# Meta-Learning Framework for Schema Matching

Elizabeth Witten

CS 7290

# Background on Schema Matching

# Schema Matching

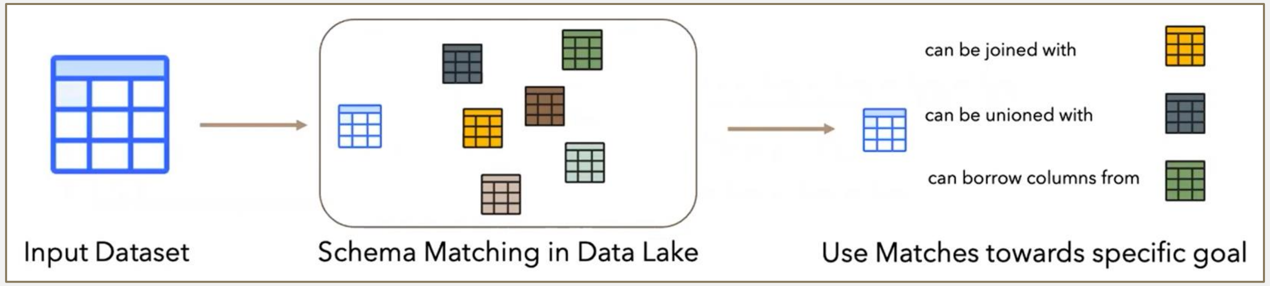
**Motivation**

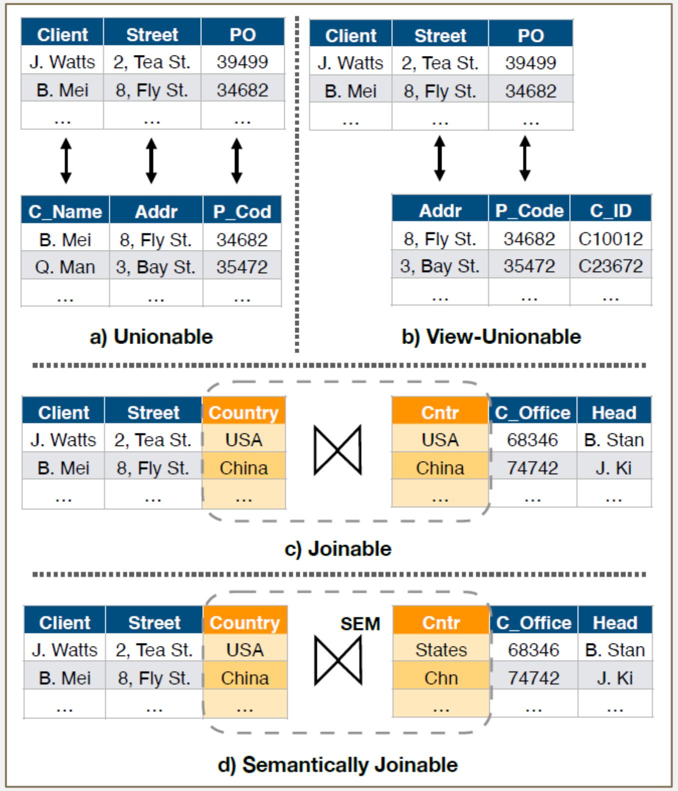
- **Goal**: Find relevant datasets among numerous data sources
- Data lakes store a lot of heterogeneous data
- Relevant data sources are unlinked
- Data integration and dataset discovery are essential for data science tasks
  - Searching for joinable tables
  - Augmenting given table with more data entries or attributes

# Schema Matching

**Problem Definition**

- Find matching pairs of columns between a source and target schema
- Capture relationships between elements of different schemas
- Schema- and Instance-based matchers



# Valentine

**Valentine: Evaluating Matching Techniques for Dataset Discovery**

- Experiment suite to execute automated matching experiments
- Motivated by an abundance of matching methods and lack of comparison
- Implements seminal schema matching methods
- Provides evaluation datasets

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

C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, et al., "Valentine: Evaluating matching techniques for dataset discovery," in IEEE ICDE, 2021.

# Overview of Schema-Based Matchers

| Cupid | Similarity Flooding | COMA |
| --- | --- | --- |
| • Schemas translated into tree structures<br>• Weighted score of name similarity and structural similarity | • Schemas translated into directed graphs<br>• Merged into a propagation graphs where similar nodes are collapsed into map pairs | • Schemas translated into directed graphs<br>• Incorporates multiple schema-based matchers<br>* Extension includes two instance-based matchers<br>• Supports human feedback |
| J. Madhavan, P. A. Bernstein, and E. Rahm, "Generic schema matching with cupid," in VLDB, 2001. | S. Melnik, H. Garcia-Molina, and E. Rahm, "Similarity flooding: A versatile graph matching algorithm and its application to schema matching," in IEEE ICDE, 2002. | H.-H. Do and E. Rahm, "COMA: a system for flexible combination of schema matching approaches," in VLDB, 2002.<br><br>D. Engmann and S. Massmann, "Instance matching with COMA++." in BTW workshops, vol. 7, 2007, pp. 28–37. |

# Overview of Hybrid Matchers

| EmbDI | SemProp |
|---|---|
| • Method for embedding values and attribute names of relations<br>• Uses external knowledge such as synonym dictionaries<br>• Finds relationships by comparing two column embeddings | • Uses an ontology and pre-trained word embeddings<br>• Two stages:<br>    • Semantic matcher: follows links from attribute and table names to ontology classes based on word embeddings<br>    • Syntactic matcher: looks at instance values |
| R. Cappuzzo, P. Papotti, and S. Thirumuruganathan, "Creating embeddings of heterogeneous relational datasets for data integration tasks," in SIGMOD, 2020. | R. C. Fernandez, E. Mansour et al., "Seeping semantics: Linking datasets using word embeddings for data discovery," in IEEE ICDE, 2018. |

# Overview of Instance-Based Matchers

| Distribution-Based | Jaccard-Levenshtein |
|---|---|
| • Clusters relational attributes using column value distribution similarity<br>• Outputs disjoint clusters whose relational attributes are considered related | • Naïve matcher that computes all pairwise column similarities<br>• Computes Jaccard similarity, and considers two values as identical if their Levenshtein distance is below a given threshold<br>• Outputs ranked list of column pairs |
| M. Zhang, M. Hadjieleftheriou, B. C. Ooi et al., “Automatic discovery of attributes in relational databases,” in ACM SIGMOD, 2011. | |

# Meta-Learning Implementation

# Meta-Learned Matcher

- Casts schema matching into a binary sequence classification task
- Takes the values of two columns as input
- Applies the InvDA operator to augment the training samples
- Uses the Rotom meta-learning framework to optimize data augmentation

# Data Pre-Processing

- Consider each column pair between different tables
  - Skip pairs containing numeric columns
  - Sample 15 random tokens from each column
  - Limit the number of negative samples

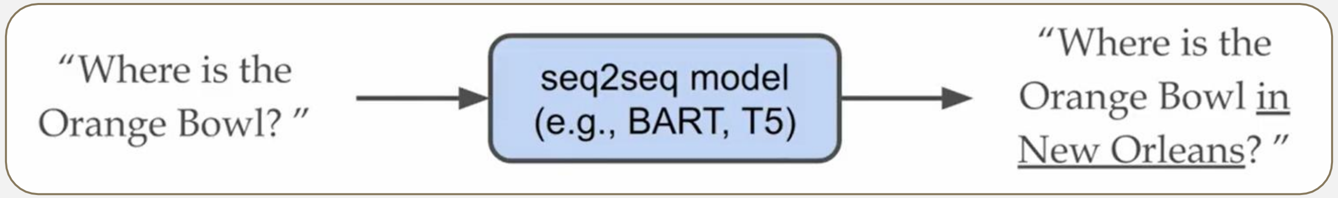
# Data Serialization

- Serialize each column to be concatenated
  [COL] $a_1$ $a_2$ … $a_n$ [SEP] [COL] $b_1$ $b_2$ … $b_m$
- Label with $1$ if the columns are considered a match, $0$ otherwise

```
[COL] Stockport MBC Stockport Stockport Stockport MBC
Stockport Stockport MBC Stockport MBC MBC Stockport MBC MBC [SEP]
[COL] Human Soft & Arts Fuel Soft Communication Equipment
Vehicle Care Supplies Furnishings Resources Furniture Community
```

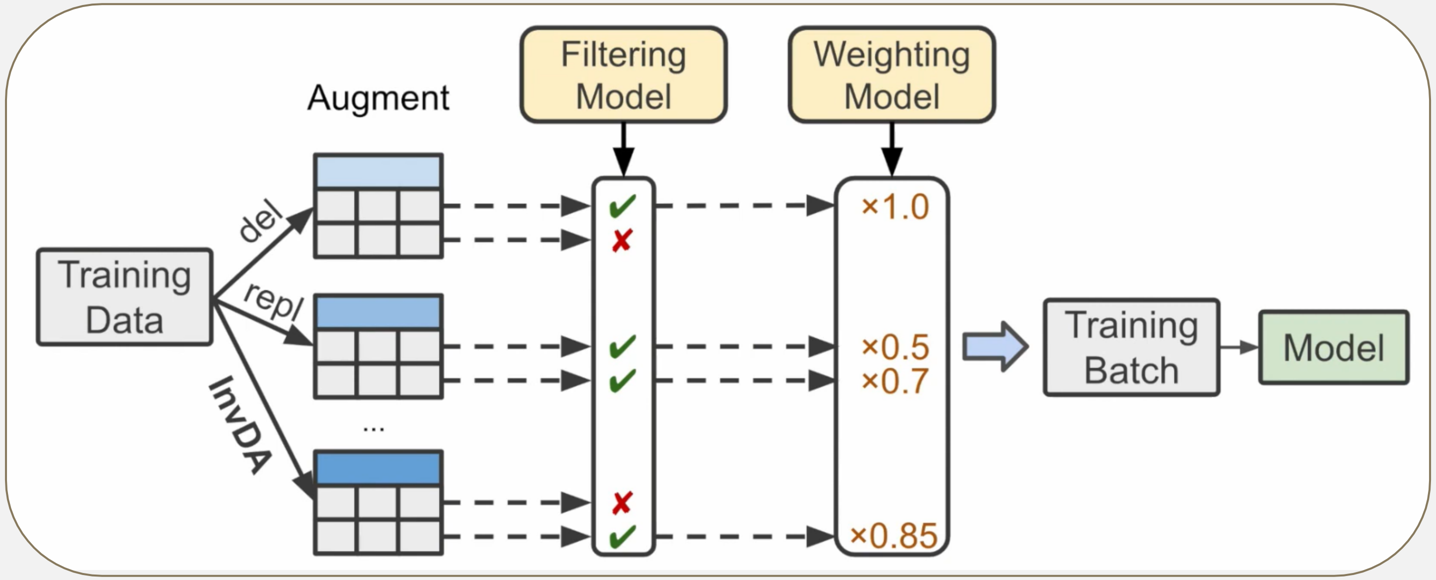
# Data Augmentation with InvDA

- Train the InvDA seq2seq model (T5) on the labeled training data
- Use the trained model to generate augmented training files




| |
|---|
| **[COL] Appropriations Higher Reform rights; and Management Beverage on Special Serving Industry for on Funds Sugar** |
| [COL] Appropriations Higher Reform rights; and Management Beverage Beer to Special Serving Industry for on Funds Sugar |
| [COL] Appropriations Higher Reform rights; Management On and Beverage Special Serving Industry for on Funds Sugar |

# Meta-Learning Framework




- Use the augmented training data in the meta-learning algorithm to train the target model

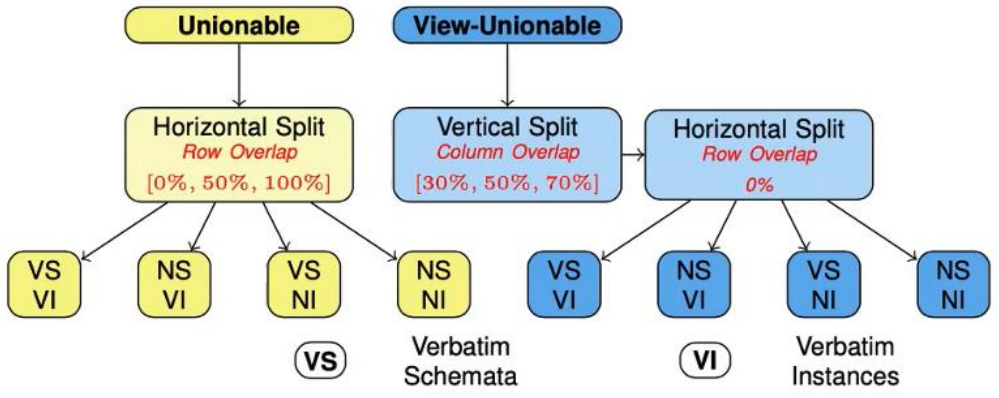
# Datasets

# SANTOS

- Sampled a subset of the data lake tables
  - 34 sets of unionable tables (e.g., albums, animal_tag_data)
- Ground truth determined by matching column names
- Pre-processed and serialized
- Split between Rotom training data and test data

# Valentine Datasets

- Valentine's Fabrication Module
  - Created by systematically splitting existing tables (horizontally or vertically)
  - Optionally add noise in schema information and instance values
  - Original table holds the ground truth




https://github.com/delftdata/valentine-data-fabricator

# Open Data Dataset Pair

- Fabricated dataset pair published with Valentine
- Parameters:
  - Horizontal split with 50% row overlap
  - Verbatim schema and instances (no noise added)
- Pre-processed and serialized
- Used as test data

# Experiments

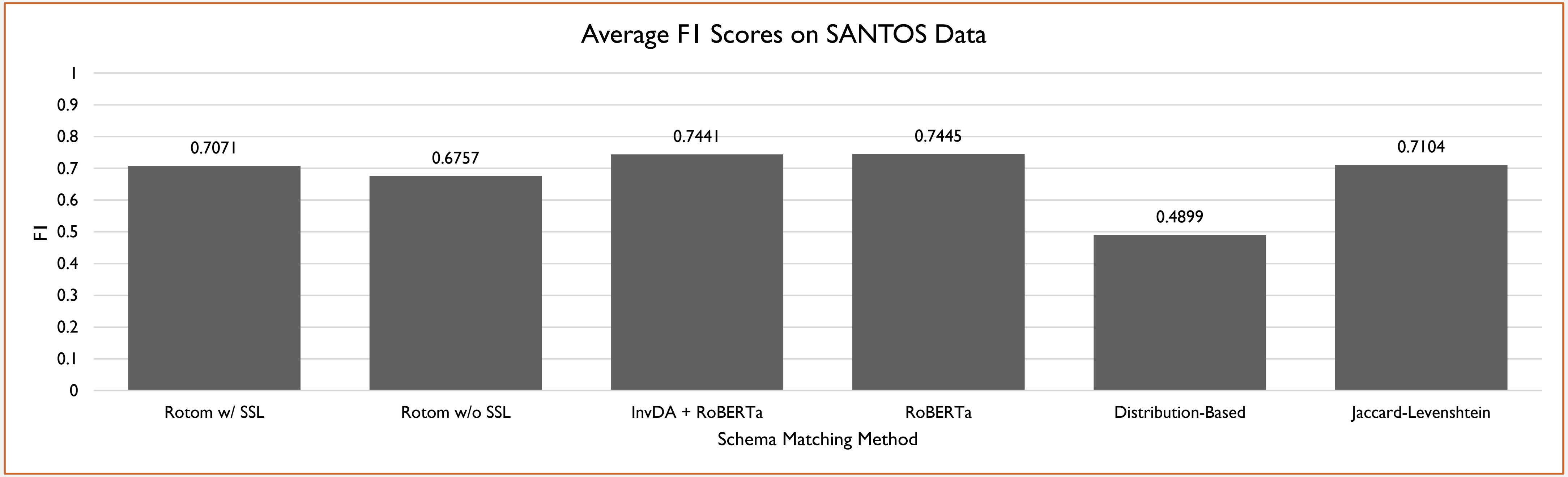
# Methodology

- Trained 4 Rotom-based models on the same training data
  - Rotom (full meta-learning framework) with semi-supervised learning
  - Rotom (full meta-learning framework) without semi-supervised learning
  - InvDA data augmentation and fined-tuned RoBERTa
  - Fine-tuned RoBERTa

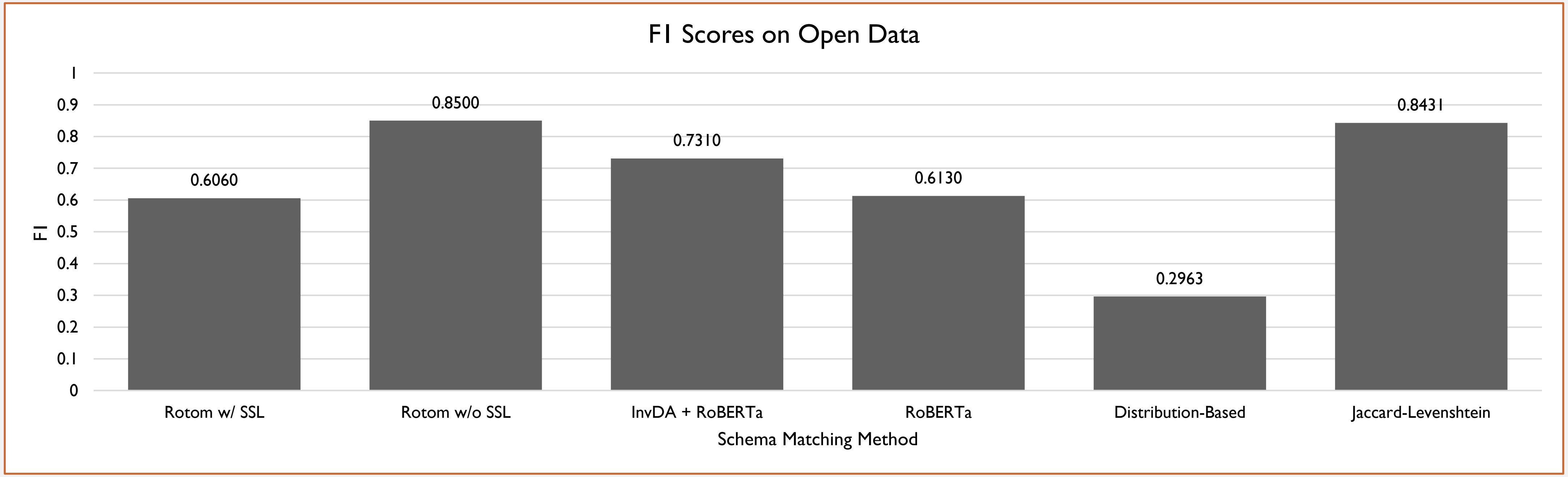
# Methodology

- Evaluated each model on the SANTOS data and the Open Data dataset from Valentine
  - 4 experimental Rotom-based models
  - Distribution-based matcher (implemented by Valentine)
  - Jaccard-Levenshtein matcher (implemented by Valentine)

# SANTOS Results



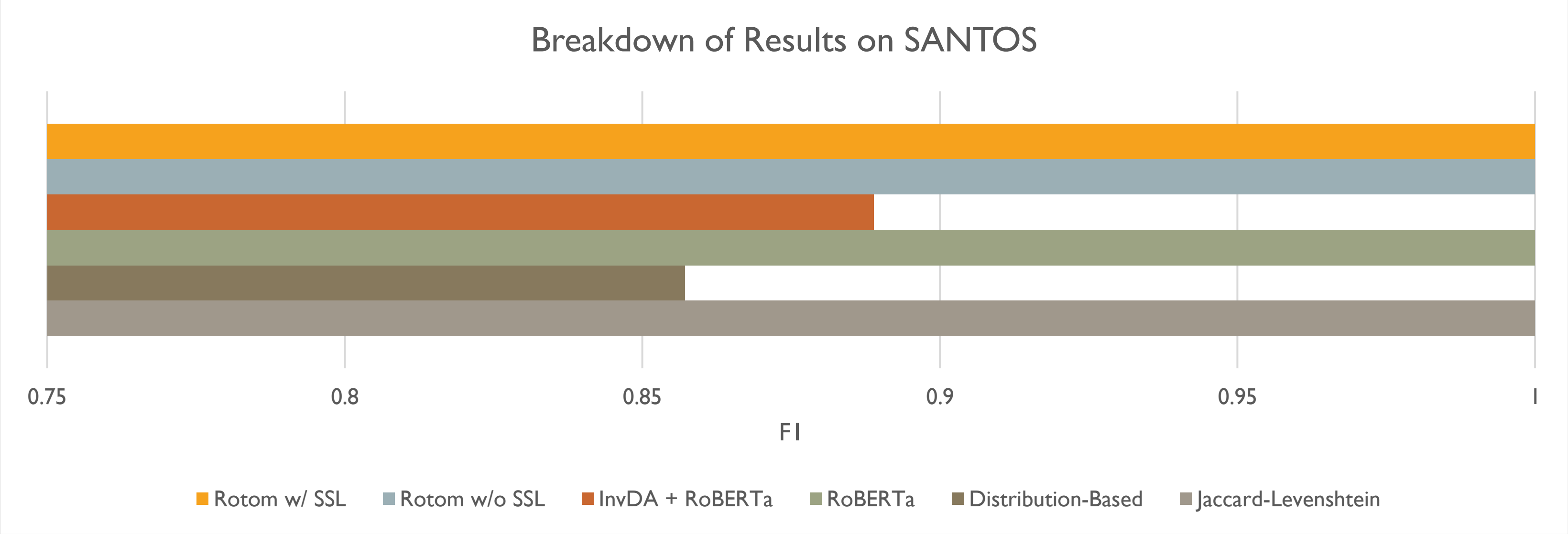
# Open Data Results



# Discussion of Results

# Results in Detail

Rotom + Valentine models mostly performed well on **lane_description**



# Results in Detail

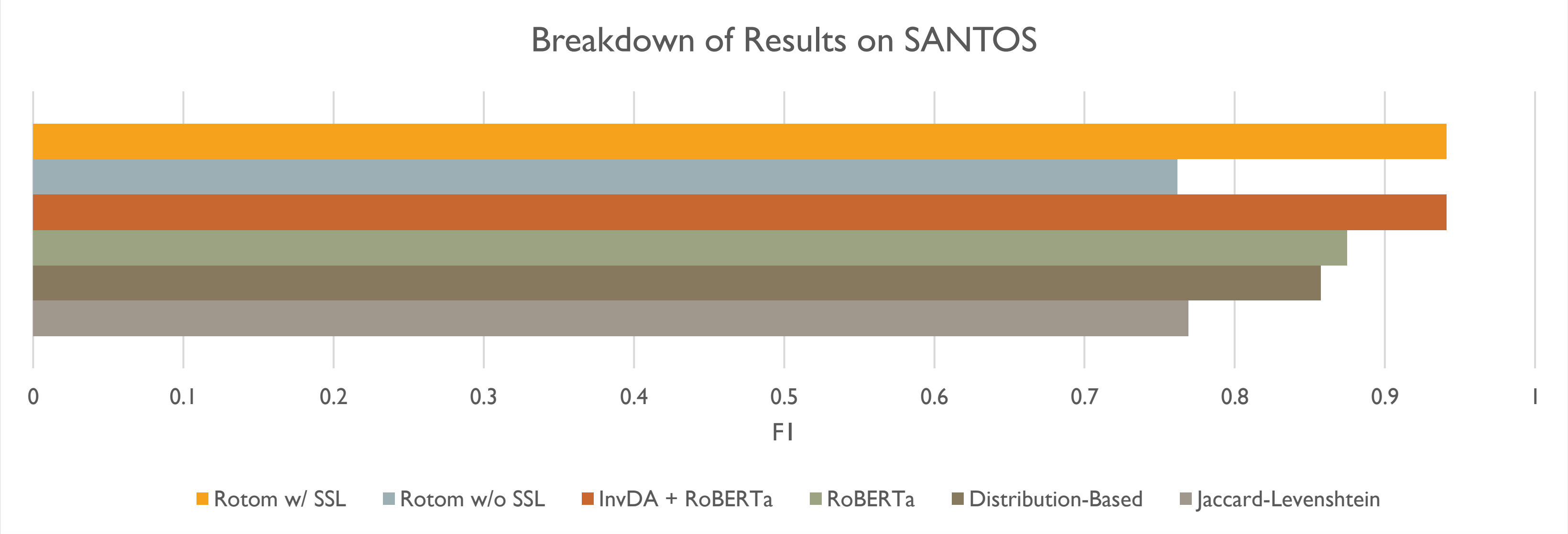
Rotom + Valentine models performed well on **lane_description**

| Sdate | LaneDescription | DirectionDescription | Flag Text |
|---|---|---|---|
| 19/07/2012 00:00 | Northbound | NorthEast | Bad data |
| 19/07/2012 00:00 | Southbound | SouthWest | Bad data |
| 19/07/2012 01:00 | Northbound | NorthEast | Bad data |

- The non-numeric columns have very little variety
- Data augmentation likely did not help due to the small set of values

# Results in Detail

Rotom + Valentine models performed well on **311_calls_historic_data**



# Results in Detail

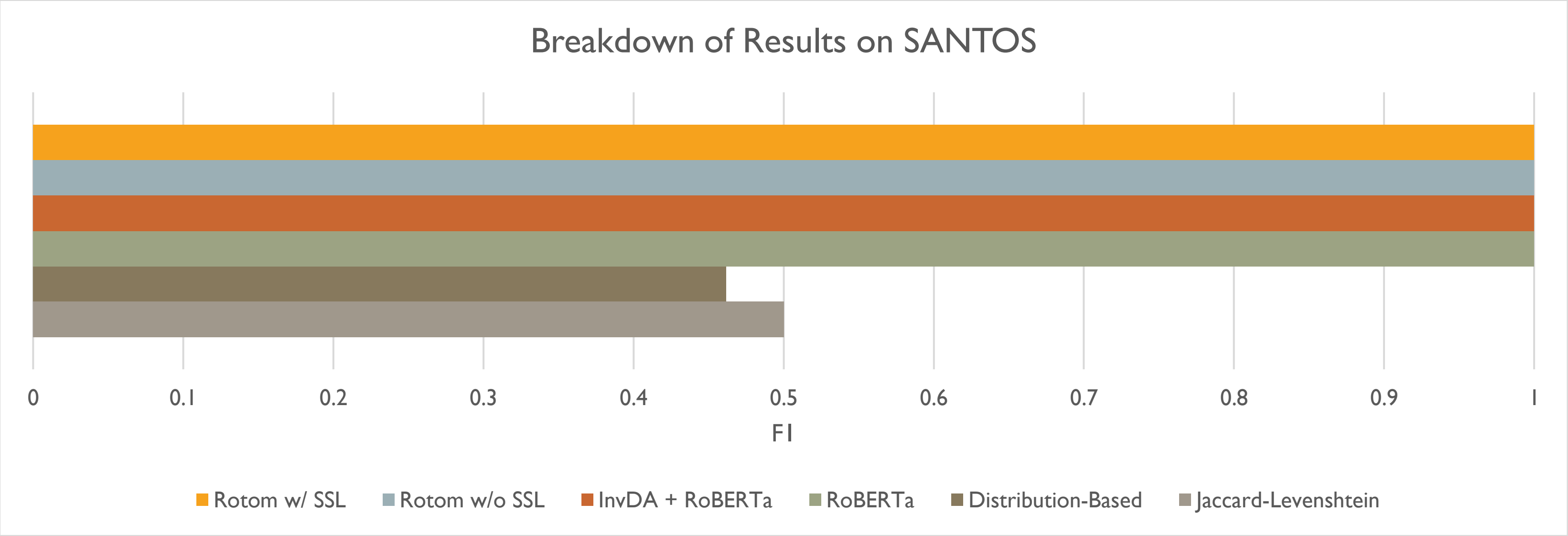
Rotom + Valentine models performed well on **311_calls_historic_data**

| issue_type | ticket_created_ date_time | ticket_closed_ date_time | ticket_ status | neighborhood_ district | city | state | case_title |
|---|---|---|---|---|---|---|---|
| Pothole/Roadway Surface Repair | 7/28/2015 8:23:00 AM | 7/29/2015 12:47:38 PM | Closed | MID-CITY | NEW ORLEANS | LA | Roadway Surface Repair - Pothole |
| Trash/Garbage Pickup | 7/23/2012 10:22:46 AM | 11/30/2012 5:15:08 PM | Closed | LAKEVIEW | NEW ORLEANS | LA | Start Trash Service |
| Trash/Garbage Pickup | 10/23/2017 11:01:04 AM | 10/25/2017 1:41:54 PM | Closed | LAKEWOOD | NEW ORLEANS | LA | Start Trash Service |

- The non-numeric columns also have little variety
- But more variety than **lane_description**, so InvDA was more effective

# Results in Detail

Rotom outperformed Valentine models on **animal_tag_data**



# Results in Detail

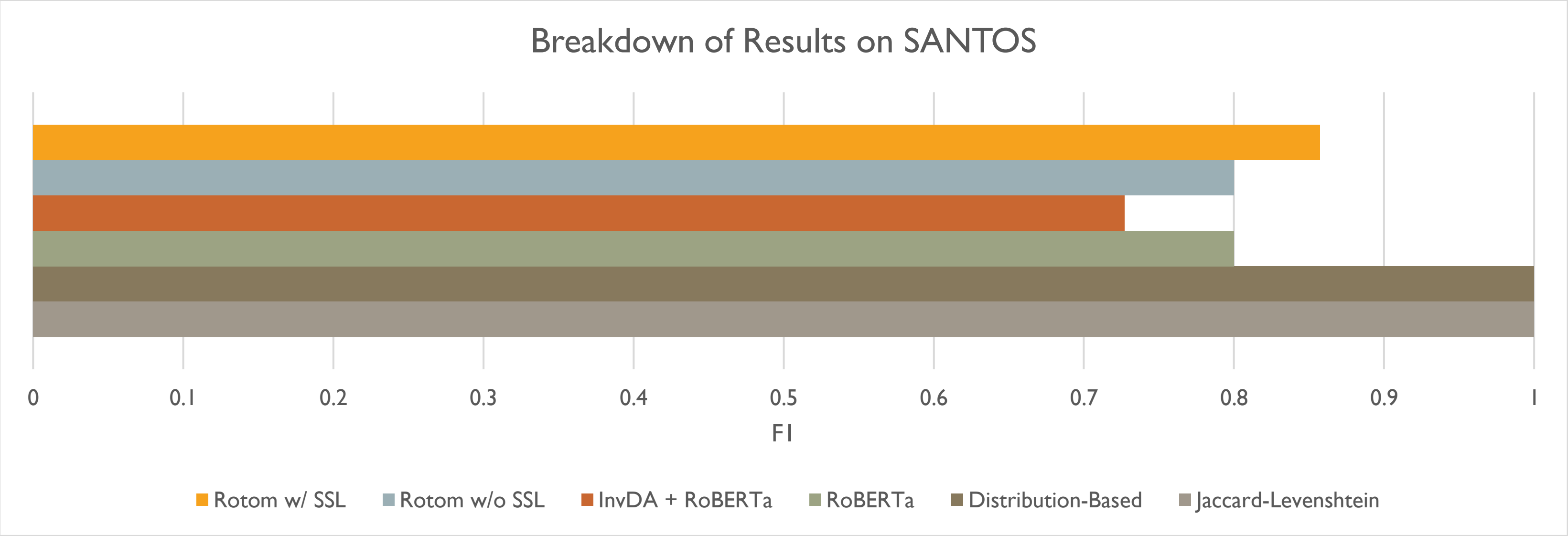
Rotom outperformed Valentine models on **animal_tag_data**

| animal_id | animal_type | breed_group | primary_breed | tag_no | tag_type | tag_subtype | tag_stat |
|---|---|---|---|---|---|---|---|
| A577425 | DOG | SETTER/RETRIEVE | GOLDEN RETR | L16-044079 | LIC ALTERED | HLP WEB | RENEWED |
| A583973 | CAT | SHORTHAIR | DOMESTIC SH | L16-056016 | LIC ALTERED | HLP SCAN | RENEWED |
| A562321 | DOG | TOY | SHIH TZU | U15-003015 | RABIES CERT | HLP IMPORT | CURRENT |

- This table also has limited values for each column
- Unclear why the Valentine models did not perform as well

# Results in Detail

Valentine models outperformed Rotom-based models on **contributors_parties**



# Results in Detail

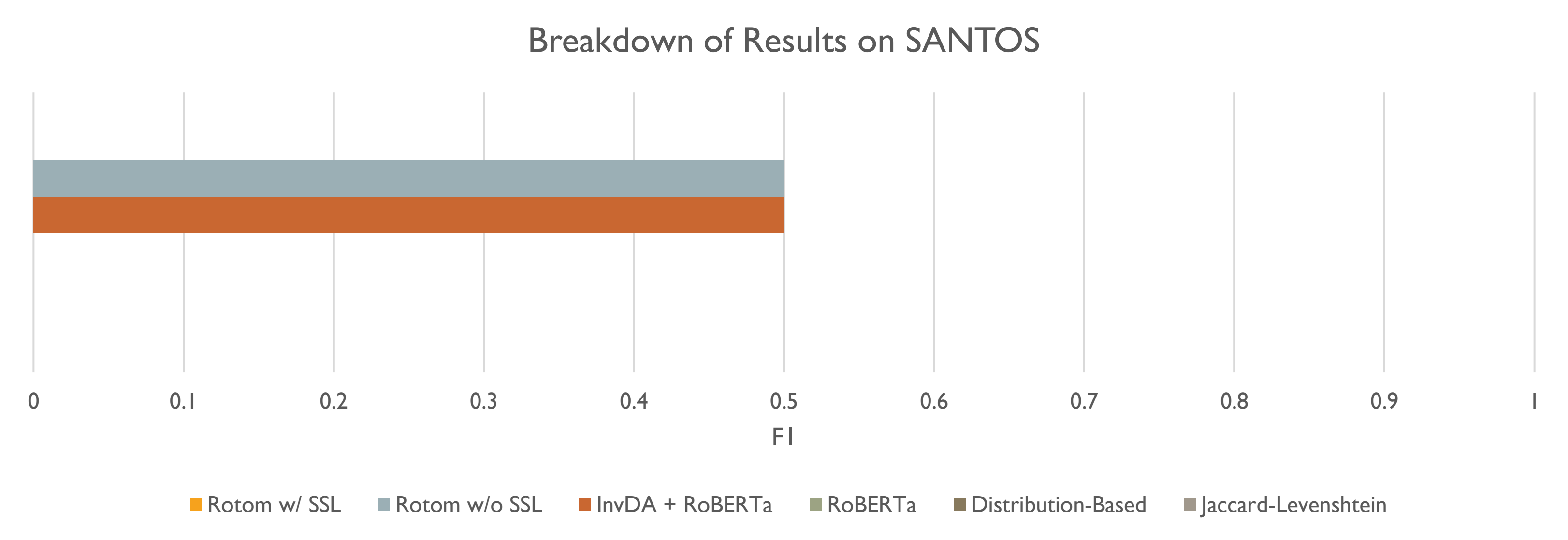
Valentine models outperformed Rotom-based models on **contributors_parties**

| Political Entity | Recipient | Electoral event | Fiscal date |
|---|---|---|---|
| Registered Party | Canadian Reform Conservative Alliance | Annual | 12/31/2002 |
| Registered Party | Canadian Reform Conservative Alliance | Annual | 12/31/2003 |
| Registered Party | New Democratic Party | Annual | 12/31/2003 |

- Rotom’s low effectiveness may be due to the sampling of column tokens during serialization

# Results in Detail

All models performed poorly on **albums**



# Results in Detail

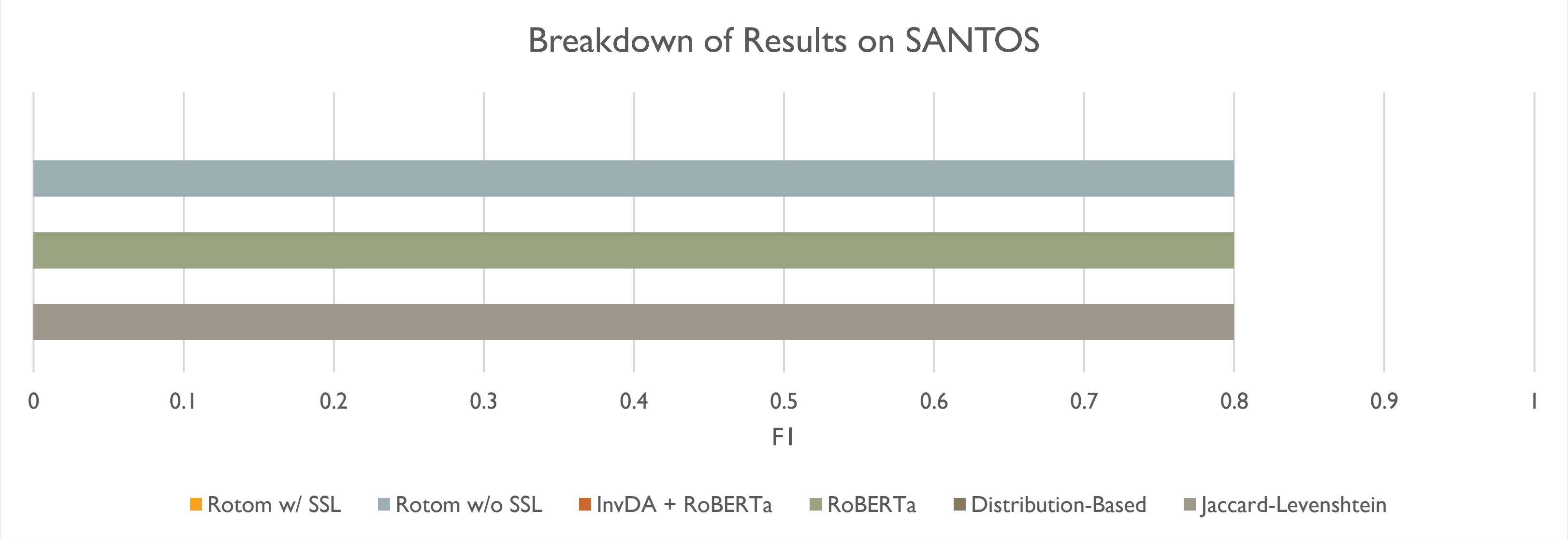
All models performed poorly on **albums**

| **title** | **album** |
|---|---|
| mechanical ape! | charge!! |
| end credits (jfk homage) | songs in the key of springfield |
| stand up (for it) | stand up |

- Very difficult to distinguish between title and album
- Lots of overlap in values

# Results in Detail

There was difficulty on **biodiversity**



# Results in Detail

There was difficulty on **biodiversity**

| scientific name | family name | common name |
|---|---|---|
| symphyotrichum novae-angliae | asteraceae | new england aster |
| phoradendron leucarpum | viscaceae | oak mistletoe |
| acer macrophyllum | aceraceae | bigleaf maple |

- Domain-specific language adds complexity
- Data augmentation did not help

# Overall Discussion

- Results vary greatly from dataset to dataset
- No matching method consistently performs better than the others
- On average, the Rotom meta-learning framework decreases effectiveness compared to the RoBERTa baseline
- On average, data augmentation with InvDA does not improve effectiveness compared to the RoBERTa baseline

# Future Work

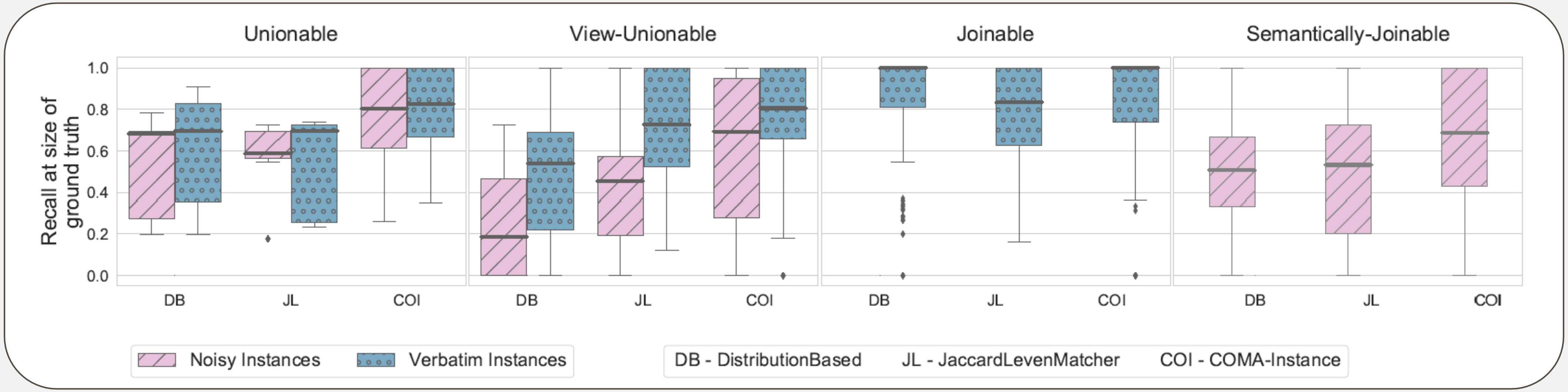
- Refine the pre-processing used for training and test data
  - Token sampling
  - Increase number of samples
- Improve use of Valentine
  - Test against more of the Valentine datasets
  - Tune the matchers' parameters instead of using the default

# Thank you!

# Additional Slides

| Dataset | Rotom w/ SSL | Rotom w/o SSL | InvDA + RoBERTa | RoBERTa | Distribution-Based | Jaccard-Levenshtein |
|---|---|---|---|---|---|---|
| 311_calls_historic_data | 0.941176471 | 0.761904762 | 0.941176471 | 0.875 | 0.857142857 | 0.769230769 |
| abandoned_wells | 1 | 0.769230769 | 1 | 1 | 0.666666667 | 0.8 |
| albums | 0 | 0.5 | 0.5 | 0 | 0 | 0 |
| animal_tag_data | 1 | 1 | 1 | 1 | 0.461538462 | 0.5 |
| biodiversity | 0 | 0.8 | 0 | 0.8 | 0 | 0.8 |
| business_rates | 0.848484848 | 0.612244898 | 0.914285714 | 0.692307692 | | |
| cdc_nutrition_physical_activity | 0.6 | 0.736842105 | 0.8 | 0.833333333 | 0.6 | 0.727272727 |
| cihr_co-applicant | 0.888888889 | 0.571428571 | 0.75 | 0.857142857 | 0.461538462 | 0.823529412 |
| civic_building_locations | 0.5 | 0.615384615 | 0.666666667 | 0.857142857 | | |
| complaint_by_practice | 1 | 0.545454545 | 0.8 | 0.461538462 | 0 | 0.857142857 |
| contributors_parties | 0.857142857 | 0.8 | 0.727272727 | 0.8 | 1 | 1 |
| data_mill | 0.75 | 0.666666667 | 0.875 | 0.666666667 | | |
| deaths_2012_2018 | 0.842105263 | 0.833333333 | 0.869565217 | 0.740740741 | | |
| film_locations_in_san_francisco | 0 | 0 | 0.5 | 0.833333333 | 0 | 0.285714286 |
| immigration_records | 0.8 | 0.5 | 0.714285714 | 0.857142857 | 0.666666667 | 0.666666667 |
| ipopayments | 0.4 | 0.7 | 0.8 | 0.888888889 | 0.545454545 | 0.666666667 |
| job_pay_scales | 0.727272727 | 0.615384615 | 0.769230769 | 0.923076923 | 0.444444444 | 0.923076923 |
| lane_description | 1 | 1 | 0.888888889 | 1 | 0.857142857 | 1 |
| mines | 0.857142857 | 0.857142857 | 1 | 0.75 | 0.857142857 | 0.666666667 |
| monthly_data_feed | 0.5 | 0.666666667 | 0.444444444 | 0.444444444 | 0 | 0 |
| new_york_city_restaurant_inspec | 0.842105263 | 0.571428571 | 0.64 | 0.689655172 | 0.461538462 | 0.75 |
| oil_and_gas_summary_production_ | 0.888888889 | 0.909090909 | 0.8 | 0.666666667 | 0.571428571 | 0.75 |
| practice_reference | 0.8 | 0.666666667 | 0.769230769 | 0.705882353 | 0.5 | 1 |
| prescribing | 0.5 | 0.8 | 0.666666667 | 0.666666667 | 0.5 | 0.909090909 |
| psyckes_antipsychotic_polypharm | 0.75 | 0.8 | 0.833333333 | 0.714285714 | 0.571428571 | 1 |
| purchasing_card | 0.875 | 0.761904762 | 0.842105263 | 0.64 | 0.5 | 0.714285714 |
| report_card_discipline_for_2015 | 0.8 | 0.75 | 0.8 | 0.75 | 0.571428571 | 0.615384615 |
| senior_officials_expenses | 0.2 | 0.592592593 | 0.592592593 | 0.56 | 0 | 0 |
| stockport_contracts | 0.642857143 | 0.682926829 | 0.64 | 0.62745098 | 0.272727273 | 0.944444444 |
| time_spent_watching_vcr_movies | 1 | 0.5 | 0.857142857 | 0.75 | 1 | 1 |
| tuition_assistance_program_tap_ | 0.888888889 | 0.545454545 | 0.714285714 | 0.714285714 | 0.75 | 0.888888889 |
| wholesale_markets | 0.727272727 | 0.823529412 | 0.714285714 | 0.736842105 | 0.6 | 0.833333333 |
| workforce_management_information | 0.845070423 | 0.352941176 | 0.845070423 | 0.987341772 | | |
| ydn_spending_data | 0.769230769 | 0.666666667 | 0.625 | 0.823529412 | | |
| **Average** | **0.707103765** | **0.675731957** | **0.744133234** | **0.744510753** | **0.489867474** | **0.71040696** |

# Valentine Instance-Based Matcher Results




- Effectiveness results of instance-based matching methods for each dataset relatedness scenario
- Results on fabricated datasets