

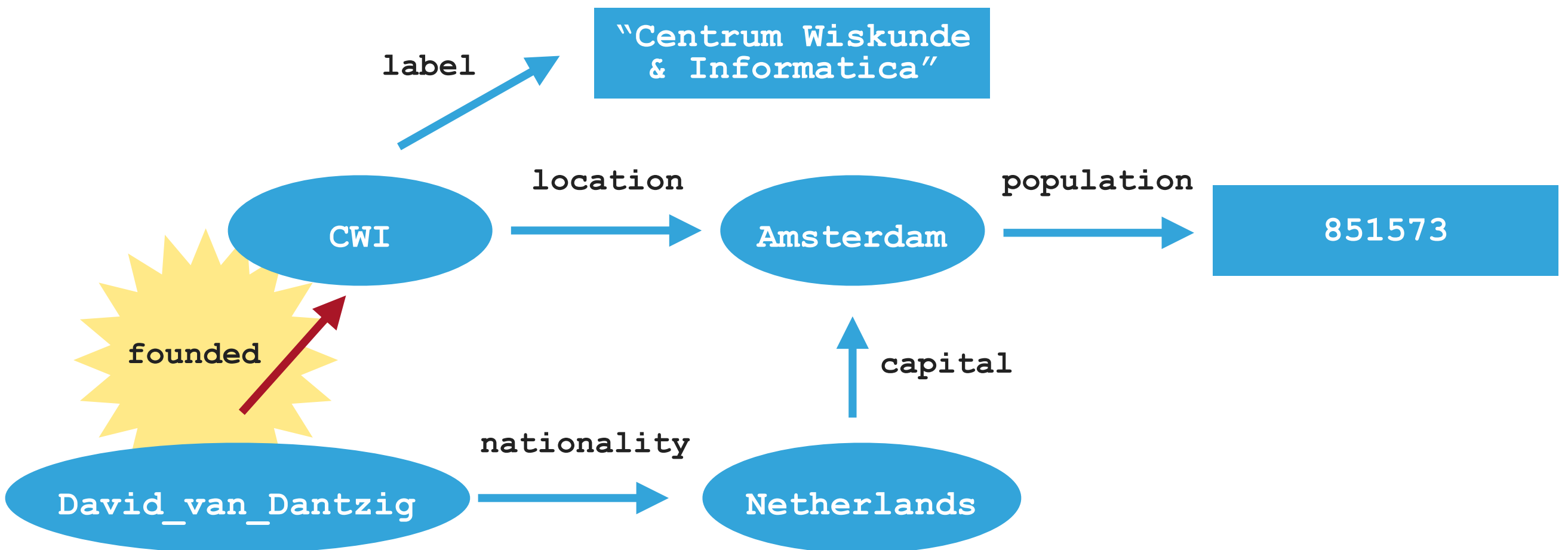


BENNO KRUIT, PETER BONCZ, JACOPO URBANI

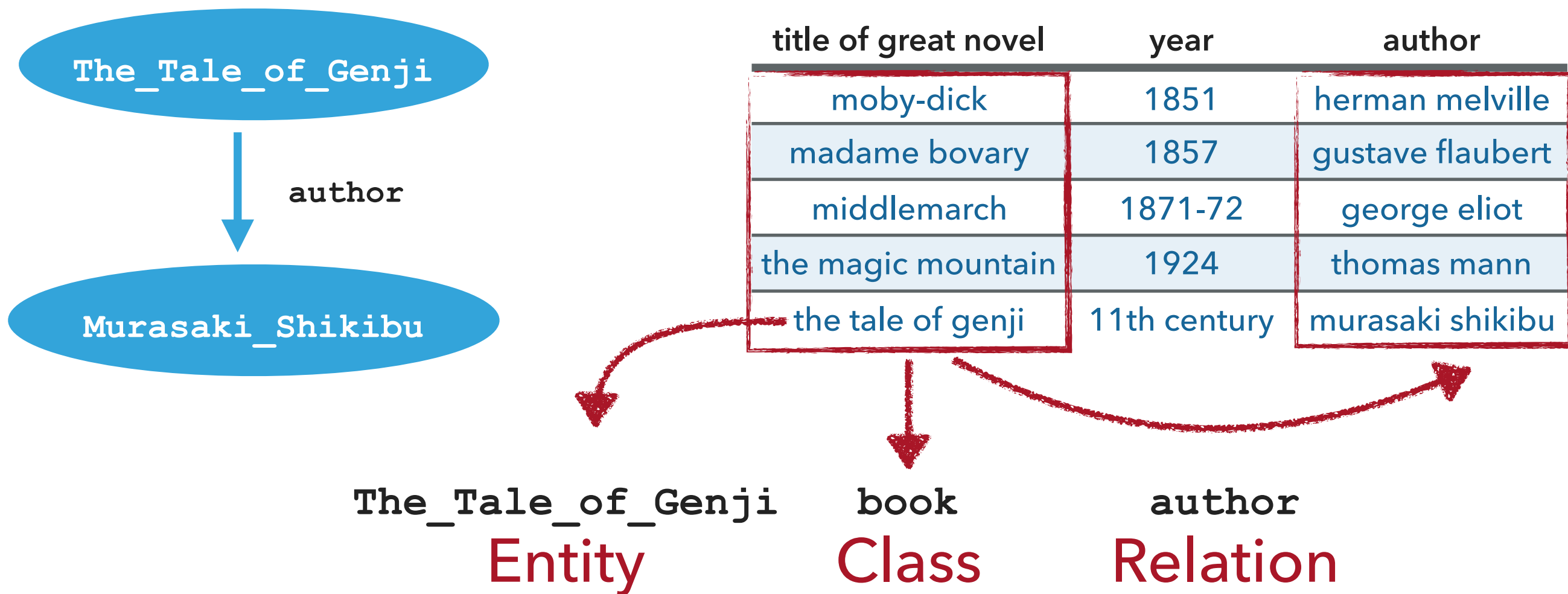
EXTRACTING NOVEL FACTS FROM TABLES FOR KNOWLEDGE GRAPH COMPLETION

MOTIVATION: Web Tables For Knowledge Graph Completion

- ▶ Knowledge Graphs are very big, but still highly incomplete
- ▶ We'd like to add links using web tables: slot filling



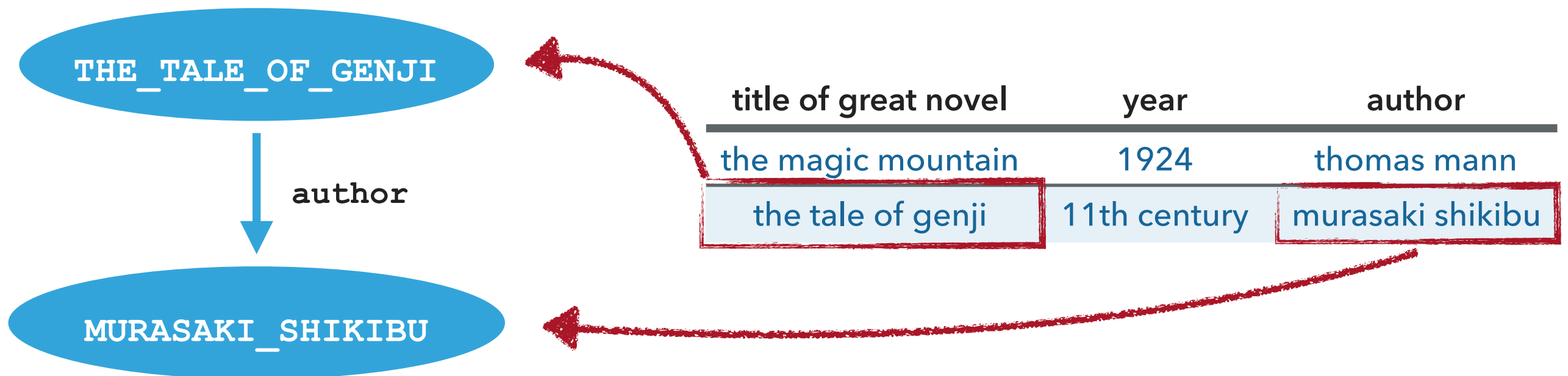
PROBLEM: Interpret Web Tables For Fact Extraction



- ▶ Disambiguate cells, classes and relations
- ▶ Add KG links using disambiguations

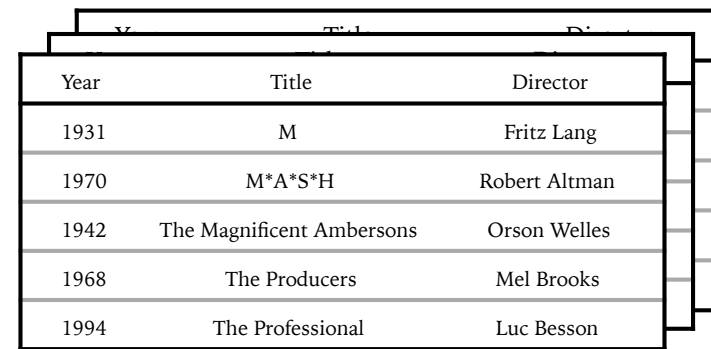
STATE OF THE ART for Table Interpretation

- ▶ T2K [Ritze et al. 2015]
- ▶ TableMiner⁺ [Zhang 2017]
- ▶ Both systems interpret tables based on **KG support**
 - ▶ More support = more confidence = *less novelty*?
- ▶ Both built for specific KGs



TAKCO: Table-based Knowledge graph Completion

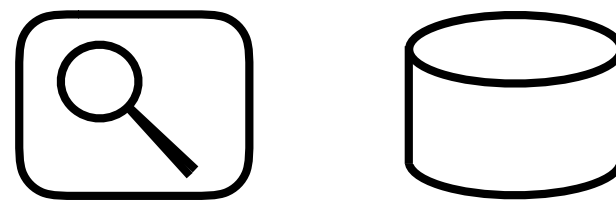
- ▶ **Novelty-oriented**
 - ▶ Recall first
 - ▶ Precision later
- ▶ KG-agnostic



A stack of three tables. The top table is visible and contains the following data:

Year	Title	Director
1931	M	Fritz Lang
1970	M*A*S*H	Robert Altman
1942	The Magnificent Ambersons	Orson Welles
1968	The Producers	Mel Brooks
1994	The Professional	Luc Besson

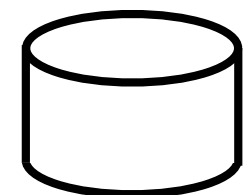
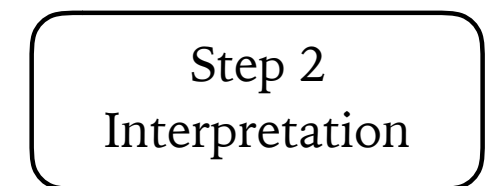
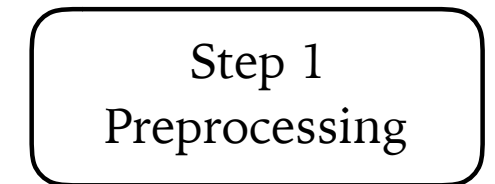
Tables



Label Index and Knowledge Graph



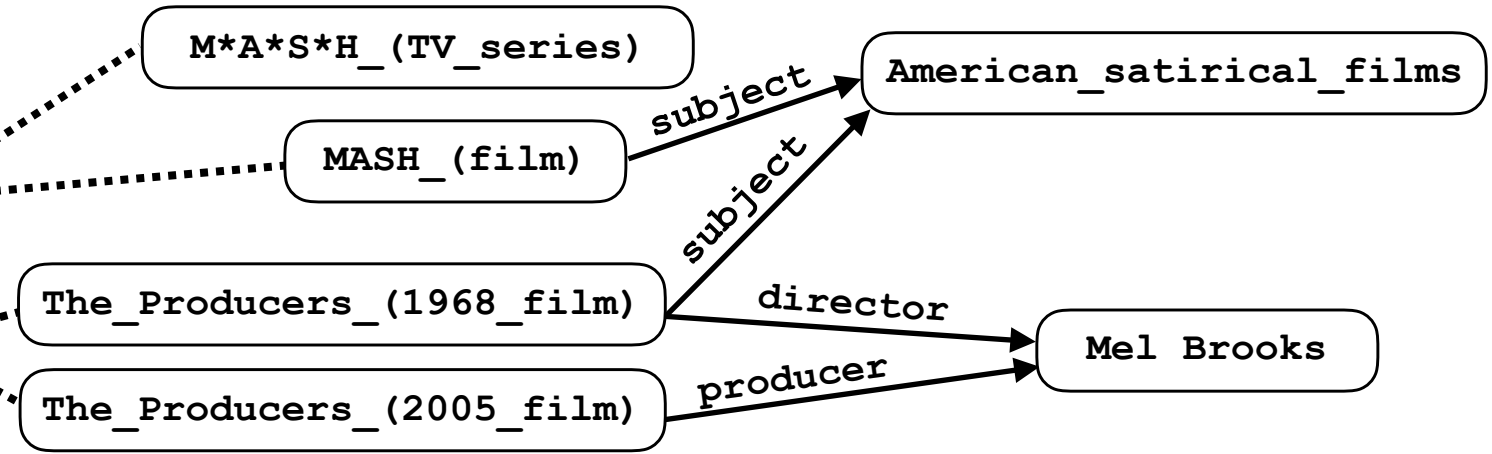
KG Embeddings



KG Completion

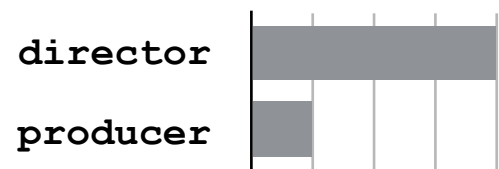
CELL DISAMBIGUATION

Director	Title
Fritz Lang	M
Robert Altman	M*A*S*H
Orson Welles	The Magnificent Ambersons
Mel Brooks	The Producers
Luc Besson	The Professional

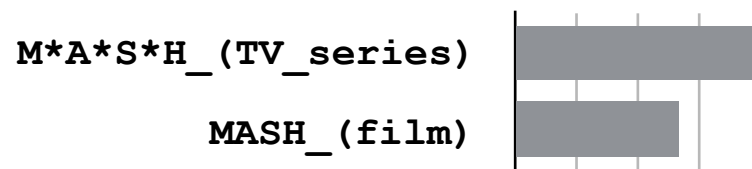


Cell Candidates

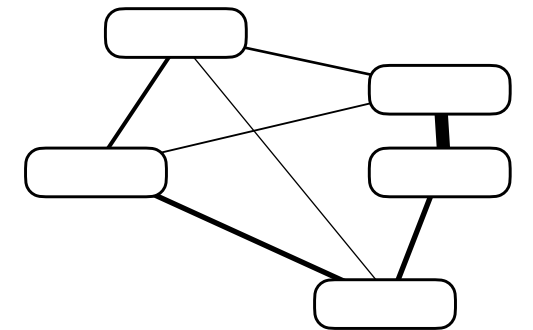
Attributes in the KB



Column-Predicate Scores

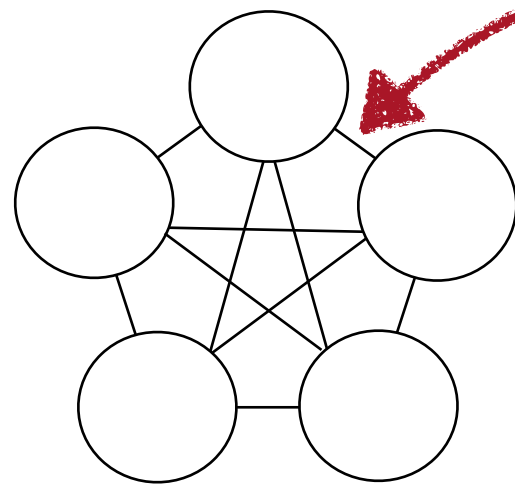


Cell-Candidate Scores



Entity Similarities

CELL DISAMBIGUATION



Loopy Belief Propagation

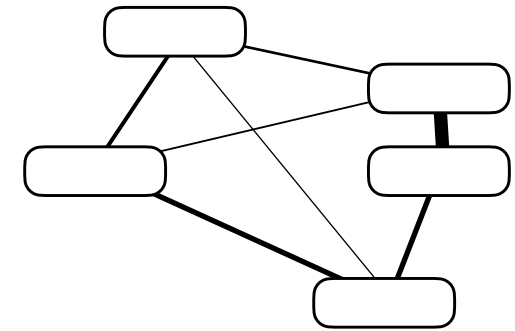
M*A*S*H_(TV_series)

MASH_(film)



Nodes:

Cell-Candidate Scores **L**



Edges:

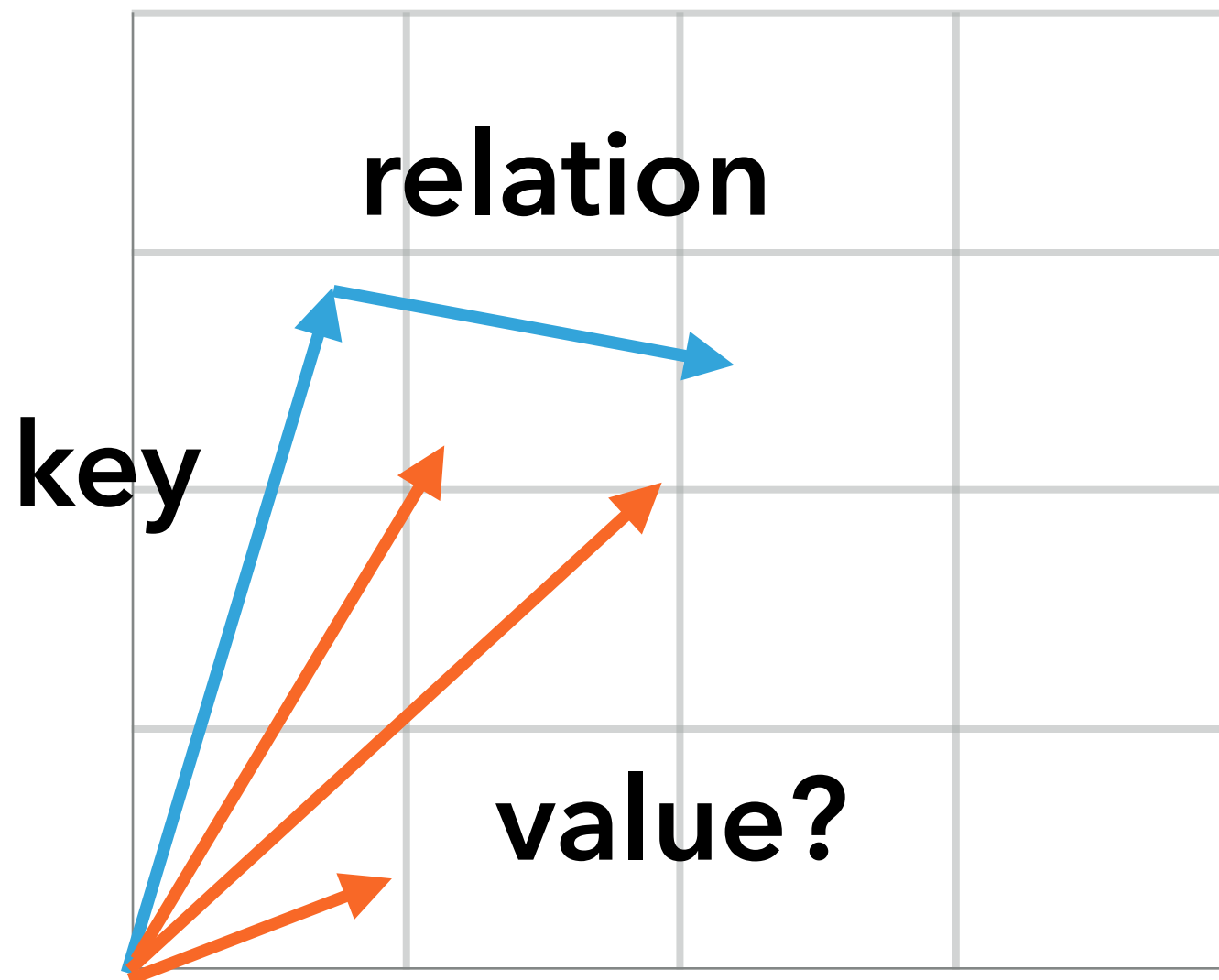
Entity Similarities **S**

$$q = \prod_{\text{row}} \left(\begin{array}{c} \text{Entities} \\ \text{Rows} \\ \mathbf{L} \end{array} \times \begin{array}{c} \text{Entities} \\ \text{Entities} \\ \mathbf{S} \end{array} \right)$$

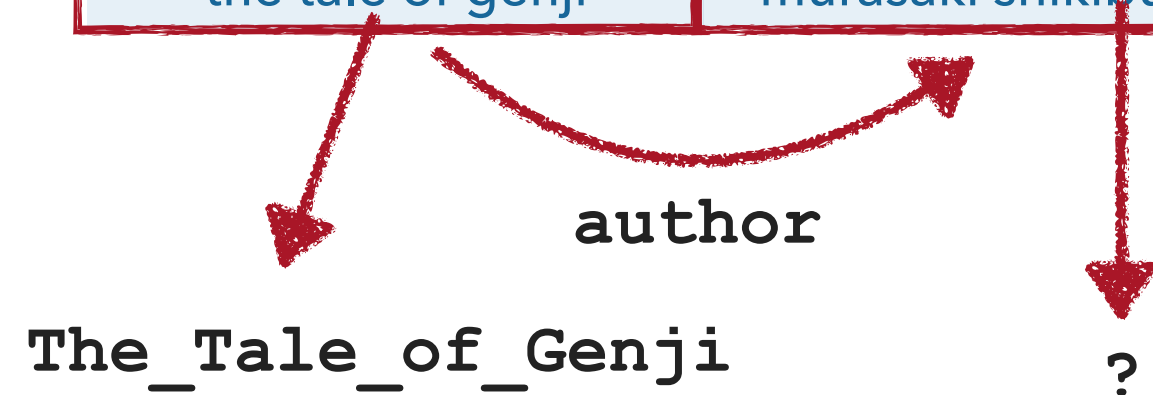
Coherence score: best *cluster* of entities

SLOT-FILLING

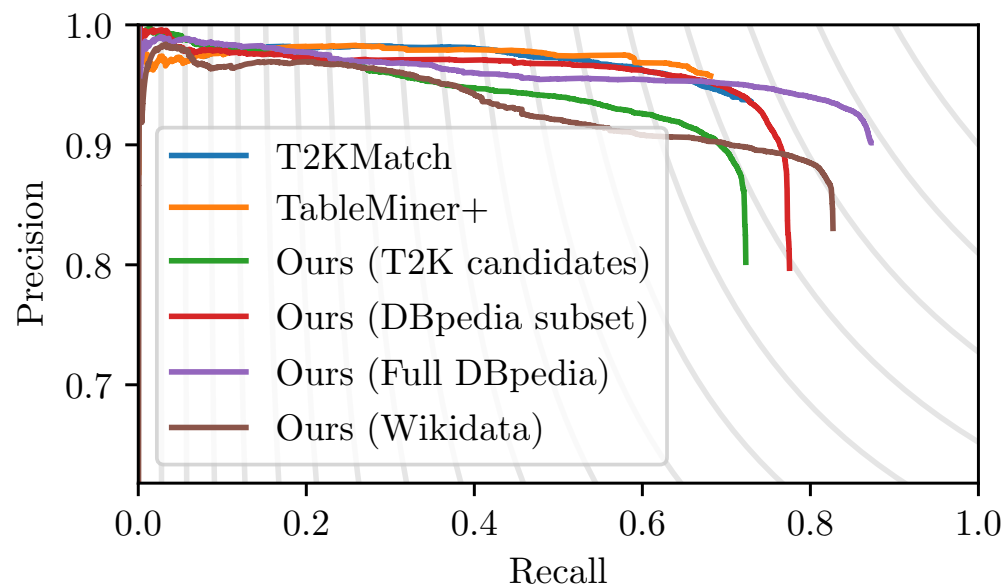
Rank possible disambiguated values by distance



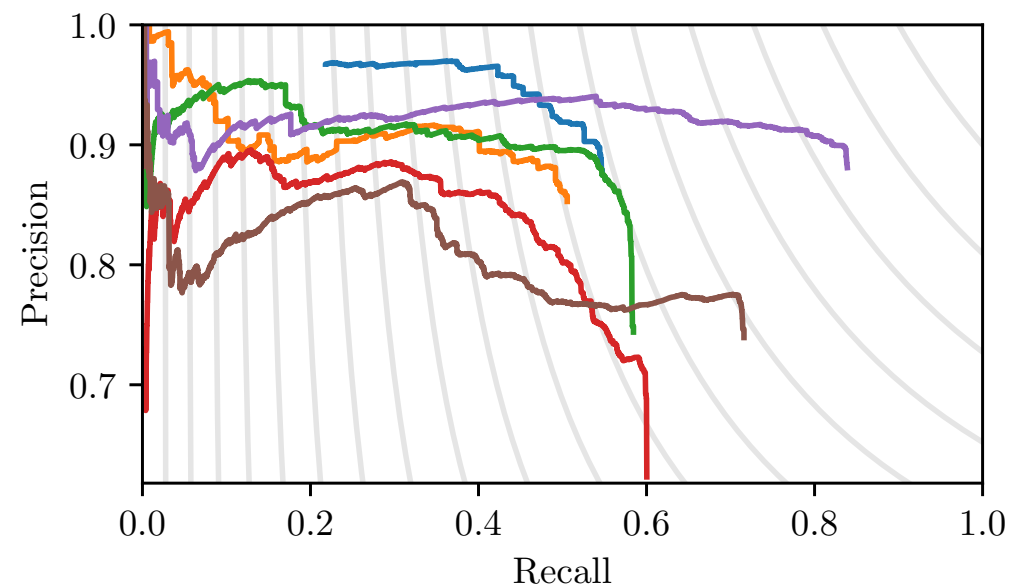
title of great novel	author
don quixote	miguel de cervantes
war and peace	leo tolstoy
ulysses	james joyce
in search of lost time	marcel proust
the brothers karamazov	feodor dostoevsky
moby-dick	herman melville
madame bovary	gustave flaubert
middlemarch	george eliot
the magic mountain	thomas mann
the tale of genji	murasaki shikibu



RESULTS: CELL DISAMBIGUATION



Performance tradeoff, T2D-v2



Performance tradeoff, Webaroo

System	Pr.	Re.	F_1
T2KMatch	.94	.73	.82
TableMiner+	.96	.68	.80
Ours (T2K candidates)	.88	.72	.79
Ours (DBpedia subset)	.90	.76	.83
Ours (Full DBpedia)	.92	.86	.89
Ours (Wikidata)	.87	.82	.84

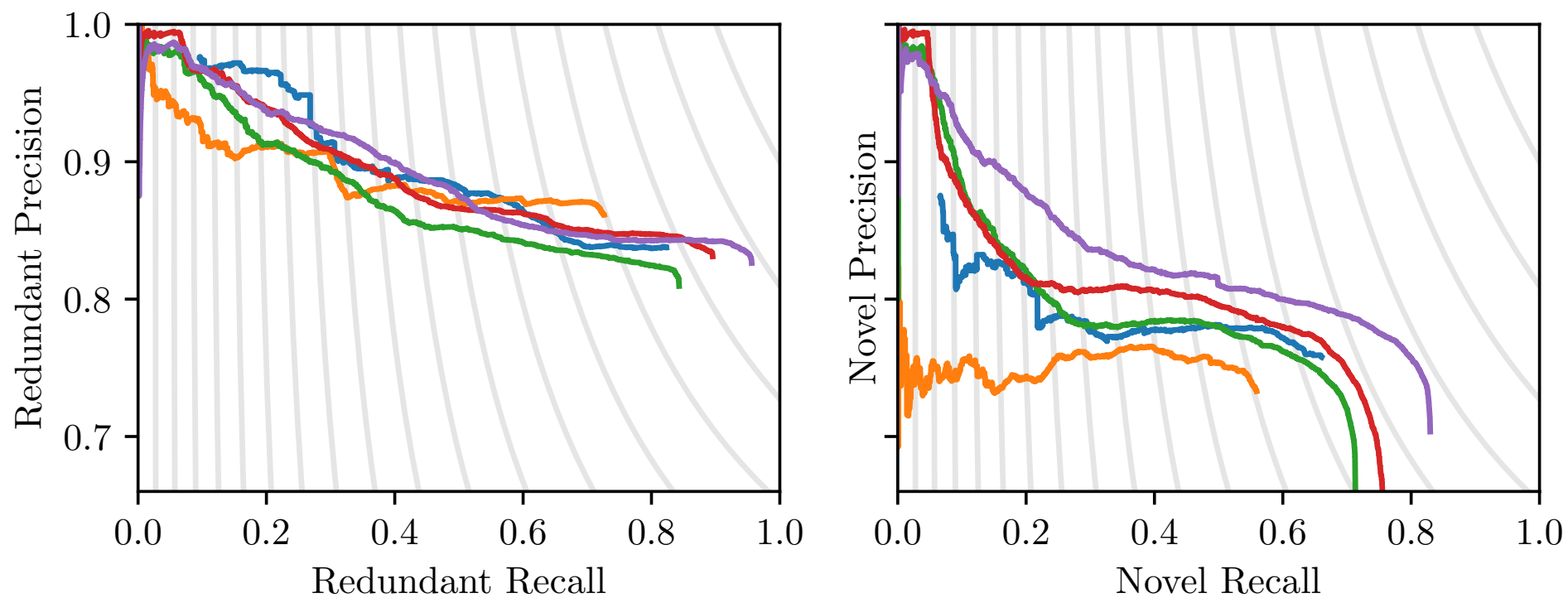
Row-entity evaluation, T2D-v2

System	Pr.	Re.	F_1
T2KMatch	.88	.55	.67
TableMiner+	.85	.51	.63
Ours (T2K candidates)	.74	.58	.65
Ours (DBpedia subset)	.72	.59	.65
Ours (Full DBpedia)	.88	.84	.86
Ours (Wikidata)	.77	.71	.74

Row-entity evaluation, Webaroo

RESULTS: NOVEL FACT EXTRACTION

System	Redundant			Novel		
	Pr.	Re.	F_1	Pr.	Re.	F_1
T2KMatch	.84	.82	.83	.76	.66	.71
TableMiner +	.86	.73	.79	.73	.56	.63
Ours (T2K candidates)	.81	.84	.83	.61	.71	.66
Ours (DBpedia subset)	.83	.90	.86	.59	.76	.66
Ours (Full DBpedia)	.83	.96	.89	.70	.83	.76



RESULTS: SLOT-FILLING ATTRIBUTE DISAMBIGUATION

Ranking	Dataset	Prec@1	Prec@3
Only Label Index (TF-IDF score)	Wikitable	0.37	0.42
	T2D-v2	0.24	0.31
Labels + Embeddings (TransE)	Wikitable	0.61	0.72
	T2D-v2	0.62	0.74

Precision of slot-filling with/out KG embeddings.

Wikitable: 1.6M tables → 786K entity tables → 498K linked

2.8M statements (1.8M slots, 823K known); 307K redundant

CONCLUSION

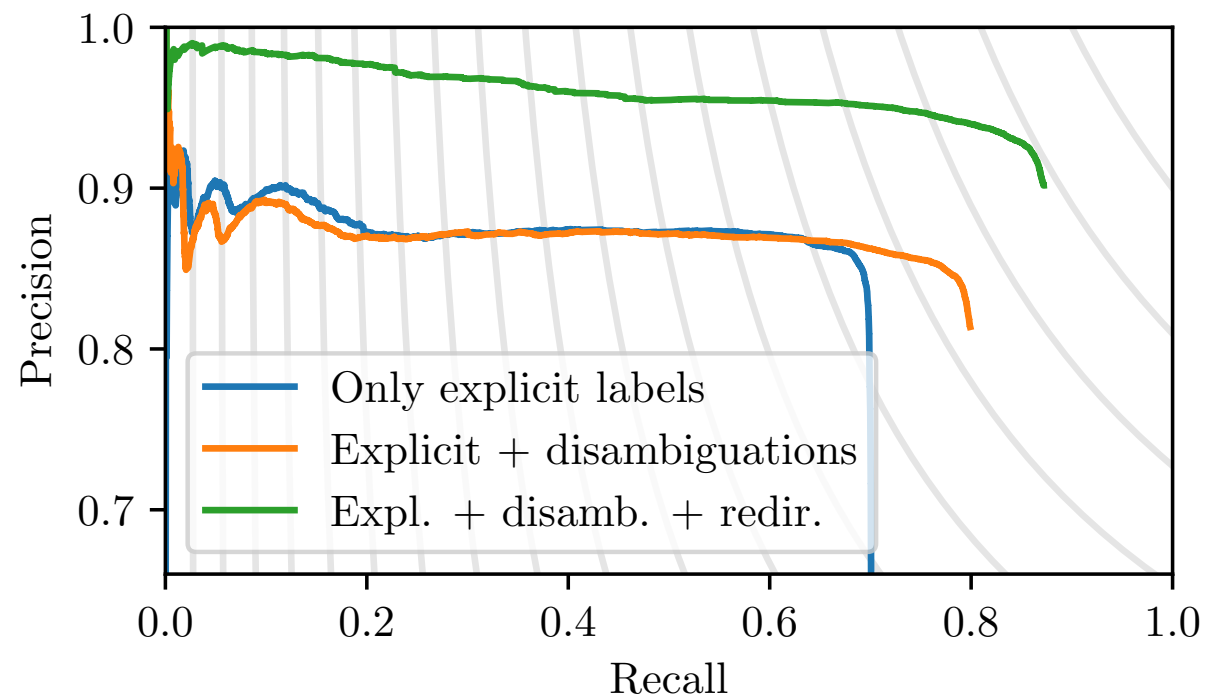
- ▶ **TAKCO: Novelty-oriented table interpretation**
 - ▶ Better disambiguation of incomplete entities
 - ▶ github.com/karmaresearch/takco
- ▶ Evaluation of novel triples provides new perspective
- ▶ Future work:
 - ▶ Use statistics from large corpus for generalisation
 - ▶ Extract n-ary relations

THANK YOU!

LIMITATIONS

party	05-02-08	min	max
unity list	34	17	22
socialist party	60	108	134
social democrats	258	245	255
social liberal party	92	57	70
christian democrats	17	6	12
new alliance		38	72
liberal party	290	245	257
conservative party	103	84	101
danish people's party	133	122	134

RESULTS: ENTITY LINKING - DIFFERENT LABEL INDEX

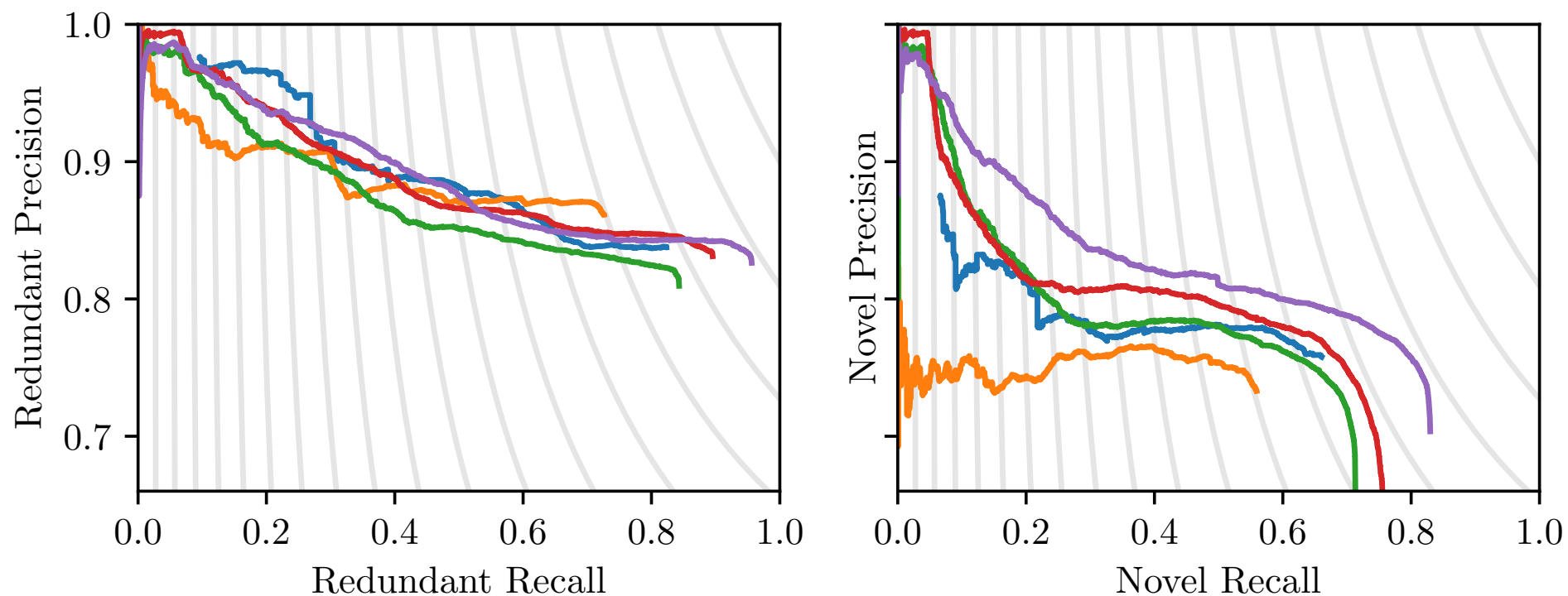


System	Pr.	Re.	F_1
Only explicit labels	.85	.69	.76
Explicit + disambiguations	.84	.79	.81
Expl. + disamb. + redir.	.92	.86	.89

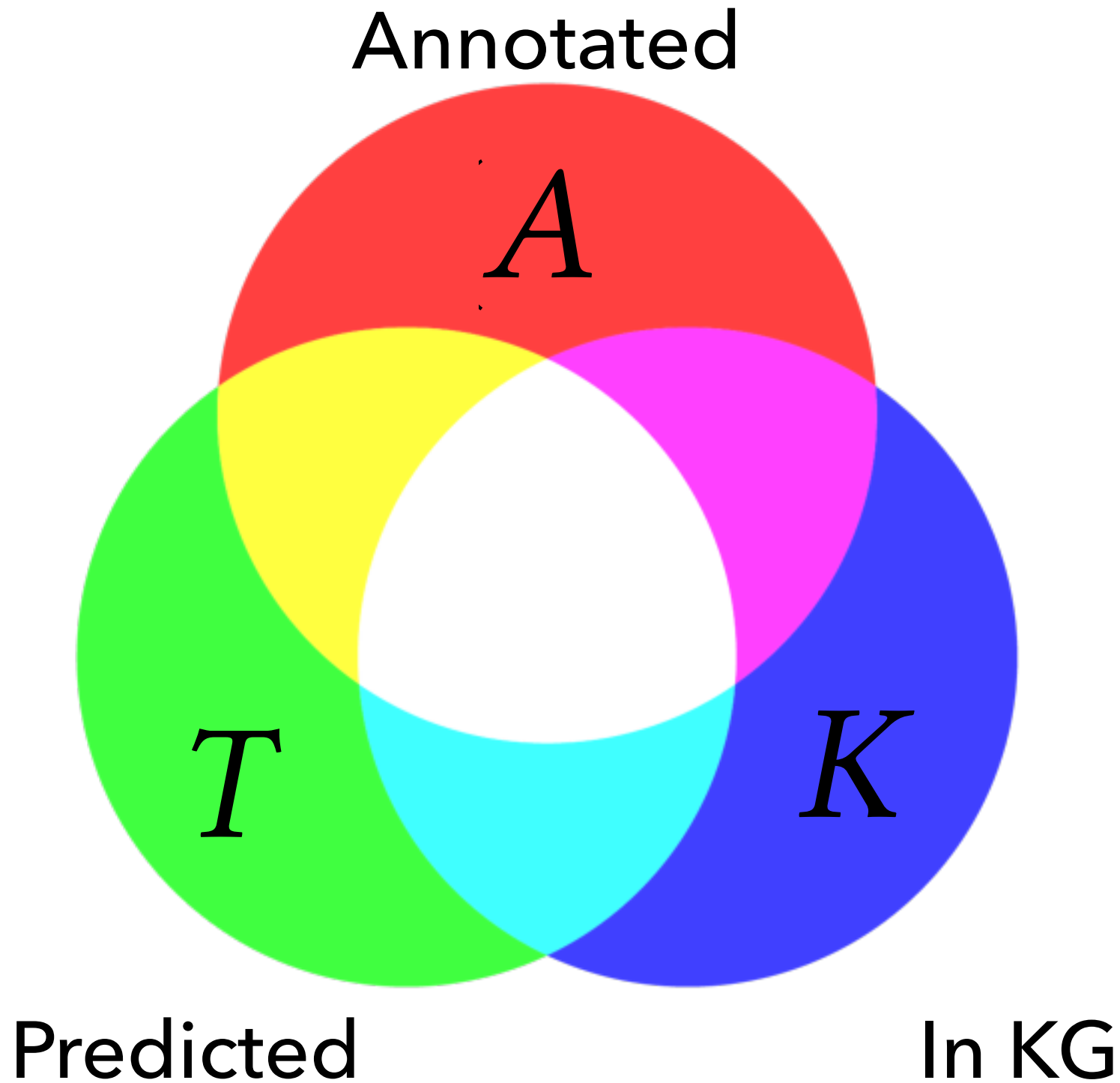
Label set is very important!

RESULTS: NOVEL FACT EXTRACTION

System	Redundant			Novel			R^+	R^-
	Pr.	Re.	F_1	Pr.	Re.	F_1		
T2KMatch	.84	.82	.83	.76	.66	.71	0.55	0.44
TableMiner +	.86	.73	.79	.73	.56	.63	0.57	0.39
Ours (T2K candidates)	.81	.84	.83	.61	.71	.66	0.55	0.44
Ours (DBpedia subset)	.83	.90	.86	.59	.76	.66	0.55	0.31
Ours (Full DBpedia)	.83	.96	.89	.70	.83	.76	0.55	0.25



REDUNDANCY OF EXTRACTED FACTS



Positive Redundancy

$$R_p(T, A) = \frac{|(A \cap T) \cap K|}{|A \cap T|}$$

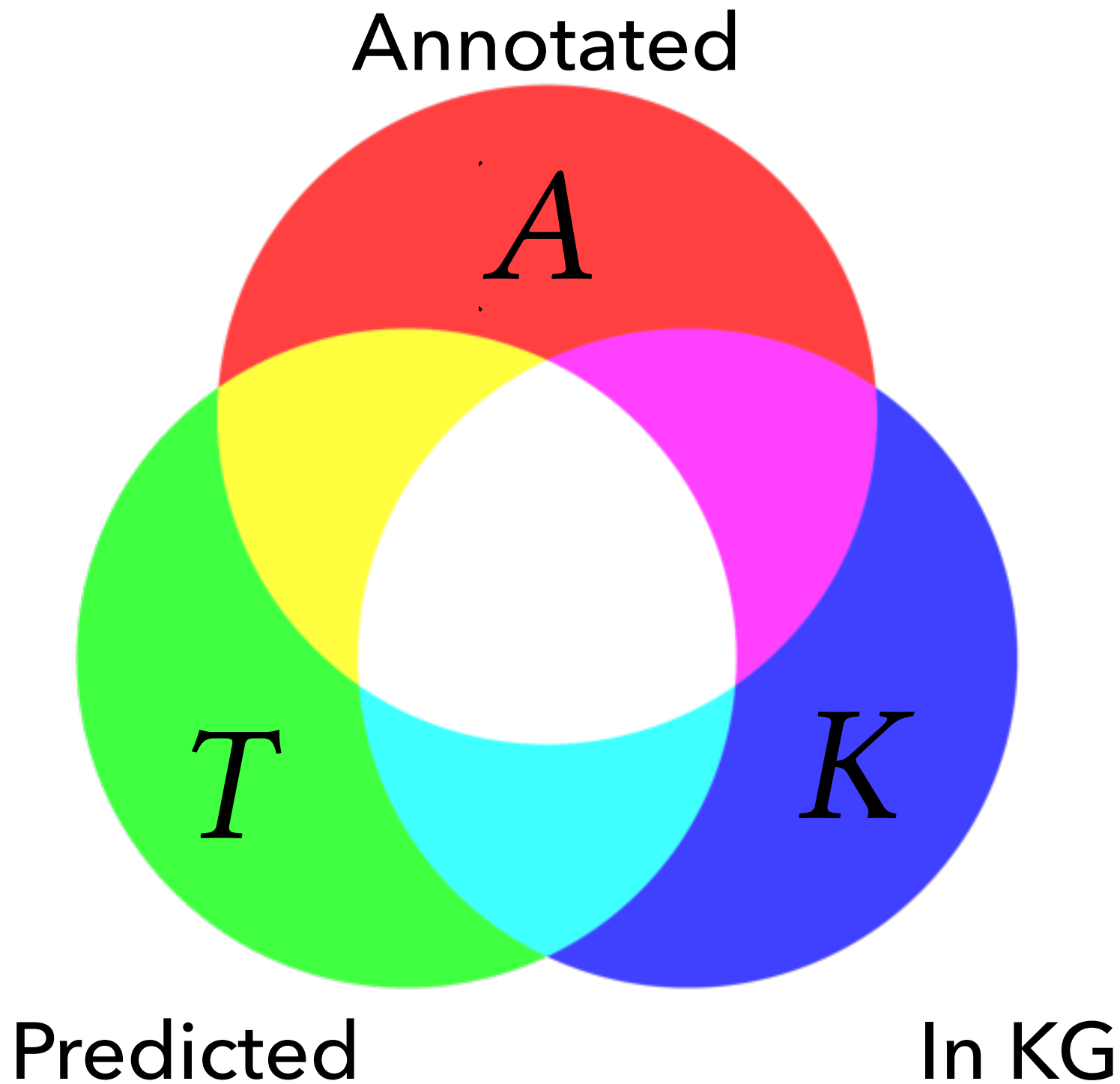
Negative Redundancy

$$R_n(T, A) = \frac{|(A \setminus T) \cap K|}{|A \setminus T|}$$

Hypothesis:

$$R_p(T, A) > R_n(T, A)$$

RECALL OF EXTRACTED FACTS



Novel Recall

$$Q_n(T, A) = \frac{|T \cap (A \setminus K)|}{|A \setminus K|}$$

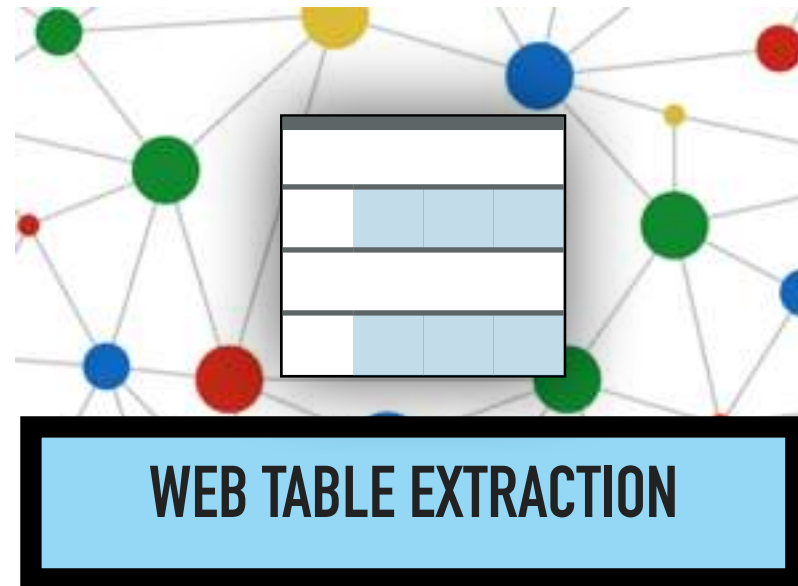
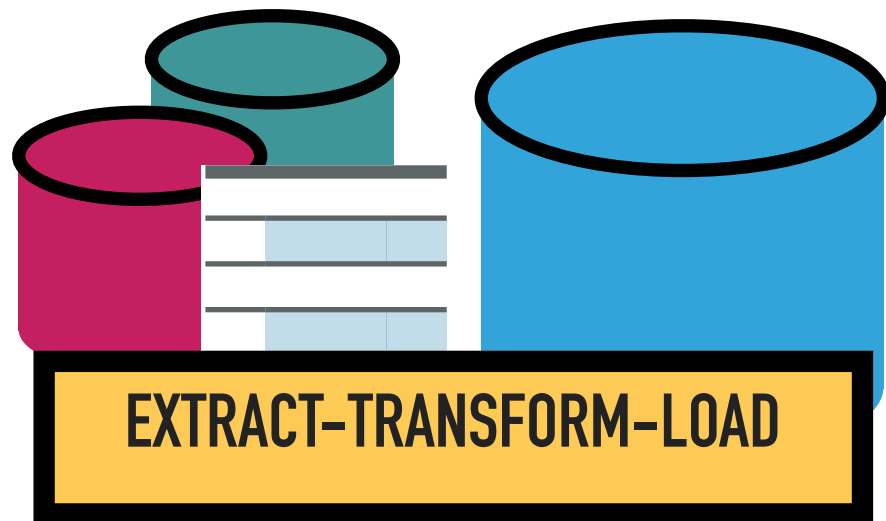
Redundant Recall

$$Q_r(T, A) = \frac{|T \cap (A \cap K)|}{|A \cap K|}$$

Hypothesis:

$$Q_n(T, A) < Q_r(T, A)$$

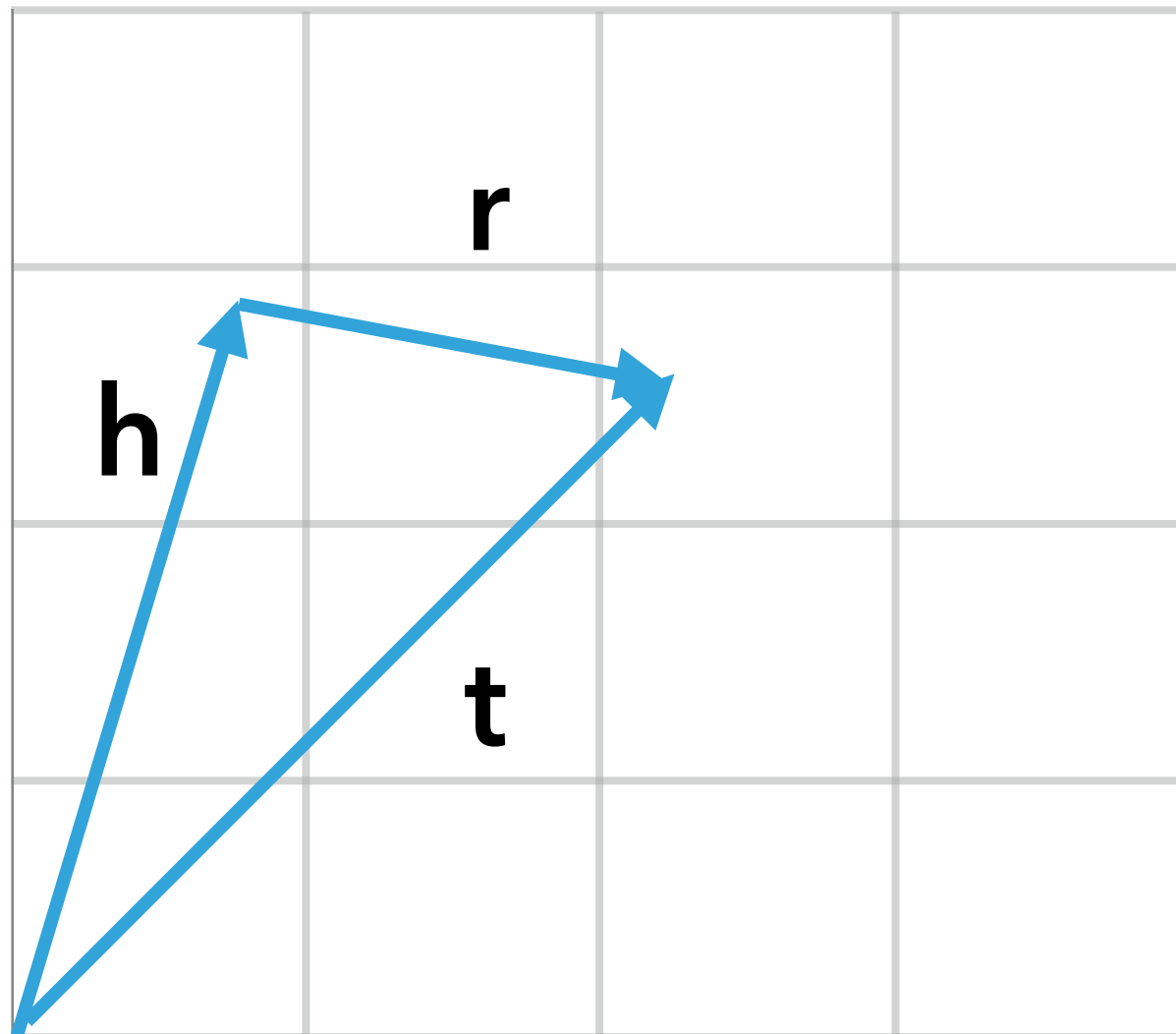
DATA INTEGRATION



- ▶ Structured sources
- ▶ Closed domain
- ▶ Writing queries
- ▶ Enterprise solutions
- ▶ Semi-structured sources
- ▶ Large domain ontology
- ▶ KG matching
- ▶ Slot filling candidates
- ▶ Unstructured sources
- ▶ Open domain
- ▶ Machine Learning
- ▶ Approximate predictions

METHOD: SLOT-FILLING

Entity Embeddings: TransE



FEATURES FOR MAKING PREDICTIONS

- ▶ Labels
- ▶ Explicit facts
- ▶ Ontology
 - ▶ hierarchy of types
 - ▶ range and domain of relations
- ▶ Statistics
 - ▶ "soft" ontology: empirical observations
- ▶ External data



Model
Knowns



Model
Unknowns

EVALUATION

- ▶ Manually annotated tables
 - ▶ ... but different data sources,
 - ▶ ... and different knowledge bases
 - ▶ ... and different features
- ▶ Measure precision & recall of predictions
 - ▶ ... but is this most useful for slot filling?