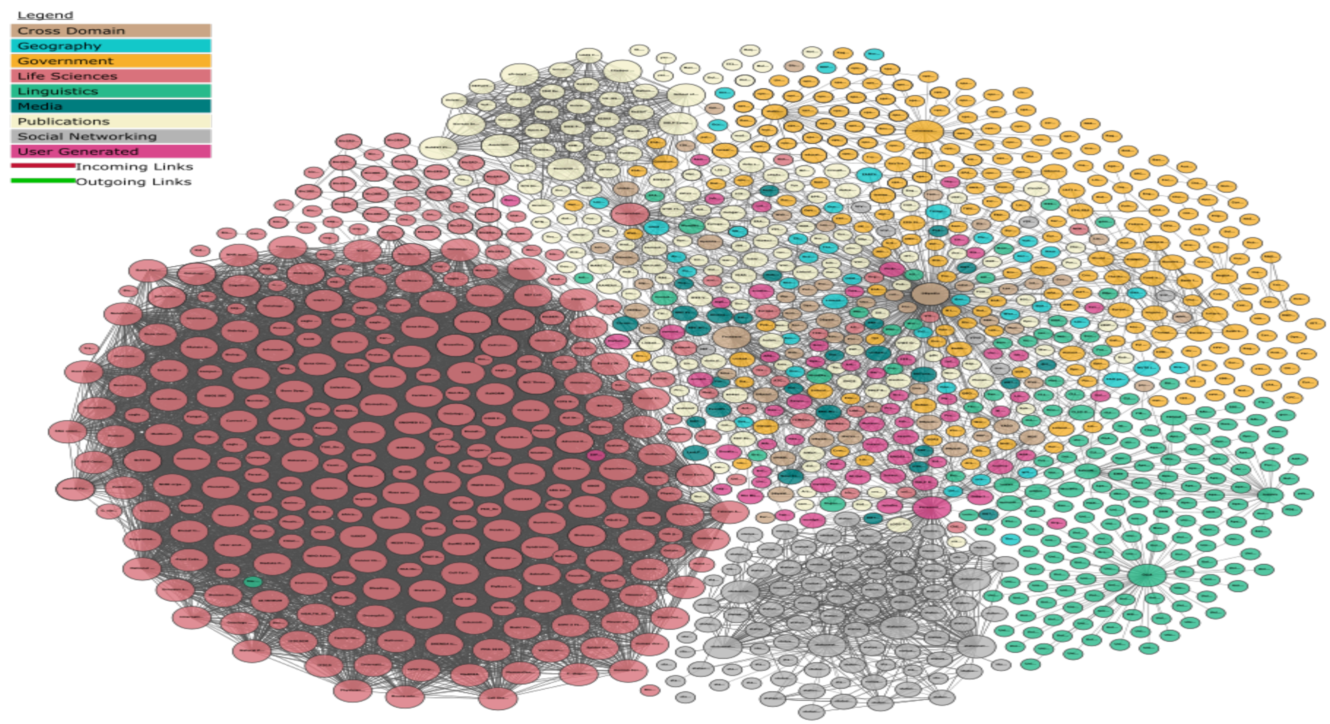
SEMANTIC LABELING

DSCI 558: Building Knowledge Graphs

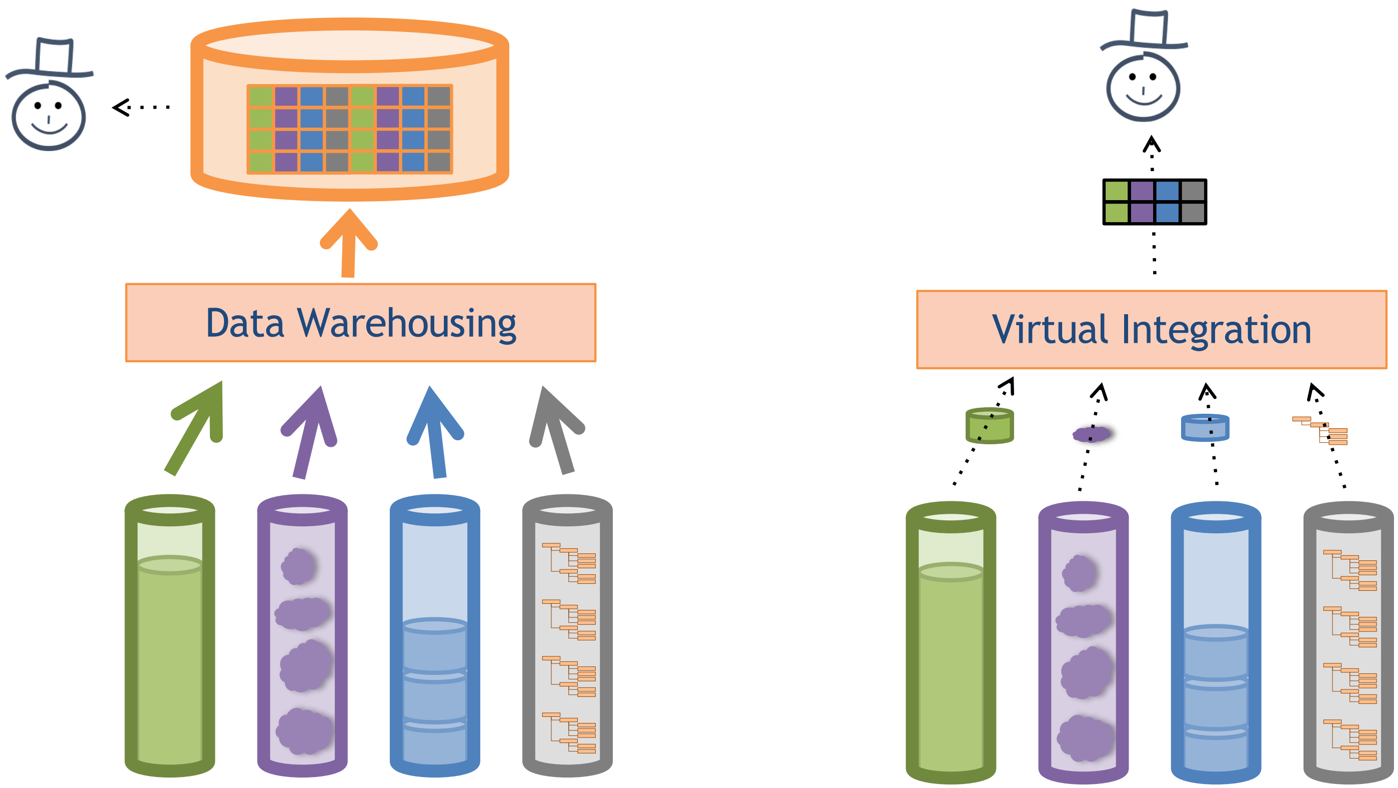
Craig Knoblock

Based on slides by Pedro Szekely, Minh Pham, & S.K. Ramnandan

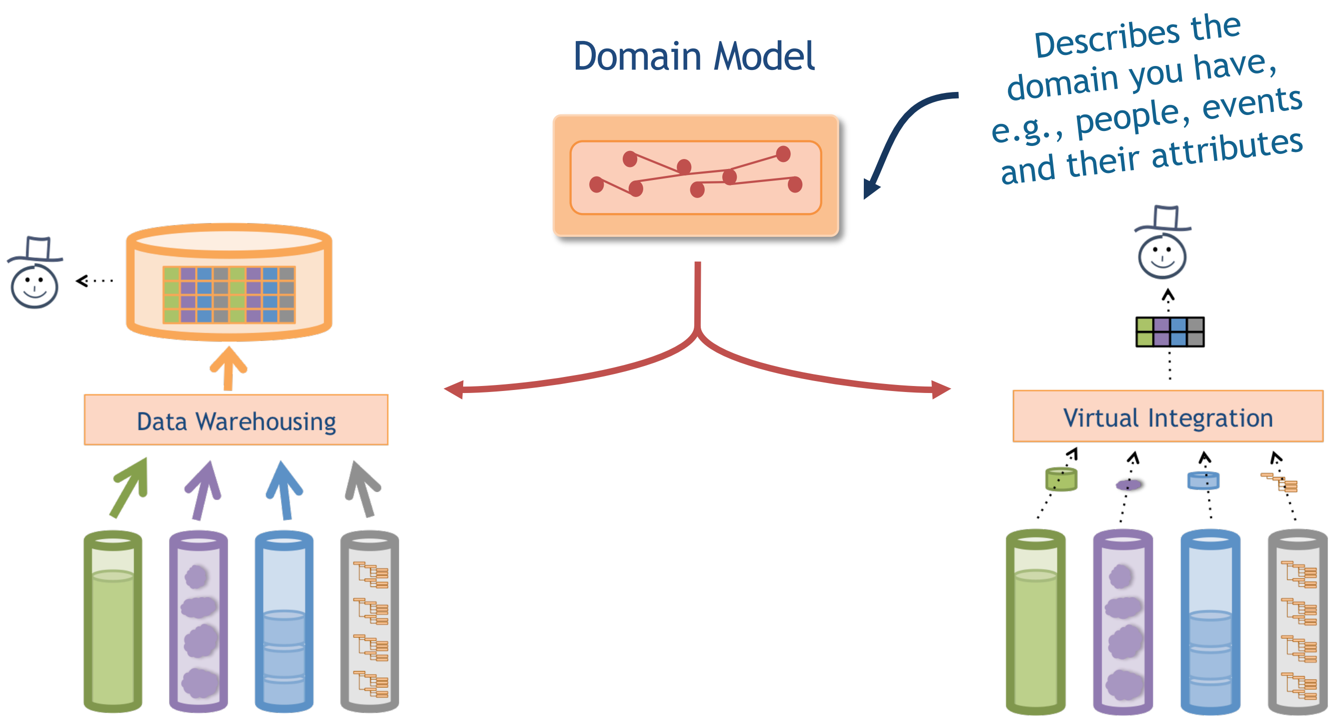
Introduction



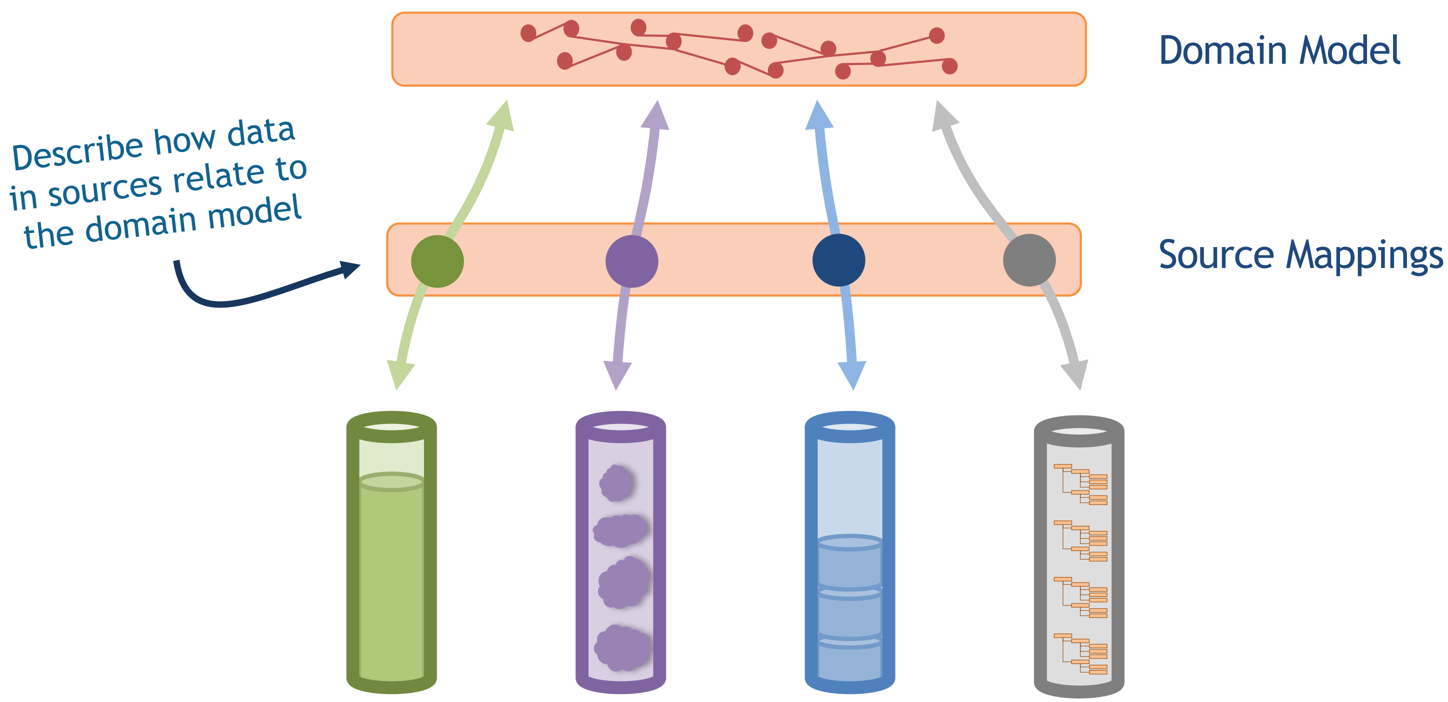
Data Integration Approaches



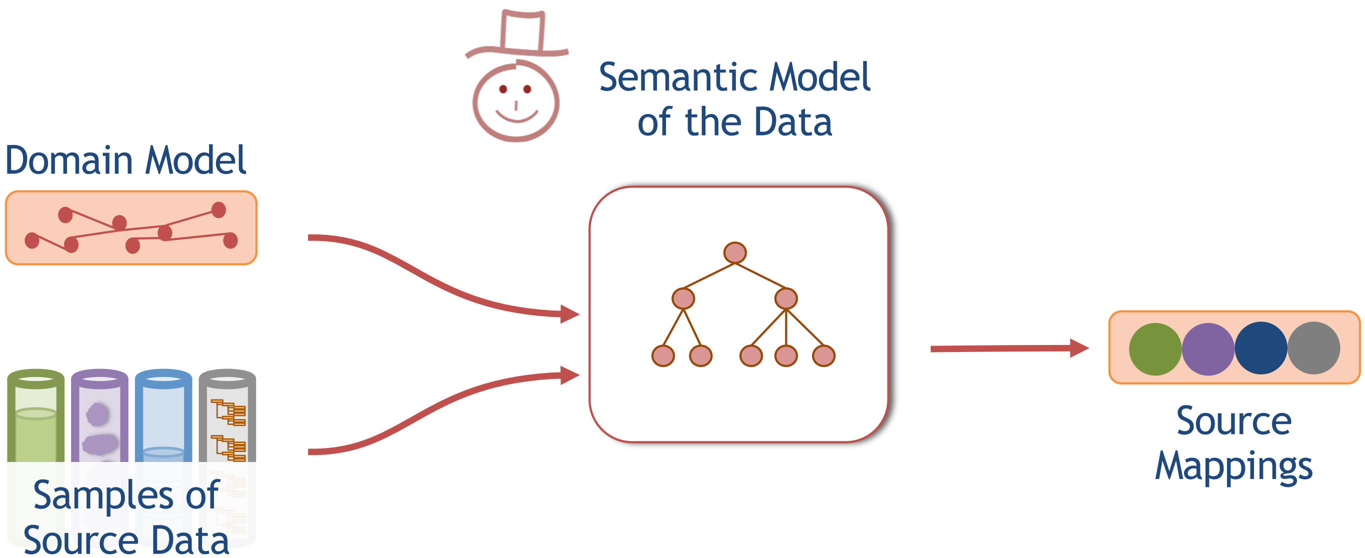
Domain Model



Key Ingredient: Source Mappings



Automatic Source Modeling

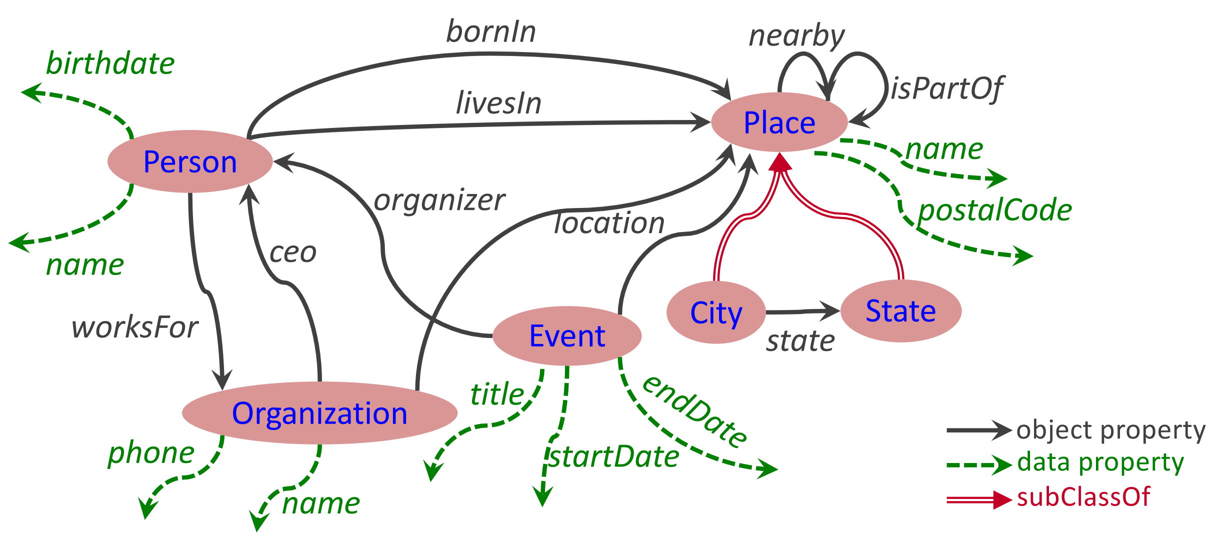


What is a Semantic Model?

Source

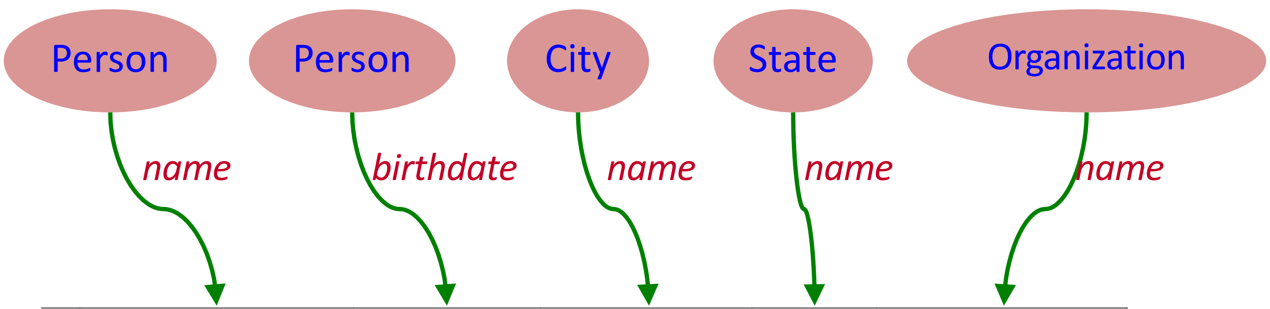
	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Domain Model



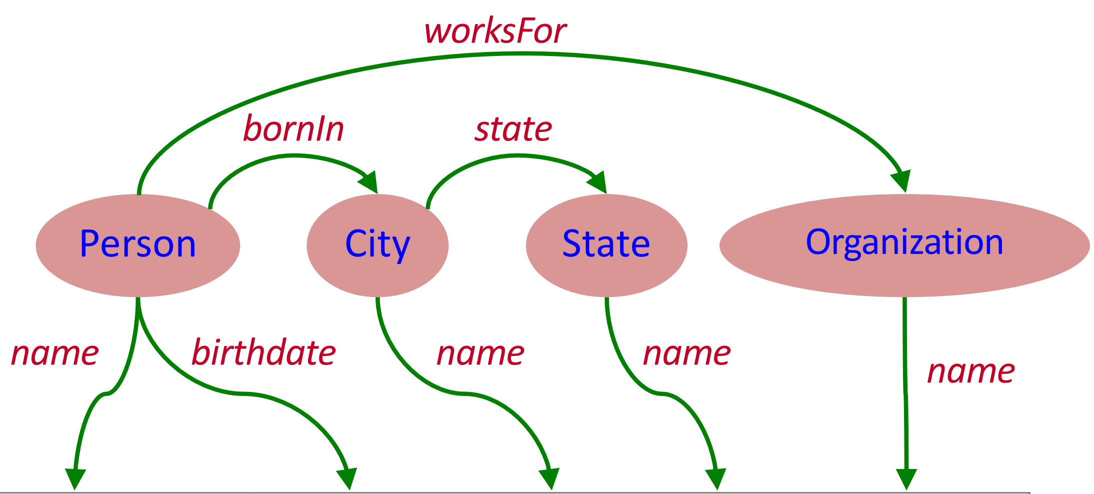
Describe sources using classes & relationships in an ontology

Semantic Types



	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Relationships

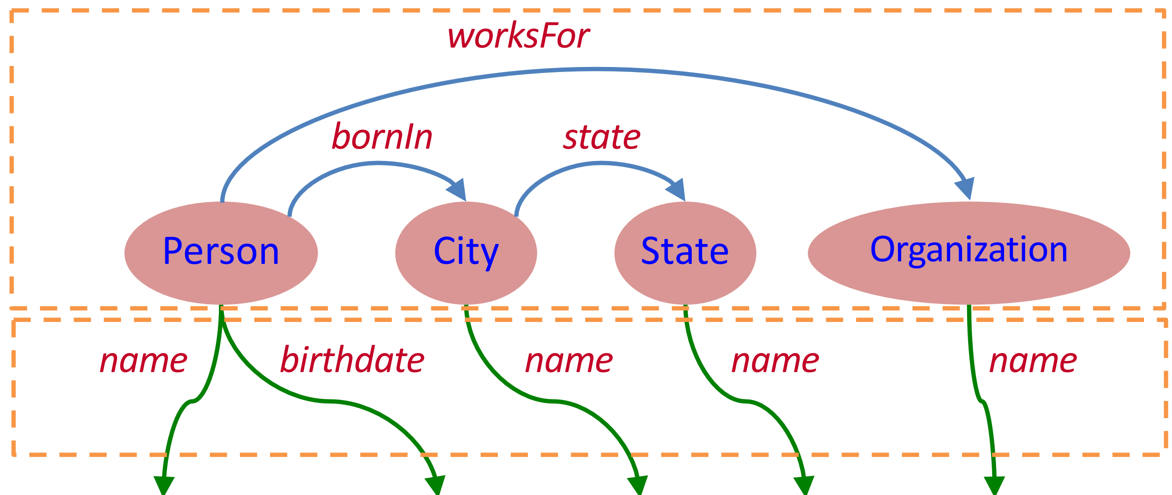


	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Source Modeling Problems

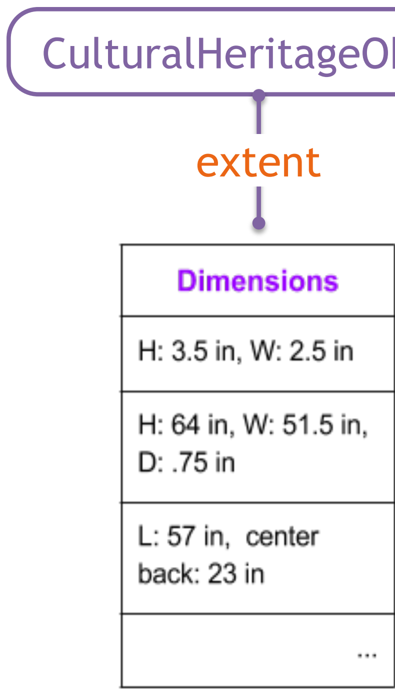
Source Modeling

Semantic Labeling



	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

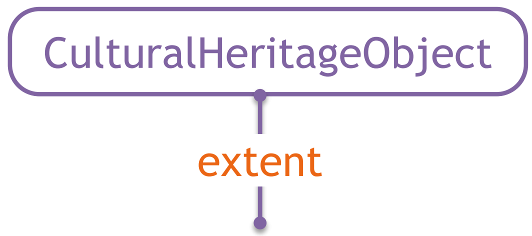
Learning Semantic Types



1- User specifies

2- System learns

Learning Semantic Types



Dimensions
H: 3.5 in, W: 2.5 in
H: 64 in, W: 51.5 in, D: .75 in
L: 57 in, center back: 23 in
...

Extent
52.1 x 71.4 cm (20 1/2 x 28 1/8 in.)
9 3/4 x 7 9/16 in.
H: 19 x W: 15 1/4 x D: 8 1/4 in.
...

Learning Semantic Types

CulturalHeritageObject

extent

Dimensions
H: 3.5 in, W: 2.5 in
H: 64 in, W: 51.5 in, D: .75 in
L: 57 in, center back: 23 in
...

CulturalHeritageObject

extent

Extent
52.1 x 71.4 cm (20 1/2 x 28 1/8 in.)
9 3/4 x 7 9/16 in.
H: 19 x W: 15 1/4 x D: 8 1/4 in.
...

Requirements

- Learn from a small number of examples
- Work on both textual and numeric values
- Learn quickly and highly scalable to large number of semantic types

RULE-BASED APPROACH

Assigning Semantic Labels to Data Sources

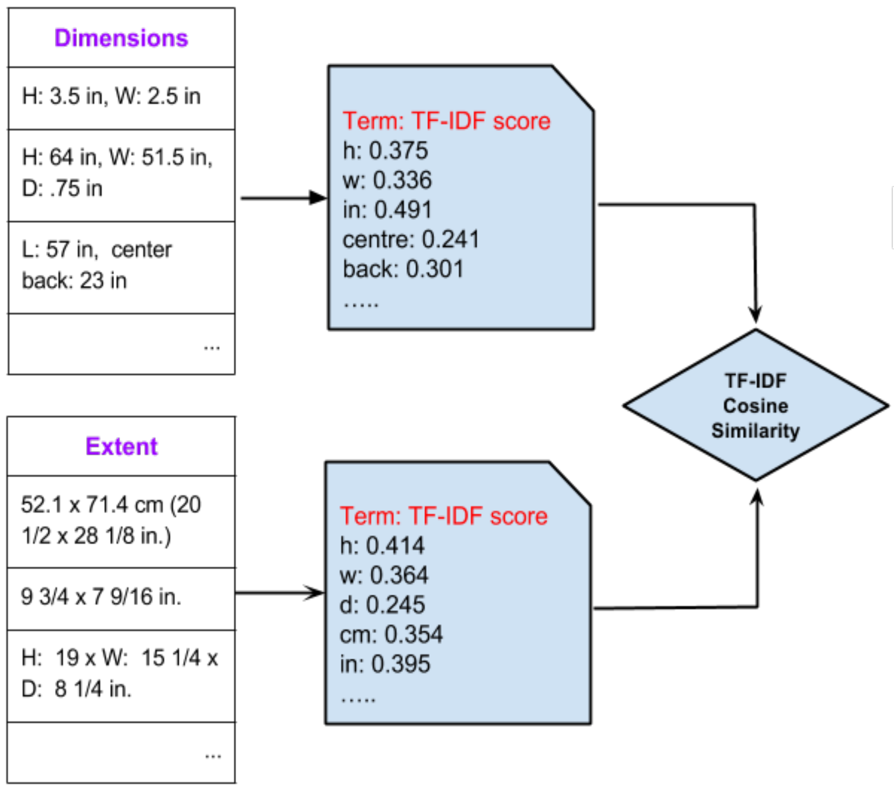
Ramnandan, S.K.; Mittal, A.; Knoblock, C. A.; and Szekely, P.

Approach for Textual Data

- Document: each column of data
- Label: each semantic type
- Use Apache Lucene to index the labeled documents
- Compute TF/IDF vectors for documents
- Compare documents using Cosine Similarity between TF/IDF vectors

Dimensions
H: 3.5 in, W: 2.5 in
H: 64 in, W: 51.5 in, D: .75 in
L: 57 in, center back: 23 in
...

Approach for Textual Data



$$tf(t,d) = frequency^{1/2}$$

$$idf(t) = 1 + log(\frac{numDocs}{docFreq + 1})$$

$$sim(q,d) = \frac{V(q).V(d)}{|V(q)|.|V(d)|}$$

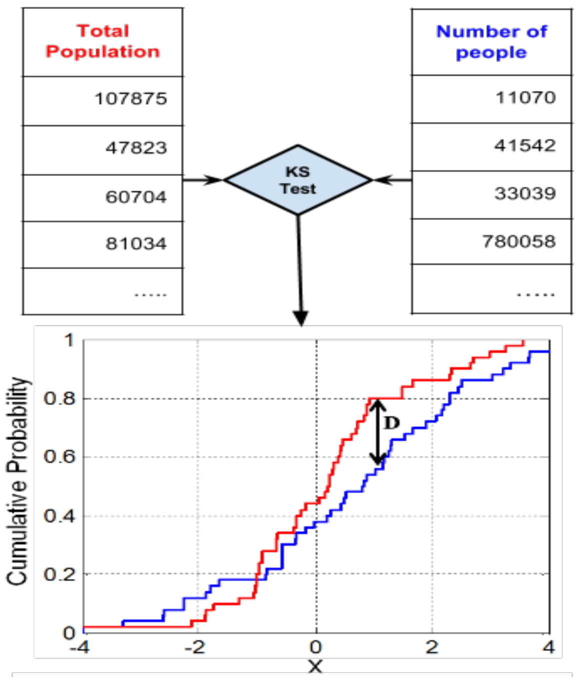
Approach for Numeric Data

- Distribution of values in different semantic types is different, e.g., temperature vs. population
- Use Statistical Hypothesis Testing to see which distribution fits best
- Welch's T-test, Mann-Whitney U-test and Kolmogorov-Smirnov Test

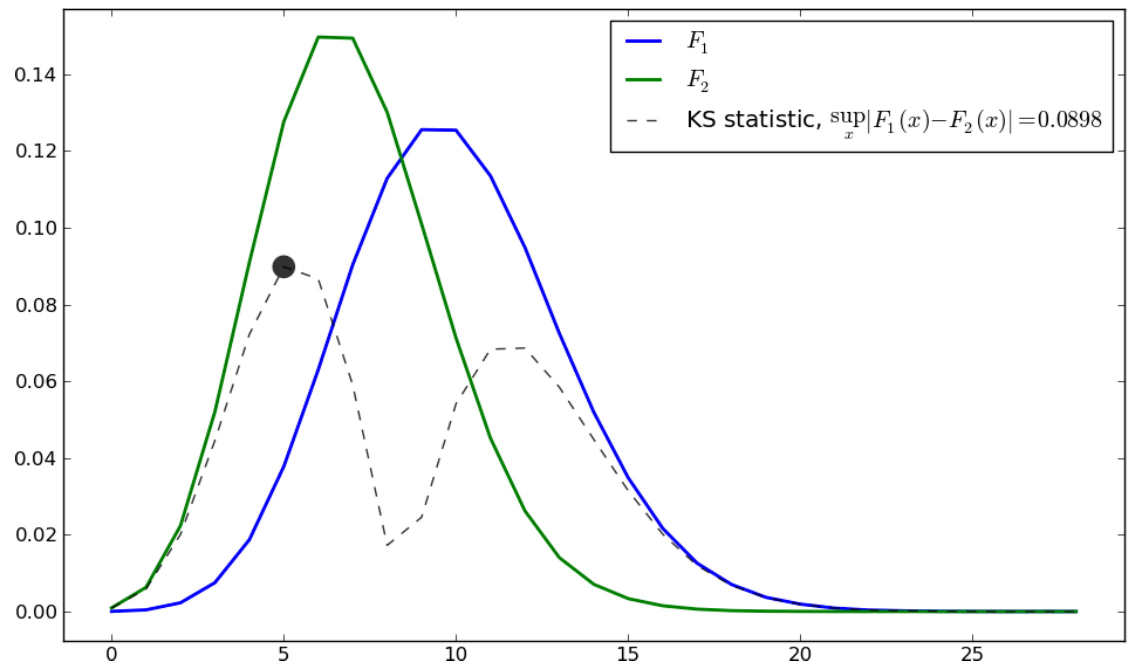
Total Population
107875
47823
60704
81034
.....

Number of people
11070
41542
33039
780058
.....

Approach for Numeric Data



$$D_{N_1,N_2} = \sup_x |F_{1,N_1}(x) - F_{2,N_2}(x)|$$

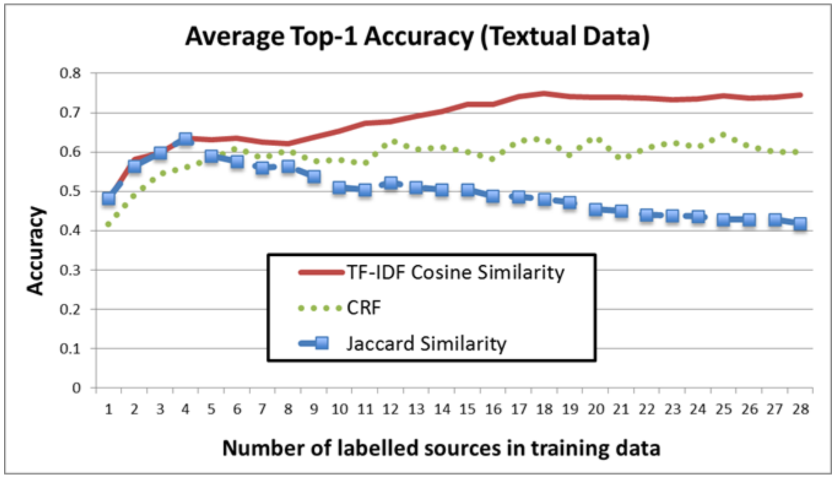


Combined Approach

- Training
  - Add new example data as training for either textual or numeric types
  - If ambiguous, train as both textual and numeric
- Testing
  - If textual, apply tf/idf
  - If numeric apply KS-test
  - If ambiguous and at least 70% numeric apply KS-test, otherwise tf/idf

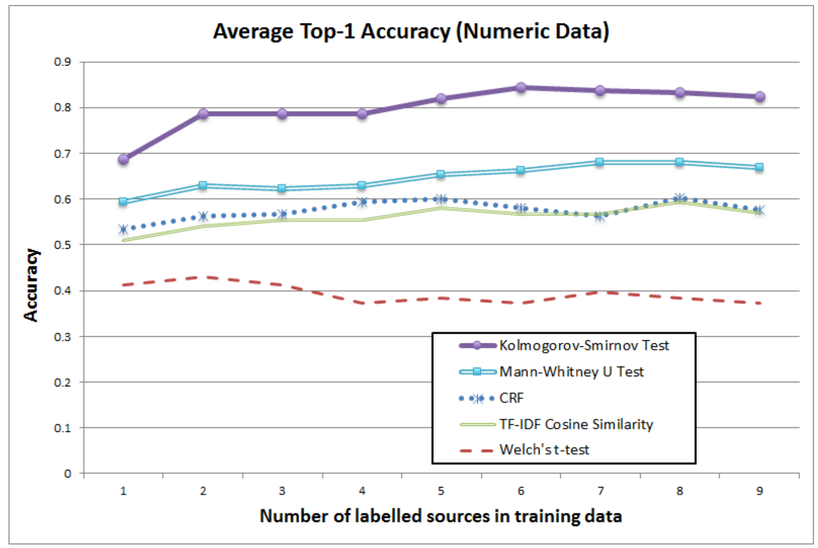
Return Top-k suggestions based on the confidence scores

Evaluation of Semantic Typing



Combined approach achieves 97% accuracy on the top-4 accuracy

Reduced the training time from 110s to 0.45s

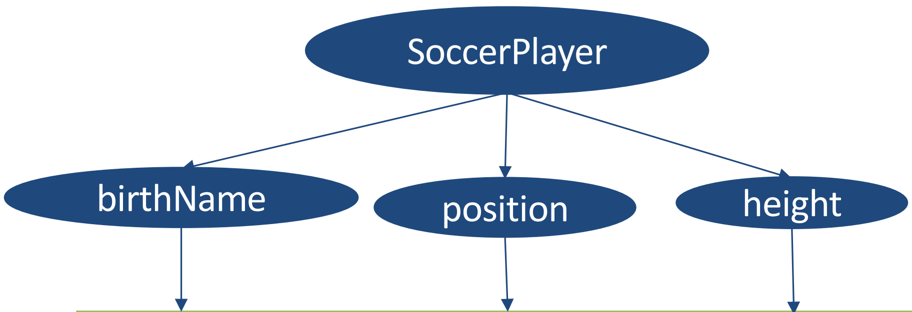


LEARNING-BASED APPROACH: CLASSIFICATION

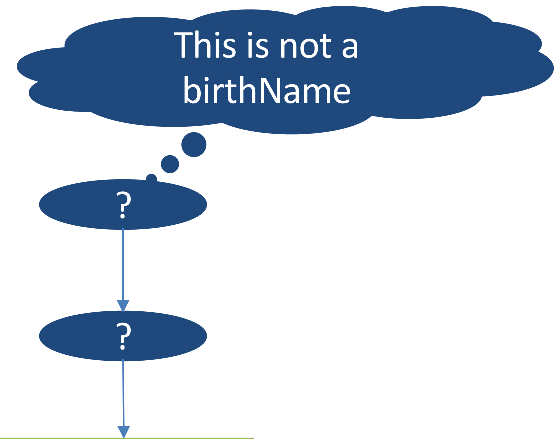
Semantic labeling: a domain-independent approach

Minh Pham, Suresh Alse, Craig Knoblock, Pedro Szekely

General idea

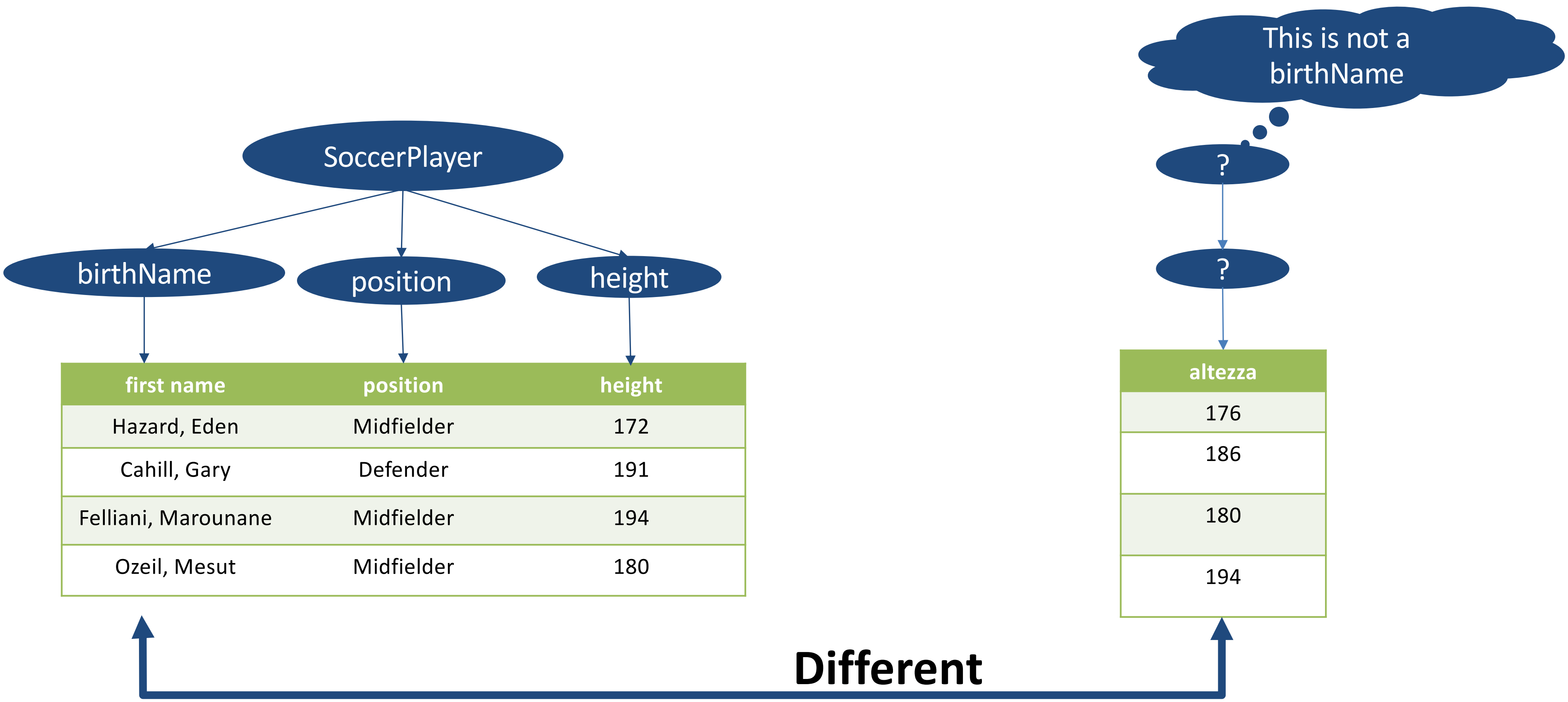


first name	position	height
Hazard, Eden	Midfielder	172
Cahill, Gary	Defender	191
Felliani, Marounane	Midfielder	194
Ozeil, Mesut	Midfielder	180



height
176
186
180
194

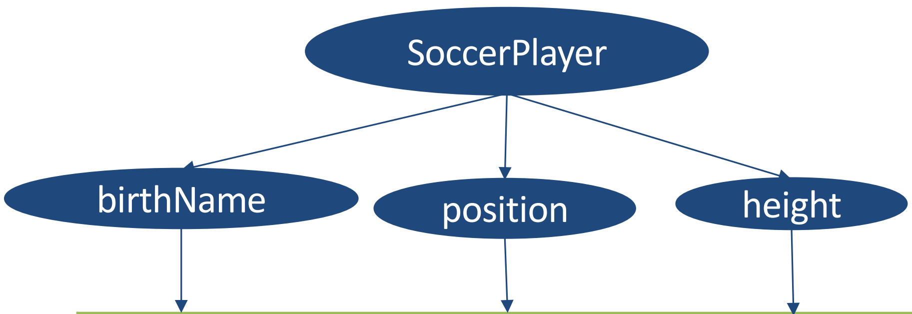
General idea



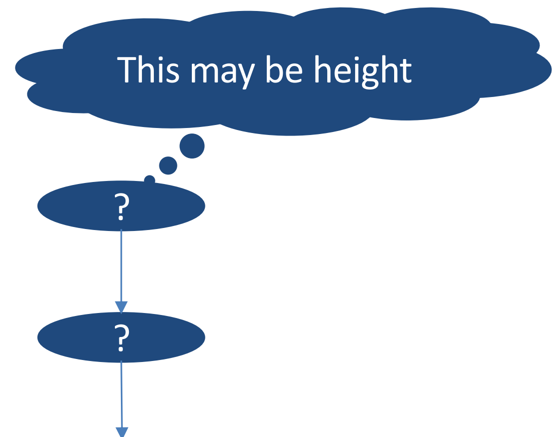
first name	position	height
Hazard, Eden	Midfielder	172
Cahill, Gary	Defender	191
Felliani, Marounane	Midfielder	194
Ozeil, Mesut	Midfielder	180

altezza
176
186
180
194

General idea



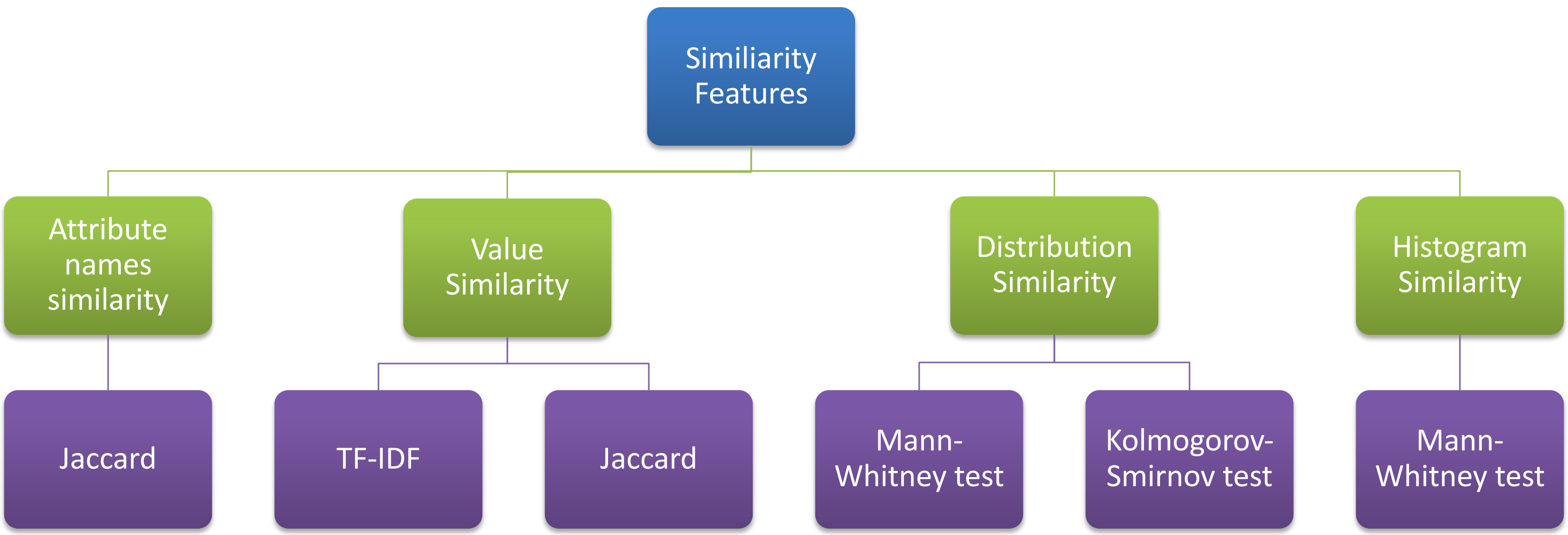
first name	position	height
Hazard, Eden	Midfielder	172
Cahill, Gary	Defender	191
Felliani, Marounane	Midfielder	194
Ozeil, Mesut	Midfielder	180



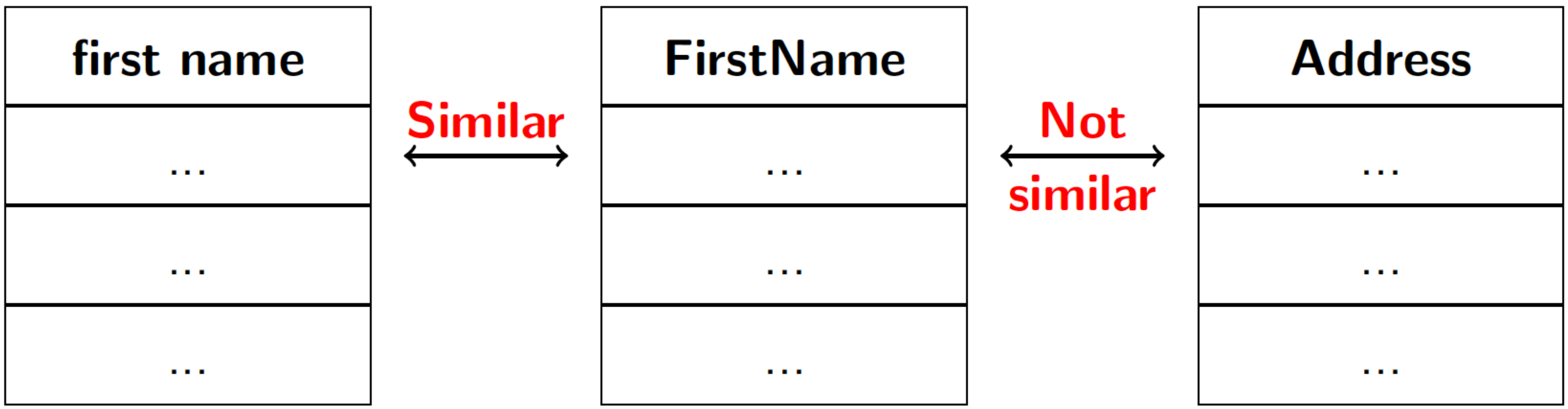
altezza
176
186
180
194

Similar

Similarity features



Attribute name similarity



Value similarity

Player name
Gary Cahill
Metsul Ozeil
Juan Mata

Similar ↔

Name
Juan Quin
De Gea
Tim Cahill

Not similar ↔

Club name
Chelsea
Real Madrid
Barcelona

Value similarity

game played
4
…
18
23

goal scored
3
…
11
22

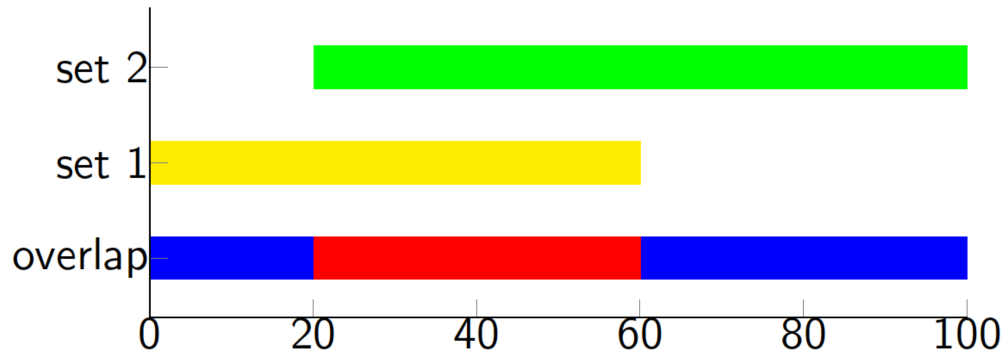
Overlapping values is not enough

Value range similarity

Numeric Jaccard Similiarity

Given 2 numeric sets of values A, B ranged in $[a_s, a_e]$ and $[b_s, b_e]$:

$$numJaccardSim(A, B) = \frac{|[a_s, a_e] \cap [b_s, b_e]|}{|[a_s, a_e] \cup [b_s, b_e]|}$$



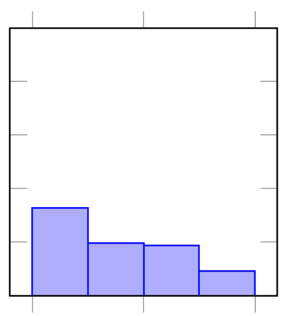
Distribution similarity

game played
4
...
18
23

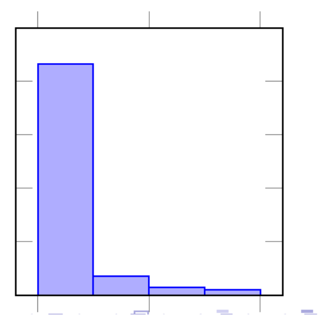
Similar

goal scored
3
...
11
22

Similarity measure: KS test



Not similar



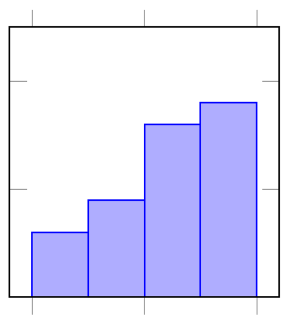
Histogram similarity

position
1
4
2
4

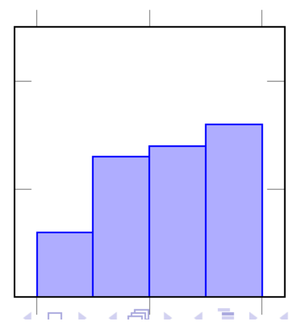
Not similar

ps
GK
MF
DF
FW

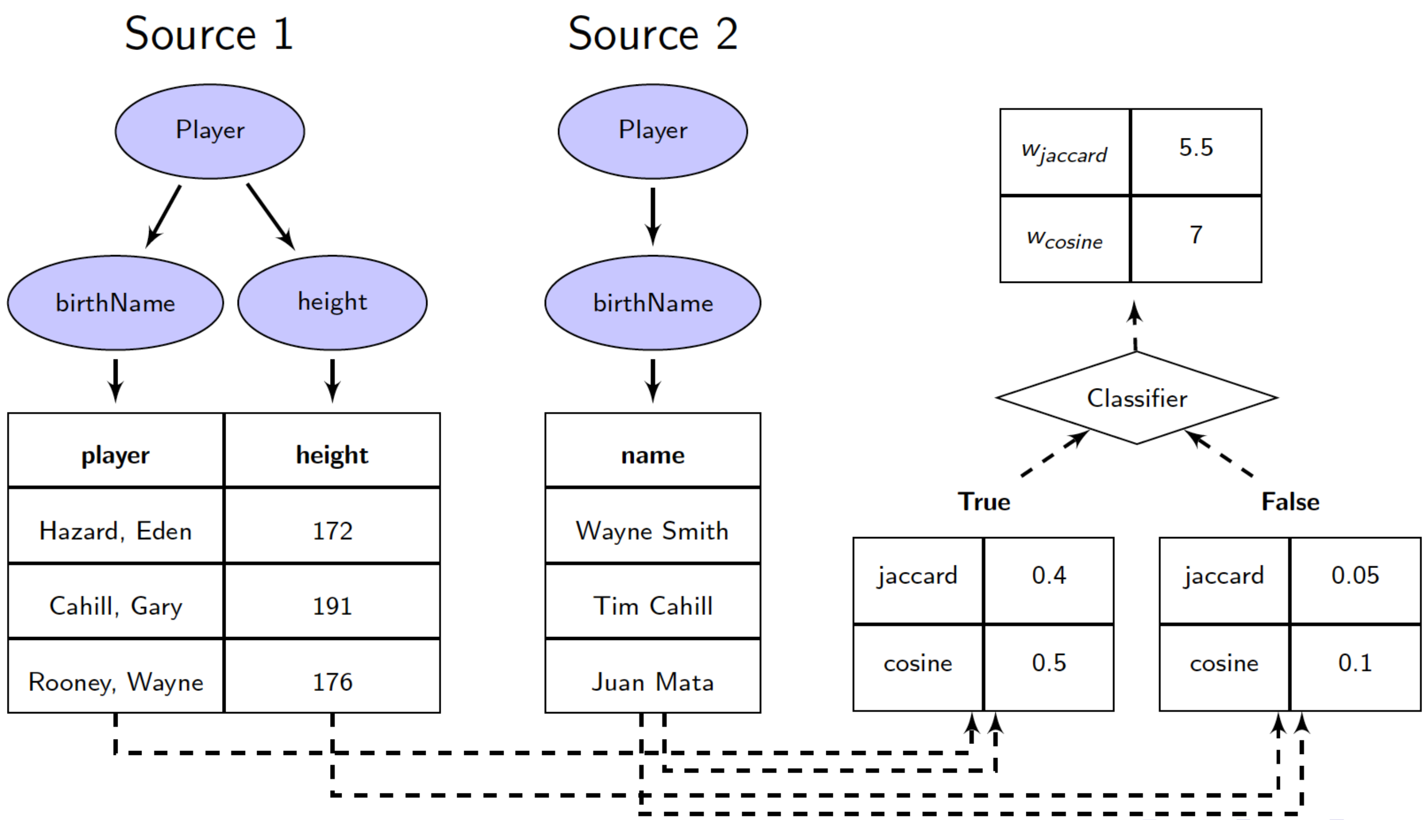
↓ Similarity measure: MW test ↓



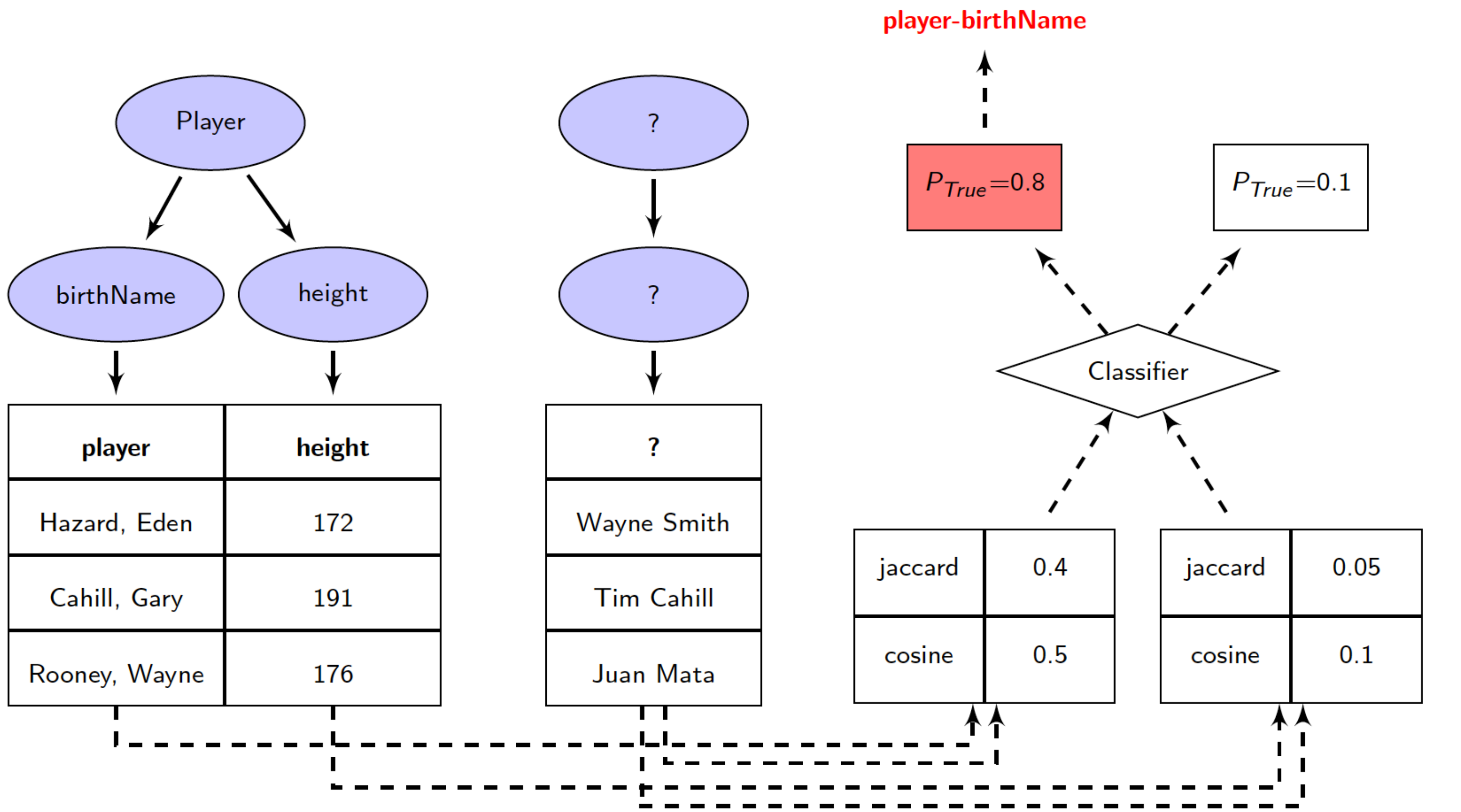
Similar



Training machine learning model



Predicting new attribute



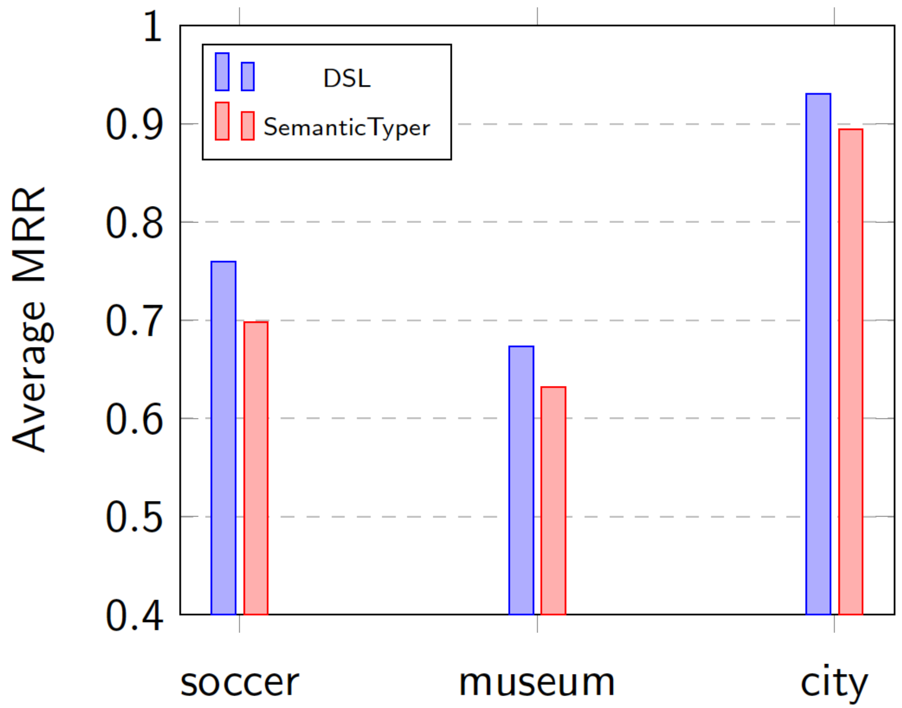
Evaluation

Data sets:

Domain data	# sources	# semantic types	# attributes
soccer	12	14	97
museum	29	20	217
city	10	52	520
weather	4	11	44
T2D Gold	1748	7983	?

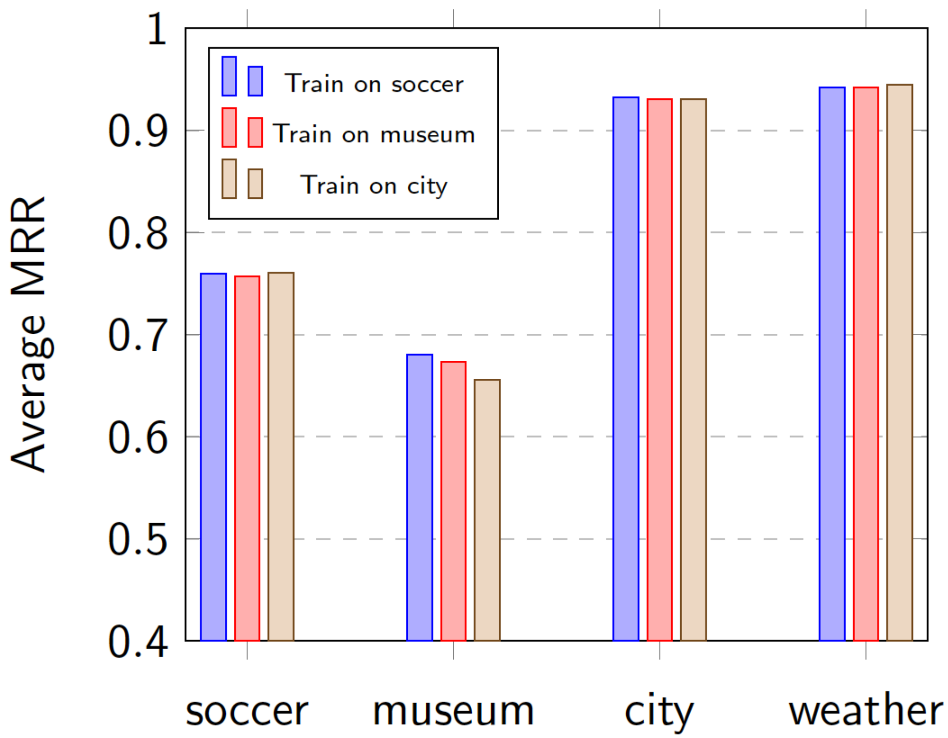
Measurements: Mean Reciprocal Rank (MRR)
Evaluating systems: DSL (our approach), SemanticTyper (Ramnandan et al, 2015), T2K (Ritze et al, 2015)

Evaluation



1
0.9
0.8
0.7
0.6
0.5
0.4
Average MRR
DSL
SemanticTyper
soccer
museum
city

Evaluation


Train on soccer
Train on museum
Train on city
Average MRR
1
0.9
0.8
0.7
0.6
0.5
0.4
soccer
museum
city
weather

Conclusion

	Rule-based Approach	Learning-based Classification
Pros	+ fast + scalable + easy to implement	+ fast + scalable + easy to extend + works in lots of domains
Cons	+ requires heuristics + difficult to extend + may not work in every domain	+ no use of relationship + need to train machine learning on a general domain