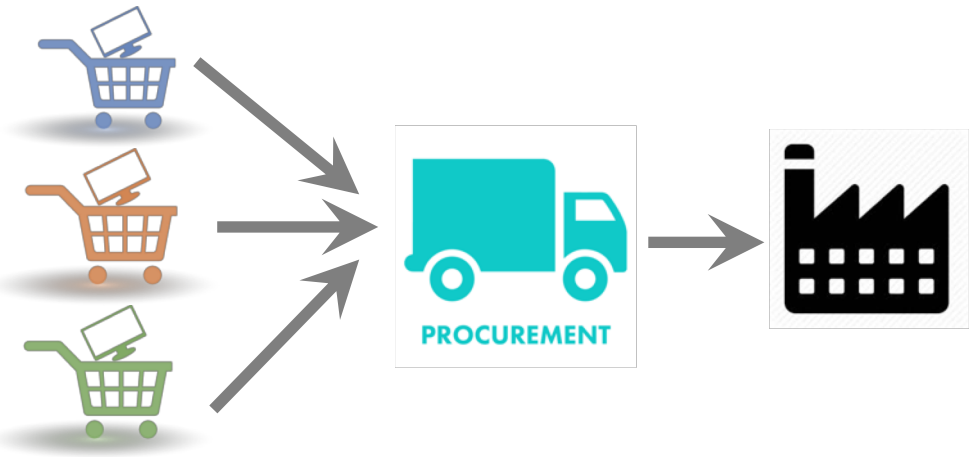
Unsupervised String Transformation Learning for Entity Consolidation

Dong Deng, Wenbo Tao, Ziawasch Abedjan, Ahmed Elmagarmid, Ihab Ilyas, Guoliang Li, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, Nan Tang

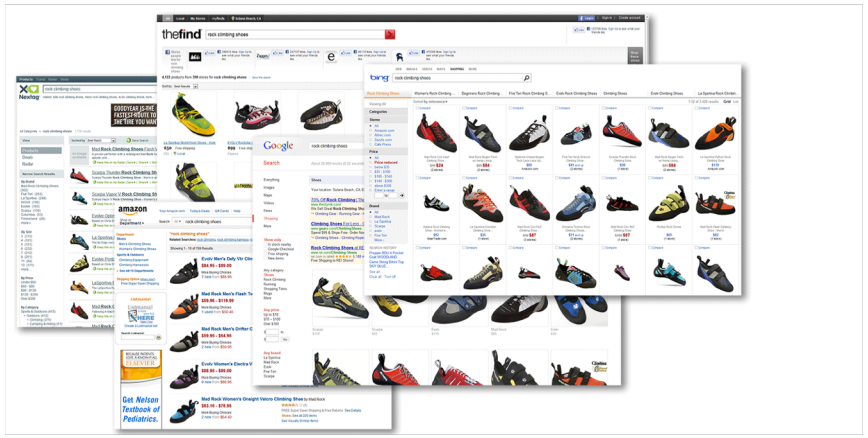
Data Integration is Ubiquitous

- Fundamental problem in numerous applications



"GE estimates they would save $100M/year by integrating orders for better pricing"

FORTUNE



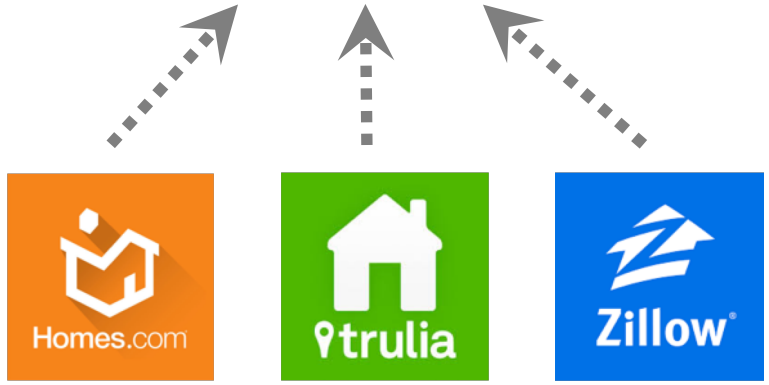
Comparison Shopping

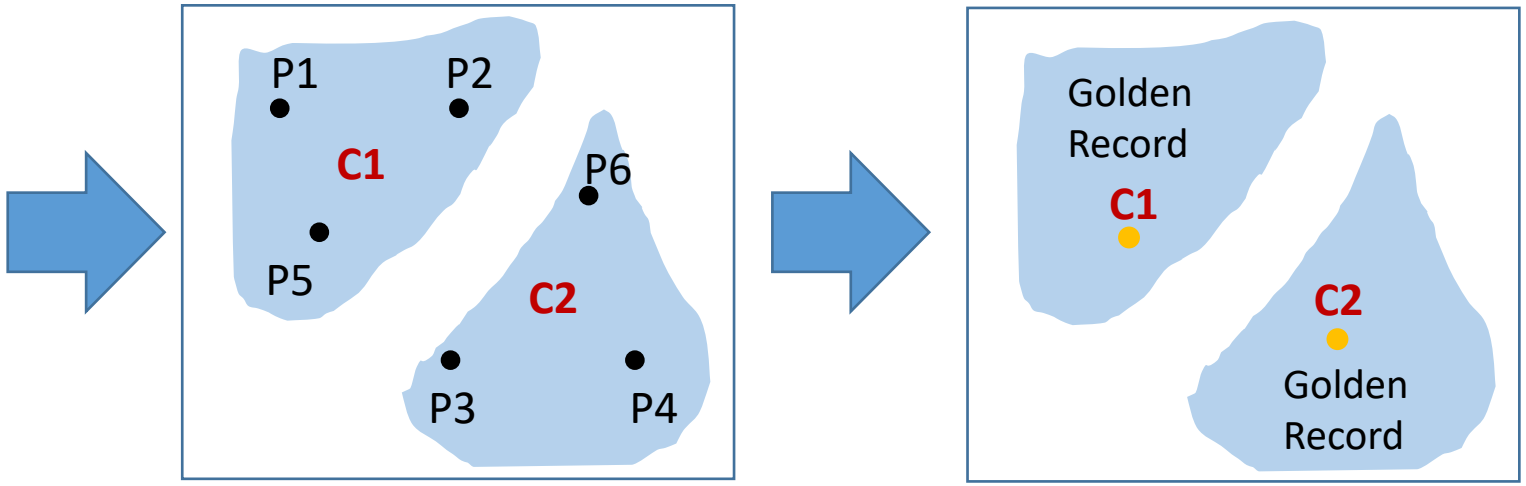
A Closer Look at Data Integration

Entity Resolution:
Find Duplicate Records

Entity Consolidation:
Merge Duplicate Records

ID	Name	Address	Telephone
P1	Mary Lee	9 St, Wisconsin	(718) 453-0681
P2	M. Lee	9th St, WI	7184530681
P3	James Smith	3rd E Ave, CA	212-213-2888x264
P4	J. Smith	3 E Avenue, CA	(212) 213-2888
P5	Lee, Mary	9 Street, WI	+1-718-453-0681
P6	Smith, James	5th Street, WA	+1-212-213-2888





Entity Consolidation: *Merge Duplicate Records*

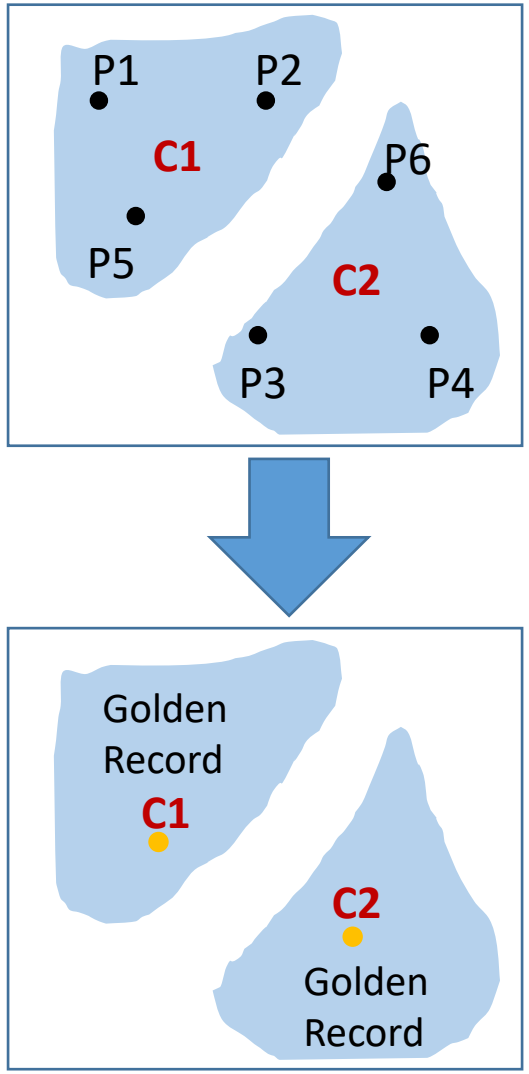
Clusters of Duplicate Records

ID	Name	Address
P1	Mary Lee	9 St, Wisconsin
P2	M. Lee	9th St, WI
P5	Lee, Mary	9 Street, WI
P3	Smith, James	5th Street, WA
P4	James Smith	3rd E Ave, California
P6	J. Smith	3 E Avenue, CA

Conflict Value Pairs

Variant Value Pairs

ID	Name	Address
C1	Mary Lee	9th Street, WI
C2	James Smith	3rd E Avenue, CA



Entity Consolidation: *Merge Duplicate Records*

Clusters of Duplicate Records

ID	Name	Address
P1	Mary Lee	9 St, Wisconsin
P2	M. Lee	9th St, WI
P5	Lee, Mary	9 Street, WI
P3	Smith, James	5th Street, WA
P4	James Smith	3rd E Ave, California
P6	J. Smith	3 E Avenue, CA

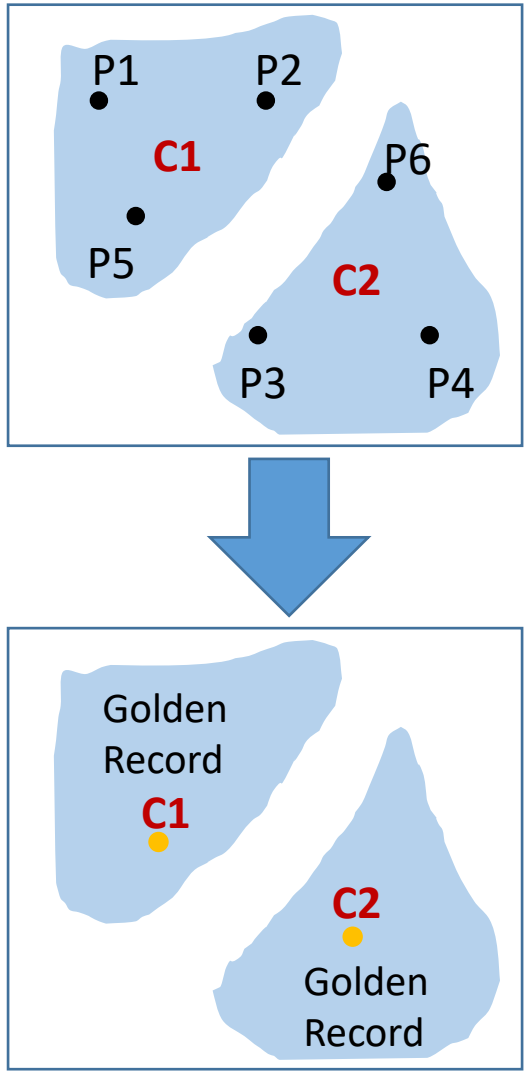
ID	Name	Address
C1	Mary Lee	9th Street, WI
C2	James Smith	3rd E Avenue, CA

majority vote
truth discovery
data fusion, etc

Conflict Value Pairs

Variant Value Pairs

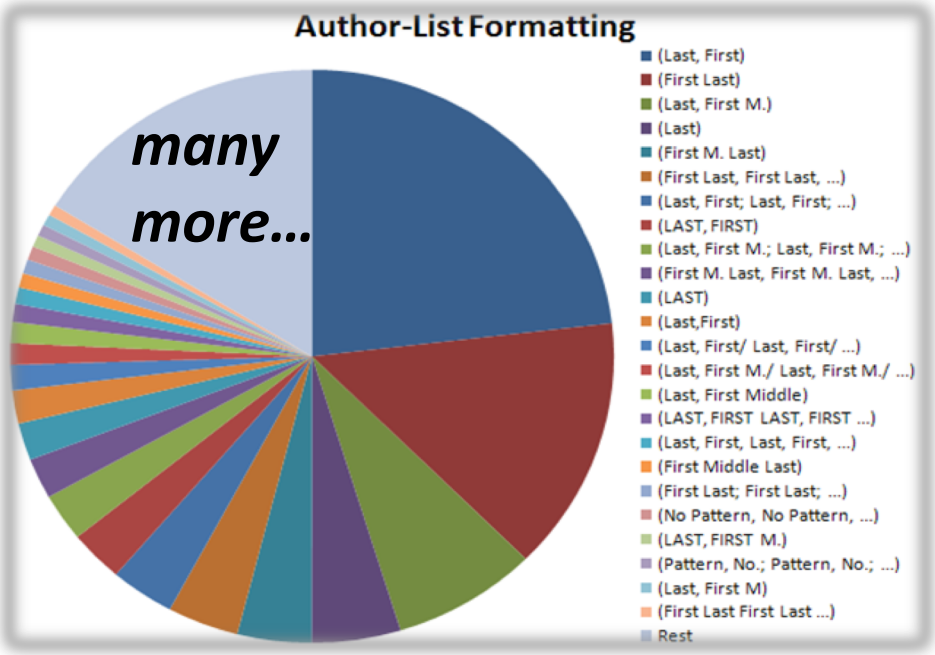
the focus of this work



Data Variety and Inconsistency

- A big issue in data integration and entity consolidation

AbeBooks.com®



- Largely done by hand, labor intensive & error prone

Example: Match & Merge Affiliations

"I've ***created a number of rules*** to map together alternative organization names and misspellings …"

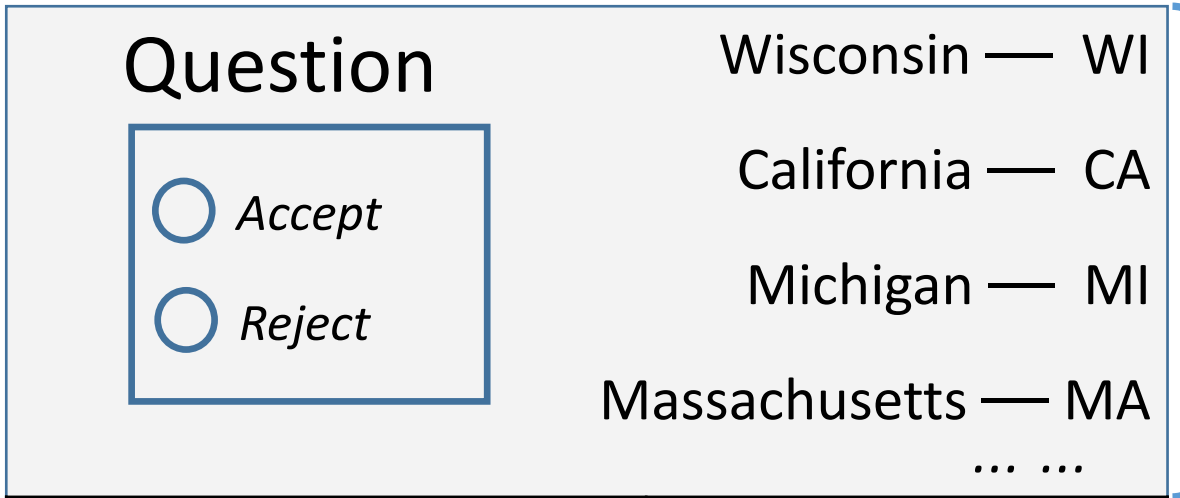
"I wrote a bunch of ***manual patterns*** to map names to canonical versions, although it is likely that I **still missed some cases** …"

"There are 10 different ways "Google" is represented and 11 different versions of "IBM," so that required some ***manual scrubbing*** …"

"I ***manually collapse*** …"

http://www.marekrei.com/blog/ml-nlp-publications-in-2017/
https://unsupervisedmethods.com/nips-accepted-papers-stats-26f124843aa0

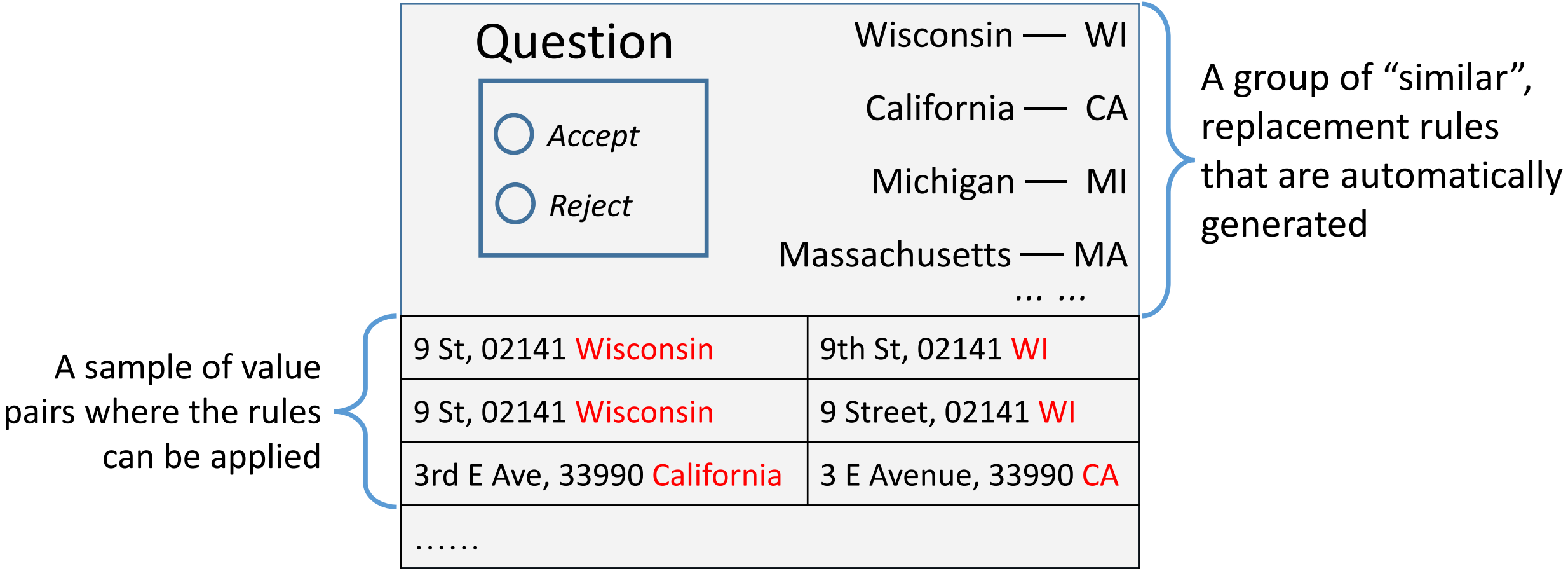
Human Answer Automatic Generated Questions



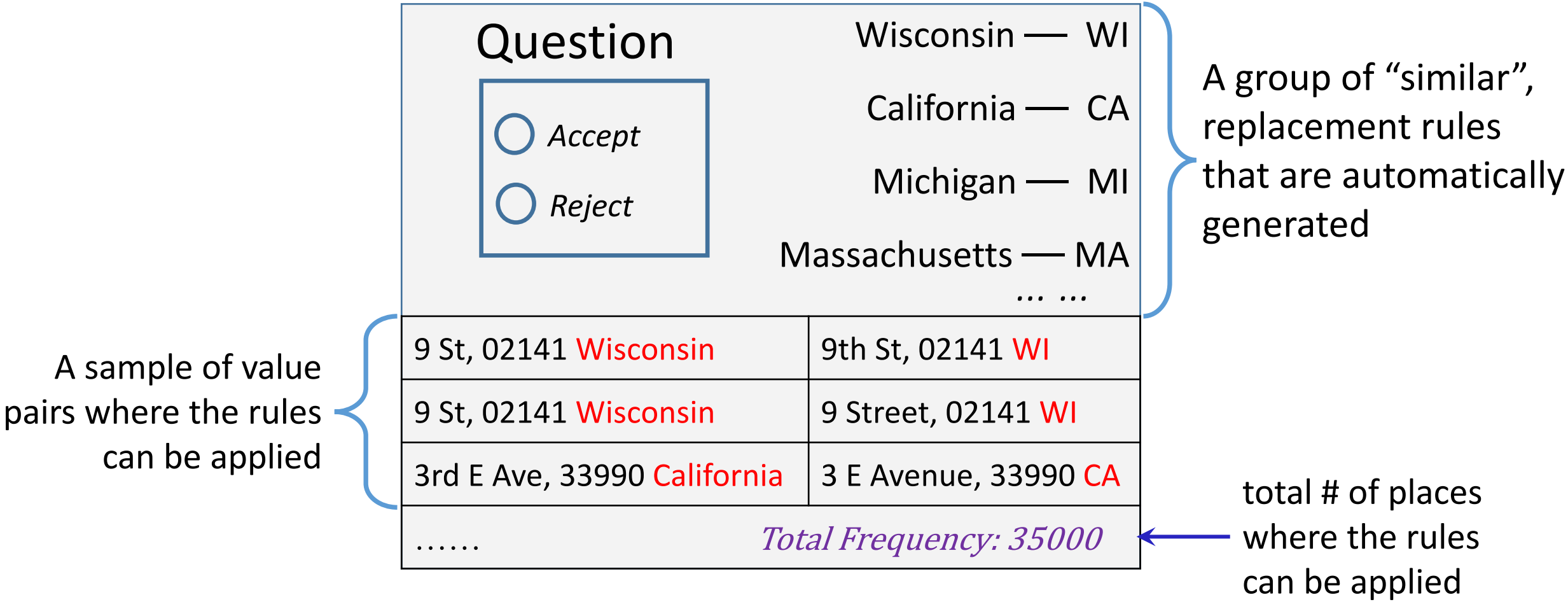
9 St, 02141 Wisconsin	9th St, 02141 WI
9 St, 02141 Wisconsin	9 Street, 02141 WI
3rd E Ave, 33990 California	3 E Avenue, 33990 CA
.....	

A group of “similar”, replacement rules that are automatically generated

Human Answer Automatic Generated Questions



Human Answer Automatic Generated Questions



Questions are asked in frequency decreasing order

Generating & Grouping Rules

Name
Mary Lee
M. Lee
Lee, Mary
Smith, James
James Smith
J. Smith
S. David
Brown, Alex
Alex Brown

Mary Lee — M. Lee
Lee, Mary — Mary Lee
Lee, Mary — M. Lee
Smith, James — James Smith
Smith, James — J. Smith
James Smith — J. Smith
Brown, Alex — S. David
Brown, Alex — Alex Brown
Alex Brown — S. David

Clusters on 1 Column

Candidate Replacement Rules

Generating & Grouping Rules

Name
Mary Lee
M. Lee
Lee, Mary
Smith, James
James Smith
J. Smith
S. David
Brown, Alex
Alex Brown

Clusters on 1 Column

Mary Lee — M. Lee
Lee, Mary — Mary Lee
Lee, Mary — M. Lee
Smith, James — James Smith
Smith, James — J. Smith
James Smith — J. Smith
Brown, Alex — S. David
Brown, Alex — Alex Brown
Alex Brown — S. David

Candidate Replacement Rules

Group by way of transforming

Lee, Mary — Mary Lee
Smith, James — James Smith
Brown, Alex — Alex Brown

Lee, Mary — M. Lee
Smith, James — J. Smith

Mary Lee — M. Lee
James Smith — J. Smith

Brown, Alex — S. David

Alex Brown — S. David

Replacement Rule Groups

Transformation Program [Gulwani. POPL-11]

P_C **the *beginning* of *1st capital* Token**

P_D **the *ending* of *1st lowercase* Token**

direction *k* *predefined regex*

Transformation Program [Gulwani. POPL-11]

P_C **the *beginning* of *1st capital* Token**

P_D **the *ending* of *1st lowercase* Token**

direction *k* *predefined regex*

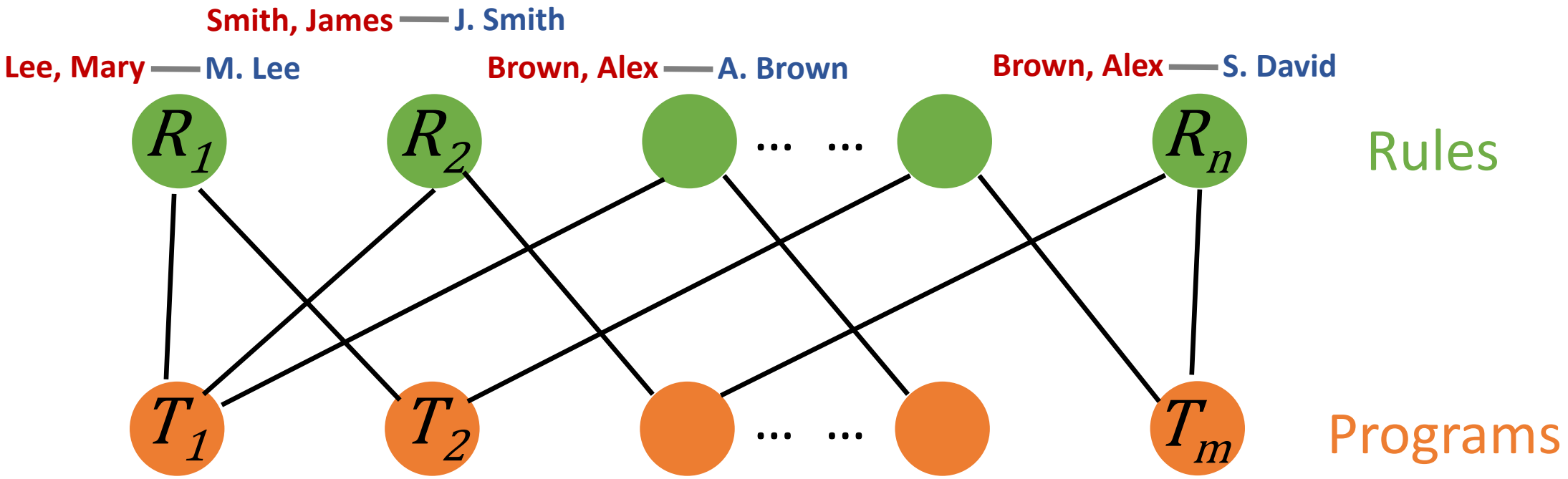
Position: Pos(*Token*, *k*, *dir*)

Substring: $Substr(Pos_1, Pos_2)$

Program: $Substr_1 + Substr_2 + Constant_1$ ...

concatenate substrings and constant strings

Many-to-Many Relationship



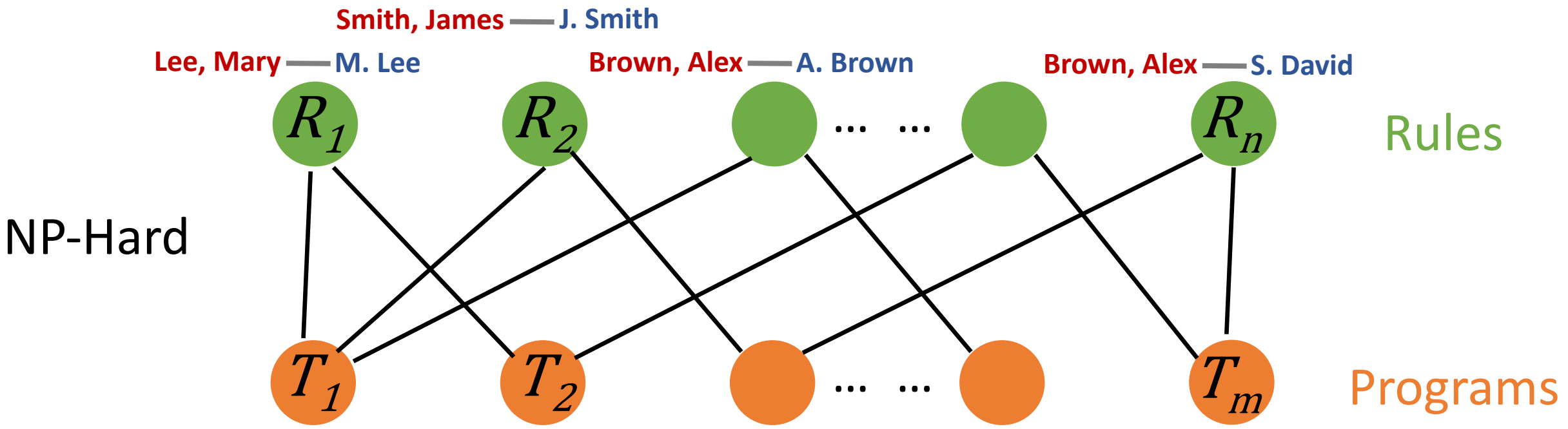
Substr(P_1, P_2) + Constant(". ") + Substr(P_3, P_4)

Rule Grouping Problem

Partition all the candidate rules such that

(i) Replacements in the same partition share a program

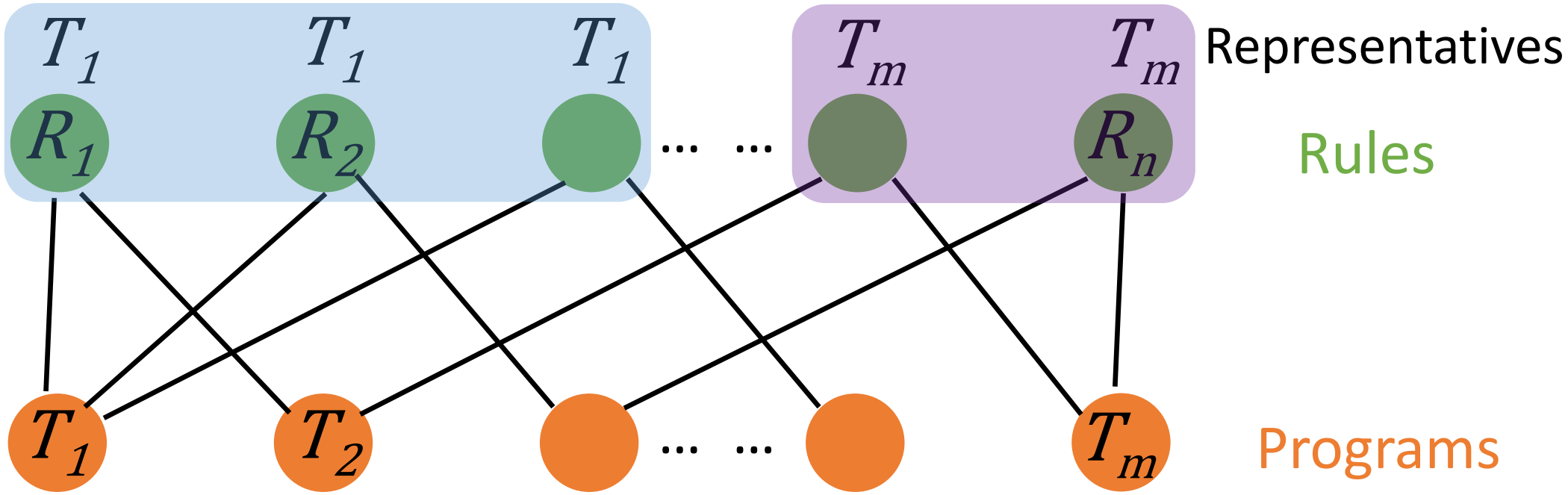
(ii) The number of partitions is minimum



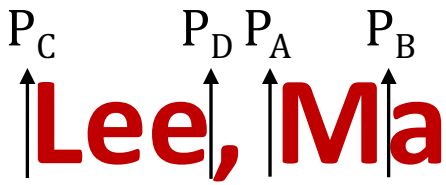
Greedy Algorithm for Rule Grouping

(1) For each rule, pick a representative from all its programs

(2) Rules with the same representative are grouped



Data-Driven Representative Program



M. Lee

Constant("M")+ Constant(". ") + Substr(P_C, P_D)

or

Substr(P_A, P_B) + Constant(". ")+ Substr(P_C, P_D)

Lee, Mary	M. Lee
Smith, James	M. Smith
Brown, Alex	M. Brown

Lee, Mary	M. Lee
Smith, James	J. Smith
Brown, Alex	A. Brown

Representative Program: the one shared by most of candidate rules

Experiment Results

100 questions

Dataset	# Records	# Clusters	# Distinct Value Pairs	Variants	Conflicts	Precision	Recall
Author-List	33,971	1,265	51,538	26.5%	73.5%	.994	.503
Address	17,497	3,038	80,451	18%	82%	.990	.744
Journal-Title	55,617	31,023	81,350	74%	26%	.991	.665

Ground Truth:

Variant Pairs

Conflict Pairs

Identical Value Pairs

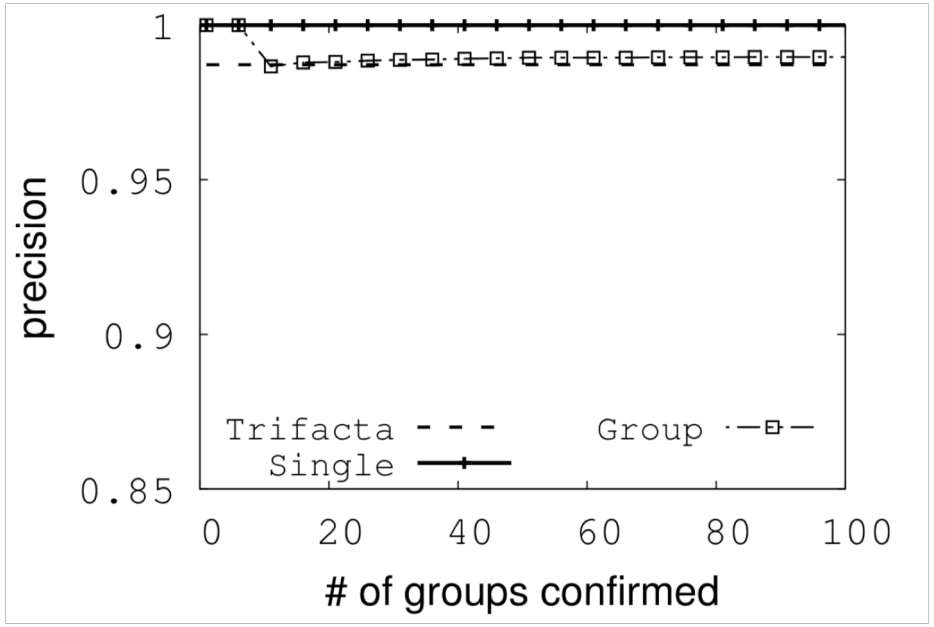
Original Clusters	Updated Clusters
9 St, Wisconsin	9 St, WI
9th St, WI	9th St, WI
9 Street, WI	9 St, WI
5th Street, WA	5th Street, WA
3rd E Ave, California	3rd E Ave, CA
3 E Avenue, CA	3rd E Ave, CA

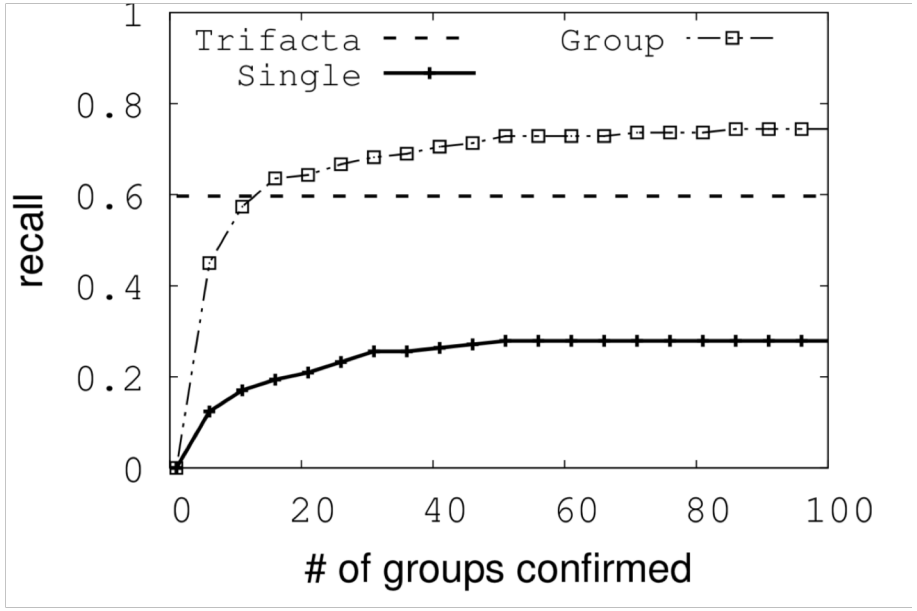
	Identical	Not Identical
Variant Pairs	True Positive	True Negative
Conflict Pairs	False Positive	False Negative

Experiment Results

100 questions

Dataset	# Records	# Clusters	# Distinct Value Pairs	Variants	Conflicts	Precision	Recall
Author-List	33,971	1,265	51,538	26.5%	73.5%	.994	.503
Address	17,497	3,038	80,451	18%	82%	.990	.744
Journal-Title	55,617	31,023	81,350	74%	26%	.991	.665





Address Dataset

Experiment Results

100 questions

Dataset	# Records	# Clusters	# Distinct Value Pairs	Variants	Conflicts	Precision	Recall
Author-List	33,971	1,265	51,538	26.5%	73.5%	.994	.503
Address	17,497	3,038	80,451	18%	82%	.990	.744
Journal-Title	55,617	31,023	81,350	74%	26%	.991	.665

TABLE VI
PRECISION IMPROVEMENT FOR MC

	AUTHORLIST	ADDRESS	JOURNALTITLE
before	.51	.32	.335
after	.65	.47	.840

Take Away

- A semi-automatic way to standardize string formats
- Data-driven group generating
- Achieved very high precision and good recall with small human effort