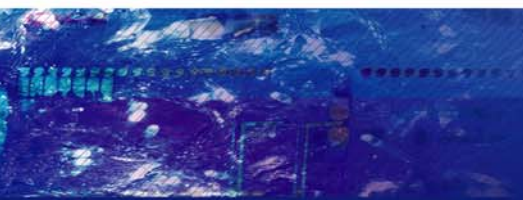
NTT

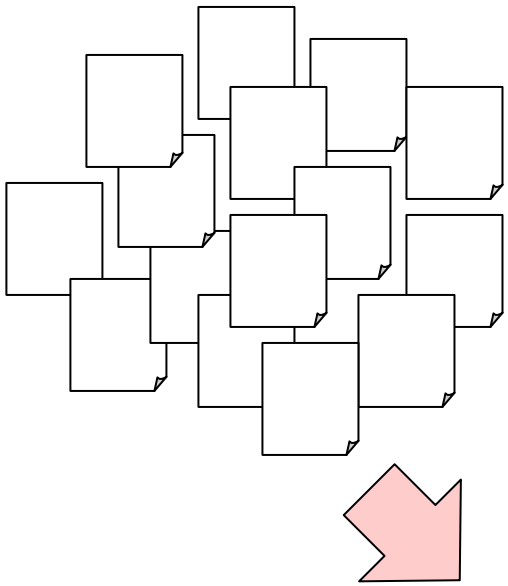
NTT Communication Science Laboratories

隠れた構造をあぶりだす
–トピックモデルに基づく潜在意味解析–

NTTコミュニケーション科学基礎研究所
岩田具治



大量の文書群に隠れた構造



同じ文書で現れやすい
単語のまとまり（トピック）がある

スポーツ	政治	経済
野球	国会	株
勝敗	首相	金融
ボール	法案	不景気
試合	大統領	資本
球場	選挙	業績
：	：	：

各文書は少数のトピックを持つ

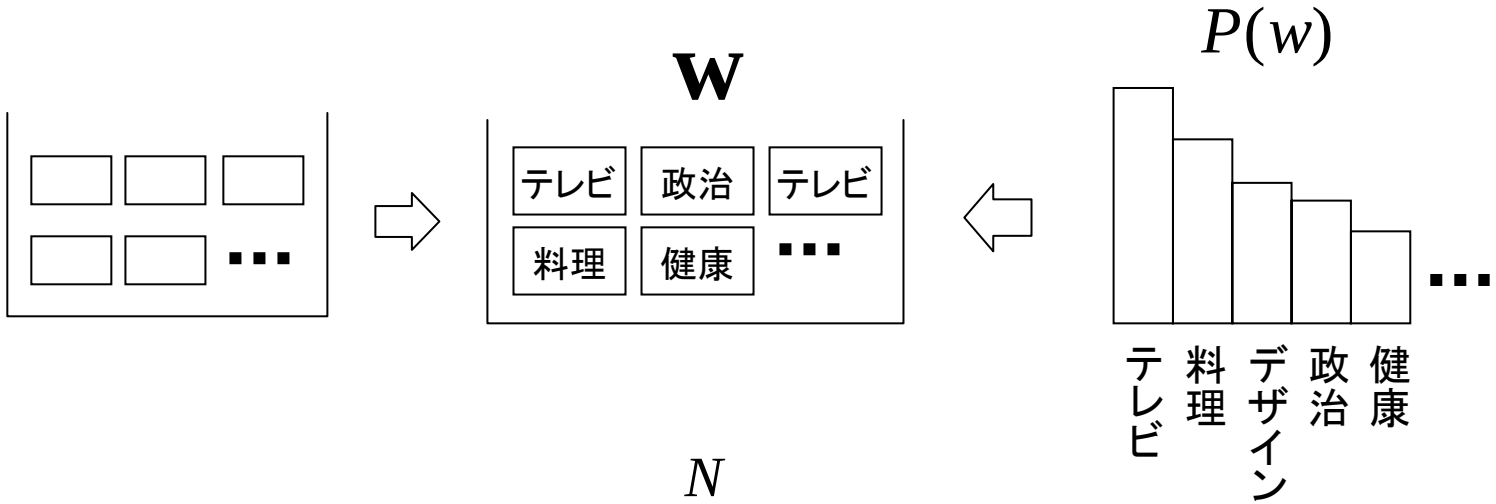
中日がナゴヤドームで行われた試合に勝利…	衆議院選挙で〇〇党が大躍進…	ねじれ国会で景気対策足踏み…	▪▪▪▪▪
スポーツ	政治	政治＋経済	

トピックモデルとは

- 文書の生成過程を確率的に表現したモデル
- 様々な離散データで有効性が確認
  - 文書、購買履歴、グラフ構造、画像、音楽
- 幅広い応用範囲
  - 情報検索、画像認識、推薦システム、音声認識
- 拡張が容易
  - 確率論の枠組みで異種情報をきれいに統合できる
- 実装が簡単

多項分布と
混合多項分布と
トピックモデル

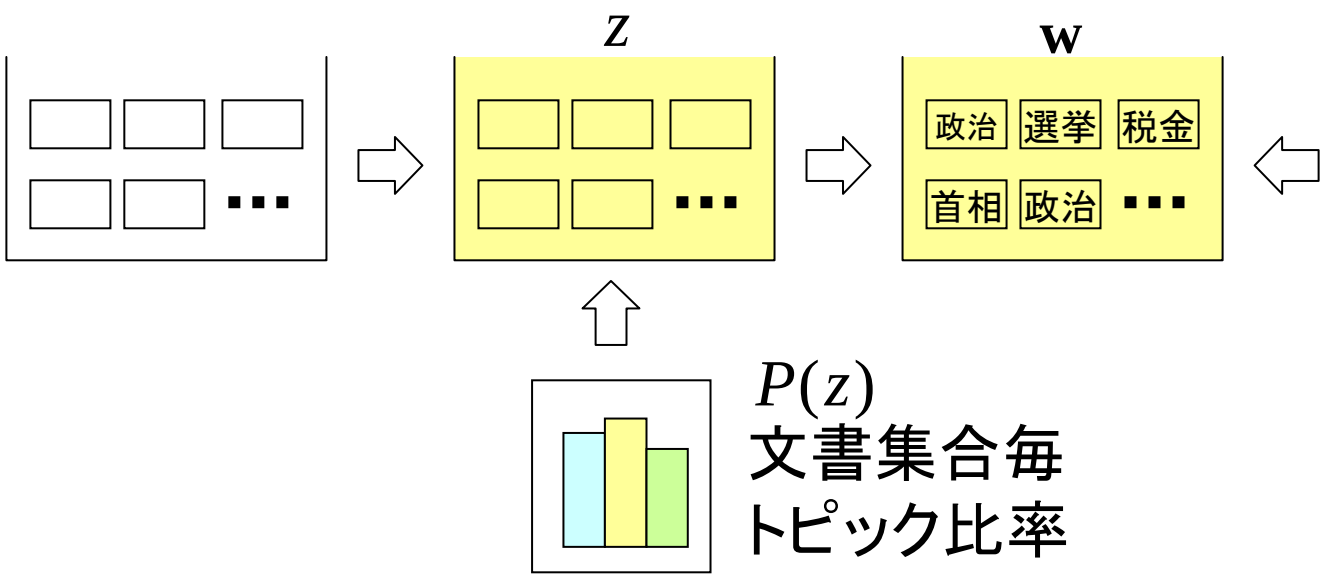
多項分布



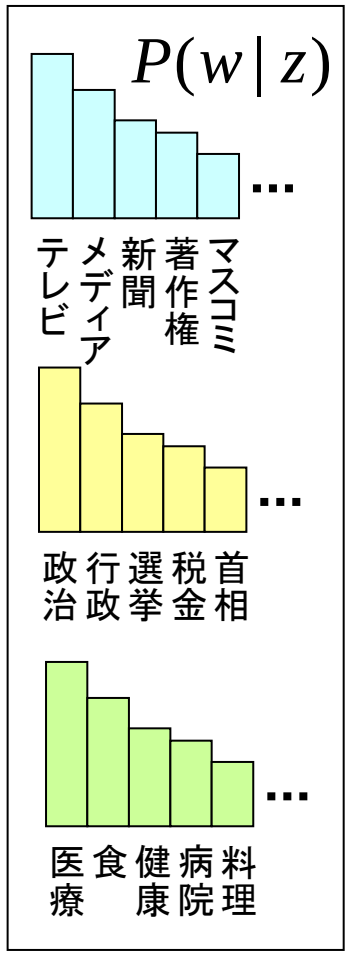
$$P(\mathbf{w}) = \prod_{n=1}^{N} P(w_n)$$

- 全文書の単語が同一の分布から生成

混合多項分布



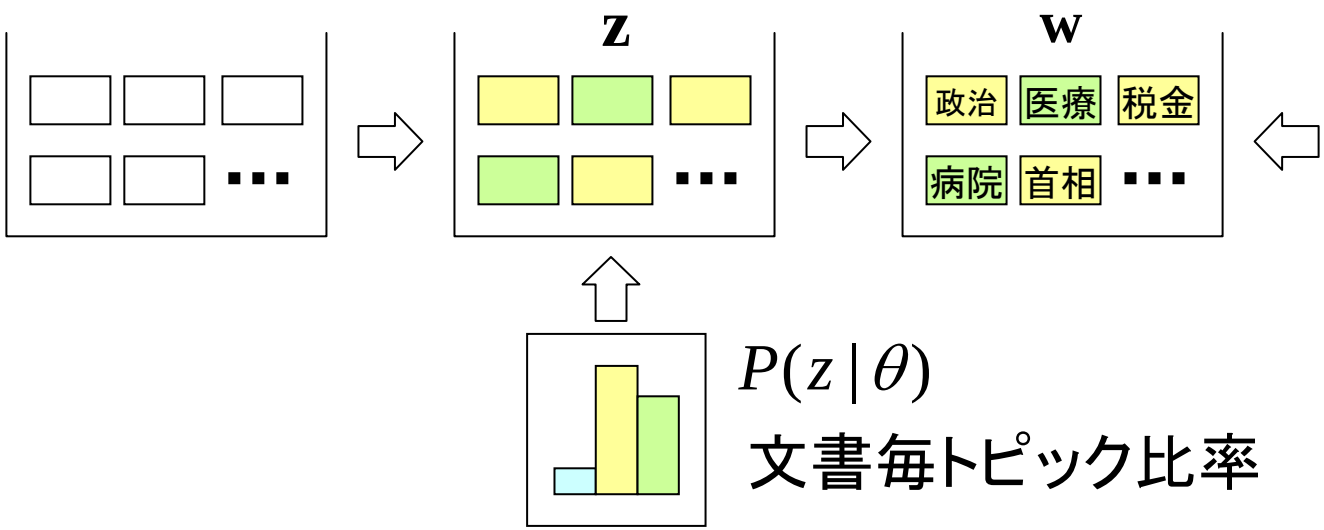
$P(z)$
文書集合毎
トピック比率



$$P(\mathbf{w}) = \sum_{z} P(z) \prod_{n=1}^{N} P(w_n \mid z)$$

- 文書毎にトピックを選択
- 1文書の単語が同一の分布から生成

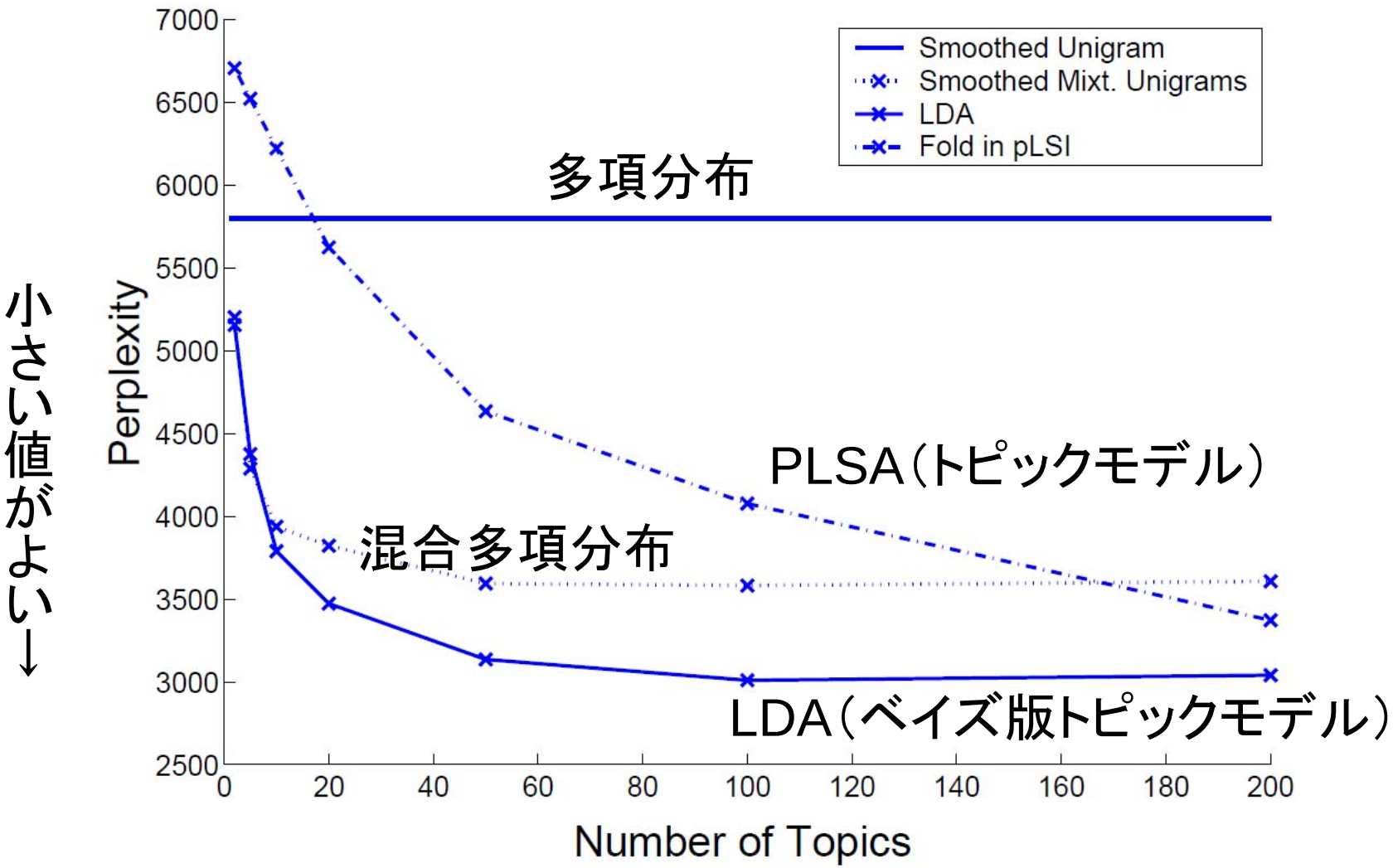
トピックモデル



$$P(\mathbf{w}) = \prod_{n=1}^{N} \sum_{z} P(z|\theta) P(w_n|z)$$

- 単語毎にトピックを選択
- １文書の単語が複数の分布から生成

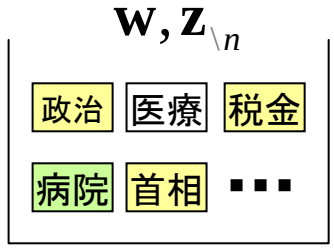
文書データを用いた比較実験



D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003

トピックモデル実装

- 入力
  - 文書データ、トピック数K
- 初期化
  - 各単語にランダムにトピック（1～K）を割り当てる
- あとは、各単語のトピックを下の確率を使って割り当てなおすだけ



$$P(z_{dn} = k) \propto \left(N_{dk\backslash n} + \alpha\right) \cdot \frac{N_{kw_n\backslash n} + \beta}{N_{k\backslash n} + \beta V}$$

文書dのn番目のトピックがkの確率

文書dのなかのトピックkの数

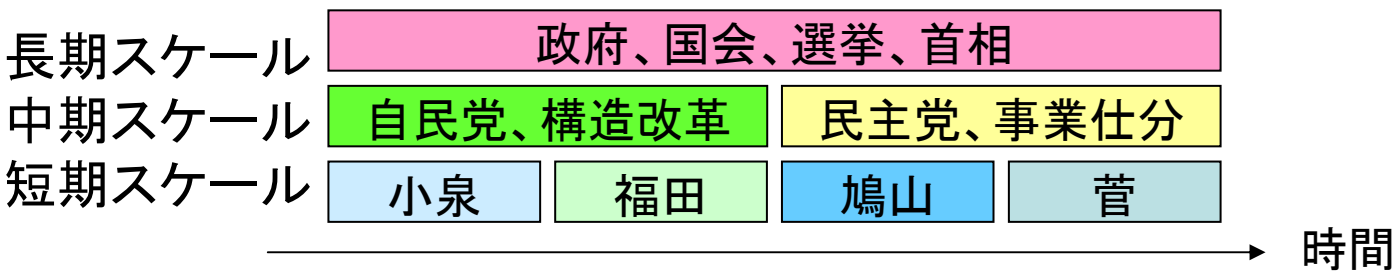
トピックkのなかの単語w_nの割合

（n番目の単語を除いたときの）

トピックモデルを使ったデータ解析

多重スケールトピック発展解析

- 時間発展する文書データにおける、複数の時間スケールでの変化をモデル化
  - 長期間流行したもの、短期間流行したものを高精度・効率的に抽出可能



Tomoharu Iwata, Takeshi Yamada, Yasushi Sakurai, Nanonori Ueda, “Online Multiscale Dynamic Topic Models,” KDD2010

ソーシャルアノテーション解析

- 内容と関連しないタグを自動抽出
  例：あとで読む、これはすごい、nikon

内容

タグ

NTT
あとで読む
研究
機械学習
これはすごい

関連する

NTT
研究
機械学習

関連しない

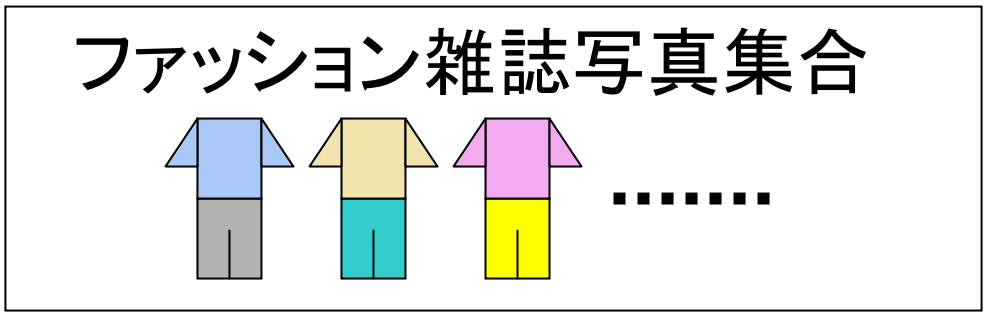
あとで読む

これはすごい

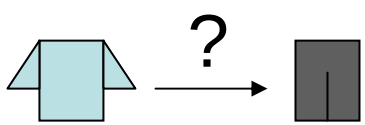
Tomoharu Iwata, Takeshi Yamada, Naonori Ueda, “Modeling Social Annotation Data with Content Relevance using a Topic Model,” NIPS2009

ファッションコーディネート推薦

- 学習データ



- トピックモデルを使って上衣・下衣の関連を学習
  - ファッション雑誌では、この上衣では、このようなトピックの下衣が合わせられることが多い
- 上衣に適した下衣を推薦



Tomoharu Iwata, Shinji Watanabe, Hiroshi Sawada, “Fashion Coordinates Recommender System using Photographs from Fashion Magazines,” IJCAI2011