

CS 589 Fall 2020

Maximum likelihood estimation

Expectation maximization

**Instructor: Susan Liu
TA: Huihui Liu**

Stevens Institute of Technology

Recap of Lecture 2

- RSJ: no parameter
- BM25: Due to the formulation of two-Poisson, parameters are difficult to estimate, so use a parameter free version to replace it
- Language model
 - Leave-one-out
 - EM algorithm

Maximum likelihood estimation

- RSJ:

PRP: rank documents by $p(rel = 1|q, d)$

$$p(rel = 1|q, d) \propto p(d|rel = 1, q)p(rel = 1)$$

$$\begin{aligned}\alpha_i &= p(w_i = 1|q, rel = 1) \\ &= \frac{\text{count}(w_i = 1, rel = 1) + 0.5}{\text{count}(rel = 1) + 1}\end{aligned}$$

$$\begin{aligned}\beta_i &= p(w_i = 0|q, rel = 0) \\ &= \frac{\text{count}(w_i = 0, rel = 0) + 0.5}{\text{count}(rel = 0) + 1}\end{aligned}$$

- Language model

$$\hat{\mu} = \operatorname{argmax}_{\mu} \sum_{w_i=1}^V \sum_d \log p(w_i|d; w_i \notin d)$$

Today's lecture

- Maximum likelihood estimation
- Expectation maximization
 - Coin-topic problem
 - Using EM algorithm to remove stop words
- Mixture of topic models
 - Probabilistic latent semantic analysis
 - PLSA with partial labels

Maximum likelihood estimation

$$\max_{\theta} \sum_i \log P(\mathbf{x}_i; \theta)$$

\mathbf{x}_i **observations** e.g., mice weights

$P(\mathbf{x}_i; \theta)$ **likelihood** e.g., $\mathcal{N}(x_i; \mu, \sigma^2)$

θ **parameters**

If the optimal solution is within θ 's space \mathcal{S} : $\frac{\partial \sum_i \log P(\mathbf{x}_i; \theta)}{\partial \theta} = 0$ at $\theta = \hat{\theta}_{ML}$

Expectation maximization algorithm

- How to estimate the optimal θ ?
- Expectation maximization (EM) algorithm:
 - Relies on the concept of **complete data** space
 - Iterative and alternative between conditional expectation and maximization steps

Incomplete data space: observation, e.g., documents

$$p(\mathbf{x}; \theta)$$

Complete data space: observation + latent variables, e.g., topic

$$p(\mathbf{x}, z|\theta)$$

Expectation maximization algorithm

- Estimating the incomplete probability using the complete space

$$p(\mathbf{x}; \theta) = \sum_{k=1}^K p(\mathbf{x}|z; \theta)p(z = k) \quad \text{discrete space}$$

$$p(\mathbf{x}; \theta) = \int_z p(\mathbf{x}|z; \theta)dp(z) \quad \text{continuous space}$$

- EM algorithm: repeat n=1...N:

E step: $Q(\theta|\hat{\theta}^{(n)}) = \mathbb{E}_{\hat{\theta}^{(n)}}[\log p(\mathbf{x}, z|\theta)]$

M step: $\hat{\theta}^{(n+1)} = \arg \max_{\theta \in \mathcal{S}} Q(\theta|\hat{\theta}^{(n)})$

Expectation maximization: convergence guarantee

- **Theorem:** the likelihood of observation, $\log p(\mathbf{x}; \theta^{(n)})$, monotonously increases with n

$$p(z, \mathbf{x}|\theta) = p(z|\mathbf{x}, \theta)p(\mathbf{x}|\theta)$$

$$\log p(z, \mathbf{x}|\theta) = \log p(z|\mathbf{x}, \theta) + \log p(\mathbf{x}|\theta)$$

$$l(\theta) = l_{cd}(\theta) - \log p(z|\mathbf{x}; \theta)$$

$$l(\theta^{(n+1)}) - l(\theta^{(n)}) = l_{cd}(\theta^{(n+1)}) - l_{cd}(\theta^{(n)}) + \log [p(z|\mathbf{x}, \theta^{(n)})/p(z|\mathbf{x}, \theta^{(n+1)})]$$

Expectation maximization: convergence guarantee

- Take the expectation over $p(z|\mathbf{x}, \theta^{(n)})$ on both side

$$l(\theta^{(n+1)}) - l(\theta^{(n)}) = l_{cd}(\theta^{(n+1)}) - l_{cd}(\theta^{(n)}) + \log [p(z|\mathbf{x}, \theta^{(n)})/p(z|\mathbf{x}, \theta^{(n+1)})]$$

$$\Rightarrow l(\theta^{(n+1)}) - l(\theta^{(n)}) = \mathbb{E}_{p(z|\mathbf{x}, \theta^{(n)})}[l_{cd}(\theta^{(n+1)})] - \mathbb{E}_{p(z|\mathbf{x}, \theta^{(n)})}[l_{cd}(\theta^{(n)})] + D_{KL}(p(z|\mathbf{x}, \theta^{(n+1)} \| p(z|\mathbf{x}, \theta^{(n)}))$$

$Q(\theta^{(n+1)} | \hat{\theta}^{(n)}) - Q(\theta^{(n)} | \hat{\theta}^{(n)})$

EM chooses $\theta^{(n+1)}$ to maximize $Q(\theta^{(n+1)} | \hat{\theta}^{(n)})$

KL divergence always nonneg

$$\Rightarrow l(\theta^{(n+1)}) \geq l(\theta^{(n)})$$

An example problem: Coin-topic problem

- Author H and author T are co-authoring a paper in the following way:
 - At each time, they toss a coin to write the next word. If it's "head", author H writes the next word, if it's "tail", author T writes the next word. The probability for "head" is λ
 - The head author selects the next word by randomly sampling from $p(w|H)$, so does the tail author
- **Problem:** estimating the parameters that maximizes the document likelihood

Coin-topic problem: known topic, unknown λ

- Maximum likelihood estimation:

$$\max_{\lambda} \sum_i \sum_{w=1}^V \log (\lambda p(w|H) + (1 - \lambda)p(w|T))$$

- Suppose both head and tail distributions are known, e.g.:

	the	computer	data	baseball	game	interesting
$p(w T)$	0.2	0.05	0.05	0.4	0.2	0.1
$p(w H)$	0.25	0.2	0.2	0.15	0.1	0.1

Expectation maximization

- We use $p(Z|v)$ to represent the hidden variable, i.e., whether the topic for word v is head or tail topic

$$\begin{aligned} & \log p(\mathbf{D}|\lambda) \\ &= \sum_i \sum_v (p(Z = 0|v) \log \lambda p(v|H) + (1 - p(Z = 0|v)) \log (1 - \lambda) p(v|T)) \end{aligned}$$

- Take the derivative of $\log p(\mathbf{D}|\lambda)$:

$$\begin{aligned} & \sum_i \sum_v (p(Z = 0|v) \cdot \frac{1}{\lambda} + (1 - p(Z = 0|v)) \cdot \frac{1}{1 - \lambda}) = 0 \\ & \Rightarrow \lambda^{(n+1)} = \frac{1}{|D|} \sum_v p^{(n)}(Z = 0|v) \end{aligned} \quad (\text{M step})$$

Expectation maximization

- We use $p(Z|v)$ to represent the hidden variable, i.e., whether the topic for word v is head or tail topic

$$\begin{aligned} & \log p(\mathbf{D}|\lambda) \\ &= \sum_d \sum_v (p(Z = 0|v) \log \lambda p(v|H) + (1 - p(Z = 0|v)) \log (1 - \lambda) p(v|T)) \end{aligned}$$

- E step: the standard derivation is to apply Bayes theorem:

$$p^{(n+1)}(Z = 0|v; \mathbf{D}) \propto p(v|Z = 0)p(Z = 0) = p(v|T)\lambda^{(n)}$$

$$p^{(n+1)}(Z = 1|v; \mathbf{D}) \propto p(v|Z = 1)p(Z = 1) = p(v|H)(1 - \lambda^{(n)})$$

Coin-topic problem: unknown topic, known λ

- For the same coin topic problem, assume lambda is known whereas $p(w|H)$ is unknown, estimate $p(w|H)$

$$\max_{\lambda} p(w|H) \sum_d \sum_{w=1}^V \log (\lambda p(w|H) + (1 - \lambda)p(w|T)) + \eta \sum_v (p(v|H) - 1)$$

- Take the derivative and set to 0, we can get (M step):

$$p(v|H) \propto \sum_i \sum_v \mathbb{1}[w_i == v] p(Z = 0|v)$$
$$\Rightarrow p^{(n+1)}(v|H) = \frac{\sum_i \mathbb{1}[w_i == v] \cdot p^{(n)}(Z = 0|v)}{\sum_u \sum_i \mathbb{1}[w_i == u] \cdot p^{(n)}(Z = 0|u)}$$

- E step follows the same posterior estimation as previous slide

Coin-topic problem: unknown topic, known λ

- For the same coin topic problem, assume lambda is known whereas $p(w|H)$ is unknown, estimate $p(w|H)$
- Application: removing background topic
 - Suppose $p(w|H)$ is the main topic (computer game)
 - $p(w|T)$ is the background topic: the: 0.3, a: 0.2, ...,
 - The mixture of head and tail topic is dominated by background words:
 - After stop words removal, the true topic $p(w|T)$ is “revealed”:

Coin-topic problem: unknown topic and λ

- Suppose both $p(w|H)$ and λ are unknown:

$$\Rightarrow \lambda^{(n+1)} = \frac{1}{|D|} \sum_v p^{(n)}(Z = 0|v)$$

$$p(v|H) \propto \sum_i \sum_v [w_i == v] p(Z = 0|v) \quad (\text{M step})$$

$$\Rightarrow p^{(n+1)}(v|H) = \frac{\sum_i \mathbb{1}[w_i == v] \cdot p^{(n)}(Z = 0|v)}{\sum_u \sum_i \mathbb{1}[w_i == u] \cdot p^{(n)}(Z = 0|u)}$$

$$p^{(n+1)}(Z = 0|v; \mathbf{D}) \propto p(v|Z = 0)p(Z = 0) = p(v|T)\lambda^{(n)} \quad (\text{E step})$$

$$p^{(n+1)}(Z = 1|v; \mathbf{D}) \propto p(v|Z = 1)p(Z = 1) = p(v|H)(1 - \lambda^{(n)})$$

Applications of Coin Topic Problem for Text Mining

- Application Scenarios:

- $p(w|H)$ & $p(w|T)$ are known; estimate λ
- $p(w|H)$ & λ are known; estimate $p(w|T)$
- $p(w|H)$ is known; estimate λ & $p(w|T)$
- λ is known; estimate $p(w|H)$ & $p(w|T)$
- Estimate λ , $p(w|H)$, $p(w|T)$

how much percent of the document is about computer game?

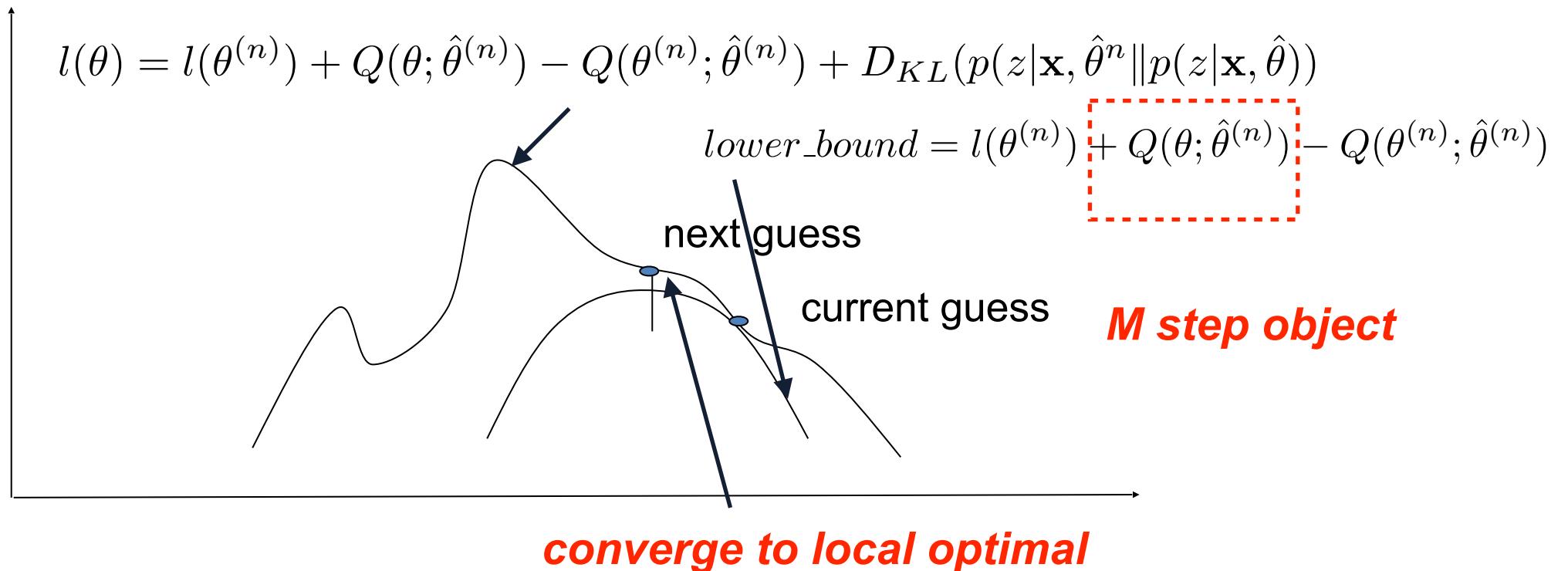
30% of the doc is about computer game, what's the

The doc is about computer game, is it also about some other topic, and if so to what extent?

30% of the doc is about one topic and 70% is about another, what are these two topics?

The doc is about two subtopics, find out what these two subtopics are and to what extent the doc covers each.

Expectation maximization as hill climbing



EM algorithm in action

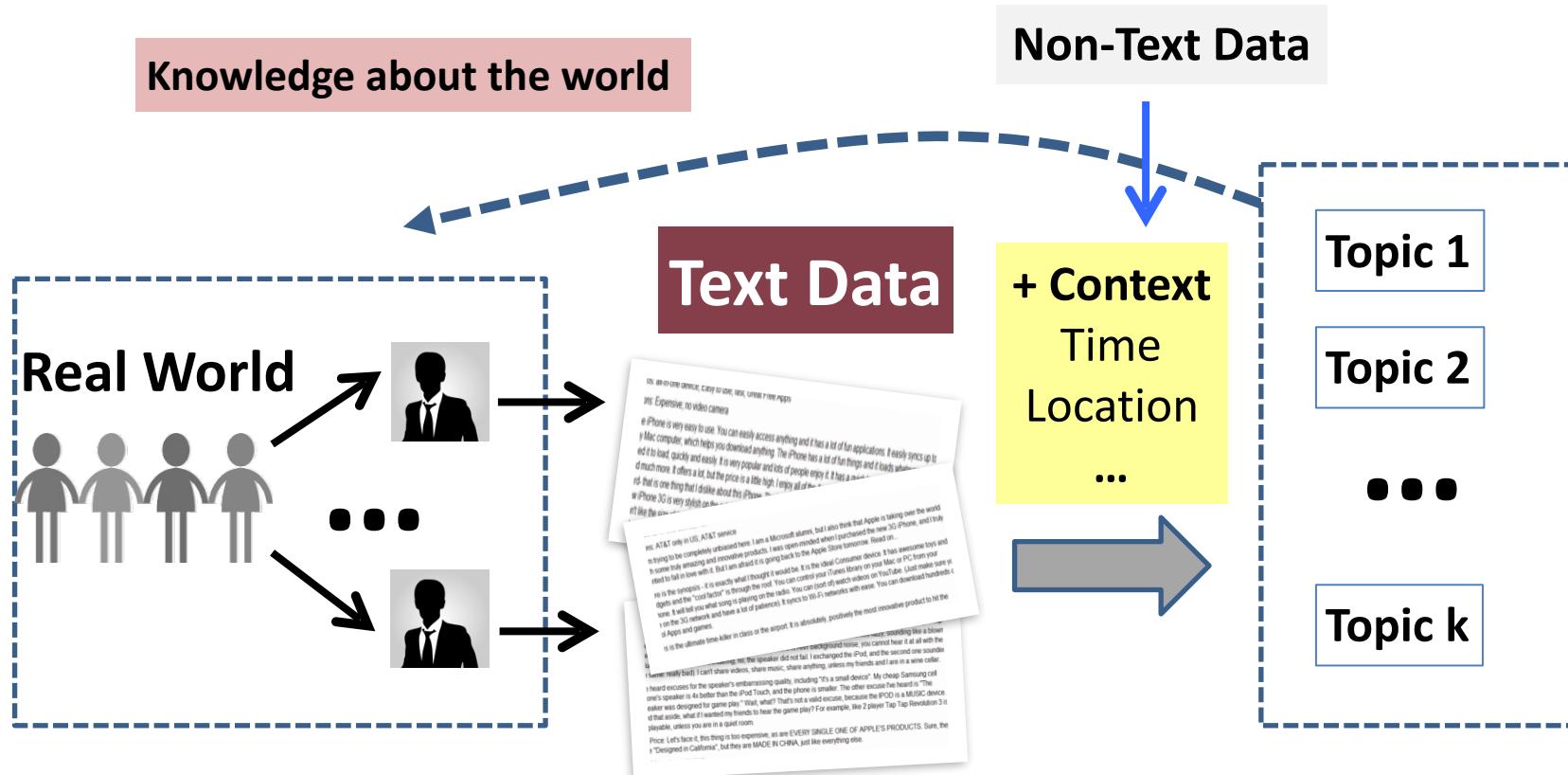
- Log likelihood increases:

Word	#	$p(w \theta_B)$	Iteration 1		Iteration 2		Iteration 3	
			$P(w \theta)$	$p(z=0 w)$	$P(w \theta)$	$p(z=0 w)$	$P(w \theta)$	$p(z=0 w)$
The	4	0.5	0.25	0.33	0.20	0.29	0.18	0.26
Paper	2	0.3	0.25	0.45	0.14	0.32	0.10	0.25
Text	4	0.1	0.25	0.71	0.44	0.81	0.50	0.93
Mining	2	0.1	0.25	0.71	0.22	0.69	0.22	0.69
Log-Likelihood			-16.96		-16.13		-16.02	

Topic models and analysis

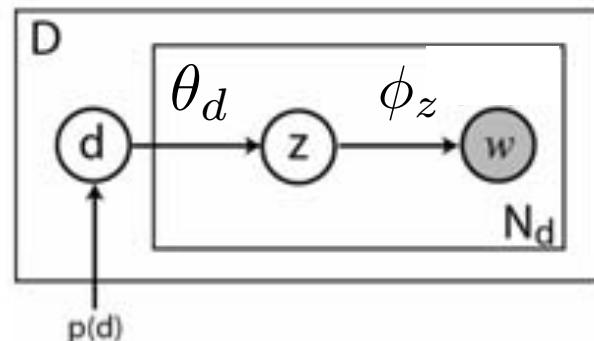
- Topic ≈ main idea discussed in text data
 - Theme/subject of a discussion or conversation
 - Different granularities (e.g., topic of a sentence, an article, etc.)
- Many applications require discovery of topics in text
 - What are Twitter users talking about today?
 - What are the current research topics in data mining? How are they different from those 5 years ago?
 - What do people like about the iPhone 6? What do they dislike?
 - What were the major topics debated in 2012 presidential election?

Lifecycle of topic and text data



A generative process of documents

- Assume documents are generated by sampling words from k latent topics
- For each document d :
 - For each token position i
 - Choose a topic $z \sim \text{Multinomial}(\theta_d)$
 - Choose a term $w \sim \text{Multinomial}(\phi_z)$



Document as a Sample of Mixed Topics

Topic θ_1

government
0.3
response 0.2

Topic θ_2

...

Topic θ_k

city 0.2
new 0.1
orleans
0.05
...

donate 0.1
relief 0.05
help 0.02
...

Background θ_B

the 0.04
a 0.03
...

Blog article about “Hurricane Katrina”

[Criticism of government response to the hurricane primarily consisted of criticism of its response to the approach of the storm and its aftermath, specifically in the delayed response] to the [flooding of New Orleans. ... 80% of the 1.3 million residents of the greater New Orleans metropolitan area evacuated] ...[Over seventy countries pledged monetary donations or other assistance]. ...

Probabilistic latent semantic analysis

$$p(d_i = w | \Phi, \theta_d) = \sum_{z=1}^T \phi_{z,w} \theta_{d,z}$$

$$p(\mathcal{W} | \Phi, \Theta)$$

$$= \prod_{d=1}^D \prod_{d_i=1}^{N_d} \sum_{z=1}^T \phi_{z,w} \theta_{d,z}$$

$$= \prod_{d=1}^D \prod_{w=1}^V \left(\sum_{z=1}^T \phi_{z,w} \theta_{d,z} \right)^{count(d,w)}$$

$$\arg \max_{\Phi, \Theta} [\log p(\mathcal{W} | \Phi, \Theta) + \sum_{d=1}^D \lambda_d (1 - \sum_{z=1}^T \theta_{(d,z)}) + \sum_{z=1}^T \sigma_k (1 - \sum_{w=1}^V \phi_{z,w})]$$

Probabilistic latent semantic analysis

- Use $R_{(w_{d_i}, z)}$ to represent which token d_i in document d comes from (repeated same tokens are from the same topic)

$$\begin{aligned}\mathcal{L} = \log p(\mathcal{W} | \mathbf{R}, \Phi, \Theta) &= \sum_d^D \sum_{d_i}^{N_d} \sum_z^T R_{(w_{d_i}, z)} (\log \phi_{(z, w_{d_i})} + \log \theta_{(d, z)}) \cdot \mathbb{1}[z_{d,i} == z] \\ &\quad + \left(\sum_{d=1}^D \lambda_d \left(1 - \sum_{z=1}^T \theta_{(d, z)} \right) + \sum_{z=1}^T \sigma_k \left(1 - \sum_{w=1}^V \phi_{z, w} \right) \right)\end{aligned}$$

- **M step: set the derivative of L to 0:**

$$\theta_{d,z} \propto \sum_{\substack{v=1 \\ v \neq D}}^V R_{d,v,z} \text{count}(v, d)$$

$$\phi_{z,v} \propto \sum_{d=1}^D \mathbb{1}[z_{d,v} == z] \text{count}(v, d)$$

E step:

$$R_{(w_{d_i}, z)} \propto \phi_{z,v} \theta_{d,z}$$

Probabilistic latent semantic analysis: partially available labels

- Generalized topic modeling:
 - Each document can contain just one topic, e.g., short documents
 - That is, topic inference = topic classification
- If we already know the document tags for a part of the documents, does the partial labels help us make better predictions for the entire corpus? (**Homework 3**)
- Example: news tagging, StackOverflow question tagging

Probabilistic latent semantic analysis: partially available labels

Trends for you · Change

#RAF100

The RAF celebrates 100th anniversary

#TuesdayThoughts

@NWMCblog is Tweeting about this

#ThailandCaveRescue

237K Tweets

Thai Navy Seal

120K Tweets

All 12

All 12 boys and coach rescued from Thai cave

#IgniteB2B

1,850 Tweets

#WildBoars

4,934 Tweets

Science

159K Tweets

George Clooney

George Clooney injured in motorcycle accident in Italy

#NationalPinaColadaDay

1,870 Tweets

T Tagger News tags

1. Intel AMT Checker for Linux (github.com) **Security** **Linux**
116 points by laamalif 4 hours ago | 34 comments
2. Fwaf – Machine Learning Driven Web Application Firewall (fsecurify.com) **AI/Machine Learning** **Security** **Data Science**
46 points by Faizann20 7 hours ago | 9 comments
3. How to Spot a Spook (1974) (cryptome.org) **Politics**
30 points by mercer 3 hours ago | 6 comments
4. LittleTable: A Relational Time-Series Database at Cisco Meraki (acm.org) **Databases**
26 points by rodionos 7 hours ago | 2 comments
5. Wcry ransomware is reborn without its killswitch, starts spreading anew (boingboing.net) **Security**
34 points by rbanffy 4 hours ago | 4 comments
6. Using Deep Learning at Scale in Twitter's Timelines (twitter.com) **AI/Machine Learning**
39 points by hunglee2 8 hours ago | 23 comments
7. Cyberattacks in 12 Nations Said to Use Leaked N.S.A. Hacking Tool (nytimes.com) **Security**
1200 points by ghosh 21 hours ago | 468 comments
8. What it's like to be in the most automated job in the United States (qz.com) **AI/Machine Learning**
28 points by billifuduo 5 hours ago | 17 comments
9. The Right to Read (1997) (gnu.org) **Blockchain**
196 points by tobysullivan 21 hours ago | 62 comments
10. Spends 10 Years Mastering Microsoft Paint to Illustrate His Book (boredpanda.com) **Microsoft**
59 points by prncpinto 8 hours ago | 12 comments
11. Visual Studio 2017 now fully supports Python and Django (visualstudio.com) **Python** **Microsoft**
91 points by vanflymen 14 hours ago | 38 comments
12. Your tl;dr by an ai: a deep reinforced model for abstractive summarization (metamind.io) **AI/Machine Learning**
99 points by etiam 18 hours ago | 18 comments
13. Lessons scaling from 10 to 20 people (josephwalla.com) **Startups**
60 points by gkop 12 hours ago | 15 comments
14. Happy nations don't focus on growth (bloomberg.com) **Politics**
133 points by smollett 21 hours ago | 45 comments
15. Ublock vs. ublock origin (reddit.com) **Web Development**
39 points by based2 4 hours ago | 12 comments
16. Rejection Letter (antipope.org) **Security**
475 points by cstruss 2 hours ago | 63 comments
17. Video Solves Mystery of How Narwhals Use Their Tusks (nationalgeographic.com) **Science**
105 points by clouddrover 8 hours ago | 7 comments
18. Netflix confirms it is blocking rooted/unlocked Android devices (androidpolice.com) **Mobile**
184 points by msq 14 hours ago | 167 comments
19. Fuzzing Irrsi (irssi.org) **Security**
136 points by jbisch 2 hours ago | 13 comments

Homework 3

- Suppose each document has only 1 topic. We have two document set: S1 contains all the tagged documents; S2 contains all the untagged documents. Each tag is a topic, ranging from 1 to T
- (Part 1): derive the EM algorithm that maximizes the probability of the observed document, given the known topics from S1
- (Part 2): implement your EM algorithm, output the predicted topic for each document in S2

PLSA applications

- Topic modeling approach can be used for
 - Interpreting content of corpora
 - Clustering documents, predicting topics
 - Time series/trend analysis

Interpreting content of corpora [Mei et al. 07]

- How do users interpret a learned topic?
 - Human generated labels, but cannot scale up
- What makes a good label?
 - Semantically close (**relevance**)
 - **Understandable** – phrases?
 - High **coverage** inside topic
 - **Discriminative** across topics

government
0.3
response 0.2

...
city 0.2
new 0.1
orleans
0.05

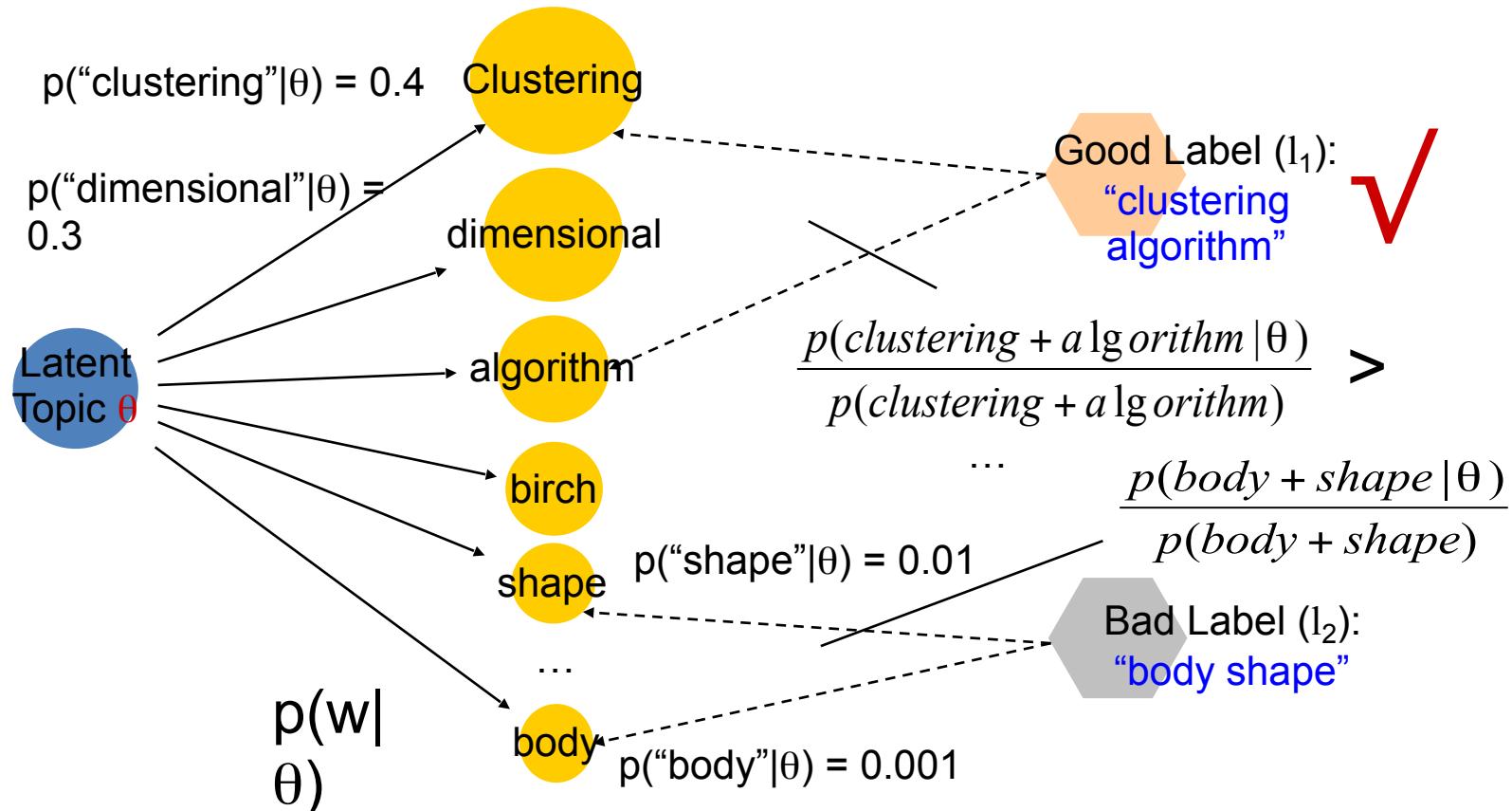
...
donate 0.1
relief 0.05
help 0.02

the 0.04
a 0.03

... 30

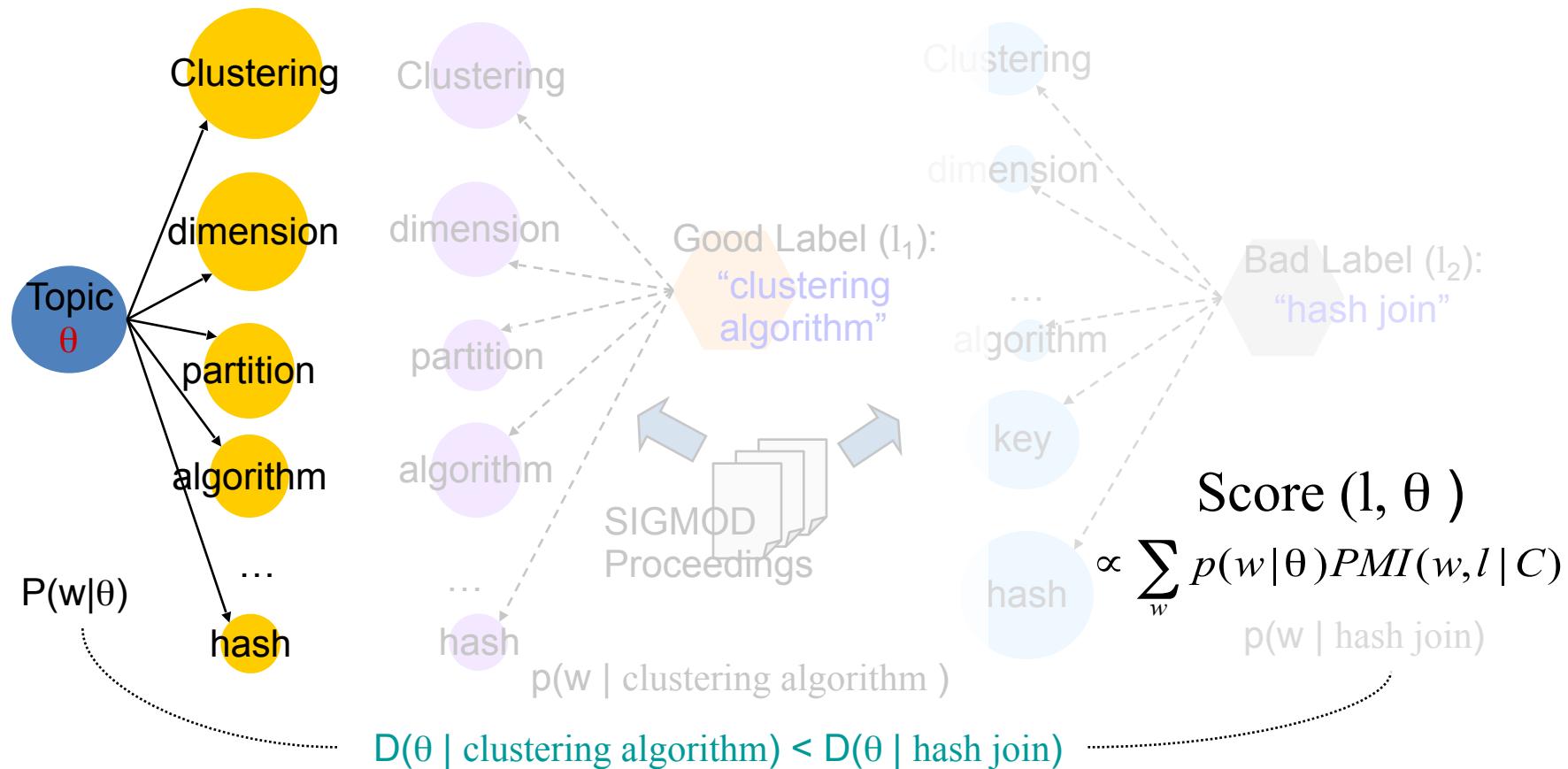
Relevance: the Zero-Order Score [Mei et al. 07]

- Intuition: prefer phrases well covering top words



Relevance: the First-Order Score [Mei et al. 07]

- Intuition: prefer phrases with similar context (distribution)



Topic labels [Mei et al. 07]

sampling 0.06
estimation 0.04
approximate 0.04
histograms 0.03
selectivity 0.03
histogram 0.02
answers 0.02
accurate 0.02

clustering algorithm
clustering structure
...

large data, data
quality, high data,
data application, ...

selectivity
estimation ...

the, of, a, and,
to, data, > 0.02
...
clustering 0.02
time 0.01
clusters 0.01
databases 0.01
large 0.01
performance 0.01
quality 0.005

north 0.02
case 0.01
trial 0.01
iran 0.01
documents 0.01
walsh 0.009
reagan 0.009
charges 0.007

r tree
b tree ...

indexing
methods

tree 0.09
trees 0.08
spatial 0.08
b 0.05
r 0.04
disk 0.02
array 0.01
cache 0.01

iran contra
...

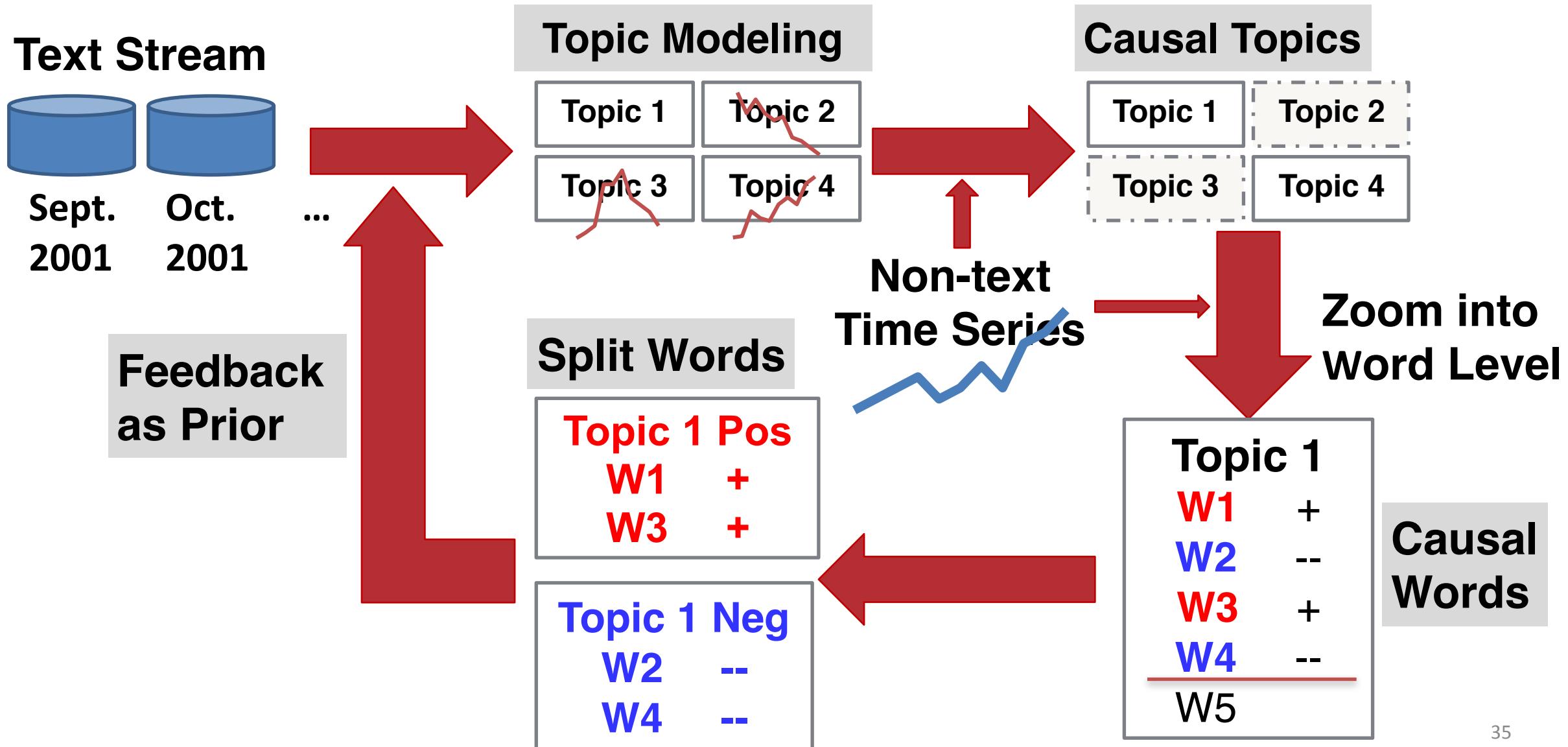
Text mining for understanding time series



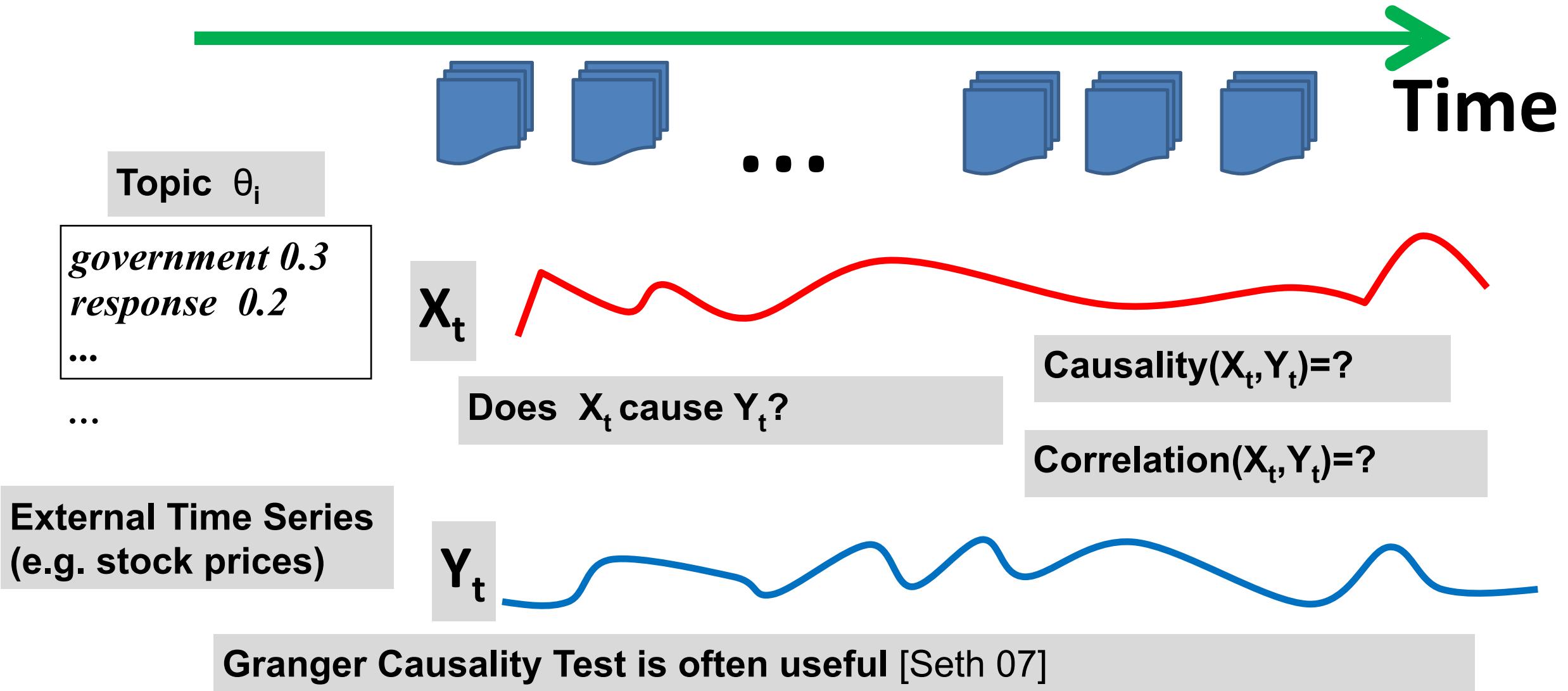
Any clues in the companion news stream?

Dow Jones Industrial Average [Source: Yahoo Finance]

Iterative Causal Topic Modeling [Kim et al. 13]



Measuring Causality (Correlation)



Topics in NY Times Correlated with Stocks

[Kim et al. 13]: June 2000 ~ Dec. 2011

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

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

tax cut 1
screen pataki guiliani
enthusiasm door symbolic
oil energy prices
news w top
pres al vice
love tucker presented
partial abortion privatization
court supreme abortion
gun control 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