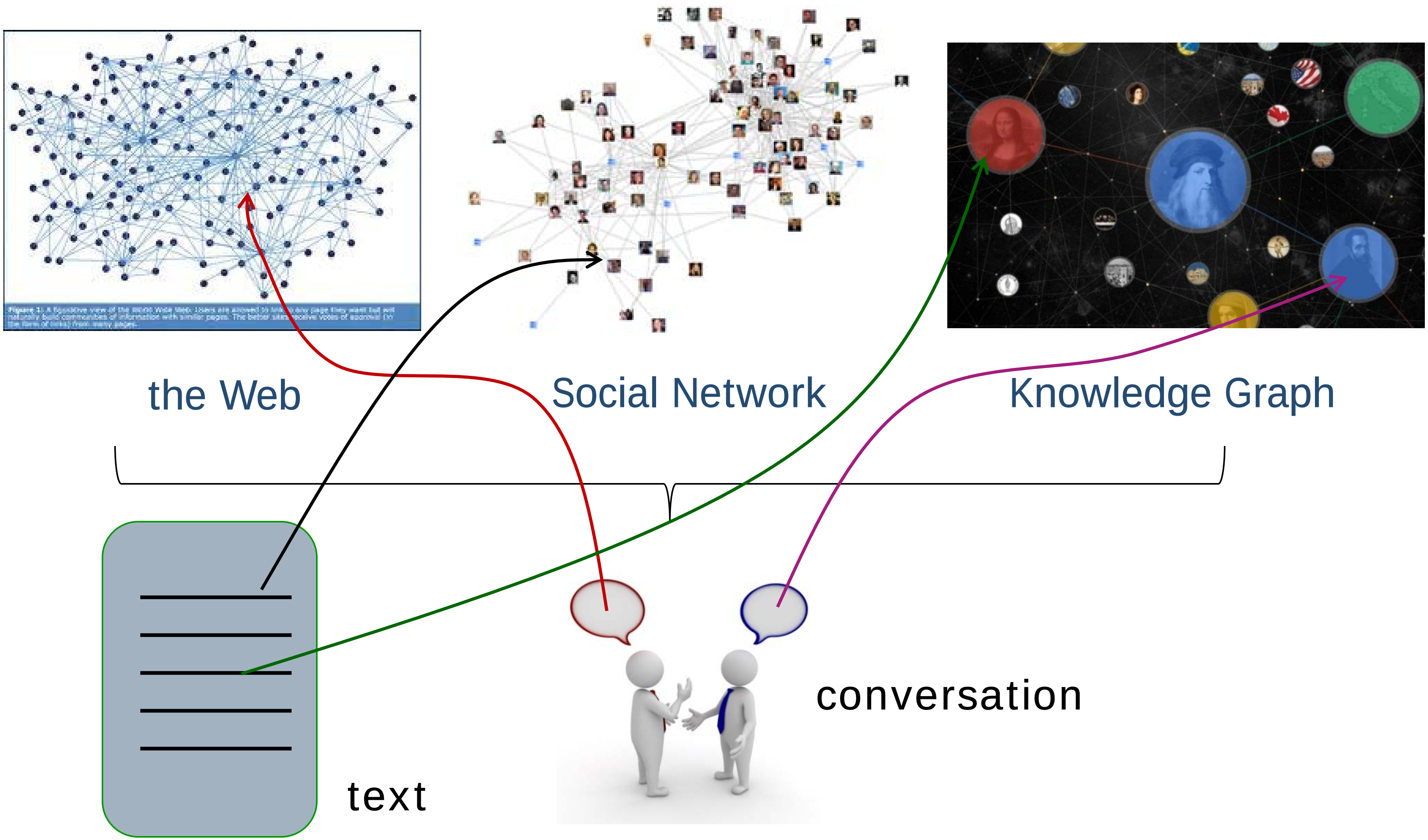
Question Answering Benchmark and Semantic Parsing

Xifeng Yan

University of California, Santa Barbara

Thanks to Yu Su, Semih Yavuz, Izzeddin Gur, Huan Sun, who did the work

Graph Data

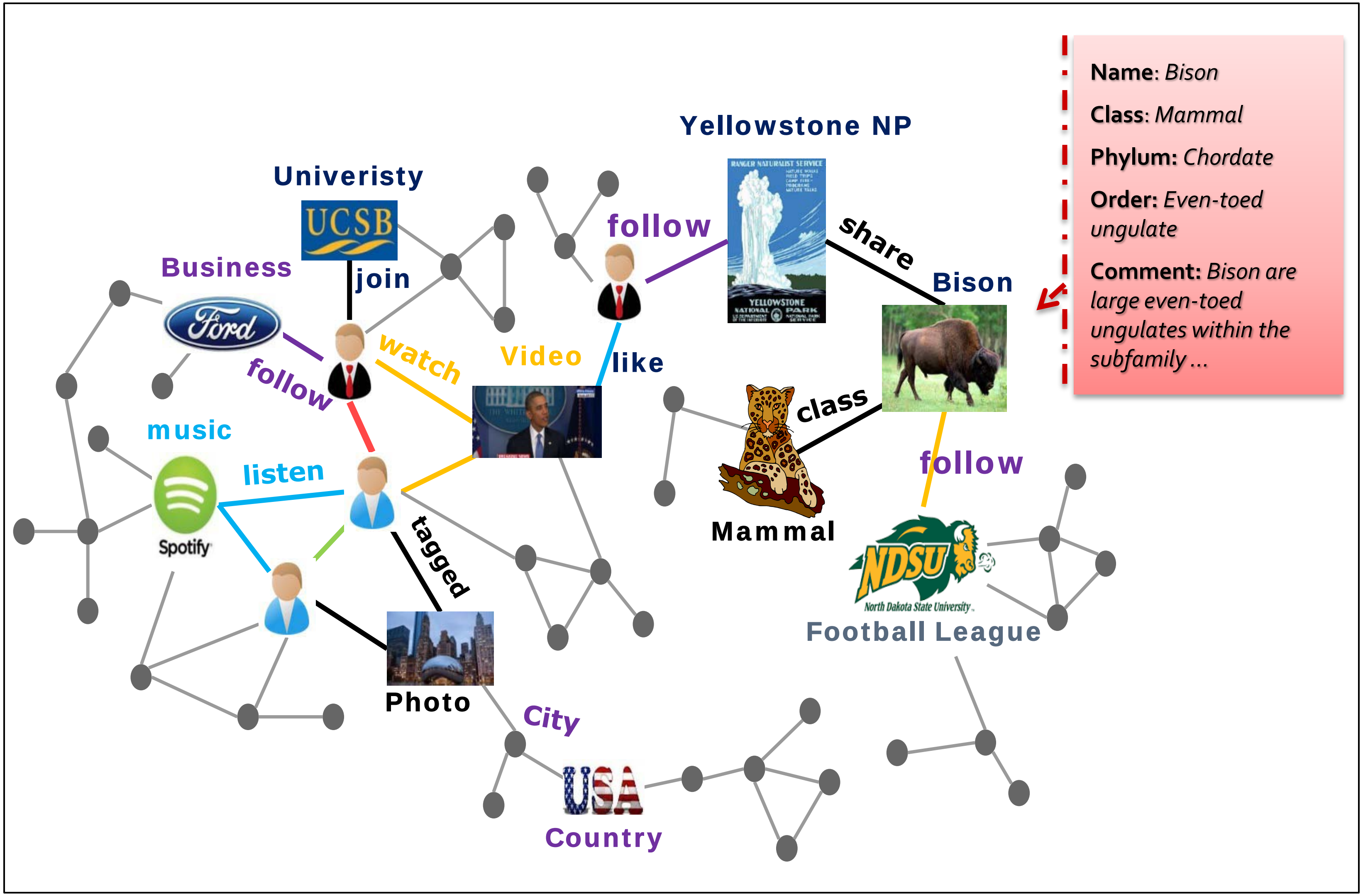


Question Answering

Natural Language Interface

- Text Based Question Answering
- Knowledge Graph Based Question Answering
- FAQ/Quora Style Question Answering

Knowledge Graph



Query Graph Data

```
g.V().has("person","name","gremlin").
out("knows").out("created").
hasLabel("project").
```

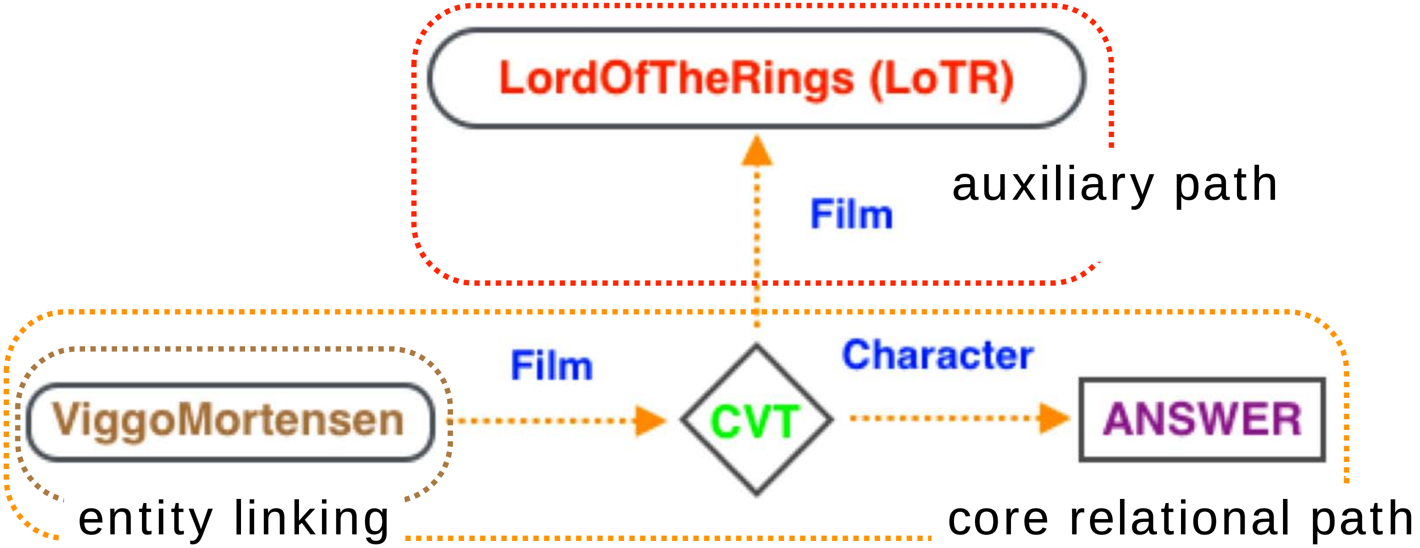
Gremlin in Titan

```
SELECT AVG(?stars)
WHERE { ?a v:label person .
?a v:name "gremlin" .
?a e:knows ?b .
?b e:created ?c .
?c v:label "project" .
?c v:stars ?stars }
```

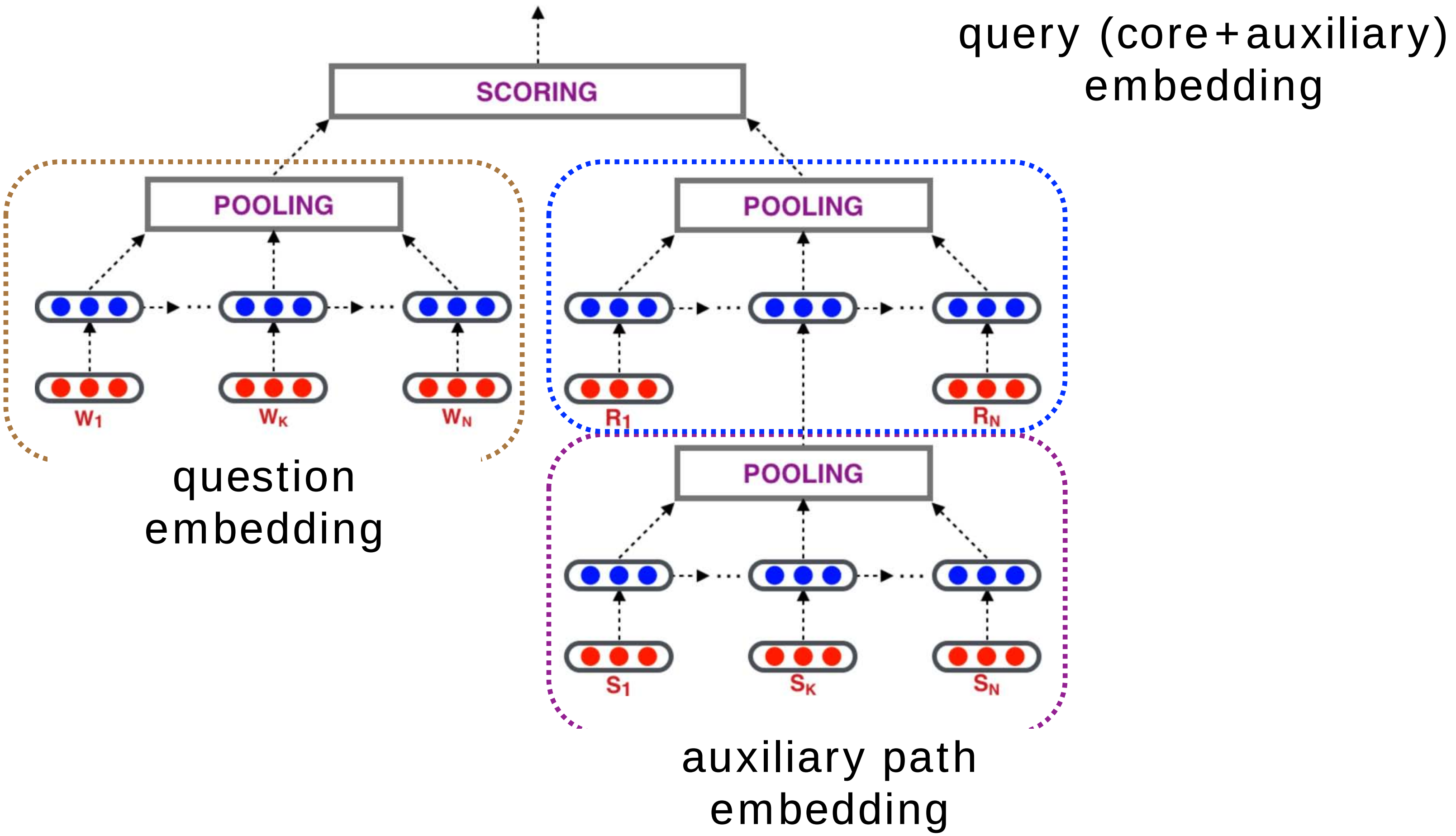
SPARQL

NL Question to Knowledge Graph Mapping

Who did [Viggo Mortensen] play in Lord of the Rings ?



Using Recurrent Neural Network



On Generating Characteristic-rich Question Sets for QA Evaluation (EMNLP’16) , with Yu Su

Characteristic-rich Question Sets (EMNLP16): Motivation

- Existing datasets for semantic parsing/question answering (QA) over knowledge bases mainly concern simple questions (question=utterance)

"Where was Obama born?"

"What party did Clay establish?"

"What kind of money to take to Bahamas?"

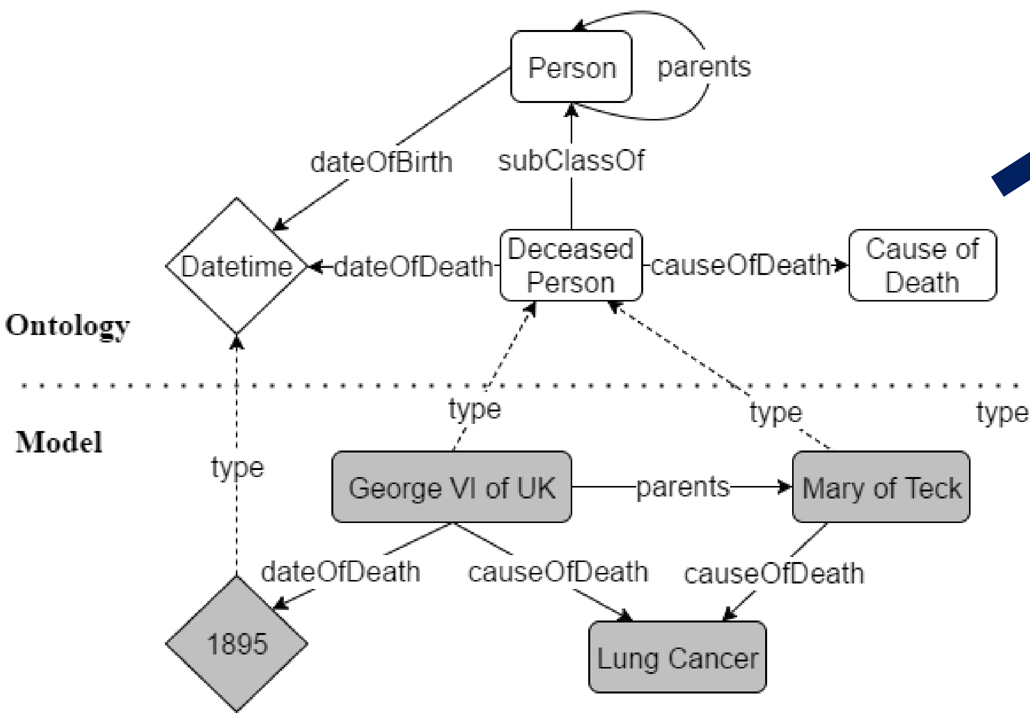
… …

Real-world questions have rich characteristics

- Structural complexity
  - *"Who was the coach when Michael Jordan stopped playing for the Chicago Bulls?"*
- Quantitative analysis (functions)
  - *"What is the best-selling smartphone in 2015?"*
- Commonness
  - *"Where was Obama born?"* vs.
  - *"What is the tilt of axis of Polestar?"*
- Paraphrase
  - *"What is the nutritional composition of coca-cola?"*
  - *"What is the supplement information for coca-cola?"*
  - *"What kind of nutrient does coke have?"*
- …

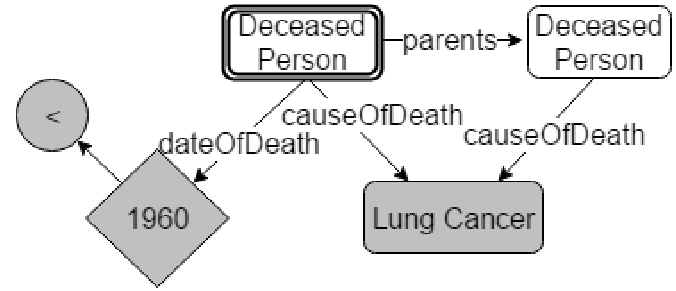
Can we generate questions with rich characteristics from a knowledge base?

Logical Form: Graph Query



Knowledge Base

24K classes, 65K relations, 41M entities, 596M facts

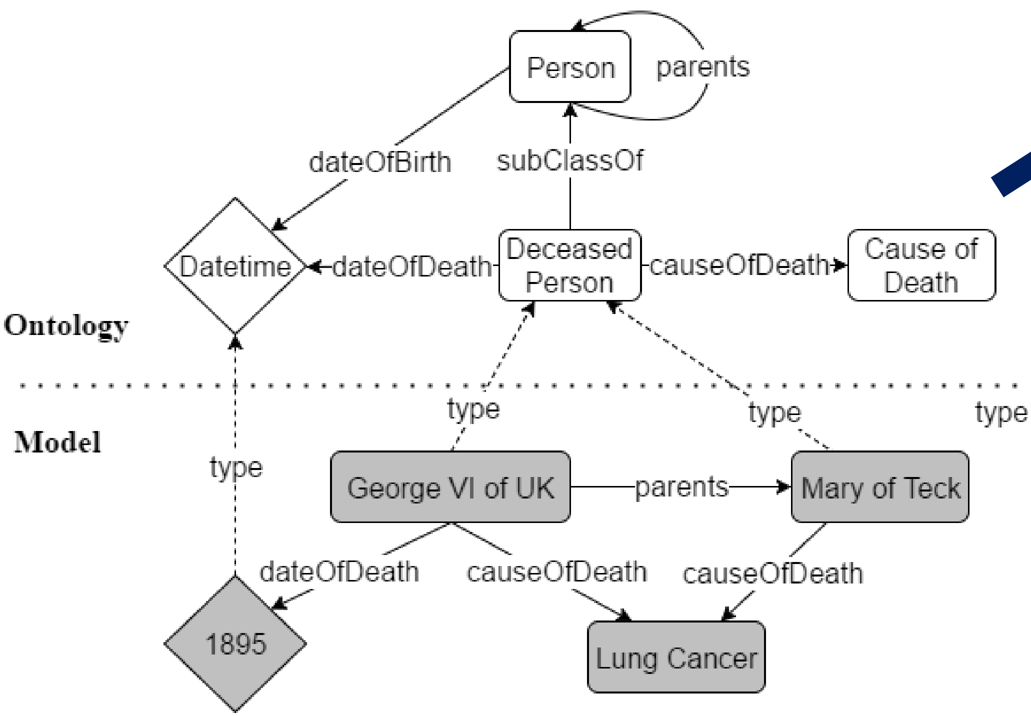


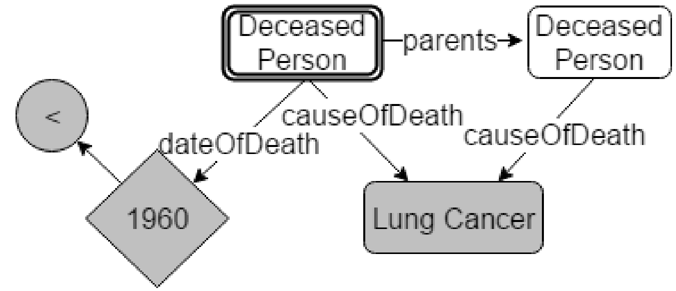
Graph Query

$$
\begin{aligned}
\lambda x.\exists y.\exists z.&\texttt{type}(x, \texttt{DeceasedPerson}) \\
&\wedge \texttt{type}(y, \texttt{DeceasedPerson}) \\
&\wedge \texttt{type}(z, \texttt{Datetime}) \wedge \texttt{parents}(x, y) \\
&\wedge \texttt{causeOfDeath}(x, \texttt{LungCancer}) \\
&\wedge \texttt{causeOfDeath}(y, \texttt{LungCancer}) \\
&\wedge \texttt{dateOfDeath}(x, z) \wedge < (z, \texttt{1960}).
\end{aligned}
$$

λ-calculus

Logical Form: Graph Query





Graph Query

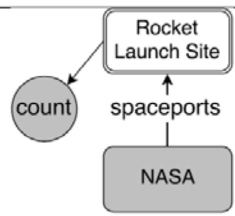
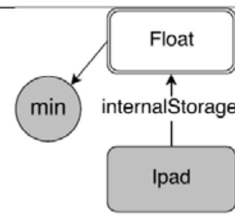
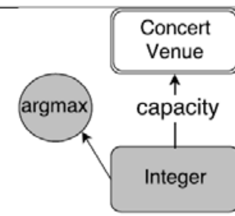
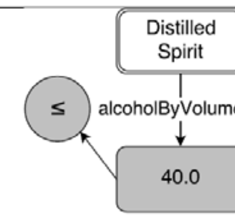
"Find people who died from <lung cancer> before <1960> and whose parent died for the same reason."

Knowledge Base

24K classes, 65K relations, 41M entities, 596M facts

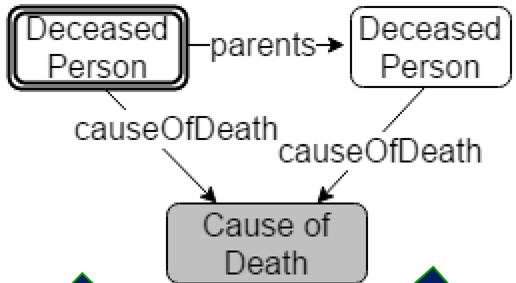
Utterance

Functions (Scott Yih et al.)

Category	Counting	Superlative		Comparative
Functions	`count`	`max` and `min`	`argmax` and `argmin`	$<, >, \leq$, and $\geq$
Domain	Question node	Question node of numeric class	Template/grounded node of numeric class	Template/grounded node of numeric class
Example				
Question	How many launch sites does nasa have?	What's the smallest internal storage of ipad?	Find the largest concert venue.	List distilled spirits with no more than 40.0% abv.

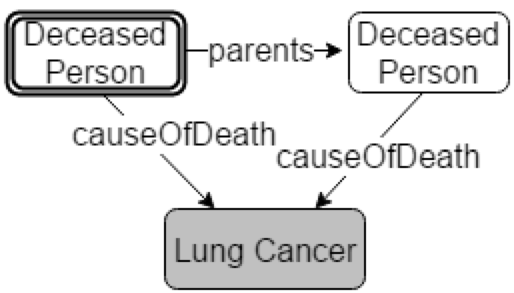
Query template and graph query generation

Query Template

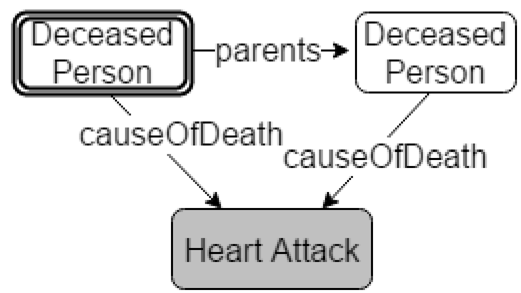


Why query template?

Graph Query 1



Graph Query 2

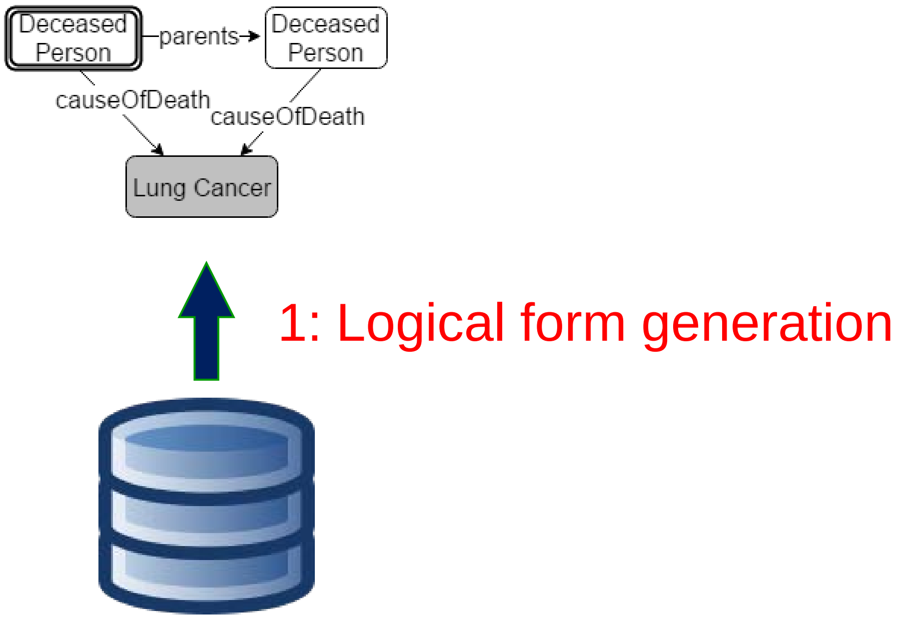


"Find people who died from <lung cancer>, same as their parent did."

"Find people who died from <heart attack>, same as their parent did."

Too many graph queries

- Freebase: 24K classes, 65K relations, 41M entities, 596M facts
- Easily generate millions of graph queries
- Which graph queries correspond to *relevant* questions?

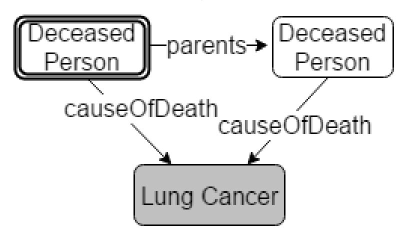


Commonness checking

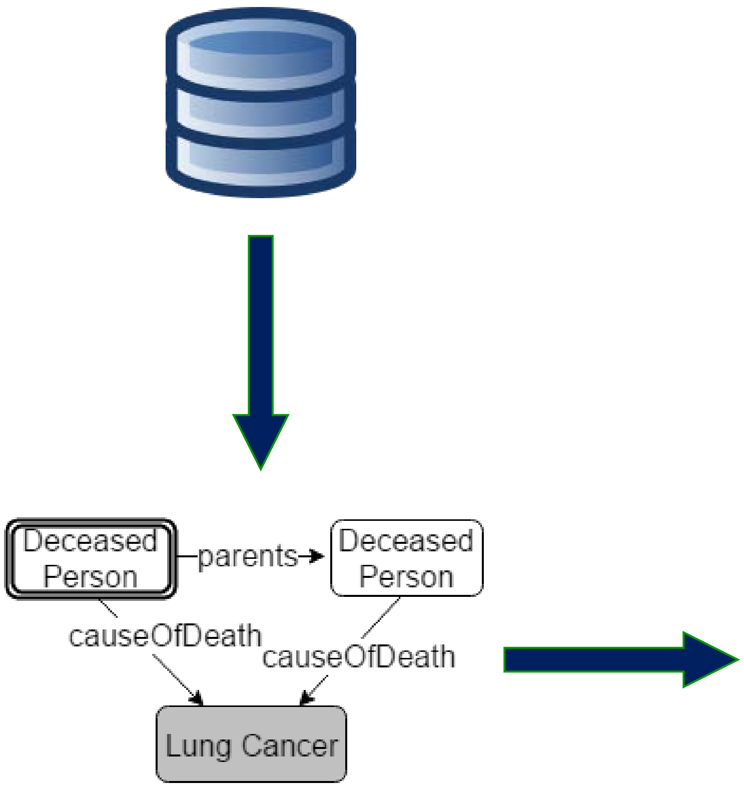
ClueWeb+FACC1:
1B documents, 10B entity mentions

Entity Probabilities	
USA	0.025
…	…
James_Southam	10^{-8}

Relation Probabilities	
Location.contains	0.08
…	…
Chromosome.identifier	0.0

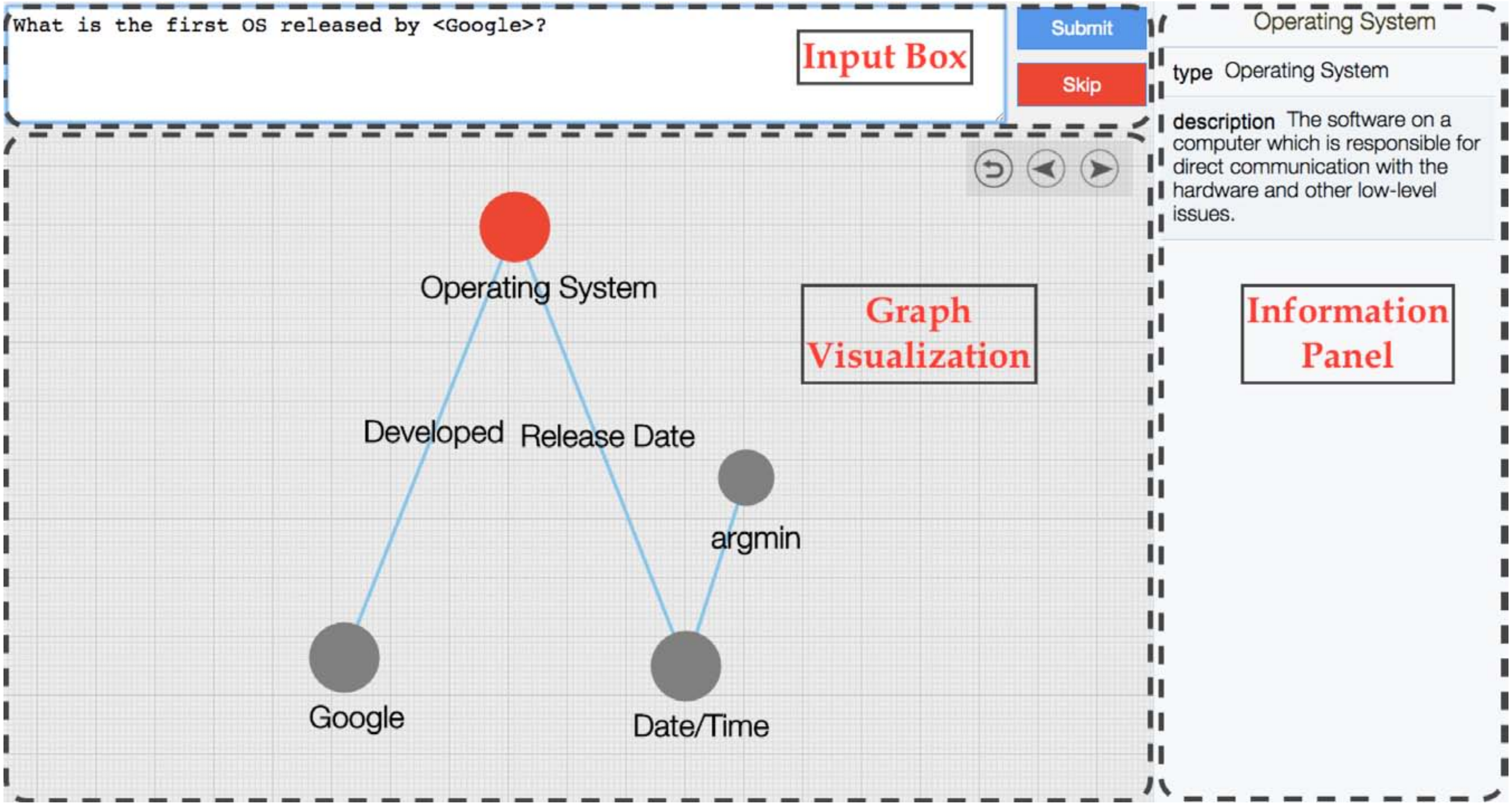
$p($  $)$

Canonical utterance generation

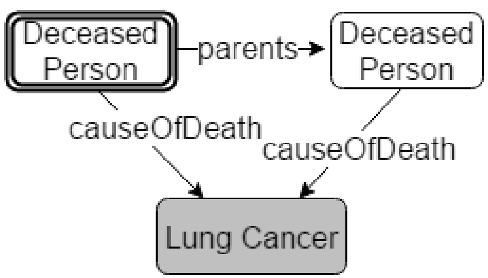


"Find people who died from <lung cancer>, same as their parent did."

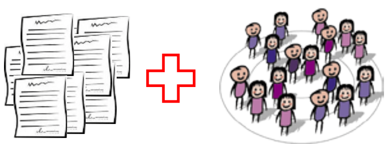
UI for canonical command generation



Paraphrasing



"Find people who died from <lung cancer>, same as their parent did."



"Find people who died from <lung tumor>, same as their parent did."

" Who died from <lung cancer>, the same cause of death of their parent?"

Dataset

- GraphQuestions
  - 5166 questions, 148 domains, 506 classes, 596 relations

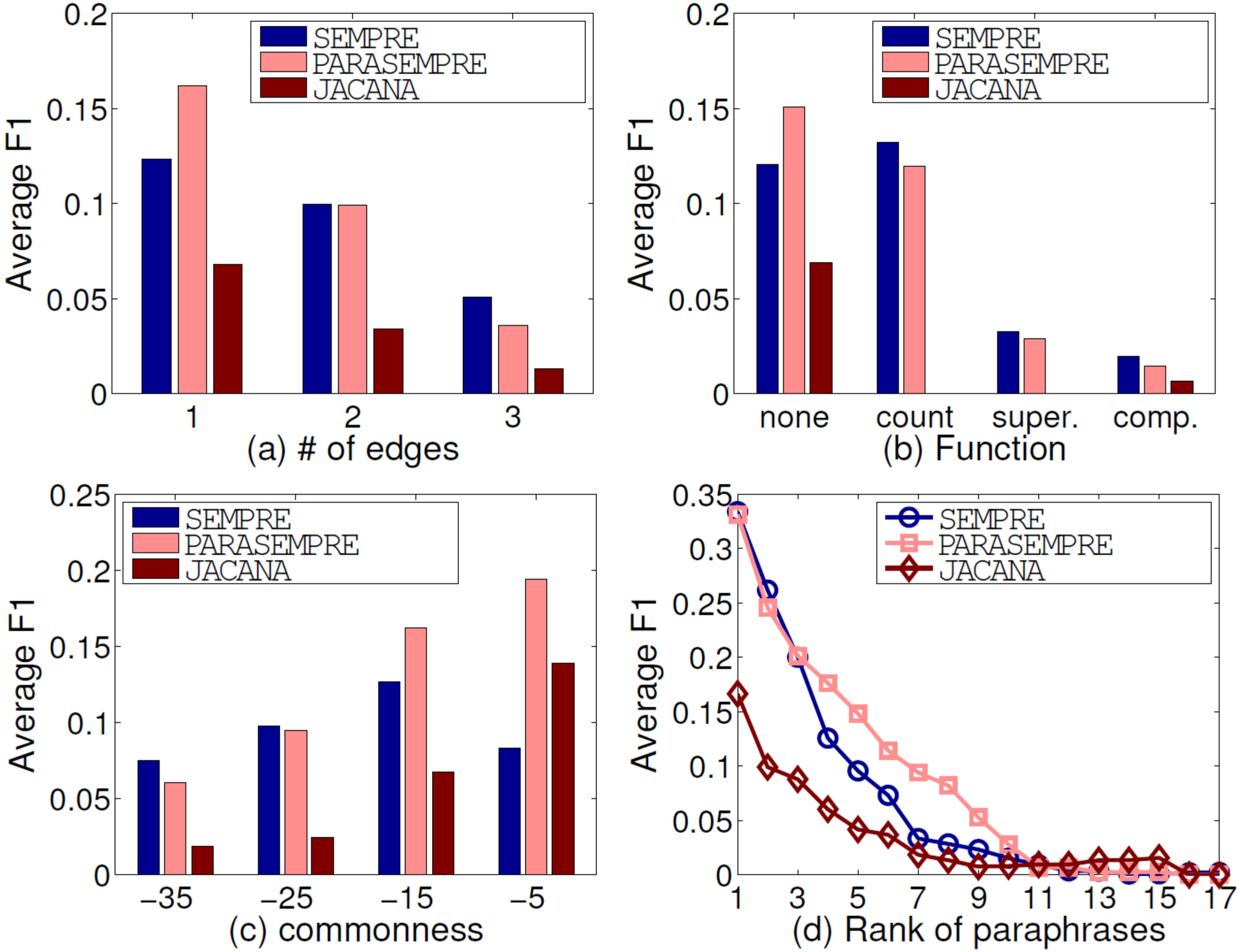
Question	Domain	Answer	# of edges	Function	$\log_{10}(p(q))$	$\|A\|$
Find terrorist organizations involved in **September 11 attacks**.	Terrorism	alQaeda	1	none	-16.67	1
The **September 11 attacks** were carried out with the involvement of what terrorist organizations?						
Who did **nine eleven**?						
How many children of **Eddard Stark** were born in **Winterfell**?	Fictional Universe	3	2	count	-23.34	1
Winterfell is the home of how many of **Eddard Stark**'s children?						
What's the number of **Ned Stark**'s children whose birthplace is **Winterfell**?						
In which month does the average rainfall of **New York City** exceed **86** mm?	Travel	March, August ...	3	comp.	-37.84	7
Rainfall averages more than **86** mm in **New York City** during which months?						
List the calendar months when **NYC** averages in excess of **86** millimeters of rain?						

Evaluation

- SEMPRE (Berant et al. EMNLP'13) semantic parsing
  - *Bottom-up beam parsing, log-linear function w/ linguistic features*
- PARASEMPRE (Berant and Liang ACL'14) semantic parsing
  - *How well the canonical utter. paraphrases the input utter.?*
- JACANA (Yao and Van Durme ACL'14) information extraction
  - *Binary classifier w/ linguistic features: which neighboring entity of the topic entity is the correct answer?*

System	F1	Time/s
SEMPRE	10.80	56.19
PARASEMPRE	12.79	18.43
JACANA	5.08	2.01

Decomposition by characteristics



Next generation KBQA

- Complex questions
  - Better exploration of the huge candidate space
  - Imitation/reinforcement learning, partial logical form evaluation
- More expressive
  - Support of functions like superlatives and comparatives
  - Open domain is more challenging ("*older*", "*best-selling*", …)
- Better handling of paraphrasing
- Better performance at the tail

Improving Semantic Parsing via Answer Type Inference (EMNLP’16) with Semih Yavuz et al.

Semantic Parsing

What did Joe Biden study in college?

Semantic Parsing

What did Joe Biden study in college?

semantic parsing

MajorFieldOfStudy.(Education.JoeBiden)

Semantic Parsing

What did Joe Biden study in college?

semantic parsing

MajorFieldOfStudy.(Education.JoeBiden)

execute logical form

Political Science

Semantic Parsing

What did Joe Biden study in college?

$\Big\downarrow$ semantic parsing

MajorFieldOfStudy.(Education.JoeBiden)

$\Big\downarrow$ execute logical form

Political Science

Motivation: An interface between natural language and structured knowledge bases (Freebase, DBPedia, Yago, ...)

Answer Type Helps

What are Taylor Swift's albums?

AgendaIL (Berant and Liang, 2015) top-10 logical forms

Relation	Answer Type	Prob	F1
people.person.profession	people.profession	0.12	0
people.person.profession	people.profession	0.12	0
music.artist.album	**music.album**	**0.05**	**0.5**
music.artist.album	music.album	0.05	0.5
music.artist.album	music.album	0.02	0.5
music.artist.album	music.album	0.02	0.5
film.actor.film/film.performance.film	film.film	0.01	0
music.artist.origin	location.location	0.01	0
film.actor.film/film.performance.film	film.film	0.01	0
music.artist.origin	location.location	0.01	0

Answer Type Helps

What college did Magic Johnson play for?

AgendaIL (Berant and Liang, 2015) top-10 logical forms

Relation	Answer Type	Prob	F1
basketball.basketball_player.position_s	basketball.basketball_position	0.39	0
basketball.basketball_player.former_teams	basketball.basketball_team	0.1	0
people.person.education / education.education.institution	**education.university**	**0.09**	**1.0**
people.person.education / education.education.institution	education.educational_instituition	0.07	0.667
government.government_position_held.office_holder	government.government_office_or_title	0.04	0
organization.organization.founders	organization.organization	0.03	0
sports.sports_team.roster/sports.sports_team_roster.player	sports.sports_team	0.03	0
sports.sports_award_winner.awards/sports.sports_award.season	sports.sports_league_season	0.02	0
sports.sports_team.roster/sports.sports_team_roster.player	sports.sports_team	0.02	0
people.person.education / education.education.institution	education.university	0.02	1.0

Observation

music.album

answer type

What are Taylor Swift's albums?

logical form

music.artist.album

Filters top-2 wrong answers!

education.university

answer type

What college did Magic Johnson play for?

logical form

people.person.education / education.education.institution

Filters top-2 wrong answers!

Room for Improvement by Answer Type

Ranking	$F1$	# Improved Qs
AgendaIL	49.7	-
w/ Oracle Types@10	57.3	+234
w/ Oracle Types@20	58.7	+282
w/ Oracle Types@50	60.1	+331
w/ Oracle Types@All	60.5	+345

Table 1: What if the correct answer type is enforced? On WebQuestions, we remove those with incorrect answer types in the top-k logical forms returned by AgendaIL (Berant and Liang, 2015), a leading semantic parsing system, and report the new average F1 score as well as the number of questions with an improved F1 score.

Outline

- Background
- Motivation
- Answer Type Inference
- Future Work

Answer Type Inference

- Setup
- Question Abstraction
- Conversion to Statement Form
- Inferring Answer Types
- Reranking Logical Forms by Answer Type

Pipeline

When did [Shaq] come into the NBA?

Abstraction

When did [drafted athlete] come into the NBA?

Conversion

[drafted athlete] come when into the NBA

Answer Type Inference

SportsLeagueDraft

Answer Type Inference

- Setup
- Question Abstraction
- Conversion to Statement Form
- Inferring Answer Types
- Reranking Logical Forms by Answer Type
- Dataset Creation
- Experiments

Question Abstraction

Intuition: Answer type remains invariant as the topic entity changes within the same category (e.g., drafted athlete).

When did [Shaq] come into the NBA?

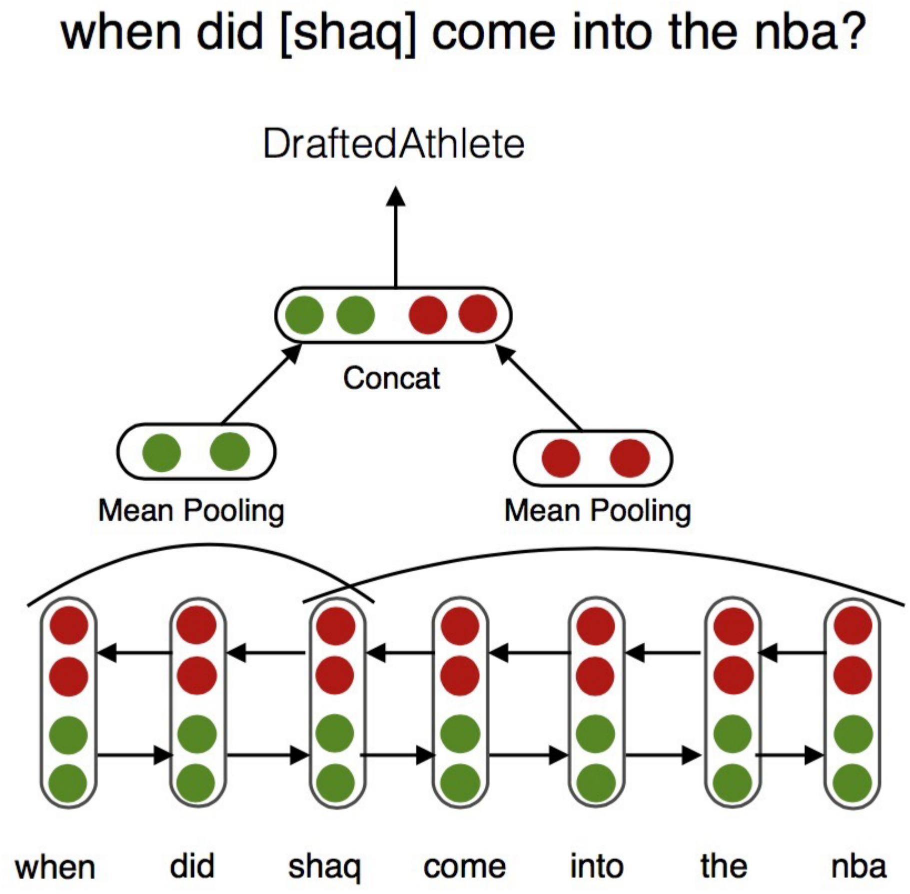
Abstraction

When did [drafted athlete] come into the NBA?

Objective:

1. Find the right KB type that represent the topic entity in the question context.
2. Form the abstract question by replacing the topic entity with this representative type.

Bidirectional LSTM Model



Output: Network outputs a **probability distribution** over KB types denoting the likelihood for being topic entity (e.g. shaq) type in the question context

Answer Type Inference

- Setup
- Question Abstraction
- Conversion to Statement Form
- Inferring Answer Types
- Reranking Logical Forms by Answer Type
- Dataset Creation
- Experiments

Conversion to Statement Form

What boarding school did Mark Zuckerberg go to?

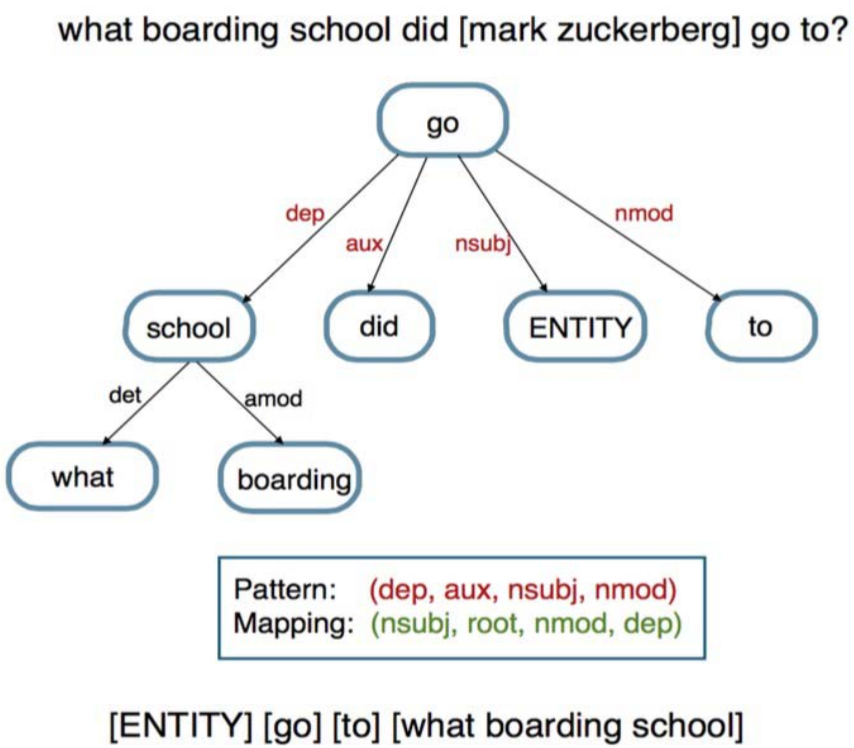
Conversion

Mark Zuckerberg go to what boarding school?

Objective:

- Canonicalize question form into statement (subject-relation-object) form by reordering the words of question

Pattern-based Approach



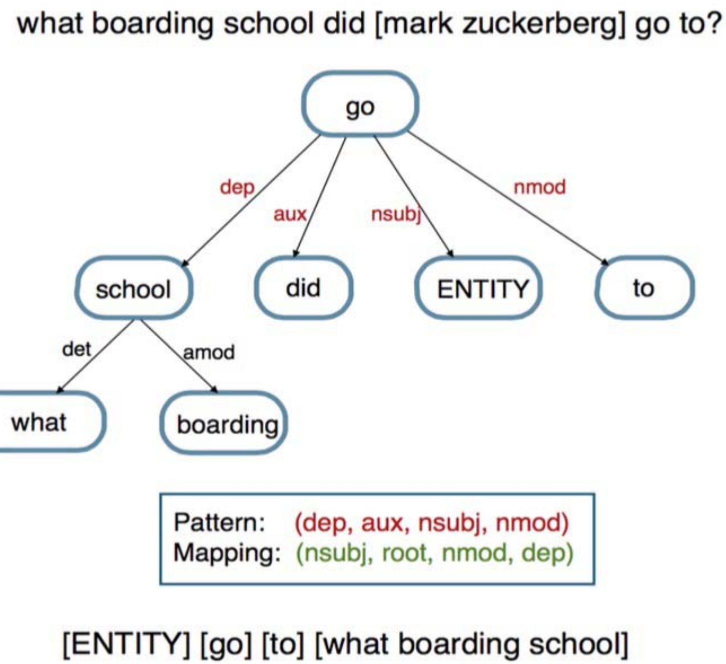
- Retrieve the named entity (NER) tags of the question tokens.
- Replace tokens corresponding to the **named entity** with a single special token ENTITY.
- Obtain the dependency parse tree of the simplified question.
- Represent each question by a **pattern**: the root's dependency relations to its sub-trees in the original order.

Conversion Mapping:

1. Cluster question representation patterns
2. Manually map frequent patterns* to their corresponding conversions (Pattern vs. Mapping)

*Frequent patterns are the ones with at least 5 occurrences in training data

Reordering Words based on Mapping



Reordering Words:

- Conversion mapping determines **the order** in which the **sub-trees** of the root is **recomposed**
- Original order of sub-trees and question tokens:
  - (dep, aux, nsubj, nmod)
  - [what boarding school] [did] [ENTITY] [to]
- Reordered sub-trees and question tokens:
  - (nsubj, root, nmod, dep)
  - [ENTITY] [go] [to] [what boarding school]

Answer Type Inference

- Setup
- Question Abstraction
- Conversion to Statement Form
- Inferring Answer Types
- Reranking Logical Forms by Answer Type
- Dataset Creation
- Experiments

Inferring Answer Type

Intuition: Question word (e.g., “when”) along with its directed left and right contexts provides the clues for answer type

[drafted athlete] come when into the NBA?

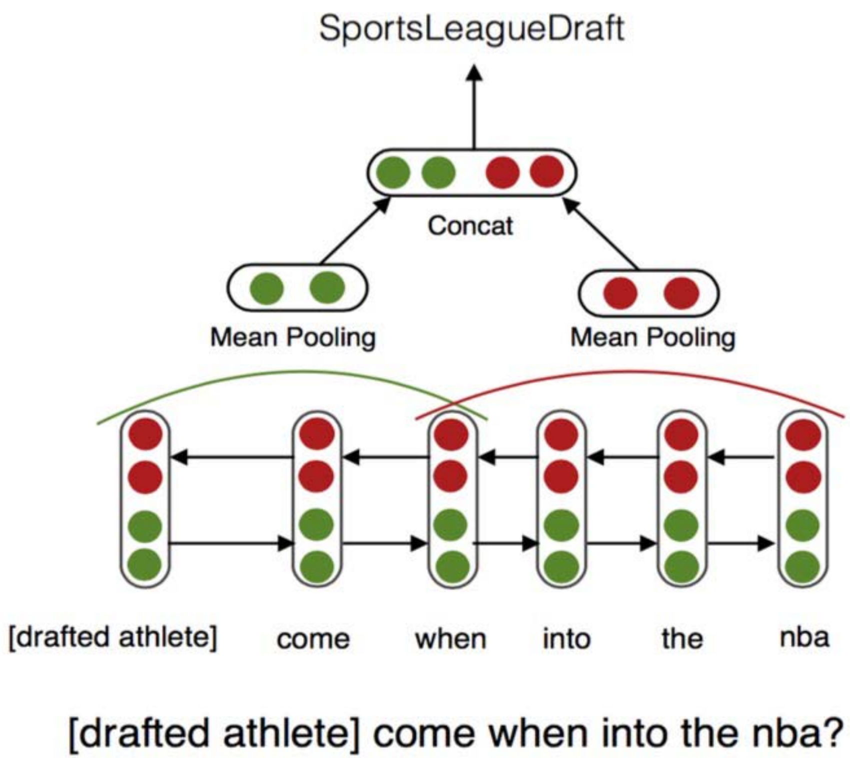
Answer Type Inference

SportsLeagueDraft

Objective:

- For a given abstract question, assign a score/probability to each target KB type denoting its likelihood of being the answer type. (for reranking candidate answers)

Bidirectional LSTM Model for Answer Type Inference



- **Output:** Network outputs a **score / probability distribution over** KB types denoting the likelihood of being the answer type
- **Output Node:** Question word (e.q., “when”)
- **Left Context:** {[drafted athlete], come, when}
- **Right Context:** {when, into, the, nba}

Answer Type Inference

- Setup
- Question Abstraction
- Conversion to Statement Form
- Inferring Answer Types
- Reranking Logical Forms by Answer Type
- Dataset Creation
- Experiments

Main Result

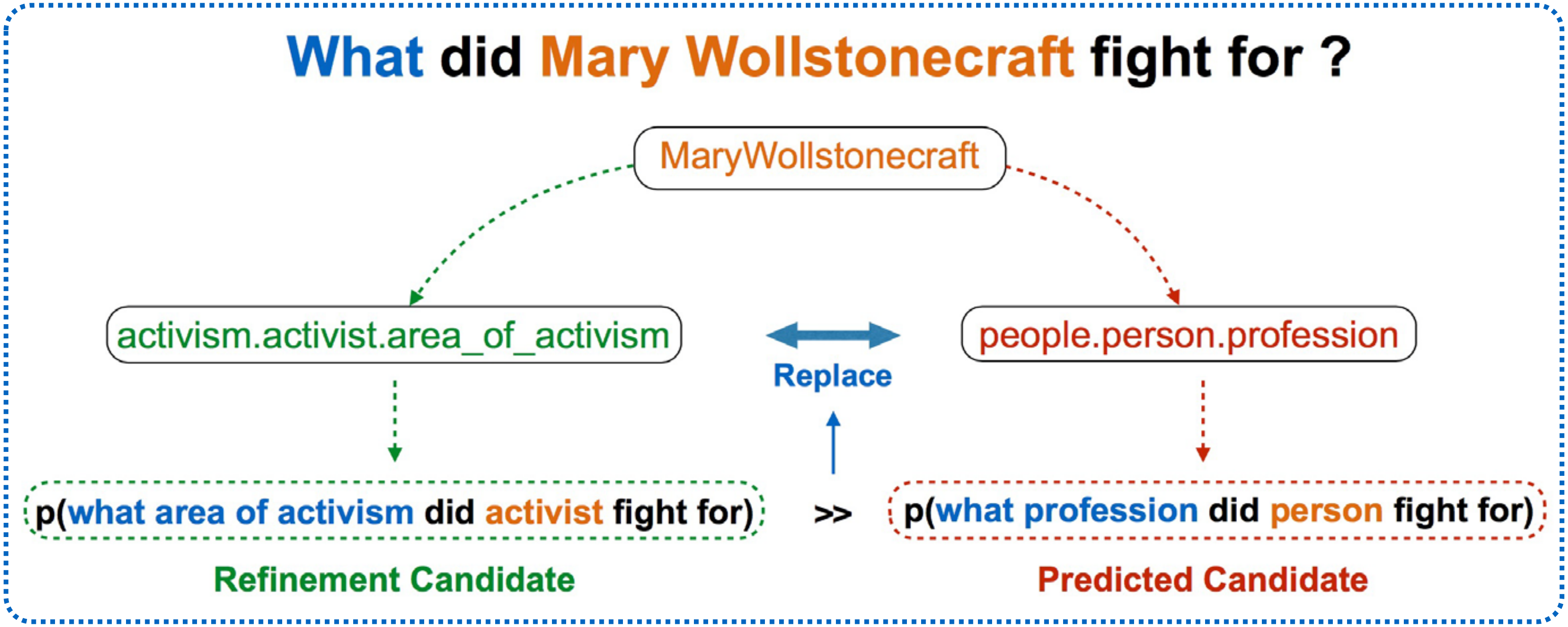
Model	F1
(Berant and Liang, 2015)	49.7
(Yih et al., 2015)	52.5
(Xu et al., 2016)	**53.3**
(Yih et al., 2015) (w/ Freebase API)	48.4
(Yih et al., 2015) (w/o ClueWeb)	50.9
(Xu et al., 2016) (w/o Wikipedia)	47.1
Our Approach	**52.6**

Table 3: Comparison of our reranking-by-type system with several existing works on WebQuestions.

Recovering Question Answering Errors via Query Revision (EMNLP'17), with Semih Yavuz et al.

Motivation & Goal

Cross-check the answer by inserting it back in the question



STAGG (Base QA)	52.5	-
Upper Bound	58.9	+639

Question Revisions

Freebase Relation	Subject Type	Object Type	Relation Text
activism.activist.area_of_activism	activist	activism issue	area of activism

Q: What did Mary Wollstonecraft fight for?

EC: what did activist fight for?

EAC: what activism issue did activist fight for?

ERC: what area of activism did activist fight for?

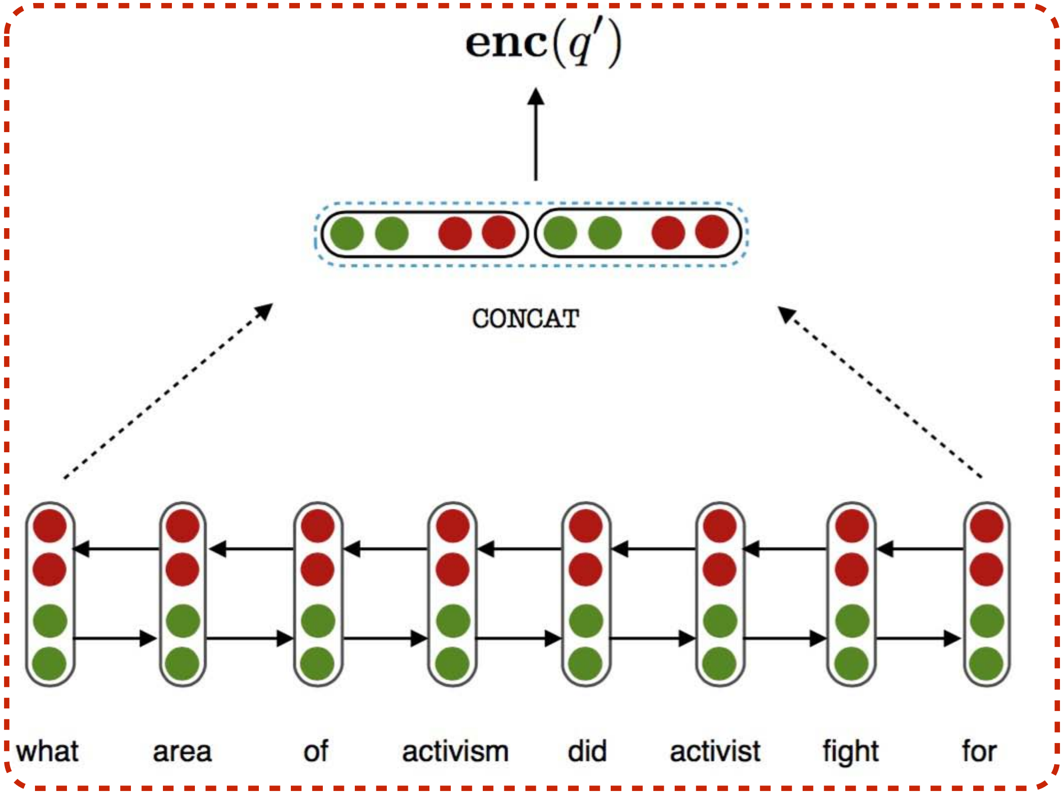
Abstraction ↑
1. EC: Entity Centric
2. EAC: Entity-Answer Centric
3. ERC: Entity-Relation Centric

↓ Expressivity

Encoding and Scoring Revised Questions

$q' = \text{What area of activism did activist fight for ?}$

Revised Question



Bidirectional LSTM Encoder

$$s(q') = \mathbf{w}^T \mathbf{enc}(q')$$

Revised Question Scoring

Margin-based Training Objective

Score(Revised Question by Correct Relation) >> Score(Revised Question by Wrong Relation)

What did Mary Wollstonecraft fight for?

r_{pos} = **activism.activist.area_of_activism**

r_{neg} = **people.person.profession**

$\mathbf{q}'_{pos}$ = **what area of activism did activist fight for**

$\mathbf{q}'_{neg}$ = **what profession did person fight for**

$$s(\mathbf{q}'_{pos}) >> s(\mathbf{q}'_{neg})$$

Minimize hinge-loss

$$\sum_{q \in Q} \sum_{(r_{pos}, r_{neg})} max\left(0, \delta(r_{pos}, r_{neg}) - [s(q'_{pos}) - s(q'_{neg})]\right)$$

Experiments

Quantitative

Comparison with Most Recent Work

(Dong et al., 2015)	40.8
(Yao, 2015)	44.3
(Berant and Liang, 2015)	49.7
(Yih et al., 2015) - **STAGG**	52.5
(Reddy et al., 2016)	50.3
(Xu et al., 2016a)	53.3
(Xu et al., 2016b)	**53.8**
Our Approach on STAGG	**53.9**

Variants of Our Model

WebQ	EC	52.9
	EAC	53.5
	ERC	53.2
	EAC + ERC	53.3
+SimpleQ	EC	52.8
	EAC	53.6
	ERC	53.8
	EAC + ERC	**53.9**

Qualitative

Questions and Refinement Candidates	KB Relations	IsR
1. where does the **zambezi river** start ?		
Prediction (ERC) : where mouth does the river start	river.mouth	
Refinement (ERC): where origin does the river start	river.origin	✓
2. what did **mary wollstonecraft** fight for ?		
Prediction (EAC) : what profession did person fight for	person.profession	
Refinement (EAC): what activism issue did activist fight for	activist.area_of_activism	✓
3. where did the **iroquois indians** come from?		
Prediction (EAC) : where ethnicity did the ethnicity come from	ethnicity.included_in_group	
Refinement (EAC): where location did the ethnicity come from	ethnicity.geographic_distribution	✗
Prefiction (ERC) : where included in group(s) did the ethnicity come from	ethnicity.included_in_group	
Refinement (ERC): where geographic distribution did the ethnicity come from	ethnicity.geographic_distribution	✓
EAC + ERC	ethnicity.geographic_distribution	✓

IsR denotes whether the model decides to refine: ✗: False, ✓: True

Future Work

Continue working on

- Knowledge Graph Based Question Answering
- Text Based Question Answering
- FAQ/Quora Style Question Answering

- Better models and commercial applications

Acknowledgements (Our Students)



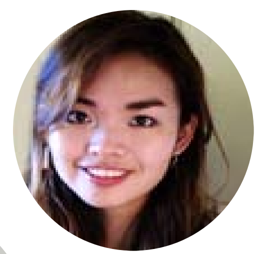
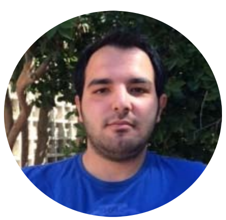


NETWORK SCIENCE CTA





NSF


Thank You