

Latent Dirichlet Allocation

-Topic model

Contents

01 방법론 소개

02 Bayseian Network

03 Gibbs Sampler

04 Model Structure

05 Some Computations

06 Conclusion

01 방법론 소개



History

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

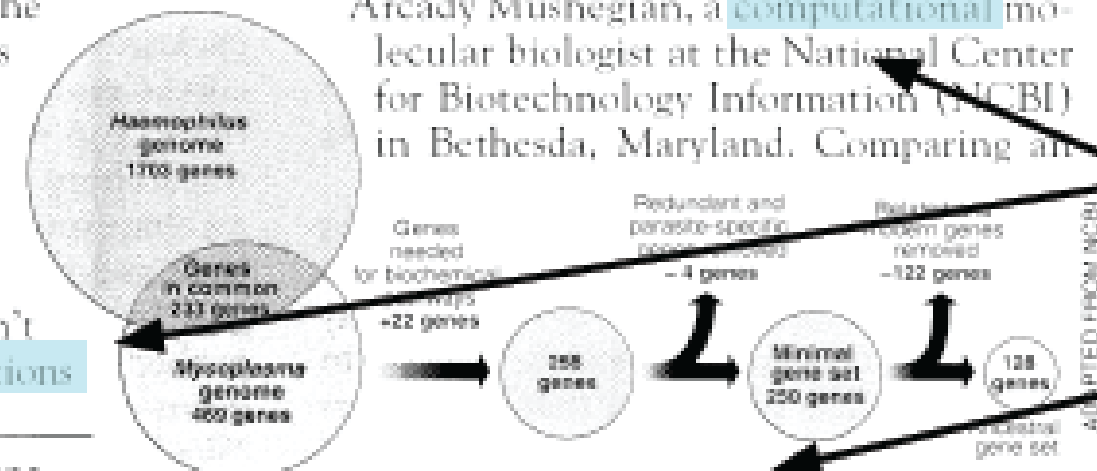
Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

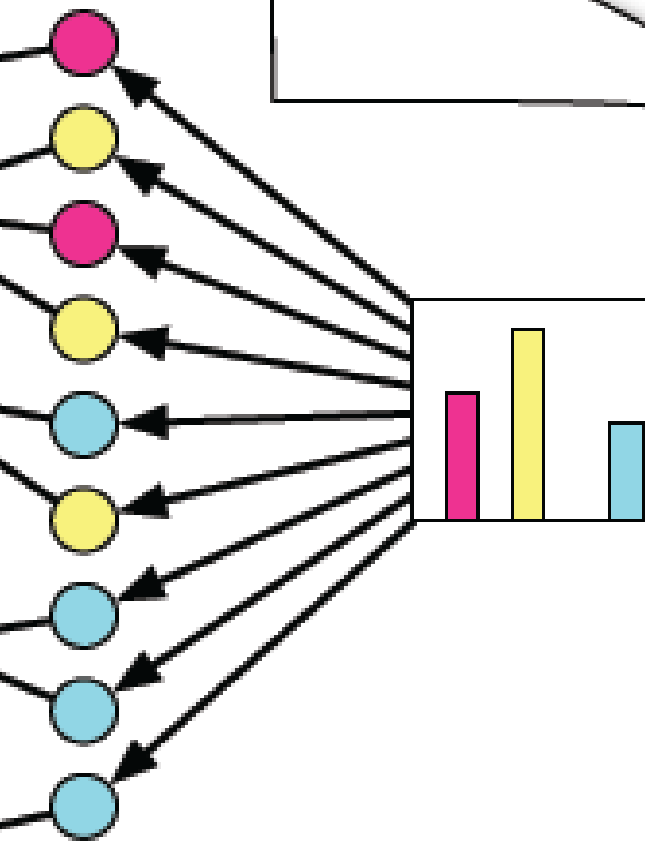


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. **Computer analysis** yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



- David Blei, Andrew Ng, and Michael I. Jordan in 2003
- Natural Language Process(NLP) 모델

Characteristics

- Document, Topic
 - Document : 다양한 Topic들의 혼합
 - Topic : Word들의 분포
- Generative
 - Observed로 모수를 학습하고 다시 Observed를 만든다.
 - 잘 알고 있는 GAN ~ image generative
- Statistical
 - word에 topic 배정시 Gibbs sampler 등

Objective

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

- How to assign a topic to a word ?

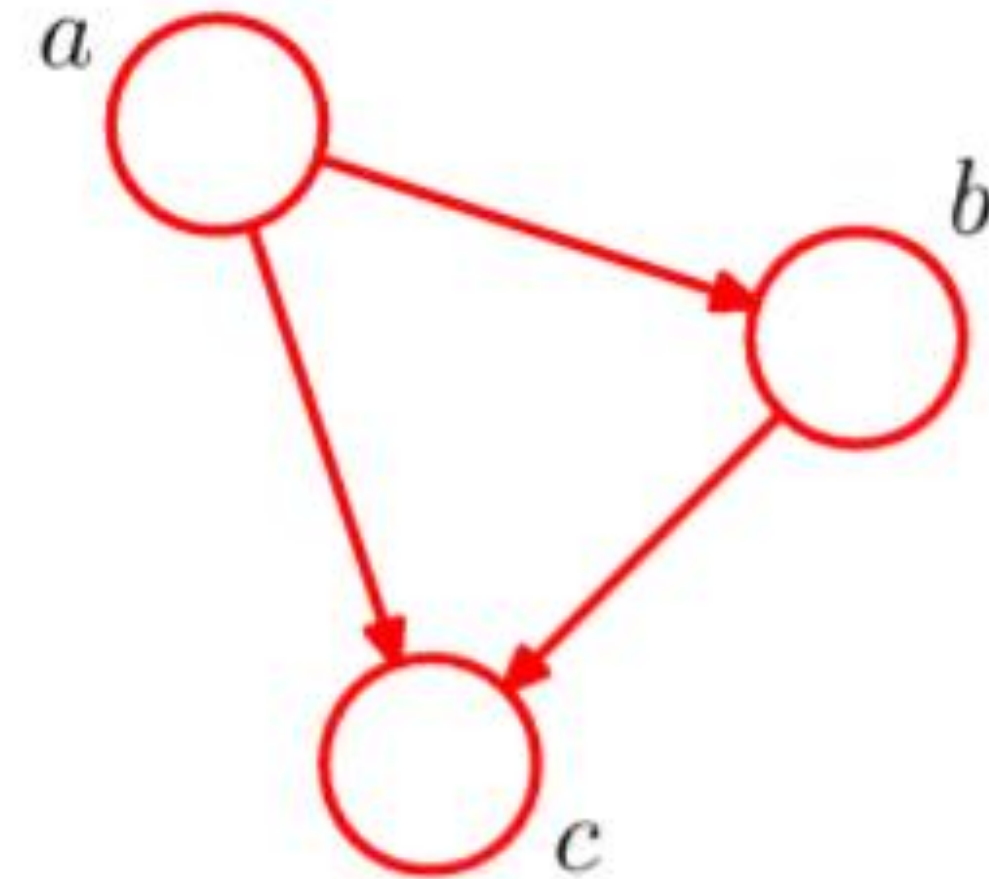
02

Bayesian Network



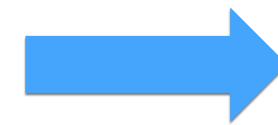
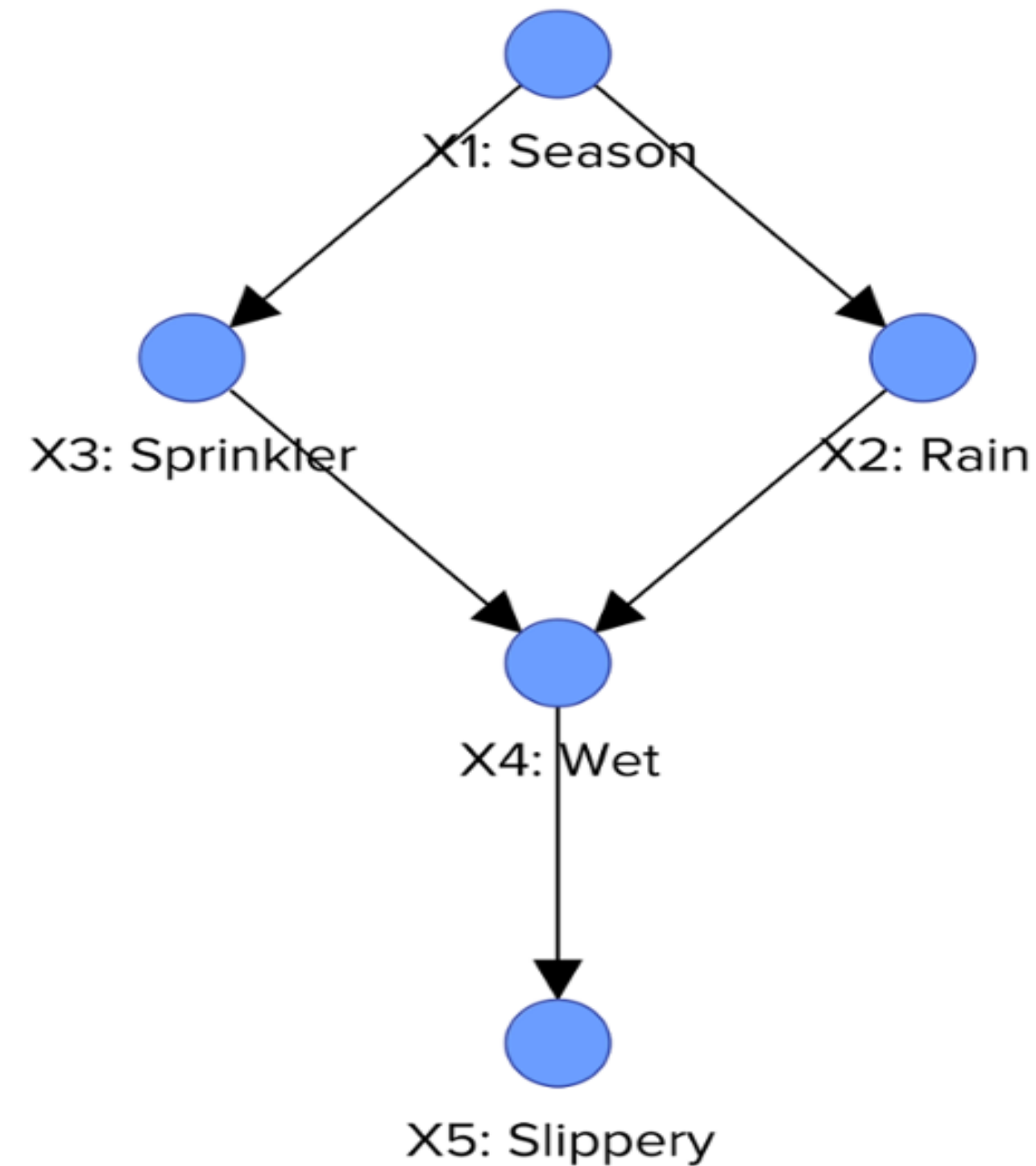
Graphical Representation

$$\begin{aligned} p(a, b, c) &= p(c|a, b)p(a, b) \\ &= p(c|a, b)p(b|a)p(a) \end{aligned}$$



- 확률분포를 그래픽적 모형으로 나타낼 수 있다.
- 시각적 직관적으로 모델을 파악할 수 있다.
- node는 확률변수, edge는 변수들 사이의 조건부 관계를 의미한다.

Bayesian Network



$$p(x_1, x_2, x_3, x_4, x_5) \\ = p(x_3|x_1)p(x_2|x_1)p(x_4|x_2, x_3)p(x_5|x_4)$$

- Design our **Belief** to graphical structure
- 결합확률분포를 얻을 수 있다.

03

Gibbs Sampling



Intuition

$$\theta_1^{(j)} \sim p\left(\theta_1 | \theta_2^{(j-1)}, \dots, \theta_K^{(j-1)}\right)$$

$$\theta_2^{(j)} \sim p\left(\theta_2 | \theta_1^{(j)}, \theta_3^{(j-1)}, \dots, \theta_K^{(j-1)}\right)$$

⋮

$$\theta_k^{(j)} \sim p\left(\theta_k | \theta_1^{(j)}, \dots, \theta_{k-1}^{(j)}, \theta_{k+1}^{(j-1)}, \dots, \theta_K^{(j-1)}\right)$$

⋮

$$\theta_K^{(j)} \sim p\left(\theta_K | \theta_1^{(j)}, \dots, \theta_{K-1}^{(j)}\right)$$

- 고차원 결합확률 분포는 계산하기 어렵다.
- 거기서 sample들을 뽑아보자
- 그런데 한 번에 뽑기 어려우니 다른 것들이 주어 졌다고 치고 하나씩 뽑자
- 다 뽑으면 그것이 하나의 sample이다.
- 여기 까지만 알자.

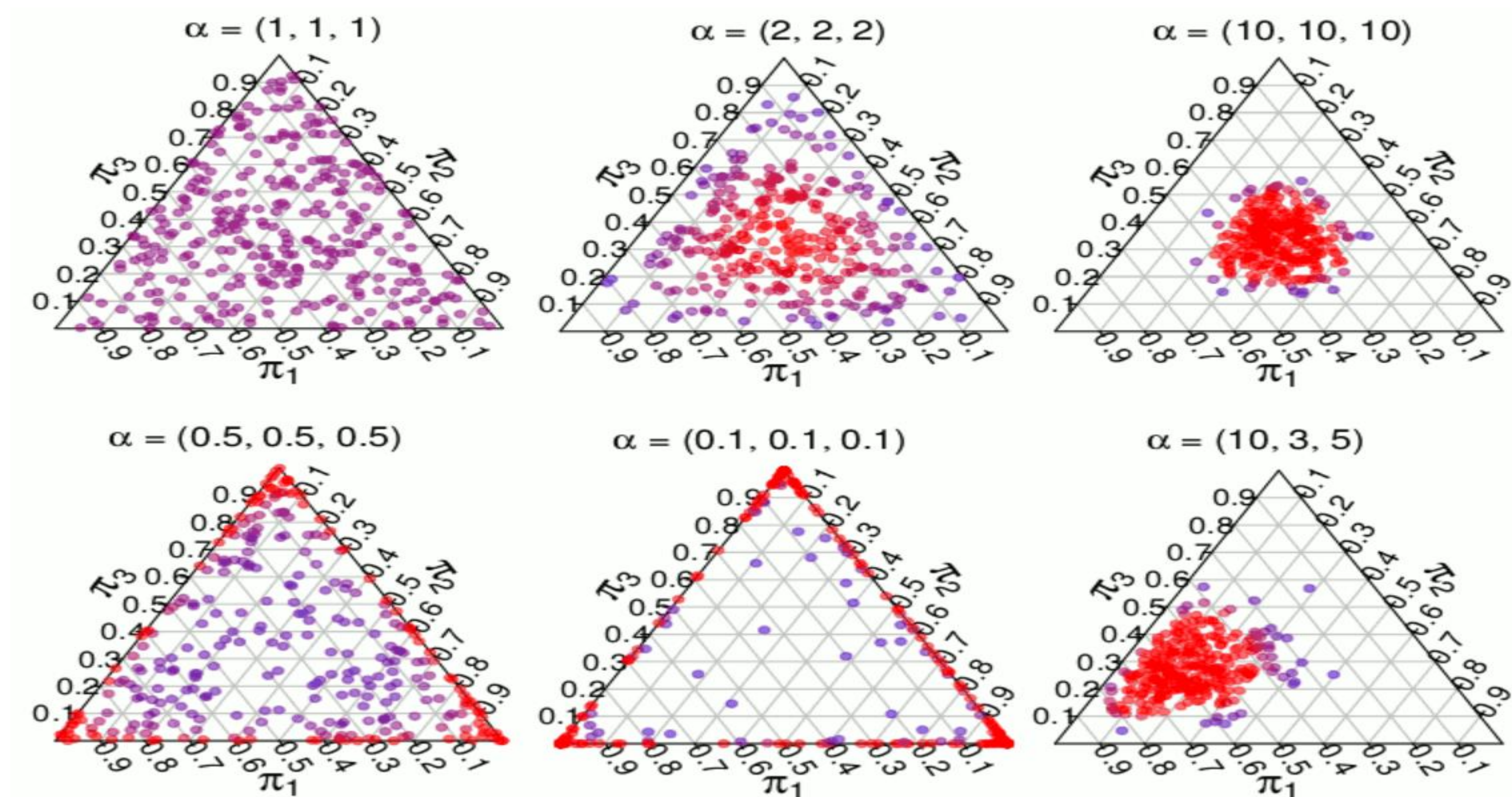
04

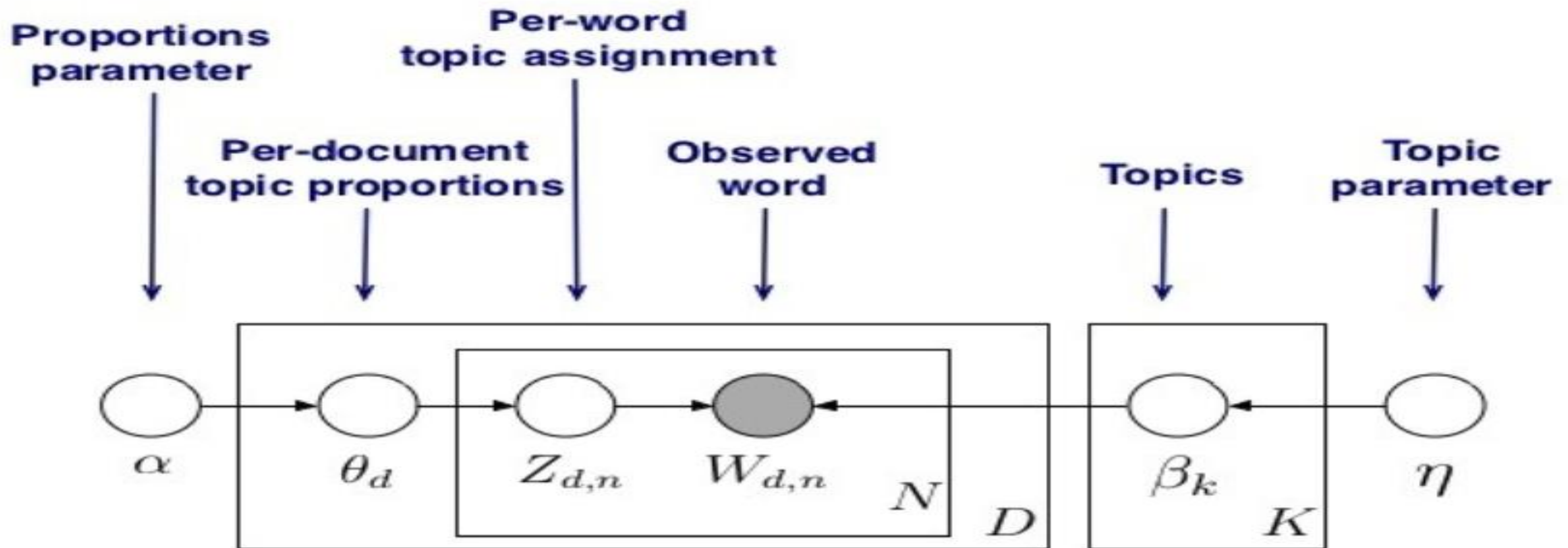
Model Structure



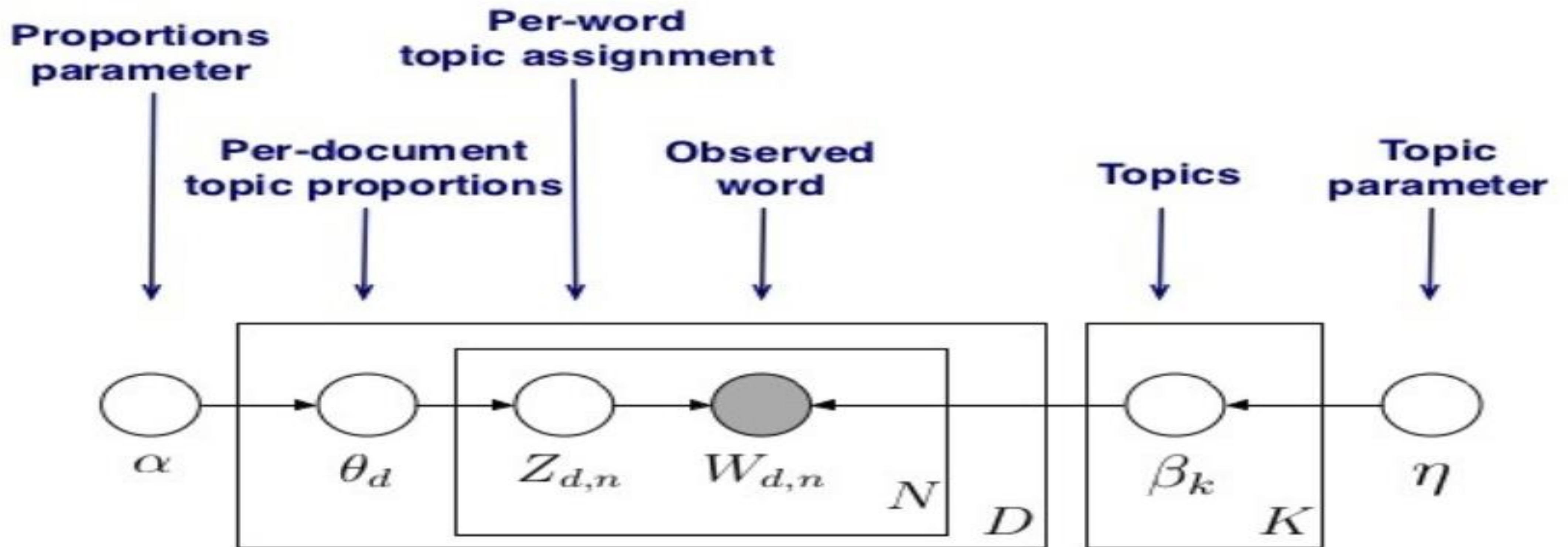
Dirichlet Distribution

$$f(x_1, \dots, x_k; \alpha_1, \dots, \alpha_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)} \prod_{i=1}^k x_i^{\alpha_i - 1}, \sum_{i=1}^k x_i = 1, x_i \geq 0$$





- $\alpha \in \mathbb{R}^K$: 문서별 주제에 대한 Dirichlet 사전 분포 모수
- $\eta \in \mathbb{R}^V$: 주제별 단어에 대한 Dirichlet 사전 분포
- $\theta_d \in \mathbb{R}^K$: 문서 d 에 대한 주제 분포
- $\beta_k \in \mathbb{R}^V$: 주제 k 에 대한 단어 분포
- $Z_{d,n}$: 문서 d 의 n 번째 단어의 주제
- $W_{d,n}$: 문서 d 의 n 번째 단어

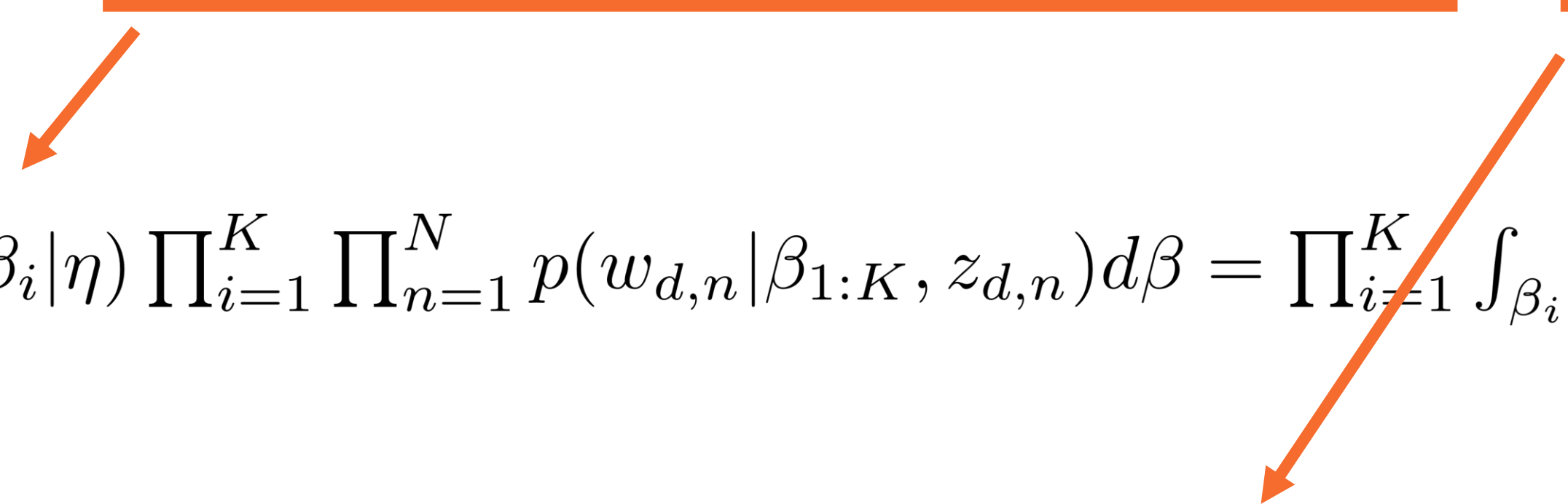


$$p(\beta, \theta, z, w | \alpha, \eta) = \prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

05 Tedious Computation

1. Marginalization

Start with initializing latent topic to all words

$$P(z, w | \alpha, \eta) = \int_{\theta} \int_{\beta} p(\beta, \theta, z, w | \alpha, \eta) d\beta d\theta$$
$$= \int_{\beta} \prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D \prod_{n=1}^N p(w_{d,n} | \beta_{1:K}, z_{d,n}) d\beta \times \int_{\theta} \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) d\theta$$


$$\int_{\beta} \prod_{i=1}^K p(\beta_i | \eta) \prod_{i=1}^K \prod_{n=1}^N p(w_{d,n} | \beta_{1:K}, z_{d,n}) d\beta = \prod_{i=1}^K \int_{\beta_i} p(\beta_i | \eta) \prod_{d=1}^D \prod_{n=1}^N p(w_{d,n} | \beta_{1:K}, z_{d,n}) d\beta_i$$

$$\int_{\theta} \prod_{d=1}^D p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) d\theta = \prod_{d=1}^D \int_{\theta_d} p(\theta_d | \alpha) \prod_{n=1}^N p(z_{d,n} | \theta_d) d\theta_d$$

2. Build Gibbs Sampler

$$P(Z, W | \alpha, \eta) = \prod_{d=1}^D \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \frac{\prod_{i=1}^K \Gamma(n_d^i + \alpha_i)}{\Gamma(\sum_{i=1}^K (n_d^i + \alpha_i))} \times \prod_{i=1}^K \frac{\Gamma(\sum_{v=1}^V \eta_v)}{\prod_{v=1}^V \Gamma(\eta_v)} \frac{\prod_{v=1}^V \Gamma(n_v^i + \eta_v)}{\Gamma(\sum_{v=1}^V (n_v^i + \eta_v))}$$

This enable you calculate below form

$$\begin{aligned} P(Z_{(d,n)} = k | Z_{-(d,n)}, W, \alpha, \eta) &\propto P(Z_{(d,n)} = k, Z_{-(d,n)}, W, \alpha, \eta) \\ &\propto (n_{d,-(d,n)}^k + \alpha_k) \times \frac{n_{v,-(d,n)}^k + \eta_v}{\sum_{v'=1}^V (n_{v',-(d,n)}^k + \eta_{v'})} \end{aligned}$$

- $n_{d,-(d,n)}^k + \alpha_k$: $W_{d,n}$ 를 제외하고 문서 d 에서 topic k 인 단어의 개수 + 문서 d 가 topic k 인 경향에 비례한다.
- $n_{v,-(d,n)}^k + \eta_v$: $W_{d,n}$ 를 제외하고 단어 뭉치 중 하나인 v 가 topic k 인 개수 + 단어 v 가 나오는 경향
- $\sum_{v'=1}^V (n_{v',-(d,n)}^k + \eta_{v'})$: 확률화

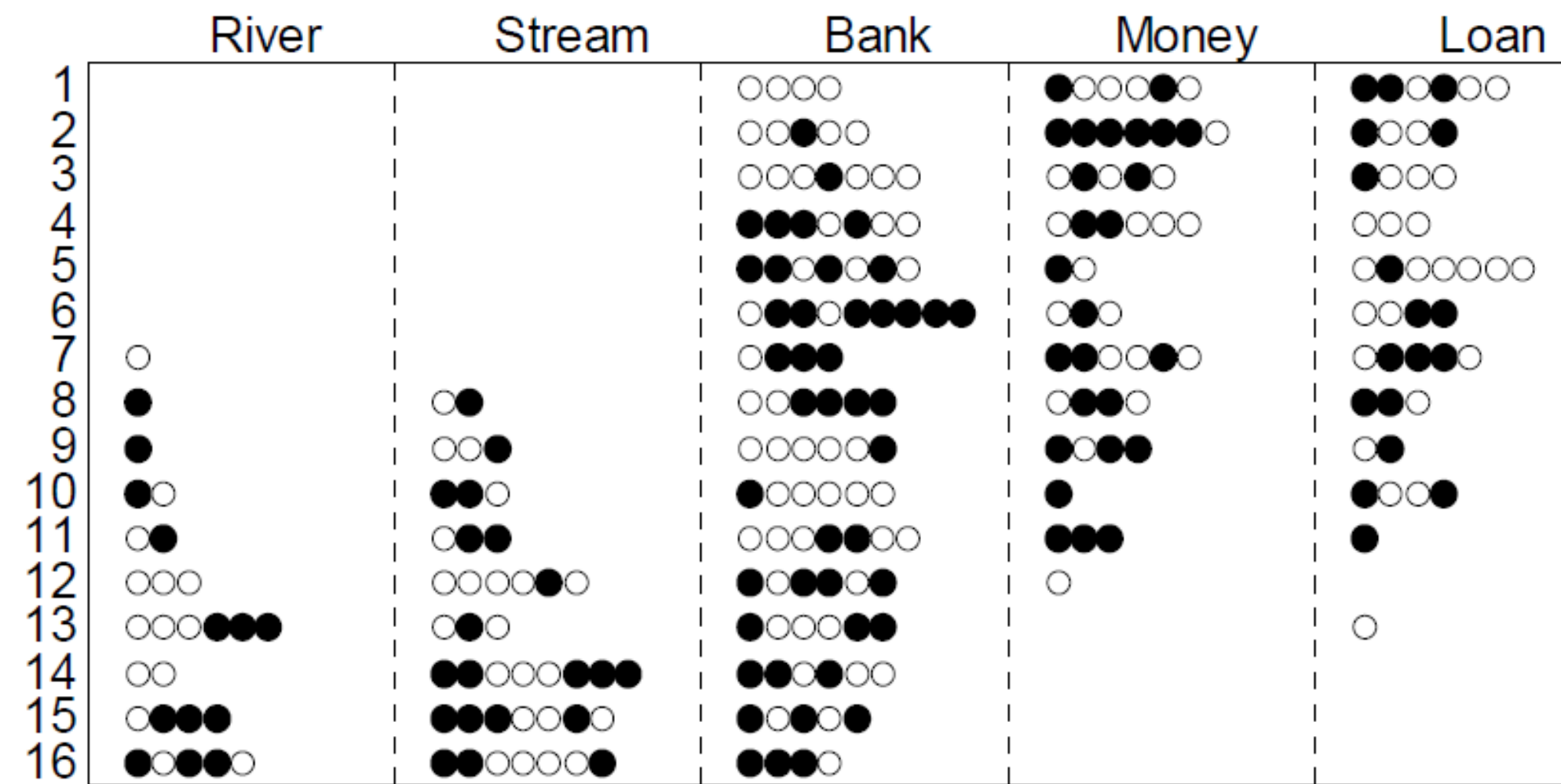
Compute Other Distribution

After all, we have latent topic for each word.

For computing topic-word distribution and document-topic distribution, just counting is enough.

- topic k 의 단어 분포 $\in \mathbb{R}^V : \beta_k = \frac{n_v^k + \eta}{\sum_{v=1}^V n_v^k + V\eta}$
 - 문서 전체에서 topic k 로 할당된 단어 킷치 속의 단어 수
- 문서 d 의 주제 분포 $\in \mathbb{R}^K : \theta_d = \frac{n_d^k + \alpha}{\sum_{i=1}^K n_d^i + K\alpha}$
 - 문서 내에서 topic k 로 할당된 단어의 수

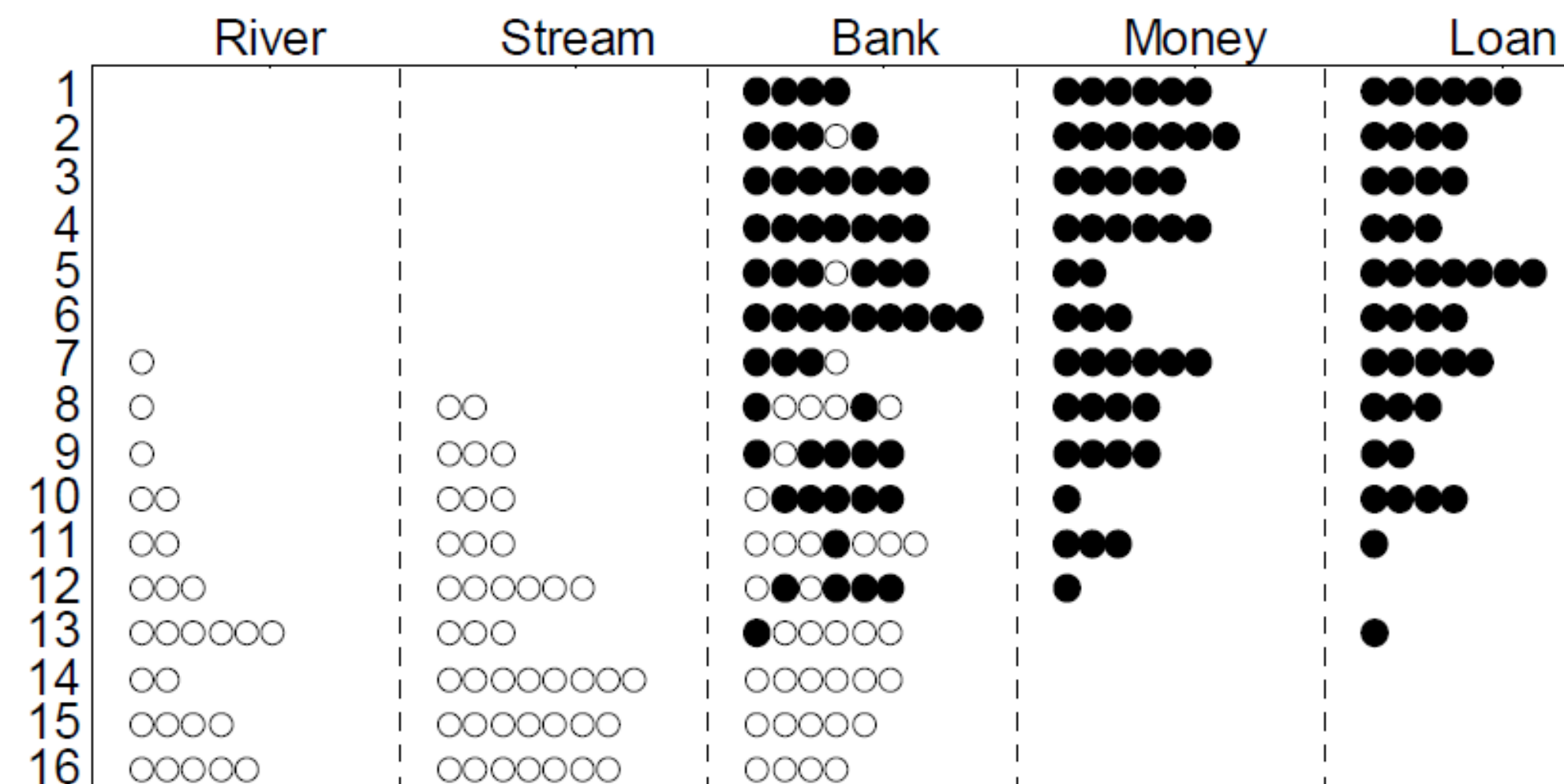
EXAMPLE



θ, β

estimation

Iteration



Variational Inference

- approximation to distribution

$$P(Z|X) \approx Q(Z)$$

- minimize distance between two distributions

$$D_{KL}(Q||P) = \sum_Z Q(Z) \log \frac{Q(Z)}{P(Z|X)}$$

$$D_{KL}(Q||P) = \sum_Z Q(Z) [\log \frac{Q(Z)}{P(Z,X)} + \log P(X)]$$

Variational Inference

- X is already observed i.e. **constant**

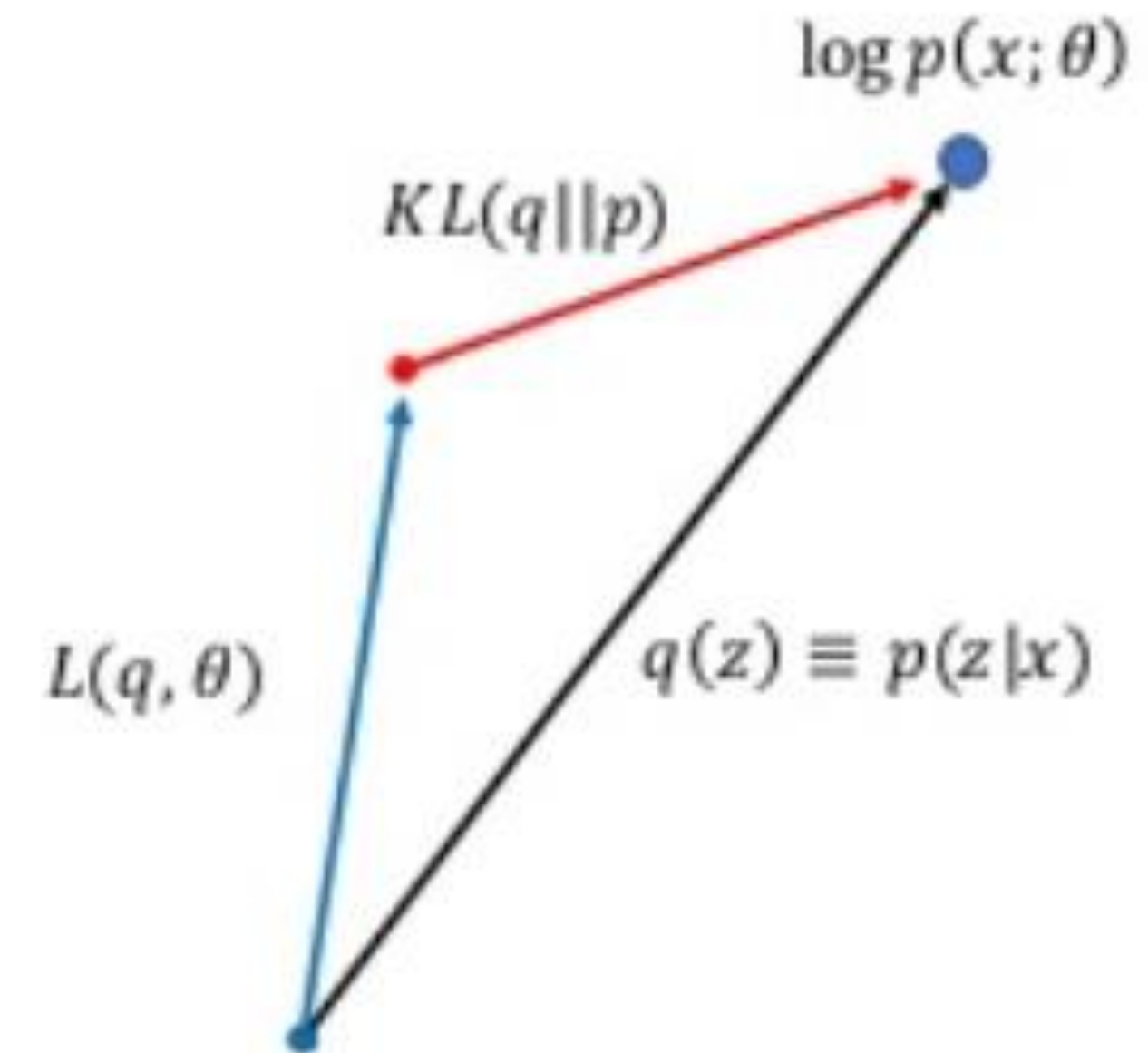
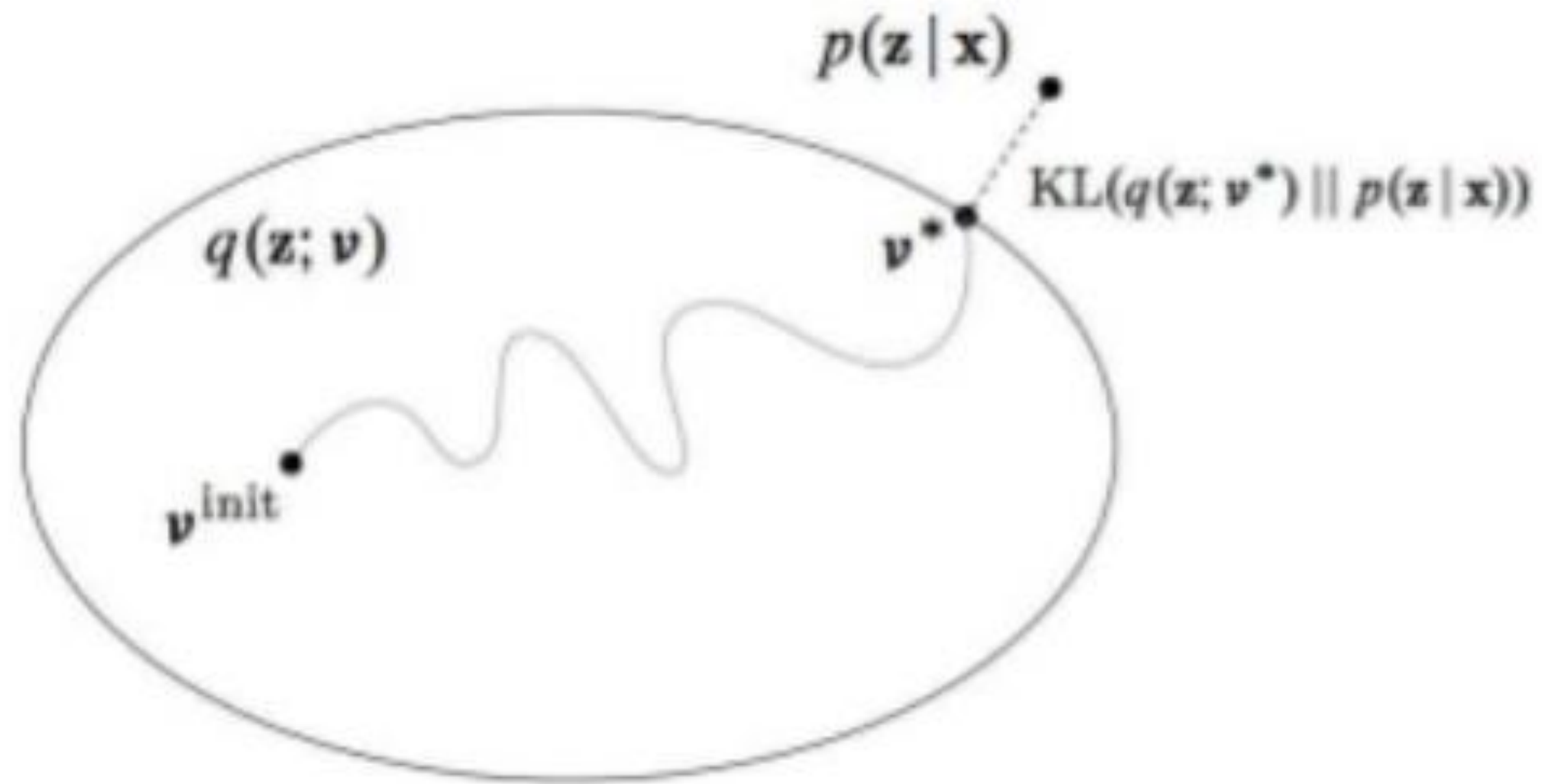
$$P(X)$$

- To minimizing distance between two distributions, maximize right term

$$D_{KL}(Q||P) = \sum_Z Q(Z) [\log \frac{Q(Z)}{P(Z,X)} + \log P(X)]$$

$$\log P(X) = D_{KL}(Q||P) - \sum_Z Q(Z) [\log \frac{Q(Z)}{P(Z,X)}]$$

Variational Inference



06 Conclusion



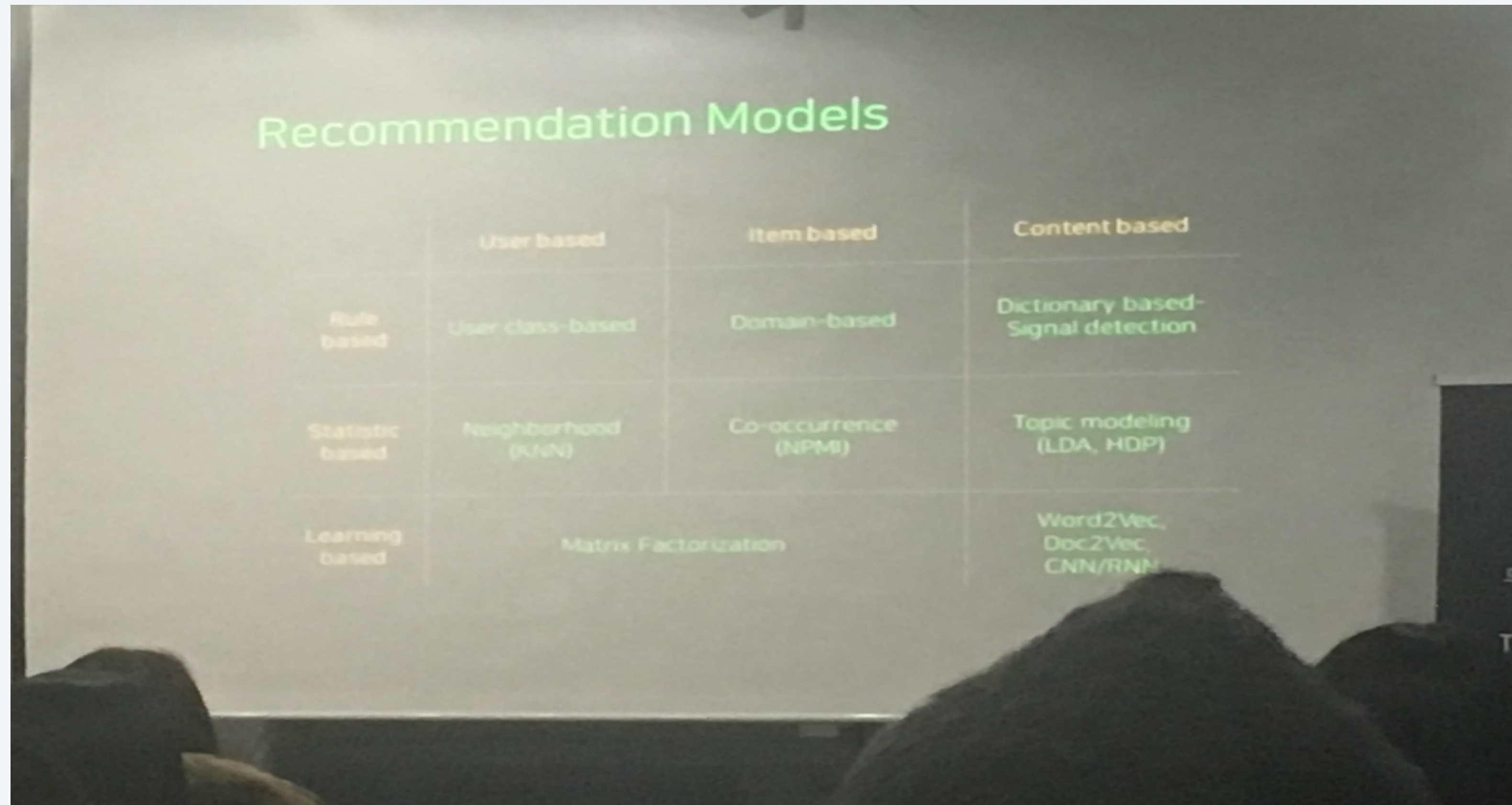
Again Objective

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

- We now have assigned topics to words

Application



The image shows a presentation slide titled "Recommendation Models" in green text. The slide contains a table with three columns: "User based", "Item based", and "Content based". The rows are categorized by model type: "Rule based", "Statistic based", and "Learning based".

	User based	Item based	Content based
Rule based	User class-based	Domain-based	Dictionary based-Signal detection
Statistic based	Neighborhood (KNN)	Co-occurrence (NPM)	Topic modeling (LDA, HDP)
Learning based	Matrix Factorization		Word2Vec, Doc2Vec, CNN/RNN

- various field including NLP, experiment

Thank You