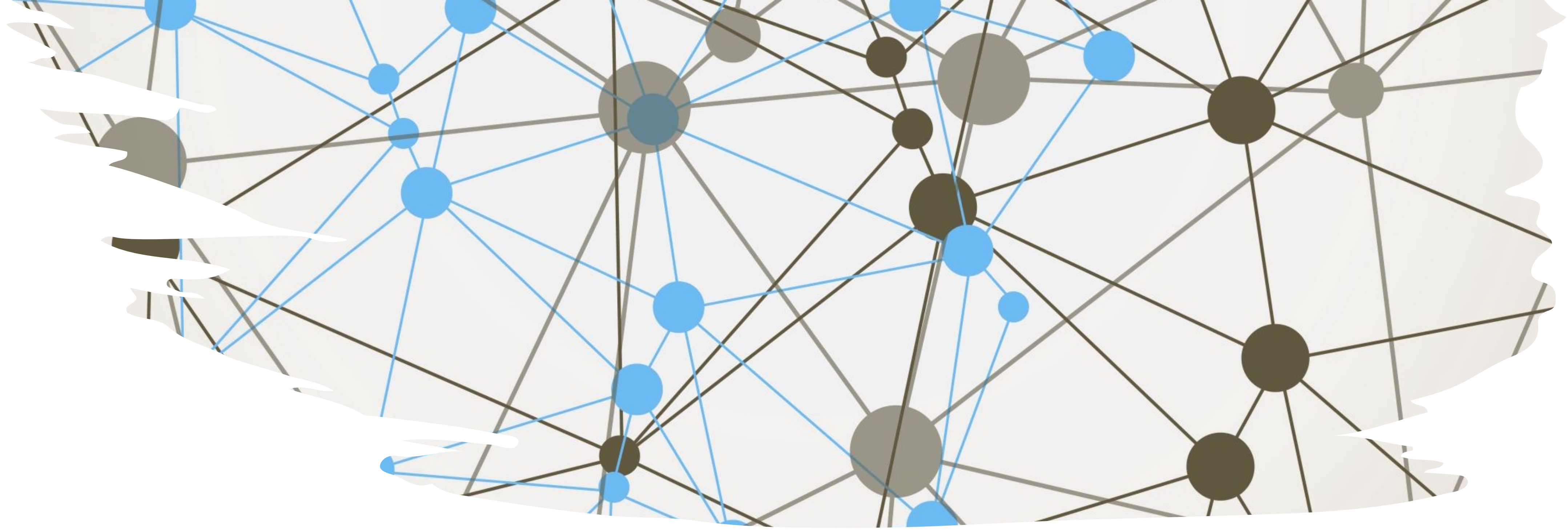


Evaluating Matching Techniques with Valentine*

CHRISTOS KOUTRAS

TU Delft

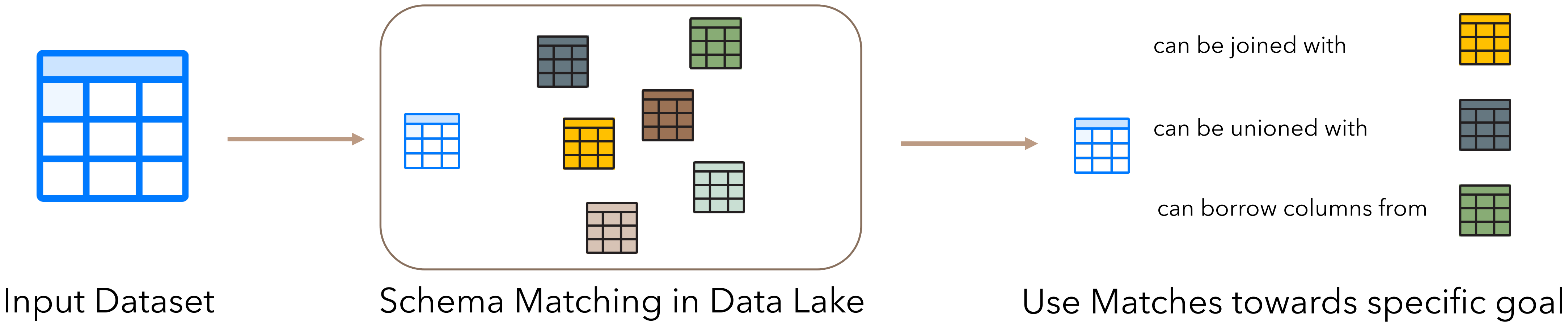
*Work presented in IEEE ICDE 2021

From Data Integration to Schema Matching

- Organizations gather heterogeneous data into *data lakes*
  - Data scientists spend most of their time on capturing relevance
  - Data Integration Problem: Relevant data sources are unlinked

- Need for *Schema Matching*
  - (Semi-)Automated methods with the goal of finding links among datasets
  - E.g. Related columns among Tabular Data

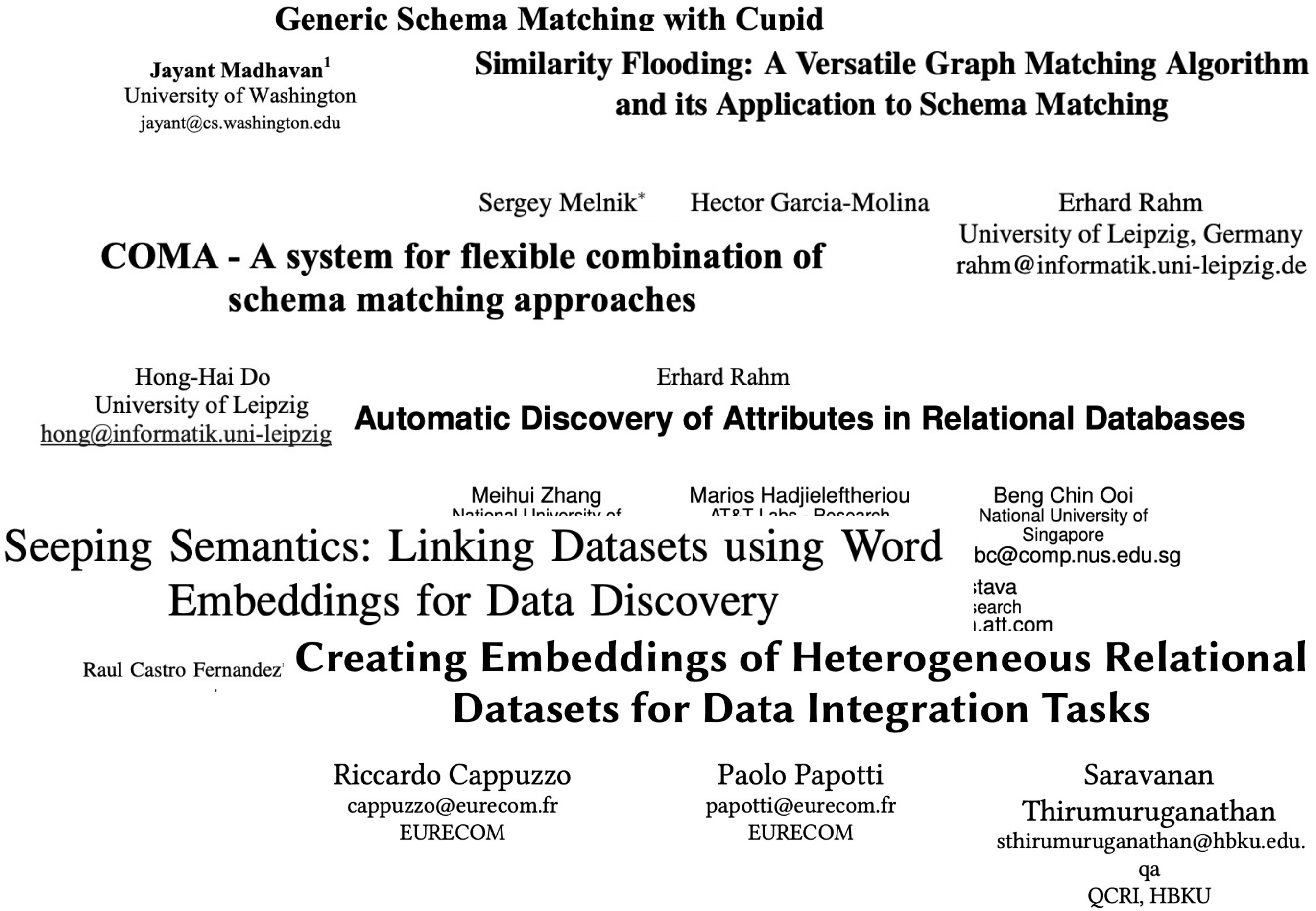
From Schema Matching to Dataset Discovery



- *Dataset Discovery* is a critical task for organizing a data lake
  - Navigate numerous data sources to find relationships for a given dataset
  - Schema Matching is a **core** component of any modern dataset discovery pipeline

Schema Matching in research

- Abundance of matching methods
- No comparison in 20 years
- No evaluation datasets
- Outdated metrics



A missing link with Dataset Discovery literature

- Result: none of the dataset discovery methods (~last 10 years) employ them!

- Dataset Discovery methods typically implement their own matching methods

Finding Related Tables

Anish Das Sarma, Lujun Fang, Nitin Gupta, Alon Halevy,
Hongrae Lee, Fei Wu, Reynold Xin, Cong Yu

InfoGather: Entity Augmentation and Attribute Discovery By Holistic Matching with Web Tables

Mohamed Yakout*
Purdue University
myakout@cs.purdue.edu

Kris Ganjam
Microsoft Research
krisgan@microsoft.com

Stitching Web Tables for Improving Matching Quality

Oliver Lehmberg, Christian Bizer
Data and Web Science Group, Universität Mannheim

Aurum: A Data Discovery System

…n, Sam Madden, Michael Stonebraker

Table Union Search on Open Data

Fatemeh Nargesian
University of Toronto
fnargesian@cs.toronto.edu

Erkang Zhu
University of Toronto
ekzhu@cs.toronto.edu

Dataset Discovery in Data Lakes

…orman W. Paton, Nikolaos Konstantinou
…ity of Manchester, Manchester, UK
…anchester.ac.uk

Finding Related Tables in Data Lakes for Interactive Data Science

Yi Zhang
yizhang5@cis.upenn.edu
University of Pennsylvania
Philadelphia, PA

Zachary G. Ives
zives@cis.upenn.edu
University of Pennsylvania
Philadelphia, PA

Valentine to the rescue

Current limitations

- No comparative study
- No specific relatedness scenarios
- No evaluation datasets
- No open-sourced methods
- Tough/impossible deployment for Dataset Discovery

Our contributions

- Most comprehensive effectiveness/efficiency study to date
- Relatedness scenarios accustomed to Dataset Discovery
- Dataset fabrication
- 6 SotA methods + a baseline
- Easily deployed and extensible

Matching in Dataset Discovery

- Six categories of matchers are used:

Attribute Overlaps | Value Overlaps | Semantic Overlaps

Data Types | Distributions | Embeddings

- Valentine brings the best of schema matching
  - Covers all matcher categories
  - Sophisticated methods that employ several intuitions and techniques

A new way of evaluating schema matching

Table 1

Client	Street	PO	Country
…	…	…	…

Ground Truth

Table 2

C_ID	Addr	P_Code	Cntr
…	…	…	…

- Ranked Matches serve better the needs of Dataset Discovery
- Recall @ ground truth shows the quality of the ranking a method returns rather than its ability to filter out irrelevant matches

Conventional Schema Matching Evaluation

Match Results

Client <-> C_ID
Country <-> Cntr

Evaluation

Precision: 0.5
Recall: 1/3

Valentine's Novel Evaluation

Ranked Match Results

Client - C_ID: 0.85
Country – Cntr: 0.67
PO – P_Code: 0.35
…

Evaluation

Recall @ ground truth: 1

Dataset Relatedness Scenarios

- Evaluate on dataset pairs that respect specific relatedness semantics

Joinable or Semantically Joinable

Client	Street	Country
J. Watts	2, Tea St.	USA
B. Mei	8, Fly St.	China
...	...	...

⋈

Cntr	C_Office	Head
USA	68346	B. Stan
China	74742	J. Ki
...	...	...

Client	Street	Country
J. Watts	2, Tea St.	USA
B. Mei	8, Fly St.	China
...	...	...

SEM ⋈

Cntr	C_Office	Head
States	68346	B. Stan
Chn	74742	J. Ki
...	...	...

Unionable or View-Unionable

Client	Street	PO
J. Watts	2, Tea St.	39499
B. Mei	8, Fly St.	34682
...	...	...

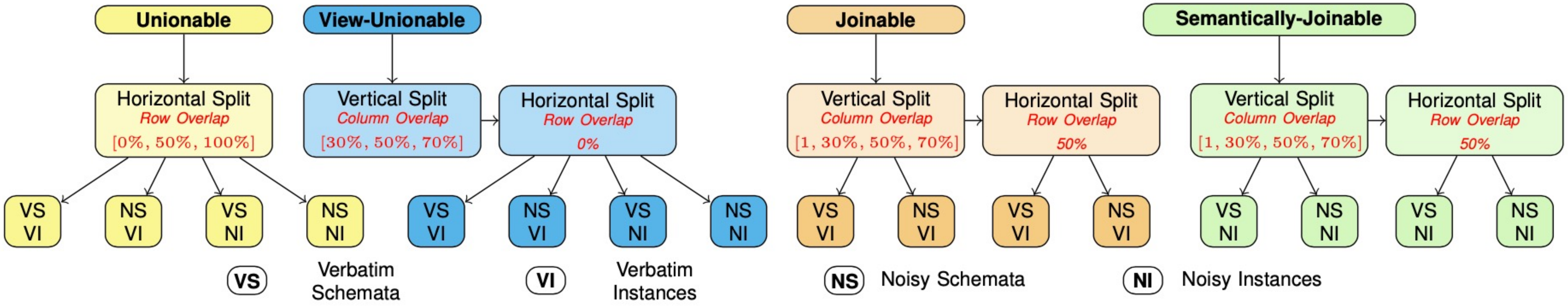
C_Name	Addr	P_Cod
B. Mei	8, Fly St.	34682
Q. Man	3, Bay St.	35472
...	...	...

Client	Street	PO
J. Watts	2, Tea St.	39499
B. Mei	8, Fly St.	34682
...	...	...

Addr	P_Code	C_ID
8, Fly St.	34682	C10012
3, Bay St.	35472	C23672
...	...	...

Dedicated Fabrication Process

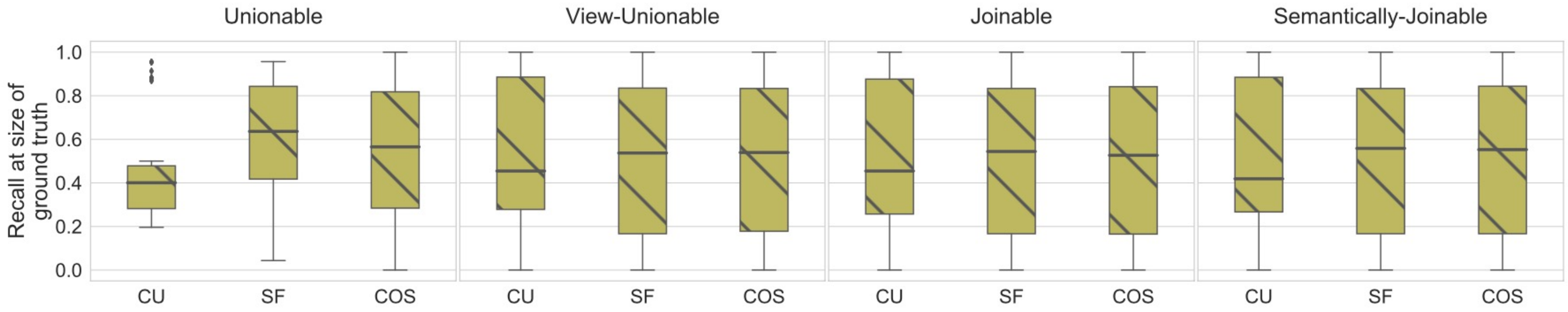
- Fabricate dataset pairs that follow the relatedness scenarios
- Based on a source table create pairs by employing
  - Horizontal/Vertical splits
  - Noise injection in Schemata/Instances



Valentine’s Schema Matching Methods

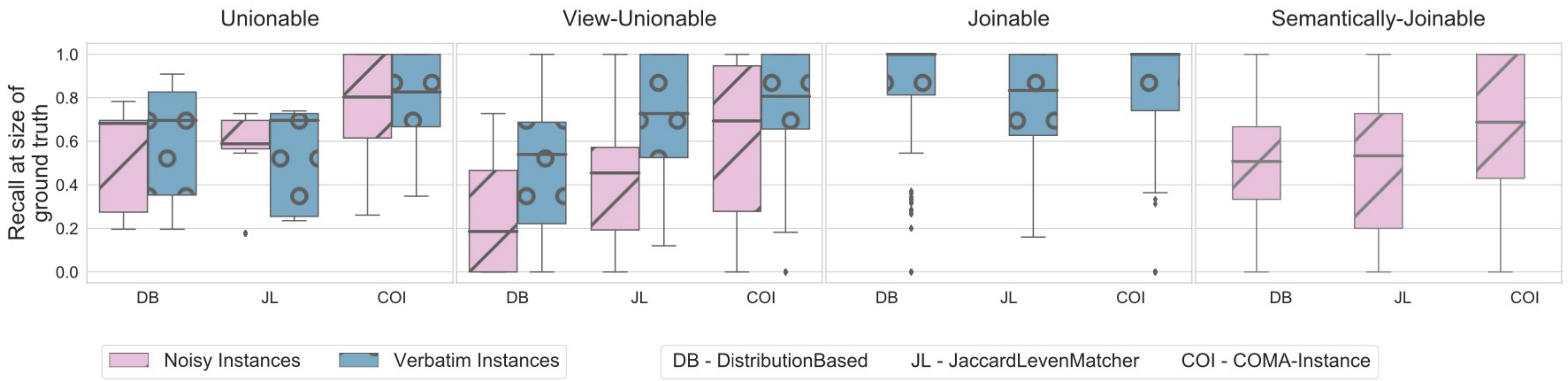
- Consolidates the best of schema matching efforts (last 20 years)
- Schema-based
  - Cupid – Similarity Flooding – COMA
- Instance-based
  - Distribution-based – COMA instance
  - Baseline based on approximate instance set overlaps
- Hybrid
  - SemProp - EmbDI

Findings – Schema Based Methods



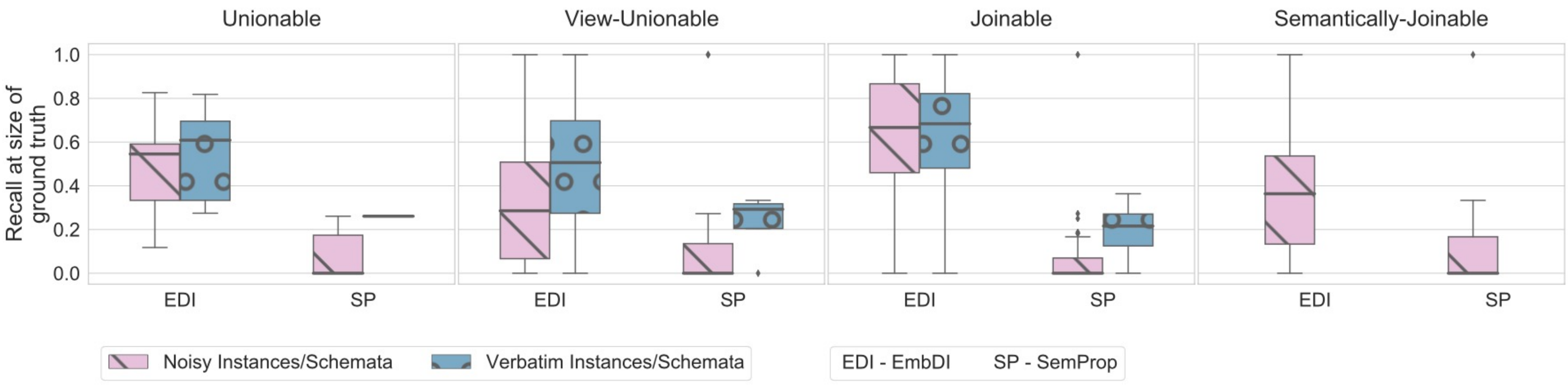
- Noisy schemata critically affect effectiveness and consistency
- Schema information (e.g. data types) and contextual information not insightful

Findings – Instance Based Methods



- View unionable and semantically-joinable scenarios are considerably harder
- High skew in effectiveness for all methods/scenarios

Findings – Hybrid Methods



- Low effectiveness and high skew
- Embeddings – pretrained or local ones – are still not a trustworthy standalone tool for matching

Findings on ING Data

Methods	ING #1	ING #2
Cupid	0.71	0.5
Sim Flooding	0.36	0.44
COMA Schema	0.79	0.12
COMA Instance	0.79	0.14
Dist based	**0.86**	**0.88**
Baseline	0.79	0.62
EmbDI	0.71	0.23

- Distributions of values can bring helpful insights
- Schema-based methods have very low effectiveness

Efficiency Results

Methods	Avg Runtime (sec)
Cupid	9.64
Sim Flooding	7.09
COMA Schema	1.67
COMA Instance	318.07
Dist based	71.16
SemProp	735.25
Baseline	522.94
EmbDI	4817.87

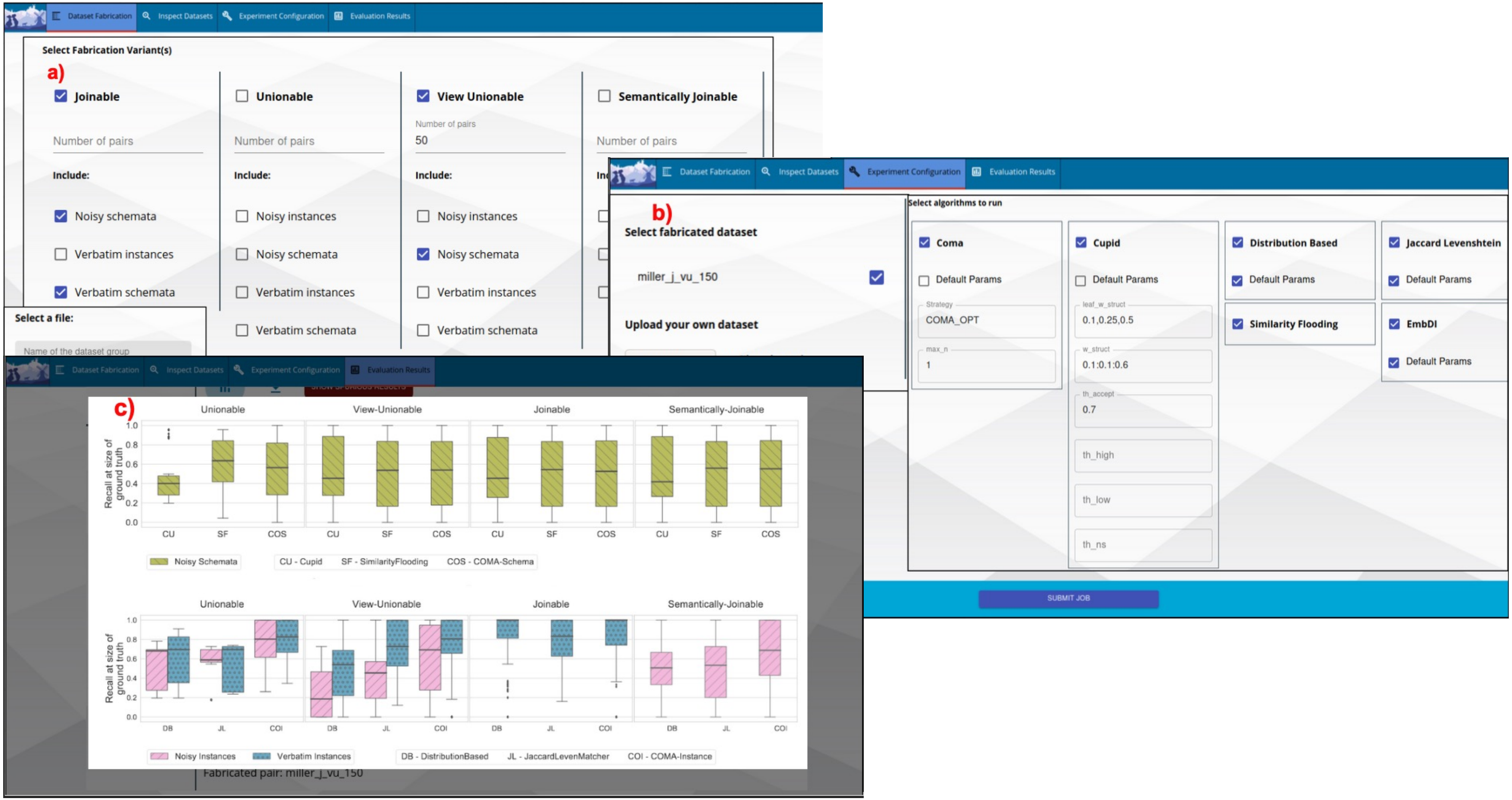
- Schema-based considerably faster
- Training embeddings can be very inefficient

Lessons Learned

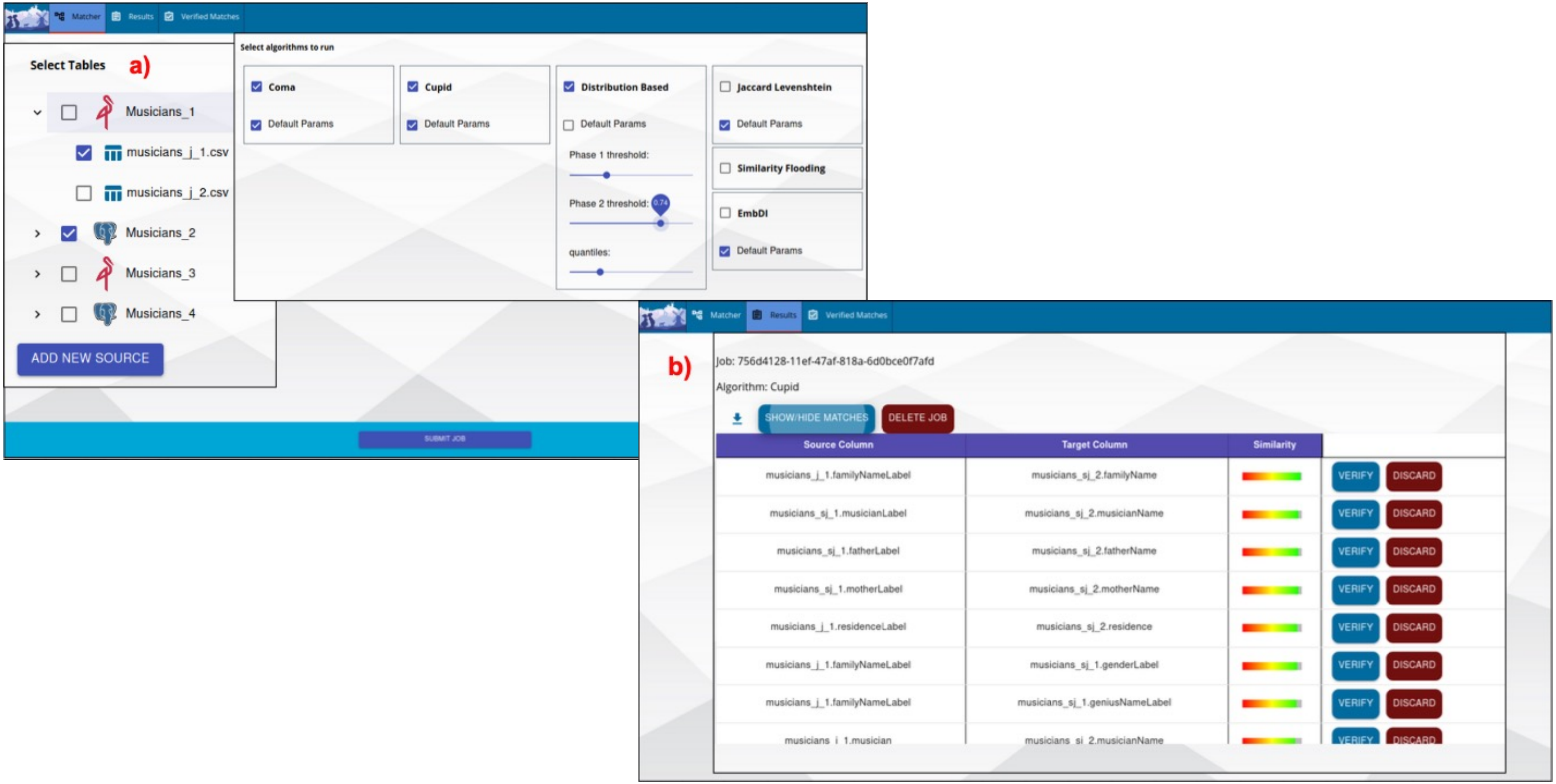
- There is no matching method that is consistently the best
- Embeddings should only be used together with other techniques
- Parameterization can be a daunting task – availability of ground truth can help
- Baselines can perform well
- Humans should be incorporated
- Schema Matching doesn't scale – expensive to deploy

Valentine in Action [Demo in VLDB 2021]

- Schema Matching systems come with an outdated GUI
- No deployment at scale – deployment in a data lake
- No easy-to-use and complete evaluation system available
- Offer Valentine's utilities through a user-friendly GUI
  - Offer a scenario-driven dataset fabricator
  - Enable users to conduct extensive experiments
- Enable users to deploy Valentine for holistic matching in a data lake



Dataset Fabrication
Inspect Datasets
Experiment Configuration
Evaluation Results
Select Fabrication Variant(s)
a)
Joinable
Unionable
View Unionable
Semantically Joinable
Number of pairs
50
Include:
Noisy schemata
Verbatim instances
Verbatim schemata
Noisy instances
Select a file:
Name of the dataset group
b)
Select fabricated dataset
miller_j_vu_150
Upload your own dataset
Select algorithms to run
Coma
Default Params
Strategy
COMA_OPT
max_n
1
Cupid
leaf_w_struct
0.1,0.25,0.5
w_struct
0.1:0.1:0.6
th_accept
0.7
th_high
th_low
th_ns
Distribution Based
Similarity Flooding
Jaccard Levenshtein
EmbDI
SUBMIT JOB
c)
Unionable
View-Unionable
Joinable
Semantically-Joinable
Recall at size of ground truth
CU
SF
COS
Noisy Schemata
CU - Cupid
SF - SimilarityFlooding
COS - COMA-Schema
DB
JL
COI
Noisy Instances
Verbatim Instances
DB - DistributionBased
JL - JaccardLevenMatcher
COI - COMA-Instance
Fabricated pair: miller_j_vu_150


Matcher
Results
Verified Matches
Select Tables
a)
Musicians_1
musicians_j_1.csv
musicians_j_2.csv
Musicians_2
Musicians_3
Musicians_4
ADD NEW SOURCE
Select algorithms to run
Coma
Default Params
Cupid
Default Params
Distribution Based
Default Params
Phase 1 threshold:
Phase 2 threshold:
quantiles:
Jaccard Levenshtein
Default Params
Similarity Flooding
EmbDI
Default Params
b)
Job: 756d4128-11ef-47af-818a-6d0bce0f7afd
Algorithm: Cupid
SHOW/HIDE MATCHES
DELETE JOB
Source Column
Target Column
Similarity
musicians_j_1.familyNameLabel
musicians_sj_2.familyName
musicians_sj_1.musicianLabel
musicians_sj_2.musicianName
musicians_sj_1.fatherLabel
musicians_sj_2.fatherName
musicians_sj_1.motherLabel
musicians_sj_2.motherName
musicians_j_1.residenceLabel
musicians_sj_2.residence
musicians_j_1.familyNameLabel
musicians_sj_1.genderLabel
musicians_j_1.familyNameLabel
musicians_sj_1.geniusNameLabel
VERIFY
DISCARD

Lessons Learned: Matching in a Data Lake

- Deploying matching in a data lake is a daunting task
  - Resource expensive
  - All-pairs comparison is inefficient / SOTA methods difficult to scale
- Automated matching techniques are not always reliable
  - They work under specific assumptions about the data
  - Such assumptions may not apply in big data repositories
  - **Human refinement is necessary**

Prospects in Large-scale Matching

- Incorporate human knowledge in development of methods
  - Instead of using humans in refinement, use them in the beginning
  - There always exists partial knowledge of the underlying data

- Build robust models
  - Model human knowledge in order to leverage modern DL methods
  - Can generalize well

Visit https://delftdata.github.io/valentine/ !

- Links to our GitHub Repos
  - Code for deployment
  - Code for dataset fabrication
  - In detail experimental results
- All fabricated dataset pairs used in the paper
- Updates regarding Valentine