APEX 数据和知识管理实验室
DATA & KNOWLEDGE MANAGEMENT LAB

An Effective Rule Miner for Instance Matching in a Web of Data

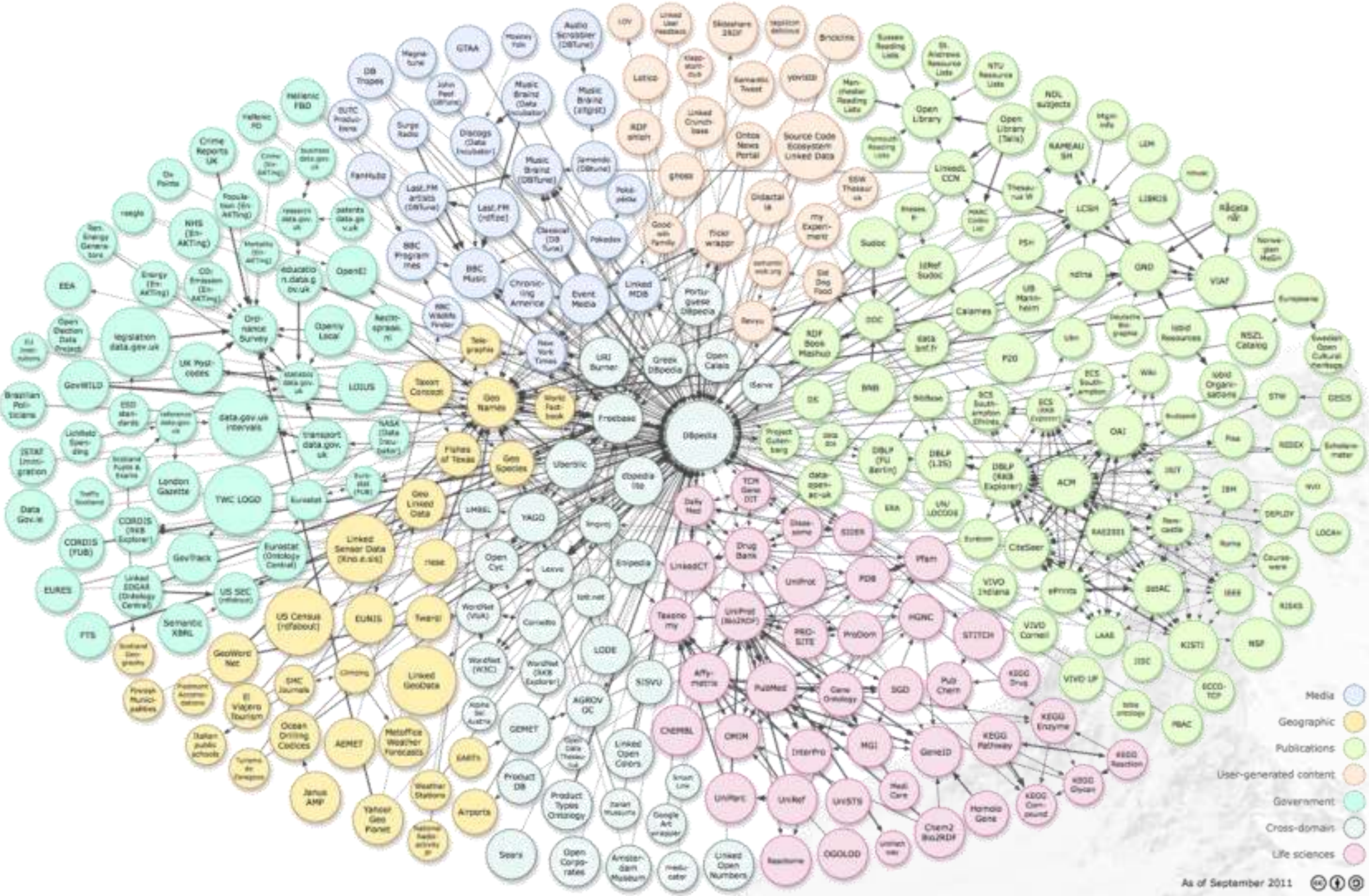
Xing Niu, Shu Rong, Haofen Wang and Yong Yu
Shanghai Jiao Tong University

2012.10.31 @ CIKM

Agenda

- Introduction
- Workflow
- Experiments
- Summary

Introduction - the Linking Open Data Project



Introduction - Equivalent Instances

- DBpedia

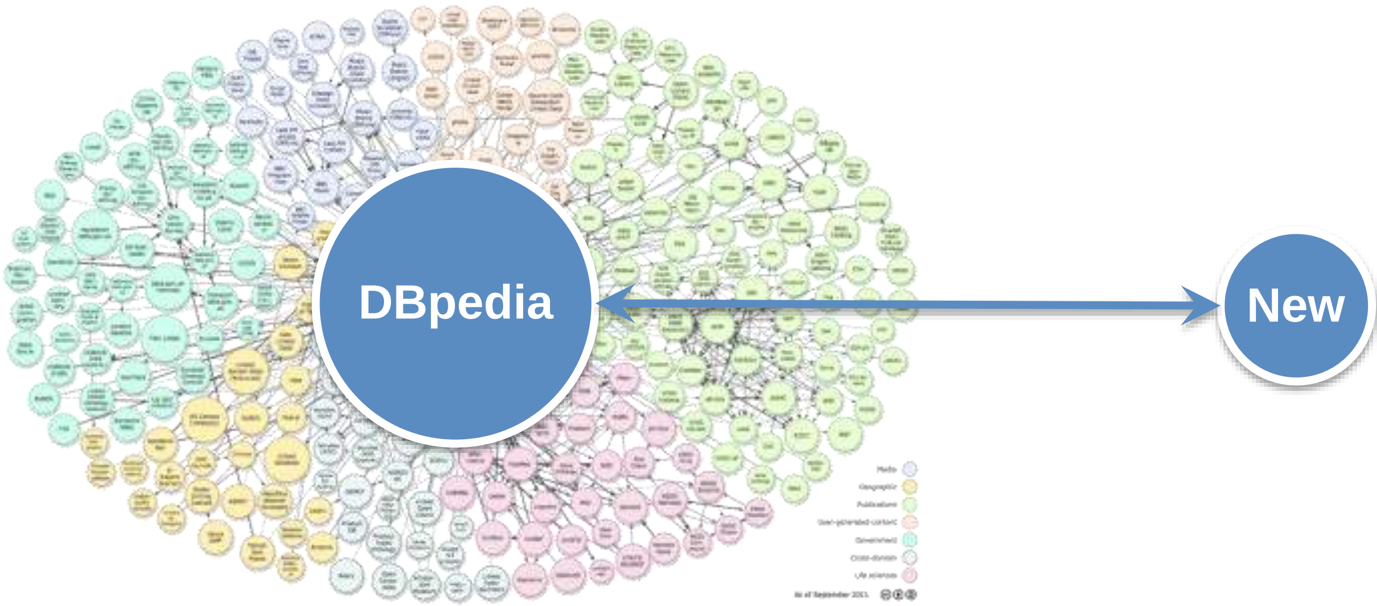
dbpedia:Nene_(bird)	
foaf:name	“Nene”
dbpprop:binomial	“Branta sandvicensis”
dbpedia-owl:phylum	dbpedia:Chordate
…	

- GeoSpecies

gs:nene	
gs:hasCommonName	“Nene”
gs:hasCanonicalName	“Branta sandvicensis”
gs:inPhylum	gs:Chordate
…	

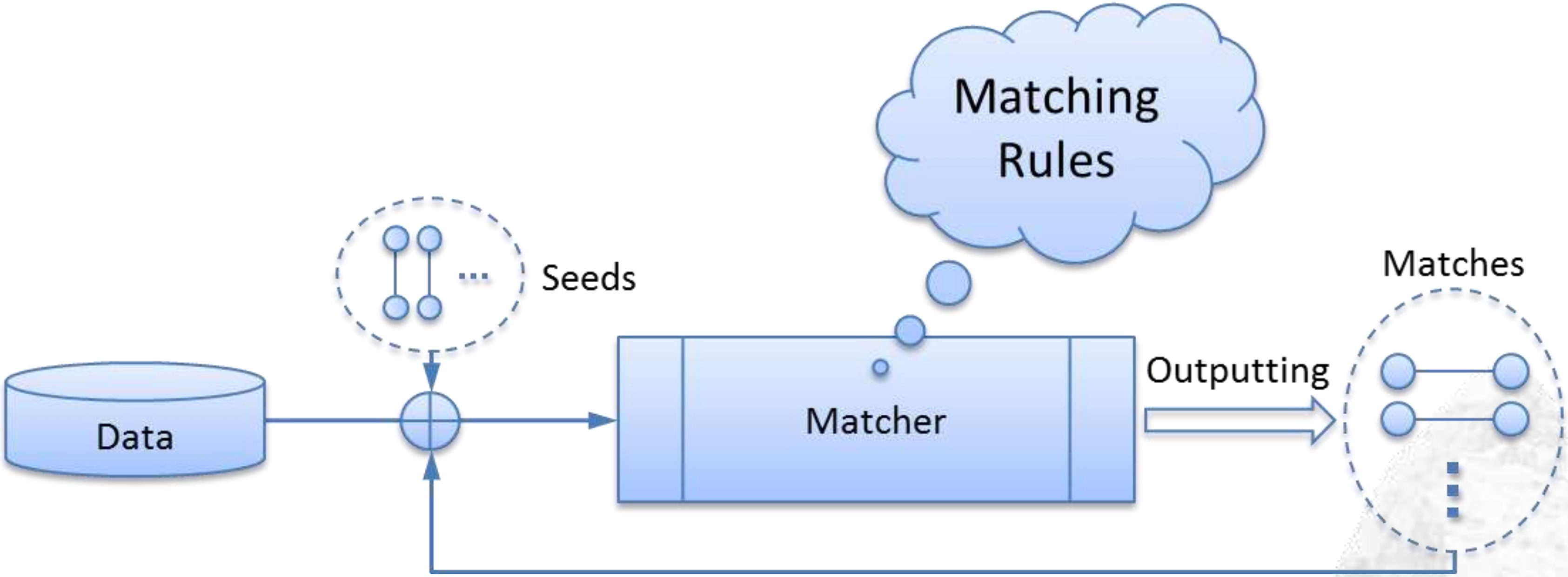
*photo by Brenda Zaun (form Wikipedia)

Introduction – Problem Definition



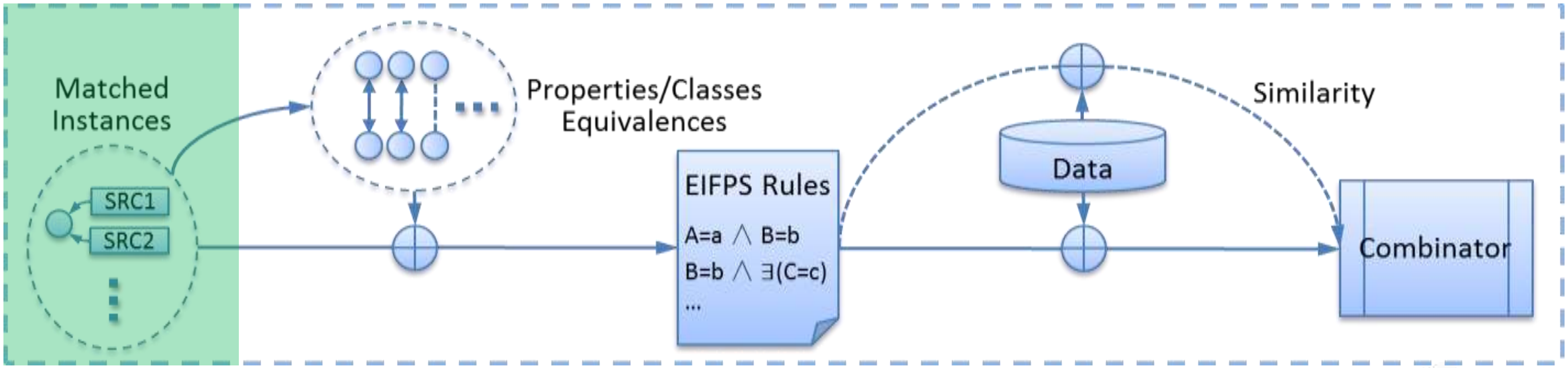
- Domain/dataset-specific methods
- General-purpose approaches
  - Requiring manually defined matching rules (i.e. link specifications)
  - Focusing on similarity metrics

Introduction – Our Solution



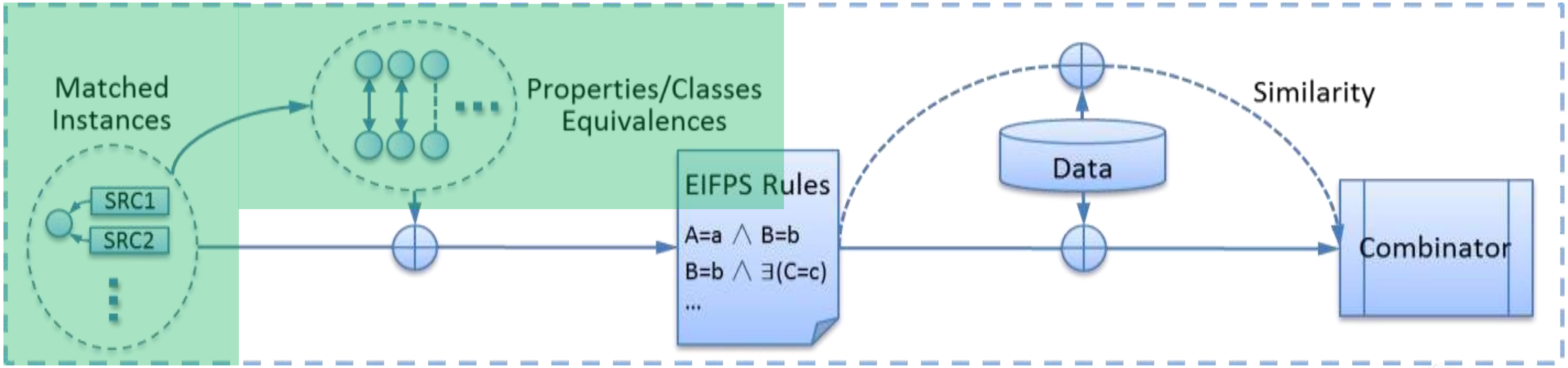
- Automatically discovering and refining dataset-specific matching rules in iterations
  - Deriving these rules by finding the most discriminative data characteristics for a given data source pair.

Workflow – Pre-processing



- For each pair of existing matched instances, their property-value pairs are merged.
- E.g. dbpedia:Nene_(bird) = gs:nene
  - foaf:name:"Nene"
  - dbpprop:binomial:"Branta sandvicensis"
  - dbpedia-owl:phylum:dbpedia:Chordate
  - gs:hasCommonName:"Nene"
  - gs:hasCanonicalName:"Branta sandvicensis"
  - gs:inPhylum:gs:Chordate

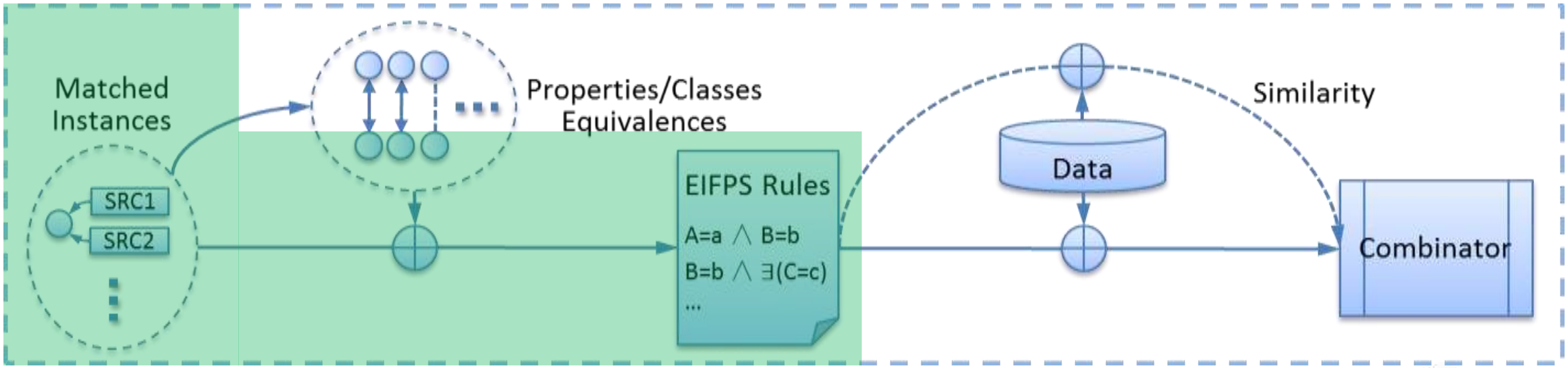
Workflow – Mining Properties Equivalences



- Statistical Schema Induction [Völker et.al., ESWC2011]
- E.g. dbpedia:Nene_(bird) = gs:nene

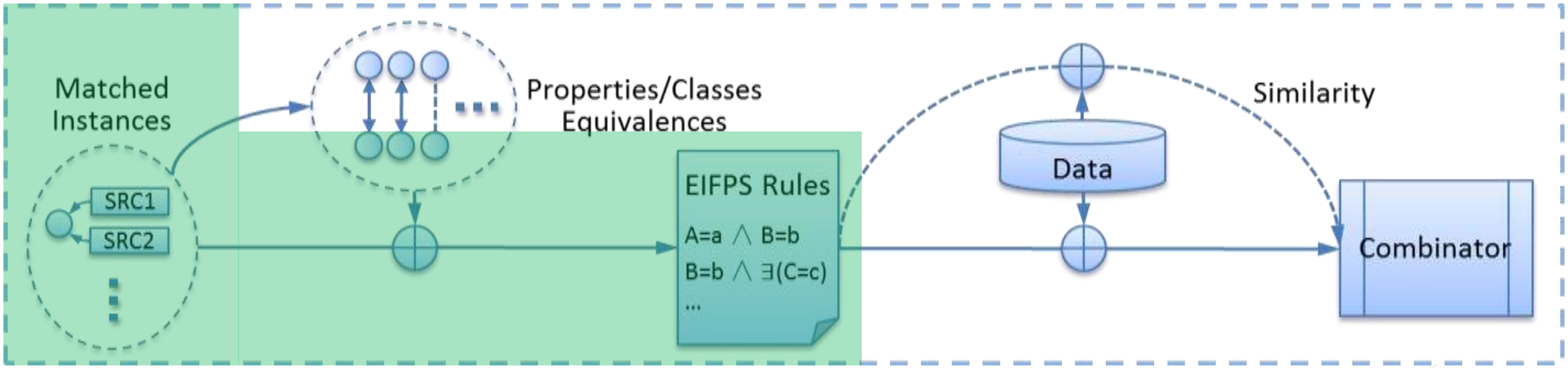
Values	Property_1	Property_2
“Nene”	foaf:name	gs:hasCommonName
“Branta sandvicensis”	dbpprop:binomial	gs:hasCanonicalName
“Panda”	foaf:name	gs:hasCommonName
“Coconut”	foaf:name	gs:hasCommonName
…	…	…

Workflow – Mining Matching Rules



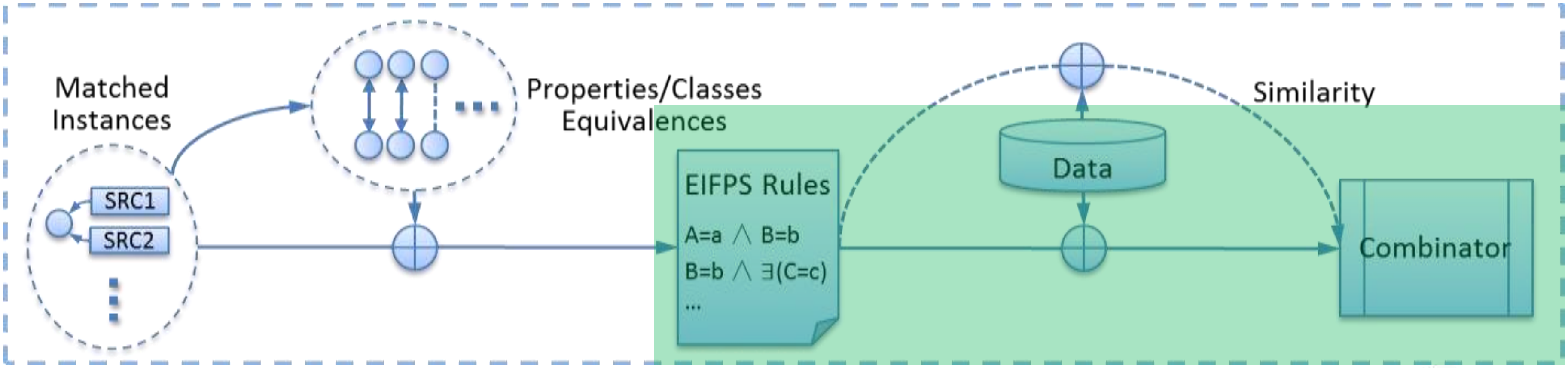
- Association rule mining, agian.
- E.g. for the transaction dbpedia:Nene_(bird) = gs:nene, it has items:
  - valueOf(foaf:name) = valueOf(gs:hasCommonName)
  - valueOf(dbpprop:binomial) = valueOf(gs:hasCanonicalName)
  - valueOf(dbpedia-owl:phylum) = valueOf(gs:inPhylum)
  - …

Workflow – Mining Matching Rules (con't)



- Matching rule:
  - dbpedia:x and gs:x are matched, iff.
  - valueOf(foaf:name) = valueOf(gs:hasCommonName)
  - and
  - valueOf(dbpprop:binomial) = valueOf(gs:hasCanonicalName)
  - and
  - valueOf(dbpedia-owl:phylum) = valueOf(gs:inPhylum)
- Each matching rule brings a confidence value with it. We will introduce it later.

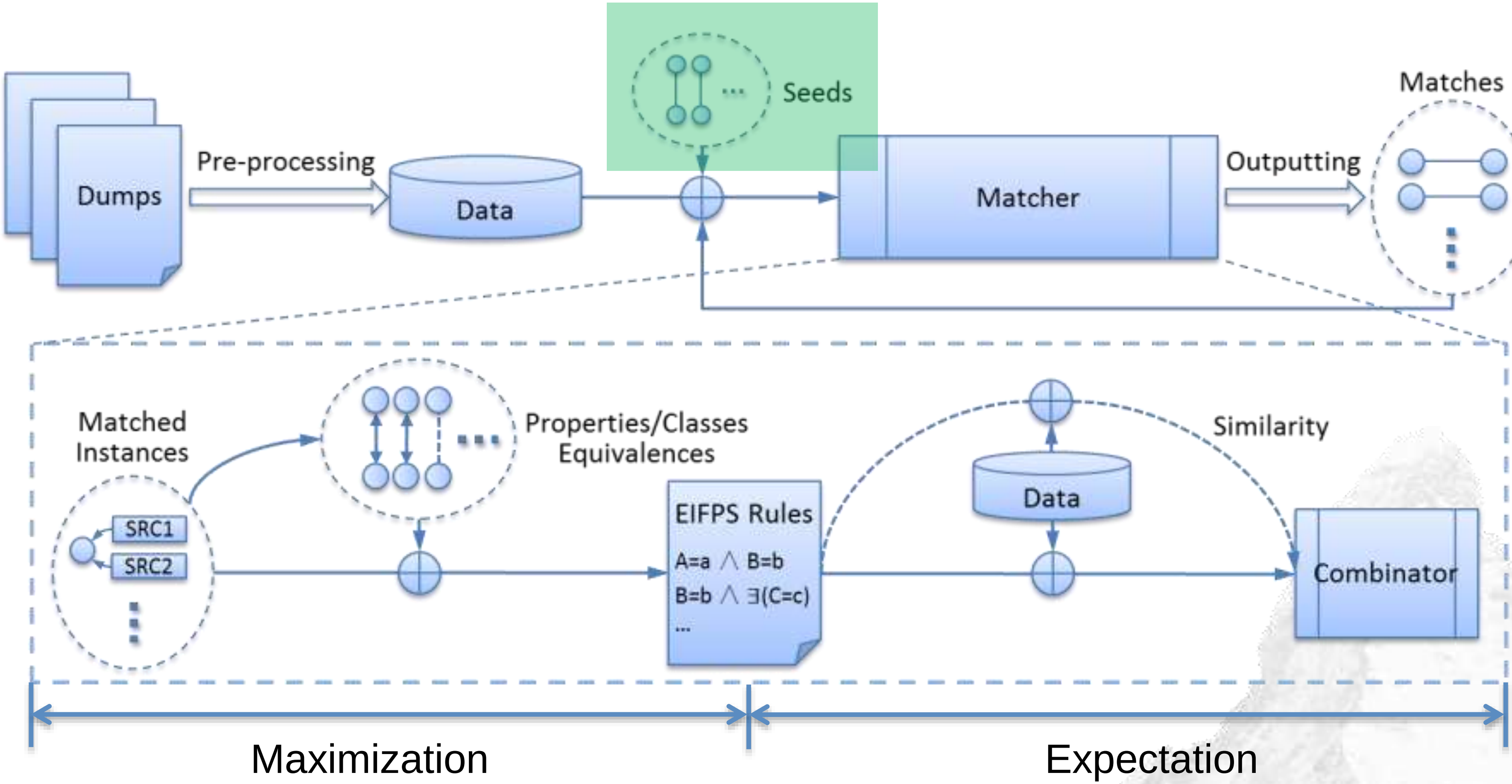
Workflow – Generating Matches



- Applying the obtained rule(s) on the unlabeled data to generate matches' candidates.
  - Each candidate also has a confidence value which equals to the confidence of its corresponding rule.

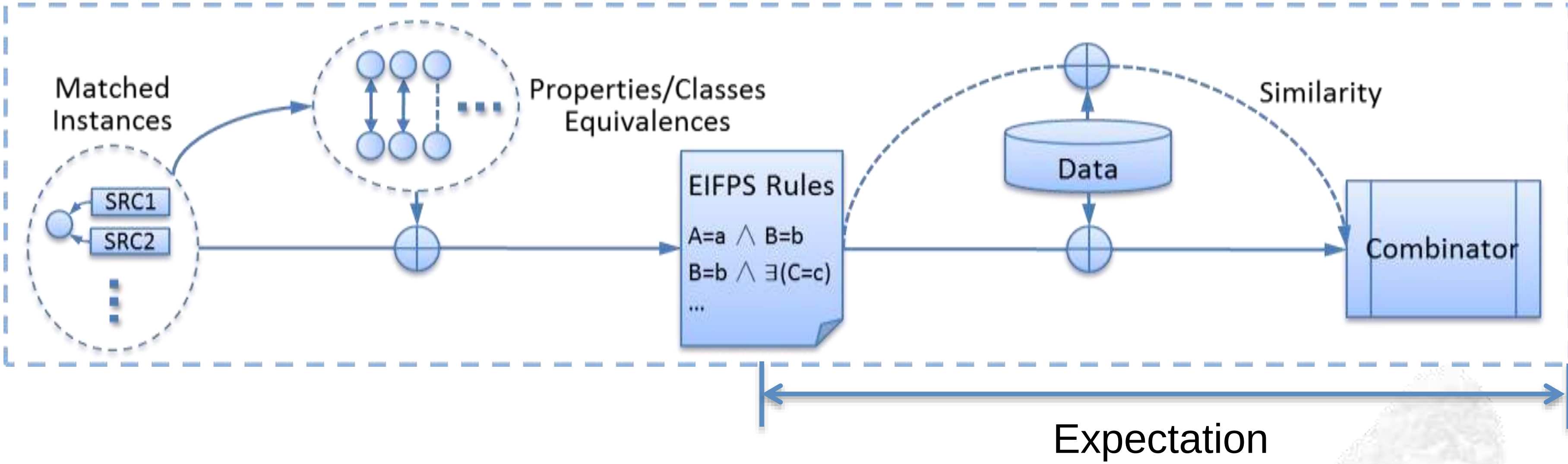
- The combiner is used to combine confidence values of a match's candidate.
  - Dempster's rule [G. Shafer, 1976]

Workflow – the Wrapper Algorithm



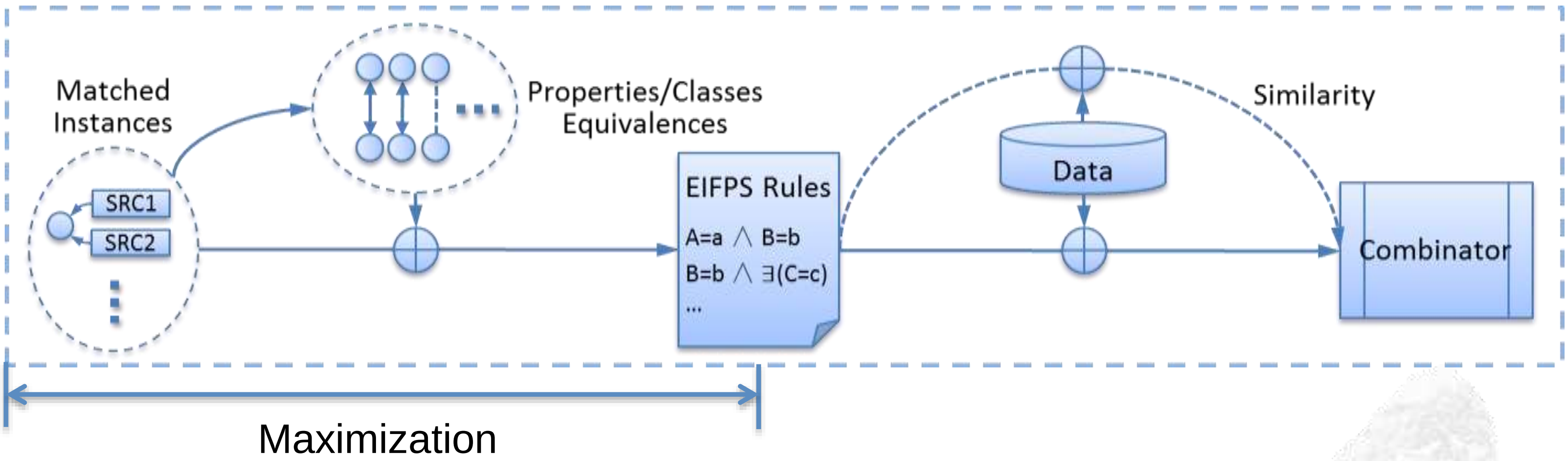
- The wrapper is an implementation of Expectation-Maximization iterations.
  - Refining matching rules, discovering new matches

Workflow – E-step



- The E-step estimates the missing data using the observed data and the current estimate for the parameters.
- The E-step estimates the matches using the current estimate for the matching rules.

Workflow – M-step



- The M-step computes parameters maximizing the likelihood function as the data estimated in E-step are used in lieu of the actual missing data.
  - *M*: matches
  - *Θ*: parameters

$$L(\boldsymbol{\theta}; \boldsymbol{M}) = \Pr(\boldsymbol{M}|\boldsymbol{\theta}).$$

Workflow – The Likelihood Function

- For a set of given matches, the proximity is reflected in two aspects:
  - correctness (precision)
  - completeness (recall)

- Without complete reference matches (in real-world data), it is difficult to precisely evaluate either of them.
- An alternative measurement: optimizing the precision takes priority and obtaining all potential matches on the premise of that precision value.

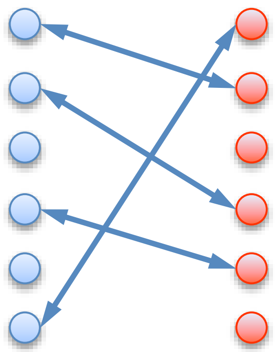
- The likelihood function can be continued as:

$$L(\boldsymbol{\theta}; \boldsymbol{M}) \approx \mathrm{Precision}(\boldsymbol{M} | \boldsymbol{\theta}).$$

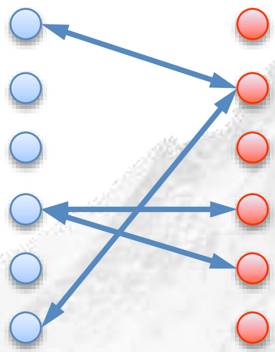
Workflow - Approximate Precision

$$\frac{|\mathrm{ConnectedComponent}(M)|}{|\mathrm{Edge}(M)|}$$

- Assuming that no equivalent instances exist in a single data source, we can infer that an instance is equivalent to at most one another from the other data source.
- Incorrect matches in *M* may result in a node connecting to more than one other node, which is contrary to the assumption.



P=4/4=1



P=2/4=0.5

Experiments - Datasets

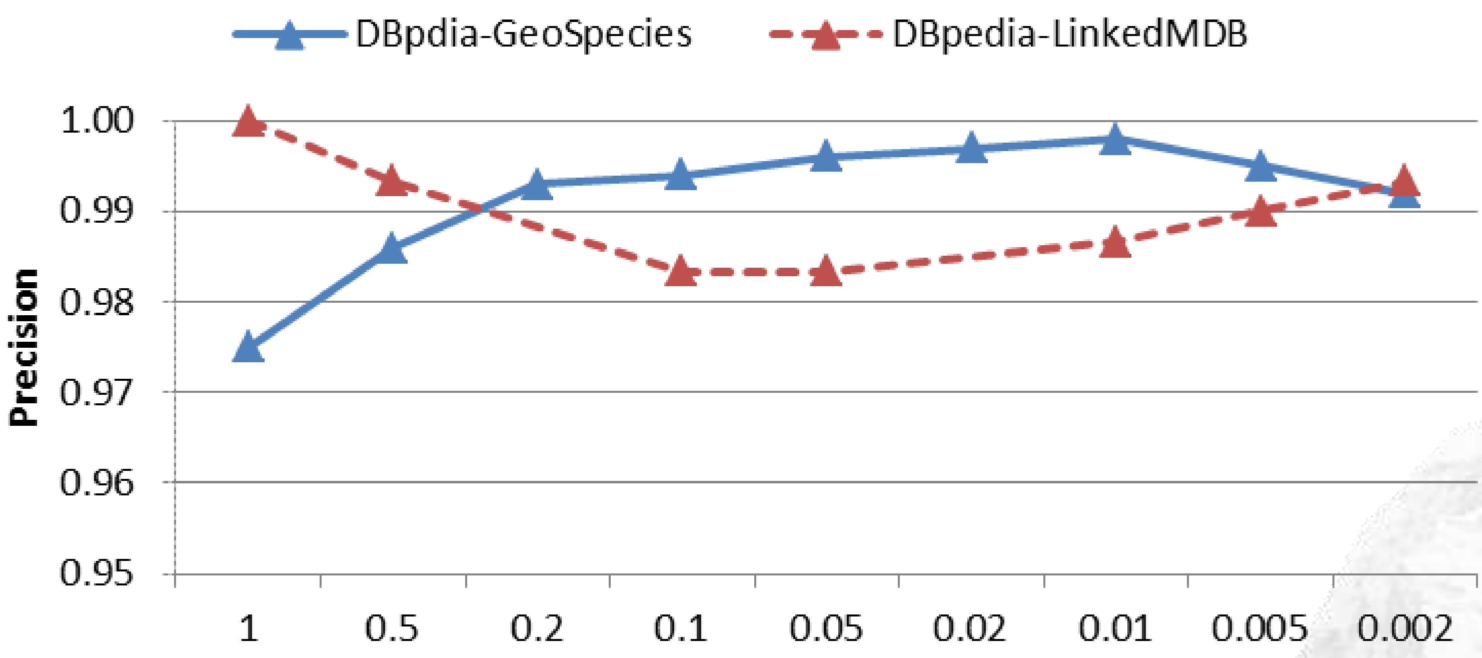
- **DBpedia** is a hub data source in LOD. It structures Wikipedia knowledge and make this structured information available on the Web.
- **GeoNames, LinkedMDB and GeoSpecies**

- **Task**: discovering matches between DBpedia and the other three domain-specific data sources.

- Statistics

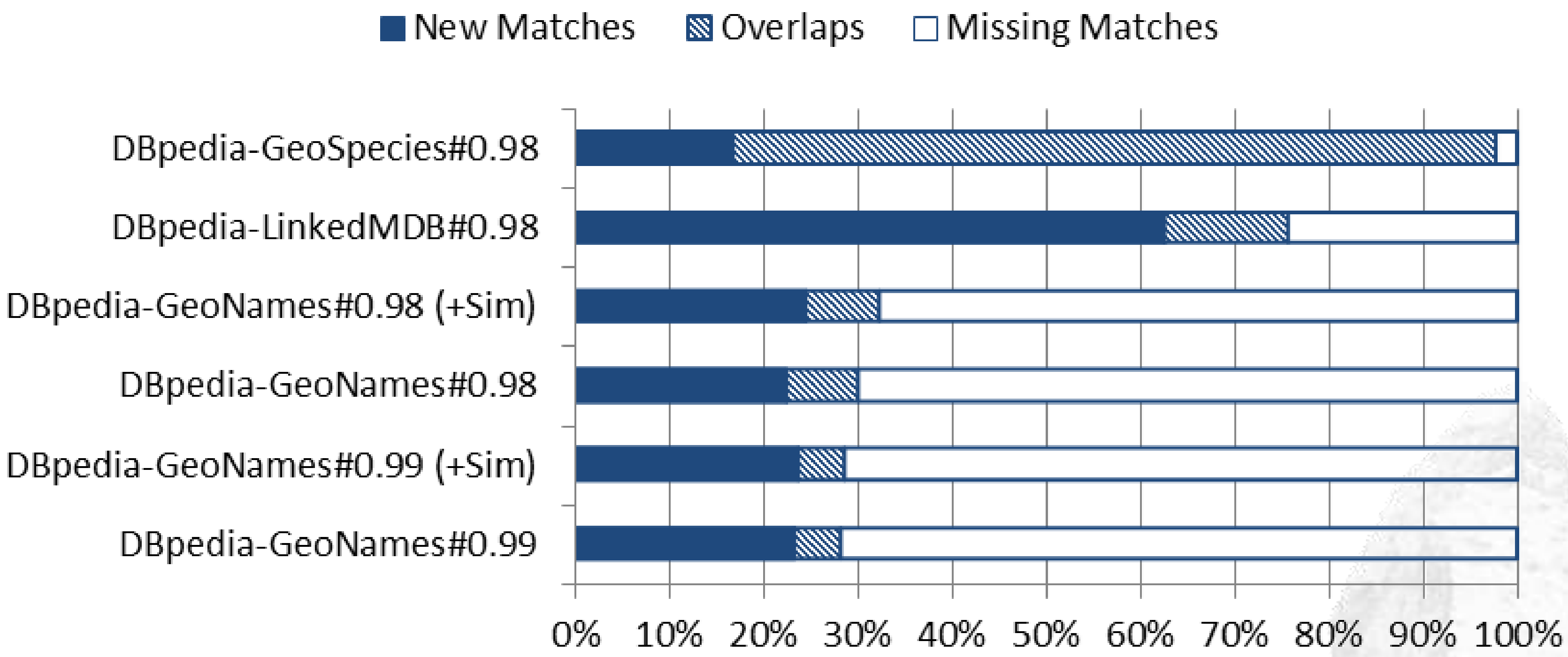
Datasets	Instances	References	Baselines
DBpedia	4,071,600	-	-
GeoNames	8,147,136	317,433	manually
LinkedMDB (Film)	97,471	16,447	ODDLinker
GeoSpecies	20,939	11,490	unknown

Experiments – Precisions



- Sampling a certain number of output matches.
- The X-axis indicates the proportions of selected seeds in complete reference matches.

Experiments – Newly Found Matches



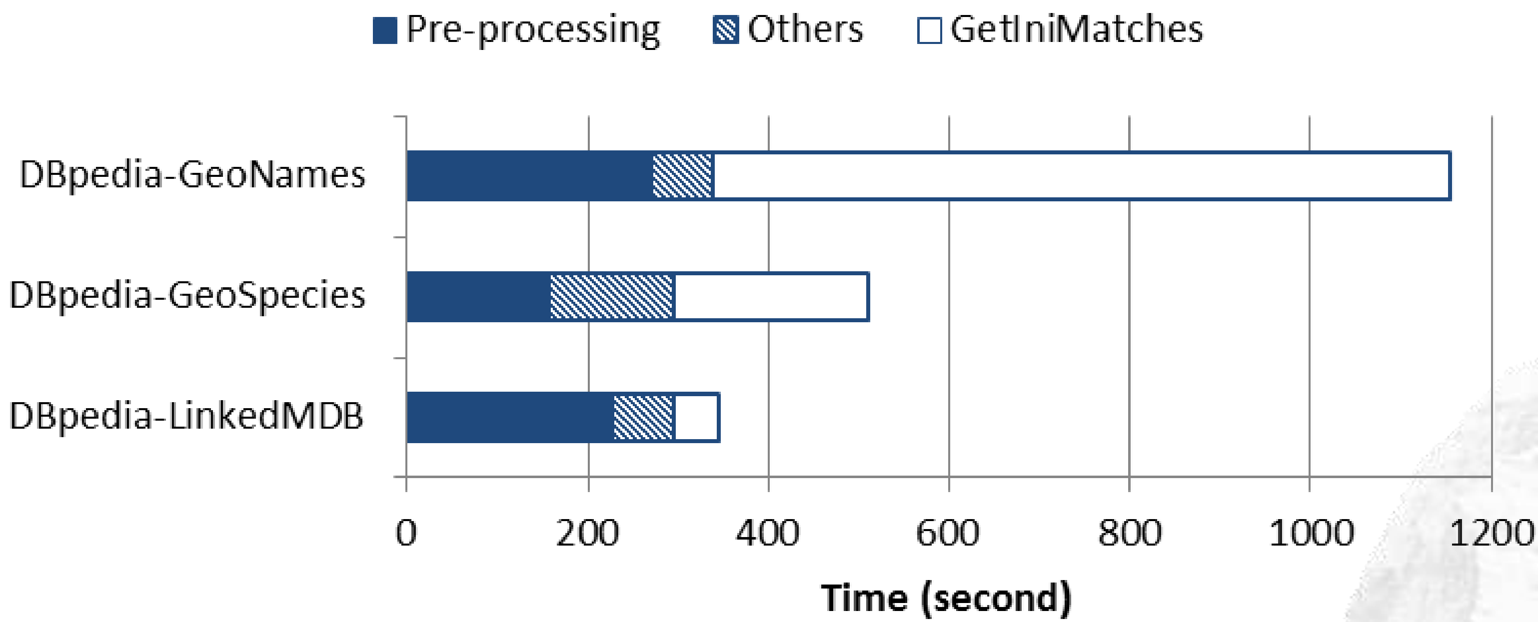
- The match space constituted by reference matches and newly found matches

Experiments – Running Time

- Java@Map/Reduce
  - The Hadoop cluster contains 40 nodes.
  - Each node is a PC (Intel Core 2 Quad 2.66GHz CPU, 2GB RAM) can run 3 Maps + 3 Reduces simultaneously.
  - This is a shared cluster and we occupy 50 slots.

- We sample some typical running times for a single iteration.

Experiments – Running Time (con't)



- The most time-consuming phases are "data pre-processing" and "getting initial matches".

Summary

- Contributions
  - We proposed a general-purpose approach to automatically mine dataset-specific matching rules based on the EM algorithm.
  - We introduced a graph-based metric to estimate likelihood (precision) and Dempster's rule to combine confidence values.
  - *We discussed some extensions to our approach in order to fit the different requirements of various practical applications.

- Conclusions
  - We carried out experiments on several real-world datasets. The results demonstrated the correctness of matches discovered by our approach (precision >0.96 in most cases).
  - We also shown more matches are found than existing references.
  - The whole process can be implemented parallel.

Thanks!