京都大学情報学研究科　統計知能分野（統計数理分野）
教授 下平英寿、准教授 本多淳也、助教 山際宏明

統計学，機械学習，データサイエンスの手法と理論を探求

ビッグデータ、データマイニング、人工知能の流行を支える理論的基盤として統計学・機械学習は重要な役割を果たしています。とくにランダムネスを考慮してデータから帰納的推論を行う方法論を提供することが統計学の大きな特徴です。研究室では、人工知能やデータサイエンスの基礎となる統計学や機械学習を融合する形でのデータ駆動型の帰納的推論の探求、すなわち「統計的方法論によるデータからの理解や思考の探求」を目指しています。

現実のデータにとりくんで，新たな理論を作る

かつて遺伝学においてR. A. Fisherが統計学を飛躍的に発展させたように、現実と向き合うことが方法論の発展をもたらします。近年ではデータ収集や処理が容易になり、大量のデータを扱える方法論の必要性が増しています。そしてTransformerのようなニューラルネットによる言語モデルを学習させることによって理解や思考といった知能が実現しつつあり、ChatGPTなどの普及によってもはや身近なものとなっています。言語に限らず画像やロボティックス等の様々な分野でこのような生成モデルによる革新が進行中です。しかし、ニューラルネットがなぜ優れた性能を発揮するのかについては未解明の部分が多く、さらなる発展のためにはその仕組みの理解が重要です。

研究室では、たとえば「自然言語処理のニューラルネットにおいて単語の意味がどのように表現されているか」を数理統計学の観点で解明して、言語モデルにおける「思考」を理解するための研究を行っています。また、ベイズや頻度論といった統計的方法論の基本的な研究に取り組んでいます。新薬の開発や新商品の推薦といった試行錯誤を要する動的意思決定問題にも取り組んでおり、バンディット問題とよばれる定式化を通じた理論限界の追求から強化学習を用いた実用的なアルゴリズムの構築まで理論・実践の両面から研究を行っています。

数学とプログラミング，どちらも重要

研究で最も重要なのはアイデアとデータです。そして数学とプログラミングは力です。定理の証明とコーディングは似た作業ですね。数学に自信のある人、Python，R，C++のスキルがある人は活躍するチャンスがあるし、やる気さえあれば研究を通して実力はつくものです。

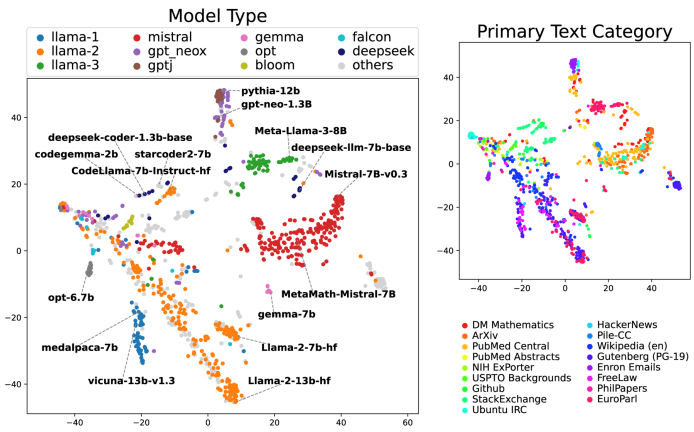
たとえば，こんな研究やってます

情報幾何学で確率分布をユークリッド空間に埋め込むデータ解析手法を考える

$2\,\mathrm{KL}(p_i, p_j) \approx \mathrm{Var}_{x\sim p_0}\left(\ell_i(x) - \ell_j(x)\right)$　下平, 統計数理, 1993

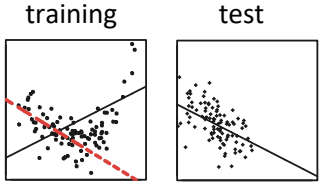
埋もれてしまうが、ふと思い出す

生成モデル全盛時代になり復活し、1000個の言語モデルを埋め込んでみる　⇐いまココ！



Mapping 1,000+ Language Models via the Log-Likelihood Vector
Oyama, Yamagiwa, Takase, Shimodaira (2025), arXiv:2502.16173

学習時とテスト時でデータの分布が変化する場合の統計理論を考える



確率密度比(density ratio)

$$w(x) = \frac{f_{\mathrm{test}}(x)}{f_{\mathrm{train}}(x)}$$

Shimodaira, Journal of Statistical Planning and Inference, 2000

covariate shift (共変量シフト)と命名！

機械学習の分野で転移学習の基本形となる

Shimodaira (2000)の論文被引用数は2500くらい

しらないうちに．．．

ニューラルネットのディープラーニングを加速する手法の原理として組み込まれる

Batch normalization: Accelerating deep network training by reducing internal covariate shift (Ioffe and Szegedy, ICML 2015)

彼らの論文被引用数はすでに60000くらい．．．

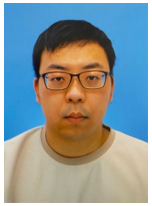
http://stat.sys.i.kyoto-u.ac.jp/　研究室ウェブサイト

統計知能分野（システム科学）
統計数理分野　（データ科学）

統計学　機械学習　データサイエンス

教授 下平英寿，准教授 本多淳也，助教 山際宏明





・京都大学情報学研究科　システム科学コース　統計知能分野　シー5
・京都大学情報学研究科　データ科学コース　統計数理分野　デー1
・京都大学工学部情報学科　数理工学コース　統計知能分野

教授 下平英寿，准教授 本多淳也，助教 山際宏明

研究トピックの紹介

- **対数尤度ベクトルに基づく言語モデルの地図**
- **言語や画像の埋め込みに共通する「形」の発見**
- **埋め込みのノルム（長さ）は「意味の強さ」を表す**
- **分散表現の論理演算**
- **単語埋め込みで薬剤と遺伝子間のアナロジータスクを解く**
- **不均衡最適輸送を用いた意味変化検出**
- **バンディット問題**
- **強化学習を用いたバンディットアルゴリズム**
- **ベイズ最適化**
- **グラフ埋め込みの理論**
- **グラフ埋め込みの応用**
- **マルチスケール k-近傍法**
- **複雑ネットワークの統計推測**
- **ブートストラップと選択的推測**

http://stat.sys.i.kyoto-u.ac.jp/　研究室ウェブサイト

対数尤度ベクトルに基づく言語モデルの地図

Oyama, Yamagiwa, Takase, Shimodaira (arXiv:2502.16173) Mapping 1,000+ Language Models via the Log-Likelihood Vector

① 背景と目的

・Hugging Face では20万近い言語モデルが公開
・言語モデルの類似関係を調べたい
・タスク性能の予測などに利用したい
・言語モデルなど生成モデルがなぜ良いのかを解明したい

②理論と方法

・言語モデル等の生成モデルは、結局は確率モデル
・テキスト生成の確率モデル $p_i(x)$
・情報幾何学の理論（指数型分布族の性質）から２つの言語モデル$p_i(x), p_j(x)$のカルバック・ライブラーダイバージェンスと対数尤度の分散の関係が示せる

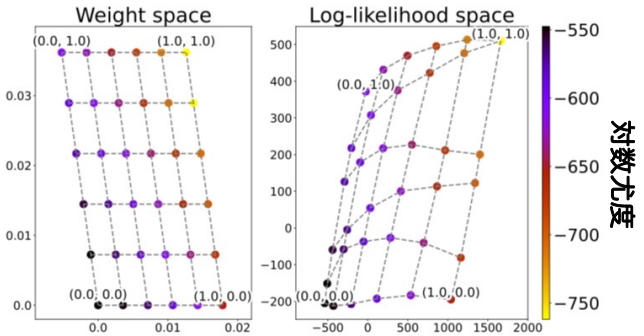
$$2\mathrm{KL}(p_i, p_j) \approx \operatorname*{Var}_{x\sim p_0}\left(\log p_i(x) - \log p_j(x)\right)$$

[下平, 1993; Shimodaira+1998]

・事前に用意したテキスト集合$x_1, \ldots, x_N$についての対数尤度ベクトル$(\log p_i(x_1), \ldots, \log p_i(x_N))$をモデル$p_i(x)$の特徴量とし、これを中心化したベクトルを$\boldsymbol{q}_i$とすると、

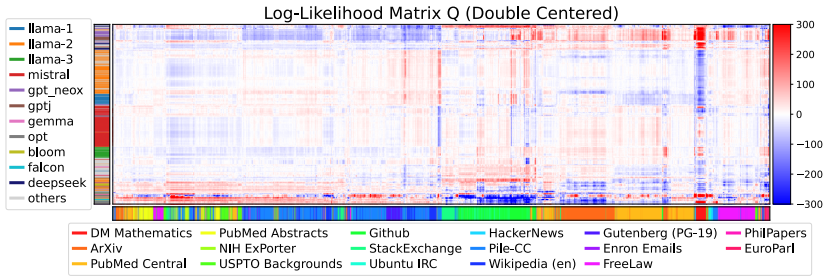
$$2\mathrm{KL}(p_i, p_j) \approx \frac{1}{N}\left\|\boldsymbol{q}_i - \boldsymbol{q}_j\right\|^2$$

・言語モデル$p_i(x)$のユークリッド座標を$\boldsymbol{q}_i$とすれば、座標間の距離の２乗がKLダイバージェンスを表すから、言語モデルを確率分布の空間で表した図が得られる



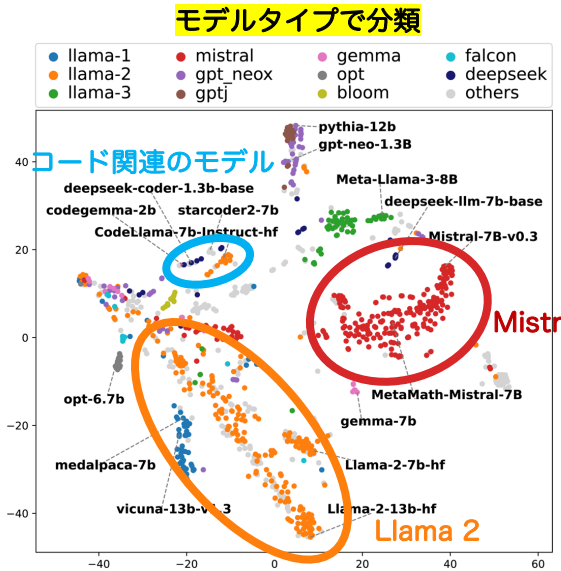
ニューラルネットの重み（左）と対数尤度ベクトル（右）のPCAによる可視化

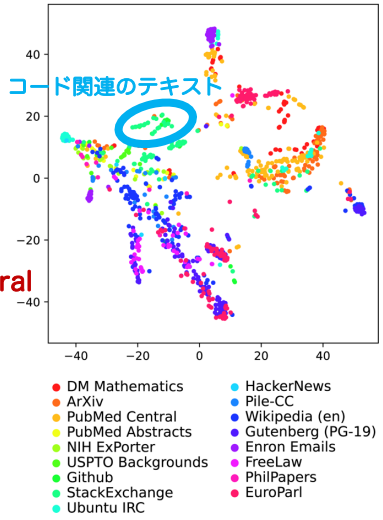
③1018個の言語モデルで10000個のテキストの対数尤度を計算

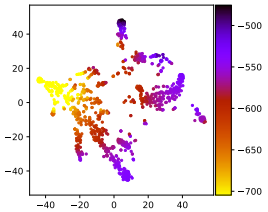


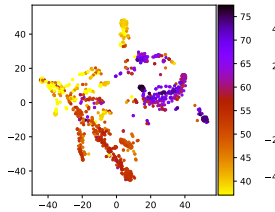
④1018個の言語モデルの地図からモデル達の全体像を把握

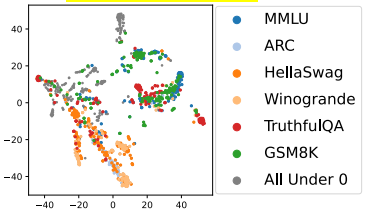
・対数尤度ベクトルのt-SNEによる可視化











言語や画像の埋め込みに共通する「形」の発見

Yamagiwa, Oyama, Shimodaira (EMNLP 2023) Discovering Universal Geometry in Embeddings with ICA

① 背景と目的

・言語モデル等の埋め込みは「分散表現」であり，ベクトルの成分に特に意味はない
・解釈性を向上したい
・言語モデルがなぜ良いのかを解明したい
・次元削減に広く利用されている主成分分析(PCA)は「白色化」で成分を無相関にする

$$\mathbf{Y} = \mathbf{X}\mathbf{A}$$

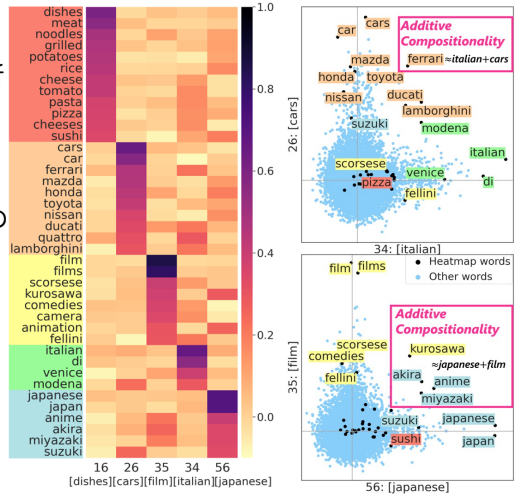
・独立成分分析 (ICA)のFastICAは各成分の非ガウス性が最大（エントロピーが最小）となるようにさらに直交変換して，成分を確率変数として独立に近づける

$$\mathbf{S} = \mathbf{X}\mathbf{A}\mathbf{R}_{\mathrm{ica}}$$

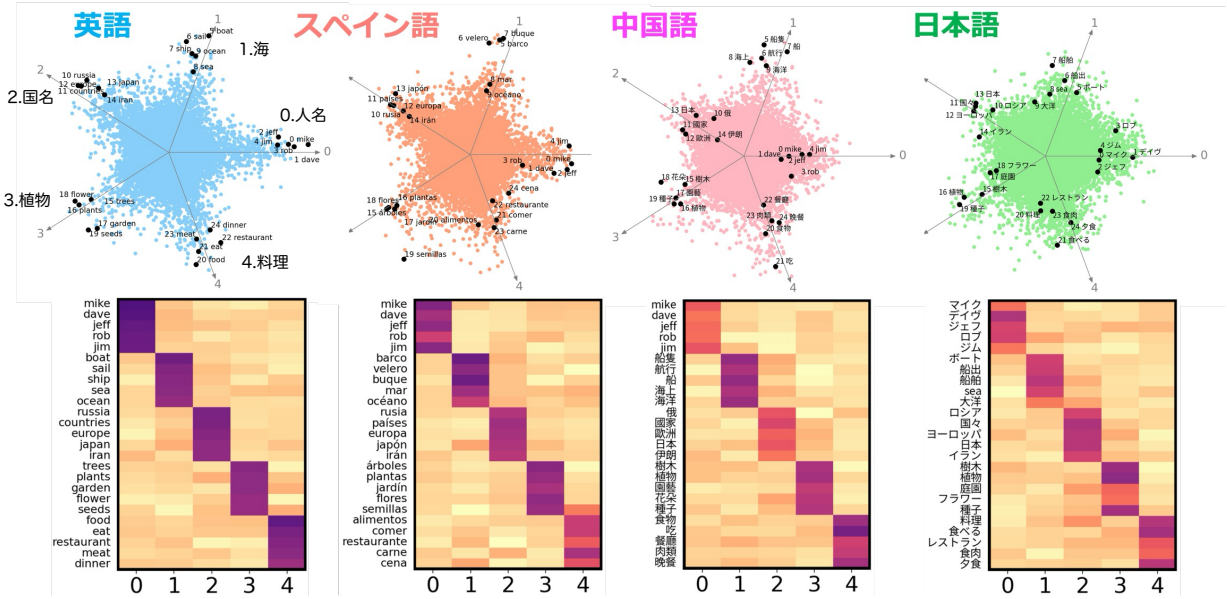
・埋め込みをICA変換する先行研究はほとんどなかった

② ICA 変換後の埋め込みのスパース性

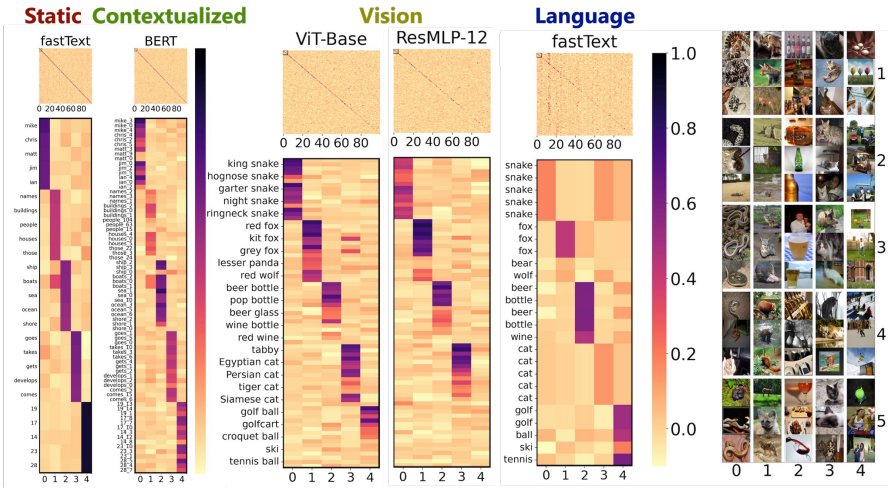
・各成分が意味の独立な軸となる
・意味の軸方向に尖った形の分布を形成



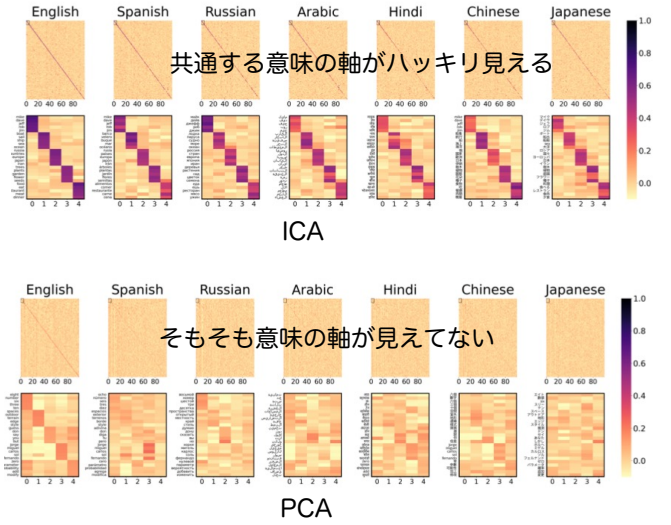
③ ICA 変換後の埋め込みには言語横断的に共通の意味の軸が存在



④動的・画像埋め込みでも共通の意味の軸を発見



⑤ICAは軸がよく見えるがPCAは見えない



埋め込みのノルム（長さ）は「意味の強さ」を表す

Oyama, Yokoi, Shimodaira (EMNLP 2023) Norm of Word Embedding Encodes Information Gain

① 背景，目的，結果

・言語モデル等で中核的な存在の埋め込みを理解したい
・言語モデルがなぜ良いのかを解明したい
・埋め込みの方向は意味を表すが，長さは何を表す？
・単語埋め込みのノルムの２乗と単語のKL情報量が線形に関係していることを発見

※実際は理論つくってから実験した

② KL情報量

・「分布仮説」（Harris, 1954）：単語wの意味は文書データにおける周辺単語w'の分布p(w'|w)で定まる
・文書データ全体の単語分布p(w')からのズレをKL情報量で測る

※ベイズの意味で事前分布から事後分布への情報ゲインに相当し，「意味の強さ」を表すと解釈する

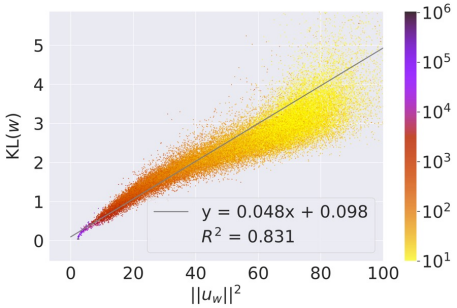
$$\mathrm{KL}\big(p(\cdot\,|w)\,\|\,p(\cdot)\big) = \sum_{w'\in V} p(w'|w)\log\frac{p(w'|w)}{p(w')}$$

$p(\cdot\,|w)$
単語 w の周辺単語分布
wを与えたときの事後分布

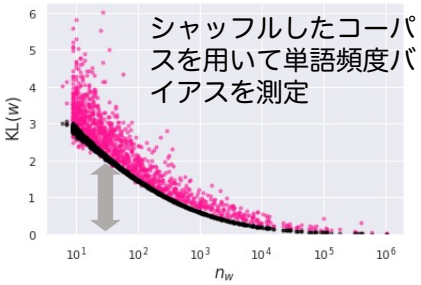
$p(\cdot)$
ユニグラム分布
事前分布

Top 10		Bottom 10	
word	$\mathrm{KL}(w)$	word	$\mathrm{KL}(w)$
rajonas	11.31	the	0.04
rajons	10.82	in	0.04
dicrostonyx	10.31	and	0.04
dasyprocta	10.27	of	0.05
stenella	10.24	a	0.07
pesce	10.22	to	0.09
audita	10.09	by	0.09
landesverband	10.05	with	0.10
auditum	9.96	for	0.10
factum	9.84	s	0.10

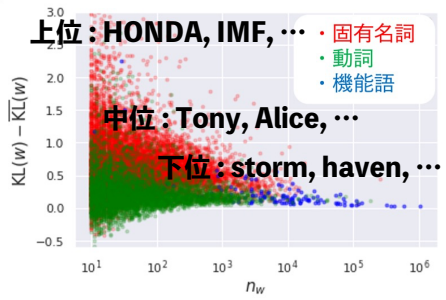
KL情報量のトップ10とワースト10



単語ベクトルの長さの２乗はKL情報量と強い相関



単語出現頻度によるKL情報量のバイアス



KL情報量をバイアス補正したもの

③指数型分布族の理論で証明

・SGNSモデルは指数型分布族でかける
・言語モデルのsoftmax層も指数型分布族で書ける
・中心化と白色化で適切にパラメータ変換
・情報幾何学 (Amari 1982)の観点からみればノルムとKL情報量の関係はほぼ自明

KLは適切に白色化した埋め込みのノルムの２乗に比例

$$2\mathrm{KL}(w) \simeq \|\tilde{u}_w\|^2$$

Skip-Gramモデル $p(w'|w) = \nu q(w') e^{\langle u_w, v_{w'}\rangle}$

埋め込みの中心化 $\hat{u}_w := u_w - \bar{u}$

埋め込みの（特殊な）白色化 $\tilde{u}_w := G^{\frac{1}{2}}\hat{u}_w$

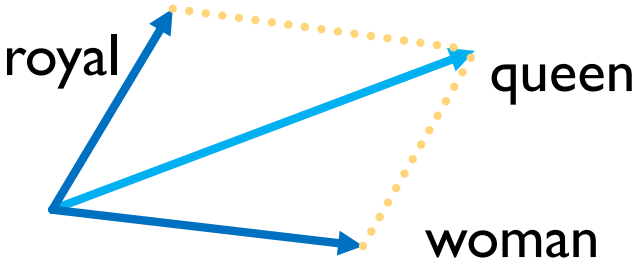
大山，横井，下平（2022）「単語ベクトルの長さは意味の強さを表す」言語処理学会第28回年次大会(NLP2022)において若手奨励賞

単語埋め込みでは，文書データから教師なし学習によって「単語の意味」の分散表現（単語ベクトル）を計算する．ベクトルの加減算で類推などの「意味の計算」ができることが知られている（例：king - man + woman = queen）．**将来，高度な思考を実現するには，複雑な意味の計算が必要になる**．その基礎となる**加法構成性**では，ベクトルの和が構成要素の**ANDの意味**に相当する．

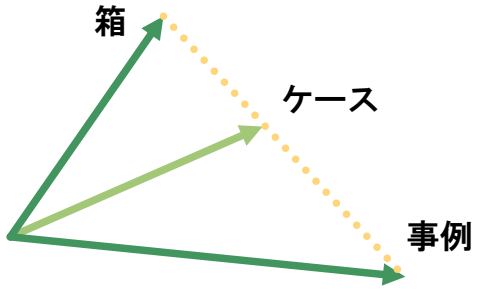
- それでは，**ORやNOTはどのように計算するのか？**
- 単語頻度による重み付き平均が**ORの意味**に相当
- NOTは むずかしい(母の反対は娘か父か？)
- 対象となる単語集合に応じて「条件付き埋め込み」の概念を導入し，ベクトルの負方向が**NOTの意味**に相当することを明らかにした

queen = royal + woman
king = royal + man

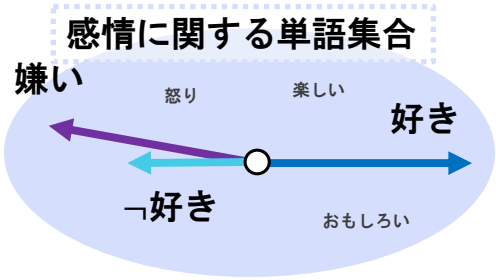
加法構成性の例．これらからroyalを消去すると， 類推の計算（king - man + woman = queen）が得られる



AND（加法構成性）



OR（ベクトルの重心）



NOT（集合内で負方向）

内藤，横井，下平（2021）「単語埋め込みによる論理演算」
言語処理学会第27回年次大会(NLP2021)において委員特別賞

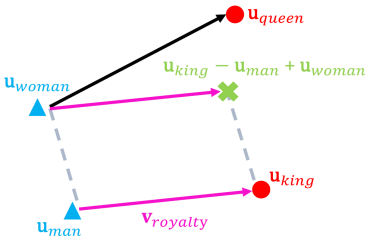
単語埋め込みで薬剤と遺伝子間のアナロジータスクを解く

Yamagiwa et al. (2024): Predicting Drug-Gene Relations via Analogy Tasks with Word Embeddings (arXiv:2406.00984)

① 背景

単語埋め込み（単語ベクトル）の加減算による意味類推タスク

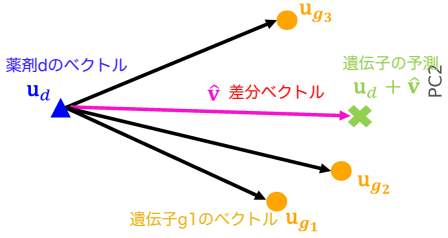
$$\mathbf{u}_{king} - \mathbf{u}_{man} + \mathbf{u}_{woman} \approx \mathbf{u}_{queen}$$



② 研究の目的

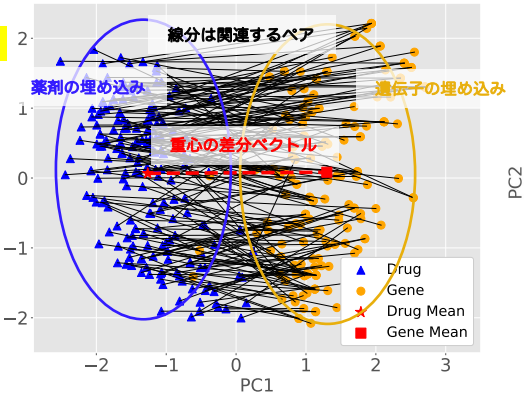
・ベクトルの足し算だけで薬剤から関連遺伝子を予測したい
（引き算で遺伝子から薬剤の予測）

$$\mathbf{u}_d + \hat{\mathbf{v}} \approx \mathbf{u}_g$$



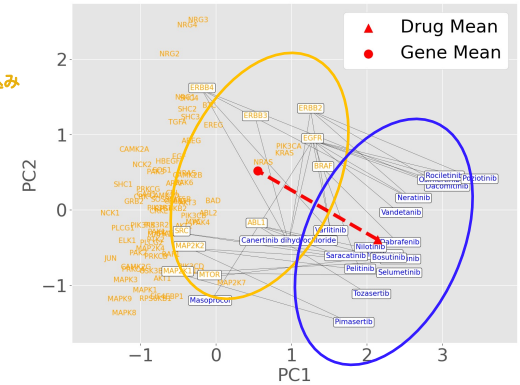
③薬剤と遺伝子の意味は平行だった

200ペアの埋め込みをPCAで可視化



④パスウェイ毎にみるとより正確

ErbB signaling pathway の薬剤・遺伝子



⑤ 実験設定と性能評価（最近までのデータ使用）

・事前学習された埋め込み：BioConceptVec (Chen et al. 2020)
・本研究ではSkip-Gram (Mikolov et al. 2013)モデルで学習
・データ：PubMedの3500万アブストラクト (2023年まで)
・遺伝子数 28000，薬剤数 51000，関連性 6000
・パスウェイ情報はKEGGから抽出，関係性情報はAsuratDBから抽出

・パスウェイ情報で精度向上
・Top1精度 = 60%
・Top10精度 = 86%
・想像以上に高精度
・大規模言語モデルがなぜ良いか、動作の仕組みを示唆

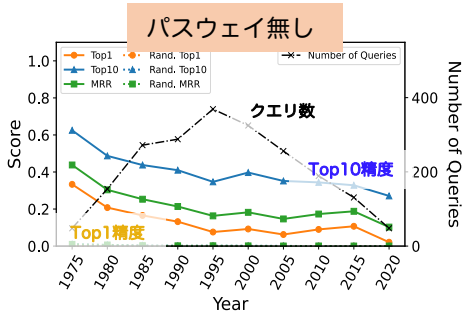
モデル	Top1精度	Top10精度
BioConceptVec（パスウェイ無し）	0.304	0.645
BioConceptVec（パスウェイあり）	0.499	0.790
Skip-Gram（パスウェイ無し）	0.300	0.686
Skip-Gram（パスウェイあり）	**0.589**	**0.862**
GPT-4o（大規模言語モデル）	0.760	0.887

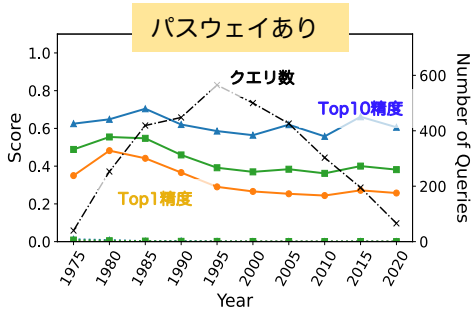
予測の具体例（正解遺伝子と関連遺伝子）

Drug	Setting	Prediction Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8	Top9	Top10	Answer Set
Bosutinib	G	ABL1	TXK*	BCR	BTK	PDGFRB	FYN	FLT3	HCK	KIT	CSK	{ABL1, SRC}
	P1	ABL1	TXK*	EGFR	BCR	ALK	RUNX1	FLT3	KIT	JAK2*	ERBB2	{ABL1, SRC}
	P2	ABL1	TXK*	EGFR	BCR	ALK	RUNX1	FLT3	KIT	JAK2*	ERBB2	{ABL1, SRC}
Masoprocol	G	ALOX5*	ALOX12	PTGS2	–	PTGS1	ALOX15B	–	PTGES	COMT	HDAC6	{EGFR, IGF1}
	P1	ALOX5*	ERBB2	TP53*	PTGS2	EGFR	HSP90AA1	TERT	COX2	EREG	GSK3A	{EGFR}
	P2	ALOX5*	ERBB2	TP53*	PTGS2	EGFR	HSP90AA1	TERT	COX2	EREG	GSK3A	{EGFR, IGF1}
Poziotinib	G	ERBB3	EML4*	FGFR2	ROS1*	PIK3CA	ALK*	MET	ERBB4	FGFR4	NRAS	{EGFR, ERBB2}
	P1	EGFR	ERBB2	ERBB3	ROS1*	ALK*	PIK3CA	MET	NRAS	FGFR2	MAP2K7	{EGFR, ERBB2}
	P2	EGFR	ERBB2	ERBB3	ROS1*	ALK*	PIK3CA	MET	NRAS	FGFR2	MAP2K7	{EGFR, ERBB2}
Selumetinib	G	MAP2K7	MAP2K2	CDK4	RAF1	PIK3CA*	PIK3R1	MAP2K1	CDK6	NRAS	BRAF*	{MAP2K1, MAP2K2}
	P1	MAP2K7	BRAF*	PIK3CA*	KRAS	MAP2K1	EGFR	NRAS	CDK4	MAPK1	PTEN	{MAP2K1, MAP2K2}
	P2	MAP2K7	BRAF*	PIK3CA*	KRAS	MAP2K1	EGFR	NRAS	CDK4	MAPK1	PTEN	{MAP2K1, MAP2K2}
Trametinib	G	MAP2K7	MAP2K2	CDK6	CDK4	MAP2K1	RAF1	MAPK1	BRAF*	MAPK3	PIK3CA*	{MAP2K1, MAP2K2}
	P1	MAP2K7	EGFR	BRAF*	MAPK1	KRAS	MAP2K1	MAPK3	CDK4	CD274	CDK6	{MAP2K1, MAP2K2}
	P2	MAP2K7	EGFR	BRAF*	MAPK1	KRAS	MAP2K1	MAPK3	CDK4	CD274	CDK6	{MAP2K1, MAP2K2}

⑥過去のデータから未知の関係性を予測

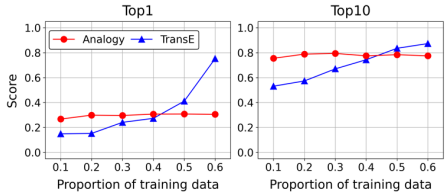
・PubMedのアブストラクトをy年以前とy年より後に分割 (y=1975,…,2020)
・過去のPubMedおよび過去の関係性のみを利用して，未来の未知の関係性を予測
・各y毎にSkip-Gramによる埋込を学習した
・関係性の初出年はPubMedアブストラクト中の単語共起から推定した
・Top10精度はパスウェイ無しで40%~50%，パスウェイありで60%程度





⑦教師ありの手法（知識グラフ埋め込み）との比較

・知識グラフ埋め込みのTransEで(薬剤,遺伝子)ペアの教師データを学習させた手法と比較
・(薬剤,遺伝子)ペアを明示的に教えなくても、単語埋め込みのアナロジータスクは関係性を予測できる



不均衡最適輸送を用いた意味変化検出

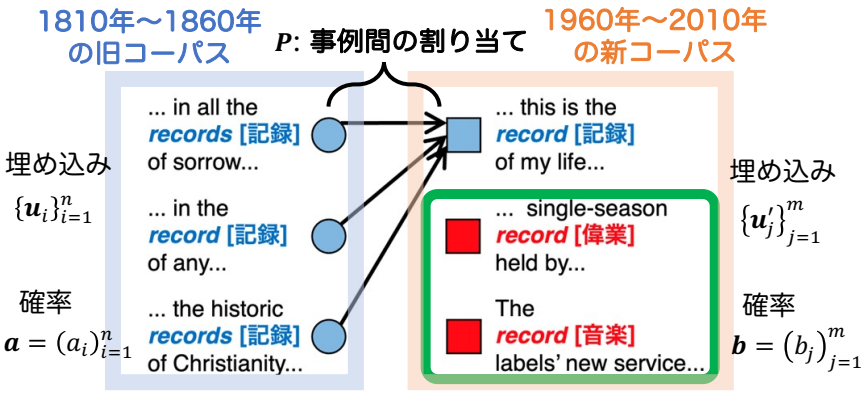
Kishino, Yamagiwa, Nagata, Yokoi, Shimodaira (arXiv:2412.12569) Quantifying Lexical Semantic Shift via Unbalanced Optimal Transport

① 背景と目的

・時代経過で単語の意味は変化する（消失、出現）
・異なる年代コーパスでの語義の変化を調べたい

年代	使用事例	語義	意味変化
1960–2010	... So did Sire ***Records***...	[音楽]	0.47
1960–2010	... a team with the third-worst ***record***...	[偉業]	0.45
1960–2010	... the AMCU single-season ***record***...	[偉業]	0.45
1810–1860	... interpretations of the Mosaic ***record***...	[記録]	-0.23
1810–1860	... the ***records*** of a professed revelation...	[記録]	-0.24
1810–1860	... the ***record*** of whose wisdom is included in...	[記録]	-0.25

②旧・新コーパスの事例の埋め込み間で不均衡最適輸送を計算



出現した語義への輸送がない

$$\min_{P \in \mathbb{R}_{\geq 0}^{n \times m}} \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}P_{ij} + \lambda\|P\mathbf{1}_m - a\|^2 + \lambda\|P^\top \mathbf{1}_n - b\|^2$$

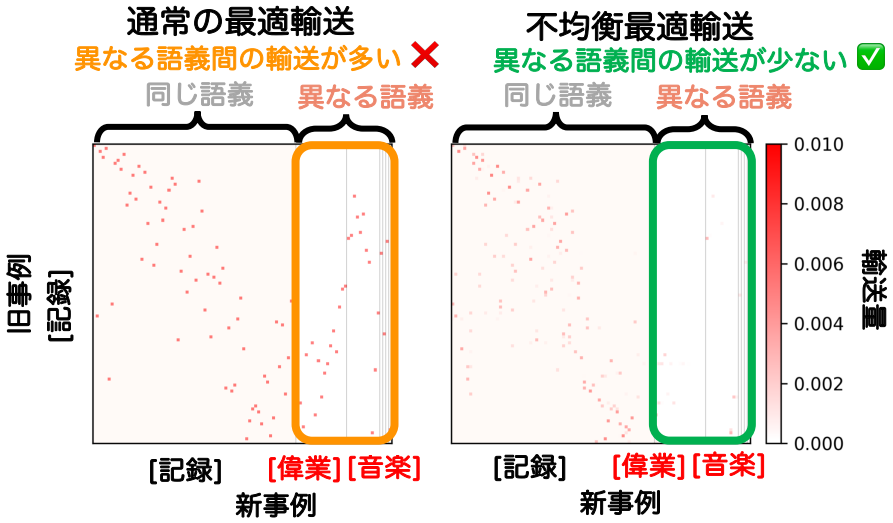
C_{ij} = u_i, u'_j 間のコスト関数　　重みの過不足によるコスト

③輸送量の歪みから意味変化の指標 (SUS)を定義

$$\mathrm{SUS}(u_i) = -\left(a_i - \sum_j P_{ij}\right)/a_i \qquad \mathrm{SUS}(u'_j) = \left(b_j - \sum_i P_{ij}\right)/b_j$$

・単語の使用事例毎に、その語義での使用頻度の変化を測定

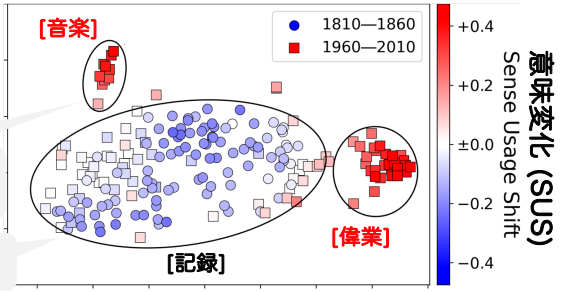
④通常の最適輸送と不均衡最適輸送の輸送量の比較



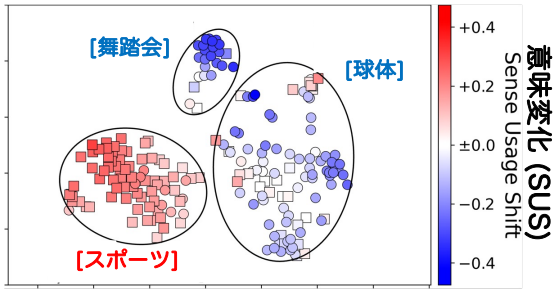
⑤旧・新コーパスの事例の埋め込みをt-SNEで可視化

record の埋め込み
・[記録]は減少
・[音楽],[偉業]が出現

SUS が大きい & 旧事例 ●が無い
SUSが小さい & 新事例 ■が少ない

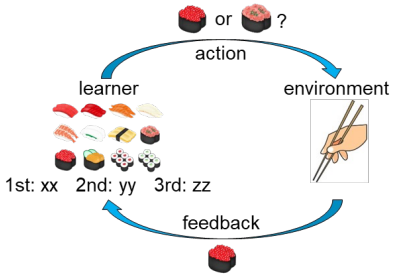


ball の埋め込み
・[舞踏会],[球体]が減少
・[スポーツ]が増加

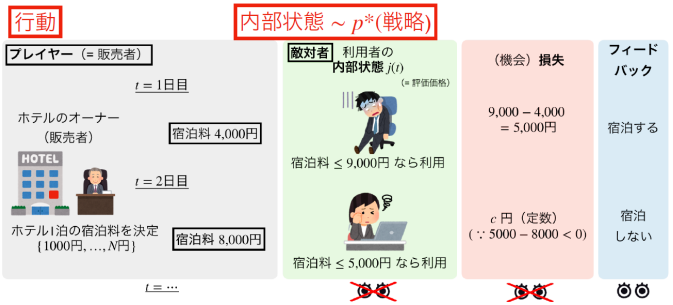


岸野, 山際, 永田, 横井, 下平 (2025)
言語処理学会第31回年次大会(NLP2025)において委員特別賞

線形モデル

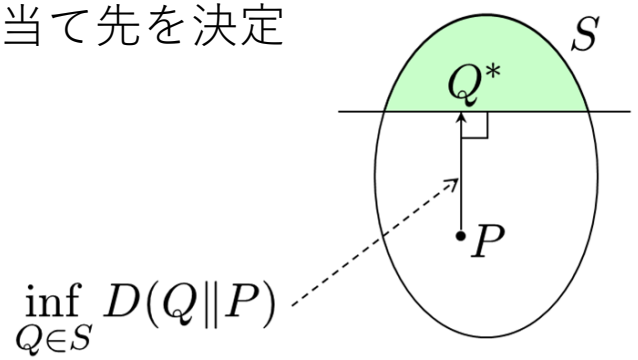


相対比較モデル

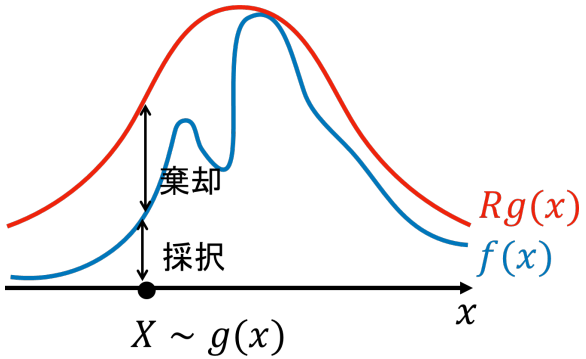


価格決定モデル

- **バンディット問題**：広告推薦や治験，価格決定問題など，事前にデータがない状態で逐次的にデータを集めながら次の割り当て先を決定
- どうやって現在の推定の不確かさを考慮する？
- 考える設定に応じて多様なモデリングや定式化



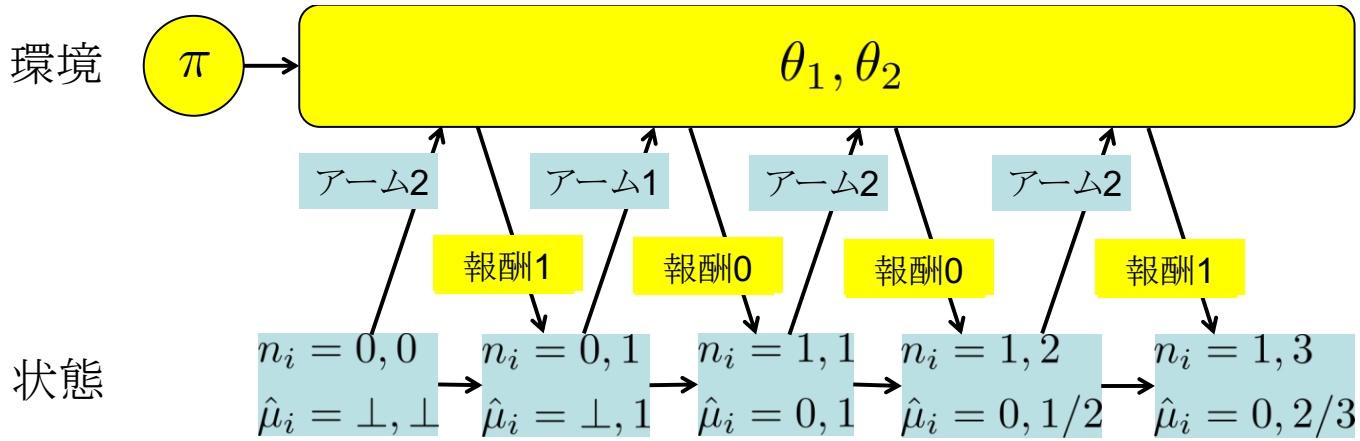
- 真の分布の候補をKL情報量を用いて抽出する！
- 理論限界を達成するアルゴリズムを
さまざまな設定で構築可能



KL情報量を明示的に用いると計算が大変

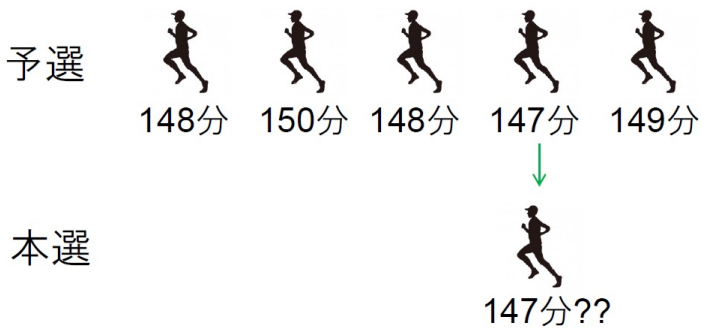
ベイズ統計に基づく事後分布サンプリングを適切に用いて複雑な計算を回避！

バンディット問題



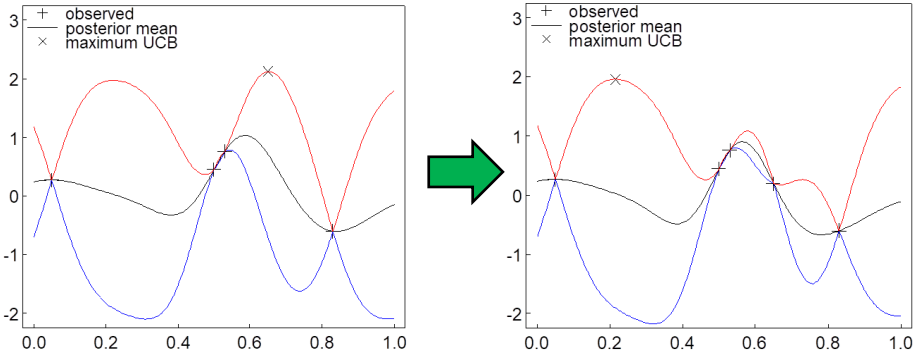
- **バンディット問題と強化学習**： バンディット問題は強化学習の特別な場合だが，実際は標準的な手法が大きく異なる
- 強化学習の手法を用いてもバンディット問題のアルゴリズムを作れる？
- 実際はランダム性の大きさ等の要因から単純なタスクを除いて学習が難しい
- 学習しやすい報酬を適切に設定して，複雑なモデル上でもうまく動くアルゴリズムを強化学習で自動設計！

- バンディットアルゴリズムを通じて得られたデータからはバイアスのない推定が難しい（選択的推定）
- 強化学習の手法を用いて推定可能性を検証する！

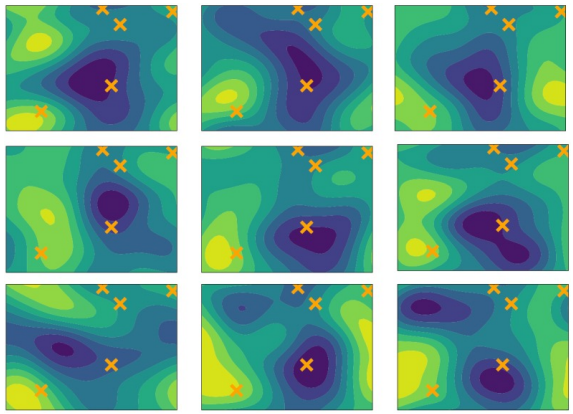


強化学習を用いたバンディットアルゴリズム

- **ベイズ最適化**： 複雑な形状の未知関数をガウス過程により表現，なるべく少ない関数評価で最大値を発見
- 各点での関数値の不確かさを定量化
- 深層学習のパラメータチューニングから物性の予測までさまざまな関数・タスクを表現できる！

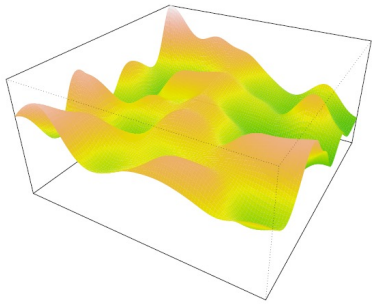


- 異なる観測コストと精度が選択できる場合，入力可能な点が事前に分からない場合，空間センシングなど時間変化がある場合等にも柔軟に対応可能

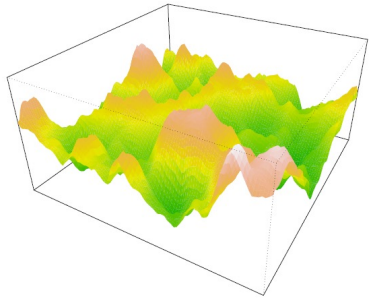


ガウス過程による電気伝導度の推定

カーネル関数を適切に設計して
関数の滑らかさをモデリング



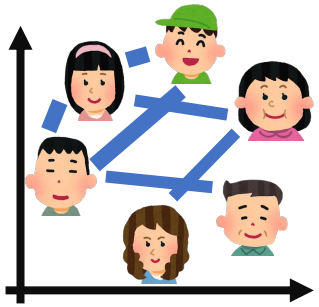
ガウスカーネル



マターンカーネル

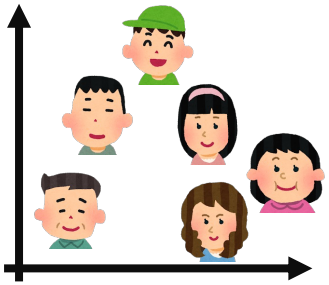
座標とリンク

（リンクだけでも良い）



ニューラルネット

座標

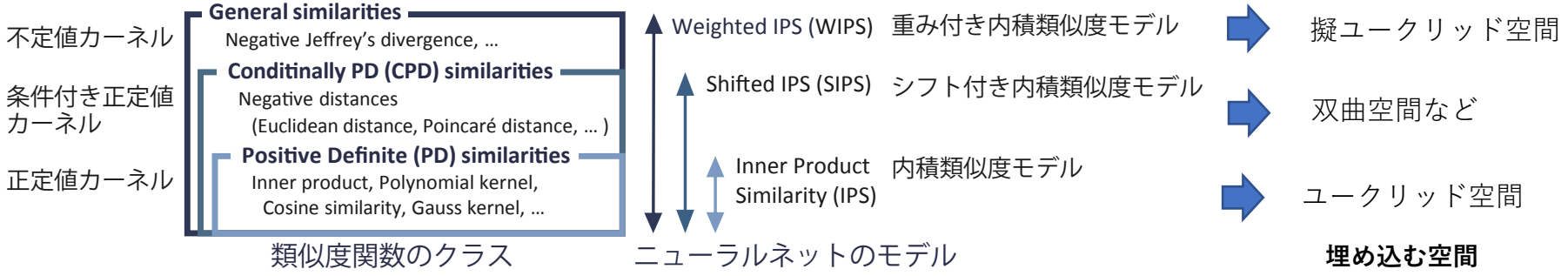


グラフのリンク構造をなるべく保存するようにノードの座標を計算する．統計学の多変量解析の拡張になっていて，機械学習で応用がたくさんある．

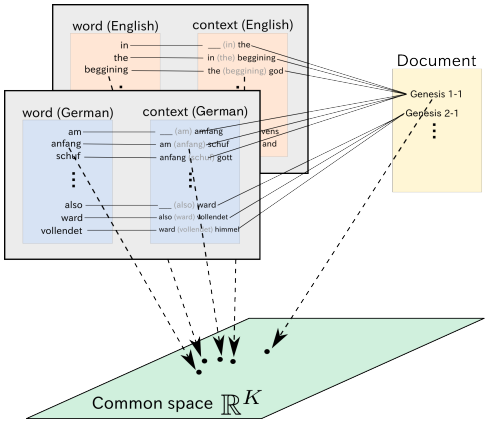
- 一般にノードの類似度をベクトルの内積やコサイン類似度で測る
- 類似度関数を変えたら，もっと性能あがるのでは？

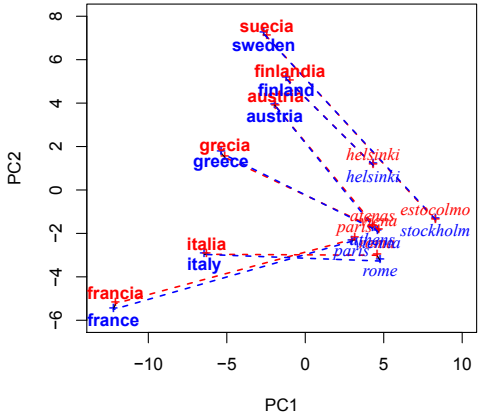
- 内積だけでも任意の正定値カーネルが表現できる！
- シフト項をいれるだけで，双曲空間への埋め込みが表現できる！
- 重み付き内積モデルにすれば，もっと表現できる！

エッシャーの絵は双曲幾何の例
（ポアンカレ埋め込み）

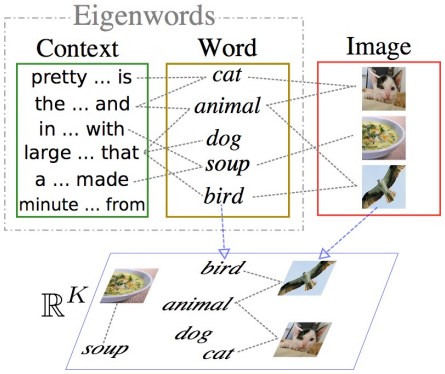


グラフ埋め込みの理論

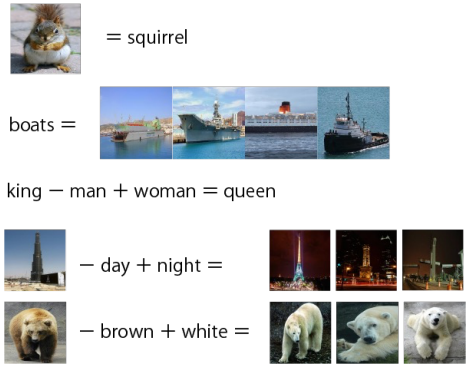




- 対訳コーパスのリンクだけから単語埋め込み
- 辞書を教えなくても，２言語の翻訳単語の座標は近くなる



- 文書データ，ソーシャルメディアの写真とタグから，単語と画像を同時埋め込み
- 画像と単語の相互検索や，意味の加減算による検索ができる



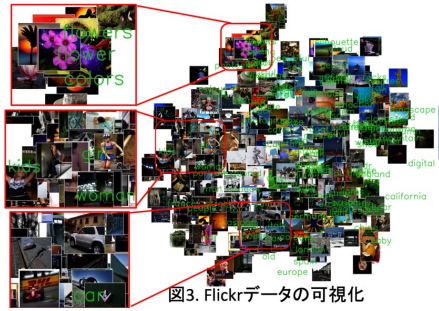


図3. Flickrデータの可視化

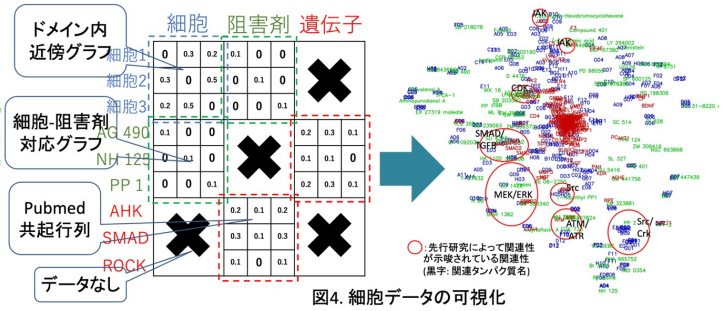


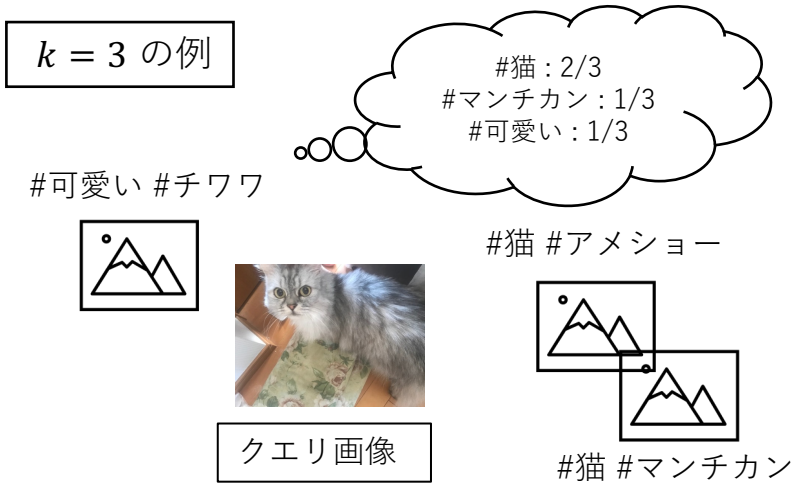
図4. 細胞データの可視化

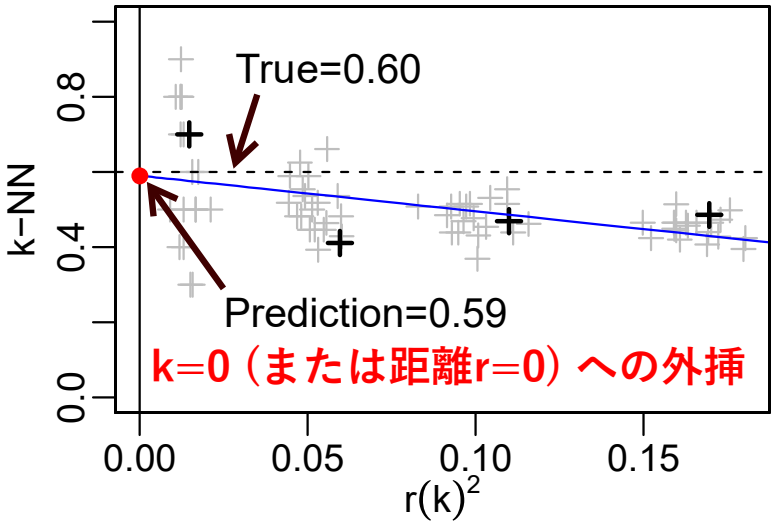
- 文書，写真，遺伝子データなど混在した関連性データを扱える
- t-SNEなどの可視化を拡張

グラフ埋め込みの応用

k-近傍法は最も単純な判別，分類の手法．たとえば，クエリ画像に特徴量ベクトルが最も近い画像をk個選び，そのラベルの平均値を出力する．**小さいkはバイアスが小さく，大きいkは分散が小さくなり，両者の影響はトレードオフの関係**にある．どうやって最適なkを選ぶかが問題になっていた．

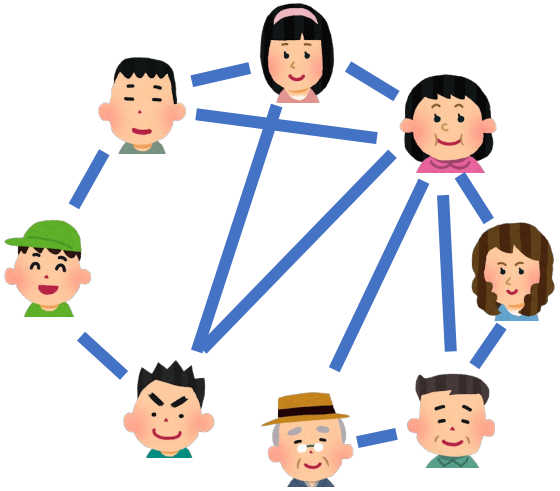
- 理論上はk=0とすればバイアスがゼロになるハズ（分散は無限になってしまう）
- **複数のkの値でk-近傍法を実行し，それをk=0に外挿すればよいのでは？**
- 架空の「0-近傍法」を計算するマルチスケールk近傍法を提案
- これが収束レートに関して最適解であることを証明
- 外挿法の工夫によって改良し，Flickr画像のタグ推定に応用





Okuno, Shimodaira (2020) “Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate” (NeurIPS 2020)
田中，奥野，下平 (2021)「マルチスケールk-近傍法による画像のExtreme Multi-Label分類」(JSAI2021)
Cao，田中，奥野，下平 (2021)「マルチスケールk-近傍法における回帰関数および損失関数の検討」(JSAI2021)

マルチスケール k-近傍法



- 成長するネットワーク（グラフ）の確率モデル
- 新しいリンクの確率が高いのは？

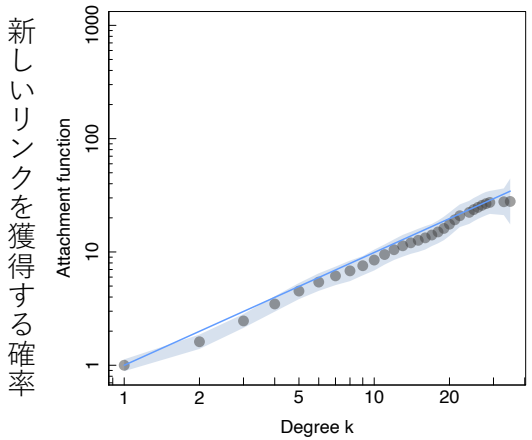
1. リンク数の多い人

（実は新人なのにすでに２人）

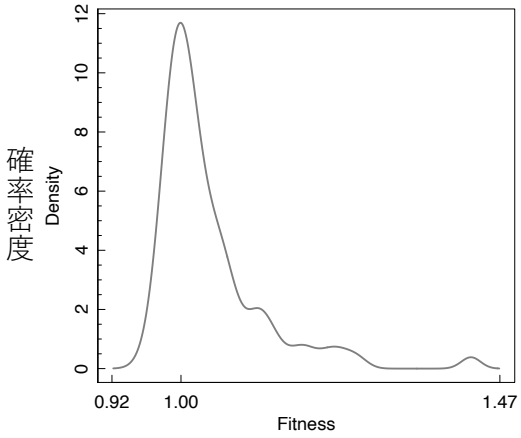
2. ポテンシャルの高い人

3. 共通する友達の多いペア

これらの効果を分離して推定する統計手法とソフトウエア(PAFit, FoFaF)の開発



現在のリンク数



新しいリンクを獲得する潜在的能力

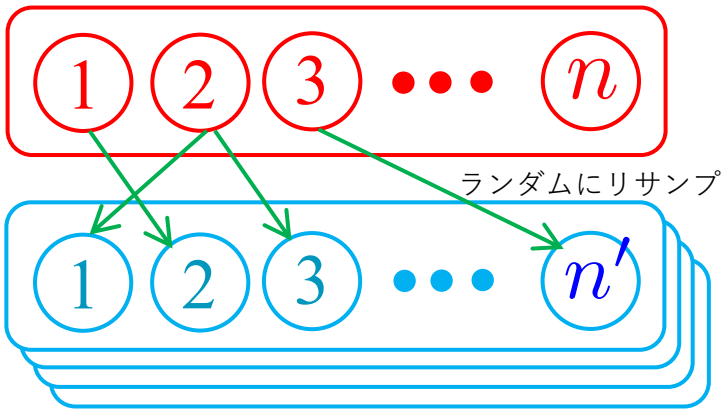
Rank	Estimated fitness	Name
1	1.42	BARABASI, A
2	1.35	NEWMAN, M
3	1.26	JEONG, H
4	1.25	LATORA, V
5	1.24	ALON, U
6	1.23	OLTVAI, Z
7	1.23	YOUNG, M
8	1.22	WANG, B
9	1.21	SOLE, R
10	1.21	BOCCALETTI, S

「コラボする潜在能力」の高い研究者トップ10

Pham, Sheridan, Shimodaira (2020)

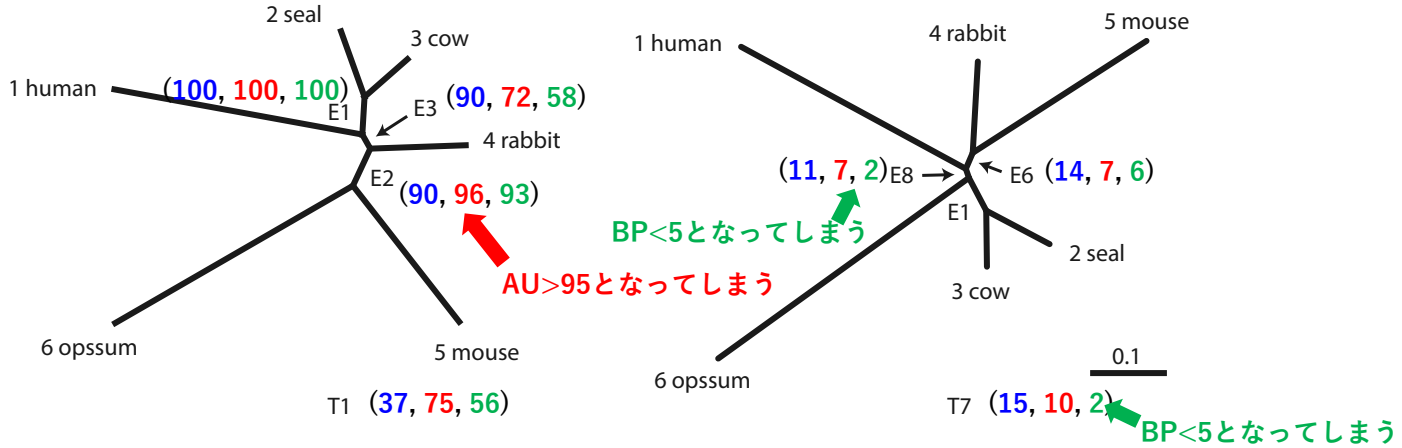
論文共著者ネットワークデータ（複雑ネットワーク分野）からの推定結果

- ブートストラップ法は，データからリサンプリング（復元抽出）
- 擬似的にたくさんデータができるので，何回もデータ解析すれば，そのバラツキもわかる
- 簡単なアルゴリズムだが，背後には確率分布の空間における距離とか曲率など高度な数理統計学



- 同じ結果が出た回数を素朴に数える(BP)は，「ニセの発見」になりやすい！
- マルチスケール・ブートストラップ法(AU)は，**n’ = -n** としてバイアスゼロ！
- じつは「膨大な仮説集合から仮説を選んでるバイアス」を直してなかった
- そのバイアスも新たに開発したばかりの選択的推測(SI)で解決できた！

SI=Selective Inference 選択的推測（提案法），AU=Approximately Unbised 近似的に不偏な検定(従来法），BP=ブートストラップ確率（ベイズ）



系統樹推定の例．T1はデータが支持する系統樹だが，実際にはT7が真実の系統樹と考えられる．