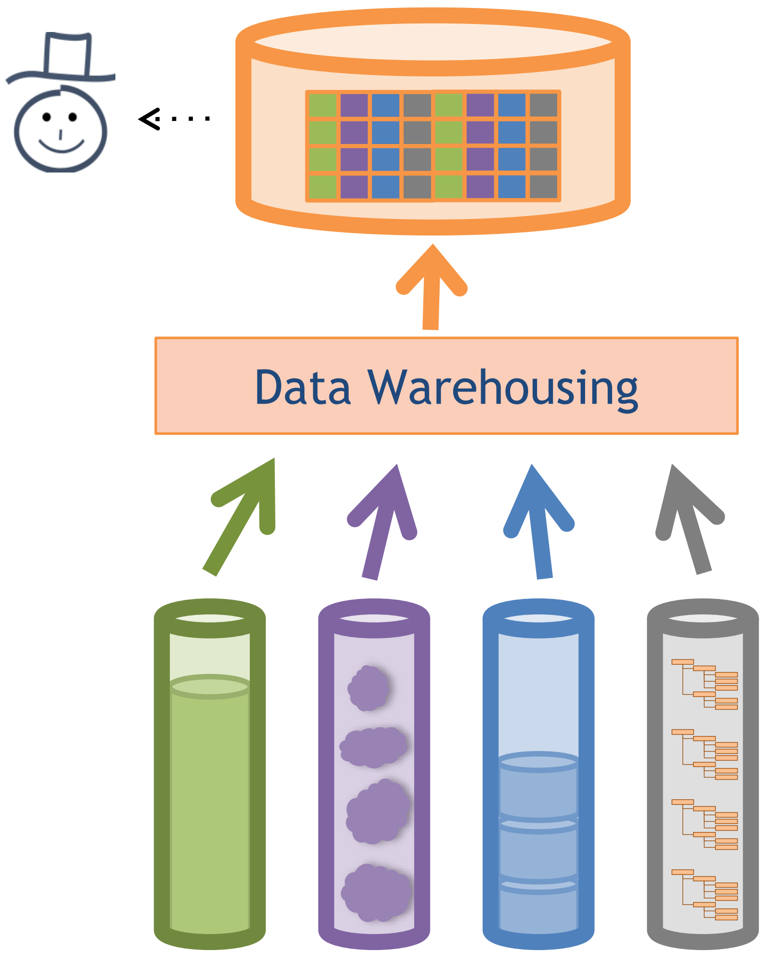
Aligning and Integrating Data in Karma

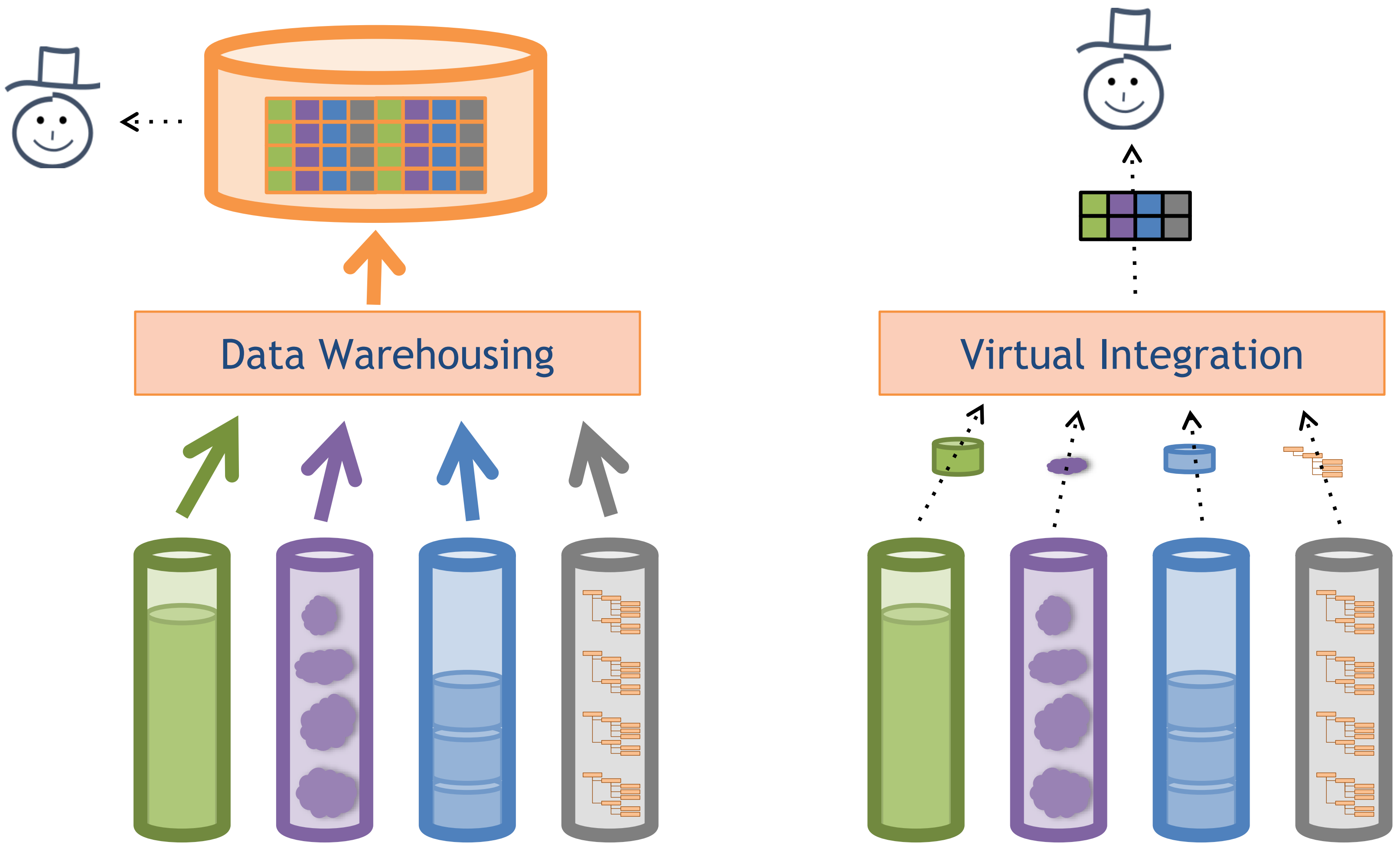
Craig Knoblock
University of Southern California

Data Integration Approaches

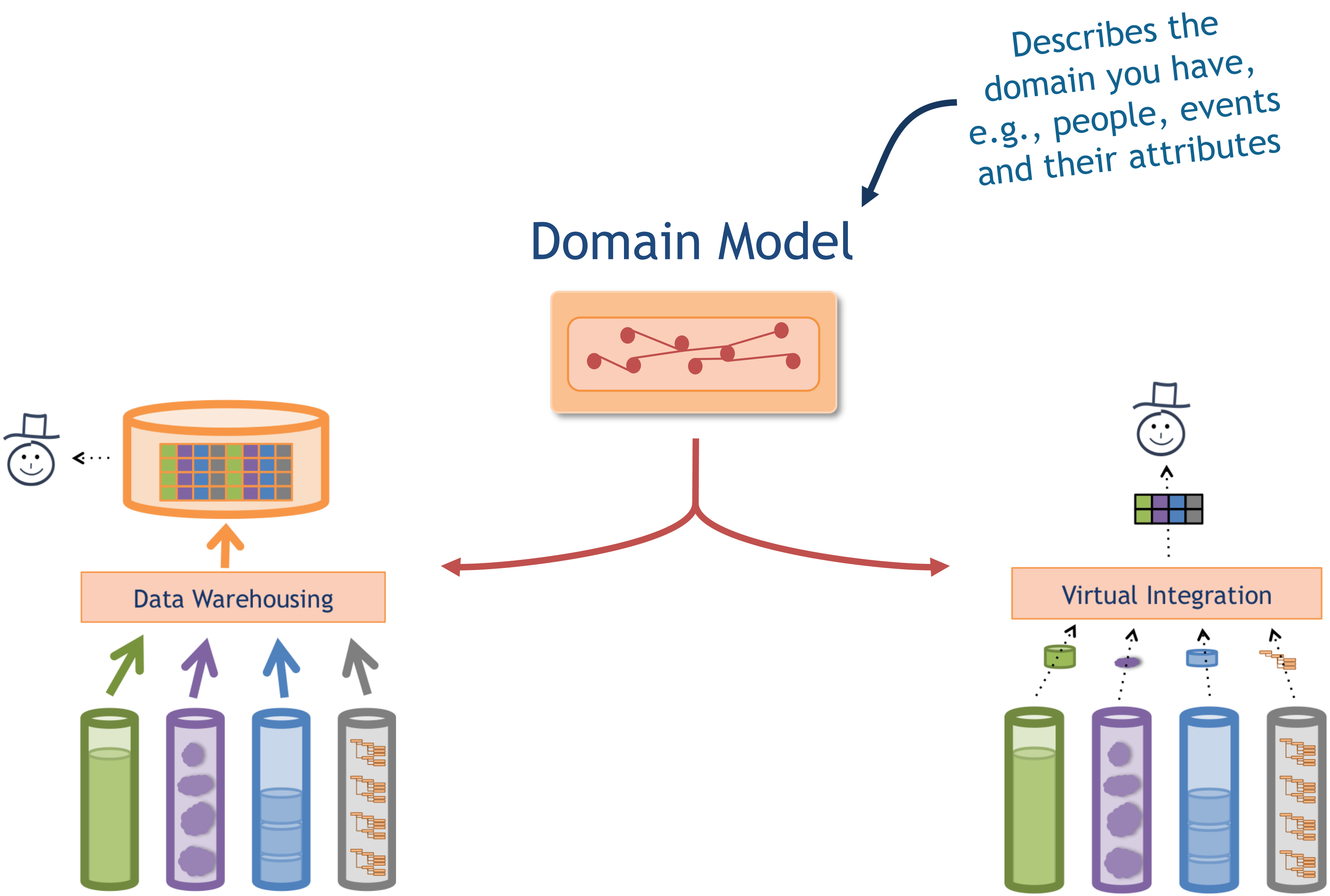
Data Integration Approaches



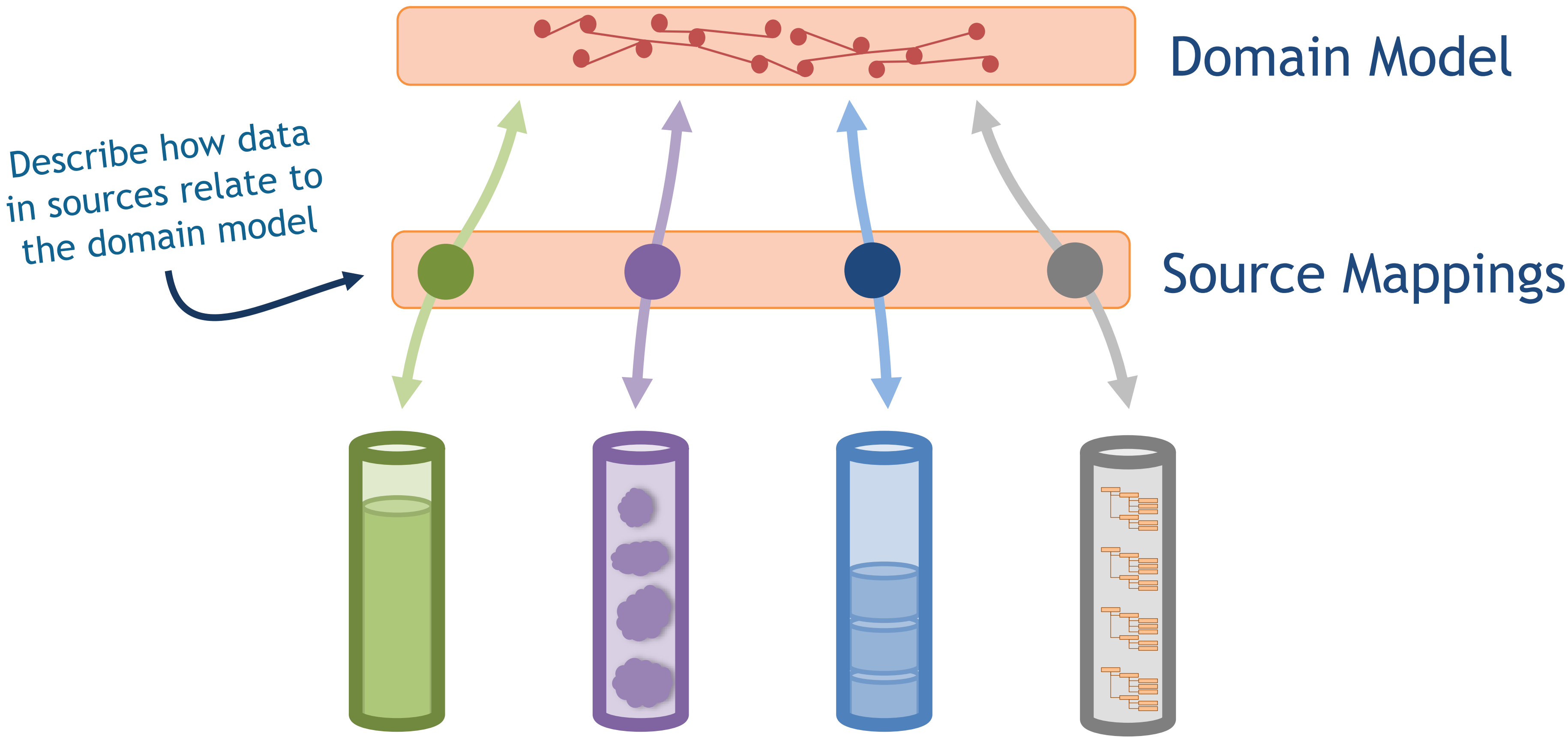
Data Integration Approaches



Domain Model



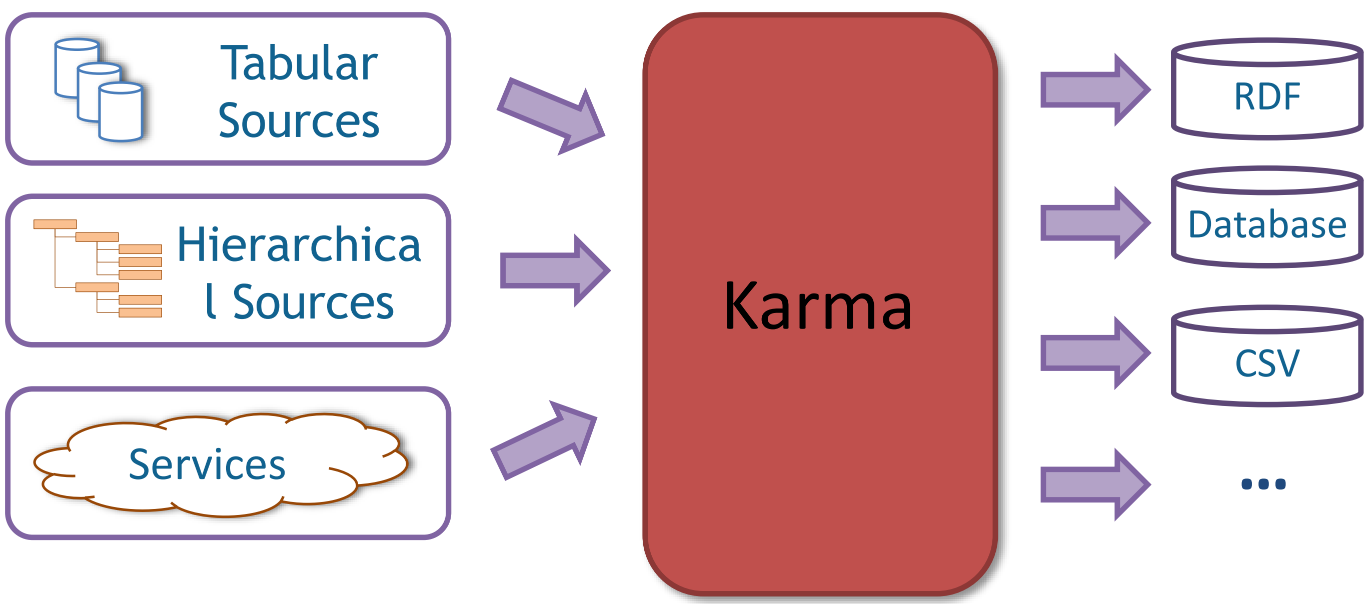
Key Ingredient: Source Mappings



Karma: A Data Integration Tool

Karma

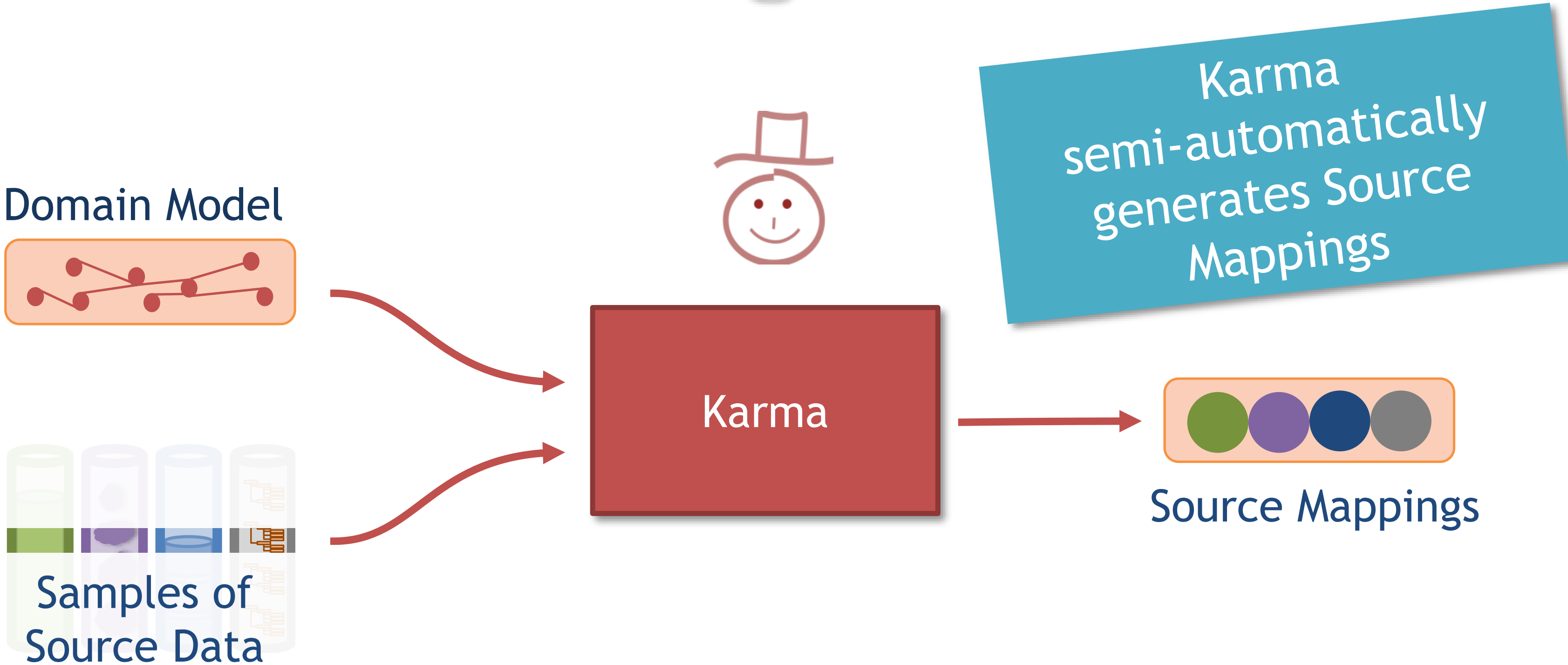
Interactive tool for rapidly extracting, cleaning, transforming, integrating and publishing data



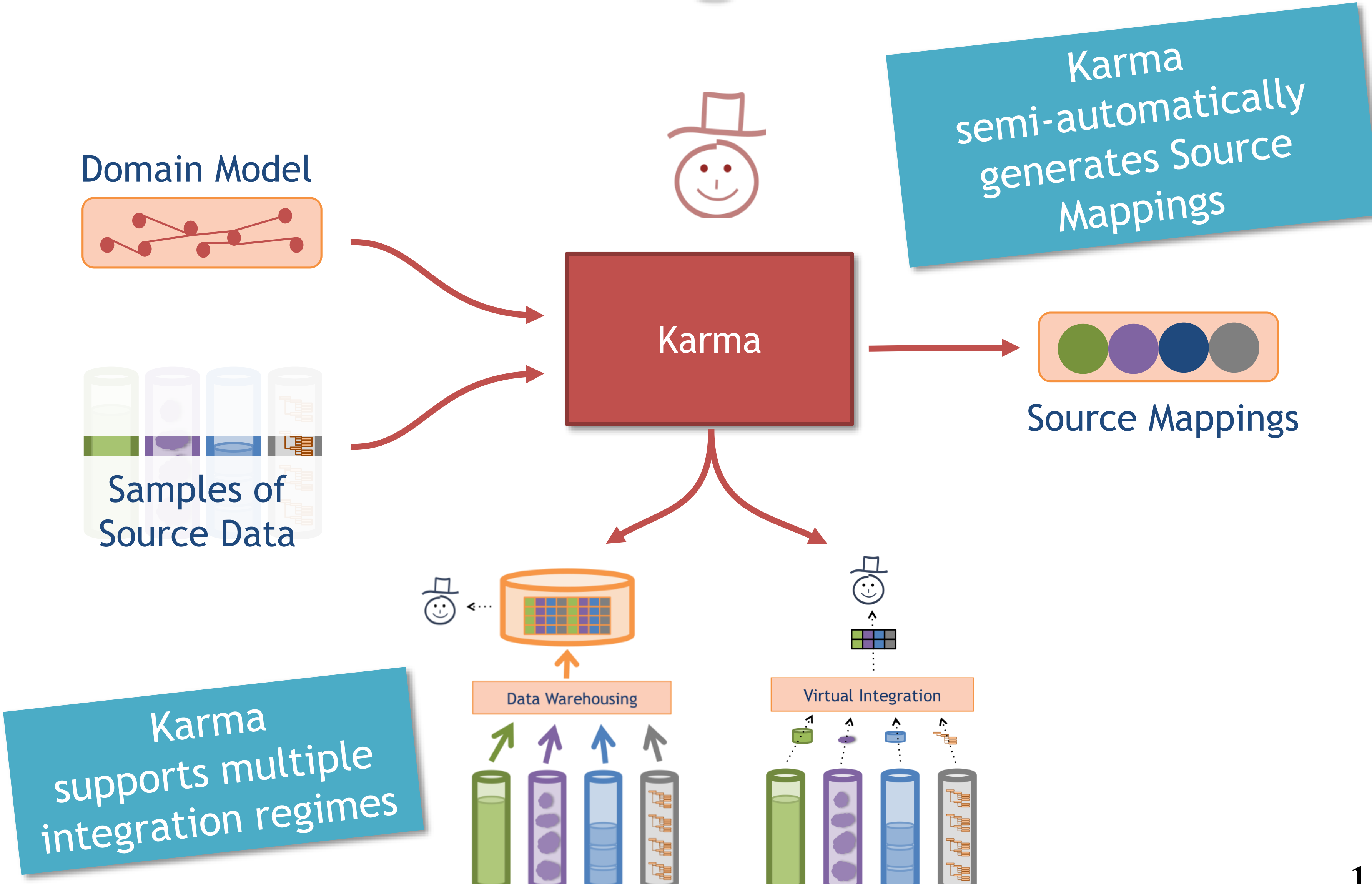
http://www.isi.edu/integration/karma

@KarmaSemWeb

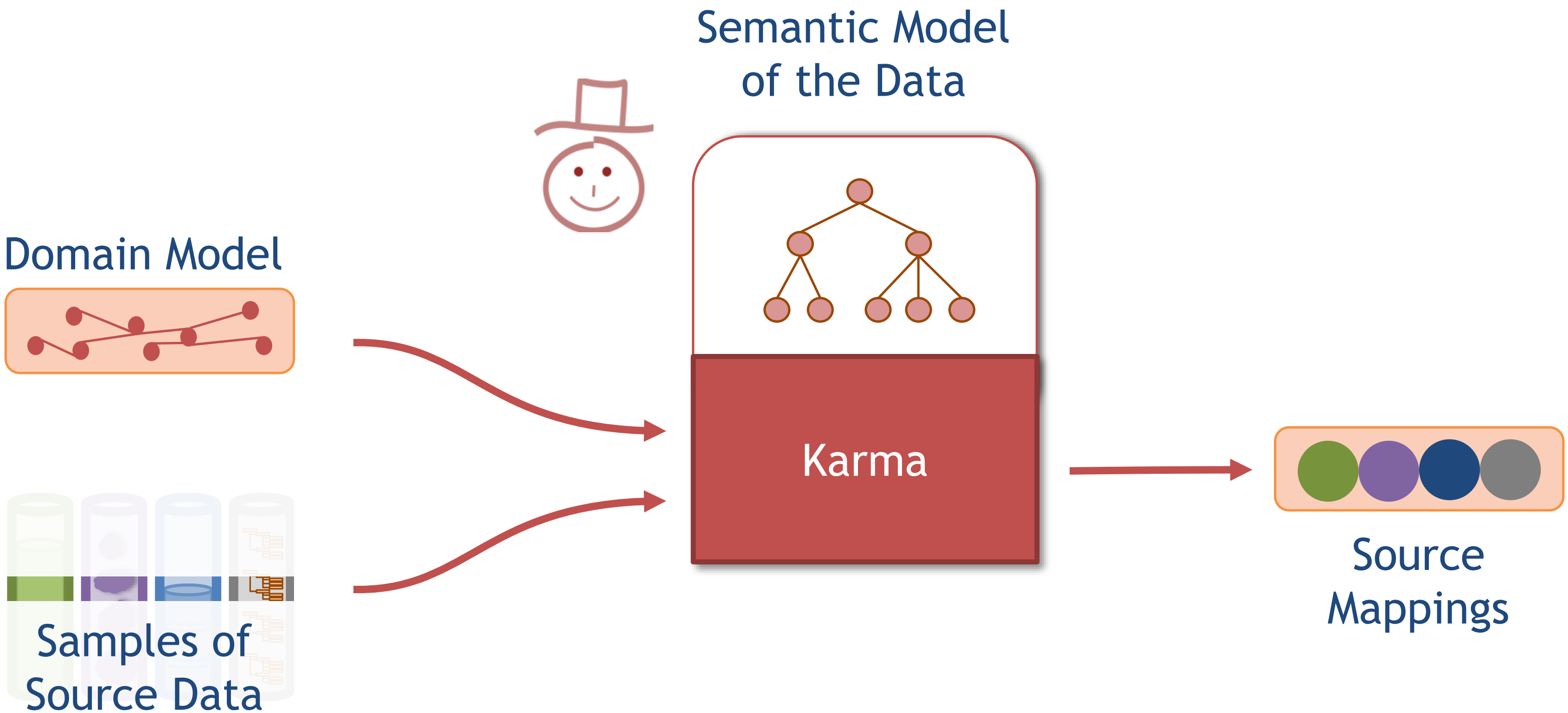
Information Integration in Karma



Information Integration in Karma



Secret Sauce: Karma Understands Your Data



Karma semi-automatically builds a semantic model of your data

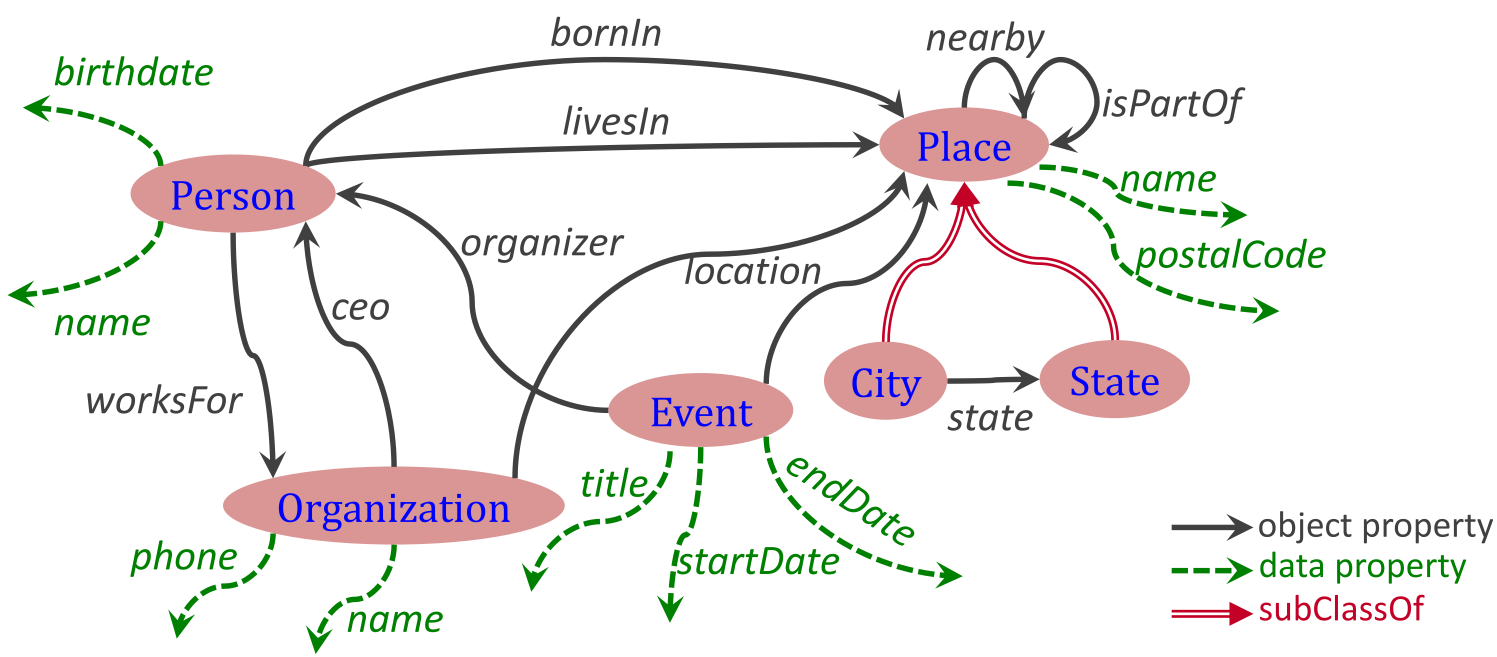
What is a Semantic Model?

Describe sources using classes & relationships in an ontology

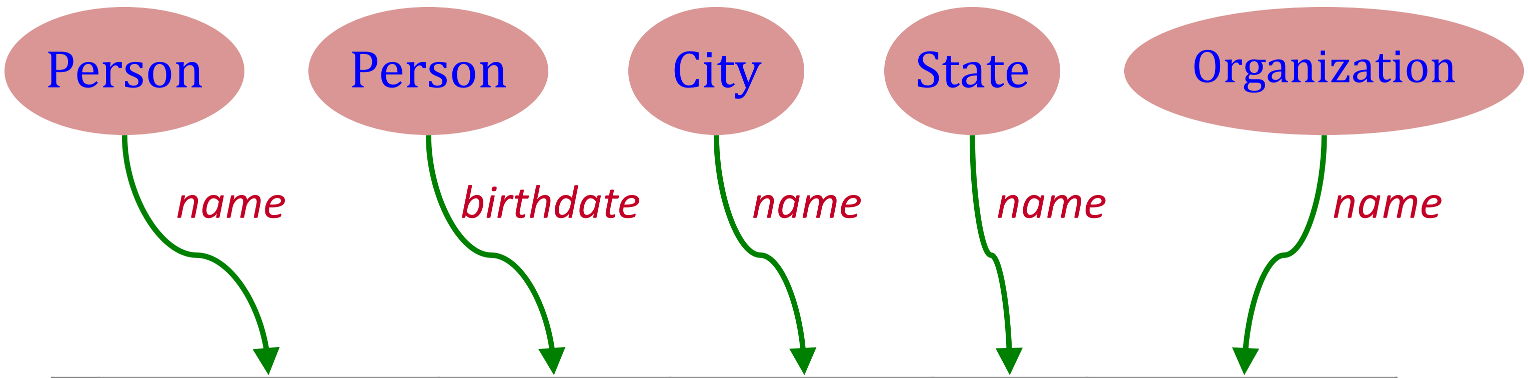
Source

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

Domain Model

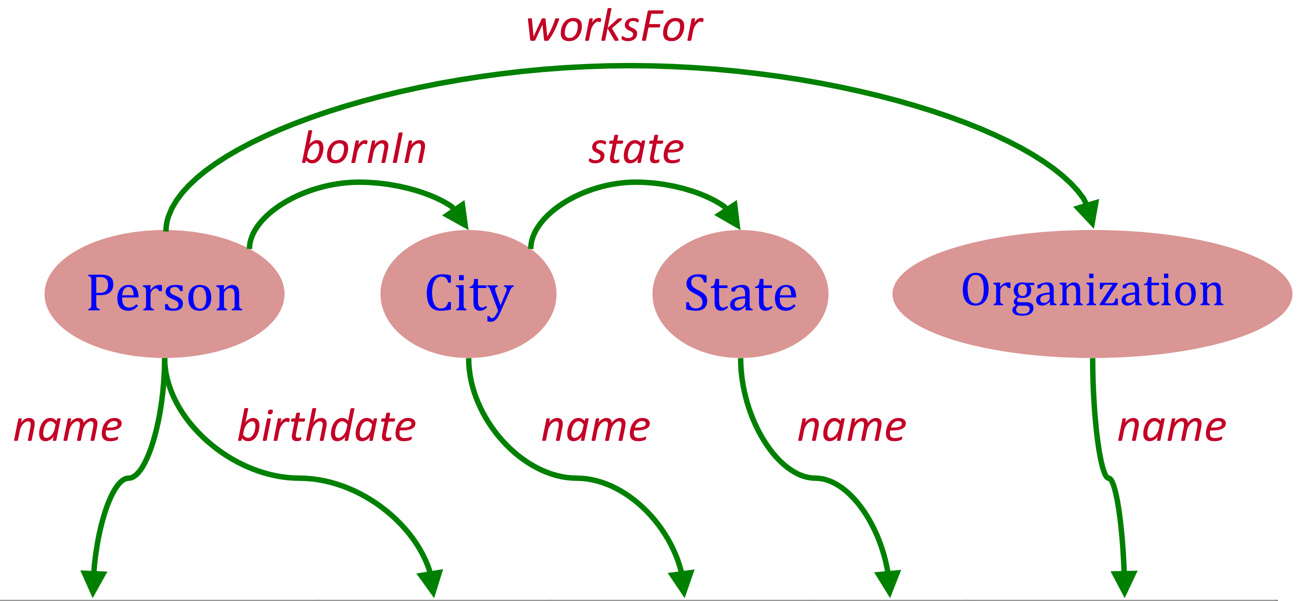


Semantic Types



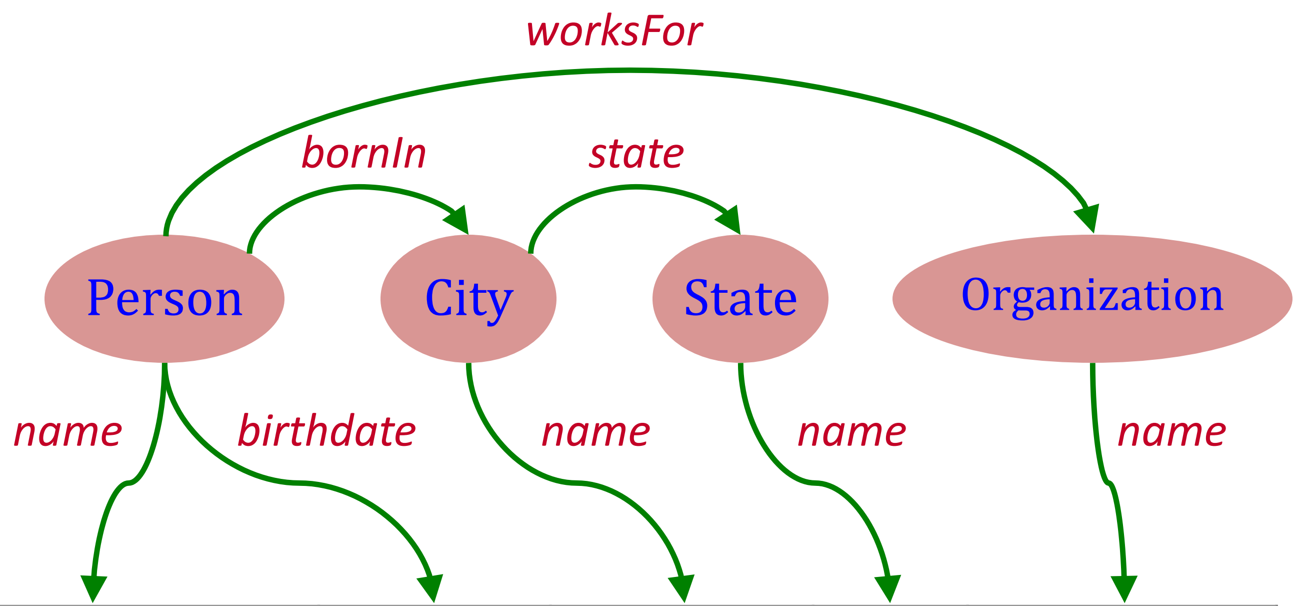
	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Relationships



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

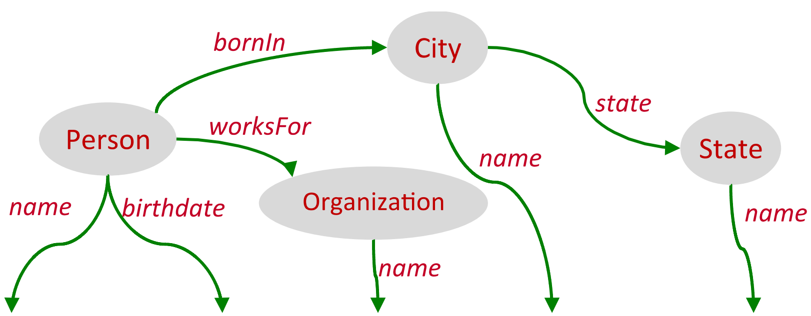
Semantic Model



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

Semantic models will be formalized as Source Mappings

Key ingredient to automate <u>source discovery</u>, <u>data integration</u>, and <u>publishing semantic data</u> (RDF triples) so what?

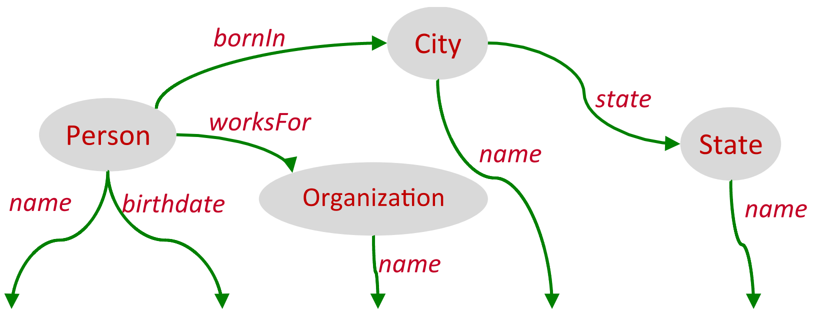


Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

Knowledge Graphs

Karma uses semantic models to create knowledge grap

Karma semi-automatically builds semantic models



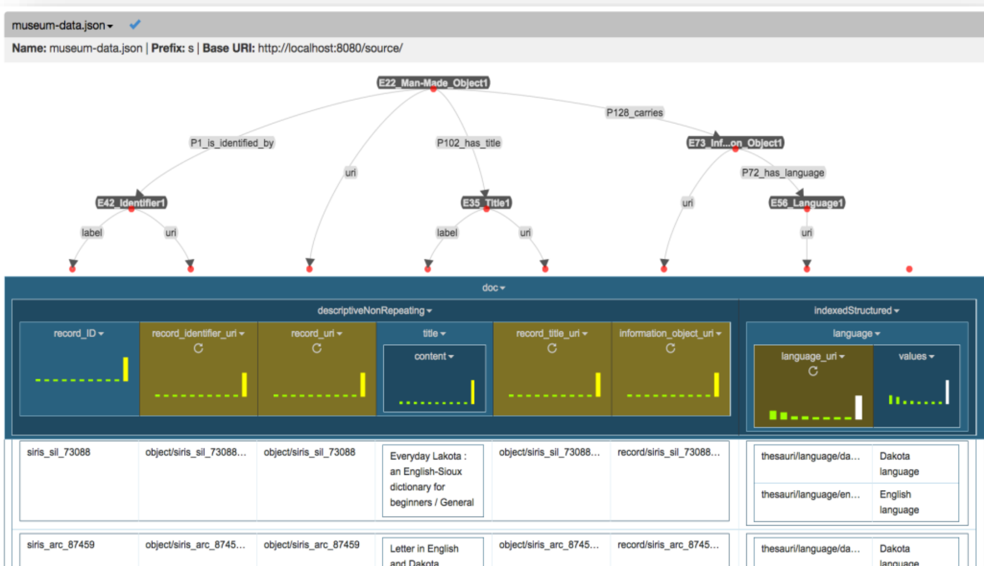
Column 1	Column 2	Column 3	Column 4	Column 5
Bill Gates	Oct 1955	Microsoft	Seattle	WA
Mark Zuckerberg	May 1984	Facebook	White Plains	NY
Larry Page	Mar 1973	Google	East Lansing	MI

Knowledge Graphs

Karma uses semantic models to create knowledge grap

Karma semi-automatically builds semantic models

… and provides a nice GUI to edit them

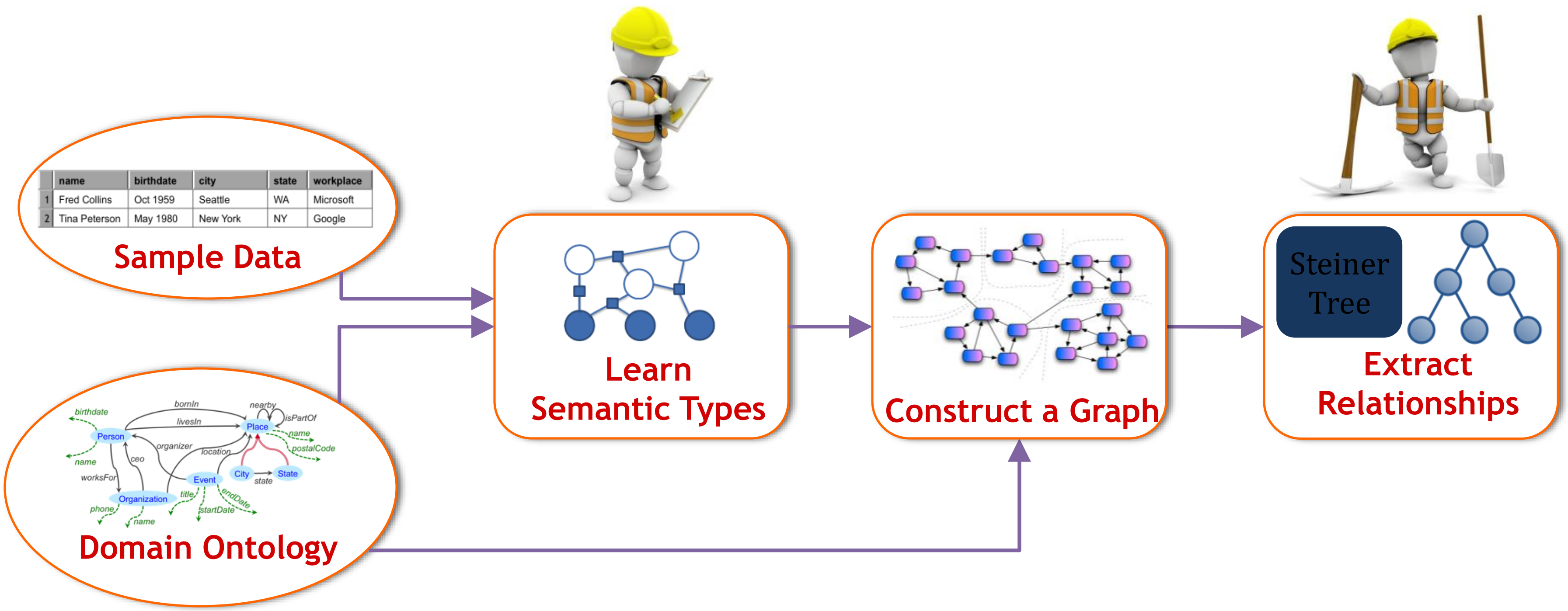


Knowledge Graphs

Karma uses semantic models to create knowledge grap

Semi-automatically Building Semantic Models in Karma

Approach
[Knoblock et al, ESWC 2012]

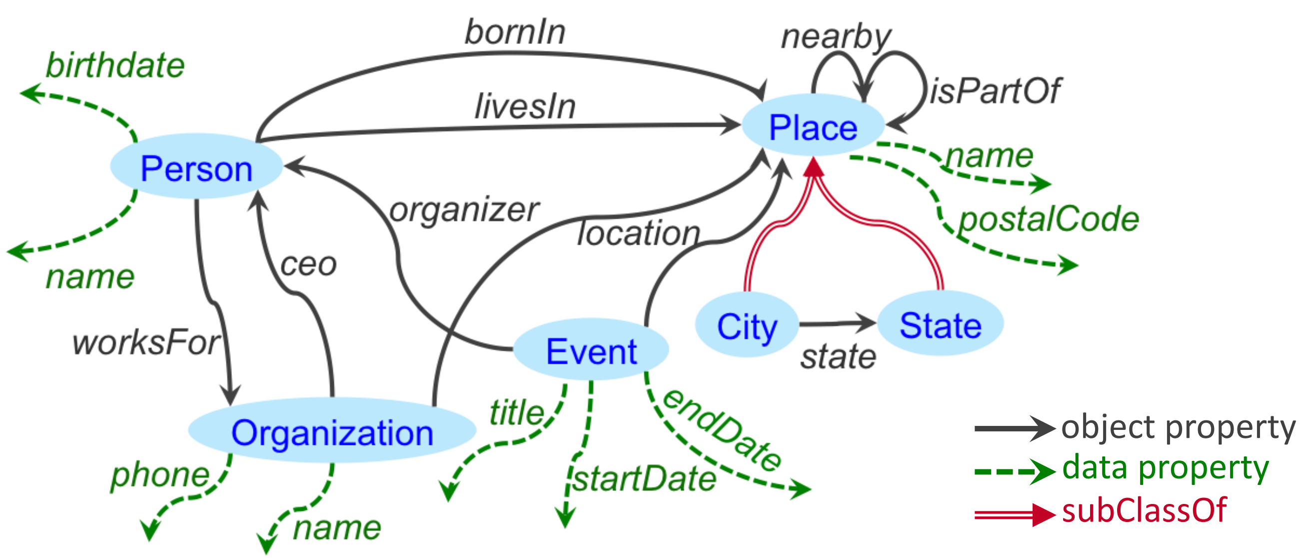


Example

Source

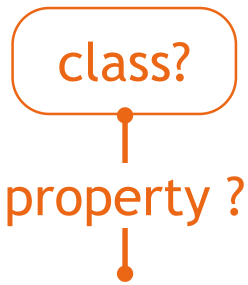
	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google

Domain Ontology



Find a semantic model for the source (map the source to the ontology)

Learning Semantic Types
[Krishnamurthy et al., ESWC 2015]



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
...

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
...

1- User specifies

2- System learns

Learning Semantic Types

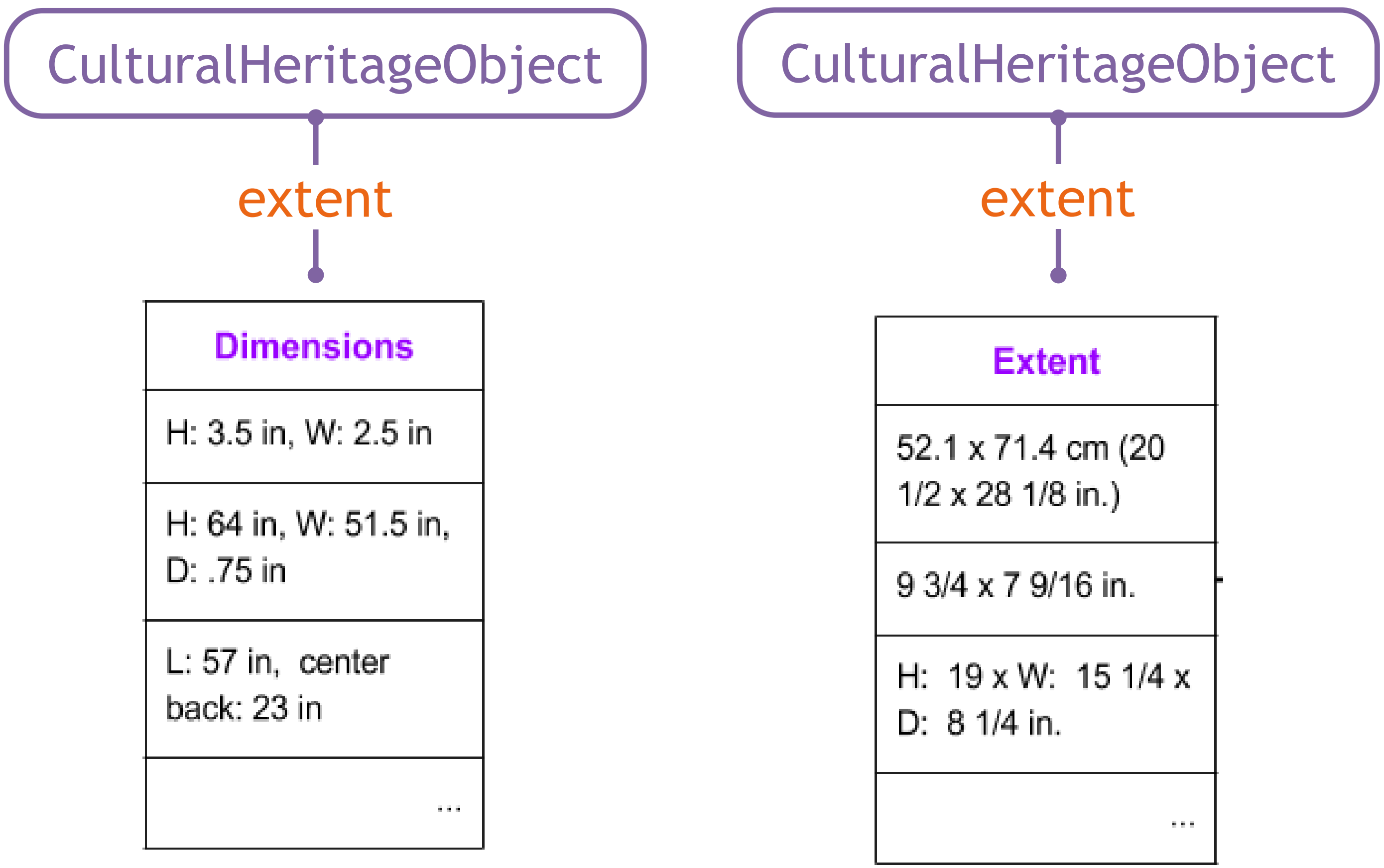
CulturalHeritageObject

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
...

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



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

Approach for Textual Data

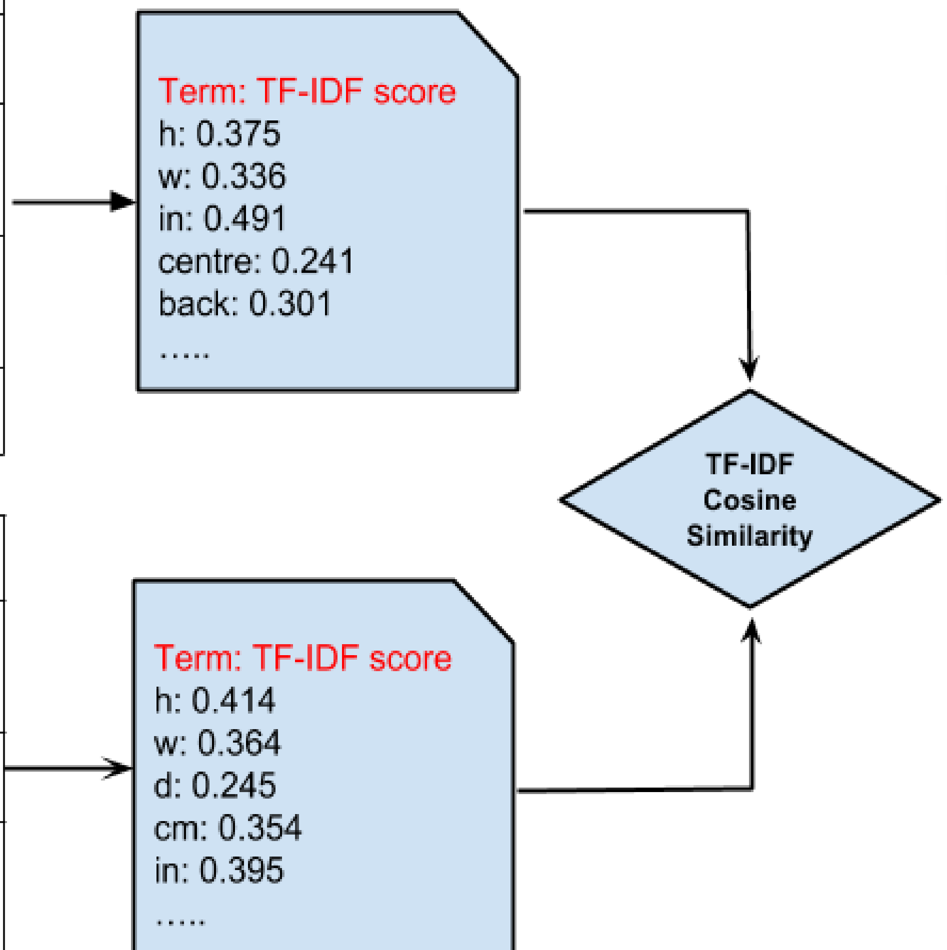
- Document: each column of data
- Label: each semantic type
- Use Apache Lucene to index the labeled documents
- Compute TF/IDF vectors for documents
- Compare documents using Cosine Similarity between TF/IDF vectors

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

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.
...



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

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

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

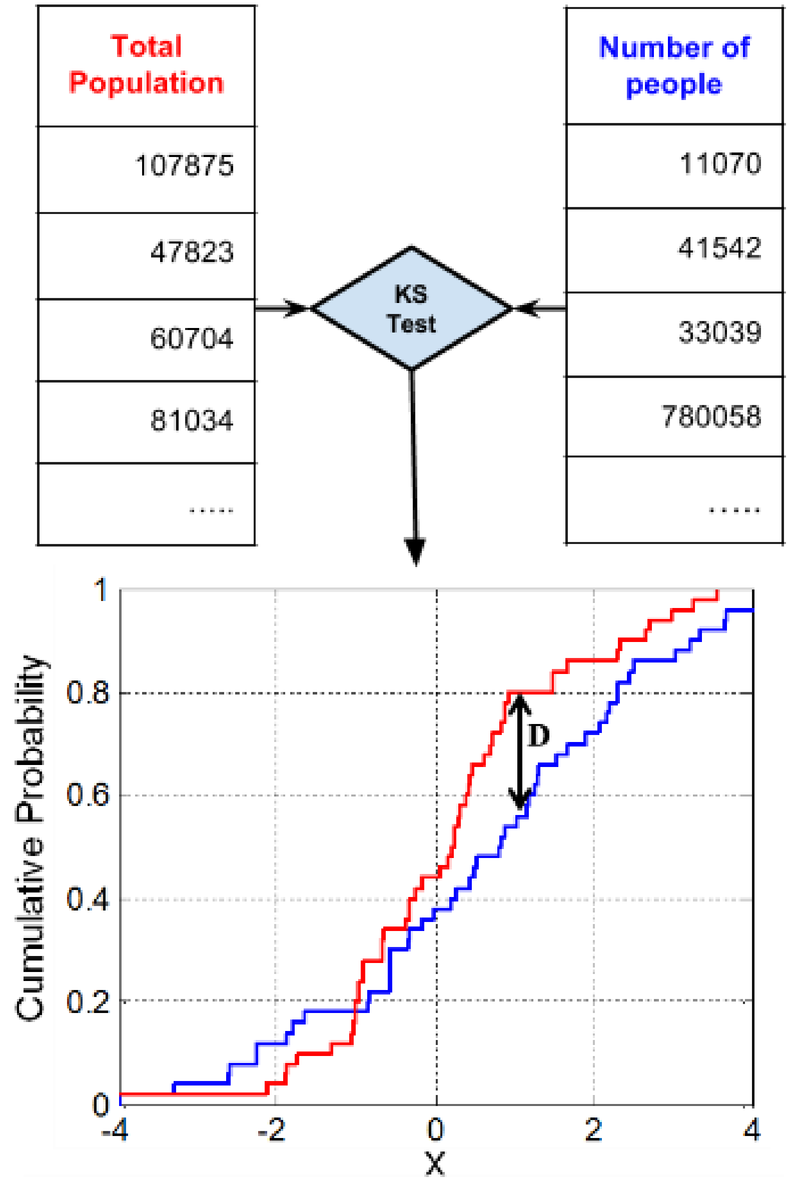
Approach for Numeric Data

- 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
.....

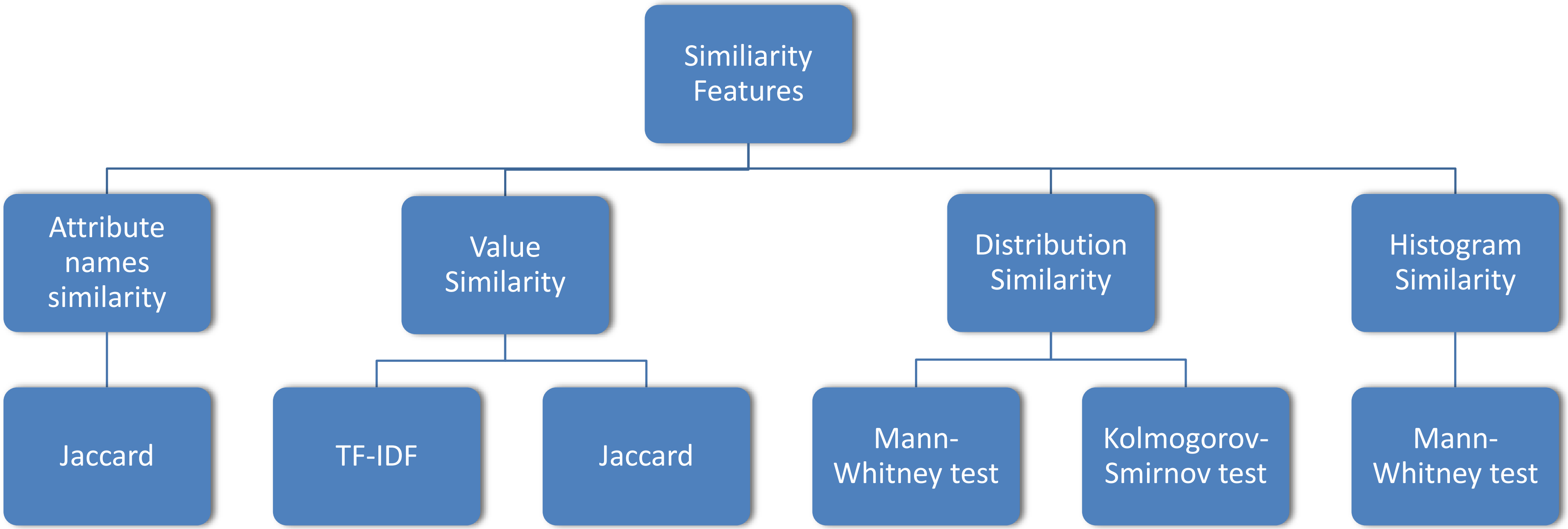
Number of people
11070
41542
33039
780058
.....

Approach for Numeric Data

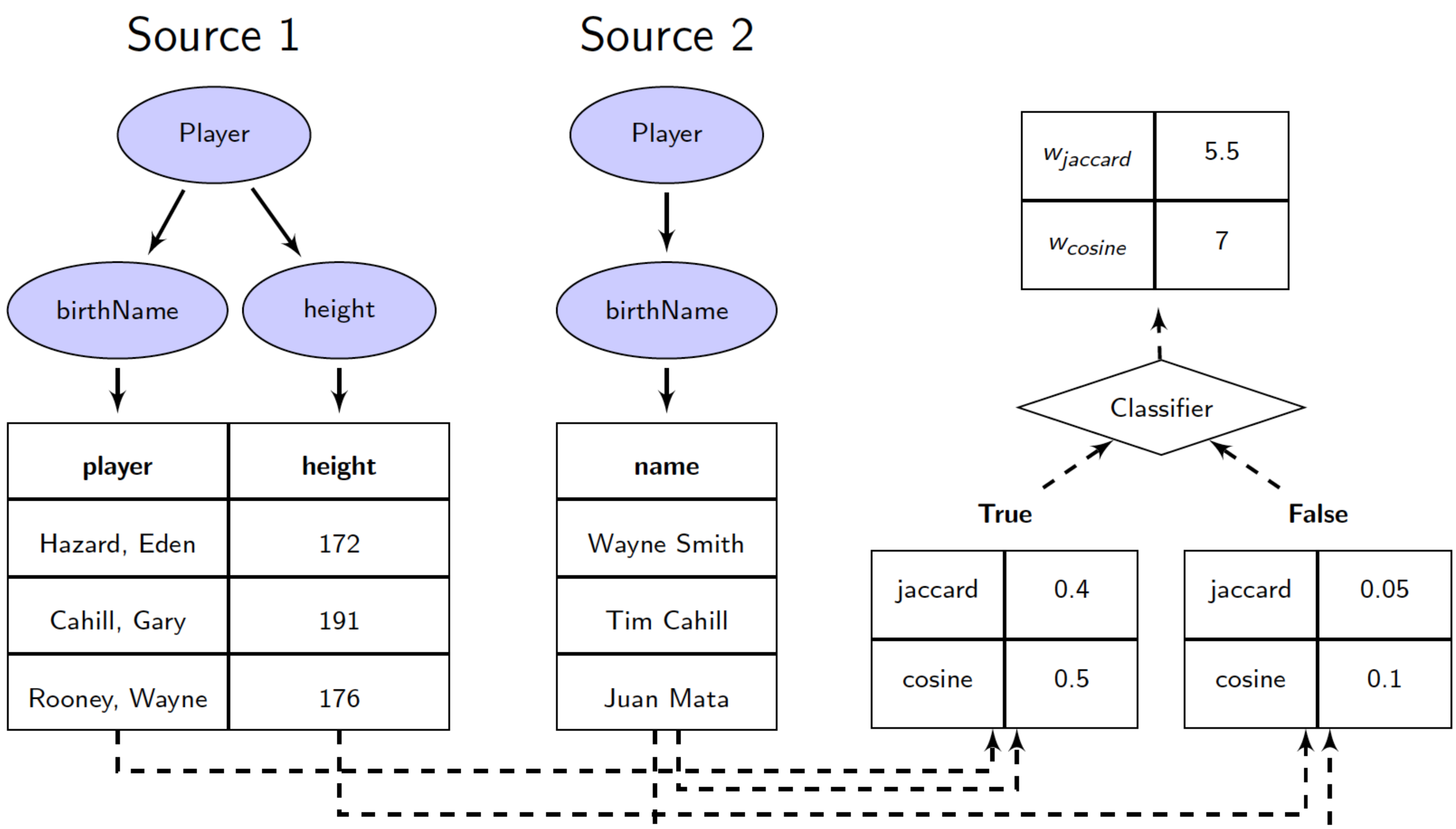


$$D_{N_1,N_2} = \sup_x |F_{1,N_1}(x) - F_{2,N_2}(x)|$$

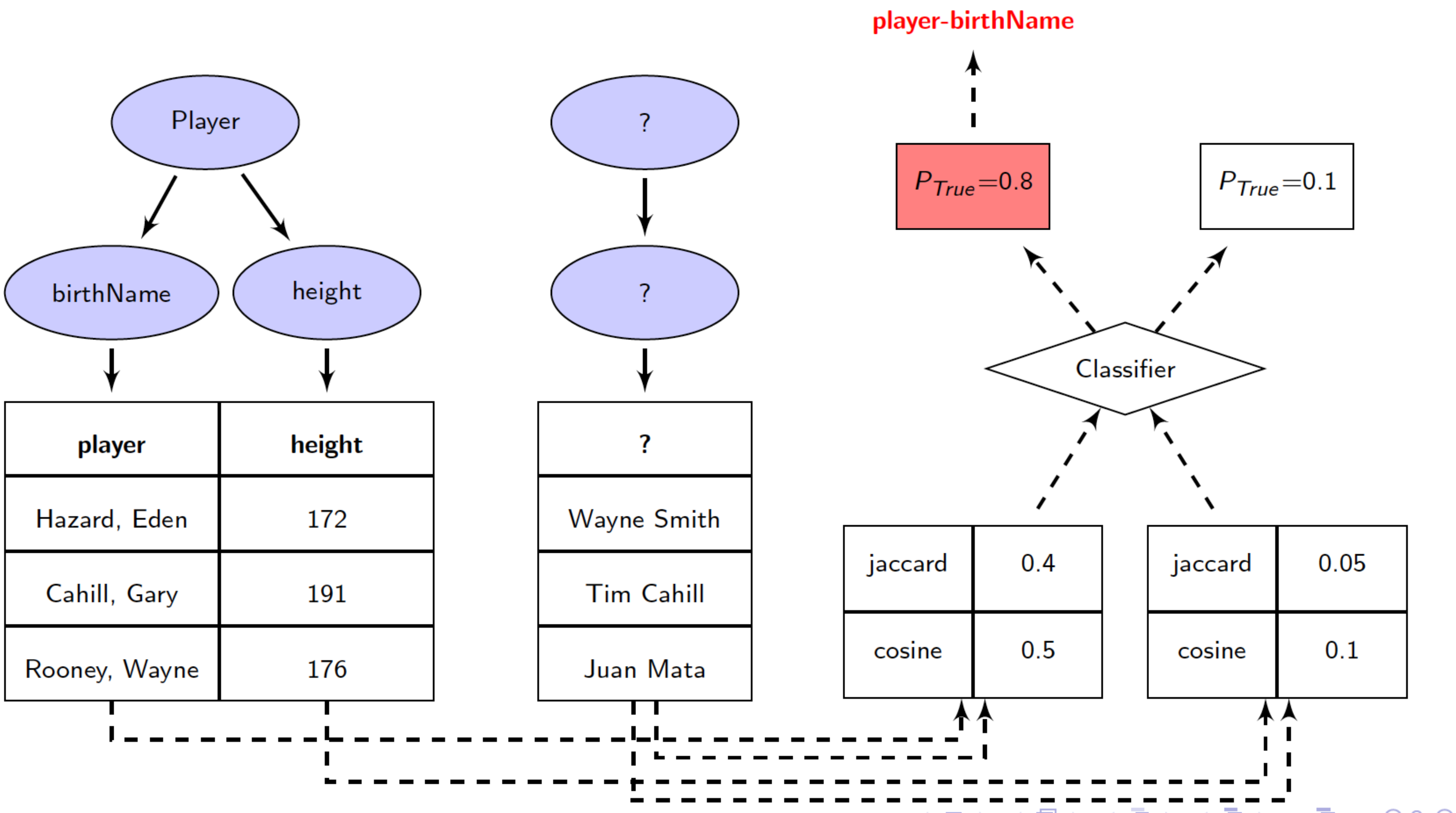
Similarity features



Training machine learning model
[Pham et al., ISWC 2016]

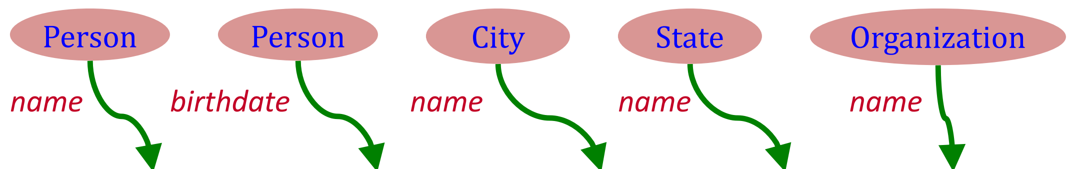


Predicting new attribute

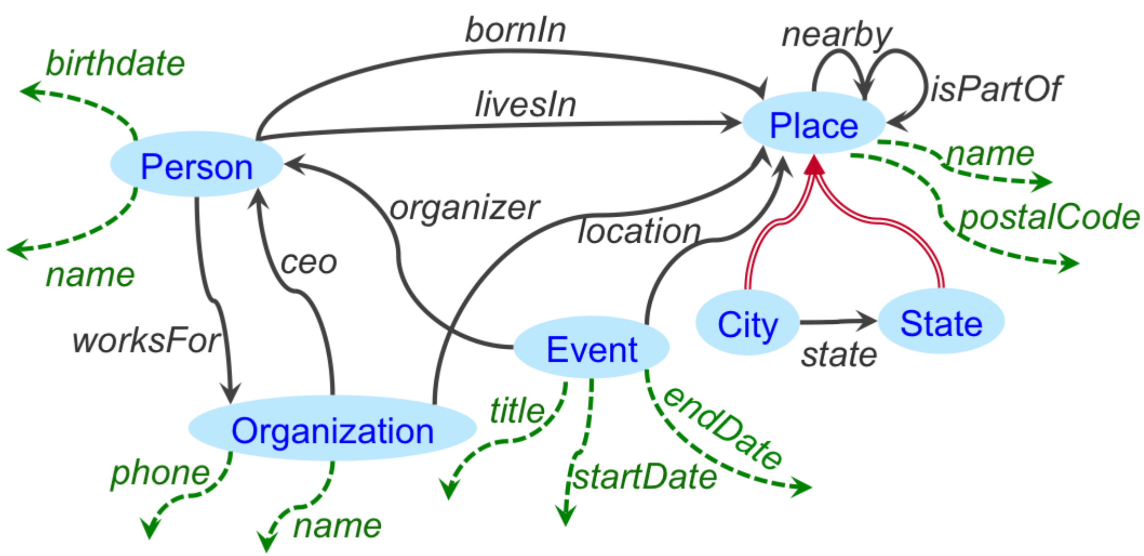


Construct a Graph

Construct a graph from semantic types and ontology

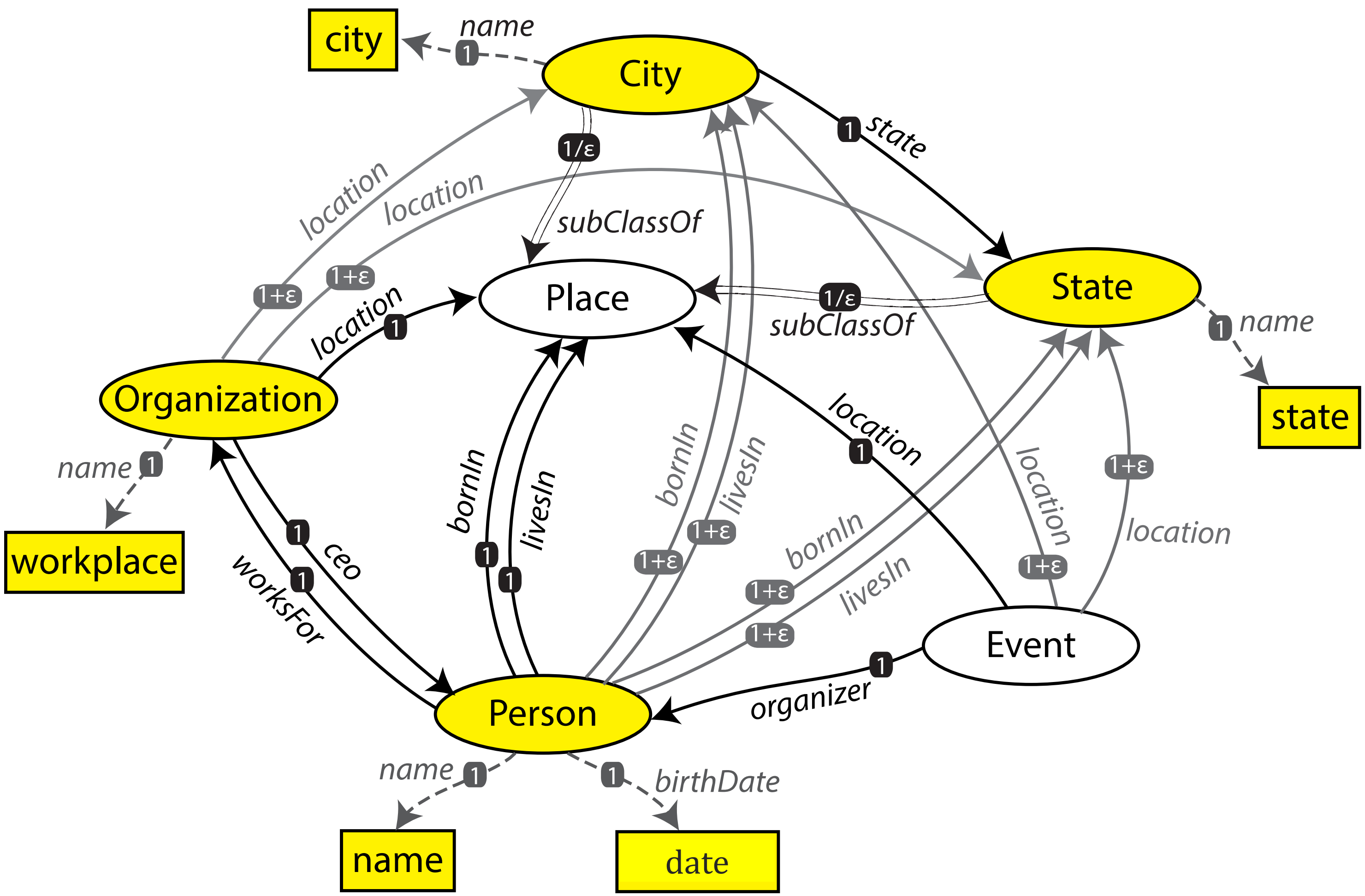


	name	date	city	state	workplace
1	Fred Collins	Oct 1959	Seattle	WA	Microsoft
2	Tina Peterson	May 1980	New York	NY	Google



Construct a Graph

Construct a graph from semantic types and ontology



Inferring the Relationships

- Search for minimal explanation

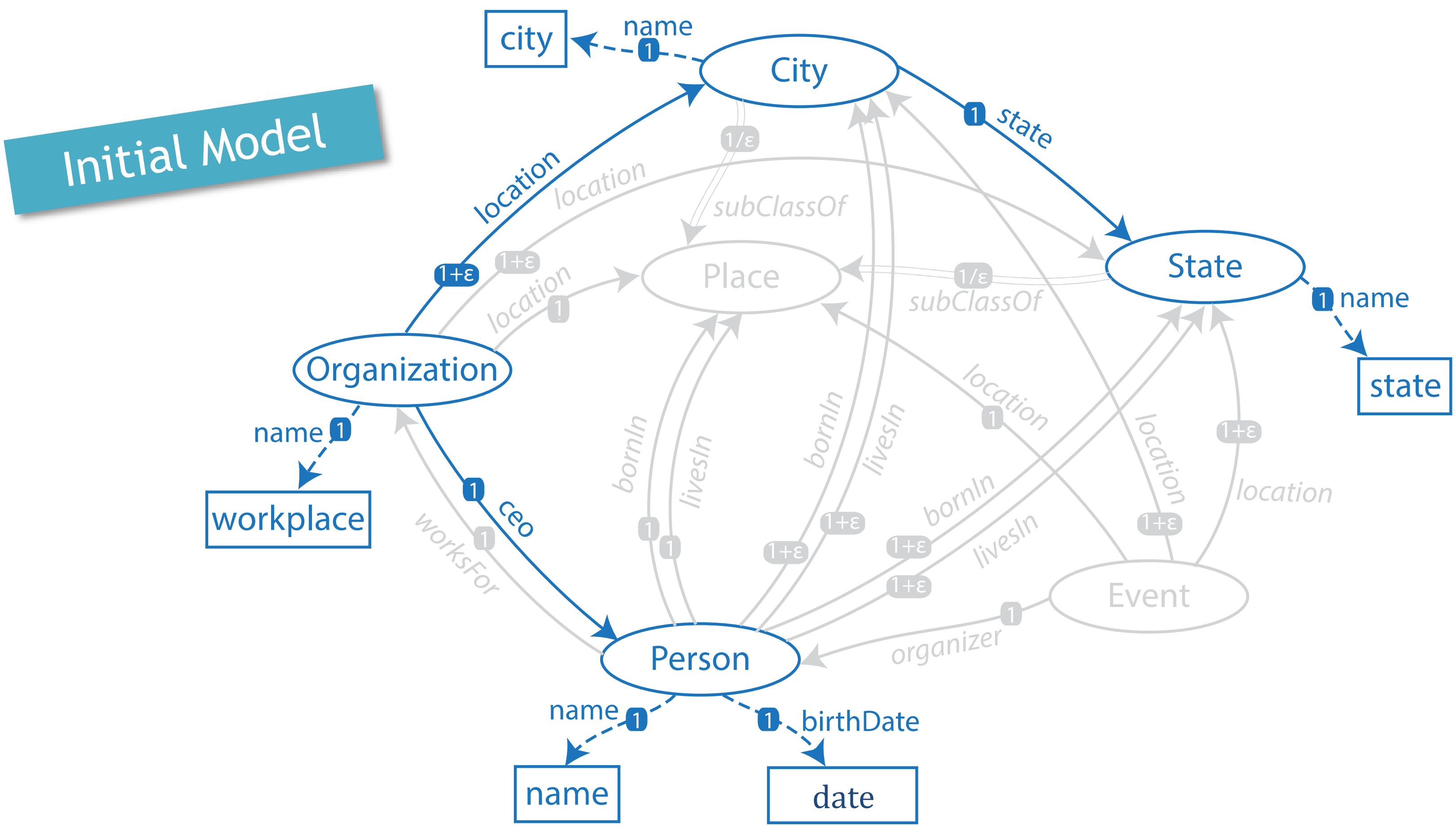
- Steiner tree connecting semantic types over ontology graph
  - Given graph G=(V,E), nodes $S \subset V$, cost $c: E \rightarrow \Re$
  - Find a tree of G that spans S with minimal total cost
  - Unfortunately, NP-complete

- Approximation Algorithm [Kou et al., 1981]
  - Worst-case time complexity: $O(|V|^2|S|)$
  - Approximation Ratio: less than 2

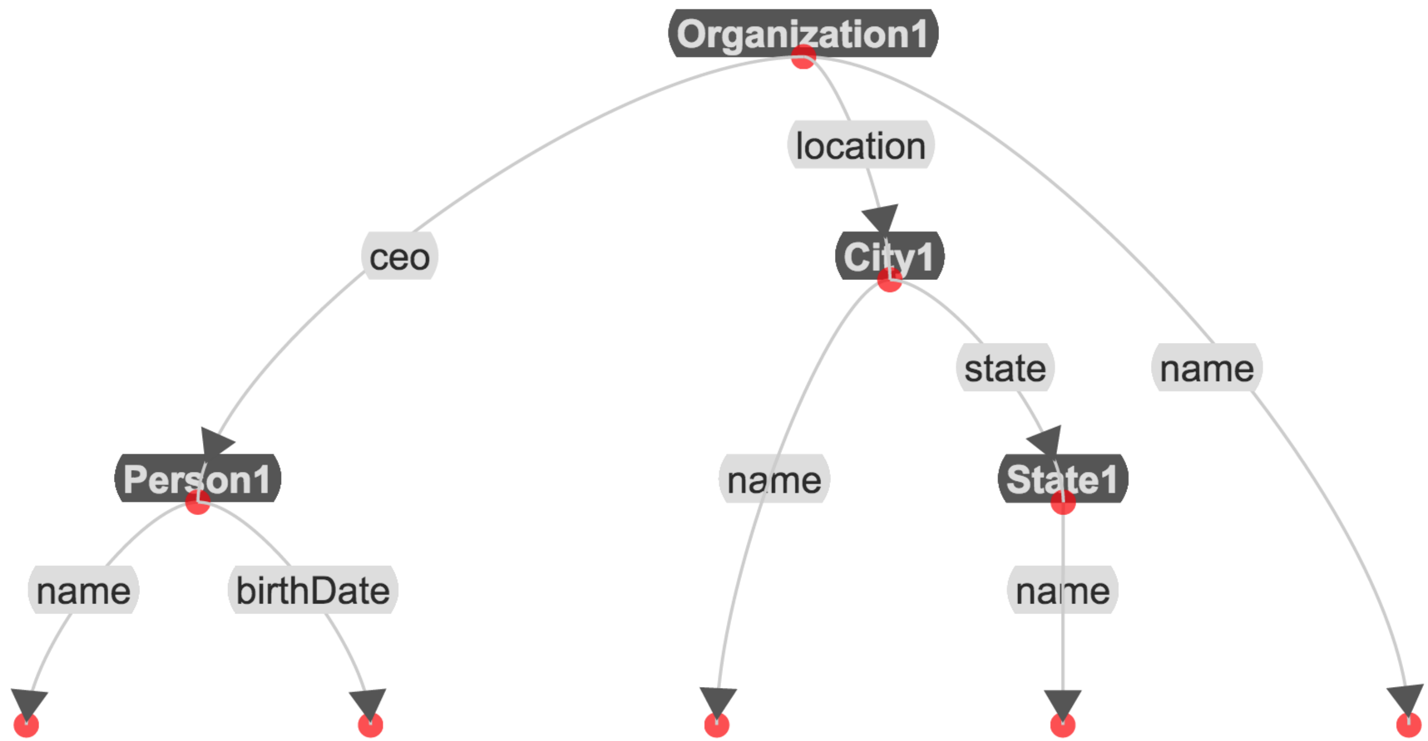
Inferring the Relationships

Select minimal tree that connects all semantic types

- A customized **Steiner tree algorithm**



Result in Karma

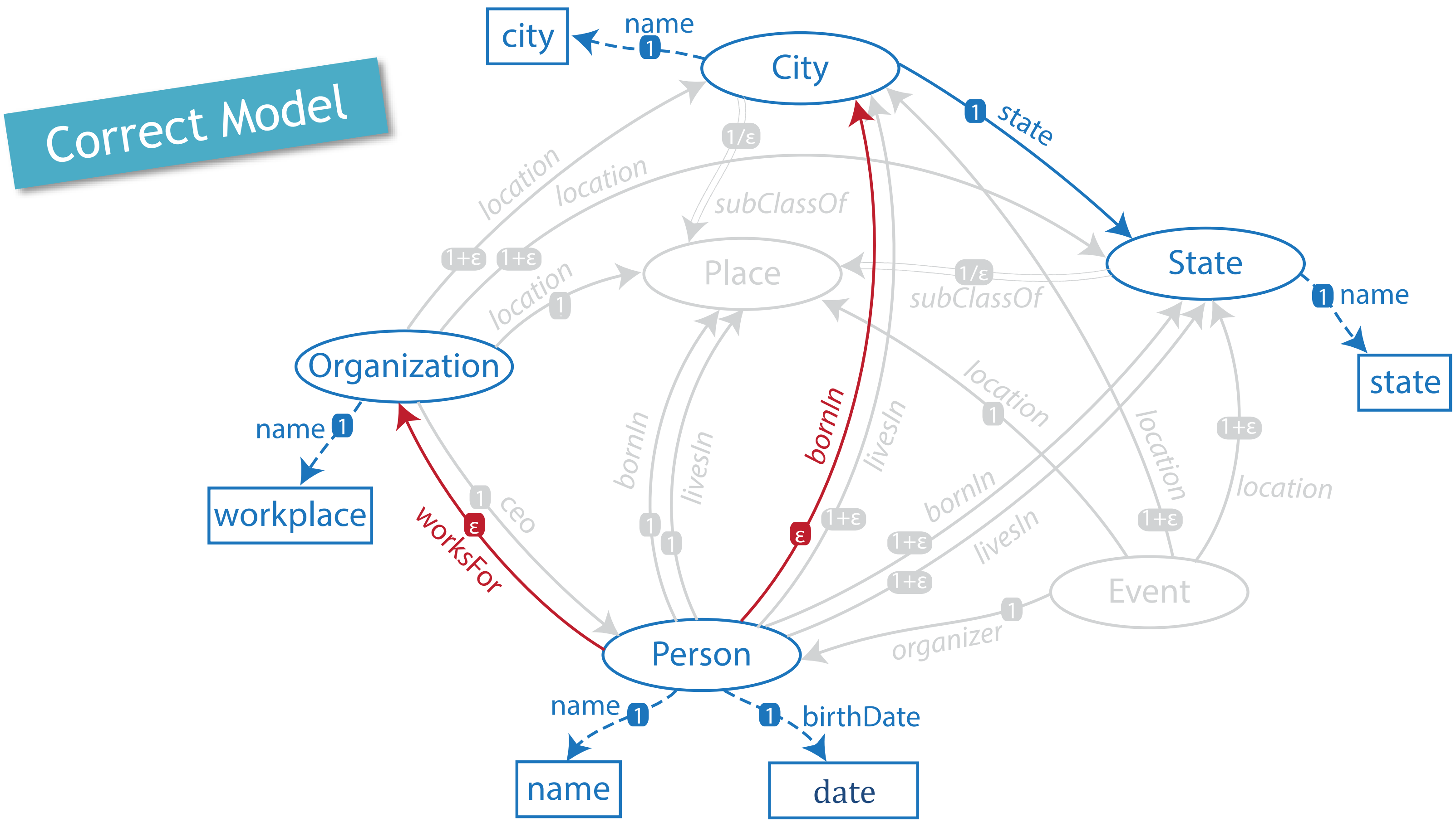


name	birthdate	city	state	workplace
Fred Collins	Oct 1959	Seattle	WA	Microsoft
Tina Peterson	May 1980	New York	NY	Google
Richard Smith	Feb 1975	Los Angeles	CA	Apple

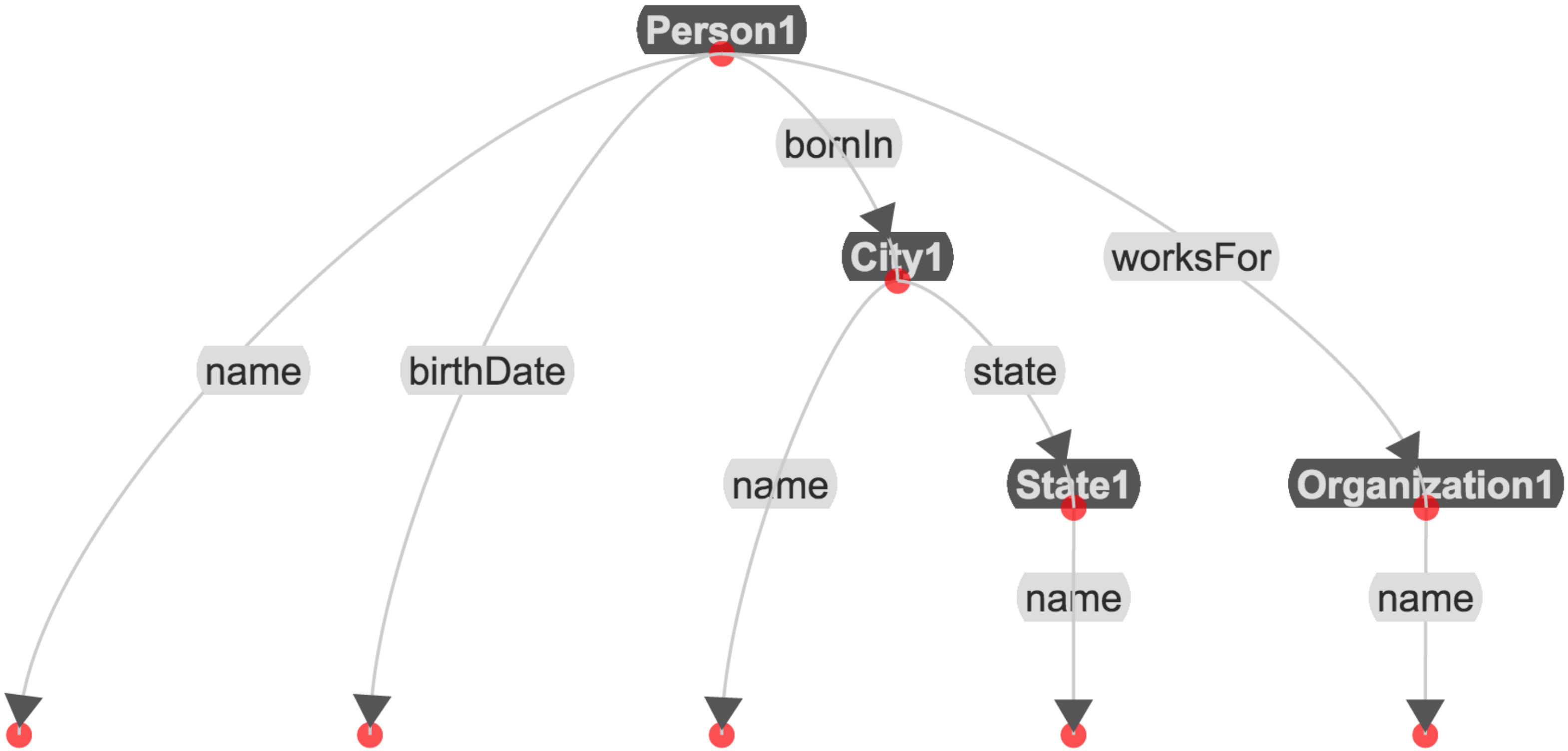
Refining the Model

Impose constraints on Steiner Tree Algorithm

- Change weight of selected links to ε
- Add source and target of selected link to Steiner nodes



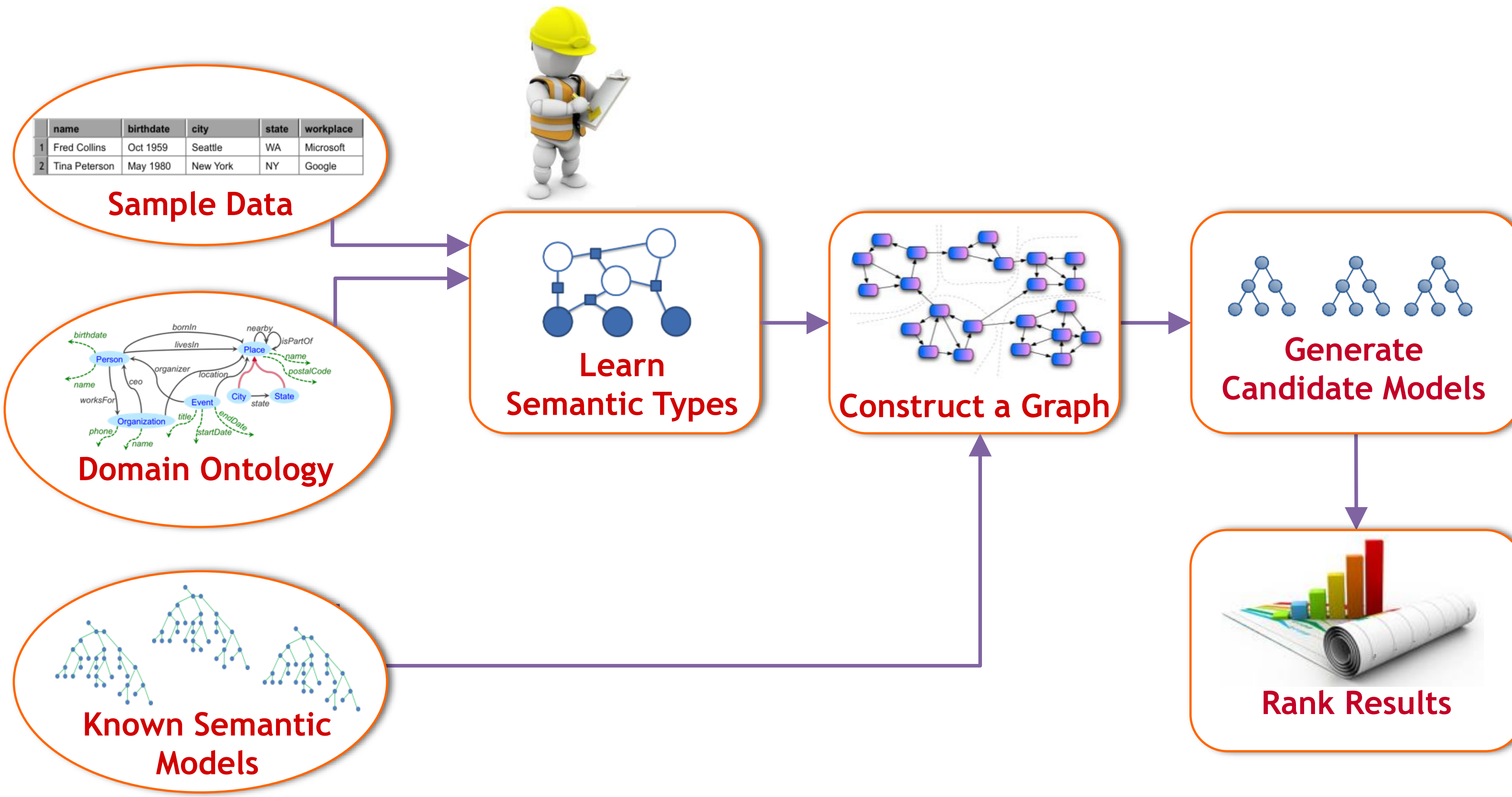
Final Semantic Model



name	birthdate	city	state	workplace
Fred Collins	Oct 1959	Seattle	WA	Microsoft
Tina Peterson	May 1980	New York	NY	Google
Richard Smith	Feb 1975	Los Angeles	CA	Apple

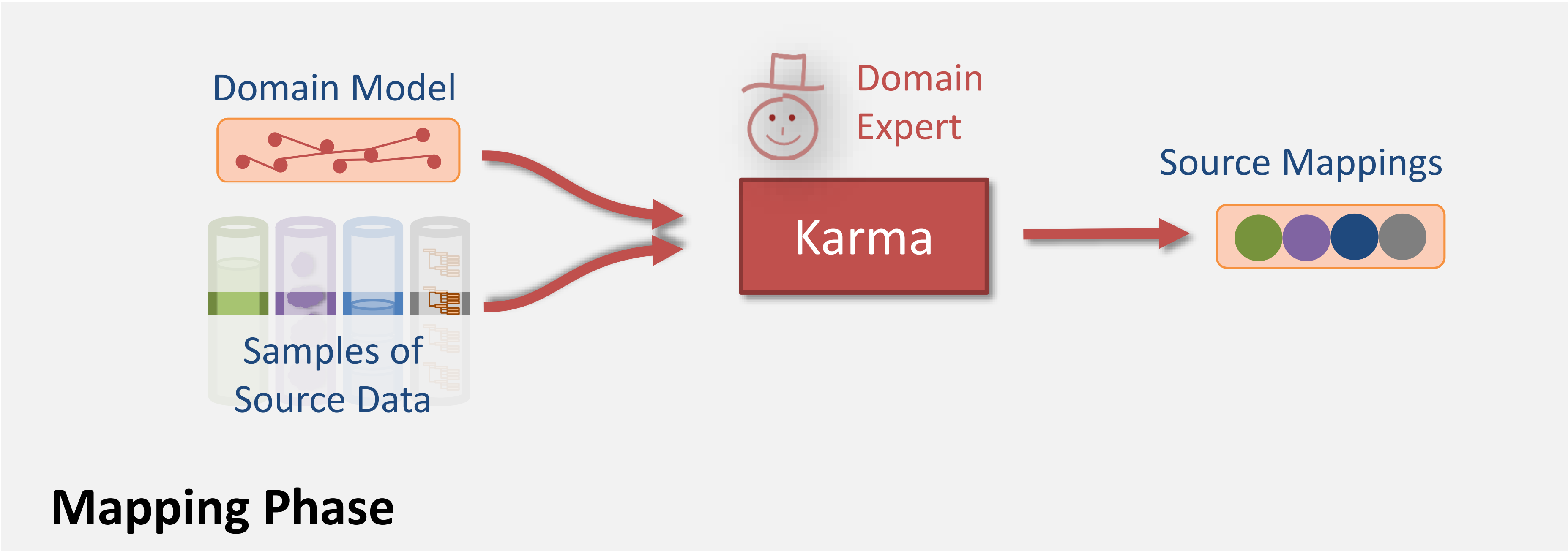
Karma Learns the Source Models

Taheriyan et al., ISWC 2013, ICSC 2014

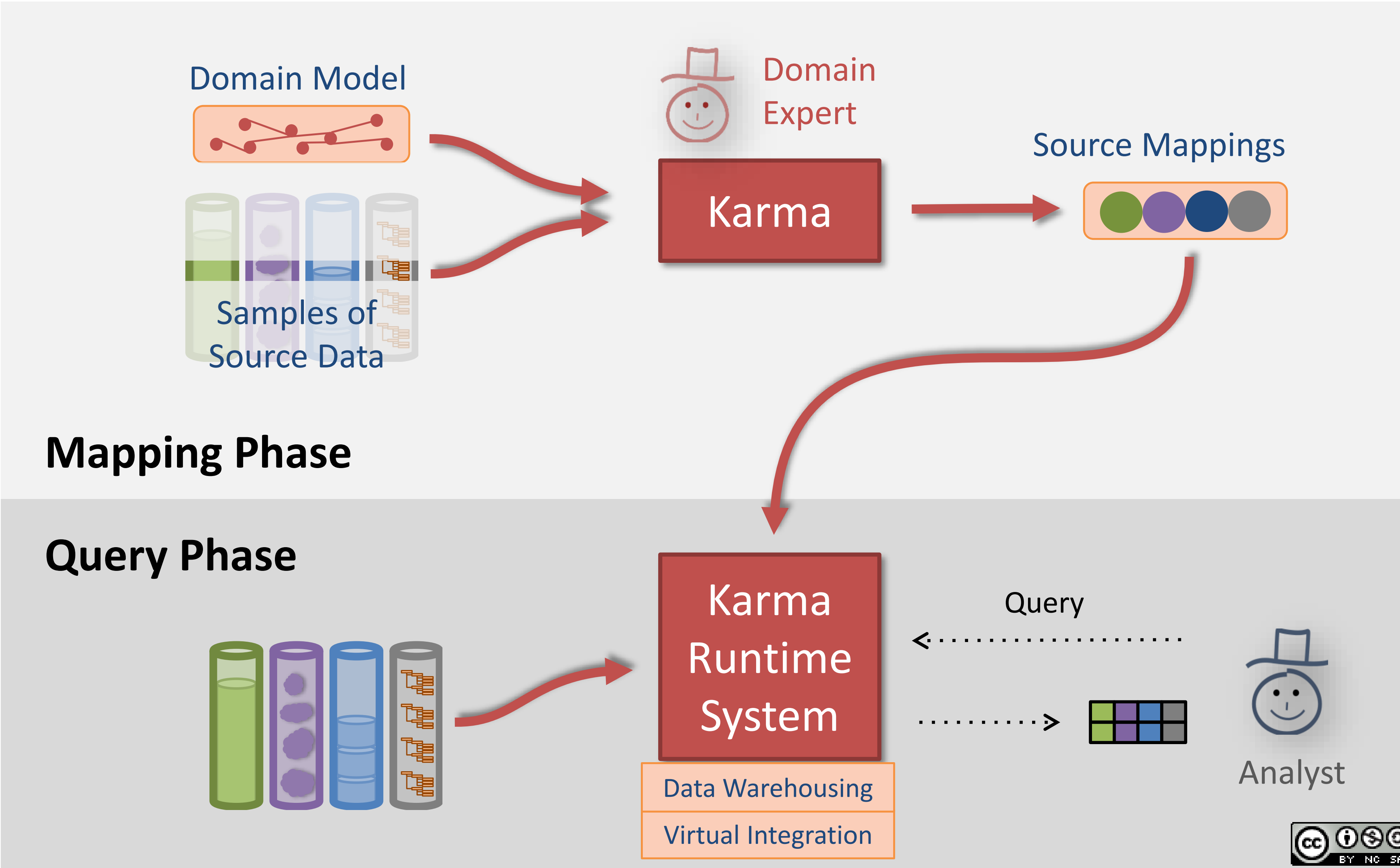


Karma Use Cases

Source Mapping Phase



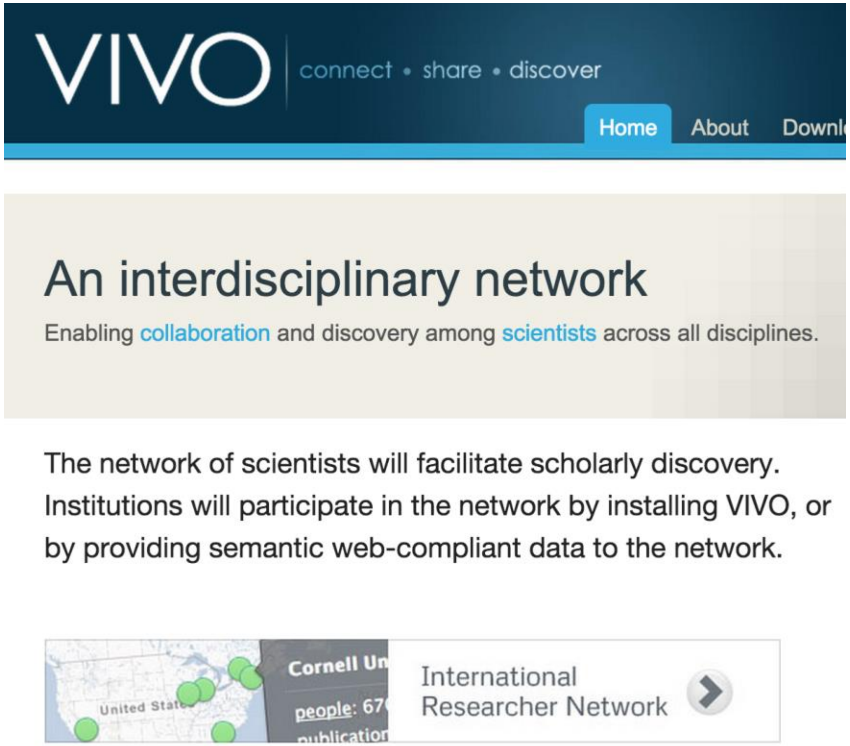
Source Mapping and Query Time



Pedro Szekely and Craig Knoblock

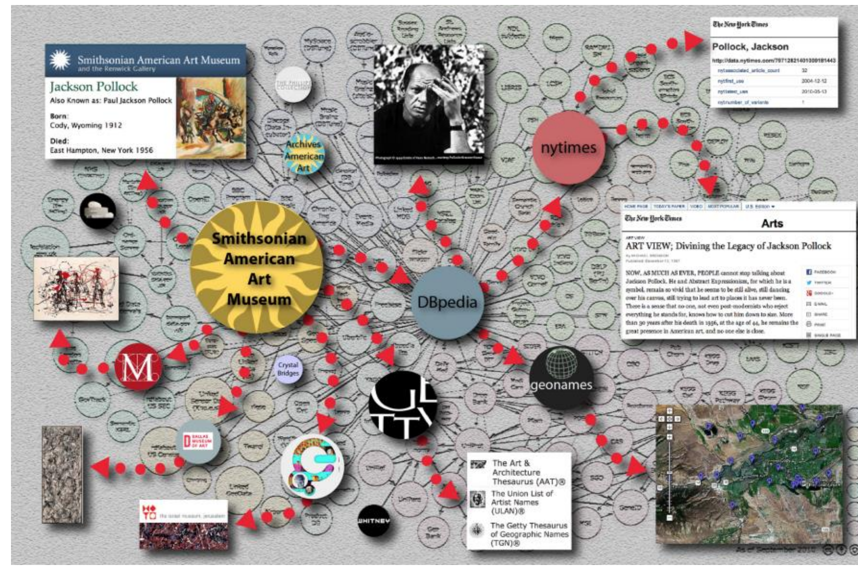
VIVO

- VIVO is a system to build researcher networks across institutions
- Used Karma to map the data about USC faculty to VIVO ontology and publish it as RDF
- VIVO ingest the RDF data
- Video



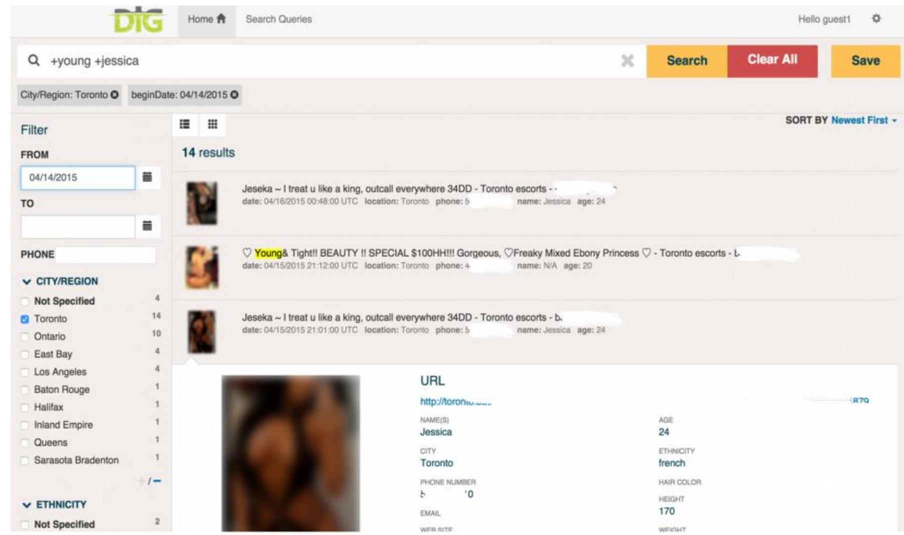
Smithsonian American Art Museum

- Used Karma to convert data of 44000 museum objects to Linked Open Data
- Modeled according to Europeana Data Model (EDM)
- Linked the generated RDF to DBPedia, ULAN, NY Times Linked Data
- News: USC press, Viterbi
- Video



DIG

- DIG: Domain-specific Insight Graphs
- Building and using knowledge graphs to combat human trafficking
- Used Karma to map extracted data and structured sources to shared domain ontology
- News: Forbes, Wired.co.uk



Demo

Using Karma to map museum data to the CIDOC CRM ontology

https://www.youtube.com/watch?v=h3_yiBhAJIc

Discussion

- Automatically build rich semantic descriptions of data sources
- Exploit the background knowledge from (i) the domain ontology, and (ii) the known source models
- Semantic descriptions are the key ingredients to automate many tasks, e.g.,
  - Source Discovery
  - Data Integration
  - Service Composition

More Info

karma.isi.edu