

WWT: A system for query-driven relation extraction from the semi- structured web








Sunita Sarawagi

Rahul Gupta Girija Limaye Prashant Borole
Rakesh Pimplikar Aditya Somani Soumen Chakrabarti

IIT Bombay

The semi-structured web

Table

Ouagadougou	 Burkina Faso
Pago Pago	 American Samoa
Palikir	 Federated States of Micronesia
Panama City	 Panama
Papeete	 French Polynesia
Paramaribo	 Suriname
Paris	 France

Regular page

Airports in Germany

Berlin-Tegel Airport
Berliner Flughafen-Gesellschaft mbH, Flughafen Tegel, 13405 Berlin, Germany
Airport Code: TXL

Cologne - Bonn Airport
Postfach 98 01 20, 51129 Cologne, Germany
Airport Code: CGN

Munich Airport
PO Box 23 17 55, 85326 Munich, Germany
Airport Code: MUC

Hamburg Airport
Flughafenstrasse 1-3, 22335 Hamburg, Germany
Airport Code: HAM

List

1. Gulf War oil spill, Persian Gulf, January 23 1991
2. Ixtoc oil well, SGulf of Mexico, June 3, 1979
3. Nowruz oil field, Persian Gulf, February, 1983
4. Atlantic Empress and Aegean Captain collision, off Trinidad and Tobago,
5. Castillo de Bellver, off Cape Town, South Africa, August 6, 1983
6. Amoco Cadiz (BP/Amoco, USA) - Brittany, France, March 16 1979
7. Torrey Canyon, South England, March 18 1967
8. Sea Star, Gulf of Oman, December 19, 1972
9. Urquiola, La Coruna, Spain, May 12, 1976
10. Hawaiian Patriot, N Pacific February 26, 1977
11. Othello, Tralhavet Bay, Sweden, March 20, 1970

Formatted list

- [Braer - Shetland Islands, January 5, 1993](#)
- [Prestige - Galicia, Spain, November 13, 2002](#)
- [Aegean Sea, off N Spain, December 3, 1992](#)
- [Sea Empress - Wales, February 15, 1996](#)
- [World Glory, off South Africa, June 13, 1968](#)
- [Corinthos Delaware River, Marcus Hook, Pennsylvania, January 31, 1975](#)
- [Burmah Agate Galveston Bay, Texas, November 1, 1979](#)
- [Exxon Valdez \(Exxon, USA\) - Prince William Sound, Alaska, March 24, 1989](#)
- [Keo, off MA, November 5, 1969](#)
- [Storage Tank, Sewaren NJ, November 4, 1969](#)
- [Ekofisk oil field, North Sea April 22, 1977](#)
- [Erika - Bay of Biscay, December 12, 1999](#)
- [Tasman Spirit, Karachi, Pakistan, July 28, 2003](#)

Queries in WWT

- Query by example

Alan Turing	Turing Machine
E. F. Codd	Relational Databases

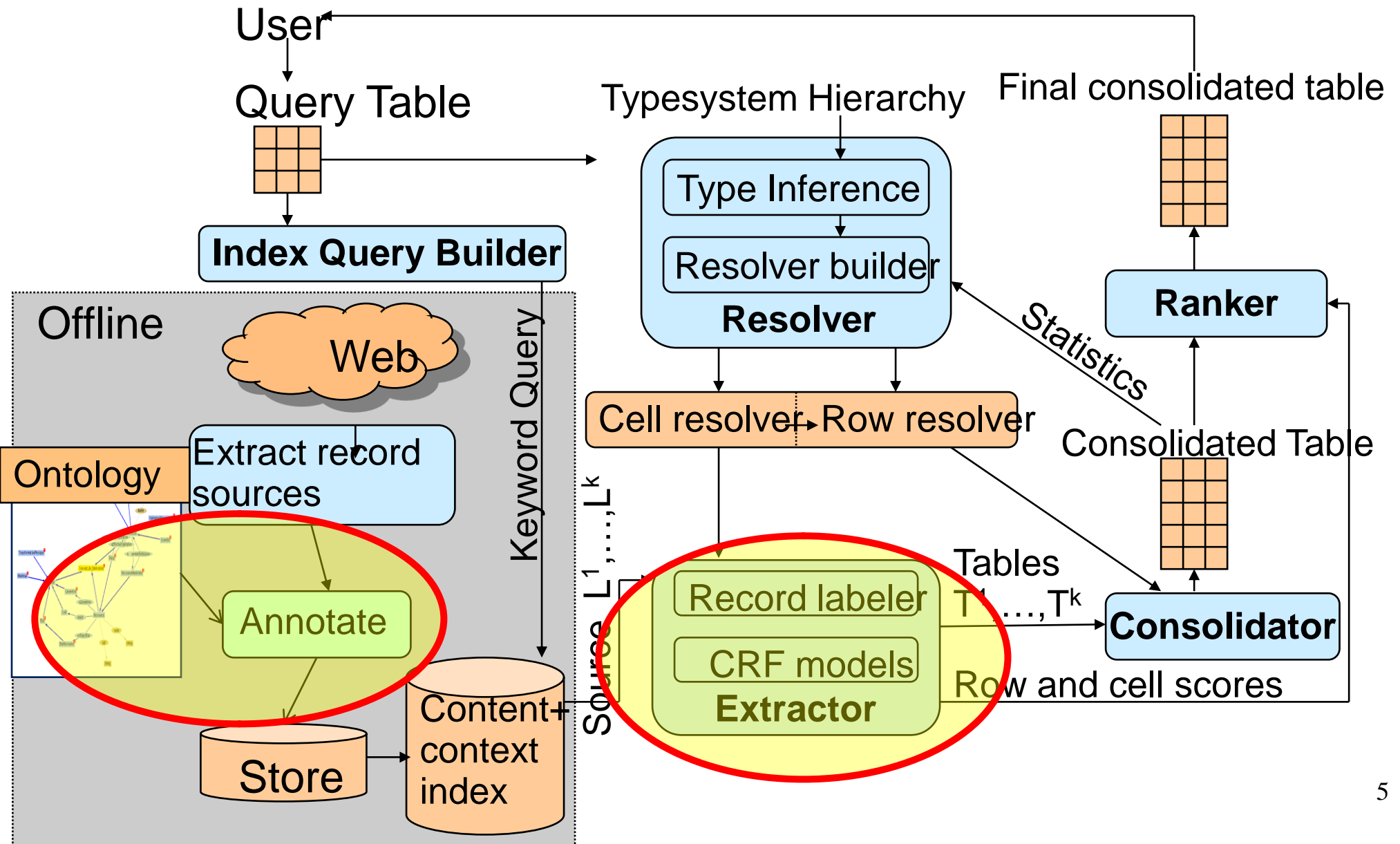
- Query by description

Inventor	Computer science concept
-----------------	---------------------------------

Answer: Table with ranked rows

Inventor	Computer Science Concept
Alan Turing	Turing Machine
Seymour Cray	Supercomputer
E. F. Codd	Relational Databases
Tim Berners-Lee	WWW
Charles Babbage	Babbage Engine

WWT Architecture



Query By Example

User



Gran Torino	Walt Kowalski	2008
Dirty Harry	Harry Callahan	1971

- 19. [Sudden Impact](#) (1983) [Harry Callahan](#)
- 20. [Honkytonk Man](#) (1982) [Red Stovall](#)
- 21. [Firefox](#) (1982) [Mitchell Gant](#)
- 22. [Any Which Way You Can](#) (1980) [Philo Beddoe](#)
- 23. [Bronco Billy](#) (1980) [Bronco Billy](#)
- 24. [Escape from Alcatraz](#) (1979) [Frank Morris](#)

- [The Dead Pool](#) (1988)
- [Heartbreak Ridge](#) (1983)
- [Pale Rider](#) (1985)
- [Tightrope](#) (1984)
- [City Heat](#) (1984)
- [Sudden Impact](#) (1983)

- [Joe Kidd \(1972\)](#) Joe Kidd
- [Dirty Harry \(1971\)](#) Inspector Harry Callahan
- [Play Misty for Me \(1971\)](#) Dave
- [The Beguiled \(1971\)](#) John McBurney
- [Kelly's Heroes \(1970\)](#) Kelly

Extract



Firefox	Mitchell Gant	1982
...
...

City Heat	-	1984
...	-	...
...	-	...

Joe Kidd	Joe Kidd	1972
...
...

Merge & de-duplicate, Rank, Display to the user

Query-time extraction

Two differences from classical IE

1. **Supervision: limited and indirect**
 - Need robust techniques for labeling unstructured lists from Q
 - Our solution:
 1. Similarity scores between a segment in a list and a cell in Q
 2. Segment list to maximize similarity (Gupta, VLDB 2009)
2. **Multiple sources with overlapping text segments**

Content Overlap



Segment Overlap: Partial, across varying number of sources

Options for exploiting overlap

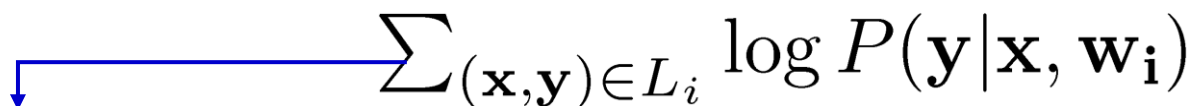
- Collective inference
 - Only helps instances with overlap
- Bootstrapping or staged training
 - Danger of error propagation or drift.

Collective Training

Goal

Input: S data sources, each source i has
Labeled records L_i , Unlabeled records U_i
Set $\mathcal{A} \equiv$ Cliques over all unlabeled records

Goal: Train CRF weights \mathbf{w}_i for each source $i = 1..S$

$$\max_{\{\mathbf{w}_1, \dots, \mathbf{w}_S\}} \sum_{i=1}^S \boxed{\text{LogLikelihood}(L_i | \mathbf{w}_i)} + \text{AgreementLikelihood}(\mathcal{A}, U_1, \dots, U_S | \mathbf{w}_1, \dots, \mathbf{w}_S)$$


Goal

Marginal prob that i^{th} model labels \mathcal{A} with $\mathbf{y}_{\mathcal{A}}$

$$\max_{\{\mathbf{w}_1, \dots, \mathbf{w}_S\}} \sum_{i=1}^S LL(L_i | \mathbf{w}_i) + C \cdot \log \underbrace{\sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i)}_{\substack{\uparrow \\ \text{Marginal prob that } i^{th} \text{ model labels } \mathcal{A} \text{ with } \mathbf{y}_{\mathcal{A}} \\ \downarrow \\ \text{Joint prob that all} \\ \text{models label } \mathcal{A} \text{ with } \mathbf{y}_{\mathcal{A}}}}$$

$$\log \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i)$$

Neither convex nor concave in the weights
Intractable because of exponential summations

Agreement Term = Log Partition

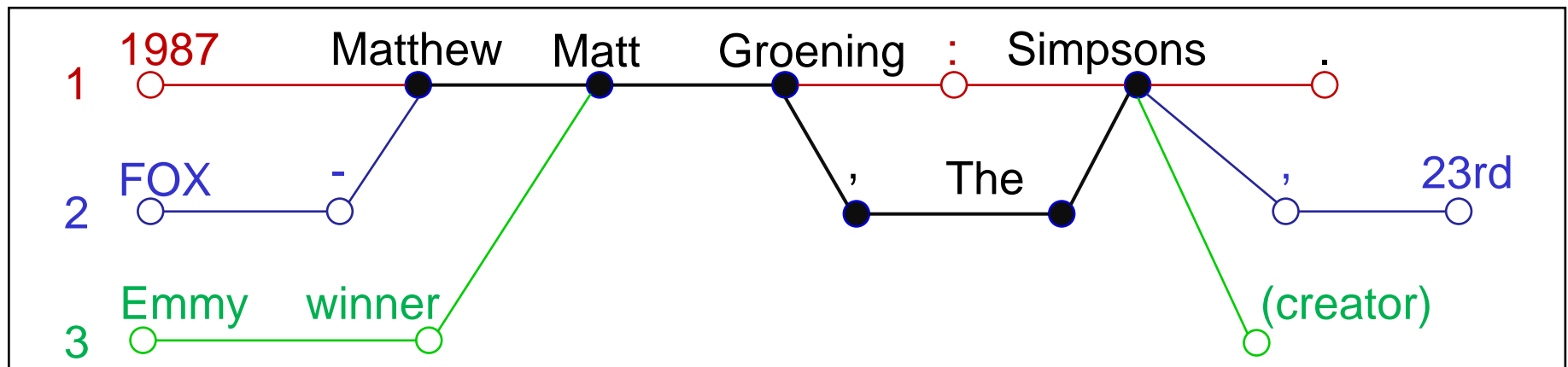
1987 Matthew “Matt” Groening : Simpsons .

FOX – Matthew “Matt” Groening , The Simpsons , 23rd

Emmy winner Matt Groening , The Simpsons (creator)

Four Cliques:

- Matthew “Matt” Groening (1,2),
- Matt Groening (1,2,3),
- Matt Groening , The Simpsons (2,3),
- Simpsons (1,2,3)



“Fused” Graph: Fuse shared segments, add potentials¹⁴

Agreement Term = Log Partition

$$\log \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i) = \log Z_{\text{fused}} - \sum_{i=1}^S \log Z_i$$

$$\nabla \log \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i) = E_{\text{fused}}[\mathbf{f}_1, \dots, \mathbf{f}_S] - \sum_{i=1}^S E_{p_i}[\mathbf{f}_i]$$

Approximating the Agreement Term

$$\log \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i) = \log Z_{\text{fused}} - \sum_{i=1}^S \log Z_i$$

- Fused graph can be arbitrarily complex
- Approximating the log-partition function
 - Belief propagation (BP)/TRW-S on the fused graph
 - Approximate BP tailored to chains ([Liang et.al. '09](#))

Two key problems

We get pseudo-marginals instead of marginals
A few wrong cliques => Completely different fused graph!

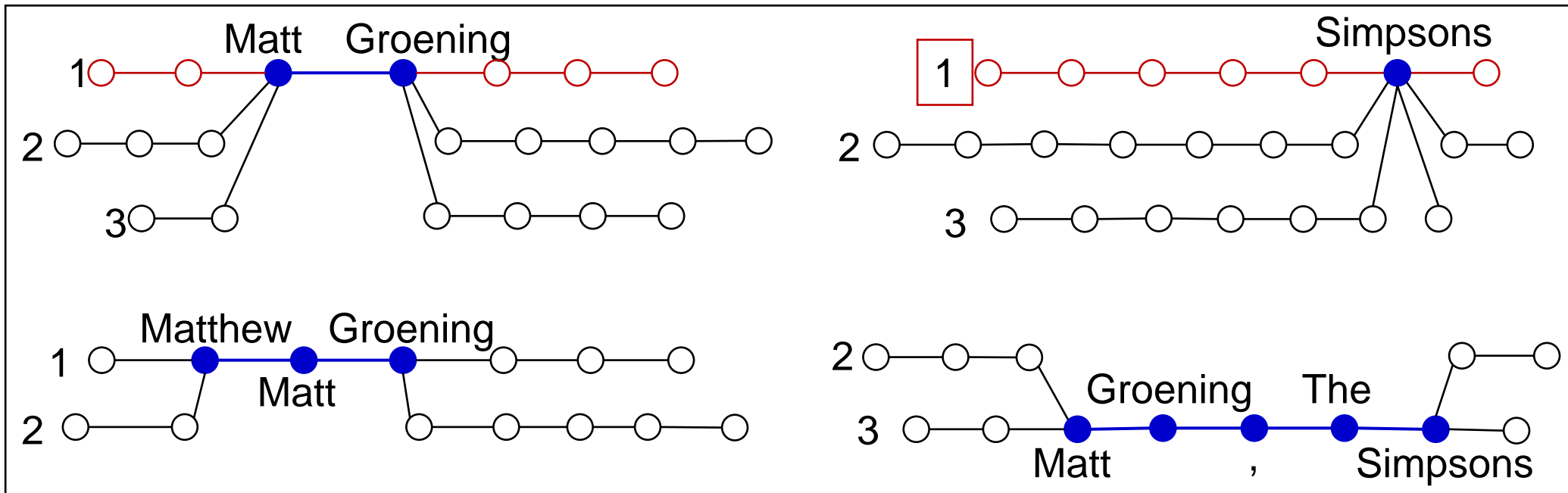
Decomposing the Agreement Term

Idea: Fused graph = Sum of small but easier graphs

$$\log \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i) \approx \sum_{T \in \mathcal{G}} \underbrace{\log \sum_{\mathbf{y}_{\mathcal{T}}} \prod_{i=1}^S p(\mathbf{y}_{\mathcal{T}} | \mathbf{w}_1, \dots, \mathbf{w}_S)}_{\log Z_{\text{fused}}(T) - \sum_{i \in T} \log Z_i}$$

T varies over low tree-width components that span G

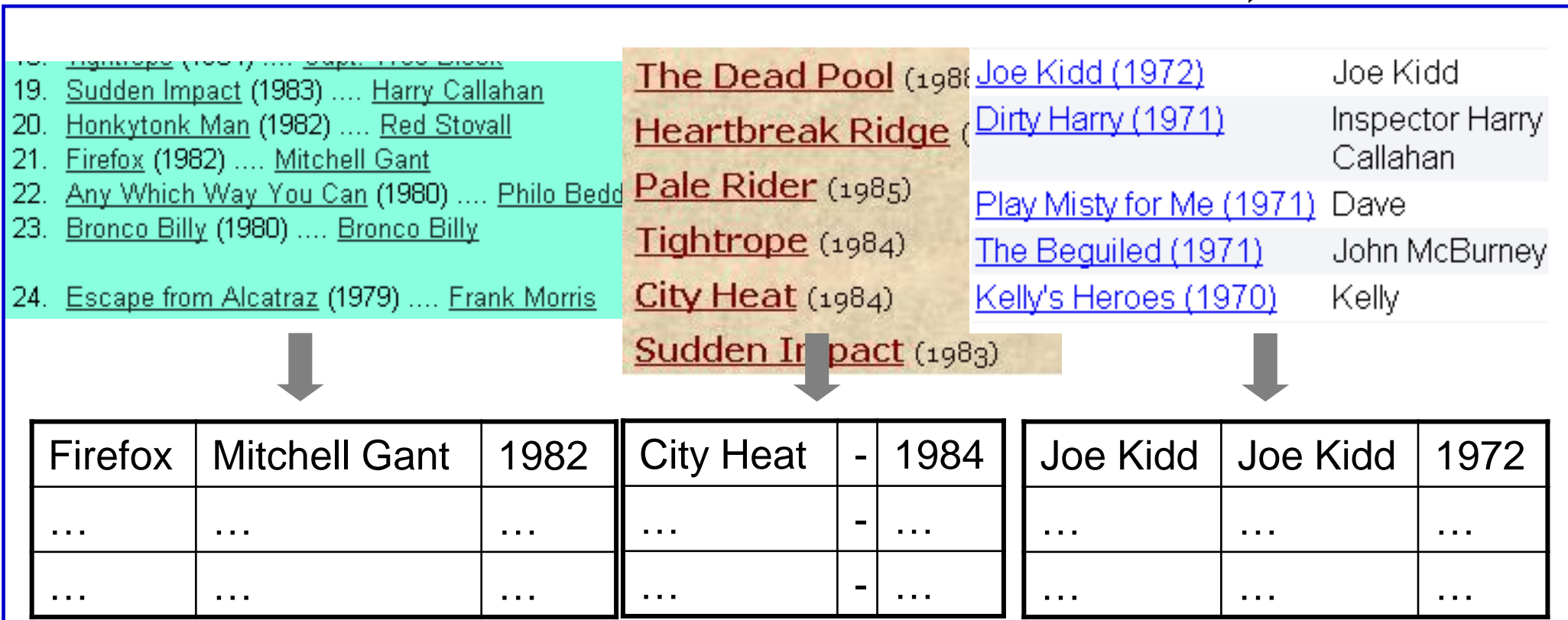
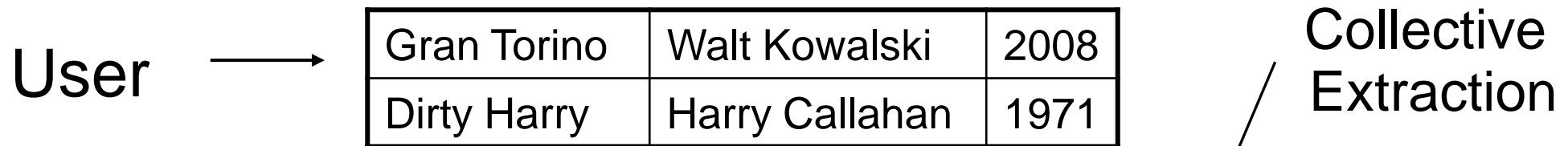
Clique-based Decomposition



$$\log \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i) \approx \sum_{c \in \mathcal{A}} \log \sum_{\mathbf{y}_c} \prod_{i=1}^S p_i(\mathbf{y}_c | \mathbf{w}_i)$$

- A component = One clique + Chains incident to the clique
- Each component is a tree, so exact logZ and gradient easy

Experiments: Structured Queries



Merge & de-duplicate, Rank, Display to the user

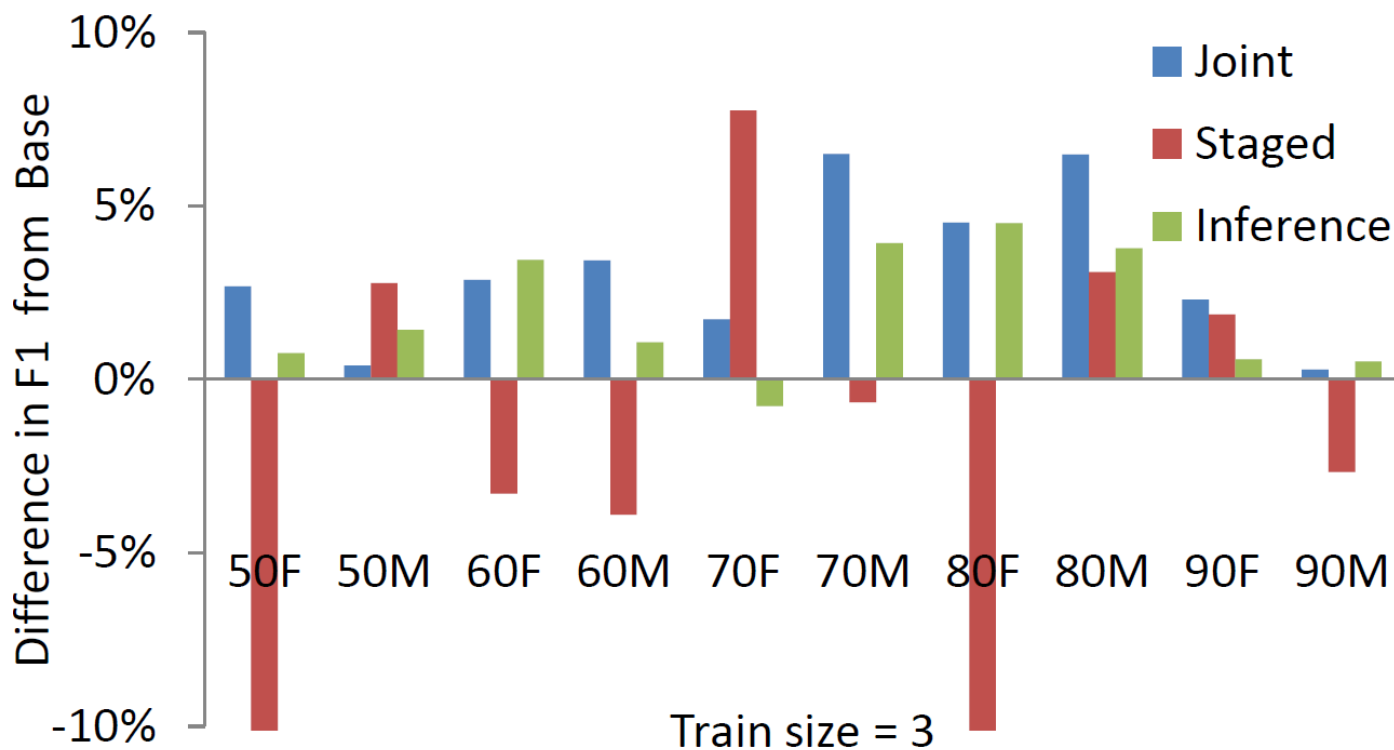
Experiments

- Aim: Reconstruct Wikipedia tables from only a few sample rows.
- Sample queries
 - TV Series: Character name, Actor name, Season
 - Oil spills: Tanker, Region, Time
 - Golden Globe Awards: Actor, Movie, Year
 - Dadasaheb Phalke Awards: Person, Year
 - Parrots: common name, scientific name, family

Experimental Setting

- IE on 58 datasets
 - Each dataset = 2-20 HTML list sources from a 500M crawl
 - Wide range of #labels, #sources, #records, #cliques, base accuracy, noise
 - Handful (~3) labeled records per list source
 - F1 measured using manually annotated ground truth
- Datasets binned by **Base model F1** and **Average Clique Incidence** for ease of presentation

Experiments: Other Frameworks



- Label transfer cascade-prone; 10% drop for some datasets
- Collective inference better, boosts 83.3% to 86.1%
- Joint training best, boosts to 87.5%
 - Even with 7 training records, boosts F1 from 87.4% to 89.2%

Experiments: Approximations

Data	Base	Agreement					PR EM
		Cliq	Nod	Pair	Full	TR1	
Train size = 3							
All	83.7	4.2	3.9	2.6	2.6	2.1	3.7
50F	55.2	2.7	3.5	2.9	2.9	3.5	1.0
50M	54.6	0.6	0.9	1.3	1.1	4.5	3.6
60F	66.9	2.9	2.6	0.8	0.6	1.5	1.5
60M	67.3	3.4	2.3	1.8	2.2	-0.1	3.4
70F	73.5	1.7	1.2	1.4	1.0	0.7	1.1
70M	76.1	6.5	5.8	3.8	4.5	3.7	6.9
80F	85.6	4.5	4.1	3.7	3.5	0.2	4.4
80M	86.6	6.5	6.0	3.8	3.4	3.6	4.5
90F	93.1	2.3	2.1	0.5	1.1	1.2	1.7
90M	96.1	0.3	0.6	-0.1	0.0	0.6	0.4

Decomposition methods >> approximation methods

- Exact gradient computation the key reason
- Clique decomposition also more tolerant to clique-noise

Collective Training: Take homes

- Collective training ideal for query-time extractions
 - Supervision limited
 - Redundancy and overlap abundant
- Challenges
 - Arbitrary overlap pattern → Intractable likelihoods
- Tractable decompositions >> approximations to intractable objectives

Queries in WWT

- Query by content

Alan Turing	Turing Machine
E. F. Codd	Relational Databases

- Query by description

Inventor	Computer science concept
-----------------	---------------------------------

Chemical Elements	Atomic Number
--------------------------	----------------------

Extraction: Description queries

Chemical Elements

Atomic Numbers

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

No headers

Lithium	3
Sodium	11
Beryllium	4

... This is a list of the **elements**, sorted by density measured at standard temperature and pressure.

Atomic Number Density (g/cmA³)
Description/Mohs' hardness ...

2

Atomic Number	Density (g/cmA ³)	Description/Mohs' hardness
1	0.00008988	gas
2	0.0001785	noble gas
10	0.0008999	noble gas
7	0.0012506	gas

Element	Symbol	Atomic We
Hydrogen	H	1.008
Helium	He	4.003
Lithium	Li	6.939
Carbon	C	12.011
Sodium	Na	22.990
Silicon	Si	28.090

... Element Symbol Atomic Weight

The table below gives some of the **atomic** weights of **chemical elements** accepted internationally ...

... A table of **chemical elements** ordered by **atomic number** and color coded according to type of **element**. Given is each **element's** name, **element** symbol, group and period, **Chemical** series, and **atomic** mass (or most stable isotope). ...

Accessing relevant tables

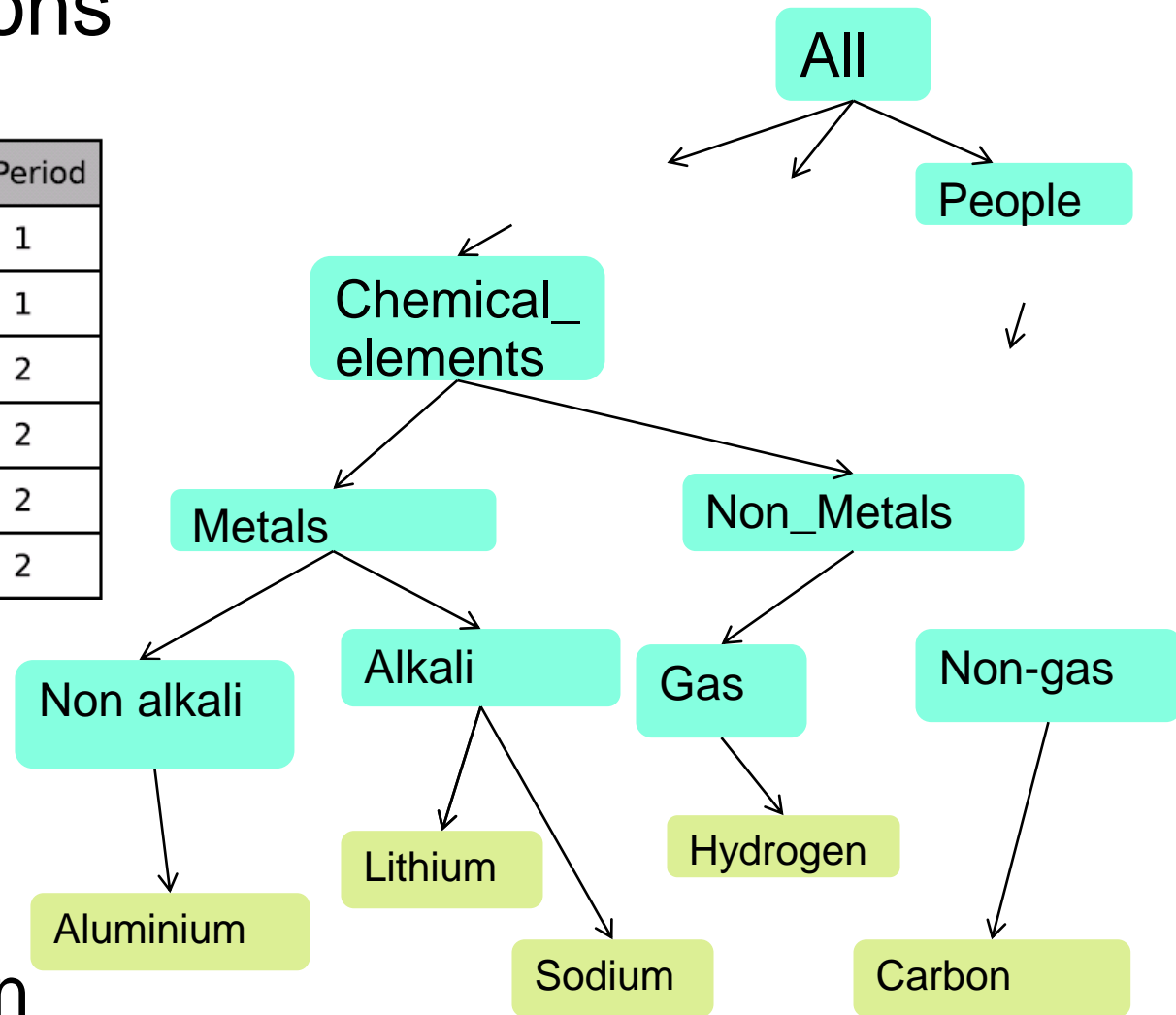
- Ontological annotations

Alkali	
Lithium	3
Sodium	11
Beryllium	4

Z	Chemical element	m	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

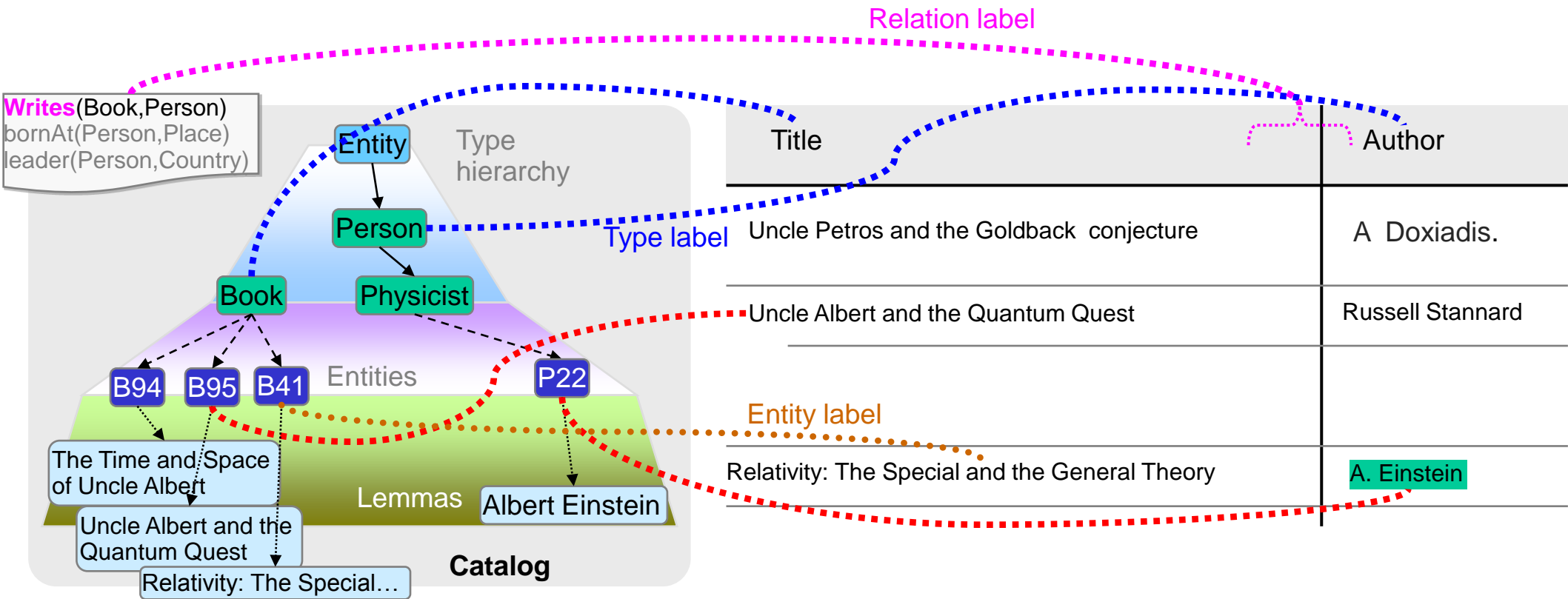
- Combine clues from

- Text around tables
- Headers
- Ontology labels when present



Annotating Tables with Ontological links

Entity, Type, and Relation links



Challenges

- Ambiguity of entity names
 - “Hydrogen” both a chemical element and a place name
- Noisy mentions of entity names
 - A. Einstein Vs Albert Einstein
- Multiple labels
 - YAGO Ontology has average 2.2 types per entity
- Missing type links in Ontology → cannot use least common ancestor
 - Universities in Toronto → Universities in Ontario.
 - Satyajit Ray → Indian film directors

Collective labeling via graphical models

- Variables

e_{rc} = Entity label in row r column c

t_c = Type label of column c

$b_{cc'}$ = Relation between columns c and c'

- Potentials

- Entity $\phi_1(r, c, e_{rc}) = \exp(\mathbf{w}_1^\top \mathbf{f}_1(r, c, e_{rc}))$.

- Similarity between cell (r, c) in table and lemmas of entity e_{rc} in catalog

- Type $\phi_2(c, t_c) = \exp(\mathbf{w}_2^\top \mathbf{f}_2(c, t_c))$

- Similarity between header & context of column c in table and lemmas of type t_c in catalog

- The specificity of a type t_c : $1/|\text{Entities in } t_c|$

Potentials (continued)

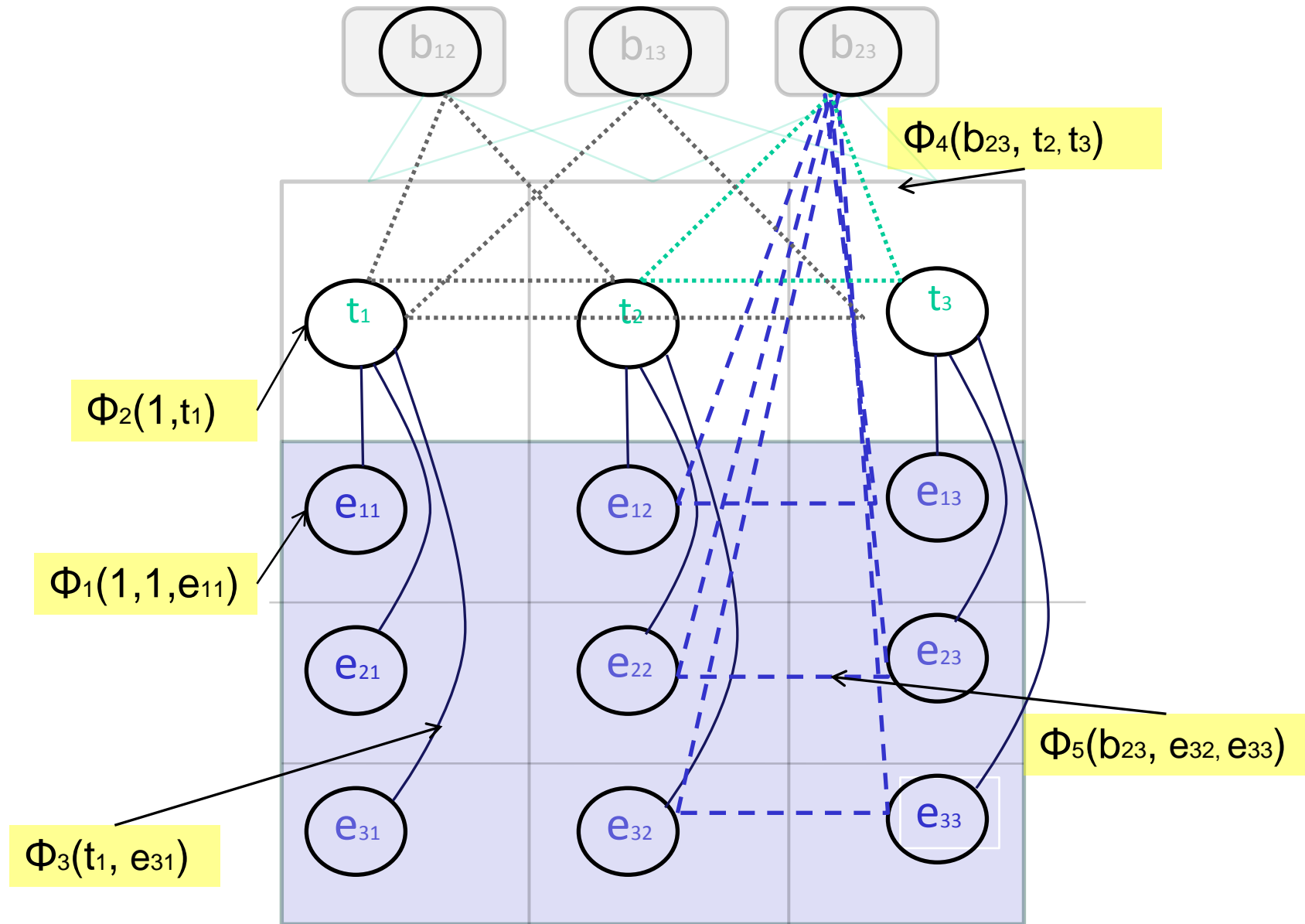
Entity-Type $\phi_3(t_c, e_{rc}) = \exp(\mathbf{w}_3^\top \mathbf{f}_3(t_c, e_{rc}))$

- 1▷ e_{rc} has path to t_c : (Einstein, Physicist)
 - Inverse distance between them
 - Penalizes over-generalization
- 2▷ e_{rc} has no path to t_c : (Julius Plucker, German Physicist)
 - Fraction of e_{rc} 's siblings with t_c
 - Julius Plucker is a Mathematician, many mathematicians are physicists
 - Newton is an English Physicist, overlap with German Physicists zero.
- 3▷ e_{rc} not in the catalog
 - Constant.
 - Allows NA label for columns with many unmatched entities

Potentials (continued)

- **Relation-Type-Type** $\phi_4(b_{cc'}, t_c, t_{c'}) = \exp(\mathbf{w}_4^\top \mathbf{f}_4(b_{cc'}, t_c, t_{c'}))$
 - Frequency of occurrence of this relationship in the catalog
- **Relation-Entity-Entity**
 $\phi_5(b_{cc'}, e_{rc}, e_{rc'}) = \exp(\mathbf{w}_5^\top \mathbf{f}_4(b_{cc'}, e_{rc}, e_{rc'}))$
 - 1 if this triple exists in the catalog
 - -1: if the catalog refutes this triple
 - 0: if the catalog is neutral

Inference



Exact: NP-hard, Belief propagation on factor graph

Accuracy of joint labeling

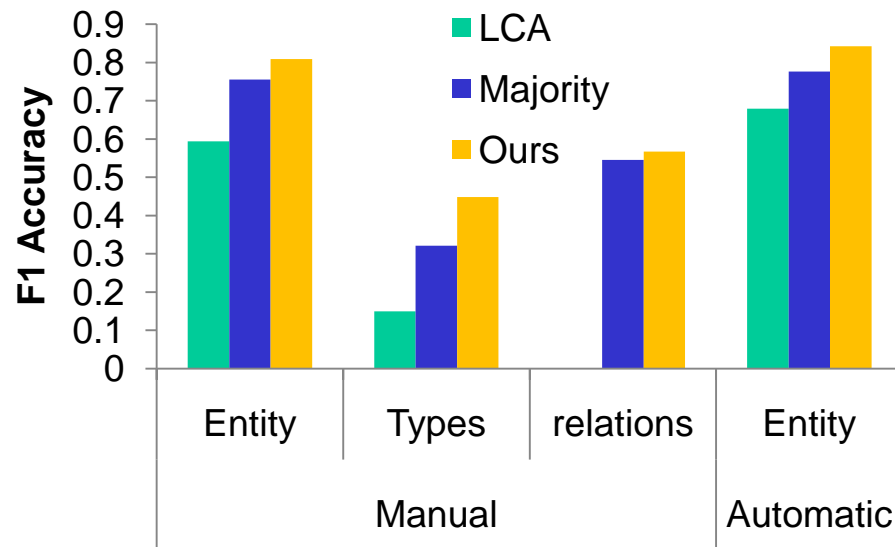
- Dataset

- Manually labeled

- 450 tables spanning general Web and Wikipedia

- Automatically labeled

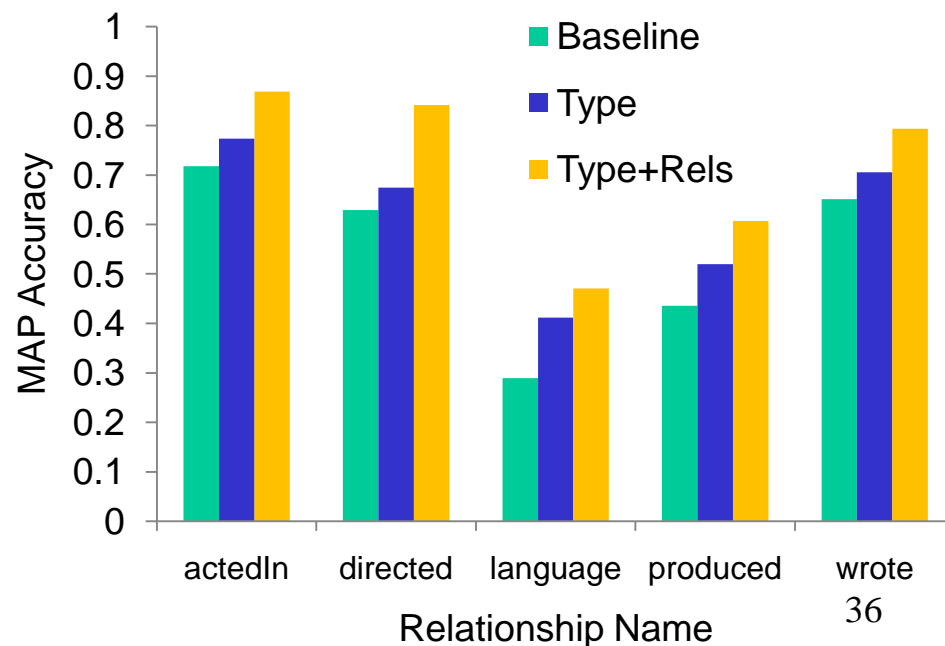
- 650 tables from Wikipedia where cells have entity links



Impact on query accuracy

Given inputs $R, T_1, T_2, E_2 \in^+ T_2$, return all $E_1 \in^+ T_1$ such that $R(E_1, E_2)$ holds.

- Movies: directed-by person="George Clooney"
- Countries: hasOfficial language="Spanish"
- Workload:
 - Five relations,
 - 40 queries per relation
- Ground truth: DBPedia



Summary

- Amazing amount of quality information on the semi-structured web.
- WWT
 - Online: structure interpretation at query time
 - Domain-independent methods for extraction and coreference resolution
 - Relies heavily on unsupervised statistical learning
 - Joint training for exploiting overlap during extraction
 - Graphical model for table annotation
 - Collective column labeling for descriptive queries
 - Bayesian network for consolidation
 - Page rank + confidence from a probabilistic extractor for ranking

Thank you.

Summary

- Enriched organically created web tables with precious schema information from reference catalogs
- Joint graphical model combines
 - Local text similarity
 - Global type and relationship consistency
- Better response to attribute-value queries
- Ongoing work:
 - Augmenting catalog with collective information from web tables.

What next?

- Designing plans for non-trivial ways of combining of sources
- Better ranking and user-interaction models.
- Expanding query set
 - Aggregate queries: tables are rich in quantities
 - Point queries: attribute value and relationship queries
- Interplay between semi-structured web & Ontologies
 - Augmenting one with the other.
- Quantify information in structured sources vis-à-vis text sources on typical query workloads

Collective Training: Related Work

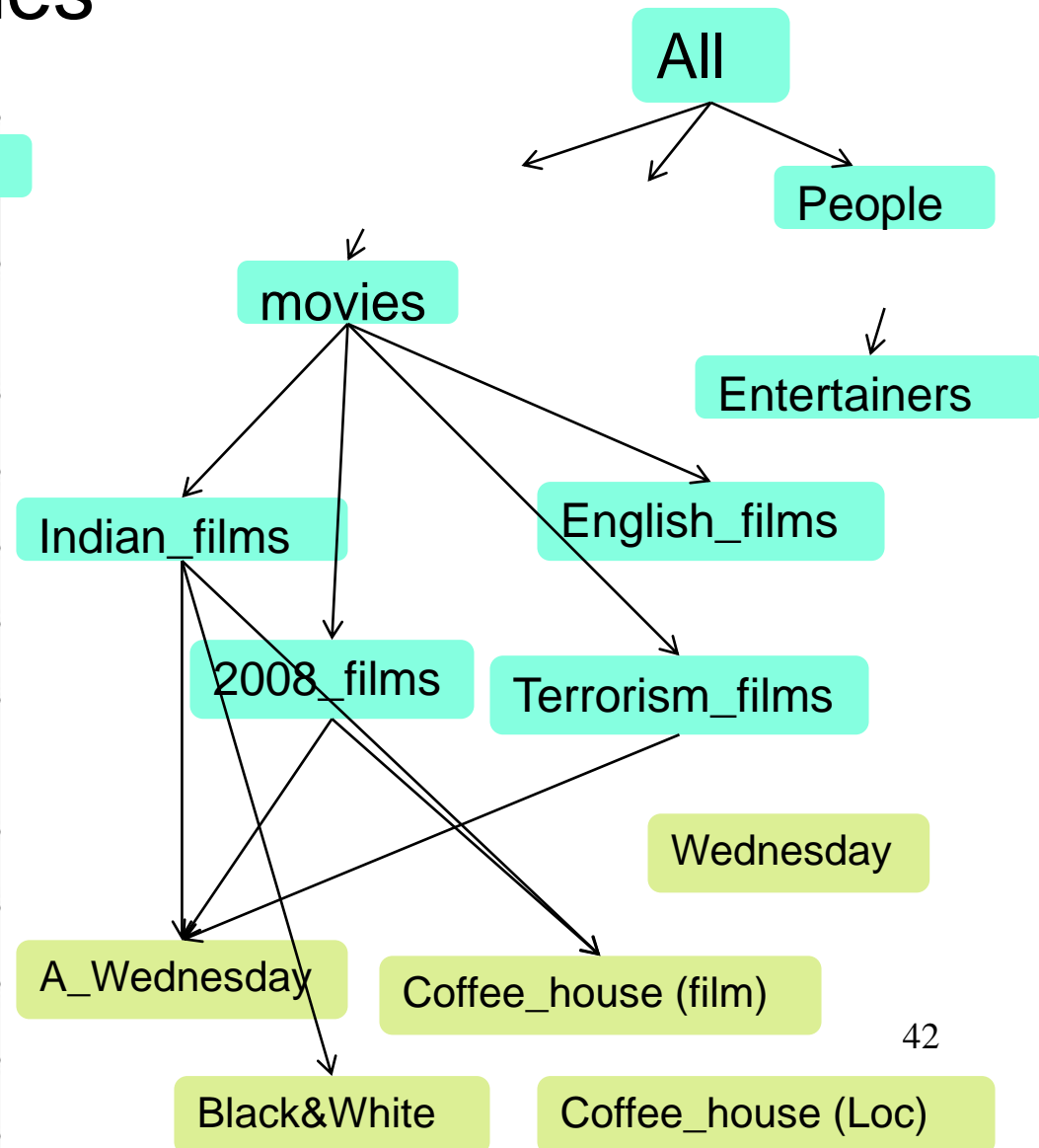
- Agreement based learning (Liang et.al. '09)
 - EM-based scheme applied on two sources with clean cliques
- Posterior Regularization (Ganchev et.al. '08)
 - Different agreement term; Used in multi-view
- Two-view {perceptron, regression}, co-{training, boosting, SVMs} (Brefeld et.al. '05, Blum & Mitchell '98, Collins & Singer '99, Sindhwani et.al. '05, Kakade & Foster '07)
 - Two source and/or hard label transfer
- Multi-task learning (Ando & Zhang '05)
 - Single source, shared features sought
- Semi-supervised learning (Chapelle et.al. '06)
 - No training, no support for partially structured overlaps
- Co-regularization, Pooling (Suzuki et.al. '07)

Annotating to an Ontology

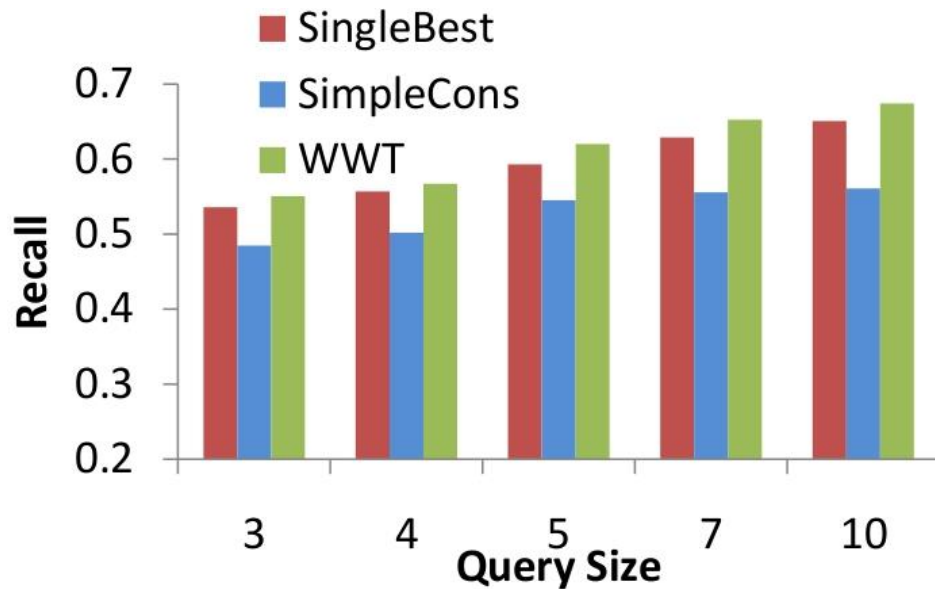
Annotate table cells with entity nodes and table columns with type nodes

2008

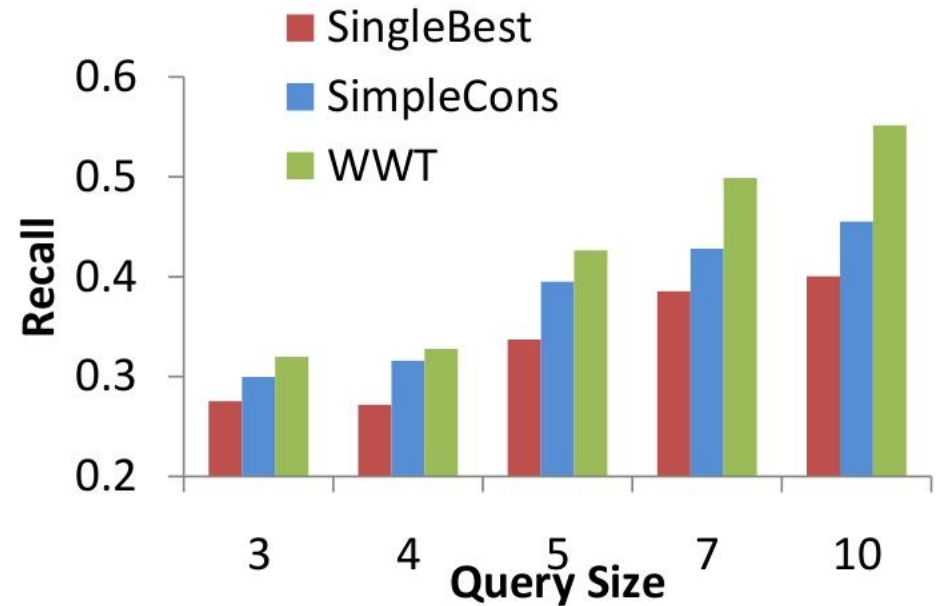
Indian_films		Indian_directors
2008_films	Title	Director
	1920	Vikram Bhatt
	Anamika	Anant Mahadevan
	Aamir	Rajkumar Gupta
A_Wednesday	day	Neeraj Pandey
	Bachna Ae Haseeno	Siddharth Anand
	Bhootnath	Vivek Sharma
Black&White	White	Subhash Ghai
	Bombay to Bangkok	Nagesh Kukunoor
	Chamku	Kabeer Kaushik
Coffee_house (film)		Gurbeer Garewal



Overall performance



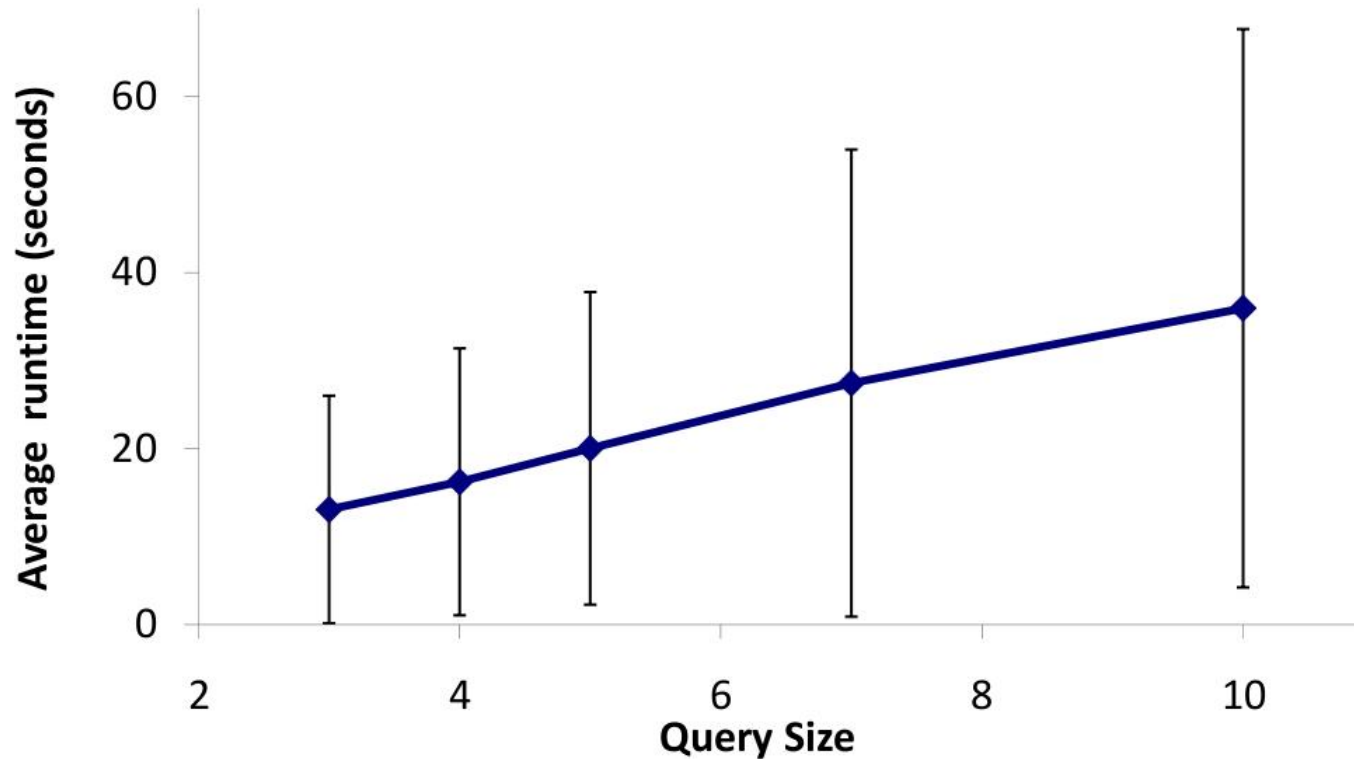
All Queries



Difficult Queries

- Justify sophisticated consolidation and resolution. So compare with:
 - Processing only the **magically known** single best list
=> no consolidation/resolution required.
 - Simple consolidation. No merging of approximate duplicates.
- WWT has > 55% recall, beats others. Gain bigger for difficult queries.

Running time



- < 30 seconds with 3 query records.
 - Can be improved by processing sources in parallel.
- Variance high because time depends on number of columns, record length etc.

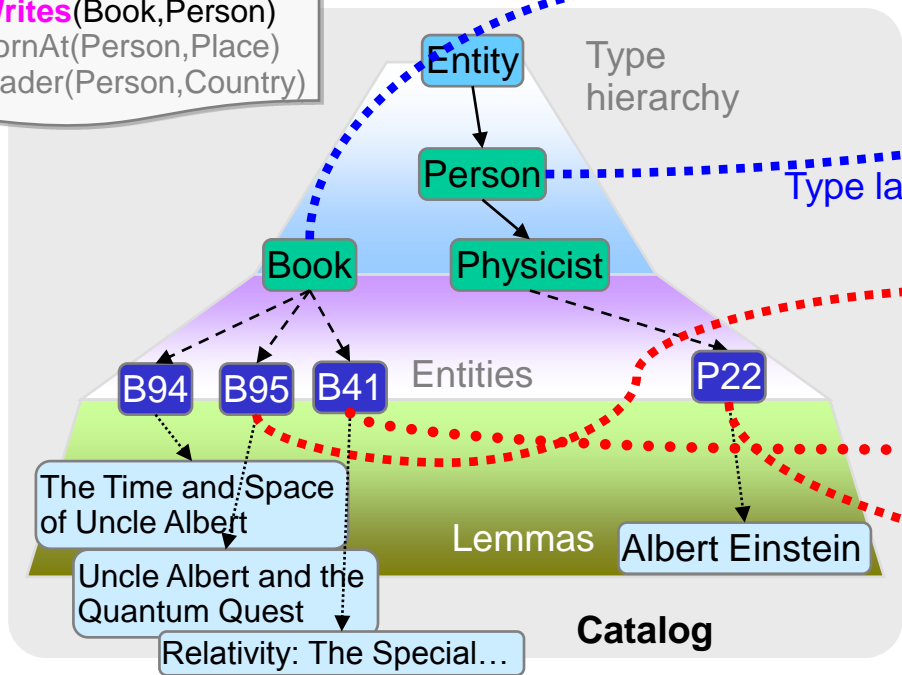
Re-writing the Agreement Term

$$\begin{aligned} & \sum_{\mathbf{y}_{\mathcal{A}}} \prod_{i=1}^S p_i(\mathbf{y}_{\mathcal{A}} | \mathbf{w}_i) \\ &= \sum_{\mathbf{y}_{\mathcal{A}}} \text{Prob that all models label } \mathcal{A} \text{ with } \mathbf{y}_{\mathcal{A}} \\ &= \sum_{\mathbf{y}_1, \dots, \mathbf{y}_S} \underbrace{\prod_i p_i(\mathbf{y}_i | \mathbf{w}_i)}_{\text{Zero if } \mathbf{y}_1, \dots, \mathbf{y}_S \text{ disagree}} \cdot \mathbf{1}[\mathbf{y}_1, \dots, \mathbf{y}_S \text{ agree on } \mathcal{A}] \\ &= \sum_{\mathbf{y}_1, \dots, \mathbf{y}_S} \text{function}(\mathbf{y}_1, \dots, \mathbf{y}_S) \end{aligned}$$

Labeling of chain i

Zero if $\mathbf{y}_1, \dots, \mathbf{y}_S$ disagree

Writes(Book,Person)
 bornAt(Person,Place)
 leader(Person,Country)



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

Relation label

Type label

Entity label

Catalog