

# 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

















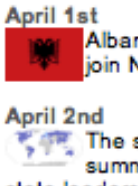




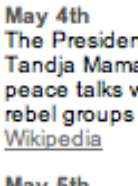



















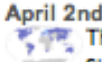

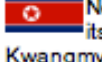

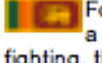





← → ↺ ↻ ☆ http://newstimeline.googlelabs.com/

Google news timeline labs News  Add query About Timeline Sign in

Example: financial crisis

News ☒ Time Magazine ☒ Wikipedia Events ☒ Add More Queries

Show:  Size:  Date:

|  | February 2009   | March 2009   | April 2009   | May 2009  | June 2009  | July 2009   | August 2009   |
|--|---|--|--|---|--|---|---|
|  | <br><br><br><br>   | <br><br><br><br>   | <br><br><br><br>  | <br><br><br><br>  | <br><br><br><br>   | <br><br><br><br>  | <br><br><br><br>  |
| <p>pan, key, and some their Nations <a href="#">Wikipedia</a></p> <p>Republic the of the pean <a href="#">Wikipedia</a></p> <p>people a fire in during New <a href="#">Wikipedia</a></p> <p>he capital y, can and Vilnius re of Culture.</p> | <p><b>February 1st</b><br/>  Jóhanna Sigurðardóttir is appointed as the new Prime Minister of Iceland, becoming the world's first openly gay head of government. <a href="#">Wikipedia</a></p> <p><b>February 1st</b><br/>  Patriarch Kirill I of Moscow is enthroned as the Patriarch of the Russian Orthodox Church. <a href="#">Wikipedia</a></p> <p><b>February 1st</b><br/>  Jóhanna Sigurðardóttir is appointed as the new Prime Minister of Iceland, becoming the world's first openly lesbian head of government. <a href="#">Wikipedia</a></p> <p><b>February 2nd</b></p> | <p><b>March 2nd</b><br/> President of Guinea-Bissau João Bernardo Vieira is assassinated during an armed attack on his residence in Bissau. <a href="#">Wikipedia</a></p> <p><b>March 3rd</b><br/> Gunmen attack a bus carrying Sri Lankan cricketers in Lahore, Pakistan, killing six policemen and two civilians, injuring six team members, and critic... <a href="#">Wikipedia</a></p> <p><b>March 4th</b><br/> The International Criminal Court (ICC) issues an arrest warrant for Sudanese</p> | <p><b>April 1st</b><br/>  Albania and Croatia join NATO. <a href="#">Wikipedia</a></p> <p><b>April 2nd</b><br/>  The second G-20 summit (involving state leaders rather than the usual finance ministers) meets in London. Its main focus is the global financial crisi... <a href="#">Wikipedia</a></p> <p><b>April 3rd</b><br/>  The 21st NATO Summit is held, 60 years after the founding of the organization. Danish PM, Anders Fogh Rasmussen is appointed new secretary general of ... <a href="#">Wikipedia</a></p> <p><b>April 5th</b><br/>  North Korea launches its controversial Kwangmyŏngsŏng-2 rocket. The satellite passes over mainland Japan, promoting</p> | <p><b>May 4th</b><br/> The President of Niger, Tandja Mamadou, holds peace talks with the Tuareg rebel groups in north Niger. <a href="#">Wikipedia</a></p> <p><b>May 5th</b><br/> A military revolt occurs in Georgia, near the capital, Tbilisi. <a href="#">Wikipedia</a></p> <p><b>May 9th</b><br/>  Chadian forces defeat a large column of advancing rebels. <a href="#">Wikipedia</a></p> <p><b>May 18th</b><br/> The third C40 Large Cities Climate Leadership Group meets in Seoul. <a href="#">Wikipedia</a></p> <p><b>May 18th</b><br/>  Following more than a quarter-century of fighting, the Sri Lankan Civil War ends with the total</p> | <p><b>June 1st</b><br/>  Air France Flight 447, en route from Rio de Janeiro, Brazil to Paris, crashes into the Atlantic Ocean, killing all 228 on board. <a href="#">Wikipedia</a></p> <p><b>June 11th</b><br/> The outbreak of the H1N1 influenza strain, commonly referred to as "swine flu", is deemed a global pandemic, becoming the first condition since the Ho... <a href="#">Wikipedia</a></p> <p><b>June 12th</b><br/> Mahmoud Ahmadinejad is reelected as the president of Iran. Over the following</p> | <p><b>July 1st</b><br/>  Sweden assumes the presidency of the European Union. <a href="#">Wikipedia</a></p> <p><b>July 4th</b><br/>  The Organization of American States suspends Honduras due to the country's recent political crisis after its refusal to reinstate President Zelaya. <a href="#">Wikipedia</a></p> <p><b>July 5th</b><br/>  when a few thousand ethnic Uyghurs target local Han Chinese during major rioting in Ürümqi, Xinjiang. <a href="#">Wikipedia</a></p> <p><b>July 8th</b><br/> The 35th G8 summit is held in L'Aquila, Italy. <a href="#">Wikipedia</a></p> <p><b>July 15th</b><br/> Caspian Airlines Flight 7908 crashes near Qazvin, Iran,</p> | <p><b>August 3rd</b><br/>  Bolivia becomes the first South American country to declare the right of indigenous people to govern themselves. <a href="#">Wikipedia</a></p> <p><b>August 4th</b><br/> North Korean leader Kim Jong-il pardons two American journalists, who had been arrested and imprisoned for illegal entry earlier in the year, after fo... <a href="#">Wikipedia</a></p> <p><b>August 7th</b><br/> Typhoon Morakot hits Taiwan, killing 500 and stranding more than 1,000 via the worst flooding on the</p> |

©2010 Google - [Google News Terms of Use](#) - [Google Labs Terms of Use](#) - [Privacy Policy](#) - [Report an Issue](#)

Timeline results are generated by a computer program, and we don't guarantee the completeness or accuracy of the information you may see. Dates may be wrong.

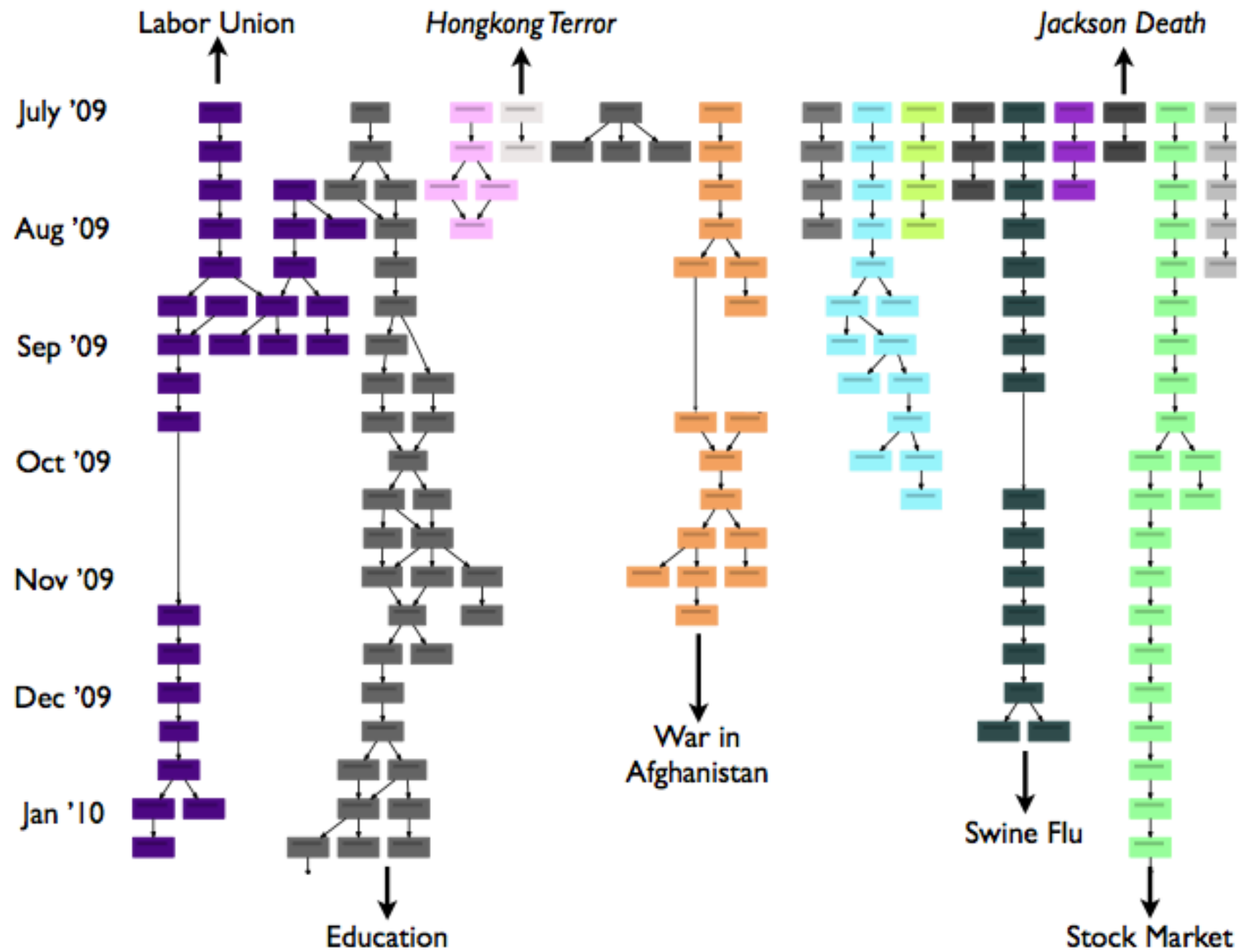
News at a glance

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)

---

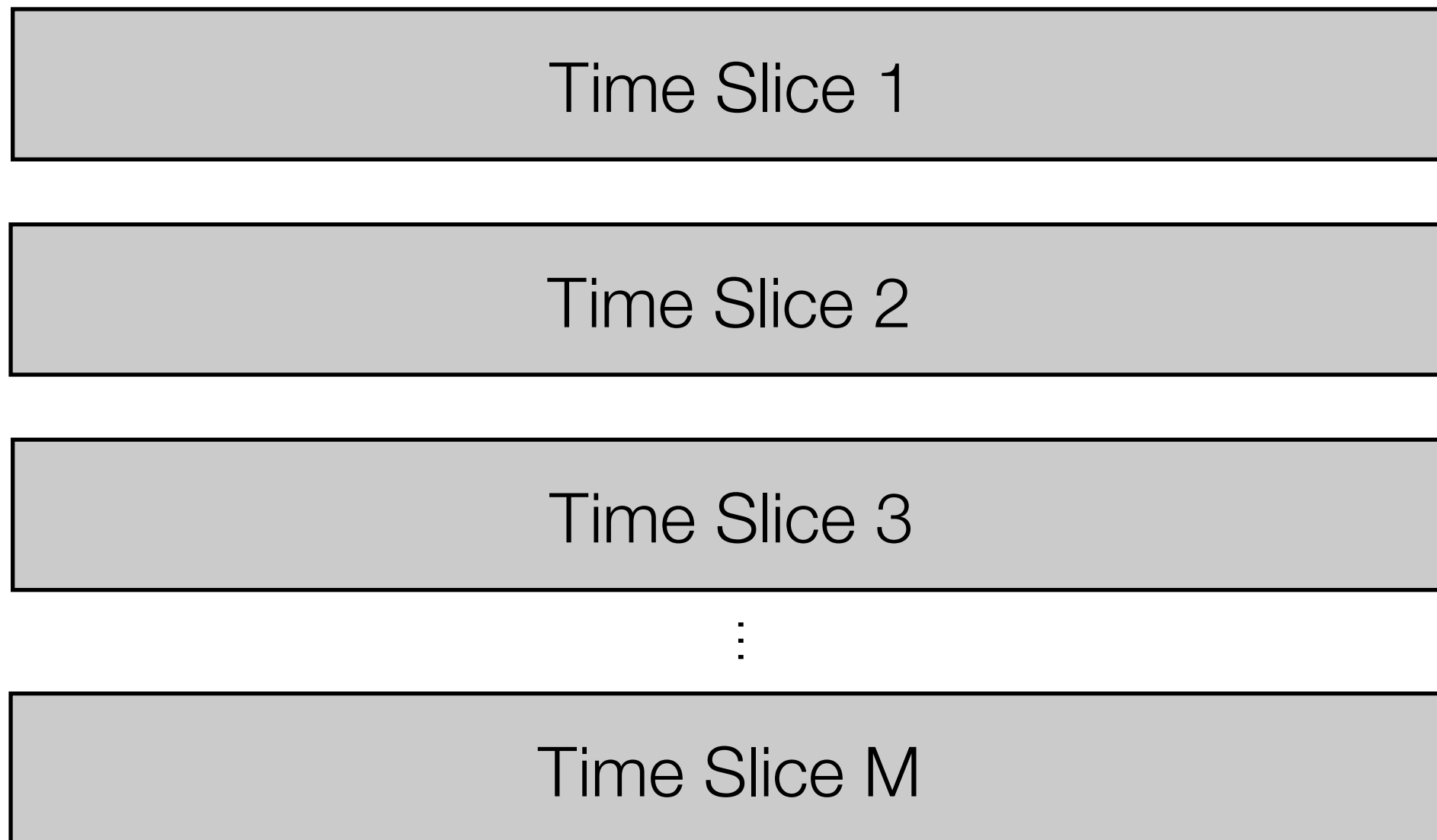


**Corpus**

1. Collect the sequential articles

## Plan to Solve the Problem (2)

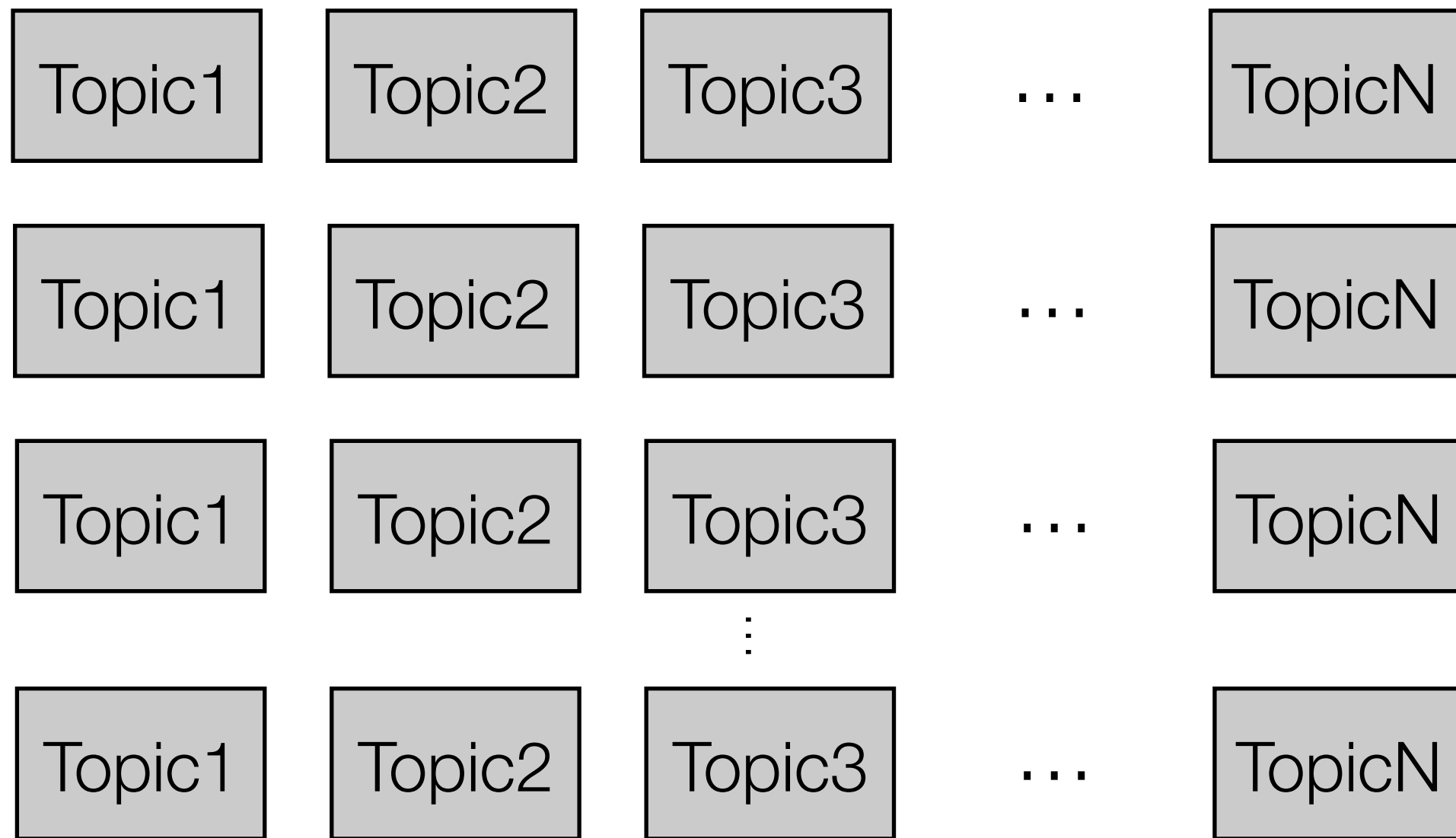
---



2. Divide the articles into M time slices

## Plan to Solve the Problem (3)

---

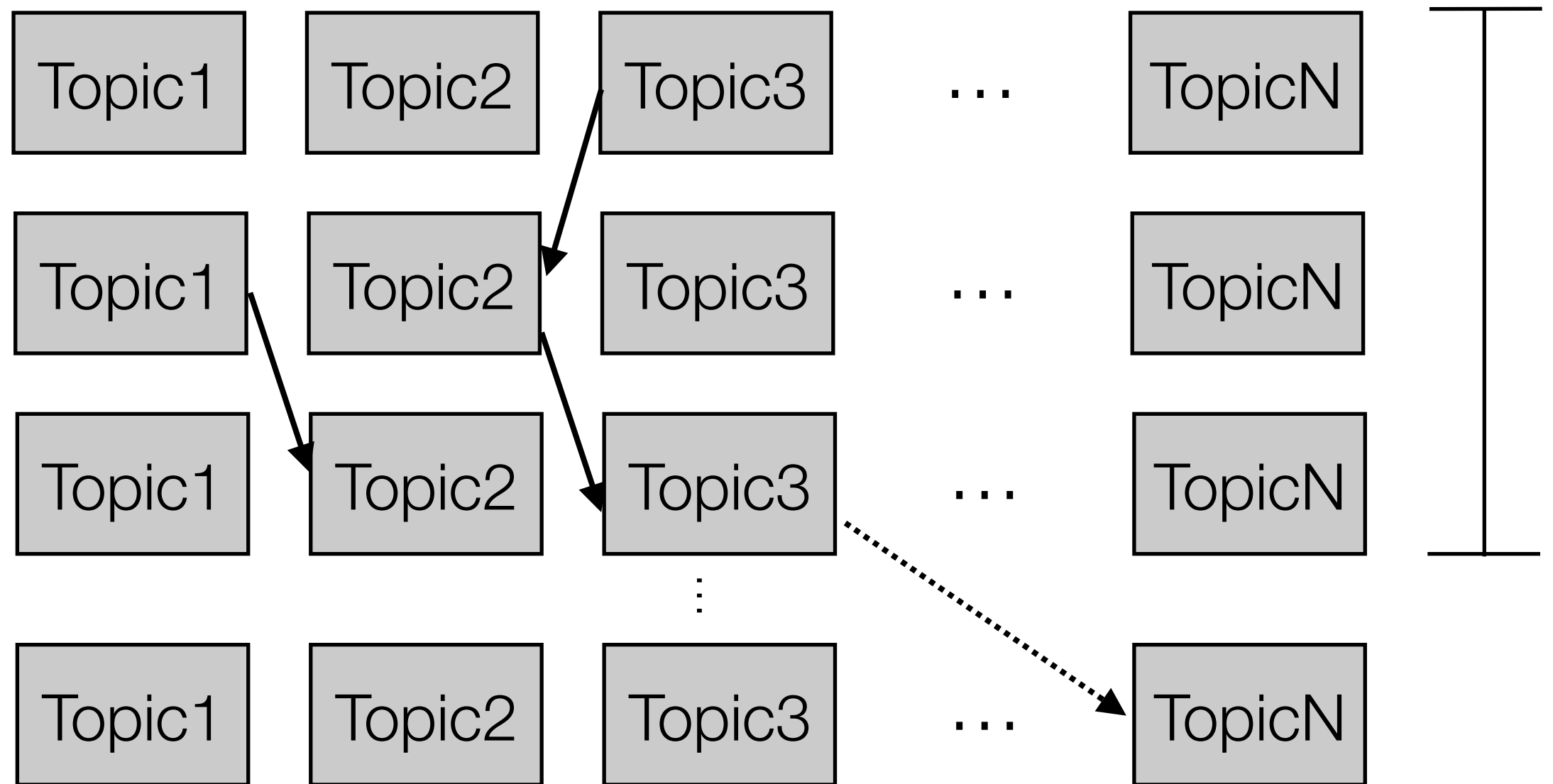


3. Find topics for each time slice using LDA



## Plan to Solve the Problem (4)

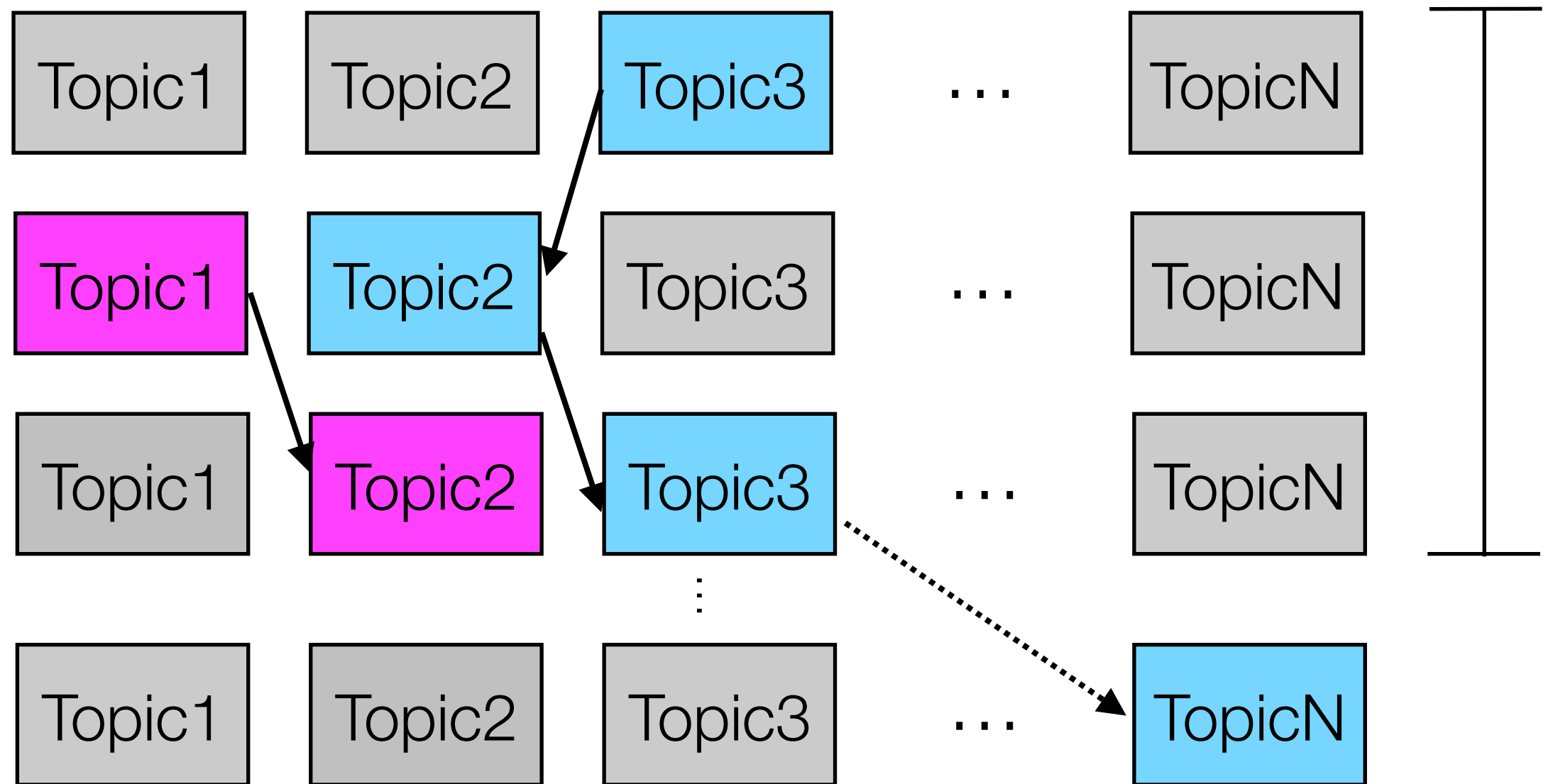
---



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

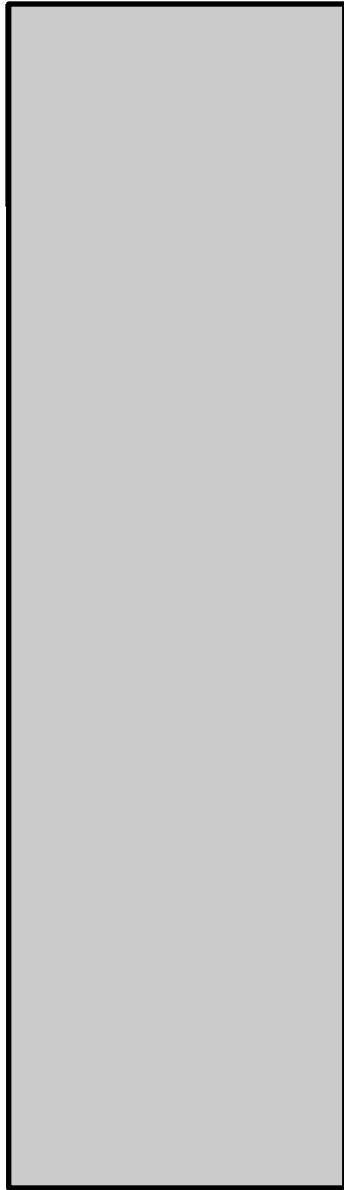
*Kendall's Tau; DCG*

- A topic is a set of top-probability words

*Jaccard's coefficient*

# Finding Best Similarity Metric (1)

---



Time Slice T

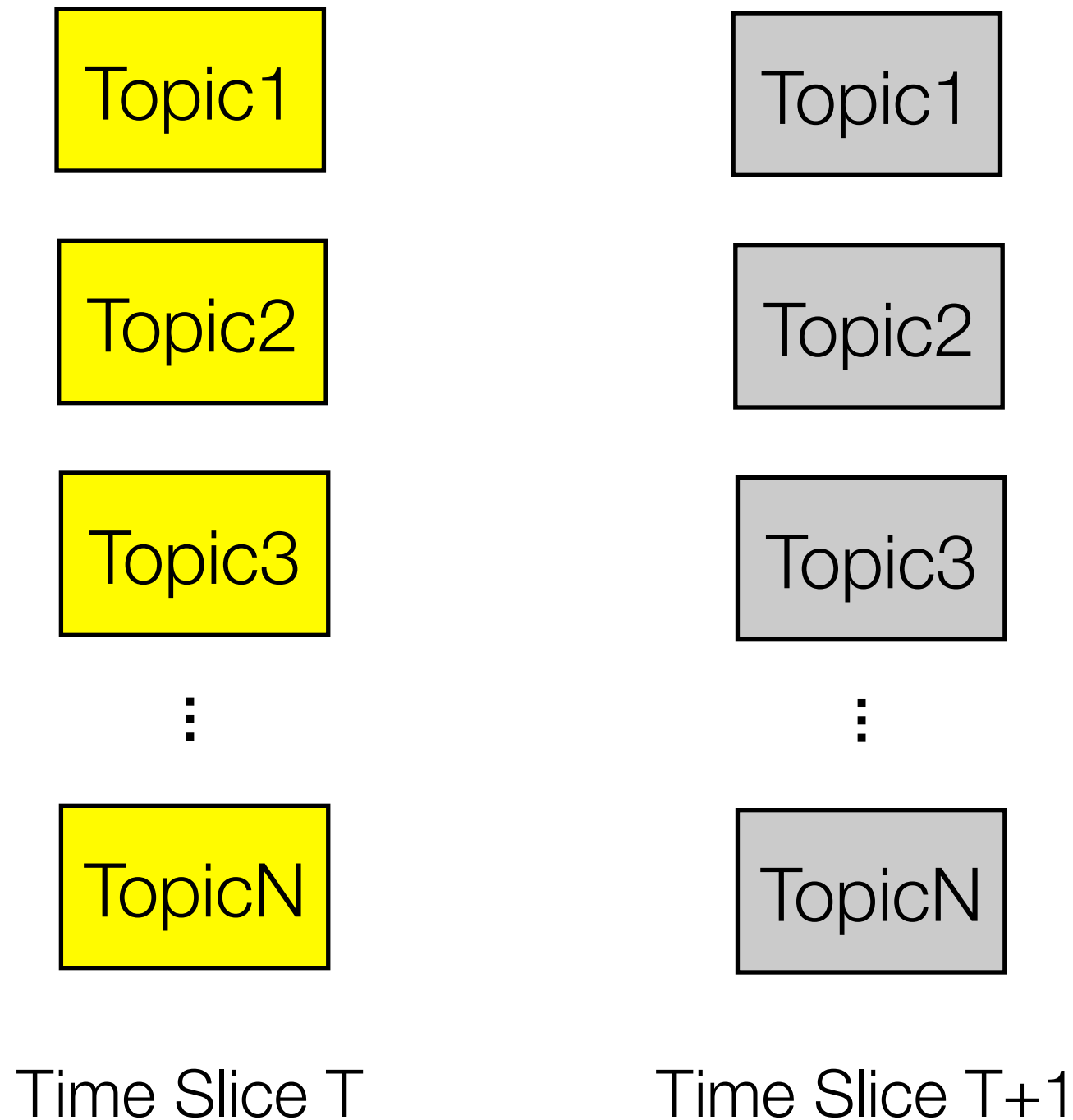


Time Slice T+1



## Finding Best Similarity Metric (2)

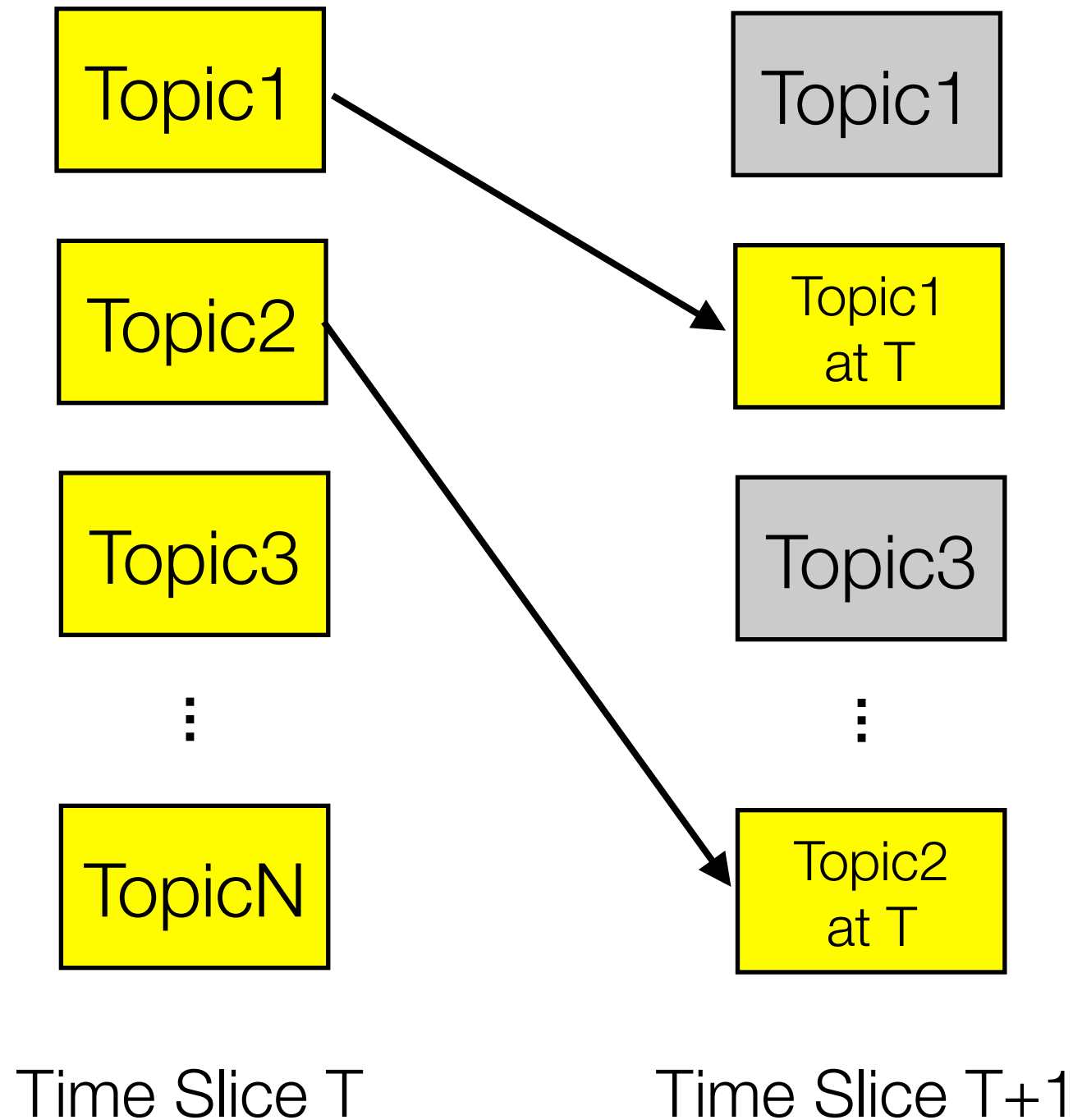
---



1. Calculate similarity  
between all pair of topics

## Finding Best Similarity Metric (3)

---

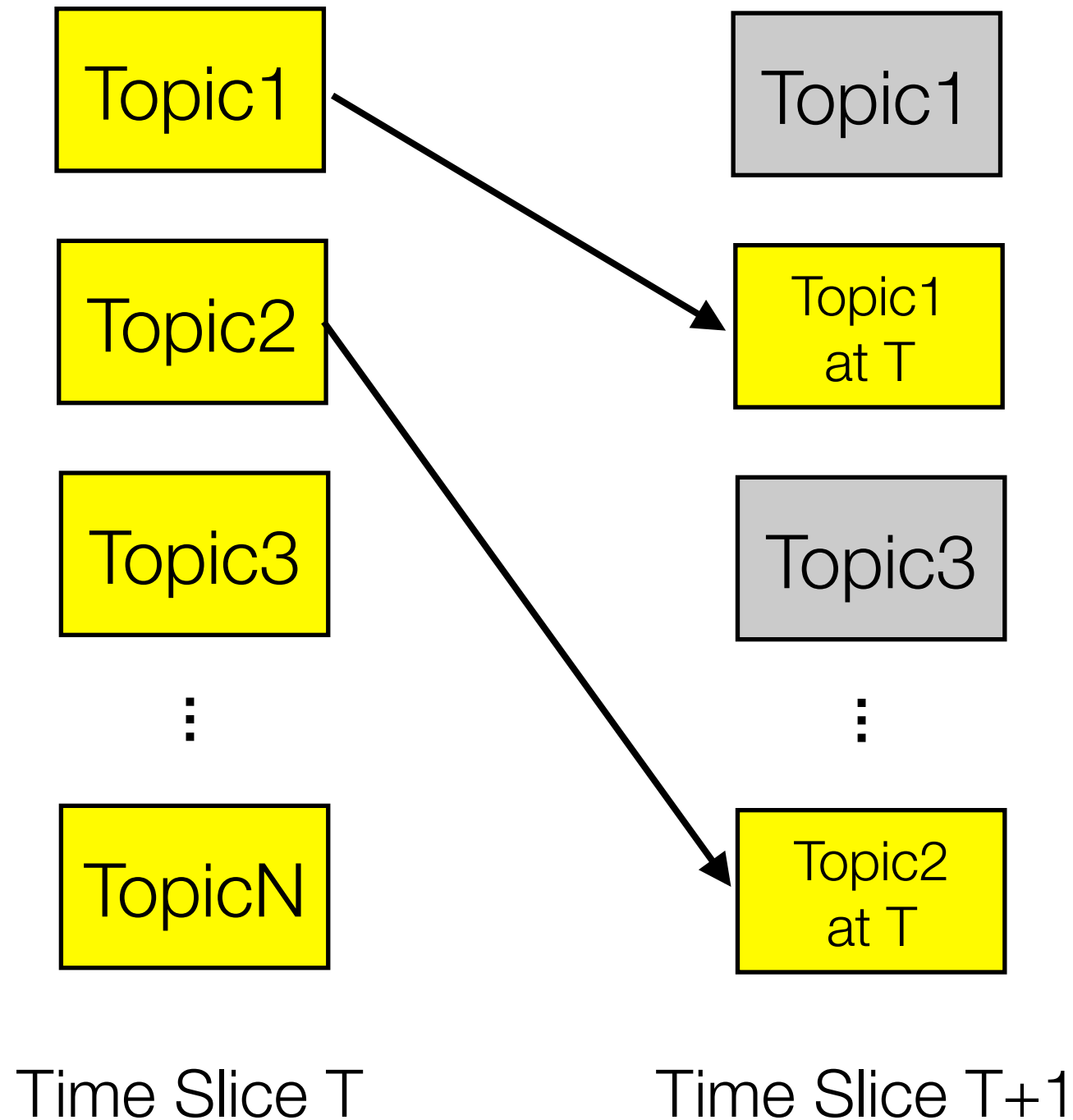


1. Calculate similarity between all pair of topics

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

## Finding Best Similarity Metric (4)

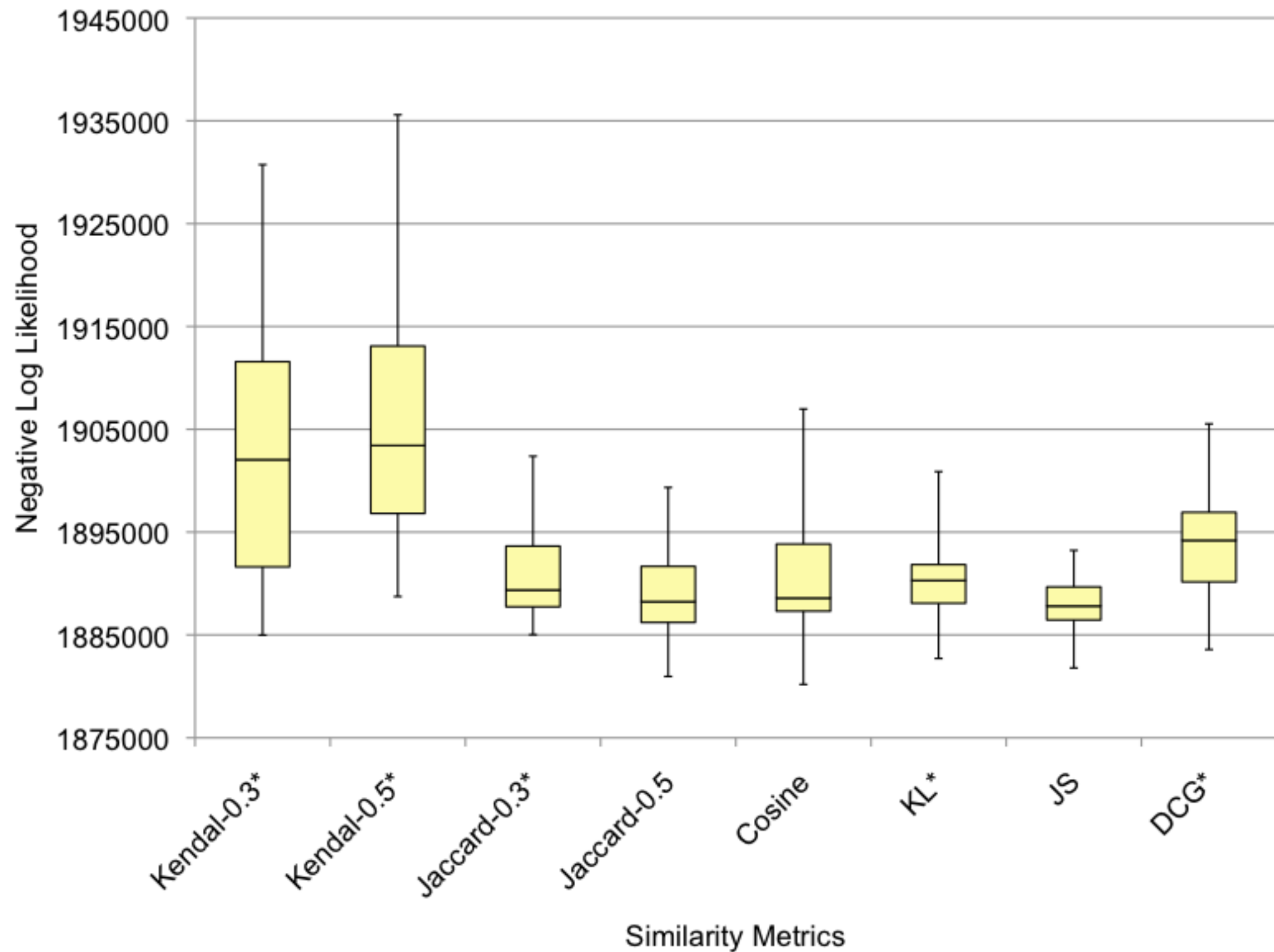
---



1. Calculate similarity between all pair of topics

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

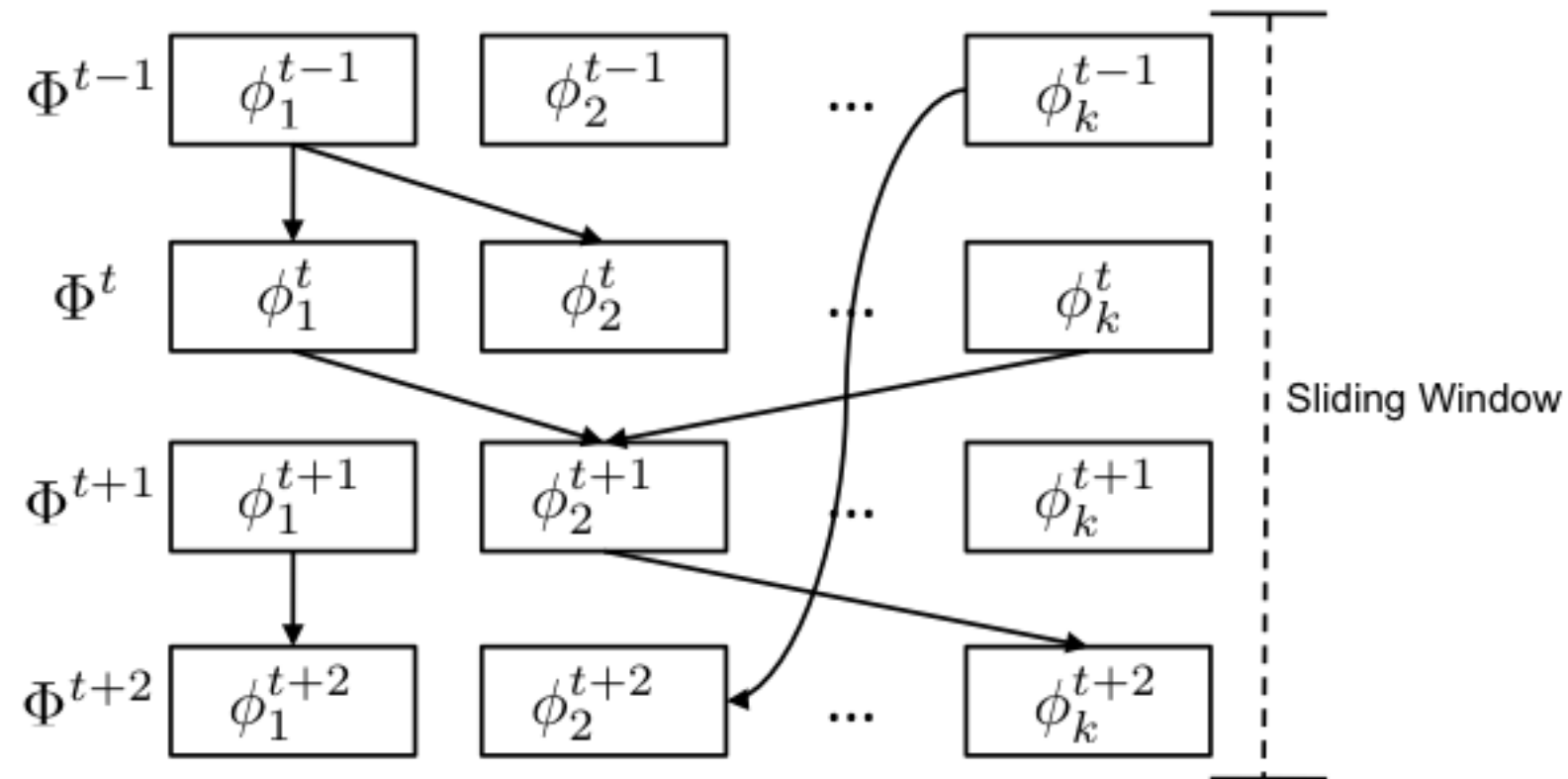
3. Calculate the likelihood of Time Slice T+1



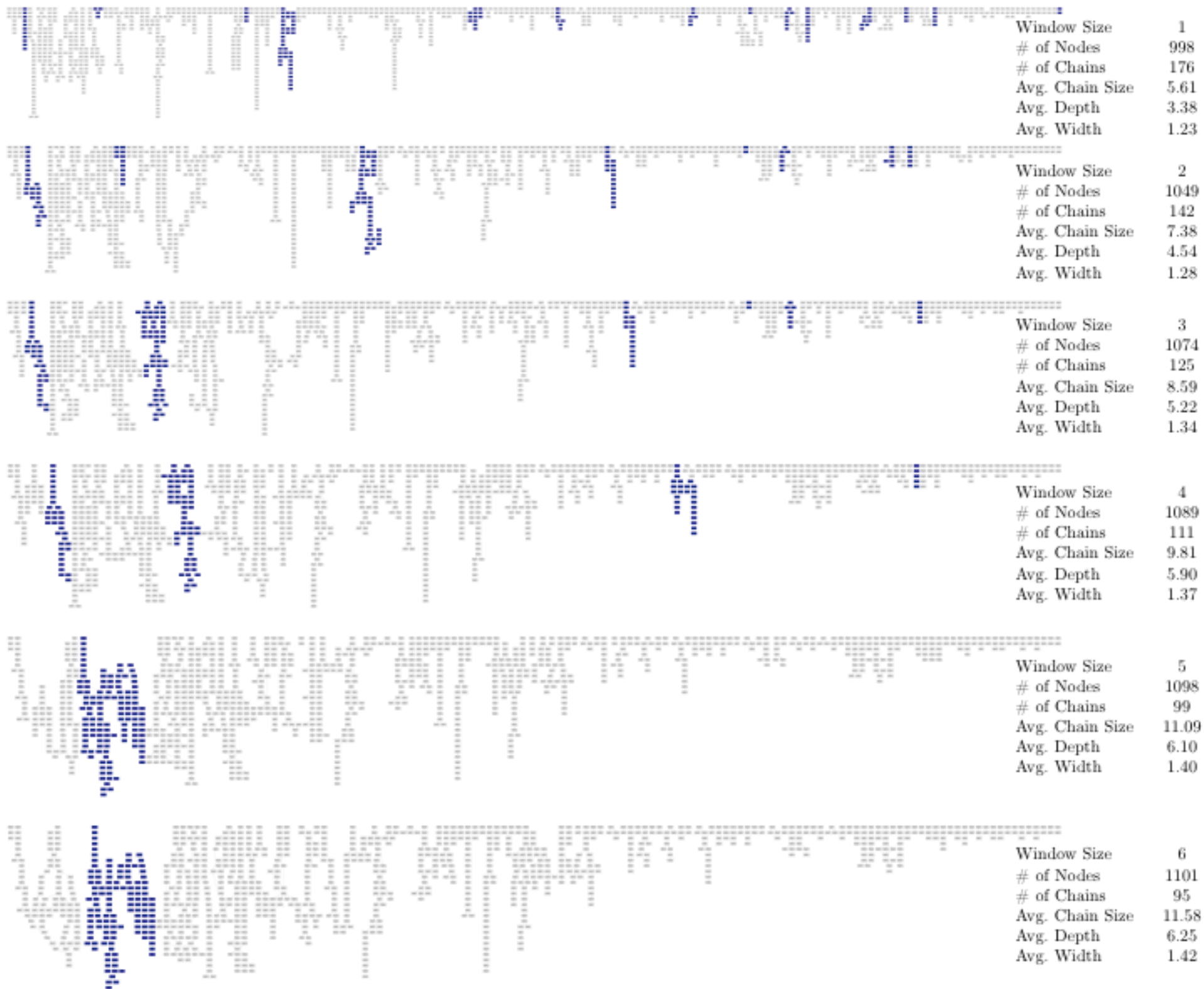
Comparison of frequently  
used similarity metrics

KL divergence is not the best

# Deciding the Range of **Neighboring** Time Slices



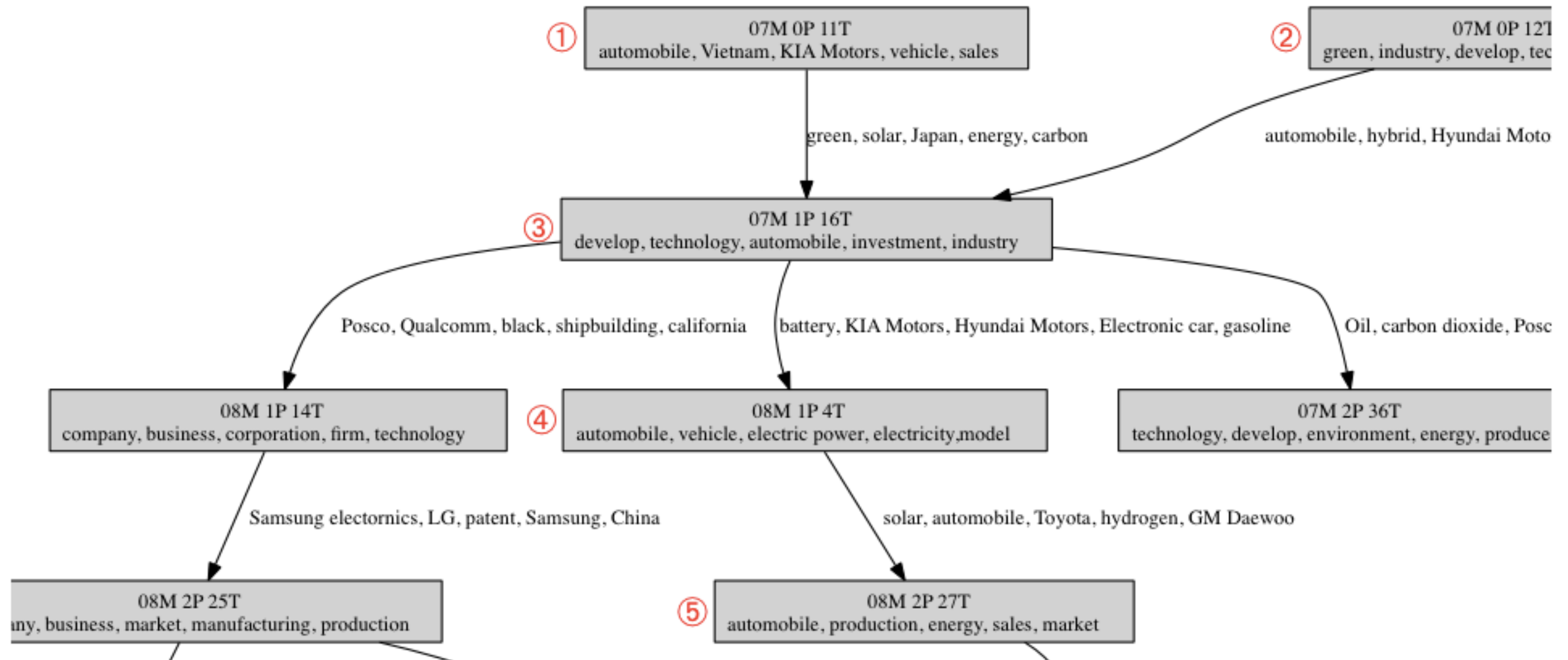
- Looking back one time slice -- might miss some chains that can be formed between two non-consecutive topics



A larger sliding window  
produces larger chains

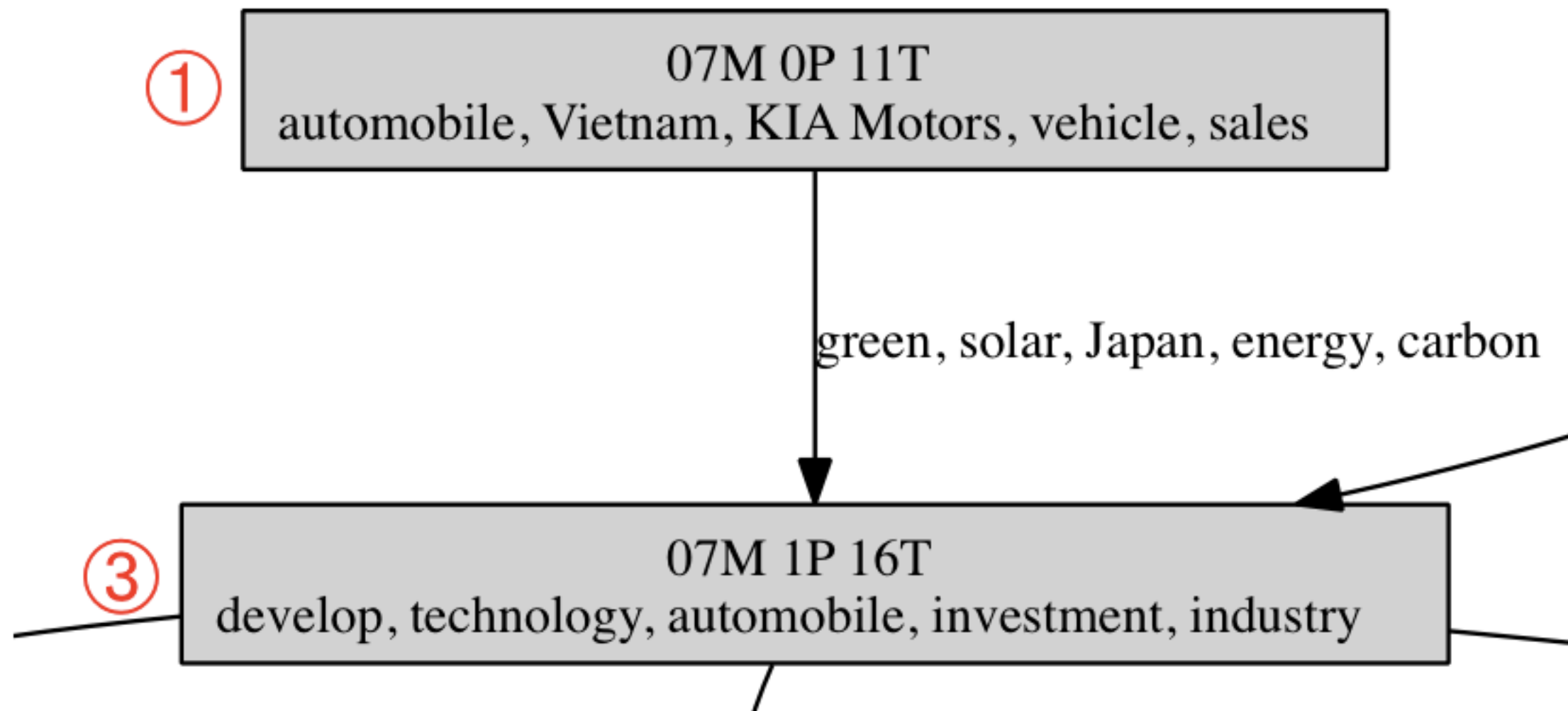
and they tend to be more abstract





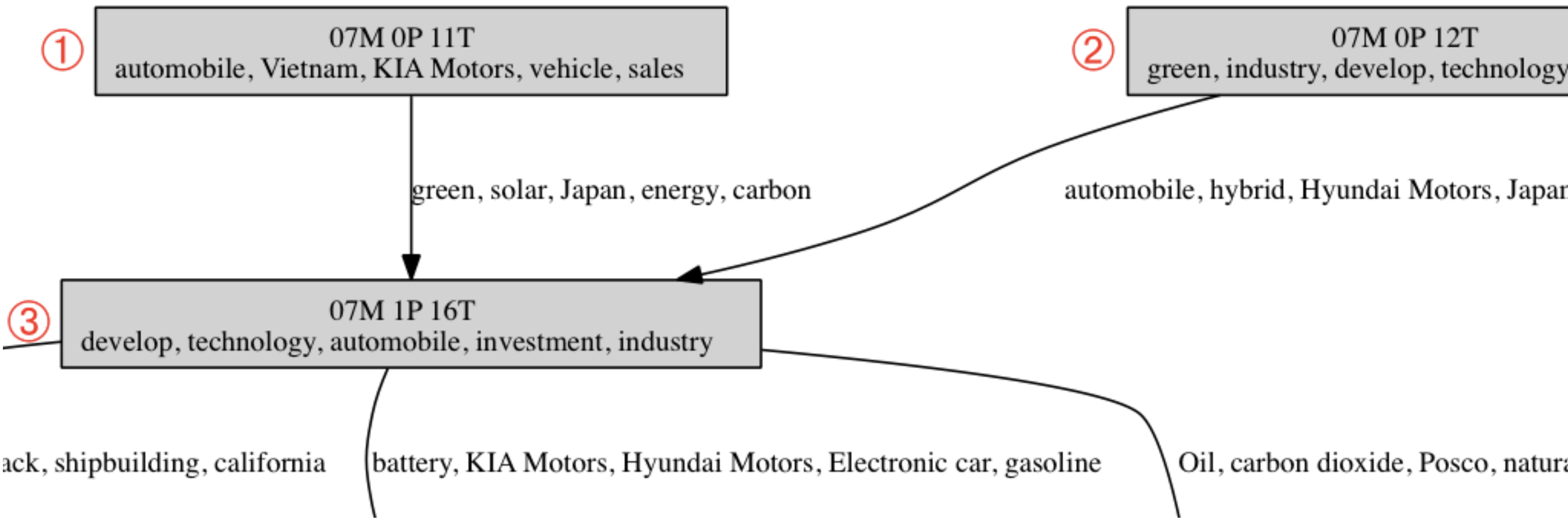
How the focus shifts in a  
topic chain

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