Scalable Construction and Querying of Massive Knowledge Bases

Part II: Schema-agnostic Knowledge Base Querying

Yu Su
Department of Computer Science
University of California, Santa Barbara

Growing Gap between Human and Data

What disease does the patient have?

- (EMR) Similar patients?
- (Literature) New findings?
- (Gene sequence) Suspicious mutations?
- … …

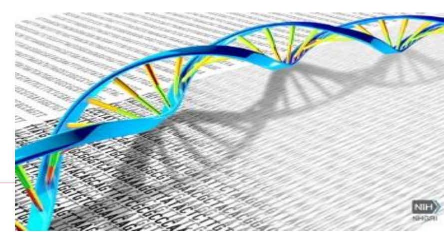
Ad-hoc information needs for on-demand decision making

Massive, heterogeneous data

86.9% adoption (NEHRS 2015)

27M+ papers, >1M new/year (PubMed)

$1000 gene sequencing



24x7 monitoring

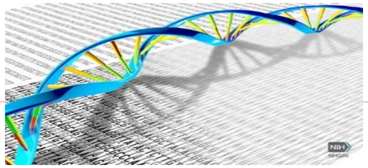
How can AI Bridge the Gap?

Insights
Discoveries
Solutions

Bottleneck #2: Access

Bottleneck #3: Reasoning

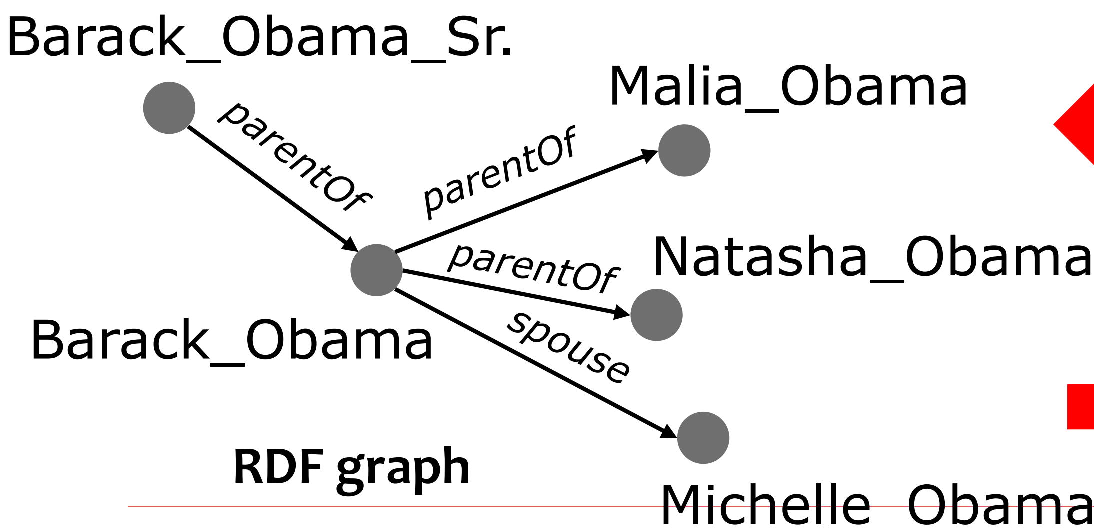
Bottleneck #1: Knowledge



Structured Query: RDF + SPARQL

Triples in an RDF

Subject	Predicate	Object
Barack_Obama	parentOf	Malia_Obama
Barack_Obama	parentOf	Natasha_Obama
Barack_Obama	spouse	Michelle_Obama
Barack_Obama_Sr.	parentOf	Barack_Obama



RDF graph

SPARQL query

```
SELECT ?x WHERE
{
Barack_Obama_Sr. parentOf ?y.
?y parentOf ?x.
}
```

Answer

<Malia_Obama>
<Natasha_Obama>

Why Structured Query Falls Short?

Knowledge Base	# Entities	# Triples	# Classes	# Relations
Freebase	45M	3B	**53K**	**35K**
DBpedia	6.6M	13B	760	2.8K
Google Knowledge Graph*	570M	18B	1.5K	**35K**
YAGO	10M	120M	**350K**	100
Knowledge Vault	45M	1.6B	1.1K	4.5K

* as of 2014

- It's more than large: High heterogeneity of KBs
- *If it's hard to write SQL on simple relational tables, it's only harder to write SPARQL on large knowledge bases*
  - Even harder on automatically constructed KBs with a loosely-defined schema

Not Everyone Can Program…

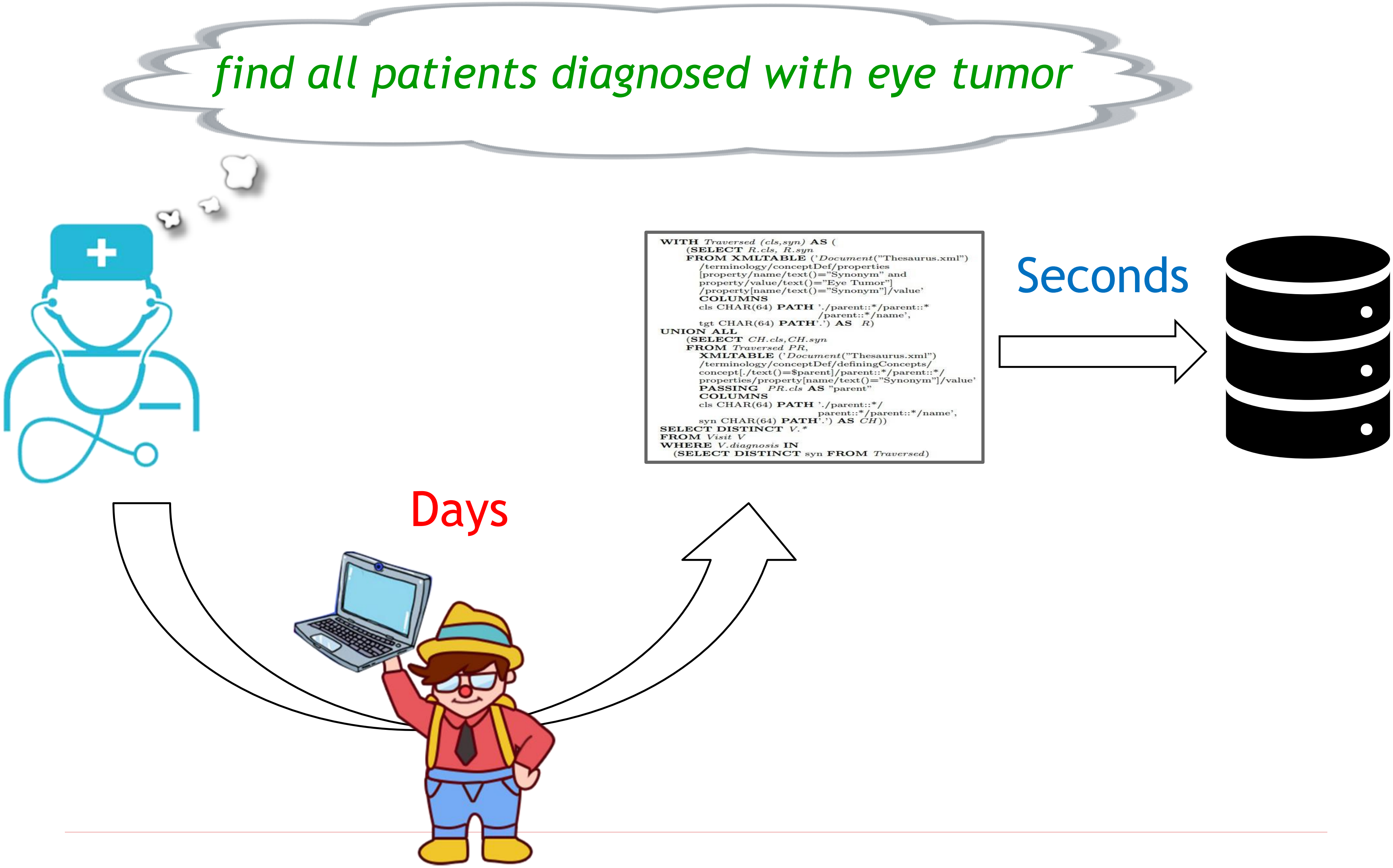
"find all patients diagnosed with eye tumor"

```
WITH Traversed (cls,syn) AS (
    (SELECT R.cls, R.syn
    FROM XMLTABLE ('Document("Thesaurus.xml")
        /terminology/conceptDef/properties
        [property/name/text()="Synonym" and
        property/value/text()="Eye Tumor"]
        /property[name/text()="Synonym"]/value'
        COLUMNS
        cls CHAR(64) PATH './parent::*/parent::*
                         /parent::*/name',
        tgt CHAR(64) PATH'.') AS  R)
UNION ALL
    (SELECT CH.cls,CH.syn
    FROM Traversed PR,
        XMLTABLE ('Document("Thesaurus.xml")
        /terminology/conceptDef/definingConcepts/
        concept[./text()=$parent]/parent::*/parent::*/
        properties/property[name/text()="Synonym"]/value'
        PASSING  PR.cls AS "parent"
        COLUMNS
        cls CHAR(64) PATH './parent::*/
                         parent::*/parent::*/name',
        syn CHAR(64) PATH'.') AS CH))
SELECT DISTINCT V.*
FROM Visit V
WHERE V.diagnosis IN
  (SELECT DISTINCT syn FROM Traversed)
```

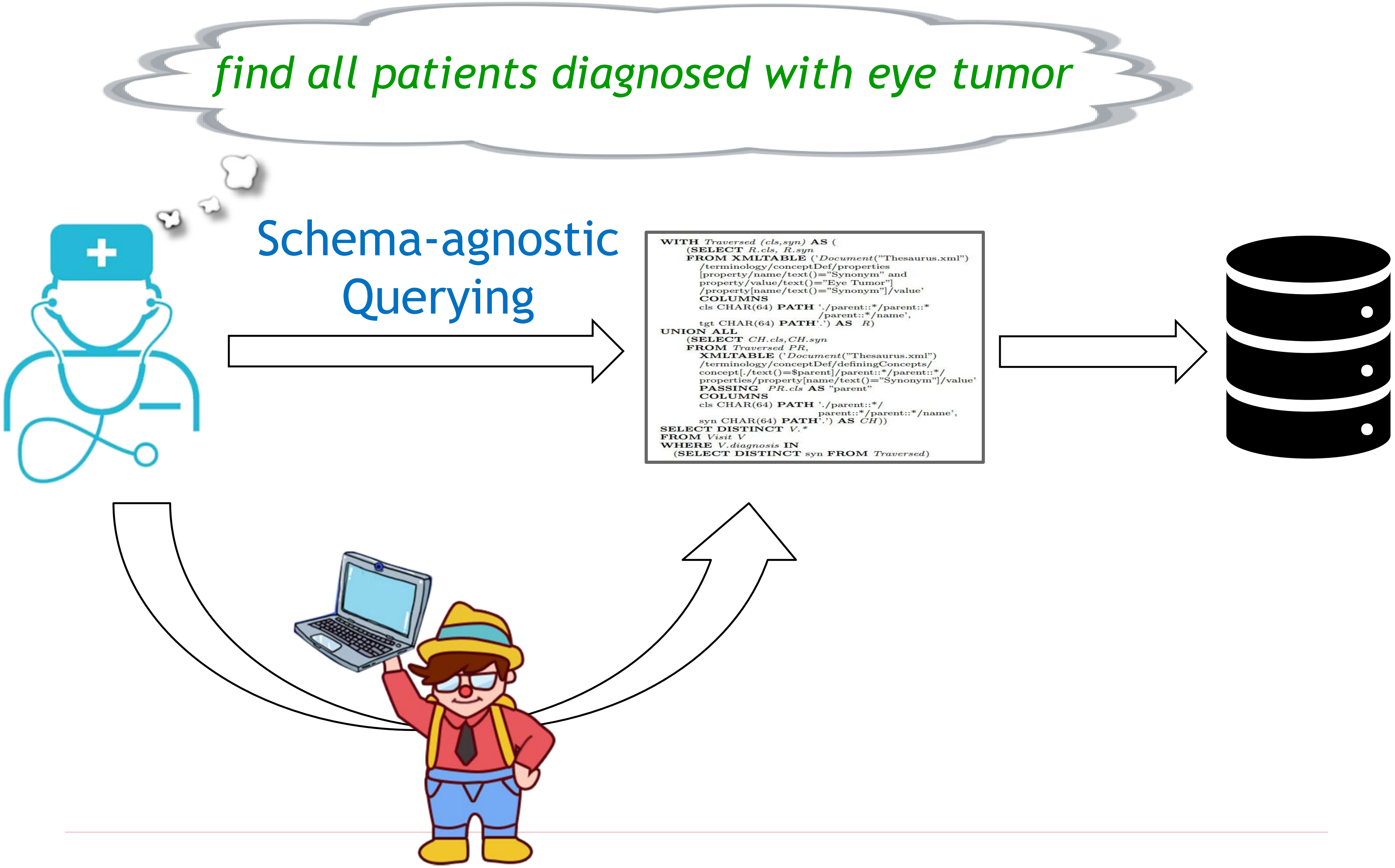
NCI thesaurus

"Semantic queries by example",
Lim et al., EDBT (2014)

In Pursue of Efficiency



In Pursue of Efficiency



Outline

- Schema-agnostic Graph Query
- Natural Language Interface (a.k.a., Semantic Parsing)
  - A little history
  - Cold-start with crowdsourcing
  - Cold-start with neural transfer learning

Schemaless and Structureless Graph Querying

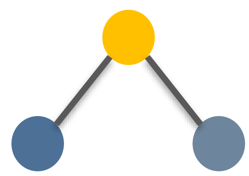
Shengqi Yang, Yinghui Wu, Huan Sun and Xifeng Yan
UC Santa Barbara

VLDB’14

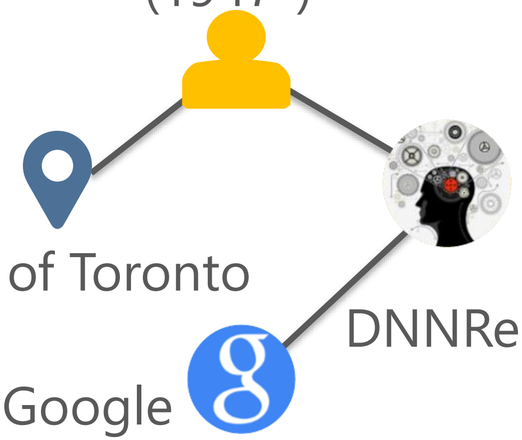
Graph Query

"Find a professor, ~70 yrs., who works in Toronto and joined Google recently."

Search intent



Graph query

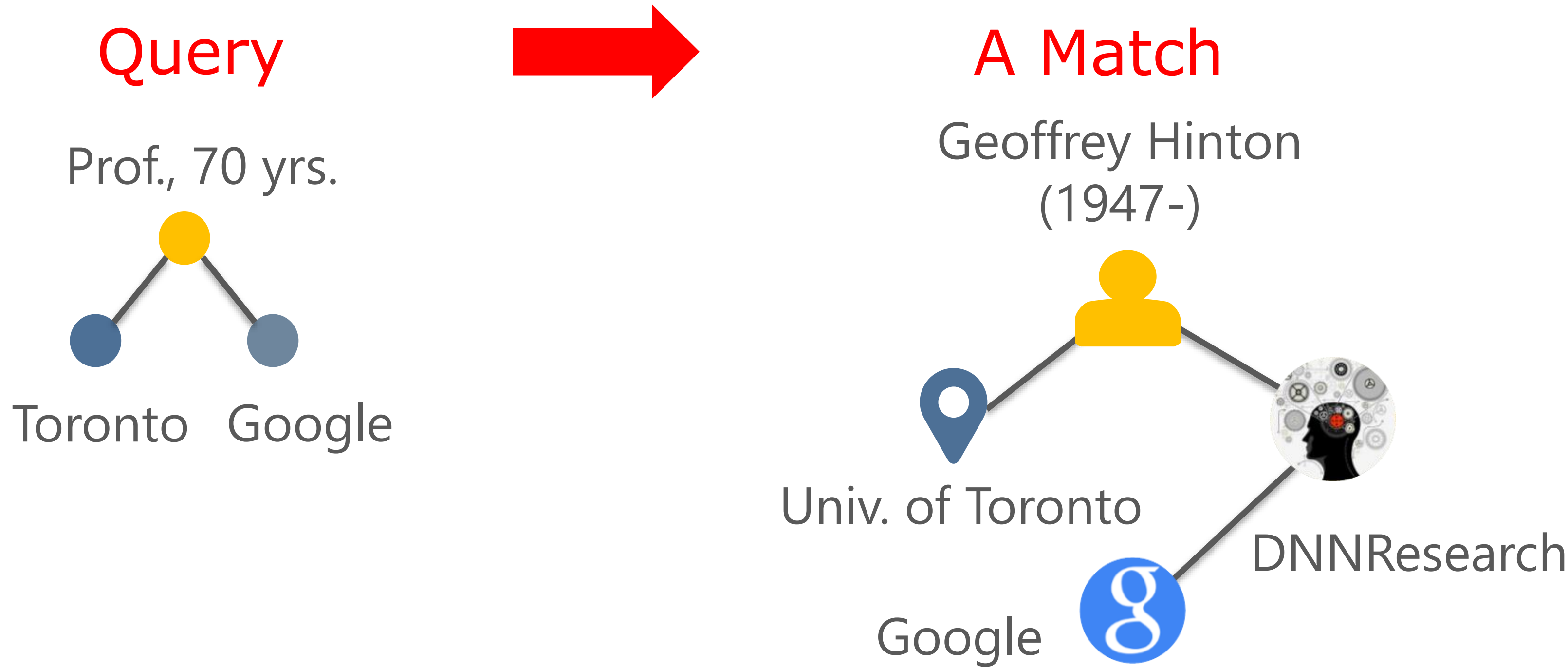


A match (result)

Query-KB Mismatch

Knowledge Base	Query
"University of Washington"	"UW"
"neoplasm"	"tumor"
"Doctor"	"Dr."
"Barack Obama"	"Obama"
"Jeffrey Jacob Abrams"	"J. J. Abrams"
"teacher"	"educator"
"1980"	"~30"
"3 mi"	"4.8 km"
"Hinton" - "DNNresearch" - "Google"	"Hinton" - "Google"
…	…

Schemaless Graph Querying (SLQ)

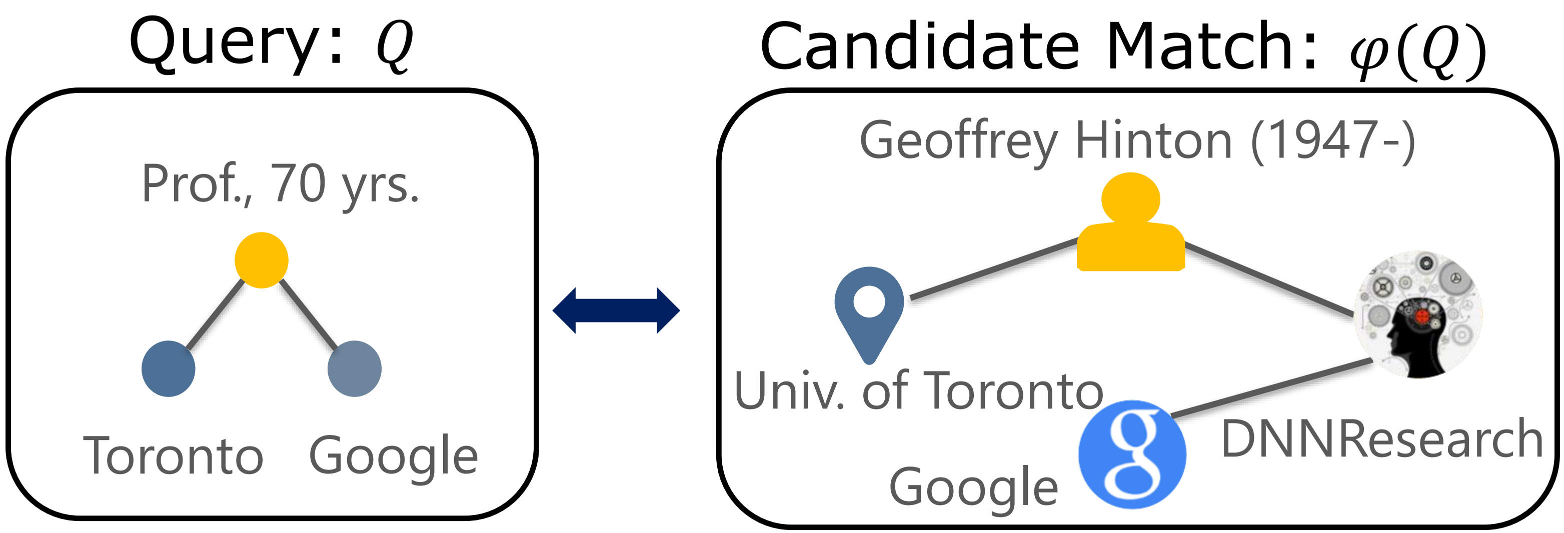


✓**Acronym transformation: '**UT**' → '**University of Toronto**'**
✓**Abbreviation transformation: '**Prof.**' → '**Professor**'**
✓**Numeric transformation: '**~70**' → '**1947**'**
✓**Structural transformation:** an edge → a path

Transformations for KB-Query Mismatch

Transformation	Category	Example
First/Last token	String	"Barack Obama" > "Obama"
Abbreviation	String	"Jeffrey Jacob Abrams" > "J. J. Abrams"
Prefix	String	"Doctor" > "Dr"
Acronym	String	"International Business Machines" > "IBM"
Synonym	Semantic	"tumor" > "neoplasm"
Ontology	Semantic	"teacher" > "educator"
Range	Numeric	"~30" > "1980"
Unit Conversion	Numeric	"3 mi" > "4.8 km"
Distance	Topology	"Pine" - "M:I" > "Pine" - "J.J. Abrams" - "M:I"
...	...	...

Candidate Match Ranking



- Features
  - Node matching features: $F_V(v,\varphi(v)) = \sum_i \alpha_i f_i(v,\varphi(v))$
  - Edge matching features: $F_E(e,\varphi(e)) = \sum_j \beta_j g_j(e,\varphi(e))$
- Overall Matching Score

Conditional Random Field

$$P(\varphi(Q) \mid Q) \propto \exp\left(\sum_{v \in V_Q} F_V(v,\varphi(v)) + \sum_{e \in E_Q} F_E(e,\varphi(e))\right)$$

Exploiting Relevance Feedback in Knowledge Graph Search

Yu Su, Shengqi Yang, Huan Sun, Mudhakar Srivatsa, Sue Kase, Michelle Vanni, and Xifeng Yan

UC Santa Barbara, IBM Research, Army Research Lab

KDD'15

Query-specific Ranking via Relevance Feedback

- Generic ranking: sub-optimal for specific queries
  - By “Washington”, user A means *Washington D.C.*, while user B might mean *University of Washington*
- Query-specific ranking: tailored for each query
  - But need additional query-specific information for further disambiguation

Relevance Feedback:
1. Given user query, generate initial ranking results
2.1. Explicit feedback: Users indicate the **(ir)relevance** of a handful of answers
2.2. Pseudo feedback: Bilindly assume top-10 initial results are correct
3. Improve ranking with feedback information

Problem Definition

$Q:$ A graph query

G: A knowledge graph

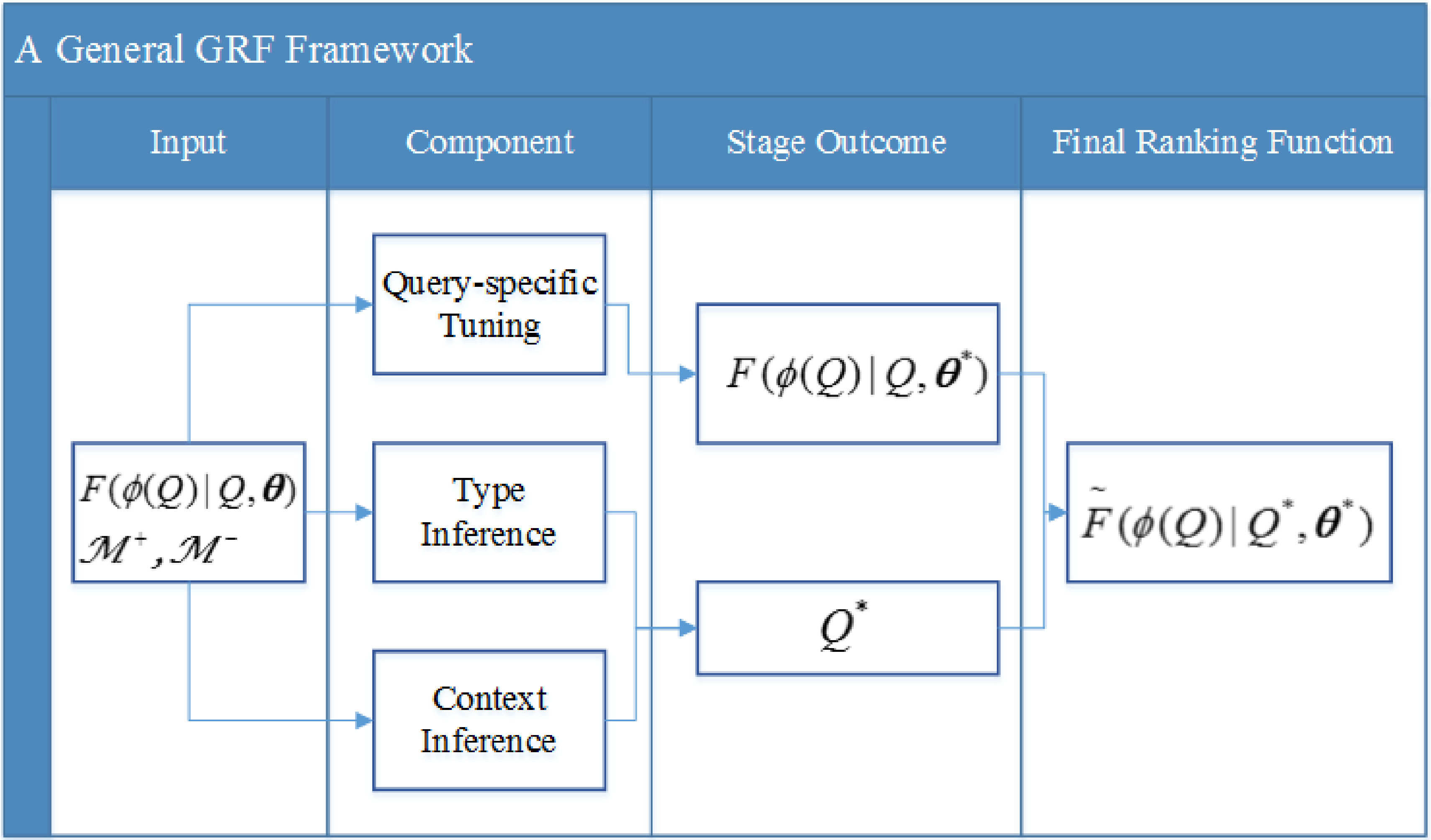
$\phi(Q):$ A candidate match to Q

$F(\phi(Q) \mid Q, \boldsymbol{\theta}):$ A generic ranking function

$\mathcal{M}^+:$ A set of positive/relevant matches of Q

$\mathcal{M}^-:$ A set of negative/non-relevant matches of Q

Graph Relevance Feedback (GRF):
Generate a query-specific ranking function $\tilde{F}$ for Q based on $\mathcal{M}^+$ and $\mathcal{M}^-$


A General GRF Framework
Input
Component
Stage Outcome
Final Ranking Function
$F(\phi(Q)|Q,\boldsymbol{\theta})$
$\mathcal{M}^+, \mathcal{M}^-$
Query-specific Tuning
Type Inference
Context Inference
$F(\phi(Q)|Q,\boldsymbol{\theta}^*)$
Q^*
$\tilde{F}(\phi(Q)|Q^*,\boldsymbol{\theta}^*)$

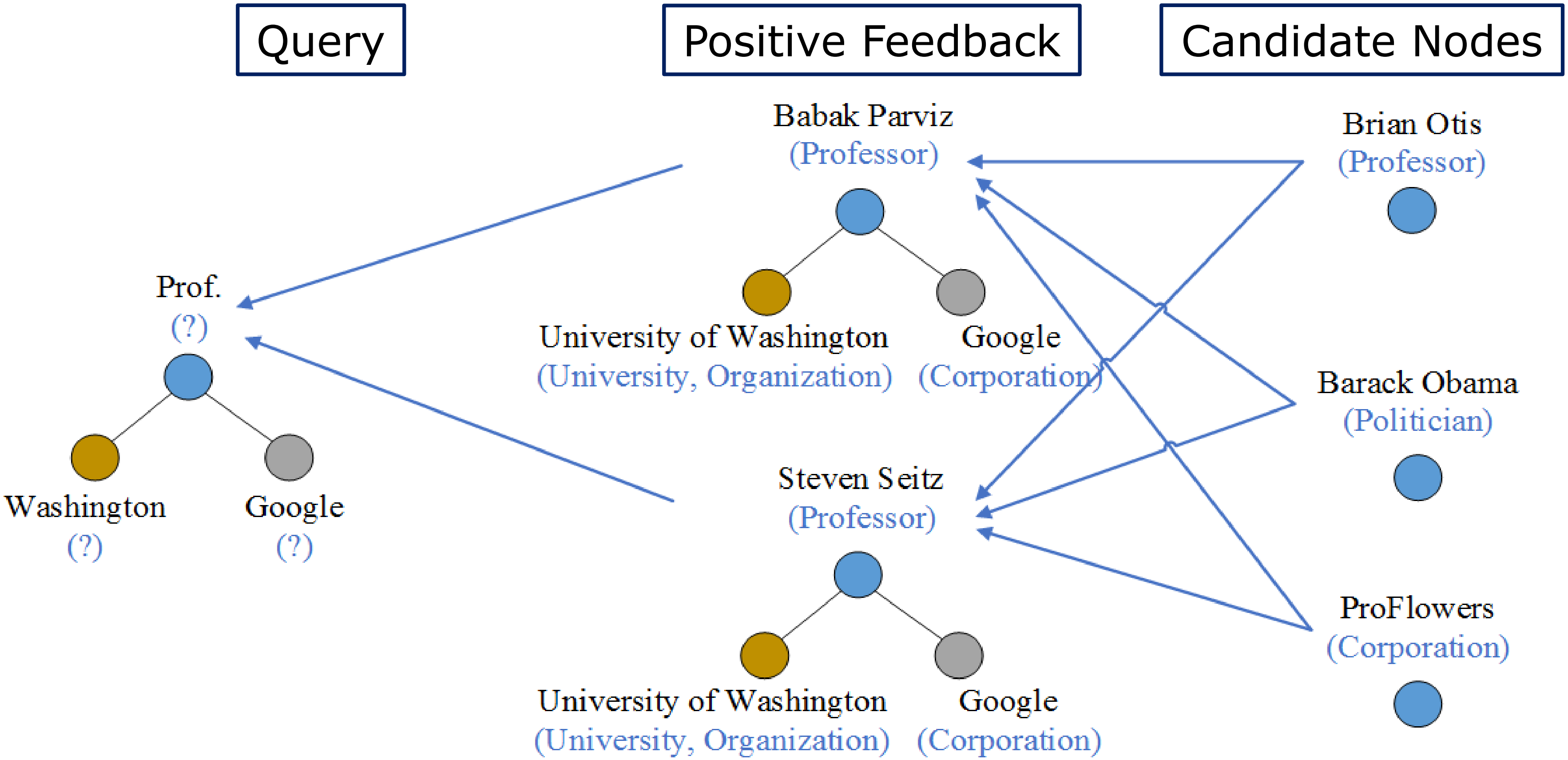
Query-specific Tuning

- θ represents (query-independent) feature weights. However, each query carries its own view of feature importance
- Find query-specific θ^* that better aligned with the query using user feedback

$$g(\boldsymbol{\theta}^*) = (1-\lambda)\underbrace{\left(\frac{\sum_{\phi(Q)\in\mathcal{M}^+} F(\phi(Q)\,|\,Q,\boldsymbol{\theta}^*)}{|\mathcal{M}^+|} - \frac{\sum_{\phi(Q)\in\mathcal{M}^-} F(\phi(Q)\,|\,Q,\boldsymbol{\theta}^*)}{|\mathcal{M}^-|}\right)}_{\text{User Feedback}} + \underbrace{\lambda R(\boldsymbol{\theta},\boldsymbol{\theta}^*)}_{\text{Regularization}}$$

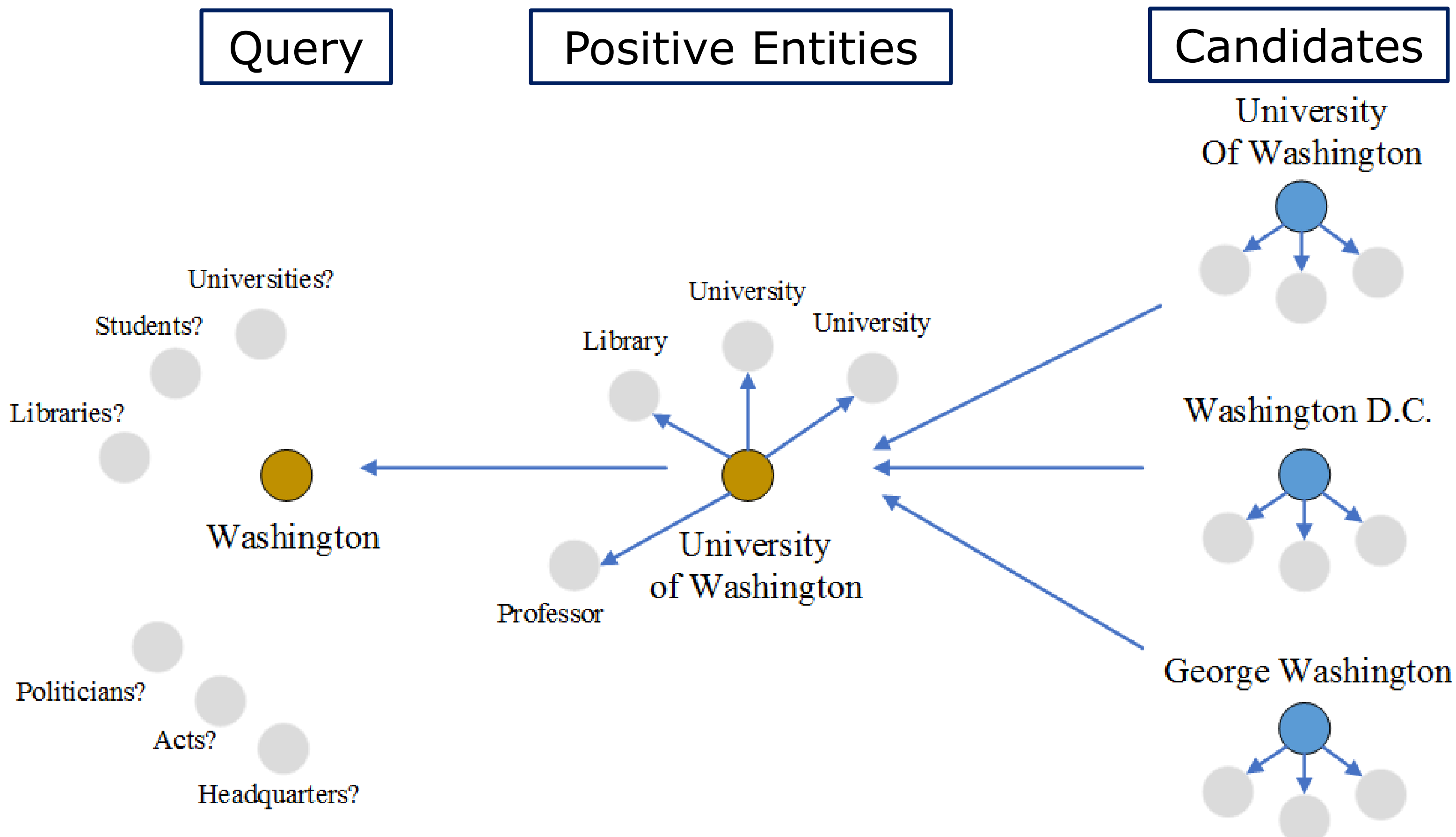
Type Inference

- Infer the implicit type of each query node
- The types of the positive entities constitute a composite type for each query node



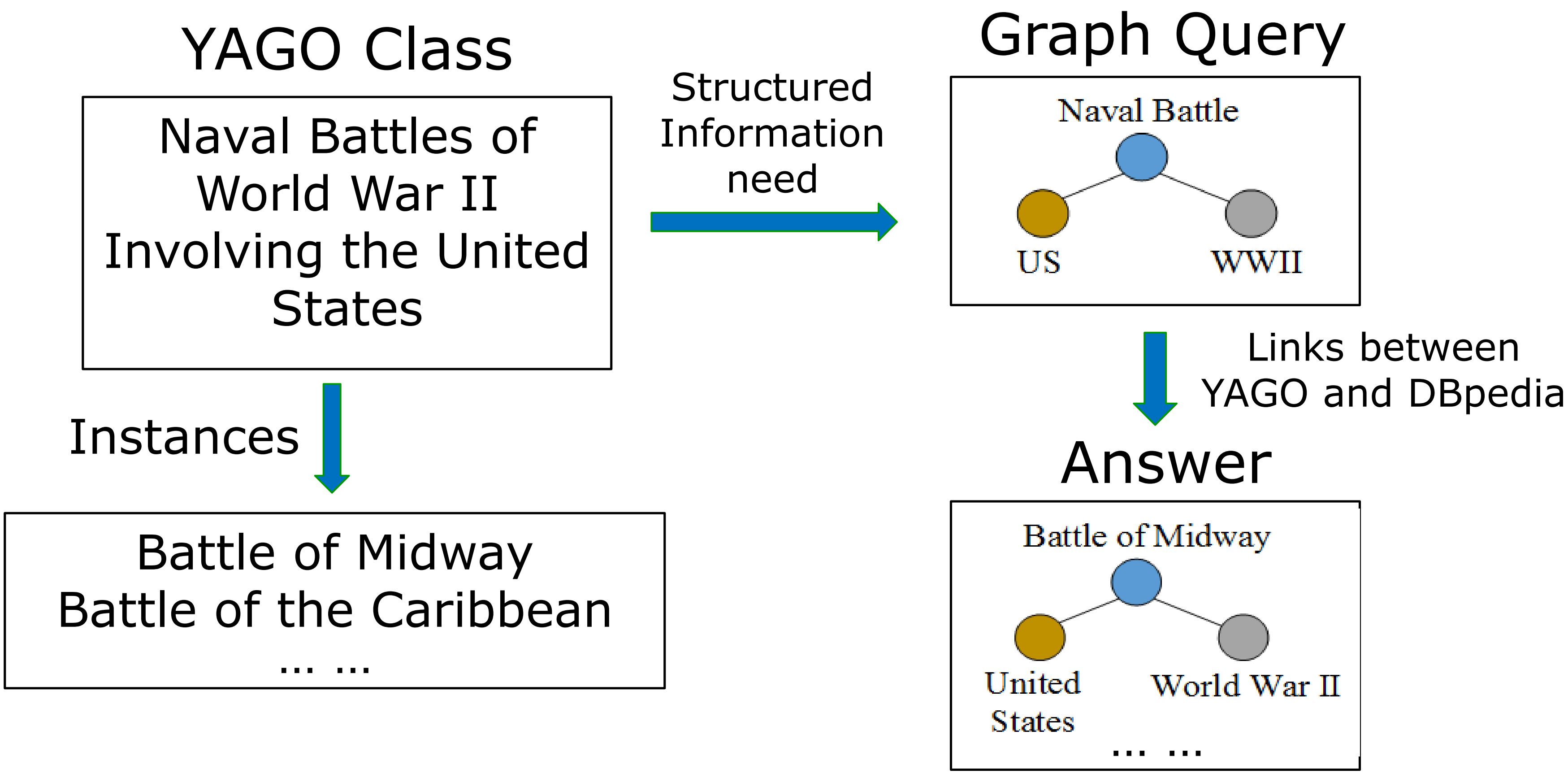
Context Inference

- *Entity context*: 1-hop neighborhood of the entity
- The contexts of the positive entities constitute a composite context for each query node



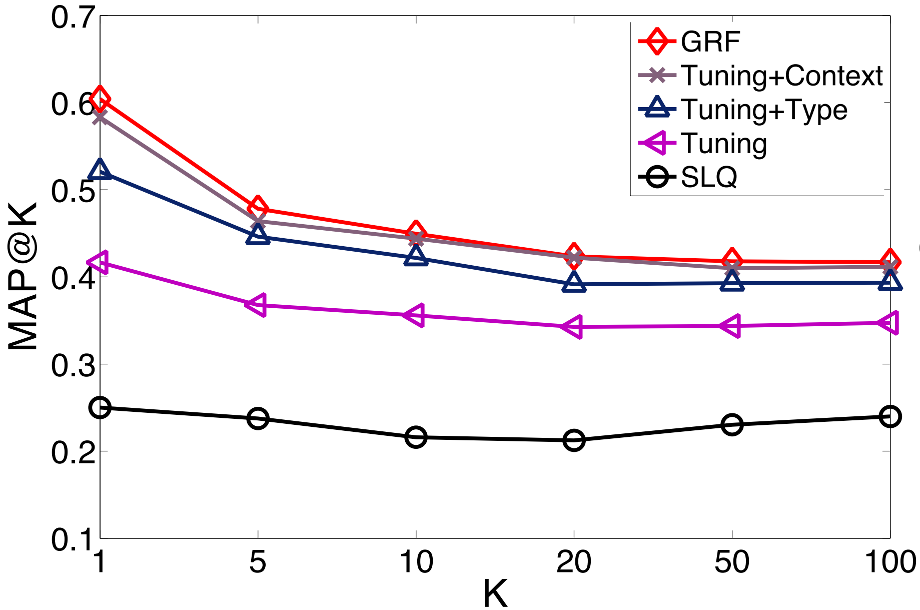
Experimental Setup

- Knowledge base: DBpedia (4.6M nodes, 100M edges)
- Graph query sets: WIKI and YAGO

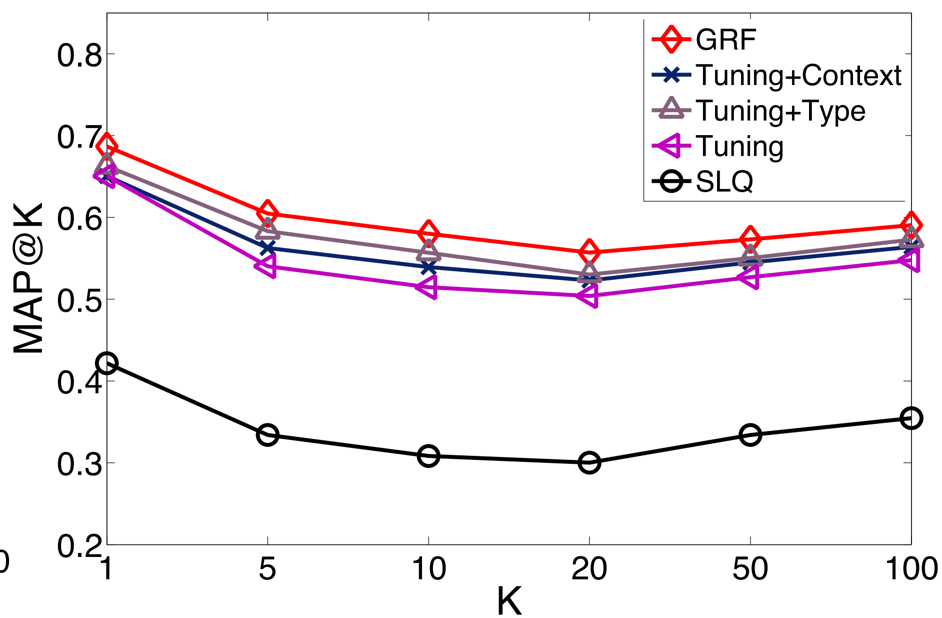


Evaluation with Explicit Feedback

- Explicit feedback: User gives relevance feedback on top-10 results
- GRF boosts SLQ by over 100%
- Three GRF components complement each other



(a) WIKI



(b) YAGO

Metric: mean average precision (MAP)@K

Evaluation with Pseudo Feedback

- Pseudo feedback: Blindly assume top-10 results from initial run are correct
- Erroneous feedback information but zero user effort

MAP@K	1	5	10	20	50	100
SLQ_WIKI	0.23	0.21	0.24	0.25	0.27	0.28
GRF_WIKI	**0.73**	**0.58**	**0.52**	**0.50**	**0.49**	**0.49**
SLQ_YAGO	0.40	0.35	0.33	0.32	0.36	0.39
GRF_YAGO	**0.82**	**0.66**	**0.60**	**0.57**	**0.58**	**0.61**

Outline

- Schema-agnostic Graph Query
- Natural Language Interface (a.k.a., Semantic Parsing)
  - A little history
  - Cold-start with crowdsourcing
  - Cold-start with neural transfer learning

Natural Language Interface ≈ Model-Theoretic Semantics

Language Variations

Utterance

find the first child of Queen Elizabeth II

semantic parsing

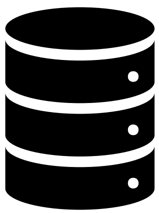
Symbol Grounding

Executable logical form
(SQL, λ-calculus, …)

argmin(child(Elizabeth), birthdate)

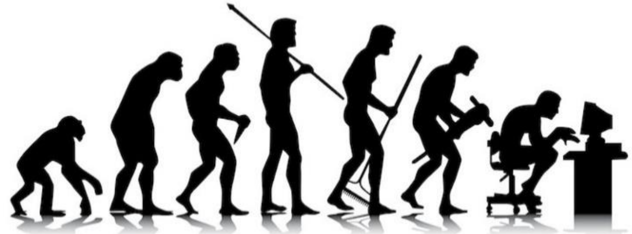
execution

World
(knowledge base,
database, …)



Denotation

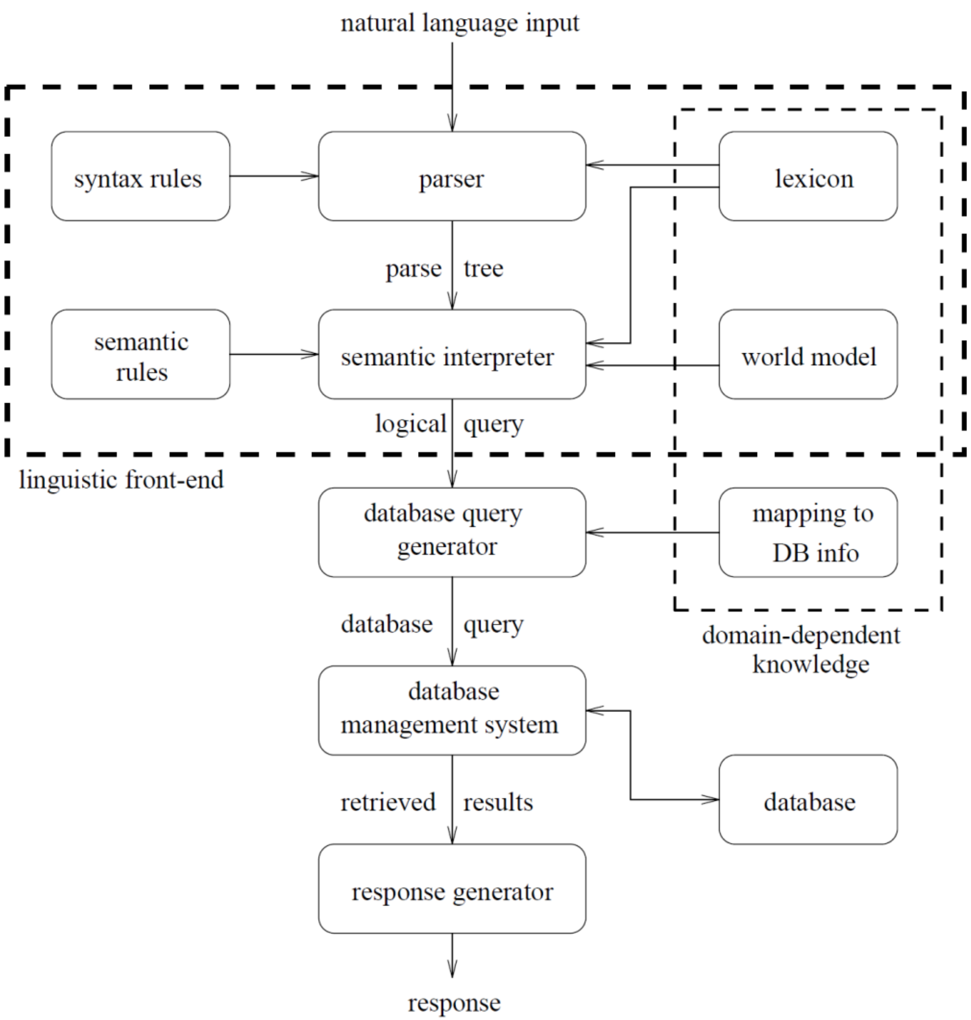
Charles, Prince of Wales



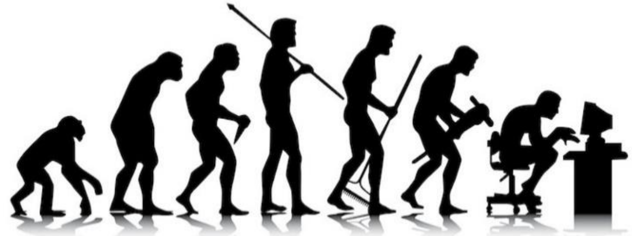
1960s-1990s | 1990s-2010s | 2015-present

	Rule-based	Statistical	Neural
Semantic Mapping			
Natural-ness			
Training Data			
Portability			

Rule-based Natural Language Interface

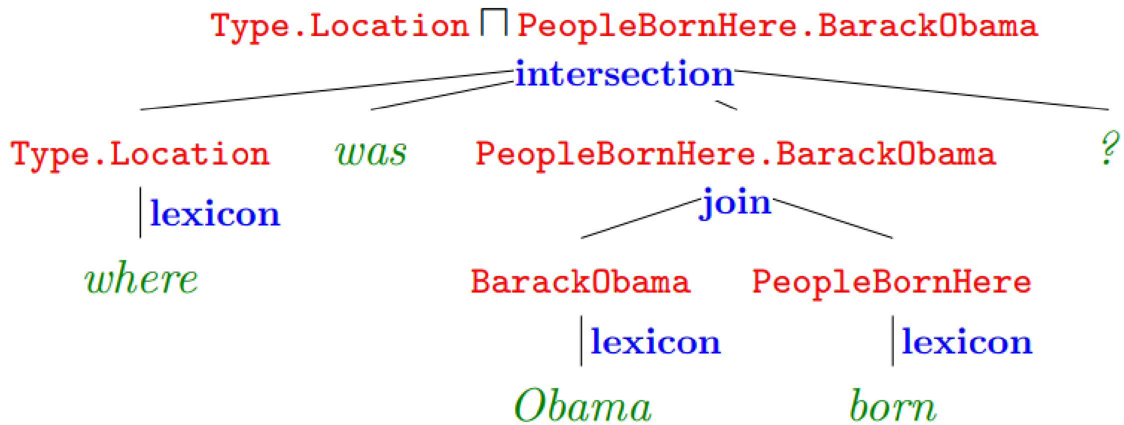



```
editor> add verb
what is your verb ? exceed
what is its third sing. pres ? exceeds
what is its past form ? exceeded
what is its perfect form ? exceeded
what is its participle form ? exceeding
to what set does the subject belong ? numeric
is there a direct object ? yes
to what set does it belong ? numeric
is there an indirect object ? no
is it linked to a complement ? no
what is its predicate ? greater_than
do you really wish to add this verb? y
```

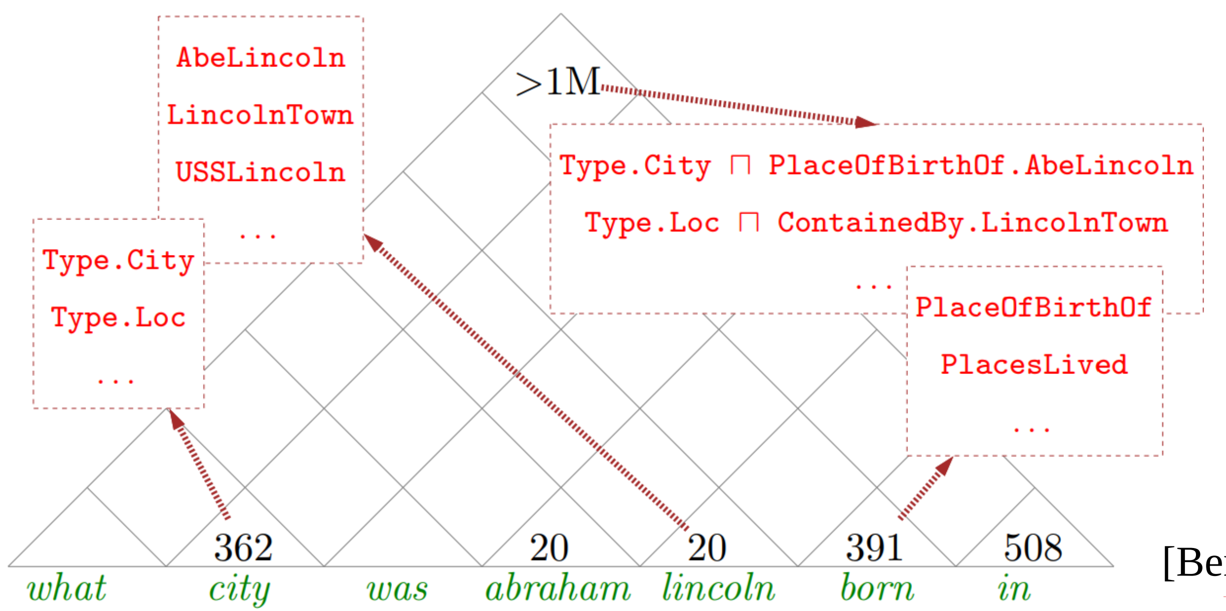


	1960s-1990s	1990s-2010s	2015-present
	Rule-based	**Statistical**	**Neural**
Semantic Mapping	• Manually designed rules • Deterministic		
Natural-ness	• Low		
Training Data	• Few		
Portability	• Low • Mostly applied on relational databases		

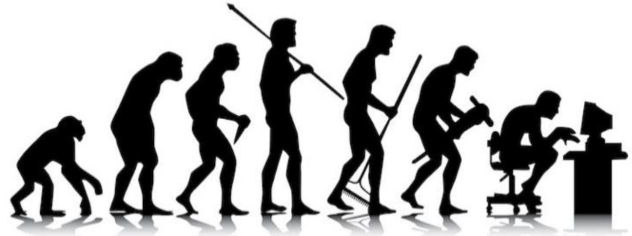
Statistical Natural Language Interface



[Berant et al., 2013]



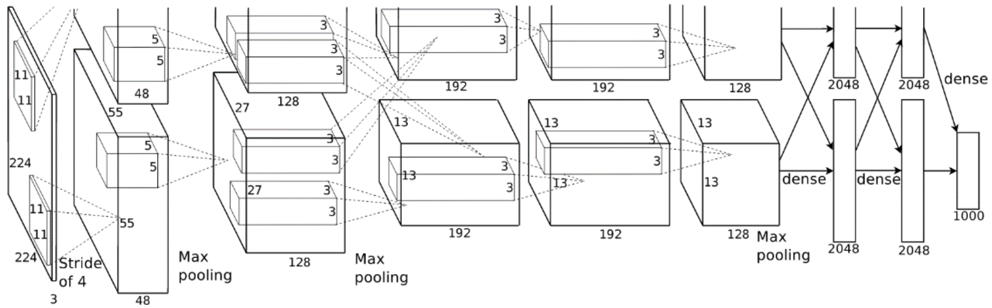
[Berant and Liang, 2015]



	1960s-1990s	1990s-2010s	2015-present
	Rule-based	**Statistical**	**Neural**
Semantic Mapping	• Manually designed rules • Deterministic	• Manually designed rules/features • Learn weights from data	
Natural-ness	• Low	• Better	
Training Data	• Few	• More	
Portability	• Low • Mostly applied on relational databases	• Better • Relational databases, knowledge bases	

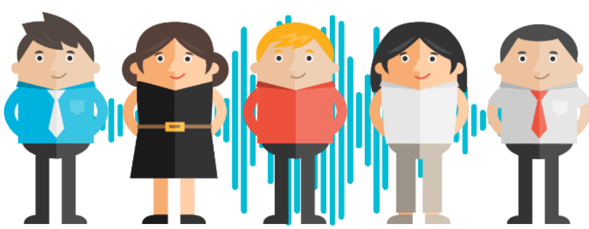
Deep Learning — Accurate, Generic, Simple

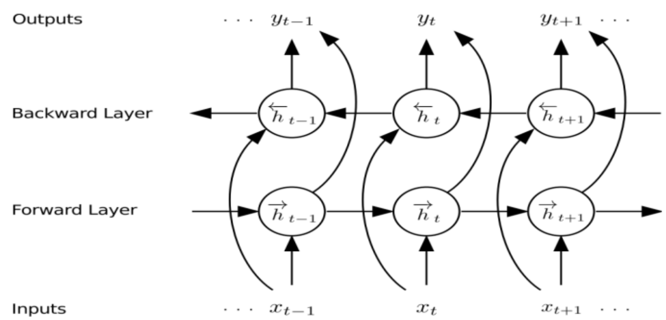
Object recognition: Krizhevsky, Sutskever, Hinton 2012



Cathedral

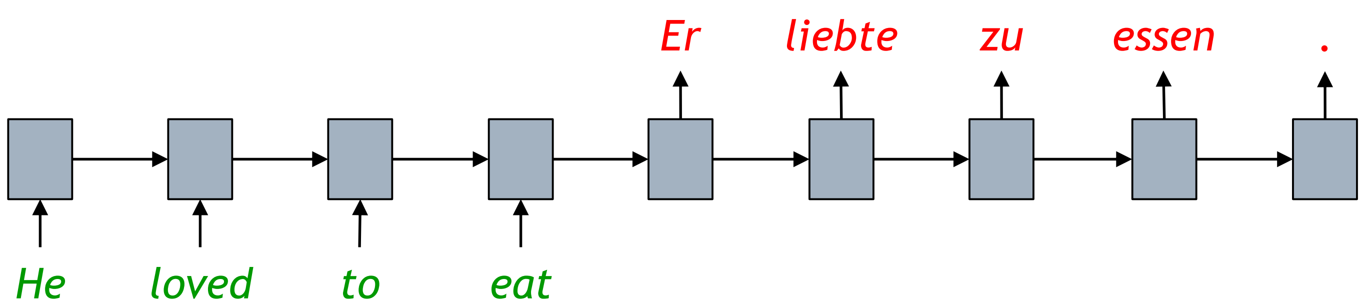
Speech recognition: Graves, Mohamed, Hinton 2013



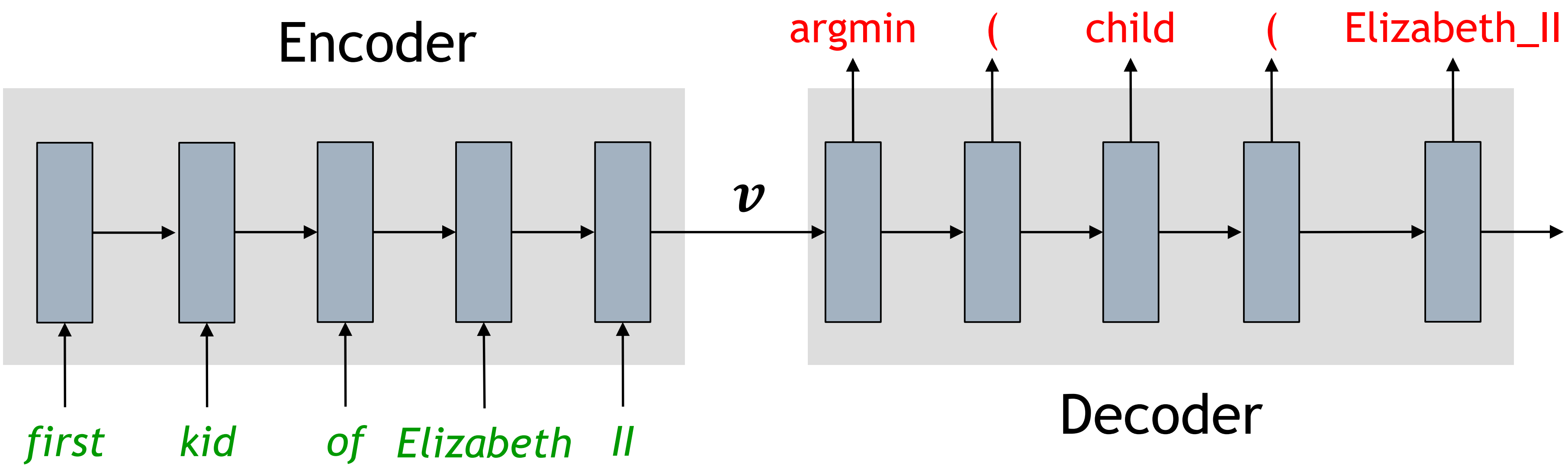


"Hey Siri, play some jazz music"

Machine translation: Sutskever, Vinyals, Le 2014



Neural Natural Language Interface



eldest kid of Elizabeth II

Her Majesty's first child

Queen Elizabeth's firstborn

$$p(y_1 = \text{"argmin"} \mid \boldsymbol{v})$$

$$\downarrow$$

$$p(y_2 = \text{"("} \mid \boldsymbol{v}, \boldsymbol{y_1})$$

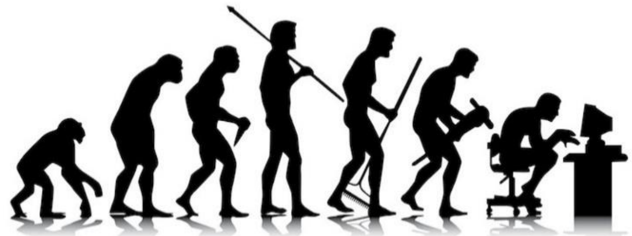
$$\downarrow$$

$$p(y_3 = \text{"child"} \mid \boldsymbol{v}, \boldsymbol{y_{1:2}})$$

$$\downarrow$$

$$\ldots$$

[Dong & Lapata 2016], [Jia & Liang 2016], [Mei et al., 2016]



	1960s-1990s	1990s-2010s	2015-present
	Rule-based	**Statistical**	**Neural**
Semantic mapping	• Manually designed rules • Deterministic	• Manually designed Rules/features • Learn weights from data	• Both features and weights learned from data
Natural-ness	• Low	• Better	• Best
Training Data	• Few	• More	• A LOT more
Portability	• Low • Mostly applied on relational databases	• Better • Relational databases, knowledge bases	• Best • Relational databases, knowledge bases, web tables, APIs, …

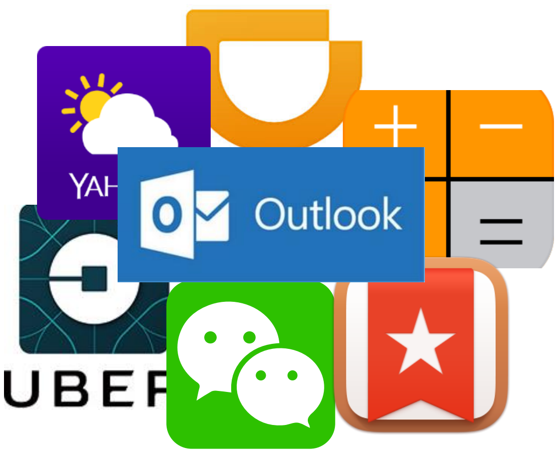
Outline

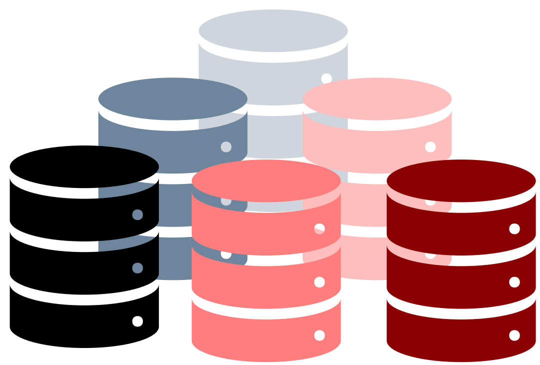
- Schema-agnostic Graph Query
- Natural Language Interface (a.k.a., Semantic Parsing)
  - A little history
  - Cold-start with crowdsourcing
  - Cold-start with neural transfer learning

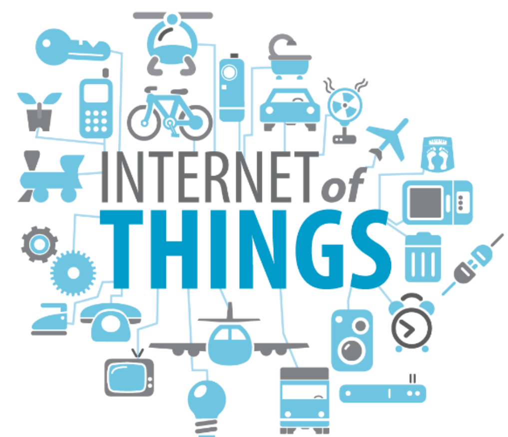
Portability: the Cold Start Problem

Portability: the Cold Start Problem

"I want to build an NLI for my domain, but I don't have any user and training data"







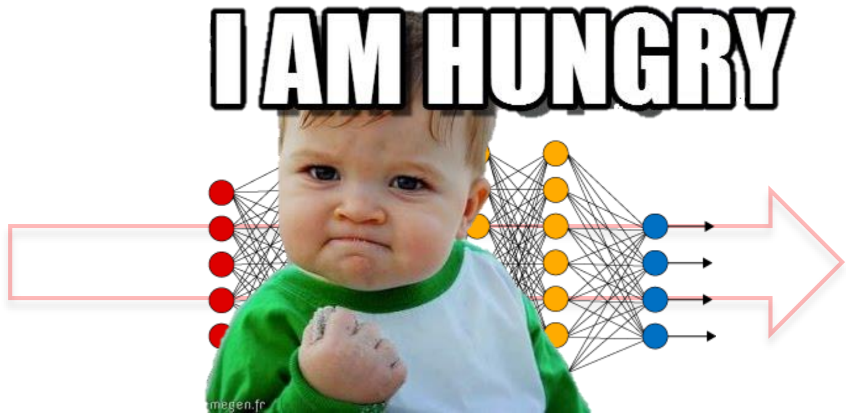
How to Build NLI for New Domain

- 1950s-1990s: Rule engineering (for rule-based NLI)
- 1990s-2010s: Feature engineering (for statistical NLI)
- 2015-present: Data engineering (for neural NLI)
  - Crowdsourcing
  - Neural transfer learning

```
editor> add verb
what is your verb ? exceed
what is its third sing. pres ? exceeds
what is its past form ? exceeded
what is its perfect form ? exceeded
what is its participle form ? exceeding
to what set does the subject belong ? numeric
is there a direct object ? yes
to what set does it belong ? numeric
is there an indirect object ? no
is it linked to a complement ? no
what is its predicate ? greater_than
do you really wish to add this verb? y
```

[Auxerre and Inder, 1986]

Deep Learning with Weak Supervision

Text  Knowledge

Strong Supervision

- In-domain, on-task

Weak Supervision

- In-domain, off-task
- Out-of-domain, on-task
- Out-of-domain, off-task

How to Collect NLI Training Data?

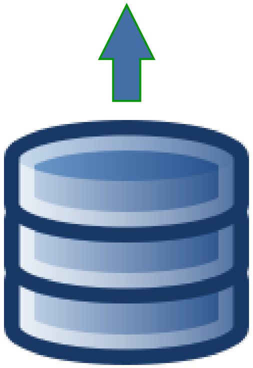
- Training data: {(utterance, logical form)}

(*"Who did nine-eleven?"*, λx.involved_in_attach(x, September_11_attacks))

(*"How many children of Eddard Stark were born in Winterfell?"*,

count(λx.children(Eddard_Stark, x) ∧ place_of_birth(x, Winterfell)))

...



How to Collect NLI Training Data?

- If we already have utterances (questions/commands/queries/…) from users…

“How many children of Eddard Stark were born in Winterfell?”

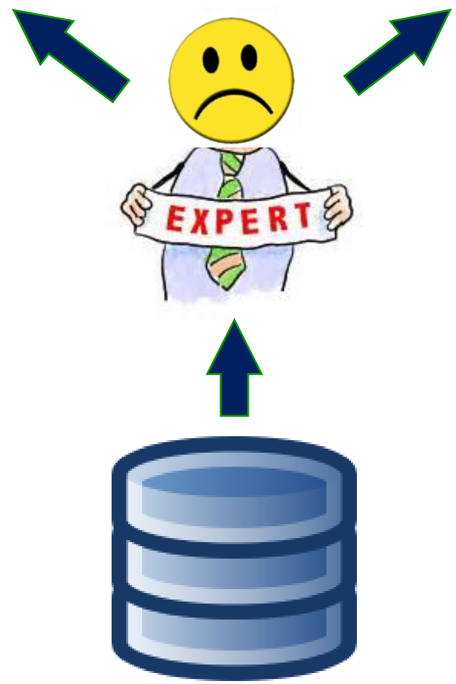
count(λx.children(Eddard_Stark, x) ∧ place_of_birth(x, Winterfell))

How to Collect NLI Training Data?

- But for most domains we are interested in, there is yet any user, nor any utterance
- Ask domain experts to do everything?

"How many children of Eddard Stark were born in Winterfell?"

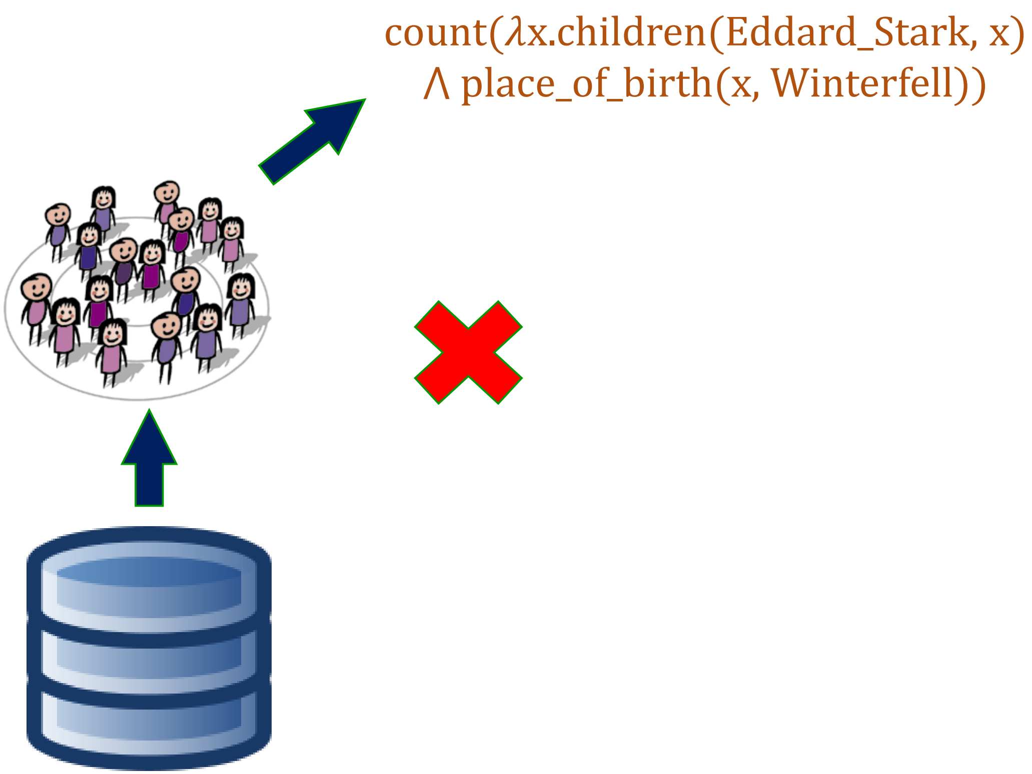
count(λx.children(Eddard_Stark, x) $\wedge$ place_of_birth(x, Winterfell))



- Do not scale
- Not representative

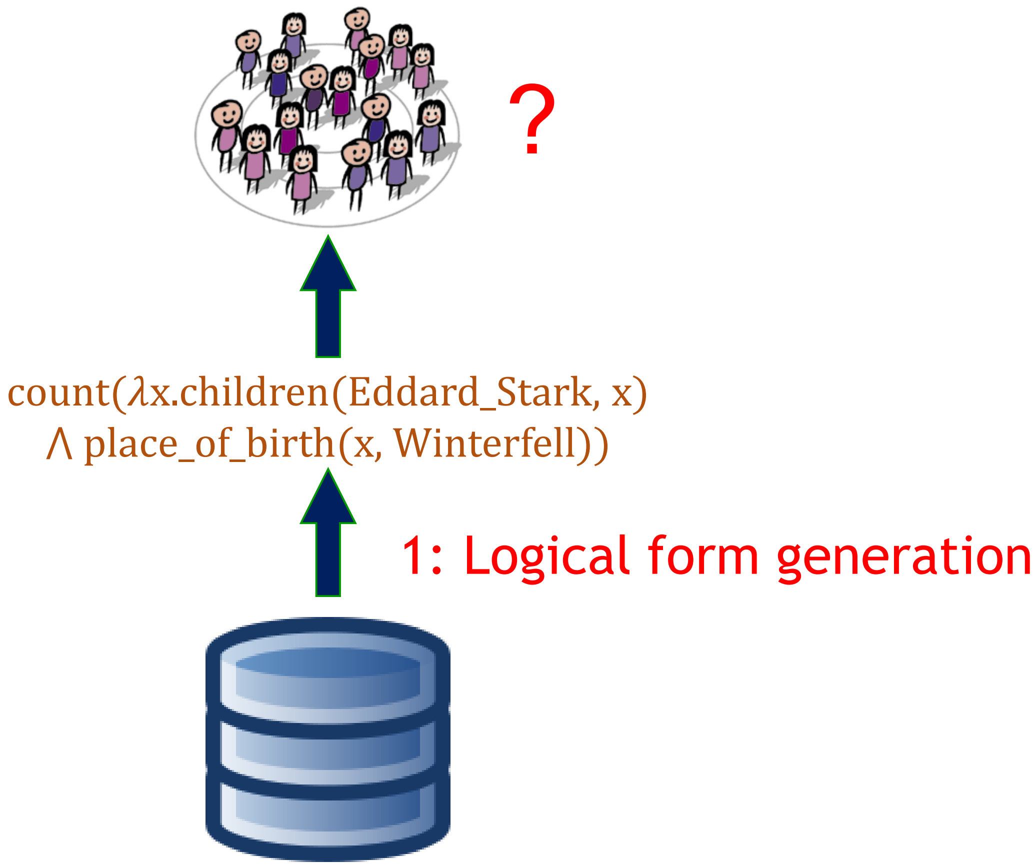
How to Collect NLI Training Data?

- Can we only use crowd workers?
- Crowd workers do not understand formal languages!



How to Collect NLI Training Data?

- Can we only use crowd workers?
- Crowd workers do not understand formal languages!



A General Framework for Crowdsourcing NLI Data

"How many children of Eddard Stark were born in Winterfell?"

3: Paraphrasing via crowdsourcing

"What is the number of person who is born in Winterfell, and who is child of Eddard Stark?"

2: Canonical utterance generation

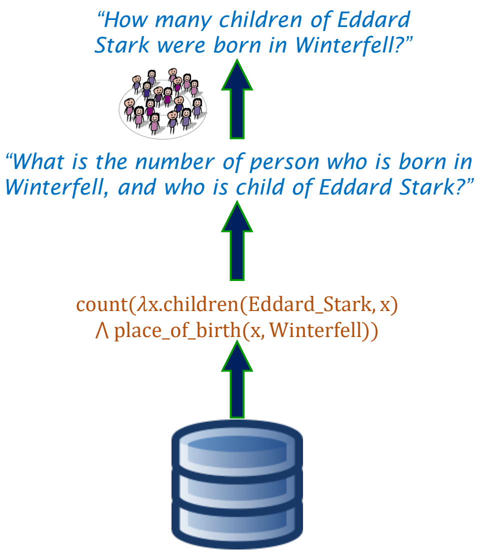
$$\mathrm{count}(\lambda x.\mathrm{children}(\mathrm{Eddard\_Stark}, x) \\ \wedge \mathrm{place\_of\_birth}(x, \mathrm{Winterfell}))$$

1: Logical form generation

[Building a Semantic Parser Overnight, Wang et al. 2015]

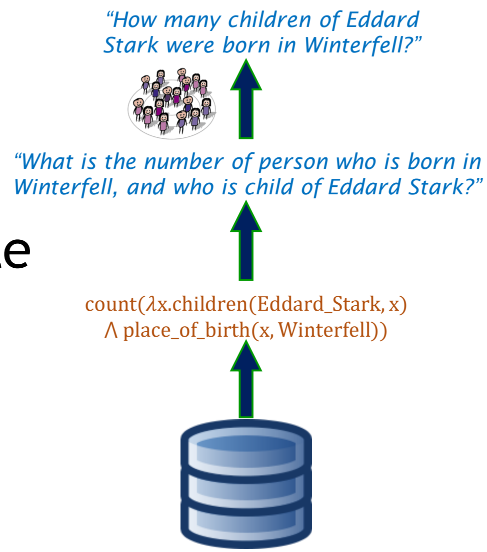
Advantages

- Scalable
  - Low-cost annotation, applicable to many domains
- Configurable
  - Full control on what to annotate and how many to get
- Complete coverage
  - Fully exercise the formal language and data
- Representative (partially)
  - Natural wording
  - Do not capture distribution of user interests



Challenges

- Logical form generation
  - How to automate and configure?
  - What logical forms are “relevant”?
  - How many to generate (huge candidate space)
- Canonical utterance generation
  - How to minimize the expertise requirement and workload for grammar design
- Paraphrasing via crowdsourcing
  - How to optimize the crowdsourcing process, i.e., select the right logical forms to annotate
  - How to control and improve result quality
  - How to encourage diversity



On Generating Characteristic-rich Question Sets for QA Evaluation

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa,
Izzeddin Gur, Zenghui Yan, Xifeng Yan

UCSB, OSU, Army Research Lab, IBM Research

EMNLP'16

Motivation

- Existing datasets for knowledge based question answering (KBQA) mainly contain *simple questions*
  - WebQuestions, SimpleQuestions, etc.

“*Where was Obama born?*”

“*What party did Clay establish?*”

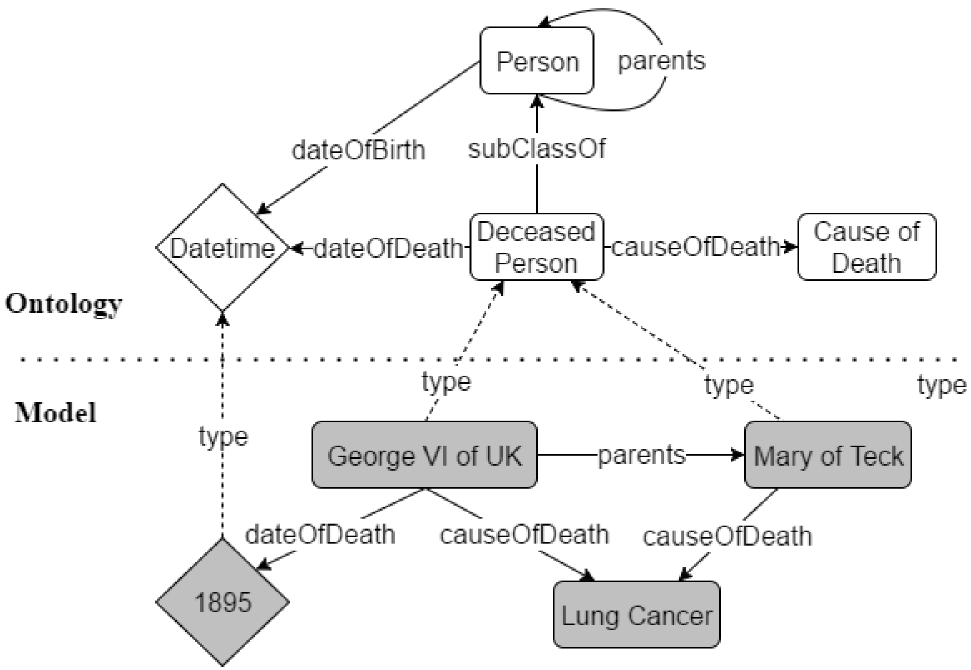
“What kind of money to take to bahamas?”

… …

Multi-dimensional Benchmarking

- Structural complexity
  - *“People who are on a gluten-free diet can’t eat what cereal grain that is used to make challah?”*
- Quantitative analysis (function)
  - *“In which month does the average rainfall of New York City exceed 86 mm?”*
- 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?”*
- …

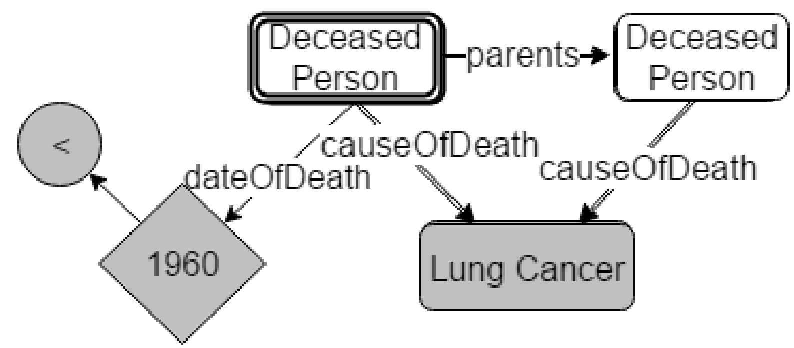
Configurable Benchmark Construction



Freebase

53K classes, 35K relations, 45M entities, 3B facts

Configurable, Quality Control



Logical Form

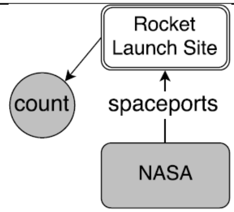
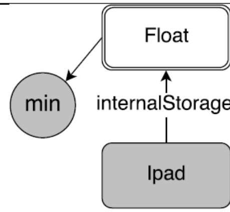
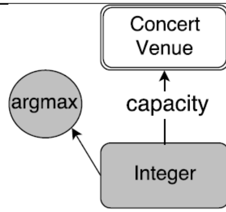
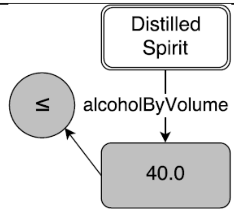
Natural Language Paraphrases

- *"Find people who died from lung cancer, same as their parent."*
- *"From those lung cancer deaths, list the ones whose parent has the same cause of death"*

V1: Graduate students

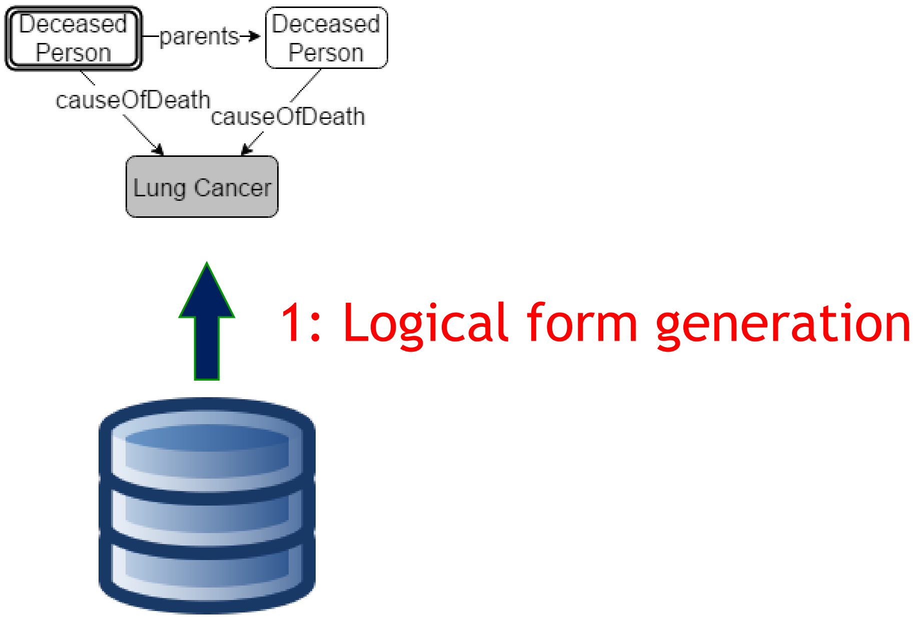
V2: Crowdsourcing (multi-stage quality control), 10x scale

Functions

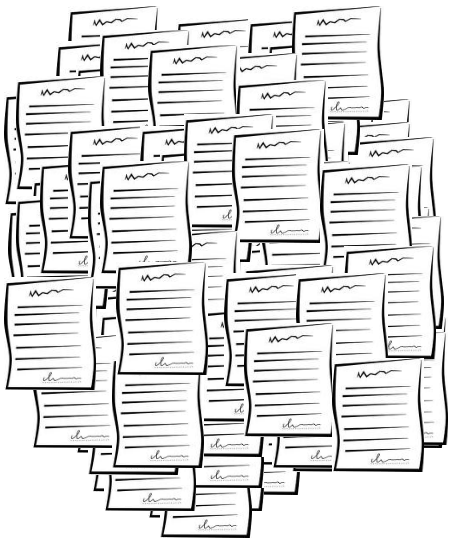
<table>
<tr><th>Category</th><th>Counting</th><th colspan="2">Superlative</th><th>Comparative</th></tr>
<tr><td>Functions</td><td><code>count</code></td><td><code>max</code> and <code>min</code></td><td><code>argmax</code> and <code>argmin</code></td><td>$<, >, \leq$, and $\geq$</td></tr>
<tr><td>Domain</td><td>Question node</td><td>Question node of numeric class</td><td>Template/grounded node of numeric class</td><td>Template/grounded node of numeric class</td></tr>
<tr><td>Example</td><td>
</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Question</td><td>How many launch sites does nasa have?</td><td>What's the smallest internal storage of ipad?</td><td>Find the largest concert venue.</td><td>List distilled spirits with no more than 40.0% abv.</td></tr>
</table>

Too Many Graph Queries

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



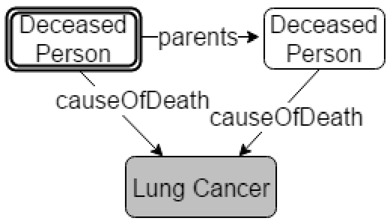
Commonness Checking



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

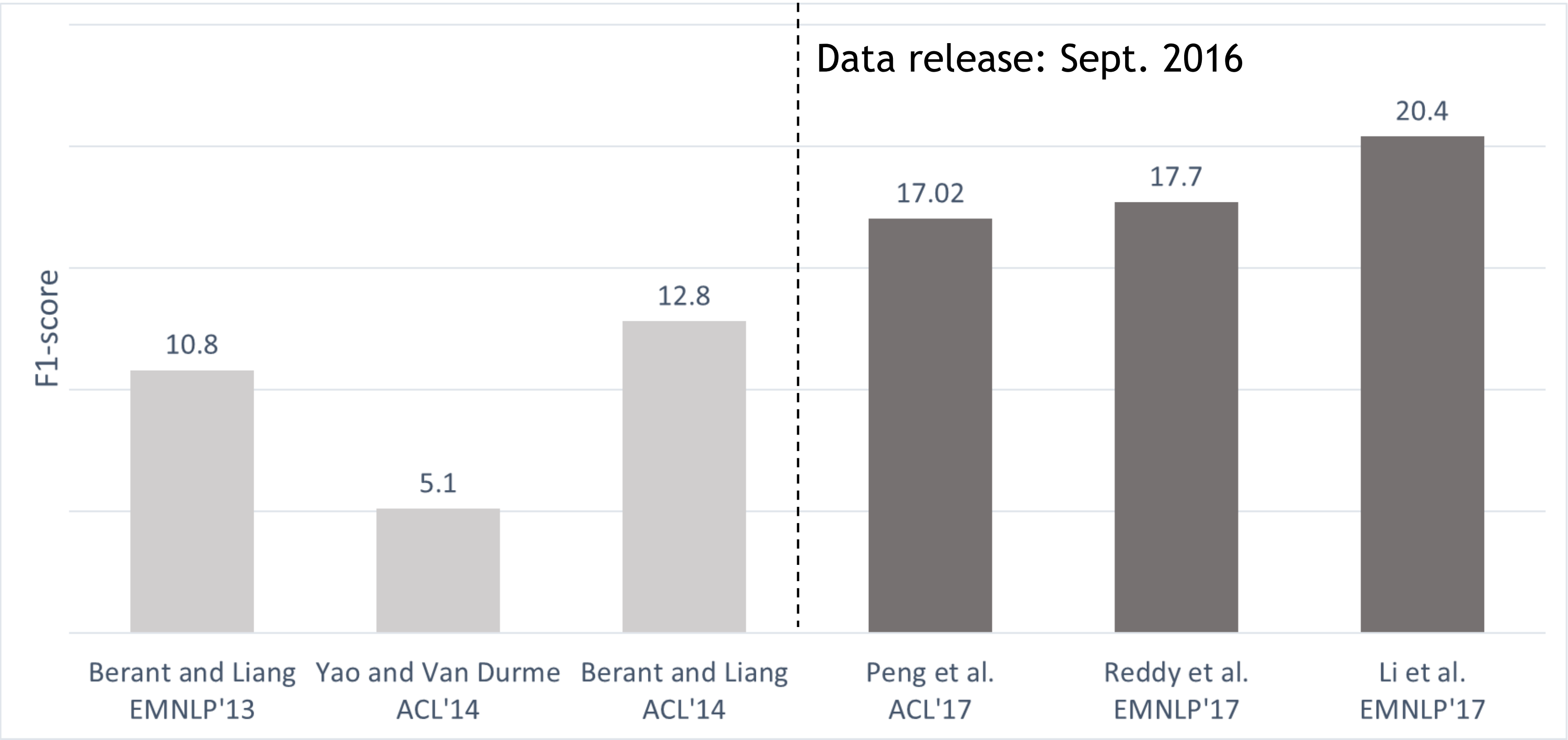
$p($  $)$

GraphQuestions

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

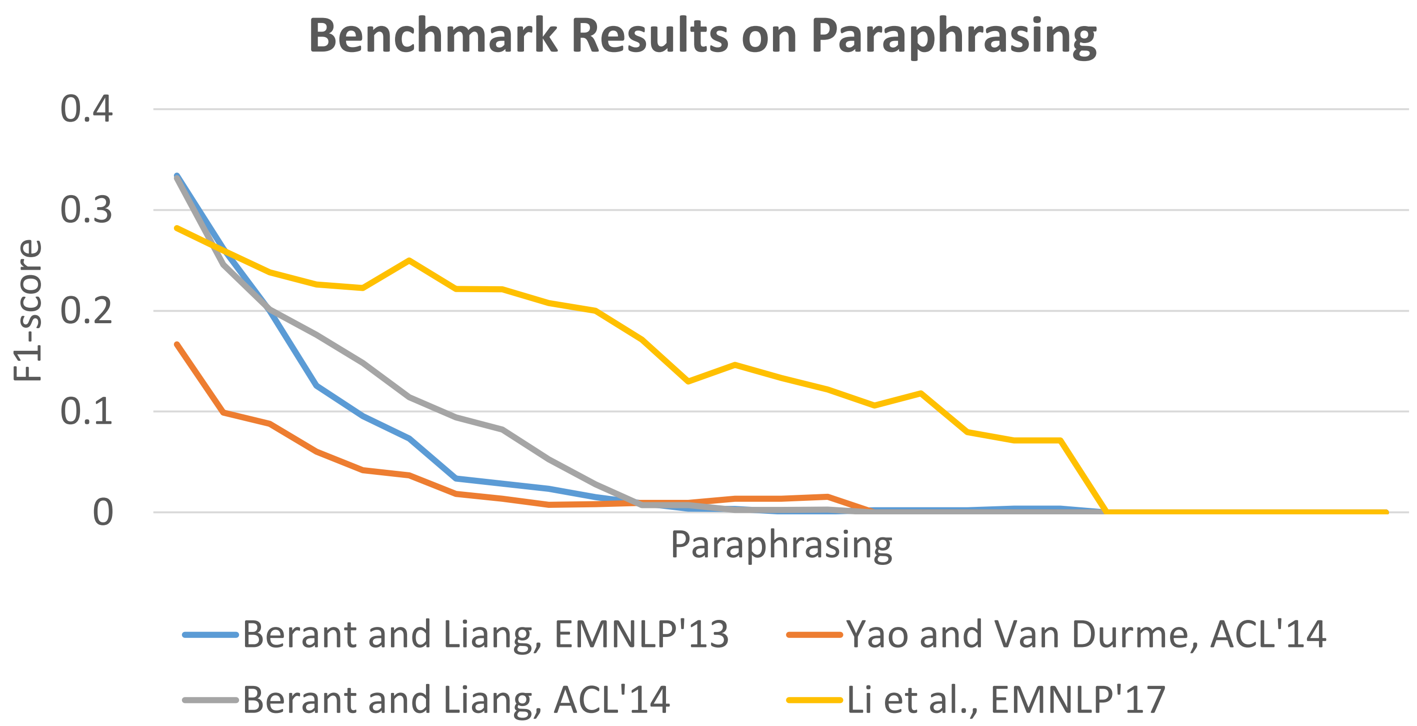
Question	Domain	Answer	# of edges	Function	$\log_{10}(p(q))$	$\|A\|$
Find terrorist organizations involved in **September 11 attacks**.	Terrorism	alQaeda	1	none	-16.67	1
The **September 11 attacks** were carried out with the involvement of what terrorist organizations?						
Who did **nine eleven**?						
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?						

Testbest for Research Progress



Pointing out Future Directions

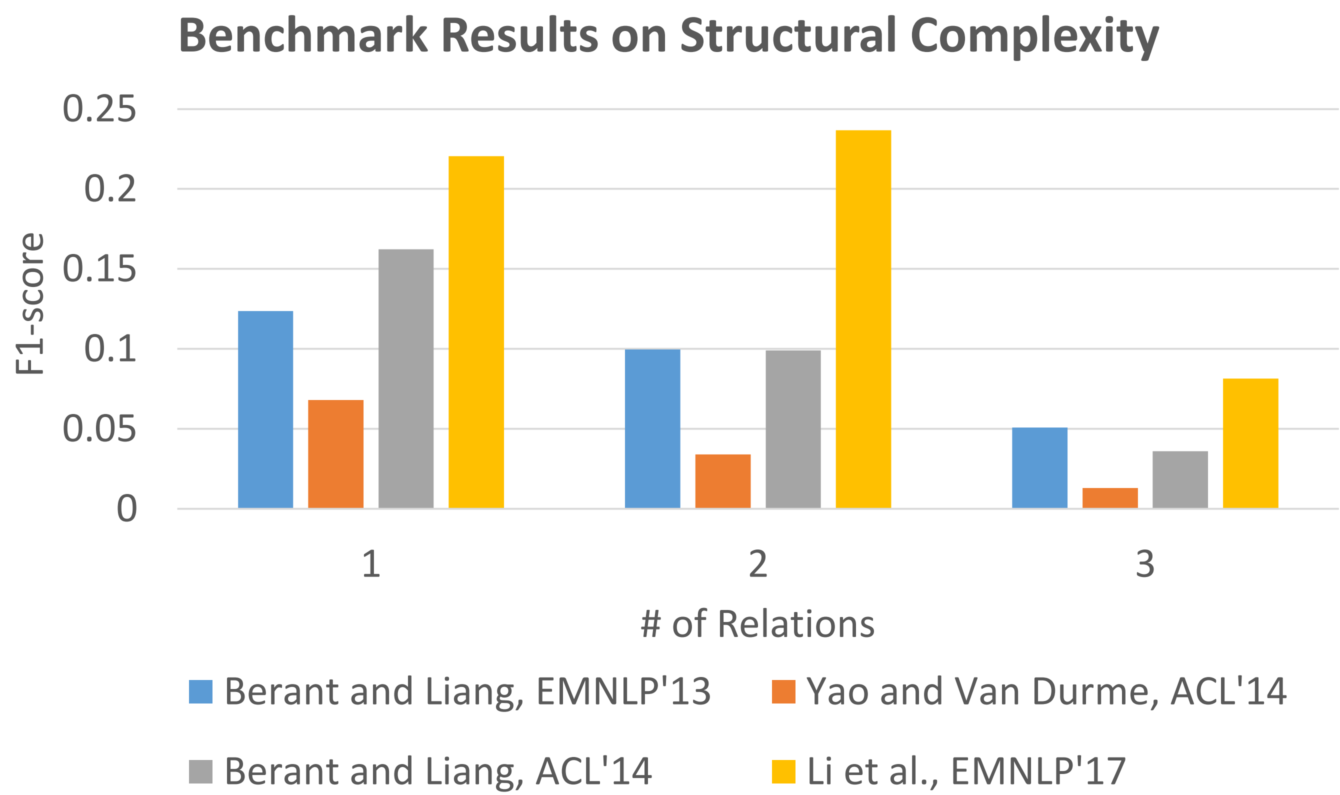
"What is the nutritional composition of coca-cola?"
"What is the supplement information for coca-cola?"
"What kind of nutrient does coke have?"



"Learning to Paraphrase for Question Answering"
Dong et al., *EMNLP* (2017)
(Su et al., 2016)

The Quest of Compositionality

[*people who are on a gluten-free diet*]$_{rel1}$ [*can't eat*]$_{rel2}$ [*what cereal grain that is used to make challah*]$_{rel3}$

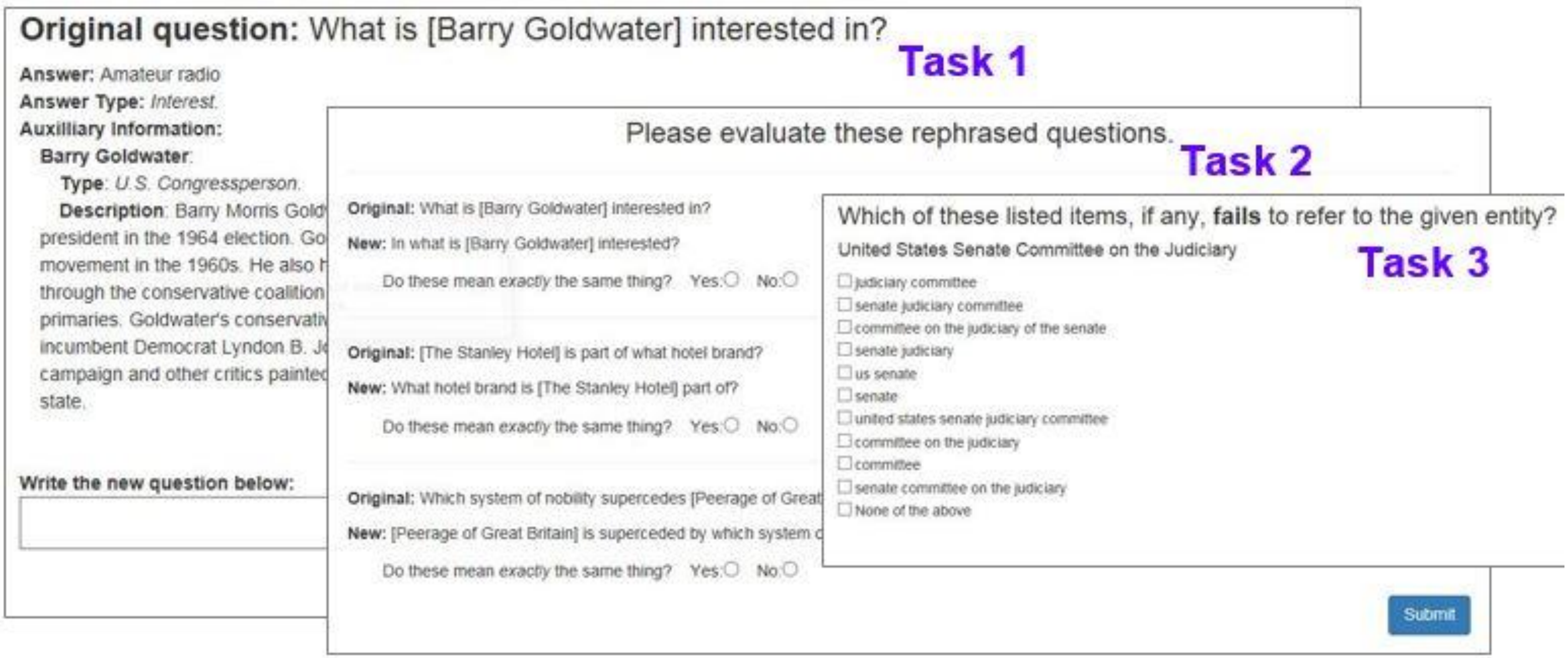


Further study on compositionality in CIKM'17 and SIGIR'18 (under review)

(Su et al., 2016)

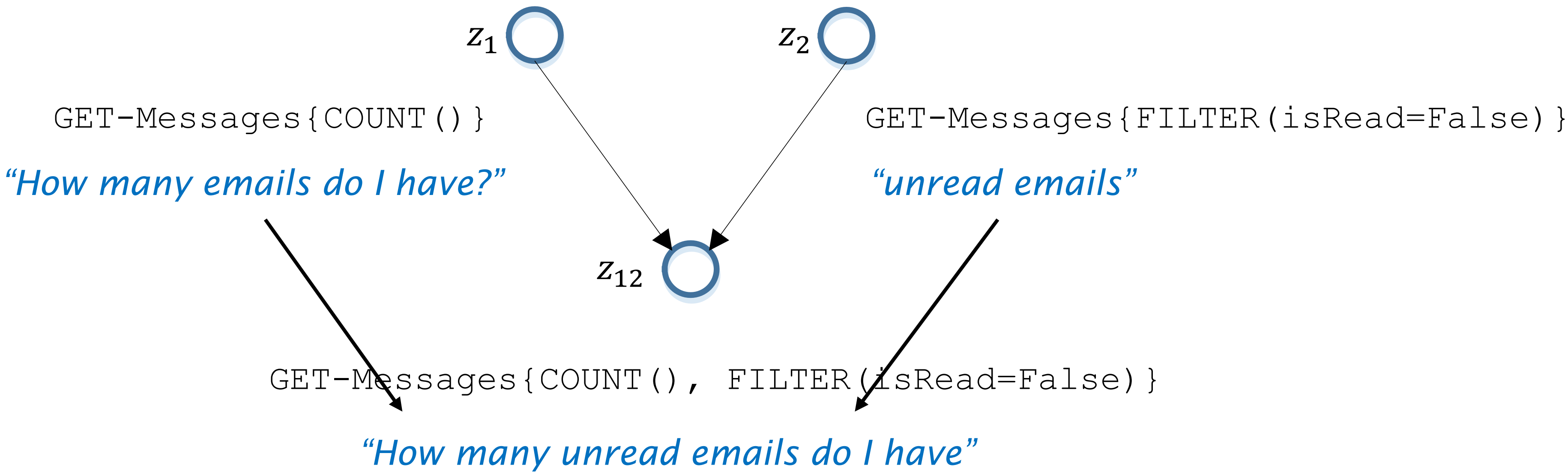
GraphQuestions V2 (coming soon)

- 10 to 20 times larger in scale
- Support more benchmarking scenarios
  - Cross-domain transfer learning, few- or zero-shot learning, compositionality, etc.



Crowdsourcing Optimization

- Which logical forms are of a high value for training NLI?



Utterances follow the composition structure of API calls

Predict the language model of an API call **without annotating it!**

Crowdsourcing optimization

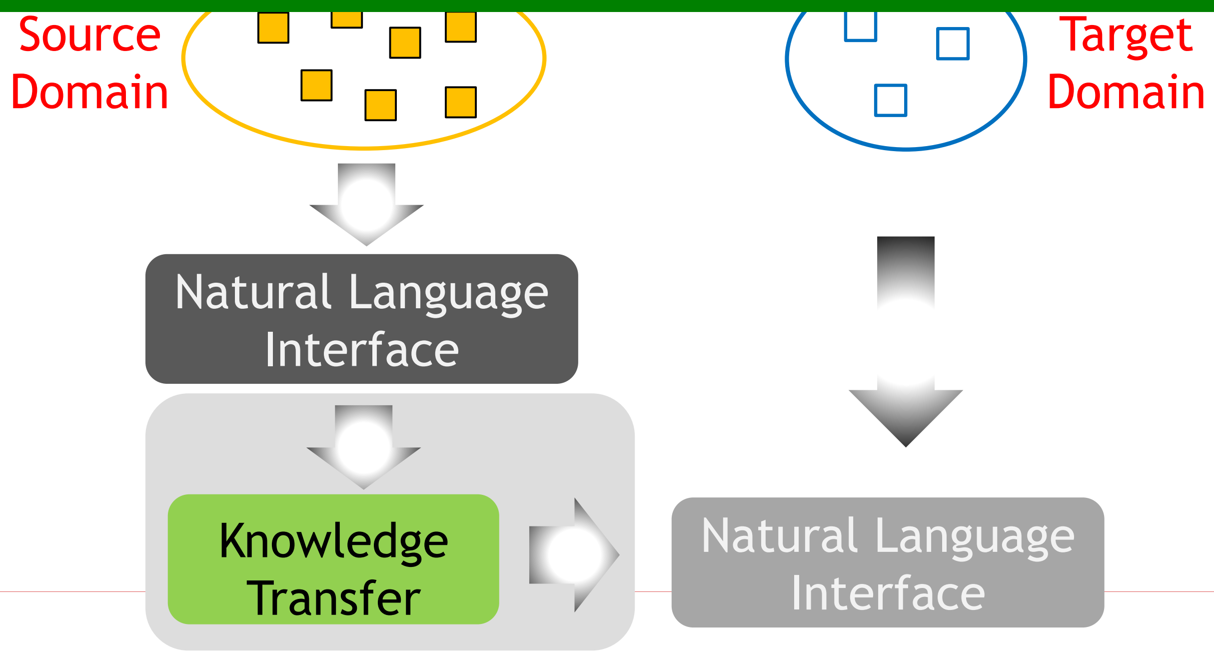
Outline

- Schema-agnostic Graph Query
- Natural Language Interface (a.k.a., Semantic Parsing)
  - A little history
  - Cold-start with crowdsourcing
  - Cold-start with neural transfer learning

How to Build NLI for New Domain

- 1950s-1990s: Rule engineering (for rule-based NLI)
- 1990s-2010s: Feature engineering (for statistical NLI)
- 2015-present: Data engineering (for neural NLI)
  - Crowdsourcing
  - Neural transfer learning

Out-of-domain, on-task supervision!



What is Transferrable in NLI across Domains?

Source Domain: Basketball

In which season did Kobe Bryant play for the Lakers? → $\mathbf{R}$[season]. (player.KobeBryant ⊓ team.Lakers)

p(relation1|"*play for*")

p(team|"*play for*")

p(relation2|"*work for*")

p(employer|"*work for*")

Target Domain: Social

When did Alice start working for Mckinsey? → $\mathbf{R}$[start]. (employee.Alice ⊓ employer.Mckinsey)

[EMNLP'17]

Cross-domain NLI via Paraphrasing

$\mathbf{R}[\text{season}].(\text{player.KobeBryant} \sqcap \text{team.Lakers})$

automatic

In which season did Kobe Bryant play for the Lakers?

Season of Player Kobe Bryant whose team is Lakers

$p($"*whose team is*"$|$"*play for*"$)$

play ≈ *work*, *team* ≈ *employer*

$p($"*whose employer is*"$|$"*work for*"$)$

When did Alice start working for Mckinsey?

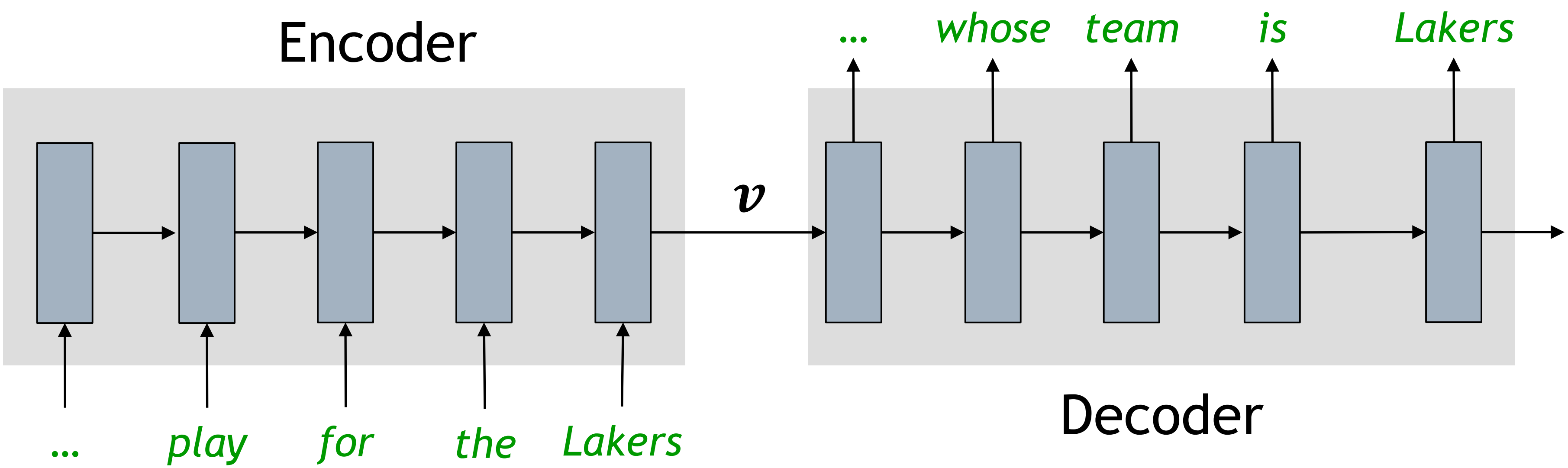
Start date of employee Alice whose employer is Mckinsey

automatic

$\mathbf{R}[\text{start}].(\text{employee.Alice} \sqcap \text{employer.Mckinsey})$

[EMNLP’17], inspired by [Berant & Liang, 2014], [Wang et al., 2015]

Seq2Seq Model for Paraphrasing

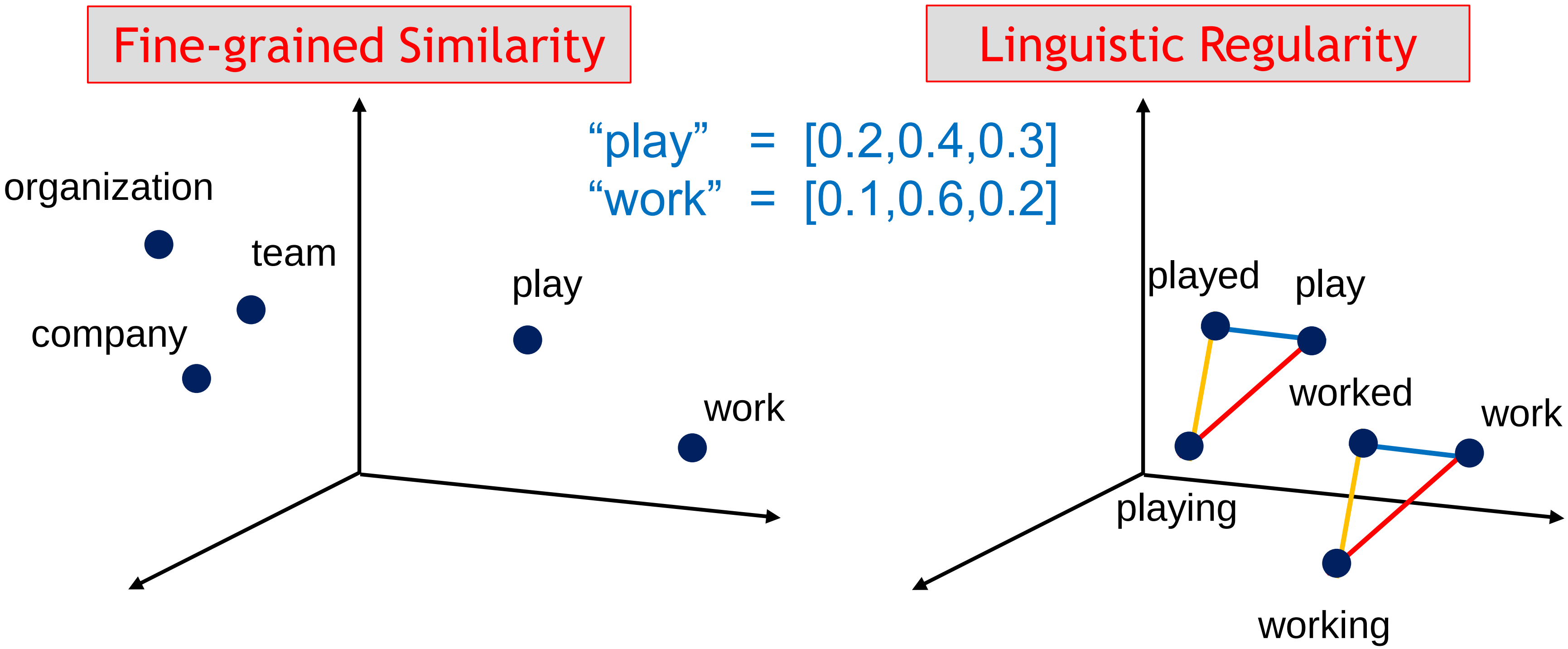


- Seq2Seq + Bi-directional encoder + Attentive decoder
- Learn to predict whether input utterance paraphrases canonical utterance
- Deterministic mapping between canonical utterance and logical form

[EMNLP'17]

Word Embedding

- Word ≜ Dense vector (typically 50-1000 dimensional)
- Word similarity ≜ Vector similarity
- Pre-trained on huge external text corpora



Out-of-domain, off-task supervision!

Pre-trained Word Embedding Alleviates Vocabulary Shifting

- Vocabulary shifting: Only 45%~70% target domain vocabulary are covered by source domains[1]
- Pre-trained word embedding can alleviate the vocabulary shifting problem
  - Word2vec: 300-d vectors pre-trained on the 100B-token Google News Corpus; vocabulary size = 3M

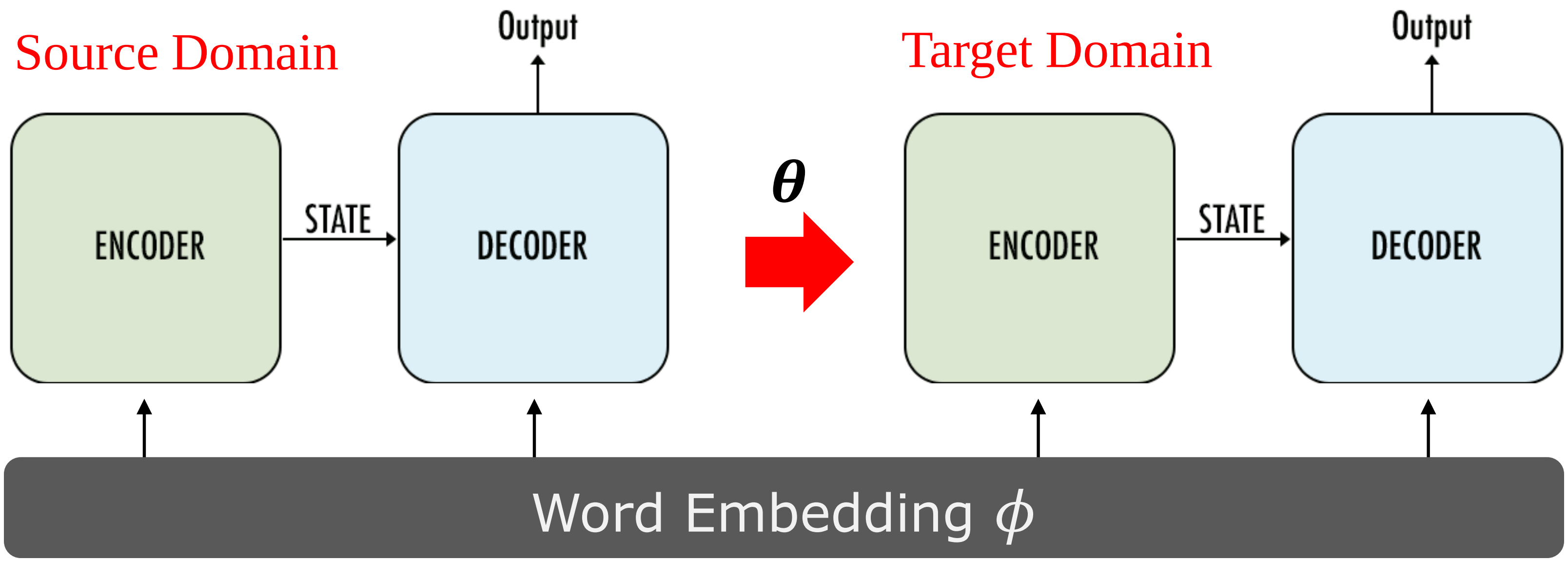
	Calendar	Housing	Restaurants	Social	Publications	Recipes	Basketball	Blocks
Coverage	71.1	60.7	55.8	46.0	65.6	71.9	45.6	61.7
+word2vec	**93.9**	**90.9**	**90.4**	**89.3**	**95.6**	**97.3**	**89.4**	**93.8**

Overnight Dataset: 8 KBs

[1] Wang et al. Building a Semantic Parser Overnight. 2015

(Su et al., 2017)

Neural Transfer Learning for NLI



- Input utterance $\boldsymbol{x} = (x_1, \ldots, x_m)$, canonical utterance $\boldsymbol{y} = (y_1, \ldots, y_n)$
- Embedding: $\phi(\boldsymbol{x}) = (\phi(x_1), \ldots, \phi(x_m))$, $\phi(\boldsymbol{y}) = (\phi(y_1), \ldots, \phi(y_n))$
- Learning on source domain: $p(\phi(\boldsymbol{y})|\phi(\boldsymbol{x}), \boldsymbol{\theta})$
- Warm start on target domain: $p(\phi(\boldsymbol{y})|\phi(\boldsymbol{x}), \boldsymbol{\theta})$
- Fine-tuning on target domain: $p(\phi(\boldsymbol{y})|\phi(\boldsymbol{x}), \boldsymbol{\theta}^*)$

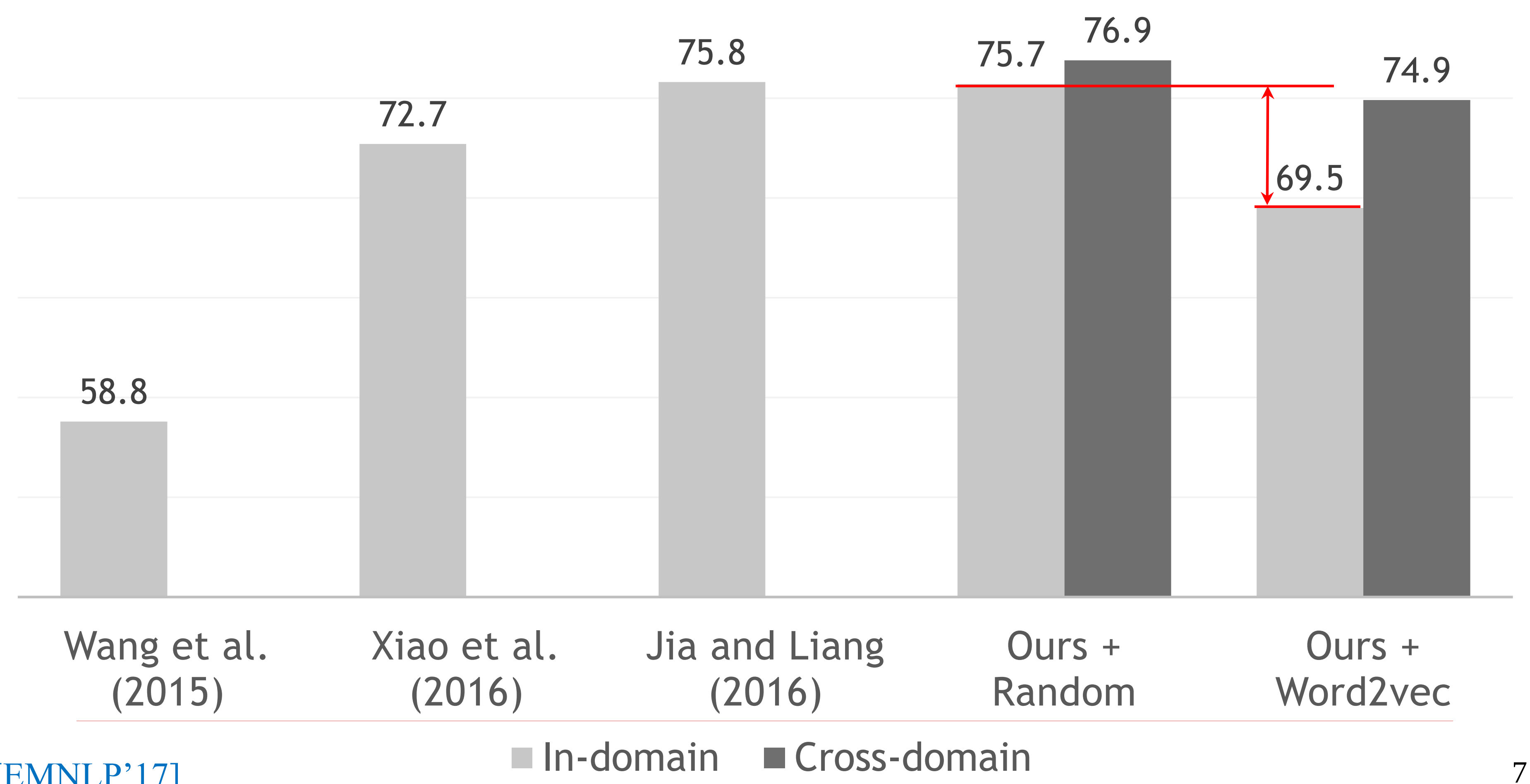
[EMNLP'17]

Experimental Setup

- Dataset: Overnight [Wang et al., 2015]
  - 8 domains (Social, Basketball, Restaurant, etc.)
- Metric: average accuracy
- Transfer learning setup
  - For each target domain, use the other 7 domains as source
- Word embedding initialization
  - **Random**: Randomly draw from uniform distribution with unit variance $U(-\sqrt{3}, \sqrt{3})$
  - **Word2vec**: 300-dimensional word2vec (skip-gram) embedding pre-trained on 100B-word News corpus

Direct Use of Word2vec Fails Dramatically…

- Transfer learning works (new state of the art)
- Word2vec brings 6.2% absolute decrease in accuracy



[EMNLP'17]

Problems of Pre-trained Word Embedding

- Small *micro variance*: hurt optimization
  - Activation variances ≈ input variances [Glorot & Bengio, 2010]
  - Small input variance implies poor exploration in parameter space
- Large *macro variance*: hurt generalization
  - Distribution discrepancy between training and testing

$$\left.\overbrace{\begin{bmatrix} | & | & \cdots & | \\ \boldsymbol{w}_1 & \boldsymbol{w}_2 & \cdots & \boldsymbol{w}_n \\ | & | & \cdots & | \end{bmatrix}}^{n=|V|}\right\} m$$

Initialization	L2 norm	Variance	Cosine Sim.
Random	17.3 ± 0.45	1.00 ± 0.05	0.00 ± 0.06
WORD2VEC	2.04 ± 1.08	0.02 ± 0.02	0.13 ± 0.11

Micro Variance
Variance of the values comprising a vector

Macro Variance
$$\frac{\sum_{i=1}^{n} var(\boldsymbol{w}_i)}{n}$$
Variance among different vectors

Proposed Solution: Standardization

- Standardize each word vector to unit variance
- But it was unclear before why standardization should be applied on pre-trained word embedding
  - Obvious downside: make loss function of word embedding sub-optimal

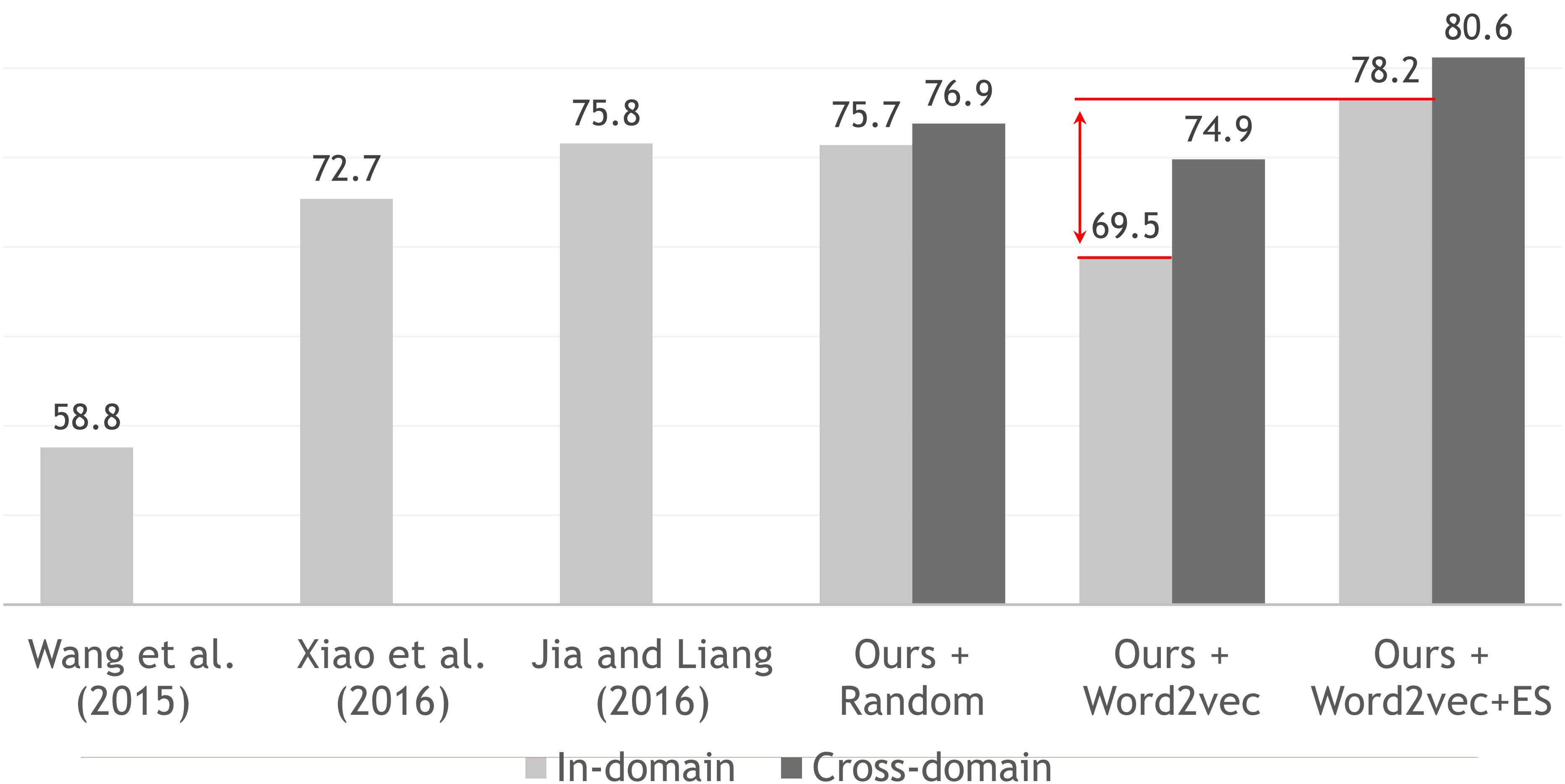
Initialization	L2 norm	Variance	Cosine Sim.
Random	17.3 ± 0.45	1.00 ± 0.05	0.00 ± 0.06
WORD2VEC	2.04 ± 1.08	0.02 ± 0.02	0.13 ± 0.11
WORD2VEC + ES	17.3 ± 0.05	1.00 ± 0.00	0.13 ± 0.11

Random: randomly draw from uniform distribution with unit variance
Word2vec: pre-trained word2vec embedding
ES: per-example standardization (per column)

[EMNLP'17]

Standardization Fixes the Variance Problems

- Standardization brings 8.7% absolute increase
- Transfer learning brings another 2.4% increase



[EMNLP'17]

Recap

- *"I want to build an NLI for my domain, but I don't have any training data"*
- Can I collect training data via crowdsourcing?
  - Yes, and it's not so expansive
  - Cost can be further reduced by crowdsourcing optimization
- Can I leverage existing training data from other domains?
  - Yes, if you turn it into a paraphrasing problem
  - Pre-trained word embedding can greatly help neural transfer learning, but only when properly standardized

FUTURE RESEARCH

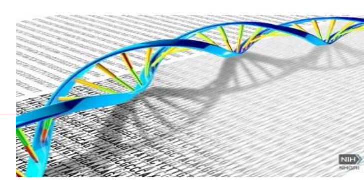
How can AI Bridge the Gap?

Insights
Discoveries
Solutions

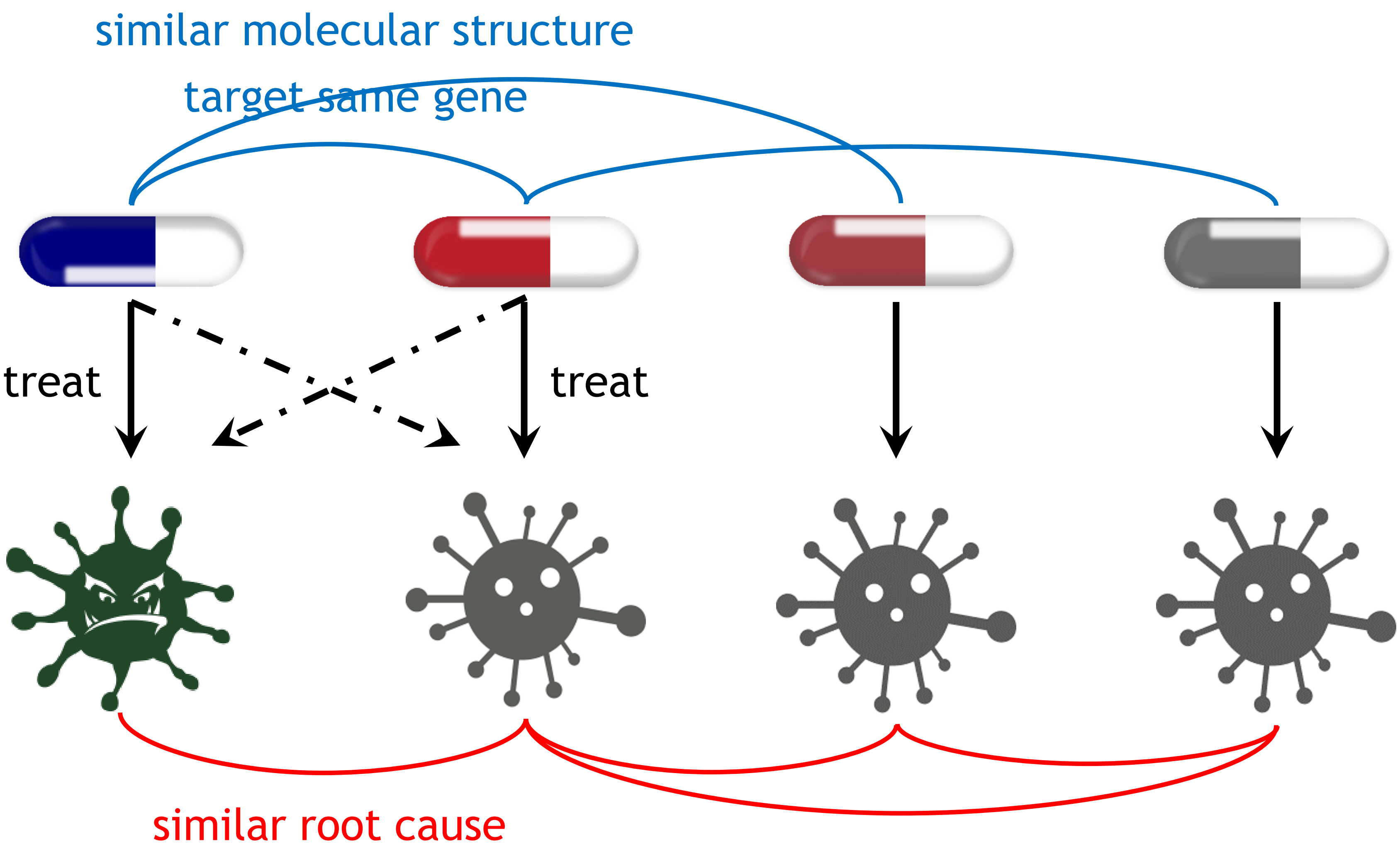
Bottleneck #2: Access

Bottleneck #3: Reasoning

Bottleneck #1: Knowledge



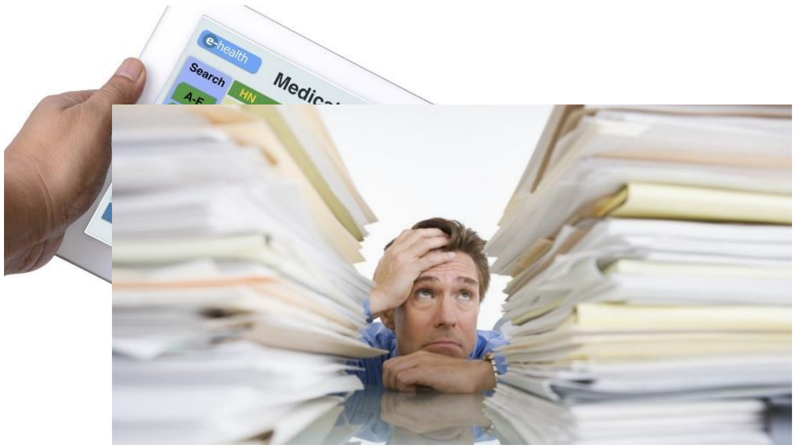
#3: Knowledge-based Machine Reasoning



Methodological Exploration

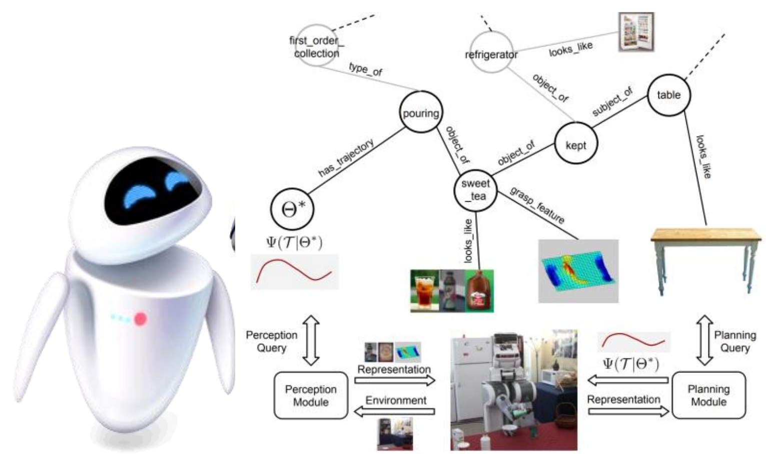
- Inherent structure of the NLI problem space
  - Strong prior for learning
  - Key: compositionality of natural & formal languages
- Integration of neural and symbolic computation
  - Neural network modularized over symbolic structures
  - (Cognitive science) neural encoding of symbolic structures
- Goal-oriented human-computer conversation
  - Accommodate dynamic hypothesis generation and verification in a natural conversation

End-to-end General-purpose Knowledge Engine



"Which cement stocks go up the most when a Category 3 hurricane hits Florida?"

KENSHO



Thanks &