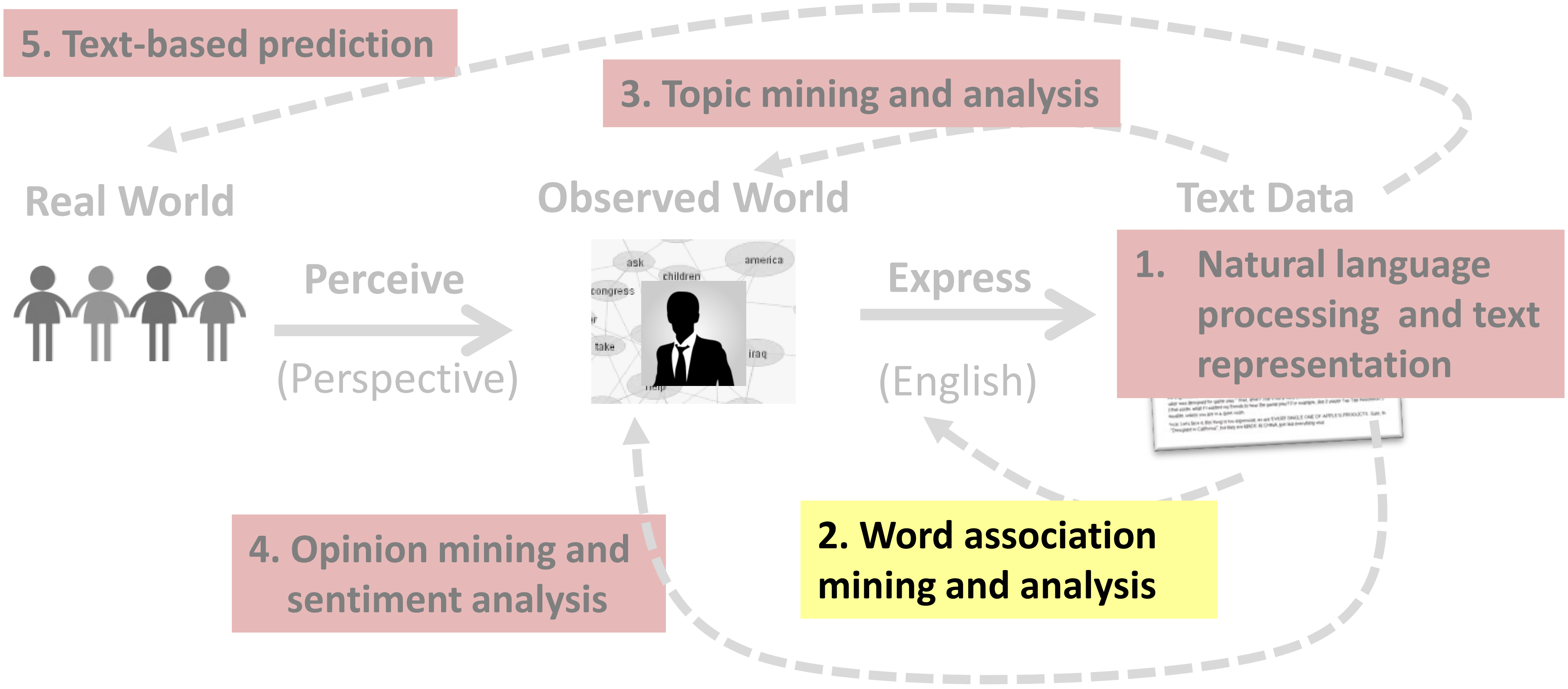
# Syntagmatic Relation Discovery: Entropy

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

# Syntagmatic Relation Discovery: 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

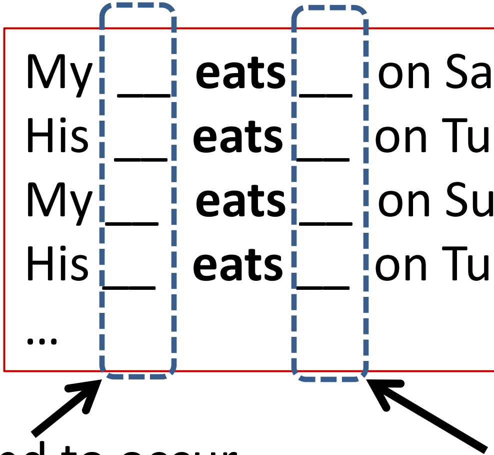
4. Opinion mining and sentiment analysis

2. Word association mining and analysis

# 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
…



My __ **eats** __ on Saturday
His __ **eats** __ on Tuesday
My __ **eats** __ on Sunday
His __ **eats** __ on Tuesday
…

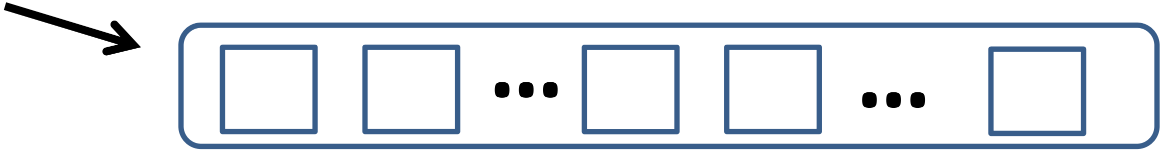
What words tend to occur to the **left** of "**eats**"?

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?**

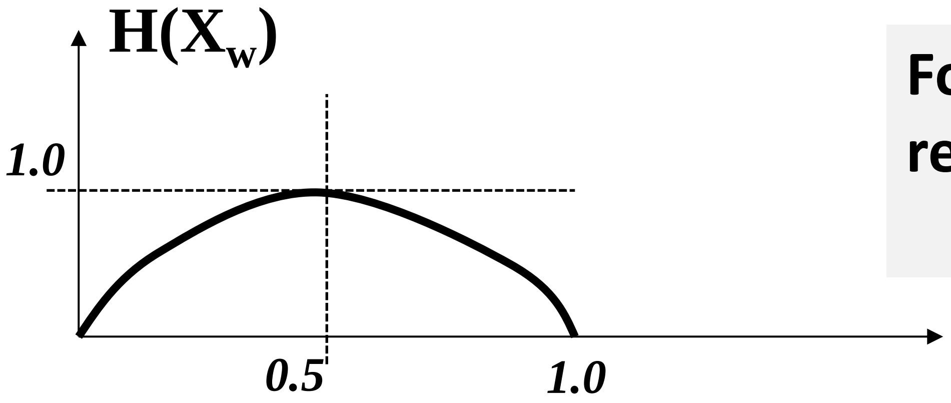
# Entropy H(X) Measures Randomness of X

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

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

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

Define $0\log_2 0 = 0$



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

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

**P(Xw=1)**

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

# Entropy H(X): Coin Tossing

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

**$X_{\text{coin}}$: tossing a coin**

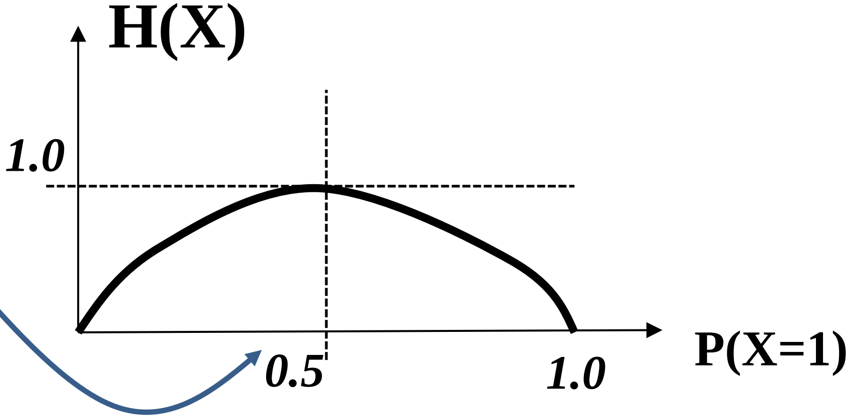
$$X_{\text{coin}} = \begin{cases} 1 & \text{Head} \\ 0 & \text{Tail} \end{cases}$$

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

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

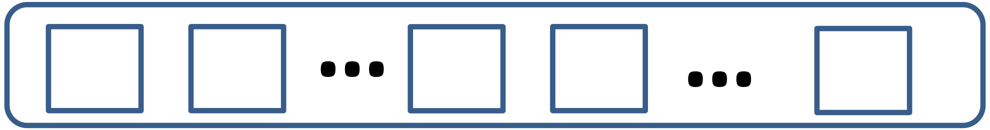
**Completely biased: p(X=1)=1**

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



# 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**? **$H(X_{\text{meat}})$, $H(X_{\text{the}})$, or $H(X_{\text{unicorn}})$?**

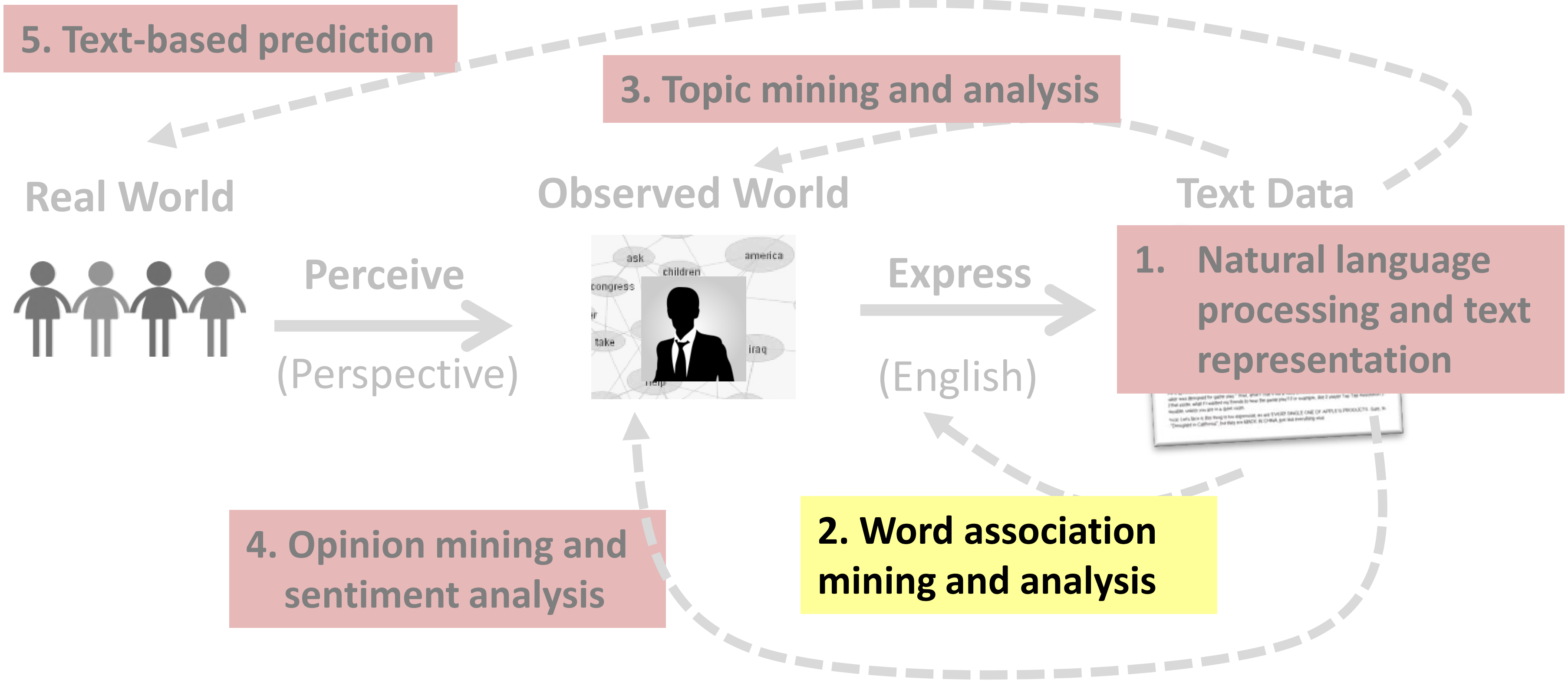
**$H(X_{\text{the}}) \approx 0$ ➔ no uncertainty since $p(X_{\text{the}}=1) \approx 1$**

**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)

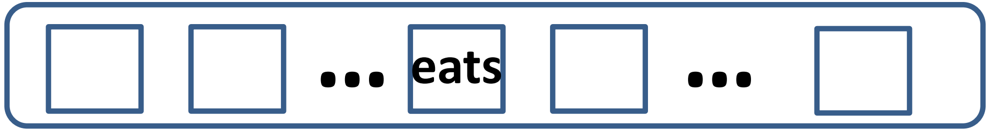
1. Natural language processing and text representation

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., $\mathbf{H(X_{meat})}$**?**

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)**

$$p(\boldsymbol{X}_{meat}=1) \dashrightarrow p(\boldsymbol{X}_{meat}=1 \mid \boldsymbol{X}_{eats}=1)$$

$$p(\boldsymbol{X}_{meat}=0) \dashrightarrow p(\boldsymbol{X}_{meat}=0 \mid \boldsymbol{X}_{eats}=1)$$

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

$$\Downarrow$$

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

$$\boldsymbol{-}\,p(\boldsymbol{X}_{meat}=1 \mid \boldsymbol{X}_{eats}=\boldsymbol{1})\,\boldsymbol{log_2}\,p(\boldsymbol{X}_{meat}=1 \mid \boldsymbol{X}_{eats}=\boldsymbol{1})$$

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

# Conditional Entropy: Complete Definition

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

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

**In general, for any discrete random variables X and Y, we have H(X) $\geq$ H(X|Y)**

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

# Conditional Entropy to Capture Syntagmatic Relation

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

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

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

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

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

# Conditional Entropy for Mining Syntagmatic Relations

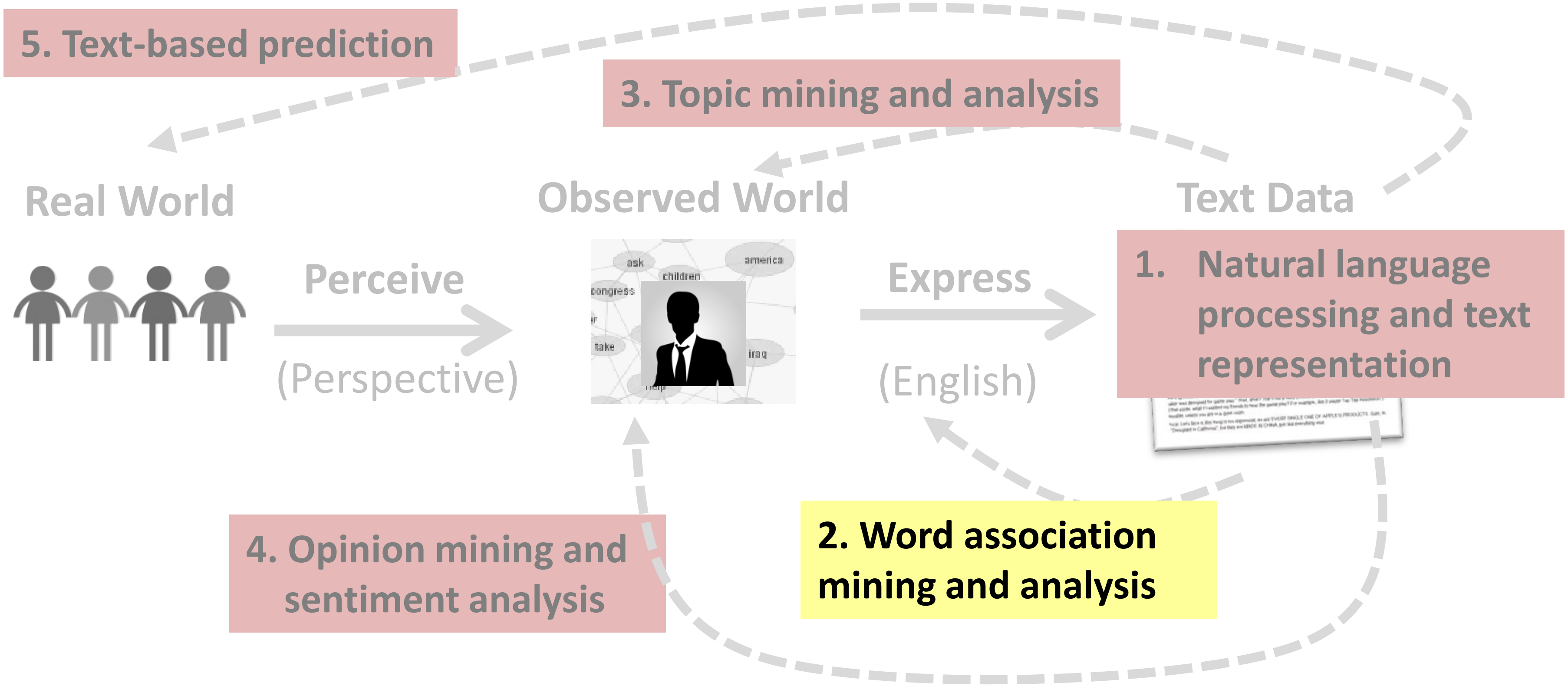
- For each word W1
  - For every other word W2, compute conditional entropy $H(X_{W1}|X_{W2})$
  - Sort all the candidate words in ascending order of $H(X_{W1}|X_{W2})$
  - Take the top-ranked candidate words as words that have potential syntagmatic relations with W1
  - Need to use a threshold for each 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: I(X; Y)= H(X) – H(X|Y) = H(Y)-H(Y|X)**

Properties:

- Non-negative: I(X;Y)≥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**"?

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

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

# 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:

$$\underbrace{\mathbf{p(X_{W1}=1, X_{W2}=1)}}_{\text{Both w1 \& w2 occur}} + \underbrace{\mathbf{p(X_{W1}=1, X_{W2}=0)}}_{\text{Only w1 occurs}} + \underbrace{\mathbf{p(X_{W1}=0, X_{W2}=1)}}_{\text{Only w2 occurs}} + \underbrace{\mathbf{p(X_{W1}=0, X_{W2}=0)}}_{\text{None of them occurs}} = \mathbf{1}$$

# 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: $\boxed{p(X_{W1}=1)} + p(X_{W1}=0) = 1$

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

Co-occurrences of w1 and w2:

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

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

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

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

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

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

**<u>Smoothing</u>: 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.

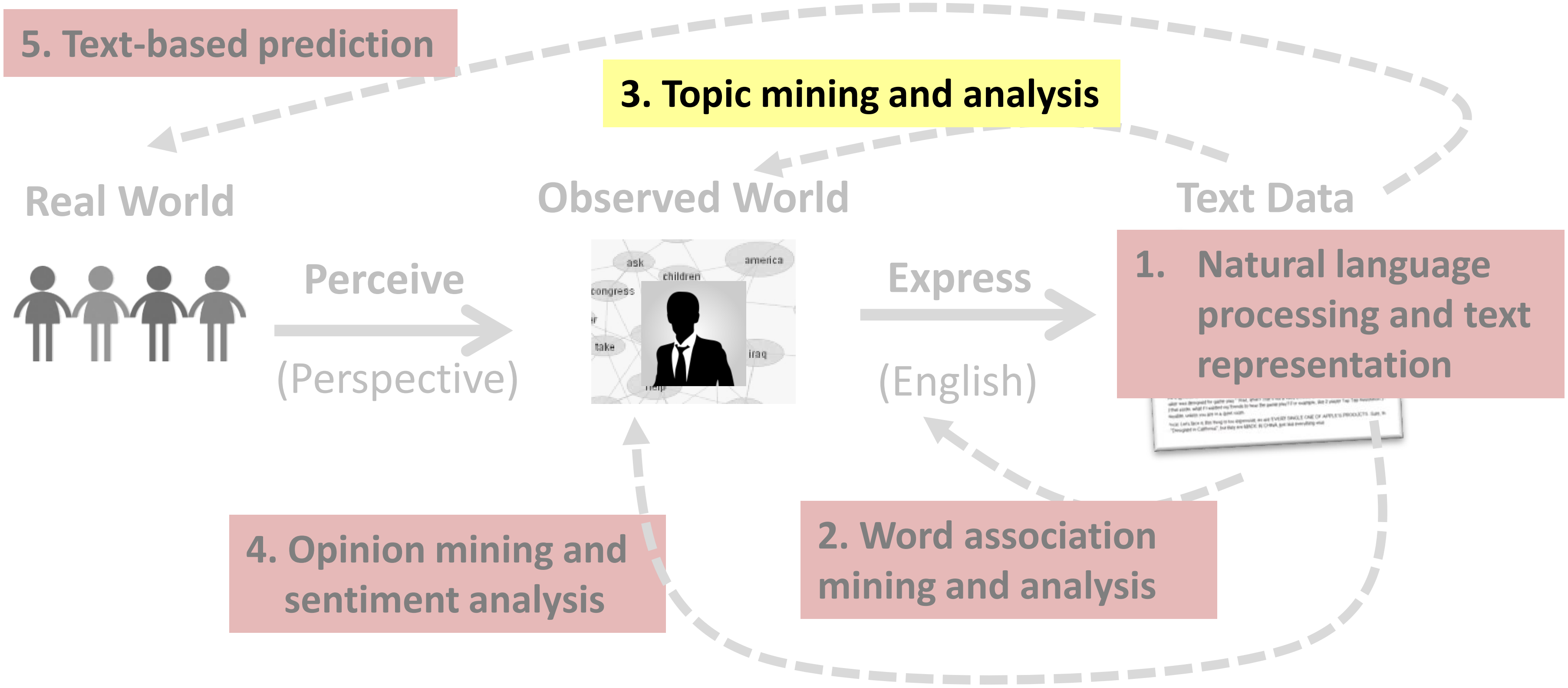
# Additional Reading

- 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



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

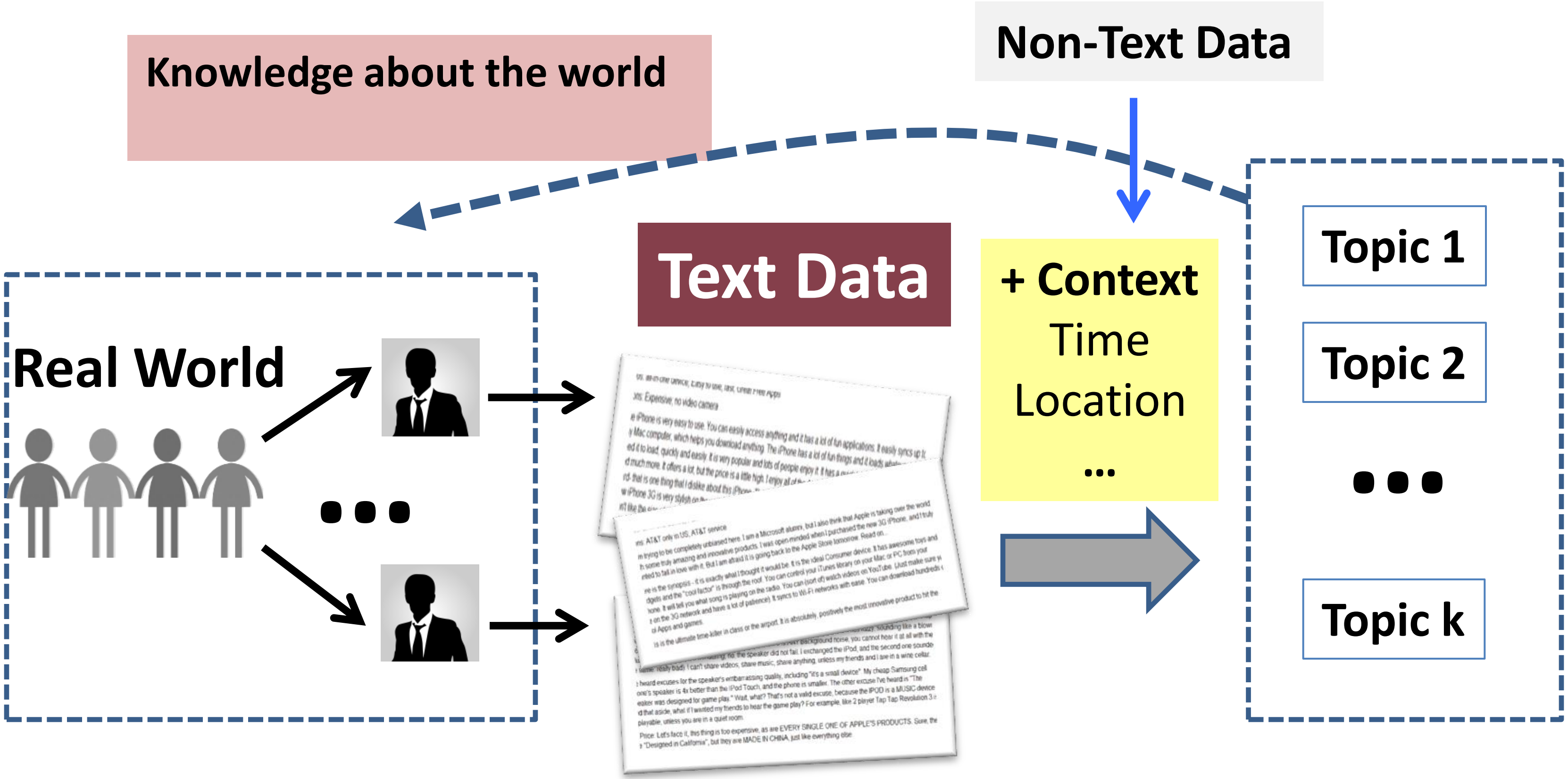
2. Word association mining and analysis

4. Opinion mining and sentiment analysis

# 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

# Tasks of Topic Mining and Analysis

**Task 2: Figure out which documents cover which topics**

**Text Data**

os: all-in-one device, Easy to use, fast, Great Free Apps

ns: Expensive, no video camera

e iPhone is very easy to use. You can easily access anything and it has a lot of fun applications. It easily syncs up t
y Mac computer, which helps you download anything. The iPhone has a lot of fun things and it loads wha
ed it to load, quickly and easily. It is very popular and lots of people enjoy it. It has a
d much more. It offers a lot, but the price is a little high. I enjoy all of th
rd- that is one thing that I dislike about this iPhone
w iPhone 3G is very stylish on th
n't like the si

ns: AT&T only in US, AT&T service

m trying to be completely unbiased here. I am a Microsoft alumni, but I also think that Apple is taking over the world
h some truly amazing and innovative products. I was open-minded when I purchased the new 3G iPhone, and I truly
nted to fall in love with it. But I am afraid it is going back to the Apple Store tomorrow. Read on...

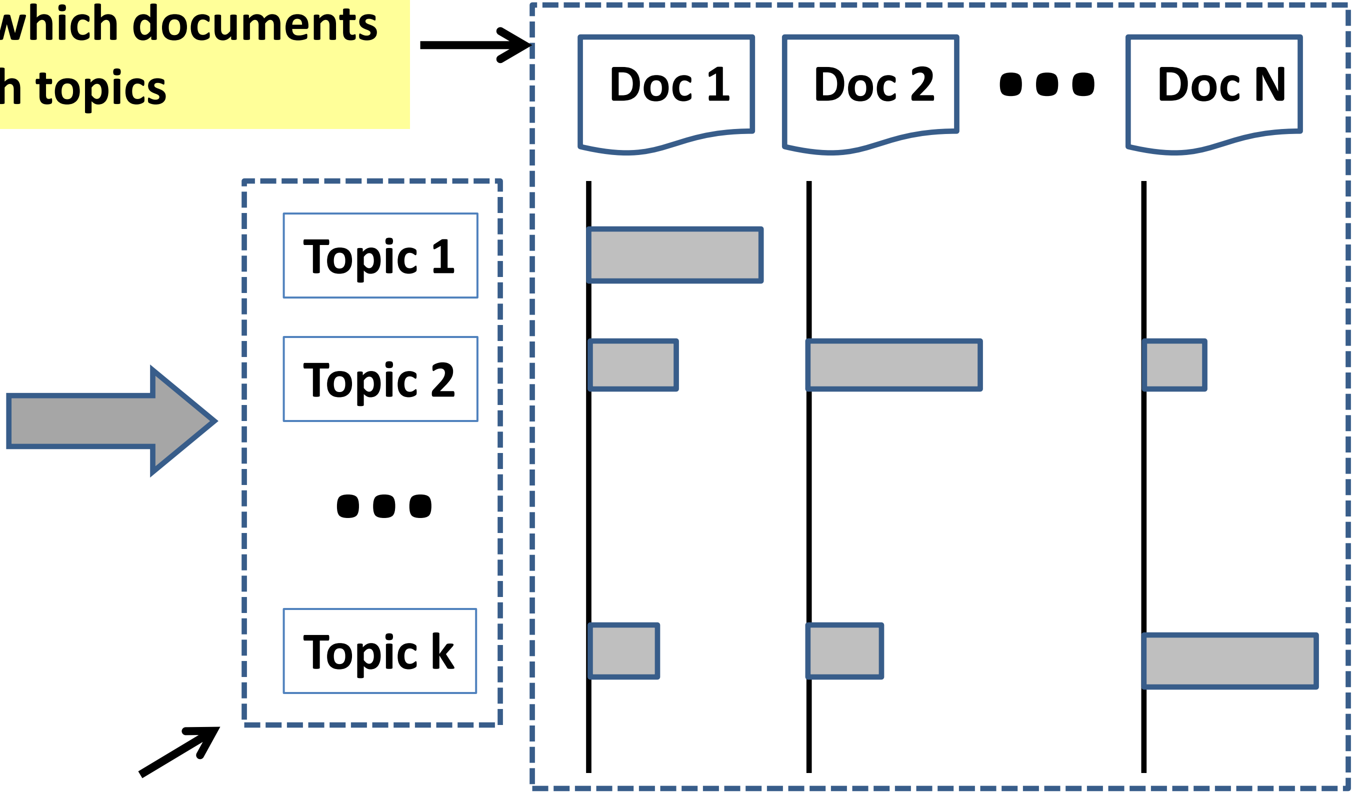
re is the synopsis - it is exactly what I thought it would be. It is the ideal Consumer device. It has awesome toys and
dgets and the "cool factor" is through the roof. You can control your iTunes library on your Mac or PC from your
hone. It will tell you what song is playing on the radio. You can (sort of) watch videos on YouTube. (Just make sure yo
t on the 3G network and have a lot of patience). It syncs to Wi-Fi networks with ease. You can download hundreds o
of Apps and games.

is is the ultimate time-killer in class or the airport. It is absolutely, positively the most innovative product to hit the

crazy. sounding like a blow
background noise, you cannot hear it at all with the
ing, no, the speaker did not fail. I exchanged the iPod, and the second one sounde
same. really bad). I can't share videos, share music, share anything, unless my friends and I are in a wine cellar.

e heard excuses for the speaker's embarrassing quality, including "it's a small device". My cheap Samsung cell
one's speaker is 4x better than the iPod Touch, and the phone is smaller. The other excuse I've heard is "The
eaker was designed for game play." Wait, what? That's not a valid excuse, because the IPOD is a MUSIC device
d that aside, what if I wanted my friends to hear the game play? For example, like 2 player Tap Tap Revolution 3 i
playable, unless you are in a quiet room.

Price: Let's face it, this thing is too expensive, as are EVERY SINGLE ONE OF APPLE'S PRODUCTS. Sure, the
e "Designed in California", but they are MADE IN CHINA, just like everything else.



**Task 1: Discover k topics**

# 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 $d_i$**: $\{ \pi_{i1}, \ldots, \pi_{ik} \}$
  - $\pi_{ij}$ = prob. of $d_i$ covering topic $\theta_j$

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

**How to define $\theta_i$ ?**

# 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

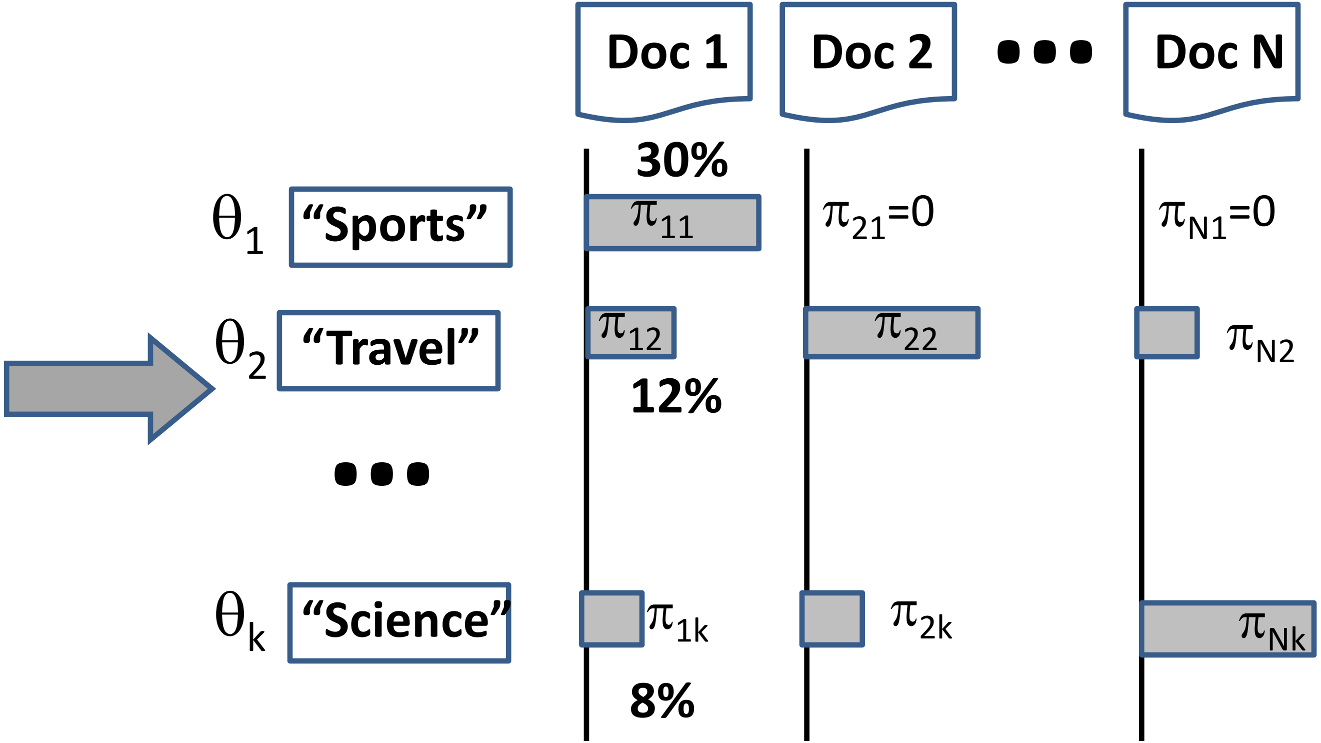
- Input
  - A **collection** of **N** text documents **$C=\{d_1, \ldots, d_N\}$**
  - **Number of topics**: **k**
- Output
  - **k topics**: **$\{ \theta_1, \ldots, \theta_k \}$**
  - **Coverage of topics in each $d_i$**: **$\{ \pi_{i1}, \ldots, \pi_{ik} \}$**
  - $\pi_{ij}$=prob. of $d_i$ covering topic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

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

**How to define $\theta_i$ ?**

# Initial Idea: Topic = Term

**Text Data**



$\theta_1$ **"Sports"**

$\theta_2$ **"Travel"**

**•••**

$\theta_k$ **"Science"**

**Doc 1** **Doc 2** **•••** **Doc N**

**30%** $\pi_{11}$

$\pi_{12}$ **12%**

$\pi_{1k}$ **8%**

$\pi_{21}=0$

$\pi_{22}$

$\pi_{2k}$

$\pi_{N1}=0$

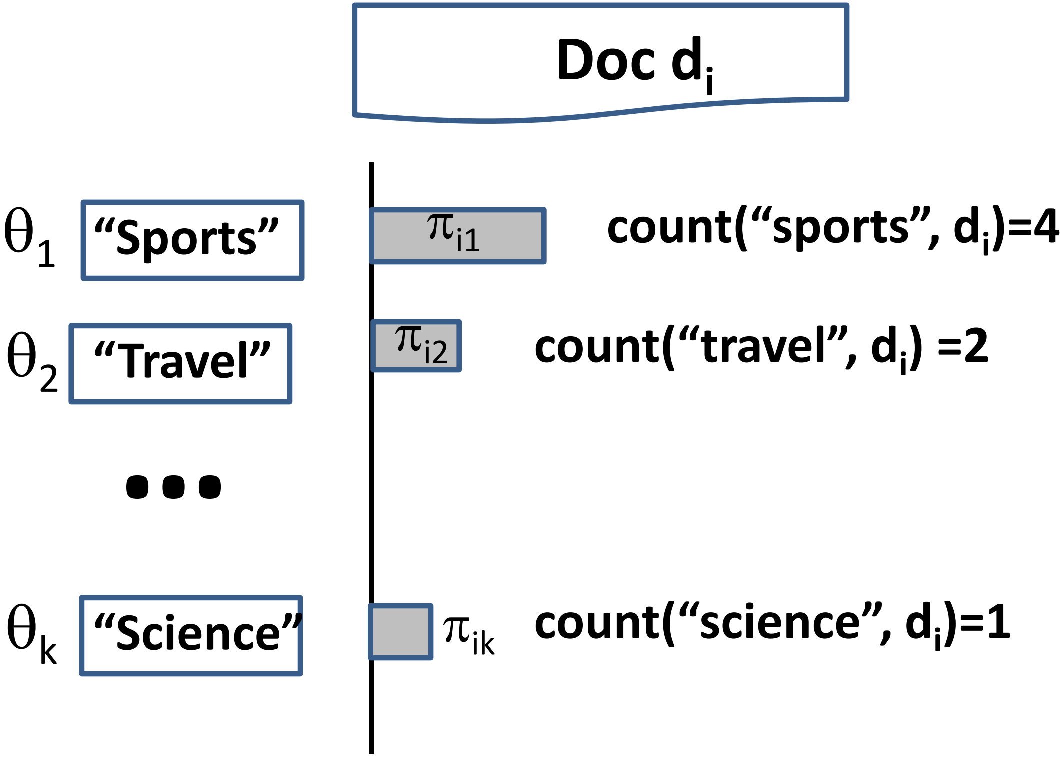
$\pi_{N2}$

$\pi_{Nk}$

# 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: $\pi_{ij}$



count("sports", $d_i$)=4

count("travel", $d_i$) =2

count("science", $d_i$)=1

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

# How Well Does This Approach Work?

**Doc $d_i$**

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

$\theta_1$ **"Sports"** $\pi_{i1} \propto c(\text{"sports"}, d_i) = 0$

$\theta_2$ **"Travel"** $\pi_{i2} \propto c(\text{"travel"}, d_i) = 1 > 0$

…

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

**1. Need to count related words also!**

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

**3. Mine complicated topics?**

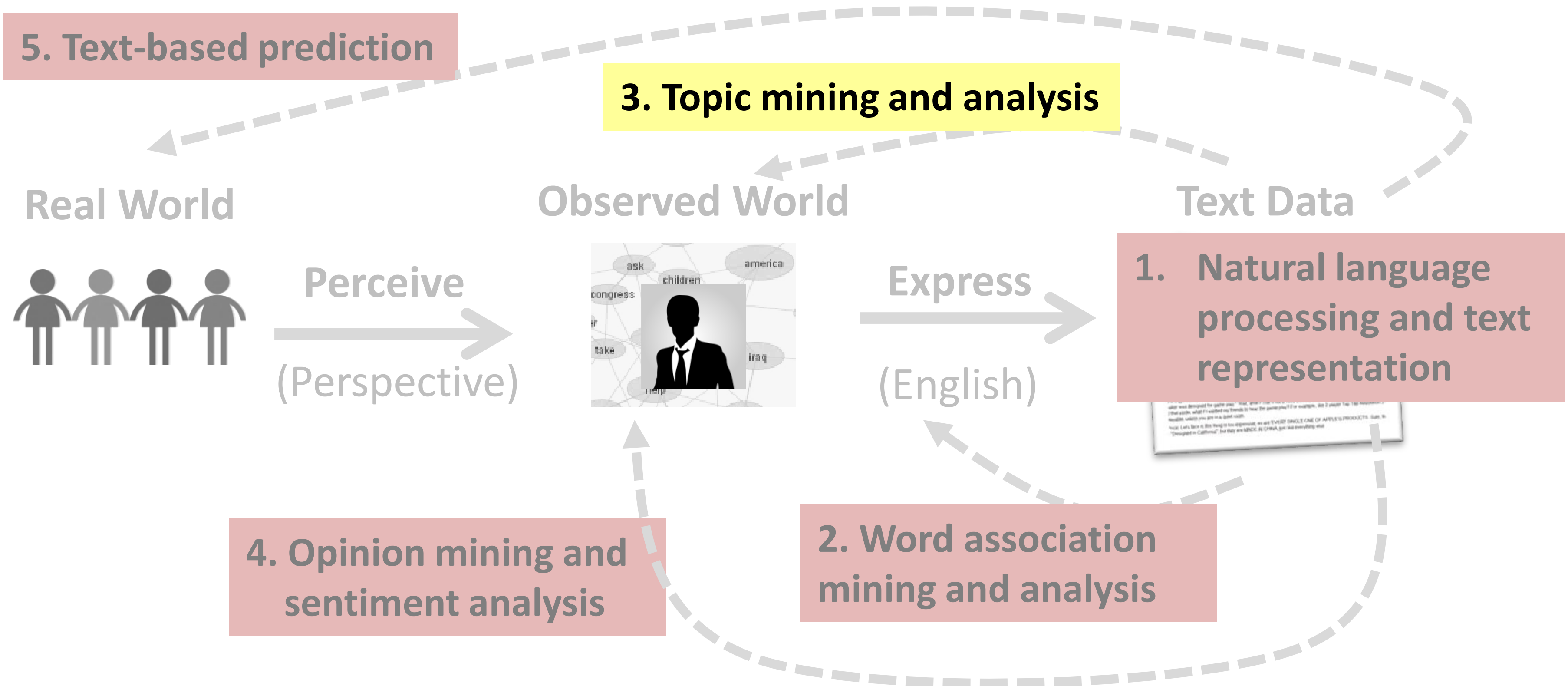
# Problems with "Term as Topic"

- 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



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

# Problems with “Term as Topic”

- Lack of expressive power
  - Can only represent simple/general topics
  - Can’t represent complicated topics

**➔ Topic = {Multiple Words}**

- Incompleteness in vocabulary coverage
  - Can’t capture variations of vocabulary (e.g., related words)

**+ weights on words**

- Word sense ambiguity
  - A topical term or related term can be ambiguous (e.g., basketball star vs. star in the sky)

**➔ Split an ambiguous word**

**A probabilistic topic model can do all these!**

# Improved Idea: Topic = Word Distribution

$\theta_1$ "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 |
| … | |

$\theta_2$ "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 |
| … | |

…

$\theta_k$ "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 **$C=\{d_1, \ldots, d_N\}$**
  - **Vocabulary set: $V=\{w_1, \ldots, w_M\}$**
  - **Number of topics**: **k**
- Output
  - **k topics, each a word distribution**: **$\{ \theta_1, \ldots, \theta_k \}$**
  - **Coverage of topics in each $d_i$**: **$\{ \pi_{i1}, \ldots, \pi_{ik} \}$**
  - $\pi_{ij}$=prob. of $d_i$ covering topic $\theta_j$

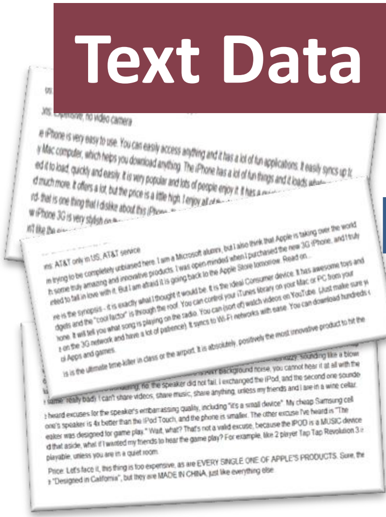
$$\sum_{w \in V} p(w \mid \theta_i) = 1$$

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

# The Computation Task

**INPUT: C, k, V**

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



Text Data

ons: expensive, no video camera

e iPhone is very easy to use. You can easily access anything and it has a lot of fun applications. It easily syncs up t
y Mac computer, which helps you download anything. The iPhone has a lot of fun things and it loads wha
ed it to load, quickly and easily. It is very popular and lots of people enjoy it. It has a
d much more. It offers a lot, but the price is a little high. I enjoy all of th
rd. that is one thing that I dislike about this iPhone
w iPhone 3G is very stylish on th
n't like the

ons: AT&T only in US, AT&T service

m trying to be completely unbiased here. I am a Microsoft alumni, but I also think that Apple is taking over the world
th some truly amazing and innovative products. I was open-minded when I purchased the new 3G iPhone, and I truly
anted to fall in love with it. But I am afraid it is going back to the Apple Store tomorrow. Read on...

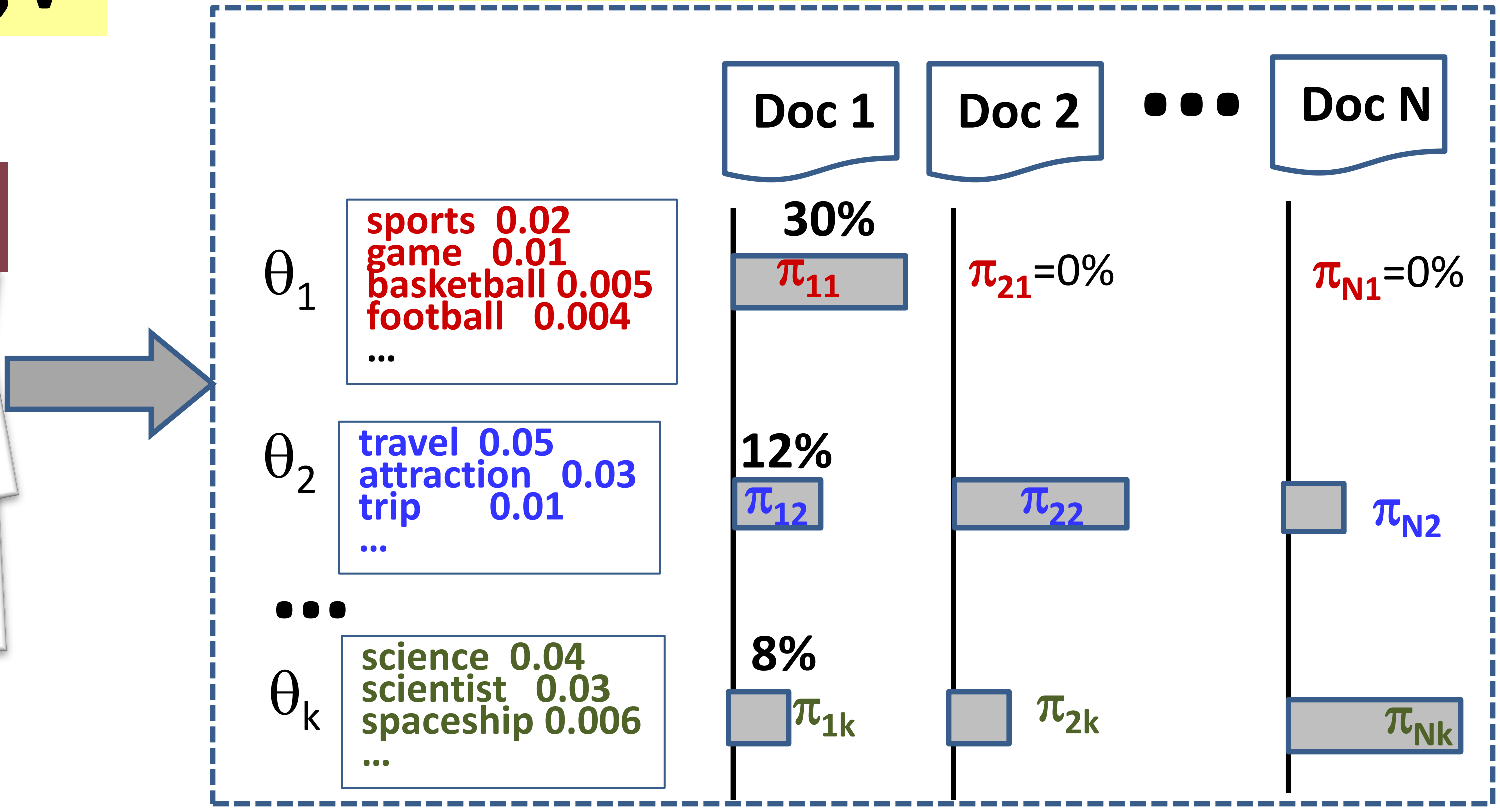
ere is the synopsis - it is exactly what I thought it would be. It is the ideal Consumer device. It has awesome toys and
dgets and the "cool factor" is through the roof. You can control your iTunes library on your Mac or PC from your
hone. It will tell you what song is playing on the radio. You can (sort of) watch videos on YouTube. (Just make sure yo
t on the 3G network and have a lot of patience). It syncs to Wi-Fi networks with ease. You can download hundreds o
ol Apps and games.

is is the ultimate time-killer in class or the airport. It is absolutely, positively the most innovative product to hit the

fuzzy, sounding like a blow
ANY background noise, you cannot hear it at all with the
wondering, no, the speaker did not fail. I exchanged the iPod, and the second one sounde
e same: really bad). I can't share videos, share music, share anything, unless my friends and I are in a wine cellar.

e heard excuses for the speaker's embarrassing quality, including "it's a small device". My cheap Samsung cell
one's speaker is 4x better than the iPod Touch, and the phone is smaller. The other excuse I've heard is "The
eaker was designed for game play." Wait, what? That's not a valid excuse, because the IPOD is a MUSIC device
d that aside, what if I wanted my friends to hear the game play? For example, like 2 player Tap Tap Revolution 3 i
playable, unless you are in a quiet room.

Price: Let's face it, this thing is too expensive, as are EVERY SINGLE ONE OF APPLE'S PRODUCTS. Sure, the
e "Designed in California", but they are MADE IN CHINA, just like everything else.

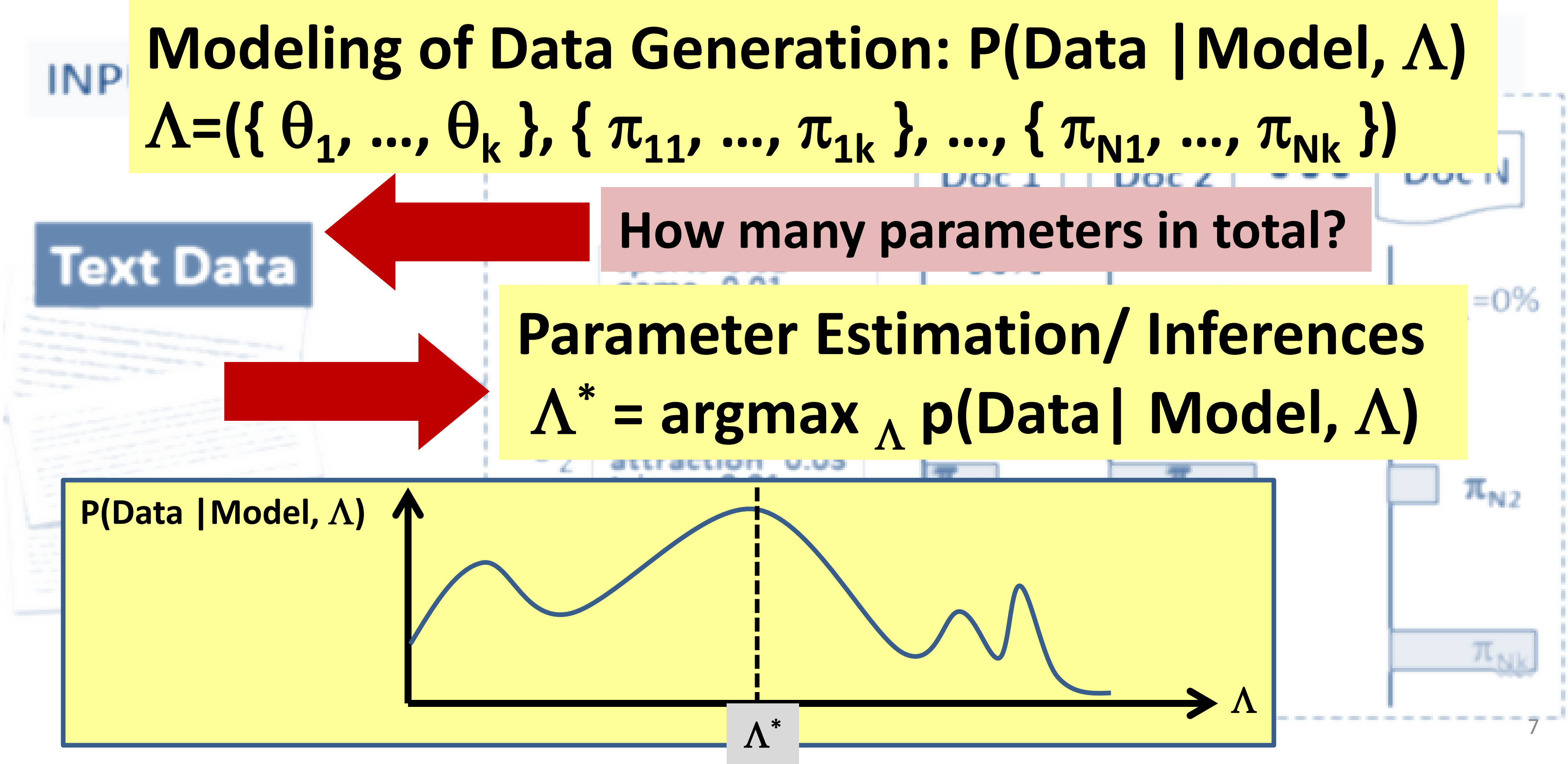


sports 0.02
game 0.01
basketball 0.005
football 0.004
…

travel 0.05
attraction 0.03
trip 0.01
…

science 0.04
scientist 0.03
spaceship 0.006
…

# Generative Model for Text Mining



**Modeling of Data Generation: P(Data |Model, $\Lambda$)**

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

**How many parameters in total?**

**Parameter Estimation/ Inferences**

**$\Lambda^*$ = argmax $_\Lambda$ p(Data| 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

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

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

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

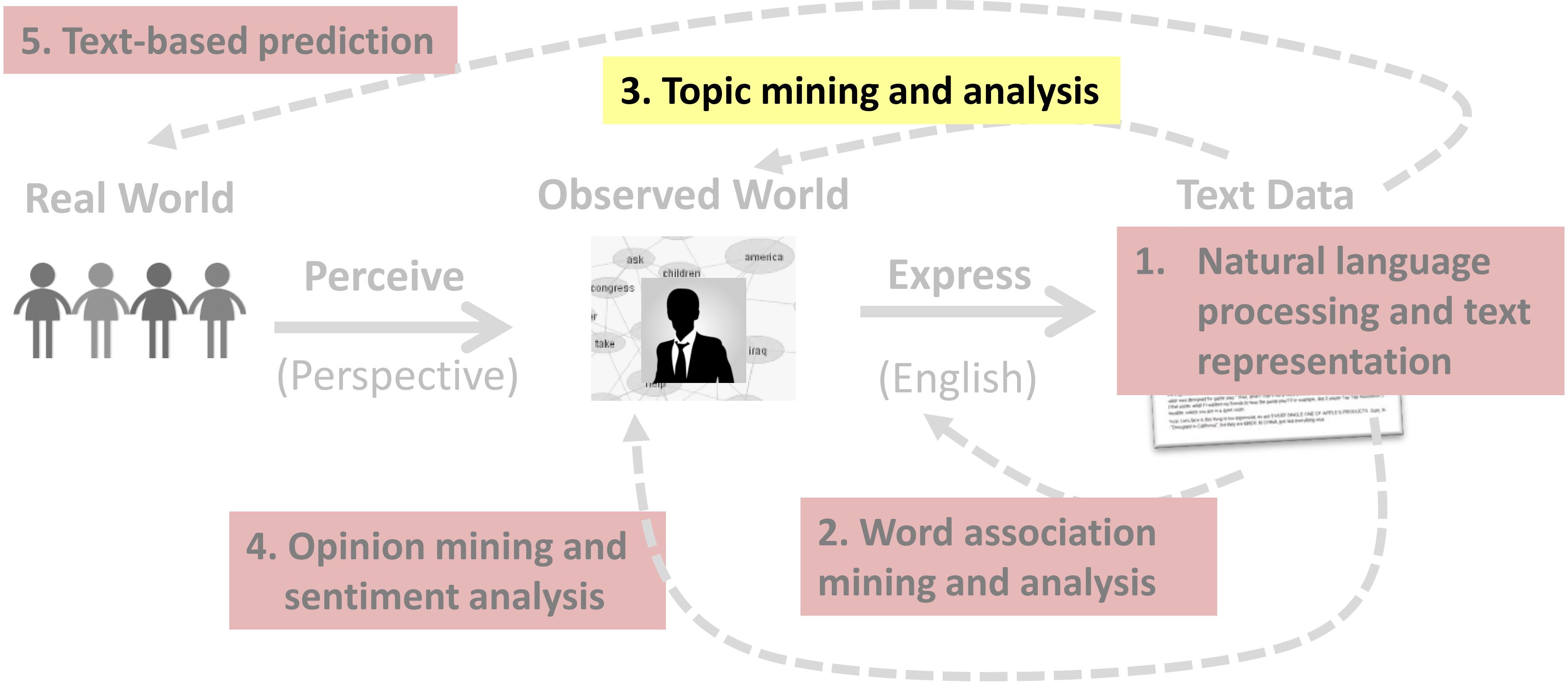
# Summary (cont.)

- **Generative model** for text mining
  - **Model data generation** with a prob. model: **P(Data |Model, $\Lambda$)**
  - **Infer the most likely parameter values $\Lambda^*$** given a particular data set**:** $\Lambda^*$ **= argmax $_\Lambda$ p(Data| Model, $\Lambda$)**
  - **Take $\Lambda^*$ 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



5. Text-based prediction

3. Topic mining and analysis

Real World

Observed World

Text Data

Perceive

(Perspective)

Express

(English)

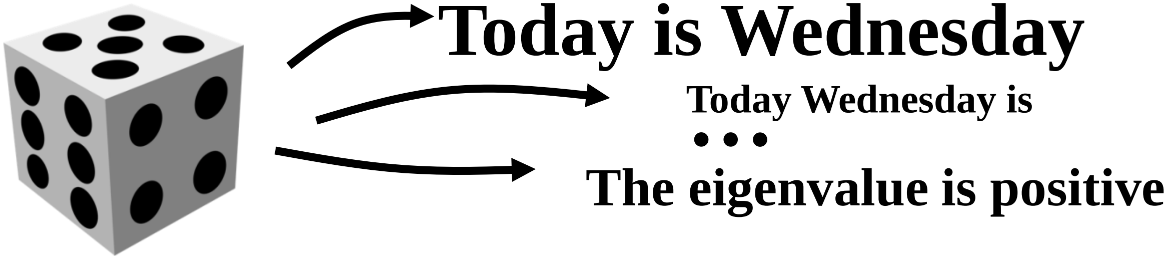
1. Natural language processing and text representation

4. Opinion mining and sentiment analysis

2. Word association mining and analysis

# What 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



**Today is Wednesday**

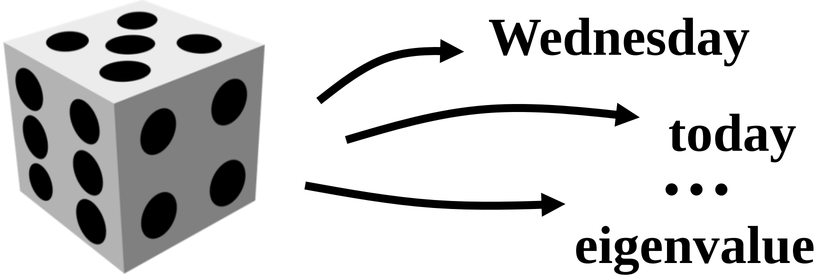
**Today Wednesday is**

**• • •**

**The eigenvalue is positive**

# 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**



Wednesday

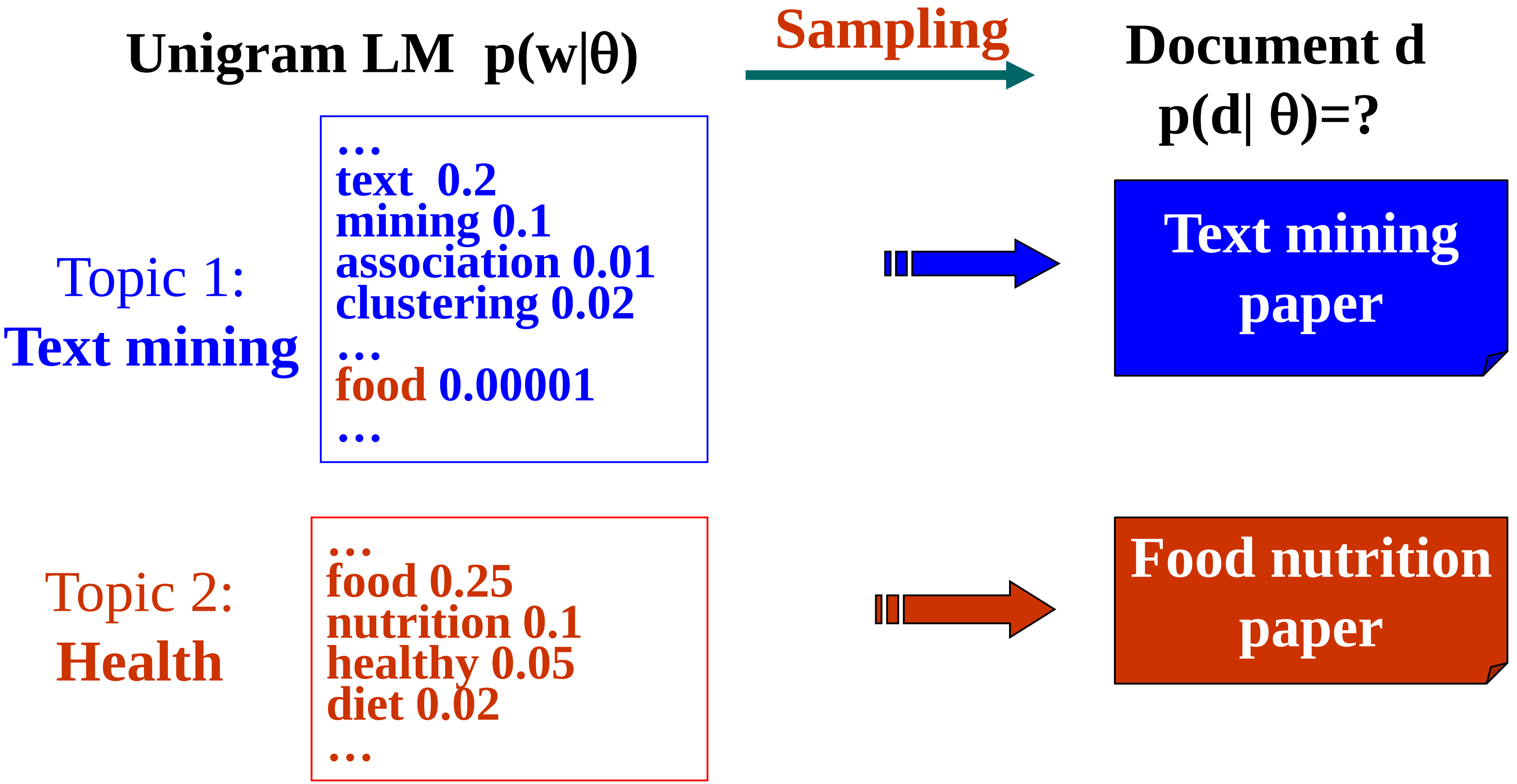
today

…

eigenvalue

$$p(\text{“today is Wed”}) = p(\text{“today”})p(\text{“is”})p(\text{“Wed”}) = 0.0002 \times 0.001 \times 0.000015$$

# 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
…

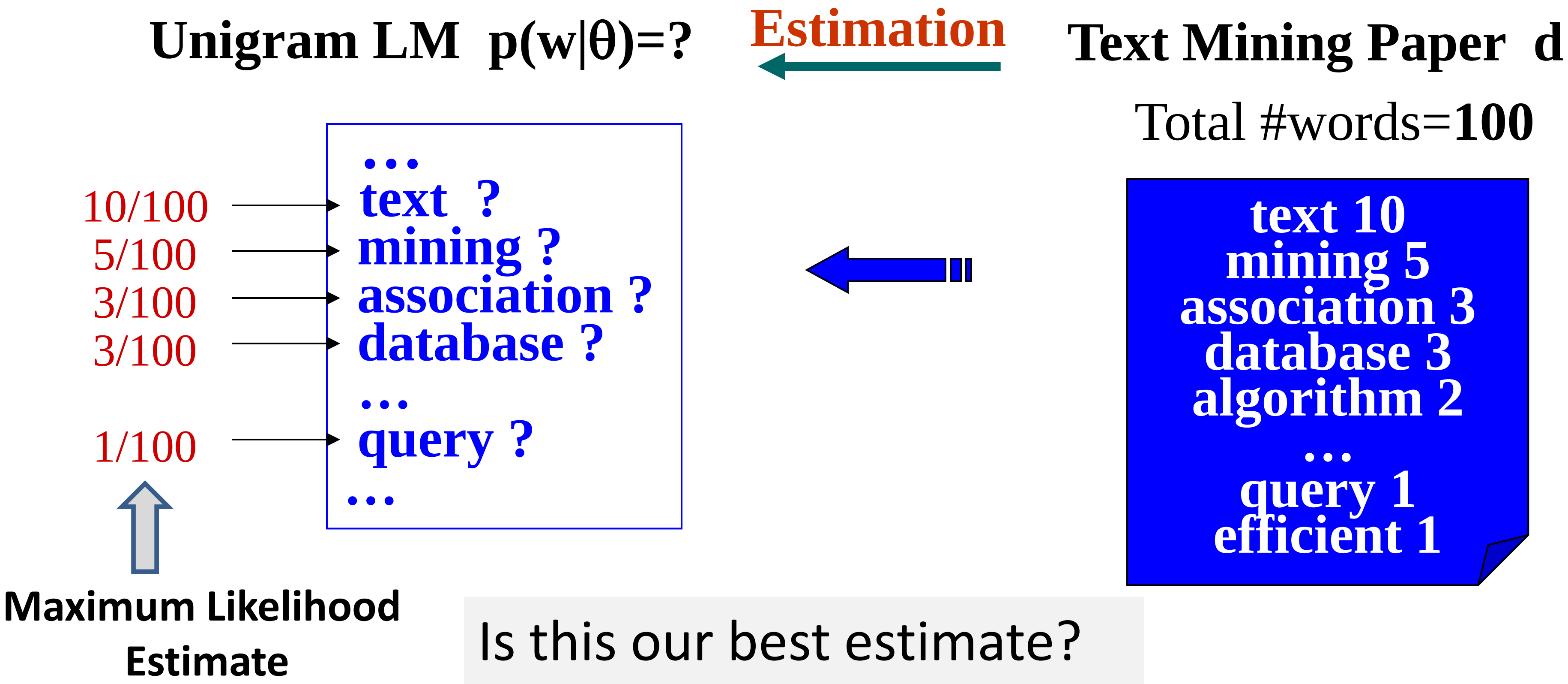
Text mining paper

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|θ)=?

Estimation

Text Mining Paper d

Total #words=**100**

…
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

**Maximum Likelihood Estimate**

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

**Bayes Rule**

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

- Bayesian estimation:
  - "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**

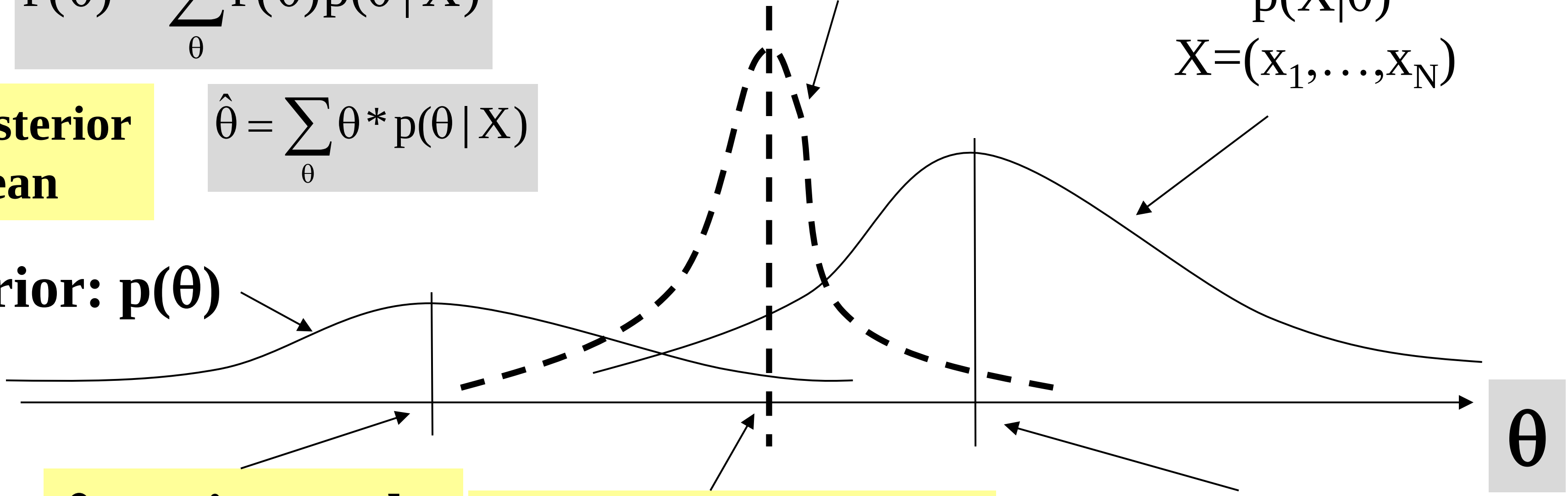
# Illustration of Bayesian Estimation

**Bayesian inference: f(θ)=?**

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

**Posterior Mean**

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



**Posterior:**
**$p(\theta|X) \propto p(X|\theta)p(\theta)$**

**Likelihood:**
$p(X|\theta)$
$X=(x_1,\ldots,x_N)$

**Prior: $p(\theta)$**

**$\theta$**

**$\theta_0$: prior mode**

**$\theta_1$: posterior mode**

**$\theta_{ml}$: ML estimate**

# Summary

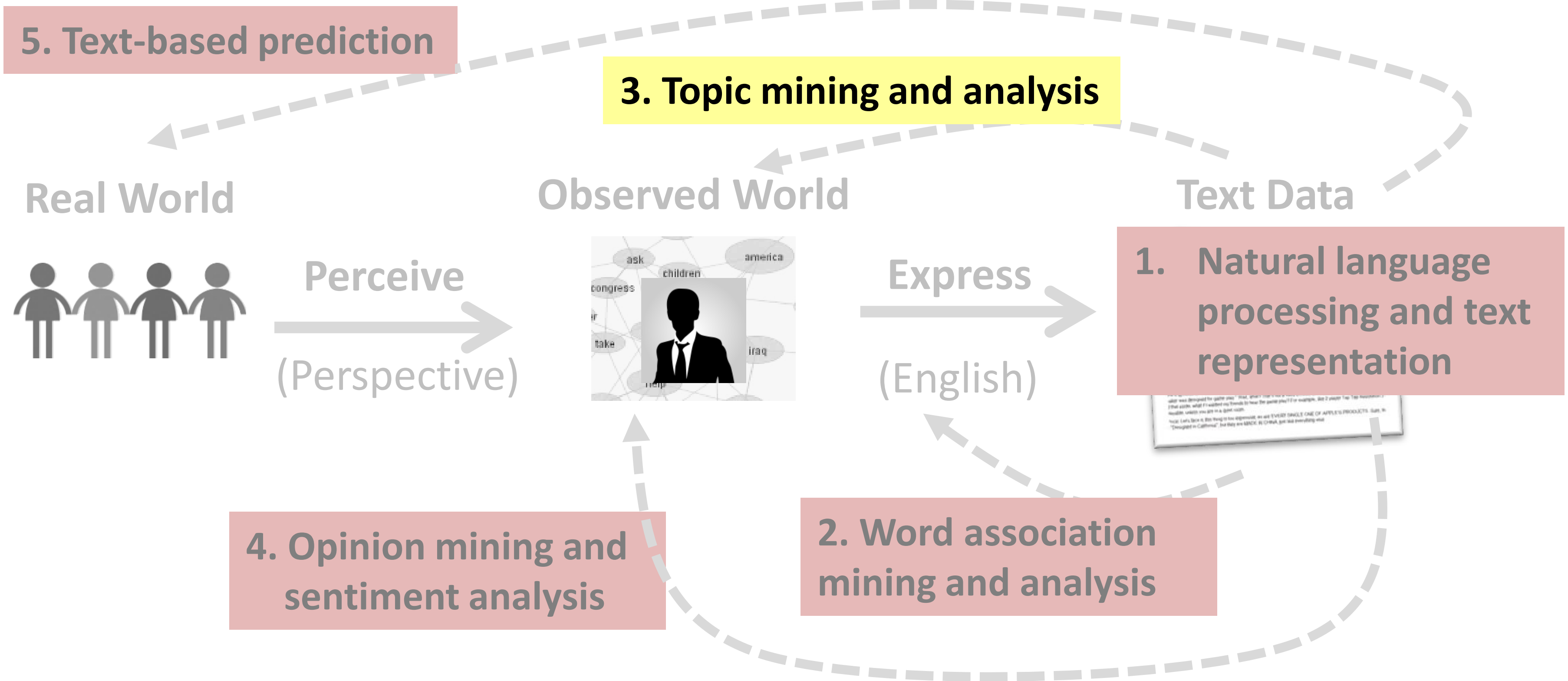
- **Language Model** = probability distribution over text = generative model for text data
- **Unigram** Language Model = **word distribution**
- **Likelihood** function: **p(X|θ)**
  - **Given θ** ➔ which X has a higher likelihood?
  - **Given X** ➔ which θ maximizes p(X| θ)? **[ML estimate]**
- **Bayesian**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

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

# Topic Mining and Analysis: Mining One Topic

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

# 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

2. Word association mining and analysis

4. Opinion mining and sentiment analysis

# Simplest Case of Topic Model: Mining One Topic

**INPUT: C={d}, V**

**Text Data**

ay, we all know the good things the iPod Touch has to offer, and those reasons are why I am BARELY holding on
e thing. I am still considering taking it back. Here's why.

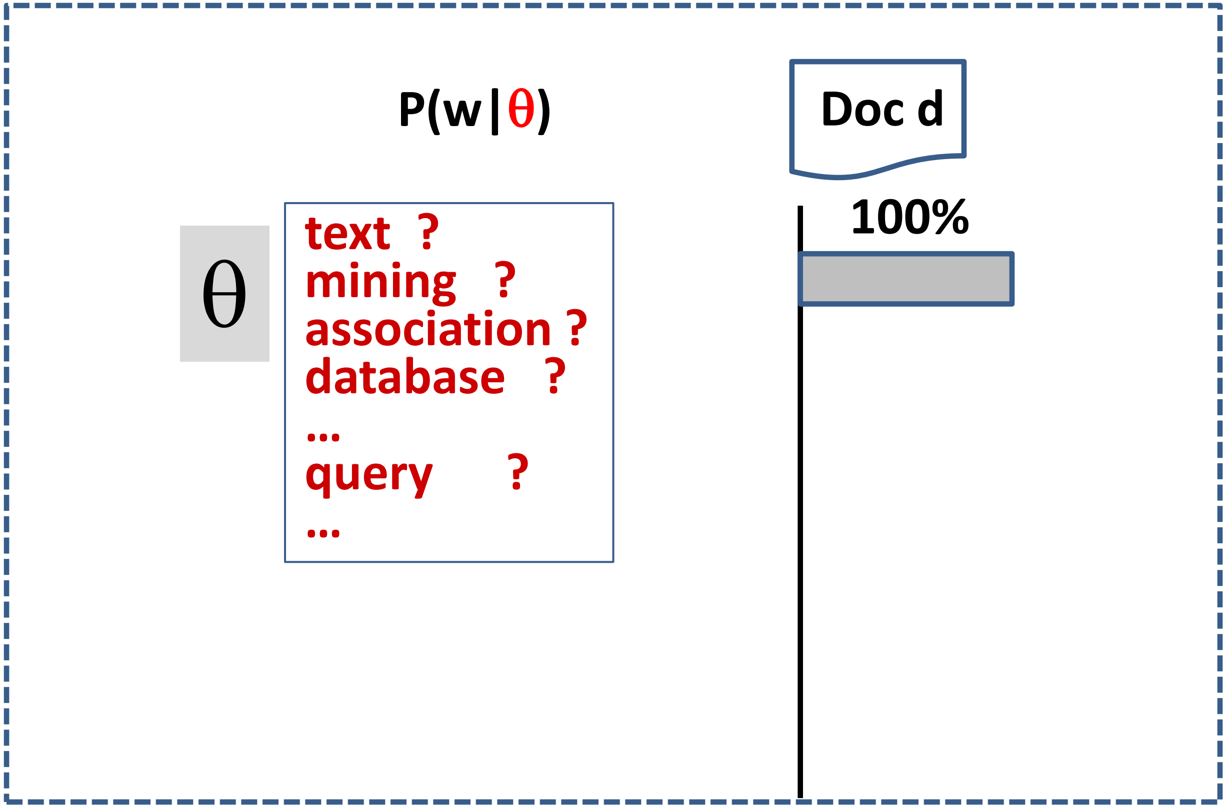
Speaker quality is ABSOLUTELY HORRENDOUS. The speaker is simply not functional, unless you are in perhap
ecording studio. When you turn it up all the way (because you can't hear it) it becomes fuzzy, sounding like a blow
eaker. What is this thing, the size of a flea? And if there is ANY background noise, you cannot hear it at all with the
lume up (in case you're wondering, no, the speaker did not fail. I exchanged the iPod, and the second one sounde
e same: really bad). I can't share videos, share music, share anything, unless my friends and I are in a wine cellar.

e heard excuses for the speaker's embarrassing quality, including "it's a small device". My cheap Samsung cell
one's speaker is 4x better than the iPod Touch, and the phone is smaller. The other excuse I've heard is "The
eaker was designed for game play." Wait, what? That's not a valid excuse, because the IPOD is a MUSIC device
d that aside, what if I wanted my friends to hear the game play? For example, like 2 player Tap Tap Revolution 3 i
playable, unless you are in a quiet room.

Price: Let's face it, this thing is too expensive, as are EVERY SINGLE ONE OF APPLE'S PRODUCTS. Sure, the
e "Designed in California", but they are MADE IN CHINA, just like everything else.

Microphone: WHY IS more »

**OUTPUT: { θ}**



P(w|θ)

θ

text ?
mining ?
association ?
database ?
…
query ?
…

# Language Model Setup

- **Data**: Document d= $x_1$ $x_2$ … $x_{|d|}$ , $x_i \in V=\{w_1 ,…, w_M\}$ is a word

- **Model**: Unigram LM $\theta$(=topic) : $\{\theta_i=p(w_i \mid \theta)\}$, i=1, …, M; $\theta_1+…+\theta_M=1$

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

# 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:** $$\sum_{i=1}^{M} \theta_i = 1$$

Use Lagrange multiplier approach

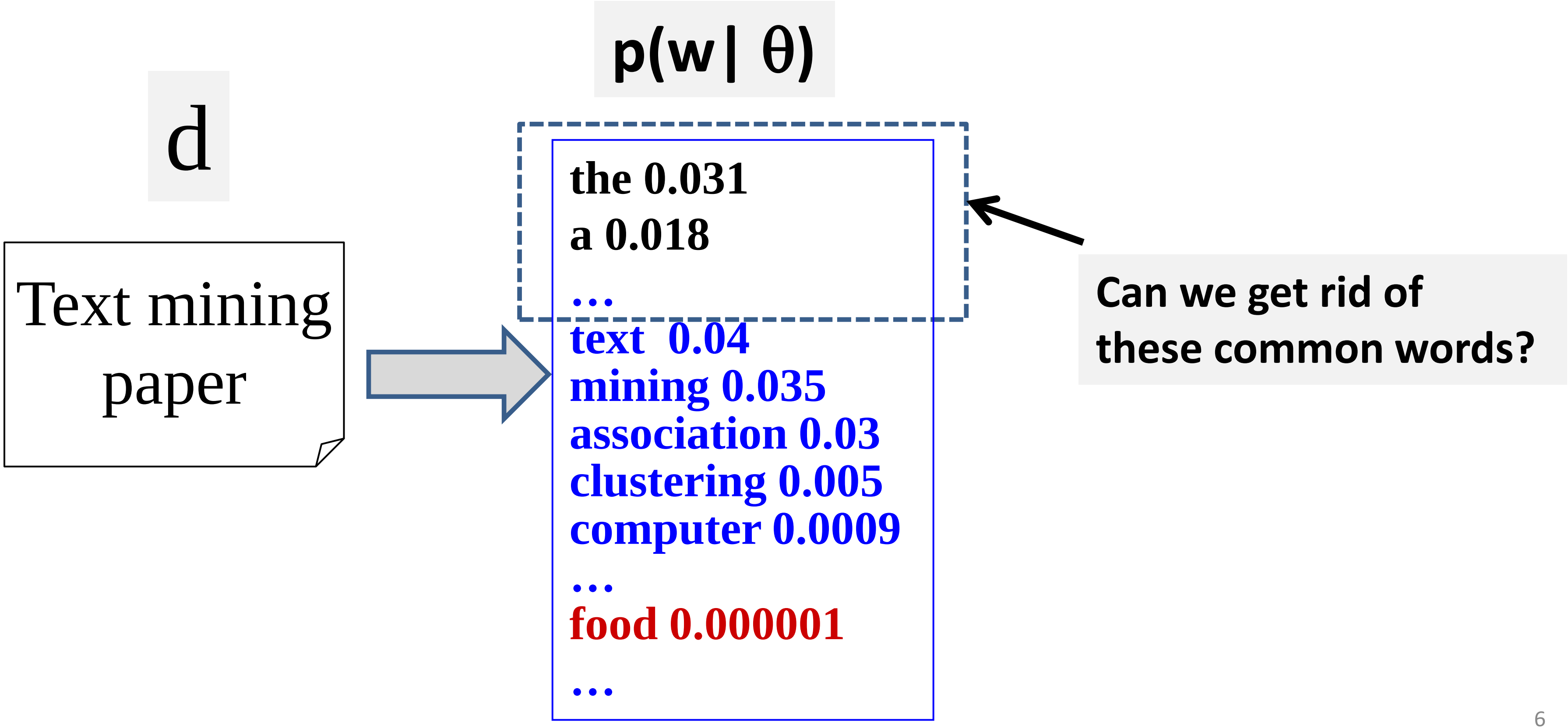
Lagrange function: $f(q\,|\,d) = \sum_{i=1}^{M} c(w_i,d)\log q_i + /\,(\sum_{i=1}^{M} q_i - 1)$

$$\frac{\partial f(q\,|\,d)}{\partial q_i} = \frac{c(w_i,d)}{q_i} + / = 0 \quad \rightarrow \quad q_i = -\frac{c(w_i,d)}{/}$$

**Normalized Counts**

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

# What Does the Topic Look Like?



d

Text mining paper

p(w| θ)

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?