Learning to Map between Ontologies on the Semantic Web

AnHai Doan, Jayant Madhavan,
Pedro Domingos, and Alon Halevy

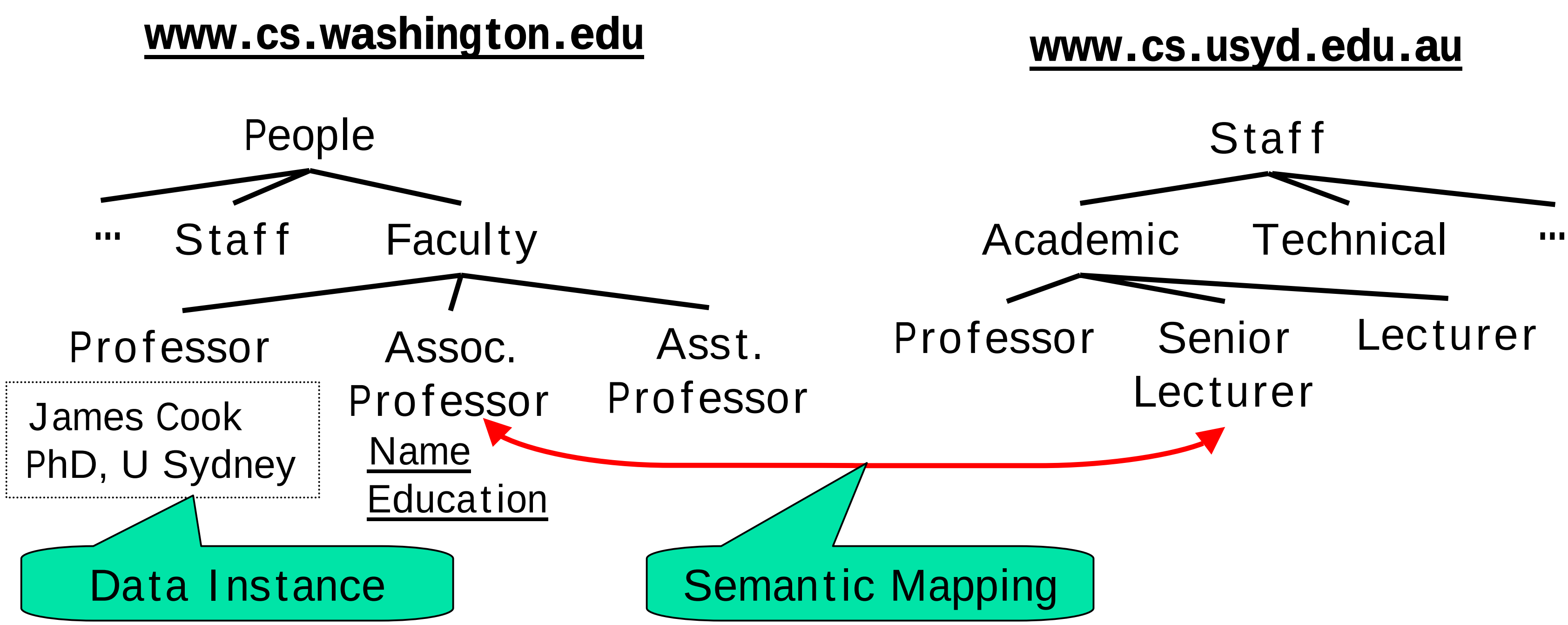
Databases and Data Mining group
University of Washington

Semantic Web

- Mark-up data on the web using ontologies
- Enable intelligent information processing over the web
  - Personal software agents
  - Queries over multiple web pages

  ...

An Example



Find Prof. Cook, a professor in a Seattle college, earlier an assoc. professor at his alma mater in Australia

Semantic Mappings allow information processing across ontologies

Semantic Web: State of the Art

- Languages for ontologies
  - RDF, DAML+OIL,...

- Ontology learning and Ontology design tools
  - [Maedche'02], Protégé, Ontolingua,...

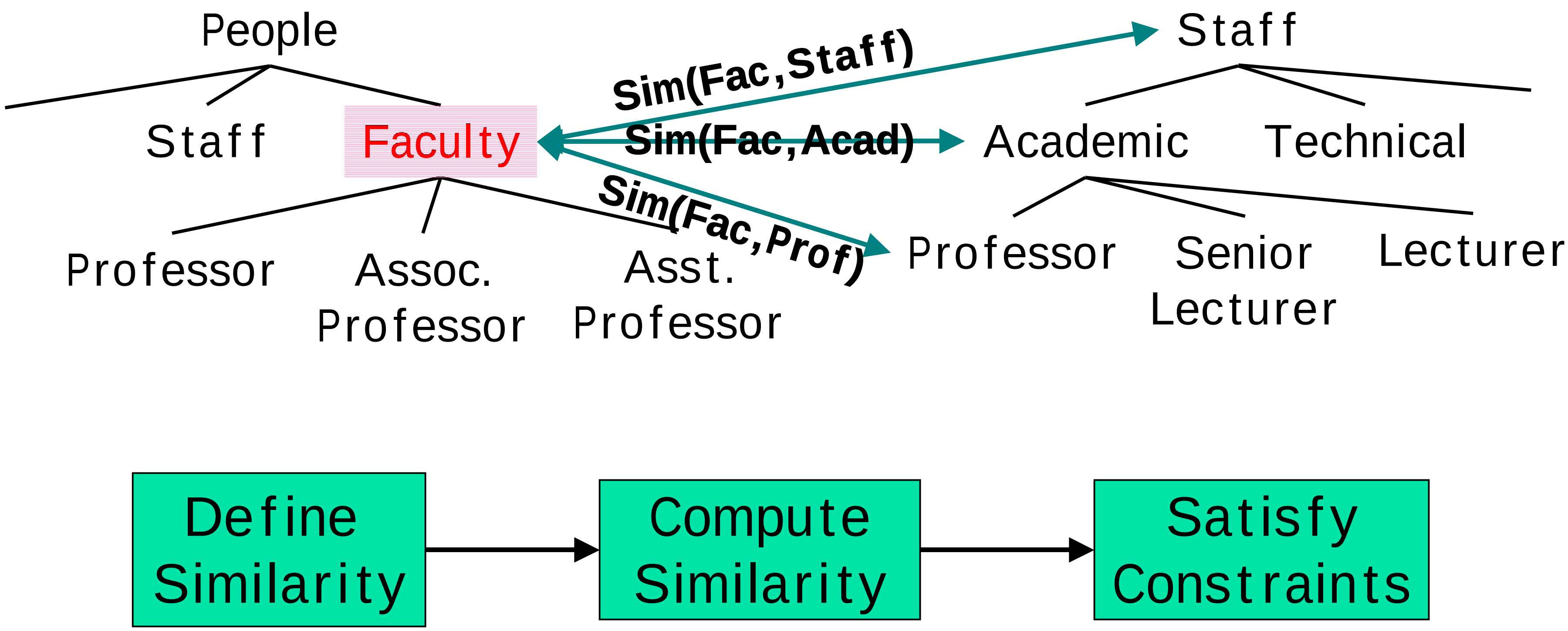
- Semantic Mappings crucial to the SW vision
  - [Uscold'01, Berners-Lee, et al.'01]

Without semantic mappings...Tower of Babel !!!

Semantic Mapping Challenges

- Ontologies can be very different
  - Different vocabularies, different design principles
  - Overlap, but not coincide
- Semantic Mapping information
  - Data instances marked up with ontologies
  - Concept names and taxonomic structure
  - Constraints on the mapping

Overview

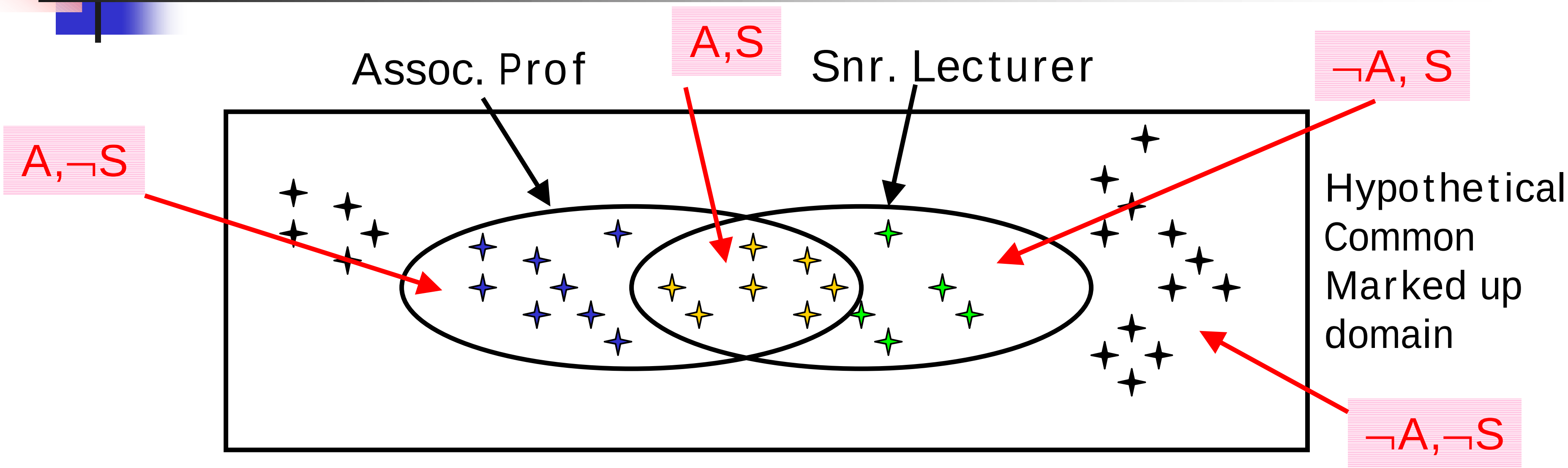


People
Staff
Faculty
Professor
Assoc. Professor
Asst. Professor
Sim(Fac,Staff)
Sim(Fac,Acad)
Sim(Fac,Prof)
Staff
Academic
Technical
Professor
Senior Lecturer
Lecturer
Define Similarity
Compute Similarity
Satisfy Constraints

Our Contributions

- An automatic solution to taxonomy matching
  - Handles different similarity notions
  - Exploits information in data instances and taxonomic structure, *using multi-strategy learning*
- Extend solution to handle wide variety of constraints, *using Relaxation Labeling*
- An implementation, our **GLUE** system, and experiments on real-world taxonomies
  - High accuracy (68-98%) on large taxonomies (100-330 concepts)

Defining Similarity



$$\text{Sim(Assoc. Prof., Snr. Lect.)} = \frac{P(\mathbf{A} \cap \mathbf{S})}{P(\mathbf{A} \cup \mathbf{S})} = \frac{P(A,S)}{P(A,\neg S) + P(A,S) + P(\neg A,S)}$$

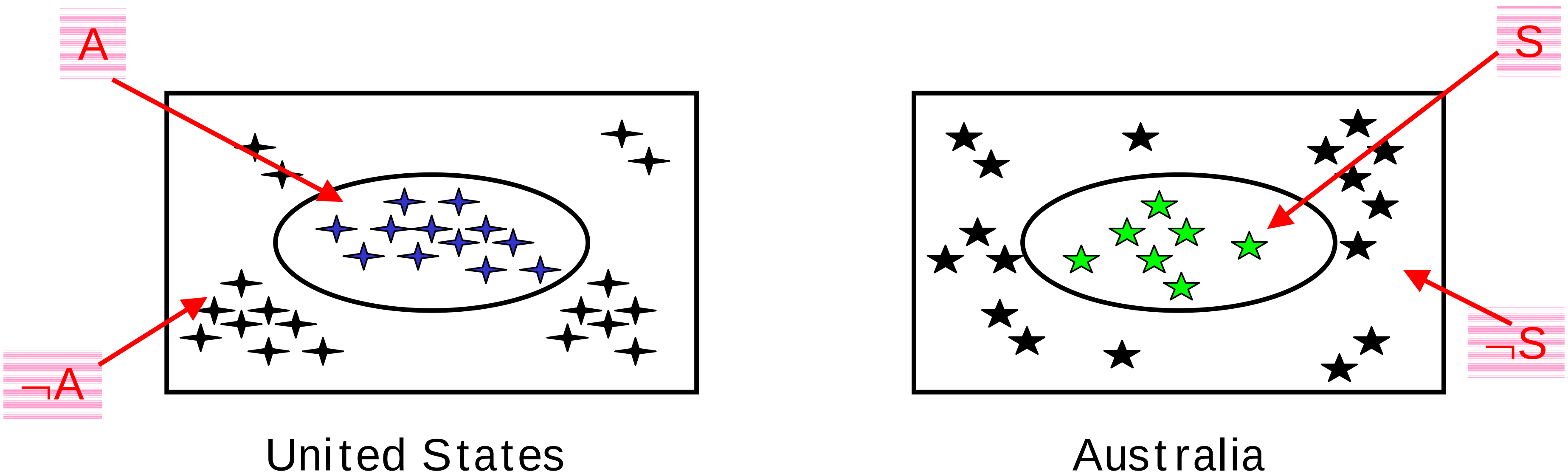
[Jaccard, 1908]

Joint Probability Distribution: **P(A,S),P(¬A,S),P(A,¬S),P(¬A,¬S)**

Multiple Similarity measures in terms of the JPD

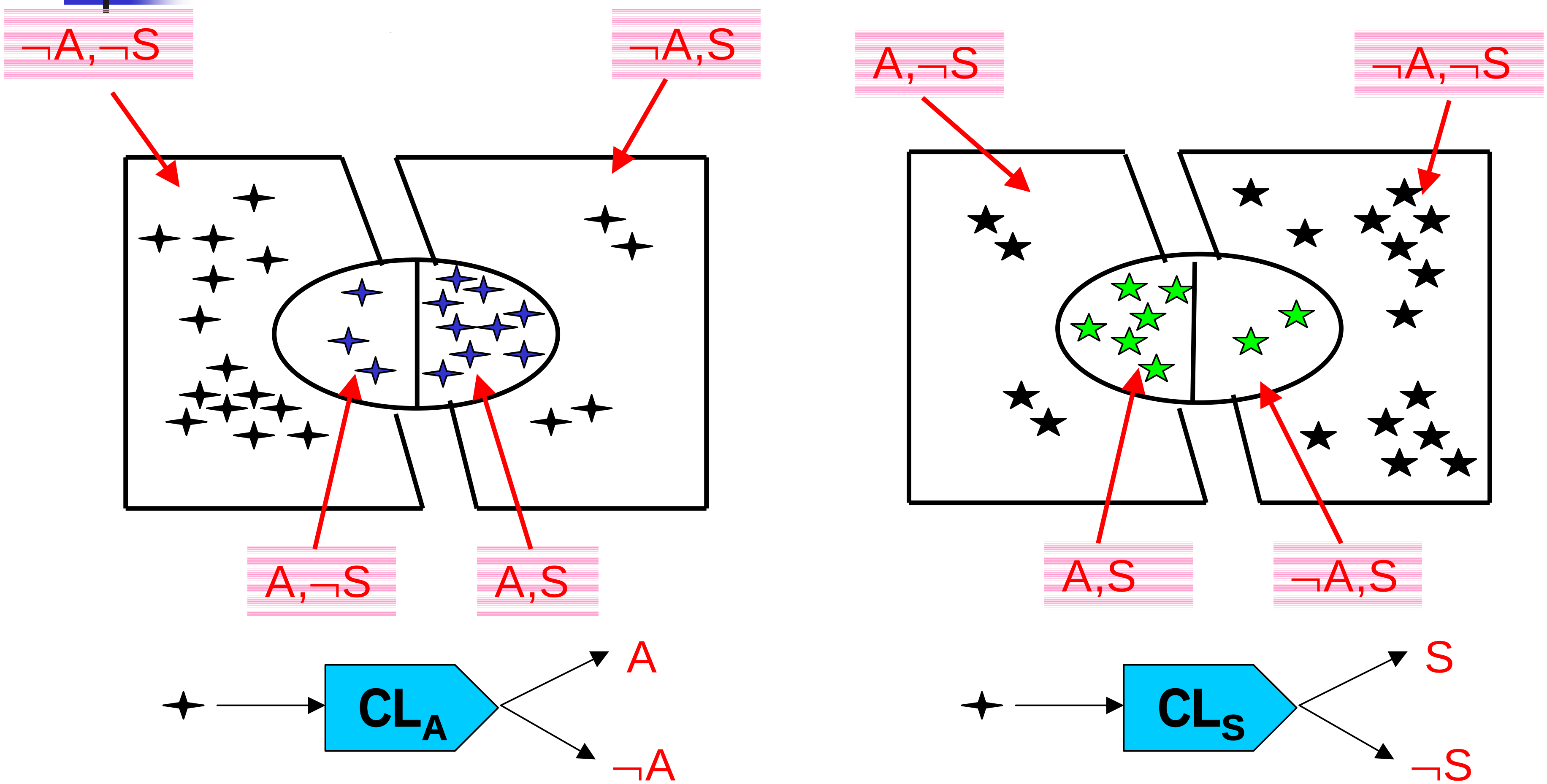
No common data instances

In practice, not easy to find data tagged with both ontologies !



Solution: Use Machine Learning

Machine Learning for computing similarities

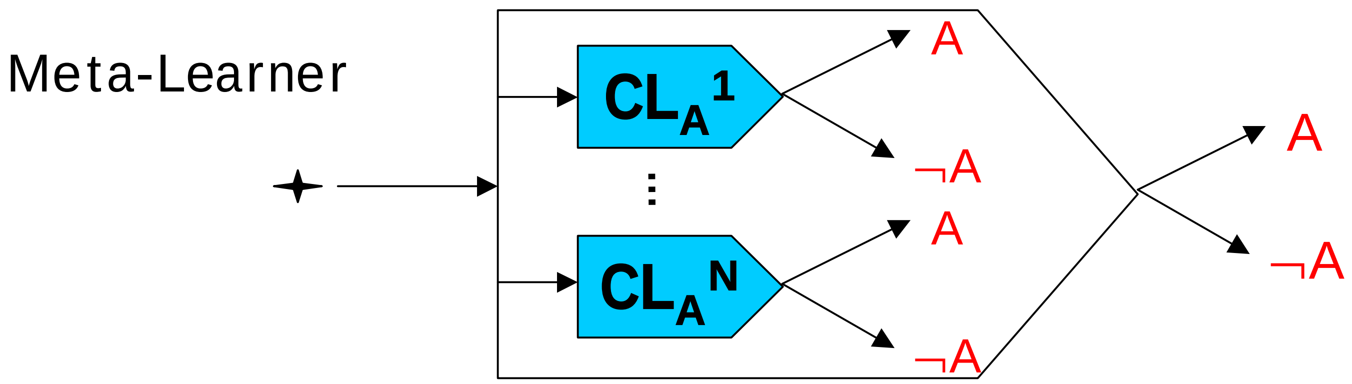


JPD estimated by counting the sizes of the partitions

Improve Predictive Accuracy – Use Multi-Strategy Learning

Single Classifier cannot exploit all available information

Combine the prediction of multiple classifiers



- Content Learner
  - Frequencies on different words in the text in the data instances
- Name Learner
  - Words used in the names of concepts in the taxonomy
- Others ...

So far…

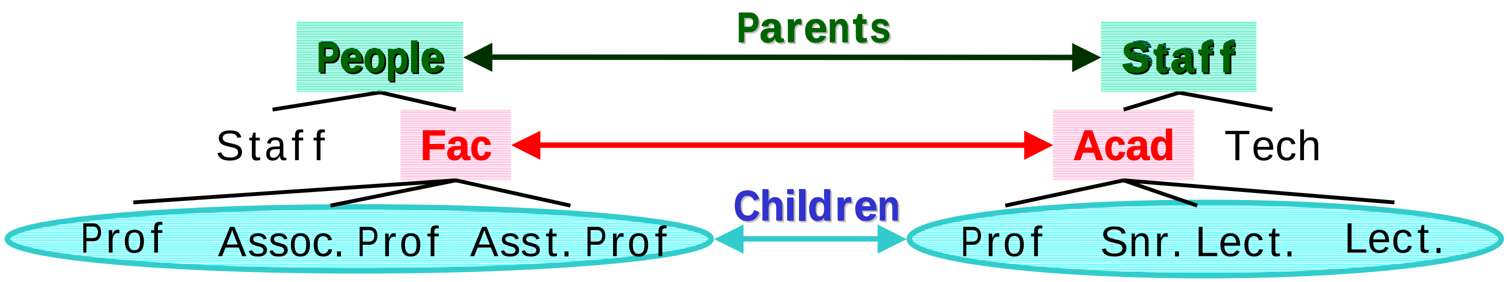
Define Similarity → Compute Similarity → Satisfy Constraints

Joint Probability Distribution

Multi-strategy Learning

Next Step: Exploit Constraints

- Constraints due to the taxonomy structure



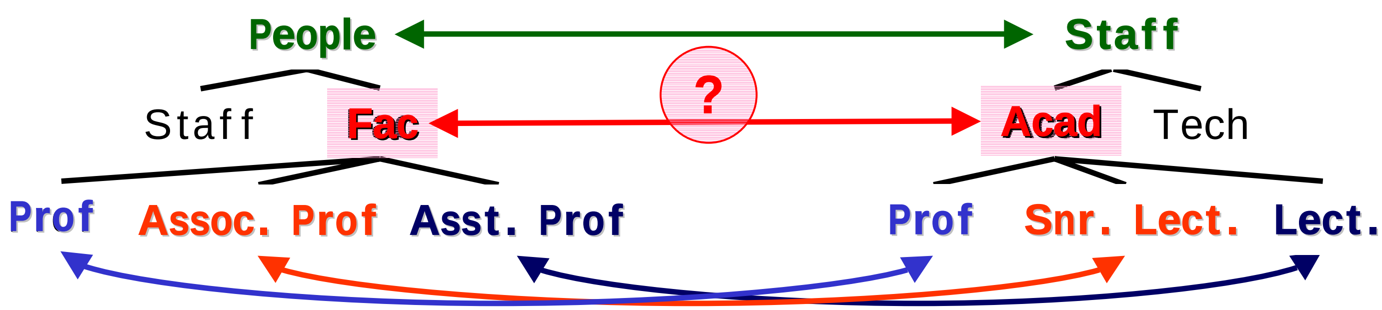
- Domain specific constraints
  - *Department-Chair* can only map to a unique concept

- Numerous constraints of different types

Extended Relaxation Labeling to ontology matching

Solution: Relaxation Labeling

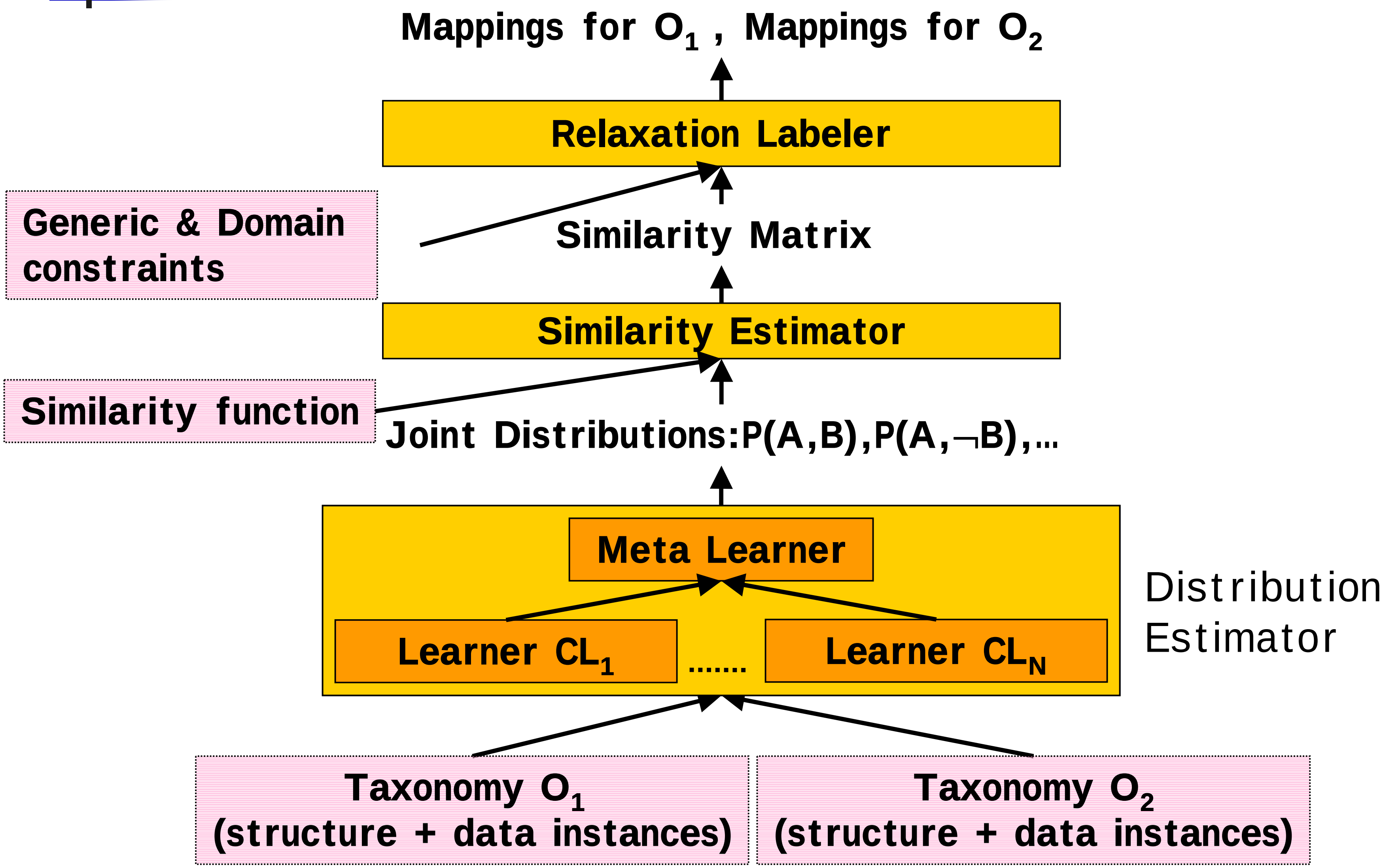
Find the best label assignment given a set of constraints



- Start with an initial label assignment
- Iteratively improves labels, given constraints

- Standard Relaxation Labeling not applicable
  - Extended in many ways

Putting it all together GLUE System



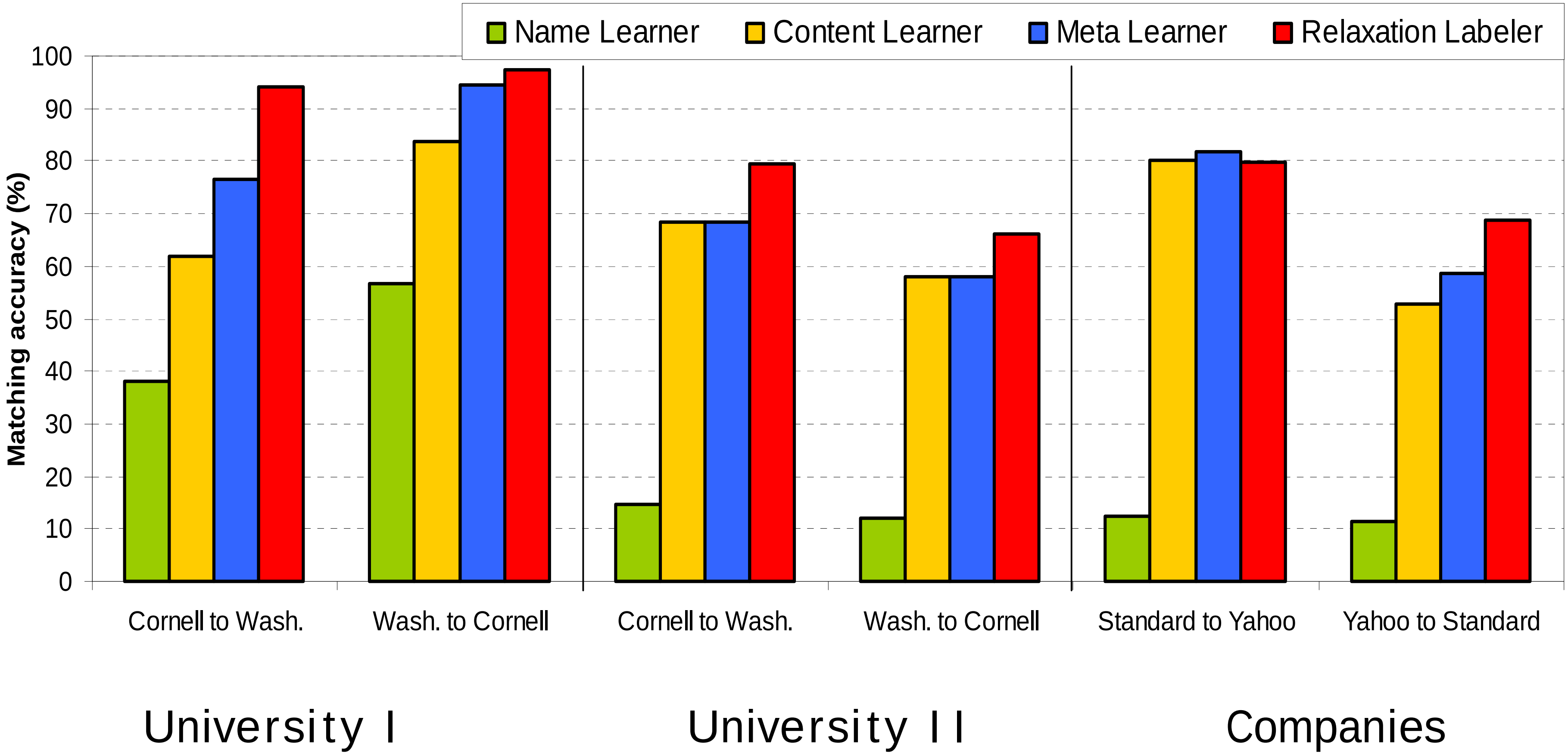
Real World Experiments

- Taxonomies on the web
  - University classes (UW and Cornell)
  - Companies (Yahoo and The Standard)

- For each taxonomy
  - Extracted data instances – course descriptions, and company profiles
  - Trivial data cleaning
  - 100 – 300 concepts per taxonomy
  - 3-4 depth of taxonomies
  - 10-90 average data instances per concept

- Evaluation against manual mappings as the gold standard

Results


Name Learner
Content Learner
Meta Learner
Relaxation Labeler
Matching accuracy (%)
100
90
80
70
60
50
40
30
20
10
0
Cornell to Wash.
Wash. to Cornell
Cornell to Wash.
Wash. to Cornell
Standard to Yahoo
Yahoo to Standard
University I
University II
Companies

Related Work

- Our LSD schema matching system [Doan, Domingos, Halevy '01]
  - GLUE handles taxonomies, richer models, and a much richer set of constraints
- Other Ontology and Schema Matching work [Noy, Musen'01], [Melnik, et al.'02], [Ichise, et al.'01]
  - Mostly heuristics, or single machine learning techniques
- Relaxation Labeling for constraint satisfaction [Hummel, Zucker'83], [Chakrabarti, et al.'00]
  - Significantly extend this approach

Conclusions & Future Work

- An automated solution to taxonomy matching
  - Handles multiple notions of similarity
  - Exploits data instances and taxonomy structure
  - Incorporates generic and domain-specific constraints
  - Produces high accuracy results
- Future Work
  - More expressive models
  - Complex Mappings
  - Automated reasoning about mappings between models