NTT

トピックモデルの応用：イントロダクション

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

- まずは本日の講義のイントロダクションです

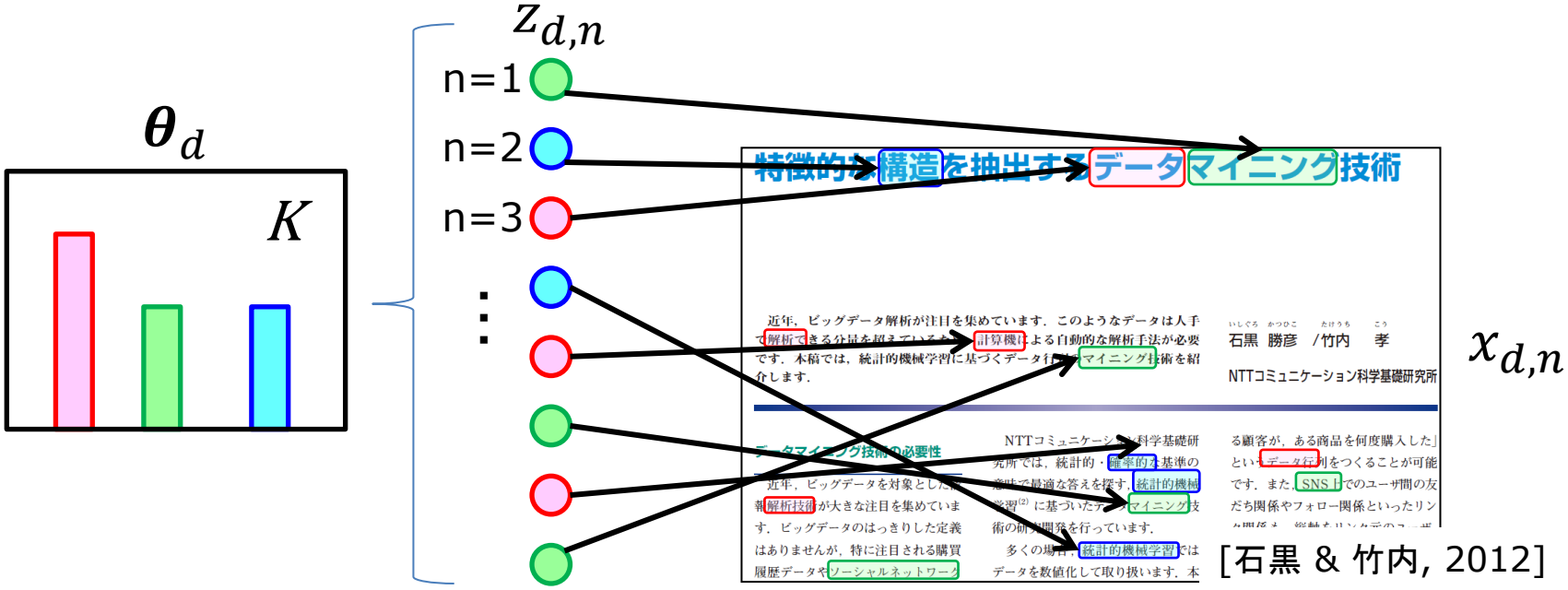
- LDAの復習
  - 使用する変数, モデルなど
  - 解法: Gibbs sampler, VB-EM
- 今日の講義全体に関する注意
  - notation policyなど
  - 参考文献

本日の予定

- LDAの拡張手法と応用について紹介します
  - 基本的には、様々な論文の説明になります
- 午前: トピックモデル(LDA)の拡張モデル
- 午後: トピックモデルの各種ドメインデータへの応用

Latent Dirichlet Allocaion [Blei, 2003]

- 様々な離散データに隠された潜在的なトピックを推定するベイジアンモデル



[石黒 & 竹内, 2012]

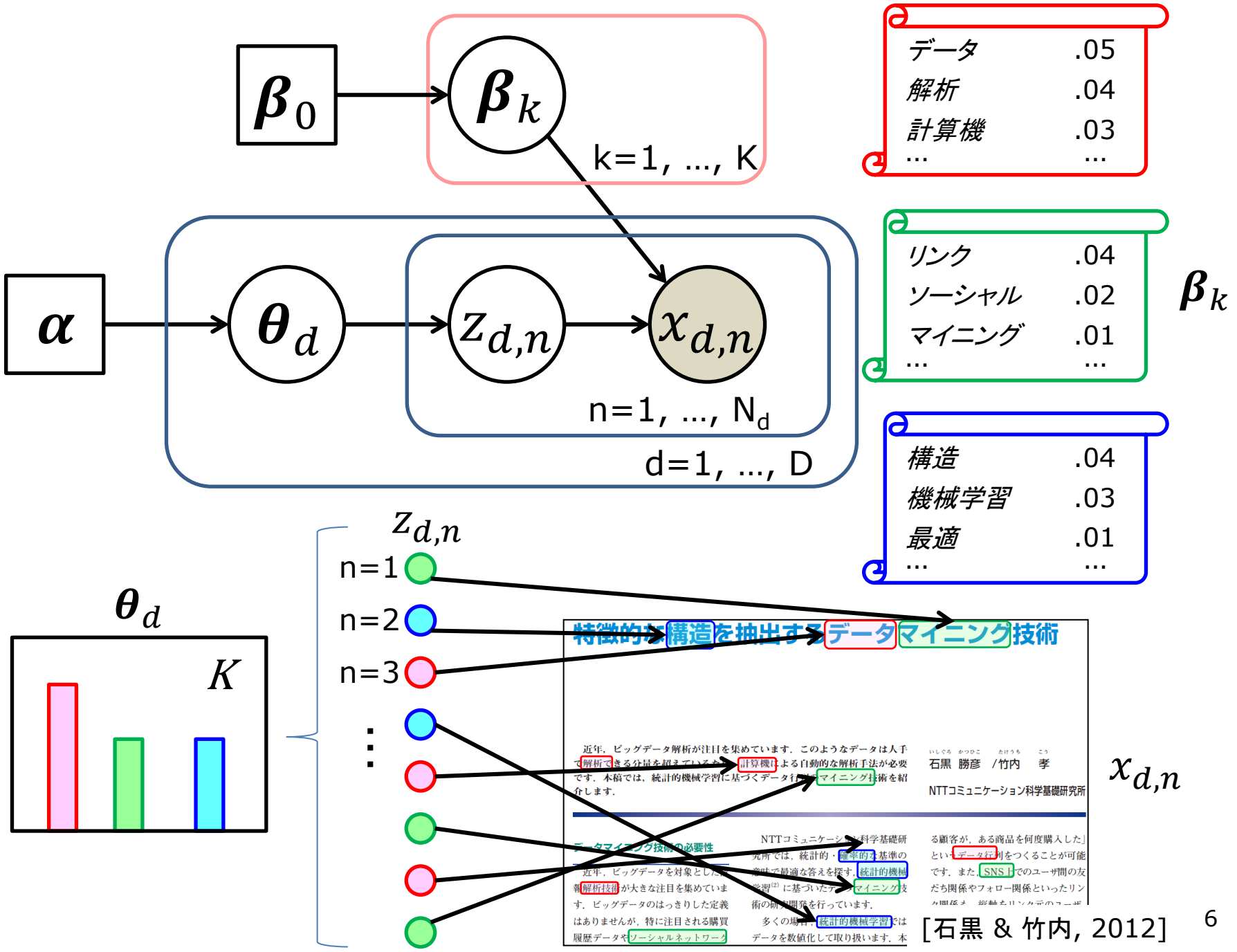
インデックス、定数、変数

- 文書インデックス d
- トピックインデックス k
- 単語インデックス n

- 文章数 D
- トピック数 K
- 文書 d 中の単語数 N_d
- 単語の種類 V

- 観測された単語 x
- 単語のtopic assignment z
- 文書のtopic proportion θ
- トピックのword proportion β

- θの事前分布パラメータ α
- βの事前分布パラメータ β_0


β_0
β_k
k=1, …, K
α
θ_d
$z_{d,n}$
$x_{d,n}$
n=1, …, N_d
d=1, …, D
データ .05
解析 .04
計算機 .03
… …
リンク .04
ソーシャル .02
マイニング .01
… …
β_k
構造 .04
機械学習 .03
最適 .01
… …
$z_{d,n}$
n=1
n=2
n=3
θ_d
K
特徴的な構造を抽出するデータマイニング技術
$x_{d,n}$

[石黒 & 竹内, 2012]

生成モデル

for 文書d = 1, 2, …, D_t

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-word assignment

$$z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$$

word observation

$$x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k|\boldsymbol{\beta}_0\sim\mathrm{Dir}(\boldsymbol{\beta}_0)$

topic-word proportionは事前分布を仮定しない場合も多々あります

Gibbs sampler

- 最も正確な解を得ることができる解法です
- 各文書dの単語nを一つずつトピックkに割り当てていきます

$$p\left(z_{d,n}=k|x_{d,n}=w,\boldsymbol{X}_{\neg(d,n)},\boldsymbol{Z}_{\neg(d,n)},\boldsymbol{\alpha},\boldsymbol{\beta_0}\right)\propto \frac{m_{dk}+\alpha_k}{\sum_{k'}(m_{dk'}+\alpha_{k'})}\frac{m_{kw}+\beta_{0,w}}{\sum_{w'}\left(m_{kw'}+\beta_{0,w'}\right)}$$

文書 dから トピック kが 生成される確率

トピック kから 単語 wが 生成される確率

m_{dk}: 文書 d 内でトピック k にアサインされた観測量のカウント($x_{d,n}$を除く)

m_{kw}: 文書全体でトピック k にアサインされた観測量のうち単語 w だった観測量のカウント($x_{d,n}$を除く)

変分ベイズ法(VB-EM)

- 文書データの周辺化尤度をJensen不等式で下から押さえて、それを最大化する変分事後分布$q()$を求める手法です

$$\log p(\boldsymbol{X}) = \iint \log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})\, d\boldsymbol{Z} d\boldsymbol{\theta}$$
$$\geq \iint q(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X}) \log \frac{p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta})}{q(\boldsymbol{Z}, \boldsymbol{\theta}|\boldsymbol{X})}\, d\boldsymbol{Z} d\boldsymbol{\theta}$$

$$q\left(z_{d,n} = k\right) \propto \beta_{k,x_{d,n}} \exp\left(\Psi\left(\hat{\theta}_{d,k}\right)\right)$$

$$\hat{\theta}_{d,k} = \alpha_k + \sum_n q\left(z_{d,n} = k\right)$$

Notationについて

- 直観的な理解のために、できる限り同じ意味を持つ量は同じ名前の変数・インデックスで表します
- また、一部のモデルについてはよりわかりやすく等価なモデルで説明します

- したがって、原論文とは変数の名前やグラフィカルモデルの形が違う可能性があります

全体を通じての重要な参考文献

- Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
  - LDAの原論文です

全体を通じての重要な参考文献

- ビショップ、”パターン認識と機械学習”, 丸善出版, 2012
  - “PRML本”: 和書の中では、現状、機械学習に関するもっとも良い本の一つです

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [PRML] Blei and Lafferty, “A Correlated Topic Model of Science”, The Annals of Applied Statistics, Vol. 1(1), pp. 17-35, 2007.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.

NTT

トピックモデルの応用：相関・構造をもつトピックモデル

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

- 機械学習の研究分野では、日々新しい、より柔軟で表現力の高い(≒複雑な☹)トピックモデルが提案されています

- このスライドでは、それらのうち、特に構造化に関する仕事を厳選してご紹介します

トピックモデルの大きな特長はモデルの単純さです

- 誤解を恐れずにいえば、単純な混合ガウシアンモデル(GMM)が理解できれば、LDAは理解できます

- GMMがその単純さゆえに非常に幅広いドメインの連続データで有効なように、LDAも幅広いドメインの離散データで有効です

トピックモデルの問題点も モデルの単純さです

- モデルが単純ということは、大胆な仮定を置いてデータを表現していることになります
- 実際のデータと明らかに合わない仮定の場合、これを正す必要があります

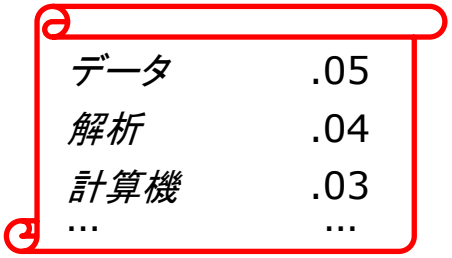
- 沢山の複雑化したトピックモデルが提案されています

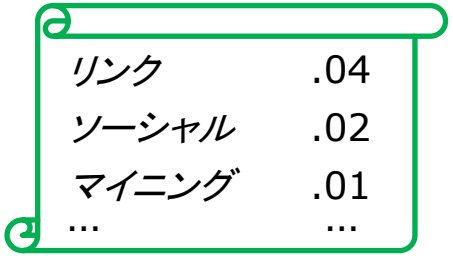
Correlated Topic Models
[Blei & Lafferty, 2007]

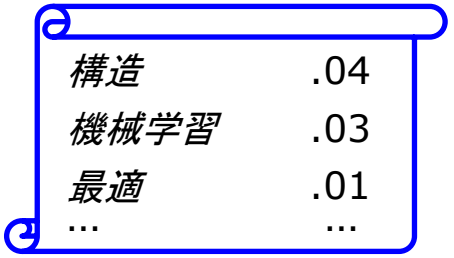
Blei and Lafferty,
"A Correlated Topic Model of Science",
The Annals of Applied Statistics,
Vol. 1(1), pp. 17-35, 2007.

トピックモデルの大前提の仮定: トピックは独立

- 簡単にいうと:「各トピック k の間には相関がない」
- 通常のGMMでも共有される考え方です
- これのおかげで各種モデル推論が簡単になっています







トピックは本当に独立なのか？

"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

[Blei, 2003]

TREC APコーパス: AP通信ニュースの文書データセット

トピックは本当に独立なのか？

- 先の例だけで分かるように、これは成り立たないことが多々ありそうです

- すなわち、「本当は相関のあるトピック」を無理やり「相関のないトピック」に分割している可能性が高いです

提案法: Correlated Topic Models (CTM)

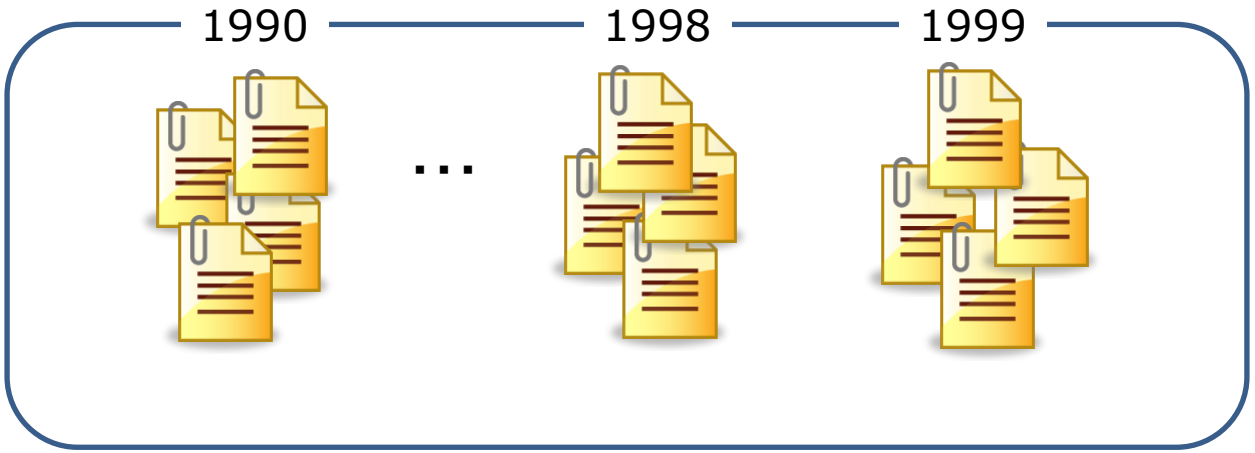
- ☺ 以後のトピックモデル研究に非常に大きな影響を与えたモデルです

- 科学誌ScienceのOCRデータを用いて、科学論文のトピック解析を行います
- トピック間の相関（正・負）をexplicitにモデル化します

- 推論は少々面倒になります

対象データ: Science誌

- 1880年にエジソンによって刊行された、非常に著名な科学論文誌
- OCRされた論文誌データ(JSTOR)を利用
- 実験では1990年代の論文を対象とします



近代の科学は分野横断的です

- 先進のbioinformaticsは高度な統計学の知識とデータマイニングの手法が必要です
- 分子動力学法は物理学に則っていますが、化学・生物学の多様な系に応用されます

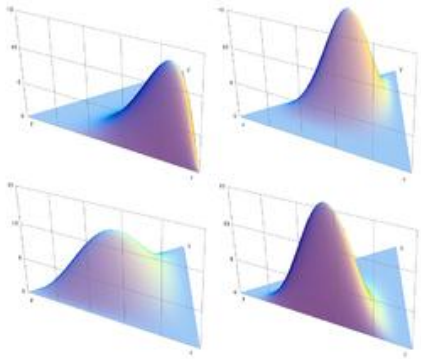
- Science誌は専門分野の論文誌ではないため、このような分野間の相関構造が強く表れるはずです

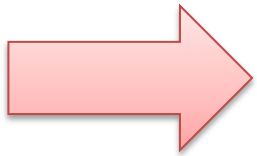
提案法のアイデア: とにかく簡単に相関を埋め込みます

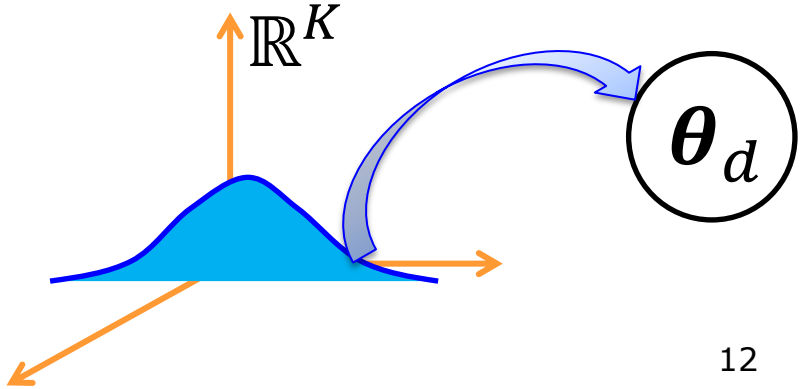
- 目的: 文書 d のtopic proportion θ_d を生成する際に、トピックの相関を埋め込む
- 解: 一番簡単な相関を持つ分布といえば多次元正規分布なので、素直にそれを使う

Dirichlet**分布: 足して1にするだけ**

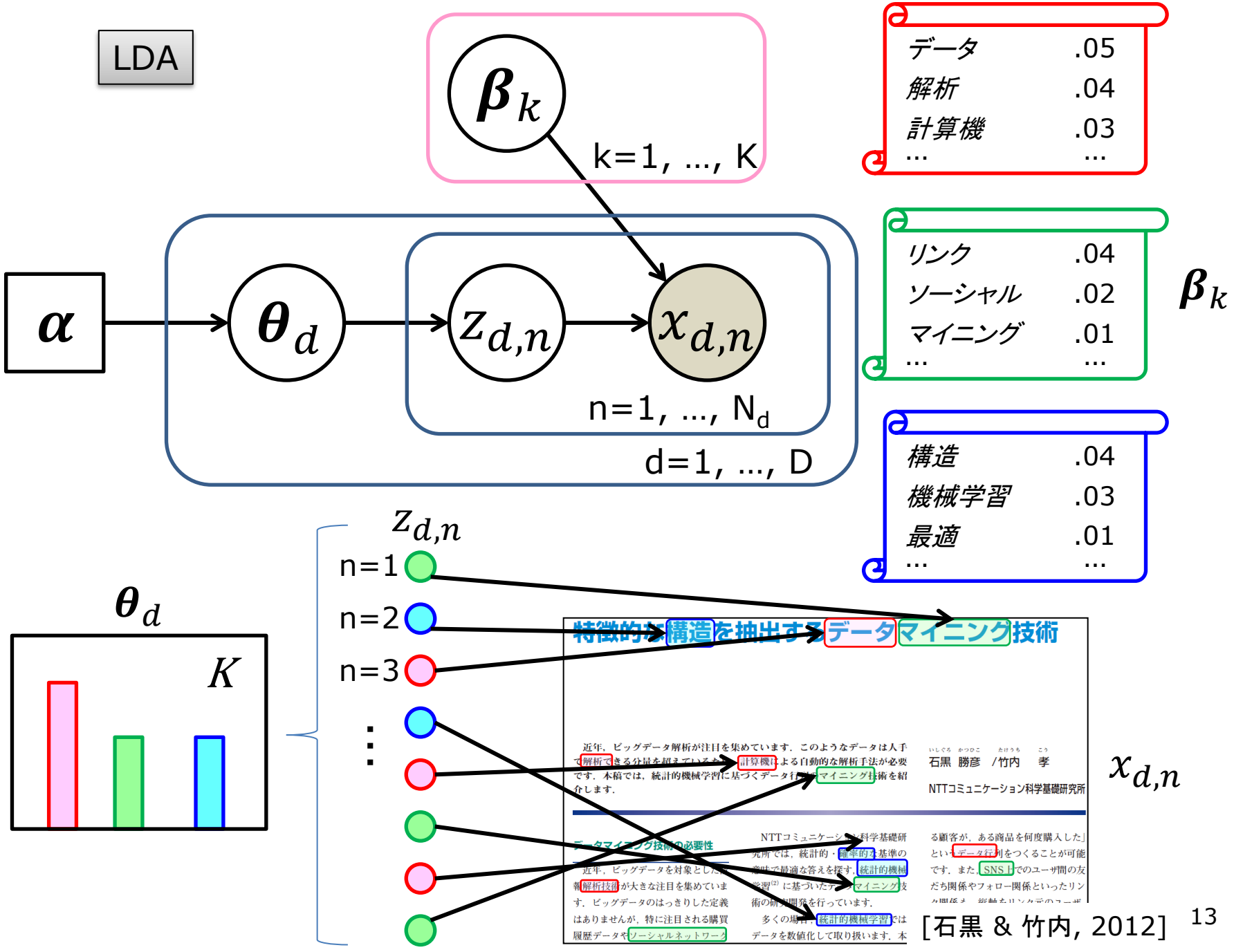
多次元正規分布: "一緒に値が動く" "片方が増えるともう片方が減る" などを表現できる





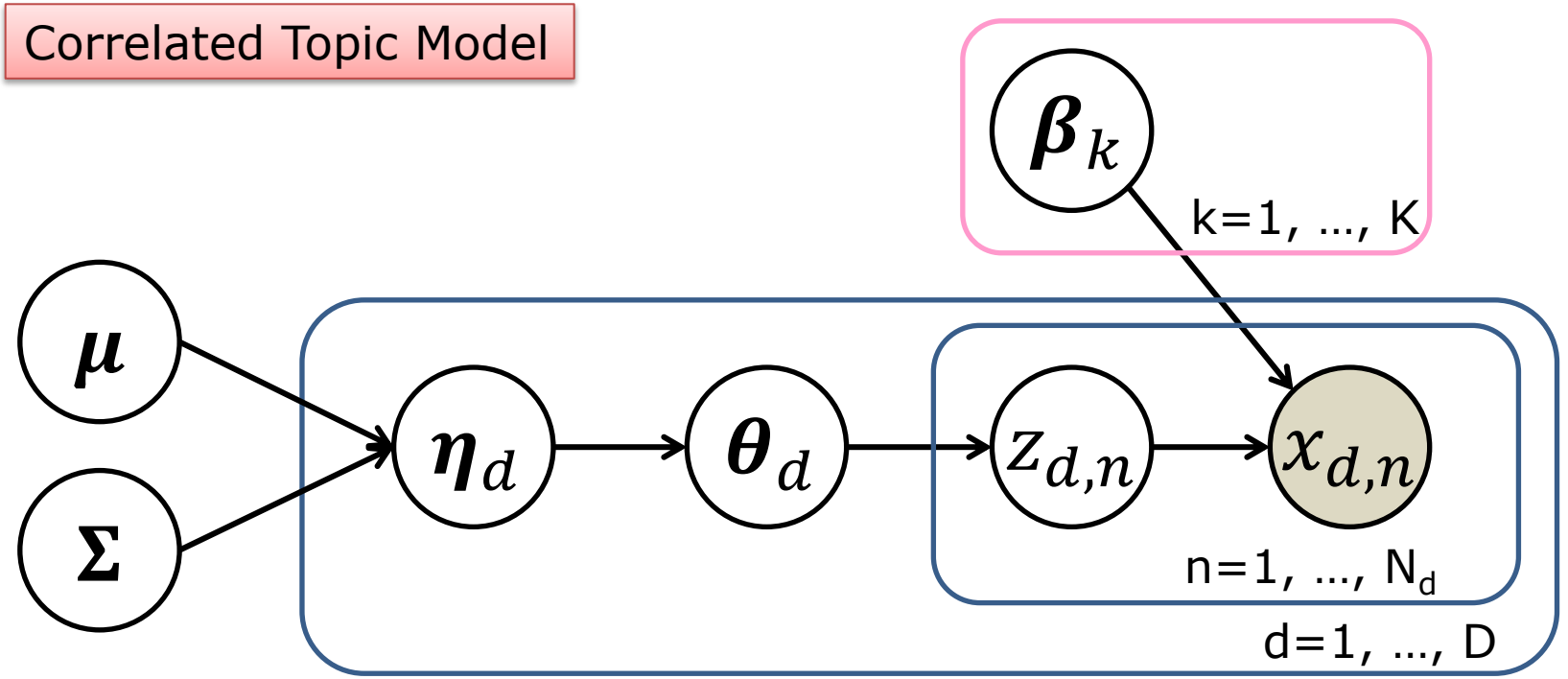


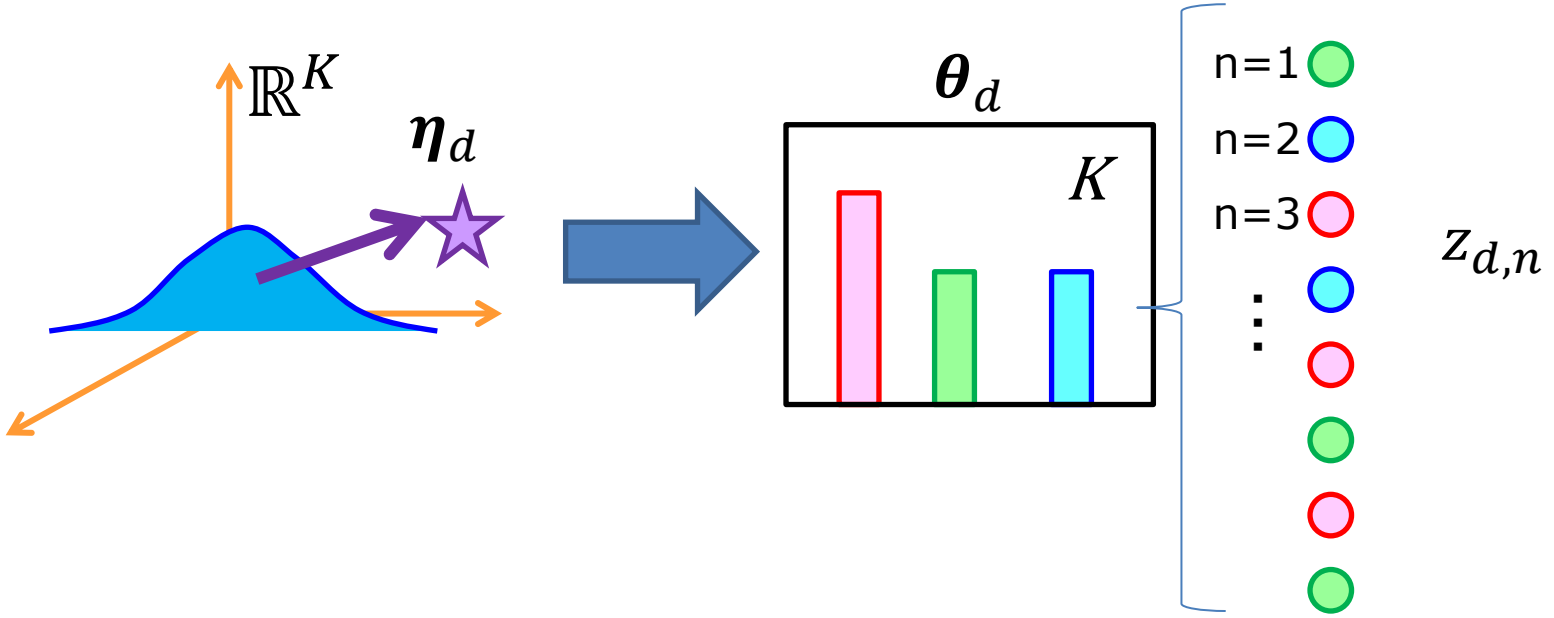
LDA



[石黒 & 竹内, 2012]

Correlated Topic Model





生成モデル

for 文書d = 1, 2, …, D_t

topic proportion

$$\boldsymbol{\eta}_d|\boldsymbol{\mu},\boldsymbol{\Sigma}\sim\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$$

$$\boldsymbol{\theta}_d|\boldsymbol{\eta}_d=\pi(\boldsymbol{\eta}_d)$$

for 単語n = 1, 2, …, N_d

topic-word assignment

$$z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$$

word observation

$$x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$$

for トピック k = 1, 2, …, K

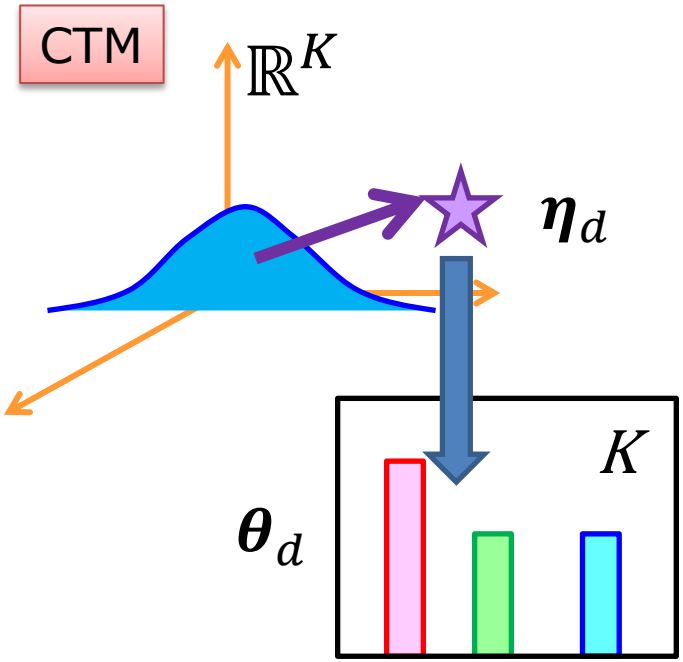
topic-word proportion $\boldsymbol{\beta}_k$

πはsoft-max関数

$$\pi(\boldsymbol{v})=\frac{\exp(v_k)}{\sum_l\exp(v_l)}$$

トピック間の相関: 多次元正規分布

- 共分散行列 Σ の効果でトピック分布に相関が生まれます
- 正規分布から生成される量はそのままは使えないので、Soft-maxで足して1にします



$$\eta_d|\mu,\Sigma \sim \mathrm{N}(\mu,\Sigma)$$

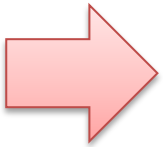
$$\theta_d|\eta_d = \pi(\eta_d)$$

πはsoft-max関数

$$\pi(v) = \frac{\exp(v_k)}{\sum_l \exp(v_l)}$$

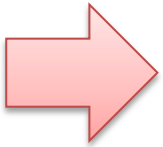
隠れ変数・パラメータの推定：難しくなります

- 原因1: Soft-max関数のため、共役性(conjugate)を利用できません ☹



(collapsed) Gibbs sampling**が非効率になるため、変分ベイズ法が候補になります**

- 原因2: 変分下限の評価も難しくなります



２段階での近似が必要になります

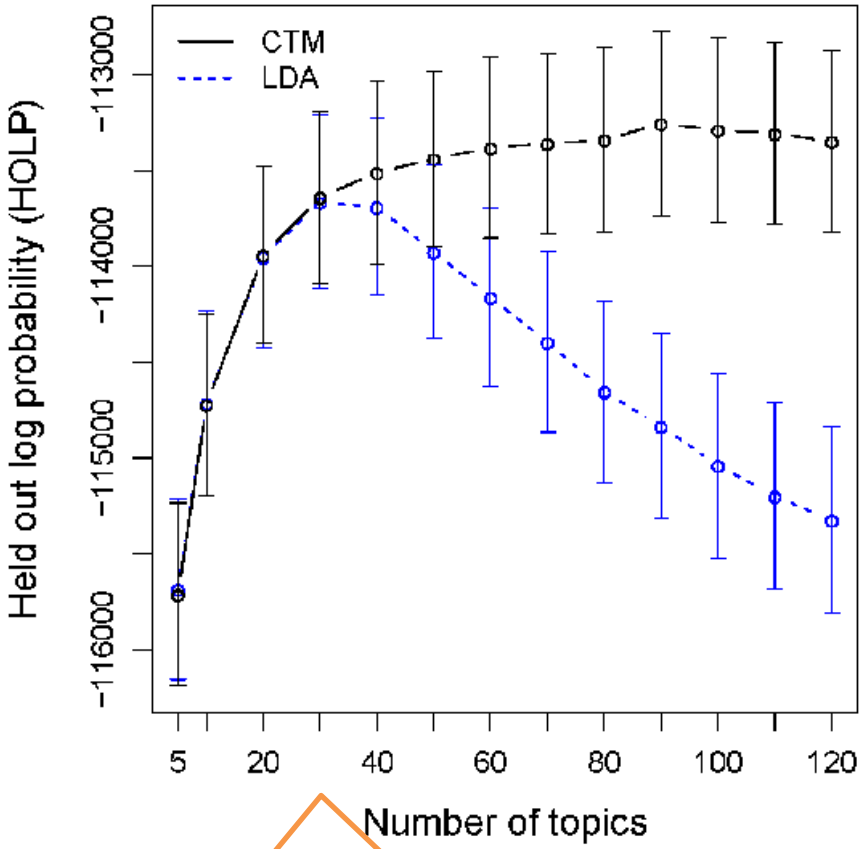
変分下限の評価

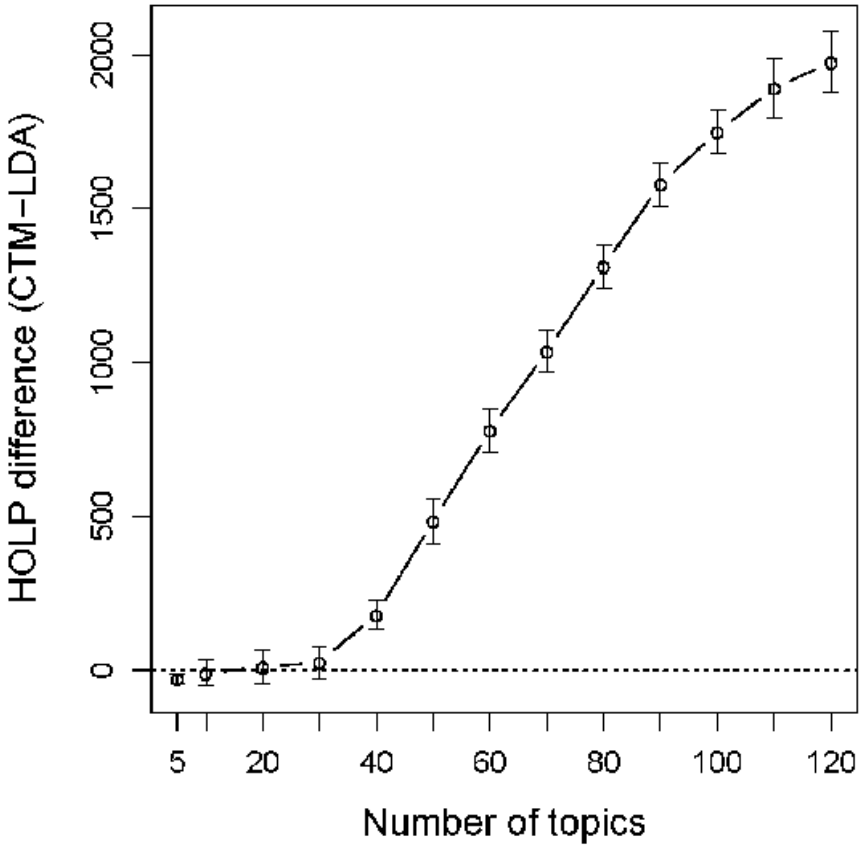
- 通常通り、Jensenによる下限を与えたあとに、さらに近似が必要です
- 詳しくは[Blei & Lafferty, 07]のAppendixをご覧ください

$$\log p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\beta}) \geq \mathrm{E}_q[\log p(\boldsymbol{\eta}|\boldsymbol{\mu},\boldsymbol{\Sigma})] + \sum_{d,n} \mathrm{E}_q\left[\log p\left(z_{d,n}|\boldsymbol{\eta}\right)\right]$$

$$+ \sum_{d,n} \mathrm{E}_q\left[\log p\left(x_{d,n}|z_{d,n},\boldsymbol{\beta}\right)\right] + H(q)$$

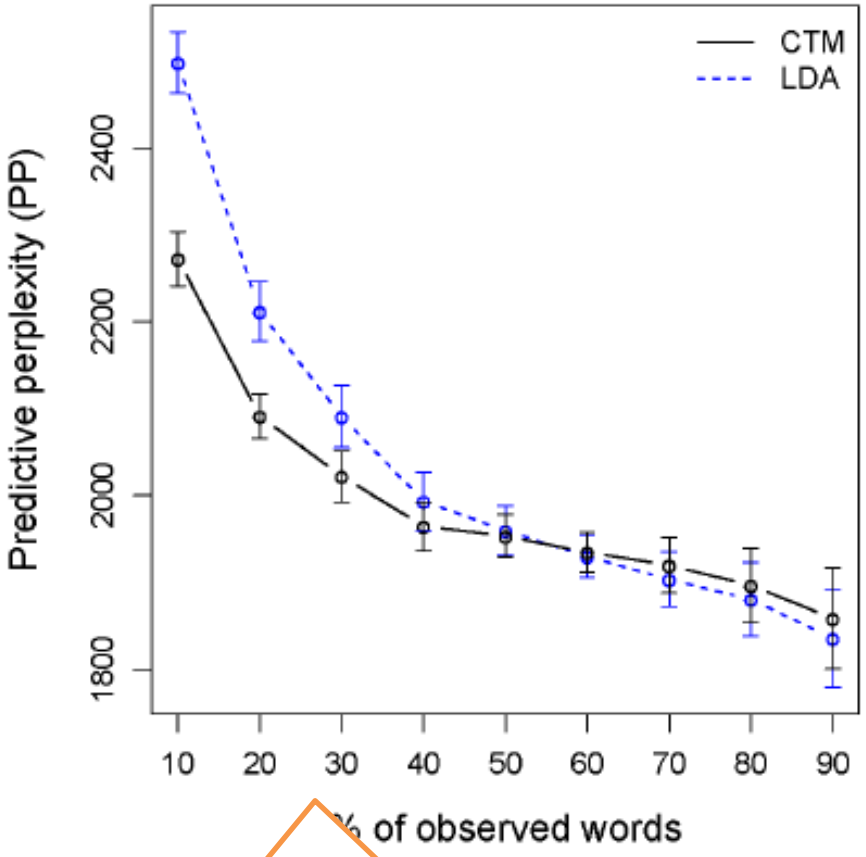
この評価でさらに近似が必要

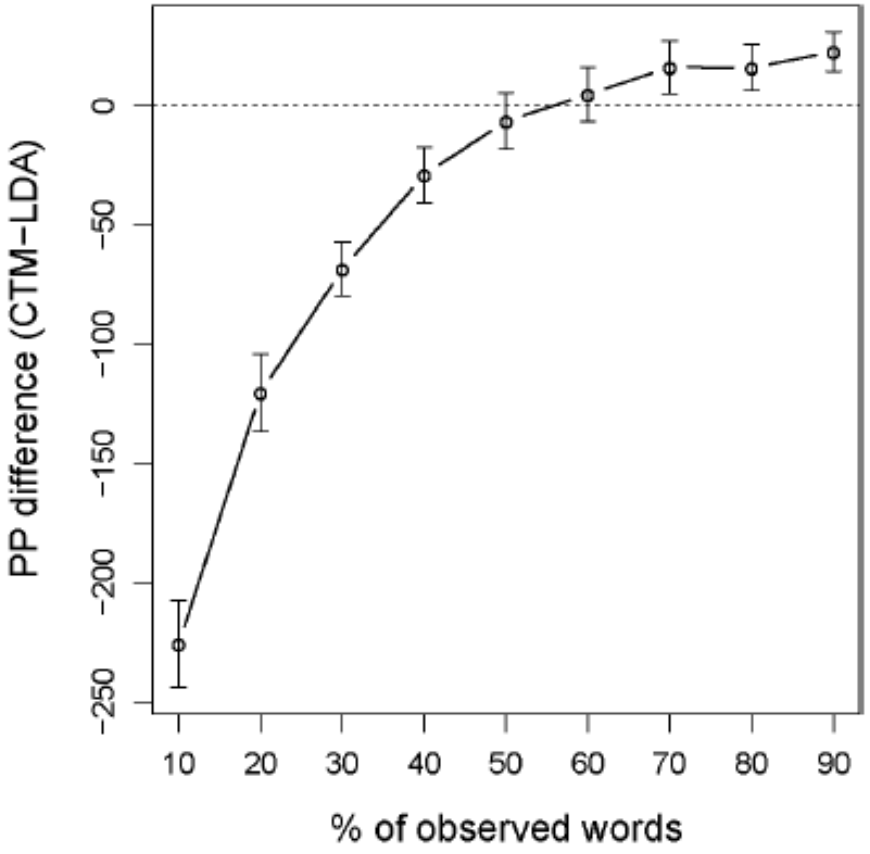




LDAは K＝30でピーク:
無理に独立なトピックを
仮定することの弊害が出てくる

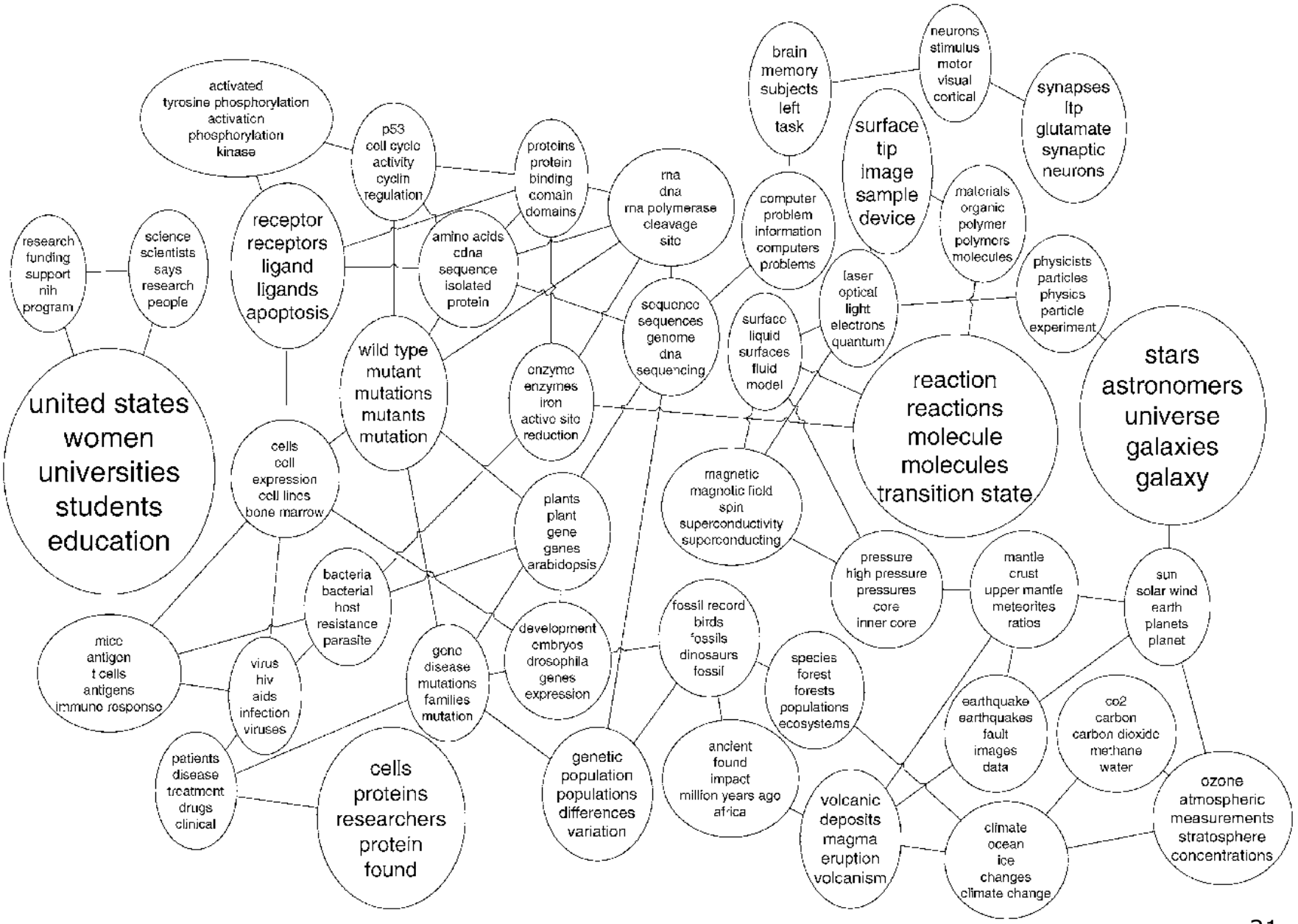
[Blei & Lafferty, 2007]





[Blei & Lafferty, 2007]

単語の観測量が少ないときにLDA よりも良い予測精度を記録



[Blei & Lafferty, 2007]

まとめ: Correlated Topic Models

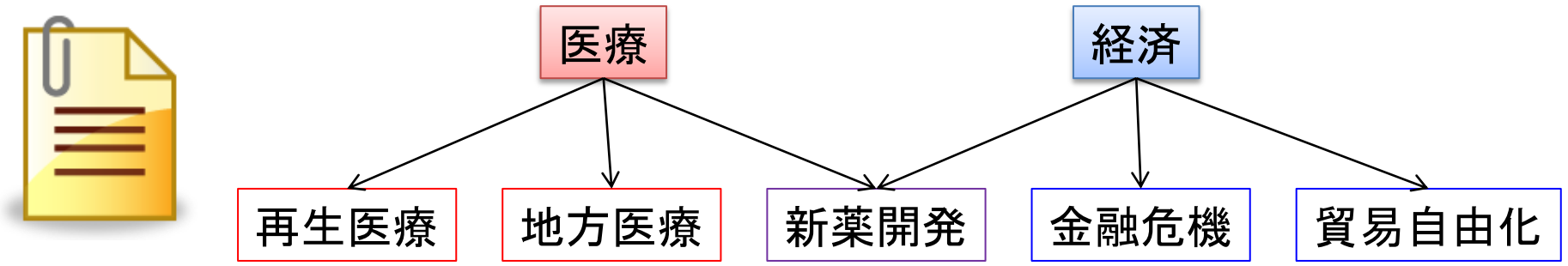
- トピック分布に、トピック間の相関を導入したモデルです
- 多次元正規分布でトピック間の関係を表現します
- 非常に有名で、後の各種トピックモデルに大きな影響を与えた仕事です。必須です。

PAM:
Pachinko Allocation Model
[Li & McCallum, 2006]

Li and McCallum,
"Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations",
in Proc. ICML, 2006.

CTMはトピックの間の階層構造が表せない

- 各トピックが同じレベルにあるからです
- トピックの階層構造（上下関係・包含関係）が適したデータの存在は容易に想像されます

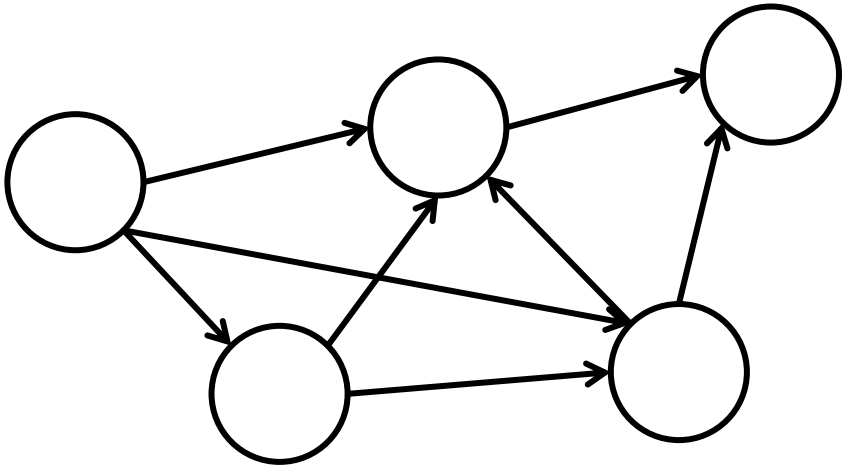


提案法: Pachinko Allocation Model (PAM)

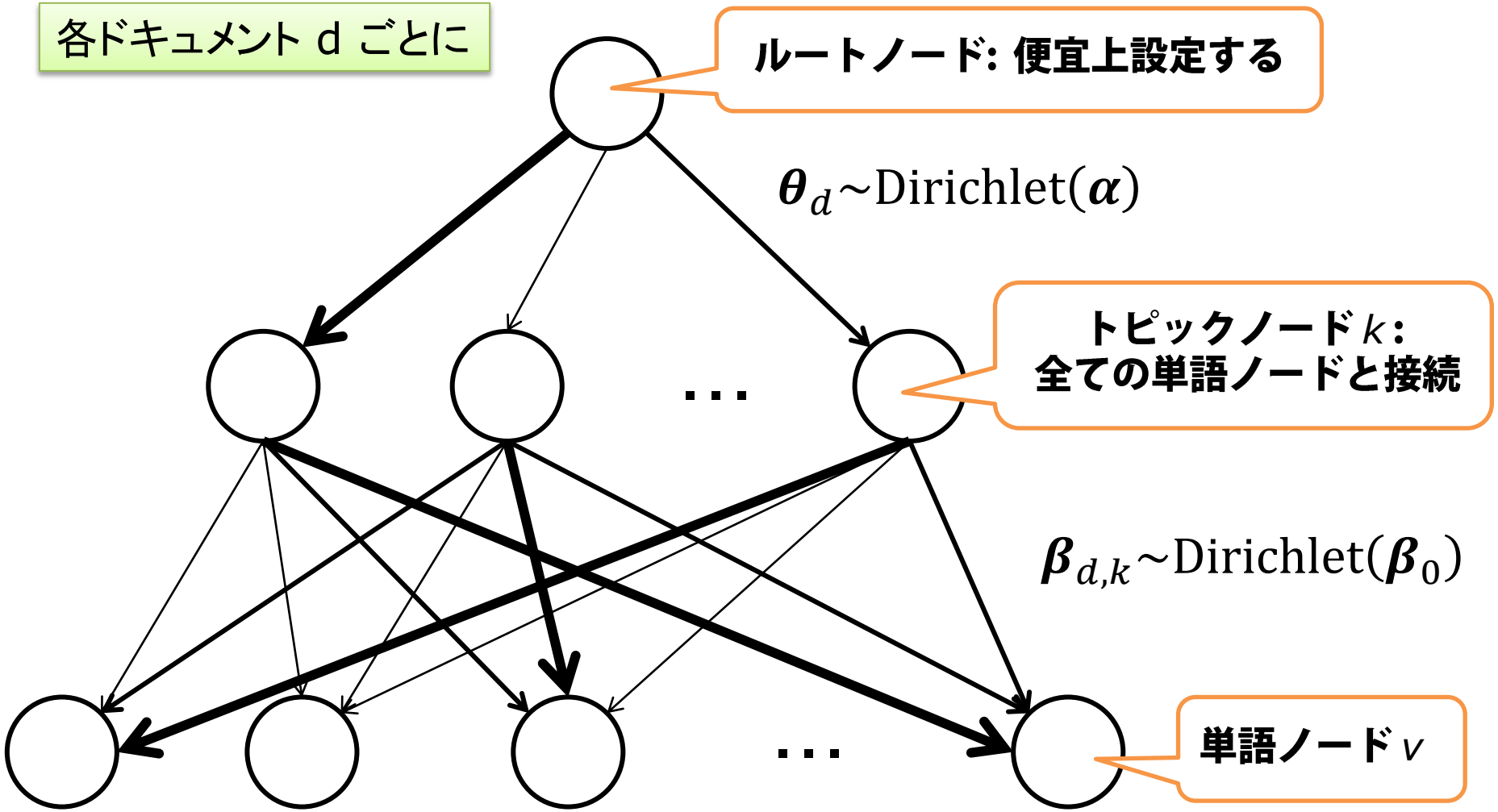
- トピック間の関係・相関を一般的に表現するモデルです
- トピック間の階層構造を基本として、パチンコ玉が落ちるように単語を生成します
- 複数トピックの共起なども表現できます

前提: 有向非巡回グラフ(DAG; Directed Acyclic Graph)

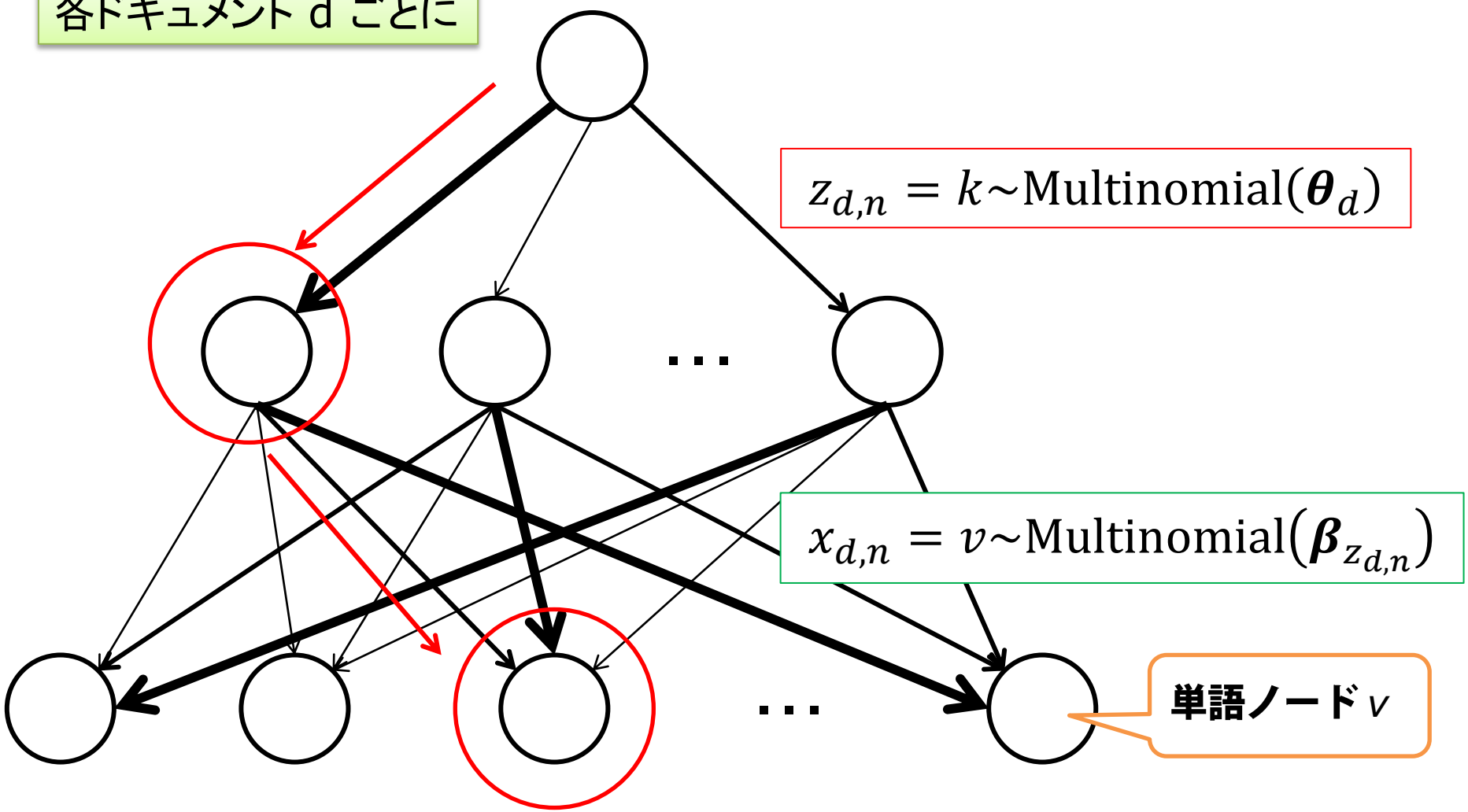
- 有向:ノード間のリンクは方向があります
- 非巡回:リンクをたどって、元のノードに戻ってくることがありません
- 木構造はDAGのさらに特殊な例です



DAGによるLDA解釈

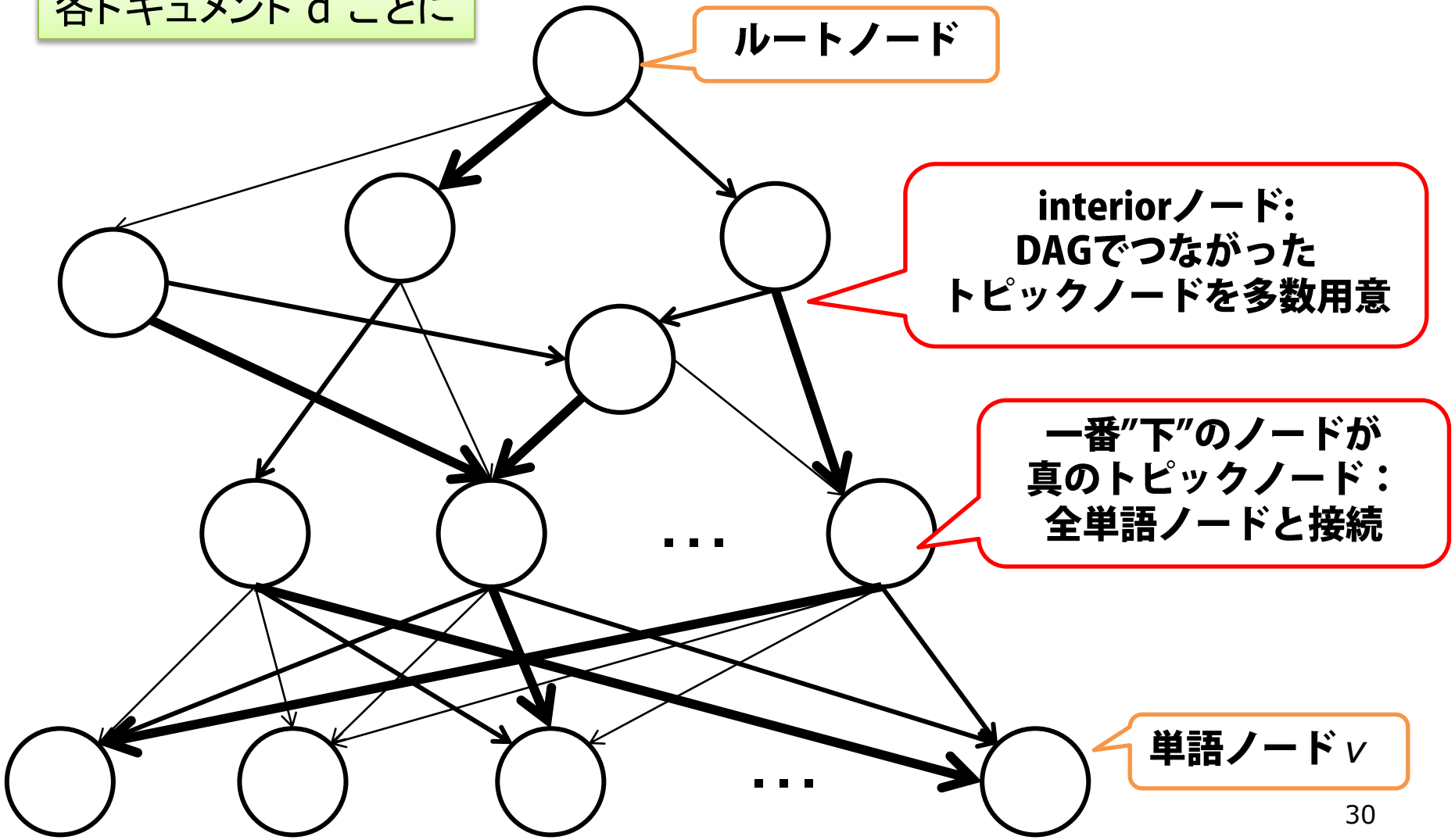


DAGによるLDA解釈



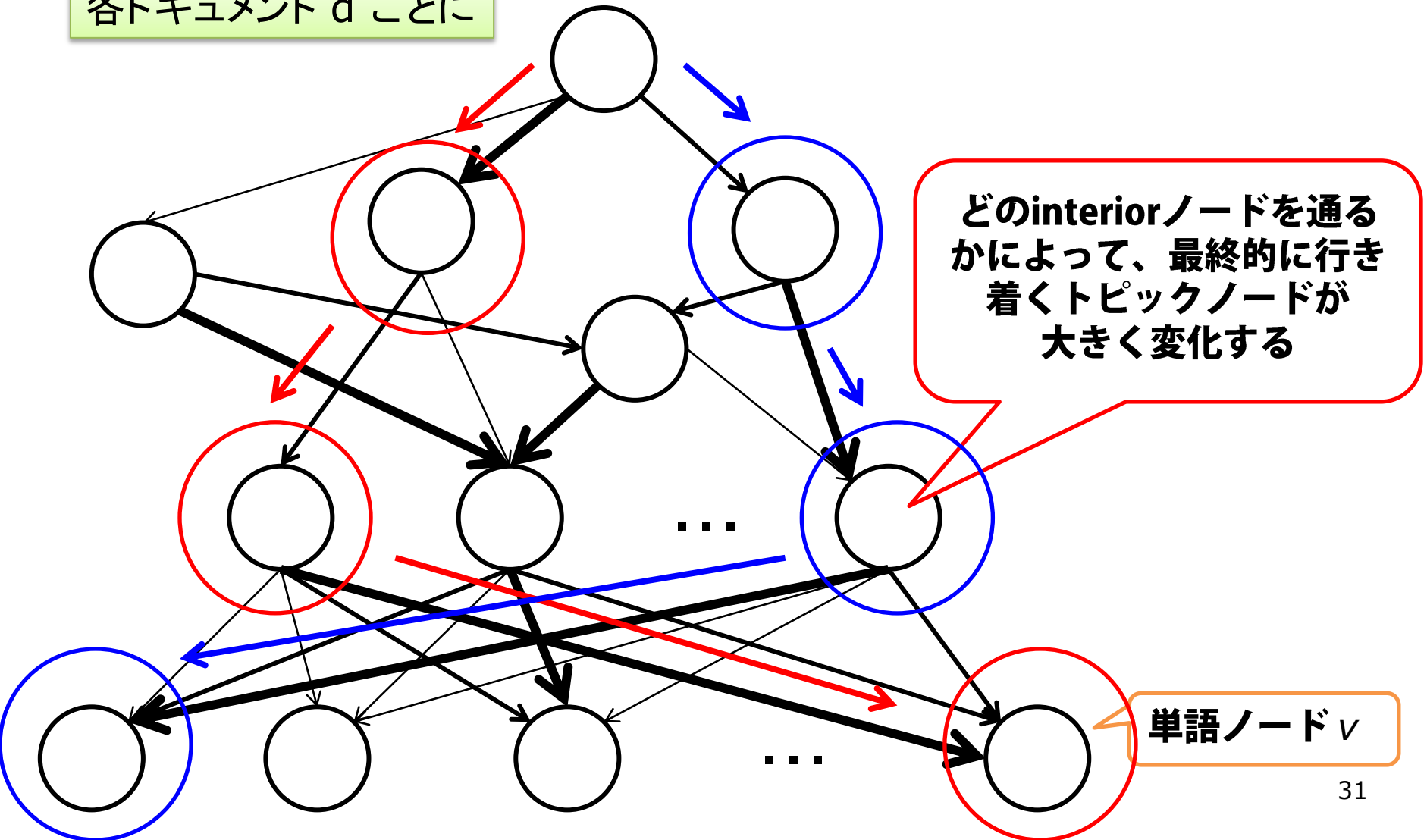
提案法のアイデア:
トピックノードをDAGで増やす

各ドキュメント d ごとに



“パチンコ玉”を落としてトピック、単語を生成します

各ドキュメント d ごとに

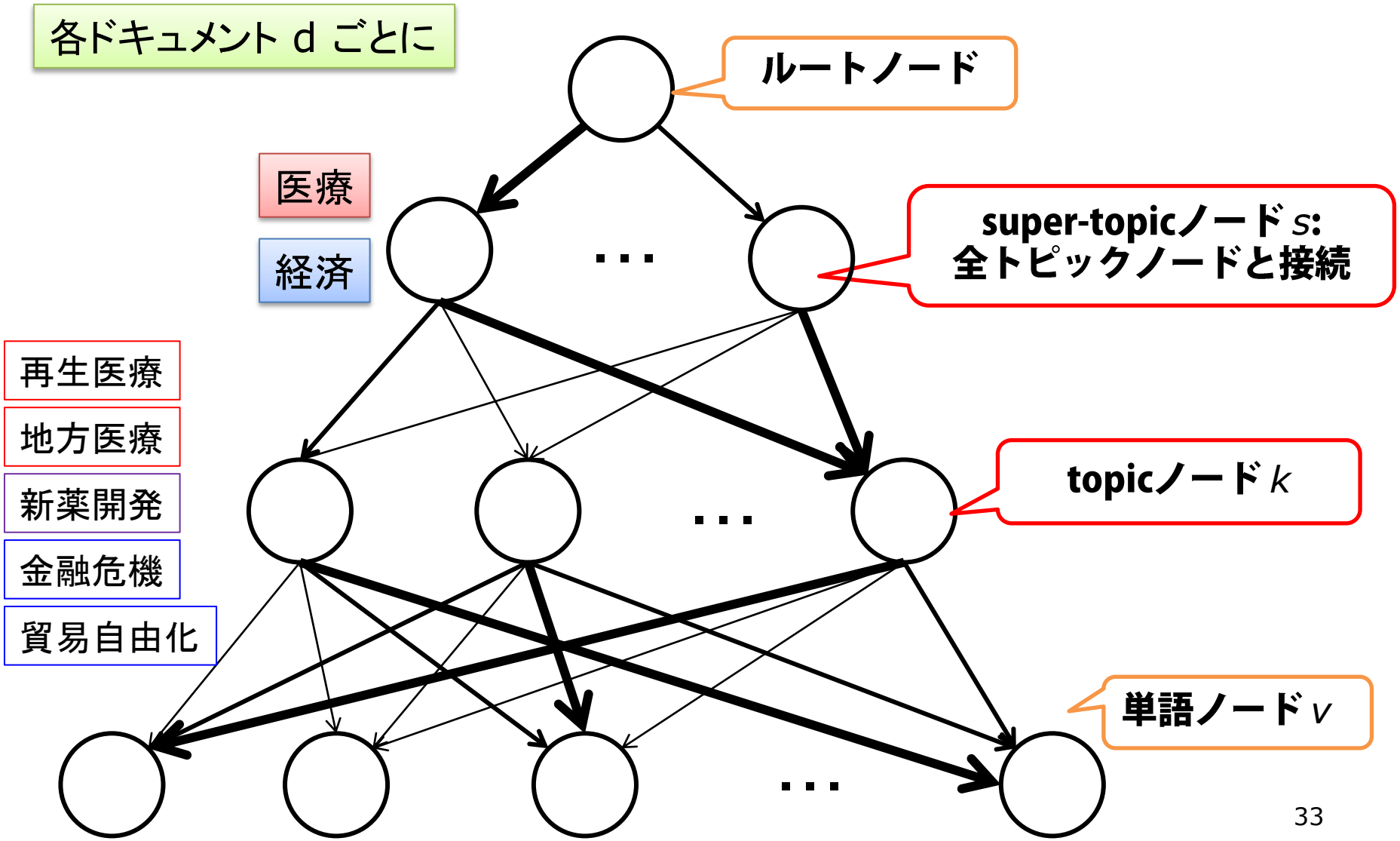


PAMの特長：表現力の高さと統一的な記述

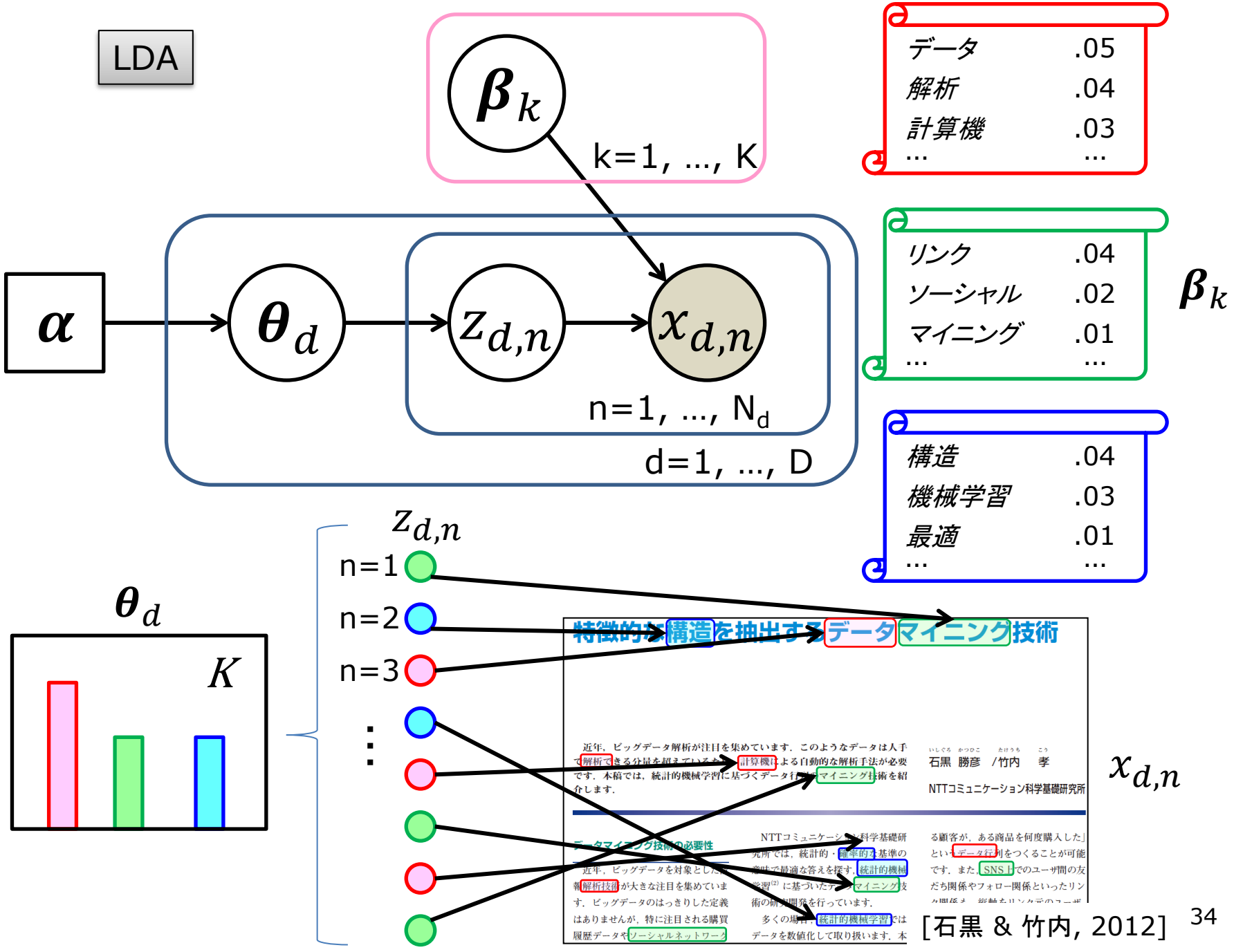
- 独立なトピックに対して、DAGで表現できるノード間の関係・構造をすべて持ち込めます
- interiorノード間の遷移確率は任意の確率分布でOK（Dir-Multが一番楽です）

- interiorノードをどのように入れても、完全に統一的な記述で表記可能です (これについては原論文を参照)

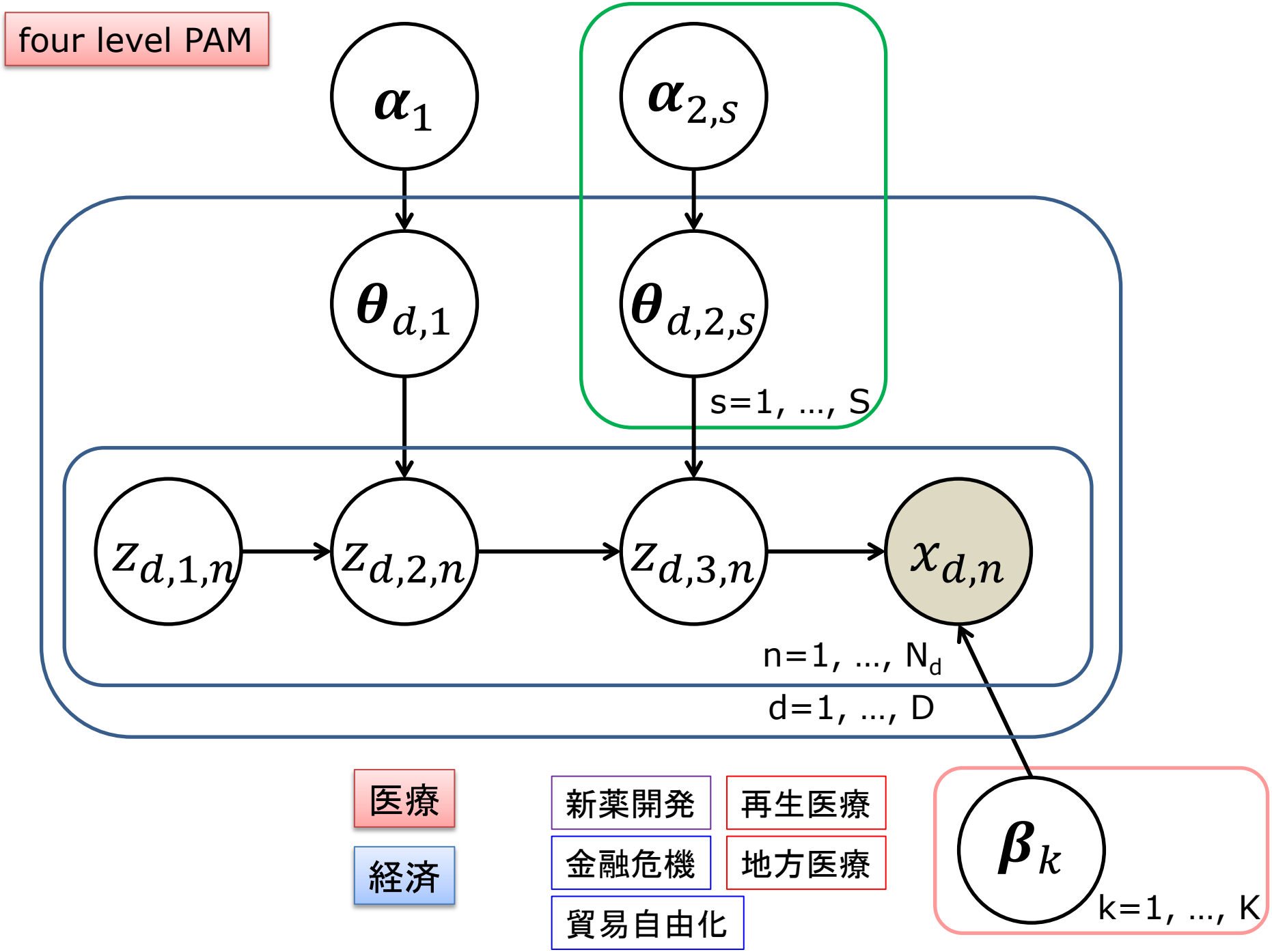
PAMの一形態: four-level PAM



LDA



[石黒 & 竹内, 2012]


four level PAM
α_1
$\alpha_{2,s}$
$\theta_{d,1}$
$\theta_{d,2,s}$
s=1, …, S
$z_{d,1,n}$
$z_{d,2,n}$
$z_{d,3,n}$
$x_{d,n}$
n=1, …, N_d
d=1, …, D
医療
経済
新薬開発
再生医療
金融危機
地方医療
貿易自由化
β_k
k=1, …, K

生成モデル

for 文書d = 1, 2, …, D_t

super-topic proportion $\boldsymbol{\theta}_{d,1}|\boldsymbol{\alpha}_1 \sim \mathrm{Dir}(\boldsymbol{\alpha}_1)$

for superトピック s = 1, 2, …, S

super-topic - topic proportion

$\boldsymbol{\theta}_{d,2,s}|\boldsymbol{\alpha}_{2,s} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{2,s})$

for 単語n = 1, 2, …, N_d

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k|\boldsymbol{\beta}_0 \sim \mathrm{Dir}(\boldsymbol{\beta}_0)$

for 文書d = 1, 2, …, D_t

super-topic proportion $\boldsymbol{\theta}_{d,1}$

for superトピック s = 1, 2, …, S

super-topic - topic proportion $\boldsymbol{\theta}_{d,2,s}$

for 単語n = 1, 2, …, N_d

root node (const.) $\boldsymbol{z}_{d,1,n}$

super-topic - word assignment

$$z_{d,2,n}|\boldsymbol{\theta}_{d,1}\sim\mathrm{Mult}(\boldsymbol{\theta}_{d,1})$$

topic-word assignment

$$z_{d,3,n}|z_{d,2,n},\boldsymbol{\theta}_{d,2,s}\sim\mathrm{Mult}(\boldsymbol{\theta}_{d,2,z_{d,2,n}})$$

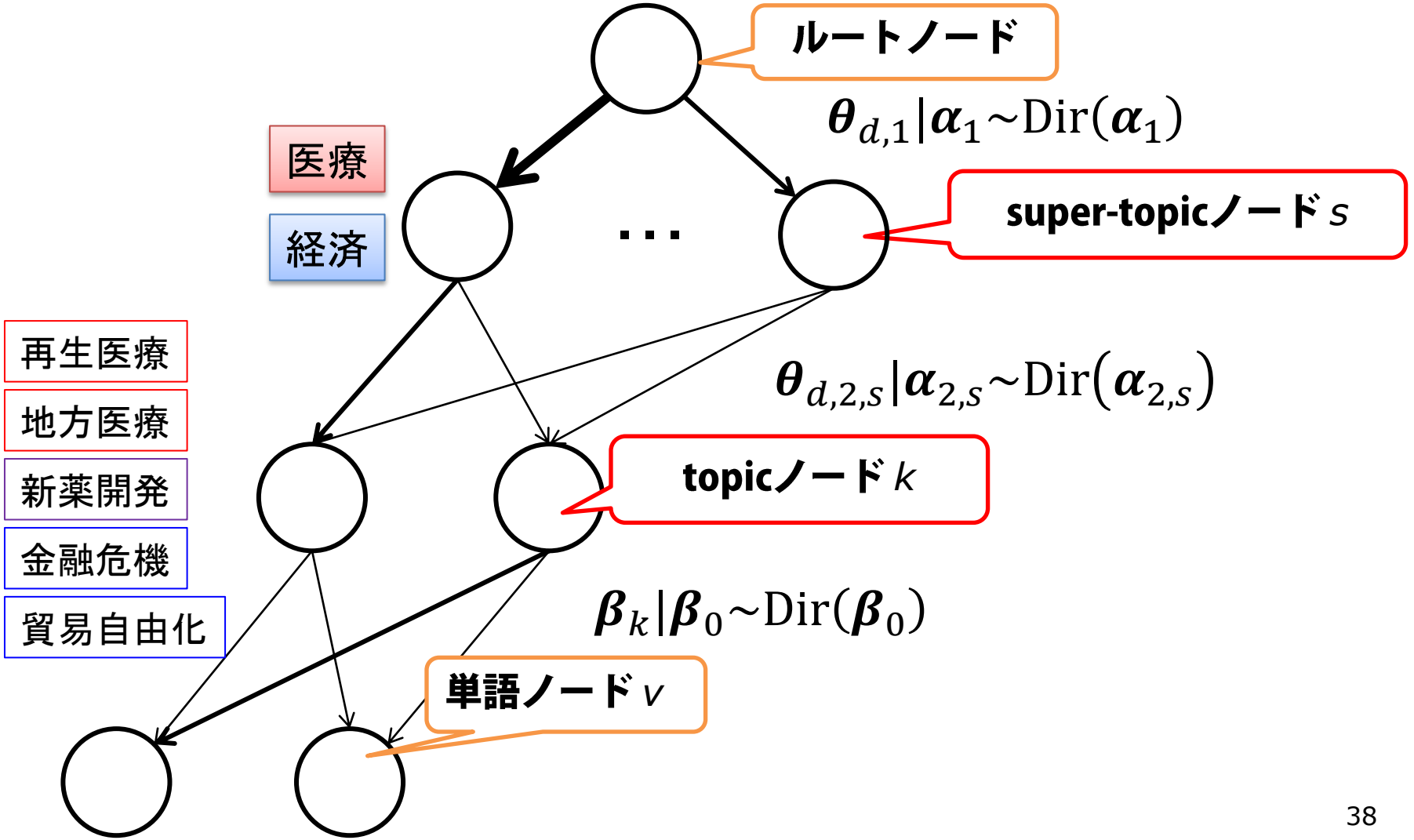
word observation

$$x_{d,n}|z_{d,3,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,3,n}})$$

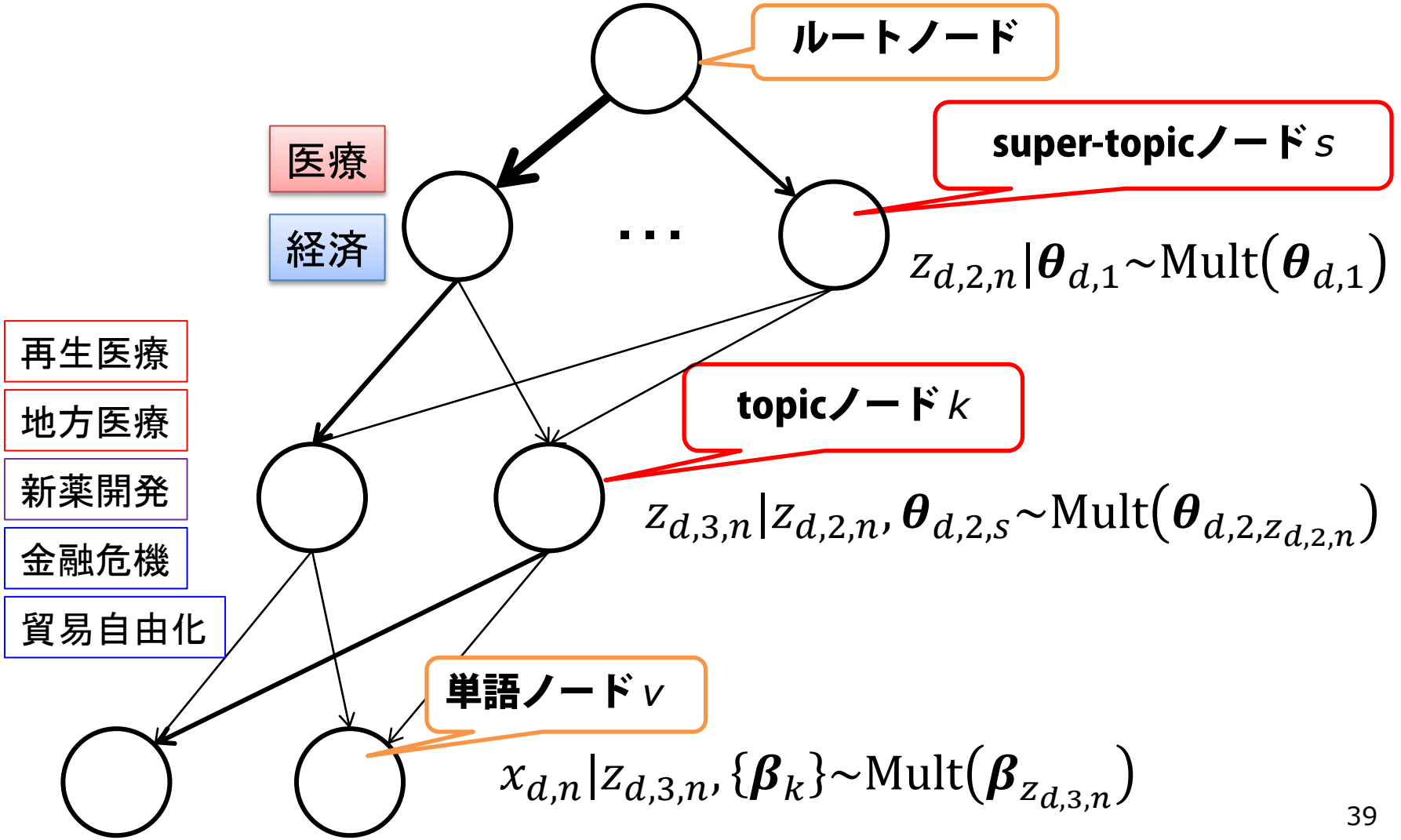
for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k$

Dirichlet-Multinomialで統一されているので簡単です



Dirichlet-Multinomialで統一されているので簡単です



Dirichlet-Multinomialで統一されているので簡単です

- つまり、この調子で何段でも階層を重ねていくことができます

- 単純ながら、有効なモデル化戦略といえるでしょう

隠れ変数・パラメータの推定

- パラメータは積分消去可能
- 隠れ変数はGibbs samplingで最適化
  - モデルの階層性から、変分ベイズでは局所解が多すぎるようです
- ハイパーパラメータ α はモーメントマッチングで最適化します

- ☺ モデルの複雑さに反して、式の導出・計算結果は非常にシンプルです

隠れ変数のGibbsサンプリング

- 通常のLDA(Dirichlet-Multinomial)と同じ構造を持ちます

$$p\left(z_{d,2,n}=s, z_{d,3,n}=k \middle| x_{d,n}=w, \boldsymbol{X}_{\neg(d,n)}, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta_0}\right) \propto$$

$$\frac{m_{ds}+\alpha_{1,s}}{\sum_{s'}\left(m_{ds'}+\alpha_{1,s'}\right)} \frac{m_{dsk}+\alpha_{2,s,k}}{\sum_{k'}\left(m_{dsk'}+\alpha_{2,s,k'}\right)} \frac{m_{kw}+\beta_{0,w}}{\sum_{w'}\left(m_{kw'}+\beta_{0,w'}\right)}$$

文書 dから superトピック sが 生成される確率

文書 dから superトピック sを経由して トピック kが 生成される確率

トピック kから 単語 wが 生成される確率

ハイパーパラメータ α の最適化

- 普通は周辺化尤度最大化やhyper priorを仮定して推定します
- 論文ではモーメントマッチを利用しています
  - 申し訳ありませんが、講師には導出の根拠がわかりませんでした ☹

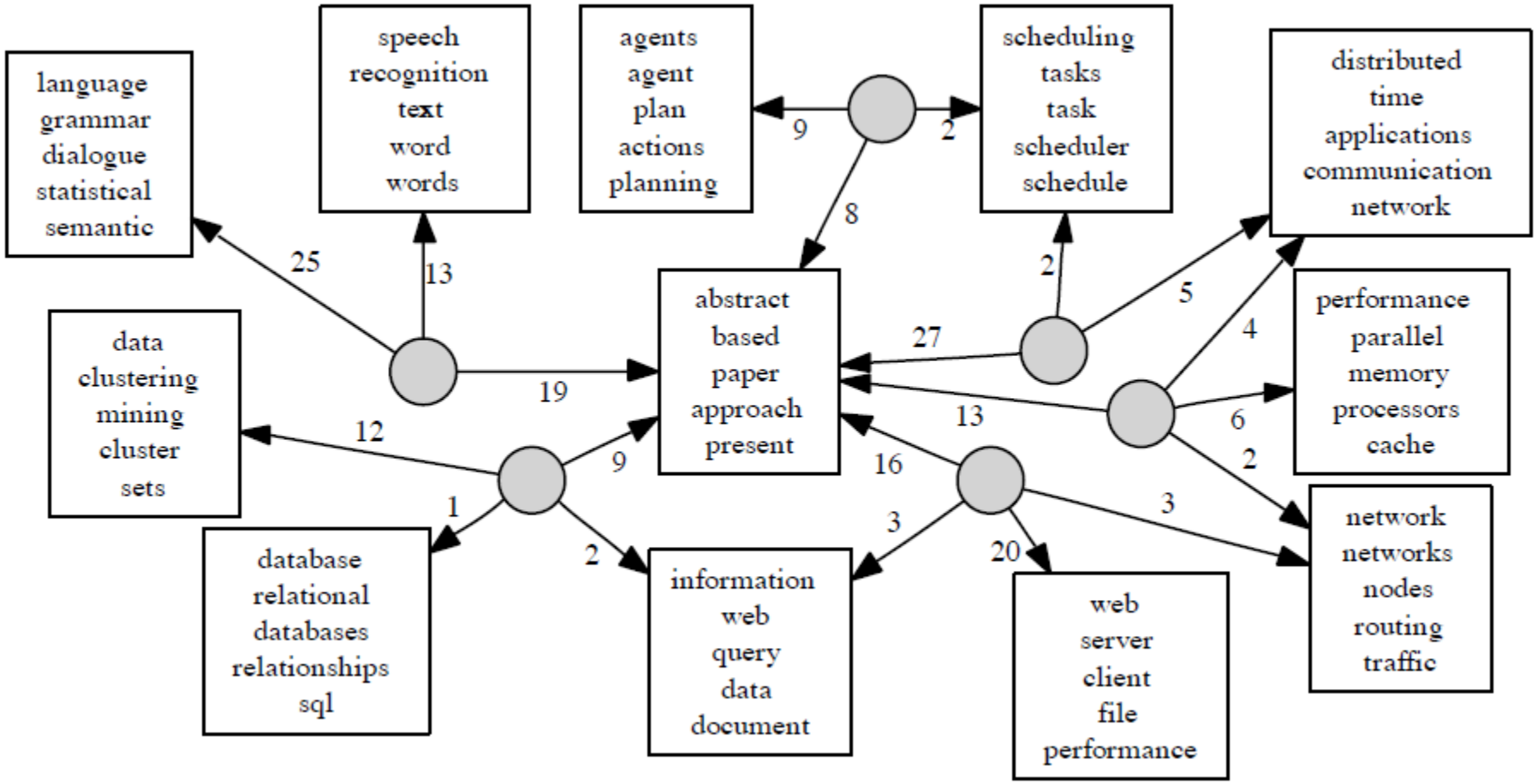
$$mean_{xy} = \frac{1}{N} \times \sum_d \frac{n_{xy}^{(d)}}{n_x^{(d)}};$$

$$var_{xy} = \frac{1}{N} \times \sum_d (\frac{n_{xy}^{(d)}}{n_x^{(d)}} - mean_{xy})^2;$$

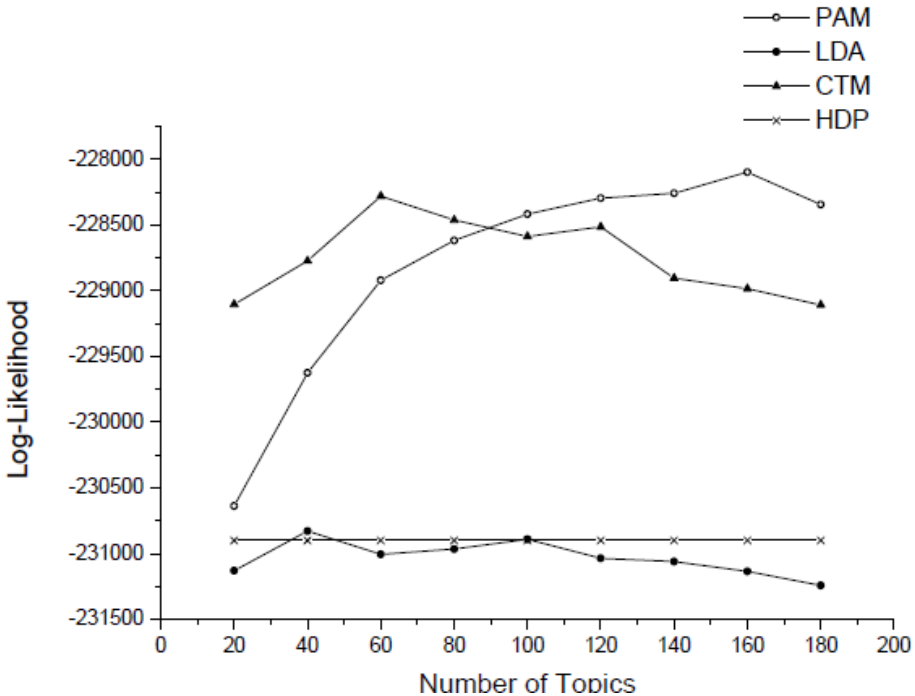
$$m_{xy} = \frac{mean_{xy} \times (1 - mean_{xy})}{var_{xy}} - 1;$$

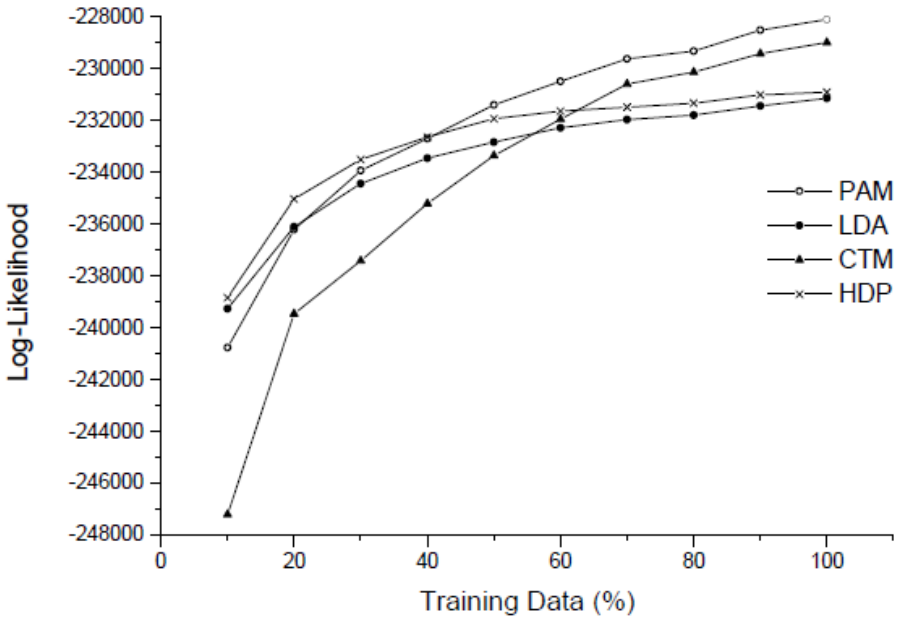
$$\alpha_{xy} \propto mean_{xy};$$

$$\sum_y \alpha_{xy} = \frac{1}{5} \times exp(\frac{\sum_y log(m_{xy})}{s_2 - 1}).$$



[Li and McCaullum, 2006]





class	# docs	LDA	PAM
graphics	243	83.95	86.83
os	239	81.59	84.10
pc	245	83.67	88.16
mac	239	86.61	89.54
windows.x	243	88.07	92.20
total	1209	84.70	87.34

Table 3. Document classification accuracy (%)

[Li and McCallum, 2006]

まとめ: Pachinko Allocation Model

- トピック間の関係構造をモデル化する、非常に柔軟かつ一般的なモデルです
- DAGによって、トピック間を任意の順番で連結してtopic pathを作ります
- 複雑なモデルですが、Gibbs samplingでシンプルかつ直観的な推論過程を実現

原論文を読むときの注意

- 論文では、一般のPAMについて生成モデルを説明しています

- また、interior nodeはすべて
  - θ ~ Dir の確率のもとで
  - z ~ Multi をサンプリングする

  として表記が統一されています

Multi-grained Topic model for aspect rating

[Titov & McDonald, 2008]

Titov and McDonald,
"Modeling Online Reviews with Multi-grained Topic Models",
in Proc. WWW, 2008.

レビュー記事のトピックモデリング

- レビュー記事はトピックモデル解析の典型的な対象です

amazon.co.jp

25 人中、23人の方が、[redacted]ています。

★★★★★ オススメです。 2012/4/10

By [redacted]

製品の品質、OSのユーザビリティともに大変良くできています。
スペックも一番安いので全く問題ありません。
配線が電源コード1本だけになるのも素晴らしい。
仕事で使ったりゲームをするのでなければWINDOWSより断然オススメできます。
ただ付属のMagic MouseはAdobeのPhotoshopやIllustratorのソフトなどをするのであればトラックパッド付属を選び、マウスは別途購入をお勧めします。

1コメント | このレビューは参考になりましたか？ はい いいえ

41 人中、36人の方が、「このレビューが参考になった」と投票しています。

★★★★★ Best of Mac 2011/8/22

By [redacted] トップ1000レビュアー VINE™ メンバー

Amazon.co.jpで購入済み

iMacはRev.Aから10台近く使っています。
アルミiMacは2台目、非常にきれいなモニター、
使いやすいワイヤレスキーボード、マジックマウスに満足しています。

マジックマウスの電池がやたら減る事も大したことじゃないし、
初心者にもハードユーザーにもお勧め。

Hotels.com

"立地は最高です"

2012/12/07 [redacted]

総合的な評価

空港から列車に乗って、駅から降りたら目の前にホテルがあるっていうのはうれしいです。朝食はビュッフェ形式でした。種類はまあまあといったところ。ただ、ヨーロッパ旅行をしていると飽きるかもです。せっかく世界遺産の広場『グランプラス』が徒歩5分くらいのところにあるので、世界遺産の広場を見ながらカフェで朝食っていうのがいいかも。朝7時くらいだとほんとに誰もいないし、広場を独占できる贅沢を味わえます。それも30分くらいのもので、7時30分には団体がきて雰囲気が台無しに。せっかく立地がいいホテルなので、早朝の広場を体感してみてください。

"駅前は最高です"

2012/10/13 [redacted]

総合的な評価

ブリュッセル中央駅出口目の前。グランプラスまで徒歩5分とかからない最高の立地です。中央駅を利用すればアントワープ、ゲント、ブルージュetc・・まで乗り換えなしで移動できるのが嬉しいです。ホテル近辺にコンビニがありませんが、駅構内に、コンビニ、カルフール、スタバなどがあり、とっても便利ですよ。ホテルは、SPGプラチナでの滞在でしたが、広いお部屋はUPGされました。基本的アメニティ(歯ブラシなし)にくわえ、バスローブ、使い捨て白いスリッパ、お水2本、ネスプレッソ4杯分、それとダンドアのスペキュロスクッキーをいただけました。 水周りは、ハンドシャワー有、ウォシュレットなし。ラウンジはないので、朝食はとりませんでしが、外へ行けば美味しいワッフル屋さんがあるので問題なしでした！！ それと、至る所に設置のある「レンタサイクル」ですが、支払いがクレカのみなのですが、日本のクレカでは使用出来ませんでした。(観光局の方に確認済み) 石畳は非常に歩きにくいため、ゴム底の楽な靴でないと、大変なことになります。総合的に、古いですが、とてもすごしやすいホテルでした。 日本語は通じませんが。。

Aspect rating

総合評価

aspects

立地

食事

観光地へのアクセス

アメニティ

Hotels.com

"立地は最高です"

2012/12/07 [redacted]

総合的な評価

空港から列車に乗って、駅から降りたら目の前にホテルがあるっていうのはうれしいです。朝食はビュッフェ形式でした。種類はまあまあといったところ。ただ、ヨーロッパ旅行をしていると飽きるかもです。せっかく「世界遺産の広場"グランプラス"」が徒歩5分くらいのところにあるので、世界遺産の広場を見ながらカフェで朝食っていうのがいいかも。朝7時くらいだとほんとに誰もいないし、広場を独占できる贅沢を味わえます。それも30分くらいのもので、7時30分には団体がきて雰囲気が台無しに。せっかく立地がいいホテルなので、早朝の広場を体感してみてください。

"駅前は最高です"

2012/10/13 [redacted]

総合的な評価

ブリュッセル中央駅出口目の前。グランプラスまで徒歩5分とかからない最高の立地です。中央駅を利用すればアントワープ、ゲント、ブルージュetc・・まで乗り換えなしで移動できるのが嬉しいです。ホテル近辺にコンビニがありませんが、駅構内に、コンビニ、カルフール、スタバなどがあり、とっても便利ですよ。ホテルは、SPGプラチナでの滞在でしたが、広いお部屋はUPGされました。基本的アメニティ(歯ブラシなし)にくわえ、バスローブ、使い捨て白いスリッパ、お水2本、ネスプレッソ4杯分、それとダンドアのスペキュロスクッキーをいただけました。 水周りは、ハンドシャワー有、ウォシュレットなし。ラウンジはないので、朝食はとりませんでしが、外へ行けば美味しいワッフル屋さんがあるので問題なしでした！！それと、至る所に設置のある「レンタサイクル」ですが、支払いがクレカのみなのですが、日本のクレカでは使用出来ませんでした。(観光局の方に確認済み) 石畳は非常に歩きにくいため、ゴム底の楽な靴でないと、大変なことになります。総合的に、古いですが、とてもすごしやすいホテルでした。 日本語は通じませんが。。

統計モデルによるAspect分析

- 統計モデルによる客観的・自動的なaspect分析が可能となると、より詳細なサービス評価・レビュー解析が可能になります

総合評価：4

ratable aspects

- 値段：A
- 性能：A
- アフターサービス：C
- 使いやすさ：B

amazon.co.jp

25 人中、23人の方 ています。

☆☆☆☆☆ オススメです。 2012/1/10

By [redacted]

製品の品質、OSのユーザビリティともに大変良くできています。
スペックも一番安いので全く問題ありません。
配線が電源コード1本だけになるのも素晴らしい。
仕事で使ったりゲームをするのでなければWINDOWSより断然オススメできます。
ただ付属のMagic MouseはAdobeのPhotoshopやIllustratorのソフトなどをお使いの場合、
トラックパッド付属を選び、マウスは別途購入をお勧めします。

1コメント | このレビューは参考になりましたか？ はい いいえ

41 人中、36人の方が、「このレビューが参考になった」と投票しています。

☆☆☆☆☆ Best of Mac 2011/8/22

By [redacted] トップ1000レビュアー VINE™ メンバー

Amazon.co.jpで購入済み

iMacはRev.Aから10台近く使っています。
アルミiMacは2台目、非常にきれいなモニター、
使いやすいワイヤレスキーボード、マジックマウスに満足しています。

マジックマウスの電池がやたら減る事も大したことじゃないし、
初心者にもハードユーザーにもお勧め。

提案法: Multi-grained Topic Models

- 主にレビュー記事を対象に、文章を全体的なトピックとaspectに関するトピックに分解するモデルです
- トピックの相関ではなく、トピックの種類を増やした構造をもつモデルです
- 推論はモデルの複雑さに反してシンプルです

提案法のアイデア:
2種類のトピックを仮定する

グローバルトピック: レビュー対象の全体的な特徴

ホテル記事ならば・・・”ロンドンのホテル” “ビジネスホテル”など

ロンドン	.05
地下鉄	.04
五輪	.03
…	…

機能的	.04
お手頃	.02
ネット接続	.01
…	…

ローカルトピック = ratable aspect

ホテル記事ならば・・・”アクセス” “値段” “食事” “立地” “部屋の広さ”など

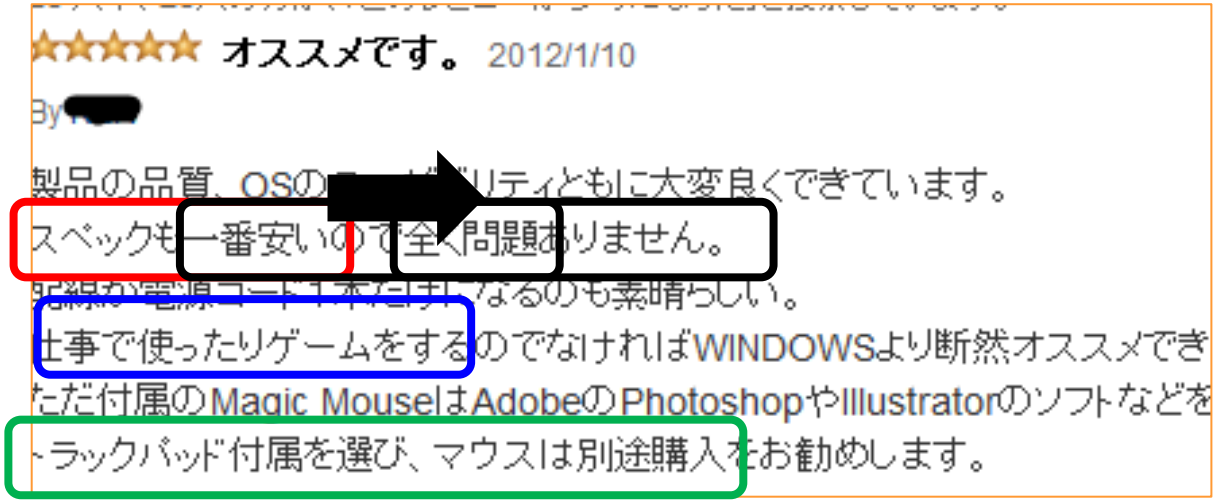
駅前	.04
徒歩圏内	.03
バス	.01
…	…

リーズナブル	.03
見合った	.03
納得できる	.02
…	…

おいしい	.07
ルームサービス	.03
味が濃い	.02
…	…

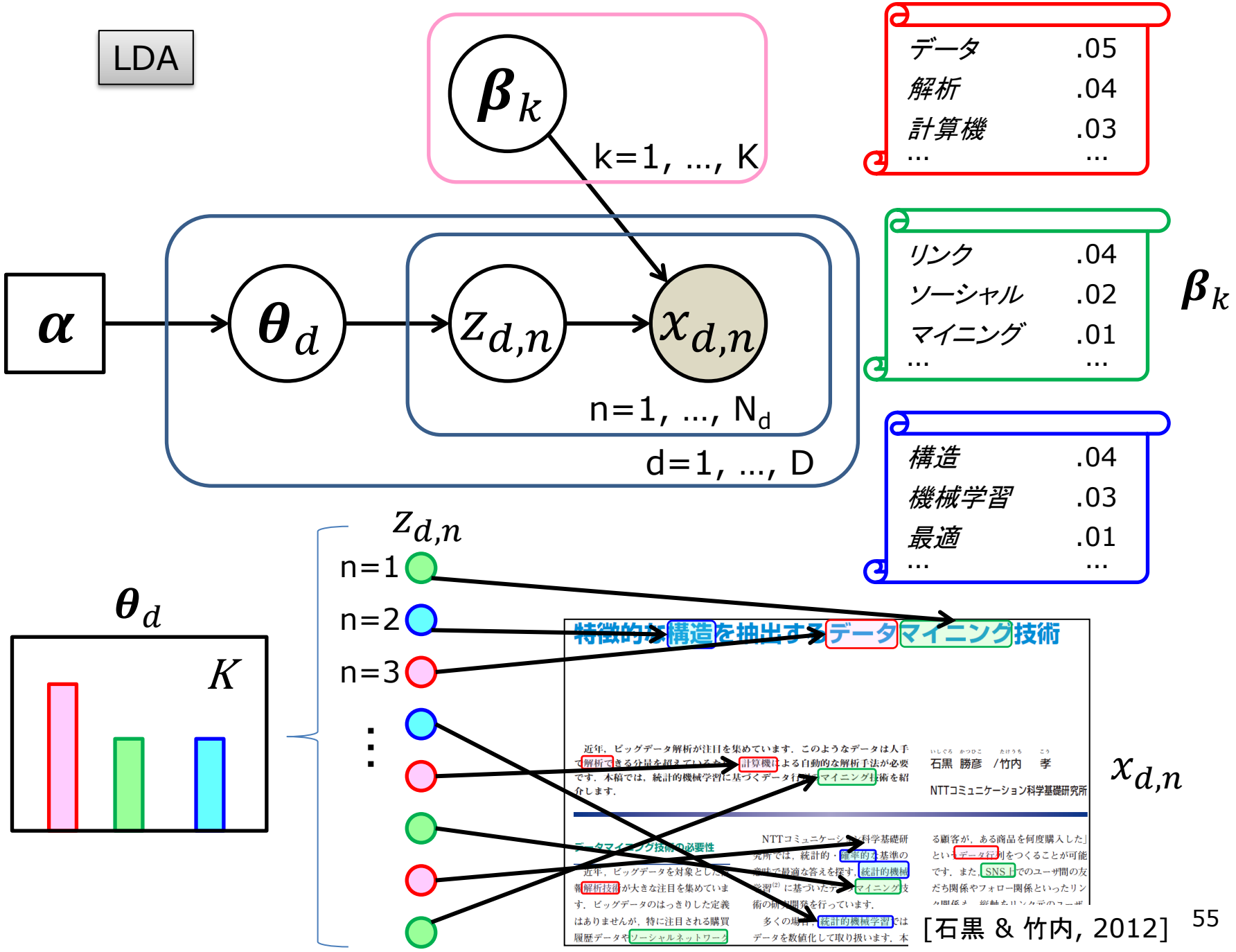
提案法のアイデア: Sliding window

- local topicは文章の一部でしか出てこないので、sliding windowで局所的にモデル化



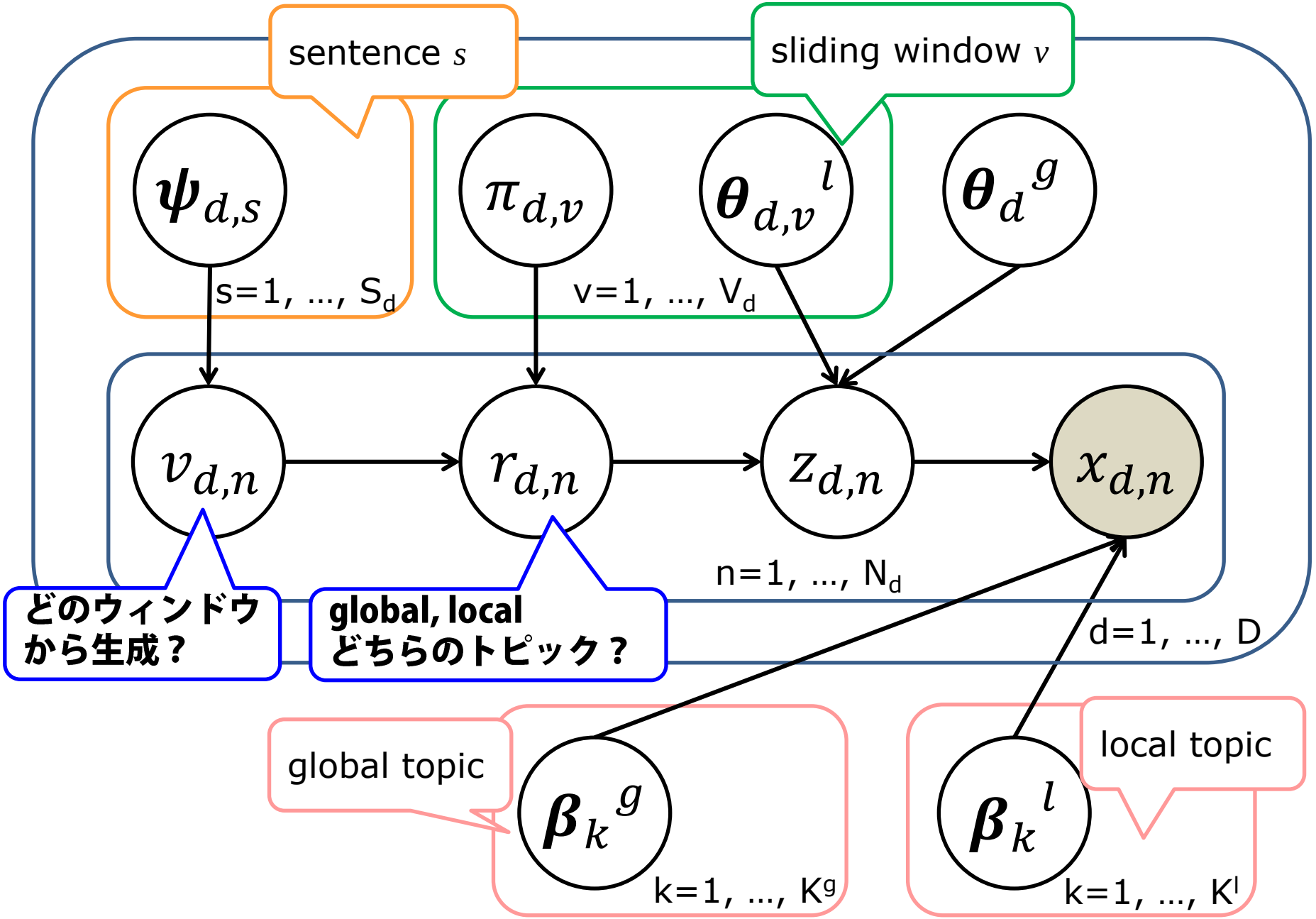
local topic 1 (値段) の割合	**小**	**大**	**中**
local topic 2 (用途)の割合	**大**	**中**	**小**
local topic 3 (付属品)の割合	**小**	**小**	**大**

LDA



[石黒 & 竹内, 2012]

Multi-grained Topic Models (ハイパーパラメータを省略)



for 文書d = 1, 2, …, D_t

for sentence s = 1, 2, …, S_d

window proportion $\boldsymbol{\psi}_{d,s}|\boldsymbol{\gamma}\sim\mathrm{Dir}(\boldsymbol{\gamma})$

for sliding window v = 1, 2, …, V_d

global-local proportion $\pi_{d,v}|\boldsymbol{\alpha}^{mix}\sim\mathrm{Beta}(\boldsymbol{\alpha}^{mix})$

local-topic proportion $\boldsymbol{\theta}_{d,v}{}^{l}|\boldsymbol{\alpha}^{l}\sim\mathrm{Dir}(\boldsymbol{\alpha}^{l})$

global-topic proportion $\boldsymbol{\theta}_{d}{}^{g}|\boldsymbol{\alpha}^{g}\sim\mathrm{Dir}(\boldsymbol{\alpha}^{g})$

for Localトピック k = 1, 2, …, K^l

local topic-word proportion $\boldsymbol{\beta}_{k}{}^{l}|\boldsymbol{\beta}_{0}{}^{l}\sim\mathrm{Dir}(\boldsymbol{\beta}_{0}{}^{l})$

for Globalトピック k = 1, 2, …, K^g

global topic-word proportion $\boldsymbol{\beta}_{k}{}^{g}|\boldsymbol{\beta}_{0}{}^{g}\sim\mathrm{Dir}(\boldsymbol{\beta}_{0}{}^{g})$

for 文書d = 1, 2, …, D_t

for sentence s = 1, 2, …, S_d window proportion $\boldsymbol{\psi}_{d,s}$

global-topic proportion $\boldsymbol{\theta}_d{}^{g}$

for sliding window v = 1, 2, …, V_d global-local proportion $\boldsymbol{\pi}_{d,v}$

local-topic proportion $\boldsymbol{\theta}_{d,v}{}^{l}$

for 単語n = 1, 2, …, N_d in sentence s

window-word assignment

$$v_{d,n}|\boldsymbol{\psi}_{d,s} \sim \mathrm{Mult}(\boldsymbol{\psi}_{d,s})$$

global/local-word assignment

$$r_{d,n}|\pi_{d,v}, v_{d,n} \sim \mathrm{Bernoulli}(\pi_{d,v_{d,n}})$$

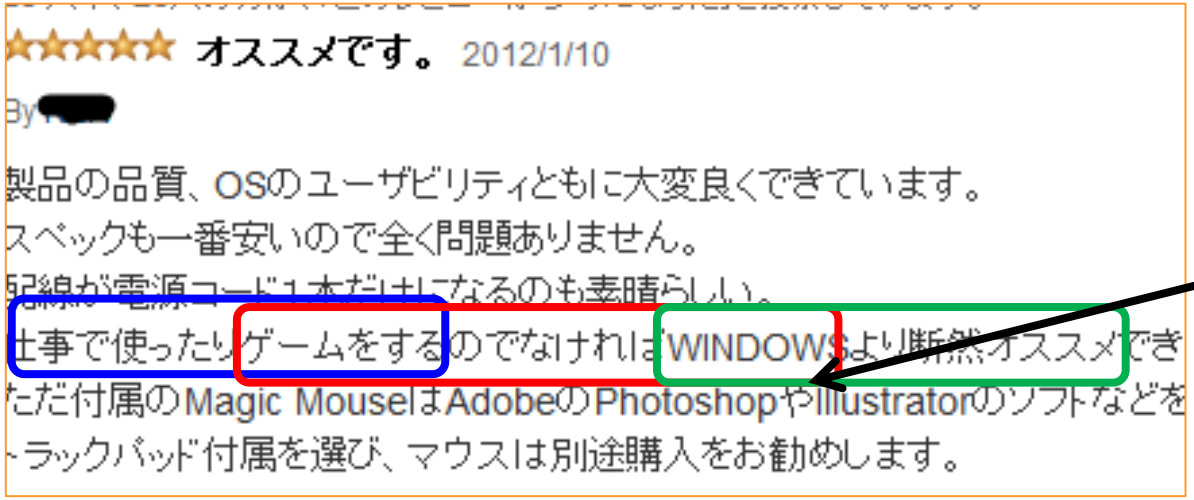
topic-word assignment

$$z_{d,n}|r_{d,n}, \boldsymbol{\theta}_d \sim \mathrm{Mult}(\boldsymbol{\theta}_{d,(v)}{}^{r_{d,n}})$$

word observation

$$x_{d,n}|z_{d,n}, r_{d,n}, \boldsymbol{\beta}_k \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}}{}^{r_{d,n}})$$

sliding windowもDirichlet-Multinomialで簡単です



sentence sの window選択確率

$$\psi_{d,s}|\gamma \sim \mathrm{Dir}(\gamma)$$

グローバルトピック

ロンドン	.05
地下鉄	.04
五輪	.03
…	…

window vのglobal/local割合

$$\pi_{d,v}|\alpha^{mix} \sim \mathrm{Beta}(\alpha^{mix})$$

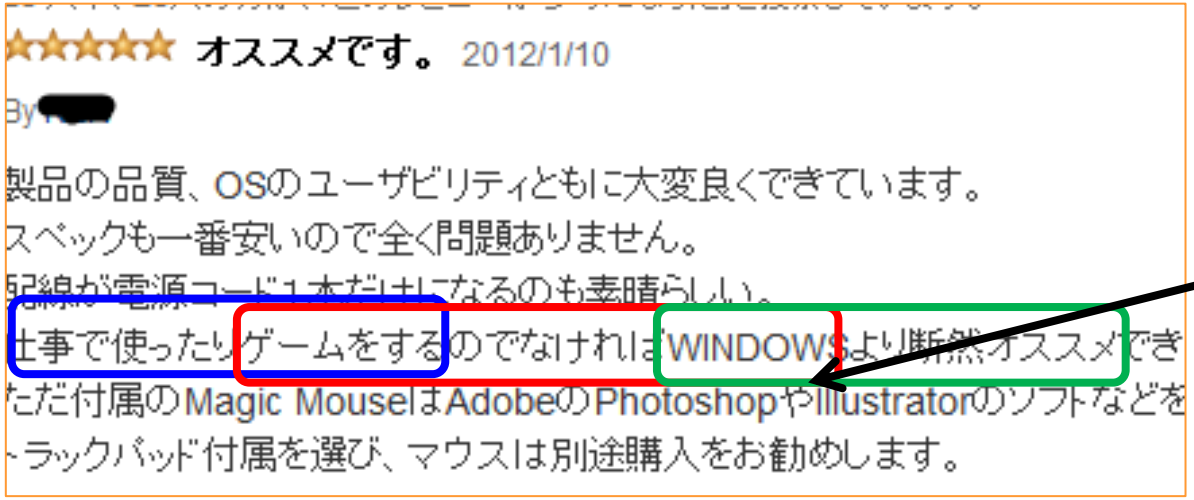
ローカルトピック

駅前	.04
徒歩圏内	.03
バス	.01
…	…

選んだトピックの種類の中で実際のトピックを決定

$$\theta_d{}^g|\alpha^g \sim \mathrm{Dir}(\alpha^g) \quad \theta_{d,v}{}^l|\alpha^l \sim \mathrm{Dir}(\alpha^l)$$

sliding windowもDirichlet-Multinomialで簡単です



sentence s内の単語 nが生成されるwindow v

$$v_{d,n}|\boldsymbol{\psi}_{d,s}\sim\mathrm{Mult}(\boldsymbol{\psi}_{d,s})$$

グローバルトピック

ロンドン	.05
地下鉄	.04
五輪	.03
…	…

ローカルトピック

駅前	.04
徒歩圏内	.03
バス	.01
…	…

window vのglobal/local割合でトピックの種類を選択

$$r_{d,n}|\pi_{d,v},v_{d,n}\sim\mathrm{Bernoulli}(\pi_{d,v_{d,n}})$$

選んだトピックの種類の中で実際のトピック kを決定

$$z_{d,n}|r_{d,n},\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_{d,(v)}{}^{r_{d,n}})$$

隠れ変数・パラメータの推定

- パラメータは積分消去可能
- 隠れ変数はGibbs samplingで最適化

- ☺ PAMと同様に、モデルの複雑さに反して、式の導出・計算結果は非常にシンプルです

隠れ変数のGibbsサンプリング

- 通常のLDA(Dirichlet-Multinomial)とほとんど同じ式を導出できます

$$p\left(v_{d,n}=v, r_{d,n}=g, z_{d,n}=k \middle| x_{d,n}=w, \{\boldsymbol{X},\boldsymbol{V},\boldsymbol{R},\boldsymbol{Z}\}_{\neg(d,n)}\right)$$

$$\propto \frac{m_{dsv}+\gamma_v}{\sum_{v'}(m_{\mathrm{ds}v'}+\gamma_{v'})} \frac{m_{dvg}+{\alpha_g}^{mix}}{\sum_{r'\in\{g,l\}}(m_{dvr'}+{\alpha_{r'}}^{mix})}$$

sentence sの単語のうち
window vから生成された単語の割合

文書 dのwindow vから
globalトピックが出てくる割合

$$\times \frac{{m_{dk}}^{g}+{\alpha_k}^{g}}{\sum_{k'}({m_{dk'}}^{g}+{\alpha_{k'}}^{g})} \frac{{m_{kv}}^{g}+{\beta_{0,w}}^{g}}{\sum_{w'}\left({m_{kw'}}^{g}+{\beta_{0,w'}}^{g}\right)}$$

文書 dの中でglobalトピック kが使われる割合

トピック kから単語 wが生成される割合

隠れ変数のGibbsサンプリング

- 通常のLDA(Dirichlet-Multinomial)とほとんど同じ式を導出できます

$$p\left(v_{d,n}=v, r_{d,n}=l, z_{d,n}=k \middle| x_{d,n}=w, \{\boldsymbol{X},\boldsymbol{V},\boldsymbol{R},\boldsymbol{Z}\}_{\neg(d,n)}\right)$$

$$\propto \frac{m_{dsv}+\gamma_v}{\sum_{v'}(m_{\mathrm{d}sv'}+\gamma_{v'})} \frac{m_{dvl}+{\alpha_l}^{mix}}{\sum_{r'\in\{g,l\}}(m_{dvr'}+{\alpha_{r'}}^{mix})}$$

sentence sの単語のうち
window vから生成された単語の割合

文書 dのwindow vから
localトピックが出てくる割合

$$\times \frac{{m_{dk}}^{l}+{\alpha_k}^{l}}{\sum_{k'}({m_{dk'}}^{l}+{\alpha_{k'}}^{l})} \frac{{m_{kv}}^{l}+{\beta_{0,w}}^{l}}{\sum_{w'}\left({m_{kw'}}^{l}+{\beta_{0,w'}}^{l}\right)}$$

文書 dの中でlocalトピック kが使われる割合

トピック kから単語 wが生成される割合

Table 2: Top words from MG-LDA and LDA topics for Mp3 players' reviews.

	label	top words
MG-LDA local (all topics)	sound quality	sound quality headphones volume bass earphones good settings ear rock excellent
	features	games features clock contacts calendar alarm notes game quiz feature extras solitaire
	connection with PC	usb pc windows port transfer computer mac software cable xp connection plug firewire
	tech. problems	reset noise backlight slow freeze turn remove playing icon creates hot cause disconnect
	appearance	case pocket silver screen plastic clip easily small blue black light white belt cover
	controls	button play track menu song buttons volume album tracks artist screen press select
	battery	battery hours life batteries charge aaa rechargeable time power lasts hour charged
	accessories	usb cable headphones adapter remote plug power charger included case firewire
	managing files	files software music computer transfer windows media cd pc drag drop file using
	radio/recording	radio fm voice recording record recorder audio mp3 microphone wma formats
MG-LDA global	iPod	ipod music apple songs use mini very just itunes like easy great time new buy really
	Creative Zen	zen creative micro touch xtra pad nomad waiting deleted labs nx sensitive 5gb eax
	Sony Walkman	sony walkman memory stick sonicstage players atrac3 mb atrac far software format
	video players	video screen videos device photos tv archos pictures camera movies dvd files view
	support	player product did just bought unit got buy work $ problem support time months
LDA (out of 40)	iPod	ipod music songs itunes mini apple battery use very computer easy time just song
	Creative	creative nomad zen xtra jukebox eax labs concert effects nx 60gb experience lyrics
	memory/battery	card memory cards sd flash batteries lyra battery aa slot compact extra mmc 32mb
	radio/recording	radio fm recording record device audio voice unit battery features usb recorder
	controls	button menu track play volume buttons player song tracks press mode screen settings
	opinion	points reviews review negative bad general none comments good please content aware
	-	player very use mp3 good sound battery great easy songs quality like just music

[Titov & McDonald, 2008]

Table 3: Top words from MG-LDA and LDA topics for hotel reviews.

	label	top words
MG-LDA local (all topics)	amenities	coffee microwave fridge tv ice room refrigerator machine kitchen maker iron dryer
	food and drink	food restaurant bar good dinner service breakfast ate eat drinks menu buffet meal
	noise/conditioning	air noise door room hear open night conditioning loud window noisy doors windows
	bathroom	shower water bathroom hot towels toilet tub bath sink pressure soap shampoo
	breakfast	breakfast coffee continental morning fruit fresh buffet included free hot juice
	spa	pool area hot tub indoor nice swimming outdoor fitness spa heated use kids
	parking	parking car park lot valet garage free street parked rental cars spaces space
	staff	staff friendly helpful very desk extremely help directions courteous concierge
	Internet	internet free access wireless use lobby high computer available speed business
	getting there	airport shuttle minutes bus took taxi train hour ride station cab driver line
	check in	early check morning arrived late hours pm ready day hour flight wait
	smells/stains	room smoking bathroom smoke carpet wall smell walls light ceiling dirty
	comfort	room bed beds bathroom comfortable large size tv king small double bedroom
	location	walk walking restaurants distance street away close location shopping shops
	pricing	$ night rate price paid worth pay cost charge extra day fee parking
MG-LDA global	beach resorts	beach ocean view hilton balcony resort ritz island head club pool oceanfront
	Las Vegas	vegas strip casino las rock hard station palace pool circus renaissance
LDA (out of 45)	beach resorts	beach great pool very place ocean stay view just nice stayed clean beautiful
	Las Vegas	vegas strip great casino $ good hotel food las rock room very pool nice
	smells/stains	room did smoking bed night stay got went like desk smoke non-smoking smell
	getting there	airport hotel shuttle bus very minutes flight hour free did taxi train car
	breakfast	breakfast coffee fruit room juice fresh eggs continental very toast morning
	location	hotel rooms very centre situated well location excellent city comfortable good
	pricing	card credit $ charged hotel night room charge money deposit stay pay cash did
	front desk	room hotel told desk did manager asked said service called stay rooms
	noise	room very hotel night noise did hear sleep bed door stay floor time just like
	opinion	hotel best stay hotels stayed reviews service great time really just say rooms
	cleanliness	hotel room dirty stay bathroom rooms like place carpet old very worst bed
	-	motel rooms nice hotel like place stay parking price $ santa stayed good

[Titov & McDonald, 2008]

Table 4: Multi-aspect ranking experiments with the PRanking algorithm for hotel reviews.

Unigram features only

Model	Overall	Check-in	Service	Value	Location	Rooms	Cleanliness
Baseline	1.118	1.126	1.208	1.272	0.742	1.356	1.002
PRank	0.774	0.831	0.799	0.793	0.707	0.798	0.715
PRank + LDA	0.735	0.786	0.762	0.749	0.677	0.746	0.690
PRank + MG-LDA	**0.706**	**0.748**	**0.731**	**0.725**	**0.635**	**0.719**	**0.676**

Unigram, bigram and trigram features

Model	Overall	Check-in	Service	Value	Location	Rooms	Cleanliness
PRank	0.689	0.735	0.725	0.710	0.627	0.700	0.637
PRank + LDA	0.682	0.728	0.717	0.705	0.620	0.684	0.637
PRank + MG-LDA	**0.669**	**0.717**	**0.700**	**0.696**	**0.607**	**0.672**	0.636

[Titov & McDonald, 2008]

まとめ: Multi-grained Topic Models

- 主にレビュー記事を対象に、文章を全体的なトピックとaspectに関するトピックに分解するモデルです
- Sliding windowを基準にしたトピックモデルになります
- 推論はモデルの複雑さに反してシンプルです

その他の構造トピックモデル

- Blei et al., “Hierarchical Topic Models and the Nested Chinese Restaurant Process”, in Advances in Neural Information Processing Systems 16 (Proc. NIPS), 2003.

- Titov and McDonald, “A Joint Model of Text and Aspect Ratings for Sentiment Summarization”, in Proc. ACL, 2008.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Blei & Lafferty, 2007] Blei and Lafferty, “A Correlated Topic Model of Science”, The Annals of Applied Statistics, Vol. 1(1), pp. 17-35, 2007.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.
- [Li & McCallum, 2006] Li and McCallum, “Pachinko Allocation: DAG-Structured Mixture Models of Topic Correlations”, in Proc. ICML, 2006.
- [Titov & McDonald, 2008] Titov and McDonald, “Modeling Online Reviews with Multi-grained Topic Models”, in Proc. WWW, 2008.

NTT

トピックモデルの応用：時系列データ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

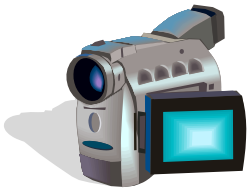
このスライドの“トピック”

- 購買データや科学論文など、時間変化をそもそも内包するデータは多数存在します

- 従って、時系列（時間変化）データ内のトピックの解析も多数試みがなされています

時系列データは数多く存在します

動画像



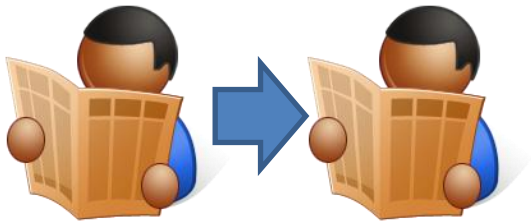
脳波・生体信号

購買履歴・市場インデックス

音楽・音響信号

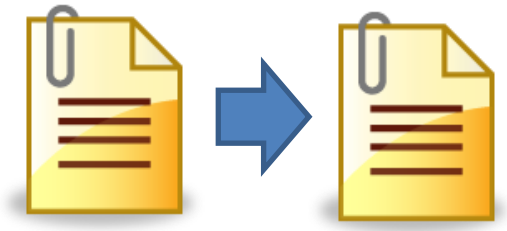
新聞・ニュース記事

01/15 01/16



科学論文・特許

2012 2013



☺ トピックモデルの有用性が明白になったため、これら時系列データへトピックモデルを応用した研究が多数発表されています

時系列データモデリングのキモ：どこに時間依存性を入れるか？

- マルコフ性：前の時刻に依存して現在の時刻の状態が変化する

- 多くの時系列データでは、モデルのどの部分にマルコフ性のアイデアを導入するか、がポイントとなります

- これはトピックモデルの時系列データモデルでも同様です

Dynamic Topic Model
[Blei & Lafferty, 2006]

Blei and Lafferty,
"Dynamic Topic Models",
in Proc. ICML, 2006.

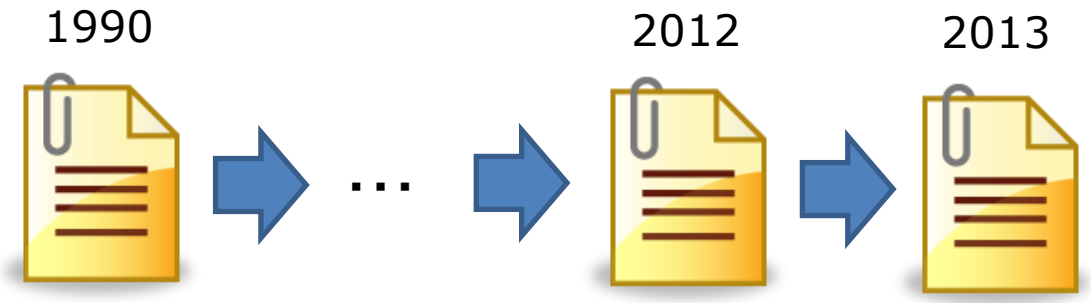
トピックモデルの大前提の仮定: exchangeability

- 簡単にいうと:「各文書 d, 各単語 wのインデックスはただのシンボルで順番や名前には意味が無い」

- これのおかげで各種モデル推論が簡単になっています

文書コレクションの時系列データを考えると、これは問題です

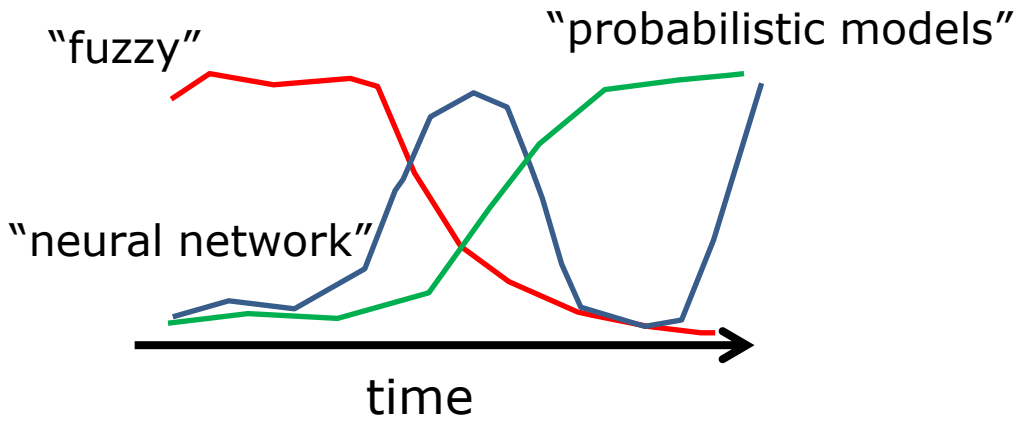
- 新聞記事は昨日までの報道の流れを汲んでいます
- 論文は先行研究の作った技術トレンドにのっています
- すなわち、文書 d は一般にはexchangeableではありません！



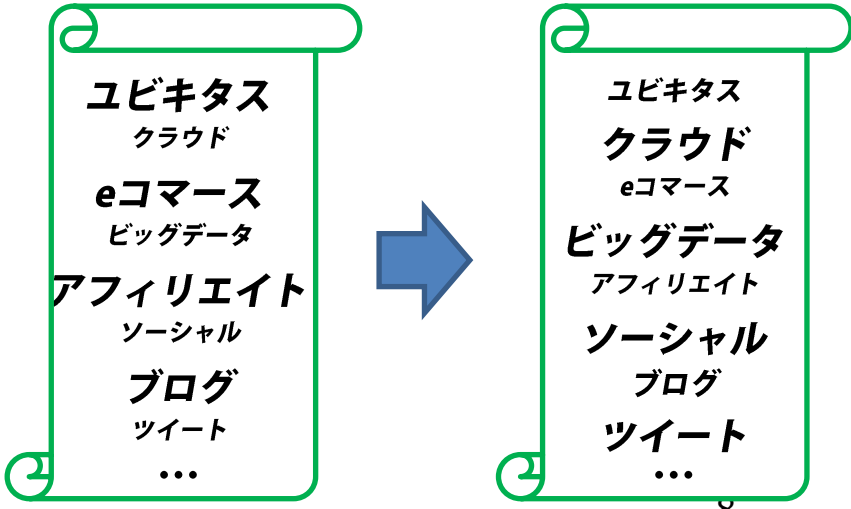
注目する時間依存性：トピック

- 1: 話題(topic)には流行り廃りがあります
- 2: トピックの中での言葉づかいも変化します
- これら2種類の「トピックの変化」を解析するモデルを考えたい

トピックの流行り廃り



トピックの中での言葉づかい

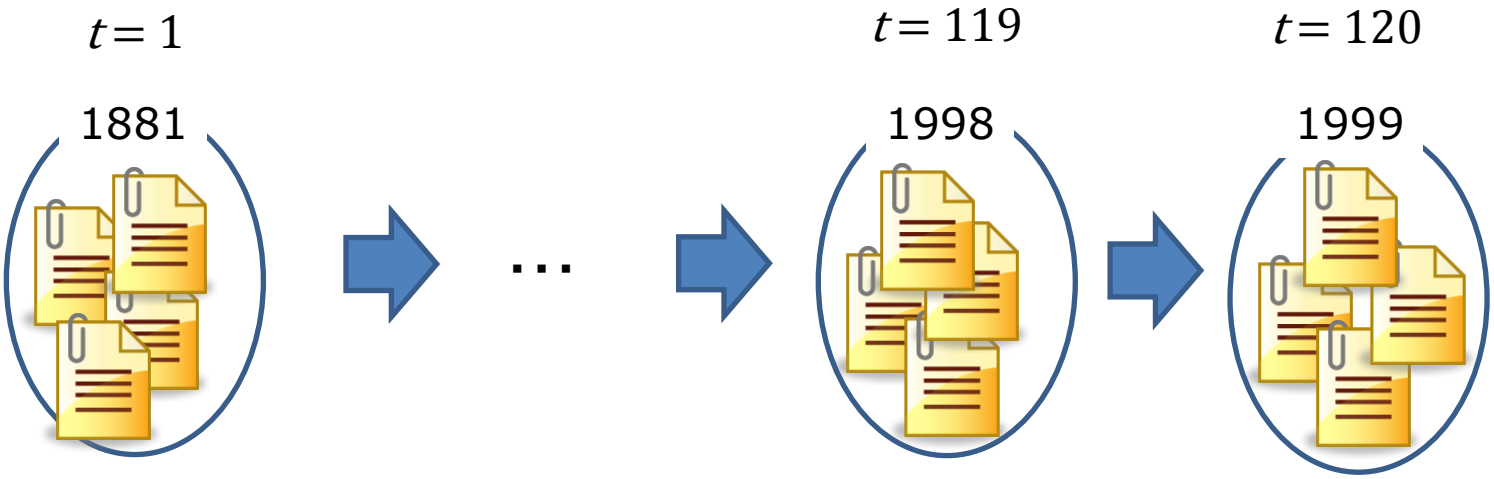


提案法: Dynamic Topic Models

- ☺ 非常に有名な時系列トピックモデルです
- 科学誌ScienceのOCRデータを用いて、科学論文の時系列トピック解析を行います
- topic proportionとtopic-word proportionに時間マルコフ性を入れたものです
- 推論は非常に難しいです

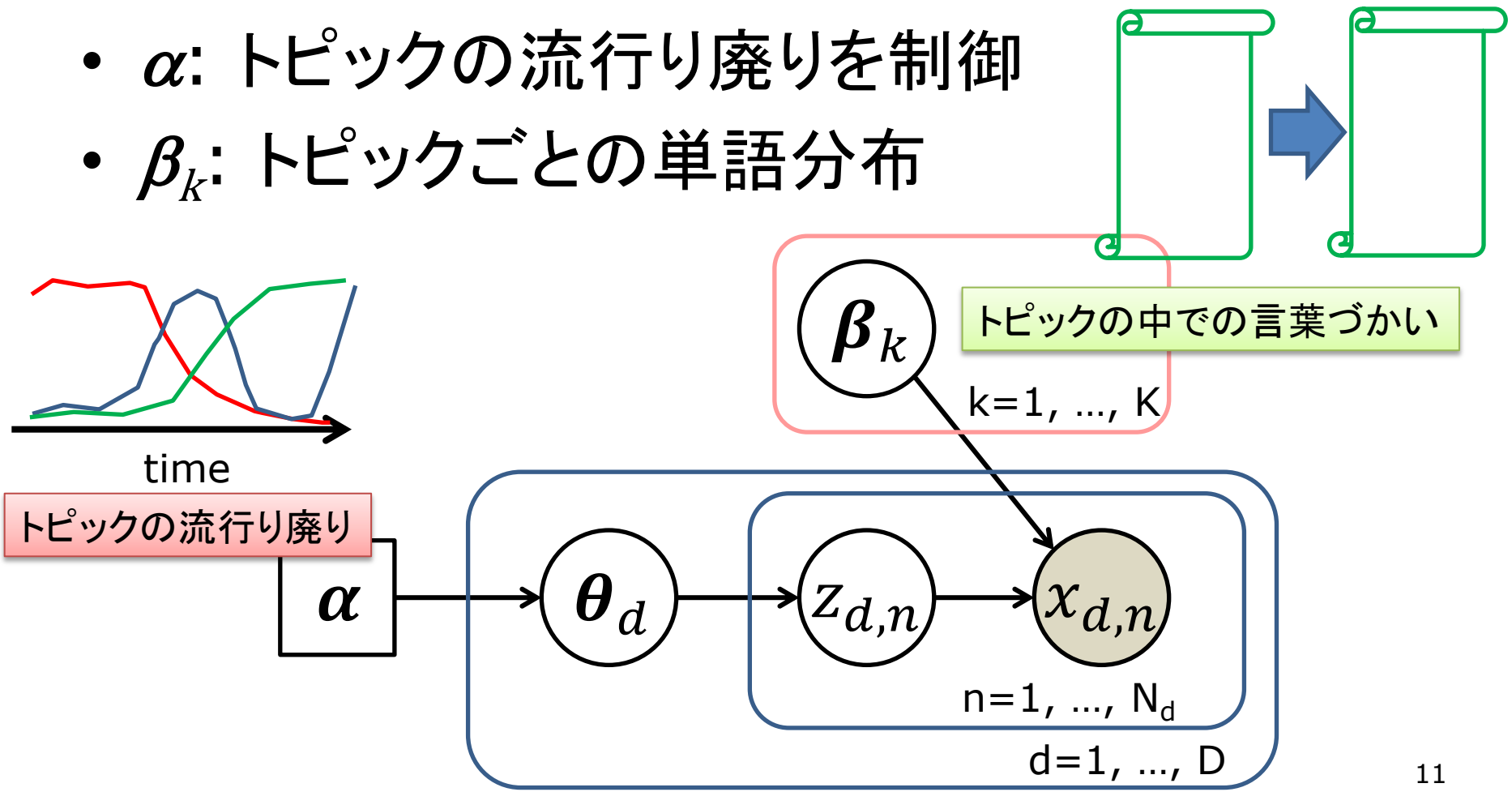
対象データ: Science誌

- 1880年にエジソンによって刊行された、非常に著名な科学論文誌
- OCRされた論文誌データ(JSTOR)を利用して、発行年度ごとの文書時系列データを作成



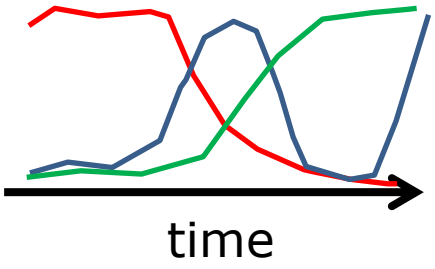
提案法のアイデア

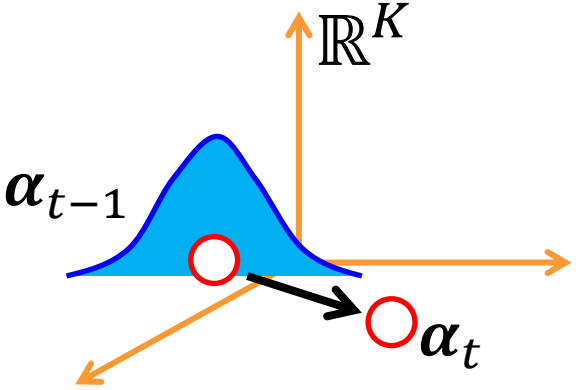
- 以下の2点を時間発展させます
- α: トピックの流行り廃りを制御
- β_k: トピックごとの単語分布

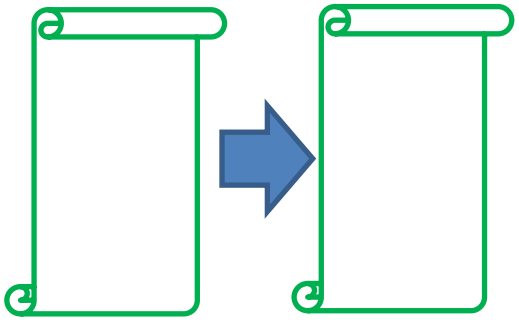


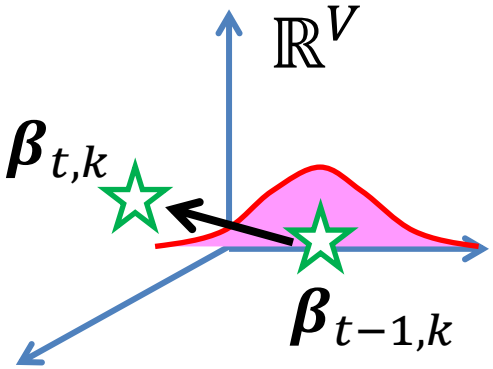
時間発展のモデル：正規ノイズでdrift

- 最も単純な時間発展モデルと言えます
- パラメータは前の時刻を中心に少しずつしか動かない、という想定です

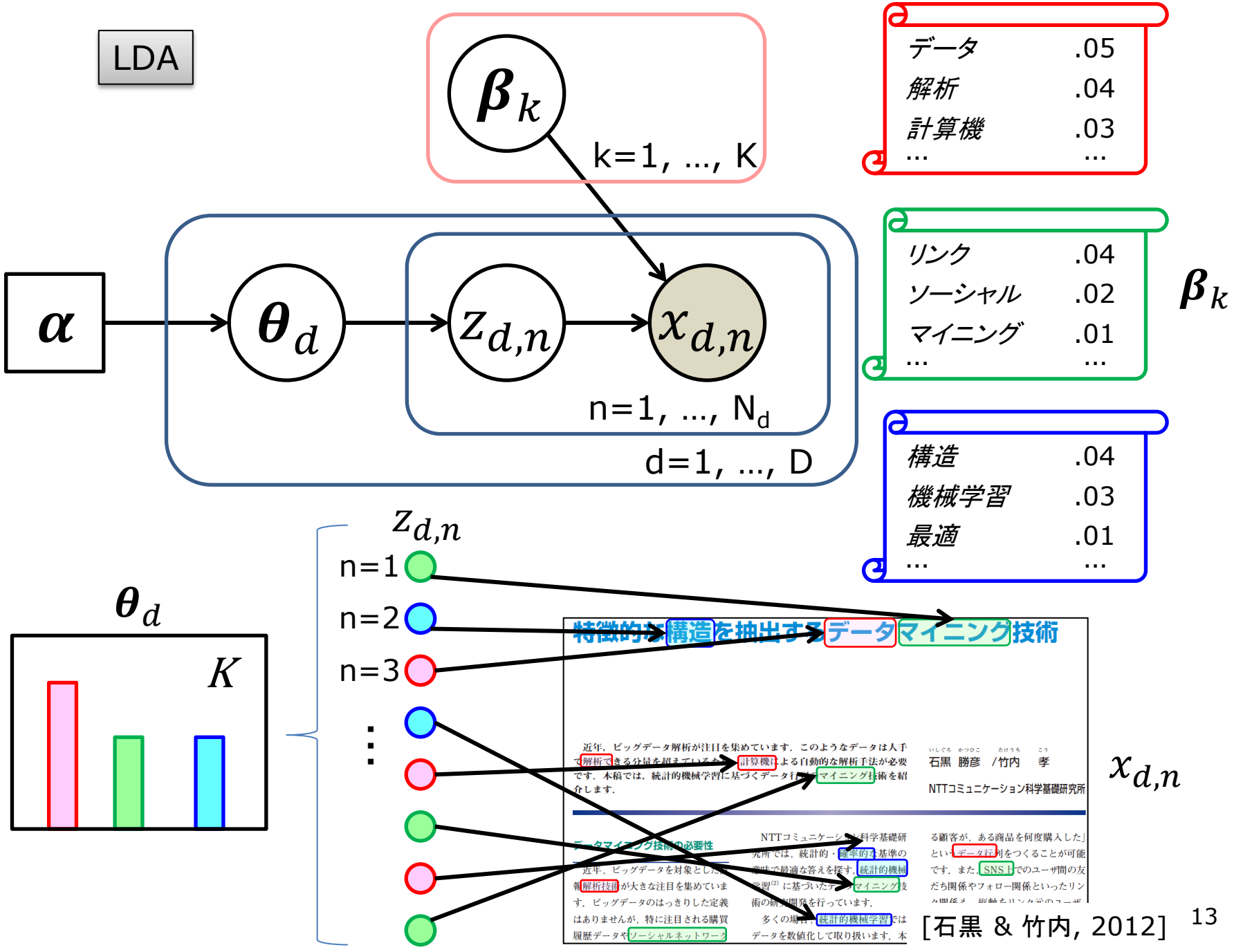






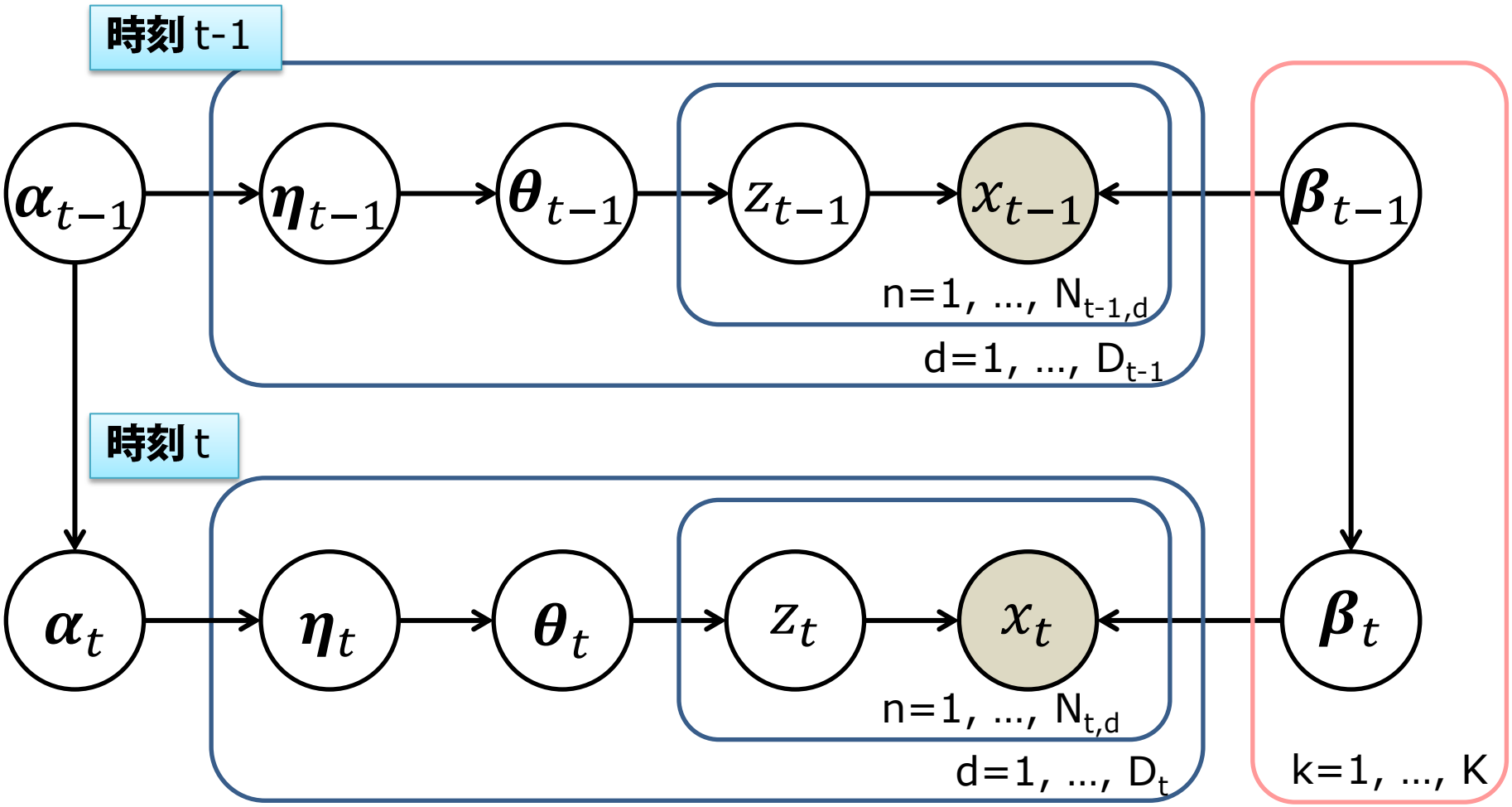


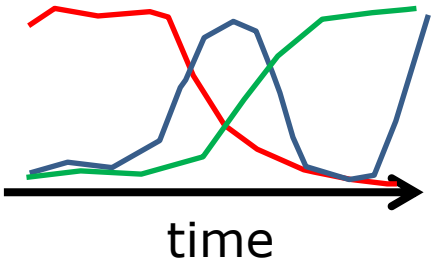
LDA

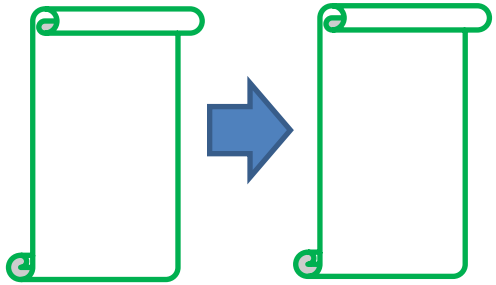


[石黒 & 竹内, 2012]

Dynamic Topic Model (添え字d,n,kは省略)







生成モデル

for 時間t = 1, 2, …, T

for theme (topic) k = 1, 2, …, K

topic-word proportion drift $\boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k}\sim\mathrm{N}\left(\boldsymbol{\beta}_{t-1,k},\sigma^2\boldsymbol{I}\right)$

topic proportion parameter drift $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}\sim\mathrm{N}(\boldsymbol{\alpha}_{t-1},\delta^2\boldsymbol{I})$

for 文書d = 1, 2, …, D_t

topic proportion

for 単語n = 1, 2, …, $N_{t,d}$

topic-word assignment

word observation

for 時間t = 1, 2, …, T

$$\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}\sim\mathrm{N}(\boldsymbol{\alpha}_{t-1},\delta^2\boldsymbol{I}) \qquad \boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1,k},\sigma^2\boldsymbol{I})$$

for 文書d = 1, 2, …, D_t

topic proportion

$$\boldsymbol{\eta}_{t,d}|\boldsymbol{\alpha}_t\sim\mathrm{N}(\boldsymbol{\alpha}_t,a^2\boldsymbol{I})$$

$$\boldsymbol{\theta}_{t,d}|\boldsymbol{\eta}_{t,d}=\pi(\boldsymbol{\eta}_{t,d})$$

for 単語n = 1, 2, …, N_d

topic-word assignment

$$z_{t,d,n}|\boldsymbol{\theta}_{t,d}\sim\mathrm{Multinomial}(\boldsymbol{\theta}_{t,d})$$

word observation

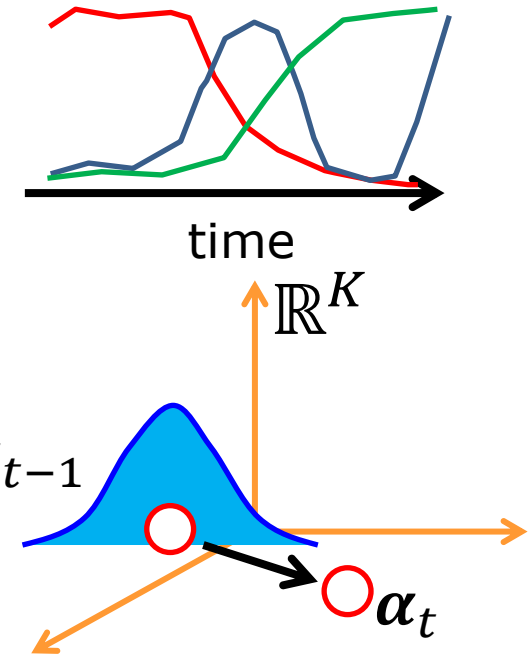
$$x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_{t,k}\}\sim\mathrm{Multinomial}\left(\pi(\boldsymbol{\beta}_{t,z_{d,n}})\right)$$

πはsoft-max関数 $\pi(\boldsymbol{v})=\dfrac{\exp(v_k)}{\sum_l\exp(v_l)}$

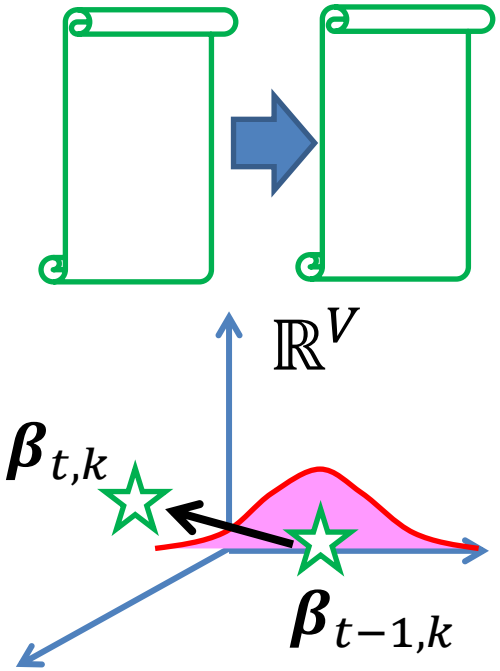
時間発展のモデル

- 正規分布を使って、1時刻のパラメータ遷移(drift)をモデル化します

$\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1}\sim\mathrm{N}(\boldsymbol{\alpha}_{t-1},\delta^2\boldsymbol{I})$



$\boldsymbol{\beta}_{t,k}|\boldsymbol{\beta}_{t-1,k}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1,k},\sigma^2\boldsymbol{I})$



トピックモデルへの適合

- 正規分布からは実数ベクトルが生成されるため、そのままでは多項分布(Multinomial)に使えません
- Soft-max関数を利用して変換します

$$\pi(\boldsymbol{v}) = \frac{\exp(v_k)}{\sum_l \exp(v_l)}$$

$$\boldsymbol{\eta}_{t,d}|\boldsymbol{\alpha}_t \sim \mathrm{N}(\boldsymbol{\alpha}_t, a^2\boldsymbol{I})$$ 時刻t, 文書dのtopic proportion

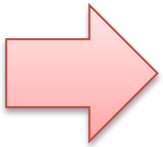
$$\boldsymbol{\theta}_{t,d}|\boldsymbol{\eta}_{t,d} = \pi(\boldsymbol{\eta}_{t,d})$$ Soft-max

$$z_{t,d,n}|\boldsymbol{\theta}_{t,d} \sim \mathrm{Multinomial}(\boldsymbol{\theta}_{t,d})$$ topic-word assign.

$$x_{d,n}|z_{d,n}, \{\boldsymbol{\beta}_{t,k}\} \sim \mathrm{Multinomial}\left(\pi(\boldsymbol{\beta}_{t,z_{d,n}})\right)$$

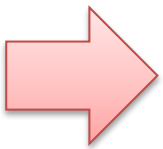
隠れ変数・パラメータの推定：非常に難しくなります

- 原因1: Soft-max関数のため、共役性(conjugate)を利用できません ☹



(collapsed) Gibbs sampling**が不可能になるため、変分ベイズ法が候補になります**

- 原因2: 時刻 t が前時刻 t-1 に依存するため、時間依存性を考慮した推定が必要になります ☹



時間発展するパラメータは、時刻依存性を考慮して変分ベイズ法を構築する必要があります

時間発展パラメータの推定: 状態空間モデル解釈[北川, 2005]

- 連続なパラメータの時間変化を追いかける定番の手法です
- DTMの時間発展部分も状態空間モデルとして解釈できます

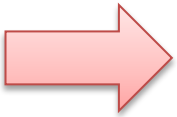
	一般の状態空間モデル	DTM(k, d, zなどを省略)
状態モデル	$y_t \mid y_{t-1} \sim f(y_{t-1}, \theta)$	$\boldsymbol{\beta}_t \mid \boldsymbol{\beta}_{t-1} \sim \mathrm{N}(\boldsymbol{\beta}_{t-1}, \sigma^2 \boldsymbol{I})$
観測モデル	$x_t \mid y_t \sim g(x_t, \varphi)$	$x_{t,n} \mid \boldsymbol{\beta}_t \sim \mathrm{Mult}\left(\pi(\boldsymbol{\beta}_t)\right)$

変分近似によるKalman filter
[Kalman, 1960]

- 状態モデル、観測モデルの双方が正規分布の場合, Kalman Filterを用いてexactな解が計算できます
- 変分事後分布として、観測モデルに正規分布を“強引に”仮定して推論します

$$\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1},\sigma^2\boldsymbol{I})$$

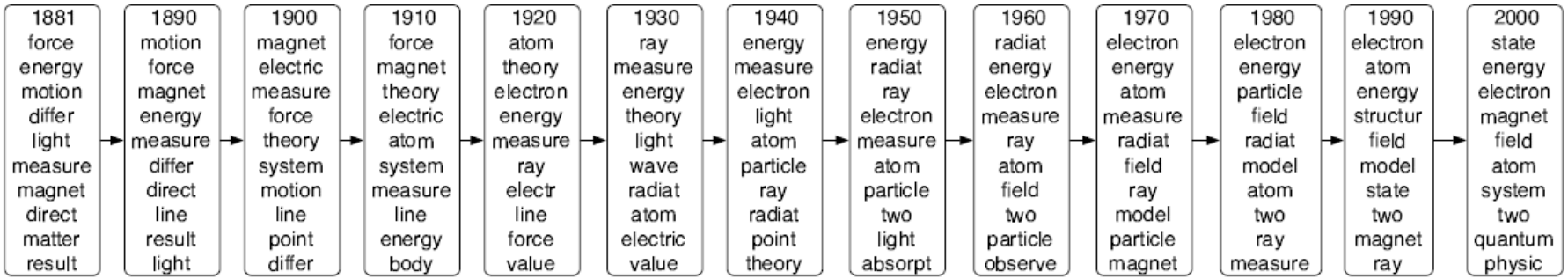
$$x_{t,n}|\boldsymbol{\beta}_t\sim\mathrm{Mult}\big(\pi(\boldsymbol{\beta}_t)\big)$$



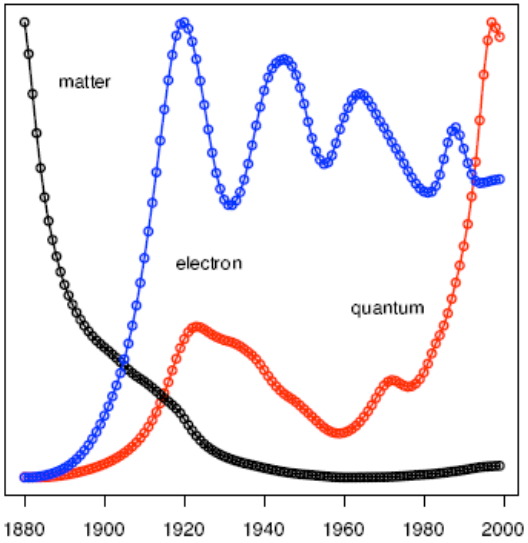
$$\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}\sim\mathrm{N}(\boldsymbol{\beta}_{t-1},\sigma^2\boldsymbol{I})$$

$$\widehat{\boldsymbol{\beta}}_t|\boldsymbol{\beta}_t\sim\mathrm{N}(\boldsymbol{\beta}_t,\hat{v}_t\boldsymbol{I})$$

変分観測量

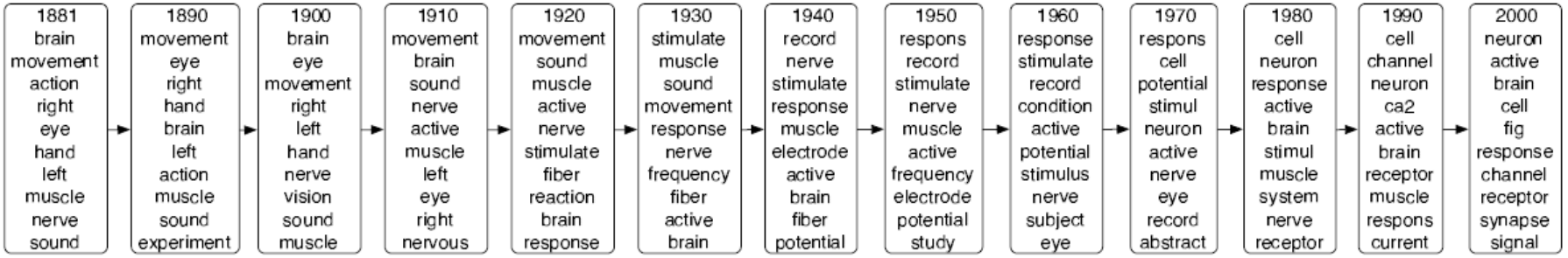


"Atomic Physics"

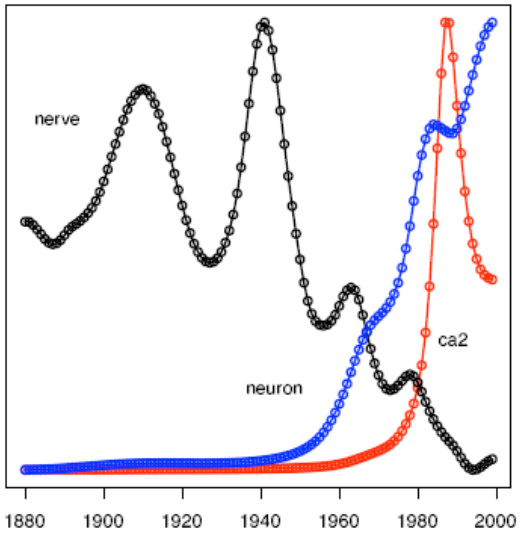


1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 ``Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

[Blei & Lafferty, 2006]

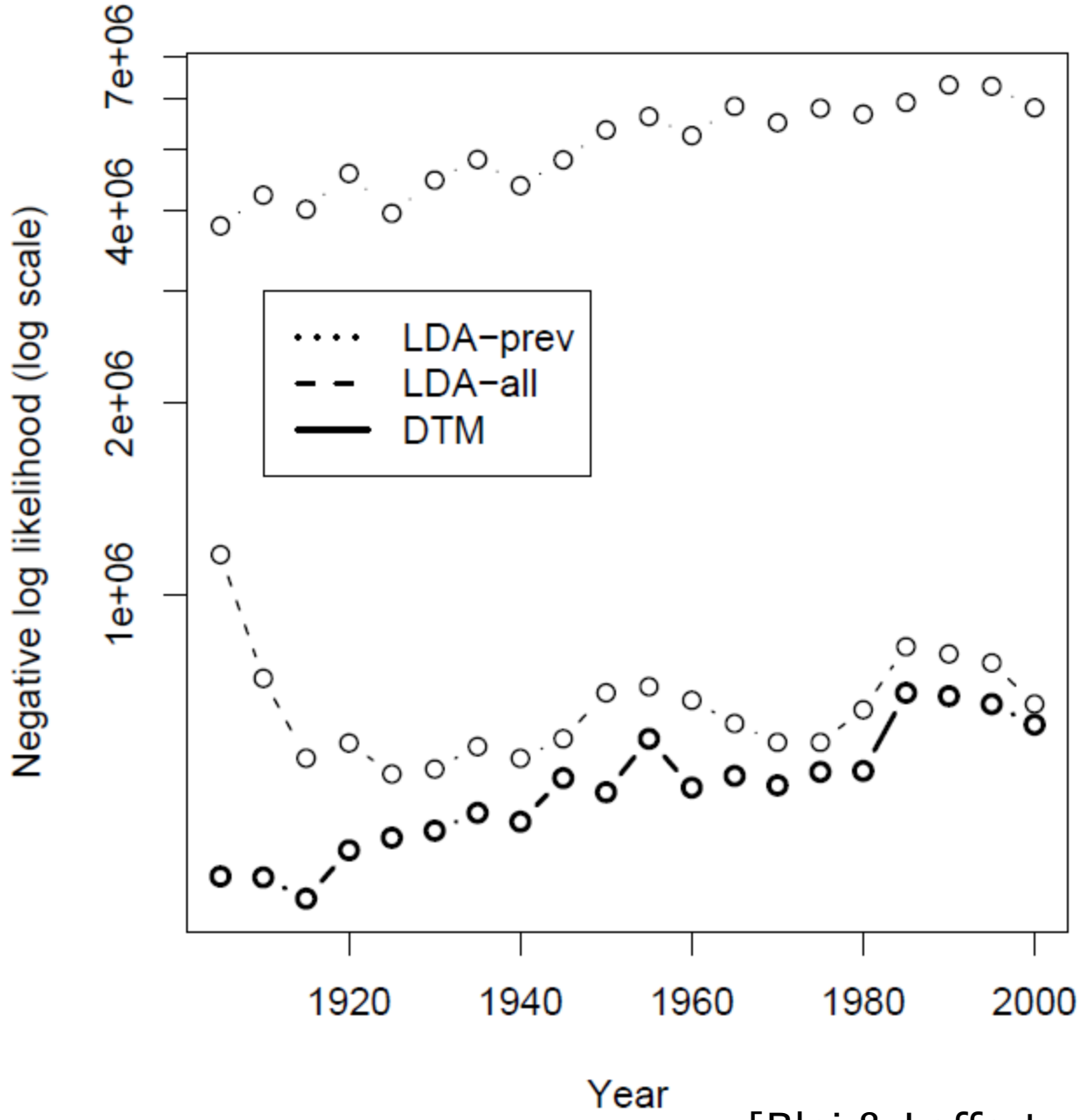


"Neuroscience"



1887 Mental Science
1900 Hemianopsia in Migraine
1912 A Defence of the ``New Phrenology''
1921 The Synchronal Flashing of Fireflies
1932 Myoesthesis and Imageless Thought
1943 Acetylcholine and the Physiology of the Nervous System
1952 Brain Waves and Unit Discharge in Cerebral Cortex
1963 Errorless Discrimination Learning in the Pigeon
1974 Temporal Summation of Light by a Vertebrate Visual Receptor
1983 Hysteresis in the Force-Calcium Relation in Muscle
1993 GABA-Activated Chloride Channels in Secretory Nerve Endings

[Blei & Lafferty, 2006]



[Blei & Lafferty, 2006]

まとめ: Dynamic Topic Models

- トピックごとの単語分布、トピックの割合の二つを時間発展させたトピックモデルです
- 正規分布によるdriftで時間遷移を表現します
- ☺ 非常に有名なので、時間モデルでは必ず押さえる必要がある論文です

Topic Tracking Model
[Iwata, 2009]

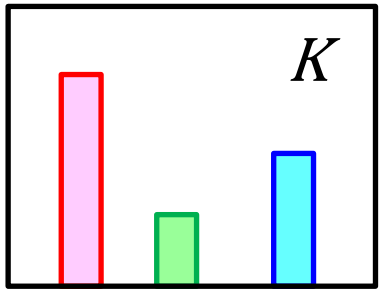
Iwata et al,
"Topic Tracking Model for Analyzing Consumer Purchase Behavior",
in Proc. IJCAI, 2009.

購買履歴データへの トピックモデル応用

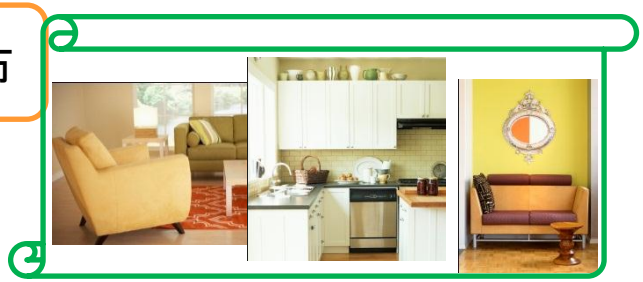
- PLSIなどのように、潜在変数モデルを使った購買履歴データのモデリングは多数存在します (e.g. [Jin, 2004])

- 当然、トピックモデルによる購買履歴データモデリングを考慮することもできます

購買履歴データの
ナイーブなトピック化



興味トピック k ごとのアイテム分布



ユーザ d の
興味トピック分布

$\boldsymbol{\beta}_k$

k=1, …, K

$\boldsymbol{\alpha}$

$\boldsymbol{\theta}_d$

$z_{d,n}$

$x_{d,n}$

n=1, …, N_d

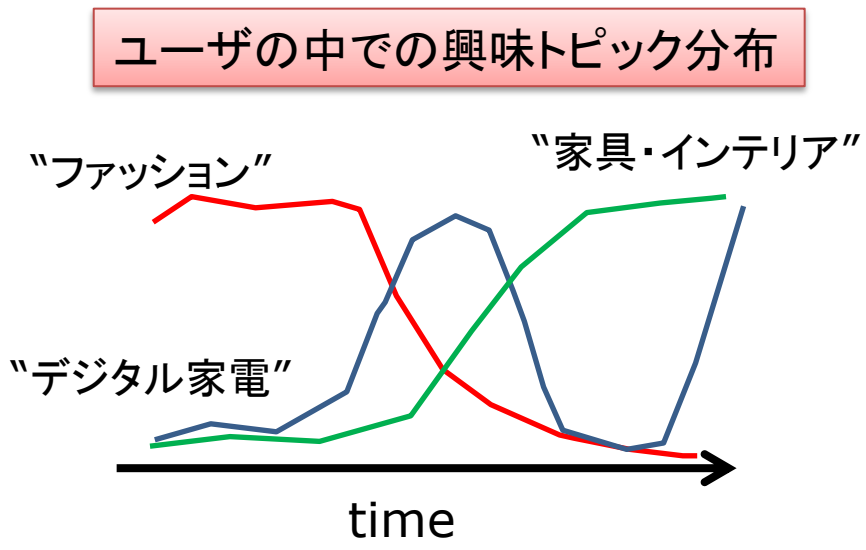
d=1, …, D

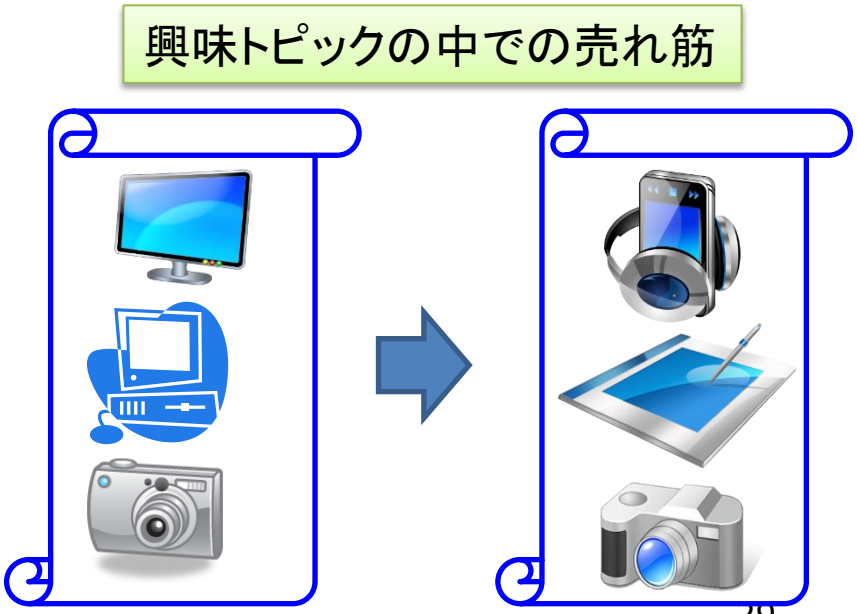
購買行動 n の潜在興味トピック

購買行動 n

注目する時間依存性：ユーザ＋トピック

- 1: ユーザの興味は少しづつ変わります
- 2: 興味トピックの中でのアイテムの売れ筋も変化します



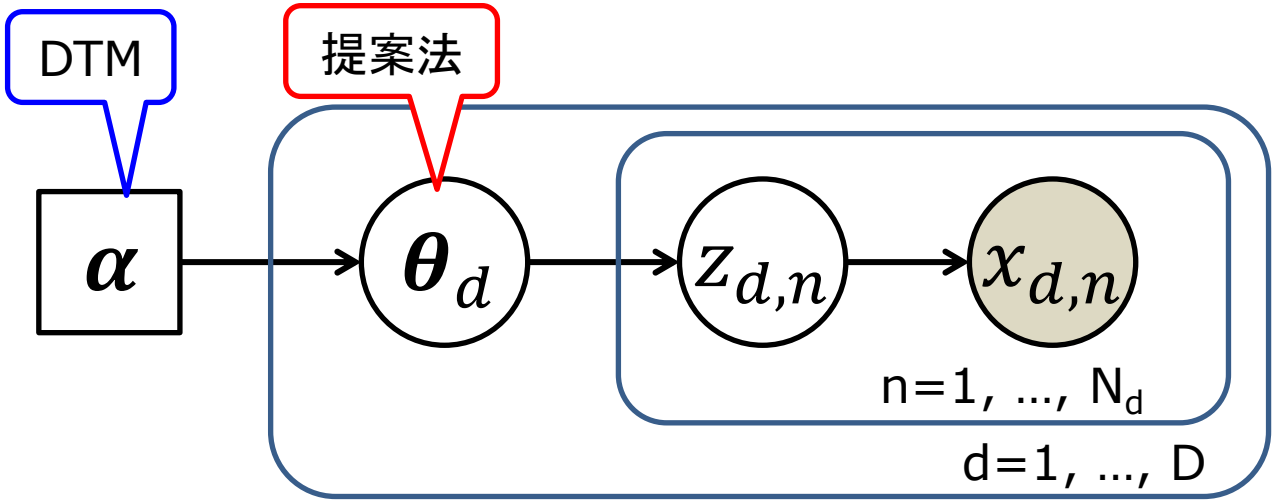


提案法: Topic Tracking Model

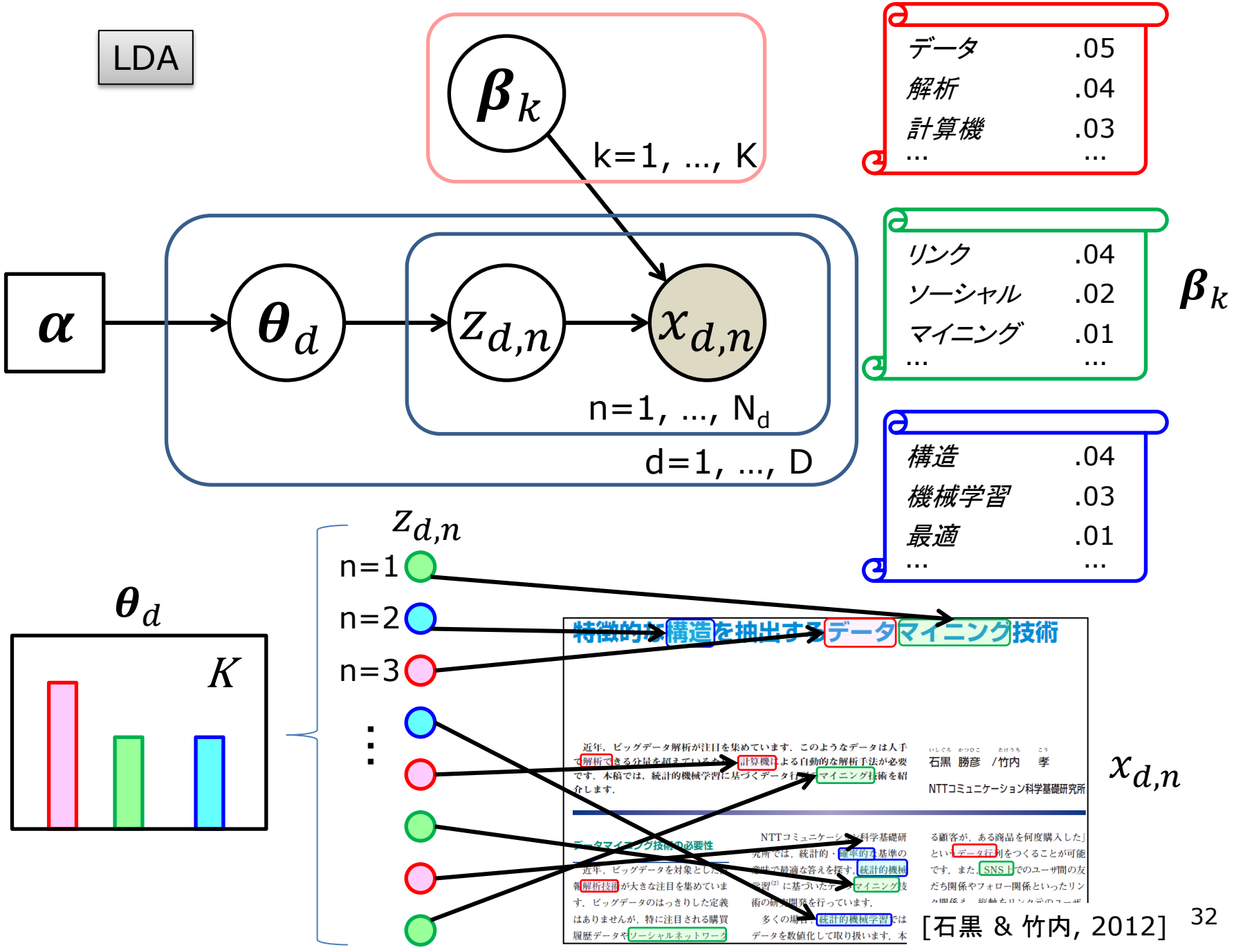
- Dynamic Topic Model(DTM)とはまた違う時系列トピックモデルです
- 文書（ユーザ）ごとのトピック分布と、トピックの単語（アイテム）分布が時間遷移します
- 推論はDTMに比べて少し簡単になるように工夫されています

提案法のアイデア: DTMとモデリングの観点が違います

- DTM: 各年度での話題の隆盛が知りたい → トピック分布制御パラメータ α を時間依存
- 提案法:ユーザの興味の変化が知りたい → 各ユーザ(文書)のトピック分布 θ を時間依存させる

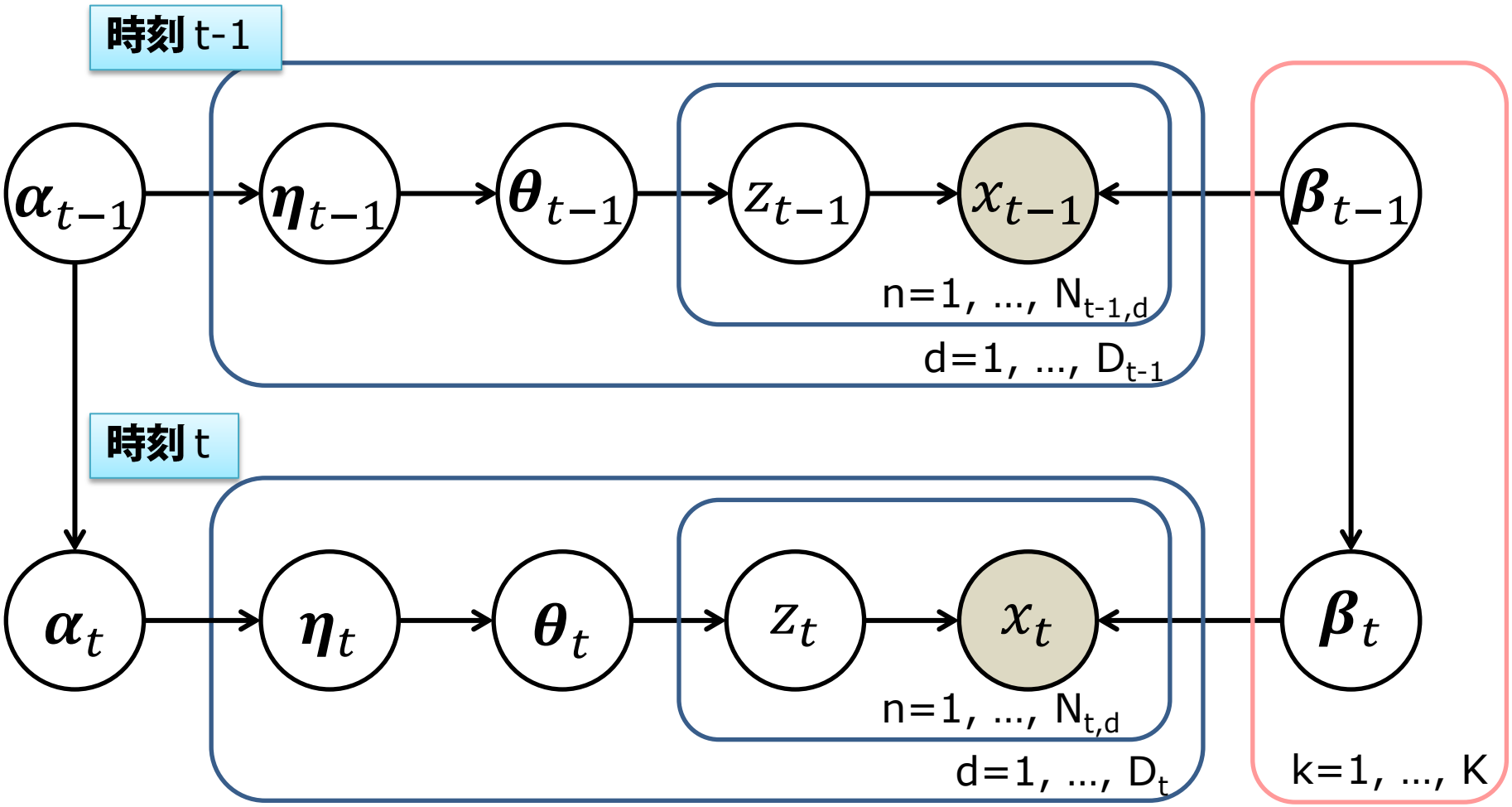


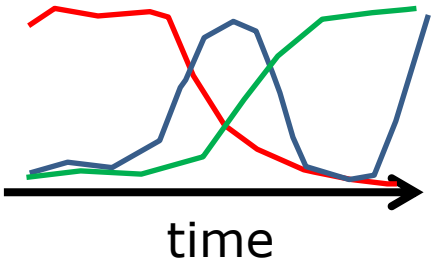
LDA

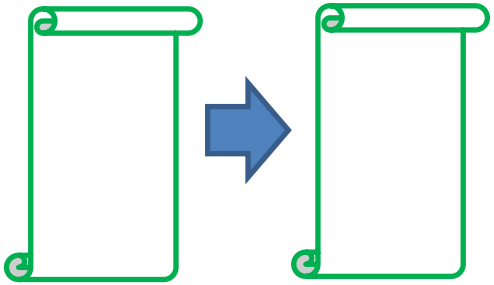


[石黒 & 竹内, 2012]

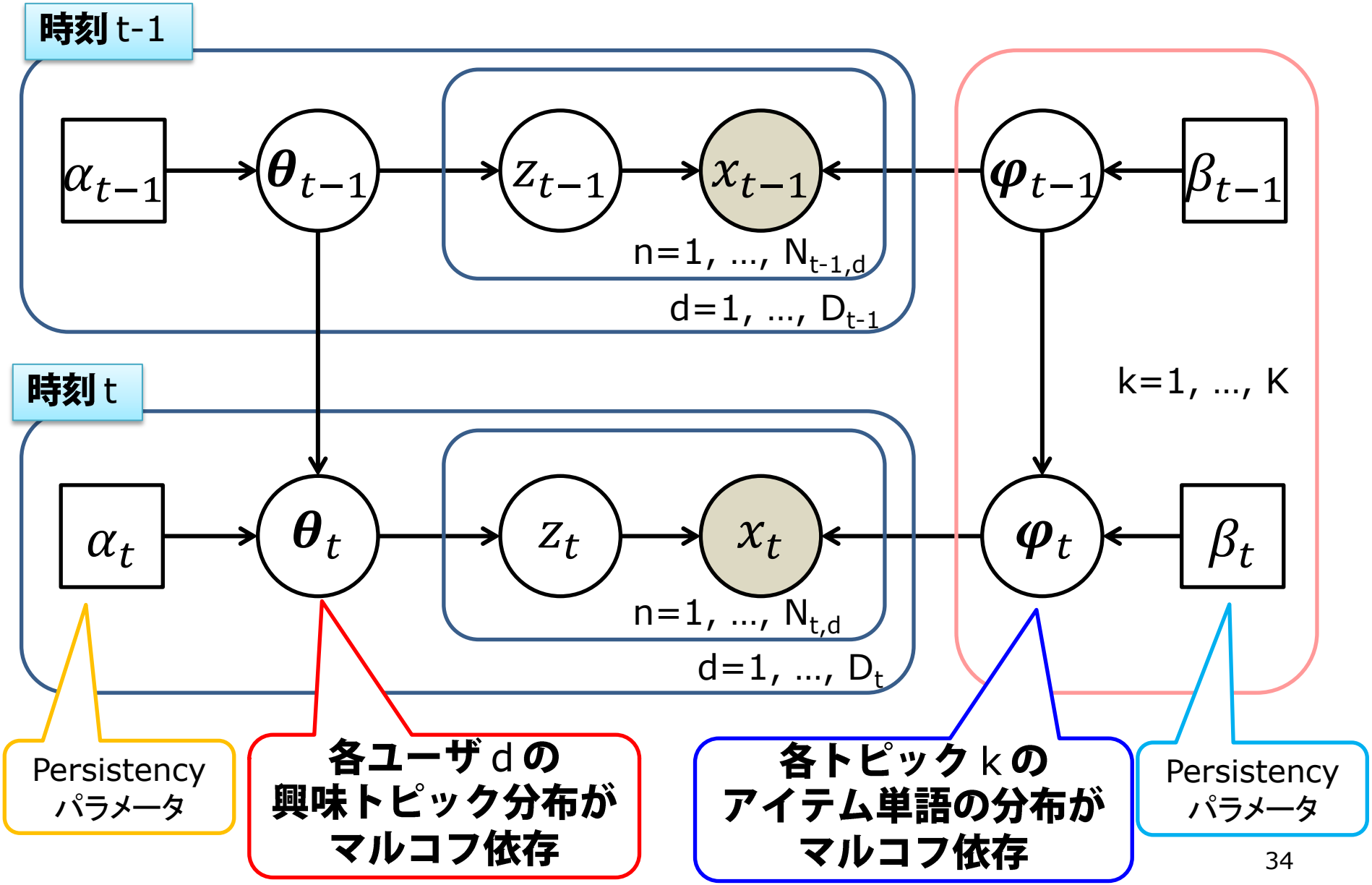
Dynamic Topic Model (添え字d,n,kは省略)







Topic Tracing Model （添え字d,n,kは省略）



生成モデル

for 時間t = 1, 2, …, T

for 興味topic k = 1, 2, …, K

topic-item word proportion parameter $\boldsymbol{\beta}_{t,k}$

for ユーザ d = 1, 2, …, D_t

topic proportion parameter t $\boldsymbol{\alpha}_{t,d}$

topic proportion

for 購買行動 n = 1, 2, …, $N_{t,d}$

topic-item word assignment

item word observation

for 時間t = 1, 2, …, T

for 興味topic k = 1, 2, …, K

topic-item word proportion evolution

$$\boldsymbol{\varphi}_{t,k}|\hat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\beta_{t,k}\hat{\boldsymbol{\varphi}}_{t-1,k}\right)$$

for ユーザ d = 1, 2, …, D_t

topic proportion evolution

$$\boldsymbol{\theta}_{t,d}|\hat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\alpha_{t,d}\hat{\boldsymbol{\theta}}_{t-1,d}\right)$$

for 購買行動 n = 1, 2, …, N_d

topic-item word assignment

$$z_{t,d,n}|\boldsymbol{\theta}_{t,d}\sim\mathrm{Mult}\left(\boldsymbol{\theta}_{t,d}\right)$$

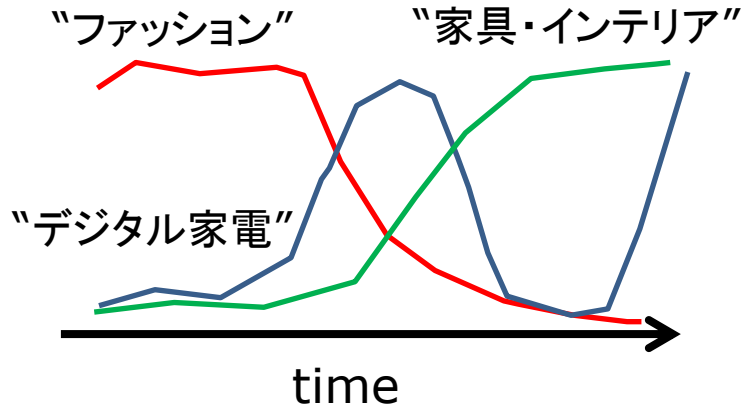
item word observation

$$x_{t,d,n}|z_{t,d,n},\left\{\boldsymbol{\varphi}_{t,k}\right\}\sim\mathrm{Mult}\left(\boldsymbol{\varphi}_{t,z_{t,d,n}}\right)$$

$\hat{\cdot}$ は”事後分布での期待値”を表す

興味トピック分布のモデル

- DTMと違い、ディリクレ分布を利用して時間発展をモデル化しています
- ユーザ、時間ごとに、興味トピックの持続度(persistency)もモデル化します



平均 $\hat{\theta}_{t-1,d}$ のディリクレ分布

$$\boldsymbol{\theta}_{t,d}|\hat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\alpha_{t,d}\hat{\boldsymbol{\theta}}_{t-1,d}\right)$$

$\hat{\boldsymbol{\theta}}_{t-1,d}$ **: 前時刻の事後分布期待値**

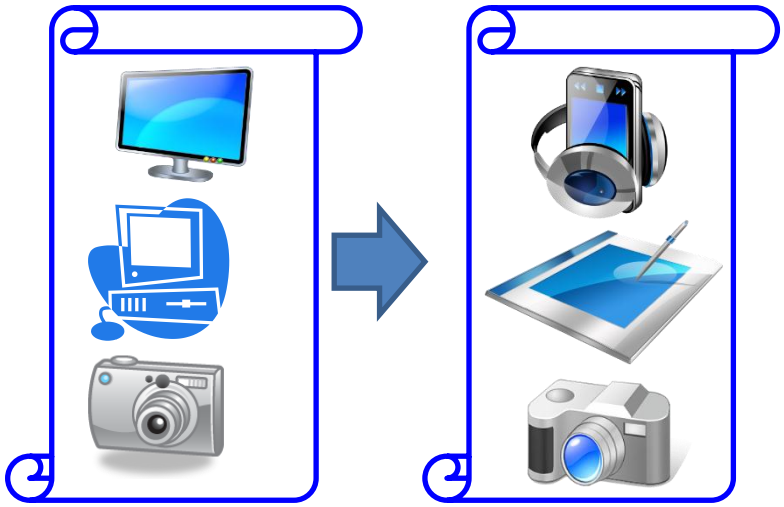
$\alpha_{t,d}$**: 持続度パラメータ**

α大 ＝ θ_tの分散小 → 小さな時間変化

α小 ＝ θ_tの分散大 → 大きな時間変化

トピック-アイテム（単語）分布のモデル

- 興味トピックと同様です



平均 $\hat{\varphi}_{t-1,d}$ のディリクレ分布

$$\boldsymbol{\varphi}_{t,k}|\hat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\beta_{t,k}\hat{\boldsymbol{\varphi}}_{t-1,k}\right)$$

$\hat{\varphi}_{t-1,d}$：前時刻の事後分布期待値

$\beta_{t,d}$：持続度パラメータ

β大 ＝ ϕ_tの分散小→ 小さな時間変化

β小 ＝ ϕ_tの分散大 → 大きな時間変化

長期時間依存モデル

- 1時刻前に依存するだけでなく、数ステップ前までに依存する形への拡張も簡単です

1ステップ前からの依存関係モデル

$$\boldsymbol{\varphi}_{t,k}|\widehat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\beta_{t,k}\widehat{\boldsymbol{\varphi}}_{t-1,k}\right)$$

$$\boldsymbol{\theta}_{t,d}|\widehat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\alpha_{t,d}\widehat{\boldsymbol{\theta}}_{t-1,d}\right)$$

Lステップ前からの依存関係モデル

$$\boldsymbol{\varphi}_{t,k}|\widehat{\boldsymbol{\varphi}}_{t-1,k},\beta_{t,k}\sim\mathrm{Dir}\left(\sum_{l=1}^{L}\beta_{t,k,l}\widehat{\boldsymbol{\varphi}}_{t-l,k}\right)$$

$$\boldsymbol{\theta}_{t,d}|\widehat{\boldsymbol{\theta}}_{t-1,d},\alpha_{t,d}\sim\mathrm{Dir}\left(\sum_{l=1}^{L}\alpha_{t,d,l}\widehat{\boldsymbol{\theta}}_{t-l,d}\right)$$

隠れ変数・パラメータの推定

- ☺ 非常に簡単な逐次推定アルゴリズムが導出できます
  - 正規分布やsoft-maxがないため！！
  - LDAのGibbs, VB (EM) を導出したことがある方にとっては自明な解が得られます
- ☹ ただし、DTMのように系列としての最適解は得られません

Table 1: Average N-best accuracies (%) over time. The digit in the bracket is the standard deviation.

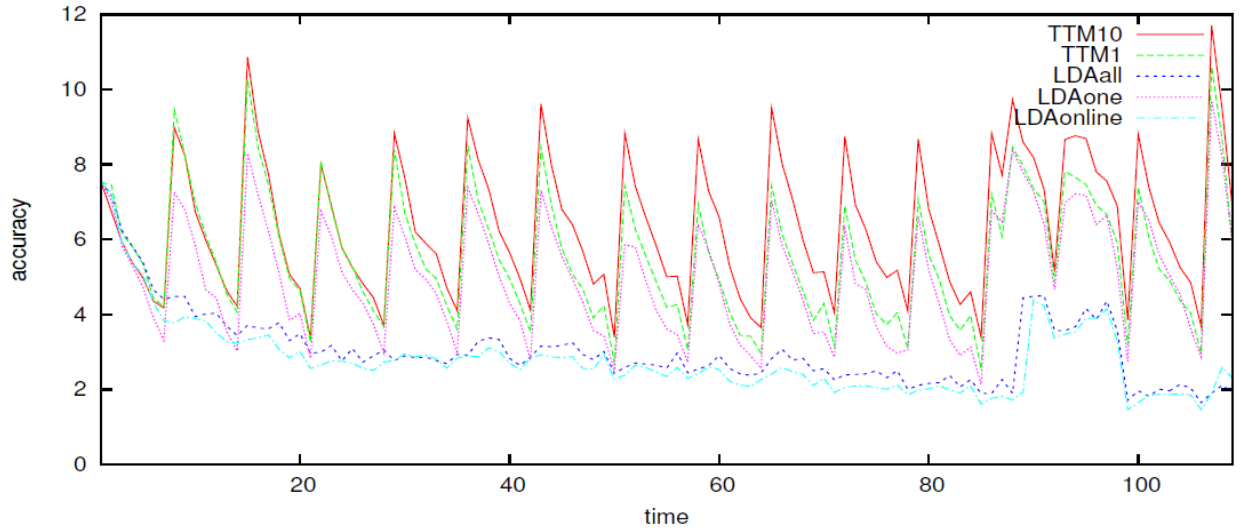
(a) movie

N	LDAall	LDAonline	LDAone	TTM1	TTM10
1	1.21 (0.61)	1.08 (0.54)	1.91 (0.78)	2.22 (0.91)	**2.46** (0.92)
2	2.18 (0.79)	2.00 (0.78)	3.52 (1.22)	3.99 (1.33)	**4.47** (1.36)
3	3.06 (1.04)	2.81 (1.02)	5.04 (1.64)	5.60 (1.75)	**6.35** (1.85)
4	3.90 (1.27)	3.56 (1.24)	6.24 (1.90)	6.82 (2.01)	**7.82** (2.15)
5	4.70 (1.51)	4.26 (1.44)	7.37 (2.20)	7.92 (2.26)	**9.20** (2.42)

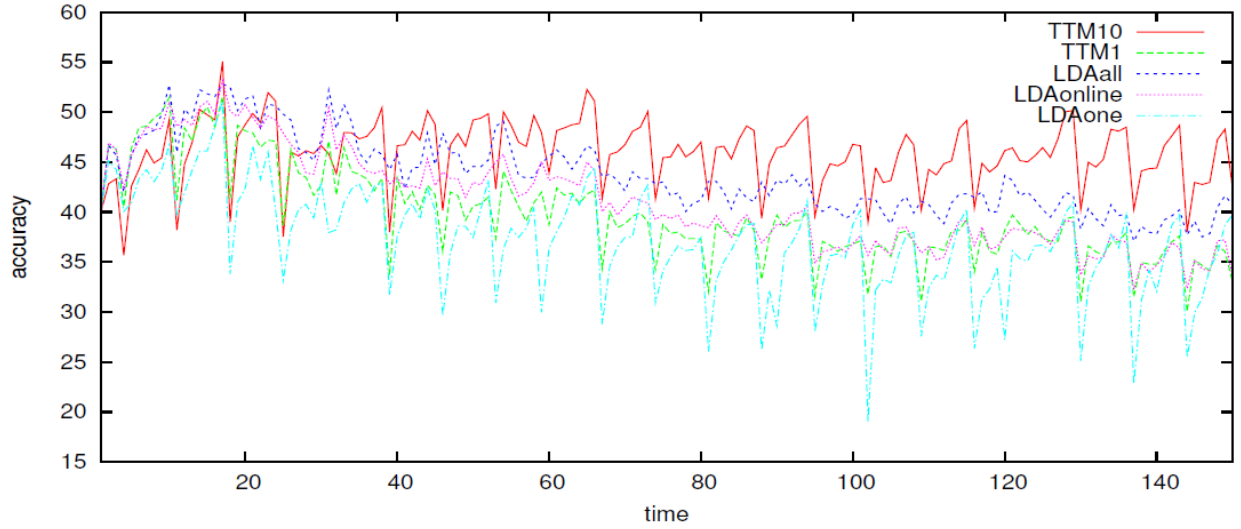
(b) cartoon

N	LDAall	LDAonline	LDAone	TTM1	TTM10
1	27.0 (3.3)	26.0 (3.5)	24.8 (4.5)	26.8 (4.2)	**30.5** (3.4)
2	37.3 (3.6)	35.1 (4.2)	32.4 (4.9)	34.2 (4.5)	**39.9** (3.5)
3	43.7 (3.9)	41.1 (4.8)	37.2 (5.3)	39.8 (4.6)	**45.9** (3.3)
4	48.5 (4.0)	45.8 (5.1)	40.9 (5.3)	44.5 (4.6)	**50.6** (3.2)
5	52.4 (4.2)	49.6 (5.4)	44.1 (5.4)	48.5 (4.6)	**54.4** (3.0)

[Iwata, 2009]



(a) movie



(b) cartoon

[Iwata, 2009]

Figure 3: Three-best accuracies (%) for each day.

まとめ: Topic Tracking model

- ユーザ(文書)ごとのトピック分布、トピックの単語分布を時間発展させたトピックモデル

- Dirichletで時間遷移を表現したことで、非常に簡単に解を導出できます

その他の時系列データ応用

- Wang and McCallum, “Topics over Time: A Non-Markov Continuous-Time model of Topical Trends”, in Proc. KDD, 2006.
- Iwata et al., “Sequential Modeling of Topic Dynamics with Multiple Timescales”, ACM Trans. on Knowledge Discovery from Data. Vol. 5(4). pp. 19:1-19:27, 2012.
- Pruteanu-Malinici, et al., “Hierarchical Bayesian Modeling of Topics in Time-Stamped Documents”, IEEE Trans. PAMI, Vol. 32(6), pp.996-1011, 2010.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Blei & Lafferty, 2006], Blei and Lafferty, “Dynamic Topic Models”, in Proc. ICML, 2006.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.
- [北川, 2005] 北川, “時系列解析入門”, 岩波書店, 2005.
- [Kalman, 1960] Kalman, “A New Approach to Linear Filtering and Prediction Problems”, Journal of Basic Engineering, 1960.
- [Iwata, 2009] Iwata et al, “Topic Tracking Model for Analyzing Consumer Purchase Behavior”, Proc. in IJCAI, 2009.
- [Jin, 2004] Jin et al, “Web Usage Mining based on Probabilistic Latent Semantic Anlysis”, Proc. in KDD, 2004.

NTT

トピックモデルの応用：教師情報・補助情報つきモデル

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

- いわゆる文書データ以外の補助情報・クラス情報が得られる場合のトピックモデル活用法の例です

教師なし学習 (unsupervised learning)

- 「正解」信号となる情報がない設定でモデルを学習したりすることです
- LDA(トピックモデル)は一般に教師なし学習のフレームワークで使われます
  - 文書データだけが与えらえた状態で、まったく未知のトピックを学習しています
- 教師なし学習は基本的に難しいので、高い精度を出すLDAは重宝されます

しかし、教師情報・補助情報があるなら使えばいいのです

- 全てをLDAで、つまり教師なし学習でまとめる必要はありません

- 教師信号・補助情報があるならば、モデル全体の「部品」としてトピックモデルを利用すれば十分です

Supervised LDA
[Blei & McAuliffe, 2008]

Blei and McAuliffe,
"Supervised Topic Models",
in Advances in Neural Information Processing Systems 20
(Proc. NIPS), 2008.

補助情報が観測可能な文書データ

- 典型的には各種レビュー記事を想像すると分かり易いと思います
  - その他、文章のsentiment analysis, 学生のレポート採点データ、・・・
- 当然、文書をモデル化する「だけ」がお仕事のLDAでは表現できません

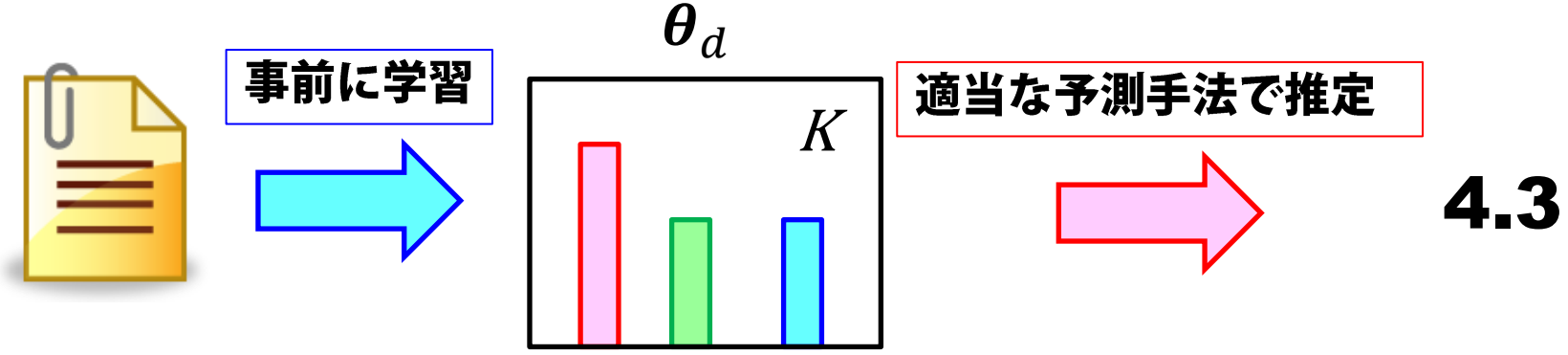
文書と単語の山

補助情報・評価情報・…

4.3

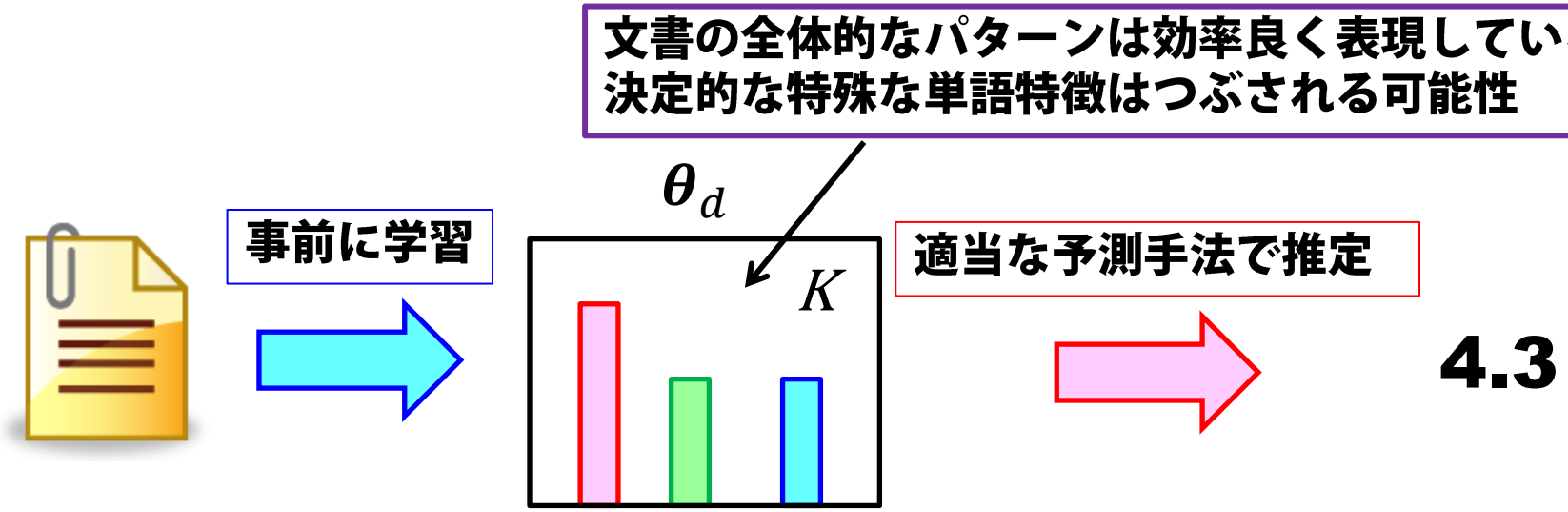
「２段構え」ではダメですか？

- LDAによるトピック学習ののち、各文書のトピックで補助情報を回帰・予測 [Blei, 2003]
- 期待: トピックは文書の中身を上手く圧縮表現しているから、上手く行くはず！ ☺
- 実際: 上手くいかない ☹



理由： 識別に必要な情報がつぶされる

- LDAを含む次元削減手法は、データの全体的な分布を効率よく表現するために、小さい情報量を削除する
- 評価値の予測に必要な特徴量がごく一部だと、「小さい方」に入って削除される可能性



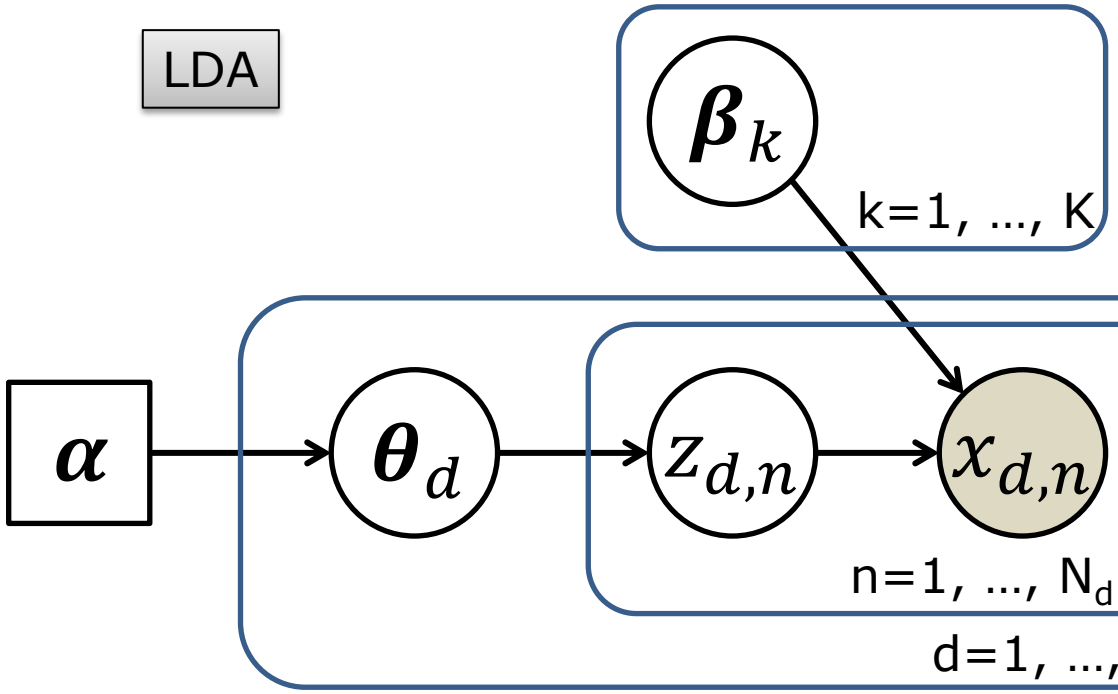
提案法: Supervised LDA

- 文書データそれぞれに対して、連続値の補助情報（教師情報）が付与されているモデル

- 補助情報の影響も考慮して、トピックを学習すると同時に補助情報の分布を学習

- ☺ 教師無し学習のモデルであるLDAを、教師情報有りデータへも応用する道を開いた重要な論文

LDA



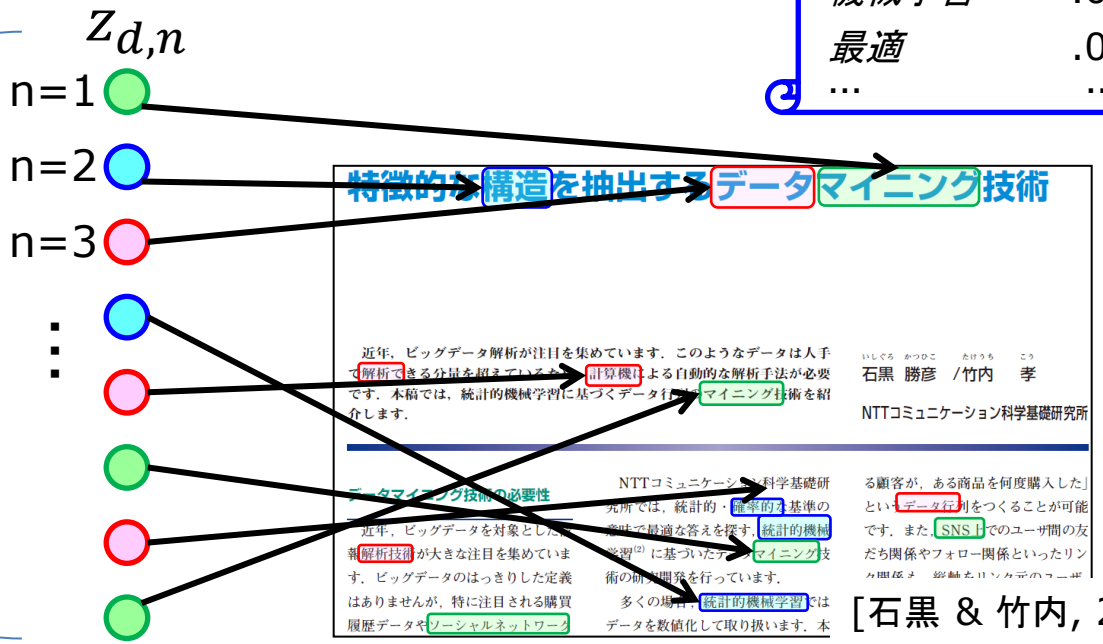
データ	.05
解析	.04
計算機	.03
…	…

$\boldsymbol{\beta}_k$

リンク	.04
ソーシャル	.02
マイニング	.01
…	…

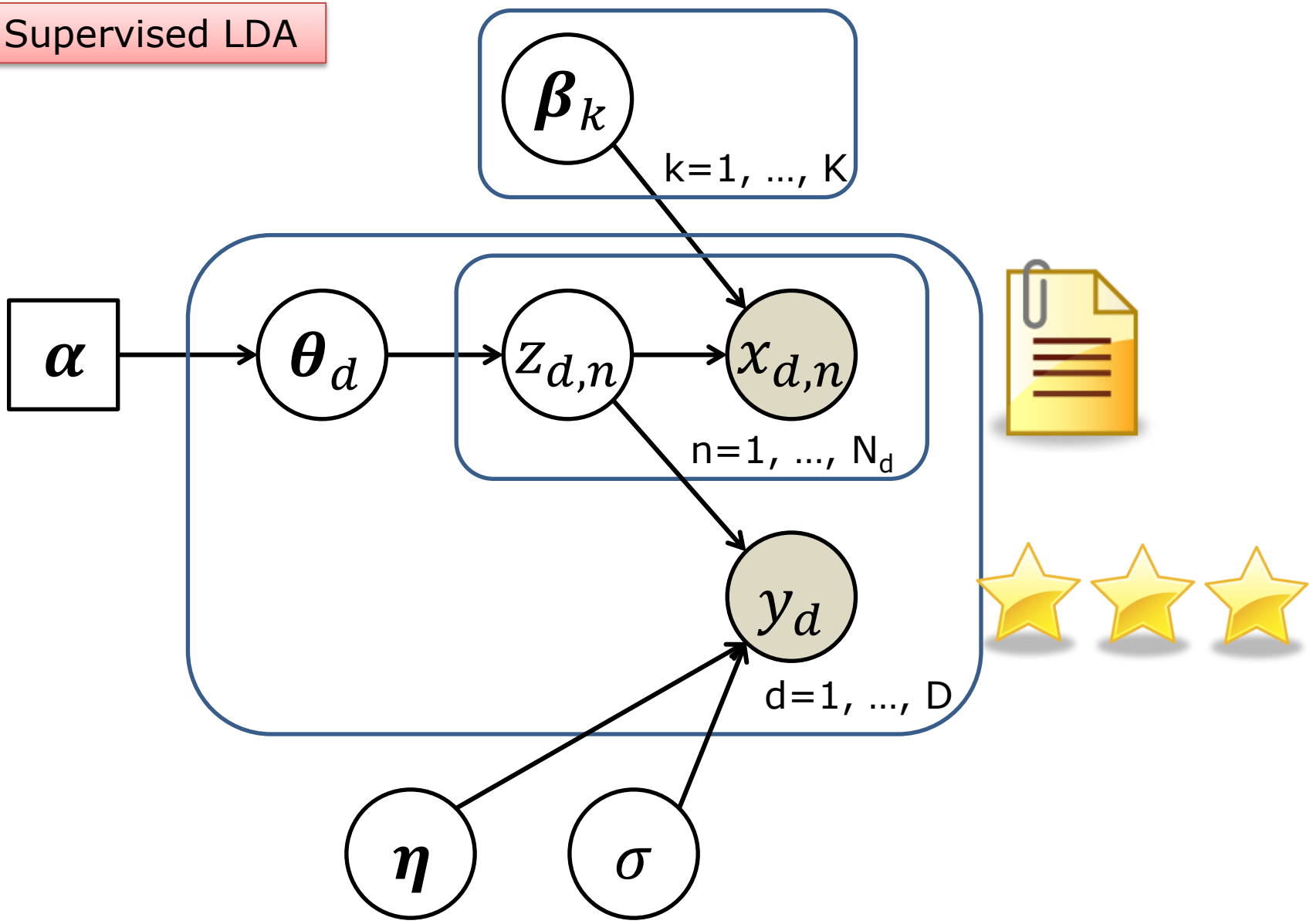
構造	.04
機械学習	.03
最適	.01
…	…

$\boldsymbol{\theta}_d$



[石黒 & 竹内, 2012]

Supervised LDA



生成モデル

for 法案d = 1, 2, …, D_t

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-word assignment

$z_{d,n}|\boldsymbol{\theta}_d \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$

word observation

$x_{d,n}|z_{d,n}, \{\boldsymbol{\beta}_k\} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$

response variable $y_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}, \sigma \sim \mathrm{N}(\boldsymbol{\eta}^T \bar{\boldsymbol{z}}_d, \sigma^2)$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k$

提案法のポイント

- 観測量として、文書データ X と補助情報 Y があり、対等な関係になっています
- 推論時にはXとYの分布を同時に満足するようにZを学習するため、LDAからの２段構えよりも良いモデルが期待できます

LDA

$p(Z|X) \longrightarrow Y$

supervised LDA

$p(Z|X,Y)$

経験トピックに基づく補助情報の回帰

- 文書ごとの経験トピック分布（平均値）でパラメータを生成します
  - topic proportion θ_d を使わない理由は不明 ☹
- 正規分布を使っているので、補助情報は連続値と仮定しています

$$\bar{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathbf{z}_{d,n}$$

$z_{d,n}$をK次元ベクトルとして見ています

$$y_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}, \sigma \sim \mathrm{N}\left(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \sigma^2\right)$$

離散値への対応

- なお、正規分布を一般化線形モデルに変更することで離散値の補助情報にも対応できる･･･そうです

- 離散値への対応は別の手法(例えば[Lacoste-Julien, 2009] など)も参考になさった方が良いと思います

隠れ変数・パラメータの推定

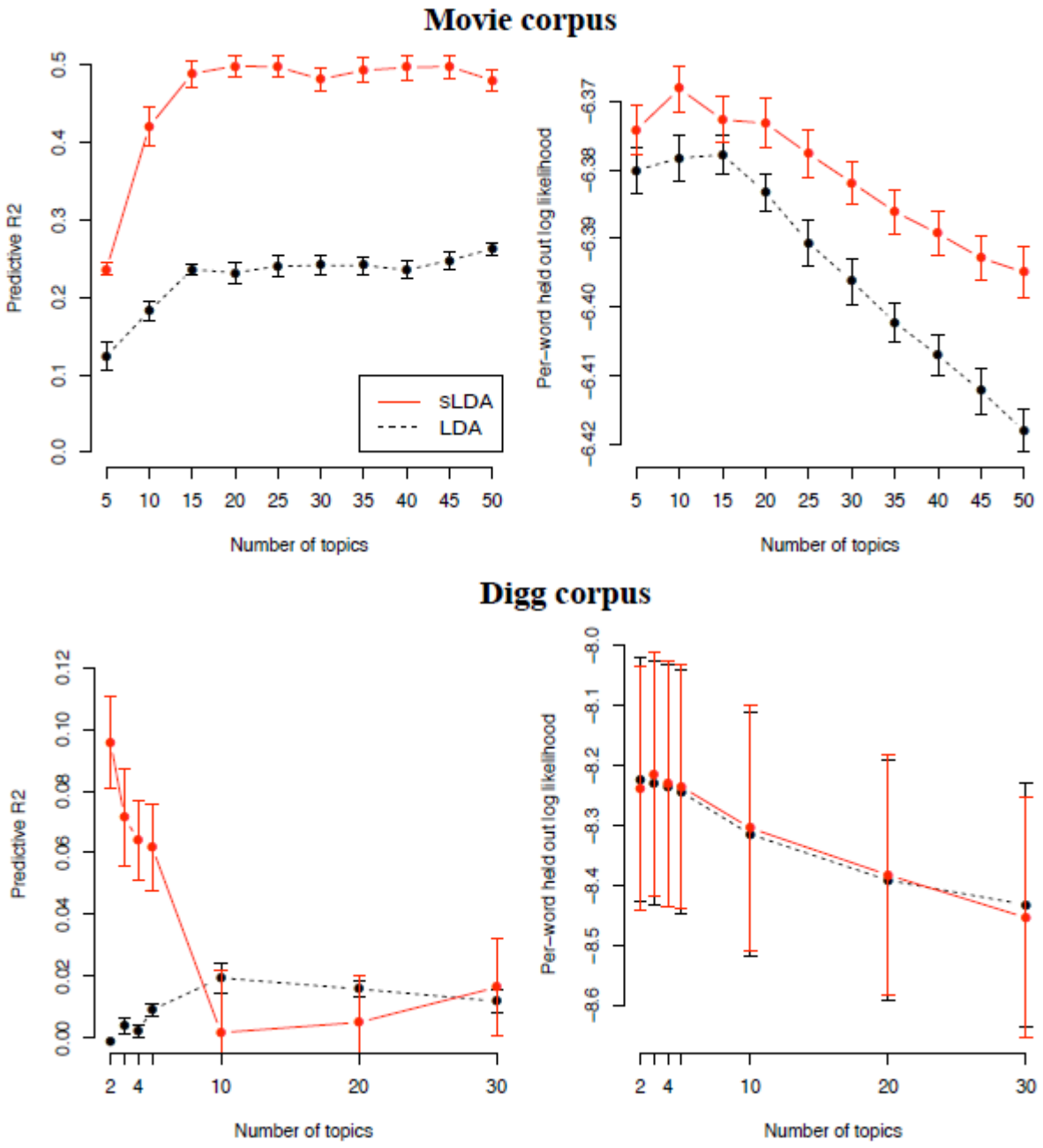
- 任意の手法で推定してかまいませんが、変分ベイズ法をお勧め

$$\log p(\boldsymbol{\theta}, \boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2) \geq H(q)$$

$$+\sum_d E_q[\log p(\boldsymbol{\theta}_d|\boldsymbol{\alpha})] + \sum_{d,n} \mathrm{E}_q\left[\log p\left(z_{d,n}|\boldsymbol{\theta}_d\right)\right]$$

$$+\sum_d E_q[\log p(y_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}, \sigma^2)] + \sum_{d,n} \mathrm{E}_q\left[\log p\left(x_{d,n}|z_{d.n}, \boldsymbol{\beta}\right)\right]$$

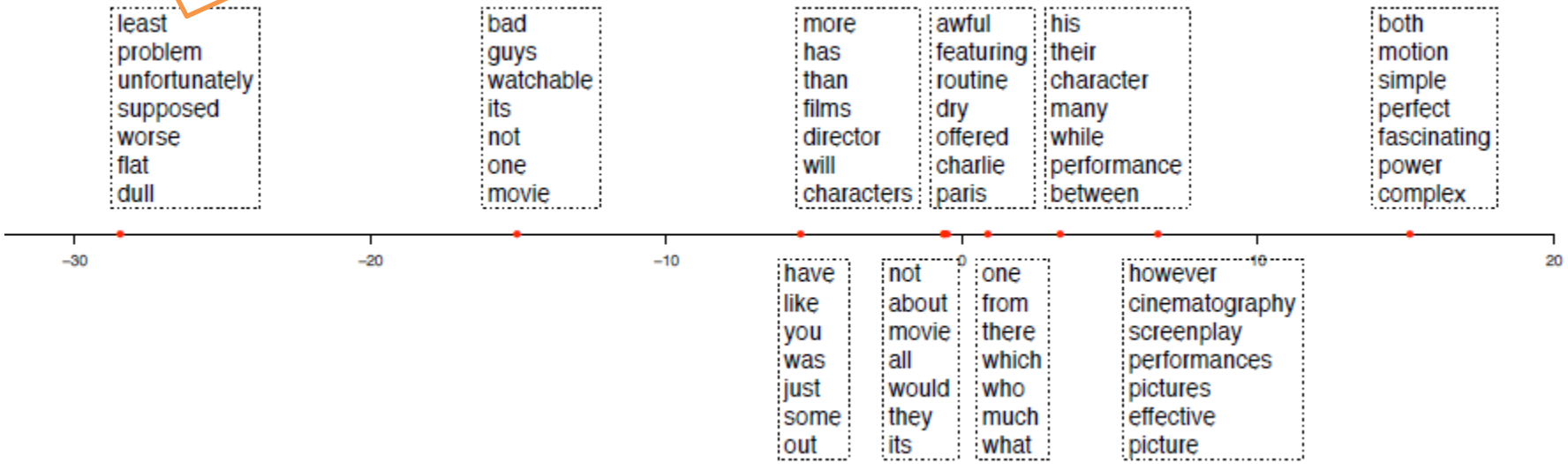
ポイントの部分で述べたように、XとYの両方が効きます



[Blei & McAuliffe, 2008]

η_kの期待値が負になるトピックの頻出単語

η_kの期待値が大きなトピックの頻出単語



$$y_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}, \sigma \sim \mathrm{N}\left(\boldsymbol{\eta}^T \bar{\mathbf{z}}_d, \sigma^2\right)$$

[Blei & McAuliffe, 2008]

まとめ: Supervised LDA

- 教師情報（観測可能な補助情報）有りのトピックモデルの先駆けです
- 非常に多くのモデルの基礎となっている、重要な拡張LDAモデルです

Ideal Point Topic Model
[Gerrish & Blei, 2011]

Gerrish and Blei,
"Predicting Legislative Roll Calls from Text",
in Proc. ICML, 2011.

数理モデルによる政治解析

- Quantitative political science: 政治系の話題においても数理手法を用いた解析・研究がおこなわれています

先日のアメリカ大統領選挙

"New York Timesの選挙予測専門家、ネイト・シルバーは昨夜、大統領選の勝敗を全50州で的中させた。その一方で、いわゆる政治専門家たちの予想はほとんどが外れた。…(中略)…シルバーは今回も彼の作った数理的予測モデルが古臭い専門家の勘や生半可な統計に基づく推測より圧倒的に優れていたことを証明した。"

G. Ferenstein, TechCrunch Japan, Nov 8 2012.
(http://jp.techcrunch.com/archives/20121107pundit-forecasts-all-wrong-silver-perfectly-right-is-punditry-dead/)

議員の投票行動モデル: ideal point model [Clinton, 2004]

- 議員の政治的信条と法案をそれぞれ潜在空間に射影、お互いの位置関係で賛成確率をモデル化

- ☹ 法案の中身（文章データ）を利用してない
- ☹ 新規法案に対する投票予測が不可能

提案法: Ideal point topic model

- 投票行動のみを扱っていたIdeal point modelを拡張

- 各法案の内容をトピックモデルでモデル化

- ☺ 文書のトピック・言葉づかいと投票行動の関係を調査できる
- ☺ 新規法案でもトピック(文書内容)から各議員の投票行動を予測できる

対象データ

- 1997-2011年のU.S. Congress Roll call data
  - 各法案(bill, resolution)はn-gramを”単語”とするBow表現
  - 各議員(legislator)は0/1のvotesを持つ

法案 dのBoW表現 $x_{d,n}$　**法案 dに対する議員 uの投票結果 $v_{d,u}$**

法案　d

u = 1: yes(1)　u = 2: no(0)　u = 3: no(0)

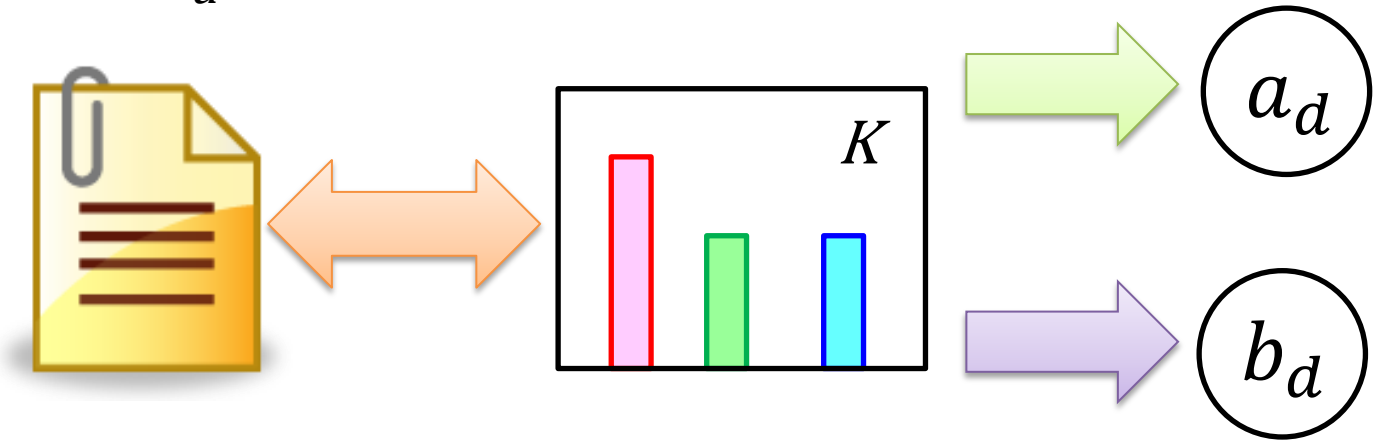
Ideal topic model: 法案の投票傾向を２つのパラメータで表現

- a_d bill difficulty: 誰でも賛成(反対)するような法案だと正の(負の)大きい値をとる
- b_d bill discriminaty: 法案の"立ち位置"
- y_u: 議員の理想とする"立ち位置"

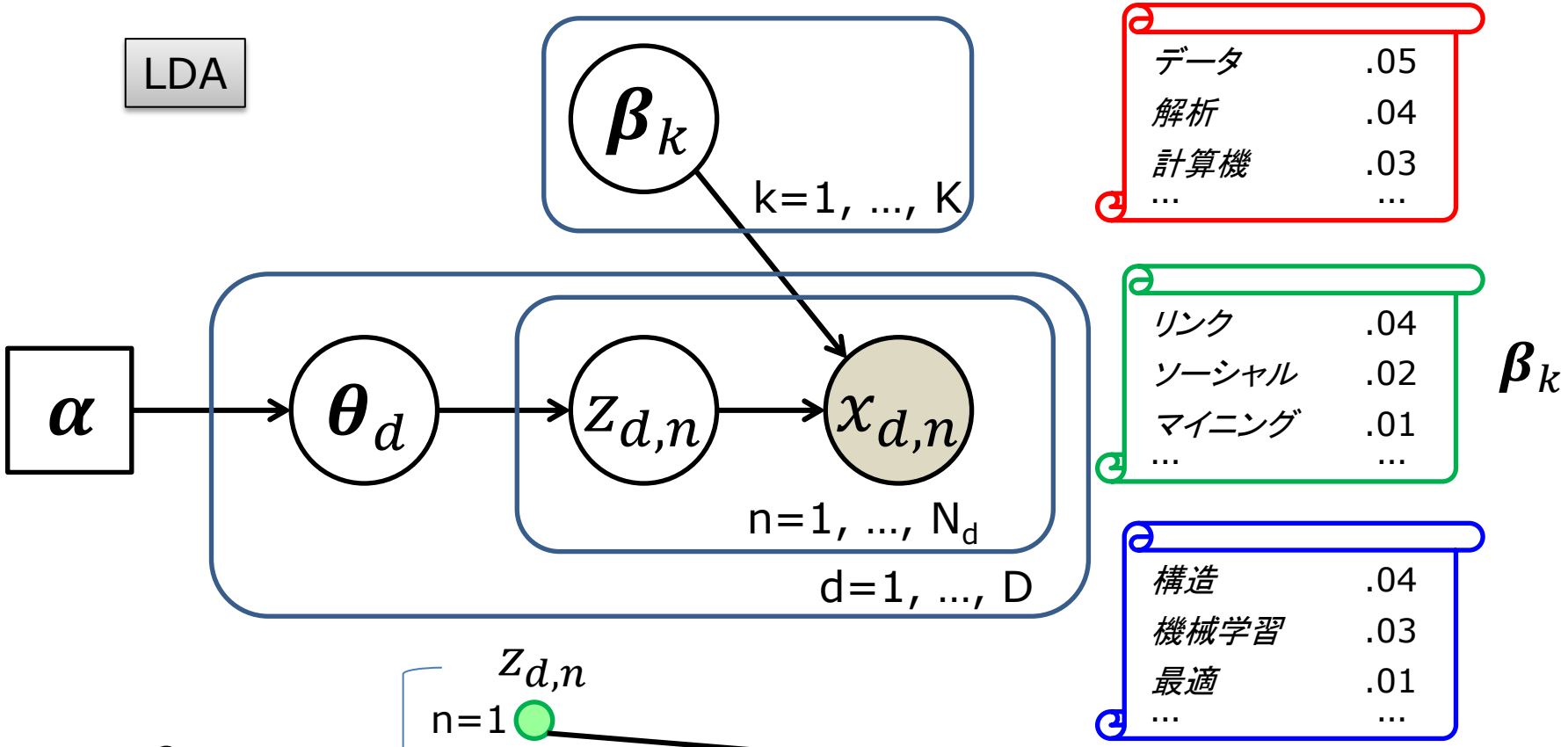
$$p(v_{du} = 1) = \sigma(a_d + b_d y_u) \qquad \sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$$

提案法のアイデア: 文書トピックによるideal point modelの制御

- 法案のパラメータは、内容＝文書のトピックによって制御されると考える
  - 例1) WBC優勝を褒め称える議案は皆賛成するのでa_dが大きな正の値
  - 例2) 予算関係は起案者の意見が色濃くでるのでb_dが特徴的な値をとる



LDA

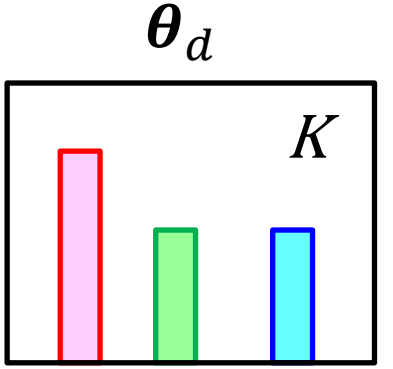


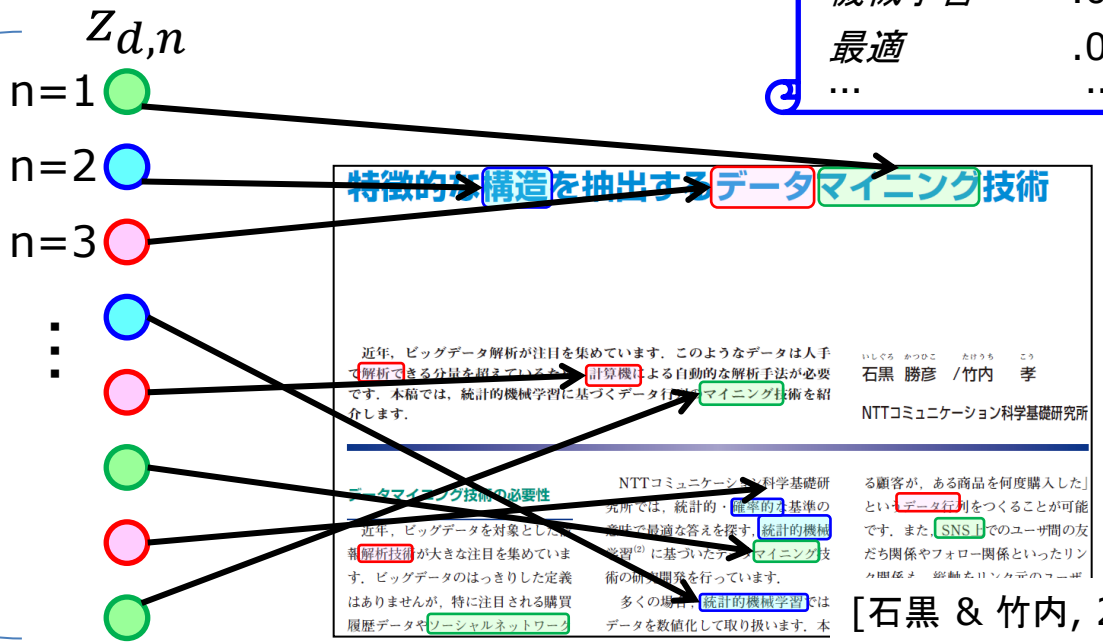
データ	.05
解析	.04
計算機	.03
…	…

リンク	.04
ソーシャル	.02
マイニング	.01
…	…

$\boldsymbol{\beta}_k$

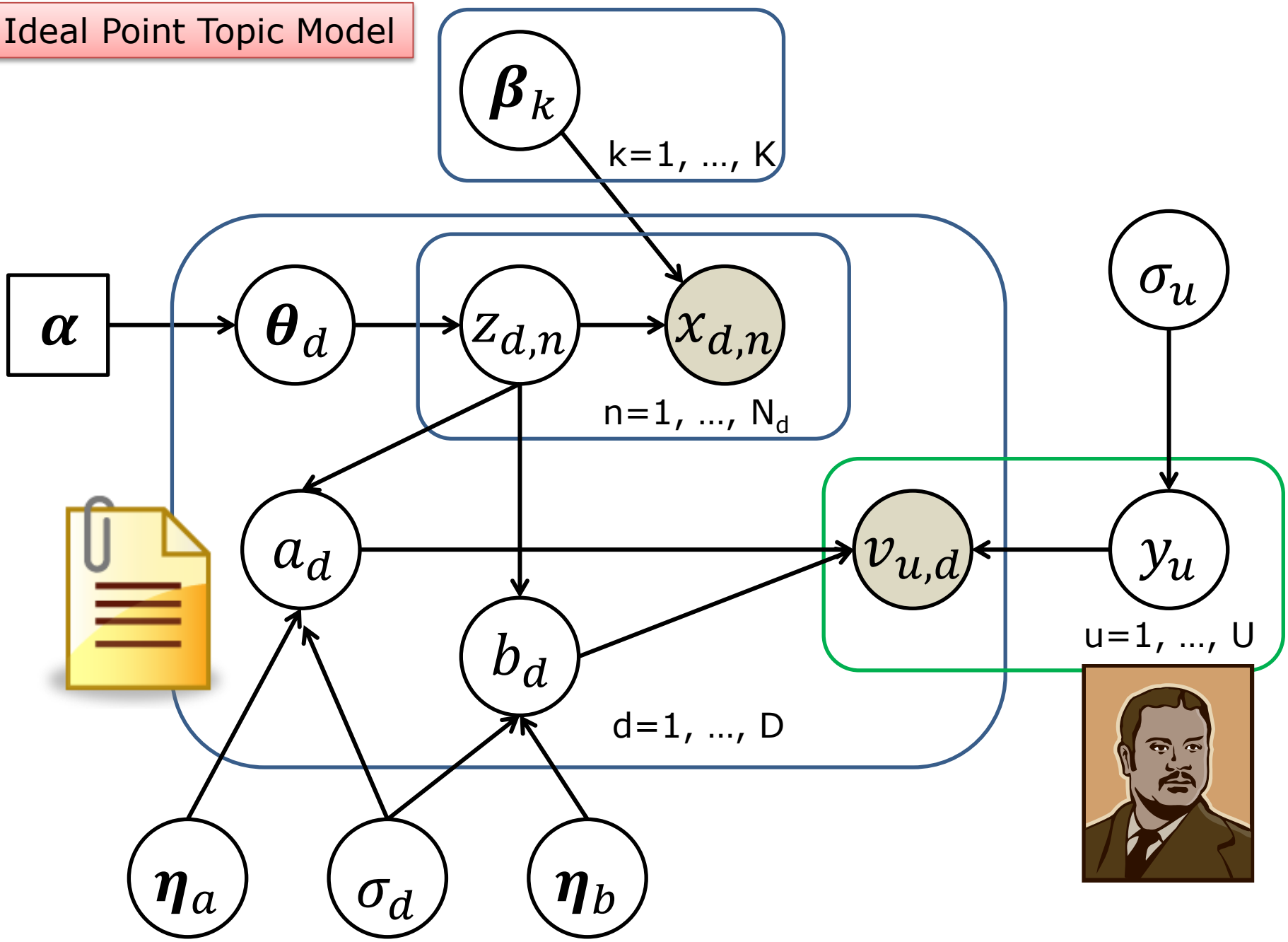
構造	.04
機械学習	.03
最適	.01
…	…





[石黒 & 竹内, 2012]

Ideal Point Topic Model



生成モデル

for 法案d = 1, 2, …, D_t

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha} \sim \mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-word assignment

$z_{d,n}|\boldsymbol{\theta}_d \sim \mathrm{Mult}(\boldsymbol{\theta}_d)$

word observation

$x_{d,n}|z_{d,n}, \{\boldsymbol{\beta}_k\} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k$

生成モデル

for 議員 u = 1, 2, …, U

legislator ideal point $y_u|\sigma_u \sim \boldsymbol{N}(0, \sigma_u)$

for 法案 d = 1, 2, …, D_t

bill difficulty param. $a_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}_a, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_a{}^T \bar{\boldsymbol{z}}_d, \sigma_d{}^{\boldsymbol{2}}\right)$

bill discrimination param. $b_d|\bar{\boldsymbol{z}}_d, \boldsymbol{\eta}_b, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_b{}^T \bar{\boldsymbol{z}}_d, \sigma_d{}^{\boldsymbol{2}}\right)$

for 議員 u = 1, 2, …, U

vote observation

$$p\left(v_{d,u} = 1|a_d, b_d, y_u\right) = \sigma(a_d + b_d y_u)$$

σはlogistic function $\sigma(x) = \dfrac{\exp(x)}{1 + \exp(x)}$

経験トピックに基づく Ideal pointパラメータモデリング

- 実際には θ_d ではなく、経験トピック分布でパラメータを生成します

$$\bar{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d} \mathbf{z}_{d,n}$$

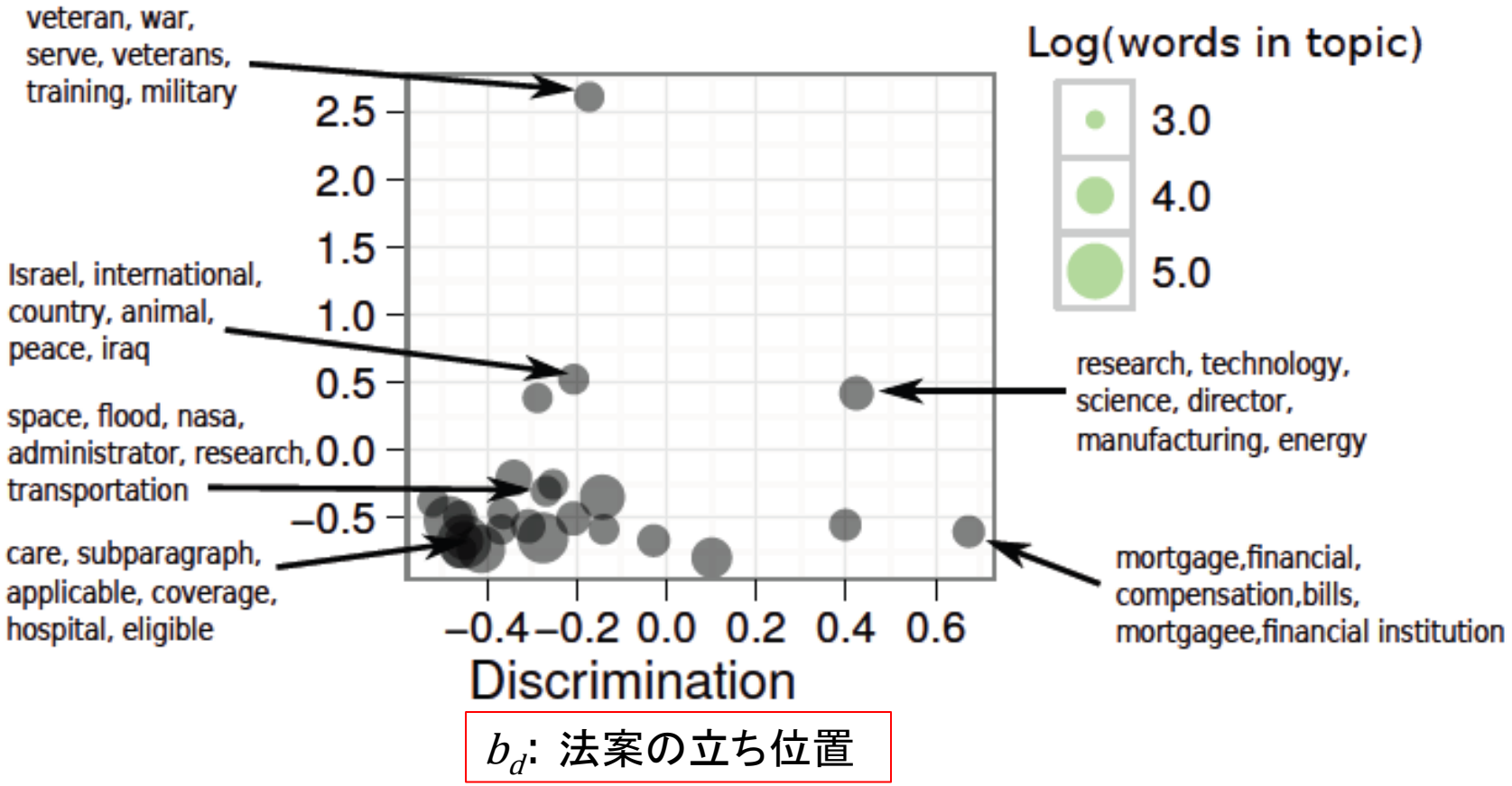
$z_{d,n}$をK次元ベクトルとして見ています

bill difficulty $a_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}_a, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_a{}^T \bar{\mathbf{z}}_d, \sigma_d{}^2\right)$

bill discriminarity $b_d|\bar{\mathbf{z}}_d, \boldsymbol{\eta}_b, \sigma_d \sim \mathrm{N}\left(\boldsymbol{\eta}_b{}^T \bar{\mathbf{z}}_d, \sigma_d{}^2\right)$

隠れ変数・パラメータの推定

- 任意の手法で推定してかまいませんが、変分ベイズ法をお勧め
- 解の不定性
  - b ＊ yのせいで、b, yの値の符号は反転可能
  - そこで、共和党・民主党の重鎮だけは極端な正負の値に固定します

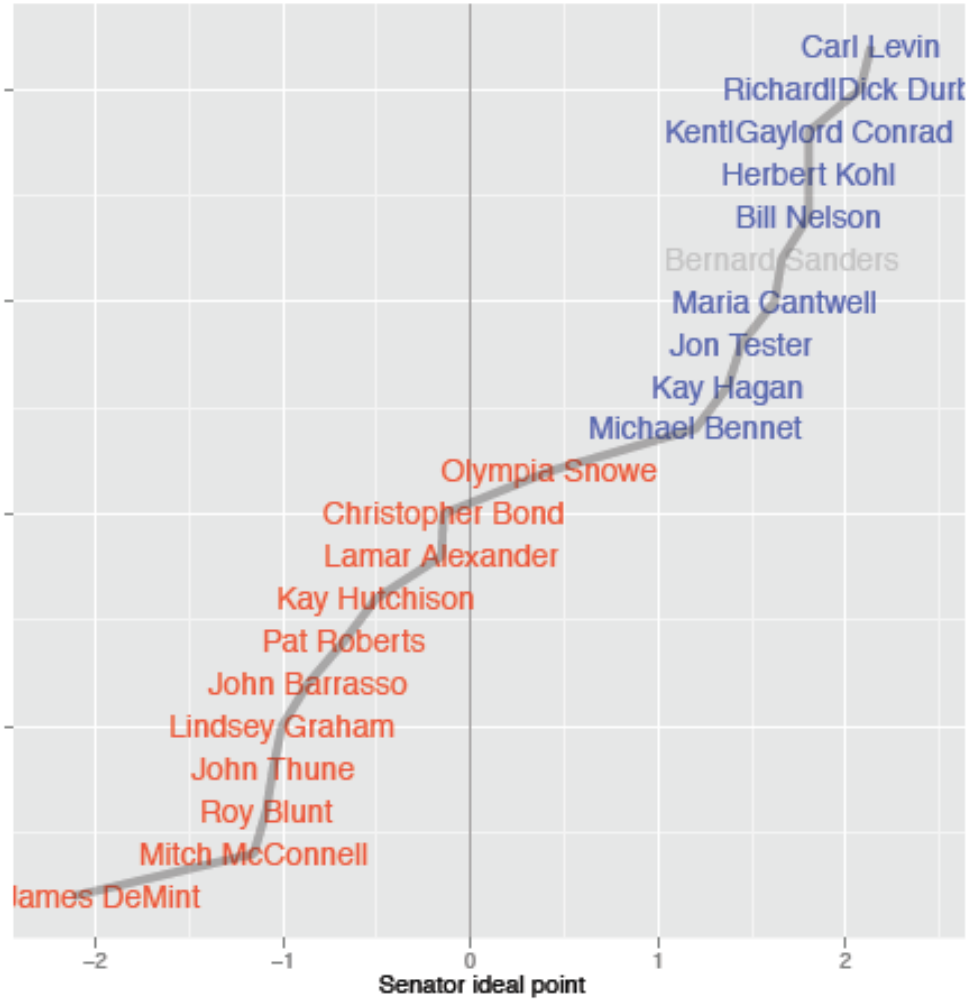


a_d: 法案のバイアス項

b_d: 法案の立ち位置

[Gerrish & Blei, 2011]

$$p(v_{du} = 1) = \sigma(a_d + b_d y_u)$$



$$p(v_{du} = 1) = \sigma(a_d + b_d y_u)$$

[Gerrish & Blei, 2011]

まとめ: Ideal Point Topic model

- 法案に対する投票行動のモデル化です
- 法案の文書内容だけでなく、各法案に対する議員の投票結果が観測できています
- トピックモデルを法案の内容を表す“部品”として、既存のideal point modelと組み合わせています

その他の教師あり・補助情報ありトピックモデル

- Flaherty et al., “A Latent Variable Model for Chemogenomic Profiling”, Bioinformatics, Vol. 21(15), pp.3286-3293, 2005.
- Lacoste-Julien et al., “DiscLDA: Discriminative Learning for Dimensionality Reductioin and Classification”, Advances in Neural Information Processing Systems 21 (Proc. NIPS), 2009.
- Ramage et al., “Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora”, in Proc. EMNLP, 2009.
- Gerrish and Blei, “How They Vote: Issue-Adjusted Models of Legislative Behavior”, in Proc. NIPS, 2012.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Blei & McAuliffe, 2008] Blei and McAuliffe, “Supervised Topic Models”, Advances in Neural Information Processing Systems 20 (Proc. NIPS), 2008.
- [Lacoste-Julien, 2009] Lacoste-Julien et al., “DiscLDA: Discriminative Learning for Dimensionality Reduction and Classification”, Advances in Neural Information Processing Systems 21 (Proc. NIPS), 2009.
- [Gerrish & Blei, 2011] Gerrish and Blei, “Predicting Legislative Roll Calls from Text”, in Proc. ICML, 2011.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.

NTT

トピックモデルの応用：関係データ、ネットワークデータ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

- 文書や著者の間に「関係」のネットワーク（グラフ）が想定されるデータセットが対象です

- お互いの関係をどのようにモデルに取り入れるかがポイントです

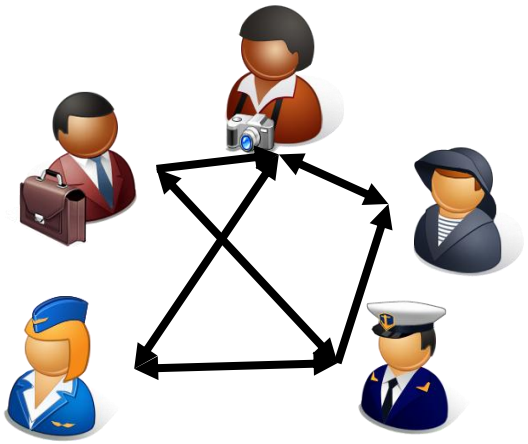
関係データ

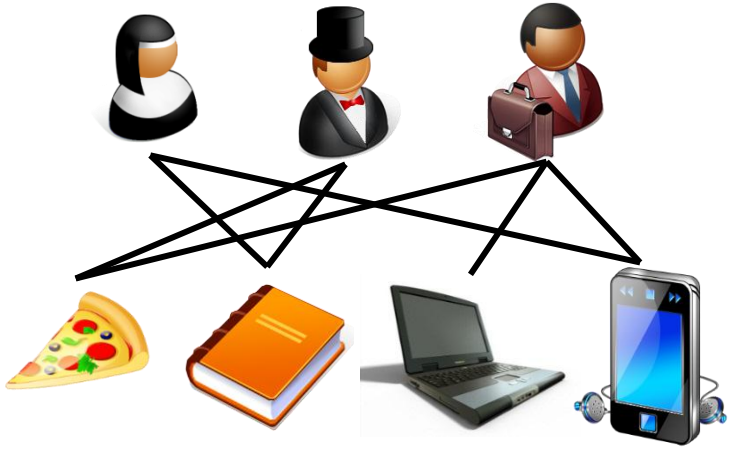
- 複数のオブジェクト（ノード）の間にリンク(エッジ）があってつながっているデータです
- 数学的には、いわゆる「グラフ」です

$$G = (V, E)$$

V(vertex): オブジェクト、ノード

E(edge): リンク、エッジ

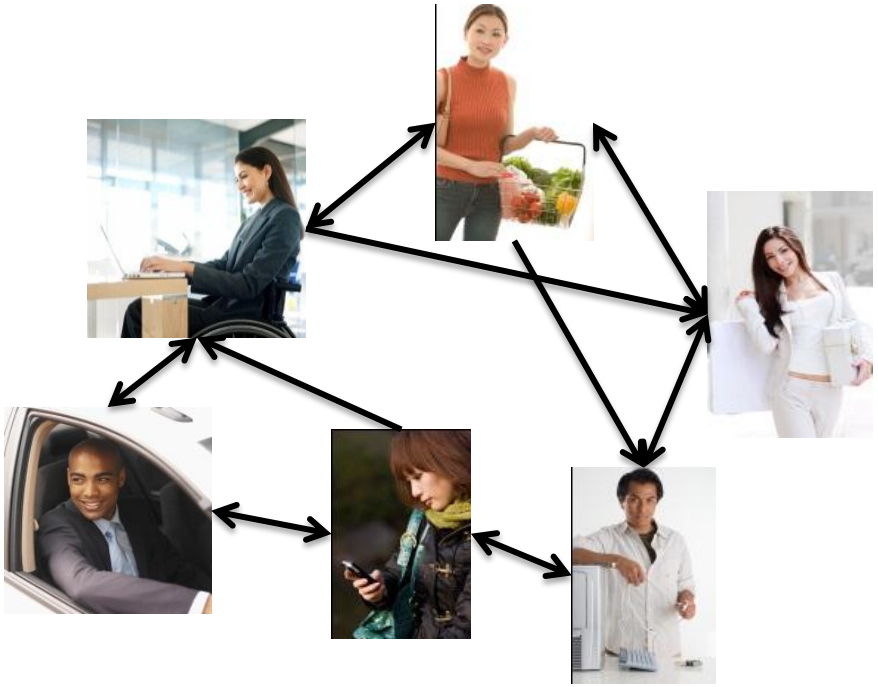




どんな関係データがありますか？

- ソーシャルネットワークサービス(SNS)上の友達関係、フォロー関係

$$G = (V, E) = (\text{ユーザ}, \text{フォロー})$$



SNS内のコミュニティ発見

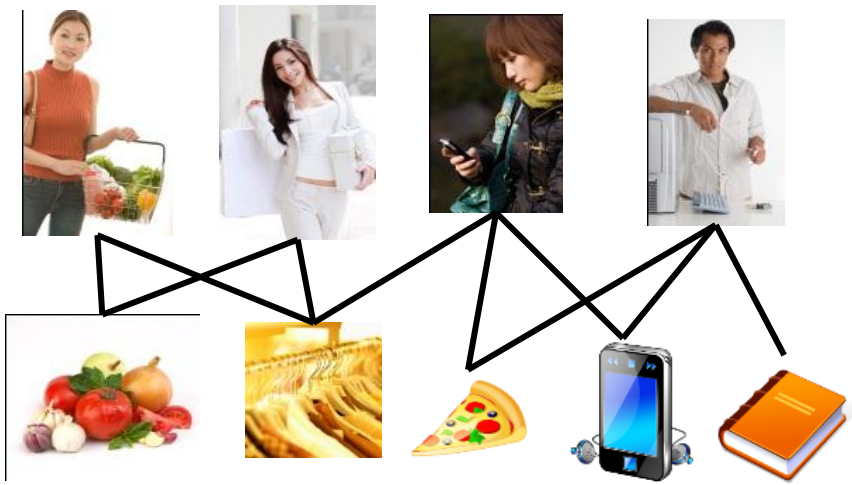
影響力の大きなユーザの発見

口コミ情報の伝搬範囲の最大化

どんな関係データがありますか？

- ネットショッピングなどの購買データ

$$G = (V, E) = (\text{ユーザ} \times \text{商品}, \text{販売実績})$$



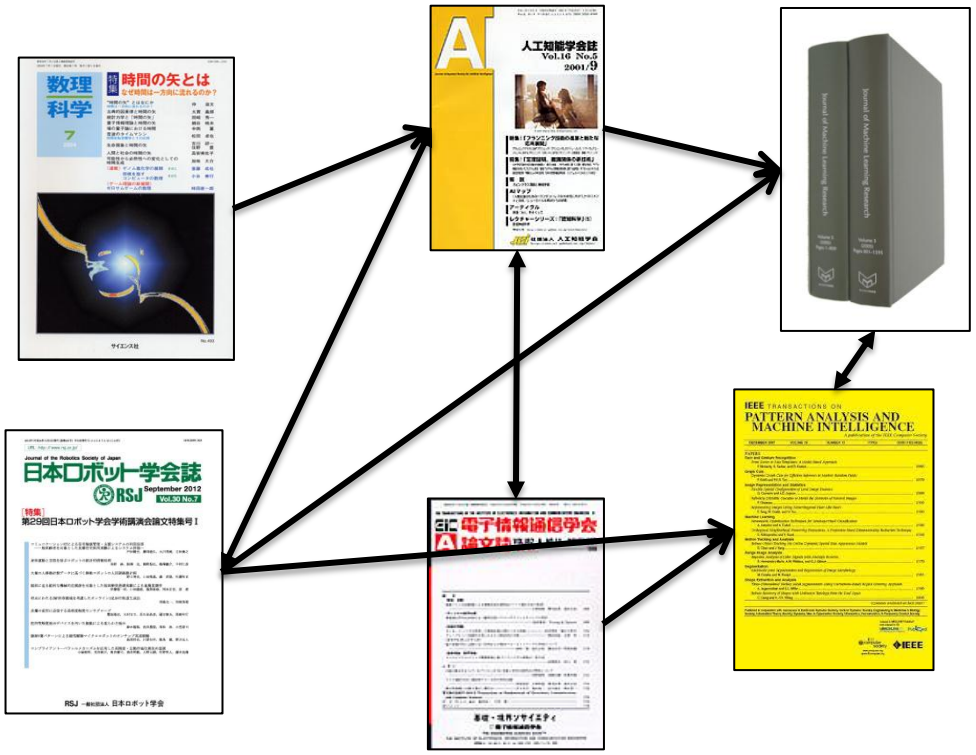
（販売実績に基づく）顧客のセグメント解析

商品のレコメンデーション（協調フィルタリング）

どんな関係データがありますか？

- 特許・技術論文の引用関係

$$G = (V, E) = (\text{特許・論文}, \text{引用・参照})$$



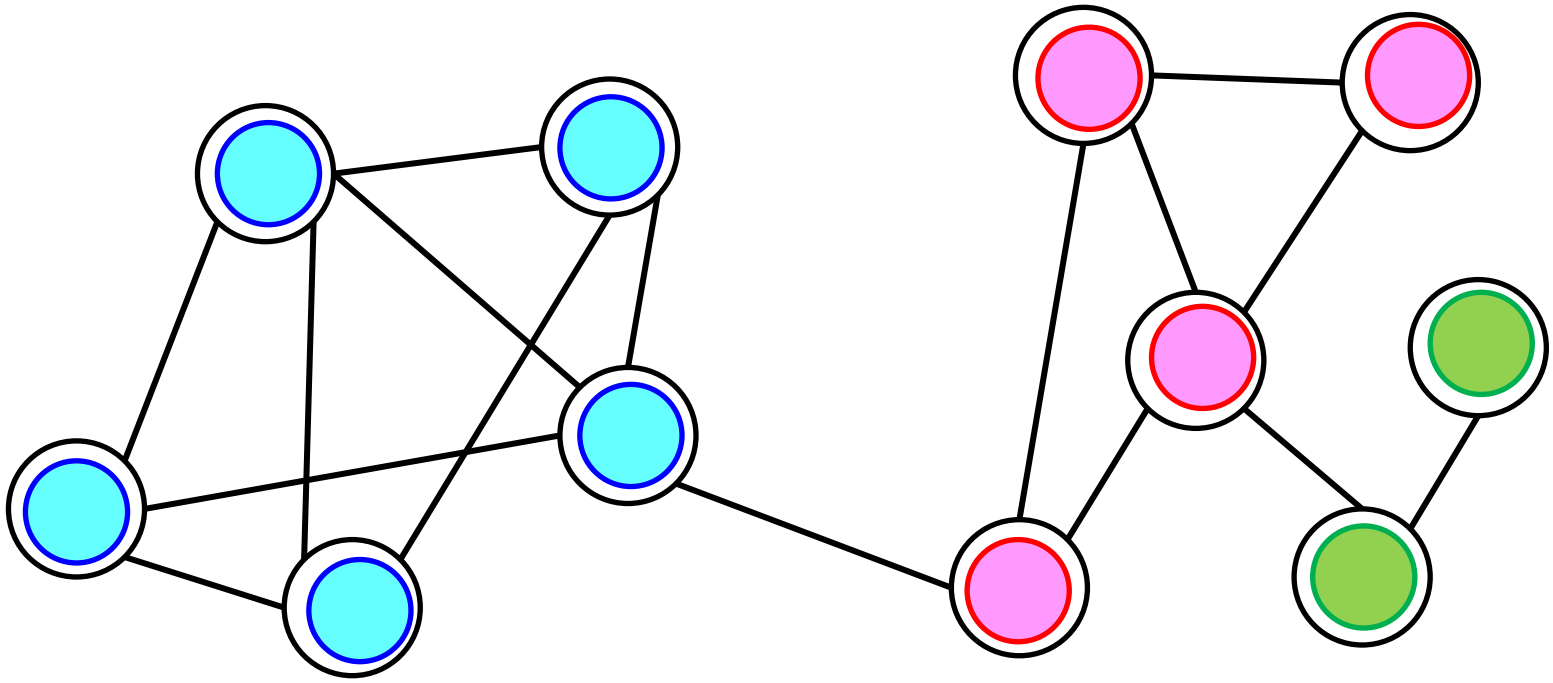
関連特許の検索

関連論文のおすすめ

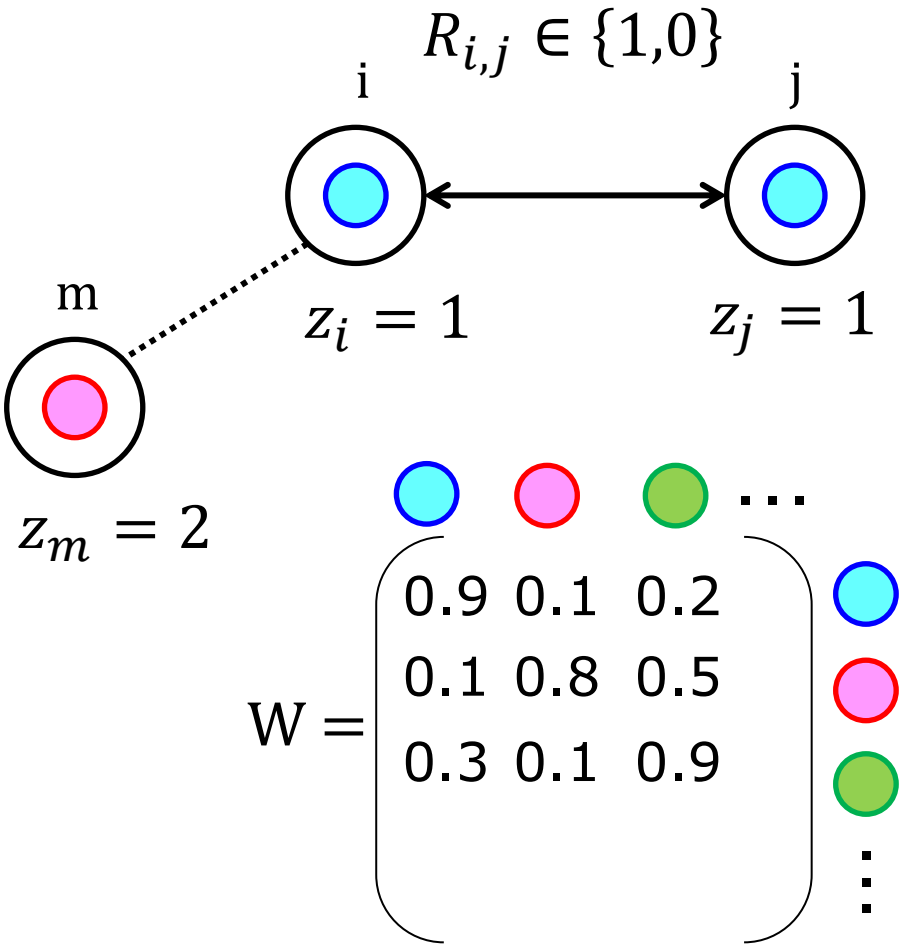
論文・特許の影響力可視化

関係データモデリング手法の例：無限関係モデル(IRM) [Kemp, 2006]

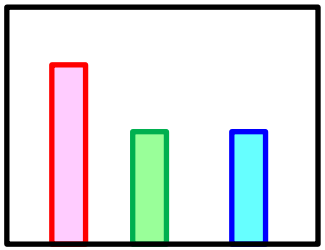
- シンプルで有効性の高い関係データモデル
- グラフのリンク構造から、オブジェクトをクラスタリング（カテゴライズ）してくれます



IRMの生成モデル



$\boldsymbol{\alpha} \sim \mathrm{Stick}(\gamma)$



$z_i = k \sim \mathrm{Mult}(\boldsymbol{\alpha})$

$w_{k,l} \sim \mathrm{Beta}(a, b)$

$R_{i,j} \sim \mathrm{Bernoulli}\left(w_{z_i,z_j}\right)$

→ 1　　⋯→ 0

IRMの問題点：
グラフの構造だけしか使わない

- 各オブジェクトは様々な情報・特徴をもっているはず → 使わない手はない

- ユーザの性別・年齢・プロフィール文
- 商品の値段・成分・キャッチコピー
- 論文（特許）の内容・請求項・キーワード

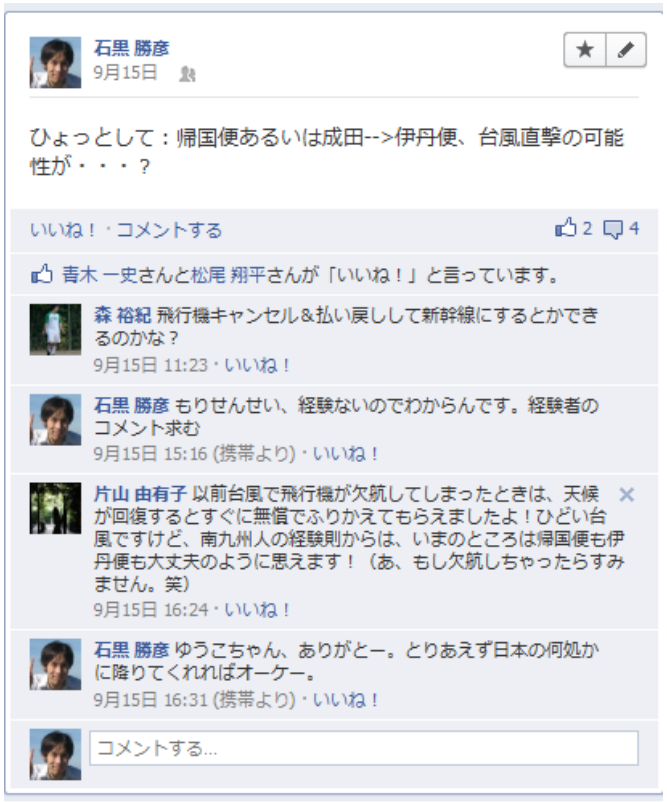
トピックモデル！！

Relational Topic Models
[Chang and Blei, 2009]

Chang and Blei,
"Relational Topic Models for Document Networks",
in Proc. AISTATS, 2009.

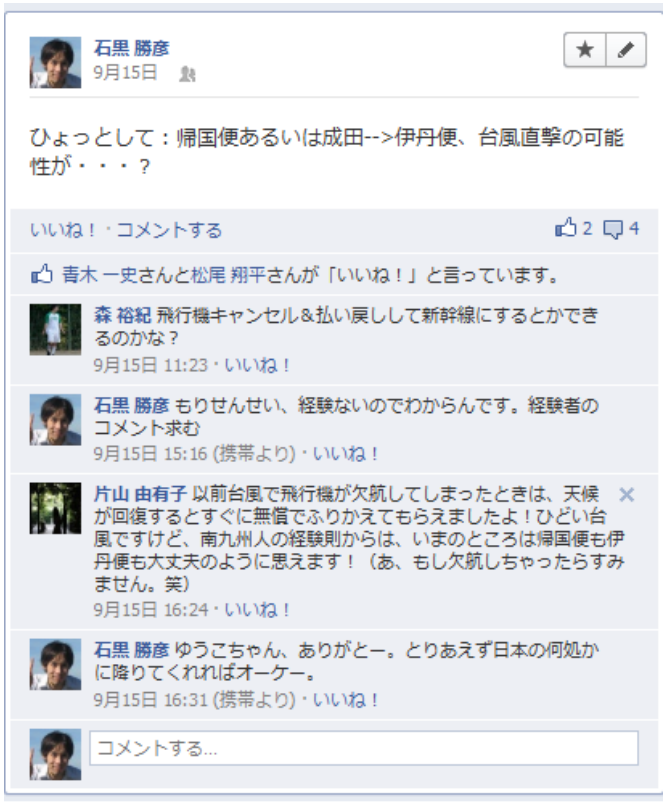
文書をリンクする情報は世の中沢山あります

- SNSでの返信、ブログの引用、特許の関連文献、論文のreference, …



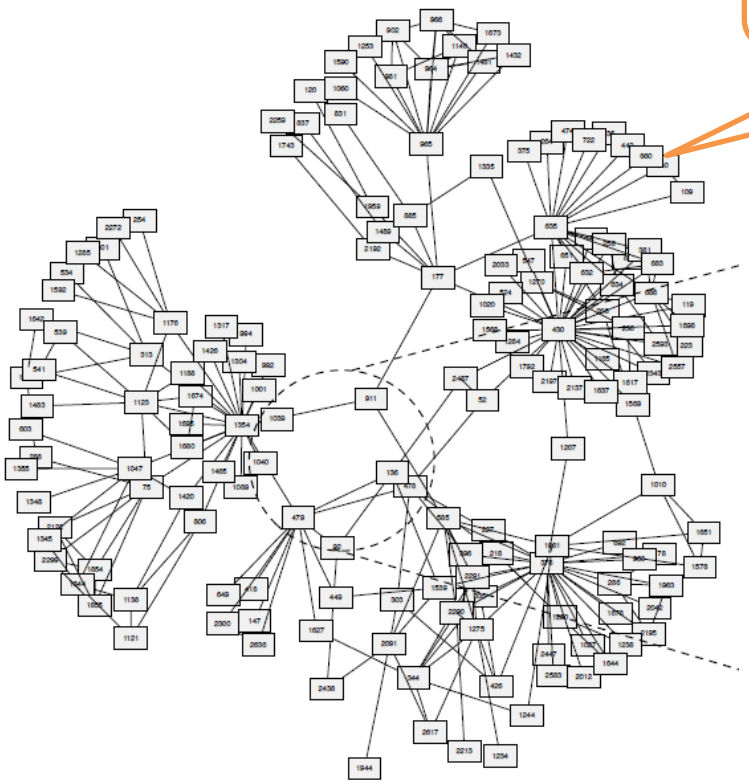
モデル化したくなります

- 関連する論文や、リツイートしたくなるようなつぶやきを自動的に発見できます



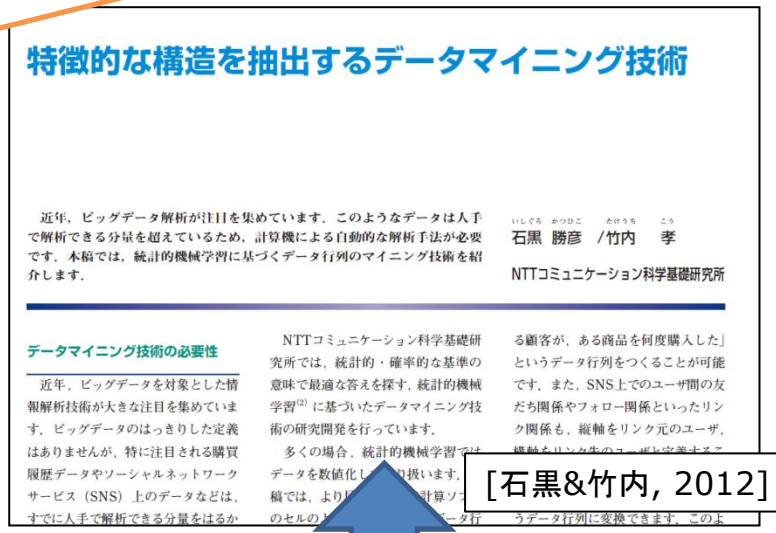
典型的なデータ構造のイメージ

論文引用ネットワーク



[Chang and Blei, 2009]

リンク＝引用した・された
オブジェクト＝文書（論文）：BoW表現



特徴的な構造を抽出するデータマイニング技術

近年，ビッグデータ解析が注目を集めています．このようなデータは人手で解析できる分量を超えているため，計算機による自動的な解析手法が必要です．本稿では，統計的機械学習に基づくデータ行列のマイニング技術を紹介します．

石黒 勝彦 /竹内 孝

NTTコミュニケーション科学基礎研究所

データマイニング技術の必要性

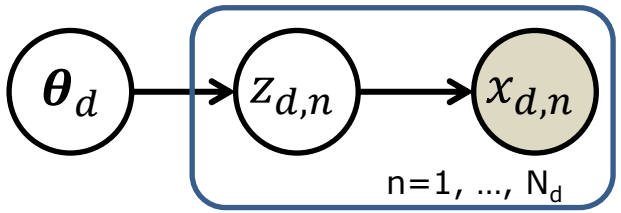
近年，ビッグデータを対象とした情報解析技術が大きな注目を集めています．ビッグデータのはっきりした定義はありませんが，特に注目される購買履歴データやソーシャルネットワークサービス（SNS）上のデータなどは，すでに人手で解析できる分量をはるか

NTTコミュニケーション科学基礎研究所では，統計的・確率的な基準の意味で最適な答えを探す，統計的機械学習[2]に基づいたデータマイニング技術の研究開発を行っています．

多くの場合，統計的機械学習では

る顧客が，ある商品を何度購入した」というデータ行列をつくることが可能です．また，SNS上でのユーザ間の友だち関係やフォロー関係といったリンク関係も，縦軸をリンク元のユーザ，

[石黒&竹内, 2012]



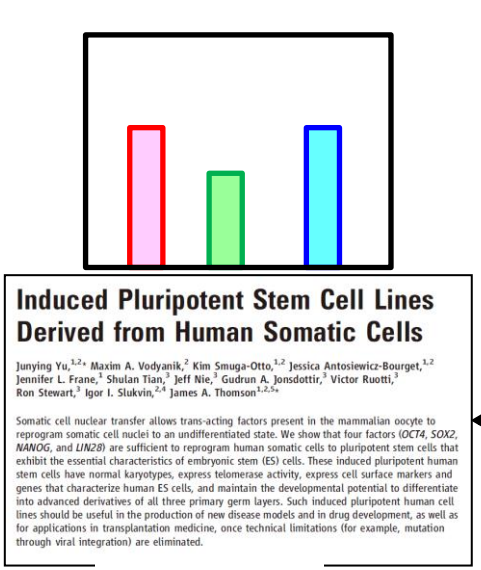
提案法: Relational Topic Model (RTM)

- 「リンク」を活かしたトピックモデル
  - 文書の中身だけでなく、文書間のリンクの生成過程も同時に確率モデル化
  - 具体的には論文や特許データを想定

- 文書のリンク推定: 論文の内容(BoW)から、関連がある論文を発見
- 文書のトピック推定: 特許の引用情報から、自分の特許とのバッティング度合を推定

手法のアイデア

- 内容（トピック）が似ている→引用（リンク）が発生する
- 文書のもつトピック分布の類似度に応じて、文書の間のリンク発生確率が変わる

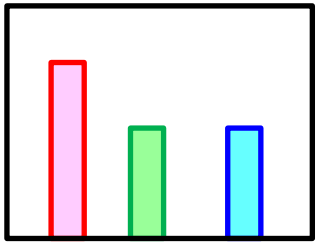


Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells

Junying Yu,[1,2,*] Maxim A. Vodyanik,[2] Kim Smuga-Otto,[1,2] Jessica Antosiewicz-Bourget,[1,2] Jennifer L. Frane,[1] Shulan Tian,[3] Jeff Nie,[3] Gudrun A. Jonsdottir,[3] Victor Ruotti,[3] Ron Stewart,[3] Igor I. Slukvin,[2,3] James A. Thomson[1,2,3,*]

Somatic cell nuclear transfer allows trans-acting factors present in the mammalian oocyte to reprogram somatic cell nuclei to an undifferentiated state. We show that four factors (*OCT4, SOX2, NANOG,* and *LIN28*) are sufficient to reprogram human somatic cells to pluripotent stem cells that exhibit the essential characteristics of embryonic stem (ES) cells. These induced pluripotent human stem cells have normal karyotypes, express telomerase activity, express cell surface markers and genes that characterize human ES cells, and maintain the developmental potential to differentiate into advanced derivatives of all three primary germ layers. Such induced pluripotent human cell lines should be useful in the production of new disease models and in drug development, as well as for applications in transplantation medicine, once technical limitations (for example, mutation through viral integration) are eliminated.

[Yu, 2007]



Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors

Kazutoshi Takahashi[1] and Shinya Yamanaka[1,2,*]

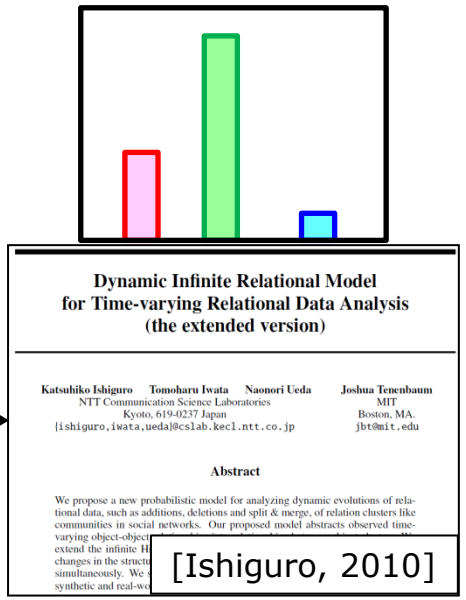
[1] Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan
[2] CREST, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan
*Contact: yamanaka@frontier.kyoto-u.ac.jp
DOI 10.1016/j.cell.2006.07.024

SUMMARY

Differentiated cells can be reprogrammed to an embryonic-like state by transfer of nuclear contents into oocytes or by fusion with embryonic stem (ES) cells. Little is known about factors

or by fusion with ES cells (Cowan et al., 2005; Tada et al., 2001), indicating that unfertilized eggs and ES cells contain factors that can confer totipotency or pluripotency to somatic cells. We hypothesized that the factors that play important roles in the maintenance of ES cell identity also play pivotal roles in the induction of pluripotency in

[Takahashi & Yamanaka, 2006]



Dynamic Infinite Relational Model for Time-varying Relational Data Analysis (the extended version)

Katsuhiko Ishiguro Tomoharu Iwata Naonori Ueda
NTT Communication Science Laboratories
Kyoto, 619-0237 Japan
{ishiguro,iwata,ueda}@cslab.kecl.ntt.co.jp

Joshua Tenenbaum
MIT
Boston, MA.
jbt@mit.edu

Abstract

We propose a new probabilistic model for analyzing dynamic evolutions of relational data, such as additions, deletions and split & merge, of relation clusters like communities in social networks. Our proposed model abstracts observed time-varying object-object

extend the infinite H

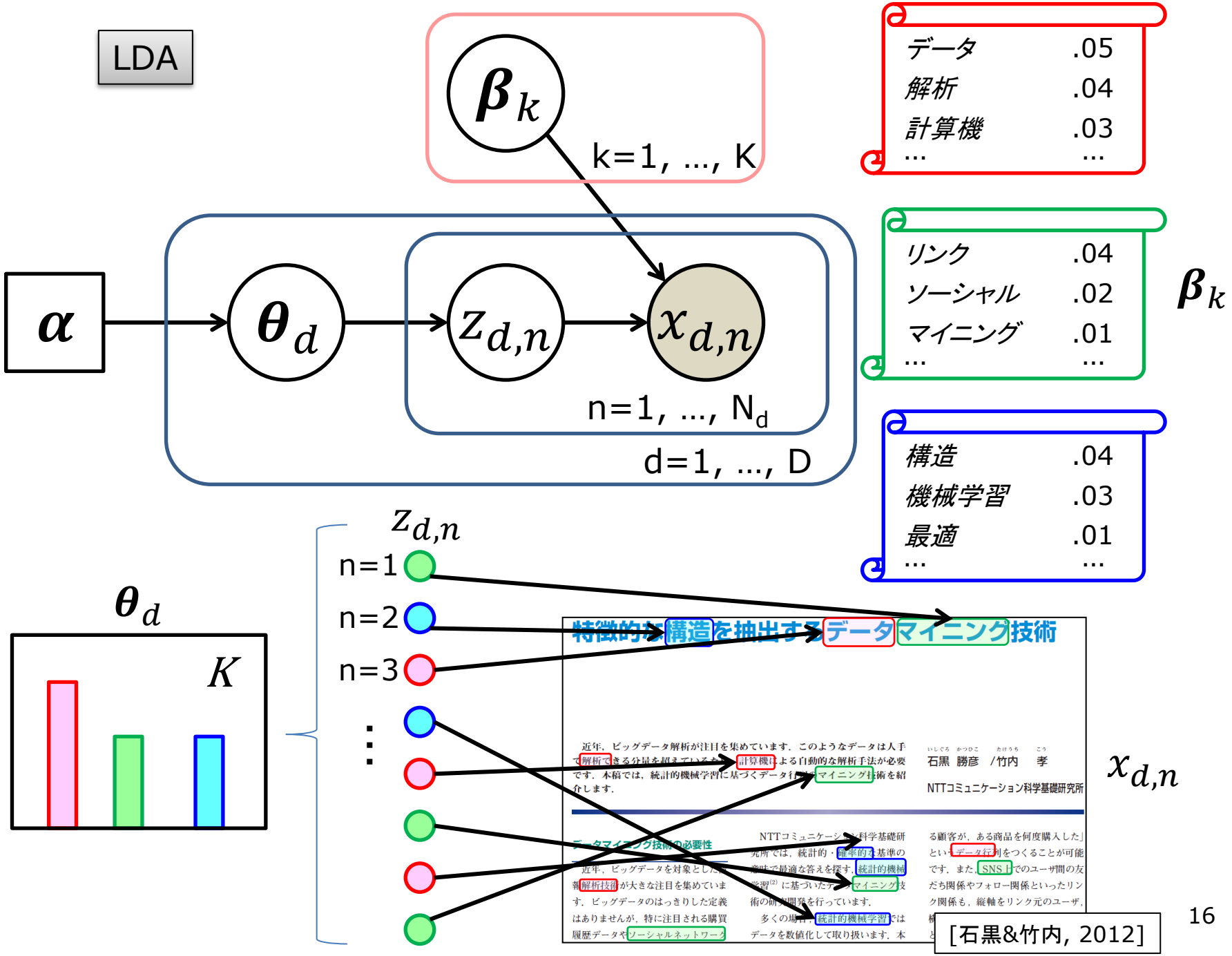
changes in the structu

simultaneously. We

synthetic and real-wo

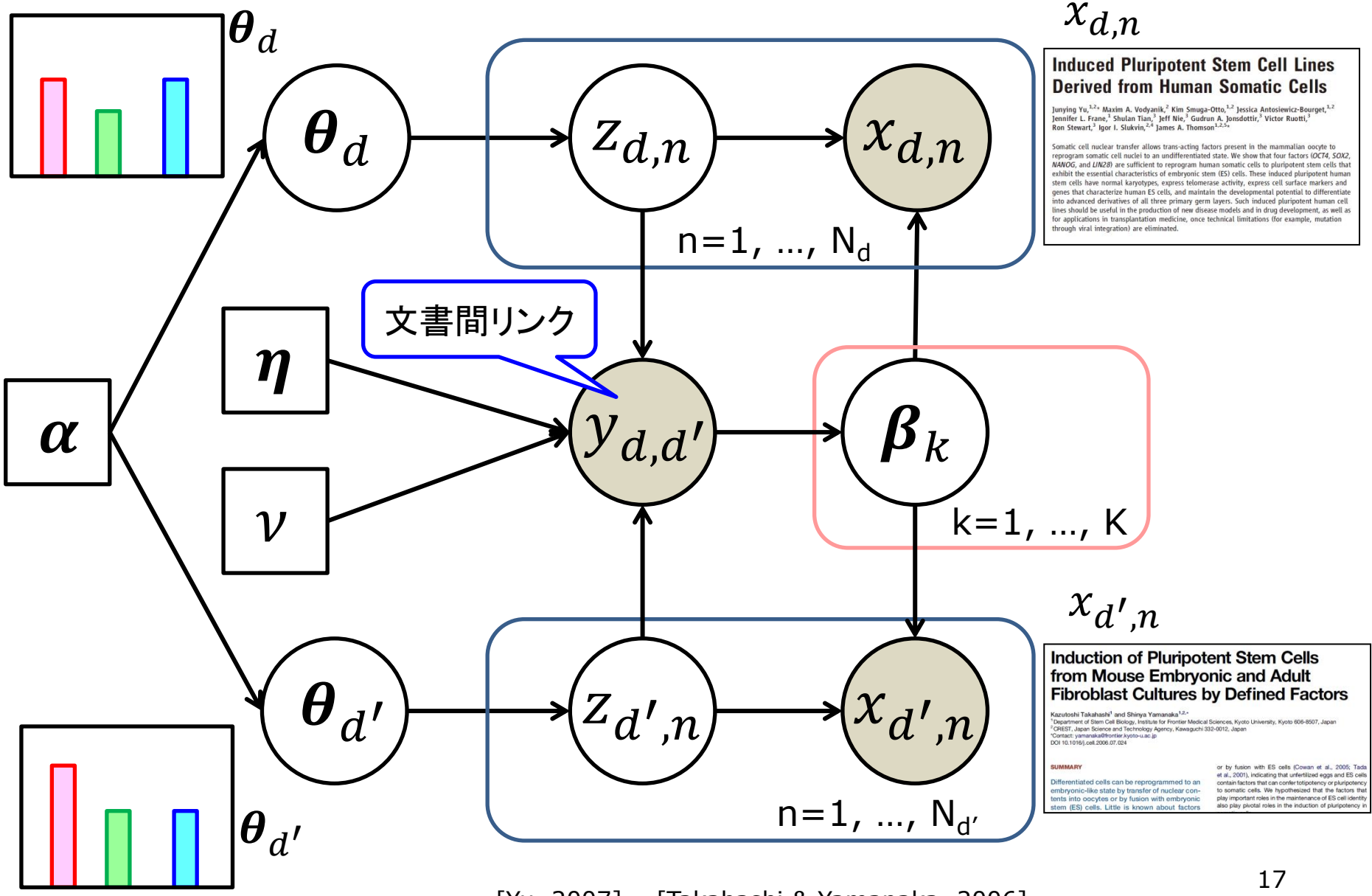
[Ishiguro, 2010]

LDA



[石黒&竹内, 2012]

Relational Topic Model（d, d'に関するプレートは省略）



$x_{d,n}$

Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells

Junying Yu,[1,2,*] Maxim A. Vodyanik,[2] Kim Smuga-Otto,[1,2] Jessica Antosiewicz-Bourget,[1,2] Jennifer L. Frane,[1] Shulan Tian,[3] Jeff Nie,[3] Gudrun A. Jonsdottir,[3] Victor Ruotti,[3] Ron Stewart,[3] Igor I. Slukvin,[2,4] James A. Thomson[1,2,5,*]

Somatic cell nuclear transfer allows trans-acting factors present in the mammalian oocyte to reprogram somatic cell nuclei to an undifferentiated state. We show that four factors (*OCT4*, *SOX2*, *NANOG*, and *LIN28*) are sufficient to reprogram human somatic cells to pluripotent stem cells that exhibit the essential characteristics of embryonic stem (ES) cells. These induced pluripotent human stem cells have normal karyotypes, express telomerase activity, express cell surface markers and genes that characterize human ES cells, and maintain the developmental potential to differentiate into advanced derivatives of all three primary germ layers. Such induced pluripotent human cell lines should be useful in the production of new disease models and in drug development, as well as for applications in transplantation medicine, once technical limitations (for example, mutation through viral integration) are eliminated.

$x_{d',n}$

Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors

Kazutoshi Takahashi[1] and Shinya Yamanaka[1,2,*]
[1]Department of Stem Cell Biology, Institute for Frontier Medical Sciences, Kyoto University, Kyoto 606-8507, Japan
[2]CREST, Japan Science and Technology Agency, Kawaguchi 332-0012, Japan
*Contact: yamanaka@frontier.kyoto-u.ac.jp
DOI 10.1016/j.cell.2006.07.024

SUMMARY

Differentiated cells can be reprogrammed to an embryonic-like state by transfer of nuclear contents into oocytes or by fusion with embryonic stem (ES) cells. Little is known about factors

or by fusion with ES cells (Cowan et al., 2005; Tada et al., 2001), indicating that unfertilized eggs and ES cells contain factors that can confer totipotency or pluripotency to somatic cells. We hypothesized that the factors that play important roles in the maintenance of ES cell identity also play pivotal roles in the induction of pluripotency in

[Yu, 2007]　[Takahashi & Yamanaka, 2006]

生成モデル

for 文書d = 1, 2, …, D

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-word assignment $z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$

word observation $x_{d,n}|z_{d,n},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$

for 文書ペア d = 1, 2, …, D, d′ = 1, 2, …, D

doc-doc link observation

$$y_{d,d'}|\boldsymbol{z}_d,\boldsymbol{z}_{d'},\boldsymbol{\eta},\nu\sim\mathrm{Bernoulli}\left(\psi\left(y_{d,d'}|\boldsymbol{z}_d,\boldsymbol{z}_{d'},\boldsymbol{\eta},\nu\right)\right)$$

for トピック k = 1, 2, …, K

topic-word proportion $\boldsymbol{\beta}_k$

文書-文書リンクの接続確率

- 各文書のトピックヒストグラム（の平均）を使う
  → 内容の要約情報を計算

$$\bar{\mathbf{z}}_d = \frac{1}{N_d}\sum_{n=1}^{N_d}\mathbf{z}_{d,n}$$

$z_{d,n}$をK次元ベクトルとして見ています

シグモイドモデル $\psi\left(y_{d,d'}|\mathbf{z}_d,\mathbf{z}_{d'},\boldsymbol{\eta},\nu\right)=\sigma(\boldsymbol{\eta}^T(\bar{\mathbf{z}}_d\circ\bar{\mathbf{z}}_{d'})+\nu)$

指数モデル $\psi\left(y_{d,d'}|\mathbf{z}_d,\mathbf{z}_{d'},\boldsymbol{\eta},\nu\right)=\exp(\boldsymbol{\eta}^T(\bar{\mathbf{z}}_d\circ\bar{\mathbf{z}}_{d'})+\nu)$

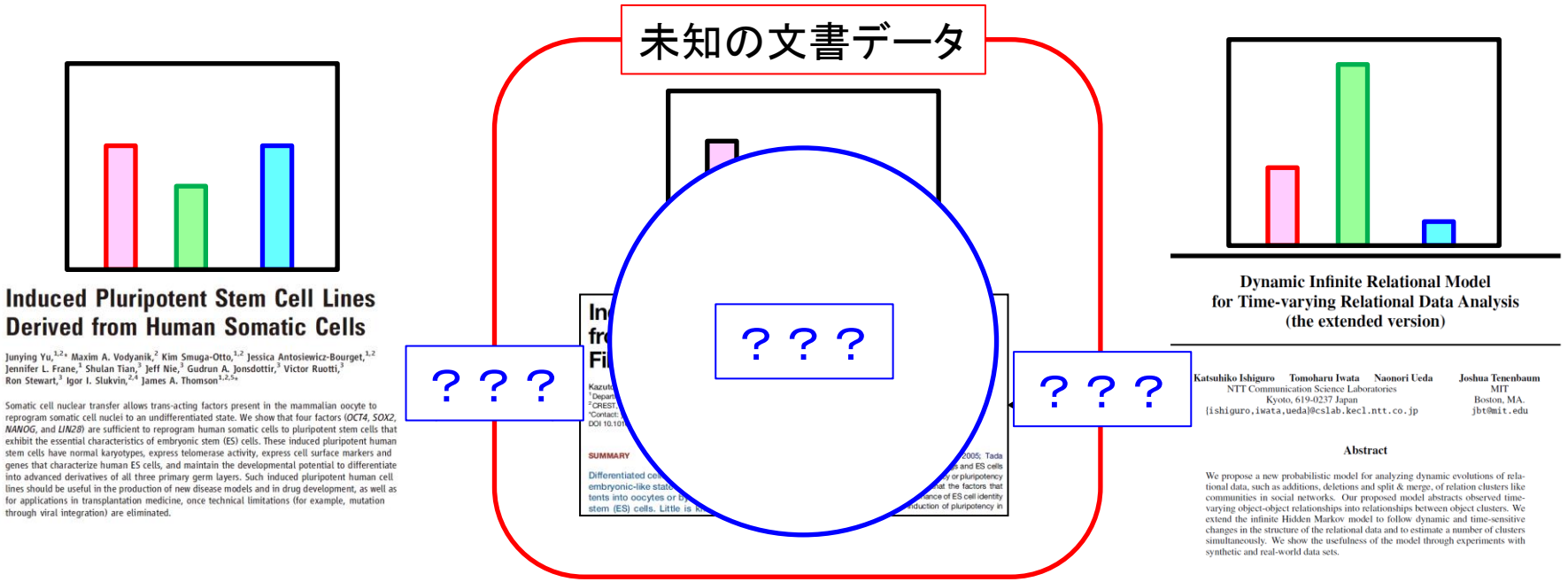
隠れ変数、パラメータの推定

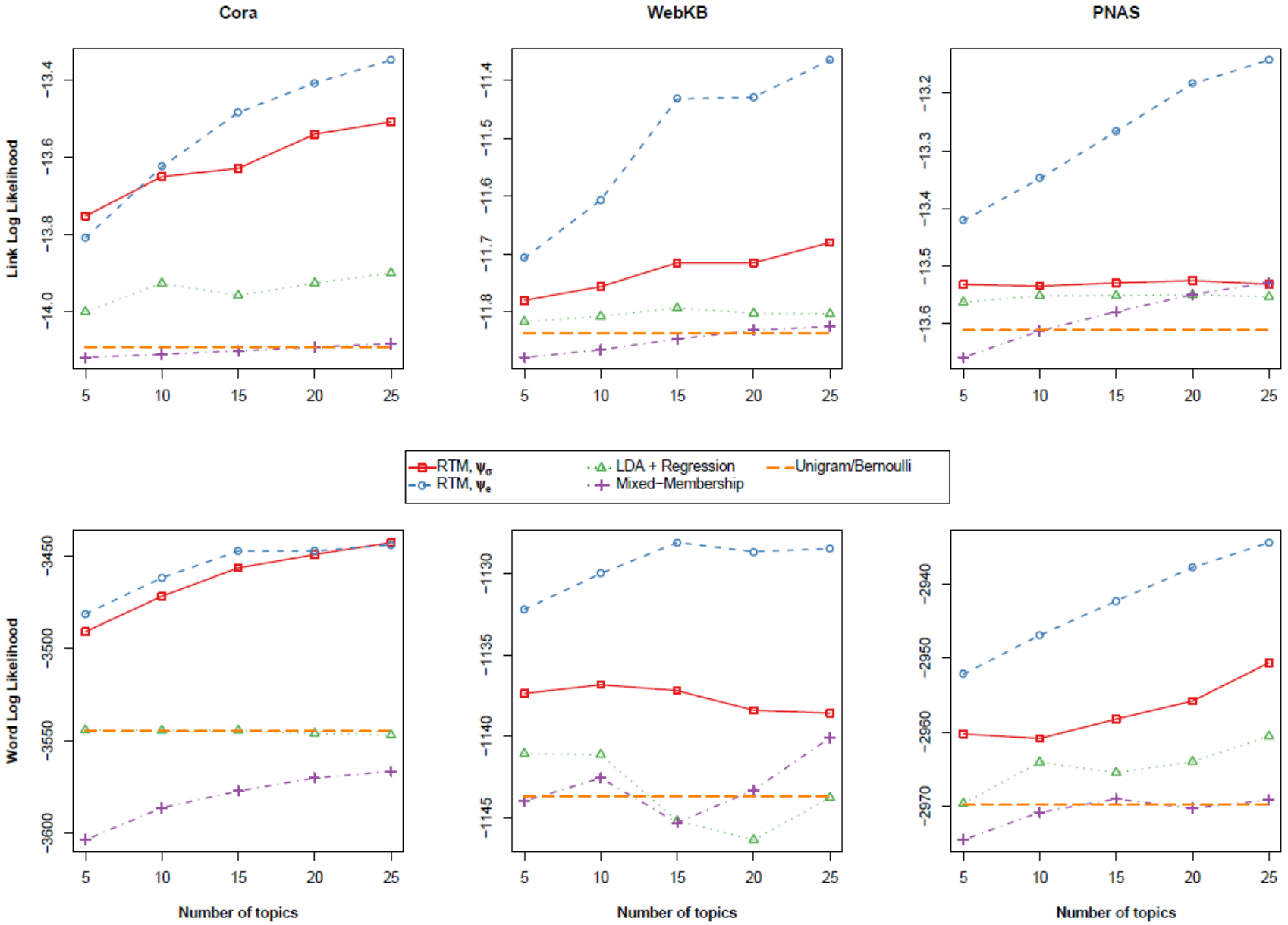
- 論文中では変分ベイズ(VB)による解法が導出されています

- 詳細はひとまず割愛します・・・

予測

- 学習が完了した提案モデルは、2種類の予測タスクに利用できます
  - リンク予測タスク
  - 内容（トピック）予測タスク





[Chang and Blei, 2009]

赤、青：提案法（詳細が少し違う）緑：トピックモデル→リンク予測
紫：文書情報を無視　オレンジ：文書情報と関係情報を別々にモデル化

Markov chain Monte Carlo convergence diagnostics: A comparative review	
Minorization conditions and convergence rates for Markov chain Monte Carlo Rates of convergence of the Hastings and Metropolis algorithms **Possible biases induced by MCMC convergence diagnostics** Bounding convergence time of the Gibbs sampler in Bayesian image restoration Self regenerative Markov chain Monte Carlo Auxiliary variable methods for Markov chain Monte Carlo with applications **Rate of Convergence of the Gibbs Sampler by Gaussian Approximation** Diagnosing convergence of Markov chain Monte Carlo algorithms	**RTM (ψ_e)**
Exact Bound for the Convergence of Metropolis Chains Self regenerative Markov chain Monte Carlo **Minorization conditions and convergence rates for Markov chain Monte Carlo** Gibbs-markov models Auxiliary variable methods for Markov chain Monte Carlo with applications Markov Chain Monte Carlo Model Determination for Hierarchical and Graphical Models Mediating instrumental variables A qualitative framework for probabilistic inference Adaptation for Self Regenerative MCMC	**LDA + Regression**

[Chang and Blei, 2009]

Relational Topic Model: まとめ

- 文書と文書の間にリンクがあるデータセットのモデル化
- 文書のトピックが似ているとリンクが張られやすくなるようにモデルを立てている
- リンク予測や内容予測、お勧め論文など

他の関係データトピックモデル

- Liu et al., “Topic-link LDA: Joint models of topic and author community”, in Proc. ICML, 2009.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Kemp, 2006] Kemp et al., “Learning Systems of Concepts with an Infinite Relational Model”, in Proc. AAAI, 2006.
- [Chang and Blei, 2009] Chang and Blei, “Relational Topic Models for Document Networks”, in Proc. AISTATS, 2009.
- [Takahashi & Yamanaka, 2006] Takahashi and Yamanaka, “Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors”, Cell, Vol. 126, pp. 663-676, 2006.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.
- [Ishiguro, 2010] Ishiguro et al, “Dynamic Infinite Relational Model for Time-varying Relational Data Analysis”, in Proc. NIPS, 2010.

引用及び参考文献

- [Yu, 2007] Yu et al., “Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells”, Science, Vol. 318, pp. 1917-1920, 2007.

NTT

トピックモデルの応用：画像・動画像データ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

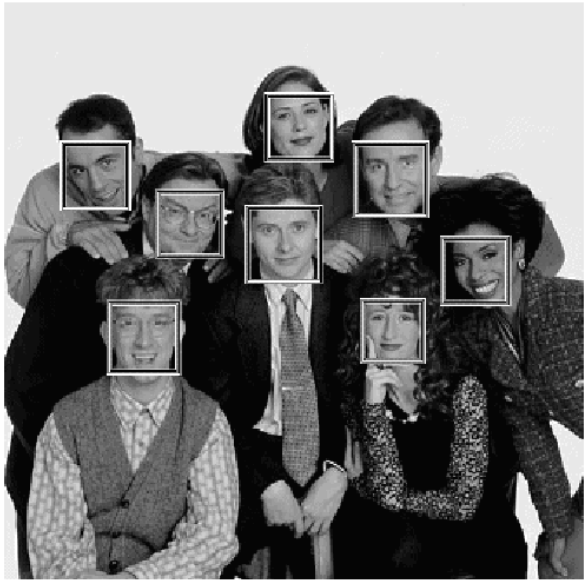
このスライドの“トピック”

- （動）画像認識・処理(Computer Vision)は最先端の機械学習技術が素早く導入される分野です

- この分野でもトピックモデルは猛威を振るっています（いました？）

Computer Vision（機械視覚・・・？）

- 人工知能の黎明期から、機械学習・パターン認識のもっとも分かり易い応用分野です



[Viola & Jones, 2001]

[Shimizu, 2008]

機械学習とcv

- 最先端のパターン認識・機械学習研究との親和性が非常に高いです
  - 大きなデータサイズ、リッチなコンテキスト
  - 実世界の複雑さ
  - 色々なことが可能

すべてを概観することは不可能なので

- 今回はトピックモデル応用がある論文まわりだけです

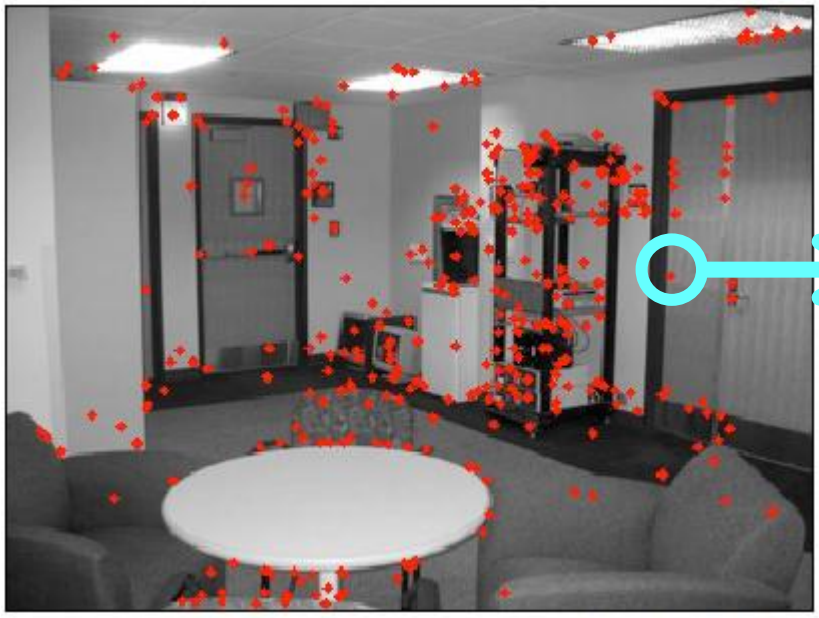
- その前に、あらゆる手法でほぼ共通して利用される基本的なアイデアを説明します

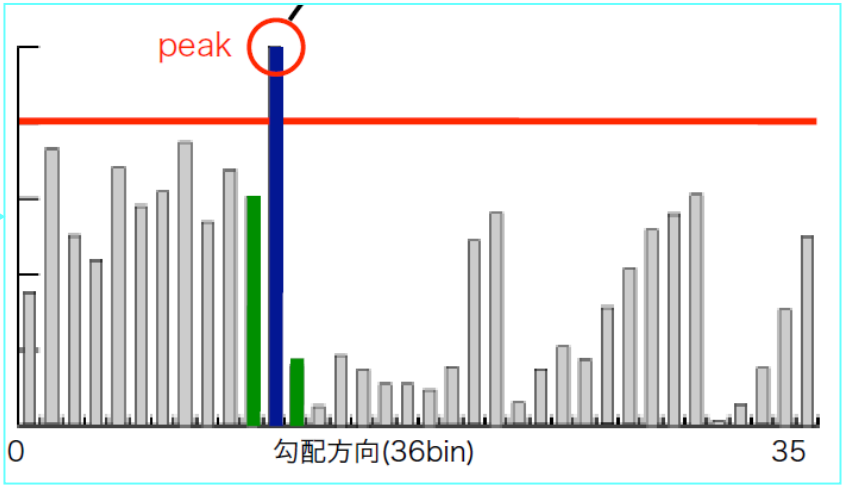
SIFT [Lowe, 2004]

- 画像認識系研究のデファクトスタンダードな特徴量

被引用数: 16268 (as of 2012/10/30, Google Scholar)

- 画素値勾配変化の局所極大・極小点を検出
- 注目点付近の画像勾配分布を計算

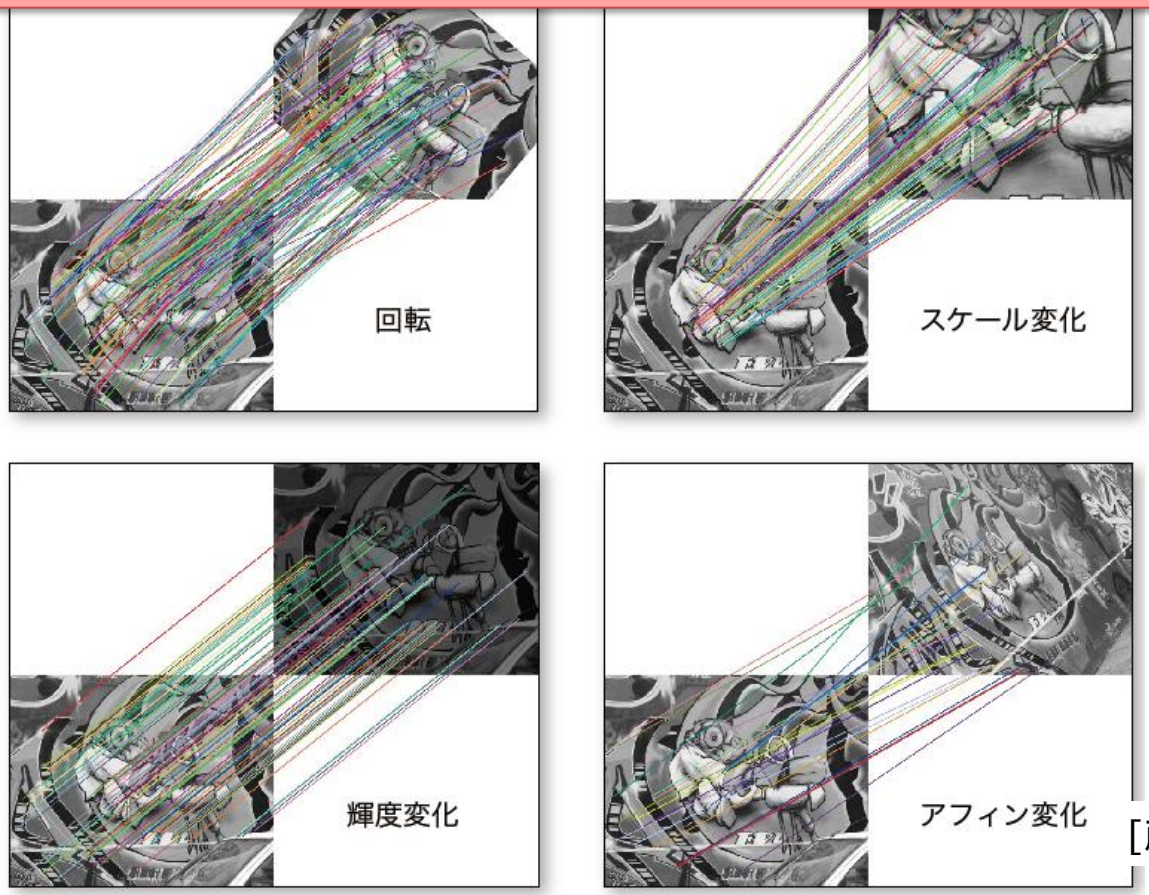




[藤吉, 2007]

- 画像の回転･スケール変化に対し不変
- 照明変化に対してロバスト

画像間の対応点検出に理想的な性質
→画像のパノラマ合成、物体認識、画像分類など
ありとあらゆる認識系のタスクで利用されている☺



[藤吉, 2007]

SIFT [Lowe, 2004]

- 近年は高速化・特徴圧縮した拡張法・関連法もよく利用される
- ☺ 中部大学・藤吉先生による日本語チュートリアルをお勧めします

藤吉, "Gradientベースの特徴抽出 - SIFTとHOG - ",
情報処理学会 研究報告 CVIM 160, pp. 211-224, 2007.

(参考文献にも載せてあります)

Bag of Visual Words: 画像データの「文書化」

- 局所画像特徴量(含むSIFT)は1枚の画像からたくさん計算できます

- 1つ1つの特徴量は高次元ベクトルです
  - (SIFTだと128次元実数ベクトル)

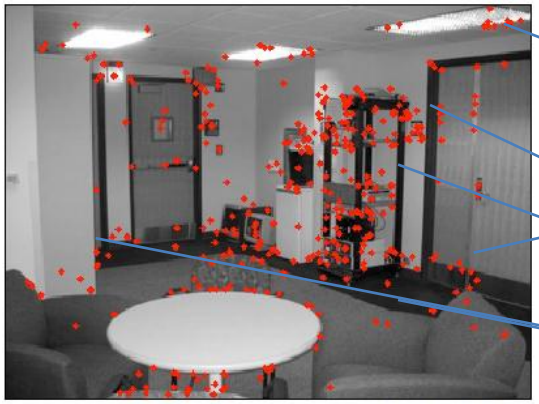
- そこで、適当にサボった表現がほしくなります
  → Bag Of Visual Words

Bag of Visual Words: 画像データの「文書化」

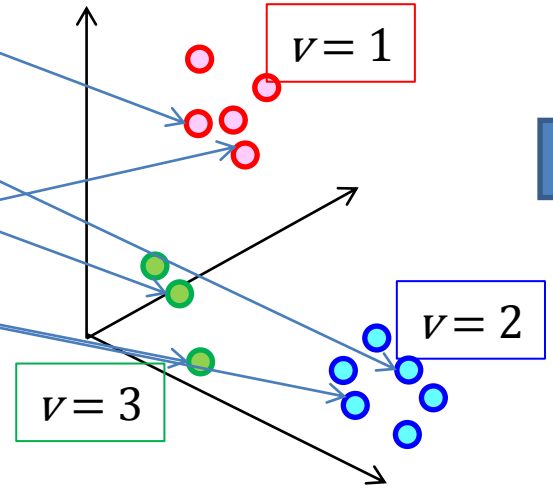
局所特徴を抽出

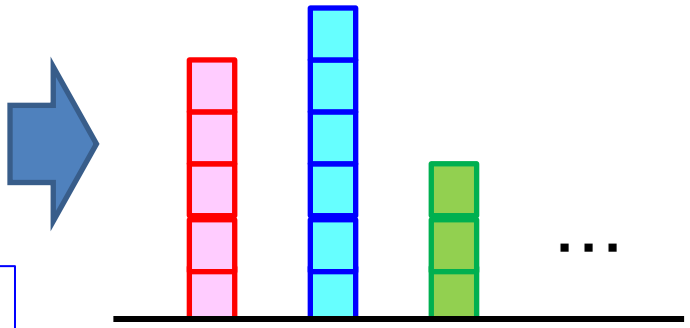
K-meansなどによる量子化

V次元のヒストグラム



[藤吉, 2007]





Visual Words:
単語に相当

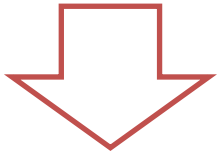
Bag of Visual Words:
文書に相当

☺ 機械学習・自然言語処理で開発された
各種学習モデルをそのまま利用できる
☺ メモリ・計算量削減

SIFT + Bag of Visual Words = 安心

- ここ数年は、SIFTをBoVWで量子化したものを観測特徴量として利用する研究ばかりです

- SIFT（あるいはその進化系）: 高性能な基本特徴量
- Bo(V)W: 計算量削減、機械学習技術との連携が容易



画像データでもトピックモデル！！

Scene Recognition
[Fei-Fei and Perona, 2005]

Fei-Fei and Perona,
"A Bayesian Hierarchical Model for Learning
Natural Scene Categories",
in Proc. CVPR. 2005.

“mountain”

“ocean”

“forest”

“ocean mountain”

人間はどのように sceneを認識しているのか？

- 画像中の全ての物体・背景を個別に認識して、それらから推測されるコンテキストとしてsceneを認識する [Treisman & Gelade, 1980]

- No!!

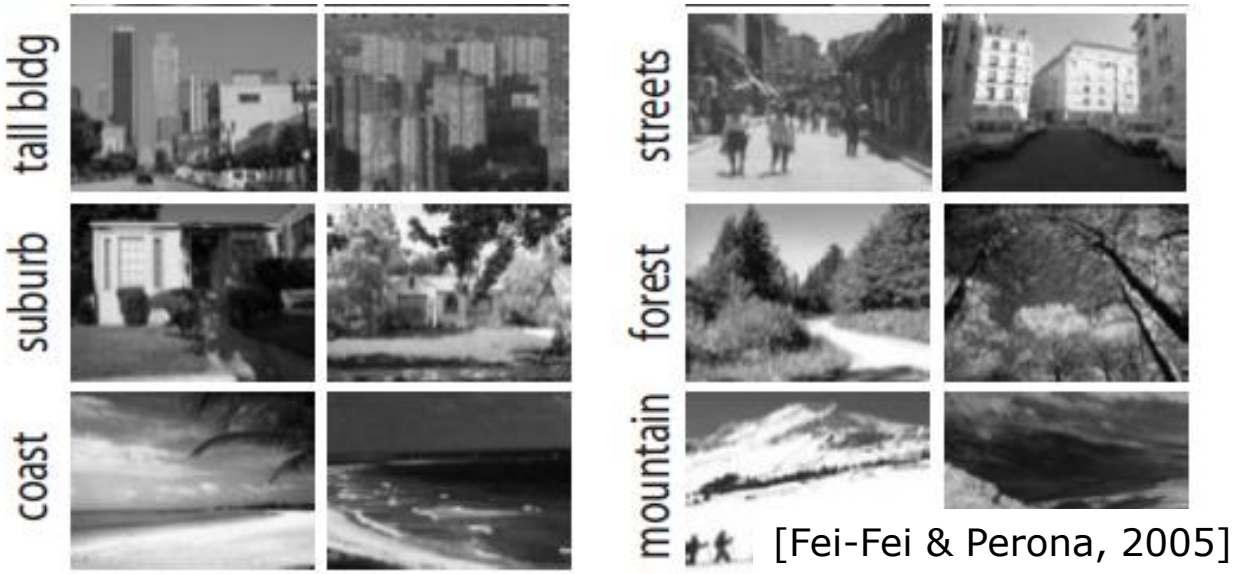
- 人間はほとんど注視をすることなく、画像のsceneを識別できる[Li, 2002]

Computer visionでも そうですか？

- 色々な研究者がいろいろ試しました
- 結論：生の画像情報だけでなく、それを「意味のある良いパターン」に抽象化した表現が作れるといい性能が得られる
- 問題：「良いパターン」を人間が設計するのは非常に大変

提案法: Theme model for scene recognition

- Scene recognitionにおいて、「意味のあるパターン＝theme」を使うと良い
- トピック＝themeと仮定してトピックモデルを適用してみた→うまくいった！☺

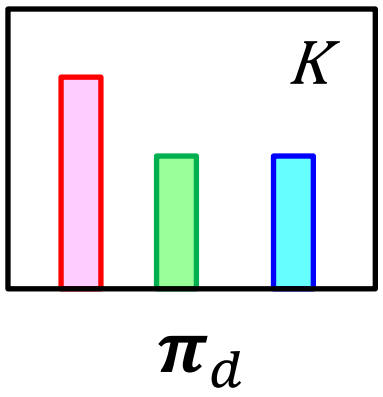


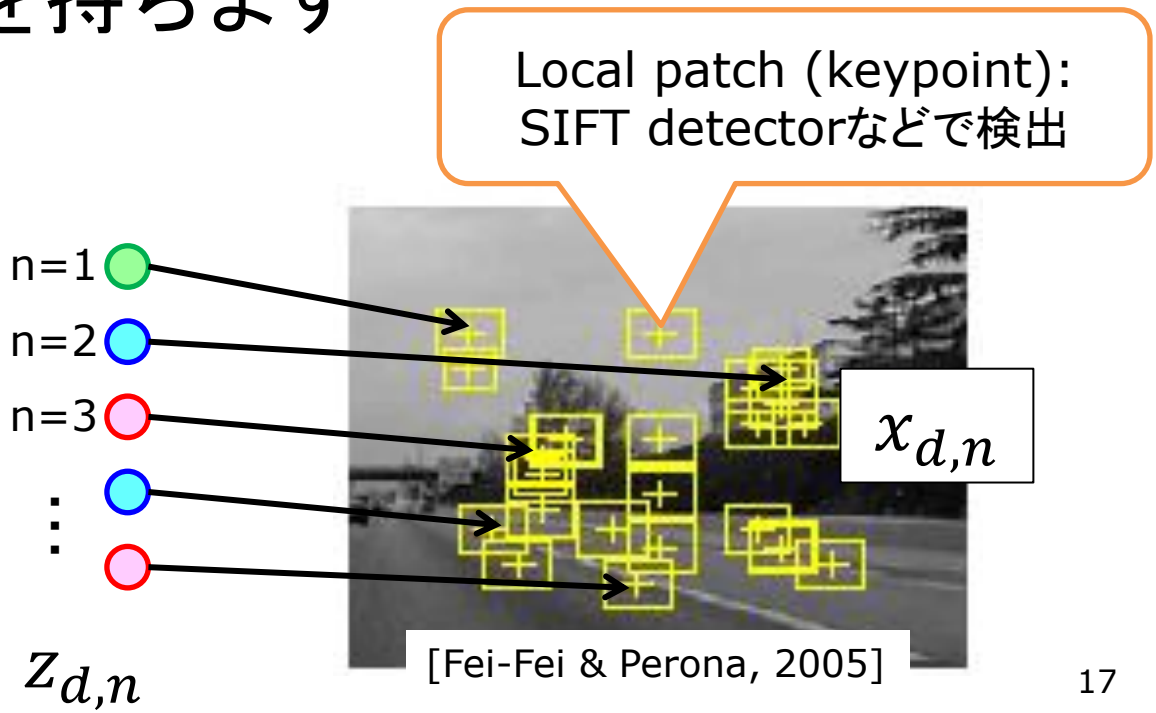
[Fei-Fei & Perona, 2005]

提案法のアイデア: 画像の表現

- 画像が文書で、単語は局所的なパッチ（小画像）のVisual Wordです
- 各画像ごとに特有のトピック分布、つまりパッチのパターンを持ちます

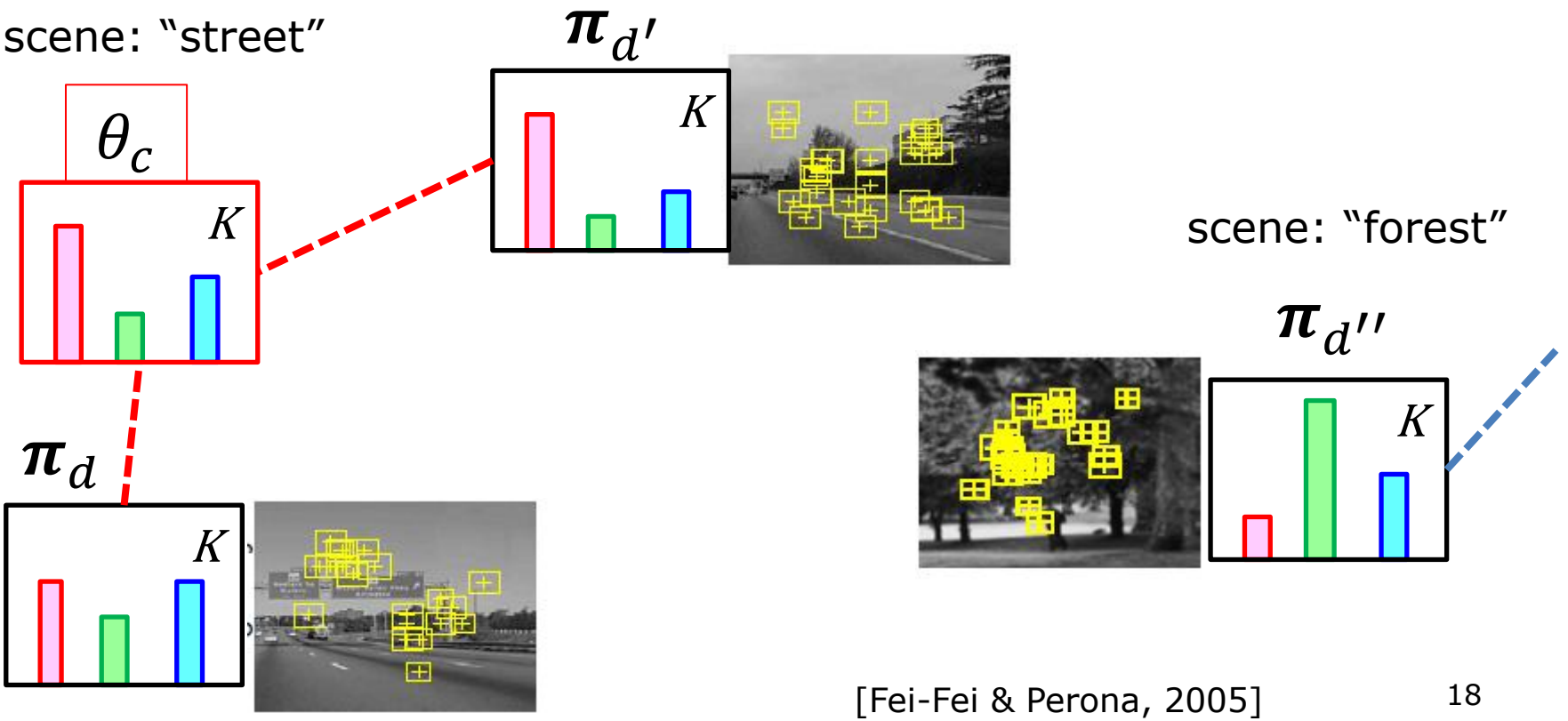
画像(文書) d





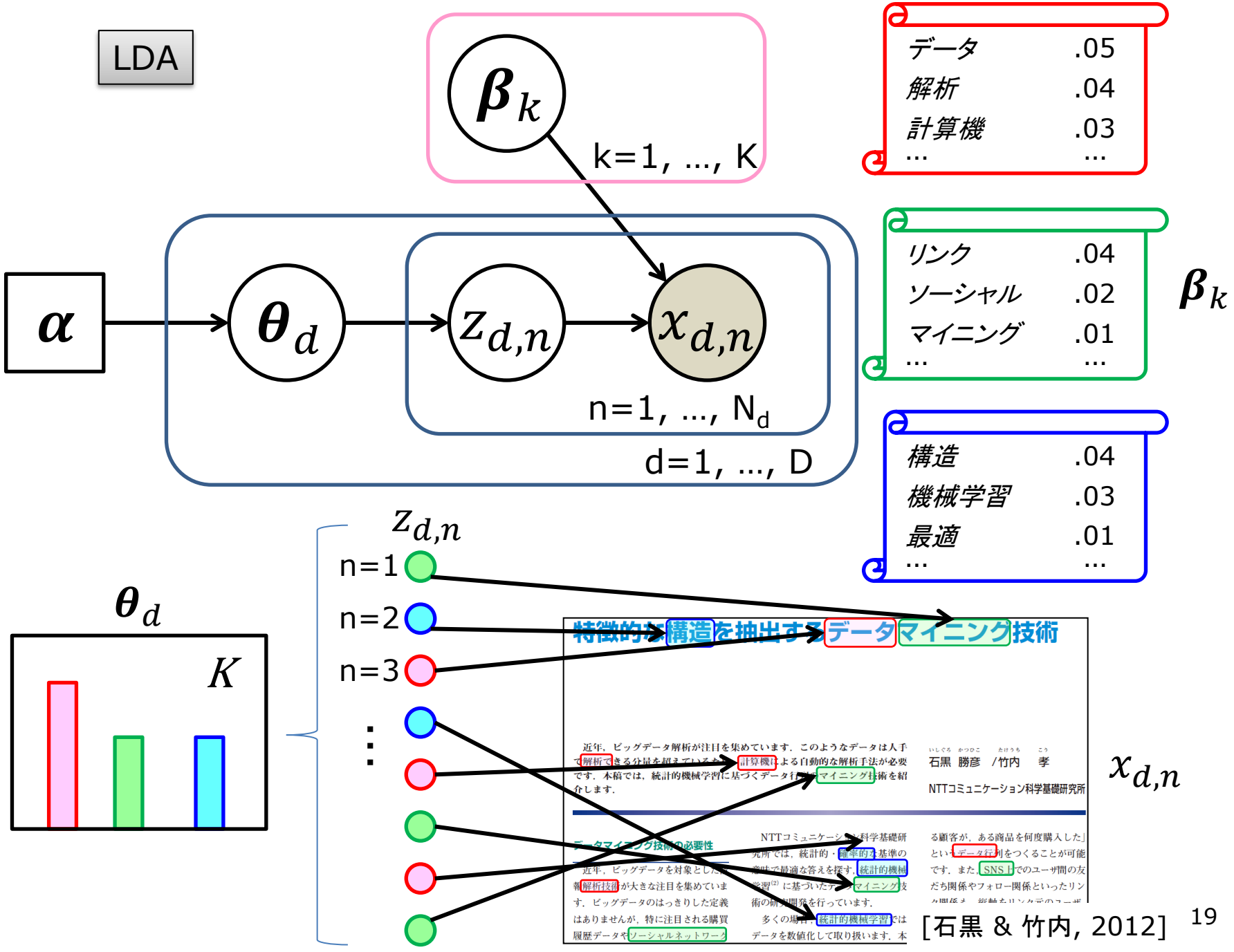
提案法のアイデア: sceneカテゴリのトピック中心

- sceneとしてカテゴライズできる以上、同じsceneカテゴリの画像は相関があるはず



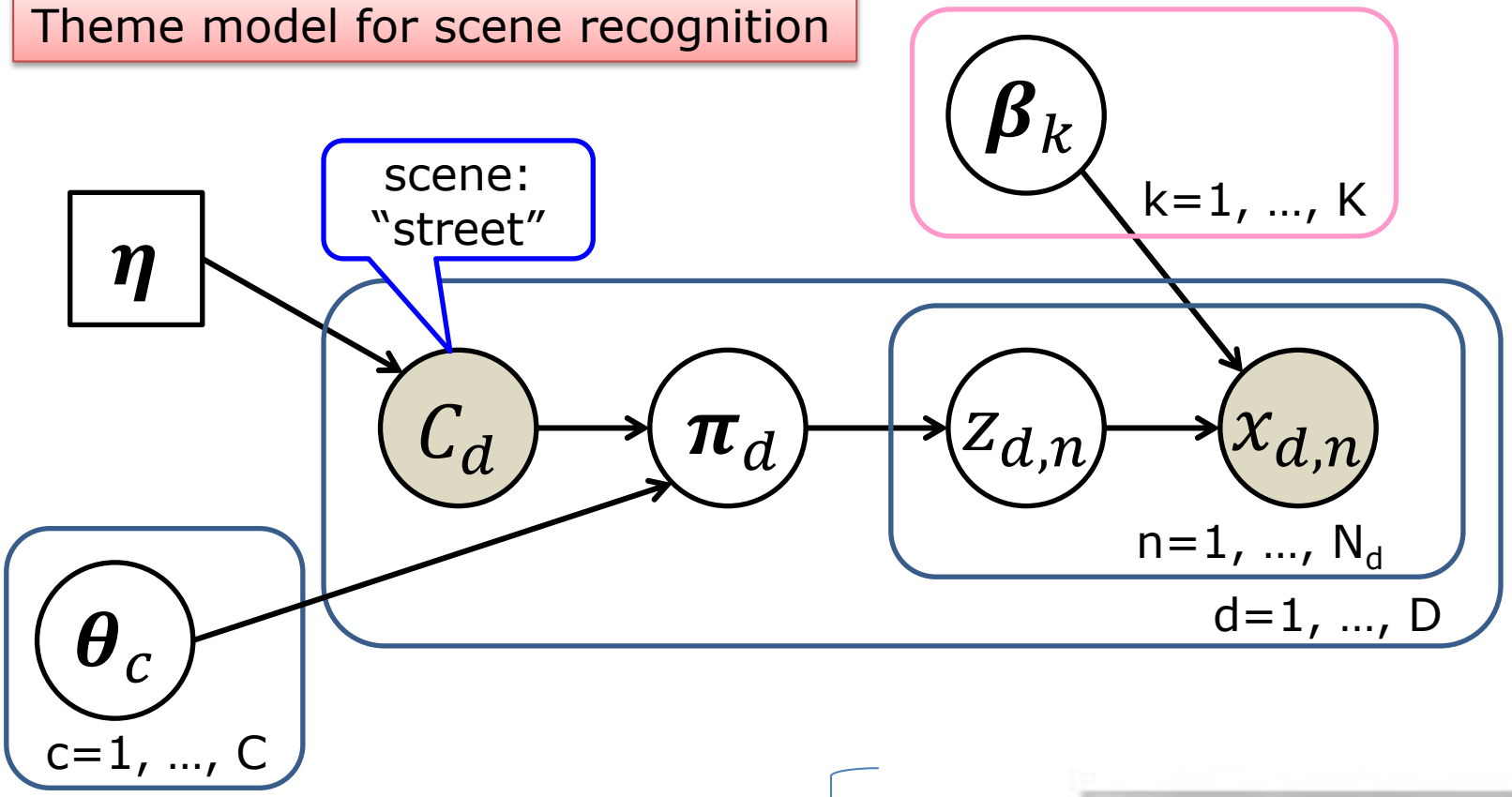
[Fei-Fei & Perona, 2005]

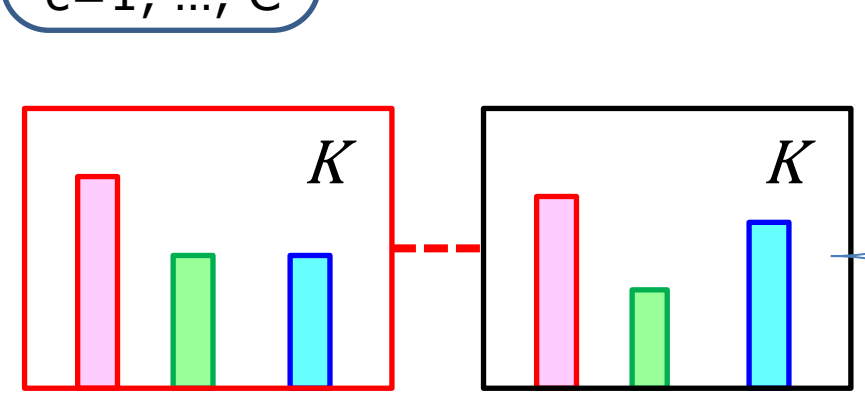
LDA

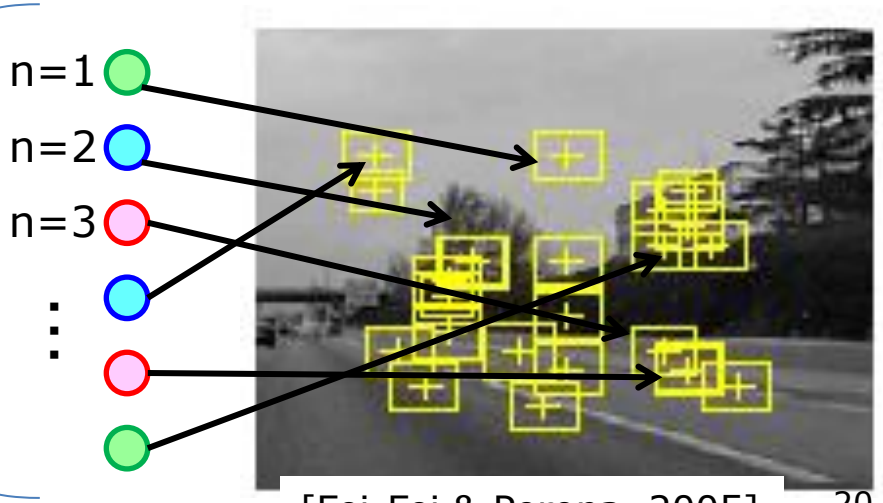


[石黒 & 竹内, 2012]

Theme model for scene recognition







[Fei-Fei & Perona, 2005]

生成モデル

for 画像d = 1, 2, …, D

scene category label $c_d|\boldsymbol{\eta} \sim \mathrm{Mult}(\boldsymbol{\eta})$

topic proportion $\boldsymbol{\pi}_d|\boldsymbol{\theta}_{c_d} \sim \mathrm{Dir}(\boldsymbol{\theta}_{c_d})$

for 単語n = 1, 2, …, N_d

topic-VW assignment $z_{d,n}|\boldsymbol{\pi}_d \sim \mathrm{Mult}(\boldsymbol{\pi}_d)$

VW observation $x_{d,n}|z_{d,n}, \{\boldsymbol{\beta}_k\} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{d,n}})$

for theme (topic) k = 1, 2, …, K

topic-VW proportion $\boldsymbol{\beta}_k$

for sceneカテゴリ c = 1, 2, …, C

“average” topic proportion $\boldsymbol{\theta}_c$

パラメータ、隠れ変数の推定

- 論文では変分ベイズ(VB)による推定法が紹介されています

- 通常のLDAの解法を参考にすれば、解は比較的簡単に導出できるようなので、ここでは割愛します

未知画像の識別

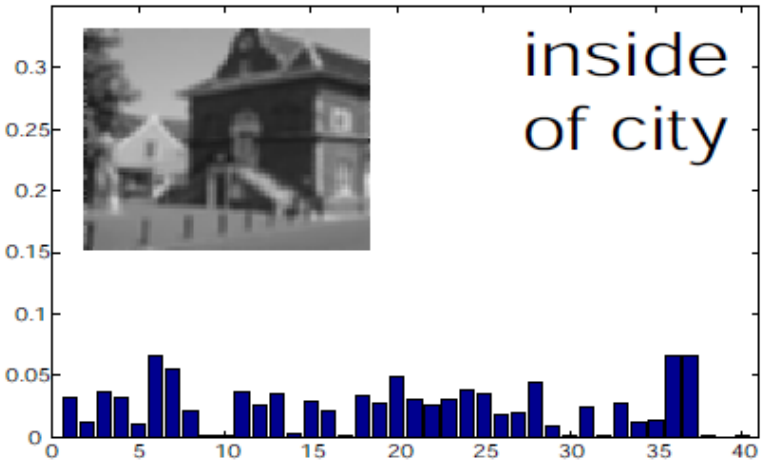
- 学習し終わったモデルに対して、未知画像 d のsceneカテゴリ c を推定します
- 最尤推定で計算しますが、正確な計算は不可能です
  - 適切な近似が必要ですが、これもLDAの元論文等を参考にしてください

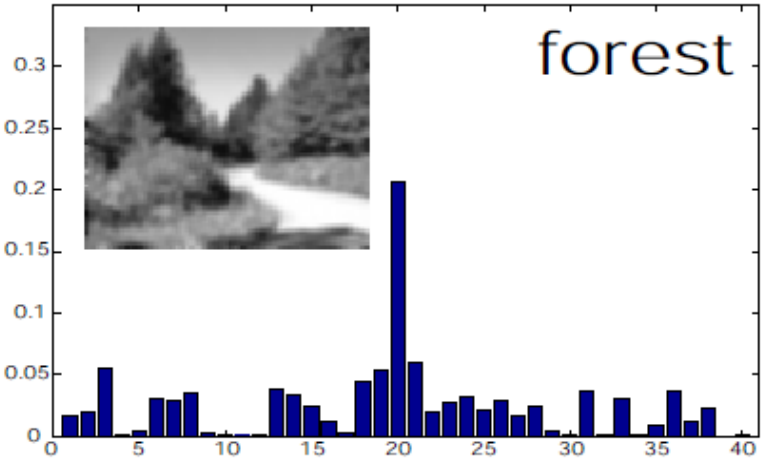
$$c_d = \arg\max_c p(X_d|c, \boldsymbol{\theta}, \boldsymbol{\beta})$$

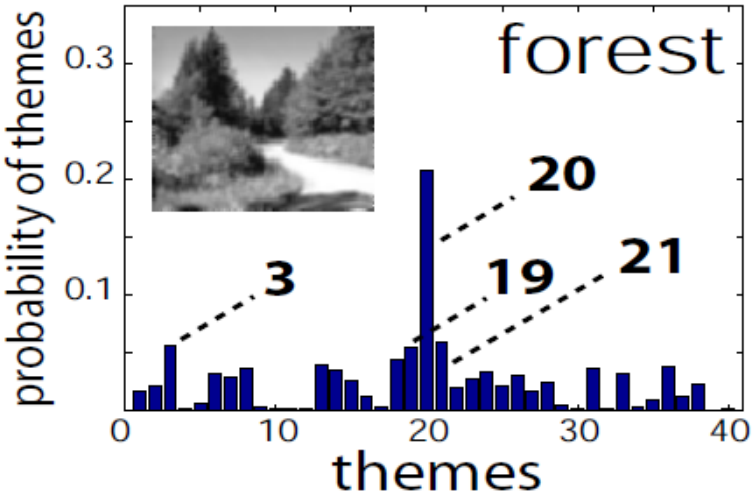
$$= \int p(\boldsymbol{\pi}_d|c, \boldsymbol{\theta}) \left\{ \prod_n \sum_k p(z_{d,n} = k|\boldsymbol{\pi}_d) p(x_{d,n}|z_{d,n}, \boldsymbol{\beta}) \right\} d\boldsymbol{\pi}_d$$

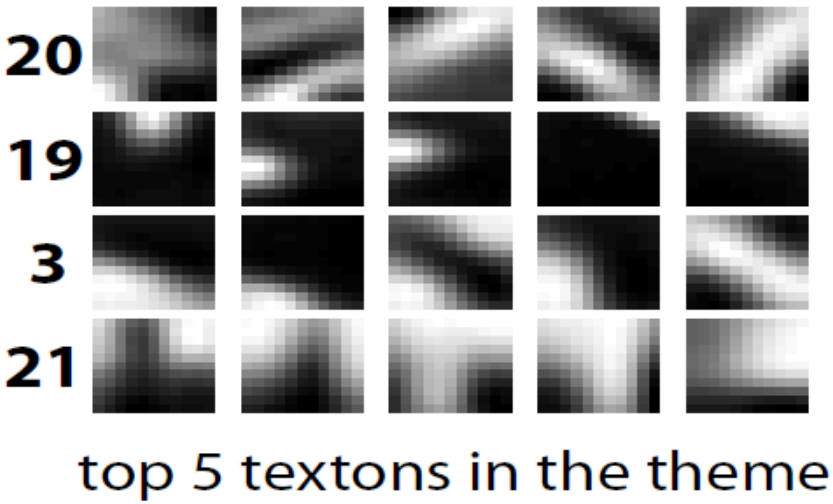
	# of categ.	training # per categ.	training requirements	perf. (%)
Theme Model 1	13	100	unsupervised	76
[17]	6	$\sim$ 100	human annotation of 9 semantic concepts for 60, 000 patches	77
[9]	8	250 $\sim$ 300	human annotation of 6 properties for thousands of scenes	89

[Fei-Fei & Perona, 2005]









[Fei-Fei & Perona, 2005]

まとめ: Theme model for Scene Recognition

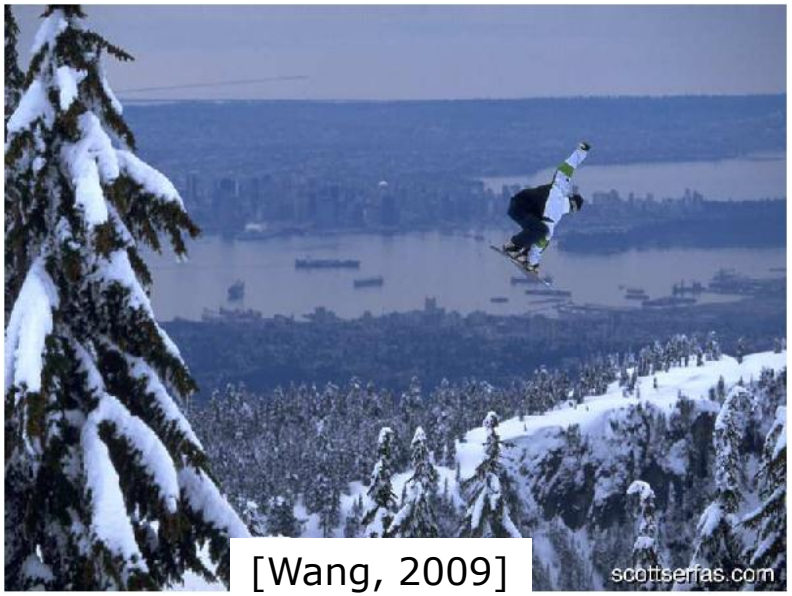
- BoVWを素直に使った（だけ）の、トピックモデル応用
- 13カテゴリの識別問題に対して、事前にパッチやthemeを人手で与える必要がない
- Theme (topic) に意味があるかは・・・？？

Scene Classification with Annotation [Wang, 2009]

Wang, Blei and Fei-Fei,
"Simultaneous al Model for Learning
Natural Scene Categories",
in Proc. CVPR. 2009.

画像カテゴリの識別問題 (image classification)

- 「この画像は何か当てて下さい」
- 画像に対して、有限のクラスのうち最も適切なものを当てる問題
- 古くからの画像認識問題



[Wang, 2009]

“running”

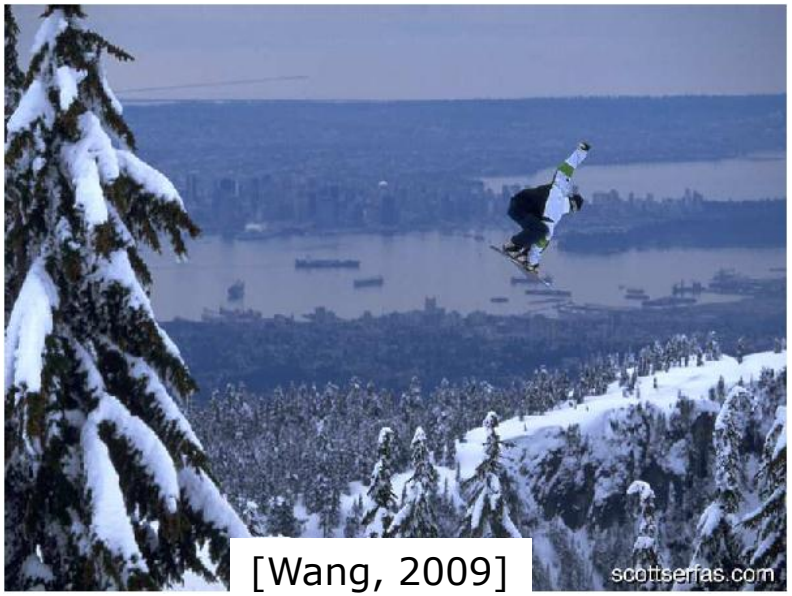
“skiing”

“snowboarding”

“painting”

画像のアノテーション付与問題 (image annotation)

- 「この画像に適切なタグを付与してください」
- 画像に対して、関連するタグ（単語）を複数付与する問題
- 画像検索の文脈などで幅広く研究



[Wang, 2009]

○ “forest”

○ “snowboarder”

○ “sky”

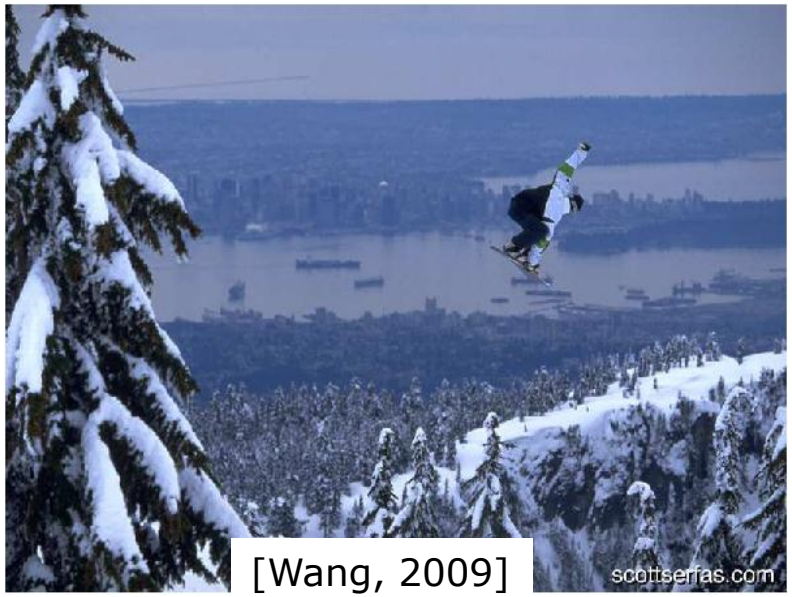
○ “jump”

× “track”

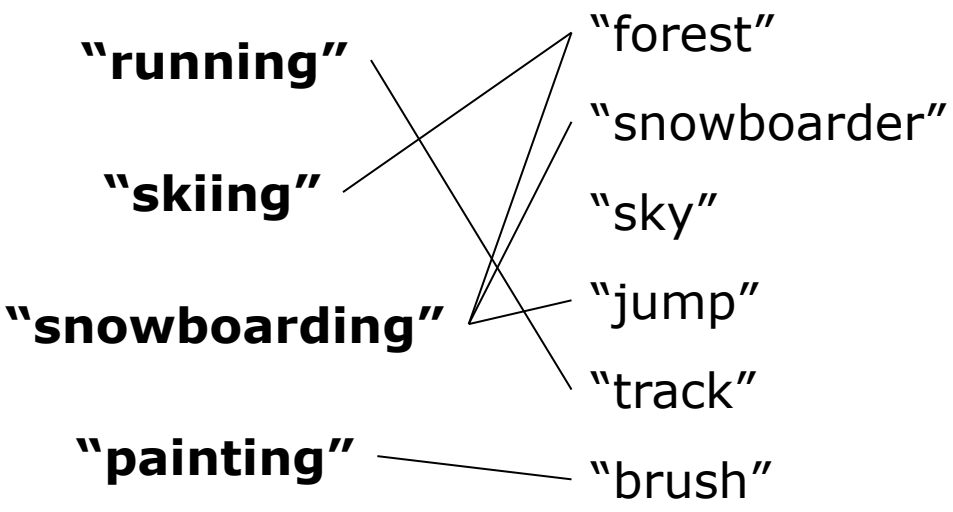
× “brush”

カテゴリ識別とアノテーション、実はすごく近い問題？

- クラスラベルとタグは間違いなく相関がある
- 両方を用いることで、より高精度に両問題を解ける？



[Wang, 2009]



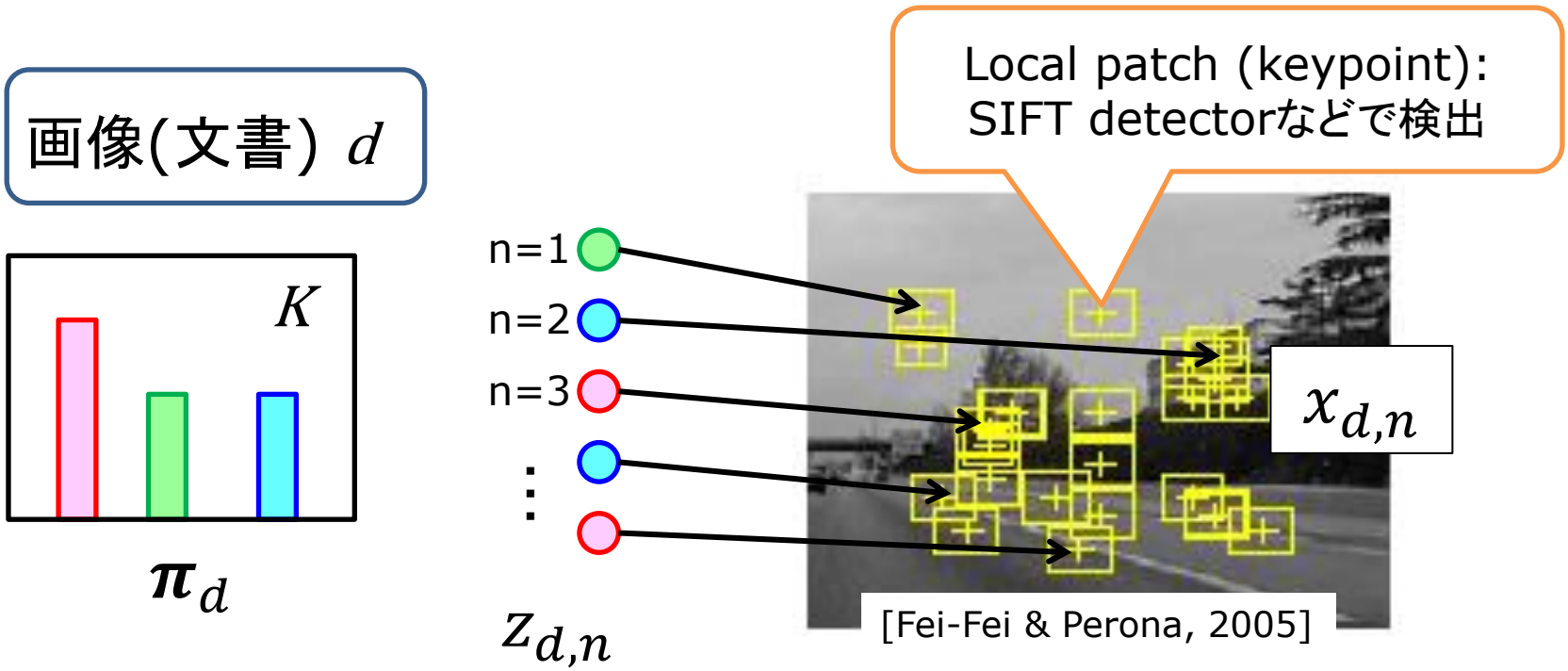
提案法: Multi-class sLDA with annotation

- 画像のカテゴリ識別(classification)とタグ当て(annotation)を、一つのトピックモデルで同時にモデル化

- 単一のモデルよりも両課題について高精度
- annotation無しでも、多クラス識別可能なトピックモデルとして新規性・応用性が高い

提案法のアイデア：画像の表現

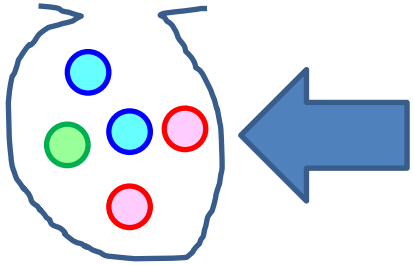
- 先ほどのモデルと同じです
- Local patchがN_d個あります



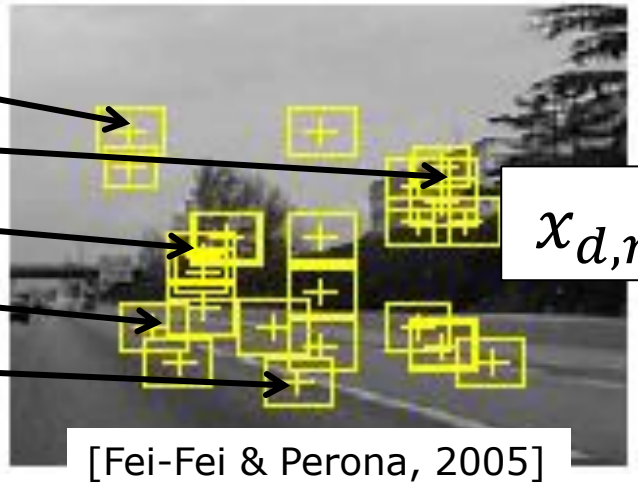
提案法のアイデア：画像とクラスラベルの関係

- 画像のクラスラベルは、画像全体のトピック割り当てから決定します

$$\bar{z}_d = \frac{1}{N_d}\sum_n z_{d,n}$$



n=1
n=2
n=3
⋮



$x_{d,n}$

[Fei-Fei & Perona, 2005]

Soft-max

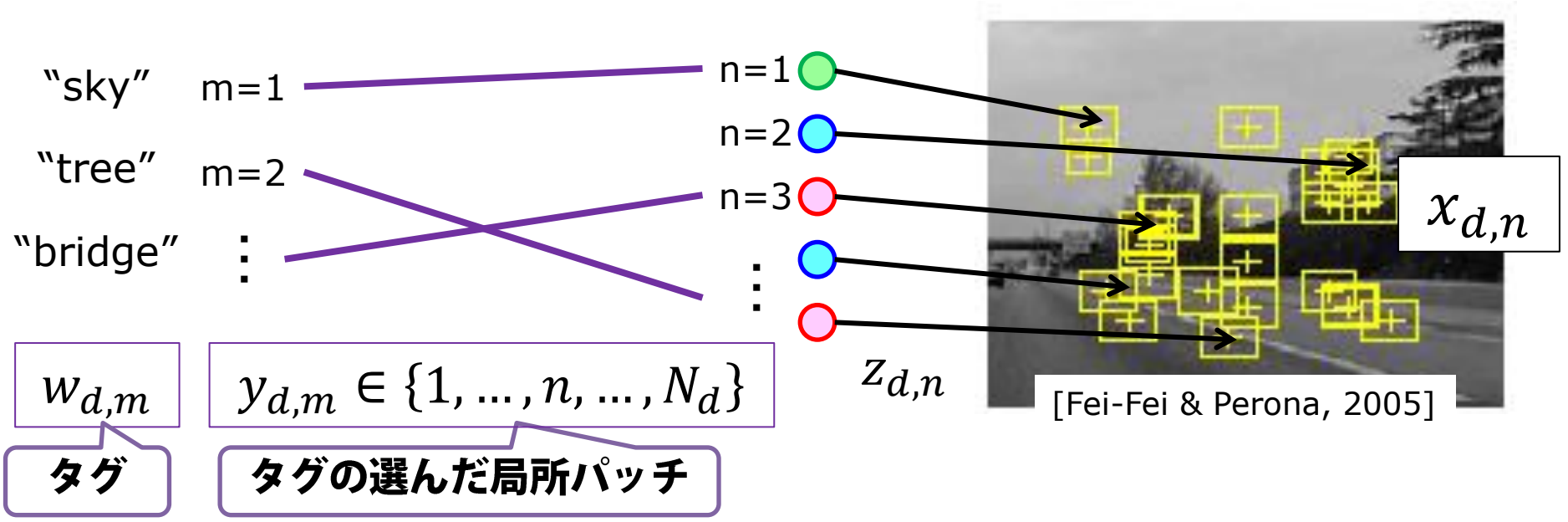
$z_{d,n}$

$c_d =$ “mountain” “water” “city” “street”

画像 d のクラスラベル

提案法のアイデア：画像とタグの関係

- 画像中の局所パッチが、あるタグの「生成元」になっていると考えて、トピックも共有します



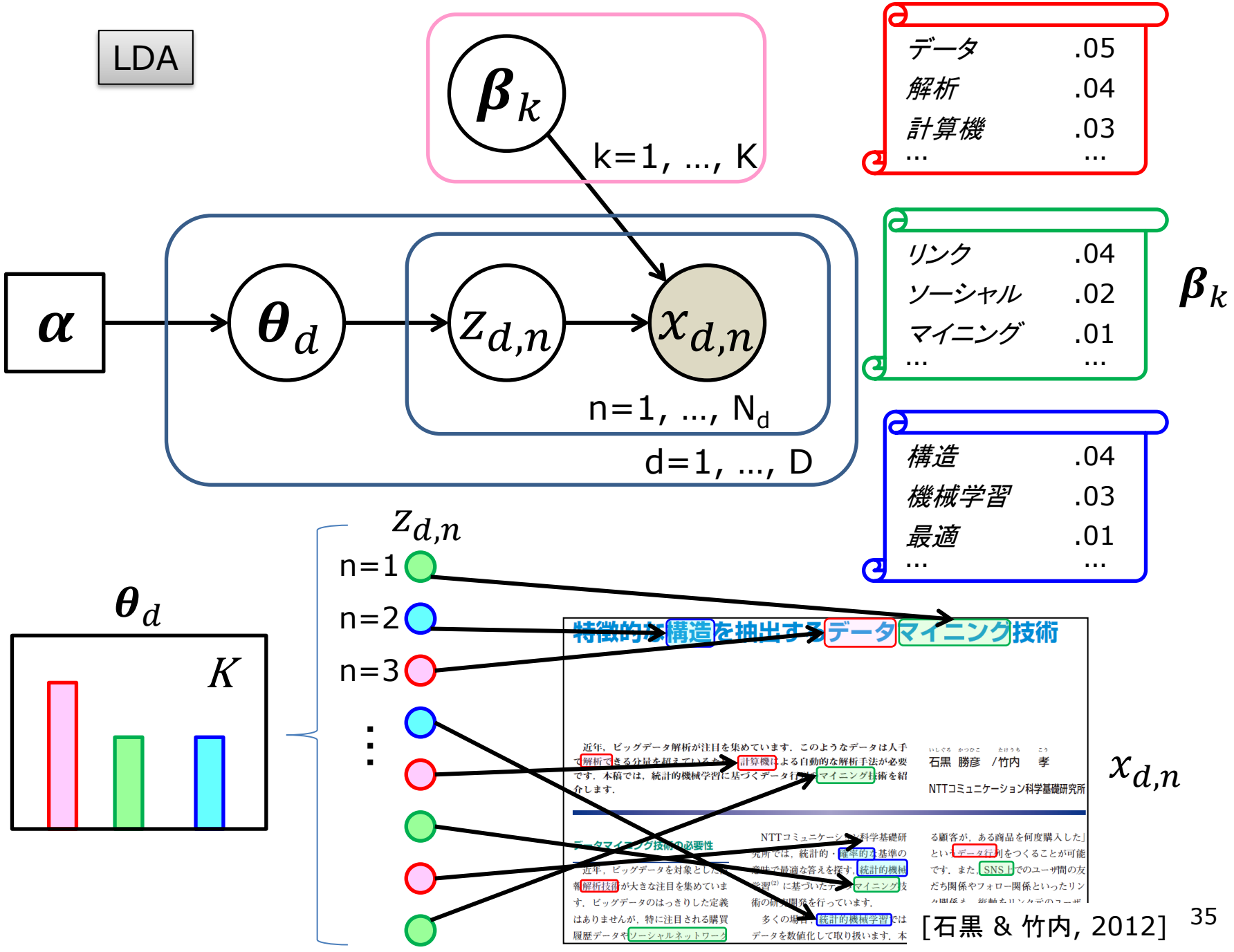
β_k

highway	.04
bridge	.03
…	…

sky	.04
cloud	.02
…	…

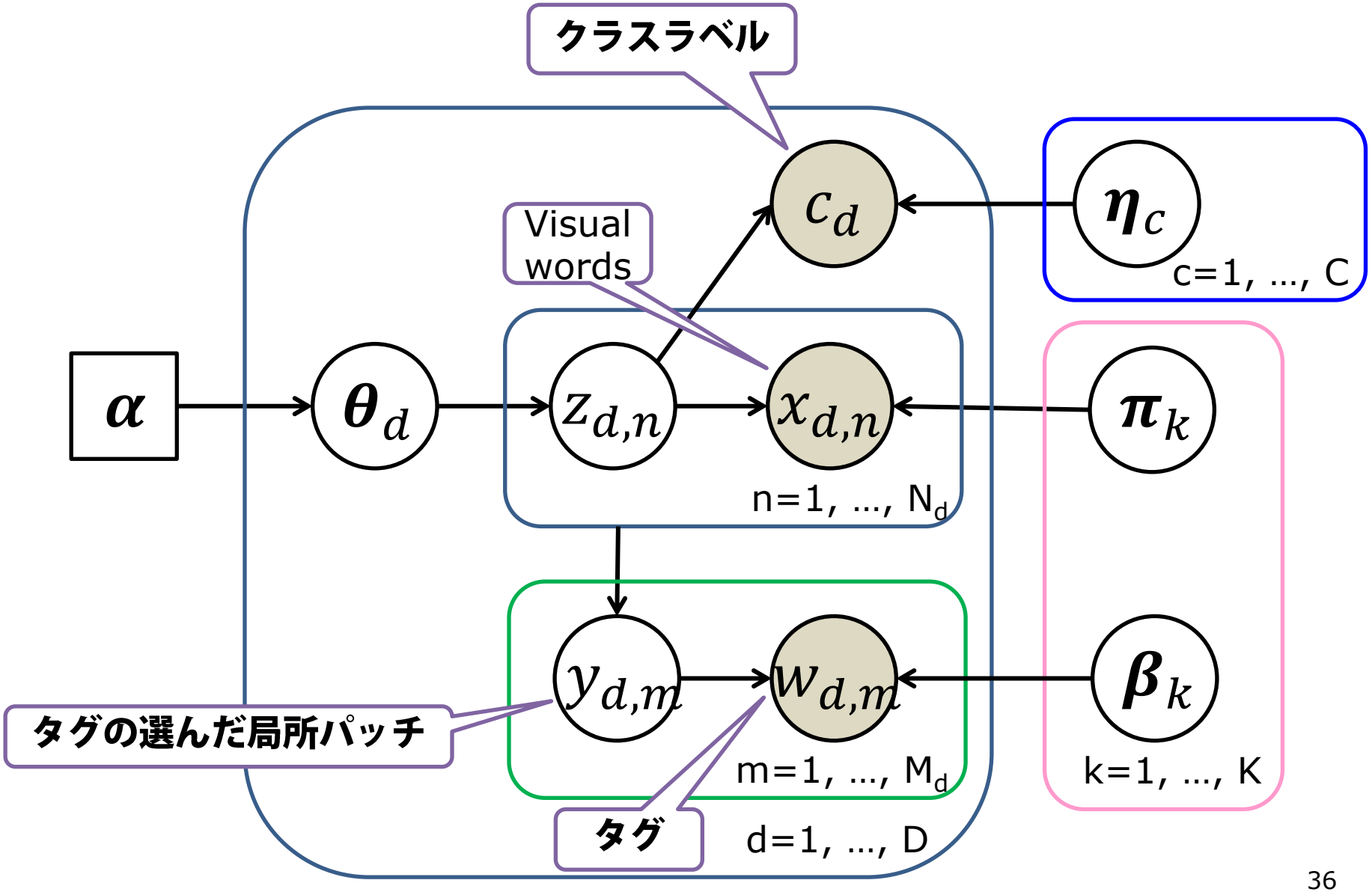
tree	.05
forest	.04
…	…

LDA



[石黒 & 竹内, 2012]

Multi-class sLDA with annotation



for 画像d = 1, 2, …, D

topic proportion $\boldsymbol{\theta}_d|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

for 単語n = 1, 2, …, N_d

topic-VW assignment $z_{d,n}|\boldsymbol{\theta}_d\sim\mathrm{Mult}(\boldsymbol{\theta}_d)$

VW observation $x_{d,n}|z_{d,n},\{\boldsymbol{\pi}_k\}\sim\mathrm{Mult}\left(\boldsymbol{\pi}_{z_{d,n}}\right)$

for タグm = 1, 2, …, M_d

tag-patch assignment $y_{d,m}\sim\mathrm{Uniform}(\{1,2,\cdots,N_d\})$

word observation $w_{d,m}|y_{d,m},\{z_{d,n}\},\{\boldsymbol{\beta}_k\}\sim\mathrm{Mult}\left(\boldsymbol{\beta}_{z_{d,y_{d,m}}}\right)$

class label $c_d|\{\boldsymbol{\eta}_c\},\{z_{d,n}\}\sim\mathrm{soft-max}\left(\{\boldsymbol{\eta}_c\},\{z_{d,n}\}\right)$

for トピック k = 1, 2, …, K

topic-VW proportion $\boldsymbol{\pi}_k$

topic-tag proportion $\boldsymbol{\beta}_k$

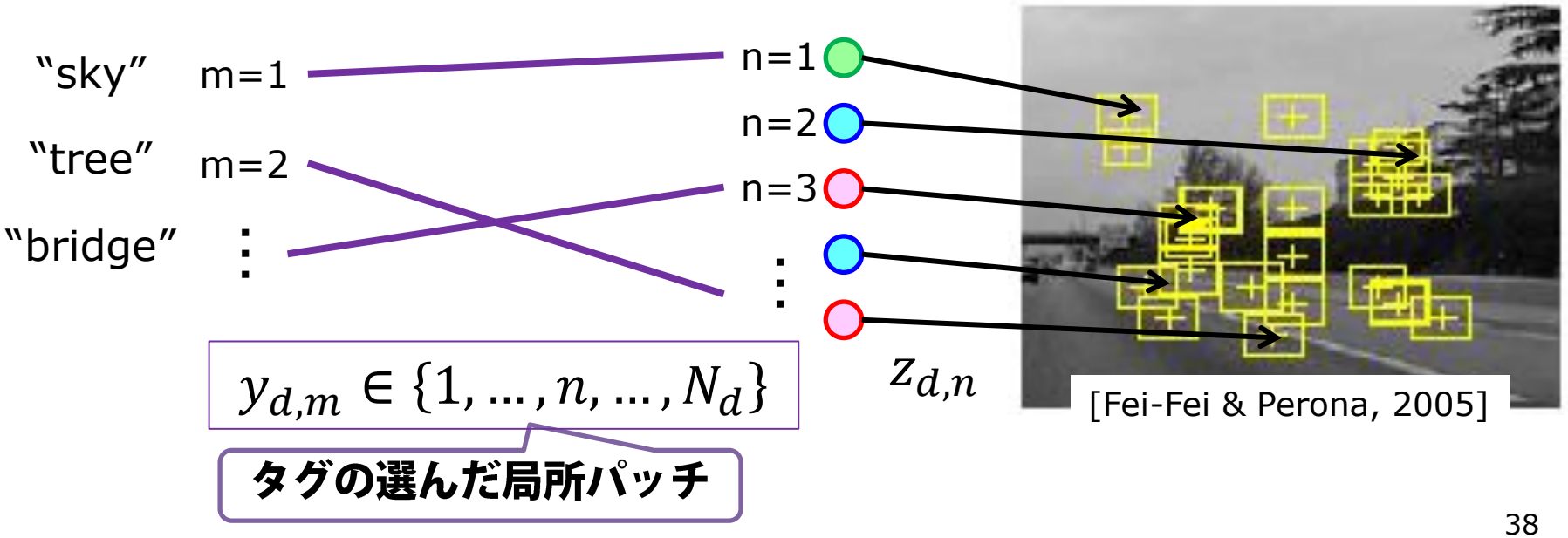
for クラスラベル c = 1, 2, …, C

class parameter for soft-max $\boldsymbol{\eta}_c$

パッチ-タグ対応

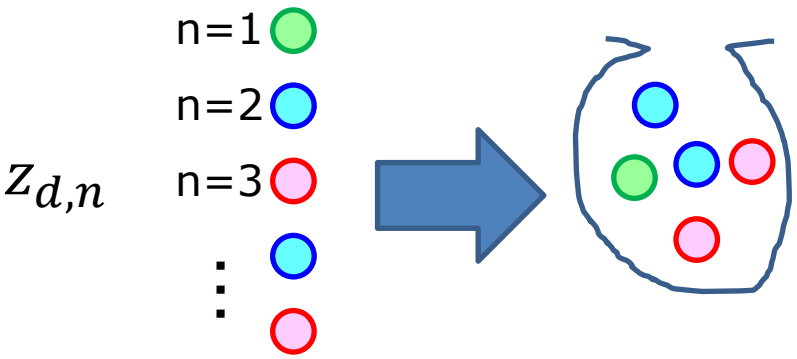
- 局所パッチとタグの対応は、何もモデルが立てられないので一様ランダムに決めます

$$y_{d,m} \sim \mathrm{Uniform}(\{1,2,\cdots,N_d\})$$



multi-class sLDAの クラスラベルモデル

- クラスラベル用のパラメータ η を使って、soft-maxによる多項分布サンプリング



文書 d の経験トピック分布

$$\bar{z}_d = \frac{1}{N_d}\sum_n z_{d,n}$$

$c_d =$ “mountain” “water” “city” “street”

$$c_d|\{\boldsymbol{\eta}_c\},\{z_{d,n}\} \sim \text{soft-max}\left(\{\boldsymbol{\eta}_c\},\{z_{d,n}\}\right)$$

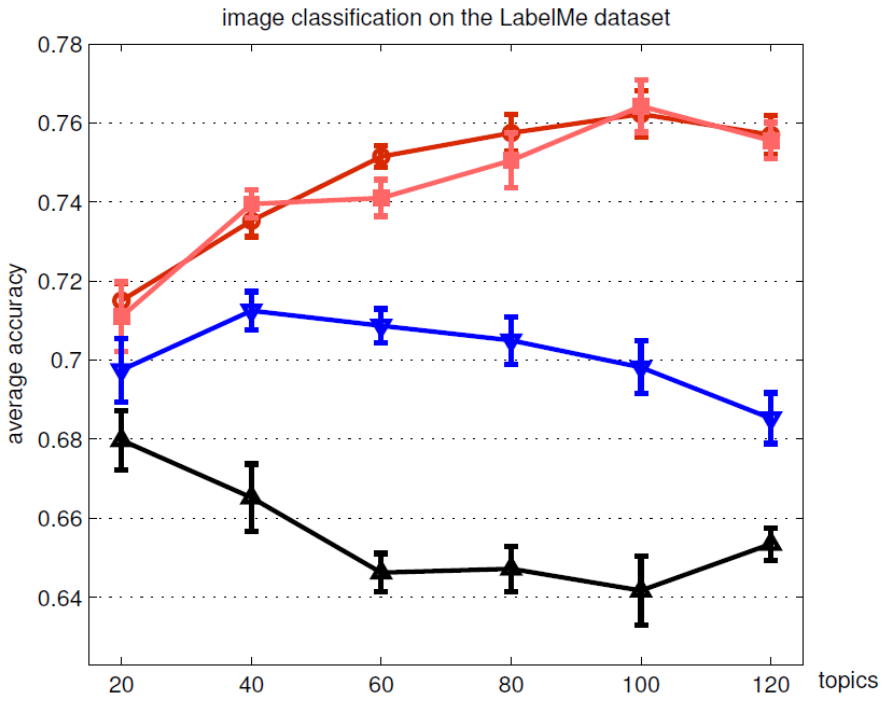
$$p\left(c_d = c|\boldsymbol{\eta}_c,\{z_{d,n}\}\right) = \frac{\exp(\boldsymbol{\eta}_c{}^T\bar{\boldsymbol{z}}_d)}{\sum_{l=1}^{C}\exp(\boldsymbol{\eta}_l{}^T\bar{\boldsymbol{z}}_d)}$$

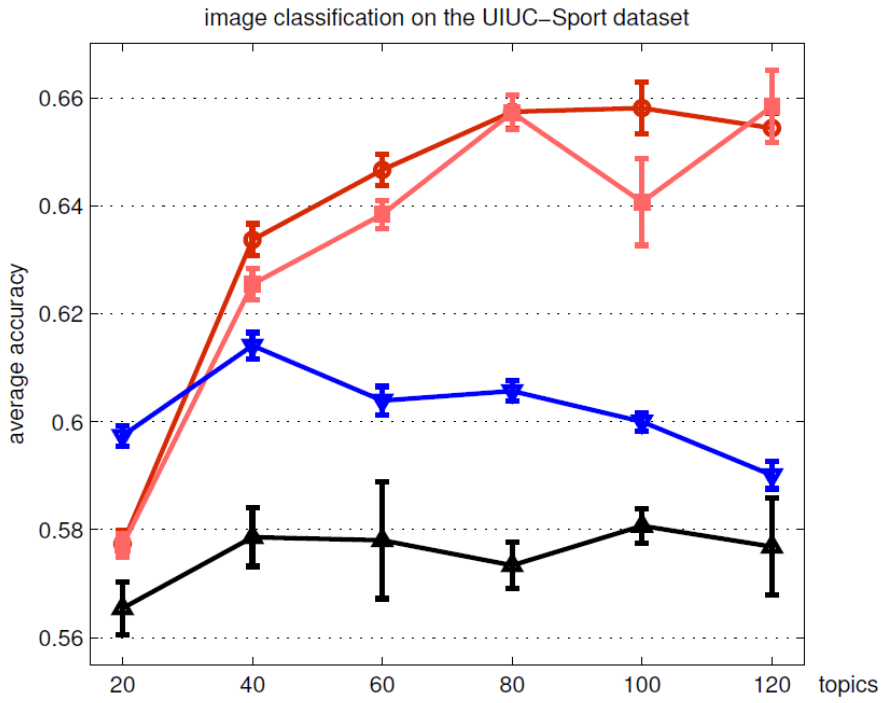
パラメータ、隠れ変数の推定

- 論文では変分ベイズ(VB)による推定法が紹介されています
- ☹ 少々複雑になります・・・

未知画像の識別

- 学習し終わったモデルに対して、未知画像 d のクラスラベル c とタグ w を推定します
- まず、画像 d の変分事後トピック分布を計算します
- これを用いて、cとwを選びますが、正確な計算は不可能なので、色々近似が必要です。
- 詳しくは論文を読んでください





[Wang, 2009]

赤: 提案法(Multi-class sLDA with annotation)

ピンク: 提案法(Multi-class sLDA)

青: Fei-Fei & Perona, 2005 (各クラスごとに個別にLDAを学習)

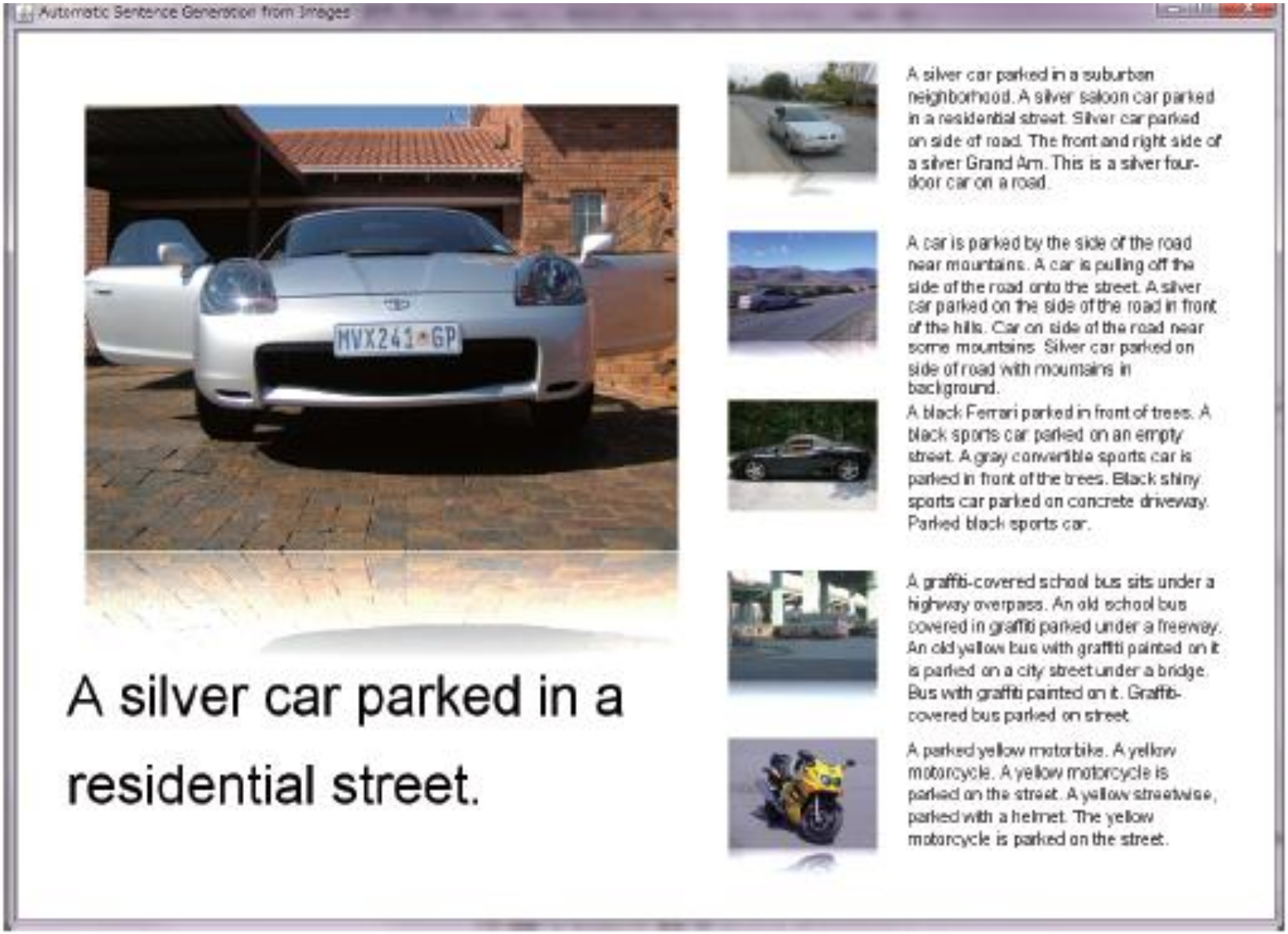
黒: Bosch, 2006 (LDA + kNN)

Correct classification with predicted annotations		**Incorrect classification (correct class) with predicted annotations**	
highway car, sign, road			*coast (highway)* car, sand beach, tree
inside city buildings, car, sidewalk			*street (inside city)* window, tree, building occluded
tall building trees, buildings occluded, window			*inside city (tall building)* tree, car, sidewalk

[Wang, 2009]

street tree, car, sidewalk			*highway (street)* car, window, tree
forest tree trunk, trees, ground grass			*mountain (forest)* snowy mountain, tree trunk
coast sand beach, cloud			*open country (coast)* sea water, buildings
mountain snowy mountain, sea water, field			*highway (mountain)* tree, snowy mountain

[Wang, 2009]


Automatic Sentence Generation from Images
A silver car parked in a residential street.
A silver car parked in a suburban neighborhood. A silver saloon car parked in a residential street. Silver car parked on side of road. The front and right side of a silver Grand Am. This is a silver four-door car on a road.
A car is parked by the side of the road near mountains. A car is pulling off the side of the road onto the street. A silver car parked on the side of the road in front of the hills. Car on side of the road near some mountains. Silver car parked on side of road with mountains in background.
A black Ferrari parked in front of trees. A black sports car parked on an empty street. A gray convertible sports car is parked in front of the trees. Black shiny sports car parked on concrete driveway. Parked black sports car.
A graffiti-covered school bus sits under a highway overpass. An old school bus covered in graffiti parked under a freeway. An old yellow bus with graffiti painted on it is parked on a city street under a bridge. Bus with graffiti painted on it. Graffiti-covered bus parked on street.
A parked yellow motorbike. A yellow motorcycle. A yellow motorcycle is parked on the street. A yellow streetwise, parked with a helmet. The yellow motorcycle is parked on the street.

[Ushiku, 2011]

まとめ: Multi-class sLDA with annotation

- BoVWベースのトピックモデルで、画像のクラスラベル、さらにタグ（アノテーション）付与まで同時にモデル化

- 画像のトピック分布から直接クラス識別
- 局所パッチからタグを生成

- アノテーション無しのMulti-class sLDA自体も多クラス識別モデルとして新規性あり
- ただのBoVW-LDAよりもはるかに高い性能

他の（動）画像応用

- BoVW: 多すぎてフォローできません

- Niebles et al., “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words”, Int. J. Computer Vision, Vol. 79, pp. 299-318, 2008.

- Rodriguez et al., “Tracking in Unstructured Crowded Scenes”, in Proc. ICCV, 2009.

- Sivic et al., “Unsupervised Discovery of Visual Object Class Hierarchies”, in Proc. CVPR, 2008.

引用及び参考文献

- [Blei, 2003] Blei et al, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.
- [Shimizu, 2008] Shimizu et al, “Super-Resolution from Image Sequence under Influence of Hot-Air Optical Turbulence”, in Proc. CVPR, 2008.
- [Viola & Jones, 2001] Viola and Jones, “Robust Real-time Object Detection”, in Proc. CVPR, 2001.
- [Lowe, 2004] Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision, Vol. 60, pp. 91-110, 2004.
- [藤吉, 2007] "Gradientベースの特徴抽出 - SIFTとHOG - ", 情報処理学会 研究報告 CVIM 160, pp. 211-224, 2007.
- [Fei-Fei & Perona, 2005] Fei-Fei and Perona, “A Bayesian Hierarchical Model for Learning Natural Scene Categories”, in Proc. CVPR. 2005.

引用及び参考文献

- [Treisman & Gelade, 1980] Treisman and Gelade, “A feature-integration theory of attention”, *Cognitive Psychology*, Vol. 12. pp. 97–136, 1980.
- [Li, 2002] Li et al., “Natural scene categorization in the near absence of attention”, PNAS, Vol. 99, No. 14. pp.9596–9601, 2002.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.
- [Wang, 2009] Wang et al., “Simultaneous Image Classification and Annotation”, in Proc. CVPR, 2009.
- [Ushiku, 2011] Ushiku et al., “Understanding Images with Natural Sentences”, in Proc. ACM Multimedia, 2011.

NTT

トピックモデルの応用：音声・音響データ

NTT コミュニケーション科学基礎研究所
石黒　勝彦

2013/01/15-16 統計数理研究所　会議室1

このスライドの“トピック”

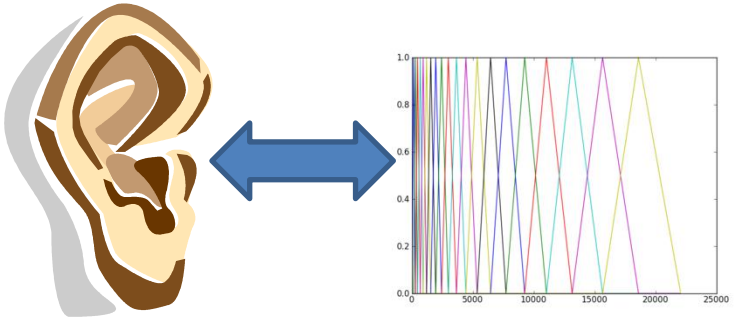
- 画像認識系から少し遅れますが、最近では音声・音響データに対してもトピックモデルが利用されるようになっています

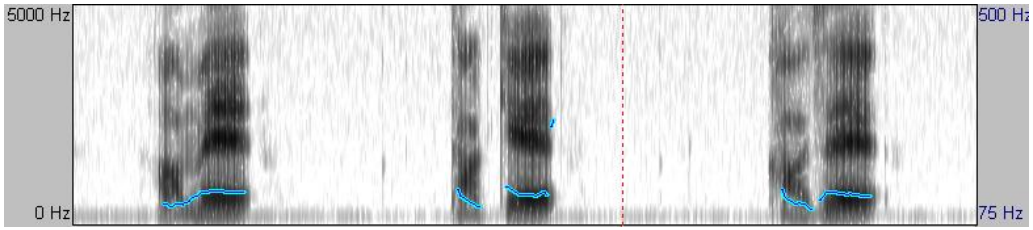
音声・音響信号で考えるべき問題

- 1. どの特徴量を利用するか？
- 2. 時系列性をどう扱うか？

音声・音響信号からは多彩な特徴量が抽出できます

- どの特徴量を利用して、どうやってBoW形式に変換するかを検討する必要があります
  - MFCC: 音声認識などで広い範囲で利用される
  - F0: 発話のイントネーションやメロディを表現





MFCC: 人間の
音声知覚を反映した(とされる)特徴

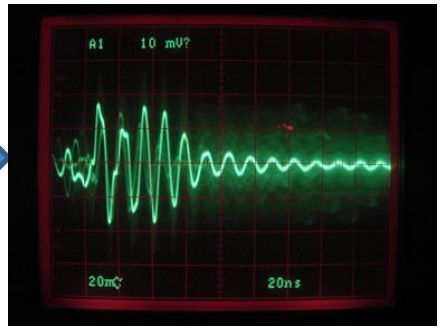
F0: 波形の基本周波数。ピッチ。

http://d.hatena.ne.jp/aidiary/20120225/1330179868
https://sites.google.com/site/utsakr/Home/courses/tsukuba2010/contents/13-f0

音声・音響信号は
複雑な時系列信号です

- マルコフ性を仮定する時系列モデルを利用するのが王道ですが、その必要があるかどうかの検討も必要です

$$f(t) = \int g(t-\tau)h(\tau)d\tau$$



http://plaza.rakuten.co.jp/DFLYRoom/diary/200808030000/

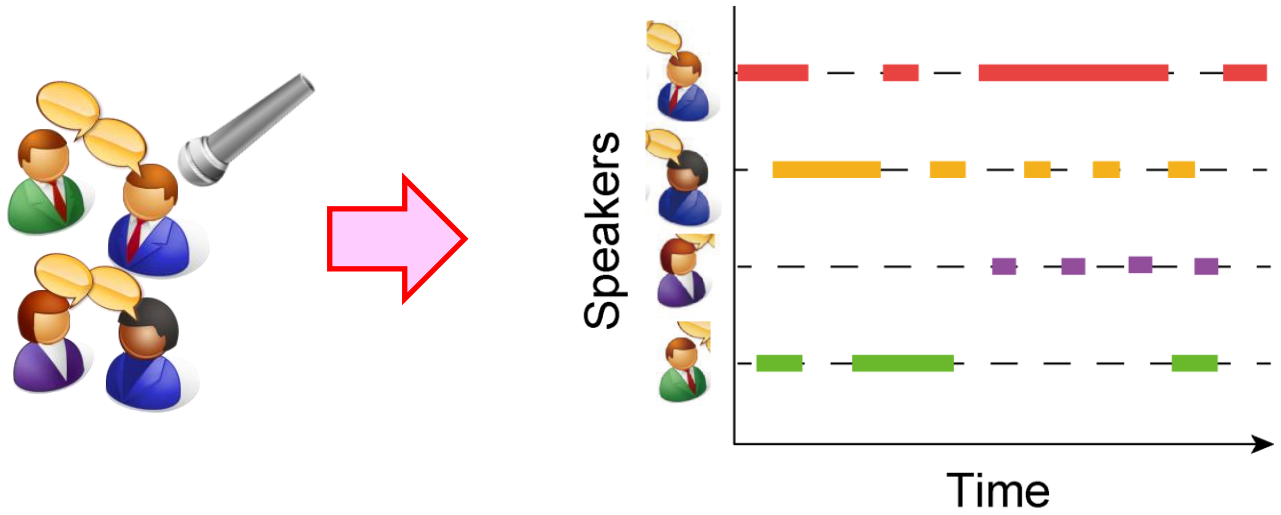
Topic Model for speaker diarization

[Ishiguro, 2012]

Ishiguro et al. ,
"Probabilistic Speaker Diarization with Bag-of-Words Representations of Speaker Angle Information",
IEEE Trans. ASLP, Vol. 20(2), pp. 447-460, 2012.

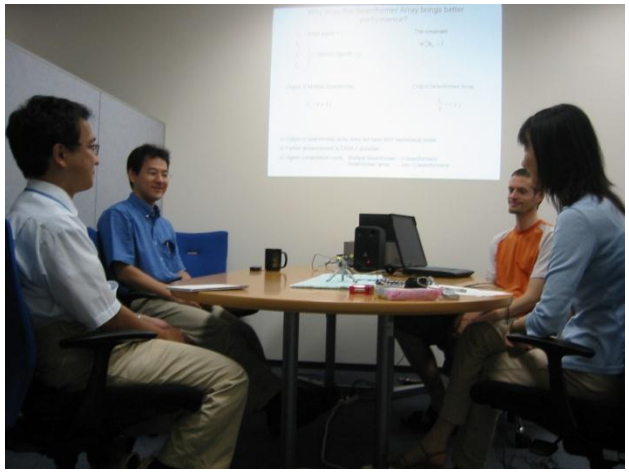
speaker diarization
“誰がいつ発話したか”

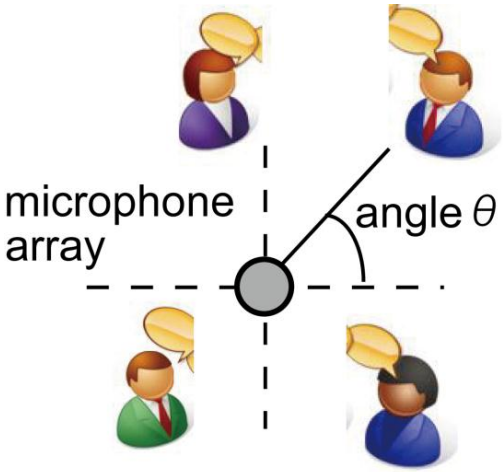
- 複数の音源があるときに、各音源がいつ信号を発信したかを決定
- 応用範囲: 会議の自動議事録作成、テレビ電話における発話者音声強調、ロボットと人間のインタラクションなど



会議状況のdiarization

- テーブルにマイクを置いて、会議状況をdiarizationします
- 一般に何人の話者がどこに座るかは事前にわかりません → 話者は潜在的な隠れ要素です
- その時々によって発話者が代わります → 各話者の発話状況は時間変化します



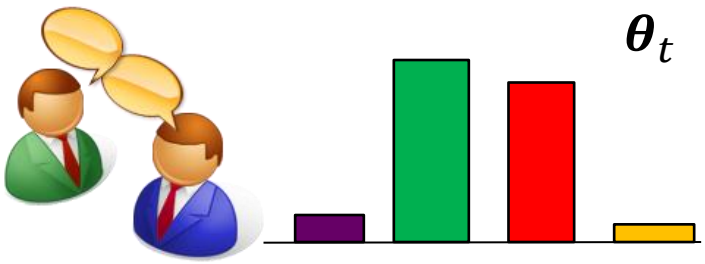


トピックモデルによるdiarization

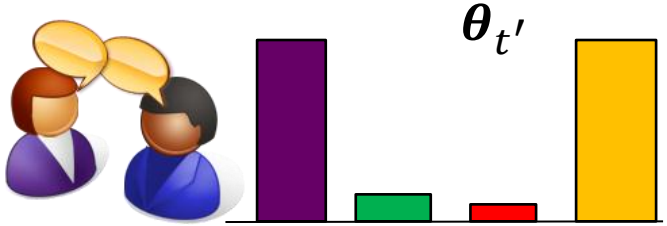
☺ diarizationタスクは自然にトピックモデルで記述できます

- 時刻 ＝ 文書と考えると、各時刻の発話は複数の潜在トピック ＝ 話者で表現できます
- トピック（話者）はわからないので推定します
- トピック分布に発話状況が反映されます

時刻 t



時刻 t′



提案法の概要

- diarizationとトピックモデルの共通点に気付いたことで、「話者＝トピック」と「各時刻の発話状態＝文書のトピック分布」を同時に推定できます

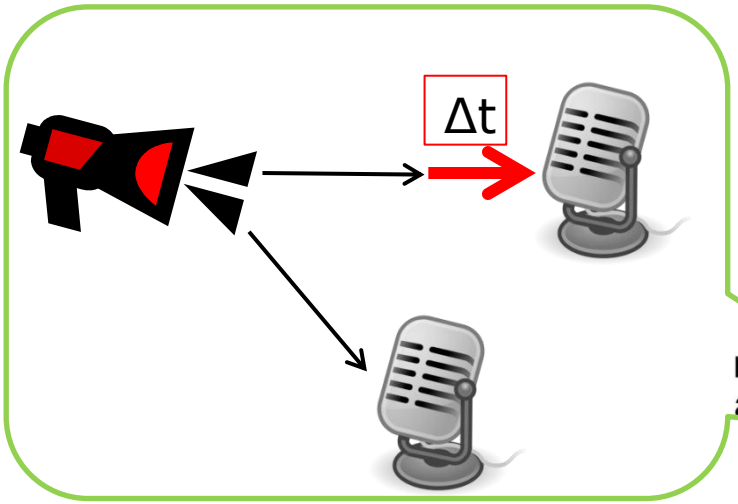
- diarizationに対するベイジアンモデルを提案できます

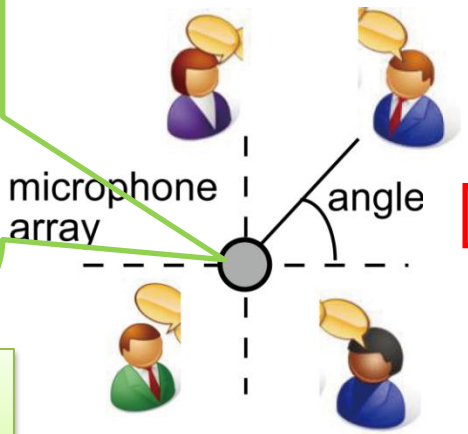
提案法のアイデア

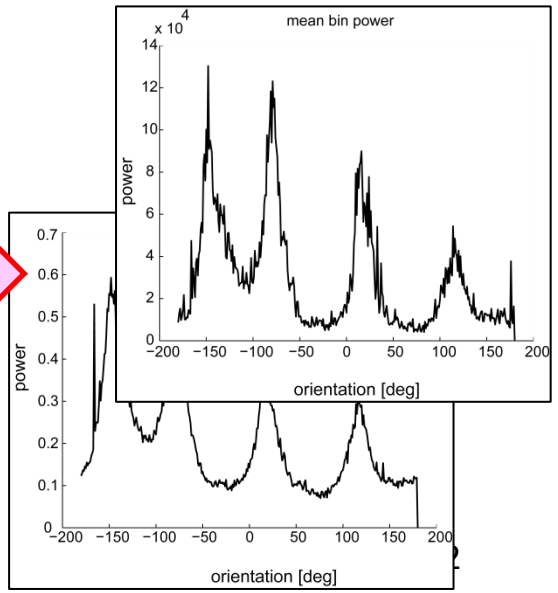
- 考えるべき2つの問題に以下のように対応します
- 特徴量: 方向情報 (DOA) → Bag of Angle Words
- 時系列性: 非定常な話者分布変化 → トピック分布の線形補間モデル

話者はどこにいるのか: DOAクラスタリング [cf. Araki, 2008]

- DOA: 音の聞こえてくる方向の特徴量
- クラスタリングによって、「話者がどこにいるのか」を推定できることが分かっています



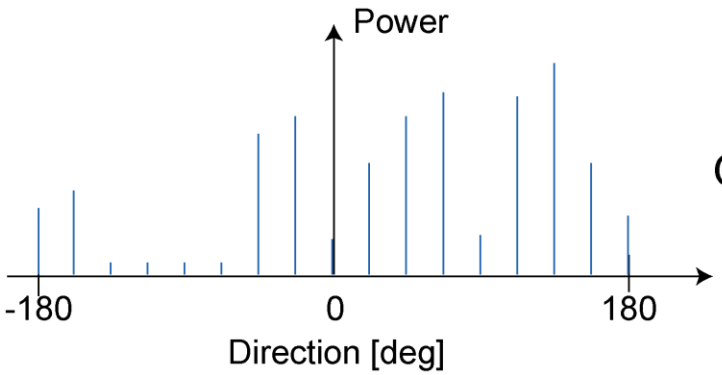




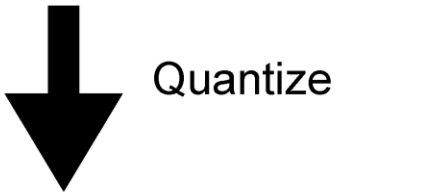
信号の受信時間差とマイク配置から音源位置が計算できる

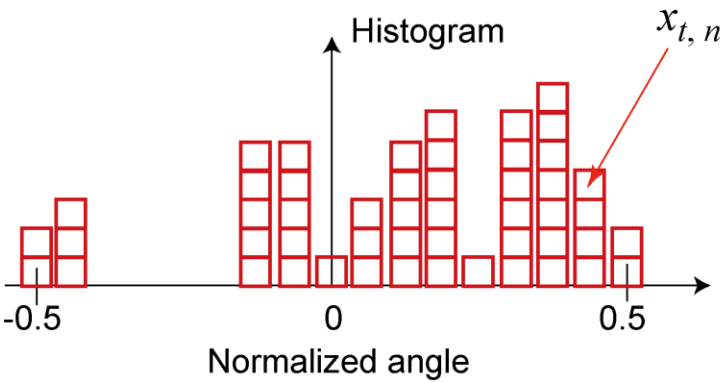
特徴量: Bag of Angle Words

☺ DOA特徴量を離散化、トピックモデルに使えるようにします



Observed DOA power





Bag of Angle words

[Ishiguro, 2012]

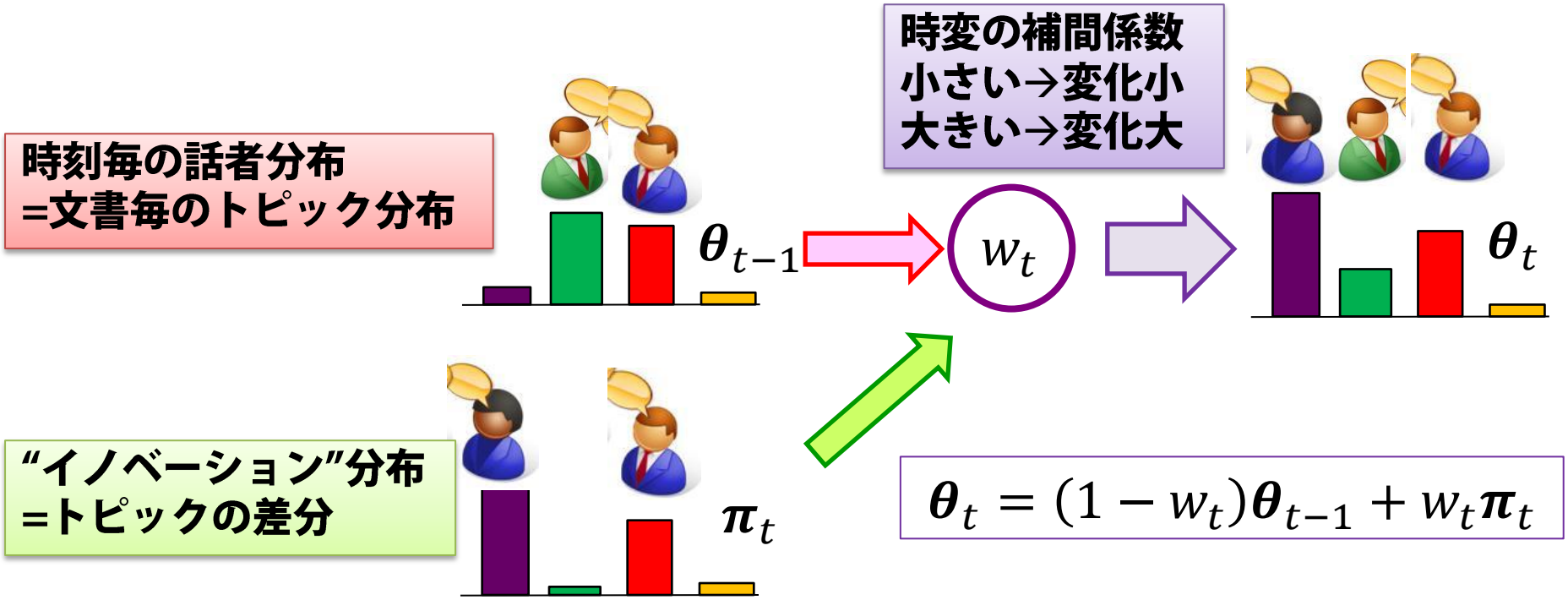
非定常な話者分布変化

- 時間連続性: ミリ秒単位の時間ステップでは、話者の発話分布は変わりません
- 時間非連続性: 発言を受けての応答など、会議の流れにそって話者分布が変化します (turn-taking)

- つまり、話者の発話状態の変化自体が非定常になっています

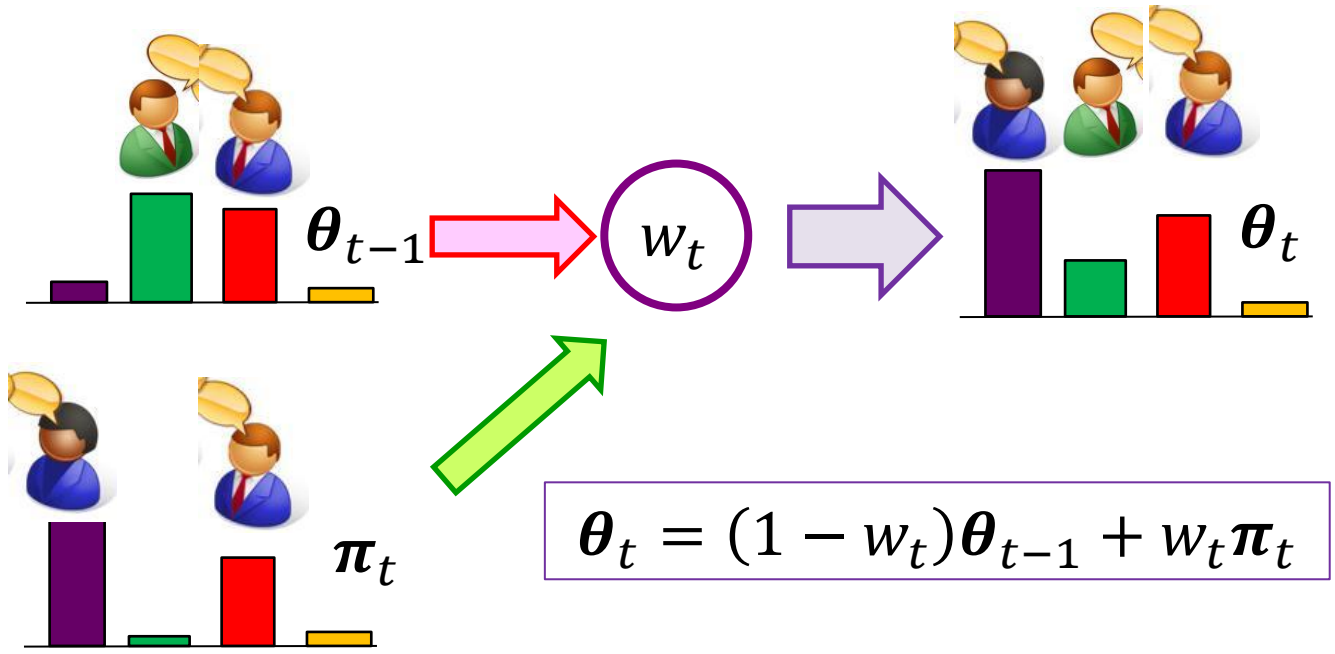
時系列性: トピック分布の線形補間モデル

- 話者分布の時間変化の非定常性を表すために、時変の補間係数を導入します



時系列性: トピック分布の線形補間モデル

- ☺簡単な線形モデルによるLDAの時間発展モデル
- ☺小規模～大幅な話者変化を表現可能
  - 前時刻との依存度をw_tで制御する



相互に独立なモデルへの変換

- 各時刻の話者分布θは、時刻ごとに独立なπの組み合わせで表現できます
- ☺ マルコフ性が消えて推論が簡単になります

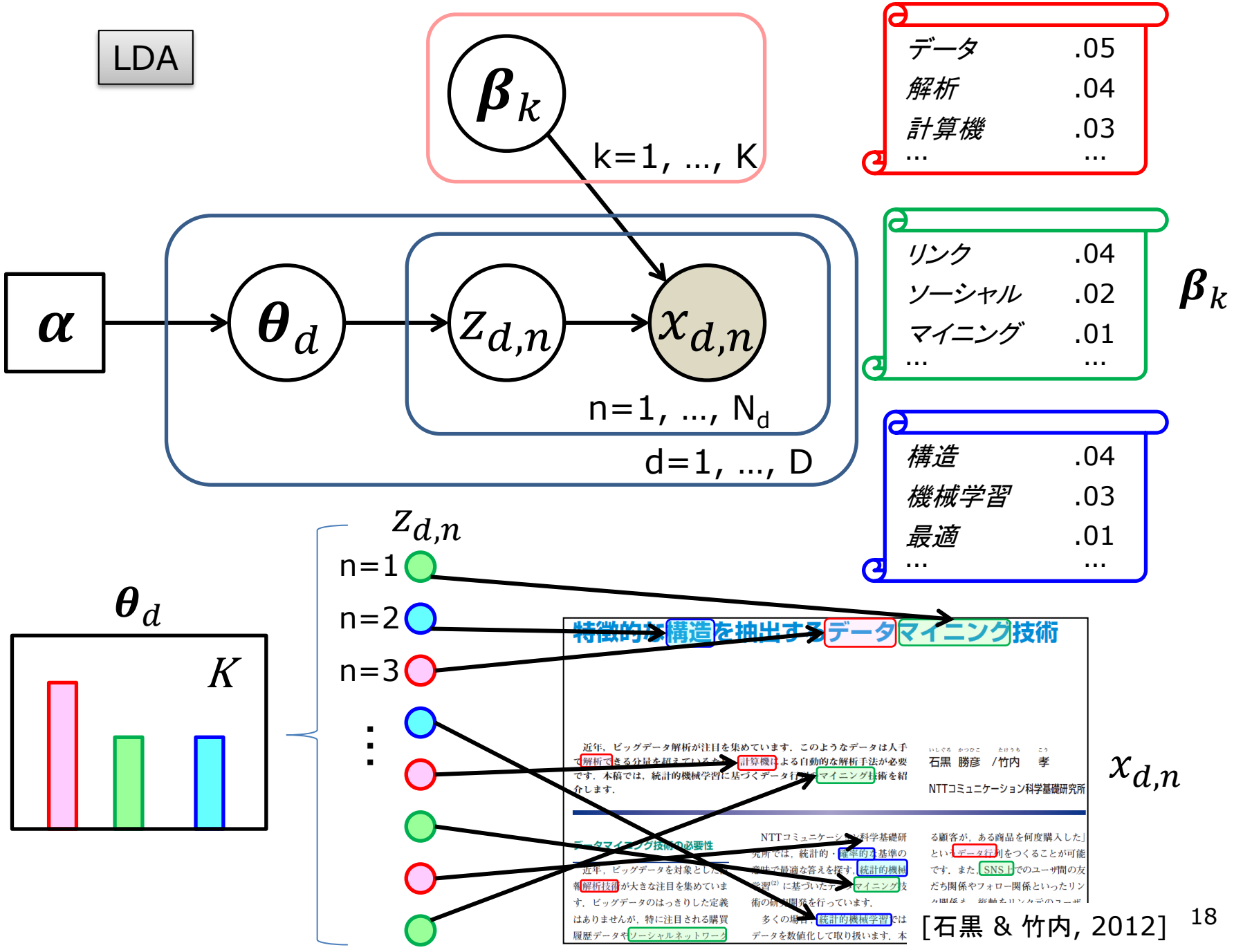
$$\boldsymbol{\theta}_t = (1 - w_t)\boldsymbol{\theta}_{t-1} + w_t\boldsymbol{\pi}_t$$

$$= (1 - w_t)\{(1 - w_{t-1})\boldsymbol{\theta}_{t-2} + w_{t-1}\boldsymbol{\pi}_{t-1}\} + w_t\boldsymbol{\pi}_t$$

$$\cdots$$

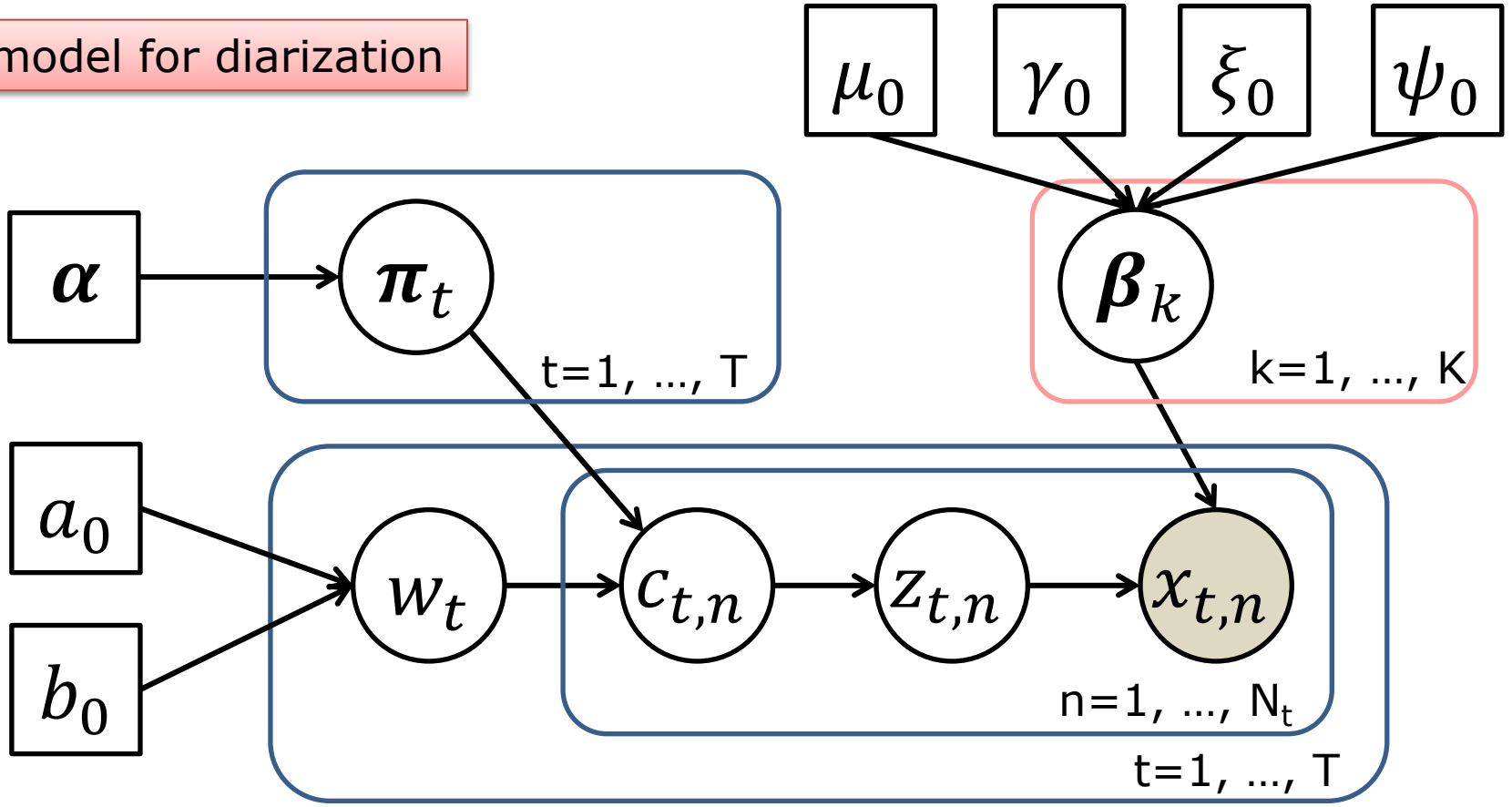
$$= \sum_{l=1}^{t} v_{tl}\boldsymbol{\pi}_l \qquad v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$

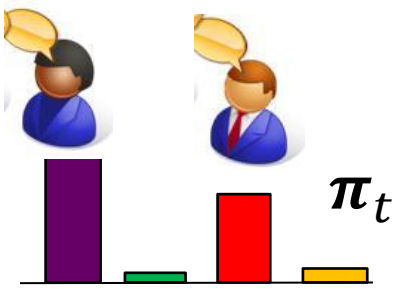
LDA



[石黒 & 竹内, 2012]

Topic model for diarization

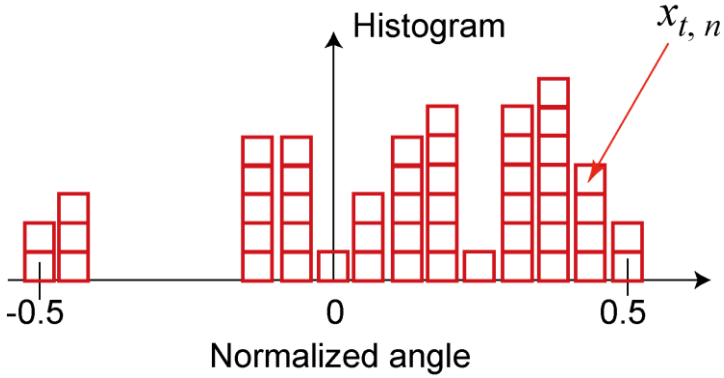




$$\boldsymbol{\theta}_t = (1 - w_t)\boldsymbol{\theta}_{t-1} + w_t\boldsymbol{\pi}_t$$

$$= \sum_{l=1}^{t} v_{tl}\boldsymbol{\pi}_l$$

$$v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$



生成モデル

for 時間t = 1, 2, …, T

"innovation" topic proportion $\boldsymbol{\pi}_t|\boldsymbol{\alpha}\sim\mathrm{Dir}(\boldsymbol{\alpha})$

interpolation factor $w_t|a_0,b_0\sim\mathrm{Beta}(a_0,b_0)$

for l = 1, 2, …, t

$$v_{tl}=w_l\prod_{m=l+1}^{t}(1-w_m)$$

for 単語n = 1, 2, …, $N_{t,d}$

for speaker (topic) k = 1, 2, …, K

topic-Angle word proportion

$$\boldsymbol{\beta}_k|\mu_0,\gamma_0,\xi_0,\psi_0\sim\mathrm{NormalGamma}(\mu_0,\gamma_0,\xi_0,\psi_0)$$

生成モデル

for 時間t = 1, 2, …, T

$$\boldsymbol{\pi}_t|\boldsymbol{\alpha}\sim\mathrm{Dirichlet}(\boldsymbol{\alpha}) \qquad v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$

for 単語n = 1, 2, …, $N_{t,d}$

innovation topic dist.-word assignment

$$c_{t,n}|\boldsymbol{v}_t\sim\mathrm{Mult}(\boldsymbol{v}_t)$$

speaker-angle word assignment

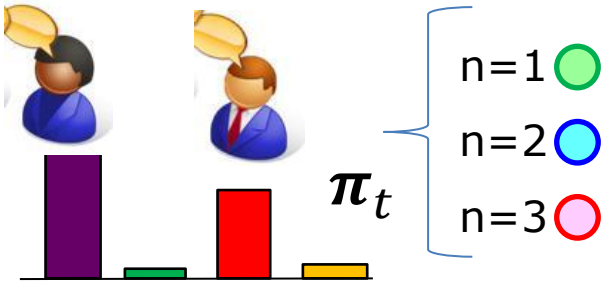
$$z_{t,n}|c_{t,n},\{\boldsymbol{\pi}_t\}\sim\mathrm{Mult}\left(\boldsymbol{\pi}_{c_{t,n}}\right)$$

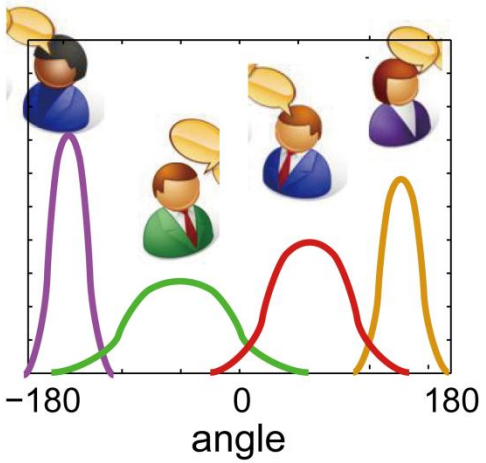
Angle word observation

$$x_{t,n}|z_{t,n},\left\{\boldsymbol{\beta}_{t,k}\right\}\sim\mathrm{N}\left(\boldsymbol{\beta}_{t,z_{d,n}}\right)$$

混合分布による疑似的なBag of Angle Words

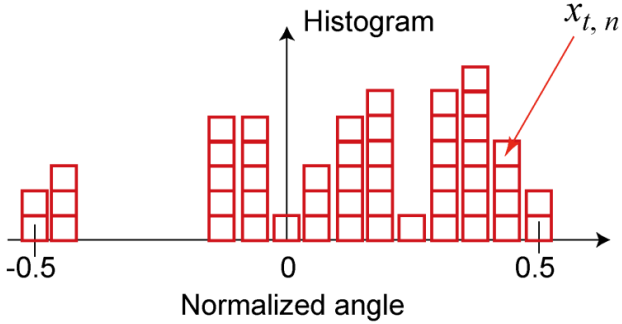
- Angle wordの値（角度・位置）には意味があるのでNormalから生成します





$$z_{t,n}|c_{t,n},\{\boldsymbol{\pi}_t\}\sim\mathrm{Multi}\left(\boldsymbol{\pi}_{c_{t,n}}\right) \qquad \boldsymbol{\beta}_k=\{\mu_k,\sigma^2\}$$

$$x_{t,n}|z_{t,n},\{\boldsymbol{\beta}_{t,k}\}\sim\mathrm{N}\left(\boldsymbol{\beta}_{t,z_{d,n}}\right)$$



話者数の自動推定

- ☺ 自動的に話者数も推定できます
- 発話していない話者に対応するトピックの重み $z_{t,n,k}$ は学習と共に0に近づきます
- 従って"存在しない"話者に対応するトピック k' は以下を満たすかで判定できます

"存在する"話者に対応するトピック k

$$\frac{1}{K} \leq \sum_{t,n} z_{t,n,k}$$

"存在しない"話者に対応するトピック k'

$$\frac{1}{K} > \sum_{t,n} z_{t,n,k'}$$ (実際にはほぼ0になります)

隠れ変数とパラメータの推定

- 論文では変分ベイズ法(VB-EM)による解法が提案されています

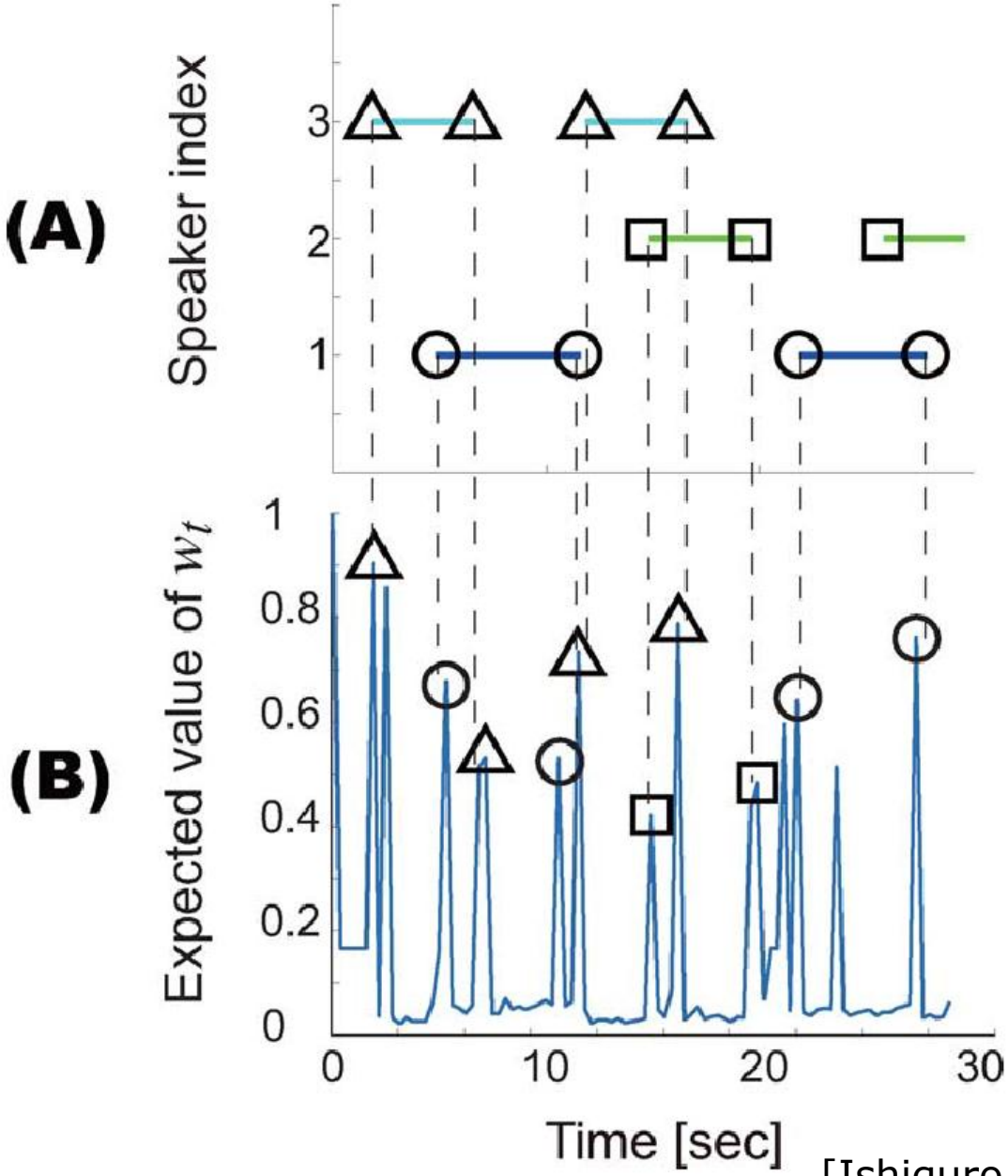
- 具体的な式は煩雑になるので省略します。必要な方は論文をチェックしてください

オンライン学習が自然に導かれます

- v_{tl} の定義から、θ_t (時刻tの話者分布)の学習には昔の分布の情報はほとんど影響しません
- すなわち、直近の情報だけを用いたオンライン(逐次)学習が可能となります

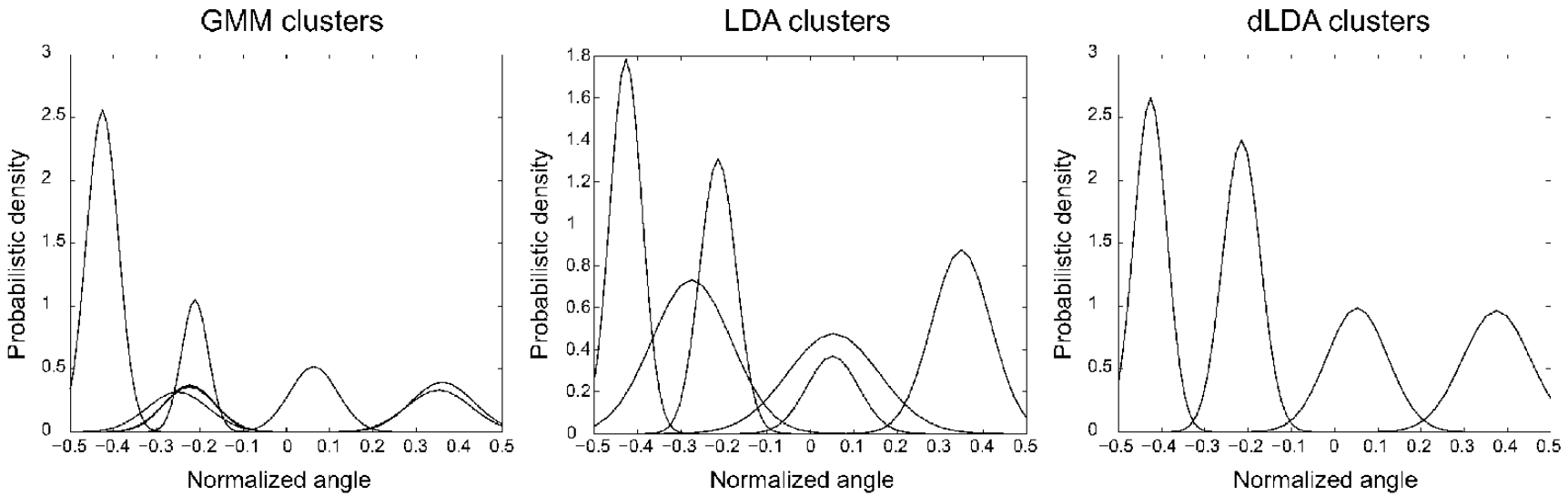
$$\boldsymbol{\theta}_t = (1 - w_t)\boldsymbol{\theta}_{t-1} + w_t\boldsymbol{\pi}_t$$

$$= \sum_{l=1}^{t} v_{tl}\boldsymbol{\pi}_l \qquad v_{tl} = w_l \prod_{m=l+1}^{t} (1 - w_m)$$



[Ishiguro, 2012]

話者4人のデータからのspeaker (topic)学習結果



[Ishiguro, 2012]

Dataset	[13]	[8]	GMM	LDA	dLDA(proposed)
CP1	21.9	(37.1)	55.8	32.7	**21.7**
*CP2	25.0	(35.8)	32.8	24.5	**19.7**
DC	**29.9**	(47.0)	60.6	48.0	31.0
CN	34.3	(56.4)	57.3	48.5	**34.1**
IS1000a	41.9	46.26	35.2	76.9	**32.2**
IS1001a	31.7	30.58	26.7	33.8	**23.7**
IS1001c	32.2	**12.07**	68.2	40.7	27.2
IS1006d	64.3	**54.56**	67.4	69.9	69.7
IS1008a	**13.1**	**5.13**	77.8	65.3	62.7
IS1008b	19.6	**16.47**	57.8	55.9	23.1
IS1008c	22.6	**12.09**	30.1	30.8	20.4
*IS1008d	**15.8**	20.83	21.9	32.1	**13.6**

[Ishiguro, 2012]

まとめ: Topic model for speaker diarizatoin

- トピックモデルにより、speaker diarization タスクを解決できます
- 簡単な時間発展モデルで話者の切り替わり(turn-taking)も自然にモデル化
- state-of-the-artの作りこんだモデルとcomparableの性能

その他の音声・音響データ応用

- Ohtsuka et al., “Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution”, in Proc. AAAI, 2012.

- Yoshii and Goto, “A Nonparametric Bayesian Multiple Analyzer Based on Infinite Latent Harmonic Allocation”, IEEE Trans. ASLP, Vol. 20(3), pp. 717-730, 2012.

引用及び参考文献

- [Ishiguro, 2012] Ishiguro et al. , “Probabilistic Speaker Diarization with Bag-of-Words Representations of Speaker Angle Information”, IEEE Trans. ASLP, Vol. 20(2), pp. 447-460, 2012.
- [Araki, 2008] Araki et al., “A DOA based Speaker Diarization System for Real Meetings”, in Proc. Joint Workshop Hndns-Free Speech Comm. Microphone Arrays, 2008.
- [石黒 & 竹内, 2012] 石黒, 竹内, “特徴的な構造を抽出するデータマイニング技術”, NTT技術ジャーナル, Vol. 24, No. 9, 2012.