Assigning Semantic Labels to Data Sources

Authors:

S.K. Ramnandan[1], Amol Mittal[2], Craig Knoblock[3], Pedro Szekely[3]

[1] Indian Institute of Technology - Madras
[2] Indian Institute of Technology - Delhi
[3] University of Southern California

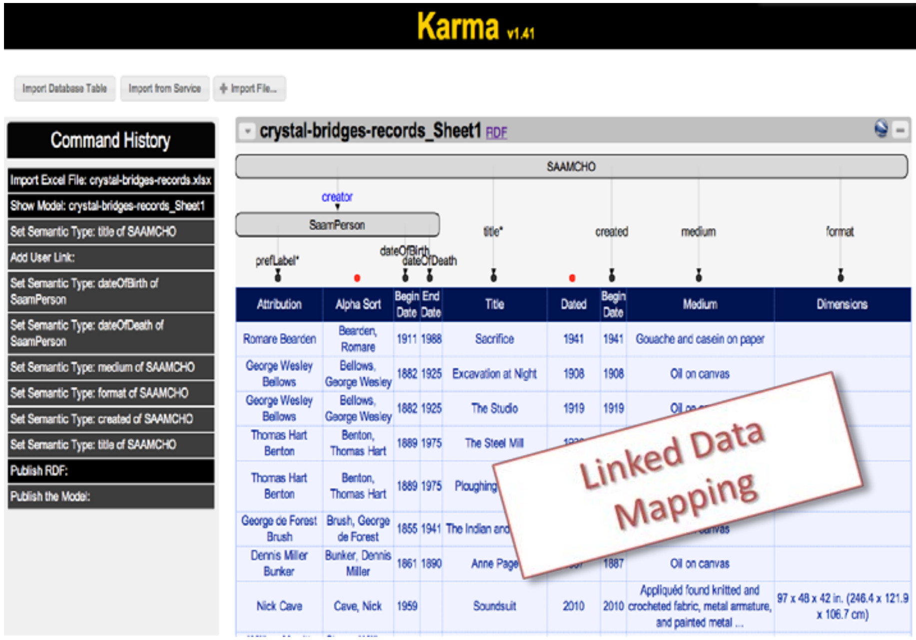
Introduction

Motivation:

- To automatically construct a semantic model of a set of data sources using domain ontologies selected by user

Applications:

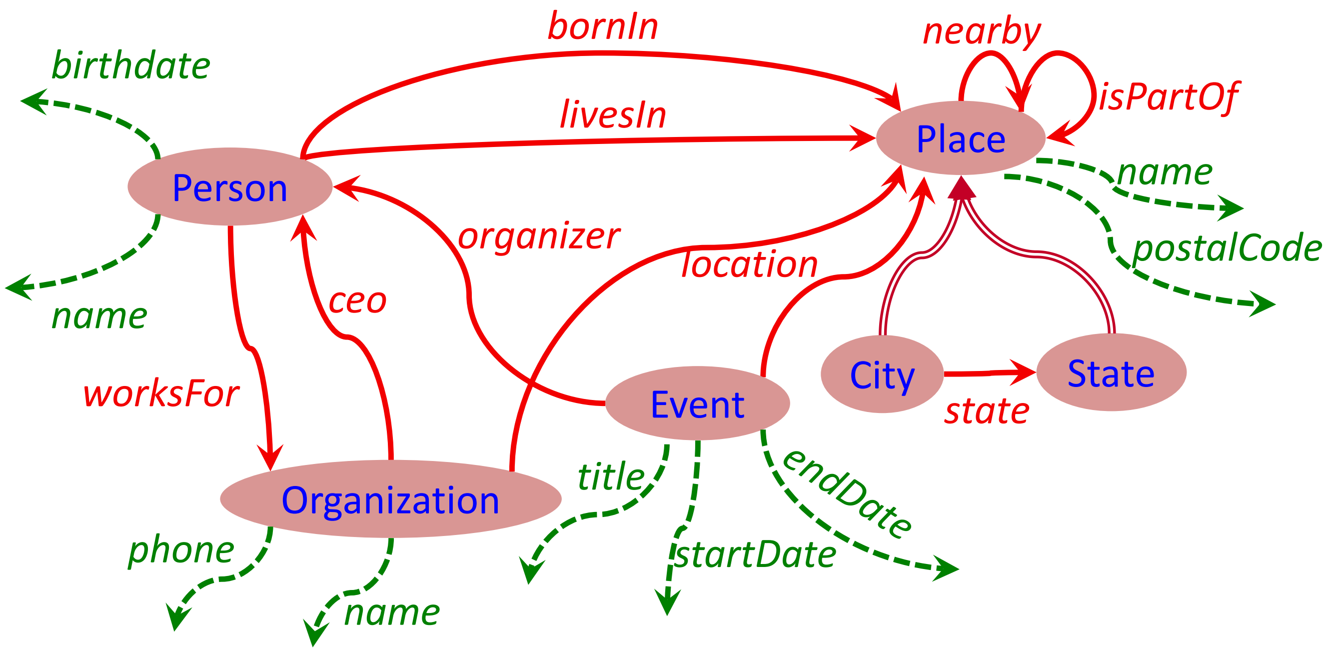
- Provides support to automate many tasks
  - Data integration
  - Source discovery
  - Service composition
  - Building knowledge graphs
- Manual description
  - tedious & time-consuming



What is a semantic model?

Description of the source in terms of the concepts and relationships defined by the **domain ontology**

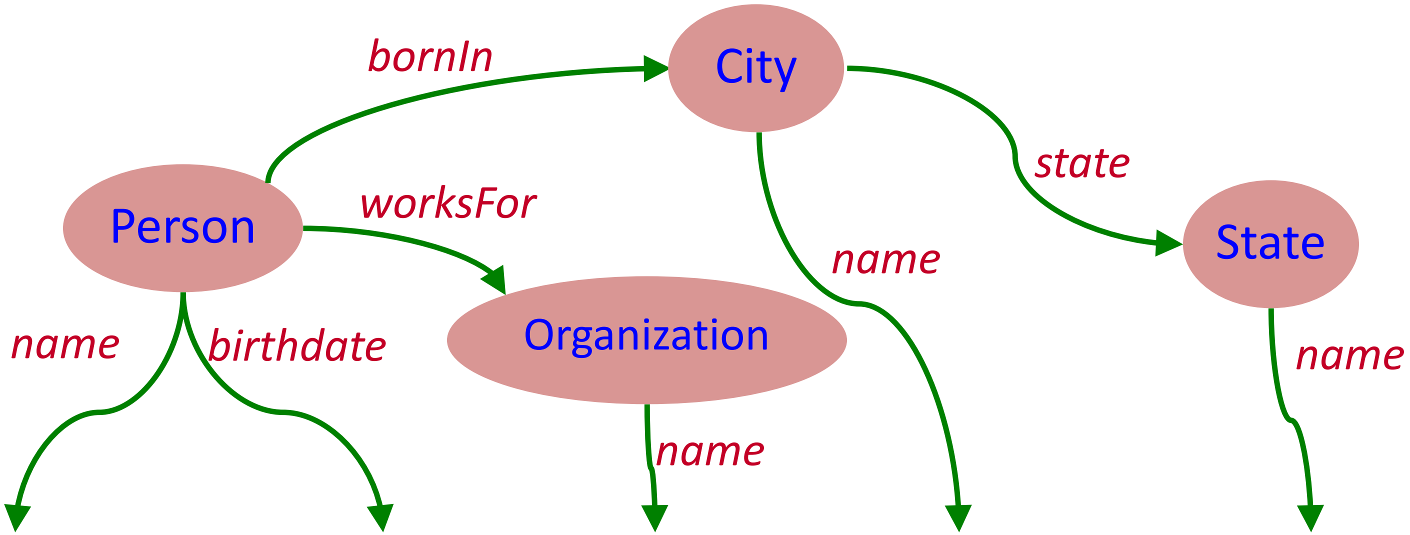
Domain Ontology



Data Source

Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

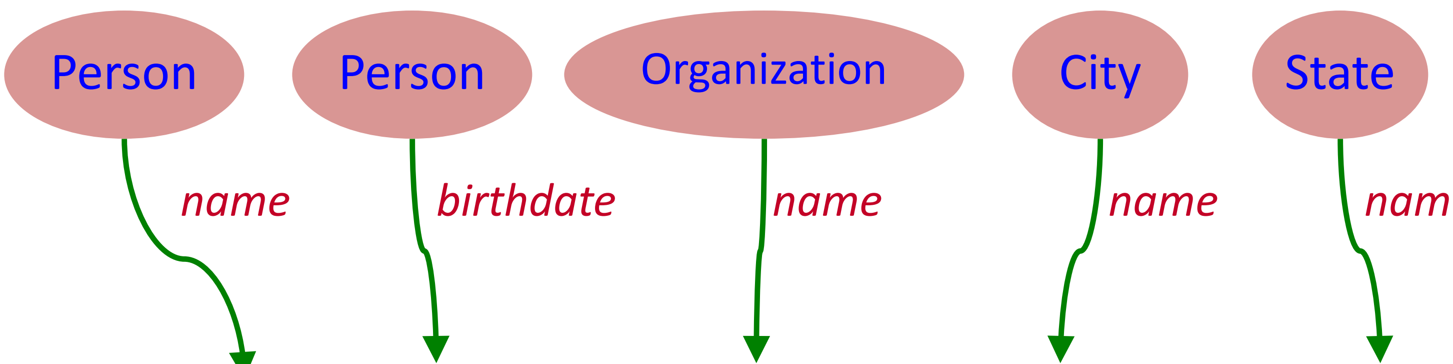
Example semantic model



Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

Semantic Labeling Step

Assigning a class or data property (semantic type) from the ontology to each attribute in the source



Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

Overall approach - semantic modeling

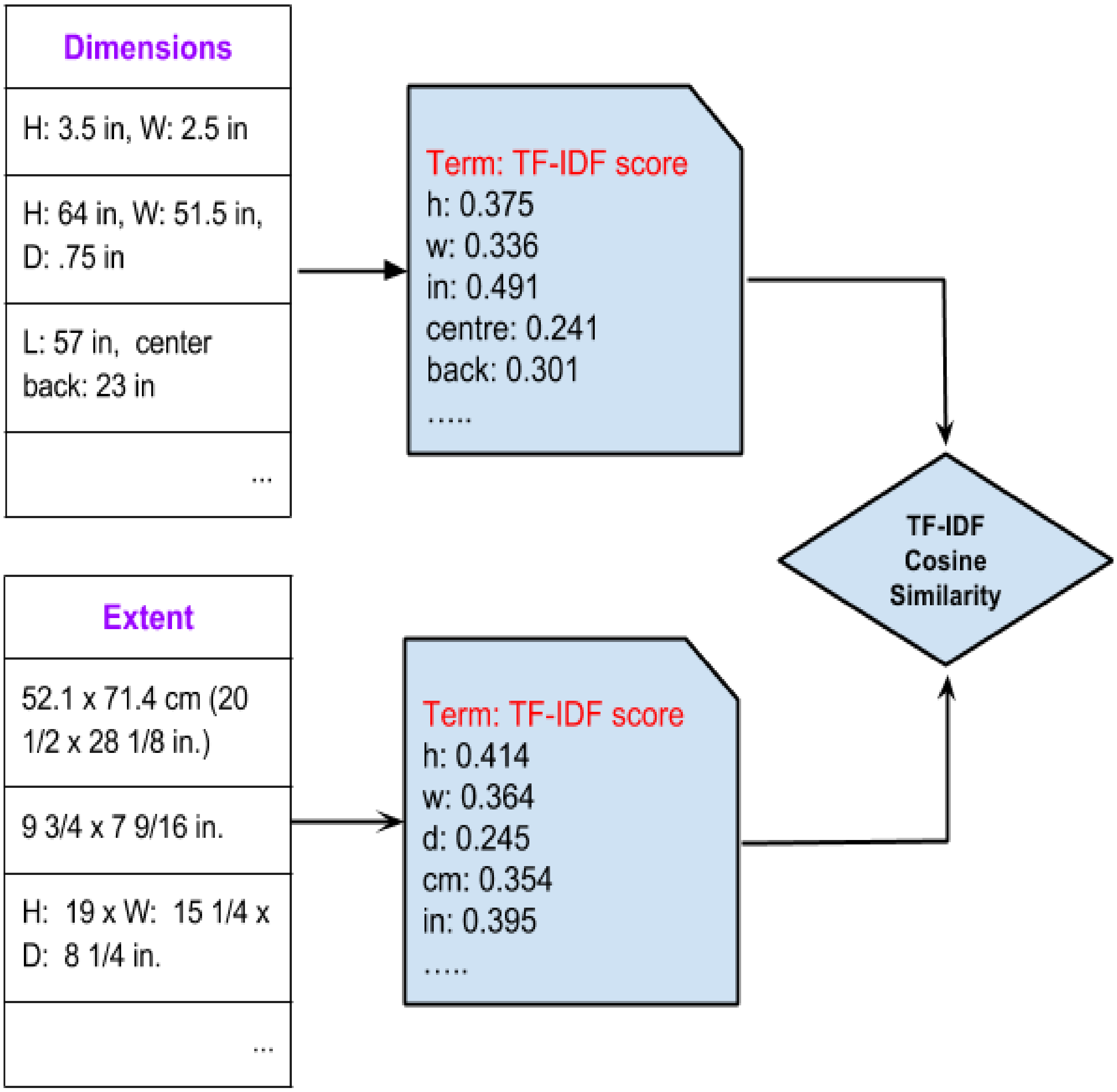
- Taheriyan et al., ISWC 2013, ICSC 2014

- Problems with model-based machine learning techniques (like CRF):
  - Low prediction accuracy for numeric data
  - Training time scales poorly as no. of ontology data properties increases

Overall Approach (SemTyper)

- Holistic view of data values to capture characteristic property of semantic type
- Textual Data : TF-IDF Cosine Similarity
- Numeric Data: Kolmogorov-Smirnov Test
- Top-k suggestions returned to the user based on the confidence scores

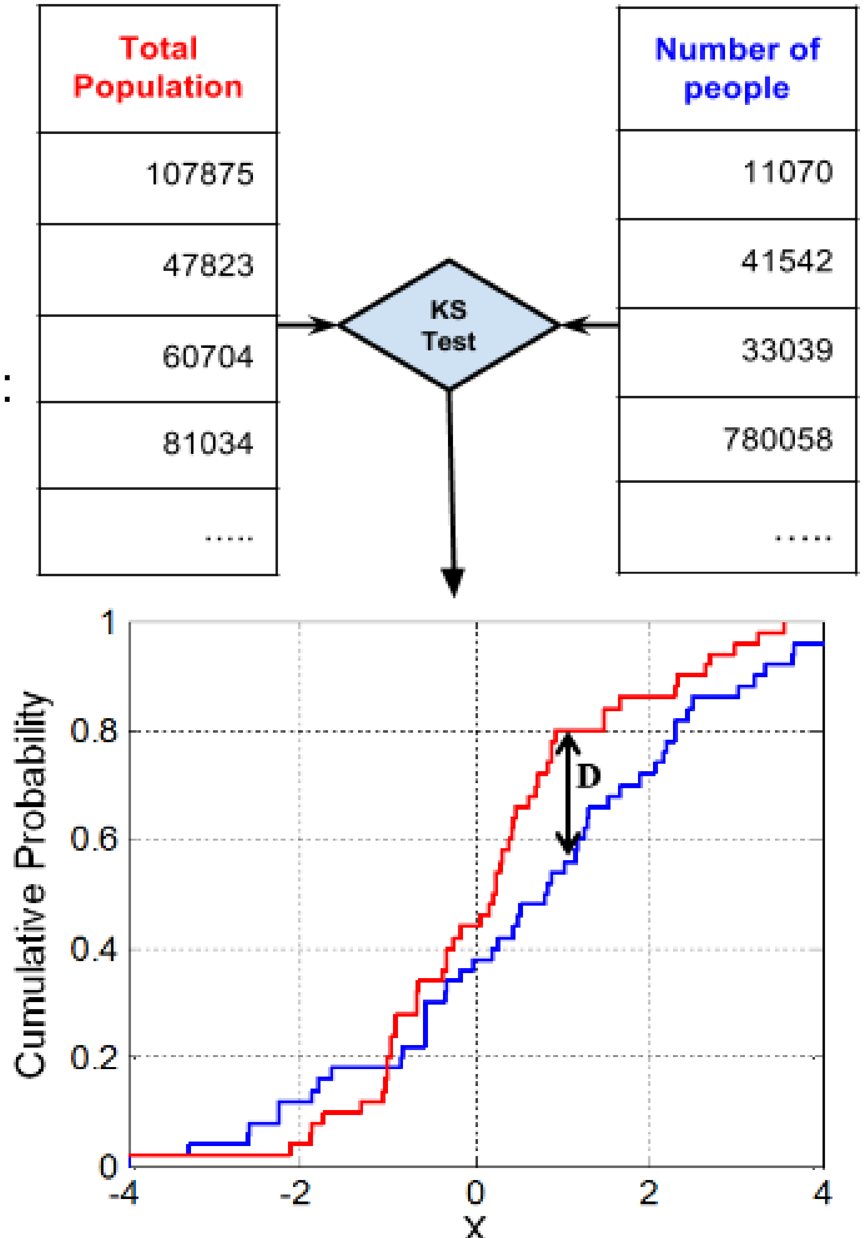
Approach to Textual Data



Approach to Numeric Data

Candidate Statistical Hypothesis tests:

- Welch's t-test
- Mann-Whitney U-test
- **Kolmogorov-Smirnov Test**



Handling noisy datasets

- How to infer if data is textual or numeric in a noisy source?
  - **Training time**: fraction of numeric values
    - < 60% - trained as purely textual
    - > 80% - trained as purely numeric
    - else - trained as both textual and numeric
  - **Prediction time**: fraction of numeric values
    - > 70% - tested as numeric data
    - else - tested as textual data
- Thresholds empirically chosen using coarse grid search
  - Measuring label prediction accuracy on held out set

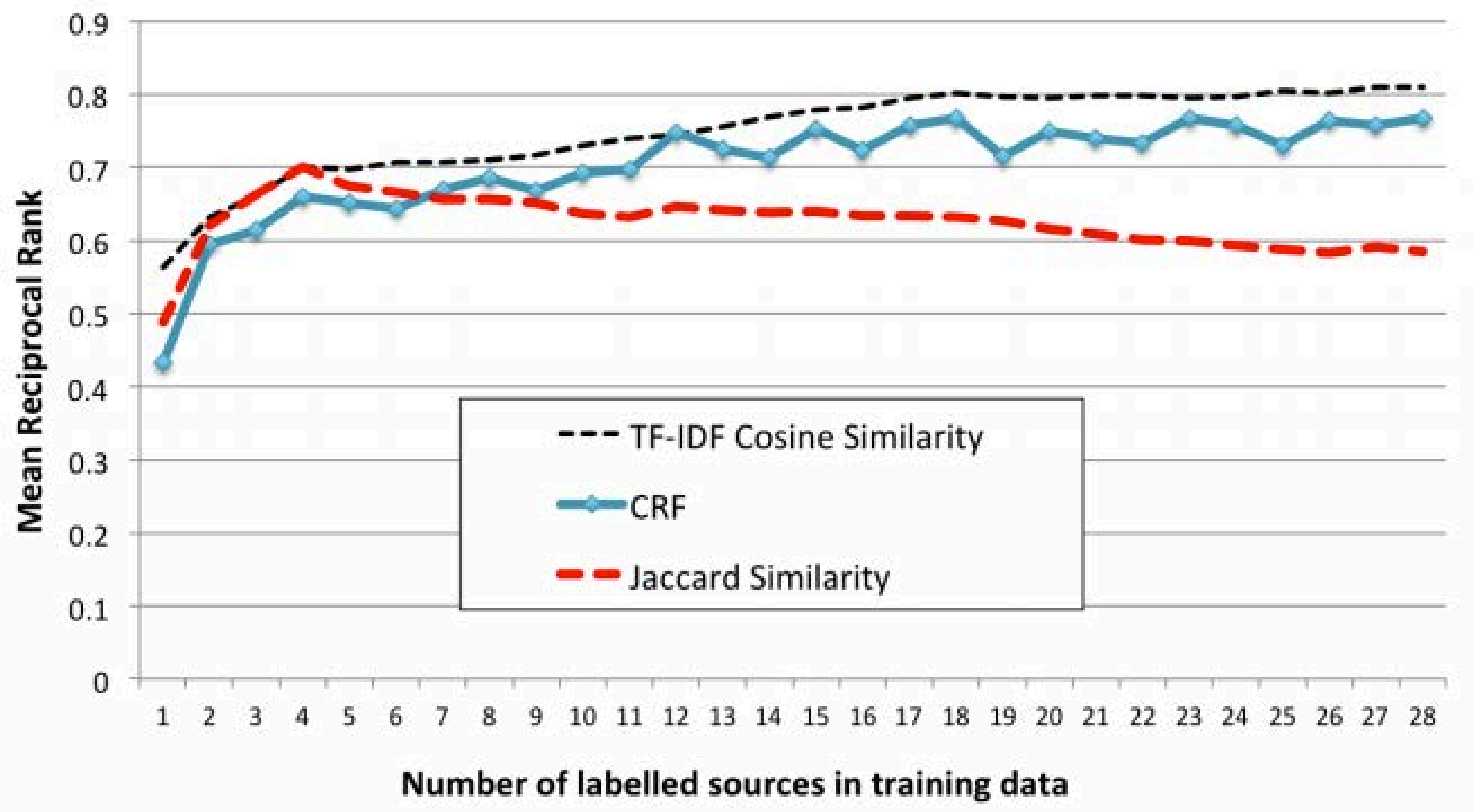
Datasets (Evaluation)

- Purely textual data
  - Museum domain: 29 museum data sources (Taheriyan et al.)
- Purely numeric data
  - City domain:
    - 30 numeric data properties from City class in Dbpedia
    - Partitioned into 10 data sources
- Mixture of textual & numeric data
  - City domain:
    - 52 data properties from City class in DBpedia
  - Weather, phone directory and flight status domains (Ambite et al.)

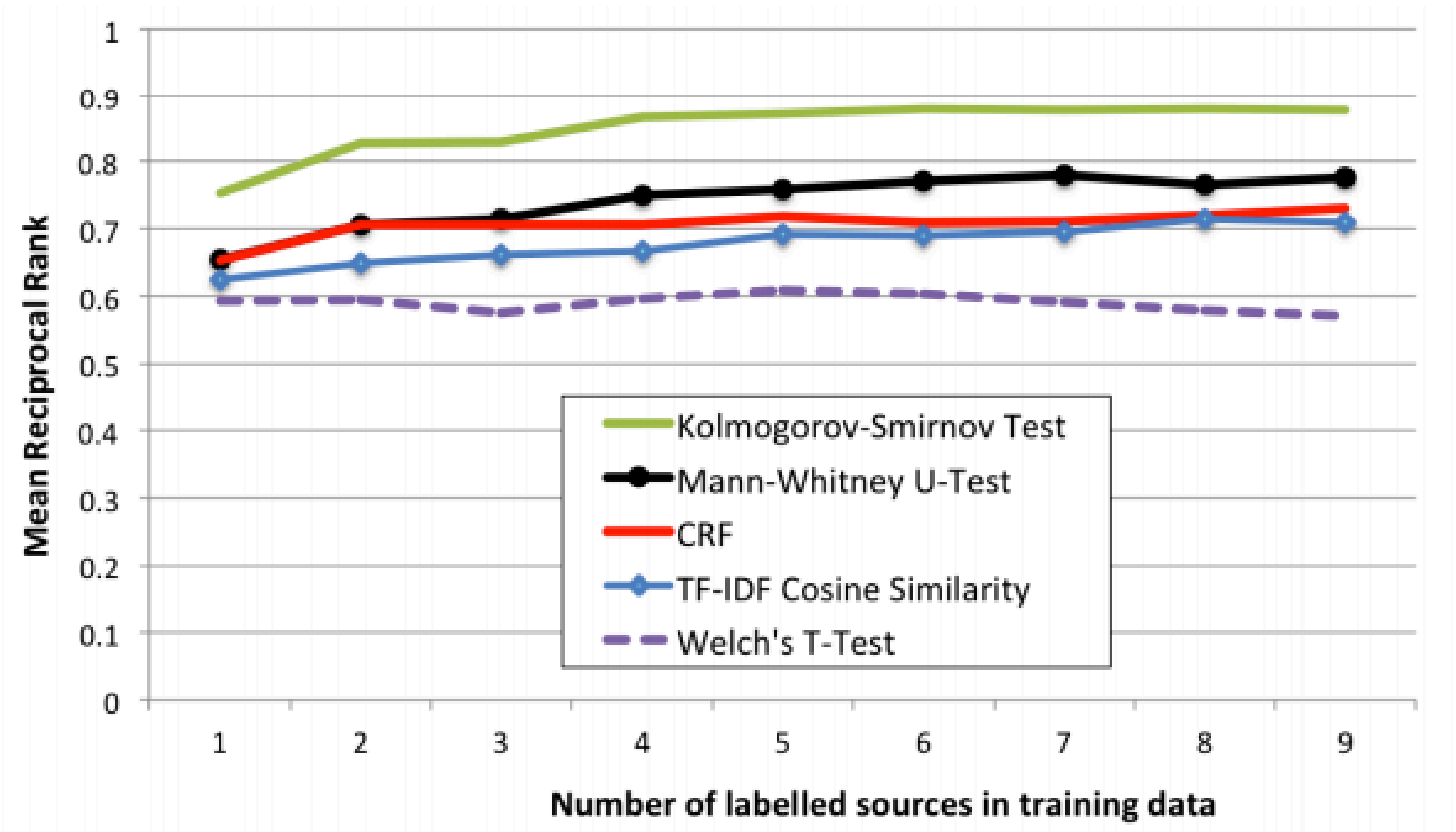
Metrics (Evaluation)

- Mean Reciprocal Rank
  - Interested in rank at which correct semantic label is predicted
- Average Training Time

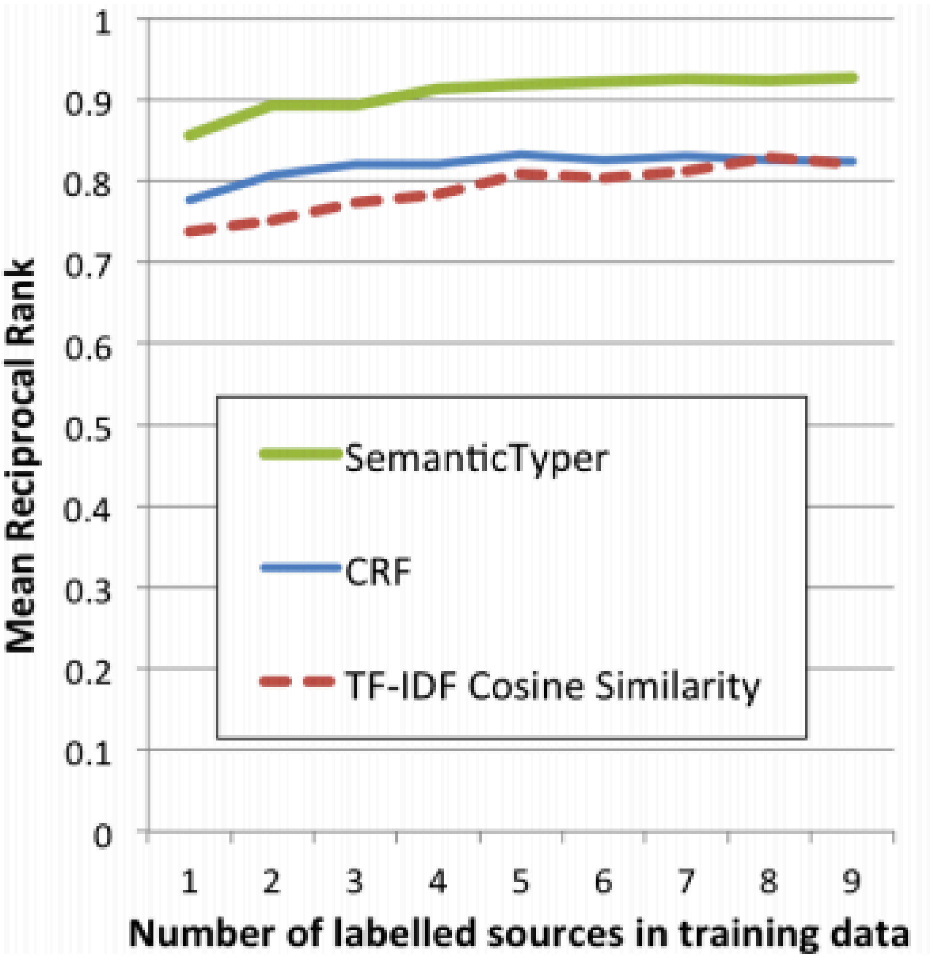
Evaluation (Textual data- Museum domain)

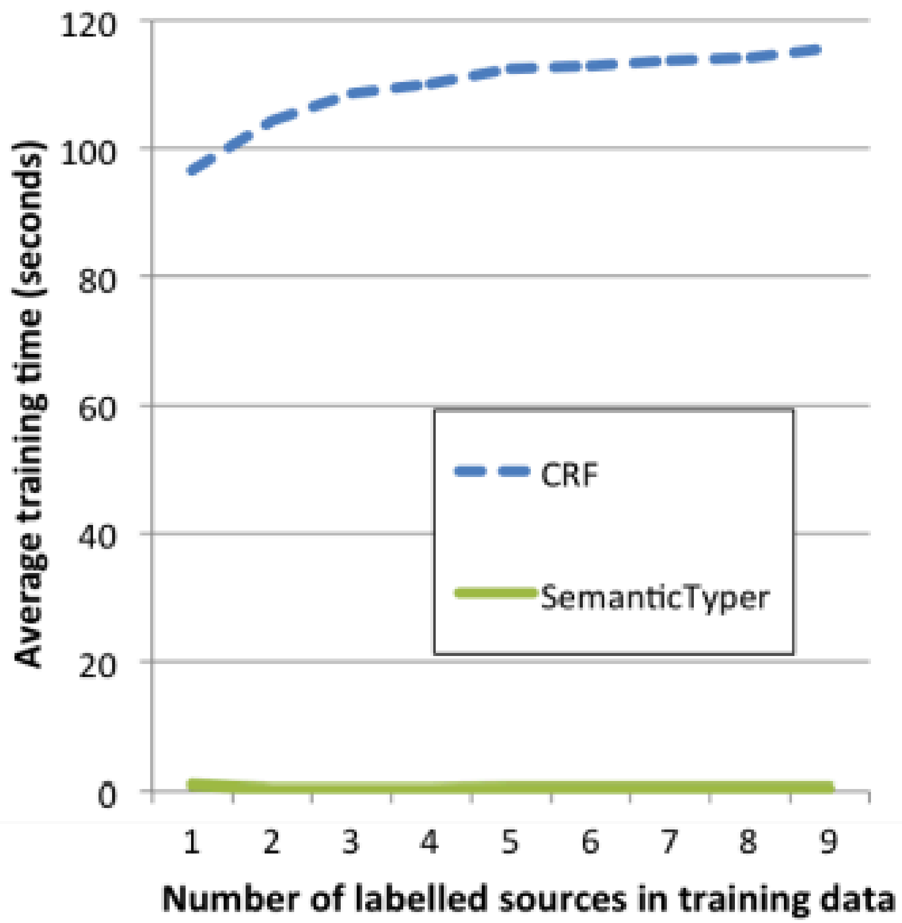


Evaluation (Numeric data- City domain)



Evaluation (Mixture data- City domain)





Evaluation (Mixture data- other domains)

Domain	No.of sources	No.of textual labels/source	No.of numeric labels/source	Max. MRR		
				CRF	TF-IDF	SemTyper
Weather	4	7	4	0.875	0.943	0.955
Flight Status	2	6	3	0.421	0.590	0.646
Phone Directory	3	8	1	0.704	0.831	0.831

Related Work

- Using model-based machine learning techniques
  - Goel et al. (ICAI 2012), Limaye et al. (PVLDB 2010), Mulwad et al. (ISWC 2013)
  - ✓ Extract features from individual data values and build graphical model
  - ✓ Do not extract characteristic properties of column data as a whole
  - ✓ Training graphical models not scalable – explosion of search space

- Using external knowledge
  - Venetis et al. (VLDB 2011), Syed et al. (SWSC 2010)
  - ✓ Leverage knowledge on Web to label individual data values
  - ✓ Restricted to domains and ontologies - huge amount of extracted data
  - ✓ Highly ontology specific – models generated from specific ontologies

- Stonebraker et al. (CIDR 2013)
  - ✓ Address problem of schema matching
  - ✓ Draw inspiration in combining collection of experts

Conclusion

- Label Prediction Accuracy
  - Our approach improves on accuracy of competing approaches on wide variety of domains
- Efficiency & Scalability
  - About 250 times faster than Conditional Random Fields based semantic labeling technique
- Capable of handling noisy datasets
- Ontology agnostic
  - Learns semantic labeling function with respect to ontologies selected by users for their application

Thank You

Questions?