INFORMATION SCIENCES INSTITUTE
• • • agent of innovation • • •

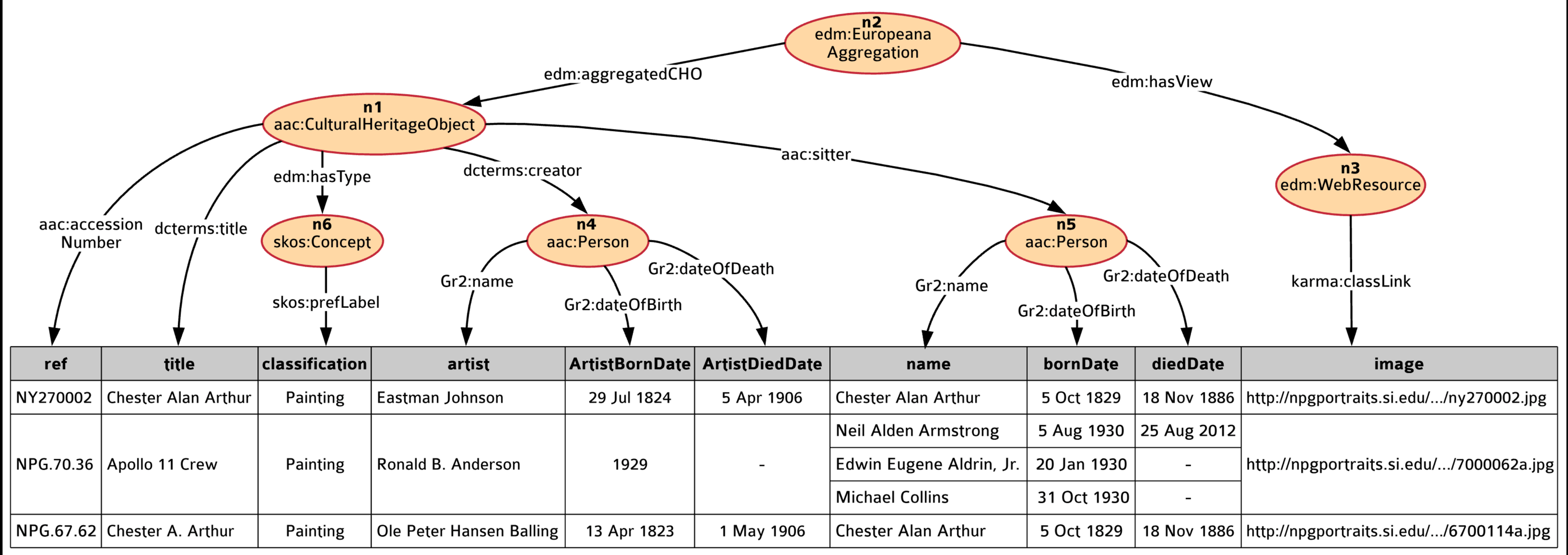
Learning Semantic Models of Data Sources using Probabilistic Graphical Models

Binh Vu, Craig A. Knoblock and Jay Pujara

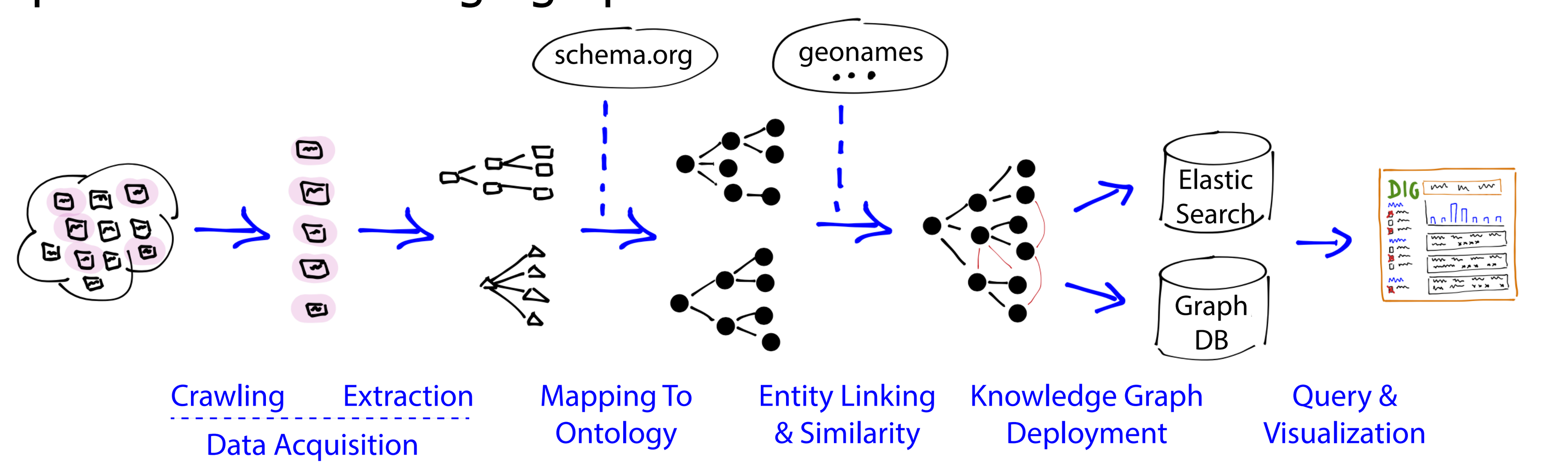
USC University of Southern California

Overview

- **Problem**: harvesting information from data sources can be challenging because data is published in different data formats and using different conventions
- **Goal**: build a semantic model that describes the data source
  - Input:
    - ❑ A set of domain ontologies $\boldsymbol{O}$
    - ❑ A target data source $s(a_1, a_2, \ldots, a_n)$: a_i is a source attribute
  - Output:
    - ❑ A semantic model $sm(s)$

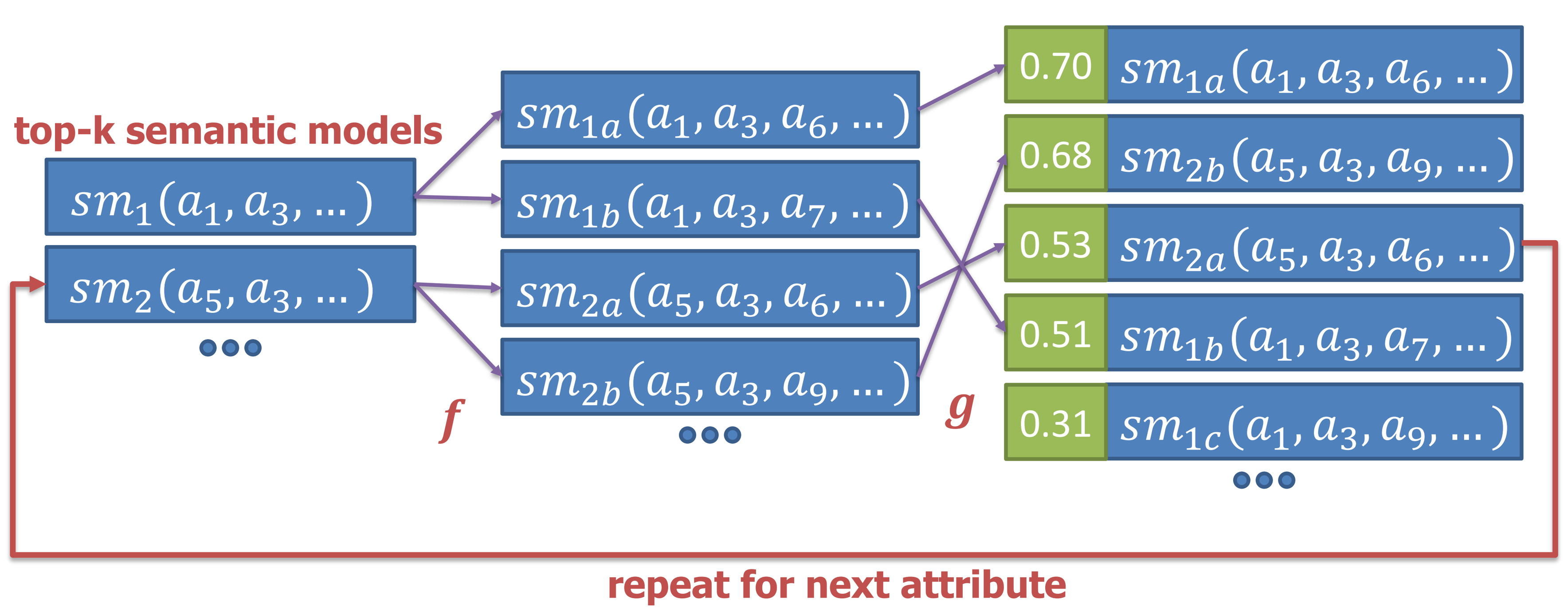


- **Application**: automatically convert data sources to RDF triples to publish to knowledge graphs



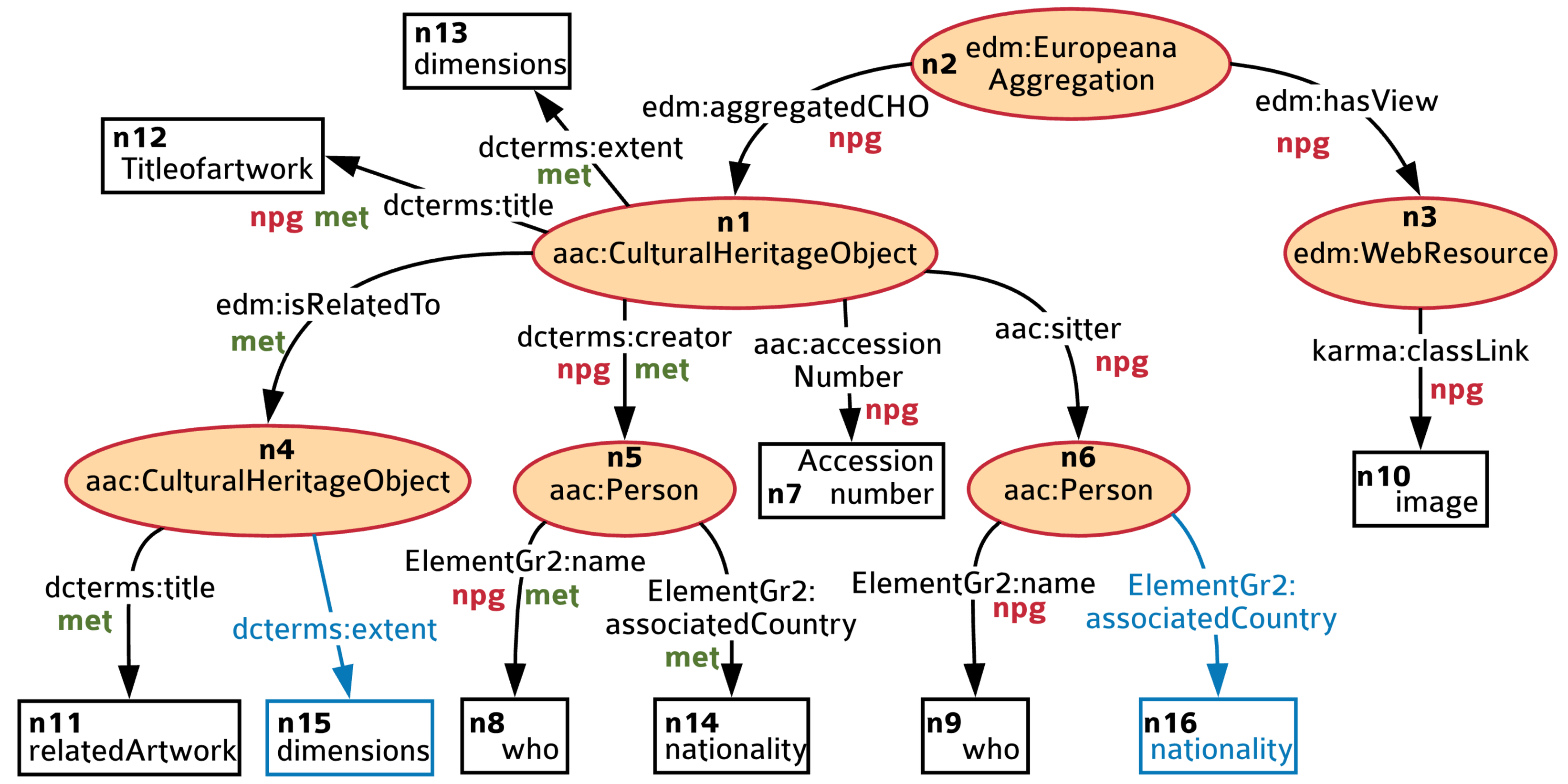
Overall search-based approach

- Use beam search to find the most probable semantic model
  - Navigate in the combinatorial space using a transition function f
  - Rank and select modeling options using a graphical model g

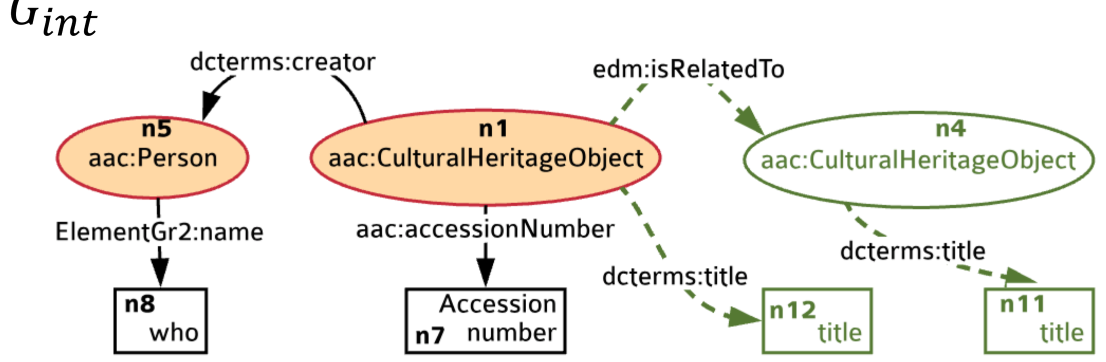


How to navigate in the search space

- Construct a space of possible semantic models (G_{int})

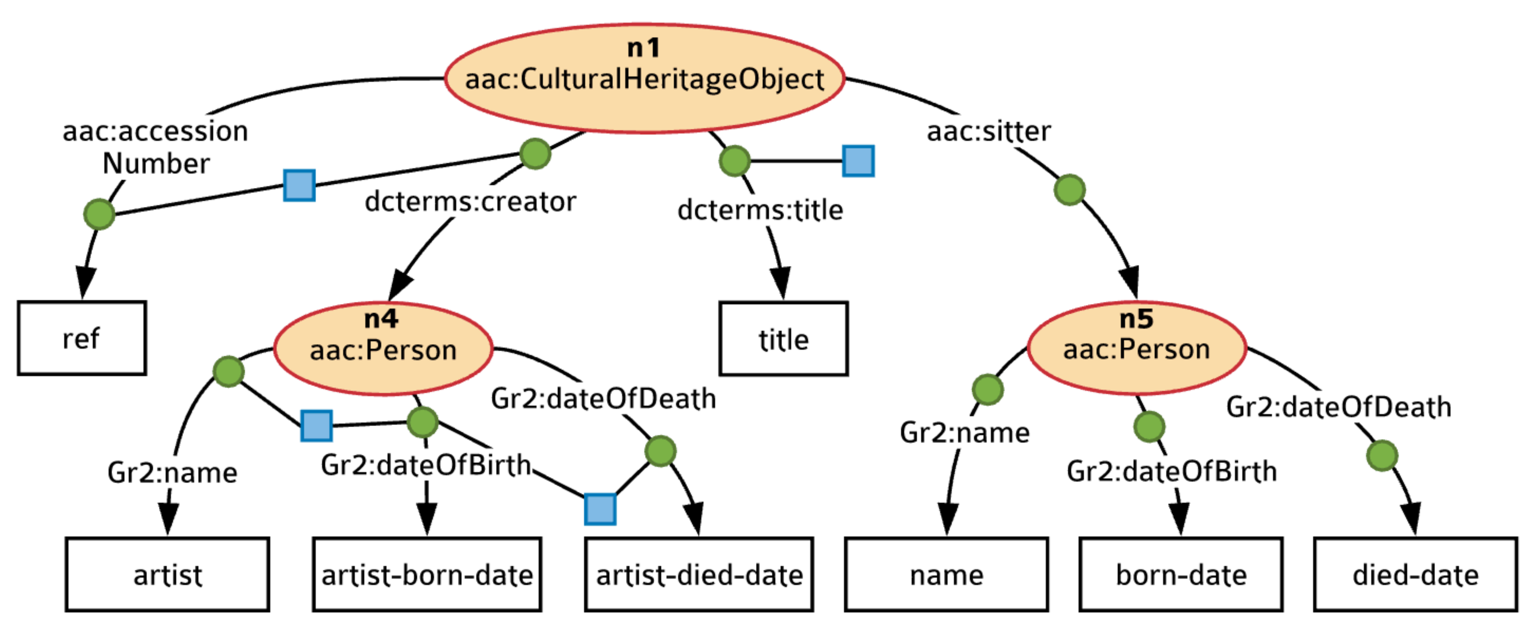


- Transition function f: merge new attribute to existing semantic model according to G_{int}



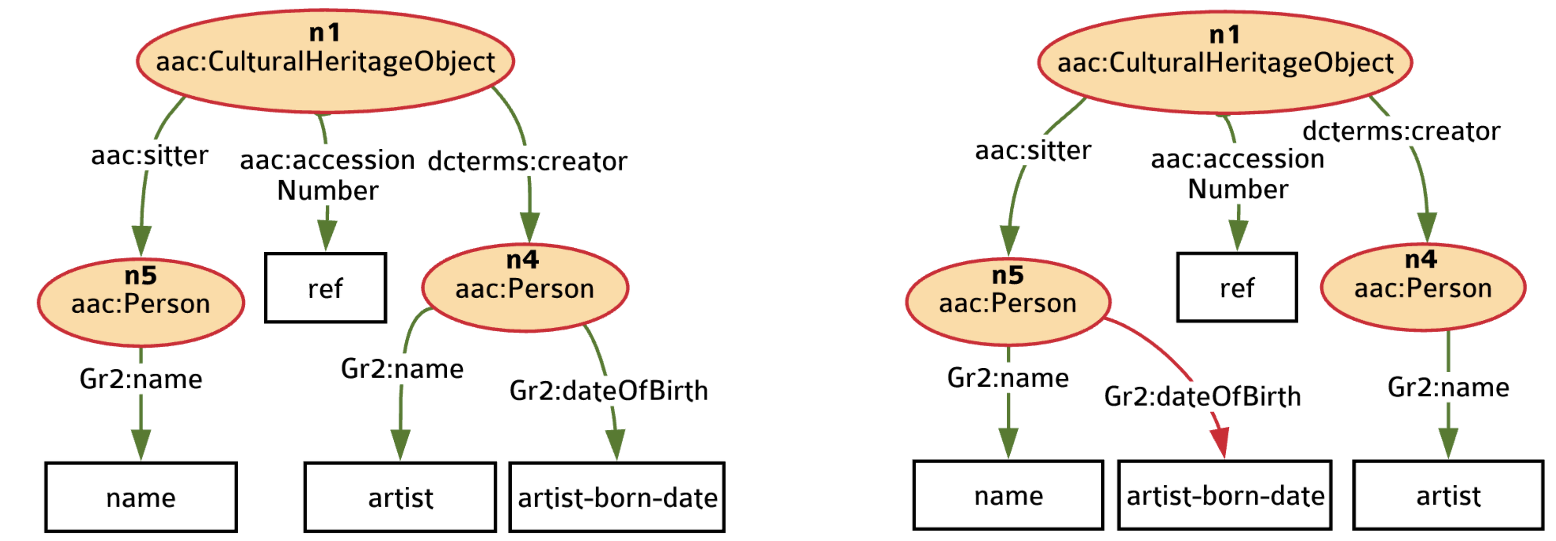
How to rank semantic models

- Likelihood of a semantic model - $P(\forall y \in \boldsymbol{y}; y = \text{true} | \boldsymbol{x})$ - indicates the quality of the model
- Estimate the likelihood using a conditional random field (CRF)
  - $P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(x_c)} \prod_{C_p \in \mathcal{C}} \prod_{\Psi_c \in C_p} \Psi_c(\boldsymbol{y}_c, \boldsymbol{x}_c; \theta_p)$
    - $Z(\boldsymbol{x}_c) = \sum_{\boldsymbol{y}} \prod_{C_p \in \mathcal{C}} \prod_{\Psi_c \in C_p} \Psi_c(\boldsymbol{y}_c, \boldsymbol{x}_c; \theta_p)$
    - $\Psi_c(\boldsymbol{y}_c, \boldsymbol{x}_c; \theta_p) = \exp\{\sum_k \theta_{pk} f_{pk}(\boldsymbol{y}_c, \boldsymbol{x}_c)\}$
- Features:
  - Link confidence
  - Cardinality relationships between source attributes
  - Structural similarity
  - …



Training the Graphical Model

- Create labeled data from sample of possible semantic models
- Train to identify correct/incorrect links in the models

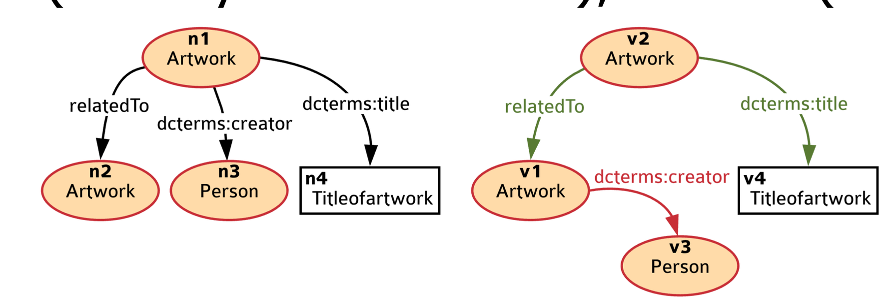


Evaluation

- **Datasets**: museum-crm and museum-edm (Taheriyan 2016)
- **Semantic labelers**: DSL (Pham et al. 2016), SemTyper (Ramnandan et al. 2015), Serene (Rummele et al. 2018)

Dataset	SemTyper	DSL	Serene
ds_{edm}	0.830	0.886	0.912
ds_{crm}	0.628	0.896	0.910

- **Baselines**: (Taheriyan et al. 2016), Serene (Una et al. 2018)



Methods	ds_{edm}				ds_{crm}			
	SemTyper	DSL	Serene	Oracle	SemTyper	DSL	Serene	Oracle
Taheriyan	0.712	0.635	0.803	0.887	0.618	0.695	0.774	0.857
Serene	0.693	0.719	0.789	0.885	0.663	0.698	0.753	0.840
PGM-SM	0.768	0.815	0.829	0.935	0.725	0.844	0.880	0.944

Conclusion and Future Work

- By exploiting relationships within the data sources and semantic models, our approach:
  - generates better semantic models
  - is more robust to noise
- Future work:
  - Minimize user effort by leveraging Linked Open Data
  - End-to-end system from web extraction to semantic model
  - Integrate with interactive modeling system (Karma)