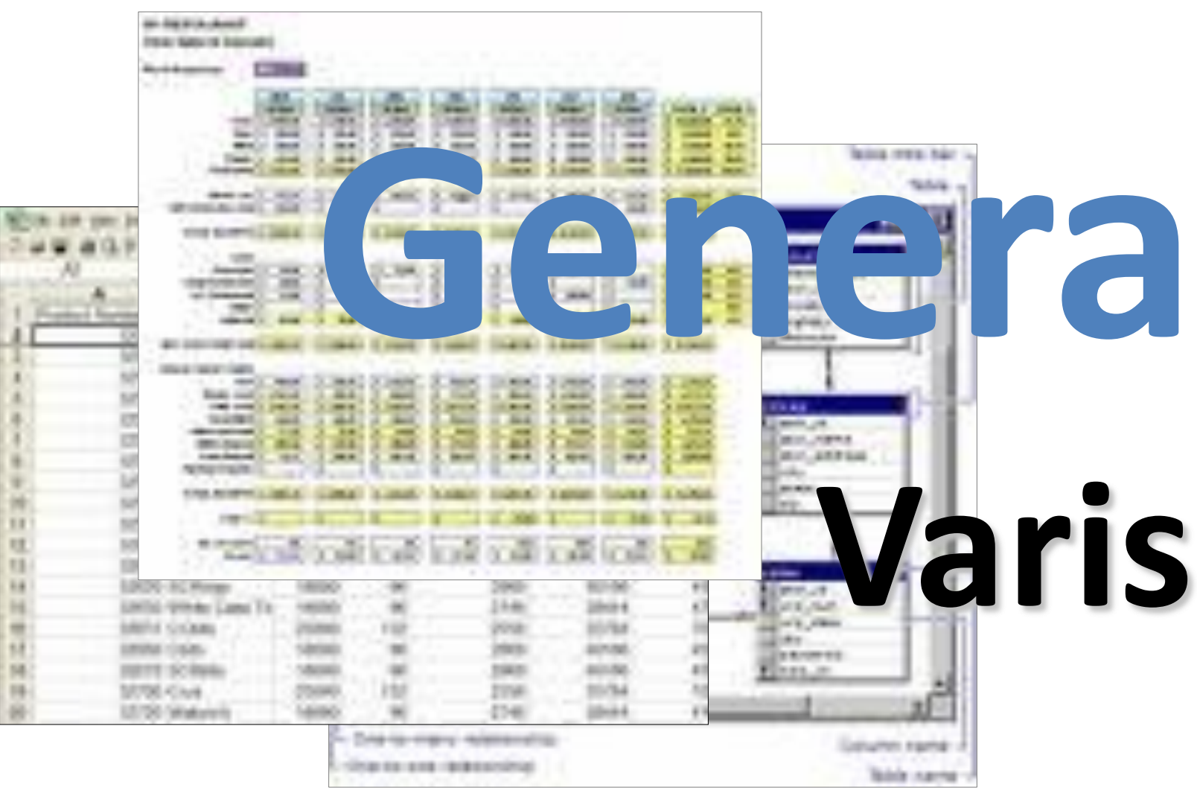
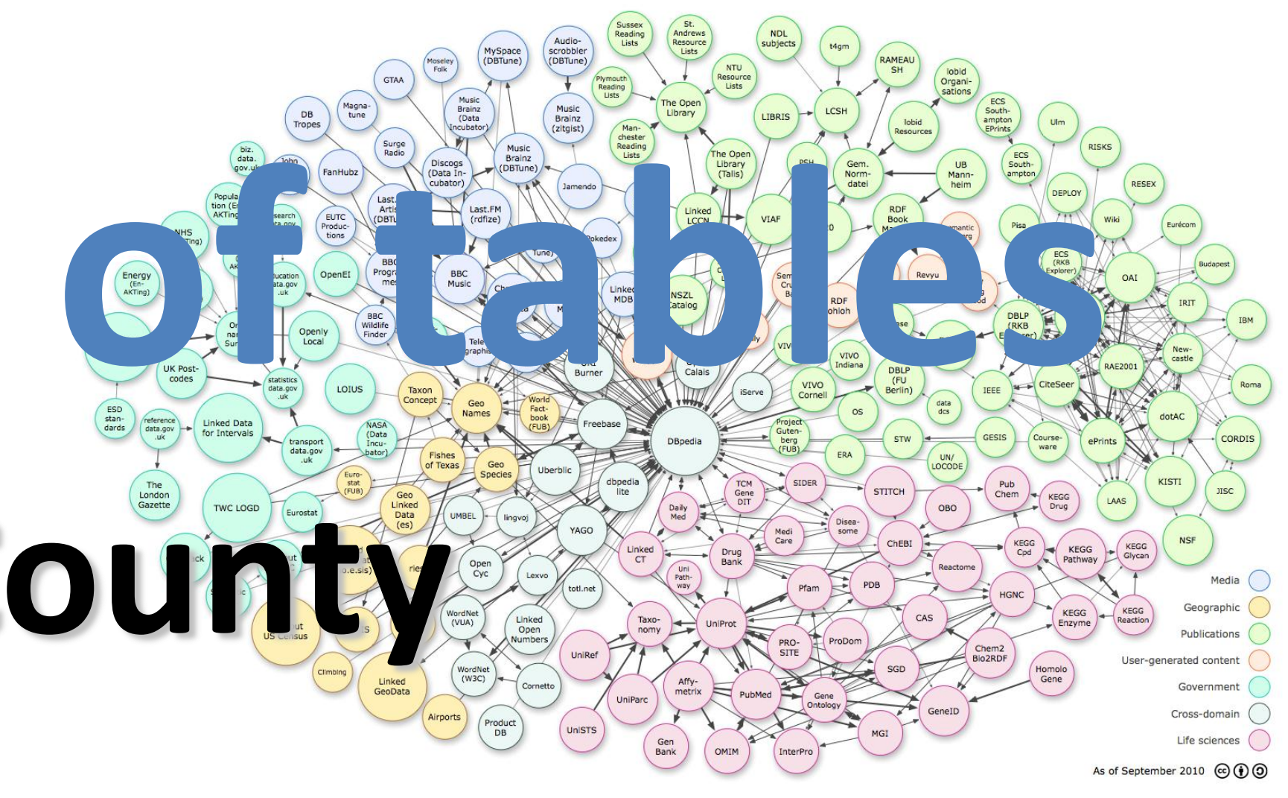
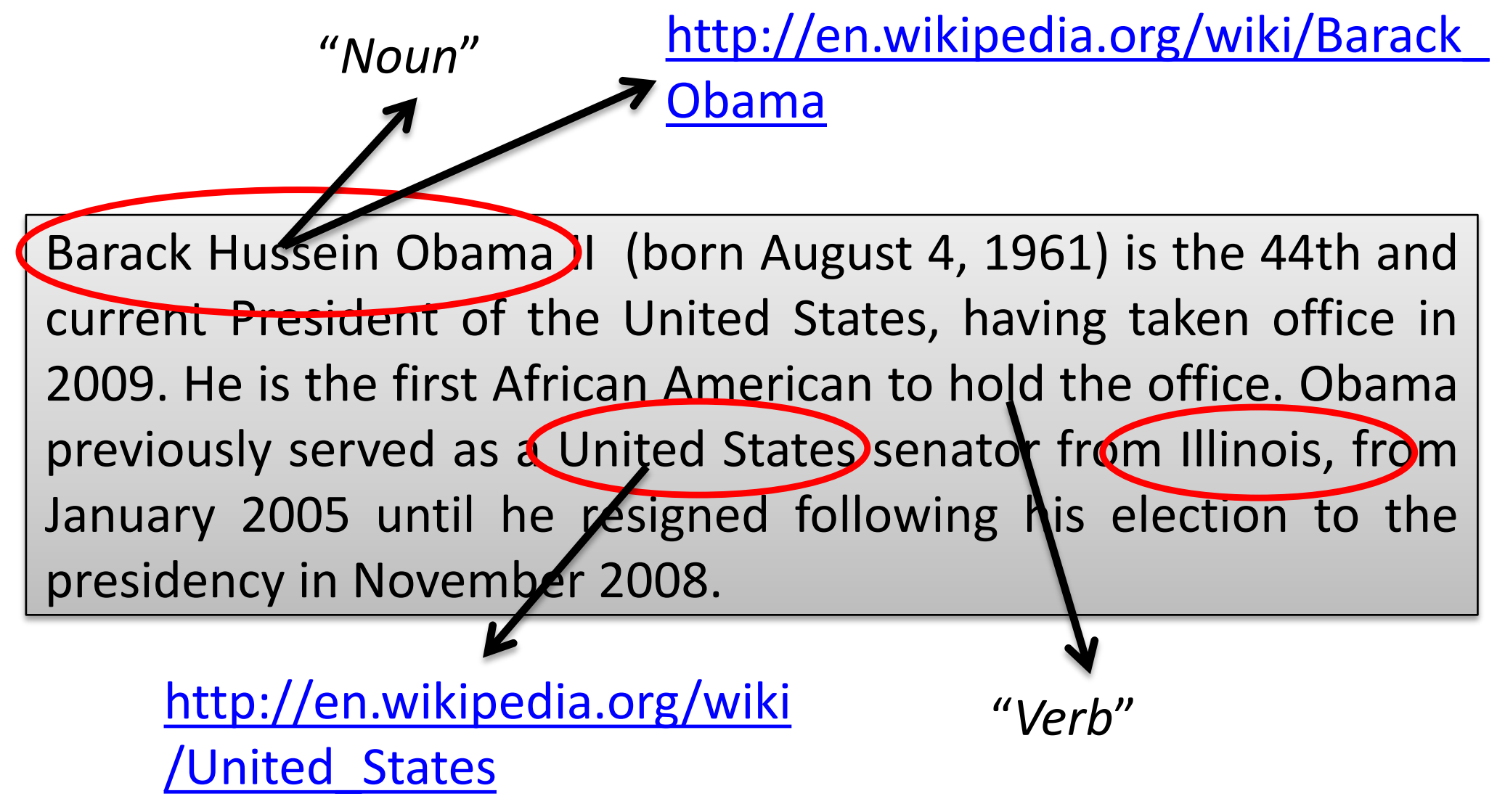




Generating Linked Data by inferring the semantics of tables

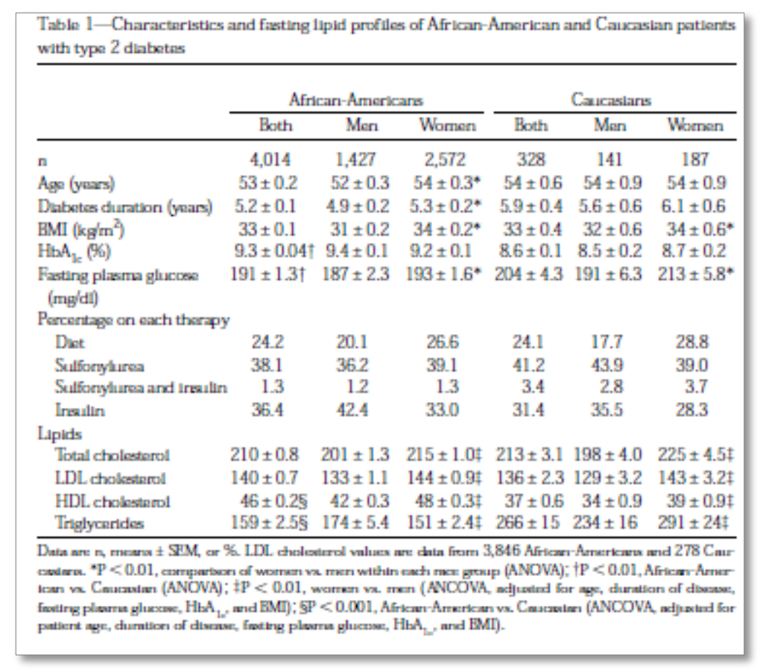
Varish Mulwad (Computer Science), University of Maryland, Baltimore County

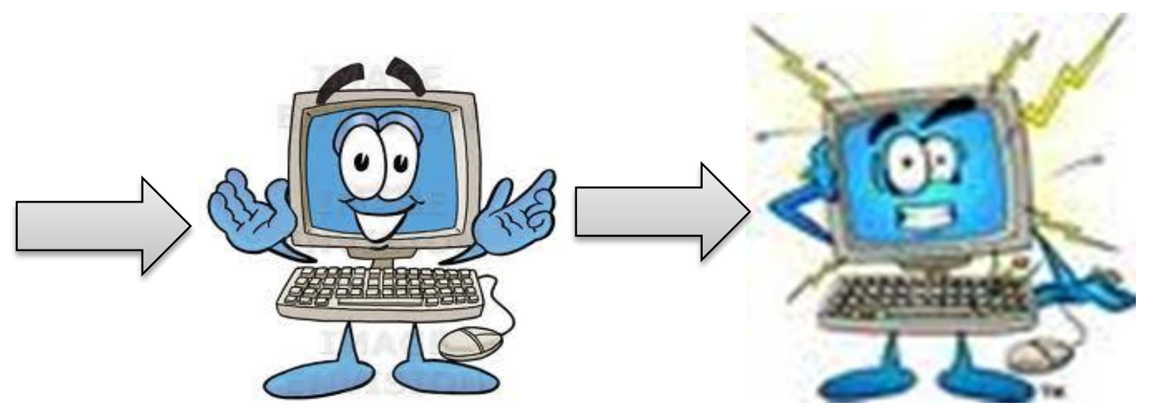
~ 50 years



"Noun"
http://en.wikipedia.org/wiki/Barack Obama
http://en.wikipedia.org/wiki/United_States
"Verb"

Barack Hussein Obama II (born August 4, 1961) is the 44th and current President of the United States, having taken office in 2009. He is the first African American to hold the office. Obama previously served as a United States senator from Illinois, from January 2005 until he resigned following his election to the presidency in November 2008.

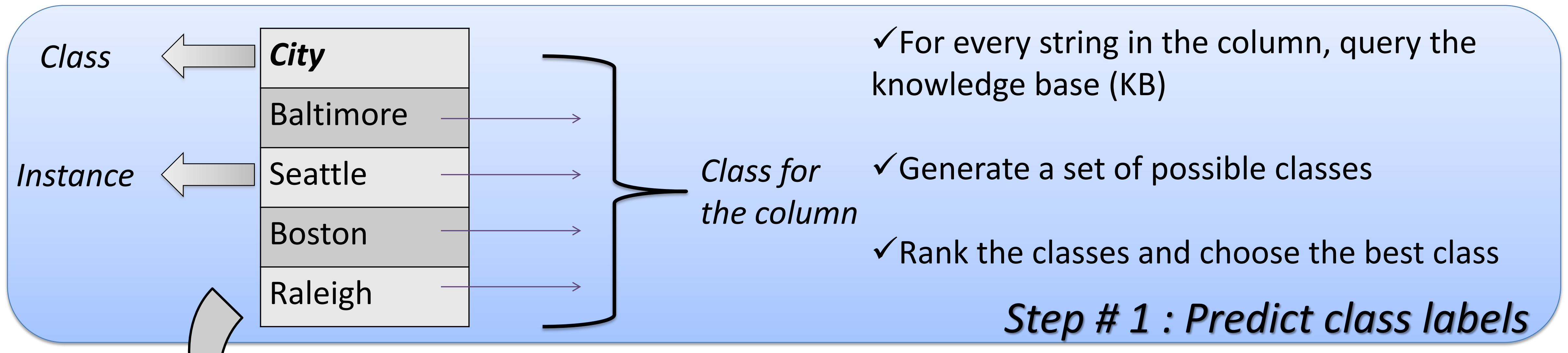




- ❑ Text analysis techniques do not work well with tables
- ❑ Structure may be captured, but what about semantics ?

Tables are everywhere !

- ❑ 154 million high quality relational tables on the web
- ❑ 14 nations sharing data as spreadsheets
- ❑ Can we build systems to exploit this knowledge ?



City
Baltimore
Seattle
Boston
Raleigh

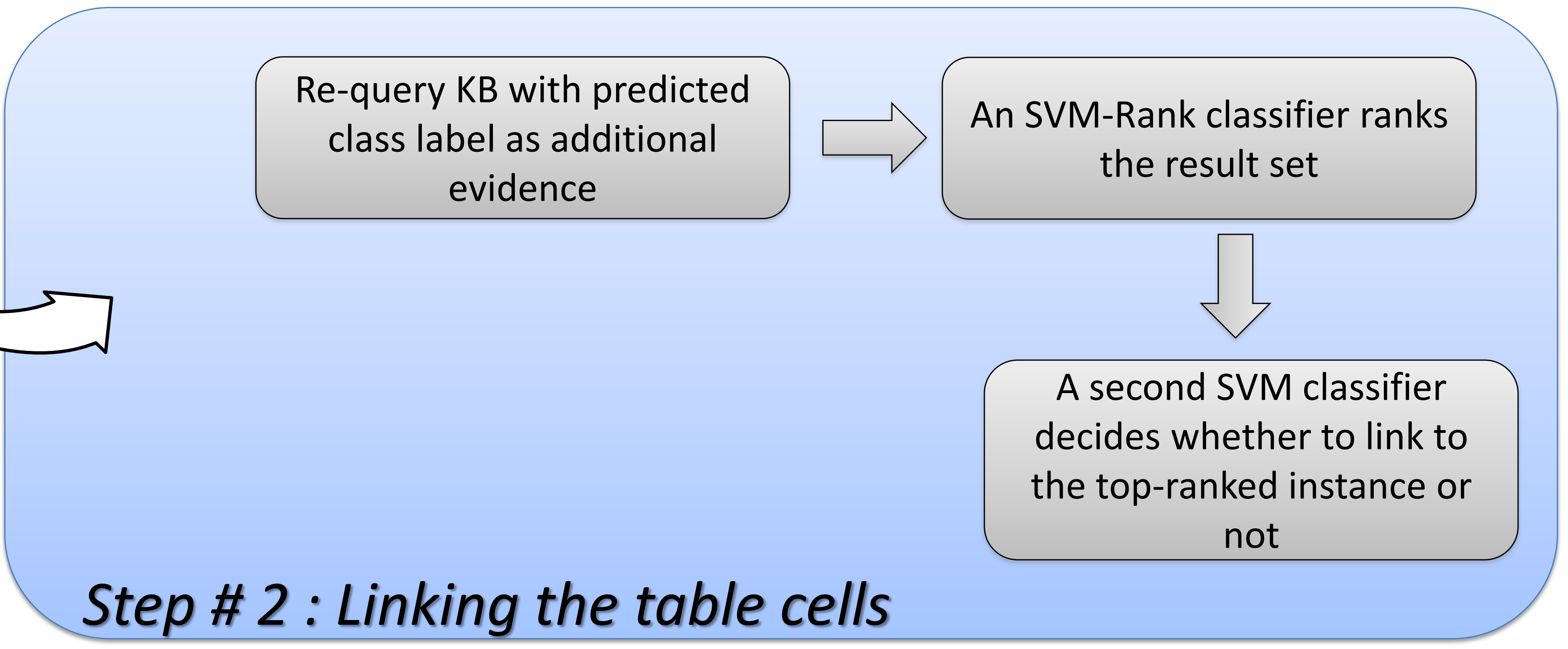
Class

Instance

Class for the column

- ✓For every string in the column, query the knowledge base (KB)
- ✓Generate a set of possible classes
- ✓Rank the classes and choose the best class

Step # 1 : Predict class labels

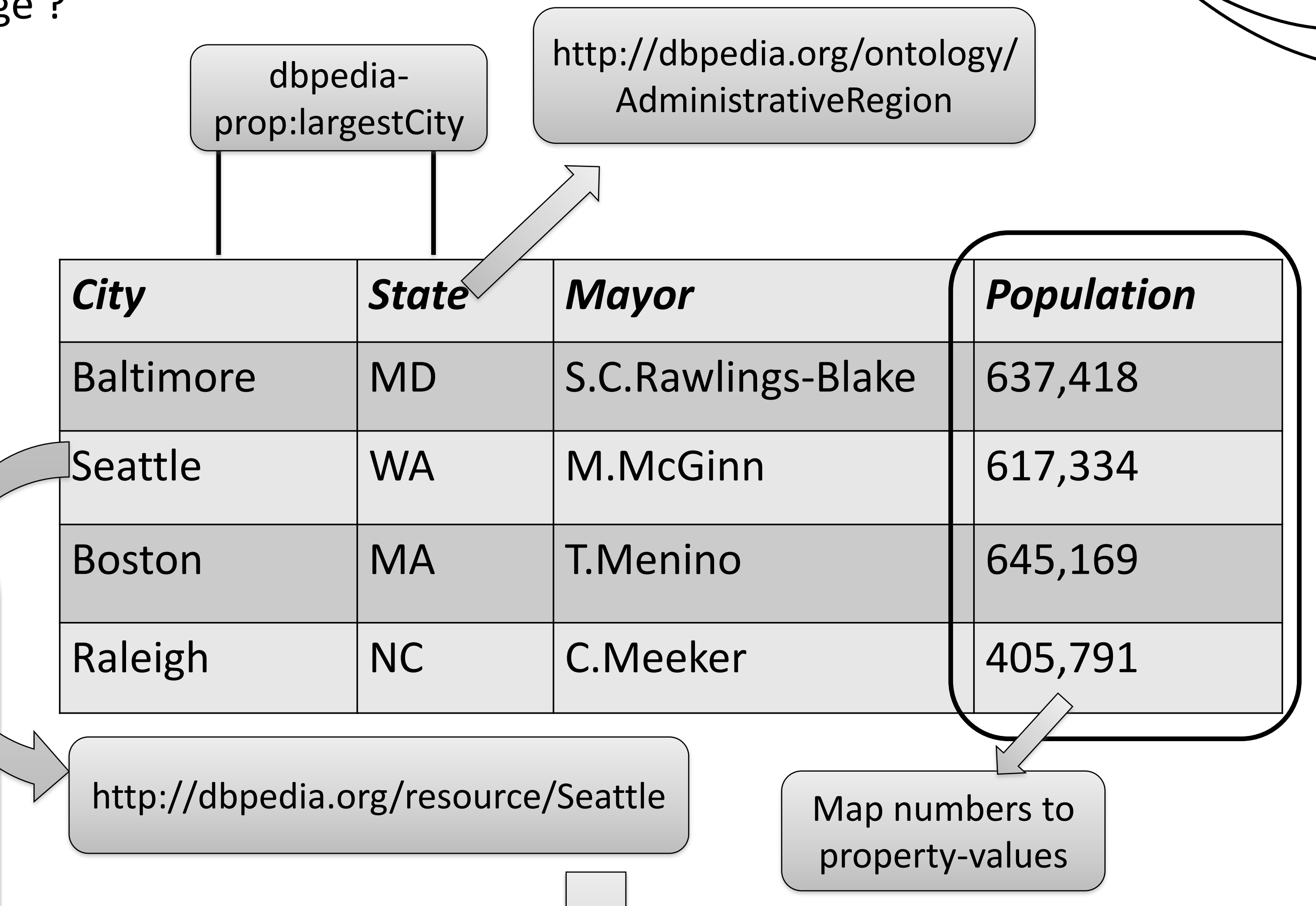


Re-query KB with predicted class label as additional evidence

An SVM-Rank classifier ranks the result set

A second SVM classifier decides whether to link to the top-ranked instance or not

Step # 2 : Linking the table cells



dbpedia-prop:largestCity

http://dbpedia.org/ontology/AdministrativeRegion

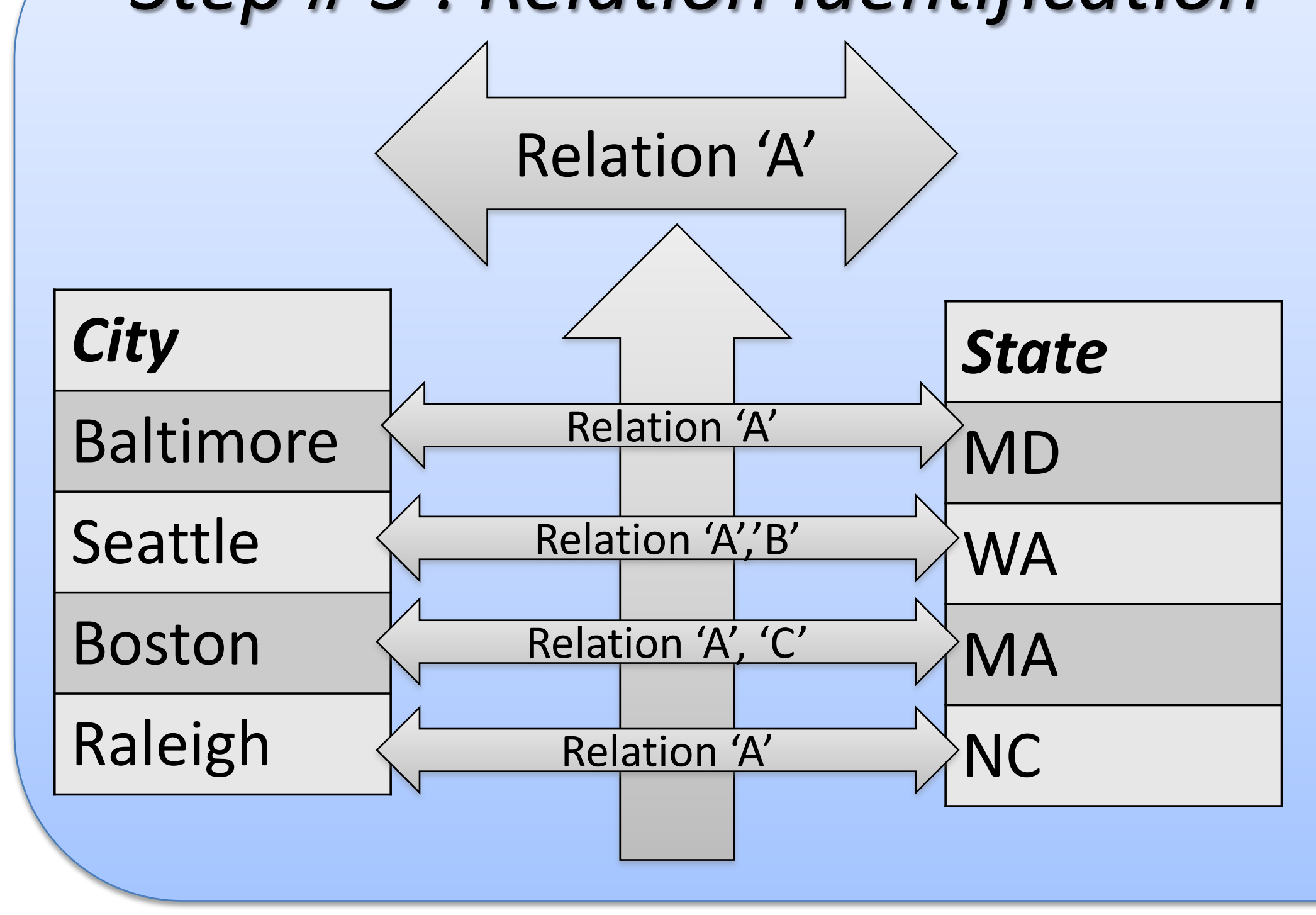
City	State	Mayor	Population
Baltimore	MD	S.C.Rawlings-Blake	637,418
Seattle	WA	M.McGinn	617,334
Boston	MA	T.Menino	645,169
Raleigh	NC	C.Meeker	405,791

http://dbpedia.org/resource/Seattle

Map numbers to property-values

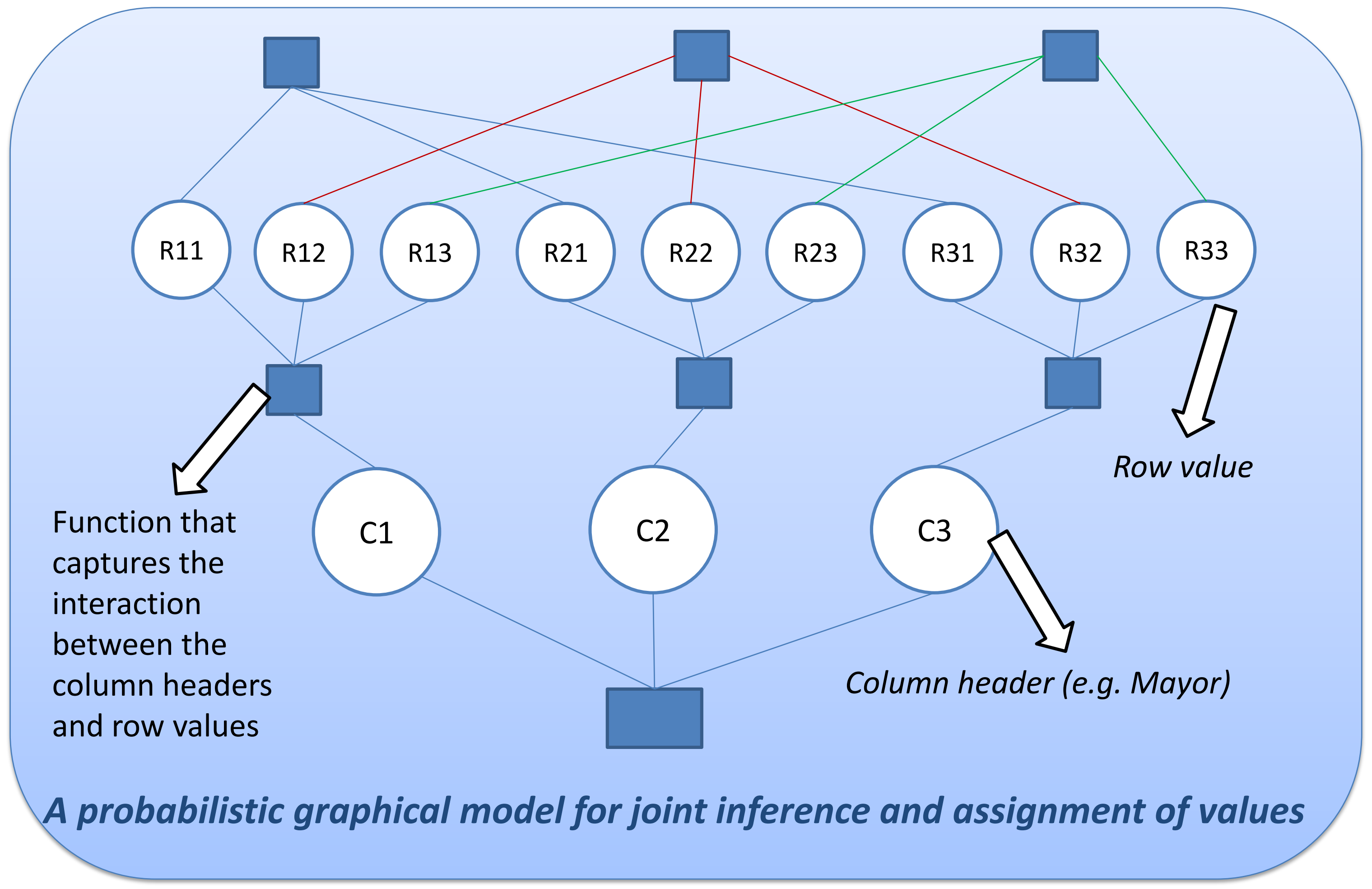
```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .
@prefix dbpprop: <http://dbpedia.org/property/> .
"City"@en is rdfs:label of dbpedia-owl:City .
"State"@en is rdfs:label of dbpedia-owl:AdminstrativeRegion .
"Baltimore"@en is rdfs:label of dbpedia:Baltimore .
dbpedia:Baltimore a dbpedia-owl:City .
"MD"@en is rdfs:label of dbpedia:Maryland .
dbpedia:Maryland a dbpedia-owl:AdministrativeRegion .
dbpprop:LargestCity rdfs:domain dbpedia-owl:AdminstrativeRegion .
dbpprop:LargestCity rdfs:range dbpedia-owl:City .
```

Step # 3 : Relation Identification



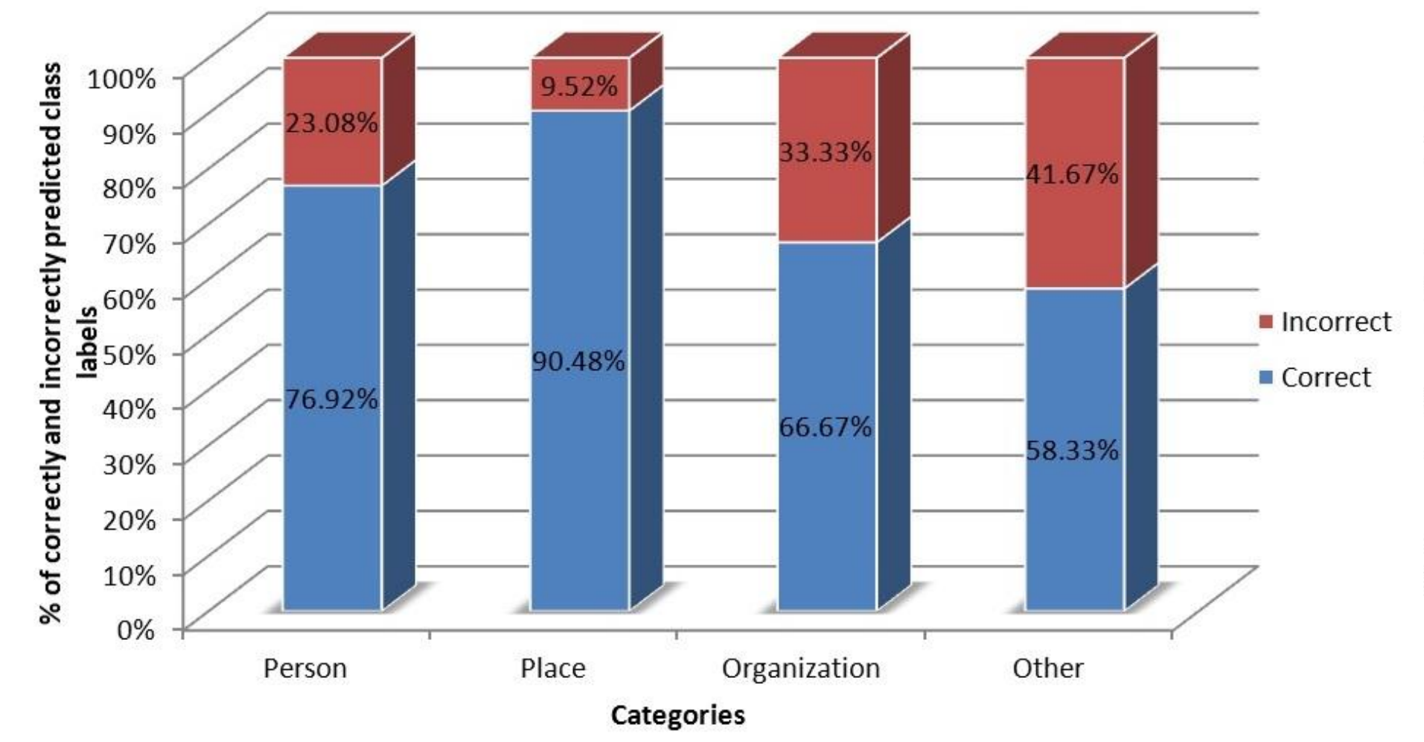
City		State
Baltimore	Relation 'A'	MD
Seattle	Relation 'A','B'	WA
Boston	Relation 'A', 'C'	MA
Raleigh	Relation 'A'	NC

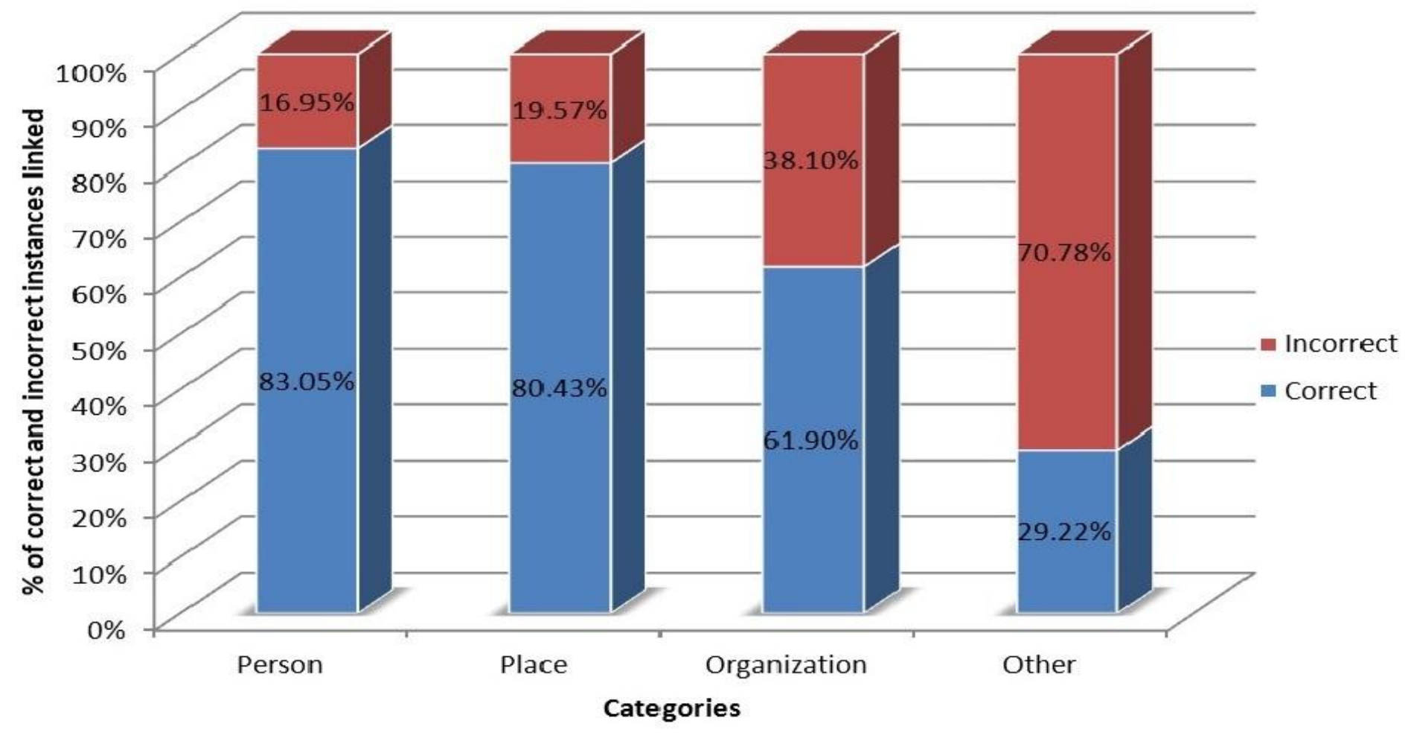
- ✓ For every pair of linked strings in the two column, query the knowledge base (KB)
- ✓ Generate a set of possible relations
- ✓ Rank the relations and choose the best relation



R11
R12
R13
R21
R22
R23
R31
R32
R33
C1
C2
C3
Row value
Function that captures the interaction between the column headers and row values
Column header (e.g. Mayor)

A probabilistic graphical model for joint inference and assignment of values





- ➢ *Column Correctness – 76.92 %*
- ➢ *Entity Linking – 66.12 %*

Column – Nationality ✗ Prediction – MilitaryConflict

Column – Birth Place ✓ Prediction – PopulatedPlace

Research Spending
($68 billion)

- ■ Defense ($12.3 billion)
- 🟩 Health/Bio ($33.47 billion)
- 🟦 Science and Tech ($21.45 billion)
- Other ($0.85 billion)

- ➢ The idea behind ***Evidence-based Medicine*** is to judge the efficacy of treatments or tests by meta-analyses or reviews of clinical trials. Key information in such trials is encoded in tables.

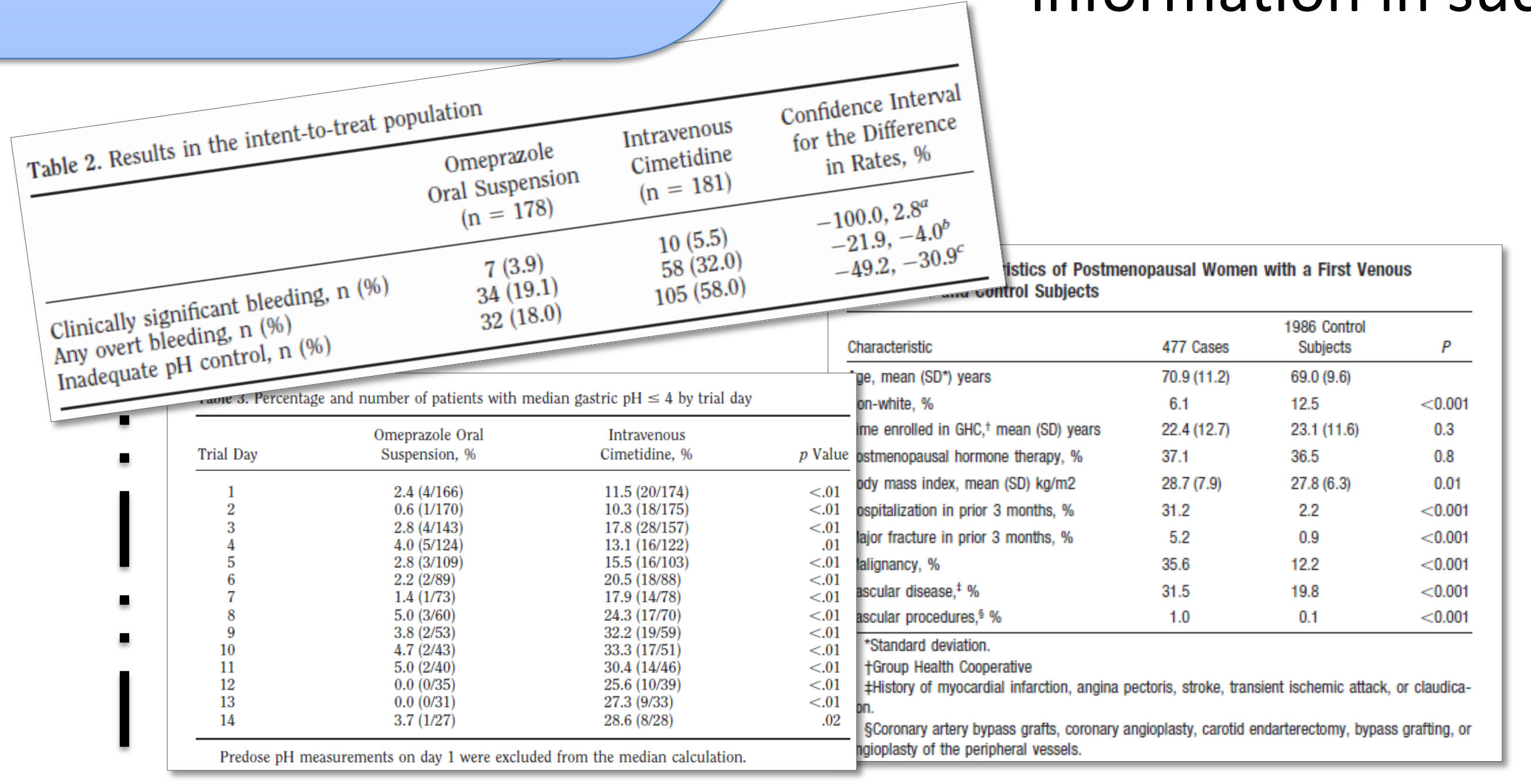


Table 2. Results in the intent-to-treat population

	Omeprazole Oral Suspension (n = 178)	Intravenous Cimetidine (n = 181)	Confidence Interval for the Difference in Rates, %
Clinically significant bleeding, n (%)	7 (3.9)	10 (5.5)	−100.0, 2.8[a]
Any overt bleeding, n (%)	34 (19.1)	58 (32.0)	−21.9, −4.0[b]
Inadequate pH control, n (%)	32 (18.0)	105 (58.0)	−49.2, −30.9[c]

Table 3. Percentage and number of patients with median gastric pH ≤ 4 by trial day

Trial Day	Omeprazole Oral Suspension, %	Intravenous Cimetidine, %	*p* Value
1	2.4 (4/166)	11.5 (20/174)	<.01
2	0.6 (1/170)	10.3 (18/175)	<.01
3	2.8 (4/143)	17.8 (28/157)	<.01
4	4.0 (5/124)	13.1 (16/122)	.01
5	2.8 (3/109)	15.5 (16/103)	<.01
6	2.2 (2/89)	20.5 (18/88)	<.01
7	1.4 (1/73)	17.9 (14/78)	<.01
8	5.0 (3/60)	24.3 (17/70)	<.01
9	3.8 (2/53)	32.2 (19/59)	<.01
10	4.7 (2/43)	33.3 (17/51)	<.01
11	5.0 (2/40)	30.4 (14/46)	<.01
12	0.0 (0/35)	25.6 (10/39)	<.01
13	0.0 (0/31)	27.3 (9/33)	<.01
14	3.7 (1/27)	28.6 (8/28)	.02

Predose pH measurements on day 1 were excluded from the median calculation.

istics of Postmenopausal Women with a First Venous
Control Subjects

Characteristic	477 Cases	1986 Control Subjects	*P*
ge, mean (SD*) years	70.9 (11.2)	69.0 (9.6)	
on-white, %	6.1	12.5	<0.001
ime enrolled in GHC,† mean (SD) years	22.4 (12.7)	23.1 (11.6)	0.3
ostmenopausal hormone therapy, %	37.1	36.5	0.8
ody mass index, mean (SD) kg/m2	28.7 (7.9)	27.8 (6.3)	0.01
ospitalization in prior 3 months, %	31.2	2.2	<0.001
ajor fracture in prior 3 months, %	5.2	0.9	<0.001
alignancy, %	35.6	12.2	<0.001
ascular disease,‡ %	31.5	19.8	<0.001
ascular procedures,§ %	1.0	0.1	<0.001

*Standard deviation.
†Group Health Cooperative
‡History of myocardial infarction, angina pectoris, stroke, transient ischemic attack, or claudication.
§Coronary artery bypass grafts, coronary angioplasty, carotid endarterectomy, bypass grafting, or angioplasty of the peripheral vessels.

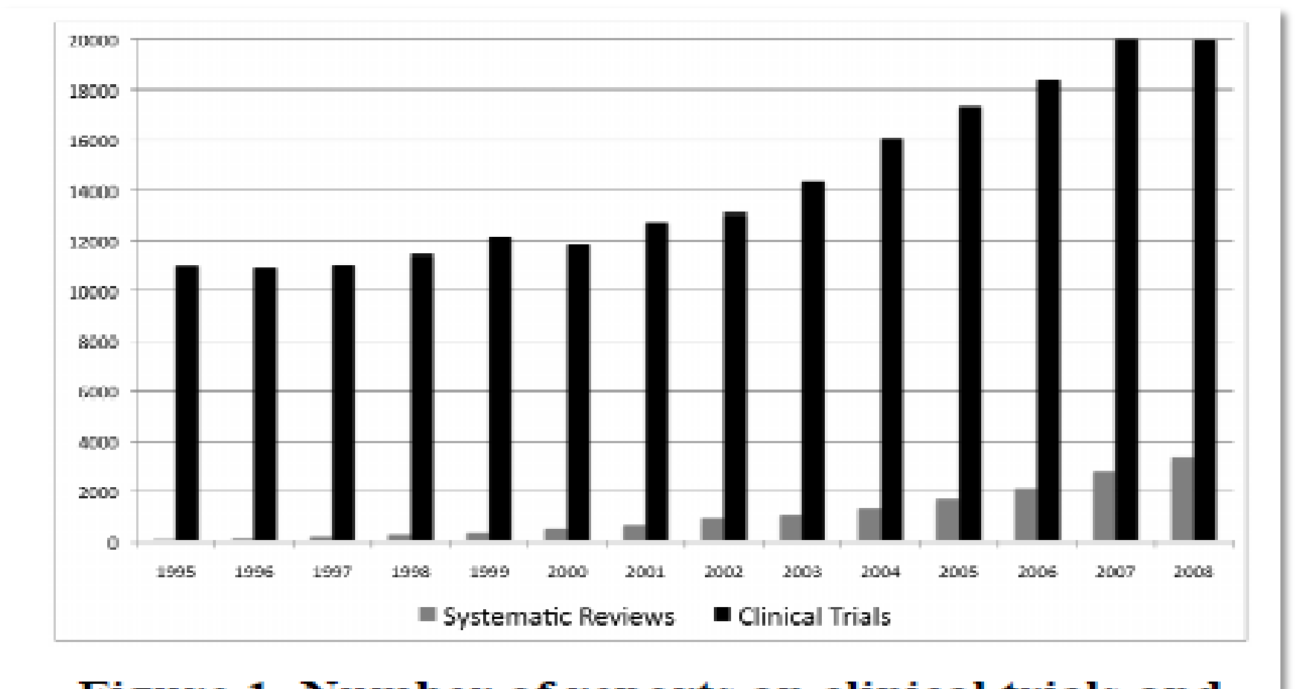


Figure 1. Number of reports on clinical trials and systematic reviews indexed in MEDLINE by year.

However, the rate at which meta-analyses are published remains very low ... hampers effective health care treatment ...

Under the guidance of

UMBC ebiquity