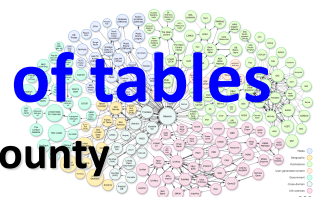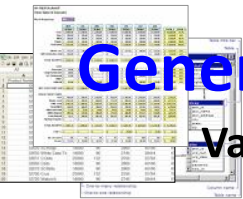# Generating Linked Data by inferring the semantics of tables

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

~ 50 years

"Noun" http://en.wikipedia.org/wiki/Barack_Obama

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.

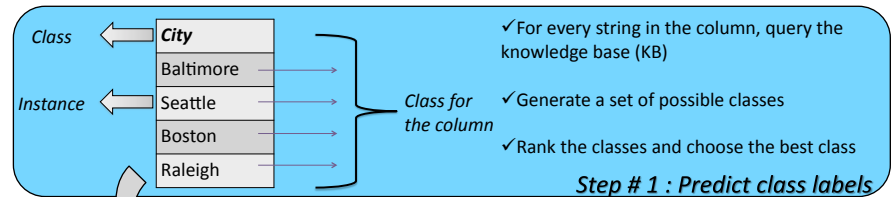http://en.wikipedia.org/wiki/United_States

"Verb"

☐ Text analysis techniques do not work well with tables
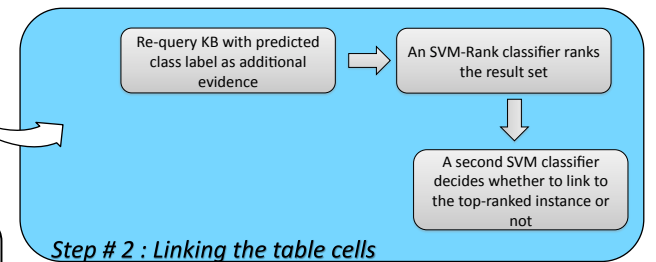
☐ Structure may be captured, but what about semantics ?
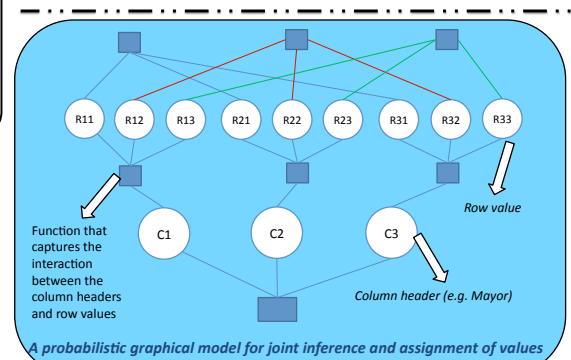
Tables are everywhere !

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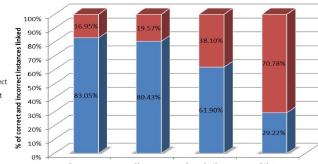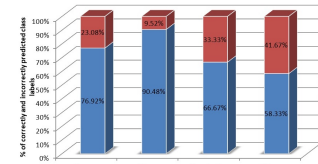☐ Can we build systems to exploit this knowledge ?

Step # 3 : Relation Identification

Relation 'A'

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

Class → City
Instance →

City
Baltimore
Seattle
Boston
Raleigh

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 .
```

R11 R12 R13 R21 R22 R23 R31 R32 R33

Row value

Function that captures the interaction between the column headers and row values

C1 C2 C3

Column header (e.g. Mayor)

A probabilistic graphical model for joint inference and assignment of values

➢ 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.
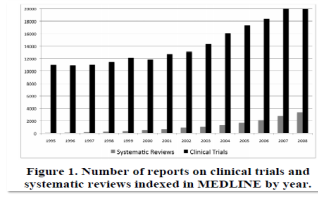
➢ Column Correctness – 76.92 %

Column – Nationality ✗	Column – Birth Place ✓
Prediction – MilitaryConflict	Prediction – PopulatedPlace

➢ Entity Linking – 66.12 %

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

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

Under the guidance of

UMBC ebiquity