Learning High Accuracy Rules for Object Identification

Sheila Tejada

Wednesday, December 12, 2001

Committee Chair: Craig A. Knoblock

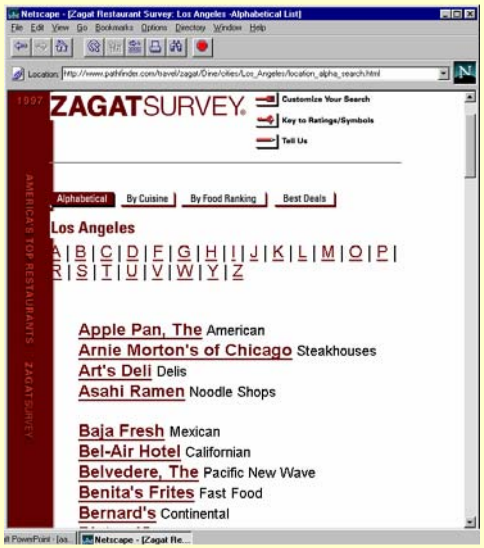
Committee: Dr. George Bekey, Dr. Kevin Knight,

Dr. Steven Minton, Dr. Daniel O'Leary

Integrating Restaurant Sources

Zagat's Restaurant Guide Source

Department of Health Restaurant Rating Source



ARIADNE
Information Mediator

Question: What is the **Review** and **Rating** for the Restaurant "**Art's Deli**"?

Ariadne Information Mediator

ARIADNE
Information Mediator

Zagat's Wrapper

User Query

Dept. of Health Wrapper

Extract web objects in the form of database records

Zagat's

Name	Street	Phone
Art's Deli	12224 Ventura Boulevard	818-756-4124
Teresa's	80 Montague St.	718-520-2910
Steakhouse The	128 Fremont St.	702-382-1600
Les Celebrites	155 W. 58th St.	212-484-5113

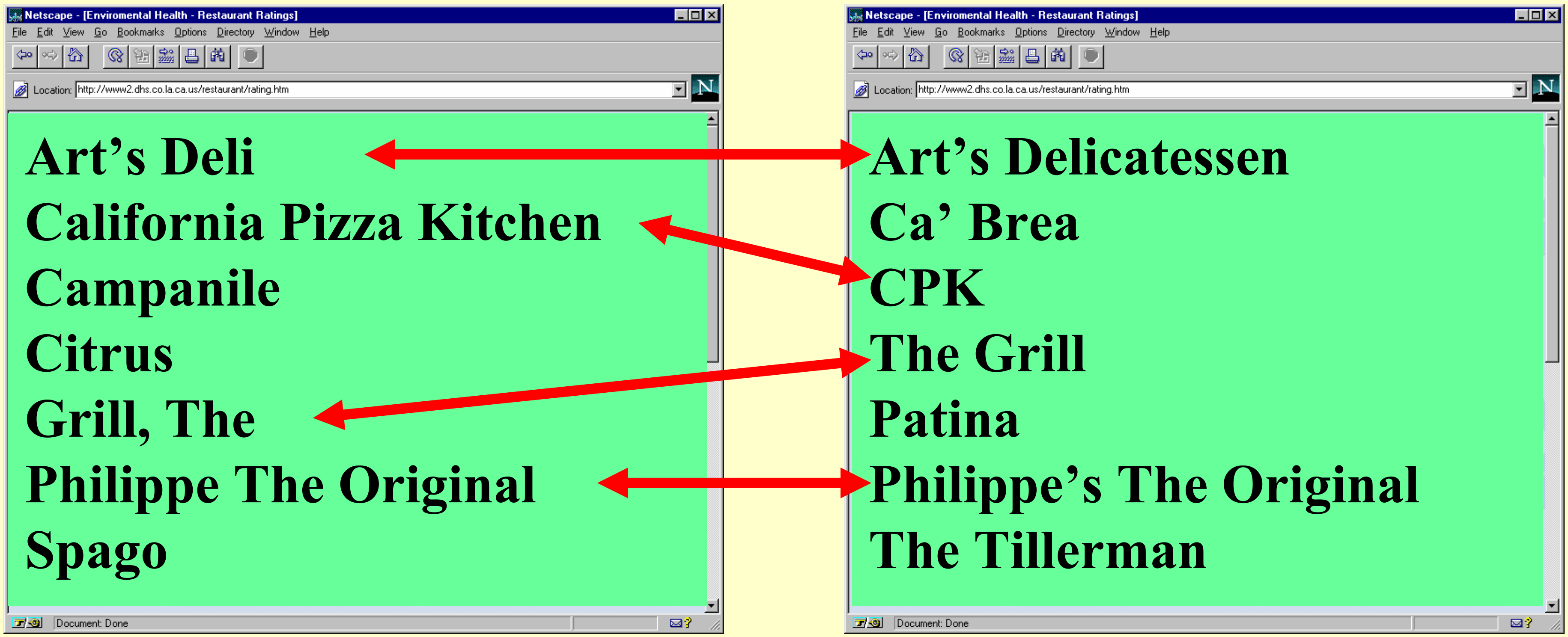
Dept of Health

Name	Street	Phone
Art's Delicatessen	12224 Ventura Blvd.	818/755-4100
Teresa's	103 1st Ave. between 6th and 7th Sts.	212/228-0604
Binion's Coffee Shop	128 Fremont St.	702/382-1600
Les Celebrites	5432 Sunset Blvd	212/484-5113

Multi-Source Inconsistency

Zagat's Restaurant Guide Source

Department of Health Restaurant Source



How can the same objects be identified when they are stored in inconsistent text formats?

Application Dependent Mapping

Observations:

- Mapping objects can be application dependent
- Example:

Mapped?

Steakhouse The	128 Fremont Street	702-382-1600

Binion's Coffee Shop	128 Fremont St.	702/382-1600

- The mapping is in the application, not the data
- User input is needed to increase accuracy of the mapping

Key Ideas for Mapping Objects

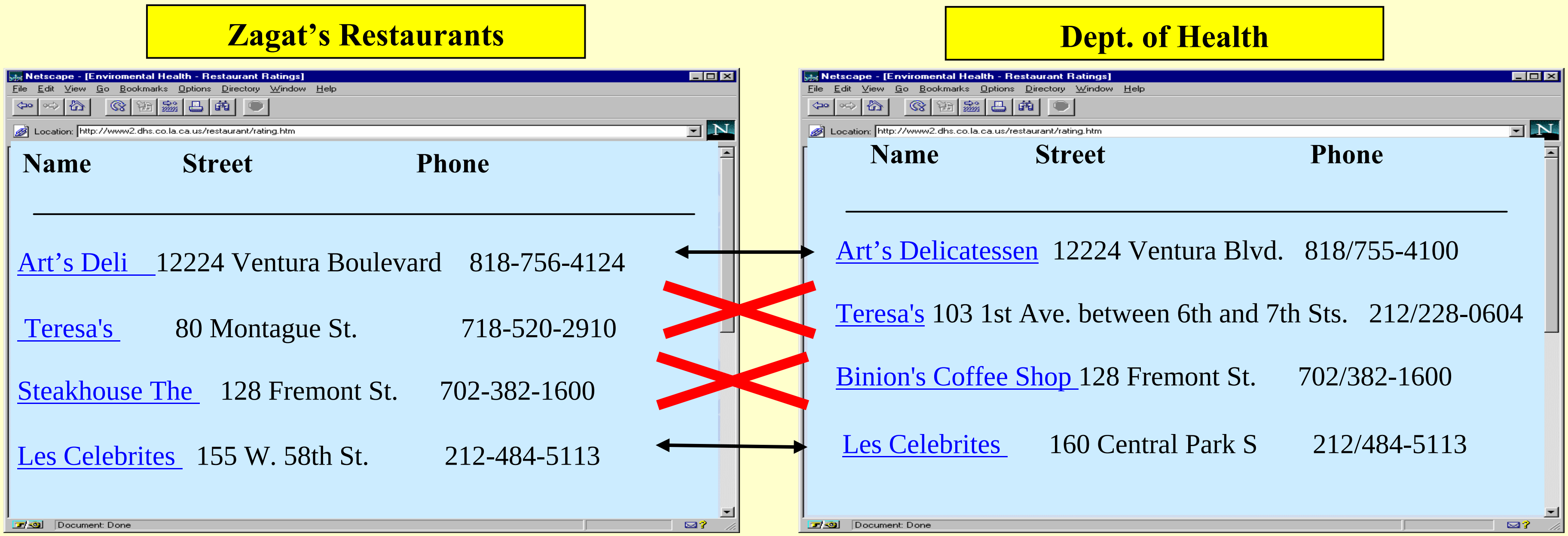
- Learning important attributes for determining a mapping

	Name	Street	Phone
Zagat's	Art's Deli	12224 Ventura Boulevard	818-756-4124
Dept of Health	Art's Delicatessen	12224 Ventura Blvd.	818/755-4100

- Learning general transformations to recognize objects

Zagat's	**Transformations**	**Dept of Health**
Art's Deli	Prefix	Art's Delicatessen
California Pizza Kitchen	Acronym	CPK
Philippe The Original	Stemming	Philippe's The Original

Mapping Rules



Mapping rules:

Name > .9 & Street > .87 => mapped

Name > .95 & Phone > .96 => mapped

Transformation Weights

- Transformations can be more appropriate for a specific application domain
  - Restaurants, Companies or Airports
- Or for different attributes within an application domain
  - Acronym more appropriate for the attribute Restaurant Name than for the Phone attribute
- Learn likelihood that if transformation is applied then the objects are mapped

Transformation Weight = P(mapped | transformation)

Thesis Statement

By simultaneously learning to tailor mapping rules and transformation weights to a specific domain, an object identification system can achieve high accuracy without sacrificing domain independence.

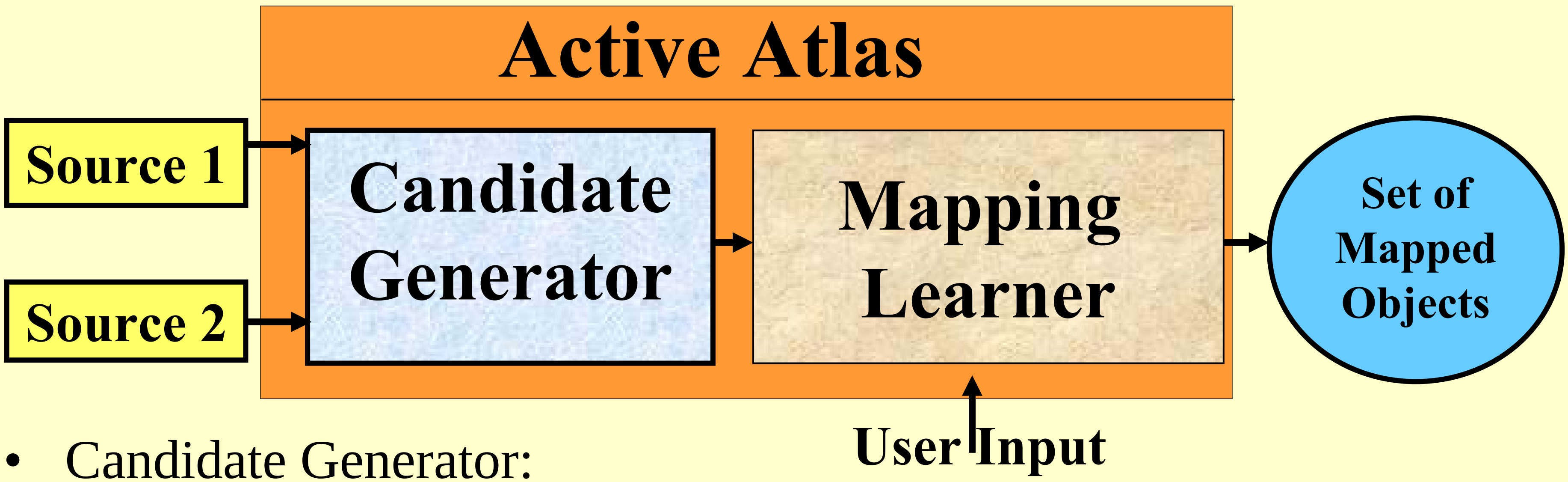
Contributions

- Approach to learning mapping rules that achieve high accuracy mapping while minimizing user involvement
- Only approach developed to tailor a general set of transformations to a specific domain application
- Novel method to combine both forms of learning to create a robust object identification system

Overview

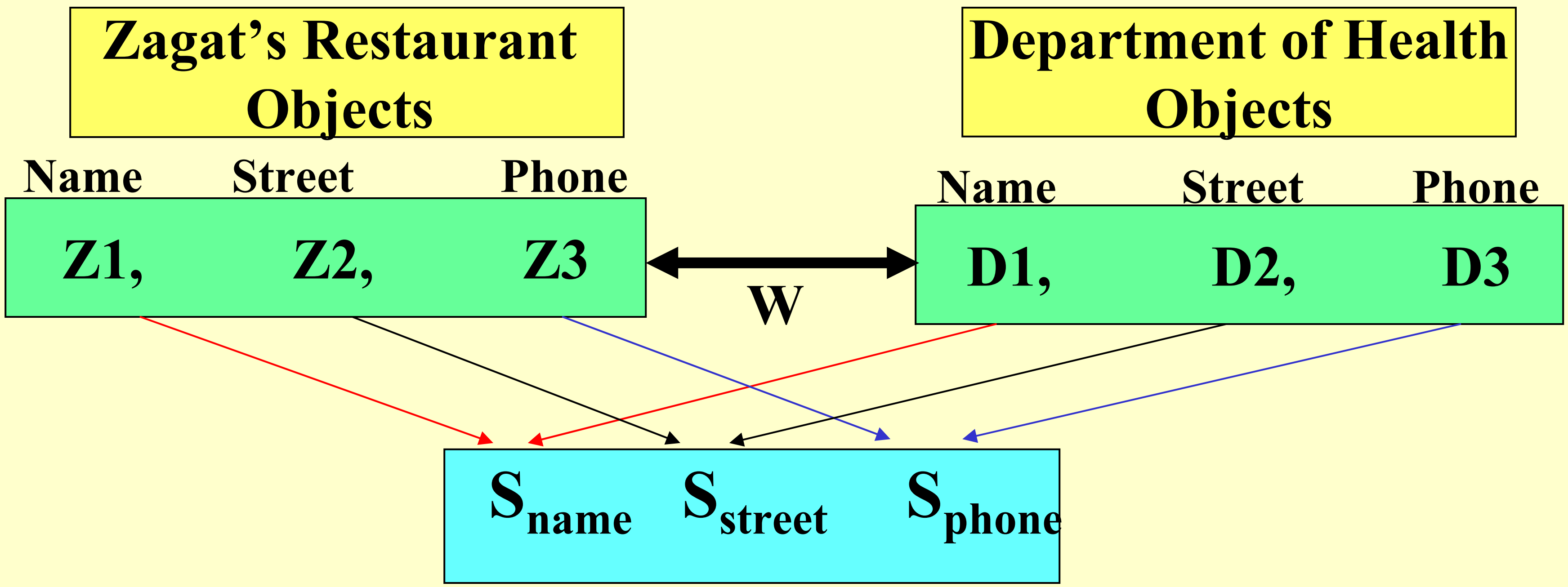
- Approach
  - Computing textual similarity
  - Learning important attributes for mapping
    - Mapping rule learning
  - Learning transformation weights
- Experimental Results
- Related Work on Object Identification
- Conclusions & Future Work

Learning Object Mappings



- Candidate Generator:
  - Judge textual similarity of mappings
  - Reduce number of mappings considered for classification
- Mapping Learner:
  - Active learning technique to learn mapping rules and transformation weights
  - Minimize the amount of user interaction

Computing Textual Similarity



- Candidate Generator returns sets of similarity scores

Name	Street	Phone
.9	.79	.4
.17	.3	.74
	. . .	

Types of Transformations

Type I Transformations

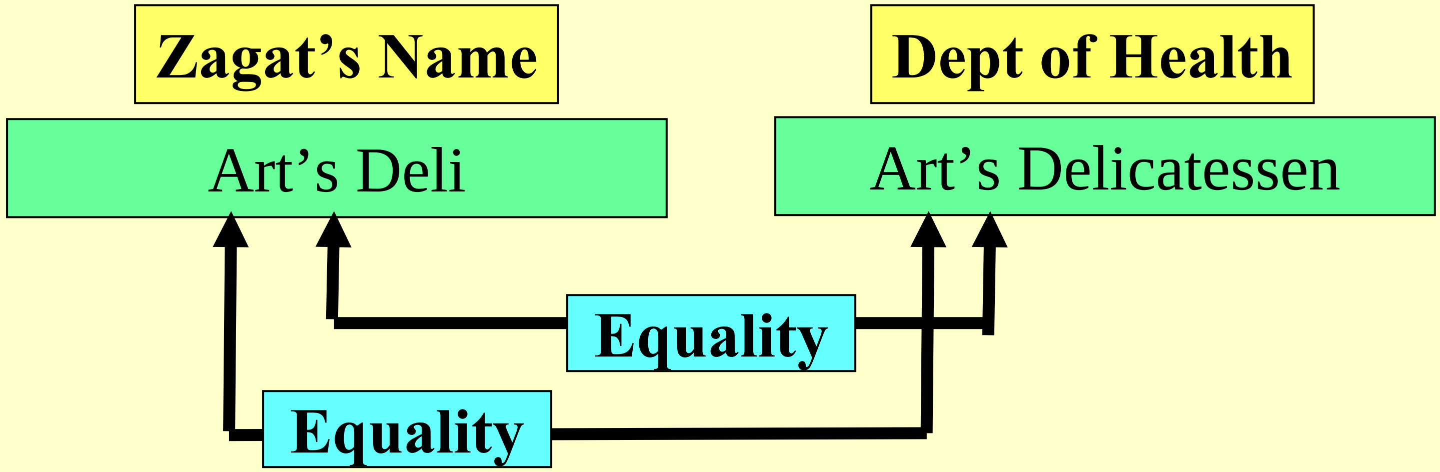
- **Equality (Exact match)**
- **Stemming**
- **Soundex (e.g. “Celebrites” => “C453”)**
- **Abbreviation (e.g. “3rd” => “third”)**

Type II Transformations

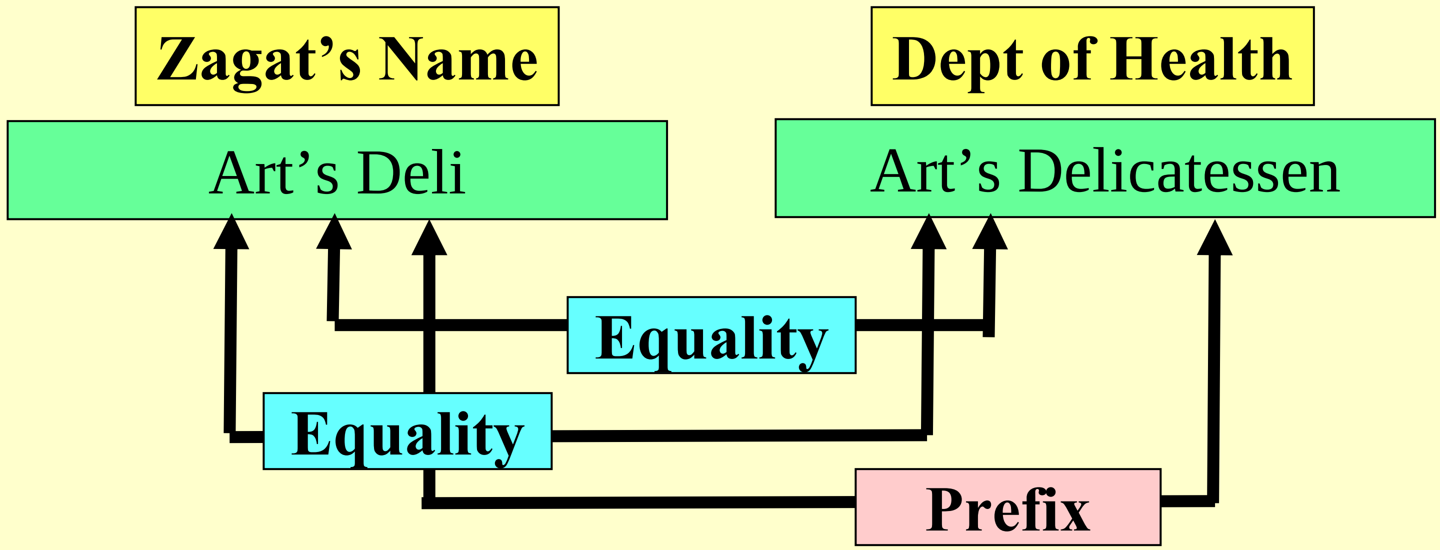
- **Initial**
- **Prefix (e.g. “Deli” & “Delicatessen”)**
- **Suffix**
- **Substring**
- **Acronym (e.g. “California Pizza Kitchen” & “CPK”)**
- **Drop Word**

Applying Type I Transformations

- Employs Information Retrieval Techniques
- One set of attribute values broken into words or tokens
  - "Art" "s" "Delicatessen"
- Apply Type I transformations to tokens
  - "Art" "A630" "s" "S000" "Delicatessen" "D423"
- Enter tokens into inverted index
- Tokens from second set used to query the index
  - Transformed query set: "Art" "A630" "s" "S000" "Deli" "Del" "D400"



Applying Type II Transformations



- Type II transformations improve measurement of similarity

Attribute Similarity Function

- Transformations determine similarity of attribute values
- Each attribute value is represented as a vector

  < 2 4 3 0 5 6 6 0 0 0 0 0 5 0 0 0 0 . . .>
- Attribute Similarity Function:
  - Cosine Measure with a TFIDF

Similarity (A, B) =

$$\frac{\sum_{i=1}^{t} (w_{ia} \times w_{ij})}{\sqrt{\sum_{i=1}^{t} (w_{ia})^2 \times \sum_{i=1}^{t} (w_{ij})^2}}$$

$w_{ia} = (0.5 + 0.5\, freq_{ia}) \times IDF_i$

$w_{ij} = freq_{ij} \times IDF_i$

$freq_{ia}$ = frequency of term *i* for attribute value *a*

IDF_i = IDF of term *i* in the entire collection

$freq_{ij}$ = frequency of term *i* in attribute value *j*

Total Object Similarity Scores

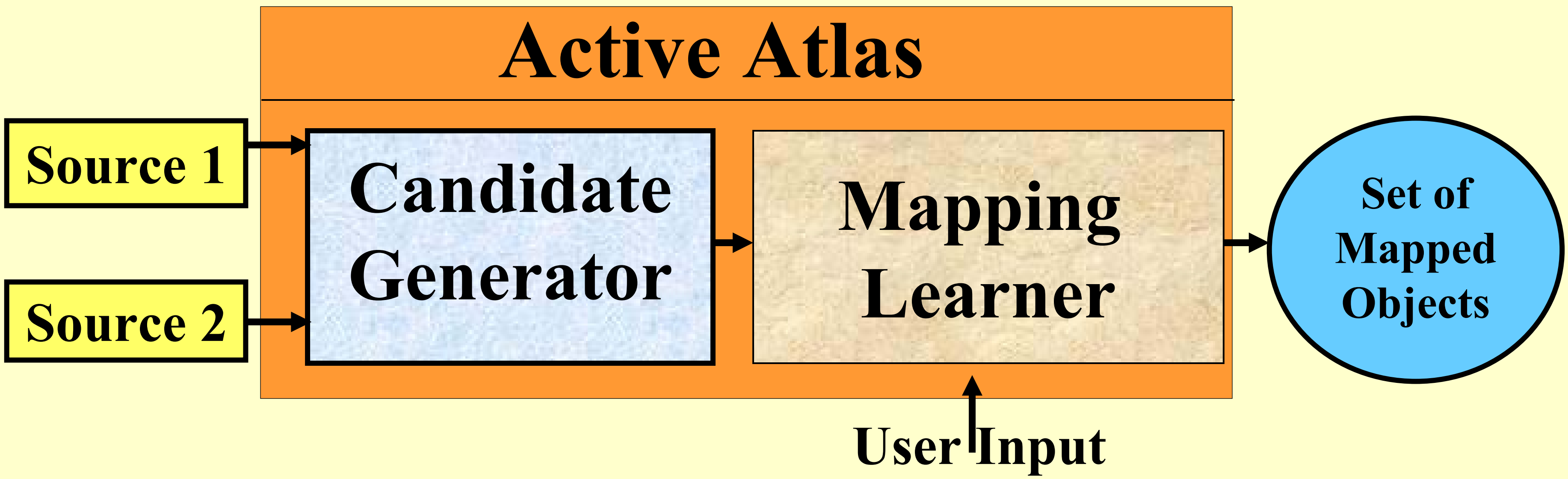
	Name	Street	Phone
Zagat's	Art's Deli	12224 Ventura Boulevard	818-756-4124
Dept of Health	Art's Delicatessen	12224 Ventura Blvd.	818/755-4100

Candidate Mapping Similarity Scores:

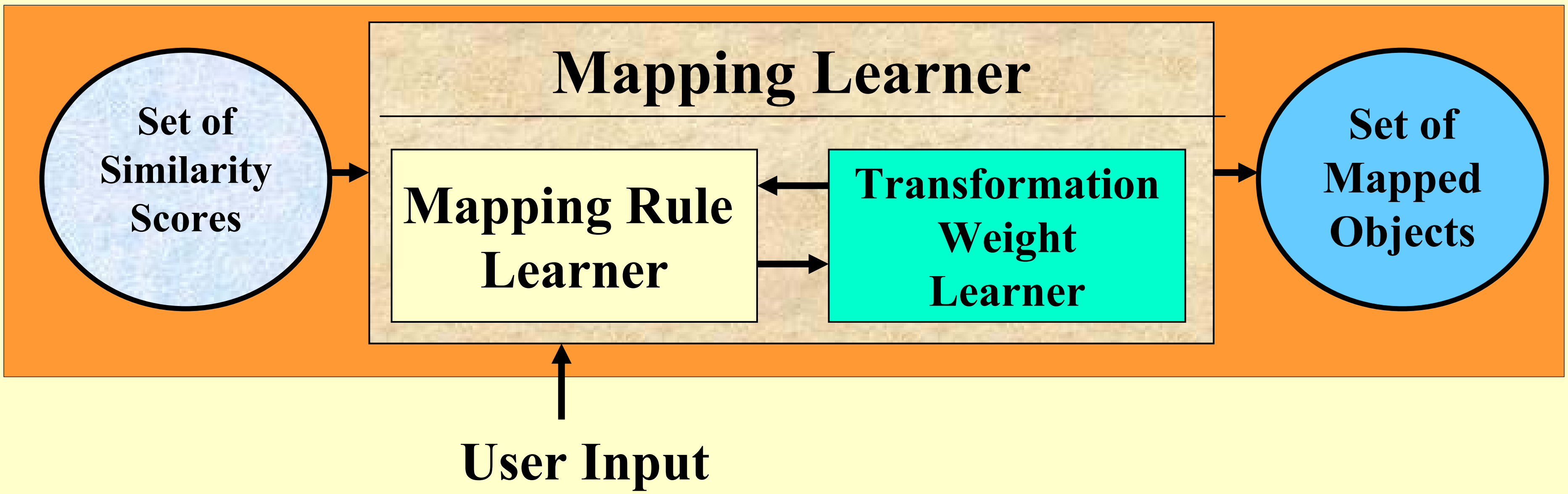
Name	Street	Phone	Total Score
.967	.973	.3	2.034
.17	.3	.74	1.182
.8	.5	.49	1.749

. . .

Learning Object Mappings



Learning Object Mappings



- The goal is to classify with high accuracy the proposed mappings while minimizing user input
  - Active learning technique
    - System chooses most informative example for the user to label

Mapping Rules

Set of Similarity Scores

Name	Street	Phone
.967	.973	.3
.17	.3	.74
.8	.542	.49
.95	.97	.67
	…	

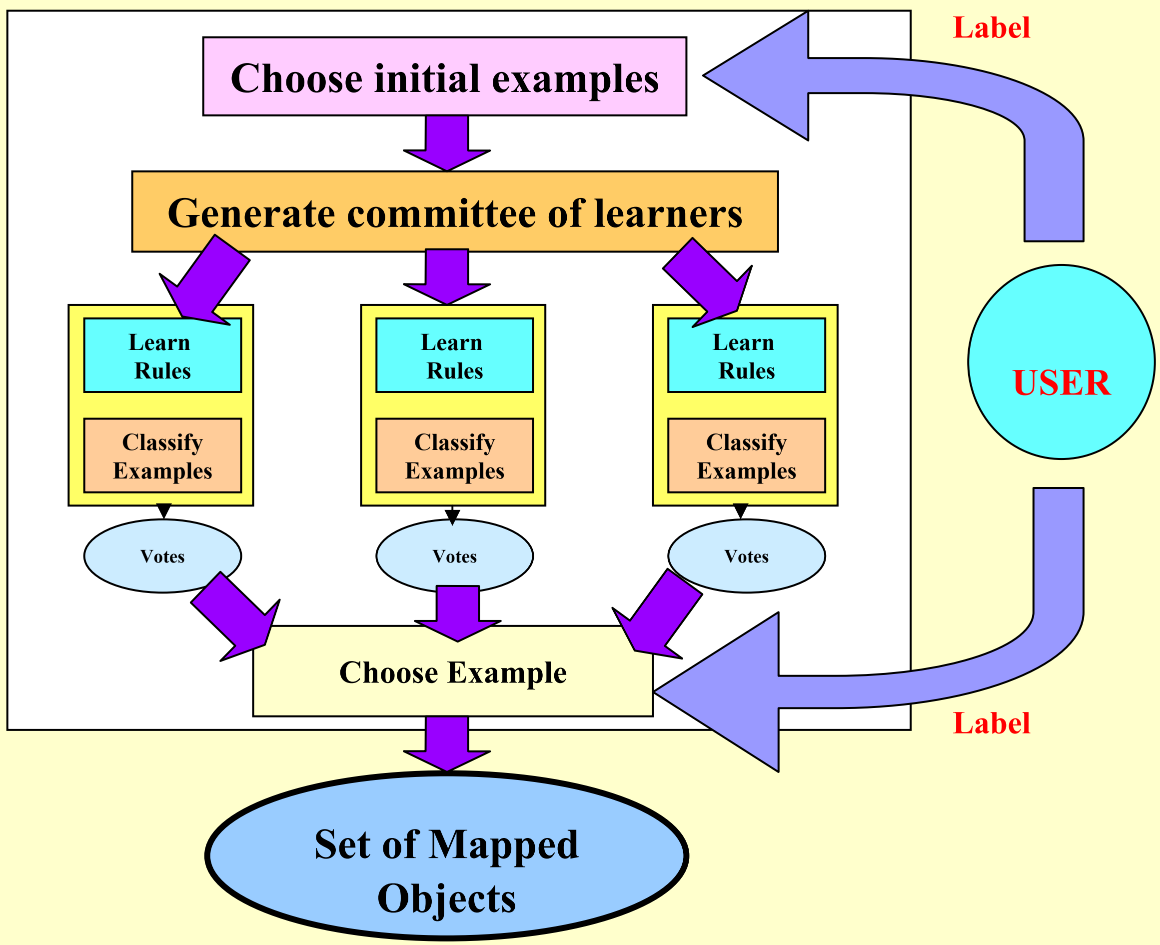
Mapping Rules

Name > .8 & Street > .79 => mapped

Name > .89 => mapped

Street < .57 => not mapped

Mapping Rule Learner



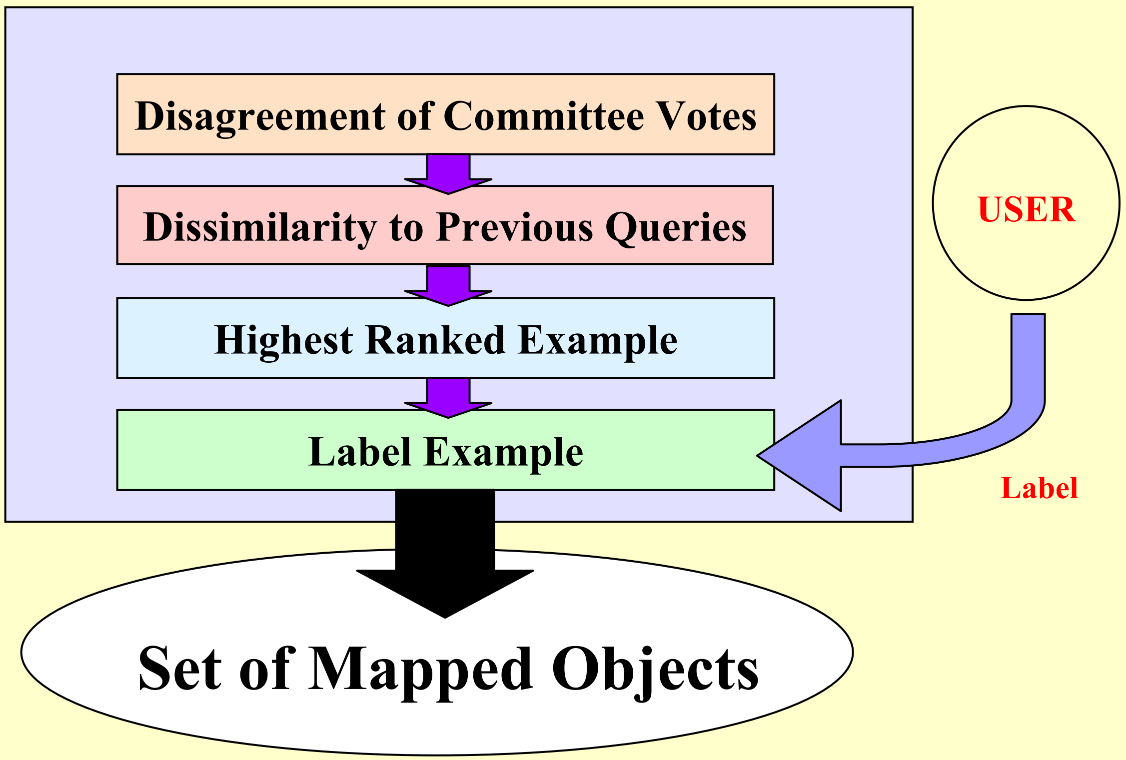
Committee Disagreement

- Chooses an example based on the disagreement of the query committee

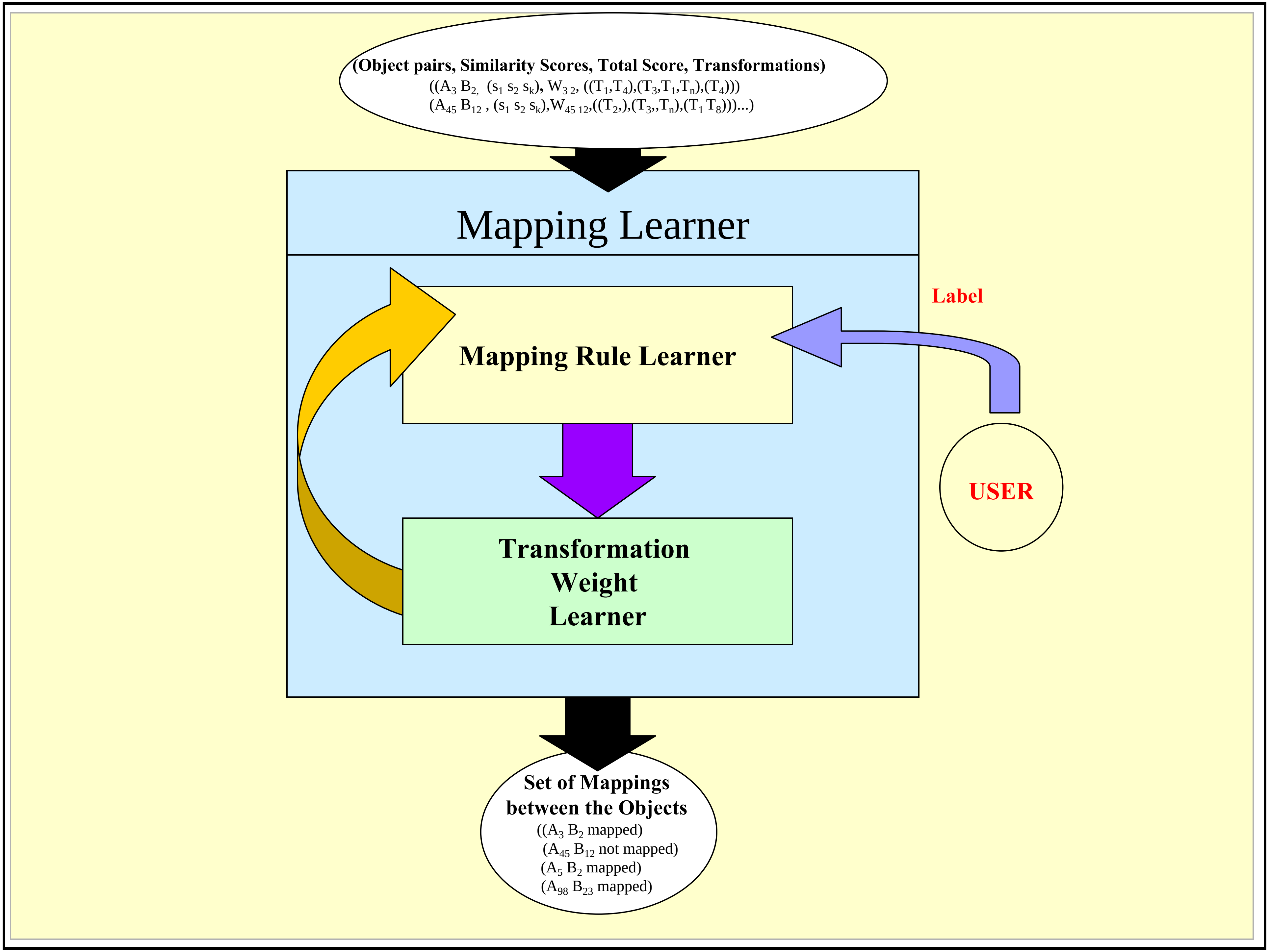
Examples	Committee M1	Committee M2	Committee M3
Art's Deli, Art's Delicatessen	Yes	Yes	Yes
CPK, California Pizza Kitchen	Yes	No	Yes
Ca'Brea, La Brea Bakery	No	No	No

- In this case CPK, California Pizza Kitchen is the most informative example based on disagreement

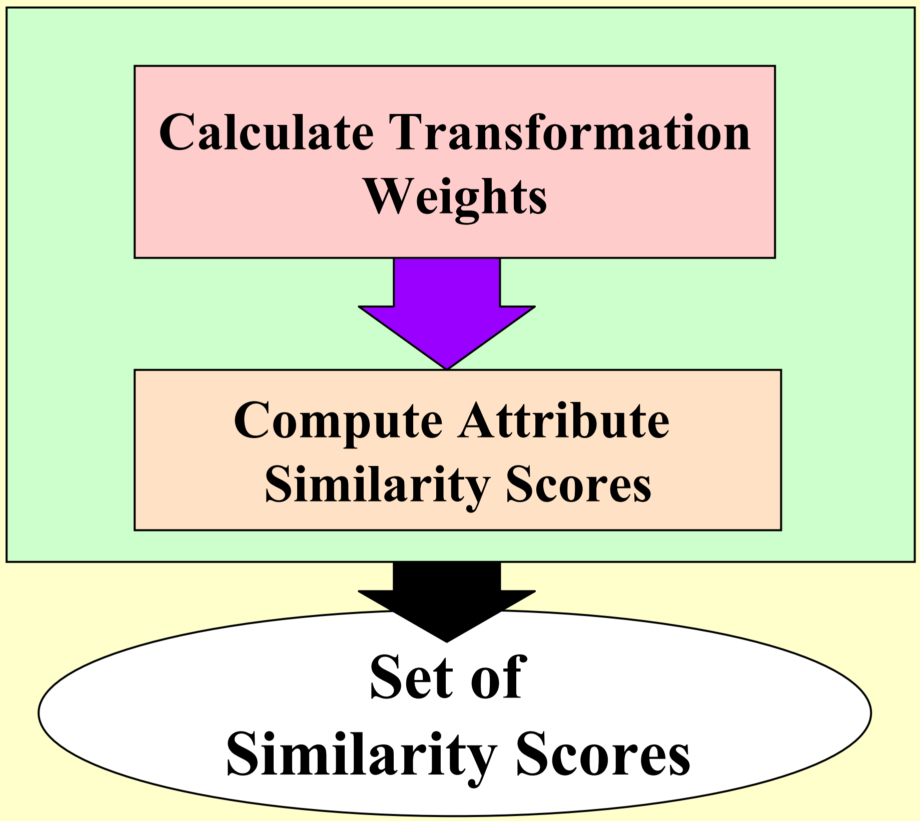
Choosing Next Example



- The user labels the example, and the system updates the committee
- Mapping Rule Learner outputs classified examples


(Object pairs, Similarity Scores, Total Score, Transformations)
((A_3 B_2, (s_1 s_2 s_k), $W_{3\,2}$, ((T_1,T_4),(T_3,T_1,T_n),(T_4)))
(A_{45} B_{12} , (s_1 s_2 s_k),$W_{45\,12}$,((T_2,),(T_3,,T_n),(T_1 T_8)))...)
Mapping Learner
Mapping Rule Learner
Label
USER
Transformation Weight Learner
Set of Mappings between the Objects
((A_3 B_2 mapped)
(A_{45} B_{12} not mapped)
(A_5 B_2 mapped)
(A_{98} B_{23} mapped)

Transformation Weight Learner



Calculate Transformation Weights

$$\boxed{P(\text{mapped} \mid \text{transformation})} = \frac{P(\text{transformation} \mid \text{mapped})\, P(\text{mapped})}{P(\text{transformation})}$$

Examples	Classification	Labeled by
Art's Deli, Art's Delicatessen	Mapped	Learner
CPK, California Pizza Kitchen	Mapped	User
Ca'Brea, La Brea Bakery	Not Mapped	Learner

Recalculating Similarity Scores

Transformation	Mapped	Not Mapped
(EQUAL "Art" "Art")	.8	.2
(EQUAL "s" "s")	.8	.2
(PREFIX "Deli" "Delicatessen")	.1	.9

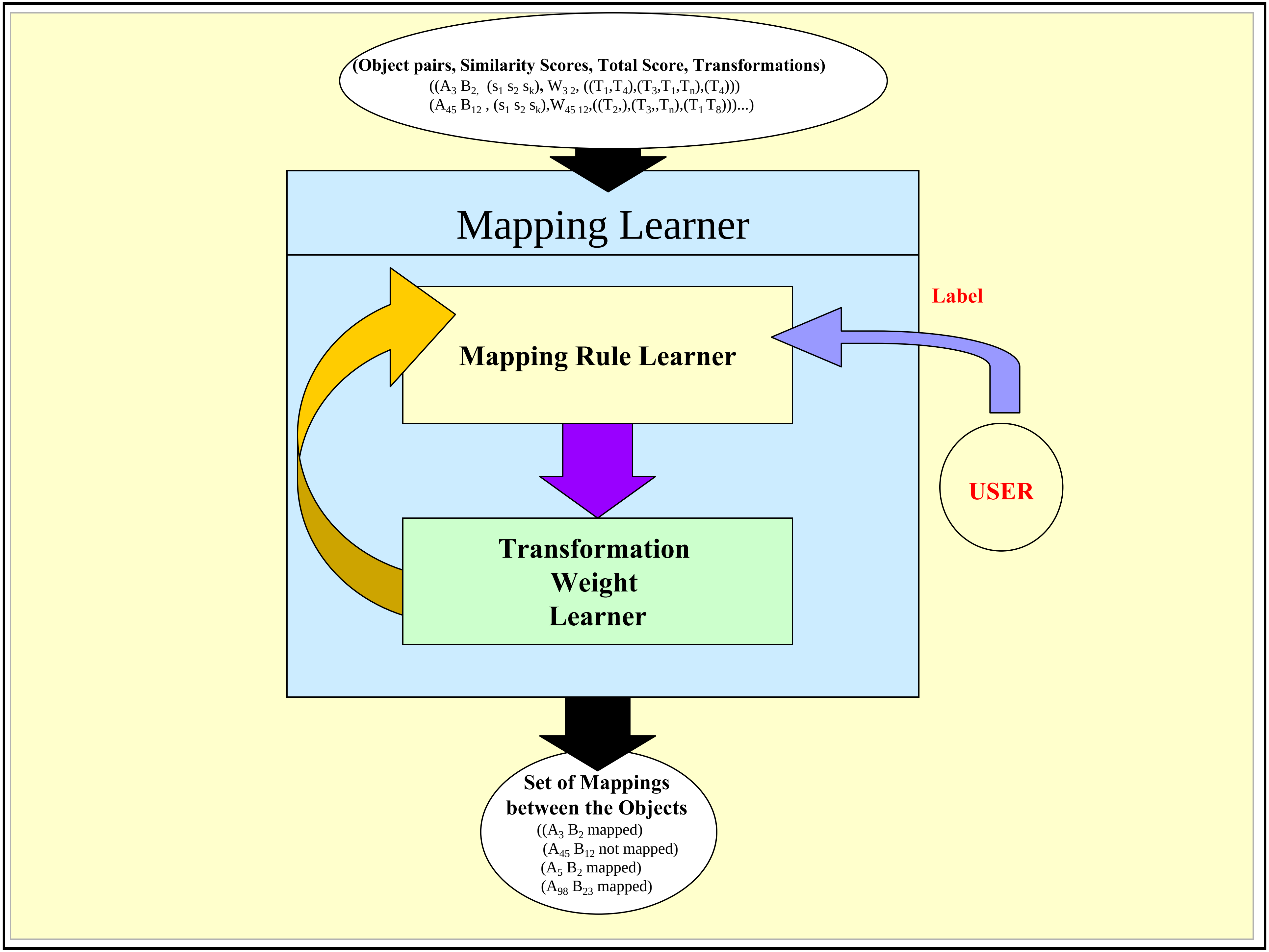
Total mapped score m = .064

Total not mapped score n = .004

Normalized Attribute Similarity Score = m/(m + n)

= .064/ (.064 + .004)

Attribute Similarity Score = .941

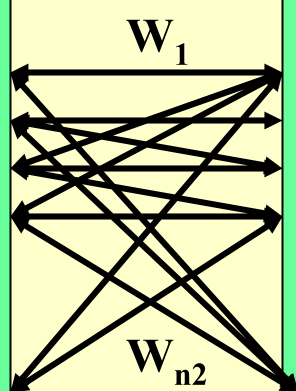

(Object pairs, Similarity Scores, Total Score, Transformations)
((A3 B2, (s1 s2 sk), W3 2, ((T1,T4),(T3,T1,Tn),(T4)))
(A45 B12 , (s1 s2 sk),W45 12,((T2,),(T3,,Tn),(T1 T8)))...)
Mapping Learner
Mapping Rule Learner
Label
USER
Transformation Weight Learner
Set of Mappings between the Objects
((A3 B2 mapped)
(A45 B12 not mapped)
(A5 B2 mapped)
(A98 B23 mapped)

Enforcing One-to-One Relationship

- Viewed as weighted bipartite matching problem

Zagat's

(Name,	Street,	City)
(Art's Deli,	1745 Ventura Boulevard,	Encino)
(Citrus,	267 Citrus Ave.,	LA)
(Spago,	456 Sunset Bl.	LA)
(Z1,	Z2,	Z3)
	.	
	.	
	.	
(	not in source	)
	.	



Dept of Health

(Name,	Street,	City)
(Art's Delicatessen,	1745 Ventura Blvd,	Encino)
(Ca' Brea,	6743 La Brea Ave.,	LA)
(Patina,	342 Melrose Ave.,	LA)
(D1,	D2,	D3)
	.	
	.	
	.	
(	not in source	)
	.	

Given weights W, matching method determines mostly likely Matching Assignment

Experimental Results

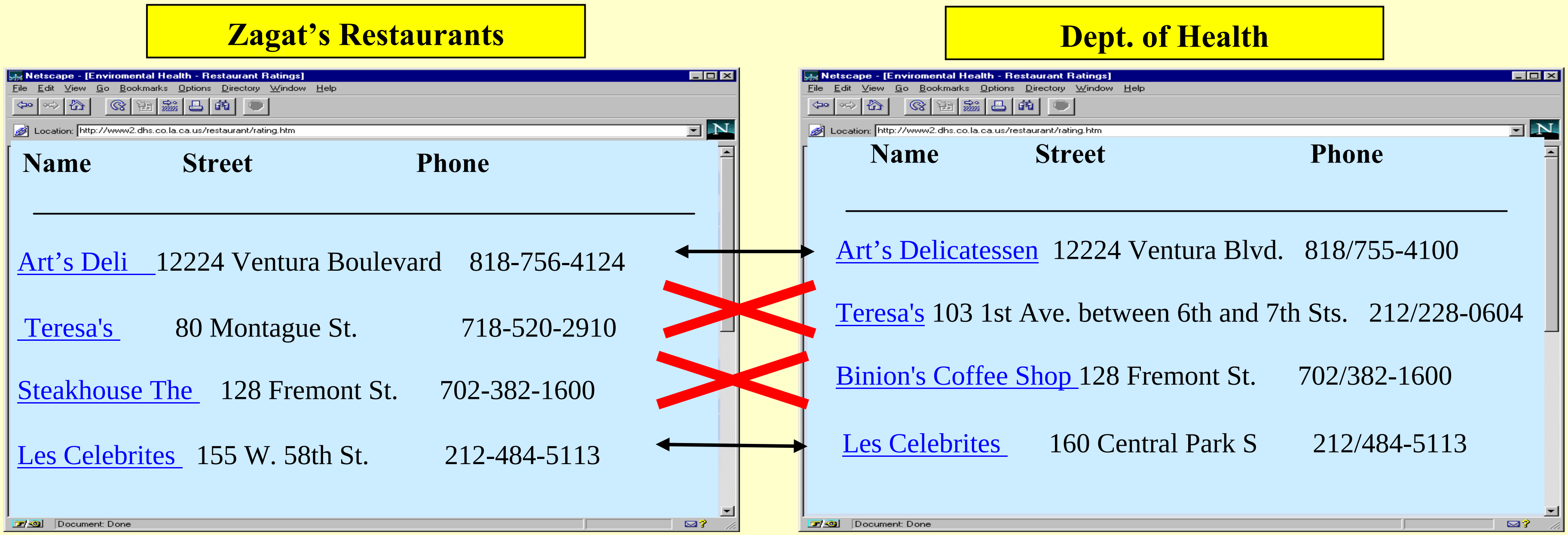
- Three domains: Restaurant, Company, Airport
- Three types of experiments
  - Active Atlas (Mapping Learner)

    CG — Mapping Learner
  - Passive Atlas (Decision tree learner)

    CG — Decision tree learner
  - Candidate Generator (Baseline)

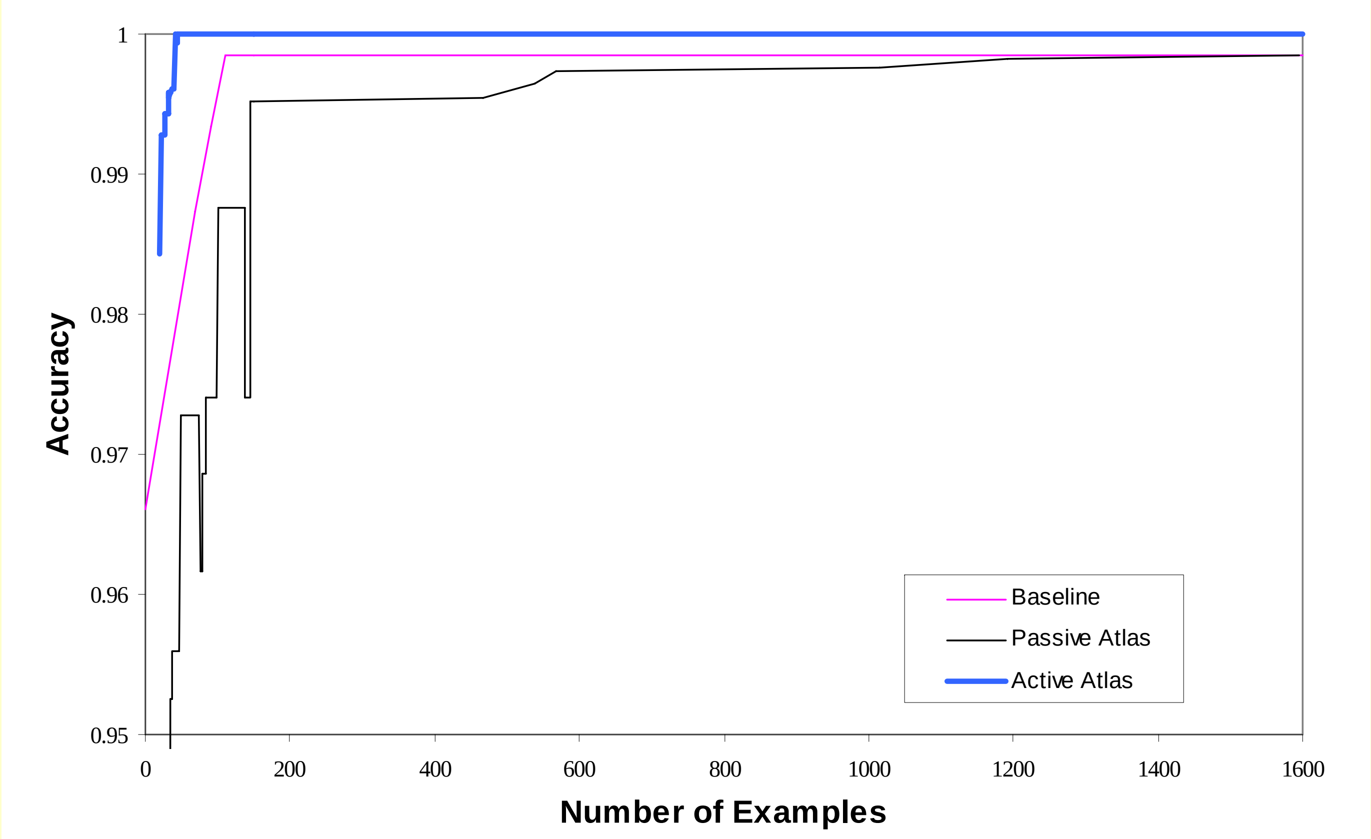
    CG **(only Stemming)**
- Three Variations of Active Atlas
  - Without Transformation weight learning
  - Without using Dissimilarity for choosing queries
  - Without enforcing One-to-One Relationship
- Learned Weights and Rules

Restaurant Domain

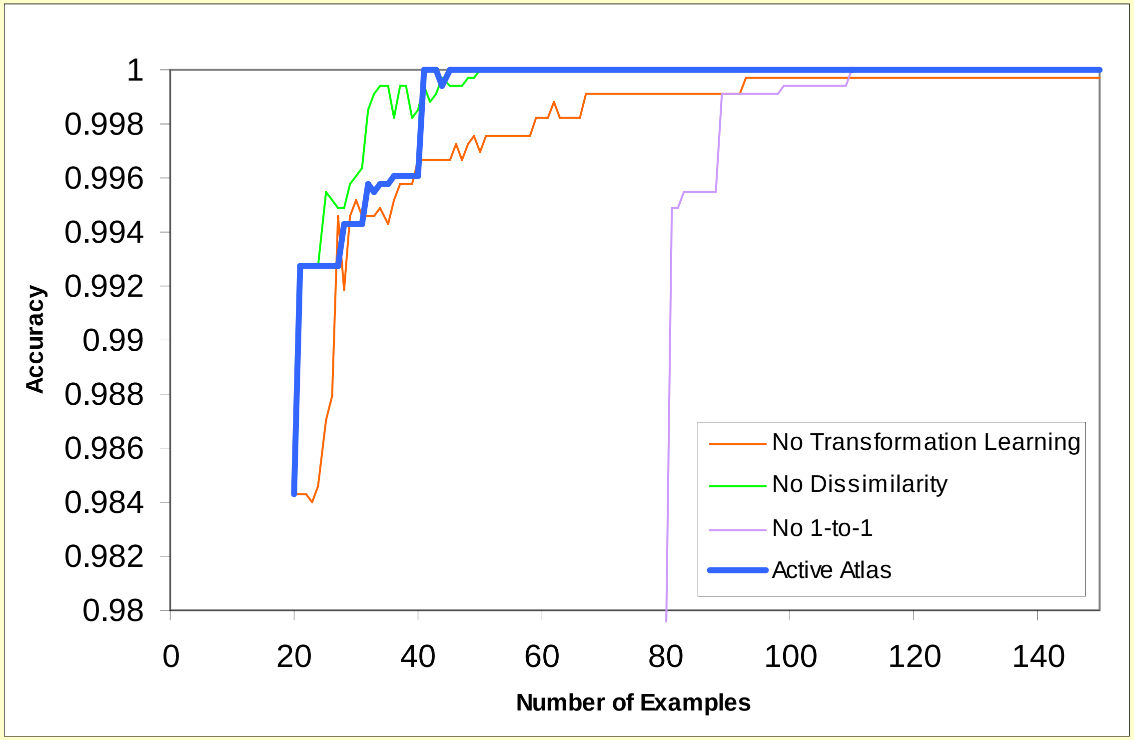


112 mapped objects / 3310 mappings proposed

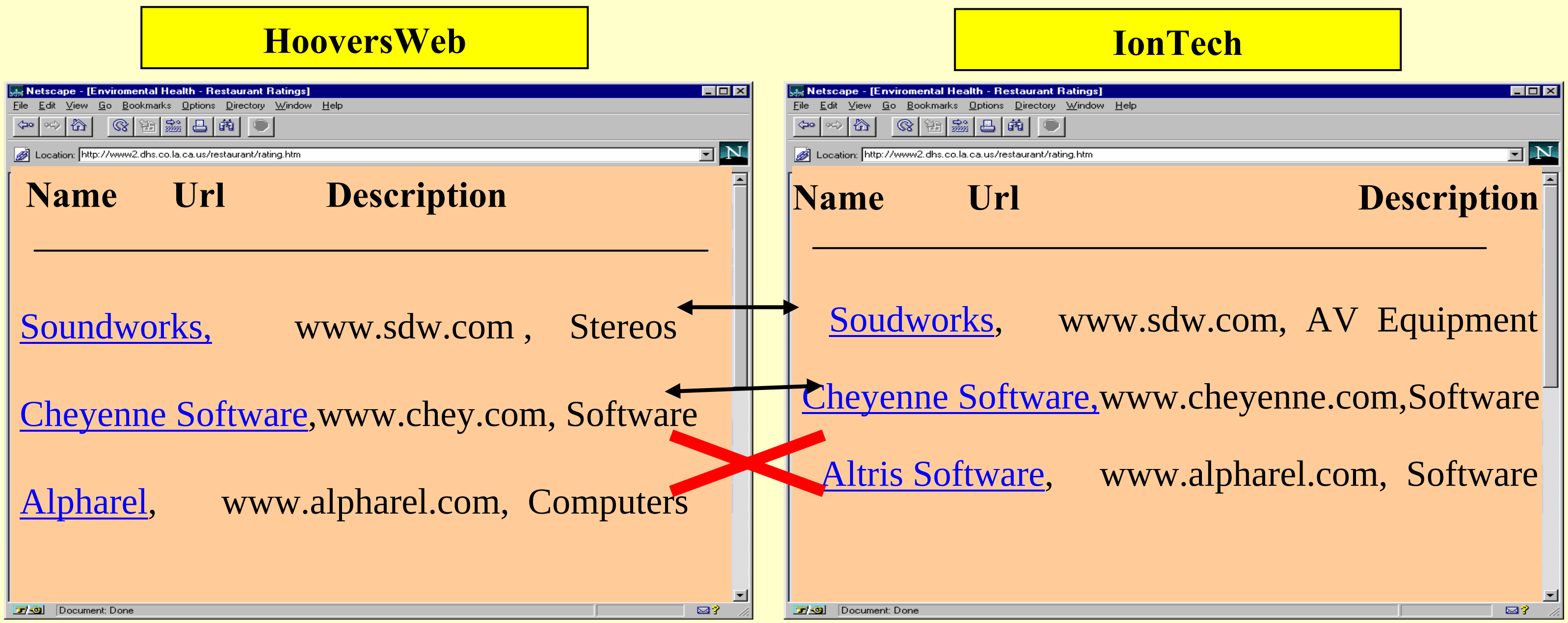
Restaurant Results



Active Atlas Results

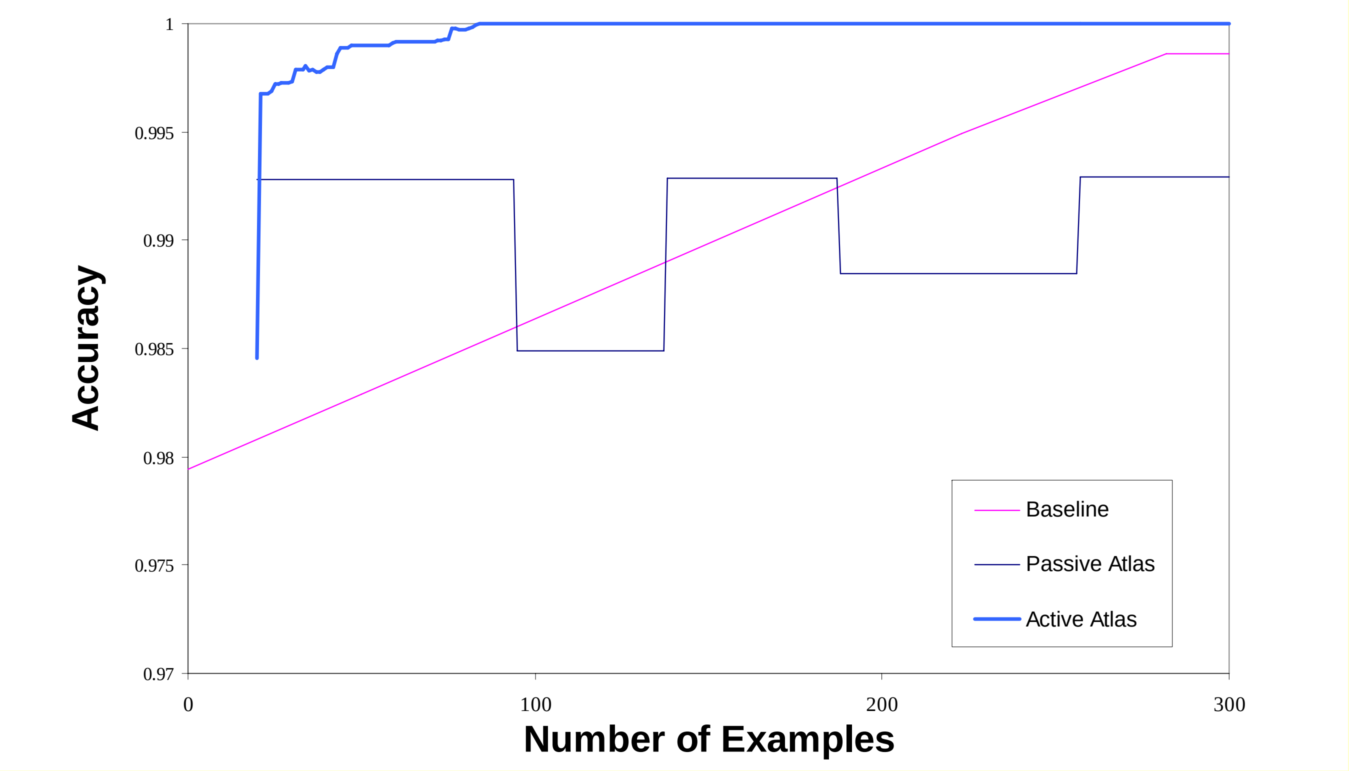


Company Domain

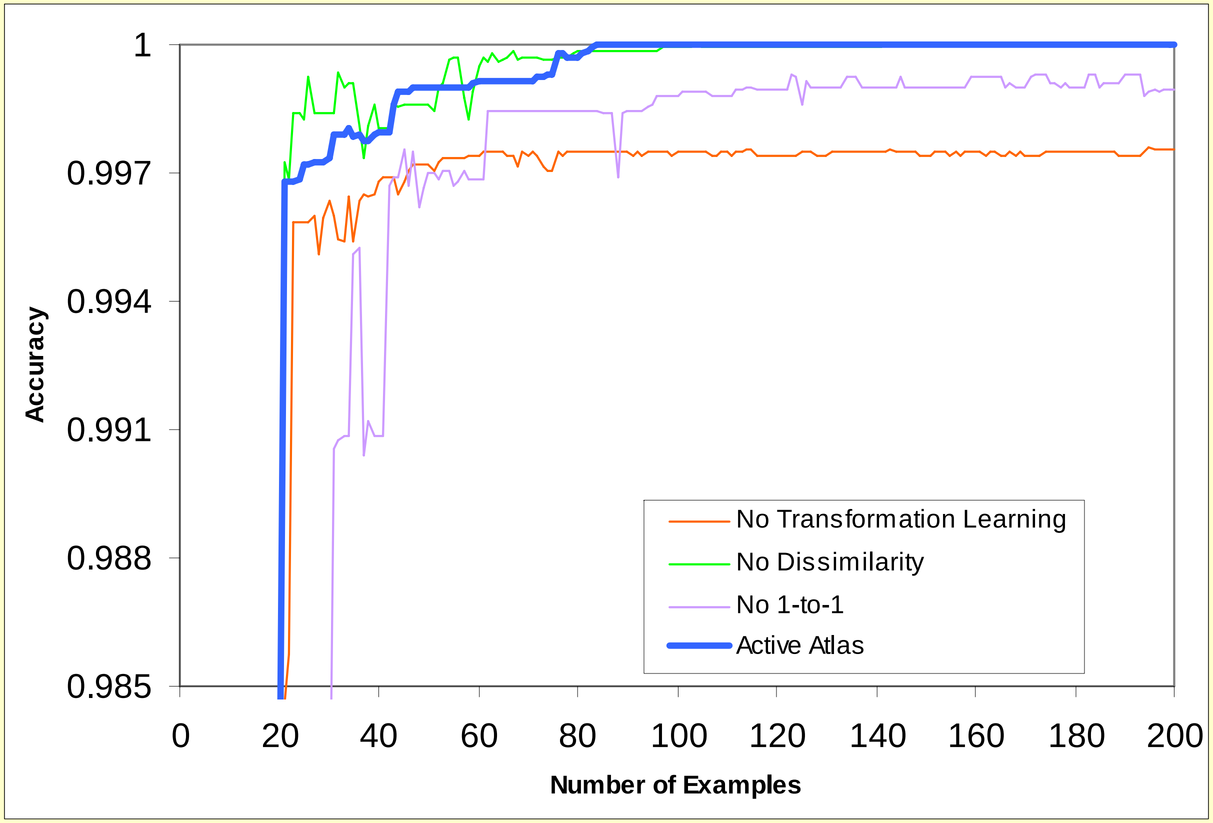


294 mapped objects / 14303 mappings proposed

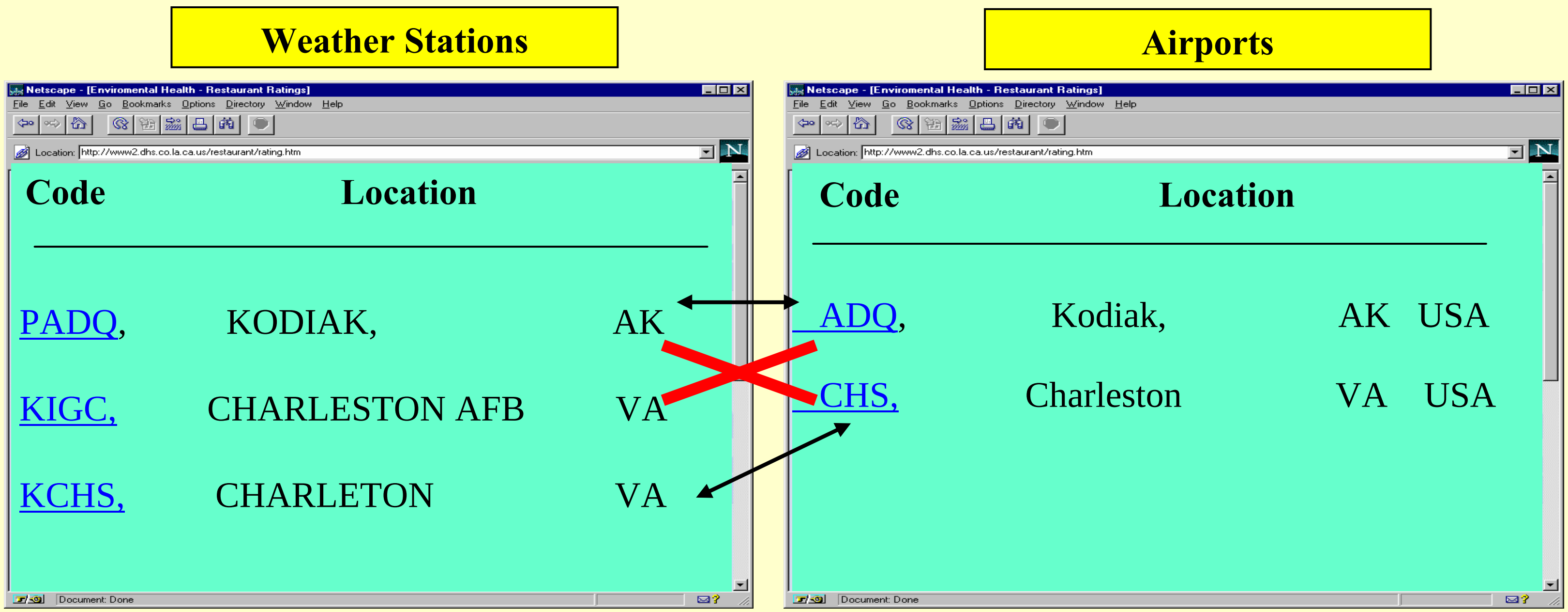
Company Results



Active Atlas Results

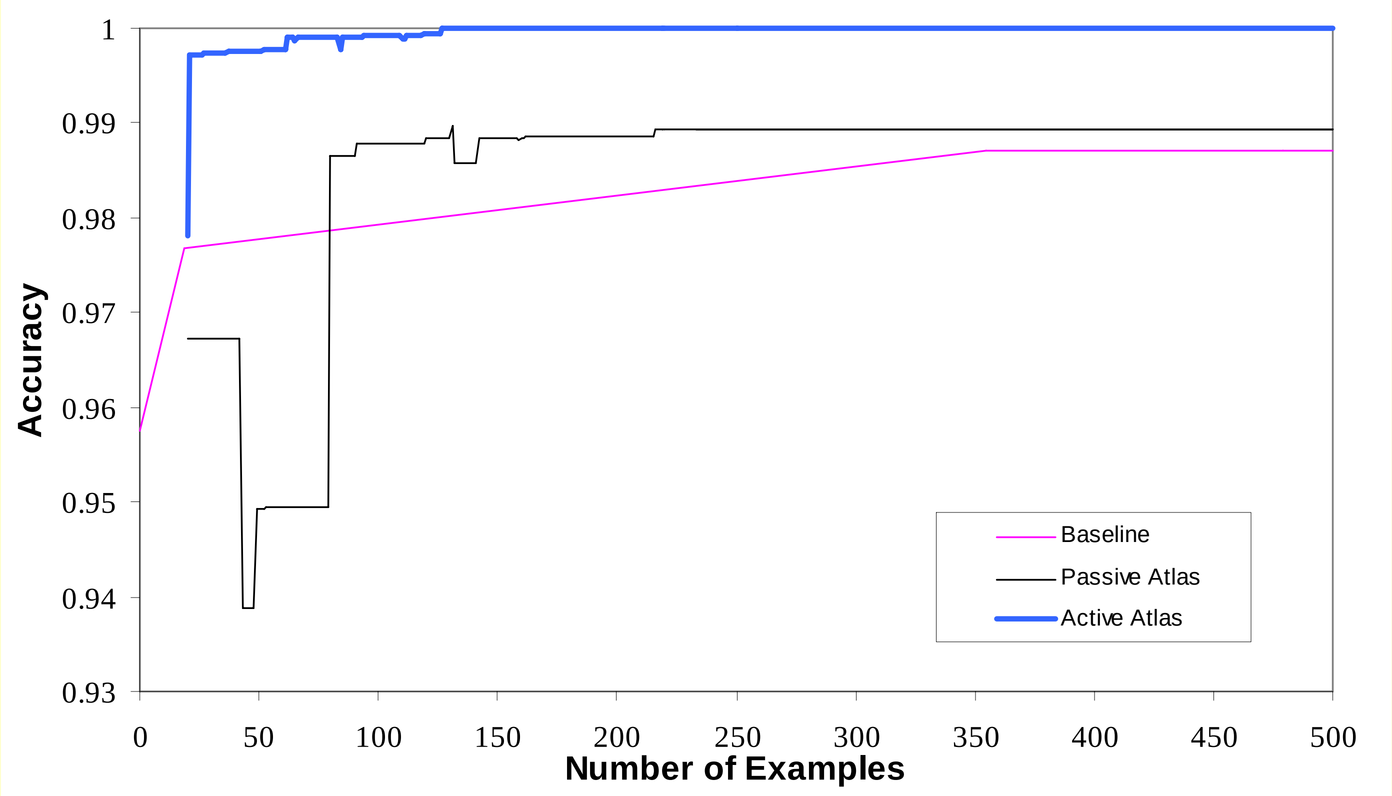


Airport/Weather Domain

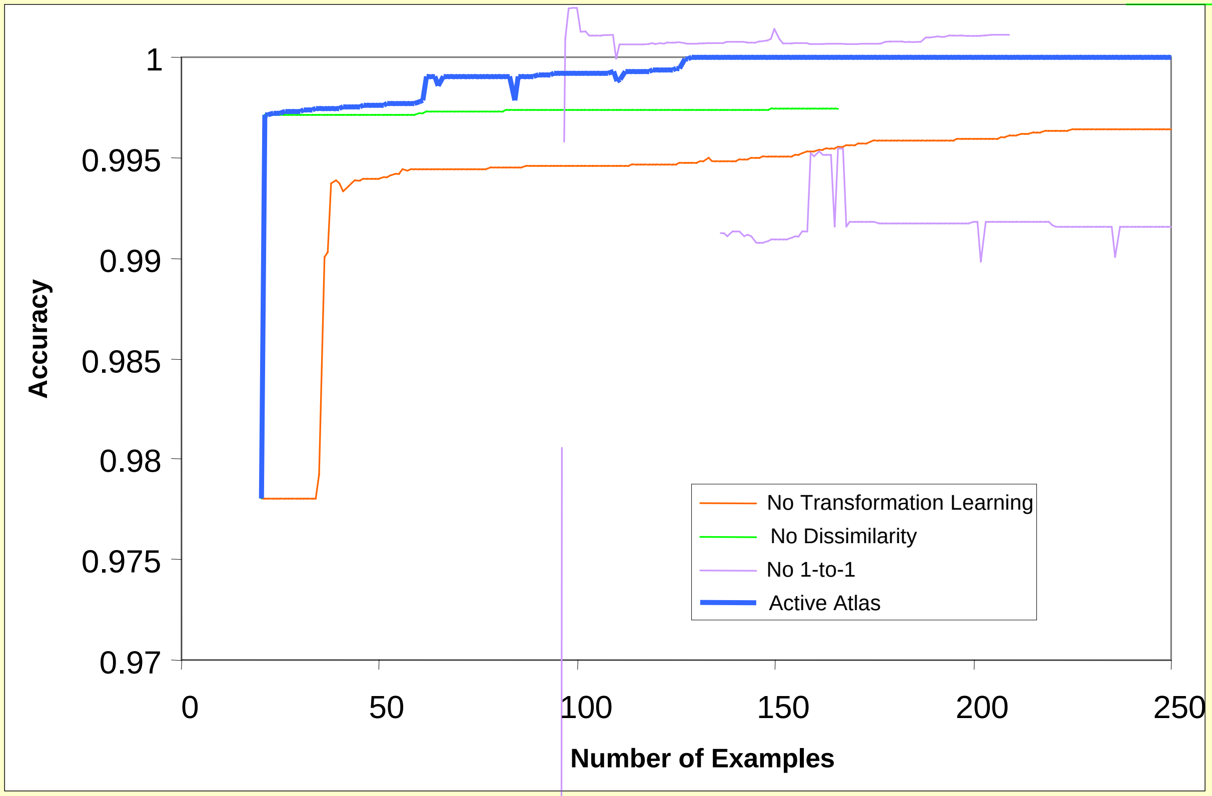


418 mapped objects / 17120 mappings proposed

Airport/Weather Results



Active Atlas Results



Applying Learned Weights & Rules

Application Domain	Total Number of Examples	Total Number of Test Examples	Average Accuracy
Restaurant	3310	662	.9989
Company	14303	2861	.9995
Airport	17120	3624	.9960

Related Work

- Key characteristics
  - Manual methods to customize rules for each domain
  - User-applied fixed threshold to match objects
  - No transformation weight learning
- Related work areas
  - Database Community (Ganesh et al, Monge&Elkan)
  - Information Retrieval (Cohen)
  - Sensor Fusion (Huang & Russell)
  - Record Linkage (Jaro, Winkler)

Database Community

- Removing duplicate records
  - Hernandez&Stolfo, Monge&Elkan
  - User-defined transformations
  - Manual generated mapping rules
- Data Mining
  - Work conducted by Pinheiro&Sun
    - User-defined transformations
    - Learned attribute model (supervised learning)
  - Work by Ganesh et al
    - Learned mapping rules (decision tree learner)

Information Retrieval

- Whirl Information Retrieval System (Cohen)
- Stemming is the only transformation
  - "CPK" would not match "California Pizza Kitchen."
- The user reviews ranked set of objects to determine the threshold of the match

Sensor Fusion & Record Linkage

- Appearance Model (Huang & Russell)
  - Appearance probabilities will not be helpful for an attribute with a unique set of instances
  - (“Art’s Deli”, “Art’s Delicatessen”)
- Record Linkage community (Jaro, Winkler)
  - Hand tailored domain specific transformations
  - The EM algorithm is applied to classify the data into three classes:
    - Matched
    - Not matched
    - To be reviewed
  - Unsupervised learning technique

Conclusions

- Novel approach combines both mapping rule learning and transformation weight learning to create a robust object identification system
- Learns to classify examples with 100% accuracy
- Requires less user involvement than other baseline techniques (Passive Atlas & Information Retrieal)

Future Work

- Noise in User Labels
- Learning Specific Transformations Weights
- Learning to Generate New Transformations
- Scaling: Approach currently applied to sets of examples on the order of 10,000. What are the issues for millions?
- Reconciling textual differences