Interpreting and Representing Tables as Linked Data

Varish Mulwad, Tim Finin, Zareen Syed and Anupam Joshi
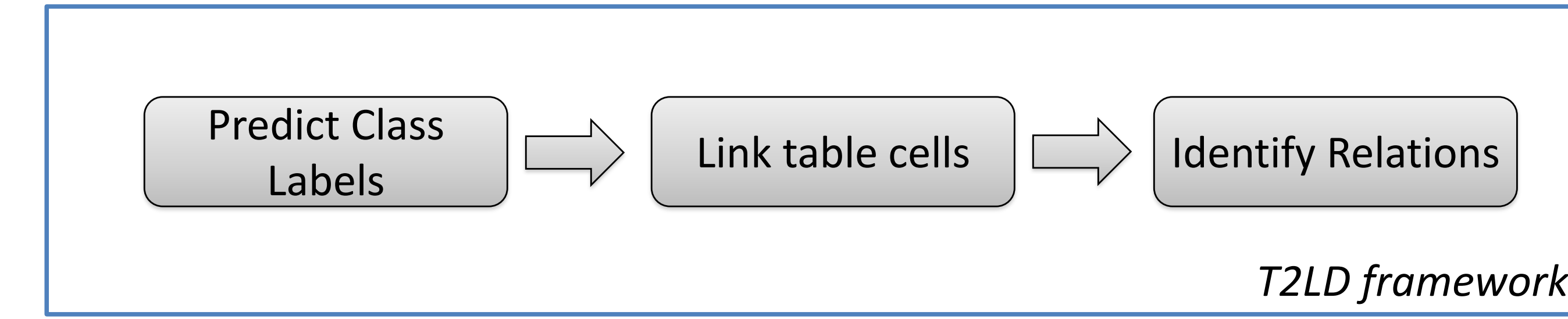## University of Maryland, Baltimore County

1

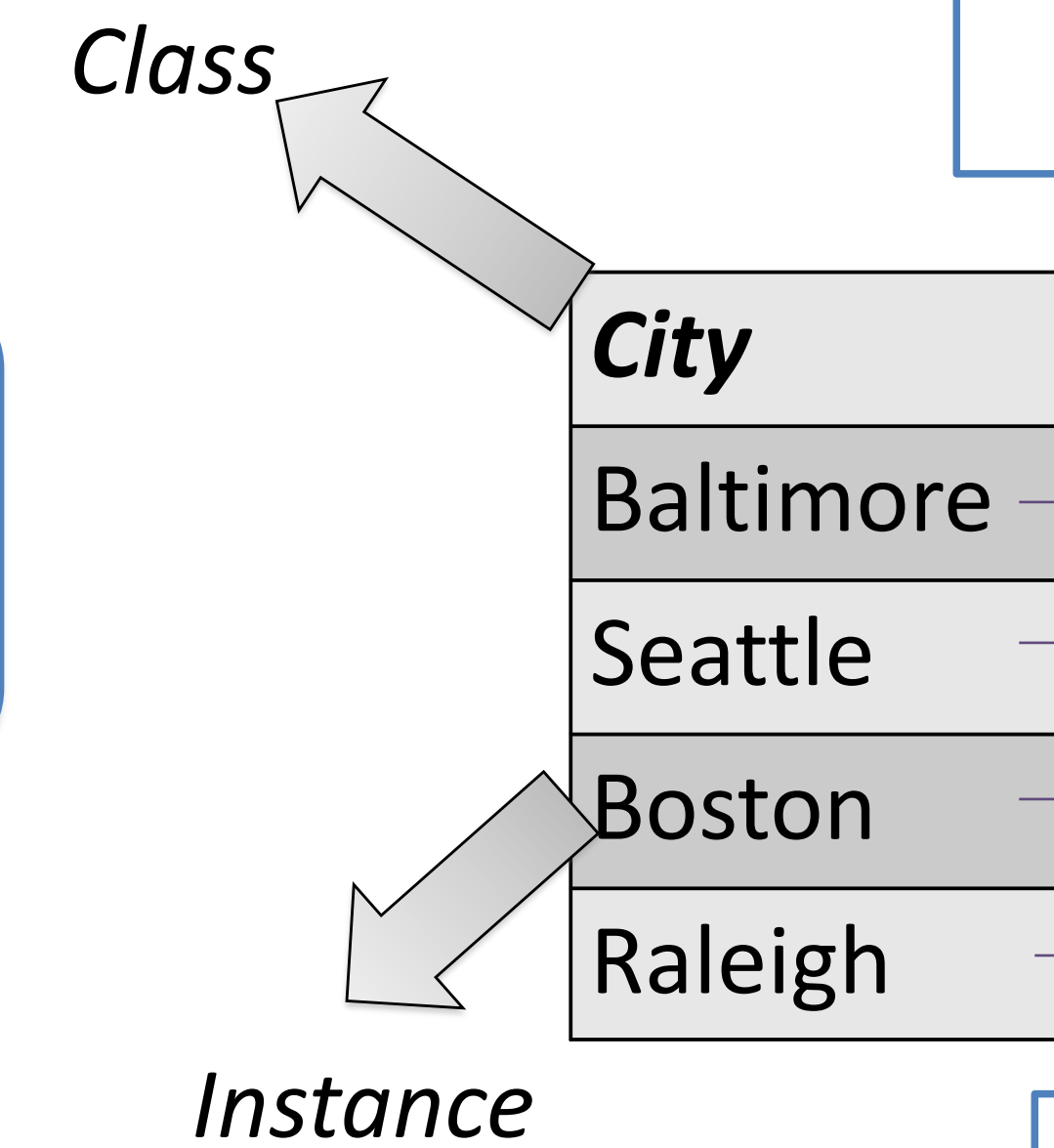- ✓ **We are laying a strong foundation for the Semantic Web**

- ✓ … but an old problem haunts us …. Chicken ? Egg ?
- ✓ Lack of structured data on the Semantic web

We need systems that can (semi) automatically convert and represent data for the Semantic Web !
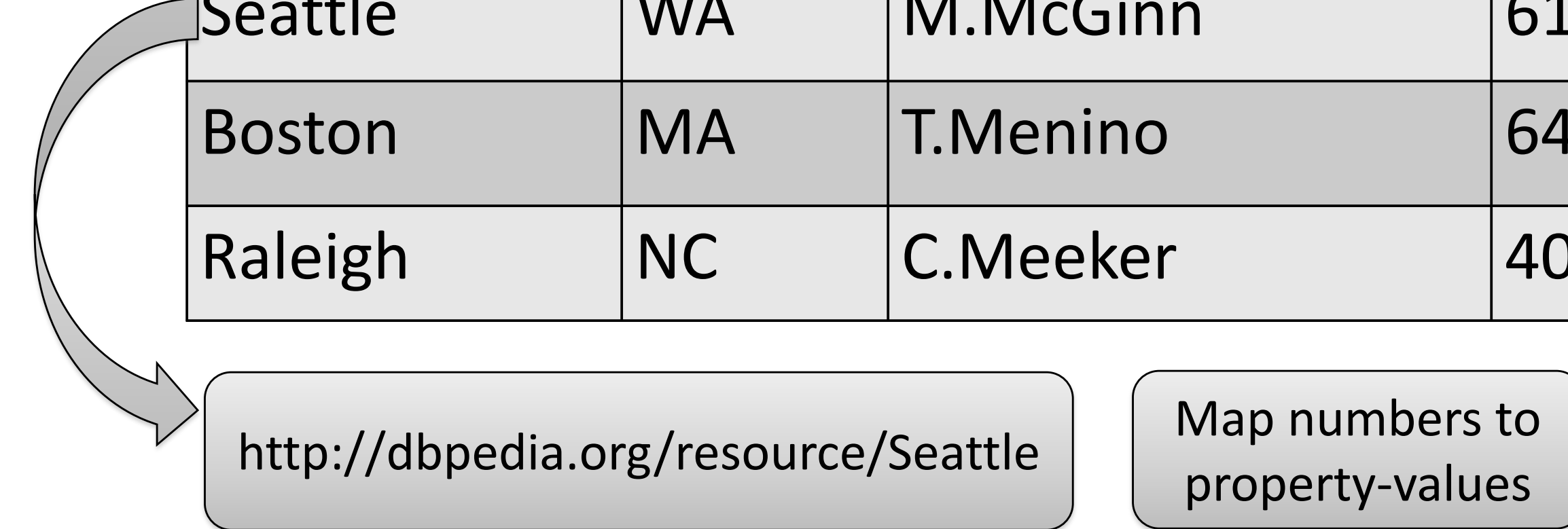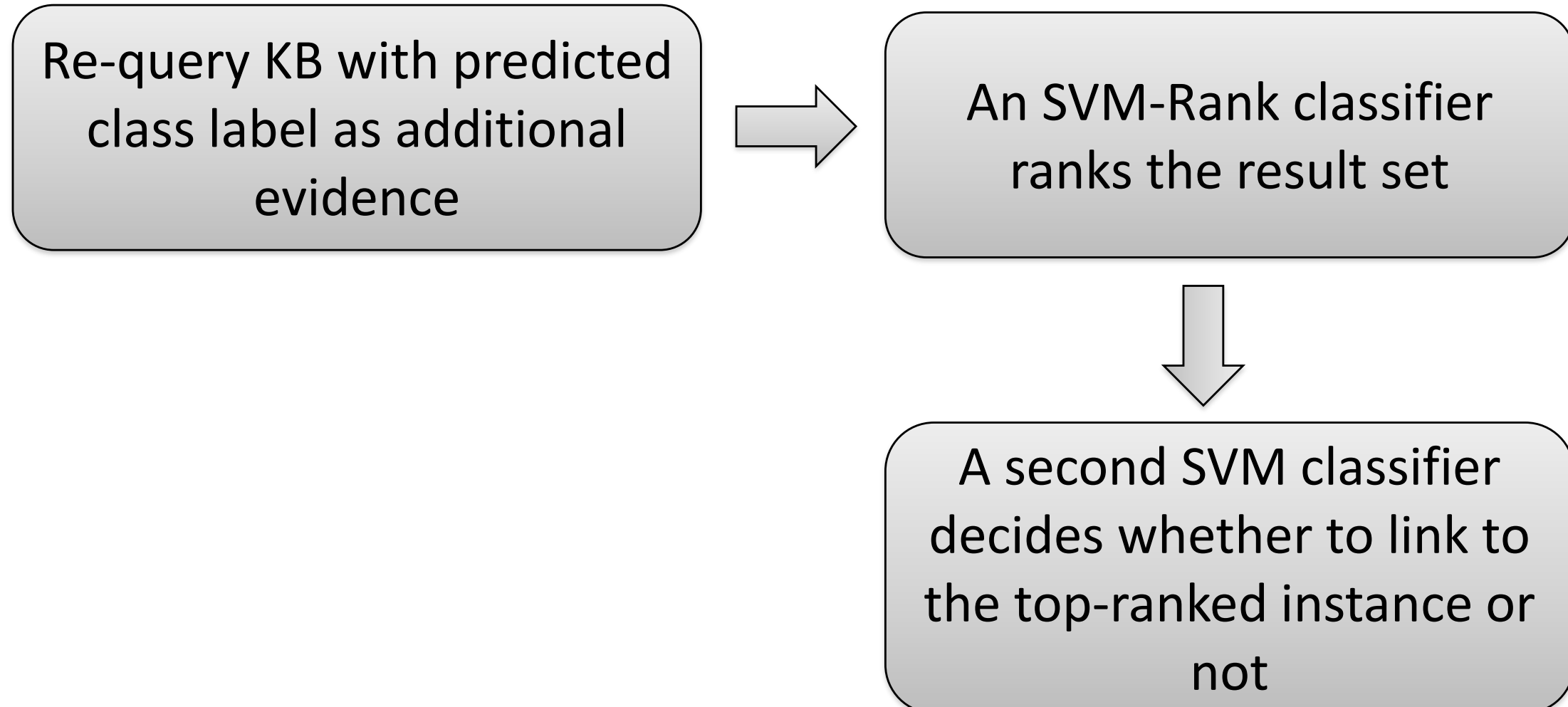
- ✓ More than a trillion documents on the web
- ✓ ~ 14.1 billion tables, *154 million* with *high quality relational data*
- ✓ 305,632 Datasets available on Data.gov (US) + *7 Other nations* establishing open data; not all is RDF
- ✓ **Is it practical to convert this into RDF manually ? NO !!**

2

- ✓ We develop a multi-stage framework for this task

| Predict Class Labels | → | Link table cells | → | Identify Relations |

T2LD framework

Class

Instance

Class for the column

City
Baltimore
Seattle
Boston
Raleigh

http://dbpedia.org/ontology/AdministrativeRegion

dbpedia-prop:largestCity

- ✓ For every string in the column, query the knowledge base (KB)
- ✓ Generate a set of possible classes
- ✓ Rank the classes and choose the best

Wikitology — WordNet — Yago — DBpedia — Freebase

- ✓ Wikitology (our KB) is a hybrid KB consisting of structured and unstructured data from Wikipedia, Dbpedia, Freebase, WordNet and Yago

City	State	Mayor	Population
Baltimore	MD	S.C.Rawlings-Blake	637,418
Seattle	WA	M.McGinn	617,334
Boston	MA	T.Menino	645,169
Raleigh	NC	C.Meeker	405,791

http://dbpedia.org/resource/Seattle

Map numbers to property-values

3

- ✓ *Linking table cells* – A machine learning based approach

| Re-query KB with predicted class label as additional evidence | → | An SVM-Rank classifier ranks the result set |

↓

| A second SVM classifier decides whether to link to the top-ranked instance or not |

- ✓ *Relation identification*

Relation 'A'

City		State
Baltimore	Relation 'A'	MD
Seattle	Relation 'A','B'	WA
Boston	Relation 'A', 'C'	MA
Raleigh	Relation 'A'	NC

- ✓ For every pair of linked strings in the two column, query the knowledge base (KB)
- ✓ Generate a set of possible relations
- ✓ Rank the relations and choose the relation

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
@prefix dbpprop: <http://dbpedia.org/property/> .
"City"@en is rdfs:label of dbpedia-owl:City .
"State"@en is rdfs:label of dbpedia-owl:AdminstrativeRegion .
"Baltimore"@en is rdfs:label of dbpedia:Baltimore .
dbpedia:Baltimore a dbpedia-owl:City .
"MD"@en is rdfs:label of dbpedia:Maryland .
dbpedia:Maryland a dbpedia-owl:AdministrativeRegion .
dbpprop:LargestCity rdfs:domain dbpedia-owl:AdminstrativeRegion .
dbpprop:LargestCity rdfs:range dbpedia-owl:City .
```
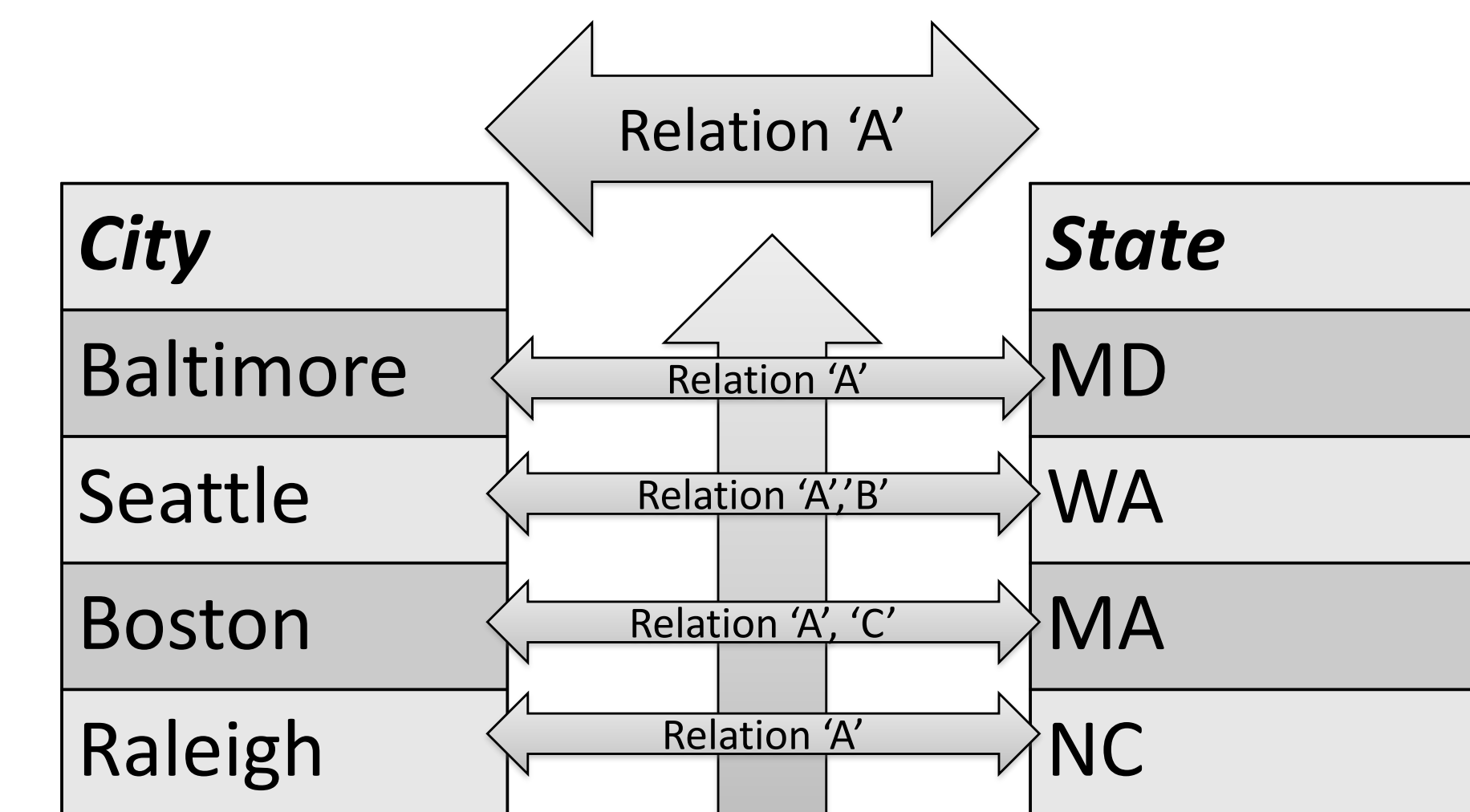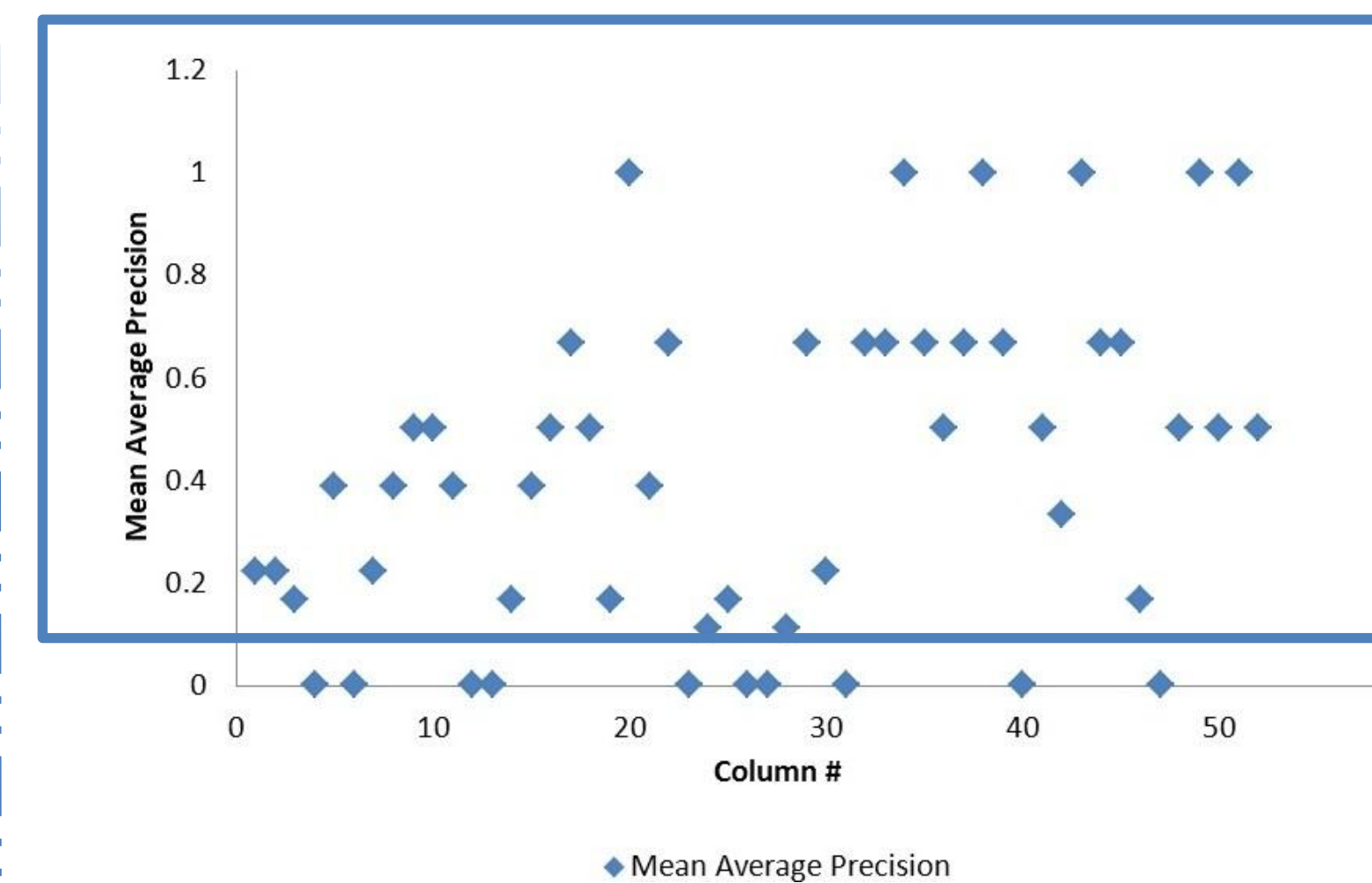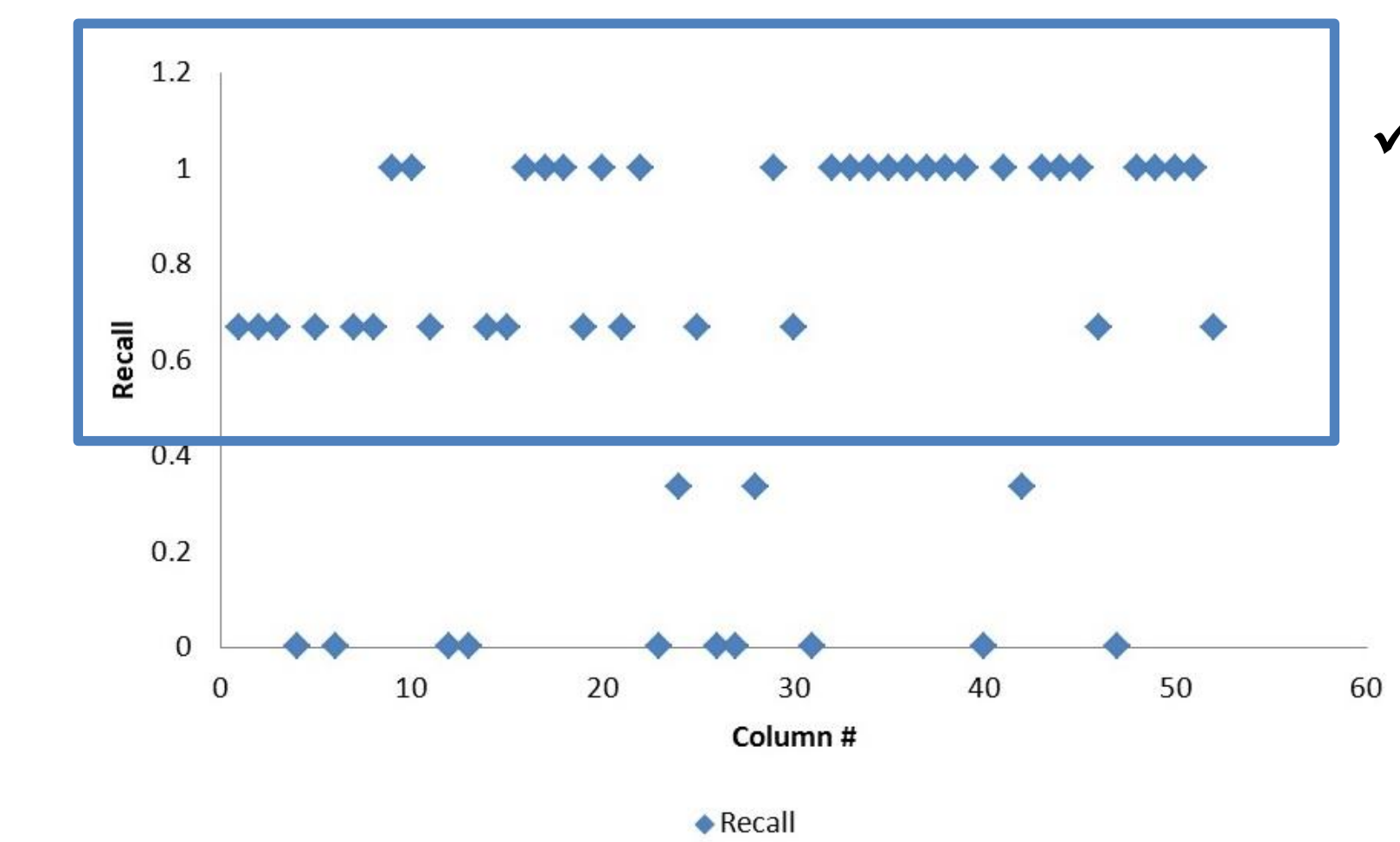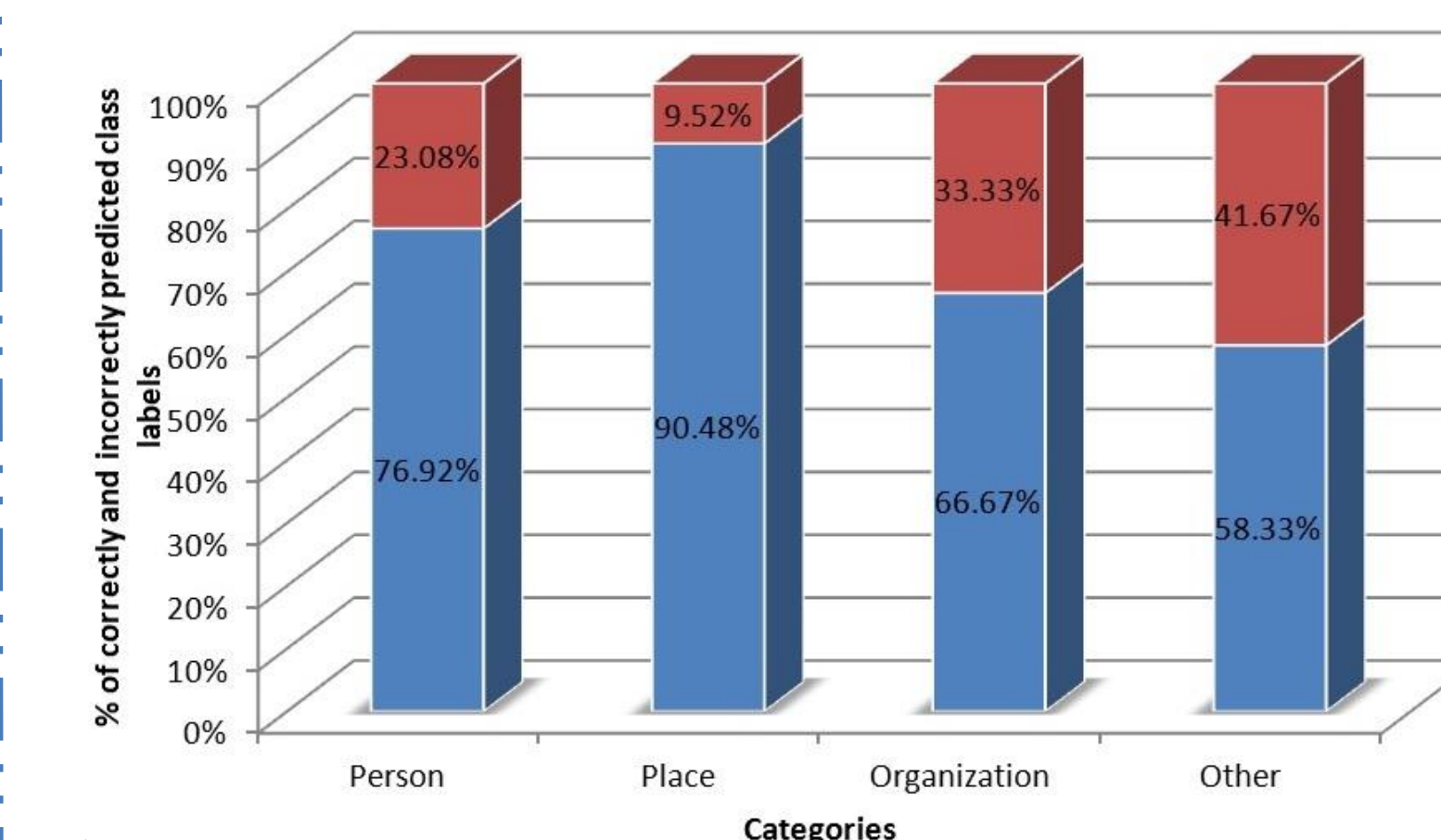
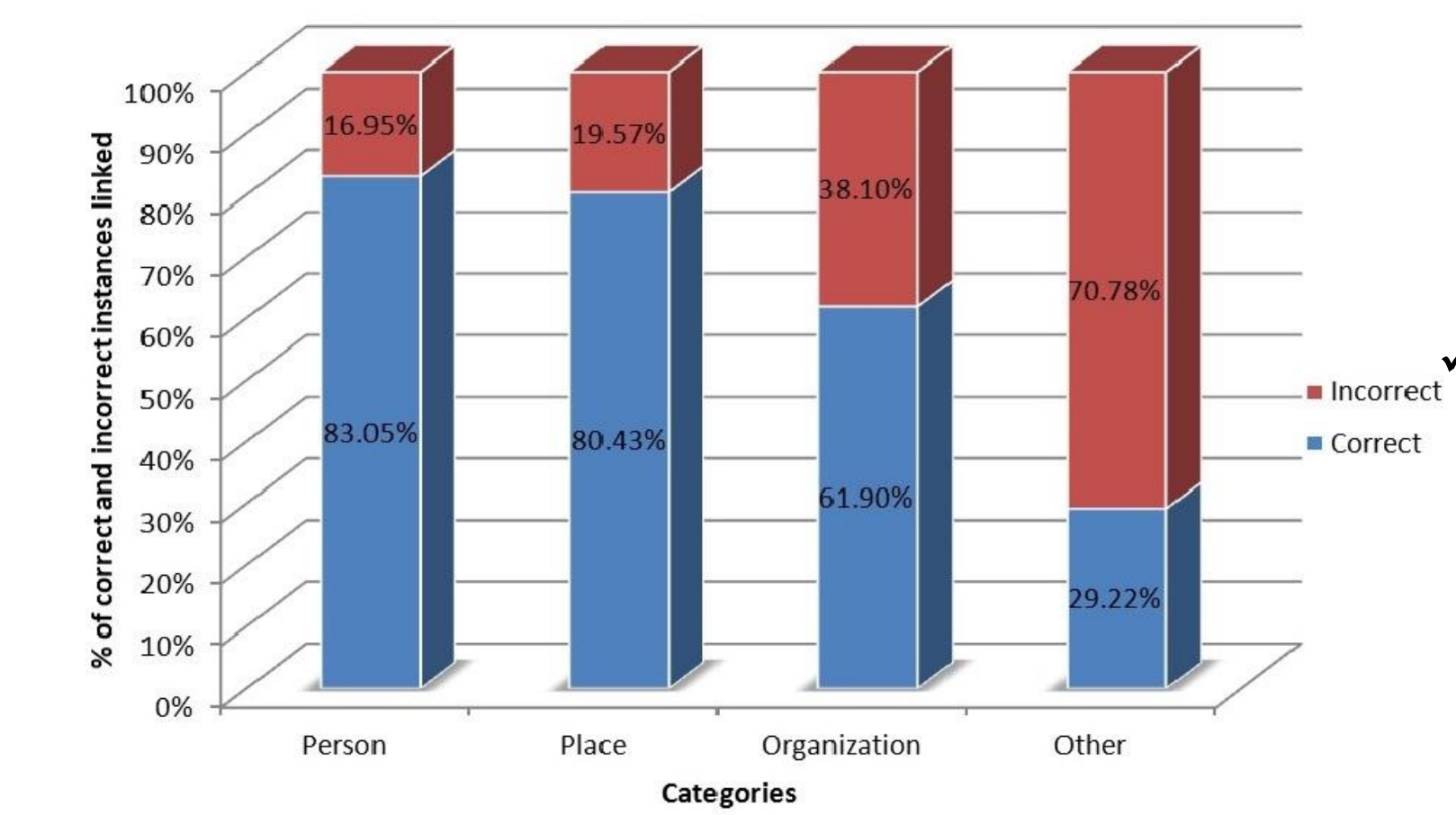- ✓ *A template for representing tables as linked data*

4

# of Tables	15
# of Columns	52
# of Entities	611

Dataset

- ✓ MAP > 0 for 81 % of the columns

- ✓ Recall >= 0.6 for 75 % of the columns

Mean Average Precision

Recall

- ✓ *Column Correctness*

	Person	Place	Organization	Other
Incorrect	23.08%	9.52%	33.33%	41.67%
Correct	76.92%	90.48%	66.67%	58.33%

% of correctly and incorrectly predicted class labels — Categories

- ✓ *Entity Linking*

	Person	Place	Organization	Other
Incorrect	16.95%	19.57%	18.10%	70.78%
Correct	83.05%	80.43%	61.90%	29.22%

% of correct and incorrect instances linked — Categories

- ✓ *Column Correctness*

| Column – Nationality Prediction – MilitaryConflict ❌ | Column – Birth Place Prediction – PopulatedPlace ✓ |

- ✓ *Relation Identification*

| Preliminary results for relation identification – 25 % accuracy 🙁 |

UMBC ebiquity
Microsoft Research