

## 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)

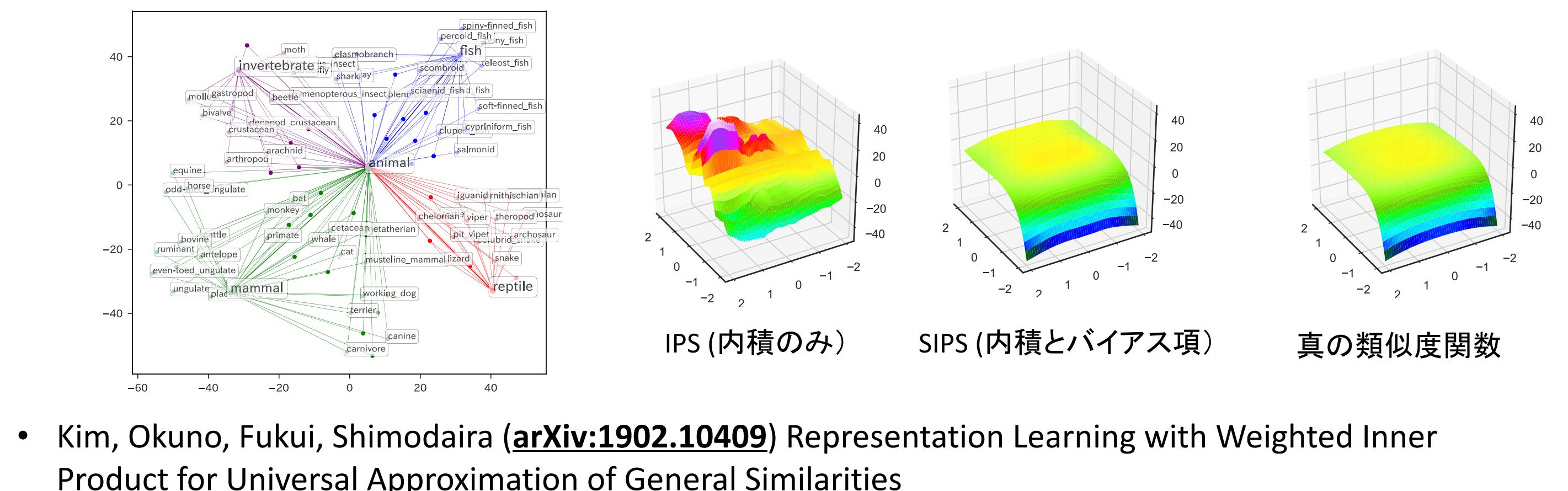
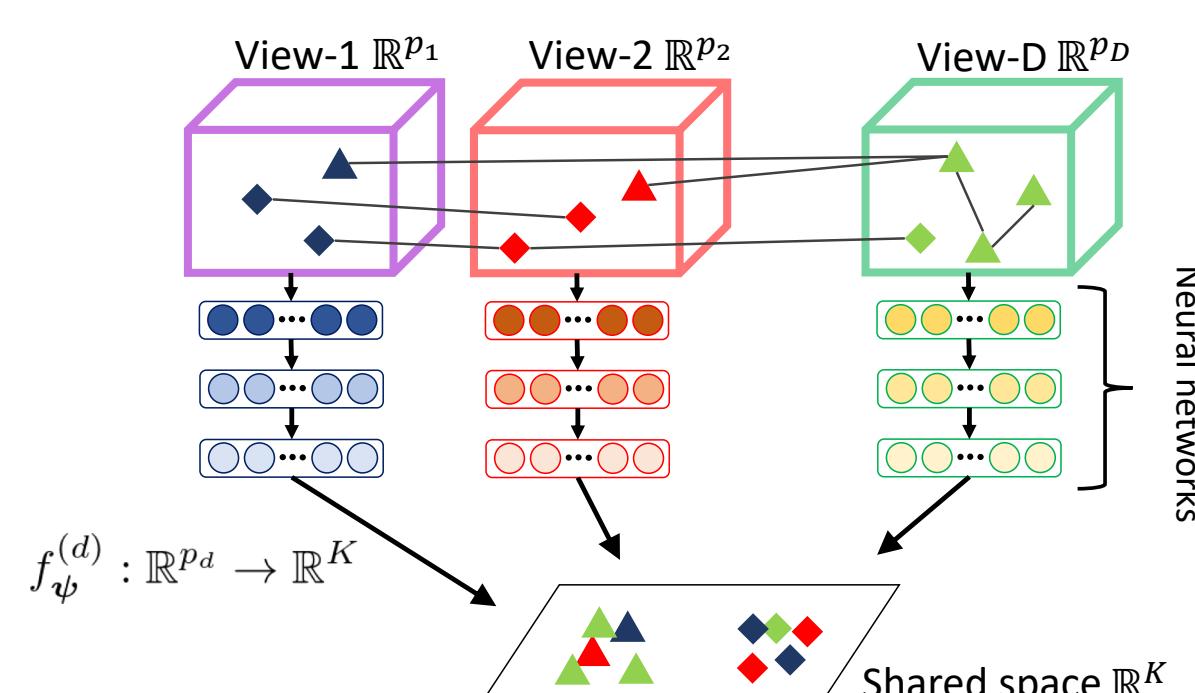
下平英寿(チームリーダー), Thong Pham(特別研究員), 福井一輝(パートタイマー), 奥野彰文(パートタイマー), 井上雅章(パートタイマー), KIM GEEWOOK(パートタイマー), 田中卓磨(パートタイマー), 岩田具治(客員研究員), 寺田吉亮(客員研究員), 伊森晋平(客員研究員), 廣瀬慧(客員研究員), 白石友一(客員研究員), 劉言(客員研究員)

### 統計学と機械学習

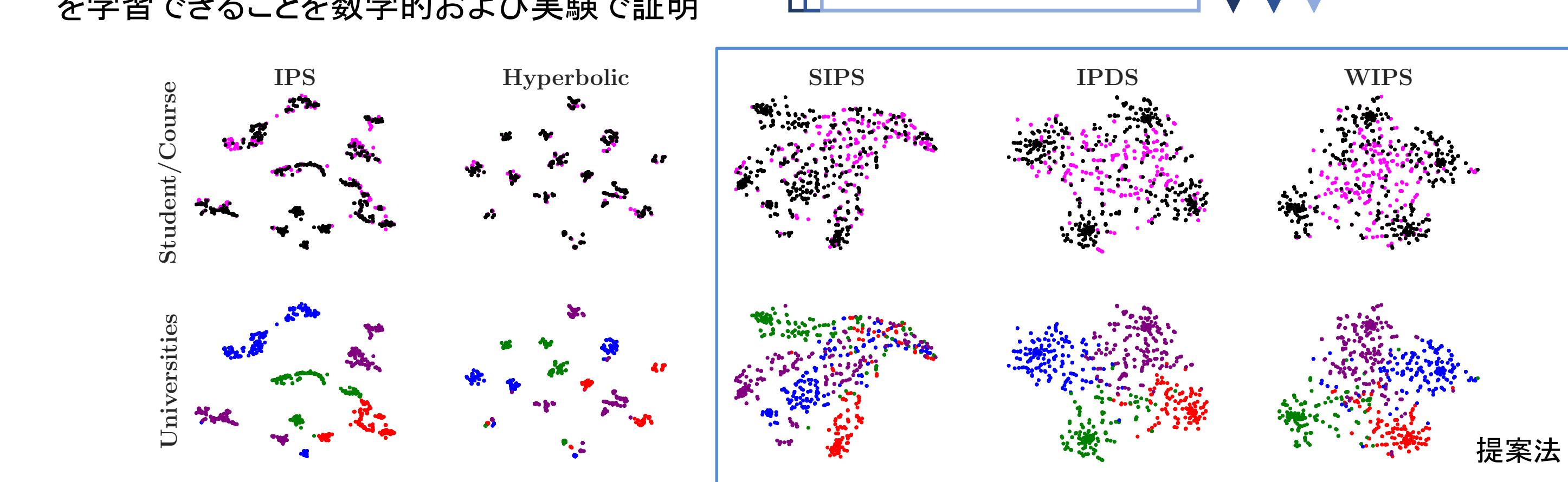
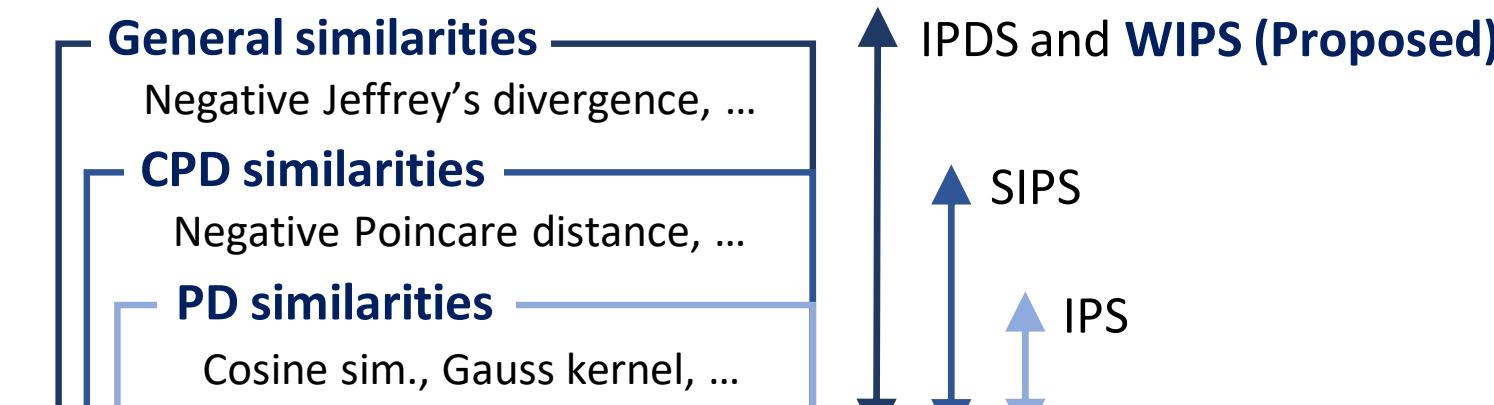
- 統計的仮説検定, 領域の検定, 選択的推測(selective inference)の理論
- マルチスケール・ブートストラップ法などリサンプリング法
- 汎化誤差を共変量シフト, 不完全観測などの設定で推定する情報量規準
- マルチドメインデータの多変量解析, ニューラルネットによるグラフ埋め込み
- 単語埋め込みなど自然言語処理, 画像と単語の埋め込みと相互検索
- 成長ネットワークの統計的推測
- 系統樹, 遺伝子発現データ, ゲノムデータの解析

## 高い表現力をもつ類似度関数の学習理論

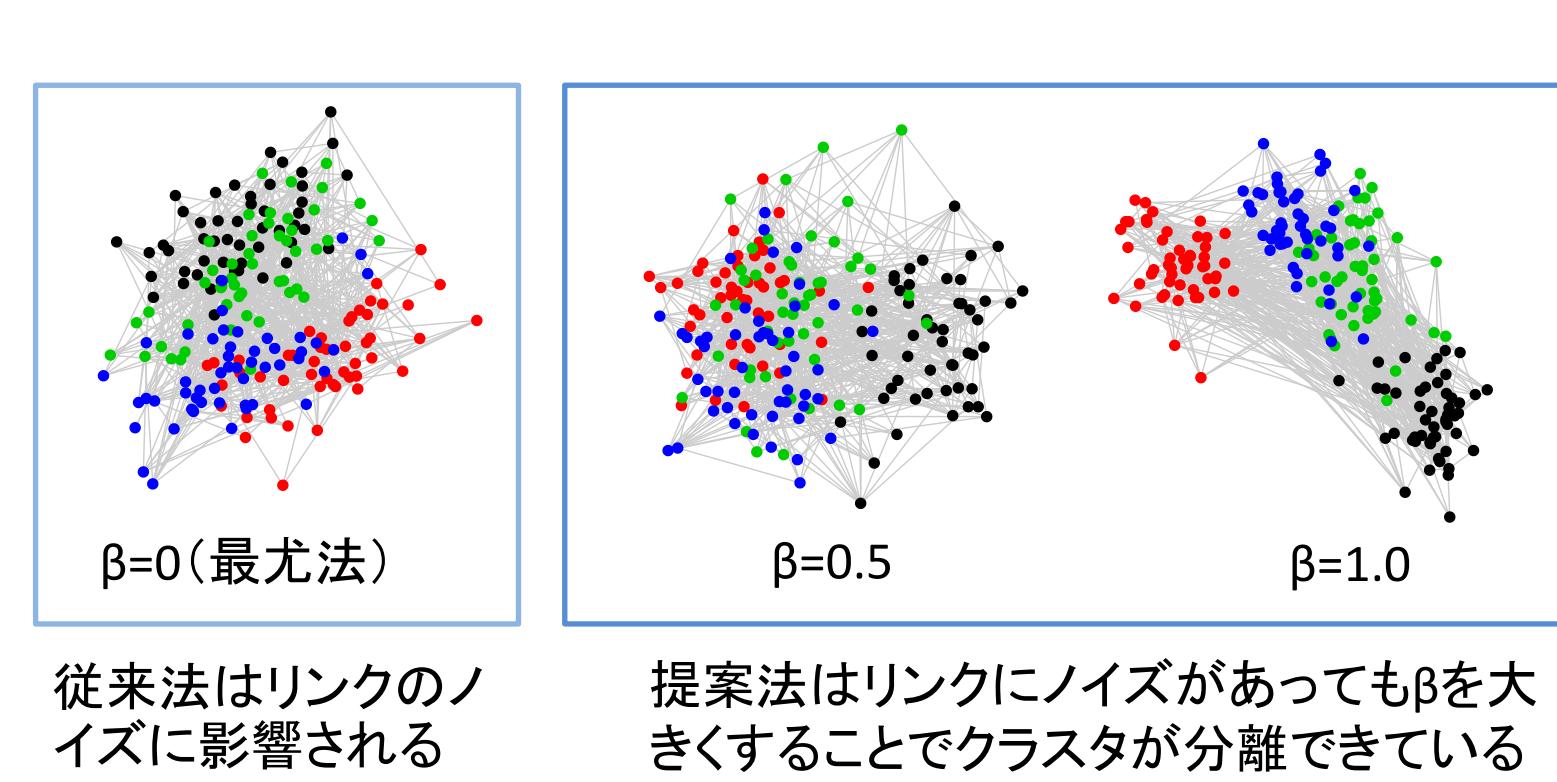
- Okuno, Hada, Shimodaira (**ICML 2018**) A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks
- 内積だけでも十分に大きなニューラルネットワークを併用することで広いクラスの類似度関数を学習できることを数学的に証明(Universal Approximation定理とMercer定理, 推定量の漸近的一致性の証明)
- グラフ埋め込みでの学習アルゴリズム(mini-batch SGD)
- 画像, 単語などマルチドメインの多変量解析への拡張
- 自然言語処理(単語埋め込み), 画像検索などの応用
- Okuno, Shimodaira (**ICML workshop Theoretical Foundations and Applications of Deep Generative Models 2018**) On representation power of neural network-based graph embedding and beyond
- Okuno, Kim, Shimodaira (to appear **AISTATS 2019**) Graph Embedding with Shifted Inner Product Similarity and Its Improved Approximation Capability
- 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(例)コサイン類似度
- 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, (一部の)Wasserstein距離など, 機械学習でよく利用される距離がたいてい表現できる
- 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



- Kim, Okuno, Fukui, Shimodaira (**arXiv:1902.10409**) Representation Learning with Weighted Inner Product for Universal Approximation of General Similarities
- 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および実験で証明

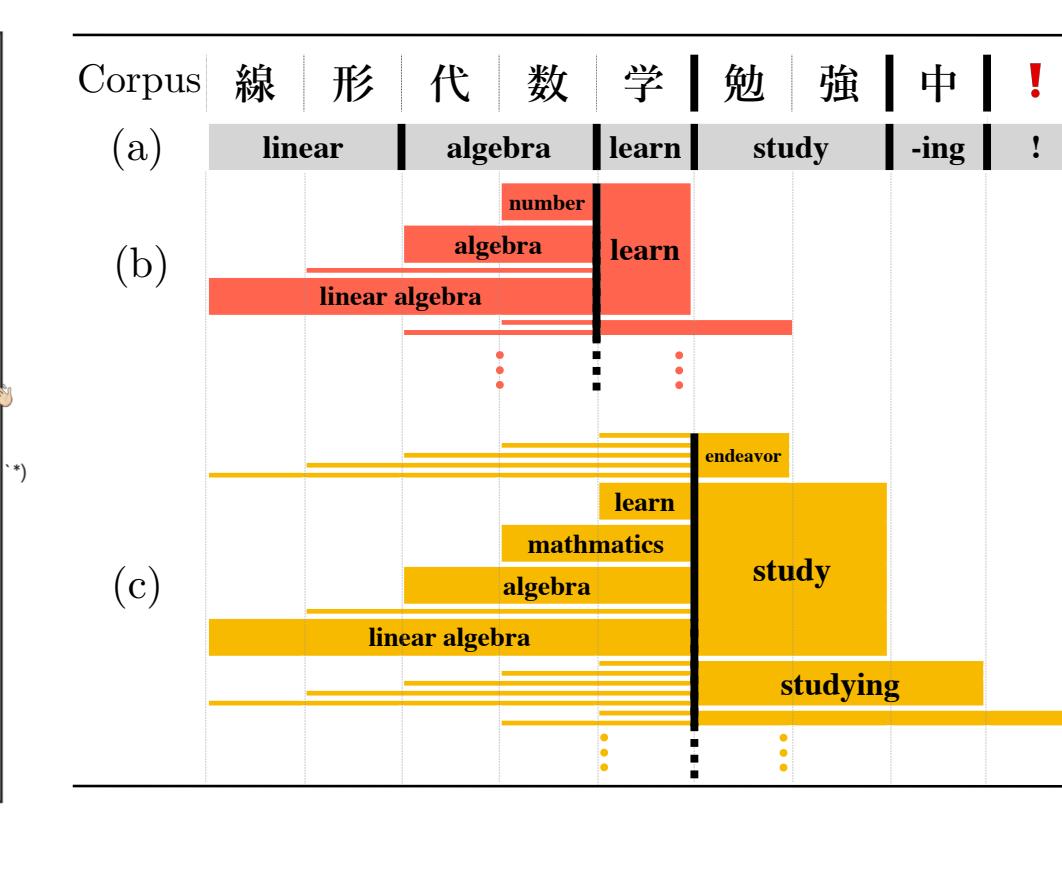
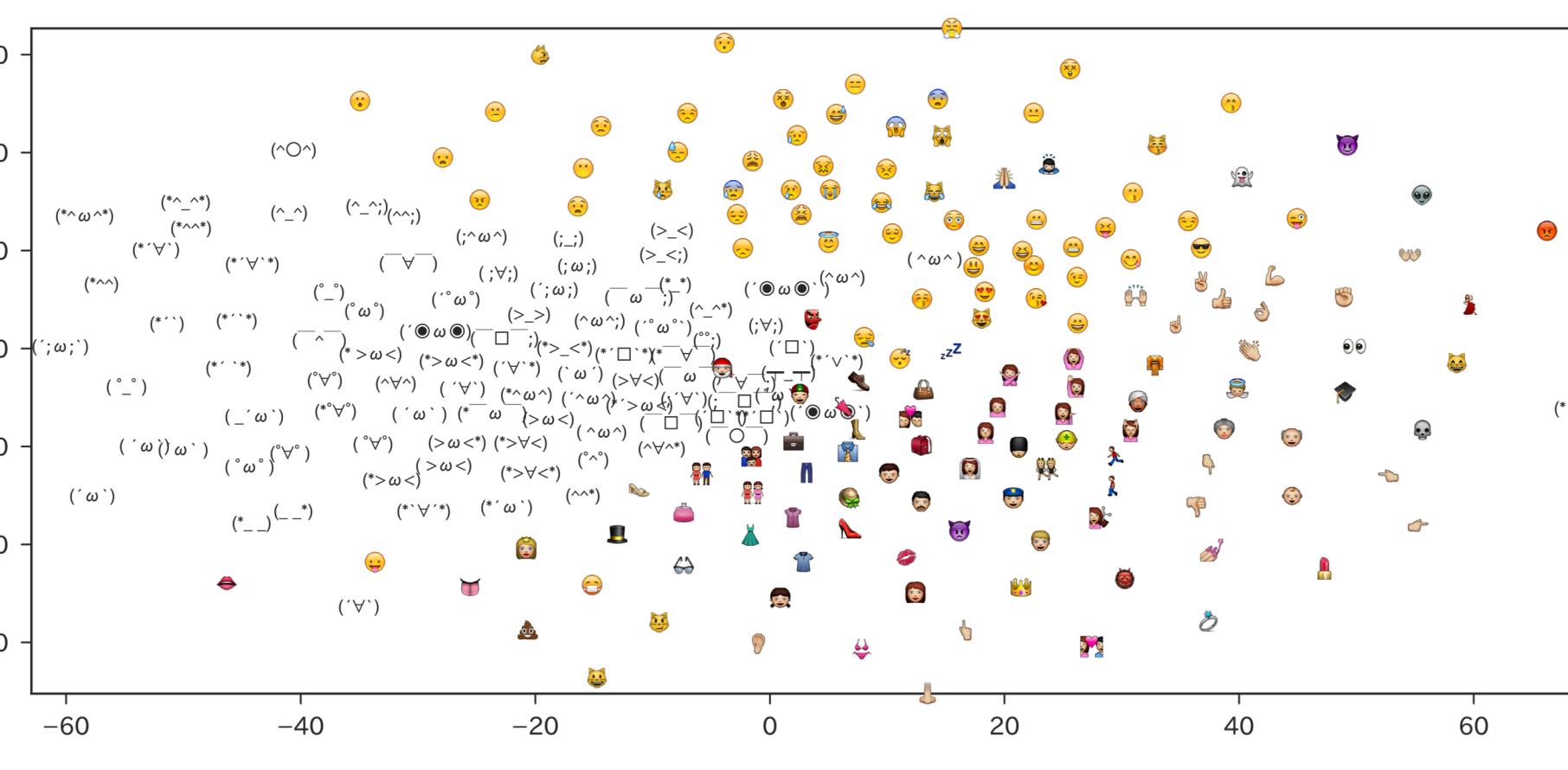
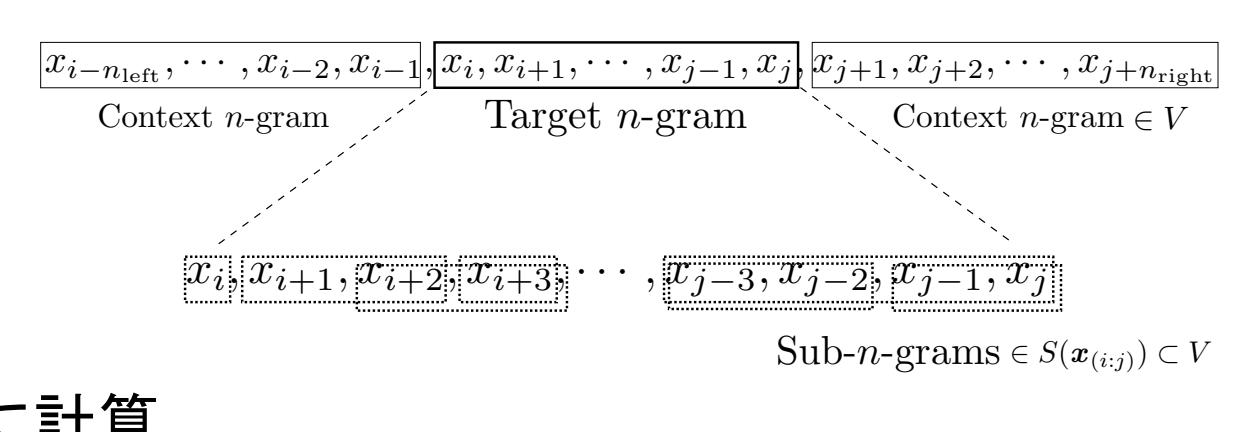


- Okuno, Shimodaira (to appear **AISTATS 2019**) Robust Graph Embedding with Noisy Link Weights
- 類似度関数学習やグラフ埋め込みがノイズに強くなるようにロバスト化する手法の提案
- 提案手法( $\beta$ -グラフ埋め込み)は新規に考案した損失関数(経験積率 $\beta$ -スコア)の最小化を実行
- ロバスト統計学におけるdensity power divergenceの理論を発展させたもの



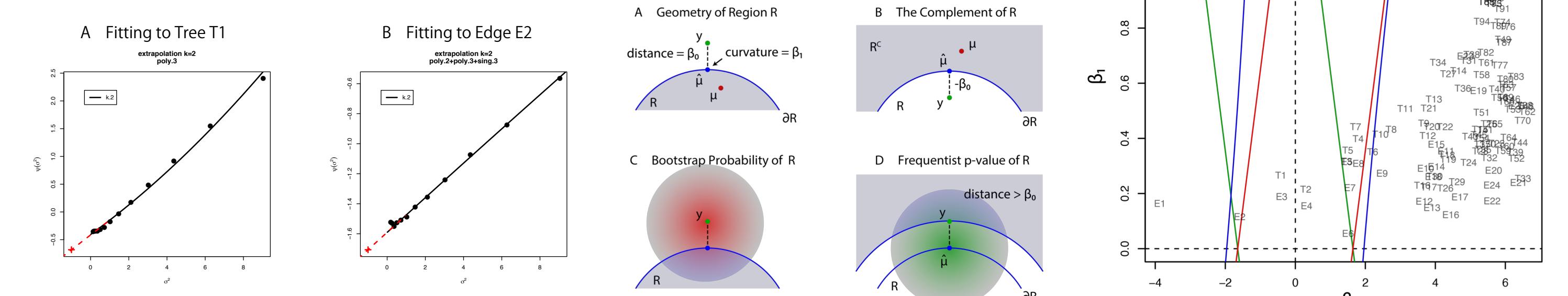
## 単語分割を必要としない単語埋め込み

- Kim, Fukui, Shimodaira (**EMNLP Workshop on Noisy User-generated Text W-NUT 2018**), Word-like character  $n$ -gram embedding
- Kim, Fukui, Shimodaira (to appear **NAACL-HLT 2019**), Segmentation-free compositional  $n$ -gram embedding
- 既存の単語埋め込み法では, 文書データを前処理によって分割して, 単語の列を生成する必要があった。英語などスペースで区切られた言語では問題ないが, 日本語, 中国語, 韓国語では形態素解析による前処理が必要であり, これが障害となることがあった
- とくにソーシャルメディアでは表記ゆれ, 新語の問題がある
- そこで「単語分割を全く行わない単語埋め込み法」の提案
- 単語, フレーズ, 文に対して連続的に適用できる
- コーパスのすべての部分文字列に対して埋め込みを効率的に計算



## 選択的推測の理論とその系統樹推定への応用

- Shimodaira, Terada (**arXiv:1902.04964**), Selective Inference for Testing Trees and Edges in Phylogenetics
- 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してきた。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)の理論が構築されつつある
- 弊チームの先行研究(Terada, Shimodaira **arXiv:1711.00949**)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
- 本研究ではその手法をもとに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



A: Projection of data point X to tree models  
B: Regions for trees with data point X

Estimated  $\beta_0$  and  $\beta_1$  for Mammal Dataset

C: Fitting to Tree T1  
D: Fitting to Edge E2

E: Geometry of Region R  
F: The Complement of R

G: Bootstrap Probability of R  
H: Frequentist p-value of R

I: Estimated  $\beta_0$  and  $\beta_1$  for Mammal Dataset



Trees T1, T2, T3, T4, T5, T6, T7, T8, T9, T10, T11, T12, T13, T14, T15, T16, T17, T18, T19, T20, T21, T22, T23, T24, T25, T26, T27, T28, T29, T30, T31, T32, T33, T34, T35, T36, T37, T38, T39, T40, T41, T42, T43, T44, T45, T46, T47, T48, T49, T50, T51, T52, T53, T54, T55, T56, T57, T58, T59, T60, T61, T62, T63, T64, T65, T66, T67, T68, T69, T70, T71, T72, T73, T74, T75, T76, T77, T78, T79, T80, T81, T82, T83, T84, T85, T86, T87, T88, T89, T90, T91, T92, T93, T94, T95, T96, T97, T98, T99, T100, T101, T102, T103, T104, T105, T106, T107, T108, T109, T110, T111, T112, T113, T114, T115, T116, T117, T118, T119, T120, T121, T122, T123, T124, T125, T126, T127, T128, T129, T130, T131, T132, T133, T134, T135, T136, T137, T138, T139, T140, T141, T142, T143, T144, T145, T146, T147, T148, T149, T150, T151, T152, T153, T154, T155, T156, T157, T158, T159, T160, T161, T162, T163, T164, T165, T166, T167, T168, T169, T170, T171, T172, T173, T174, T175, T176, T177, T178, T179, T180, T181, T182, T183, T184, T185, T186, T187, T188, T189, T190, T191, T192, T193, T194, T195, T196, T197, T198, T199, T200, T201, T202, T203, T204, T205, T206, T207, T208, T209, T210, T211, T212, T213, T214, T215, T216, T217, T218, T219, T220, T221, T222, T223, T224, T225, T226, T227, T228, T229, T220, T221, T222, T223, T224, T225, T226, T227, T228, T229, T230, T231, T232, T233, T234, T235, T236, T237, T238, T239, T230, T231, T232, T233, T234, T235, T236, T237, T238, T239, T240, T241, T242, T243, T244, T245, T246, T247, T248, T249, T240, T241, T242, T243, T244, T245, T246, T247, T248, T249, T250, T251, T252, T253, T254, T255, T256, T257, T258, T259, T250, T251, T252, T253, T254, T255, T256, T257, T258, T259, T260, T261, T262, T263, T264, T265, T266, T267, T268, T269, T260, T261, T262, T263, T264, T265, T266, T267, T268, T269, T270, T271, T272, T273, T274, T275, T276, T277, T278, T279, T270, T271, T272, T273, T274, T275, T276, T277, T278, T279, T280, T281, T282, T283, T284, T285, T286, T287, T288, T289, T280, T281, T282, T283, T284, T285, T286, T287, T288, T289, T290, T291, T292, T293, T294, T295, T296, T297, T298, T299, T290, T291, T292, T293, T294, T295, T296, T297, T298, T299, T300, T301, T302, T303, T304, T305, T306, T307, T308, T309, T300, T301, T302, T303, T304, T305, T306, T307, T308, T309, T310, T311, T312, T313, T314, T315, T316, T317, T318, T319, T310, T311, T312, T313, T314, T315, T316, T317, T318, T319, T320, T321, T322, T323, T324, T325, T326, T327, T328, T329, T320, T321, T322, T323, T324, T325, T326, T327, T328, T329, T330, T331, T332, T333, T334, T335, T336, T337, T338, T339, T330, T331, T332, T333, T334, T335, T336, T337, T338, T339, T340, T341, T342, T343, T344, T345, T346, T347, T348, T349, T340, T341, T342, T343, T344, T345, T346, T347, T348, T349, T350, T351, T352, T353, T354, T355, T356, T357, T358, T359, T350, T351, T352, T353, T354, T355, T356, T357, T358, T359, T360, T361, T362, T363, T364, T365, T366, T367, T368, T369, T360, T361, T362, T363, T364, T365, T366, T367, T368, T369, T370, T371, T372, T373, T374, T375, T376, T377, T378, T379, T370, T371, T372, T373, T374, T375, T376, T377, T378, T379, T380, T381, T382, T383, T384, T385, T386, T387, T388, T389, T380, T381, T382, T383, T384, T385, T386, T387, T388, T389, T390, T391, T392, T393, T394, T395, T396, T397, T398, T399, T390, T391, T392, T393, T394, T395, T396, T397, T398, T399, T400, T401, T402, T403, T404, T405, T406, T407, T408, T409, T400, T401, T402, T403, T404, T405, T406, T407, T408, T409, T410, T411, T412, T413, T414, T415, T416, T417, T418, T419, T410, T411, T412, T413, T414, T415, T416, T417, T418, T419, T420, T421, T422, T423, T424, T425, T426, T427, T428, T429, T420, T421, T422, T423, T424, T425, T426, T427, T428, T429, T430, T431, T432, T433, T434, T435, T436, T437, T438, T439, T430, T431, T432, T433, T434, T435, T436, T437, T438, T439, T440, T441, T442, T443, T444, T445, T446, T447, T448, T449, T440, T441, T442, T443, T444, T445, T446, T447, T448, T449, T450, T451, T452, T453, T454, T455, T456, T457, T458, T459, T450, T451, T452, T453, T454, T455, T456, T457, T458, T459, T460, T461, T462, T463, T464, T465, T466, T467, T468, T469, T460, T461, T462, T463, T464, T465, T466, T467, T468, T469, T470, T471, T472, T473, T474, T475, T476, T477, T478, T479, T470, T471, T472, T473, T474, T475, T476, T477, T478, T479, T480, T481, T482, T483, T484, T485, T486, T487, T488, T489, T480, T481, T482, T483, T484, T485, T486, T487, T488, T489, T490, T491, T492, T493, T494, T495, T496, T497, T498, T499, T490, T491, T492, T493, T494, T495, T496, T497, T498, T499, T500, T501, T502, T503, T504, T505, T506, T507