自然言語処理・機械学習における企業との連携

持橋大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

「統計的機械学習」の中核としての統計数理
シンポジウム
2023-5-25 (木)

自己紹介

- 持橋大地
  統計数理研究所 数理・推論研究系 /
  総合研究大学院大学 統計科学専攻
- 1998年 東京大学教養学部基礎科学科第二卒
  2005年 奈良先端科学技術大学院大学博士後期修了
  2011年～統数研
- 専門：統計的自然言語処理、ベイズ統計的機械学習

産学官における連携

- 日本学術振興会
  グローバル学術情報センター
  (2015～)
  学術情報分析センター
  (2018～)
- 分析研究員 (定員3名) の一人を継続して務めています

産学官における連携 (2)

- 科研費の審査では、審査点の分布を擬似的に正規分布に近づける“Tスコア”という補正方式が使われている
  - 物理の主任研究員が考案したもの
  - 各審査員の特徴などを見ていない形式的補正のため、ボーダーライン付近で多数の同点が発生
  - 科研費の審査員は、どうしてもこのスコアに判断がバイアスされる

$$\text{Tスコア：}\quad y' = 3 + \frac{y - \mu_j}{\sigma_j} \times 0.6$$

産学官における連携 (3)

- 心理統計学における項目反応理論(IRT)による科研費審査点の補正・連続モデル化
  - 潜在的な「良さ」θ を連続値として推定

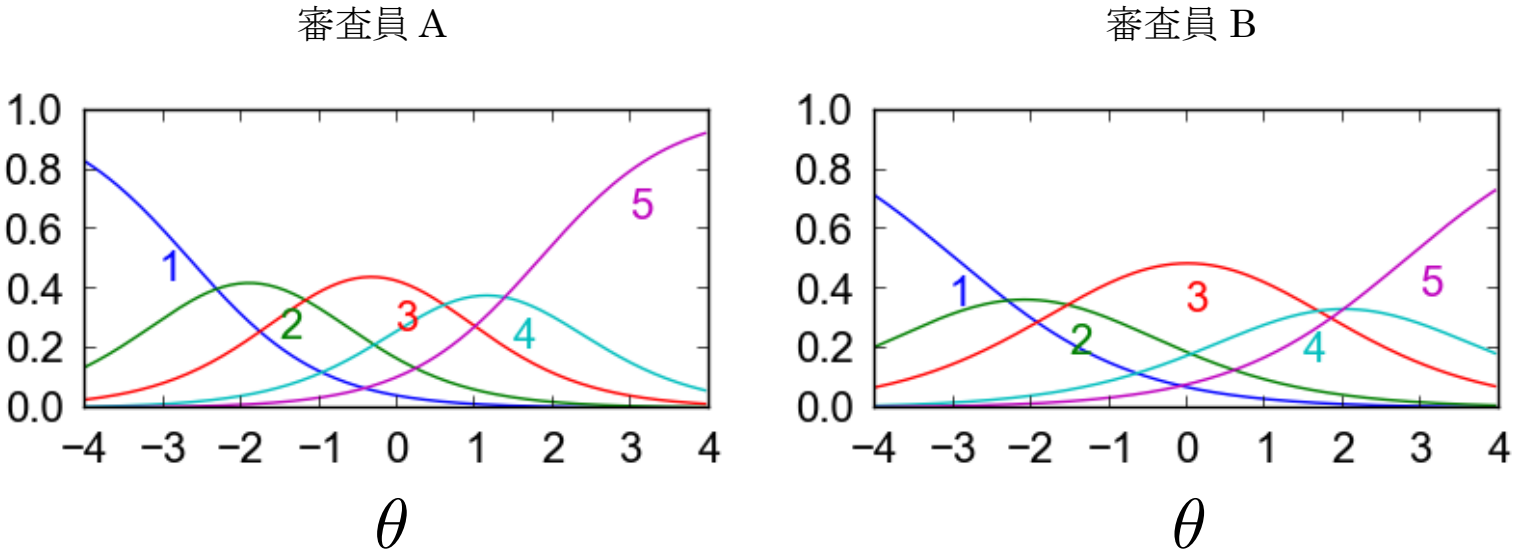
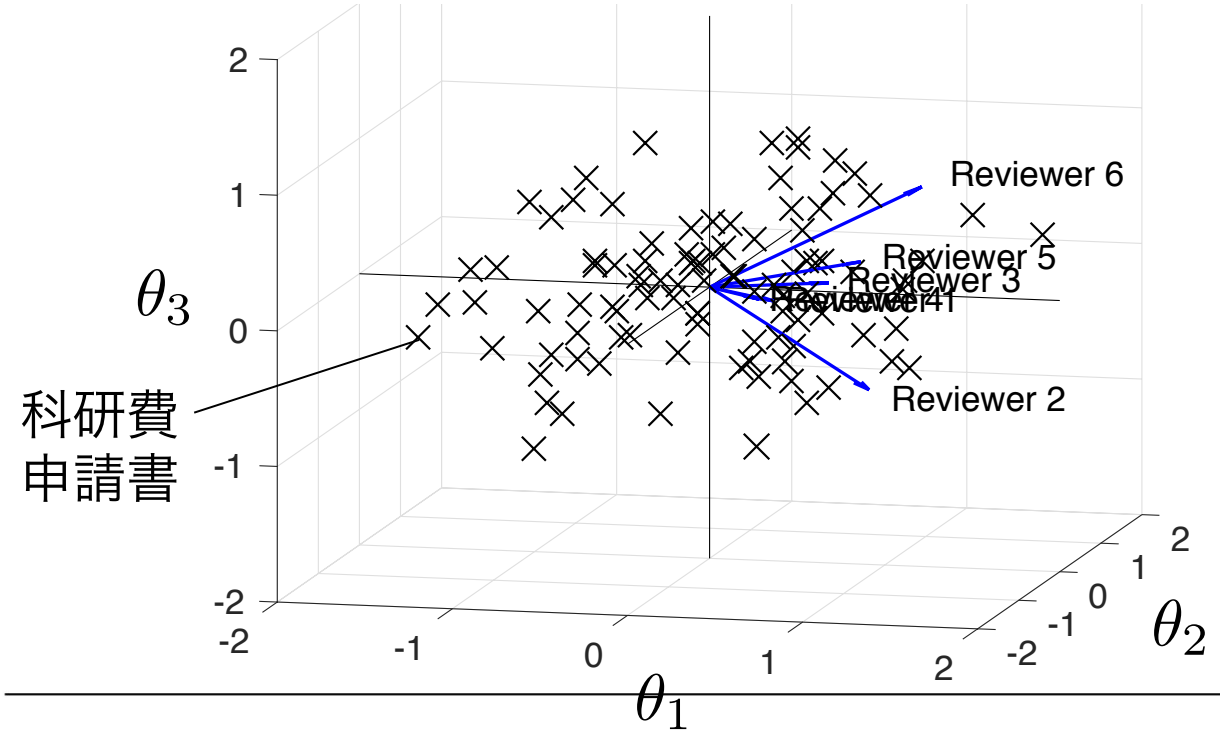


図 1. IRT による科研費申請に対するスコアの応答曲線の例。審査員によって各曲線の形状が異なり、そのパラメータはデータから統計的に学習される。

産学官における連携 (4)

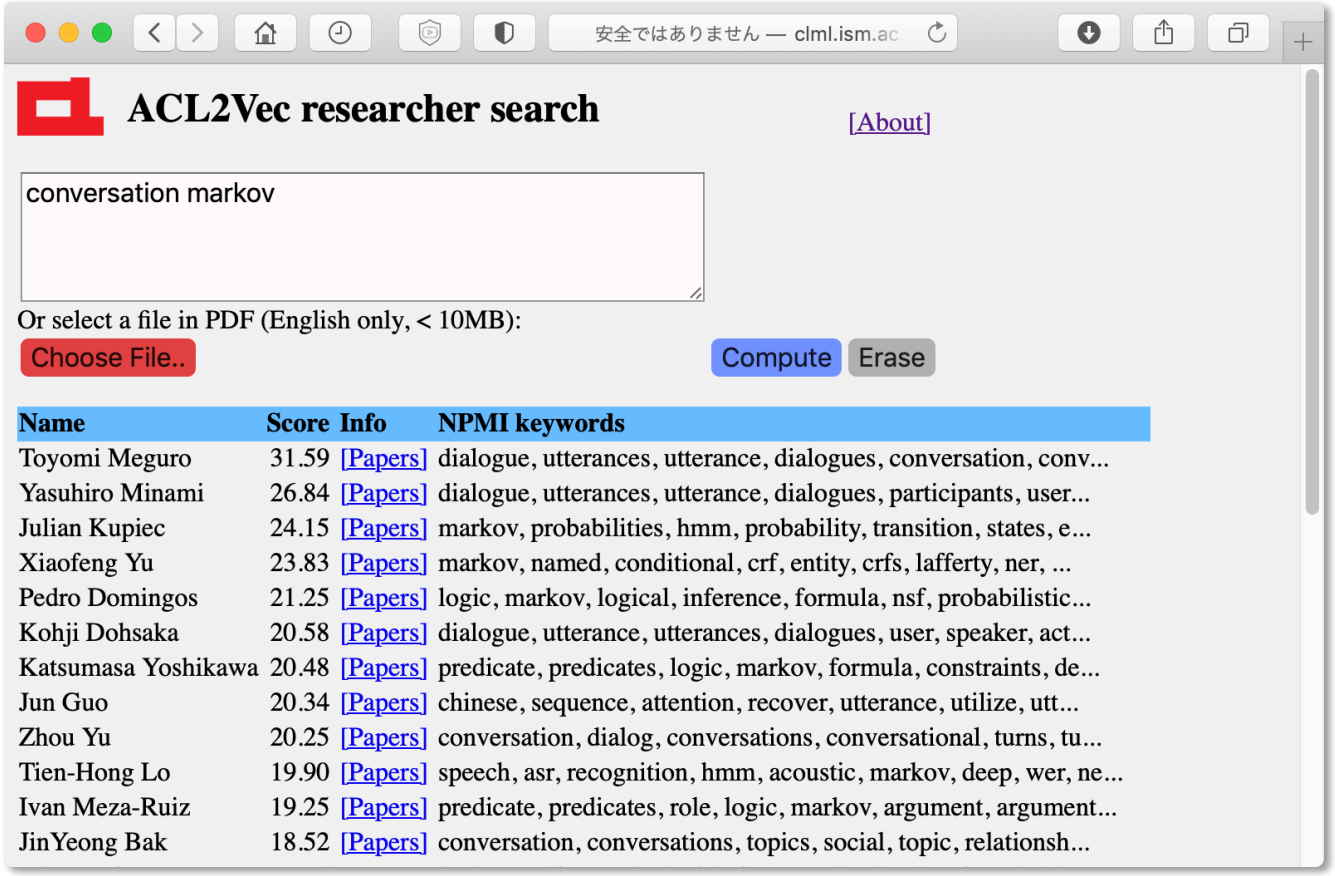
- θを多次元に→多様な審査の観点を統計モデル化
  - 各審査員は、潜在空間でその人の「審査軸」のベクトルを持っている
  - 多次元項目反応理論により、MCMCで計算可能



ある物理分野での
審査点から導出した
潜在評価点と
審査員の評価軸

産学官における連携 (5)

- 潜在意味解析およびニューラル文書ベクトルによる、各申請書の審査員候補の推薦 (中身はSVD+線形回帰)



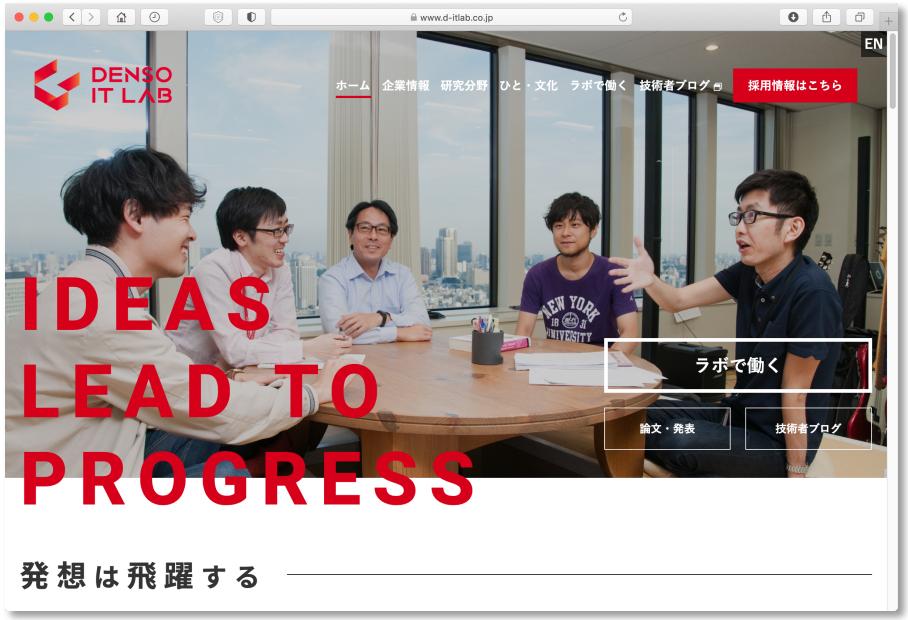
Name	Score	Info	NPMI keywords
Toyomi Meguro	31.59	[Papers]	dialogue, utterances, utterance, dialogues, conversation, conv...
Yasuhiro Minami	26.84	[Papers]	dialogue, utterances, utterance, dialogues, participants, user...
Julian Kupiec	24.15	[Papers]	markov, probabilities, hmm, probability, transition, states, e...
Xiaofeng Yu	23.83	[Papers]	markov, named, conditional, crf, entity, crfs, lafferty, ner, ...
Pedro Domingos	21.25	[Papers]	logic, markov, logical, inference, formula, nsf, probabilistic...
Kohji Dohsaka	20.58	[Papers]	dialogue, utterance, utterances, dialogues, user, speaker, act...
Katsumasa Yoshikawa	20.48	[Papers]	predicate, predicates, logic, markov, formula, constraints, de...
Jun Guo	20.34	[Papers]	chinese, sequence, attention, recover, utterance, utilize, utt...
Zhou Yu	20.25	[Papers]	conversation, dialog, conversations, conversational, turns, tu...
Tien-Hong Lo	19.90	[Papers]	speech, asr, recognition, hmm, acoustic, markov, deep, wer, ne...
Ivan Meza-Ruiz	19.25	[Papers]	predicate, predicates, role, logic, markov, argument, argument...
JinYeong Bak	18.52	[Papers]	conversation, conversations, topics, social, topic, relationsh...

企業との主な共同研究

- NTTコミュニケーション科学基礎研究所 (2011～)
- デンソーITラボラトリ (2014～)
- 博報堂 研究開発局 (2015～)
- AGC(旭硝子) (2018～) →統計科学専攻入学
- D2C (2018～) (ドコモ&電通)
- ブリヂストン (2019～) →統計科学専攻入学
- コーセー (2021～) →統計科学専攻入学
- 三井住友海上火災 (2022～)
- KDDI総合研究所 (2023～)
- 横河電機 (2023～)

デンソーITラボラトリとの共同研究

- (株)デンソーの持つ独立研究所、渋谷
- 共同研究のきっかけ:
  ITLabの内海さんが、
  半教師あり形態素解析
  についてうまくいかない
  点があり、共同研究を
  依頼された



- 自動車の車内でのコミュニケーションのための
  基礎研究

デンソーITラボラトリ (2)

- 当初持ち込まれたSHD-CRFによる半教師あり学習は、非常に局所解に陥りやすいことを指摘
  ↓
  局所解に陥らないベイズ学習に移行
- 任意の言語の文字列から、単語とその品詞をすべて教師なし学習する PYHSMM (Pitman-Yor Hidden Semi-Markov Model) を提案
- 自然言語処理のトップ国際会議 ACL 2015 に本会議論文として採択

“非教科書的”な言語の解析

今个学期第一次来图书馆呀～嘻嘻～感觉好好呀～稳紧写论文既资料呀～收获唔多，睇来要上网“刮”料拉
活动精彩照片集锦8-行进途中，璀璨夜色
到底这次是不是真的要走啦！！！tnnd敢不敢不要再变来变去了[愤怒]
我一直很开心的，，不是吗？？杭州的几个，我很想你们，，空了就来。。一定的，。。。你们懂！！！玲姐，，生意兴旺！！你们等着我！！！！继续我的骄傲，不可一世！！！！！
#小月月~月月~蓉荣#真月月，假月月，真假月月谁记得。浮云，一切都是浮云！

- 中国語のWeibo(SNS)の文の例

教師なし形態素解析 (持橋2009)

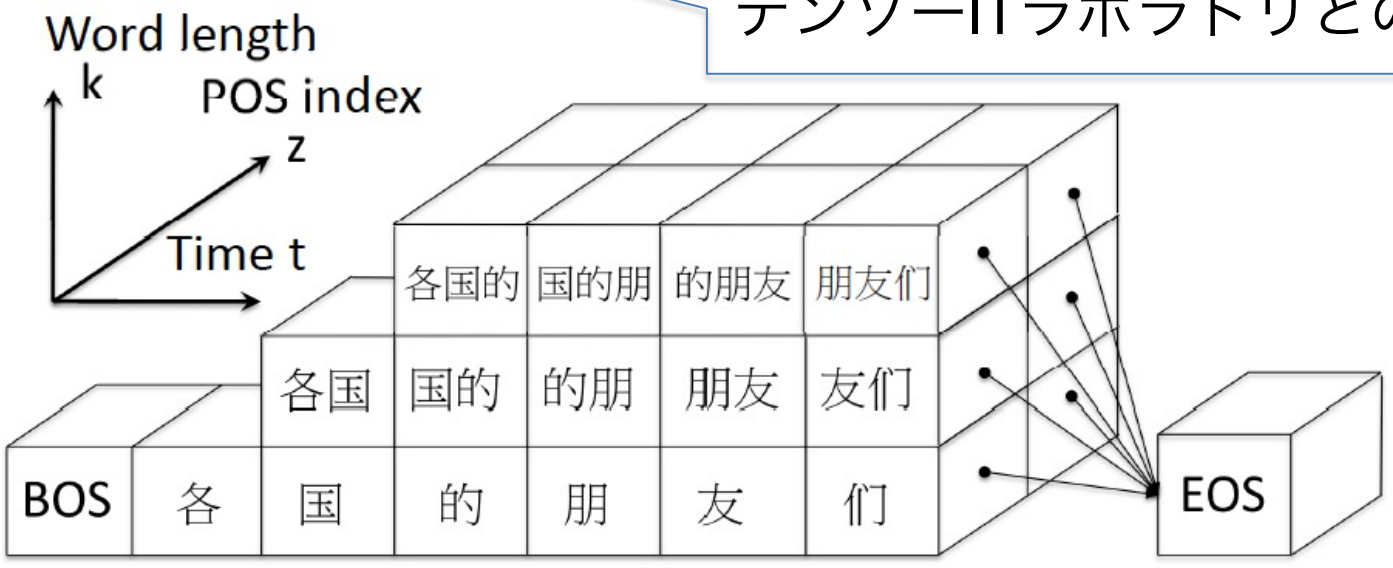
first,shedreamedoflittlealiceherself,andonceagainthetinyhandswereclaspedupo
nherknee,andthebrighteagereyeswerelookingupintohersshecouldheartheveryto
nesofhervoice,andseethatqueerlittletossofherheadtokeepbackthewanderinghai
rthatwouldalwaysgetintohereyesandstillasshelistened,orseemedtolisten,thewho
leplacearoundherbecamealivethestrangecreaturesofherlittlesister'sdream.thelo
Nggrassrustledatherfeetasthewhiterabbithurriedbythefrightenedmousesplashed
Hiswaythroughtheneighbouringpoolshecouldheartherattleoftheteacupsasthemar
chhareandhisfriendssharedtheirneverendingmeal,andtheshrillvoiceofthequeen…

first, she dream ed of little alice herself ,and once again the tiny hand s were clasped upon her knee ,and the bright eager eyes were looking up into hers -- shecould hearthe very tone s of her voice , and see that queer little toss of herhead to keep back the wandering hair that would always get into hereyes -- and still as she listened , or seemed to listen , thewhole place a round her became alive the strange creatures of her little sister 'sdream. thelong grass rustled ather feet as thewhitera bbit hurried by -- the frightened mouse splashed his way through the neighbour ing pool -- shecould hearthe rattle ofthe tea cups as the marchhare and his friends shared their never -endingme a l ,and the …

隠れた「品詞」との同時学習 (ACL 2015)

- 単語だけでなく、その“品詞”も知りたい→同時学習
  - MCMCの前向き確率は $\alpha[t][k][z]$(k:単語長, z:品詞)
  - Amazon EC2と並列化を使った膨大な計算量

デンソーITラボラトリとの共同研究



各国的朋友们
“friends of each country”

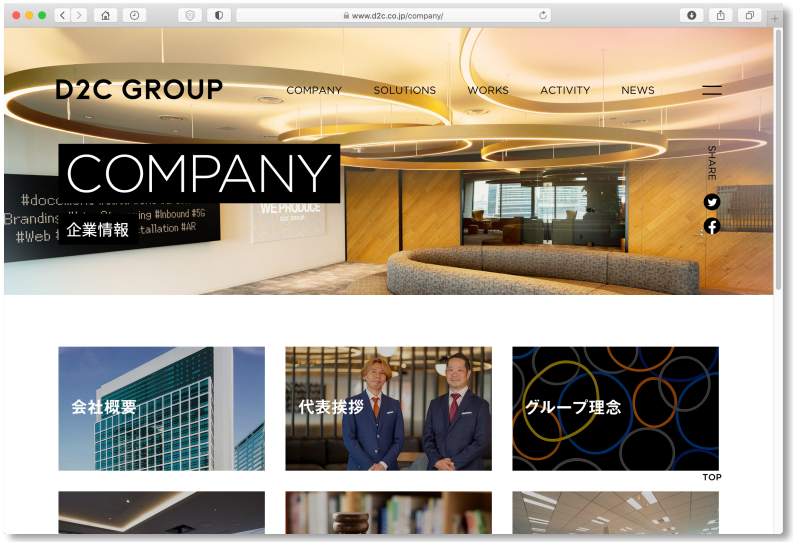
日本語方言の教師なし解析例

- 三河弁のTwitterアカウントから教師なし学習
  - 方言特有の語尾、顔文字、局所的な地名などがそれぞれ別の「品詞」に自動的に学習されている

z	Induced words
2	の 、 は に が で と も を 「
3	ぞん かん ね のん だに だん りん か ん だのん
9	(*^^*) !(^-^; (^_^;) (^^;; !(^^;;
10	。 ! !! ? 」 (≧∇≦) !!」「
11	楽 入 ど寒 大丈夫 会 受 停電 良 美味 台風が
13	に ら わ な よ ね だら じゃんね え ぁ
41	豊橋 名古屋 三河 西三河 名古屋弁 名古屋人 大阪

(株) D2Cとの共同研究

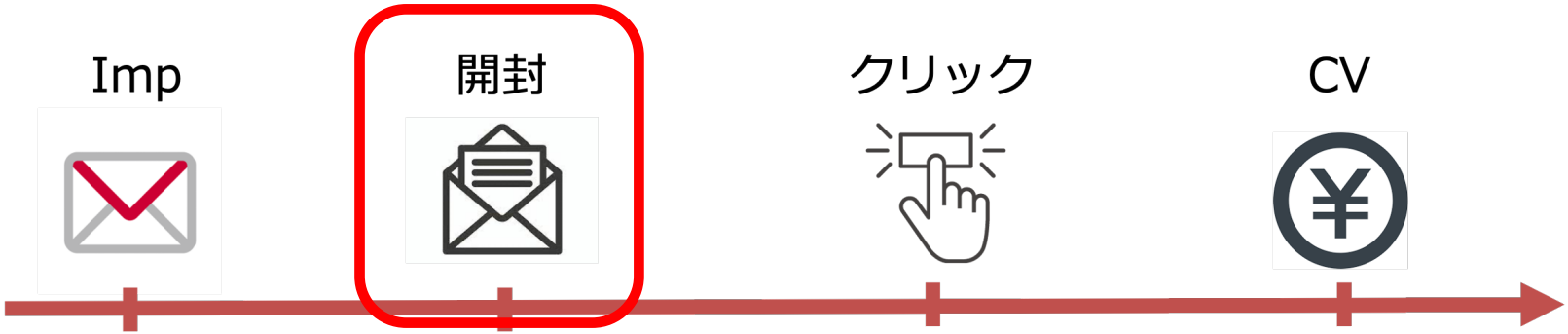
- NTTドコモと電通の合同のマーケティング企業
- 共同研究のきっかけ：
  先方からの依頼
  (理研/東大の杉山教授と私の2名)

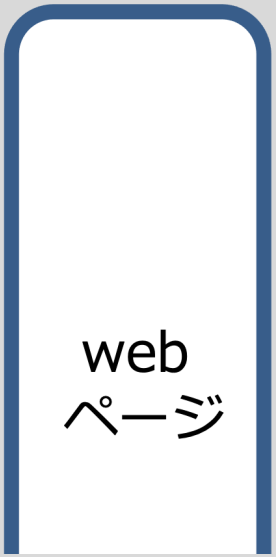


- 先方の目的：
  長期的な視点で、自然言語処理に関する基礎体力を高めたい
- 「メール型広告におけるタイトルが開封に与える影響」, 情報処理学会SIG-IFAT研究会, 2022.

(株) D2C (2)

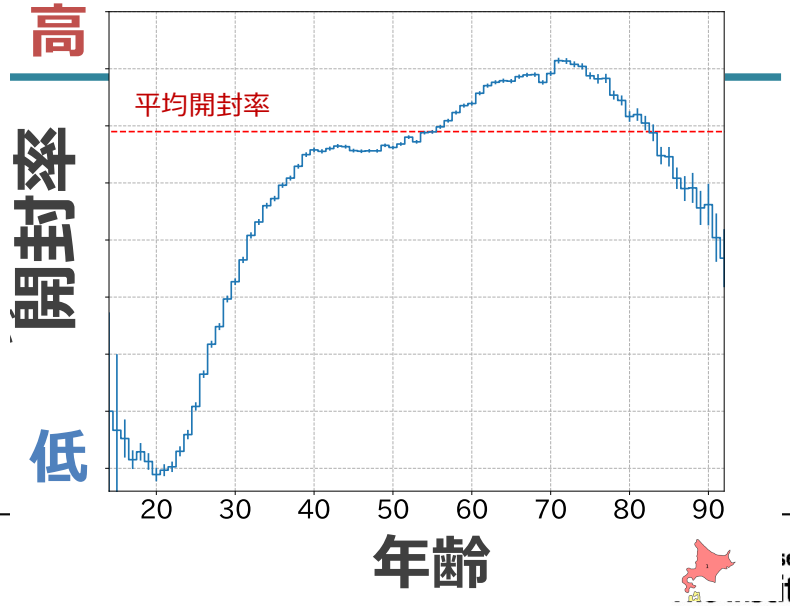
- ドコモのメッセージSにおける開封





(株) D2C (3)

- ドコモの“メッセージS”は、登録者3300万人
- クリック以前に、そもそもメールを開封してもらう必要
- 特徴量: タイトル文字列のみ
- ユーザーによって、どんなタイトルのメッセージを開封するかが異なる



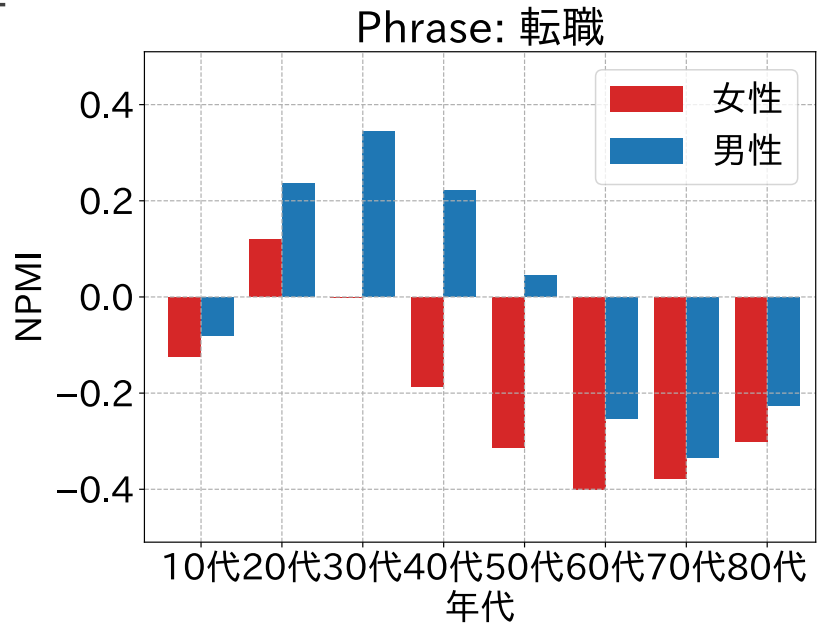
(株) D2C (4)

メッセージSの例

- 研究その1:
  タイトル文字列を統計的にフレーズ化
  → L_1ロジスティック回帰で開封予測
- 研究その2:
  各フレーズについて、開封したユーザの特徴との
  正規化自己相互情報量 (Normalized PMI) を計算

(株) D2C (5)

- 相互情報量を使ったタイトルの評価
  - 学習データのうち開封がされたログについて、フレーズと特徴(年代×性別)のNPMI(p,f)を計算



(株) D2C (6)

- Normalized PMI (正規化自己相互情報量)

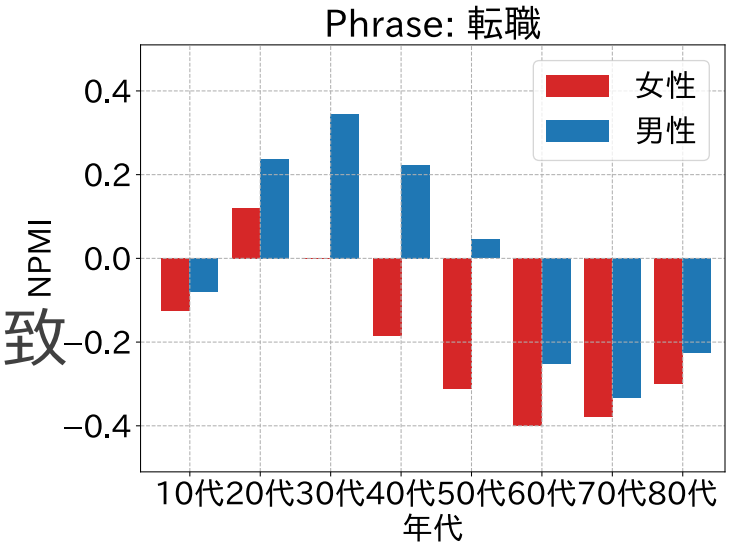
$$\mathrm{NPMI}(p, f) = \log \frac{p(f|p)}{p(f)} \Big/ (-\log p(p, f))$$

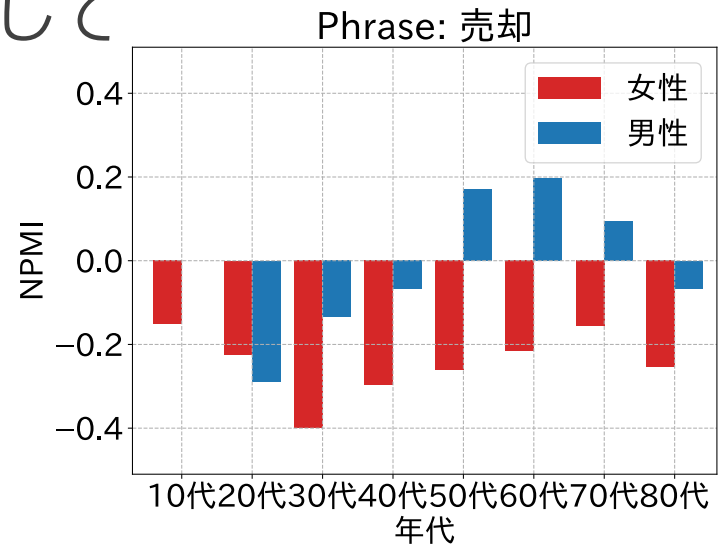
$$-1 \leq \mathrm{NPMI}(p, f) \leq 1$$

  - p(f|p) : フレーズpを含むタイトルを開封した中で特徴fを持つユーザーの割合
  - p(f) : ユーザーが特徴fを持つ事前確率
  - p(p,f) : 特徴fのユーザーがタイトルにフレーズpを含むメッセージを開封した確率

◆結果

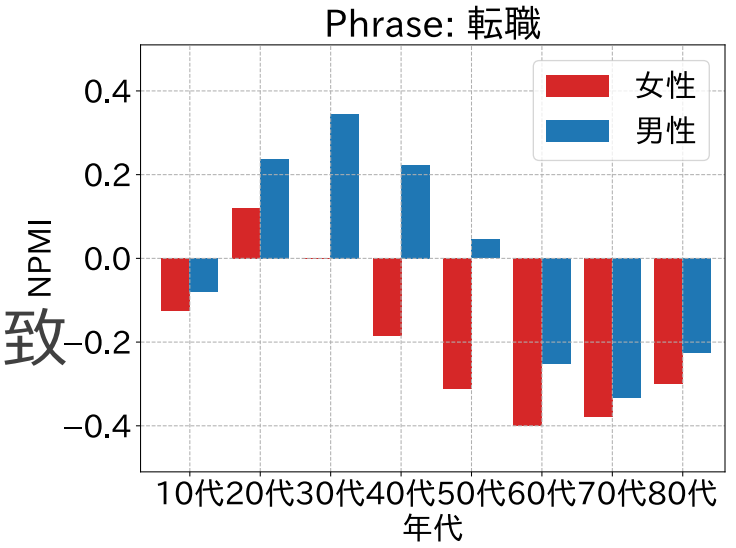
- 男性に効果的なフレーズ例
  - “転職”
    - 若年層のほうが高齢世代より転職活動に積極的という解釈に一致
  - “売却”
    - 不動産や車の売却で使用されるフレーズで高齢男性が資産が成熟して管理する立場にあることを示唆
  - フレーズと特徴の共起度から直感と矛盾のない知見が得られている

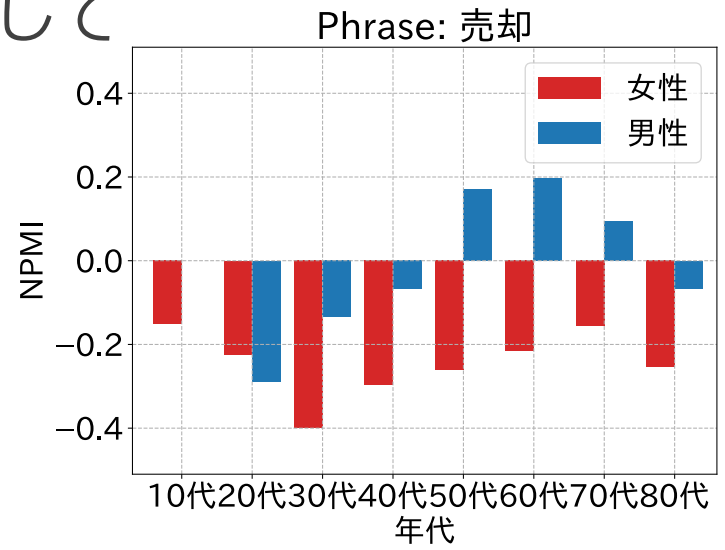




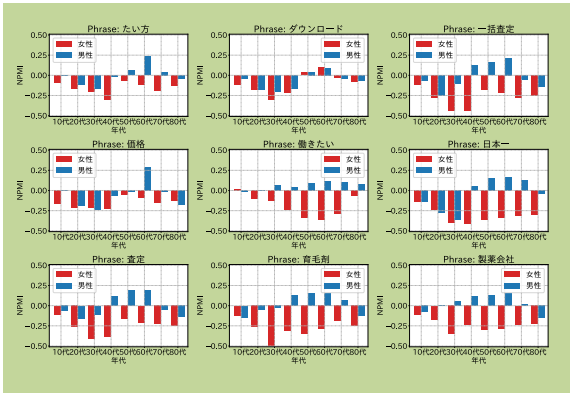
◆結果

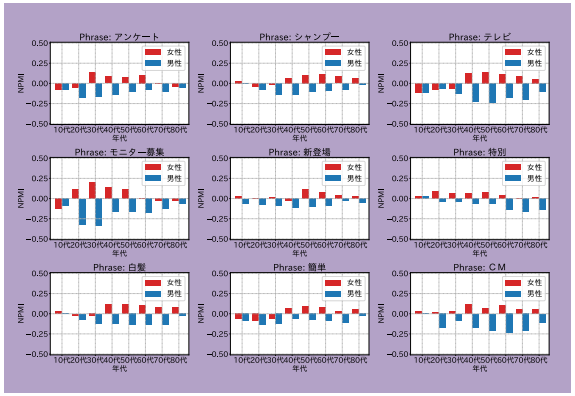
- 男性に効果的なフレーズ例
  - “転職”
    - 若年層のほうが高齢世代より転職活動に積極的という解釈に一致
  - “売却”
    - 不動産や車の売却で使用されるフレーズで高齢男性が資産が成熟して管理する立場にあることを示唆
  - フレーズと特徴の共起度から直感と矛盾のない知見が得られている

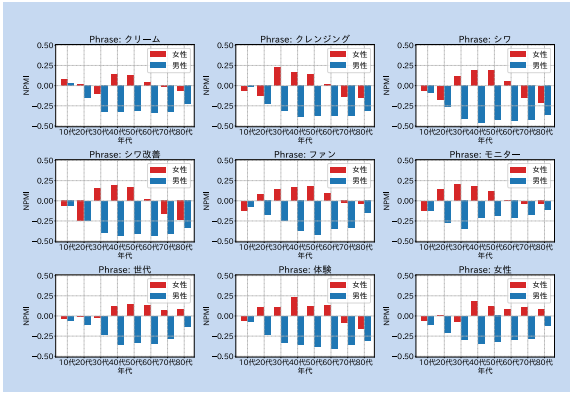


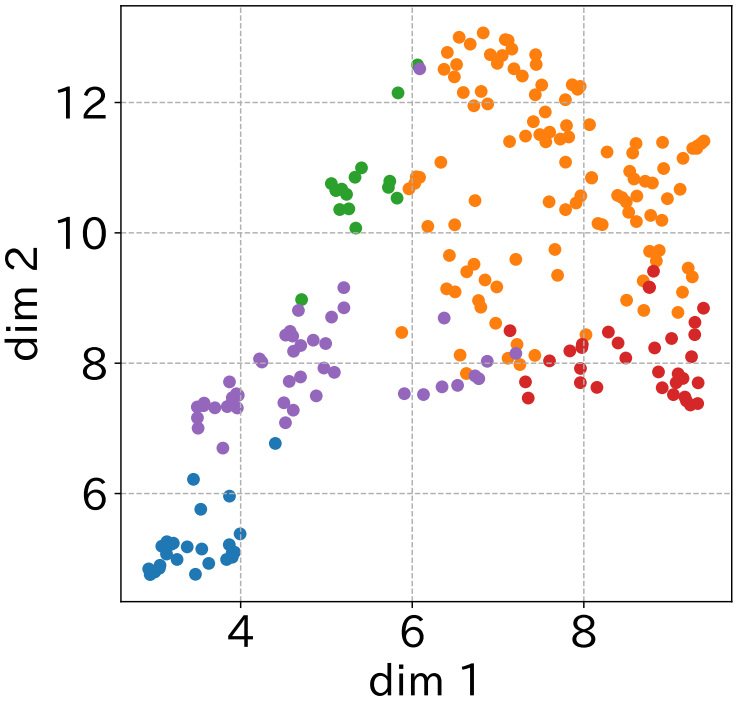


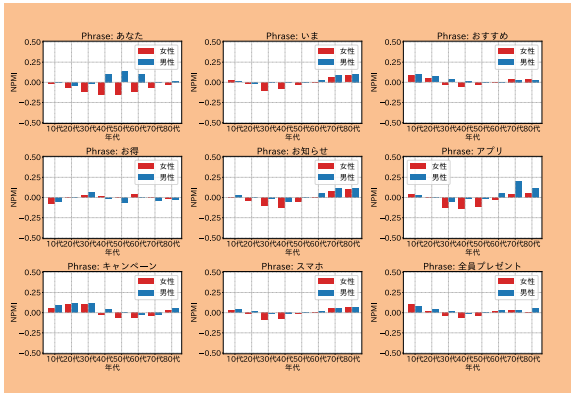
◆結果

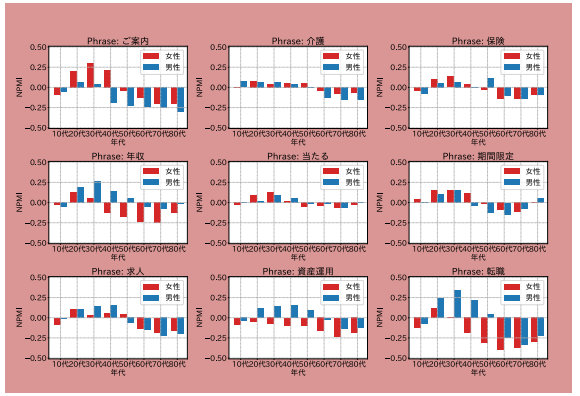






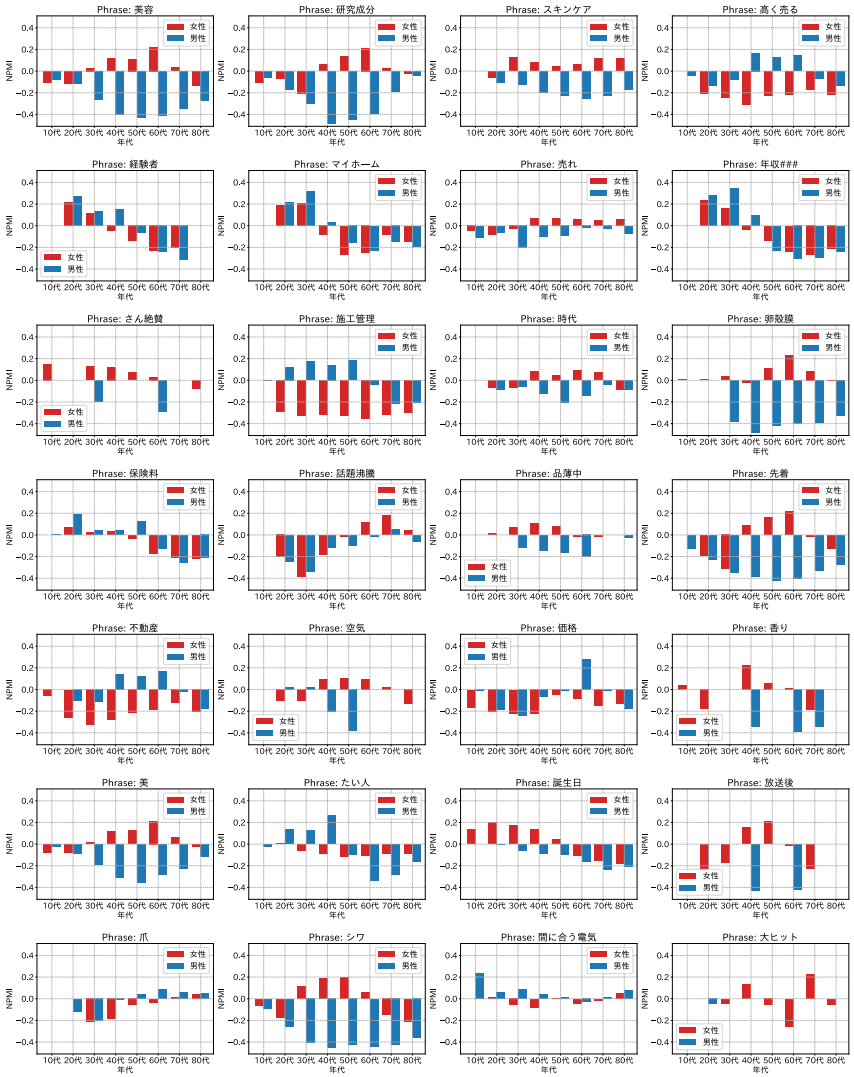


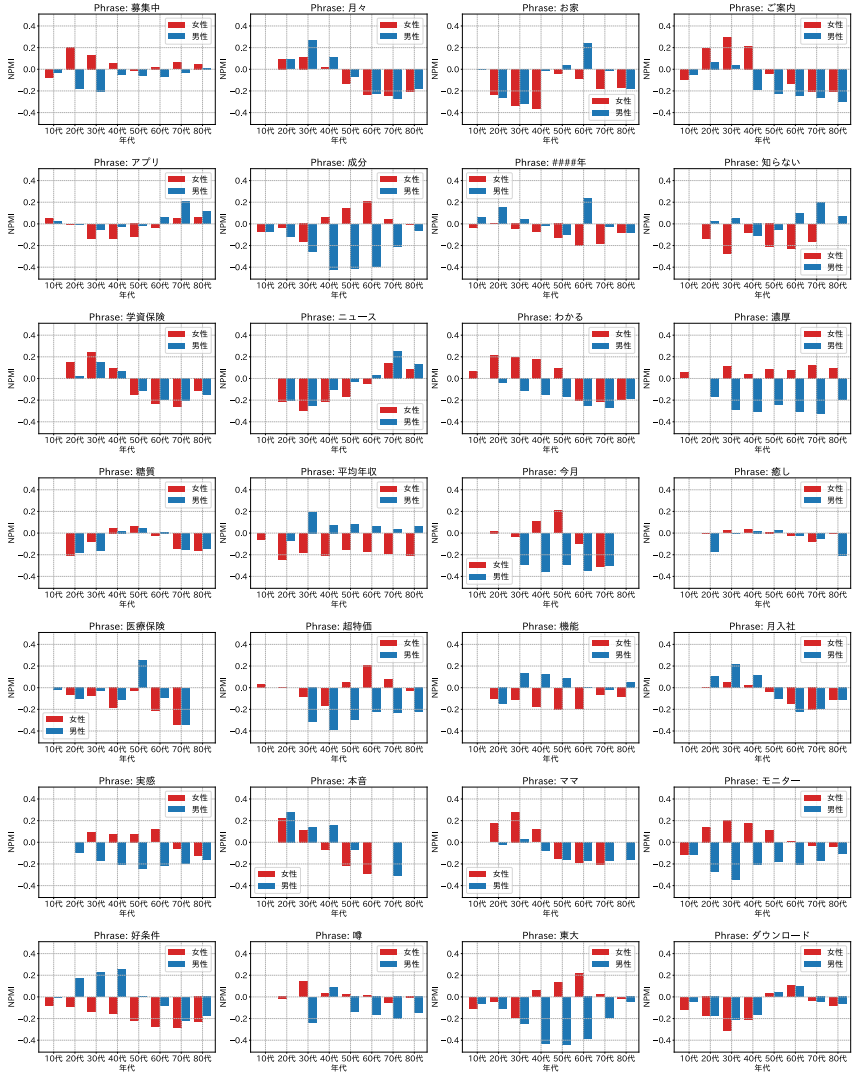




- 男性にやや肯定的、女性に否定的
- 男性にやや否定的、女性にやや肯定的
- 男性に否定的、女性にやや肯定的
- 開封への影響は弱い
- 若い世代に肯定的、高齢層に否定的

おまけ


Phrase: 美容
Phrase: 研究成分
Phrase: スキンケア
Phrase: 高く売る
Phrase: 経験者
Phrase: マイホーム
Phrase: 売れ
Phrase: 年収###
Phrase: さん絶賛
Phrase: 施工管理
Phrase: 時代
Phrase: 卵殻膜
Phrase: 保険料
Phrase: 話題沸騰
Phrase: 品薄中
Phrase: 先着
Phrase: 不動産
Phrase: 空気
Phrase: 価格
Phrase: 香り
Phrase: 美
Phrase: たい人
Phrase: 誕生日
Phrase: 放送後
Phrase: 爪
Phrase: シワ
Phrase: 間に合う電気
Phrase: 大ヒット
女性
男性
NPMI
年代

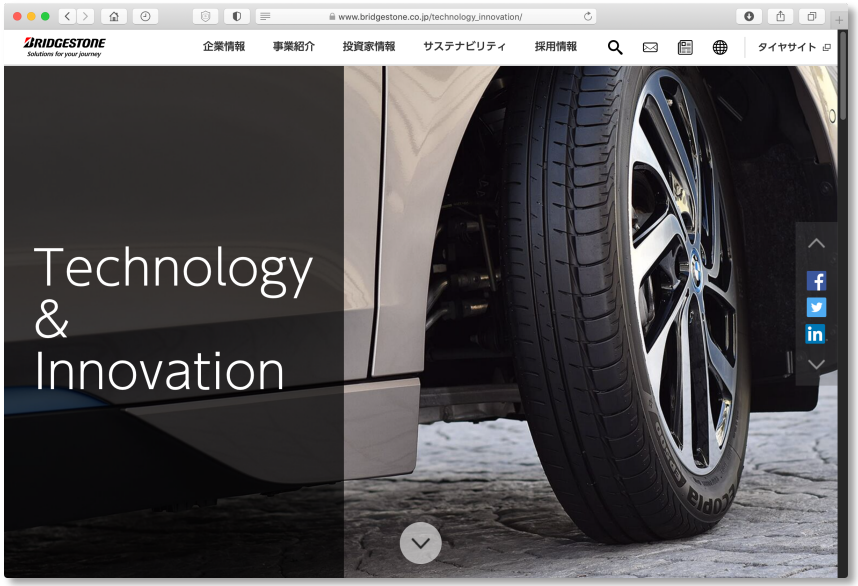

Phrase: 募集中
Phrase: 月々
Phrase: お家
Phrase: ご案内
Phrase: アプリ
Phrase: 成分
Phrase: ####年
Phrase: 知らない
Phrase: 学資保険
Phrase: ニュース
Phrase: わかる
Phrase: 遺厚
Phrase: 糖質
Phrase: 平均年収
Phrase: 今月
Phrase: 癒し
Phrase: 医療保険
Phrase: 超特価
Phrase: 機能
Phrase: 月入社
Phrase: 実感
Phrase: 本音
Phrase: ママ
Phrase: モニター
Phrase: 好条件
Phrase: 噂
Phrase: 東大
Phrase: ダウンロード
女性
男性
NPMI
年代

(株)D2Cとの共同研究

- 数千万人のユーザーを抱えるNTTドコモでしかできない、実際のユーザーの嗜好をとらえる研究
  - 大学では、こうした現実のデータは入手不能
- 正規化自己相互情報量を計算することで、どんな言葉がどんなユーザーの開封に繋がるのかを統計的に分析できた
- あまり教科書には書かれていないため、専門家の一部しか使えない手法 →教科書執筆中

(株)ブリヂストンとの共同研究

- 共同研究のきっかけ:
  URAを介した依頼
  - 統数研とは、以前より別内容で共同研究を行っていた



- 営業等から多数上がる
  社内文書の資源化
  → 文書の統計的可視化についての研究
- ECML-PKDD 2022に採択
- 統計科学専攻に入学 (主指導: 藤澤教授)

NPDV

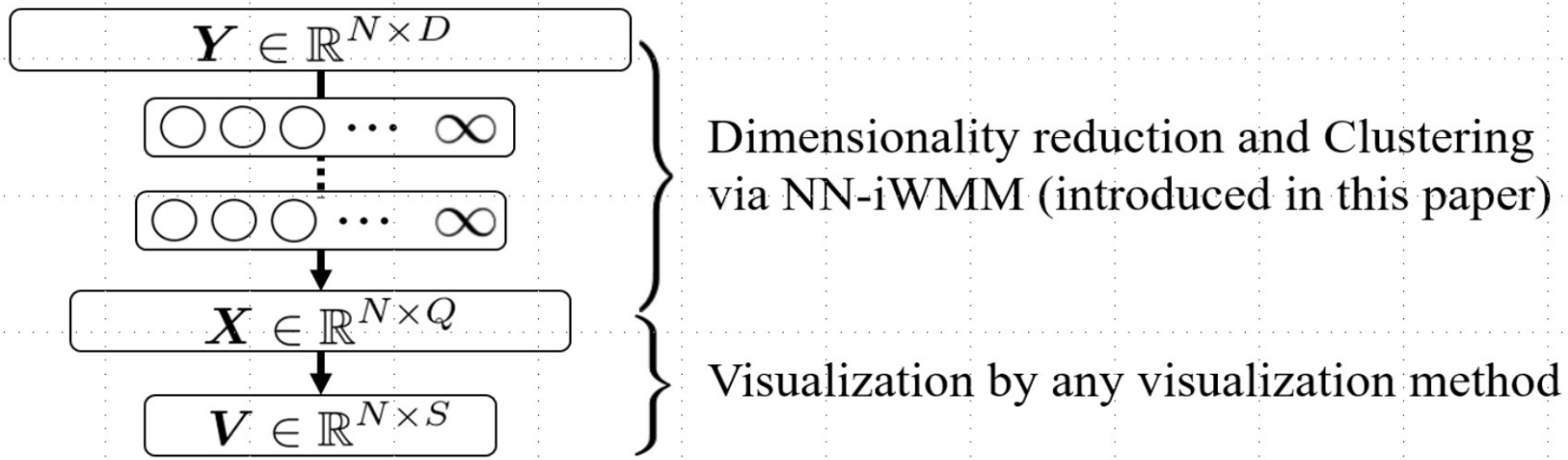
- NPDV: Nonparametric Bayesian Deep Visualization

NPDV t-SNE (2008) PaCMAP (2021)

  - 各点は一つの文書
  - 潜在空間で無限ガウス混合モデルでクラスタリングも同時に行う

Nonparametric Bayeisna Deep Visualization: Overview

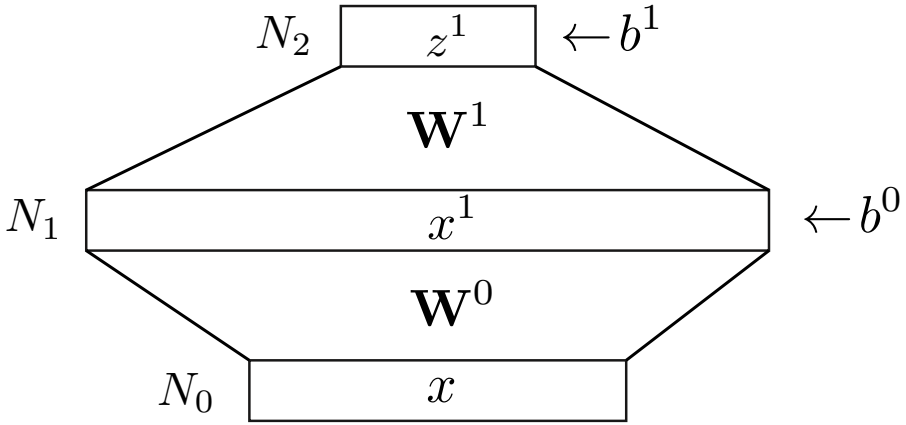
- Integrate NN-iWMM and a visualization method into a Bayesian model
  - Infer ***X*** so as to be optimal to estimate ***V***



Neural Network = Gaussian process (Neal 1994)

- Single hidden layer NN with inputs layer x
- i'th output $z_i(x)$is written as:

$$z_i^1(x) = b_i^1 + \sum_{j=1}^{N_1} W_{ij}^1 x_j^1(x), \quad x_j^1(x) = \phi\left(b_j^0 + \sum_{k=1}^{K} W_{jk}^0 x_k\right)$$



Connection

$$W_{ij}^\ell \sim \mathcal{N}(0, \sigma_w/N_\ell)$$

Bias

$$b_i^\ell \sim \mathcal{N}(0, \sigma_b)$$

NNGP (Neural Network Gaussian Process) (Lee+ 2017)

- Assume $\ell-1$ layer output $z_j^{\ell-1}$ is GP:

$$z_i^\ell(x) = b_i^\ell + \sum_{j=1}^{N_\ell} W_{ij}^\ell \, x_j^\ell(x), \quad x_j^\ell(x) = \phi(z_j^{\ell-1}(x))$$

- ℓ-layer Mean is 0, variance is

$$\begin{aligned} K^\ell(x, x') &\equiv \mathbb{E}\left[z_i^\ell(x) z_i^\ell(x')\right] \\ &= \sigma_b^2 + \sigma_w^2 \, \mathbb{E}_{z_i^{\ell-1} \sim \mathrm{GP}(0, K^{\ell-1})} \left[\phi(z_i^{\ell-1}(x)) \phi(z_i^{\ell-1}(x'))\right] \end{aligned}$$

  - This expectation can be computed by:
    (1) GP regression (2) Numerical approximation
    (3) Analytical solution for specific φ like ReLU

NNGP (Neural Network Gaussian Process) (2)

- When φ is ReLU (Cho&Saul 2009, Lee+ 2017) :

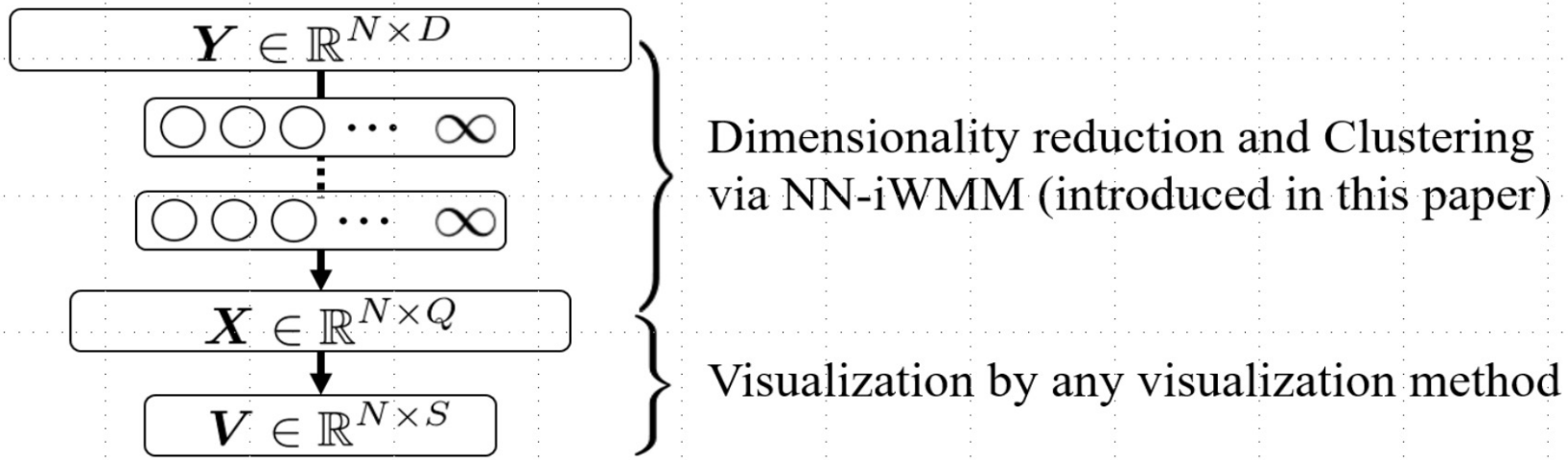
$$K^{\ell}(x,x') = \sigma_b^2 + \frac{\sigma_w^2}{2\pi}\sqrt{K^{\ell-1}(x,x)K^{\ell-1}(x',x')} \times \left(\sin\theta_{x,x'}^{\ell-1} + (\pi - \theta_{x,x'}^{\ell-1})\cos\theta_{x,x'}^{\ell-1}\right)$$

$$\theta_{x,x'}^{\ell} = \cos^{-1}\left(\frac{K^{\ell}(x,x')}{\sqrt{K^{\ell}(x,x)K^{\ell}(x',x')}}\right)$$

- Multi-layer NN is obtained just by matrix multiplications!

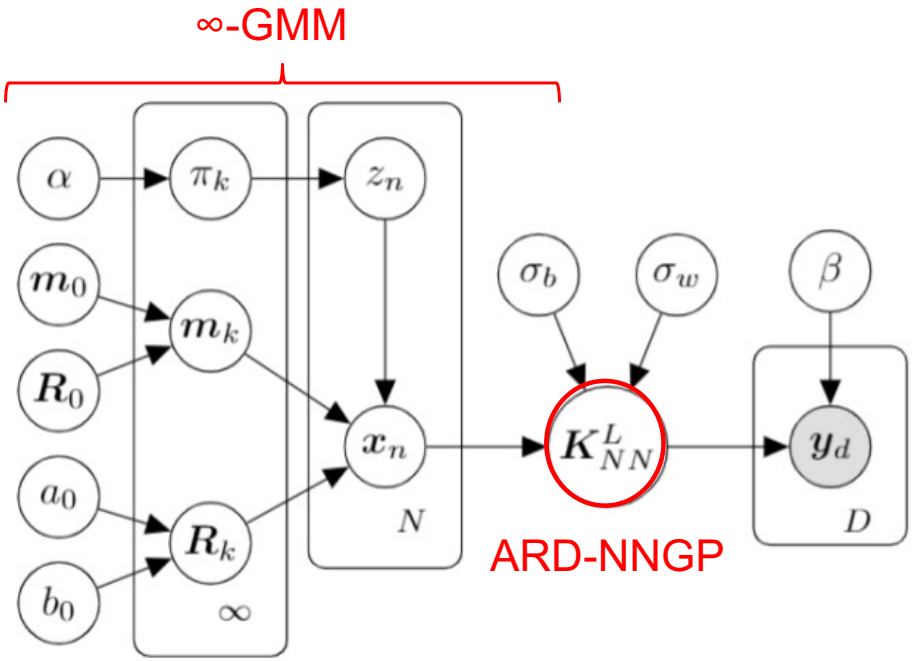
Nonparametric Bayeisna Deep Visualization: Overview

- Integrate NN-iWMM and a visualization method into a Bayesian model
  - Infer ***X*** so as to be optimal to estimate ***V***



NN-iWMM: Generative process

- Incorporate ARD-NNGP and ∞~GMM into a latent variable model



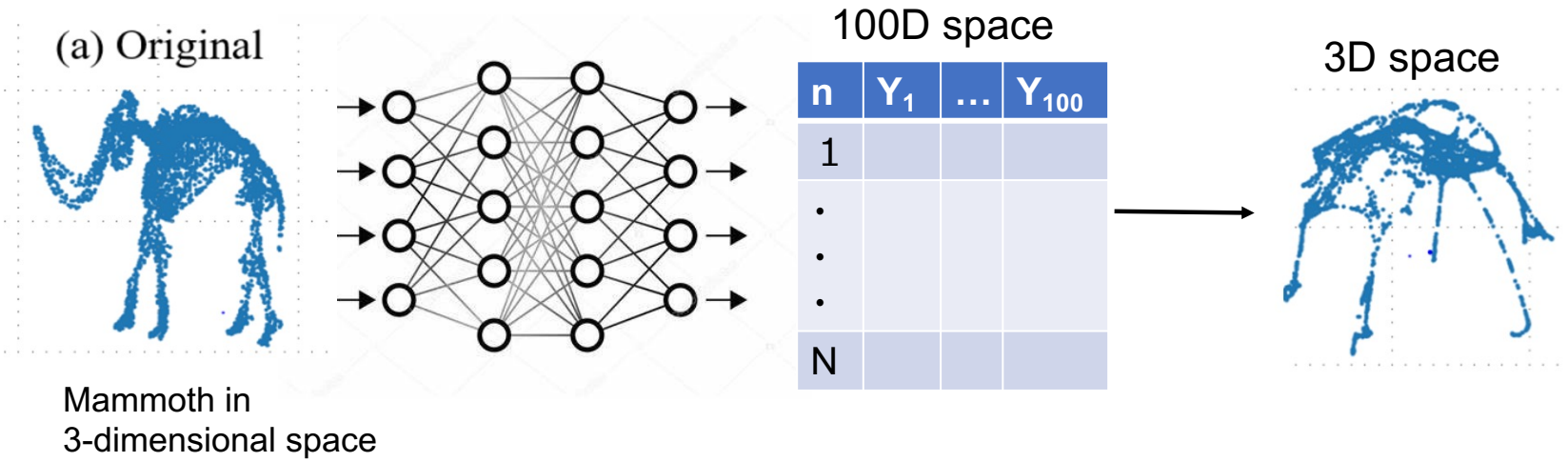
×: to be optimized

✓:no need to be optimized

	NN-iWMM
# of layers	×
# of units	✓
# of latent dims	✓
# of clusters	✓

Data generation setup

- Generate 100-dimensional data through NN transformation
- Apply 4 methods including NPDV (MF) to recover the original data



Visualization results

- NPDV (MF) could recover the original mammoth shape accurately
- Some body parts (horn, pow) was found through estimated clusters

(a) NPDV (MF) (b) *t*-SNE (2008) (c) Trimap (2018) (d) PaCMAP (2021)

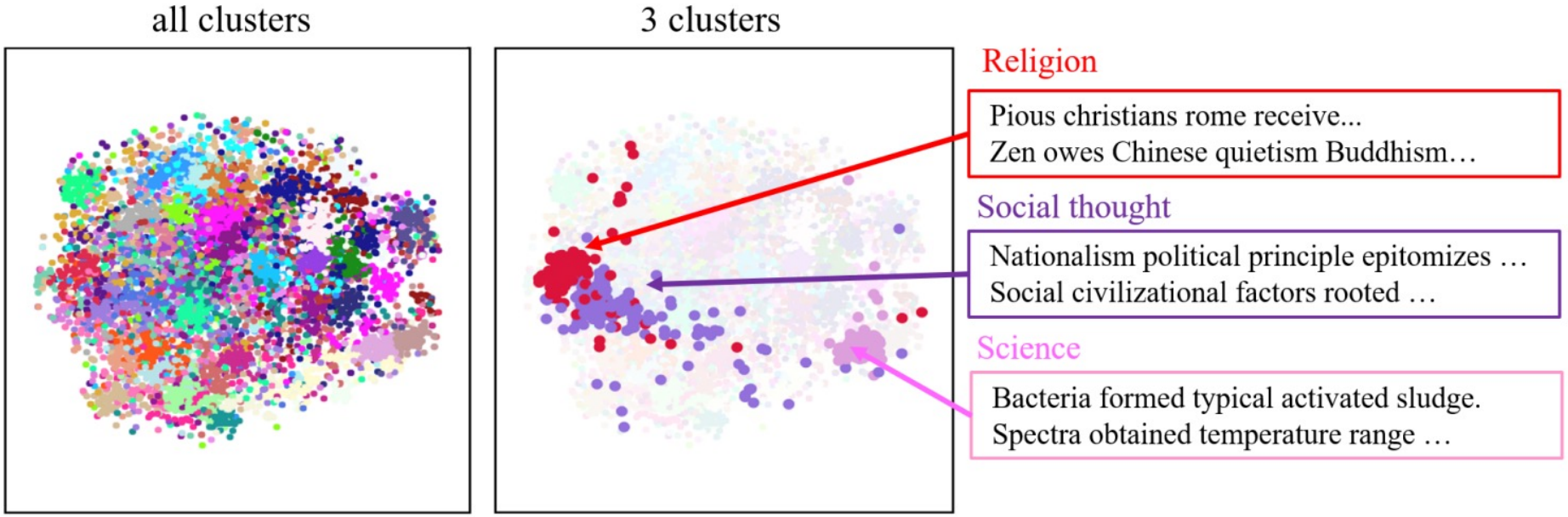
Qualitative comparison @ 20 newsgroups

- Coloring provides intuitive understanding for the cluster structure
- NPDV shows better cluster separation than (b)

(a) NPDV(t-SNE) (b) VSB-DVM (c) t-SNE (2008) (d) PaCMAP (2021)

Latent cluster discovery @ Brown corpus

- Visualize plausible topics and their relationships through NPDV



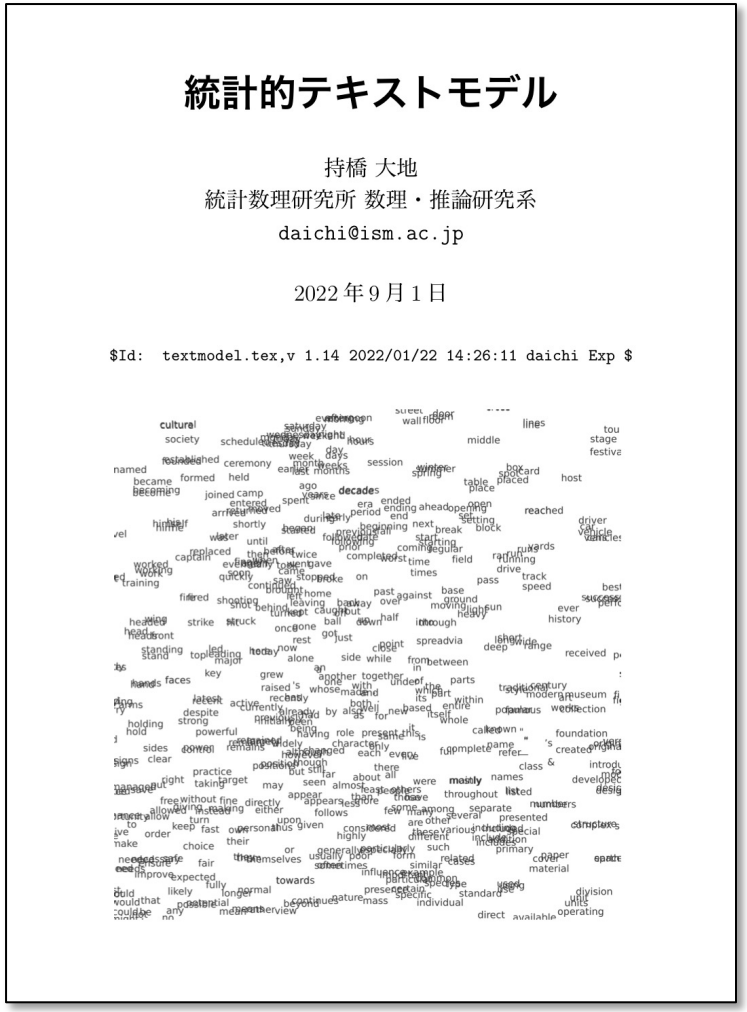
(株)ブリヂストンとの共同研究: まとめ

- 社内文書の可視化のための、新しいニューラル+統計モデルによる可視化法: NPDV
  - 潜在層は無限ガウス混合モデル
  - 潜在層→観測次元へはニューラルネットガウス過程(NNと等価だが、重みがすべて積分消去されており最適化不要)
- 統計的に考えることで、可視化に余計なパラメータ推定を必要とせず、クラスタ数も自動推定
- 担当者の高い理解力・実装力

全体のまとめ

- 政府機関や企業しか持っていないデータ・実問題に対する共同研究 → 統計学の知見が有効
- 共同研究でアドバイスをしないと、実用的な速度で動かないことが多い (実装のKnow-How)
- 企業側の担当者の理論面での理解力・実装力の高さ →トップ国際会議にも採択
- 最近忘れられかけている、統計的機械学習の基本を共有していく必要を実感

教科書 (刊行予定)



- 岩波書店より、2023年中に発売予定 (現在284p)
- 他の教科書にあまり詳しく書かれていない話も扱っています
- http://chasen.org/~daiti-m/textmodel/ で草稿を公開・意見募集中