Diachrony-aware Induction of Binary Latent Representations from Typological Features

MURAWAKI Yūgo
Kyoto University

The source code is available at
https://github.com/murawaki/latent-typology/

IJCNLP 2017 @ Taipei, Taiwan 29 Nov. 2017

IJC*NLP*? No Text Processing!

And I abandoned a ***neural network*** model I proposed 2 years ago, and switched back to a ***Bayesian*** model

IJC***NLP***? No Text Processing!

And I abandoned a ***neural network*** model I proposed 2 years ago, and switched back to a ***Bayesian*** model

- What I am playing with is matrix X
  - $L = 2{,}679,\ N = 104$

N categorical features

L languages				
	1	1	⋯	2
	2	4	⋯	3
	1	1	⋯	3
	⋮	⋮	⋮	⋮
	3	1	⋯	1

X

IJC*NLP*? No Text Processing!

And I abandoned a ***neural network*** model I proposed 2 years ago, and switched back to a ***Bayesian*** model

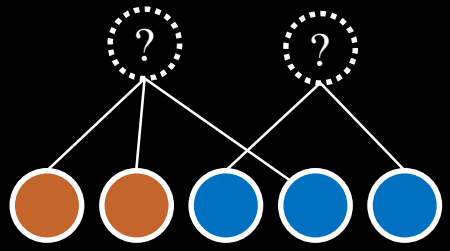
- What I am playing with is matrix X
  - $L = 2{,}679,\ N = 104$
- And make use of two additional resources

N categorical features

L languages

1	1	⋯	2
2	4	⋯	3
1	1	⋯	3
⋮	⋮	⋮	⋮
3	1	⋯	1

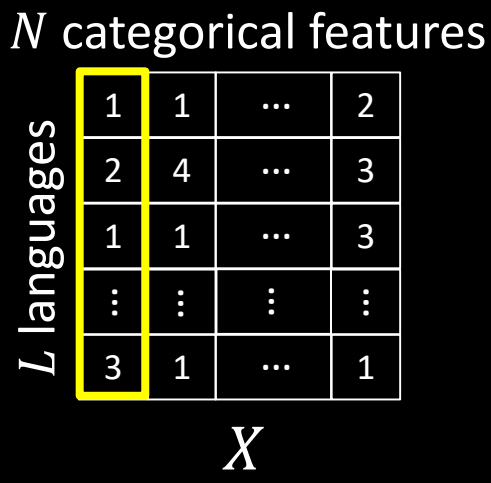
X



Phylogenetic groupings of languages

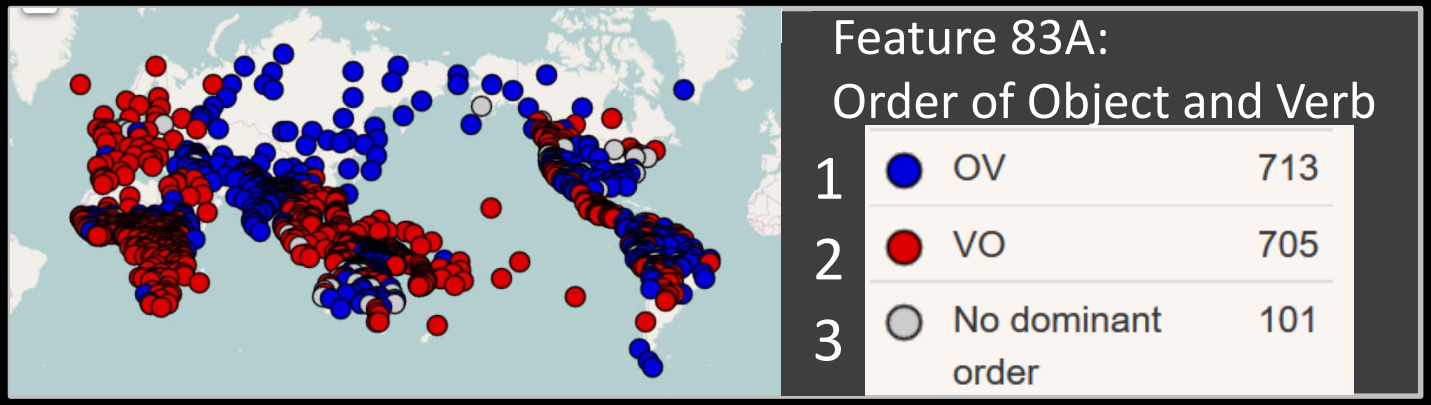
Spatial locations of languages

Features of Linguistic Typology



Source: Online Edition of WALS [Haspelmath+, 2005]

Features of Linguistic Typology



N categorical features

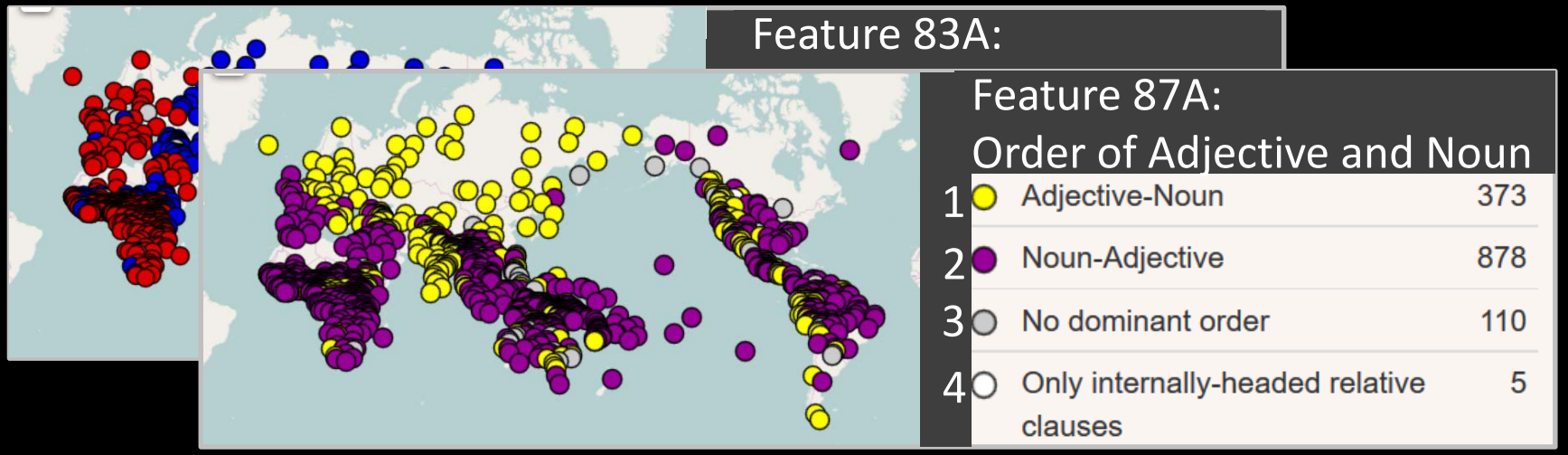
L languages

1	1	⋯	2
2	4	⋯	3
1	1	⋯	3
⋮	⋮	⋮	⋮
3	1	⋯	1

X

Source: Online Edition of WALS [Haspelmath+, 2005]

Features of Linguistic Typology



N categorical features

1	1	⋯	2
2	4	⋯	3
1	1	⋯	3
⋮	⋮	⋮	⋮
3	1	⋯	1

L languages

X

Source: Online Edition of WALS [Haspelmath+, 2005]

Questions (Not Addressed in This Talk)

1. How have these features changed over time?
2. Can we use these features to trace the deep history of languages?

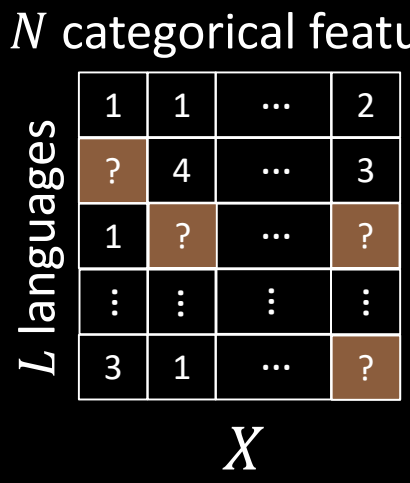
N categorical features

L languages	1	1	⋯	2
	2	4	⋯	3
	1	1	⋯	3
	⋮	⋮	⋮	⋮
	3	1	⋯	1

X

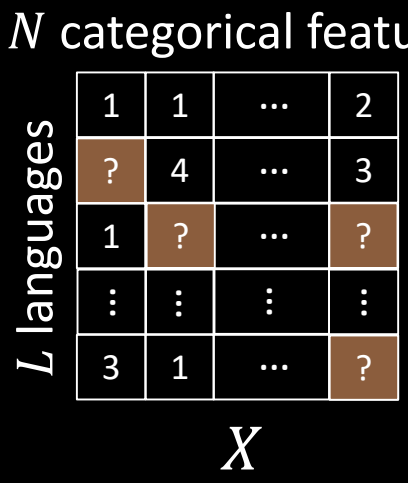
Missing Values are a Real Problem

- $\mathbf{73.1\%}$ of items are missing



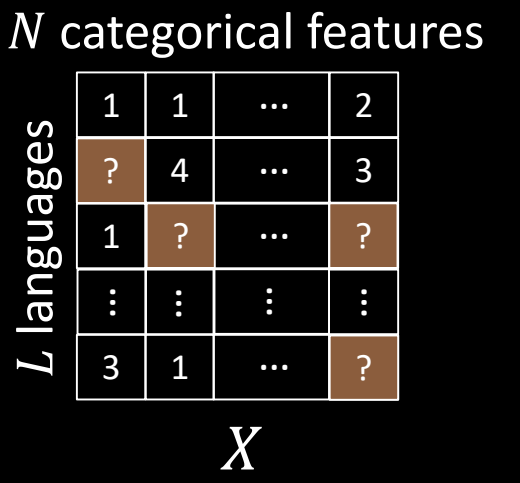
Missing Values are a Real Problem

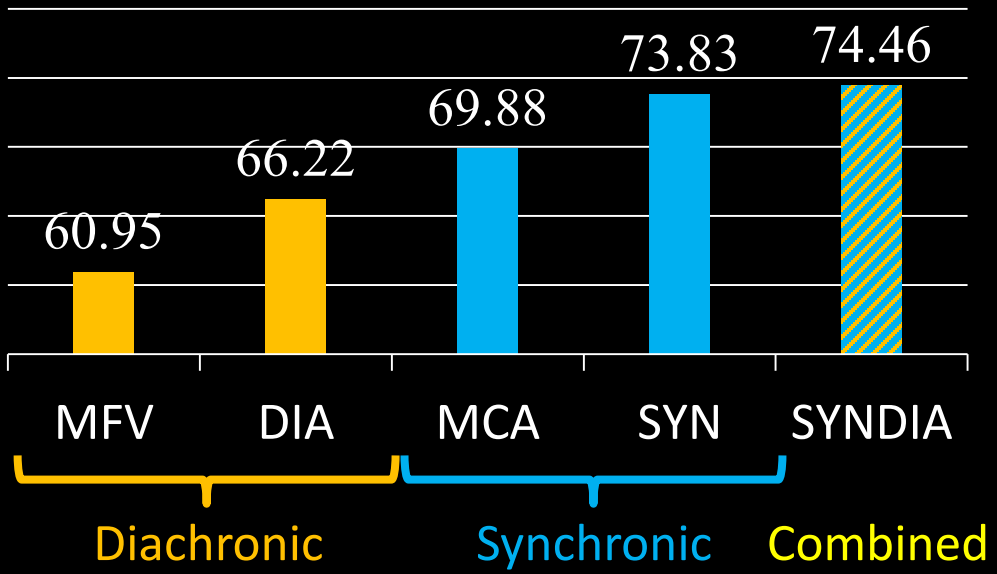
- 73.1% of items are missing
- My model imputes missing values by combining
  1. Synchronic clues (inter-feature dependencies)
  2. Diachronic clues (inter-language dependencies)



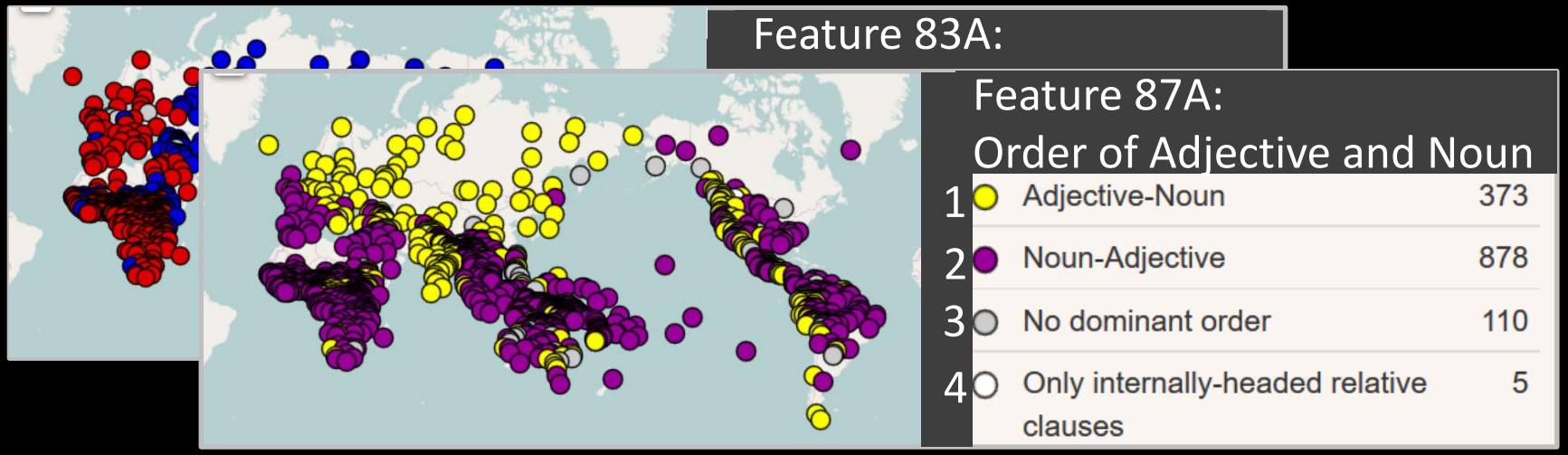
Missing Values are a Real Problem

- 73.1% of items are missing
- My model imputes missing values by combining
  1. Synchronic clues (inter-feature dependencies)
  2. Diachronic clues (inter-language dependencies)





Synchronic Clues: Inter-feature Dependencies



N categorical features

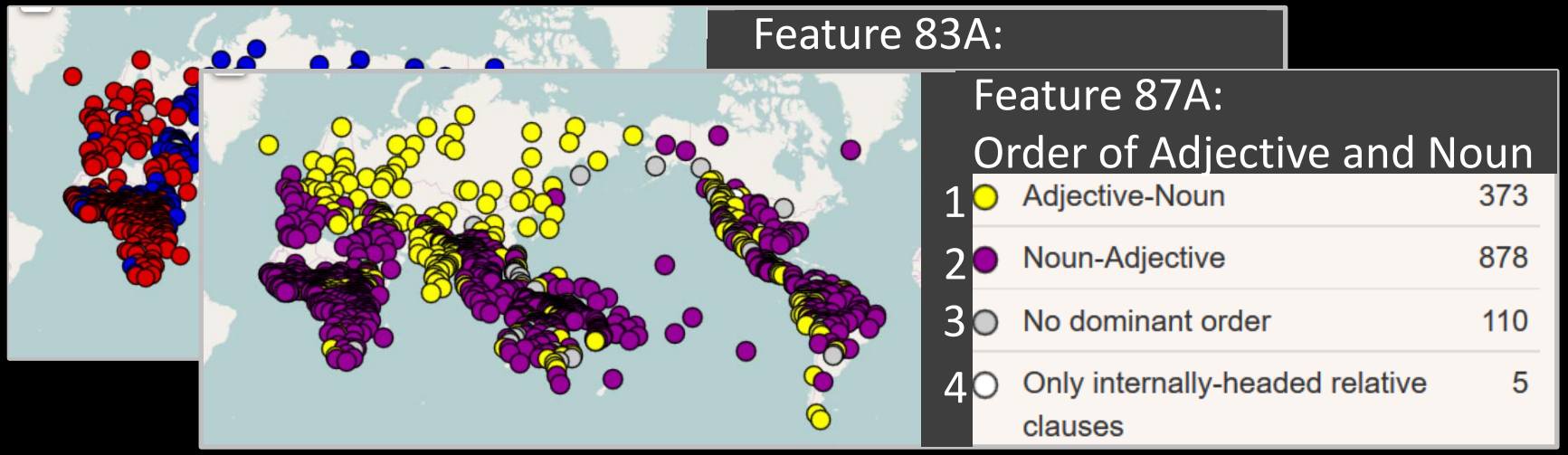
L languages

1	1	…	2
?	4	…	3
1	?	…	?
⋮	⋮	⋮	⋮
3	1	…	?

X

[Greenberg, 1963]

Synchronic Clues: Inter-feature Dependencies



N categorical features

L languages

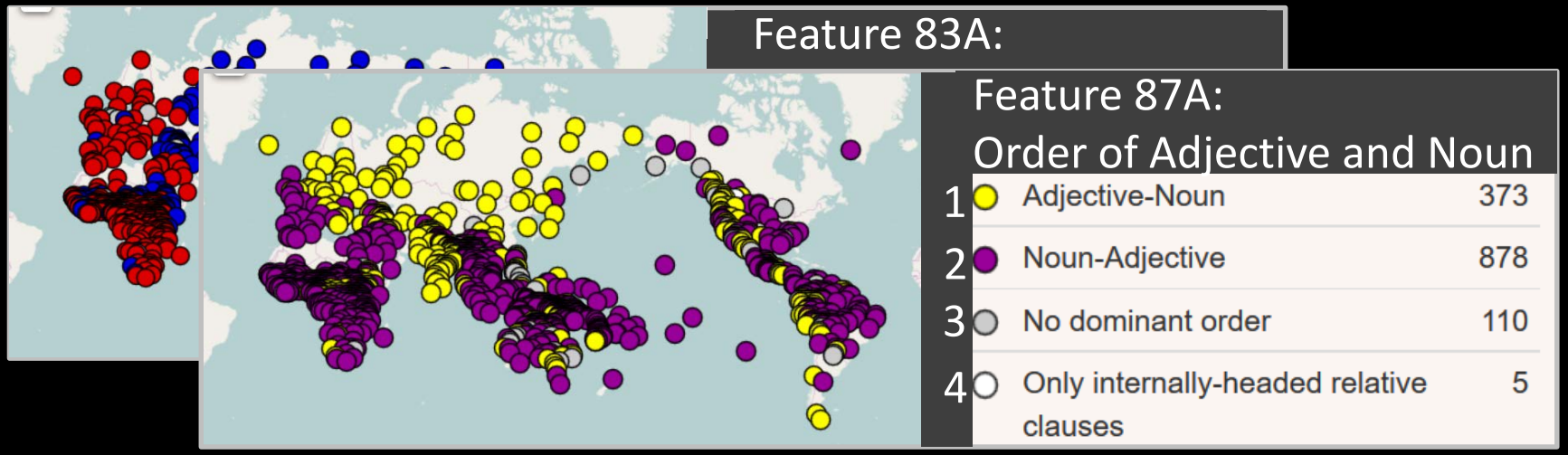
1	1	…	2
?	4	…	3
1	?	…	?
⋮	⋮	⋮	⋮
3	1	…	?

X

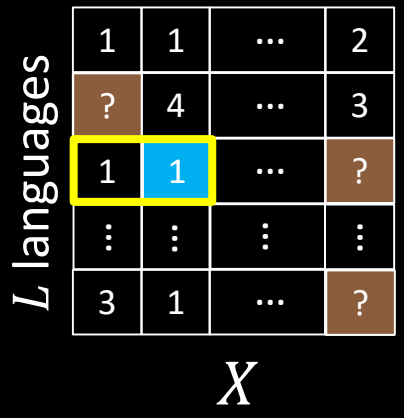
If Feature 83A is 1 (OV), then Feature 87A is likely to be 1 (Adjective-Noun)

[Greenberg, 1963]

Synchronic Clues: Inter-feature Dependencies



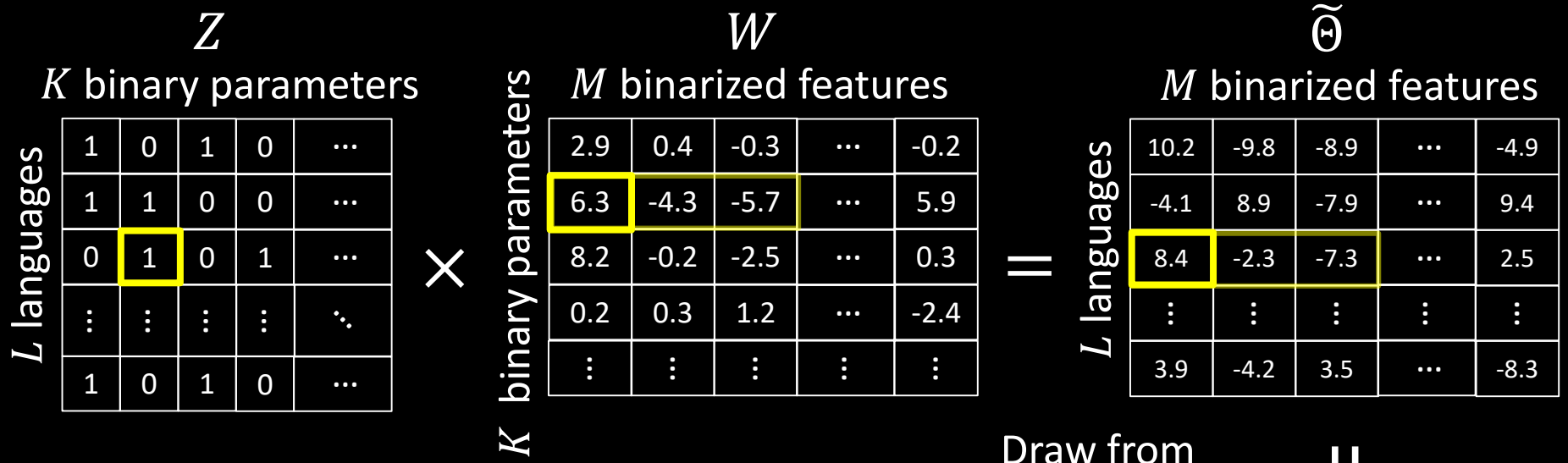
N categorical features



If Feature 83A is 1 (OV), then Feature 87A is likely to be 1 (Adjective-Noun)

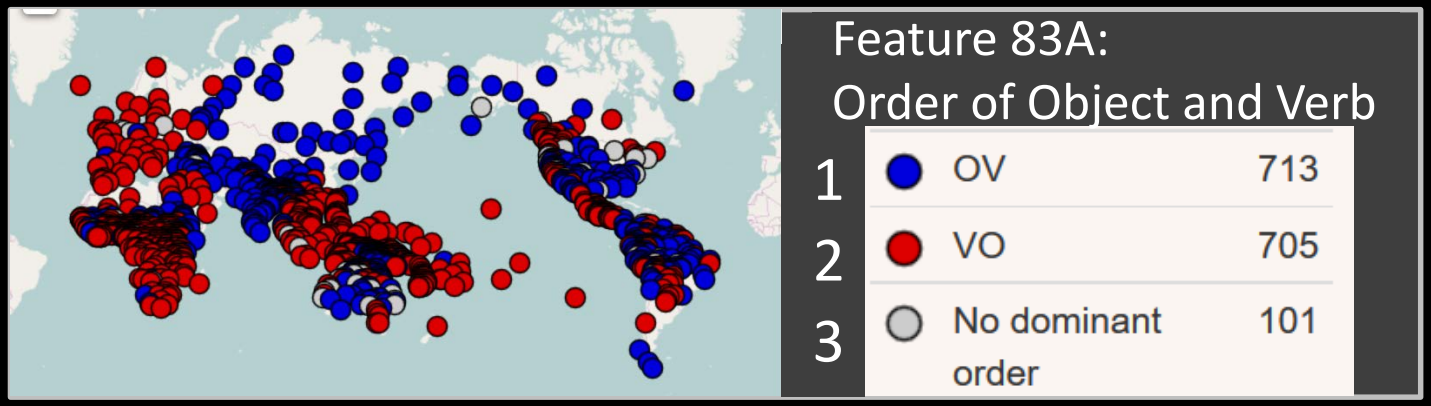
[Greenberg, 1963]

Matrix Decomposition

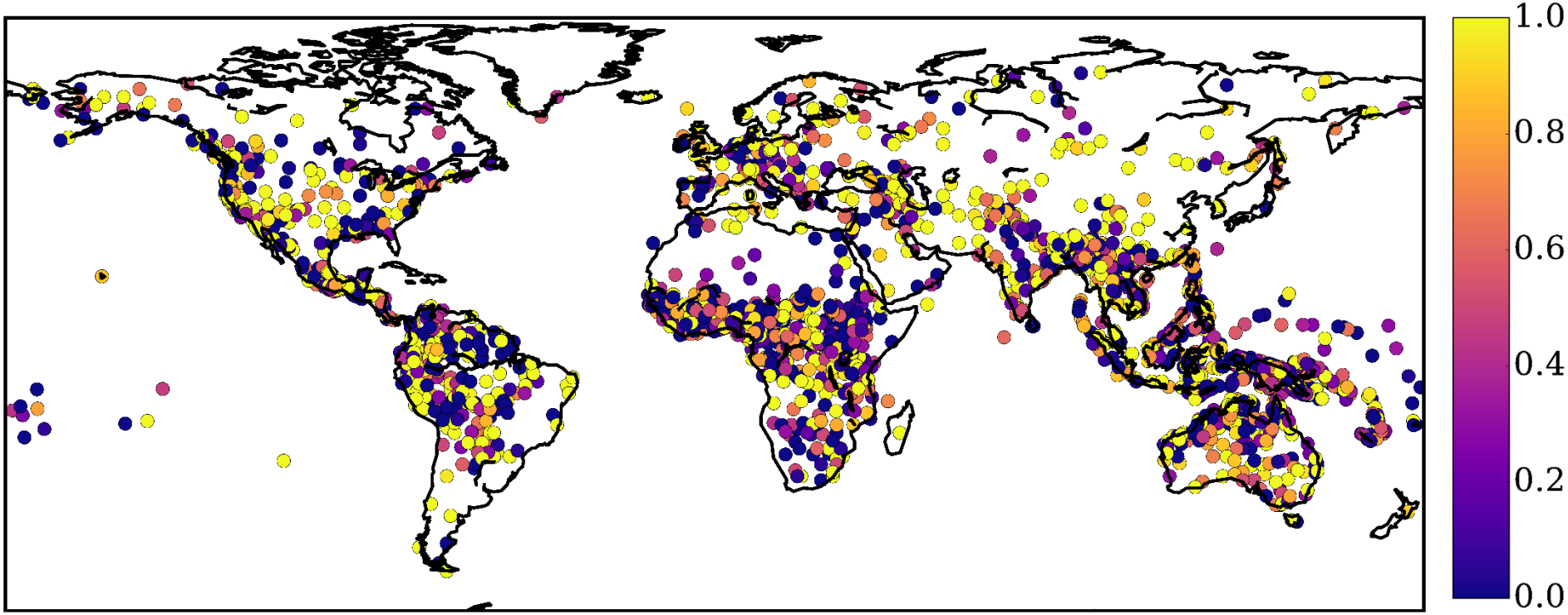


- Z is a latent representation of languages (**parameters**)
- W captures co-occurrences of feature values

Geographic Distribution of an Induced Parameter



Geographic Distribution of an Induced Parameter



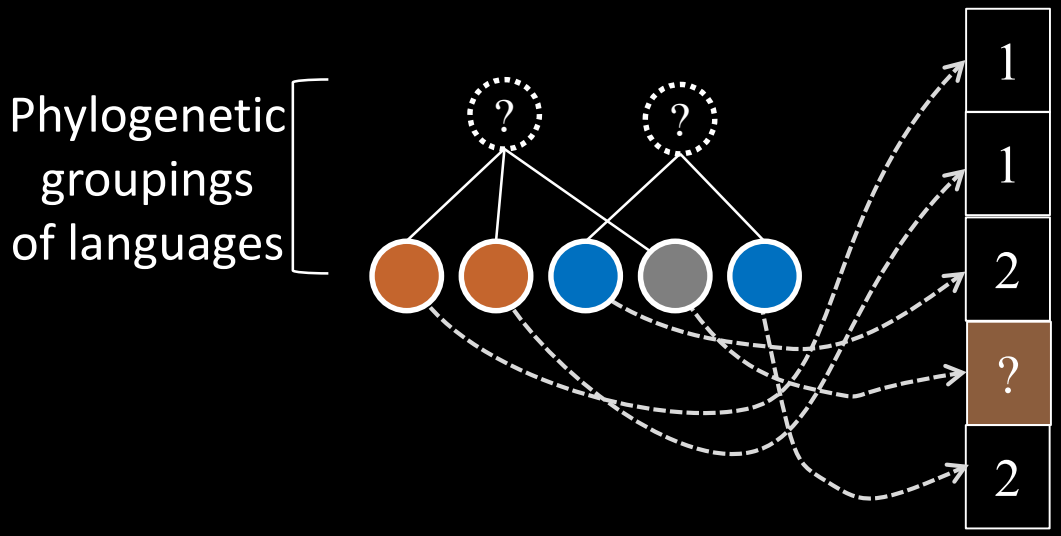
Geographic signals observed in surface features are **somehow lost** during the induction

Diachronic Clues: Inter-language Dependencies

A ***column***
of the matrix

Languages may share the same value because they …

Diachronic Clues: Inter-language Dependencies

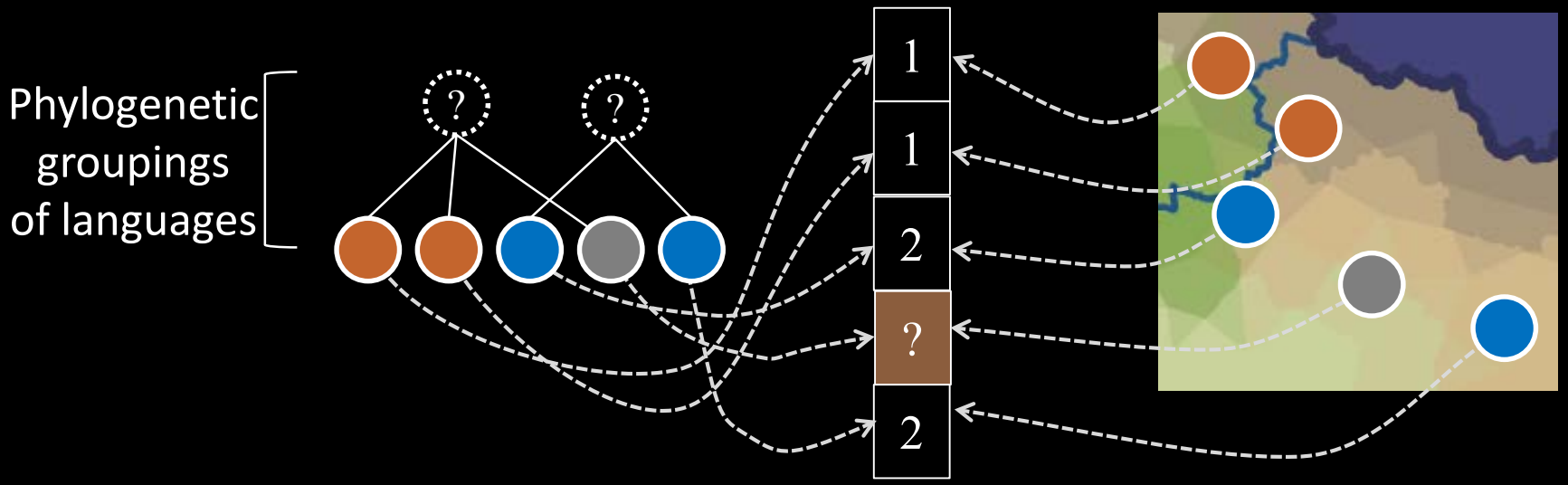


A ***column*** of the matrix

Languages may share the same value because they …

1. are phylogenetically related,

Diachronic Clues: Inter-language Dependencies

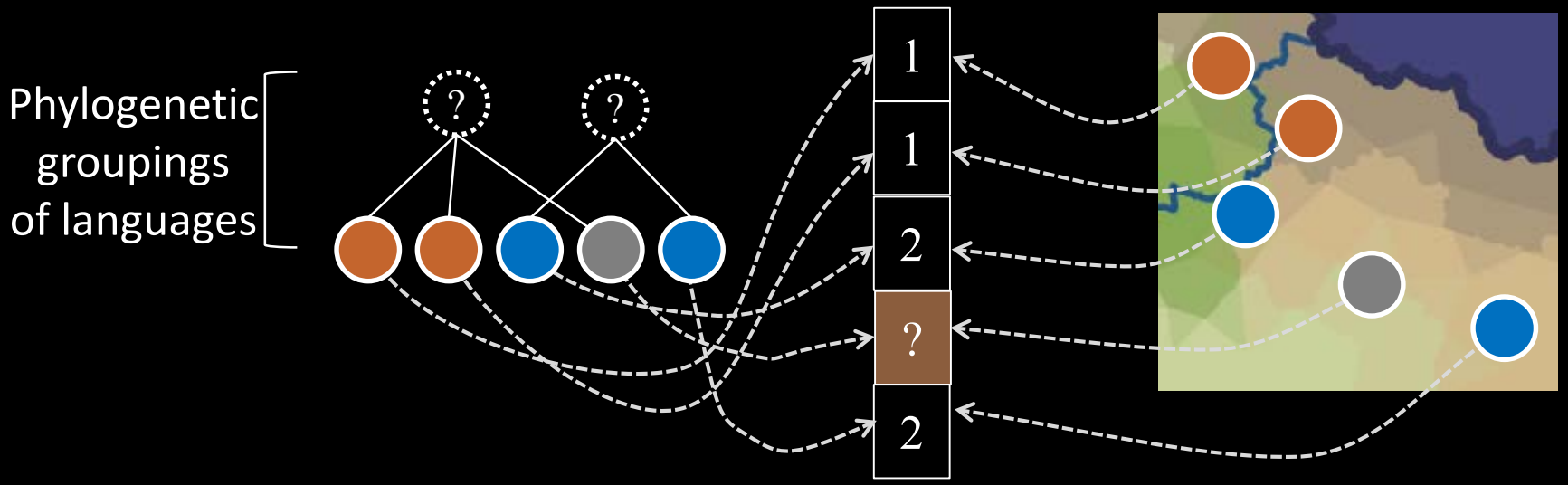


A ***column*** of the matrix

Languages may share the same value because they …

1. are phylogenetically related,
2. are spatially close (in contact), and/or

Diachronic Clues: Inter-language Dependencies

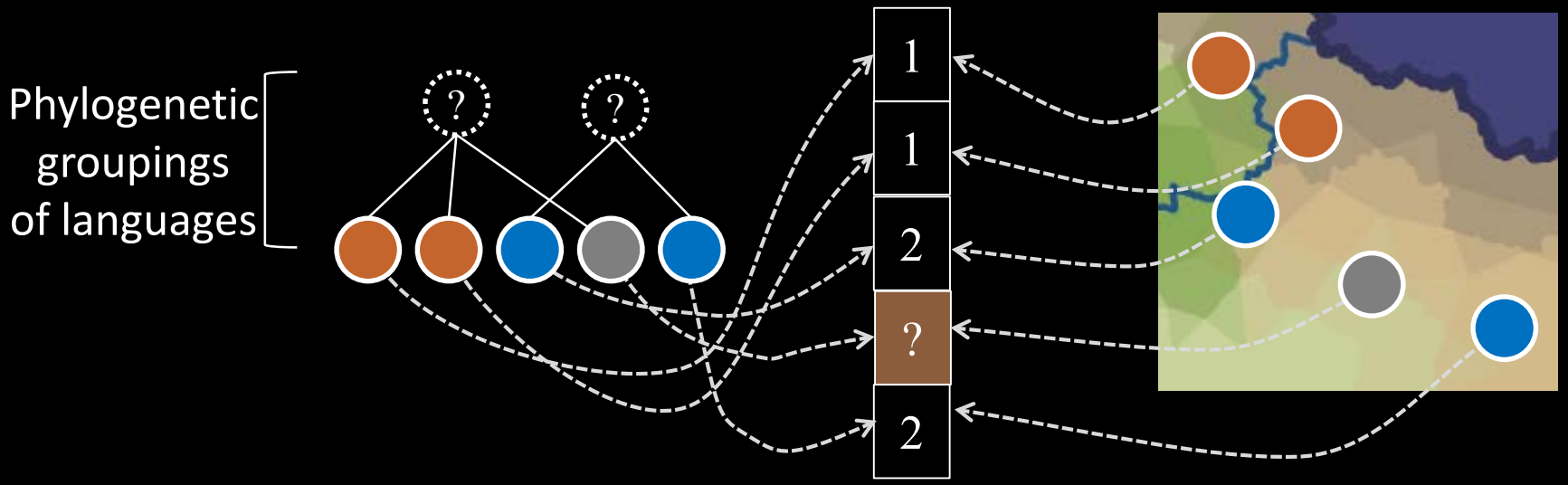


A ***column*** of the matrix

Languages may share the same value because they …

1. are phylogenetically related,
2. are spatially close (in contact), and/or
3. just take a universally (globally) frequent value

Diachronic Clues: Inter-language Dependencies



A ***column*** of the matrix

Languages may share the same value because they …

1. are phylogenetically related,
2. are spatially close (in contact), and/or
3. just take a universally (globally) frequent value

Strength controlled by

- vertical stability
- horizontal diffusibility
- universality

Autologistic Model

Phylogenetic groupings of languages

spatial locations of languages

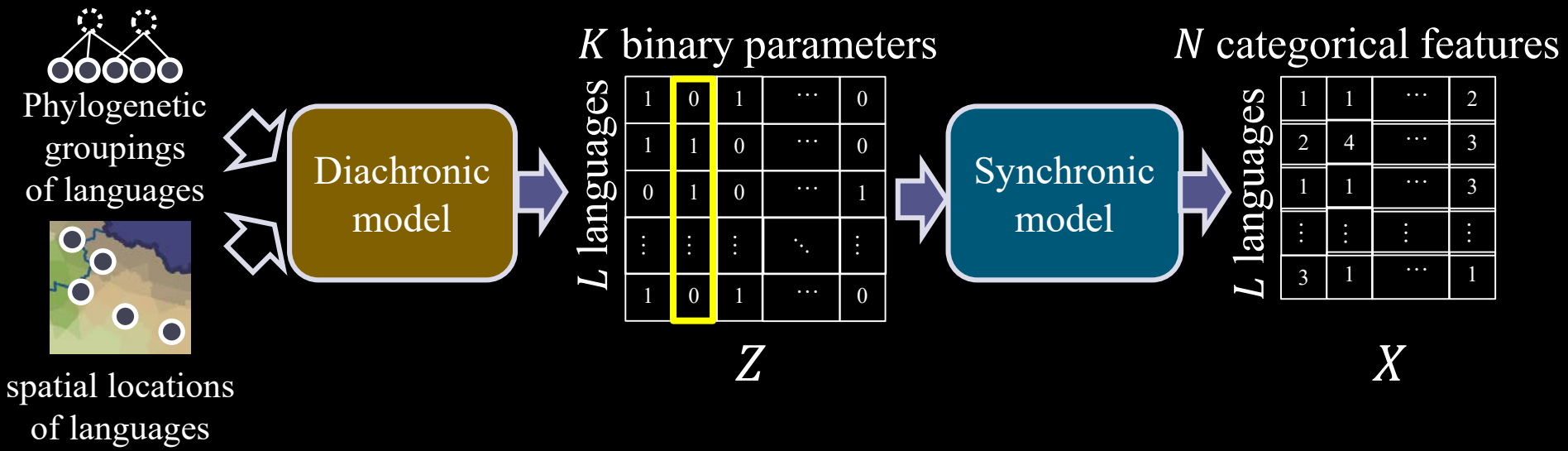
The prob. of language l taking value j for feature i is

$$P(x_{l,i} = j \mid \mathrm{x}_{-l,i}, v_i, h_i, \mathrm{u}_i) \propto \exp(v_i V_{l,i,j} + h_i H_{l,i,j} + u_{i,j})$$

- $v_i V_{l,i,j}$: vertical stability; # of l's phylogenetic neighbors taking value j
- $h_i H_{l,i,j}$: horizontal diffusibility; # of l's spatial neighbors taking value j
- $u_{i,j}$: universality

[Yamauchi+, COLING2016]
A Bayesian extension will appear in the Journal of Language Evolution

Combined Model:
Diachronic model works in the latent space

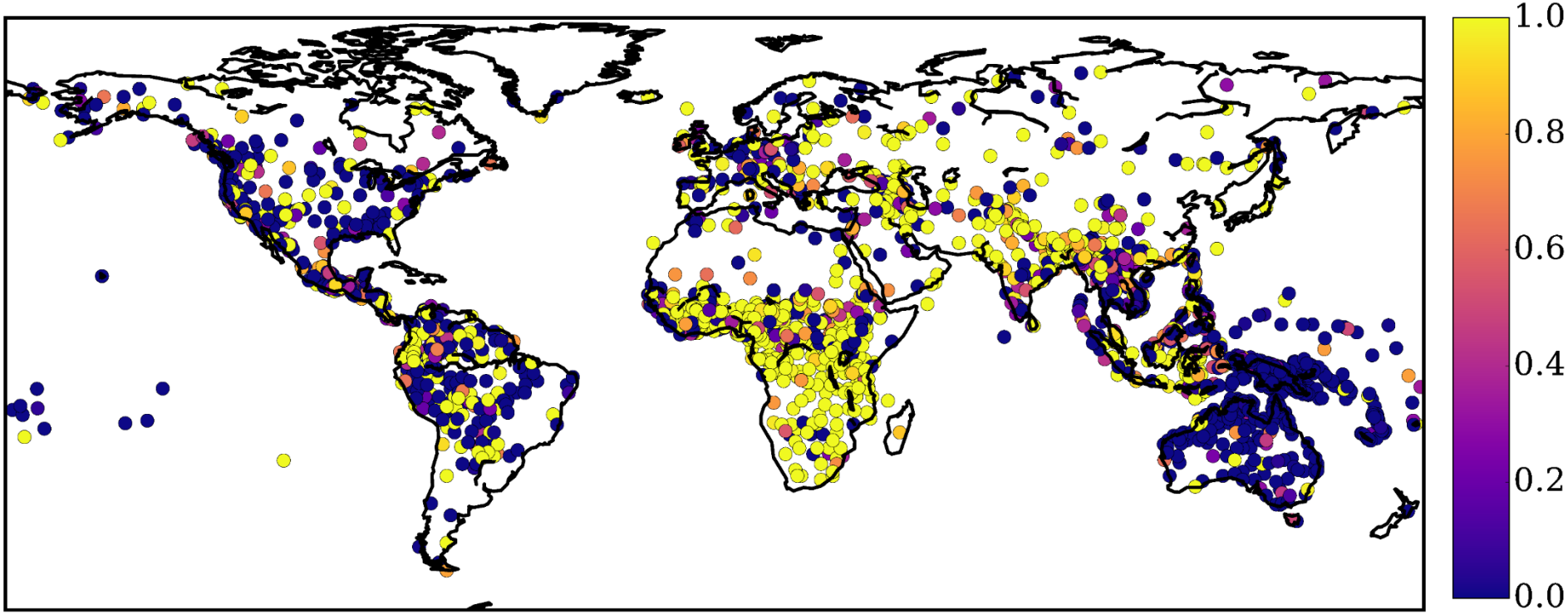


$$P(A, Z, W, X) = P(A)P(Z|A)P(W)P(X|Z, W)$$

Diachronic model

Synchronic model

Geographic Distribution of an Induced Parameter, Revised



Some parameters appear to preserve geographic signals observed in surface features

Conclusions

- Proposed a Bayesian model that combines synchronic and diachronic clues
- Induced a binary latent representation that appears to preserve geographic signals
- Remaining question: Can the latent representation help uncovering the deep history of languages?