# Querying unNorMaLiZED and Inc_mpl_te Knowledge Bases

Percy Liang

Stanford University


Automated Knowledge Base Construction (AKBC) 2016

June 17, 2016

# Computing the answer

*What is the second most populous city in California?*

# Computing the answer

*What is the second most populous city in California?*

semantic parsing

$\text{argmax}(\text{Type.City} \sqcap \text{ContainedBy.CA}, \text{Population}, 2)$

# Computing the answer

*What is the second most populous city in California?*

semantic parsing

$\text{argmax}(\text{Type.City} \sqcap \text{ContainedBy.CA}, \text{Population}, 2)$

execute

San Diego

# Computing the answer

*Which states' capitals are also their largest cities by area?*

# Computing the answer

*Which states' capitals are also their largest cities by area?*

semantic parsing

$\mu x.\text{Type.USState} \sqcap \text{Capital.argmax}(\text{Type.City} \sqcap \text{ContainedBy}.x, \text{Area})$

# Computing the answer

*Which states' capitals are also their largest cities by area?*

semantic parsing

$\mu x$.Type.USState ⊓ Capital.argmax(Type.City ⊓ ContainedBy.$x$, Area)

execute

Arizona,Hawaii,Idaho,Indiana,Iowa,Oklahoma,Utah

# Computing the answer

*Which states' capitals are also their largest cities by area?*

semantic parsing

$\mu x.$Type.USState $\sqcap$ Capital.argmax(Type.City $\sqcap$ ContainedBy.$x$, Area)

execute

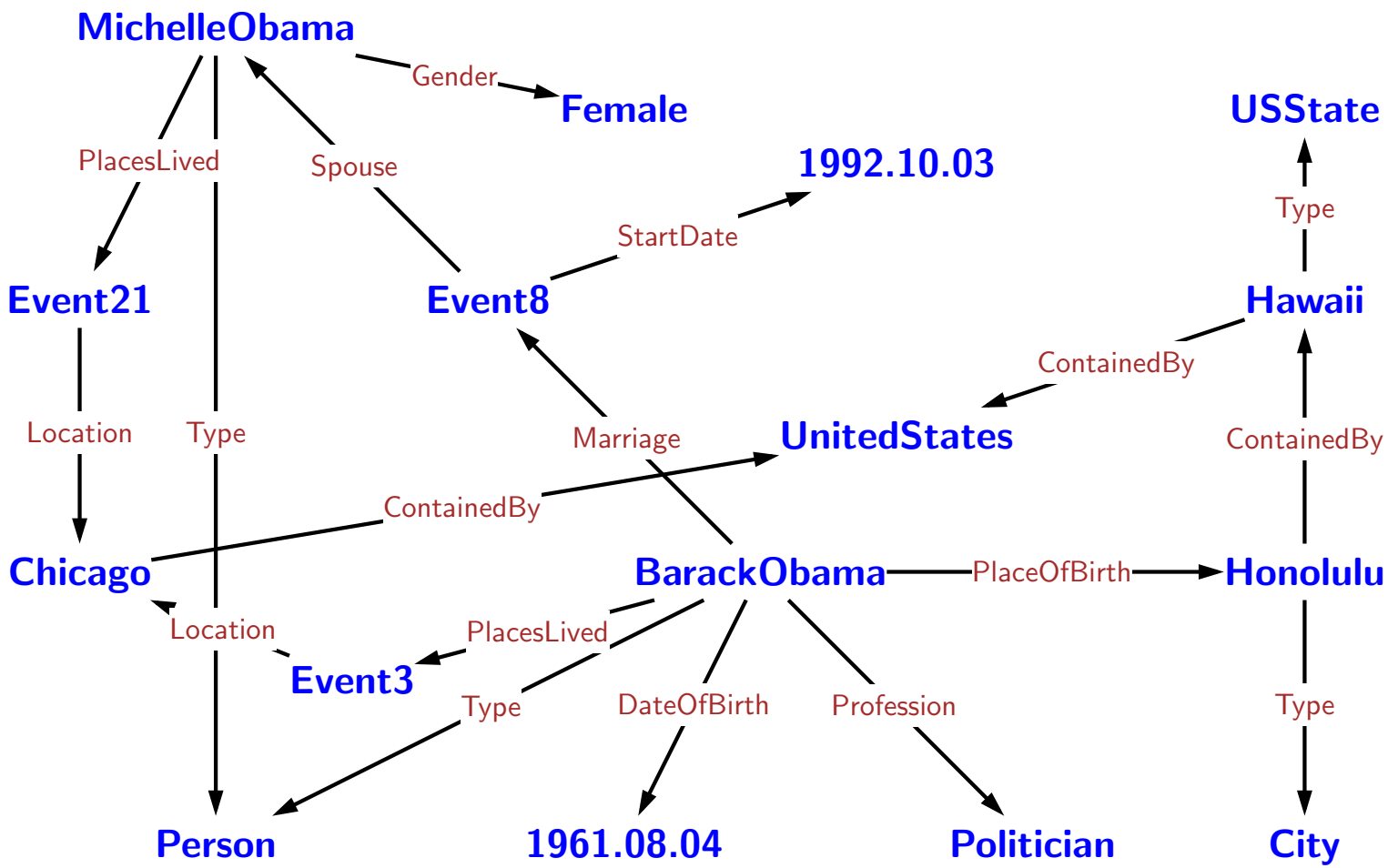
Arizona,Hawaii,Idaho,Indiana,Iowa,Oklahoma,Utah

**Strongly leverages KB structure!**

[Bollacker, 2008; Google, 2013]

# Freebase

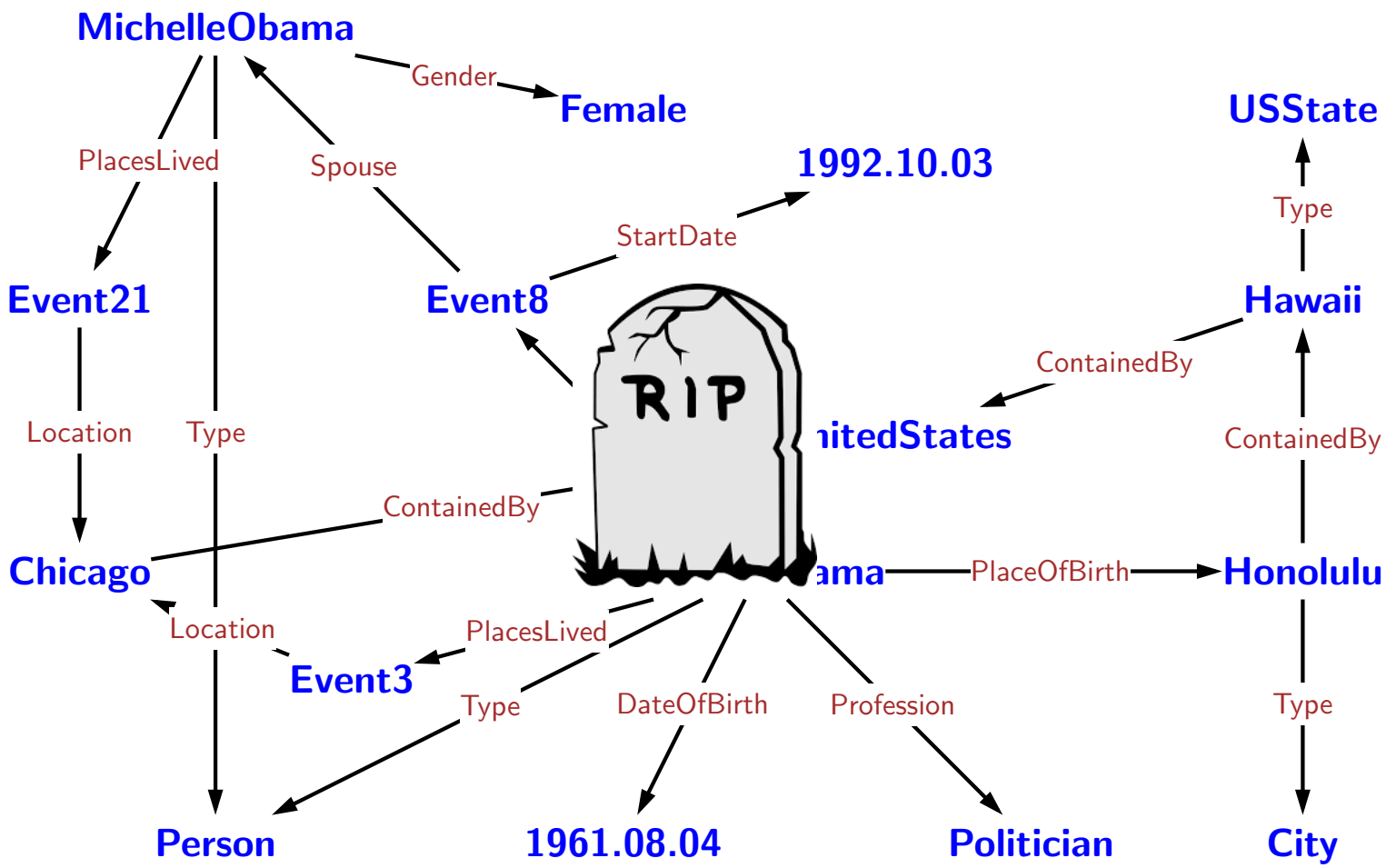
100M entities (nodes)  1B assertions (edges)



[Bollacker, 2008; Google, 2013]

# Freebase

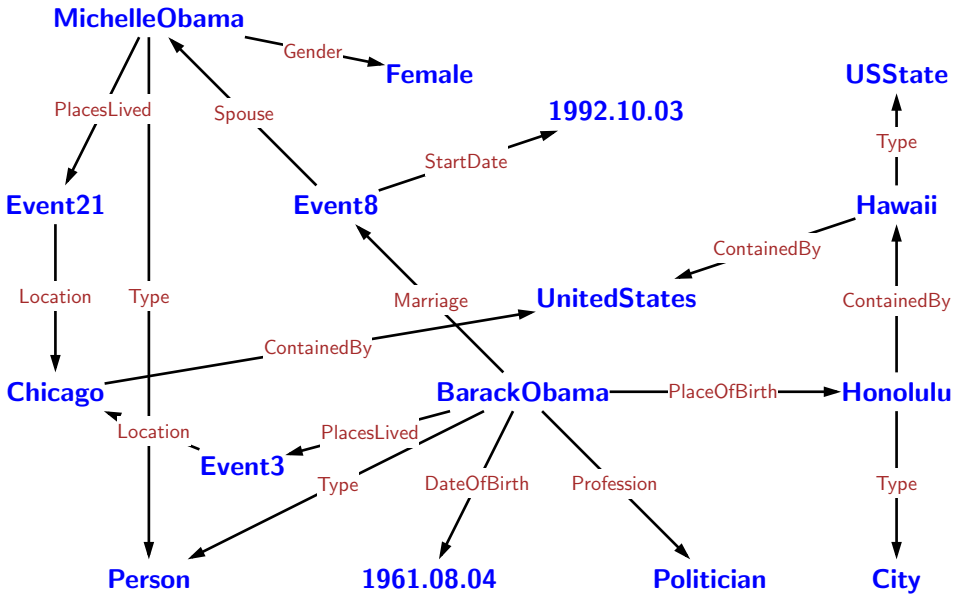
100M entities (nodes)  1B assertions (edges)



*hiking trails near Palo Alto*

*dishes at Oren's Hummus*

*ACL 2014 papers*



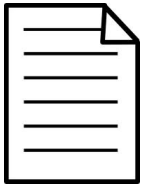
*hiking trails near Palo Alto*

*dishes at Oren's Hummus*

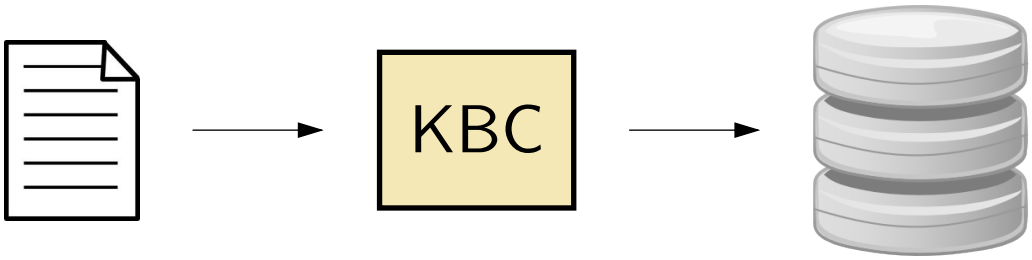
*ACL 2014 papers*

Fewer than 10% of WebQuestions answerable via Freebase

# Obtaining better KBs



# Obtaining better KBs



# Obtaining better KBs

# Obtaining better KBs


# Philosophy

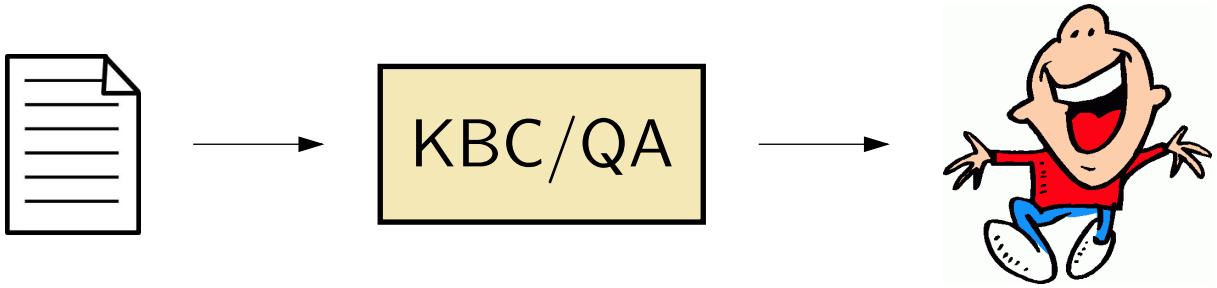
Focus on the **end-to-end** task of question answering.

Let that end goal drive learning and construction of **intermediate** knowledge representations.

# Philosophy

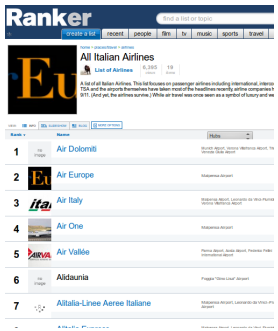
Focus on the **end-to-end** task of question answering.

Let that end goal drive learning and construction of **intermediate** knowledge representations.



# Outline

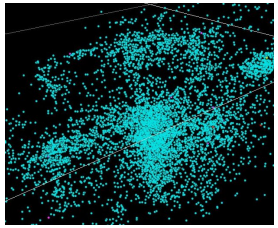
**On web pages**




On tables

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | | | 1st | Medley relay | 1:50.46 |
| | European Junior Championships | Grosseto, Italy | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

In vector space



# Simple semantic parsing on web pages

Panupong (Ice) Pasupat

ACL 2014

# Semantic parsing on the web

Input:

- query $x$

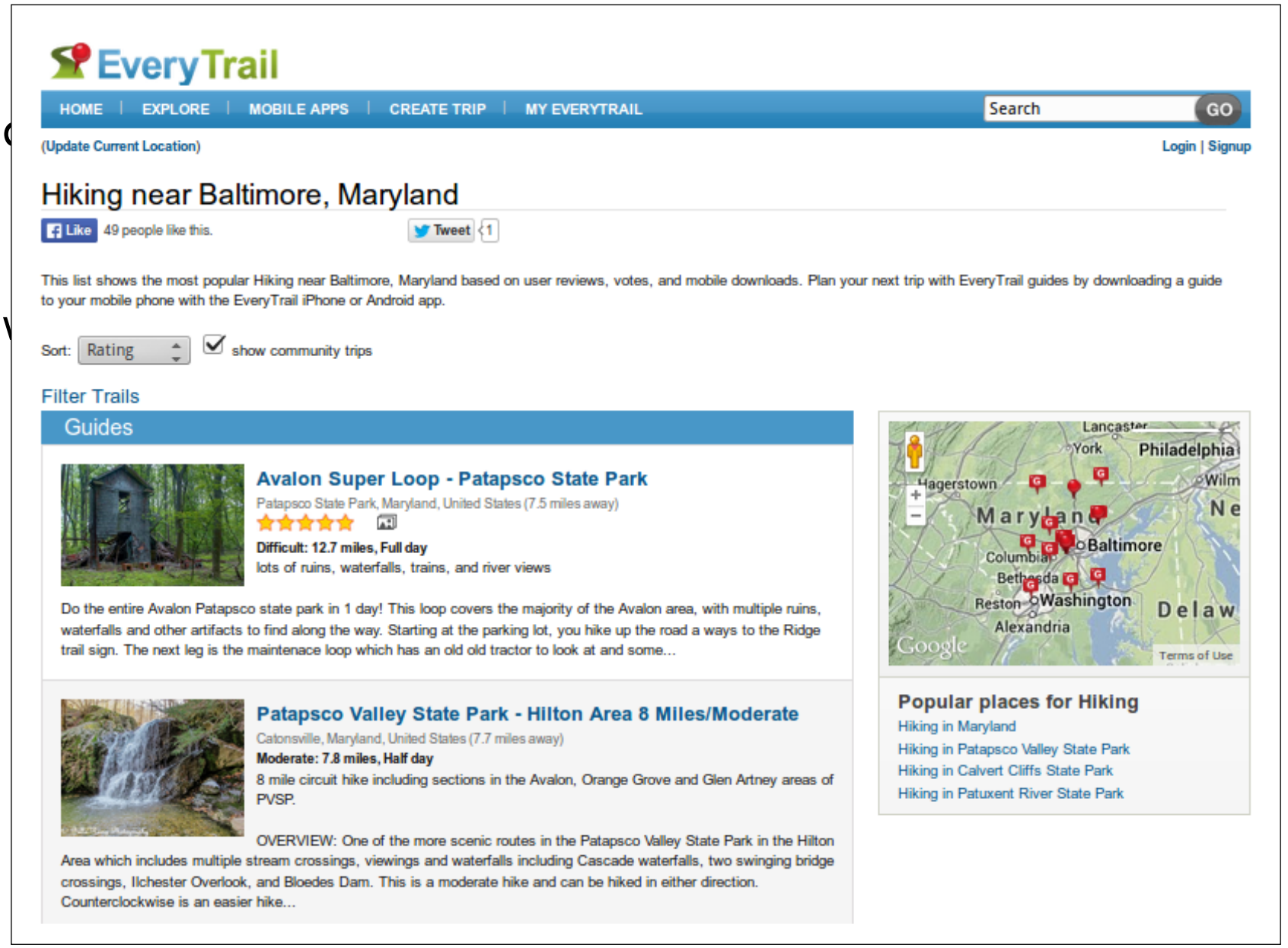
*hiking trails near Baltimore*

- web page $w$

# Semantic parsing on the web

Input:

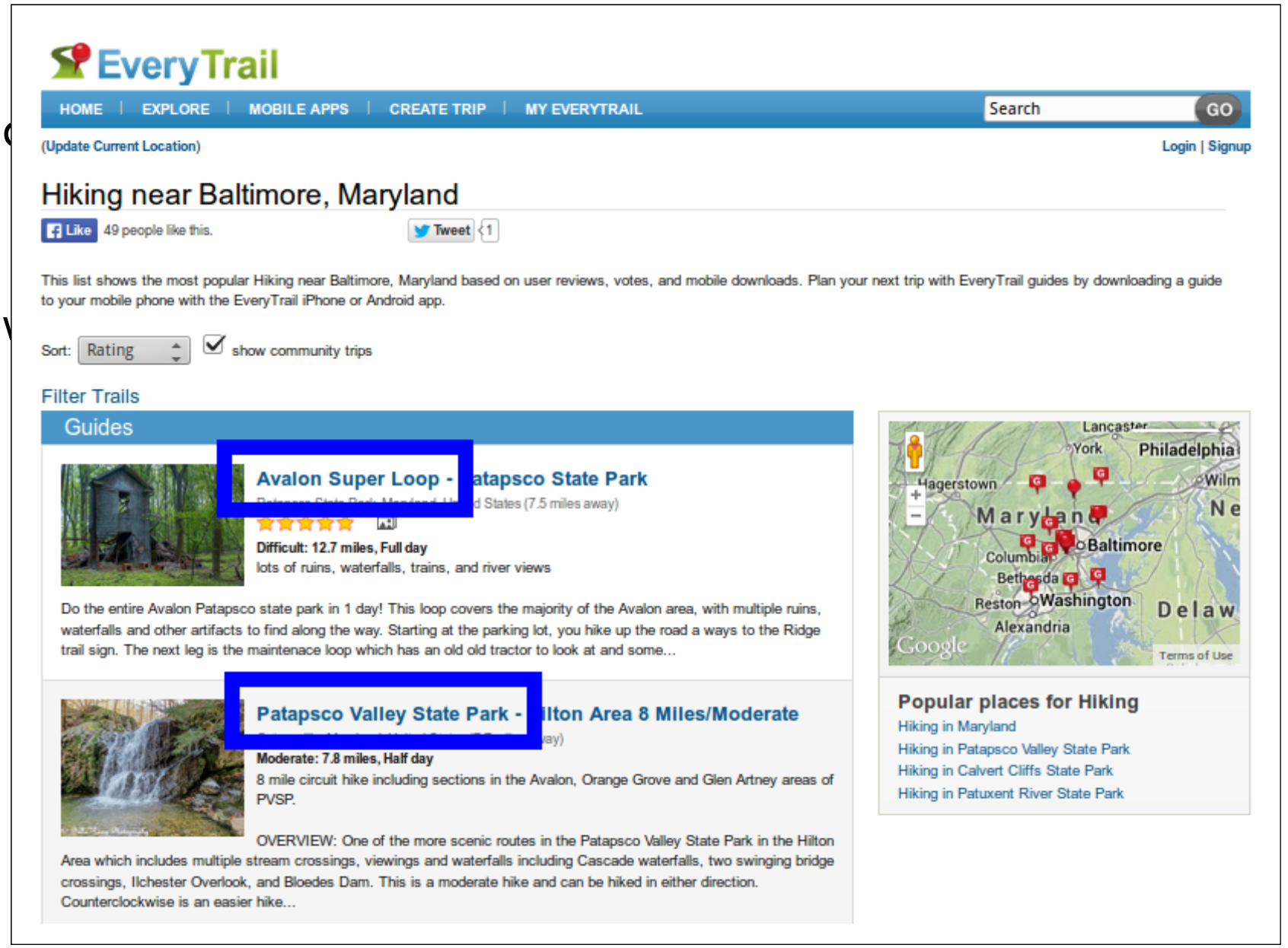
- C
- V



# Semantic parsing on the web

Input:

- C
- V



# Semantic parsing on the web

Input:

- query $x$

*hiking trails near Baltimore*

- web page $w$

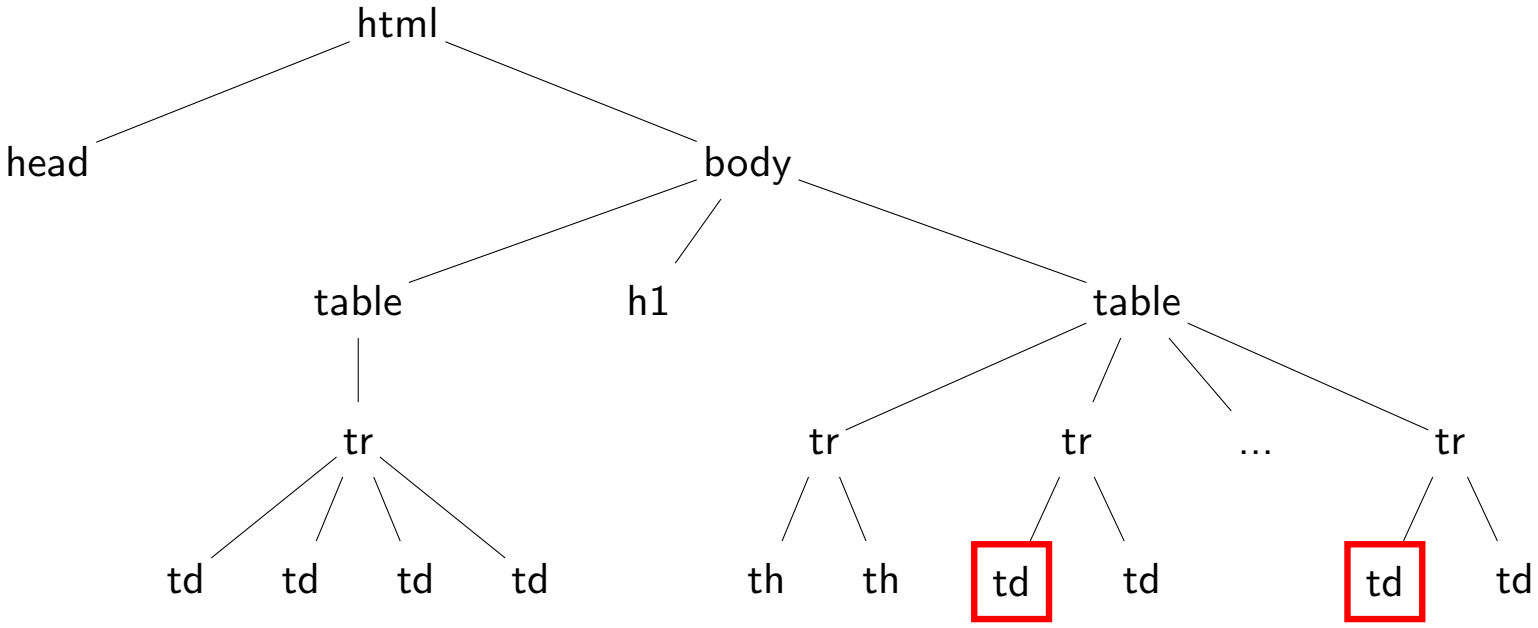
Output:

- list of entities $y$

[Avalon Super Loop, Patapsco Valley State Park, ...]

[Sahuguet and Azavant, 1999; Liu et al., 2000; Crescenzi et al., 2001]

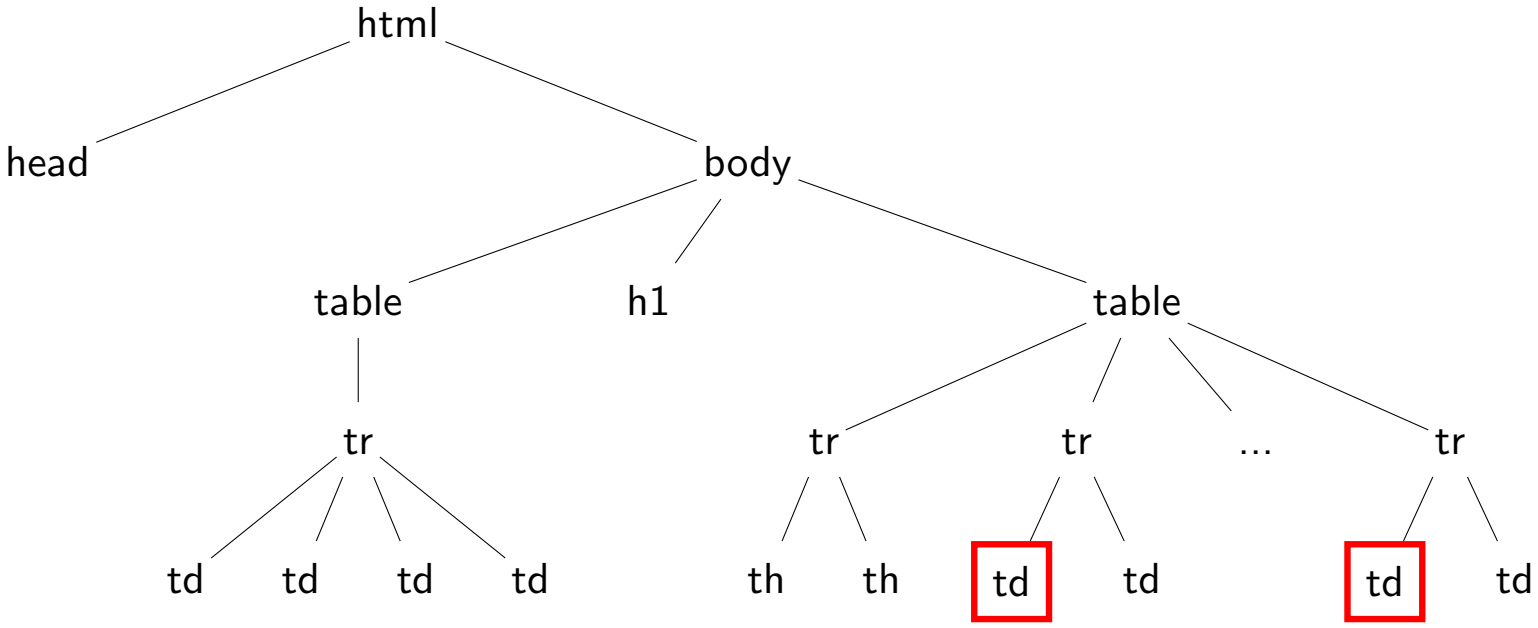
# Logical forms: XPath expressions




$z$ = /html[1]/body[1]/table[2]/tr/td[1]

[Sahuguet and Azavant, 1999; Liu et al., 2000; Crescenzi et al., 2001]

# Logical forms: XPath expressions

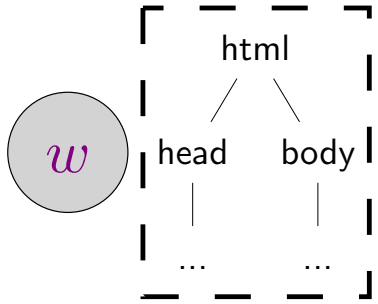



$z$ = /html[1]/body[1]/table[2]/tr/td[1]

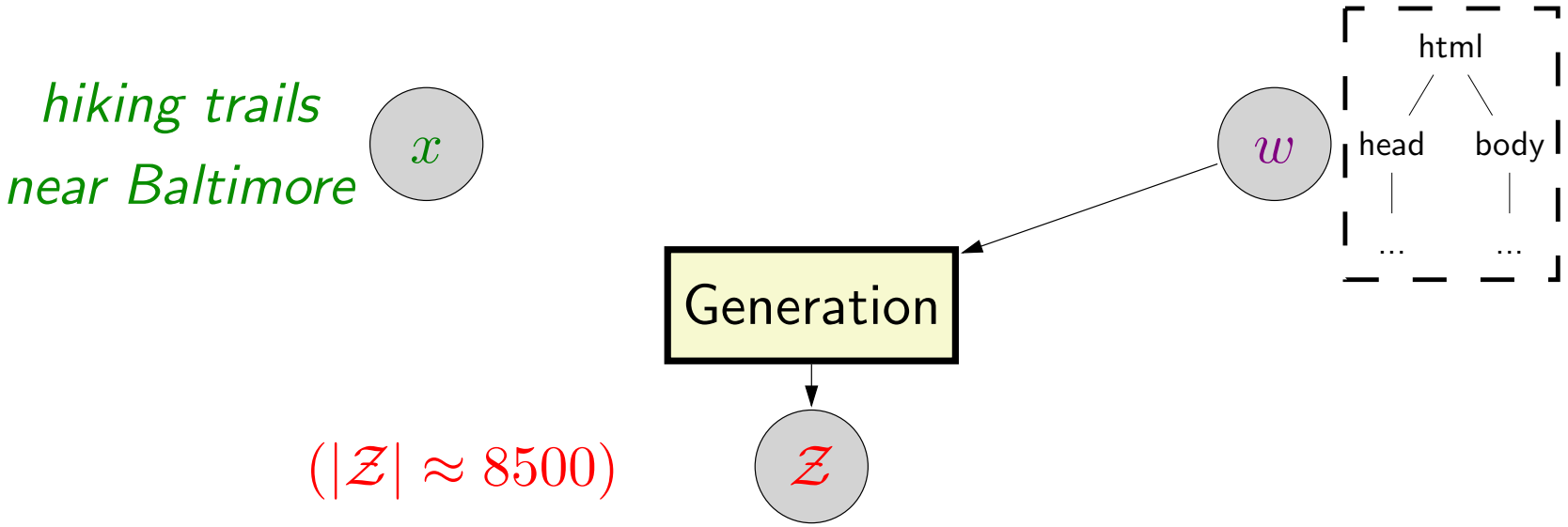
**A low-level KB**

# Framework

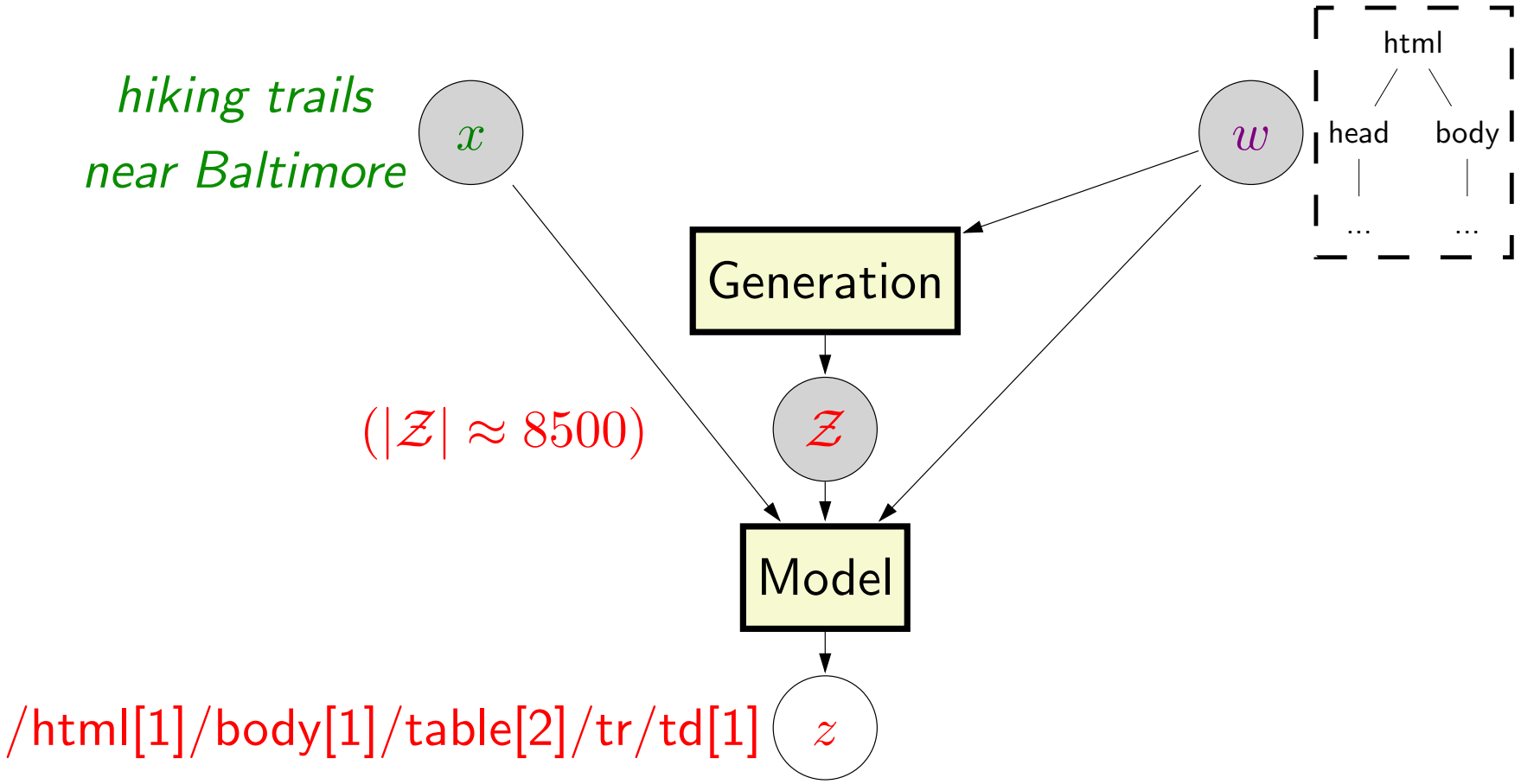
*hiking trails near Baltimore* $x$



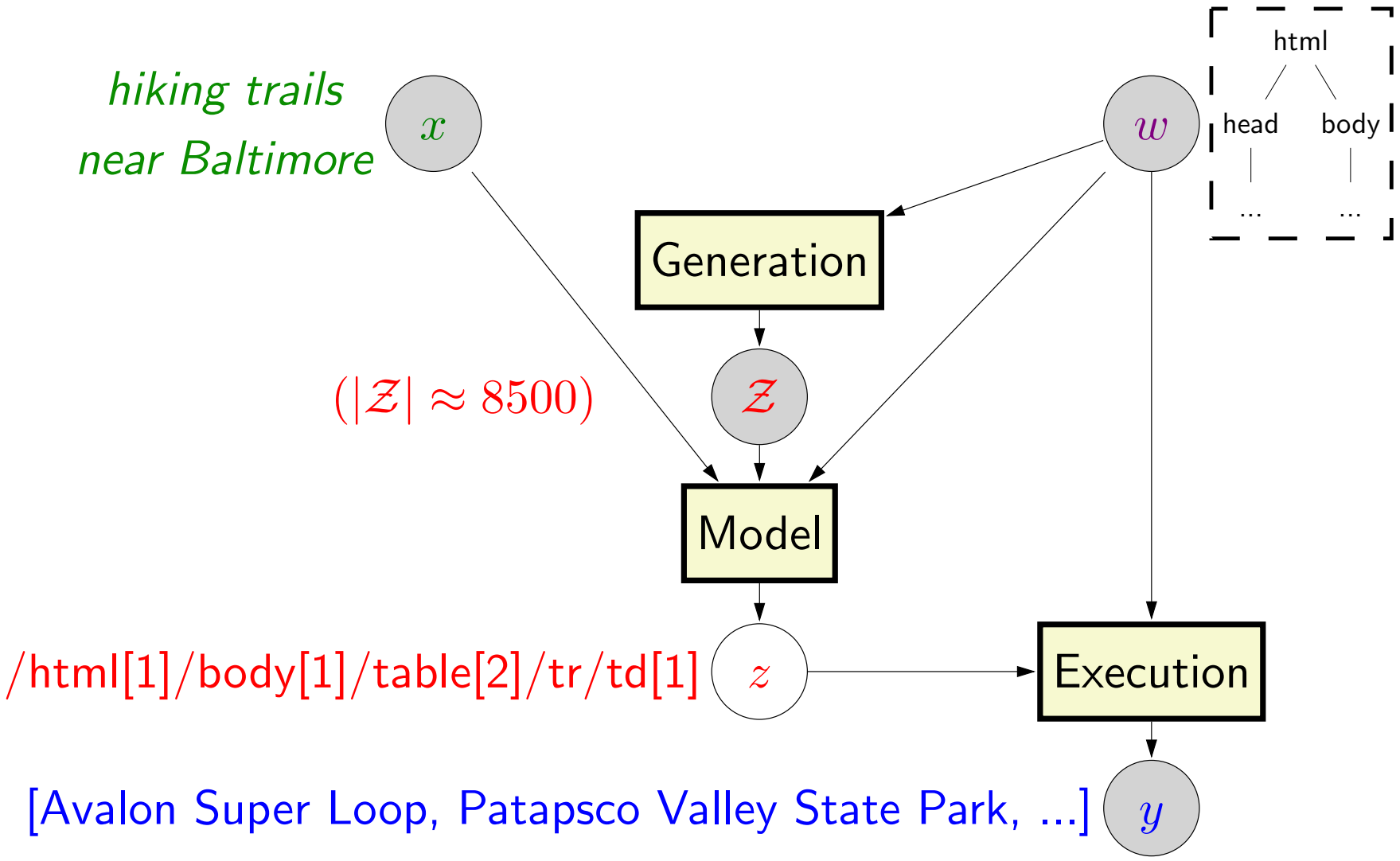
# Framework


hiking trails
near Baltimore
$x$
$w$
html
head
body
...
...
Generation
$(|\mathcal{Z}| \approx 8500)$
$\mathcal{Z}$

# Framework


hiking trails
near Baltimore
$x$
$w$
html
head
body
...
...
Generation
$(|\mathcal{Z}| \approx 8500)$
$\mathcal{Z}$
Model
/html[1]/body[1]/table[2]/tr/td[1]
$z$

# Framework


hiking trails
near Baltimore
$x$
$w$
html
head
body
...
...
Generation
$(|\mathcal{Z}| \approx 8500)$
$\mathcal{Z}$
Model
/html[1]/body[1]/table[2]/tr/td[1]
$z$
Execution
[Avalon Super Loop, Patapsco Valley State Park, ...]
$y$

# Dataset

*airlines of italy*

*natural causes of global warming*

*lsu football coaches*

*bf3 submachine guns*

*badminton tournaments*

*foods high in dha*

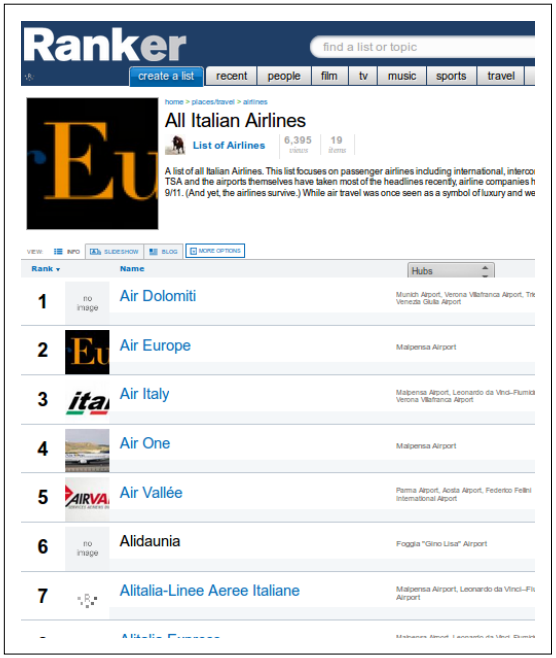
*technical colleges in south carolina*

*songs on glee season 5*

*singers who use auto tune*

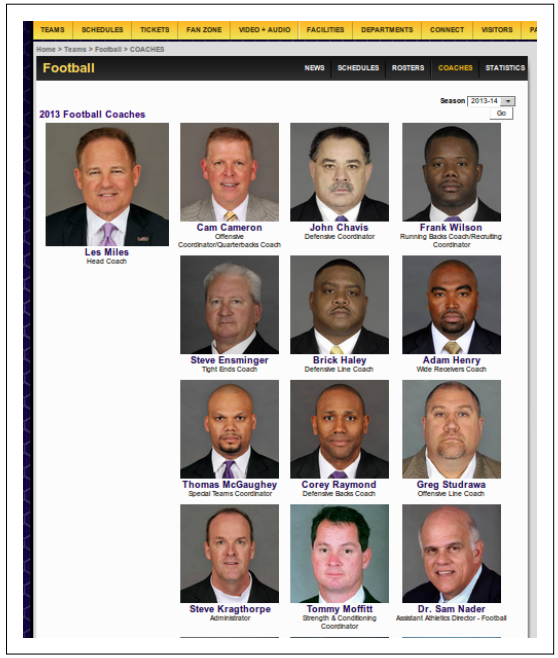
*san francisco radio stations*

# Dataset




*airlines of italy*


*natural causes of global warming*




*lsu football coaches*

# Dataset statistics

2773 examples

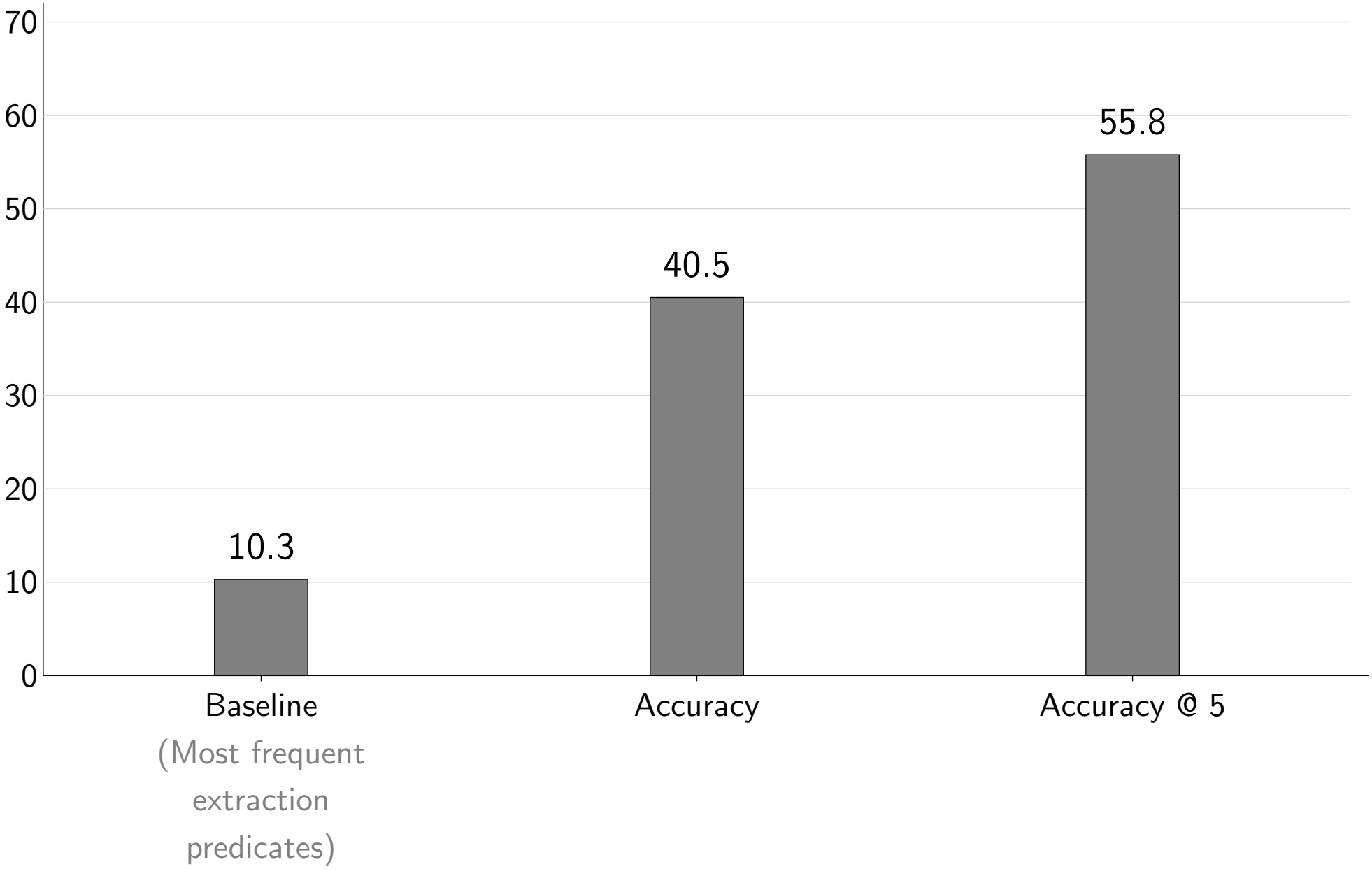
2269 unique queries

894 unique headwords $\leftarrow$ long tail!

1483 unique web domains $\leftarrow$ long tail!

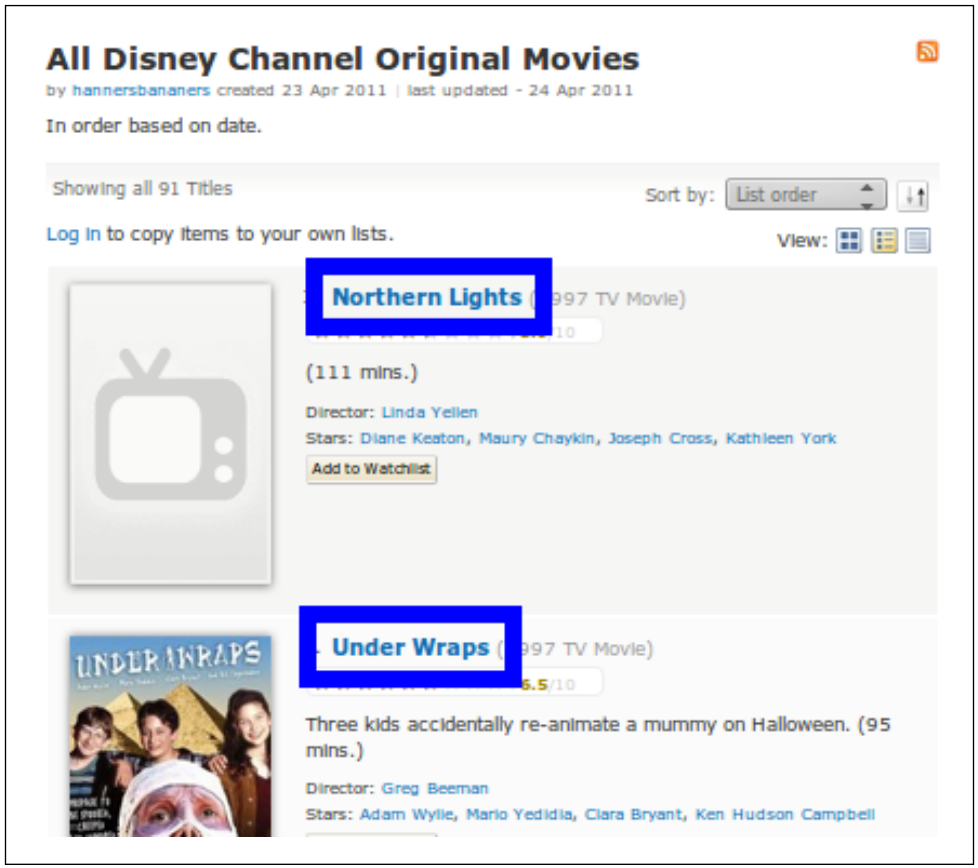
($\neq$ wrapper induction)

# Results


70
60
50
40
30
20
10
0
10.3
40.5
55.8
Baseline
(Most frequent extraction predicates)
Accuracy
Accuracy @ 5

# Correct prediction

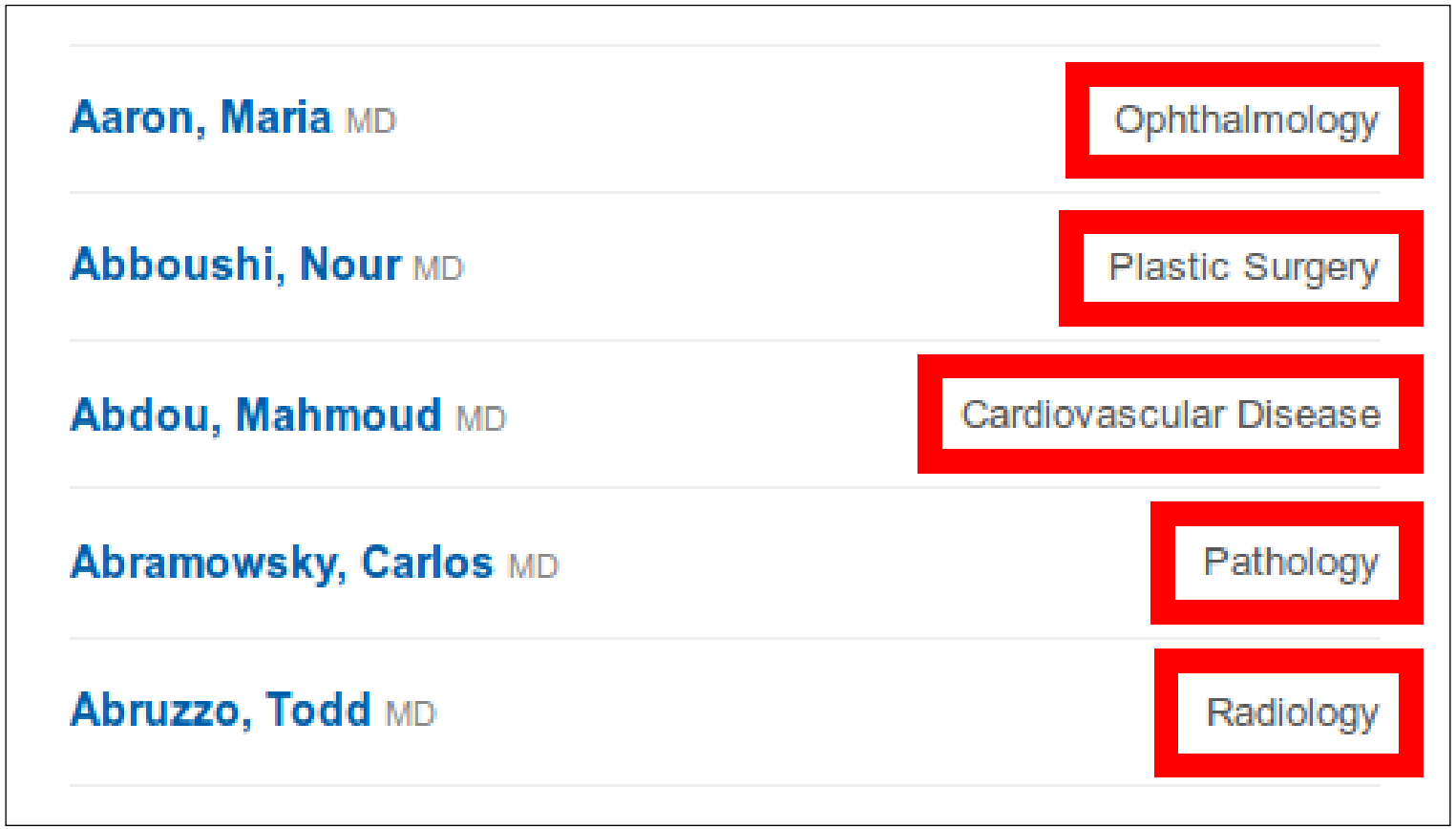
Query: *disney channel movies*




/html[1]/body/div[2]/div/div/div[3]/div[1]/div/div/div/div/b

# Ranking error

Query: *doctors at emory*




/html/body/div[3]/div[4]/table/tbody/tr/td[2]

**Need better understanding of entities/categories**

# Coverage error

Query: *hedge funds in new york*

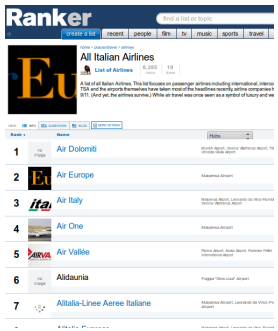
| Rank | Firm | Headquarters |
|---|---|---|
| 1 | Bridgewater Associates | Westport, CT |
| 2 | Man Group | London |
| 3 | J.P. Morgan Asset Management | New York |
| 4 | Brevan Howard Asset Management | London |
| 5 | Och-Ziff Capital Management Group | New York |
| 6 | Paulson & Co. | New York |
| 7 | BlackRock Advisors | New York |

/html/body/div[3]/div[3]/div[4]/.../table/tbody/tr/td[2]/a

**Need compositionality**

# Outline

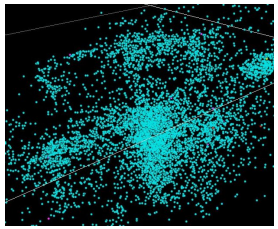
On web pages




**On tables**

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | European Junior Championships | Grosseto, Italy | 1st | Medley relay | 1:50.46 |
| | | | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

In vector space



# Semantic parsing on tables

Panupong (Ice) Pasupat

ACL 2015, ACL 2016

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | | | 1st | Medley relay | 1:50.46 |
| | European Junior Championships | Grosseto, Italy | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

*In what city did Piotr's last 1st place finish occur?*

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | | | 1st | Medley relay | 1:50.46 |
| | European Junior Championships | Grosseto, Italy | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

*How long did it take this competitor to finish the 4x400 meter relay at Universiade in 2005?*

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | | | 1st | Medley relay | 1:50.46 |
| | European Junior Championships | Grosseto, Italy | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

*Where was the competition held immediately before the one in Turkey?*

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | | | 1st | Medley relay | 1:50.46 |
| | European Junior Championships | Grosseto, Italy | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

*How many times has this competitor placed 5th or better in competition?*

# Dataset

Statistics:

- 22000 question/answers
- 2100 tables
- 6.3 columns and 27.5 rows per table

# Dataset

Statistics:

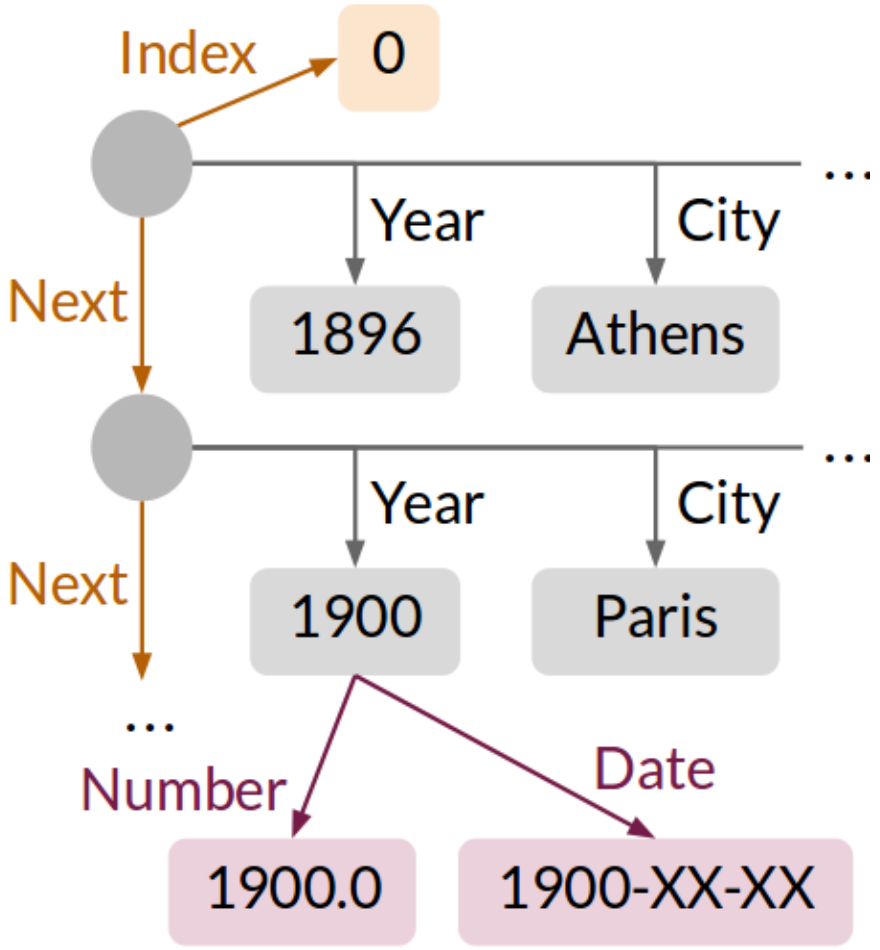
- 22000 question/answers
- 2100 tables
- 6.3 columns and 27.5 rows per table

Challenges:

- High logical complexity (conjunction, disjunction, superlatives, comparatives, aggregation, arithmetic)
- Tables are unnormalized
- Train and test tables are distinct; need to generalize!

# Knowledge representation

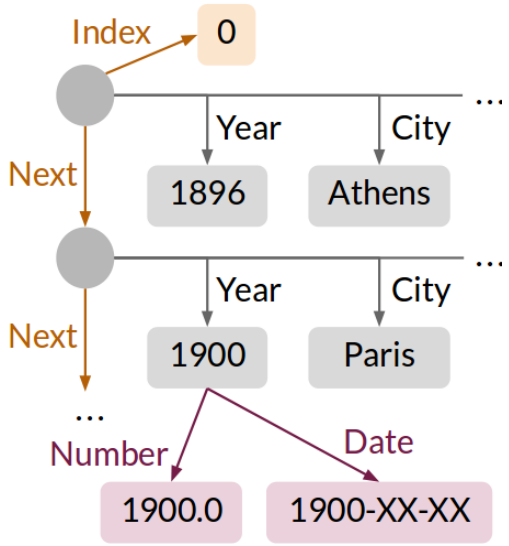
| Year | City | Country | Nations |
|---|---|---|---|
| 1896 | Athens | Greece | 14 |
| 1900 | Paris | France | 24 |
| 1904 | St. Louis | USA | 12 |
| ... | ... | ... | ... |
| 2004 | Athens | Greece | 201 |
| 2008 | Beijing | China | 204 |
| 2012 | London | UK | 204 |




Add normalization / auxiliary edges (custom functions), push resolution to semantic parsing

# Model

| Year | City | Country | Nations |
|---|---|---|---|
| 1896 | Athens | Greece | 14 |
| 1900 | Paris | France | 24 |
| 1904 | St. Louis | USA | 12 |
| ... | ... | ... | ... |
| 2004 | Athens | Greece | 201 |
| 2008 | Beijing | China | 204 |
| 2012 | London | UK | 204 |

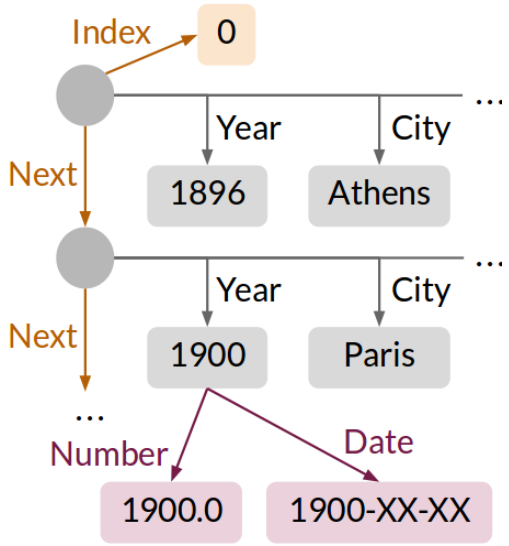



*Greece held its last Summer Olympics in which year?*

2004

# Model

| Year | City | Country | Nations |
|---|---|---|---|
| 1896 | Athens | Greece | 14 |
| 1900 | Paris | France | 24 |
| 1904 | St. Louis | USA | 12 |
| ... | ... | ... | ... |
| 2004 | Athens | Greece | 201 |
| 2008 | Beijing | China | 204 |
| 2012 | London | UK | 204 |




*Greece held its last Summer Olympics in which year?*

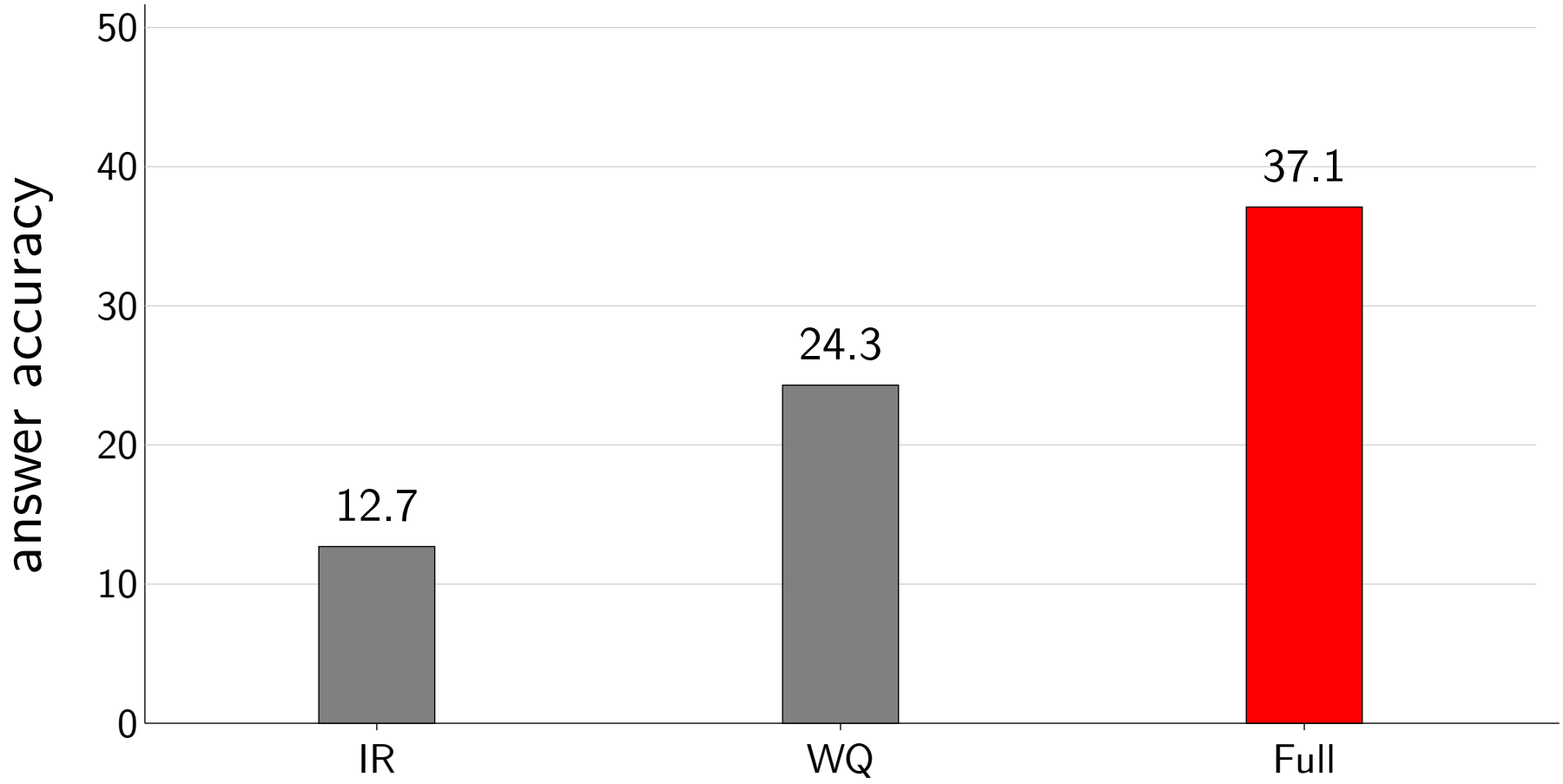
$$\mathbf{R}[\text{Date}].\mathbf{R}[\text{Year}].\text{argmax}(\text{Country.Greece}, \text{Index})$$

2004

# Results

IR: Train classifer to pick answer directly from table.

WQ: Use logical complexity of our previous Freebase work.

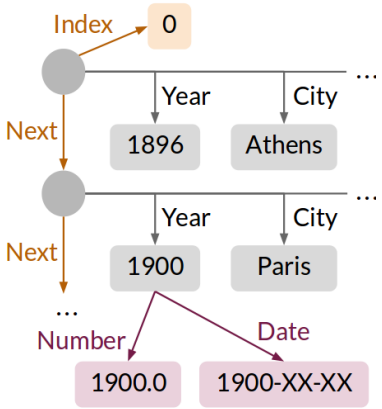


# Oracle accuracy

Can the system even generate a set of candidates containing the answer?

*How many times did Greece hold the summer olympics?*

| Year | City | Country | Nations |
|---|---|---|---|
| 1896 | Athens | Greece | 14 |
| 1900 | Paris | France | 24 |
| 1904 | St. Louis | USA | 12 |
| ... | ... | ... | ... |
| 2004 | Athens | Greece | 201 |
| 2008 | Beijing | China | 204 |
| 2012 | London | UK | 204 |



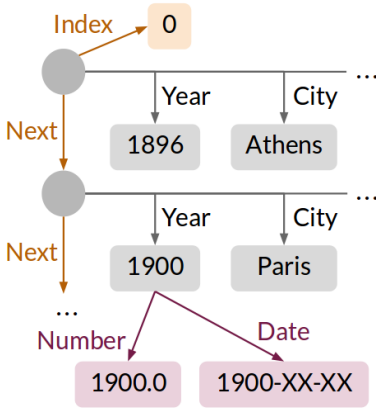

2

# Oracle accuracy

Can the system even generate a set of candidates containing the answer?

*How many times did Greece hold the summer olympics?*

| Year | City | Country | Nations |
|---|---|---|---|
| 1896 | Athens | Greece | 14 |
| 1900 | Paris | France | 24 |
| 1904 | St. Louis | USA | 12 |
| ... | ... | ... | ... |
| 2004 | Athens | Greece | 201 |
| 2008 | Beijing | China | 204 |
| 2012 | London | UK | 204 |




2

| **Method** | **LF accuracy** |
|---|---|
| ACL 2015 | 53.5% |
| ACL 2016 (dynamic prog. on denotations) | 76.0% |

# Error analysis

Unhandled operations (19%):

- *Was there more gold medals won than silver?*
- *Which movies were number 1 for at least two* **consecutive** *weeks?*
- *How many titles had the* **same** *author listed as the illustrator?*

# Error analysis

Unhandled operations (19%):

- *Was there more gold medals won than silver?*
- *Which movies were number 1 for at least two* **consecutive** *weeks?*
- *How many titles had the* **same** *author listed as the illustrator?*

Table normalization:

- *In what city did Piotr's last 1st place finish occur?* ...[Bangkok, Thailand]...
- *How long does the show defcon 3 last?* ...[2pm-3pm]...

# Error analysis

Unhandled operations (19%):

- *Was there more gold medals won than silver?*
- *Which movies were number 1 for at least two* **consecutive** *weeks?*
- *How many titles had the* **same** *author listed as the illustrator?*

Table normalization:

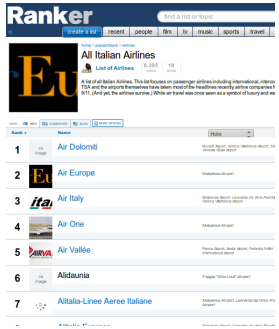
- *In what city did Piotr's last 1st place finish occur?* ...[Bangkok, Thailand]...
- *How long does the show defcon 3 last?* ...[2pm-3pm]...

Lexical mismatch:

- *Mexican* $\Rightarrow$ Mexico, *airplane* $\Rightarrow$ Model

# Outline

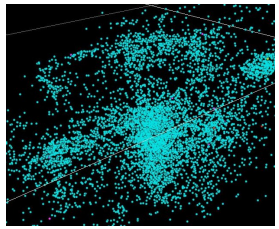
On web pages




On tables

| Year | Competition | Venue | Position | Event | Notes |
|---|---|---|---|---|---|
| Representing Poland | | | | | |
| 2001 | World Youth Championships | Debrecen, Hungary | 2nd | 400 m | 47.12 |
| | | | 1st | Medley relay | 1:50.46 |
| | European Junior Championships | Grosseto, Italy | 1st | 4x400 m relay | 3:06.12 |
| 2003 | European Junior Championships | Tampere, Finland | 3rd | 400 m | 46.69 |
| | | | 2nd | 4x400 m relay | 3:08.62 |
| 2005 | European U23 Championships | Erfurt, Germany | 11th (sf) | 400 m | 46.62 |
| | | | 1st | 4x400 m relay | 3:04.41 |
| | Universiade | Izmir, Turkey | 7th | 400 m | 46.89 |
| | | | 1st | 4x400 m relay | 3:02.57 |
| 2006 | World Indoor Championships | Moscow, Russia | 2nd (h) | 4x400 m relay | 3:06.10 |
| | European Championships | Gothenburg, Sweden | 3rd | 4x400 m relay | 3:01.73 |
| 2007 | European Indoor Championships | Birmingham, United Kingdom | 3rd | 4x400 m relay | 3:08.14 |
| | Universiade | Bangkok, Thailand | 7th | 400 m | 46.85 |
| | | | 1st | 4x400 m relay | 3:02.05 |
| 2008 | World Indoor Championships | Valencia, Spain | 4th | 4x400 m relay | 3:08.76 |
| | Olympic Games | Beijing, China | 7th | 4x400 m relay | 3:00.32 |
| 2009 | Universiade | Belgrade, Serbia | 2nd | 4x400 m relay | 3:05.69 |

**In vector space**



# Compositional Queries in Vector Space

Kelvin Guu

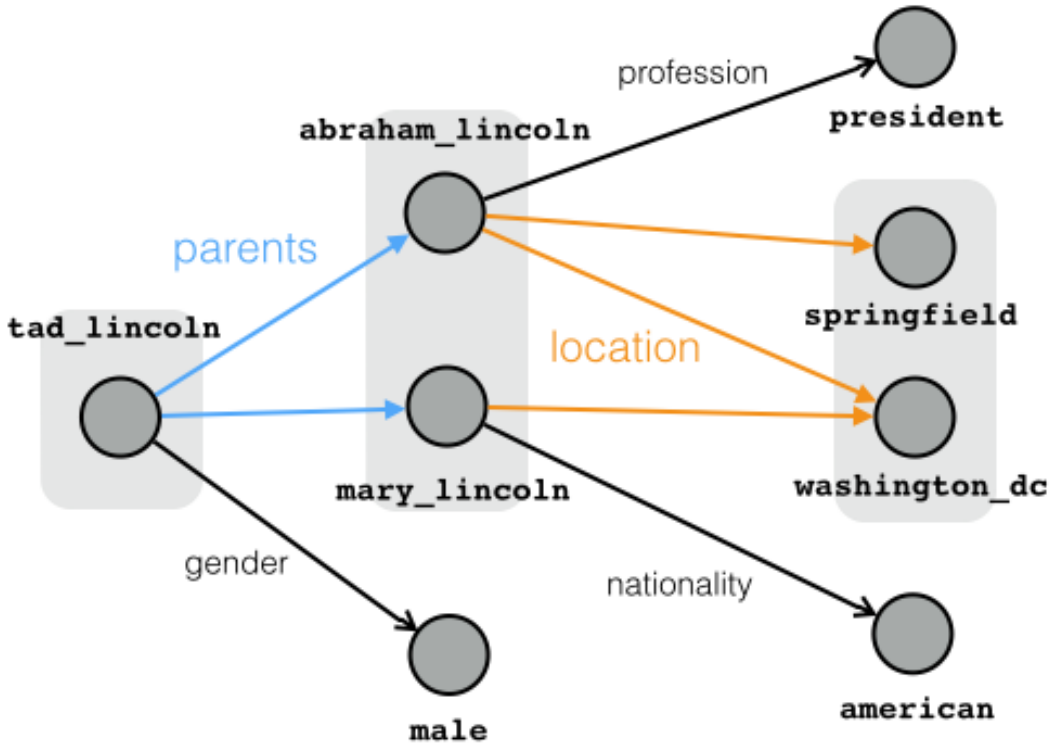
John Miller

# Focus on path queries

TadLincoln/Parents/Location

# Focus on path queries

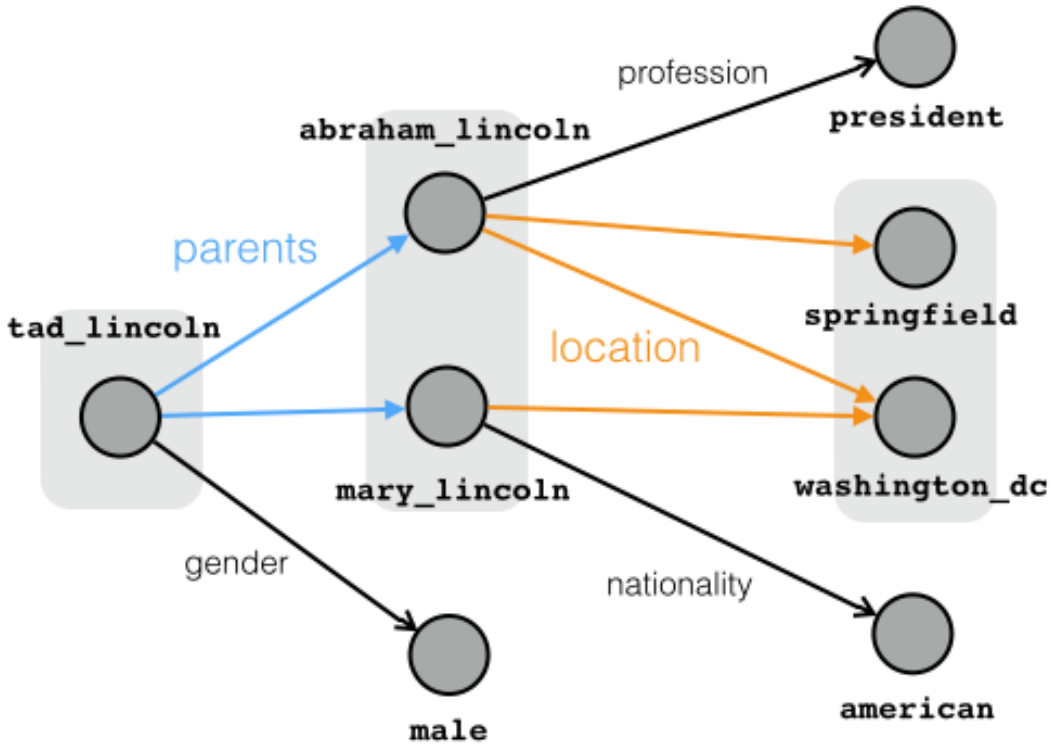
## TadLincoln/Parents/Location




Strength: compositionality

# Focus on path queries

TadLincoln/Parents/Location




Strength: compositionality

Weaknesses: can't handle fact incompleteness, hypotheticals

abraham_lincoln/daughter/ethnicity

# Vector space relation modeling

$$\text{Score}(\text{entity1}, \text{relation}, \text{entity2})$$

# Vector space relation modeling

$$\text{Score}(\text{entity1}, \text{relation}, \text{entity2})$$

Prior work:

- Tensor factorization [Nickel et al., 2011]
- Neural Tensor Network [Socher et al., 2013]
- TransE [Bordes et al., 2013]
- Universal schema [Riedel et al., 2013]
- General framework + comparison [Yang et al., 2015]
- Compositional embedding of paths [Neelakantan et al., 2015]

# Vector space relation modeling

$$\text{Score}(\text{entity1}, \text{relation}, \text{entity2})$$

Prior work:

- Tensor factorization [Nickel et al., 2011]
- Neural Tensor Network [Socher et al., 2013]
- TransE [Bordes et al., 2013]
- Universal schema [Riedel et al., 2013]
- General framework + comparison [Yang et al., 2015]
- Compositional embedding of paths [Neelakantan et al., 2015]

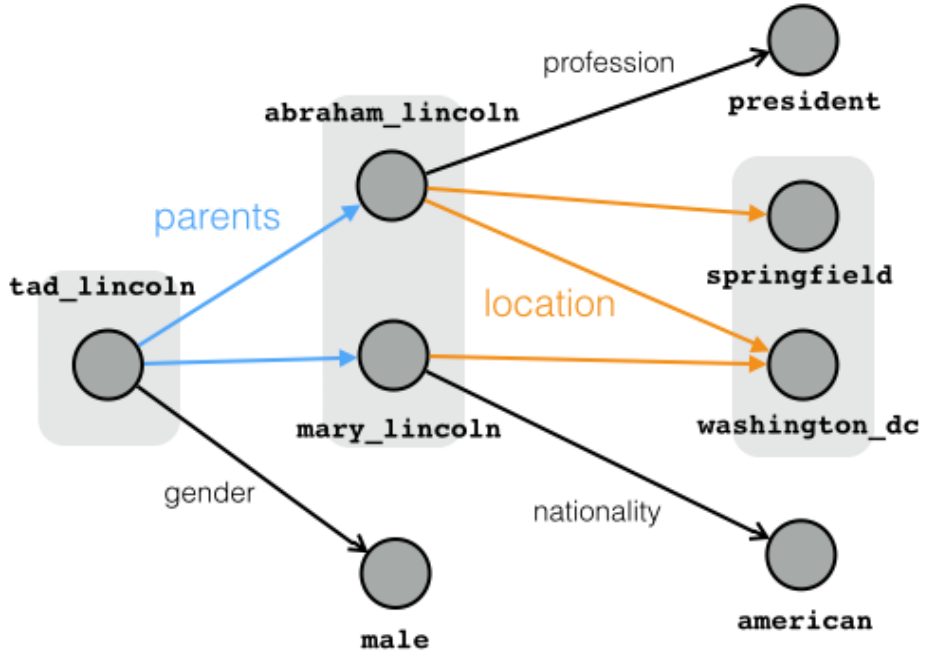
Strength: handles incompleteness

Weakness: no compositionality

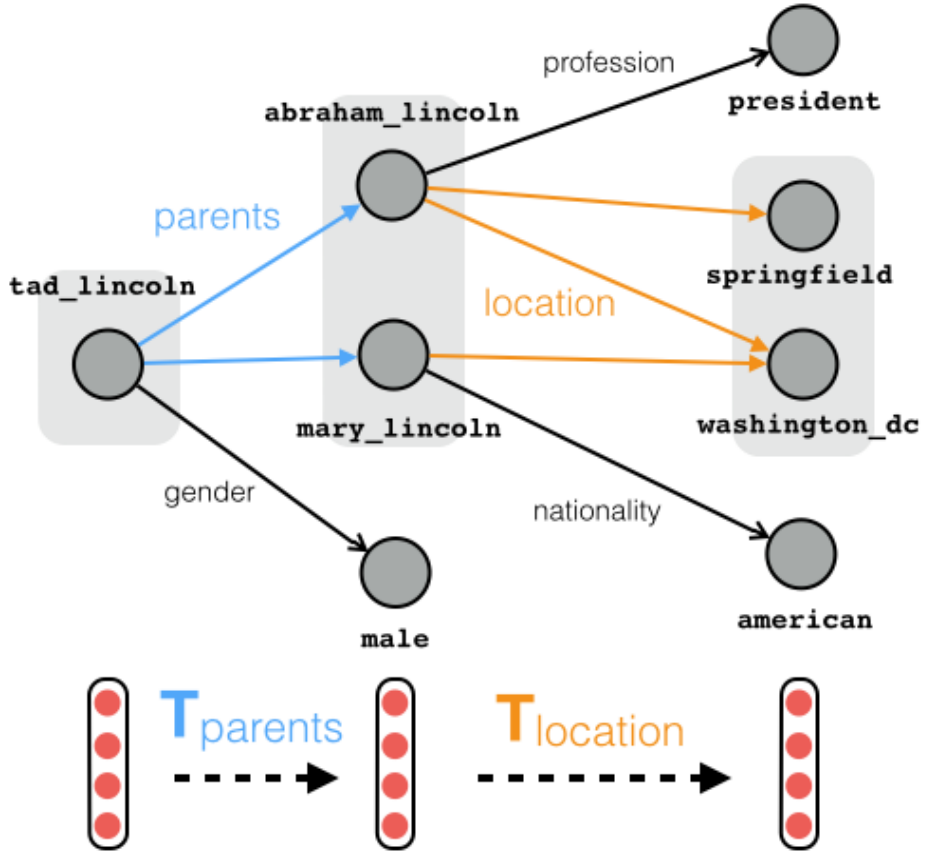
# Goal

Compositionality + Handle Incompleteness

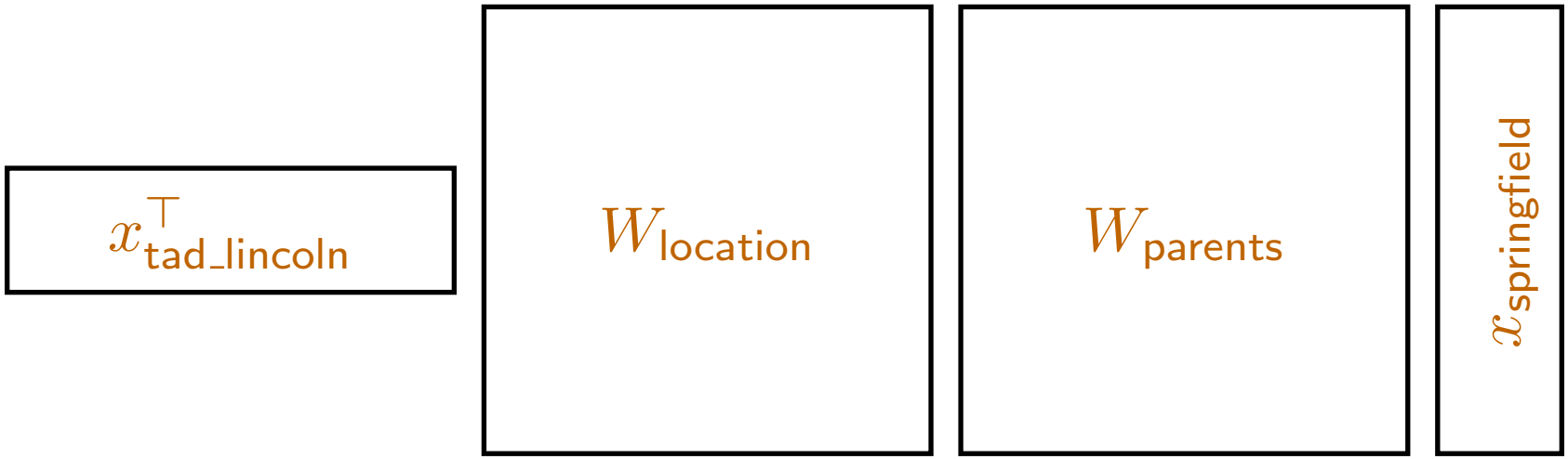
# Path queries in vector space



# Path queries in vector space



# Path queries in vector space: example



# Experimental setup: models


Bilinear [Nickel et al., 2012]:

$$\mathbb{T}_r(v) = v^\top W_r$$

# Experimental setup: models


Bilinear [Nickel et al., 2012]:

$$\mathbb{T}_r(v) = v^\top W_r$$

Bilinear-Diag [Yang et al., 2015]:

$$\mathbb{T}_r(v) = v^\top \mathsf{diag}(w_r)$$

TransE [Bordes et al., 2013]:

$$\mathbb{T}_r(v) = v + w_r$$

# Experimental setup: datasets

| | | WordNet | Freebase |
|---|---|---|---|
| **Relations** | | 11 | 13 |
| **Entities** | | 38,696 | 75,043 |
| **Base** | Train | 112,581 | 316,232 |
| | Test | 10,544 | 23,733 |
| **Paths** | Train | 2,129,539 | 6,266,058 |
| | Test | 46,577 | 109,557 |

(paths length 1-5 generated randomly)

# Experimental setup: training

SINGLE: standard training

| | | |
|---|---|---|
| abraham_lincoln/location | $\Rightarrow$ | springfield |
| stephen_curry/birthdate | $\Rightarrow$ | 1988 |
| ... | | ... |

# Experimental setup: training

SINGLE: standard training

| | | |
|---|---|---|
| abraham_lincoln/location | $\Rightarrow$ | springfield |
| stephen_curry/birthdate | $\Rightarrow$ | 1988 |
| ... | | ... |

COMP: compositional training

| | | |
|---|---|---|
| abraham_lincoln/location | $\Rightarrow$ | springfield |
| stephen_curry/birthdate | $\Rightarrow$ | 1988 |
| ... | | ... |

# Experimental setup: training

SINGLE: standard training

| | | |
|---|---|---|
| abraham_lincoln/location | $\Rightarrow$ | springfield |
| stephen_curry/birthdate | $\Rightarrow$ | 1988 |
| ... | | ... |

COMP: compositional training

| | | |
|---|---|---|
| abraham_lincoln/location | $\Rightarrow$ | springfield |
| stephen_curry/birthdate | $\Rightarrow$ | 1988 |
| ... | | ... |
| tad_lincoln/parents/location | $\Rightarrow$ | springfield |
| stephen_curry/wife/birthdate | $\Rightarrow$ | 1989 |
| ... | | ... |

# Experiments: knowledge base completion

abraham_lincoln/location ⇒ springfield

| | | Bilinear | | | Bilinear-Diag | | | TransE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Path query task** | | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) |
| WordNet | MQ | 84.7 | 89.4 | **30.7** | 59.7 | 90.4 | **76.2** | 83.7 | 93.3 | **58.9** |
| | H@10 | 43.6 | 54.3 | **19.0** | 7.9 | 31.1 | **25.4** | 13.8 | 43.5 | **34.5** |
| Freebase | MQ | 58.0 | 83.5 | **60.7** | 57.9 | 84.8 | **63.9** | 86.2 | 88 | **13.0** |
| | H@10 | 25.9 | 42.1 | **21.9** | 23.1 | 38.6 | **20.2** | 45.4 | 50.5 | **9.3** |
| **KBC task** | | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) |
| WordNet | MQ | 76.1 | 82.0 | **24.7** | 76.5 | 84.3 | **33.2** | 75.5 | 86.1 | **43.3** |
| | H@10 | 19.2 | 27.3 | **10.0** | 12.9 | 14.4 | **1.72** | 4.6 | 16.5 | **12.5** |
| Freebase | MQ | 85.3 | 91.0 | **38.8** | 84.6 | 89.1 | **29.2** | 92.7 | 92.8 | **1.37** |
| | H@10 | 70.2 | 76.4 | **20.8** | 63.2 | 67.0 | **10.3** | 78.8 | 78.6 | -0.9 |

# Experiments: knowledge base completion

abraham_lincoln/location ⇒ springfield

| | | Bilinear | | | Bilinear-Diag | | | TransE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Path query task** | | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) |
| WordNet | MQ | 84.7 | 89.4 | **30.7** | 59.7 | 90.4 | **76.2** | 83.7 | 93.3 | **58.9** |
| | H@10 | 43.6 | 54.3 | **19.0** | 7.9 | 31.1 | **25.4** | 13.8 | 43.5 | **34.5** |
| Freebase | MQ | 58.0 | 83.5 | **60.7** | 57.9 | 84.8 | **63.9** | 86.2 | 88 | **13.0** |
| | H@10 | 25.9 | 42.1 | **21.9** | 23.1 | 38.6 | **20.2** | 45.4 | 50.5 | **9.3** |
| **KBC task** | | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) |
| WordNet | MQ | 76.1 | 82.0 | **24.7** | 76.5 | 84.3 | **33.2** | 75.5 | 86.1 | **43.3** |
| | H@10 | 19.2 | 27.3 | **10.0** | 12.9 | 14.4 | **1.72** | 4.6 | 16.5 | **12.5** |
| Freebase | MQ | 85.3 | 91.0 | **38.8** | 84.6 | 89.1 | **29.2** | 92.7 | 92.8 | **1.37** |
| | H@10 | 70.2 | 76.4 | **20.8** | 63.2 | 67.0 | **10.3** | 78.8 | 78.6 | -0.9 |

**Compositional training improves KBC!**

Surprising it works: train on paths, test on single edges...

# Experiments: knowledge base completion

abraham_lincoln/location ⇒ springfield

| | | Bilinear | | | Bilinear-Diag | | | TransE | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Path query task** | | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) |
| WordNet | MQ | 84.7 | 89.4 | **30.7** | 59.7 | 90.4 | **76.2** | 83.7 | 93.3 | **58.9** |
| | H@10 | 43.6 | 54.3 | **19.0** | 7.9 | 31.1 | **25.4** | 13.8 | 43.5 | **34.5** |
| Freebase | MQ | 58.0 | 83.5 | **60.7** | 57.9 | 84.8 | **63.9** | 86.2 | 88 | **13.0** |
| | H@10 | 25.9 | 42.1 | **21.9** | 23.1 | 38.6 | **20.2** | 45.4 | 50.5 | **9.3** |
| **KBC task** | | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) | SINGLE | COMP | (%red) |
| WordNet | MQ | 76.1 | 82.0 | **24.7** | 76.5 | 84.3 | **33.2** | 75.5 | 86.1 | **43.3** |
| | H@10 | 19.2 | 27.3 | **10.0** | 12.9 | 14.4 | **1.72** | 4.6 | 16.5 | **12.5** |
| Freebase | MQ | 85.3 | 91.0 | **38.8** | 84.6 | 89.1 | **29.2** | 92.7 | 92.8 | **1.37** |
| | H@10 | 70.2 | 76.4 | **20.8** | 63.2 | 67.0 | **10.3** | 78.8 | 78.6 | -0.9 |

**Compositional training improves KBC!**

Surprising it works: train on paths, test on single edges...

Think of it as a form of **path regularization**

*Final remarks*

# KBC + QA

- Knowledge bases provide structure for difficult **aggregation** questions (computing the answer)

# KBC + QA

- Knowledge bases provide structure for difficult **aggregation** questions (computing the answer)

- Focus on end task of question answering; knowledge bases are just **caching**

# KBC + QA

- Knowledge bases provide structure for difficult **aggregation** questions (computing the answer)

- Focus on end task of question answering; knowledge bases are just **caching**

**Question-dependent KB**

# Question-dependent KB

- Multiple levels of compositionality (think $n$-gram backoff)

*dog-friendly hiking trails near Palo Alto*

# Question-dependent KB

- Multiple levels of compositionality (think $n$-gram backoff)

*[dog-friendly] [hiking trails near Palo Alto]*

# Question-dependent KB

- Multiple levels of compositionality (think $n$-gram backoff)

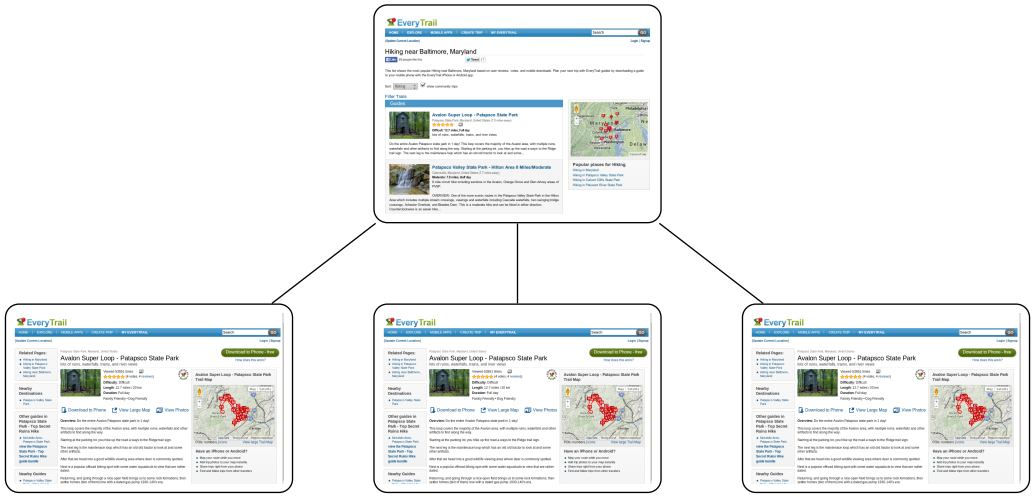
*[dog-friendly] [hiking trails] [near Palo Alto]*

# Question-dependent KB

- Multiple levels of compositionality (think $n$-gram backoff)

*[dog-friendly] [hiking trails] [near Palo Alto]*

/body/div[2]/(CLICK)/body/div/(SEARCH)/body/div[3]



**Query a low-level KB**

What about unstructured text?


# Reading comprehension dataset


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**


# Reading comprehension dataset


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

- 107,785 question/answer pairs
- 536 articles
- Answer is **span** of paragraph


# Reading comprehension dataset


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

- 107,785 question/answer pairs
- 536 articles
- Answer is **span** of paragraph

| **Method** | **F1** |
|---|---|
| Sliding window | 20.0% |
| Log. reg | 51.0% |
| Human | 86.8% |


# Reading comprehension dataset


In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
**gravity**

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
**graupel**

Where do water droplets collide with ice crystals to form precipitation?
**within a cloud**

- 107,785 question/answer pairs
- 536 articles
- Answer is **span** of paragraph

| **Method** | **F1** |
|---|---|
| Sliding window | 20.0% |
| Log. reg | 51.0% |
| Human | 86.8% |

stanford-qa.com

## Code and data


worksheets.codalab.org

## Collaborators

Panupong Pasupat
Kelvin Guu
John Miller

## Funding

Google
Microsoft
DARPA

**Thank you!**