Summary of Word Association Mining

- Two basic associations: paradigmatic and syntagmatic
  - Generally applicable to any items in any language (e.g., phrases or entities as units)
- Pure statistical approaches are available for discovering both (can be combined to perform joint analysis).
  - Generally applicable to any text with no human effort
  - Different ways to define "context" and "segment" lead to interesting variations of applications
- Discovered associations can support many other applications.

7:03 / 9:42

Recommended Reading

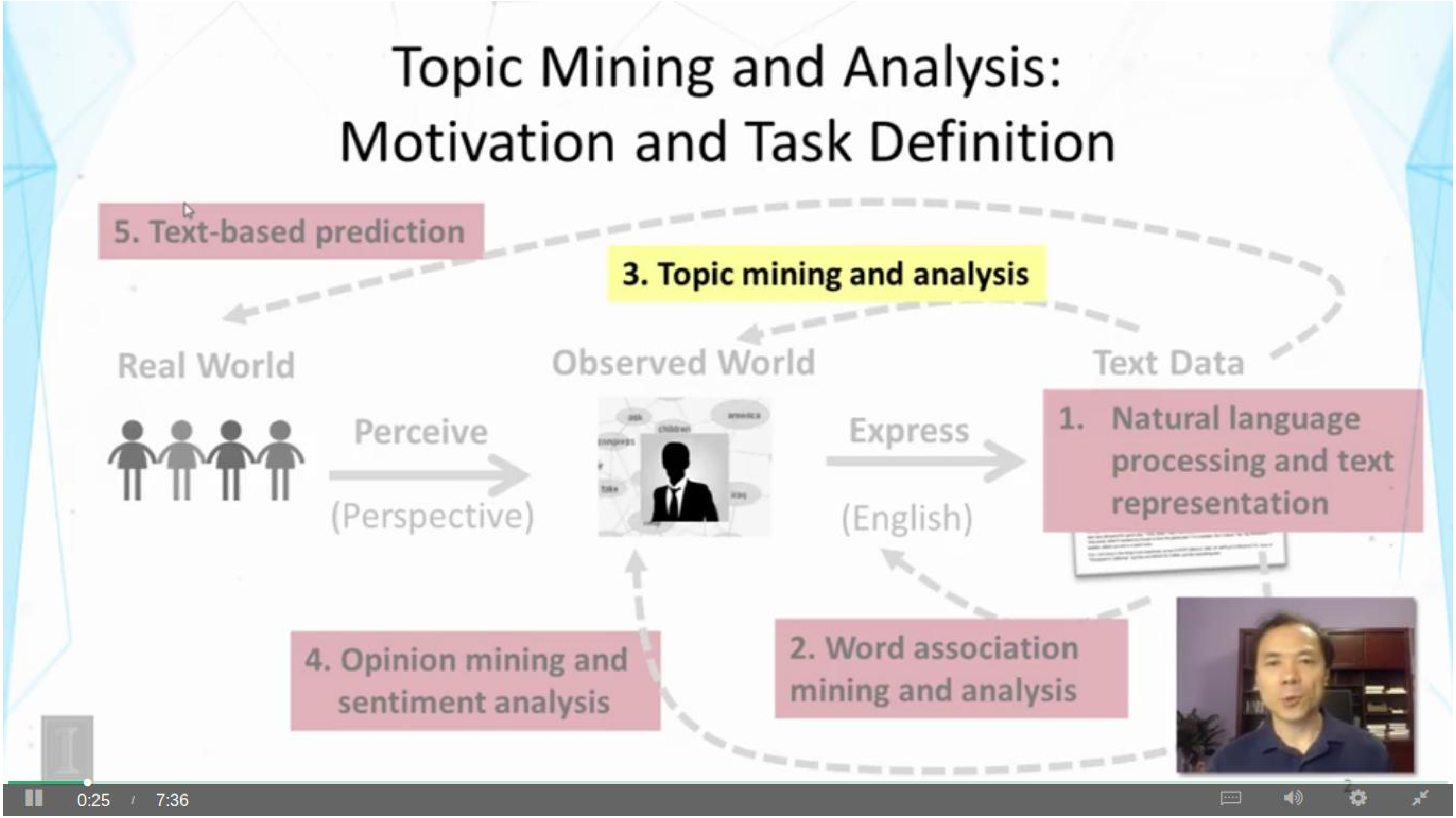
- Chris Manning and Hinrich Schütze, *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA: May 1999. (Chapter 5 on collocations)
- Chengxiang Zhai, Exploiting context to identify lexical atoms: A statistical view of linguistic context. Proceedings of the International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97), Rio de Janeiro, Brazil, Feb. 4-6, 1997. pp. 119-129.
- Shan Jiang and ChengXiang Zhai, Random walks on adjacency graphs for mining lexical relations from big text data. Proceedings of IEEE BigData Conference 2014, pp. 549-554.

8:44 / 9:42

Topic Mining and Analysis: Motivation and Task Definition

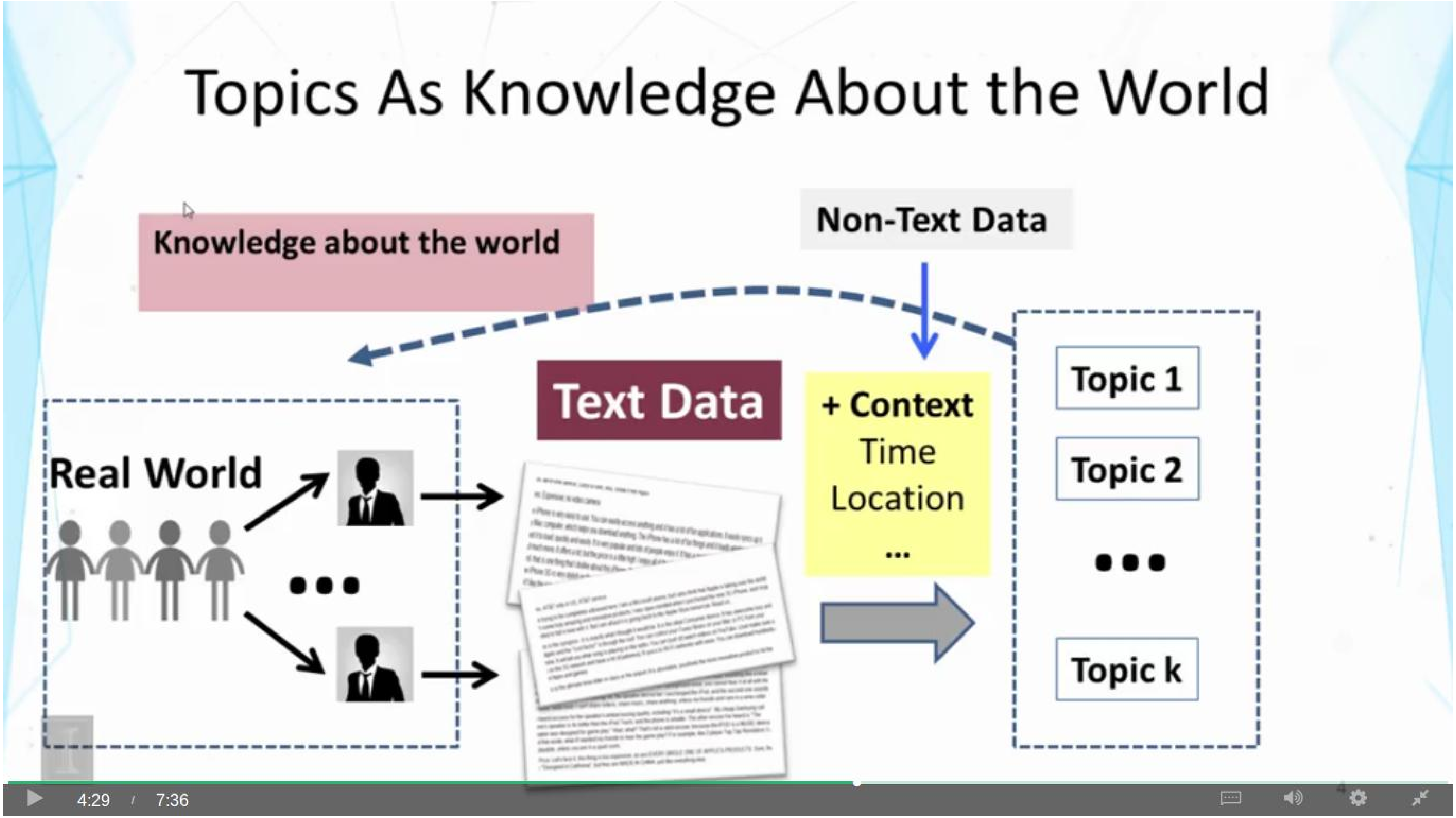
ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

0:09 / 7:36


Topic Mining and Analysis:
Motivation and Task Definition
5. Text-based prediction
3. Topic mining and analysis
Real World
Observed World
Text Data
Perceive
(Perspective)
Express
(English)
1. Natural language processing and text representation
4. Opinion mining and sentiment analysis
2. Word association mining and analysis
0:25 / 7:36

Topic Mining and Analysis: Motivation

- 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?


Topics As Knowledge About the World
Knowledge about the world
Non-Text Data
Text Data
+ Context
Time
Location
...
Real World
Topic 1
Topic 2
Topic k
4:29 / 7:36

Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **$C=\{d_1, \ldots, d_N\}$**
  - **Number of topics**: **k**
- Output
  - **k topics**: **$\{ \theta_1, \ldots, \theta_k \}$**
  - **Coverage of topics in each d_i**: **$\{ \pi_{i1}, \ldots, \pi_{ik} \}$**
  - π_{ij} = prob. of d_i covering topic θ_j

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

How to define θ_i ?

Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **$C=\{d_1, \ldots, d_N\}$**
  - **Number of topics**: **k**
- Output
  - **k topics**: **$\{\theta_1, \ldots, \theta_k\}$**
  - **Coverage of topics in each d_i**: **$\{\pi_{i1}, \ldots, \pi_{ik}\}$**
  - π_{ij} = prob. of d_i covering topic θ_j

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

How to define θ_i ?

Topic Mining and Analysis: Term as Topic

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

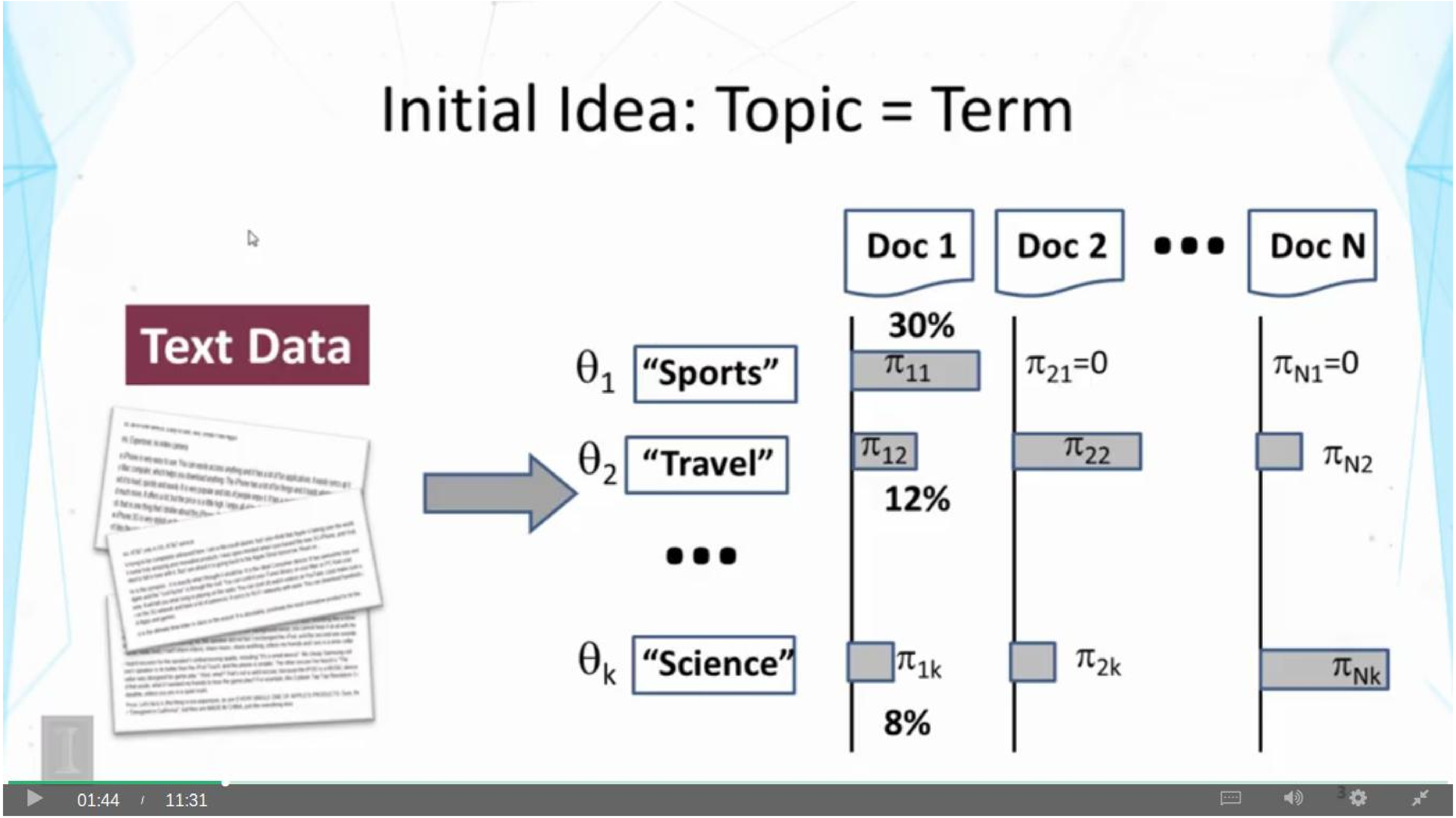
Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents $C=\{d_1, \ldots, d_N\}$
  - **Number of topics**: **k**
- Output
  - **k topics**: $\{ \theta_1, \ldots, \theta_k \}$
  - **Coverage of topics in each d_i**: $\{ \pi_{i1}, \ldots, \pi_{ik} \}$
  - π_{ij}=prob. of d_i covering topic θ_j

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

How to define θ_i ?

00:27 / 11:31



Initial Idea: Topic = Term
Text Data
θ1 "Sports"
θ2 "Travel"
θk "Science"
Doc 1
Doc 2
Doc N
30%
π11
π12
12%
π1k
8%
π21=0
π22
π2k
πN1=0
πN2
πNk
01:44 / 11:31

Mining k Topical Terms from Collection C

- Parse text in C to obtain candidate terms (e.g., term = word).

02:27 / 11:31

Mining k Topical Terms from Collection C

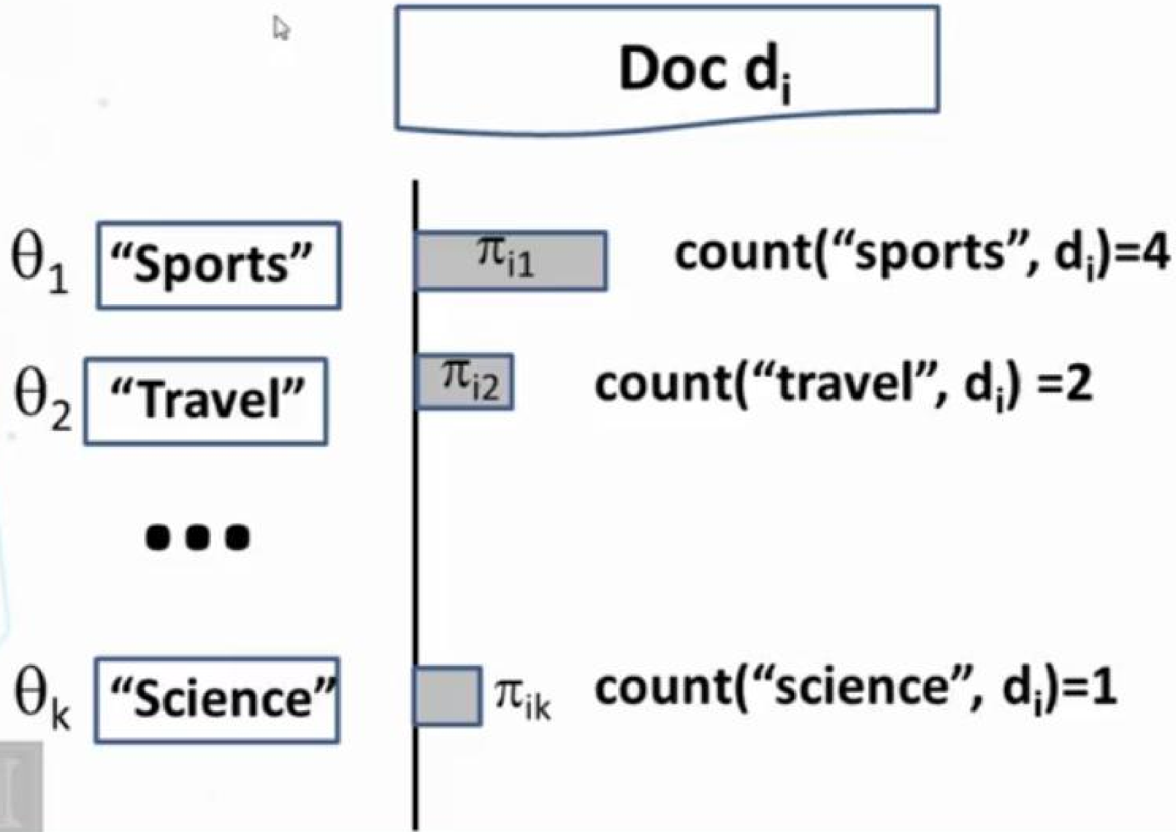
- Parse text in C to obtain candidate terms (e.g., term = word).
- Design a scoring function to measure how good each term is as a topic.
  - Favor a representative term (high frequency is favored)
  - Avoid words that are too frequent (e.g., “the”, “a”).
  - TF-IDF weighting from retrieval can be very useful.
  - Domain-specific heuristics are possible (e.g., favor title words, hashtags in tweets).

02:53 / 11:31

Mining k Topical Terms from Collection C

- Parse text in C to obtain candidate terms (e.g., term = word).
- Design a scoring function to measure how good each term is as a topic.
  - Favor a representative term (high frequency is favored)
  - Avoid words that are too frequent (e.g., "the", "a").
  - TF-IDF weighting from retrieval can be very useful.
  - Domain-specific heuristics are possible (e.g., favor title words, hashtags in tweets).
- Pick k terms with the highest scores but try to minimize redundancy.
  - If multiple terms are very similar or closely related, pick only one of them and ignore others.

Computing Topic Coverage: π_{ij}



$$\pi_{ij} = \frac{\text{count}(\theta_j, d_i)}{\sum_{L=1}^{k} \text{count}(\theta_L, d_i)}$$

How Well Does This Approach Work?

Doc d_i

Cavaliers vs. Golden State Warriors: NBA playoff finals … basketball game … travel to Cleveland … star …

θ_1 **"Sports"** $\pi_{i1} \propto c(\text{"sports"}, d_i) = 0$

1. Need to count related words also!

θ_2 **"Travel"** $\pi_{i2} \propto c(\text{"travel"}, d_i) = 1 > 0$

2. "Star" can be ambiguous (e.g., star in the sky).

…

θ_k **"Science"** $\pi_{ik} \propto c(\text{"science"}, d_i) = 0$

3. Mine complicated topics?

Problems with "Term as Topic"

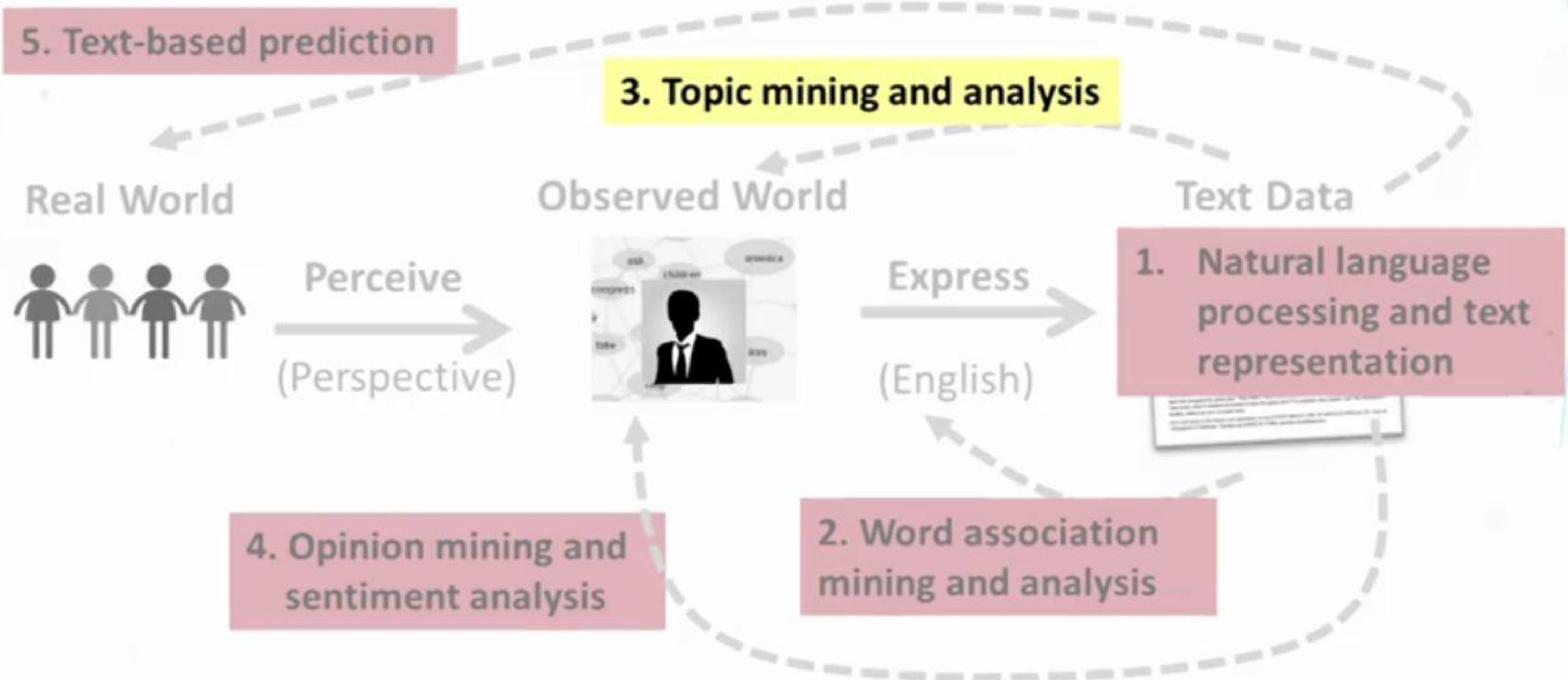
- Lack of expressive power
  - Can only represent simple/general topics
  - Can't represent complicated topics
- Incompleteness in vocabulary coverage
  - Can't capture variations of vocabulary (e.g., related words)
- Word sense ambiguity
  - A topical term or related term can be ambiguous (e.g., basketball star vs. star in the sky)

Topic Mining and Analysis: Probabilistic Topic Models

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

00:07 / 14:17

Topic Mining and Analysis: Probabilistic Topic Models



Problems with "Term as Topic"

- Lack of expressive power **➔ Topic = {Multiple Words}**
  - Can only represent simple/general topics
  - Can't represent complicated topics
- Incompleteness in vocabulary coverage **+ weights on words**
  - Can't capture variations of vocabulary (e.g., related words)
- Word sense ambiguity **➔ Split an ambiguous word**
  - A topical term or related term can be ambiguous (e.g., basketball star vs. star in the sky)

A probabilistic topic model can do all these!

Improved Idea: Topic = Word Distribution

θ_1 "Sports"

$P(w|\theta_1)$

sports	0.02
game	0.01
basketball	0.005
football	0.004
play	0.003
star	0.003
…	
nba	0.001
…	
travel	0.0005
…	

θ_2 "Travel"

$P(w|\theta_2)$

travel	0.05
attraction	0.03
trip	0.01
flight	0.004
hotel	0.003
island	0.003
…	
culture	0.001
…	
play	0.0002
…	

•••

θ_k "Science"

$P(w|\theta_k)$

science	0.04
scientist	0.03
spaceship	0.006
telescope	0.004
genomics	0.004
star	0.002
…	
genetics	0.001
…	
travel	0.00001
…	

$$\sum_{w \in V} p(w|\theta_i) = 1$$

Vocabulary Set: V={w1, w2,….}

04:31 / 14:17

Probabilistic Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **C={d_1, …, d_N}**
  - **Vocabulary set: V={w_1, …, w_M}**
  - **Number of topics**: **k**
- Output
  - **k topics, each a word distribution**: **{ θ_1, …, θ_k }**
  - **Coverage of topics in each d_i**: **{ π_{i1}, …, π_{ik} }**
  - π_{ij}=prob. of d_i covering topic θ_j

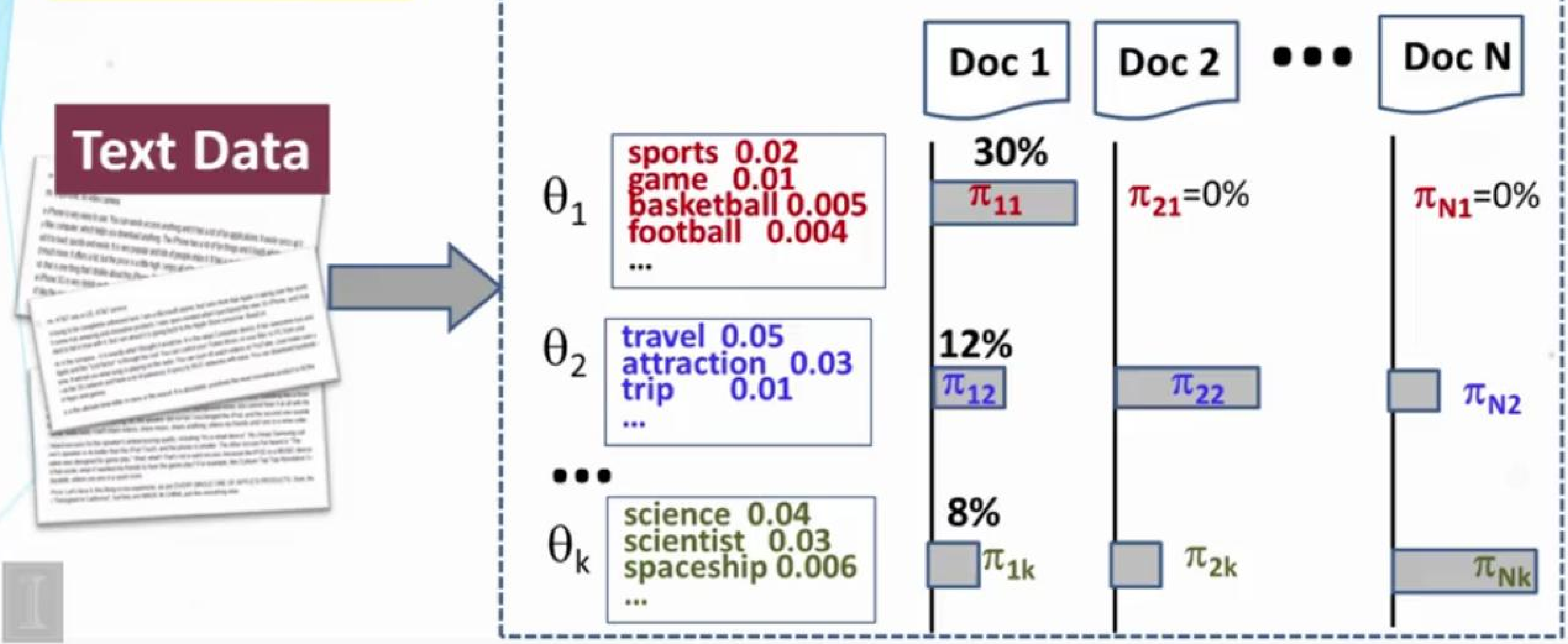
$$\sum_{w \in V} p(w \mid \theta_i) = 1$$

$$\sum_{j=1}^{k} \pi_{ij} = 1$$

The Computation Task

INPUT: C, k, V

OUTPUT: $\{ \theta_1, \ldots, \theta_k \}, \{ \pi_{i1}, \ldots, \pi_{ik} \}$



Generative Model for Text Mining

Modeling of Data Generation: P(Data |Model, Λ)

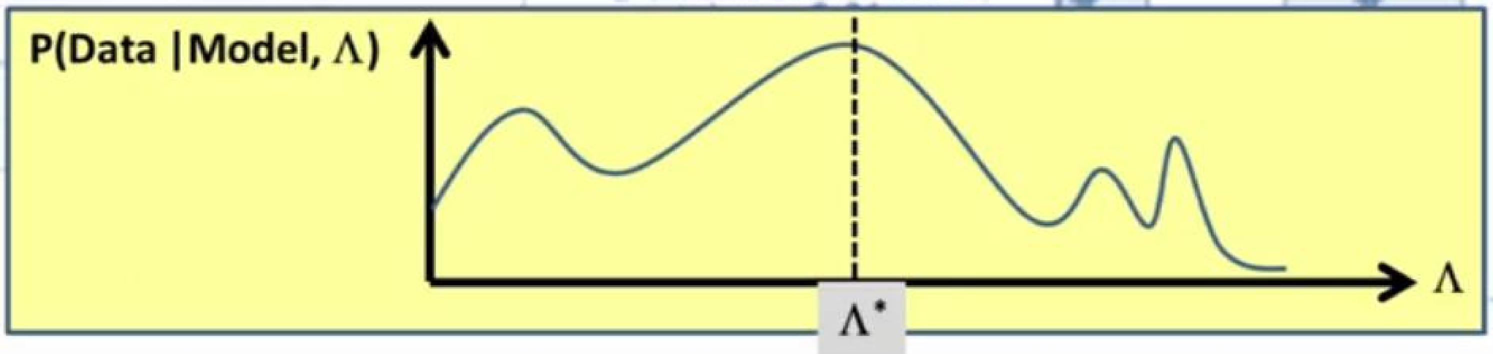
Λ=({ θ_1, …, θ_k }, { π_{11}, …, π_{1k} }, …, { π_{N1}, …, π_{Nk} })

Text Data

How many parameters in total?

Parameter Estimation/ Inferences

$\Lambda^* = \text{argmax}_{\Lambda}\ p(\text{Data} |\ \text{Model}, \Lambda)$



Summary

- Topic represented as word distribution
  - Multiple words: allow for describing a complicated topic
  - Weights on words: model subtle semantic variations of a topic
- Task of topic mining and analysis
  - Input: collection C, number of topics k, vocabulary set V
  - Output: a set of topics, each a word distribution; coverage of all topics in each document

$$\Lambda=(\{\theta_1, \ldots, \theta_k\}, \{\pi_{11}, \ldots, \pi_{1k}\}, \ldots, \{\pi_{N1}, \ldots, \pi_{Nk}\})$$

$$\forall j \in [1,k], \sum_{w \in V} p(w \mid \theta_j) = 1$$

$$\forall i \in [1,N], \sum_{j=1}^{k} \pi_{ij} = 1$$

Summary (cont.)

- **Generative model** for text mining
  - **Model data generation** with a prob. model: **P(Data |Model, Λ)**
  - **Infer the most likely parameter values Λ^*** given a particular data set: $\Lambda^* = \text{argmax}_{\Lambda}\, p(\text{Data} \mid \text{Model}, \Lambda)$
  - **Take Λ^* as the "knowledge"** to be mined for the text mining problem
  - **Adjust** the design of the model to discover different knowledge

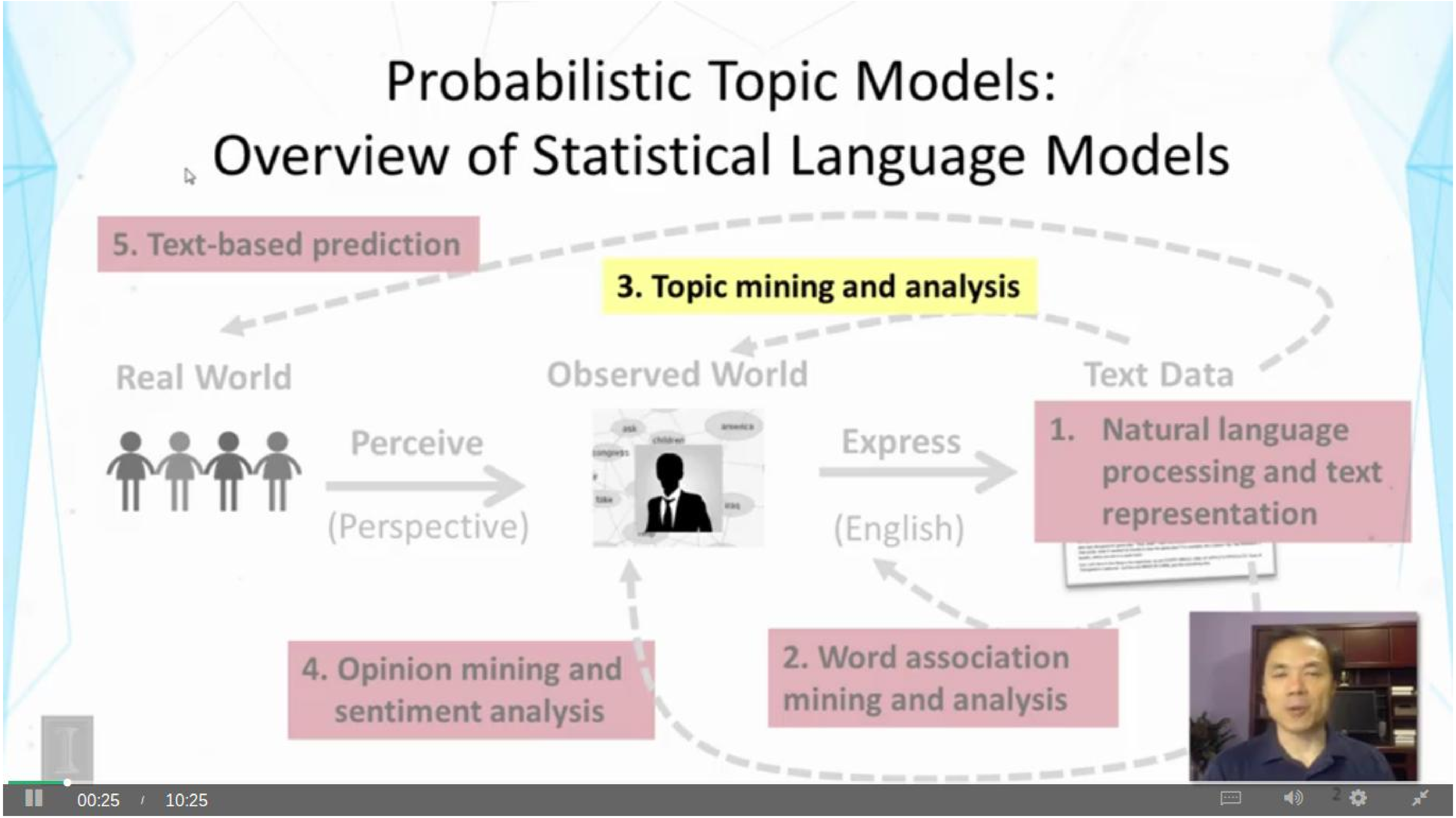
13:38 / 14:17

Topic Mining and Analysis: Overview of Statistical Language Models

Part 1

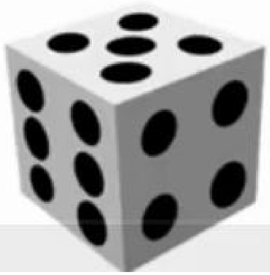
ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

00:14 / 10:25


Probabilistic Topic Models:
Overview of Statistical Language Models
5. Text-based prediction
3. Topic mining and analysis
Real World
Observed World
Text Data
Perceive
(Perspective)
Express
(English)
1. Natural language processing and text representation
4. Opinion mining and sentiment analysis
2. Word association mining and analysis
00:25 / 10:25

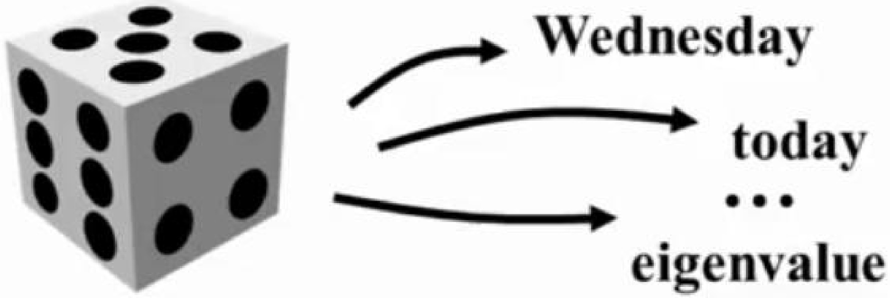
What Is a Statistical Language Model (LM)?

- A probability distribution over word sequences
  - p("*Today is Wednesday*") ≈ 0.001
  - p("*Today Wednesday is*") ≈ 0.0000000000001
  - p("*The eigenvalue is positive*") ≈ 0.00001
- Context-dependent!
- Can also be regarded as a probabilistic mechanism for "generating" text – thus also called a "generative" model



The Simplest Language Model: Unigram LM

- Generate text by generating each word INDEPENDENTLY
- Thus, $p(w_1\ w_2\ \ldots\ w_n)=p(w_1)p(w_2)\ldots p(w_n)$
- Parameters: $\{p(w_i)\}$ $p(w_1)+\ldots+p(w_N)=1$ (N is voc. size)
- Text = sample drawn according to this **word distribution**



p("today is Wed")

$= p(\text{"today"})p(\text{"is"})p(\text{"Wed"})$

$= 0.0002 \times 0.001 \times 0.000015$

04:48 4:59
04:48 / 10:25

Text Generation with Unigram LM

Unigram LM $p(w|\theta)$ —— **Sampling** ——> **Document d**

$p(d|\theta)$=?

Topic 1:
Text mining

…
text 0.2
mining 0.1
association 0.01
clustering 0.02
…
food 0.00001
…

→ **Text mining paper**

Topic 2:
Health

…
food 0.25
nutrition 0.1
healthy 0.05
diet 0.02
…

→ **Food nutrition paper**

Estimation of Unigram LM

Unigram LM $p(w|\theta)$=? ← **Estimation** — **Text Mining Paper d**

Total #words=**100**

Estimate	Word
	…
10/100	text ?
5/100	mining ?
3/100	association ?
3/100	database ?
	…
1/100	query ?
	…

text 10
mining 5
association 3
database 3
algorithm 2
…
query 1
efficient 1

Is this our best estimate?
How do we define "best"?

Topic Mining and Analysis: Overview of Statistical Language Models

Part 2

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Maximum Likelihood vs. Bayesian

- Maximum likelihood estimation
  - "Best" means "data likelihood reaches maximum"

$$\hat{\theta} = \arg\max_{\theta} P(X \mid \theta)$$

  - Problem: Small sample
- Bayesian estimation:

Bayes Rule $p(X \mid Y) = \dfrac{p(Y \mid X)p(X)}{p(Y)}$

  - "Best" means being consistent with our "prior" knowledge and explaining data well

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid X) = \arg\max_{\theta} P(X \mid \theta) P(\theta)$$

  - Problem: How to define prior?

Maximum a Posteriori (MAP) estimate

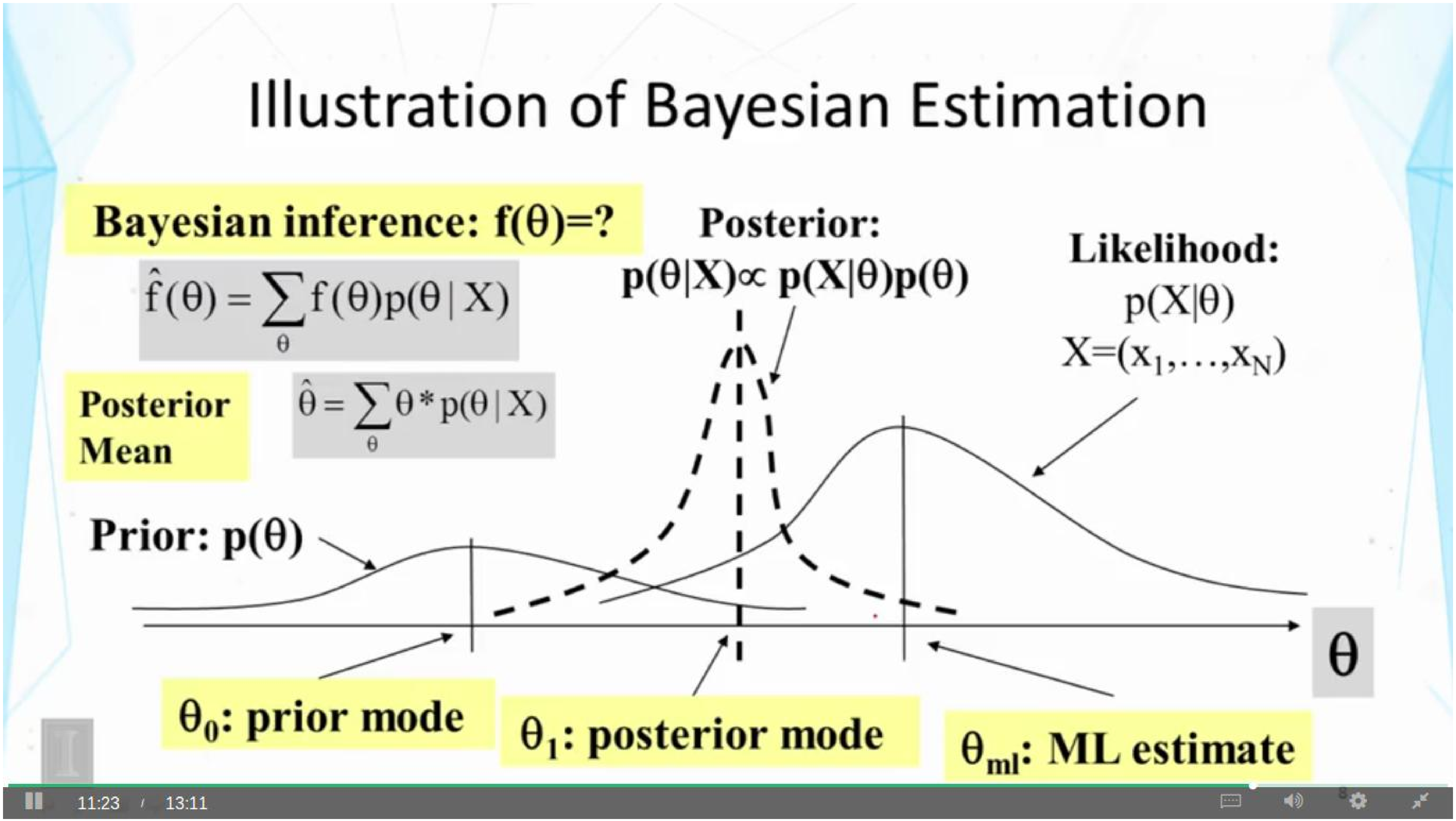


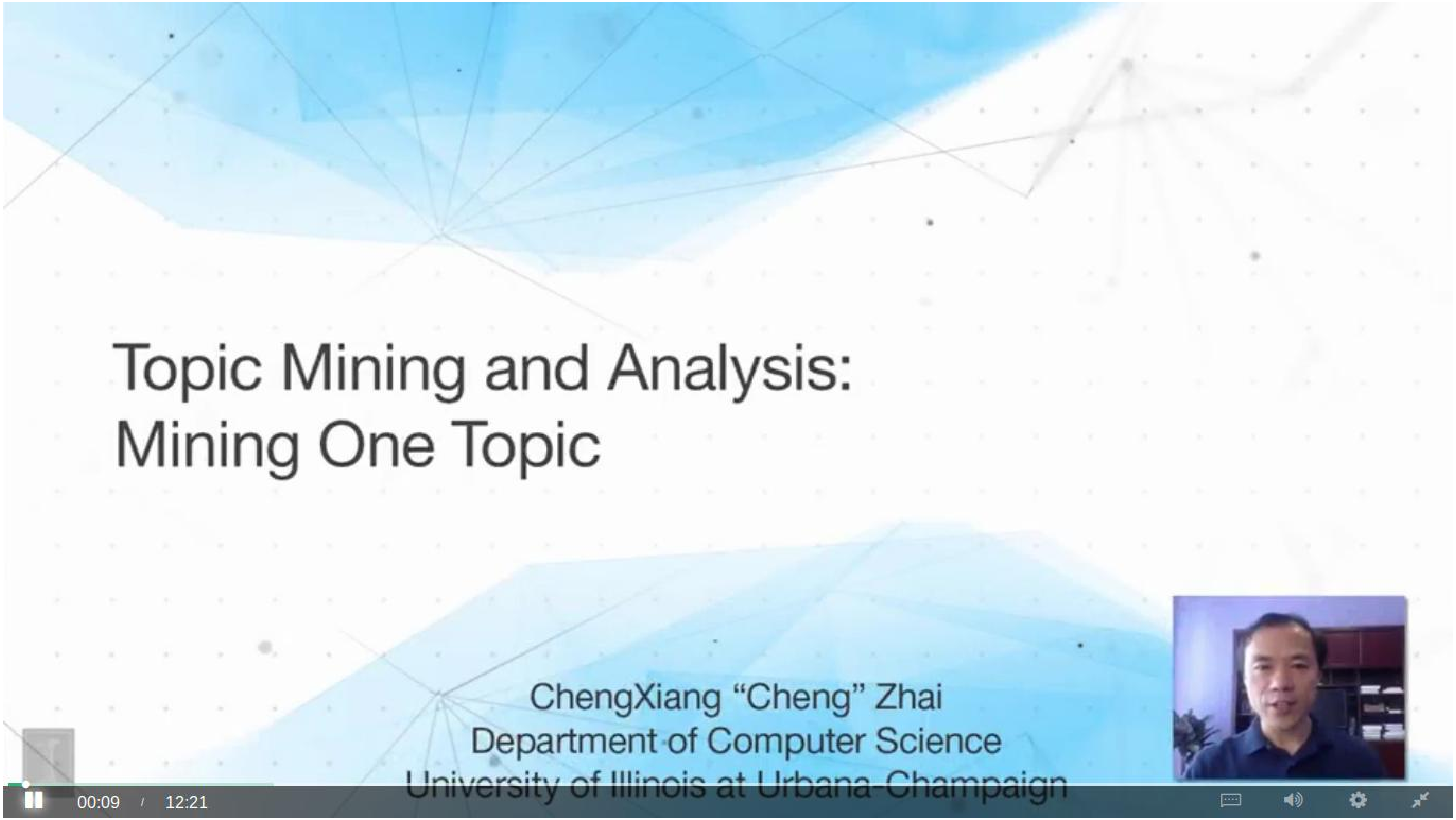
Illustration of Bayesian Estimation
Bayesian inference: f(θ)=?
$\hat{f}(\theta) = \sum_{\theta} f(\theta) p(\theta \mid X)$
Posterior Mean
$\hat{\theta} = \sum_{\theta} \theta * p(\theta \mid X)$
Posterior:
$p(\theta|X) \propto p(X|\theta)p(\theta)$
Likelihood:
$p(X|\theta)$
$X=(x_1,\ldots,x_N)$
Prior: p(θ)
θ
θ_0: prior mode
θ_1: posterior mode
θ_{ml}: ML estimate
11:23 / 13:11

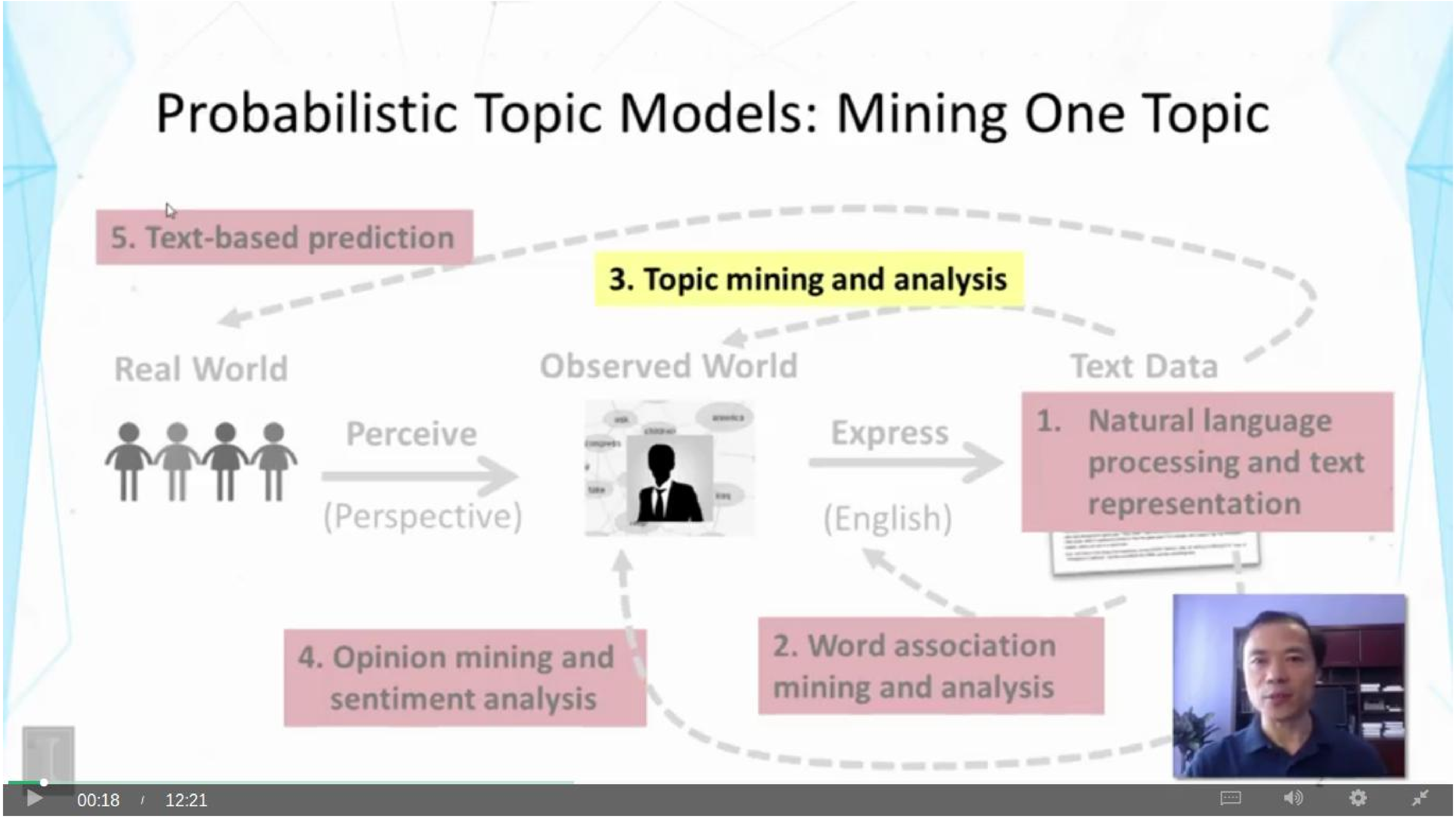
Summary

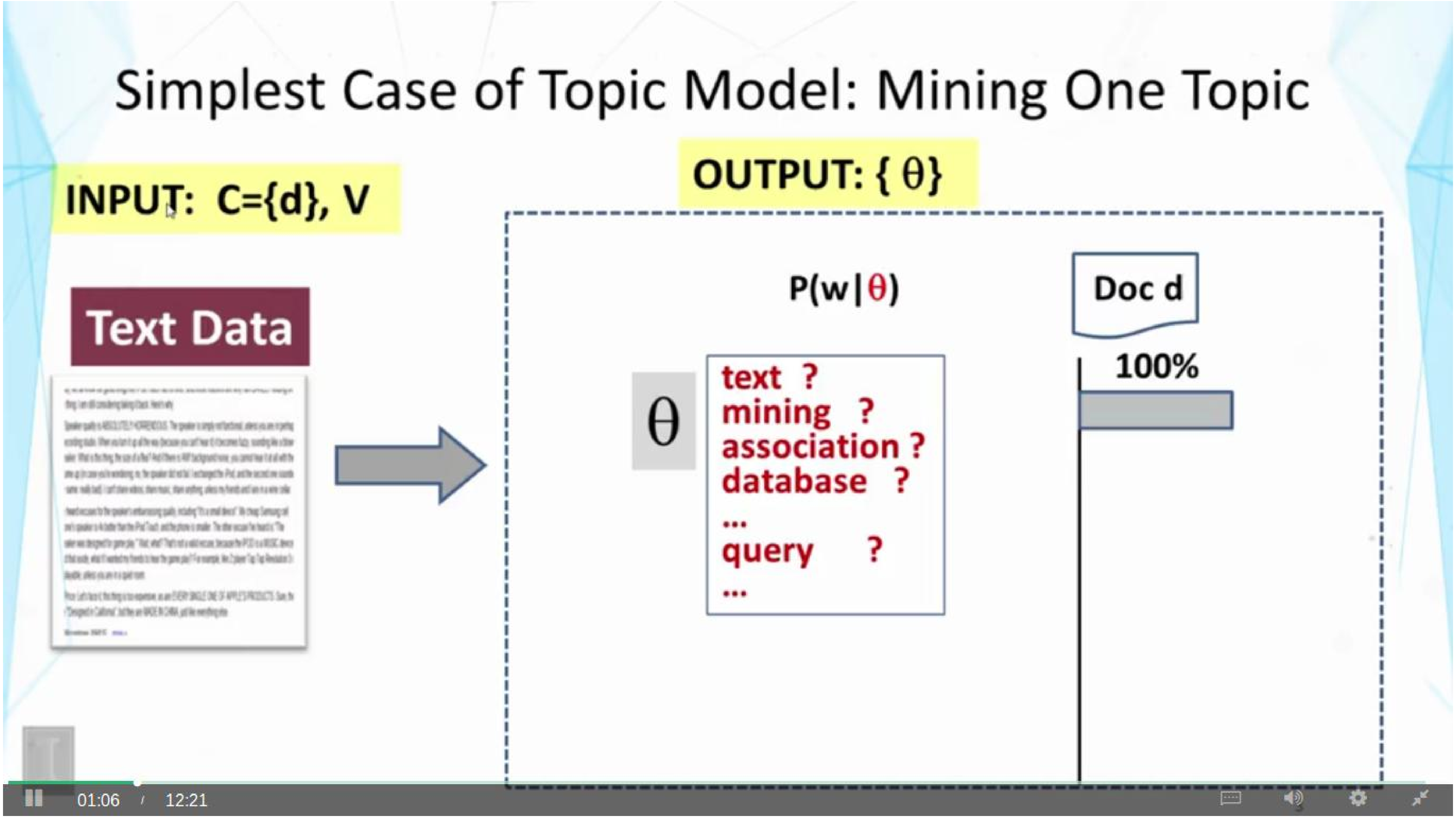
- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: $\mathbf{p(X|\theta)}$
  - **Given** θ ➔ which X has a higher likelihood?
  - **Given X** ➔ which θ maximizes $p(X|\theta)$? **[ML estimate]**
- **Bayesian** estimation/inference
  - Must define a **prior:** $\mathbf{p(\theta)}$
  - **Posterior** distribution: $\mathbf{p(\theta|X) \propto p(X|\theta)p(\theta)}$

  ➔ **Allows for inferring any "derived value" from** θ**!**

12:12 / 13:11


Topic Mining and Analysis:
Mining One Topic
ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign
00:09 / 12:21


Probabilistic Topic Models: Mining One Topic
5. Text-based prediction
3. Topic mining and analysis
Real World
Observed World
Text Data
Perceive
(Perspective)
Express
(English)
1. Natural language processing and text representation
4. Opinion mining and sentiment analysis
2. Word association mining and analysis
00:18 / 12:21


Simplest Case of Topic Model: Mining One Topic
INPUT: C={d}, V
OUTPUT: { θ}
Text Data
P(w|θ)
Doc d
θ
text ?
mining ?
association ?
database ?
...
query ?
...
100%
01:06 / 12:21

Language Model Setup

- **Data**: Document d= $x_1\ x_2 \ldots x_{|d|}$, $x_i \in V=\{w_1, \ldots, w_M\}$ is a word
- **Model**: Unigram LM θ(=topic) : $\{\theta_i=p(w_i \mid \theta)\}$, i=1, …, M; $\theta_1+\ldots+\theta_M=1$
- **Likelihood** function: $p(d \mid \theta) = p(x_1 \mid \theta) \times \ldots \times p(x_{|d|} \mid \theta)$

$$= p(w_1 \mid \theta)^{c(w_1,d)} \times \ldots \times p(w_M \mid \theta)^{c(w_M,d)}$$

$$= \prod_{i=1}^{M} p(w_i \mid \theta)^{c(w_i,d)} = \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$$

- **ML estimate:** $(\hat{\theta}_1, \ldots, \hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d \mid \theta) = \arg\max_{\theta_1,\ldots,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$

04:05 / 12:21

Computation of Maximum Likelihood Estimate

Maximize p(d|θ) $$(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d\,|\,\theta) = \arg\max_{\theta_1,\ldots,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$$

Max. Log-Likelihood $$(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} \log[p(d\,|\,\theta)] = \arg\max_{\theta_1,\ldots,\theta_M} \sum_{i=1}^{M} c(w_i,d)\log\theta_i$$

Subject to constraint: $$\boxed{\sum_{i=1}^{M}\theta_i = 1}$$

Use Lagrange multiplier approach

Lagrange function: $$f(\theta\,|\,d) = \sum_{i=1}^{M} c(w_i,d)\log\theta_i + \lambda\left(\sum_{i=1}^{M}\theta_i - 1\right)$$

$$\frac{\partial f(\theta\,|\,d)}{\partial\theta_i} = \frac{c(w_i,d)}{\theta_i} + \lambda = 0 \;\rightarrow\; \theta_i = -\frac{c(w_i,d)}{\lambda}$$

$$\sum_{i=1}^{M} -\frac{c(w_i,d)}{\lambda} = 1 \rightarrow \lambda = -\sum_{i=1}^{N} c(w_i,d) \rightarrow \hat{\theta}_i = p(w_i\,|\,\hat{\theta}) = \frac{c(w_i,d)}{\sum_{i=1}^{M} c(w_i,d)} = \frac{c(w_i,d)}{|d|}$$

07:50 / 12:21

Computation of Maximum Likelihood Estimate

Maximize p(d|θ) $(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} p(d\mid\theta) = \arg\max_{\theta_1,\ldots,\theta_M} \prod_{i=1}^{M} \theta_i^{c(w_i,d)}$

Max. Log-Likelihood $(\hat{\theta}_1,\ldots,\hat{\theta}_M) = \arg\max_{\theta_1,\ldots,\theta_M} \log[p(d\mid\theta)] = \arg\max_{\theta_1,\ldots,\theta_M} \sum_{i=1}^{M} c(w_i,d)\log\theta_i$

Subject to constraint: $\boxed{\sum_{i=1}^{M}\theta_i = 1}$

Use Lagrange multiplier approach

Lagrange function: $f(\theta\mid d) = \sum_{i=1}^{M} c(w_i,d)\log\theta_i + \lambda(\sum_{i=1}^{M}\theta_i - 1)$

Normalized Counts

$$\frac{\partial f(\theta\mid d)}{\partial\theta_i} = \frac{c(w_i,d)}{\theta_i} + \lambda = 0 \quad\rightarrow\quad \theta_i = -\frac{c(w_i,d)}{\lambda}$$

$$\sum_{i=1}^{M} -\frac{c(w_i,d)}{\lambda} = 1 \rightarrow \lambda = -\sum_{i=1}^{N} c(w_i,d) \rightarrow \hat{\theta}_i = p(w_i\mid\hat{\theta}) = \frac{c(w_i,d)}{\sum_{i=1}^{M} c(w_i,d)} = \frac{c(w_i,d)}{|d|}$$

10:24 / 12:21

What Does the Topic Look Like?

$p(w|\ \theta)$

d

Text mining paper

the 0.031
a 0.018
…
text 0.04
mining 0.035
association 0.03
clustering 0.005
computer 0.0009
…
food 0.000001
…

Can we get rid of these common words?

11:54 / 12:21