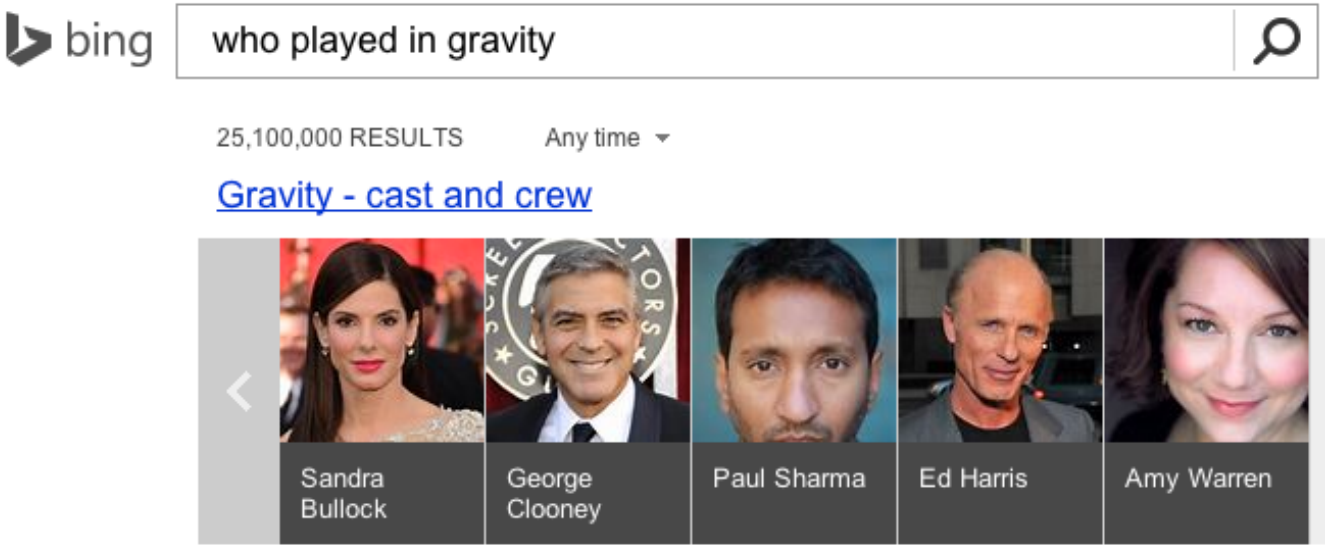
Information Extraction over Structured Data: Question Answering with Freebase

Xuchen Yao and Benjamin Van Durme

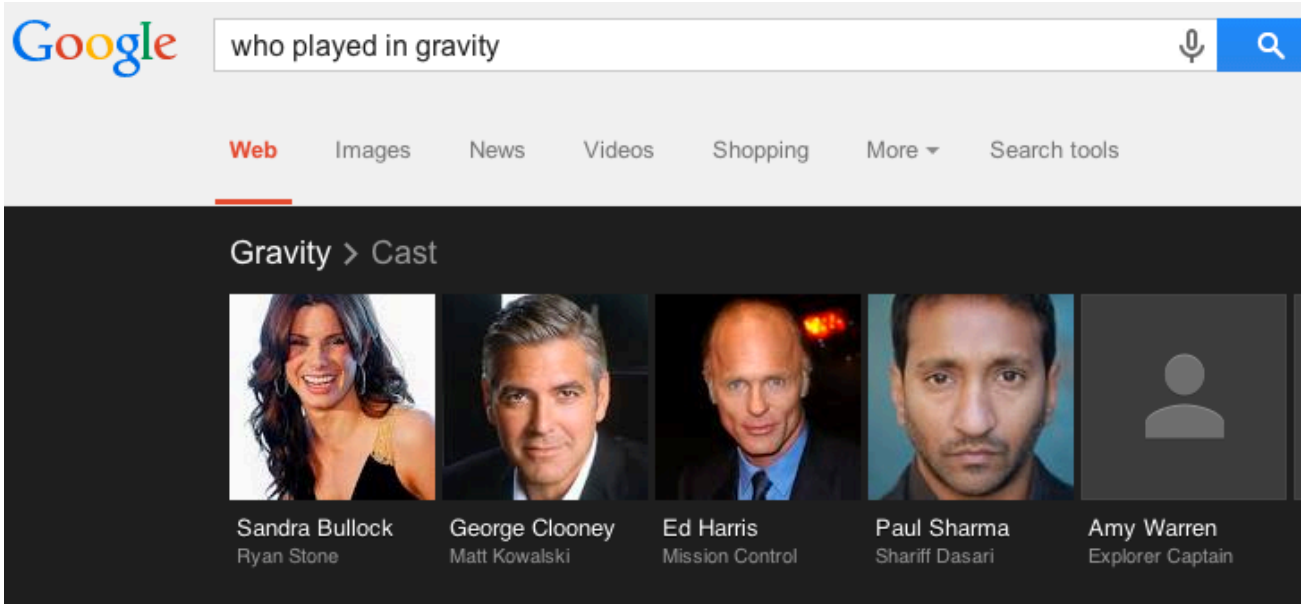
The Center For Language and Speech Processing
at the Johns Hopkins University

human language technology
center of excellence

"Who played in Gravity?"



- Bing: Satori



- Google: knowledge graph, Freebase

Gravity

Theatrical release poster

Directed by	Alfonso Cuarón
Produced by	Alfonso Cuarón David Heyman
Written by	Alfonso Cuarón Jonás Cuarón
Starring	Sandra Bullock George Clooney

Answering from a Knowledge Base

Performances /film/film/starring

Actor	Character
George Clooney	Matt
Sandra Bullock	Ryan

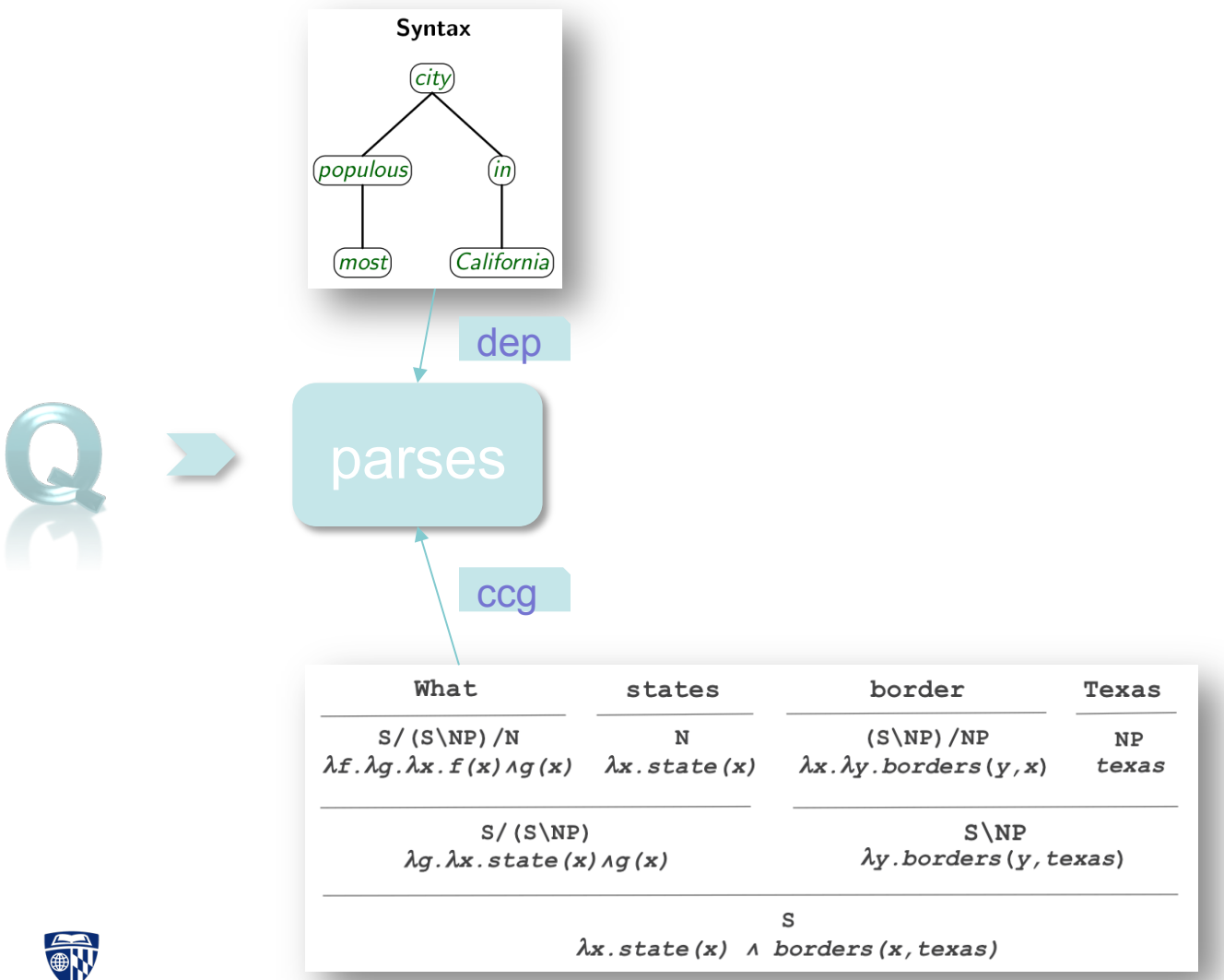
- the model challenge
- the data challenge

QA from KB

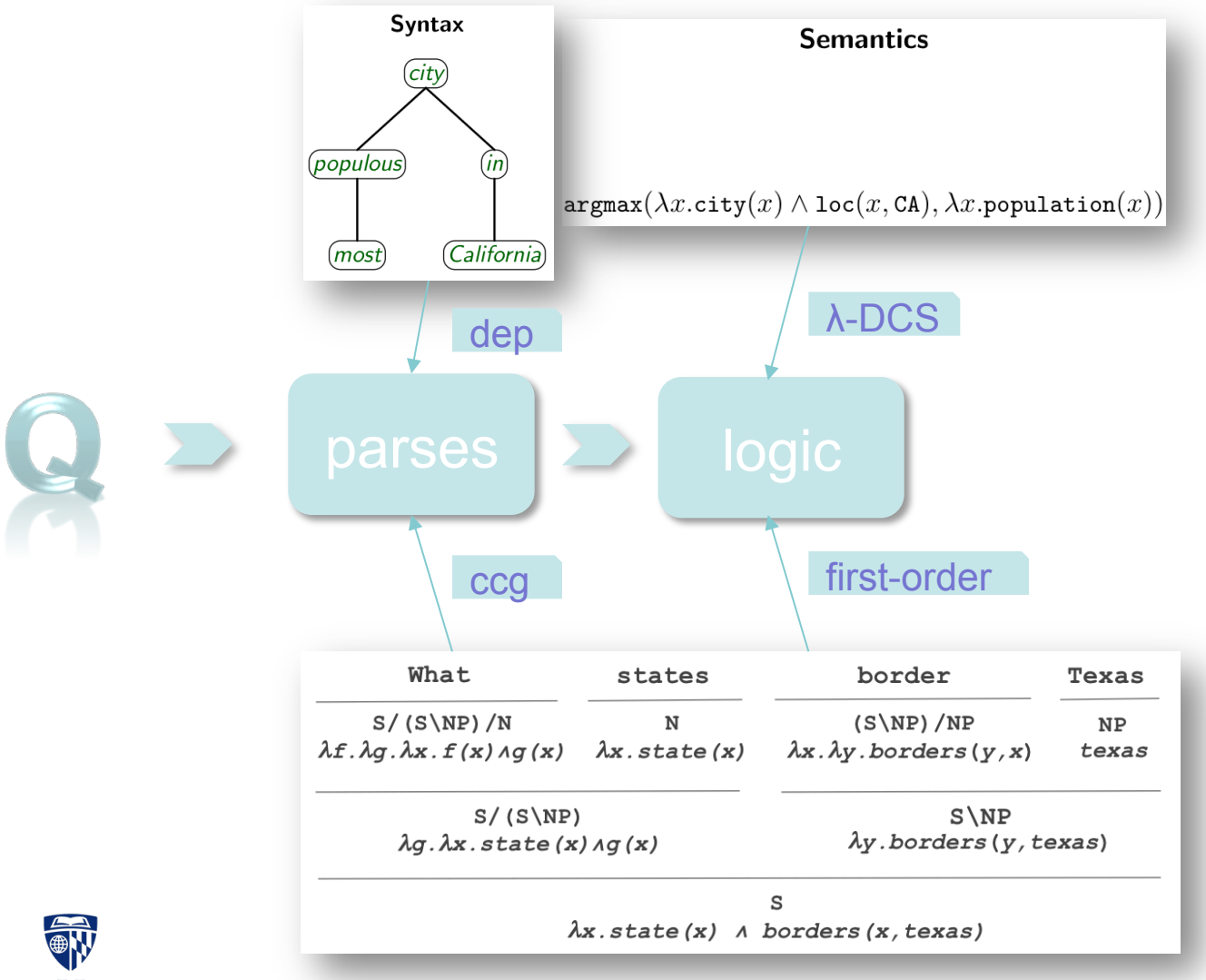
The Model Challenge

Previous Approach: Semantic Parsing

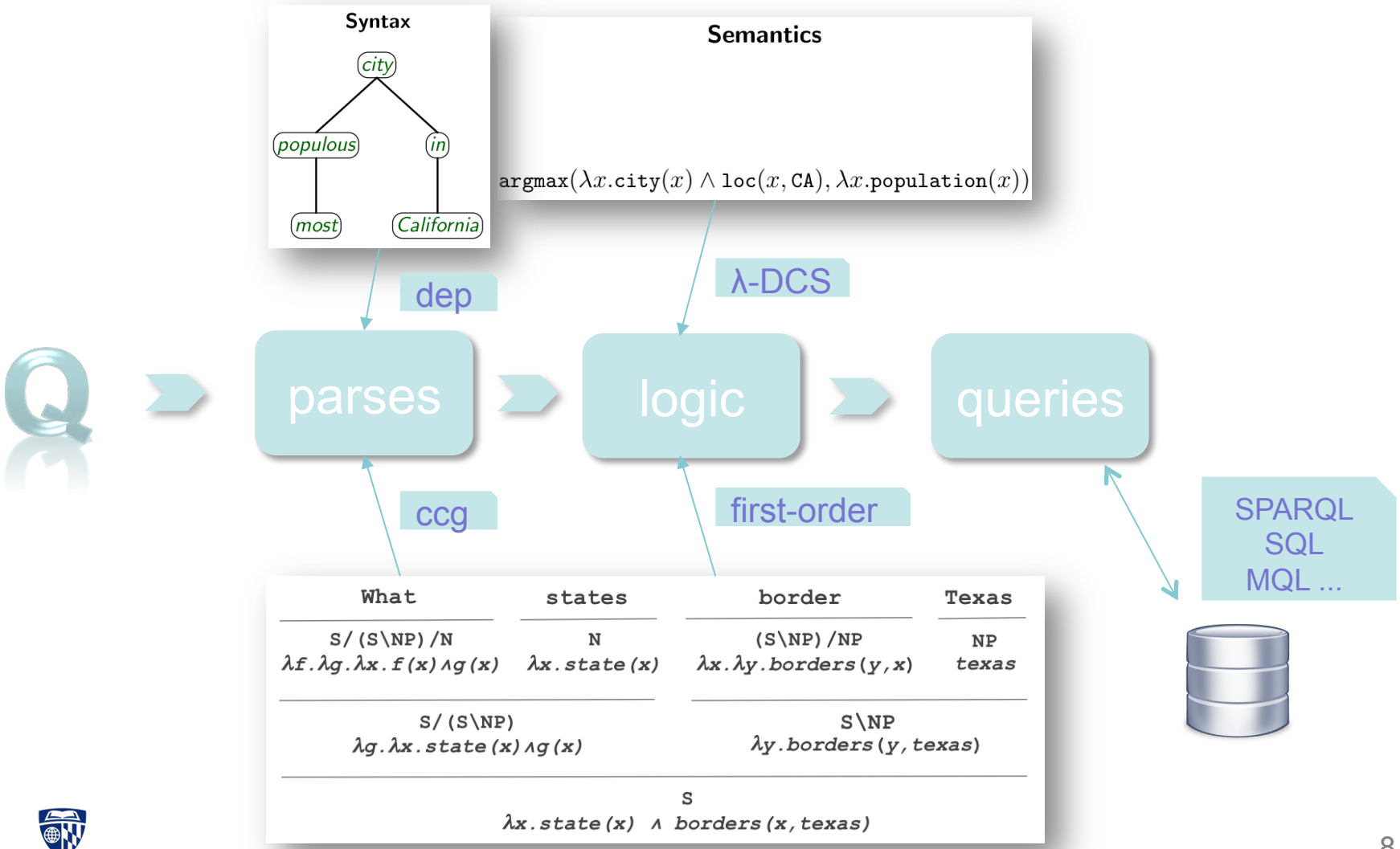
Previous Approach: Semantic Parsing



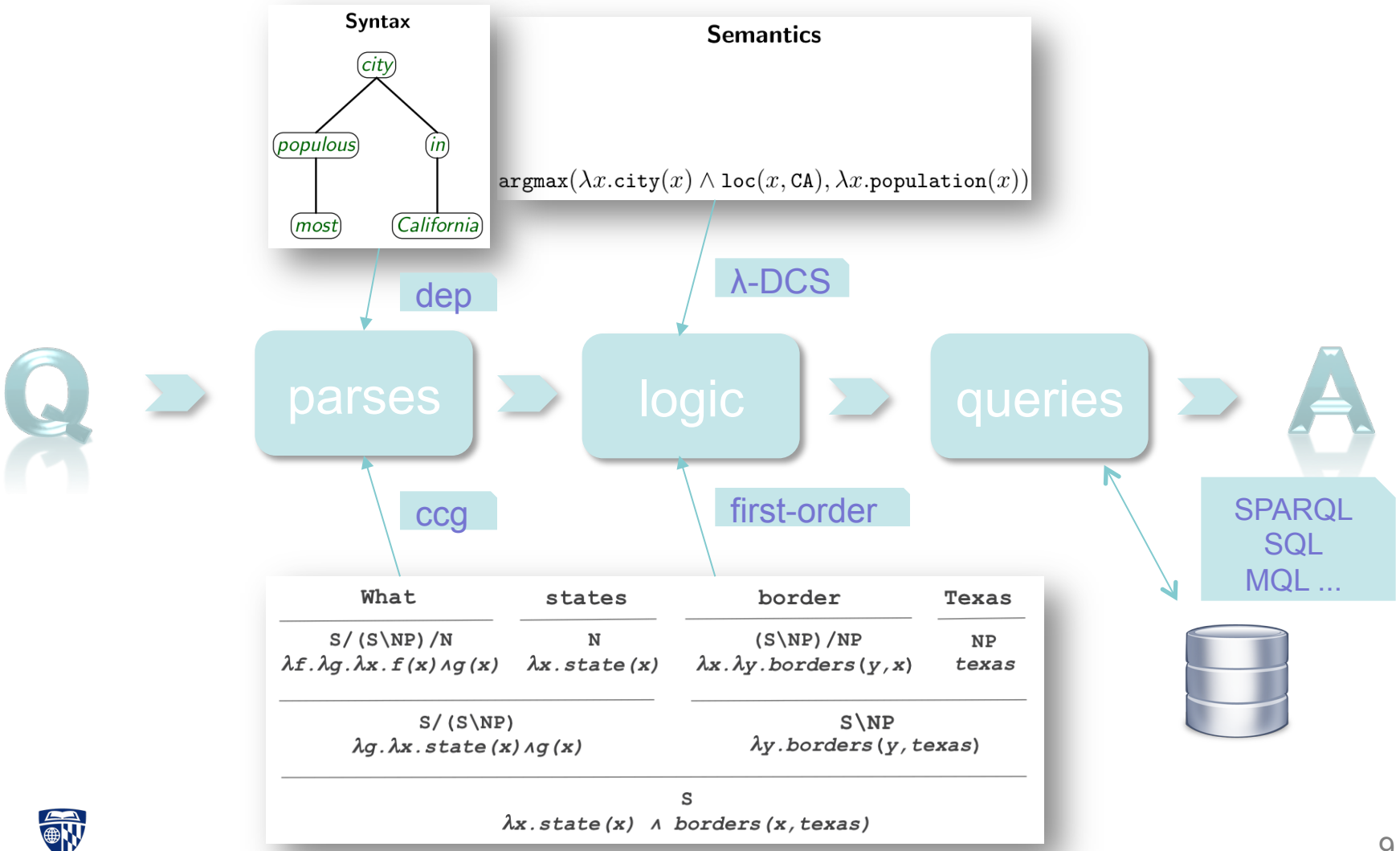
Previous Approach: Semantic Parsing



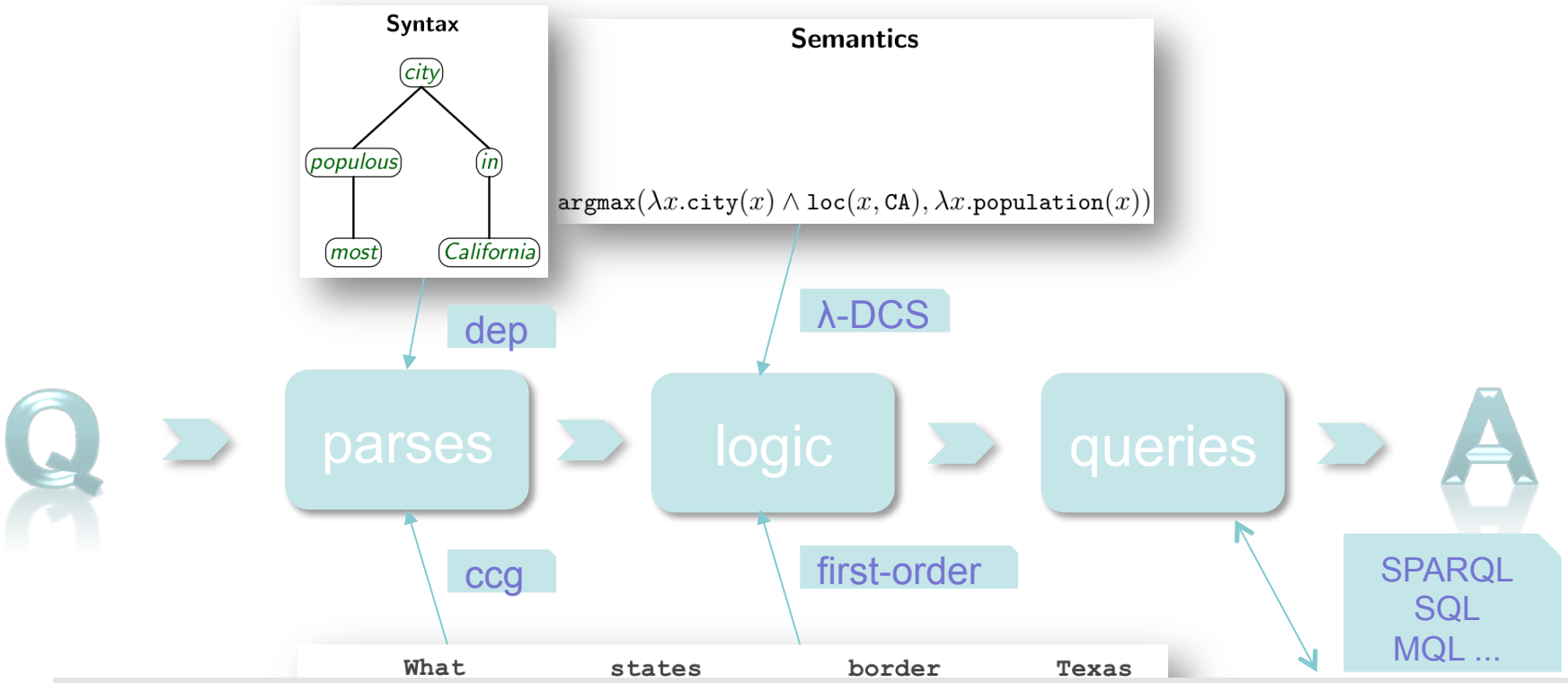
Previous Approach: Semantic Parsing



Previous Approach: Semantic Parsing

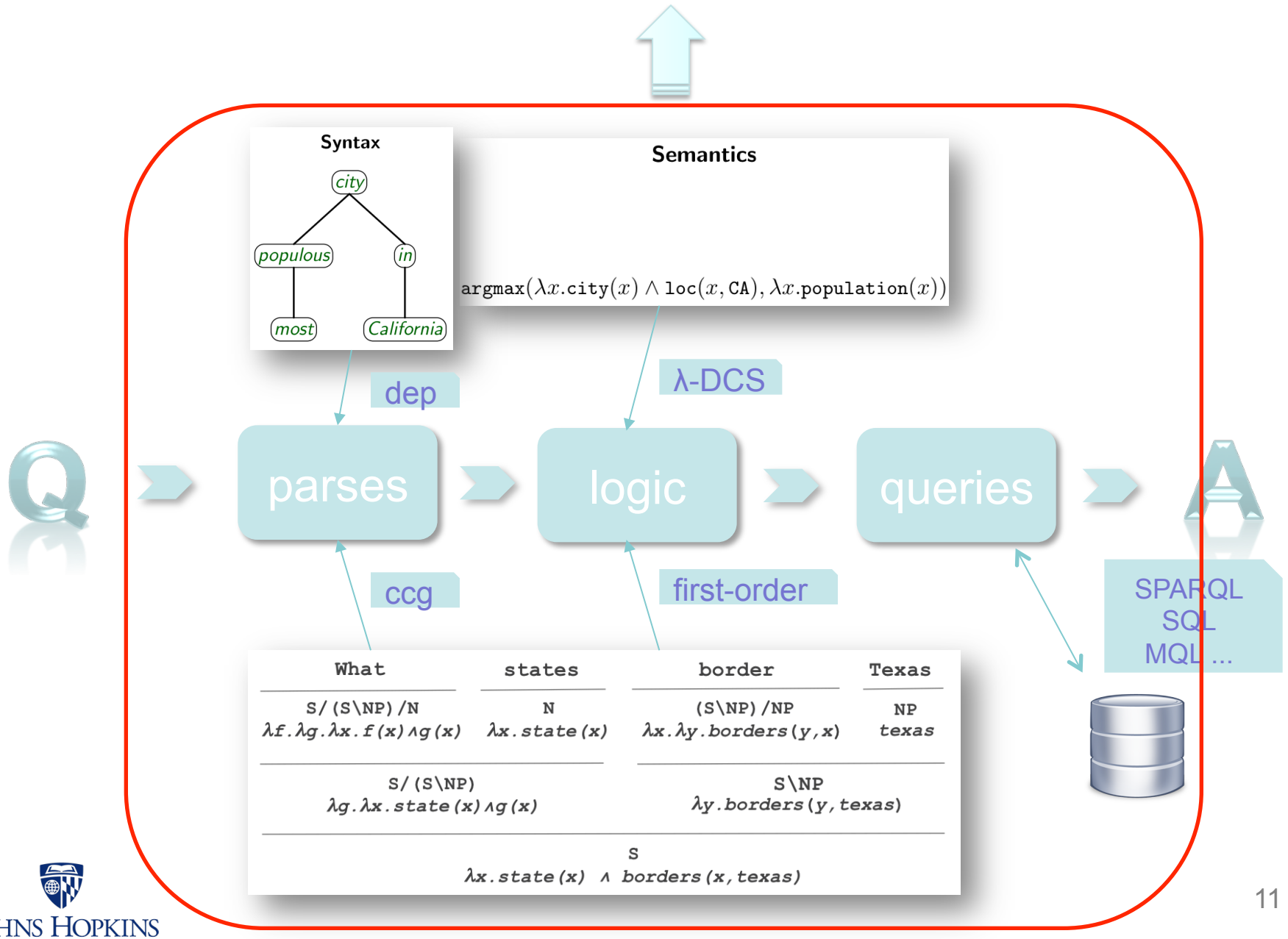


Previous Approach: Semantic Parsing



Author	Title	Year
Chen and Mooney	{Learning to Interpret Natural Language Navigation Instructions fro...	2011
Wong and Mooney	Learning synchronous grammars for semantic parsing with lambd...	2007
Kate and Mooney	Using string-kernels for learning semantic parsers	2006
Ge and Mooney	A statistical semantic parser that integrates syntax and semantics	2005
Thompson and Mooney	Acquiring word-meaning mappings for natural language interfaces	2003
Tang and Mooney	Using multiple clause constructors in inductive logic programming...	2001
Zelle and Mooney	Learning to parse database queries using inductive logic program...	1996

Is this how YOU find the answer?



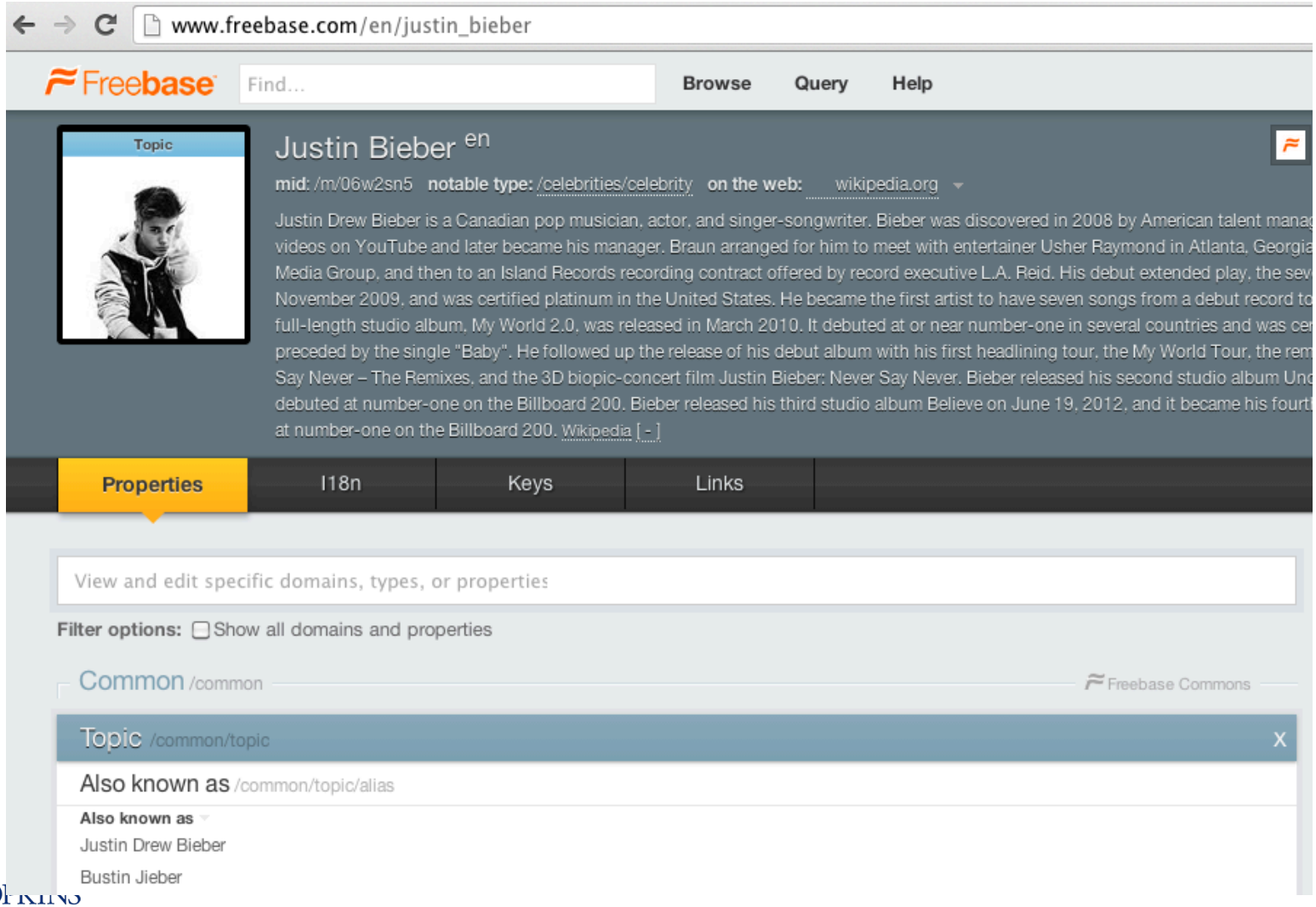
JOHNS HOPKINS UNIVERSITY

this instead might be how you find the answer

Question:

Who is the brother of Justin Bieber?

who is the brother of Justin Bieber?
1st step: go to JB's Freebase page



who is the brother of Justin Bieber?

2nd step: maybe wander around a bit?

Celebrity /celebrities/celebrity

Romantic relationships (with celebrities) /celebrities/celebrity/sexual_relationships

Celebrity	Relationship type	Start date	End date
Caitlin Beadles	Dated	2007	2007
Selena Gomez	Dated	2011 February	-

Sexual orientation /celebrities/celebrity/sexual_orientation

Sexual orientation	Start	End

Legal entanglements /celebrities/celebrity/legal_entanglements

Offense	Location	Date
Driving under the influence	Miami	1/23/2014
Driving under the influence	Miami Beach	1/23/2014

Substance abuse problems /celebrities/celebrity/substance_abuse_problems

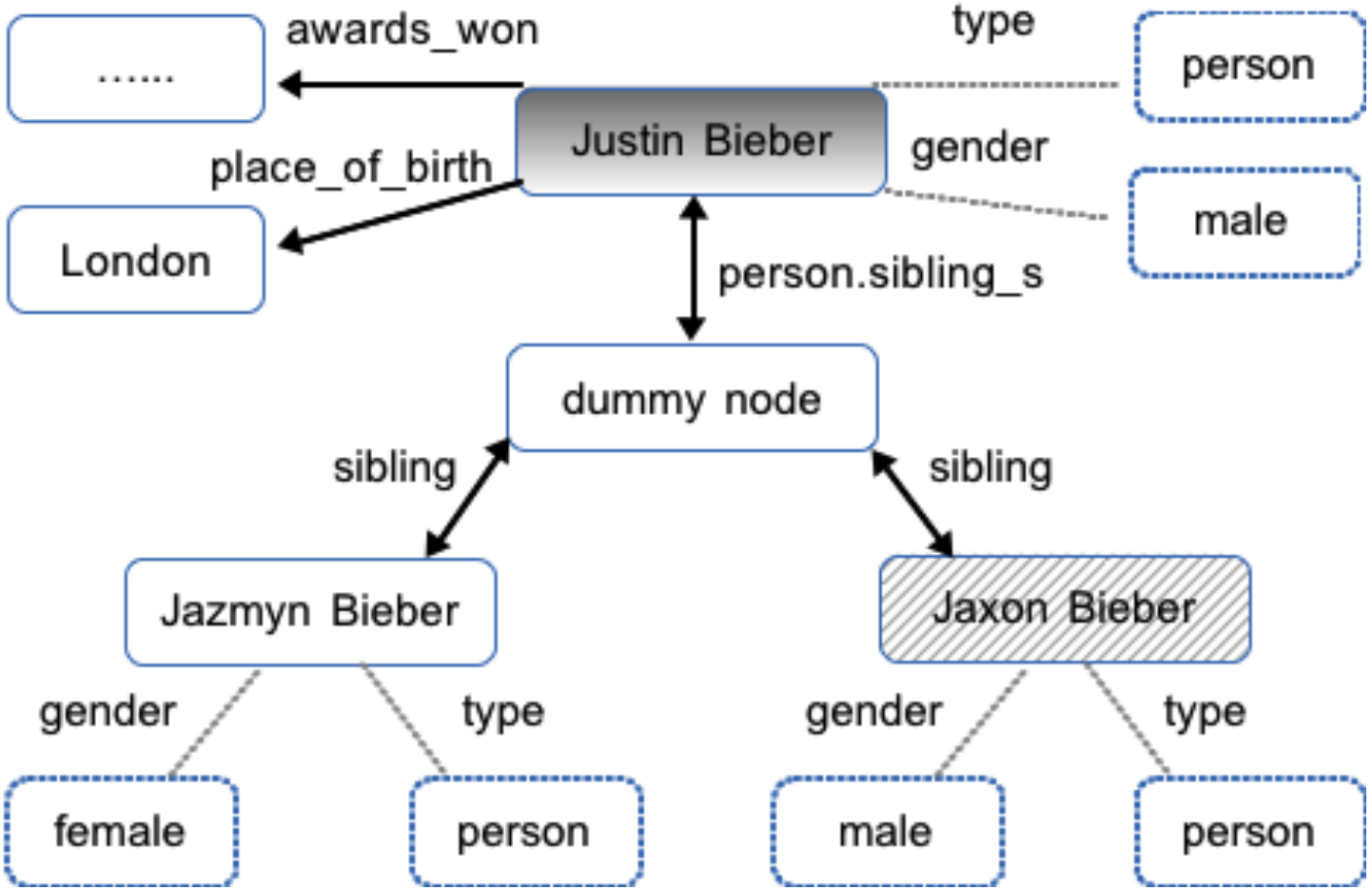
Substance	Start	End

who is the brother of Justin Bieber?

finally: oh yeah, his brother

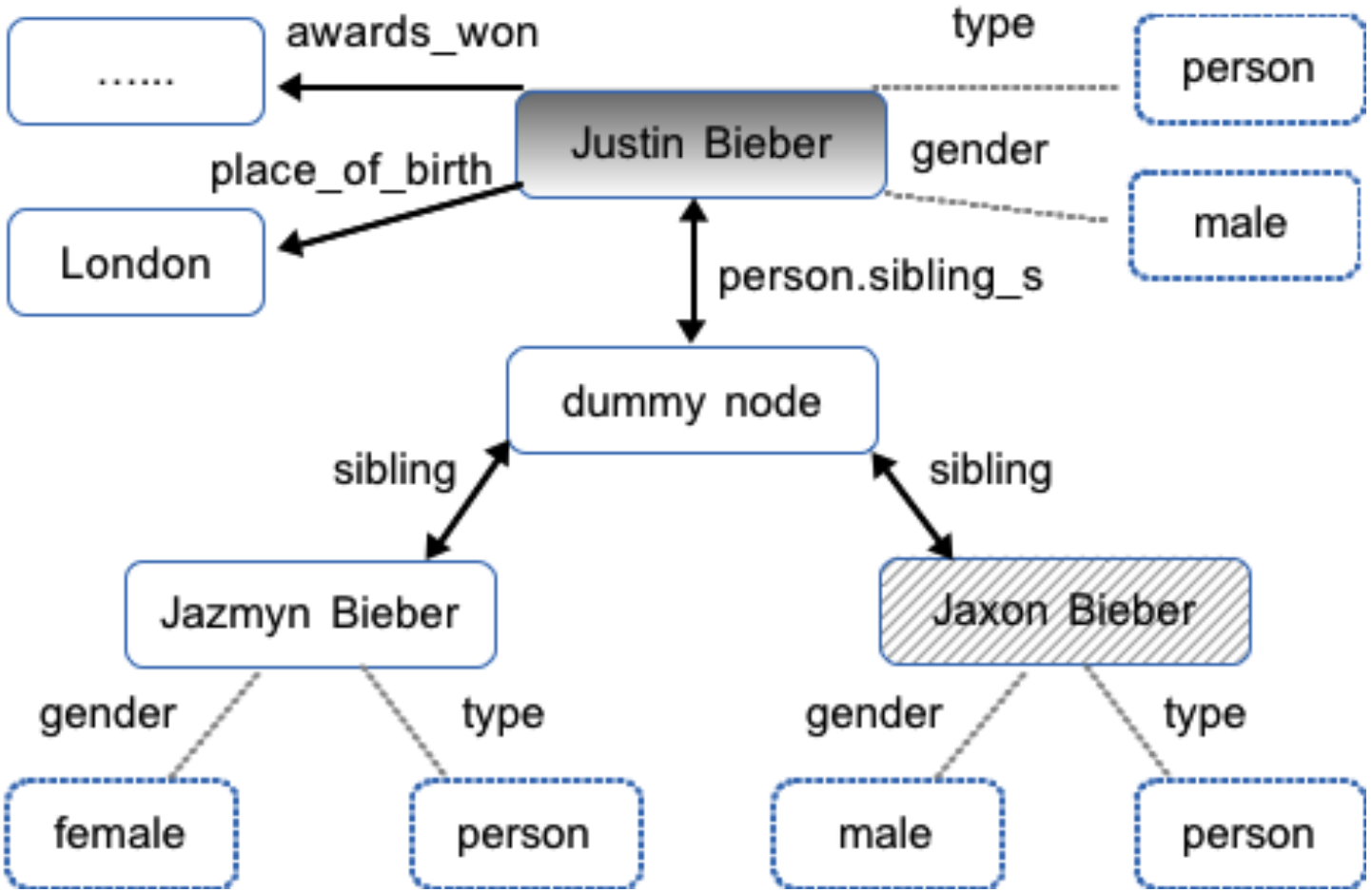
Freebase Topic Graph

we know **just enough** about the answer from the following view:



Freebase Topic Graph

who is the brother of Justin Bieber?



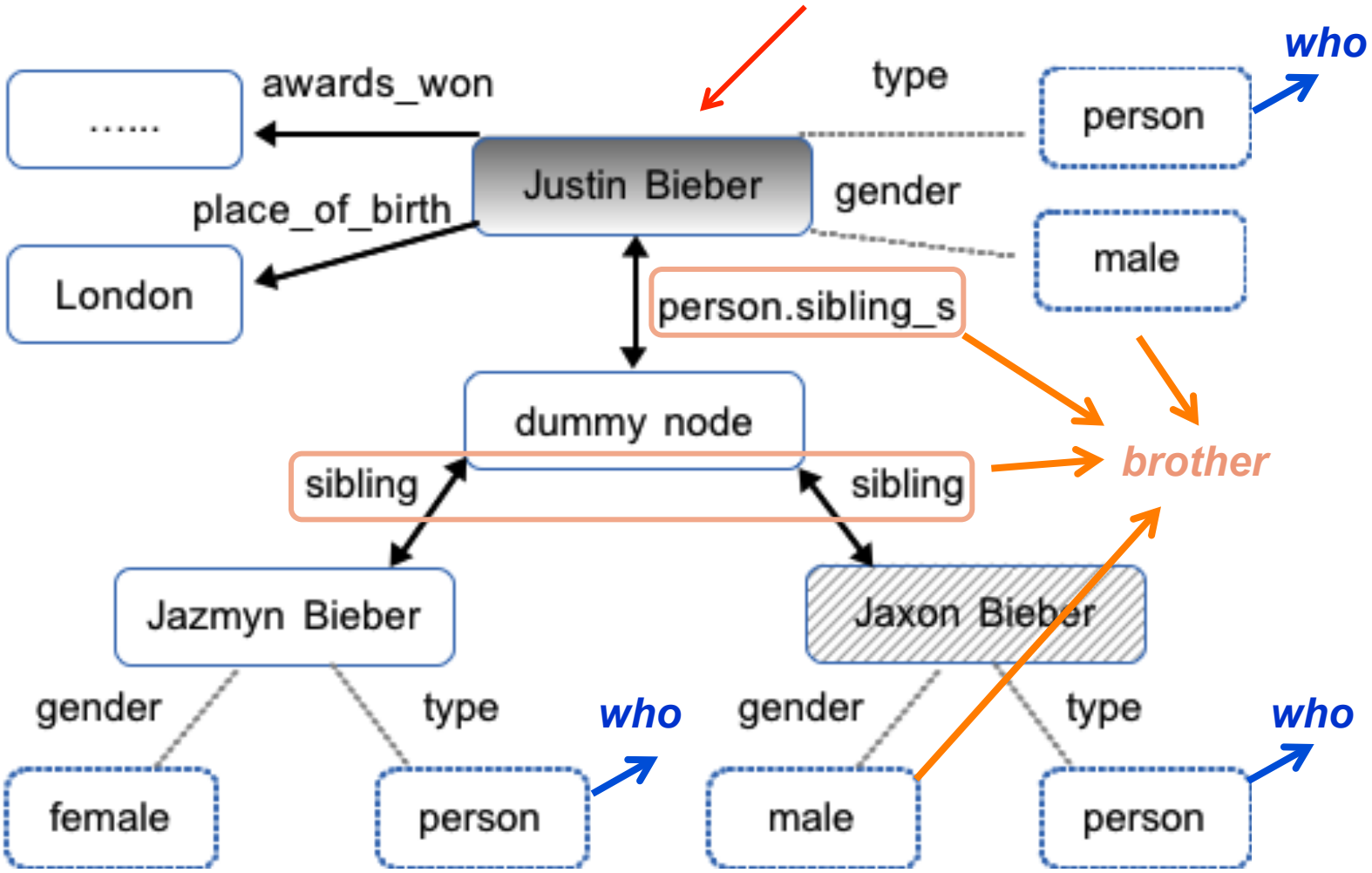
Signals!

Major challenge for Question Answering:

finding indicative (linguistic) signals for answers

Freebase Topic Graph

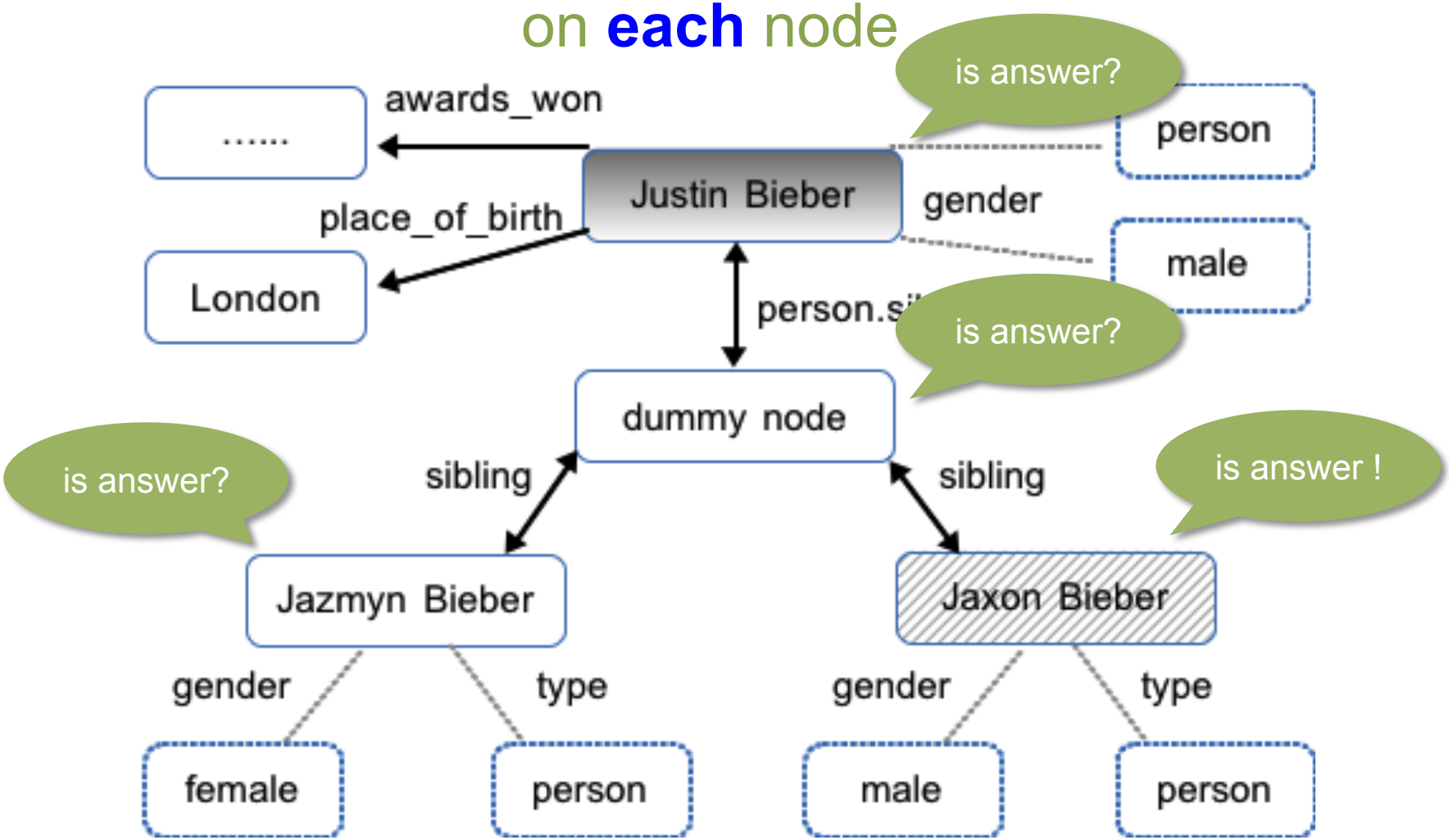
who is the brother of Justin Bieber?



JOHNS HOPKINS UNIVERSITY

QA on Freebase

is now a binary classification problem on **each** node



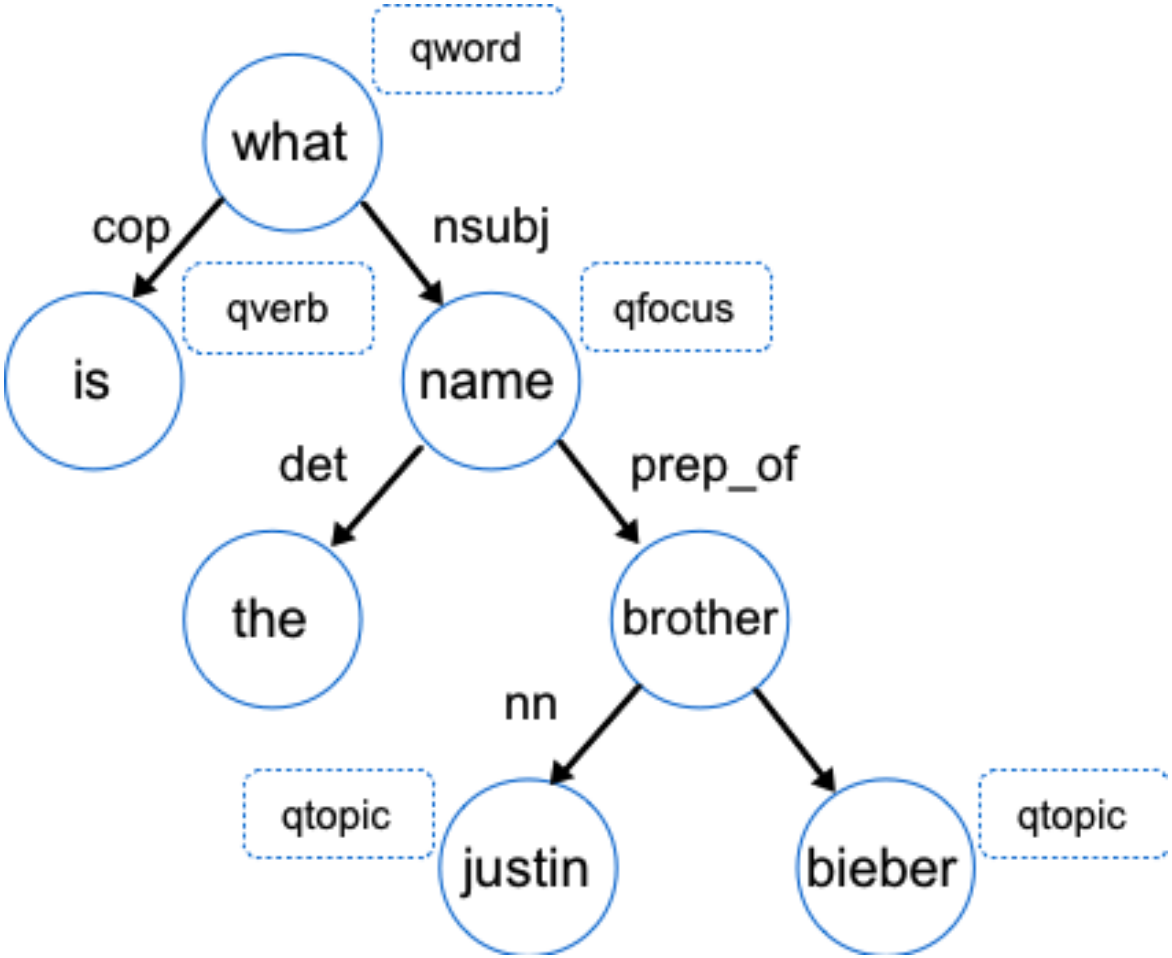
Features on Graph

extract features for *each* node

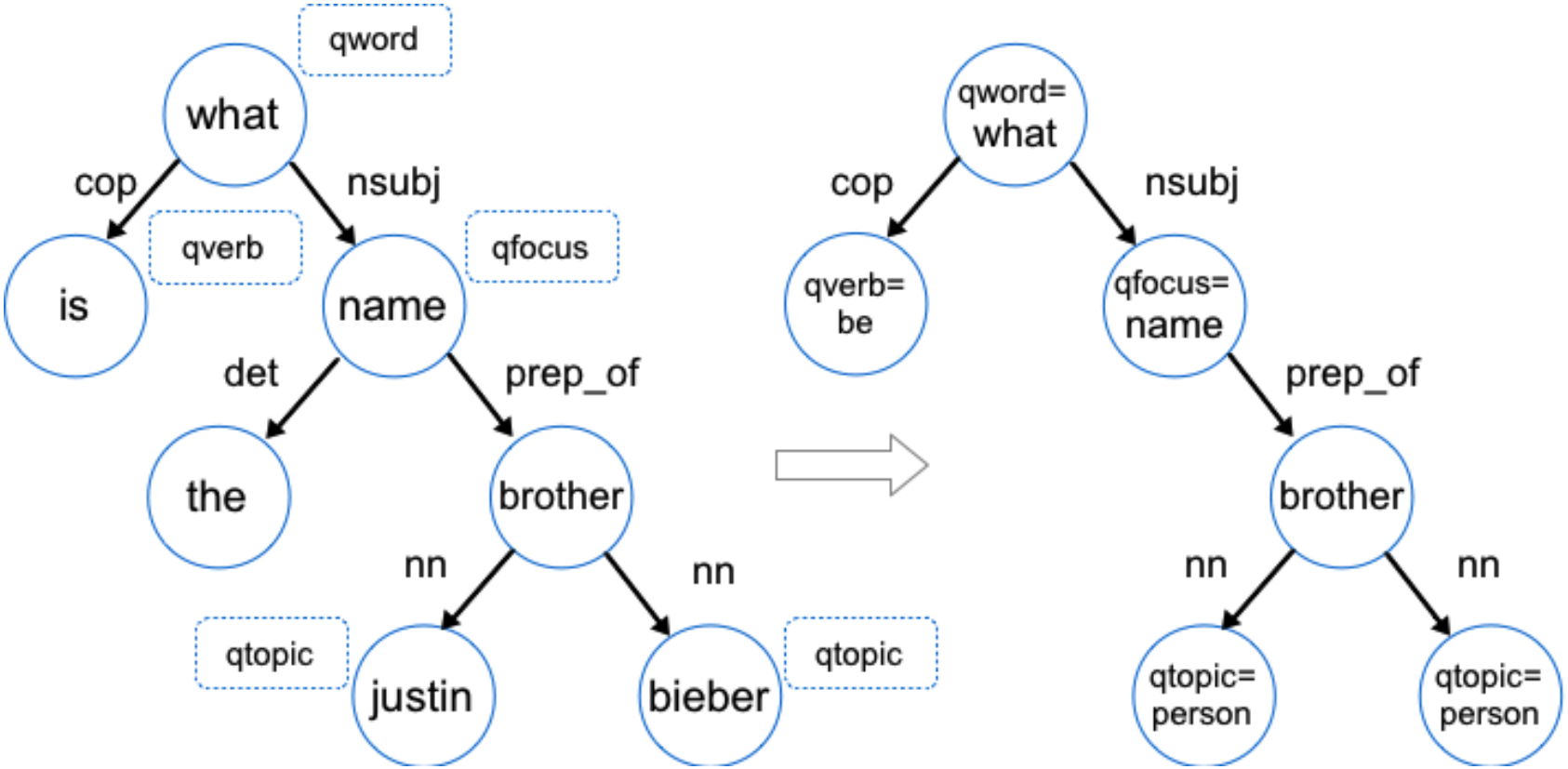
Justin Bieber	Jazmyn Bieber	Jaxon Bieber
• has:awards_won • has:place_of_birth • has:sibling • type:person • …	• has:sibling • gender:female • type:person • …	• has:sibling • gender:male • type:person • …

brown: relation; relations connect to other nodes
blue: property; properties have literal values.

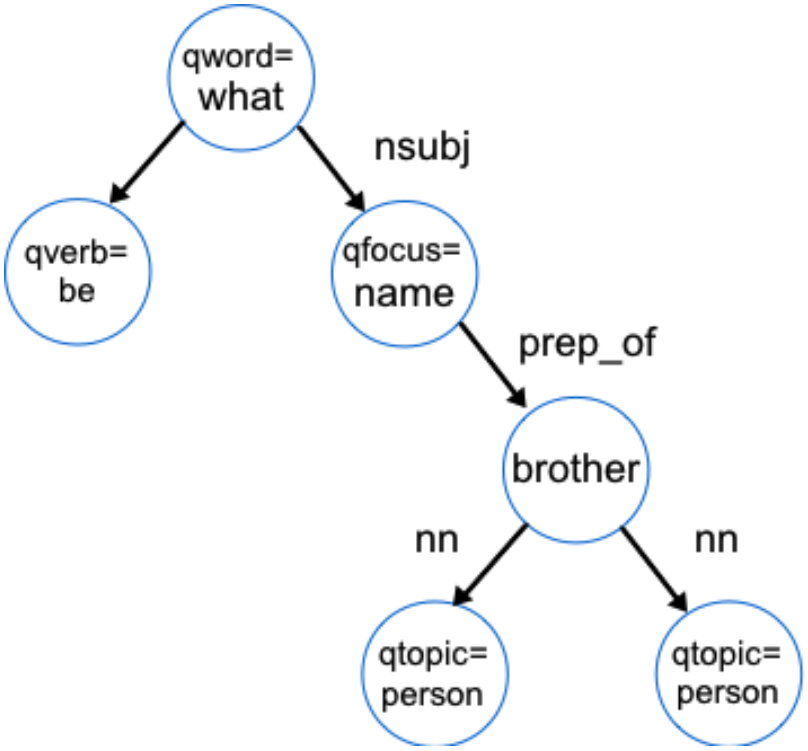
What do we know about the question?



What do we know about the question?



Features on Question

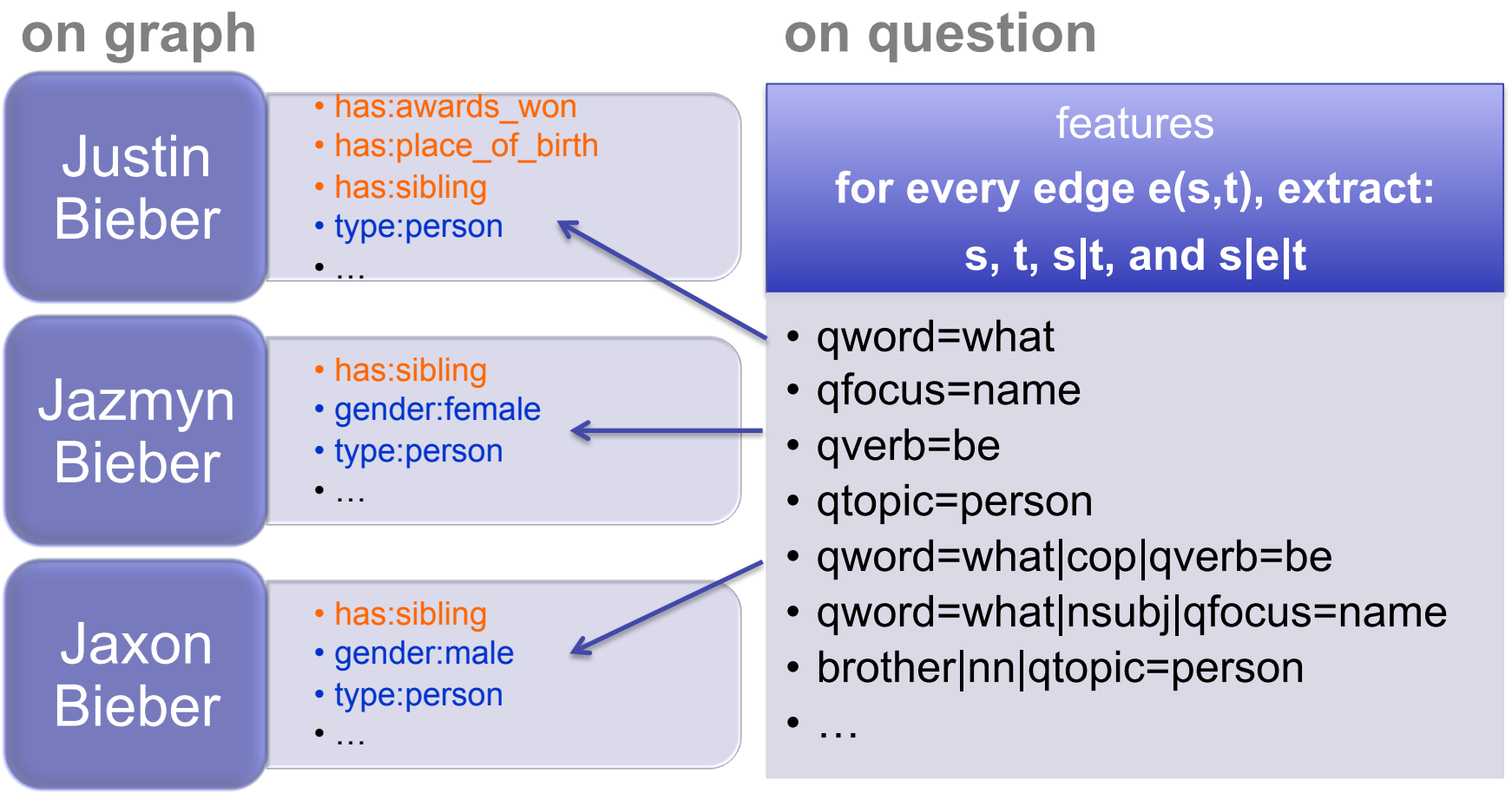


features

for every edge e(s,t), extract: s, t, s|t, and s|e|t

- qword=what
- qfocus=name
- qverb=be
- qtopic=person
- qword=what|cop|qverb=be
- qword=what|nsubj|qfocus=name
- brother|nn|qtopic=person
- …

Combining Graph Features with Question Features



Some Combined Features Can Be Helpful

Justin Bieber	expected weights
• has:awards_won | qword=what	• medium
• has:place_of_birth | qword=what	• low
• has:sibling | qfocus=name	• low
• type:person | qfocus=name	• medium
• …	• …

Jazmyn Bieber	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:female | brother | nn | qtopic=person	• low
• type:person | brother | nn | qtopic=person	• high
• …	• …

Jaxon Bieber (is answer)	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:male | brother | nn | qtopic=person	• high
• type:person | qword=what | nsubj | qfocus=name	• high
• …	• …

brown: relation; relations connect to other nodes
blue: property; properties have literal values.
red: question features.

JOHNS HOPKINS UNIVERSITY

Some Combined Features Can Be Helpful

Justin Bieber	expected weights
• has:awards_won | qword=what	• medium
• has:place_of_birth | qword=what	• low
• has:sibling | qfocus=name	• low
• type:person | qfocus=name	• medium
• …	• …

Jazmyn Bieber	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:female | brother | nn | qtopic=person	• low
• type:person | brother | nn | qtopic=person	• high
• …	• …

Jaxon Bieber (is answer)	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:male | brother | nn | qtopic=person	• high
• type:person | qword=what | nsubj | qfocus=name	• high
• …	• …

brown: relation; relations connect to other nodes
blue: property; properties have literal values.
red: question features.

JOHNS HOPKINS UNIVERSITY

Some Combined Features Can Be Helpful

Justin Bieber	expected weights
• has:awards_won | qword=what	• medium
• has:place_of_birth | qword=what	• low
• has:sibling | qfocus=name	• low
• type:person | qfocus=name	• medium
• …	• …

Jazmyn Bieber	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:female | brother | nn | qtopic=person	• low
• type:person | brother | nn | qtopic=person	• high
• …	• …

Jaxon Bieber (is answer)	expected weights
• has:sibling | brother | nn | qtopic=person	• high
• gender:male | brother | nn | qtopic=person	• high
• type:person | qword=what | nsubj | qfocus=name	• high
• …	• …

brown: relation; relations connect to other nodes
blue: property; properties have literal values.
red: question features.

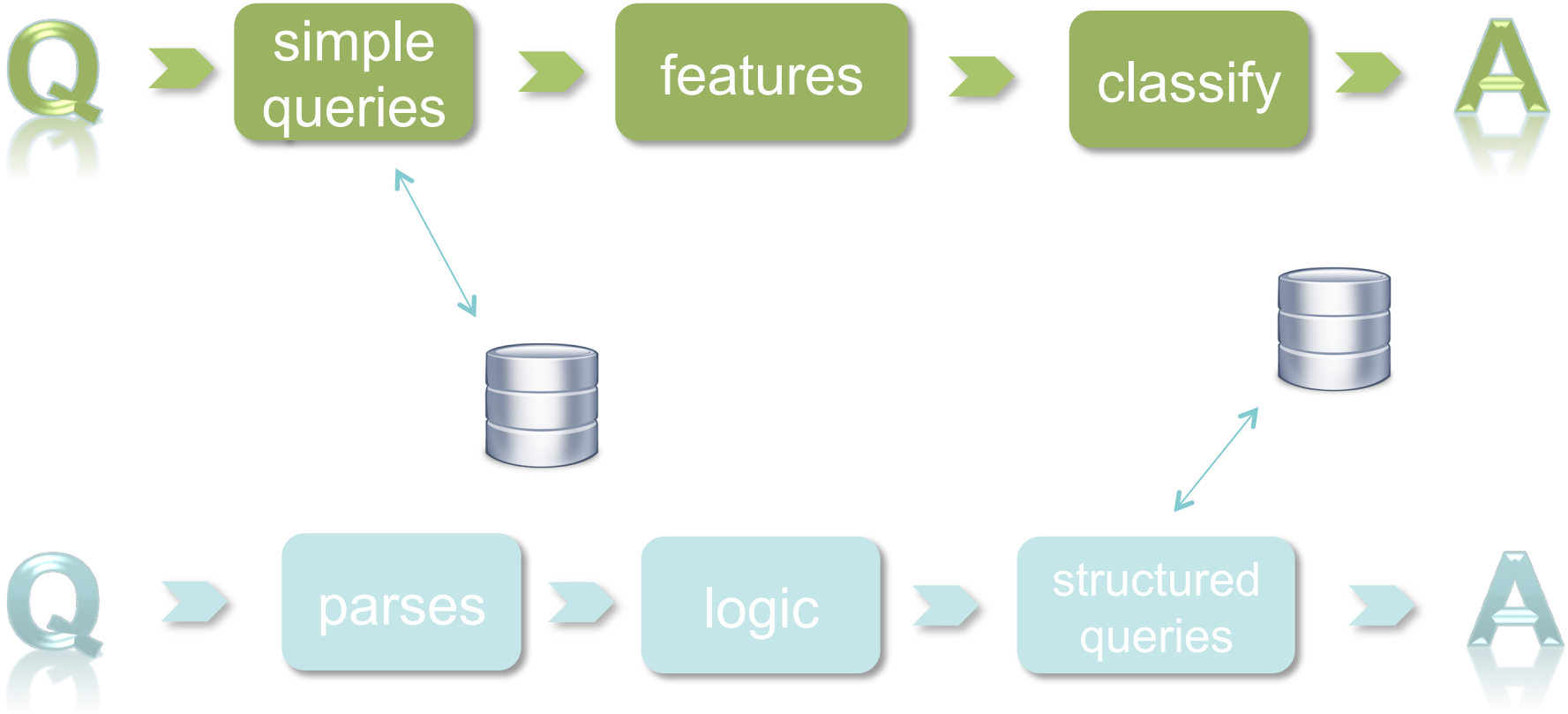
JOHNS HOPKINS UNIVERSITY

Information Extraction

***who** is the **brother** of **Justin Bieber**?*

Justin Bieber?
Jaxon Bieber
Jasmin Bieber?
…

Information Extraction
vs. Semantic Parsing

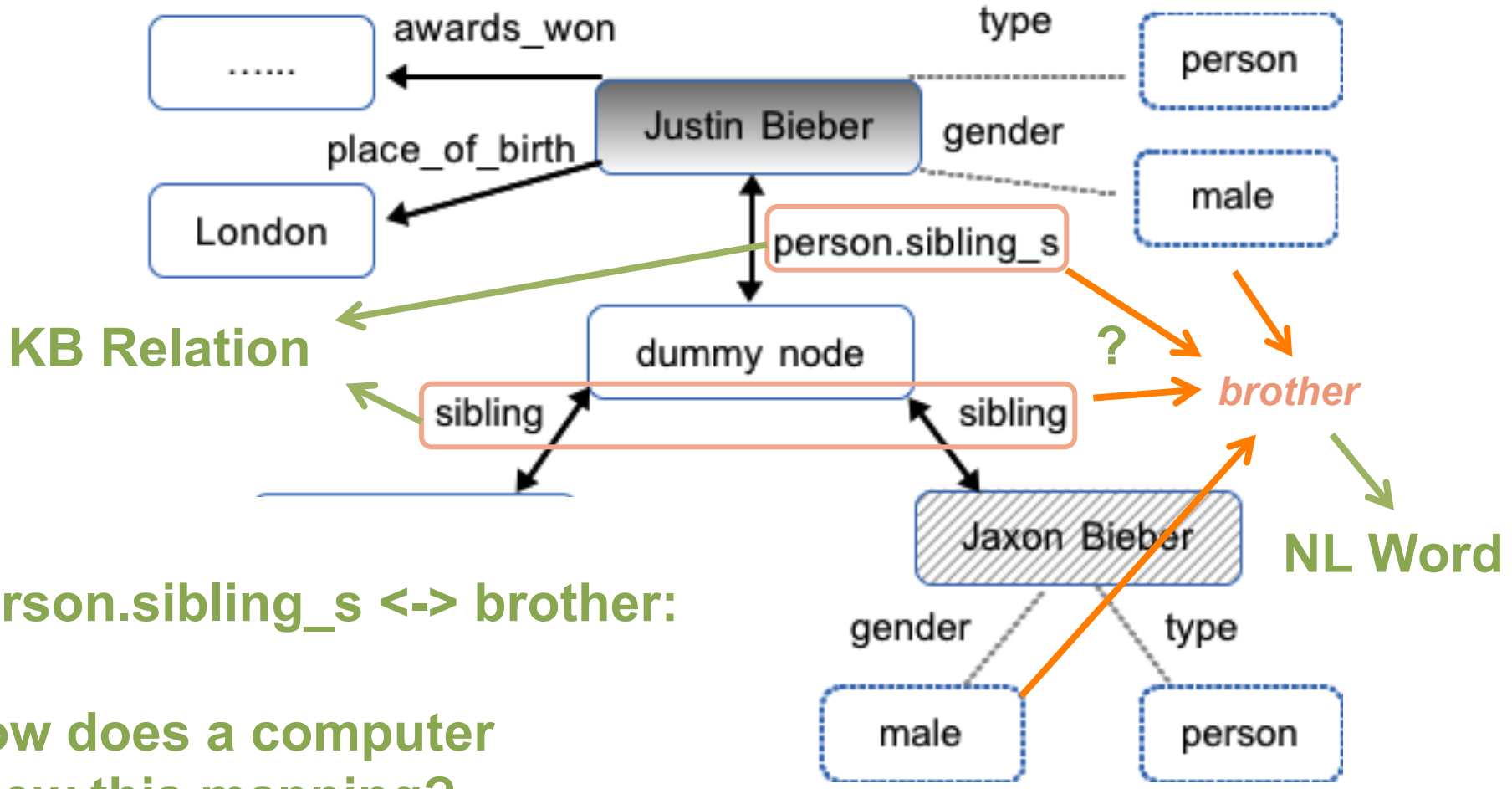


QA from KB

The Data Challenge

Freebase Topic Graph

*who is the **brother** of Justin Bieber?*



person.sibling_s <-> brother:

How does a computer know this mapping?

The Challenge
aligning KB relations with NL words

- KB entry:
  - **film/starring** (Gravity, Bullock/Clooney)

- How questions can be asked:
  - what's **the cast of** Gravity?
  - who **played/acted in** Gravity?
  - who **starred in** Gravity?
  - show me **the actors in** Gravity.

Aligning KB Relations with NL Words

- annotated ClueWeb (with Freebase entities), released by Google
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity

Aligning KB Relations with NL Words

- annotated ClueWeb (with Freebase entities), thanks to Google
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity

Aligning KB Relations with NL Words

- Input: **film/starring** (Gravity, Sandra Bullock)
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity
- Task: find NL words that express **film/starring**

Aligning KB Relations with NL Words

- Input: **film/starring** (Gravity, Sandra Bullock)
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity
- Task: find NL words that express **film/starring**

Aligning KB Relations with NL Words

- maps the NL phrases to KB relations film/starring:
  - Sandra then was cast in Gravity, a two actor spotlight film
  - Sandra Bullock plays an astronaut hurtling through space in new blockbuster "Gravity"
  - Sandra Bullock stars/acts in Gravity
  - Sandra Bullock conquered her fears to play the lead in Gravity

- in massive scale:
  - Freebase: 40 million entities, 2.5 billion facts
  - ClueWeb09 Annotation: 5 billion entities in 340 million documents (5TB compressed)

- very simple solution:
  - treat it as an alignment problem (IBM Model 1)
  - fire up GIZA++ and hundreds of computers

Samples of CluewebMapping

film.actor	• won, star, among, show, …
film.directed_by	• director, direct, by, with, …
celebrity.infidelity.victim	• Jennifer Aniston…
celebrity.infidelity.participant	• you know who…

Samples of CluewebMapping

film.actor	• won, star, among, show, …
film.directed_by	• director, direct, by, with, …
celebrity.infidelity.victim	• Jennifer Aniston…
celebrity.infidelity.participant	• you know who…

Samples of CluewebMapping

film.actor	• won, star, among, show, …
film.directed_by	• director, direct, by, with, …
celebrity.infidelity.victim	• Jennifer Aniston…
celebrity.infidelity.participant	• you know who…

Samples of CluewebMapping

film.actor	• won, star, among, show, …
film.directed_by	• director, direct, by, with, …
celebrity.infidelity.victim	• Jennifer Aniston…
celebrity.infidelity.participant	• you know who…

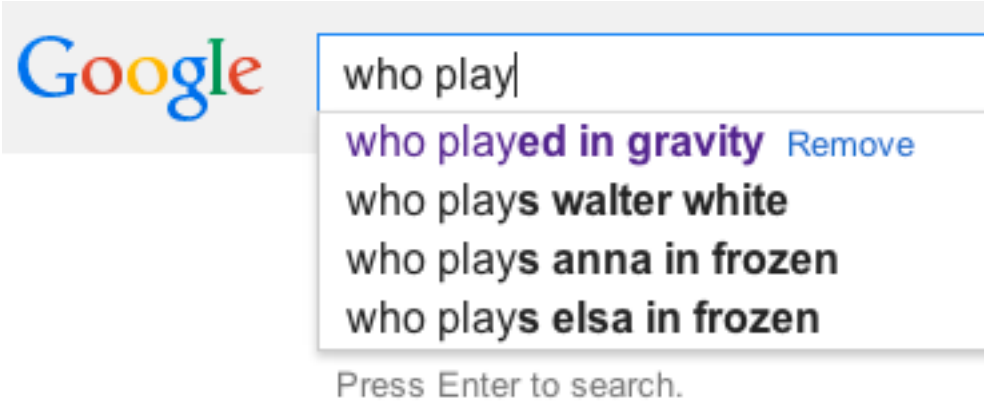
Using KB Alignment as Features

- Who is the brother of Justin Bieber?
- predictions from KB alignment:
  - /people/sibling_relationship/sibling
  - /fictional_universe/ sibling_relationship_of_fictional_characters/siblings
  - …

- Features: the rank (top 1/3/5/50…) of node's relation predicted by KB alignment

Evaluation

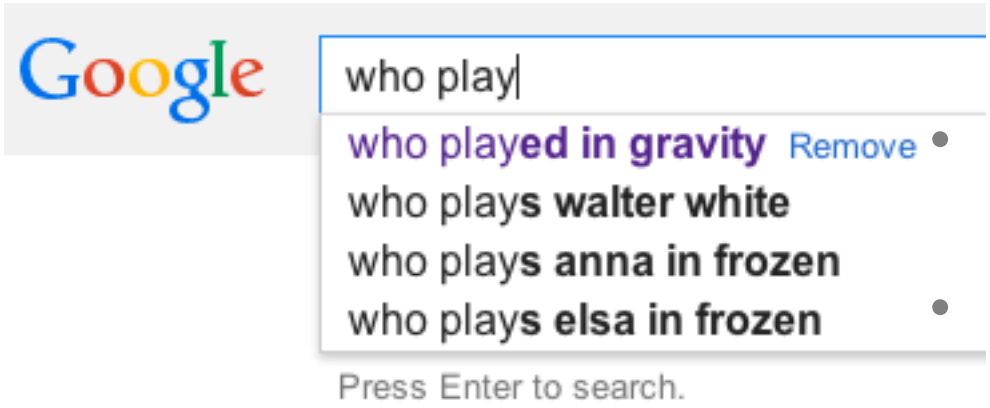
- Data: WebQuestions
- Berant et. al. (2013)
- 5810 questions annotated from 1 million crawled off Google Suggest



JOHNS HOPKINS UNIVERSITY

Evaluation

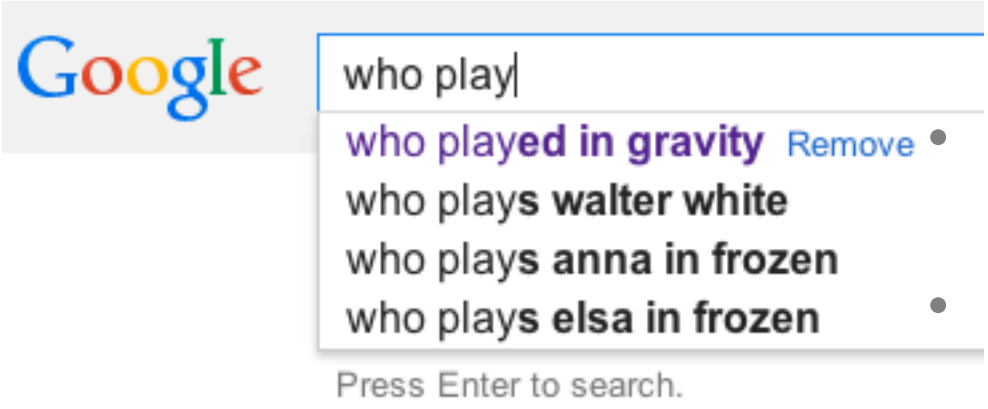
- Data: WebQuestions
- Berant et. al. (2013)
- 5810 questions annotated from crawling off Google Suggest



- which states does the connecticut river flow through?
- who does david james play for 2011?
- what date was john adams elected president?
- what kind of currency does cuba use?
- who owns the cleveland browns?

Evaluation

- Data: WebQuestions
- Berant et. al. (2013)
- 5810 questions annotated from crawling off Google Suggest



- which states does the connecticut river flow through?
- who does david james play for 2011?
- what date was john adams elected president?
- what kind of currency does cuba use?
- who owns the cleveland browns?

JOHNS HOPKINS UNIVERSITY

Evaluation

- Tag named entities with Stanford CoreNLP (caseless model)
- Search named entities using the Freebase Search API
- Retrieve topics using the Freebase Topic API

- which states does the connecticut river flow through?
- who does david james play for 2011?
- what date was john adams elected president?
- what kind of currency does cuba use?
- who owns the cleveland browns?

Evaluation

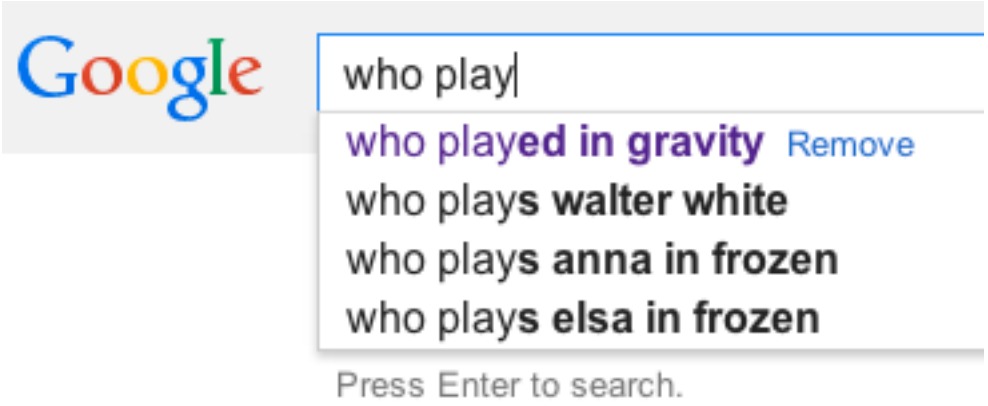
- Tag named entities with Stanford CoreNLP (caseless model)
- Search named entities using the Freebase Search API
- Retrieve topics using the Freebase Topic API

who did natalie portman play in star wars?

topic	score
natalie_portman	722
star_wars	233
saturday_night_live_season_31	56
clone_wars	51
lego_star_wars	38
star_wars_music	37
star_wars_episode_iv_a_new_hope	36
star_wars_episode_i_the_phantom_menace	35

Evaluation & Training

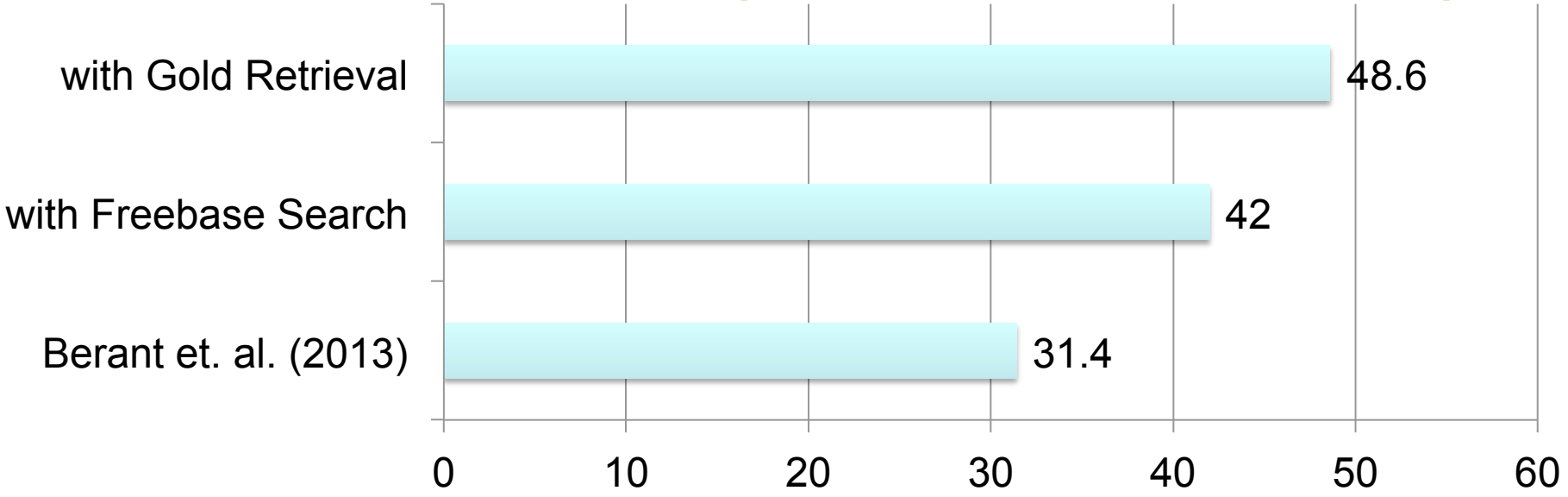
- Data: WebQuestions
- Berant et. al. (2013)
- 5810 questions annotated from crawling off Google Suggest



Training

- L1 regularized Logistic Regression with high-performance ML tool: Classias (Okazaki, 2009)
- original: 7 million feature types extracted
- training: 4 hours
- after: 30 thousand features with non-zero weight

F1 on TEST (2032 questions)



feature	weight	feature	weight
qfocus=religion|type=Religion	8.60	qword=when|type=datetime	5.11
qfocus=money|type=Currency	5.56	qverb=border|rel=location.adjoins	4.56
qverb=die|type=CauseOfDeath	5.35	qverb=go|qtopic=location| type=Tourist attraction	2.94

Free917?

(Cai and Yates 2013)

- jacana-freebase was not designed to handle **argmax()**, **count()** operators out of the box
  - **how many** companies are traded by the nyse
  - **how many** teams participate in the uefa
  - **how many** ships has nathanael herreshoff designed
  - in what year did motorola have the **most** revenue

Conclusion

- Model: an Information Extraction approach with massive features
  - a naïve baseline for semantic parsing based QA!
- Data: helps with mapping between NL words and KB relations
  - CluewebMapping: ~3000 Freebase relations <-> 10,000 words
- Wednesday 10:45–11:10 Semantics III
  - **Semantic Parsing via Paraphrasing.**
  - Jonathan Berant and Percy Liang
- Thursday, Semantic Parsing Workshop
  - **Freebase QA: Information Extraction or Semantic Parsing?**
  - Xuchen Yao, Jonathan Berant and Benjamin Van Durme

thank you

Error Analysis

20%~25%: Annotation Error

Question	MTurk answer
what state does selena gomez?	New York City
how old is sacha baron cohen?	a URL
what two countries invaded poland in the beginning of ww2?	Germany
which countries border the us?	Canada
where is rome italy located on a map?	Rome
how much did adriana lima gain during pregnancy?	Spike Guys' Choice Awards
what does thai mean?	Language
which wife did king henry behead?	Anne of the Thousand Days
what are the major cities in france?	Paris
what season did tony soprano get shot?	The Sopranos

15%~20%: “complicated” questions

- what did james k polk do before he was president?
- what is the oregon ducks 2012 football schedule?
- what country did germany invade first in ww1?
- who is governor of ohio 2011?
- when did charles goodyear invented rubber?
- who did france surrender to in ww2?
- who did george w. bush run against for the second term?
- who was the leader of soviet union during wwii?

5%~10%: answer typing failure

- what things did martin luther king do?
- what town was martin luther king assassinated in?
- what electorate does anna bligh represent?
- what channel is the usa pageant on?
- what are some of the traditions of islam?
- what is the state flower of arizona?
- what did the islamic people believe in?
- what did the scientist chadwick discovered?

Other errors

- Freebase search error (10%)
- ill-formed web text (2% ~ 3%)