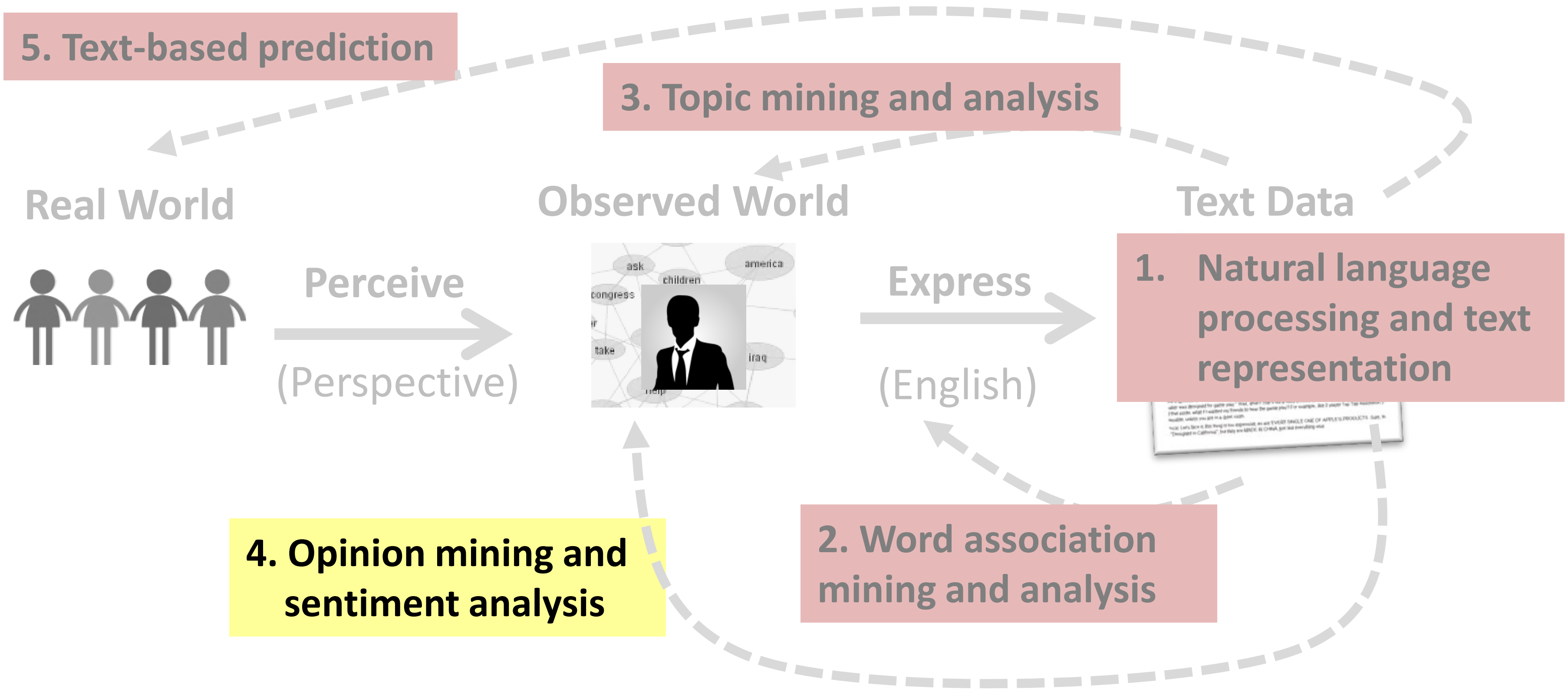
Opinion Mining and Sentiment Analysis: Latent Aspect Rating Analysis

Slide adapted from ChengXiang Zhai's presentation

Opinion Mining and Sentiment Analysis: Latent Aspect Rating Analysis



Slide adapted from ChengXiang Zhai's presentation

Motivation

How to infer aspect ratings?

<u>Hotel XYX</u>

Reviewer 1: ★★★★★

"Great location + spacious room = happy traveler"

Stayed for a weekend in July. Walked everywhere, enjoyed the comfy bed and quiet hallways....

★★★★	**Value**
★★★★★	**Rooms**
★★★★★	**Location**
★★★★	**Service**

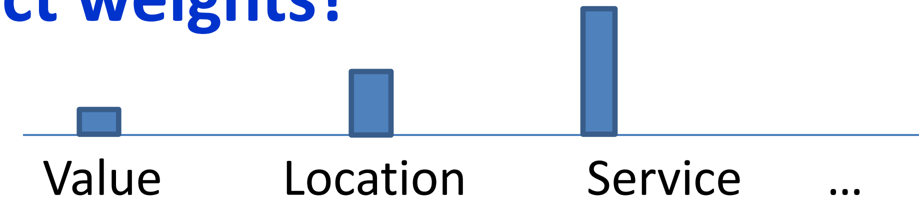
Reviewer 2: ★★★★★

"Terrific service and gorgeous facility"

I stayed at the hotel wiht my young daughter for three nights June 17-20, 2010 and absolutely loved the hotel. The room was one of the nicest I've ever stayed in ...

★★★★	**Value**
★★★★★	**Rooms**
★★★★★	**Location**
★★★★	**Service**

How to infer aspect weights?



Slide adapted from ChengXiang Zhai's presentation

Latent Aspect Rating Analysis [Wang et al. 10]

- Given a set of review articles about a topic with overall ratings
- Output
  – Major aspects commented on in the reviews
  – Ratings on each aspect
  – Relative weights placed on different aspects by reviewers
- Many applications
  – Opinion-based entity ranking
  – Aspect-level opinion summarization
  – Reviewer preference analysis
  – Personalized recommendation of products
  – …

Slide adapted from ChengXiang Zhai's presentation

Solving LARA in Two Stages

Aspect Segmentation + *Latent Rating Regression*

r_d

★★★★★

"A friend and I stayed at the Hotel … The hotel was very nice. The **location** was **amazing**. We could **walk** almost anywhere, but … **far**. The **room** was very **nicely appointed** and the **bed** was sooo **comfortable**. Even though the **bathroom door** did not close all the way, it was still pretty private. … But what I **liked best** about the hotel was the **staff**. They were soooo **nice** and **accommodating** …"

Observed

$c_i(w,d)$ Aspect segments		$\beta_{i,w}$ Term Weights	$r_i(d)$ Aspect Rating		$\alpha_i(d)$ Aspect Weight
location:1		0.0			
amazing:1	⊗	3.9			
walk:1		0.1	3.8	✖	0.2
far:1		-0.2			
room:1		0.1			
nicely:1	⊗	1.7			
appointed:1		0.1	4.8	✖	0.2
comfortable:1		3.9			
nice:1		2.1			
accommod.:1	⊗	1.2	5.8	✖	0.6
smile:1		1.7			
friendliness:1		1.2			
attentiveness:1		0.6			

Latent!

Slide adapted from ChengXiang Zhai's presentation

Latent Rating Regression [Wang et al. 10]

- Data: a set of review documents with overall ratings: $C=\{(d, r_d)\}$
  - d is pre-segmented into k aspect segments
  - $c_i(w, d)$ = count of word w in aspect segment i (zero if w didn't occur)
- Model: predict rating based on d: $p(r_d \mid d)$

Multivariate Gaussian Prior

Overall Rating = Weighted Average of Aspect Ratings

$$r_d \sim N\left(\sum_{i=1}^{k} \alpha_i(d) r_i(d), \underline{\delta^2}\right), \qquad \vec{\alpha}(d) \sim N(\underline{\vec{\mu}, \Sigma})$$

$$r_i(d) = \sum_{w \in V} c_i(w, d) \underline{\beta_{i,w}} \qquad \beta_{i,w} \in \Re$$

Aspect-Specific Sentiment of w

Aspect Rating = Sum of sentiment weights of words in the aspect

Latent Rating Regression (cont.)

- Maximum Likelihood Estimate
  - Parameters: $\Lambda = (\{\beta_{i,w}\}, \vec{\mu}, \Sigma, \delta^2)$
  - ML estimate: $\Lambda^* = \arg\max_{\Lambda} \prod_{d \in C} p(r_d \mid d, \Lambda)$
- Aspect Rating for aspect i

$$r_i(d) = \sum_{w \in V} c_i(w,d)\beta_{i,w}$$

c$_i$(w,d)=0 for words not occurring in aspect segment i

- Aspect Weights: α_i(d) =weight on aspect i

$$\vec{\alpha}(d)^* = \arg\max_{\vec{\alpha}(d)} p(\vec{\alpha}(d) \mid \mu, \Sigma) p(r_d \mid d, \{\beta_{i,w}\}, \delta^2, \vec{\alpha}(d))$$

Maximum a Posteriori

Prior

Likelihood

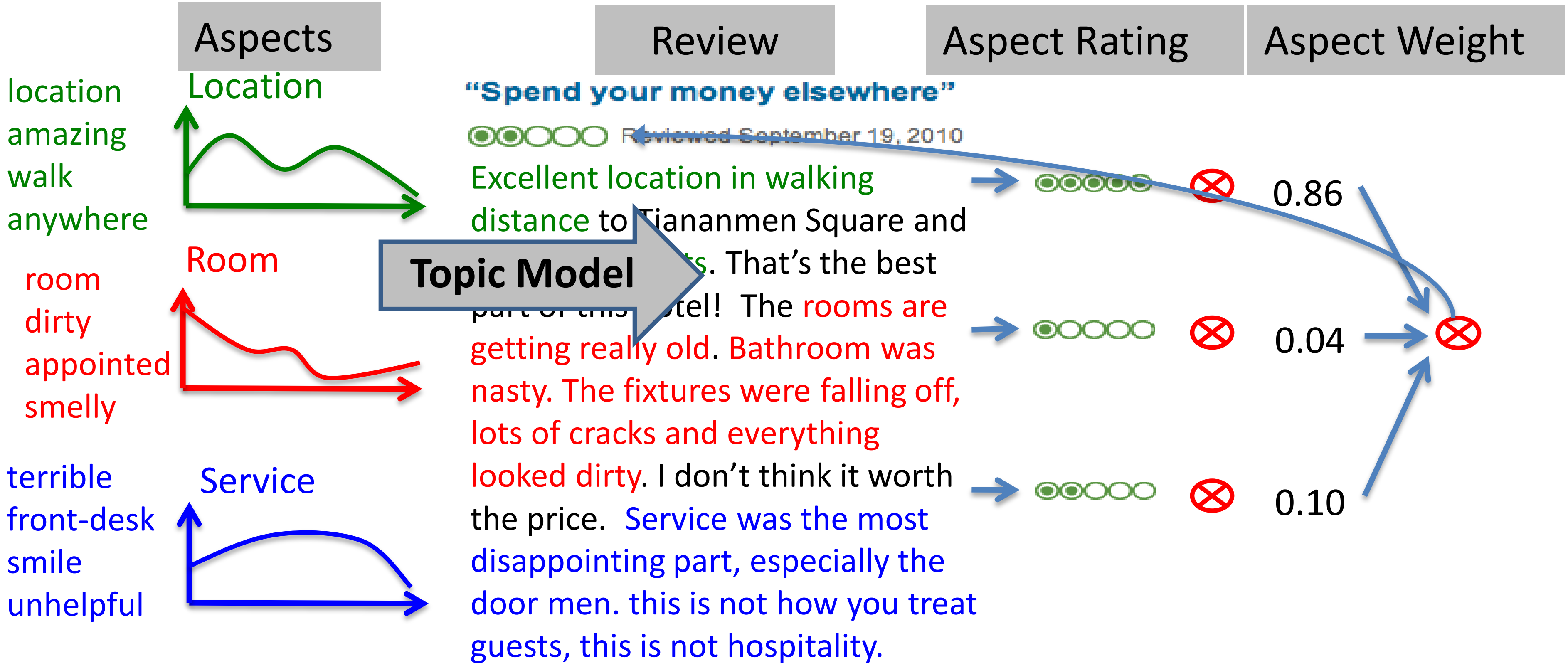
Slide adapted from ChengXiang Zhai's presentation

Suggested Reading

- [Wang et al. 10] Hongning Wang, Yue Lu, and ChengXiang Zhai, Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of ACM KDD 2010,* pp. 783-792, 2010. DOI=10.1145/1835804.1835903

A Unified Generative Model for LARA [Wang et al. 11]

Any Entity



Slide adapted from ChengXiang Zhai's presentation

Sample Result 1: Rating Decomposition [Wang et al. 10]

- Hotels with the same overall rating but different aspect ratings

(All 5 Stars hotels, ground-truth in parenthesis)

Hotel	*Value*	*Room*	*Location*	*Cleanliness*
HOTEL 1	4.2(4.7)	3.8(3.1)	4.0(4.2)	4.1(4.2)
HOTEL 2	4.3(4.0)	3.9(3.3)	3.7(3.1)	4.2(4.7)
HOTEL 3	3.7(3.8)	4.4(3.8)	4.1(4.9)	4.5(4.8)

- Reveal detailed opinions at the aspect level

Slide adapted from ChengXiang Zhai's presentation

Sample Result 2: Comparison of Reviewers [Wang et al. 10]

- Per-Reviewer Analysis
  - Different reviewers' ratings on the same hotel

Reviewer	*Value*	*Room*	*Location*	*Cleanliness*
Reviewer 1	3.7(4.0)	3.5(4.0)	3.7(4.0)	5.8(5.0)
Reviewer 2	5.0(5.0)	3.0(3.0)	5.0(4.0)	3.5(4.0)

  - Reveal differences in opinions of different reviewers

Sample Result 3: Aspect-Specific Sentiment Lexicon [Wang et al. 10]

Value	***Rooms***	***Location***	***Cleanliness***
resort 22.80	view 28.05	restaurant 24.47	clean 55.35
value 19.64	comfortable 23.15	walk 18.89	smell 14.38
excellent 19.54	modern 15.82	bus 14.32	linen 14.25
worth 19.20	quiet 15.37	beach 14.11	maintain 13.51
bad -24.09	*carpet -9.88*	*wall -11.70*	*smelly -0.53*
money -11.02	*smell -8.83*	*bad -5.40*	*urine -0.43*
terrible -10.01	*dirty -7.85*	*road -2.90*	*filthy -0.42*
overprice -9.06	*stain -5.85*	*website -1.67*	*dingy -0.38*

Learn sentimental information directly from the data.

Slide adapted from ChengXiang Zhai's presentation

Sample Result 4: Validating Preference Weights [Wang et al. 10]

Top-10: Reviewers with the highest Val/X ratio (emphasize "value")

Bot-10: Reviewers with the lowest Val/X ratio (emphasize a non-value aspect)

City	*Avg. Price*	*Group*	*Val/Loc*	*Val/Rm*	*Val/Ser*
Amsterdam	241.6	top-10	190.7	214.9	221.1
		bot-10	**270.8**	**333.9**	**236.2**
San Francisco	261.3	top-10	214.5	249.0	225.3
		bot-10	**321.1**	**311.1**	**311.4**
Florence	272.1	top-10	269.4	248.9	220.3
		bot-10	**298.9**	**293.4**	**292.6**

Higher!

Slide adapted from ChengXiang Zhai's presentation

Application 1: Rated Aspect Summarization

Aspect	*Summary*	*Rating*
Value	Truly unique character and a great location at a reasonable price Hotel Max was an excellent choice for our recent three night stay in Seattle.	**3.1**
	Overall not a negative experience; however, considering that the hotel industry is very much in the impressing business, there was a lot of room for improvement.	**1.7**
Location	The location, a short walk to downtown and Pike Place market, made the hotel a good choice.	**3.7**
	When you visit a big metropolitan city, be prepared to hear a little traffic outside!	**1.2**
Business Service	You can pay for wireless by the day or use the complimentary Internet in the business center behind the lobby, though.	**2.7**
	My only complaint is the daily charge for Internet access when you can pretty much connect to wireless on the streets anymore.	**0.9**

Slide adapted from ChengXiang Zhai's presentation

Application 2: Discover Consumer Preferences [Wang et al. 2011]

- Amazon reviews: No guidance

Table 2: Topical Aspects Learned on MP3 Reviews

Low Overall Ratings			High Overall Ratings		
unit	jack	service	files	player	vision
usb	headphone	charge	format	music	video
battery	warranty	problem	included	download	player
charger	replacement	support	easy	headphones	quality
reset	problem	hours	convert	button	great
time	player	months	mp3	set	product
hours	back	weeks	videos	hours	sound
work	months	back	file	buds	radio
thing	buy	customer	wall	volume	accessory
wall	amazon	time	hours	ear	fm
battery life	accessory	service	file format	volume	video

Slide adapted from ChengXiang Zhai's presentation

Application 3: User Rating Behavior Analysis [Wang et al. 10]

	Expensive Hotel		*Cheap Hotel*	
	5 Stars	*3 Stars*	*5 Stars*	*1 Star*
Value	0.134	0.148	**0.171**	0.093
Room	0.098	0.162	0.126	0.121
Location	0.171	0.074	0.161	0.082
Cleanliness	0.081	**0.163**	0.116	**0.294**
Service	**0.251**	0.101	0.101	0.049

People like expensive hotels because of good service.

People like cheap hotels because of good value.

Application 4: Personalized Ranking of Entities [Wang et al. 10]

Query: 0.9 value
0.1 others

	Hotel	Overall Rating	Price	Location
Approach 1	Majestic Colonial	5.0	339	Punta Cana
	Agua Resort	5.0	753	Punta Cana
	Majestic Elegance	5.0	537	Punta Cana
	Grand Palladium	5.0	277	Punta Cana
	Iberostar	5.0	157	Punta Cana
Approach 2	Elan Hotel Modern	5.0	216	Los Angeles
	Marriott San Juan Resort	4.0	354	San Juan
	Punta Cana Club	5.0	409	Punta Cana
	Comfort Inn	5.0	155	Boston
	Hotel Commonwealth	4.5	313	Boston

Non-personalized → Approach 1

Personalized

(Query-specific) → Approach 2

Summary of Opinion Mining

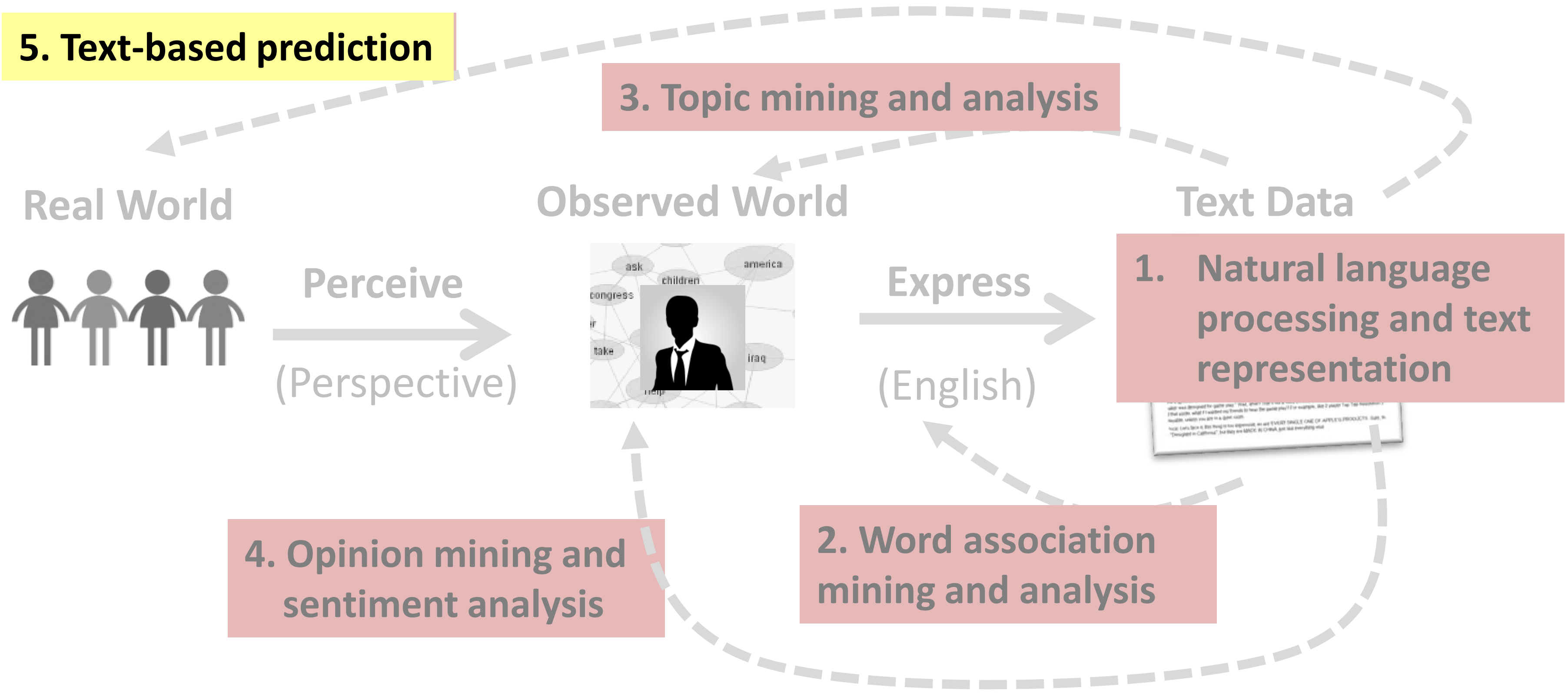
- Very important with a lot of applications!
- Sentiment analysis can be done using text categorization techniques
  - With enriched feature representation
  - With consideration of ordering of the categories
- Generative models are powerful for mining latent user preferences
- Most approaches were proposed for product reviews
- Opinion mining from news and social media remains challenging

Suggested Reading

- Bing Liu, *Sentiment analysis and opinion mining*, Morgan & Claypool Publishers, 2012.
- Bo Pang and Lillian Lee, Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval* 2(1-2), pp. 1–135, 2008.
- Hongning Wang, Yue Lu, and ChengXiang Zhai, Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of ACM KDD 2010,* pp. 783-792, 2010. DOI=10.1145/1835804.1835903
- Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of ACM KDD 2011*, pp. 618-626. DOI=10.1145/2020408.2020505

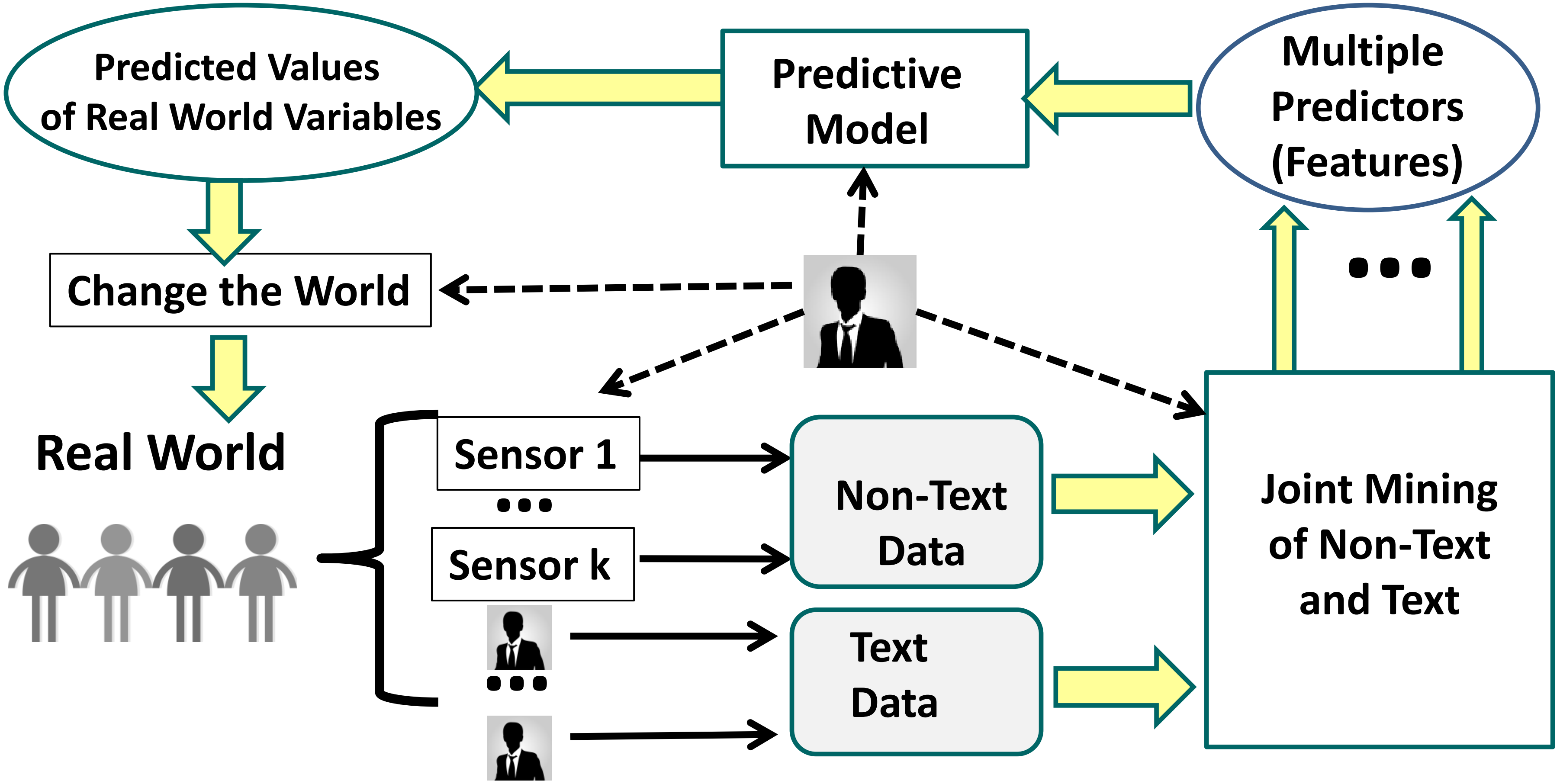
Text-Based Prediction

Text-Based Prediction



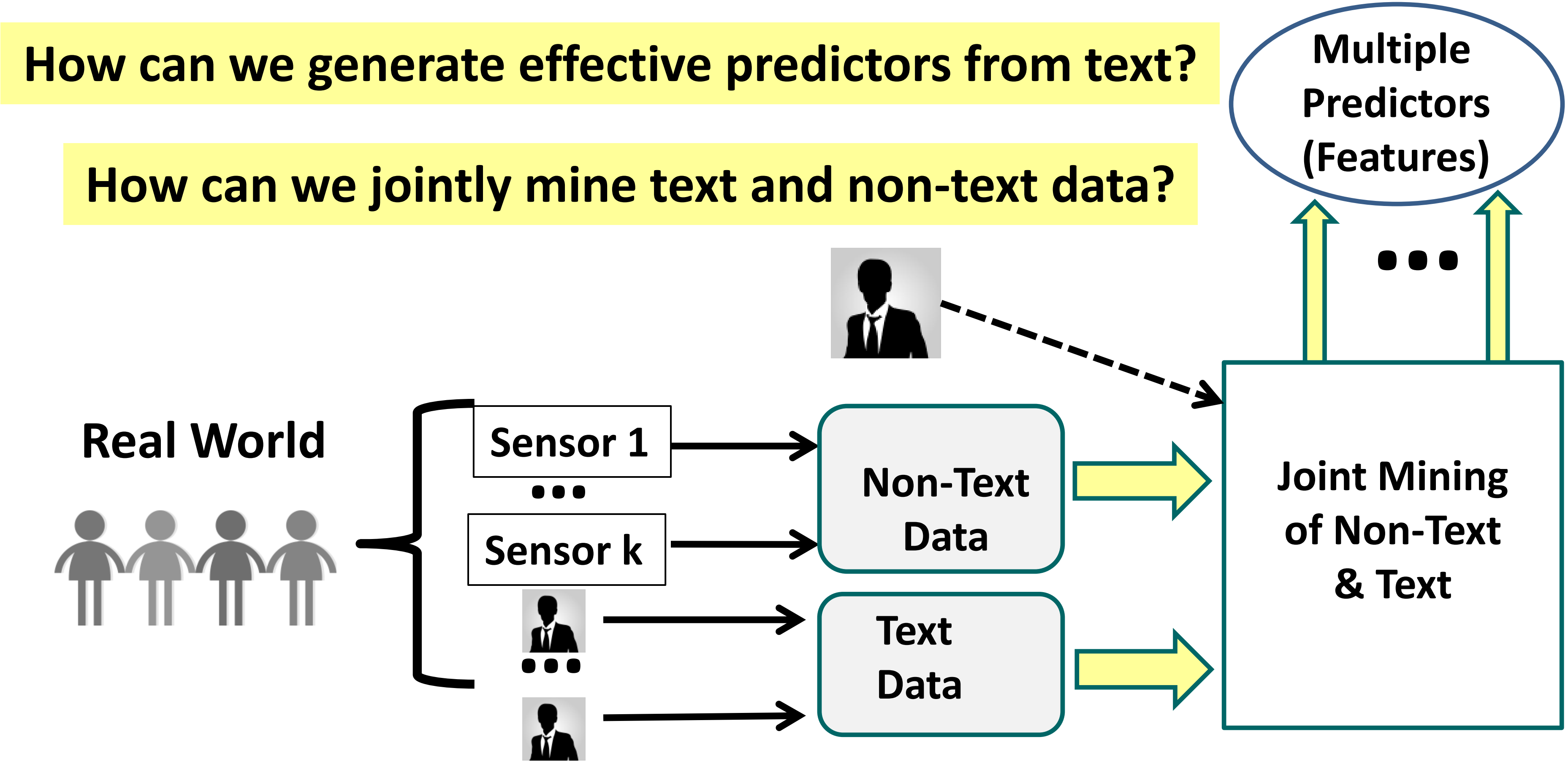
Slide adapted from ChengXiang Zhai's presentation

The Big Picture of Prediction: Data Mining Loop



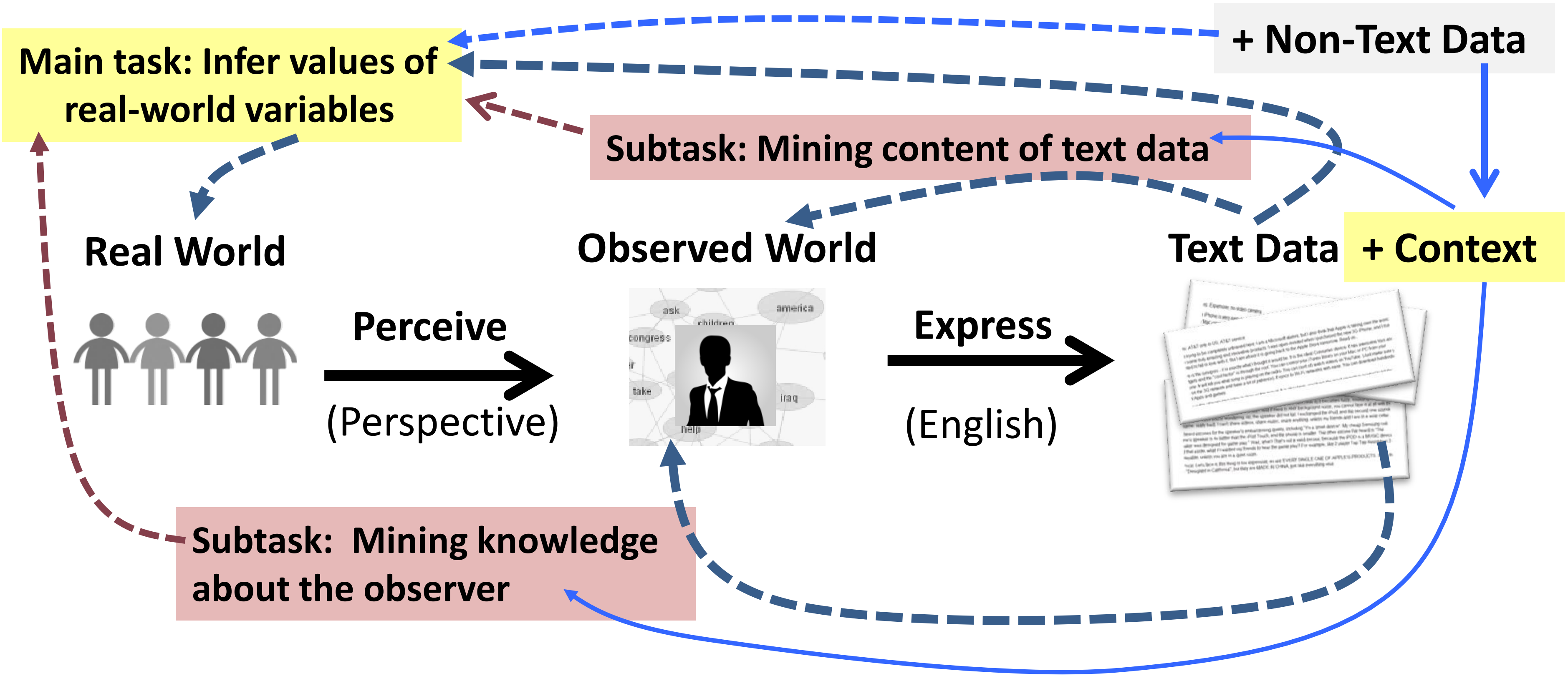
Slide adapted from ChengXiang Zhai's presentation

Text-Based Prediction



Slide adapted from ChengXiang Zhai's presentation

Text-Based Prediction = a Unified View of Text Mining and Analysis



Slide adapted from ChengXiang Zhai's presentation

Joint Mining and Analysis of Text and Non-Text Data

- Non-text data help text mining
  - Non-text data provide context for mining text data
  - **Contextual Text Mining**: Mining text in the context defined by non-text data (see [Mei 2009] for a large body of work)
- Text data help non-text data mining
  - Text data help interpret patterns discovered from non-text data
  - **Pattern Annotation**: Using text data to interpret patterns found in non-text data (see [Mei et al. 2006] for detail)

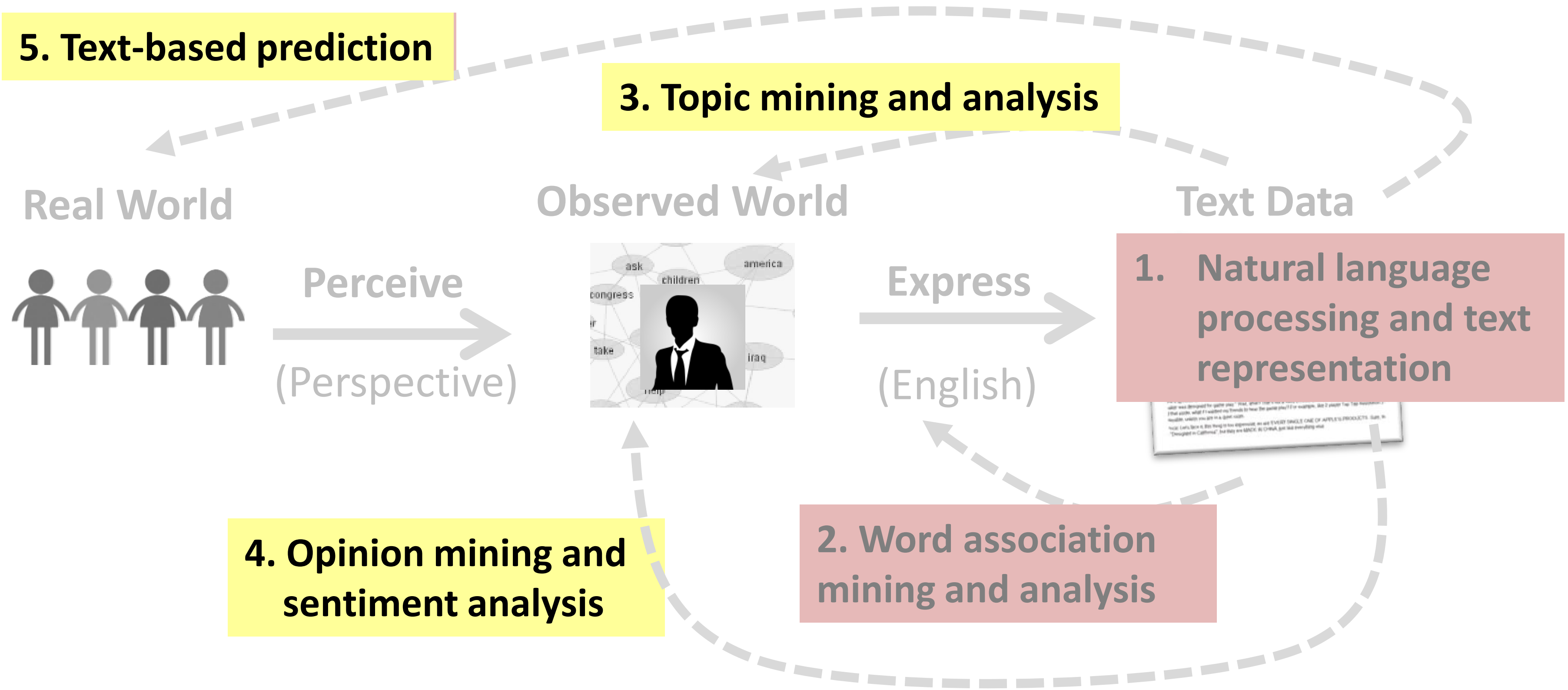
Suggested Reading

- **[Mei et al. 2006]** Qiaozhu Mei, Dong Xin, Hong Cheng, Jiawei Han, and ChengXiang Zhai. 2006. Generating semantic annotations for frequent patterns with context analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD 2006). ACM, New York, NY, USA, 337-346. DOI=10.1145/1150402.1150441
- **[Mei 2009]** Qiaozhu Mei, Contextual Text Mining, Ph.D. Thesis, University of Illinois at Urbana-Champaign, 2009. http://hdl.handle.net/2142/14707

Slide adapted from ChengXiang Zhai's presentation

Contextual Text Mining: Motivation

Contextual Text Mining

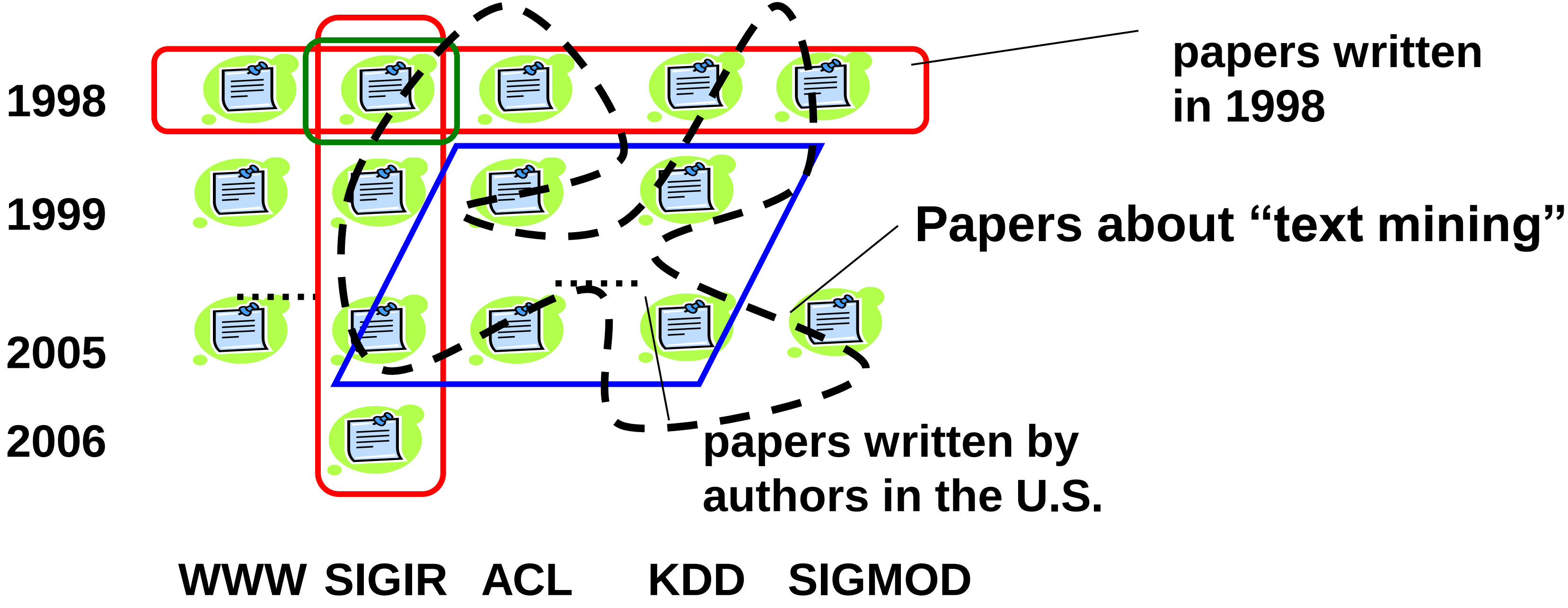


Slide adapted from ChengXiang Zhai's presentation

Contextual Text Mining: Motivation

- Text often has rich context information
  - Direct context (Meta-Data): time, location, authors, source, …
  - Indirect context (additional data related to meta-data): social network of the author, author's age, other text from the same source, etc.
  - Any related data can be regarded as context
- Context can be used to
  - Partition text data for comparative analysis
  - Provide meaning to the discovered topics

Context = Partitioning of Text



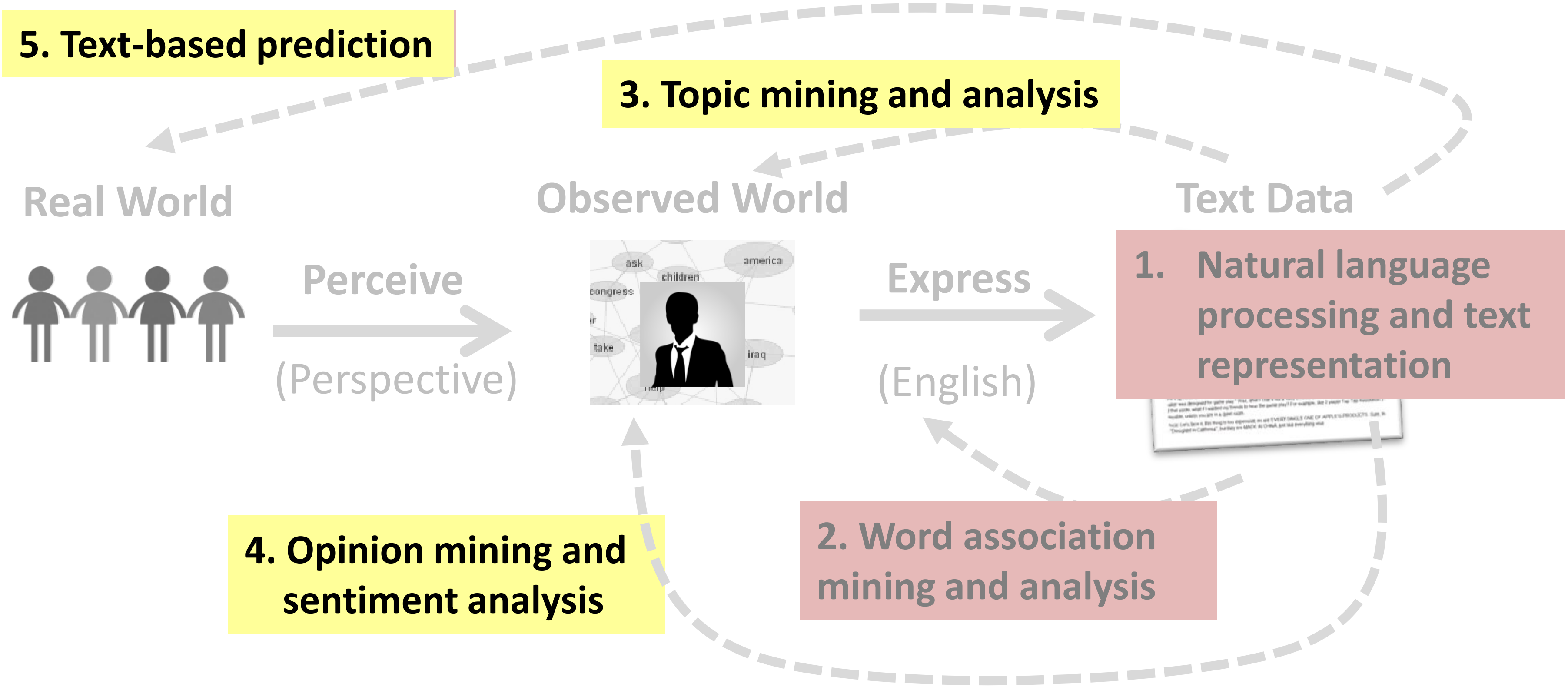
Enables discovery of knowledge associated with different context as needed

Many Interesting Questions Require Contextual Text Mining

- What topics have been gaining increasing attention recently in data mining research? (time as context)
- Is there any difference in the responses of people in different regions to the event? (location as context)
- What are the common research interests of two researchers? (authors as context)
- Is there any difference in the research topics published by authors in the USA and those outside? (author's affiliation and location as context)
- Is there any difference in the opinions about a topic expressed on one social network and another? (social network of authors and topic as context)
- Are there topics in news data that are correlated with sudden changes in stock prices? (time series as context)
- What issues "mattered" in the 2012 presidential election? (time series as context)

Contextual Text Mining:
Contextual Probabilistic Latent Semantic Analysis

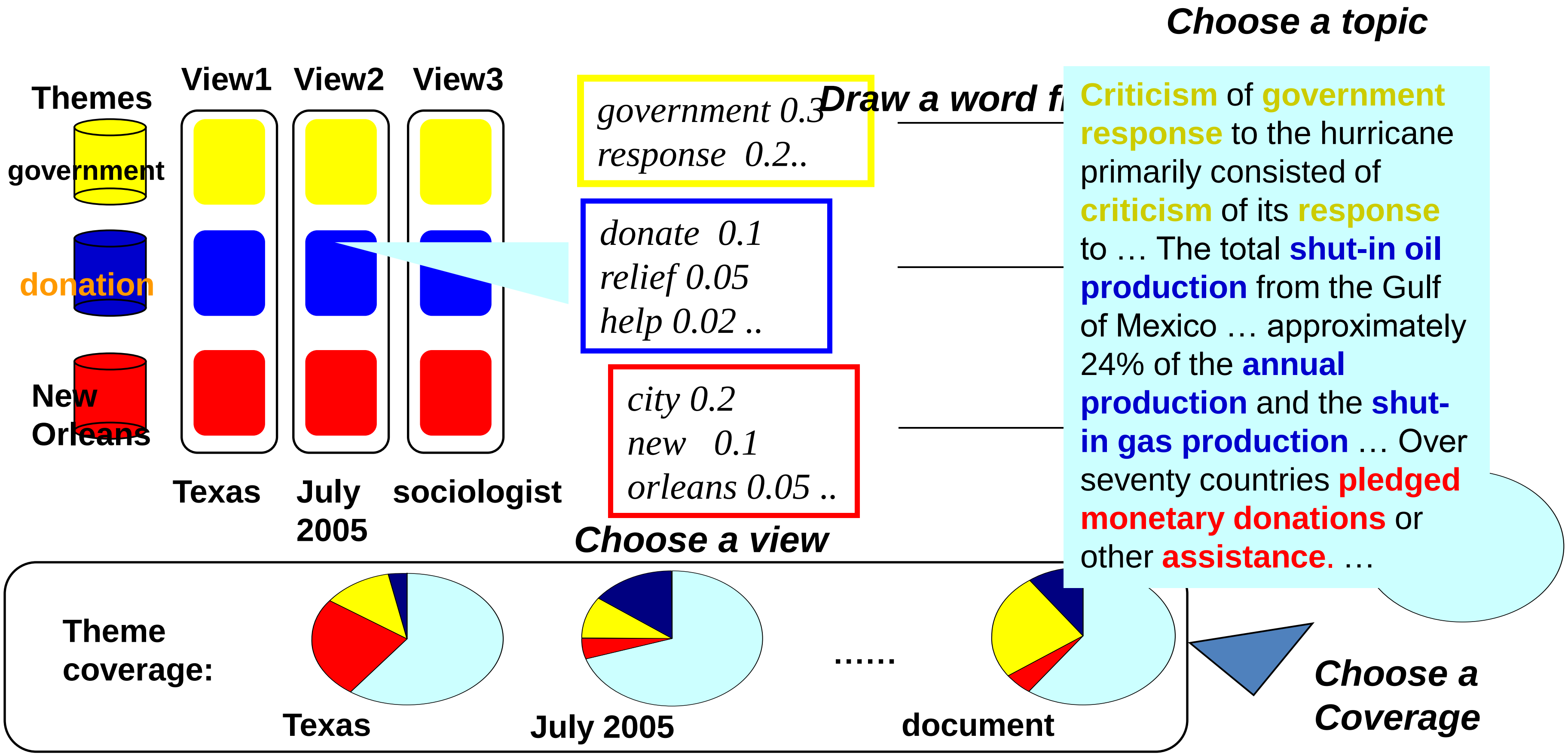
Contextual Text Mining: Contextual Probabilistic Latent Semantic Analysis



Contextual Probabilistic Latent Semantic Analysis (CPLSA) [Mei & Zhai 06]

- General idea:
  - Explicitly add interesting context variables into a generative model (➔ enable discovery contextualized topics)
  - Context influences both coverage and content variation of topics
- As an extension of PLSA
  - Model the conditional likelihood of text given context
  - Assume context-dependent views of a topic
  - Assume context-dependent topic coverage
  - EM algorithm can still be used for parameter estimation
  - Estimated parameters naturally contain context variables, enabling contextual text mining

Generation Process of CPLSA



Slide adapted from ChengXiang Zhai's presentation

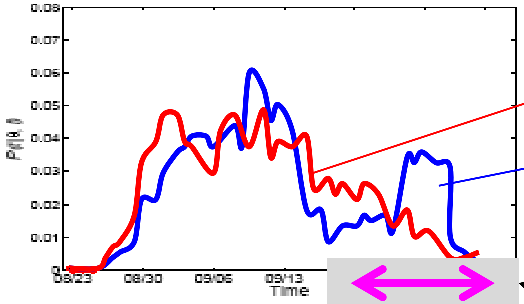
Comparing News Articles [Zhai et al. 04]
Iraq War (30 articles) vs. Afghan War (26 articles)

The common theme indicates that "United Nations" is involved in both wars

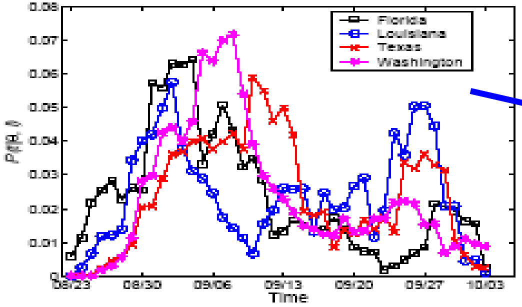
	Cluster 1	**Cluster 2**	**Cluster 3**
Common Theme	united 0.042 nations 0.04 …	killed 0.035 month 0.032 deaths 0.023 …	…
Iraq Theme	n 0.03 Weapons 0.024 Inspections 0.023 …	troops 0.016 hoon 0.015 sanches 0.012 …	…
Afghan Theme	Northern 0.04 alliance 0.04 kabul 0.03 taleban 0.025 aid 0.02 …	taleban 0.026 rumsfeld 0.02 hotel 0.012 front 0.011 …	…

Collection-specific themes indicate different roles of "United Nations" in the two wars

Theme Life Cycles in Blog Articles About "Hurricane Katrina" [Mei et al. 06]



(a) Theme life cycles in Texas (Hurricane Katrina)



(b) Theme "New Orleans" over states (Hurricane Katrina)

Hurricane Rita

Oil Price

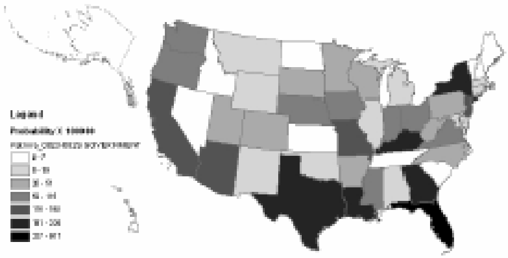
price 0.0772
oil 0.0643
gas 0.0454
increase 0.0210
product 0.0203
fuel 0.0188
company 0.0182
…

New Orleans

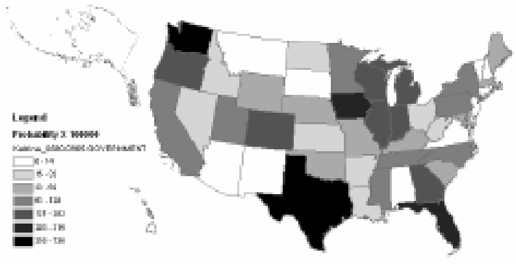
city 0.0634
orleans 0.0541
new 0.0342
louisiana 0.0235
flood 0.0227
evacuate 0.0211
storm 0.0177
…

Slide adapted from ChengXiang Zhai's presentation

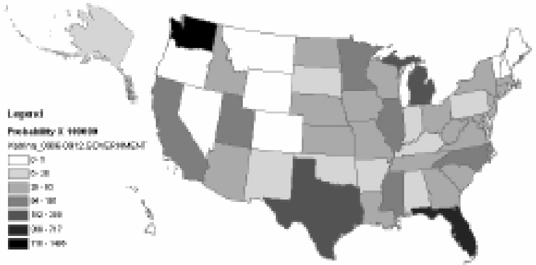
Spatial Distribution of the Topic “Government Response” in Blog Articles About Hurricane Katrina [Mei et al. 06]



(a) Week1: 08/23-08/29

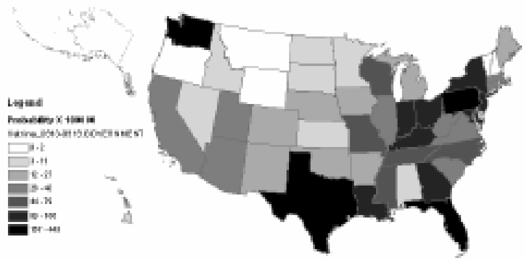


(b) Week Two: 08/30-09/05

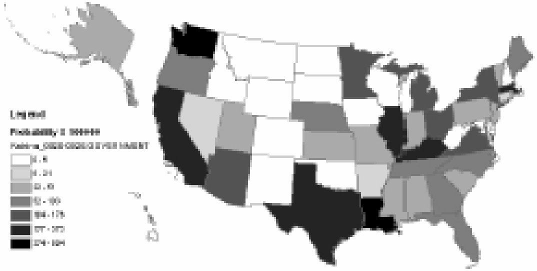


(c) Week Three:09/06-09/12

Theme 1
Government Response
bush 0.0716374
president 0.0610942
federal 0.0514114
govern 0.0476977
fema 0.0474692
administrate 0.0233903
response 0.0208351
brown 0.0199573
blame 0.0170033
governor 0.0142153



(d) Week Four: 09/13-09/19



(e) Week Five: 09/20-09/26

Slide adapted from ChengXiang Zhai's presentation

Event Impact Analysis: IR Research [Mei & Zhai 06]

Topic: retrieval models

SIGIR papers

term	*0.1599*
relevance	*0.0752*
weight	*0.0660*
feedback	*0.0372*
independence	*0.0311*
model	*0.0310*
frequent	*0.0233*
probabilistic	*0.0188*
document	*0.0173*
…	

vector	*0.0514*
concept	*0.0298*
extend	*0.0297*
model	*0.0291*
space	*0.0236*
boolean	*0.0151*
function	*0.0123*
feedback	*0.0077*
…	

xml	*0.0678*
email	*0.0197*
model	*0.0191*
collect	*0.0187*
judgment	*0.0102*
rank	*0.0097*
subtopic	*0.0079*
…	

1992

year

Star

probabilist	*0.0778*
model	*0.0432*
logic	*0.0404*
ir	*0.0338*
boolean	*0.0281*
algebra	*0.0200*
estimate	*0.0119*
weight	*0.0111*
…	

1998

A seminal paper [Croft & Ponte 98]

model	*0.1687*
language	*0.0753*
estimate	*0.0520*
parameter	*0.0281*
distribution	*0.0268*
probable	*0.0205*
smooth	*0.0198*
markov	*0.0137*
likelihood	*0.0059*
…	

Suggested Reading

- **[Zhai et al. 04]** ChengXiang Zhai, Atulya Velivelli, and Bei Yu. 2004. A cross-collection mixture model for comparative text mining. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining* (KDD 2004). ACM, New York, NY, USA, 743-748. DOI=10.1145/1014052.1014150
- **[Mei & Zhai 06]** Qiaozhu Mei and ChengXiang Zhai. 2006. A mixture model for contextual text mining. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining* (KDD 2006). ACM, New York, NY, USA, 649-655. DOI=10.1145/1150402.1150482
- **[Mei et al. 06]** Qiaozhu Mei, Chao Liu, Hang Su, and ChengXiang Zhai. 2006. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *Proceedings of the 15th international conference on World Wide Web* (WWW 2006). ACM, New York, NY, USA, 533-542. DOI=10.1145/1135777.1135857

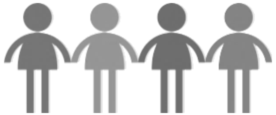
Contextual Text Mining:
Mining Topics with Social Network Context

Contextual Text Mining: Mining Topics with Social Network as Context

5. Text-based prediction

3. Topic mining and analysis

Real World



Perceive

(Perspective)

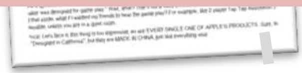
Observed World

Express

(English)

Text Data

1. Natural language processing and text representation



2. Word association mining and analysis

4. Opinion mining and sentiment analysis

Slide adapted from ChengXiang Zhai's presentation

Topic Analysis with Network Context

- The **context** of a text article can form a **network**, e.g.,
  - Authors of research articles may form **collaboration networks**
  - Authors of social media content form **social networks**
  - Locations associated with text can be connected to form a **geographic network**
- **Benefit** of **joint analysis** of text and its network context
  - Network imposes **constraints** on topics in text (**authors connected in a network tend to write about similar topics**)
  - Text helps **characterize** the content associated with each subnetwork (e.g., difference in opinions expressed in two subnetworks?)

Network Supervised Topic Modeling: General Idea [Mei et al. 08]

- Probabilistic topic modeling as optimization: maximize likelihood

$$\Lambda^* = \arg\max_{\Lambda} p(\text{TextData} \mid \Lambda)$$

- Main idea: network imposes constraints on model parameters Λ
  - The text at two adjacent nodes of the network tends to cover similar topics
  - Topic distributions are smoothed over adjacent nodes
  - Add network-induced regularizers to the likelihood objective function

Any generative model

Any network

$$\Lambda^* = \arg\max_{\Lambda} f(p(\text{TextData} \mid \Lambda), r(\Lambda, \text{Network}))$$

Any way to combine

Any regularizer

Slide adapted from ChengXiang Zhai's presentation

Instantiation: NetPLSA [Mei et al. 08]

Network-induced prior: Neighbors have similar topic distribution

Modified objective function

Text collection

PLSA log-likelihood

$$O(C,G) = (1-\lambda)\cdot\boxed{\left(\sum_{d}\sum_{w} c(w,d)\log\sum_{j=1}^{k} p(\theta_j \mid d)\,p(w \mid \theta_j)\right)}$$

Network graph

$$+\boxed{\lambda}\cdot\left(-\frac{1}{2}\sum_{\langle u,v\rangle\in E}\underline{w(u,v)}\,\underline{\sum_{j=1}^{k}\left(p(\theta_j \mid u)-p(\theta_j \mid v)\right)^2}\right)$$

Influence of network constraint

Weight of edge (u,v)

Quantify the difference in the topic coverage at node u and v

Mining 4 Topical Communities: Results of PLSA

Can't uncover the 4 communities (IR, DM, ML, Web)

Topic 1		Topic 2		Topic 3		Topic 4	
term	0.02	peer	0.02	visual	0.02	interface	0.02
question	0.02	patterns	0.01	analog	0.02	towards	0.02
protein	0.01	mining	0.01	neurons	0.02	browsing	0.02
training	0.01	clusters	0.01	vlsi	0.01	xml	0.01
weighting	0.01	stream	0.01	motion	0.01	generation	0.01
multiple	0.01	frequent	0.01	chip	0.01	design	0.01
recognition	0.01	e	0.01	natural	0.01	engine	0.01
relations	0.01	page	0.01	cortex	0.01	service	0.01
library	0.01	gene	0.01	spike	0.01	social	0.01

Slide adapted from ChengXiang Zhai's presentation

Mining 4 Topical Communities: Results of NetPLSA

Uncovers the 4 communities well

Information Retrieval		Data Mining		Machine Learning		Web	
retrieval	0.13	mining	0.11	neural	0.06	web	0.05
information	0.05	data	0.06	learning	0.02	services	0.03
document	0.03	discovery	0.03	networks	0.02	semantic	0.03
query	0.03	databases	0.02	recognition	0.02	services	0.03
text	0.03	rules	0.02	analog	0.01	peer	0.02
search	0.03	association	0.02	vlsi	0.01	ontologies	0.02
evaluation	0.02	patterns	0.02	neurons	0.01	rdf	0.02
user	0.02	frequent	0.01	gaussian	0.01	management	0.01
relevance	0.02	streams	0.01	network	0.01	ontology	0.01

Text Information Network

- In general, we can view text data that naturally "lives" in a rich information network with all other related data
- Text data can be associated with
  - Nodes of the network
  - Edges of the network
  - Paths of the network
  - Subnetworks
  - …
- Analysis of text should be using the entire network!

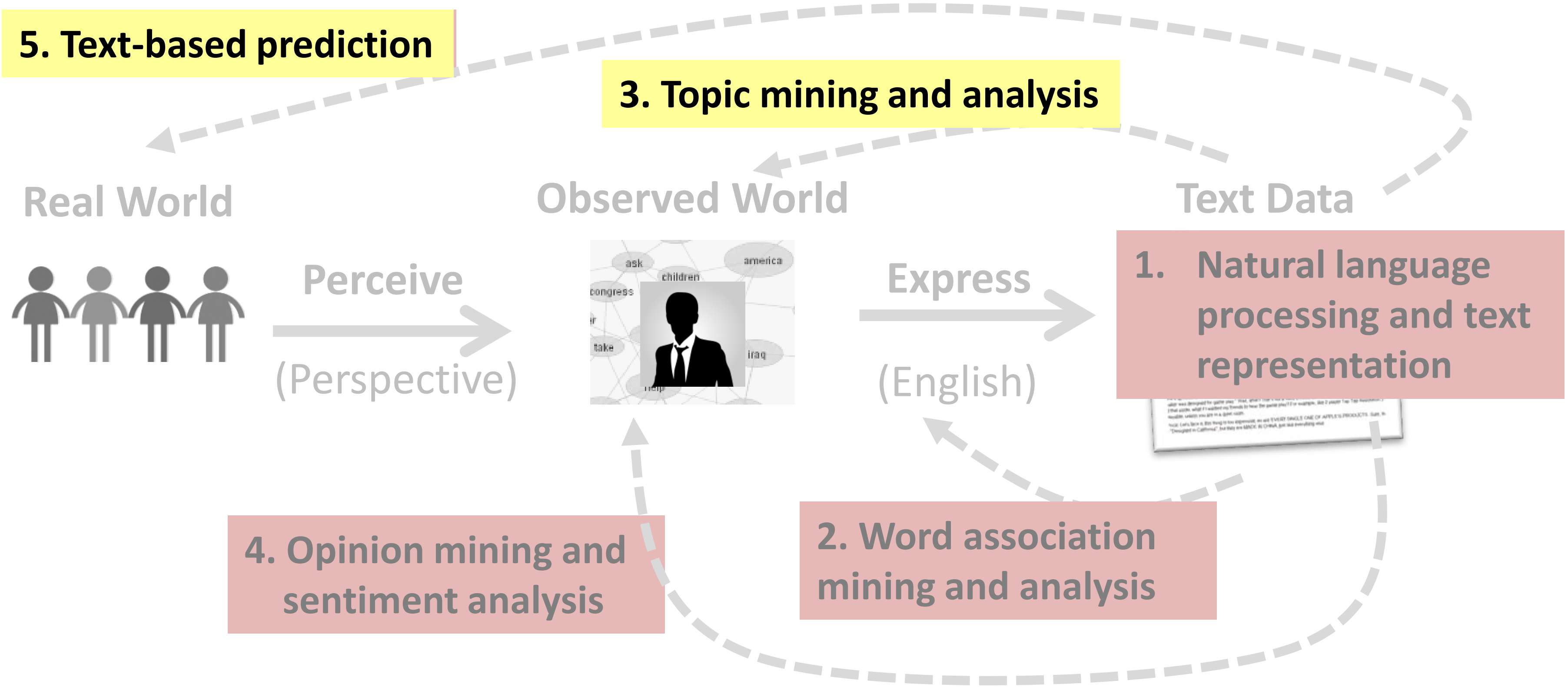
Suggested Reading

- **[Mei et al. 08]** Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *Proceedings of the 17th international conference on World Wide Web* (WWW 2008). ACM, New York, NY, USA, 101-110. DOI=10.1145/1367497.1367512

Contextual Text Mining:
Mining Causal Topics with Time Series Supervision

Slide adapted from ChengXiang Zhai's presentation

Contextual Text Mining: Mining Causal Topics with Time Series Supervision



Slide adapted from ChengXiang Zhai's presentation

Text Mining for Understanding Time Series

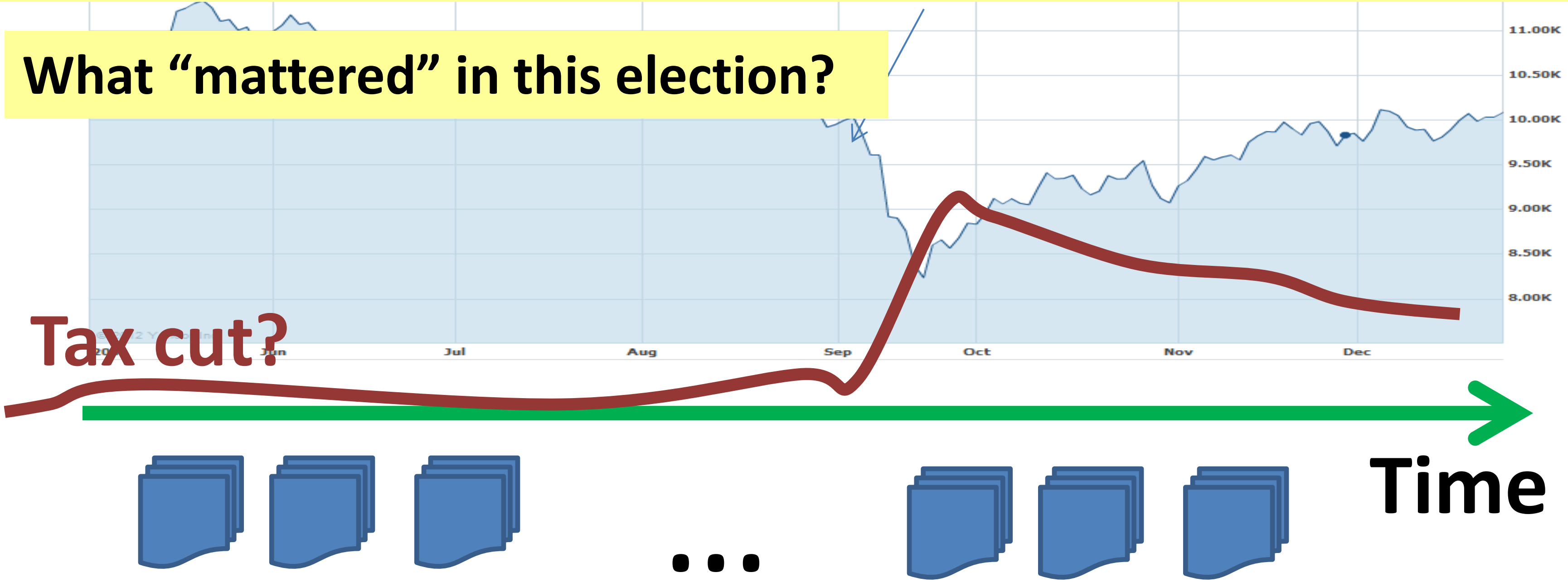
Any clues in the companion news stream?

Dow Jones Industrial Average [Source: Yahoo Finance]

Analysis of Presidential Prediction Markets

What might have caused the sudden drop of price for this candidate?

What "mattered" in this election?

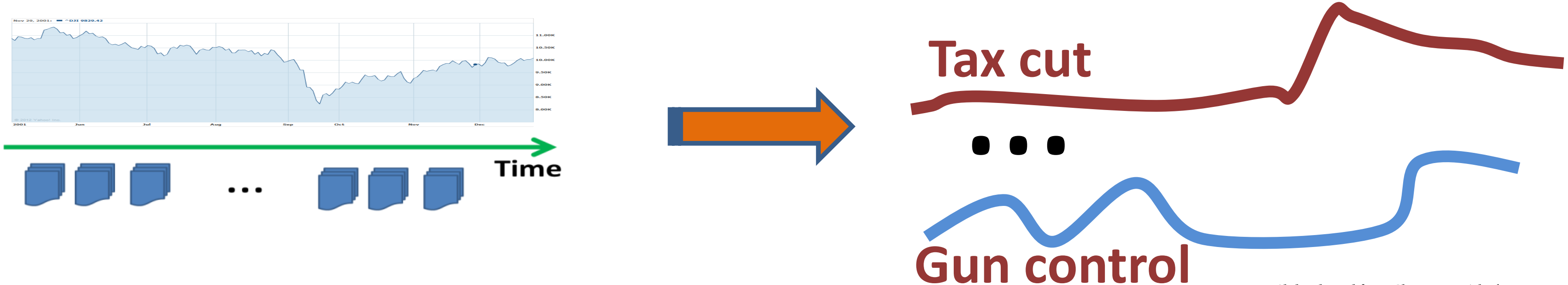


Any clues in the companion news stream?

Slide adapted from ChengXiang Zhai's presentation

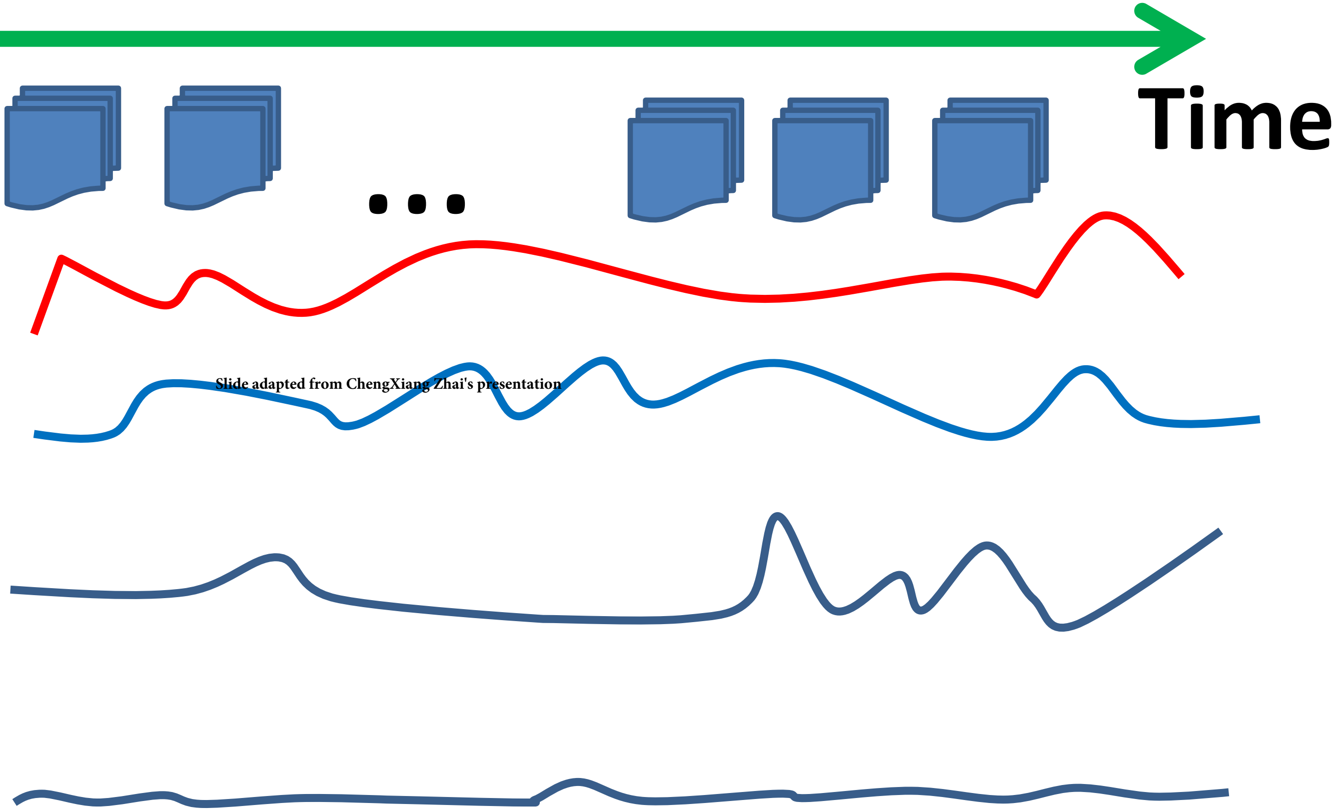
Joint Analysis of Text and Time Series to Discover "Causal Topics"

- Input:
  - Time series
  - Text data produced in a similar time period (text stream)
- Output
  - Topics whose coverage in the text stream has strong correlations with the time series ("causal" topics)



Slide adapted from ChengXiang Zhai's presentation

When a Topic Model Applied to Text Stream



Topic θ_1

government 0.3
response 0.2
…

Topic θ_2

city 0.2
new 0.1
orleans 0.05
…

…

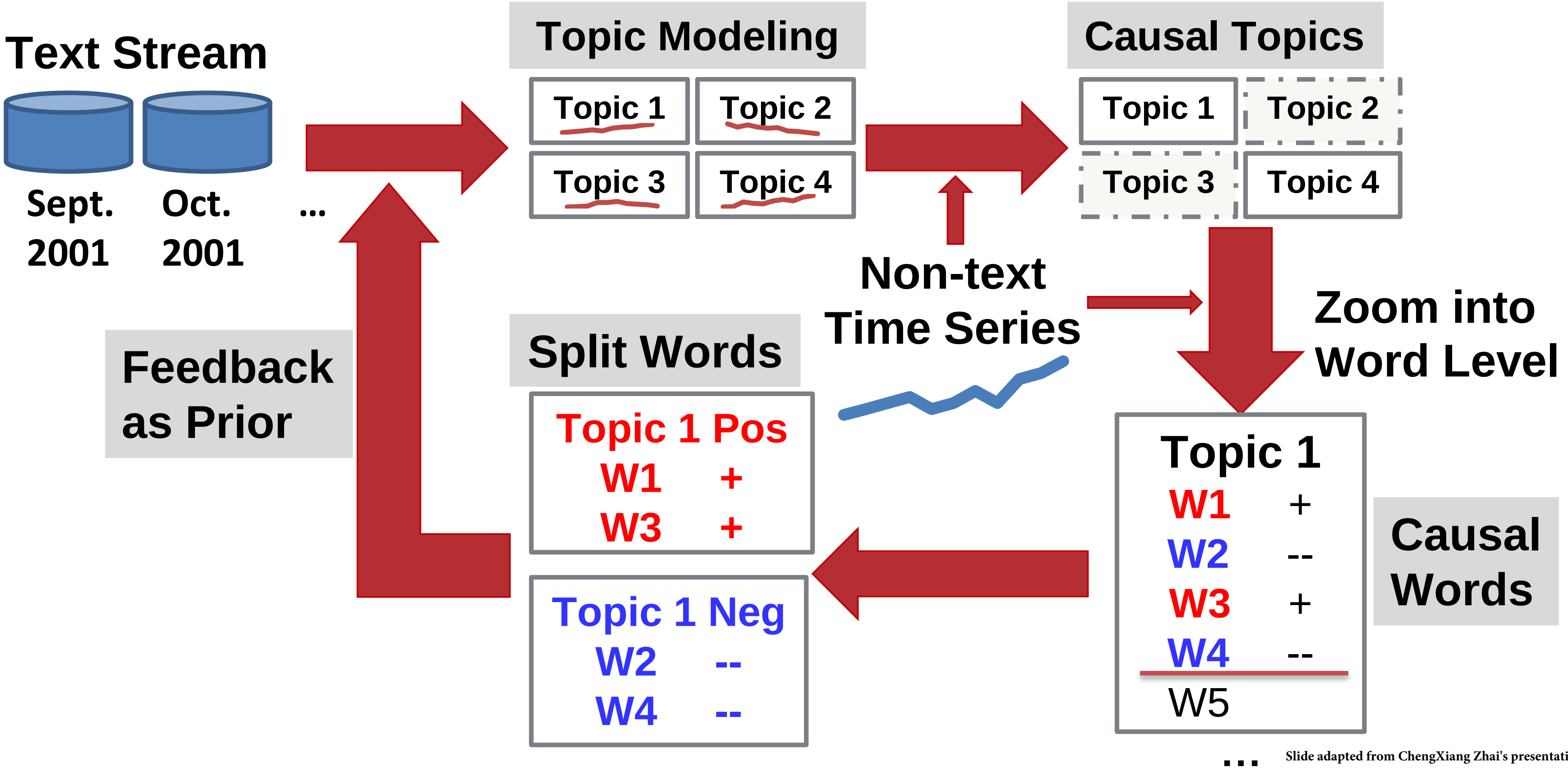
Topic θ_k

donate 0.1
relief 0.05
help 0.02
…

Background θ_k

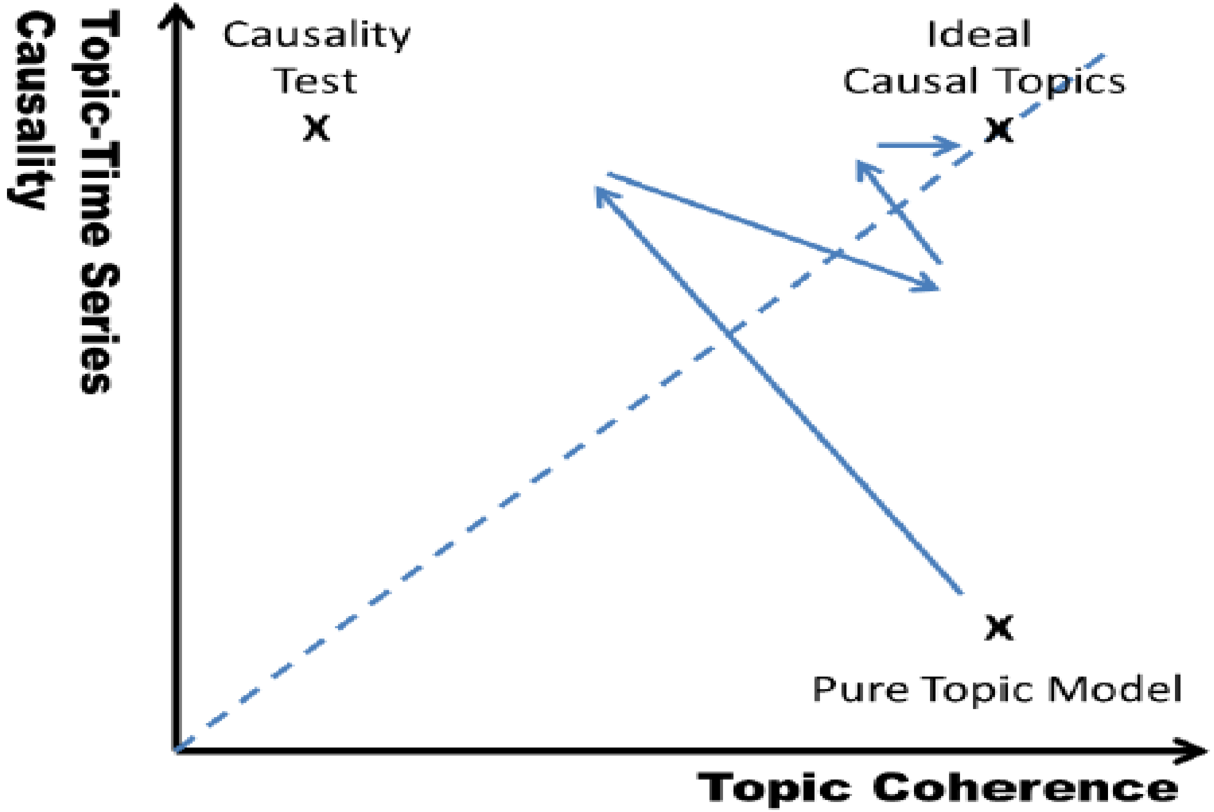
is 0.05
the 0.04
a 0.03

Iterative Causal Topic Modeling [Kim et al. 13]



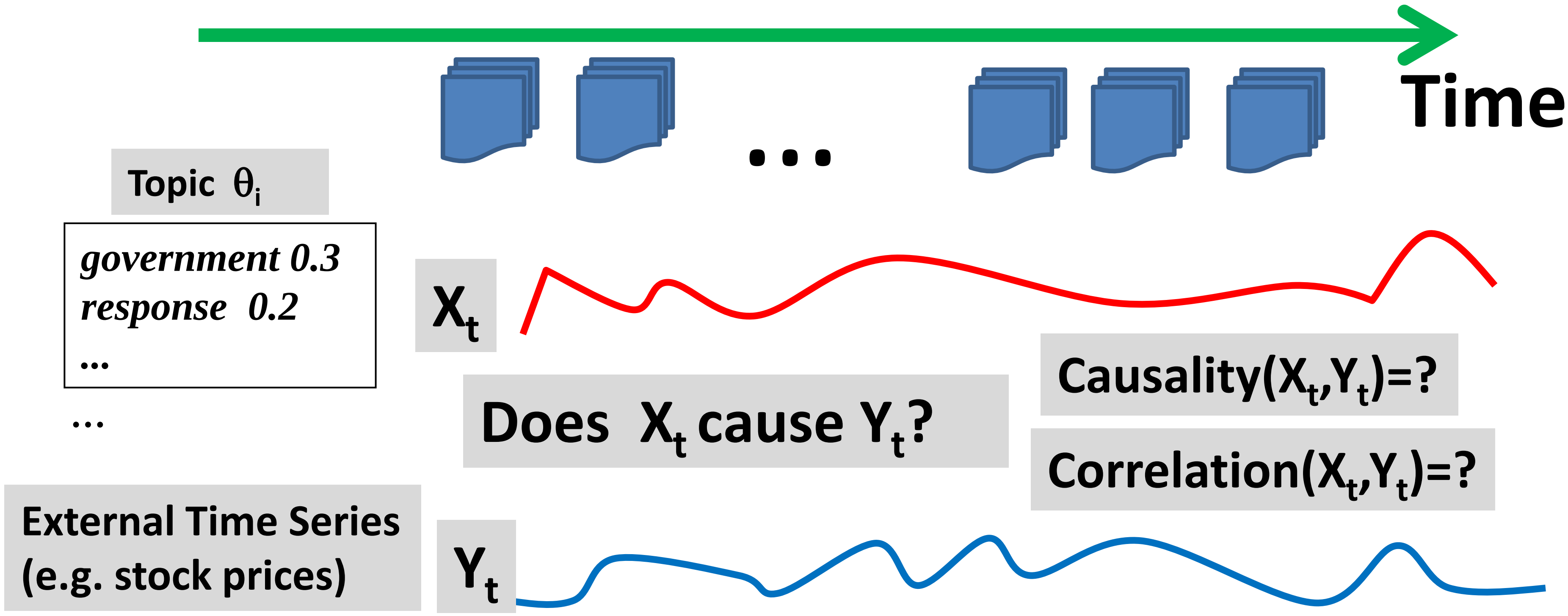
Slide adapted from ChengXiang Zhai's presentation

Heuristic Optimization of Causality + Coherence



Slide adapted from ChengXiang Zhai's presentation

Measuring Causality (Correlation)



Granger Causality Test is often useful [Seth 07]

Slide adapted from ChengXiang Zhai's presentation

Topics in NY Times Correlated with Stocks [Kim et al. 13]: June 2000 ~ Dec. 2011

AAMRQ (American Airlines)	AAPL (Apple)
russia russian putin	paid notice st
europe european germany	russia russian europe
bush gore presidential	olympic games olympics
police court judge	she her ms
airlines airport air united trade terrorism	oil ford prices
food foods cheese	black fashion blacks
nets scott basketball	**computer technology software internet com web**
tennis williams open	football giants jets
awards gay boy	japan japanese plane
moss minnesota chechnya	

Topics are biased toward each time series

Major Topics in 2000 Presidential Election [Kim et al. 13]

Top Three Words in Significant Topics from NY Times
<u>tax cut</u> 1
screen pataki guiliani
enthusiasm door symbolic
<u>oil energy</u> prices
news w top
pres al vice
love tucker presented
partial **<u>abortion</u>** privatization
court supreme **<u>abortion</u>**
<u>gun control</u> nra

Text: NY Times (May 2000 - Oct. 2000)

Time Series: Iowa Electronic Market

http://tippie.uiowa.edu/iem/

Issues known to be important in the 2000 presidential election

Suggested Reading

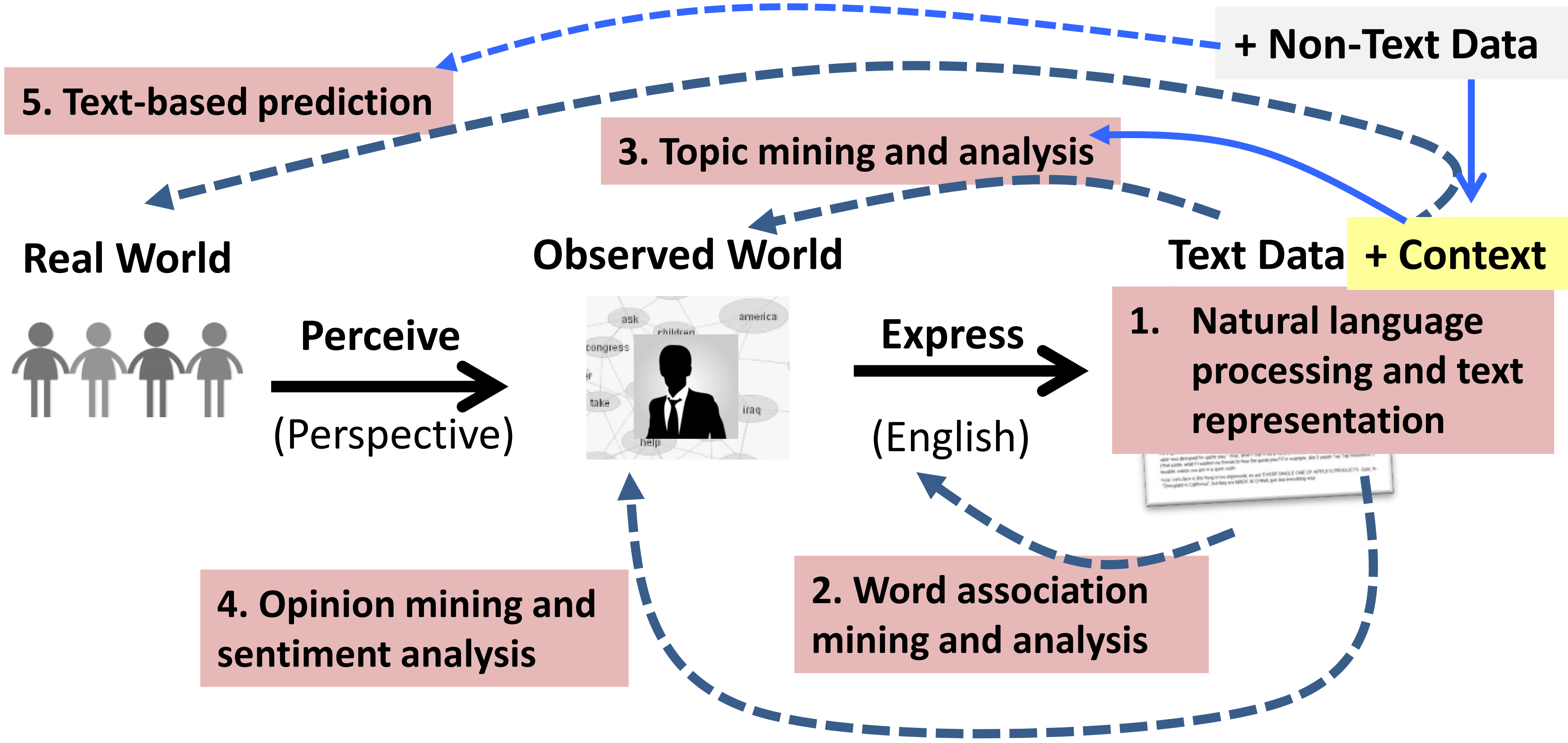
- **[Kim et al. 13]** Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. 2013. Mining causal topics in text data: Iterative topic modeling with time series feedback. In *Proceedings of the 22nd ACM international conference on information & knowledge management* (CIKM 2013). ACM, New York, NY, USA, 885-890. DOI=10.1145/2505515.2505612
- **[Seth 07]** Anil Seth, Granger Causality. 2007. *Scholarpedia*, 2(7): 1667, doi: 10.4249/scholarpedia.1667

Summary of Text-Based Prediction

- Text-based prediction is very useful for "big data" applications:
  - Inferring new knowledge about the world
  - Optimizing decision making
- Text data is often combined with non-text data for prediction
  - Joint analysis of text and non-text is necessary and useful
  - Non-text data provide context for mining text data (contextual text mining)
  - Text data help interpret patterns discovered from non-text data (pattern annotation)
- An active research topic with many open challenges

Course Summary

Topics Covered in This Course



Slide adapted from ChengXiang Zhai's presentation

Key High-Level Take-Away Messages

13. Joint mining of text and non-text
14. Contextual PLSA
15. NetPLSA
16. Causal topic mining

6. Probabilistic Topic Model (PLSA, LDA)
7. Generative model; ML estimate; EM
8. Text clustering: model vs. similarity-based
9. Text categorization: generative vs. discriminative
10. Evaluation of clustering and categorization

Perceive

(Perspective)

1. NLP ➔ Text representation ➔ Knowledge discovery
2. Robust TM = Word-based rep + Statistical analysis

(English)

11. Sentiment classification: ordinal regression
12. Latent Aspect Rating Analysis

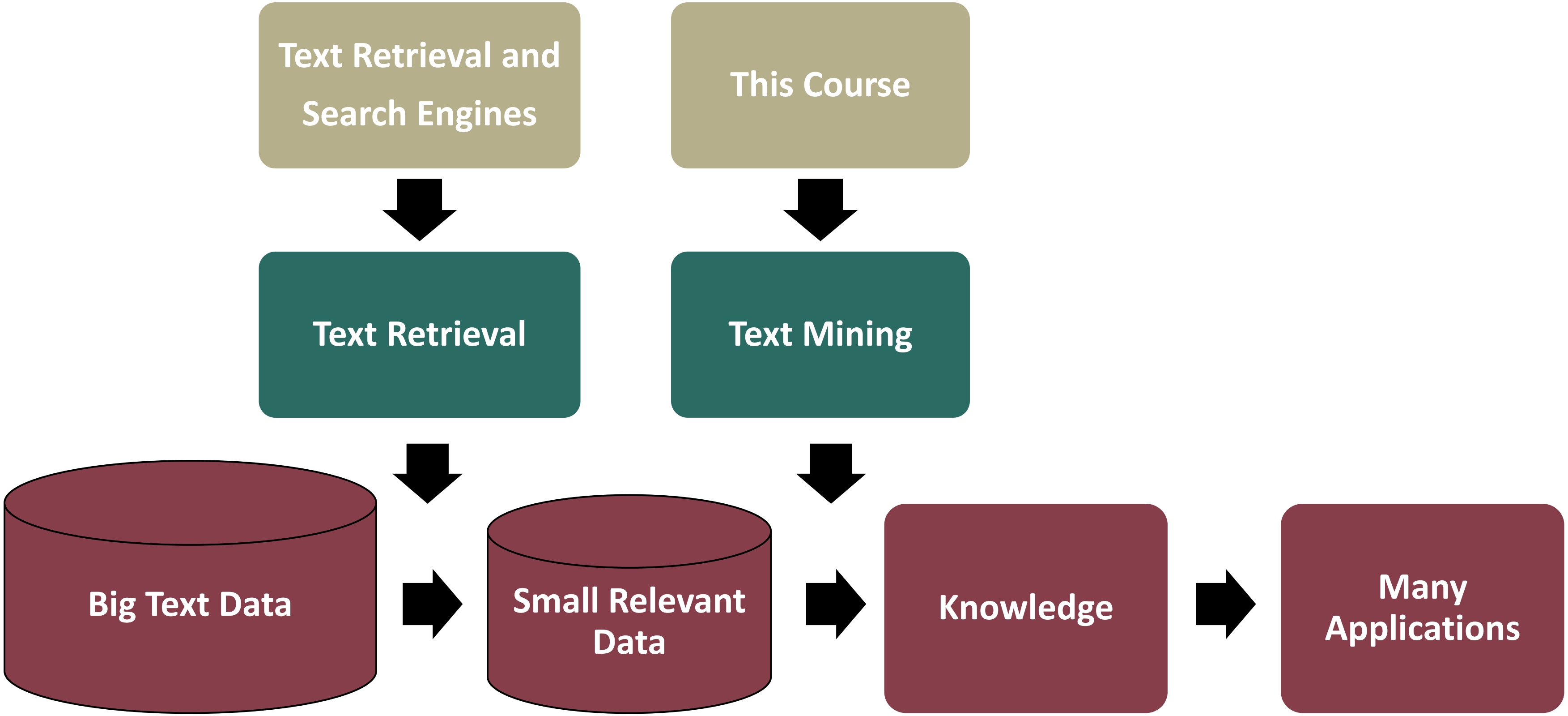
3. Paradigmatic and syntagmatic relations
4. Text similarity: Vector space, BM25
5. Co-occurrence analysis: Entropy, MI

Slide adapted from ChengXiang Zhai's presentation

What to Learn Next

- **Natural Language Processing**
  - Foundation for all text-based applications
  - More NLP ➔ Deeper knowledge discovery
- **Statistical Machine Learning**
  - Backbone techniques for NLP and text analysis
  - Key to predictive modeling and “big data” applications
- **Data Mining**
  - General data mining algorithms can always be applied to text
- **Text/Information Retrieval**
  - Essential system component in any text-based application (human in the loop)
  - Some techniques useful for text data mining

Main Techniques for Harnessing Big Text Data: Text Retrieval + Text Mining



Slide adapted from ChengXiang Zhai's presentation