Exploring Correlated Evolution with Latent Representations

MURAWAKI Yūgo
Kyoto University

The slides are available at http://bit.ly/2tTF328

Talk @ University of Zurich 20 Mar 2018

My visit to UZH is funded by Kyoto University Design School

DESIGN SCHOOL
KYOTO UNIVERSITY

ABOUT 概要 | EDUCATION 教育プログラム | ADMISSION 入試情報 | EVENT イベント | FACILITY 施設 | PEOPLE 教職員・学生紹介 | OUTCOME 成果

LATEST NEWS

- "Judgement and Strategy" "Videography of Alternative Entrep... update : 2018.3.5
- 合格者発表（平成30年度本科生） update : 2018.3.5
- 情報学領域概論科目「情報通信技... update : 2018.3.1
- Spring Design School 2018 update : 2018.2.20
- 京都大学－琉球大学合同デザイン... update : 2018.2.13
- >>> MORE NEWS

FAQ Access Contact

Google カスタム検索

FOR STUDENT FOR FACULTY

Important Information / 重要なお知らせ

Admission 入試情報

- **2018/03/05** 平成30年度プログラム履修者（本科） 合格発表/ Successful Candidates for Regular Course

Kyoto University, Japan

MURAWAKI Yūgo 村脇 有吾

Background in computer science / engineering

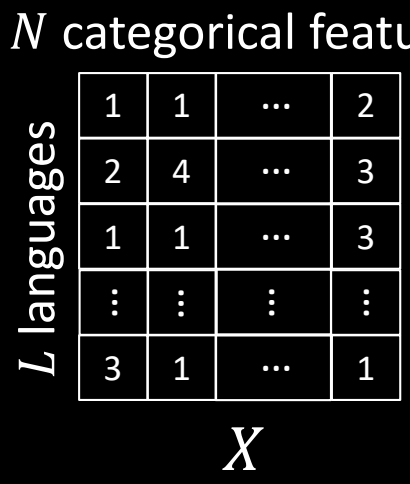
- Assistant professor at Language Media Lab.
  - Our lab is the direct successor to Wired Telecommunications Lab. of the 1950s
  - Teach undergrads in the School of Electrical and Electronic Engineering
- Devote >80% of research time in natural language processing
  - Japanese zero anaphora resolution, shallow discourse parsing, neural network-based text generation, etc.
  - Conference publication preferred to journal publication
- Bayesian since 2011
  - Topic models, nonparametric Bayesian models for unsupervised word segmentation, etc.
- No formal training as a linguist

Computational Approaches to Linguistic Questions

- Relatively new to this field
  - Started in 2014 when I was an assistant professor in Kyushu University
  - Continue as a 20% project (actually much less) after moving back to Kyoto University in 2016
- Focus on statistical modeling
  - Bayesian generative models can uncover complex latent structure of data
  - “Data beggar” relying on publically available data
    - WALS, Glottolog, etc.
- Interested in collaboration with linguists

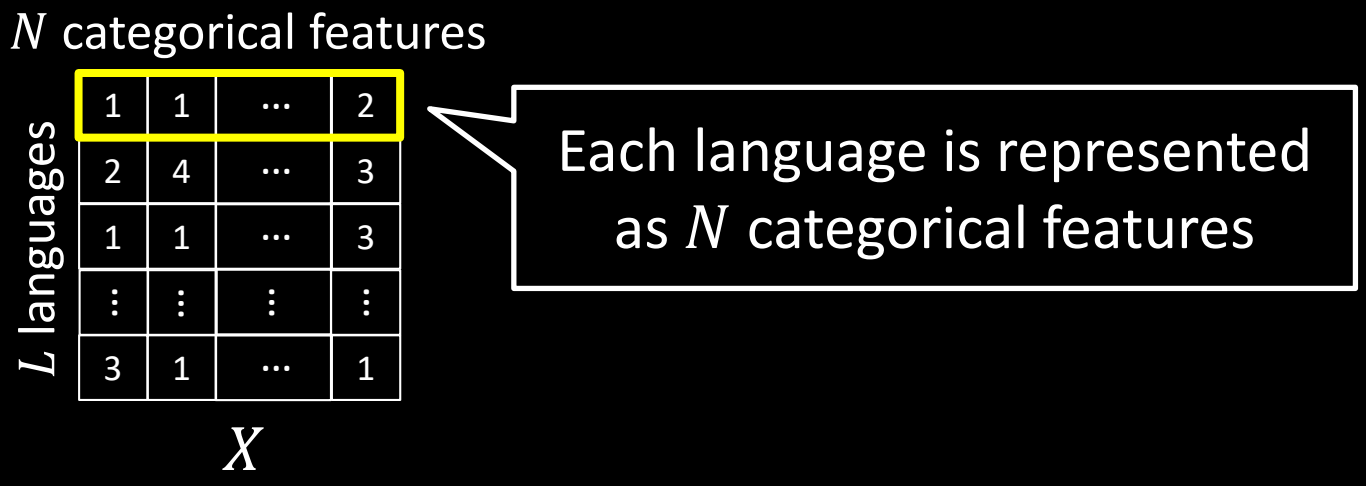
Data Used in This Talk

- World Atlas of Language Structures (WALS)
  - Matrix X with $L = 2{,}679$, $N = 104$ after preprocessing



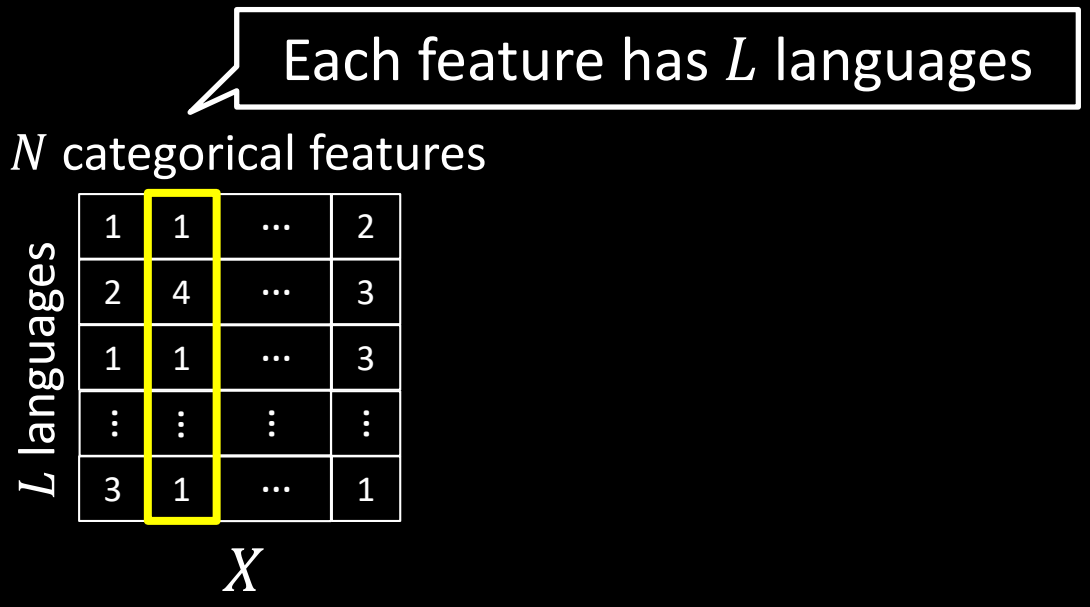
Data Used in This Talk

- World Atlas of Language Structures (WALS)
  - Matrix X with $L = 2{,}679$, $N = 104$ after preprocessing



Data Used in This Talk

- World Atlas of Language Structures (WALS)
  - Matrix X with $L = 2{,}679$, $N = 104$ after preprocessing



Example: Feature 81A
Order of Subject, Object, Verb

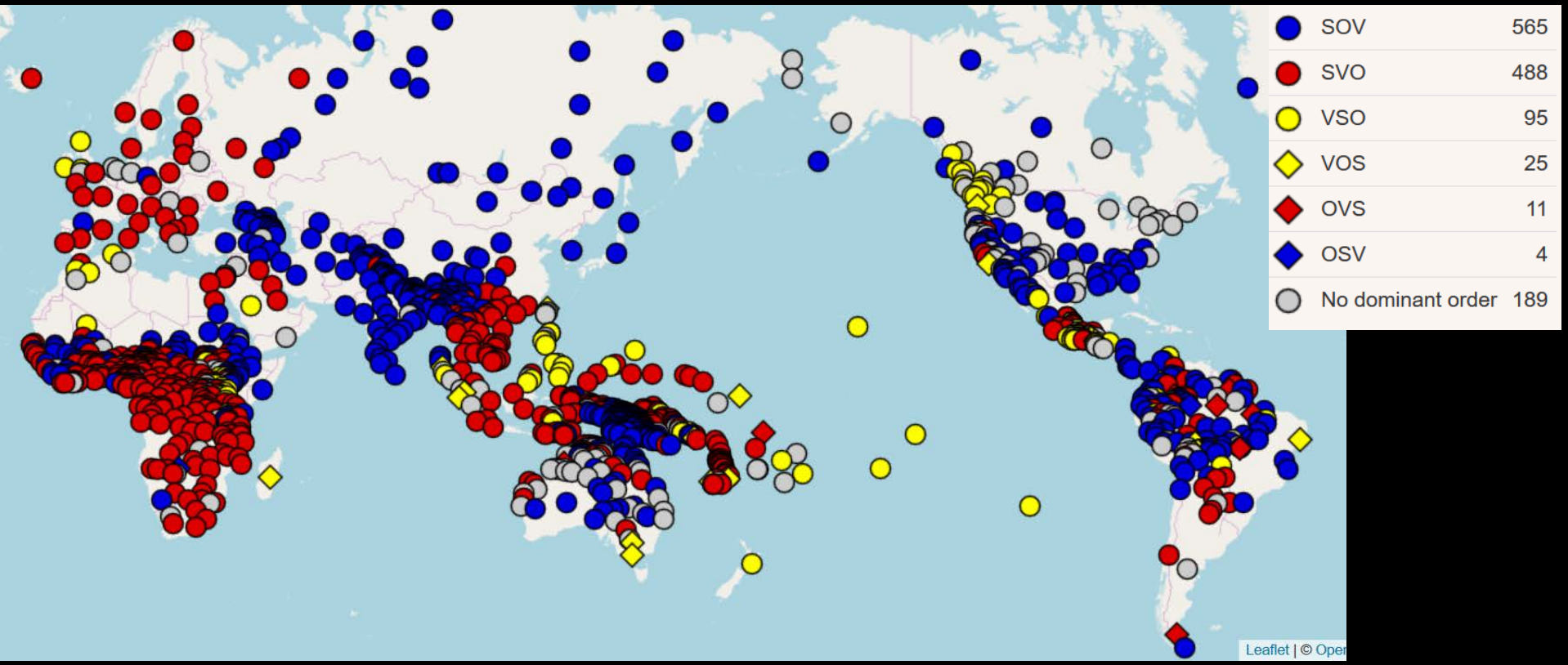
Hereafter referred to as BWO (basic word order) [Dryer, 2005]

1 SOV	Japanese	*John ga*	*tegami o*	*yon-da*
		S	O	V
2 SVO	English	*The dog*	*chased*	*the cat*
		S	V	O
3 VSO	Irish	*Léann*	*na sagairt*	*na leabhair*
		V	S	O

The priests are reading the books.

Example: Feature 81A
Order of Subject, Object, Verb

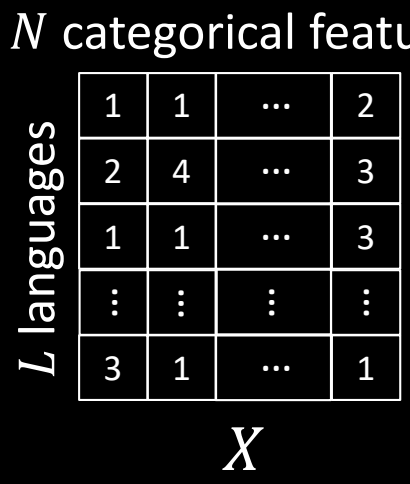
Hereafter referred to as BWO (basic word order) [Dryer, 2005]



Source: http://wals.info/feature/81A

Data Used in This Talk

- World Atlas of Language Structures (WALS)
  - Matrix X with $L = 2{,}679$, $N = 104$ after preprocessing



Data Used in This Talk

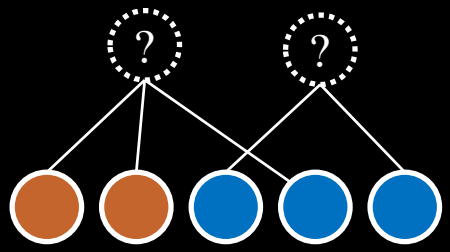
- World Atlas of Language Structures (WALS)
  - Matrix X with $L = 2{,}679$, $N = 104$ after preprocessing
  - Make use of two additional resources for part 1

N categorical features

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

L languages

X



Phylogenetic groupings of languages

Spatial locations of languages

Data Used in This Talk

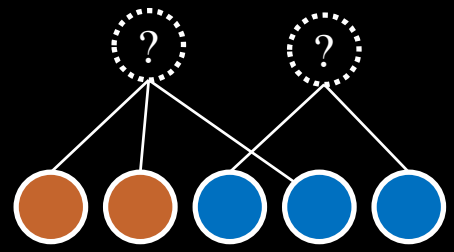
- World Atlas of Language Structures (WALS)
  - Matrix X with $L = 2{,}679$, $N = 104$ after preprocessing
  - Make use of two additional resources for part 1
- Glottolog for part 2

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

Question: How to Model Correlated Evolution

- Implicational universals [Greenberg, 1963]
  - Order of object and verb & order of noun and relative clause
    - If VO, then NRel
    - If RelN, then OV

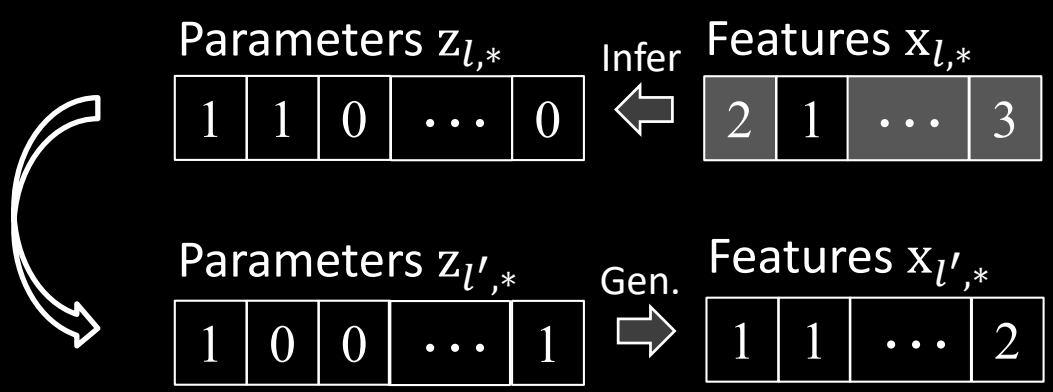
	NRel	RelN
VO	✓	×
OV	✓	✓

- A BWO change must have a profound impact on the whole grammatical system
  - English: Shift to SVO appears to correlate with the move from synthetic to analytic
  - North American langs: Rigidity of SOV reduced with morphological innovations [Mithun, 1995]

Key Idea: Latent Representations

- Reconfigure $\mathrm{N} = 104$ categorical surface features into $\mathrm{K} = 100$ binary latent parameters
  - Parameters are independent by assumption
- Inference in the latent space implicitly captures correlated evolution

Do some inference in the latent space



Outline of This Talk

Part 1: How to induce latent representations from surface features

Part 2: How to use the latent representations to analyze diachronic changes

Part 1

Diachrony-aware Induction of Binary Latent Representations from Typological Features

[Murawaki, IJCNLP2017]

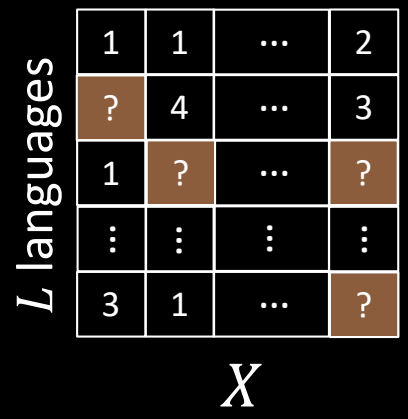
The code available at https://github.com/murawaki/latent-typology/

Missing Values are a Real Problem

also used as an indicator of the quality of latent representations

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

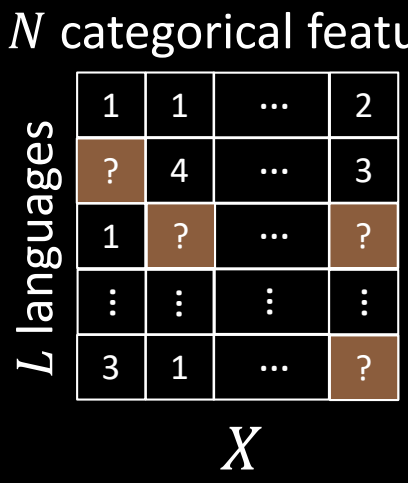
N categorical features



Missing Values are a Real Problem

also used as an indicator of the quality of latent representations

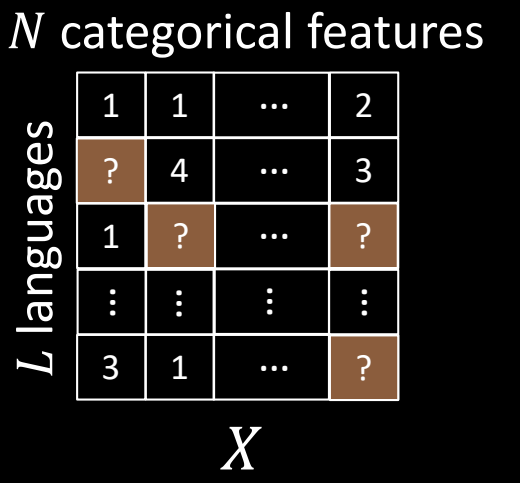
- 73.1% of items are missing
- My model imputes missing values by combining
  1. Synchronic clues (inter-feature dependencies)
  2. Diachronic clues (inter-language dependencies)

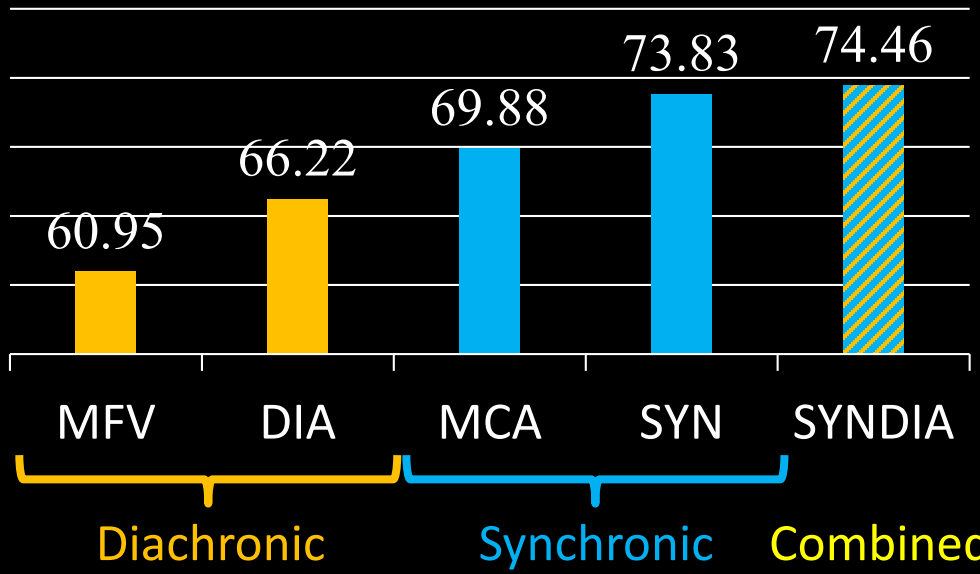


Missing Values are a Real Problem

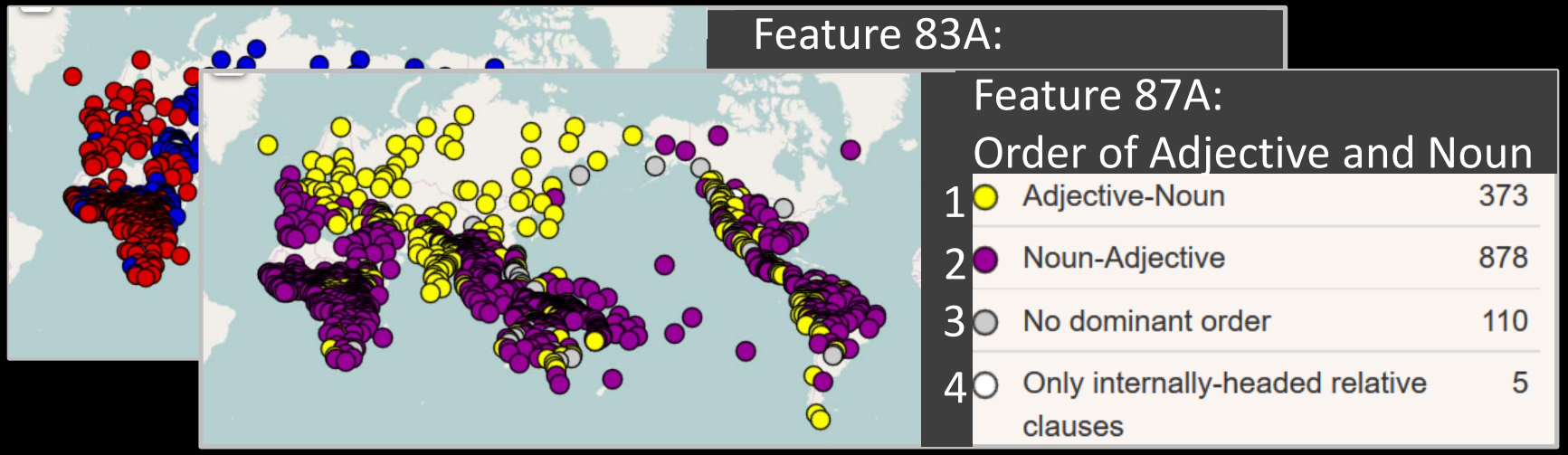
also used as an indicator of the quality of latent representations

- 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

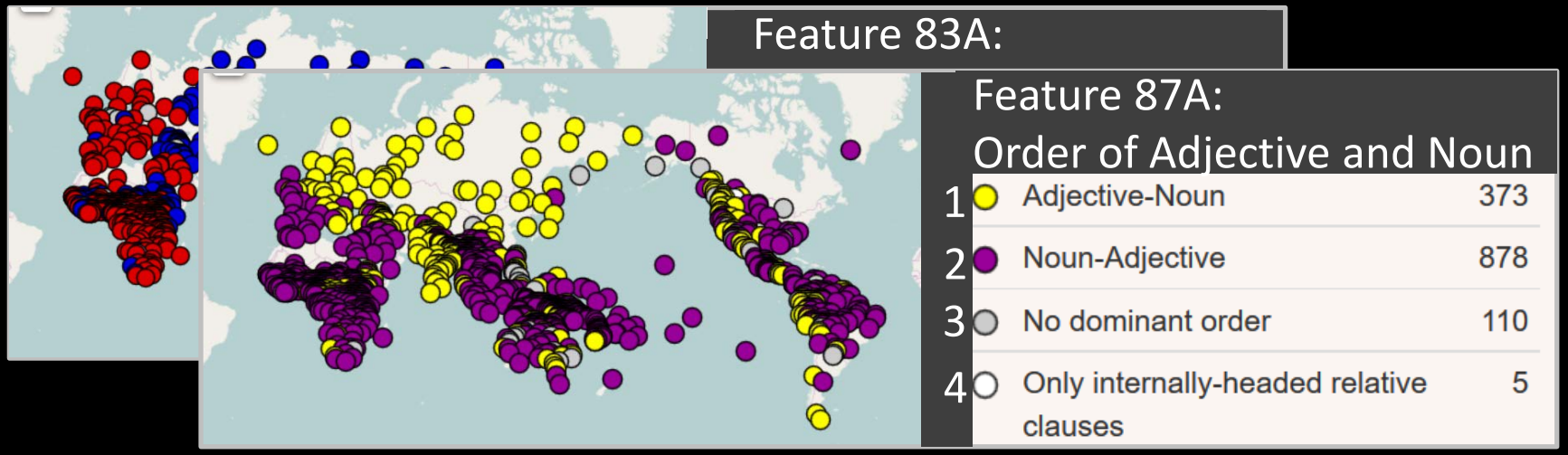
L languages

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

X

[Greenberg, 1963]

Synchronic Clues: Inter-feature Dependencies



N categorical features

L languages

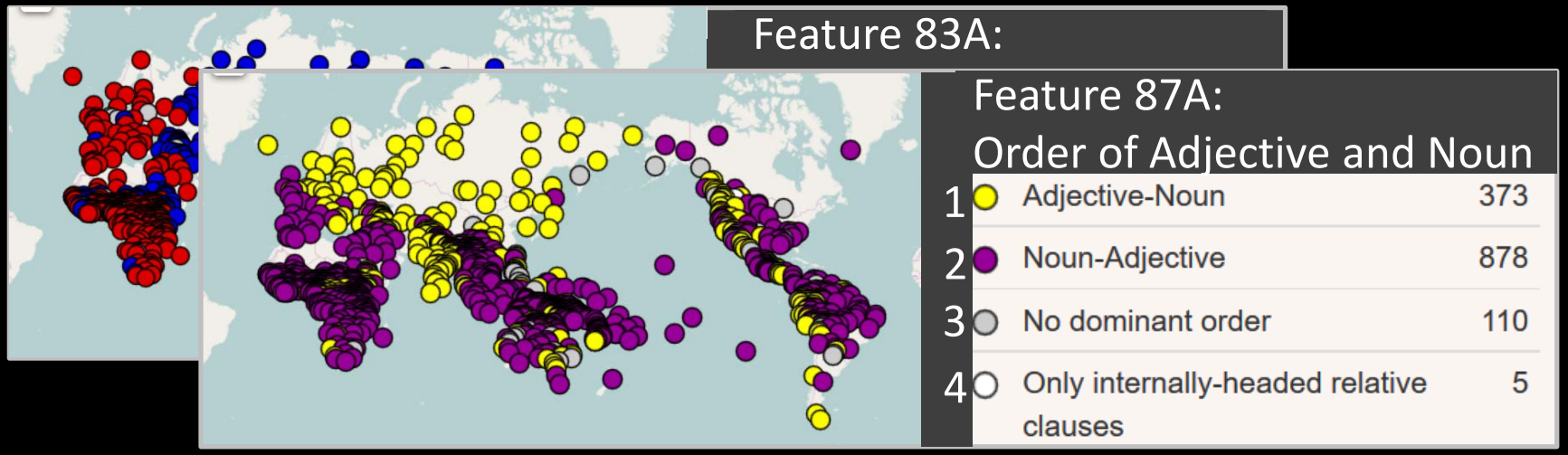
1	1	…	2
?	4	…	3
1	?	…	?
⋮	⋮	⋮	⋮
3	1	…	?

X

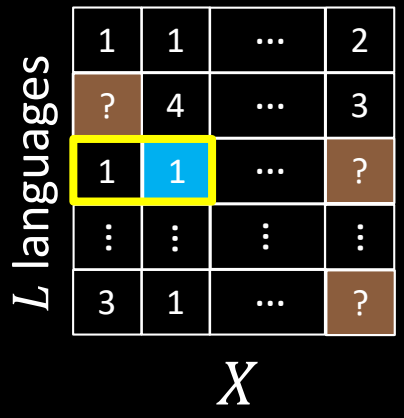
If Feature 83A is 1 (OV), then Feature 87A is likely to be 1 (Adjective-Noun)

[Greenberg, 1963]

Synchronic Clues: Inter-feature Dependencies



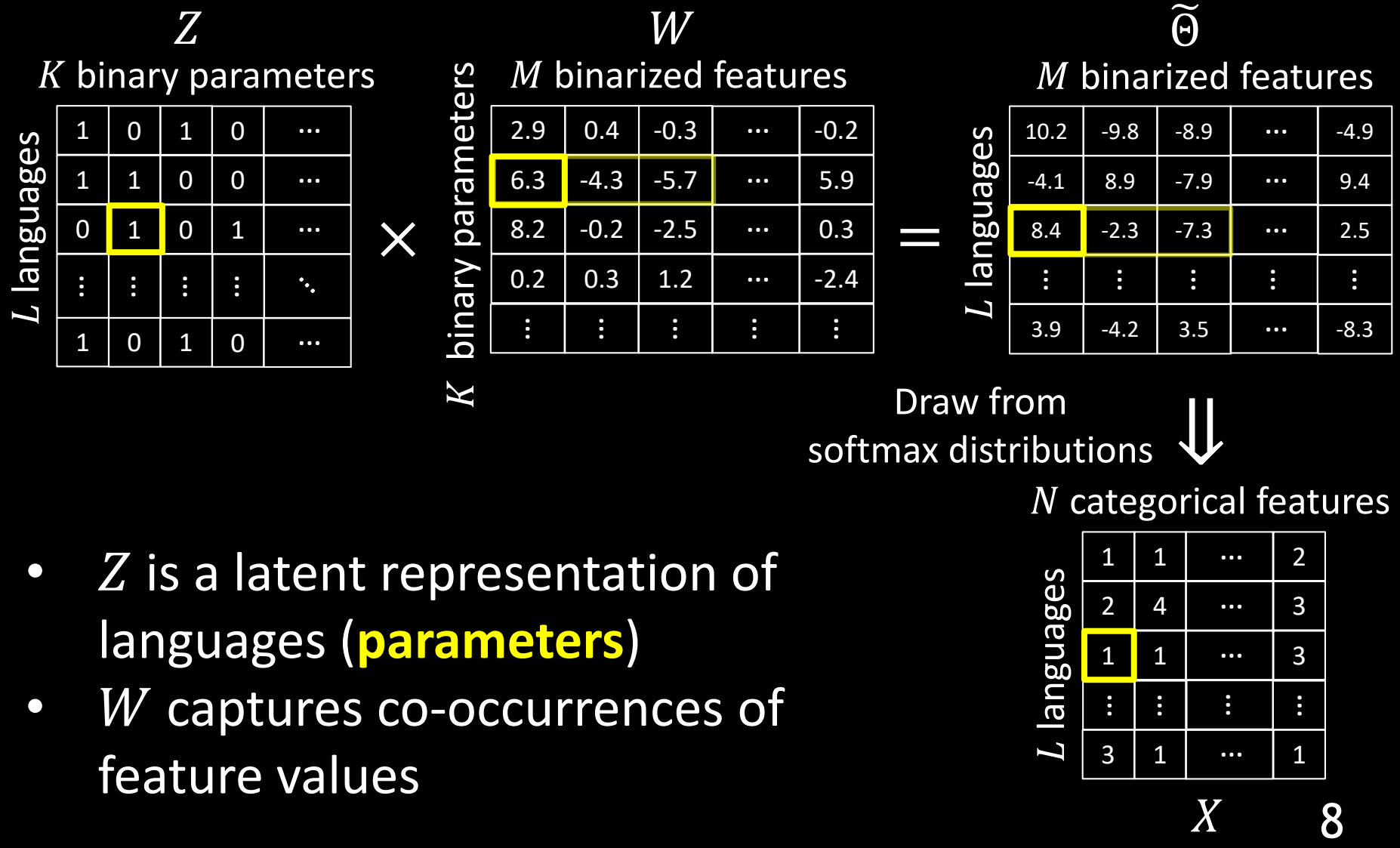
N categorical features



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

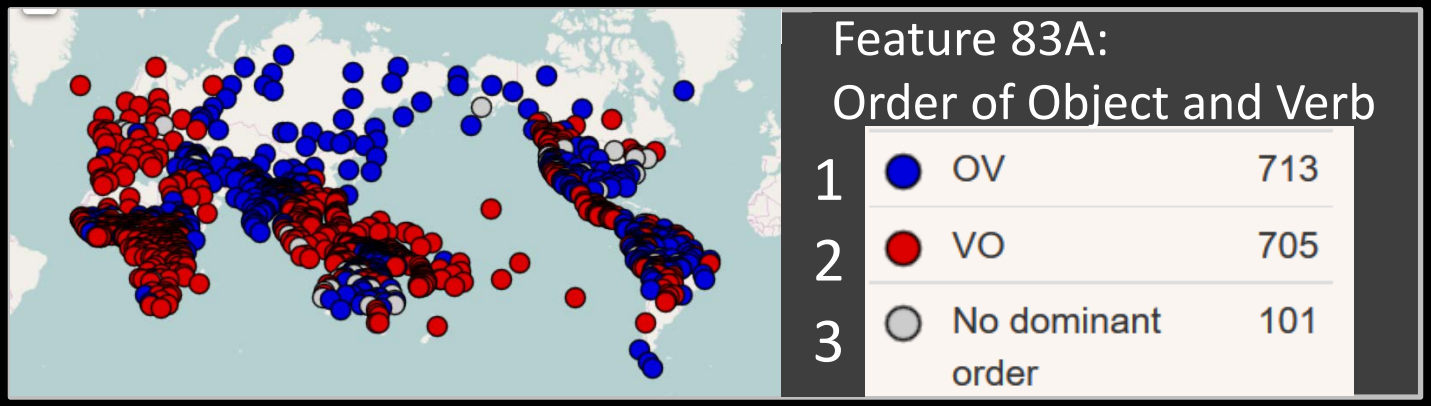
[Greenberg, 1963]

Matrix Decomposition

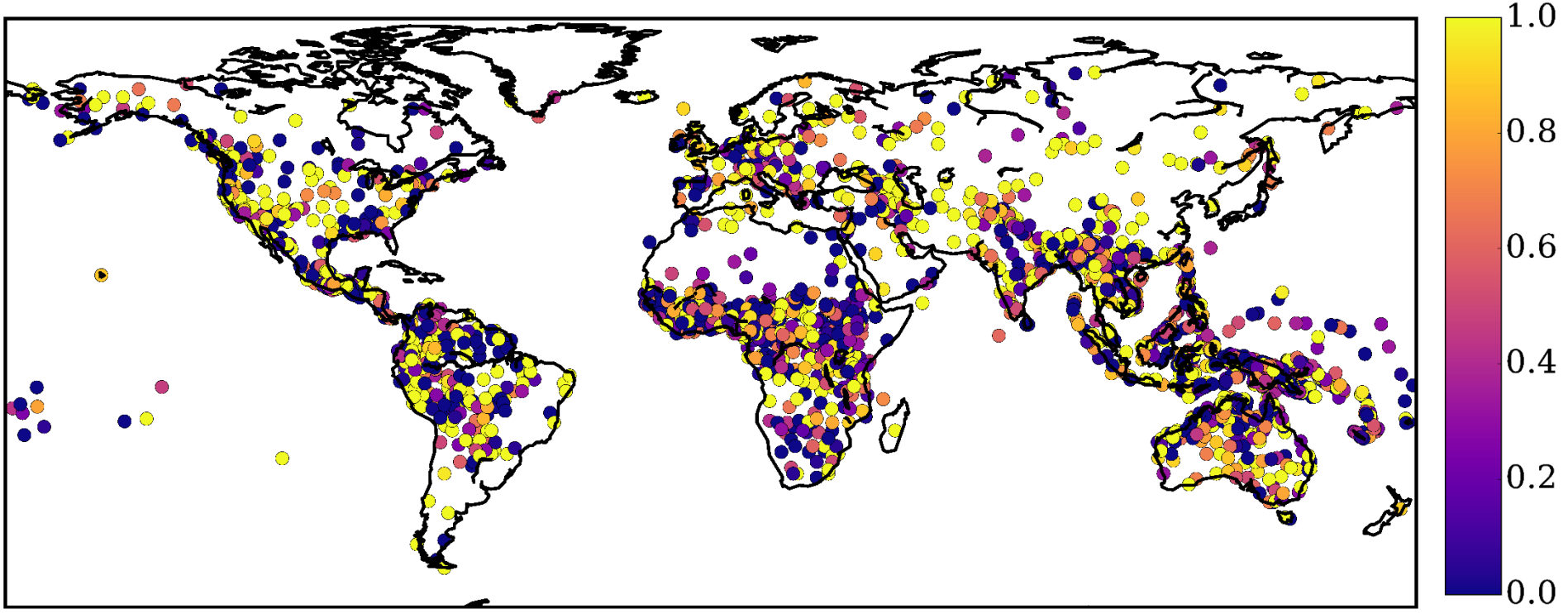


- Z is a latent representation of languages (**parameters**)
- W captures co-occurrences of feature values

Geographic Distribution of an Induced Parameter



Geographic Distribution of an Induced Parameter



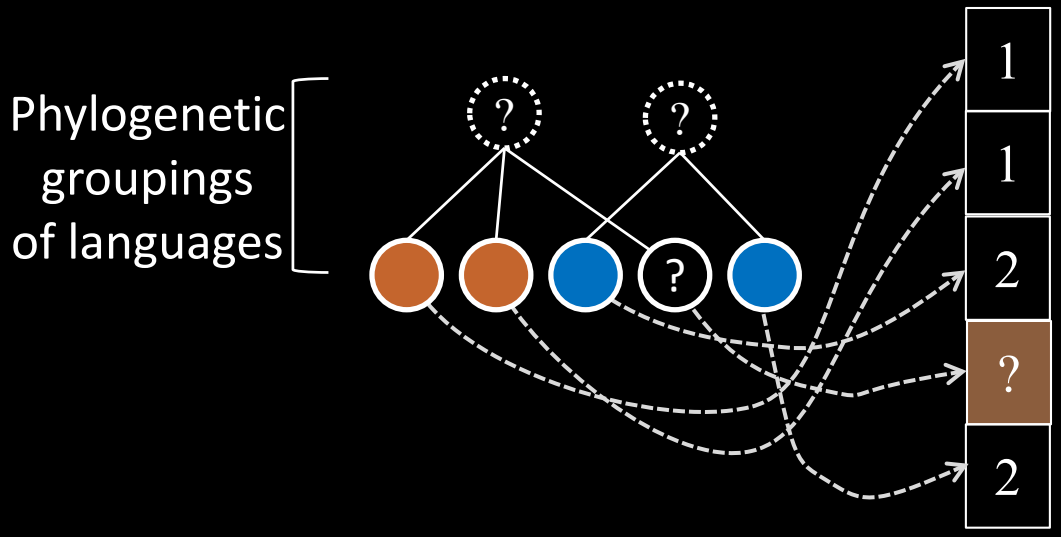
Geographic signals observed in surface features are **somehow lost** during the induction

Diachronic Clues: Inter-language Dependencies

A ***column*** of the matrix

Languages may share the same value because they …

Diachronic Clues: Inter-language Dependencies

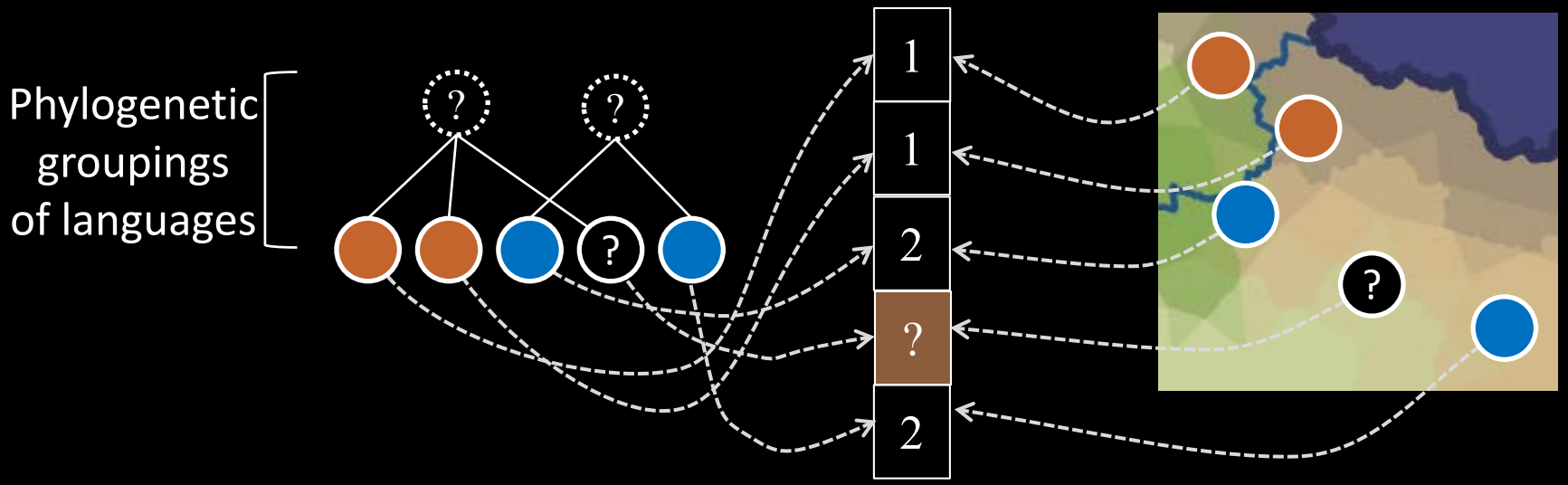


A ***column*** of the matrix

Languages may share the same value because they …

1. are phylogenetically related,

Diachronic Clues: Inter-language Dependencies

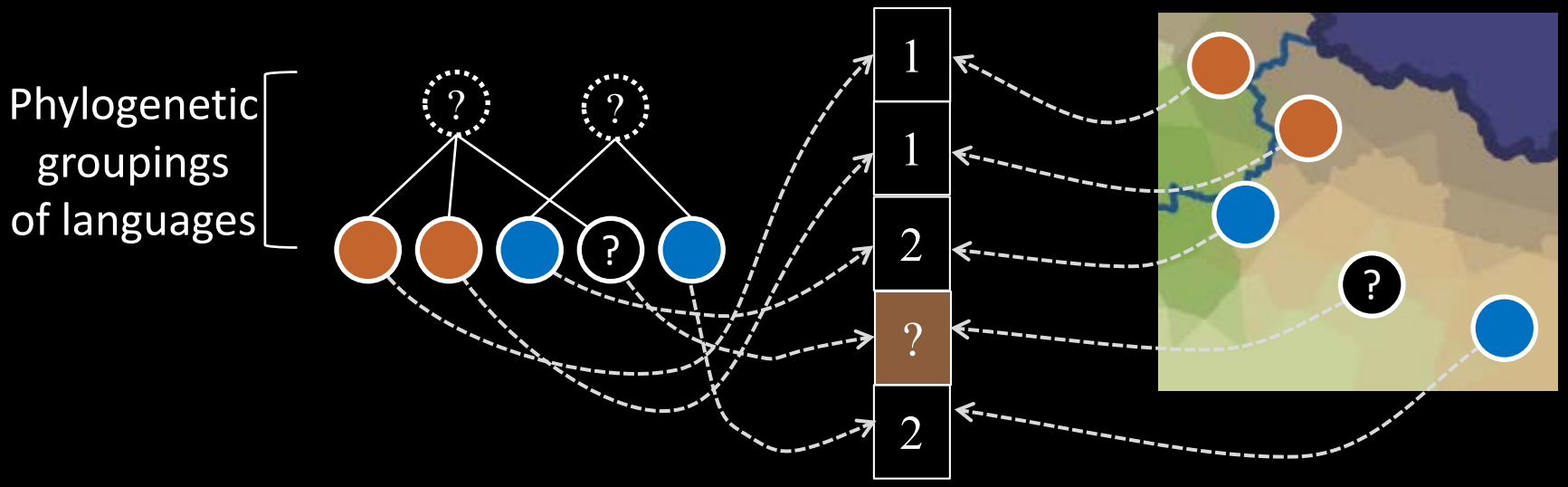


A ***column*** of the matrix

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

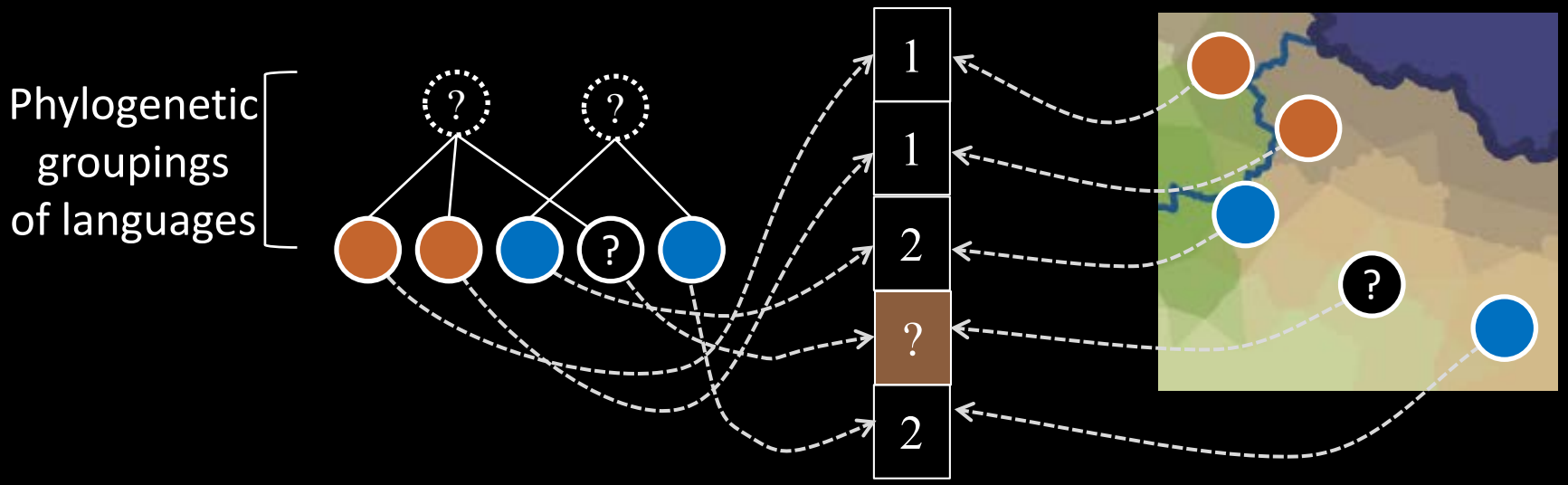


A ***column*** of the matrix

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

[Murawaki+, JoLE 2018]

Phylogenetic groupings of languages

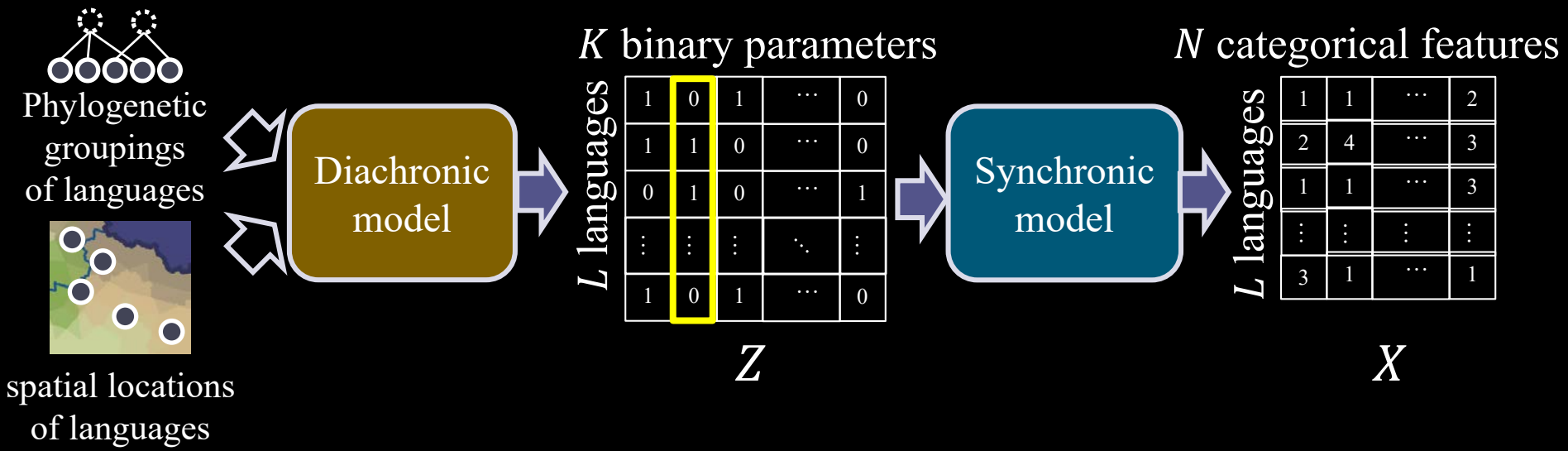
spatial locations of languages

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

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

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

Combined Model:
Diachronic model works in the latent space



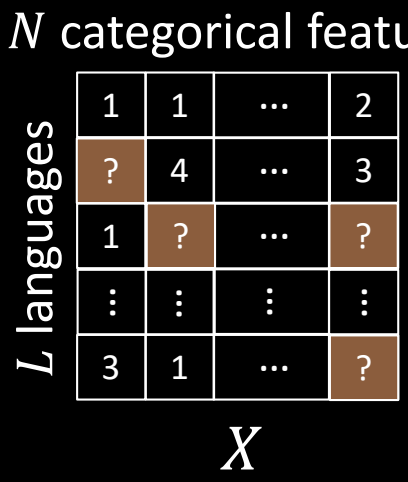
$$P(A, Z, W, X) = P(A)P(Z|A)P(W)P(X|Z, W)$$

Diachronic model

Synchronic model

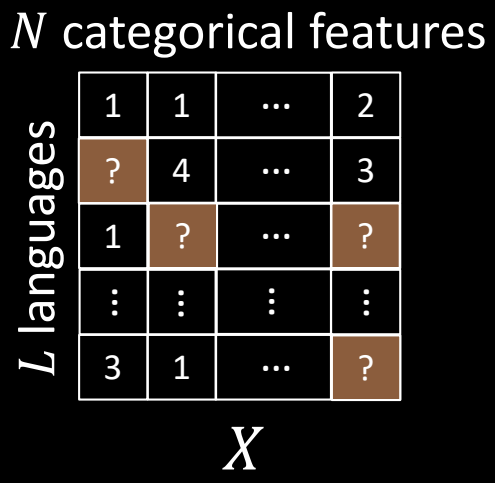
Missing Values Imputation

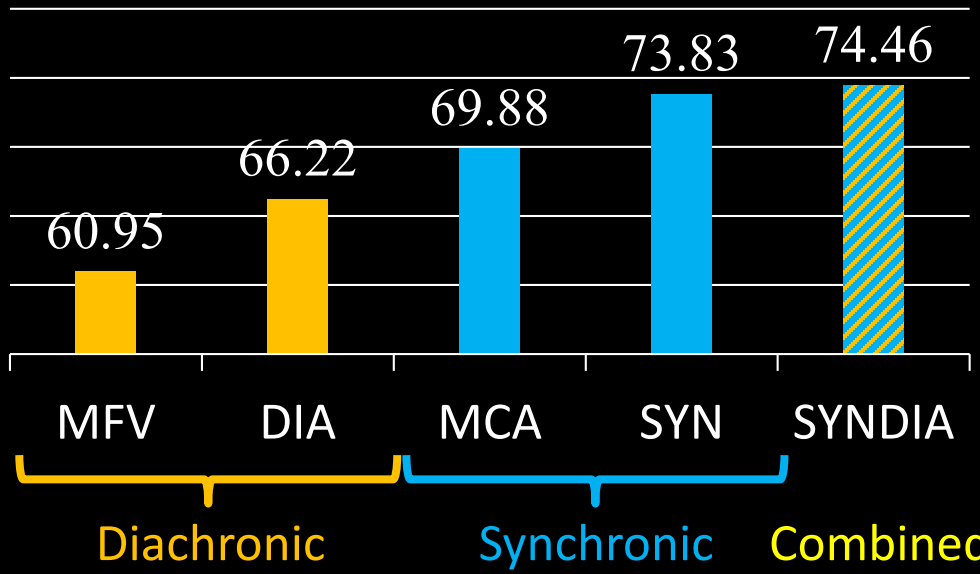
- Recap: 73.1% of items are missing
- Treat $\mathbf{10}\%$ of observed items as missing to see how well the model recovers them
- High accuracy $\doteqdot$ Good quality of latent representations



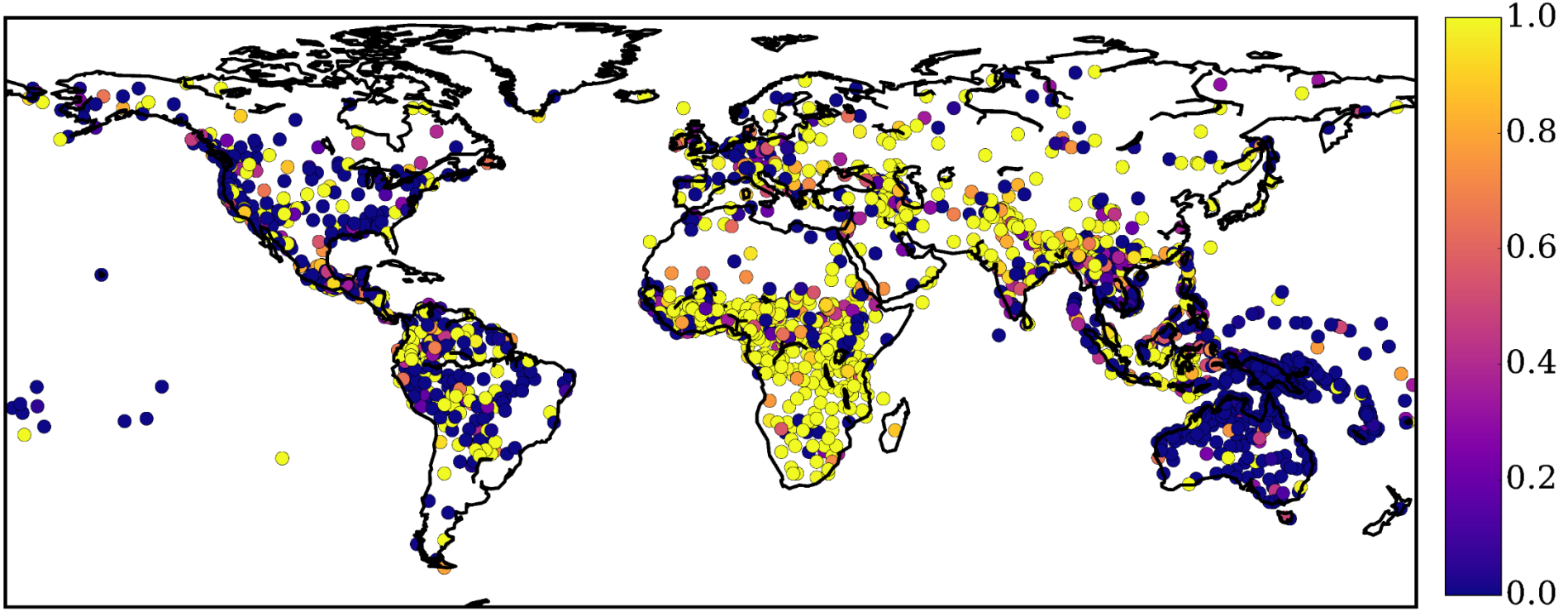
Missing Values Imputation

- Recap: 73.1% of items are missing
- Treat **10**% of observed items as missing to see how well the model recovers them
- High accuracy ≒ Good quality of latent representations



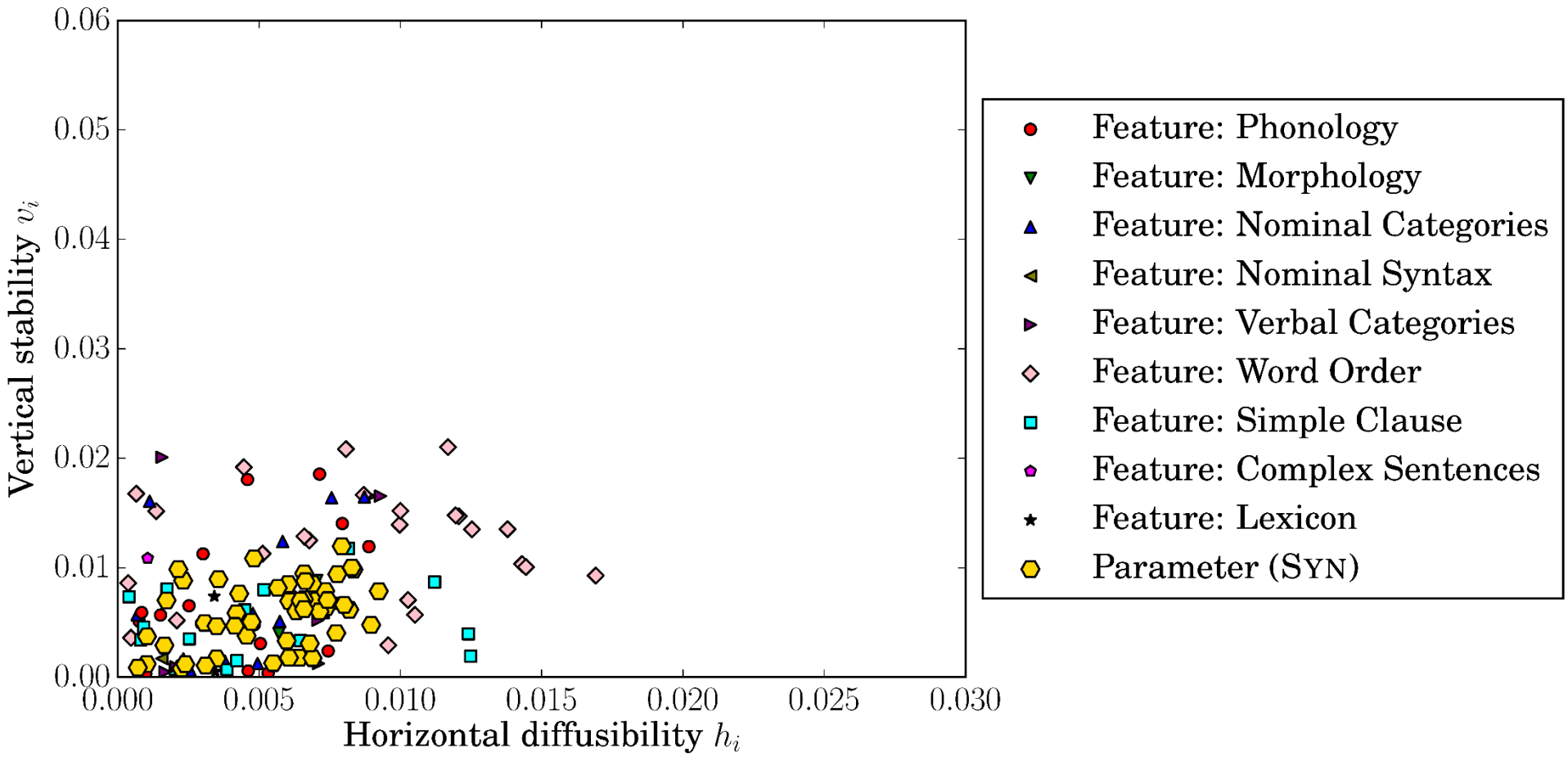


Geographic Distribution of an Induced Parameter, Revised

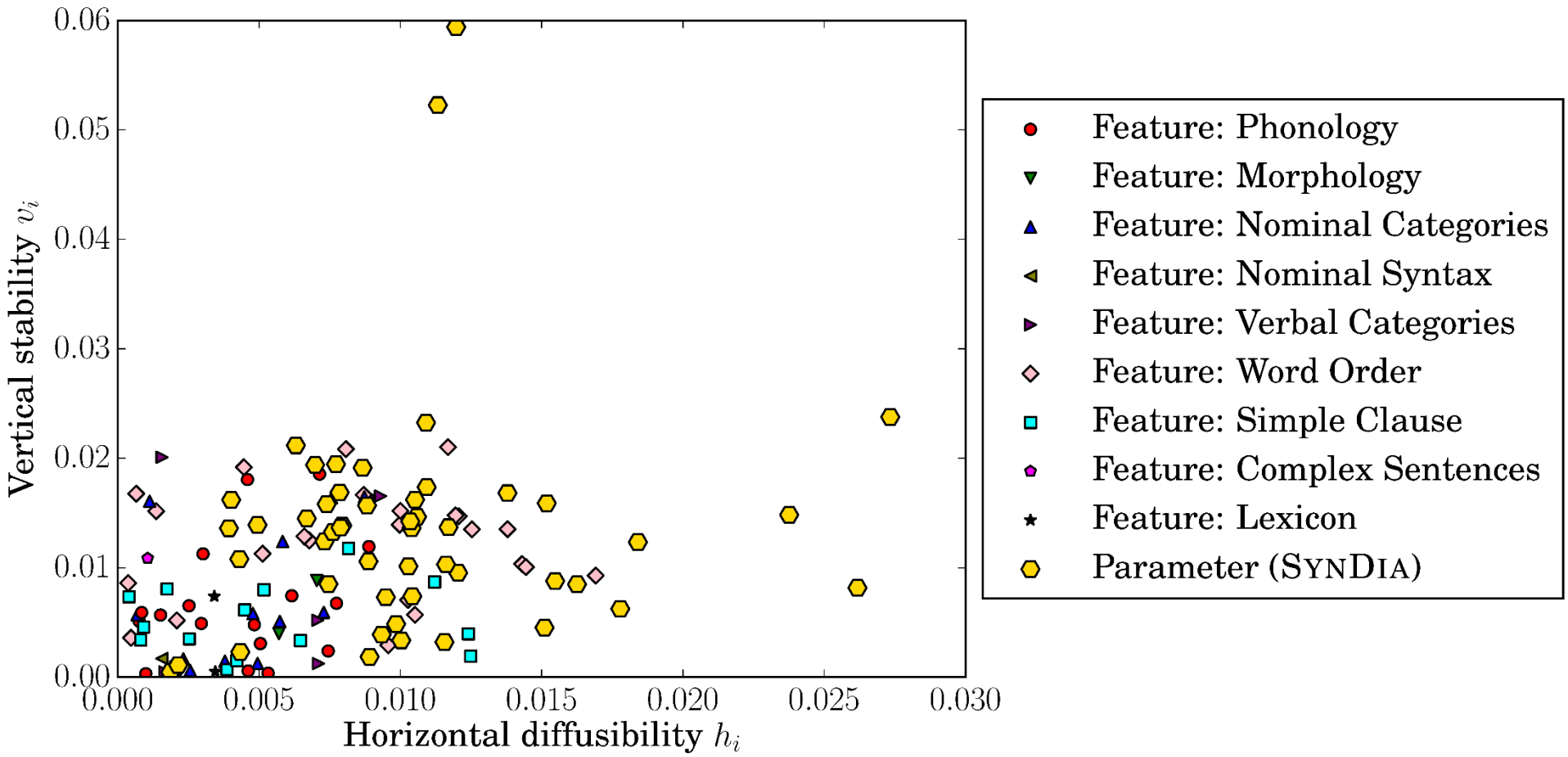


Some parameters appear to preserve geographic signals observed in surface features

Estimated Vertical Stability and Horizontal Diffusibility (SYN)



Estimated Vertical Stability and Horizontal Diffusibility (SYNDIA)



Summary of Part 1

- Proposed a Bayesian model that combines synchronic and diachronic clues
- Missing value imputation indicates a good quality of the binary latent parameters
- Some latent parameters appear to preserve geographic signals of surface features

Part 2

Latent Representation-based Analysis of Diachronic Typology

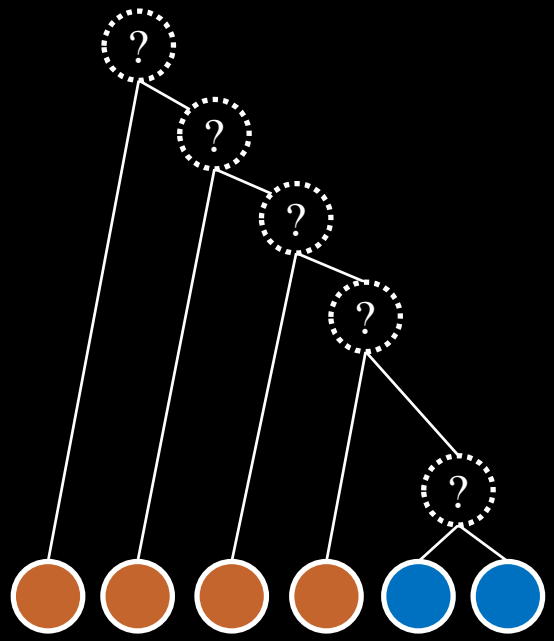
(under review)

Outline of Part 2

- Apply latent representations to diachronic inference
  - Uniformitarian hypothesis
- Directly model parent-to-child transitions
  - In part 1, we modeled both vertical and horizontal transmission but did not estimate ancestral states
  - In part 2, we estimate ancestral states but do not directly model horizontal transmission

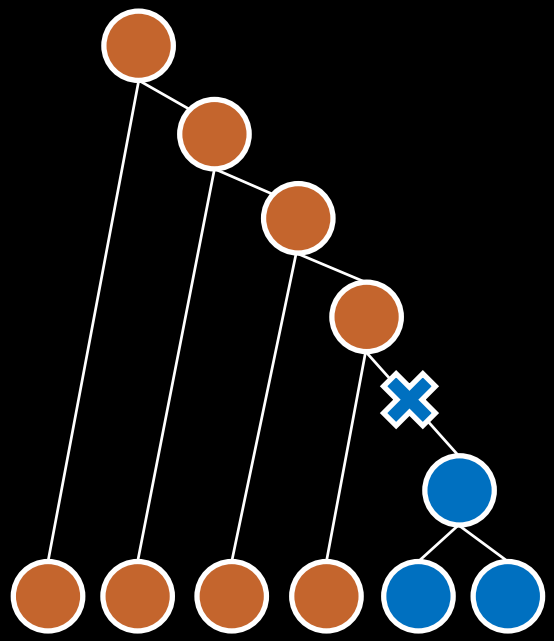
Phylogenetic Comparative Methods

- Assumption: phylogenetic trees are known (primarily based on lexical evidence)
- Able to infer ancestral states, with varying degrees of confidence



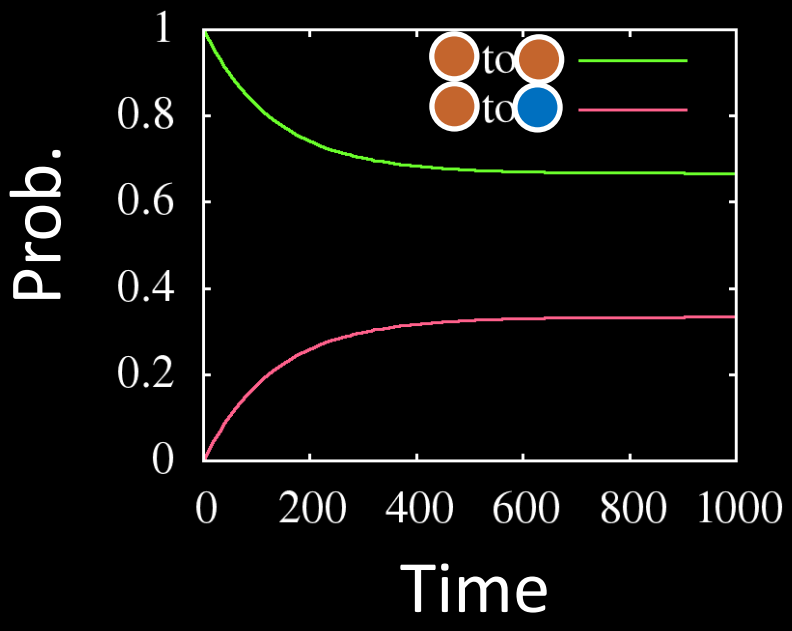
Phylogenetic Comparative Methods

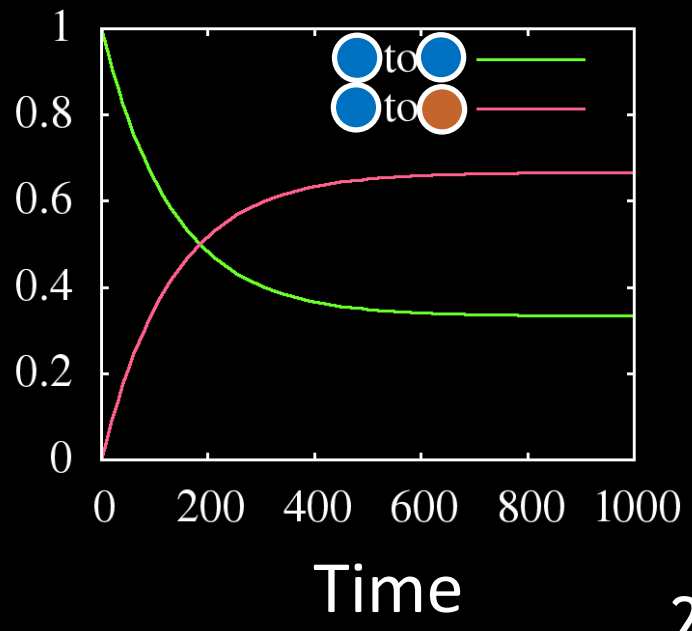
- Assumption: phylogenetic trees are known (primarily based on lexical evidence)
- Able to infer ancestral states, with varying degrees of confidence



Continuous-time Markov Chain (CTMC)

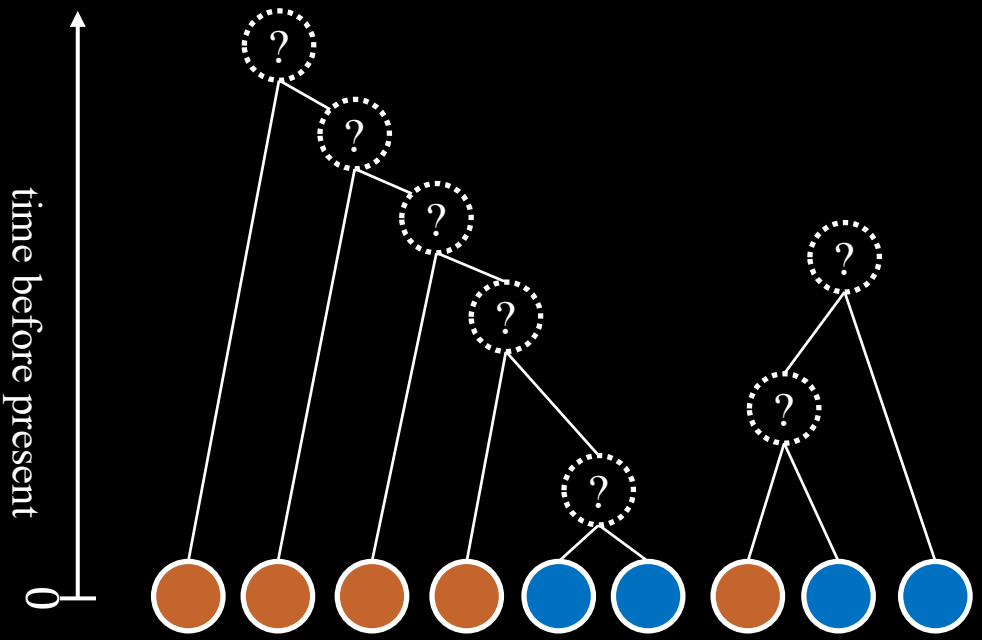
- Prob. of taking value b after time t conditioned on the present value a:
  $\exp(tQ)_{a,b}$,
  where transition rate matrix $Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$





Estimate Transition Rate Matrix (TRM) using (Time-)tree(s)

- Observed data
  - (Time-)tree(s)
  - States of leaf nodes
- Latent data
  - Transition rate matrix
  - States of internal nodes

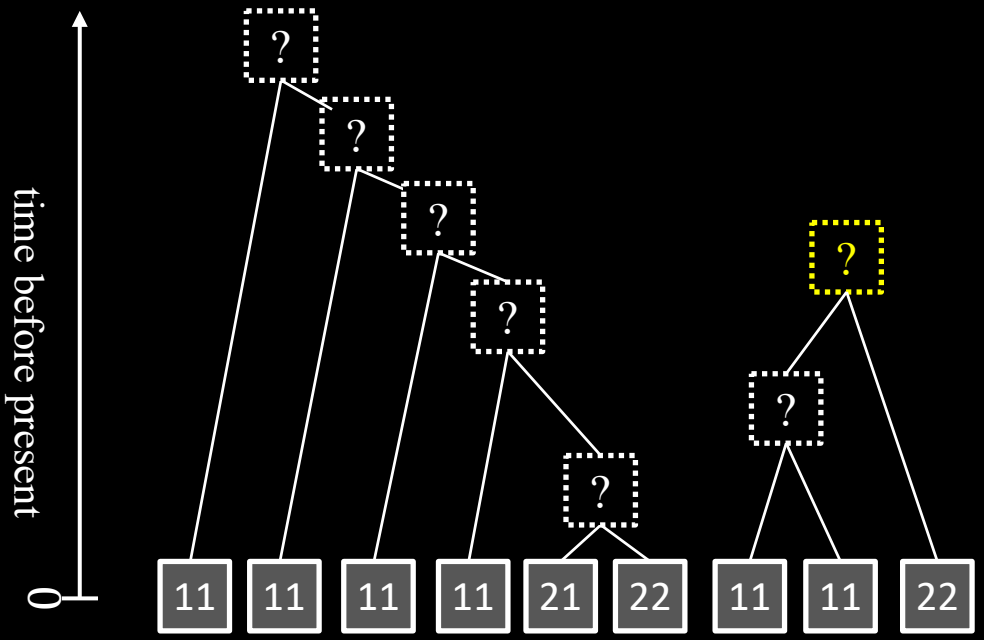


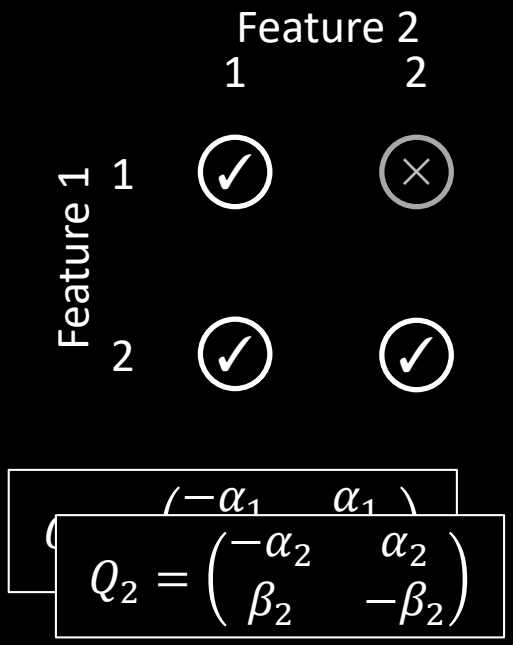
$$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

[Greenhill+, 2010]
[Maurits+, PNAS 2014]

Tying each Feature to a TRM Entails Independence Assumption

- Unable to take into account the observation that feature combination 12 is unnatural

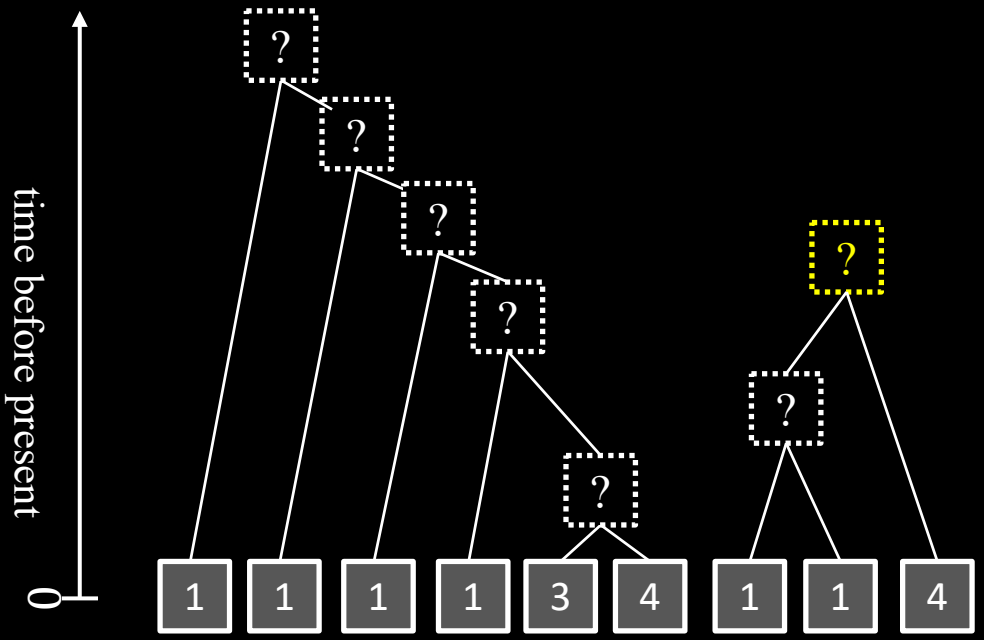


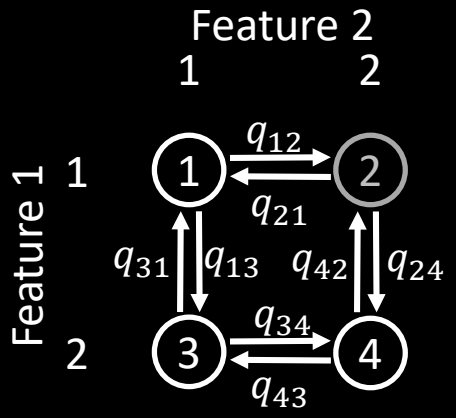


Modeling Feature Dependencies by Expanding Feature Combinations

[Dunn+, Nature 2011]

- Categorical features (BWO has 6-7 values) and interactions between more than two features cause combinatorial explosion



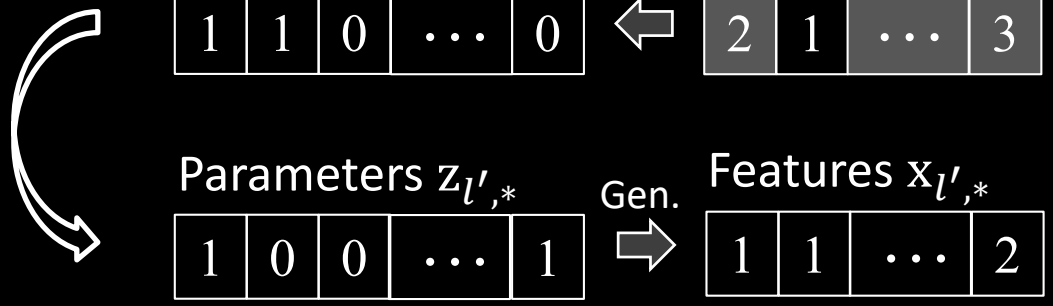


$$Q = \begin{array}{c} \\ 11 \\ 12 \\ 21 \\ 22 \end{array} \begin{array}{c} \begin{array}{cccc} 11 & 12 & 21 & 22 \end{array} \\ \begin{pmatrix} * & q_{12} & q_{13} & - \\ q_{21} & * & - & q_{24} \\ q_{31} & - & * & q_{34} \\ - & q_{42} & q_{43} & * \end{pmatrix} \end{array}$$

Latent Representations (Recap)

- Reconfigure $\mathrm{N} = 104$ categorical surface features into $\mathrm{K} = 100$ binary latent parameters
  - Parameters are independent by assumption
- Inference in the latent space implicitly captures correlated evolution

Phylogenetic inference in the latent space

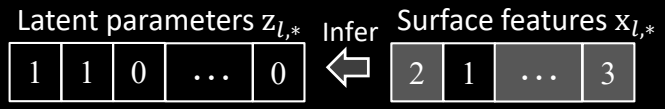


Comparison with Previous Studies

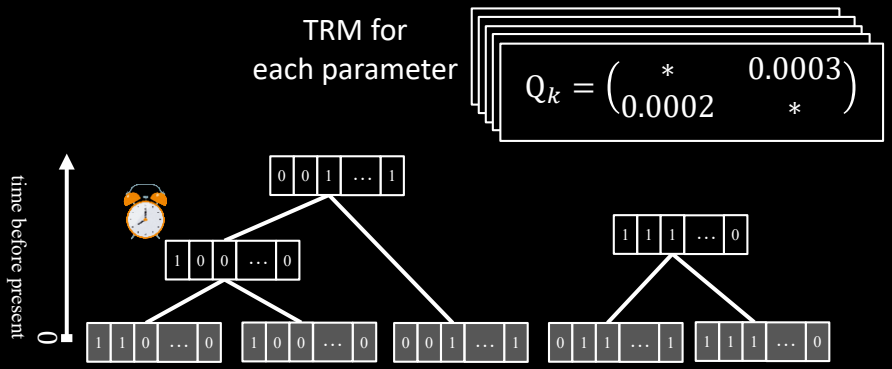
	Dep.	# of families	Tree sources	Dating	Abs.
Dediu (2010)	Single	1	Experts	Yes	No
Greenhill+ (2010)	Single	1	Cognates	No	Yes
Maurits+ (2014)	Single	1 or 7 combined	Other	No	Yes
Dunn+ (2011)	Bin. Pair	1	Cognates	No	Yes
Ours	All	**309** incl. 155 isolates	Experts	Yes	Yes

- Sample diversity problem [Croft, 2011]
- Solution: use Glottolog trees
  - Need to infer the dates of internal nodes
  - 50 calibration points collected from the secondary literature

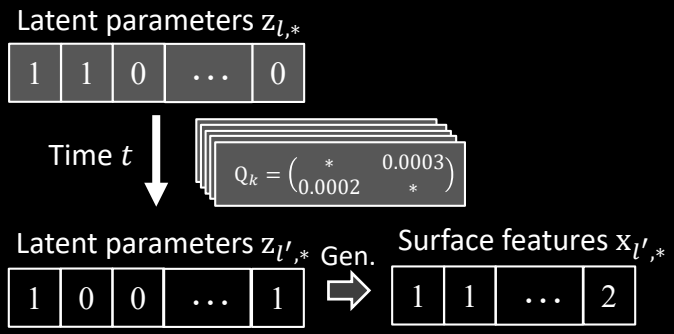
Step 1. Map each language into the latent representations



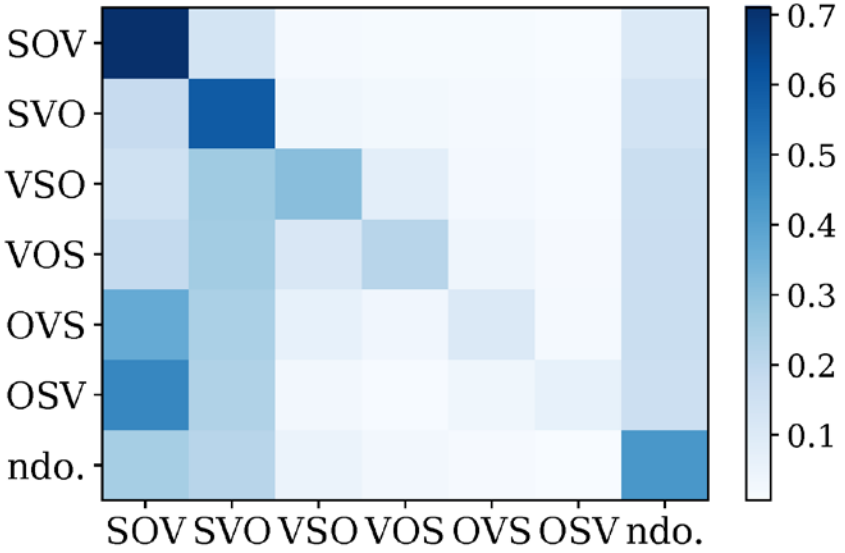
Step 2. Infer a set of TRMs using phylogenetic trees (also infer the states and dates of internal nodes)



Step 3. Simulate language evolution using TRMs



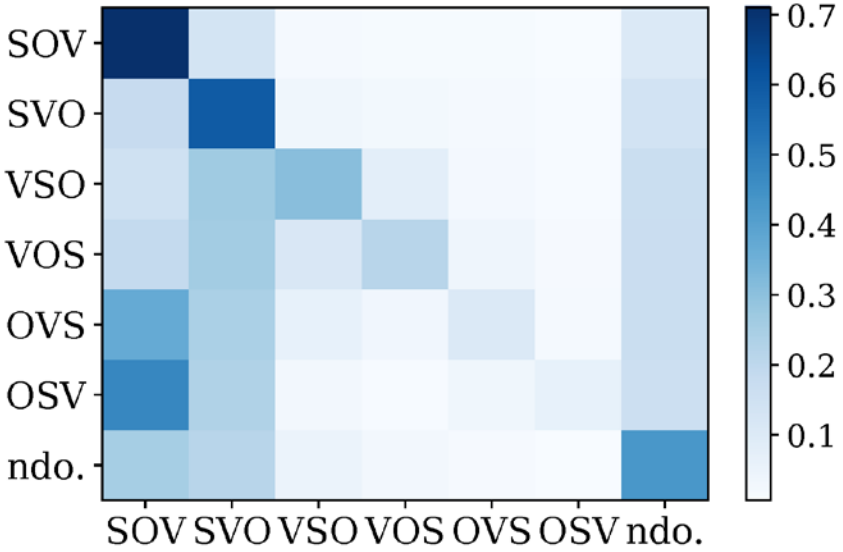
Predicting Future BWO with $t = 2{,}000$



Avg. transition prob.
of BWO

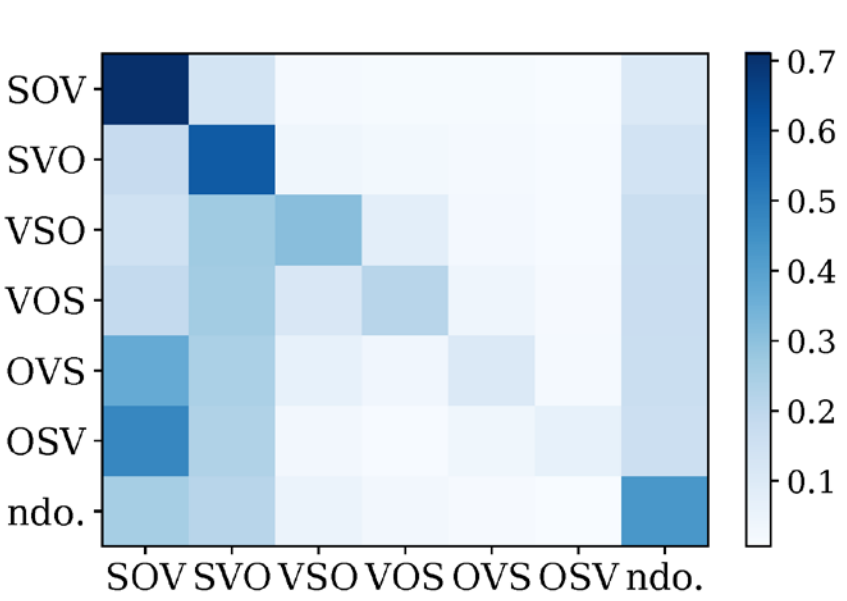
Predicting Future BWO with $t = 2{,}000$

Largely agree with the findings of Maurits+ (PNAS 2014)

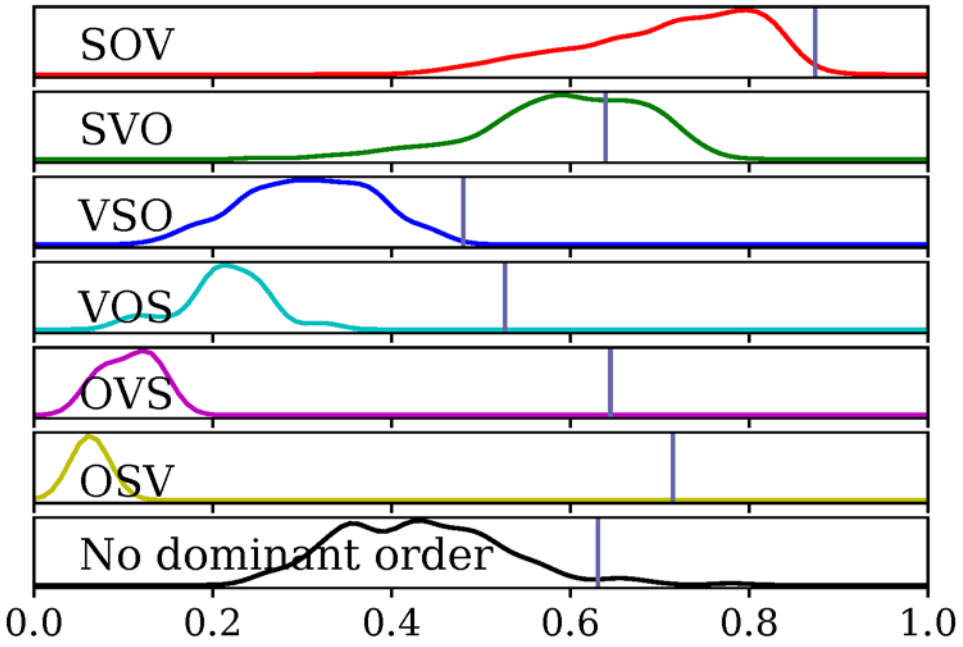


Avg. transition prob. of BWO

Predicting Future BWO with $t = 2{,}000$



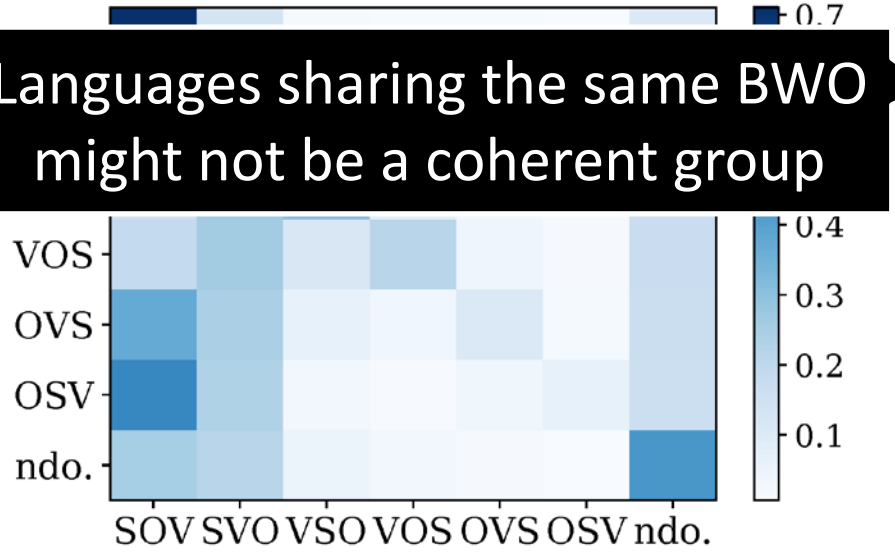
Avg. transition prob. of BWO

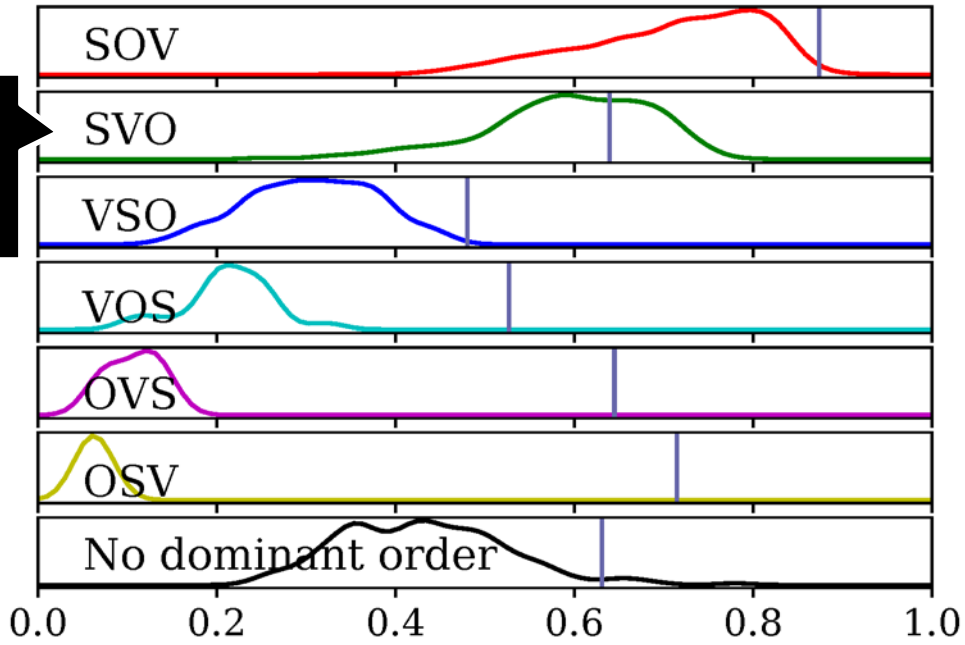


Variability in the prob. of keeping the same BWO

Predicting Future BWO with $t = 2{,}000$

Languages sharing the same BWO might not be a coherent group

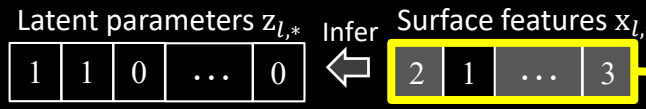




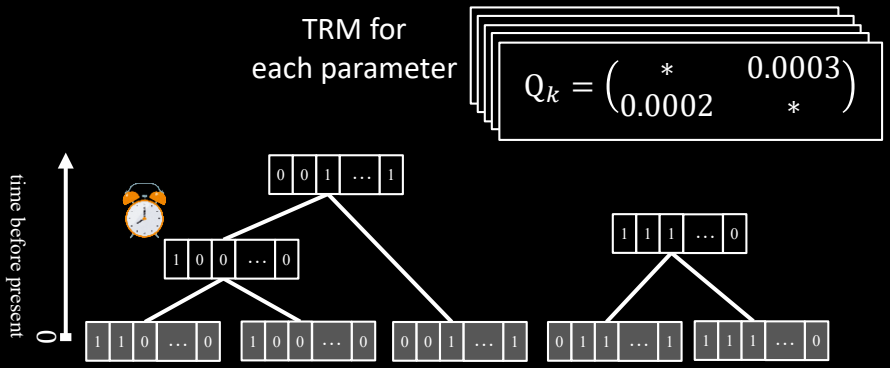
Avg. transition prob. of BWO

Variability in the prob. of keeping the same BWO

Step 1. Map each language into the latent representations

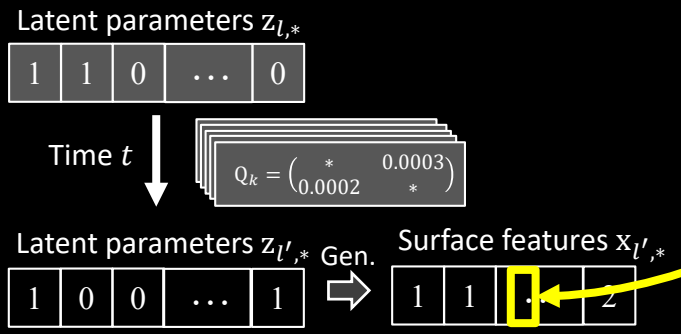


Step 2. Infer a set of TRMs using phylogenetic trees
(also infer the states and dates of internal nodes)



Linear Regression

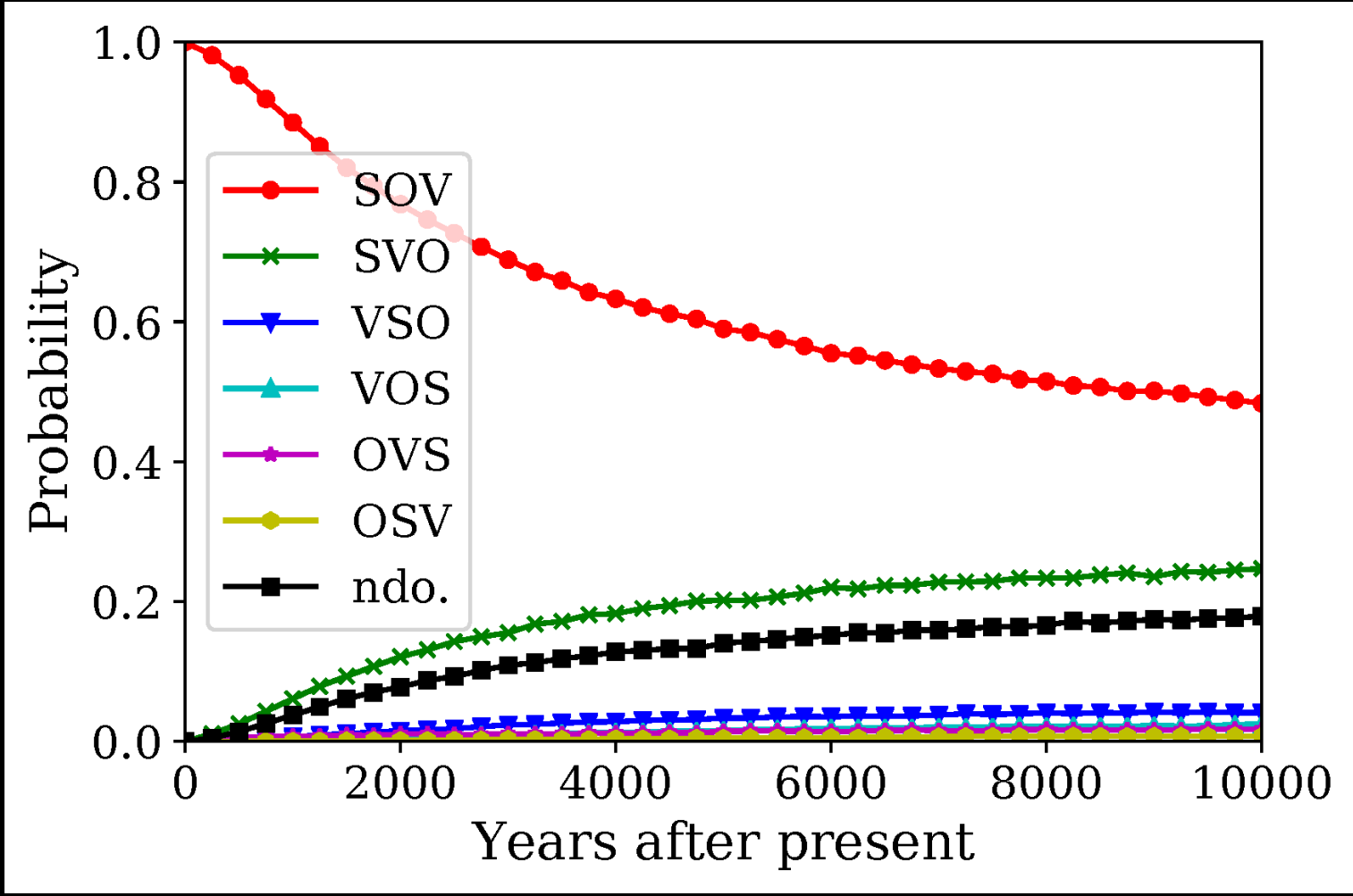
Step 3. Simulate language evolution using TRMs



Regression Analysis

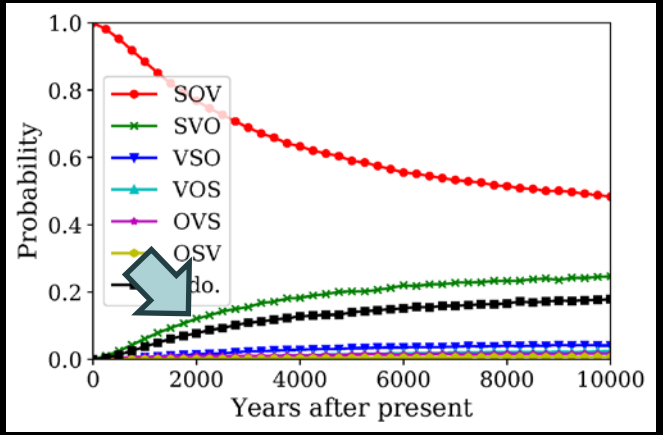
- Feature values typically associated with the specified word order have positive weights
  - SOV: genitive before noun, postpositions, RelN, suffixes
- Diachronic stability of SVO positively correlates with
  - 51A Position of Case Affixes: Case prefixes
  - 26A Prefixing vs. Suffixing in Inflectional Morphology: Little affixation
  - 26A Prefixing vs. Suffixing in Inflectional Morphology: Strong prefixing
  - 51A Position of Case Affixes: No case affixes or adpositional clitics

Simulating Japanese BWO



Regression Analysis Again

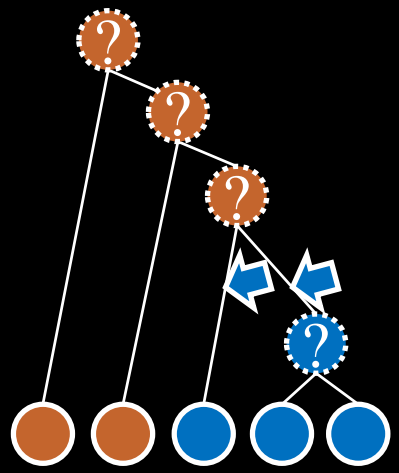
- Future Japanese with SVO order (12.1%) is characterized by:
  - 85A Order of Adposition and Noun Phrase: Prepositions
  - 51A Position of Case Affixes: No case affixes or adpositional clitics



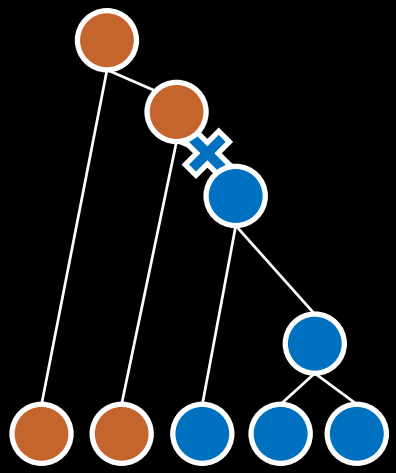
Conclusions and Future Work

- New framework of latent representation-based analysis of diachronic typology
  - Investigate correlated evolution in an exploratory manner
- Future work
  - Analyze features other than BWO
  - Manually analyze inferred ancestral states
  - Modeling contacts [Murawaki, NAACL2016]

Contact-induced Changes



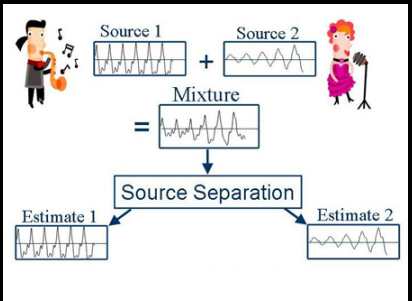
What actually happened



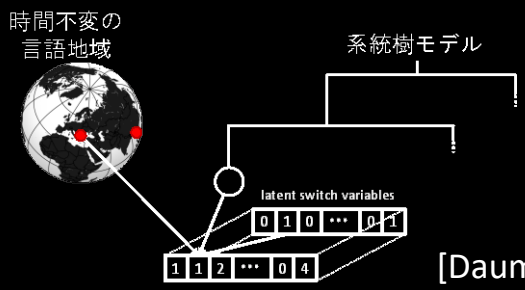
What is inferred from leaf nodes

- Example: Tone
  - All modern Sinitic languages are tonal
  - Proto-Sinitic had a complex tone system with prob. of 70.4%, according to our analysis
  - But Old Chinese was atonal [Baxter+, 2014]
- Need to collect typological profiles of past languages

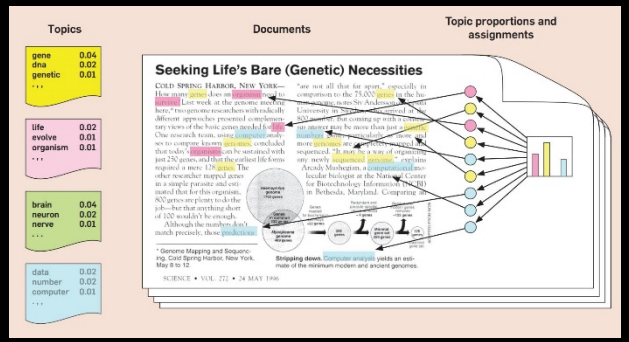
Mixture Model aka Topic Model



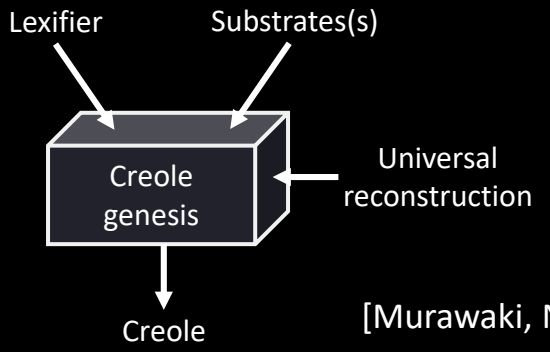
Audio source separation



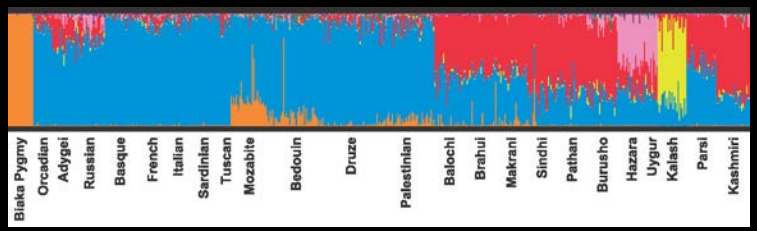
[Daumé III, NAACL2009]



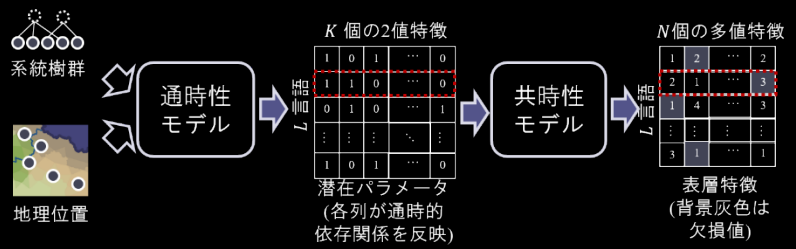
Topic model of documents [Blei+, 2012]



[Murawaki, NAACL2016]

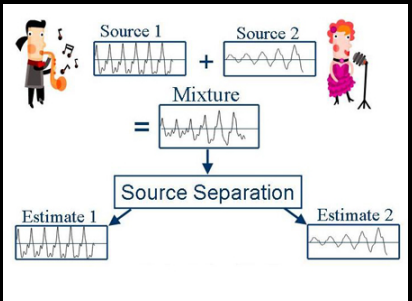


Admixture analysis
in population genetics [Pritchard+, 2000]

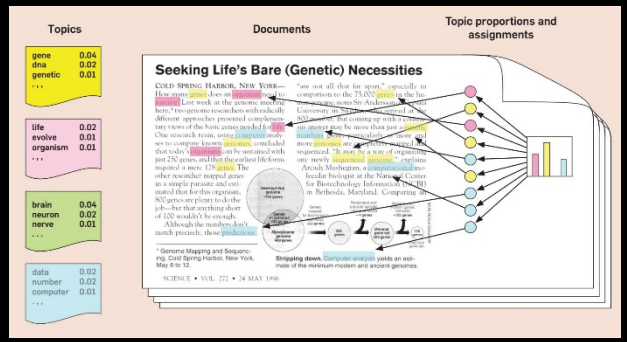


[Yamauchi+, COLING2016] [Murawaki, IJCNLP2017]

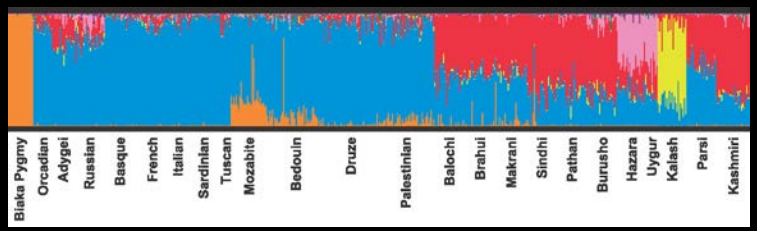
Mixture Model aka Topic Model



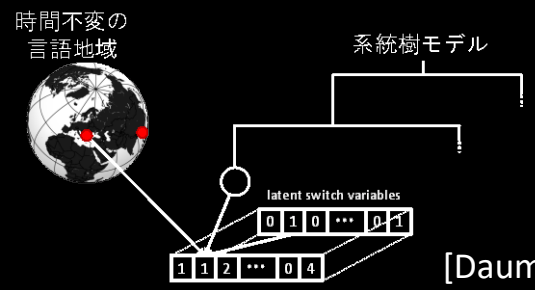
Audio source separation



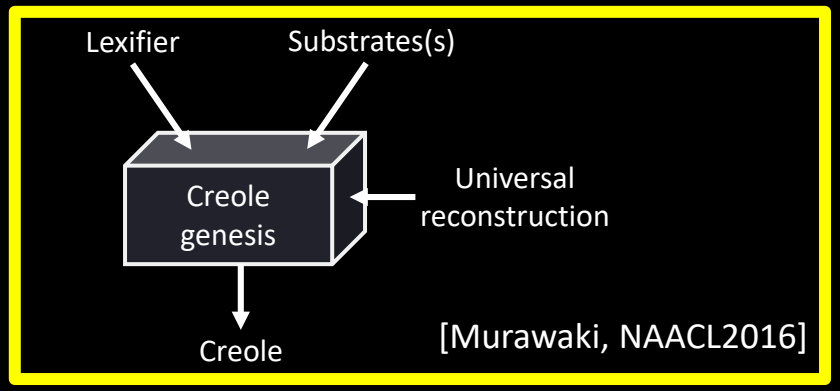
Topic model of documents [Blei+, 2012]



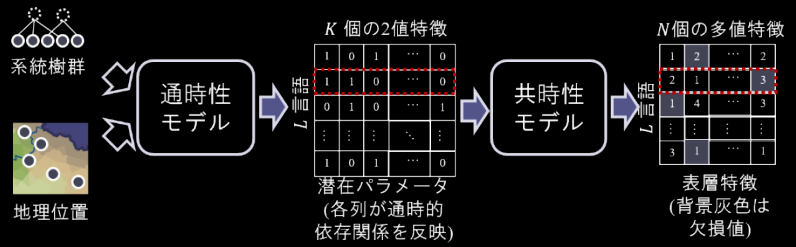
Admixture analysis
in population genetics [Pritchard+, 2000]



[Daumé III, NAACL2009]



[Murawaki, NAACL2016]



[Yamauchi+, COLING2016] [Murawaki, IJCNLP2017]