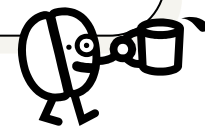


# 階層 Pitman-Yor トピックモデル



Keywords:

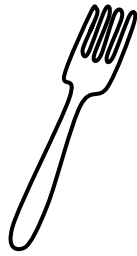
教師なし学習, ノンパラメトリックベイズ, 言語モデル  
文書の確率的生成モデル

IBISML 2010/06/14

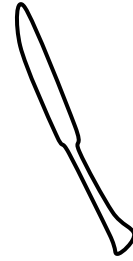
○佐藤一誠<sup>1,2</sup>, 中川裕志<sup>1</sup>

東京大学<sup>1</sup>

日本学術振興会特別研究員(DC1)<sup>2</sup>



# Today's Menu



- ◆ Background:
  - Topic model, Latent Dirichlet allocation(LDA)
- ◆ Restaurant representation for LDA
- ◆ Pitman Yor Topic model (PYTM)
- ◆ Experiments

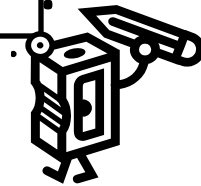
# Topic model

[Hofmann, 1999][Blei+, 2001]

The purpose of topic models is

to find short descriptions preserving the statistical relationships of words in documents,

and predict new words or

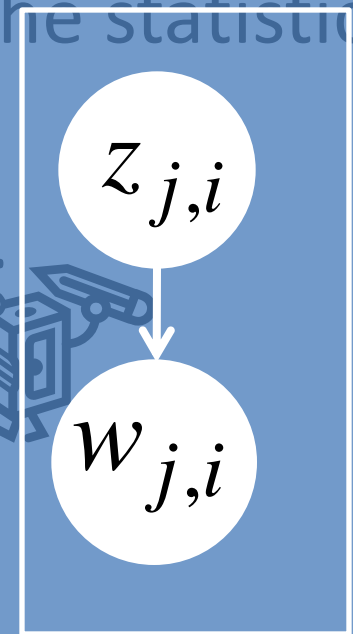


# Topic model

[Hofmann, 1999][Blei+, 2001]

The purpose of topic models is to find short descriptions preserving the statistical relationships of words in documents, and predict new words or documents.

**Latent Dirichlet Allocation(LDA)**  
[Blei+, 2001]



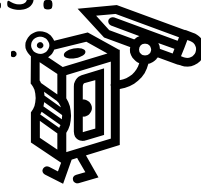
# Topic model

[Hofmann, 1999][Blei+, 2001]

The purpose of topic models is

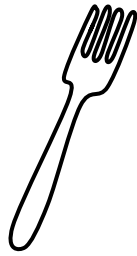
to find short descriptions preserving the statistical relationships of words in documents,  
and predict new words or documents.

$$p(w | D) = \sum_z p(w | z) p(z | D)$$

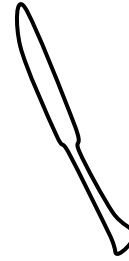


Our contribution

To capture the power-law phenomenon of a word distribution, known in linguistics as Zipf's law.



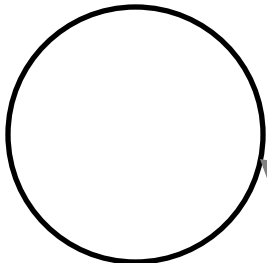
# Today's Menu



- ◆ Background:
  - Topic model, Latent Dirichlet allocation(LDA)
- ◆ Restaurant representation for LDA
- ◆ Pitman Yor Topic model (PYTM)
- ◆ Experiments

Restaurant=document

Table

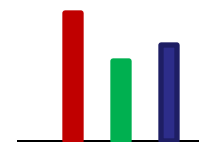
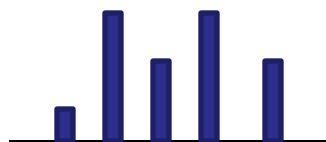
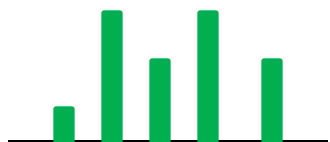
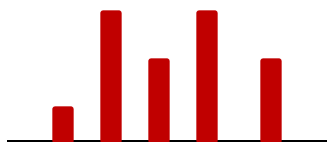


New customer

Topic 1

Topic 2

Topic 3

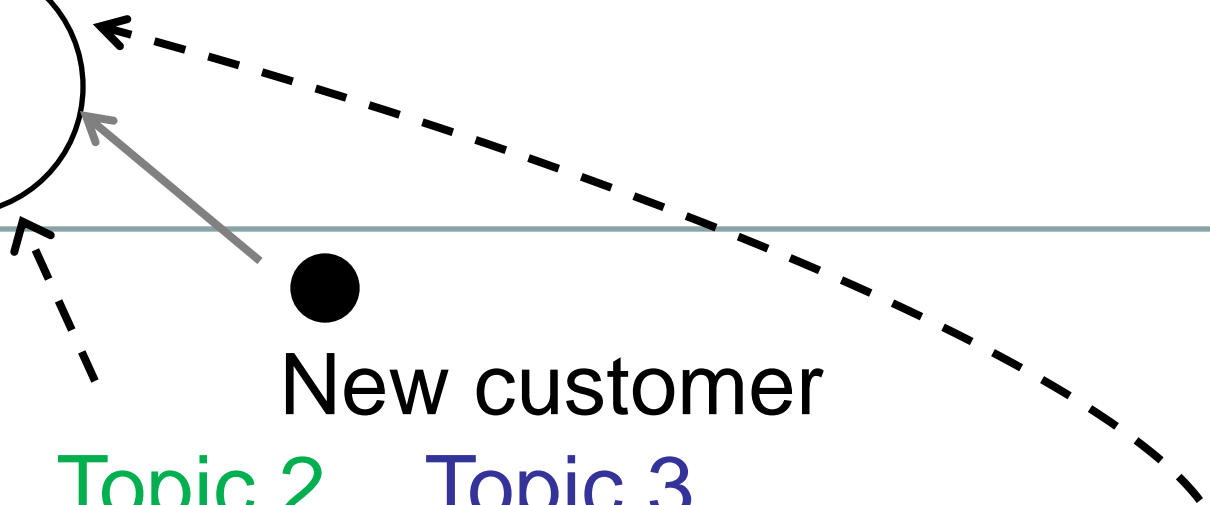


Word distributions

Topic distribution

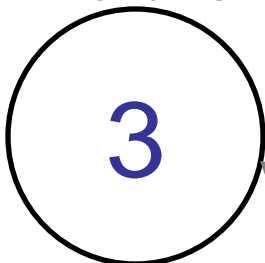
$$p(w | z)$$

$$p(z | D)^7$$



Restaurant=document

Table

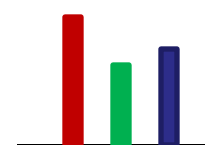
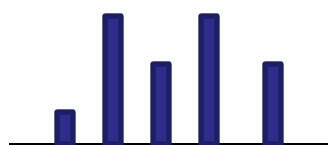
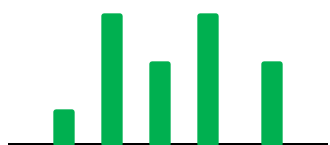
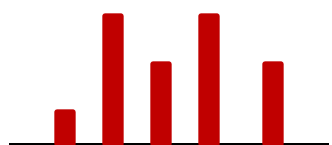


New customer

Topic 1

Topic 2

Topic 3



Word distributions

Topic distribution

$$p(w | z)$$

$$p(z | D)^8$$

Restaurant=document

Table

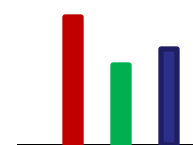
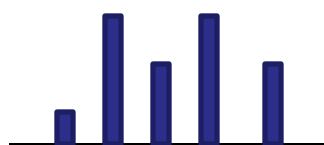
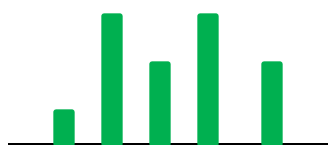
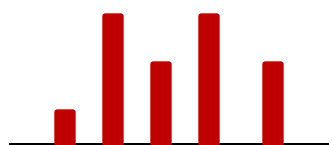
the

New customer

Topic 1

Topic 2

Topic 3

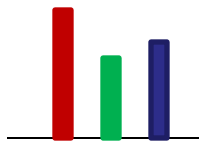
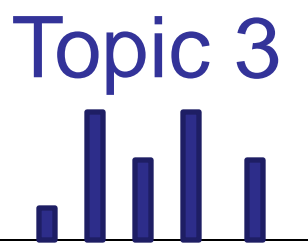
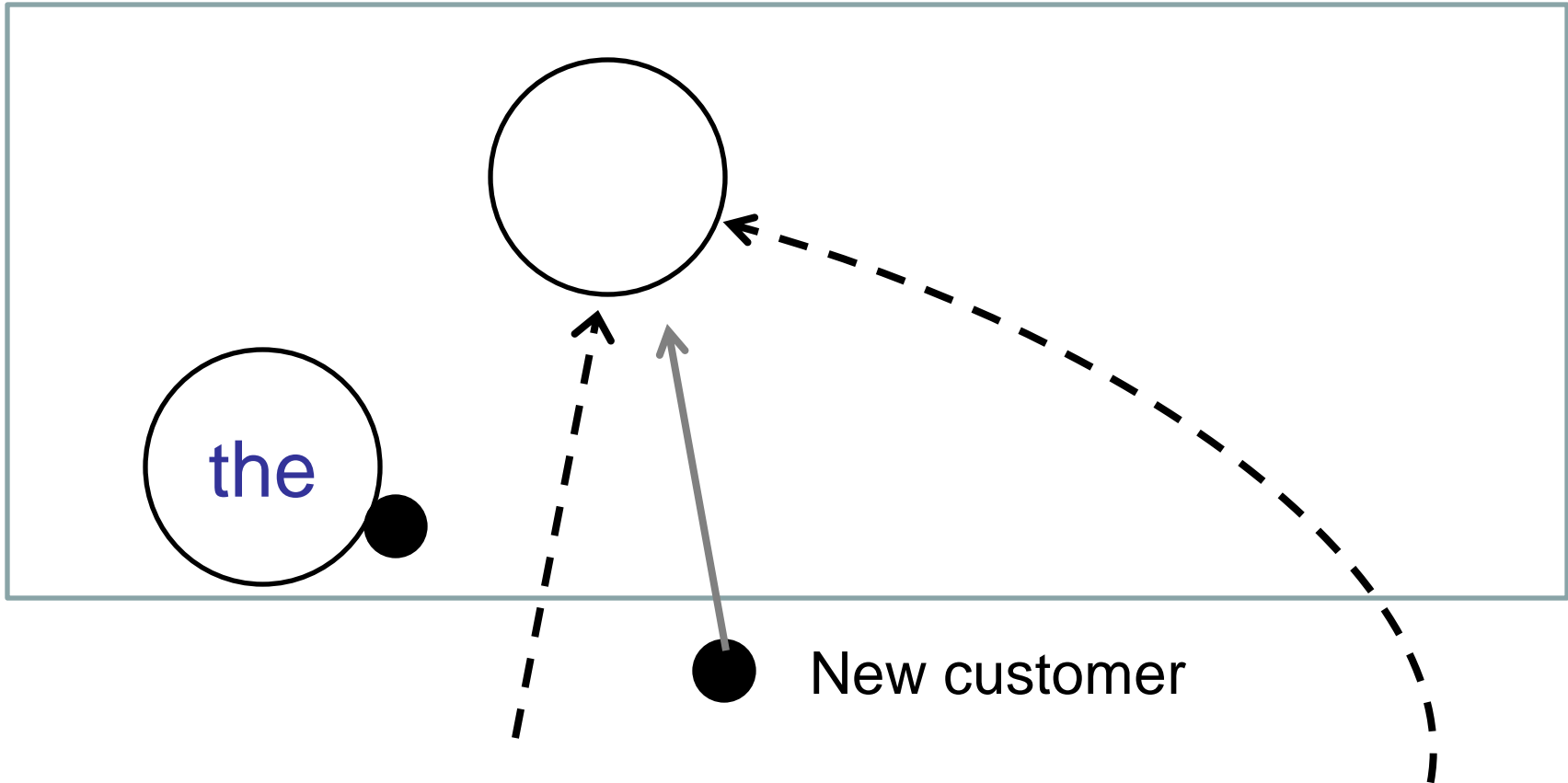


Word distributions

Topic distribution

$$p(w | z)$$

$$p(z | D)^9$$

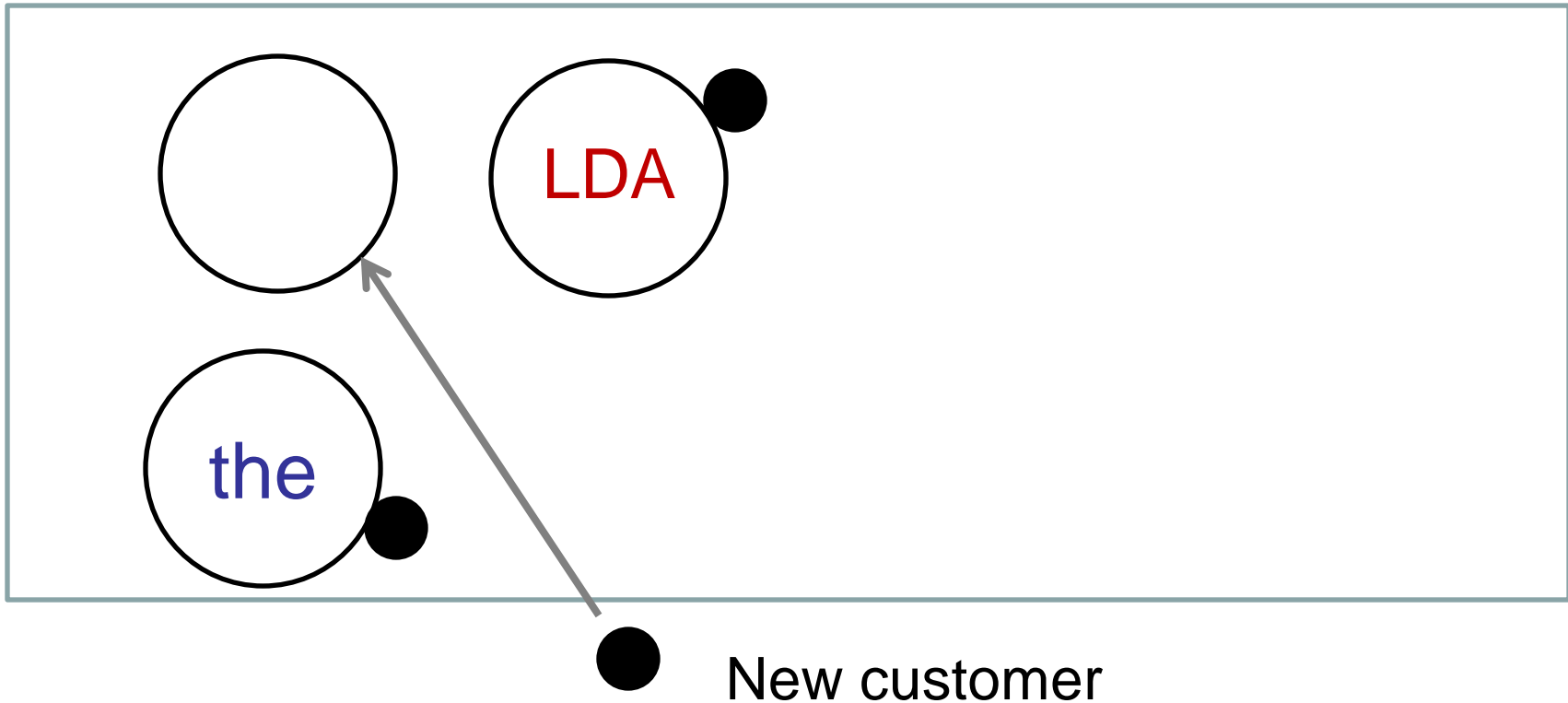


Word distributions

Topic distribution

$$p(w | z)$$

$$p(z | D)^{10}$$



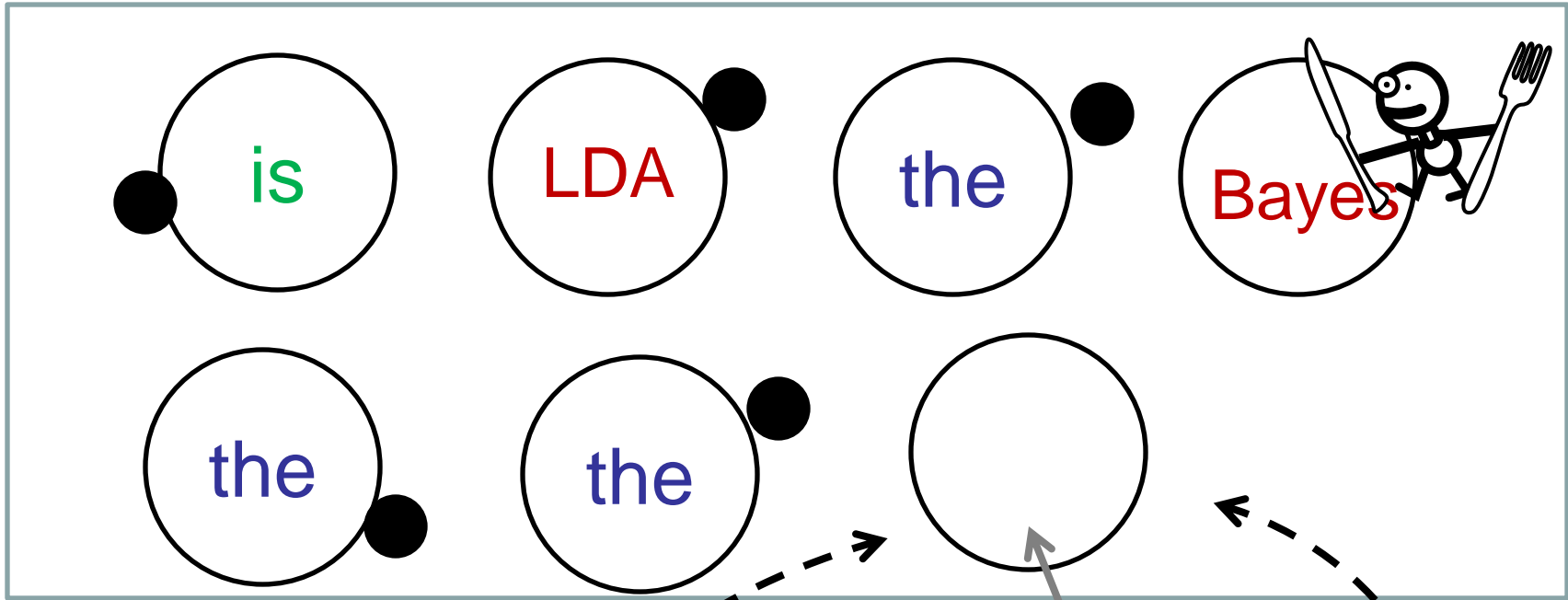
Word distributions

$$p(w | z)$$



Topic distribution

$$p(z | D)^{11}$$

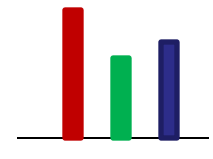
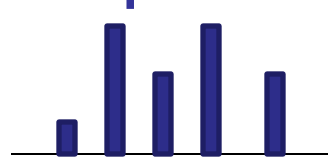
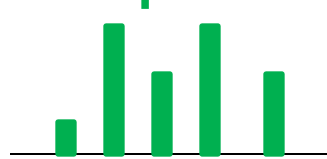
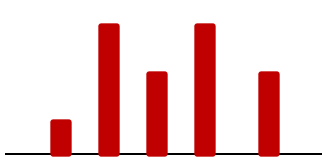


Topic 1

Topic 2

Topic 3

New customer

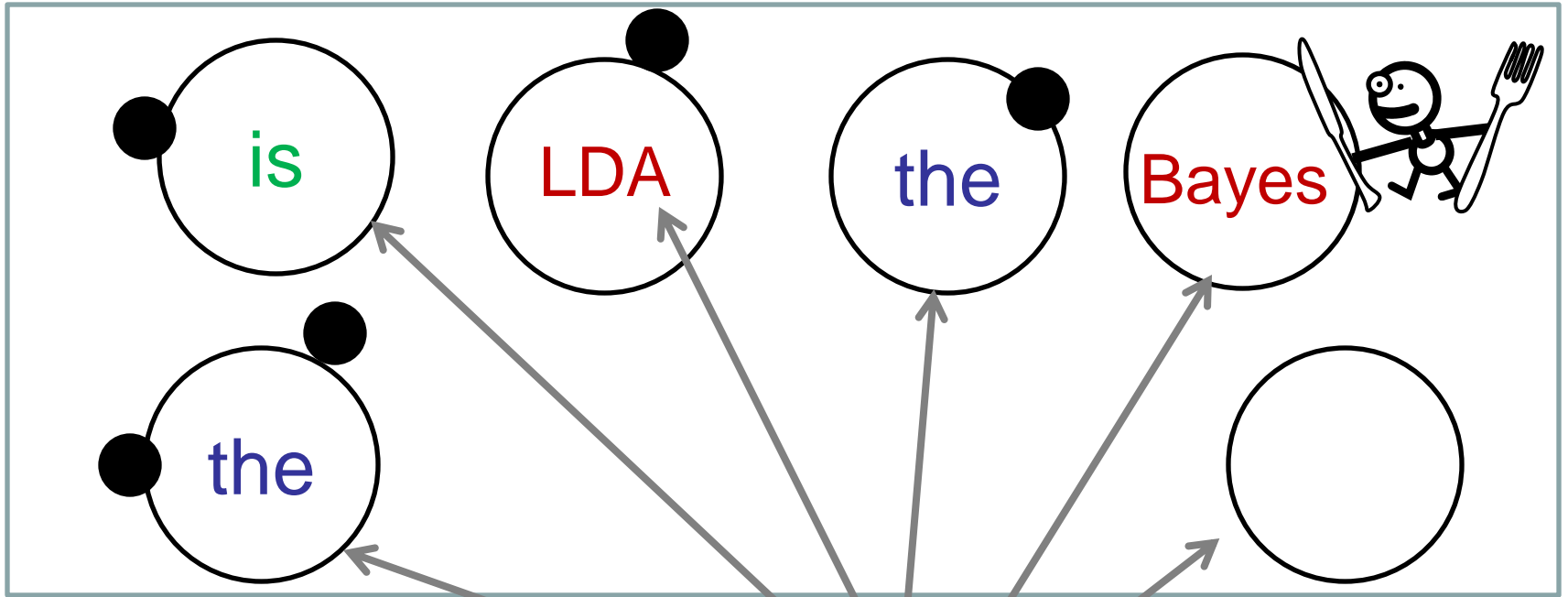


Word distributions

Topic distribution

$$p(w | z)$$

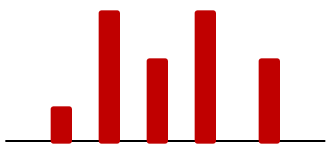
$$p(z | D)^{12}$$



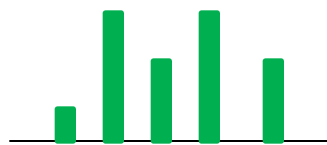
Proposed model

New customer

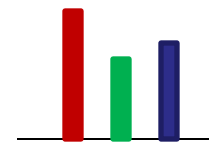
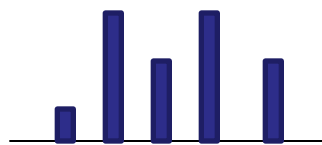
Topic 1



Topic 2

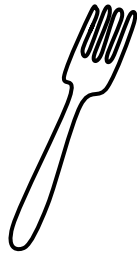


Topic 3

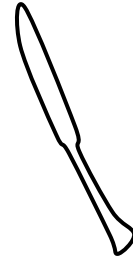


Topic distribution

Word distributions



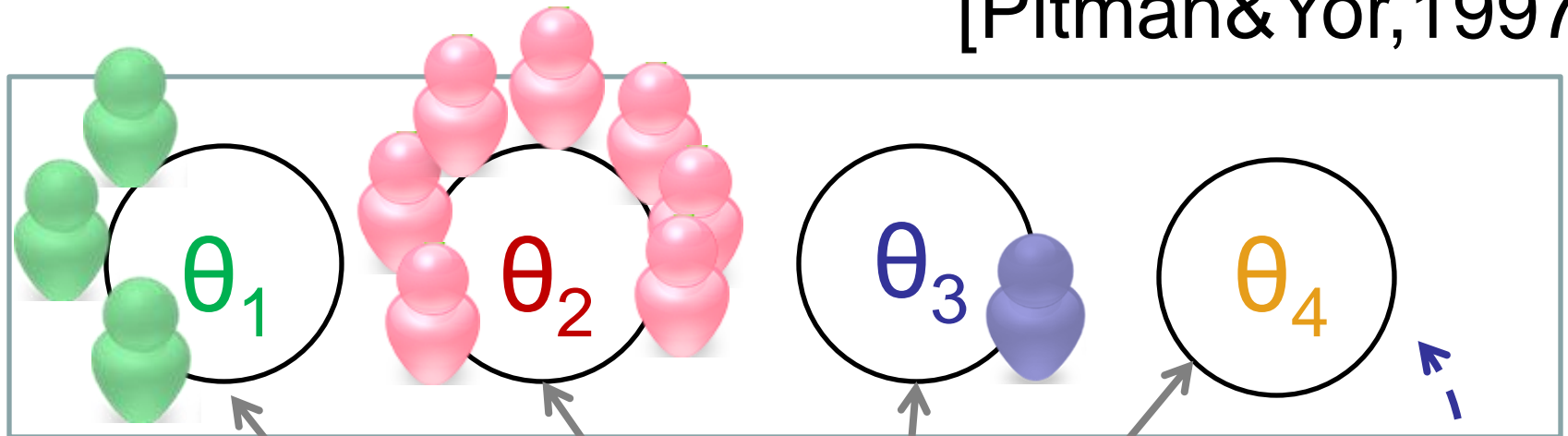
# Today's Menu



- ◆ Background:
  - Topic model, Latent Dirichlet allocation(LDA)
- ◆ Restaurant representation for LDA
- ◆ Pitman Yor Topic model (PYTM)
  - Pitman-Yor process
- ◆ Experiments

# Pitman-Yor process

[Pitman&Yor,1997]



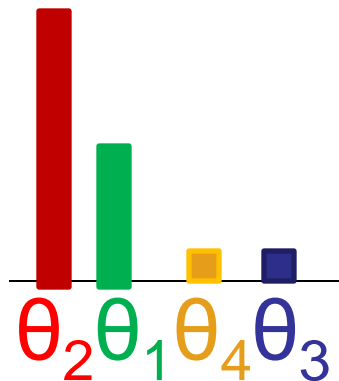
$$\frac{3-d}{\gamma+11}$$

$$\frac{7-d}{\gamma+11}$$

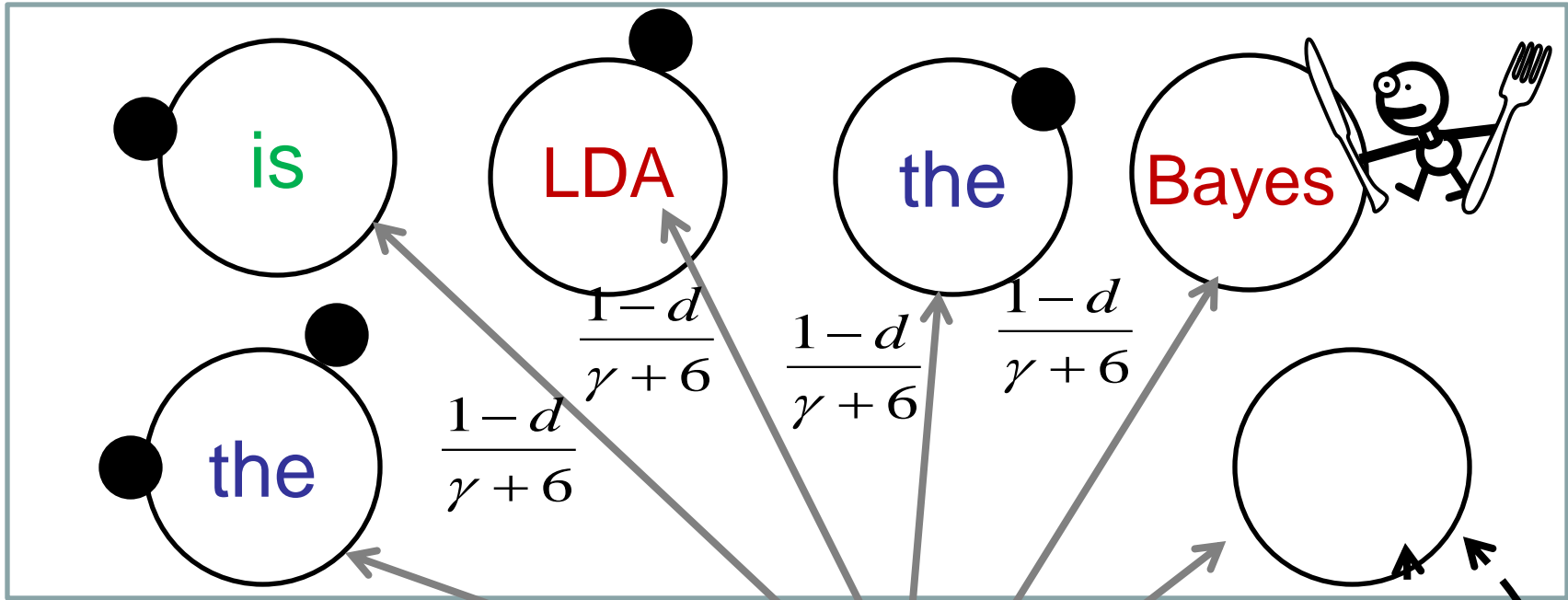
$$\frac{1-d}{\gamma+11}$$

$$\frac{\gamma+3d}{\gamma+11}$$

New customer  $\theta_4 \sim G_0$



$$G \sim PY(\gamma, d, G_0)$$



$$\frac{2-d}{\gamma+6}$$

$$\frac{1-d}{\gamma+6}$$

$$\frac{1-d}{\gamma+6}$$

$$\frac{1-d}{\gamma+6}$$

$$\frac{1-d}{\gamma+6}$$

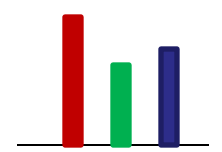
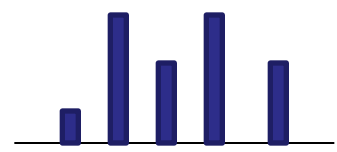
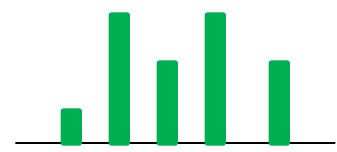
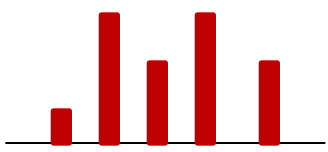
$$\frac{\gamma+5d}{\gamma+6}$$

New customer

Topic 1

Topic 2

Topic 3



Topic distribution

Word distributions

# PYTMの予測分布

$$p(w_j^{new} = v | D)$$

vのあるテーブル数

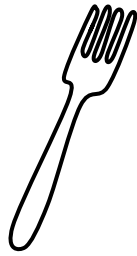
全テーブル数

LDAの予測分布

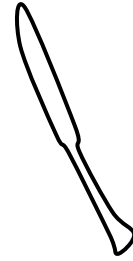
$$= \frac{N_{j,v} - dK_{j,v}}{\underbrace{\gamma + N_j}} + \frac{\gamma + dK_{j,\cdot}}{\gamma + N_j} \underbrace{p^{LDA}(w_j^{new} = v | D)}$$

単語数

単語vが既出の場合に影響のある項

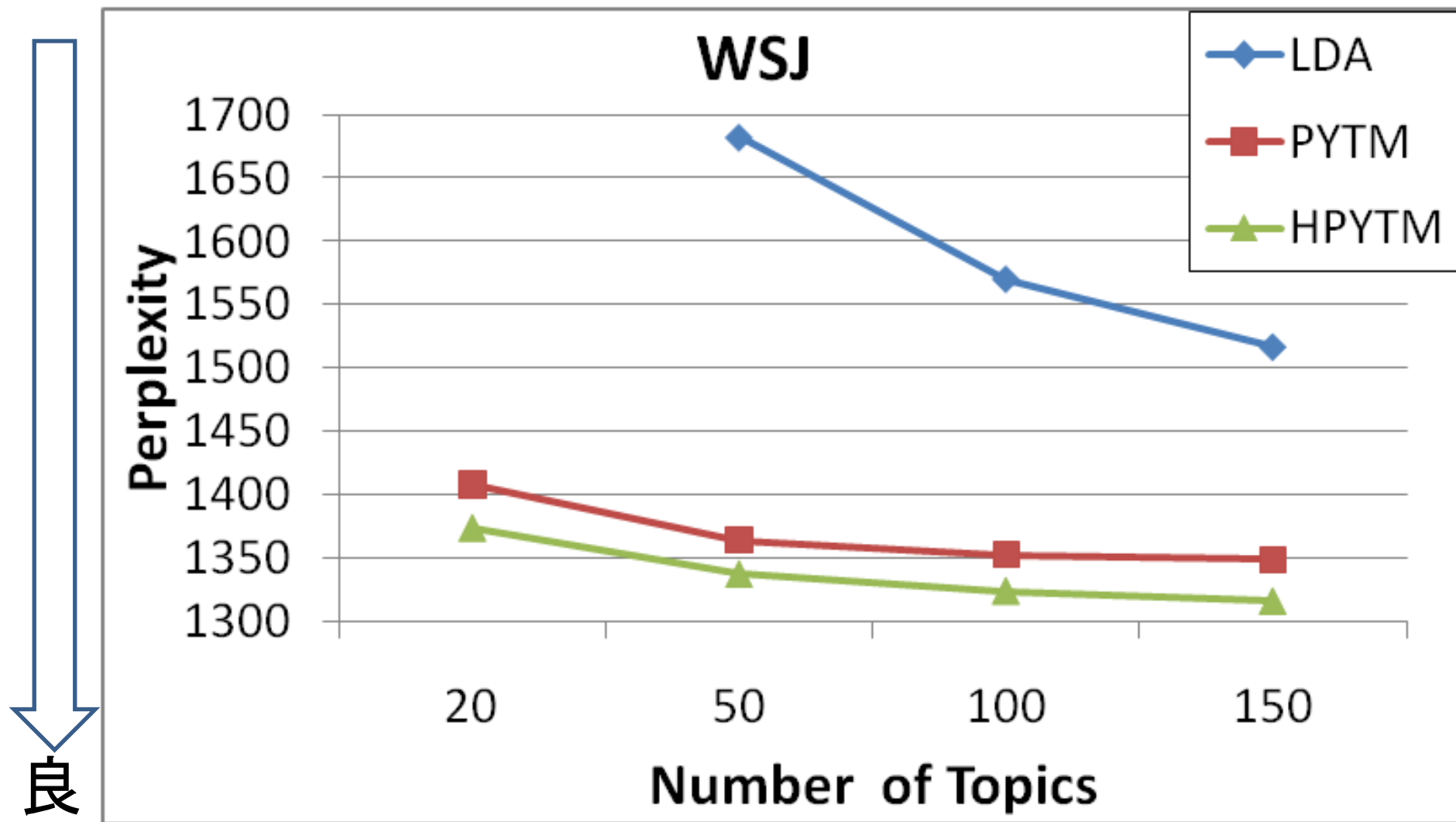


# Today's Menu



- ◆ Background:
  - Topic model, Latent Dirichlet allocation(LDA)
- ◆ Restaurant representation for LDA
- ◆ Pitman Yor Topic model (PYTM)
- ◆ Experiments

5,000 documents, training 90%, test 10%



ご清聴ありがとうございました

