# 数理統計学チーム　　下平 英寿
# Mathematical Statistics Team
## Hidetoshi Shimodaira

## 数理統計学チーム@京大クラスタ

### 下平研究室（情報学研究科，京都大学）

下平英寿（チームリーダー），Thong Pham（特別研究員），奥野彰文（テクニカルスタッフI），福井一輝（パートタイマー），井上雅章（パートタイマー），KIM GEEWOOK（パートタイマー），田中卓磨（パートタイマー），岩田具治（客員研究員），寺田吉壱（客員研究員），伊森晋平（客員研究員），廣瀬慧（客員研究員），白石友一（客員研究員）

### 統計的推測の方法論の研究

- **帰納的推論の原理の探求**：とくに多数の仮説について統計推測するときの信頼度計算の方法論．多重検定，多重比較，モデル選択，選択的推測などの手法を開発する理論研究．系統樹推定や発現解析など生命科学における標準手法になっている．今後，AIによる「発見」（仮説生成と検証）を実用化するときの基盤技術になる．
- **情報統合の多変量解析と適切な情報表現の探求**：とくにマルチモーダル関連性データのグラフ埋め込みによる表現学習．画像認識や自然言語処理での実践や擬ユークリッド空間への埋め込みや加法構成性の理論など．今後，AIによる「高度な思考」を実装するときの基盤技術になる．

### これまでの代表的な研究成果

- **マルチスケール・ブートストラップ法による信頼度計算**：複雑な機械学習による予測や推定値の信頼区間や仮説のp値でも使える汎用手法を弊研究室で開発し，遺伝子発現解析等の標準手法となっている（4論文の被引用数＞9000，50近くの国際特許でも利用）．しかし選択バイアスの問題があったが，選択的推測として理論的にほぼ解決に成功した．
- **グラフ埋め込みによる次元削減**：マルチモーダルデータの関連性を深層学習によるグラフ埋め込みとして定式化し，従来の多変量解析などを一般化．表現可能な類似度の関数クラスを明らかにし，従来の内積よりも擬ユークリッド空間，双曲空間への埋め込みは表現力を飛躍的に拡大し，実装も容易であることを証明した．

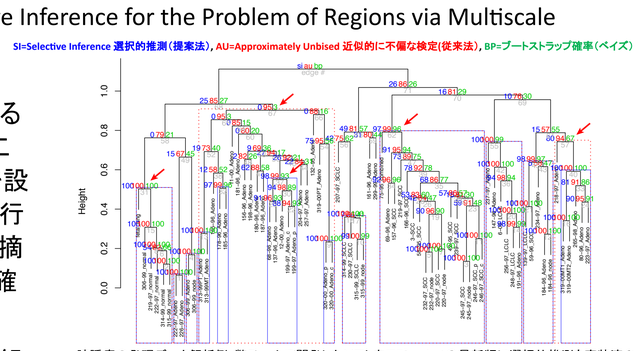
### 目指すゴール，今後の展開

- 頻度論，ベイズ，情報論的方法など従来の統計学・機械学習で帰納的推論の方法論が議論されていたが，「p値の誤用」など様々な問題が指摘されている．AIに限らずあらゆる分野で重要な「データからの推論」のより良い原理を探求する．とくに選択的推論の手法開発と応用を行う．
- 単語ベクトルで確認されている意味の演算（king − man + woman = queen）など，「構成性」に関連してNeurIPS（2019/12）でも注目されているが，このための理論にとりくみ，高度な思考を実現するステップとしたい．

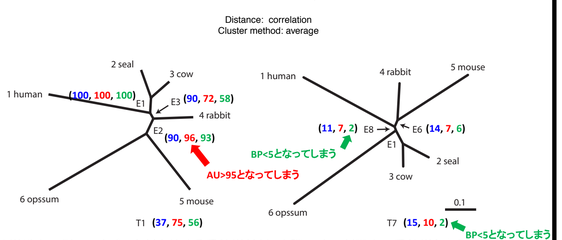
## 選択的推測の理論とクラスタリングへの応用

- Shimodaira, Terada (**arXiv:1902.04964, Frontiers in Ecology and Evolution 2019**), Selective Inference for Testing Trees and Edges in Phylogenetics
- Terada, Shimodaira (**arXiv:1711.00949v2**), Selective Inference for the Problem of Regions via Multiscale Bootstrap

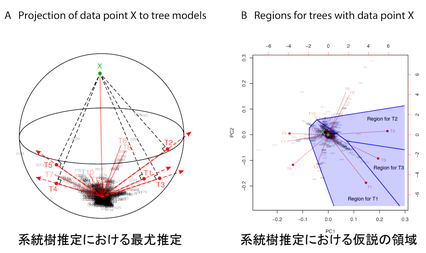
- 従来の統計的仮説検定では，事前に仮説を定める必要がある．ところが医学，科学全般で必ずしもこれが守られず，実際にはデータを見てから仮説を設定し，その同じデータを再び用いて仮説検定を実行するため，偽の発見となりやすくその危険性が指摘されてきた．頻度論のp値の代わりにベイズ事後確率を用いてもこの問題は解決しない
- 「データを見てから仮説を選択すること」を設定に組み込んだ選択的推測 (selective inference)，選択後の推測 (post selection inference) の統計手法や理論が近年，構築されつつある
- 弊チームの先行研究 (Terada, Shimodaira **arXiv:1711.00949v2**)ではマルチスケール・ブートストラップ法 (Shimodaira 2002, 2004, 2008)によってブートストラップ確率のスケーリング則から選択的推測のp-値を計算する数理統計理論を与えている．確率分布の空間における仮説領域の曲率や距離といった幾何学にもとづいてアルゴリズムが構成される
- 本研究ではその手法をもとに発現解析や分子進化系統樹を推定する問題へ実際に応用し，クラスタリングや系統樹のクレードのp値において選択的推測による調整の重要性を示した



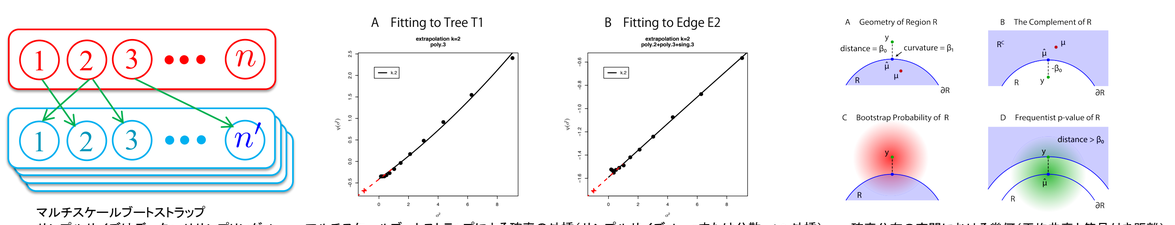
肺腫瘍の発現データ解析例．弊チームで開発したソフトウエアpvclustの最新版に選択的推測も実装済み



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．



系統樹推定における最尤推定　　系統樹推定における仮説の領域



マルチスケールブートストラップ
サンプルサイズはデータn，リサンプリングn′　　マルチスケールブートストラップによる確率の外挿（サンプルサイズn′ = -nまたは分散=-1へ外挿）　　確率分布の空間における幾何（平均曲率と符号付き距離）

## 選択的推測の理論と特徴量選択への応用

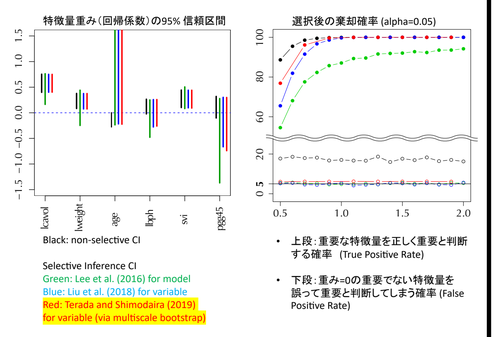
- Lim, Yamada, Jitkrittum, Terada, Matsui, Shimodaira (**arXiv:1910.06134v1, AISTATS 2020**), More Powerful Selective Kernel Tests for Feature Selection [山田チームの共同研究]
- Terada, Shimodaira (**arXiv:1905.10573v3**), Selective Inference after Feature Selection via Multiscale Bootstrap

- さまざまな特徴量（説明変数）から判別や予測を行うとき，データに当てはめて特徴量の重み（回帰係数）を推定する．すべての特徴量を使わずに，Lassoなどの手法で有用な特徴量だけを選択したほうが性能が良くなる
- 選択した特徴量の重みについて信頼区間を計算すると，重み＝0という帰無仮説が有意に棄却されやすくなるバイアスがある．これは選択バイアスの影響であり，本来は重要でない特徴量にもかかわらず，誤って重要と判断されてしまう (偽陽性)
- そこで弊チームの先行研究 (Terada, Shimodaira **arXiv:1711.00949v2**) によるマルチスケール・ブートストラップ法を用いた選択的推測の理論を特徴量選択に適用してアルゴリズムを開発した
- 従来法よりLassoにおける検出力が向上し，本来は重要な特徴量が誤って重要でないと判断されてしまう（偽陰性）をへらすことができた．また，従来法では困難であった非凸の正則化項（SCAD，MCP)に初めて適用できた
- カーネル法による特徴量選択(MMD, HSIC)にも適用できた（山田チーム）

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

全部で8個の特徴量があったが，Lasso は 6個を選択した

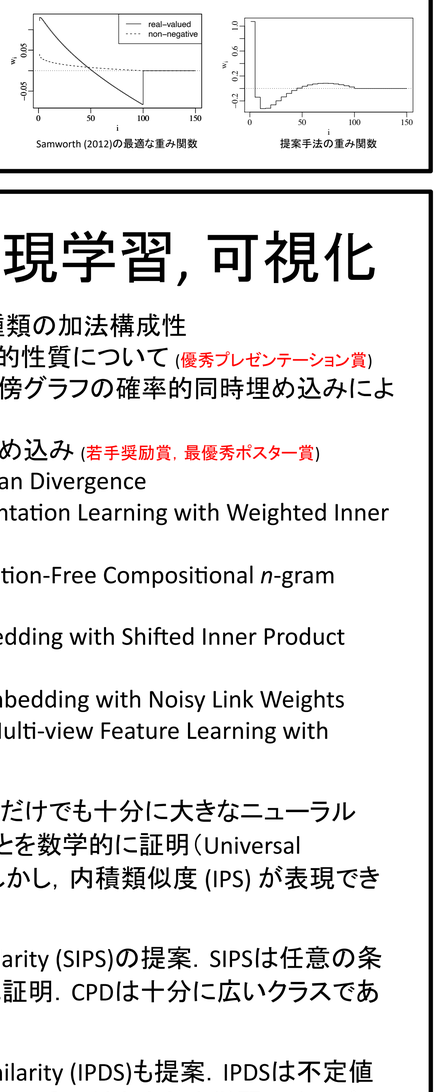
$$\{\beta_{\mathrm{lcavol}}, \beta_{\mathrm{lweight}}, \beta_{\mathrm{age}}, \beta_{\mathrm{lbph}}, \beta_{\mathrm{svi}}, \beta_{\mathrm{pgg45}}\}$$




## k-近傍法 (k-NN) のバイアスをゼロにする

- Okuno, Shimodaira (**arXiv:2002.03054**), Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate
- 田中，奥野，福井，Kim，下平 (**IBIS 2019**), マルチスケールk-近傍法を用いた画像のタグ推定

- k-NNは最も単純な判別，分類の手法．特徴量をみて最も近い過去事例を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は分散が小さくなる．両者の影響は分類誤差でトレードオフの関係にある
- 理論上はk=0とすればバイアスがゼロになる．そこで複数のkの値でk-NNを実行し，それをk=0へ外挿する．分散をおさえつつ，現実には存在しない架空の「0-NN」を計算する「マルチスケール k-近傍法」を提案
- Flickrの画像からタグ推定で有効性を確認
- 提案手法の収束レートは理論的な最適レートを達成




## グラフ埋め込みの学習理論，表現学習，可視化

- Kim, 横井, 下平 (**言語処理学会 2020 to appear**), 単語埋め込みの二種類の加法構成性
- 奥野, 矢野,下平 (**IBIS 2019**), ノンパラメトリックなリンク回帰とその理論的性質について (優秀プレゼンテーション賞)
- 水谷, 奥野, 福井, Kim, 金澤, 白石，岡田 , 下平 (**IBIS 2019**), グラフと近傍グラフの確率的同時埋め込みによるマルチモーダルデータの可視化
- Kim, 奥野, 下平 (**言語処理学会 2019**), 擬ユークリッド空間への単語埋め込み (若手奨励賞，最優秀ポスター賞)
- Okuno, Shimodaira (**arXiv:1908.02573**), Hyperlink Regression via Bregman Divergence
- Kim, Okuno, Fukui, Shimodaira (**arXiv:1902.10409, IJCAI 2019**), Representation Learning with Weighted Inner Product for Universal Approximation of General Similarities
- Kim, Fukui, Shimodaira (**arXiv:1809.00918, NAACL-HLT 2019**), Segmentation-Free Compositional *n*-gram Embedding
- Okuno, Kim, Shimodaira (**arXiv:1810.03463, AISTATS 2019**), Graph Embedding with Shifted Inner Product Similarity and Its Improved Approximation Capability
- Okuno, Shimodaira (**arXiv:1902.08440, AISTATS 2019**), Robust Graph Embedding with Noisy Link Weights
- Okuno, Hada, Shimodaira (**ICML 2018**), A Probabilistic Framework for Multi-view Feature Learning with Many-to-many Associations via Neural Networks

- 弊チームの先行研究(Okuno, Hada, Shimodaira **ICML 2018**) では，内積だけでも十分に大きなニューラルネットワークを併用することで広いクラスの類似度関数を学習できることを数学的に証明（Universal Approximation定理とMercer定理，推定量の漸近的一致性の証明）．しかし，内積類似度 (IPS) が表現できるのは 正定値関数（positive definite, PD)だけである．
- 内積だけでなくバイアス項も学習するモデルShifted Inner Product Similarity (SIPS)の提案．SIPSは任意の条件付き正定値関数 (conditionally PD, CPD)を学習できることを数学的に証明．CPDは十分に広いクラスである (例) ユークリッド空間の距離，Poincare埋め込み，双曲空間の距離
- さらに，2つの内積の差を学習するモデル Inner Product Difference Similarity (IPDS)も提案．IPDSは不定値 (indefinite)カーネルを含む類似度関数を学習できることを数学的に証明．重み付き内積を学習するモデル Weighted Inner Product Similarity (WIPS)の提案と高速でスケーラブルな学習法の実装．重みの値として負も許して学習することにより，不定値 (indefinite)カーネルを含む類似度関数を学習できることを数学的に証明および実験で検証

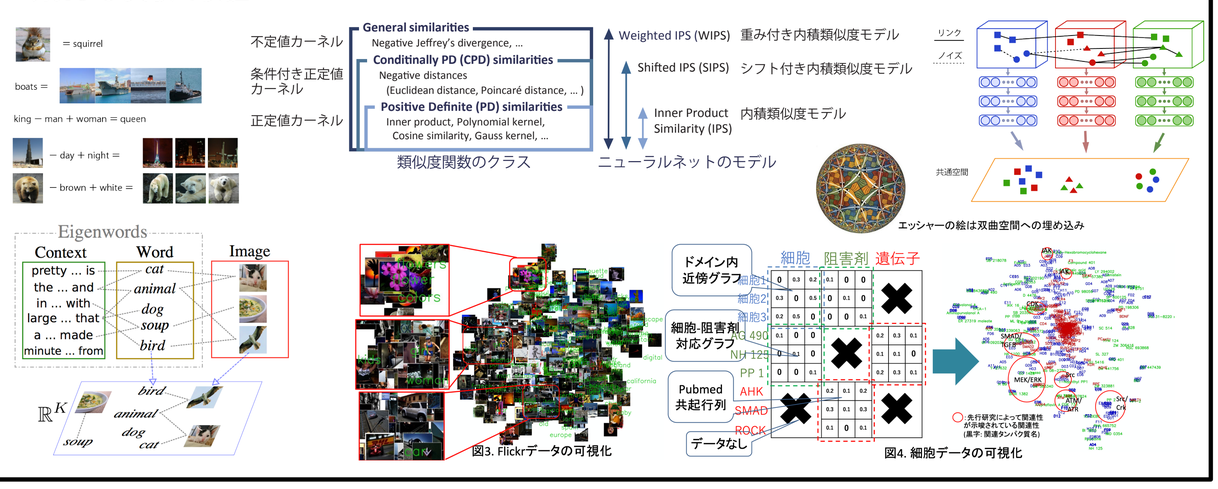


図3. Flickrデータの可視化



図4. 細胞データの可視化