On the Power of Big Data: Mining Structures from Massive, Unstructured Text Data

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

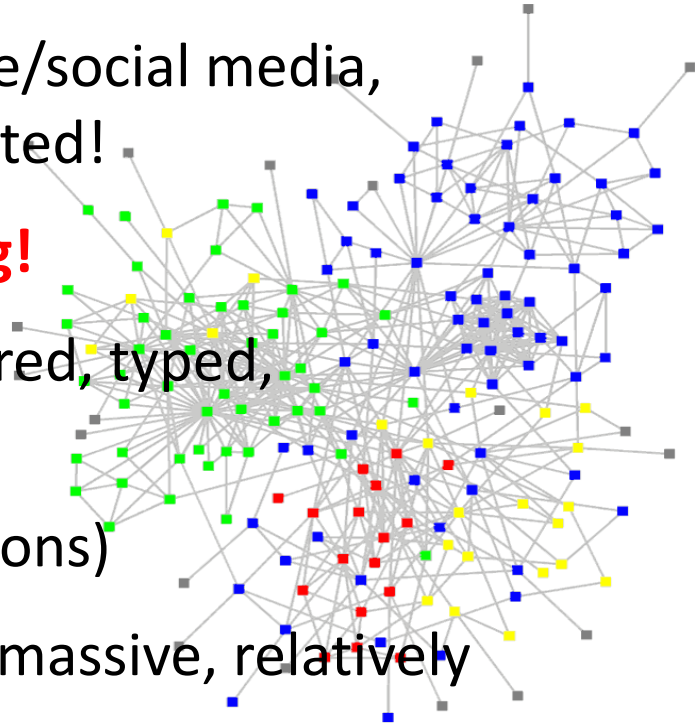
DECEMBER 7, 2016

Outline

- Mining Structures from Text: A Data-Driven Approach
- On the Power of Big Data: Structures from Massive Unstructured Text
  - Phrase Mining: ToPMine → SegPhrase → AutoPhrase
  - Entity Resolution and Typing: ClusType → PLE (Refined Typing)
  - Relationship Discovery by Network Embedding
  - LAKI: Latent Keyphrase Inference
- Data to Network to Knowledge: A Path from Data to Knowledge

Ubiquitous Unstructured, Big Data

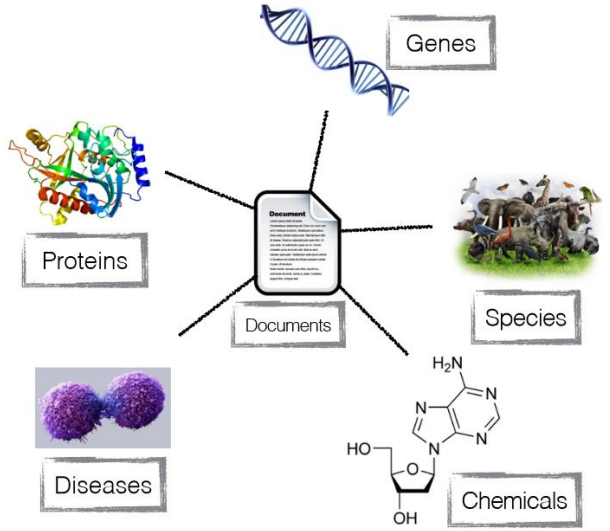
- Ubiquity of big unstructured data
  - Big Data: Over 80% of our data is from text/natural language/social media, unstructured, noisy, dynamic, unreliable, …, but interconnected!
- How to turn big data into big knowledge (BD2K)?—**Structuring!**
  - Structuring (i.e., transforming unstructured text into structured, typed, interconnected entities/relationships)
  - Networking (take advantage of massive, structured connections)
  - Mining/reasoning (e.g., induction/deduction) effectively on massive, relatively structured, interconnected networks
- Thus, our proposal: BD2K → BD2N2K (Big Data to Network to Knowledge)

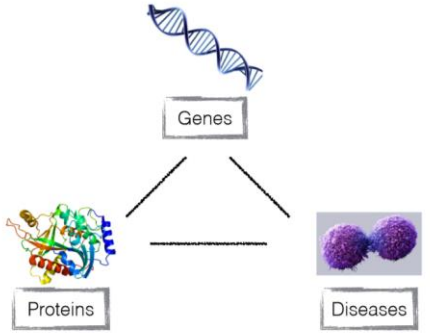


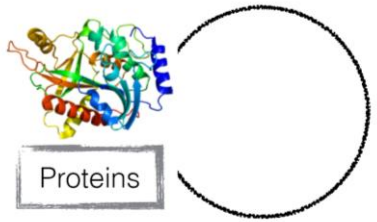
★ → Structuring and mining heterogeneous information networks

Heterogeneous Information Networks

- Heterogeneous networks: Multiple object and link types
  - Medical network: Patients, doctors, diseases, contacts, treatments
  - Bibliographic network: Publications, authors, venues (e.g., DBLP > 2 million papers)
- PubMed is a huge heterogeneous info network
  - Document links to: gene, protein, disease, drug, species, …
  - Rich knowledge can be mined from such networks





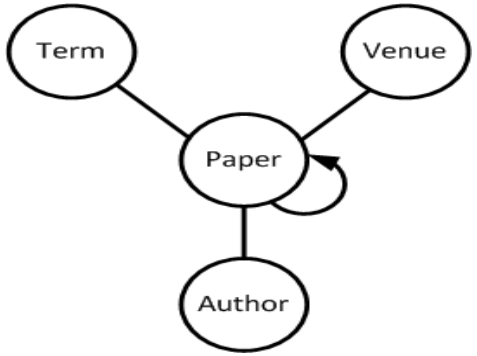


Knowledge hidden in DBLP Network	Mining Functions
Who are the **leading** researchers on Web search?	Ranking
Who are the **peer** researchers of Jure Leskovec?	Similarity Search
Whom will Christos Faloutsos collaborate with?	**Relationship Prediction**
How was the field of Data Mining **emerged** or **evolving**?	Network Evolution
Which authors are **rather different** from his/her peers in IR?	Outlier/anomaly detection

PathPredict: Meta-Path Based Relationship Prediction

- Who will be your new coauthors in the next 5 years?
  - Meta path-guided prediction of links and relationships

- Philosophy: Meta path relationships among similar typed links share similar semantics and are comparable and inferable



- Co-author prediction (A—P—A) [Sun et al., ASONAM'11]
  - Use topological features encoded by meta paths, e.g., citation relations between authors (A—P→P—A)

Meta-paths between authors of length ≤ 4

Meta-Path	Semantic Meaning
$A-P\rightarrow P-A$	a_i cites a_j
$A-P\leftarrow P-A$	a_i is cited by a_j
$A-P-V-P-A$	a_i and a_j publish in the same venues
$A-P-A-P-A$	a_i and a_j are co-authors of the same authors
$A-P-T-P-A$	a_i and a_j write the same topics
$A-P\rightarrow P\rightarrow P-A$	a_i cites papers that cite a_j
$A-P\leftarrow P\leftarrow P-A$	a_i is cited by papers that are cited by a_j
$A-P\rightarrow P\leftarrow P-A$	a_i and a_j cite the same papers
$A-P\leftarrow P\rightarrow P-A$	a_i and a_j are cited by the same papers

The Power of PathPredict: Experiment on DBLP

- Explain the prediction power of each meta-path
  - Wald Test for logistic regression

Evaluation of the prediction power of different meta-paths

- Higher prediction accuracy than using projected homogeneous network
  - **11%** higher in prediction accuracy

Prediction of new coauthors of Jian Pei in [2003-2009]

Meta Path	p-value	significance level[1]
$A-P \rightarrow P-A$	0.0378	**
$A-P \leftarrow P-A$	0.0077	***
$A-P-V-P-A$	1.2974e-174	****
$A-P-A-P-A$	1.1484e-126	****
$A-P-T-P-A$	3.4867e-51	****
$A-P \rightarrow P \rightarrow P-A$	0.7459	
$A-P \leftarrow P \leftarrow P-A$	0.0647	*
$A-P \rightarrow P \leftarrow P-A$	9.7641e-11	****
$A-P \leftarrow P \rightarrow P-A$	0.0966	*

[1] *: $p < 0.1$; **: $p < 0.05$; ***: $p < 0.01$, ****: $p < 0.001$

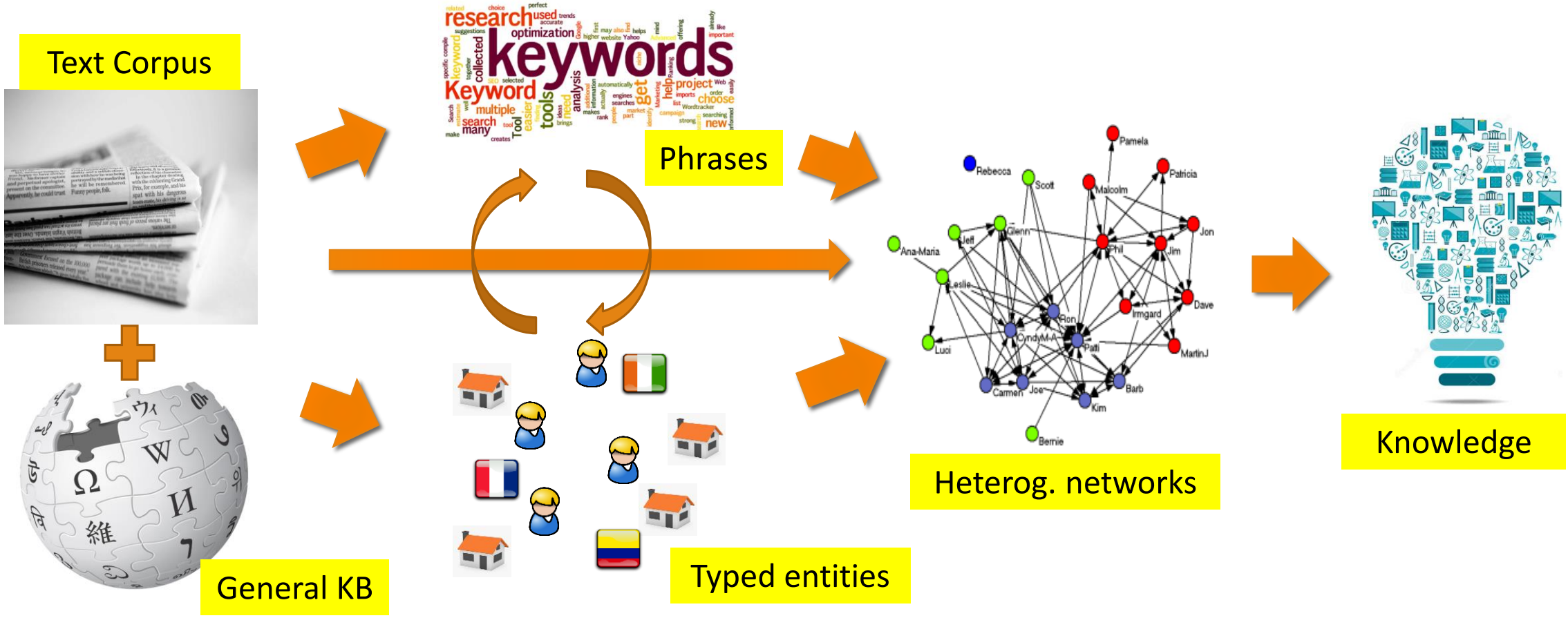
Social relations play more important role?

Rank	Hybrid heterogeneous features	# Shared authors
1	**Philip S. Yu**	**Philip S. Yu**
2	**Raymond T. Ng**	Ming-Syan Chen
3	Osmar R. Zaïane	Divesh Srivastava
4	**Ling Feng**	Kotagiri Ramamohanarao
5	**David Wai-Lok Cheung**	**Jeffrey Xu Yu**

Co-author prediction for Jian Pei: Only 42 among 4809 candidates are true first-time co-authors! (Feature collected in [1996, 2002]; Test period in [2003,2009])

Bottleneck: Where Are the Structured Networks?

- Unfortunately, most of our big data is unstructured text!
- Bottleneck: How to automatically generate structured networks from text data?
  - Automated mining of phrases, topics, entities, links and types from text corpora



Why Data Driven Approach?

- Automatic: Not relying on extensive expert "labeling" and curation!
- Data Driven: Use massive text corpora to "automatically" generate knowledge!
  - Mining phrases from massive text data (Top-Mine + SegPhrase)
  - Entity recognition and typing (ClusType)
  - Relationship extraction (Example: Biological relationship discovery)
  - Latent Keyphrase Inference (LAKI)

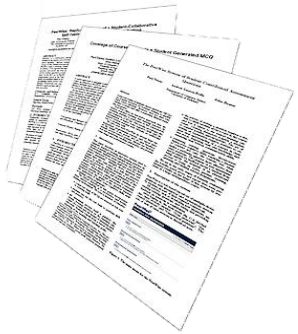
From unstructured data to structured networks:
Minimal human training or curation, maximal structure generation

Outline

- Mining Structures from Text: A Data-Driven Approach
- On the Power of Big Data: Structures from Massive Unstructured Text
  - Phrase Mining: ToPMine → SegPhrase → AutoPhrase
  - Entity Resolution and Typing: ClusType → PLE (Refined Typing)
  - Relationship Discovery by Network Embedding
  - LAKI: Latent Keyphrase Inference
- Data to Network to Knowledge: A Path from Data to Knowledge

Phrase Mining: From Raw Corpus to Quality Phrases and Segmented Corpus

Raw Corpus



Quality Phrases

Segmented Corpus

Document 1
Citation recommendation is an interesting but challenging research problem in data mining area.

Document 2
In this study, we investigate the problem in the context of heterogeneous information networks using data mining technique.

Document 3
Principal Component Analysis is a linear dimensionality reduction technique commonly used in machine learning applications.

+ A small set of labels or a general KB

Input Raw Corpus → **Quality Phrases** ↔ **Segmented Corpus**

Phrase Mining **Phrasal Segmentation**

Integrating phrase mining with phrasal segmentation

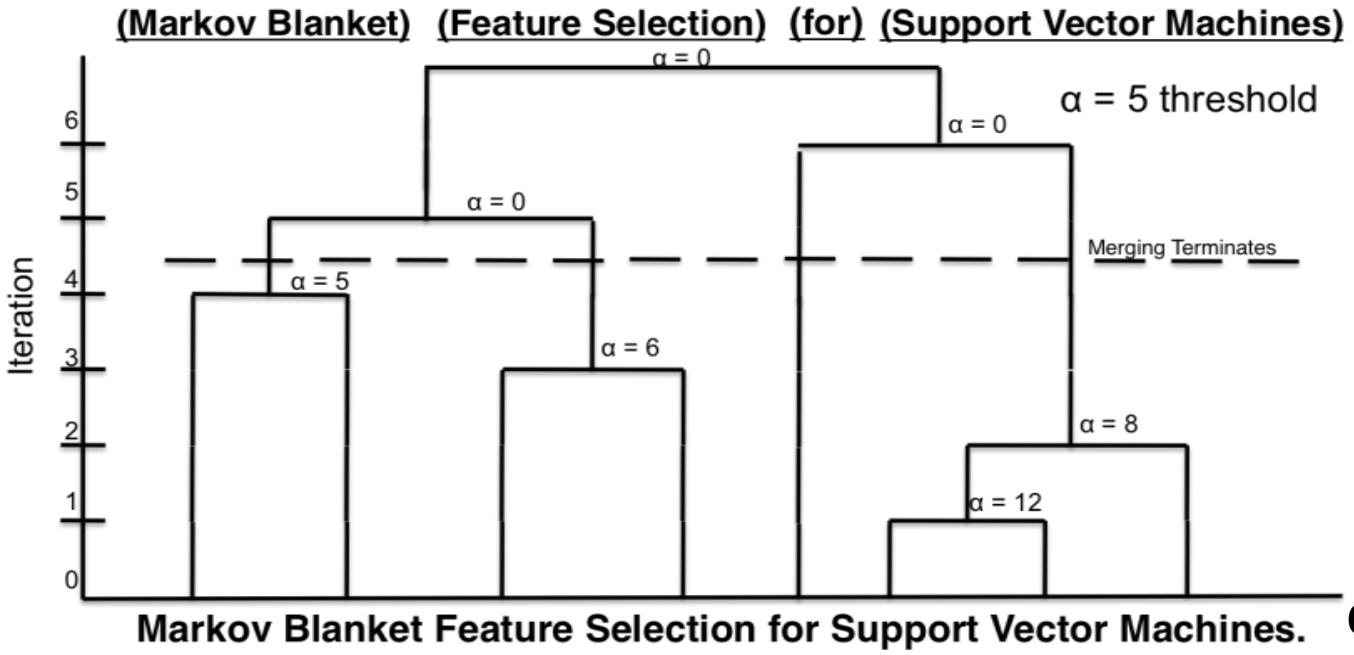
TOPMINE: A. El-Kishky, et al., Scalable Topical Phrase Mining from Text Corpora", in VLDB'15

SegPhrase: J. Liu et al., Mining Quality Phrases from Massive Text Corpora. *SIGMOD*'15 (**Grand Prize in Yelp Dataset Challenge)**

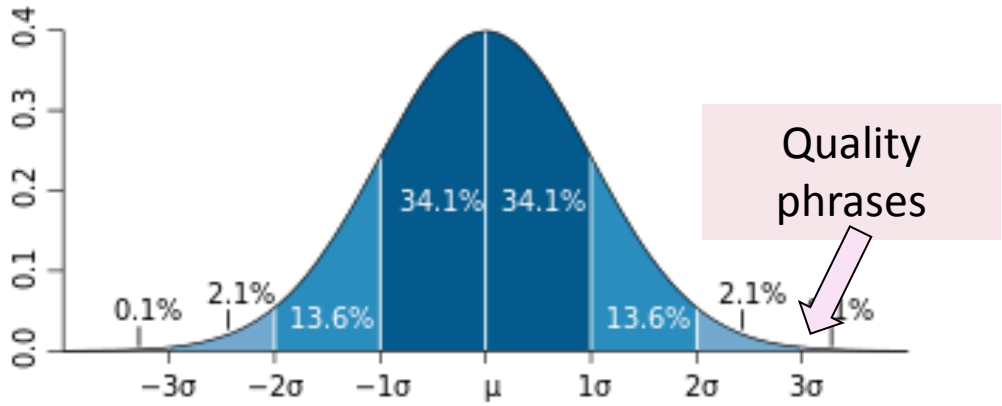
AutoPhrase: J. Shang, et al., AutoPhrase: A Unified Framework for Automated Quality Phrase Mining from Massive Text Corpora, 2016 (under evaluation)

TopMine for Phrase Mining: Frequent Pattern Mining + Statistical Analysis

First perform frequent ***contiguous pattern*** mining to extract candidate phrases and their counts



Markov Blanket Feature Selection for Support Vector Machines.



Based on significance score [Church et al.'91]:

$$\alpha(P_1, P_2) \approx (f(P_1 \bullet P_2) - \mu_0(P_1, P_2))/\sqrt{f(P_1 \bullet P_2)}$$

[Markov blanket] [feature selection] for [support vector machines]
[knowledge discovery] using [least squares] [support vector machine] [classifiers]
…[support vector] for [machine learning]…

Phrase	Raw freq.	True freq.
[support vector machine]	90	80
[vector machine]	95	0
[support vector]	100	20

What Kind of Phrases Are of "High Quality"?

- Judging the quality of phrases
  - **Popularity**
    - "information retrieval" vs. "cross-language information retrieval"
  - **Concordance**
    - "powerful tea" vs. "strong tea"
    - "active learning" vs. "learning classification"
  - **Informativeness**
    - "this paper" (frequent but not discriminative, not informative)
  - **Completeness**
    - "vector machine" vs. "support vector machine"

ToPMine: Experiments on Yelp Reviews

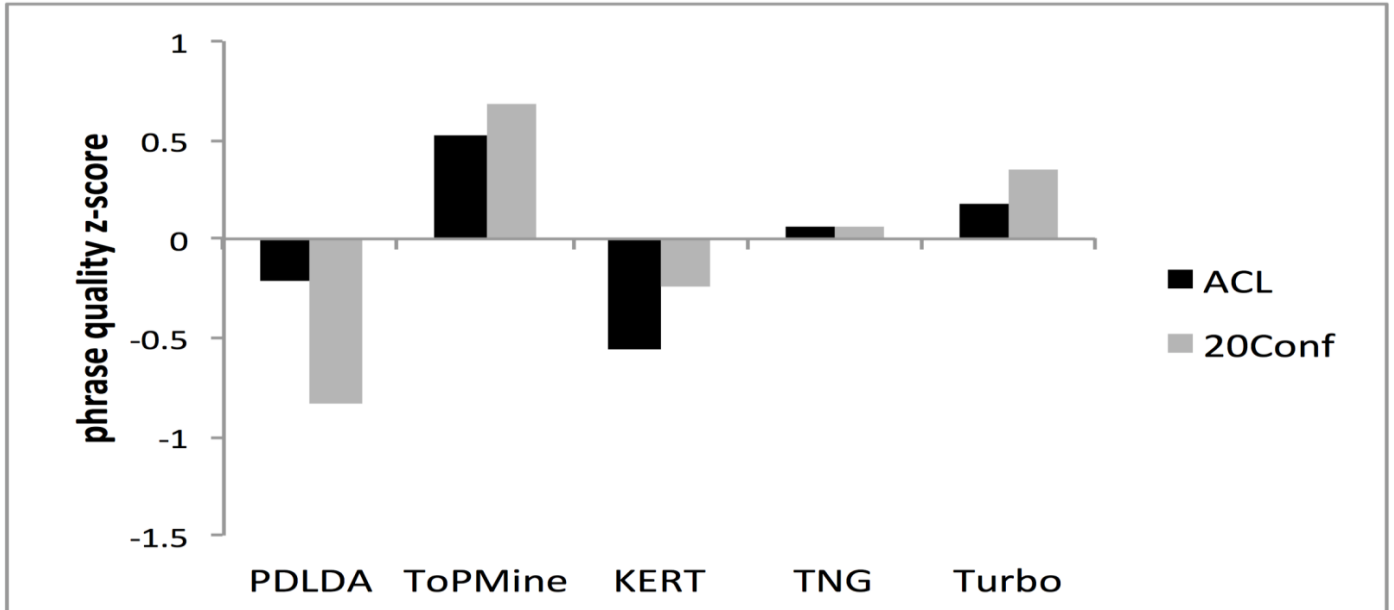
	Topic 1	*Topic 2*	*Topic 3*	*Topic 4*	*Topic 5*
unigrams	coffee	food	room	store	good
	ice	good	parking	shop	food
	cream	place	hotel	prices	place
	flavor	ordered	stay	find	burger
	egg	chicken	time	place	ordered
	chocolate	roll	nice	buy	fries
	breakfast	sushi	place	selection	chicken
	tea	restaurant	great	items	tacos
	cake	dish	area	love	cheese
	sweet	rice	pool	great	time
n-grams	ice cream	spring rolls	parking lot	grocery store	mexican food
	iced tea	food was good	front desk	great selection	chips and salsa
	french toast	fried rice	spring training	farmer's market	food was good
	hash browns	egg rolls	staying at the hotel	great prices	hot dog
	frozen yogurt	chinese food	dog park	parking lot	rice and beans
	eggs benedict	pad thai	room was clean	wal mart	sweet potato fries
	peanut butter	dim sum	pool area	shopping center	pretty good
	cup of coffee	thai food	great place	great place	carne asada
	iced coffee	pretty good	staff is friendly	prices are reasonable	mac and cheese
	scrambled eggs	lunch specials	free wifi	love this place	fish tacos

ToPMine: Faster and Generating Better Quality Phrases

Running time of different algorithms

Method	*sampled dblp titles (k=5)*	*dblp titles (k=30)*	*sampled dblp abstracts*	*dblp abstracts*
PDLDA	3.72(hrs)	~20.44(days)	1.12(days)	~95.9(days)
Turbo Topics	6.68(hrs)	>30(days)*	>10(days)*	>50(days)*
TNG	146(s)	5.57 (hrs)	853(s)	NA†
LDA	**65(s)**	3.04 (hrs)	353(s)	13.84(hours)
KERT	68(s)	3.08(hrs)	1215(s)	NA†
ToP-Mine	67(s)	**2.45(hrs)**	**340(s)**	**10.88(hrs)**

Phrase quality measured by z-score

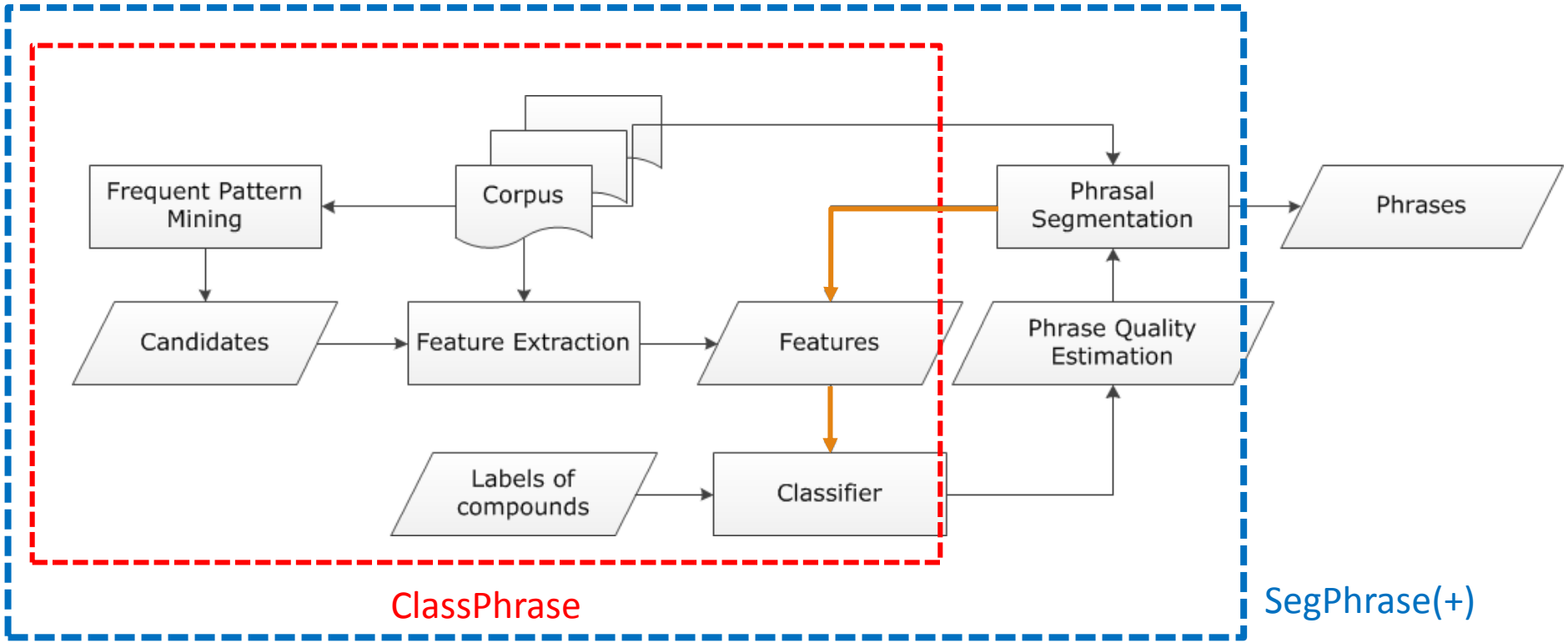


SegPhrase+: The Overall Framework

- ClassPhrase: Frequent pattern mining, feature extraction, classification
- SegPhrase: Phrasal segmentation and phrase quality estimation
- SegPhrase+: One more round to enhance mined phrase quality

SegPhrase (Classifier)

Small labeled dataset
Provided by experts
Or
A distant supervised
KB (Wiki/DBPedia)



Experimental Results: Interesting Phrases Generated (From Titles & Abstracts of SIGKDD)

Query	SIGKDD	
Method	SegPhrase+	Chunking (TF-IDF & C-Value)
1	data mining	data mining
2	data set	association rule
3	association rule	knowledge discovery
4	knowledge discovery	frequent itemset
5	**time series**	decision tree
…	…	…
51	association rule mining	search space
52	rule set	domain knowledge
53	concept drift	**important problem**
54	knowledge acquisition	concurrency control
55	**gene expression data**	conceptual graph
…	…	…
201	web content	optimal solution
202	**frequent subgraph**	semantic relationship
203	intrusion detection	**effective way**
204	**categorical attribute**	space complexity
205	user preference	**small set**
…	…	…

Only in SegPhrase+

Only in Chunking

Mining Quality Phrases in Multiple Languages

- Both ToPMine and SegPhrase+ are extensible to mining quality phrases in multiple languages
- SegPhrase+ on Chinese (From Chinese Wikipedia)
- ToPMine on Arabic (From Quran (Fus7a Arabic)(no preprocessing)
  - Experimental results of Arabic phrases:

كفروا→ Those who disbelieve

بسم الله الرحمن الرحيم→ In the name of God the Gracious and Merciful

Rank	Phrase	In English
…	…	…
62	首席_执行官	CEO
63	中间_偏右	Middle-right
…	…	…
84	百度_百科	Baidu Pedia
85	热带_气旋	Tropical cyclone
86	中国科学院_院士	Fellow of Chinese Academy of Sciences
…	…	…
1001	十大_中文_金曲	Top-10 Chinese Songs
1002	全球_资讯网	Global Info Website
1003	天一阁_藏_明代_科举_录_选刊	A Chinese book name
…	…	…
9934	国家_戏剧_院	National Theater
9935	谢谢_你	Thank you
…	…	…

Outline

- Mining Structures from Text: A Data-Driven Approach
- On the Power of Big Data: Structures from Massive Unstructured Text
  - Phrase Mining: ToPMine → SegPhrase → AutoPhrase
  - Entity Resolution and Typing: ClusType → PLE (Refined Typing)
  - Relationship Discovery by Network Embedding
  - LAKI: Latent Keyphrase Inference
- Data to Network to Knowledge: A Path from Data to Knowledge

Recognizing Typed Entities

Identifying token span as entity mentions in documents and labeling their types

FOOD
LOCATION
JOB_TITLE
EVENT
ORGANIZATION
…

Target Types

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. …

Plain text

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

Text with typed entities

Why? — Enabling structured analysis of unstructured text corpus

Social media challenge!

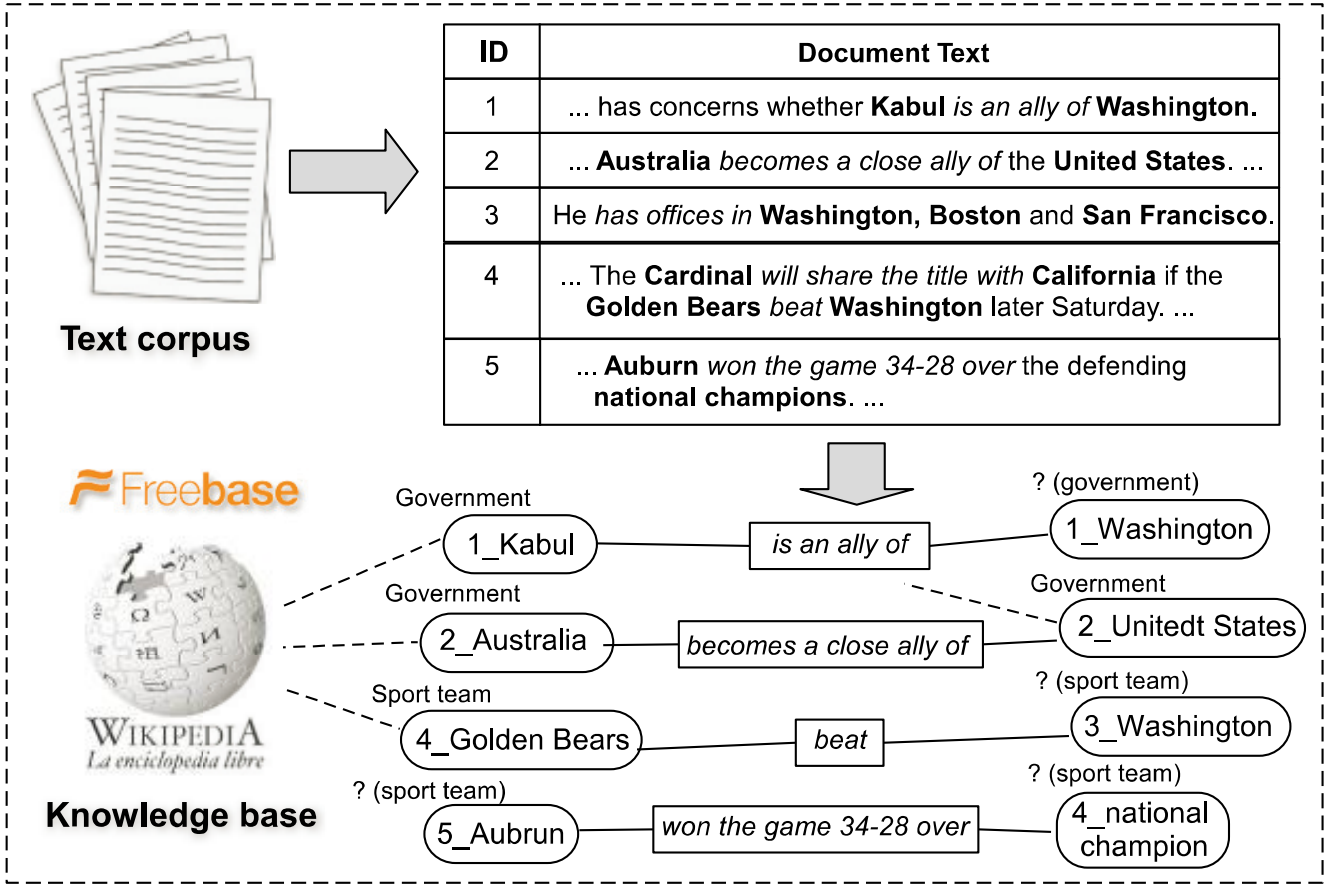
FOOD LOCATION EVENT

Traditional methods:
Expensive human labor on annotation
500 documents for entity extraction; 20,000 queries for entity linking

ClusType: A Distant Supervision Framework

Problem: *Distantly-supervised entity recognition in a domain-specific corpus*

- Given: (1) a domain-specific corpus *D,* (2) a knowledge base (e.g., Freebase), (3) a set of target types (*T*) from a KB
- Detect candidate entity mentions in D, and categorize each candidate mention by target types or Not-Of-Interest (NOI)



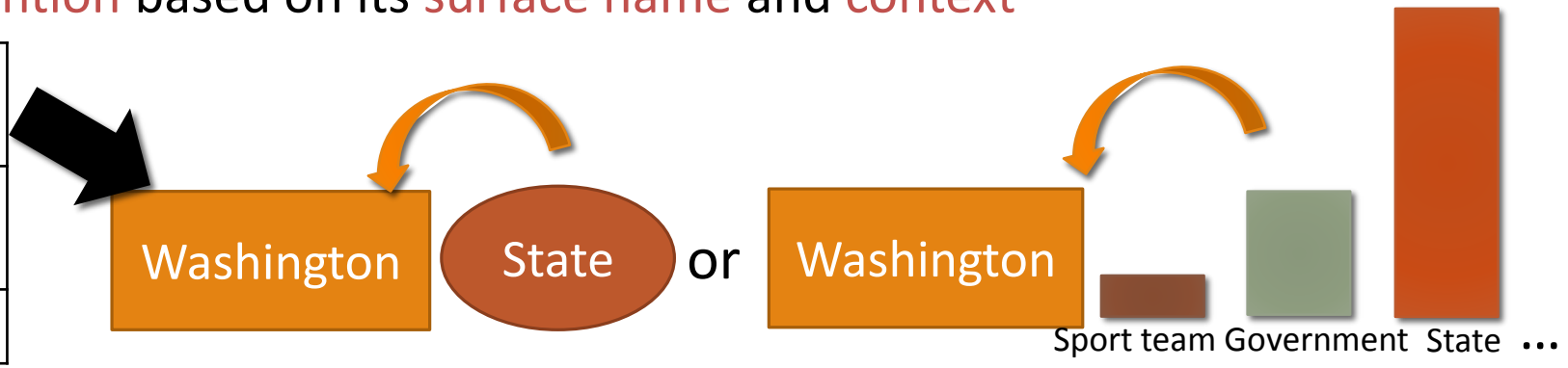
Solution [ClusType: KDD'15]:

- Detect entity mentions from text
- Map candidate mentions to KB entities of target types
- Use confidently mapped {mention, type} to infer types of remaining candidate mentions

Entity Recognition and Typing: Challenges and Solutions

- Challenge 1: Domain Restriction: Extensive training, use general-domain corpora, not work well on specific, dynamic or emerging domains (*e.g.*, tweets, Yelp reviews)
  - Solution: Domain-agnostic phrase mining: Extracts candidate entity mentions with minimal linguistic assumption (e.g., only use POS tagging)
- Challenge 2: Name ambiguity: Multiple entities may share the same surface name
  - Solution: Model each mention based on its surface name and context

While Griffin is not the part of Washington's plan on Sunday's game, …
… news from Washington indicates that the congress is going to…
It is one of the best state parks in Washington.



- Challenge 3: Context Sparsity: There are many ways to describe the same relation
  - Solution: cluster *relation phrase*, infer synonymous relation phrases

Sentence	freq
The magnitude 9.0 quake **caused widespread devastation in [Kesennuma city]**	12
… tsunami that **ravaged [northeastern Japan]** last Friday	31
The resulting tsunami **devastate [Japan]**'s northeast	244

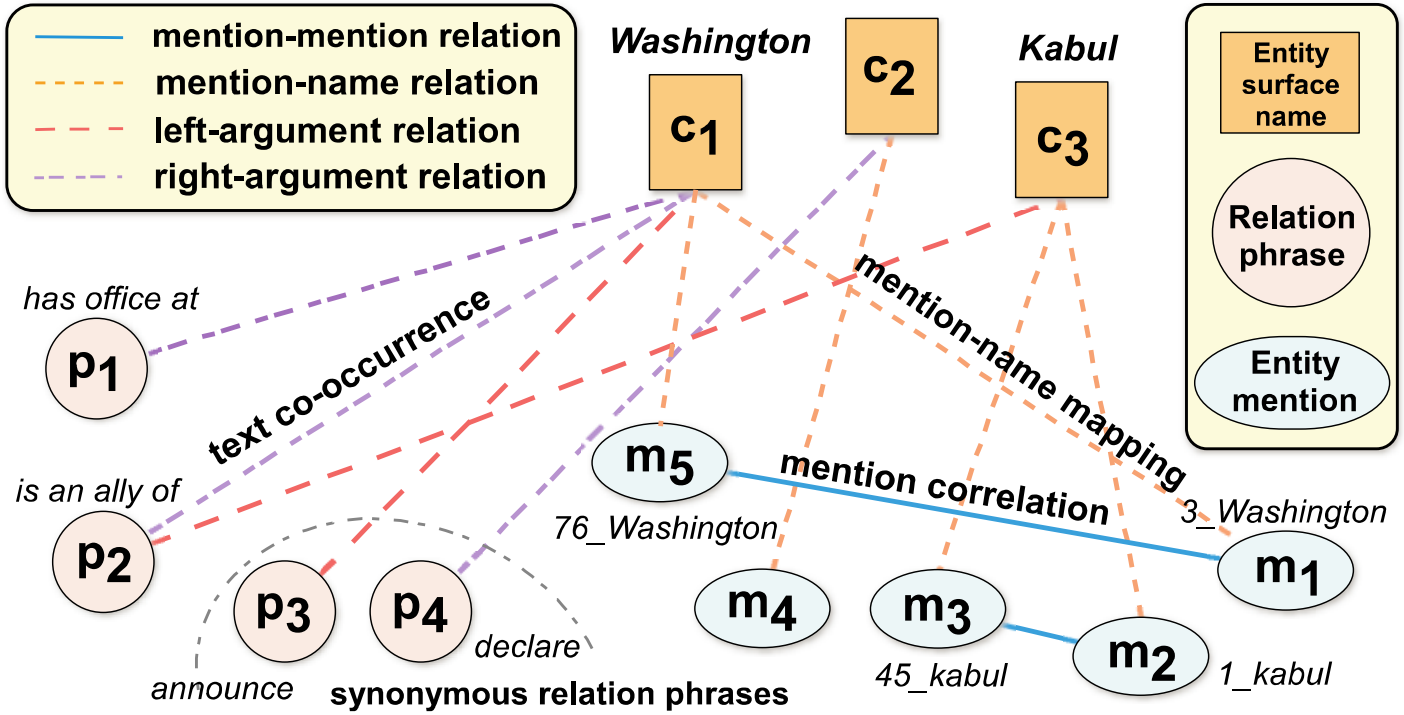
The ClusType Framework: Phrase Segmentation and Heterogeneous Graph Construction

- POS-constrained phrase segmentation for mining candidate entity mentions and relation phrases, simultaneously
- Construct a heterogeneous graph to represent available information in a unified form

Entity mentions are kept as individual objects **to be disambiguated**

Linked to entity surface names & relation phrases

Weight assignment: The more two objects are likely to share the same label, the larger the weight will be associated with their connecting edge



Candidate Generation

- Phrase mining incorporating both *corpus-level statistics* and *syntactic constraints*
  - **Global significance score**: Filter low-quality candidates; **generic POS tag patterns**: remove phrases with improper syntactic structure
  - Extend ToPMine to partition corpus into segments which meet both significance threshold and POS patterns → candidate entity mentions & relation phrases

Relation phrase: Phrase that denotes a unary or binary relation in a sentence

Pattern	Example
V	disperse; hit; struck; knock;
P	in; at; of; from; to;
V P	locate in; come from; talk to;
VW*(P)	caused major damage on; come lately

V-verb; P-prep; W-{adv | adj | noun | det | pron}
W* denotes multiple W; (P) denotes optional.

Experiment: Entity detection: Performance comparison between our method and an NP chunker

Method	NYT		Yelp		Tweet	
	Prec	Recall	Prec	Recall	Prec	Recall
Our method	**0.469**	**0.956**	**0.306**	**0.849**	0.226	**0.751**
NP chunker	0.220	0.609	0.296	0.247	**0.287**	0.181

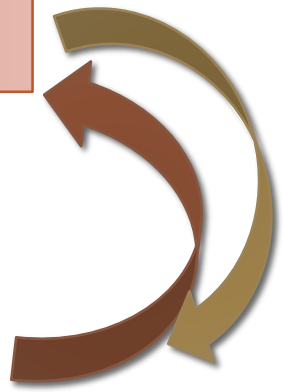
Recall is most critical for this step,since later we cannot detect the misses (i.e., false negatives)

The Framework: Mutual Enhancement of Type Propagation and Relation Phrase Clustering

- With the constructed graph, formulate a graph-based semi-supervised learning of two tasks jointly:

Type propagation on heterogeneous graph

Derived entity argument types serve as **good feature** for clustering relation phrases



Multi-view relation phrase clustering

Propagate type information among entities bridges via synonymous relation phrases

Mutually enhancing each other; leads to quality recognition of unlinkable entity mentions

Type Inference: A Joint Optimization Problem

$$\mathcal{O}_{\alpha,\gamma,\mu} = \mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) + \mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) + \Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R). \quad (2)$$

$$\mathcal{F}(\mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) = \sum_{i=1}^{n}\sum_{j=1}^{l} W_{L,ij}\left\|\frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{L,ii}}} - \frac{\mathbf{P}_{L,j}}{\sqrt{D^{(\mathcal{P})}_{L,jj}}}\right\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{l} W_{R,ij}\left\|\frac{\mathbf{C}_i}{\sqrt{D^{(\mathcal{C})}_{R,ii}}} - \frac{\mathbf{P}_{R,j}}{\sqrt{D^{(\mathcal{P})}_{R,jj}}}\right\|_2^2$$

Mention modeling & mention correlation

$$\Omega_{\gamma,\mu}(\mathbf{Y}, \mathbf{C}, \mathbf{P}_L, \mathbf{P}_R) = \big\|\mathbf{Y} - f(\Pi_{\mathcal{C}}\mathbf{C}, \Pi_L\mathbf{P}_L, \Pi_R\mathbf{P}_R)\big\|_F^2 + \frac{\gamma}{2}\sum_{c\in\mathcal{C}}\sum_{i,j=1}^{M_c} W_{ij}^{(c)}\left\|\frac{\mathbf{Y}_i}{\sqrt{D_{ii}^{(c)}}} - \frac{\mathbf{Y}_j}{\sqrt{D_{jj}^{(c)}}}\right\|_2^2 + \mu\|\mathbf{Y} - \mathbf{Y}_0\|_F^2$$

Type propagation between entity surface names and relation phrases

$$\mathcal{L}_\alpha\big(\mathbf{P}_L, \mathbf{P}_R, \{\mathbf{U}^{(v)}, \mathbf{V}^{(v)}\}, \mathbf{U}^*\big) = \sum_{v=0}^{d}\beta^{(v)}\big(\|\mathbf{F}^{(v)} - \mathbf{U}^{(v)}\mathbf{V}^{(v)T}\|_F^2 + \alpha\|\mathbf{U}^{(v)}\mathbf{Q}^{(v)} - \mathbf{U}^*\|_F^2\big). \quad (3)$$

Multi-view relation phrase clustering

ClusType: Experiment Setting

- Datasets: 2013 New York Times news (~110k docs) [event, PER, LOC, ORG]; Yelp Reviews (~230k) [Food, Job, …]; 2011 Tweets (~300k) [event, product, PER, LOC, …]
- Seed mention sets: < 7% extracted mentions are mapped to Freebase entities
- Evaluation sets: manually annotate mentions of target types for subsets of the corpora
- Evaluation metrics: Follows named entity recognition evaluation (Precision, Recall, F1)
- Compared methods
  - **Pattern**: Stanford pattern-based learning; **SemTagger**:bootstrapping method which trains contextual classifier based on seed mentions; **FIGER**: distantly-supervised sequence labeling method trained on Wiki corpus; **NNPLB**: label propagation using ReVerb assertion and seed mention; **APOLLO**: mention-level label propagation using Wiki concepts and KB entities;
  - **ClusType-NoWm**: ignore mention correlation; **ClusType-NoClus**: conducts only type propagation; **ClusType-TwpStep**: first performs hard clustering then type propagation

Comparing ClusType with Other Methods and Its Variants

Performance comparison on three datasets in terms of Precision, Recall and F1 score

Data sets	NYT			Yelp			Tweet		
Method	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Pattern [9]	0.4576	0.2247	0.3014	0.3790	0.1354	0.1996	0.2107	0.2368	0.2230
FIGER [16]	0.8668	0.8964	0.8814	0.5010	0.1237	0.1983	**0.7354**	0.1951	0.3084
SemTagger [12]	0.8667	0.2658	0.4069	0.3769	0.2440	0.2963	0.4225	0.1632	0.2355
APOLLO [29]	0.9257	0.6972	0.7954	0.3534	0.2366	0.2834	0.1471	0.2635	0.1883
NNPLB [15]	0.7487	0.5538	0.6367	0.4248	0.6397	0.5106	0.3327	0.1951	0.2459
ClusType-NoClus	0.9130	0.8685	0.8902	0.7629	0.7581	0.7605	0.3466	0.4920	0.4067
ClusType-NoWm	0.9244	0.9015	0.9128	0.7812	0.7634	0.7722	0.3539	**0.5434**	0.4286
ClusType-TwoStep	0.9257	0.9033	0.9143	0.8025	0.7629	0.7821	0.3748	0.5230	0.4367
ClusType	**0.9550**	**0.9243**	**0.9394**	**0.8333**	**0.7849**	**0.8084**	0.3956	0.5230	**0.4505**

- **vs. FIGER**: Effectiveness of our candidate generation and type propagation
- **vs**. **NNPLB** and **APOLLO**: ClusType utilizes not only semantic-rich relation phrase as type cues, but also cluster synonymous relation phrases to tackle context sparsity
- **vs**. our **variants**: (i) models mention correlation for name disambiguation; and (ii) integrates clustering in a mutually enhancing way

Comparing on Trained NER System

- Compare with Stanford NER,which is trained on general-domain corpora including ACE corpus and MUC corpus, on three types: PER, LOC, ORG

Comparing F1 measure with trained NER on large datasets

Method	NYT	Yelp	Tweet
Stanford NER*	0.6819	0.2403	0.4383
ClusType	**0.9419**	**0.5943**	**0.4717**

- ClusType and its variants outperform Stanford NER on both dynamic corpus (NYT) and domain-specific corpus (Yelp)
- ClusType has lower precision but higher Recall and F1 score on Tweet → Superior recall of ClusType mainly come from domain-independent candidate generation

* J. R. Finkel, T. Grenager and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In ACL'05.

Example Output and Relation Phrase Clusters

Example output of ClusType and the compared methods on the Yelp dataset

ClusType	SemTagger	NNPLB
The best **BBQ:Food** I've tasted in **Phoenix:LOC** ! I had the **[pulled pork sandwich]:Food** with **coleslaw:Food** and **[baked beans]:Food** for lunch. ...	The best BBQ I've tasted in **Phoenix:LOC** ! I had the pulled **[pork sandwich]:LOC** with **coleslaw:Food** and **[baked beans]:LOC** for lunch. ...	The best **BBQ:Loc** I've tasted in **Phoenix:LOC** ! I had the pulled pork **sandwich:Food** with coleslaw and baked **beans:Food** for **lunch:Food**. ...
I only go to **ihop:LOC** for **pancakes:Food** because I don't really like anything else on the menu. Ordered **[chocolate chip pancakes]:Food** and a **[hot chocolate]:Food**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered **[chocolate chip pancakes]:LOC** and a **[hot chocolate]:LOC**.	I only go to ihop for pancakes because I don't really like anything else on the menu. Ordered chocolate chip pancakes and a hot chocolate.

- Extracts more mentions and predicts types with higher accuracy

Example relation phrase clusters and corpus-wide frequency from the NYT dataset

ID	Relation phrase
1	recruited by (5.1k); employed by (3.4k); want hire by (264)
2	go against (2.4k); struggling so much against (54); run for re-election against (112); campaigned against (1.3k)
3	looking at ways around (105); pitched around (1.9k); echo around (844); present at (5.5k);

- Not only synonymous relation phrases, but also both sparse and frequent relation phrase can be clustered together
- → boosts sparse relation phrases with type information of frequent relation phrases

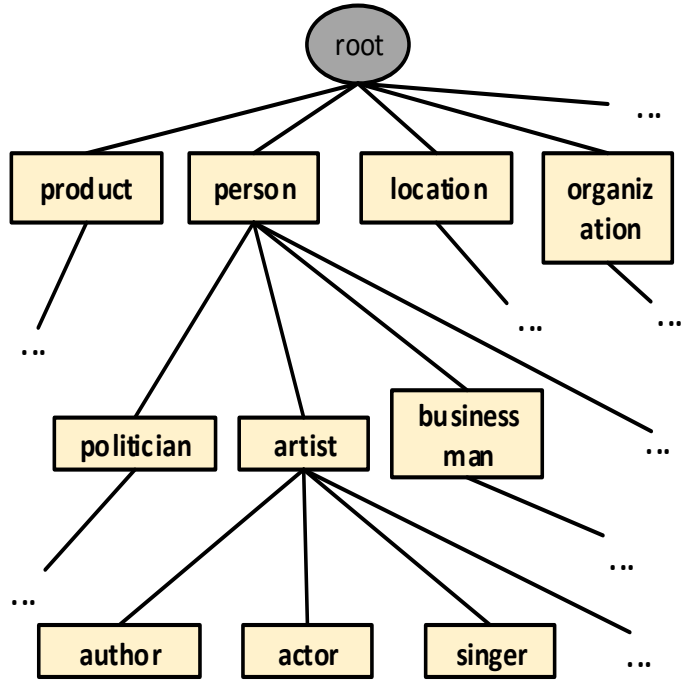
Fine-Grained Entity Typing [PLE: KDD'16]

- **Fine-grained Entity Typing**: Type labels for a mention forms a "*type-path*" (not necessarily ending in a leaf node) in a given (tree-structured) type hierarchy

ID	Sentence
S1	Republican presidential candidate ***Donald Trump*** spoke during a campaign event in Rock Hill.
S2	***Donald Trump***'s company has threatened to withhold up to $1 billion of investment if the U.K. government decides to ban his entry into the country.
S3	In ***Trump***'s TV reality show, "The Apprentice", 16 people competed for a job.
…	…

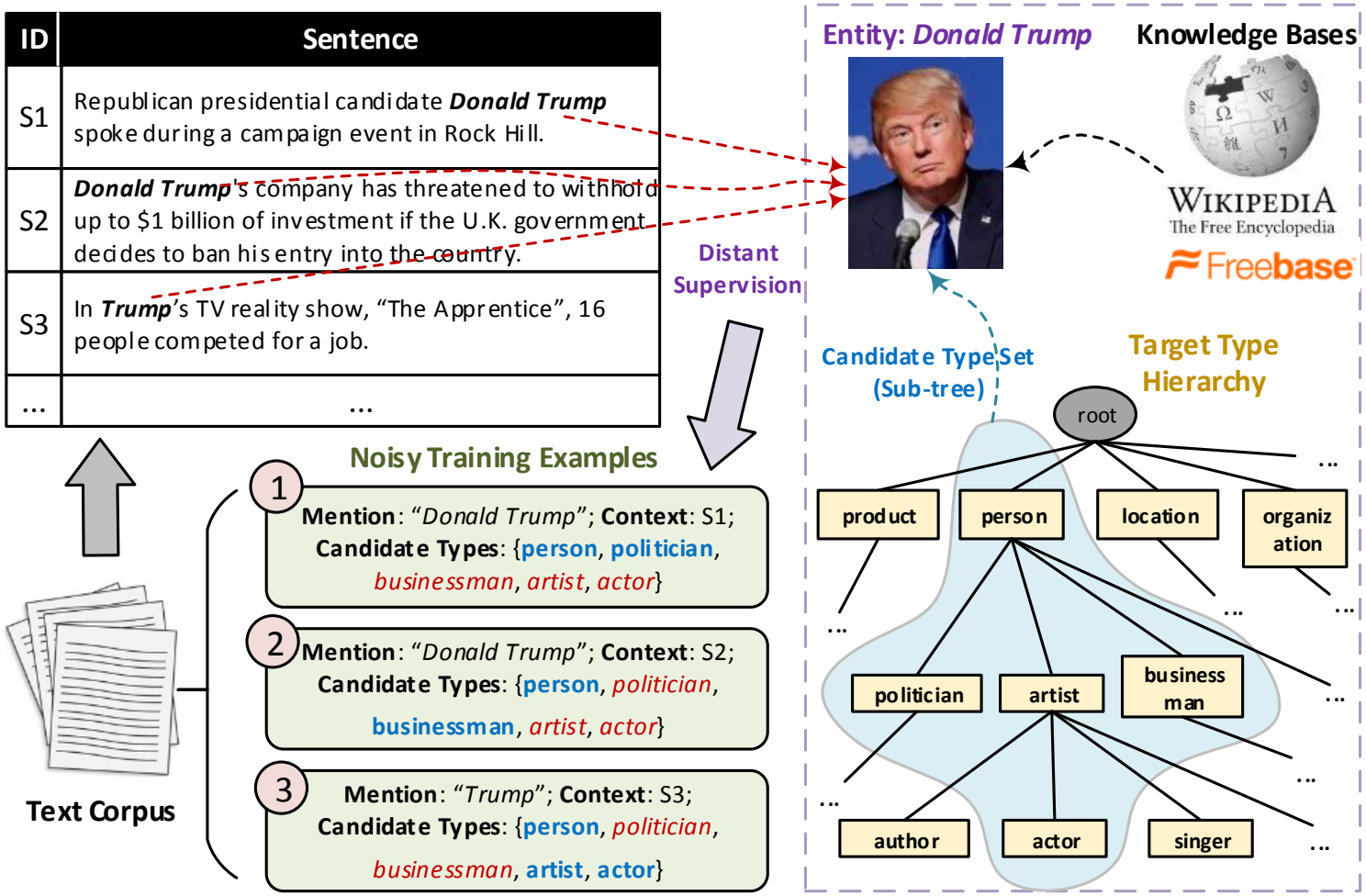
Type-path

- Person → politician
- Person → businessman
- Person → artist → actor



- Manually annotating training corpora with **100+** entity types
  - Expensive & Error-prone
- **Current practice**: Use distant supervision to ***automatically labeled training corpora***

Label Noise in Entity Typing

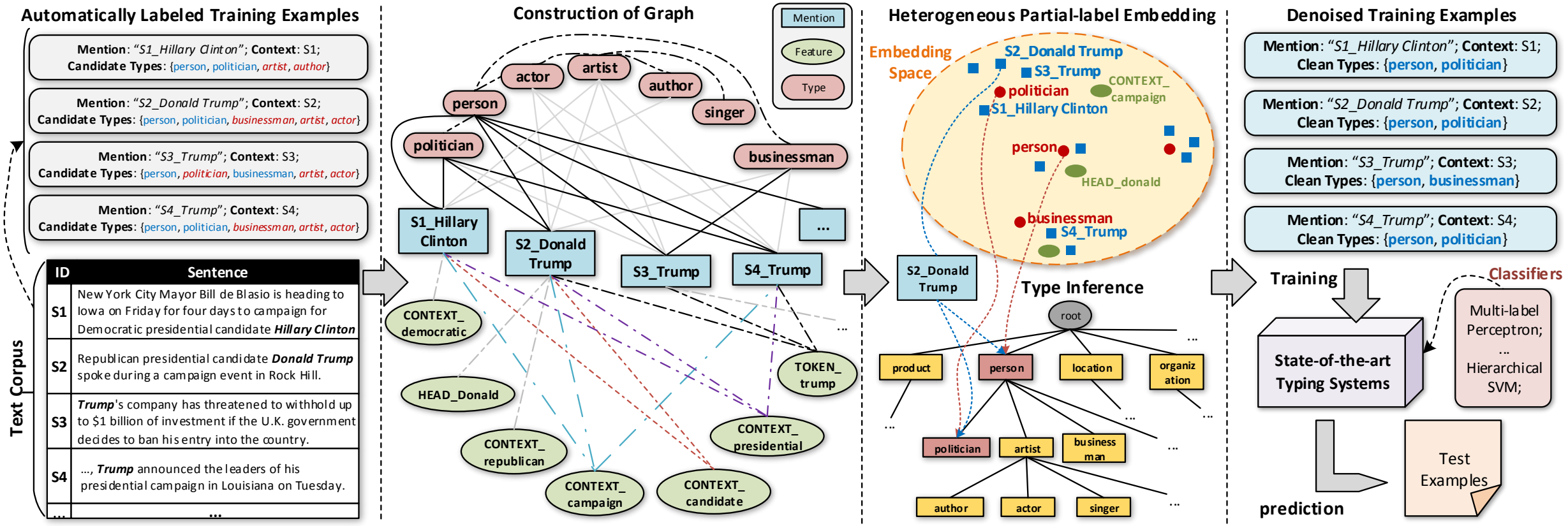


Donald Trump is mentioned in sentences S1-S3

- Distant supervision
  - Assign *same* types (blue region) to *all* the mentions
  - Does not consider ***local contexts*** when assigning type labels
  - Introduce ***label noise*** to the mentions

The types assigned to entity Trump include person, artist, actor, politician, businessman, while only {person, politician} are correct types for the **mention *"Trump"* in S1**

Label Noise Reduction: Framework Overview



1. Generate text features and construct a heterogeneous graph
2. Perform joint embedding of the constructed graph G into the same low-dimensional space
3. For each mention, search its candidate type sub-tree in a top-down manner and estimate the true type-path from learned embedding

Experiment Setting: PLE vs. Others

- Datasets:
  - (1) **Wiki**: 1.5M sentences sampled from ~780k Wikipedia articles
  - (2) **OntoNotes**: 13,109 news
  - (3) **BBN**: 2,311 Wall Street Journal articles

Data sets	Wiki	OntoNotes	BBN
#Types	113	89	47
#Documents	780,549	13,109	2,311
#Sentences	1.51M	143,709	48,899
#Training mentions	2.69M	223,342	109,090
#Ground-truth mentions	563	9,604	121,001
#Features	644,860	215,642	125,637
#Edges in graph	87M	5.9M	2.9M

- Compared Methods
  - **Sib**: removes siblings types
  - **Min**: removes types that appear only once in the document
  - **All**: first performs Sib pruning then Min pruning
  - **DeepWalk**: embedding a homogeneous graph with binary edges
  - **LINE**: second-order LINE
  - **WSABIE**: adopts WARP loss with kernel extension
  - **PTE**: applied PTE joint training algorithm on subgraphs GM F and GM
  - **PL-SVM**: uses a margin-based loss to handle label noise
  - **CLPL**: uses a linear model to encourage large average scores for candidate types

Intrinsic Experiments: Effectiveness of Label Noise Reduction

- Goal: compare how accurately PLE and the other methods can estimate the true types of mentions from its noisy candidate type set

	Wiki							OntoNotes						
Method	Acc	Ma-P	Ma-R	Ma-F1	Mi-P	Mi-R	Mi-F1	Acc	Ma-P	Ma-R	Ma-F1	Mi-P	Mi-R	Mi-F1
Raw	0.373	0.558	**0.681**	0.614	0.521	**0.719**	0.605	0.480	0.671	**0.793**	0.727	0.576	**0.786**	0.665
Sib [7]	0.373	0.583	0.636	0.608	0.578	0.653	0.613	0.487	0.710	0.732	0.721	0.675	0.702	0.688
Min [7]	0.373	0.561	0.679	0.615	0.524	0.717	0.606	0.481	0.680	0.777	0.725	0.592	0.763	0.667
All [7]	0.373	0.585	0.634	0.608	0.581	0.651	0.614	0.487	0.716	0.724	0.720	0.686	0.691	0.689
DeepWalk-Raw [21]	0.328	0.598	0.459	0.519	0.595	0.367	0.454	0.441	0.625	0.708	0.664	0.598	0.683	0.638
LINE-Raw [29]	0.349	0.600	0.596	0.598	0.590	0.610	0.600	0.549	0.699	0.770	0.733	0.677	0.754	0.714
WSABIE-Raw [34]	0.332	0.554	0.609	0.580	0.557	0.633	0.592	0.482	0.686	0.743	0.713	0.667	0.721	0.693
PTE-Raw [28]	0.419	0.678	0.597	0.635	0.686	0.607	0.644	0.529	0.687	0.754	0.719	0.657	0.733	0.693
PLE-NoCo	0.556	0.795	0.678	0.732	0.804	0.668	0.730	0.593	0.768	0.773	0.770	0.751	0.762	0.756
PLE-CoH	0.568	0.805	0.671	0.732	0.808	0.704	0.752	0.620	0.789	0.785	0.787	0.778	0.769	0.773
PLE	**0.589**	**0.840**	0.675	**0.749**	**0.833**	0.705	**0.763**	**0.639**	**0.814**	0.782	**0.798**	**0.791**	0.766	**0.778**

40.57% improvement in Accuracy and 23.89% improvement in Macro-Precision compared to the best baseline on Wiki dataset

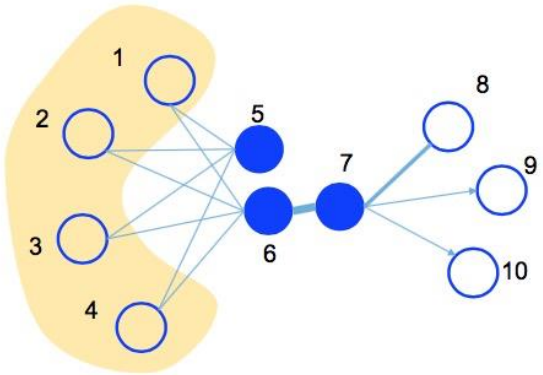
- **vs. pruning strategies**: LNR *identifies true types* from the candidate type sets instead of *aggressively deleting instances* with noisy type labels
- **vs. other embedding methods**: PLE obtains superior performance because it effectively *models the noisy type labels*
- **vs. PLE variants**: (i) PLE captures *type semantic similarity*; (ii) modeling type correlation with entity-type facts in KB yields more accurate and complete type correlation statistics than type hierarchy-based approach

Outline

- Mining Structures from Text: A Data-Driven Approach
- On the Power of Big Data: Structures from Massive Unstructured Text
  - Phrase Mining: ToPMine → SegPhrase → AutoPhrase
  - Entity Resolution and Typing: ClusType → PLE (Refined Typing)
  - Relationship Discovery by Network Embedding
  - LAKI: Latent Keyphrase Inference
- Data to Network to Knowledge: A Path from Data to Knowledge

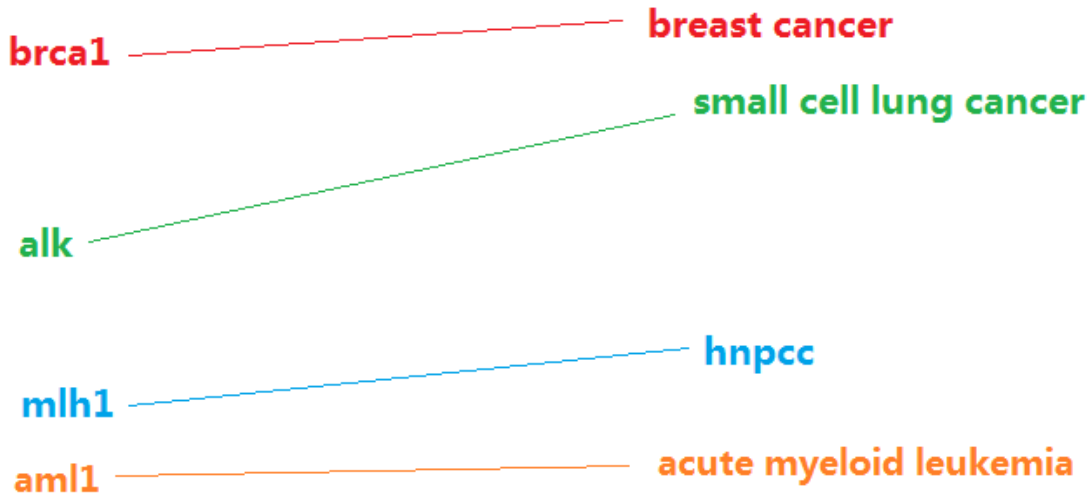
Biological Relationship Discovery for Network Building

- Automatic extraction of relationships between different biological entities from biological research papers (e.g., PubMed)
  - Gene – Disease; Drug - Disease; Drug - Pathway; Drug - Target gene
- Challenges
  - Entity recognition: Most biological entities consist of multiple words
    - E.g., Non-small Cell Lung Cancer, Acute Myeloid Leukemia
  - Sparsity: Most biological entities co-occur only a few times in research papers
    - Most relationships are not explicitly described in papers
  - Few labeled data
- Key ideas
  - Phrase mining
  - Learn phrase-based network embedding from massive data
    - Using: LINE (Tang et al., Large-scale Information Network Embedding, WWW'15)
  - Calculate network embedding



Key Property to Learn Embedding & Experiments

- Key Property to Learn Embedding
  - The lines between genes and diseases are parallel
  - Given a seed pair (A,B) and a query X, we can find an entity Y which satisfies
    - $(A,B) \approx (X, Y)$
    - $Y = Argmax\{sim(\ B - A + X\ ,\ Y\ )\}$



- Experimental Settings
  - Sample 10% Pubmed abstracts
  - Detect phrases by using a 200K phrase list
  - Build a co-occurrence network for all words and phrases
  - Learn entity embedding from the co-occurrence network

Results: Extracted Relations (from 10% PubMed Abstracts)

Relation	Seed Pair	Query Entity	Top Ranked Entities
Gene-Disease	Breast Cancer, BRCA1	Acute Myeloid Leukemia	AML1, E2A-PBX1, NPM1, RUNX1, PBX1
		Acute Lymphocytic Leukemia	E2A-PBX1, NPM1, EVI1, BCL6, ALL1
		HNPCC	MLH1, MSH6, hMSH2, hMLH1, MSH2
	BRCA1, Breast Cancer	ALK	Small Cell Lung Cancer, Non-small Cell Lung Cancer
		AML1	Leukemia, AML, CML
		MLH1	Colorectal Cancer, HNPCC, Colon Cancer
Drug-Disease	Leukemia, Doxorubicin	Small Cell Lung Cancer	Paclitaxel, Gemcitabine, Docetaxel, Cisplatin
		Depressive Disorder	Sertraline, Desvenlafaxine, Duloxetine, Paliperidone
		HIV	Zidovudine, Ritonavir, Lamivudine, Atazanavir
	Doxorubicin, Leukemia	Aspirin	Peptic Ulcer Bleeding, Venous Thromboembolic
		Sertraline	Depressive Disorder, Social Anxiety Disorder
		Penicillin	Bacterial Meningitis, Scabies, Streptococcus

Disease-* Relation Extraction

Relation and Seed Pair	Query Entity	Top Ranked Entities and Their Scores
Disease-Drug Heart Disease : Aspirin	Cerebrovascular	Clopidogrel, Anti-platelet, Ticlopidine, Ticagrelor, prasugrel 0.7170, 0.6955, 0.6922, 0.6759, 0.6661
	Ischemic Heart Disease	Anti-platelet, Clopidogrel, Ticlopidine, Aspirin-Clopidogrel, Plavix 0.7785, 0.7732, 0.7532, 0.7481, 0.7473
	Coronary Heart Disease	Clopidogrel, Anti-platelet, Aspirin-Clopidogrel, Prasugrel, Ticagrelor 0.7855, 0.7606, 0.7248, 0.7148, 0.7086
	Dilated Cardiomyopathy	Clopidogrel, Ticlopidine, Prasugrel, Plavix, ACE Inhibitor 0.7649, 0.7436, 0.6968, 0.6765, 0.6586
	Valvular Heart Disease	Anti-platelet, Ticlopidine, Clopidogrel, Aspirin-Clopidogrel, Plavix 0.7750, 0.7749, 0.7668, 0.7529, 0.7260
	Arrhythmia	Clopidogrel, Anti-platelet, Ticlopidine, Thienopyridine, Ticagrelor 0.7589, 0.7411, 0.6958, 0.6838, 0.6788
Disease-Gene Breast Cancer : brca1	Cerebrovascular	tlr9, myh15, abca1, uts2, abcg1 0.6980, 0.6952, 0.6829, 0.6790, 0.6770
	Ischemic Heart Disease	sdc2, mth1, uts2, kcnn4, hspa8 0.7624, 0.7604, 0.7443, 0.7431, 0.7390
	Coronary Heart Disease	apoc2, uts2, apoh, lox1, mth1 0.7911, 0.7765, 0.7754, 0.7718, 0.7615
	Dilated Cardiomyopathy	calm1, actn2, ankrd1, col1a2, fhl2 0.7385, 0.7370, 0.7368, 0.7314, 0.7298
	Valvular Heart Disease	col11a2, ndufs2, kcnn4, ncam1, myl1 0.6938, 0.6815, 0.6765, 0.6750, 0.6717
	Arrhythmia	atp1a2, casq2, ndufs2, gpd1l, kcne4 0.6772, 0.6745, 0.6743, 0.6713, 0.6705

Top Molecules for Cadiovascular Diseases

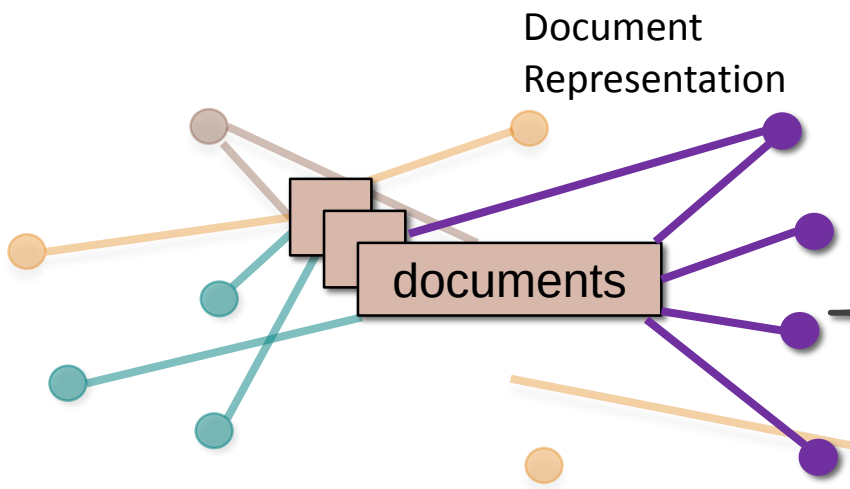
Disease	Top Ranked Molecules and their scores
Cerebrovascular Accident	Alpha-galactosidase A, Brain-derived Neurotrophic Factor, Tissue-type Plasminogen Activato Methylenetetrahydrofolate Reductase, Matrix Metalloproteinase-9 5.903, 5.595, 4.945, 2.710, 2.680
Ischemic Heart Disease	Cholesteryl Ester Transfer Protein, Apolipoprotein A-I, Adiponectin, Lipoprotein Lipase, Myeloperoxidase 4.597, 3.989, 3.651, 3.302, 3.240
Cardiomyopathy	Interferon Gamma, Interleukin-4, Interleukin-17a, Tumor Necrosis Factor, Titin 3.336, 2.809, 2.729, 2.549, 2.349
Arrhythmia	Methionine Synthase, Ryanodine Receptor 2, Platelet-Activating Factor Acetylhydrolase, Potassium Voltage-gated Channel Subfamily H Member 2, Gap Junction Alpha-1 Protein, 3.799, 3.354, 1.740, 2.730, 1.872
Valve Dysfunction	Mineralocorticoid Receptor, Elastin, Tropomyosin Alpha-1 Chain, Myosin-Binding Protein C Cardiac-type, Platelet-Activating Factor Acetylhydrolase 3.276, 2.380, 2.332, 1.704, 1.611
Congenital Heart Disease	Fibrillin-1, Plakophilin-2, Tyrosine-protein Phosphatase Non-receptor Type 11, Arachidonate 5-Lipoxygenase-activating Protein, Catechol O-methyltransferase 4.920, 3.208, 2.667, 2.036, 1.791

Outline

- Mining Structures from Text: A Data-Driven Approach
- On the Power of Big Data: Structures from Massive Unstructured Text
  - Phrase Mining: ToPMine → SegPhrase → AutoPhrase
  - Entity Resolution and Typing: ClusType → PLE (Refined Typing)
  - Relationship Discovery by Network Embedding
  - LAKI: Latent Keyphrase Inference 
- Data to Network to Knowledge: A Path from Data to Knowledge

LAKI: Representing Documents via Latent Keyphrase Inference

- Jialu Liu, Xiang Ren, Jingbo Shang, Taylor Cassidy, Clare Voss and Jiawei Han, "Representing Documents via Latent Keyphrase Inference", WWW'16
- **Document Representation**



Words:
dbscan, methods, clustering, process, …

Topics:
[k-means, clustering, clusters, dbscan, …]
[clusters, density, dbscan, clustering, …]
[machine, learning, knowledge, mining, …]

Knowledge base concepts:
data mining: /m/0blvg
clustering analysis: /m/031f5p
dbscan: /m/03cg_k1

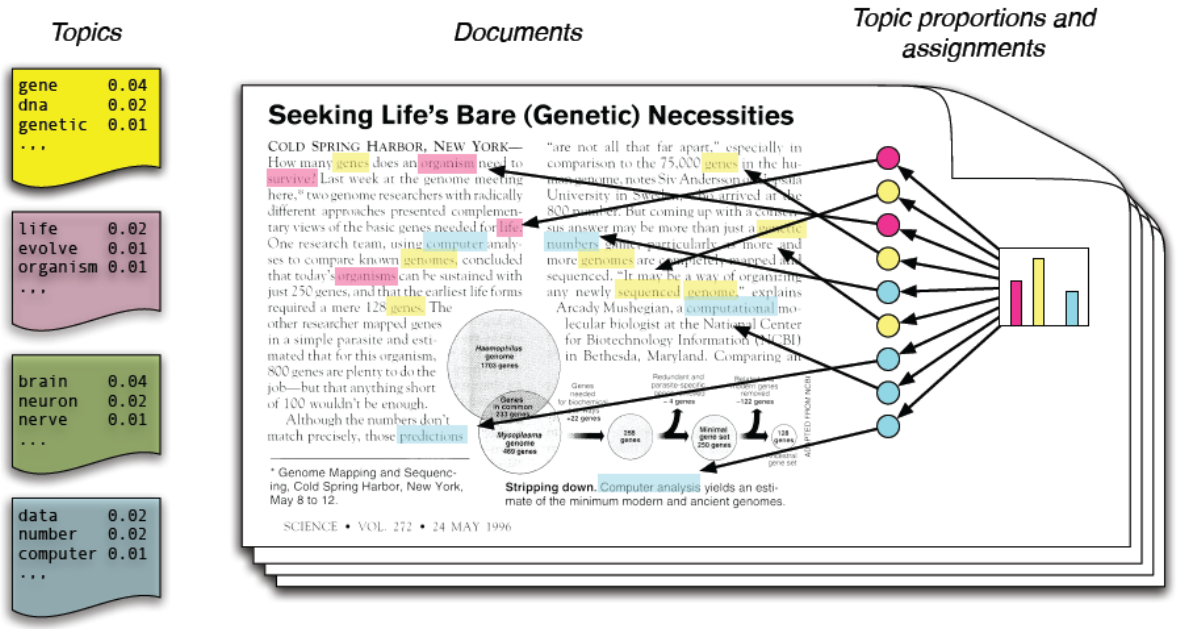
Document keyphrase:
dbscan: [dbscan, density, clustering, ...]
clustering: [clustering, clusters, partition, ...]
data mining: [data mining, knowledge, ...]

- A document can be represented by
  - A set of works, topics, KB concepts, Keyphrases, …

Document Representation: Traditional Methods

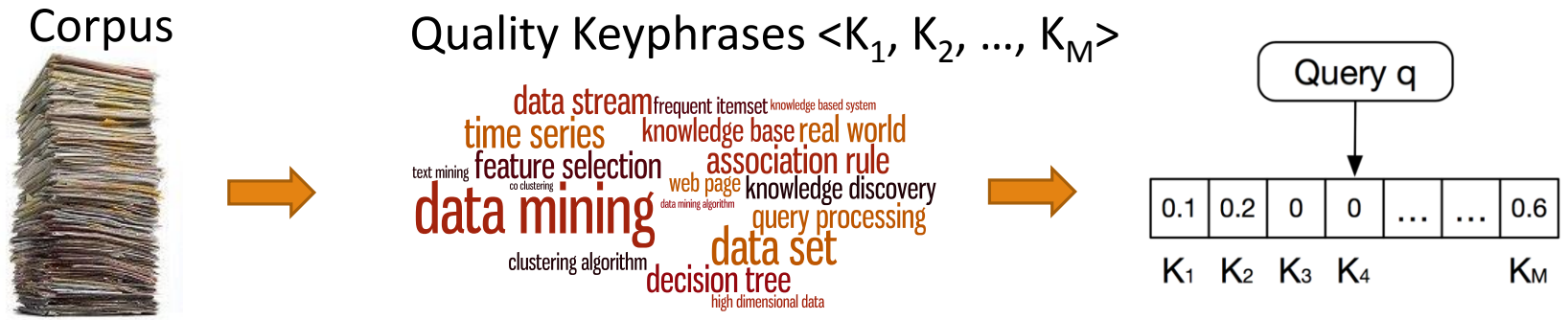
- Bag-of-Words or Bag-of-Phrases
  - Cons: Sparse on short texts
- Topic models [LDA]
  - Each **topic** is a distribution over words; each **document** is a mixture of corpus-wide topics
  - Cons: Difficult for human to infer topic semantics

	doc1	doc2	doc3
I	1	0	0
like	1	0	0
football	1	1	0
John	0	1	1
likes	0	1	1
football	1	1	0
basketball	0	0	1



Document Representation Using Keyphrases

- Use quality phrases as the entries in the vector and identify document keyphrases (subset of quality phrases) by evaluating relatedness between (doc, quality phrase)
- Unsupervised model

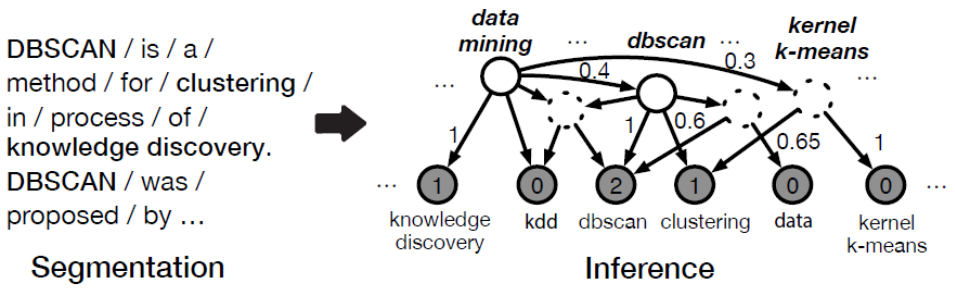


- Challenges
  - Where to get quality phrases from a given corpus?
    - Mining Quality Phrases from Massive Text Corpora [SIGMOD15]
  - How to identify document keyphrases?
    - Can be latent mentions
    - Relatedness scores
  - How to deal with relationship between quality phrases?

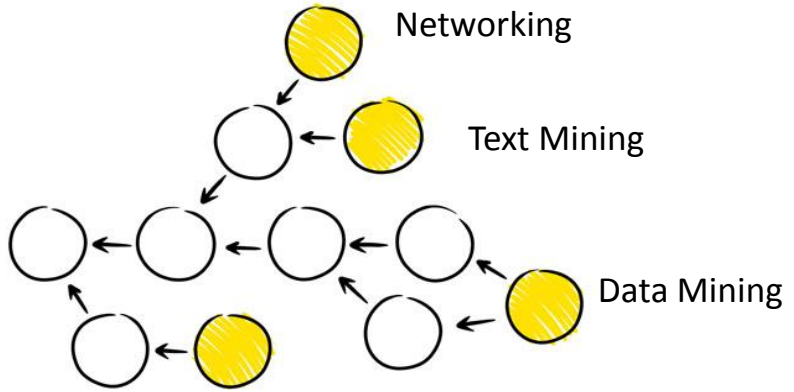
Document Representation Using Keyphrases: General Ideas

- How to identify document keyphrases?
  - Powered by Bayesian Inference on "Quality Phrase Silhouette"
    - Quality Phrase Silhouette: Topic centered on quality phrase
      - "Reverse" topic models
      - "Pseudo content" for quality phrase

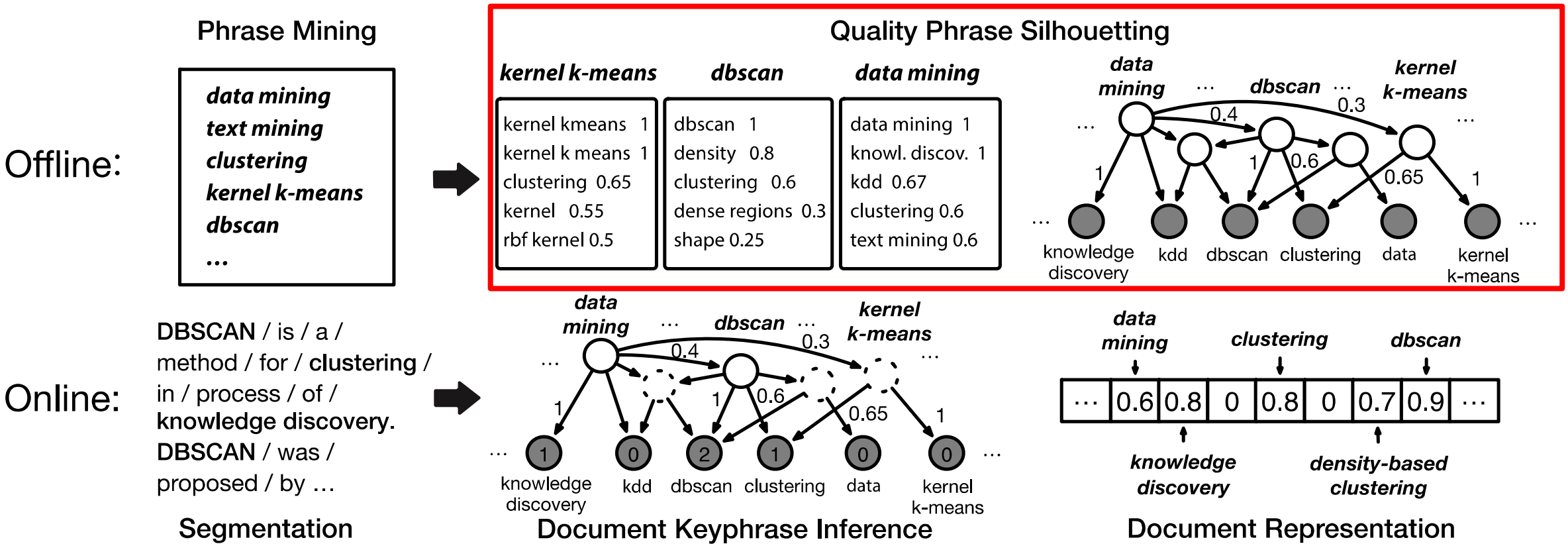
kernel k-means	*dbscan*	*data mining*
kernel kmeans 1	dbscan 1	data mining 1
kernel k means 1	density 0.8	knowl. discov. 1
clustering 0.65	clustering 0.6	kdd 0.67
kernel 0.55	dense regions 0.3	clustering 0.6
rbf kernel 0.5	shape 0.25	text mining 0.6



- How to deal with relationship between quality phrases?
  - Phrases are interconnected as a Directed Acyclic Graph

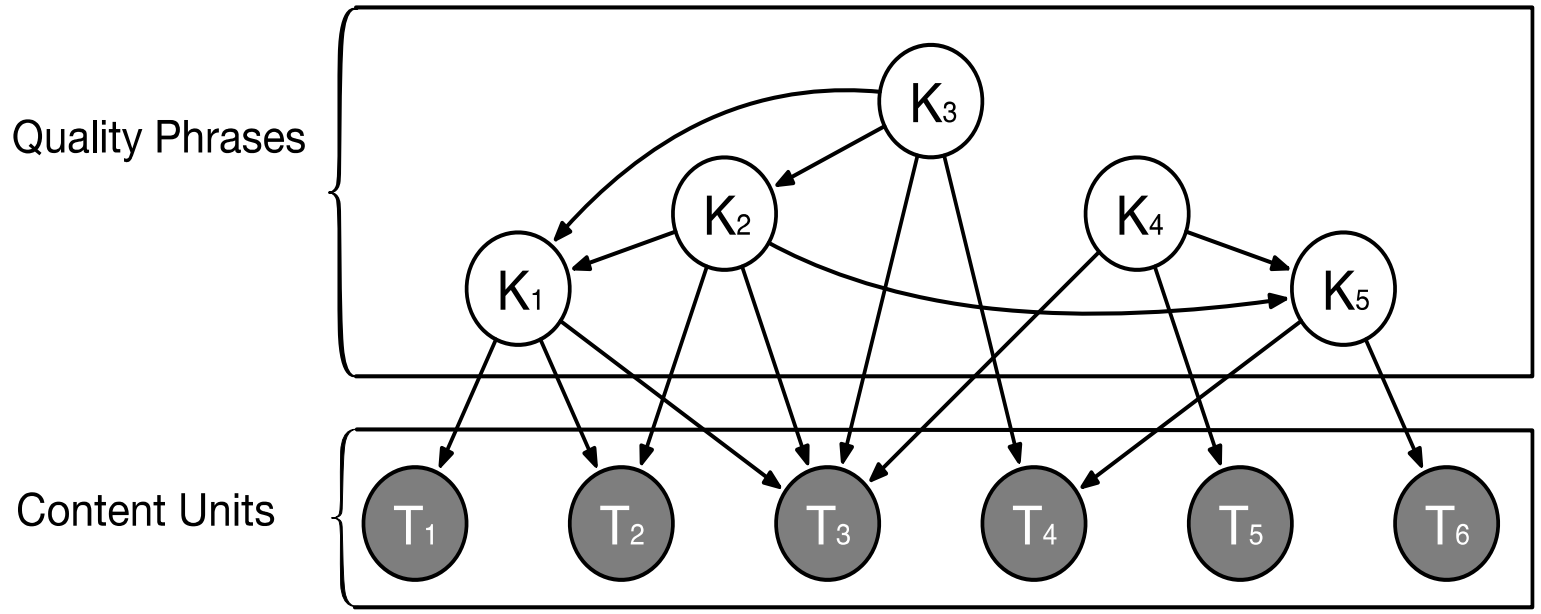


Framework for Latent Keyphrase Inference (LAKI)



LAKI: Deriving Quality Phrase Silhouette

- Learning Hierarchical Bayesian Network (DAG)



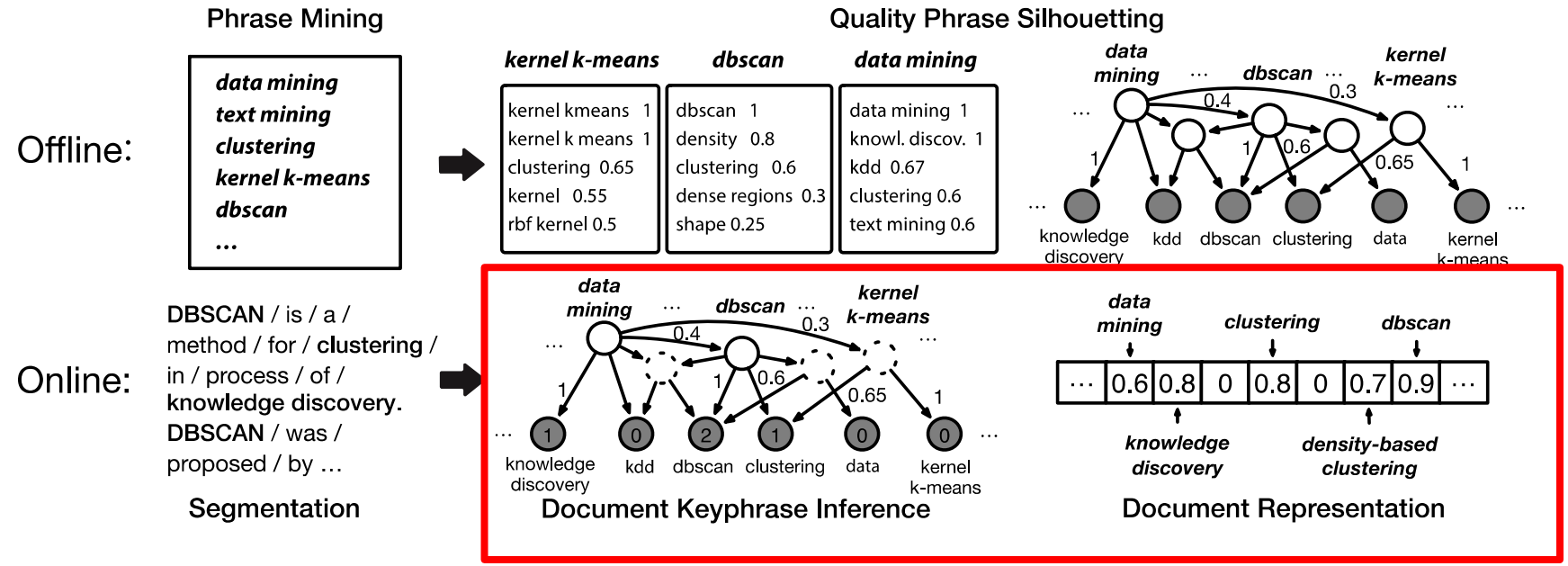
Binary Variables

Task 1: Model Learning: Learning link weights

Task 2: Structure Learning: Learning network structure

LAKI: Inference

- When do we need inference？
  - Expectation step in model learning
  - New documents
- Why is it slow?
  - NP hard to compute posterior probability for Noisy-Or networks



- Method: Approximate inference instead
  - Pruning irrelevant nodes using an efficient scoring function
  - Gibbs sampling

LAKI: Experiment Setting

- Two text-related tasks to evaluate document representation quality
  - Phrase relatedness
  - Document classification
- Two datasets:
- Methods:
  - **ESA** (Explicit Semantic Analysis)
  - **KBLink** uses link structure in Wikipedia
  - **BoW** (bag-of-words)
  - **ESA-C:** extends ESA by replacing Wiki with domain corpus
  - **LSA** (Latent Semantic Analysis)
  - **LDA** (Latent Dirichlet Allocation)
  - **Word2Vec** is a neural network computing word embeddings
  - **EKM** uses explicit keyphrase detection

Dataset	#Docs	#Words	Content type
Academia	0.43M	28M	title & abstract
Yelp	0.47M	98M	review

Method	Semantic Space	Input Source
ESA	KB concepts	KB
KBLink	KB concepts	KB
BoW	Words	-
ESA-C	Documents	Corpus
LSA	Topics	Corpus
LDA	Topics	Corpus
Word2Vec	-	Corpus
EKM	Explicit Keyphrases	Corpus
LAKI	Latent Keyphrases	Corpus

LAKI: Experimental Results

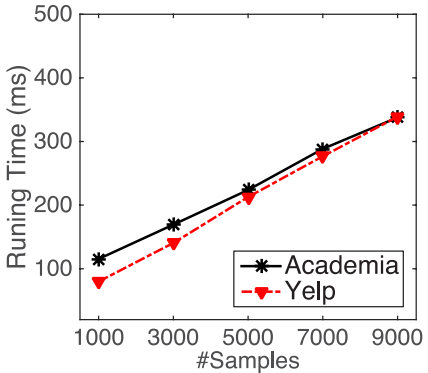
- **Phrase Relatedness Correlation**

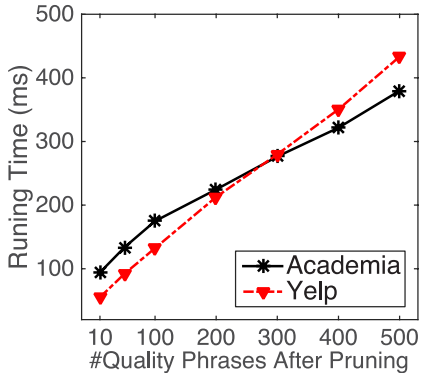
Method	Academia (w/ phrase)	Yelp (w/ phrase)
ESA	37.61 (-)	46.56 (-)
KBLink	36.37 (-)	35.94 (-)
BoW	48.05 (45.60)	51.26 (45.97)
ESA-C	39.75 (42.20)	49.13 (54.51)
LSA	72.50 (79.22)	66.55 (78.57)
LDA	77.27 (80.52)	75.55 (82.65)
EKM	45.46	40.57
LAKI	**84.42**	**90.58**

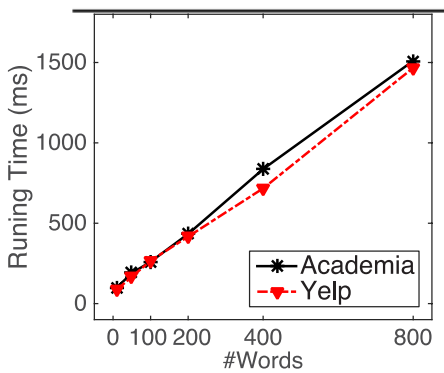
- **Document Classification**

Method	Academia (w/ phrase)	Yelp (w/ phrase)
ESA	0.4320 (-)	0.4567 (-)
KBLink	0.1878 (-)	0.4179 (-)
ESA-C	0.4905 (0.5243)	0.4655 (0.5029)
LSA	0.5877 (0.6383)	0.6700 (0.7229)
LDA	0.3610 (0.5391)	0.3928 (0.5405)
Word2Vec	0.6674 (0.7281)	0.7143 (0.7419)
LAKI	**0.7504**	**0.7609**

- **Time Complexity**







Case Study

- Query on phrases
  - Academia
  - Yelp
- Query on short documents (paper titles or sentences)
  - Academia
  - Yelp

Query	*LDA*	*BOA*
Keyphrases	linear discriminant analysis, latent dirichlet allocation, topic models, topic modeling, face recognition, sda, latent dirichlet, generative model, topic, subspace models, ...	boa steakhouse, bank of america, stripsteak, agnolotti, credit card, santa monica, restaurants, wells fargo, steakhouse, prime rib, bank, vegas, las vegas, cash, cut, dinner, bank, money, ...
Query	*LDA topic*	*BOA steak*
Keyphrases	latent dirichlet allocation, topic, topic models, topic modeling, probabilistic topic models, latent topics, topic discovery, generative model, mixture, text mining, topic distribution, plsi, ...	steak, stripsteak, boa steakhouse, steakhouse, ribeye, craftsteak, santa monica, medium rare, prime, vegas, entrees, potatoes, french fries, filet mignon, mashed potatoes, texas roadhouse, ...
Query	*SVM*	*deep dish pizza*
Keyphrases	support vector machines, svm classifier, multi class, training set, margin, knn, classification problems, kernel function, multi class svm, multi class support vector machine, support vector, ...	deep dish pizza, chicago, deep dish, amore taste of chicago, amore, pizza, oregano, chicago style, chicago style deep dish pizza, thin crust, windy city, slice, pan, oven, pepperoni, hot dog, ...

Query	*Mining Frequent Patterns without Candidate Generation*	*I am a huge fan of the All You Can Eat Chinese food buffet.*
Keyphrases	mining frequent patterns, candidate generation, frequent pattern mining, candidate, prune, fp growth, frequent pattern tree, apriori, subtrees, frequent patterns, candidate sets, ...	all you can eat, chinese food, buffet, chinese buffet, dim sum, orange chicken, chinese restaurant, asian food, asian buffet, crab legs, lunch buffet, fan, salad bar, all you can drink, ...
Query	*Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through means such as statistical pattern learning.*	*It's the perfect steakhouse for both meat and fish lovers. My table guest was completely delirious about his Kobe Beef and my lobster was perfectly cooked. Good wine list, they have a lovely Sancerre! Professional staff, quick and smooth.*
Keyphrases	text analytics, text mining, patterns, text, textual data, topic, information, text documents, information extraction, machine learning, data mining, knowledge discovery, ...	kobe beef, fish lovers, steakhouse, sancerre, wine list, guests, perfectly cooked, lobster, staff, meat, fillet, fish, lover, seafood, ribeye, filet, sea bass, risotto, starter, scallops, steak, beef, ...

Academia | Yelp

Outline

- Mining Structures from Text: A Data-Driven Approach
- On the Power of Big Data: Structures from Massive Unstructured Text
  - Phrase Mining: ToPMine → SegPhrase → AutoPhrase
  - Entity Resolution and Typing: ClusType → PLE (Refined Typing)
  - Relationship Discovery by Network Embedding
  - LAKI: Latent Keyphrase Inference
- Data to Network to Knowledge: A Path from Data to Knowledge

Conclusions

- Ubiquity of massive unstructured data
- Massive unstructured data → Massive power at turning them into structures
  - Phrase mining
  - Entity resolution and typing
  - Hidden relationship discovery
  - Latent keyphrase inference
  - Many more to be explored!
- Knowledge is power, but knowledge is hidden in massive unstructured text data!
  - Key—Turning unstructured data into massive, "relatively structured" networks!
- From data to knowledge (D2K) → D2N2K: A long but promising way to go!!

From Data Mining to Mining Structures & Heter. Info. Networks

Han, Kamber and Pei, Data Mining, 3rd ed. 2011

Yu, Han and Faloutsos (eds.), Link Mining, 2010

Sun and Han, Mining Heterogeneous Information Networks, 2012

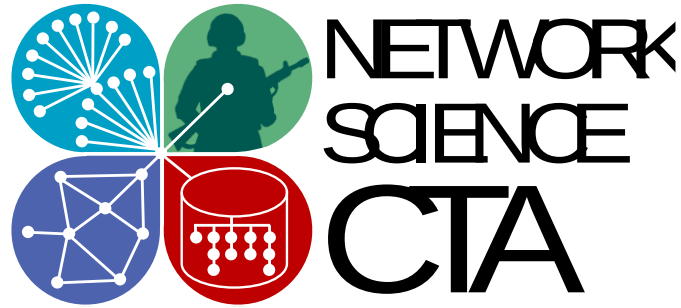
Y. Sun: SIGKDD'13 Dissertation Award

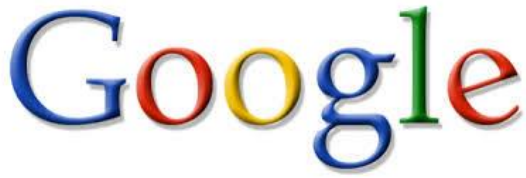
Wang and Han, Mining Latent Entity Structures, 2015

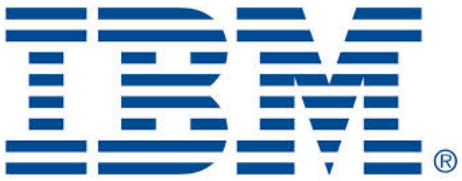
C. Wang: SIGKDD'15 Dissertation Award

Acknowledgement

- Thanks for the research support from: NIH, ARL/NSCTA, NSF, DHS, ARO, DTRA







References

- A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, "Scalable Topical Phrase Mining from Text Corpora", VLDB'15
- J. Liu, J. Shang, C. Wang, X. Ren, J. Han, "Mining Quality Phrases from Massive Text Corpora", SIGMOD'15
- J. Liu, X. Ren, J. Shang, T. Cassidy, C. Voss and J. Han, "Representing Documents via Latent Keyphrase Inference", WWW'16
- X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, H. Ji and J. Han, "ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering", KDD'15
- X. Ren, W. He, M. Qu, C. R. Voss, H. Ji, J. Han, "Label Noise Reduction in Entity Typing by Heterogeneous Partial-Label Embedding", KDD'16
- Y. Sun and J. Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Morgan & Claypool Publishers, 2012
- J. Tang, M. Qu, M. Zhang, Q. Mei, Large-scale Information Network Embedding, WWW'15
- C. Wang and J. Han, Mining Latent Entity Structures, Morgan & Claypool, 2015