Generating Linked Data by Inferring the Semantics of Tables

Varish Mulwad, Ph.D. 2015

http://ebiq.org/j/96

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

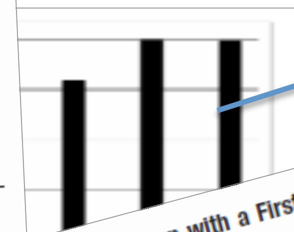
	African-Americans			Caucasians		
	Both	Men	Women	Both	Men	Women
n	4,014	1,427	2,572	328	141	187
Age (years)	53 ± 0.2	52 ± 0.3	54 ± 0.3*	54 ± 0.6	54 ± 0.9	54 ± 0.9
Diabetes duration (years)	5.2 ± 0.1	4.9 ± 0.2	5.3 ± 0.2*	5.9 ± 0.4	5.6 ± 0.6	6.1 ± 0.6
BMI (kg/m^2)	33 ± 0.1	31 ± 0.2	34 ± 0.2*	33 ± 0.4	32 ± 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



# of Clinical trials published in 2008

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)	[illegible]

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)	
Non-white, %	6.1	12.5	<0.001
Time enrolled in GHC,† mean (SD) years	22.4 (12.7)	23.1 (11.6)	0.3
Postmenopausal hormone therapy, %	37.1	36.5	0.8
Body mass index, mean (SD) kg/m2	28.7 (7.9)	27.8 (6.3)	0.01
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	<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.

# of meta analysis published in 2008

very low … hampers effective health …

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

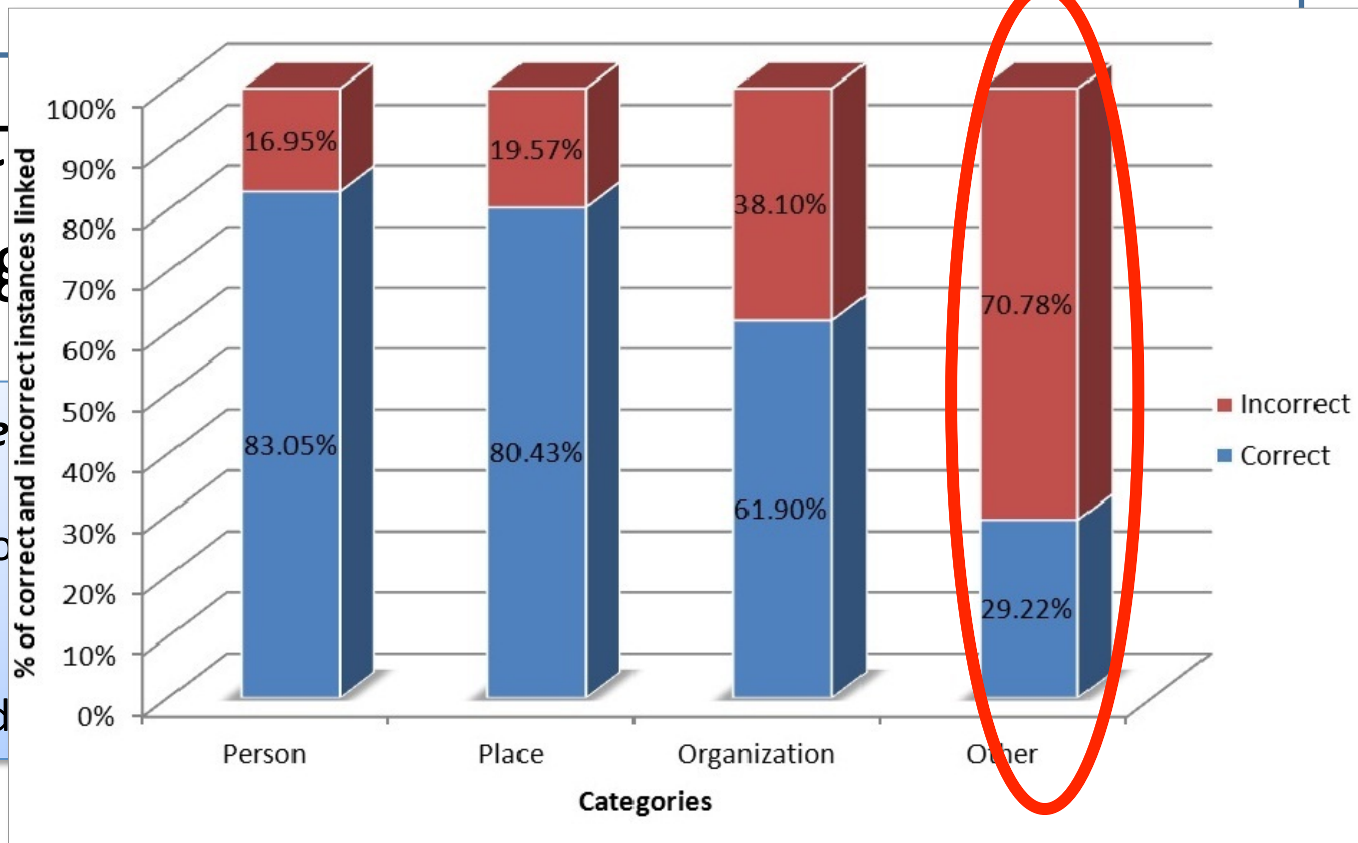
Predict Class for Columns → Linking the table cells → Identify and Discover relations

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

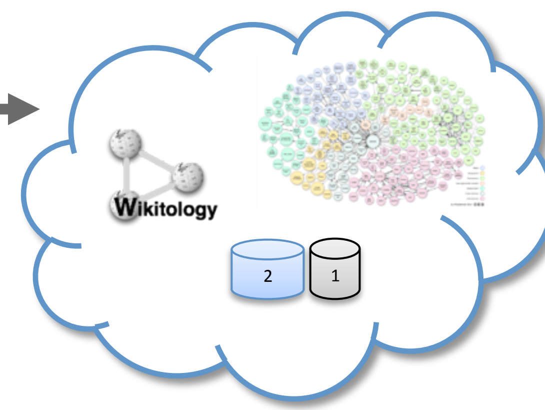
City	Mayor	State	Population
Boston	T. Menino	MA	610,000
New York	M. Bloomberg	NY	8,400,000
Philadelphia	M. Nutter	PA	1,500,000
Baltimore	S. Dixon	MD	640,000
Washington	A. Fenty	DC	595,000

Pre-processing modules

Sampling

Acronym detection

Query and generate initial mappings



Joint Inference/ Assignment

Generate Linked RDF

Verify (*optional*)

Store in a knowledge base & publish as LOD

Query Mechanism

	Team		
Michael Jordan	Chicago Bulls	Shooting Guard	1.98

Wikitology + DBpedia

{dbo:Place,dbo:City,yago:WomenArtist,yago:LivingPeople,yago:NationalBasketballAssociationTeams…}

possible types

Chicago Bulls, Chicago, Judy Chicago …

possible entities

………

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)

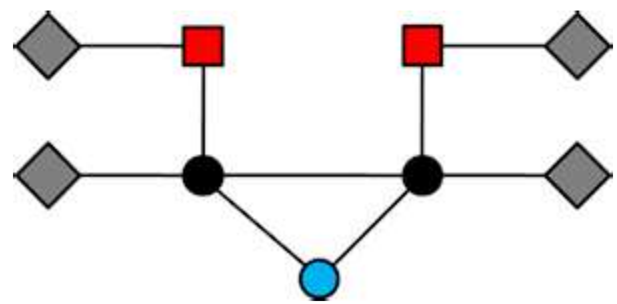
- f_2 = [Levenshtein distance(R_{ij} , E_{ij}),
  Dice Score (R_{ij} , E_{ij}),
  PageRank (E_{ij}),
  KBScore (E_{ij})
  PageLength (E_{ij})]

String similarity metrics

Popularity metrics

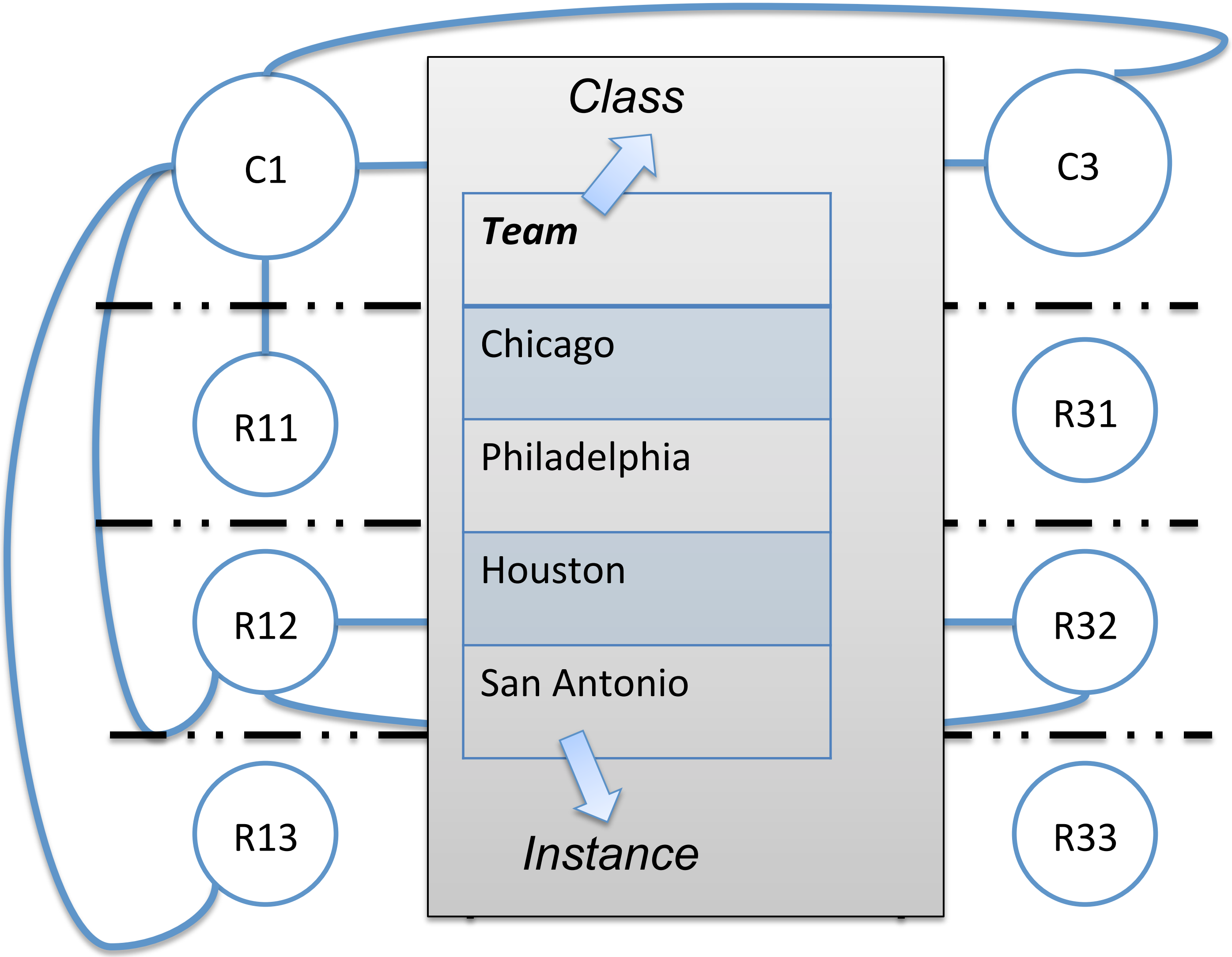
$$\psi_2 = \exp(w_2^T f_2(R_{ij}, E_{ij}))$$

Joint Inference over evidence in a table

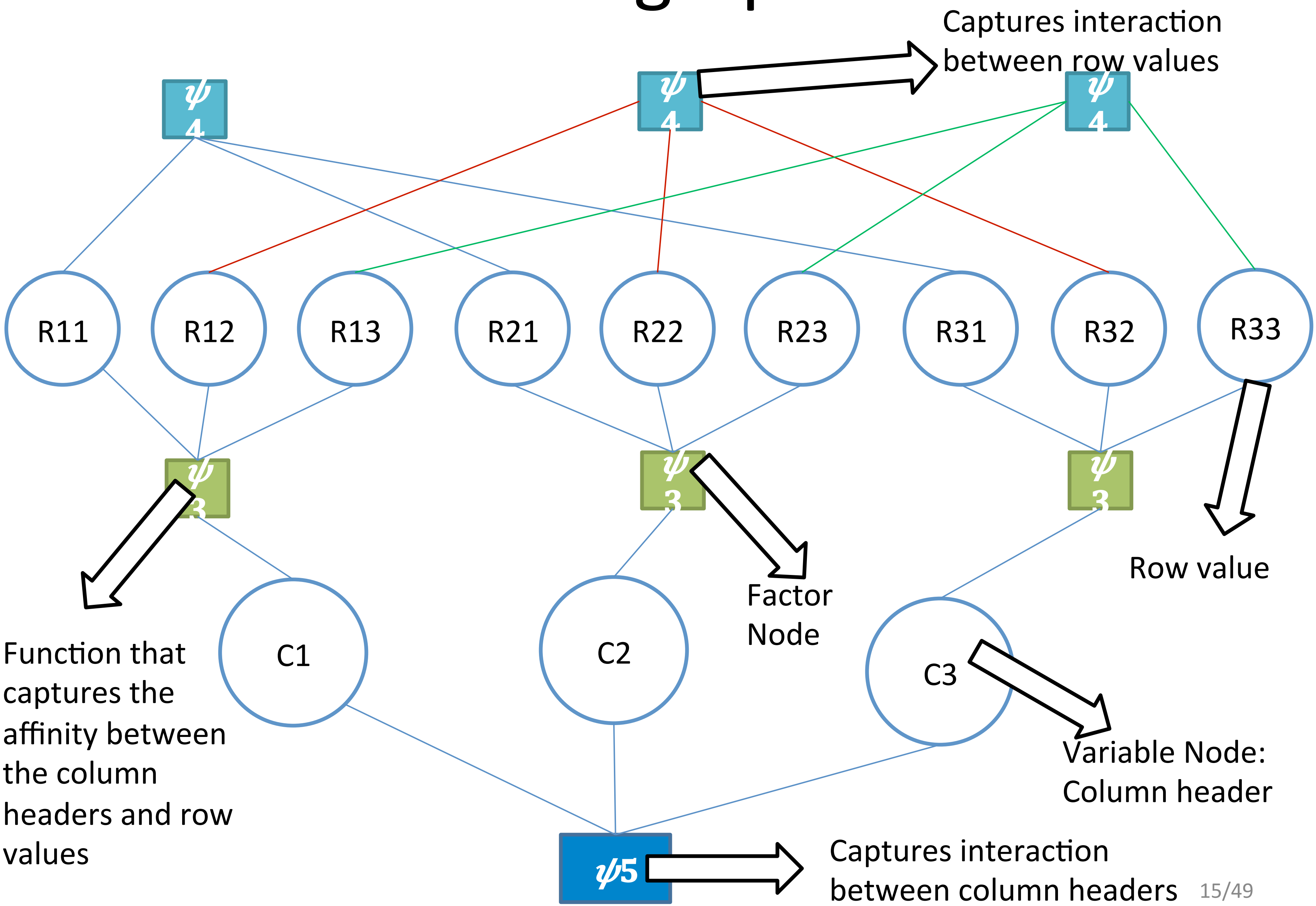


✓ Probabilistic Graphical Models

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?