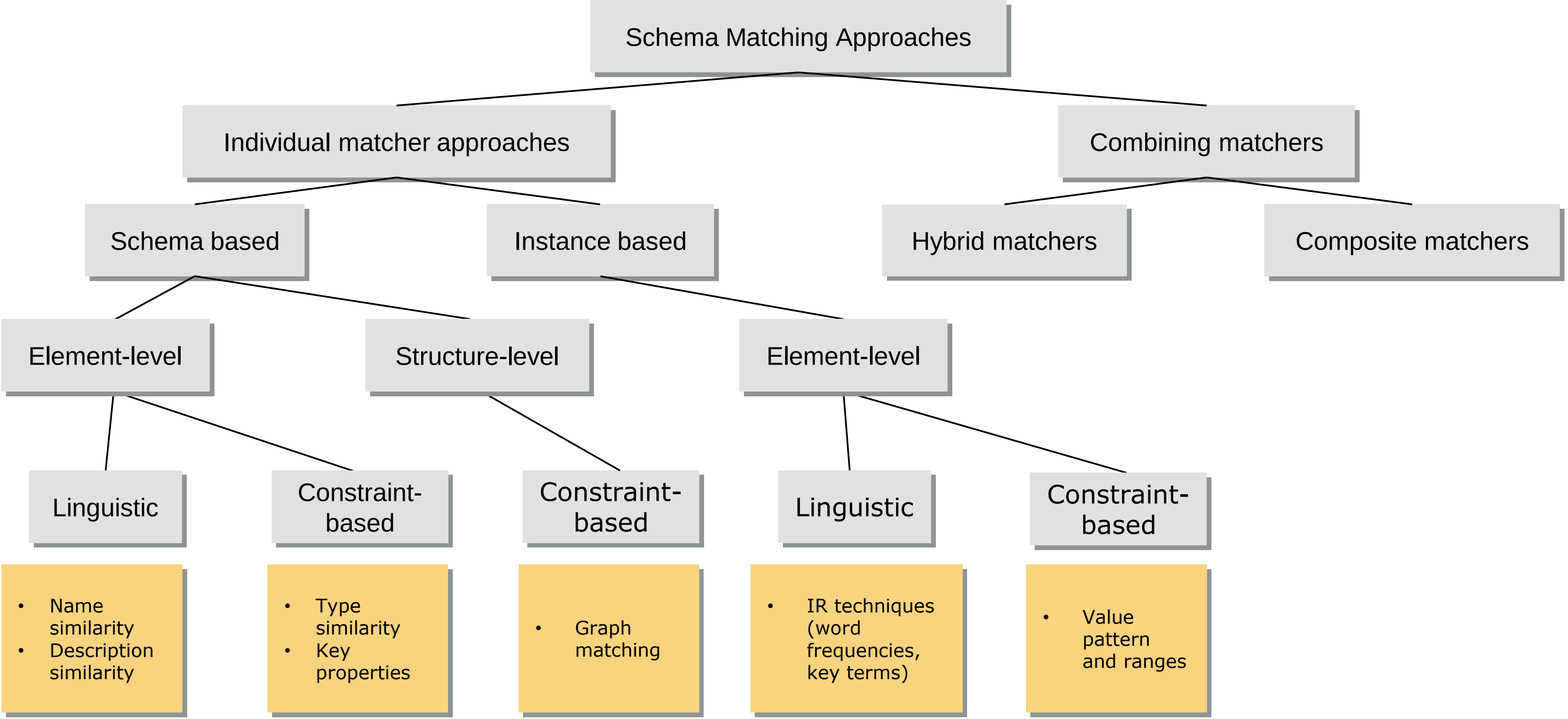
1:n

Date → Day, Month, Year

n:m

Title, Name → Title & Firstname, Surname

Schema Matching - Approaches



Schema Matching Approaches

Individual matcher approaches

Combining matchers

Schema based

Instance based

Hybrid matchers

Composite matchers

Element-level

Structure-level

Element-level

Linguistic

Constraint-based

Constraint-based

Linguistic

Constraint-based

- Name similarity
- Description similarity

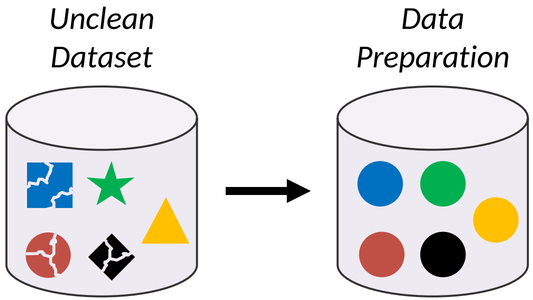
- Type similarity
- Key properties

- Graph matching

- IR techniques (word frequencies, key terms)

- Value pattern and ranges

Duplicate Elimination Pipeline



Segmentation:

Address
„33101 Miami, USA“

⬇

ZIP	City	Country
„33101“	„Miami“	„USA“

Standardization:

Unclean Value		Clean Value
„31.03.2021“	⇒	„2021-03-31“
„1st May, 2021“	⇒	„2021-05-01 “
„27.02.21“	⇒	„2021-02-27“

Cleaning:

City	Country
„M1ami“	„USB“

⬇

City	Country
„Miami“	„USA“

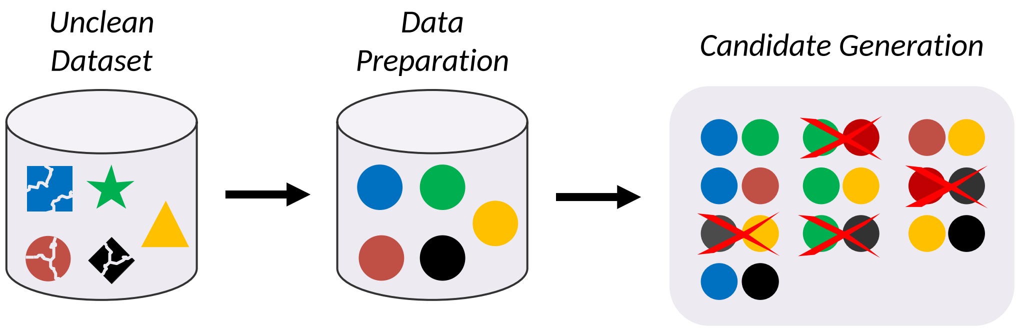
Enrichment:

City
„Miami“

⬇

City	LAT	LONG
„Miami“	25.76	-80.2

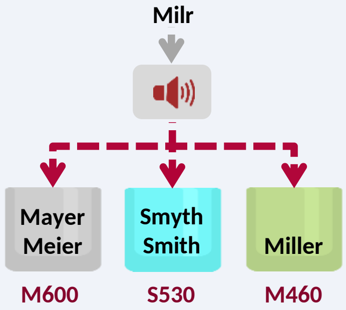
Duplicate Elimination Pipeline



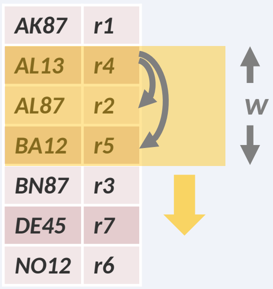
Unclean Dataset

Data Preparation

Candidate Generation

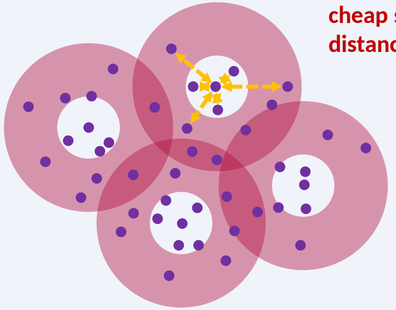


Blocking



AK87	r1
AL13	r4
AL87	r2
BA12	r5
BN87	r3
DE45	r7
NO12	r6

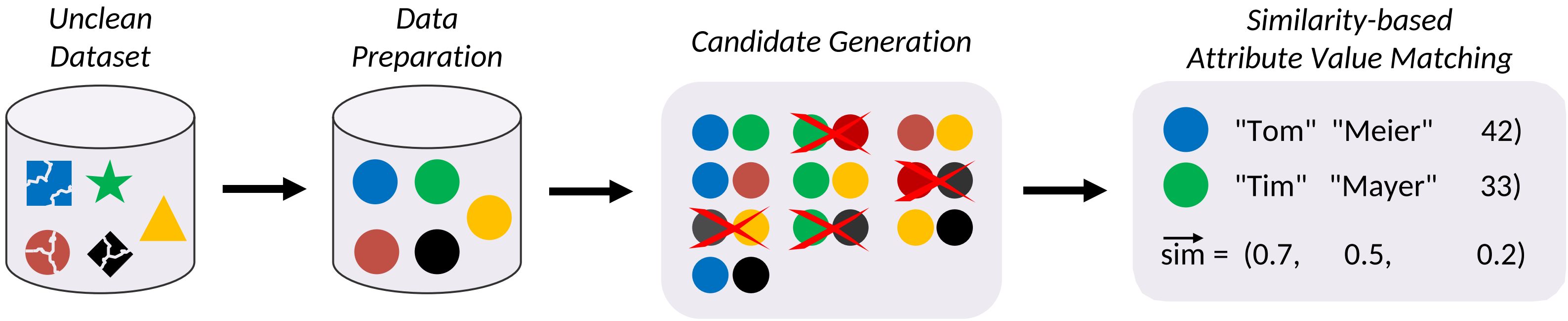
Sorted Neighborhood Method



cheap similarity / distance measure

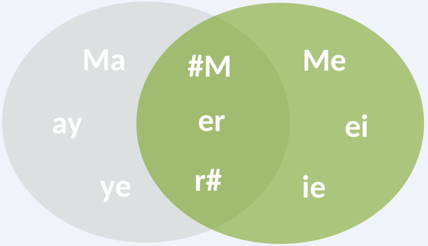
Canopy Clustering

Duplicate Elimination Pipeline



●	"Tom"	"Meier"	42)
●	"Tim"	"Mayer"	33)
$\overrightarrow{\text{sim}}$ =	(0.7,	0.5,	0.2)

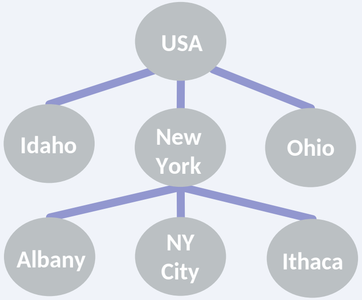
Meier vs. Mayer



Set/Token-based

	M	e	i	e	r
M	0	1	2	3	4
a	1	1	2	3	4
y	2	2	2	3	4
e	3	2	3	2	3
r	4	3	3	3	2

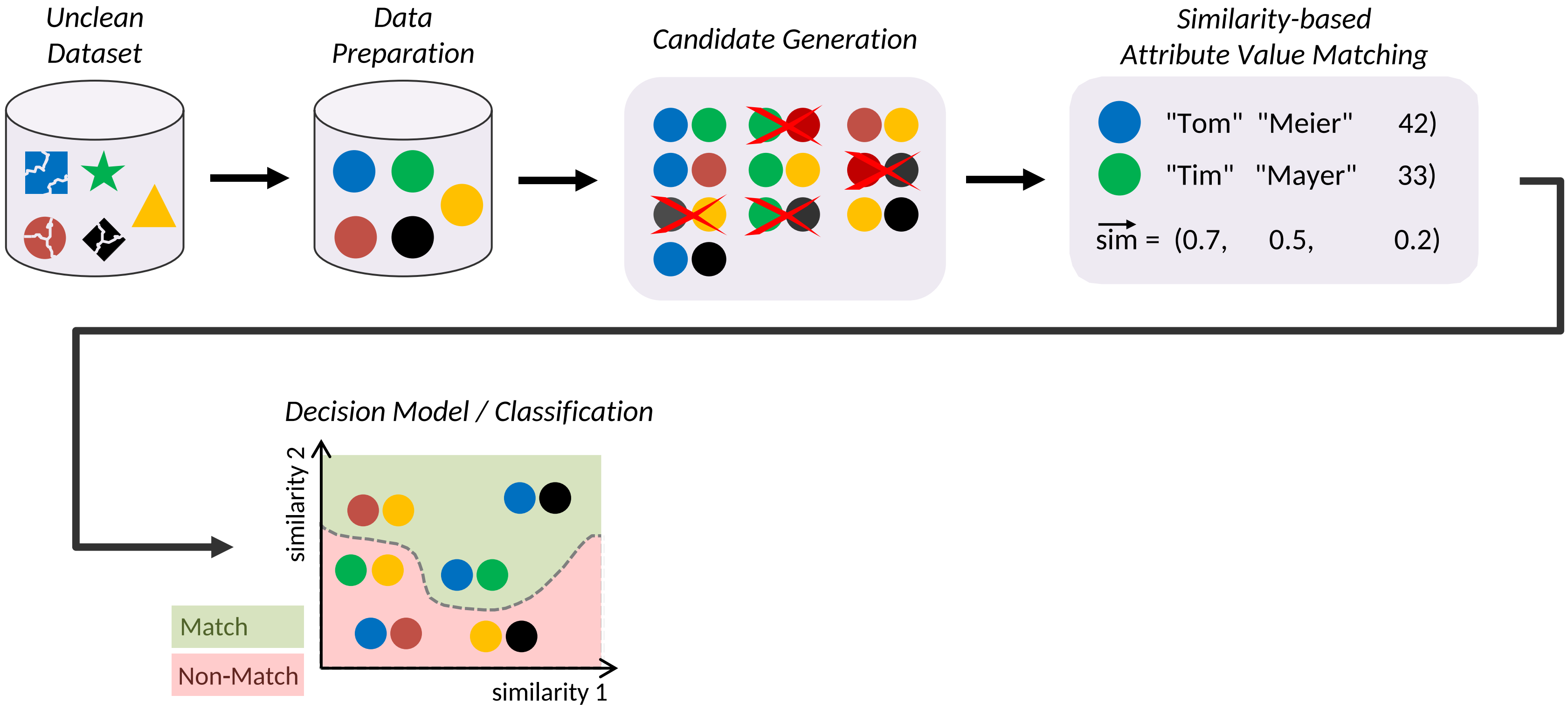
Sequence-based



Semantic

Phonetic

Duplicate Elimination Pipeline



Unclean Dataset

Data Preparation

Candidate Generation

Similarity-based Attribute Value Matching

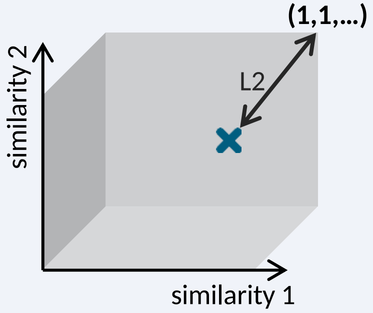
"Tom" "Meier" 42)

"Tim" "Mayer" 33)

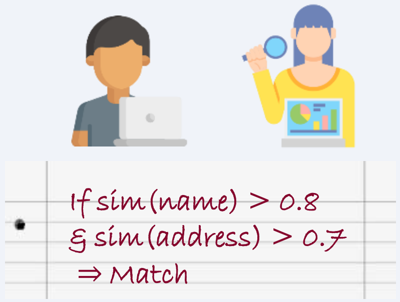
$\overrightarrow{sim}$ = (0.7, 0.5, 0.2)

Decision Model / Classification

Duplicate Elimination Pipeline

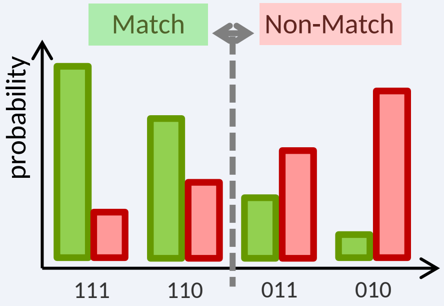


Distance-based

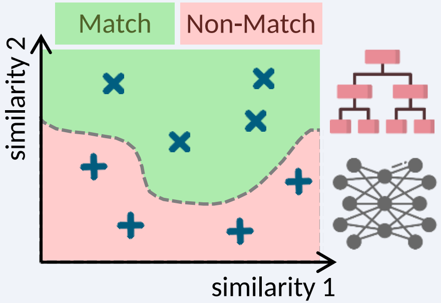


If sim(name) > 0.8
& sim(address) > 0.7
⇒ Match

Knowledge Rules

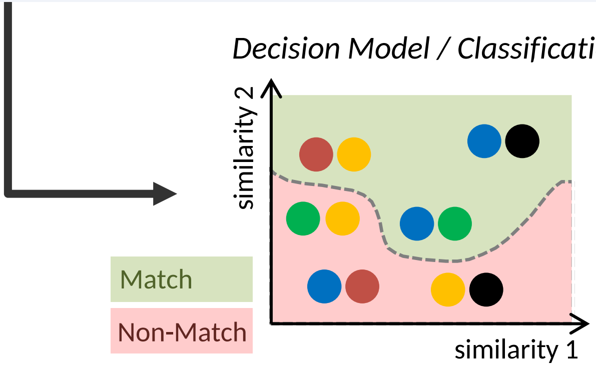


Statistical

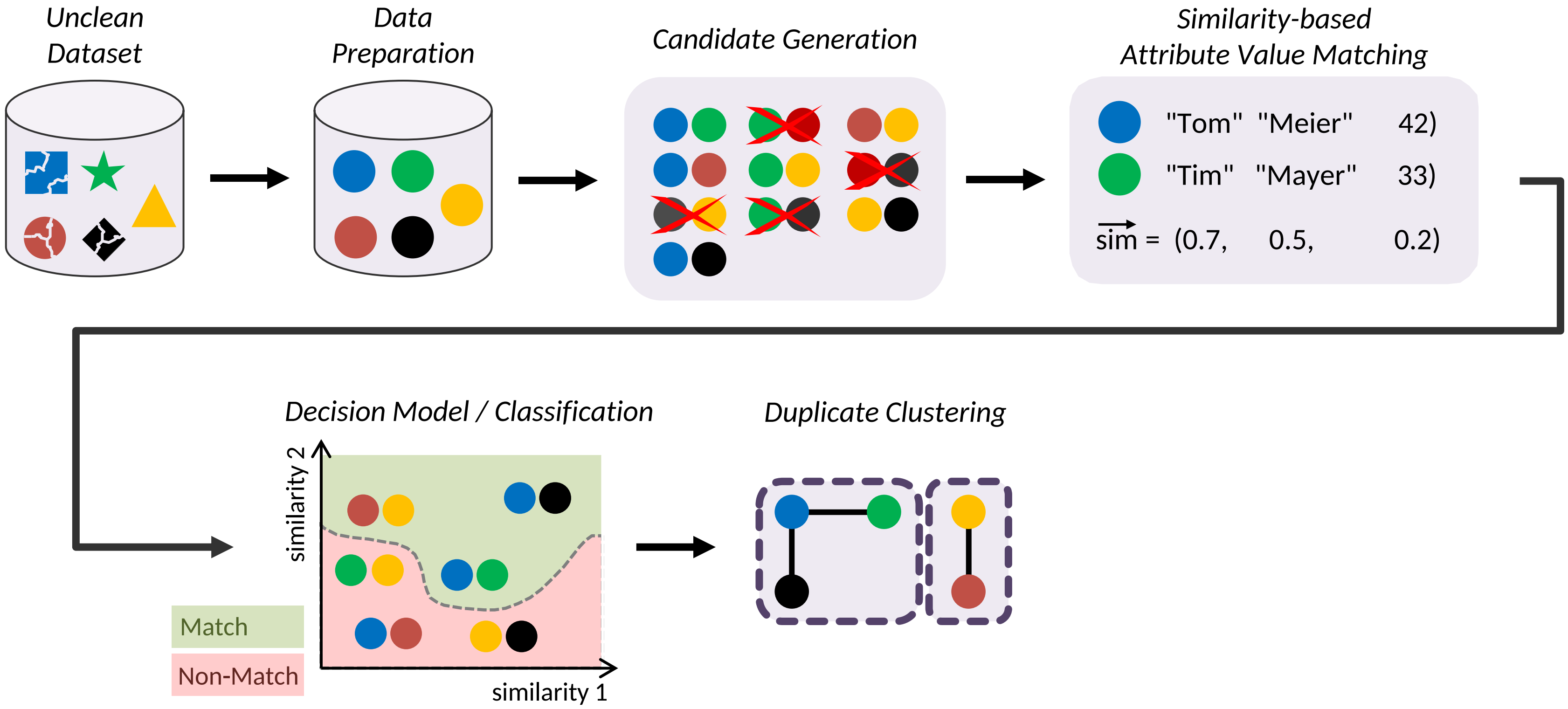


Machine Learning

Decision Model / Classification



Duplicate Elimination Pipeline



Unclean Dataset

Data Preparation

Candidate Generation

Similarity-based Attribute Value Matching

"Tom" "Meier" 42)

"Tim" "Mayer" 33)

$\overrightarrow{\text{sim}}$ = (0.7, 0.5, 0.2)

Decision Model / Classification

similarity 2

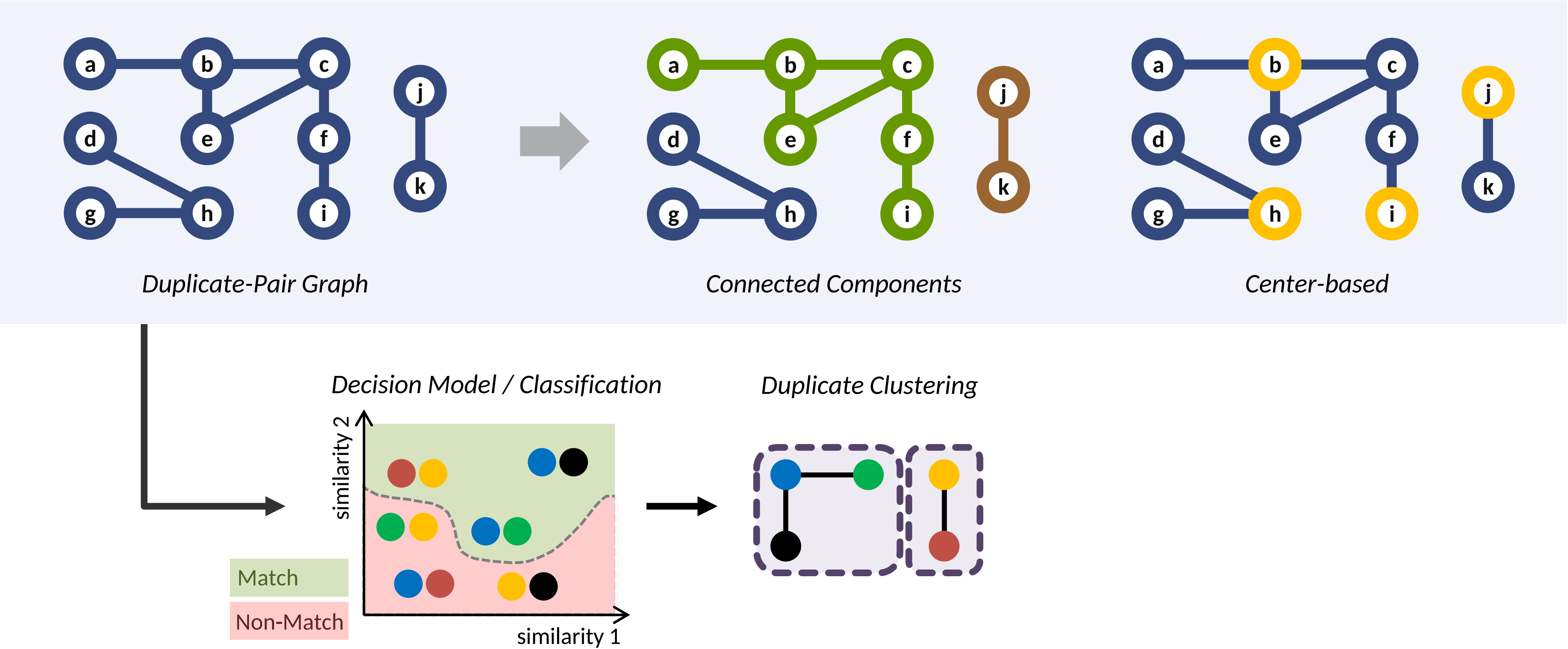
similarity 1

Match

Non-Match

Duplicate Clustering

Duplicate Elimination Pipeline



Duplicate-Pair Graph

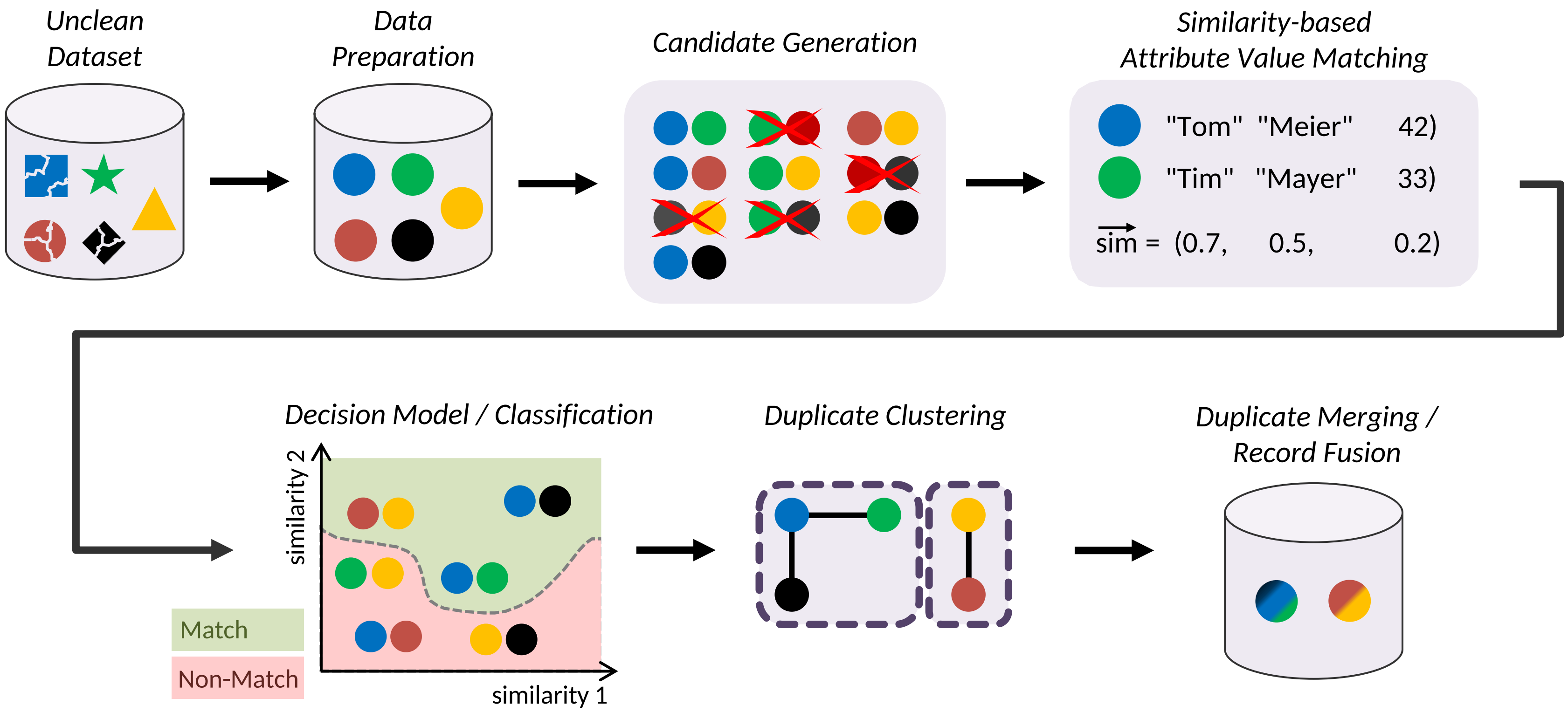
Connected Components

Center-based

Decision Model / Classification

Duplicate Clustering

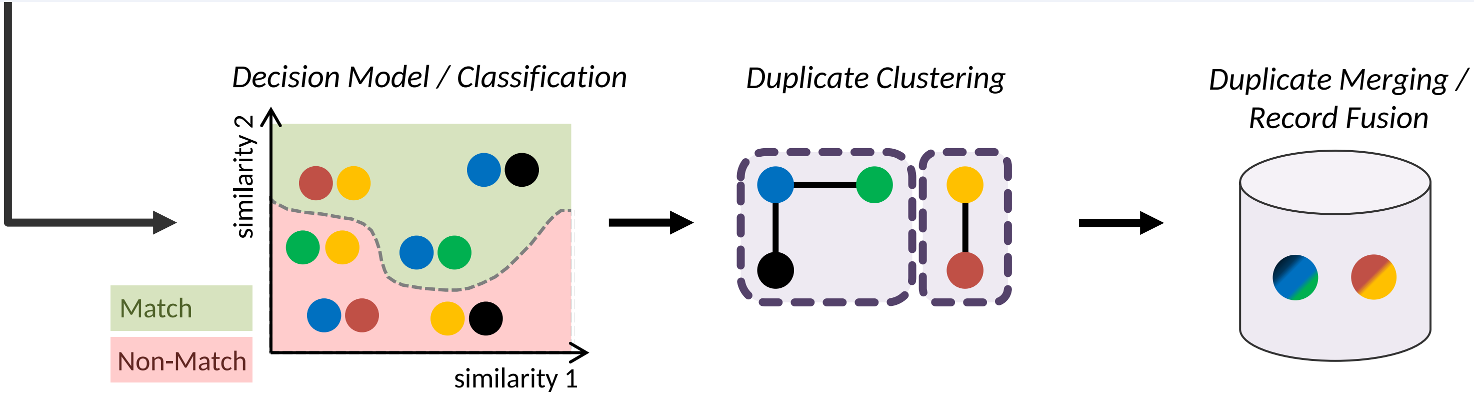
Duplicate Elimination Pipeline


Unclean Dataset
Data Preparation
Candidate Generation
Similarity-based Attribute Value Matching
"Tom" "Meier" 42)
"Tim" "Mayer" 33)
$\overrightarrow{sim}$ = (0.7, 0.5, 0.2)
Decision Model / Classification
similarity 2
similarity 1
Match
Non-Match
Duplicate Clustering
Duplicate Merging / Record Fusion

Duplicate Elimination Pipeline

longest	*voting*	*sum*	*concat*	*newest*	
Kim	Doe	5.20	James	Detroit	2007-11-09 T 11:20 UTC
K.	Doe	6.99	Jim	Chicago	2012-05-19 T 13:10 UTC
Kimberly	Smith	10.00	James	Paris	2001-10-23 T 08:32 UTC
Kimberly	Doe	22.19	James Jim	Chicago	
Deciding		*Mediating*		*Using Metadata*	

- Source dependencies
- Source trustworthiness
- Machine Learning approaches



Decision Model / Classification
similarity 2
similarity 1
Match
Non-Match
Duplicate Clustering
Duplicate Merging / Record Fusion

Deep Entity Matching

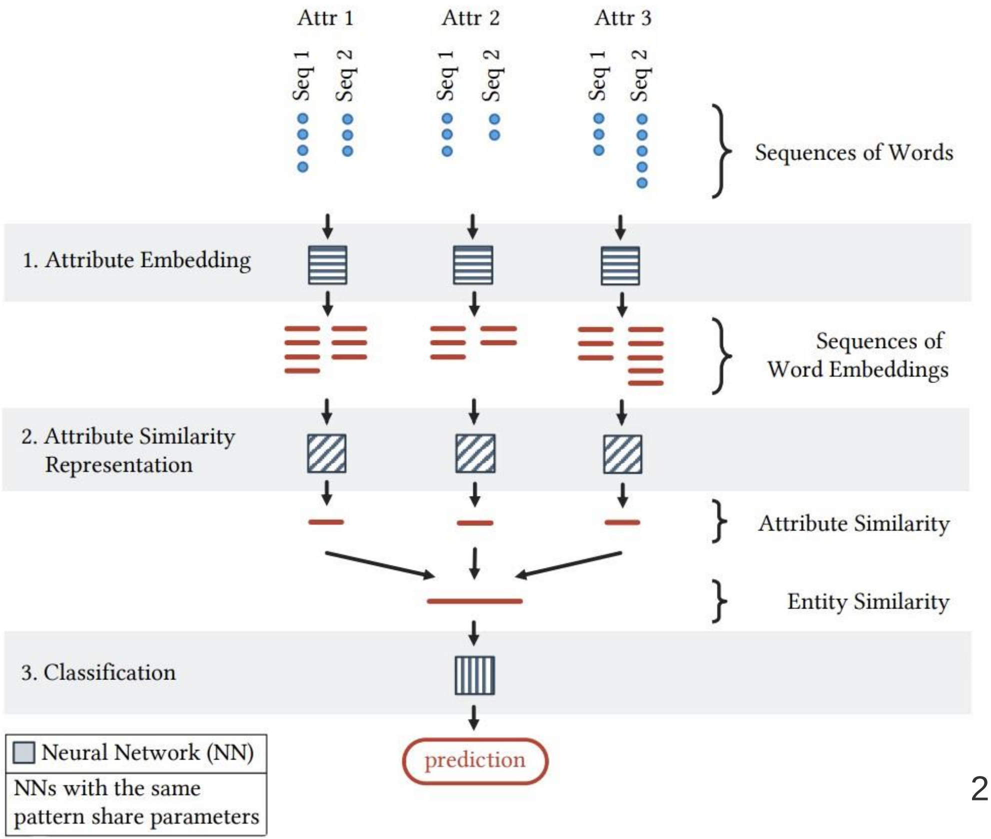
- **Why Deep Learning?**
  - **Feature Learning:** DL models can automatically learn relevant features from the data, reducing the need for handcrafted feature engineering
  - **Handling Noisy data:** DL models can handle variations in data, including misspellings, synonyms, abbreviations, and noisy data
  - **End-to-End Learning:** DL models can learn end-to-end solutions, eliminating the need for multiple stages of pre-processing and post-processing
  - **Performance:** DL models have achieved SOTA performance in EM tasks

Deep Entity Matching

- **Neural networks are used to ...**
  - **Encode textual values:**
    - Text is transformed to a numerical format, a.k.a embeddings
    - Mainly from the NLP domain
  - **Automatically detect matching patterns:**
    - Initial embeddings are transformed through multiple stacked layers in order to learn task-specific features, a.k.a, similarity embeddings
  - **Classify record pairs as match or non-match:**
    - Similarity embeddings are used as input for a binary classification task

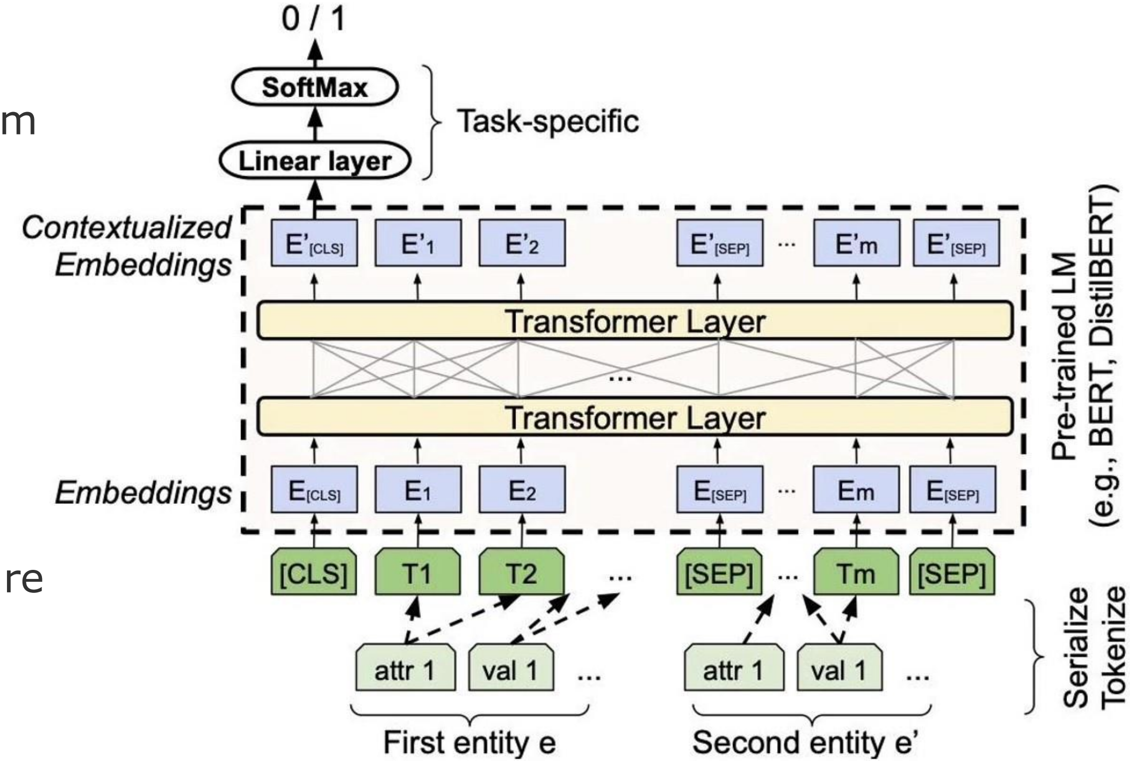
Deep Matcher

- Pioneer architecture template for DL solutions for EM
- Assumes the input records to be aligned in the schema
- Combines the values of an attribute in the two records into a single attribute embedding
- Generates record representations by combining attribute embeddings



Ditto

- Converts structured data into plain texts and integrates them with external knowledge
  - Attribute separators
  - Domain knowledge
  - Augmentation
- Encodes texts with PLMs
- Fine-tunes the PLM architecture on the EM task



Agenda

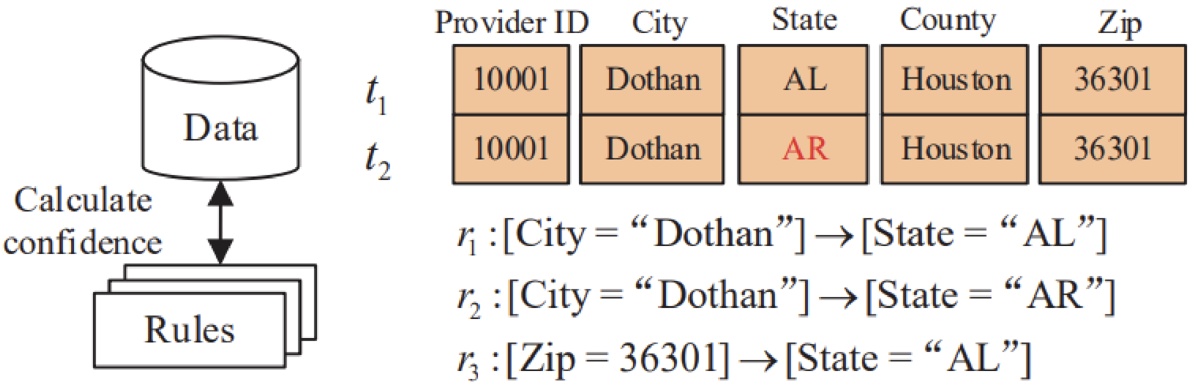
- ❑ Chair Introduction
- ❑ Organizational Information
- ❑ Data Cleaning and Integration
- ❑ Seminar Topics

Seminar Topics

1. Automatic Generation of Data Cleaning Rules with Deep Learning
2. Propagating Data Errors from Query Results to Data Sources
3. Schema Matching post-processing with Deep Learning
4. Schema Matching using Pretrained Language Models
5. Weakly supervised Entity Matching
6. Domain Adaptation for Deep Entity Resolution
7. Self-supervised training of EM models
8. Unsupervised Entity Matching
9. Entity Resolution for Complex Entities
10. Multi-purpose Data Integration Models

Topic 1: Automatic Generation of Data Cleaning Rules with Deep Learning

- **Goal:** Can we use deep learning models to automatically (e.g., with self-supervision) generate interpretable data cleaning rules?
- **Main Reference:** Jinfeng Peng, et al. “Self-supervised and Interpretable Data Cleaning with Sequence Generative Adversarial Networks”, VLDB 2022 - https://dl.acm.org/doi/abs/10.14778/3570690.3570694



	Provider ID	City	State	County	Zip
t_1	10001	Dothan	AL	Houston	36301
t_2	10001	Dothan	AR	Houston	36301

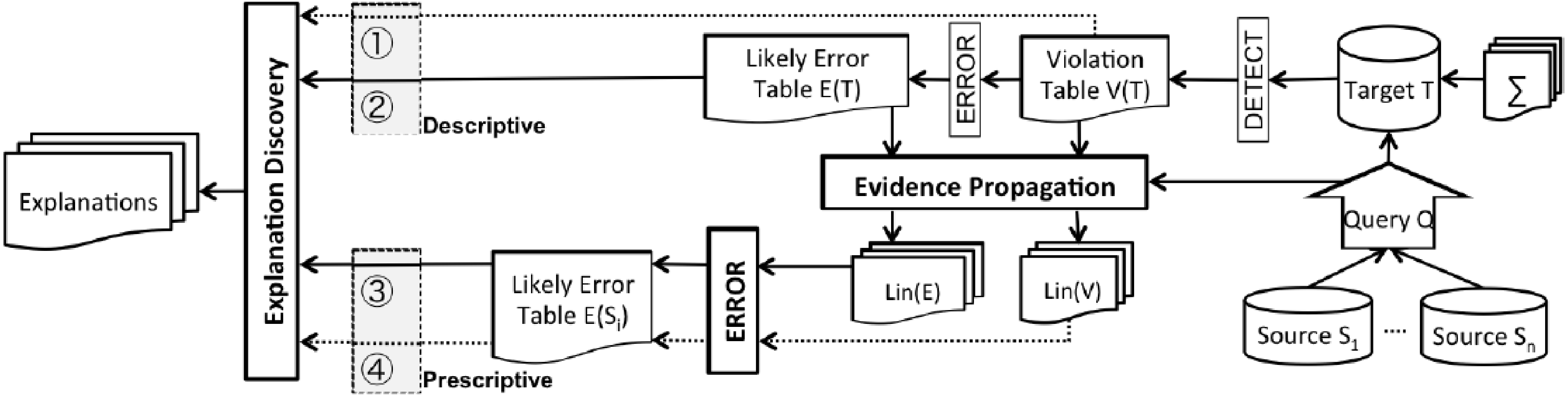
r_1 :[City = “Dothan”] → [State = “AL”]

r_2 :[City = “Dothan”] → [State = “AR”]

r_3 :[Zip = 36301] → [State = “AL”]

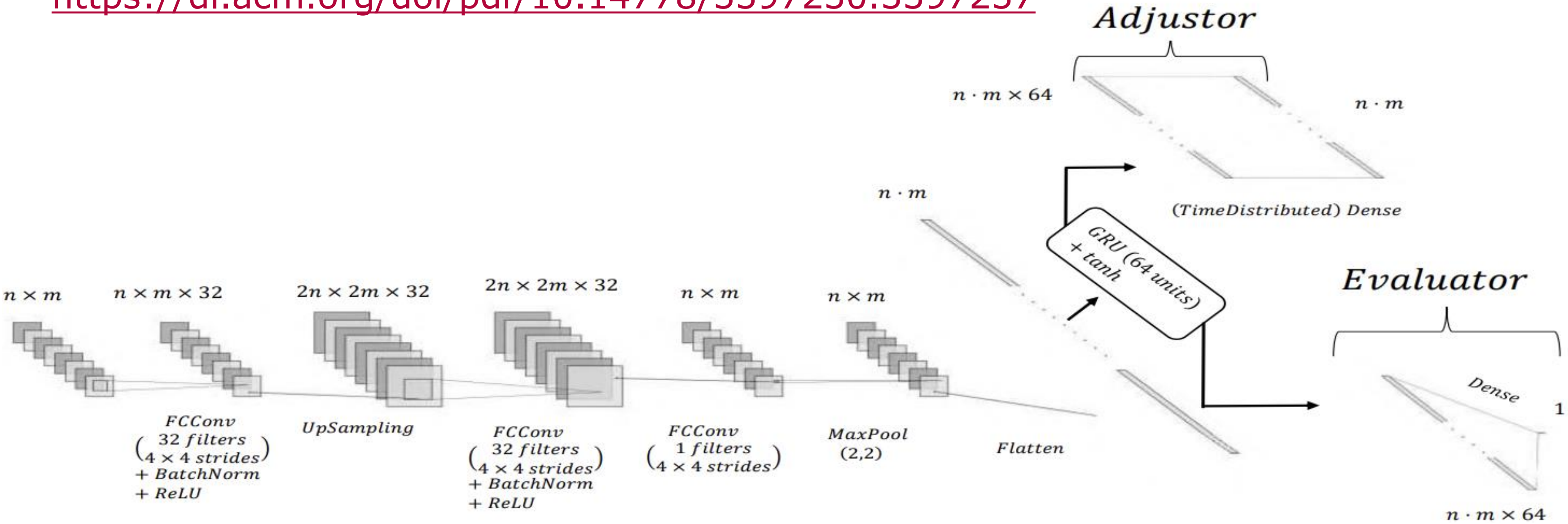
Topic 2: Propagating Data Errors from Query Results to Data Sources

- **Goal:** When a data error is found in a query result. How can we find the corresponding error in the queried data source?
- **Main Reference:** Anup Chalamalla, et al. "Descriptive and Prescriptive Data Cleaning", SIGMOD 2014 - https://cs.uwaterloo.ca/~ilyas/papers/AnupSIGMOD2014.pdf



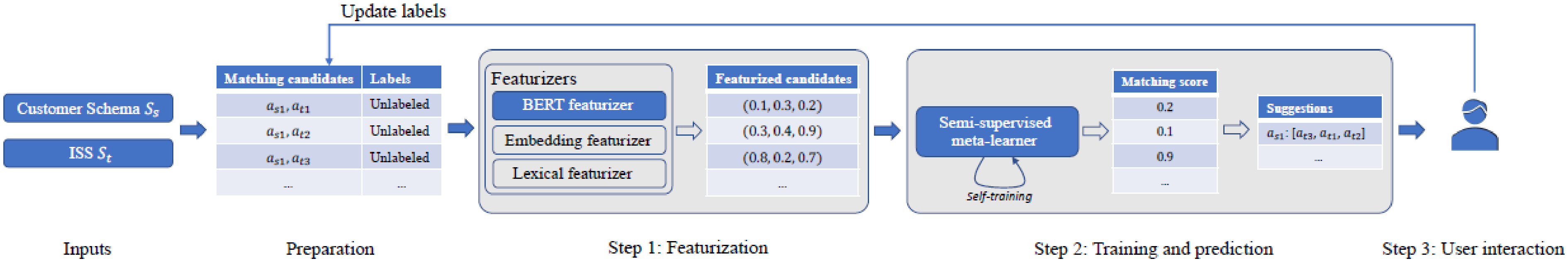
Topic 3: Schema Matching post-processing with Deep Learning

- **Goal:** Refine Schema Matching results with Deep Learning
- **Main Reference:** Roee Shraga, et al. "ADnEV: cross-domain schema matching using deep similarity matrix adjustment and evaluation", VLDB 2020 - https://dl.acm.org/doi/pdf/10.14778/3397230.3397237



Topic 4: Schema Matching using Pretrained Language Models

- **Goal:** How can we leverage LLMs to match relational schemas?
- **Main Reference:** Yunjia Zhang, et al. “Schema Matching using Pre-Trained Language Models”, ICDE 2023 - https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10184612



Update labels

Customer Schema S_s

ISS S_t

Matching candidates	Labels
a_{s1}, a_{t1}	Unlabeled
a_{s1}, a_{t2}	Unlabeled
a_{s1}, a_{t3}	Unlabeled
...	...

Featurizers

BERT featurizer

Embedding featurizer

Lexical featurizer

Featurized candidates
(0.1, 0.3, 0.2)
(0.3, 0.4, 0.9)
(0.8, 0.2, 0.7)
...

Semi-supervised meta-learner

Self-training

Matching score
0.2
0.1
0.9
...

Suggestions
a_{s1}: $[a_{t3}, a_{t1}, a_{t2}]$
...

Inputs

Preparation

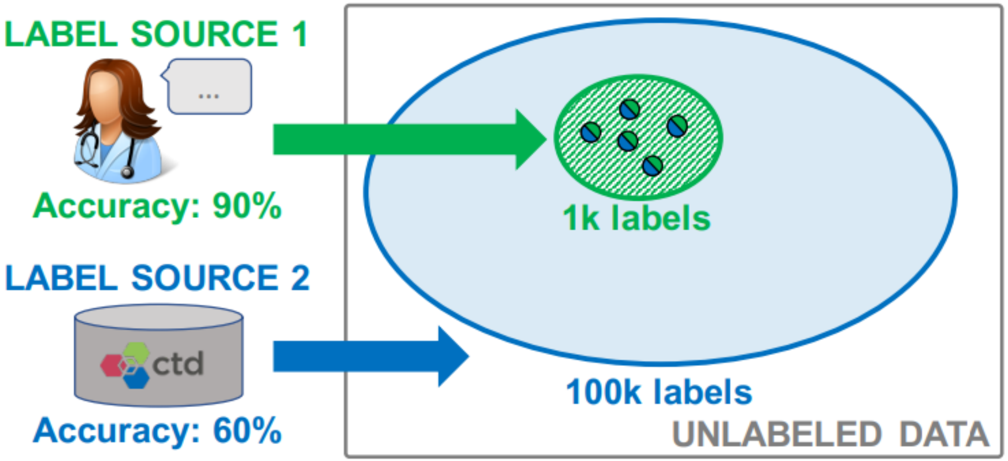
Step 1: Featurization

Step 2: Training and prediction

Step 3: User interaction

Topic 5: Weakly supervised Entity Matching

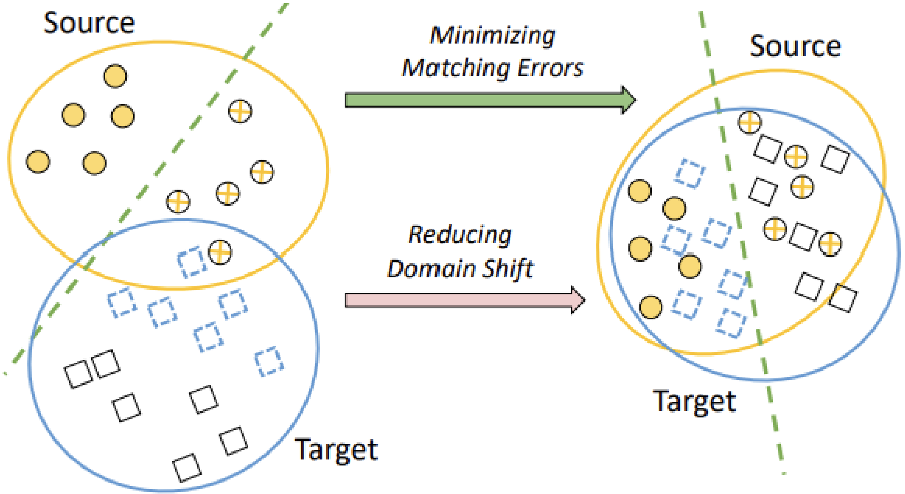
- ❑ **Goal:** Study some EM techniques that exploit weak forms of supervision (e.g., user-defined matching functions)
- ❑ **Main Reference:** Renzhi Wu, et al. "Ground Truth Inference for Weakly Supervised Entity Matching", SIGMOD 2023 - https://dl.acm.org/doi/10.1145/3588712



A. Ratner, S. Bach, et al., Snorkel: Rapid Training Data Creation with Weak Supervision, VLDB 2017

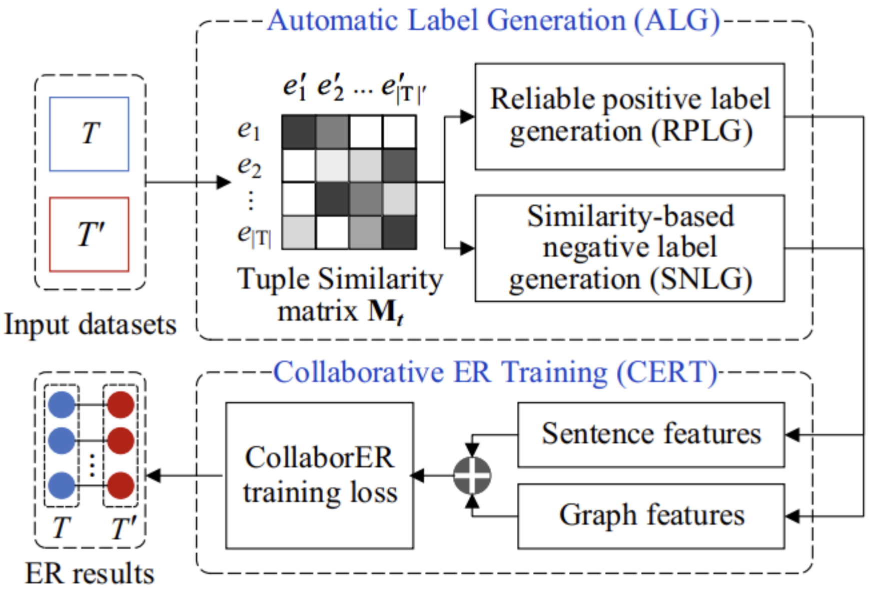
Topic 6: Domain Adaptation for Deep Entity Resolution

- **Goal:** If we have a well-labeled source ER dataset, can we train a DL-based ER model for a target dataset, without any labels or with a few labels?
- **Main Reference:** Jianhong Tu, et al. "Domain Adaptation for Deep Entity Resolution", SIGMOD 2022 – https://dl.acm.org/doi/10.1145/3514221.3517870



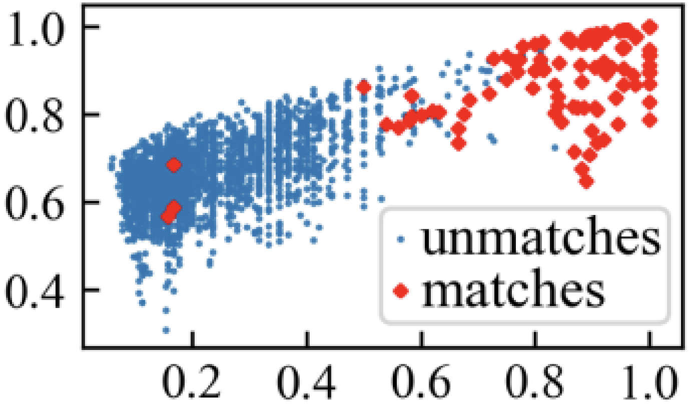
Topic 7: Self-supervised Entity Matching

- **Goal:** Study some techniques for training an EM model in a self-supervised way
- **Main Reference:** Congcong Ge, et al. "CollaborER: A Self-supervised Entity Resolution Framework Using Multi-features Collaboration", SIGMOD 2022 - https://arxiv.org/pdf/2108.08090.pdf


Automatic Label Generation (ALG)
T
T′
Input datasets
$e'_1\ e'_2 \dots e'_{|T'|}$
e_1
e_2
$e_{|T|}$
Tuple Similarity matrix $\mathbf{M}_t$
Reliable positive label generation (RPLG)
Similarity-based negative label generation (SNLG)
Collaborative ER Training (CERT)
Sentence features
Graph features
CollaborER training loss
T
T′
ER results

Topic 8: Unsupervised Entity Matching

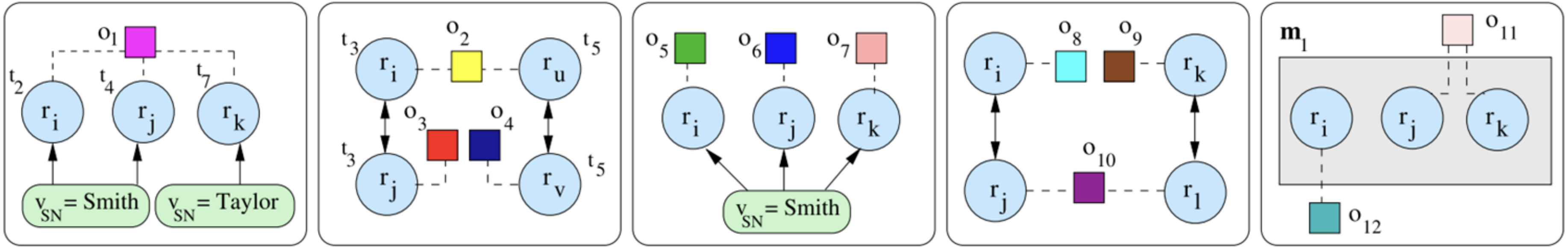
- **Goal:** Is it possible to design an effective algorithm for EM that requires zero labeled examples, yet can achieve performance comparable to supervised approaches?
- **Main Reference:** Renzhi Wu, et al. "ZeroER: Entity Resolution using Zero Labeled Examples", SIGMOD 2020 – https://dl.acm.org/doi/10.1145/3318464.3389743



Feature similarity distributions

Topic 9: Entity Resolution for Complex Entities

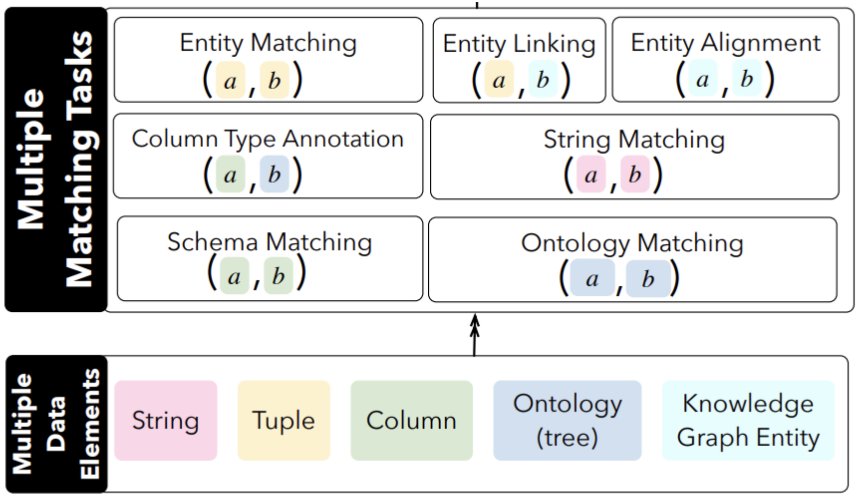
- **Goal:** Study some techniques for deduplicating complex entities that evolve over time (i.e., entity values and relationships change over time)
- **Main Reference:** Nishadi Kirielle, et al. "Unsupervised Graph-Based Entity Resolution for Complex Entities", TKDD 2023 - https://dl.acm.org/doi/10.1145/3533016



(a) Changing attribute values (b) Different relationships (c) Disambiguation problem (d) Partial match groups (e) Incorrect link problem

Topic 10: Multi-purpose Data Integration Models

- **Goal:** Study how to develop unified architectures for addressing multiple data integration tasks
- **Main Reference:** Jianhong Tu, et al. "Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration", SIGMOD 2023 - https://dl.acm.org/doi/10.1145/3588938



Further Procedure

- To apply for this seminar (binding):
  - **Email** to fabian.panse@hpi.de with **one topic choice**
  - **Deadline**: Friday 20.10.2023 11:59
  - **Notification**: Friday 20.10.2023 18:00
  - Register with the Studienreferat
- In case of too many applications, we need to choose **randomly**.

Seminar Webpage