From Unstructured Text to TextCube: Automated Construction and Multidimensional Exploration

JIAWEI HAN
COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

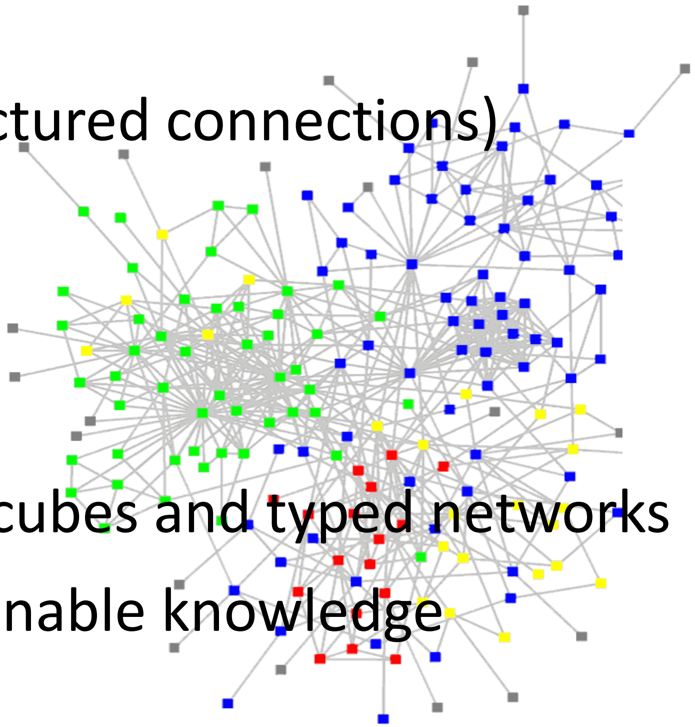
NOVEMBER 15, 2019

Outline

- On the Power of Multi-Dimensional Text Cubes
- Automated Mining of Semantic Structures from Massive Text Data
    - Phrase Mining
    - Entity/Relation Recognition and Typing
    - Meta Pattern-Directed Structure Discovery
- Automated Construction of Multidimensional Text Cubes
    - Multifaceted Taxonomy Mining
    - Doc2Cube: Constructing TextCube from Massive Documents
    - Quality Enhancement: Local and Global Joint Spherical Text Embedding
- Looking Forward

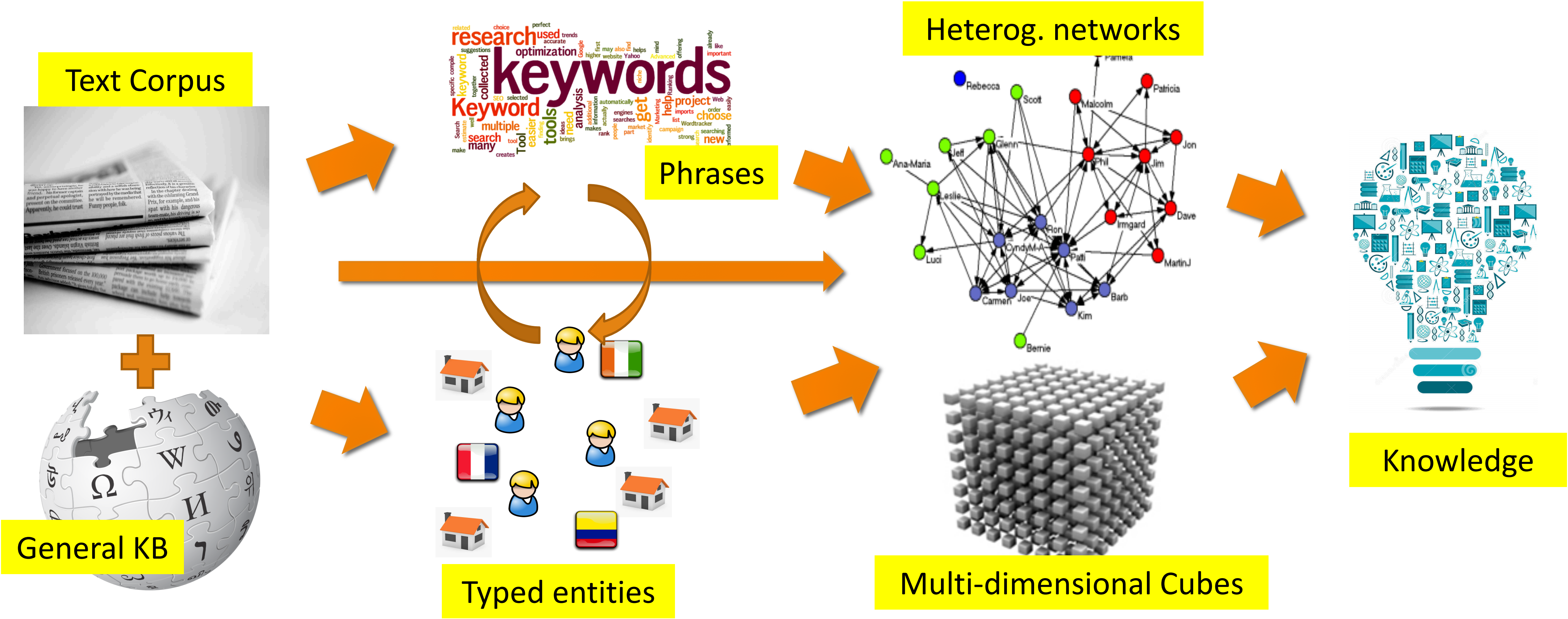
From Big Data to Big Knowledge: Taming Text is the Key

- Ubiquity of big unstructured data
  - Big Data: Over 80% of our data is from text/natural language/social media, unstructured/semi-structured, noisy, dynamic, …, but inter-related!
- How to mine such big data systematically?
  - Structuring (i.e., transforming unstructured text into structured, typed, interconnected entities/relationships)
  - Networking (take advantage of massive, structured connections)
  - Mining massive structures and networks
- Our roadmap:
  - Mining hidden structures from text data
  - Turning text data into multidimensional text-cubes and typed networks
  - Mining cubes and networks to generate actionable knowledge



Bottleneck: Mining Unstructured Text for Structures

- ❑ One of the most challenging issues at mining big data: structuring and mining text!!
- ❑ Bottleneck: How to automatically generate structures from text data?
  - ❑ Automated mining of phrases, topics, entities, links and types from text corpora



Text Corpus

General KB

Phrases

Typed entities

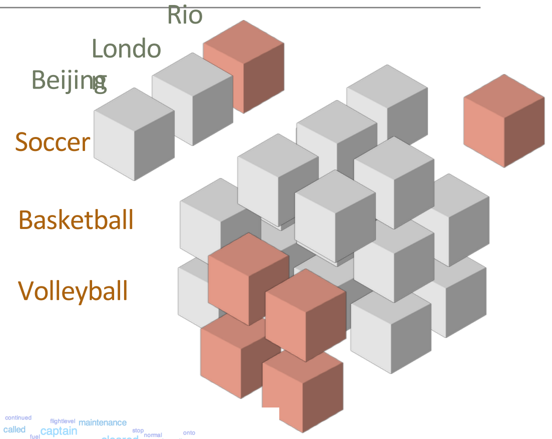
Heterog. networks

Multi-dimensional Cubes

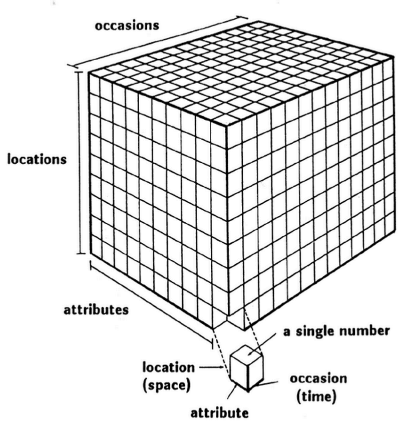
Knowledge

The Power of Text Cube: Multi-Dimensional Text Analysis

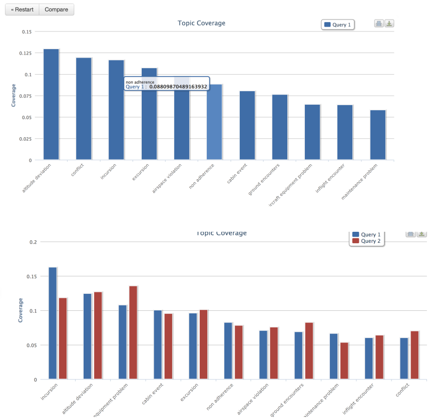
- From TextCube to EventCube [KDD’13 demo]
  - Keyword- or entity-based search or summary of documents
- CASeOLAP [EngBul’16]: Comparative summary/mining



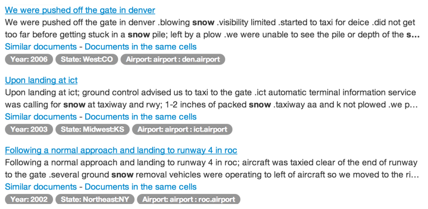
Cube Structure



Slice
Roll-up
Drill-down
Dice
...



runway
aircraft
flight
taxiway
controller
approach
captain
traffic
tower
landing
control
ground
position
system
time
feet
cleared
passenger
officer
final
short
departure
degrees
approximate
takeoff
knots
flying
miles



We were pushed off the gate in denver

We were pushed off the gate in denver .blowing **snow** .visibility limited .started to taxi for deice .did not get too far before getting stuck in a **snow** pile; left by a plow .we were unable to see the pile or depth of the **s...**

Similar documents - Documents in the same cells

Year: 2006 State: West:CO Airport: airport : den.airport

Upon landing at ict

Upon landing at ict; ground control advised us to taxi to the gate .ict automatic terminal information service was calling for **snow** at taxiway and rwy; 1-2 inches of packed **snow** .taxiway aa and k not plowed .we p...

Similar documents - Documents in the same cells

Year: 2003 State: Midwest:KS Airport: airport : ict.airport

Following a normal approach and landing to runway 4 in roc

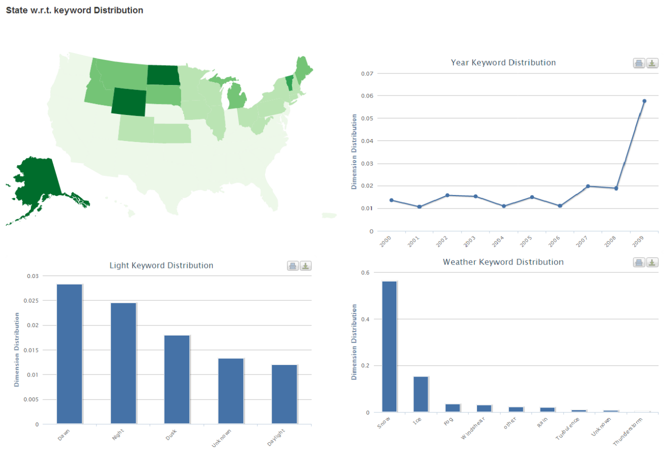
Following a normal approach and landing to runway 4 in roc; aircraft was taxied clear of the end of runway to the gate .several ground **snow** removal vehicles were operating to left of aircraft so we moved to the ri...

Similar documents - Documents in the same cells

Year: 2002 State: Northeast:NY Airport: airport : roc.airport

Textual Analysis

Text Data



Structural Analysis

Effectiveness of Comparative Summary on Real-World Cases

Contrasting analysis

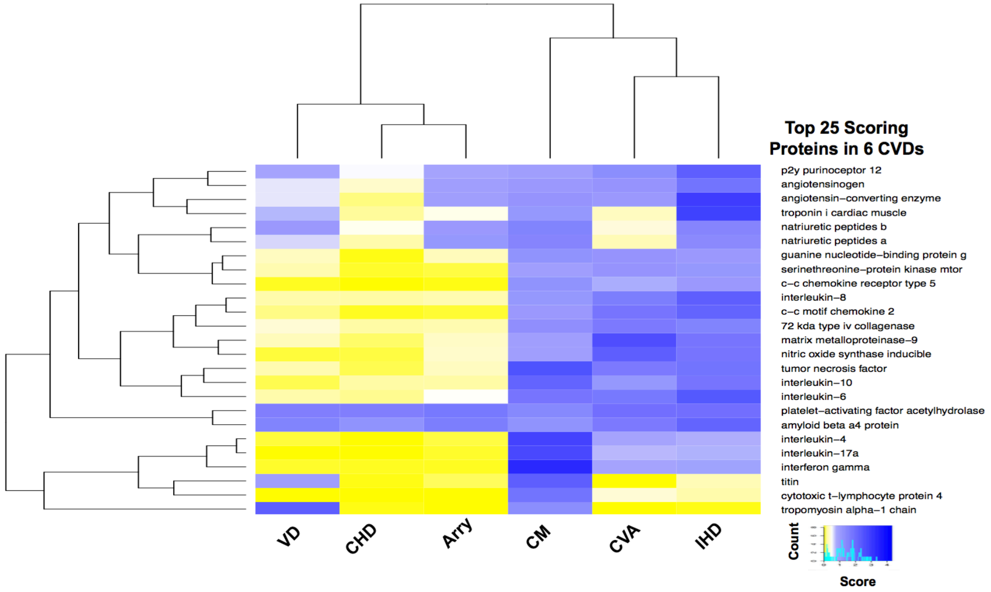
- Integrity
- Popularity
- distinctness

Mining distinct phrases: 2016 news data

⟨US, Gun Control⟩	⟨US, Immigration⟩	⟨US, Domestic Politics⟩	⟨US, Law and Crime⟩	⟨US, Military⟩
gun laws	immigration debate	gun laws	district attorney	sexual assault in the military
the national rifle association	border security	insurance plans	shot and killed	military prosecutors
gun rights	guest worker program	background check	federal court	armed services committee
background check	immigration legislation	health coverage	life in prison	armed forces
gun owners	undocumented immigrants	tax increases	death row	defense secretary
assault weapons ban	overhaul of the nation's immigration laws	the national rifle association	grand jury	military personnel
mass shootings	legal status	assault weapons ban	department of justice	sexually assaulted
high capacity magazines	path to citizenship	immigration debate	child abuse	fort meade
gun legislation	immigration status	the federal exchange	plea deal	private manning
gun control advocates	immigration reform	medicaid program	second degree murder	pentagon officials

Mining Distinct relationships between 6 subcategories of **cardiovascular diseases** and **proteins**: PubMed Abstracts

Disease	Top Ranked Molecules and their scores
Cerebrovascular Accident	Alpha-galactosidase A, Brain-derived Neurotrophic Factor, Tissue-type Plasminogen Activator Methylenetetrahydrofolate Reductase, 5.903, 5.595, 4.945, 2.710, 2.680
Ischemic Heart Disease	Cholesteryl Ester Transfer Protein, Apoli Myeloperoxidase 4.597, 3.989, 3.651, 3.302, 3.240
Cardiomyopathy	Interferon Gamma, Interleukin-4, Interl 3.336, 2.809, 2.729, 2.549, 2.349
Arrhythmia	Methionine Synthase, Ryanodine Recep Potassium Voltage-gated Channel Subfa 3.799, 3.354, 1.740, 2.730, 1.872
Valve Dysfunction	Mineralocorticoid Receptor, Elastin, Tro Myosin-Binding Protein C Cardiac-type, 3.276, 2.380, 2.332, 1.704, 1.611
Congenital Heart Disease	Fibrillin-1, Plakophilin-2, Tyrosine-prote Arachidonate 5-Lipoxygenase-activating 4.920, 3.208, 2.667, 2.036, 1.791

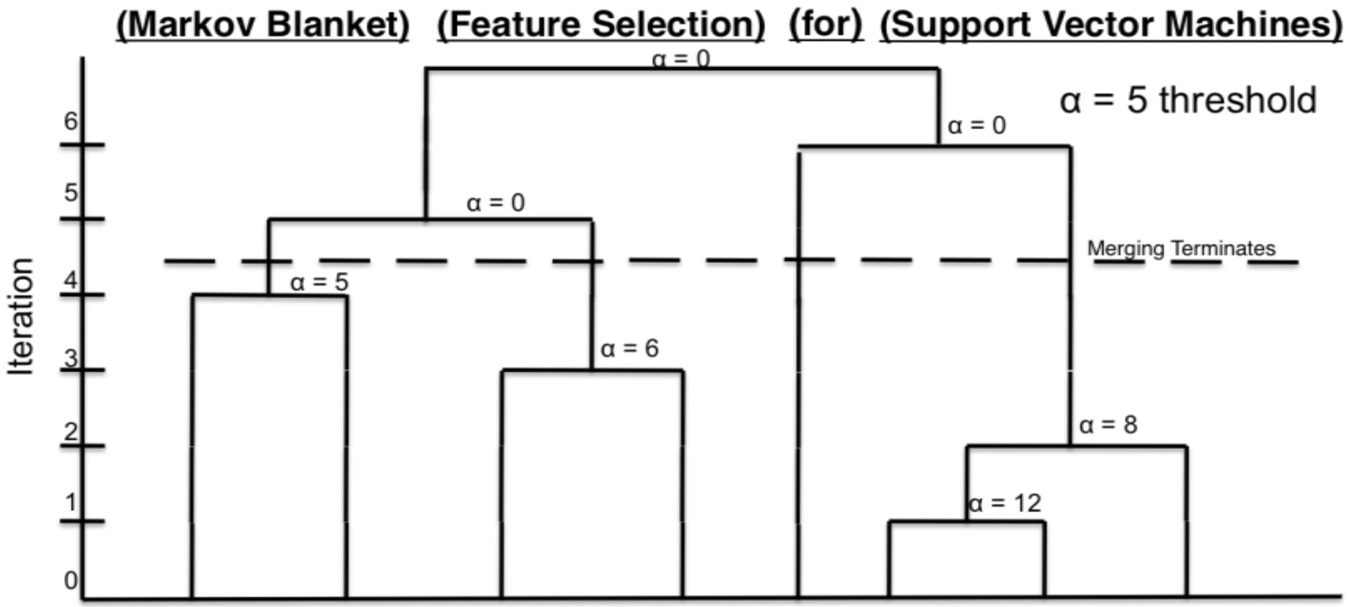


Outline

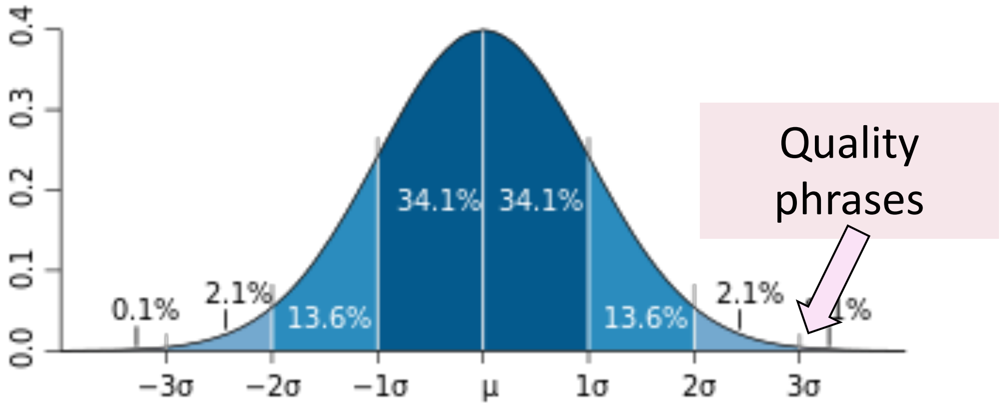
- On the Power of Multi-Dimensional Text Cubes
- Automated Mining of Semantic Structures from Massive Text Data
  - Phrase Mining
  - Entity/Relation Recognition and Typing
  - Meta Pattern-Directed Structure Discovery
- Automated Construction of Multidimensional Text Cubes
  - Multifaceted Taxonomy Mining
  - Doc2Cube: Constructing TextCube from Massive Documents
  - Quality Enhancement: Local and Global Joint Spherical Text Embedding
- Looking Forward

TopMine: 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.



Quality phrases

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

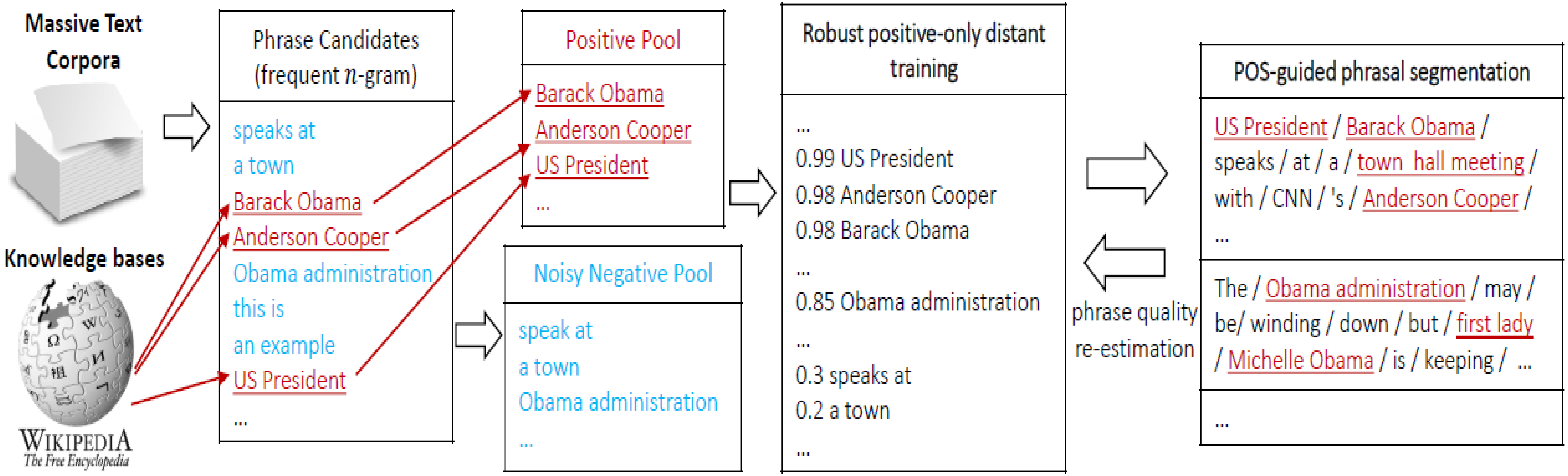
$$\alpha(P_1, P_2) \approx (f(P_1 \bullet P_2) - \mu_0(P_1, P_2)) / \surd\, f(P_1 \bullet P_2)$$

[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

AutoPhrase: Automated Phrase Mining

- **ToPMing (unsupervised) [VLDB'14] → SegPhrase (weakly supervised) [SIGMOD'15] → AutoPhrase (distantly supervised) [TKDE''18]**
- Automatic extraction of high-quality phrases (e.g., scientific terms and general entity names) in a given corpus (e.g., research papers and news)
  - No human efforts / Multiple languages / High performance—precision, recall, efficiency



Massive Text Corpora

Knowledge bases

WIKIPEDIA
The Free Encyclopedia

Phrase Candidates (frequent *n*-gram)

speaks at
a town
Barack Obama
Anderson Cooper
Obama administration
this is
an example
US President
…

Positive Pool

Barack Obama
Anderson Cooper
US President
…

Noisy Negative Pool

speak at
a town
Obama administration
…

Robust positive-only distant training

…
0.99 US President
0.98 Anderson Cooper
0.98 Barack Obama
…
0.85 Obama administration
…
0.3 speaks at
0.2 a town
…

phrase quality re-estimation

POS-guided phrasal segmentation

US President / Barack Obama / speaks / at / a / town hall meeting / with / CNN / 's / Anderson Cooper /
…

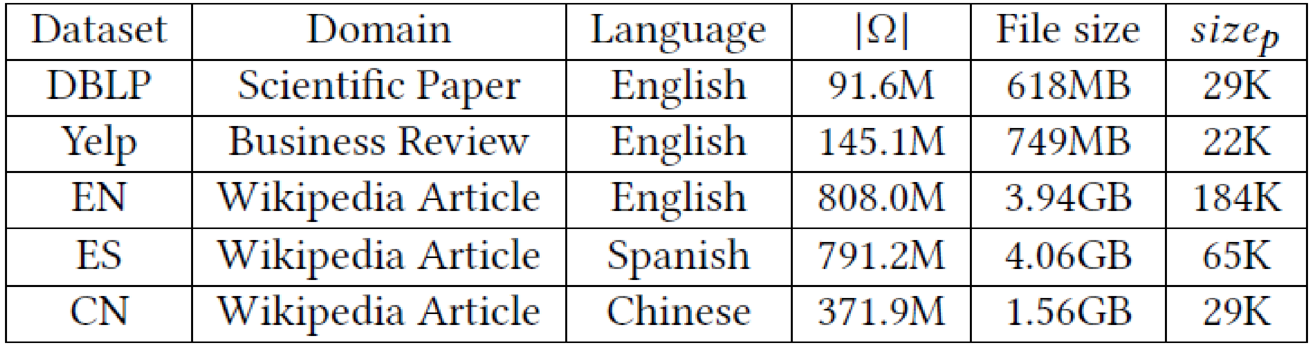
The / Obama administration / may / be/ winding / down / but / first lady / Michelle Obama / is / keeping / …

…

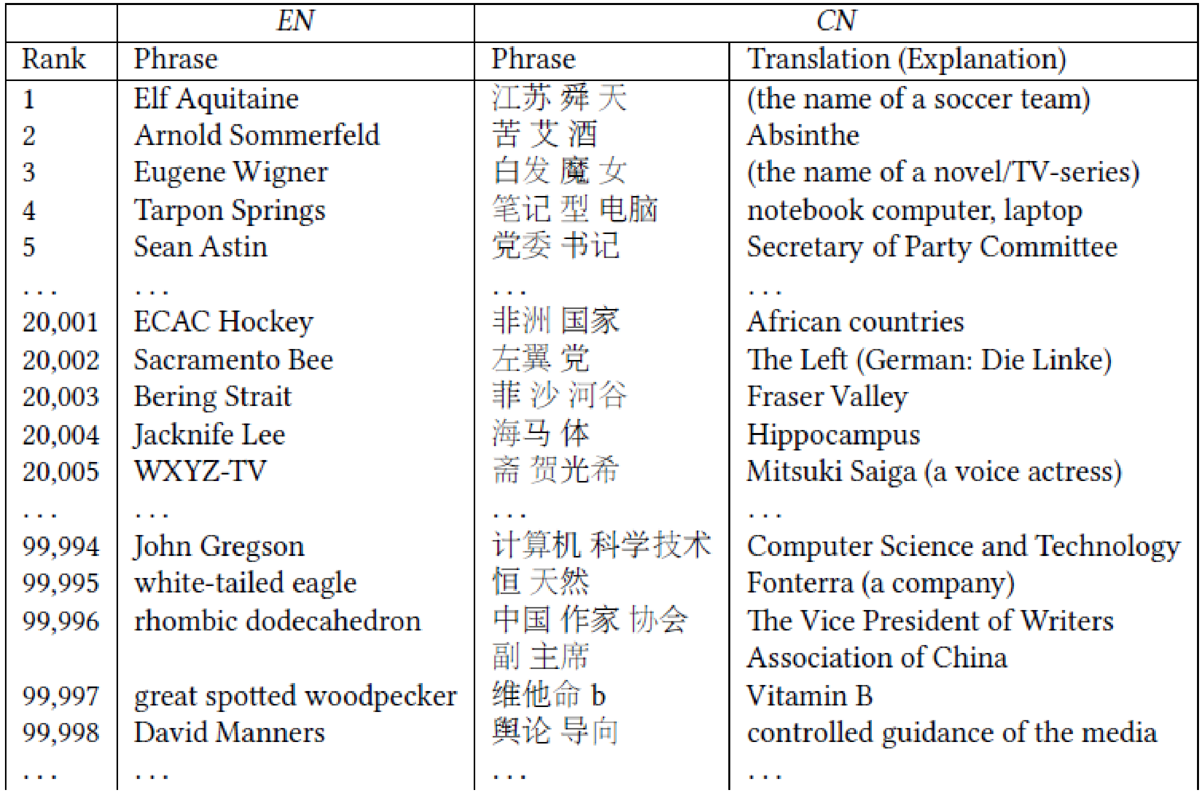
Experiments and Performance Comparison

- Datasets:

Phrase Mining Results

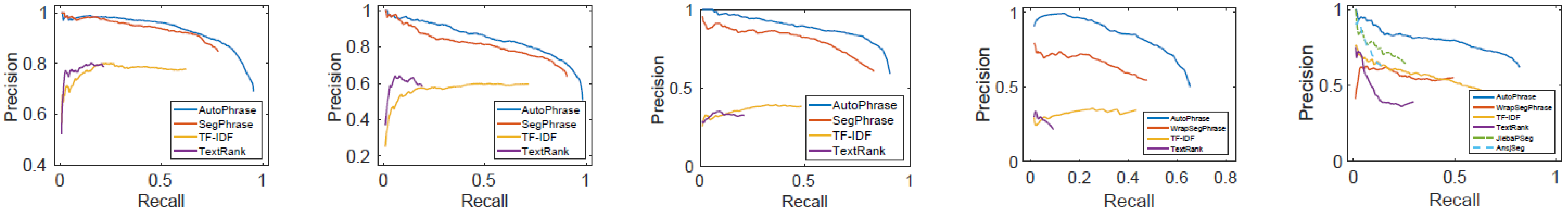


Dataset	Domain	Language	$\|\Omega\|$	File size	$size_p$
DBLP	Scientific Paper	English	91.6M	618MB	29K
Yelp	Business Review	English	145.1M	749MB	22K
EN	Wikipedia Article	English	808.0M	3.94GB	184K
ES	Wikipedia Article	Spanish	791.2M	4.06GB	65K
CN	Wikipedia Article	Chinese	371.9M	1.56GB	29K



	EN	*CN*	
Rank	Phrase	Phrase	Translation (Explanation)
1	Elf Aquitaine	江苏 舜 天	(the name of a soccer team)
2	Arnold Sommerfeld	苦 艾 酒	Absinthe
3	Eugene Wigner	白发 魔 女	(the name of a novel/TV-series)
4	Tarpon Springs	笔记 型 电脑	notebook computer, laptop
5	Sean Astin	党委 书记	Secretary of Party Committee
...	...	...	...
20,001	ECAC Hockey	非洲 国家	African countries
20,002	Sacramento Bee	左翼 党	The Left (German: Die Linke)
20,003	Bering Strait	菲 沙 河谷	Fraser Valley
20,004	Jacknife Lee	海马 体	Hippocampus
20,005	WXYZ-TV	斋 贺光希	Mitsuki Saiga (a voice actress)
...	...	...	...
99,994	John Gregson	计算机 科学技术	Computer Science and Technology
99,995	white-tailed eagle	恒 天然	Fonterra (a company)
99,996	rhombic dodecahedron	中国 作家 协会 副 主席	The Vice President of Writers Association of China
99,997	great spotted woodpecker	维他命 b	Vitamin B
99,998	David Manners	舆论 导向	controlled guidance of the media
...	...	...	...

- Comparing methods
  - SegPhrase/WrapSegPhrae (encoding preprocessing for handling non-English)
  - TF-IDF/TextRank



(a) DBLP (b) Yelp (c) EN (d) ES (e) CN

Outline

- On the Power of Multi-Dimensional Text Cubes
- Automated Mining of Semantic Structures from Massive Text Data
    - Phrase Mining
    - Entity/Relation Recognition and Typing
    - Meta Pattern-Directed Structure Discovery
- Automated Construction of Multidimensional Text Cubes
    - Multifaceted Taxonomy Mining
    - Doc2Cube: Constructing TextCube from Massive Documents
    - Quality Enhancement: Local and Global Joint Spherical Text Embedding
- Looking Forward

Recognizing Typed Entities

Identifying token span as entity mentions in documents and labeling their types
— Enabling structured analysis of unstructured text corpus

FOOD
LOCATION
JOB_TITLE
EVENT
ORGANIZATION
…

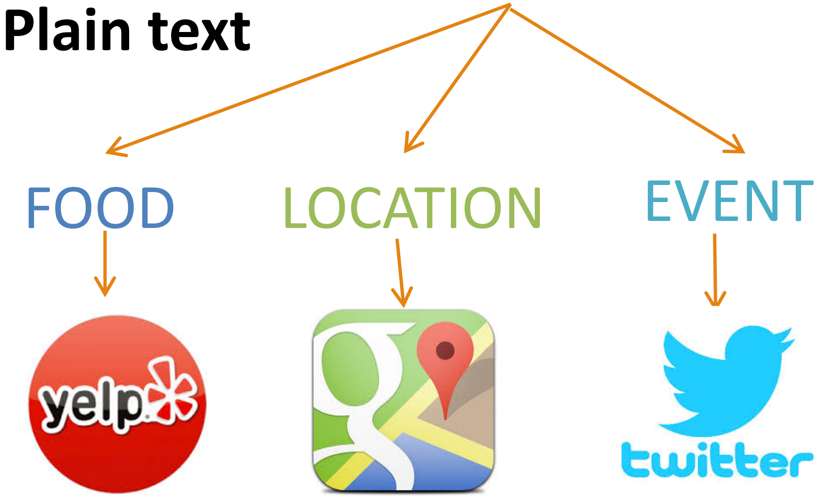
Target Types

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



Social media challenge!

Can we use the "distant labels" in the KBs?

Traditional methods:

Expensive human labor on annotation of 500 documents for entity extraction and 20,000 queries for entity linking

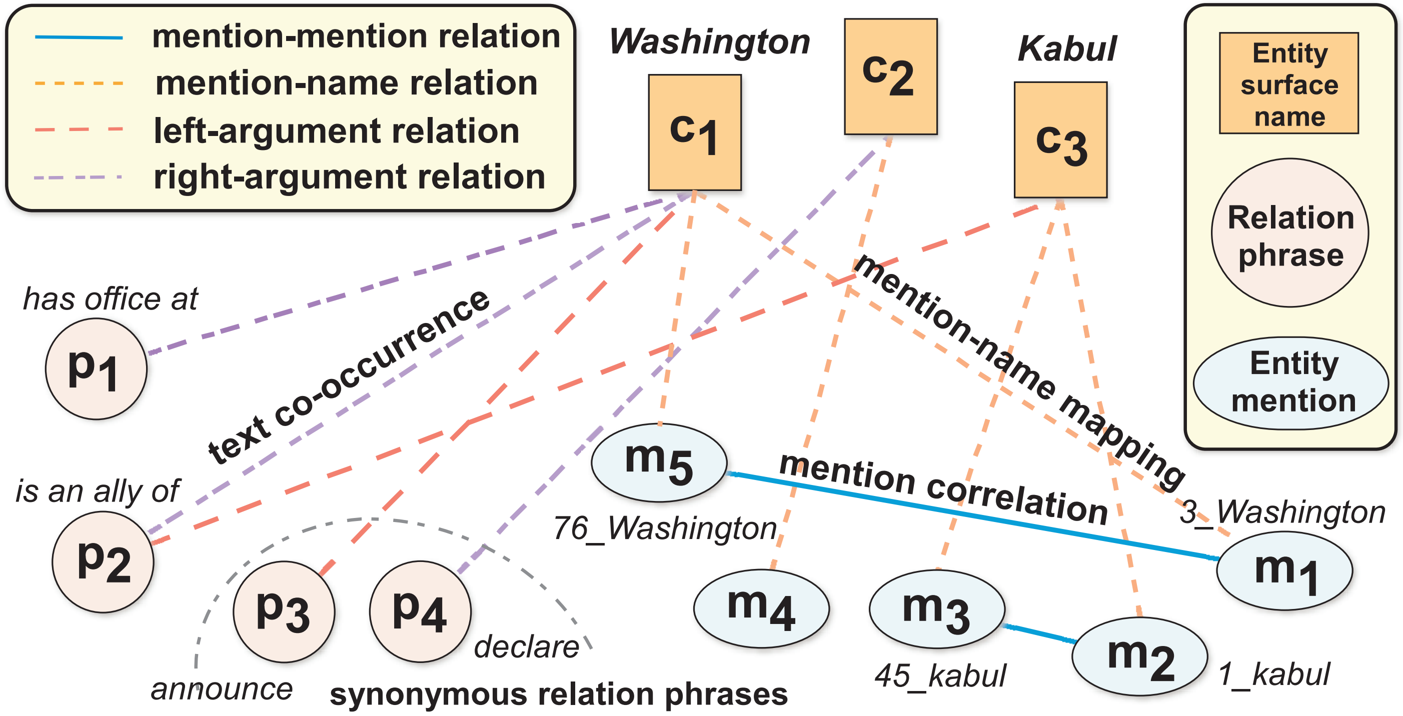
The ClusType Framework: Phrase Segmentation and Heterogeneous Graph Construction [KDD'15]

- 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**

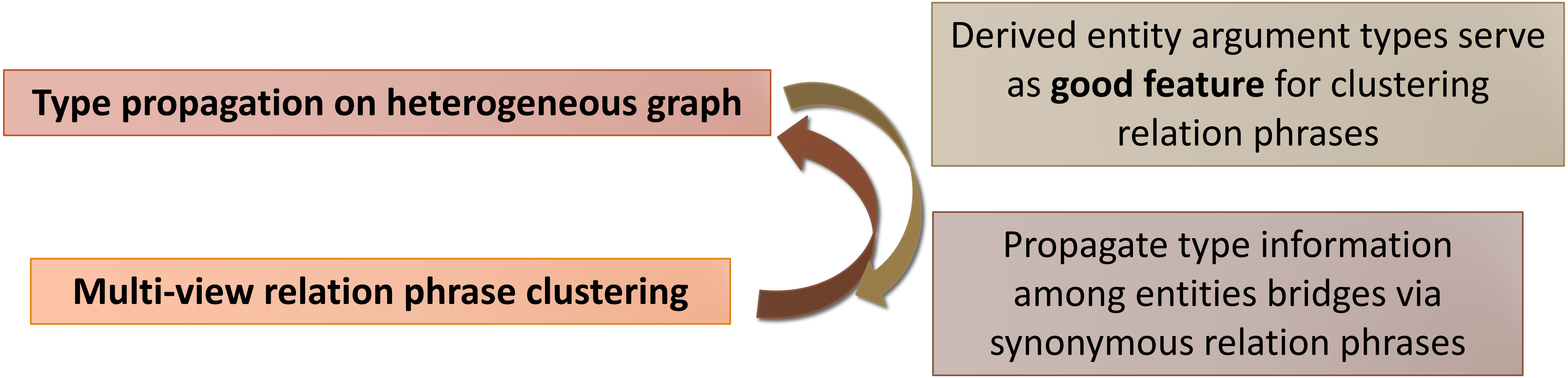
Linked to entity surface names & relation phrases

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



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

Multi-view relation phrase clustering

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

Propagate type information among entities bridges via synonymous relation phrases

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

ClusType: Comparing with State-of-the-Art Systems

Bootstrapping

Label propagation

Classifier with linguistic features

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

F1-score

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

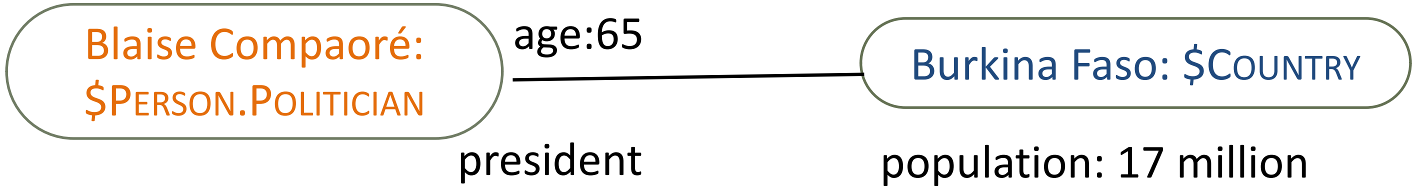
Outline

- On the Power of Multi-Dimensional Text Cubes
- Automated Mining of Semantic Structures from Massive Text Data
  - Phrase Mining
  - Entity/Relation Recognition and Typing
  - Meta Pattern-Directed Structure Discovery
- Automated Construction of Multidimensional Text Cubes
  - Multifaceted Taxonomy Mining
  - Local and Global Joint Spherical Text Embedding
- Looking Forward

MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora [KDD'17]

Given a sentence in a large corpus, "President Blaise Compaoré's government of **Burkina Faso** was founded…", …

Can we find:



Blaise Compaoré: $PERSON.POLITICIAN

age:65

president

Burkina Faso: $COUNTRY

population: 17 million

- ❑ Attribute Discovery: Two tasks

Task 1: ⟨entity, attribute name, attribute value⟩

⟨Burkina Faso, president, Blaise Compaoré⟩

⟨Burkina Faso, population, 17 million⟩

⟨Blaise Compaoré, age, 65⟩

Instance-level

Task 2: ⟨entity type, attribute name⟩

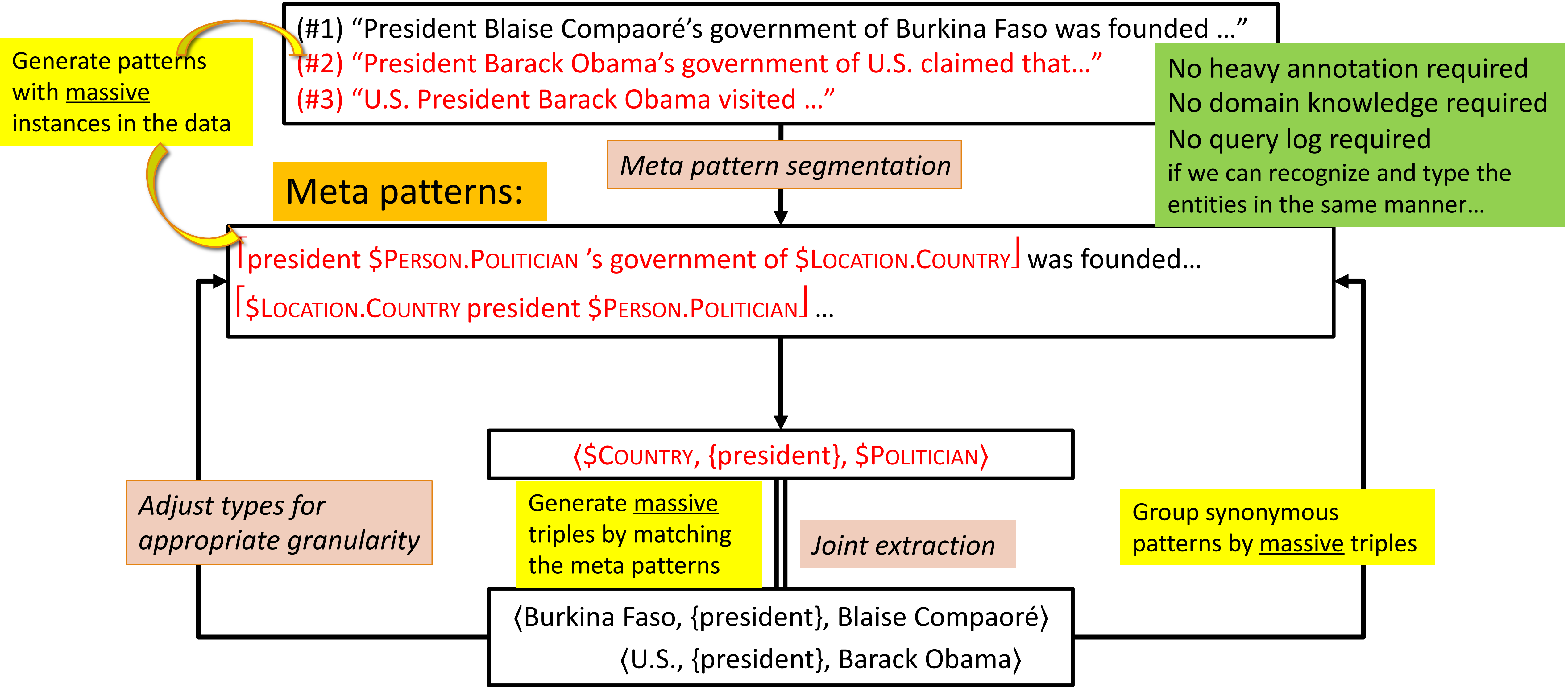
⟨$COUNTRY, president⟩

⟨$COUNTRY, population⟩

⟨$PERSON, age⟩

Type-level

The Meta-Pattern Methodology



Generate patterns with massive instances in the data

(#1) “President Blaise Compaoré’s government of Burkina Faso was founded …”
(#2) “President Barack Obama’s government of U.S. claimed that…”
(#3) “U.S. President Barack Obama visited …”

No heavy annotation required
No domain knowledge required
No query log required
if we can recognize and type the entities in the same manner…

Meta pattern segmentation

Meta patterns:

⌈president $Person.Politician ’s government of $Location.Country⌋ was founded…
⌈$Location.Country president $Person.Politician⌋ …

⟨$Country, {president}, $Politician⟩

Adjust types for appropriate granularity

Generate massive triples by matching the meta patterns

Joint extraction

Group synonymous patterns by massive triples

⟨Burkina Faso, {president}, Blaise Compaoré⟩
⟨U.S., {president}, Barack Obama⟩

Patterns, Entities and Attribute Values Found in News Corpus

Meta patterns	Entity	Attribute value
$Country President $Politician $Country’s president $Politician President $Politician of $Country … $Politician’s government of $Country	United States	Barack Obama
	Russia	Vladimir Putin
	France	Francois Hollande
	…	…
	Burkina Faso	Blaise Compaoré

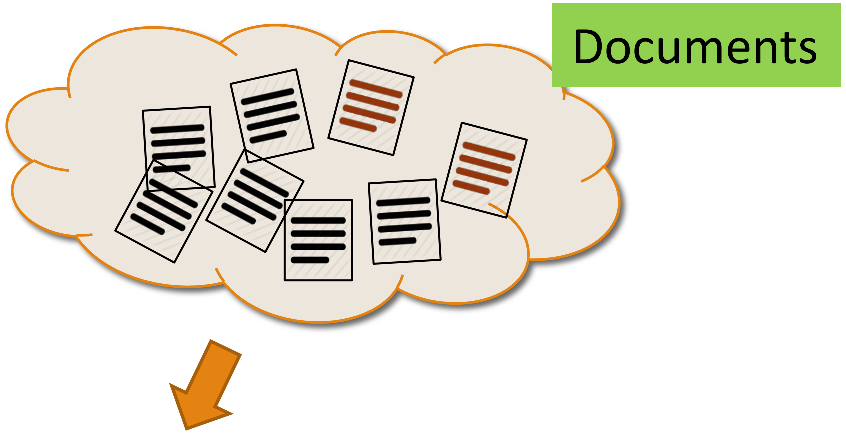
Meta patterns	Entity	Attribute value
$Company CEO $Person $Company chief executive $Person $Person, the $Company CEO, … $Company former CEO $Person $Person, the $Company former CEO,	Apple	Tim Cook
	Facebook	Mark Zuckerberg
	Hewlett-Packard	Carly Fiorina
	…	…
	Infor	Charles Phillips
	Afghan Citadel	Roya Mahboob

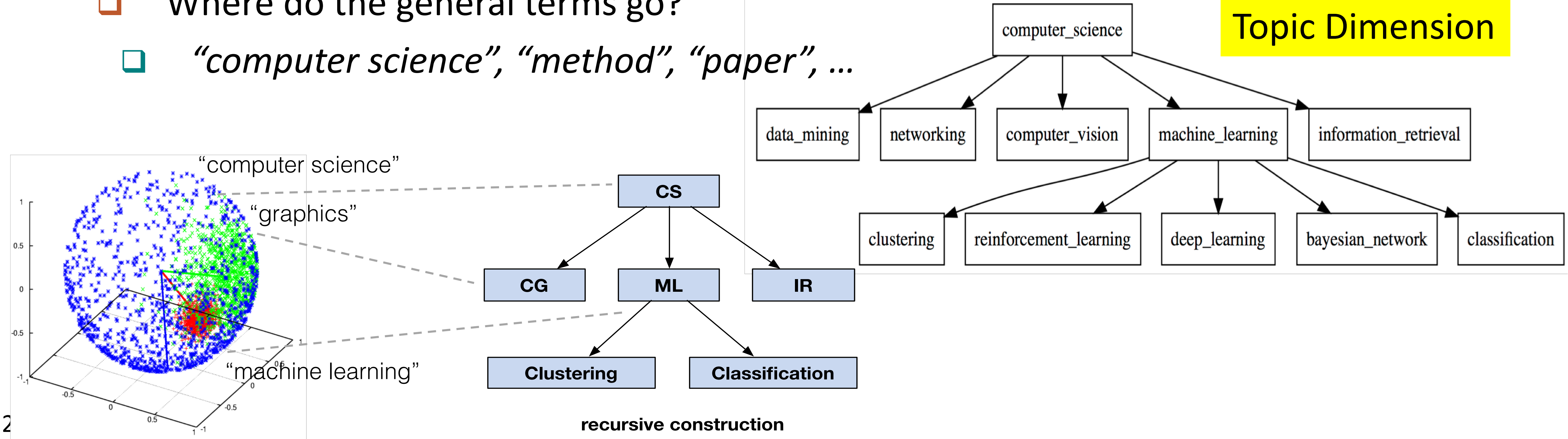
Outline

- ❑ On the Power of Multi-Dimensional Text Cubes
- ❑ Automated Mining of Semantic Structures from Massive Text Data
    - ❑ Phrase Mining
    - ❑ Entity/Relation Recognition and Typing
    - ❑ Meta Pattern-Directed Structure Discovery
- ❑ Automated Construction of Multidimensional Text Cubes
    - ❑ Multifaceted Taxonomy Mining
    - ❑ Doc2Cube: Constructing TextCube from Massive Documents
    - ❑ Quality Enhancement: Local and Global Joint Spherical Text Embedding
- ❑ Looking Forward

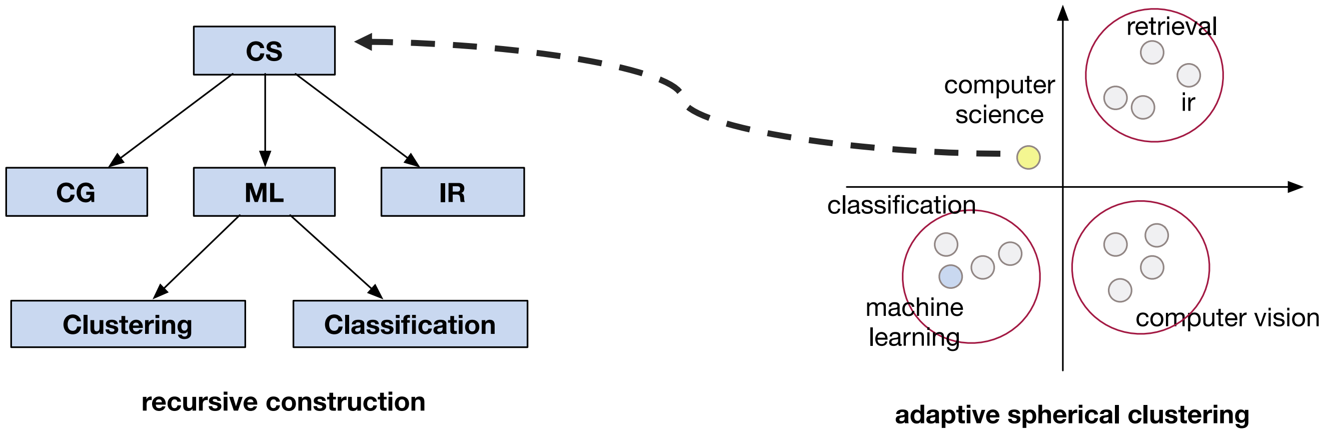
Taxonomy Generation from Massive Text Corpora

- Automated construction of topic taxonomy
- Selected method: **spherical clustering**—Use **embeddings** to find semantically consistent clusters
  - Domain-specific terms can be clustered together
    - *"machine learning", "learning algorithm", ...*
  - Where do the general terms go?
    - *"computer science", "method", "paper", ...*





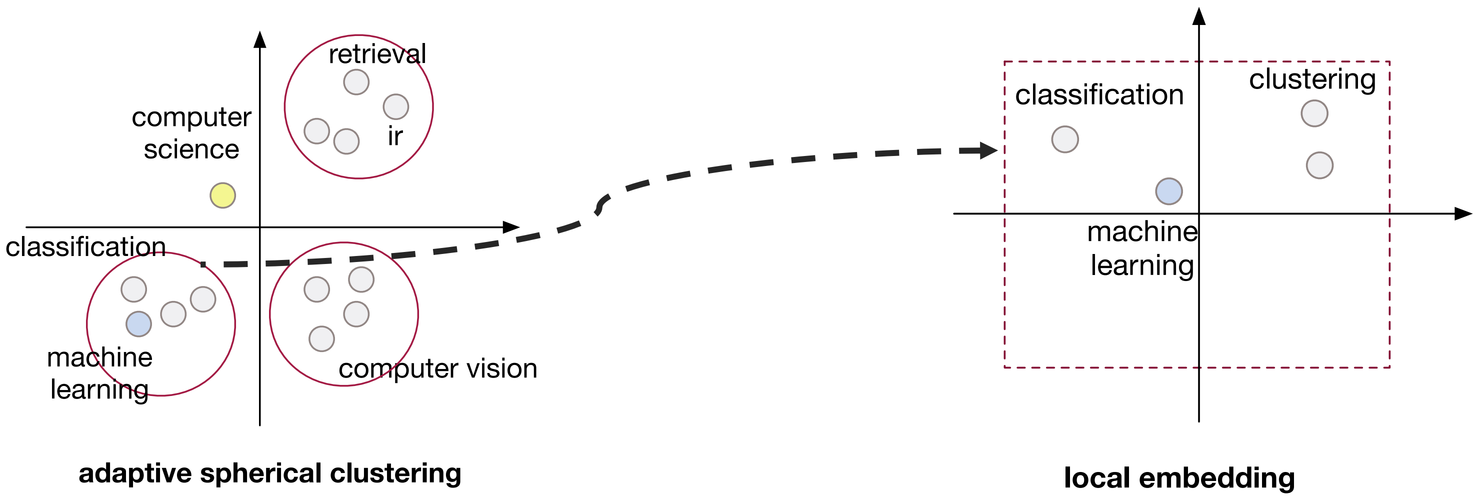
TaxoGen [KDD'18]: Adaptive Spherical Clustering



- ❑ Design a ranking module to select ***representative phrases*** for each cluster
  - ❑ Conduct comparative analysis (combining **popularity** and **concentration)**
    - ❑ Does this phrase better fit my cluster or my sliblings'?
- ❑ Push the ***background phrases*** back to the general node
  - ❑ "computer science", "paper" → the higher-level node (root node)
  - ❑ "machine learning", "ml", "classification" → the "ML" node
- ❑ The set of remaining phrases leads to more separable clustering

TaxoGen: Local Embedding vs. Global Embedding

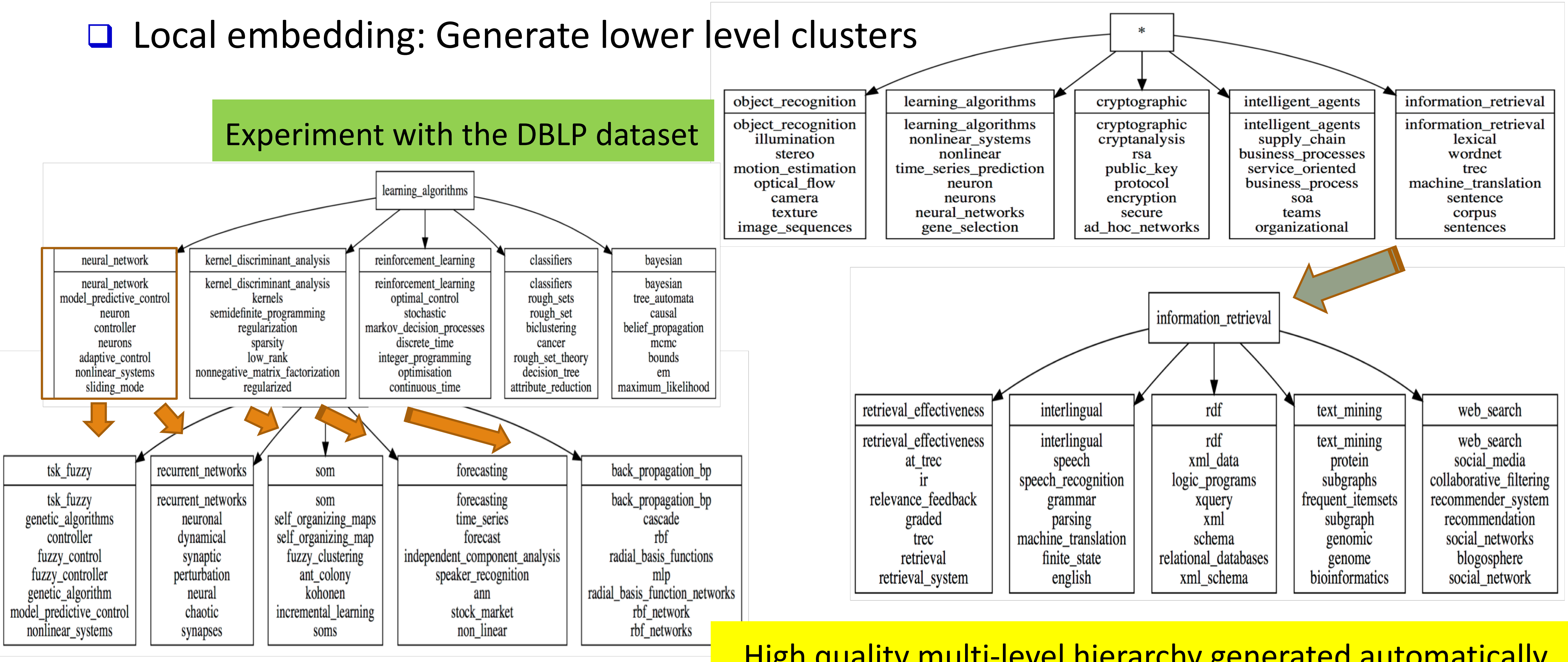
- Global embedding (embedding learning on the global dataset) does not work
  - Terms at different granularity can have close embeddings
- Ex. “Information Extraction”: similar to *“text mining”, “NLP”, “machine learning”*
- Solution: local-corpus embedding:
  - For each “sub-topic” node, learn ***local embedding*** only on relevant documents
  - Only perserve information relevant to the “sub-topic”



TaxonGen: Adaptive Spherical Clustering + Local Embedding

- ❑ Phrase mining + Adaptive spherical clustering: Generate top-level clusters
- ❑ Local embedding: Generate lower level clusters

Experiment with the DBLP dataset



*

object_recognition
object_recognition
illumination
stereo
motion_estimation
optical_flow
camera
texture
image_sequences

learning_algorithms
learning_algorithms
nonlinear_systems
nonlinear
time_series_prediction
neuron
neurons
neural_networks
gene_selection

cryptographic
cryptographic
cryptanalysis
rsa
public_key
protocol
encryption
secure
ad_hoc_networks

intelligent_agents
intelligent_agents
supply_chain
business_processes
service_oriented
business_process
soa
teams
organizational

information_retrieval
information_retrieval
lexical
wordnet
trec
machine_translation
sentence
corpus
sentences

information_retrieval

retrieval_effectiveness
retrieval_effectiveness
at_trec
ir
relevance_feedback
graded
trec
retrieval
retrieval_system

interlingual
interlingual
speech
speech_recognition
grammar
parsing
machine_translation
finite_state
english

rdf
rdf
xml_data
logic_programs
xquery
xml
schema
relational_databases
xml_schema

text_mining
text_mining
protein
subgraphs
frequent_itemsets
subgraph
genomic
genome
bioinformatics

web_search
web_search
social_media
collaborative_filtering
recommender_system
recommendation
social_networks
blogosphere
social_network

learning_algorithms

neural_network
neural_network
model_predictive_control
neuron
controller
neurons
adaptive_control
nonlinear_systems
sliding_mode

kernel_discriminant_analysis
kernel_discriminant_analysis
kernels
semidefinite_programming
regularization
sparsity
low_rank
nonnegative_matrix_factorization
regularized

reinforcement_learning
reinforcement_learning
optimal_control
stochastic
markov_decision_processes
discrete_time
integer_programming
optimisation
continuous_time

classifiers
classifiers
rough_sets
rough_set
biclustering
cancer
rough_set_theory
decision_tree
attribute_reduction

bayesian
bayesian
tree_automata
causal
belief_propagation
mcmc
bounds
em
maximum_likelihood

tsk_fuzzy
tsk_fuzzy
genetic_algorithms
controller
fuzzy_control
fuzzy_controller
genetic_algorithm
model_predictive_control
nonlinear_systems

recurrent_networks
recurrent_networks
neuronal
dynamical
synaptic
perturbation
neural
chaotic
synapses

som
som
self_organizing_maps
self_organizing_map
fuzzy_clustering
ant_colony
kohonen
incremental_learning
soms

forecasting
forecasting
time_series
forecast
independent_component_analysis
speaker_recognition
ann
stock_market
non_linear

back_propagation_bp
back_propagation_bp
cascade
rbf
radial_basis_functions
mlp
radial_basis_function_networks
rbf_network
rbf_networks

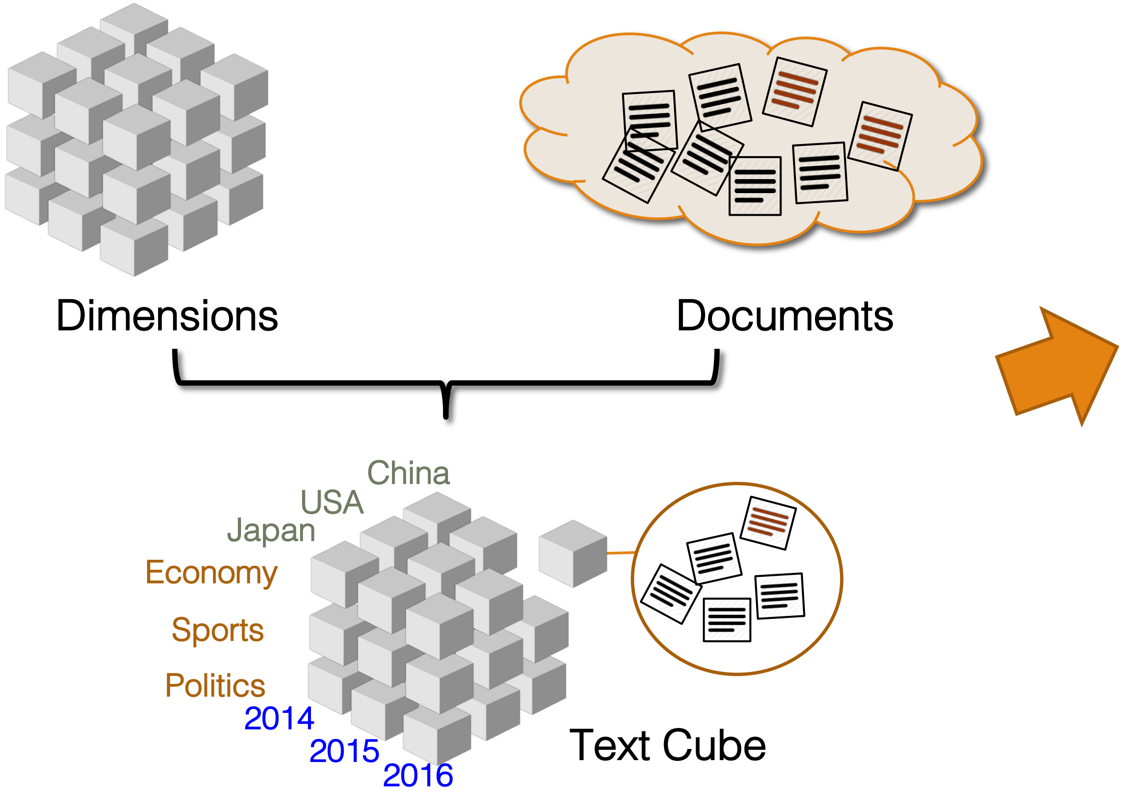
High quality multi-level hierarchy generated automatically

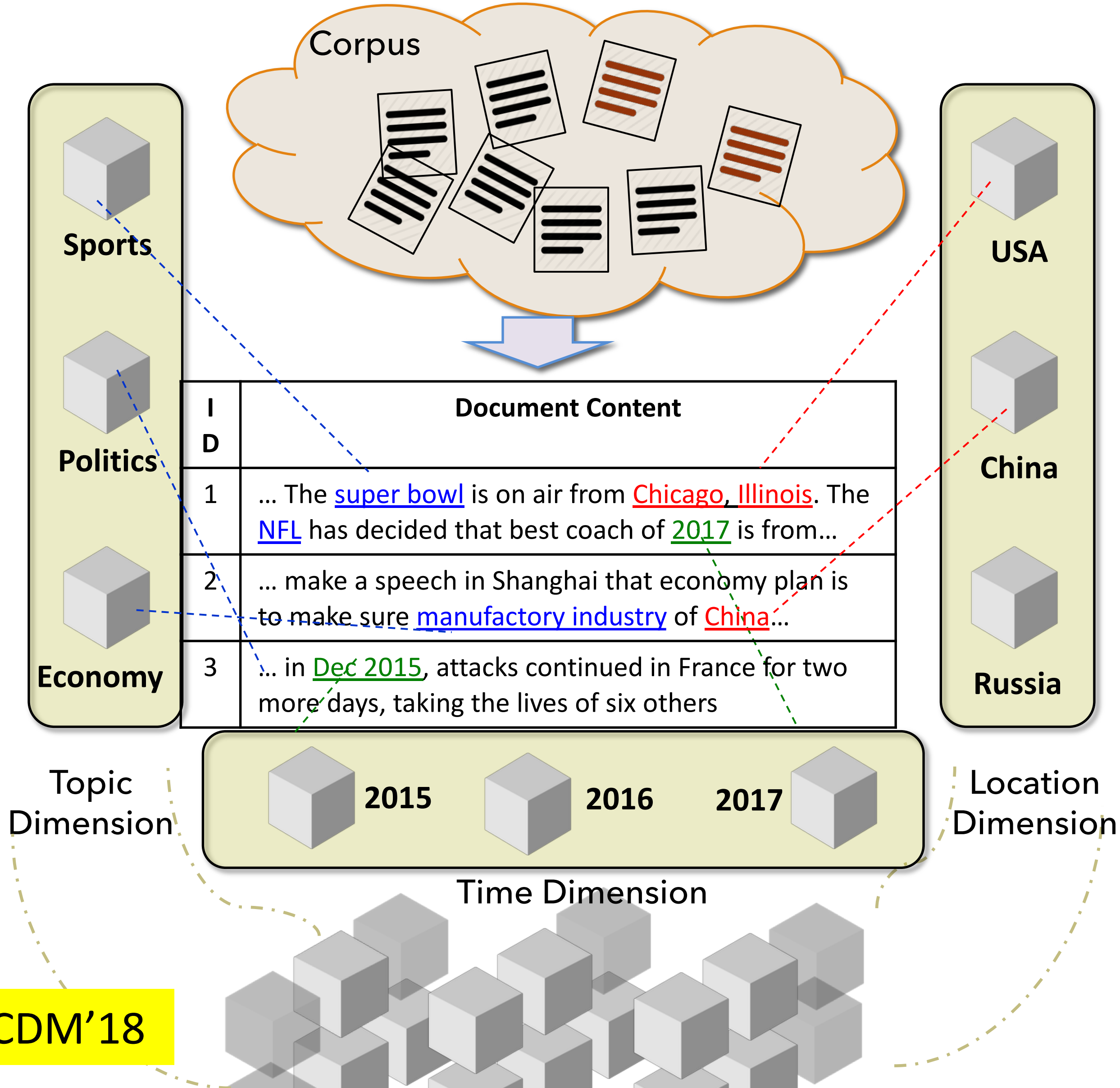
Outline

- On the Power of Multi-Dimensional Text Cubes
- Automated Mining of Semantic Structures from Massive Text Data
  - Phrase Mining
  - Entity/Relation Recognition and Typing
  - Meta Pattern-Directed Structure Discovery
- Automated Construction of Multidimensional Text Cubes
  - Multifaceted Taxonomy Mining
  - Doc2Cube: Constructing TextCube from Massive Documents
  - Quality Enhancement: Local and Global Joint Spherical Text Embedding
- Looking Forward

Cube Construction: Which Document Goes to Which Cell?

- ❑ Cell-based Document Allocation
  - ❑ Which document goes to which cell?



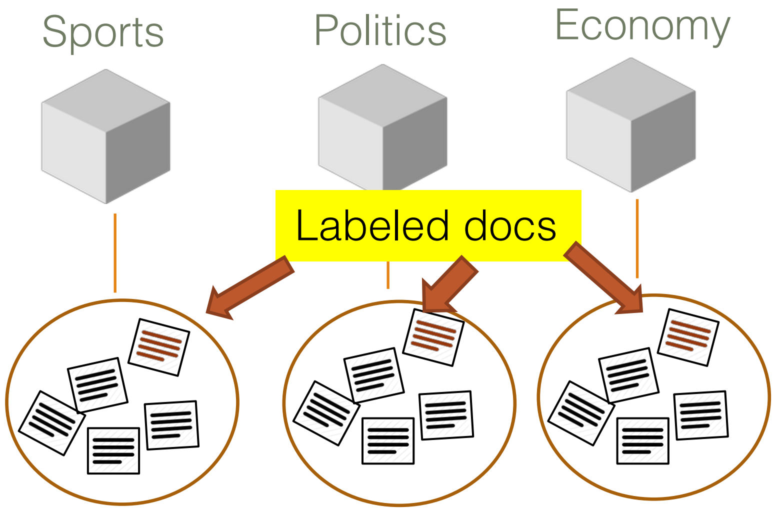


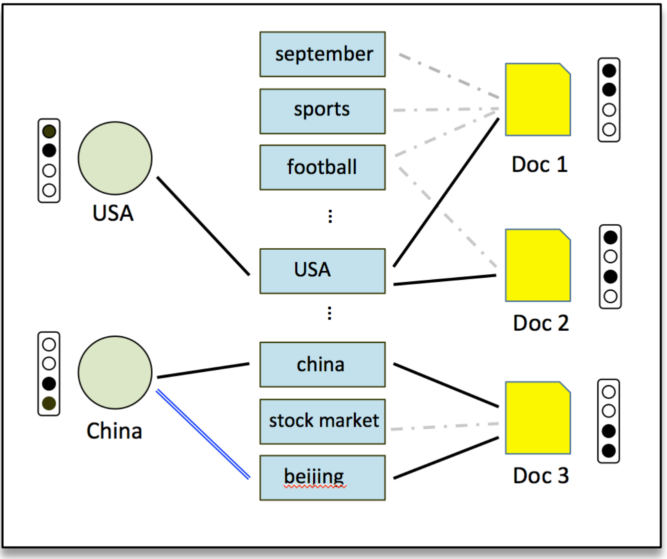
ID	Document Content
1	… The super bowl is on air from Chicago, Illinois. The NFL has decided that best coach of 2017 is from…
2	… make a speech in Shanghai that economy plan is to make sure manufactory industry of China…
3	… in Dec 2015, attacks continued in France for two more days, taking the lives of six others

Doc2Cube: Constructing Cube from Massive Docs: ICDM'18

How to Put Documents into the Right Cube Cell?

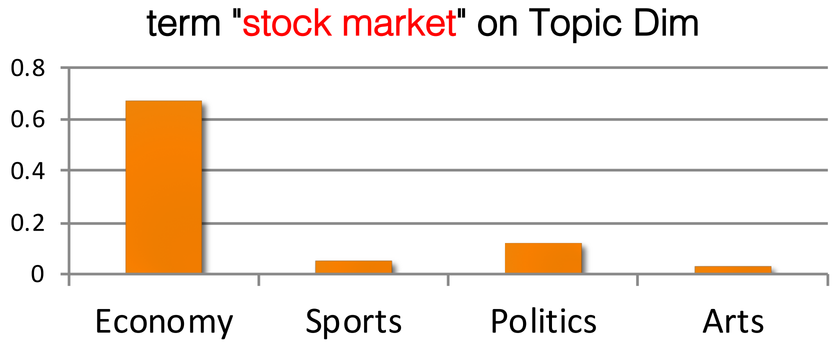
- Major challenges on putting docs into the right cell
  - Few would like label the “training sets”
    - So many cells, so many documents
  - Dimension values are often “under-represented”
    - E.g., Topic dimension: Sports, economy, politics, ….
  - Documents are often “over-represented” on single dimension
    - Ex. “ … … The super bowl is on air from Chicago, Illinois. The NFL has decided that best coach of 2017 is from …
- Our methodology: Dimension-aware joint embedding
  - Constructing an L-T-D (label-term-document) graph

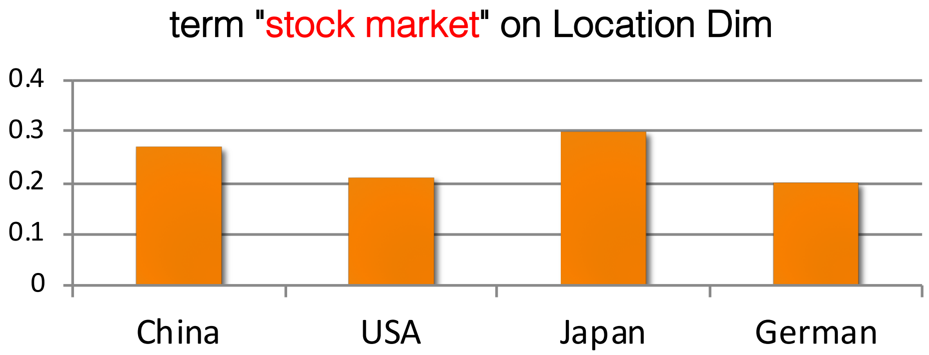




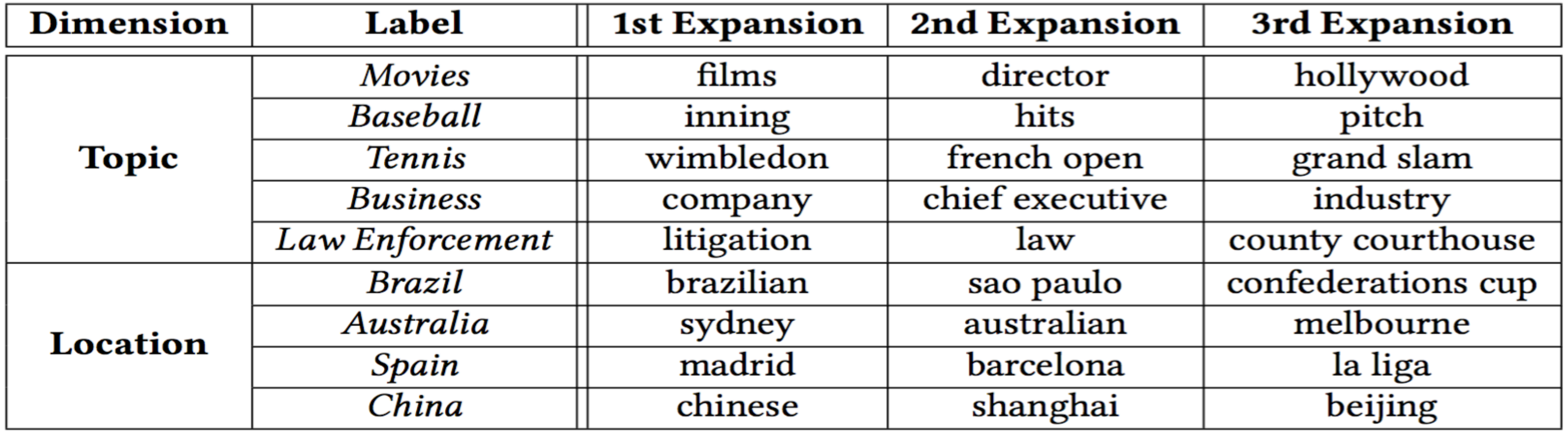
Constructing Text Cubes with Massive Data, Few Labels

- Dimension focusing—**Dimension-Focal Score**, a discriminative measure
  - A term *t* is “focal” to dimension L
    - The documents with t has very imbalanced labels (KL-divergence can be a good measure)
      - Ex.





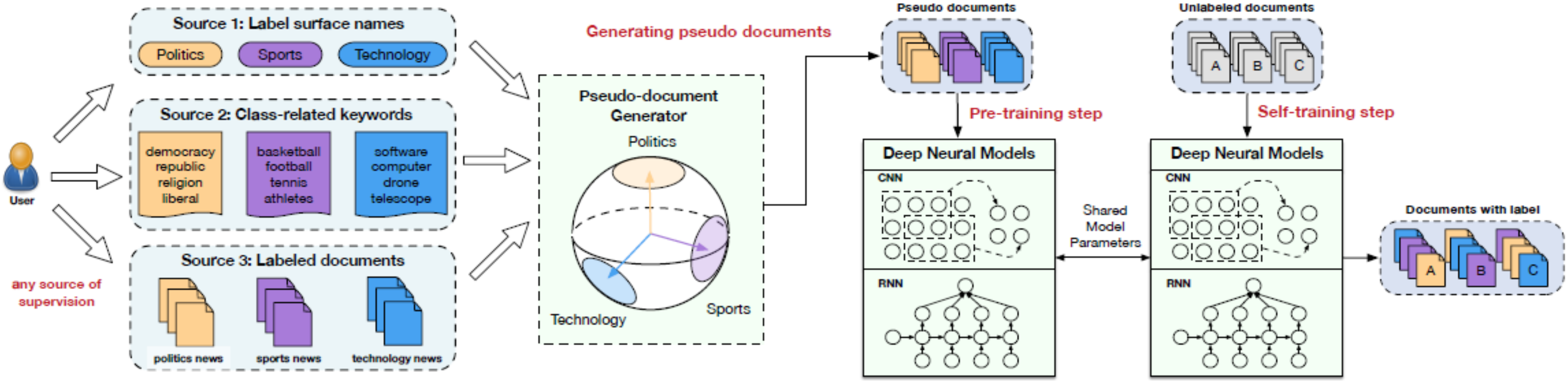
- Label expansion: Combining two measures for seed expansion
  - Discriminativeness
    - Using focal score
  - Popularity
- Example:



Dimension	Label	1st Expansion	2nd Expansion	3rd Expansion
Topic	*Movies*	films	director	hollywood
	Baseball	inning	hits	pitch
	Tennis	wimbledon	french open	grand slam
	Business	company	chief executive	industry
	Law Enforcement	litigation	law	county courthouse
Location	*Brazil*	brazilian	sao paulo	confederations cup
	Australia	sydney	australian	melbourne
	Spain	madrid	barcelona	la liga
	China	chinese	shanghai	beijing

WeSTClass: Weakly Supervised Text Classification

- Modeling class distribution in word2vec embedding space
  - Word2vec embedding captures **skip-gram (local) similarity** (i.e., words with similar local context windows are expected to have similar meanings)

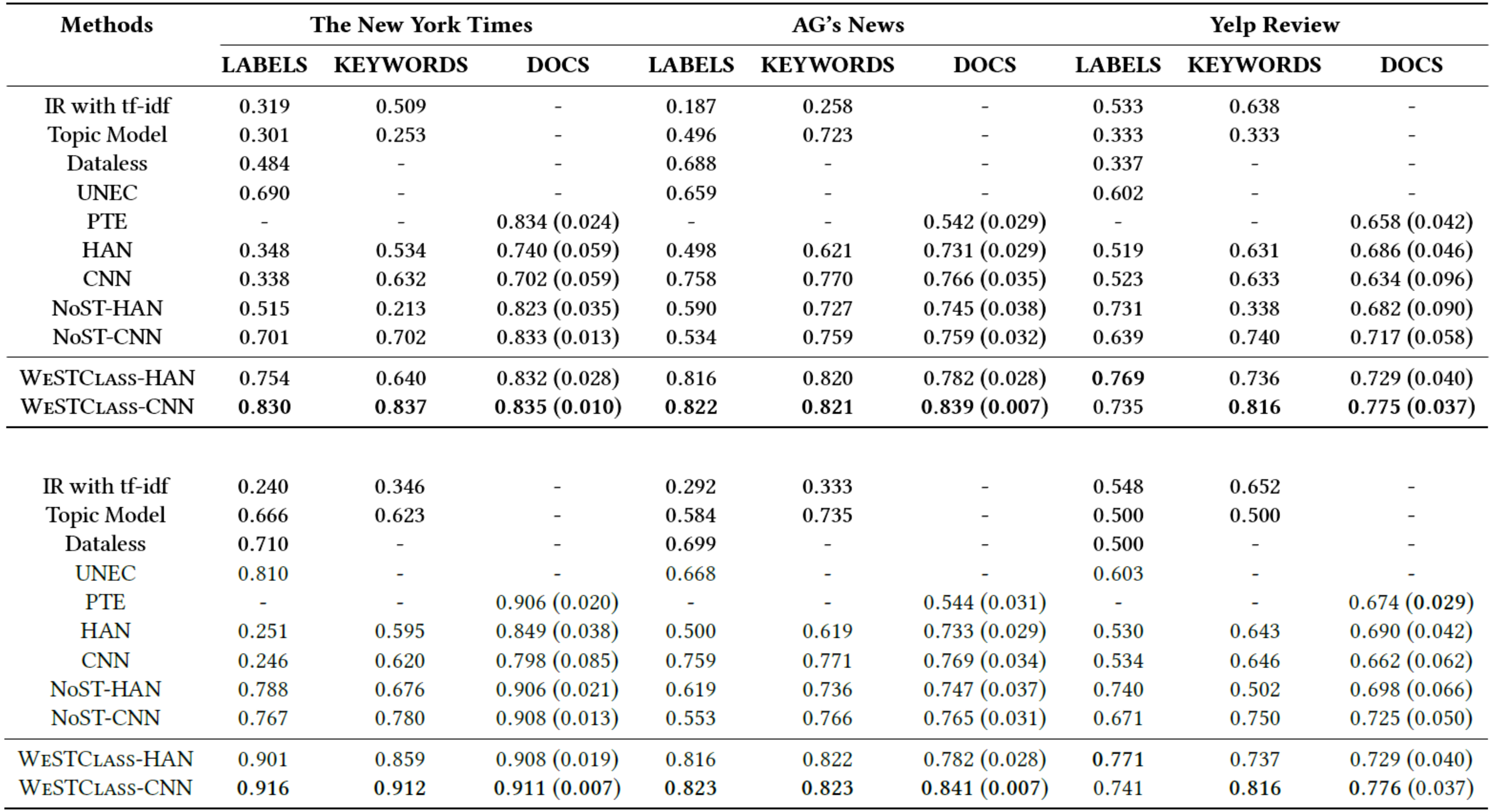


WeSTClass (Weakly Supervised Text Classification): CIKM’18

WeSHClass (Weakly Supervised Hierarchical Text Classification): AAAI’19

WeSTClass: Overall Classification Performance

- Datasets: (1) NYT, (2) AG's News, (3) Yelp
- Evaluation: use different types of weak supervision and measure accuracies



Macro-F1 scores:

Methods	The New York Times			AG's News			Yelp Review		
	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS
IR with tf-idf	0.319	0.509	-	0.187	0.258	-	0.533	0.638	-
Topic Model	0.301	0.253	-	0.496	0.723	-	0.333	0.333	-
Dataless	0.484	-	-	0.688	-	-	0.337	-	-
UNEC	0.690	-	-	0.659	-	-	0.602	-	-
PTE	-	-	0.834 (0.024)	-	-	0.542 (0.029)	-	-	0.658 (0.042)
HAN	0.348	0.534	0.740 (0.059)	0.498	0.621	0.731 (0.029)	0.519	0.631	0.686 (0.046)
CNN	0.338	0.632	0.702 (0.059)	0.758	0.770	0.766 (0.035)	0.523	0.633	0.634 (0.096)
NoST-HAN	0.515	0.213	0.823 (0.035)	0.590	0.727	0.745 (0.038)	0.731	0.338	0.682 (0.090)
NoST-CNN	0.701	0.702	0.833 (0.013)	0.534	0.759	0.759 (0.032)	0.639	0.740	0.717 (0.058)
WeSTClass-HAN	0.754	0.640	0.832 (0.028)	0.816	0.820	0.782 (0.028)	**0.769**	0.736	0.729 (0.040)
WeSTClass-CNN	**0.830**	**0.837**	**0.835 (0.010)**	**0.822**	**0.821**	**0.839 (0.007)**	0.735	**0.816**	**0.775 (0.037)**

Micro-F1 scores:

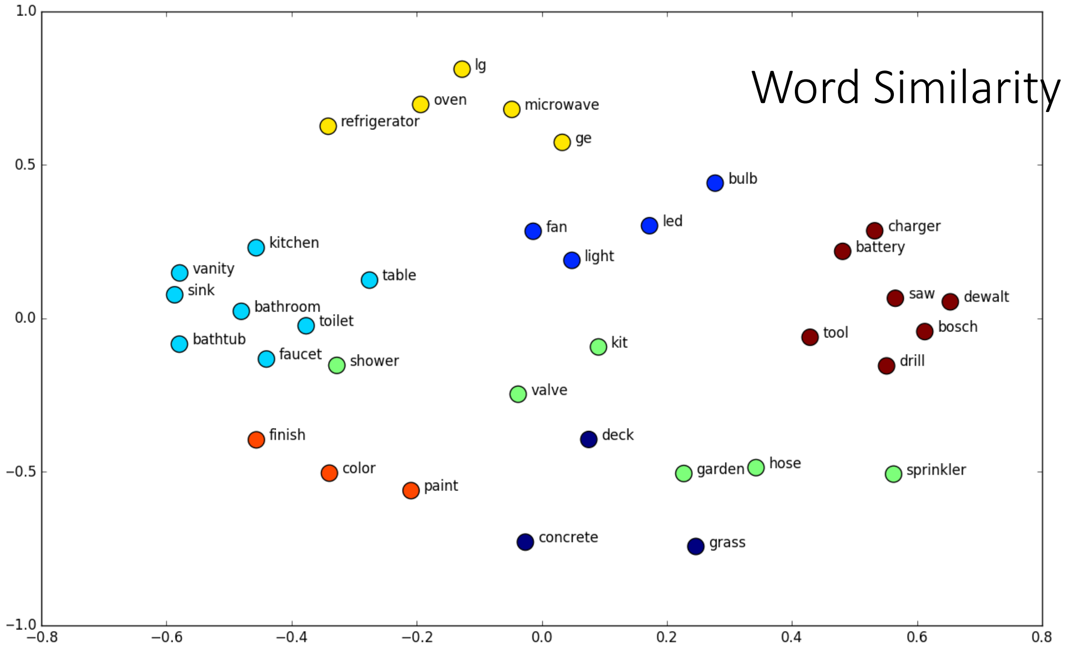
Methods	The New York Times			AG's News			Yelp Review		
	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS
IR with tf-idf	0.240	0.346	-	0.292	0.333	-	0.548	0.652	-
Topic Model	0.666	0.623	-	0.584	0.735	-	0.500	0.500	-
Dataless	0.710	-	-	0.699	-	-	0.500	-	-
UNEC	0.810	-	-	0.668	-	-	0.603	-	-
PTE	-	-	0.906 (0.020)	-	-	0.544 (0.031)	-	-	0.674 (0.029)
HAN	0.251	0.595	0.849 (0.038)	0.500	0.619	0.733 (0.029)	0.530	0.643	0.690 (0.042)
CNN	0.246	0.620	0.798 (0.085)	0.759	0.771	0.769 (0.034)	0.534	0.646	0.662 (0.062)
NoST-HAN	0.788	0.676	0.906 (0.021)	0.619	0.736	0.747 (0.037)	0.740	0.502	0.698 (0.066)
NoST-CNN	0.767	0.780	0.908 (0.013)	0.553	0.766	0.765 (0.031)	0.671	0.750	0.725 (0.050)
WeSTClass-HAN	0.901	0.859	0.908 (0.019)	0.816	0.822	0.782 (0.028)	**0.771**	0.737	0.729 (0.040)
WeSTClass-CNN	**0.916**	**0.912**	**0.911 (0.007)**	**0.823**	**0.823**	**0.841 (0.007)**	0.741	**0.816**	**0.776 (0.037)**

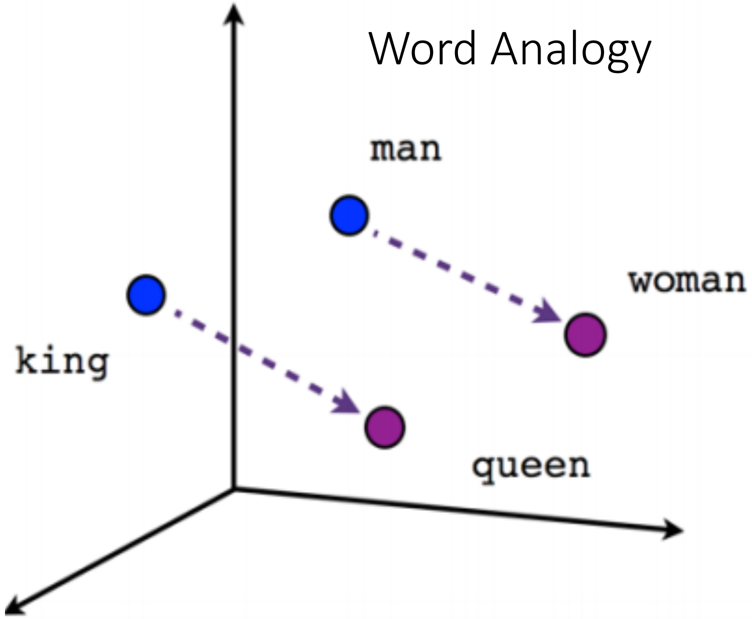
Outline

- On the Power of Multi-Dimensional Text Cubes
- Automated Mining of Semantic Structures from Massive Text Data
  - Phrase Mining
  - Entity/Relation Recognition and Typing
  - Meta Pattern-Directed Structure Discovery
- Automated Construction of Multidimensional Text Cubes
  - Multifaceted Taxonomy Mining
  - Doc2Cube: Constructing TextCube from Massive Documents
  - Quality Enhancement: Local and Global Joint Spherical Text Embedding
- Looking Forward

Text Embedding: Preliminaries

- A milestone in NLP and ML: Unsupervised learning of text representations
- Embed one-hot vectors into lower-dimens. space—Address “curse of dimensionality”
- Word embedding captures useful properties of word semantics
  - Word similarity: Words with similar meanings are embedded closer
  - Word analogy: Linear relationships between words (e.g., king – queen = man–woman)





Typical embedding methods:

Word2Vec

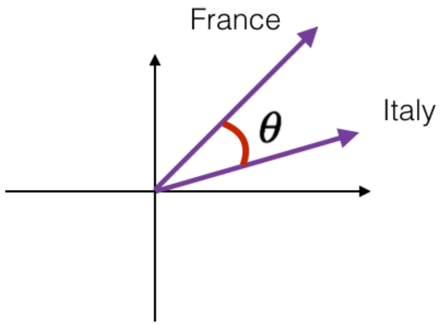
GloVe

fastText

Trained in Euclidean space

Why Spherical Text Embedding? [NeurIPS'19]

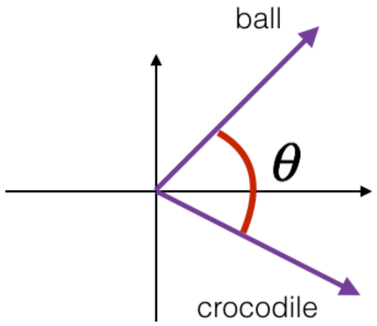
- Previous text embeddings (e.g., Word2Vec) are trained in the Euclidean space
  - But used on spherical space—Mostly directional similarity (i.e., cosine similarity)
  - Word similarity is derived using cosine similarity



France and Italy are quite similar

θ is close to 0°

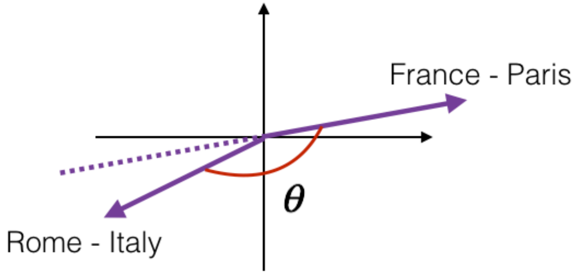
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

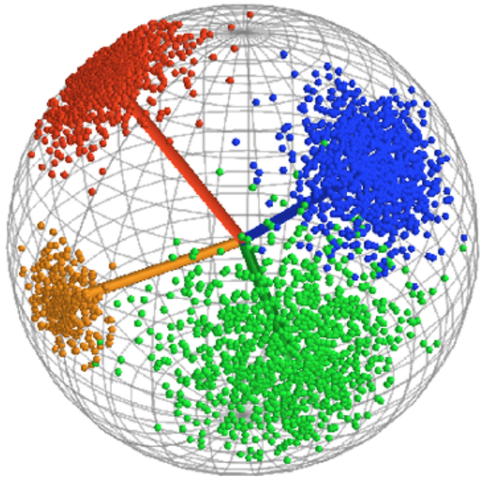
$\cos(\theta) \approx 0$



the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

θ is close to 180°

$\cos(\theta) \approx -1$

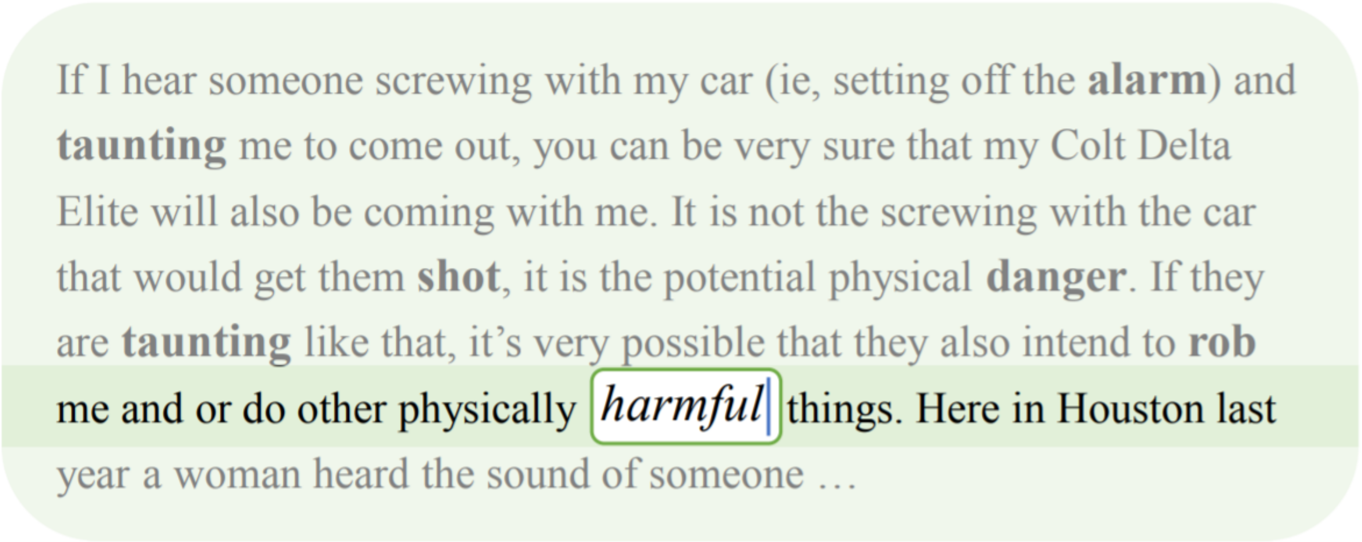


  - Word clustering (e.g., TaxoGen) is performed on a sphere
  - Better document clustering performances when embeddings are normalized and spherical clustering algorithms are used

Why Integrating Local and Global Contexts?

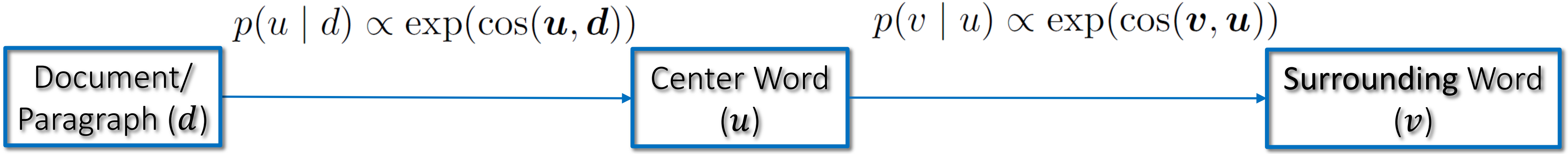
- Local contexts can only partly define word semantics in unsupervised word embedding learning

Local contexts of "harmful"



If I hear someone screwing with my car (ie, setting off the **alarm**) and **taunting** me to come out, you can be very sure that my Colt Delta Elite will also be coming with me. It is not the screwing with the car that would get them **shot**, it is the potential physical **danger**. If they are **taunting** like that, it's very possible that they also intend to **rob** me and or do other physically *harmful* things. Here in Houston last year a woman heard the sound of someone …

- Design a generative model on the sphere that follows how humans write articles:
  - First a general idea of the paragraph/doc, then start to write down each word in consistent with not only the paragraph/doc, but also the surrounding words



$p(u \mid d) \propto \exp(\cos(\boldsymbol{u}, \boldsymbol{d}))$

$p(v \mid u) \propto \exp(\cos(\boldsymbol{v}, \boldsymbol{u}))$

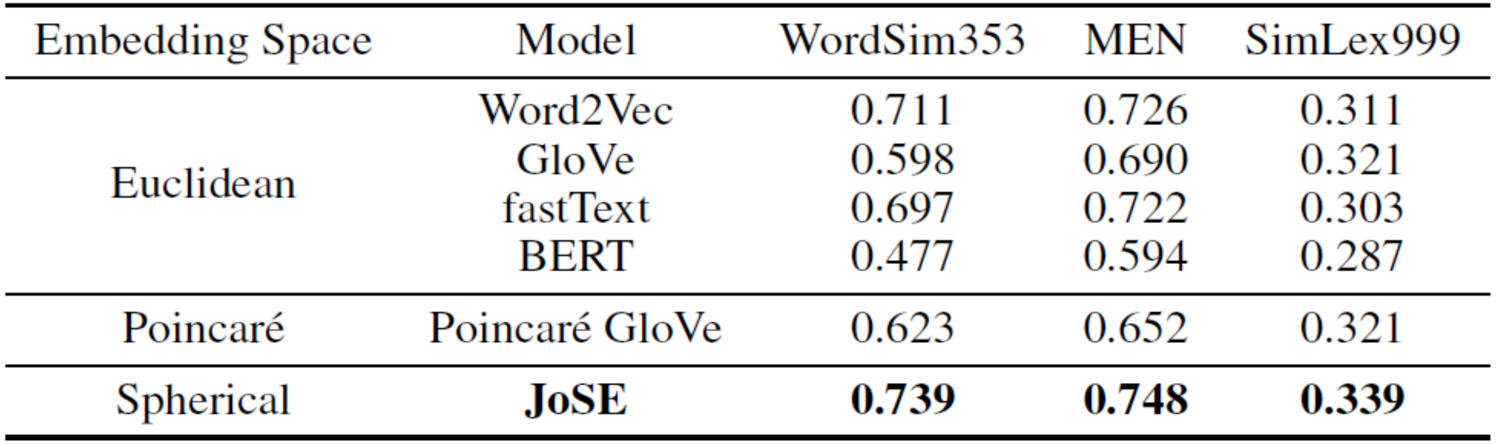
Document/ Paragraph ($\boldsymbol{d}$)
Center Word ($\boldsymbol{u}$)
Surrounding Word ($\boldsymbol{v}$)

JoSE: Performance Comparison with Recent Methods

JoSE: Joint Spherical Text Embedding [NeurIPS'19]

- Word similarity results:

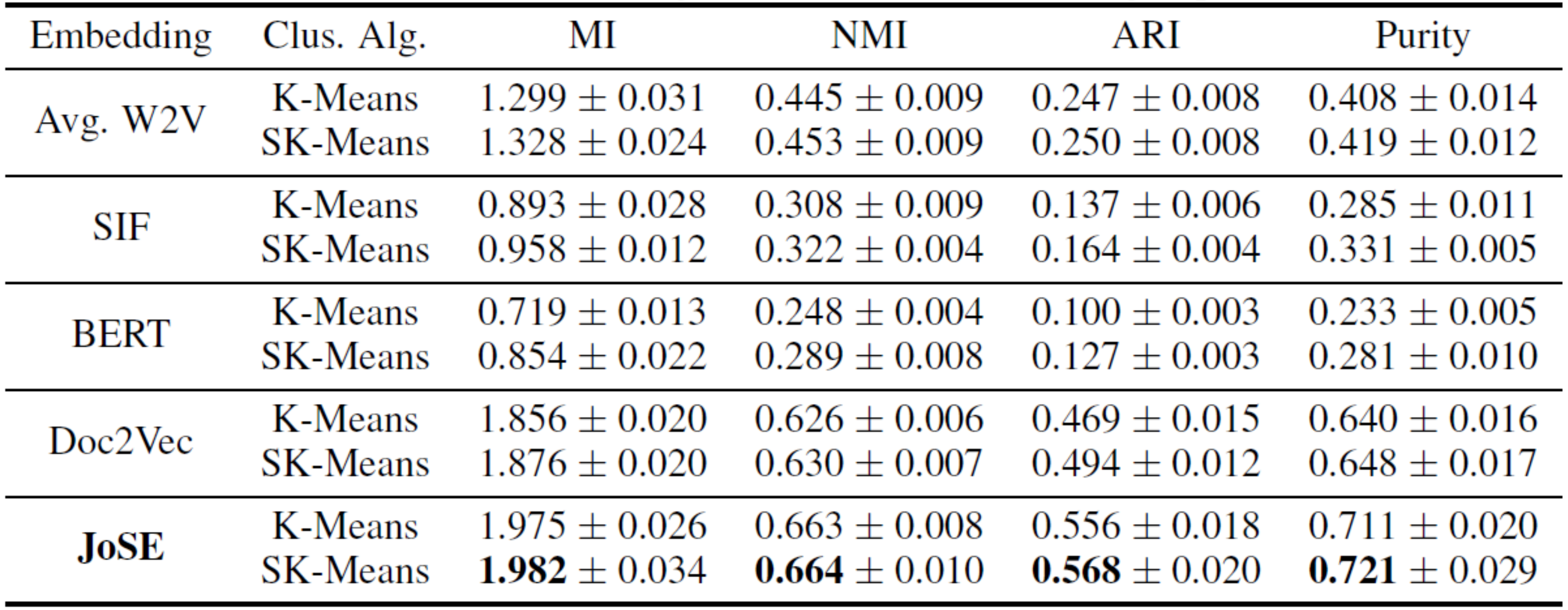
Table 1: Spearman rank correlation on word similarity evaluation.



Embedding Space	Model	WordSim353	MEN	SimLex999
Euclidean	Word2Vec	0.711	0.726	0.311
	GloVe	0.598	0.690	0.321
	fastText	0.697	0.722	0.303
	BERT	0.477	0.594	0.287
Poincaré	Poincaré GloVe	0.623	0.652	0.321
Spherical	**JoSE**	**0.739**	**0.748**	**0.339**

- Document clustering results:

Table 2: Document clustering evaluation on the 20 Newsgroup dataset.



Embedding	Clus. Alg.	MI	NMI	ARI	Purity
Avg. W2V	K-Means	1.299 ± 0.031	0.445 ± 0.009	0.247 ± 0.008	0.408 ± 0.014
	SK-Means	1.328 ± 0.024	0.453 ± 0.009	0.250 ± 0.008	0.419 ± 0.012
SIF	K-Means	0.893 ± 0.028	0.308 ± 0.009	0.137 ± 0.006	0.285 ± 0.011
	SK-Means	0.958 ± 0.012	0.322 ± 0.004	0.164 ± 0.004	0.331 ± 0.005
BERT	K-Means	0.719 ± 0.013	0.248 ± 0.004	0.100 ± 0.003	0.233 ± 0.005
	SK-Means	0.854 ± 0.022	0.289 ± 0.008	0.127 ± 0.003	0.281 ± 0.010
Doc2Vec	K-Means	1.856 ± 0.020	0.626 ± 0.006	0.469 ± 0.015	0.640 ± 0.016
	SK-Means	1.876 ± 0.020	0.630 ± 0.007	0.494 ± 0.012	0.648 ± 0.017
JoSE	K-Means	1.975 ± 0.026	0.663 ± 0.008	0.556 ± 0.018	0.711 ± 0.020
	SK-Means	**1.982** ± 0.034	**0.664** ± 0.010	**0.568** ± 0.020	**0.721** ± 0.029

JoSE: Performance & Case Studies

- Document classification results

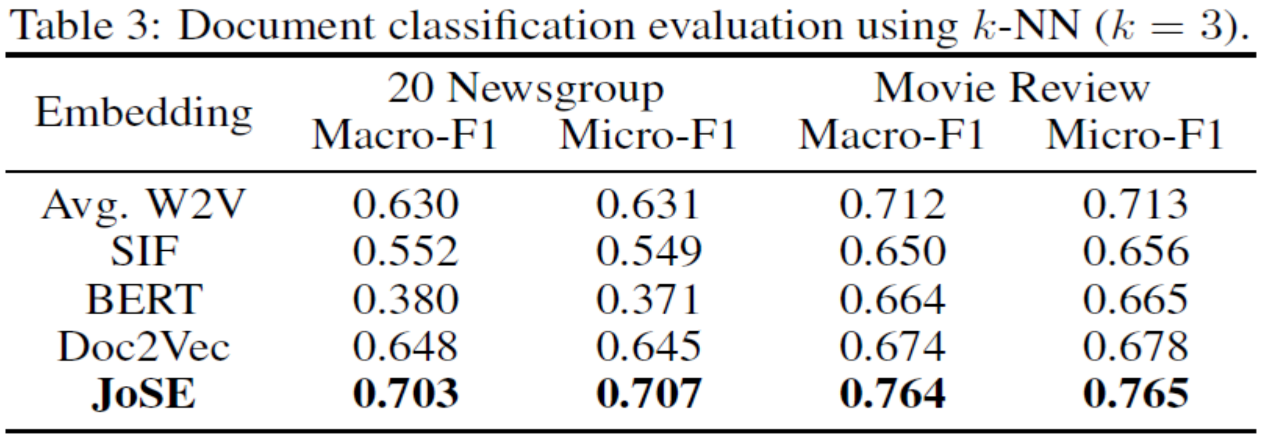


Table 3: Document classification evaluation using k-NN ($k = 3$).

Embedding	20 Newsgroup		Movie Review	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Avg. W2V	0.630	0.631	0.712	0.713
SIF	0.552	0.549	0.650	0.656
BERT	0.380	0.371	0.664	0.665
Doc2Vec	0.648	0.645	0.674	0.678
JoSE	**0.703**	**0.707**	**0.764**	**0.765**

- Training efficiency

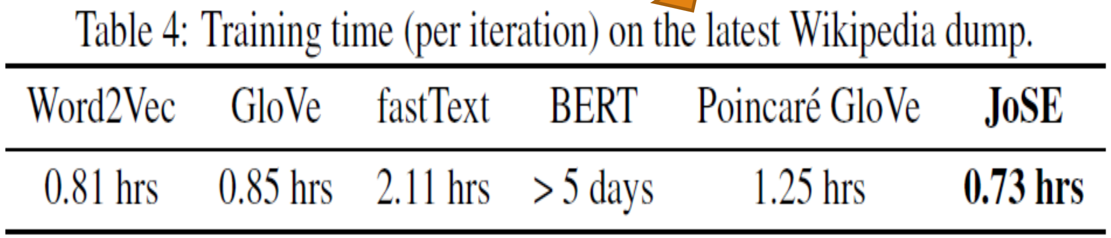


Table 4: Training time (per iteration) on the latest Wikipedia dump.

Word2Vec	GloVe	fastText	BERT	Poincaré GloVe	**JoSE**
0.81 hrs	0.85 hrs	2.11 hrs	> 5 days	1.25 hrs	**0.73 hrs**

- Acronym → similar words

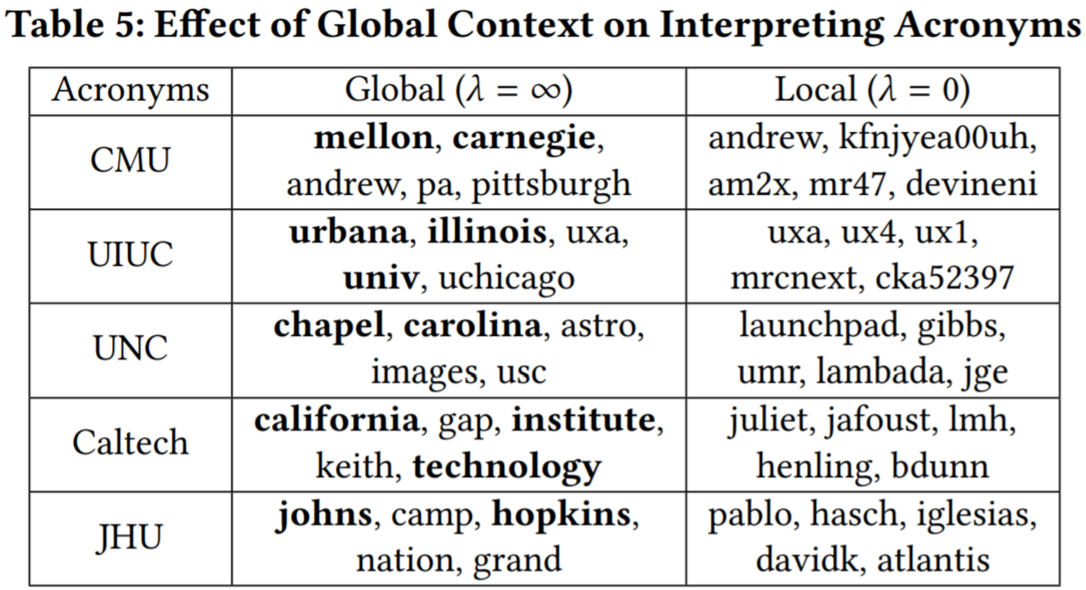


Table 5: Effect of Global Context on Interpreting Acronyms

Acronyms	Global ($\lambda = \infty$)	Local ($\lambda = 0$)
CMU	**mellon**, **carnegie**, andrew, pa, pittsburgh	andrew, kfnjyea00uh, am2x, mr47, devineni
UIUC	**urbana**, **illinois**, uxa, **univ**, uchicago	uxa, ux4, ux1, mrcnext, cka52397
UNC	**chapel**, **carolina**, astro, images, usc	launchpad, gibbs, umr, lambada, jge
Caltech	**california**, gap, **institute**, keith, **technology**	juliet, jafoust, lmh, henling, bdunn
JHU	**johns**, camp, **hopkins**, nation, grand	pablo, hasch, iglesias, davidk, atlantis

- Testing antonym similarity

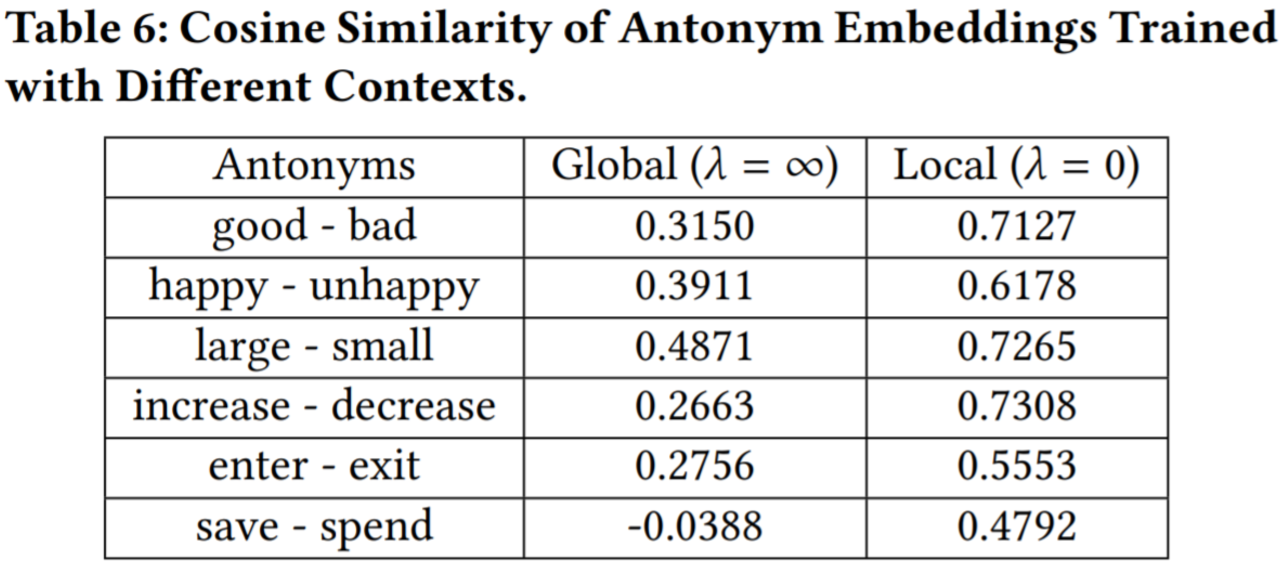


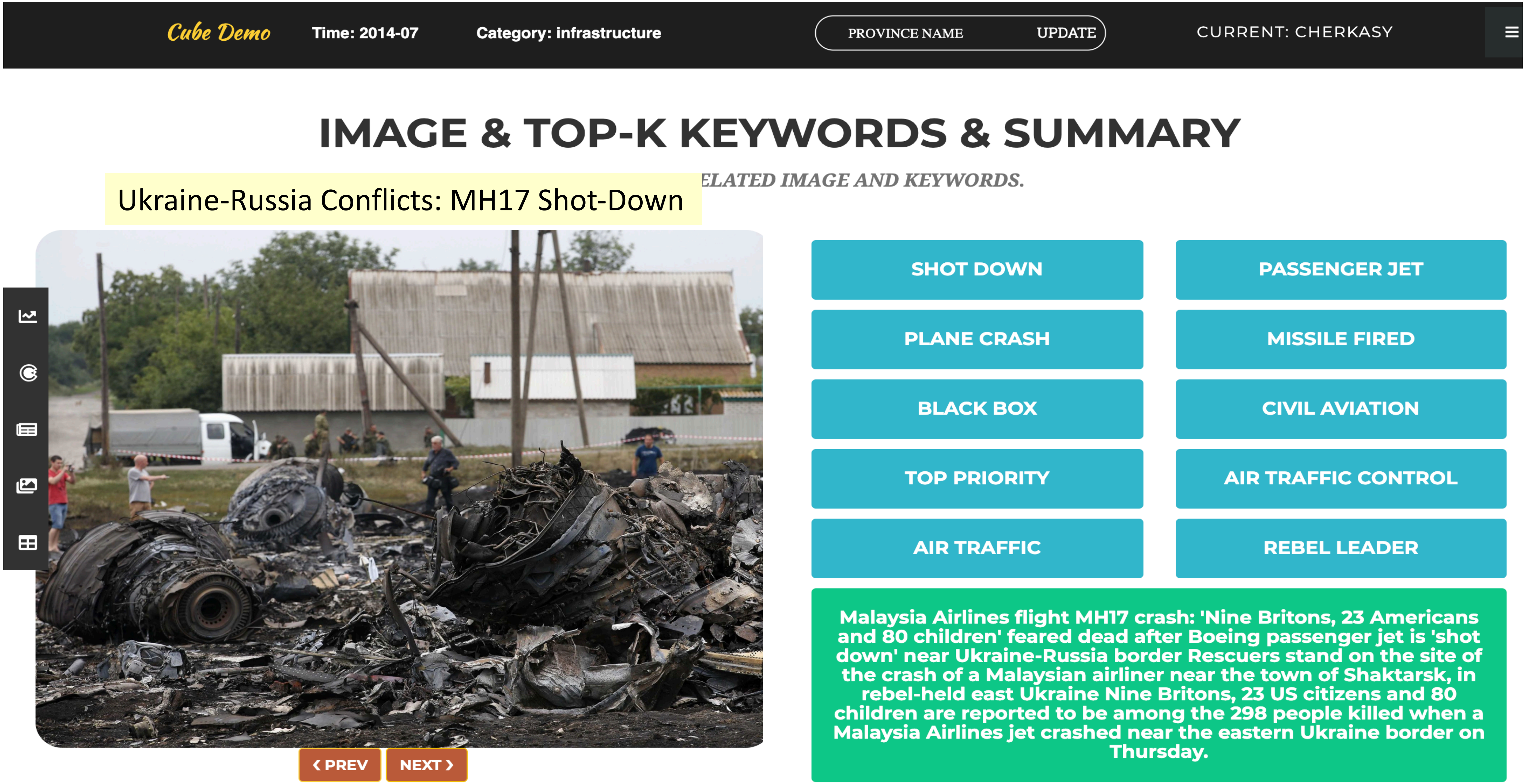
Table 6: Cosine Similarity of Antonym Embeddings Trained with Different Contexts.

Antonyms	Global ($\lambda = \infty$)	Local ($\lambda = 0$)
good - bad	0.3150	0.7127
happy - unhappy	0.3911	0.6178
large - small	0.4871	0.7265
increase - decrease	0.2663	0.7308
enter - exit	0.2756	0.5553
save - spend	-0.0388	0.4792

Outline

- ❑ On the Power of Multi-Dimensional Text Cubes
- ❑ Automated Mining of Semantic Structures from Massive Text Data
  - ❑ Phrase Mining
  - ❑ Entity/Relation Recognition and Typing
  - ❑ Meta Pattern-Directed Structure Discovery
- ❑ Automated Construction of Multidimensional Text Cubes
  - ❑ Multifaceted Taxonomy Mining
  - ❑ Doc2Cube: Constructing TextCube from Massive Documents
  - ❑ Quality Enhancement: Local and Global Joint Spherical Text Embedding
- ❑ Looking Forward

Application: Support Multi-Dimensional Text Analysis



Cube Demo Time: 2014-07 Category: infrastructure

PROVINCE NAME UPDATE CURRENT: CHERKASY

IMAGE & TOP-K KEYWORDS & SUMMARY

ELATED IMAGE AND KEYWORDS.

Ukraine-Russia Conflicts: MH17 Shot-Down

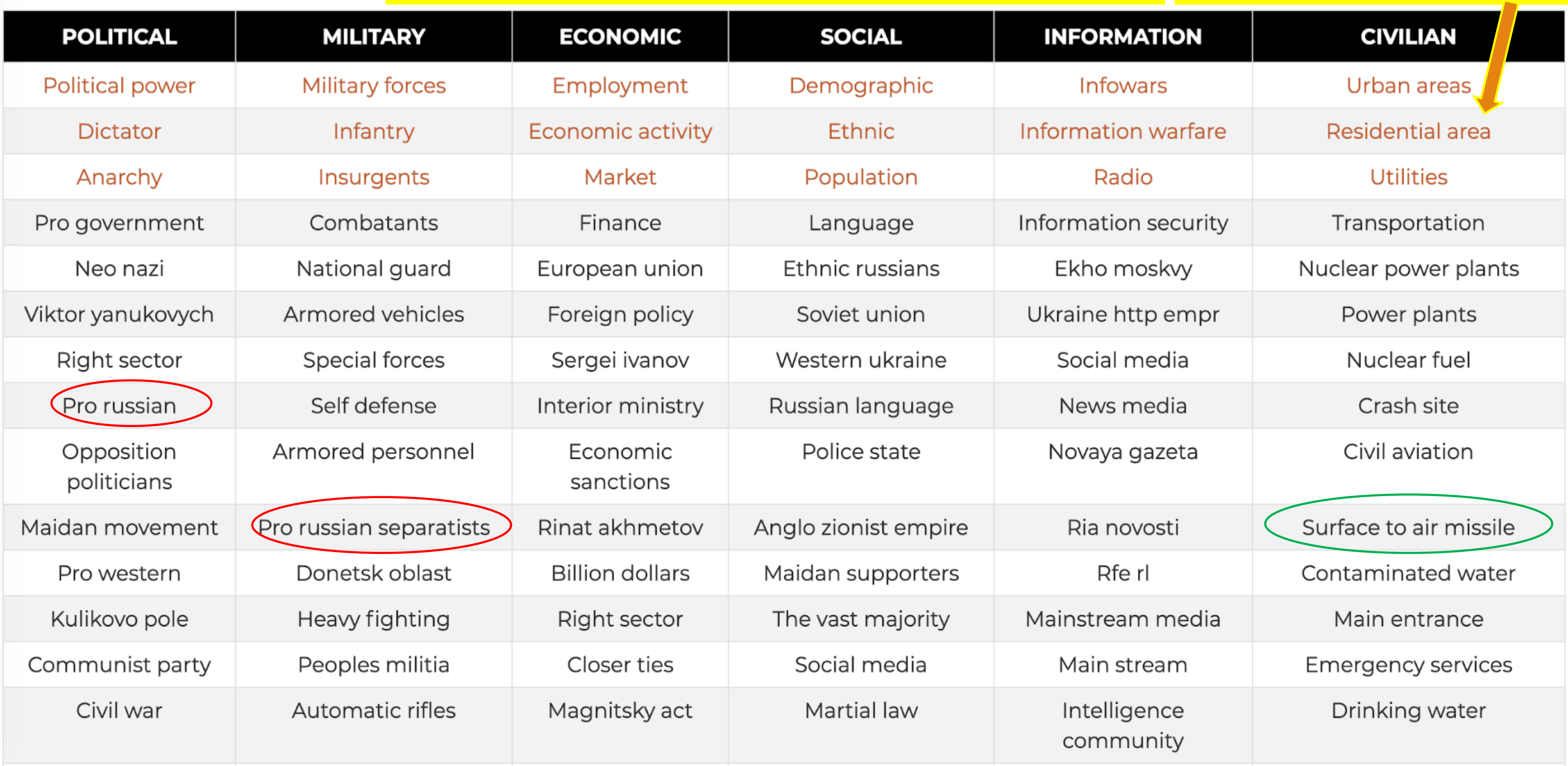
SHOT DOWN | PASSENGER JET
PLANE CRASH | MISSILE FIRED
BLACK BOX | CIVIL AVIATION
TOP PRIORITY | AIR TRAFFIC CONTROL
AIR TRAFFIC | REBEL LEADER

Malaysia Airlines flight MH17 crash: 'Nine Britons, 23 Americans and 80 children' feared dead after Boeing passenger jet is 'shot down' near Ukraine-Russia border Rescuers stand on the site of the crash of a Malaysian airliner near the town of Shaktarsk, in rebel-held east Ukraine Nine Britons, 23 US citizens and 80 children are reported to be among the 298 people killed when a Malaysia Airlines jet crashed near the eastern Ukraine border on Thursday.

Analysis of Russia-Ukraine Conflicts

Category representative phrases generated automatically

category names and three examples from the experts



POLITICAL	MILITARY	ECONOMIC	SOCIAL	INFORMATION	CIVILIAN
Political power	Military forces	Employment	Demographic	Infowars	Urban areas
Dictator	Infantry	Economic activity	Ethnic	Information warfare	Residential area
Anarchy	Insurgents	Market	Population	Radio	Utilities
Pro government	Combatants	Finance	Language	Information security	Transportation
Neo nazi	National guard	European union	Ethnic russians	Ekho moskvy	Nuclear power plants
Viktor yanukovych	Armored vehicles	Foreign policy	Soviet union	Ukraine http empr	Power plants
Right sector	Special forces	Sergei ivanov	Western ukraine	Social media	Nuclear fuel
Pro russian	Self defense	Interior ministry	Russian language	News media	Crash site
Opposition politicians	Armored personnel	Economic sanctions	Police state	Novaya gazeta	Civil aviation
Maidan movement	Pro russian separatists	Rinat akhmetov	Anglo zionist empire	Ria novosti	Surface to air missile
Pro western	Donetsk oblast	Billion dollars	Maidan supporters	Rfe rl	Contaminated water
Kulikovo pole	Heavy fighting	Right sector	The vast majority	Mainstream media	Main entrance
Communist party	Peoples militia	Closer ties	Social media	Main stream	Emergency services
Civil war	Automatic rifles	Magnitsky act	Martial law	Intelligence community	Drinking water

Cube Demo Dataset: Hong Kong Protest Time: 2019.08.12-08.18 Category: transportation City: Islands ☰

IMAGE & TOP-K KEYWORDS & SUMMARY

IT SHOWS THE RELATED IMAGE AND KEYWORDS.

ALLEGEDLY SHOT	EYE PATCHES
TEAR GAS INSIDE	PATCHES
AIRPORTS	AIRPORT SECURITY
CHASING PROTESTERS	CHARGED PROTESTERS
BEANBAG ROUND	NEWS FOOTAGE

Demonstrators don eye patches at Lantau Island hub, one of the world's busiest international airports, in anger that a girl allegedly shot with a police beanbag round could lose an eye \n Sit-in comes after night of escalated violence inside subway stations \n Demonstrators don eye patches at Lantau Island hub, one of the world's busiest international airports, in anger that a girl allegedly shot with a police beanbag round could lose an eye.

MissionCube: Analysis of Different News Data Sets: HK Protests

Analysis of Hong Kong Protests

Category representative phrases generated automatically

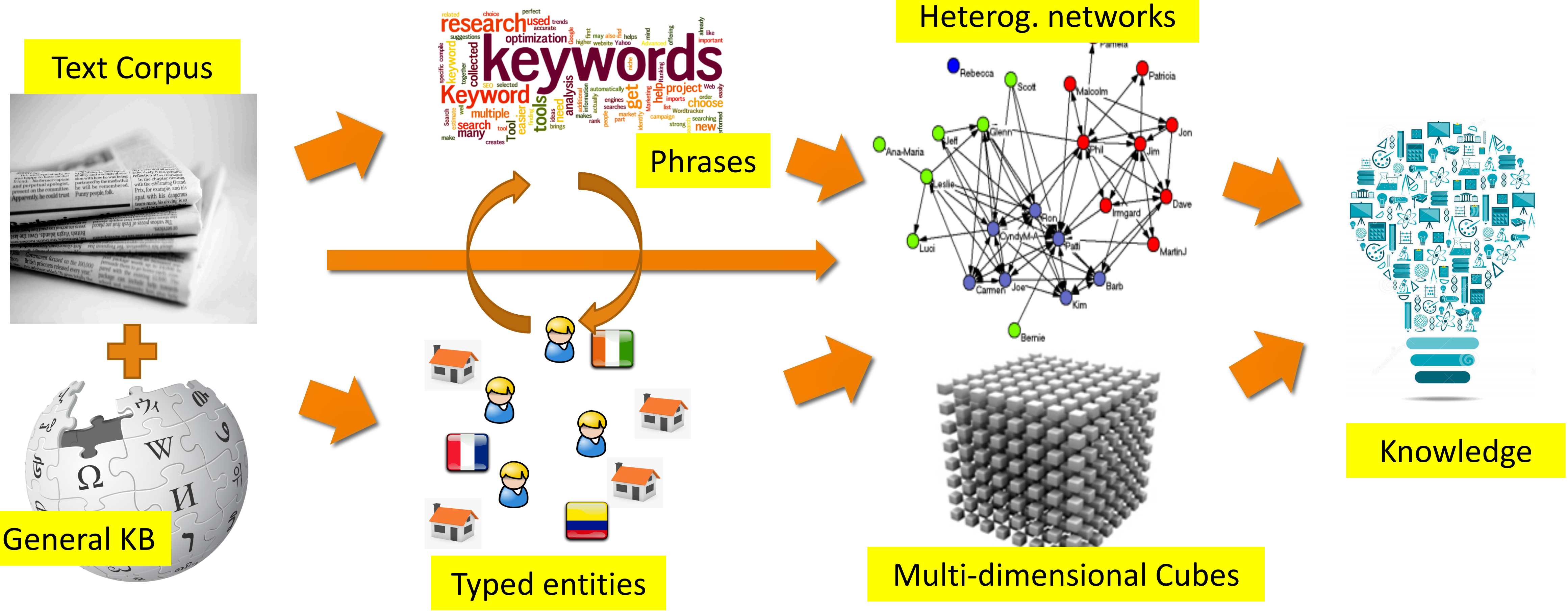
category names and three examples from the experts

IT SHOWS RELEVANT WORDS OF DIFFERENT CATEGORIES;

POLITICAL	POLICE	ECONOMIC	INFORMATION	INFRASTRUCTURE
pro democracy	tear gas	financial crisis	cbc news	hong kong university
pro beijing	hong kong police	economic downturn	cbs news	transportation
hong kong government	riot police	economic growth	fox news	international aiport
Chief executive	Water cannon	Infrastructure	Chinese state media	Mass transit railway
Mainland china	Pepper spray	Real estate	Bbc news	Lantau link
Pro establishment	Petrol bombs	Affordable housing	Global times	Flight cancellations
Mainland chinese	Hong kong government	Trade war	News media	Victoria harbour
Chief executive carrie lam	Beanbag rounds	The united states	Sina weibo	Rail operator
Carrie lam	Firing tear gas	Financial secretary	Internet censorship	Busiest airports
The chinese government	Tsuen wan	Global financial	Local media	Public transport

Looking Forward: Structural Mining of Massive Text Data

- From big data to big knowledge
  - A key problem: **Structural mining of massive text data**
  - Lots to be explored!!!



Text Corpus

General KB

Phrases

Typed entities

Heterog. networks

Multi-dimensional Cubes

Knowledge

Our Journey: From ctures & Knowledge

MORGAN & CLAYPOOL PUBLISHERS

Multidimensional Mining of Massive Text Data

Chao Zhang
Jiawei Han

C. Zhang: SIGKDD'19 Dissertation Award Runner-Up

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

MORGAN & CLAYPOOL PUBLISHERS

Mining Structures of Factual Knowledge from Text

An Effort-Light Approach

Xiang Ren
Jiawei Han

Xiang Ren: SIGKDD'18 Dissertation A

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

MORGAN & CLAYPOOL PUBLISHERS

Phrase Mining from Massive Text and Its Applications

ialu Liu
ngbo Shang
wei Han

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

HESIS LECTURES ON
ATA MINING AND KNOWLEDGE D

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, Serie

C. CLAYPOOL PUBLISHERS

ISCOVERY

Grossman, *Series Editors*

Han, Kamber
Data Mining, 3rd

and Faloutsos (ed
Link Mining, 2010

Sun and Han, Mining Heterogeneous Information Networks, 2012
Y. Sun: SIGKDD'13 Dissertation Award

Latent Entity
015
C. Wang: SIG sertation Award

Acknowledgements

- Thanks for the research support from: ARL/NSCTA, NIH, NSF, DARPA, DTRA, ……

References

- M. Jiang, J. Shang, X. Ren, T. Cassidy, L. Kaplan, T. Hanratty and J. Han, “MetaPAD: Meta Pattern-driven Attribute Discovery from Massive Text Corpora”, **KDD’17**
- Q. Li, M. Jiang, X. Zhang, M. Qu, T. Hanratty, J. Gao and J. Han, “TruePIE: Discovering Reliable Patterns in Pattern-Based Information Extraction”, **KDD’18**
- J. Liu, J. Shang, J. Han, **Phrase Mining from Massive Text and Its Applications**, M. & Claypool, **2017**
- Y. Meng, J. Huang, G. Wang, C. Zhang, H. Zhuang, L. Kaplan and J. Han, "Spherical Text Embedding", **NeurIPS’19**
- Y. Meng, J. Shen, C. Zhang and J. Han, "Weakly-Supervised Neural Text Classification", **CIKM’18**
- 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 and J. Han, **Mining Structures of Factual Knowledge from Text: An Effort-Light Approach**, Morgan & Claypool, **2018**
- F. Tao, C. Zhang, X. Chen, M. Jiang, T. Hanratty, L. Kaplan, J. Han, “Doc2Cube: Automated Document Allocation to Text Cube via Dimension-Aware Joint Embedding”, **ICDM’18**
- C. Zhang, F. Tao, X. Chen, J. Shen, M. Jiang, B. Sadler, M. Vanni and J. Han, “TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering”, **KDD’18**