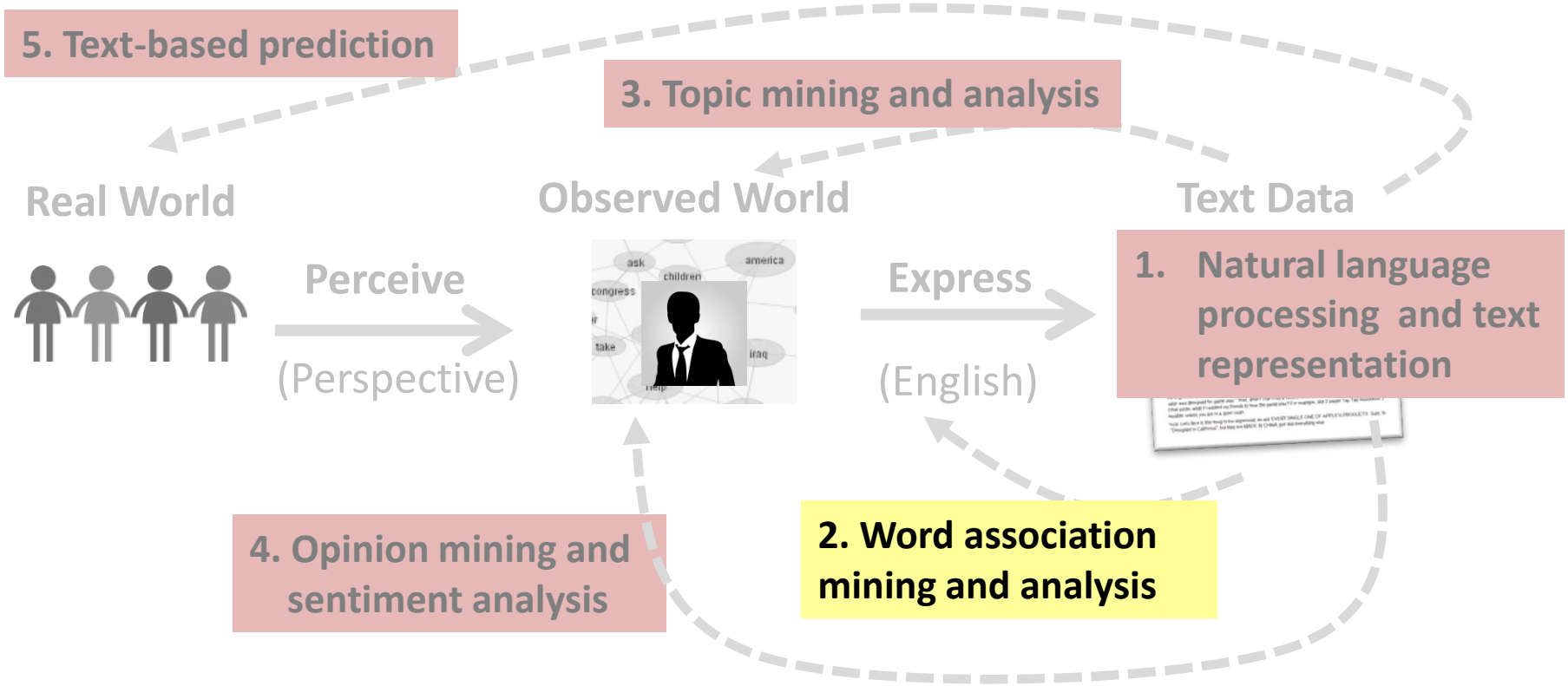
Syntagmatic Relation Discovery: Entropy

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

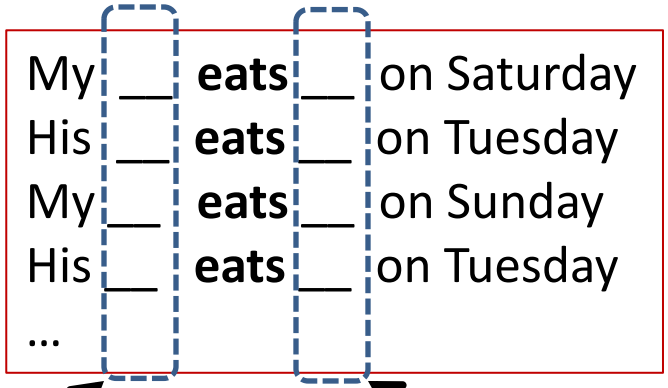
Syntagmatic Relation Discovery: Entropy



Syntagmatic Relation = Correlated Occurrences

Whenever "**eats**" occurs, what **other words** also tend to occur?

My cat **eats** fish on Saturday
His cat **eats** turkey on Tuesday
My dog **eats** meat on Sunday
His dog **eats** turkey on Tuesday
…



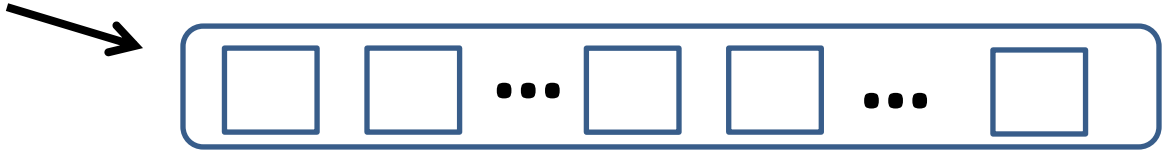
What words tend to occur to the **left** of "**eats**"**?**

What words are to the **right?**

Word Prediction: Intuition

Prediction Question: Is word **W** present (or absent) in this segment?

Text Segment (any unit, e.g., sentence, paragraph, document)



Are some words easier to predict than others?

1) W = “meat” **2) W=“the”** **3) W=“unicorn”**

Word Prediction: Formal Definition

Binary Random Variable : $X_w \in \{0, 1\}$

$$X_w = \begin{cases} 1 & w \text{ is present} \\ 0 & w \text{ is absent} \end{cases}$$

$$p(X_w = 1) + p(X_w = 0) = 1$$

The more random X_w is, the more difficult the prediction would be.

How does one quantitatively measure the "randomness" of a random variable like Xw?

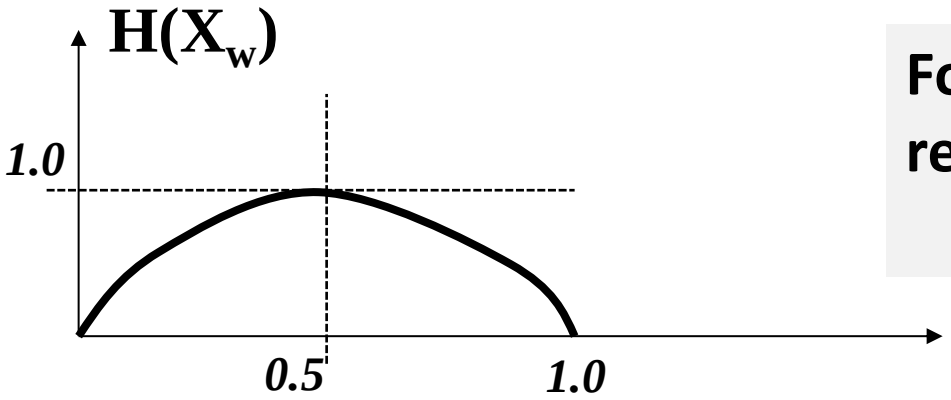
Entropy H(X) Measures Randomness of X

$$H(X_w) = \sum_{v \in \{0,1\}} -p(X_w = v)\log_2 p(X_w = v)$$

$$= -p(X_w = 0)\log_2 p(X_w = 0) - p(X_w = 1)\log_2 p(X_w = 1)$$

$$X_w = \begin{cases} 1 & w \text{ is present} \\ 0 & w \text{ is absent} \end{cases}$$

Define $0 \log_2 0 = 0$



For what X_w , does $H(X_w)$ reach maximum/minimum?

E.g., $P(X_w=1)=1$? $P(X_w=1)=0.5$?

or equivalently P(Xw=0) (Why?)

Entropy H(X): Coin Tossing

$$H(X_{coin}) = -p(X_{coin}=0)\log_2 p(X_{coin}=0) - p(X_{coin}=1)\log_2 p(X_{coin}=1)$$

X_{coin}: tossing a coin

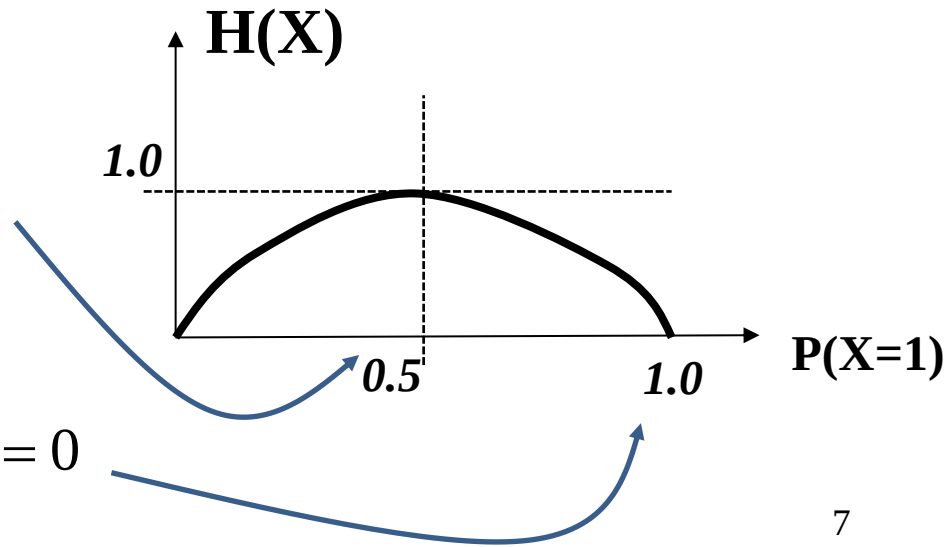
$$X_{coin} = \begin{cases} 1 & \text{Head} \\ 0 & \text{Tail} \end{cases}$$

Fair coin: p(X=1)=p(X=0)=1/2

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

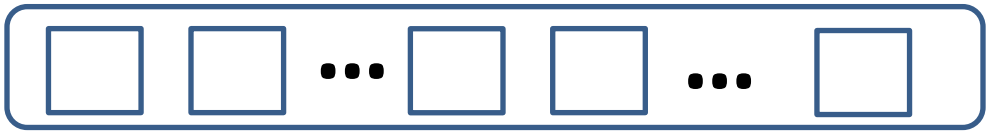
Completely biased: p(X=1)=1

$$H(X) = -0*\log_2 0 - 1*\log_2 1 = 0$$



Entropy for Word Prediction

Is word **W** present (or absent) in this segment?



1) W = "meat" 2) W = "the" 3) W = "unicorn"

Which is **high/low**? **$H(X_{meat})$, $H(X_{the})$, or $H(X_{unicorn})$?**

$H(X_{the}) \approx 0$ ➔ no uncertainty since $p(X_{the}=1) \approx 1$

High entropy words are harder to predict!

Syntagmatic Relation Discovery: Conditional Entropy

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

Syntagmatic Relation Discovery: Conditional Entropy

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

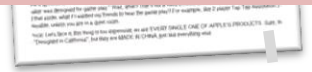
Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation



4. Opinion mining and sentiment analysis

2. Word association mining and analysis

What If We Know More About a Text Segment?

Prediction question: Is "**meat**" present (or absent) in this segment?

Does presence of "**eats**" help predict the presence of "**meat**"?
Does it **reduce** the uncertainty about "meat", i.e., **H(X_{meat})?**

What if we know of the absence of "eats"? Does it also help?

Conditional Entropy

Know nothing about the segment → **Know "eats" is present (X_{eats} =1)**

$$p(X_{meat} = 1) \dashrightarrow p(X_{meat} = 1 \mid X_{eats} = 1)$$

$$p(X_{meat} = 0) \dashrightarrow p(X_{meat} = 0 \mid X_{eats} = 1)$$

$$H(X_{meat}) = -p(X_{meat} = 0)\log_2 p(X_{meat} = 0) - p(X_{meat} = 1)\log_2 p(X_{meat} = 1)$$

$$H(X_{meat} \mid X_{eats} = 1) = -p(X_{meat} = 0 \mid X_{eats} = 1)\log_2 p(X_{meat} = 0 \mid X_{eats} = 1)$$
$$- p(X_{meat} = 1 \mid X_{eats} = 1)\log_2 p(X_{meat} = 1 \mid X_{eats} = 1)$$

$H(X_{meat} \mid X_{eats} = 0)$ **can be defined similarly**

Conditional Entropy: Complete Definition

$$H(X_{meat} \mid X_{eats}) = \sum_{u \in \{0,1\}} [p(X_{eats} = u)\ H(X_{meat} \mid X_{eats} = u)]$$

$$= \sum_{u \in \{0,1\}} [p(X_{eats} = u) \sum_{v \in \{0,1\}} [-p(X_{meat} = v \mid X_{eats} = u) \log_2 p(X_{meat} = v \mid X_{eats} = u)]]$$

In general, for any discrete random variables X and Y, we have H(X) ≥ H(X|Y)

What's the minimum possible value of H(X|Y)?

Conditional Entropy to Capture Syntagmatic Relation

$$H(X_{meat} \mid X_{eats}) = \sum_{u \in \{0,1\}} [p(X_{eats} = u)\ H(X_{meat} \mid X_{eats} = u)]$$

$$H(X_{meat} \mid X_{meat}) = ?$$

Which is smaller? $H(X_{meat} \mid X_{the})$ or $H(X_{meat} \mid X_{eats})$?

For which word w, does $H(X_{meat} \mid X_w)$ reach its minimum (i.e., 0)?

For which word w, does $H(X_{meat} \mid X_w)$ reach its maximum, $H(X_{meat})$?

Conditional Entropy for Mining Syntagmatic Relations

- For each word W1
  - For every other word W2, compute conditional entropy $H(X_{W1}|X_{W2})$
  - Sort all the candidate words in ascending order of $H(X_{W1}|X_{W2})$
  - Take the top-ranked candidate words as words that have potential syntagmatic relations with W1
  - Need to use a threshold for each W1
- However, while $H(X_{W1}|X_{W2})$ and $H(X_{W1}|X_{W3})$ are comparable, $H(X_{W1}|X_{W2})$ and $H(X_{W3}|X_{W2})$ aren't!

How can we mine the strongest K syntagmatic relations from a collection?

Syntagmatic Relation Discovery: Mutual Information

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

Syntagmatic Relation Discovery: Mutual Information

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

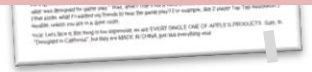
Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation



4. Opinion mining and sentiment analysis

2. Word association mining and analysis

Mutual Information I(X;Y): Measuring Entropy Reduction

How much reduction in the entropy of X can we obtain by knowing Y?

Mutual Information: $\mathbf{I(X; Y)= H(X) – H(X|Y) = H(Y)-H(Y|X)}$

Properties:

- Non-negative: $I(X;Y)\geq 0$
- Symmetric: $I(X;Y)=I(Y;X)$
- $I(X;Y)=0$ iff X & Y are independent

When we fix X to rank different Ys, I(X;Y) and H(X|Y) give the same order but I(X;Y) allows us to compare different (X,Y) pairs.

Mutual Information I(X;Y) for Syntagmatic Relation Mining

Mutual Information: $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Whenever "**eats**" occurs, what **other words** also tend to occur?

Which **words** have high mutual information with "**eats**"?

$$I(X_{eats}; X_{meats}) = I(X_{meats}; X_{eats}) \quad > \quad I(X_{eats}; X_{the}) = I(X_{the}; X_{eats})$$

$$I(X_{eats}; X_{eats}) = H(X_{eats}) \geq I(X_{eats}; X_w)$$

Rewriting Mutual Information (MI) Using KL-divergence

The observed joint distribution of X_{W1} and X_{W2}

$$I(X_{w1};X_{w2}) = \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} p(X_{w1}=u, X_{w2}=v) \log_2 \frac{p(X_{w1}=u, X_{w2}=v)}{p(X_{w1}=u)p(X_{w2}=v)}$$

The expected joint distribution of X_{W1} and X_{W2} if X_{W1} and X_{W2} were independent

MI measures the divergence of the actual joint distribution from the expected distribution under the independence assumption. The larger the divergence is, the higher the MI would be.

Probabilities Involved in Mutual Information

$$I(X_{w1};X_{w2}) = \sum_{u\in\{0,1\}}\sum_{v\in\{0,1\}} p(X_{w1}=u, X_{w2}=v)\log_2\frac{p(X_{w1}=u, X_{w2}=v)}{p(X_{w1}=u)p(X_{w2}=v)}$$

Presence & absence of w1: $p(X_{W1}=1) + p(X_{W1}=0) = 1$

Presence & absence of w2: $p(X_{W2}=1) + p(X_{W2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0) + p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = 1$$

Both w1 & w2 occur

Only w1 occurs

Only w2 occurs

None of them occurs

Relations Between Different Probabilities

Presence & absence of w1: $p(X_{w1}=1) + p(X_{w1}=0) = 1$

Presence & absence of w2: $p(X_{w2}=1) + p(X_{w2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = 1$$

Constraints:

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=1, X_{w2}=0) = p(X_{w1}=1)$$

$$p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = p(X_{w1}=0)$$

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=0, X_{w2}=1) = p(X_{w2}=1)$$

$$p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=0) = p(X_{w2}=0)$$

Computation of Mutual Information

Presence & absence of w1: $p(X_{W1}=1) + p(X_{W1}=0) = 1$

Presence & absence of w2: $p(X_{W2}=1) + p(X_{W2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0) + p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = 1$$

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0) = p(X_{W1}=1)$$

$$p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = p(X_{W1}=0)$$

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=0, X_{W2}=1) = p(X_{W2}=1)$$

$$p(X_{W1}=1, X_{W2}=0) + p(X_{W1}=0, X_{W2}=0) = p(X_{W2}=0)$$

We only need to know $p(X_{W1}=1)$, $p(X_{W2}=1)$, and $p(X_{W1}=1, X_{W2}=1)$.

Estimation of Probabilities (Depending on the Data)

$$p(X_{w1} = 1) = \frac{\text{count}(w1)}{N}$$

$$p(X_{w2} = 1) = \frac{\text{count}(w2)}{N}$$

$$p(X_{w1} = 1, X_{w2} = 1) = \frac{\text{count}(w1, w2)}{N}$$

	W1	W2	
Segment_1	**1**	**0**	Only W1 occurred
Segment_2	**1**	**1**	Both occurred
Segment_3	**1**	**1**	Both occurred
Segment_4	**0**	**0**	Neither occurred
…			
Segment_N	**0**	**1**	Only W2 occurred

Count(w1) = total number segments that contain W1
Count(w2) = total number segments that contain W2
Count(w1, w2) = total number segments that contain both W1 and W2

Smoothing: Accommodating Zero Counts

$$p(X_{w1}=1)=\frac{\text{count}(w1)+0.5}{N+1}$$

$$p(X_{w2}=1)=\frac{\text{count}(w2)+0.5}{N+1}$$

$$p(X_{w1}=1,X_{w2}=1)=\frac{\text{count}(w1,w2)+0.25}{N+1}$$

	W1	W2
¼ PseudoSeg_1	0	0
¼ PseudoSeg_2	1	0
¼ PseudoSeg_3	0	1
¼ PseudoSeg_4	1	1

Segment_1	1	0
...		
Segment_N	0	1

Actually observed data

Smoothing: Add pseudo data so that no event has zero counts (pretend we observed extra data)

Summary of Syntagmatic Relation Discovery

- Syntagmatic relation can be discovered by measuring correlations between occurrences of two words.
- Three concepts from Information Theory:
  - Entropy H(X): measures the uncertainty of a random variable X
  - Conditional entropy H(X|Y): entropy of X given we know Y
  - Mutual information I(X;Y): entropy reduction of X (or Y) due to knowing Y (or X)
- Mutual information provides a principled way for discovering syntagmatic relations.

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.

Additional 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, Brzil, 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.

Syntagmatic Relation Discovery: Mutual Information

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

Syntagmatic Relation Discovery: Mutual Information

5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

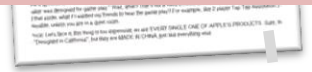
Perceive

(Perspective)

Express

(English)

1. Natural language processing and text representation



4. Opinion mining and sentiment analysis

2. Word association mining and analysis

Mutual Information I(X;Y): Measuring Entropy Reduction

How much reduction in the entropy of X can we obtain by knowing Y?

Mutual Information: $\mathbf{I(X; Y)= H(X) – H(X|Y) = H(Y)-H(Y|X)}$

Properties:

- Non-negative: $I(X;Y)\geq 0$
- Symmetric: $I(X;Y)=I(Y;X)$
- $I(X;Y)=0$ iff X & Y are independent

When we fix X to rank different Ys, I(X;Y) and H(X|Y) give the same order but I(X;Y) allows us to compare different (X,Y) pairs.

Mutual Information I(X;Y) for Syntagmatic Relation Mining

Mutual Information: $I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

Whenever "**eats**" occurs, what **other words** also tend to occur?

Which **words** have high mutual information with "**eats**"?

$$I(X_{eats}; X_{meats}) = I(X_{meats}; X_{eats}) \quad > \quad I(X_{eats}; X_{the}) = I(X_{the}; X_{eats})$$

$$I(X_{eats}; X_{eats}) = H(X_{eats}) \geq I(X_{eats}; X_w)$$

Rewriting Mutual Information (MI) Using KL-divergence

The observed joint distribution of X_{W1} and X_{W2}

$$I(X_{w1};X_{w2}) = \sum_{u\in\{0,1\}} \sum_{v\in\{0,1\}} p(X_{w1}=u, X_{w2}=v) \log_2 \frac{p(X_{w1}=u, X_{w2}=v)}{p(X_{w1}=u)p(X_{w2}=v)}$$

The expected joint distribution of X_{W1} and X_{W2} if X_{W1} and X_{W2} were independent

MI measures the divergence of the actual joint distribution from the expected distribution under the independence assumption. The larger the divergence is, the higher the MI would be.

Probabilities Involved in Mutual Information

$$I(X_{w1};X_{w2}) = \sum_{u\in\{0,1\}}\sum_{v\in\{0,1\}} p(X_{w1}=u, X_{w2}=v)\log_2\frac{p(X_{w1}=u, X_{w2}=v)}{p(X_{w1}=u)p(X_{w2}=v)}$$

Presence & absence of w1: $p(X_{W1}=1) + p(X_{W1}=0) = 1$

Presence & absence of w2: $p(X_{W2}=1) + p(X_{W2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0) + p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = 1$$

Both w1 & w2 occur

Only w1 occurs

Only w2 occurs

None of them occurs

Relations Between Different Probabilities

Presence & absence of w1: $p(X_{w1}=1) + p(X_{w1}=0) = 1$

Presence & absence of w2: $p(X_{w2}=1) + p(X_{w2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = 1$$

Constraints:

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=1, X_{w2}=0) = p(X_{w1}=1)$$

$$p(X_{w1}=0, X_{w2}=1) + p(X_{w1}=0, X_{w2}=0) = p(X_{w1}=0)$$

$$p(X_{w1}=1, X_{w2}=1) + p(X_{w1}=0, X_{w2}=1) = p(X_{w2}=1)$$

$$p(X_{w1}=1, X_{w2}=0) + p(X_{w1}=0, X_{w2}=0) = p(X_{w2}=0)$$

Computation of Mutual Information

Presence & absence of w1: $p(X_{W1}=1) + p(X_{W1}=0) = 1$

Presence & absence of w2: $p(X_{W2}=1) + p(X_{W2}=0) = 1$

Co-occurrences of w1 and w2:

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0) + p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = 1$$

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=1, X_{W2}=0) = p(X_{W1}=1)$$

$$p(X_{W1}=0, X_{W2}=1) + p(X_{W1}=0, X_{W2}=0) = p(X_{W1}=0)$$

$$p(X_{W1}=1, X_{W2}=1) + p(X_{W1}=0, X_{W2}=1) = p(X_{W2}=1)$$

$$p(X_{W1}=1, X_{W2}=0) + p(X_{W1}=0, X_{W2}=0) = p(X_{W2}=0)$$

We only need to know $p(X_{W1}=1)$, $p(X_{W2}=1)$, and $p(X_{W1}=1, X_{W2}=1)$.

Estimation of Probabilities (Depending on the Data)

$$p(X_{w1} = 1) = \frac{\text{count}(w1)}{N}$$

$$p(X_{w2} = 1) = \frac{\text{count}(w2)}{N}$$

$$p(X_{w1} = 1, X_{w2} = 1) = \frac{\text{count}(w1, w2)}{N}$$

	W1	W2	
Segment_1	**1**	**0**	Only W1 occurred
Segment_2	**1**	**1**	Both occurred
Segment_3	**1**	**1**	Both occurred
Segment_4	**0**	**0**	Neither occurred
…			
Segment_N	**0**	**1**	Only W2 occurred

Count(w1) = total number segments that contain W1
Count(w2) = total number segments that contain W2
Count(w1, w2) = total number segments that contain both W1 and W2

Smoothing: Accommodating Zero Counts

$$p(X_{w1}=1)=\frac{\text{count}(w1)+0.5}{N+1}$$

$$p(X_{w2}=1)=\frac{\text{count}(w2)+0.5}{N+1}$$

$$p(X_{w1}=1,X_{w2}=1)=\frac{\text{count}(w1,w2)+0.25}{N+1}$$

	W1	W2
¼ PseudoSeg_1	0	0
¼ PseudoSeg_2	1	0
¼ PseudoSeg_3	0	1
¼ PseudoSeg_4	1	1

Segment_1	1	0
...		
Segment_N	0	1

Actually observed data

Smoothing: Add pseudo data so that no event has zero counts (pretend we observed extra data)

Summary of Syntagmatic Relation Discovery

- Syntagmatic relation can be discovered by measuring correlations between occurrences of two words.
- Three concepts from Information Theory:
  - Entropy H(X): measures the uncertainty of a random variable X
  - Conditional entropy H(X|Y): entropy of X given we know Y
  - Mutual information I(X;Y): entropy reduction of X (or Y) due to knowing Y (or X)
- Mutual information provides a principled way for discovering syntagmatic relations.

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.

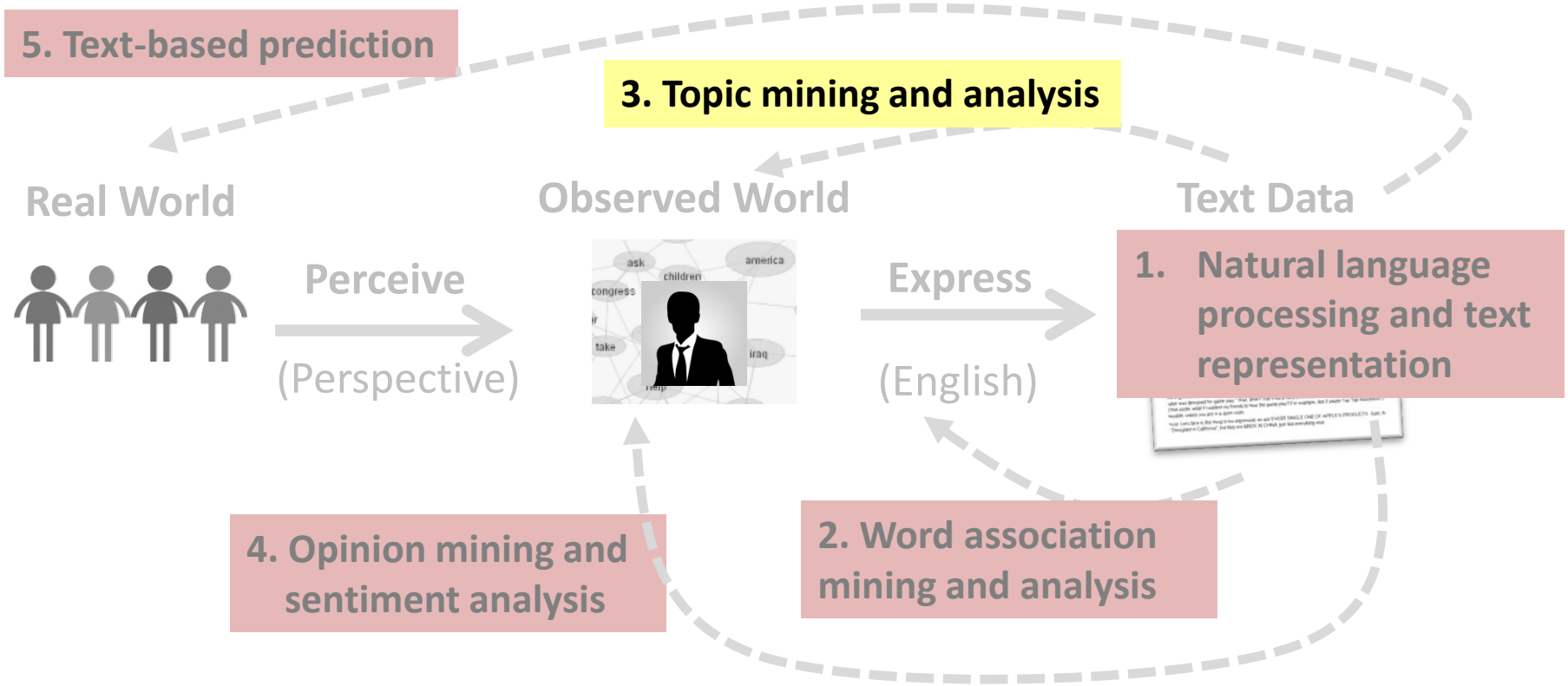
Additional Reading

- 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, Brzil, 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.

Topic Mining and Analysis: Motivation and Task Definition

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

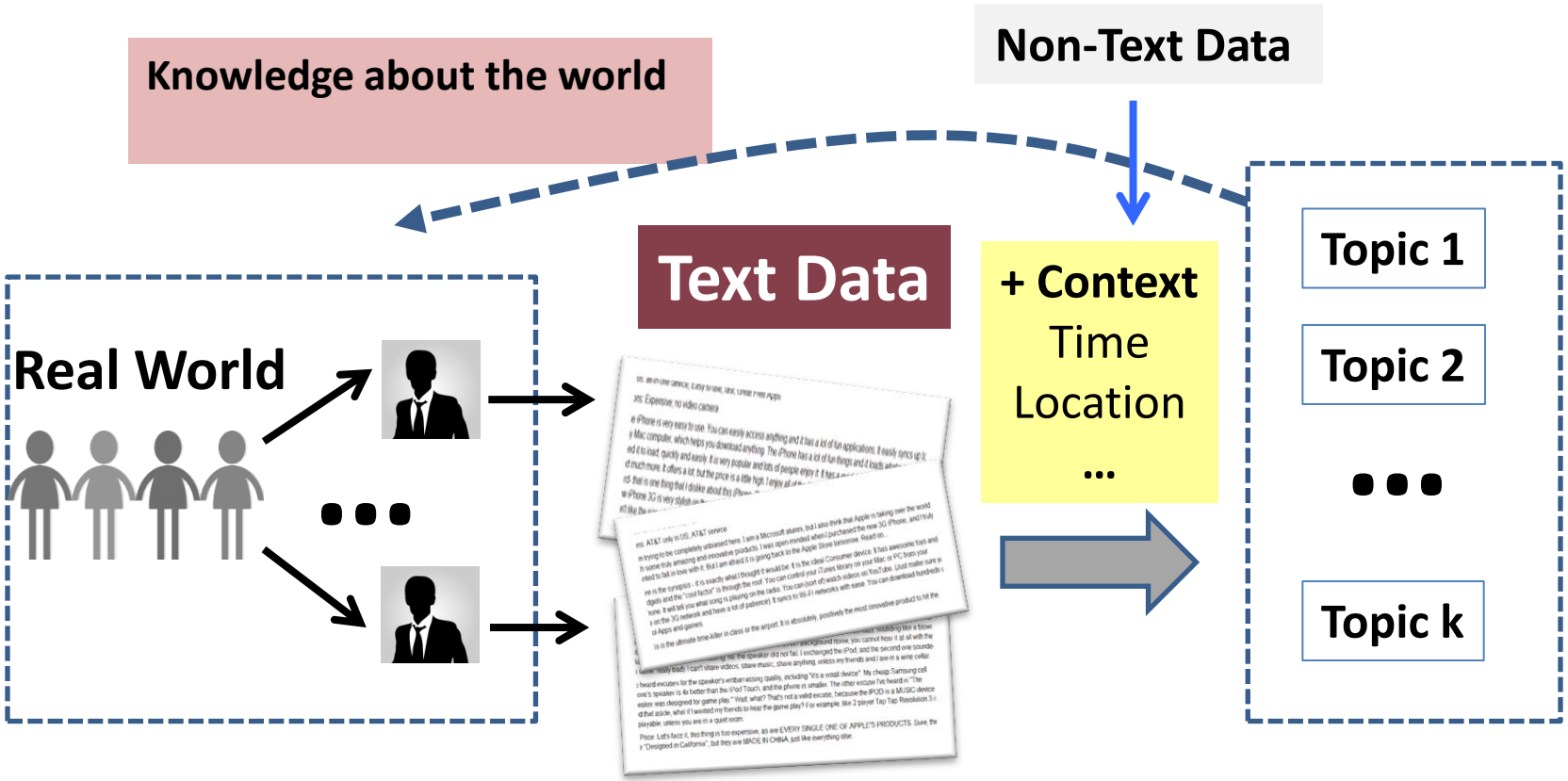
Topic Mining and Analysis: Motivation and Task Definition



Topic Mining and Analysis: Motivation

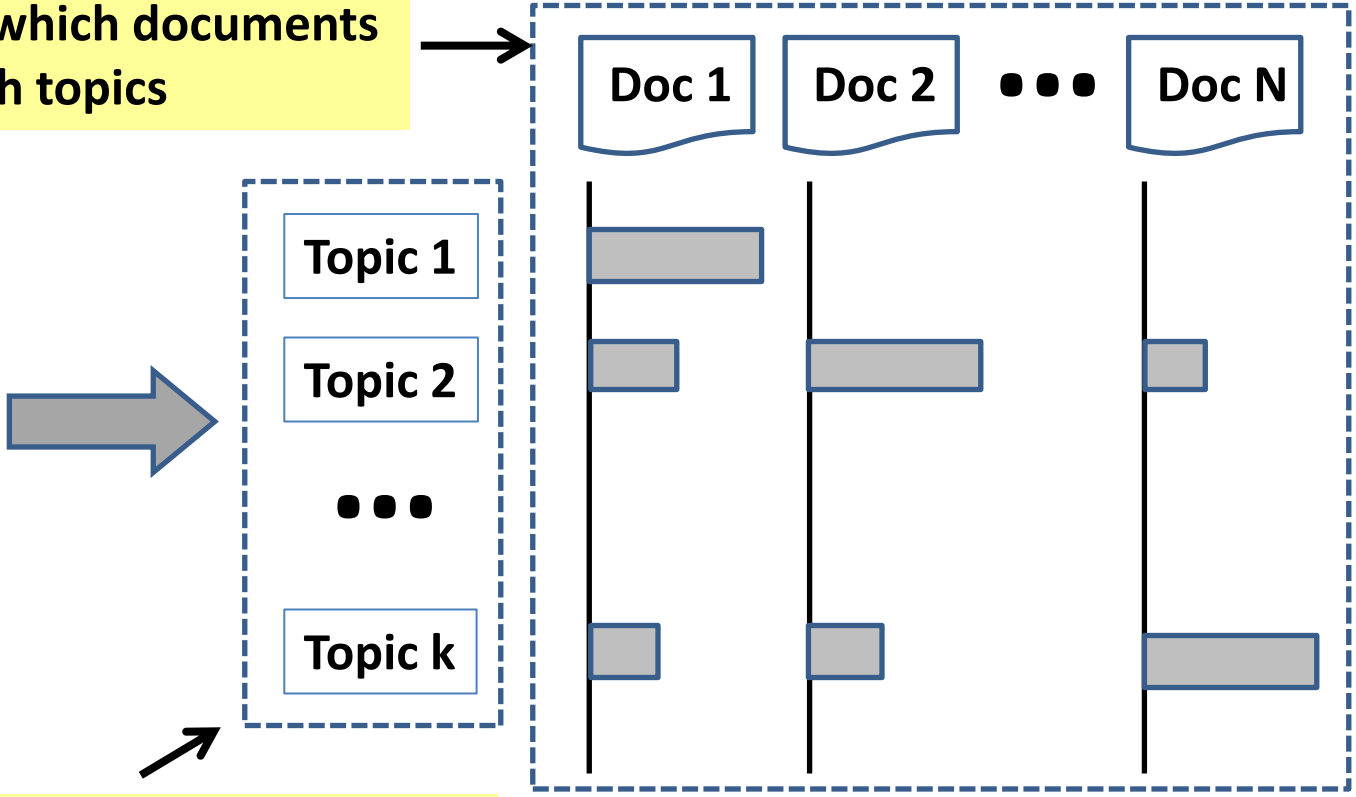
- Topic ≈ main idea discussed in text data
  - Theme/subject of a discussion or conversation
  - Different granularities (e.g., topic of a sentence, an article, etc.)
- Many applications require discovery of topics in text
  - What are Twitter users talking about today?
  - What are the current research 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



Tasks of Topic Mining and Analysis

Task 2: Figure out which documents cover which topics



Formal Definition of Topic Mining and Analysis

- Input
  - A **collection** of **N** text documents **C={d_1, …, d_N}**
  - **Number of topics**: **k**
- Output
  - **k topics**: **{ 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

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

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

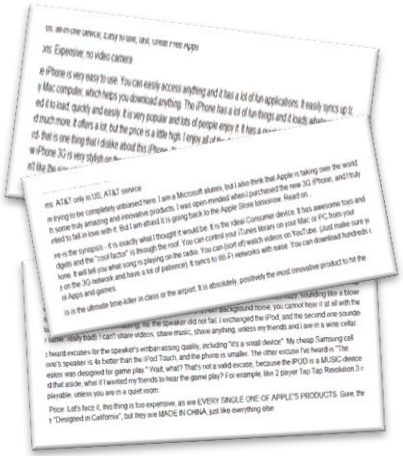
- Input
  - A **collection** of **N** text documents **C={d_1, …, d_N}**
  - **Number of topics**: **k**
- Output
  - **k topics**: **{ θ_1, …, θ_k }**
  - **Coverage of topics in each d_i**: **{ π_{i1}, …, π_{ik} }**
  - π_{ij}=prob. of d_i covering topic θ_j

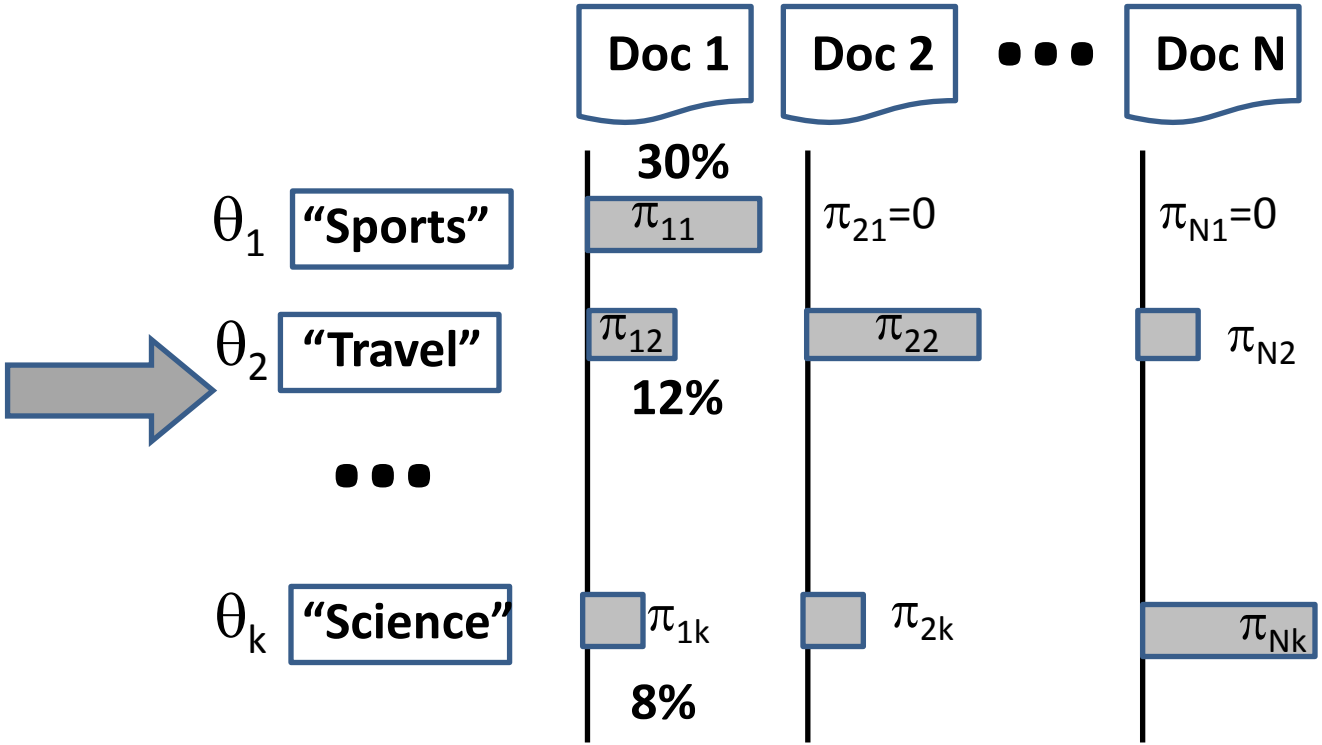
$$\sum_{j=1}^{k} \pi_{ij} = 1$$

How to define θ_i ?

Initial Idea: Topic = Term

Text Data

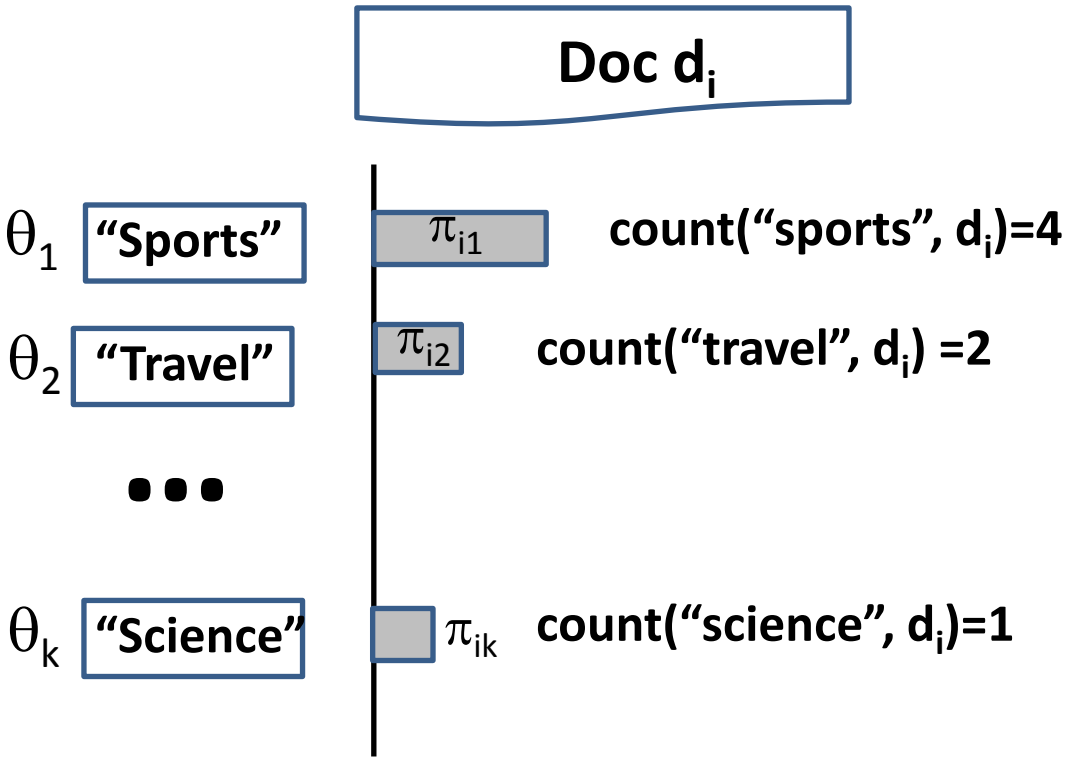




Mining k Topical Terms from Collection C

- 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}



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

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

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

…

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

1. Need to count related words also!

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

3. Mine complicated topics?

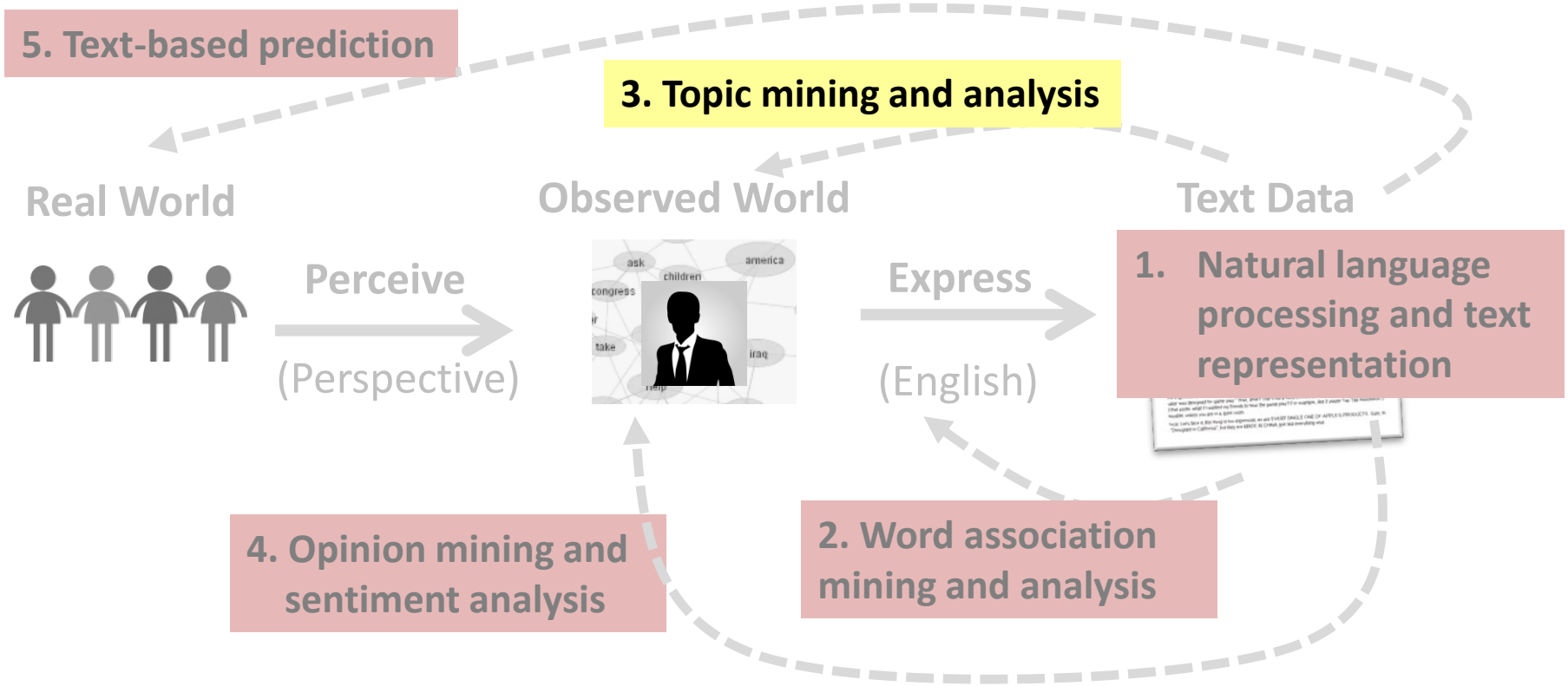
Problems with “Term as Topic”

- 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

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,....}

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

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

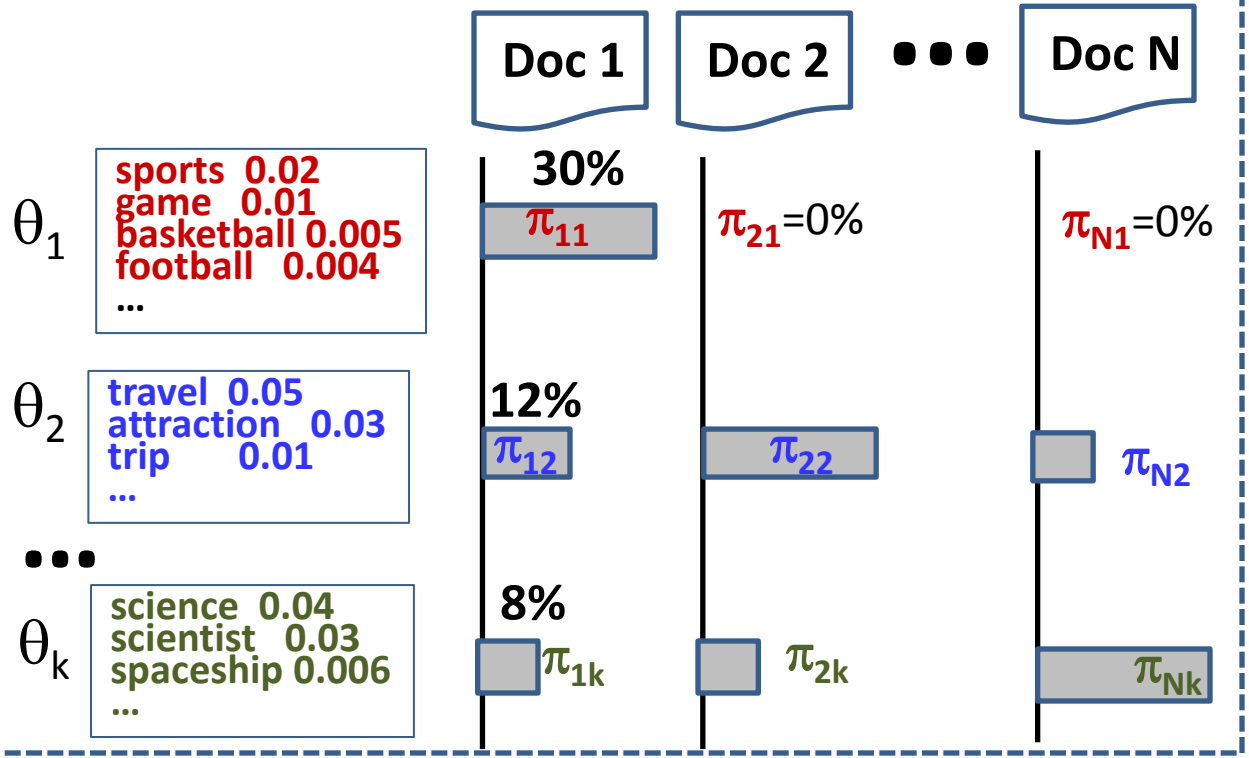
$$\sum_{j=1}^{k} \pi_{ij} = 1$$

The Computation Task

INPUT: C, k, V

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

Text Data



Generative Model for Text Mining

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

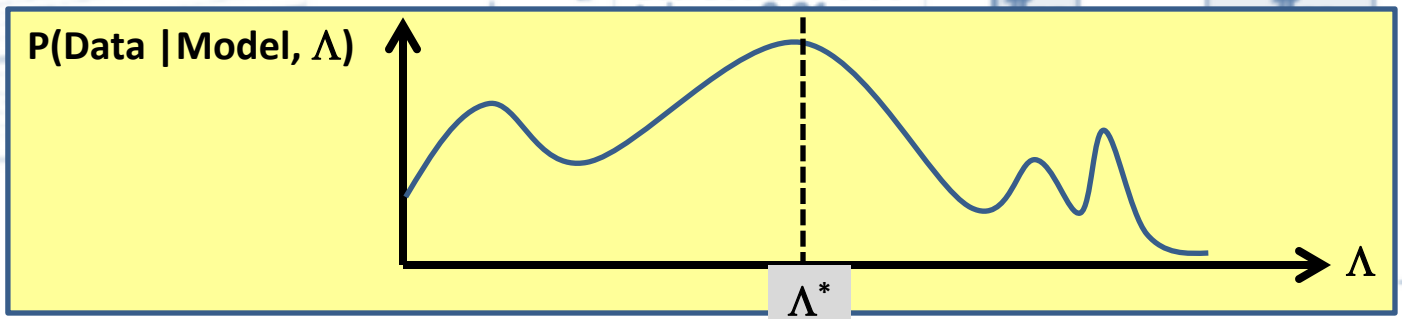
$\Lambda=(\{\theta_1, \ldots, \theta_k\}, \{\pi_{11}, \ldots, \pi_{1k}\}, \ldots, \{\pi_{N1}, \ldots, \pi_{Nk}\})$

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

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

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

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

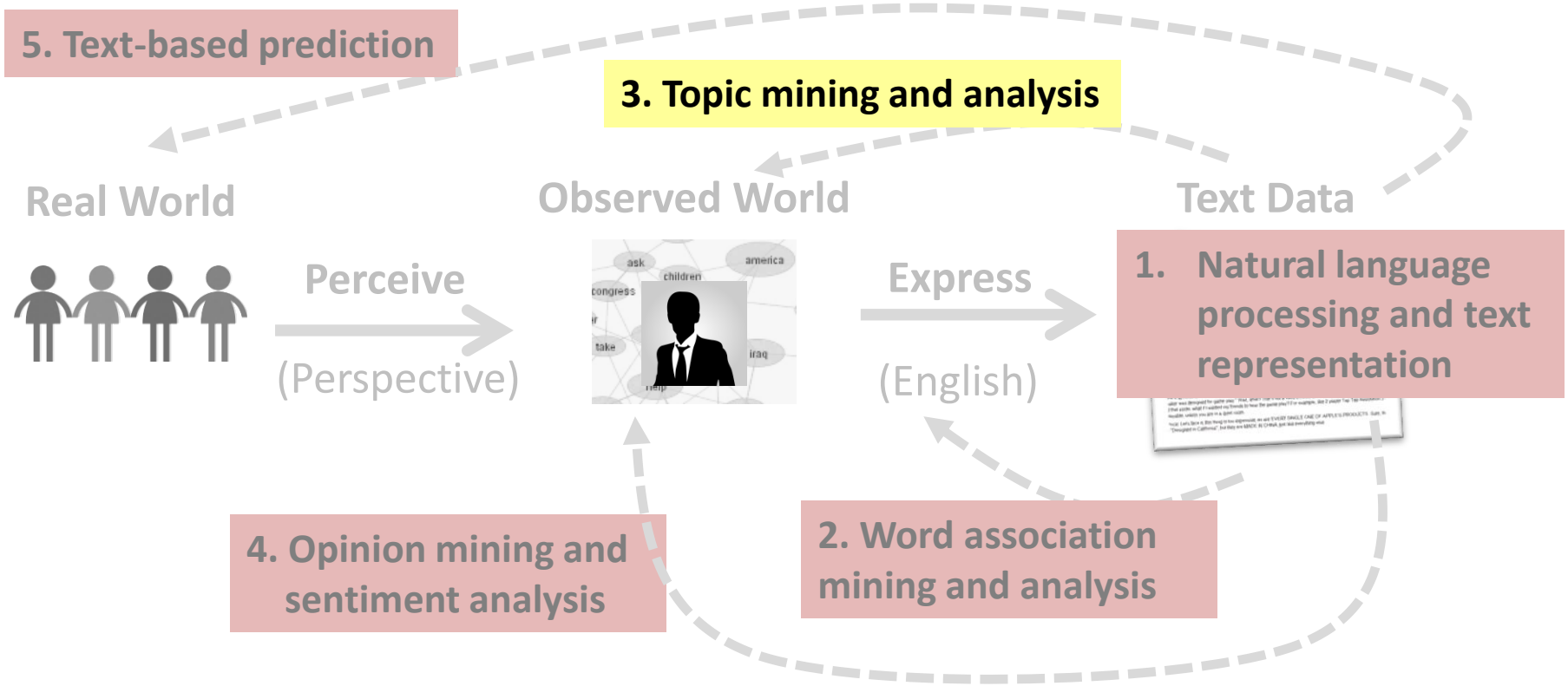
Summary (cont.)

- **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**:** $\Lambda^* = \text{argmax}_{\Lambda}\ p(\text{Data}|\ \text{Model}, \Lambda)$
  - **Take Λ^* as the "knowledge"** to be mined for the text mining problem
  - **Adjust** the design of the model to discover different knowledge

Topic Mining and Analysis: Overview of Statistical Language Models

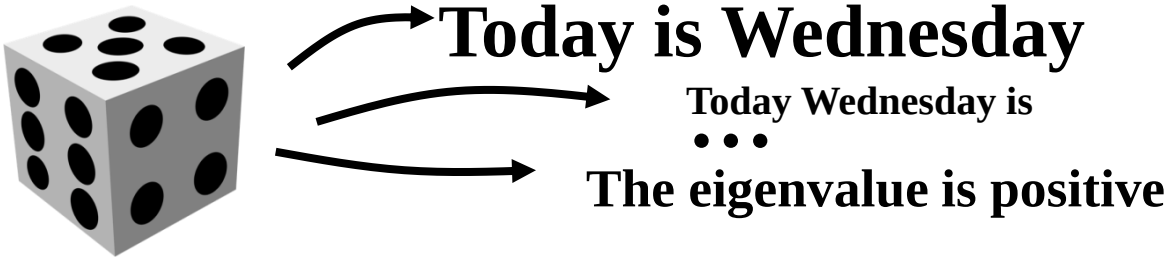
ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Overview of Statistical Language Models



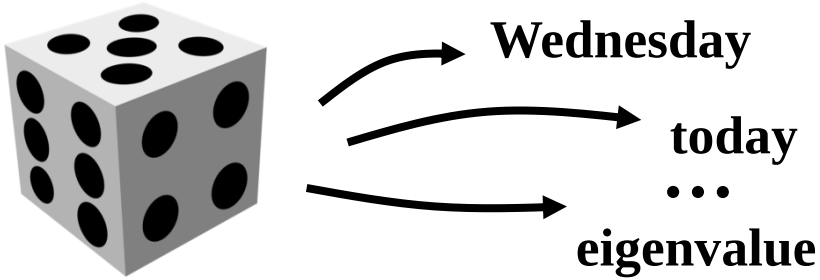
What Is a Statistical Language Model (LM)?

- 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



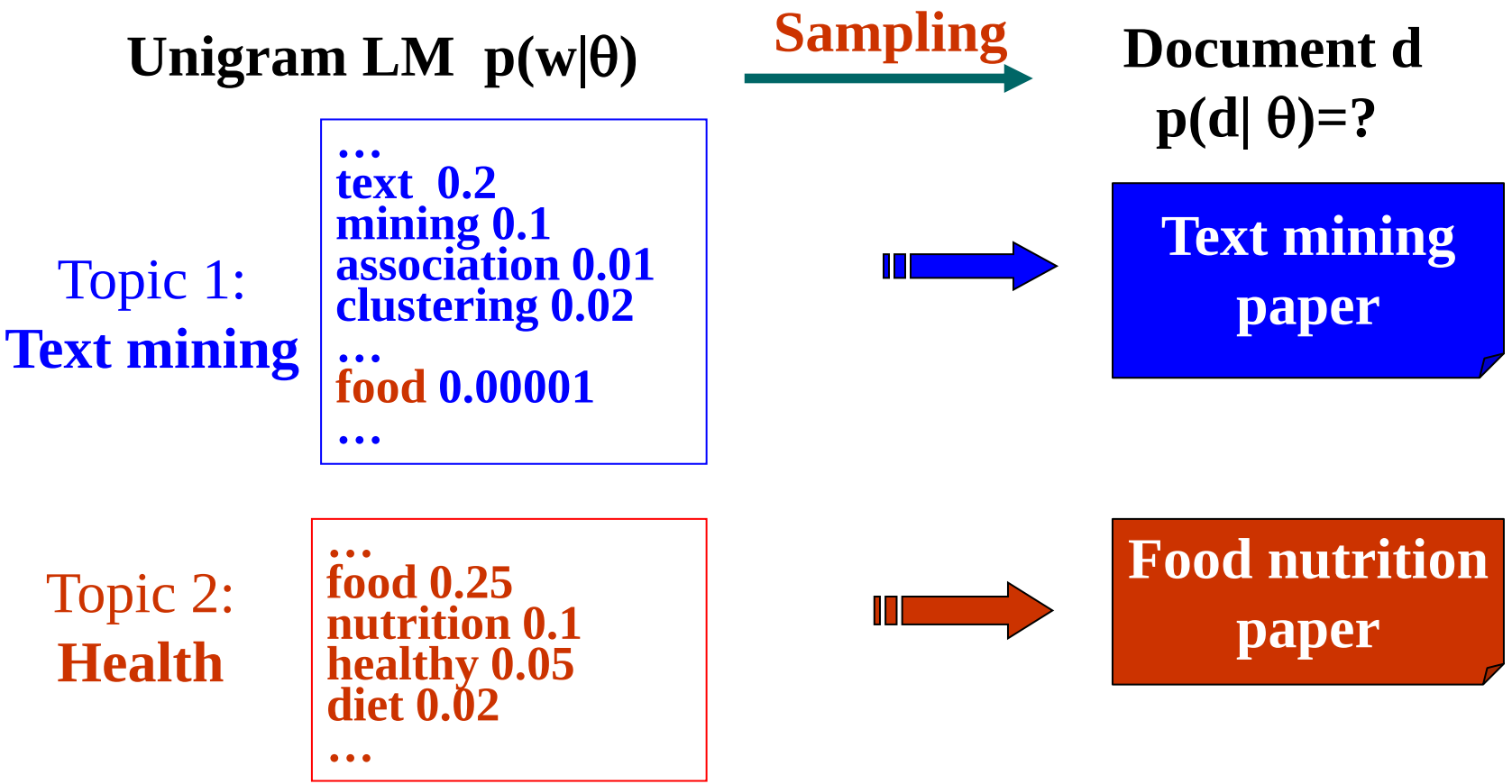
The Simplest Language Model: Unigram LM

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

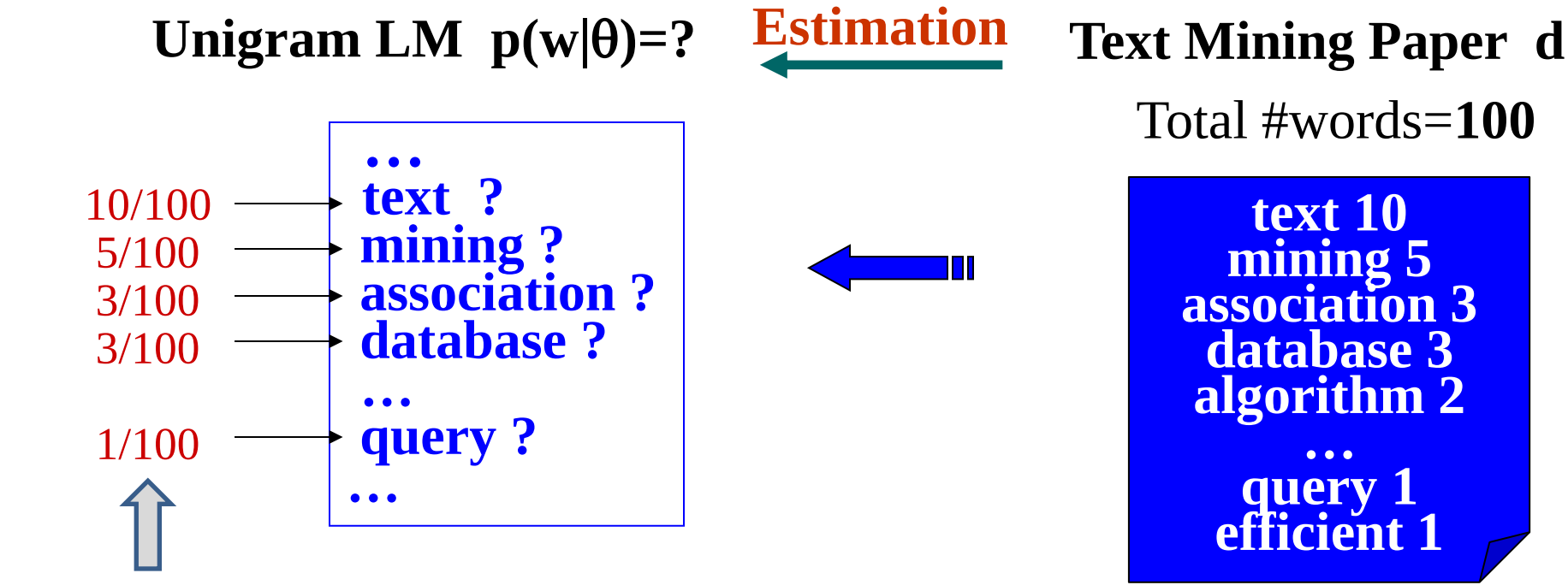


p("today is Wed")
= p("today")p("is")p("Wed")
= 0.0002 × 0.001 × 0.000015

Text Generation with Unigram LM



Estimation of Unigram LM



Is this our best estimate?

How do we define "best"?

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 $p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$

  - "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

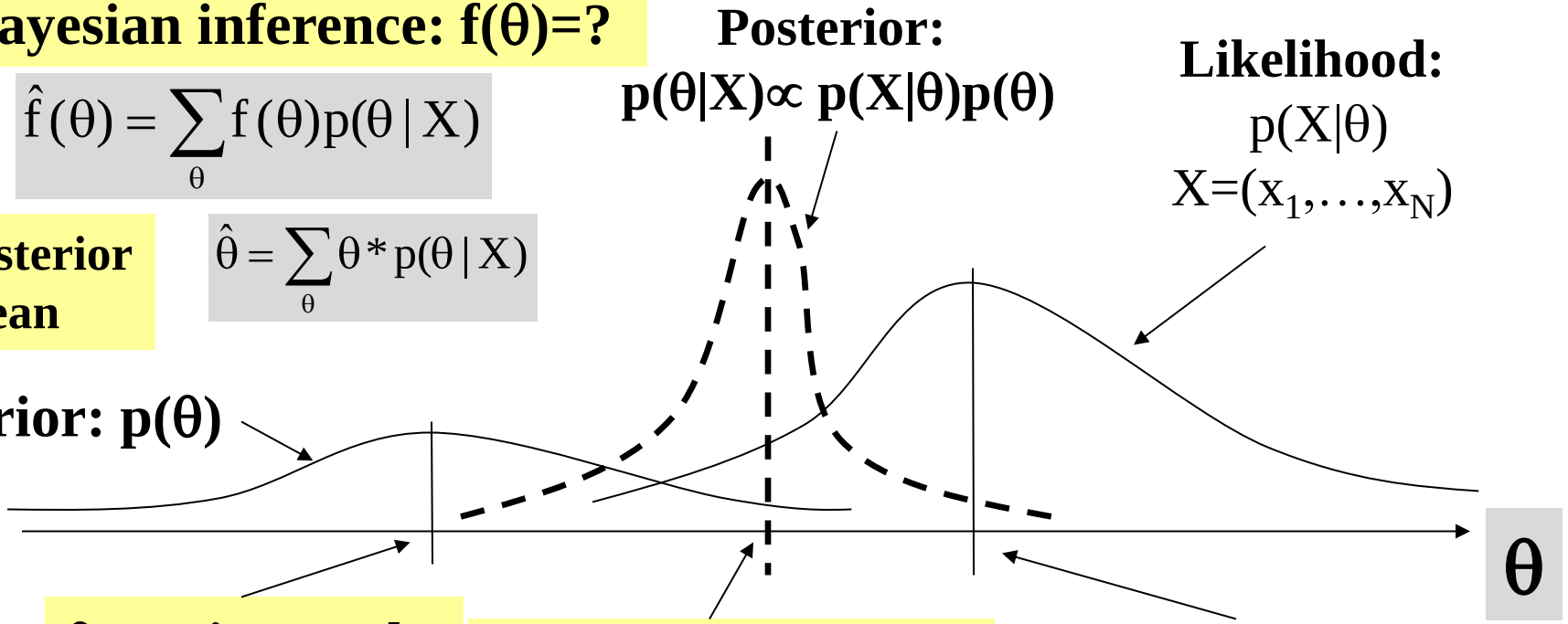
Illustration of Bayesian Estimation

Bayesian inference: f(θ)=?

$$\hat{f}(\theta) = \sum_{\theta} f(\theta) p(\theta \mid X)$$

Posterior Mean

$$\hat{\theta} = \sum_{\theta} \theta * p(\theta \mid X)$$



Summary

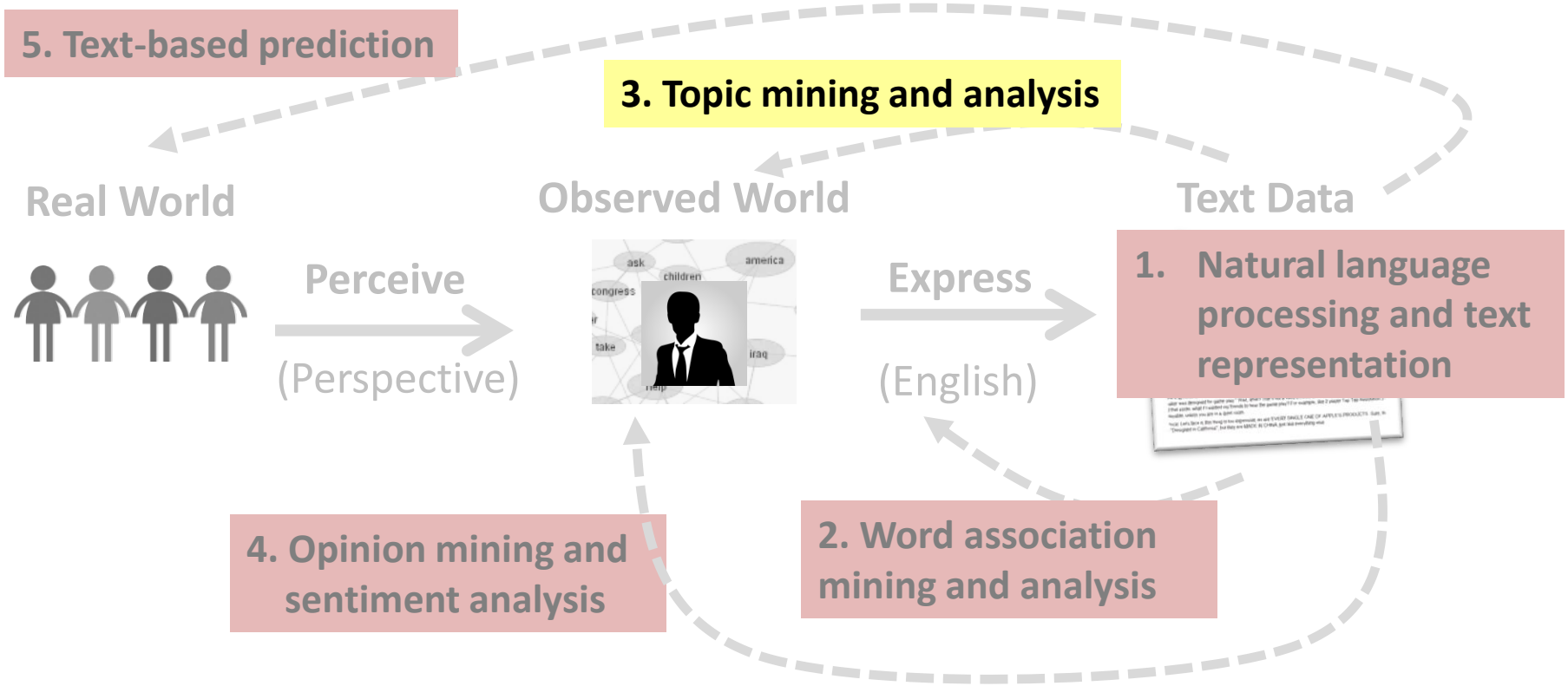
- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: **$p(X|\theta)$**
  - **Given θ ➔** which X has a higher likelihood?
  - **Given X ➔** which θ maximizes $p(X|\theta)$? **[ML estimate]**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

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

Topic Mining and Analysis: Overview of Statistical Language Models

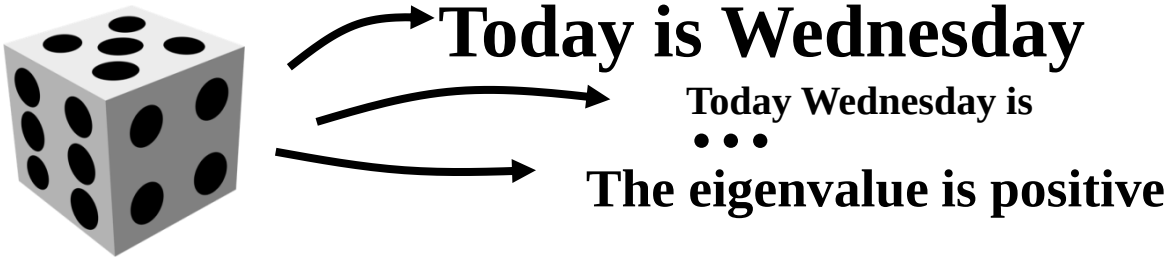
ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Overview of Statistical Language Models



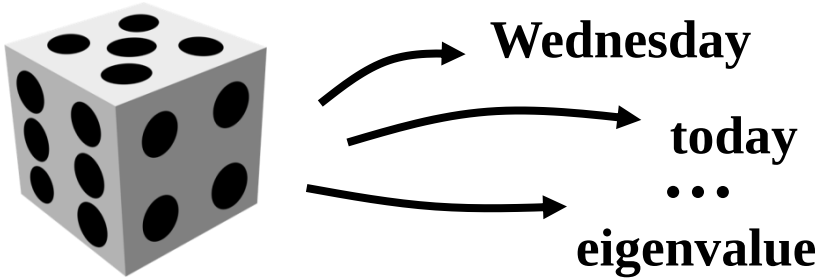
What Is a Statistical Language Model (LM)?

- 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



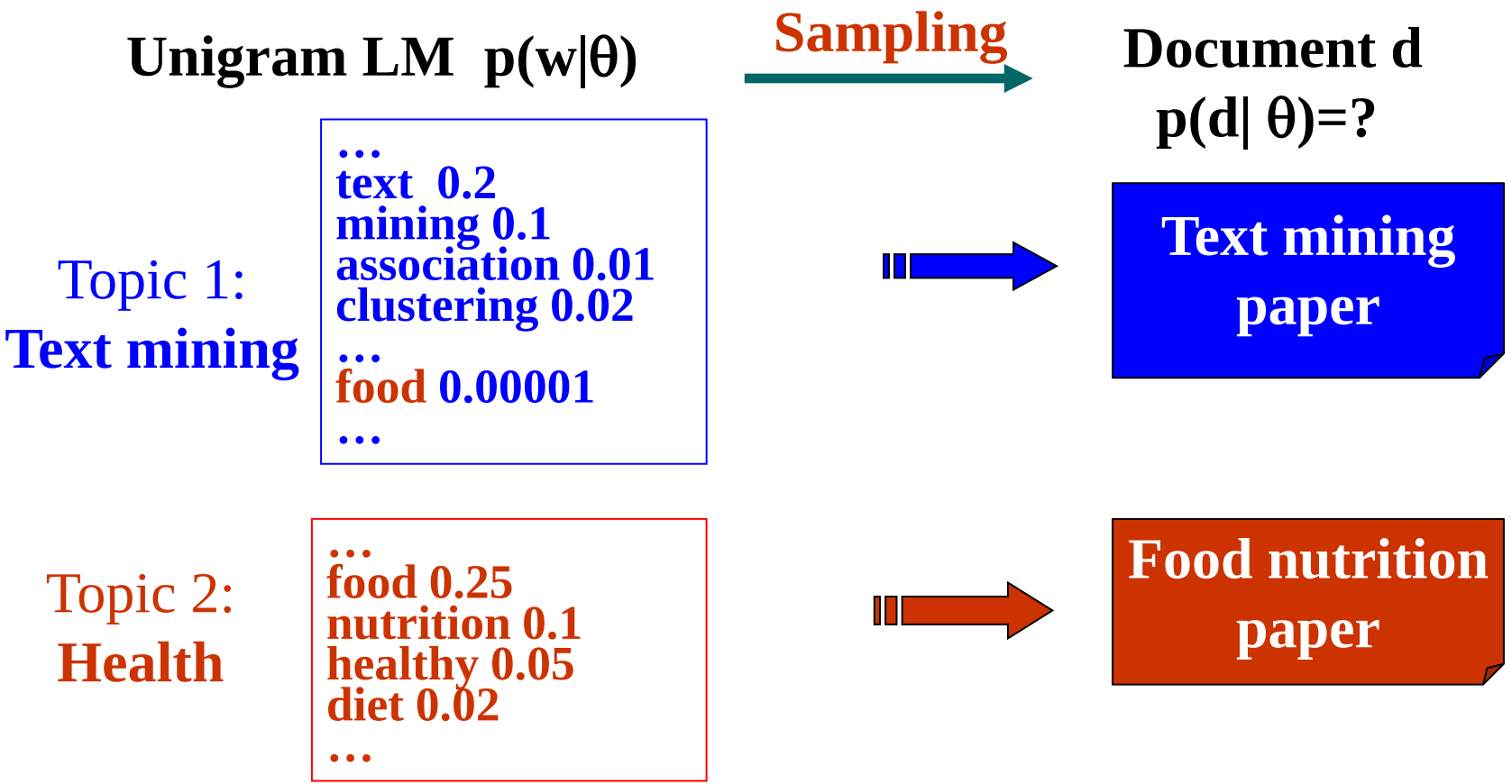
The Simplest Language Model: Unigram LM

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

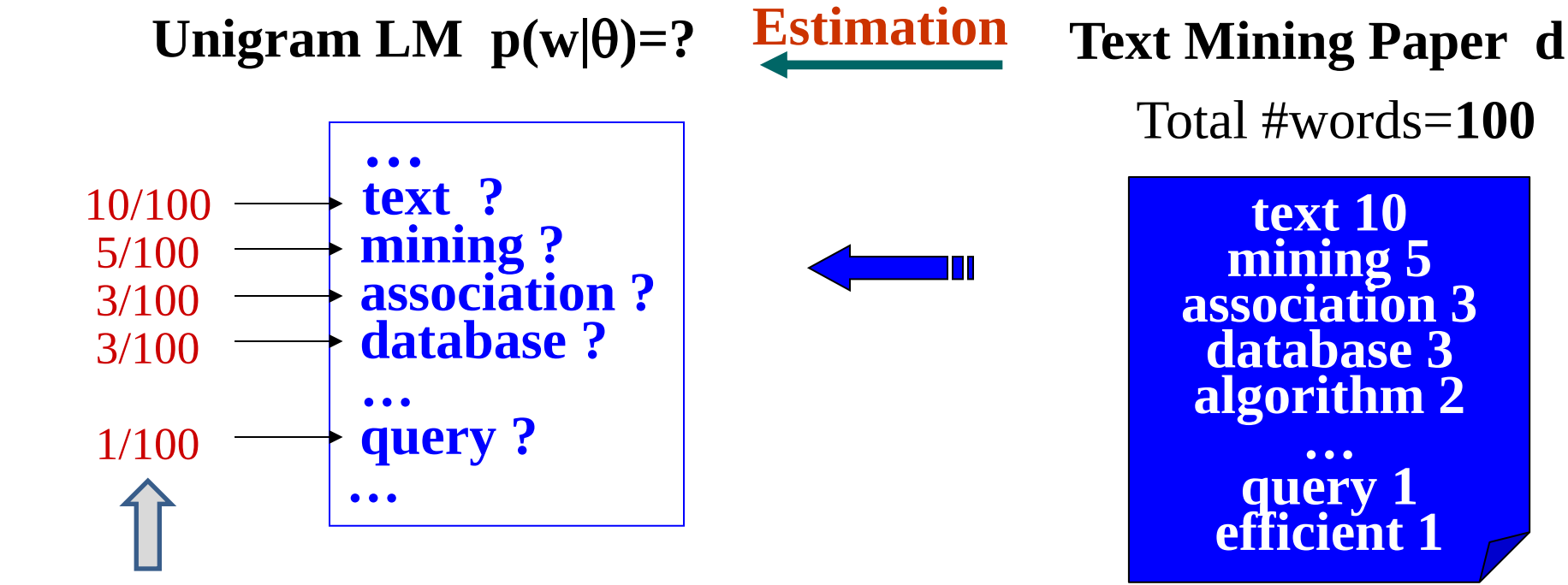


p("today is Wed")
= p("today")p("is")p("Wed")
= 0.0002 × 0.001 × 0.000015

Text Generation with Unigram LM



Estimation of Unigram LM



Is this our best estimate?

How do we define "best"?

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 $p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$

  - "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

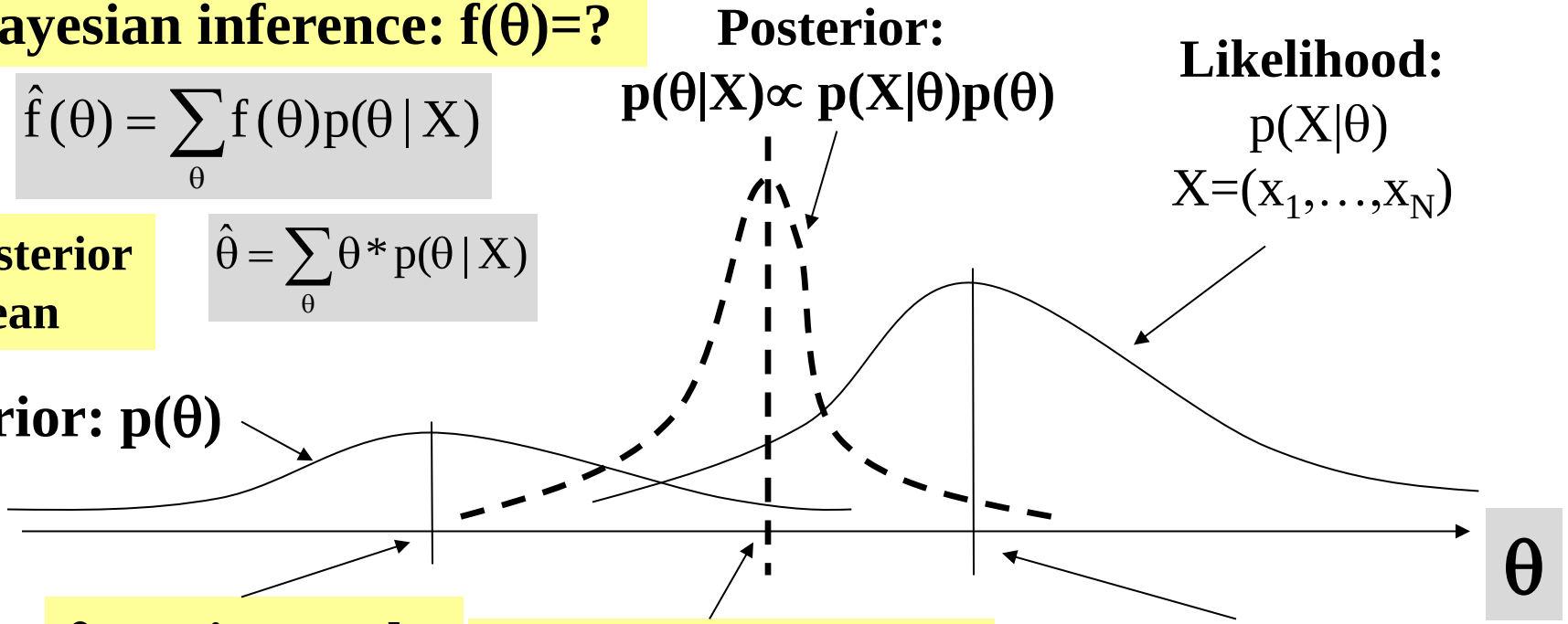
Illustration of Bayesian Estimation

Bayesian inference: f(θ)=?

$$\hat{f}(\theta) = \sum_{\theta} f(\theta) p(\theta \mid X)$$

Posterior Mean

$$\hat{\theta} = \sum_{\theta} \theta * p(\theta \mid X)$$



Summary

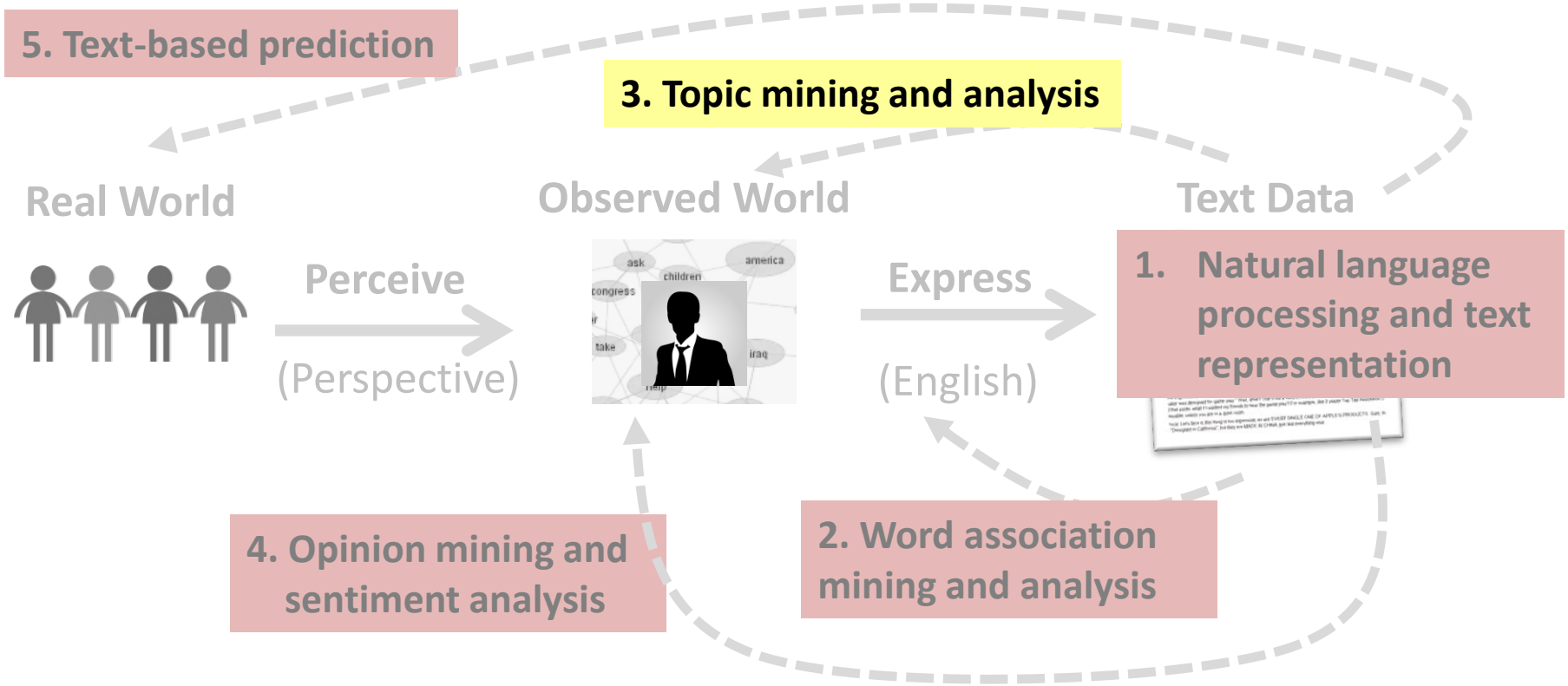
- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: **$p(X|\theta)$**
  - **Given θ ➔** which X has a higher likelihood?
  - **Given X ➔** which θ maximizes $p(X|\theta)$? **[ML estimate]**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

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

Topic Mining and Analysis: Overview of Statistical Language Models

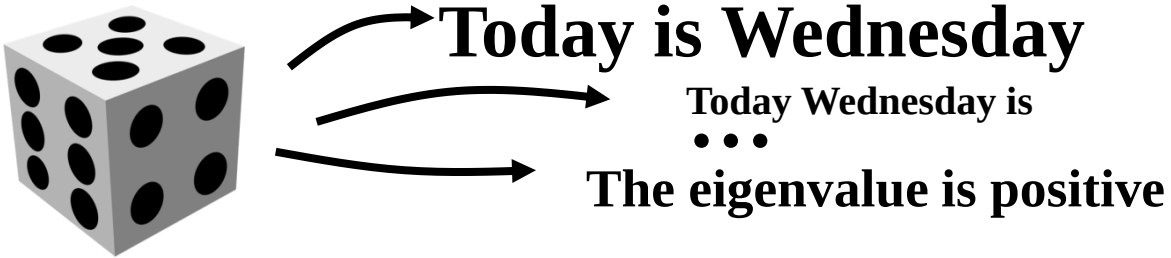
ChengXiang “Cheng” Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Probabilistic Topic Models: Overview of Statistical Language Models



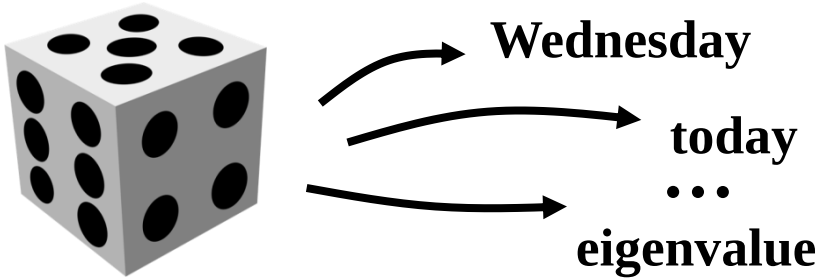
What Is a Statistical Language Model (LM)?

- 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



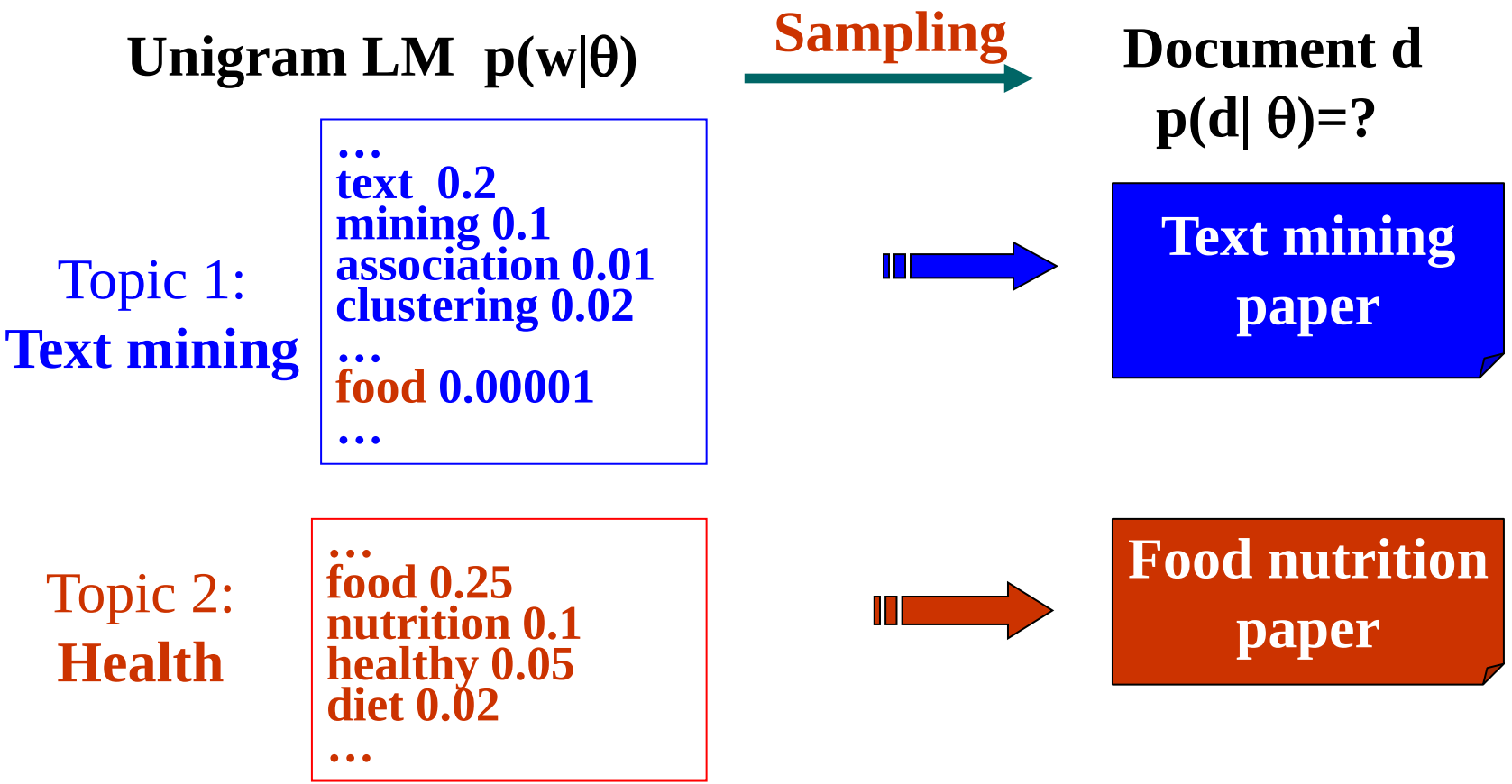
The Simplest Language Model: Unigram LM

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

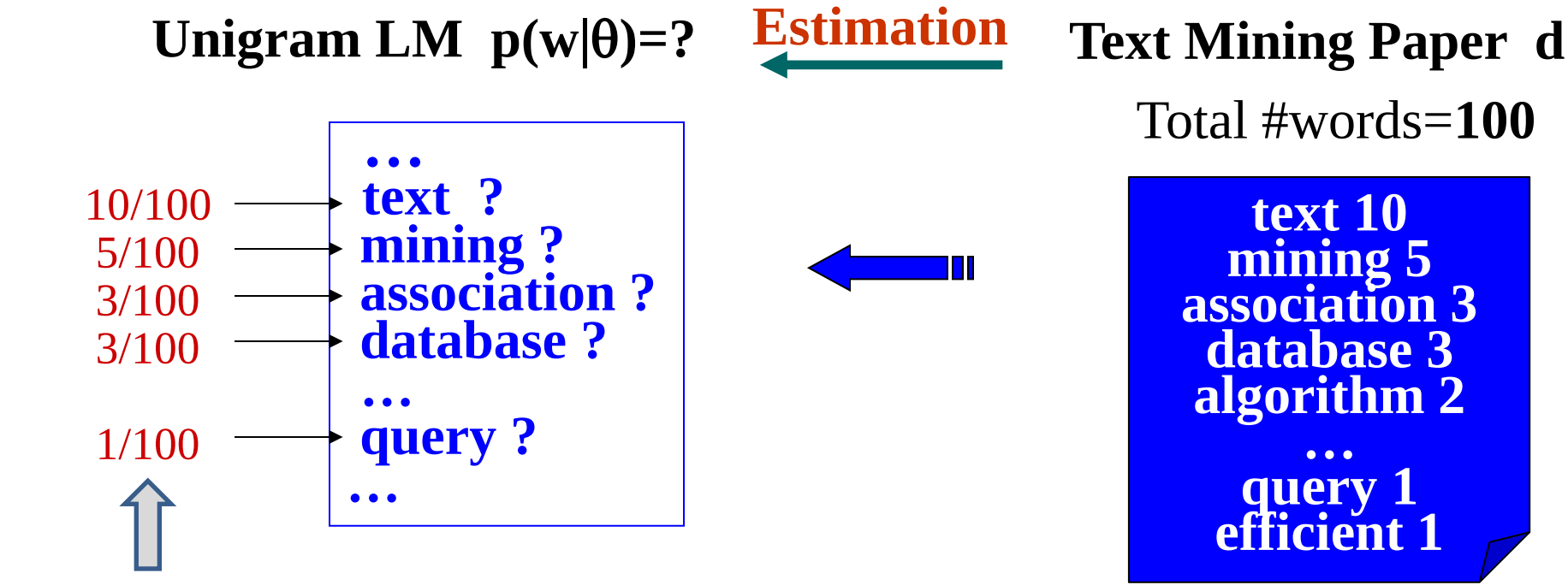


p("today is Wed")
= p("today")p("is")p("Wed")
= 0.0002 × 0.001 × 0.000015

Text Generation with Unigram LM



Estimation of Unigram LM



Is this our best estimate?

How do we define "best"?

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 $p(X \mid Y) = \frac{p(Y \mid X)p(X)}{p(Y)}$

  - "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

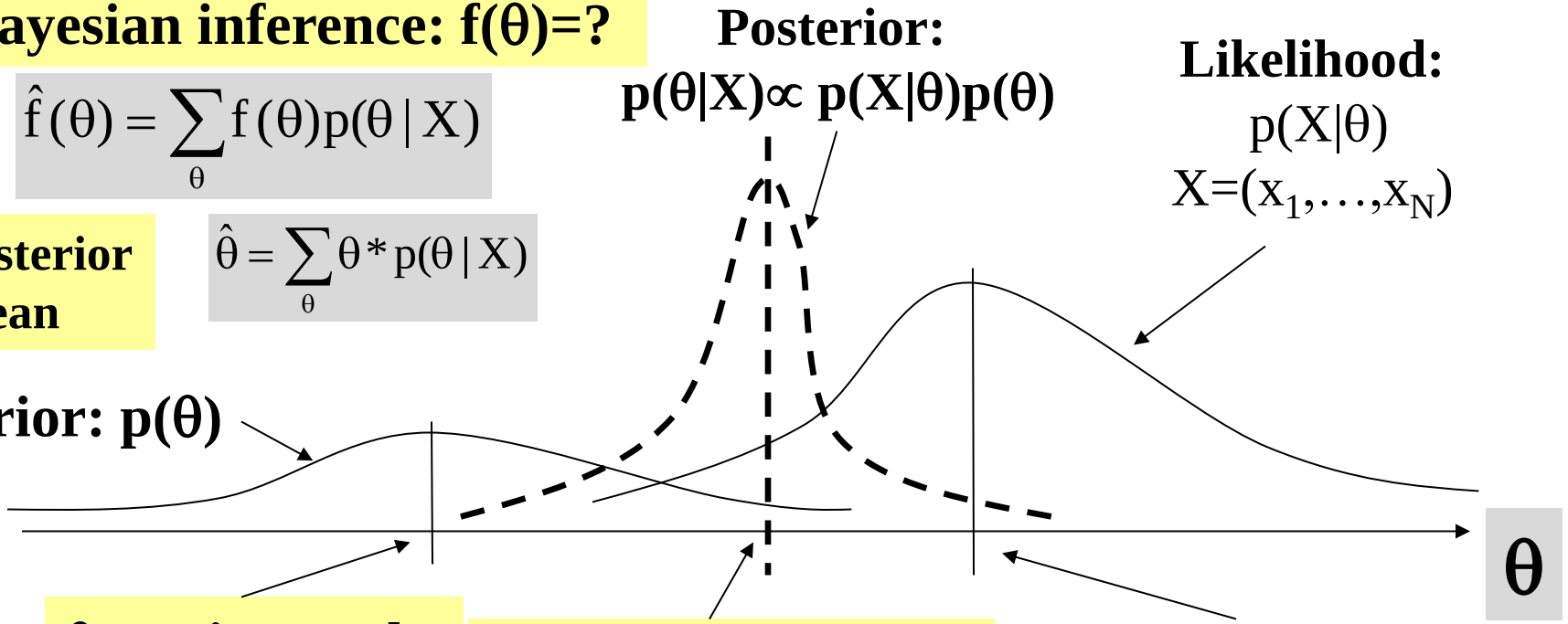
Illustration of Bayesian Estimation

Bayesian inference: f(θ)=?

$$\hat{f}(\theta) = \sum_{\theta} f(\theta) p(\theta \mid X)$$

Posterior Mean

$$\hat{\theta} = \sum_{\theta} \theta * p(\theta \mid X)$$



Summary

- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: **$p(X|\theta)$**
  - **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: $p(\theta)$**
  - **Posterior** distribution**: $p(\theta|X) \propto p(X|\theta)p(\theta)$**

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