

言語類型論の特徴からの 潜在表現の獲得と その歴史的変化の分析への応用

京都大学
村脇 有吾



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 助教
工学部電気電子工学科 兼任
- 専門：計算言語学と自然言語処理
 - －表の仕事は普通のテキスト処理
 - 単語分割、ゼロ照応解析、常識的知識の獲得ほか
 - －今日お話は裏の仕事
 - 言語の研究ですが、テキストは出てきません

今回扱うのは言語類型論のデータ

- 行列 X をひたすら解析
 - $L = 2,679, N = 104$

N 個の多値特徴

L 個の言語	1	1	...	2
	2	4	...	3
	1	1	...	3
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	3	1	...	1

X

今回扱うのは言語類型論のデータ

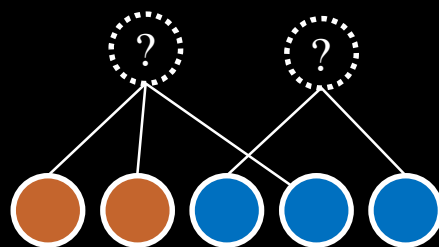
- 行列 X をひたすら解析
 - $L = 2,679$, $N = 104$
- さらに2種類の資源を利用

N 個の多値特徴

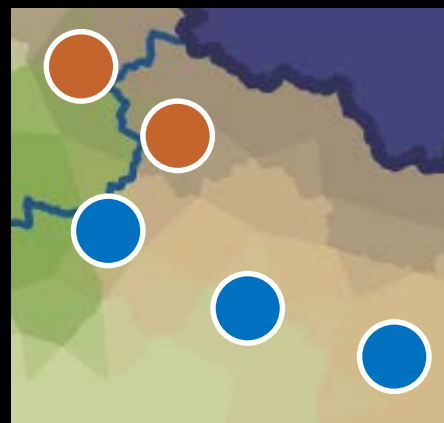
1	1	...	2
2	4	...	3
1	1	...	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$
3	1	...	1

L 個の言語

X



既知の系統群



言語の地理位置

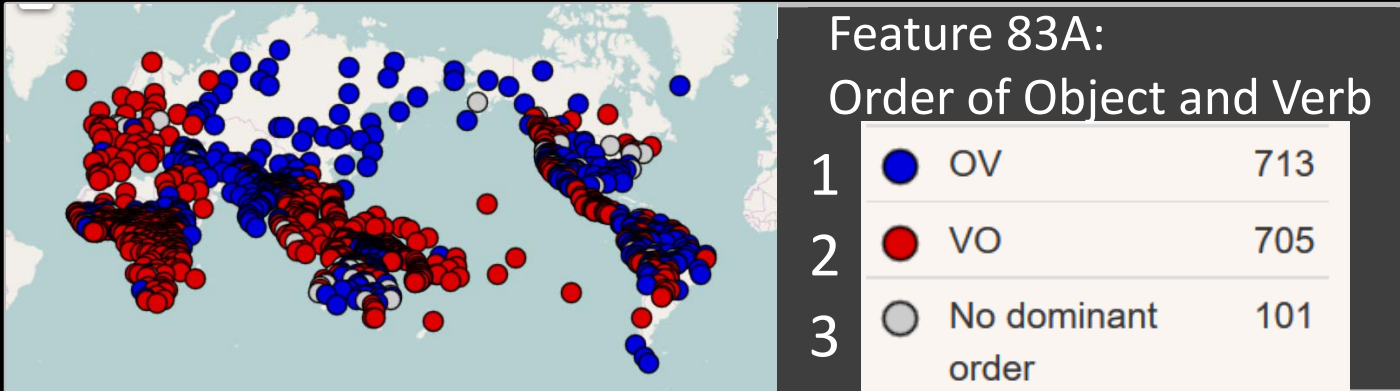
言語類型論の特徴

N 個の多値特徴

L 個の言語	1	1	...	2
	2	4	...	3
	1	1	...	3
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	3	1	...	1

X

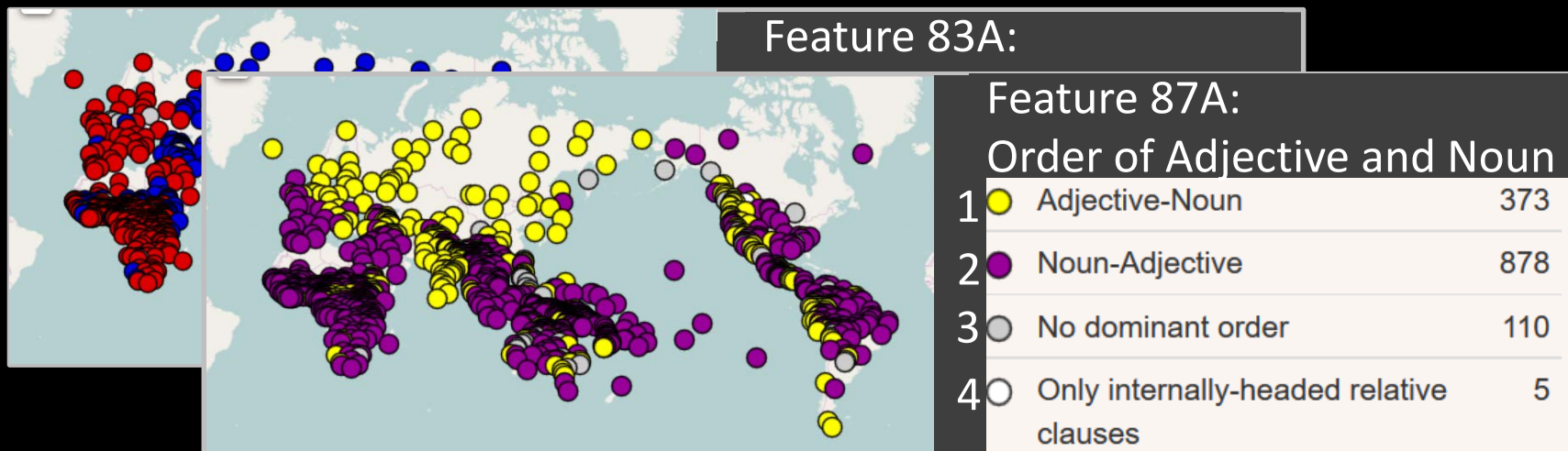
言語類型論の特徴



N 個の多値特徴

L 個の言語	1	1	...	2
	2	4	...	3
	1	1	...	3
	⋮	⋮	⋮	⋮
	3	1	...	1
	X			

言語類型論の特徴



N 個の多値特徴

L 個の言語	1	1	...	2
	2	4	...	3
	1	1	...	3
	⋮	⋮	⋮	⋮
	3	1	...	1
	X			

研究疑問

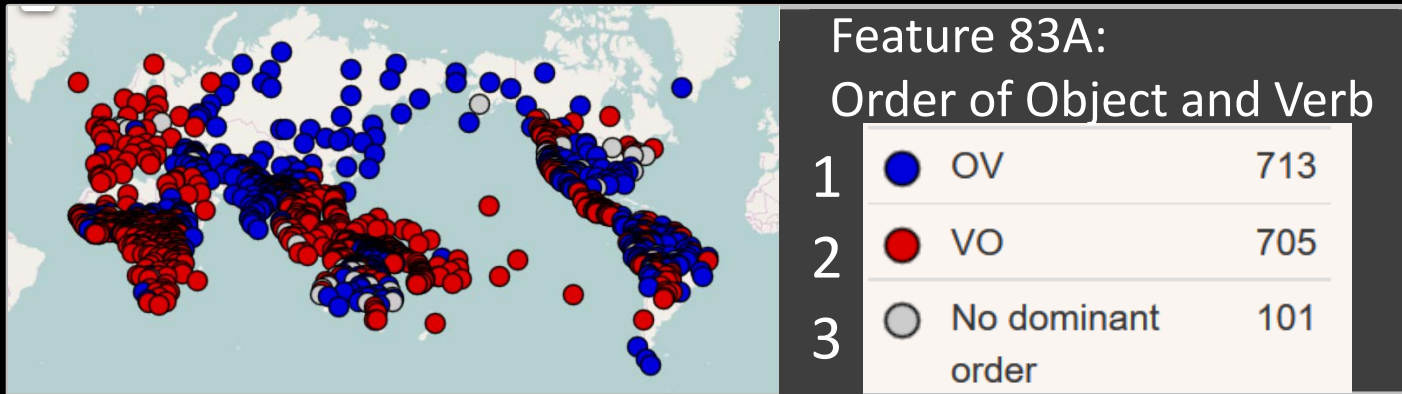
1. これらの特徴は時間とともにどのように変化してきたのか?
2. これらの特徴を用いて諸言語の深い歴史的関係を解明できるか?

N 個の多値特徴

L 個の言語	1	1	...	2
	2	4	...	3
	1	1	...	3
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	3	1	...	1

X

従来研究: 特徴の地理的分布から 過去の状態を推測



- 一部の特徴は歴史的に安定で深い歴史的関係を反映している (ように見える) [Nichols, 1994]
- 従来研究は定量化が不十分に見えるが、それこそ計算集約的統計手法が得意とするところ

その前に：欠損値が大問題

- 行列の要素の73.1%が欠損

N 個の多値特徴

L 個の言語

1	1	...	2
?	4	...	3
1	?	...	?
⋮	⋮	⋮	⋮
3	1	...	?

X

その前に：欠損値が大問題

- 行列の要素の73.1%が欠損
- 提案モデルは2種類のでがかりを用いて補完：
 1. 共時的でがかり (特徴間の依存関係)
 2. 通時的でがかり (言語間の依存関係)

N 個の多値特徴

L 個の言語	1	1	...	2
	?	4	...	3
	1	?	...	?
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	3	1	...	?
	X			

その前に：欠損値が大問題

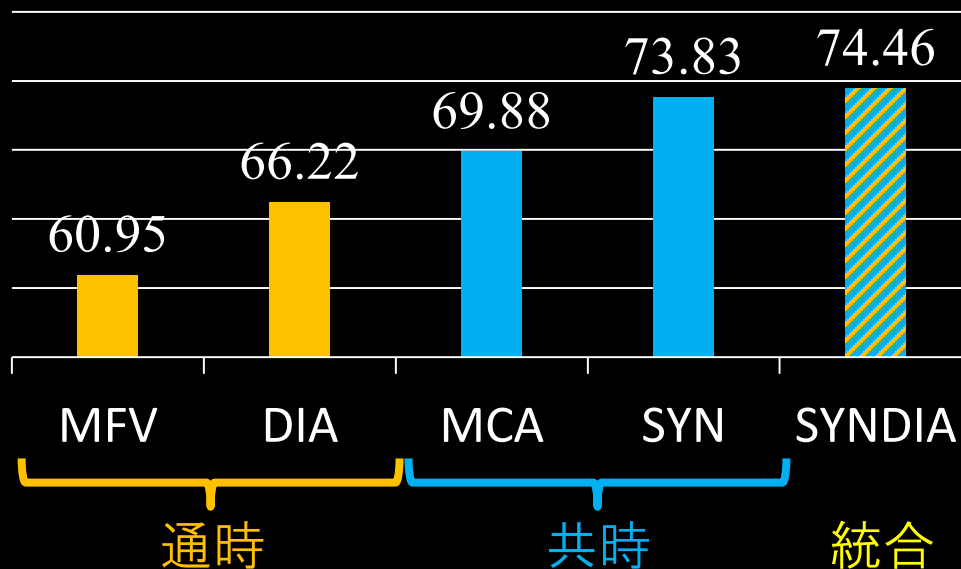
- 行列の要素の73.1%が欠損
- 提案モデルは2種類のでがかりを用いて補完：
 1. 共時的でがかり (特徴間の依存関係)
 2. 通時的でがかり (言語間の依存関係)

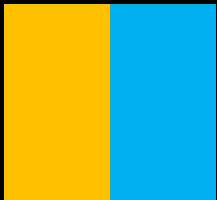
N 個の多値特徴

1	1	...	2
?	4	...	3
1	?	...	?
⋮	⋮	⋮	⋮
3	1	...	?

L 個の言語

X

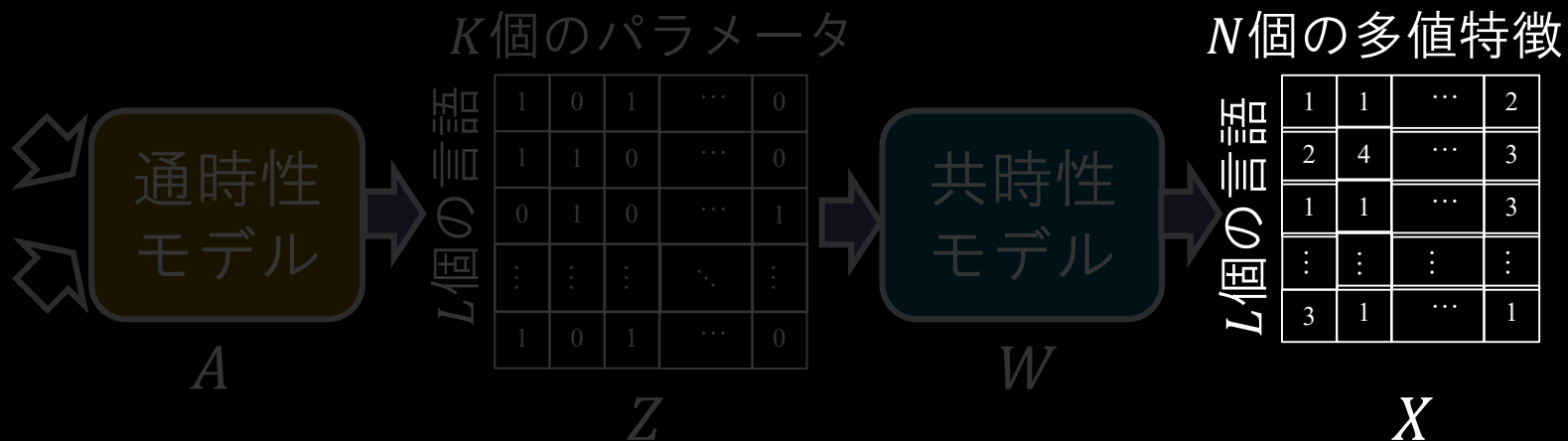


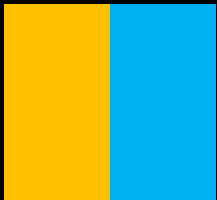


確率的生成モデル： モデル化と推論



地理位置

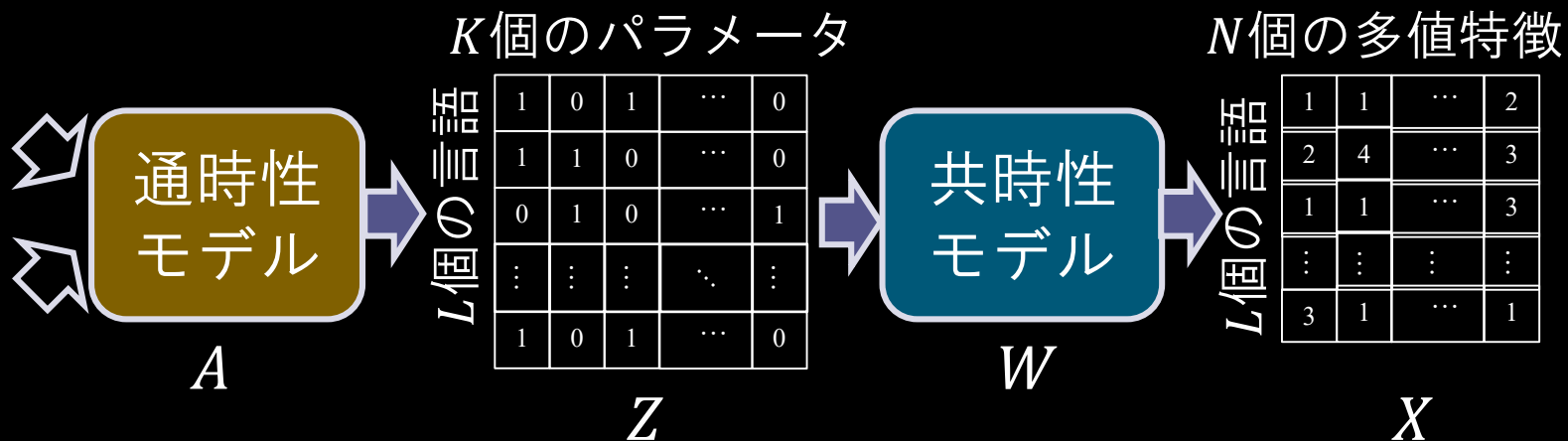




確率的生成モデル： モデル化と推論

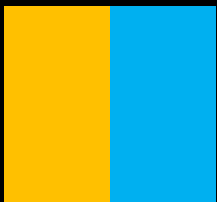


地理位置



生成モデル: $P(A, Z, W, X) = P(A)P(Z|A)P(W)P(X|Z, W)$

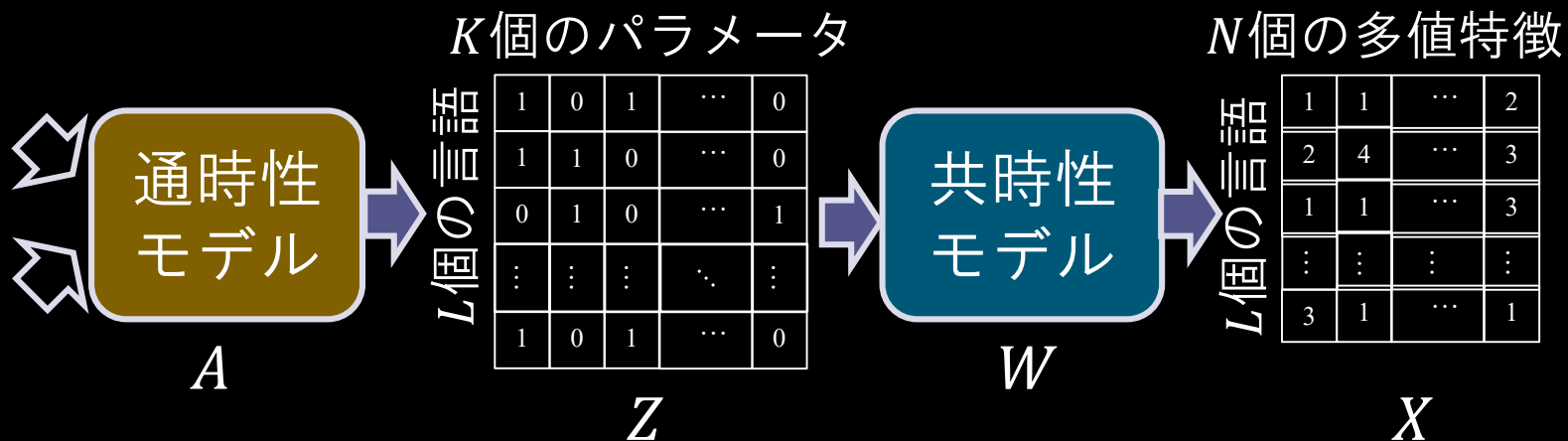
事後推論: $P(A, Z, W, X_{\text{mis}} | X_{\text{obs}})$
 $X = X_{\text{mis}} \cup X_{\text{obs}}$



確率的生成モデル： モデル化と推論

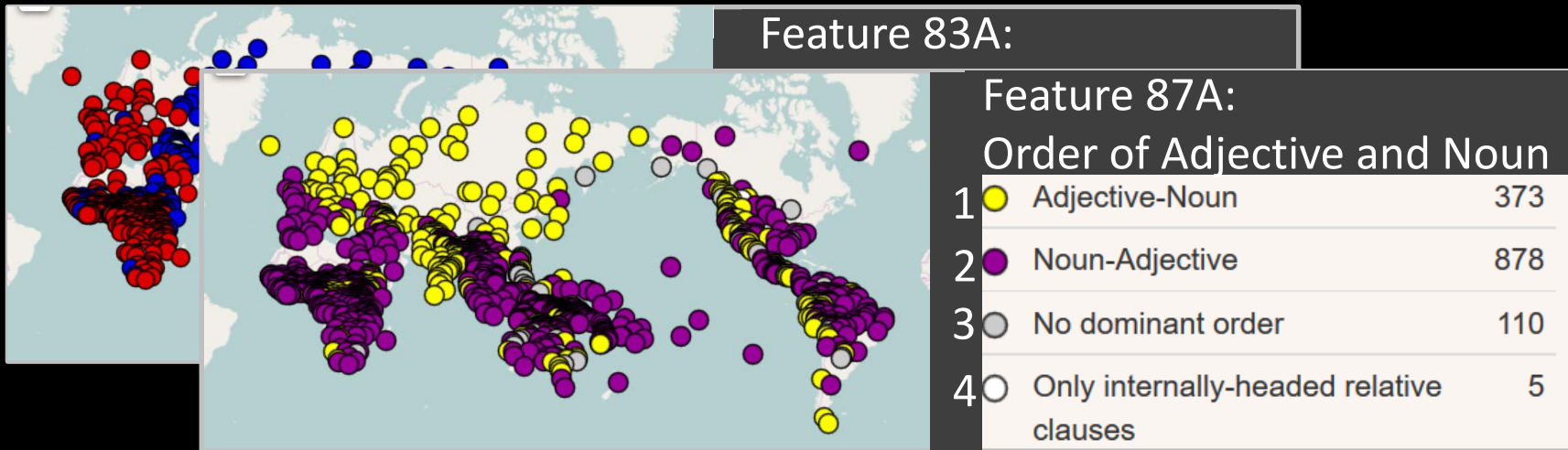


地理位置



$$\text{生成モデル: } P(A, Z, W, X) = P(A)P(Z|A)P(W)P(X|Z, W)$$

共時的でがかり： 特徴間の依存関係

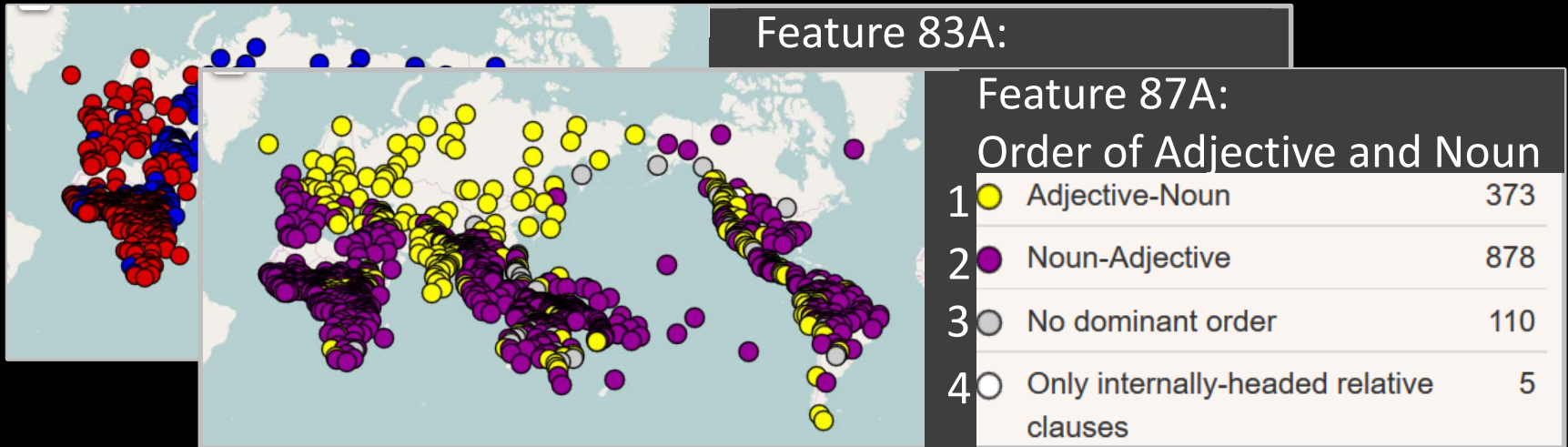


N 個の多値特徴

L 個の言語	1	1	...	2
	?	4	...	3
	1	?	...	?
	⋮	⋮	⋮	⋮
	3	1	...	?

X

共時的でがかり： 特徴間の依存関係



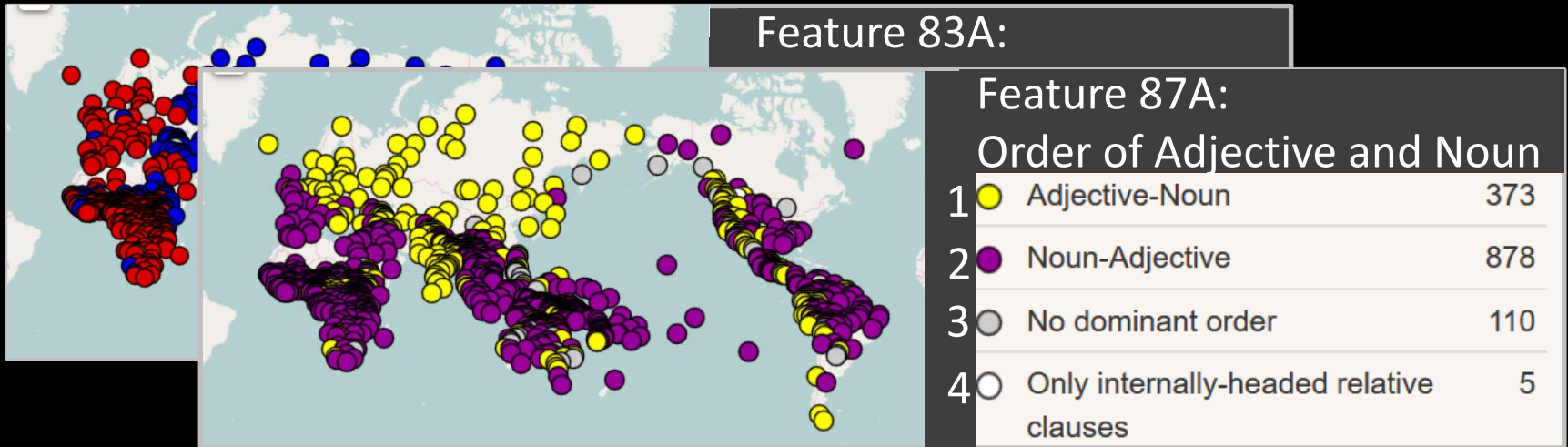
N 個の多値特徴

L 個の言語	1	1	...	2
	?	4	...	3
	1	?	...	?
	⋮	⋮	⋮	⋮
	3	1	...	?

X

Feature 83Aが1 (OV) なら、
Feature 87Aは1 (Adjective-Noun)
の可能性が高い

共時的でがかり： 特徴間の依存関係



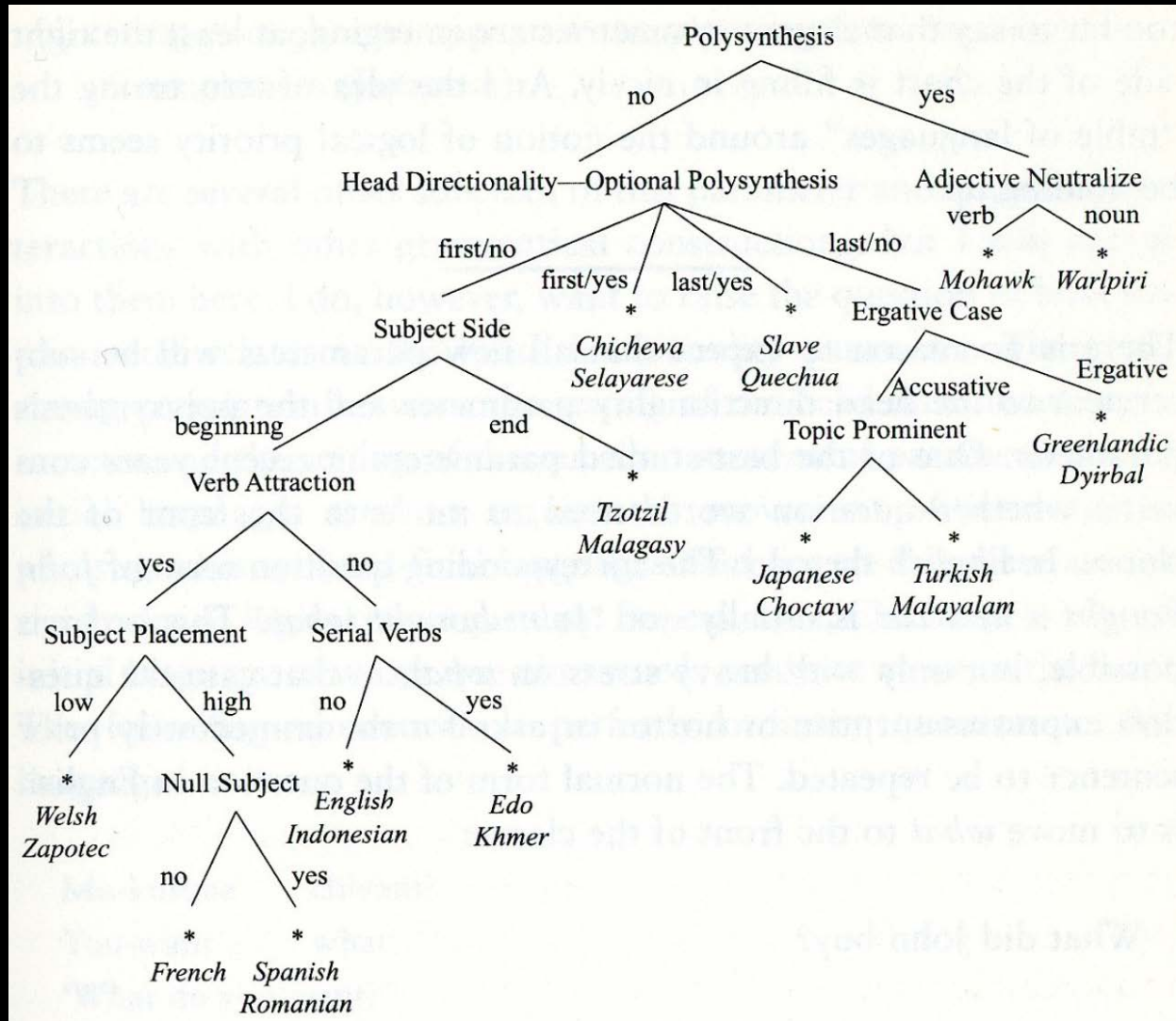
N 個の多値特徴

L 個の言語	1	1	...	2
	?	4	...	3
	1	1	...	?
	⋮	⋮	⋮	⋮
	3	1	...	?

X

Feature 83Aが1 (OV) なら、
Feature 87Aは1 (Adjective-Noun)
の可能性が高い

原理とパラメータ?



[Baker, 2001]

行列分解による定式化

Z
 K 個の2値パラメータ

L 個の言語

1	0	1	0	...
1	1	0	0	...
0	1	0	1	...
⋮	⋮	⋮	⋮	⋮
1	0	1	0	...

$\times$

K 個の2値パラメータ

W
 M 個の2値化特徴

2.9	0.4	-0.3	...	-0.2
6.3	-4.3	-5.7	...	5.9
8.2	-0.2	-2.5	...	0.3
0.2	0.3	1.2	...	-2.4
⋮	⋮	⋮	⋮	⋮

$=$

$\tilde{\Theta}$
 M 個の2値化特徴

L 個の言語

10.2	-9.8	-8.9	...	-4.9
-4.1	8.9	-7.9	...	9.4
8.4	-2.3	-7.3	...	2.5
⋮	⋮	⋮	⋮	⋮
3.9	-4.2	3.5	...	-8.3

Softmax分布から
確率的に生成



N 個の多値特徴

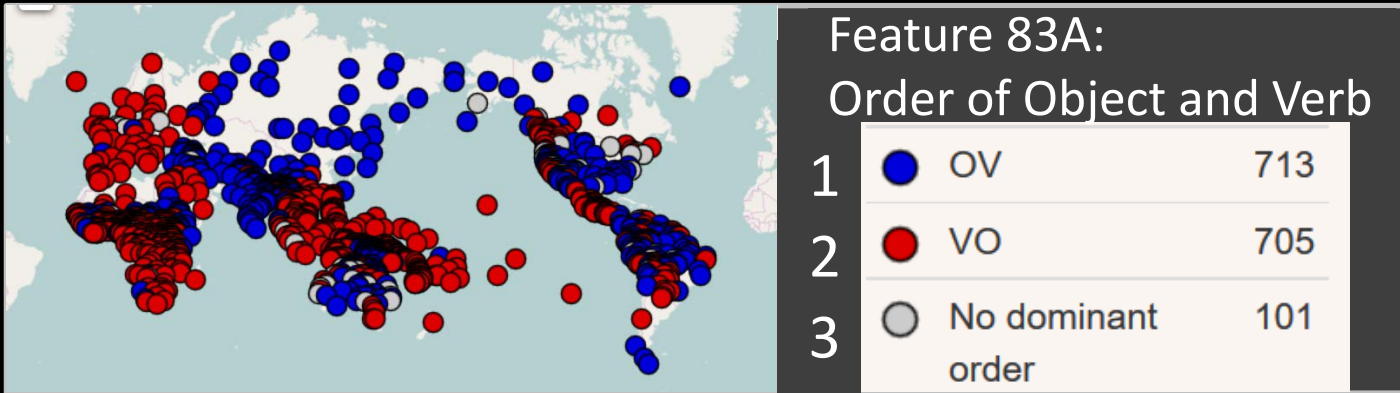
L 個の言語

1	1	...	2
2	4	...	3
1	1	...	3
⋮	⋮	⋮	⋮
3	1	...	1

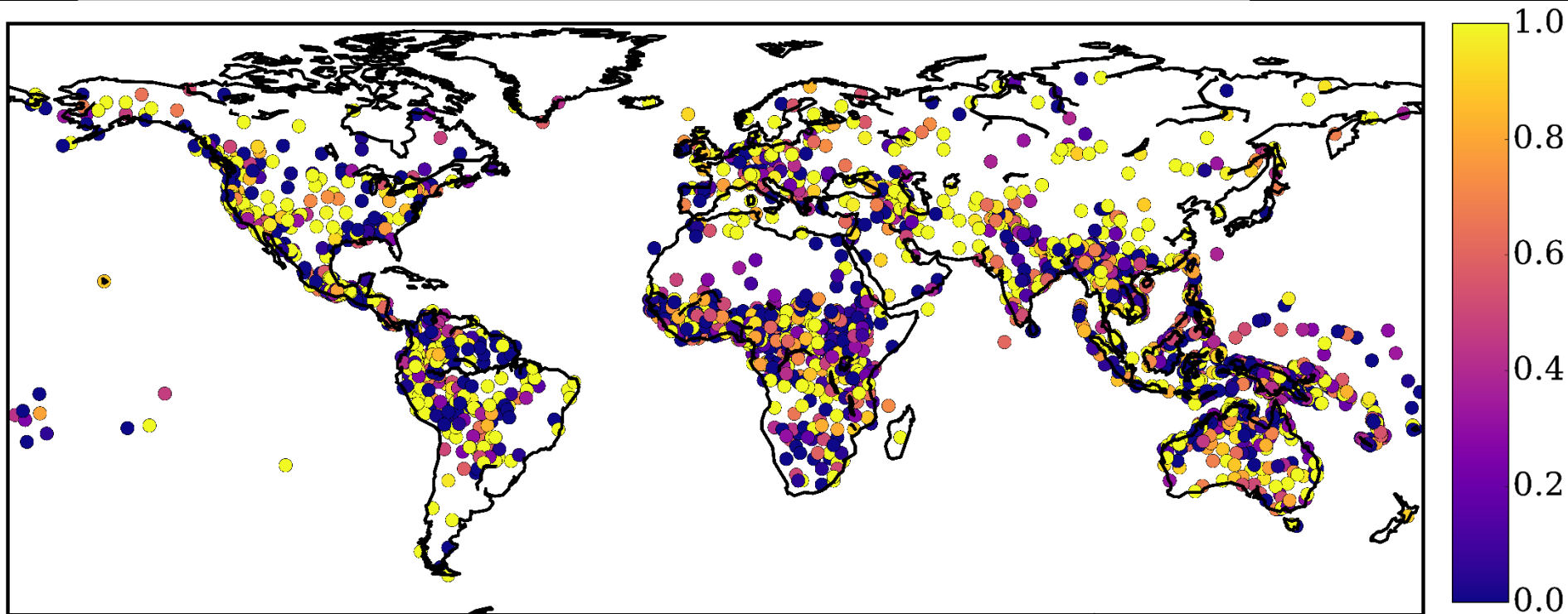
X 8

- 行列 Z は言語の潜在表現 (パラメータ列)
- 行列 W は特徴の値の共起を捉える

獲得されたパラメータの 地理的分布の例



獲得されたパラメータの 地理的分布の例



表層的特徴に見られた地理的信号が潜在的パラメータへの写像時に消えてしまう

通時的てがかり： 言語間の依存関係

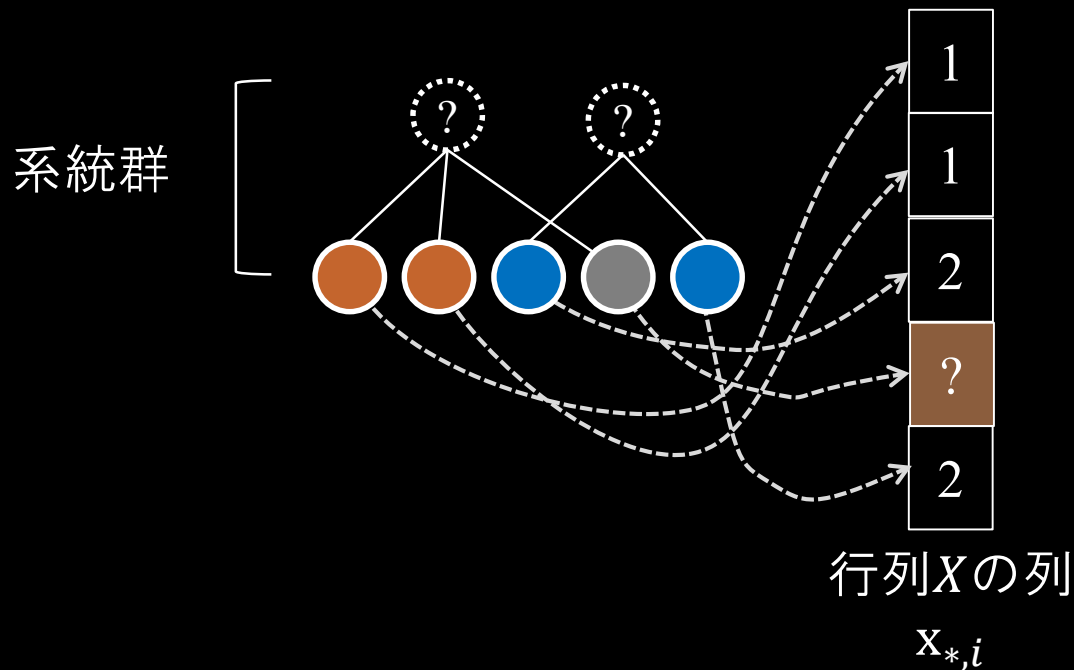


行列 X の列

$X_{*,i}$

言語が同じ値を共有している理由：

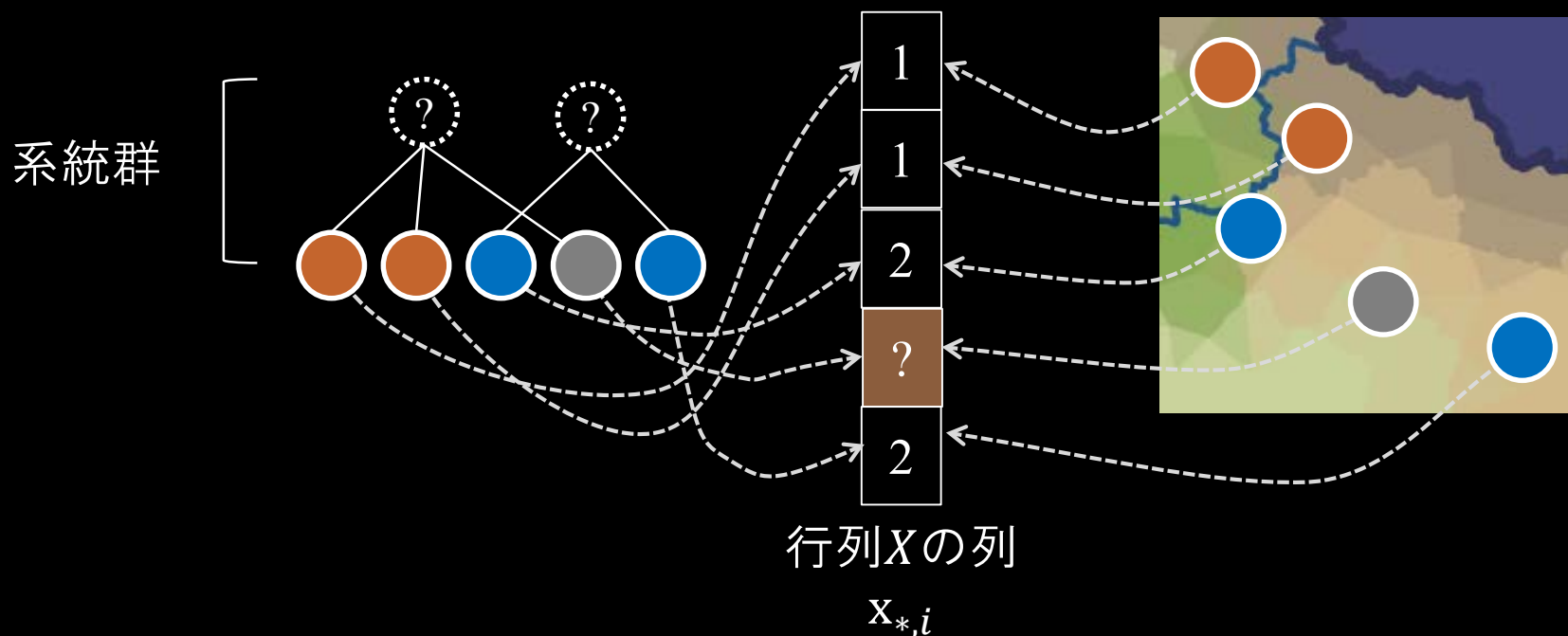
通時的てがかり： 言語間の依存関係



言語が同じ値を共有している理由：

1. 系統的関係 (祖先からの継承)

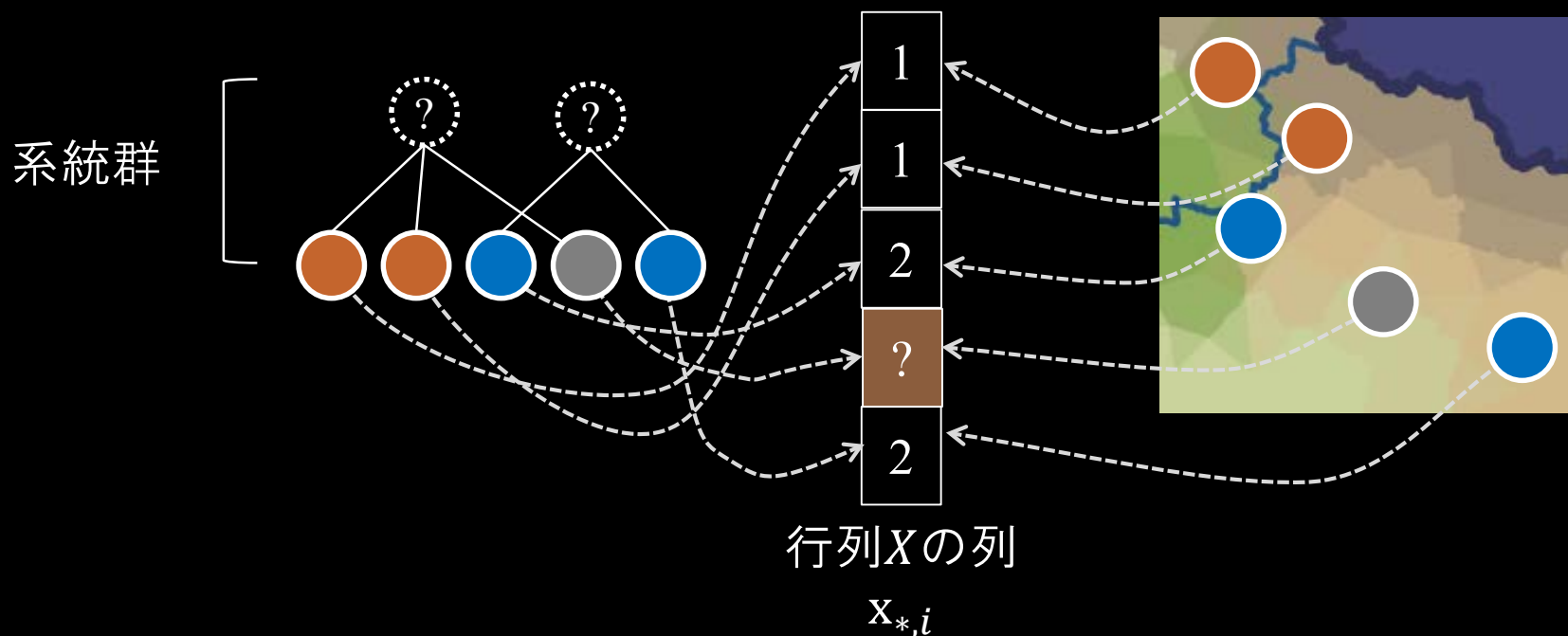
通時的てがかり： 言語間の依存関係



言語が同じ値を共有している理由:

1. 系統的関係 (祖先からの継承)
2. 空間的關係 (接触の影響)

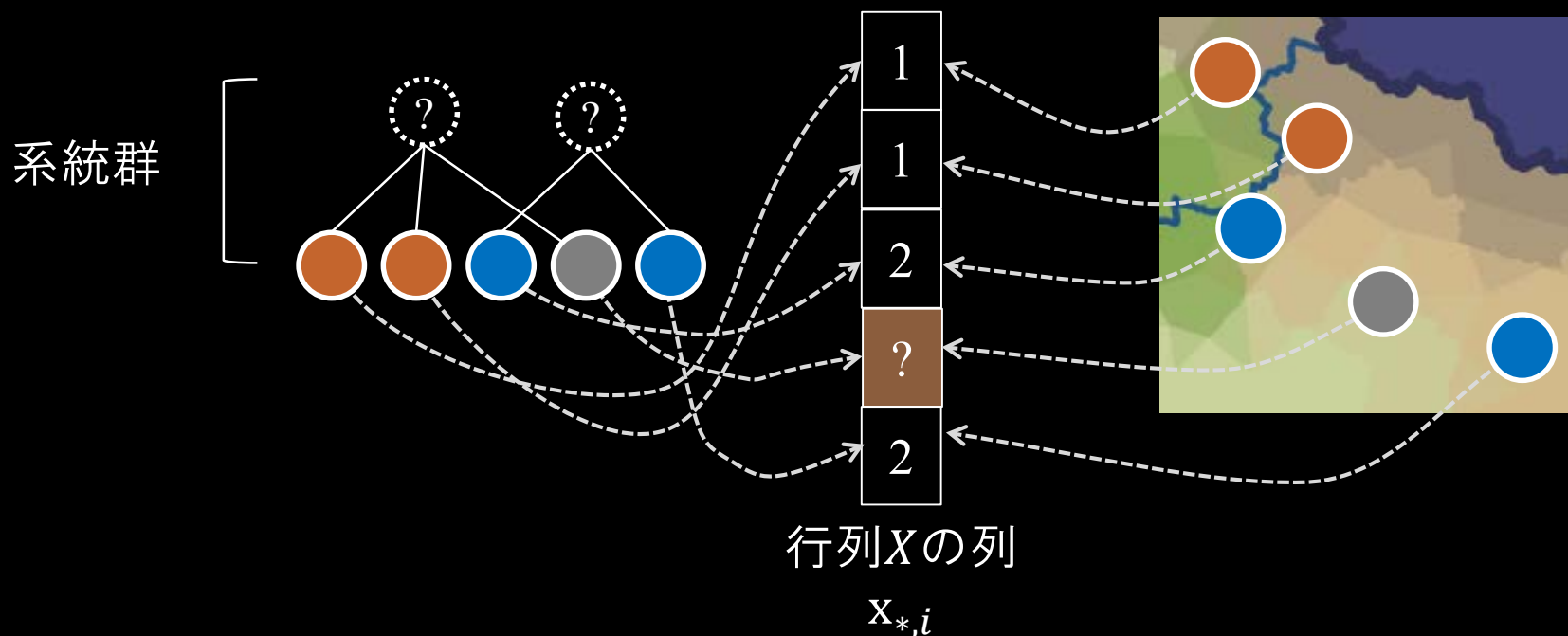
通時的てがかり： 言語間の依存関係



言語が同じ値を共有している理由:

1. 系統的関係 (祖先からの継承)
2. 空間的關係 (接触の影響)
3. そもそも普遍的に高頻度な値

通時的てがかり： 言語間の依存関係



言語が同じ値を共有している理由:

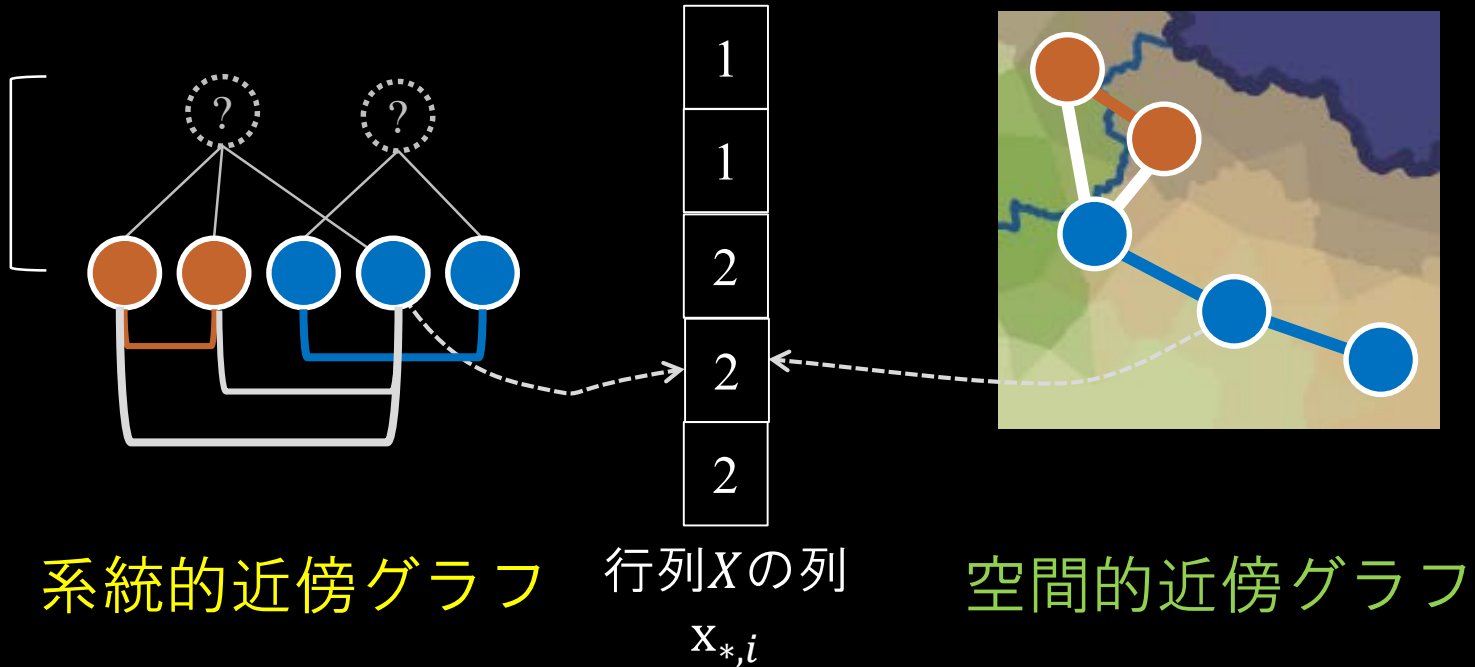
1. 系統的関係 (祖先からの継承)
2. 空間的關係 (接触の影響)
3. そもそも普遍的に高頻度な値

その強度を制御する要因

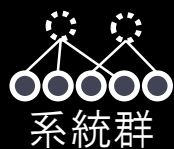
垂直安定性
水平拡散性
普遍性

2種類の近傍グラフ

系統群



自己ロジスティックモデル



言語 l が特徴 i の値として j をとる確率:

$$P(x_{l,i} = j | x_{-l,i}, v_i, h_i, u_i) \\ \propto \exp(v_i V_{l,i,j} + h_i H_{l,i,j} + u_{i,j})$$

言語 l の空間的近傍で j をとる言語の数
言語 l の系統的近傍で j をとる言語の数
普遍性
水平拡散性
垂直安定性

[Yamauchi+, COLING2016]

ベイズ的拡張がJournal of Language Evolutionに採録

自己ロジスティックモデル



行列 X の列 $x_{*,i}$ の生成確率:



$$P(x_{*,i} | v_i, h_i, u_i) = \frac{\exp(v_i V(x_{*,i}) + h_i H(x_{*,i}) + \sum_j u_{i,j} U_j(x_{*,i}))}{\sum_{x'_{*,i}} \exp(v_i V(x'_{*,i}) + h_i H(x'_{*,i}) + \sum_j u_{i,j} U_j(x'_{*,i}))}$$

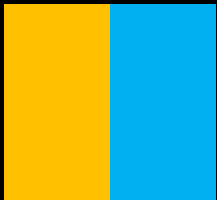
$V(x_{*,i})$: 系統的近傍グラフにおいて同じ値をとる辺の数

$H(x_{*,i})$: 空間的近傍グラフにおいて同じ値をとる辺の数

$U_j(x_{*,i})$: $x_{*,i}$ において値 j をとる言語の数

[Yamauchi+, COLING2016]

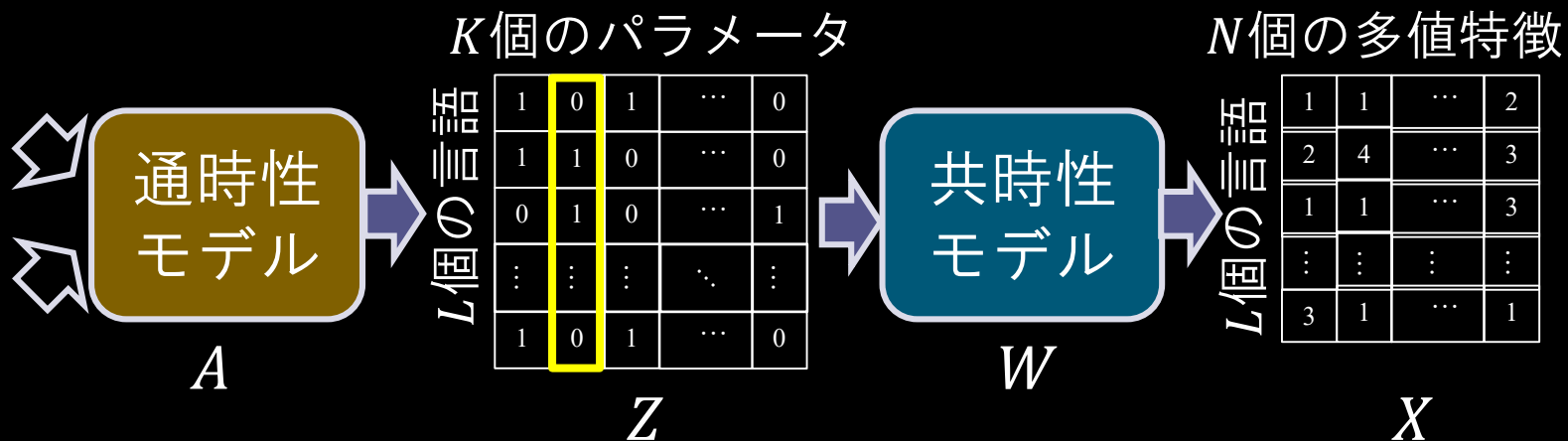
ベイズ的拡張がJournal of Language Evolutionに採録



統合モデル： 通時性モデルを潜在空間で利用



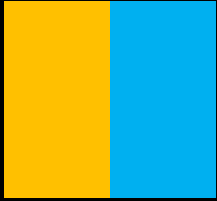
地理位置



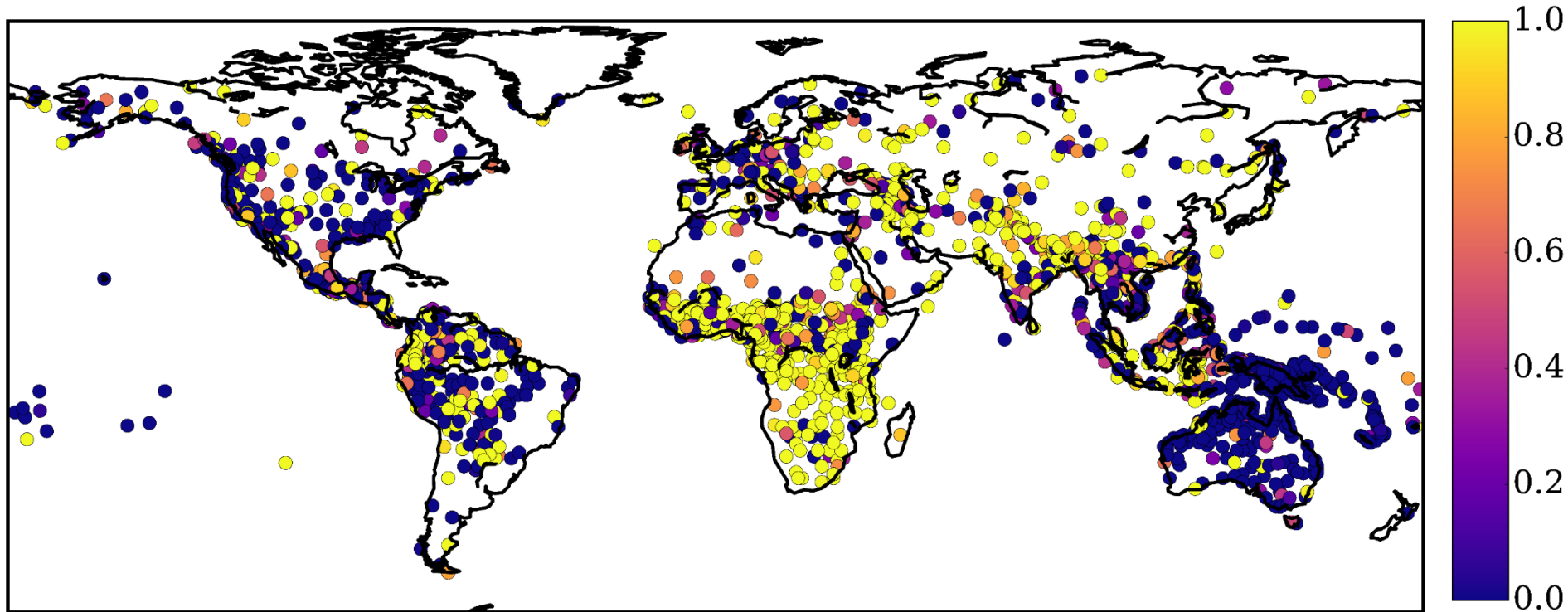
$$P(A, Z, W, X) = P(A)P(Z|A)P(W)P(X|Z, W)$$

通時性モデル

共時性モデル

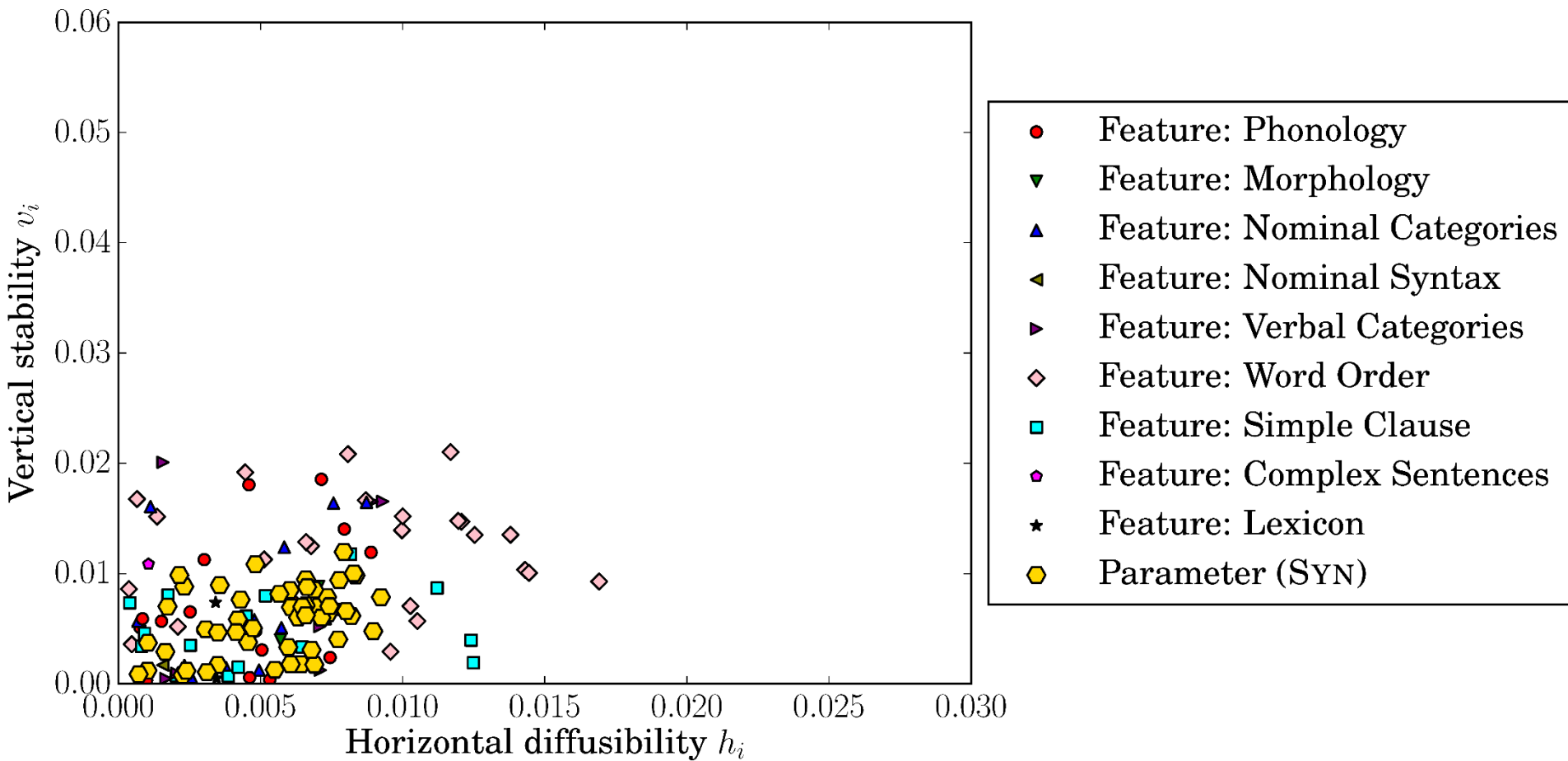


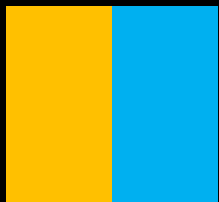
獲得されたパラメータの 地理的分布の例



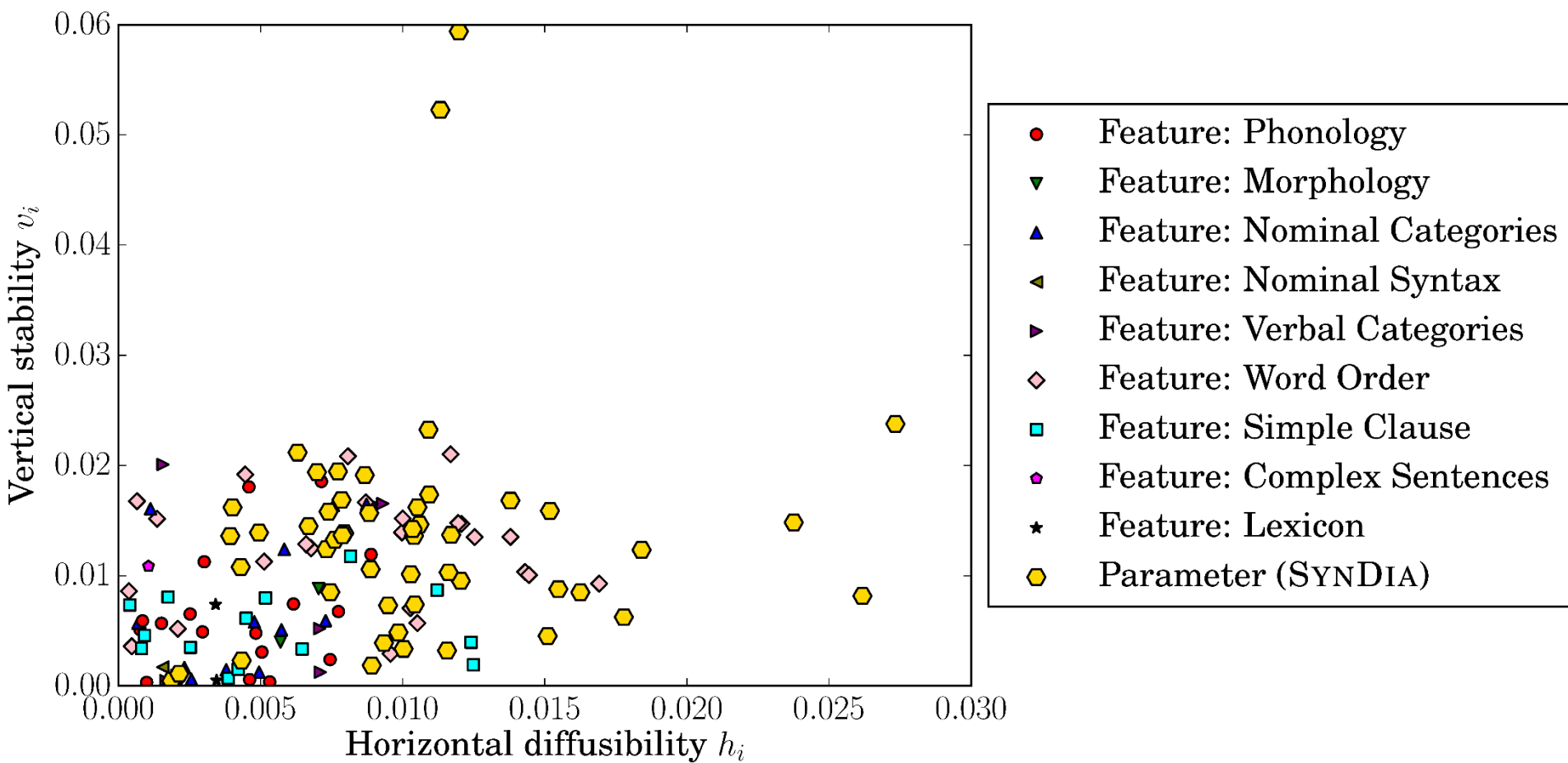
表層的特徴に見られた地理的信号を潜在的パラ
メータへ写像しても維持している (ように見える)

獲得されたパラメータの 垂直安定性と水平拡散性





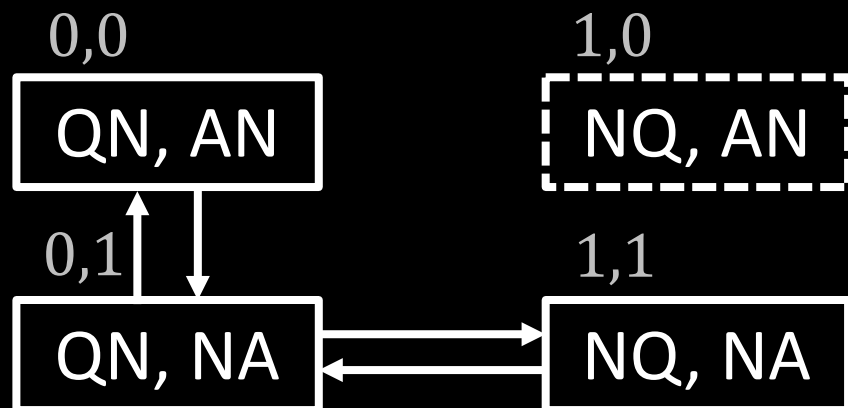
獲得されたパラメータの 垂直安定性と水平拡散性



まとめと今後の課題

- 共時性と通時性を同時に考慮するベイズ統計モデルを提案
- 獲得された潜在的2値パラメータは空間的信号を維持しているようにみえる
- 残る疑問: 潜在的2値パラメータを使って諸言語の深い歴史的関係を解明できる?

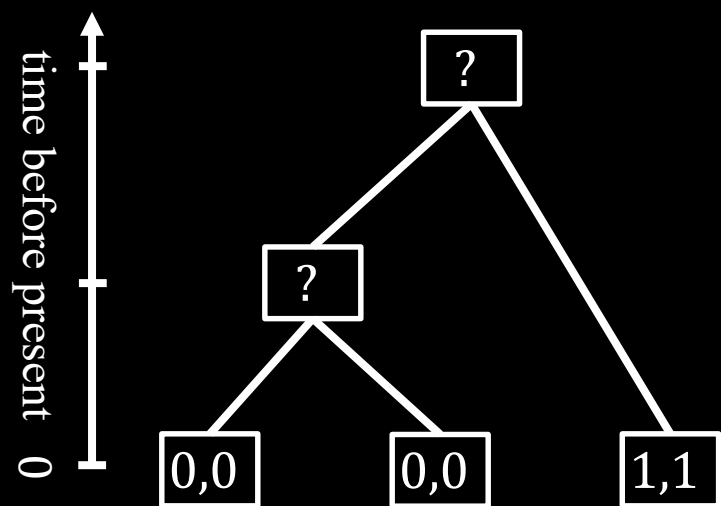
表層的特徴の状態空間モデル



QN: 数詞 + 名詞 語順
AN: 形容詞 + 名詞 語順
NQ, NAは逆の語順

- 特徴間の依存関係を適切にモデル化すれば、不自然な祖語の再構を避けられるのでは
- 従来モデルは表層的特徴を展開 [Greenberg, 1978]
 - 既知の年代つき系統樹を用いた置換モデルのパラメータ推定 [Dunn+, Nature 2011]
- 3個以上の特徴間の依存関係のモデル化が難しい

状態空間モデルの置換モデル



遷移確率: $P(x = j | \pi(x) = i, t) = \exp(tQ)_{i,j}$

$$Q = \begin{matrix} & \begin{matrix} 0,0 & 0,1 & 1,0 & 1,1 \end{matrix} \\ \begin{matrix} 0,0 \\ 0,1 \\ 1,0 \\ 1,1 \end{matrix} & \begin{pmatrix} * & q_{12} & q_{13} & - \\ q_{21} & * & - & q_{24} \\ q_{31} & - & * & q_{34} \\ - & q_{42} & q_{43} & * \end{pmatrix} \end{matrix}$$

パラメータは複数の表層的特徴の 依存関係を捉える

