

Building and Leveraging Knowledge Bases for Science

Andrew McCallum

Center for Data Science
College of Information and Computer Sciences
University of Massachusetts Amherst



Joint work with

Sebastian Riedel, Limin Yao, Arvind Neelakantan, Patrick Verga, Rajarshi Das.

Web page search



lebron james height



Web

Images

Videos

News

Shopping

More ▾

Search tools

[Insane vertical leap by **Lebron James**. Look at how far up he jumps ...](#)

www.youtube.com/watch?v=F1-YcD5pQXQ ▾ YouTube ▾

Jan 11, 2010 - **Lebron James** jumps with one leg. Look at that **height!!!** Sick. Come on Mr. James - 2011 Slam Dunk Contest!

[LeBron James - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/LeBron_James ▾ Wikipedia ▾

LeBron James vs Washington 3-30-11.jpg ... Listed **height**, 6 ft 8 in (203 cm) ... LeBron Raymone James (/ləˈbrɒn/; born December 30, 1984) is an American ...

[List of career achievements by ... - St. Vincent–St. Mary High School - Akron, Ohio](#)

[LeBron James Stats, Video, Bio, Profile | NBA.com](#)

www.nba.com/playerfile/lebron_james/ ▾ National Basketball Association ▾

Find a complete bio, stats and videos about **LeBron James**, Forward for the Miami Heat. Stay up to date ... **LeBron James**. NBA.com/Stats **Height: 6'8" / 2.03 m.**

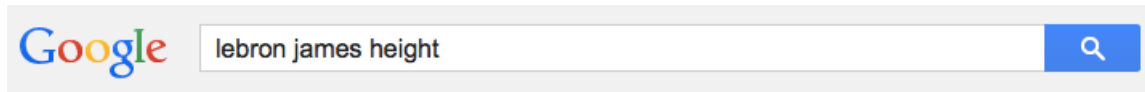
[How **LeBron James'** life changed in fourth grade - ESPN The ...](#)

espn.go.com/.../how-lebron-james-life-changed-fourth-grade-espn-... ▾ ESPN ▾

Oct 17, 2013 - ... of **LeBron James** the fourth grader, before basketball came into his life. ... to her , he saw LeBron, lean and lanky, already as **tall** as his mother, ...

~~Web page search~~ → dialog, QA, KB

Structured knowledge of world.



Web Images Videos News Shopping More Search tools

[Insane vertical leap by LeBron James. Look at how far up he jumps ...](#)

www.youtube.com/watch?v=F1-YcD5pQXQ YouTube

Jan 11, 2010 - **LeBron James** jumps with one leg. Look at that **height!!!** Sick. Come on Mr. James - 2011 Slam Dunk Contest!

[LeBron James - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/LeBron_James Wikipedia

LeBron James vs Washington 3-30-11.jpg ... Listed **height**, 6 ft 8 in (203 cm) ... LeBron Raymone James (/ləˈbrɒn/; born December 30, 1984) is an American ...

List of career achievements by ... - St. Vincent-St. Mary High School - Akron, Ohio

[LeBron James Stats, Video, Bio, Profile | NBA.com](#)

www.nba.com/playerfile/lebron_james/ National Basketball Association

Find a complete bio, stats and videos about **LeBron James**, Forward for the Miami Heat. Stay up to date ... **LeBron James**. NBA.com/Stats **Height**: 6'8"/ 2.03 m.

[How LeBron James' life changed in fourth grade - ESPN The ...](#)

espn.go.com/.../how-lebron-james-life-changed-fourth-grade-espn-... ESPN

Oct 17, 2013 - ... of **LeBron James** the fourth grader, before basketball came into his life. ... to her , he saw LeBron, lean and lanky, already as **tall** as his mother, ...

Freebase

Jeff Bezos

Entrepreneur

Jeffrey Preston "Jeff" Bezos is an American Internet entrepreneur and investor.

He is a technology entrepreneur who has played a key role in the growth of e-commerce as the founder and CEO of Amazon.com, ... [Wikipedia](#)

Born: January 12, 1964 (age 50),
Albuquerque, NM

Nationality: American

Spouse: Mackenzie Bezos (m. 1993)

Parents: Ted Jorgensen, Jacklyn Bezos,
Miguel Bezos

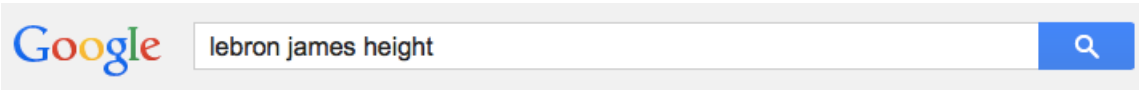
Education: Princeton University (1986),
River Oaks Elementary School, Miami
Palmetto High School



~~Web page search~~ → dialog, QA, KB

Structured knowledge of world.

Freebase



Web Images Videos News Shopping More Search tools

[Insane vertical leap by LeBron James. Look at how far up he jumps ...](#)

[www.youtube.com/watch?v=F1-YcD5pQXQ](#) YouTube

Jan 11, 2010 - **LeBron James** jumps with one leg. Look at that **height!!!** Sick. Come on Mr. James - 2011 Slam Dunk Contest!

[LeBron James - Wikipedia, the free encyclopedia](#)

[en.wikipedia.org/wiki/LeBron_James](#) Wikipedia

LeBron James vs Washington 3-30-11.jpg ... Listed **height**, 6 ft 8 in (203 cm) ... LeBron Raymone James (/ləˈbrɒn/; born December 30, 1984) is an American ...

List of career achievements by ... - St. Vincent-St. Mary High School - Akron, Ohio

[LeBron James Stats, Video, Bio, Profile | NBA.com](#)

[www.nba.com/playerfile/lebron_james/](#) National Basketball Association

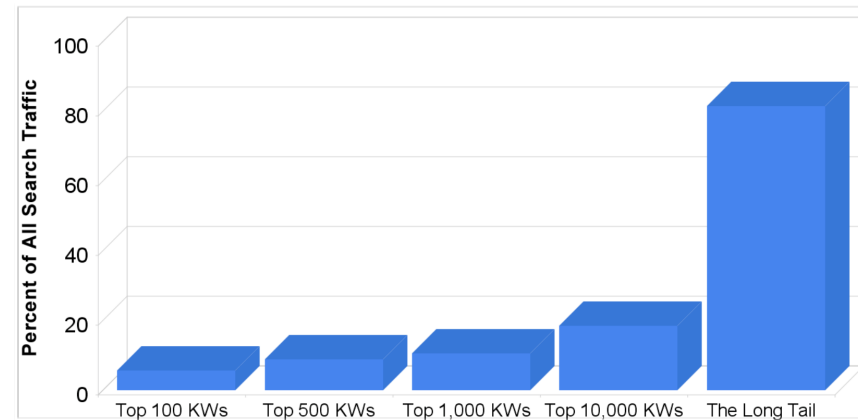
Find a complete bio, stats and videos about **LeBron James**, Forward for the Miami Heat. Stay up to date ... **LeBron James**. NBA.com/Stats **Height**: 6'8"/ 2.03 m.

[How LeBron James' life changed in fourth grade - ESPN The ...](#)

[espn.go.com/.../how-lebron-james-life-changed-fourth-grade-espn-...](#) ESPN

Oct 17, 2013 - ... of **LeBron James** the fourth grader, before basketball ... life. ... to her , he saw LeBron, lean and lanky, already

"open schema"
KB (of Science!)
with entity-relation structure?
...and reasoning?



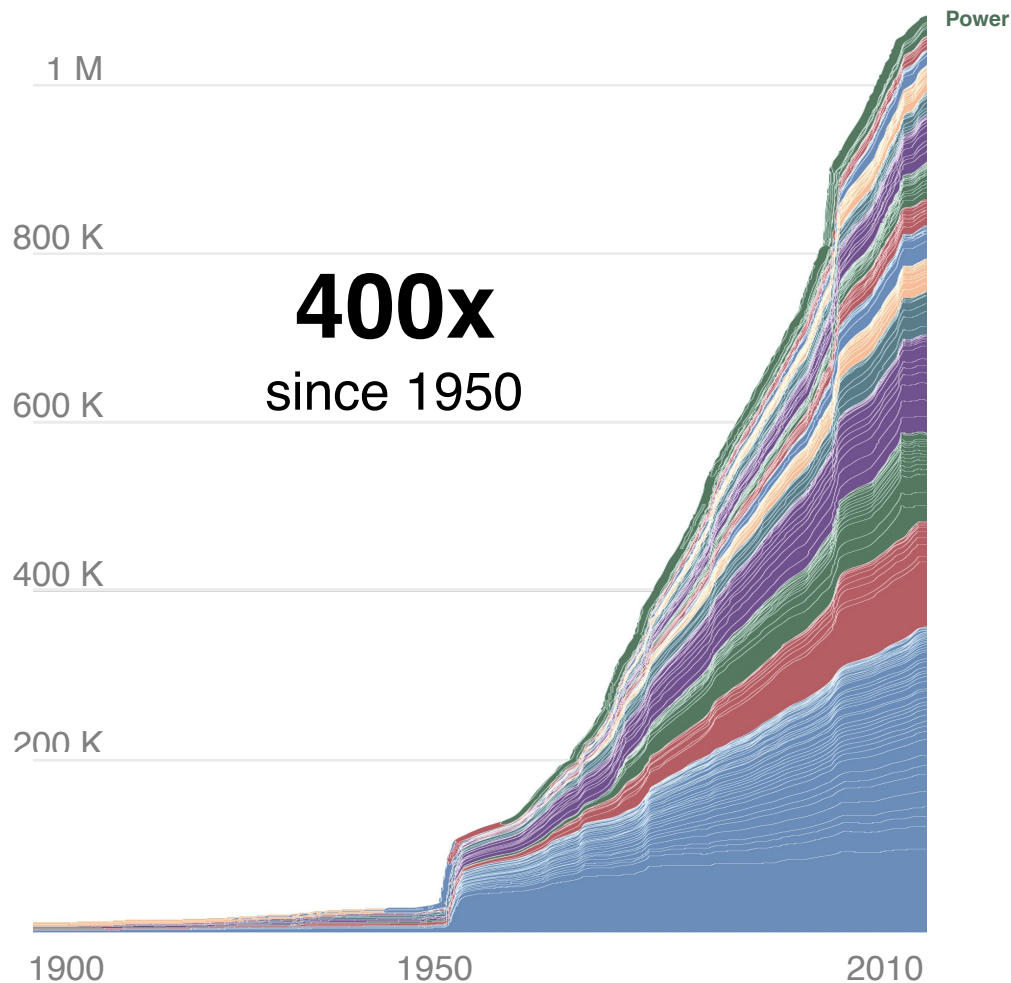
"Who did Bezos criticize?"

Spouse: Mackenzie Bezos (m. 1993)

Parents: Ted Jorgensen, Jacklyn Bezos, Miguel Bezos

Education: Princeton University (1986), River Oaks Elementary School, Miami Palmetto High School

Scientific Literature Growth



Graph: ReLX Group

26,560,336
papers since 1811

4,000+
new every day

Text



Knowledge Base



Reasoning

**Scientific
Text**



**Scientific
Knowledge Base**



**Scientific
Reasoning**

Cora: KB of Research Papers

[McCallum et al 1996]

Reinforcement Learning: A Survey

Leslie Pack Kaelbling

Michael L. Littman

*Computer Science Department, Box 1910, Brown University
Providence, RI 02912-1910 USA*

Andrew W. Moore

*Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue
Pittsburgh, PA 15213 USA*

LPK@CS.BROW

MLITTMAN@CS.BROW

AWM@CS.CM

Abstract

This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad selection of current work are summarized. Reinforcement learning is the problem faced by an agent that learns behavior through trial-and-error interactions with a dynamic environment. The work described here has a resemblance to work in psychology, but differs considerably in the details and in the use of the word "reinforcement." The paper discusses central issues of reinforcement learning including trading off exploration and exploitation, establishing the foundations of the field via Markov decision theory, learning from delayed reinforcement, constructing empirical models to accelerate learning, making use of generalization and hierarchy, and coping with hidden state. It concludes with a survey of some implemented systems and an assessment of the practical utility of current methods for reinforcement learning.

1. Introduction

Reinforcement learning dates back to the early days of cybernetics and work in psychology, neuroscience, and computer science. In the last five to ten years, it has attracted rapidly increasing interest in the machine learning and artificial intelligence communities. Its promise is beguiling—a way of programming agents by reward and punishment without needing to specify *how* the task is to be achieved. But there are formidable computational obstacles to fulfilling the promise.

This paper surveys the historical basis of reinforcement learning and some of the current work from a computer science perspective. We give a high-level overview of the field and taste of some specific approaches. It is, of course, impossible to mention all of the important work in the field; this should not be taken to be an exhaustive account.

Netscape: Cora Research Paper Search

File Edit View Go Communicator

Bookmarks Location: http://www.cora.justresearch.com/cgi-bin/cora_query.

author:boyan "search engines" Search Help

Title, author, institution and abstract are automatically extracted, and are often, but not always correct.

Number of hits found: 64

1. A Machine Learning Architecture for Optimizing Web Search Engines
Justin Boyan, Dayne Freitag, and Thorsten Joachims
Abstract: Indexing systems for the World Wide Web, such as Lycos and Alta Vista, play an essential role in making the Web useful and usable. These systems are based on Information Retrieval methods for indexing plain text documents and heuristics for adjusting their document rankings based on the special HTML structure of Web documents. In this paper, we describe a wide range of such heuristics including a novel one inspired by reinforcement learning techniques for learning rewards through a graph which can be used to affect a search engine's rankings. We then demonstrate a system that combines these heuristics automatically, based on feedback collected unintrusively from users, resulting in much improved rankings.
[Postscript](#) [Referring Page](#) [Details](#) [BibTeX Entry](#) Word Matches: boyan, search engines Score: 1

2. Value Function Based Production Scheduling
Jeff G. Schneider Justin A. Boyan Andrew W. Moore
Abstract: Production scheduling, the problem of sequentially configuring a factory to meet forecasted demands, is a difficult problem throughout the manufacturing industry. The requirement of maintaining product inventories in the face of fluctuating demand and stochastic factory output makes standard scheduling models, such as job-shop, inadequate. Current algorithms, such as simulated annealing and constraint propagation, must employ ad-hoc methods such as frequent replanning to cope with uncertainty. In this paper, we describe a Markov Decision Process (MDP) formulation of production scheduling that captures stochasticity in both production and demands. The solution to this MDP is a value function which can be used to generate optimal scheduling decisions online. A simple example illustrates the theoretical superiority of this approach over replanning-based methods. We then describe an industrial application and two reinforcement learning methods for approximating the value function in this domain. Our results demonstrate that in both deterministic and noisy scenarios, value function approximation is an effective technique.
[Postscript](#) [Referring Page](#) [Details](#) [BibTeX Entry](#) Word Matches: boyan Score: 0.6094

3. Least-Squares Temporal Difference Learning
Justin A. Boyan
Abstract: Submitted to NIPS-98 TD(0) is a popular family of algorithms for approximate policy evaluation in large state spaces by incrementally updating the value function after each observed transition. It has two major drawbacks: inefficient use of data, and it requires the user to manually tune a stepsize schedule for good performance. For value function approximations and $\gamma = 0$, the Least-Squares TD (LSTD) algorithm of Bradtko and Barto [5] eliminates these drawbacks and improves data efficiency. This paper extends Bradtko and Barto's work in three significant ways: it presents a simpler derivation of the LSTD algorithm. Second, it generalizes from $\gamma = 0$ to arbitrary values of γ ; at the same time, the resulting algorithm is shown to be a practical formulation of supervised linear regression. Third, it presents



Rexa.info

■ Research • People × Connections

Andrew McCallum • Tags • Send Invites (477) • Submit • Logout

Papers Authors Grants

Search

Optional fields include abstract: body: title: author: venue: year: tag:

Queries may use AND, OR or (). Default is OR.

W. Bruce Croft

[\[Google\]](#)[\[Edit Info\]](#)[\[Send Invite\]](#)[\[Email link\]](#)



Distinguished Professor

Department of Computer Science, University of Massachusetts

BRUCE CROFT, Amherst, MA, 01003-9264

Email: croftg@cs.umass.edu

URL: <http://ciir.cs.umass.edu/personnel/croft.html>

Publications: (1 to 40 of 233) (total 1436 citations)

Sorted by **date** | [citations](#)

- 2004
 - Donald Metzler, W. Bruce Croft. *Combining the language model and inference network approaches to retrieval*. Inf. Process. Manage. vol 40, pages 735, 2004 (1 citation)
 - Xiaoyong Liu, W. Bruce Croft. *Cluster-based retrieval using language models*. SIGIR, 2004 (0 citations)
 - Andrés Corrada-Emmanuel, W. Bruce Croft. *Answer models for question answering passage retrieval*. SIGIR, 2004 (0 citations)
 - Chirag Shah, W. Bruce Croft. *Evaluating high accuracy retrieval techniques*. SIGIR, 2004 (1 citation)
 - Haizheng Zhang, W. Bruce Croft, Brian N. Levine, Victor R. ... *A Multi-Agent Approach for Peer-to-Peer Based Information Retrieval System*. AAMAS, 2004
 - Donald Metzler, Victor R. ... W. Bruce Croft. *Formal ... models for language modeling*. SIGIR, 2004
 - Stephen Cronen-Townsend, Yu Zhou, W. Bruce Croft. *A framework for selective query expansion*. CIKM, 2004 (0 citations)
- 2003
 - W. Bruce Croft. *Language Models for Information Retrieval*. JCDL, 2003 (0 citations)

Co-authors | Cited authors | Citing authors: (1 to 40 of 257)

Sorted by **date** | [number](#) | [name](#)

- Victor Lavrenko, 2004 2003 2002 2002 2001 2001
???? ????
 - Stephen Cronen-Townsend, 2004 2002 2001 ????
 - Donald Metzler, 2004 2004 2003
 - Xiaoyong Liu, 2004 2002
 - Andrés Corrada-Emmanuel, 2004
 - Victor Lavrenko, 2004
 - Brian N. Levine, 2004
 - Chirag Shah, 2004
 - Haizheng Zhang, 2004
 - Yu Zhou, 2004
 - James P. Callan, 2003 2001 1997 1996 1996 1995
1995 1995 1995 1995 1994 1994 1994 1994 1994
1993 1993 1993 1992 1992 ???? ???? ????
 - Howard R. Turtle, 2003 1999 1997 1996 1993 1992

institutions, conferences, journals, grants, advisors,...

Application Goals

A KB of all scientists in the world

from papers, patents, web pages, newswire, press releases, tweets, blogs,...

A KB of scientific entities & relations

materials, equipment, organisms, processes, tasks, methods,...

- Better tools → Accelerate progress of science.
- Revolutionize peer review
 - “open peer review”
 - Submission, reviews & comments public.

Improving Generative Adversarial Networks with Denoising Feature Matching [pdf](#)

David Warde-Farley, Yoshua Bengio

5 Nov 2016 ICLR 2017 conference submission readers: everyone

Abstract: We propose an augmented training procedure for generative adversarial networks designed to address shortcomings of the original by directing the generator towards probable configurations of abstract discriminator features. We estimate and track the distribution of these features, as computed from data, with a denoising auto-encoder, and use it to propose high-level targets for the generator. We combine this new loss with the original and evaluate the hybrid criterion on the task of unsupervised image synthesis from datasets comprising a diverse set of visual categories, noting a qualitative and quantitative improvement in the "objectness" of the resulting samples.

TL;DR: Use a denoiser trained on discriminator features to train better generators.

Conflicts: umontreal.ca, iro.umontreal.ca, polymtl.ca, google.com

Keywords: Deep learning, Unsupervised Learning

Authorids: d.warde.farley@gmail.com, yoshua.umontreal@gmail.com

Add [Comment](#) [public review](#)

1 reply

Training Scheme and Denoising

Antonia Creswell

16 Nov 2016 ICLR 2017 conference paper580 public comment readers: everyone

Comment: The generations in this paper suggest that using extra information from features of the discriminator allows the generator to produce images with more object like features. I have some questions/comments:

- 1) In equation 5 it appears that you are training r to reconstruct a corrupted version of the features, rather than the features themselves, the reason for this is not clear?
$$\|C(\phi(x)) - r(C(\phi(x)))\|$$
 rather than $\|\phi(x) - r(C(\phi(x)))\|$
- 2) This approach involves training 3 networks. It would be interesting to know what kind of training scheme was used? Whether, D,G and r networks are trained for one iteration each, or if some networks are trained for more iterations before updating the next network?
- 3) It would also be interesting to know whether parameters l_{denoise} and l_{adv} are fixed or adjusted during training?

Add [Comment](#)

* denotes a required field

title

Brief summary of your comment.

comment

Your comment or reply.

Text

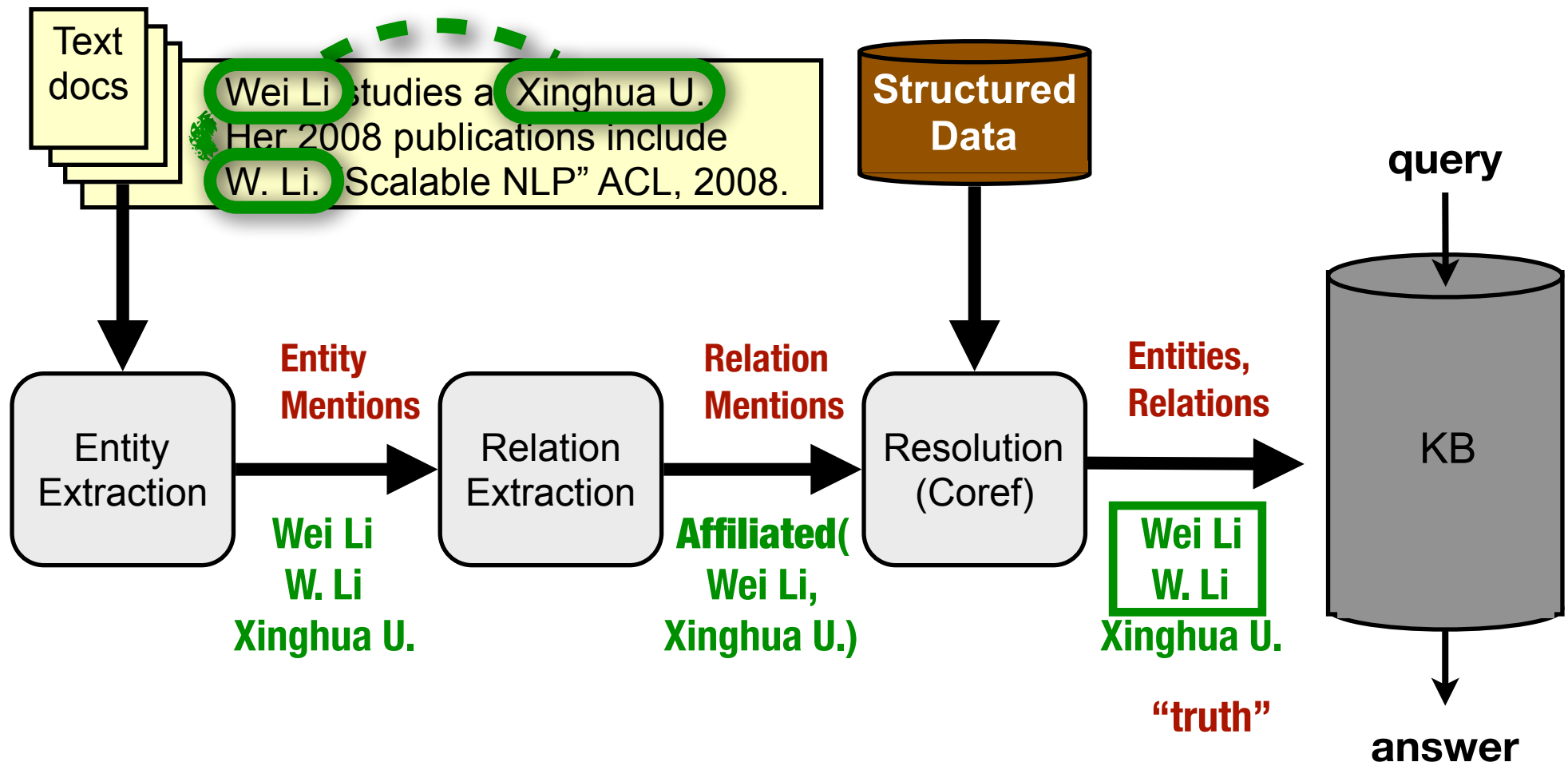


Knowledge Base



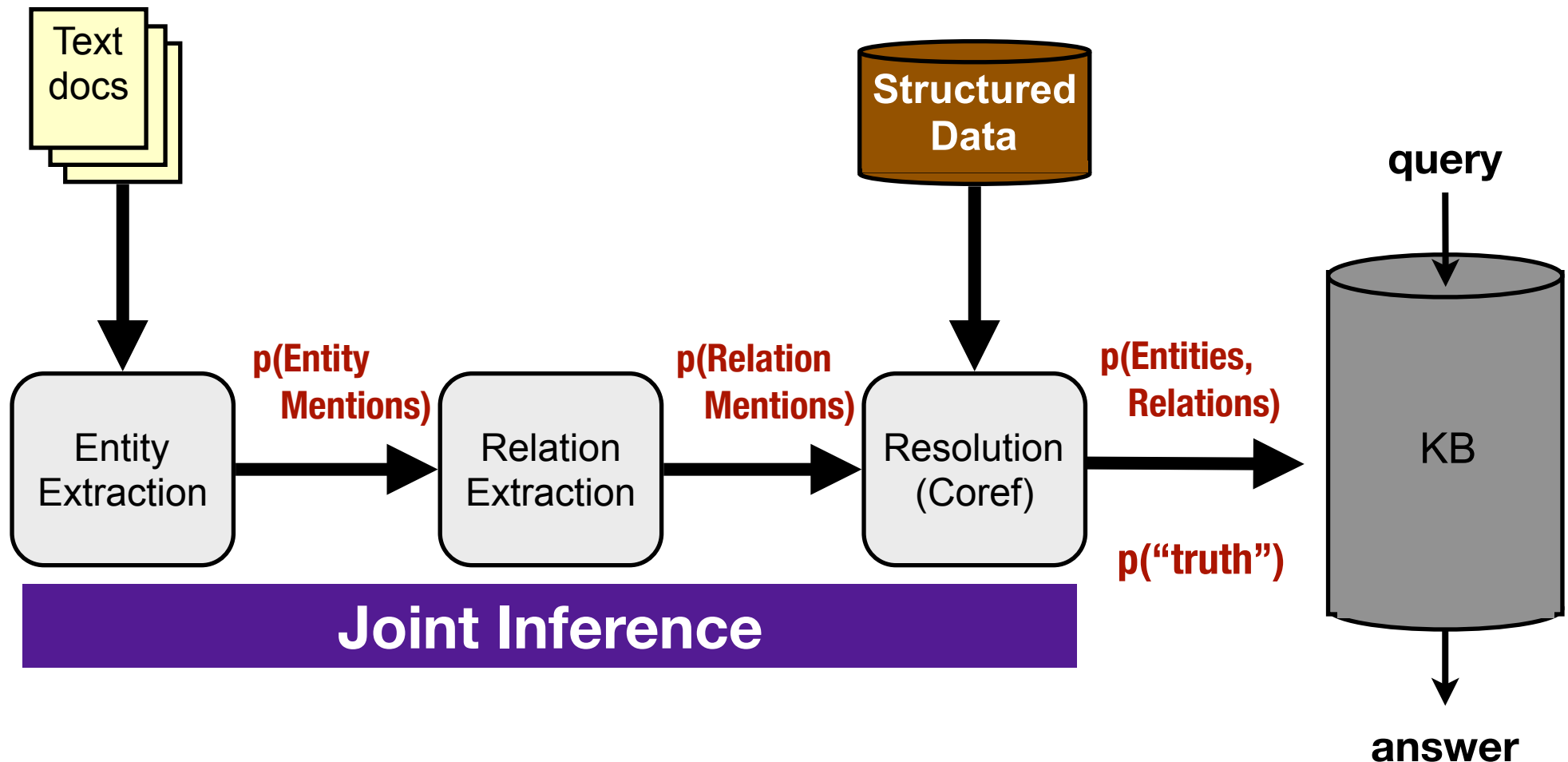
Reasoning

Knowledge Base Construction



**Information Extraction components aren't perfect.
Errors snowball.**

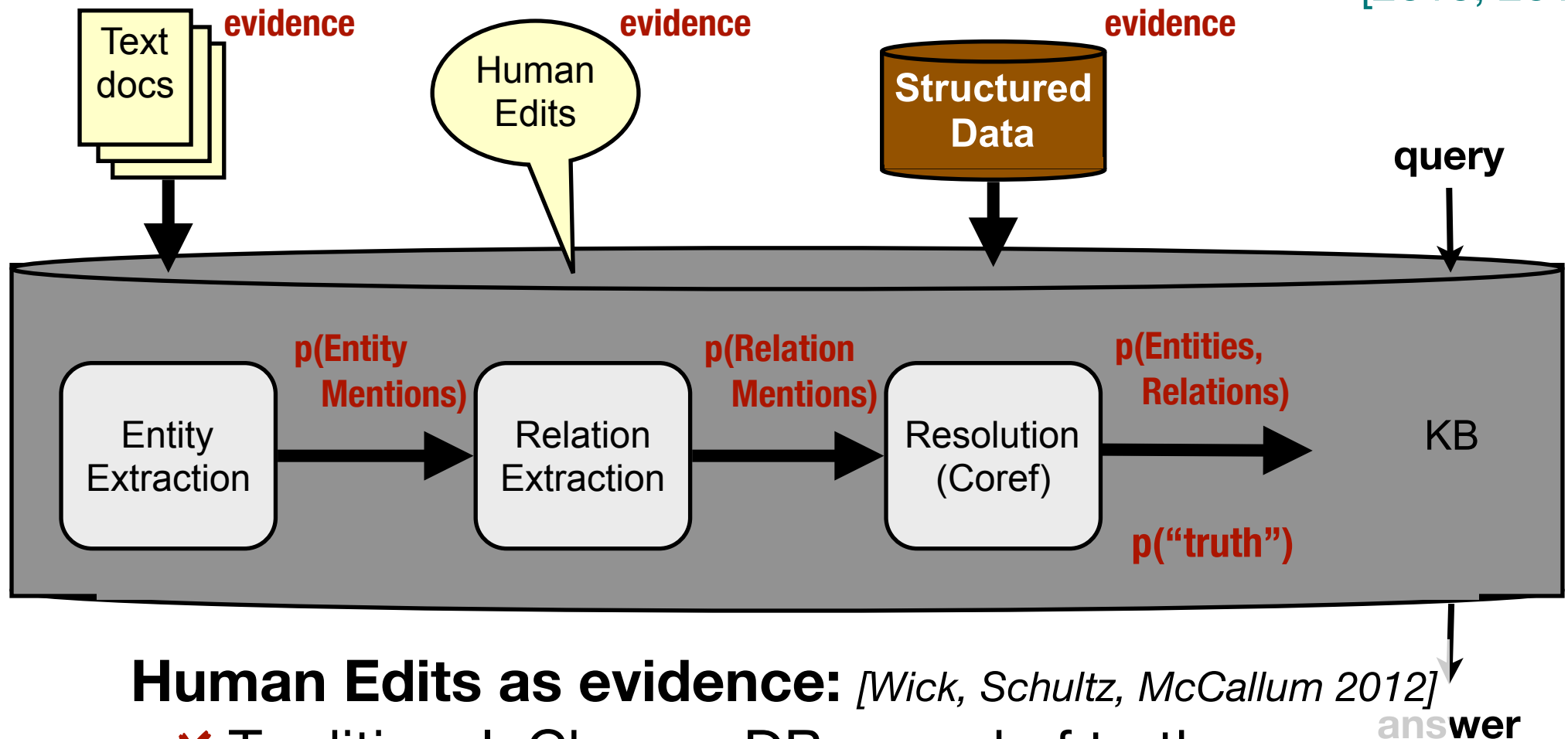
Knowledge Base Construction



1. How to represent & inject uncertainty from IE into KB?
2. How to use KB contents to aid IE?
3. IE isn't "one-shot." Add new data later; redo inference. Want KB infrastructure to manage IE.

“Epistemological Balkan Conundrum”

[2010, 2012]



Human Edits as evidence: [Wick, Schultz, McCallum 2012]

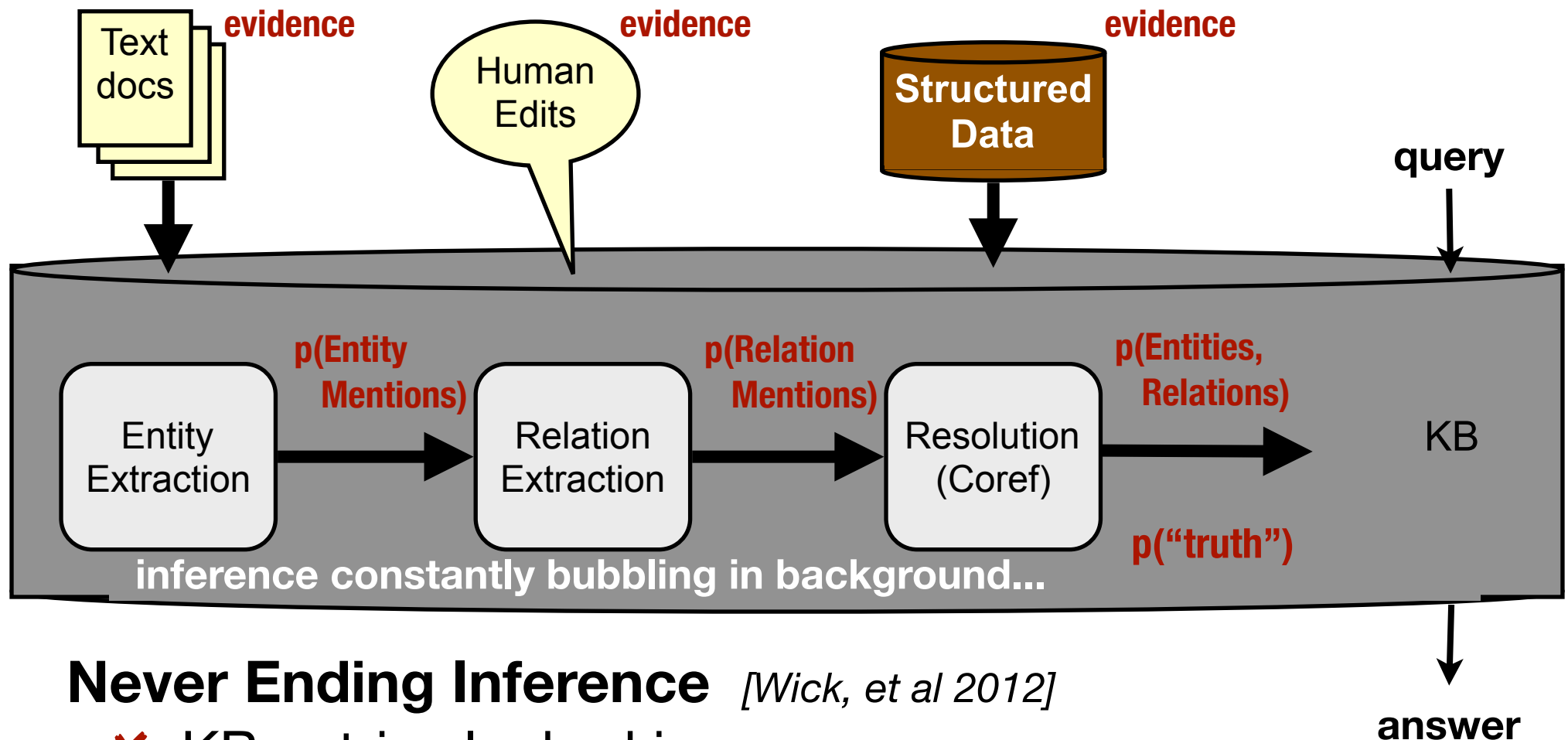
✗ Traditional: Change DB record of truth

✓ Mini-document “Nov 15: Scott said this was true”

- Sometimes humans are wrong, disagree, out-of-date.
 - Jointly reason about truth & editors' reliability/reputation.
- “Truth is inferred, not observed.”**

Epistemological Philosophy

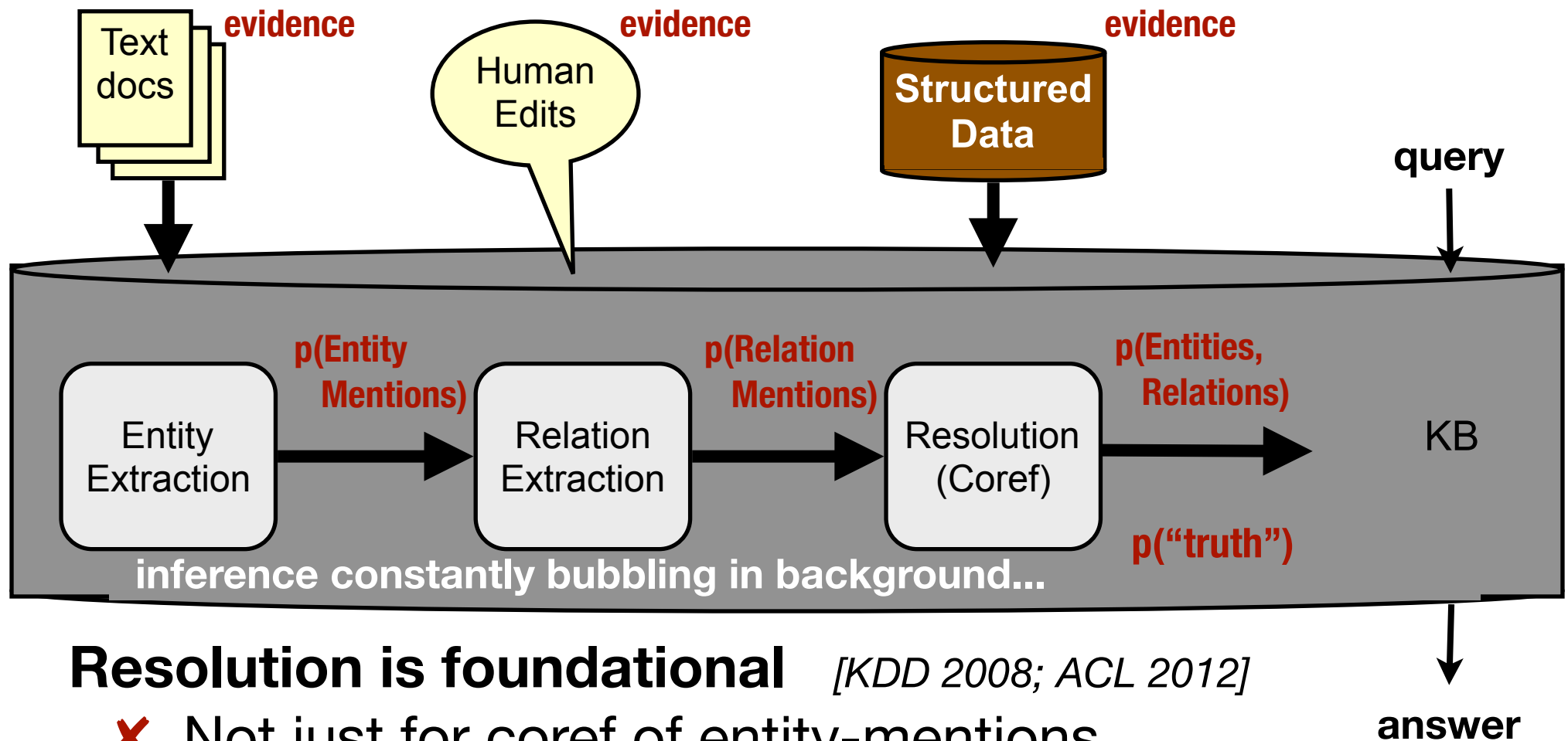
“Epistemological KnowledgeBase”



Never Ending Inference [Wick, et al 2012]

- ✗ KB entries locked in
- ✓ KB entries always reconsidered with more evidence, time,...

“Epistemological KnowledgeBase”



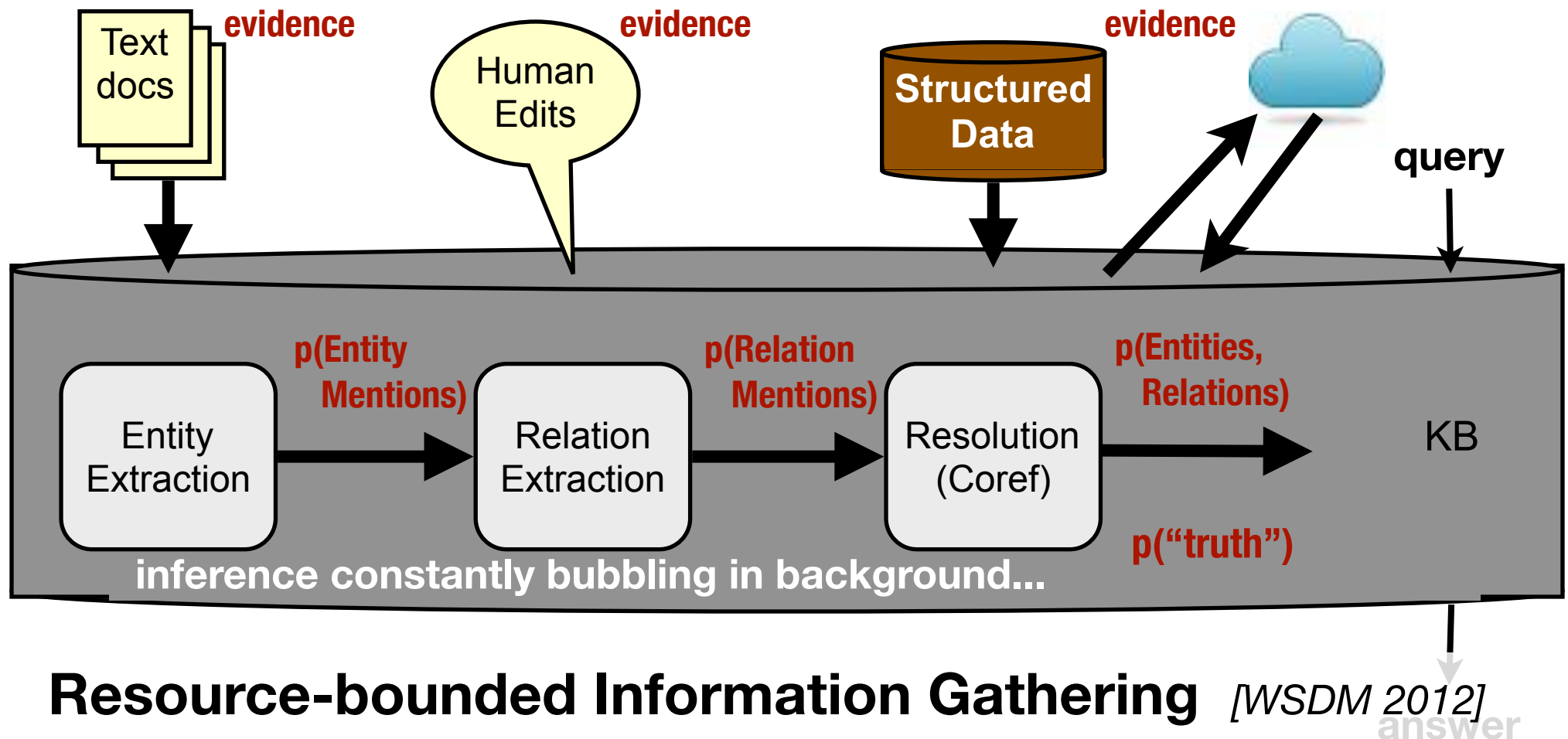
Resolution is foundational [KDD 2008; ACL 2012]

✗ Not just for coref of entity-mentions...

✓ Align values, ontologies, schemas, relations, events,...

Especially in Epistemological DB: entities/relations never input, only “mentions”

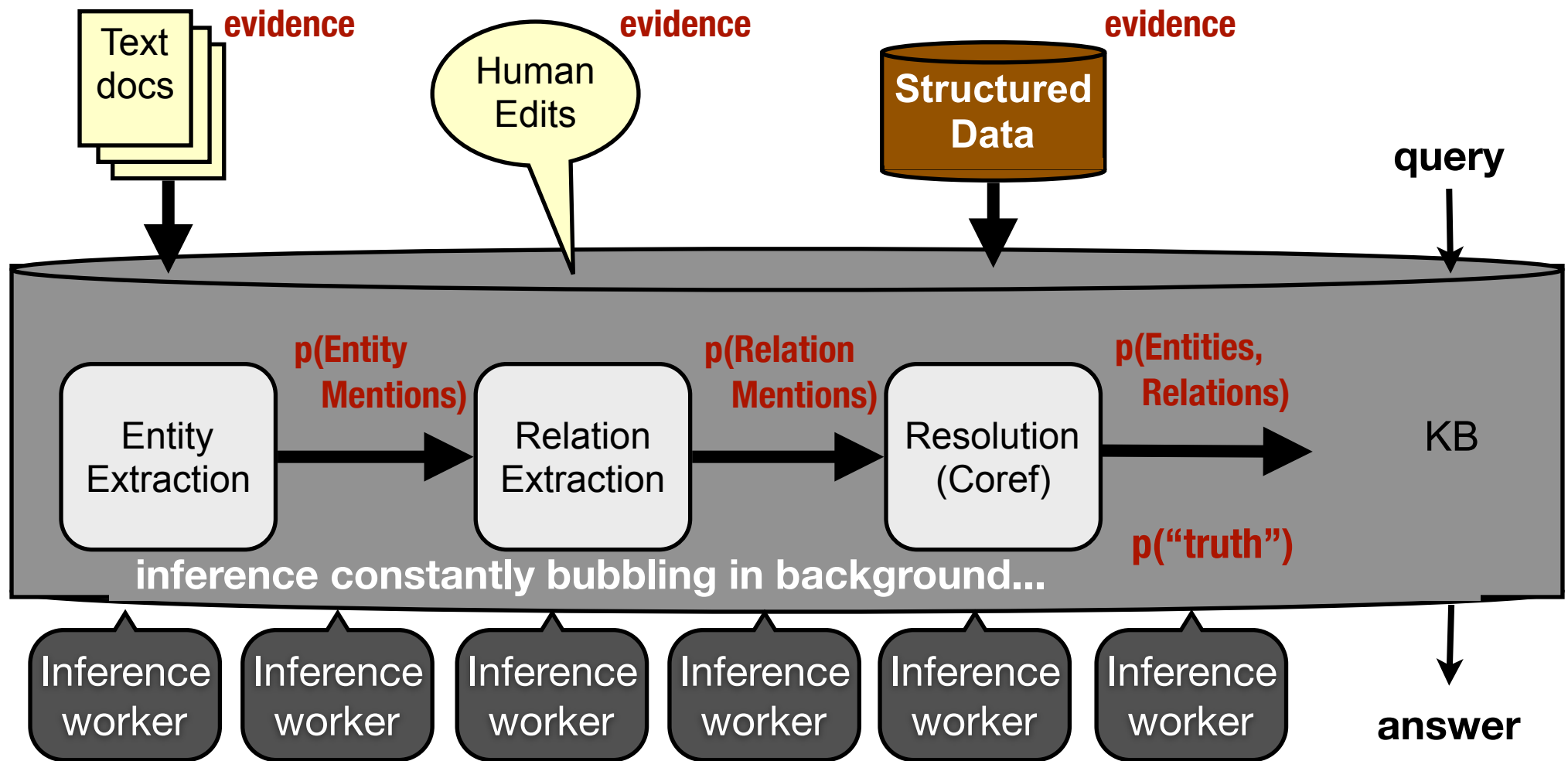
“Epistemological KnowledgeBase”



Resource-bounded Information Gathering [WSDM 2012]

- ✗ Full processing on whole web
- ✓ Focus queries and processing where needed & fruitful

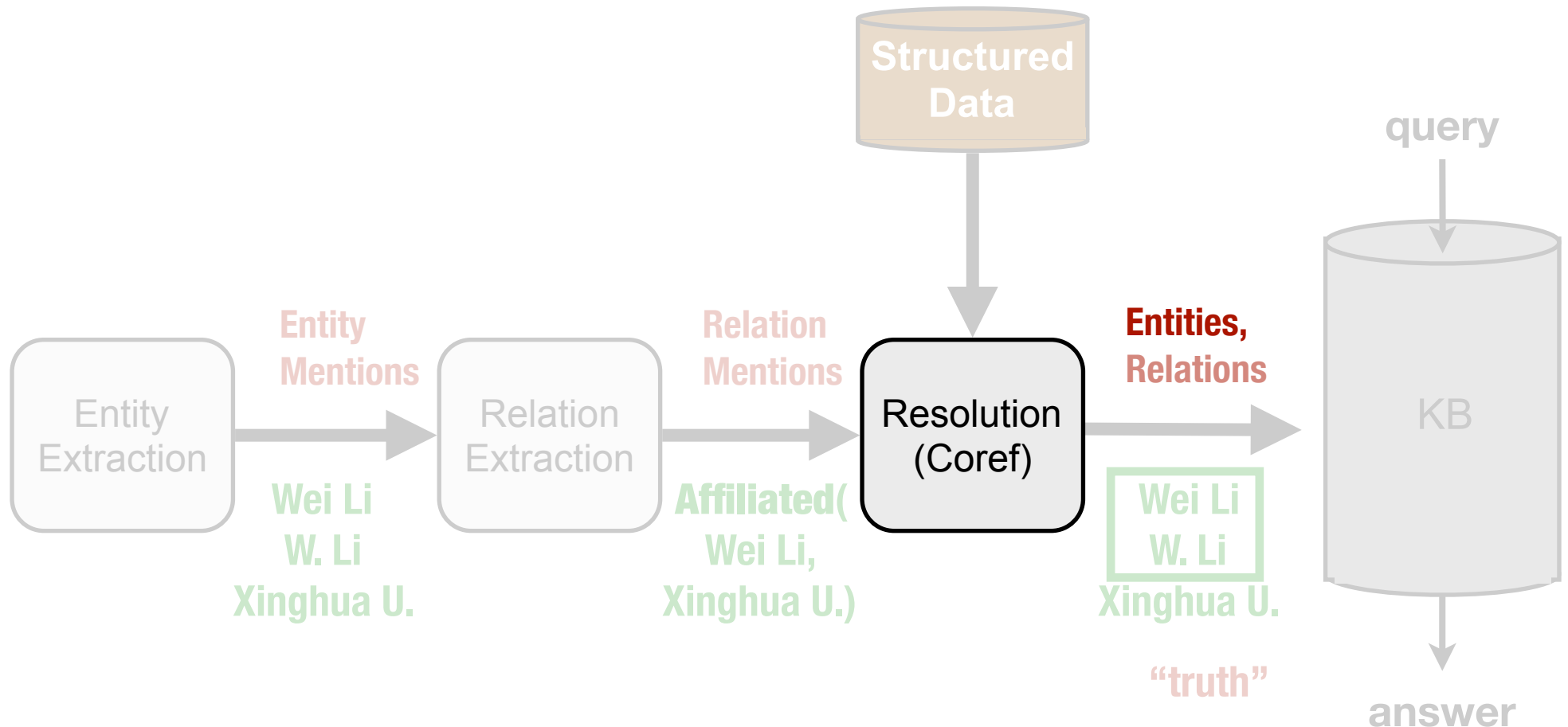
“Epistemological KnowledgeBase”



Smart Parallelism [ACL 2011; NIPS 2011]

- ✗ MapReduce, black-box
- ✓ Reason about inference & parallelism together

Entity Resolution



Author Entity Resolution

A. Banerjee, S. Chassang, E. Snowberg. *Decision Theoretic Approaches to Experiment Design and External Validity*. Handbook of Field Experiments. 2016.

Arindam Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh. *Clustering with Bregman Divergences*. JMLR. 2006.

A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra. *Clustering on the Unit Hypersphere using von Mises-Fisher Distributions*.
Journal of Machine Learning Research. 2005

Author Entity Resolution

A. Banerjee, S. Chassang, E. Snowberg. *Decision Theoretic Approaches to Experiment Design and External Validity*. Handbook of Field Experiments. 2016.

Arindam Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh. *Clustering with Bregman Divergences*. JMLR. 2006.

A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra. *Clustering on the Unit Hypersphere using von Mises-Fisher Distributions*. Journal of Machine Learning Research. 2005

arXiv:1511.06396v2 [cs.CL] 3 Mar 201

Proc. Natl. Acad. Sci. USA
Vol. 91, pp. 2395-2400, March 1994
Colloquium Paper

This paper was presented at a colloquium entitled "Changes in Human Ecology and Behavior: Effects on Infectious Diseases," organized by Bernard Rodman, held September 27 and 28, 1993, at the National Academy of Sciences.

 CrossMark
Click for updates

Risk to developed and developing countries

naVax, Inc., 250 Albany Street, Cambridge, MA 02139

viruses are members of the Flaviviridae in a cycle involving humans and

Moléculaire et Structurale, Unité Mixte de Recherche 2472/1157, Centre National de la Recherche Scientifique et National de la Recherche Agronomique, 1 Avenue de la Terrasse, 91198 Gif-sur-Yvette Cedex, France

Fig. 1. Semitransparent surface representation of the dengue virus E protein. The two subunits in the dimer are shown in different shades of blue. The backbone of the molecule is contoured in a yellow

<p>(54) ADVANCED VOICE AND DATA OPERATIONS IN A MOBILE DATA COMMUNICATION DEVICE</p> <p>(75) Inventor: Gary Monson, Waterloo (CA); Mike</p>	<p>confirmation-in-part of application No. 10/095,603, filed on Mar. 11, 2002.</p> <p>(60) Provisional application No. 60/274,408, filed on Mar. 9, 2001.</p>
---	---

No.	Author(s)	Year	Publication Classification
4	Lazaridis, Waterloo (CA); David Yach, Waterloo (CA); Raymond Vander Veen, Waterloo (CA); Harry Major, Waterloo (CA); Atul Ashiana, Waterloo (CA)	(51)	Int. Cl. <i>H04M</i> 11/00 (2006.01)

(73) Assignee: **Research in Motion Limited, Waterloo (CA)**

(21) Appl. No.: 11/458,843
(22) Filed: Jul. 20, 2006

Related U.S. Application Data

(63) Confirmation of application No. 10,830,836, which contains features that allow incoming data events to trigger outgoing voice events.

The diagram illustrates the flow of information between a Wireless Network, a Wireless Node, and a Wireless Network. The Wireless Node is connected to the Wireless Network via an Outgoing Cable and an Incoming Message. The Wireless Node is also connected to the Wireless Network via a Call Received on Voice Component and a Message Received. The Wireless Node is also connected to the Wireless Network via a Call Received on Voice Component and a Message Received.

Step 4

Call Requested on
Please Bring in Bag
Person ID - J0002
(416) 555-1212

Subject: Please Talk to Person W

Dear X:

I think the final contract details can be worked out with Person W. His number is based in Toronto as (416) 555-1212. Please call him

Step 3

ASAP as we need this solved today.

Sincerely, Y

Dual-Mode Status Device - 100

energy function of candidate labels,

COMMENTARY



Entity Resolution as Clustering

Given **mentions** $M = \{m_1, m_2, \dots, m_N\}$



Entity Resolution as Clustering

Given **mentions** $M = \{m_1, m_2, \dots, m_N\}$



A. Banerjee, S. Chassang, E. Snowberg. *Decision Theoretic Approaches to Experiment Design and External Validity*. Handbook of Field Experiments. 2016.

Entity Resolution as Clustering

Given **mentions** $M = \{m_1, m_2, \dots, m_N\}$



Entity Resolution as Clustering

Given **mentions** $M = \{m_1, m_2, \dots, m_N\}$



Partition M into **entities** $E = \{e_1, e_2, \dots, e_k\}$
where k unknown in advance

Entity Resolution as Clustering

Given **mentions** $M = \{m_1, m_2, \dots, m_N\}$

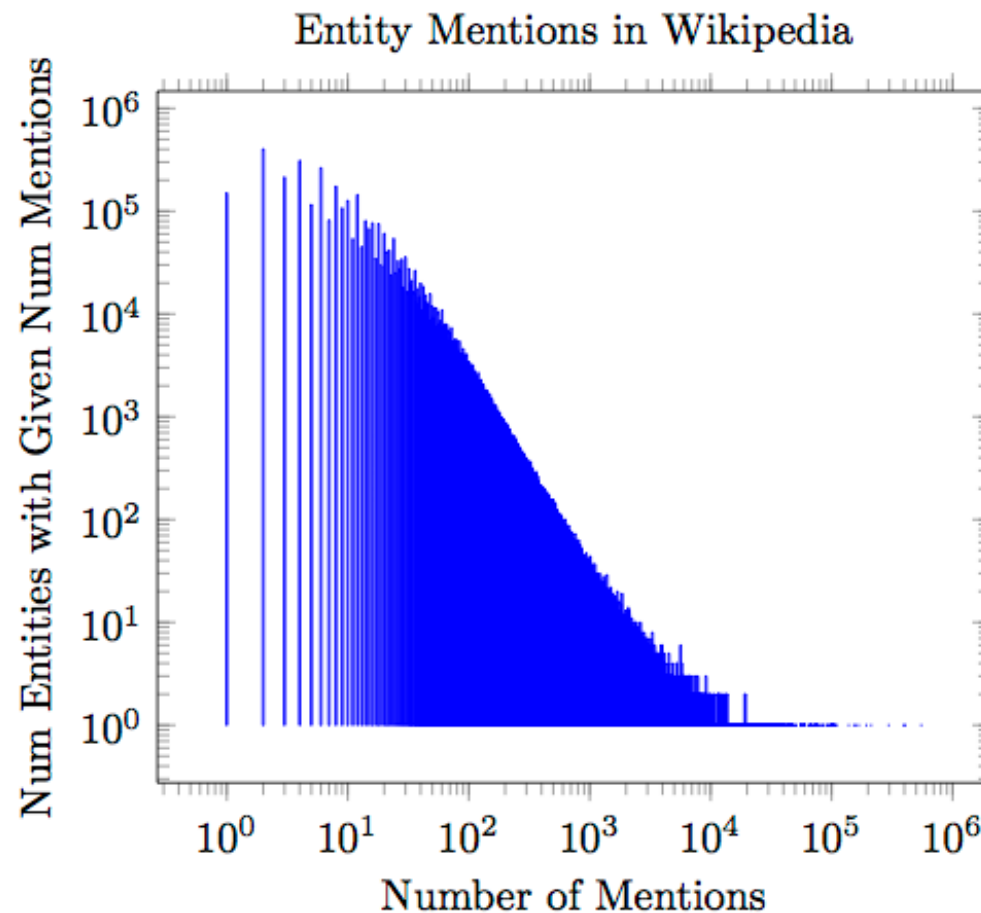


Partition M into **entities** $E = \{e_1, e_2, \dots, e_k\}$
where k unknown in advance



Entity Resolution Challenge

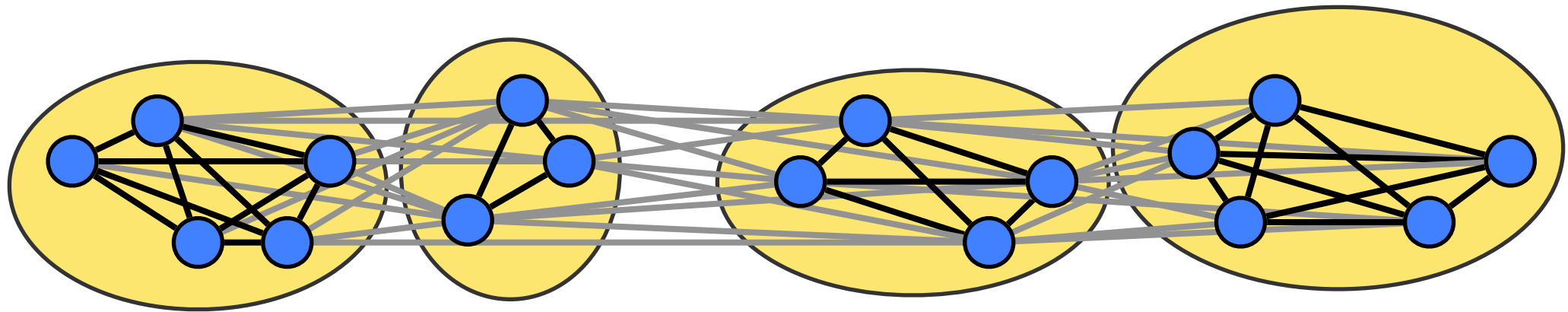
Power law of entity size



Large number of mentions (100Ks or 10Ms)
Large number of entities (many singleton clusters)

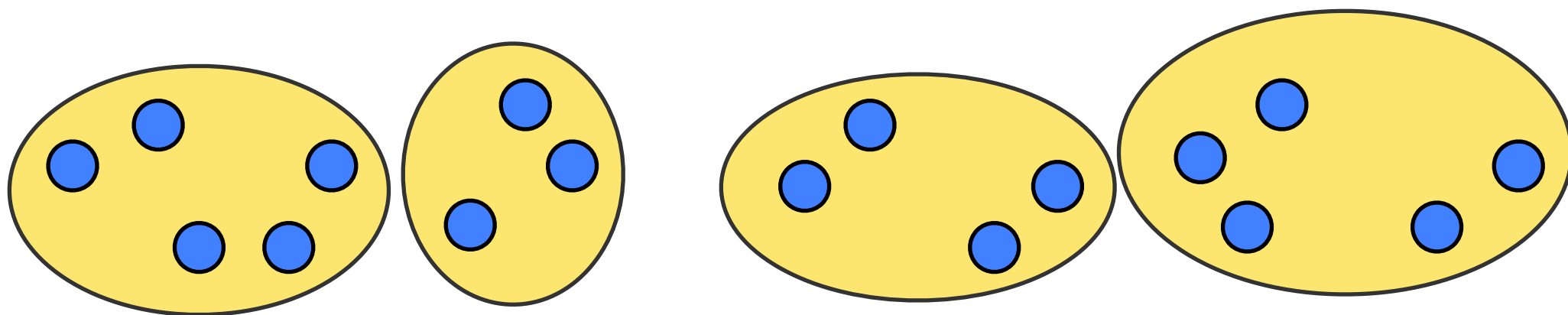
Pair-based Coref

Super-Entity
Entity
Sub-Entity
Mention

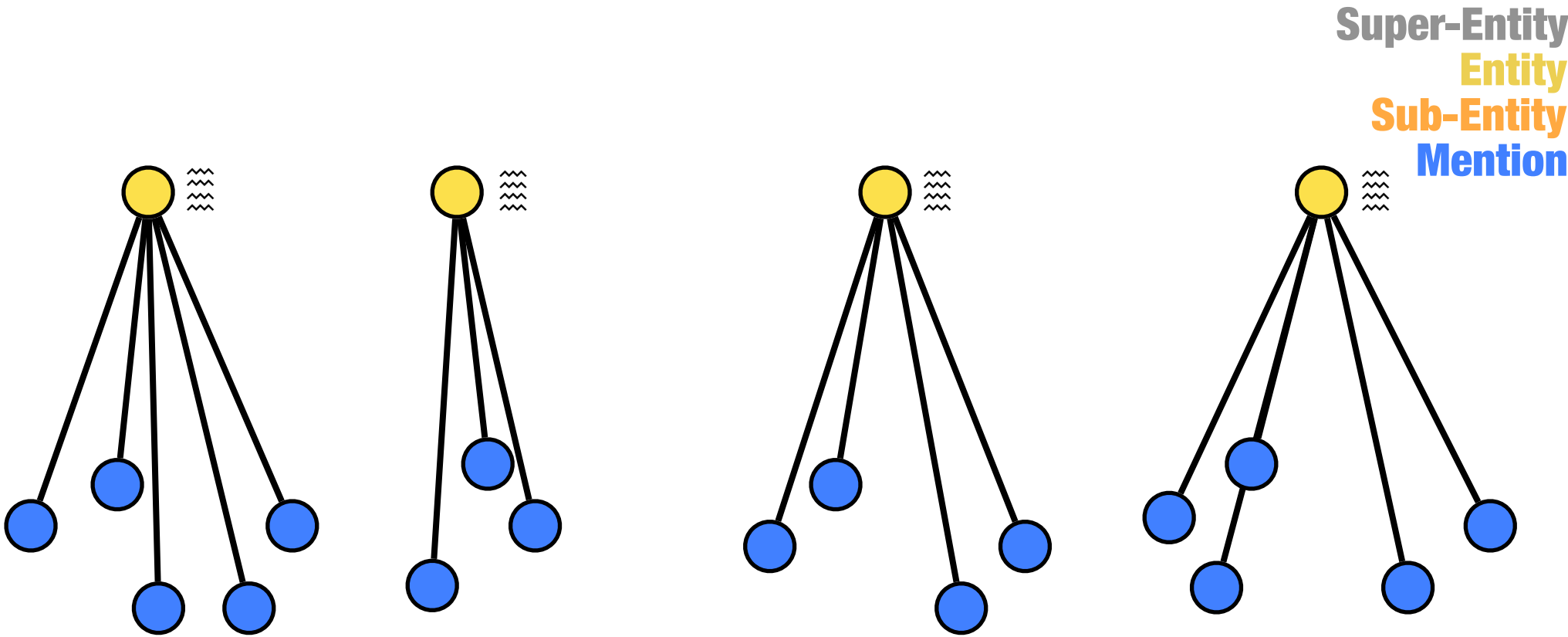


Pair-based Coref

Super-Entity
Entity
Sub-Entity
Mention

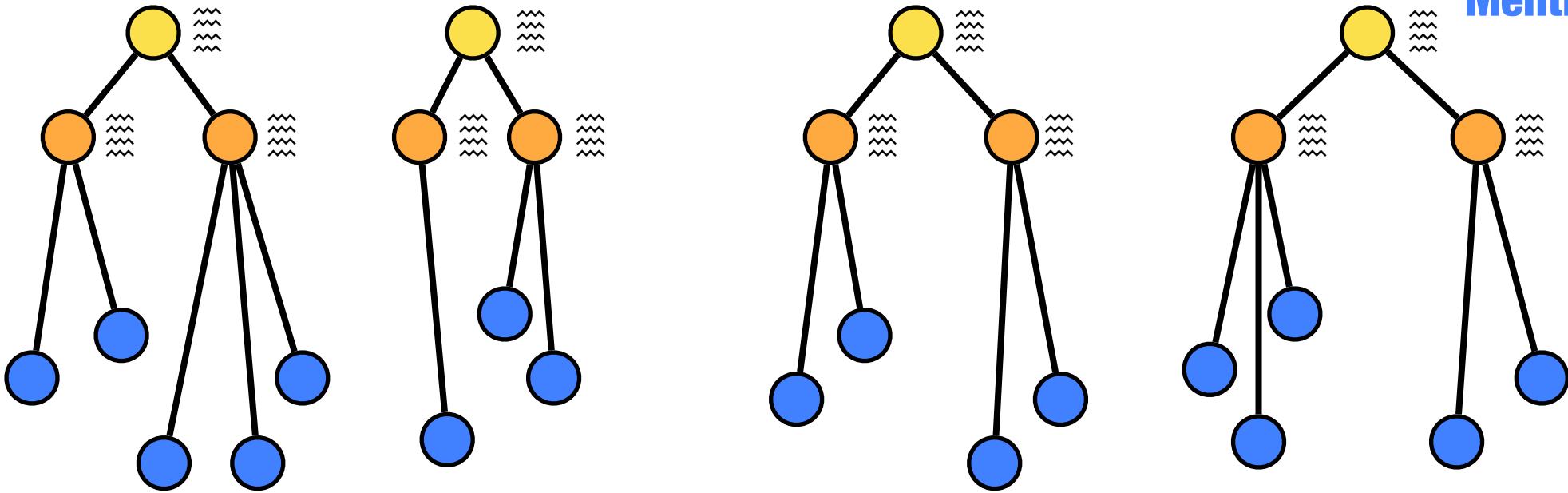


Entity-based Coref

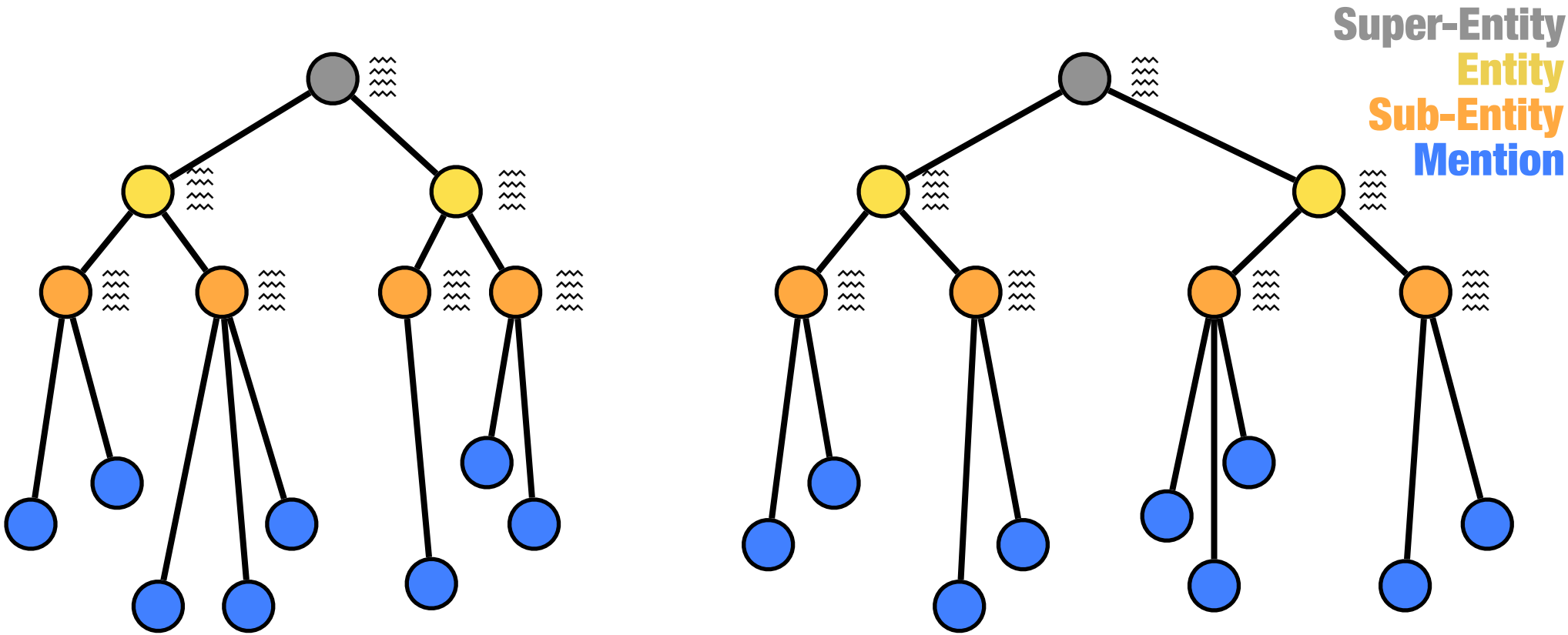


Entity-based Coref

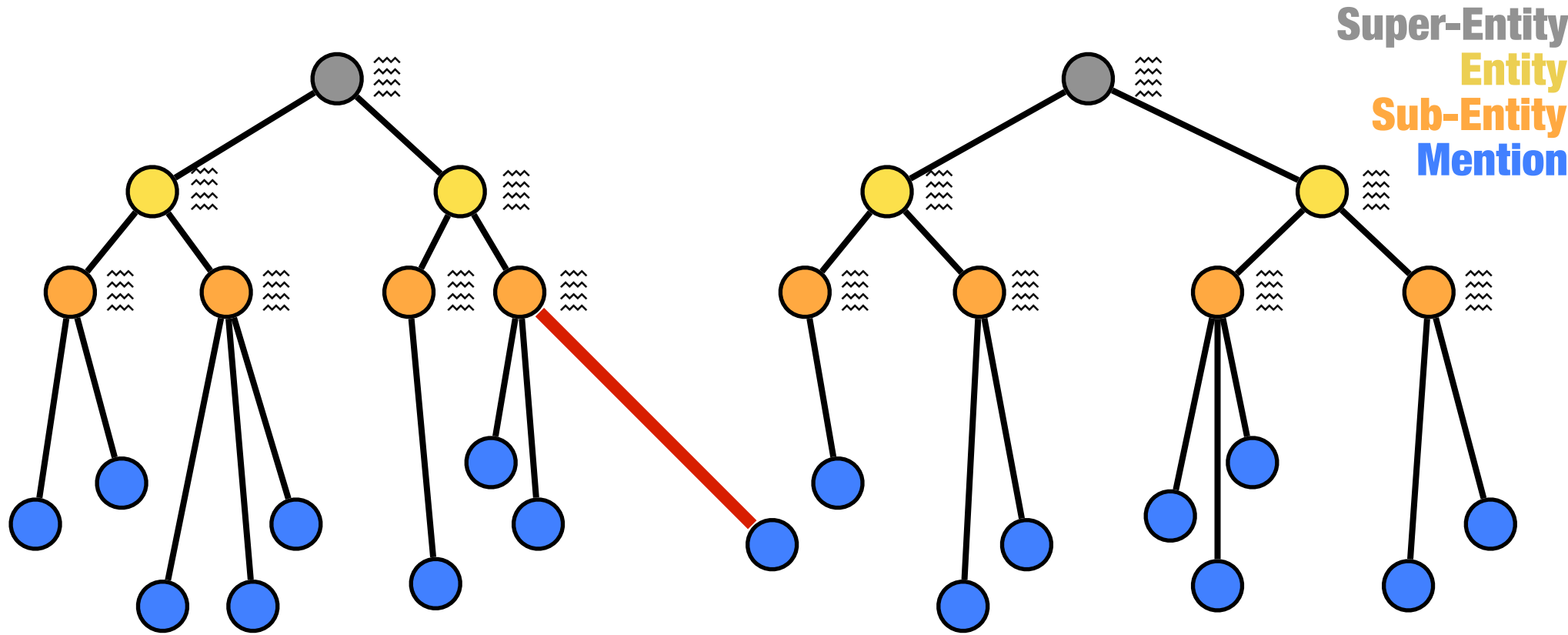
Super-Entity
Entity
Sub-Entity
Mention



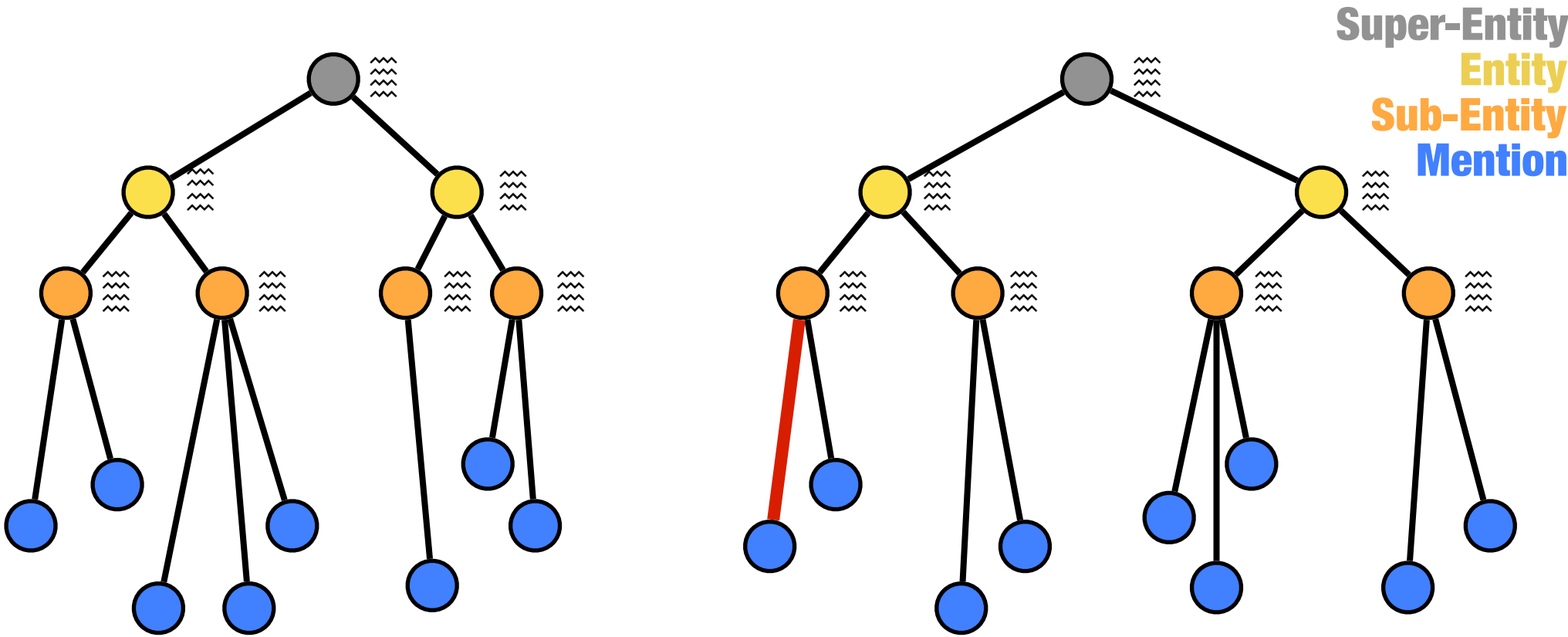
Entity-based Coref



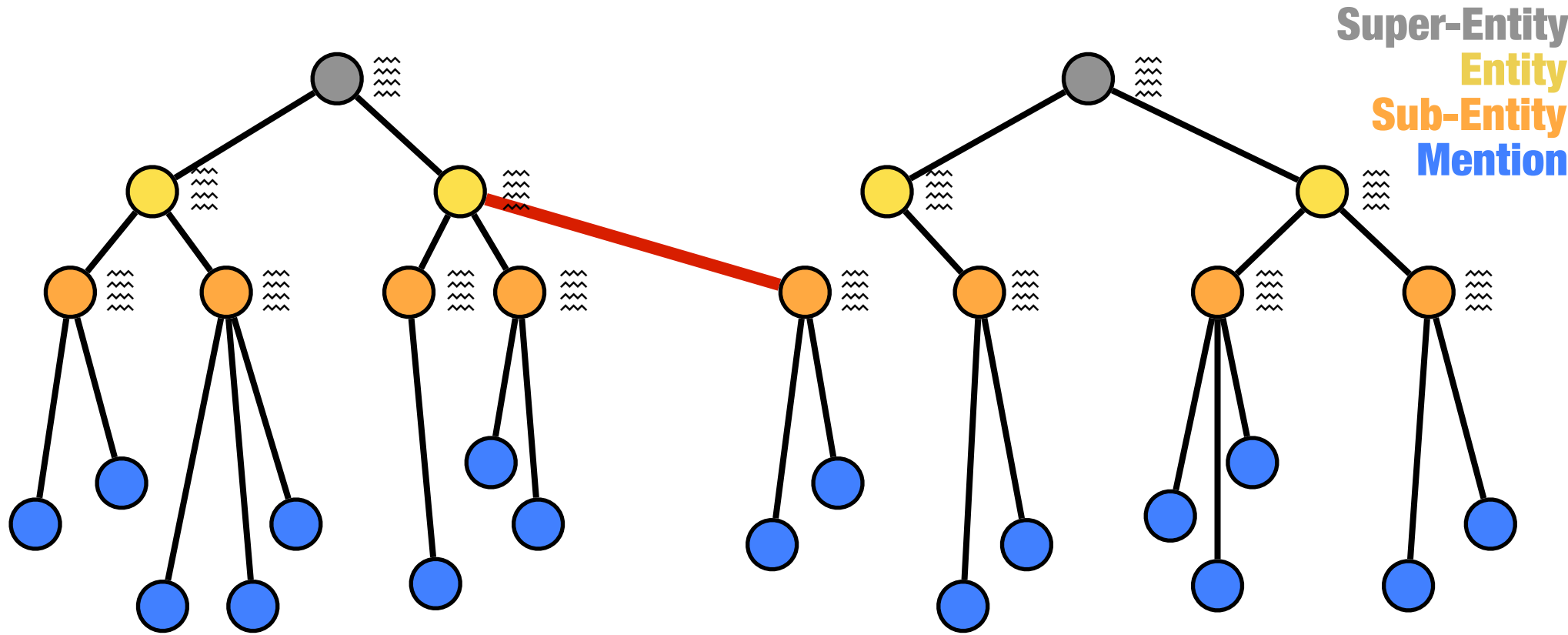
Entity-based Coref



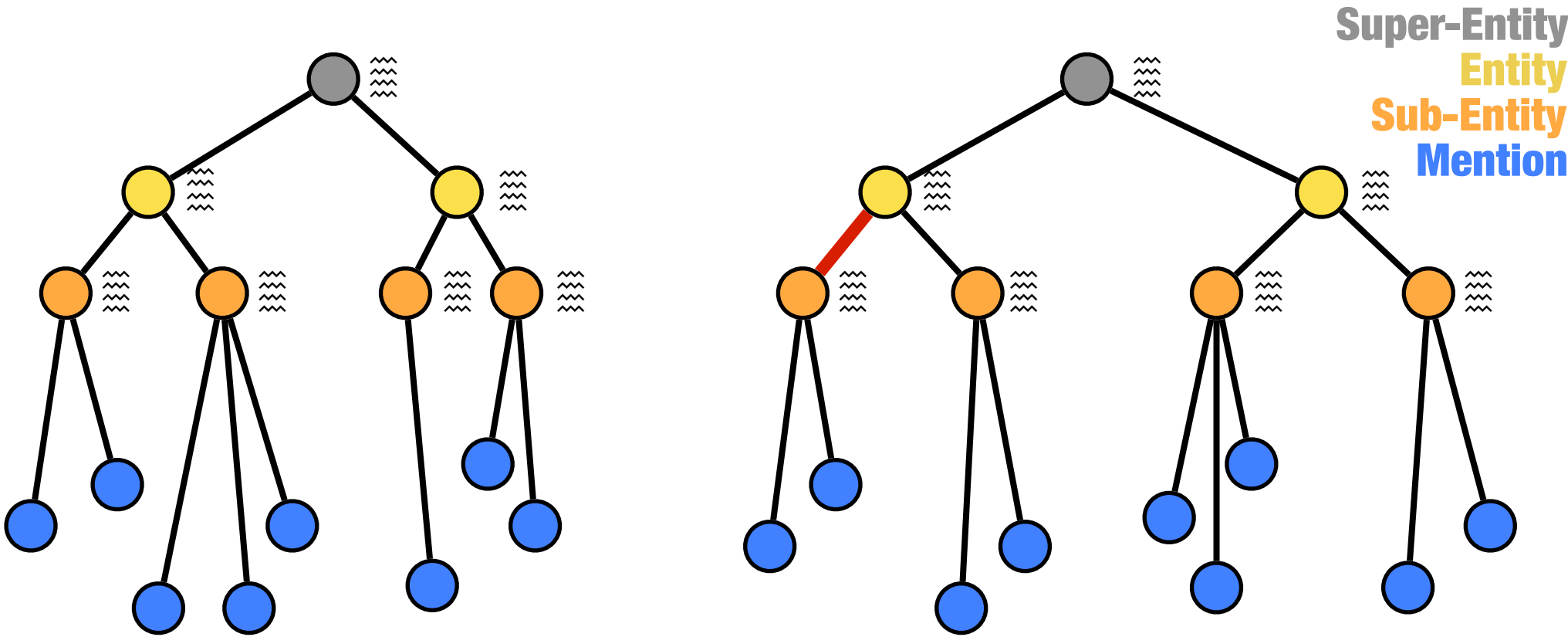
Entity-based Coref



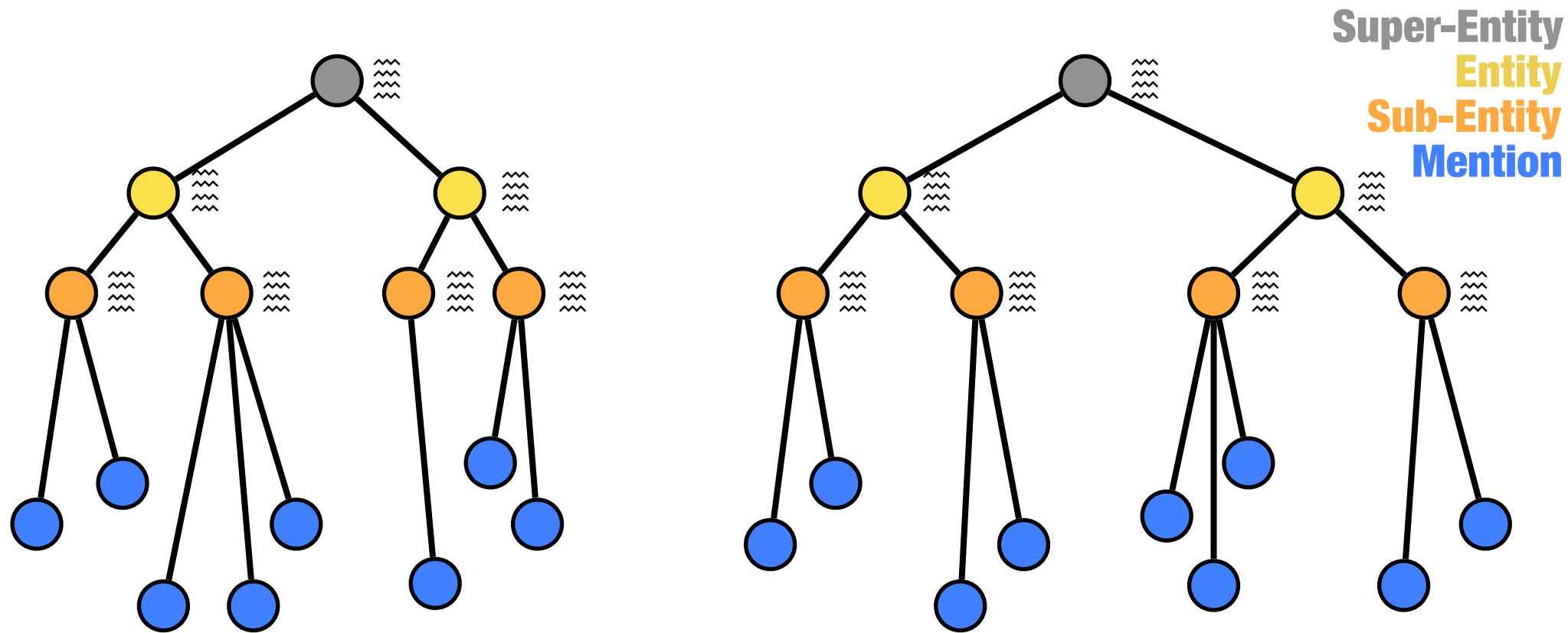
Entity-based Coref



Entity-based Coref



Entity-based Coref



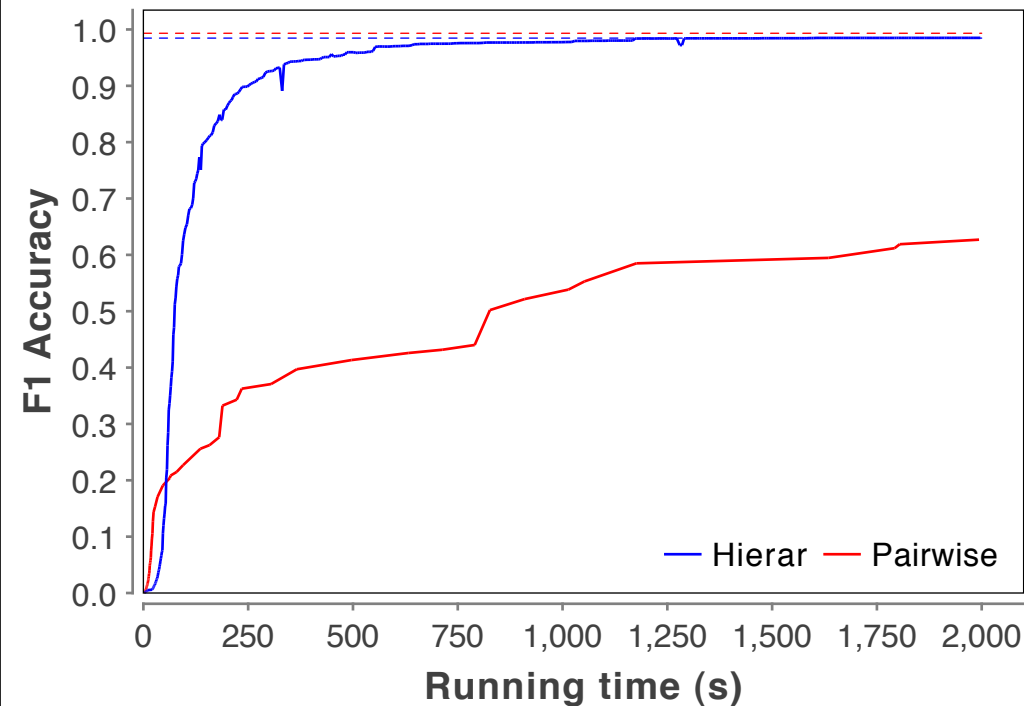
- ★ More efficient. Fewer factors; avoid N^2 .
- ★ Joint inference on all attributes of entity. Pair-wise couldn't
- ★ 100k mentions "e coli" hidden under one sub-entity.
- ★ Better supports inference about crowd-sourced edits

Hierarchical vs Pairwise Evaluation

Author Coreference (single threaded)

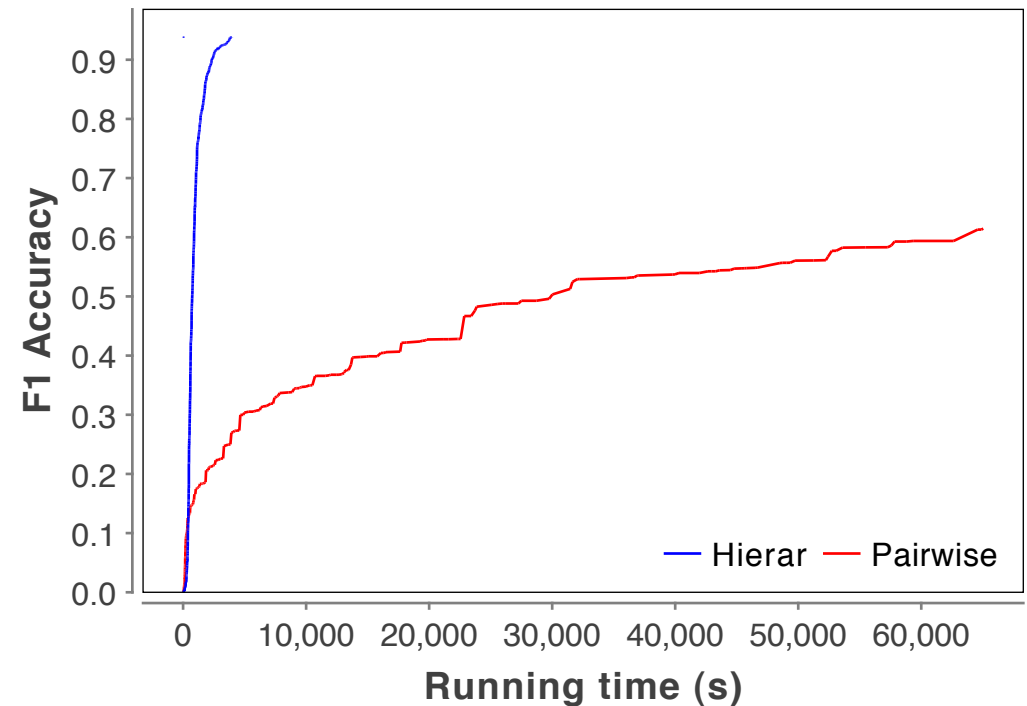
145k mentions

Accuracy versus Time



1.3m mentions

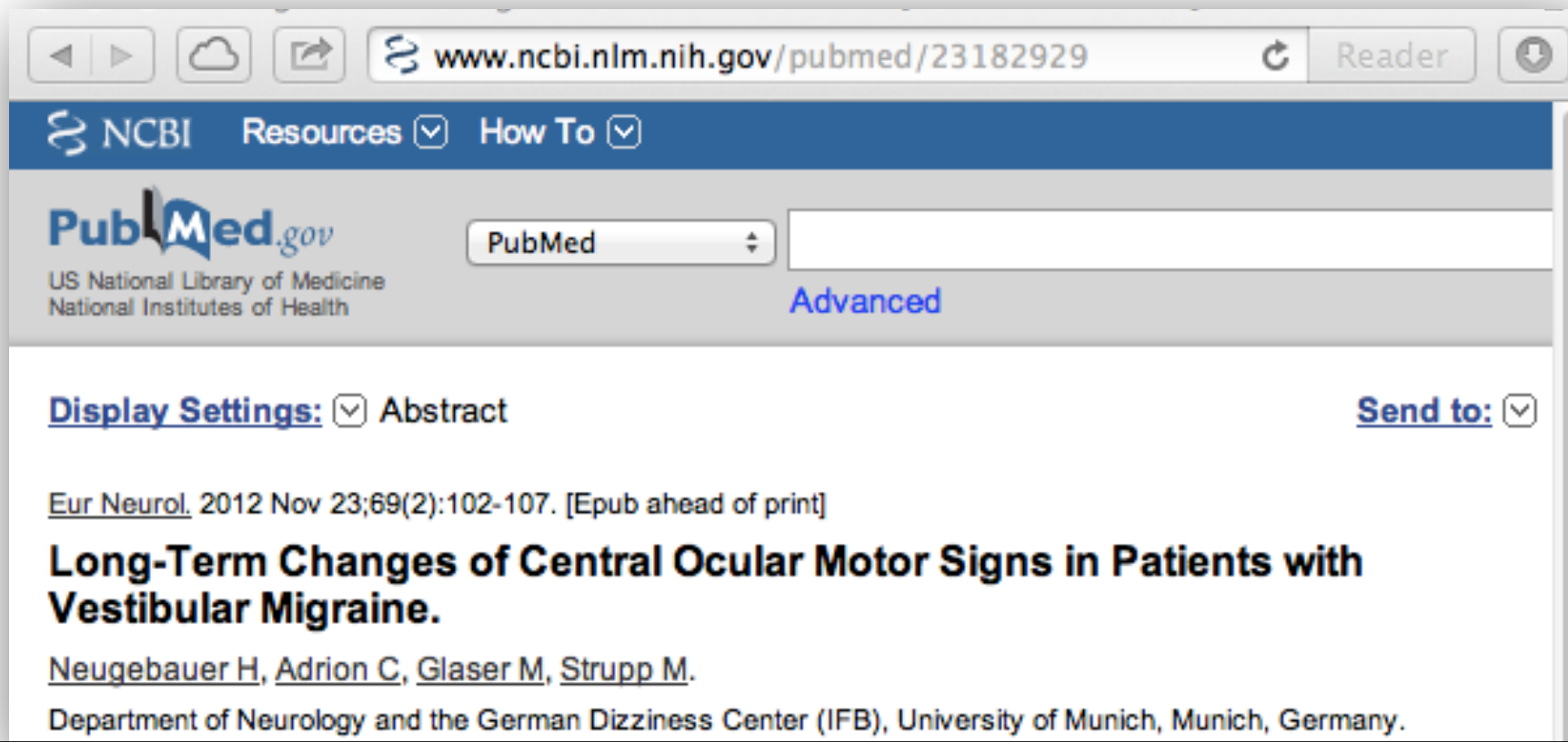
Accuracy versus Time



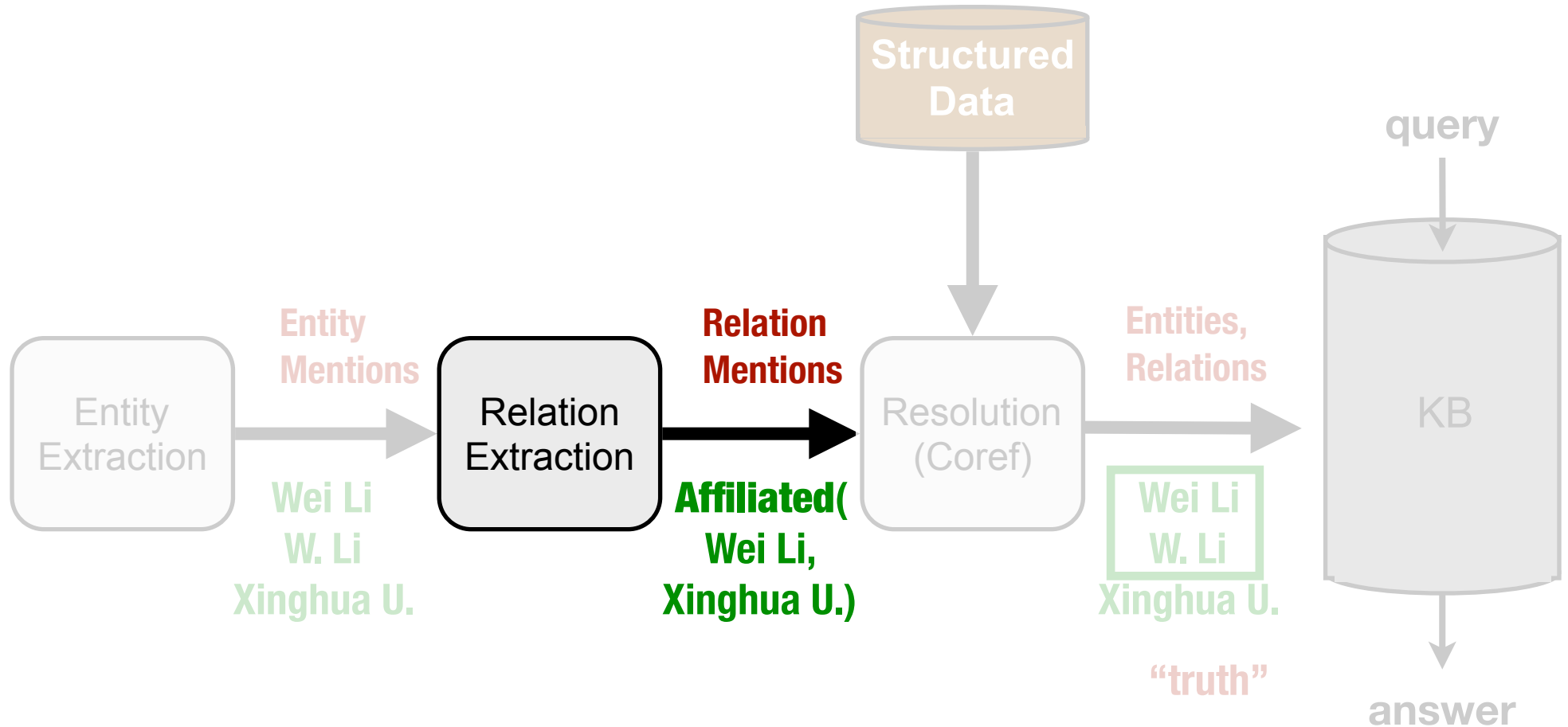
PubMed + Web of Science

- 200 million author mentions = ~400GB
- Inference speed
 - ~100k samples per second
 - ~48 hours of inference time

3 machines
48 cores



Relation Extraction



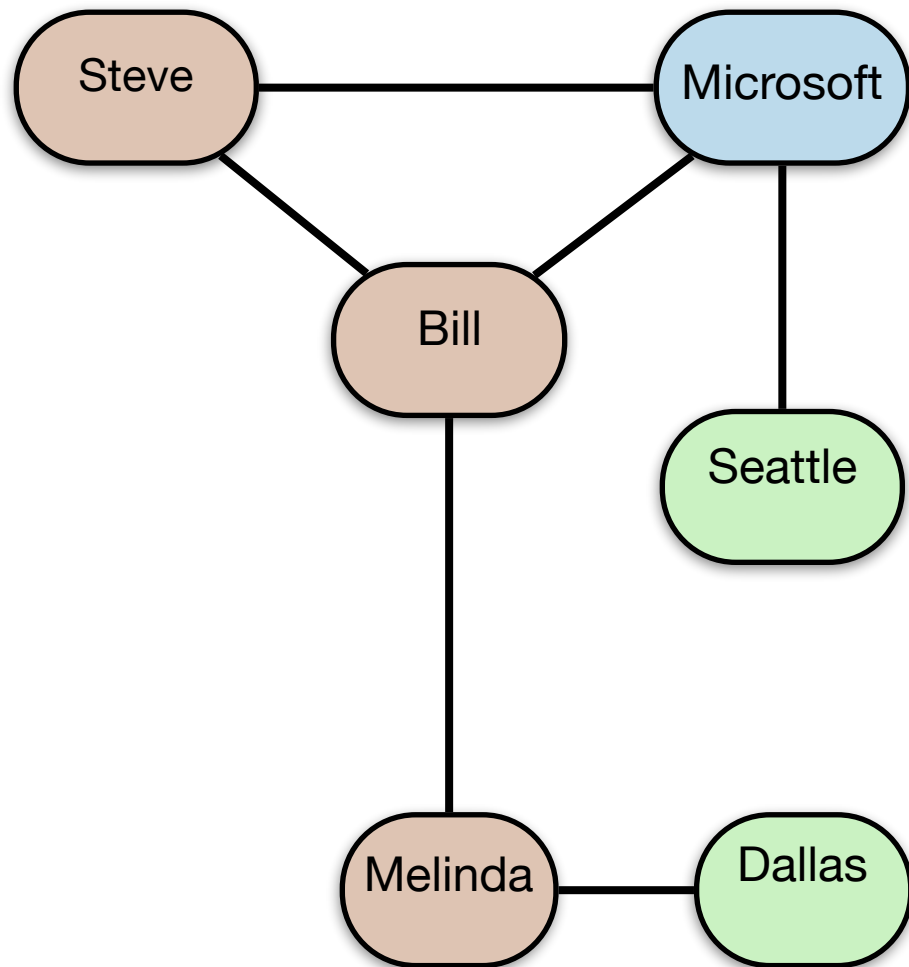
January 15, 2000

Tech pioneer Bill Gates stepped down today as chief executive officer of Microsoft, the Seattle-headquartered software giant. He will remain Chairman of the company, which rose to prominence after beating Digital Research Inc for the contract to provide an operating system for PCs. His long-time friend, Steve Balmer, will take over as CEO of Microsoft. Gates will now focus on the charitable foundation he runs with his wife Melinda French Gates. Bill and Melinda were married in a ceremony in Hawaii, rather than her hometown of Dallas. Steve Balmer was best man.

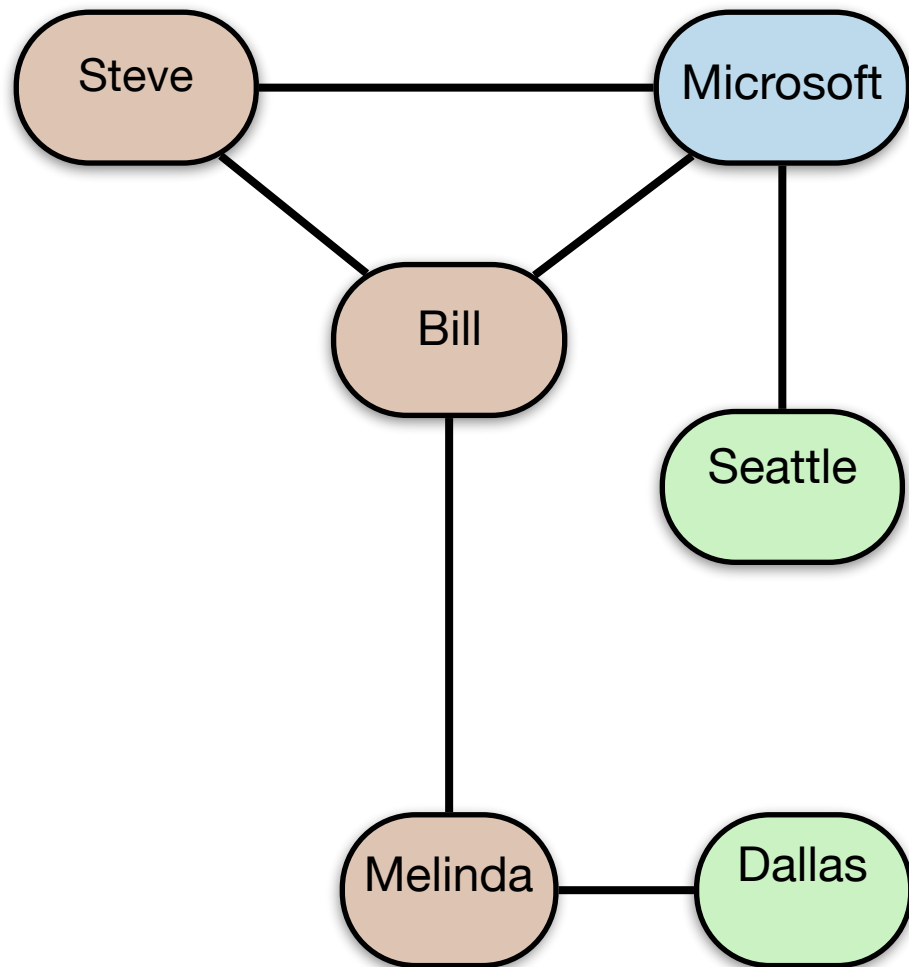
January 15, 2000

Tech pioneer **Bill Gates** stepped down today as chief executive officer of **Microsoft**, the **Seattle**-headquartered software giant. **He** will remain Chairman of **the company**, which rose to prominence after beating **Digital Research Inc** for the contract to provide an operating system for PCs. His long-time friend, **Steve Balmer**, will take over as CEO of **Microsoft**. **Gates** will now focus on the charitable foundation he runs with **his** wife **Melinda French Gates**. **Bill** and **Melinda** were married in a ceremony in **Hawaii**, rather than her hometown of **Dallas**. **Steve Balmer** was best man.

• Text → Mentions → Coref → Relations

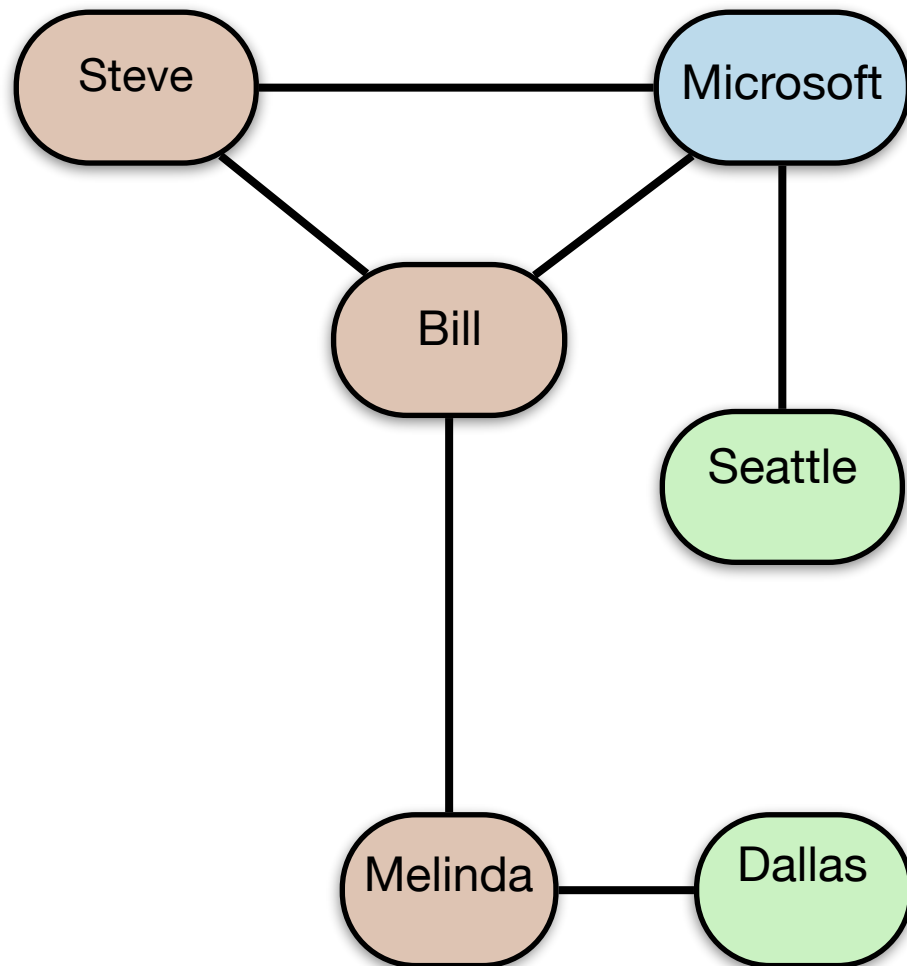


- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types



- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types

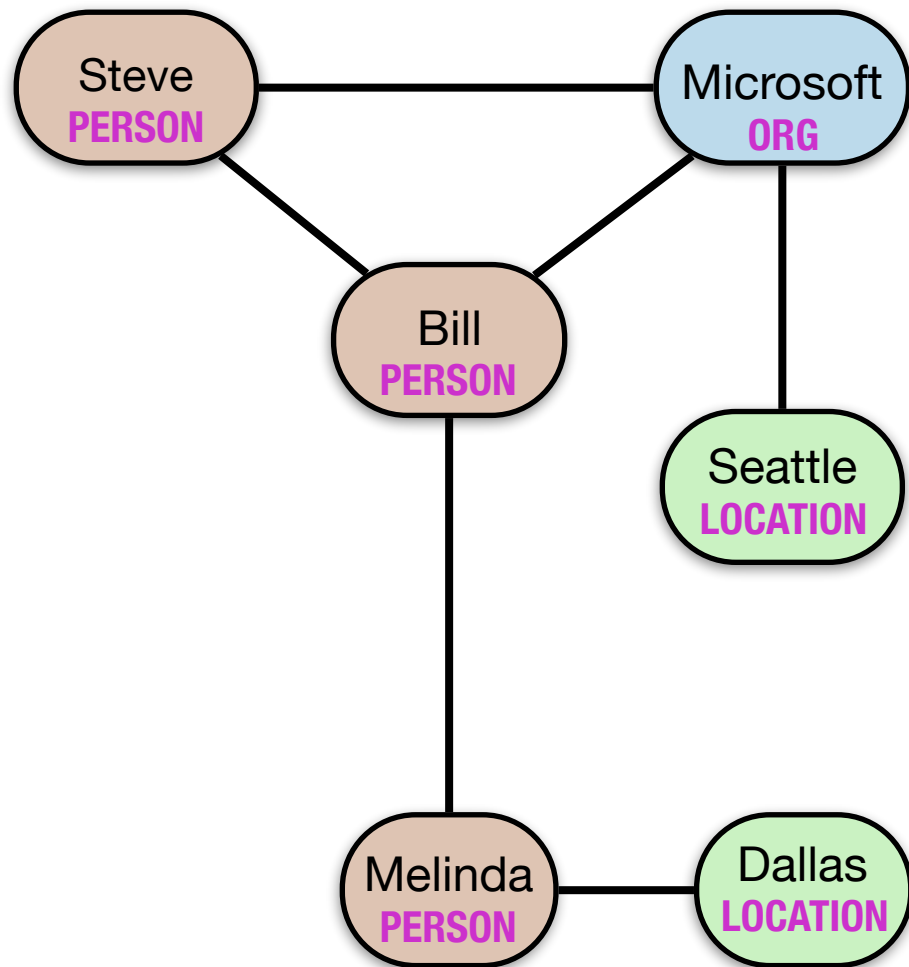
entity types



PERSON
LOCATION
ORG

- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types

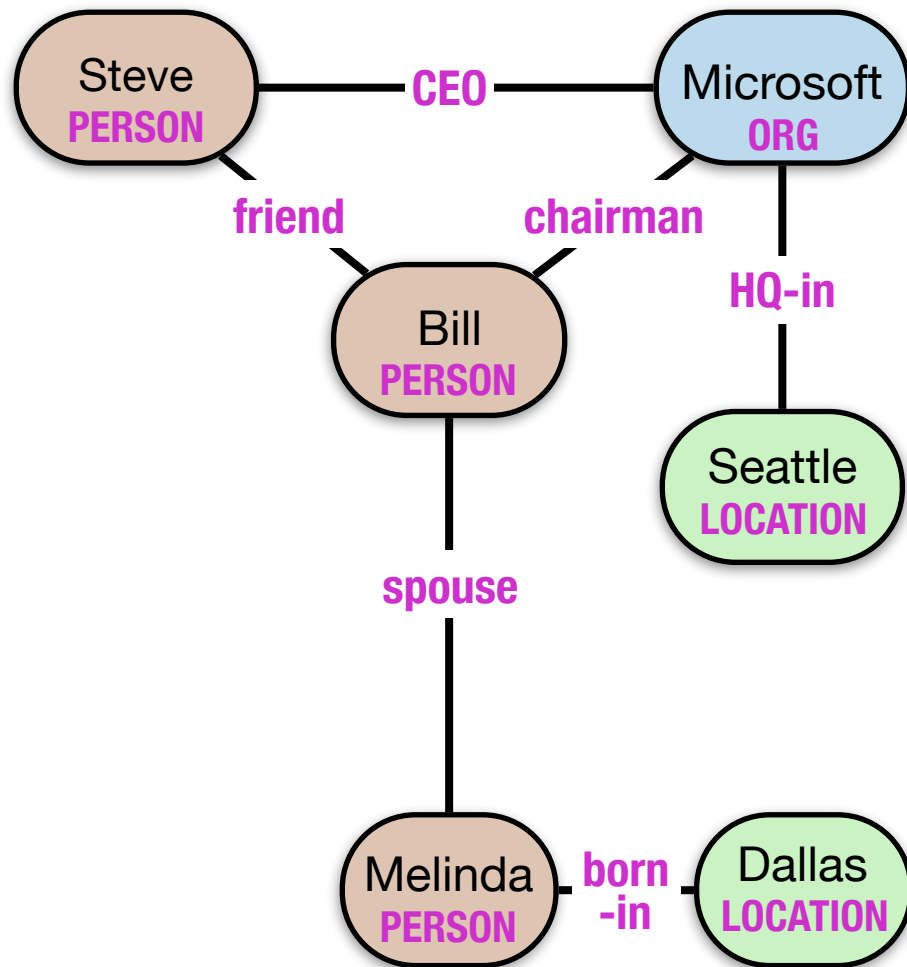
entity types



PERSON
LOCATION
ORG

- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types
 - Relation Types

entity types



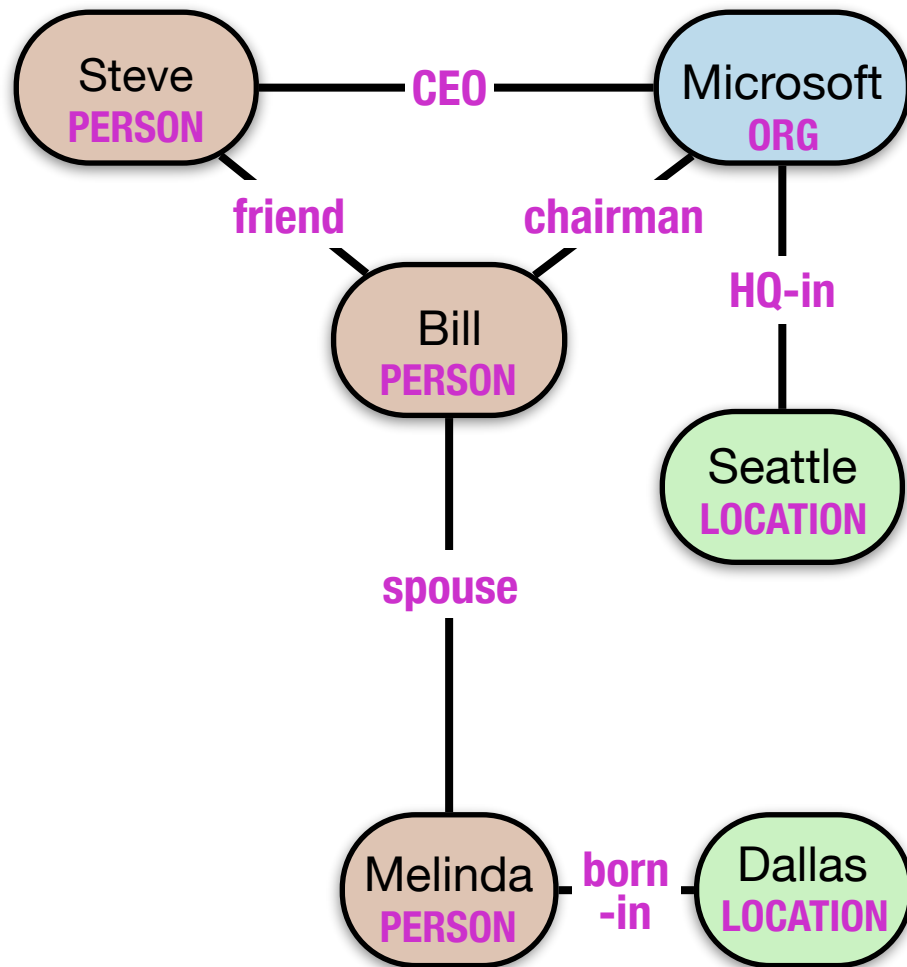
PERSON
LOCATION
ORG

relation types

CEO
friend
chairman
HQ-in
spouse
born-in

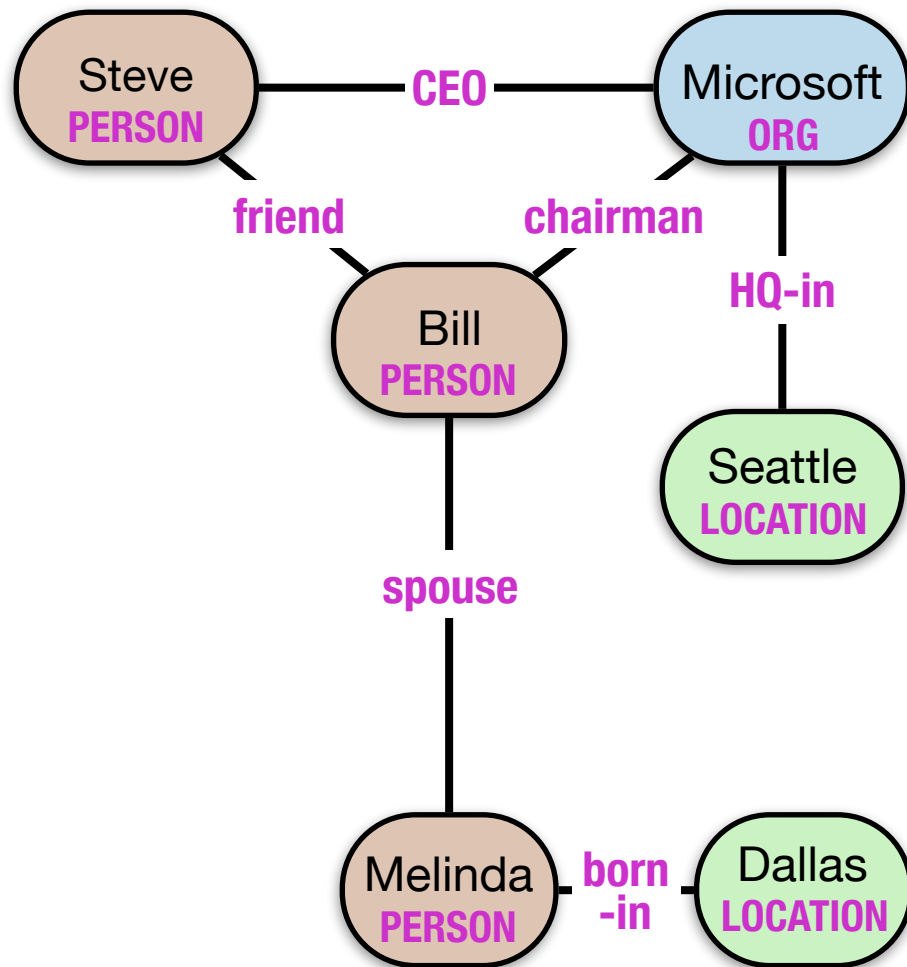
- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types
 - Relation Types

entity types



PERSON
LOCATION
ORG

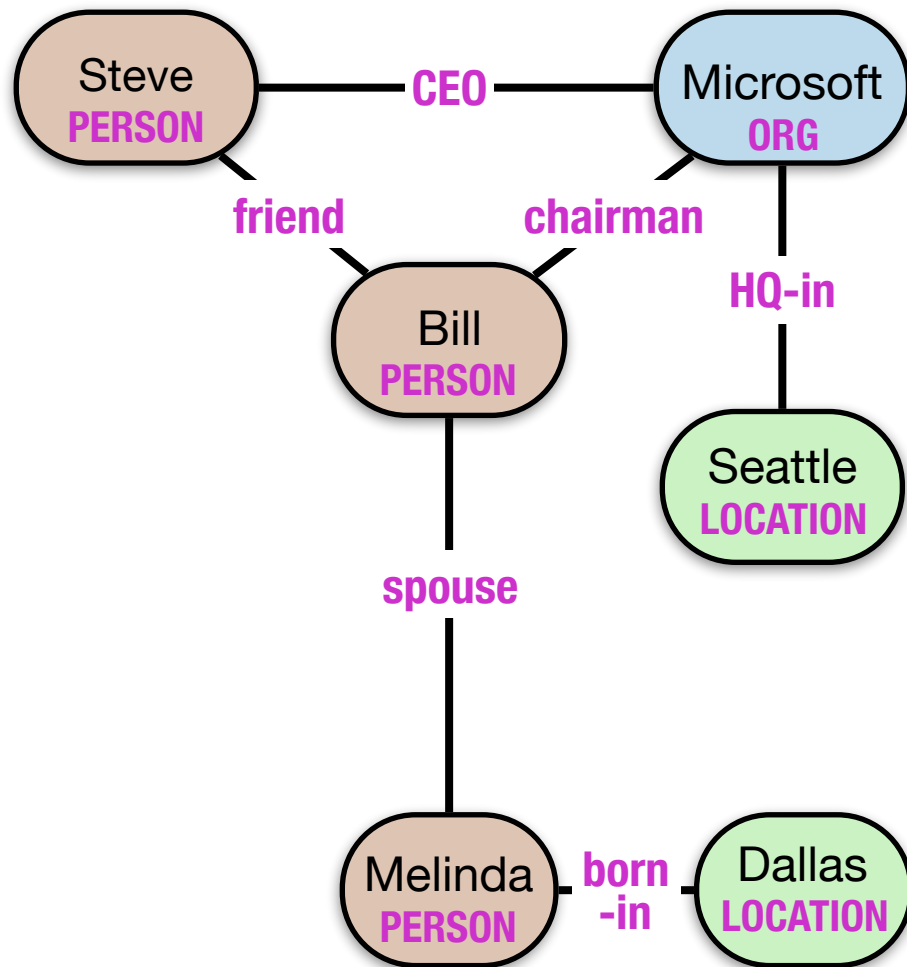
- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types
 - Relation Types



entity types

PERSON
LOCATION
ORG
/people/person
/film/subject
/location/city

- Text → Mentions → Coref → Relations
- Schema:
 - Entity Types
 - Relation Types



entity types

structured types

PERSON
LOCATION
ORG

unstructured types

/people/person
/film/subject
/location/city

company
software giant
charity

tech pioneer

programmer

executive

chairman

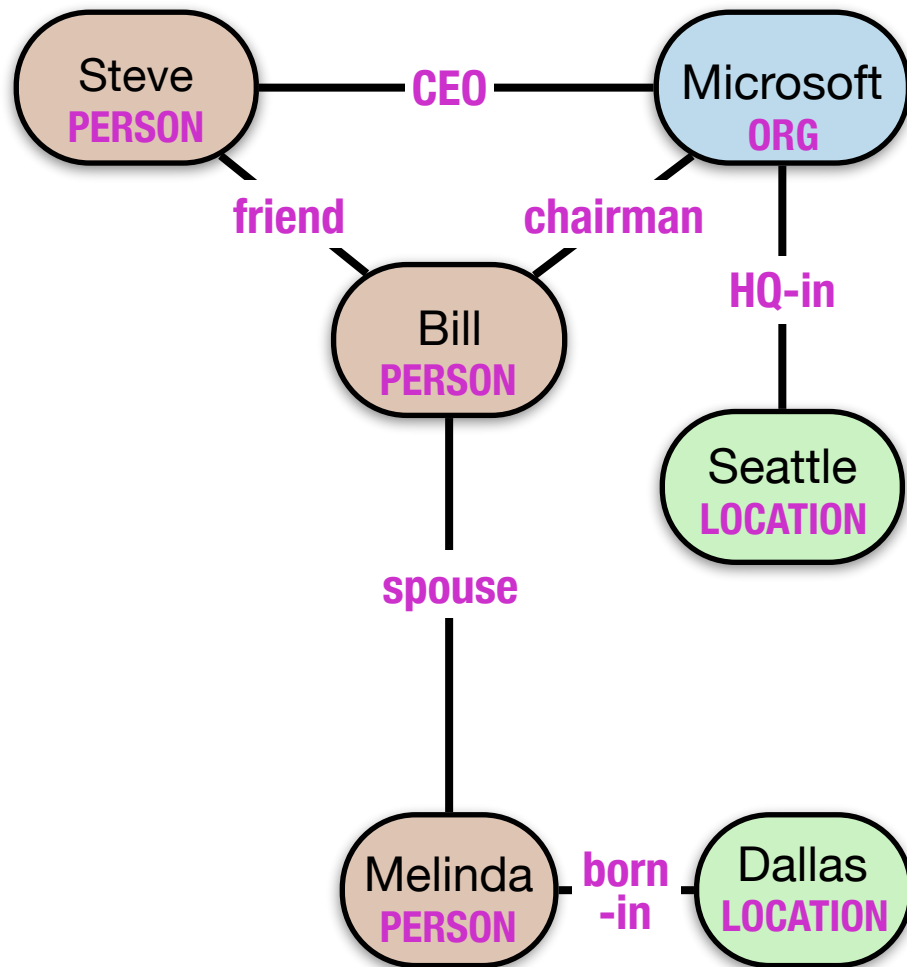
leader

50k
columns

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

Universal Schema entity types



structured types

unstructured types

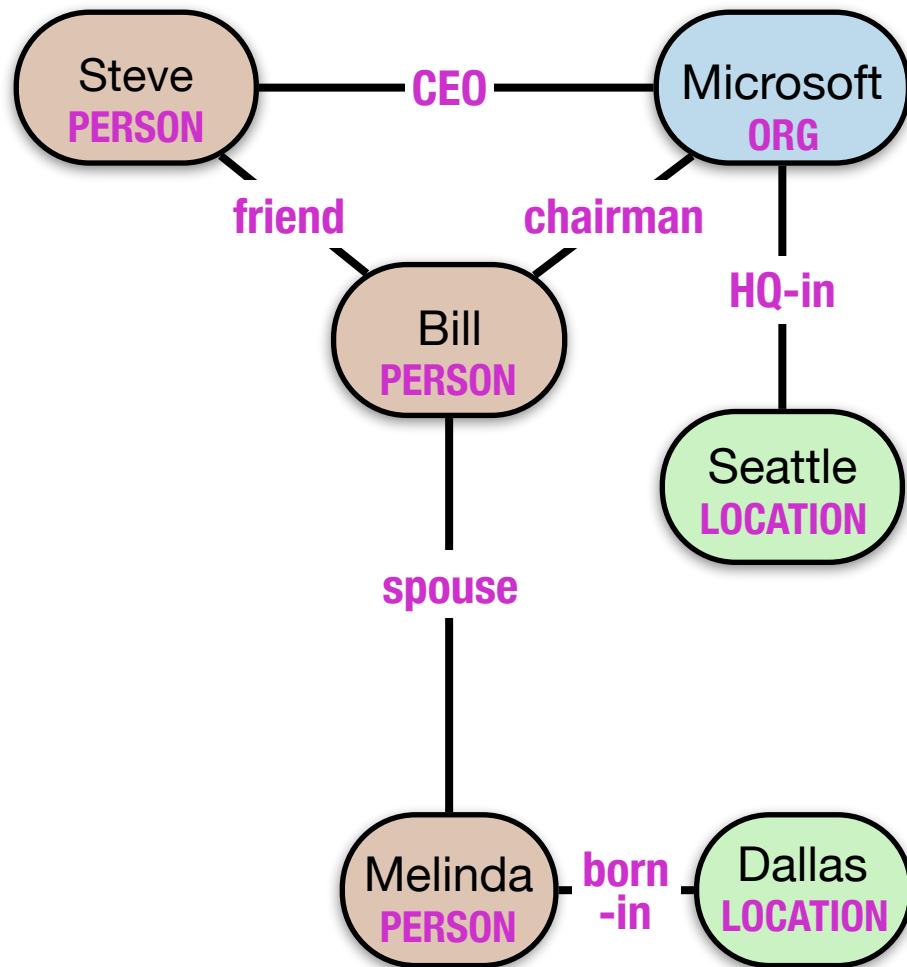
50k
columns

[illegible]

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

Universal Schema entity types



structured types

unstructured types

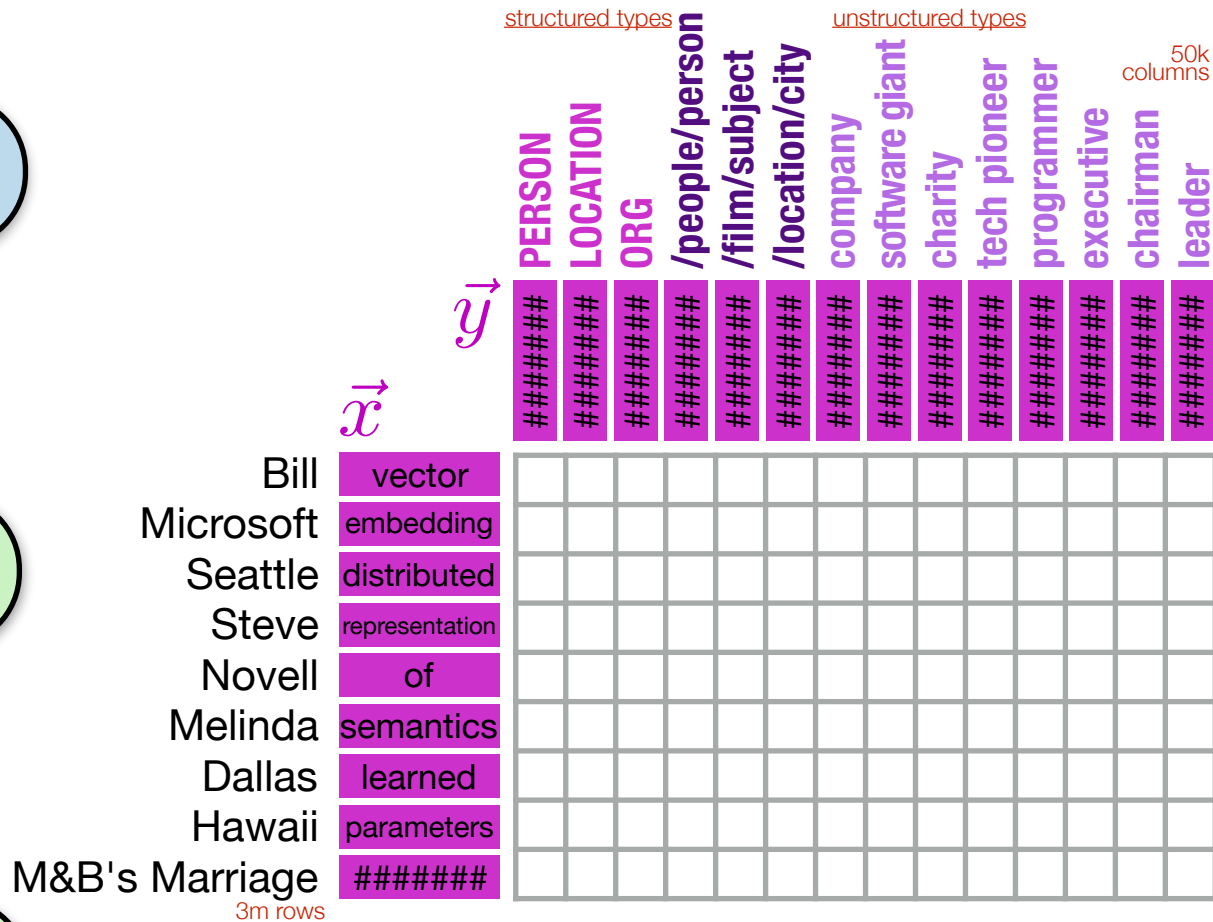
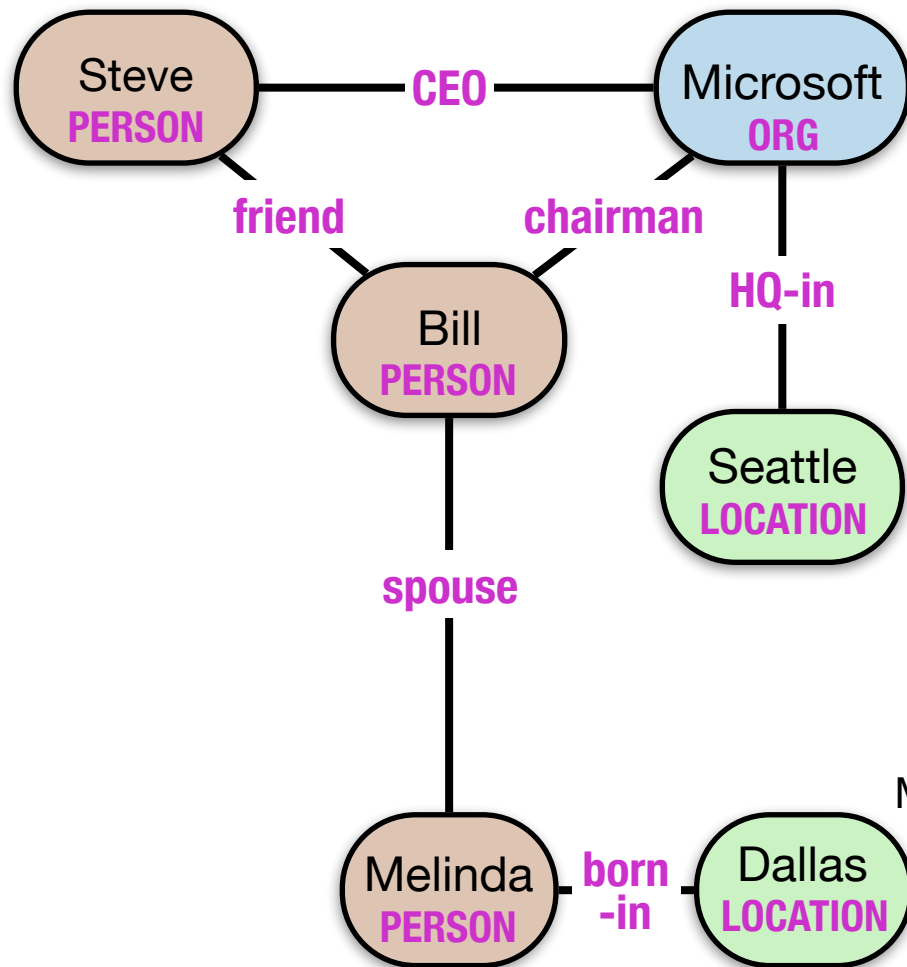
50k
columns

The diagram illustrates the difference between structured and unstructured data types. On the left, under the heading "structured types", there is a list of 14 specific categories: PERSON, LOCATION, ORG, /people/person, /film/subject, /location/city, company, software giant, charity, tech pioneer, programmer, executive, chairman, and leader. Each category is represented by a purple rectangular block containing the text "#####". On the right, under the heading "unstructured types", there is a single block labeled "50k columns", representing a large, unstructured dataset.

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

Universal Schema *entity types*



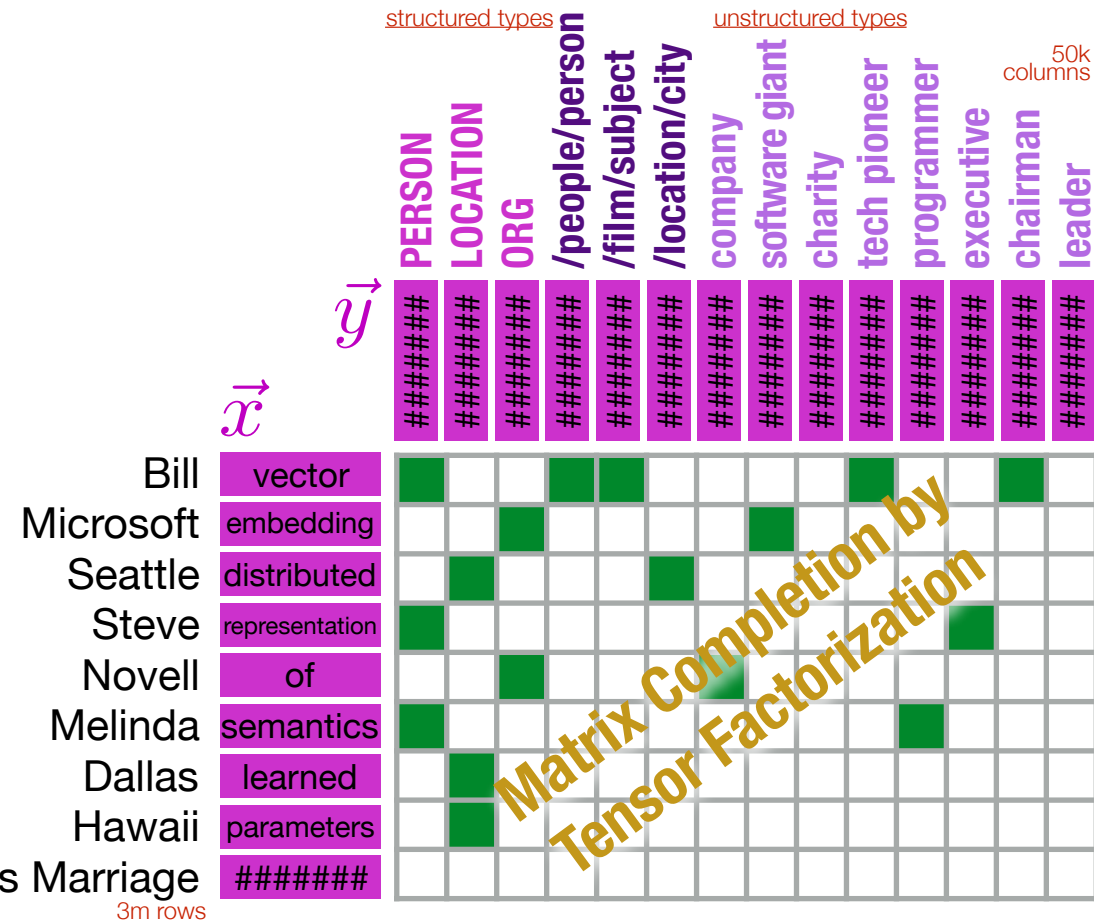
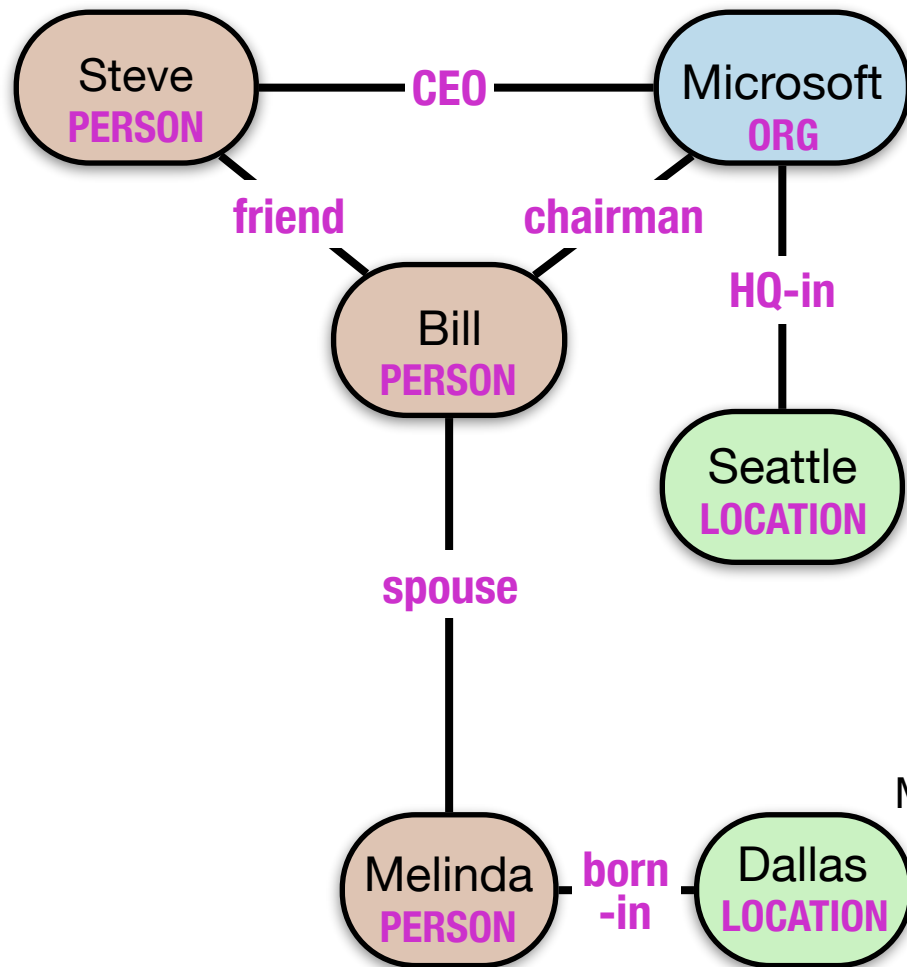
$$f_{e,t} = \sigma(\vec{x}_e \cdot \vec{y}_t)$$

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

Universal Schema *entity types*



$$f_{e,t} = \sigma(\vec{x}_e \cdot \vec{y}_t)$$

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$

- ## Universal Schema entity types

Universal Schema entity types

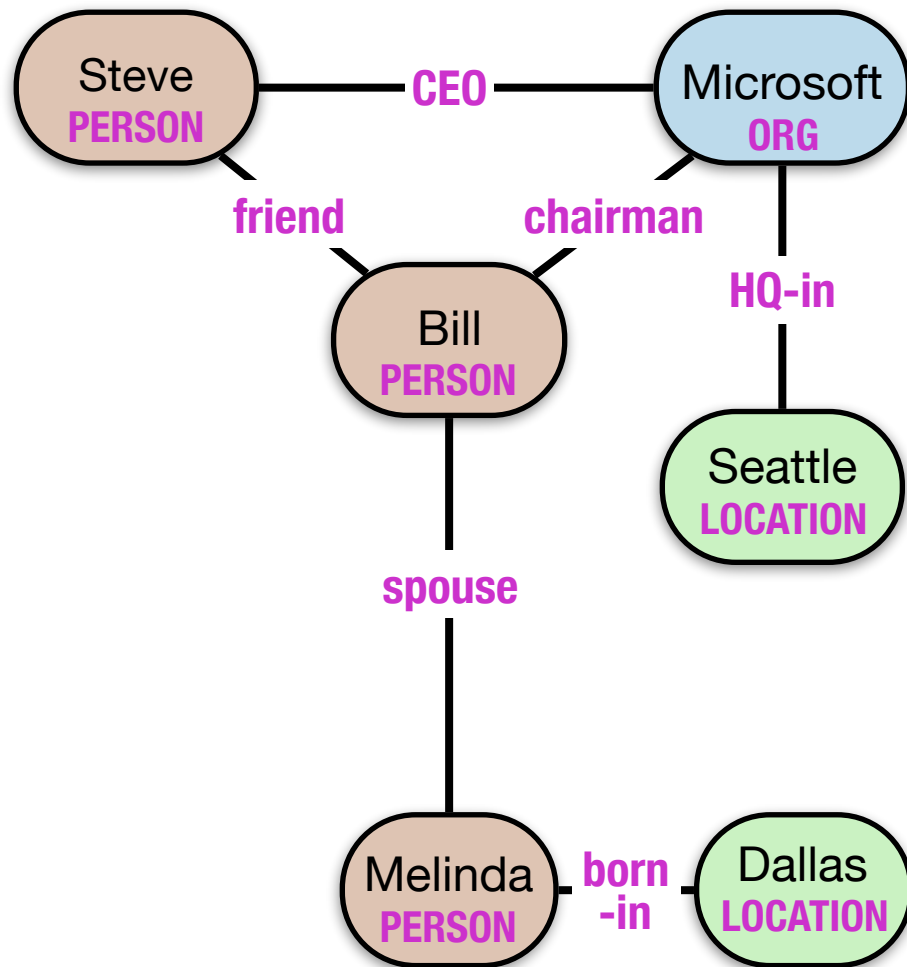

$$f_{e,t} = \sigma(\vec{x}_e \cdot \vec{y}_t)$$

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

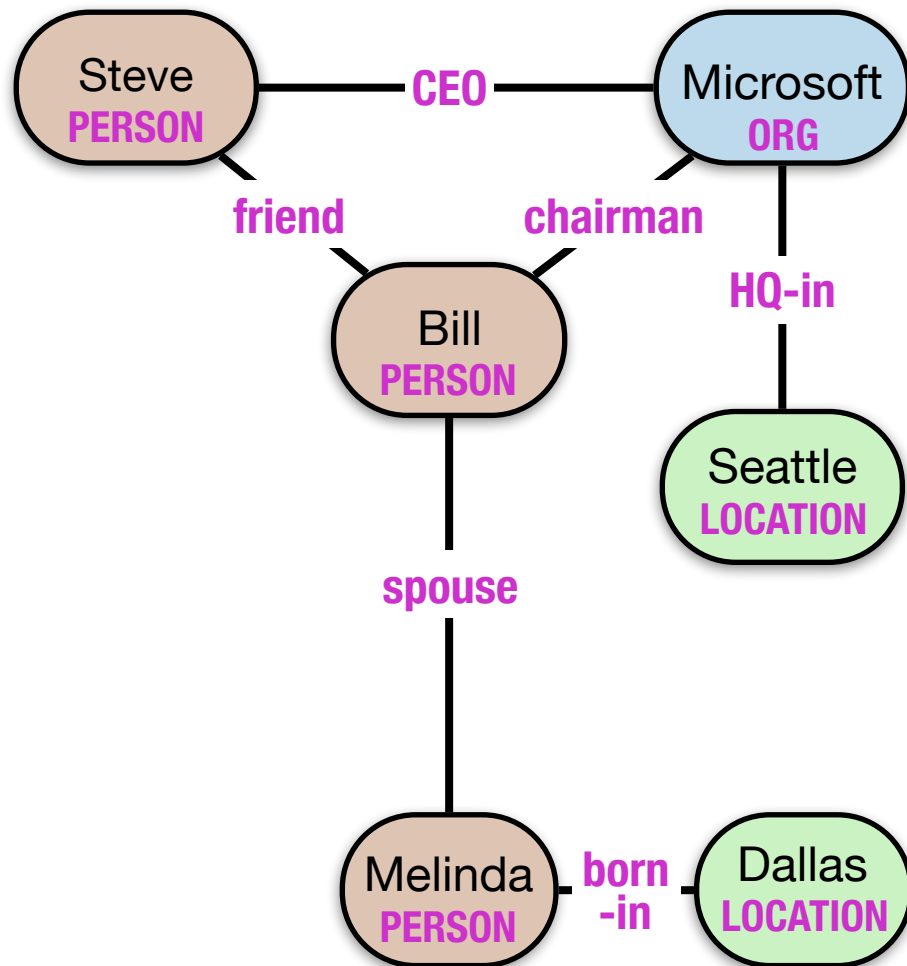
Universal Schema entity types



- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

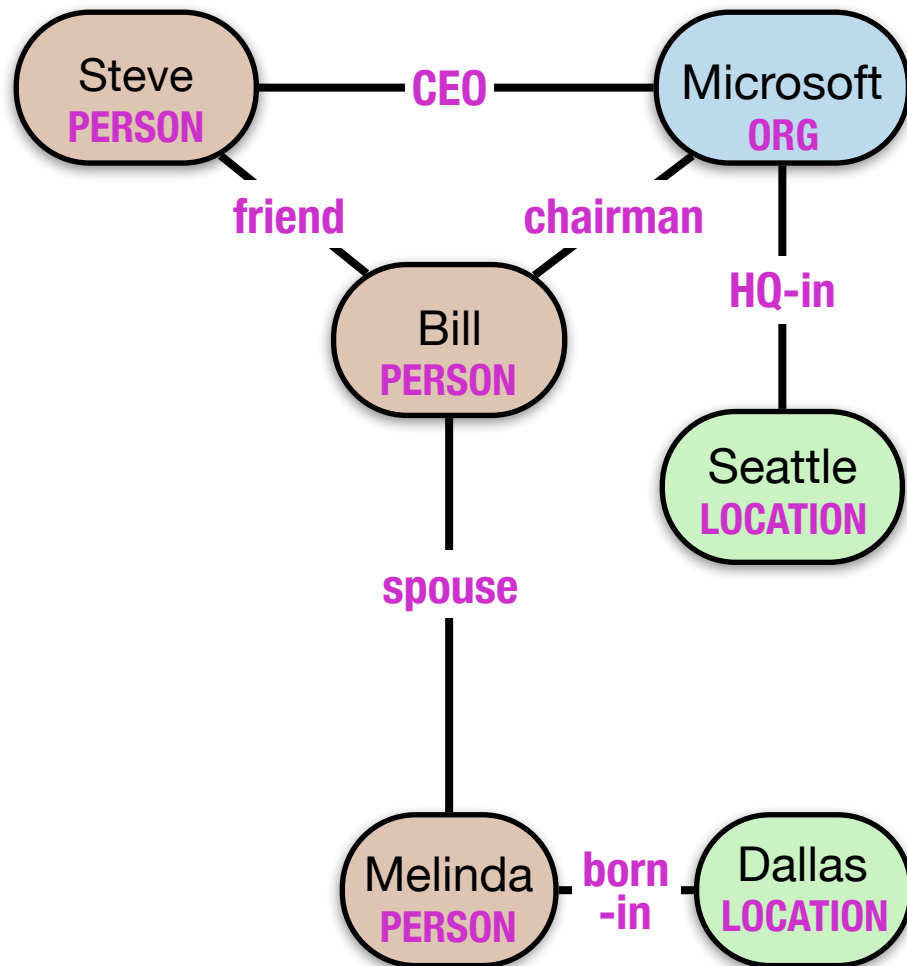
Universal Schema *relation types*



- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

[Riedel, Yao, Marlin, McCallum NAACL 2012]

Universal Schema *relation types*



spouse
born-in
friend

#####

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types

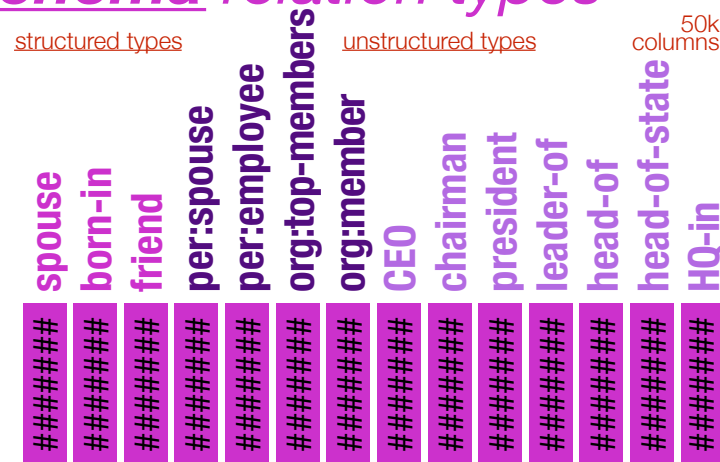
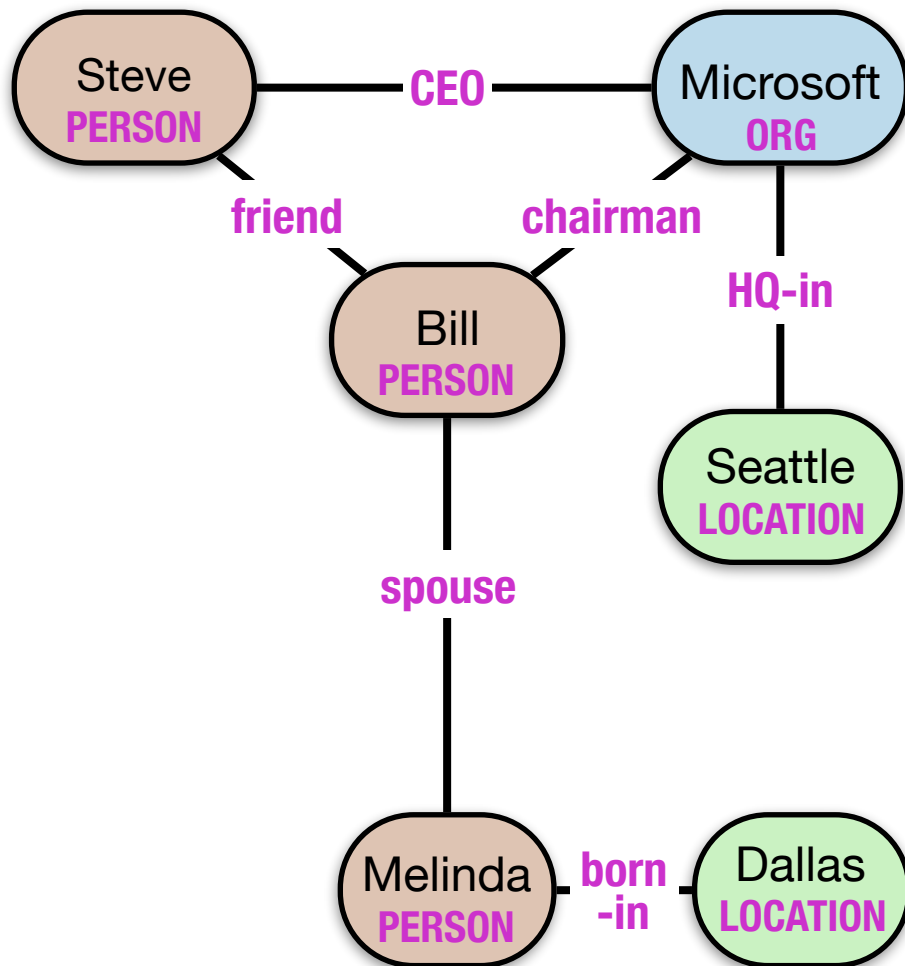
[Riedel, Yao, Marlin, McCallum NAACL 2012]

Universal Schema *relation types*

structured types

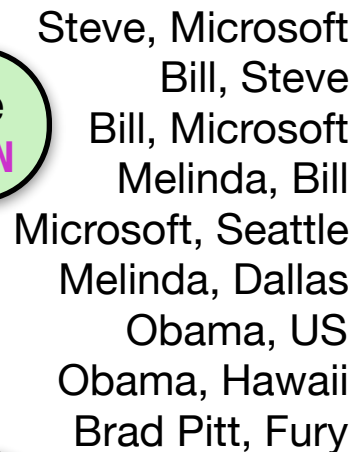
unstructured types

50k
columns



- ## Universal Schema relation types

Universal Schema relation types

 \vec{x} \bar{y}

unstructured types

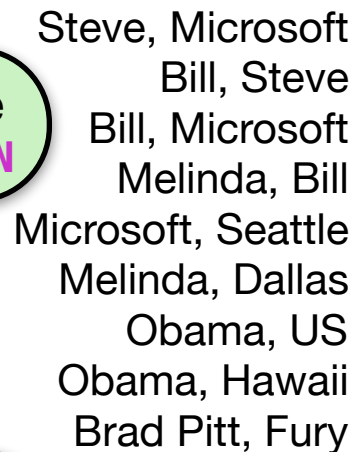
50k
columns

[illegible]

$$f_{e_1, e_2, r} = \sigma(\vec{x}_{e_1, e_2} \cdot \vec{y}_r)$$

- ## Universal Schema relation types

Universal Schema relation types

 \vec{x} \bar{y}

unstructured types

50k
columns

structured types

unstructured types

50k columns

spouse
born-in
friend
per:spouse
per:employee
org:top-members
org:member
CEO
chairman
president
leader-of
head-of
head-of-state
HQ-in

$$f_{e_1, e_2, r} = \sigma(\vec{x}_{e_1, e_2} \cdot \vec{y}_r)$$

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$

- ## Universal Schema relation types

Universal Schema relation types



50k
columns



$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$

“Universal Schema” Relation Types

	<subj< professor >prep >at>	<subj< historian >prep> at>
Kevin Boyle Ohio State		Y
R. Freeman Harvard	Y	

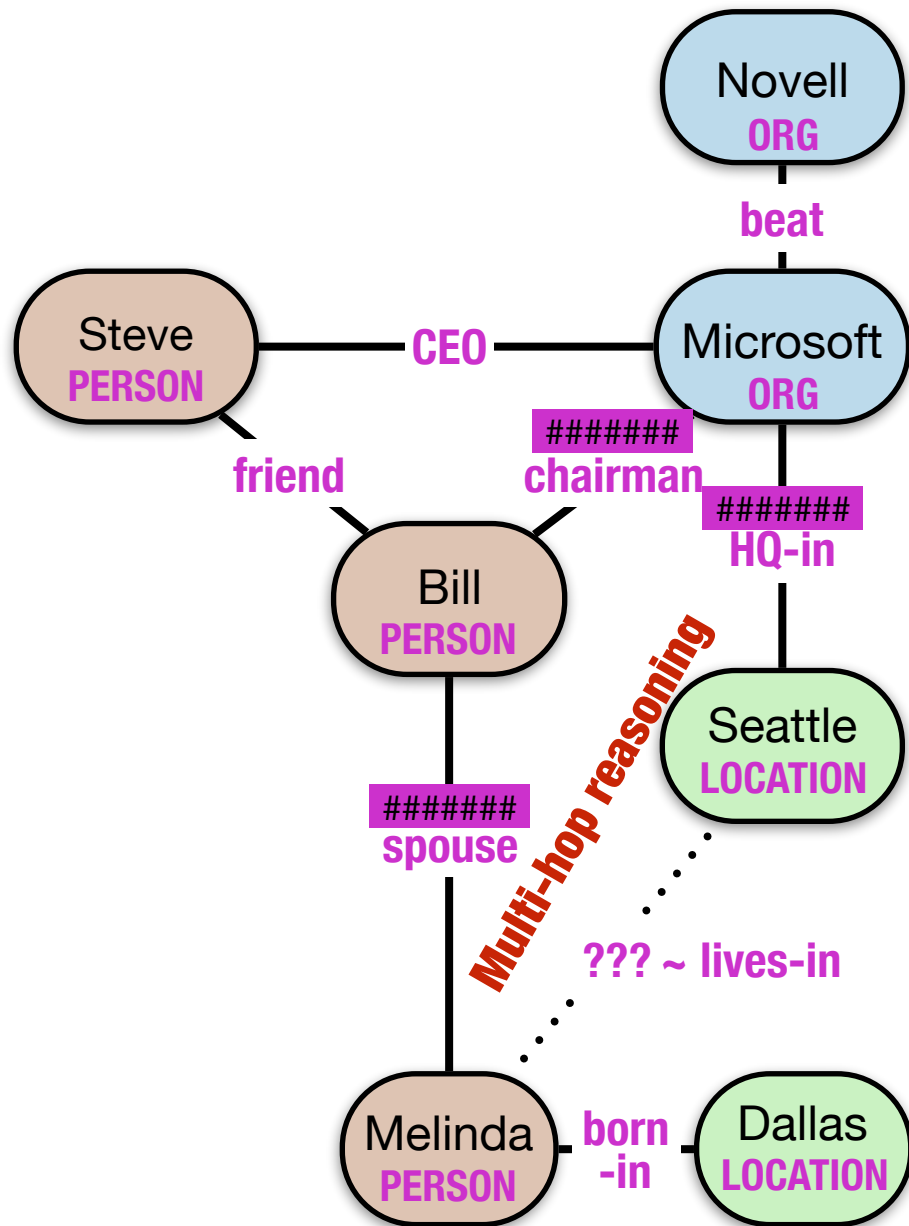
Learns asymmetric entailment:

PER historian at UNIV → PER professor at UNIV

but

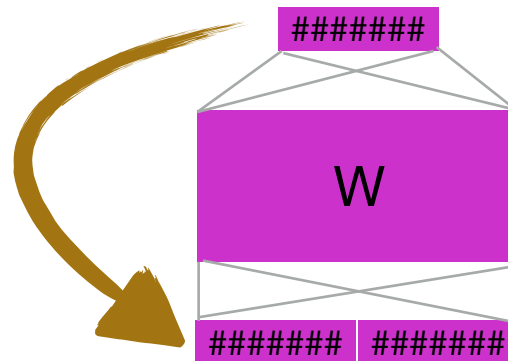
PER professor at UNIV ↛ PER historian at UNIV

- Text → Mentions → Coref → Relations
- Universal Schema: [AKBC 2012]
 - Entity Types
 - Relation Types
 - Implicature of implicit info



Spouse(A,B) & Chairman(B,C) & HQ-in(C,D) → Lives-in(A,D)
 Spouse(A,B) & CEO(B,C) & HQ-in(C,D) → Lives-in(A,D)
 Spouse(A,B) & COO(B,C) & HQ-in(C,D) → Lives-in(A,D)
 Child-of(A,B) & COO(B,C) & HQ-in(C,D) → Lives-in(A,D)

Chain of reasoning on vectors



Deep Recursive Neural Network

[Neelakantan, Roth, McCallum, 2015]

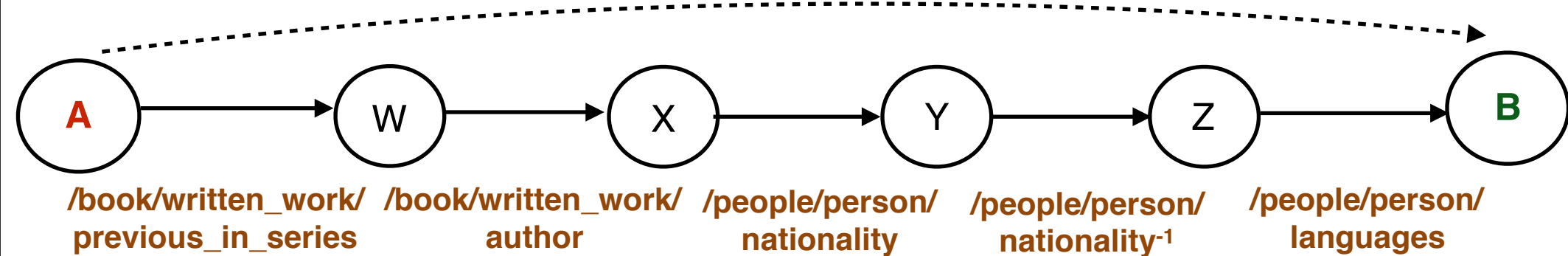
Data

<i>Entities</i>	18M
<i>Freebase triples</i>	40M
<i>ClueWeb triples</i>	12M
<i>Relation types</i>	25,994

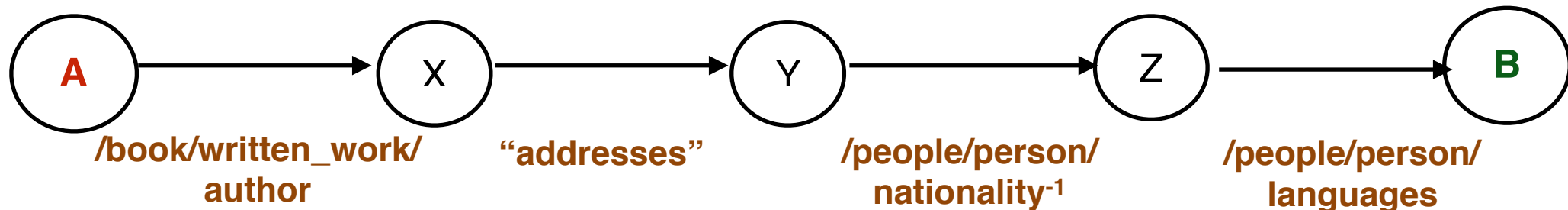
Predictive Paths

seen paths

`/book/written_work/original_language(A, B)`



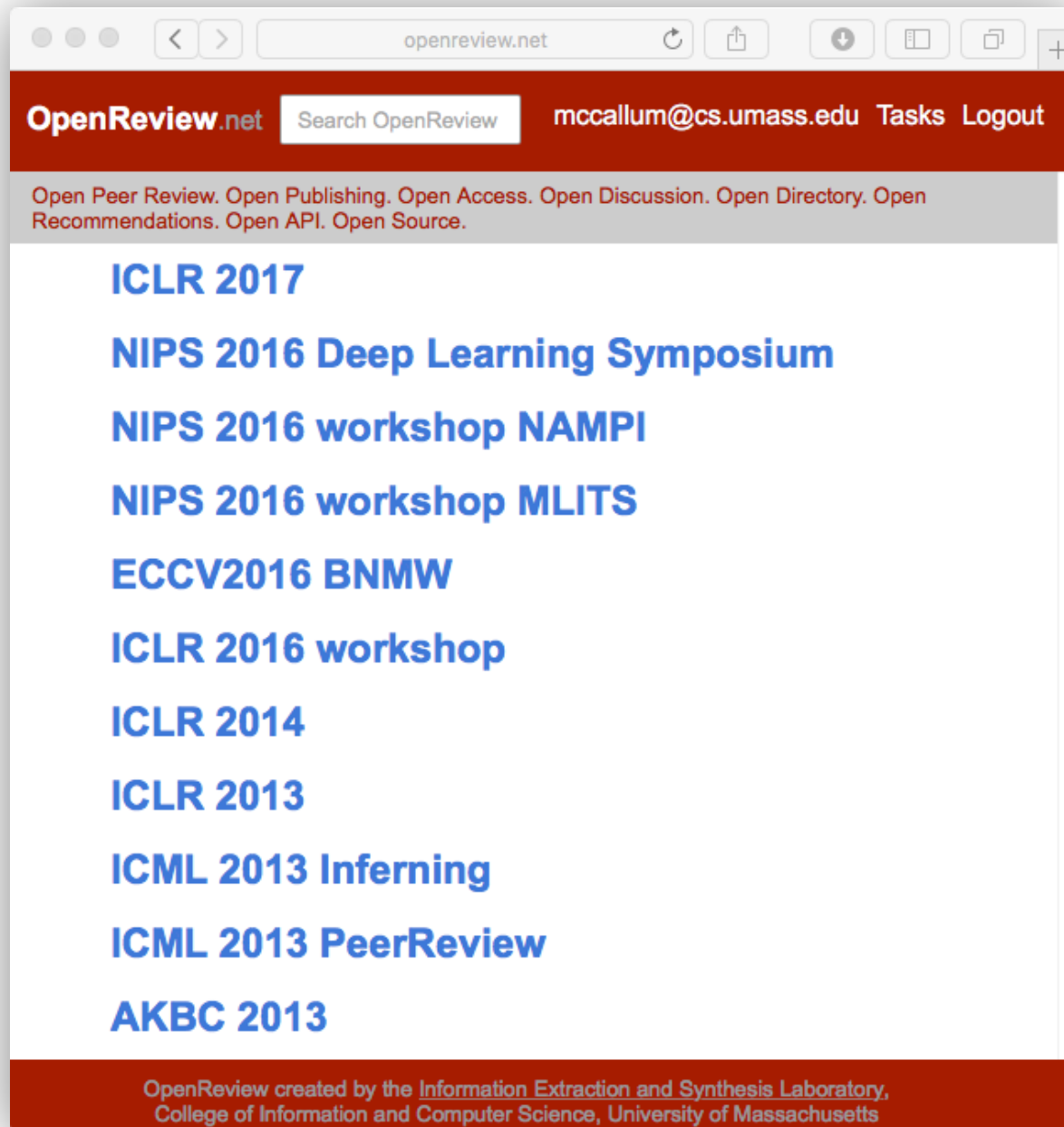
unseen paths



Applications & Collaborations

- OpenReview.net
- MIT Material Science
- US Patent Office
- Meta.com

OpenReview.net



- Backend API
- ICLR 2017
UAI 2017
...
- ArXiv overlay
- lightweight
"reviewing entities"
- Experimentation &
social science on
peer review culture

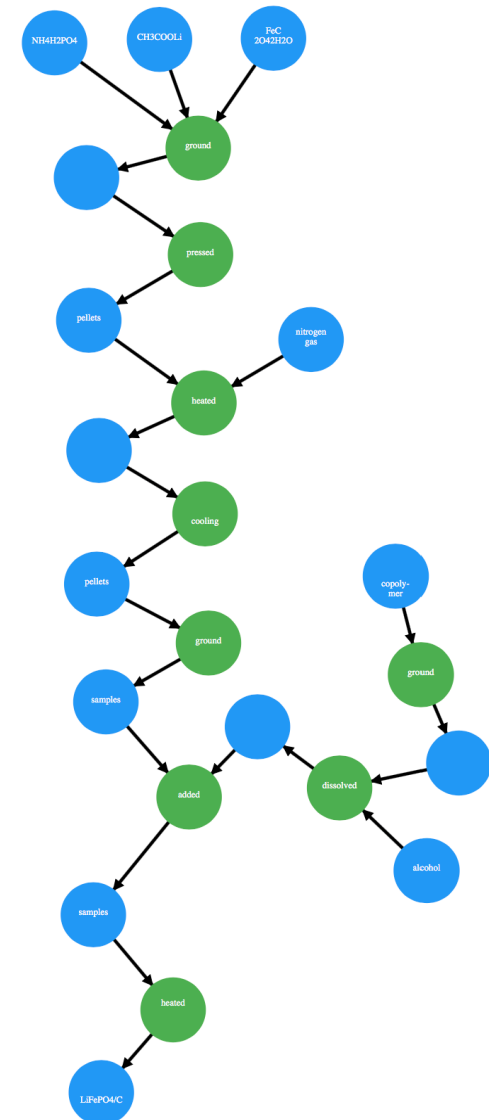
MIT Material Science

Recipe paragraphs
from 300k papers

→ Extracted recipe structure

2.1. Synthesis procedure

LiFePO_4 was synthesized from a stoichiometric mixture of reagent grade $\text{NH}_4\text{H}_2\text{PO}_4$ (Alfa-Aesar), CH_3COOLi (Aldrich), and $\text{FeC}_2\text{O}_4 \cdot 2\text{H}_2\text{O}$ (Aldrich) by a conventional solid-state reaction method. These materials were ground for 20 min, then pressed into pellets and heated at 623 K in a quartz-tube furnace with flowing nitrogen gas for 6 h. After slowly cooling to room temperature, pellets were ground again for 20 min and up to 6 wt.% copolymer (guluronic acid) was added to the samples. The guluronic acid powder was ground and dissolved in the alcohol solution. These samples were heated to 973 K at a heating rate of approximately 3 K min^{-1} and held at that temperature for 10 h in order to derive the LiFePO_4/C composite materials. After solid-state reaction, the total carbon content of LiFePO_4/C powder was measured by EA. These carbons were obtained from the synthesized precursors and guluronic acid.



New recipe ideas

USPTO PatentsView

Inventor Disambiguation Competition

The screenshot shows the USPTO PatentsView Beta search interface. At the top, there are links for API, Data Query, and Data Download. The main heading is 'PatentsView BETA' with a magnifying glass icon. Below this, a descriptive paragraph states: 'The PatentsView search tool allows audiences to interact with nearly 40 years of data on patenting activity in the US. Use the tool to explore technological, regional, and individual-level trends through several search filters and multiple view options.'

The search filters are organized into a grid. The top row is labeled 'VIEW RESULTS BY:' and includes four tabs: Patent (selected), Inventor, Assignee, and Class. Below this, the filters are arranged in two columns:

- Patent:** A text input field for 'title or number'.
- Inventor:** A text input field for 'first and/or last name'.
- Assignee, At-Issue:** A text input field for 'name'.
- USPC Patent Class:** A text input field for 'name or number'.
- Location, At-Issue:** Three input fields for 'country', 'state', and 'city'.
- Grant Date (1976–2016):** A text input field for 'yyyy, mm/yyyy, or range'.

At the bottom of the filter section, there are two buttons: 'View as a List' and 'View as a Map'.

Below the search filters, there is a section titled 'Recent Updates from the PatentsView Team' with three bullet points:

- The PatentsView database has been updated through July 15, 2016. All respective data are accessible through the [web tool](#), [API](#), and [bulk downloads](#).
- [World Intellectual Property Office \(WIPO\) technology fields](#) are now integrated into the database and can be retrieved through the [API](#) and [bulk downloads](#).
- The updated database features two new data fields – U.S. government organization name and contract and award numbers – both extracted from the government interest statement on

UMass: 1st place. Deploying at USPTO.



**Mission - Organize and Deliver All of the World's
Scientific and Technical Information.**

Founded in 2010 • Team of 25+ • Venture Backed

Toronto (HQ) • San Francisco • Montreal

**MIT
Technology
Review**

Bloomberg

WIRED

TheScientist

TECHVIBES

 **OUTSELL**

YAHOO!

VentureBeat

 **TechCrunch**

Large Commercial STEM Text-Mining Collection

37 38K 28M+14M

Major STM
Publishers

Serial Titles
(Books & Journals)

Closed Access
Full-Text Articles

Open Access
Full-Text Articles

Meta^α's Scientific Knowledge Graph:

3.5B

recommendations

1B

paper-concept
matches

422M

citations

26M

papers

16M

genetic elements

20M

concepts

14.5M

researchers

2M

antibodies

407K

drugs

427K

institutes

234K

bacteria

96K

diseases

85K

products

36K

journals

4.6B

Knowledge Graph
connections

Summary

- Building and leveraging knowledge bases for science
- **Representation**
 - Knowledge graph: entities & relations: (nodes) (edges)
 - ~~symbols~~ → *universal schema* vector embeddings (on nodes & edges)
 - Reasoning by RNN paths through network.
 - Next: efficient search for scientific reasoning by RL through this graph
- **Applications**
 - OpenReview.net (+ KB of all researchers, expertise, career path)
 - MIT Material Science
 - USPTO Patent Inventor Disambiguation
 - Meta.com