方言群の分析のための
分岐と伝播の統合的モデル化

京都大学
村脇 有吾

シンポジウム「日琉諸方言系統論の展望」 @ online　2020年12月20日

自己紹介: 村脇 有吾

- 京都大学 大学院情報学研究科
  知能情報学専攻 講師
- 表の仕事は自然言語処理

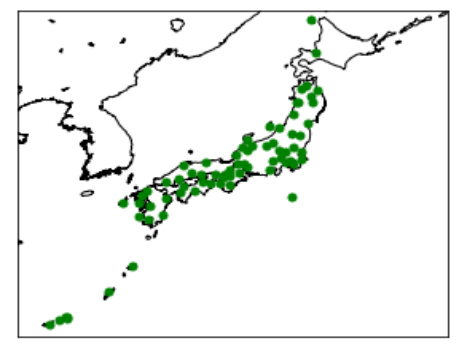
Good luck with it, but I think the location is going to be a hard sell.

native-like expression

[Harust+, 2020]

- 今日のお話は裏テーマ
- 表と裏の共通性は言語の問題を
  統計・機械学習を使って解くこと

今日のお話のあらまし

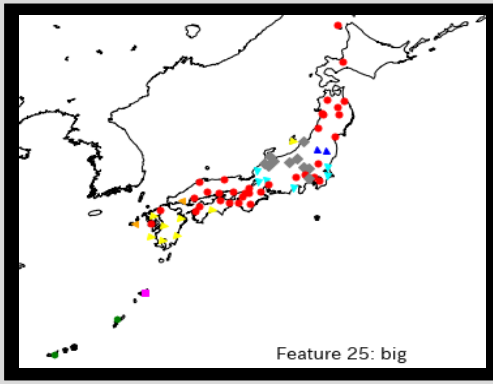


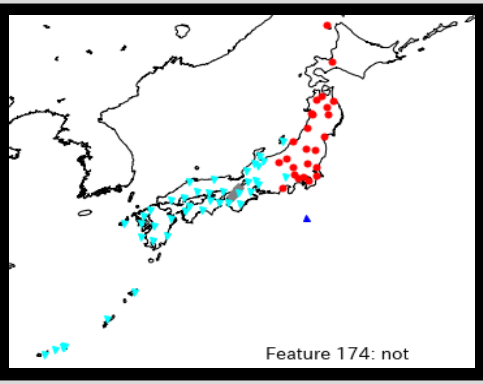
$L = 72$個の方言

方言語彙・座標データ提供: 五十嵐陽介氏@国立国語研

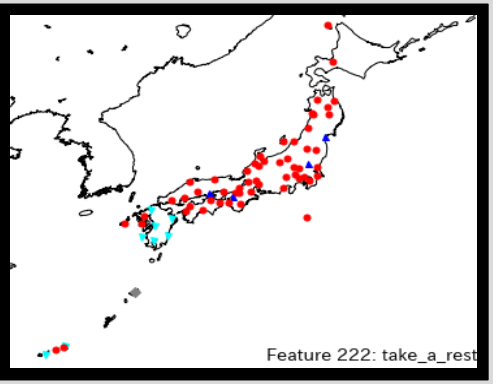
今日のお話のあらまし

観測
データ





…



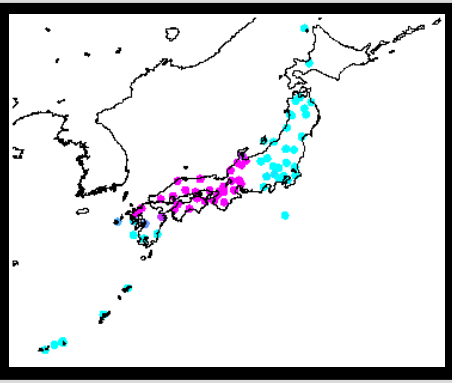
$L = 72$個の方言　　$N = 255$個の特徴 ($M = 1768$個の形質)

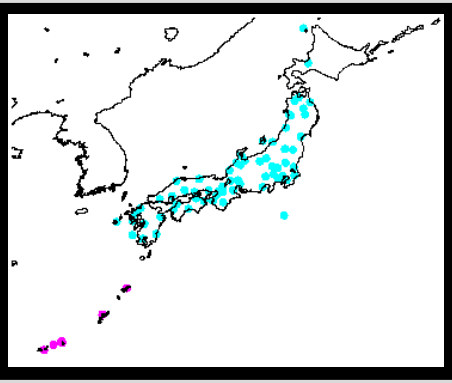
方言語彙・座標データ提供: 五十嵐陽介氏@国立国語研

今日のお話のあらまし

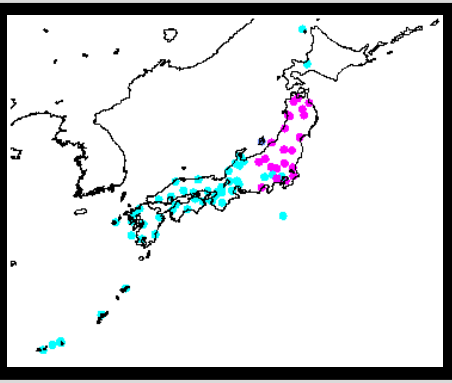
$K = 20$個の潜在的地理的分布

本発表で仮定する潜在構造

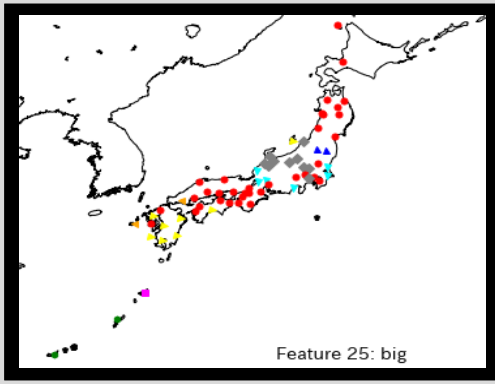


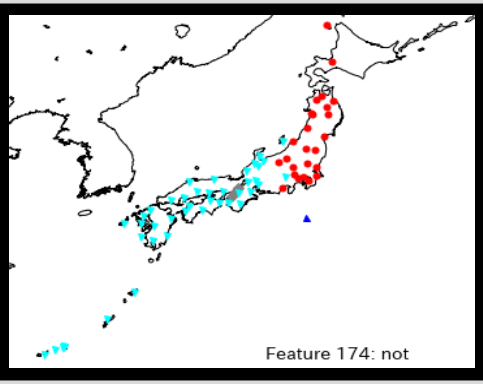


…

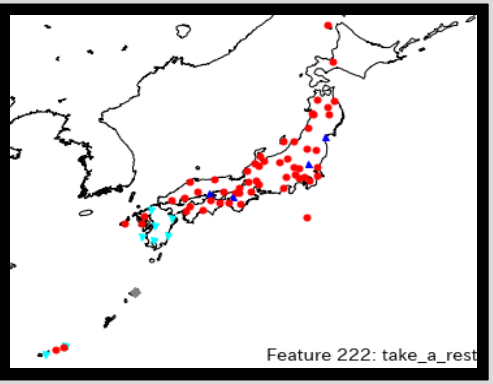


観測データ





…

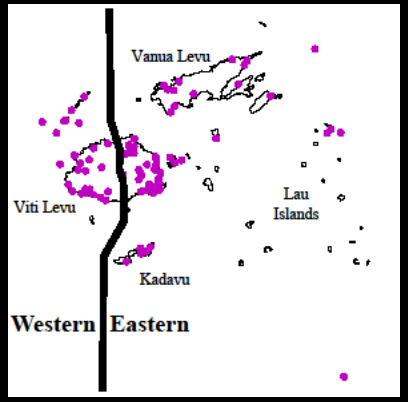


$L = 72$個の方言　　$N = 255$個の特徴 ($M = 1768$個の形質)

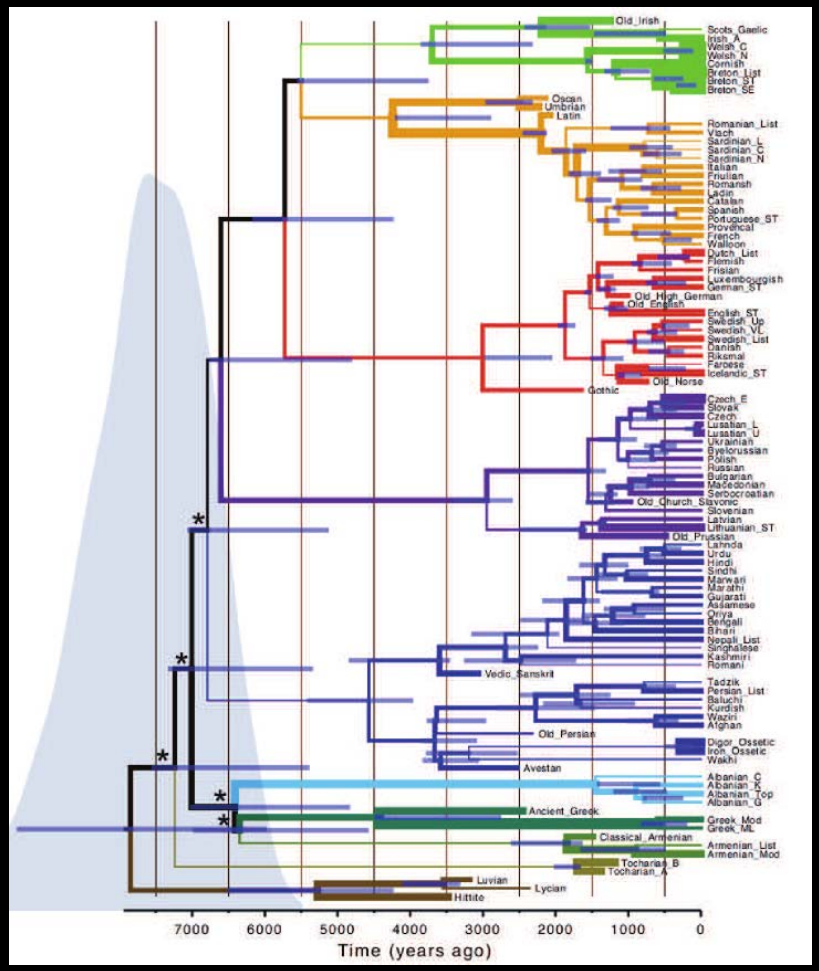
方言語彙・座標データ提供: 五十嵐陽介氏@国立国語研

おことわり: 日琉諸方言に関する新知見は (まだ) ありません

- 統計的モデルの開発と個別データセットへの適用の分業
  - 斎藤先生・木村先生のご発表は後者 (のはず)
  - 私の今回の発表は前者
- 日琉諸方言以外への提案モデルの適用を視野に
  - e.g. フィジー語諸方言 [Murawaki, 2020]



背景: ゲノムから言語への統計的モデルの転用



[Bouckaert+, Science 2012]

- Russell D. Grayらが1990年代末から
- 一部には集団遺伝学のモデルの転用事例も
- 言語独自のモデル化の試みはほとんどない

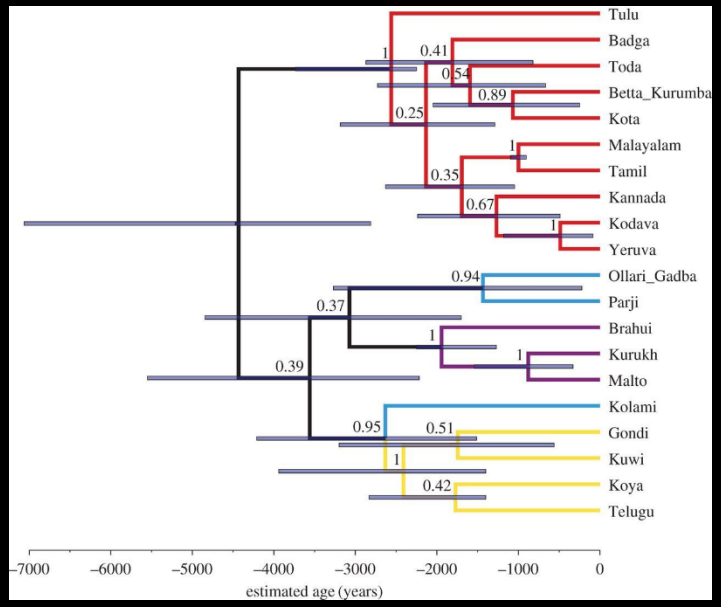
ゲノムと言語データは かなり性質が違う

	ゲノム	言語
集団	個体の集合	1個の個体 (言語)
個体数	>数十万	数十から数千
特徴数	数十万から数百万	数百
特徴列	並びに意味がある	便宜的に並べただけ
過去のデータ	直接観測可能に (古代DNA)	文献資料は限定的 かつ扱いが難しい
各点の情報量	ジャンク	豊か

言語の特性に応じた言語独自の統計的モデルの開発が有望

言語に対する2つの進化モデル

縦の継承 (木状)



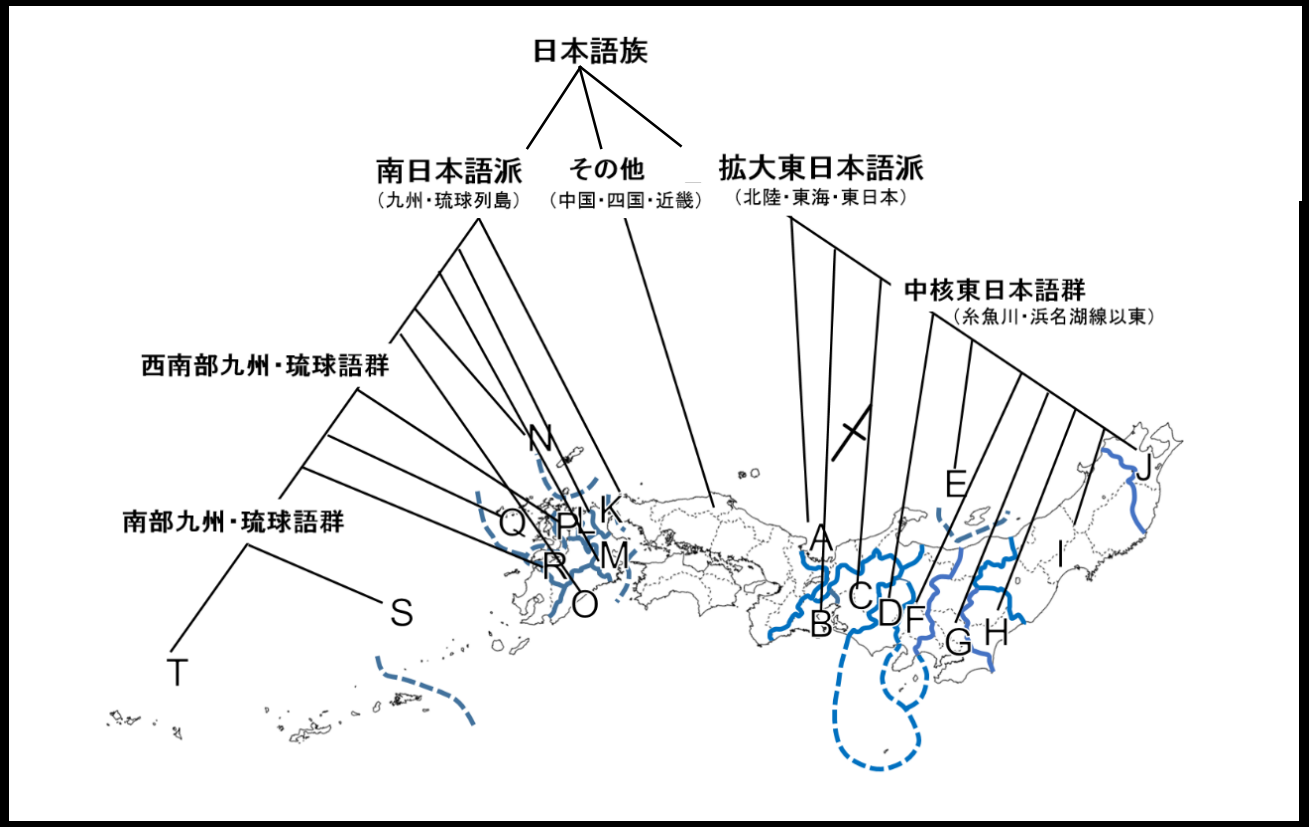
[Kolipakam+, 2018]

横の伝播 (波状)

[柳田, 1930]

接触の影響の強い方言群の進化は従来、
横の伝播によって説明されてきた

近年の研究は方言群からも
縦の継承の痕跡が見つかる可能性を示唆



[五十嵐, 2018]

人手による論証:
確実性の高い手がかりへの依存

- 2つの方言が形質を共有する理由:
  1. 偶然の一致
  2. 平行進化
  3. 共通祖先からの継承
  4. 伝播による借用
- 系統分類のための継承形質の細分類:
  1. 原始形質
  2. 派生形質

人手による論証:
確実性の高い手がかりへの依存

- 2つの方言が形質を共有する理由:
  1. 偶然の一致
  2. 平行進化
  3. 共通祖先からの継承
  4. 伝播による借用

記号の恣意性のため語彙的特徴が偶然一致する可能性はほぼ排除できる
意味の平行変化 (PERSON→MAN) に注意が必要
[Chang+, 2015]

- 系統分類のための継承形質の細分類:
  1. 原始形質
  2. 派生形質

人手による論証:
確実性の高い手がかりへの依存

- 2つの方言が形質を共有する理由:
  1. 偶然の一致
  2. 平行進化
  3. 共通祖先からの継承
  4. 伝播による借用

記号の恣意性のため語彙的特徴が偶然一致する可能性はほぼ排除できる
意味の平行変化 (PERSON→MAN) に注意が必要
[Chang+, 2015]

この2つの識別が実際には不確実！

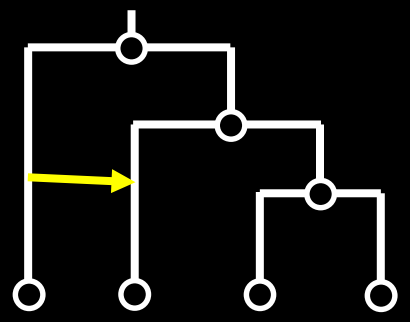
- 系統分類のための継承形質の細分類:
  1. 原始形質
  2. 派生形質

解決策: 計算集約的統計的モデル

- 不確実 → 解が絞り込めない → 組合せ爆発
- 人の手に余る数の候補は計算機に検討させればよい
- 確率統計学はある解候補の確からしさを評価するための道具

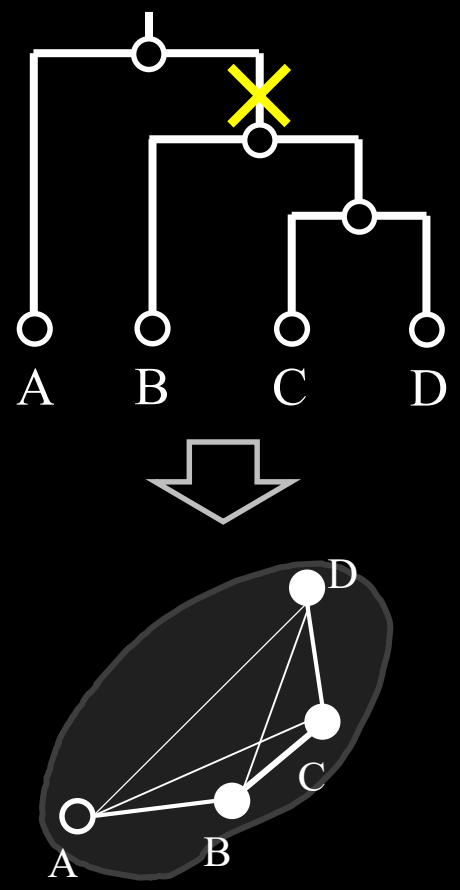
縦の継承と横の伝播の統合的モデル化?

- 縦のモデル (系統樹) を拡張して横の伝播を扱うのは非現実的
  - 自由度が高すぎて何でもありに
  - 横の伝播はノイズとして無視できると言い張ろうとする悪あがきも [Greenhill+ 2009]
- 系統樹の拡張以外の方法で、統一的モデル化できないか?



アイデア: 縦の継承と横の伝播は いずれも地理的分布を生み出す

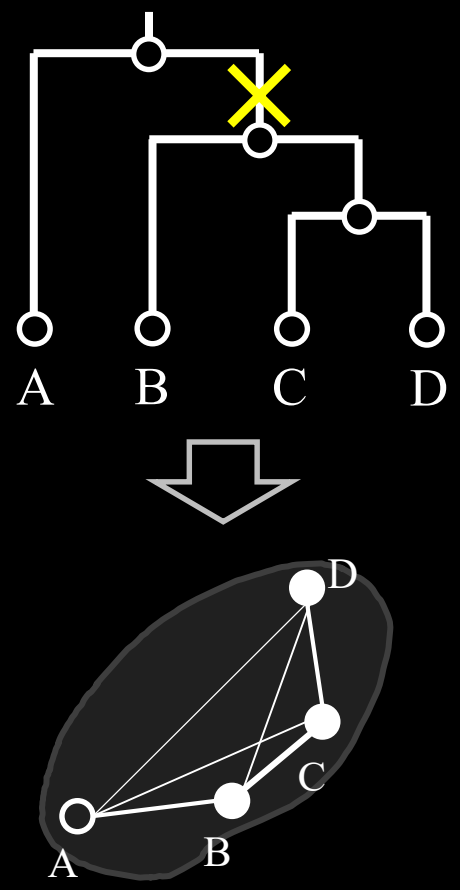
縦の継承 (木状)



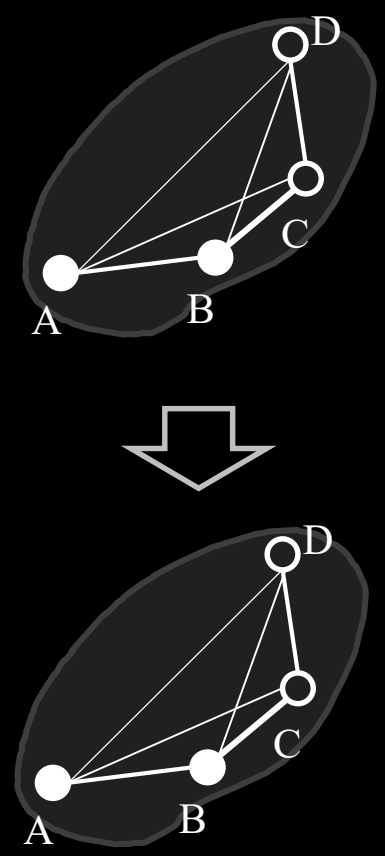
[Murawaki, 2020]

アイデア: 縦の継承と横の伝播は いずれも地理的分布を生み出す

縦の継承 (木状)

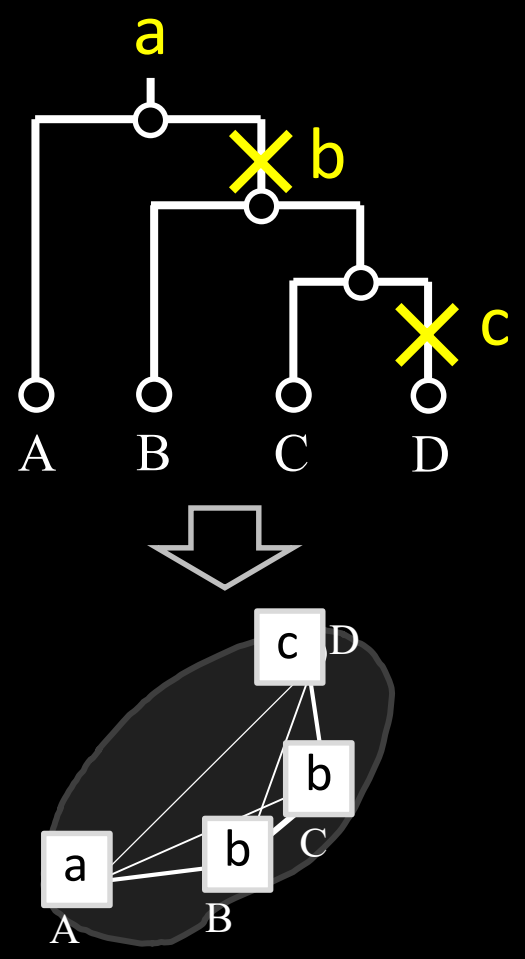


横の伝播 (波状)



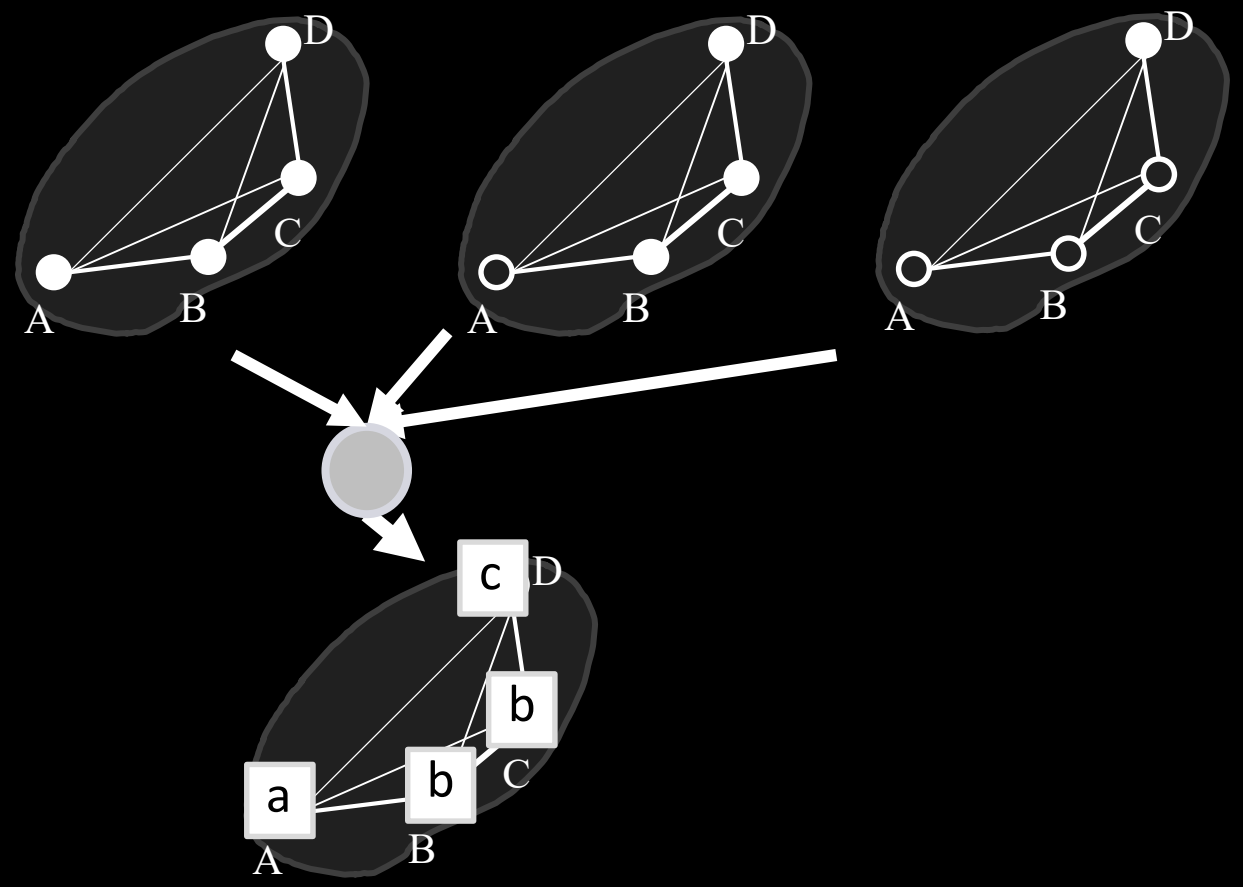
[Murawaki, 2020]

地理的分布は現在直接観測できるとは限らない
(潜在的地理的分布)



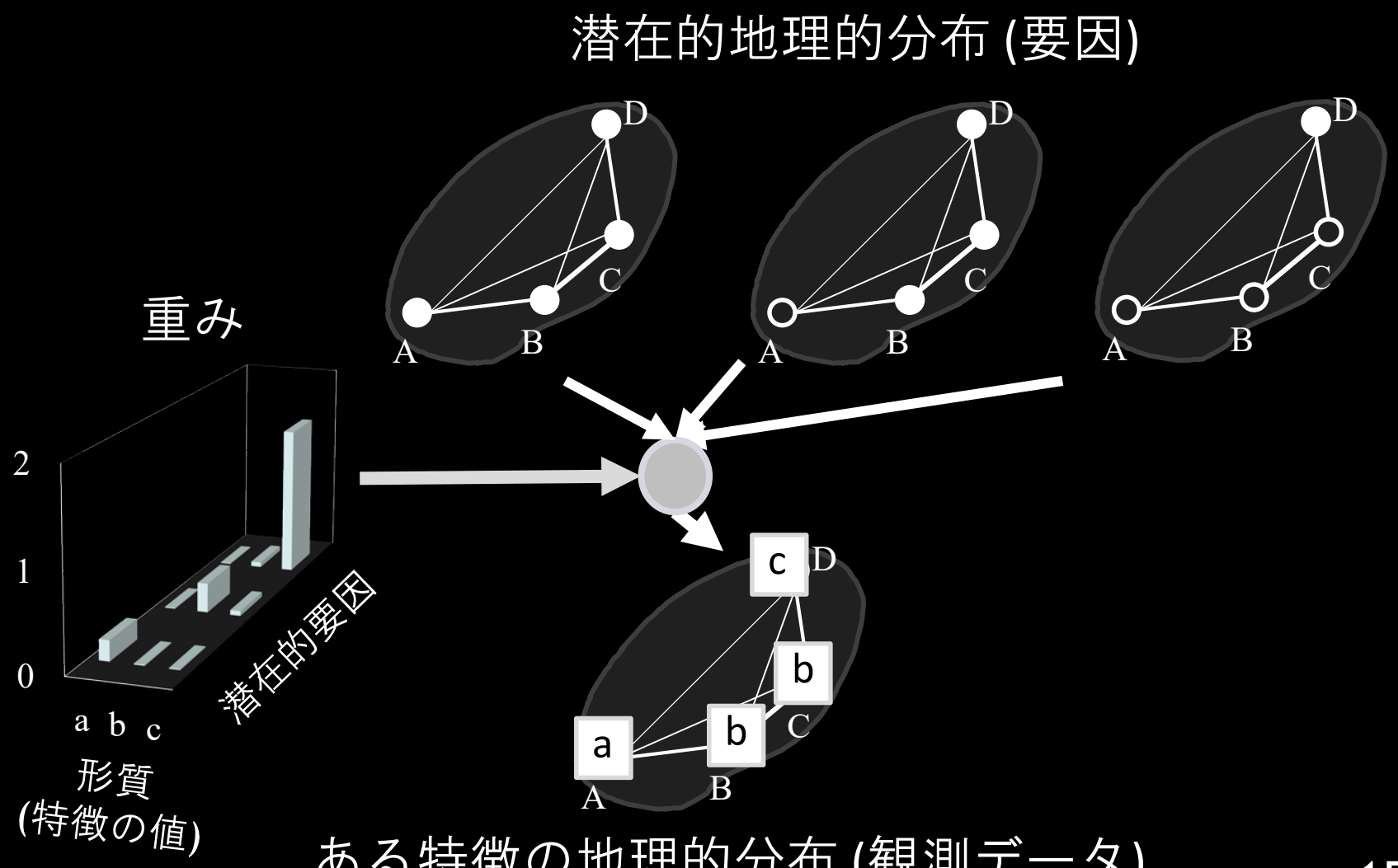
地理的分布は現在直接観測できるとは限らない (潜在的地理的分布)

潜在的地理的分布 (要因)



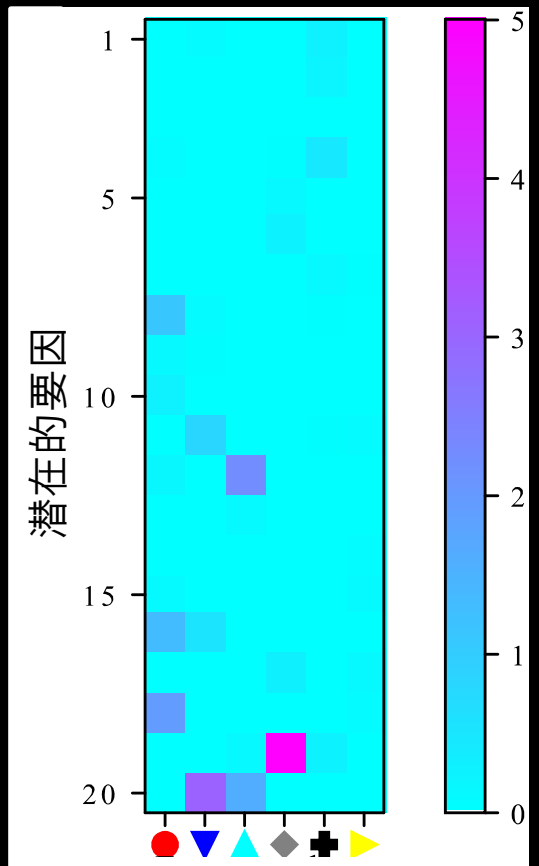
ある特徴の地理的分布 (観測データ)

潜在的地理的分布と観測される地理的分布は正の重みを使って確率的に結びつける

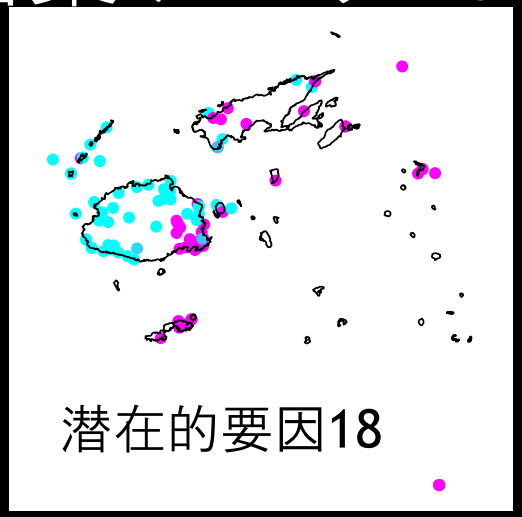


フィジー語諸方言の基礎語彙データでの実験

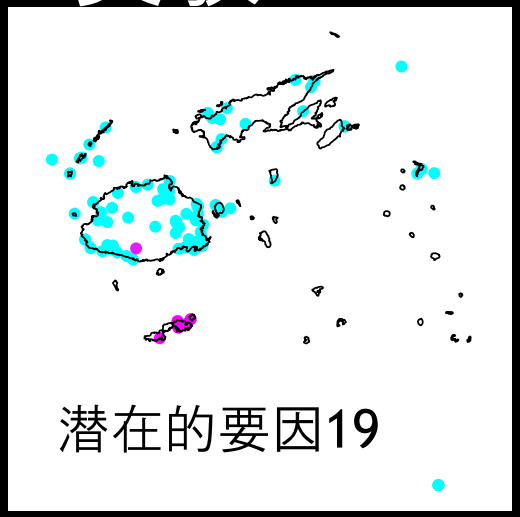
重み



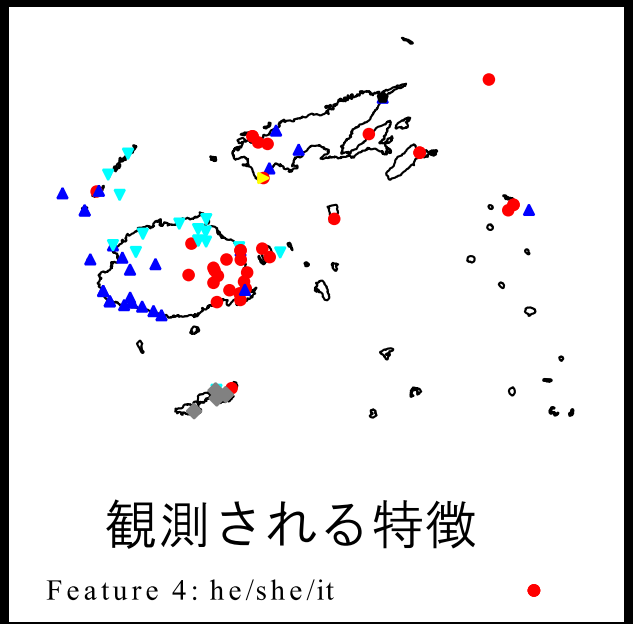
形質 (特徴の値)



潜在的要因18



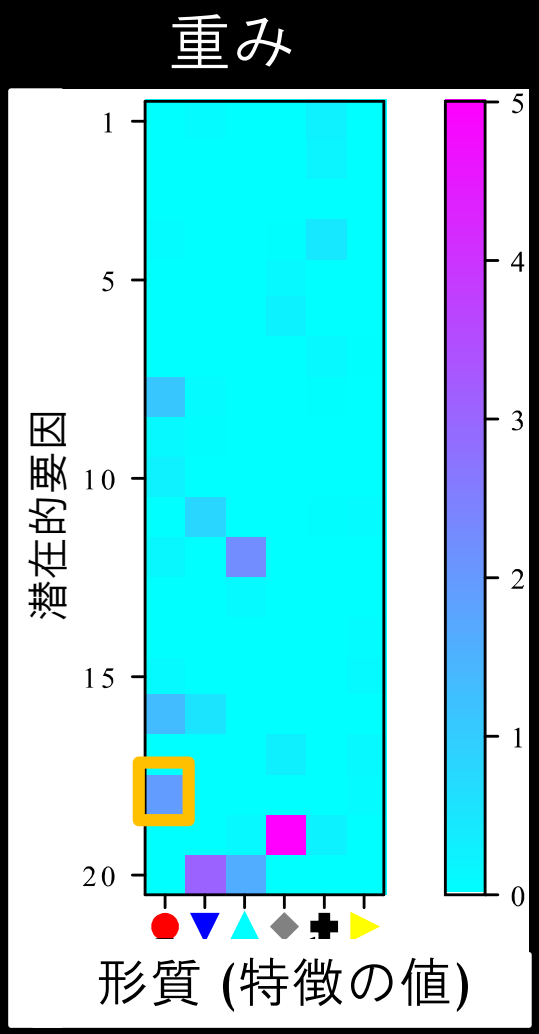
潜在的要因19

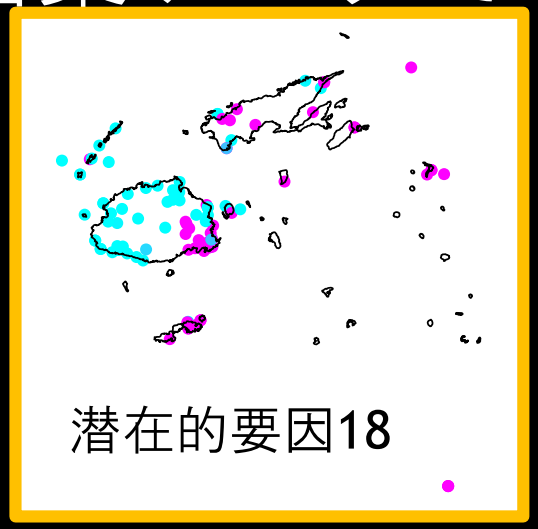


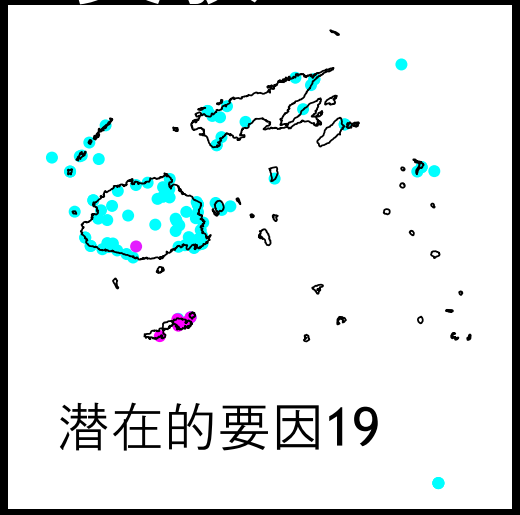
観測される特徴

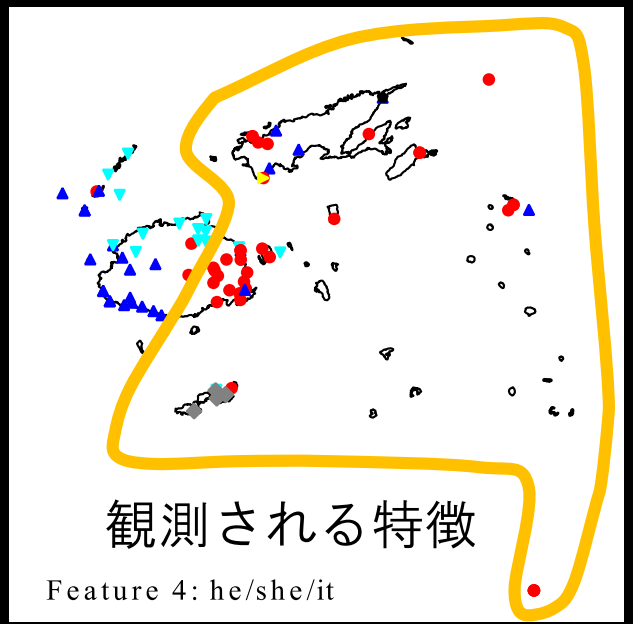
Feature 4: he/she/it

フィジー語諸方言の基礎語彙データでの実験

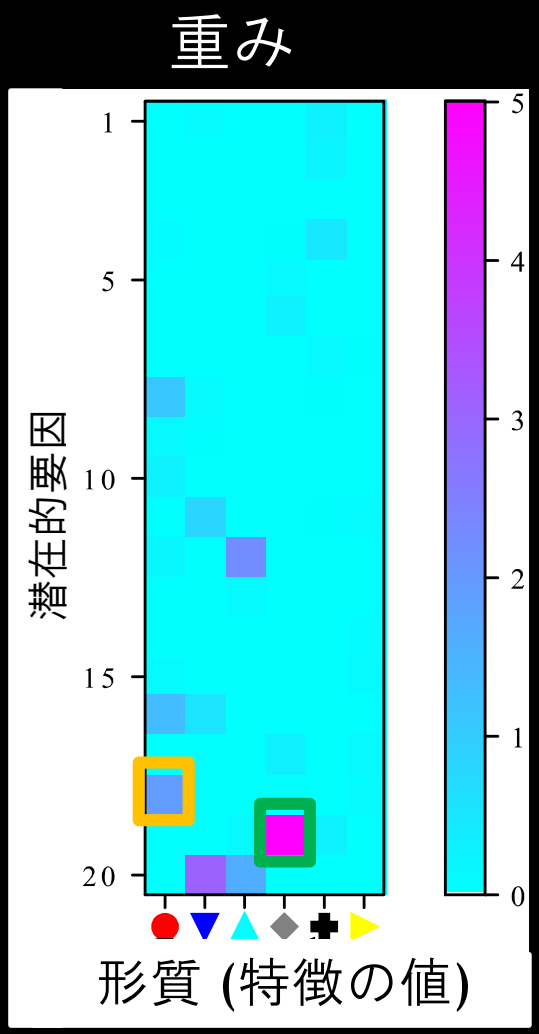


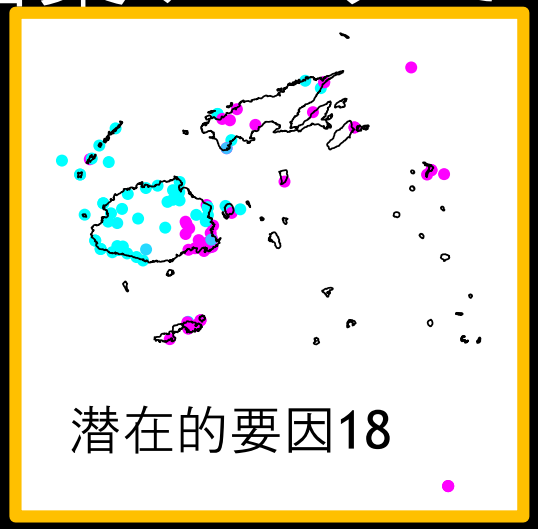


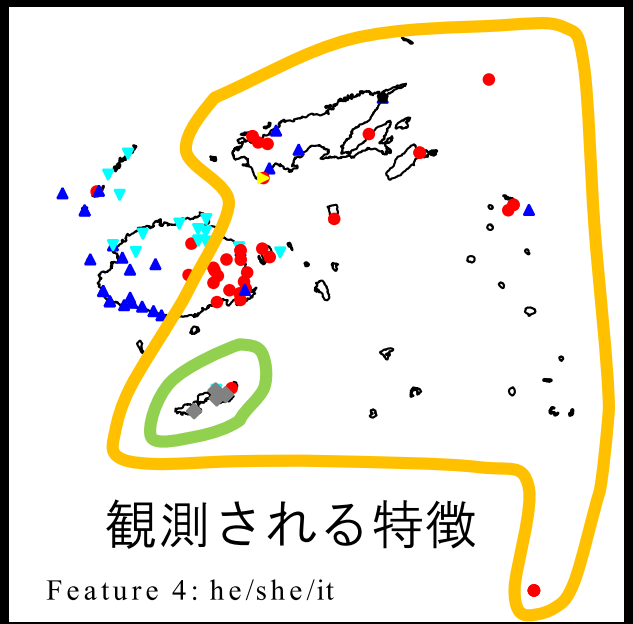




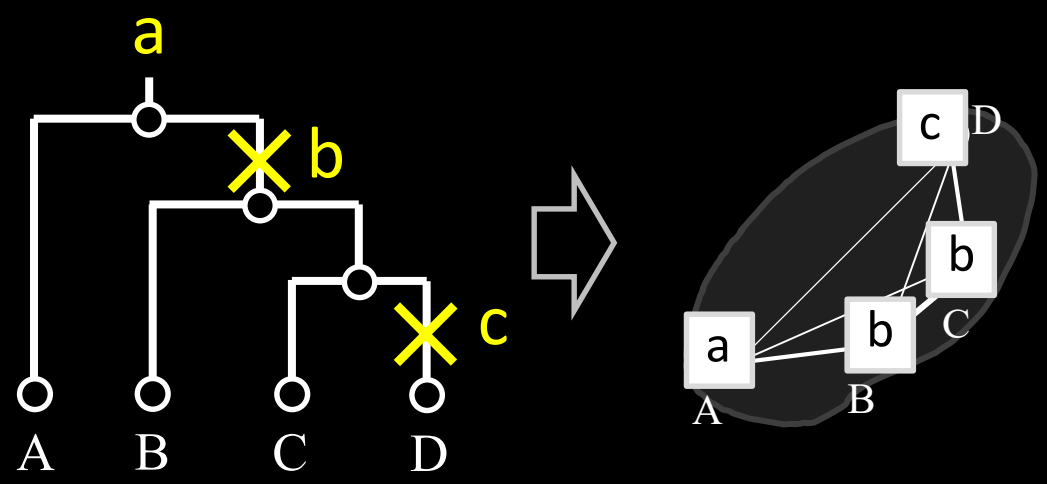
フィジー語諸方言の基礎語彙データでの実験

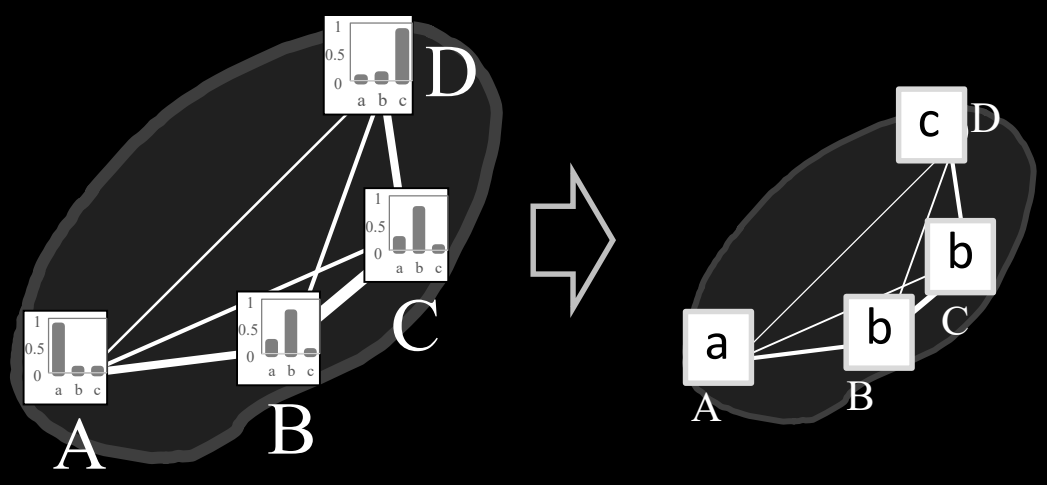






系統樹モデルと何が違うのか



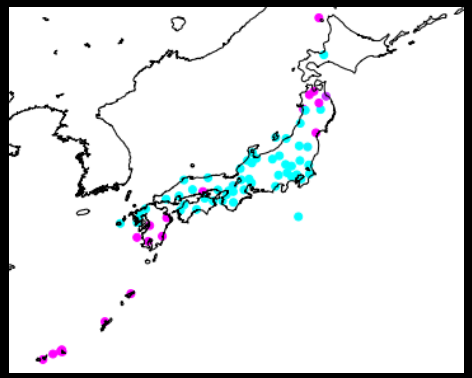


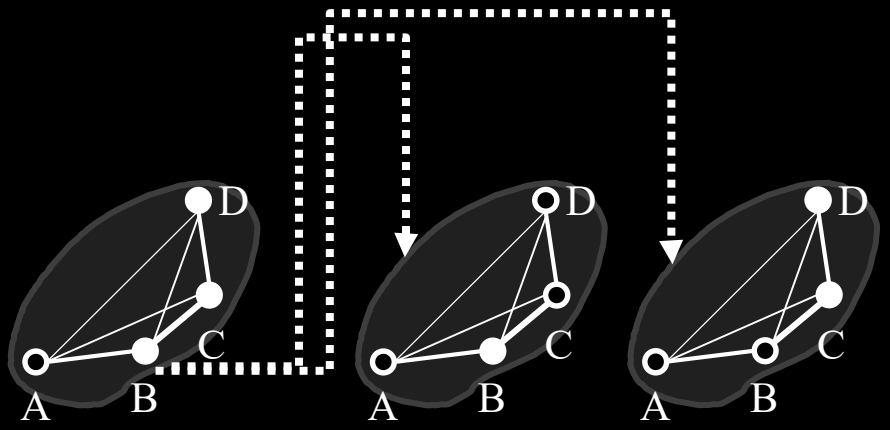
- 系統樹モデルは形質の生死を完全に追跡
- 横の伝播を取り込もうとすると破綻

- 提案手法では、各方言は潜在的にすべての形質を持っているが、確率の高い形質が結果として観測される
- 例外的現象の説明を諦めてマクロなパターンの説明に注力

木構造制約の必要性

- 現在のモデルは地理的一体性を欠いた潜在的地理的分布を推定し得る
- 縦の継承に由来する潜在的地理的分布は木構造制約を満たすべき
  - 親は2つの子を持つ
  - 2つの子が被覆する方言群は親が被覆する方言群の重複のない分割
- モデルへの木構造制約の組み込みは技術的に難しく近似が必要そう





結論と今後の課題

- 縦の継承と横の伝播を統合的に扱うための手法として、潜在的地理的分布に基づく確率的モデルを提案
- 人工データを使った定量評価とフィジー語諸法元基礎語彙データを使った定性評価により有望な結果を得た
- 潜在的要因の一部が木の制約に従うようにモデルを拡張したい