Topic Chains for Understanding a News Corpus

to uncover the underlying semantic structure of a sequential corpus of news

Dongwoo Kim and Alice Oh
KAIST, KOREA
CICLING 2011

WORLD »

- Doubts Rise in Rwanda as Election Is Held
- Grenoble Journal: Utopian Dream Becomes Battleground in France
- Netanyahu Testifies on Flotilla Raid

U.S. »

- In Superman's Hometown, a Labor Dispute Over Health
- On Education: Lesson Plan in Boston Schools: Don't Go It Alone
- Plug in Gulf Well Is Declared a Success

POLITICS »

- Colorado Races Test Voters' Anger
- Lamont Moves to Center in Connecticut Race
- Familiar Story in Nevada: Republicans on Offensive

N.Y. / REGION »

- Haitians Look to Family 1,500 Miles North for Help
- Lamont Moves to Center in Connecticut Race
- In Connecticut, a New Level of Intensity for Primaries

SCIENCE »

- Minerals Service Had a Mandate to Produce Results

BUSINESS DAY »

- European Shares Rise as Traders Look to Fed
- BlackBerry Security Stance Sows Anxiety
- Investor Appetite for Bonds in a Tepid Recovery Weighs on Rates

TECHNOLOGY »

- After Drought, Hope for Shows Made for Web
- BlackBerry Security Stance Sows Anxiety
- Docks for Apple Gadgets Help a Business Thrive

SPORTS »

- Yankees 7, Red Sox 2: Putting Ruth and Red Sox in the Rearview Mirror
- Hall of Famer's Slow Road to a Major League Bench
- Woods's Finish Looks Like Rock Bottom

OBITUARIES »

- Patricia Neal, an Oscar Winner Who Endured Tragedy, Dies at 84
- Rabbi Bruce M. Cohen, Is Dead at 65; Worked to Promote Peace
- Tony Judt, Chronicler of History, Is Dead at 62

TRAVEL »

OPINION »

- Editorial: As the Economy Slows
- Letters: Excess Radiation From CT Scans
- Op-Ed Columnist: America Goes Dark

ARTS »

- The Hand of a Master Architect
- Patricia Neal, an Oscar Winner Who Endured Tragedy, Dies at 84
- Debt Problem Has Museum Scrambling

MOVIES »

- Film: Start Poor, Spread 'Glee,' Then Try 'Eat Pray Love'
- Film: Cult Director Courts the Mass, Keeps the Crazy
- A Go-to Actor for 'That Guy' Roles

THEATER »

- London Theatergoers Have Front-Row Seats at End of the World
- Theater Review | 'Wolves': After Peter and the Wolf Comes Everyone and the Wolf
- Theater Review | 'Tales From the Tunnel': Odors and Oddities of the Underground

TIMES WIRE »

Most recent updates on NYTimes.com. See More »

35 minutes ago App Smart Extra: Restricted Diets

39 minutes ago The Early Word: Trailblazer to Fund-raiser

44 minutes ago Nabors to Buy Driller Superior Wells for $900 Million

MOST POPULAR

E-MAILED | BLOGGED | SEARCHED | VIEWED

1. But Will It Make You Happy?
2. Op-Ed Contributor: Congregations Gone Wild
3. Thomas L. Friedman: Steal This Movie
4. My Life in Therapy
5. Paul Krugman: America Goes Dark
6. Bucks: How to Find Cheaper College Textbooks
7. Across Nation, Mosque Projects Meet Opposition
8. Op-Ed Contributor: This Bedbug's Life
9. Frank Rich: How to Lose an Election Without Really Trying
10. 36 Hours in Boston

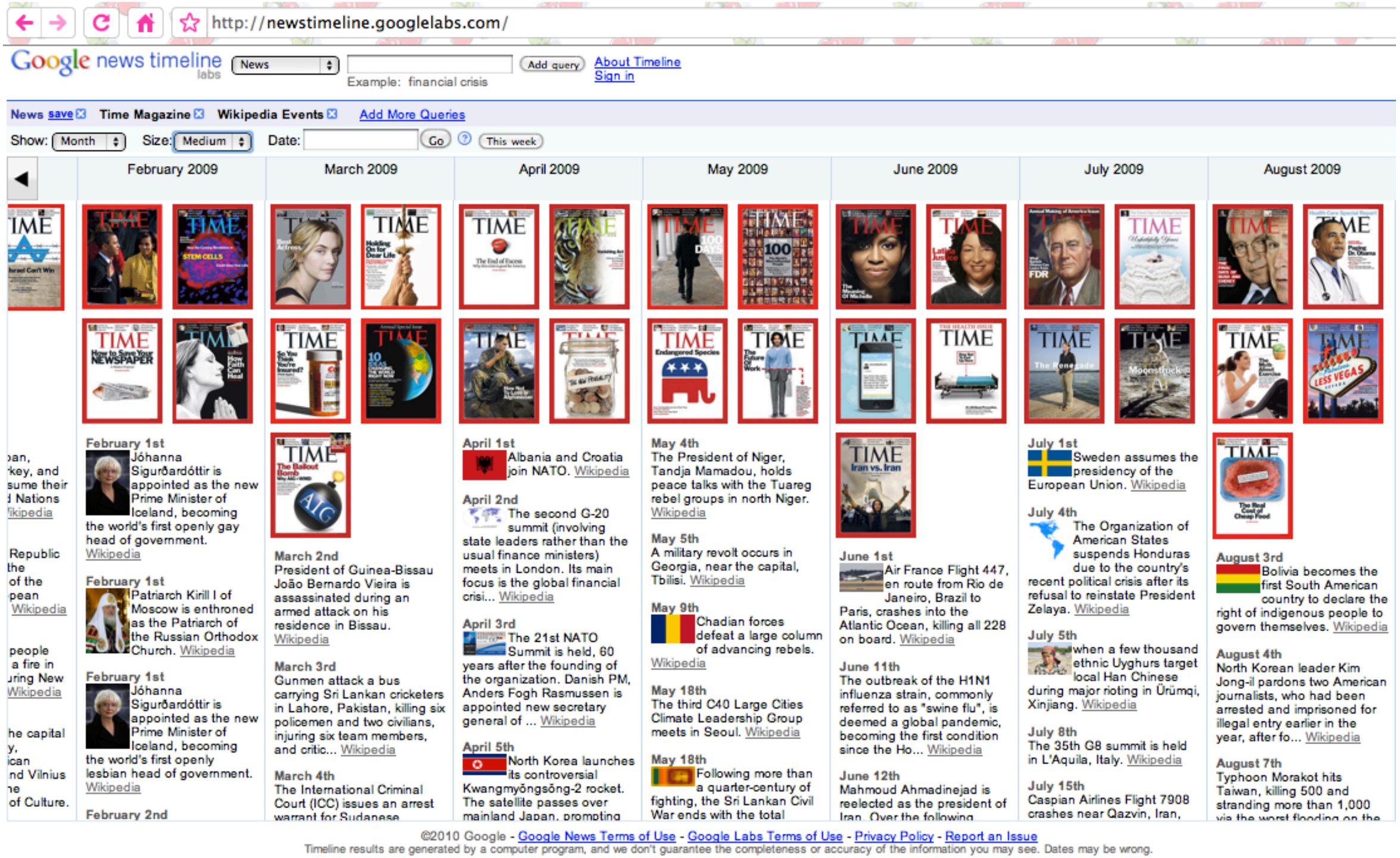
Go to Complete List »

CUSTOMIZE HEADLINES

Create a personalized list of headlines based on your interests. Login » or Register »

News at a glance

a product of hard-working editors

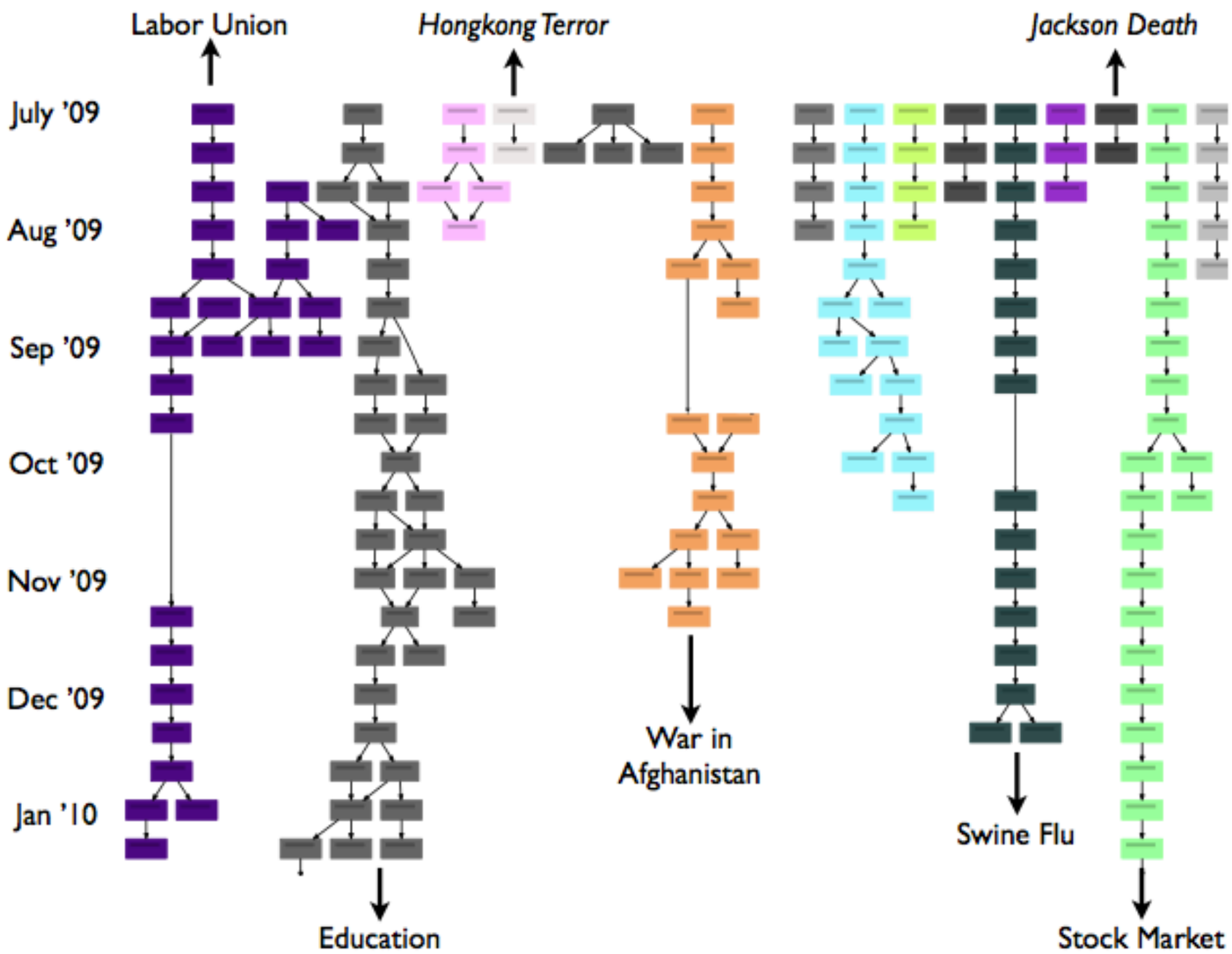


News at a glance

now this is cool

Unresolved Questions

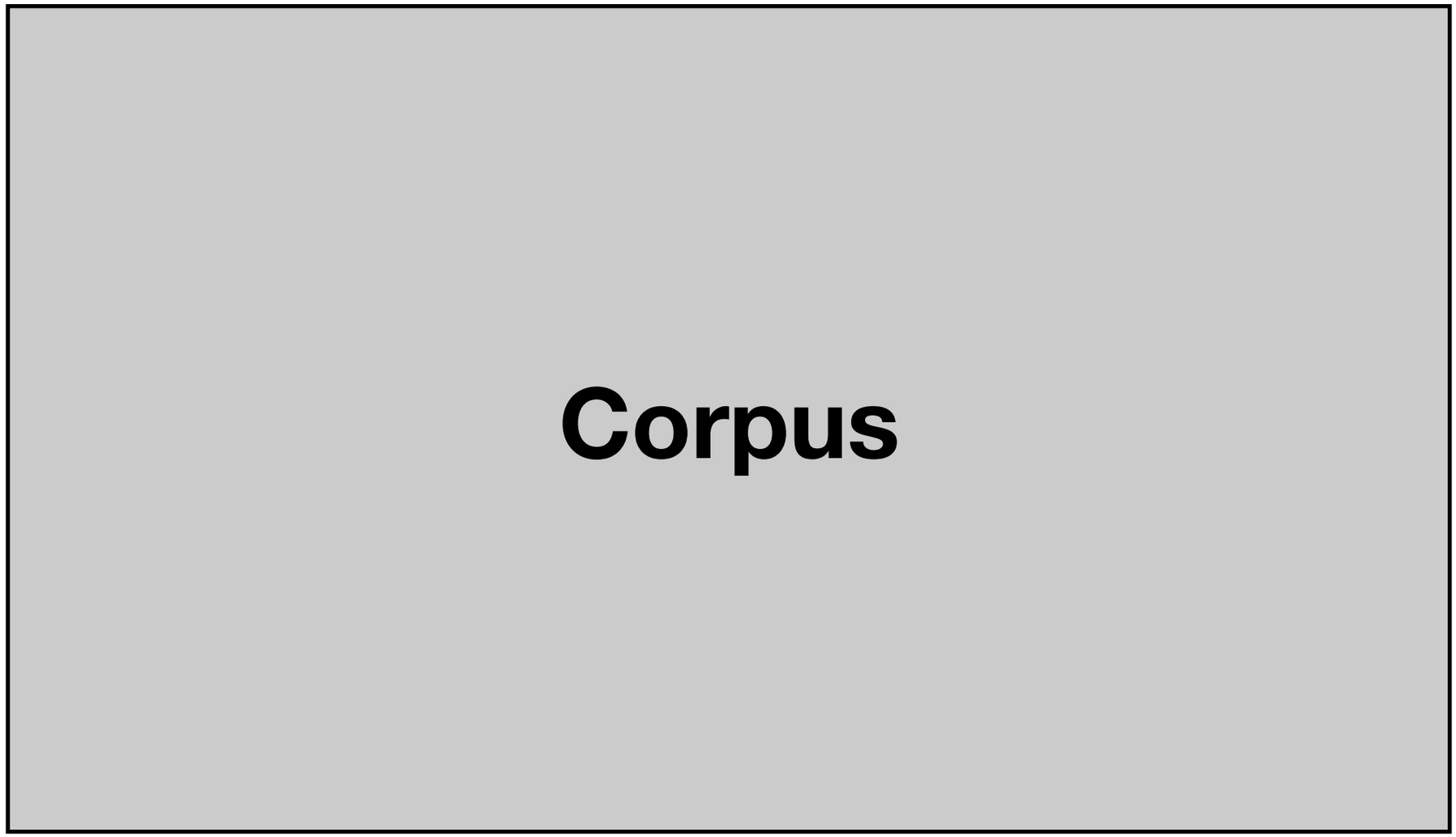
- For a given article, were there similar articles following it?
- If there were similar articles talking about the same topic, how long did that topic last in the news?
- Was that part of a general perpetual topic, such as the US economy?
- Or was it part of a temporary issue, such as the death of a famous person?
- If it is part of a general topic or a long-running topic, how did the focus of the topic change over time?



News at a glance

a bird's eye view

Plan to Solve the Problem (1)

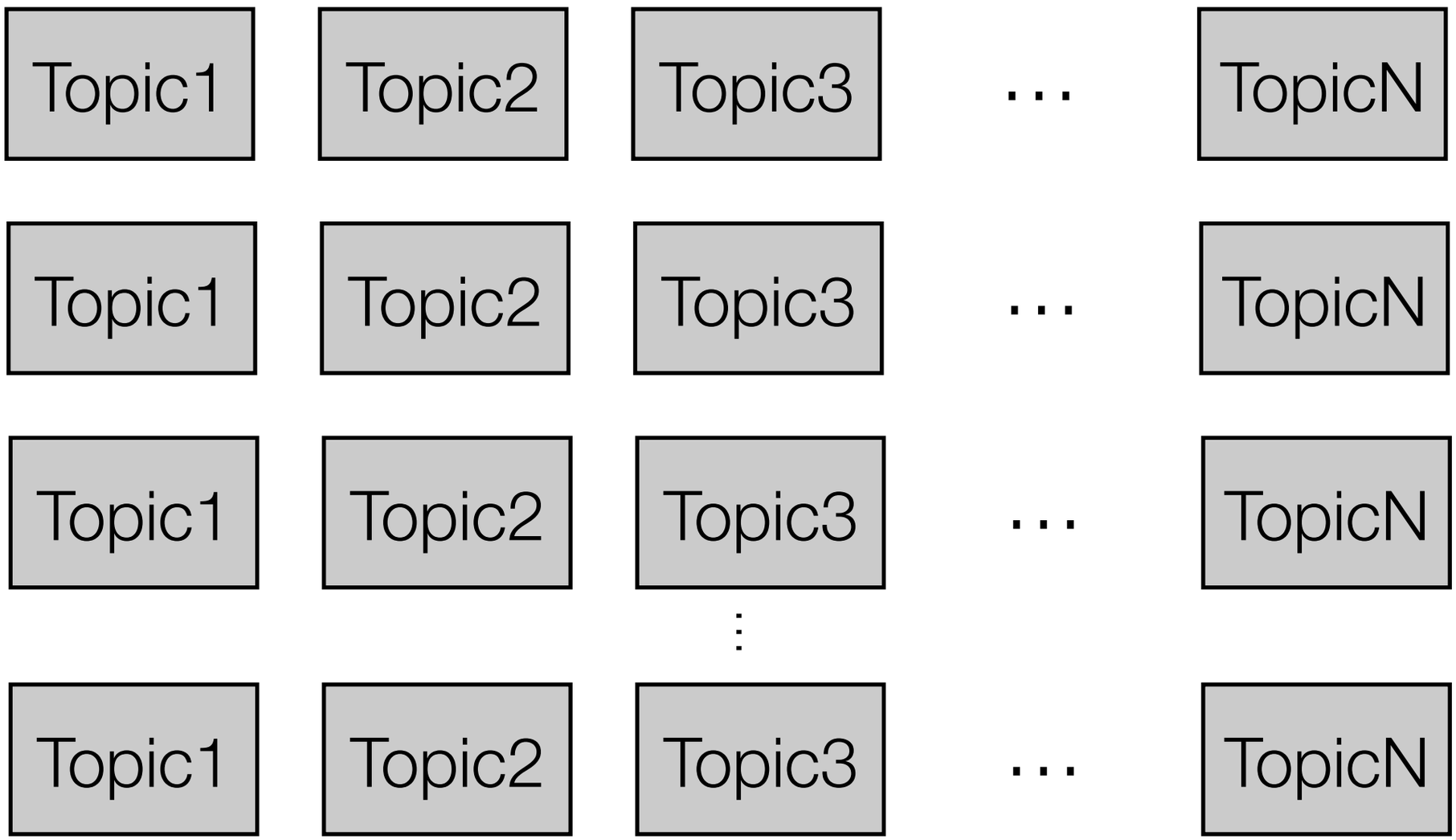


1. Collect the sequential articles

Plan to Solve the Problem (2)

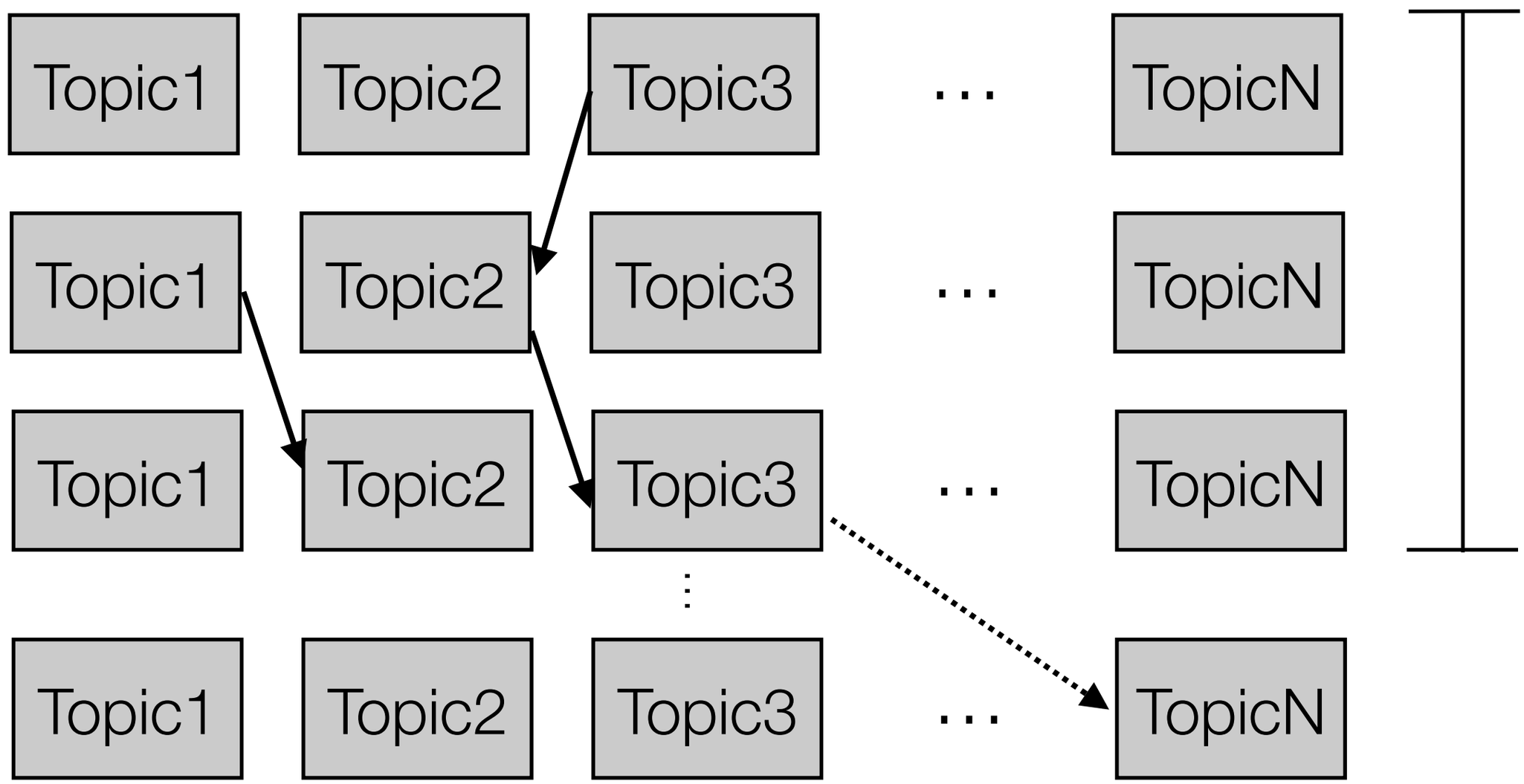
2. Divide the articles into M time slices

Plan to Solve the Problem (3)



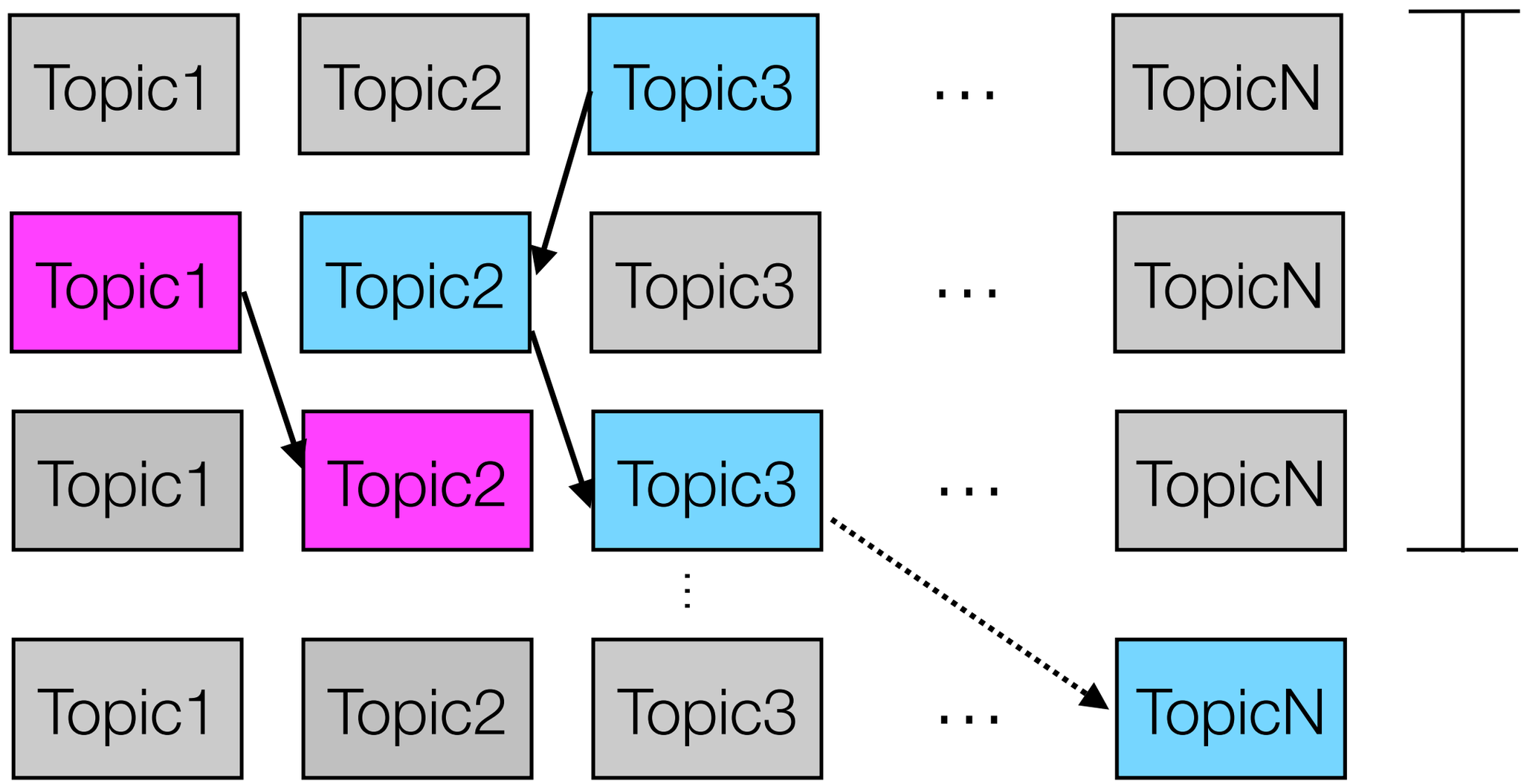
3. Find topics for each time slice using LDA

Plan to Solve the Problem (4)



4. Look for similar topics within the neighboring time slices and Connect similar topics to construct topic chains

Plan to Solve the Problem (5)



5. Identify long-term general topics and short-term temporary issues

Corpus: nine months of Korean news articles

from websites of three major newspapers

130,000+	articles
4,700+	articles per time slice
140,000+	unique words
43,000+	named entities
50	topics per time slice
1,400	topics total

Soccer	Business	Smart phones	Academia
game	growth	apple	research
player	business	smartphone	professor
league	recovery	internet	science
coach	crisis	iphone	doctorate
soccer	prospect	mobile phone	discovery
season	policy	google	analysis
leader	investment	computer	technology
competition	strategy	usage	universe
advance	market	advertise	plant
pro	consume	information	experiment

Finding Topics Using LDA

showing 4 of the 1,400 topics found

Measuring Similarity Between Two Topics

nascar	0.12	spending	0.09	sports	0.12
races	0.10	economic	0.07	team	0.11
cars	0.10	recession	0.06	game	0.10
racing	0.09	save	0.05	player	0.10
track	0.08	money	0.05	athlete	0.09
speed	0.06	cut	0.04	win	0.07

- A topic is a multinomial distribution over words

  KL divergence; JS divergence

- A topic is a vector, where each dimension is a probability of the word in the topic

  cosine similarity

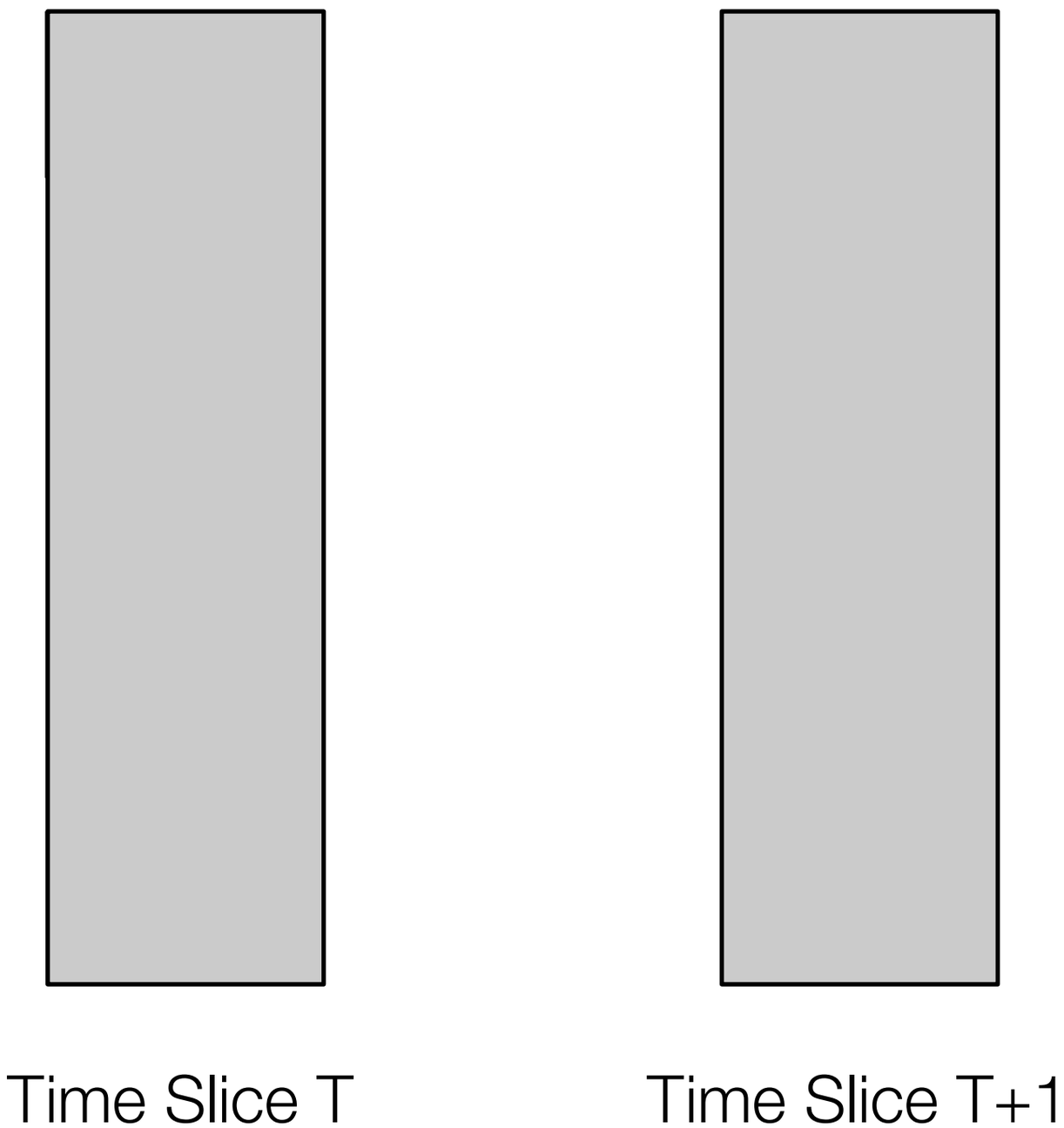
- A topic is a ranked list of words

  Kendall's Tau; DCG

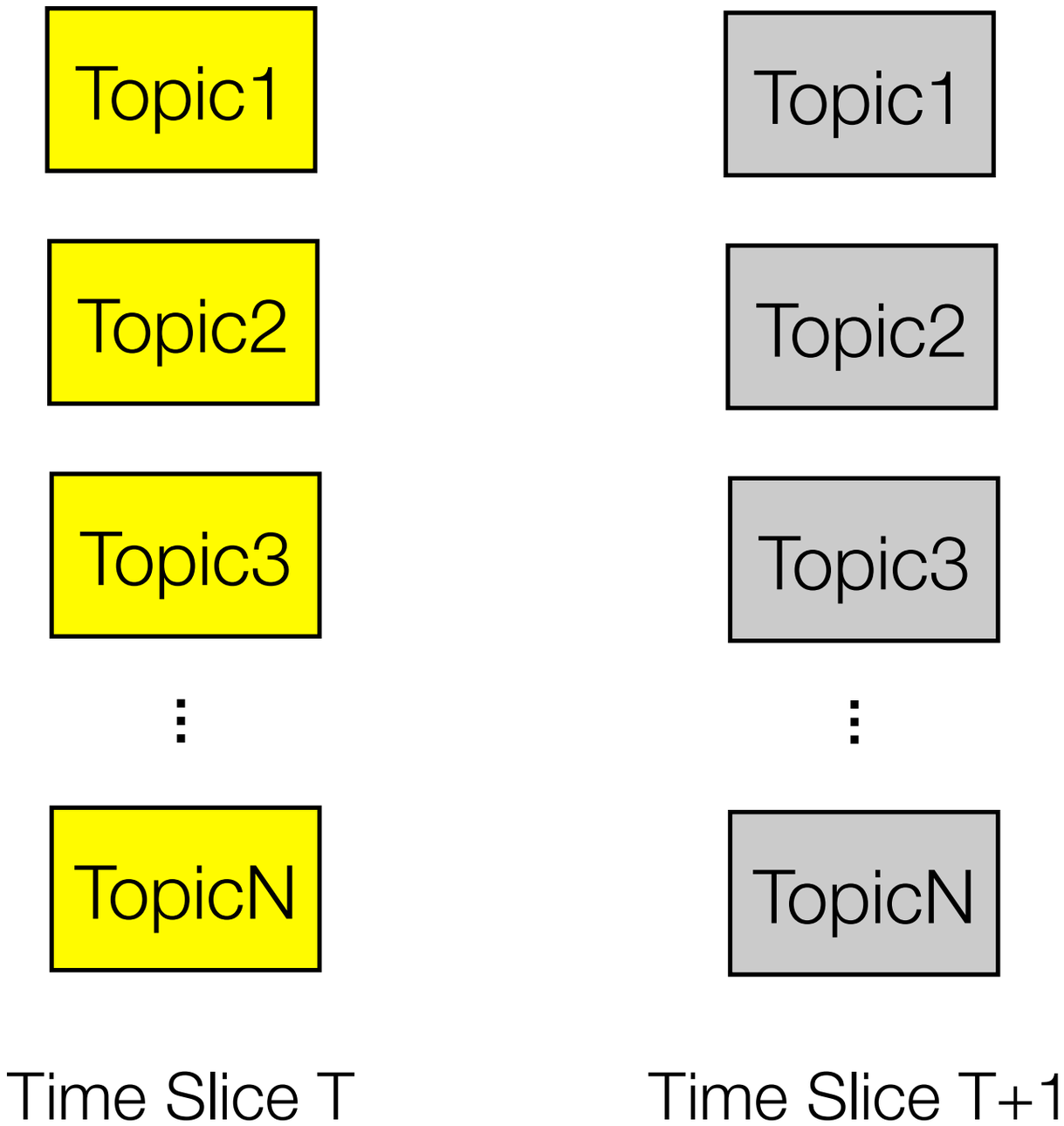
- A topic is a set of top-probability words

  Jaccard's coefficient

Finding Best Similarity Metric (1)

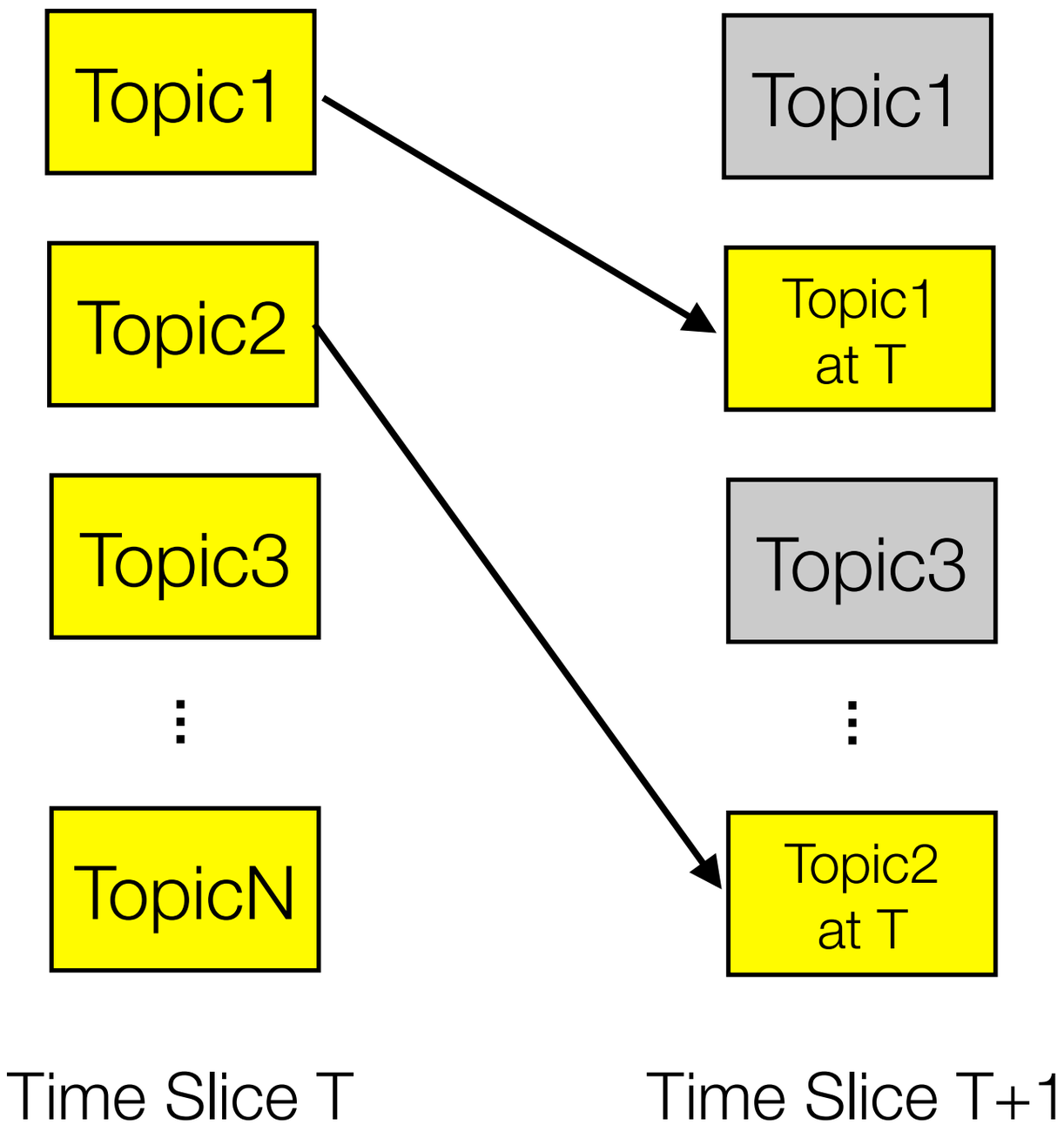


Finding Best Similarity Metric (2)



1. Calculate similarity between all pair of topics

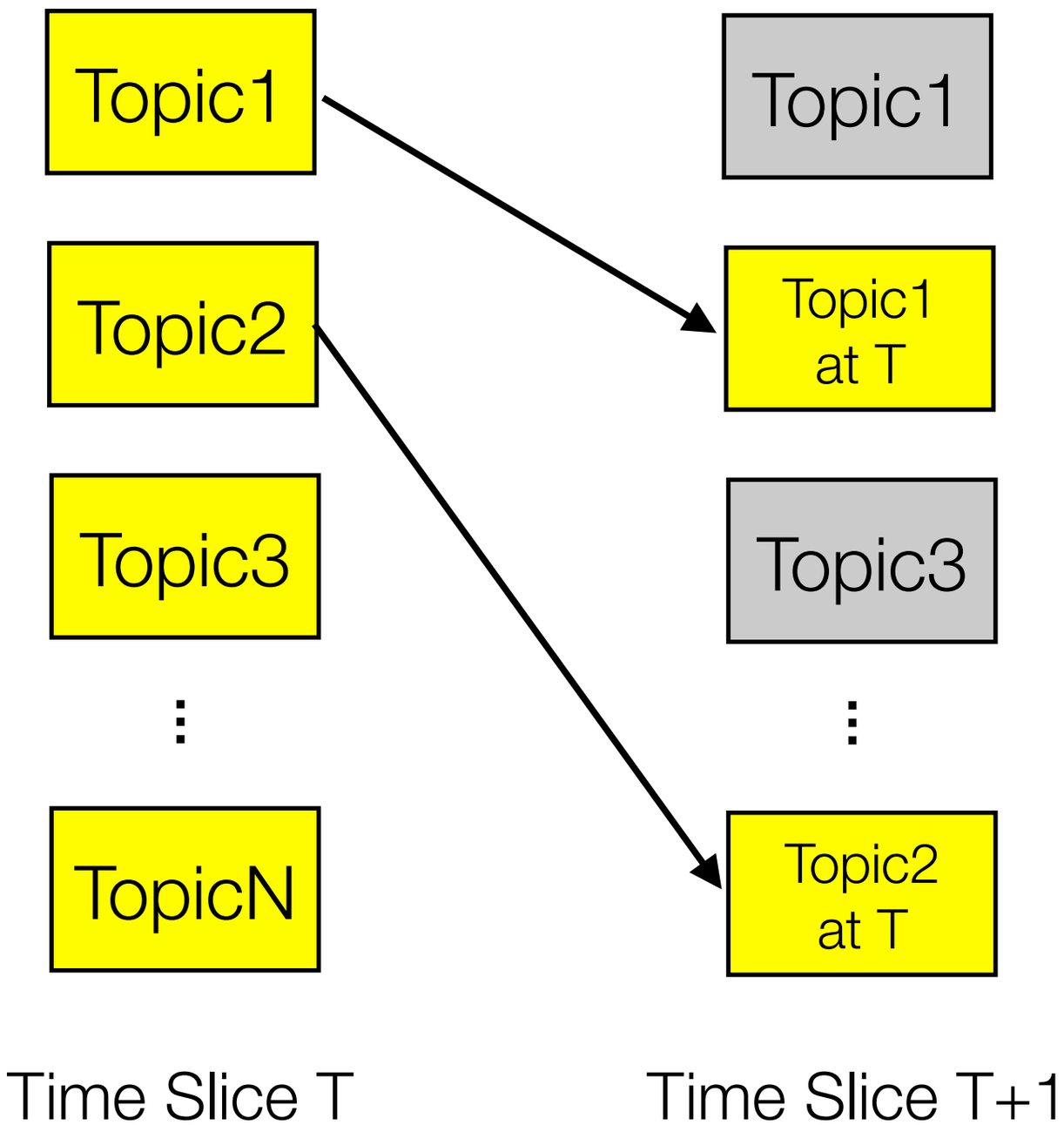
Finding Best Similarity Metric (3)



1. Calculate similarity between all pair of topics

2. Replace the most similar 5 topics from previous time slice

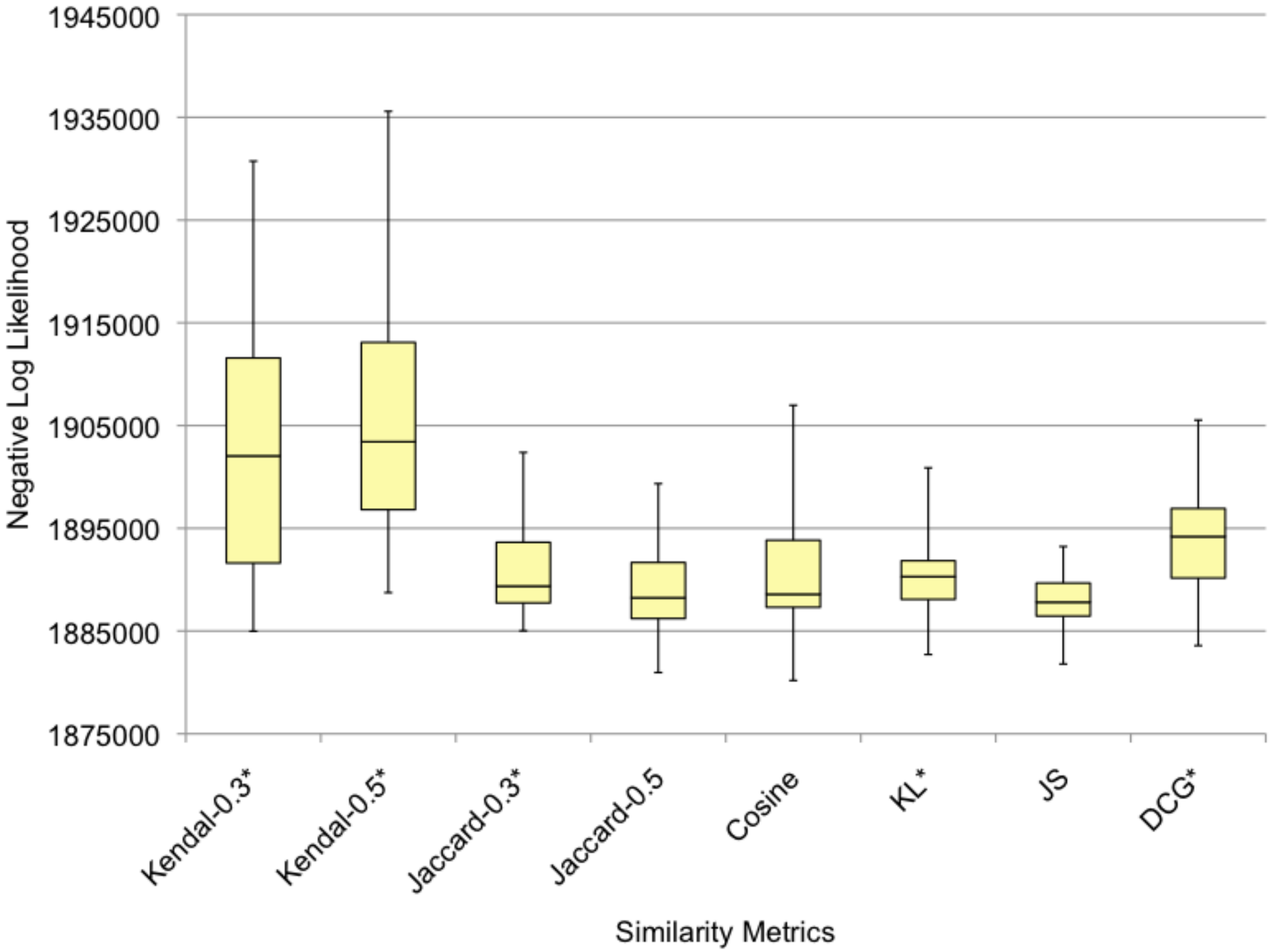
Finding Best Similarity Metric (4)



1. Calculate similarity between all pair of topics

2. Replace the most similar 5 topics from previous time slice

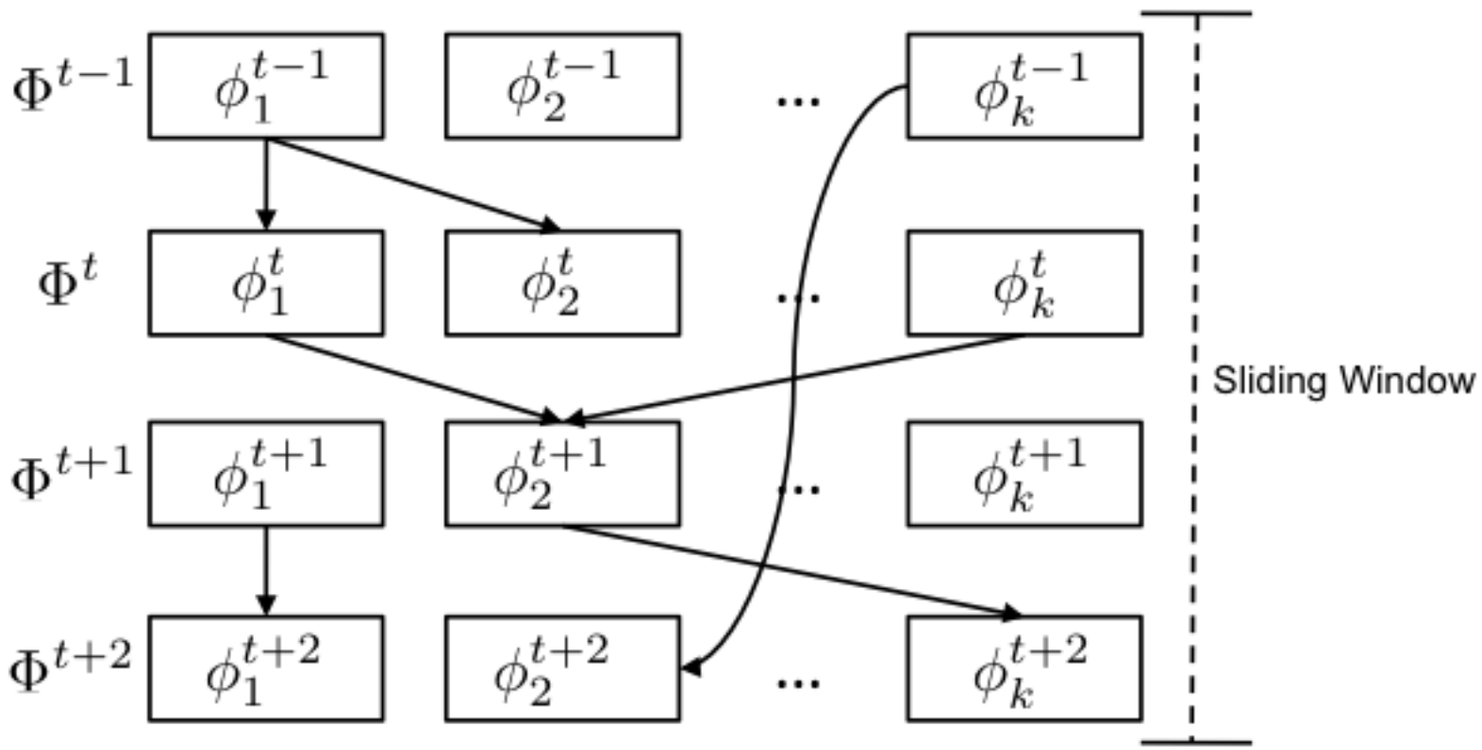
3. Calculate the likelihood of Time Slice T+1



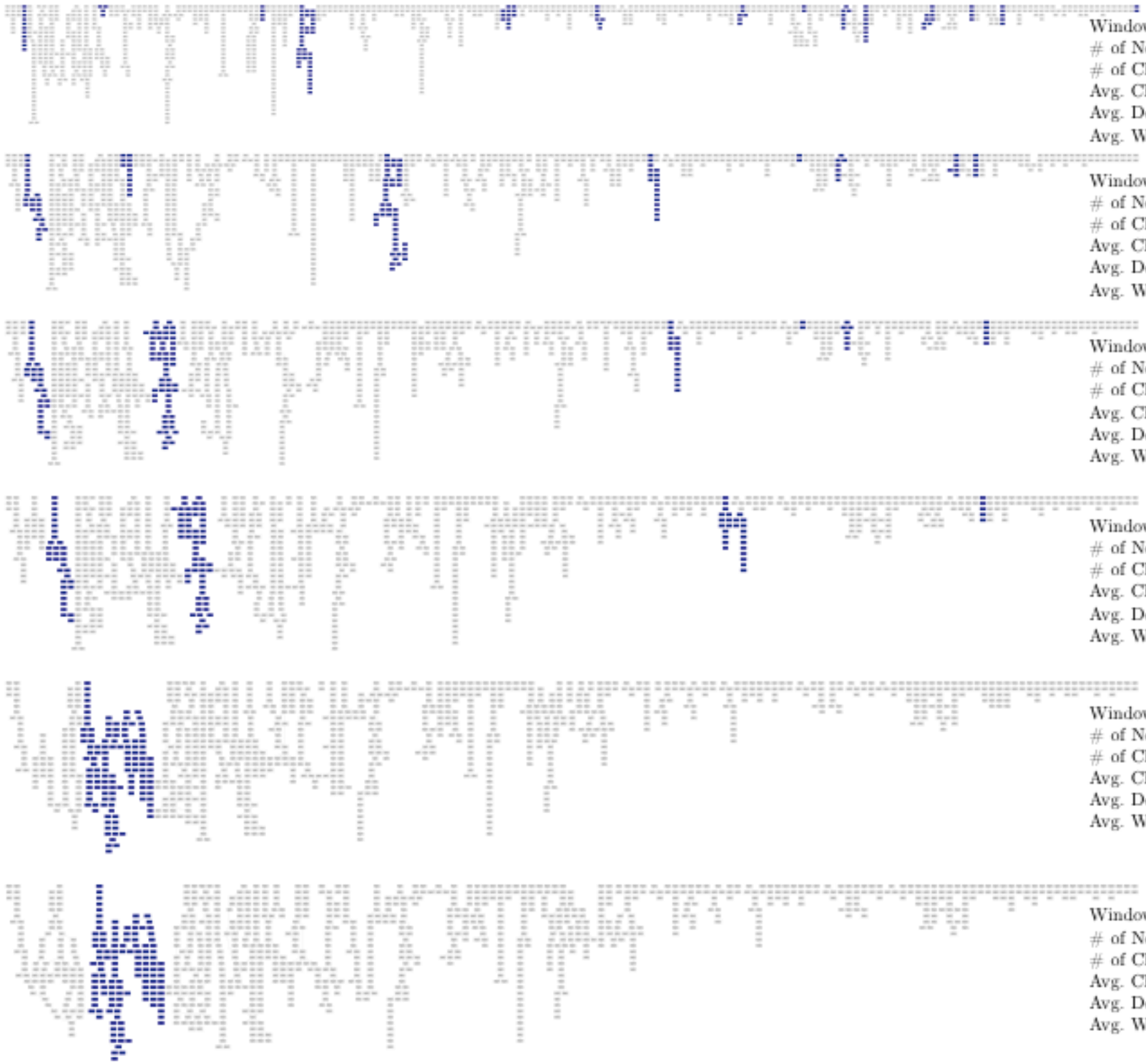
Comparison of frequently used similarity metrics

KL divergence is not the best

Deciding the Range of Neighboring Time Slices

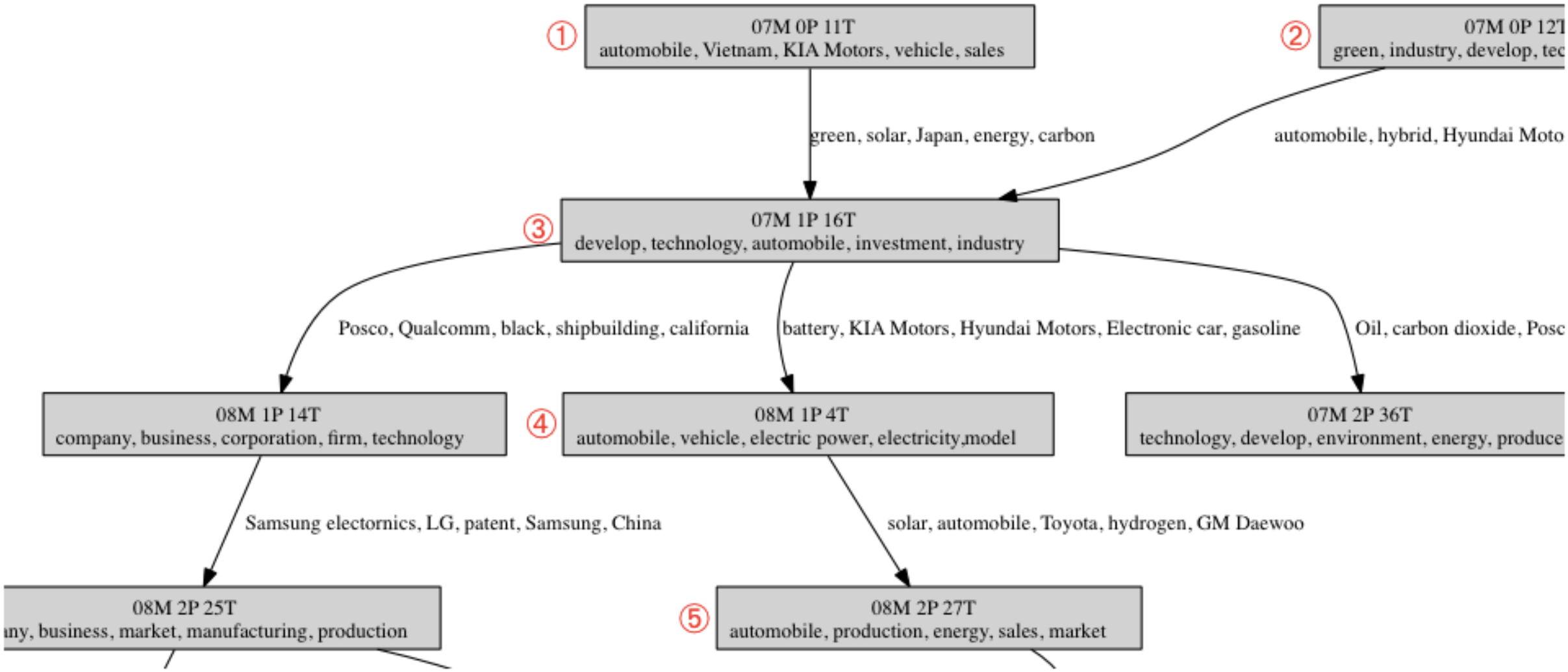


- Looking back one time slice -- might miss some chains that can be formed between two non-consecutive topics



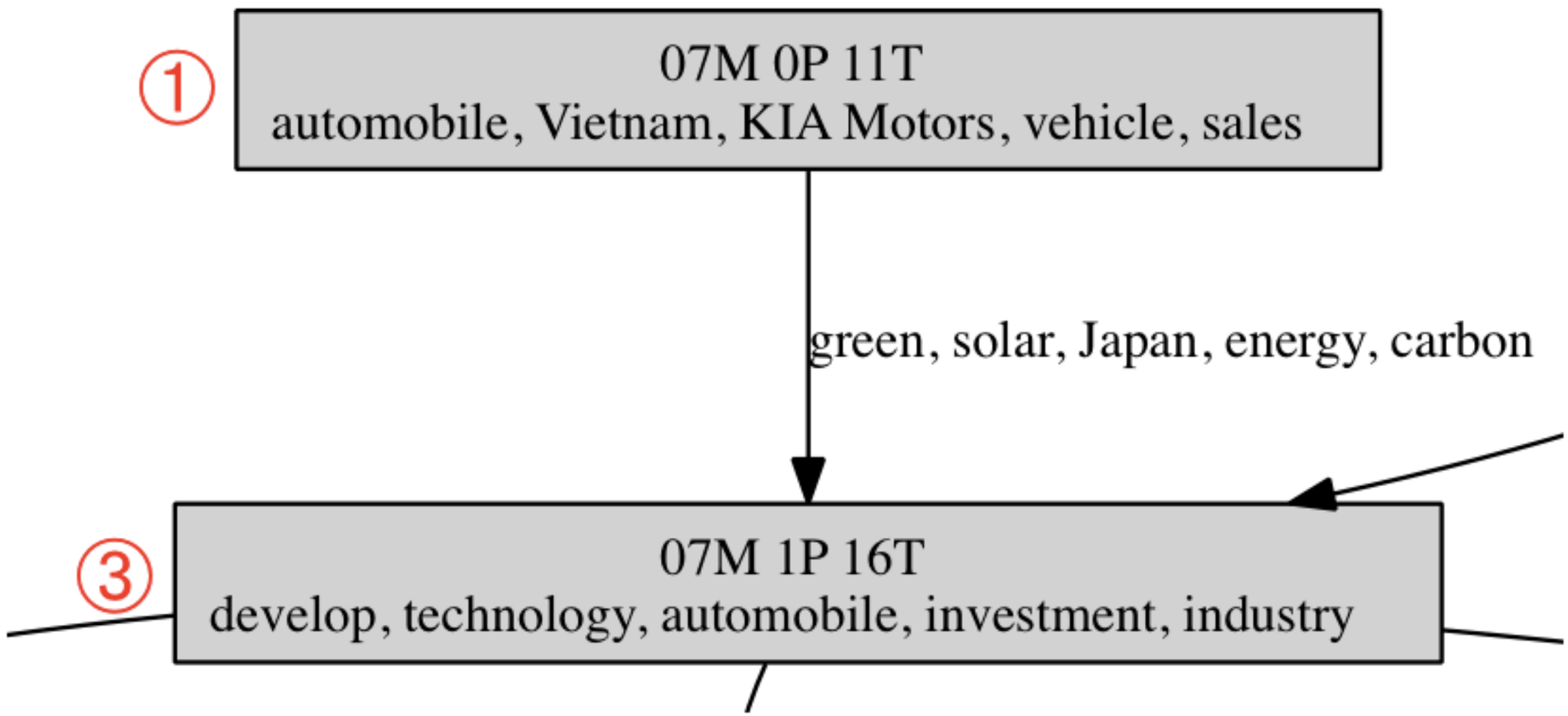
A larger sliding window produces larger chains

and they tend to be more abstract



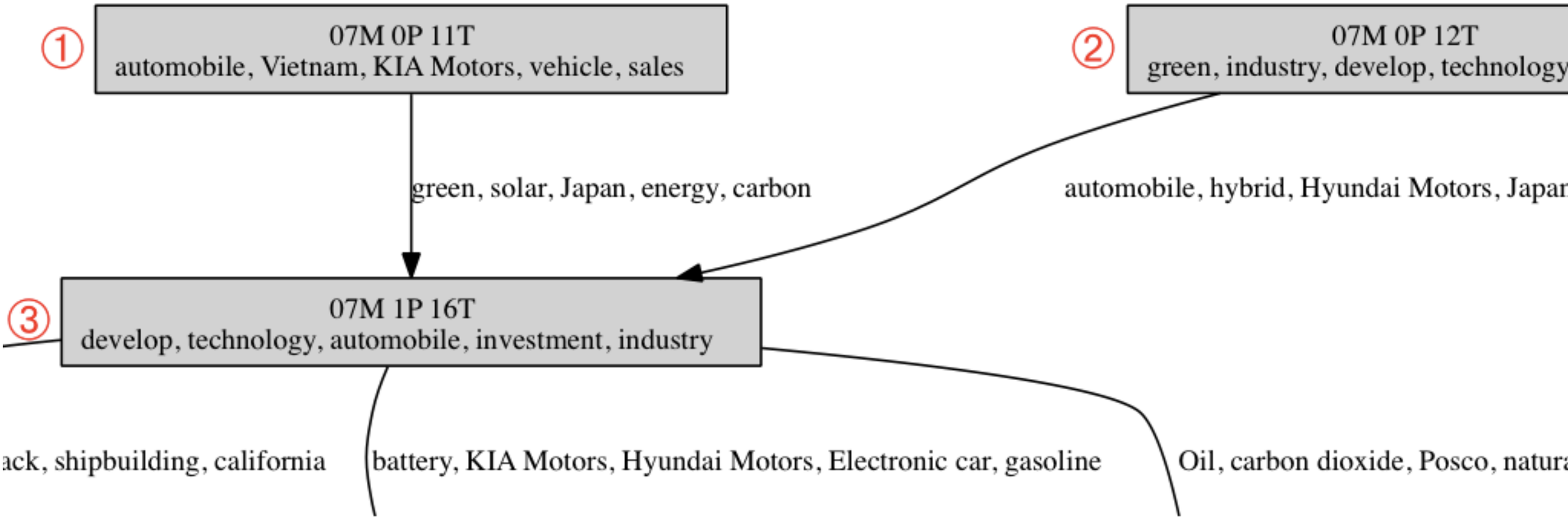
How the focus shifts in a topic chain

is shown by named entities with large probability changes



How the focus shifts in a topic chain

is shown by named entities with large probability changes



How the focus shifts in a topic chain

is shown by named entities with large probability changes

0P 07M 2009Y	North Korea, missile, launch, range, UN Security Council, ship, navy, East sea, ballistic missile
0P 07M 2009Y	Jackson, family, funeral, cherish the memory of, Michael Jackson, son, LA, publish, report, death
0P 10M 2009Y	melamine, dry milk, region, environment, investigation, food, pollution, mercury, produce, management
2P 12M 2009Y	flight, airport, passenger, airplane, search, terror, time, security, explosion, aircraft
0P 01M 2010Y	Hyesoo Kim, actor, 2010, ski, Haejin Ryu, once, 4, soul, colleague, lover
0P 04M 2010Y	tree, recover, park, culture, movement, development, environment, ecology, forest, designation
0P 02M 2010Y	Obama, Republicans, Jeju island, game, Jeju, golf, White house, Woods, gamers, budget

North Korea missile launch
death of Michael Jackson
melamine in milk scandal
heightened airport security at year-end
romance of a famous actor & actress
Arbor day

Short topic chains

represent temporal issues
or incoherent topics

Topic chains: a framework based on LDA

to uncover the underlying semantic structure of a sequential corpus of news