LOD 123: Making the semantic web easier to use

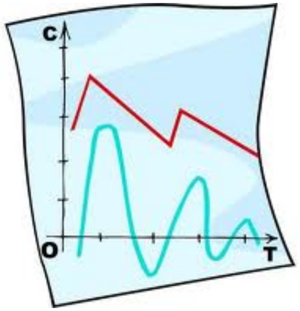
Tim Finin

University of Maryland, Baltimore County

Joint work with Lushan Han, Varish Mulwad, Anupam Joshi

UMBC ebiquity

Overview



- Linked Open Data 101
- Two ongoing UMBC dissertations
  - Varish Mulwad, Generating linked data from tables
  - Lushan Han, Querying linked data with a quasi-NL interface

Linked Open Data (LOD)

- Linked **data** is just RDF data, typically just the instances (ABOX), not schema (TBOX)
- RDF data is a graph of triples
  - URI URI string
    dbr:Barack_Obama dbo:spouse “Michelle Obama”
  - URI URI URI
    dbr:Barack_Obama dbo:spouse dbpedia:Michelle_Obama
- Best **linked** data practice prefers the 2nd pattern, using nodes rather than strings for “entities”
- Liked **open** data is just linked data freely accessible on the Web along with any required ontologies

Semantic Web

Use Semantic Web Technology
to publish shared data &
knowledge

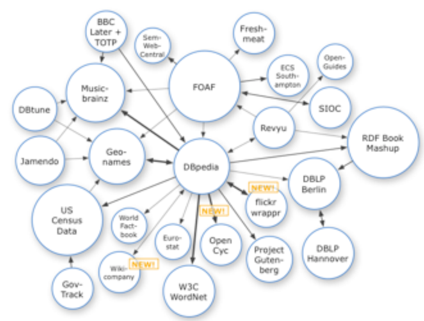
Semantic web technologies
allow machines to share data
and knowledge using common
web language and protocols.

~ 1997

Semantic Web beginning

Semantic Web => Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge



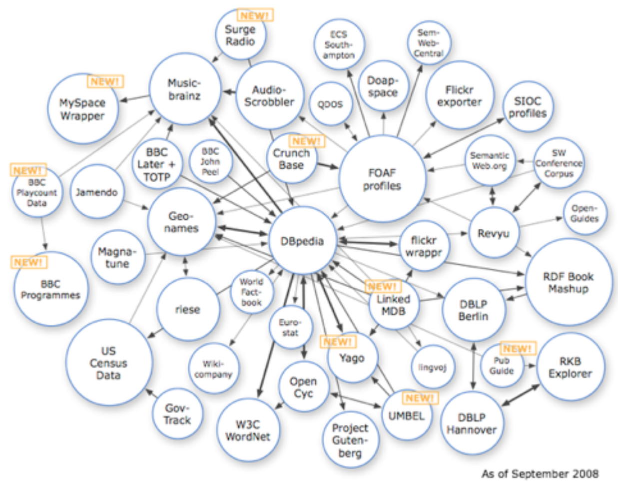
2007

Data is inter-
linked to support inte-
gration and fusion of knowledge

LOD beginning

Semantic Web => Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge



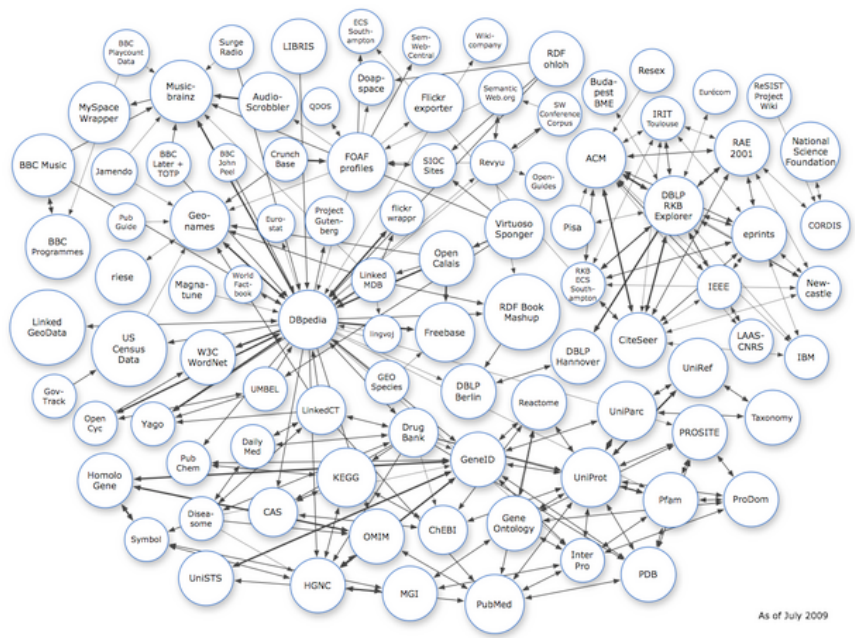
2008

Data is inter-
linked to support inte-
gration and fusion of knowledge

LOD growing

Semantic Web => Linked Open Data

Use Semantic Web Technology
to publish shared data &
knowledge



2009

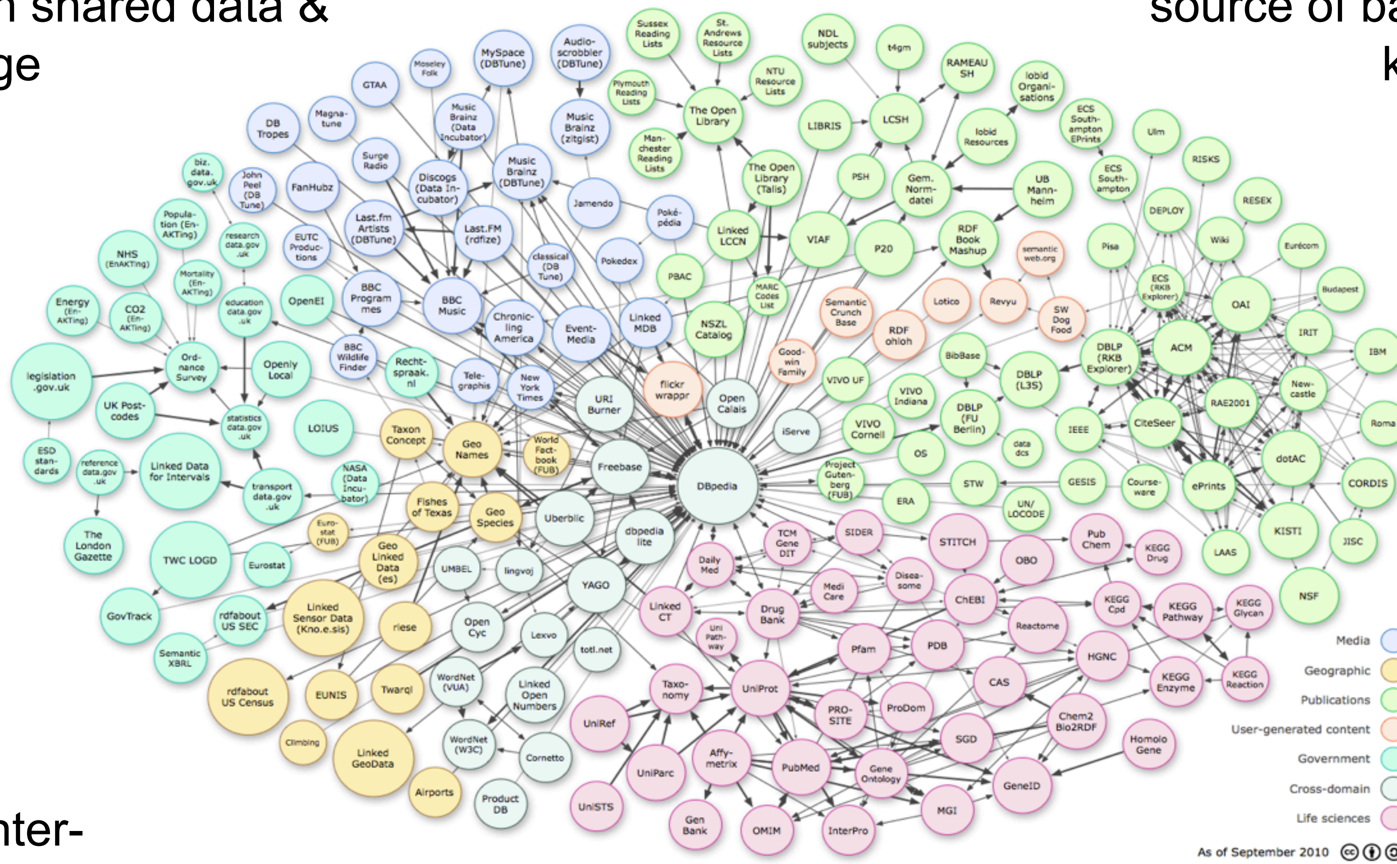
Data is inter-
linked to support inte-
gration and fusion of knowledge

... and growing

Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

LOD is the new Cyc: a common source of background knowledge



2010

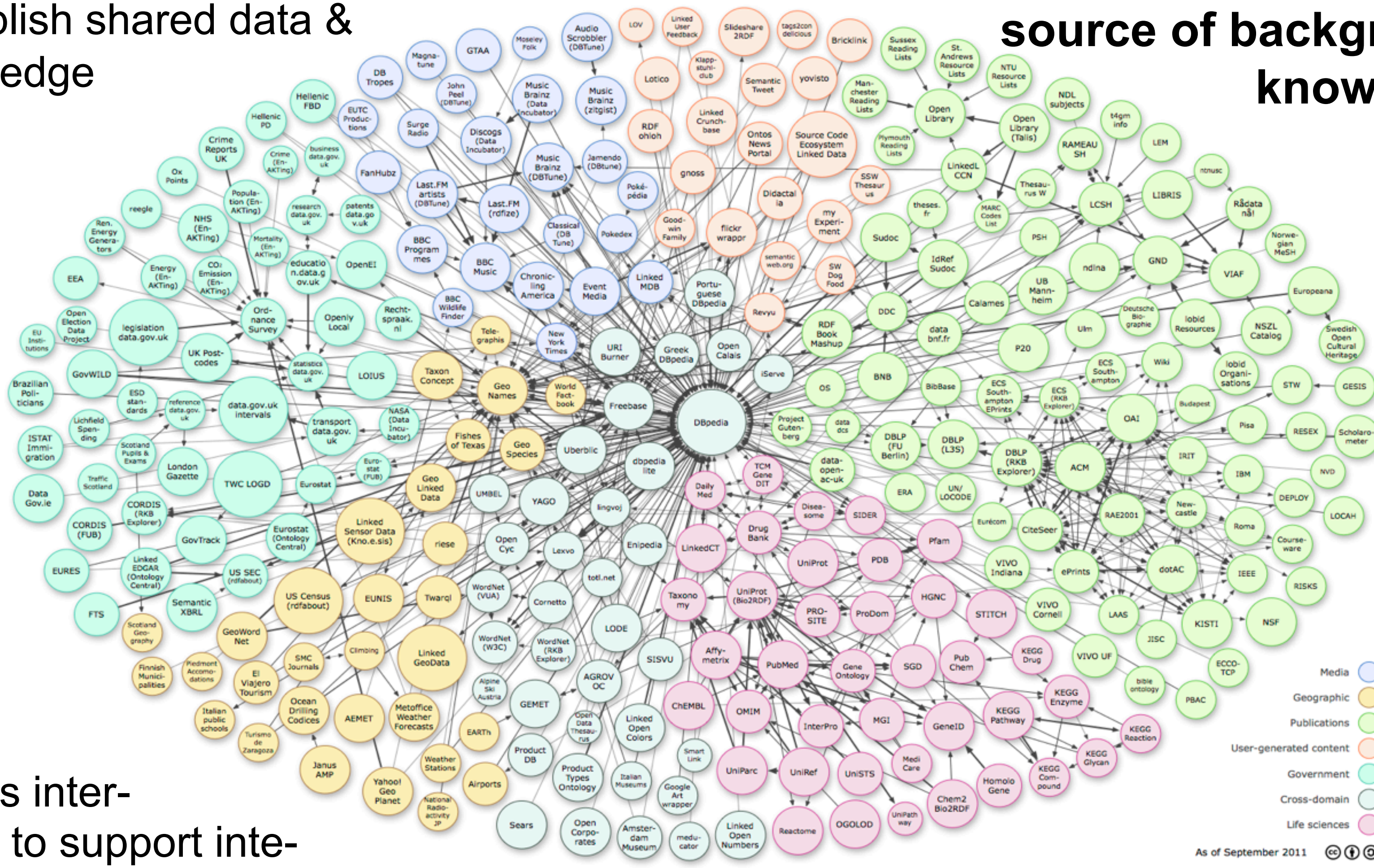
Data is inter-linked to support inte-gration and fusion of knowledge

...growing faster

Linked Open Data

Use Semantic Web Technology to publish shared data & knowledge

LOD is the new Cyc: a common source of background knowledge



Data is inter-linked to support inte-gration and fusion of knowledge

2011: 31B facts in 295 datasets interlinked by 504M assertions on ckan.net

Exploiting LOD not (yet) Easy

- Publishing or using LOD data has inherent difficulties for the potential user
  - It's difficult to explore LOD data and to query it for answers
  - It's challenging to publish data using appropriate LOD vocabularies & link it to existing data
- Problem: $O(10^4)$ schema terms, $O(10^{11})$ instances
- I'll describe two ongoing research projects that are addressing these problems

Generating Linked Data by Inferring the Semantics of Tables

Research with Varish Mulwad

http://ebiq.org/j/96

Early work

- Mapping tables to RDF led to early tools
  - D2RQ (2006) relational tables to RDF
  - RDF 123 (2007) spreadsheet to RDF
- And a recent W3C standard
  - R2RML (2012) a W3C recommendation
- But none of these can automatically generate high-quality linked data
  - They don't link to LOD classes and properties nor recognize entity mentions

Goal: Table => LOD*

dbprop:team

http://dbpedia.org/class/yago/NationalBasketballAssociationTeams

Name	Team	Position	Height
Michael Jordan	Chicago	Shooting guard	1.98
Allen Iverson	Philadelphia	Point guard	1.83
Yao Ming	Houston	Center	2.29
Tim Duncan	San Antonio	Power forward	2.11

http://dbpedia.org/resource/Allen_Iverson

Player height in meters

* DBpedia

Goal: Table => LOD*

Name	Team	Position	Height
Michael Jordan	Chicago	Shooting guard	1.98
Allen Iverson	Philadelphia	Point guard	1.83
Yao Ming	Houston		
Tim Duncan	San Ant		

RDF Linked Data

@prefix dbpedia: <http://dbpedia.org/resource/> .
@prefix dbo: <http://dbpedia.org/ontology/> .
@prefix yago: <http://dbpedia.org/class/yago/> .

"Name"@en is rdfs:label of dbo:BasketballPlayer .
"Team"@en is rdfs:label of yago:NationalBasketballAssociationTeams .

"Michael Jordan"@en is rdfs:label of dbpedia:Michael Jordan .
dbpedia:Michael Jordan a dbo:BasketballPlayer .

"Chicago Bulls"@en is rdfs:label of dbpedia:Chicago Bulls .
dbpedia:Chicago Bulls a yago:NationalBasketballAssociationTeams .

All this in a completely automated way

* DBpedia

Tables are everywhere !! … yet …

The web – ***154 million*** high quality relational tables

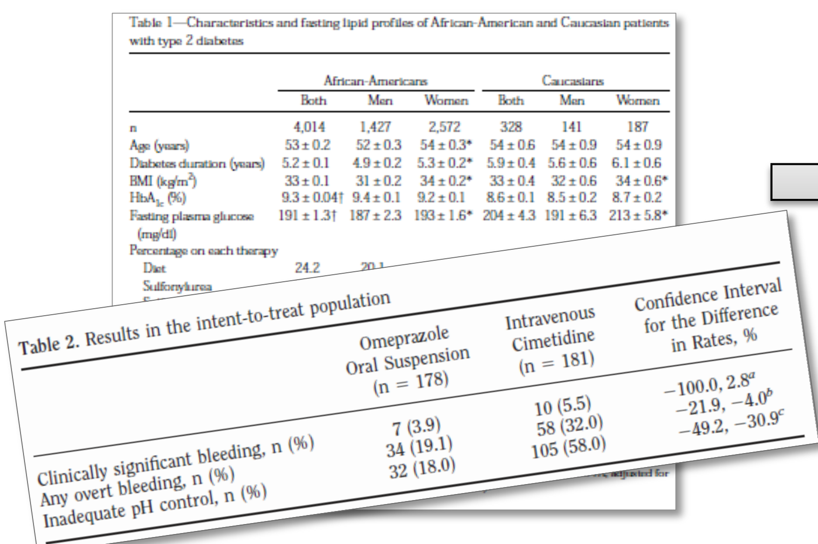
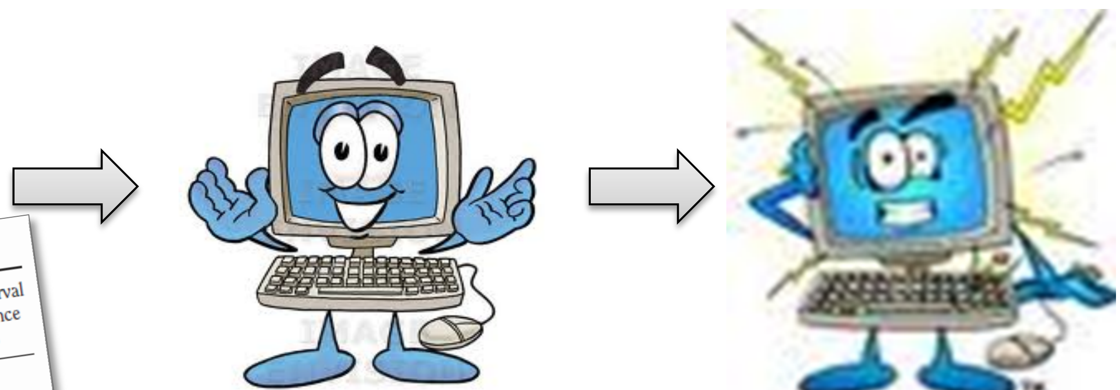


Table 1—Characteristics and fasting lipid profiles of African-American and Caucasian patients with type 2 diabetes

	African-Americans			Caucasians		
	Both	Men	Women	Both	Men	Women
n	4,014	1,427	2,572	328	141	187
Age (years)	53 ± 0.2	52 ± 0.3	54 ± 0.3*	54 ± 0.6	54 ± 0.9	54 ± 0.9
Diabetes duration (years)	5.2 ± 0.1	4.9 ± 0.2	5.3 ± 0.2*	5.9 ± 0.4	5.6 ± 0.6	6.1 ± 0.6
BMI (kg/m^2)	33 ± 0.1	31 ± 0.2	34 ± 0.2*	33 ± 0.4	32 ± 0.6	34 ± 0.6*
HbA_{1c} (%)	9.3 ± 0.04†	9.4 ± 0.1	9.2 ± 0.1	8.6 ± 0.1	8.5 ± 0.2	8.7 ± 0.2
Fasting plasma glucose (mg/dl)	191 ± 1.3†	187 ± 2.3	193 ± 1.6*	204 ± 4.3	191 ± 6.3	213 ± 5.8*
Percentage on each therapy						
Diet	24.2	20.1				
Sulfonylurea						

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]

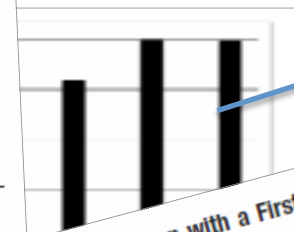


Evidence–based medicine

Evidence-based medicine judges the efficacy of treatments or tests by meta-analyses of clinical trials. Key information is often found in tables in articles

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)	



# of Clinical trials published in 2008

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.

TABLE 1. Characteristics of Postmenopausal Women with a First Venous Thrombosis and Control Subjects

Characteristic	477 Cases	1986 Control Subjects	P
Age, mean (SD*) years	70.9 (11.2)	69.0 (9.6)	<0.001
Non-white, %	6.1	12.5	0.3
Time enrolled in GHC,† mean (SD) years	22.4 (12.7)	23.1 (11.6)	0.8
Postmenopausal hormone therapy, %	37.1	36.5	0.01
Body mass index, mean (SD) kg/m2	28.7 (7.9)	27.8 (6.3)	<0.001
Hospitalization in prior 3 months, %	31.2	2.2	<0.001
Major fracture in prior 3 months, %	5.2	0.9	<0.001
Malignancy, %	35.6	12.2	<0.001
Vascular disease,‡ %	31.5	19.8	<0.001
Vascular procedures,§ %	1.0	0.1	

*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.

# of meta analysis published in 2008

Figure: Evidence-Based Medicine - the Essential Role of Systematic Reviews ... Need for Automated Text Mining Tools, IHI 2010

DATA.GOV

EMPOWERING PEOPLE

~ *400,000* datasets

~ *< 1* % in RDF

2010 Preliminary System

T2LD Framework

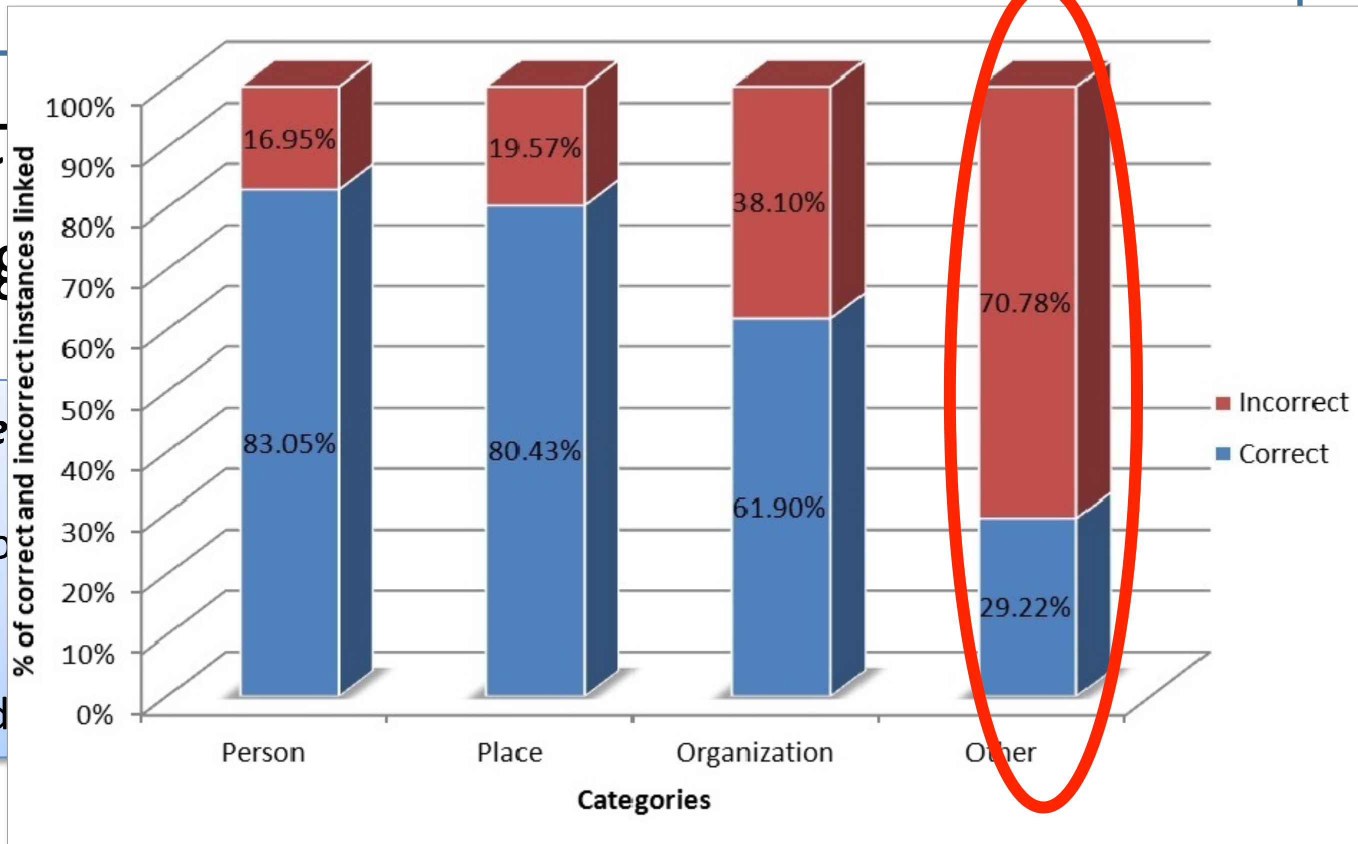
Predict Class for Columns → Linking the table cells → Identify and Discover relations

Class predict

Entity Linking

Examples of class labe
Column – Nationality
Prediction – MilitaryCo

Column – Birth Place
Prediction – Populated

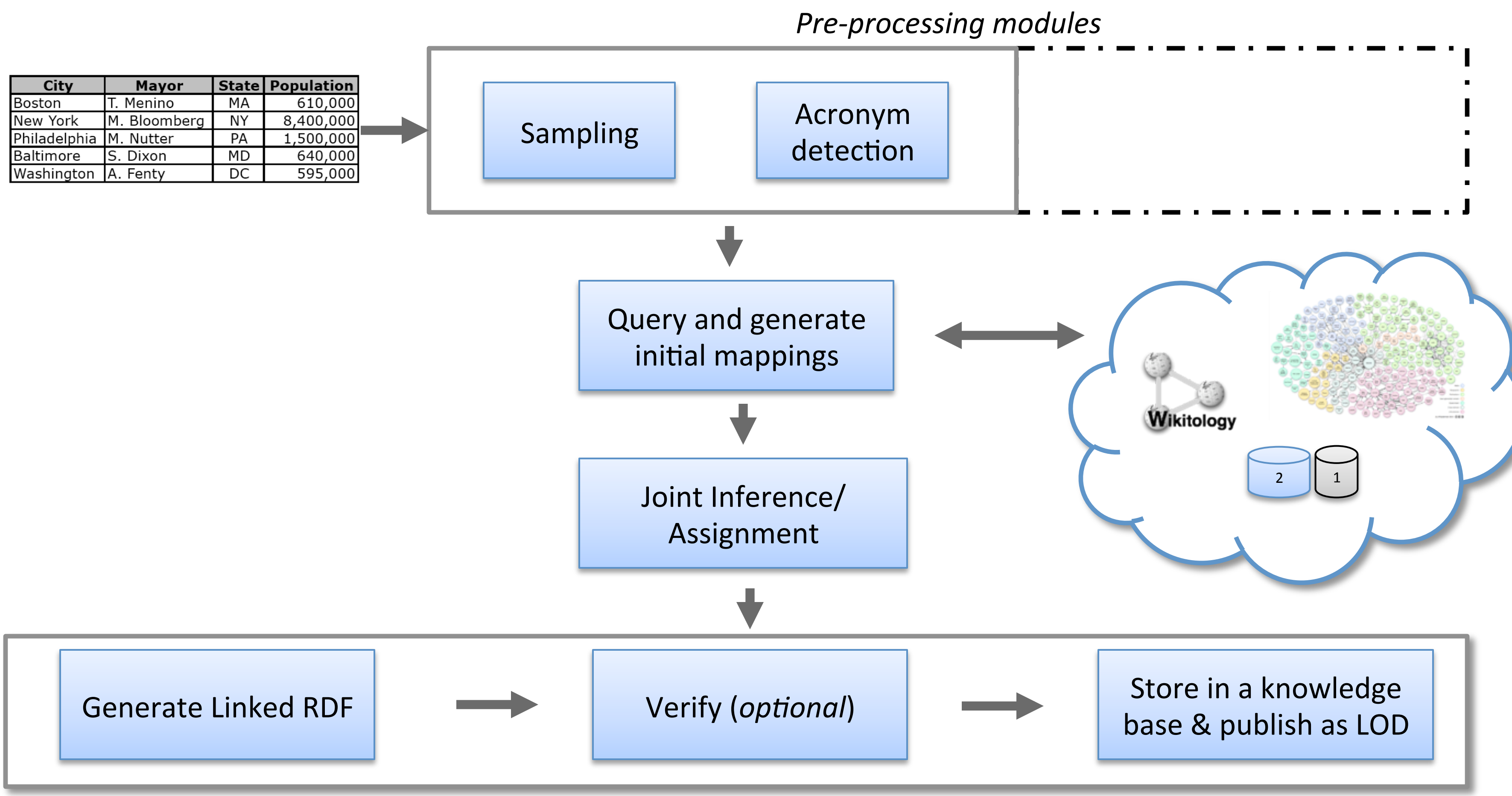


Sources of Errors

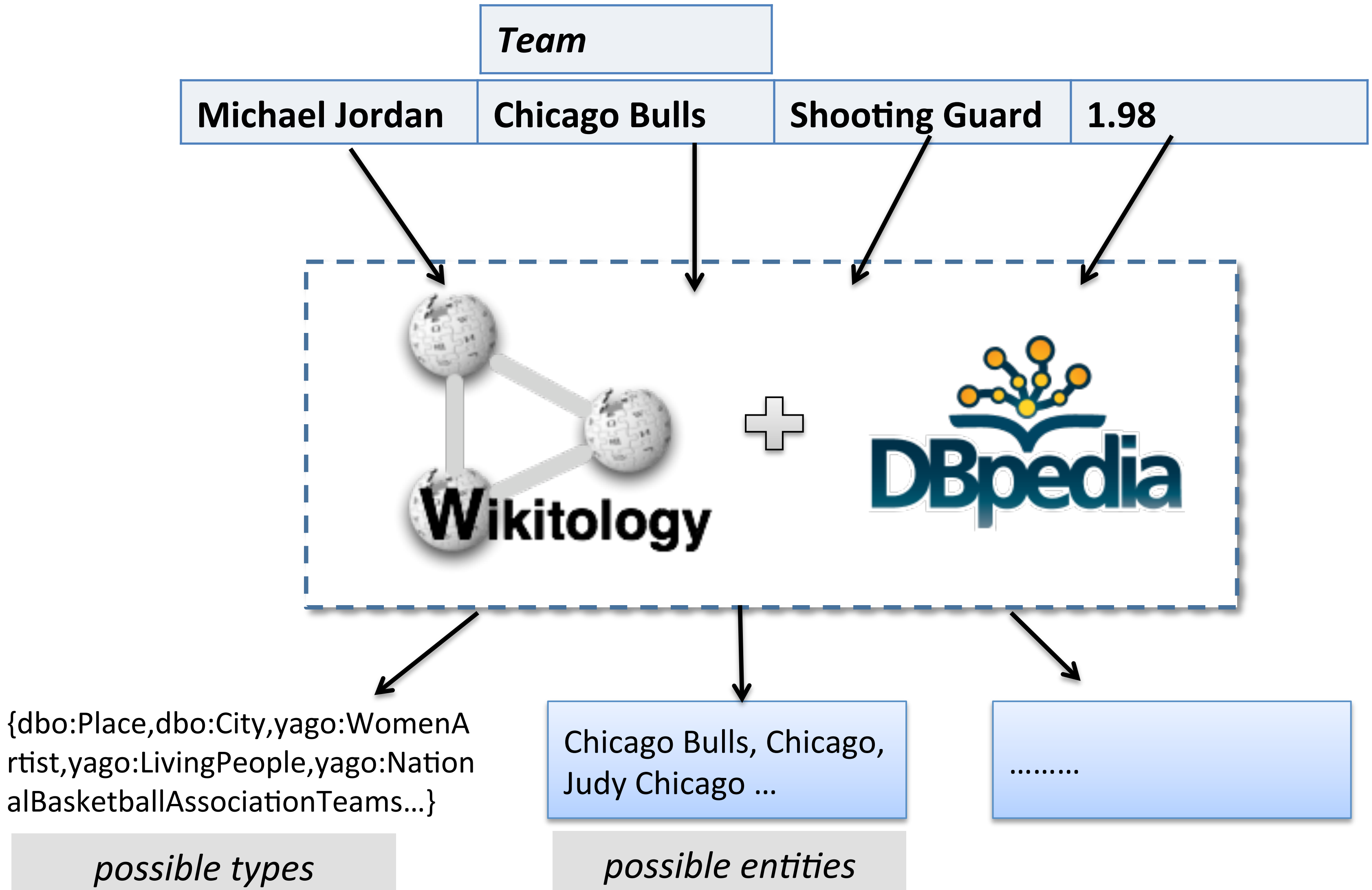
- The *sequential* approach let errors percolate from one phase to the next
- The system was biased toward predicting overly general classes over more appropriate specific ones
- Heuristics largely drive the system
- Although we consider multiple sources of evidence, we did not joint assignment

A Domain Independent Framework

City	Mayor	State	Population
Boston	T. Menino	MA	610,000
New York	M. Bloomberg	NY	8,400,000
Philadelphia	M. Nutter	PA	1,500,000
Baltimore	S. Dixon	MD	640,000
Washington	A. Fenty	DC	595,000



Query Mechanism



Ranking the candidates

- C_i = "State" ; L_{ci} = AdministrativeRegion

String in column header

Class from an ontology

- f_1 = [Levenshtein distance(C_i , L_{Ci}),
  Dice Score (C_i , L_{Ci}),
  Semantic Similarity (C_i , L_{Ci}),
  InformationGain(L_{Ci})]

String similarity metrics

- $\psi_1 = \exp(w_1^T f_1(C_i, L_{Ci}))$

Ranking the candidates

- R_{ij} = "Baltimore" ; E_{ij} = Baltimore_Maryland

String in table cell

Entity from the knowledge base (KB)

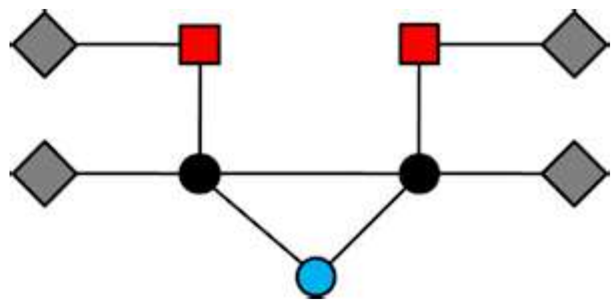
- f_2 = [Levenshtein distance(R_{ij} , E_{ij}), Dice Score (R_{ij} , E_{ij}), PageRank (E_{ij}), KBScore (E_{ij}) PageLength (E_{ij})]

String similarity metrics

Popularity metrics

$$\psi_2 = \exp(w_2^T f_2(R_{ij}, E_{ij}))$$

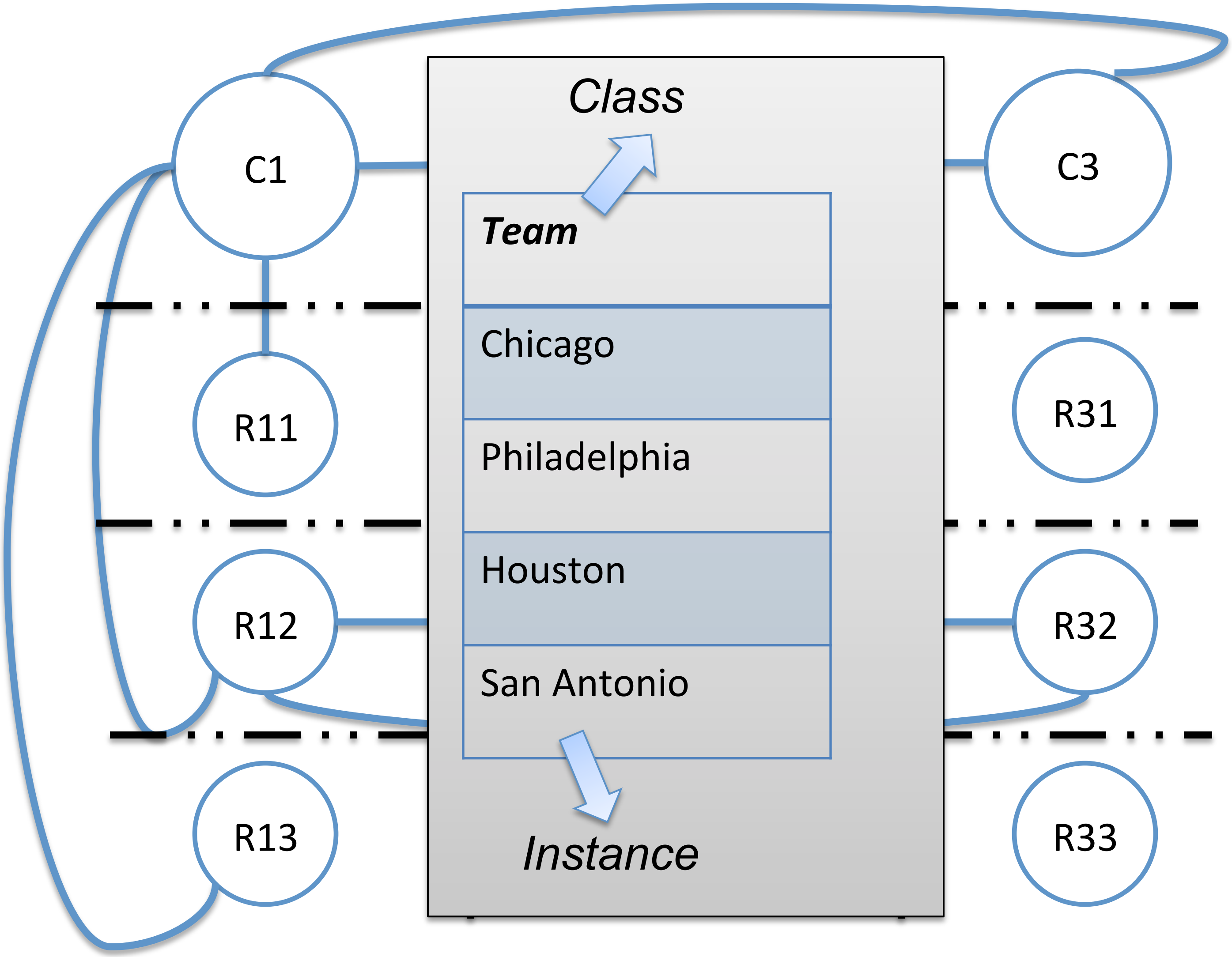
Joint Inference over evidence in a table



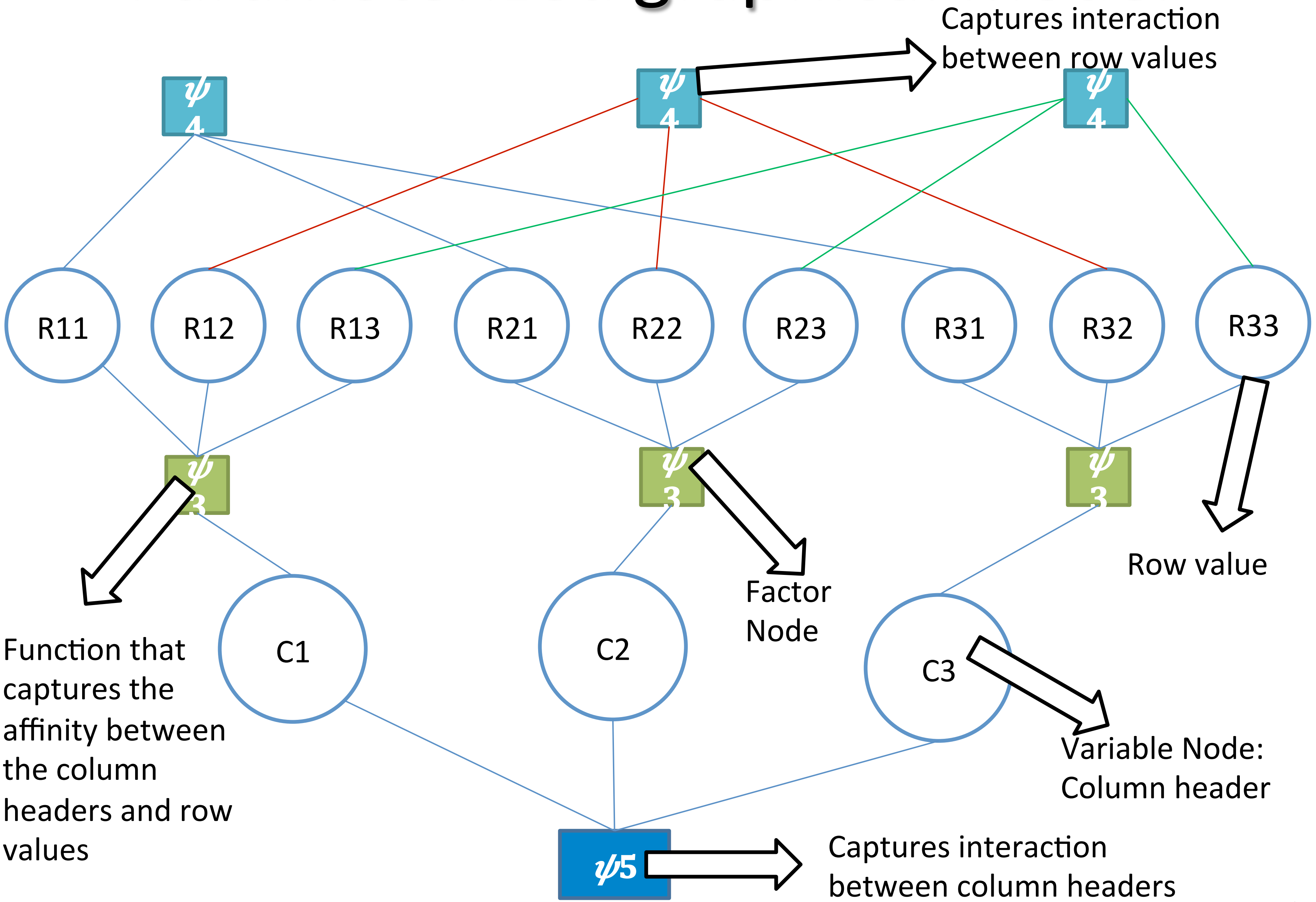
✓Probabilistic Graphical Models

A graphical model for tables

Joint inference over evidence in a table



Parameterized graphical model



Challenge: Interpreting Literals

Many columns have literals, e.g., numbers

Population?
Profit in $K ?

Population
690,000
345,000
510,020
120,000

Age
75
65
50
25

Age in years?
Percent?

- Predict properties based on cell values
- Cyc had hand coded rules: *humans don't live past 120*
- We extract *value distributions* from LOD resources
  - Differ for subclasses: age of *people* vs. *political leaders* vs. *athletes*
  - Represent as *measurements*: value + units
- Metric: possibility/probability of values given distribution

Other Challenges

- Using table *captions* and other text is associated documents to provide context
- Size of some data.gov tables (> 400K rows!) makes using full graphical model impractical
  - Sample table and run model on the subset
- Achieving acceptable accuracy may require human input
  - 100% accuracy unattainable automatically
  - How best to let humans offer advice and/or correct interpretations?

PMI as an association measure

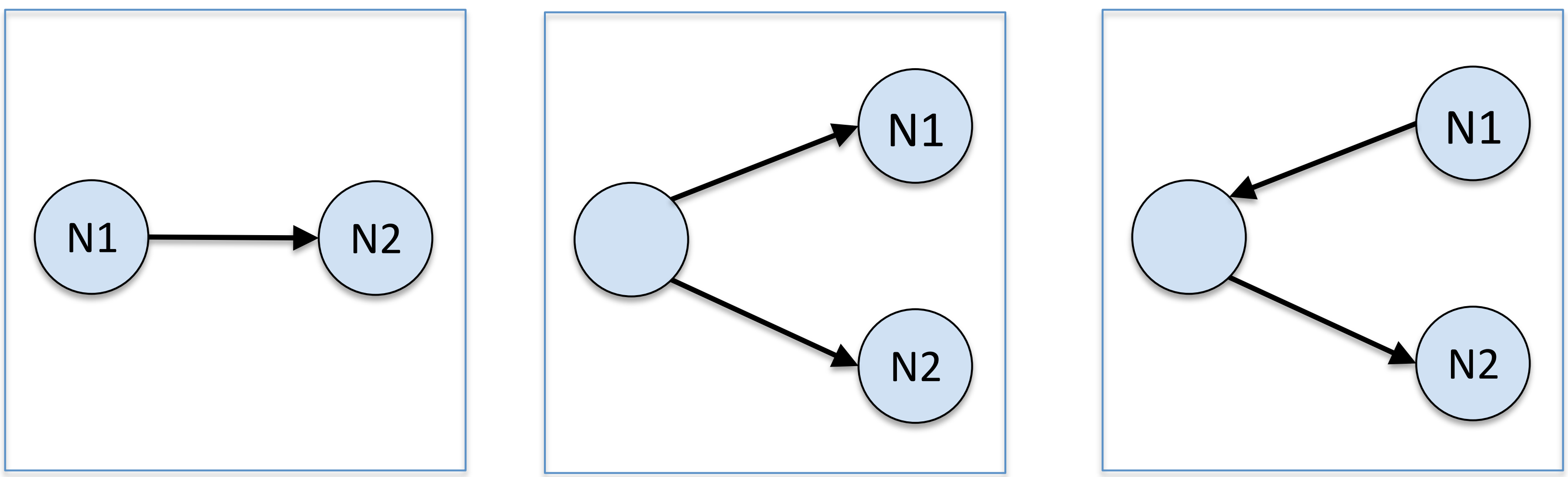
We use pointwise mutual information (pmi) to measure the association between two RDF resources (nodes)

$$\operatorname{pmi}(x;y) \equiv \log\frac{p(x,y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)} = \log\frac{p(y|x)}{p(y)}.$$

pmi is used for word association by comparing how often two words occur together in text to their expected co-occurrence if independent

PMI for RDF instances

- For text, the co-occurrence context is usually a window of some number of words (e.g, 50)
- For RDF instances, we count three graph patterns as instances of the co-occurrence of N1 and N2



- Other graph patterns can be added, but we've not evaluated their utility or cost to compute.

PMI for RDF types

- We also want to measure the association strength between *RDF types*, e.g., a dbo:Actor associated with a dbo:Film vs. a dbo:Place
- We can also measure the association of an RDF property and types, e.g. dbo:author used with a dbo:Film vs. a dbo:Book
- Such simple statistics can be efficiently computed for large RDF collections in parallel

PREFIX dbo: <http://dbpedia.org/ontology/>

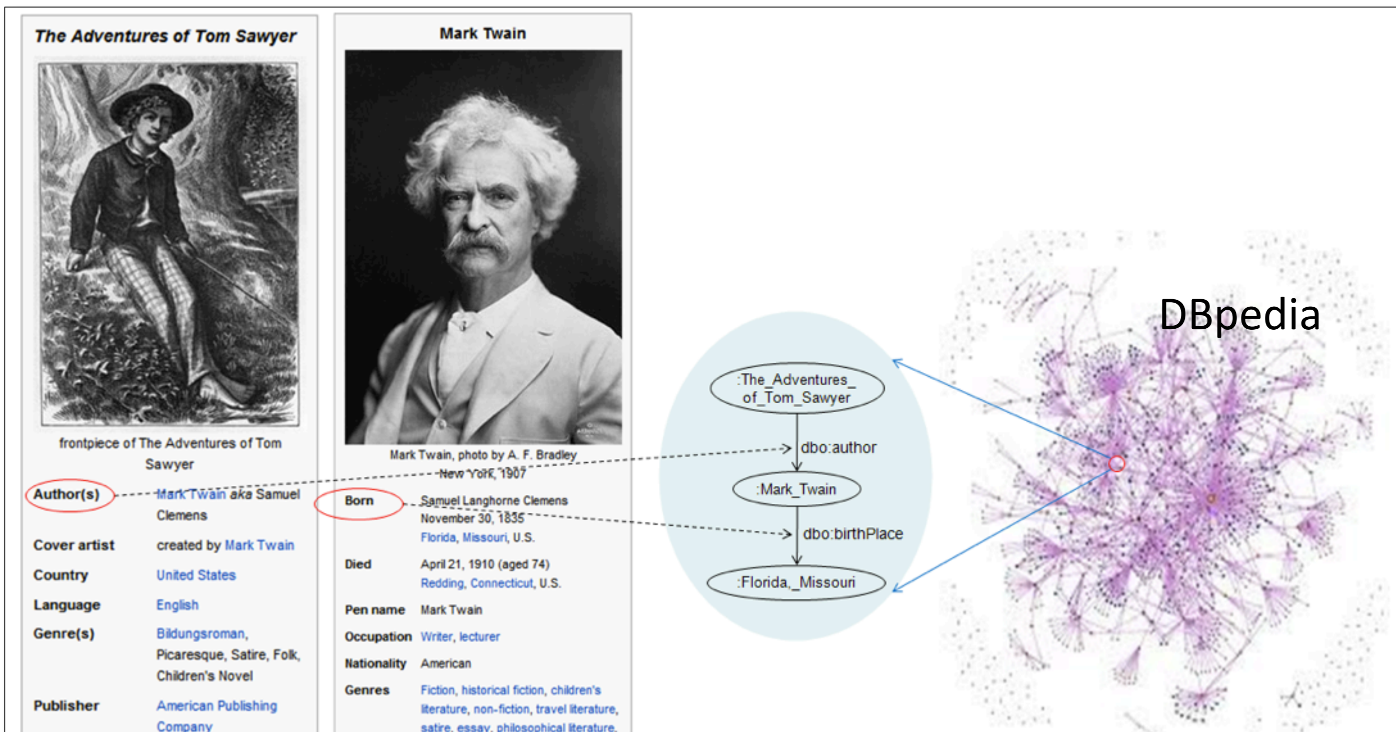
GoRelations:
Intuitive Query System for Linked Data

Research with Lushan Han

http://ebiq.org/j/93

Dbpedia is the Stereotypical LOD

- DBpedia is an important example of Linked Open Data
  - Extracts structured data from Infoboxes in Wikipedia
  - Stores in RDF using custom ontologies Yago terms
- The major integration point for the entire LOD cloud
- Explorable as HTML, but harder to query in SPARQL



Browsing DBpedia's Mark Twain

About: Mark Twain

dbpedia.org/page/Mark_Twain

G news OGK U R T S Learn util me fun ! py 2do nyt bitly Read Later l8r » Other Bookmarks

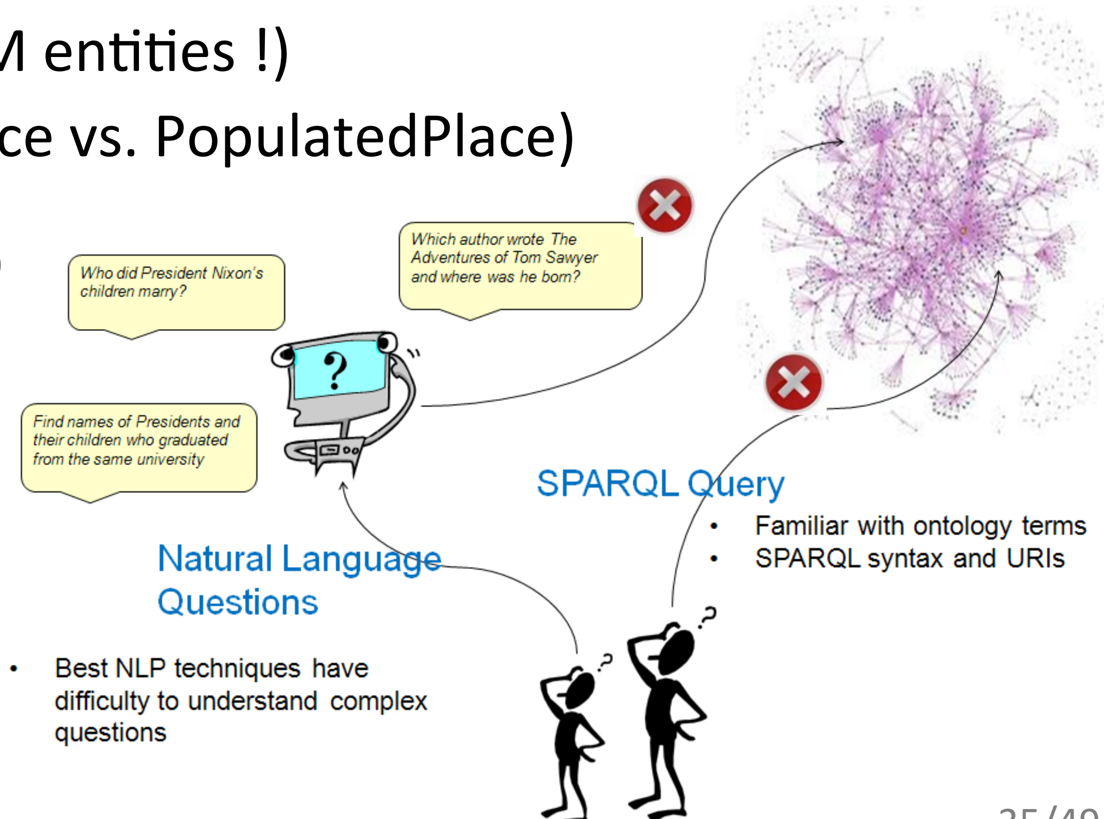
dbpedia-owl:birthDate	1835-11-30 (xsd:date)
dbpedia-owl:birthName	Samuel Langhorne Clemens
dbpedia-owl:birthPlace	dbpedia:Florida,_Missouri dbpedia:Missouri
dbpedia-owl:child	dbpedia:Jean_Clemens dbpedia:Susy_Clemens dbpedia:Clara_Clemens
dbpedia-owl:deathDate	1910-04-21 (xsd:date)
dbpedia-owl:deathPlace	dbpedia:Redding,_Connecticut dbpedia:Connecticut
dbpedia-owl:genre	dbpedia:Essay dbpedia:Historical_fiction dbpedia:Social_commentary dbpedia:Satire dbpedia:Non-fiction dbpedia:Philosophy_and_literature dbpedia:Literary_criticism dbpedia:Travel_literature dbpedia:Fiction dbpedia:Children's_literature
dbpedia-owl:notableWork	dbpedia:Adventures_of_Huckleberry_Finn dbpedia:The_Adventures_of_Tom_Sawyer
dbpedia-owl:occupation	dbpedia:Writer dbpedia:Lecture
dbpedia-owl:pseudonym	Mark Twain
dbpedia-owl:thumbnail	http://upload.wikimedia.org/wikipedia/commons/thumb/c/c5/Mark_Twain_by_AF_Bra
dbpedia-owl:wikiPageExternalLink	http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2005/08/19/MNGOBEA9JI1.DTL http://www.pbs.org/marktwain/scrapbook/index.html http://www.kamakurapens.com/TwainsConklin.html http://www.ucmerced.edu/faculty/facultybio.asp?facultyid=95

Querying LOD is Much Harder

- Querying DBpedia requires a lot of a user
  - Understand the RDF model
  - Master SPARQL, a formal query language
  - Understand ontology terms: 320 classes & 1600 properties !
  - Know instance URIs (>1M entities !)
  - Term heterogeneity (Place vs. PopulatedPlace)

- Querying large LOD sets overwhelming

- Natural language query systems still a research goal



Goal

- Allow a user with a basic understanding of RDF to query DBpedia and ultimately distributed LOD collections
  - To explore what data is in the system
  - To get answers to question
  - To create SPARQL queries for reuse or adaptation
- Desiderata
  - Easy to learn and to use
  - Good accuracy (e.g., precision and recall)
  - Fast

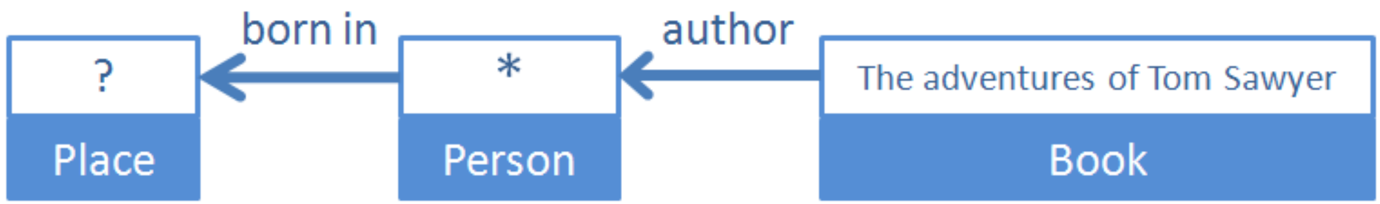
Key Idea

Structured keyword queries

Reduce problem complexity by:

- User enters a *simple graph*, and
- Annotates the nodes and arcs with *words and phrases*

Structured Keyword Queries

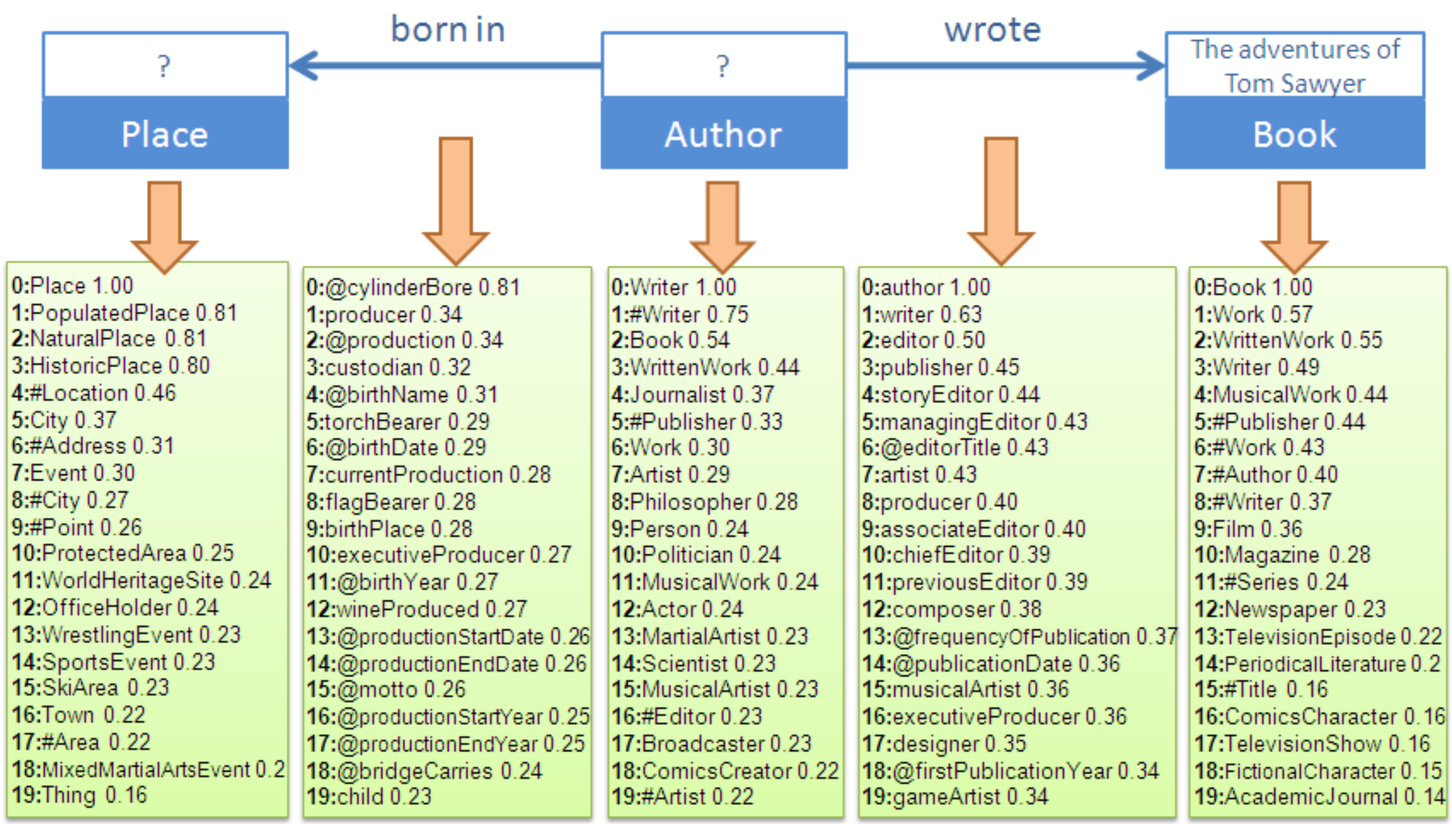


- Nodes denote entities and links binary relations
- Entities described by two unrestricted terms: *name* or value and *type* or concept
- Result entities marked with *?* and those not with *
- A compromise between a natural language Q&A system and SPARQL
  - Users provide compositional structure of the question
  - Free to use their own terms in annotating the structure

Translation – Step One

finding semantically similar ontology terms

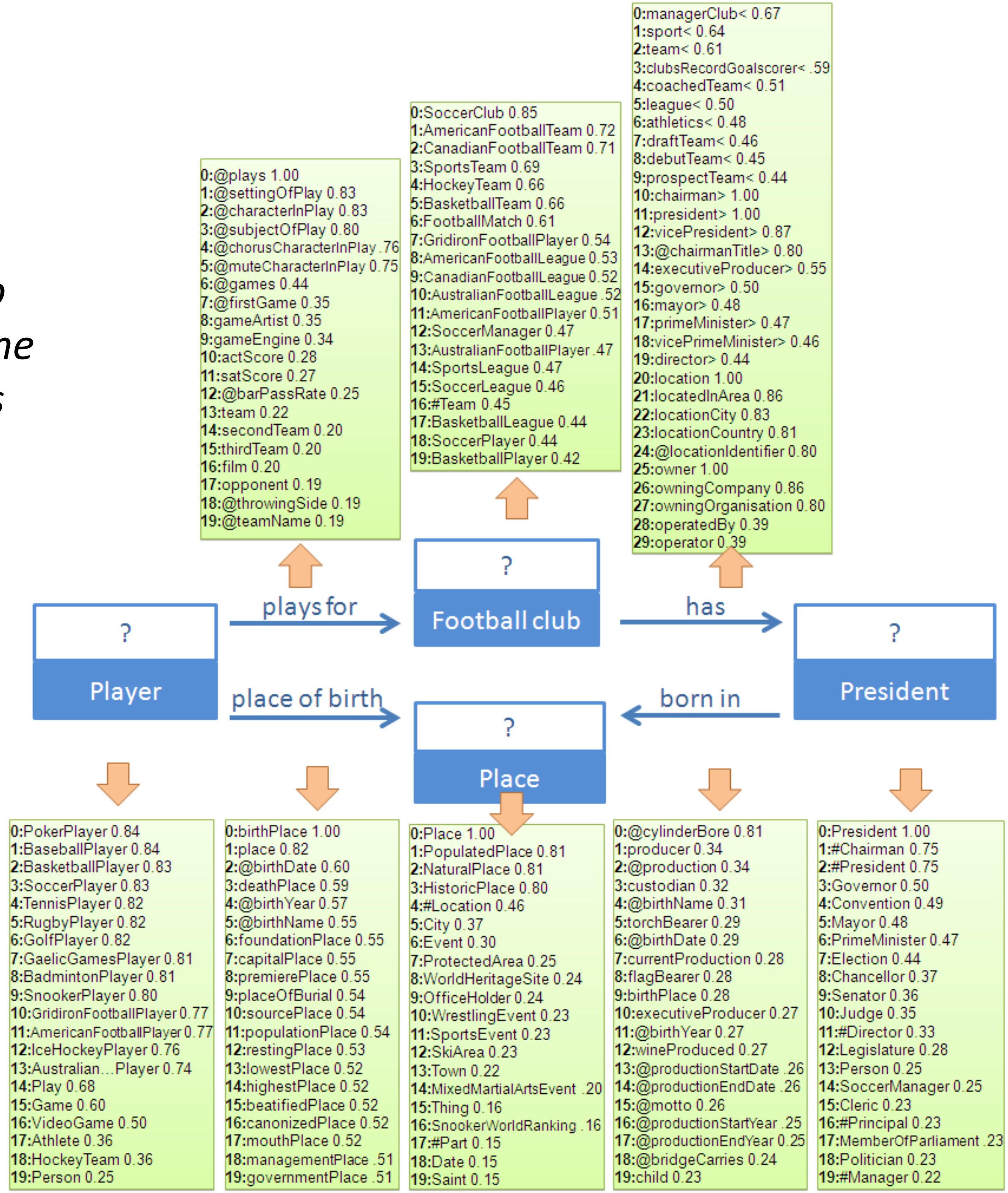
For each concept or relation in the graph, generate the *k* most semantically similar candidate ontology classes or properties



Similarity metric is distributional similarity, LSA, and WordNet.

Another Example

Football players who were born in the same place as their team's president



Translation – Step Two

disambiguation algorithm

- To assemble the best interpretation we rely on *statistics of the data*
- Primary measure is *pointwise mutual information* (PMI) between RDF terms in the LOD collection

  This measures the degree to which two RDF terms occur together in the knowledge base

- In a reasonable interpretation, *ontology terms associate* in the way that their corresponding *user terms* connect in the structured keyword query

Translation – Step Two

disambiguation algorithm

Three aspects are combined to derive an *overall goodness measure* for each candidate interpretation

Joint disambiguation

$$\operatorname*{argmax}_{p_1..p_m\, c_1..c_n \in H} \text{goodness}(G) = \operatorname*{argmax}_{p_1..p_m\, c_1..c_n \in H} \sum_{i=1}^{m} \text{goodness}(L_i) \quad (1)$$

Resolving direction

$$\begin{aligned} &\text{If } [\overrightarrow{\text{PMI}}(c(O_i), p(R_i)) + \overrightarrow{\text{PMI}}(p(R_i), c(S_i))] \\ &\qquad - [\overrightarrow{\text{PMI}}(c(S_i), p(R_i)) + \overrightarrow{\text{PMI}}(p(R_i), c(O_i))] > \alpha \\ &\text{Then } S_i' = O_i,\ O_i' = S_i \\ &\text{Else } S_i' = S_i,\ O_i' = O_i \end{aligned} \quad (2)$$

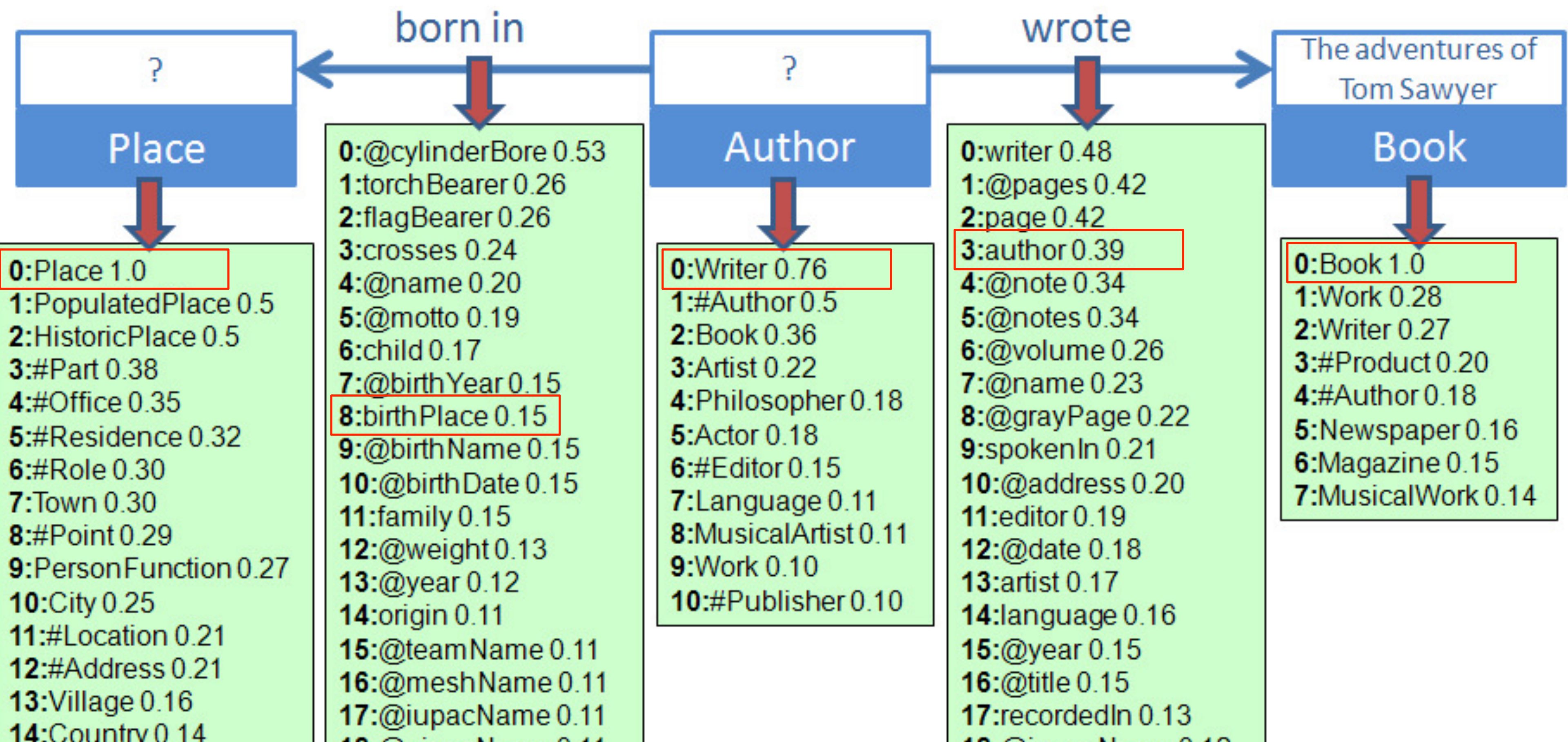
Link reasonableness

$$\begin{aligned} \text{goodness}(L_i) = \max(&\overrightarrow{\text{PMI}}(c(S_i'), p(R_i)) \cdot \text{sim}(S_i', c(S_i')) \cdot \text{sim}(R_i, p(R_i)) \\ &+ \overrightarrow{\text{PMI}}(p(R_i), c(O_i')) \cdot \text{sim}(O_i', c(O_i')) \cdot \text{sim}(R_i, p(R_i)), \beta) \\ &+ \text{PMI}(c(S_i'), c(O_i')) \cdot \text{sim}(S_i', c(S_i')) \cdot \text{sim}(O_i', c(O_i')) \end{aligned} \quad (3)$$

Example of Translation result

Concepts: Place => Place, Author => Writer, Book => Book

Properties: born in => birthPlace, wrote => author (inverse direction)



SPARQL Generation

The translation of a semantic graph query to SPARQL is straightforward given the mappings

Concepts
- Place => Place
- Author => Writer
- Book => Book

Relations
- born in => birthPlace
- wrote => author

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x, ?y WHERE {
 ?0 a dbo:Book .
 ?0 rdfs:label ?label0 .
 ?label0 bif:contains '"The adventures of Tom Sawyer"' .
 ?x a dbo:Writer .
 ?y a dbo:Place .
 {?0 dbo:author ?x} .
 {?x dbo:birthPlace ?y} .
}
```

Evaluation

- 33 test questions from 2011 *Workshop on Question Answering over Linked Data* answerable using DBpedia
- Three human subjects unfamiliar with DBpedia translated the test questions into semantic graph queries
- Compared with two top natural language QA systems: PowerAqua and True Knowledge

		Prec.	*Recall*	*F*
GoRelations	regular	0.687	0.722	0.704
	concise	0.736	0.803	0.768
PowerAqua	1st triple	0.372	0.483	0.420
	all triples	0.334	0.483	0.395
	merged	0.255	0.291	0.272
True Knowledge		0.469	0.535	0.500

GoRelations: an Intuitive Que

semanticwebarchive.cs.umbc.edu/GOR/example7.html

Manual Experiments Q&A About

Graph of Relations *alpha ver 0.1*

Please input relations in your query. One per line.

```
?x/American Football Player, date of birth, ?y/Date
?x, height, ?z/Number
```

Examples: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Query | Relaxed Query | See Translations

Message: This query gives data about the height of American football players and their date of birth. Football funs may feel it interesting to know how the height of football players changes over time.

http://ebiq.org/GOR

Powered by DBpedia and Virtuoso
2011 Ebiquity Lab, UMBC.

lod.openlinksw.com/sparql?default-graph-uri=http%3A%2F%2Fdbpedia.org&query=PREFIX+dbo%3A+<http%...

x	z	y
http://dbpedia.org/resource/Jack_Shapiro	1.5494	1907-03-22
http://dbpedia.org/resource/Josh_Betts	1.5748	1982-08-25
http://dbpedia.org/resource/Nate_Abrams	1.6256	1897-12-25
http://dbpedia.org/resource/Herbert_Clow	1.6256	1899-05-07
http://dbpedia.org/resource/Trindon_Holliday	1.651	1986-04-27
http://dbpedia.org/resource/Earl_Warweg	1.6764	1892-01-11
http://dbpedia.org/resource/Jacquizz_Rodgers	1.6764	1990-02-06
http://dbpedia.org/resource/John_Brallier	1.6764	1876-12-27
http://dbpedia.org/resource/Menz_Lindsey	1.6764	1897-07-25
http://dbpedia.org/resource/Stefan_Logan	1.6764	1981-06-02
http://dbpedia.org/resource/Cory_Ross	1.6764	1982-09-22
http://dbpedia.org/resource/Darren_Sproles	1.6764	1983-06-20
http://dbpedia.org/resource/Jack_Fleischman	1.6764	1901-08-15
http://dbpedia.org/resource/Eddie_Scharer	1.6764	1902-01-26
http://dbpedia.org/resource/John_Barrett_%28American_football%29	1.6764	1899-02-25
http://dbpedia.org/resource/Dick_Thornton_%28American_football%29	1.6764	1908-02-04
http://dbpedia.org/resource/Walter_French_%28baseball%29	1.7018	1899-07-12
http://dbpedia.org/resource/Al_Loeb	1.7018	1890-03-11
http://dbpedia.org/resource/Antwaun_Carter	1.7018	1981-09-09
http://dbpedia.org/resource/Terry_Caulley	1.7018	1984-06-22
http://dbpedia.org/resource/Kendall_Hunter	1.7018	1988-09-16
http://dbpedia.org/resource/LaRod_Stephens-Howling	1.7018	1987-04-26
http://dbpedia.org/resource/Jayson_Foster	1.7018	1985-07-22
http://dbpedia.org/resource/Mark_McMillian	1.7018	1970-04-29
http://dbpedia.org/resource/Al_Cornsweet	1.7018	1906-07-16
http://dbpedia.org/resource/Bodie_Weldon	1.7018	1895-11-07
http://dbpedia.org/resource/Bill_Stobbs	1.7018	1896-05-28
http://dbpedia.org/resource/Johnny_Bryan	1.7018	1897-02-28
http://dbpedia.org/resource/Andrew_Hawkins	1.7018	1986-03-10
http://dbpedia.org/resource/Chad_Owens	1.7018	1982-04-03
http://dbpedia.org/resource/Mark_Higgs	1.7018	1966-04-11
http://dbpedia.org/resource/Emmett_McLemore	1.7018	[illegible]
http://dbpedia.org/resource/Danny_Woodhead	1.7018	1985-01-25
http://dbpedia.org/resource/Garrett_Wolfe	1.7018	1984-08-17
http://dbpedia.org/resource/Shaud_Williams	1.7018	1980-02-10

http://ebiq.org/GOR

D GoRelations -- Translations

semanticwebarchive.cs.umbc.edu/GOR/GoRelations

Three Top Interpretations of the User Query

Top One Interpretation

Before Step Three

* ?x/American Football Player has a candidate list including 0:AmericanFootballPlayer 1.00 1:Person 0.25 (the selected choice is AmericanFootballPlayer 1.00)
* ?z/Number has a candidate list including 0:Number 1.00 1:#Group 0.31 2:#Code 0.31 3:EthnicGroup 0.25 4:#Size 0.23 5:#Series 0.23 6:Single 0.21 7:OlympicResult 0.15 8:MountainRange 0.14 (the selected choice is Number 1.00)
* ?y/Date has a candidate list including 0:Date 1.00 1:Grape 0.19 2:Holiday 0.18 3:Name 0.18 4:Place 0.15 5:Event 0.15 6:#Era 0.15 7:GivenName 0.15 8:HistoricPlace 0.14 9:WrestlingEvent 0.14 10:PopulatedPlace 0.12 11:SportsEvent 0.12 12:WorldHeritageSite 0.12 13:NaturalPlace 0.12 14:MixedMartialArtsEvent 0.11 (the selected choice is Date 1.00)

The Regular SPARQL Query

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x, ?z, ?y WHERE {
  ?x a dbo:AmericanFootballPlayer .
  ?x dbo:birthDate ?y .
  ?x dbo:height ?z .
}
```

Query

The Concise SPARQL Query

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x, ?z, ?y WHERE {
  ?x a dbo:AmericanFootballPlayer .
  ?x dbo:birthDate ?y .
  ?x dbo:height ?z .
}
```

Query

After Step Three

* ?x/American Football Player => AmericanFootballPlayer
* ?z/Number => Number
* ?y/Date => Date
& ?x/American Football Player:date of birth:?y/Date => 0:@birthDate 1.00
& ?x/American Football Player:height:?z/Number => 0:@height 1.00

The Regular SPARQL Query

```
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT DISTINCT ?x, ?z, ?y WHERE {
  ?x a dbo:AmericanFootballPlayer .
  {{?x dbo:birthDate ?y}} .
  {{?x dbo:height ?z}} .
}
```

Query

The Concise SPA

```
PREFIX dbo: <http:/

SELECT DISTINCT ?x
  ?x a dbo:American
  {{?x dbo:birthDate
  {{?x dbo:height ?z}
}
```

Top Two Interpretation

Before Step Three

* ?x/American Football Player has a candidate list including 0:AmericanFootballPlay
0.25 (the selected choice is AmericanFootballPlayer 1.00)
* ?z/Number has a candidate list including 0:Number 1.00 1:#Group 0.31 2:#Code 0.31 3:EthnicGroup 0.25 4:#Size 0.23 5:#Series 0.23 6:Single 0.21 7:OlympicResult 0.15 8:MountainRange 0.14 (the

* ?z/Number => #Series
* ?y/Date => Date
& ?x/American Football Player:date of birth:?y/Date => 0:@birthDate 1.00

http://ebiq.org/GOR

Current challenges

- Baseline system works well for DBpedia
- Current challenges we are addressing are
  - Adding direct entity matching
  - Relaxing the need for type information
  - Testing on other LOD collections and extending to a set of distributed LOD collections
  - Developing a better Web interface
  - Allowing user feedback and advice
- See http://ebiq.org/93 for more information & try our alpha version at http://ebiq.org/GOR

Final Conclusions

- Linked Data is an emerging paradigm for sharing structured and semi-structured data
  - Backed by machine-understandable semantics
  - Based on successful Web languages and protocols
- Generating and exploring Linked Data resources can be challenging
  - Schemas are large, too many URIs
- New tools for mapping tables to Linked Data and translating structured natural language queries help reduce the barriers

http://ebiq.org/