Construction and Querying of Large-scale Knowledge Bases

Xiang Ren[1]   Yu Su[2]   Xifeng Yan[2]

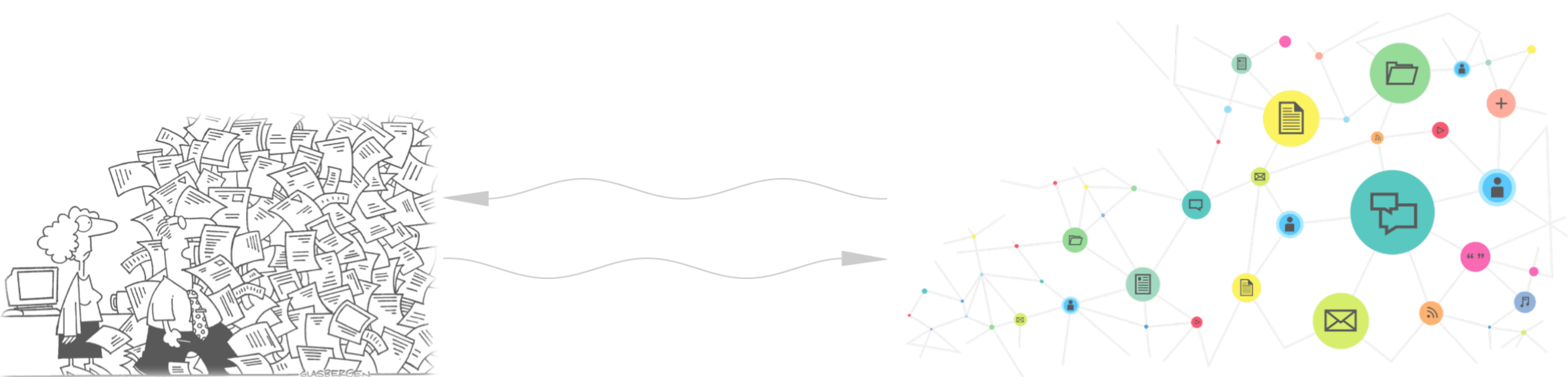
University of Southern California[1]

University of California, Santa Barbara[2]

Tutorial website:

http://xren7.web.engr.illinois.edu/tutorial-cikm17.html

Slides, code, datasets, references



Turning Unstructured Text Data into Structures

Unstructured Text Data

(account for ~80% of all data in organizations)

?

Structures

Knowledge & Insights

STORE

Store_key	City	Region
1	New York	**East**
2	Chicago	Central
3	Atlanta	**East**
4	Los Angeles	**West**
5	San Francisco	**West**
6	Philadelphia	**East**
.	.	.
.	.	.

PRODUCT

Product_key	Description	Brand
1	Beautiful Girls	MKF Studios
2	Toy Story	**Wolf**
3	Sense and Sensibility	Parabuster Inc.
4	Holiday of the Year	**Wolf**
5	Pulp Fiction	MKF Studios
6	The Juror	MKF Studios
7	From Dusk Till Dawn	Parabuster Inc.
8	Hellraiser: Bloodline	Big Studios
.	.	.
.	.	.

SALES_FACT

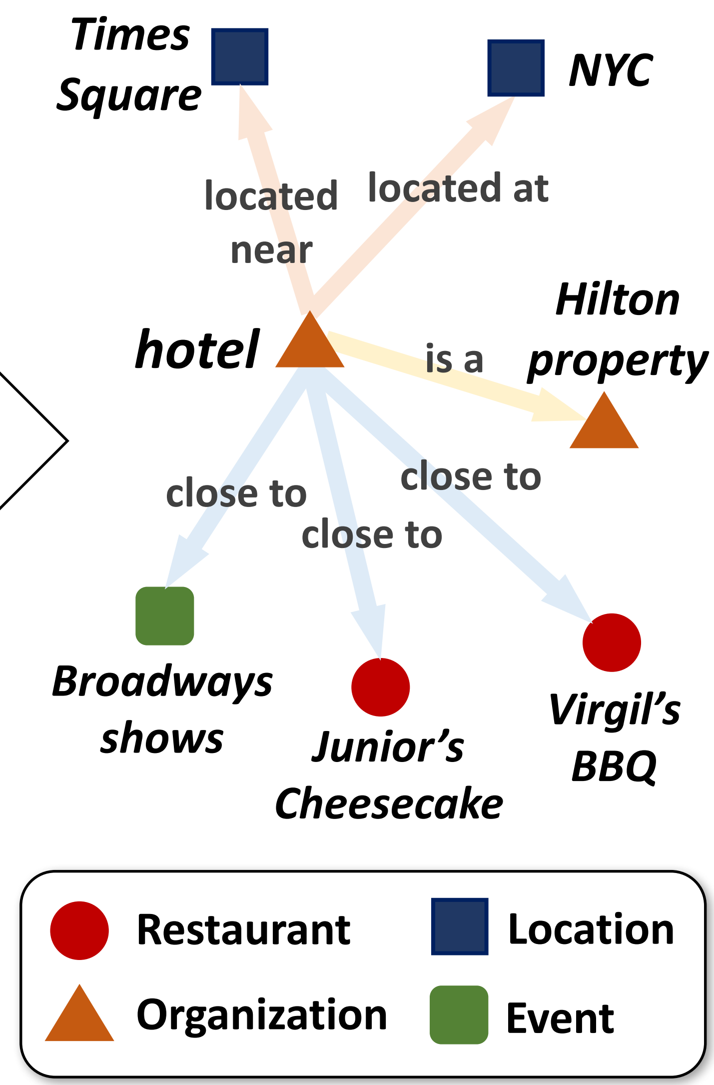
Store_key	Product_key	Sales	Cost	Profit
1	6	2.39	1.15	1.24
1	2	16.7	6.91	9.79
2	7	7.16	2.75	4.40
3	2	4.77	1.84	2.93
5	3	11.93	4.59	7.34
5	1	14.31	5.51	8.80
.	.	.	.	.
.	.	.	.	.

(Chakraborty, 2016)

Reading the reviews: From Text to Structured Facts

This **hotel** is my favorite **Hilton property** in **NYC**! It is located right on 42nd street near **Times Square**, it is close to all subways, **Broadways shows**, and next to great restaurants like **Junior's Cheesecake**, **Virgil's BBQ** and many others.

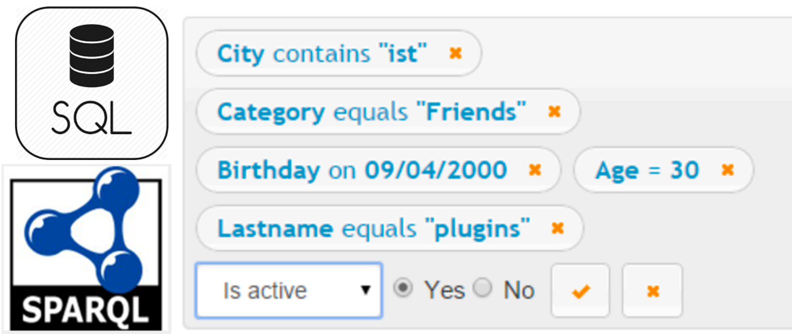
-- *TripAdvisor*



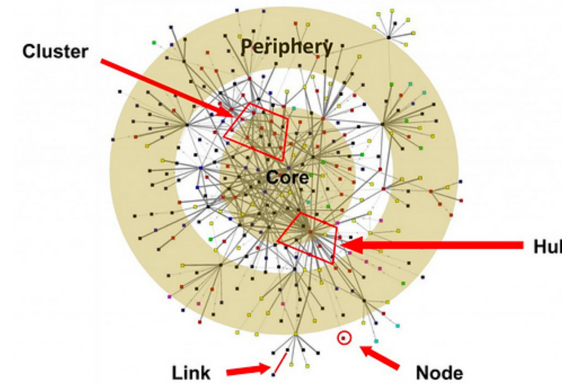
Structured Facts
1. "Typed" entities
2. "Typed" relationships

Why Text to Structures?

Structured Search & Exploration

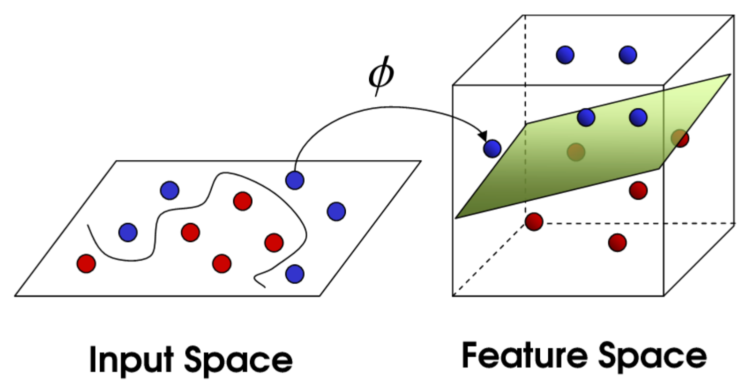


Graph Mining & Network Analysis



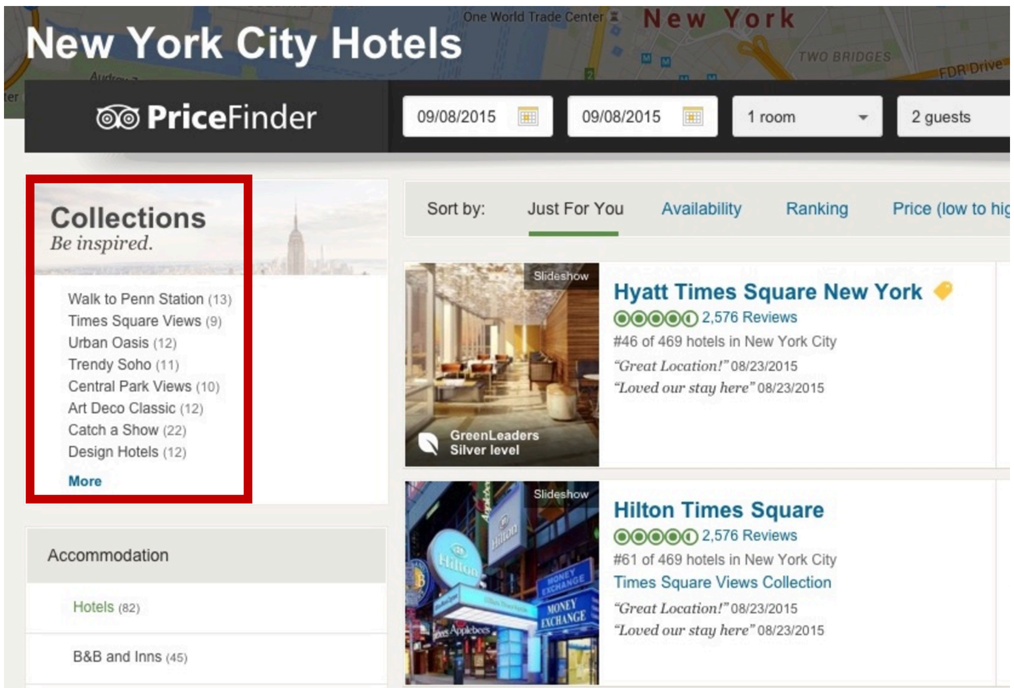
Pattern / Association Rule Mining

Structured Feature Generation



A Product Use Case: Finding “Interesting Hotel Collections”

Technology Transfer to TripAdvisor



Grouping hotels based on structured facts extracted from the review text

Features for “**Catch a Show**” collection

1	broadway shows
2	beacon theater
3	broadway dance center
4	broadway plays
5	david letterman show
6	radio city music hall
7	theatre shows

Features for “**Near The High Line**” collection

1	high line park
2	chelsea market
3	highline walkway
4	elevated park
5	meatpacking district
6	west side
7	old railway

http://engineering.tripadvisor.com/using-nlp-to-find-interesting-collections-of-hotels/

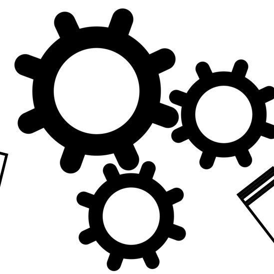
Prior Art: Extracting Structures with Repeated Human Effort

Human labeling

... We had a room facing *Times Square* and a room facing the *Empire State Building*, The location is close to everything and we love ...

Labeled data

Extraction Rules
Machine-Learning Models



This hotel is my favorite Hilton property in NYC! It is located right on 42nd street near Times Square, it is close to all subways, Broadways shows, and next to many great ...

Text Corpus

Structured Facts

Broadways shows

NYC

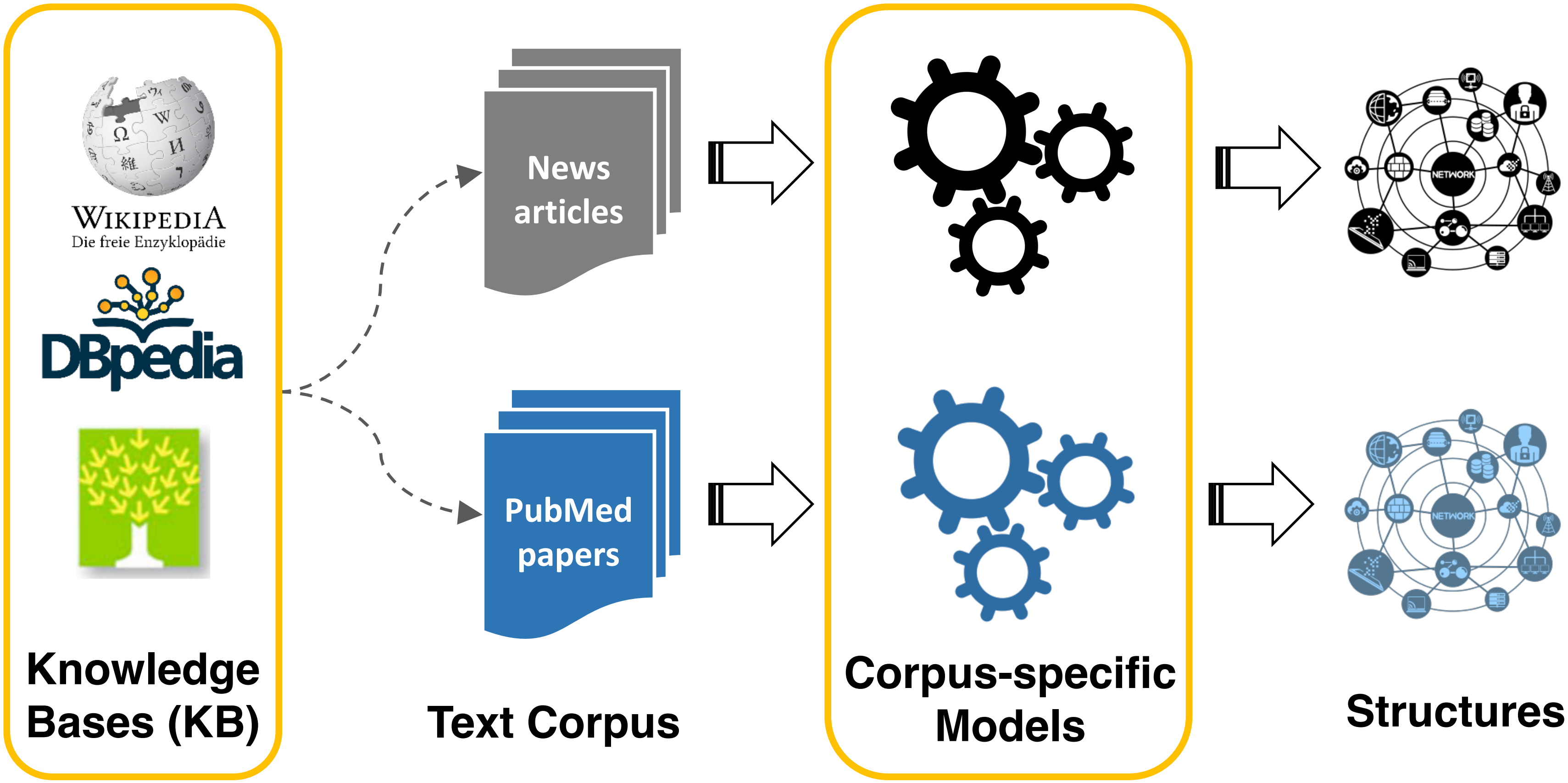
Times square

hotel

Hilton property

Stanford CoreNLP
CMU NELL
UW KnowItAll
USC AMR
IBM Alchemy APIs
Google Knowledge Graph
Microsoft Satori
...

This Tutorial: Effort-Light StructMine



- Enables ***quick*** development of applications over various corpora
- Extracts ***complex*** structures without introducing human error

Effort–Light StructMine: **Where Are We?**

Human labeling effort

Hand-crafted Systems

UCB Hearst Pattern, 1992
NYU Proteus, 1997

Supervised learning systems

Stanford CoreNLP, 2005 - present
UT Austin Dependency Kernel, 2005
IBM Watson Language APIs

Weakly-supervised learning systems

CMU NELL, 2009 - present
UW KnowItAll, Open IE, 2005 - present
Max-Planck YAGO, 2008 - present

Distantly-supervised Learning Systems

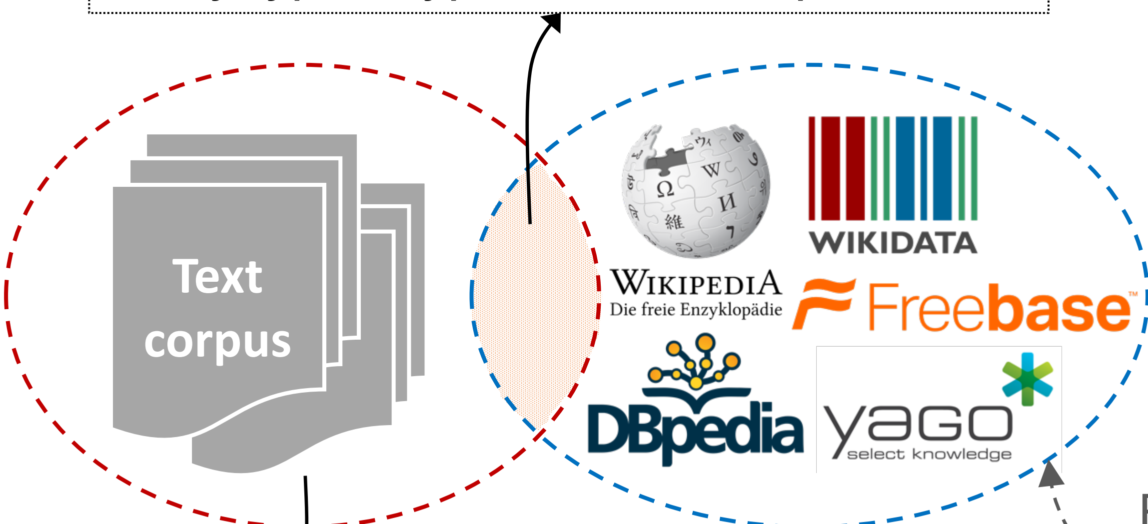
Stanford DeepDive, MIML-RE 2012 - present
UW FIGER, MultiR, 2012
Effort-Light StructMine
(WWW'15, KDD'15, KDD'16, EMNLP'16, WWW'17, …)

Feature engineering effort

A Review of Previous Efforts

“Distant” Supervision: **What Is It?**

“Matchable” structures: entity names, entity types, typed relationships ...

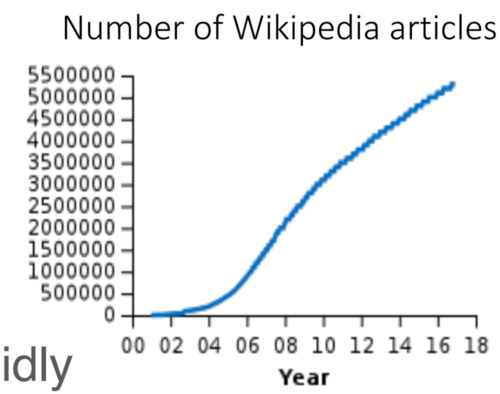


“Un-matchable”

Knowledge Bases

Freely available!

- Common knowledge
- Life sciences
- Art …



Rapidly growing!

Human crowds

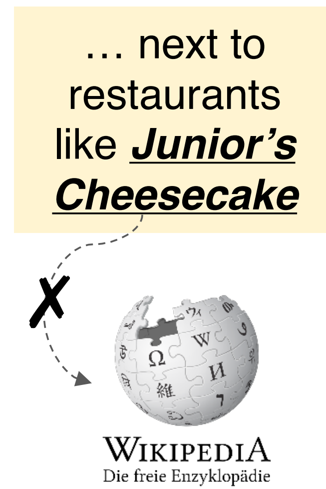
(Mintz et al., 2009) , (Riedek et al., 2010), (Lin et al., 2012), (Ling et al., 2012), (Surdeanu et al., 2012), (Xu et al., 2013), (Nagesh et al., 2014), …

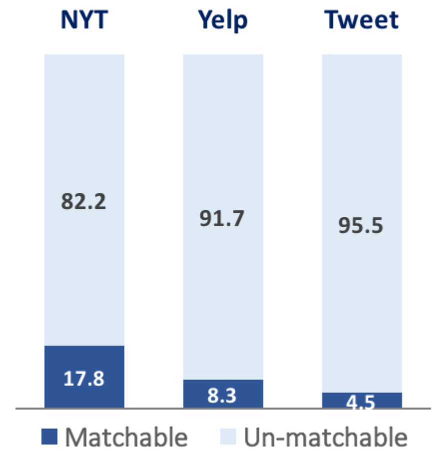
https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

Learning with Distant Supervision: **Challenges**

1. Sparsity of "Matchable"

- Incomplete knowledge bases
- Low-confidence matching

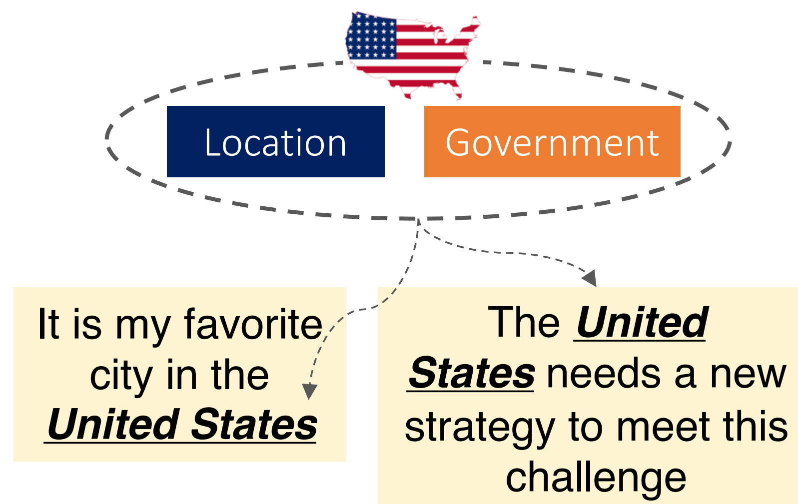




(Ren et al., KDD'15)

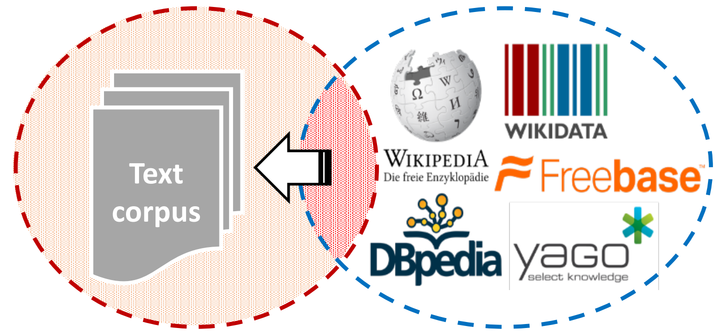
2. Accuracy of "Expansion"

- For "matchable": *Are all the labels assigned accurately?*
- For "un-matchable": *How to perform inference accurately?*

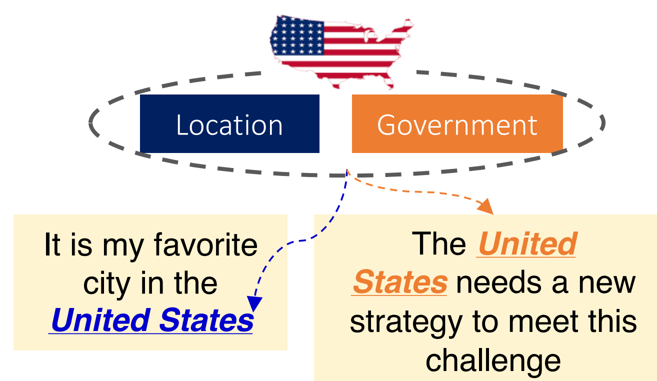


Effort-Light StructMine: **Contributions**

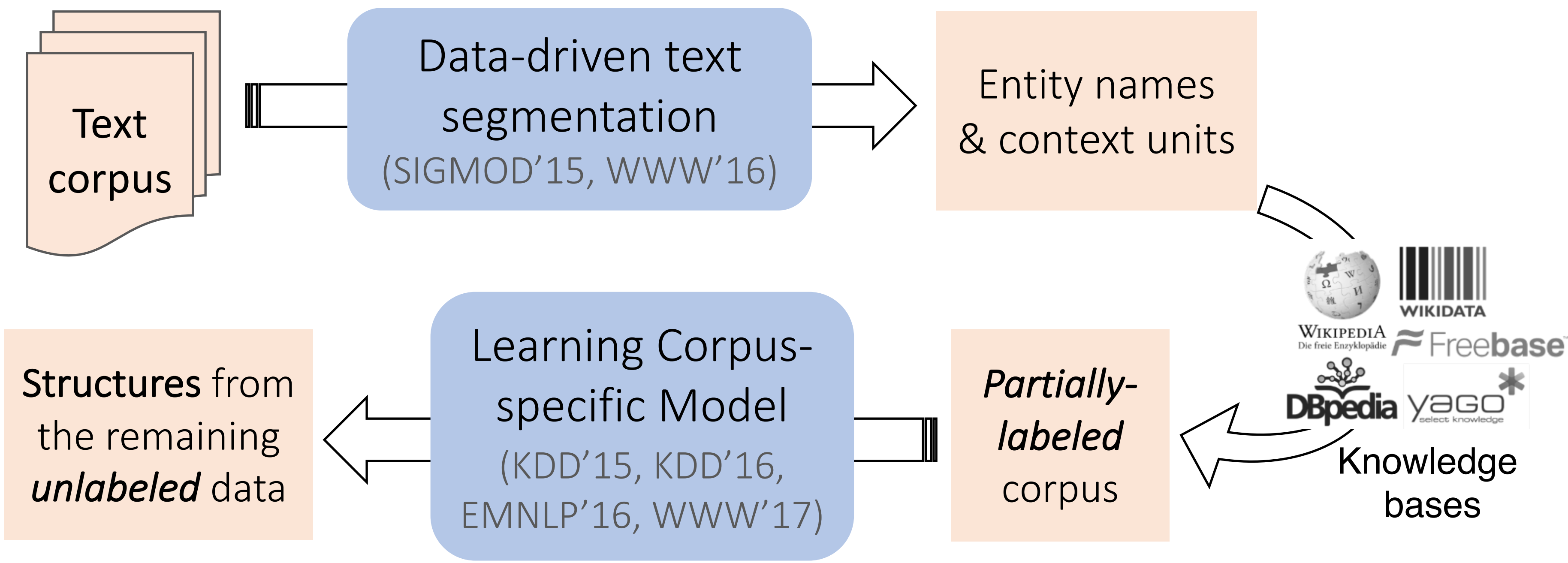
Challenge	Solution Idea
Sparsity of "Matchable"	Effective expansion from "*matchable*" to "*un-matchable*"
Accuracy of "Expansion"	Pick the "*best*" labels based on the context (for both "matchable" and "un-matchable")



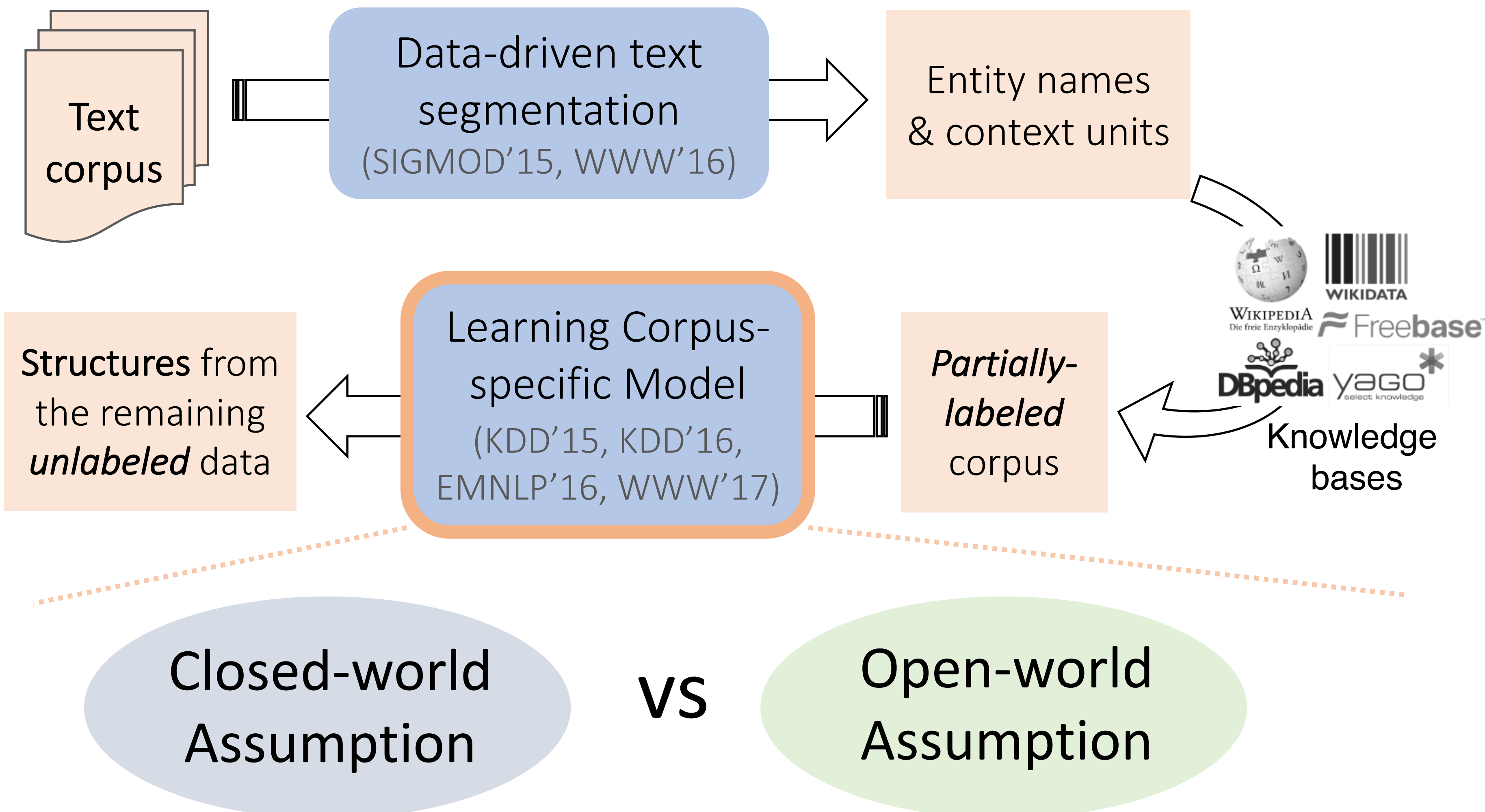
Harness the "**data redundancy**" using graph-based *joint* optimization



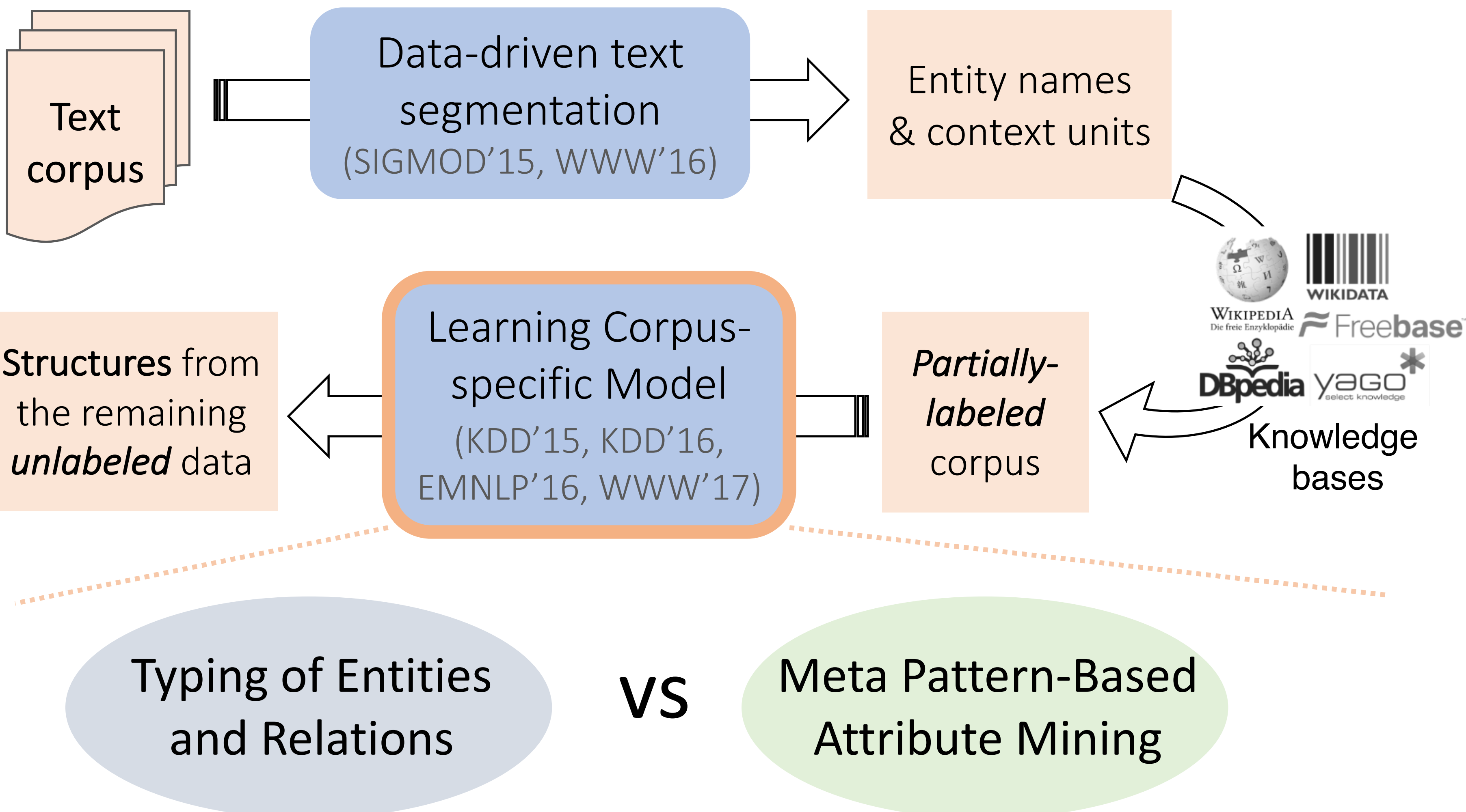
Effort-Light StructMine: **Methodology**



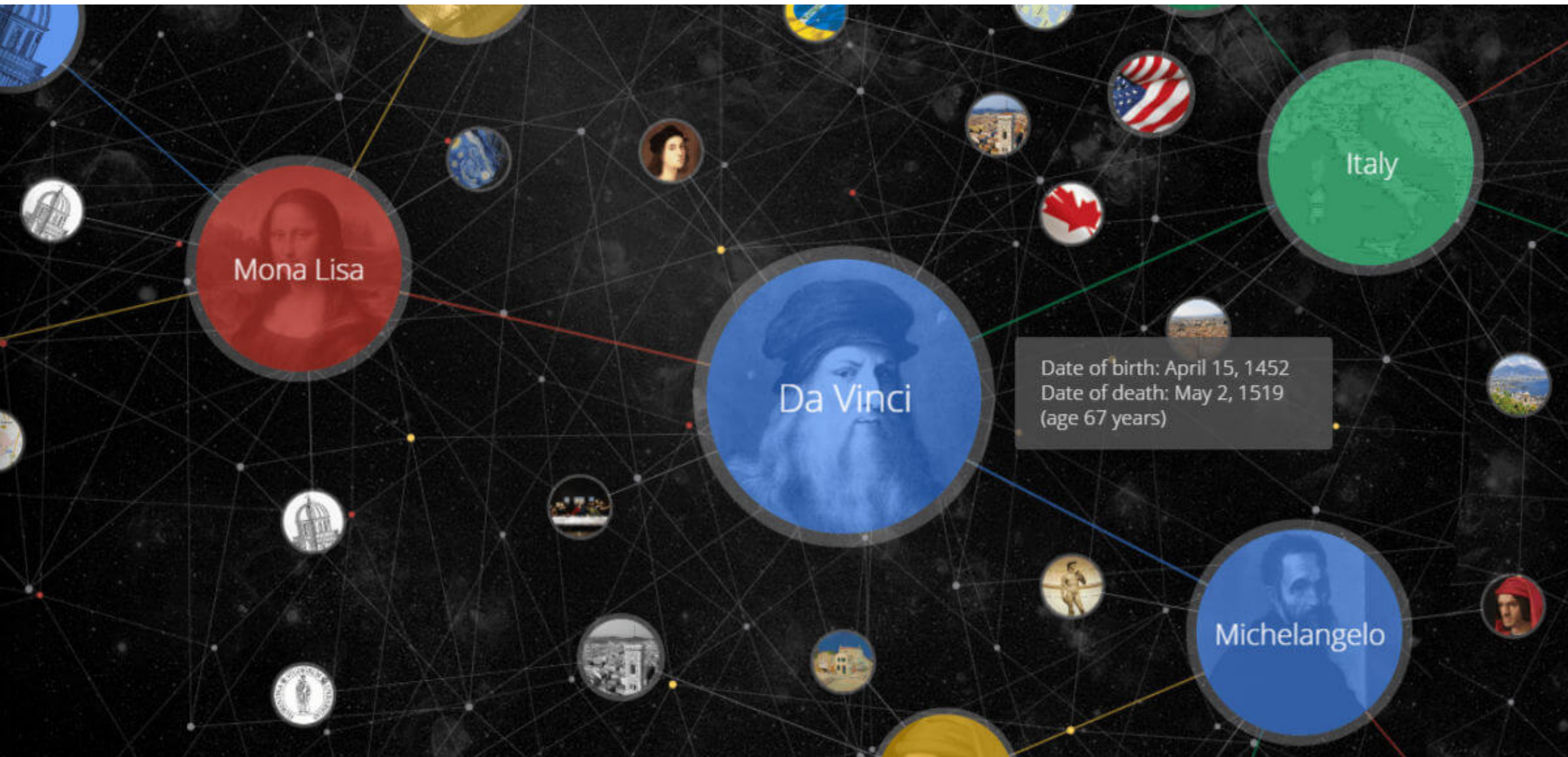
Effort-Light StructMine: **Methodology**



Effort-Light StructMine: **Methodology**



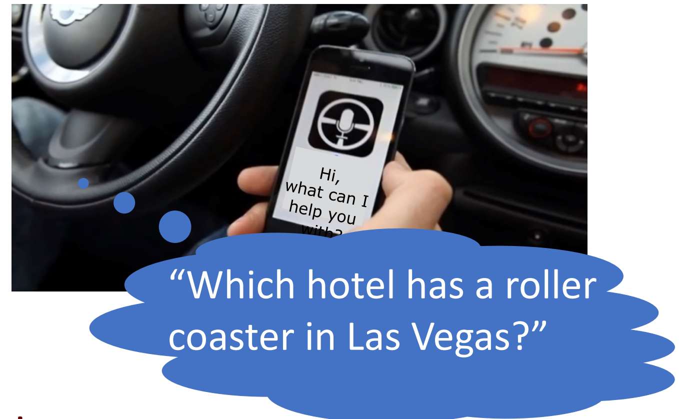
Knowledge Base Querying



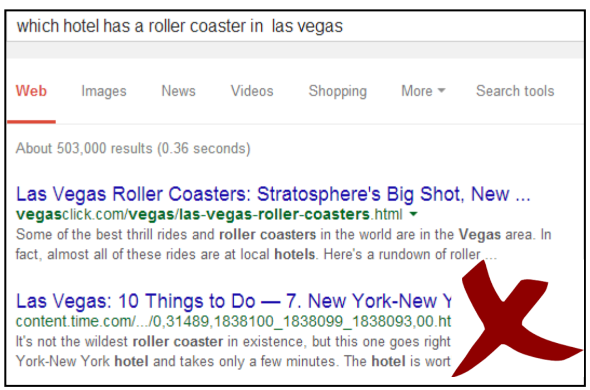
Transformation in Information Search

Desktop search

Mobile search



Lengthy Documents? Direct Answers!



which hotel has a roller coaster in las vegas

Web Images News Videos Shopping More Search tools

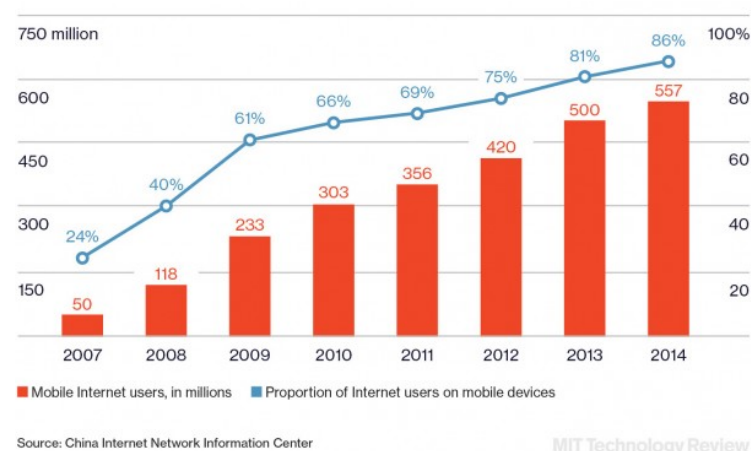
About 503,000 results (0.36 seconds)

Las Vegas Roller Coasters: Stratosphere's Big Shot, New ...
vegasclick.com/**vegas**/**las-vegas-roller-coasters**.html
Some of the best thrill rides and **roller coasters** in the world are in the **Vegas** area. In fact, almost all of these rides are at local **hotels**. Here's a rundown of roller ...

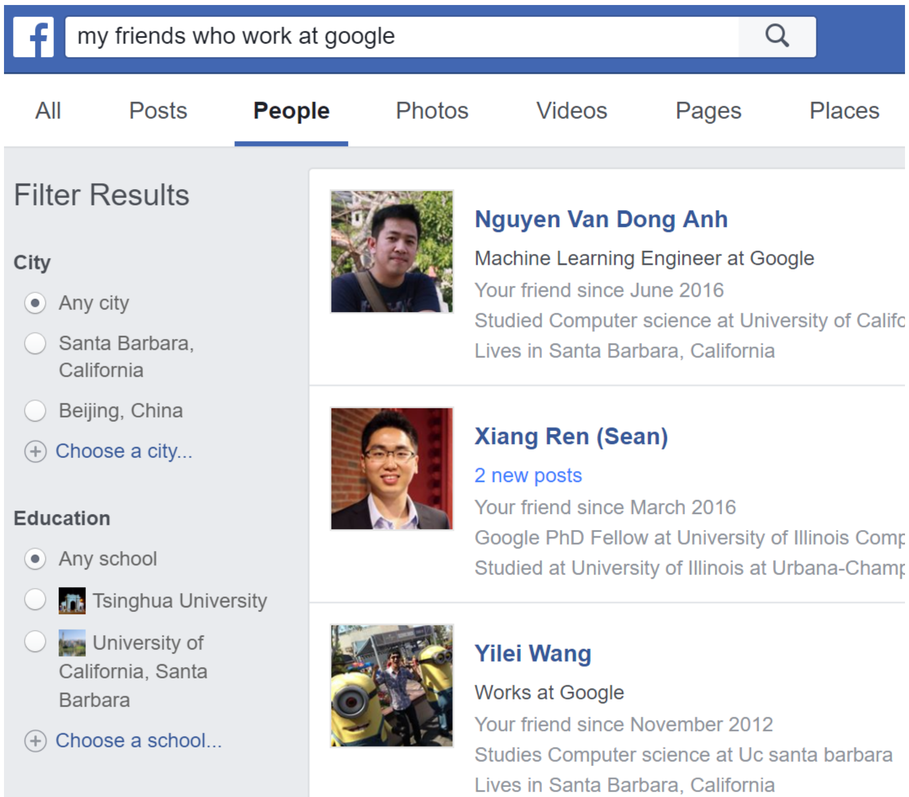
Las Vegas: 10 Things to Do — 7. New York-New Y
content.time.com/.../0,31489,1838100_1838099_1838093,00.ht
It's not the wildest **roller coaster** in existence, but this one goes right
York-New York **hotel** and takes only a few minutes. The **hotel** is wort

Answer: New York-New York hotel

Surge of mobile Internet use in China



Application: Facebook Entity Graph



People, Places, and Things

Facebook's knowledge graph (entity graph) stores as entities the users, places, pages and other objects within the Facebook.

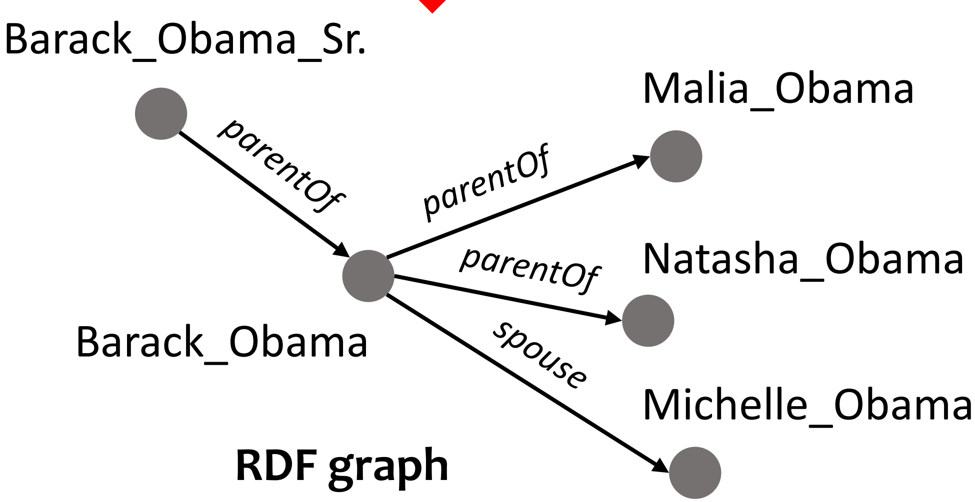
Connecting

The connections between the entities indicate the type of relationship between them, such as friend, following, photo, check-in, etc.

Structured Query: RDF + SPARQL

Triples in an RDF graph

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 massive, loosely-defined schema

Certainly, You Do Not Want to Write This!

"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)
```

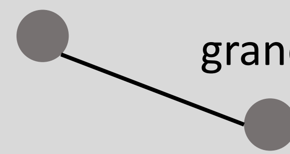
NCI thesaurus

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

Schema-agnostic KB Querying

“Barack Obama Sr. grandchildren”

Keyword query: query like search engine



Graph query: add a little structure

“Who are Barack Obama Sr.’s grandchildren?”

Natural language query: like asking a friend

<Barack Obama Sr., Malia Obama>

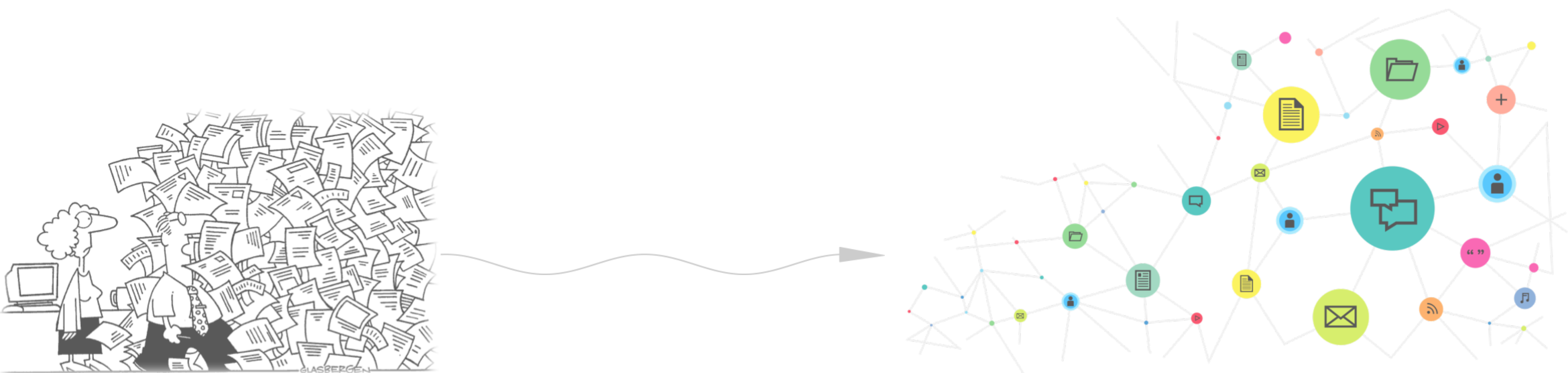
Query by example: Just show me examples

Tutorial Outline

- Introduction
- Part I: Effort–Light StructMine
  - Tea break at 3:00pm
- Part II: Schema-agnostic KB Querying
- Summary & Future Directions

Construction and Querying of Large-scale Knowledge Bases

Part I: Effort-Light StructMine for Knowledge Base Construction



Effort-Light StructMine: **Overview**

Text corpus

Data-driven text segmentation
(SIGMOD'15, WWW'16)

Entity names & context units

WIKIPEDIA
Die freie Enzyklopädie

WIKIDATA

Freebase

DBpedia

yago
select knowledge

Knowledge bases

Partially-labeled corpus

Learning Corpus-specific Model
(KDD'15, KDD'16, EMNLP'16, WWW'17)

Structures from the remaining *unlabeled* data

Entity Recognition and Coarse-grained Typing (KDD'15)

Fine-grained Entity Typing (KDD'16)

Joint Entity and Relation Extraction (WWW'17)

Corpus to Structured Network: The Roadmap

Corpus to Structured Network: **The Roadmap**

Text corpus

Data-driven text segmentation
(SIGMOD'15, WWW'16)

entity names & context units

Knowledge bases

Partially-labeled corpus

Learning Corpus-specific Model
(KDD'15, KDD'16, EMNLP'16, WWW'17)

Structures from the remaining *unlabeled* data

Entity Recognition and Coarse-grained Typing (KDD'15)	Fine-grained Entity Typing (KDD'16)	Joint Entity and Relation Extraction (WWW'17)

Recognizing Entities of Target Types in Text

The best BBQ I've tasted in Phoenix! I had the pulled pork sandwich with coleslaw and baked beans for lunch. The owner is very nice. …

The best ***BBQ*** I've tasted in ***Phoenix*** ! I had the ***pulled pork sandwich*** with ***coleslaw*** and ***baked beans*** for lunch. The **owner** is very nice. …

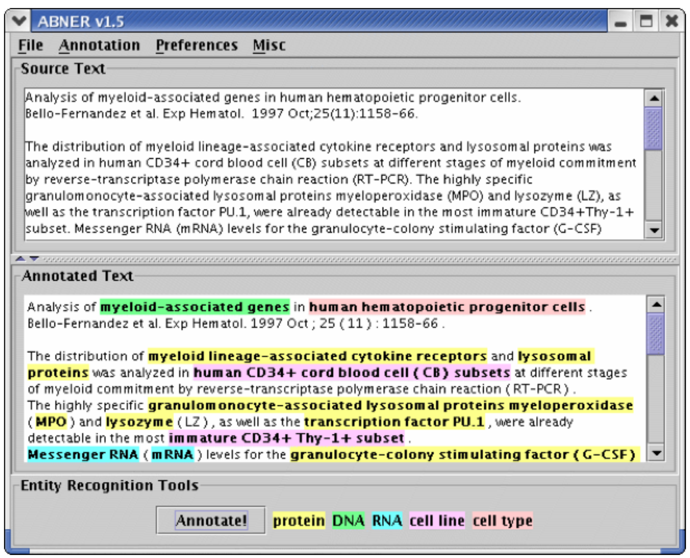
food

location

person

Traditional Named Entity Recognition (NER) Systems

- Heavy reliance on corpus-specific human labeling
- Training sequence models is slow



A manual annotation interface

The	best	BBQ	I've	tasted	in	Phoenix
O	O	Food	O	O	O	Location

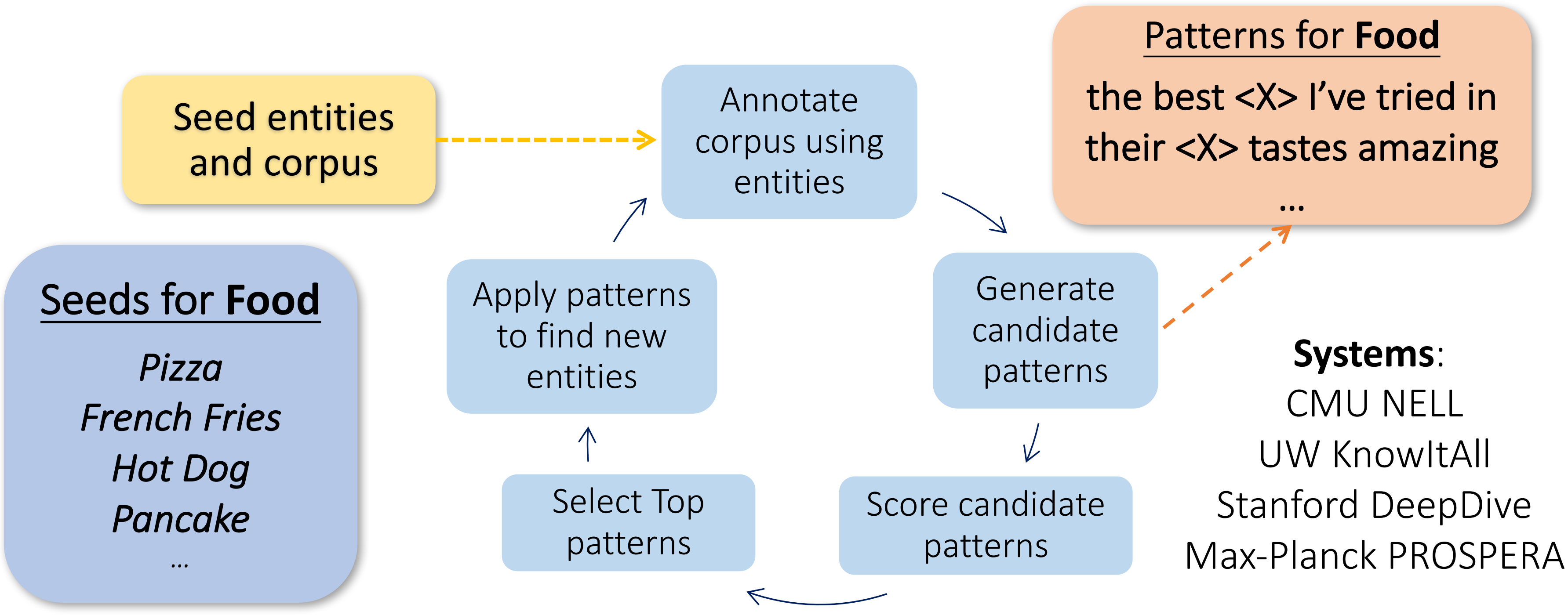
Sequence model training

NER Systems:
Stanford NER
Illinois Name Tagger
IBM Alchemy APIs
…

e.g., (McMallum & Li, 2003), (Finkel et al.,2005), (Ratinov & Roth, 2009), …

Weak-Supervision Systems: Pattern-Based Bootstrapping

- Requires manual seed selection & mid-point checking

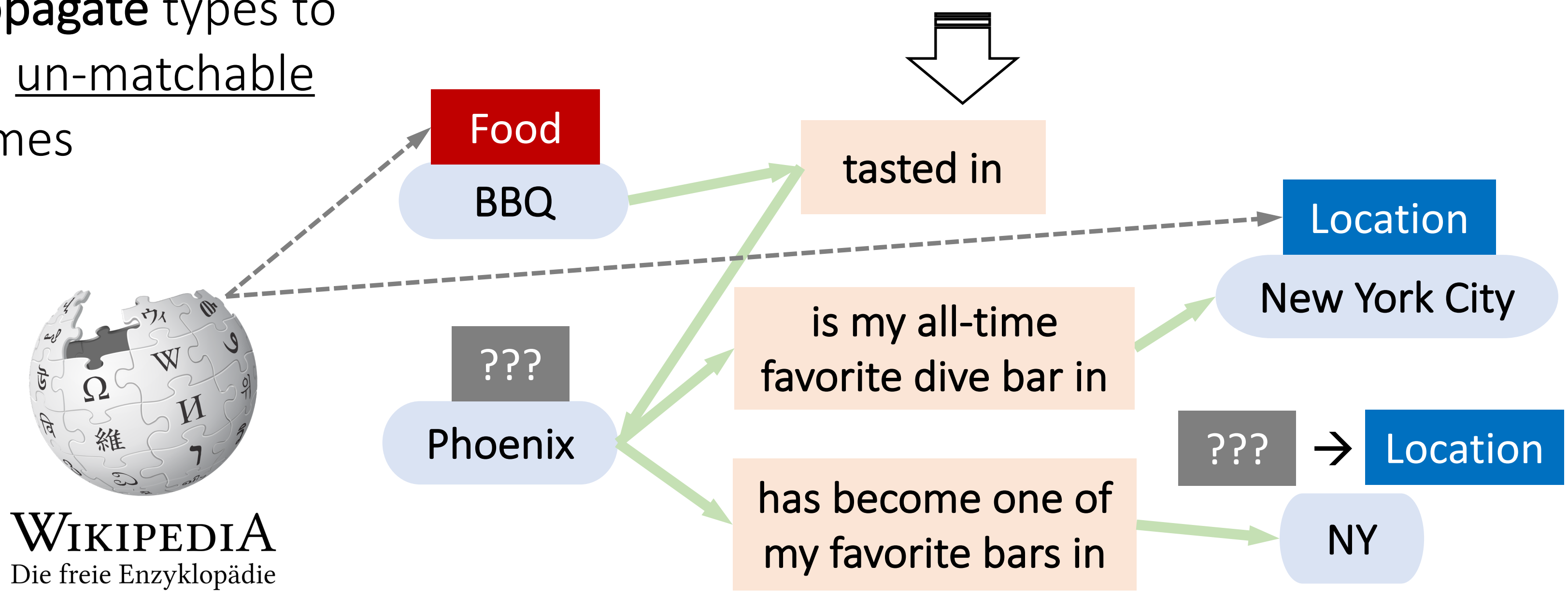


e.g., (Etzioni et al., 2005), (Talukdar et al., 2010), (Gupta et al., 2014), (Mitchell et al., 2015), ...

Leveraging Distant Supervision

1. **Detect** entity names from text
2. **Match** name strings to KB entities
3. **Propagate** types to the un-matchable names

ID	Sentence
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S2	The best *BBQ* I've tasted in *Phoenix.*
S3	*Phoenix* has become one of my favorite bars in *NY* .



(Lin et al., 2012), (Ling et al., 2012), (Nakashole et al., 2013)

Current Distant Supervision: Limitation I

1. Context-agnostic type prediction
    - Predict types for each mention regardless of context
2. Sparsity of contextual bridges

ID	Sentence
S1	*Phoenix* is my all-time favorite dive bar in ***New York City*** .
S2	The best ***BBQ*** I've tasted in ***Phoenix***.
S3	*Phoenix* has become one of my favorite bars in ***NY*** .

Current Distant Supervision: Limitation II

1. Context-agnostic type prediction
2. Sparsity of contextual bridges
   - Some **relational phrases** are **infrequent** in the corpus → ineffective type propagation

ID	Sentence
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S3	*Phoenix* has become one of my favorite bars in *NY* .

ClusType: Data-Driven Entity Mention Detection

- **Significance** of a merging between two sub-phrases

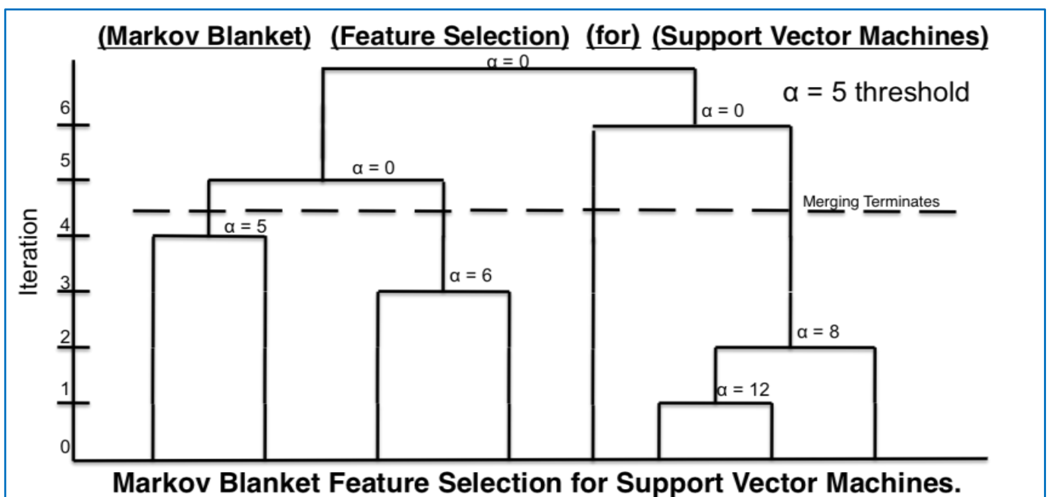
Quality of merging

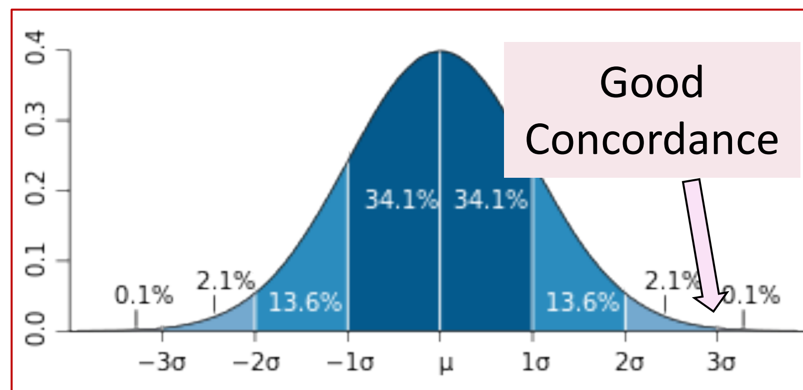
Corpus-level Concordance

Syntactic quality

$$\rho_X(S_1, S_2) = \boxed{\frac{v(S_1 \oplus S_2) - N \frac{v(S_1)}{N} \frac{v(S_2)}{N}}{\sqrt{v(S_1 \oplus S_2)}}} \cdot \boxed{I_X(S_1 \oplus S_2)}$$

Pattern	Example
(J*)N*	support vector machine
VP	tasted in, damage on
VW*(P)	train a classifier with





ClusType: Data-Driven Entity Mention Detection

- **Significance** of a merging between two sub-phrases

Quality of merging

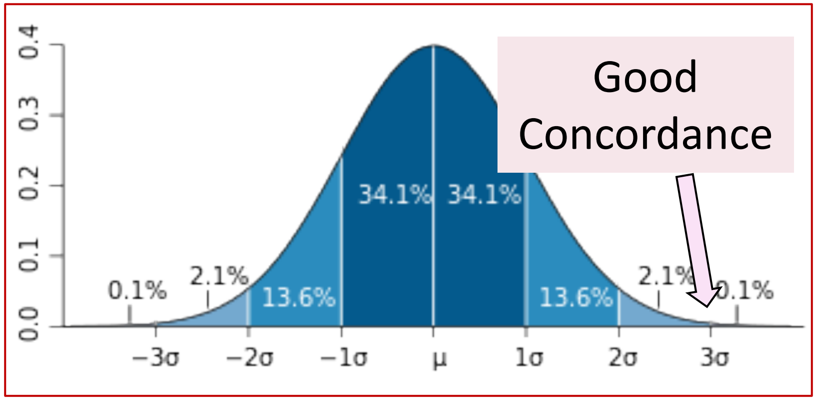
Corpus-level Concordance

Syntactic quality

$$\rho_X(S_1, S_2) = \boxed{\frac{v(S_1 \oplus S_2) - N\frac{v(S_1)}{N}\frac{v(S_2)}{N}}{\sqrt{v(S_1 \oplus S_2)}}} \cdot \boxed{I_X(S_1 \oplus S_2)}$$

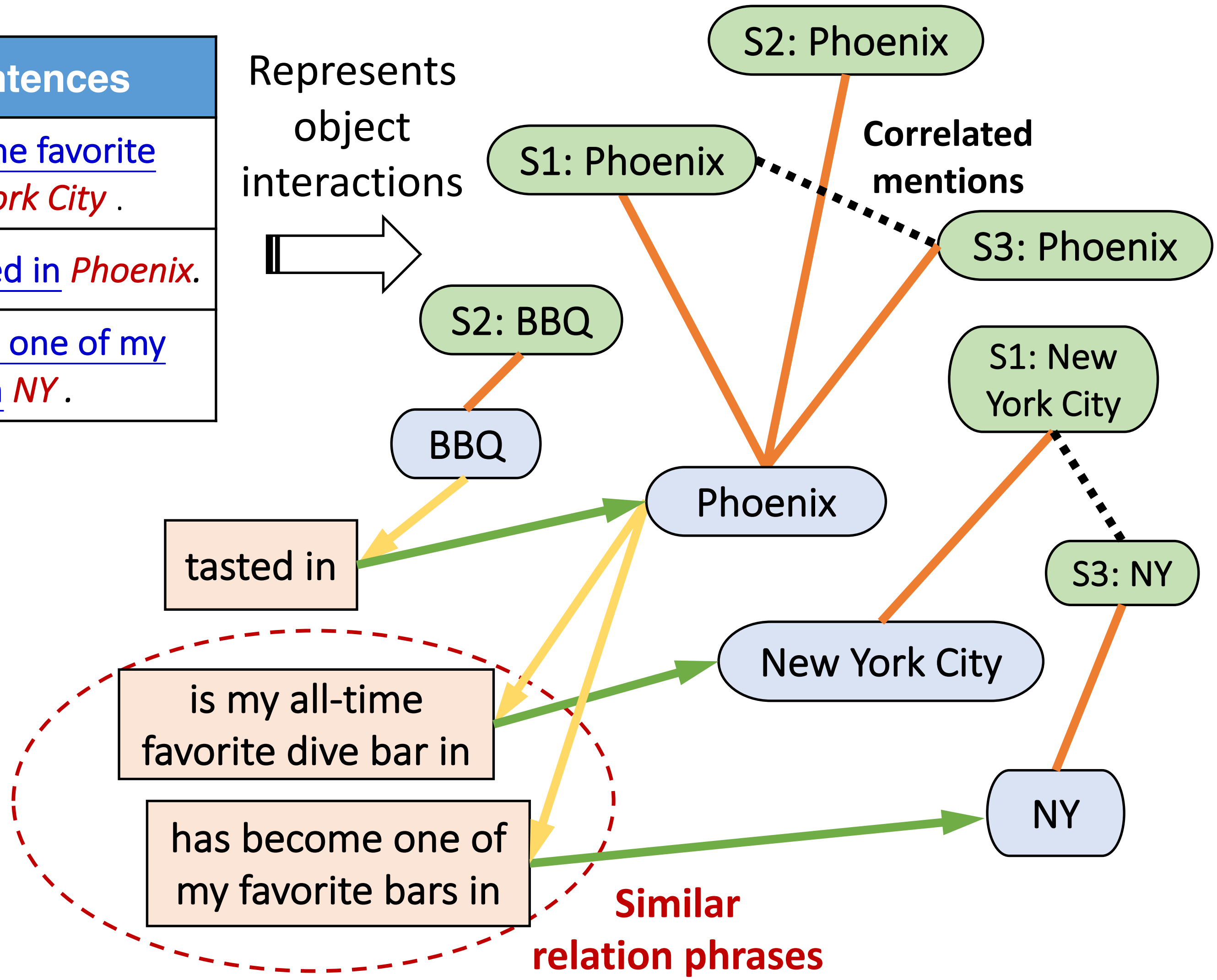
Pattern	Example
(J*)N*	support vector machine
VP	tasted in, damage on
VW*(P)	train a classifier with

The best *BBQ* I've tasted in *Phoenix* ! I had the *pulled pork sandwich* with *coleslaw* and *baked beans* for lunch. … This *place* serves up the best *cheese steak sandwich* in west of *Mississippi*.



My Solution: **ClusType** (KDD'15)

ID	Segmented Sentences
S1	*Phoenix* is my all-time favorite dive bar in *New York City* .
S2	The best *BBQ* I've tasted in *Phoenix*.
S3	*Phoenix* has become one of my favorite bars in *NY* .



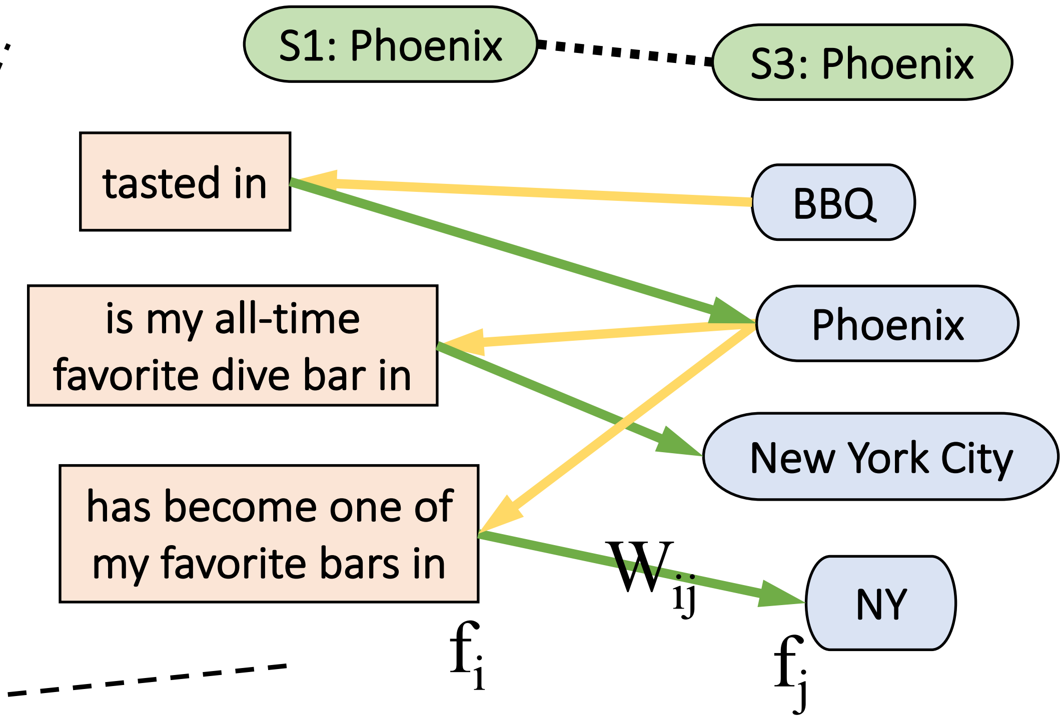
Putting two sub-tasks together:

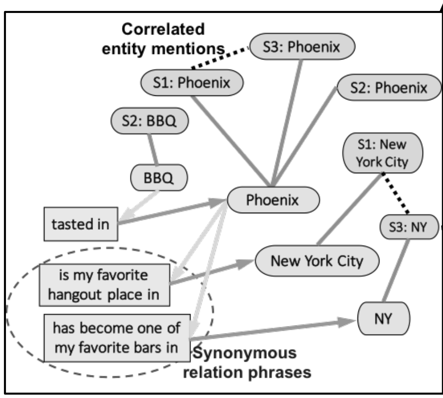
1. Type label propagation
2. Relation phrase clustering

Type Propagation in **ClusType**

Smoothness Assumption

If two objects are similar according to the graph, then their type labels should be also similar





Edge weight / object similarity

Measure of Non-Smoothness

Vector of scores for single label on nodes

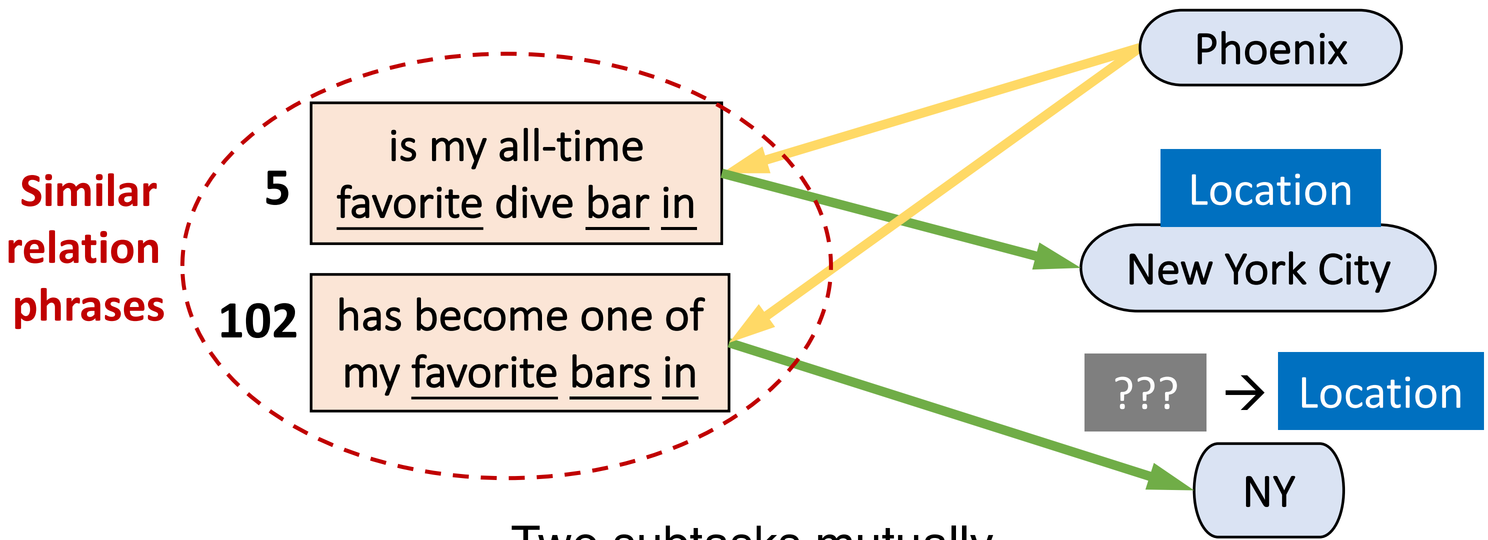
$$f^T L f = \sum_{i,j} W_{ij}(f_i - f_j)^2$$

(Belkin & Partha, NIPS'01), (Ren et al., KDD'15)

Relation Phrase Clustering in **ClusType**

- Two relation phrases should be grouped together if:
  1. Similar string
  2. Similar context
  3. Similar types for entity arguments

"Multi-view" clustering



Two subtasks mutually enhance each other

(Ren et al., KDD'15)

ClusType: Comparing with State-of-the-Art Systems (F1 Score)

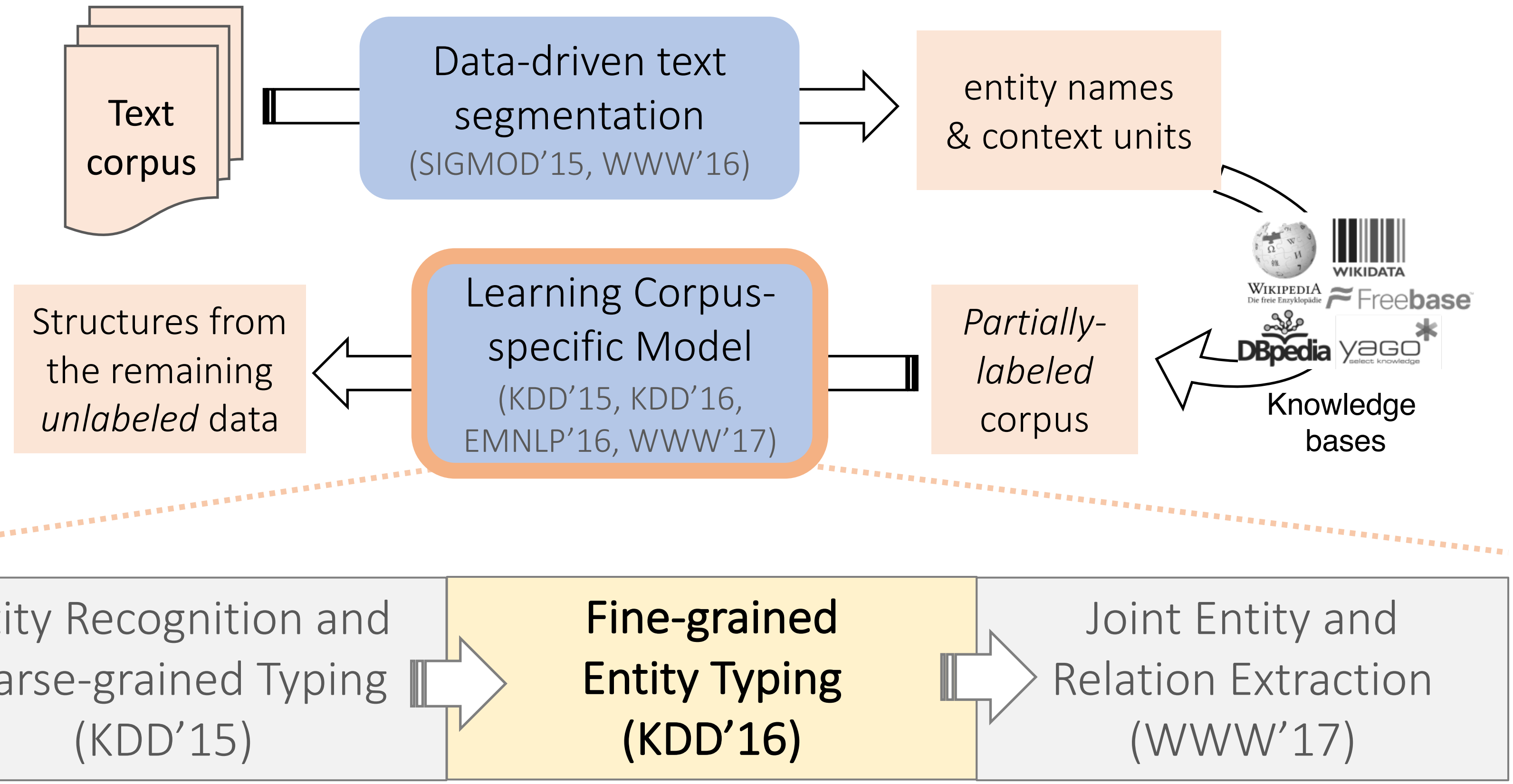
	Methods	NYT	Yelp	Tweet
Bootstrapping	**Pattern** (Stanford, CONLL'14)	0.301	0.199	0.223
	SemTagger (U Utah, ACL'10)	0.407	0.296	0.236
Label propagation	**NNPLB** (UW, EMNLP'12)	0.637	0.511	0.246
	APOLLO (THU, CIKM'12)	0.795	0.283	0.188
Classifier with linguistic features	**FIGER** (UW, AAAI'12)	0.881	0.198	0.308
	ClusType (KDD'15)	**0.939**	**0.808**	**0.451**

- vs. **bootstrapping**: context-aware prediction on "un-matchable"
- **vs. label propagation**: group similar relation phrases
- **vs. FIGER**: no reliance on complex feature engineering

NYT: 118k news articles (1k manually labeled for evaluation); **Yelp**: 230k business reviews (2.5k reviews are manually labeled for evaluation); **Tweet**: 302 tweets (3k tweets are manually labeled for evaluation)

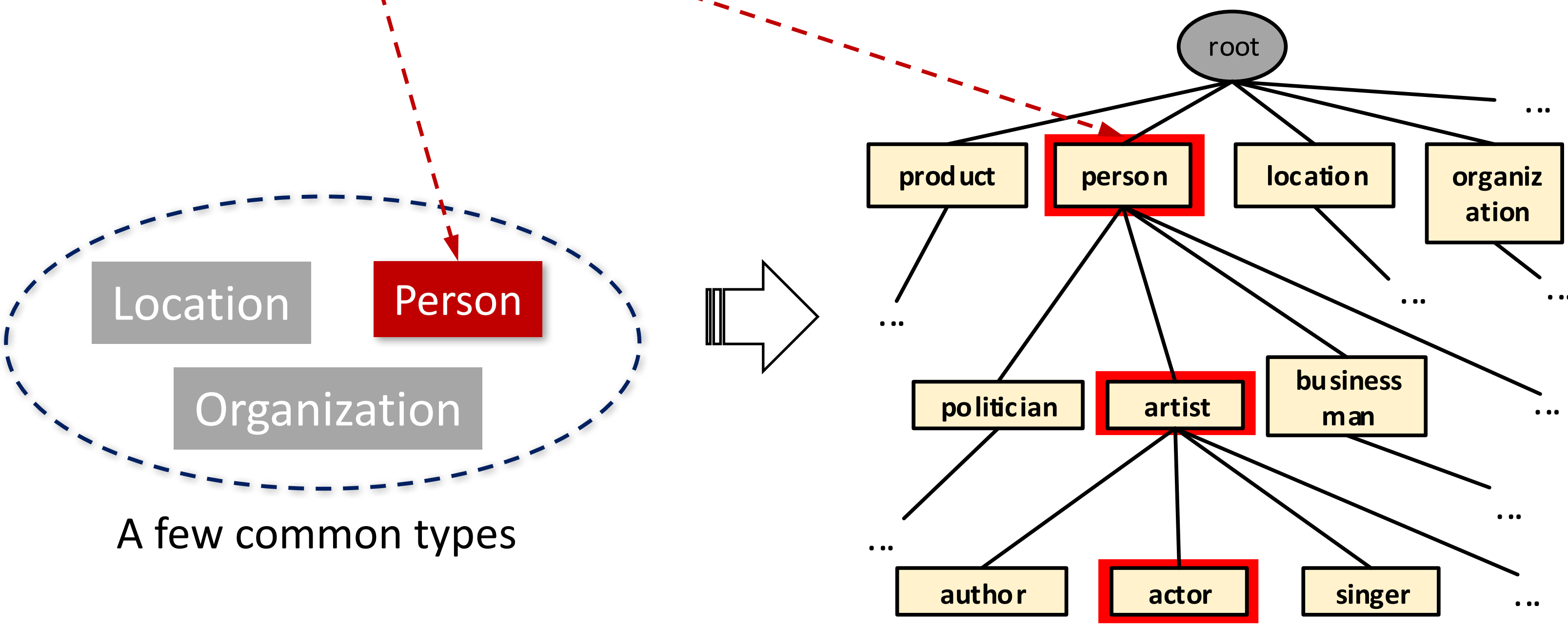
Precision (P) = $\frac{\#Correctly\text{−}typed\ mentions}{\#System\text{−}recognized\ mentions}$, Recall ($R$) = $\frac{\#Correctly\text{−}typed\ mentions}{\#ground\text{−}truth\ mentions}$, F1 score = $\frac{2(P \times R)}{(P+R)}$

Corpus to Structured Network: **The Roadmap**



From Coarse-Grained Typing to **Fine-Grained Entity Typing**

ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a game show, NBC's The Apprentice.



A few common types

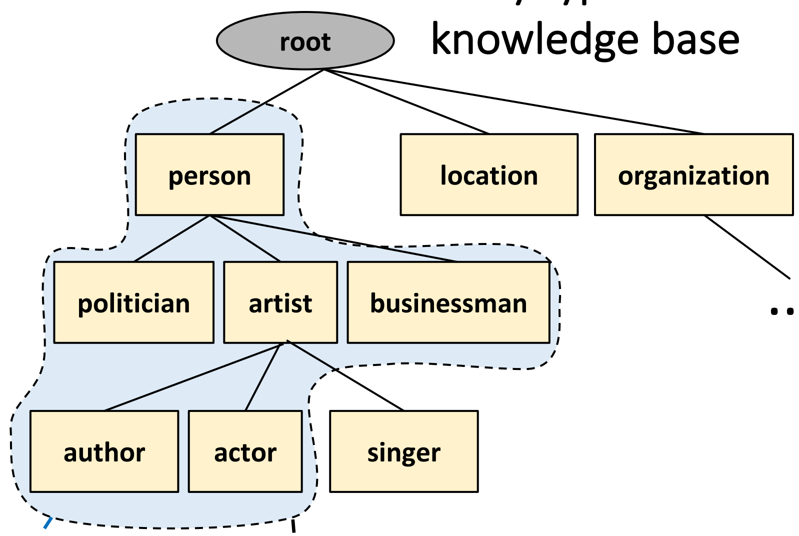
A type hierarchy with 100+ types (from knowledge base)

(Ling et al., 2012), (Nakashole et al., 2013), (Yogatama et al., 2015)

Current Distant Supervision: **Context-Agnostic Labeling**

- Inaccurate labels in **training data**
- **Prior work**: <u>all</u> labels are "<u>perfect</u>"

Entity types from knowledge base



ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a game show, NBC's The Apprentice

S1: *Donald Trump*
Entity Types: person, artist, actor, ~~author~~, ~~businessman~~, ~~politician~~

Entity:
Donald Trump

My Solution: **P**artial **L**abel **E**mbedding (KDD'16)

Extract Text Features

↓

"Label Noise Reduction" with **PLE**

↓

Train Classifiers on De-noised Data

↓

Prediction on New Data

ID	Sentence
S1	*Donald Trump* spent 14 television seasons presiding over a game show, NBC's The Apprentice

Text features: TOKEN_Donald, CONTEXT: television, CONTEXT: season, TOKEN_trump, SHAPE: AA

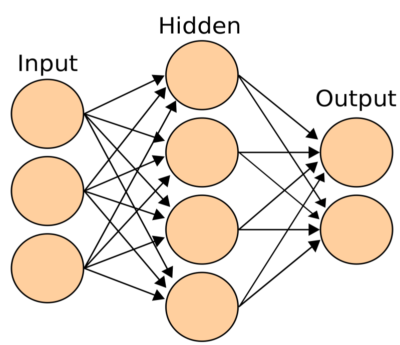
⇩

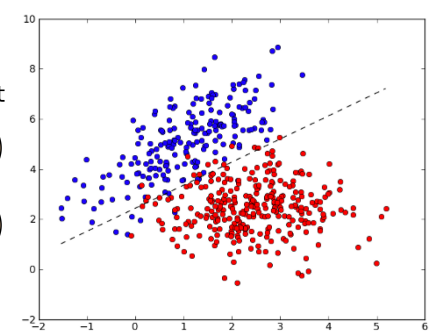
S1: *Donald Trump*

Entity Types: person, artist, actor, ~~author~~, ~~businessman~~, ~~politician~~

"De-noised" labeled data

⇩





More effective classifiers

(Ren et al., KDD'16)

PLE: Modeling Clean and Noisy Mentions Separately

For a clean mention, its "*positive types*" should be **ranked higher** than all its "*negative types*"

S_i: *Ted Cruz*
Types in KB: person, politician

ID	Noisy Entity Mention
S1	***Donald Trump*** spent 14 television seasons presiding over a game show, NBC's The Apprentice

S1: *Donald Trump*
Types in KB: person, artist, actor, author, businessman, politician

Types ranked

(+) actor
(-) singer
(-) coach
(-) doctor
(-) location
(-) organization

"**Best**" candidate type

(+) actor 0.88
(+) artist 0.74
(+) person 0.55
(+) author 0.41
(+) politician 0.33
(+) business 0.31

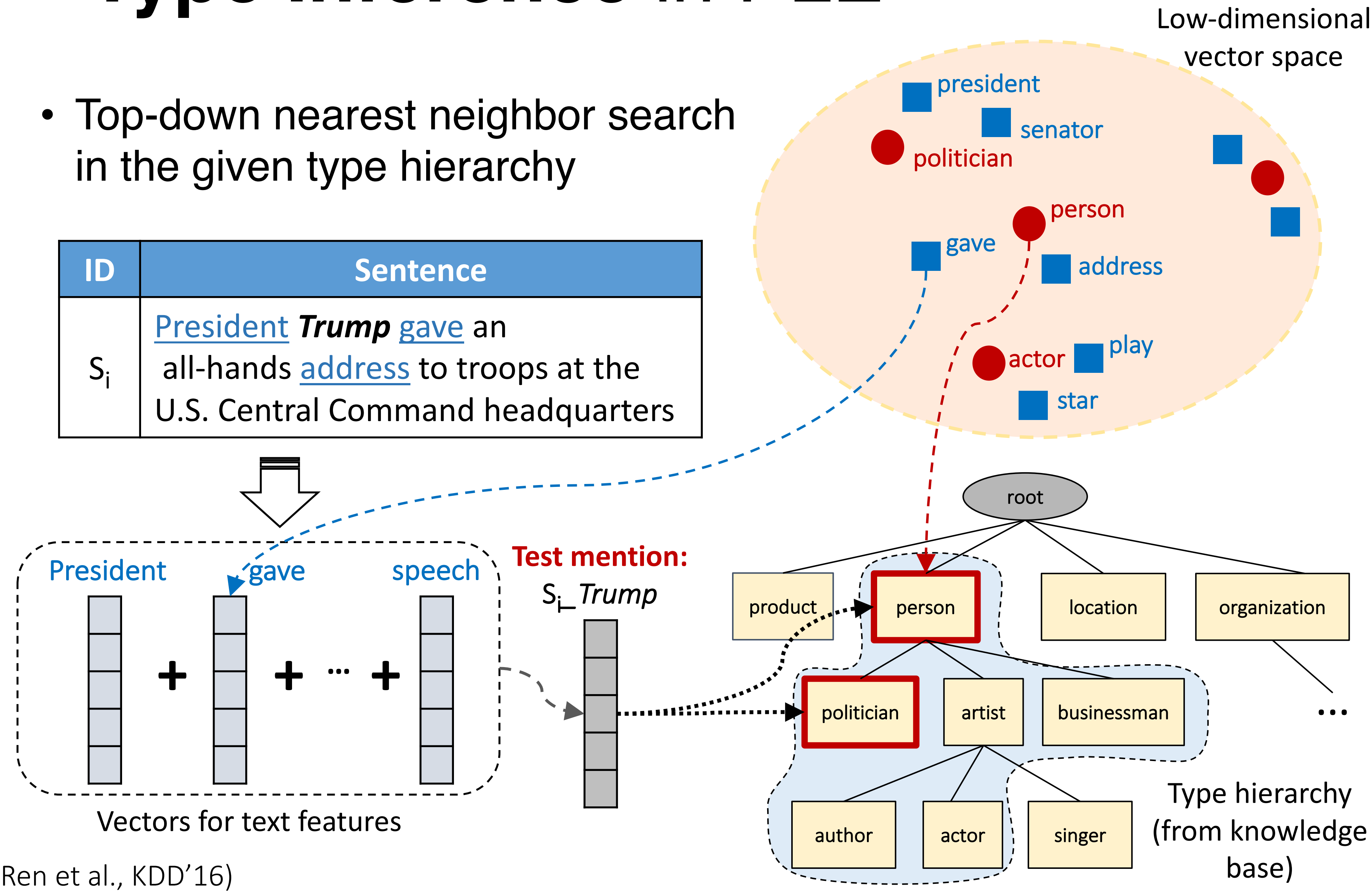
For a noisy mention, its "*best candidate type*" should be **ranked higher** than all its "*non-candidate types*"

(Ren et al., KDD'16)

Type Inference in PLE

- Top-down nearest neighbor search in the given type hierarchy

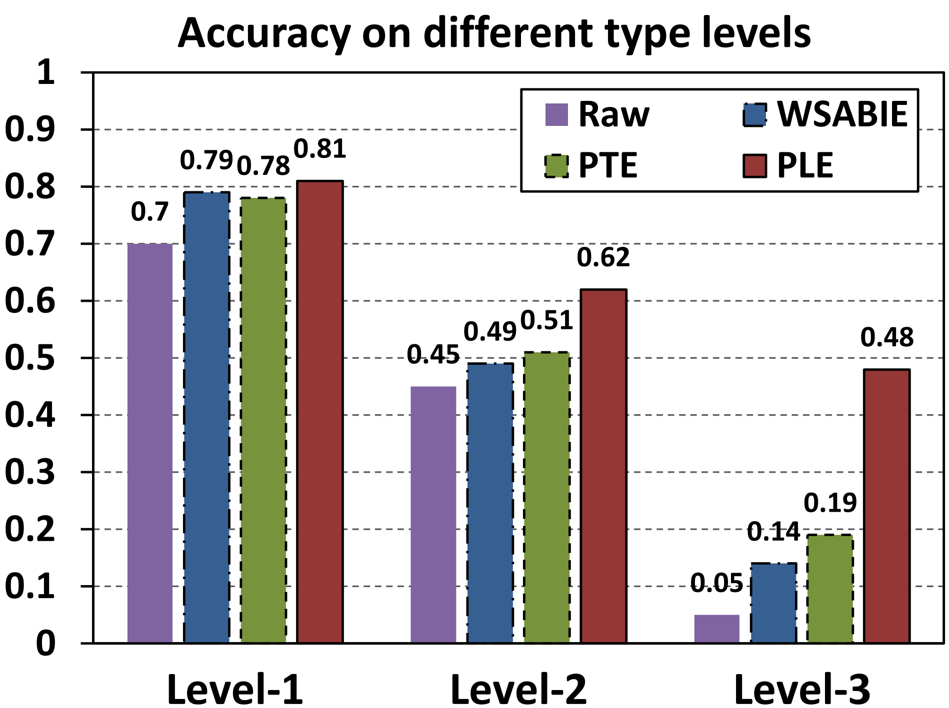
ID	Sentence
S_i	President ***Trump*** gave an all-hands address to troops at the U.S. Central Command headquarters



(Ren et al., KDD'16)

PLE: Performance of Fine-Grained Entity Typing

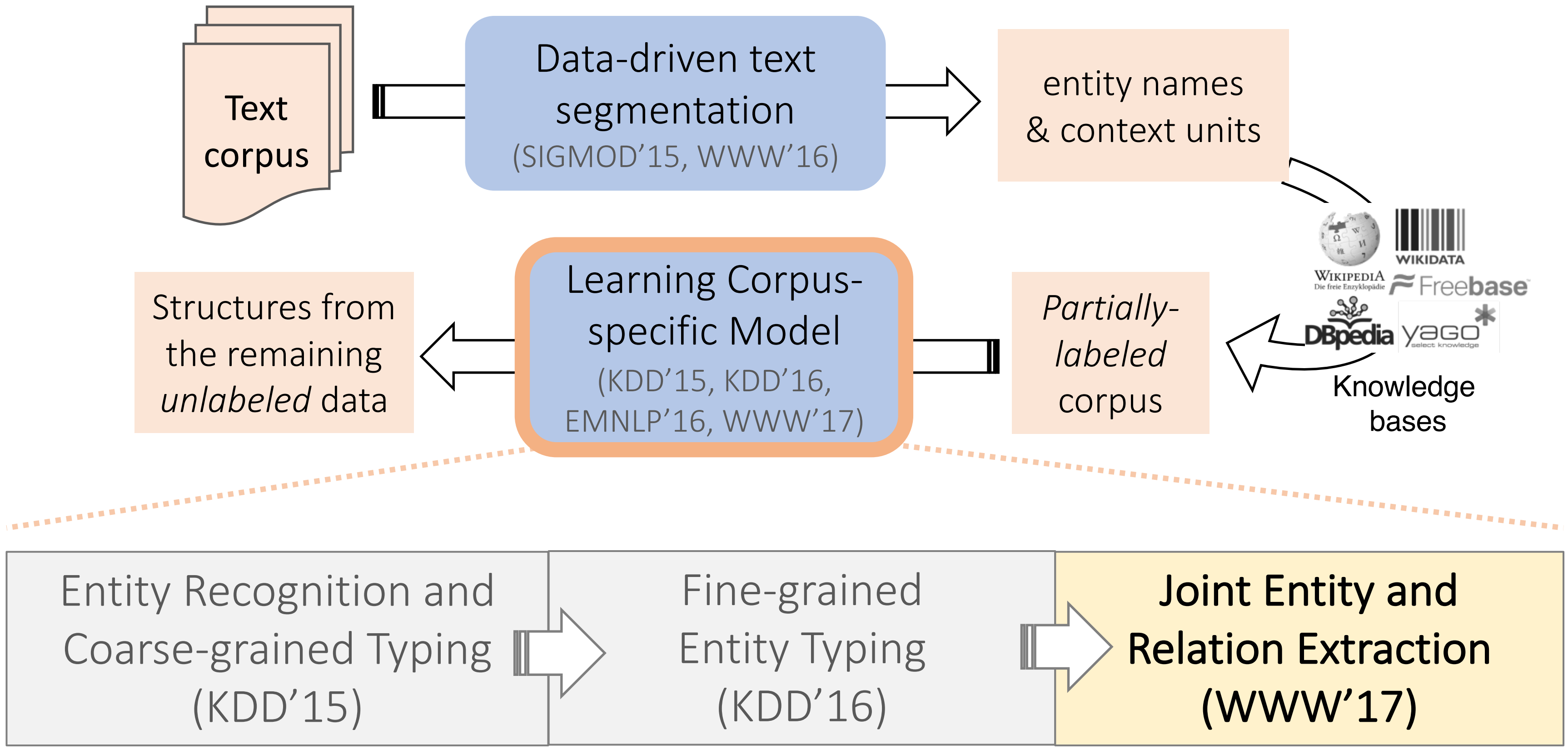
$$\text{Accuracy} = \frac{\#\ mentions\ with\ all\ types\ correctly\ predicted}{\#\ mentions\ in\ the\ test\ set}$$



- **Raw**: candidate types from distant supervision
- **WSABIE** (Google, ACL'15): joint feature and type embedding
- **P**redictive **T**ext **E**mbedding (MSR, WWW'15): joint mention, feature and type embedding
  - Both WASBIE and PTE suffer from "noisy" training labels
- **PLE** (KDD'16): partial-label loss for context-aware labeling

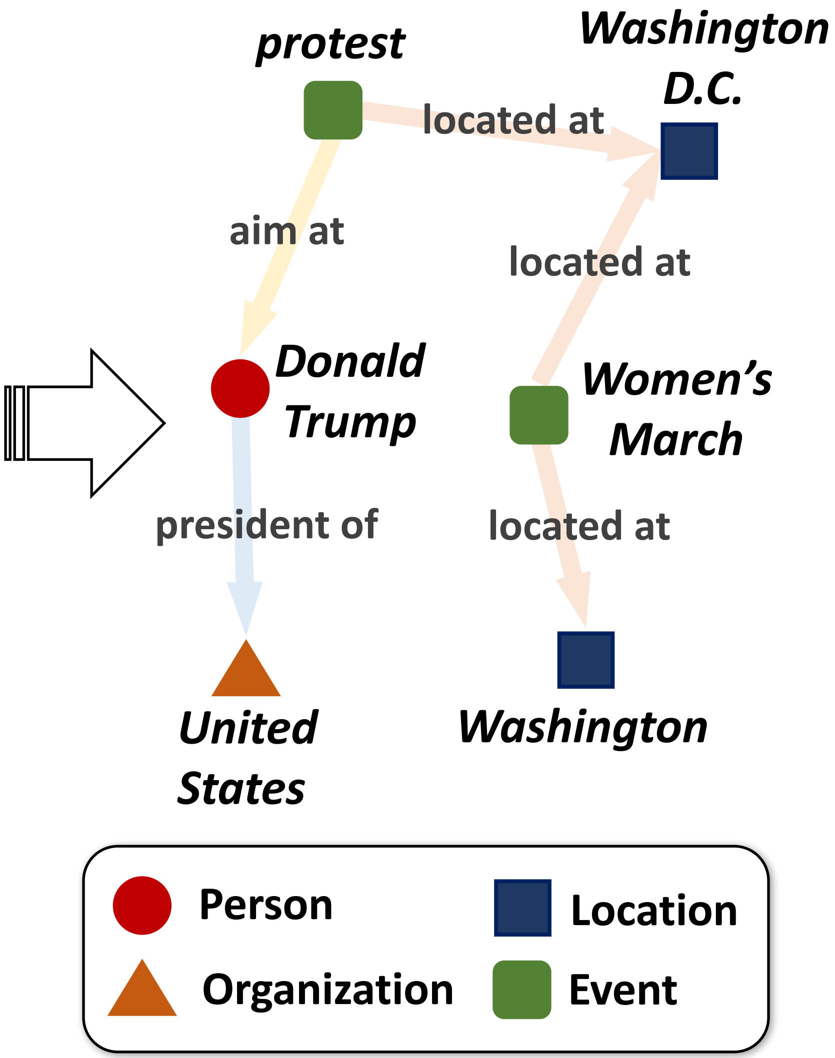
OntoNotes public dataset (Weischedel et al. 2011, Gillick et al., 2014): 13,109 news articles, 77 annotated documents, 89 entity types

Corpus to Structured Network: **The Roadmap**



Joint Extraction of Typed Entities and Relations

The **Women's March** was a worldwide **protest** on January 21, 2017. The **protest** was aimed at **Donald Trump**, the recently inaugurated president of the **United States**. The first **protest** was planned in **Washington, D.C.**, and was known as the **Women's March** on **Washington**.



Prior Work: Relation Extraction (RE)

Substantial human annotation → ***No human annotation***

Supervised RE systems ⇒ **Pattern-based bootstrapping RE systems** ⇒ **Distantly-supervised RE systems (cont.)**

Supervised RE systems:
- Hard to be ported to deal with different kinds of corpora

Pattern-based bootstrapping RE systems:
- Focus on “explicit” relation mentions
- “Semantic drift”

Distantly-supervised RE systems (cont.):
- Error propagation
- Noisy candidate type labels

Mintz et al. *Distant supervision for relation extraction without labeled data*. ACL, 2009.

Etzioni et al. *Web-scale information extraction in knowitall.* WWW, 2004.

Surdeanu et al. *Multi-instance multi-label learning for relation extraction*. EMNLP, 2012.

Prior Work: An "Incremental" System Pipeline

Error propagation cascading down the pipeline

Entity mention detection → Context-aware entity typing → Relation mention detection → Context-aware relation typing

Entity boundary errors:

The *Women* 's March was a worldwide *protest* on *January 21, 2017*.

Entity type errors:

The *Women* 's March was a worldwide *protest* on *January 21, 2017*. → person ✗

Relation mention errors:

(*women*, *protest*) ✗

(*protest*, *January 21, 2017*)

Relation type errors

(*women*, *protest*) → is a ✗

(protest, January 21, 2017)

(Mintz et al., 2009), (Riedel et al., 2010), (Hoffmann et al., 2011), (Surdeanu et al., 2012), (Nagesh et al., 2014), …

My Solution: **CoType** (WWW'17)

Entity mention detection

Context-aware entity typing

Relation mention detection

Context-aware relation typing

1. Data-driven detection of entity and relation mentions

- Data-driven text segmentation
- Syntactic pattern learning from KBs

2. Joint typing of entity and relation mentions

- **Context-aware** type modeling
- Model **entity-relation interactions**

(Ren et al. WWW'17)

Data-Driven Entity and Relation Detection

S2: The protest was aimed at Donald Trump, the recently inaugurated president of the United States.

Frequent Pattern Mining

S2: The <u>protest</u> was <u>aimed at</u> <u>Donald Trump</u>, the recently inaugurated <u>president of</u> the <u>United States</u>.

Segment Quality Estimation

Phrases quality: *United States*: 0.9, *was aimed at*: 0.4, ...
Part-of-speech (POS) patterns quality: *ADJ NN*: 0.85, *V PROP*: 0.4, …

POS-guided Segmentation

S2: The *protest* *<u>was aimed at</u>* *Donald Trump*, the recently inaugurated *<u>president of</u>* the *United States*.

Quality Re-estimation & **Re-segmentation**

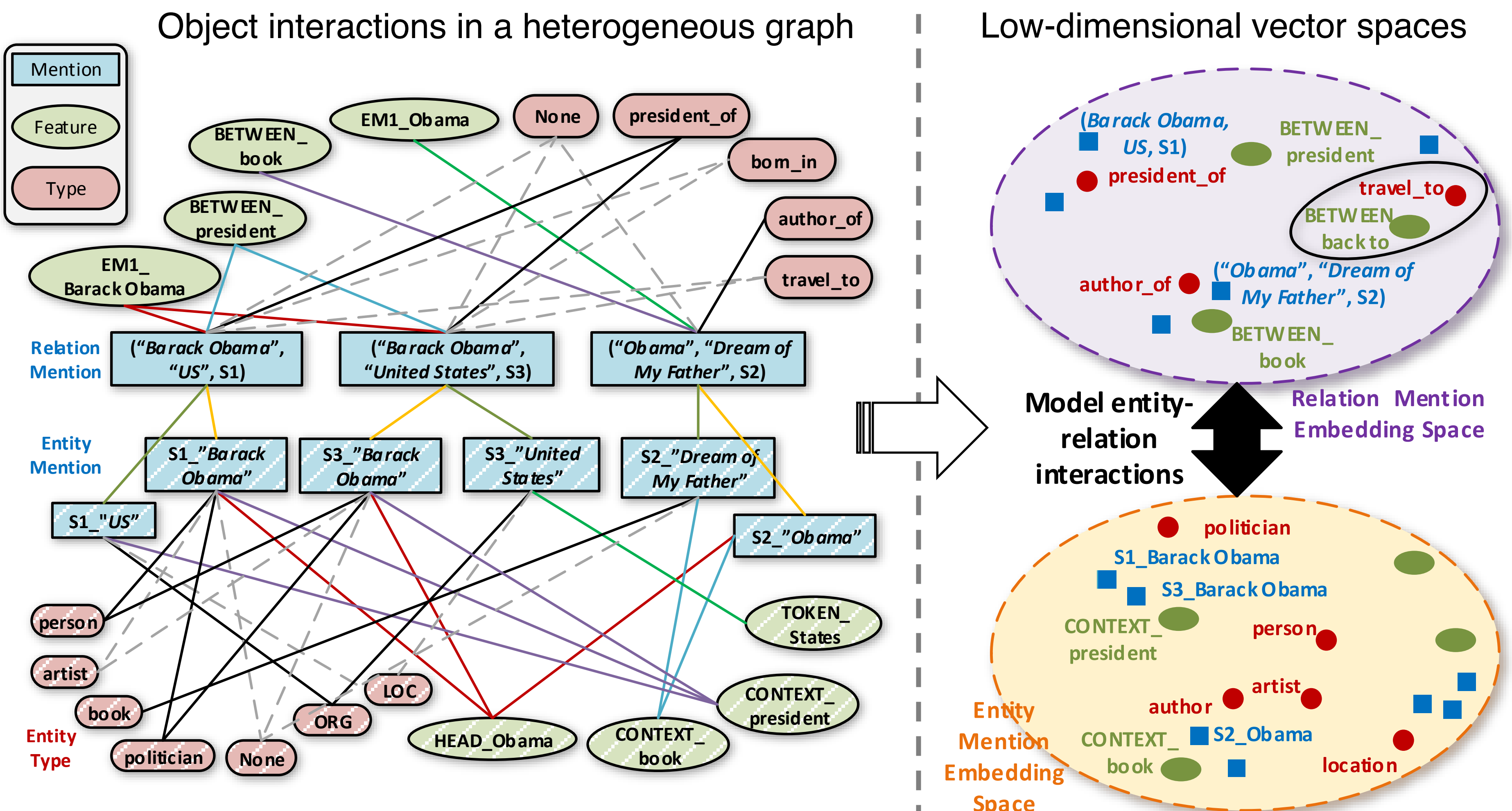
(**S2**: *protest, Donald Trump*), (**S2**: *Donald Trump, United States*)

Entity Mention Detection: Results

	POS Tag Pattern	Example
Good (high score)	*NNP NNP* *NN NN* *CD NN* *JJ NN*	San Francisco/Barack Obama/United States comedy drama/car accident/club captain seven network/seven dwarfs/2001 census crude oil/nucletic acid/baptist church
Bad (low score)	*DT JJ NND* *CD CD NN IN* *NN IN NNP NNP* *VVD RB IN*	a few miles/the early stages/the late 1980s 2 : 0 victory over/1 : 0 win over rating on rotten tomatoes worked together on/spent much of

	NYT	Wiki-KBP	BioInfer
FIGER segmenter [UW, 2012]	0.751	0.814	0.652
Our Approach	**0.837**	0.833	0.785

CoType: Co-Embedding for Typing Entities and Relations



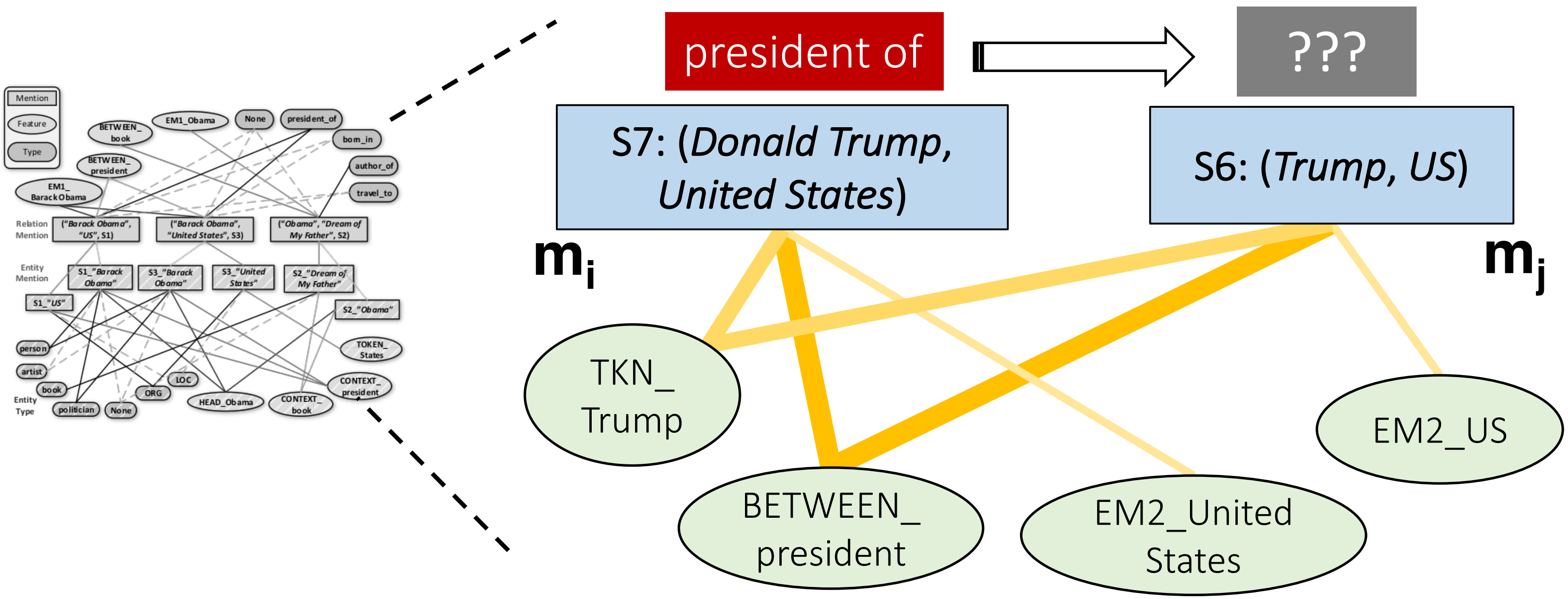
(Ren et al. WWW'17)

Modeling Mention-Feature Co-occurrences

- Second-order Proximity

  Mentions with similar distributions over text features should have similar types

Vertex $\mathbf{m_i}$ and $\mathbf{m_j}$ have a large second-order proximity



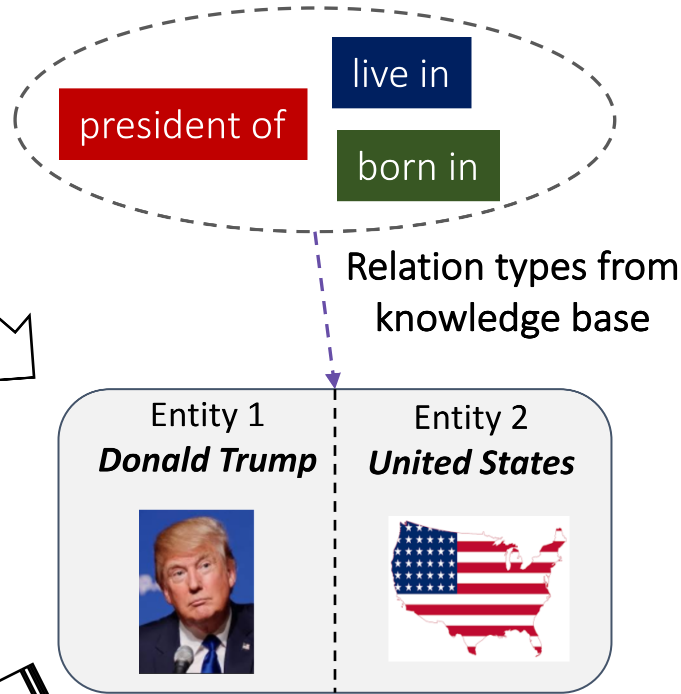
(Tang et al., WWW'15), (Ren et al. WWW'17)

Challenge: Context-Agnostic Labeling

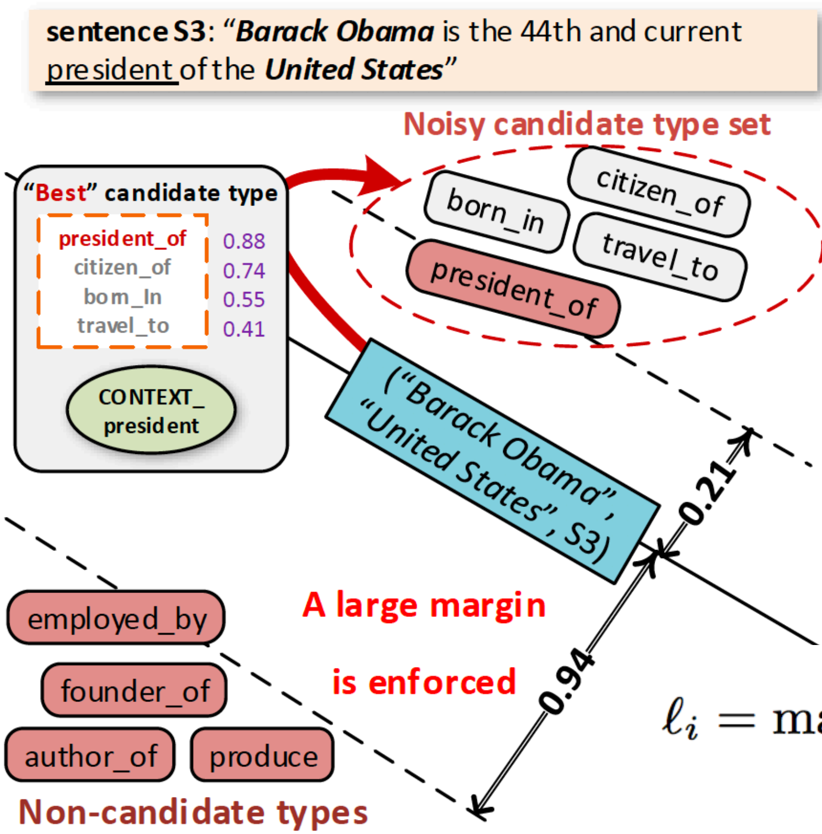
ID	Sentence
S2	The protest was aimed at *Donald Trump*, the recently inaugurated president of the *United States*.

Type labels for relation mention:

E1: *Donald J. Trump*	**E2**: *United States*
E1 Types: person, politician, businessman, author, actor	E2 Types: location, organization
Relations between E1, E2 in KB: president of, live in, born in	



Context-Aware Type Modeling



Partial-label Loss Function

- Vector representation of the relation mention should be **more similar** to its "**best**" **candidate type**, than to any other **non-candidate type**

Score for ***"best"*** **candidate** type

$$\ell_i = \max\left\{0, 1 - \left[\boxed{\max_{y\in\mathcal{Y}_i} s(m_i, y)} - \boxed{\max_{y'\in\overline{\mathcal{Y}}_i} s(m_i, y')}\right]\right\}$$

Maximal score for **non-candidate** types

(Ren et al. WWW'17)

Modeling Entity-Relation Interactions

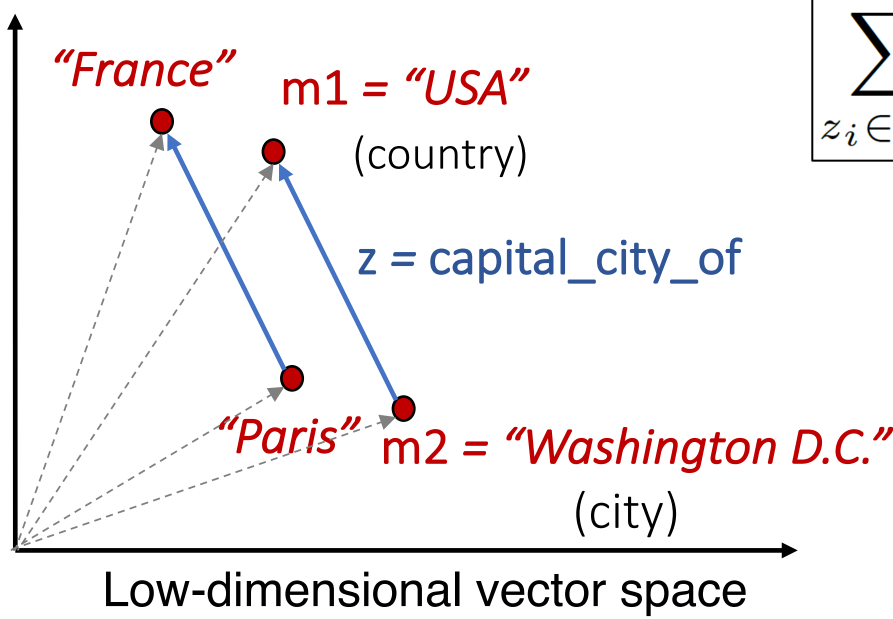
Object "Translating" Assumption

For a relation mention **z** between entity arguments **m1** and **m2**:

vec(m1) ≈ **vec**(m2) + **vec**(z)

Error on a relation triple (z, m1, m2):

$$\tau(z) = \|\mathbf{m}_1 + \mathbf{z} - \mathbf{m}_2\|_2^2$$



$$\sum_{z_i \in \mathcal{Z}_L} \sum_{v=1}^{V} \max\{0, 1 + \tau(z_i) - \tau(z_v)\}$$

positive relation triple

negative relation triple

(Bordes, NIPS'13), (Ren et al., WWW'17)

Reducing Error Propagation: A Joint Optimization Framework

Modeling **entity-relation interactions**

$$O_{ZM} = \sum_{z_i \in \mathcal{Z}_L} \sum_{v=1}^{V} \max\left\{0, 1 + \tau(z_i) - \tau(z_v)\right\}$$

$$\min\ \mathcal{O} = \boxed{\mathcal{O}_M} + \boxed{\mathcal{O}_Z} + \boxed{\mathcal{O}_{ZM}}$$

$$O_Z = \mathcal{L}_{ZF} + \sum_{i=1}^{N_L} \ell_i + \frac{\lambda}{2} \sum_{i=1}^{N_L} \|\mathbf{z}_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_r} \|\mathbf{r}_k\|_2^2.$$

Modeling types of **relation mentions**

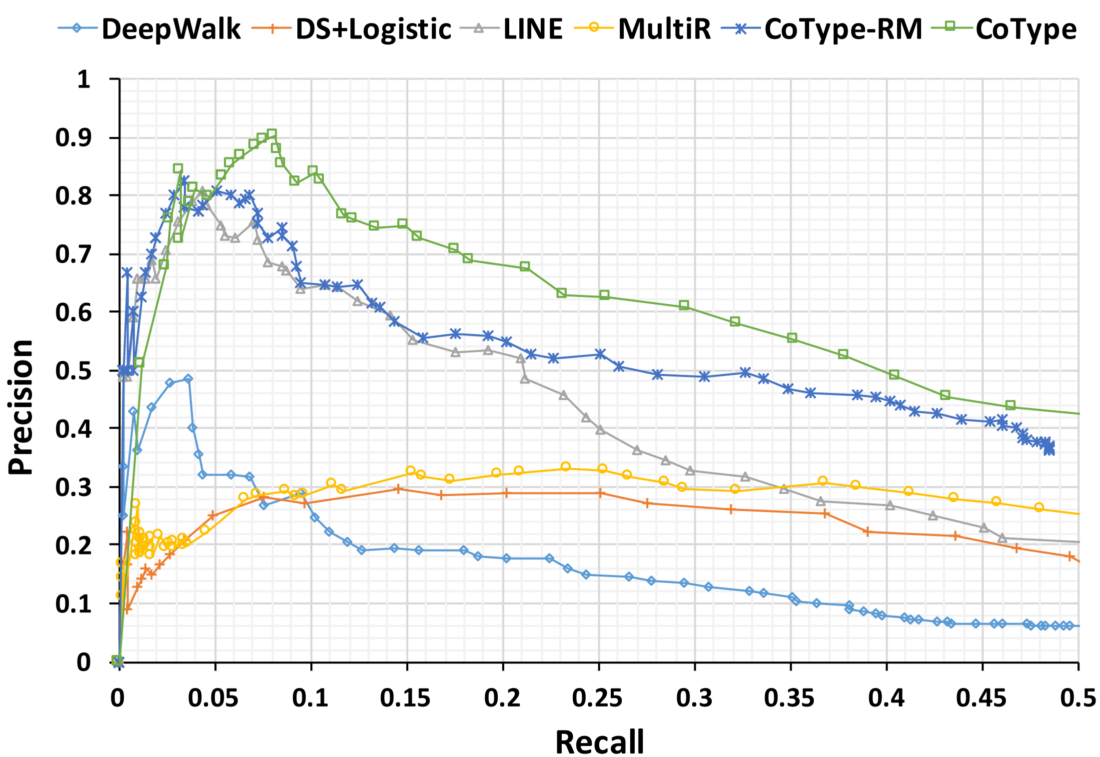
$$O_M = \mathcal{L}_{MF} + \sum_{i=1}^{N_L'} \ell_i' + \frac{\lambda}{2} \sum_{i=1}^{N_L'} \|\mathbf{m}_i\|_2^2 + \frac{\lambda}{2} \sum_{k=1}^{K_y} \|\mathbf{y}_k\|_2^2.$$

Modeling types of **entity mentions**

(Ren et al., WWW'17)

CoType: Comparing with State-of-the-Arts RE Systems

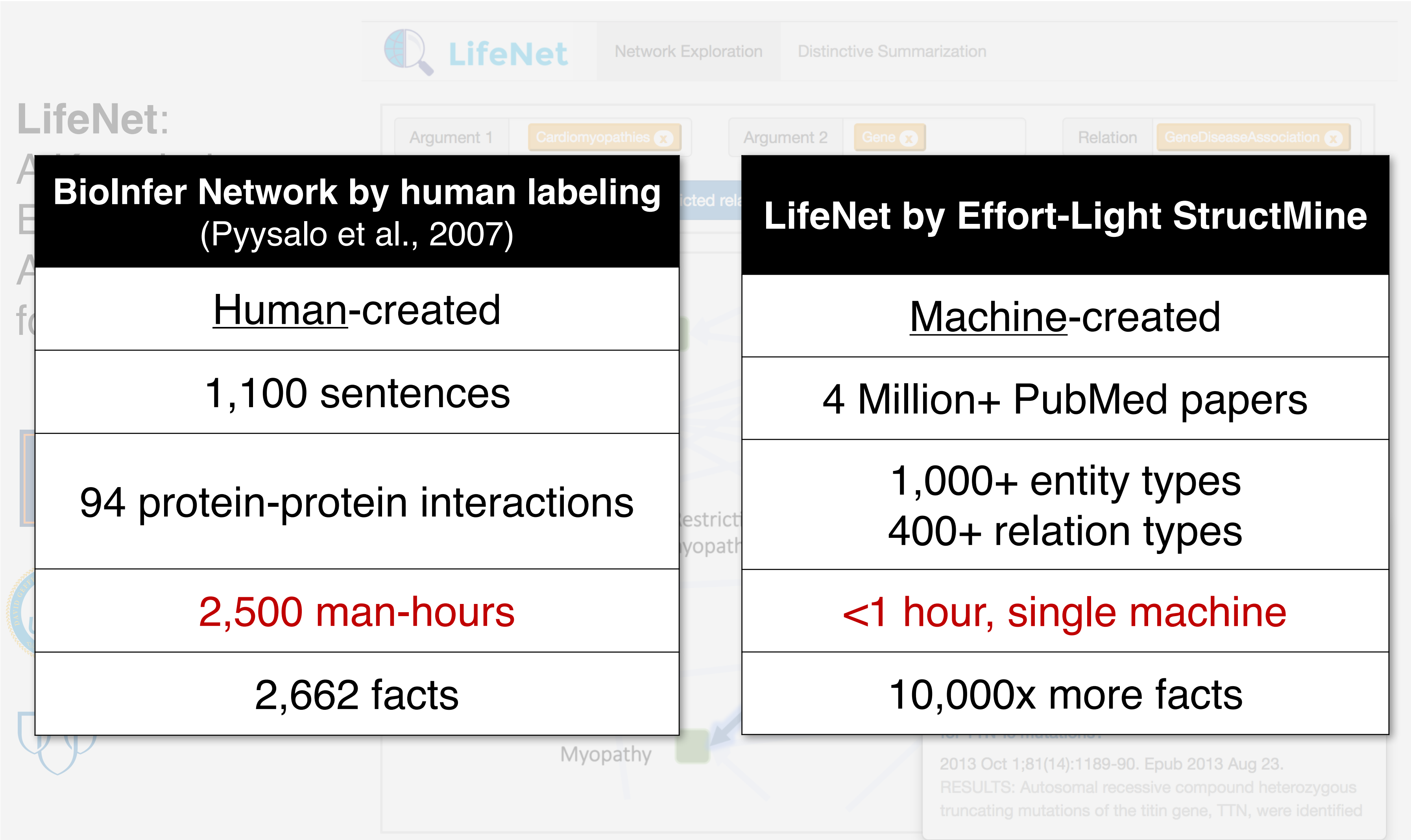
- Given candidate relation mentions, predict its relation type if it expresses a relation of interest; otherwise, output "None"



- DS+Logistic (Stanford, ACL'09): logistic classifier on DS
- MultiR (UW, ACL'11): handles inappropriate labels in DS
- DeepWalk (StonyBrook, KDD'14): homogeneous graph embedding
- LINE (MSR, WWW'15): joint feature & type embedding
- CoType-RM (WWW'17): only models relation mentions
- CoType (WWW'17): models entity-relation interactions

NYT public dataset (Riedel et al. 2010, Hoffmann et al., 2011): 1.18M sentences in the corpus, 395 manually annotated sentences for evaluation, 24 relation types

An Ongoing Application to Life Sciences



BioInfer Network by human labeling (Pyysalo et al., 2007)
<u>Human</u>-created
1,100 sentences
94 protein-protein interactions
2,500 man-hours
2,662 facts

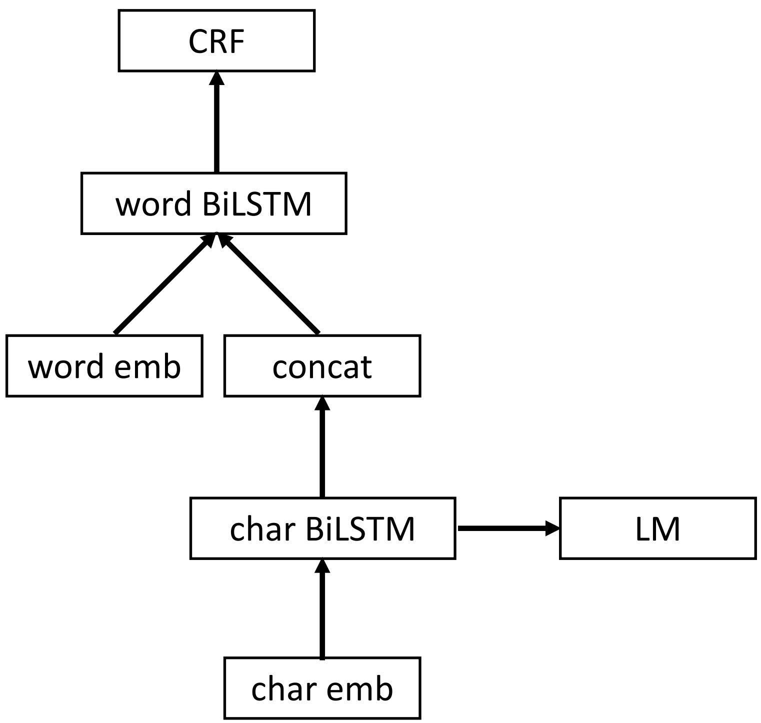
LifeNet by Effort-Light StructMine
<u>Machine</u>-created
4 Million+ PubMed papers
1,000+ entity types 400+ relation types
<1 hour, single machine
10,000x more facts

(Pyysalo et al., BMC Bioinformatics'07)
(Ren et al., ACL'17 demo, *under review*)

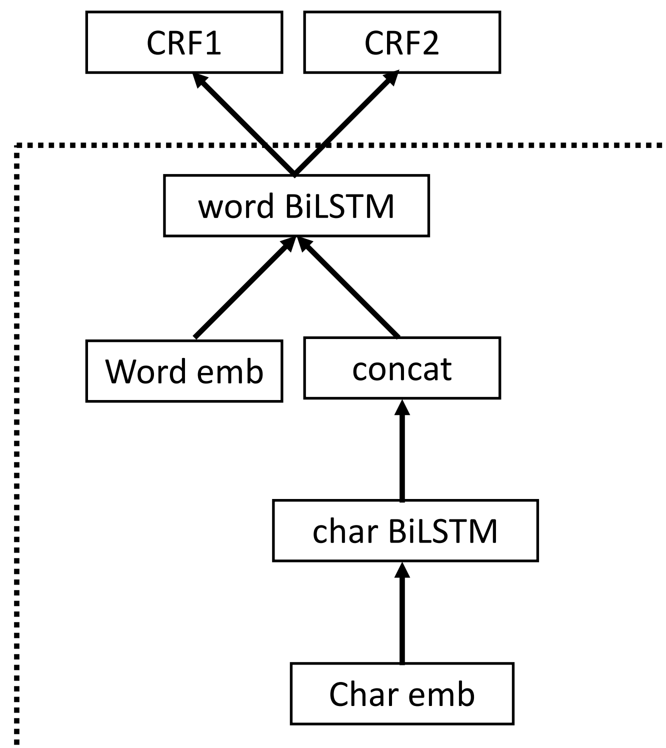
Performance evaluation on BioInfer:
Relation Classification Accuracy = 61.7%
(11%↑ over the best-performing baseline)

Biomedical Named Entity Recognition by Multi-tasking different datasets

Single-task/dataset learning



Multi-task/dataset learning



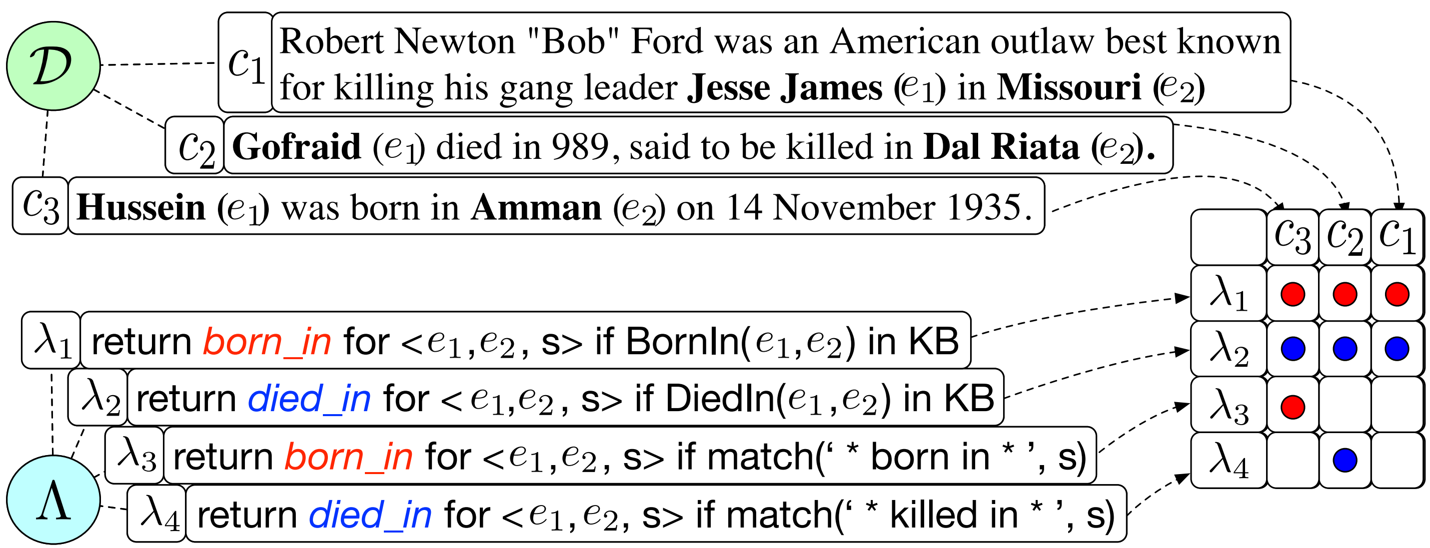
(Liu et al., AAAI'18)

Performance of NER on Biomed Benchmark Datasets

		Dataset Benchmark	Liu et al. 2017 (single-task)	Multi-task
BC2GM (gene/protein)	Prec	**88.48**	83.82	82.99
	Rec	**85.97**	82.12	83.08
	F1	87.21	82.96	83.03
BC4CHEMD (Chemical)	Prec	89.09	90.21	**90.50**
	Rec	**85.75**	84.82	85.45
	F1	87.39	87.44	87.90
BC5CDR (Chemical, Diseases)	Prec	**89.21**	85.71	87.70
	Rec	84.45	84.71	**86.63**
	F1	86.76	85.21	87.16
NCBI (Diseases)	Prec	85.10	84.06	**85.39**
	Rec	80.80	84.57	**87.44**
	F1	82.90	84.32	86.40
JNLPBA (Gene, DNA, Cell Line, etc.)	Prec	69.42	72.10	**72.89**
	Rec	75.99	**77.52**	77.17
	F1	72.55	74.72	74.97

“Heterogeneous Supervision” for Relation Extraction

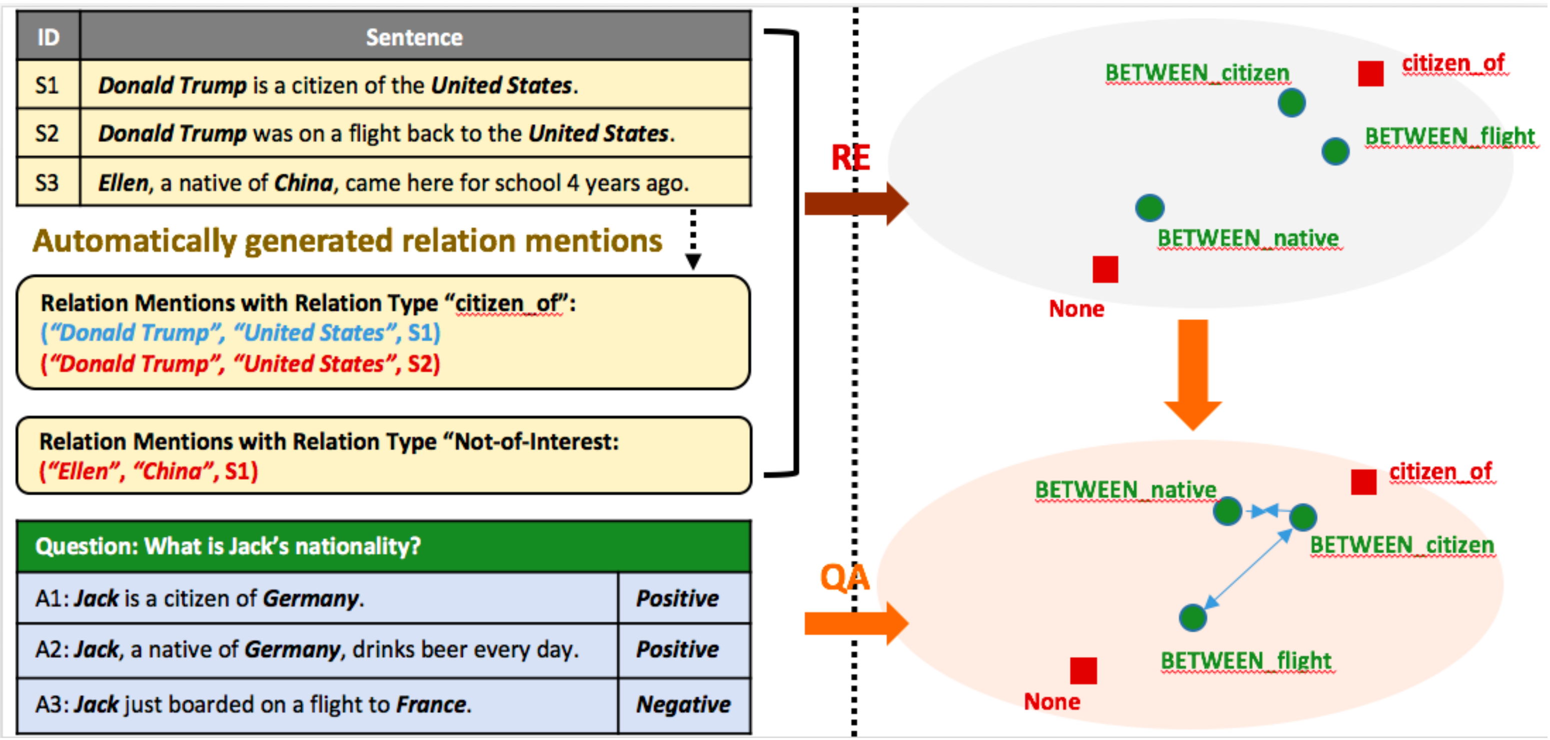
- A principled l framework to **unify** KB-supervision, manual rules, crowd-sourced labels, etc.
- Multiple “**labeling functions**” annotate one instance → resolve conflicts & redundancy → “**expertise**” of each labeling function



(Liu et al., EMNLP’17)

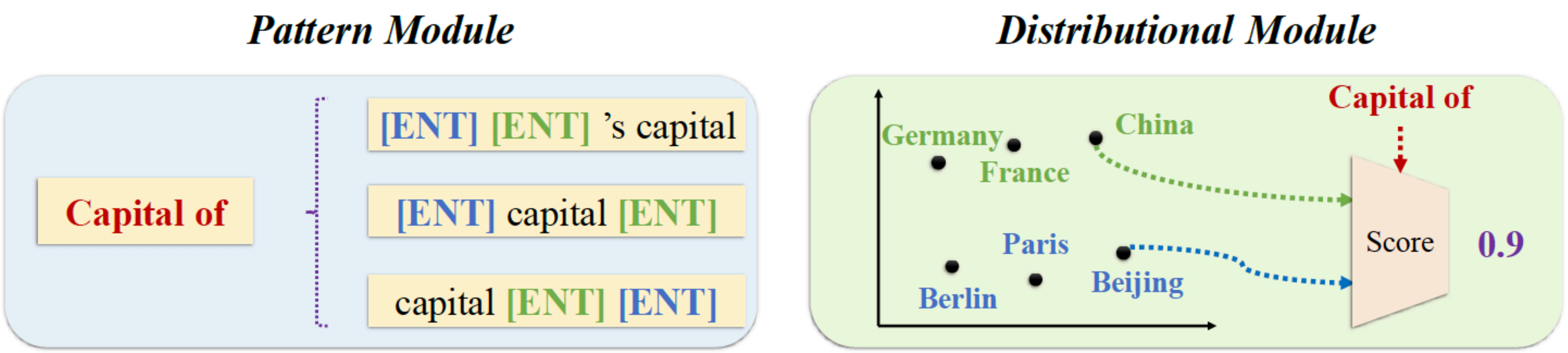
Indirect Supervision for Relation Extraction -- using QA Pairs

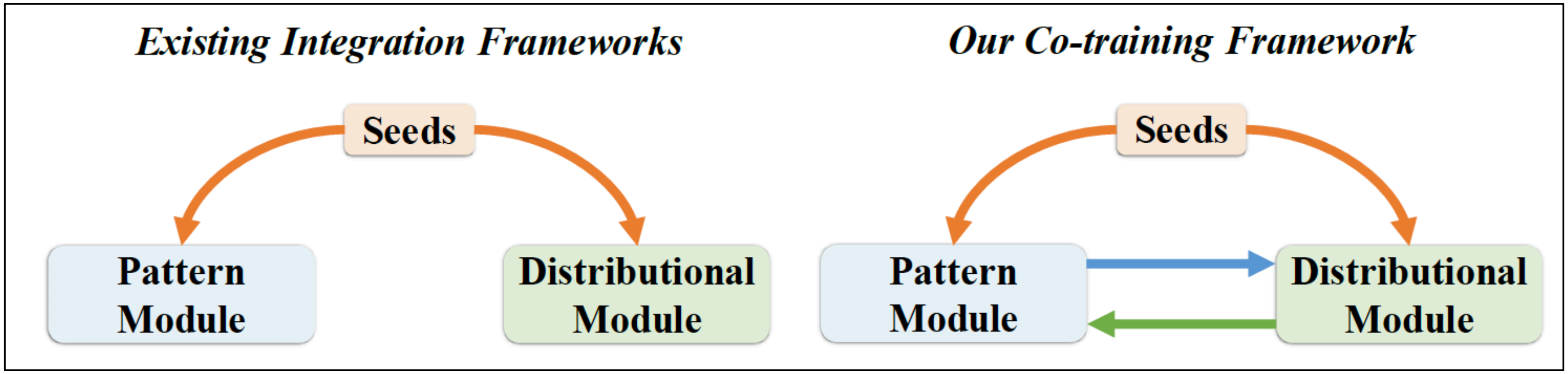
- Questions → positive / negative answers
- pos pairs → similar relation; neg pairs → distinct relations



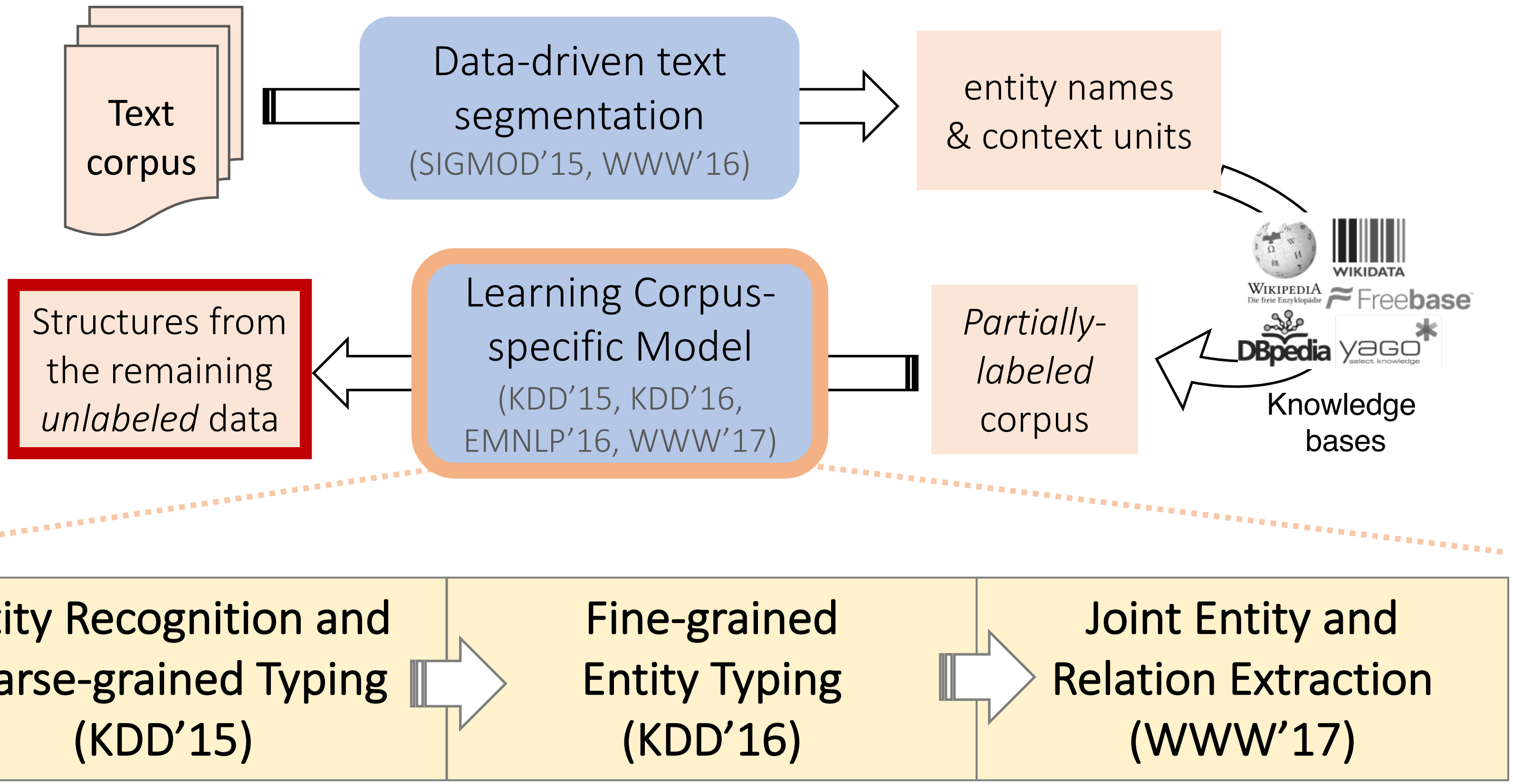
(Wu et al., WSDM'18)

Pattern-enhanced Distributional Representation Learning





Corpus to Structured Network: **The Roadmap**



References I

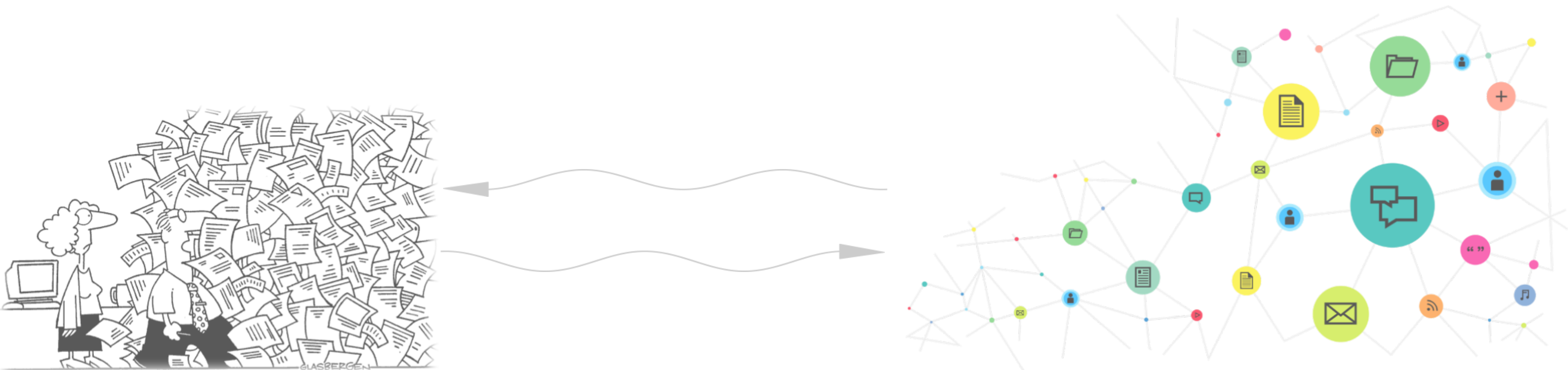
- **Xiang Ren**, Zeqiu Wu, Wenqi He, Meng Qu, Clare R. Voss, Heng Ji, Tarek F. Abdelzaher, Jiawei Han. CoType: Joint Extraction of Typed Entities and Relations with Knowledge Bases. WWW, 2017.
- **Xiang Ren**, Ahmed El-Kishky, Heng Ji, and Jiawei Han. Automatic Entity Recognition and Typing in Massive Text Data (Conference Tutorial). SIGMOD, 2016.
- **Xiang Ren***, Wenqi He*, Meng Qu, Lifu Huang, Heng Ji, Jiawei Han. AFET: Automatic Fine-Grained Entity Typing by Hierarchical Partial-Label Embedding. EMNLP, 2016.
- **Xiang Ren***, Wenqi He*, Meng Qu, Heng Ji, Clare R. Voss, Jiawei Han. Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding. KDD, 2016.
- **Xiang Ren**, Wenqi He, Ahmed El-Kishky, Clare R. Voss, Heng Ji, Meng Qu, Jiawei Han. Entity Typing: A Critical Step for Mining Structures from Massive Unstructured Text (Invited Paper). MLG, 2016.
- **Xiang Ren**, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, H. Ji, J. Han. ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering. KDD, 2015.
- **Xiang Ren**, Tao Cheng. Synonym Discovery for Structured Entities on Heterogeneous Graphs. WWW, 2015.
- Tarique A. Siddiqui*, **Xiang Ren***, Aditya Parameswaran, Jiawei Han. FacetGist: Collective Extraction of Document Facets in Large Technical Corpora. CIKM, 2016.
- Jialu Liu, Jingbo Shang, Chi Wang, **Xiang Ren**, Jiawei Han. Mining Quality Phrases from Massive Text Corpora. SIGMOD, 2015.

References II

- Marina Danilevsky, Chi Wang, Nihit Desai, **Xiang Ren**, Jingyi Guo, and Jiawei Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents. SDM, 2014
- **Xiang Ren**, Yuanhua Lv, Kuansan Wang, Jiawei Han. Comparative Document Analysis for Large Text Corpora. WSDM, 2017.
- Jialu Liu, **Xiang Ren**, Jingbo Shang, Taylor Cassidy, Clare R. Voss, Jiawei Han. Representing Documents via Latent Keyphrase Inference. WWW, 2016.
- Hyungsul Kim, **Xiang Ren**, Yizhou Sun, Chi Wang, and Jiawei Han. Semantic Frame-Based Document Representation for Comparable Corpora. ICDM, 2013.
- **Xiang Ren**, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han. ClusCite: Effective Citation Recommendation by Information Network-Based Clustering. KDD, 2014.
- X. Yu, **Xiang Ren**, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. Personalized Entity Recommendation: A Heterogeneous Information Network Approach. WSDM 2014a.
- **Xiang Ren**, Yujing Wang, Xiao Yu, Jun Yan, Zheng Chen, Jiawei Han. Heterogeneous Graph-Based Intent Learning from Queries, Web Pages and Wikipedia Concepts. WSDM 2014b.
- X. Yu, **Xiang Ren**, Y. Sun, B. Sturt, U. Khandelwal, Q. Gu, B. Norick, and J. Han. HeteRec: Entity Recommendation in Heterogeneous Information Networks with Implicit User Feedback. RecSys, 2013..
- Xiao Yu, Xiang Ren, Quanquan Gu, Yizhou Sun and Jiawei Han. Collaborative Filtering with Entity Similarity Regularization in Heterogeneous Information Networks. IJCAI-HINA, 2013.

Construction and Querying of
Large-scale Knowledge Bases

Part II: Schema-agnostic Knowledge Base Querying

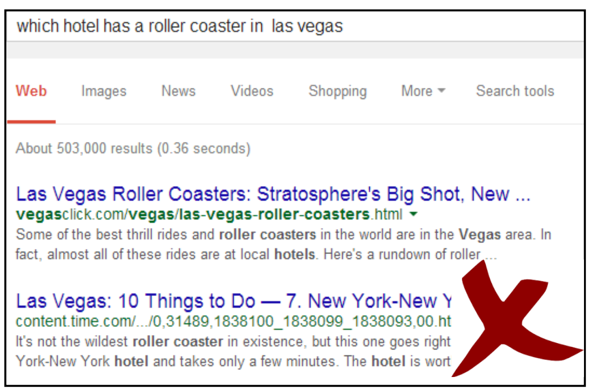


Transformation in Information Search

Desktop search

Mobile search

Lengthy Documents? Direct Answers!



which hotel has a roller coaster in las vegas

Web Images News Videos Shopping More Search tools

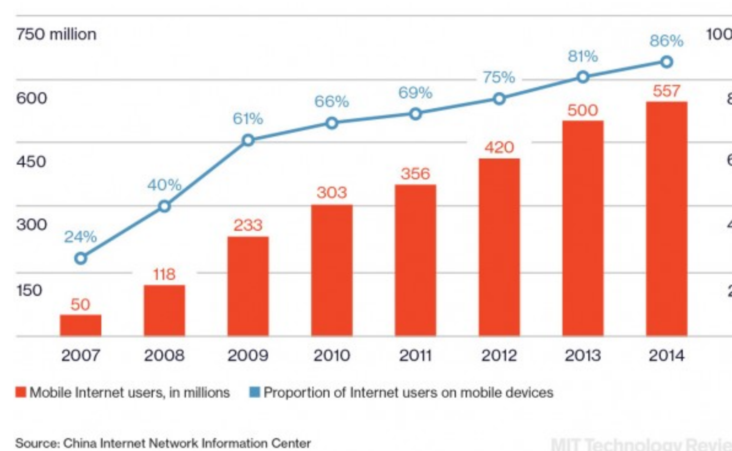
About 503,000 results (0.36 seconds)

Las Vegas Roller Coasters: Stratosphere's Big Shot, New ...
vegasclick.com/vegas/las-vegas-roller-coasters.html
Some of the best thrill rides and roller coasters in the world are in the Vegas area. In fact, almost all of these rides are at local hotels. Here's a rundown of roller ...

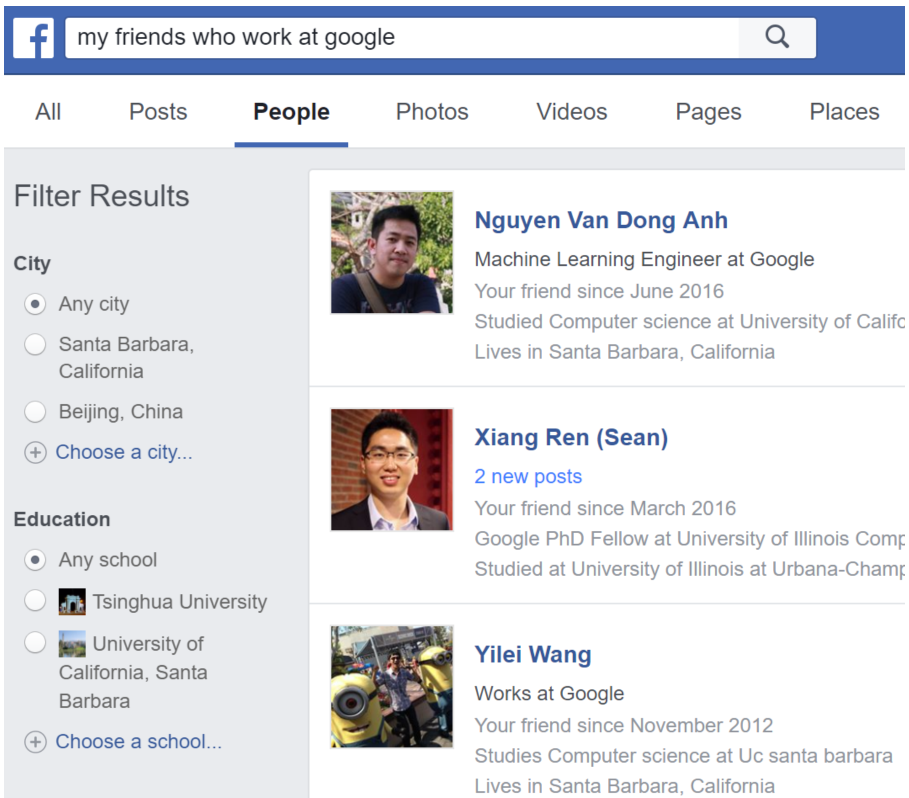
Las Vegas: 10 Things to Do — 7. New York-New Y
content.time.com/.../0,31489,1838100_1838099_1838093,00.ht
It's not the wildest roller coaster in existence, but this one goes right
York-New York hotel and takes only a few minutes. The hotel is wort

Answer: New York-New York hotel

Surge of mobile Internet use in China



Application: Facebook Entity Graph



People, Places, and Things

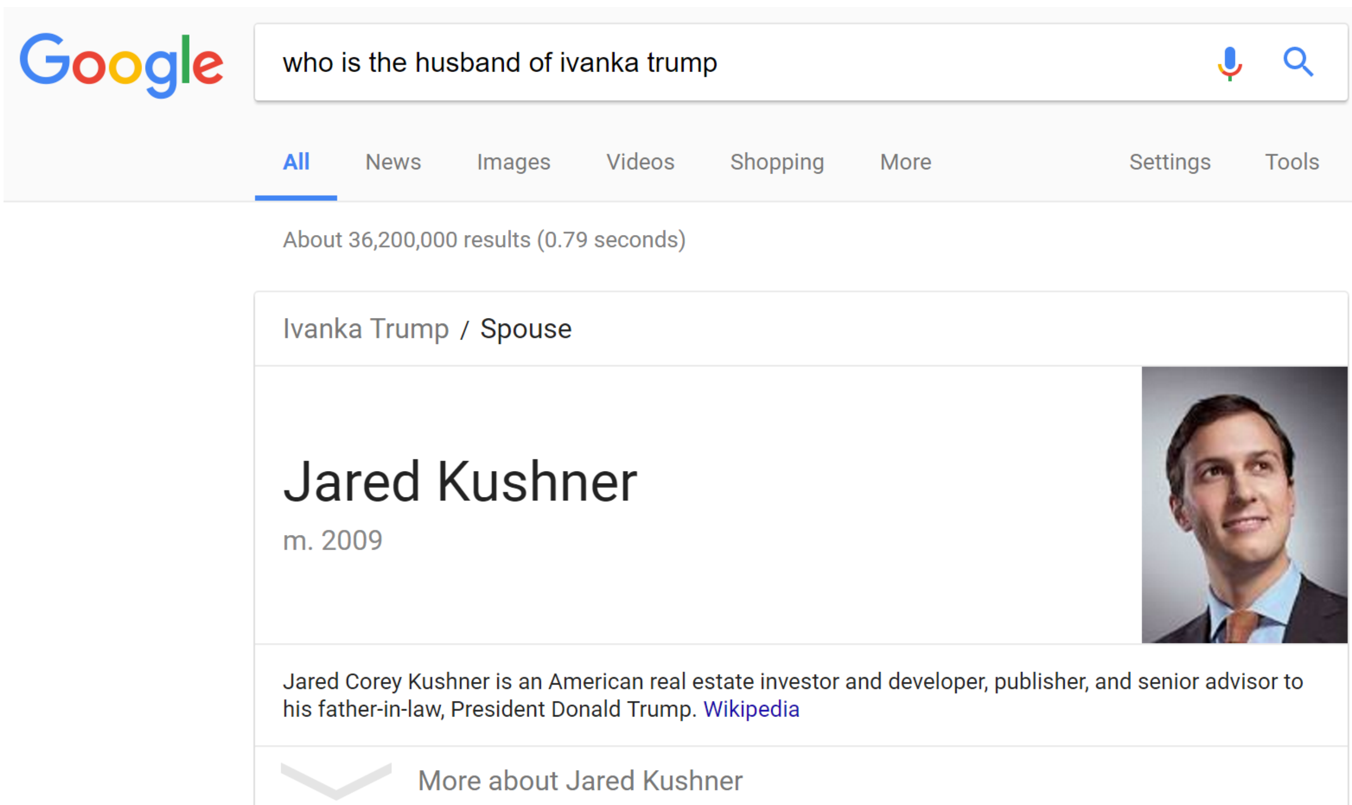
Facebook's knowledge graph (entity graph) stores as entities the users, places, pages and other objects within the Facebook.

Connecting

The connections between the entities indicate the type of relationship between them, such as friend, following, photo, check-in, etc.

QA Engine instead of Search Engine

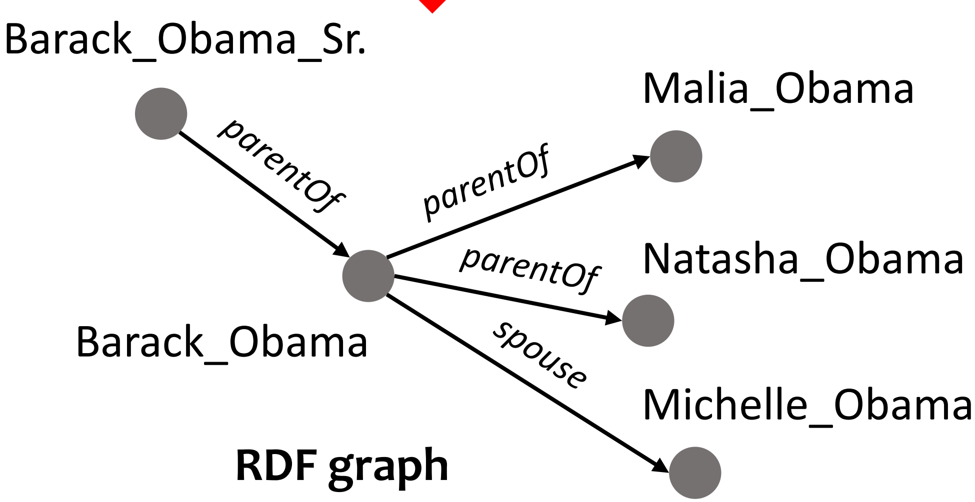
- Behind the scene: A knowledge graph with millions of entities and billions of facts



Structured Query: RDF + SPARQL

Triples in an RDF graph

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 massive, loosely-defined schema

Certainly, You Do Not Want to Write This!

"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)
```

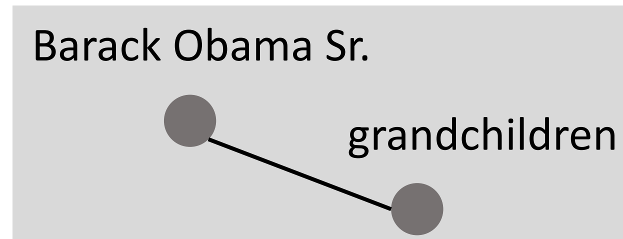
NCI thesaurus

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

Schema-agnostic KB Querying

“Barack Obama Sr. grandchildren”

Keyword query: query like search engine



Graph query: add a little structure

“Who are Barack Obama Sr.’s grandchildren?”

Natural language query: like asking a friend

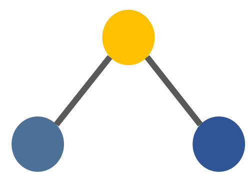
<Barack Obama Sr., Malia Obama>

Query by example: Just show me examples

Graph Query

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

Search intent



Graph query

A match (result)

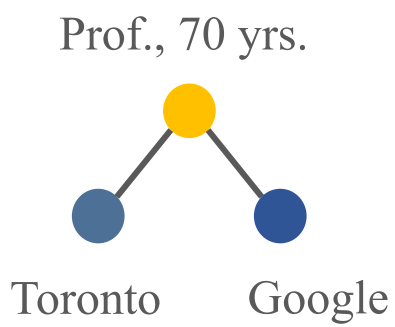
Mismatch between Knowledge Base and Query

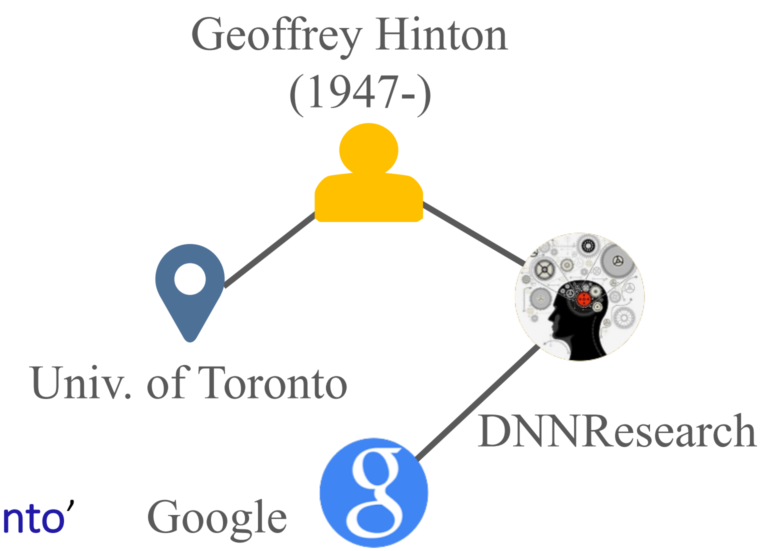
Knowledge Base	Query
“University of Washington”	“UW”
“neoplasm”	“tumor”
“Doctor”	“Dr.”
“Barack Obama”	“Obama”
“Jeffrey Jacob Abrams”	“J. J. Abrams”
“teacher”	“educator”
“1980”	“~30”
“3 mi”	“4.8 km”
“Hinton” - “DNNresearch” - “Google”	“Hinton” - “Google”
…	…

[Yang et al. VLDB'14]

Schema-less Graph Querying (SLQ)

Query → A Match





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

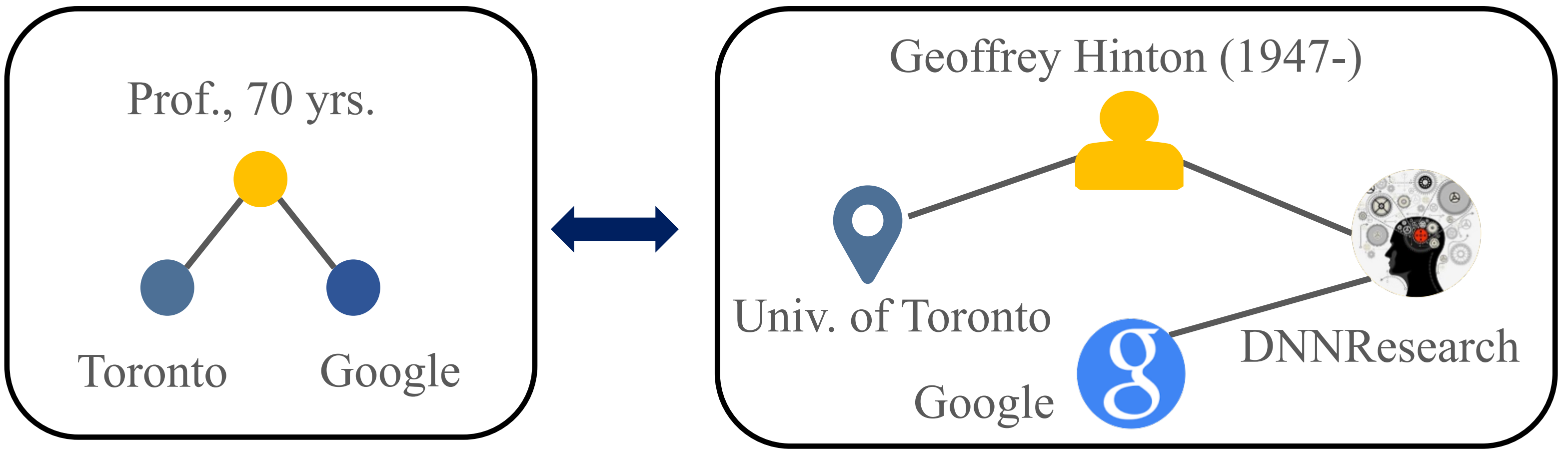
[Yang et al. VLDB'14]

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"
...	...	...

[Yang et al. VLDB'14]

Candidate Match Ranking

Query: Q Candidate Match: $\varphi(Q)$



- 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

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

[Su et al. 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:
Users indicate the **(ir)relevance** of a handful of answers

[Su et al. KDD'15]

Problem Definition

Q: A graph query

G: A knowledge graph

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

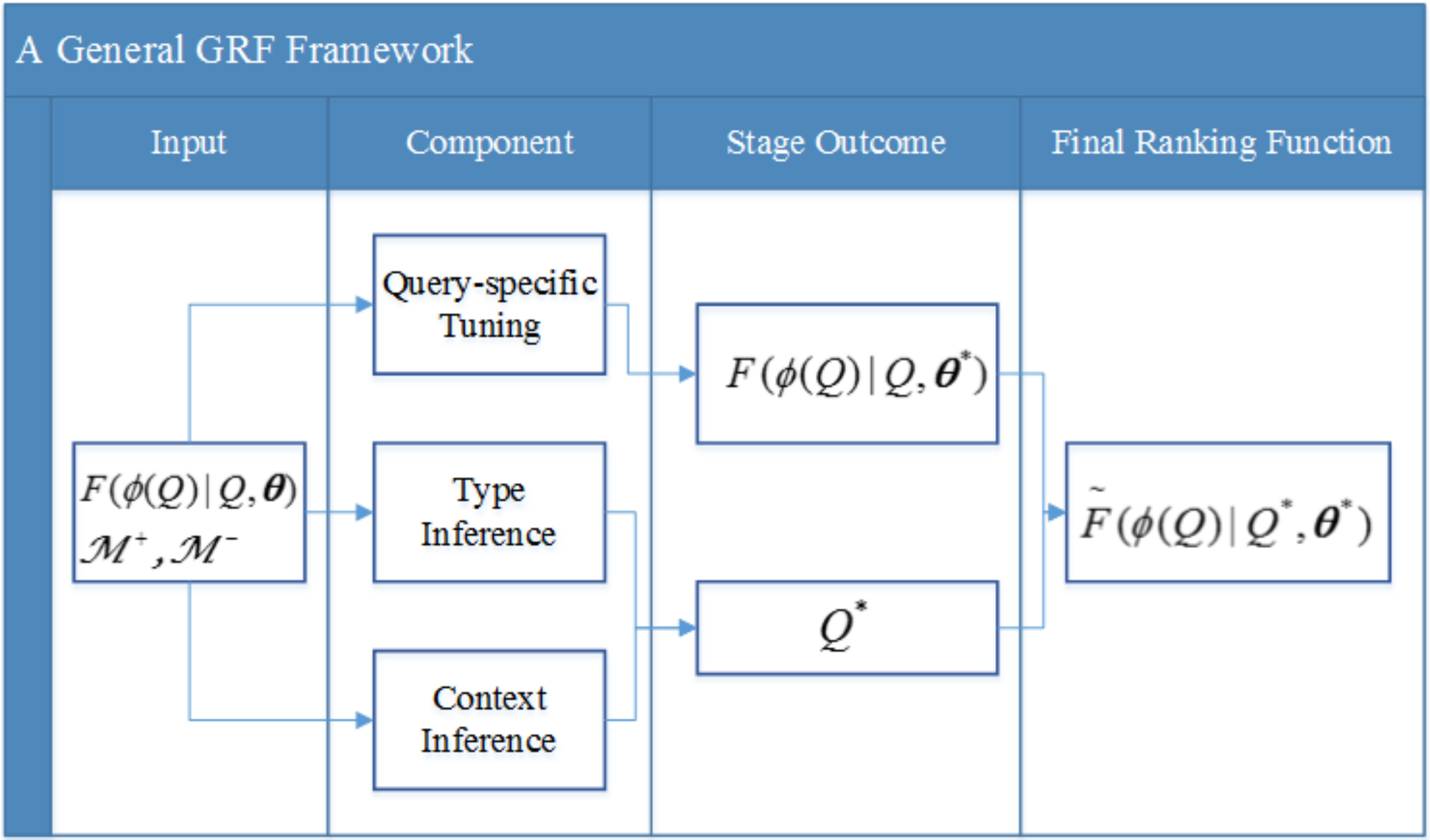
$F(\phi(Q) \mid Q, \boldsymbol{\theta})$: A generic ranking function

$\mathcal{M}^+$: 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 $\mathcal{M}^-$

[Su et al. KDD'15]



[Su et al. KDD'15]

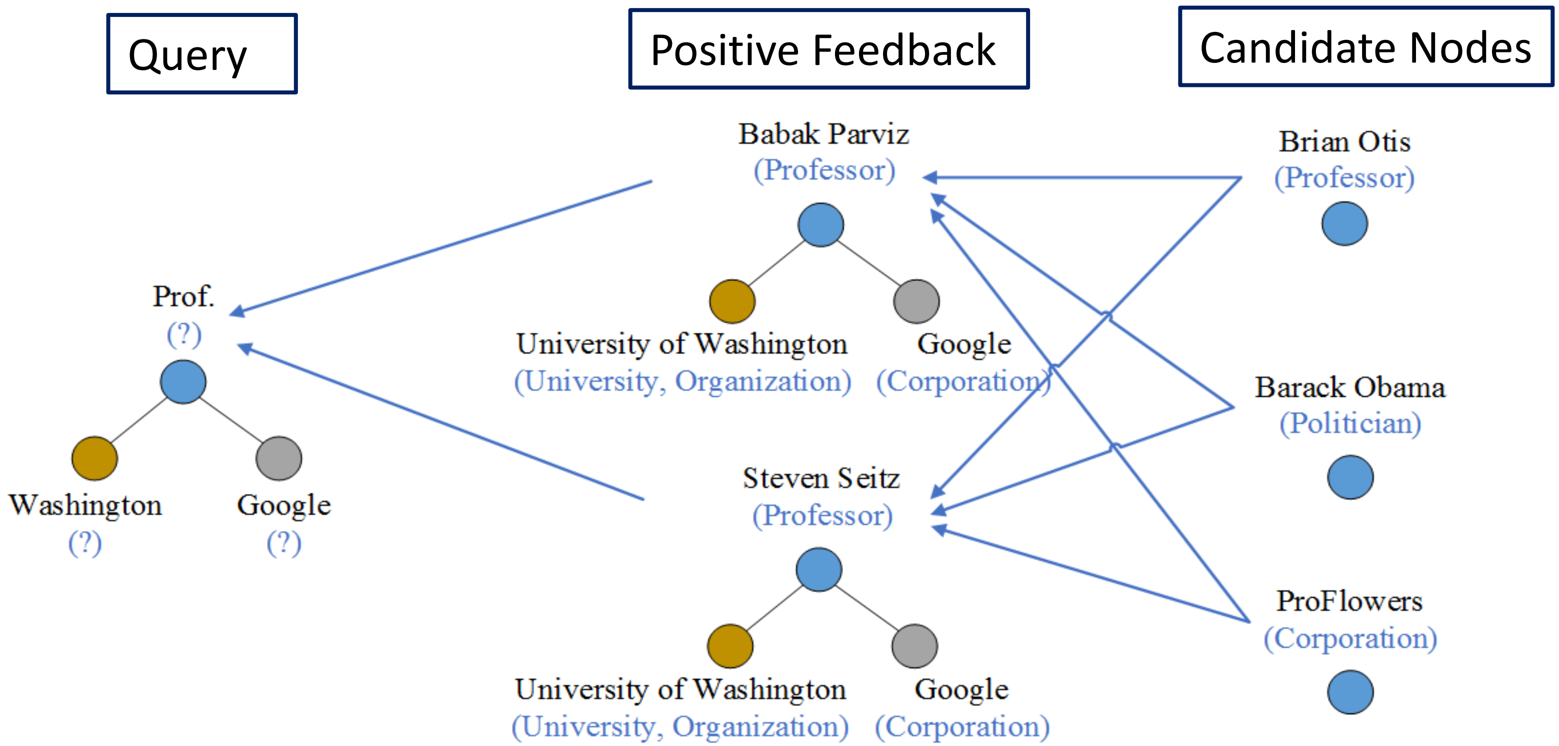
Query-specific Tuning

- The 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

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

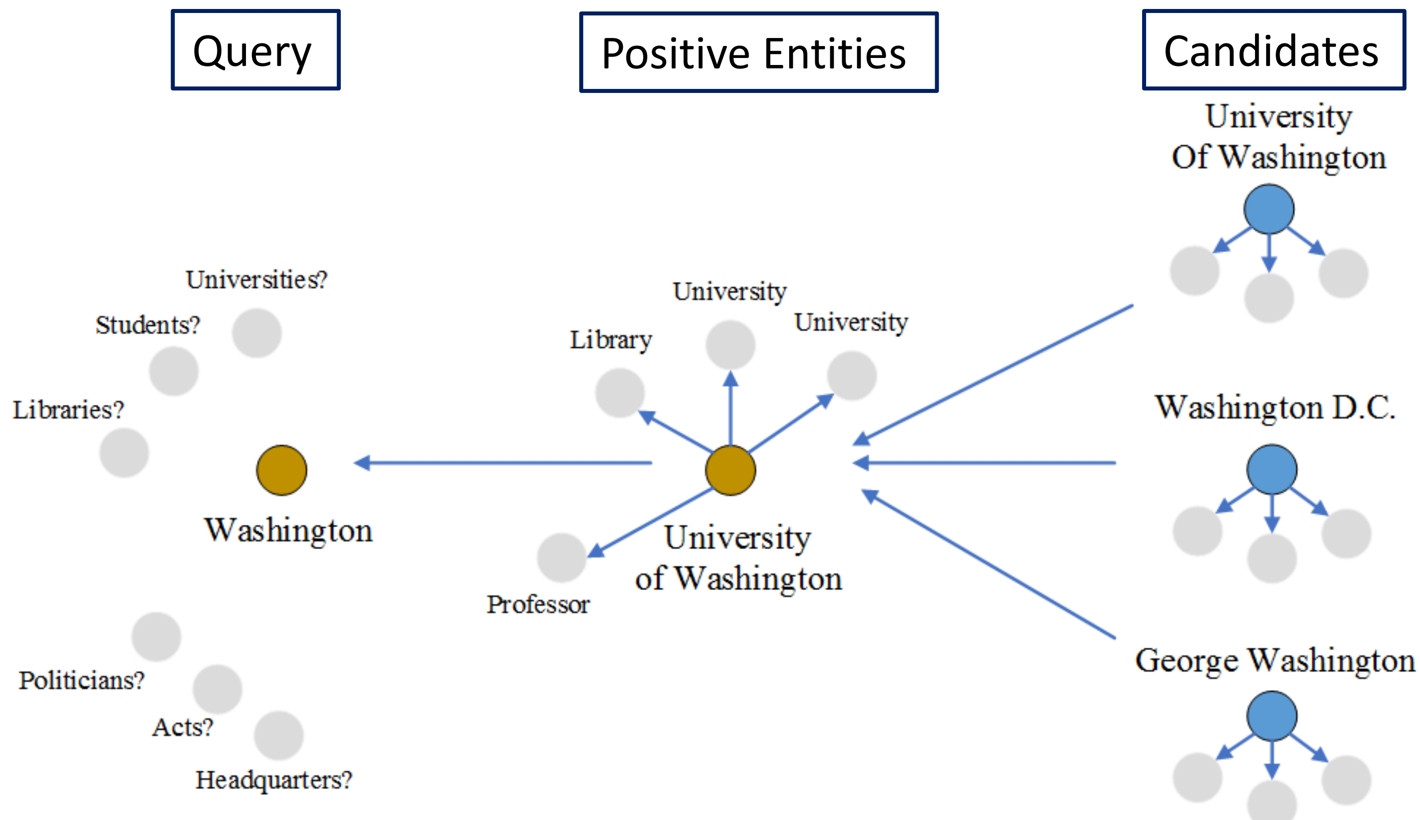
Type Inference

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



Context Inference

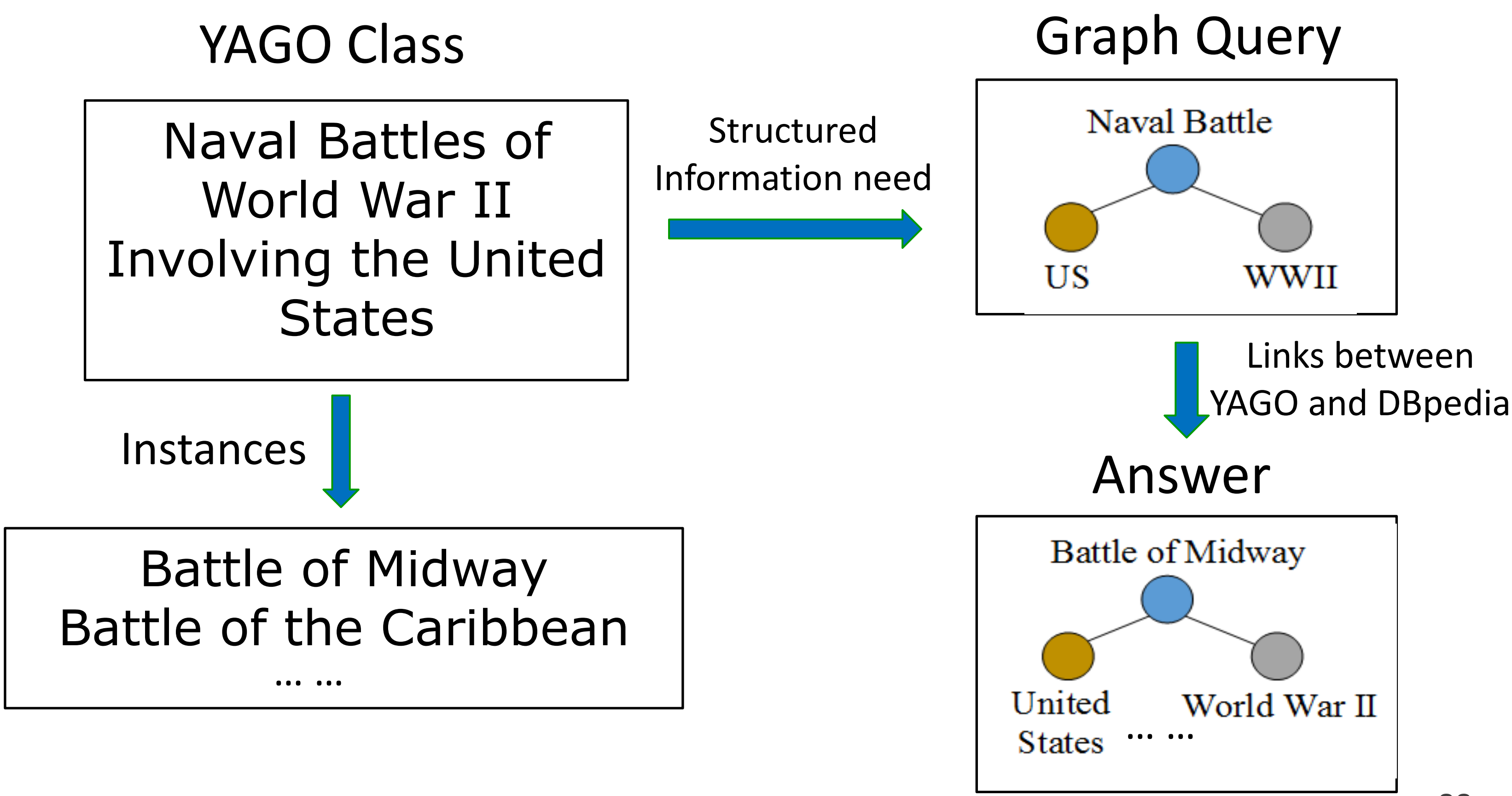
- *Entity context*: neighborhood of the entity
- The contexts of the positive entities constitute a composite context for each query node



[Su et al. KDD'15]

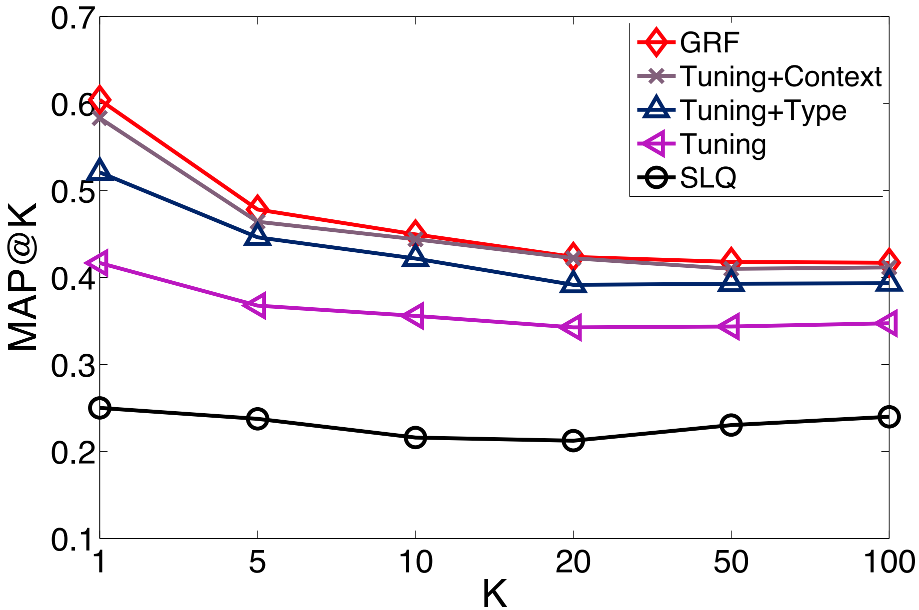
Experiment Setup

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

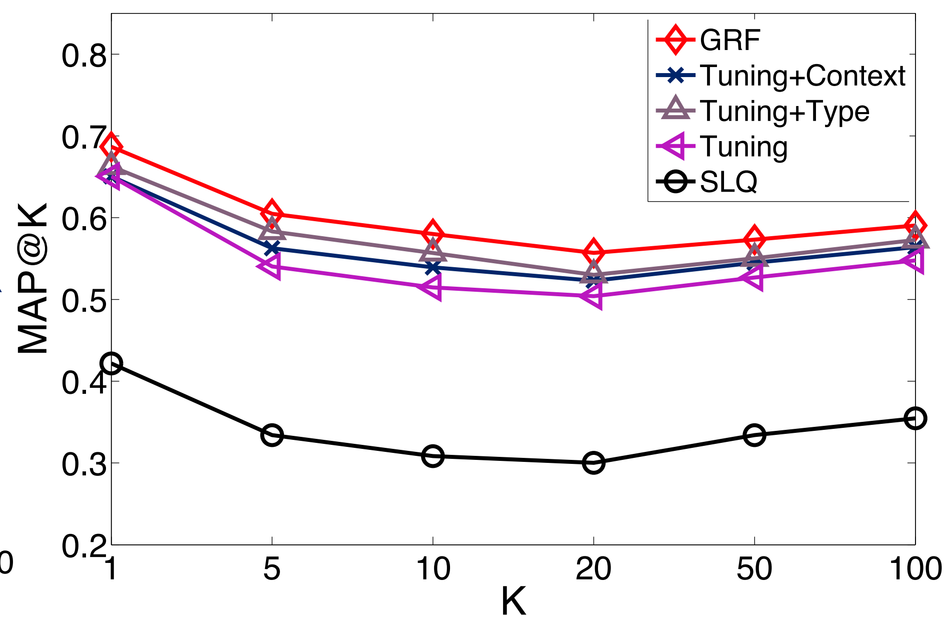


[Su et al. KDD'15]

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



(a) WIKI



(b) YAGO

Metric: mean average precision (MAP)

- Pseudo feedback: Blindly assume top-10 results are correct
- Erroneous feedback information but no additional 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**

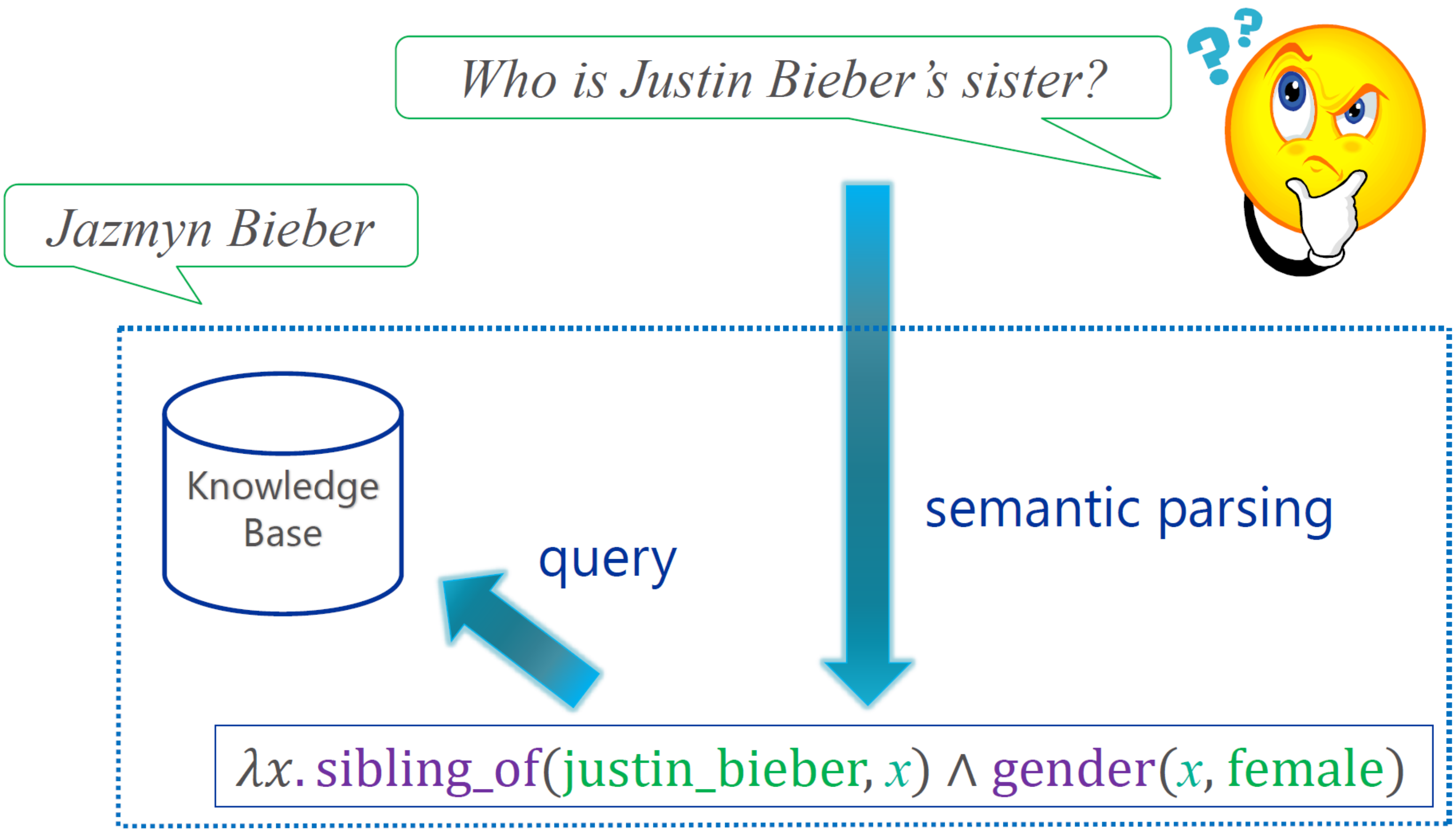


Figure credit to Scott Yih

Challenges

- Language mismatch
  - Lots of ways to ask the same question
    - *Find terrorist organizations involved in September 11 attacks*
    - *Who did September 11 attacks?*
    - *The nine eleven were carried out with the involvement of what terrorist organizations?*
  - All need to be mapped to the KB relation: `terrorist_attack`

Challenges

- Language mismatch
- Large search space
  - United_States has over 1 million neighbors in Freebase

Challenges

- Language mismatch
- Large search space
  - United_States has over 1 million neighbors in Freebase
- Scalability
  - How to scale up to more advanced inputs, and scale out to more domains?
  - KBQA data is highly domain-specific

Challenges

- Language mismatch
- Large search space
  - United_States has over 1 million neighbors in Freebase
- Scalability
  - How to scale up to more advanced inputs, and scale out to more domains?
  - KBQA data is highly domain-specific
- Compositionality
  - If a model understands relation A and B, can it answer A+B?

What will be covered

- Model
  - General pipeline
  - Semantic matching: CNN and Seq2Seq
- Data
  - Low-cost data collection via crowdsourcing
  - Cross-domain semantic parsing via neural transfer learning

General pipeline

Topic Entity Linking

Candidate Logical Form Generation

Seq2Seq:
[Jia and Liang, ACL'16]
[Liang et al. ACL'17]

Semantic Matching

CNN:
[Yih et al. ACL'15]
Seq2Seq:
[Su and Yan, EMNLP'17]

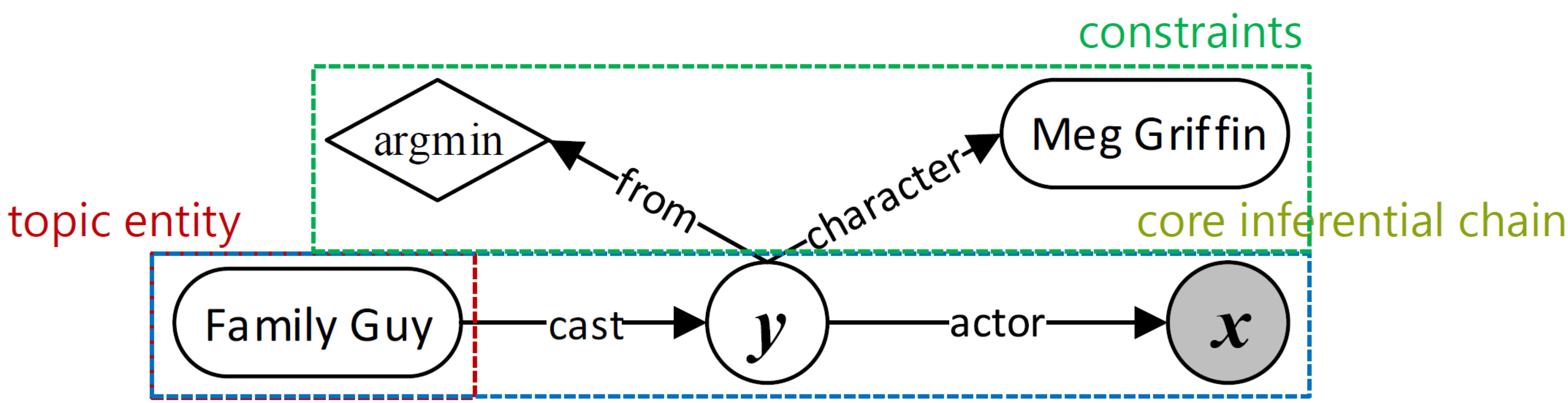
Execution

[Yih et al. ACL'15]

Query Graph

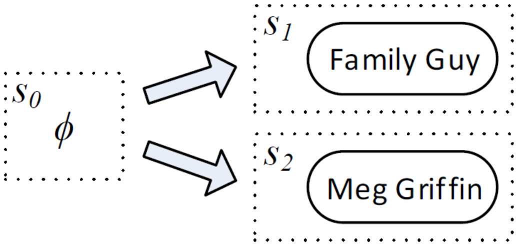
Who first voiced Meg on Family Guy?

$$\lambda x.\exists y.\,\mathrm{cast}(\mathrm{FamilyGuy}, y) \wedge \mathrm{actor}(y, x) \wedge \mathrm{character}(y, \mathrm{MegGriffin})$$



Slides adapted from Scott Yih

Topic Entity Linking

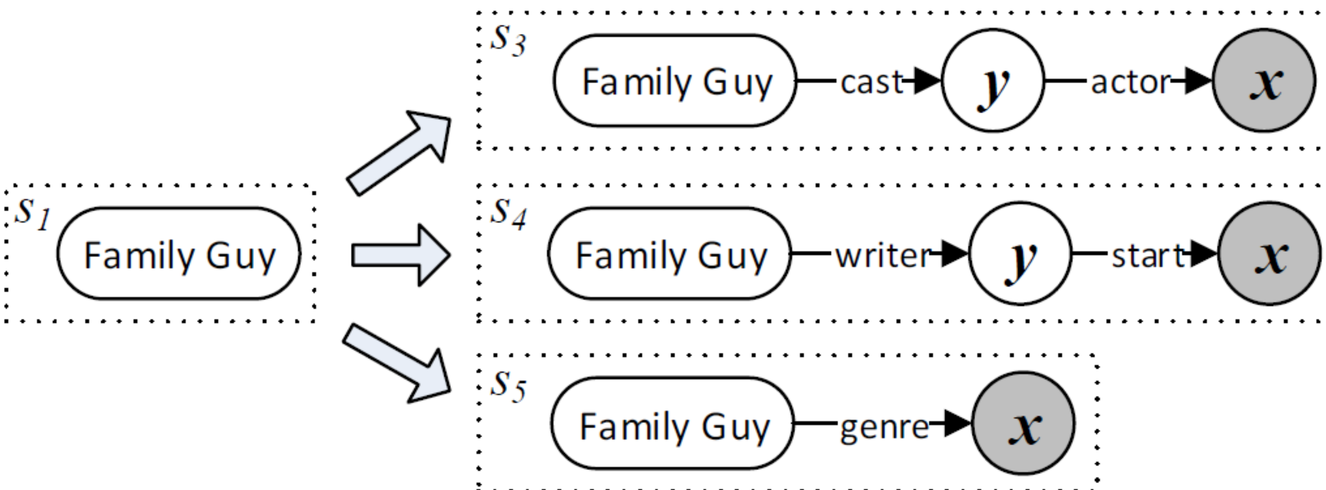


- An advanced entity linker for short text
  - Yang and Chang, "S-MART: Novel Tree-based Structured Learning Algorithms Applied on Tweet Entity Linking." ACL'15
- Prepare surface form lexicon for KB entities
- Entity mention candidates: all consecutive word sequences in lexicon
- Score entity mention candidates with the statistical model, keep top-10 entities

[Yih et al. ACL'15]

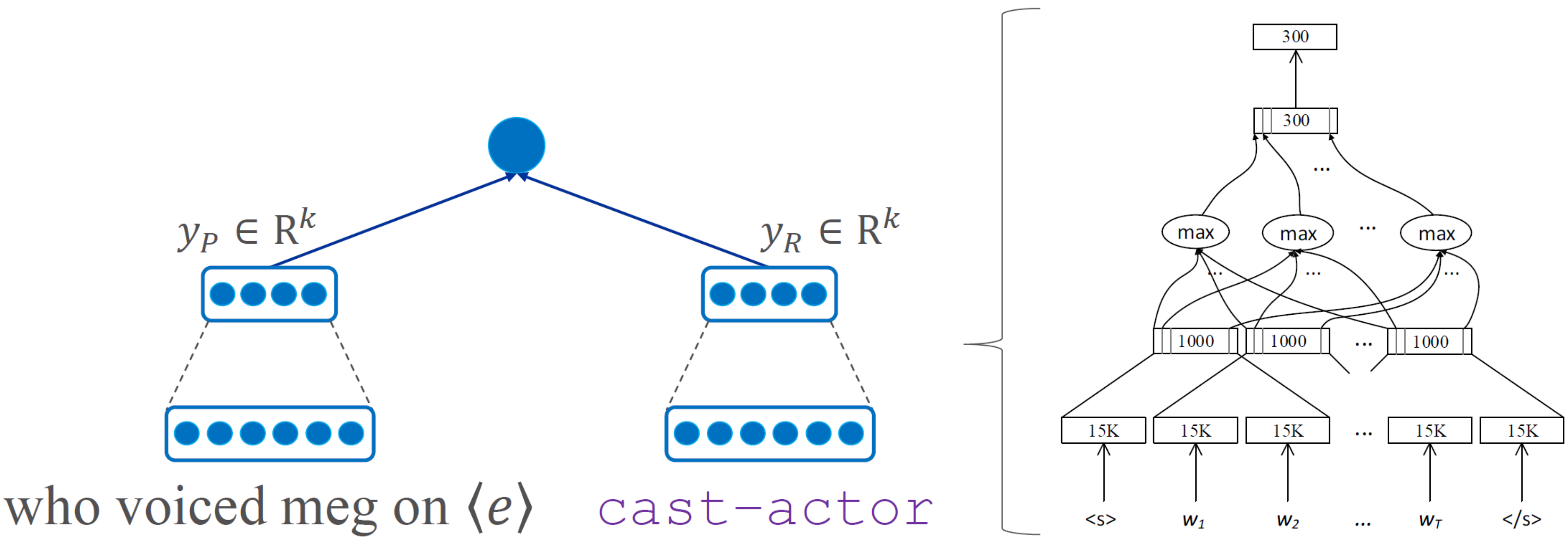
Candidate Logical Form Generation

- (Roughly) *enumerate* all admissive logical forms up to a certain complexity (2-hop)



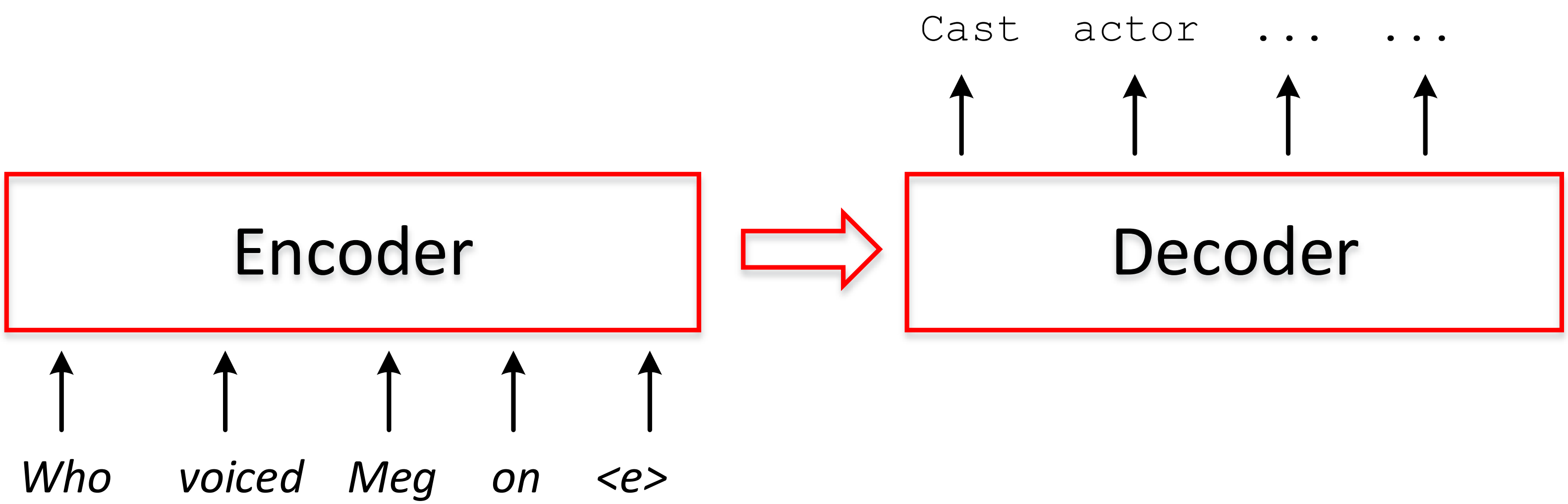
[Yih et al. ACL'15]

Discriminative model: $\mathrm{p}(R|P) = \frac{\exp(\cos(y_R, y_P))}{\sum_{R'} \exp(\cos(y_{R'}, y_P))}$



[Jia+ ACL'16, Liang+ ACL'17, Su+ EMNLP'17]

Generative model: $\mathrm{p}(R|P) = \prod_i p(R_i|P, R_{<i})$



What will be covered

- Model
  - General pipeline
  - Semantic matching: CNN and Seq2Seq
- Data
  - Low-cost data collection via crowdsourcing
  - Cross-domain semantic parsing via neural transfer learning

Scalability

- Vertical scalability
  - Scale up to more complex inputs and logical constructs

Who was the head coach when Michael Jordan started playing for the Chicago Bulls?

↑

In which season did Michael Jordan get the most points?

↑

What team did Michael Jordan play for?

Scalability

- Vertical scalability
  - Scale up to more complex inputs and logical constructs
- Horizontal scalability
  - Scale out to more domains
  - Weather, calendar, hotel, flight, restaurant, …
  - Knowledge base, relational database, API, robots, …
  - Graph, table, text, image, audio, …
- More **data** + Better (more data-efficient) model

On Generating Characteristic-rich Question Sets for QA Evaluation (EMNLP'16)
Cross-domain Semantic Parsing via Paraphrasing (EMNLP'17)
Building Natural Language Interfaces to Web APIs (CIKM'17)

Low-cost Data Collection via Crowdsourcing

"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

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

1: Logical form generation

[Wang+ ACL'15, Su+ EMNLP'16, Su+ CIKM'17]

Existing KBQA datasets mainly contain *simple questions*

"Where was Obama born?"

"What party did Clay establish?"

"What kind of money to take to bahamas?"

… …

GraphQuestions: A New KBQA Dataset with Rich Characteristics

- Structural complexity
  - *“people who are on a gluten-free diet can’t eat what cereal grain that is used to make challah?”*
- Quantitative analysis (functions)
  - *“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?”*
- …

https://github.com/ysu1989/GraphQuestions

Model	Average F1 (%)
Sempre (Berant+ EMNLP'13)	10.8
Jacana (Yao+ ACL'14)	5.1
ParaSempre (Berant+ ACL'14)	12.8
UDepLambda (Reddy+ EMNLP'17)	17.6
Para4QA (Li+ EMNLP'17)	20.4

[Su+ EMNLP'17]

Crowdsourcing is great, but…

- There is an unlimited number of application domains; prohibitive cost to collect (sufficient) training data for every one.

- **Transfer learning**: Use existing data of some source domains to help target domain

- **Problem**: KBQA data is highly domain-specific

[Su+ EMNLP’17]

What is transferrable in semantic parsing?

In which season did Kobe Bryant play for the Lakers?

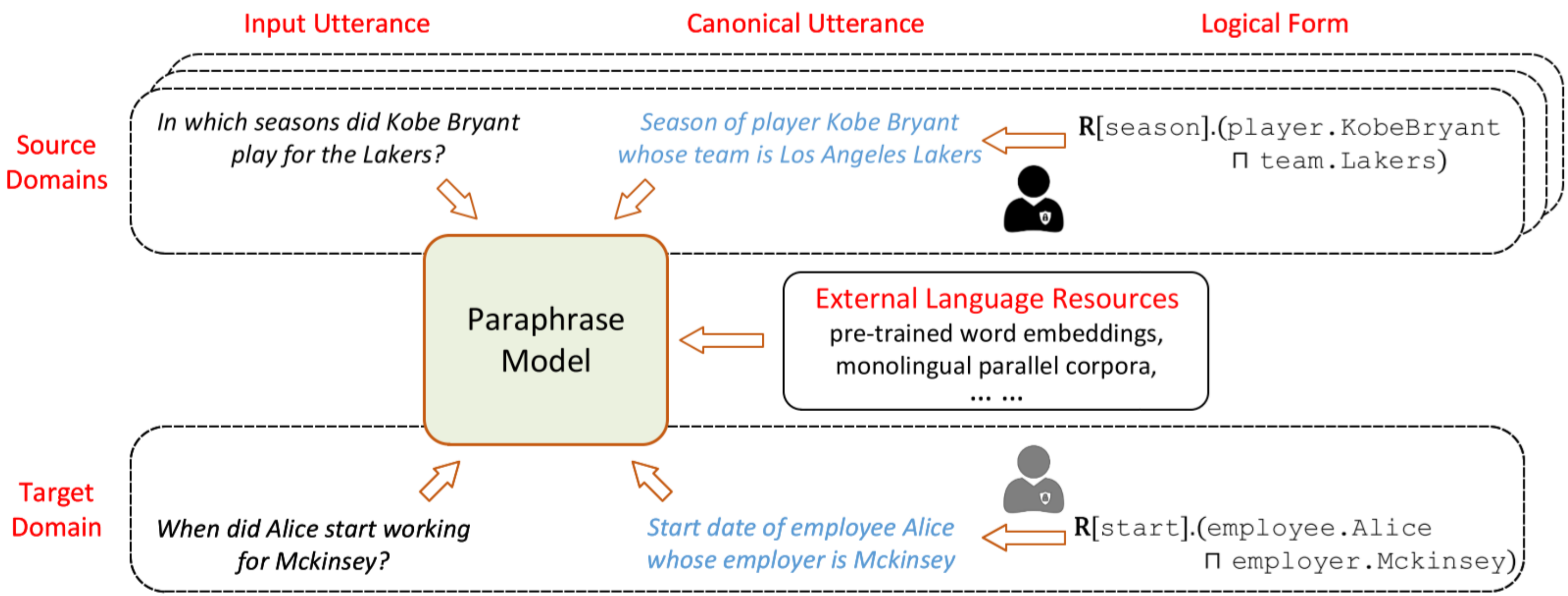
$\mathbf{R}[\texttt{season}].(\texttt{player.KobeBryant} \sqcap \texttt{team.Lakers})$

$p(\texttt{team}|\text{"}\textit{play for}\text{"})$

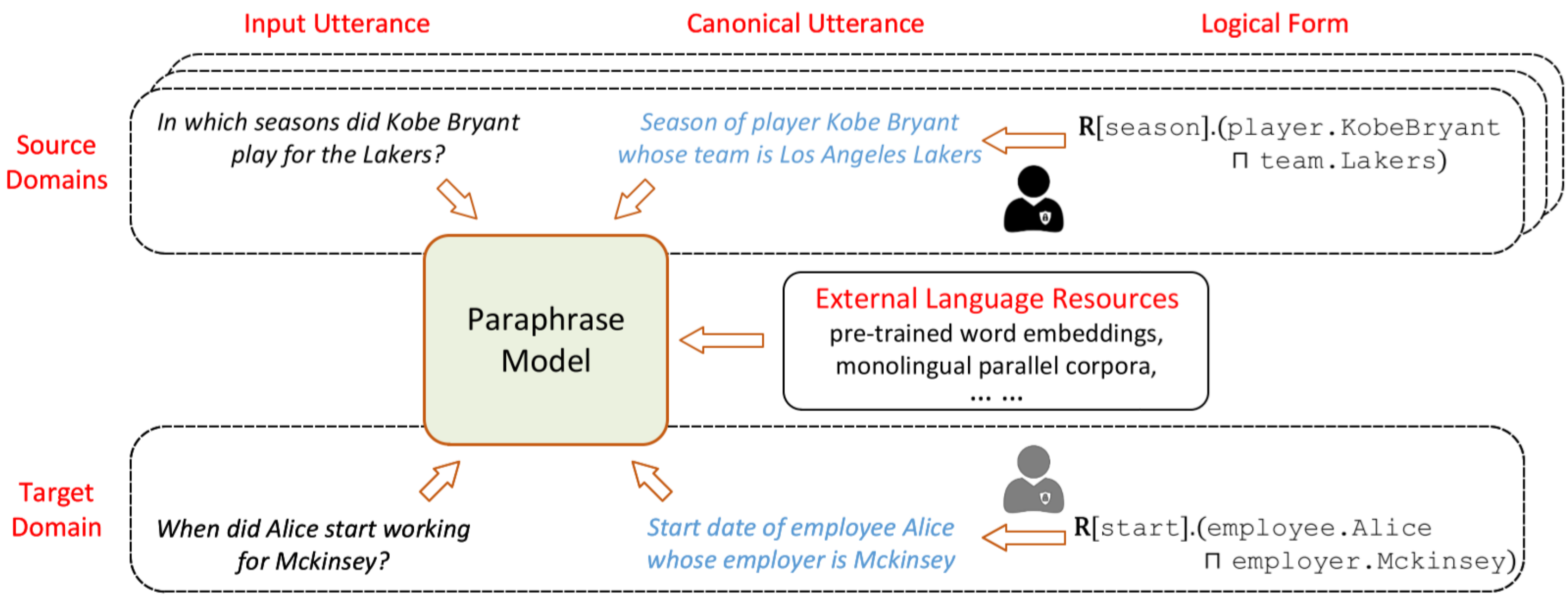
When did Alice start working for Mckinsey?

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

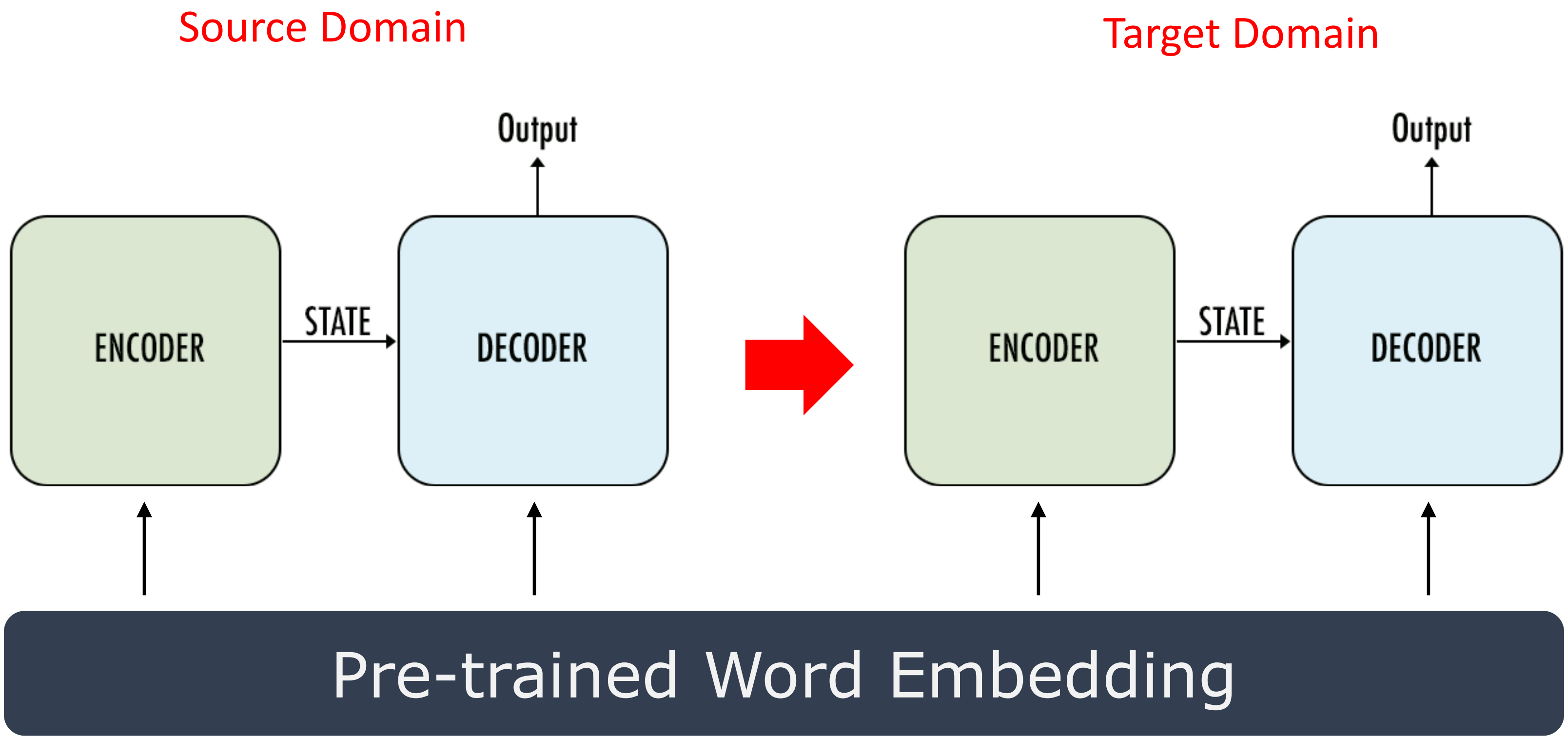
- First convert logical forms to canonical utterances
- Train a neural paraphrase model on the source domains; adapt the model to the target domain



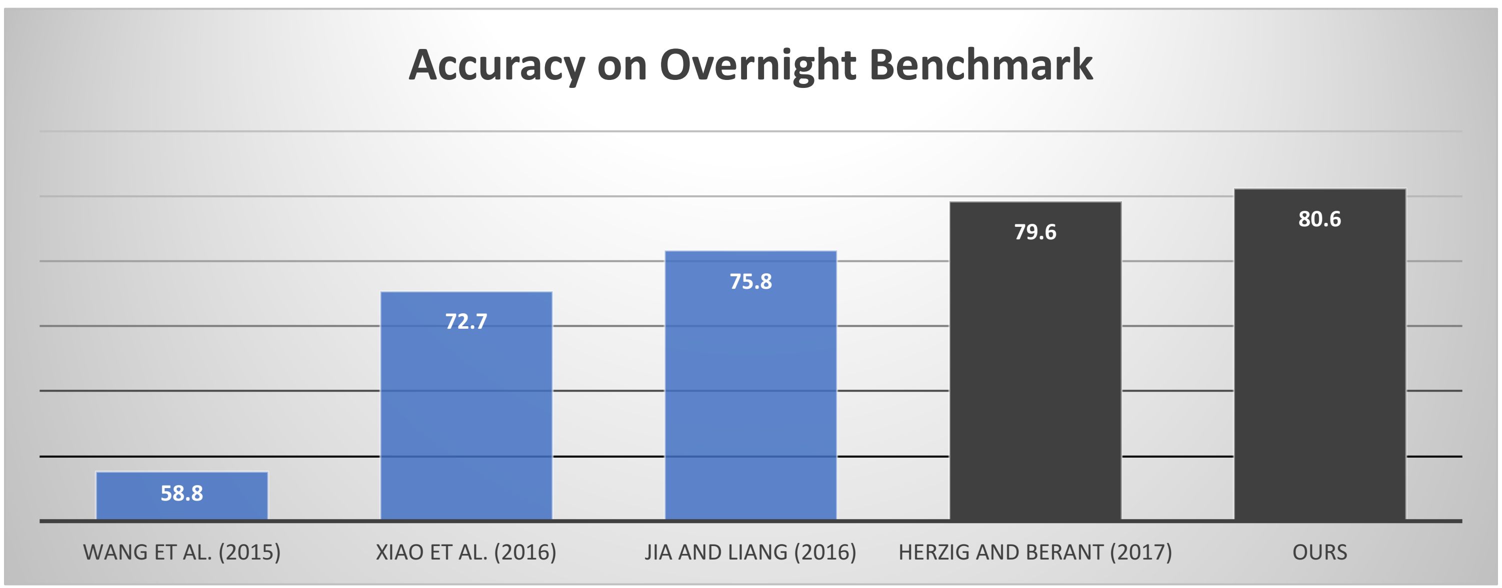
- Source domain: “play for” ⇒ “whose team is”
- Word embedding: “play” ⇒ “work”, “team” ⇒ “employer”
- Target domain: “work for” ⇒ “whose employer is”



Neural Transfer Learning for Semantic Parsing

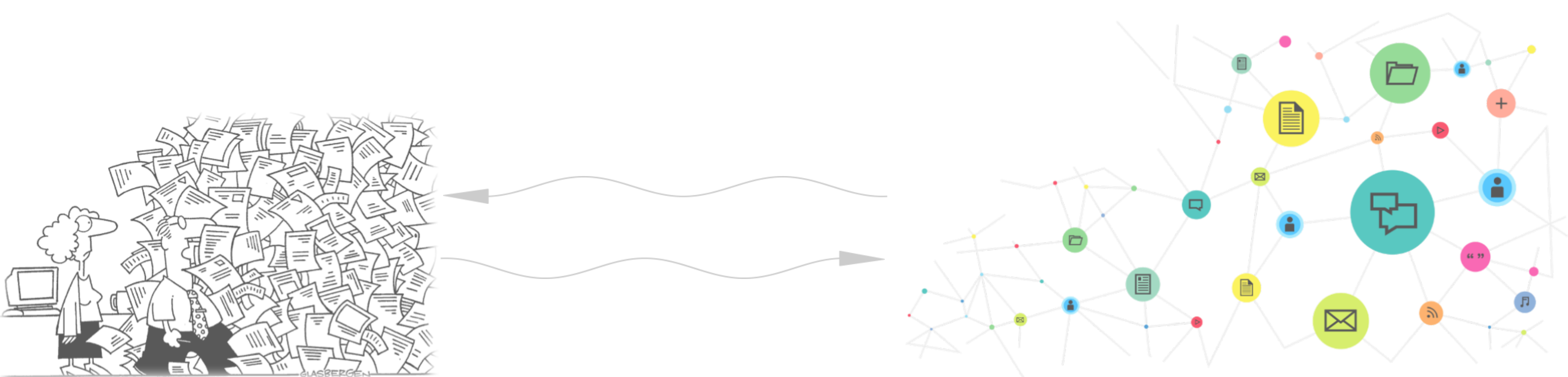


- Overnight dataset: 8 domains (basketball, calendar, etc.), each with a knowledge base
- For each target domain, use other 7 domains as source



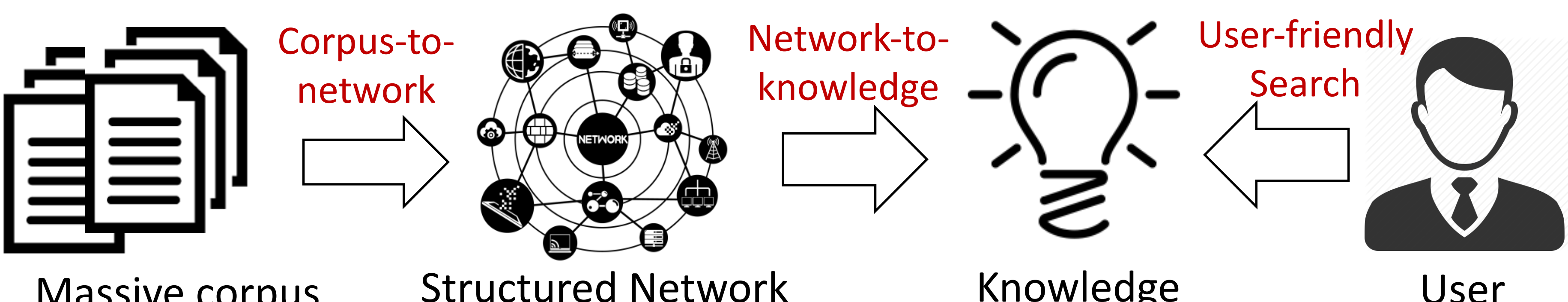
Construction and Querying of Large-scale Knowledge Bases

Summary



Overall Contributions

- **Effort-Light StructMine**: “accurate” expansion of “matchable”
  → Corpus-specific labeling free, domain/language-independent
- **Schema-agnostic Query**: query without programming
- Technology Transfer:
- A principled approach to manage, explore, analyze, and search “Big Text Data”

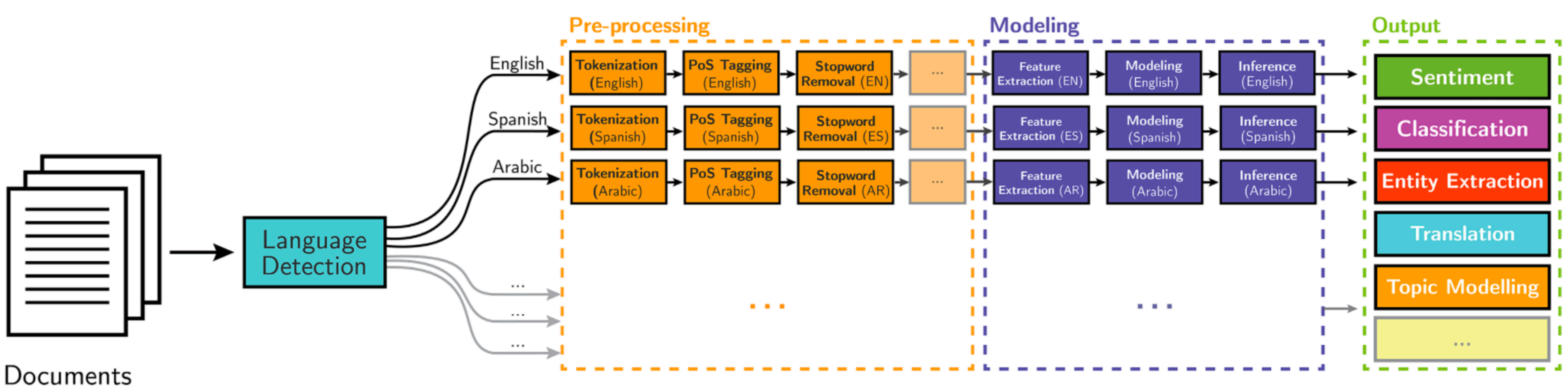


Future Work: Phrase Mining

- Refine quality phrases to entity mentions
- Further use the refined entity mention results to improve phrase mining
- Use high-quality phrases in different languages to improve the entity tagging

Future Work: Phrase Mining

- For popular languages with sufficient NLP tools
  - Incorporate more NLP features and structures
- For low-/zero- resource languages
  - Better unsupervised method

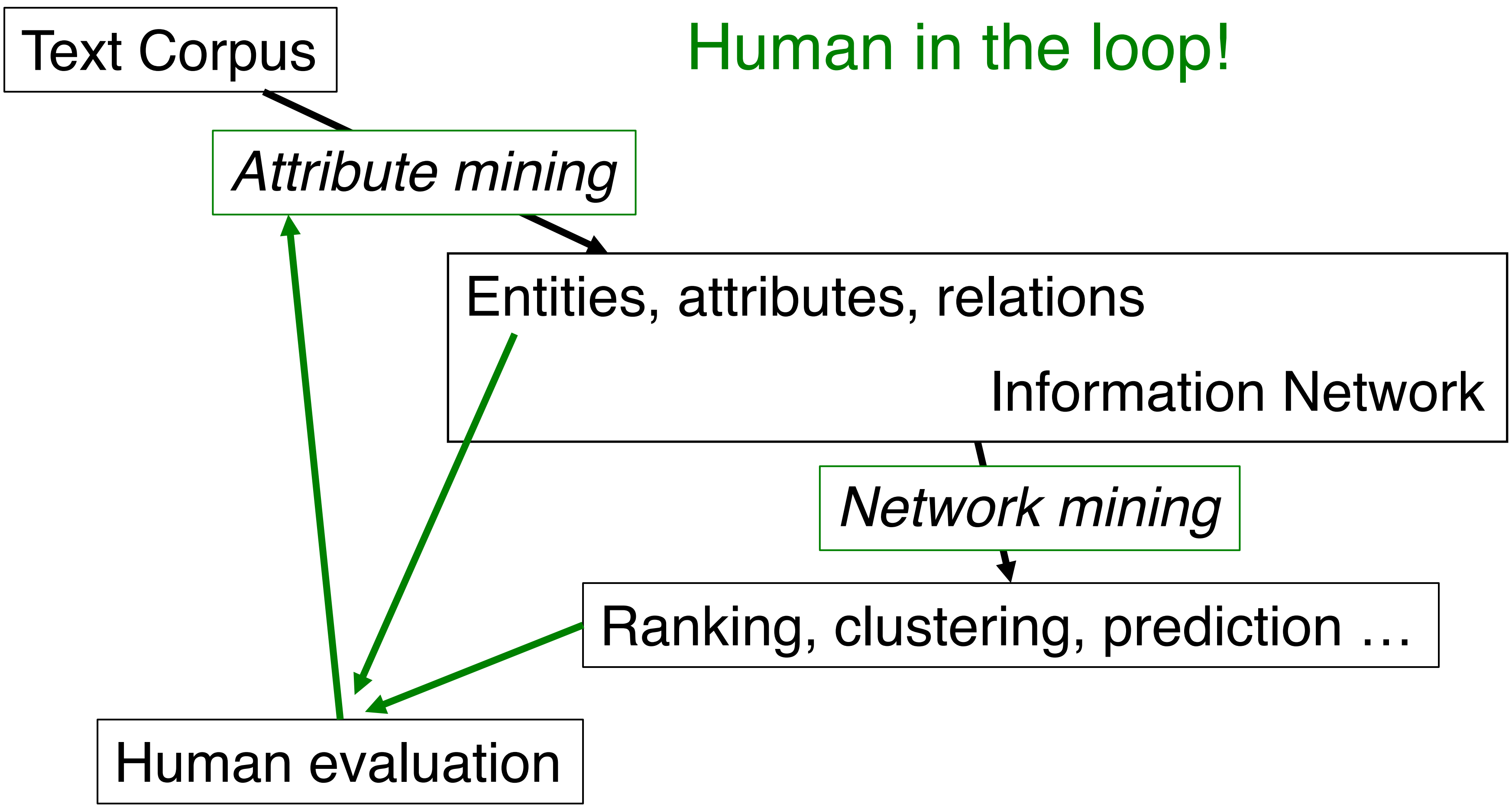


Future Work: Attribute Discovery

- Combining complementary methods towards attribute discovery from massive text corpora
  - Learning Approaches
  - Linguistic patterns using POS tagging, NP chunking, clause analysis, dependency parsing …
  - Meta pattern-driven approaches
    - Harnessing entity recognition and (fine-grained) typing systems
    - Quality assessment and meta-pattern segmentation based on contexts
    - Grouping synonymous patterns
    - Adjusting type levels for appropriate granularity

Future Work: Attribute Discovery

- Combining network mining and attribute mining



Looking Forward: **What's Next?**

Text corpus

Data-driven text segmentation
(SIGMOD'15, WWW'16, ...)

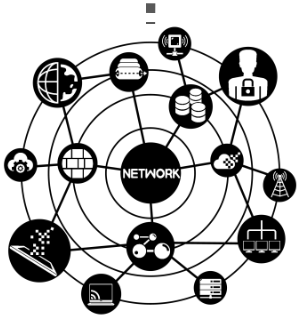
entity names & context units

Knowledge bases

Partially-labeled corpus

Learning Corpus-specific Model
(KDD'15, KDD'16, EMNLP'16, WWW'17, ...)

Structures from the remaining *unlabeled* data



Structured Network

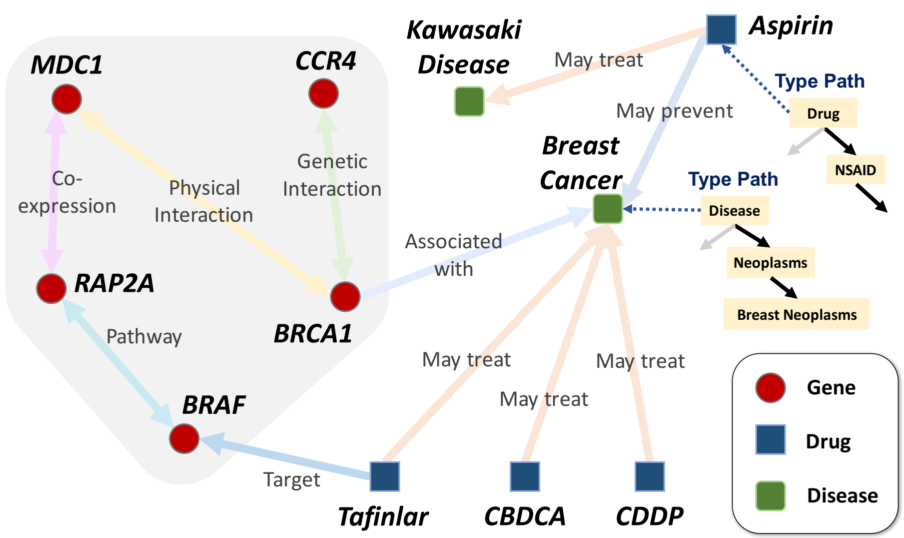
Network-to-knowledge

?

Knowledge & Insights

Looking Forward: Analyzing Literature to **Facilitate Scientific Research**

- Literature → Structured Network ?→ Scientific Discovery
- More disciplines & More structure analysis functions



Scientific Hypothesis Generation by predicting missing relationships

Gaining insights for various research tasks in different disciplines

Collaborate with life scientists, chemists, physicists, computer scientists, …

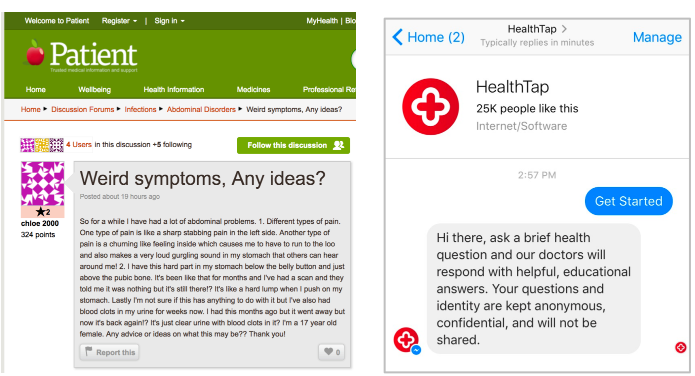
Looking Forward: Engaging with **Human Behaviors**

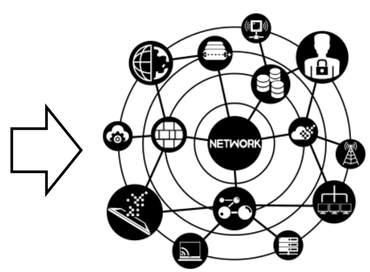
User-generated Content **(Structured Network)** + Structured Behavior Data → ? Personalized Intelligent Systems

Social media post,
Customer review,
Chats & messages

Social network,
Electronic health record,
Transaction record

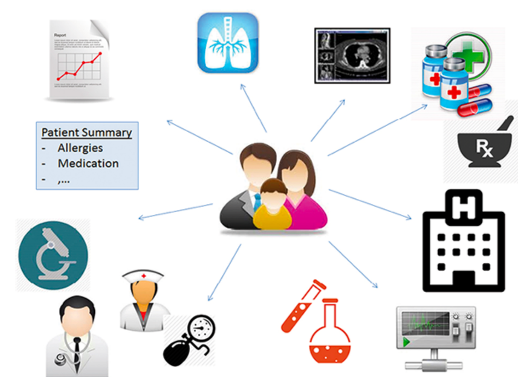
Smart Health,
Business intelligence,
Conversational agent





User content to structured network

\+



Structured information from eHealth records

Collaborate with doctors, social scientists, economists, …

Looking Forward: Integrating with **Our Physical World**

Textual Signals
(Structured Network)

+

Physical Sensor
(Network) Signals

?

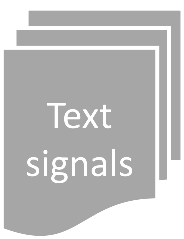
Data Analytics

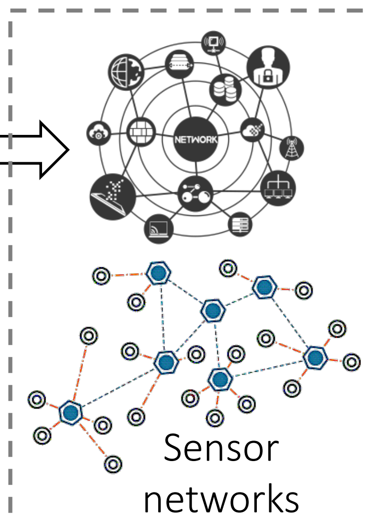
Smart-City
Operating
Systems

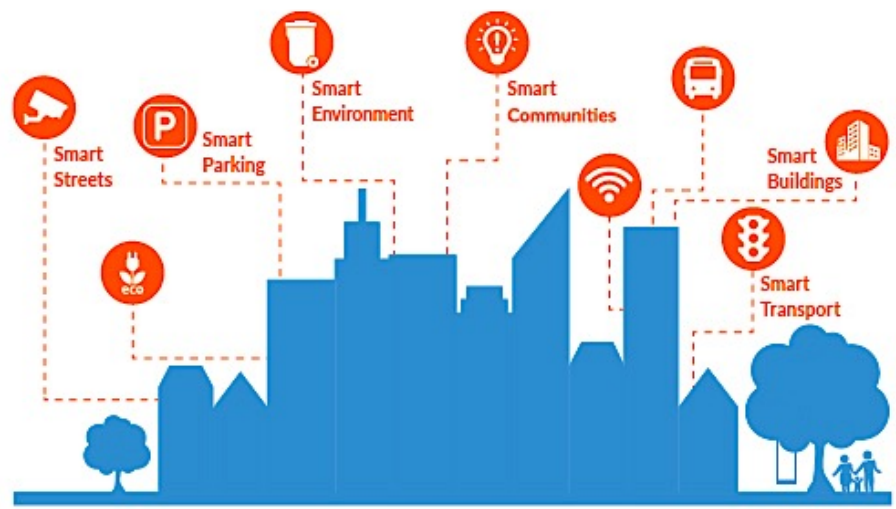
News streams,
Social media post,
Organization report

Geo-sensors
Audio/video sensors
Bio-sensors

Traffic management,
Sustainable urban system,
Cyber-physical system

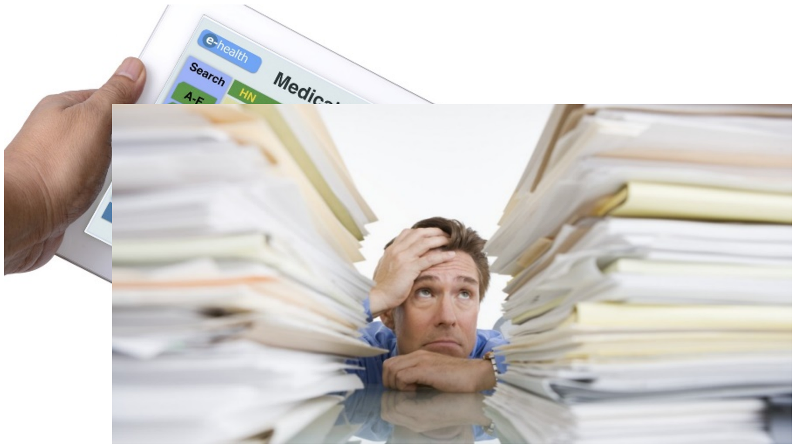






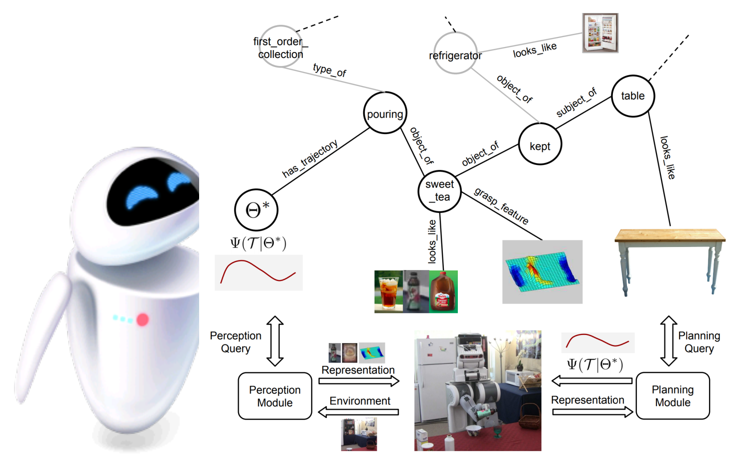
Collaborate with network & system researchers, environmental scientists, …

Application to Vertical Domains



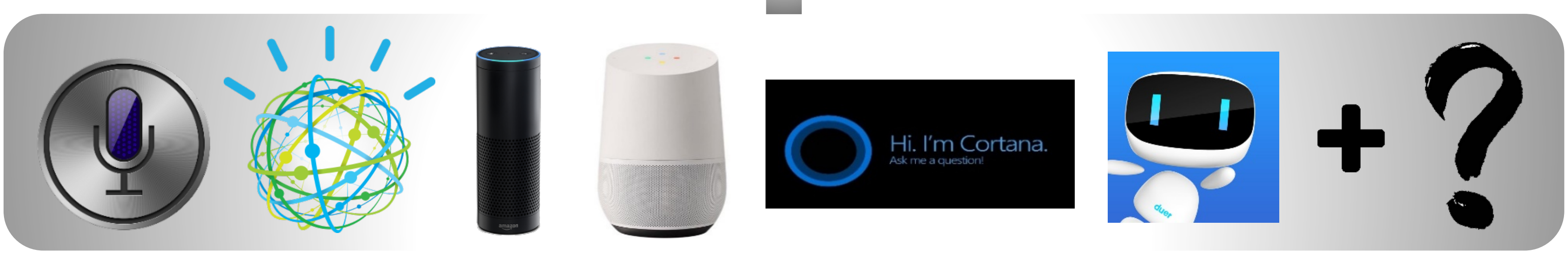
"Which cement stocks go up the most when a Category 3 hurricane hits Florida?"

KENSHO



One Interface for All

- All domains in a unified knowledge base
- Incrementally learn new domains without forgetting (or instead boosting) existing ones

Acknowledgement

- **Academic Collaborators**

  Jiawei Han (UIUC), ChengXiang Zhai (UIUC), Tarek Abdelzaher (UIUC), Aditya Parameswaran (UIUC), Saurabh Sinha (UIUC), Heng Ji (RPI), Yizhou Sun (UCLA), Peipei Ping (UCLA), David Liem (UCLA), Shih-Fu Chang (Columbia), Morteza Dehghani (USC), Richard Weinshilboum (Mayo Clinic), Clare V. Ross (ARL), Lance Kaplan (ARL), James Hendler (RPI), Xifeng Yan (UCSB), Brian Sadler (ARL), Michelle Vanni (ARL), Sue Kase (ARL), Huan Sun (OSU)

- **Industry Collaborators**

  Surajit Chaudhuri (MSR), Kuansan Wang (MSR), Kaushik Chakrabarti (MSR), Chi Wang (MSR), Hao Ma (MSR), Bin Bi (MSR), Yuanhua Lv (Microsoft Bing), Cong Yu (Google Research), Jialu Liu (Google Research), Tao Cheng (Pinterest), Mike Tung (DiffBot), Craig Schmidt (TripAdvisor), Mudhakar Srivatsa (IBM), Ahmed Awadallah (MSR), Patrick Pantel (MSR), Michael Gamon (MSR), Scott Yih (AI2)

- **Funding** ARL NSF NIH Google Microsoft IBM

Thank you! Q&A

- **Effort-Light StructMine**: "accurate" expansion of "matchable"
  → Corpus-specific labeling free, domain/language-independent
- **Schema-agnostic Query**: query without programming
- Technology Transfer:
- A principled approach to manage, explore, analyze, and search "Big Text Data"