Annotating and Searching Web Tables for Entities, Types, Quantities, and Relationships

Sunita Sarawagi

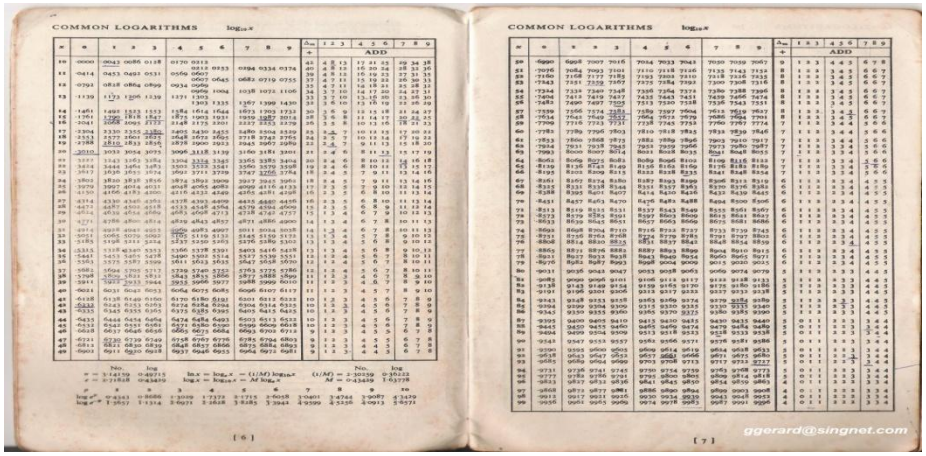
IIT Bombay

Joint work with

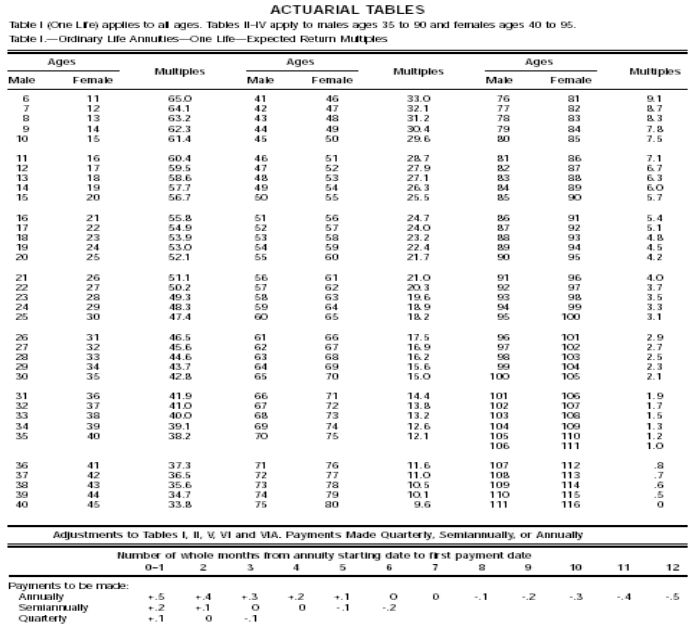
Soumen Chakrabarti, Girija Limaye, Rakesh Pimplikar

Tables: an age-old storage idiom for humans

A table from Sumaria ~1822 B.C.



Logarithm tables ~ 1200 A.D.



ACTUARIAL TABLES

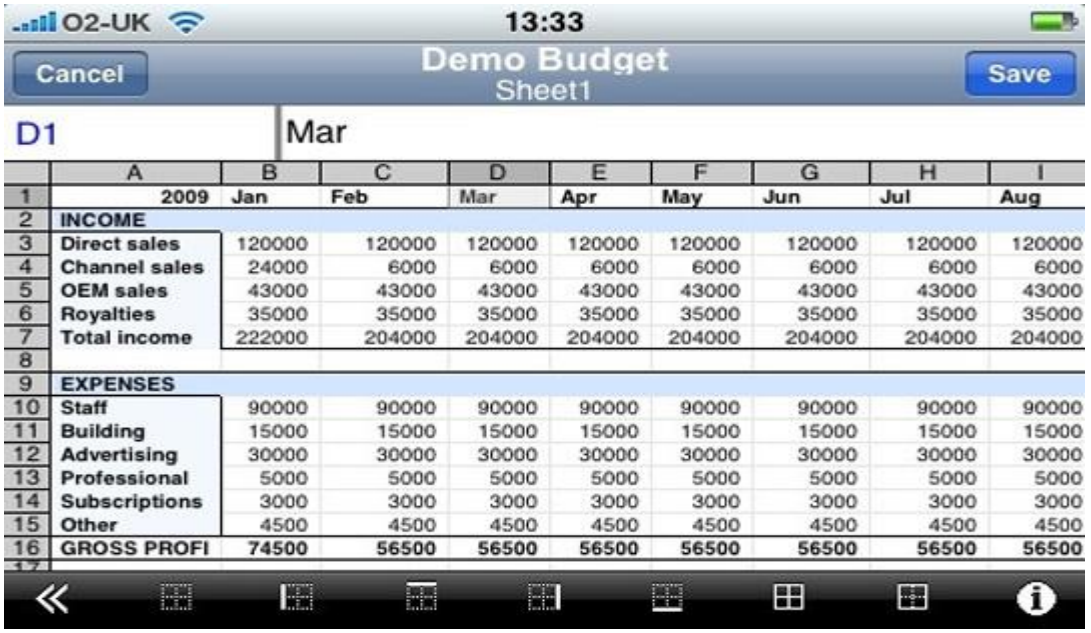
Table I (One Life) applies to all ages. Tables II–IV apply to males ages 35 to 90 and females ages 40 to 95.
Table I.—Ordinary Life Annuities—One Life—Expected Return Multiples

Ages Male	Ages Female	Multiples	Ages Male	Ages Female	Multiples	Ages Male	Ages Female	Multiples
6	11	65.0	41	46	33.0	76	81	9.1
7	12	64.1	42	47	32.1	77	82	8.7
8	13	63.2	43	48	31.2	78	83	8.3
9	14	62.3	44	49	30.4	79	84	7.8
10	15	61.4	45	50	29.6	80	85	7.5
11	16	60.4	46	51	28.7	81	86	7.1
12	17	59.5	47	52	27.9	82	87	6.7
13	18	58.6	48	53	27.1	83	88	6.3
14	19	57.7	49	54	26.3	84	89	6.0
15	20	56.7	50	55	25.5	85	90	5.7
16	21	55.8	51	56	24.7	86	91	5.4
17	22	54.9	52	57	24.0	87	92	5.1
18	23	53.9	53	58	23.2	88	93	4.8
19	24	53.0	54	59	22.4	89	94	4.5
20	25	52.1	55	60	21.7	90	95	4.2
21	26	51.1	56	61	21.0	91	96	4.0
22	27	50.2	57	62	20.3	92	97	3.7
23	28	49.3	58	63	19.6	93	98	3.5
24	29	48.3	59	64	18.9	94	99	3.3
25	30	47.4	60	65	18.2	95	100	3.1
26	31	46.5	61	66	17.5	96	101	2.9
27	32	45.6	62	67	16.9	97	102	2.7
28	33	44.6	63	68	16.2	98	103	2.5
29	34	43.7	64	69	15.6	99	104	2.3
30	35	42.8	65	70	15.0	100	105	2.1
31	36	41.9	66	71	14.4	101	106	1.9
32	37	41.0	67	72	13.8	102	107	1.7
33	38	40.0	68	73	13.2	103	108	1.5
34	39	39.1	69	74	12.6	104	109	1.3
35	40	38.2	70	75	12.1	105	110	1.2
						106	111	1.0
36	41	37.3	71	76	11.6	107	112	.8
37	42	36.5	72	77	11.0	108	113	.7
38	43	35.6	73	78	10.5	109	114	.6
39	44	34.7	74	79	10.1	110	115	.5
40	45	33.8	75	80	9.6	111	116	0

Adjustments to Tables I, II, V, VI and VIA. Payments Made Quarterly, Semiannually, or Annually

	Number of whole months from annuity starting date to first payment date: 0–1	2	3	4	5	6	7	8	9	10	11	12
Payments to be made:												
Annually	+.5	+.4	+.3	+.2	+.1	0	0	-.1	-.2	-.3	-.4	-.5
Semiannually	+.2	+.1	0	0	-.1	-.2						
Quarterly	+.1	0	-.1									

Actuarial table ~1800 A.D



	A	B	C	D	E	F	G	H	I
1	2009	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug
2	INCOME								
3	Direct sales	120000	120000	120000	120000	120000	120000	120000	120000
4	Channel sales	24000	6000	6000	6000	6000	6000	6000	6000
5	OEM sales	43000	43000	43000	43000	43000	43000	43000	43000
6	Royalties	35000	35000	35000	35000	35000	35000	35000	35000
7	Total income	222000	204000	204000	204000	204000	204000	204000	204000
8									
9	EXPENSES								
10	Staff	90000	90000	90000	90000	90000	90000	90000	90000
11	Building	15000	15000	15000	15000	15000	15000	15000	15000
12	Advertising	30000	30000	30000	30000	30000	30000	30000	30000
13	Professional	5000	5000	5000	5000	5000	5000	5000	5000
14	Subscriptions	3000	3000	3000	3000	3000	3000	3000	3000
15	Other	4500	4500	4500	4500	4500	4500	4500	4500
16	GROSS PROFI	74500	56500	56500	56500	56500	56500	56500	56500

A modern day spreadsheet

Tables on the Web

Languages by Countries

Afghanistan	Dari Persian, Pashtu (both official), other Turkic and minor languages
Albania	Albanian (Tosk is the official dialect), Greek
Algeria	Arabic (official), French, Berber dialects
Andorra	Catalán (official), French, Castilian, Portuguese
Angola	Portuguese (official), Bantu and other African languages
Antigua and Barbuda	English (official), local dialects
Argentina	Spanish (official), English, Italian, German, French
Armenia	Armenian 98%, Yezidi, Russian

Z	Name	Sym	Period
1	Hydrogen	H	1
2	Helium	He	1
3	Lithium	Li	2
4	Beryllium	Be	2
5	Boron	B	2
6	Carbon	C	2

Country	Number of Students	Language
Bosnia	4	Bosnian
Brazil	1	Portuguese
China	4	Chinese
Hong Kong	3	Chinese
India	1	Assamese
India	1	Marathi
India	1	Punjabi
India	6	Tamil
India	1	Telugu
India	1	Mahavalam

Mountain Pass	Elevation (m/ft)
Tonale pass	1884 (6181)
Colle Maniva	1669 (5476)
Auden's Col	(17552)
Crown saddle	1076

Tables and Quantities

- 40% of Web tables columns are quantities
- Long tail of quantities missing from Knowledge Bases (KBs) but present in Web Tables
  - Total user accounts of Skype in 2012 Q1
  - Energy density of magnetic fields
  - Syndicated loan volumes
  - Quantity of major oil spills

QUANTITIES OF OIL SPILT

Position	Ship name	Year	Location	Spill Size (tonnes)
1	Atlantic Empress	1979	Off Tobago, West Indies	287,000
2	ABT Summer	1991	700 nautical miles off Angola	260,000
3	Castillo de Bellver	1983	Off Saldanha Bay, South Africa	252,000
4	Amoco Cadiz	1978	Off Brittany, France	223,000
5	Haven	1991	Genoa, Italy	144,000
6	Odyssey	1988	700 nautical miles off Nova Scotia, Canada	132,000

Tables on the Web

No headers

Languages by Countries

Afghanistan	Dari Persian, Pashtu (both official), other Turkic and minor languages
Albania	Albanian (Tosk is the official dialect), Greek
Algeria	Arabic (official), French, Berber dialects
Andorra	Catalán (official), French, Castilian, Portuguese
Angola	Portuguese (official), Bantu and other African languages
Antigua and Barbuda	English (official), local dialects
Argentina	Spanish (official), English, Italian, German, French
Armenia	Armenian 98%, Yezidi, Russian

Non-informative headers

Z	Name	Sym	Period
1	Hydrogen	H	1
2	Helium	He	1
3	Lithium	Li	2
4	Beryllium	Be	2
5	Boron	B	2
6	Carbon	C	2

OK headers

Not official language.

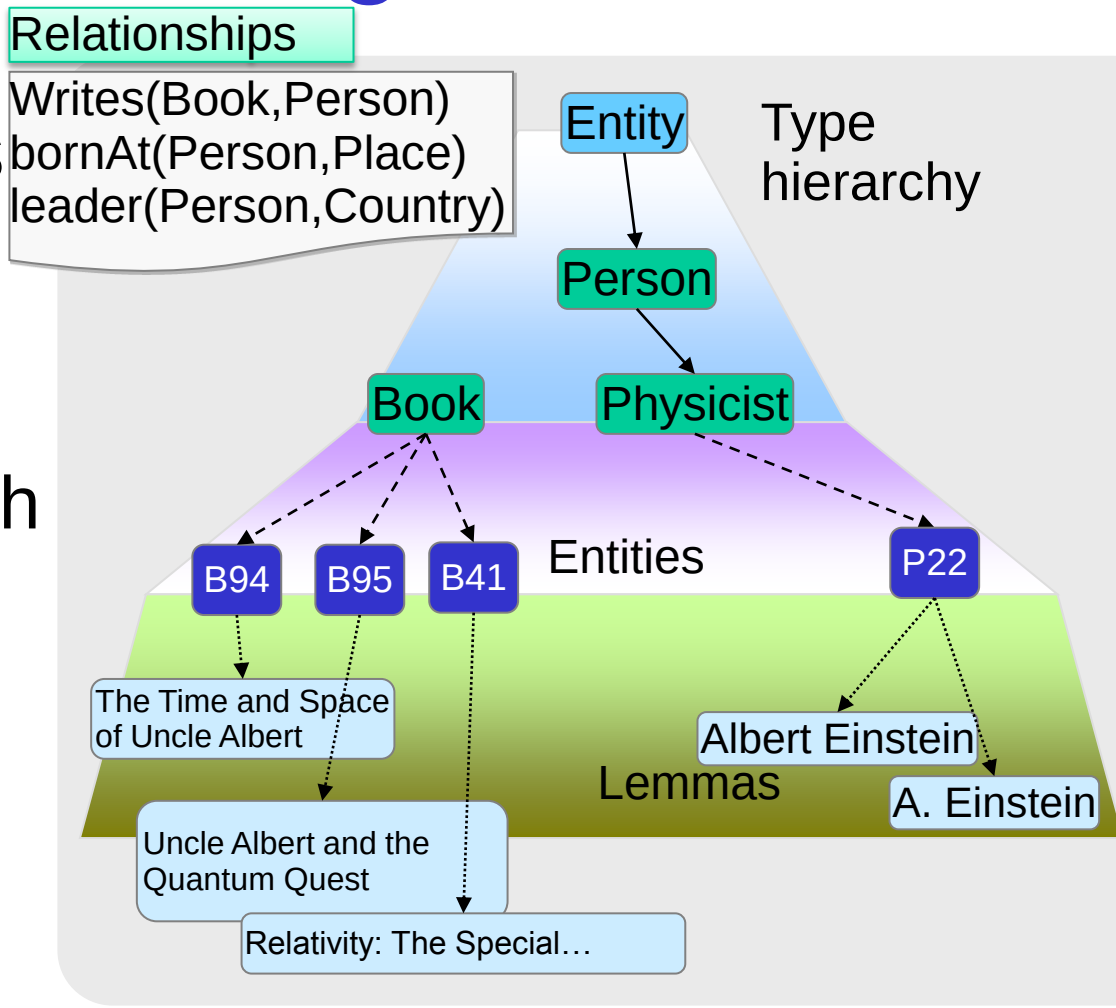
Country	Number of Students	Language
Bosnia	4	Bosnian
Brazil	1	Portuguese
China	4	Chinese
Hong Kong	3	Chinese
India	1	Assamese
India	1	Marathi
India	1	Punjabi
India	6	Tamil
India	1	Telugu
India	1	Malayalam

Useless without correct unit parsing

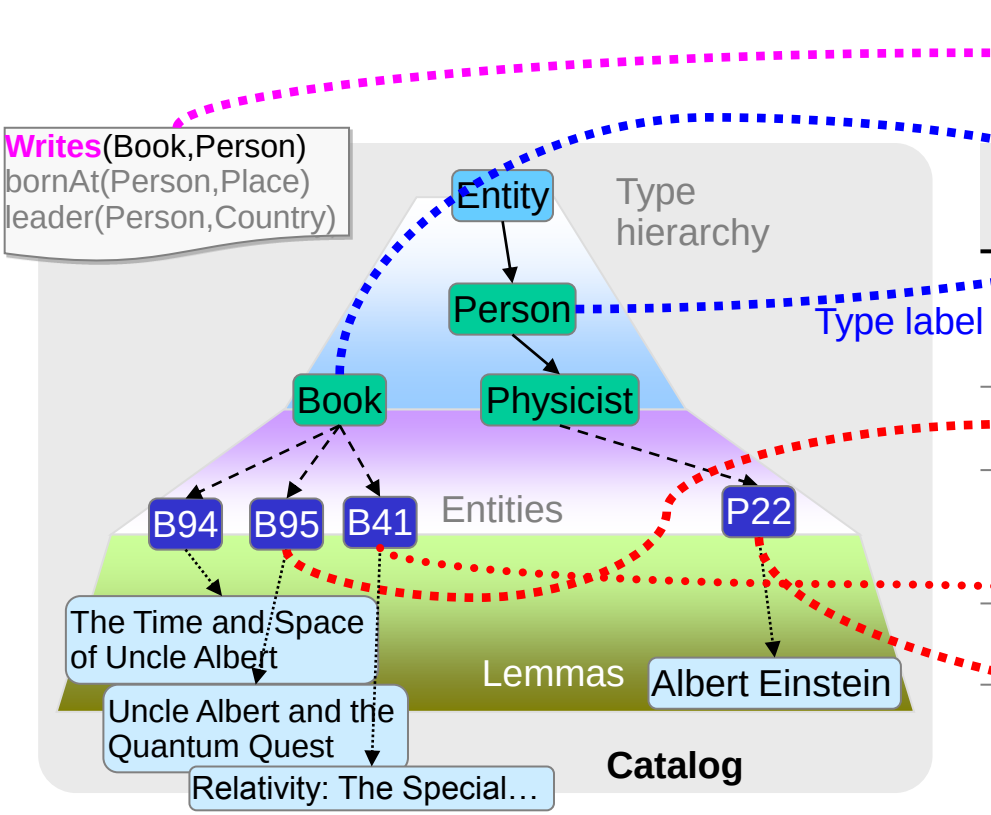
Pass	Elevation (m/ft)
Tonale pass	1884 (6181)
Colle Maniva	1669 (5476)
Auden's Col	(17552)
Crown saddle	1076

Ontological Knowledge Base

- Entities,Types,Relations
  - Freebase
  - DBPedia
  - Google Knowledge Graph
  - Microsoft Satori
  - YAGO
    - ~ 350 K types
    - ~ 10 million entities
    - ~ 100 relation types
    - ~ 120 million relation instances



Annotating Tables with Entity, Type, and Relation links



A Reference Ontology

Relation label

Type label

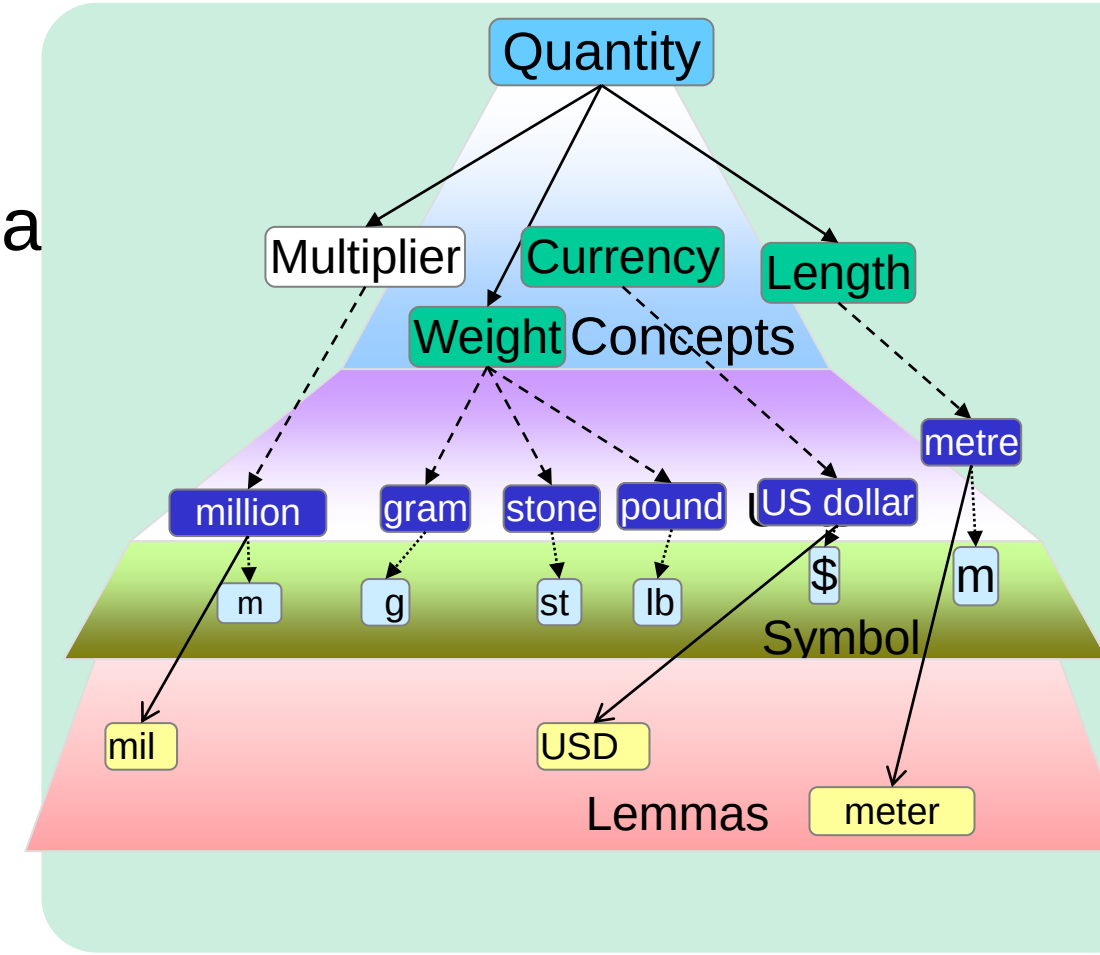
Entity label

Title	Author
Uncle Petros and the Goldback conjecture	A Doxiadis
Uncle Albert and the Quantum Quest	Russell Stannard
Relativity: The Special and the General Theory	A Einstein

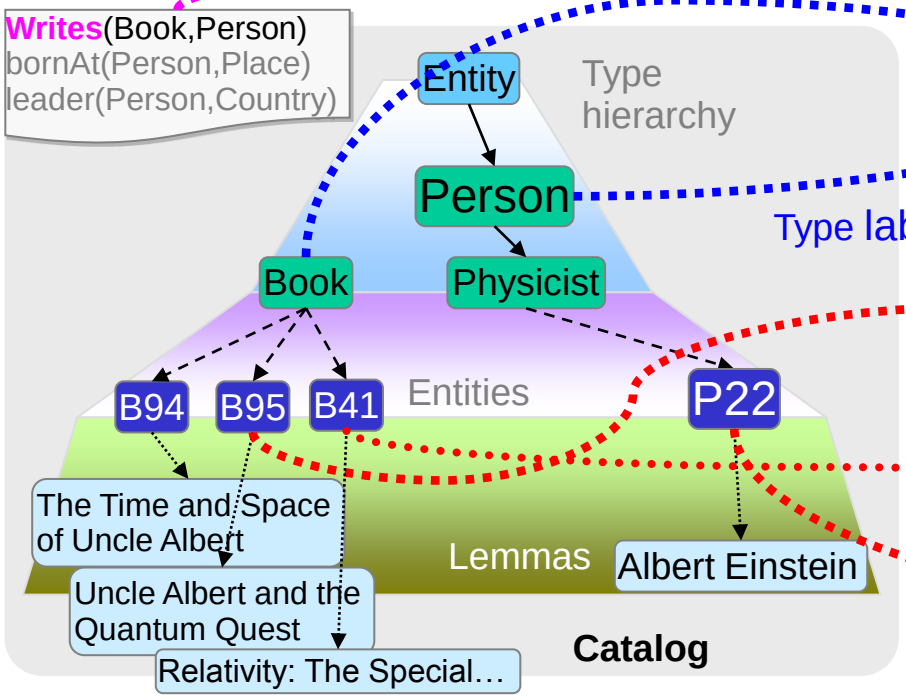
A web table

Unit Catalog

- QuTree
  - Seeded from Wikipedia
  - + lemmas from Web
    - ~ 44 concepts (types)
    - ~ 750 units (entities)



Annotating Tables with Entity, Type, Unit, and Relation links



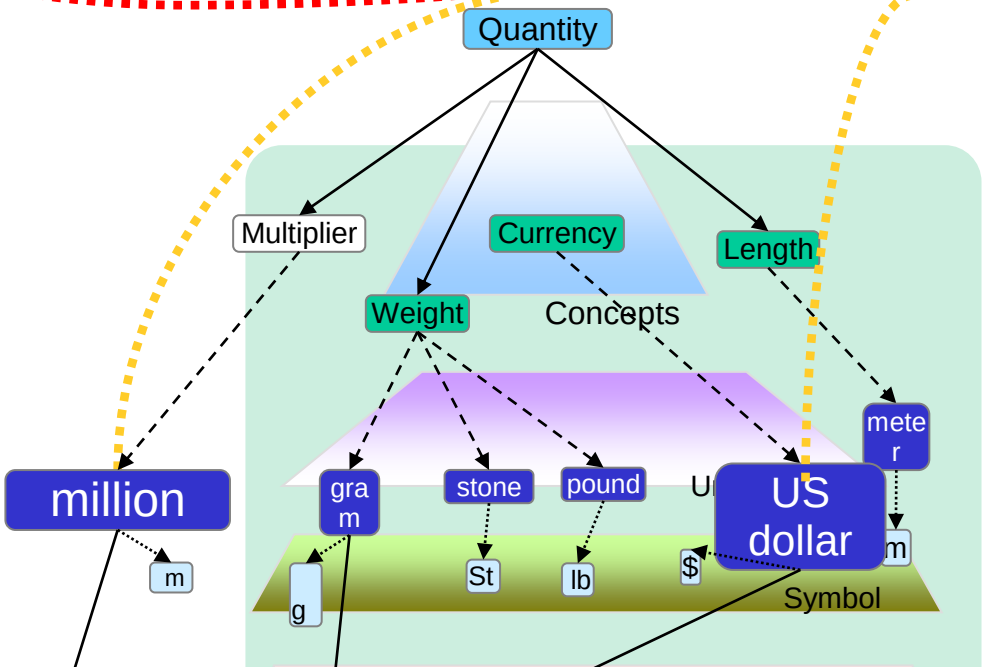
Relation label

Type label

Entity label

Unit label

Title	Author	Sales ($ mil)
Uncle Petros and the Goldback conjecture	A Doxiadis	10
Uncle Albert and the Quantum Quest	Russell Stannard	5
Relativity: The Special and the General Theory	A Einstein	25



Query-Time Types/Relationships

User Query

Name of Explorers	Nationality	Areas Explored

Index Probe Relevant Tables

Web Table 1

List of explorers - Wikipedia, the free encyclopedia

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

Web Table 2

This article lists the explorations in history. For the documentary 'Explorations, powered by Duracell', see Explorations (TV)

Exploration (Chronological order)	Who (explorer)
Sea route to India	Vasco da Gama
Caribbean	Christopher Columbus
Oceania	Abel Tasman
...	...

Web Table 3

Other Formal Reserves 1.3 Forest Reserves under the Forestry Act 1920

All areas will be available for mineral exploration and mining

Forest reserves		
ID	Name	Area
7	Shakespeare Hills	2236
9	Plains Creek	880
13	Welcome Swamp	168
...	...	...

Query-Time Types/Relationships

User Query

Name of Explorers	Nationality	Areas Explored

Web Table 1

List of explorers - Wikipedia, the free encyclopedia

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

Web Table 2

This article lists the explorations in history. For the documentary 'Explorations, powered by Duracell', see Explorations (TV)

Exploration (Chronological order)	Who (explorer)
Sea route to India	Vasco da Gama
Caribbean	Christopher Columbus
Oceania	Abel Tasman
...	...

Web Table 3

Other Formal Reserves 1.3 Forest Reserves under the Forestry Act 1920

All areas will be available for mineral exploration and mining

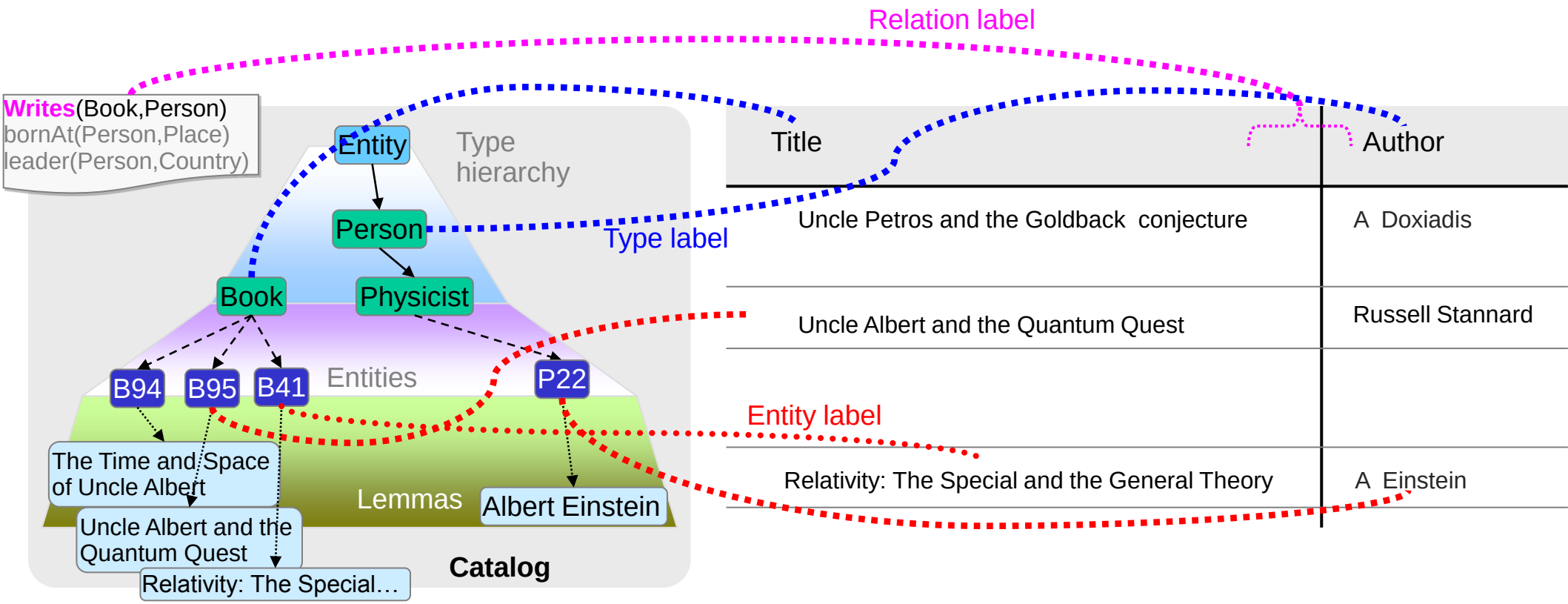
Forest reserves		
ID	Name	Area
7	Shakespeare Hills	2236
9	Plains Creek	880
13	Welcome Swamp	168
...	...	...

Covers the long tail of types + relationships not present in any ontology.

Outline

1. Entity,Type,Relationship annotation to an existing ontology
2. Query-time annotation for new types and relationships
3. Unit annotation for quantity columns

Annotating Tables with Entity, Type, and Relation links



Challenges

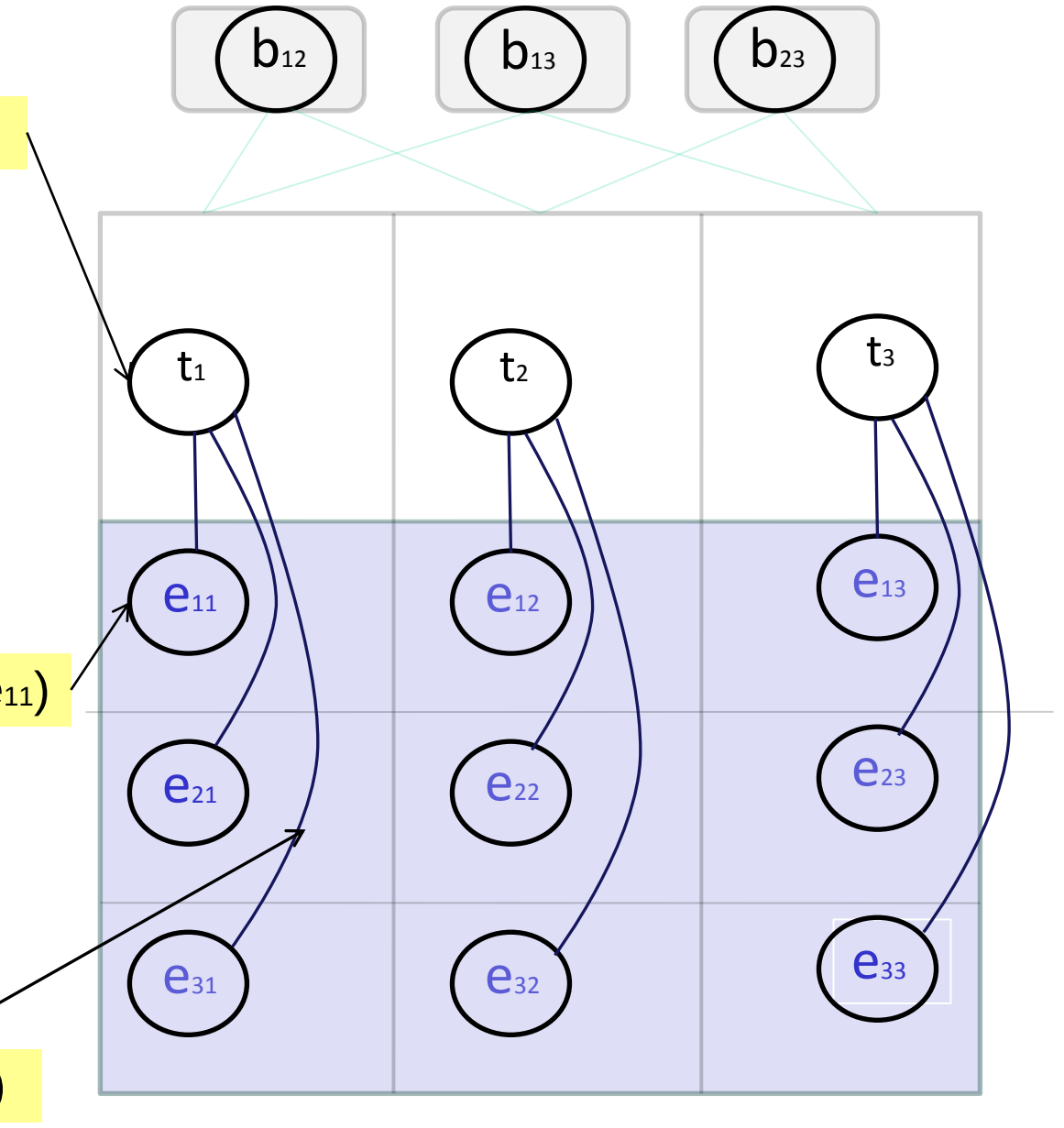
- Usual challenges of entity annotation
  - Noisy mentions
    - A. Einstien Vs Albert Einstein
  - Ambiguity
    - “Hydrogen” both a chemical element and a place name
- New challenges of type annotation
  - Multiple labels
    - YAGO has average 2.2 types per entity
  - Missing type links in Ontology
    - Universities in Toronto ➔ Universities in Ontario.
    - Satyajit Ray ➔Indian film directors

A simple approach - Least common ancestor

- Entity Link:
  - link each cell to the entity with the most similar lemma string
- Type Link:
  - Least common type ancestor of the entity.

Over generalizes to entity even with perfect entity annotation

Collective annotation approach

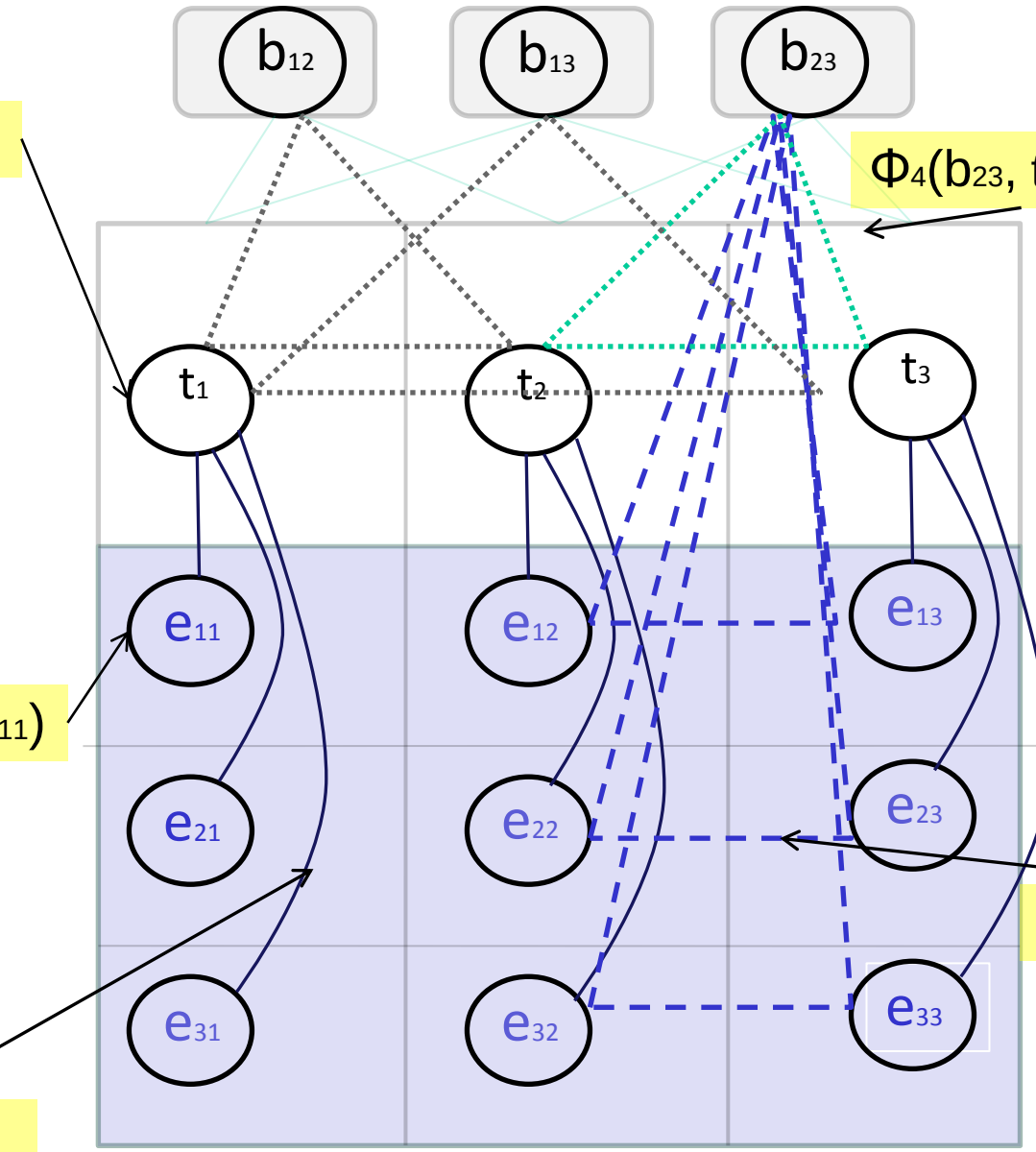


Weighted sum of string similarity between cell and entity/type lemma

Compatibility between type and entity labels (specially designed for missing links)

A table with three rows, three columns and a header row

Collective annotation approach



$\Phi_2(1,t_1)$

$\Phi_4(b_{23}, t_2, t_3)$

Triple's support in ontology

Weighted sum of string similarity between cell(1,1) and entity lemma

$\Phi_1(1,1,e_{11})$

Compatibility between type and entity labels (specially designed for missing links)

$\Phi_5(b_{23}, e_{32}, e_{33})$

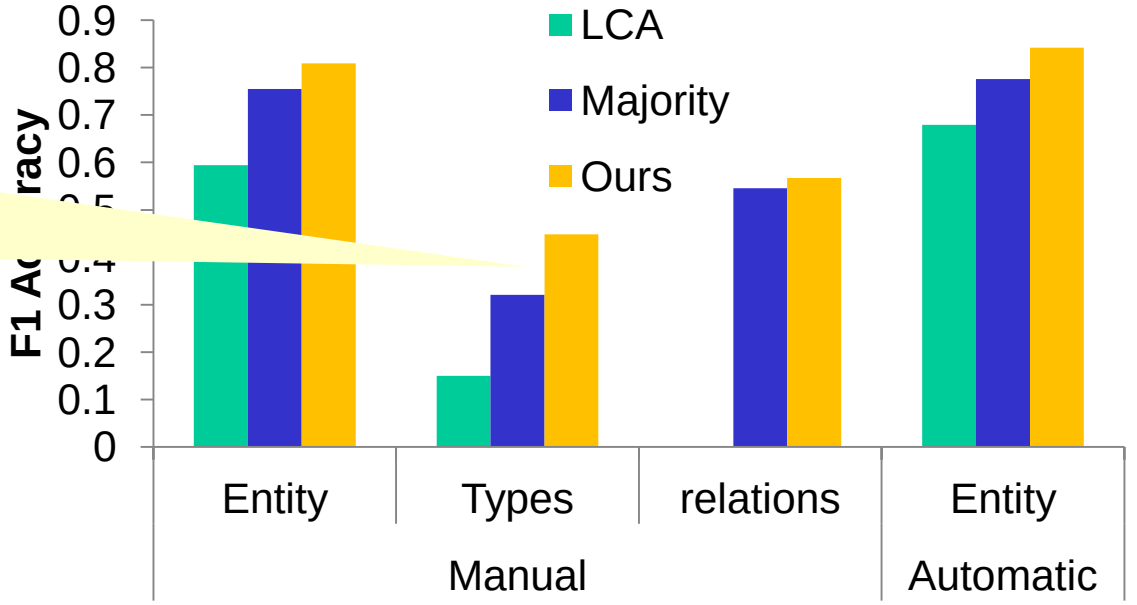
$\Phi_3(t_1, e_{31})$

Exact: NP-hard, Belief propagation on factor graph

Accuracy of joint labeling

- Dataset
  - Manually labeled
    - 450 tables spanning general Web and Wikipedia
  - Automatically labeled
    - 6500 tables from Wikipedia where cells have entity links

Type annotation improves from 32% to 44%



Dataset available at http://www.cse.iitb.ac.in/~sunita/wwt

Outline

1. Entity,Type,Relationship annotation to an existing ontology
2. Query-time annotation for new types and relationships
3. Unit annotation for quantity columns

Query-time annotation

User Query

Name of Explorers	Nationality	Areas Explored

Web Table 1

List of explorers - Wikipedia, the free encyclopedia

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

Web Table 2

This article lists the explorations in history. For the documentary 'Explorations, powered by Duracell', see Explorations (TV)

Exploration (Chronological order)	Who (explorer)
Sea route to India	Vasco da Gama
Caribbean	Christopher Columbus
Oceania	Abel Tasman
...	...

Web Table 3

Other Formal Reserves 1.3 Forest Reserves under the Forestry Act 1920

All areas will be available for mineral exploration and mining

Forest reserves		
ID	Name	Area
7	Shakespeare Hills	2236
9	Plains Creek	880
13	Welcome Swamp	168
...	...	...

Two tasks

1. Identify relevant tables
2. Annotate their columns to query columns

Challenges

- More difficult than matching to an ontology
  - No entities or lemmas on the query-side
  - Table header+context is all that we have.
- Table header can be noisy, missing, uninformative
  - HTML table header tag is not always used (80%).
  - Many tables have no headers (18%).
  - Header text is often uninformative.
  - Context does not give column specific information .

Two contributions

- Carefully designed segmented similarity model to match a query-type-string to a table column

- Collective annotation of multiple tables that leverages content overlap.

Segmented similarity

- Query :- comprises of two parts
  - Top-level type → column header
  - Modifier → context, table's content cells, header itself
- Similarity: the best match over all segmentations
- Example: Modifier in context

User Query **Nobel Prize Winners**

The present list contains laureates under the country/countries that are stated by the **Nobel Prize** committee on its website.

Year	Winners	Subject
1902	Ronald Ross	Medicine
1907	Rudyard Kipling	Literature
…	…	…

Segmented similarity

- Modifier in content cells → select a subset of rows.

Black metal bands

User Query

Band name	Country	Genre
Aarcon	Germany	Black Metal
Act of God	Russia	Melodic Black
Adragard	Italy	Black Metal
…	…	…

- Modifier in other header rows

User Query

Name of Explorers	Nationality	Areas Explored

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
…	…	…

Exploration (Chronological order)	Who (explorer)
Sea route to India	Vasco da Gama
…	…

Segmented similarity matches modifier in first while ignoring extra terms in second

Collective labeling via Graphical Models

User Query

Name of Explorers	Nationality	Areas Explored

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

N_1 N_2 N_3

- Create a node for every column

N_4 N_5 N_6

Exploration (Chronological order)	Who (explorer)	Century
Sea route to India	Vasco da Gama	15th/16th
Caribbean	Christopher Columbus	15th/16th
Oceania	Abel Tasman	17th
...	...	...

N_7 N_8 N_9

Forest reserves		
ID	Name	Area
7	Shakespeare Hills	2236
9	Plains Creek	880
13	Welcome Swamp	168
...	...	...

Collective labeling via Graphical Models

User Query

Name of Explorers	Nationality	Areas Explored

- **Possible labels for every node**
  1. Name of explorers
  2. Nationality
  3. Areas Explored
  4. NA (Not Assigned)
  5. NR (Not Relevant)

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
…	…	…

N_1 N_2 N_3

Constraint: 0/1 match per query column

Constraint: all columns NR or none

N_4 N_5 N_6

Exploration (Chronological order)	Who (explorer)	Century
Sea route to India	Vasco da Gama	15th/16th
Caribbean	Christopher Columbus	15th/16th
Oceania	Abel Tasman	17th
…	…	…

N_7 N_8 N_9

Forest reserves		
ID	Name	Area
7	Shakespeare Hills	2236
9	Plains Creek	880
13	Welcome Swamp	168
…	…	…

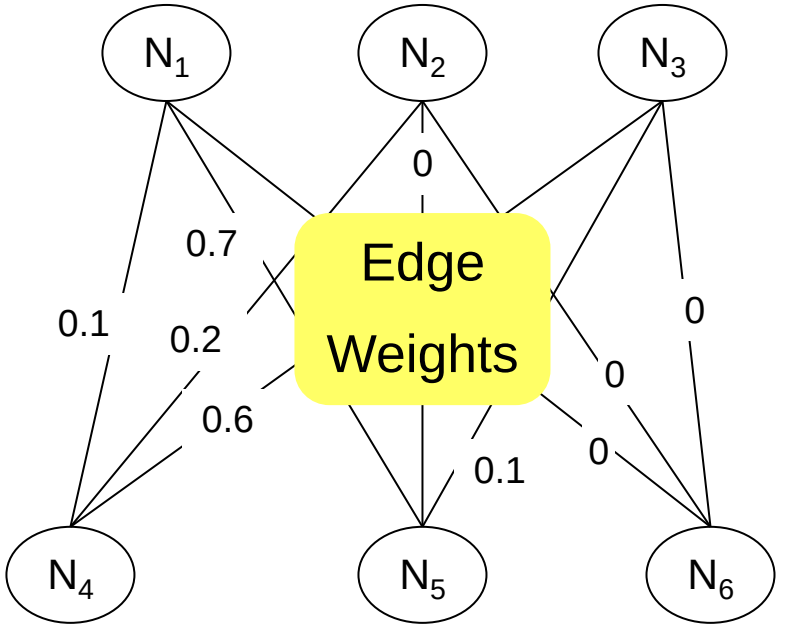
Graphical Model Approach

User Query

Name of Explorers	Nationality	Areas Explored

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

- **Edges**
  - Complete Bipartite Graph between nodes of two tables
  - Content overlap between column contents and headers
  - Maximum Bipartite Matching



Exploration (Chronological order)	Who (explorer)	Century
Sea route to India	Vasco da Gama	15th/16th
Caribbean	Christopher Columbus	15th/16th
Oceania	Abel Tasman	17th
...	...	...

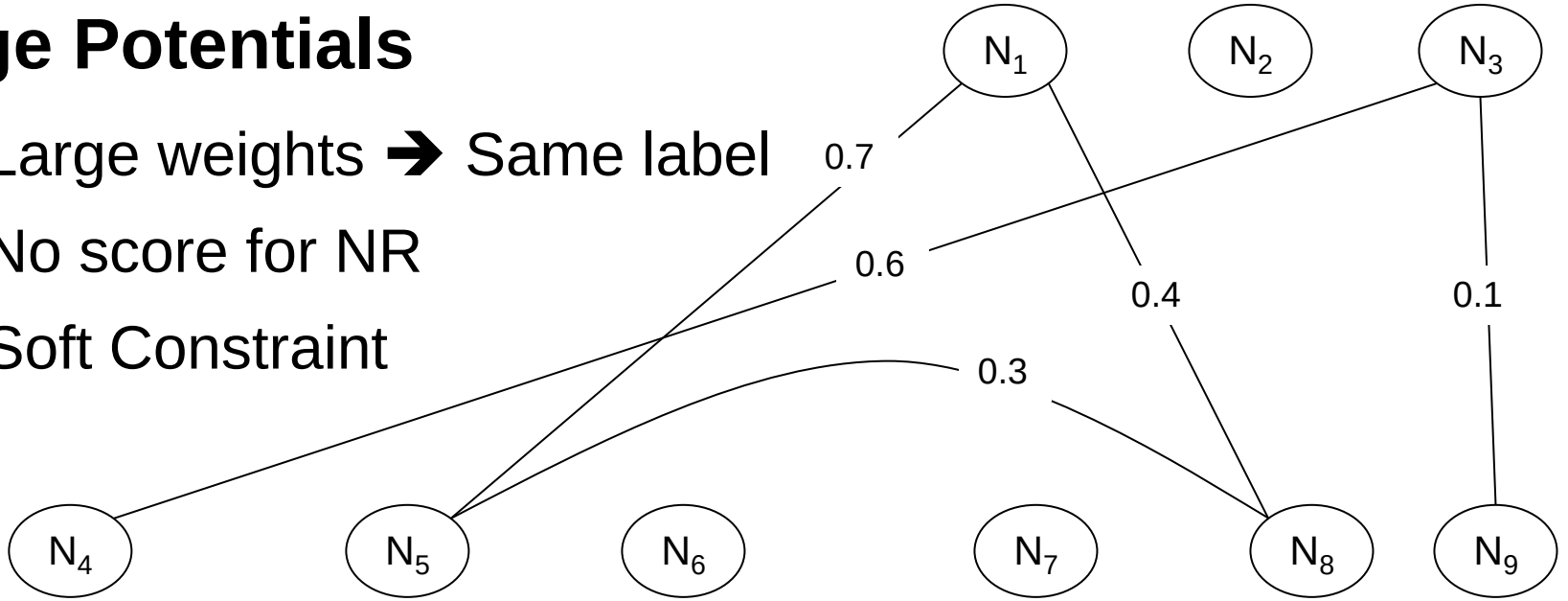
Graphical Model Approach

User Query

Name of Explorers	Nationality	Areas Explored

Name	Nationality	Main areas explored
Abel Tasman	Dutch	Oceania
Vasco da Gama	Portuguese	Sea route to India
Alexander Mackenzie	British	Canada
...	...	...

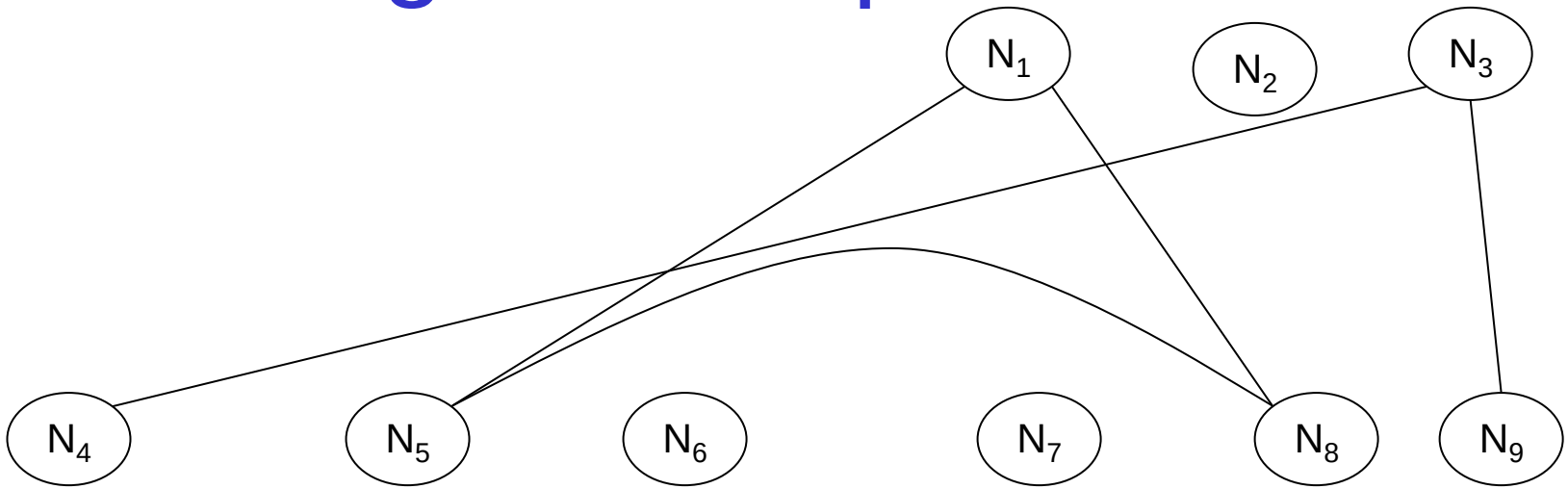
- **Edge Potentials**
  - Large weights ➔ Same label
  - No score for NR
  - Soft Constraint



Exploration (Chronological order)	Who (explorer)	Century
Sea route to India	Vasco da Gama	15th/16th
Caribbean	Christopher Columbus	15th/16th
Oceania	Abel Tasman	17th
...	...	...

Forest reserves		
ID	Name	Area
7	Shakespeare Hills	2236
9	Plains Creek	880
13	Welcome Swamp	168
...	...	...

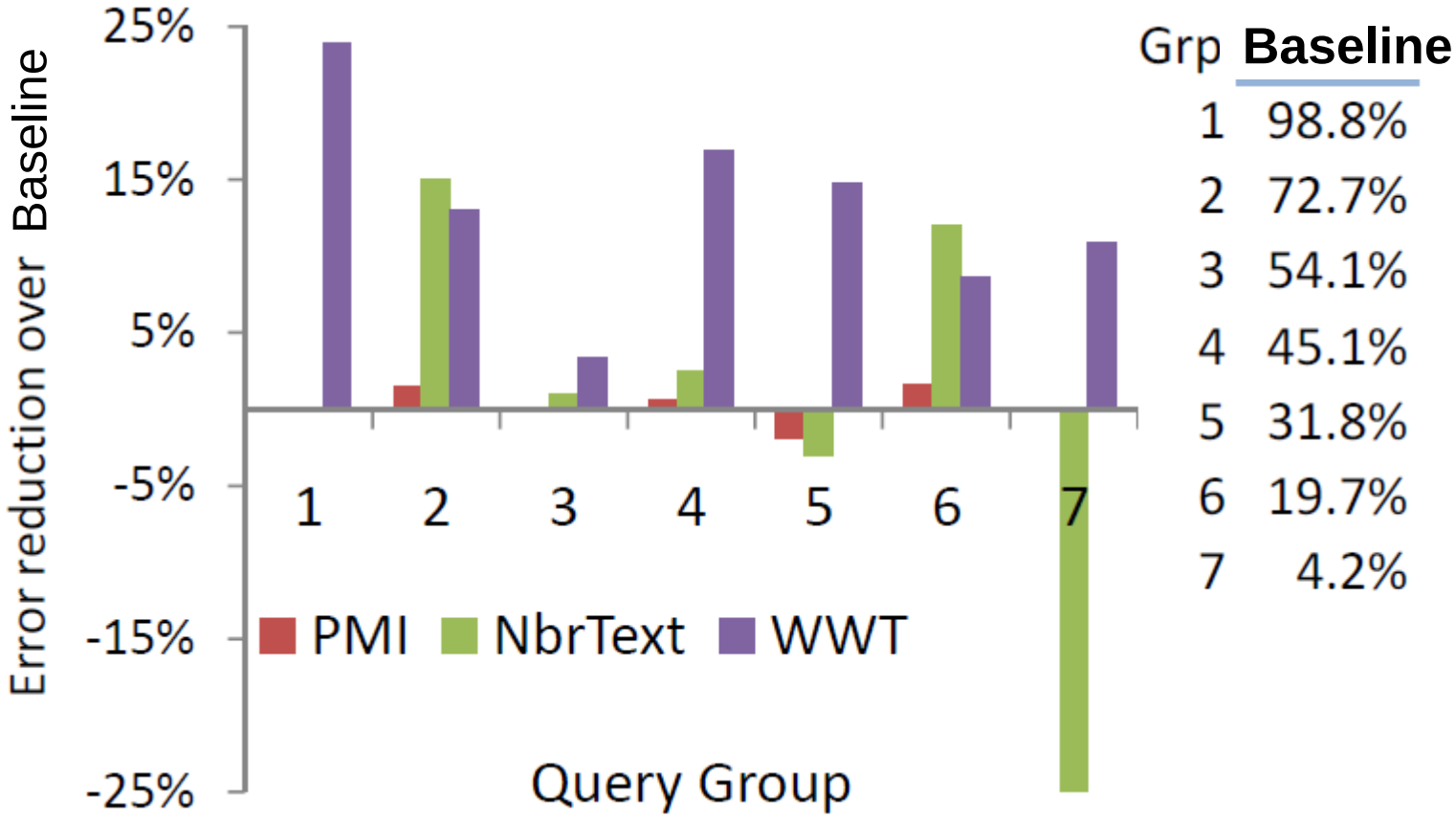
Labeling the Graphical Model



- **Jointly** assign one of |***Q***|+2 labels to each column to
  - maximize sum of node and edge potentials
  - satisfy the hard mutex constraint
    - At most one column per table assigned a query label
- NP-Hard: Modified α-Expansion Algorithm

Pimplikar and Sarawagi, VLDB 2012.

Column Mapping Methods Comparison



Grp	Baseline
1	98.8%
2	72.7%
3	54.1%
4	45.1%
5	31.8%
6	19.7%
7	4.2%

59 multi-column queries mostly collected from Amazon Mechanical Turk (AMT) service [Cafarella et al, 2009]

Enriching Web Tables

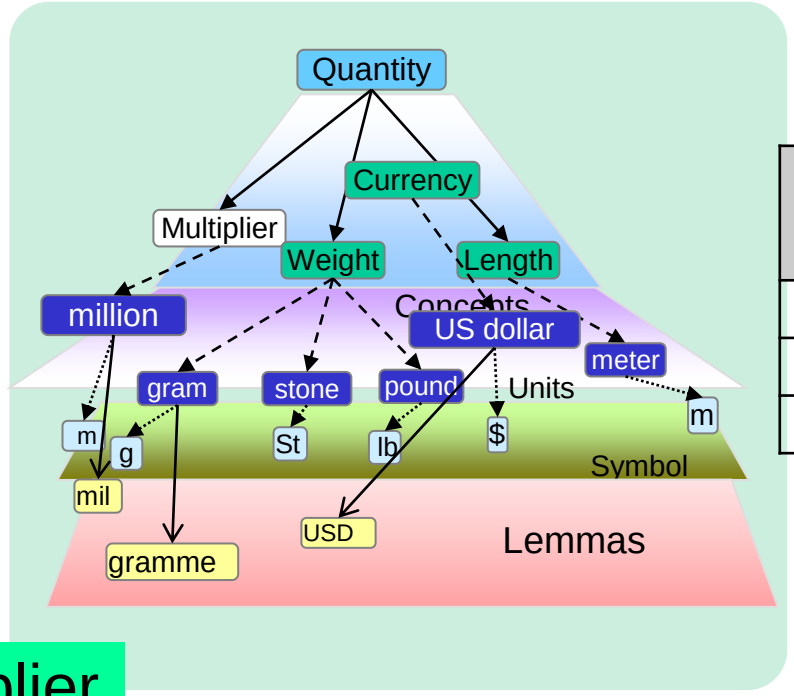
1. Entity,Type,Relationship annotation to an existing ontology
2. Query-time annotation for new types and relationships
3. Unit annotation for quantity columns

Unit Annotation

Unit list

Metre|foot

Pass	Elevation (m/ft)
Tonale pass	1884 (6181)
Colle Maniva	1669 (5476)
…	…



Ratio with new unit

Inh. /square kilometre

City	Density (inh. Per km2)
Macau	19796
Mumbai	20694
…	…

Atomic unit

Unit with multiplier

Year

British pound [million]

Year ended	Net profit/(loss) (£m)
2012	143
2011	40
…	…

Ratio unit

Mega joule/Kilogram

Storage	Energy density by mass in MJ/Kg
Liquid hydrogen	143
Energy from the sun	645,000,000
…	…

Simple rules do not work

- Word after “in” a unit e.g. Price in $, Length in km
  - Scores in last match → last is also a unit name.
  - Capacity in kt → kt = carat? Kiloton?
- Words within brackets is a unit e.g Price ($)
  - Population (Dec 2006), dec is a lemma for unit decade.
  - s: second, plural e.g. duration(s), year(s)
- Concept+unit-match e.g. Length m, Speed mph
  - Length (m:s)
    - m gets mapped to meter because of length. Actually, minute.
  - Energy density by volume (MJ/L)
    - → volume helps resolve L as liter.

... Rules not easy for compound units that need simultaneous labelling of different parts

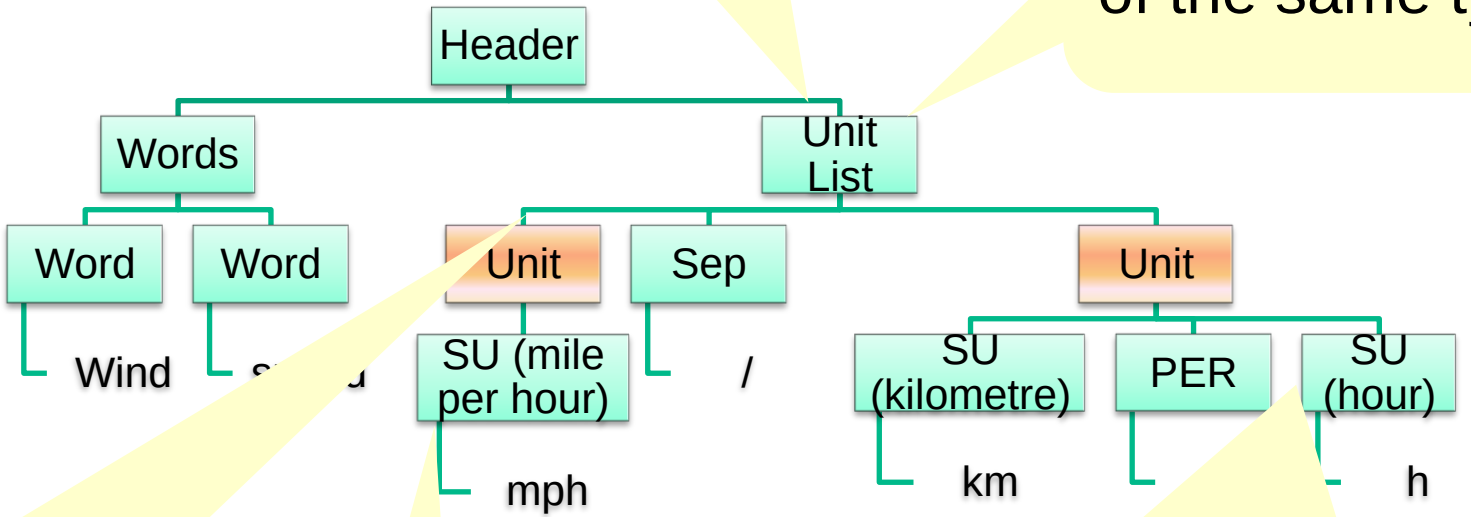
Our approach

- Probabilistic Context Free Grammar (QuantCFG)
  - Grammar: captures possible patterns of derived units
  - Diverse set of features from several resources
    - Wordnet frequencies
    - Co-occurrence statistics from unlabeled table corpus.

Header	::=	Words? Unit-List Words?
Unit	::=	CUnit | Multiplier Msep CUnit |
		CUnit Msep Multiplier | Msep Multiplier
Msep	::=	Empty | OF | IN
CUnit	::=	SimpleUnit | SimpleUnit UnitOp SimpleUnit
UnitOp	::=	Empty | PER | '/' | ×
SimpleUnit	::=	AtomUnit | Multiplier AtomUnit
AtomUnit	::=	TokenList_matching_QuTree | New_word
Multiplier	::=	Token_matching_QuTree_Multiplier | Number

An example parse

Wind speed (mph / km/h)



Brackets

Multiple units in a list tend to be of the same type

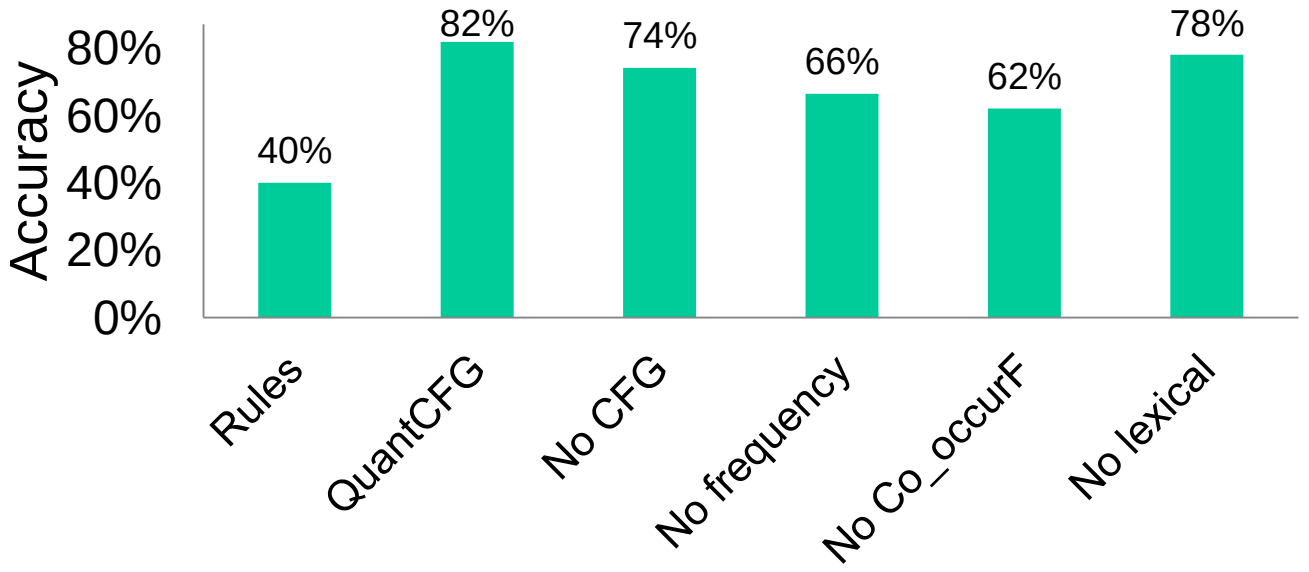
Co-occurrence of “mile per hour” with “wind”

Dictionary match

Relative frequency of h in the sense of “hour” in wordnet

Accuracy of QuantCFG

- Manually labeled 664 headers from Web tables



- Rules << QuantCFG
- Linear classifier with same features < QuantCFG
- Frequency feature very useful in NoUnit case
- Co-occurrence features somewhat useful

Summary

- Web tables: rich source of “long-tail” structured data, specifically for
  - Types: richer than “IsA” text patterns.
  - Quantities: more prevalent than in text (40%)
- This talk
  - Joint models for annotation to existing ontologies
    - Type-Entity scores to handle missing/multi-type links
  - Query-time type and relation annotation
    - Segmented similarity to query-type to column
    - Content overlap over many tables.
  - Unit annotation
    - PCFG + frequency + co-occurrence from unlabeled corpus

Thank you.