

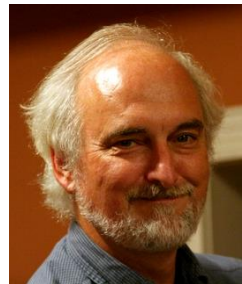


Generating *Linked Data* by *inferring* the *semantics* of *tables*

Varish Mulwad (@varish)

University of Maryland, Baltimore County

September 2, 2011

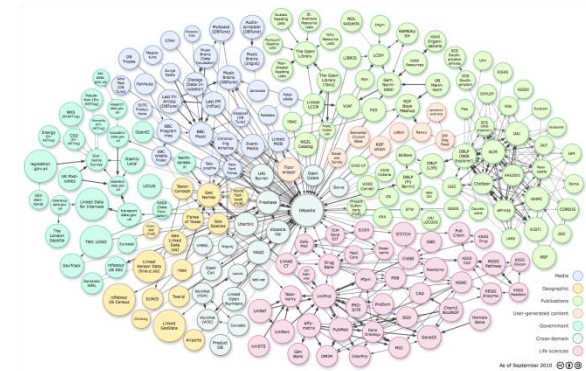


Dr. Tim Finin



Dr. Anupam Joshi

The screenshot shows a web browser window displaying a table of data. The table has multiple columns and rows, with some cells highlighted in yellow. The browser's address bar shows a URL starting with 'http://'. The table appears to be a list of items with various attributes, possibly related to a database or a data repository.



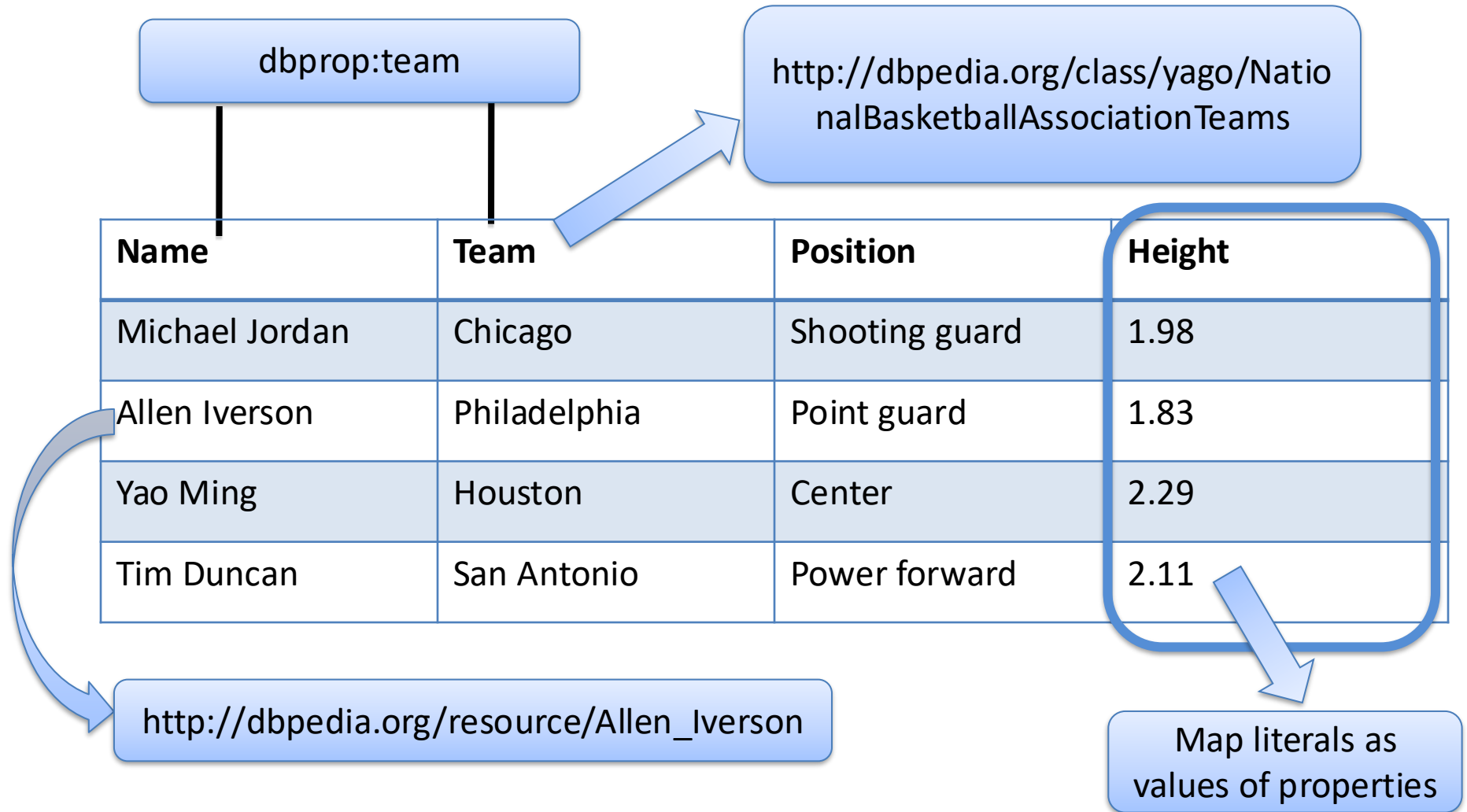
Goal

	ITT		PP	
	n	% (95% CI)	n	% (95% CI)
OAC1W	240/301	79.7 (74.8 to 83.9)	183/219	83.6 (78.1 to 87.9)
OAC2W	246/301	81.7 (77 to 85.7)	185/218	84.9 (79.5 to 89.0)
OA	136/305	44.6 (39.1 to 50.2)	96/224	42.9 (36.5 to 49.4)

ITT, intention-to-treat; PP, per protocol; OA, omeprazole 20 mg twice daily and amoxicillin 1 g twice daily and placebo for 2 weeks; OAC1W, omeprazole 20 mg twice daily and amoxicillin 1 g twice daily and clarithromycin 500 mg twice daily for 1 week, followed by omeprazole 20 mg twice daily and placebo for 1 week; OAC2W, omeprazole 20 mg twice daily and amoxicillin 1 g twice daily and clarithromycin 500 mg twice daily for 2 weeks.

Image from : Zagari RM, Bianchi-Porro G, Fiocca R, Gasbarrini G, Roda E, Bazzoli F. Comparison of 1 and 2 weeks of omeprazole, amoxicillin and clarithromycin treatment for *Helicobacter pylori* eradication: the HYPER Study. Gut. 2007;56: 475-9. [PMID: 17028126]

Contribution



Contribution

Name	Team	Position	Height
Michael Jordan	Chicago	Shooting guard	1.98
Allen Iverson	Philadelphi		
Yao Ming	Houston		
Tim Duncan	San Antonio		

@prefix dbpedia: <http://dbpedia.org/resource/> .

@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .

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

"Name"@en is rdfs:label of dbpedia-owl:BasketballPlayer .

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

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

dbpedia:Michael Jordan a dbpedia-owl:BasketballPlayer .

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

dbpedia:Chicago Bulls a yago:NationalBasketballAssociationTeams .

All this in a completely automated way !! 😊

Introduction & Motivation

Tables are everywhere !



The web – **154 million**
high quality relational
tables (Cafarella et al. 2008)



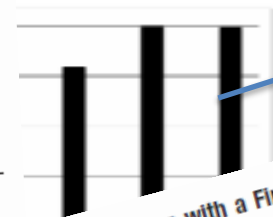
389,697 raw and
geospatial datasets

Evidence-based medicine

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)	



of Clinical trials published in 2008

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)	23.4 (11.6)	0.8
Postmenopausal hormone therapy, %	37.1	36.5	0.01
Body mass index, mean (SD) kg/m ²	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	<0.001

of meta analysis 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.

very low ... hampers effective health

Related Work

- Extracting tables from documents and web pages
 - Hurst (2006), Embley et al. (2006)
- Understanding semantics of tables
 - Wang et al. (2011), Ventis et al. (2011), Limaye et al. (2010)

Current systems

- Use 'semantically poor' knowledge bases
- Only one system focuses on complete table interpretation
- Do not generate Linked Data
- No system tackles literal data
- Critical piece of evidence for interpreting medical tables
- No system dealing with tables in specialized domains (e.g. tables found medical literature)

Building a table interpretation framework



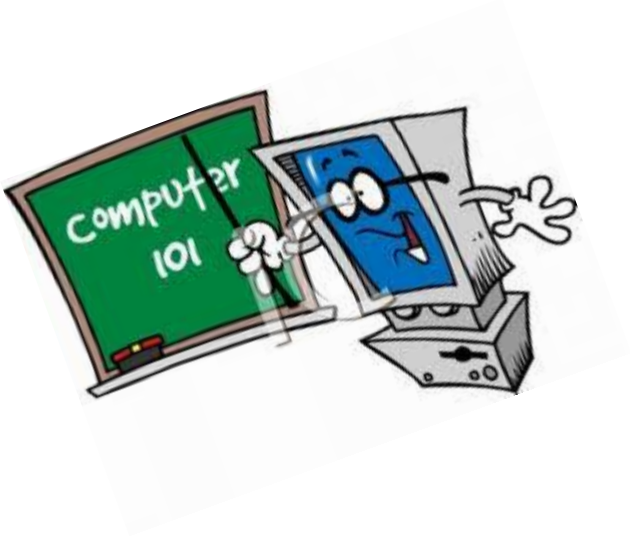
- Preliminary work / Baseline system
- Analysis and Evaluation of baseline
- Framework grounded in graphical models and probabilistic reasoning

The System's Brain (Knowledgebase)



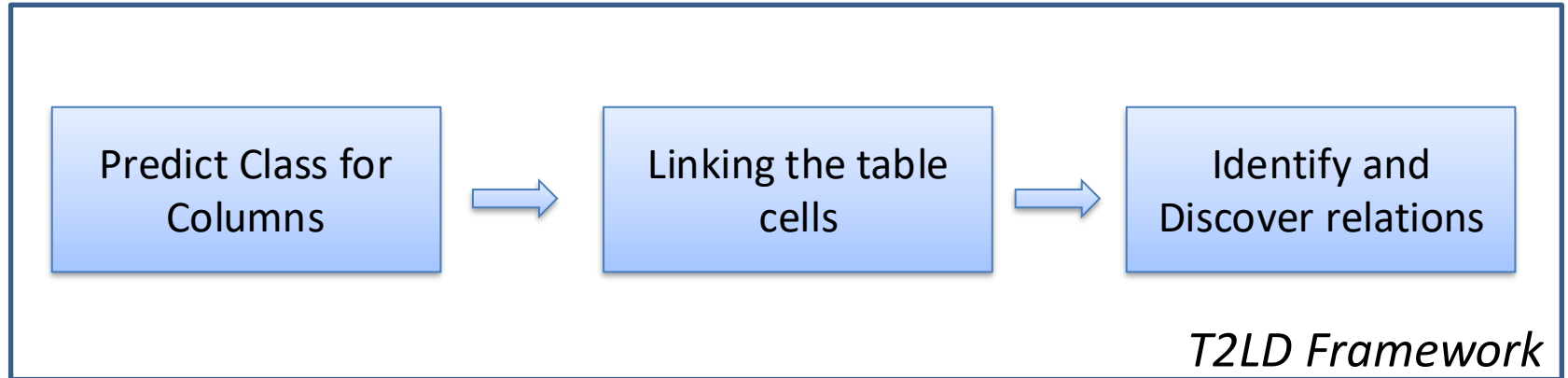
Wikitology¹ – A hybrid knowledgebase where
structured data meets unstructured data

Syed, Z., and Finin, T. 2011. Creating and Exploiting a Hybrid Knowledge Base for Linked Data, volume 129 of Revised Selected Papers Series: Communications in Computer and Information Science. Springer.

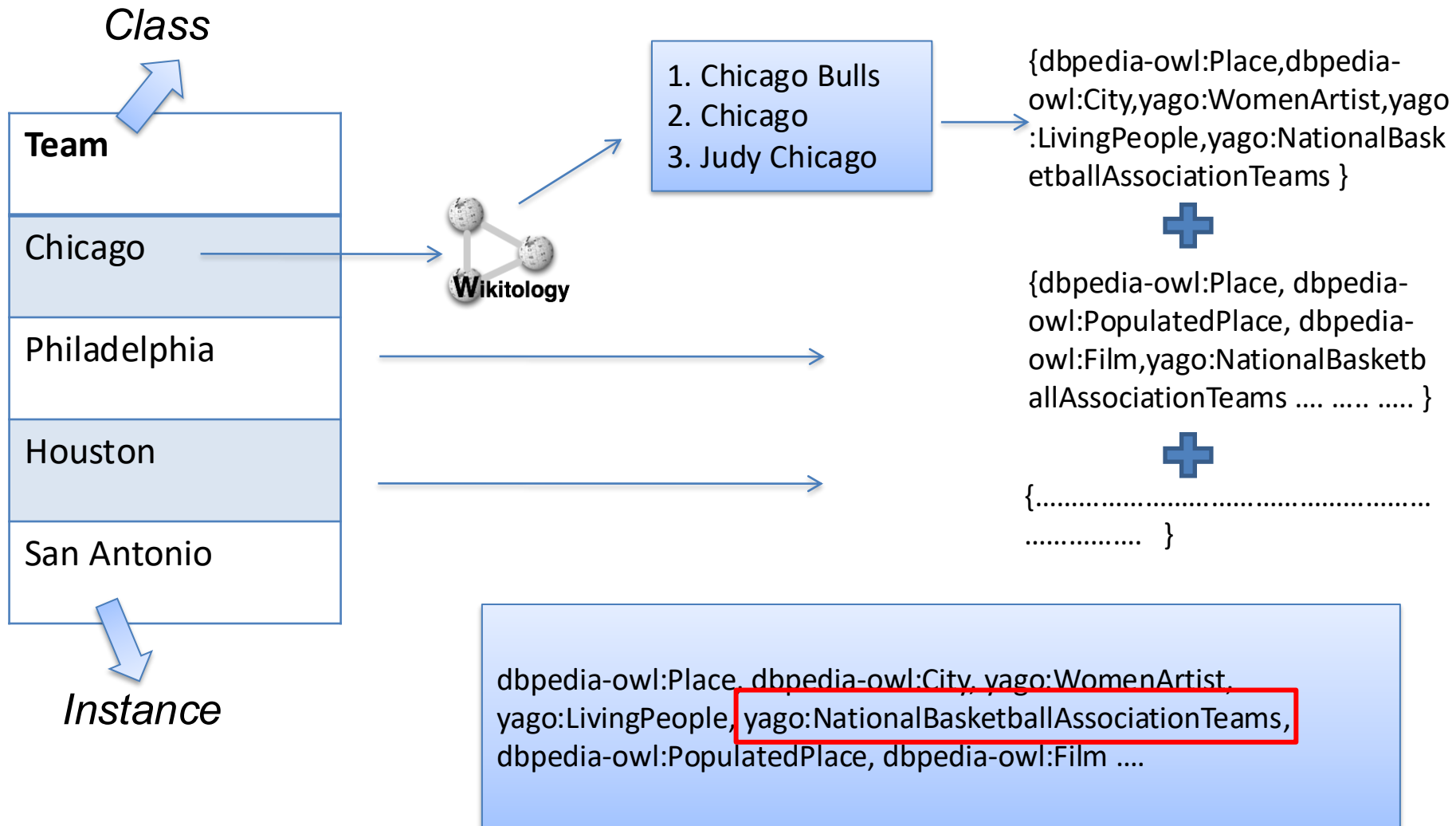


The Baseline System

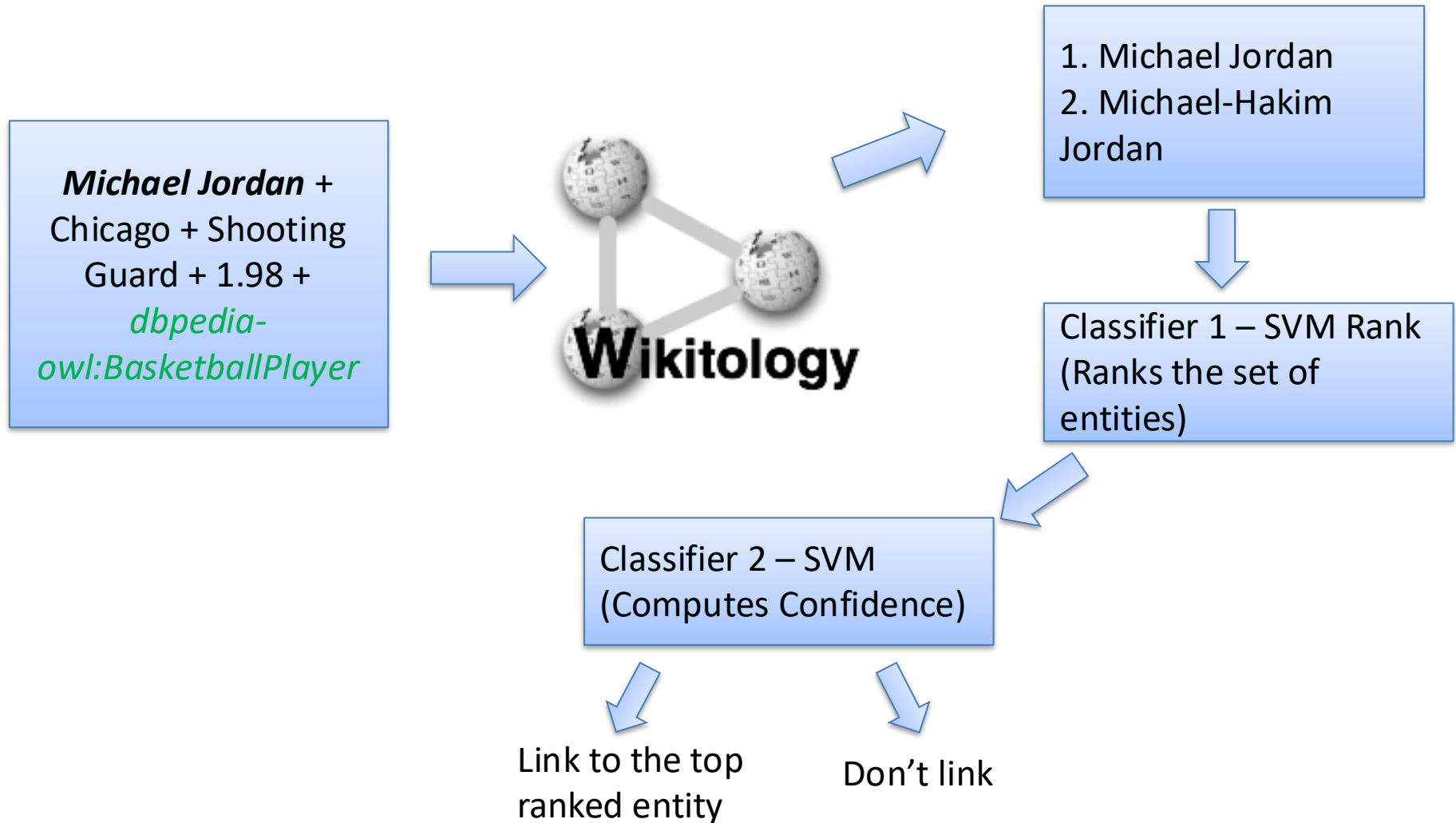
T2LD Framework



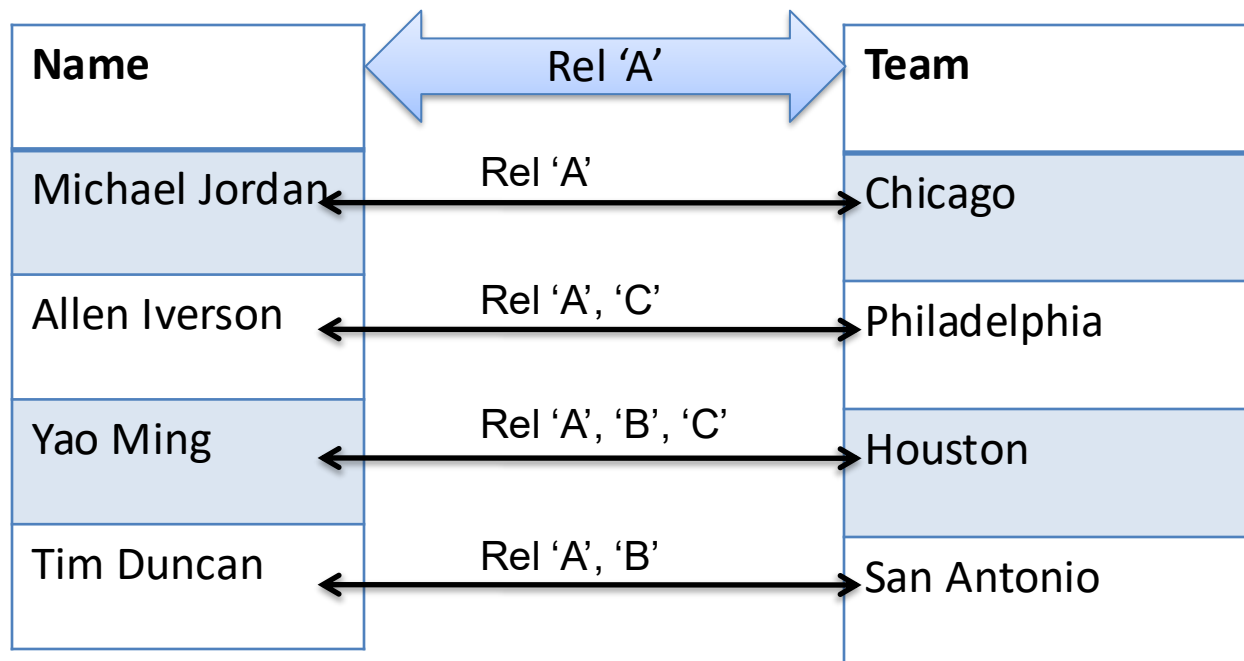
Predicting Class Labels for column



Linking table cells to entities



Identify Relations



Generating a linked RDF representation

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .  
@prefix dbpedia: <http://dbpedia.org/resource/> .  
@prefix dbpedia-owl: <http://dbpedia.org/ontology/> .  
@prefix yago: <http://dbpedia.org/class/yago/> .
```

"Name"@en is rdfs:label of dbpedia-owl:BasketballPlayer .

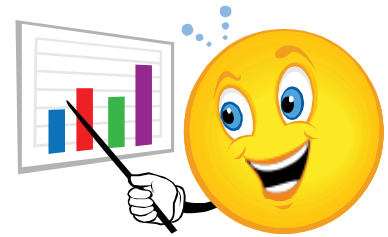
"Team"@en is rdfs:label of yago:NationalBasketballAssociationTeams .

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

dbpedia:Michael Jordan a dbpedia-owl:BasketballPlayer .

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

dbpedia:Chicago Bulls a yago:NationalBasketballAssociationTeams .



Evaluation of the baseline system

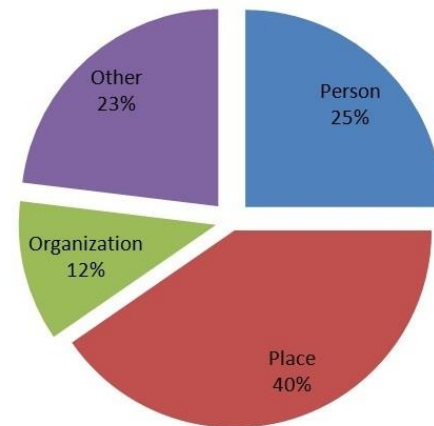
Dataset summary

Number of Tables	15
Total Number of rows	199
Total Number of columns	56 (52)
Total Number of entities	639 (611)

Distribution of Entities per Category



Distribution of Columns per Category

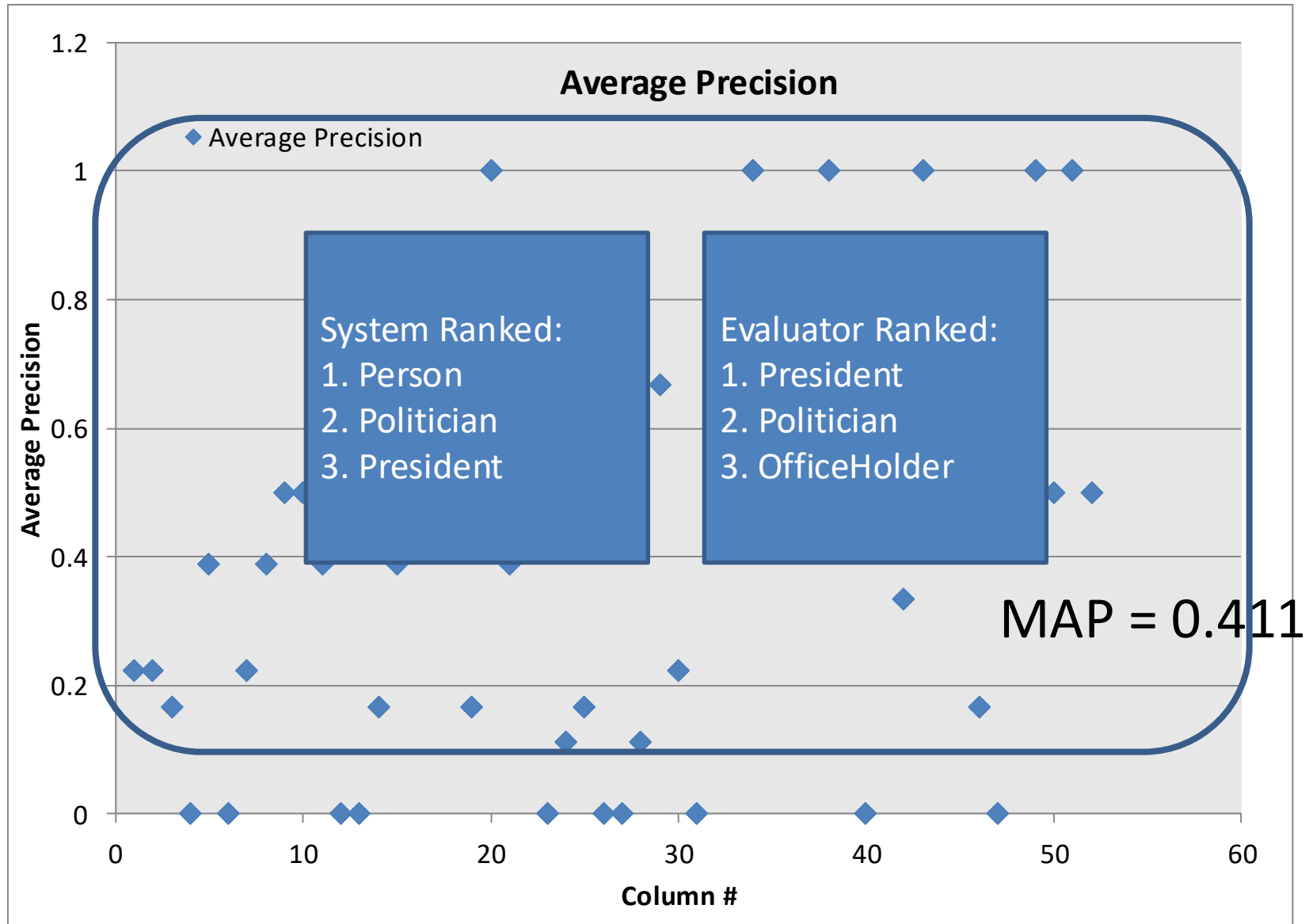


* The number in the brackets indicates # excluding columns that contained numbers

Evaluation # 1 (MAP)

- Compared the system's ranked list of labels against a human-ranked list of labels
- Metric - Average Precision (a.p.) [Mean Average Precision gives a mean over set of queries]
$$\text{AveragePrecision}(q) = \frac{\sum_{n=1}^N P(n) \times R(n)}{\text{total number of relevant labels for } q}$$
- Commonly used in the Information Retrieval domain to compare two ranked sets

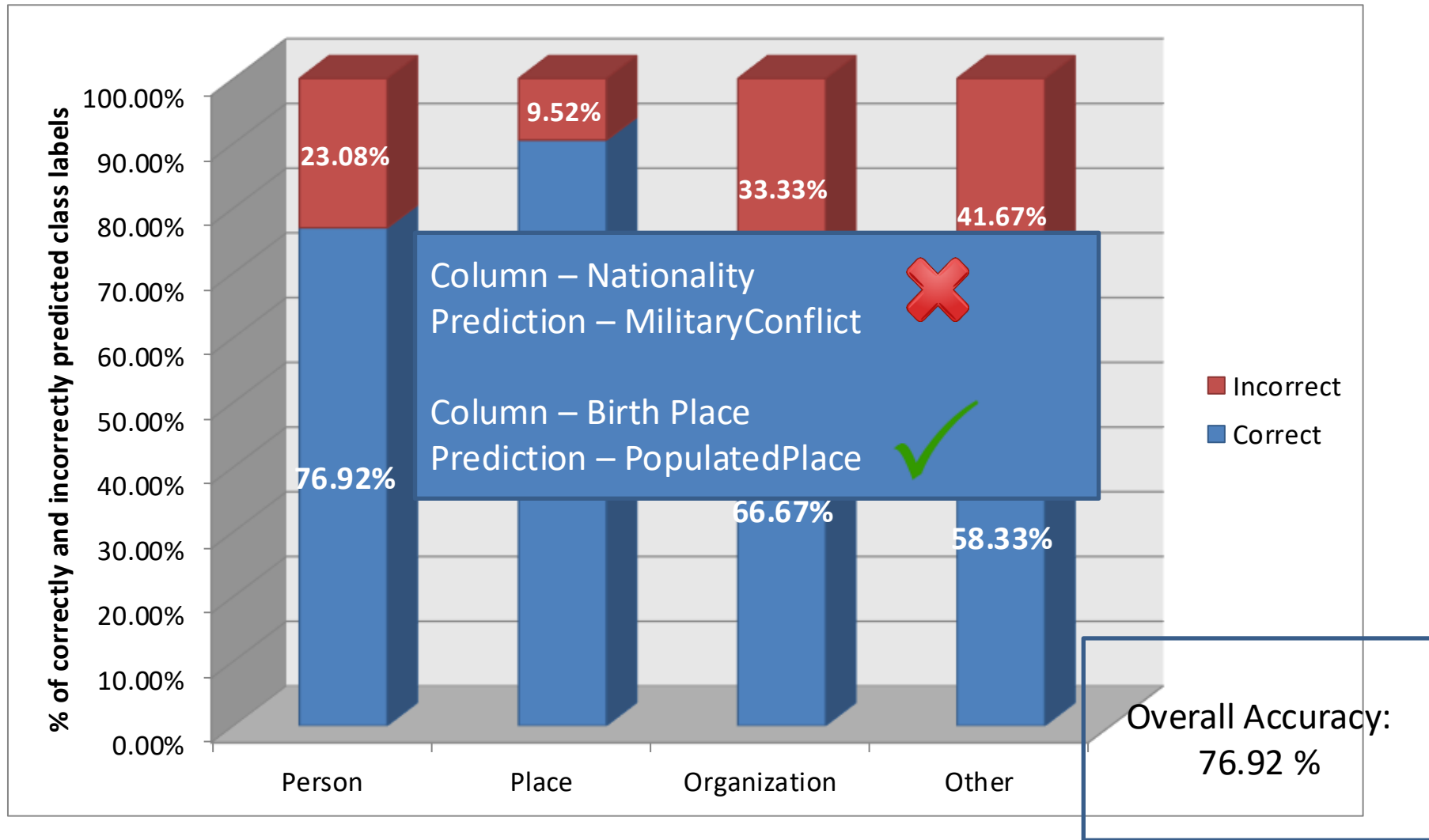
Evaluation # 1 (MAP)



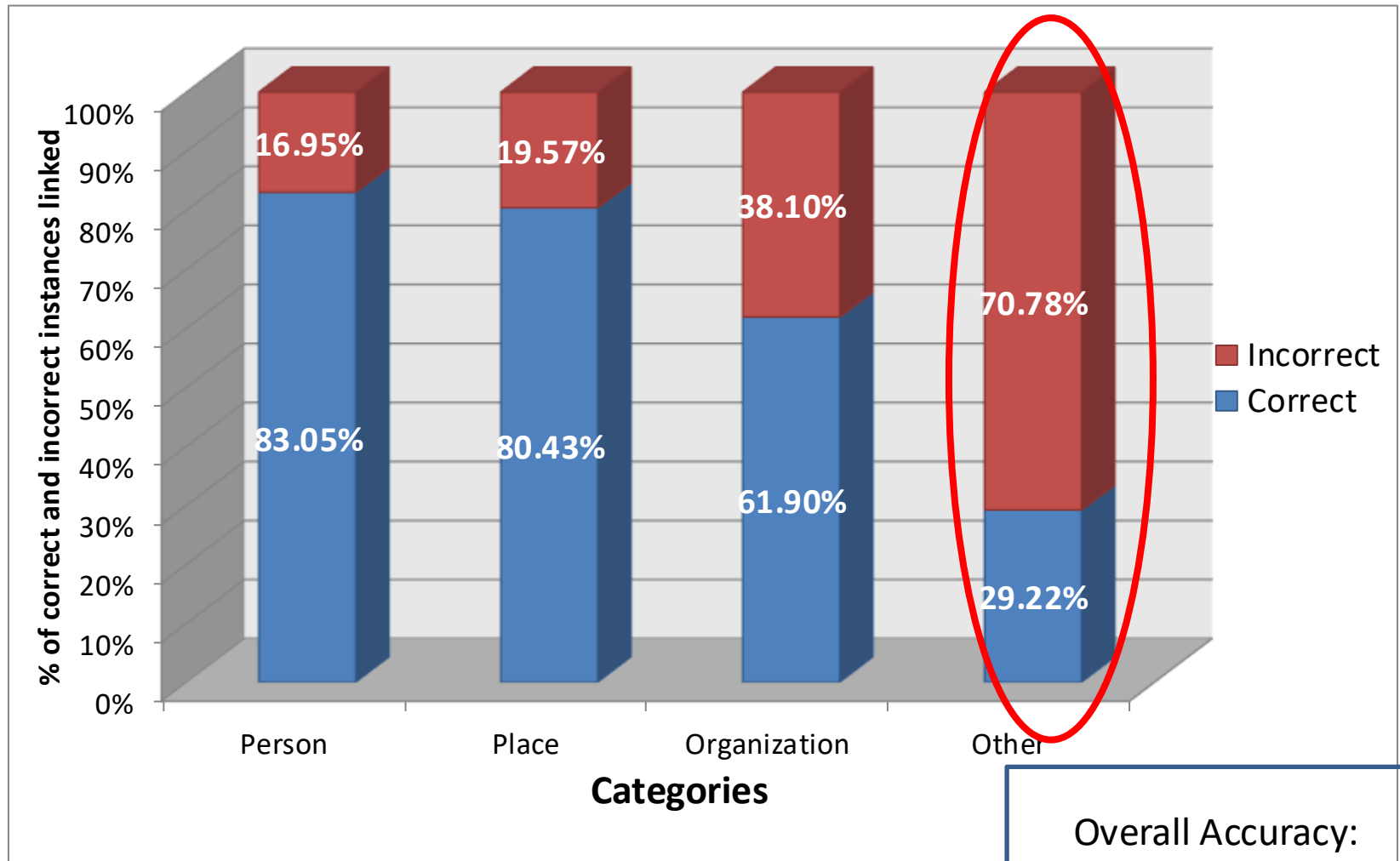
Evaluation # 2 (Correctness)

- Evaluated whether our predicted class labels were “fair and correct”
- Class label may not be the most accurate one, but may be correct
 - E.g. dbpedia:PopulatedPlace is not the most accurate, but still a correct label for column of cities
- Three human judges evaluated our predicted class labels

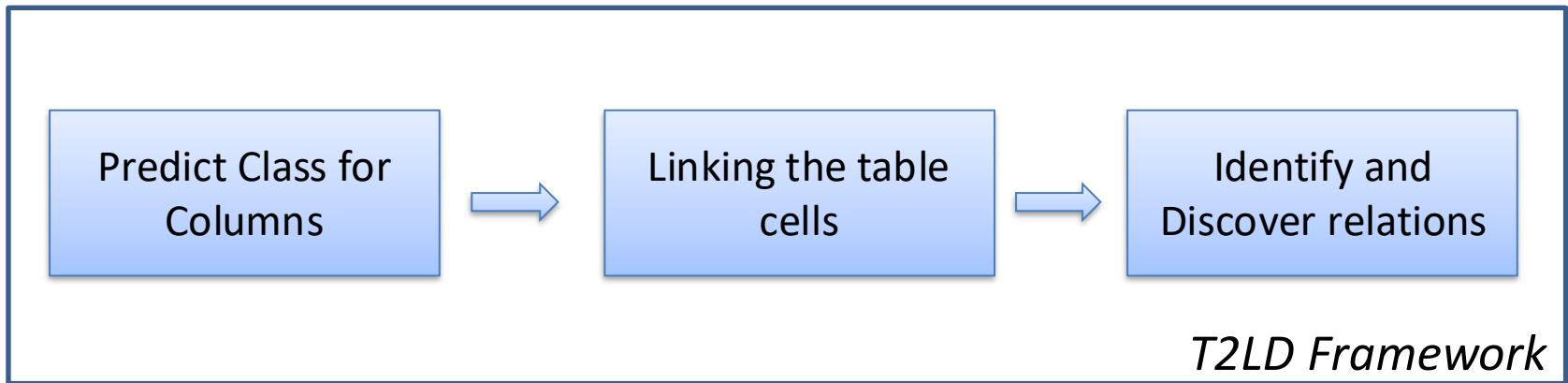
Evaluation # 2 (Correctness)



Accuracy for Entity Linking



Lessons Learnt



- Sequential System – Error percolated from one phase to the next
- Current system favors general classes over specific ones (MAP score = 0.411)
- Largely, a system driven by “heuristics”
- Although we consider evidence, we don’t do assignment jointly

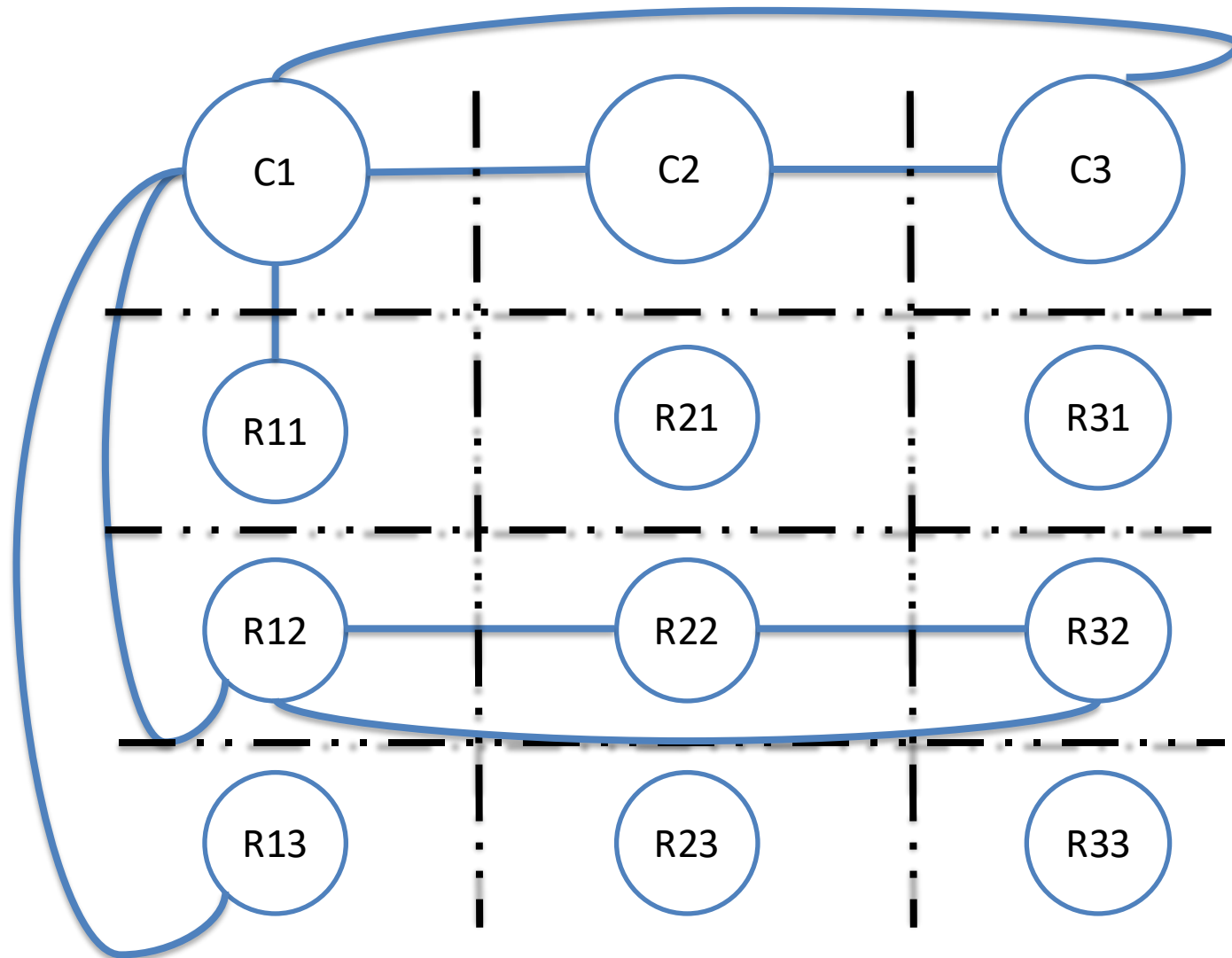
Joint Inference over evidence in a table

- ✓ Probabilistic Graphical Models

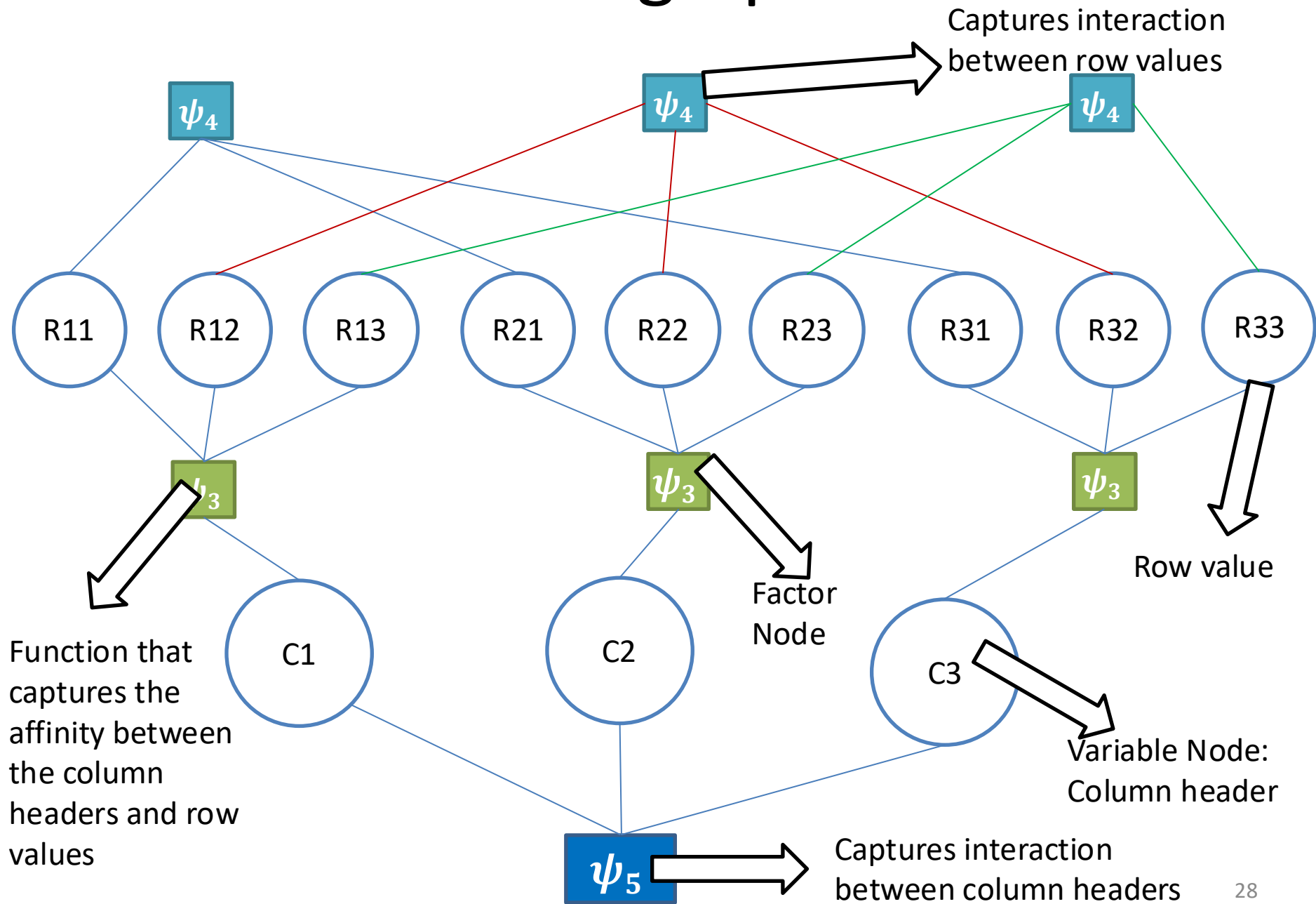
- ✓ Markov logic Networks



A graphical model for tables



Parameterized graphical model



Challenges - Abbreviations

Table 2 *H pylori* eradication rates for each treatment regimen

	ITT		PP	
	n	% (95% CI)	n	% (95% CI)
OAC1W	240/301	79.7 (74.8 to 83.9)	183/219	83.6 (78.1 to 87.9)
OAC2W	246/301	81.7 (77 to 85.7)	185/218	84.9 (79.5 to 89.0)
OA	136/305	44.6 (39.1 to 50.2)	96/224	42.9 (36.5 to 49.4)

ITT, intention-to-treat; PP, per protocol; OA, omeprazole 20 mg twice daily and amoxicillin 1 g twice daily and placebo for 2 weeks; OAC1W, omeprazole 20 mg twice daily and amoxicillin 1 g twice daily and clarithromycin 500 mg twice daily for 1 week, followed by omeprazole 20 mg twice daily and placebo for 1 week; OAC2W, omeprazole 20 mg twice daily and amoxicillin 1 g twice daily and clarithromycin 500 mg twice daily for 2 weeks.

- Other examples:
 - State Abbreviations
 - Stock Tickers
 - Airport Codes
 - Currency codes
- Preprocessing – parse and identify such columns
- Replace abbreviations with expanded forms

Challenges - Literals

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

Age
75
65
50
25

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

Conclusion

- Presented a framework for inferring the semantics of tables and generating Linked data
- Evaluation of the baseline system show feasibility in tackling the problem
- Work in progress for building framework grounded in graphical models and probabilistic reasoning
- Working on tackling challenges posed by tables from domains such as the medical and open government data

References

1. Cafarella, M. J.; Halevy, A. Y.; Wang, Z. D.; Wu, E.; and Zhang, Y. 2008. Webtables:exploring the power of tables on the web. PVLDB 1(1):538–549
2. M. Hurst. Towards a theory of tables. IJDAR,8(2-3):123-131, 2006.
3. D. W. Embley, D. P. Lopresti, and G. Nagy. Notes on contemporary table recognition. In Document Analysis Systems, pages 164-175, 2006.
4. Wang, Jingjing, Shao, Bin, Wang, Haixun, and Zhu, Kenny Q. Understanding tables on the web. Technical report, Microsoft Research Asia, 2010.
5. Venetis Petros, Halevy Alon, Madhavan Jayant, Pasca Marius, Shen Warren, Wu Fei, Miao Gengxin, and Wu Chung. Recovering semantics of tables on the web. In Proc. of the 37th Int'l Conference on Very Large Databases (VLDB), 2011.
6. Limaye Girija, Sarawagi Sunita, and Chakrabarti Soumen. Annotating and searching web tables using entities, types and relationships. In Proc. of the 36th Int'l Conference on Very Large Databases (VLDB), 2010

“A little semantics goes a long way” ~ Jim Hendler

Thank You !

Questions ?

varish1@cs.umbc.edu



@varish

Web: <http://goo.gl/NVu8N>