

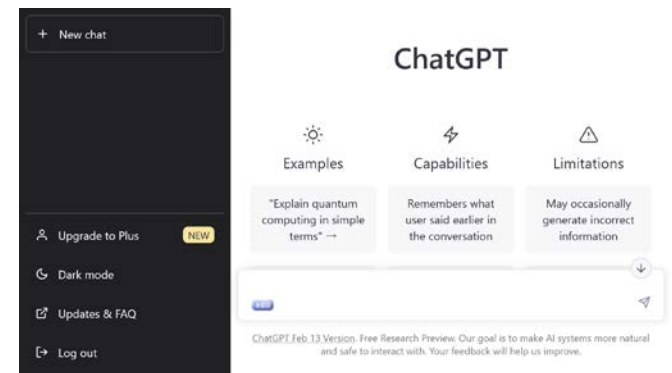
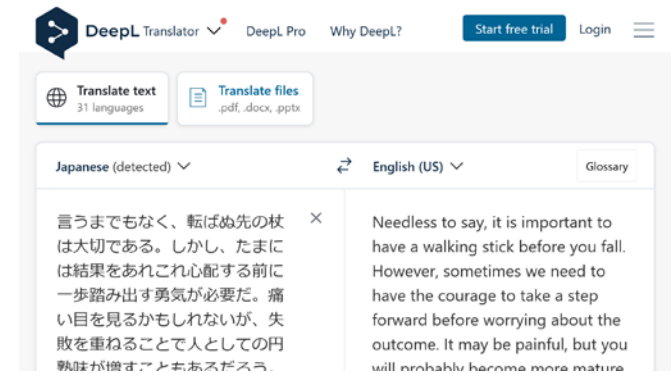
言語変化の数理

京都大学大学院情報学研究科
村脇 有吾



自己紹介: 村脇 有吾

- 研究的背景: ずっと情報系。人文系の背景はない
 - 京大工学部情報学科 → 情報学研究科
- 専門: 自然言語処理・計算言語学
 - 工学的応用: AIの急激な性能向上にともない社会的期待が膨張
 - 翻訳が長年の大課題だったはずなのに、ほとんどできてしまっている…
 - 自然なテキストの生成は夢のまた夢だったのに、ChatGPTが何でも流暢に答えてくれる…
 - 科学的探究: なぜ我々の言語はこんな風なのか？
 - 今日の発表はこちらに振り切った話題



英語は語順をひっくり返す言語？

日本語 私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語は語順をひっくり返す言語？

日本語 私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語 *I want to try on a suit I saw in a shop that's across the street from the hotel*

英語は語順をひっくり返す言語？

ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

日本語

私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

英語は語順をひっくり返す言語？



ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

日本語

私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

英語とネパール語は語順が全然違うが 祖先は共通…ということは？

ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

英語とネパール語は語順が全然違うが 祖先は共通…ということとは？

ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

インド・ヨーロッパ祖語

英語

ネパール語

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

英語とネパール語は語順が全然違うが 祖先は共通…ということは？

ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

インド・ヨーロッパ祖語

どこかの時点で
少なくとも一度
語順が変化したはず

英語

ネパール語

英語

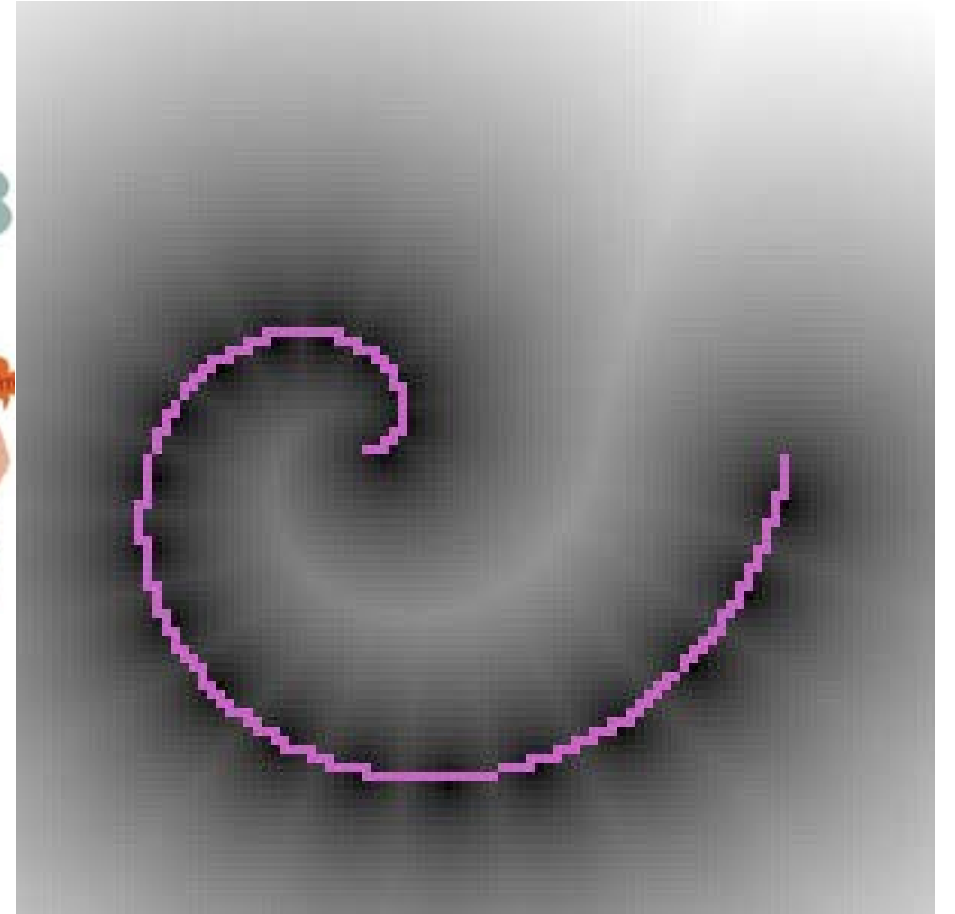
I want to try on a suit I saw in a shop that's across the street from the hotel

疑問



- どうして言語によって語順が異なるのか？
- 語順はどのように時間変化するのか？
 - 我々の子や孫が突然別の語順を使いだすとは思えない
 - 日本語は万葉集や魏志倭人伝の時代までさかのぼっても語順に変化が見られない
 - もし日本語が将来別の語順を使うようになるとしたら、どんな語順か？
 - もし日本語が有史以前に別の語順を使っていたとしたら、どんな語順か？

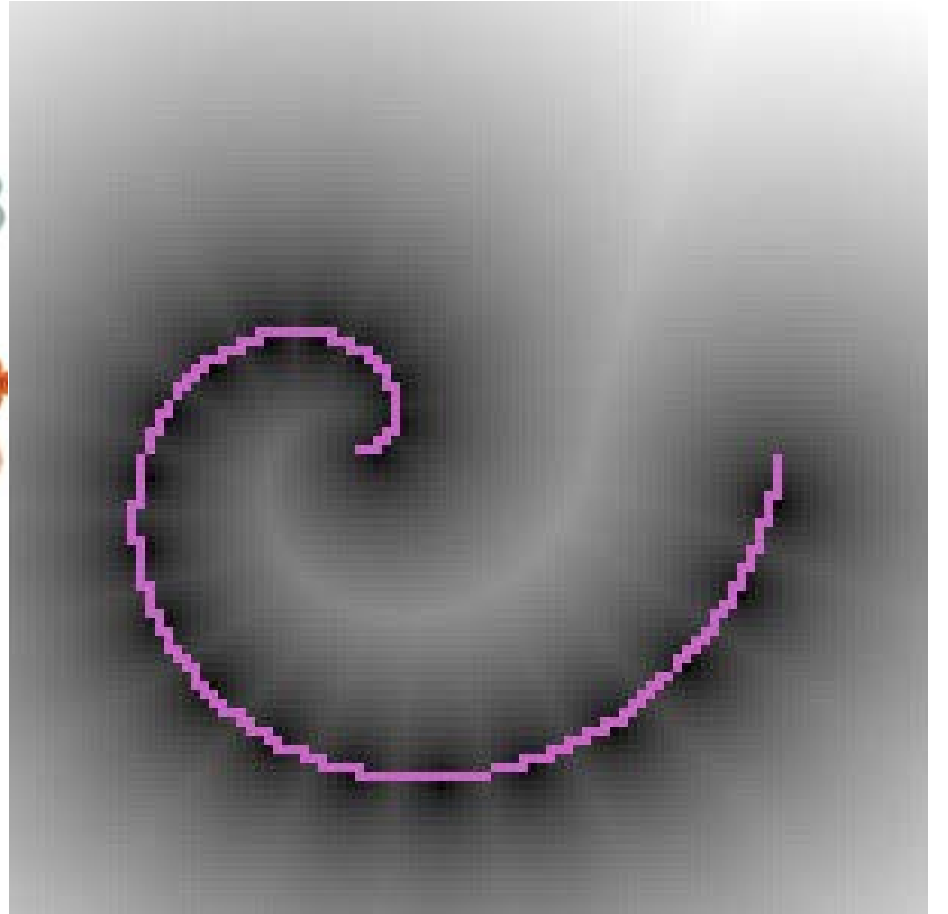
人文系の領域の問題に 情報学から取り組むとどうなる？



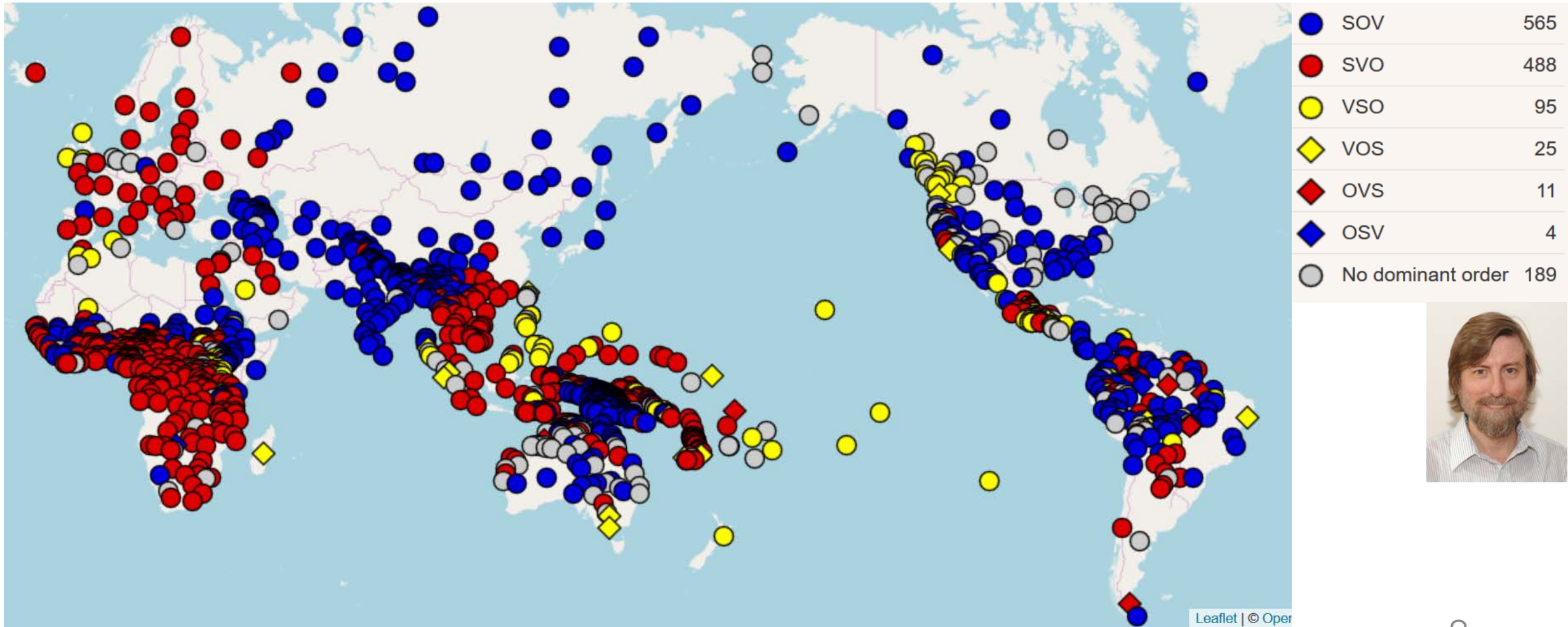
人文系の領域の問題に 情報学から取り組むとどうなる？



この2つには共通の性質が！



Subject主語-Object目的語-Verb動詞の基本語順



SOV

SVO

VSO

VOS

OVS

OSV

SV対VS と OV対VO

SOV

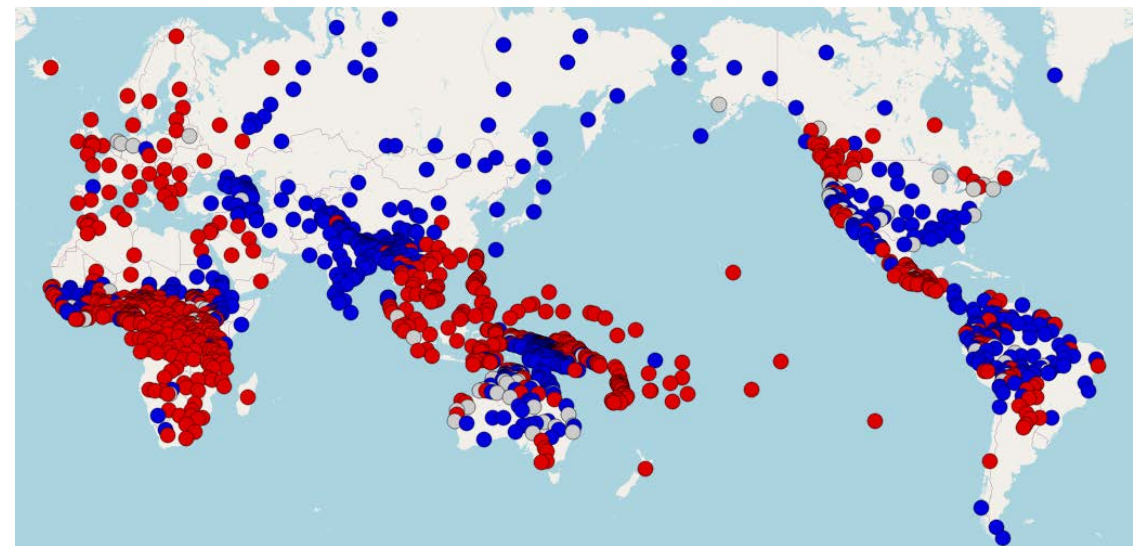
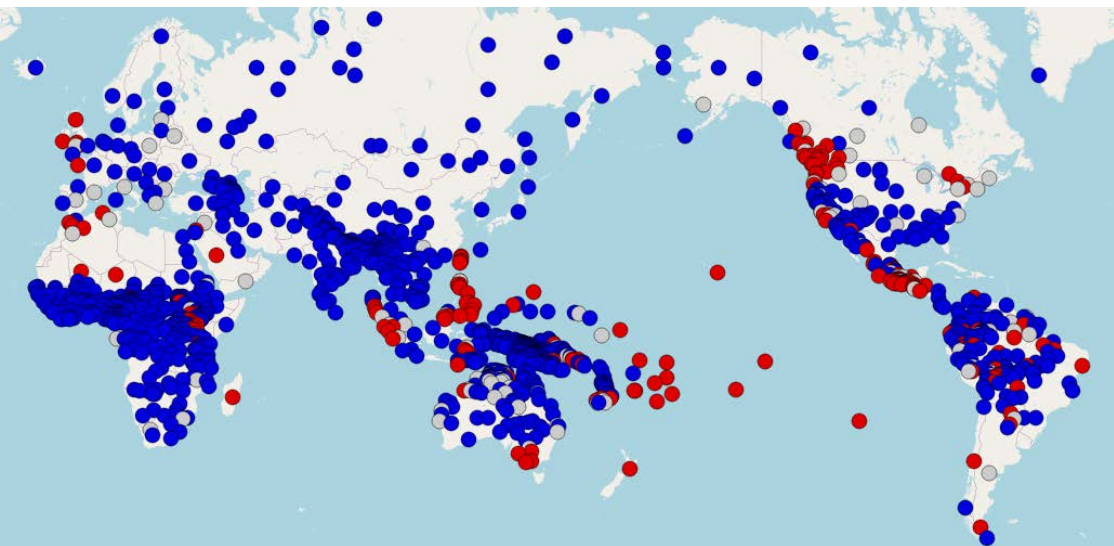
SVO

VSO

VOS

OVS

OSV



文献記録に残る語順変化を調べる

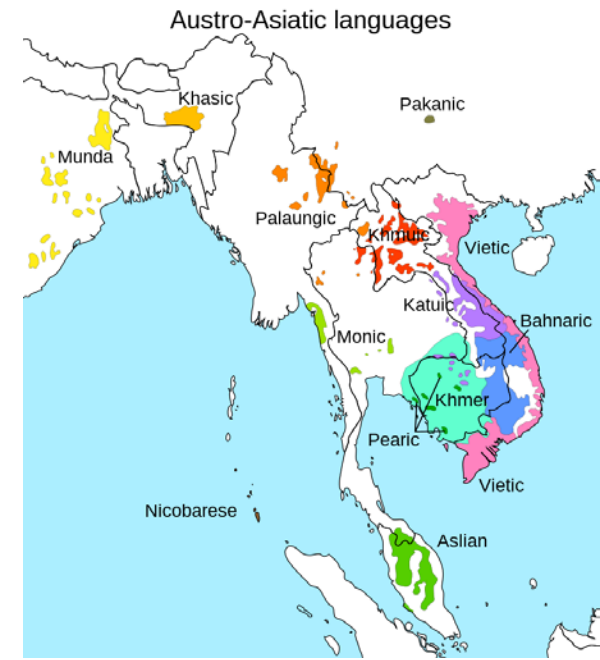
- 古英語 (450-1100) の語順は比較的自由だったが、中英語 (1100-1500) 期に現代語のような厳格なSVO語順に変化

er fela ðinga swa gerad man sceal don (OSV)
and such a wise man must do many things

(Rectitudines Singularum Personarum, 1070-1100年頃)

- 文献記録に残るのは例外的で、ほとんどの語順変化は記録されていない

記録されていない語順変化を 明らかにしたい



日本語

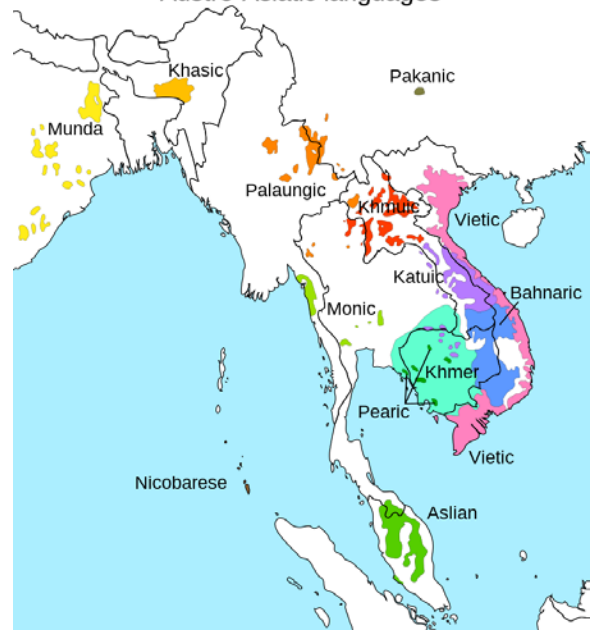
彼は私にご飯をあげたくない

ソラ語

amin dən-nɛn darəj-an ə-tiy-ben idsim-tɛ ted

クメール語

kǎət ʔət caŋ ʔaoy baay knom



記録されていない語順変化を明らかにしたい

日本語

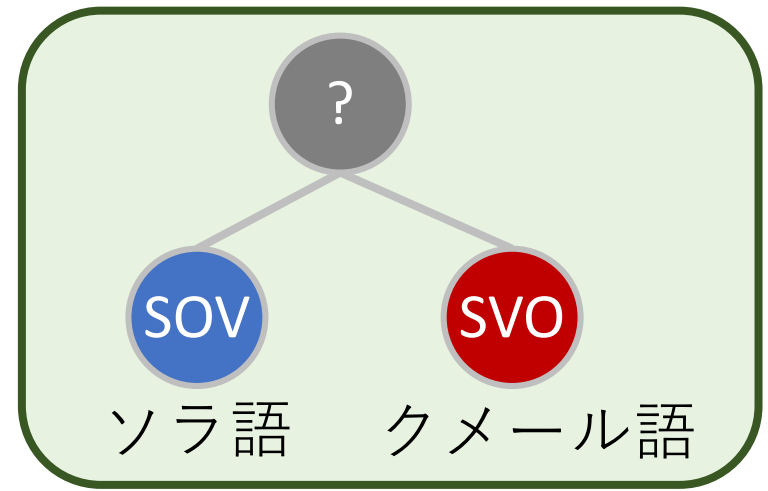
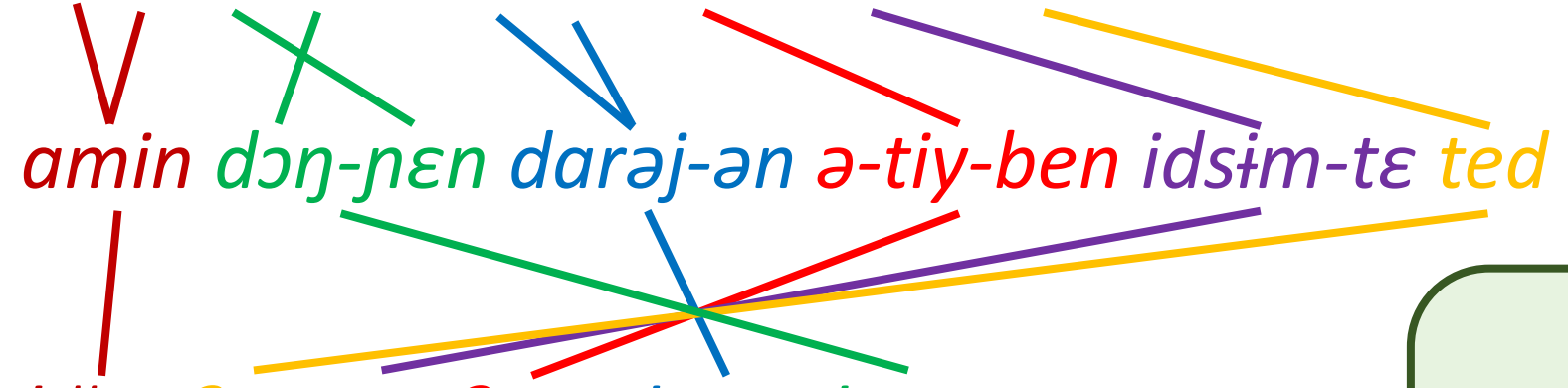
彼は私にご飯をあげたくない

ソラ語

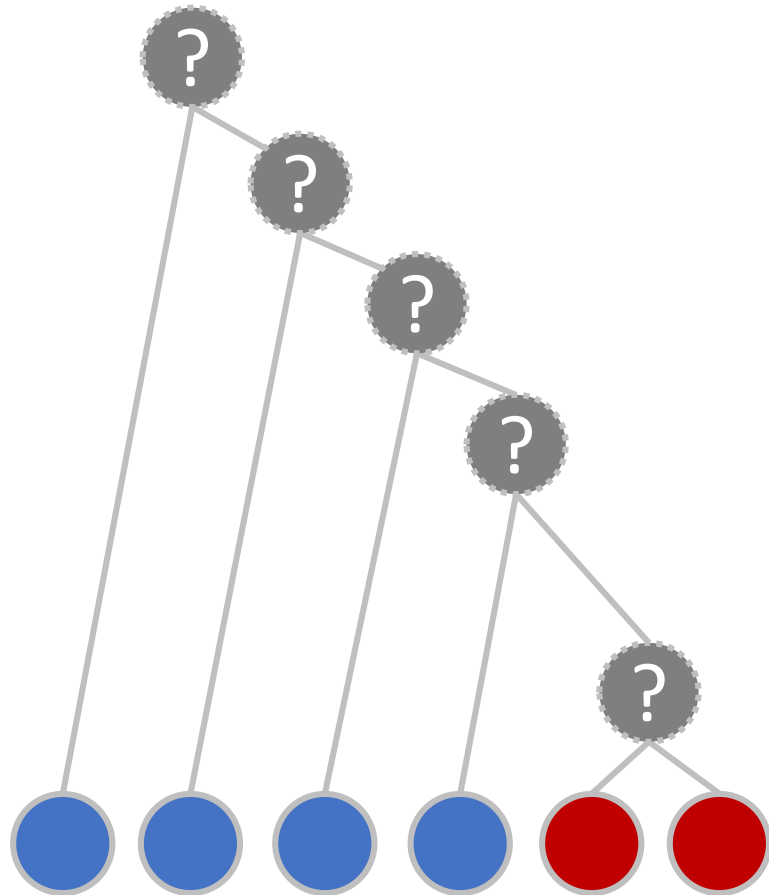
amin dən-nen darəj-an ə-tiy-ben idsim-tɛ ted

クメール語

kǎət ʔət caŋ ʔaoy baay knom

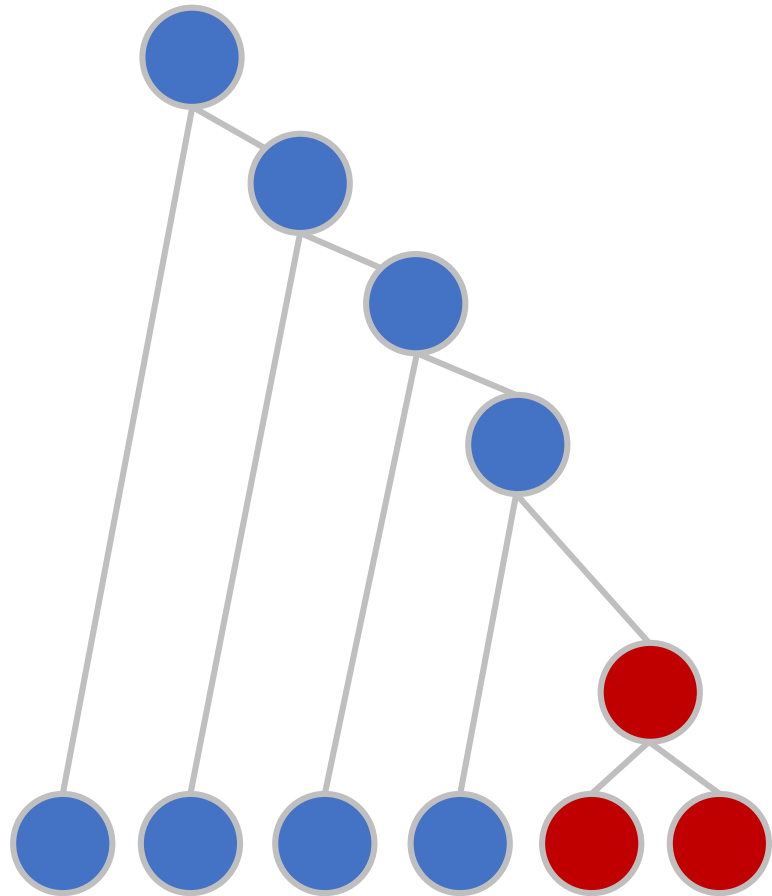


系統学的比較法



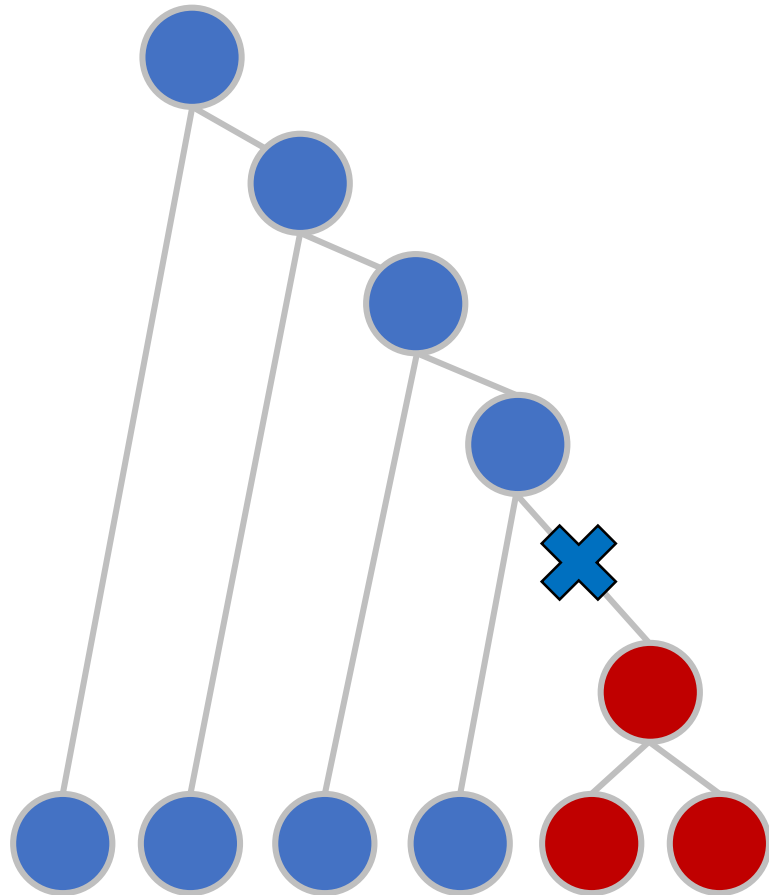
- 言語同士の歴史的関係（系統樹）は既知とする
 - 歴史比較言語学の成果によって
- 祖語の状態と変化が起きた枝が推測できる場合がある

系統学的比較法



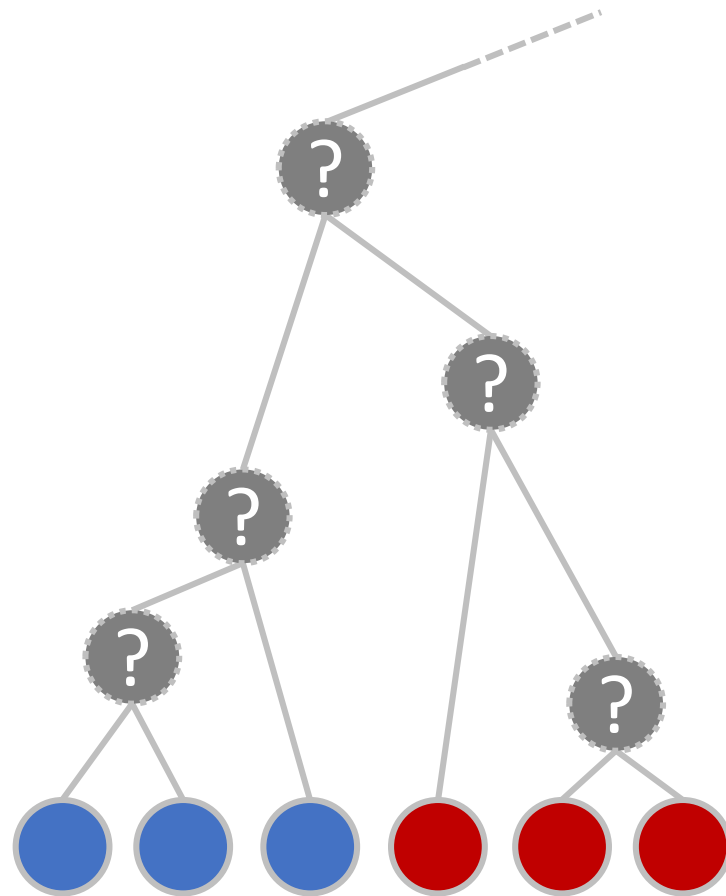
- 言語同士の歴史的関係（系統樹）は既知とする
 - 歴史比較言語学の成果によって
- 祖語の状態と変化が起きた枝が推測できる場合がある

系統学的比較法



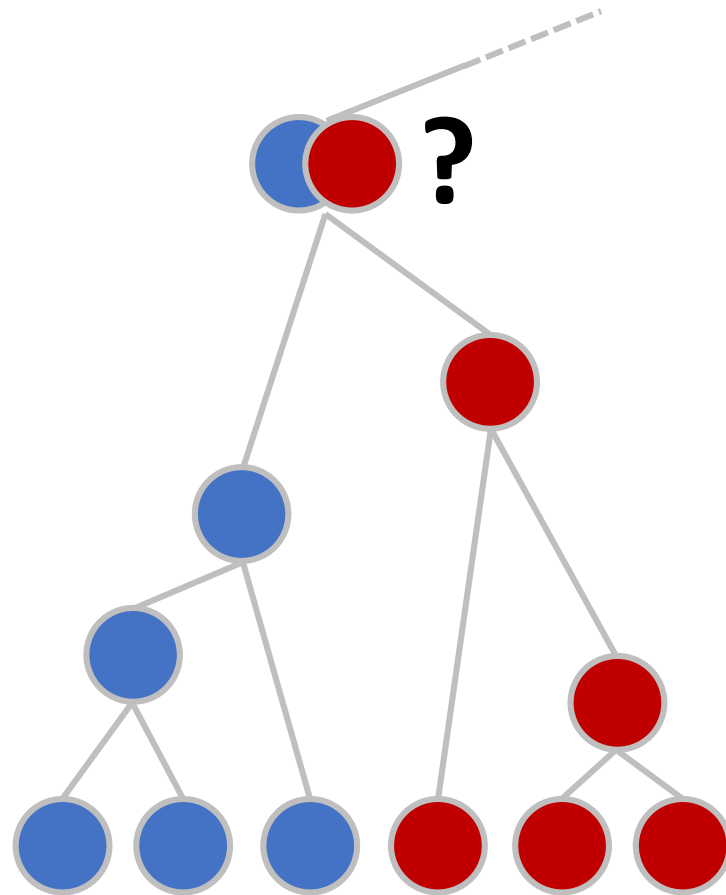
- 言語同士の歴史的関係（系統樹）は既知とする
 - 歴史比較言語学の成果によって
- 祖語の状態と変化が起きた枝が推測できる場合がある

系統学的比較法



- 現在の手がかりだけでは確信をもって推測できない場合も多い
- 人間 (言語学者) による論証は手詰まり

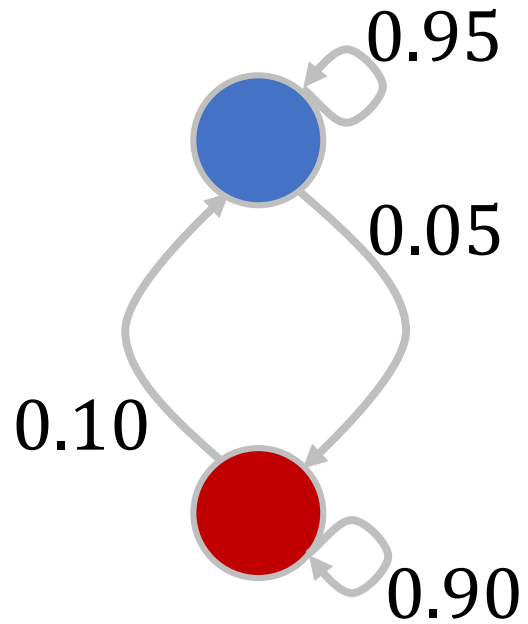
系統学的比較法



- 現在の手がかりだけでは確信をもって推測できない場合も多い
- 人間（言語学者）による論証は手詰まり

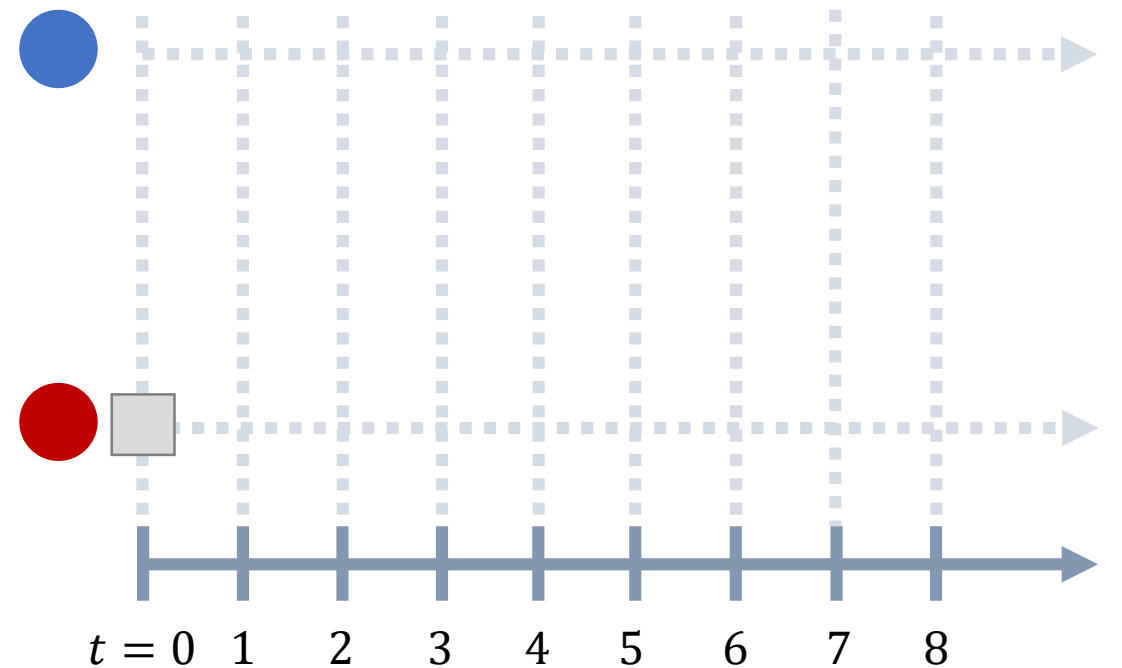
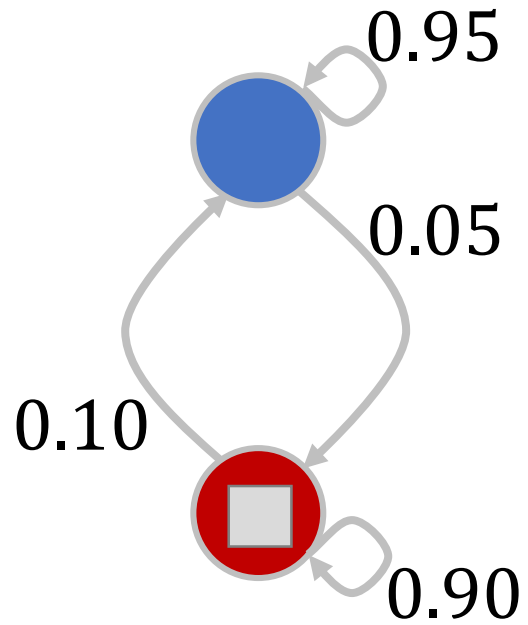
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化



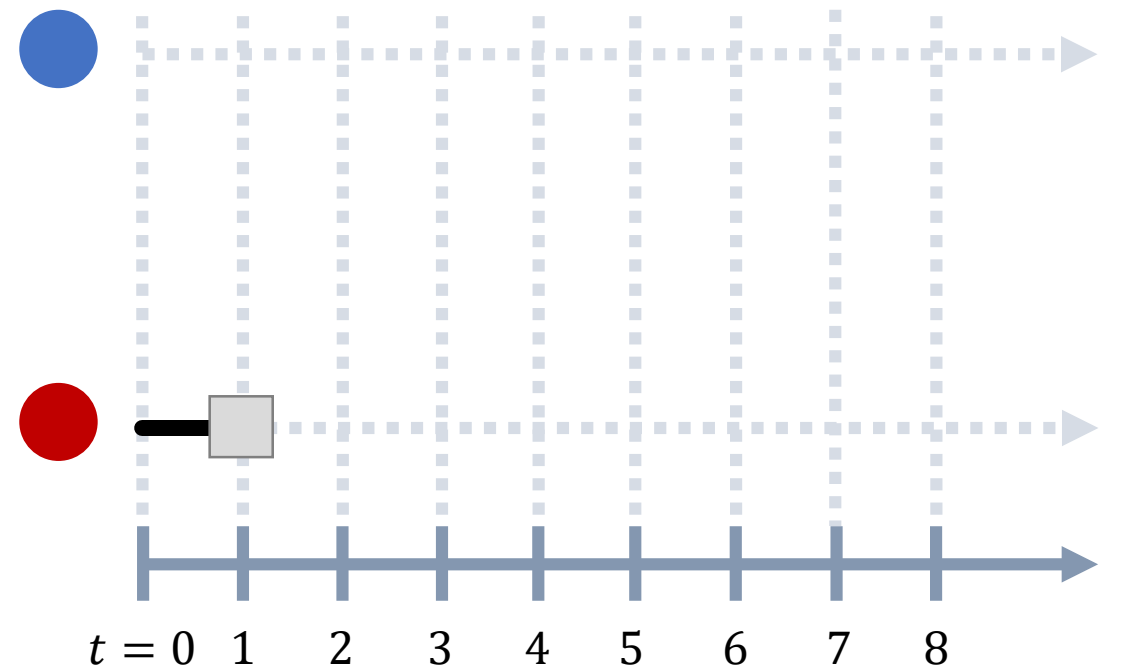
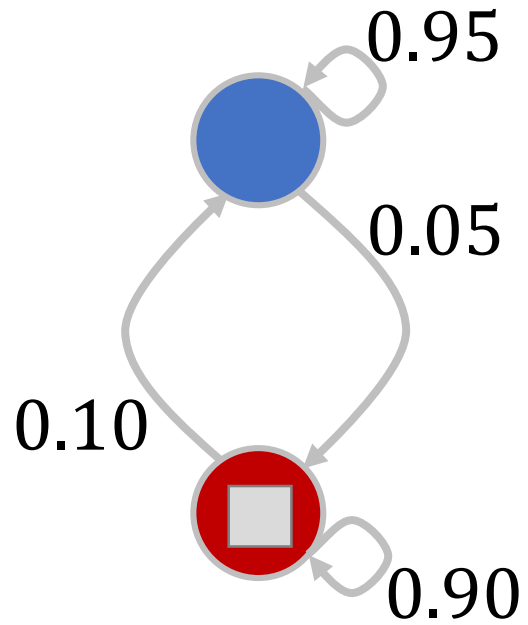
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化



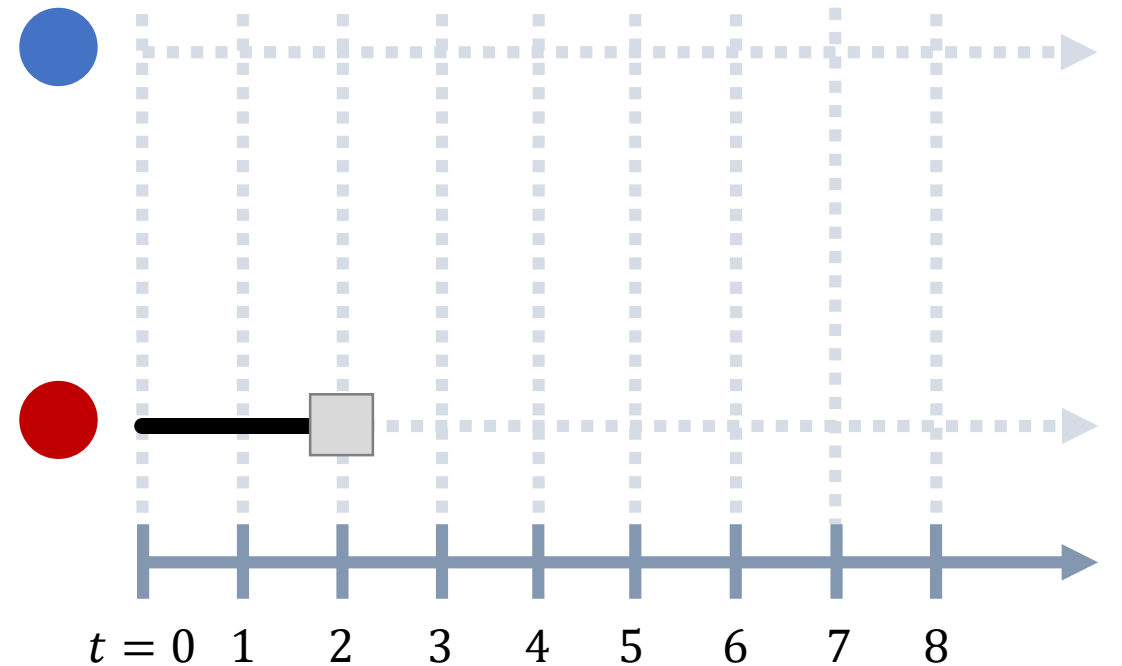
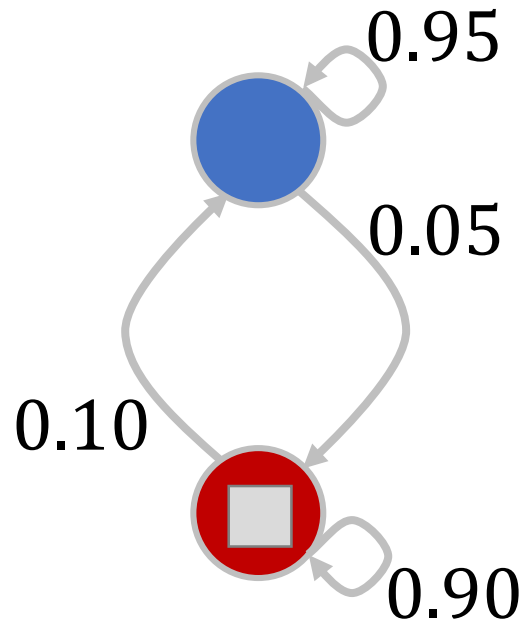
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化



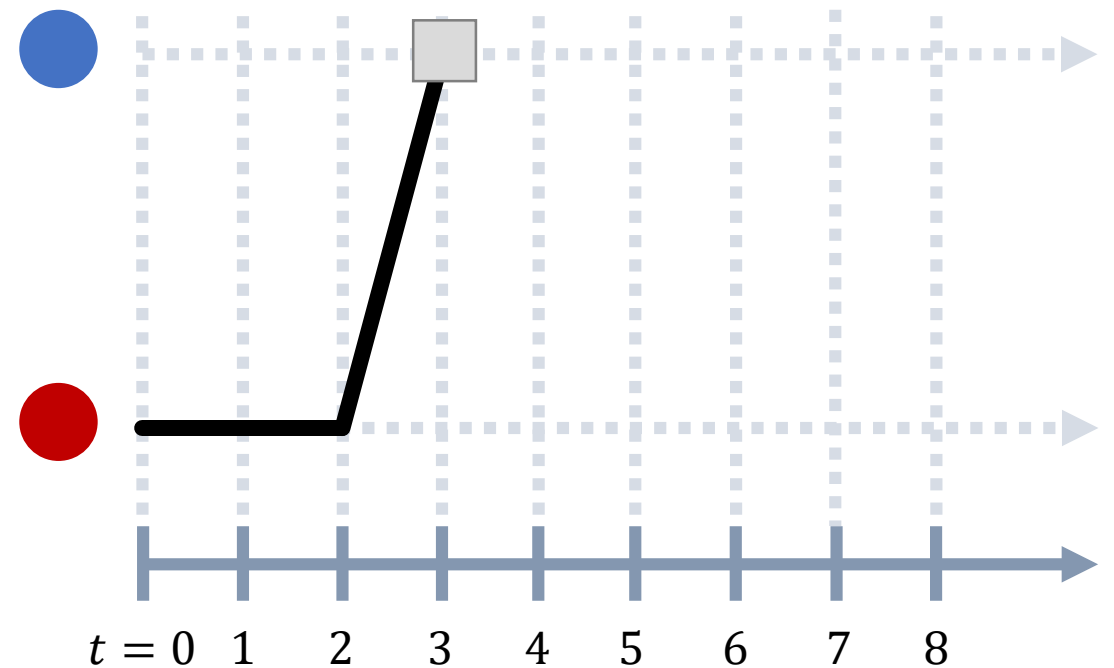
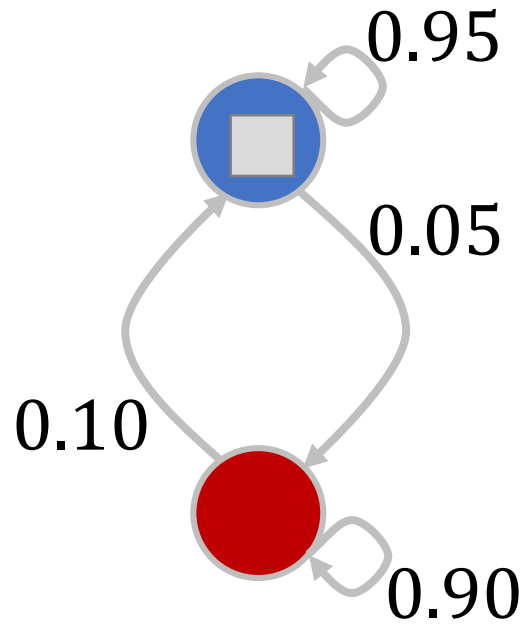
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化



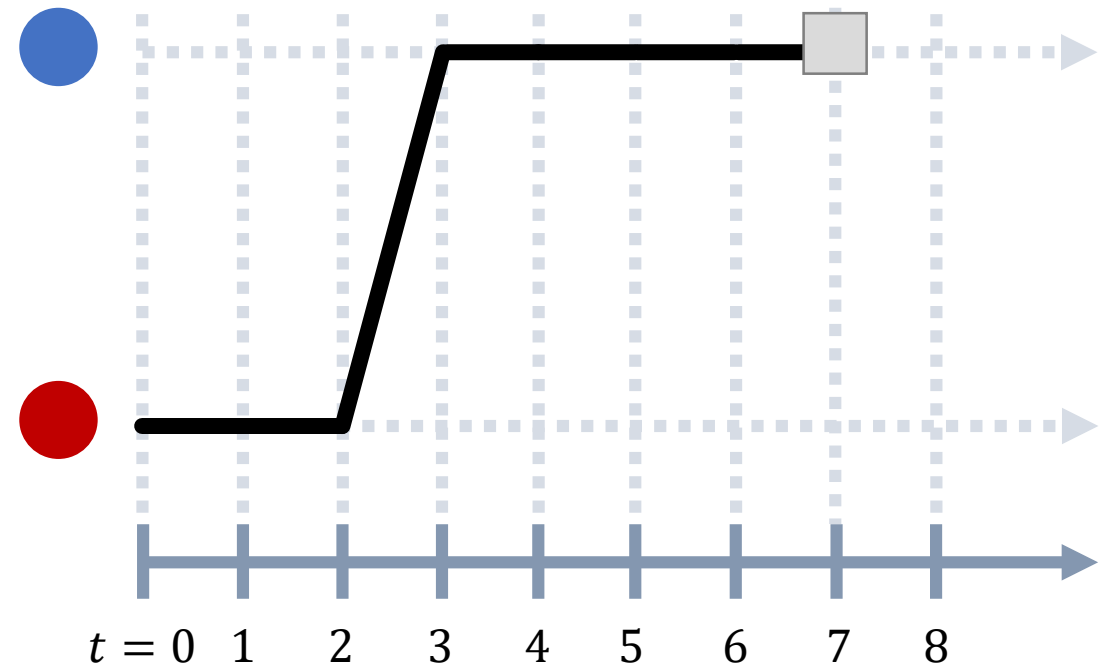
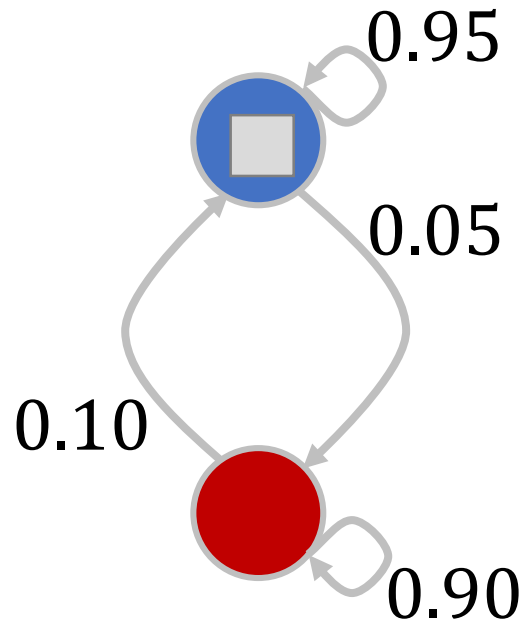
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化



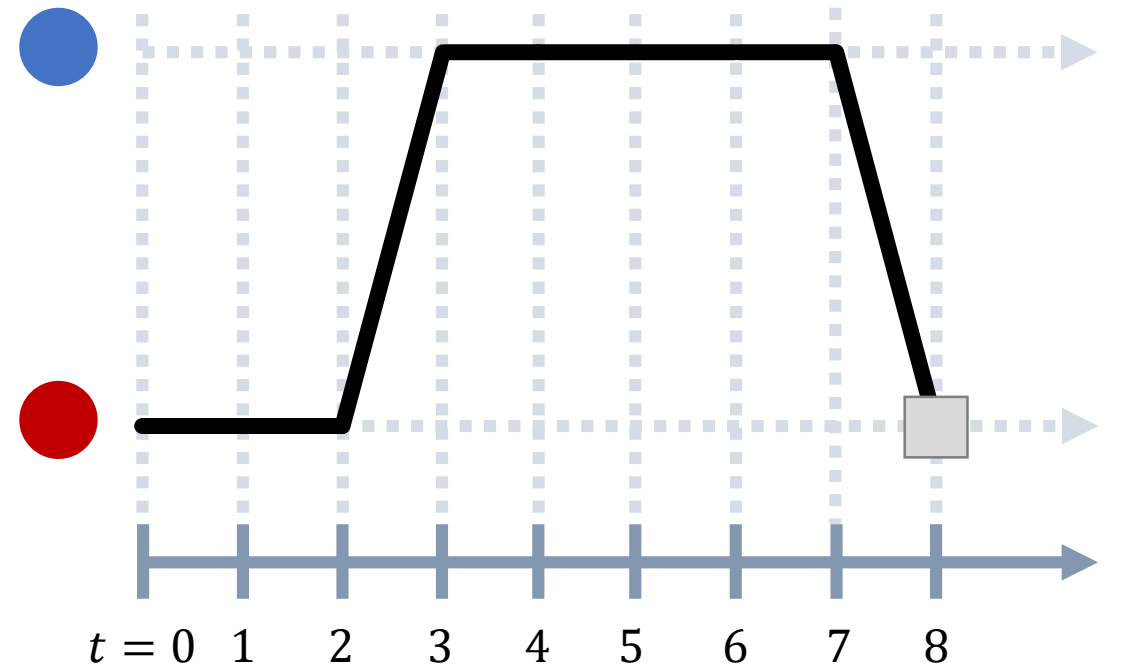
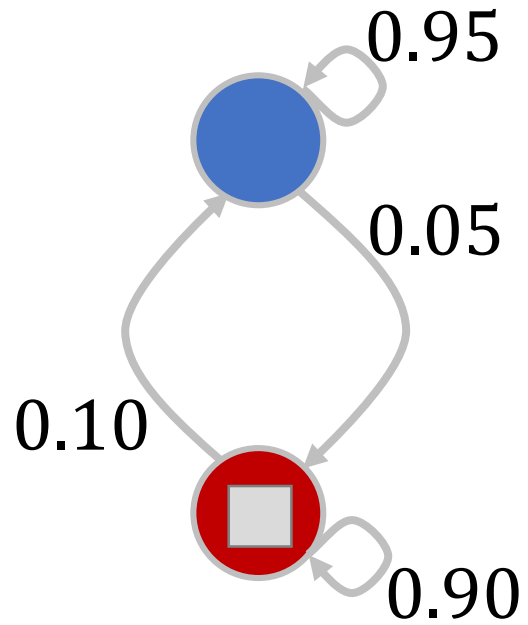
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化



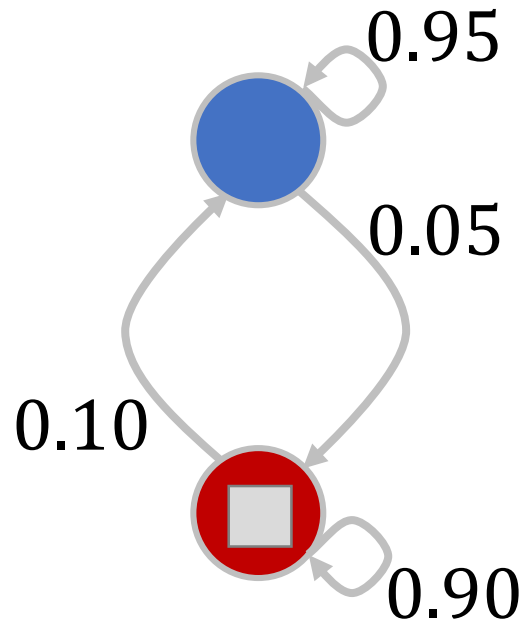
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化

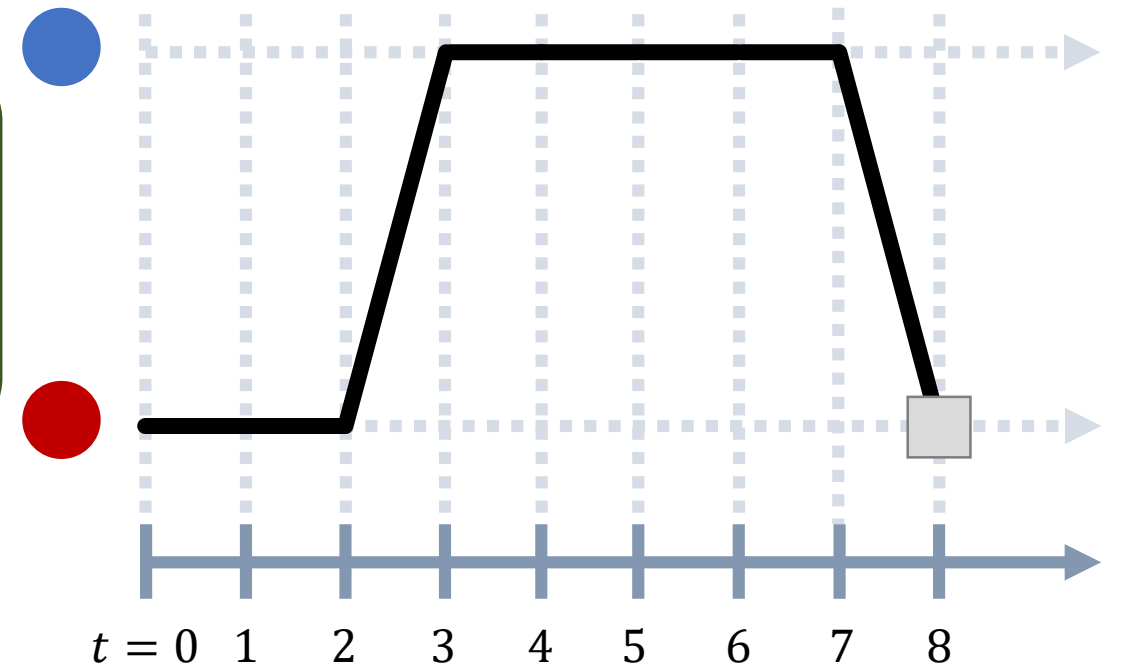


情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化

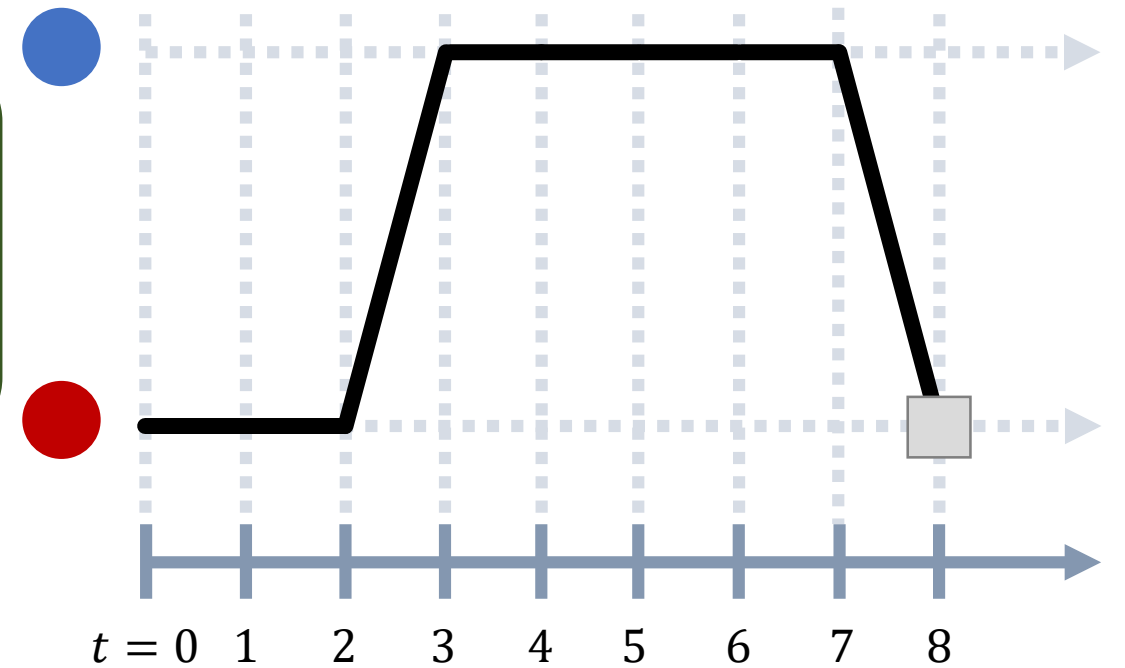
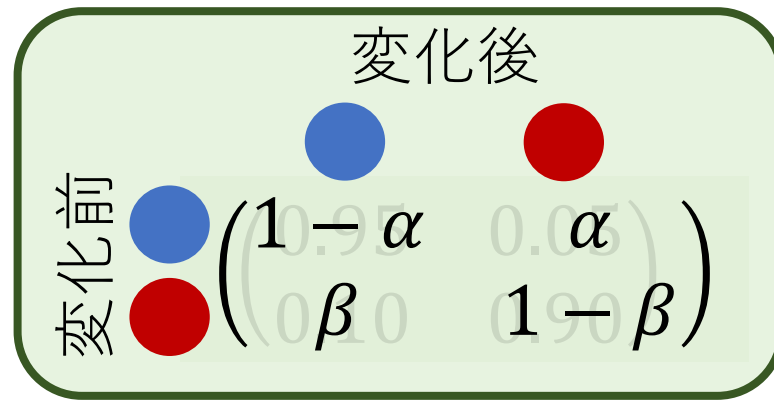
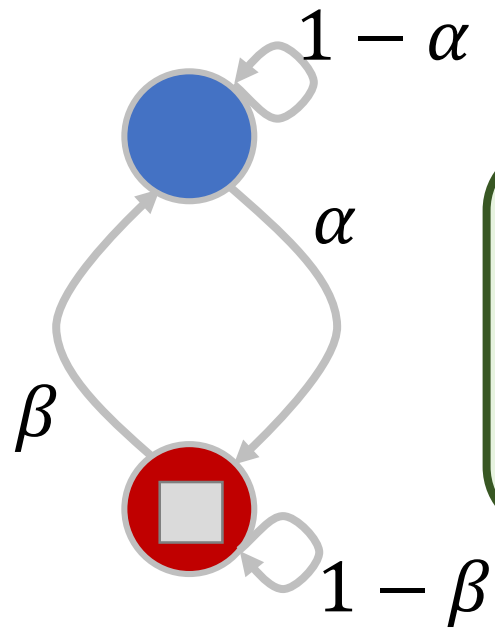


		変化後	
変化前	● (Blue)	0.95	0.05
	● (Red)	0.10	0.90



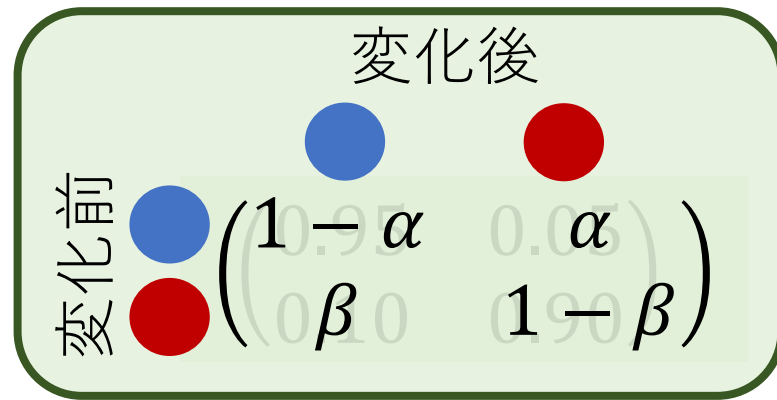
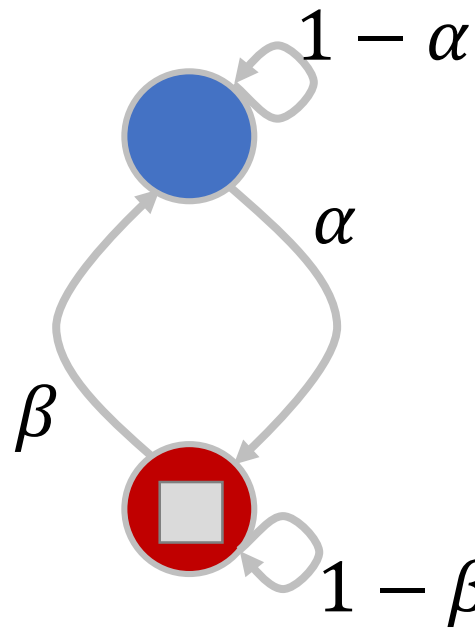
情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化

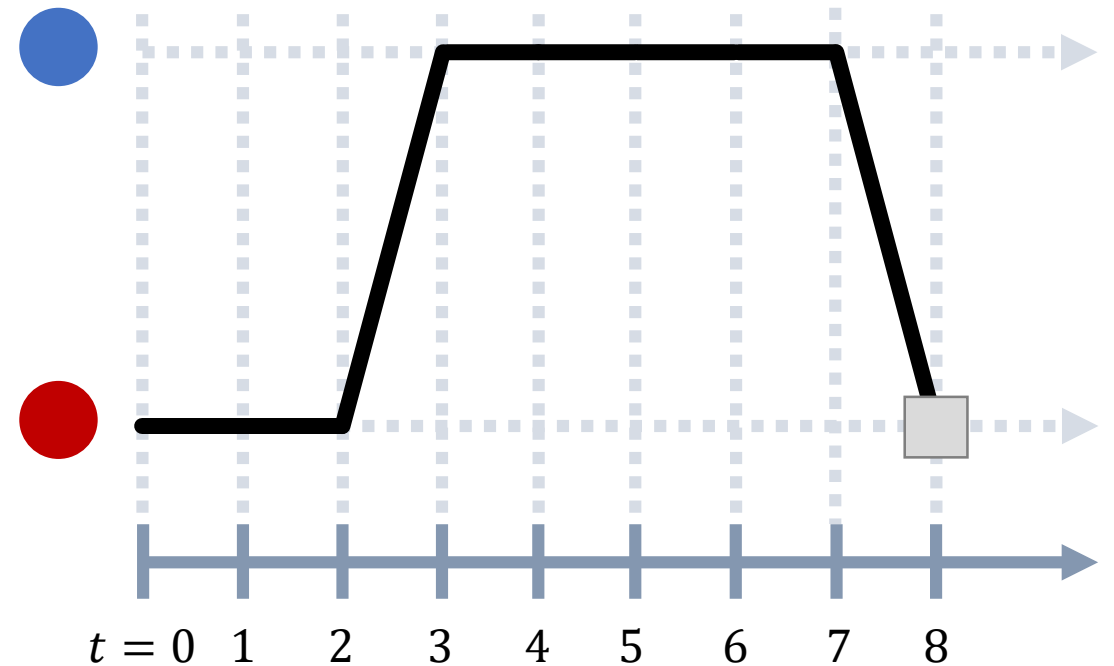


情報学の出番: 数理モデルの導入

- 統計を使って不確実性をモデル化

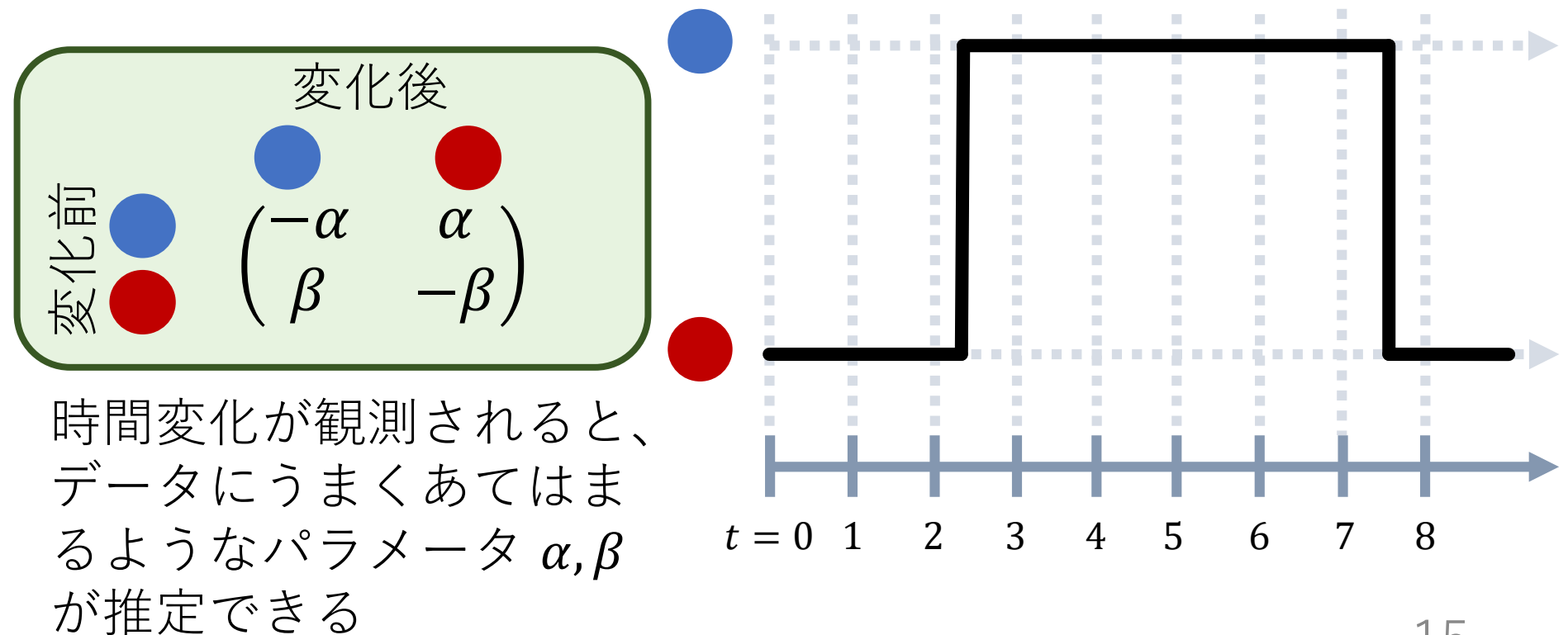


時間変化が観測されると、データにうまくあてはまるようなパラメータ α, β が推定できる



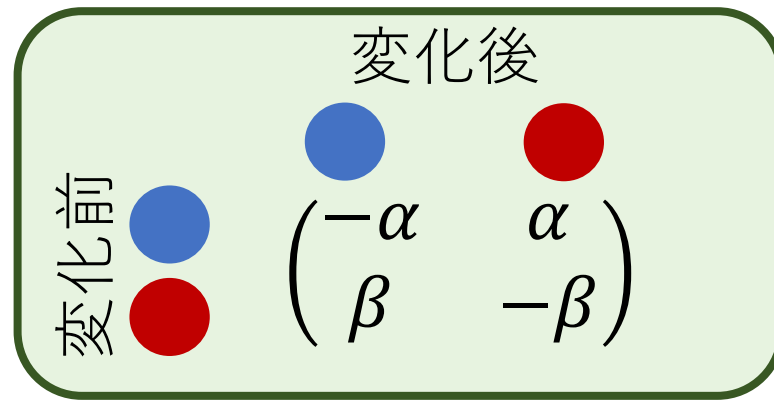
情報学の出番: 数理モデルの導入

- 離散時間から連続時間に拡張



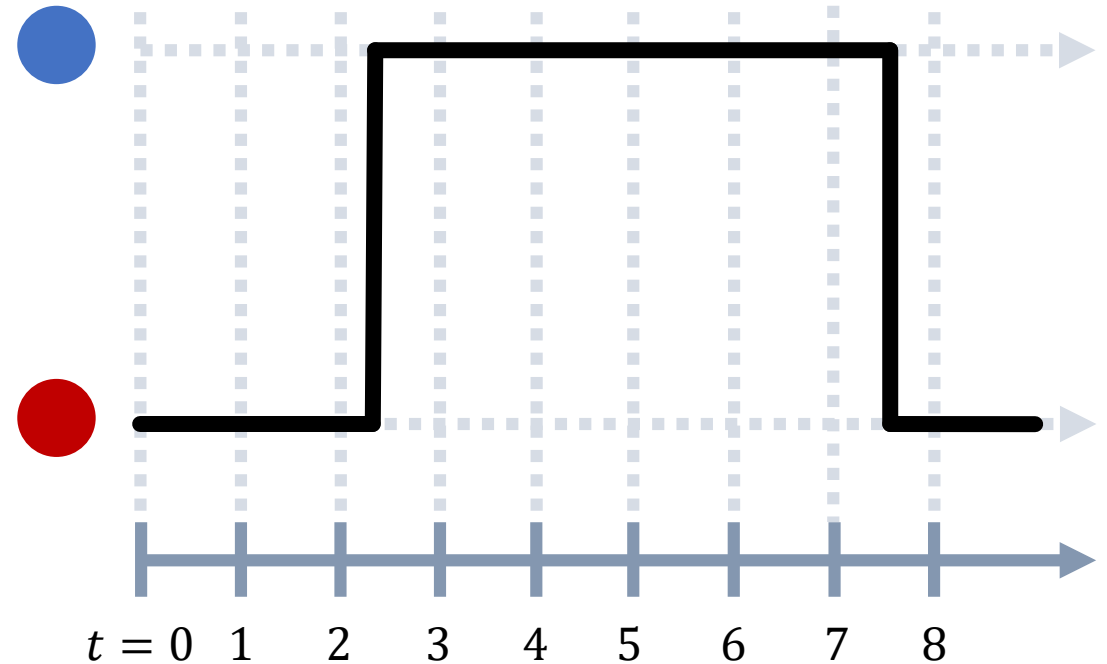
情報学の出番: 数理モデルの導入

- 離散時間から連続時間に拡張

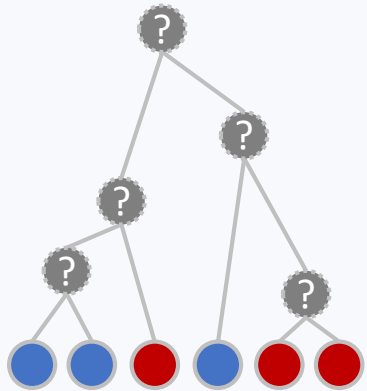


時間変化が観測されると、
データにうまくあてはまる
ようなパラメータ α, β
が推定できる

瞬間的な語順変化は
不自然では？
(あとで議論)



系統学的比較法 + 数理モデル

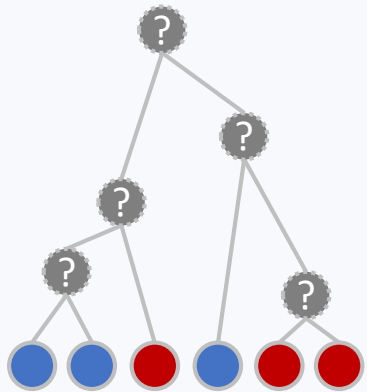


系統樹
(現代語の状態は既知)
(過去の状態は未知)

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

時間変化の
パラメータ

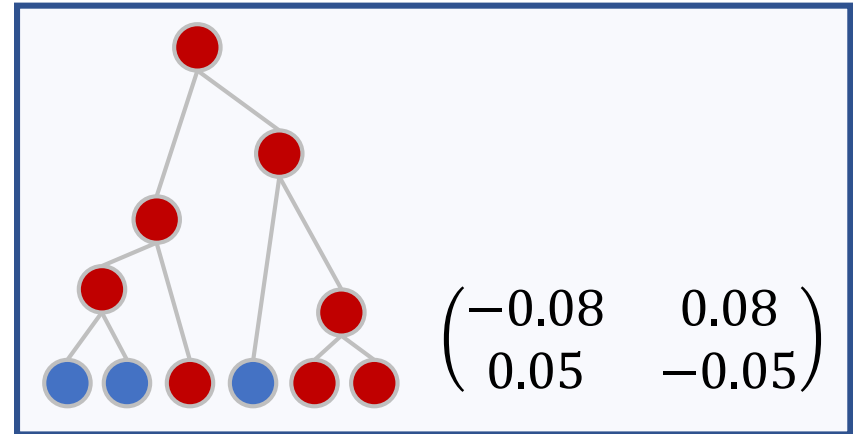
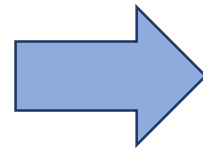
系統学的比較法 + 数理モデル



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

系統樹
(現代語の状態は既知)
(過去の状態は未知)

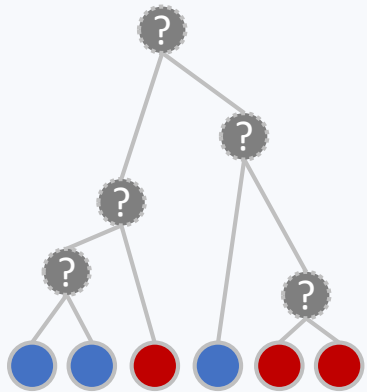
時間変化の
パラメータ



$$\begin{pmatrix} -0.08 & 0.08 \\ 0.05 & -0.05 \end{pmatrix}$$

コンピュータを
使った
シミュレーション

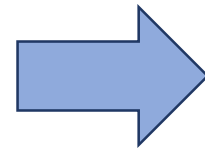
系統学的比較法 + 数理モデル



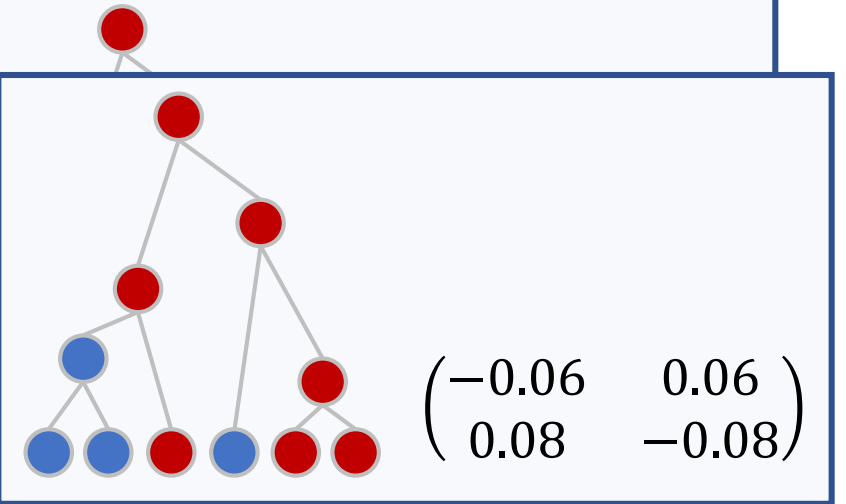
$$\begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

系統樹
(現代語の状態は既知)
(過去の状態は未知)

時間変化の
パラメータ

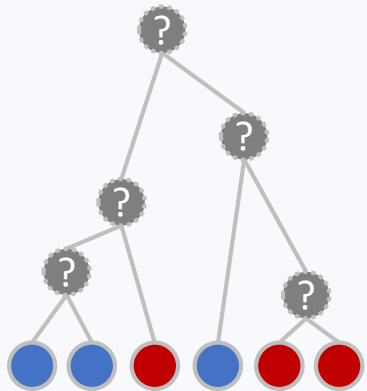


コンピュータを
使った
シミュレーション



$$\begin{pmatrix} -0.06 & 0.06 \\ 0.08 & -0.08 \end{pmatrix}$$

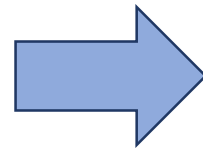
系統学的比較法 + 数理モデル



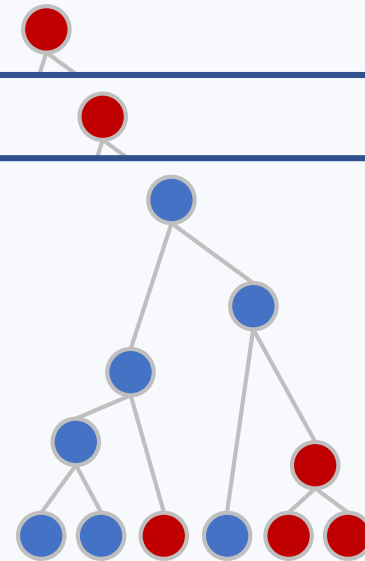
系統樹
(現代語の状態は既知)
(過去の状態は未知)

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

時間変化の
パラメータ

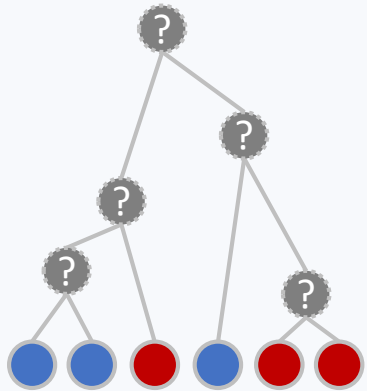


コンピュータを
使った
シミュレーション



$$\begin{pmatrix} -0.04 & 0.04 \\ 0.06 & -0.06 \end{pmatrix}$$

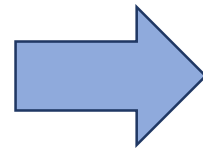
系統学的比較法 + 数理モデル



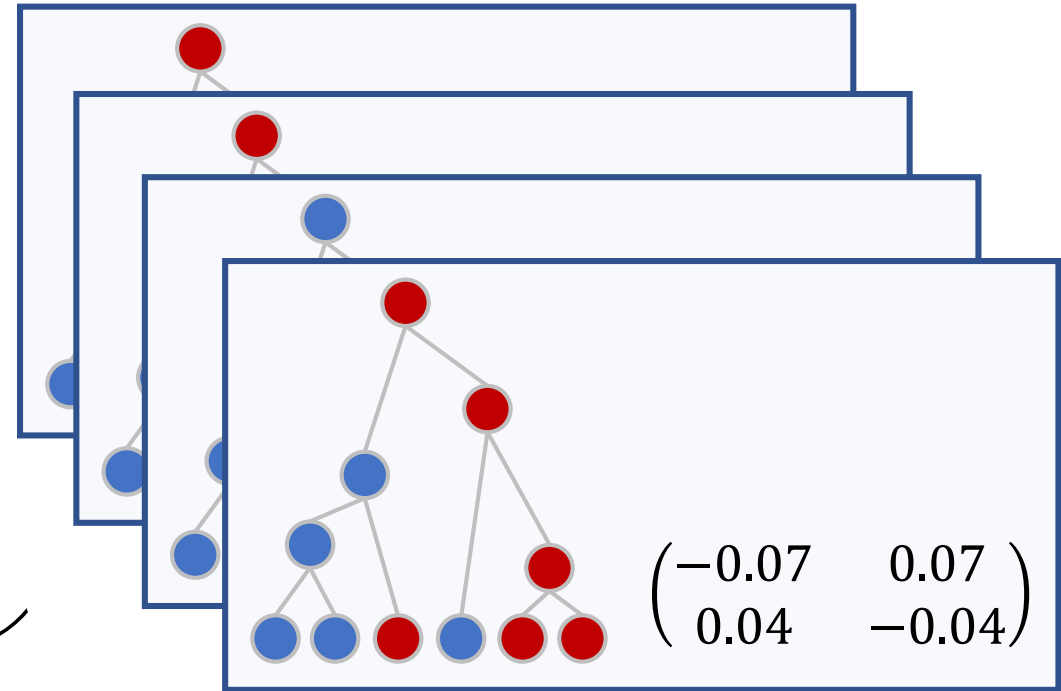
系統樹
(現代語の状態は既知)
(過去の状態は未知)

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

時間変化の
パラメータ

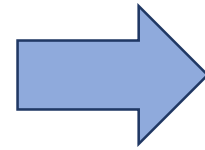
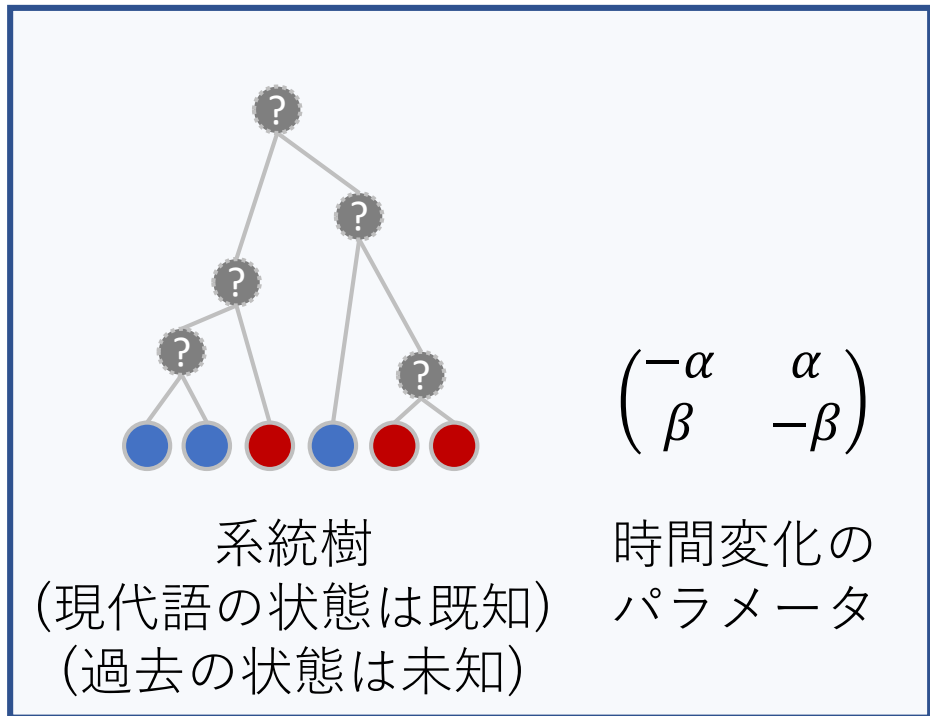


コンピュータを
使った
シミュレーション

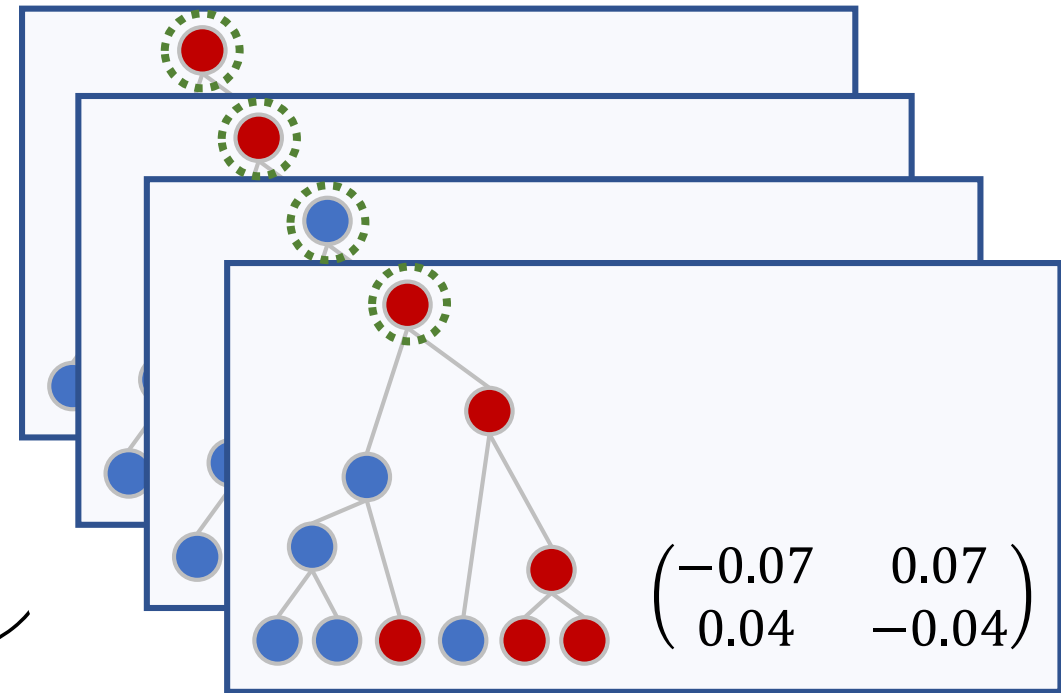


$$\begin{pmatrix} -0.07 & 0.07 \\ 0.04 & -0.04 \end{pmatrix}$$



系統学的比較法 + 数理モデル

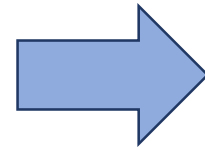
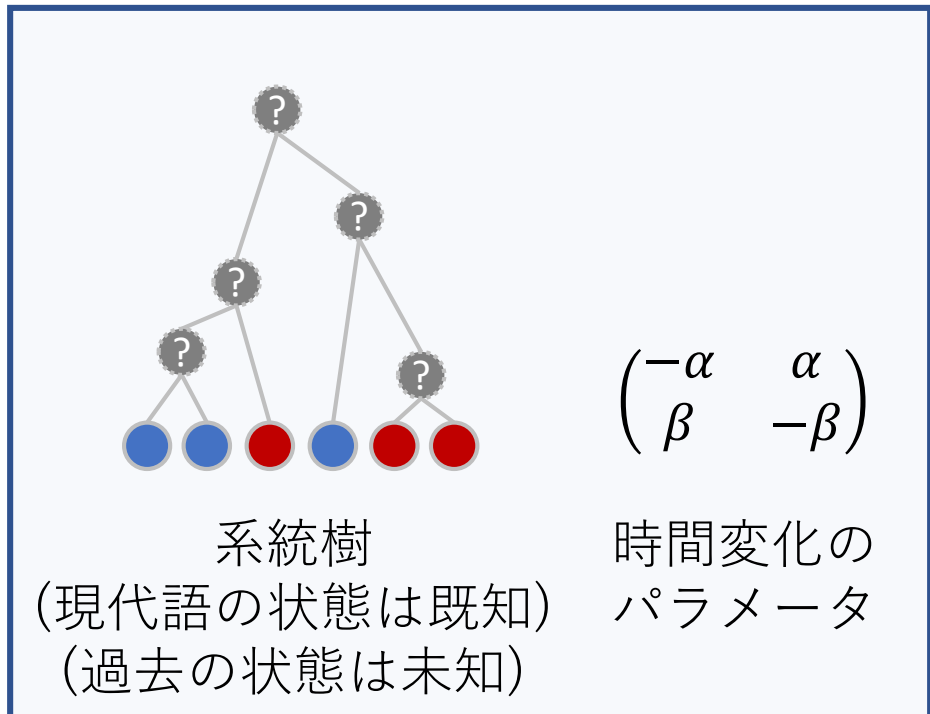


コンピュータを
使った
シミュレーション

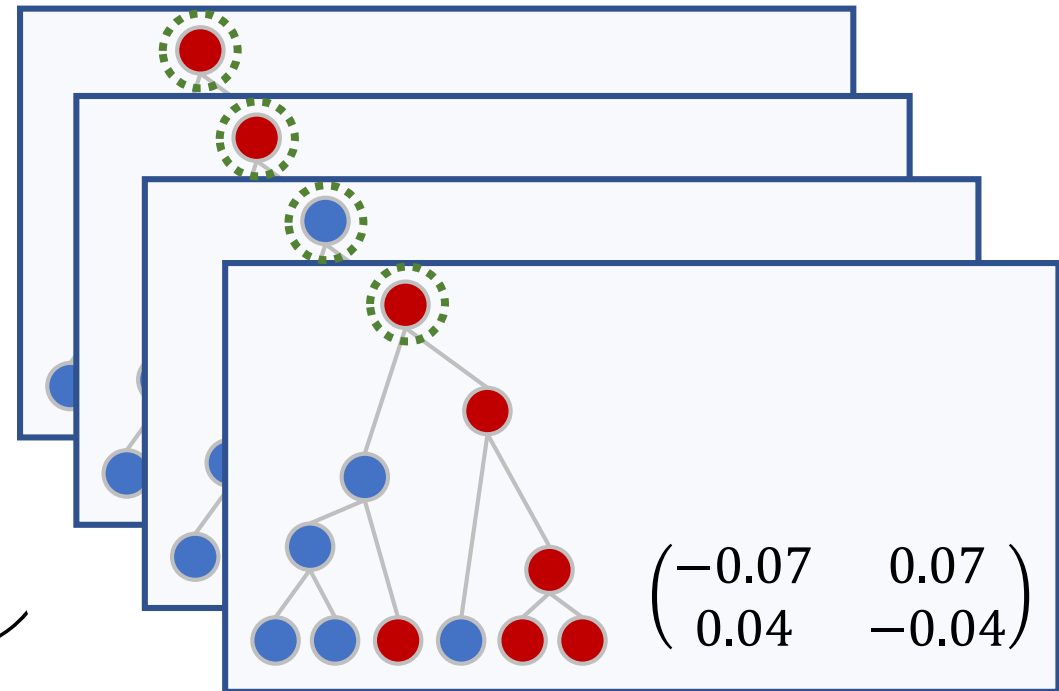


系統学的比較法 + 数理モデル

共通祖先が  だった確率は 0.25
 だった確率は 0.75



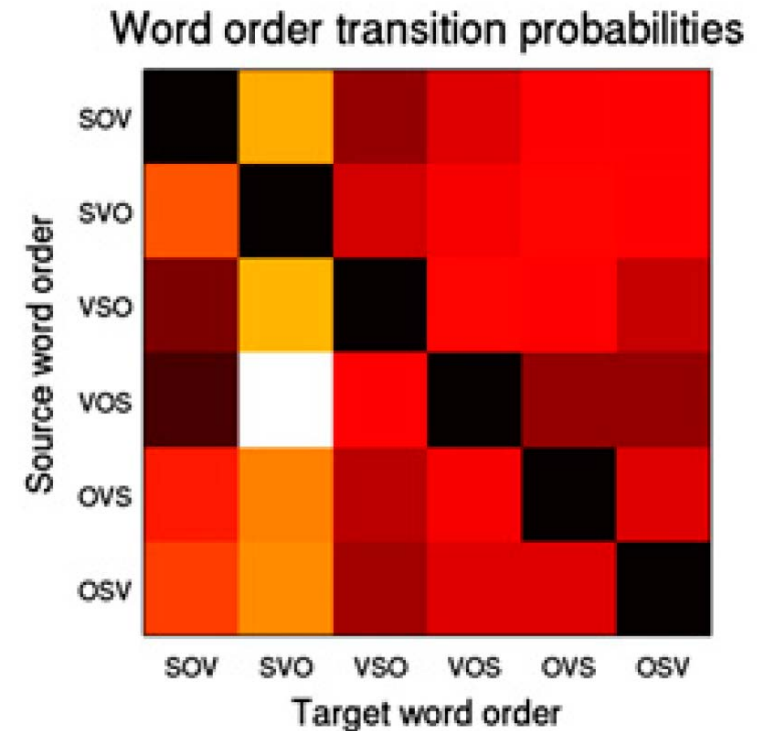
コンピュータを
使った
シミュレーション



系統学的比較法 + 数理モデル

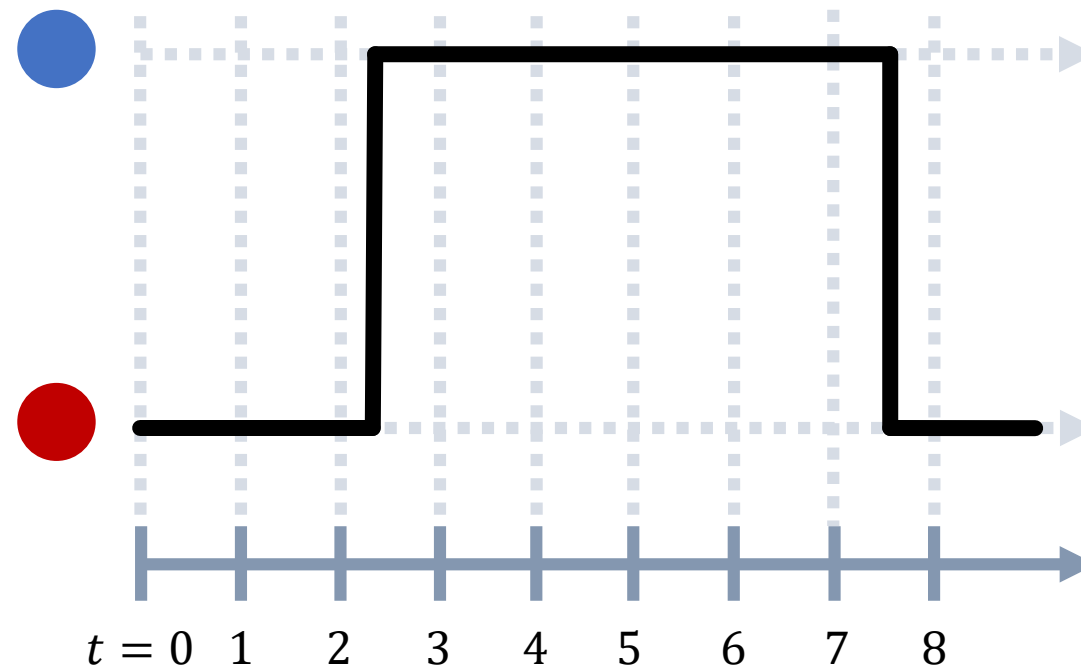
[Maurits and Griffiths, 2014]

- SOV基本語順（異なり数6）の時間変化を推定
- 世界の6個の系統樹（671言語）に適用
- 得られた知見：
 - SOVが時間的にもっとも安定だが、SVOと大差はない
 - SOVからSVOへの変化がその逆よりも起こりやすい
 - もしこれらの言語が単一の祖先を持つとしたら、その言語はSOVだった可能性がもっとも高いが、確率0.25と確信度は低い

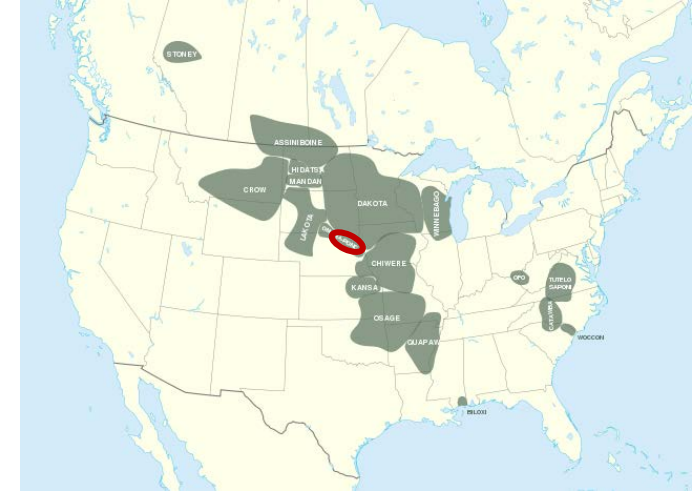


このモデルの限界

瞬間的な語形変化は
不自然



このモデルの限界



- SOV言語は実際には一枚岩ではない

日本語

男は羽飾りを取った

オマハ語

nú amá hiⁿqpé gǝ́iza-bi

オマハ語の基本語順はSOVだが、VにSやOが後続することも多い

日本語

そして男はその丈夫なひもを取った

オマハ語

ga^{n'} hájiñga áwanji ke é ǝ́iza-biamá nú aká (OVS)

このモデルの限界



- SOV言語は実際には一枚岩ではない

日本語

男は羽飾りを取った

オマハ語

nú amá hiⁿqpé gǝ́iza-bi

オマハ語の基本語順はSOVだが、VにSやOが後続することも多い

オマハ語のSOV語順は日本語よりも不安定かも

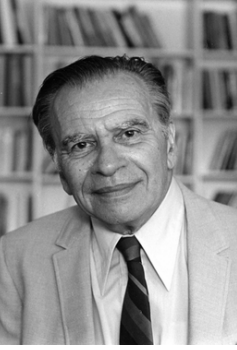
日本語

そして男はその丈夫なひもを取った

オマハ語

gaⁿ hájiñga áwanji ke é ǝ́iza-biamá nú aká (OVS)

含意的普遍性 [Greenberg, 1963]



ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

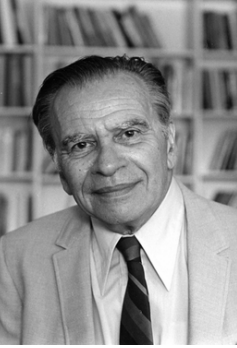
日本語

私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

含意的普遍性 [Greenberg, 1963]



ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहन्छु

日本語

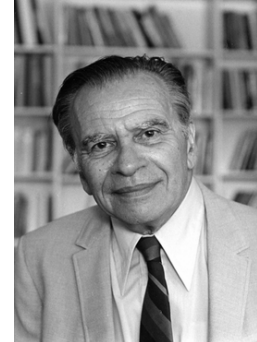
私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

含意的普遍性

[Greenberg, 1963]



ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहन्छु

日本語

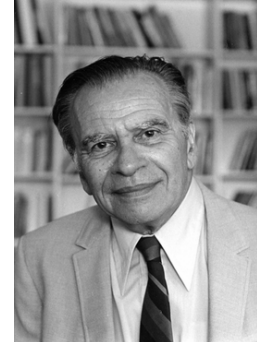
私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語

I want to try on a suit I saw in a shop that's across the street from the hotel

N名詞とRel関係節

- VOならNRel
- ReINならOV



含意的普遍性 [Greenberg, 1963]

ネパール語

म होटलको अगाडि भएको पसलमा हेरेको सुट लगाएर हेर्न चाहान्छु

日本語

私はホテルの向かいにあるお店で見たスーツを着てみたいです

英語

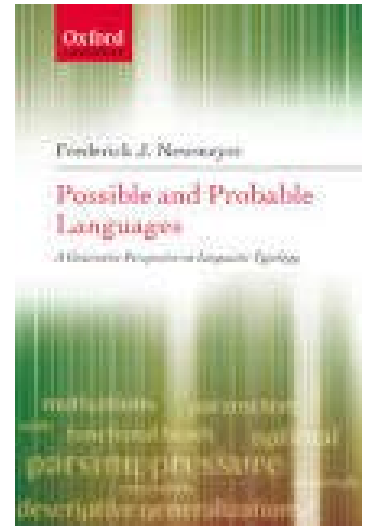
I want to try on a suit I saw in a shop that's across the street from the hotel

N名詞とRel関係節	NRel	RelN
• VOならNRel	○	×
• RelNならOV	○	○

ありえそうな言語とありえなそうな言語

- 各言語はSOV基本語順、NRel/ReIN語順やその他の**特徴**の列（並びは適当）で表現できる

ネパール語	SOV	ReIN	性2種	受動有	...
日本語	SOV	ReIN	性なし	受動有	...
英語	SVO	NRel	性なし	受動有	...



Possible and Probable Languages

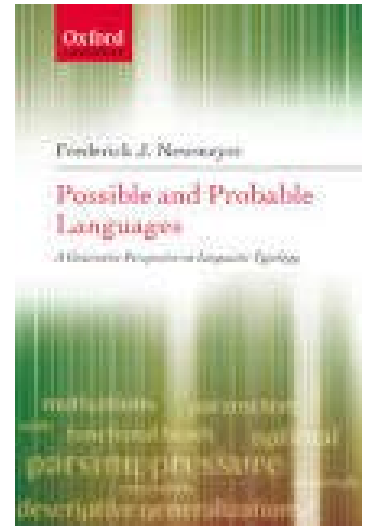
ありえそうな言語とありえなそうな言語

- 各言語はSOV基本語順、NRel/ReIN語順やその他の**特徴**の列（並びは適当）で表現できる

ネパール語	SOV	ReIN	性2種	受動有	...
日本語	SOV	ReIN	性なし	受動有	...
英語	SVO	NRel	性なし	受動有	...

- 観測された言語を一般化すれば、未知の言語がどの程度ありえそうかを推定できる

未知の言語A	SOV	ReIN	性3種	受動有	...
未知の言語B	VSO	ReIN	性2種	受動無	...



Possible and Probable Languages

ありえそうな言語とありえなそうな言語

- 各言語はSOV基本語順、NRel/ReIN語順やその他の**特徴**の列（並びは適当）で表現できる

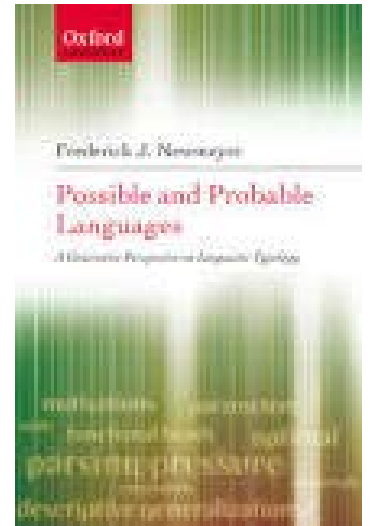
ネパール語	SOV	ReIN	性2種	受動有	...
日本語	SOV	ReIN	性なし	受動有	...
英語	SVO	NRel	性なし	受動有	...

- 観測された言語を一般化すれば、未知の言語がどの程度ありえそうかを推定できる

未知の言語A	SOV	ReIN	性3種	受動有	...
未知の言語B	VSO	ReIN	性2種	受動無	...

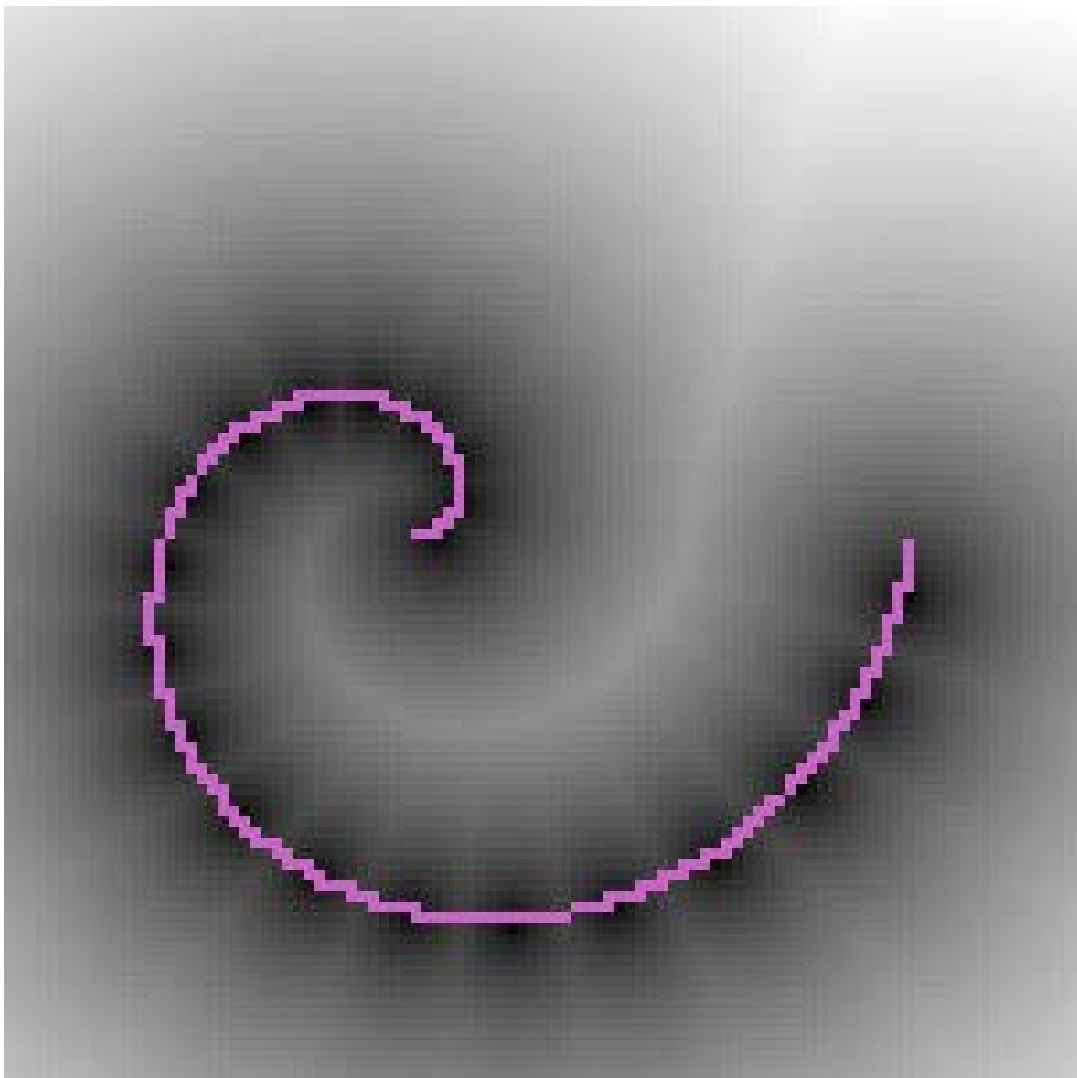
ありえそう

ありえなそう



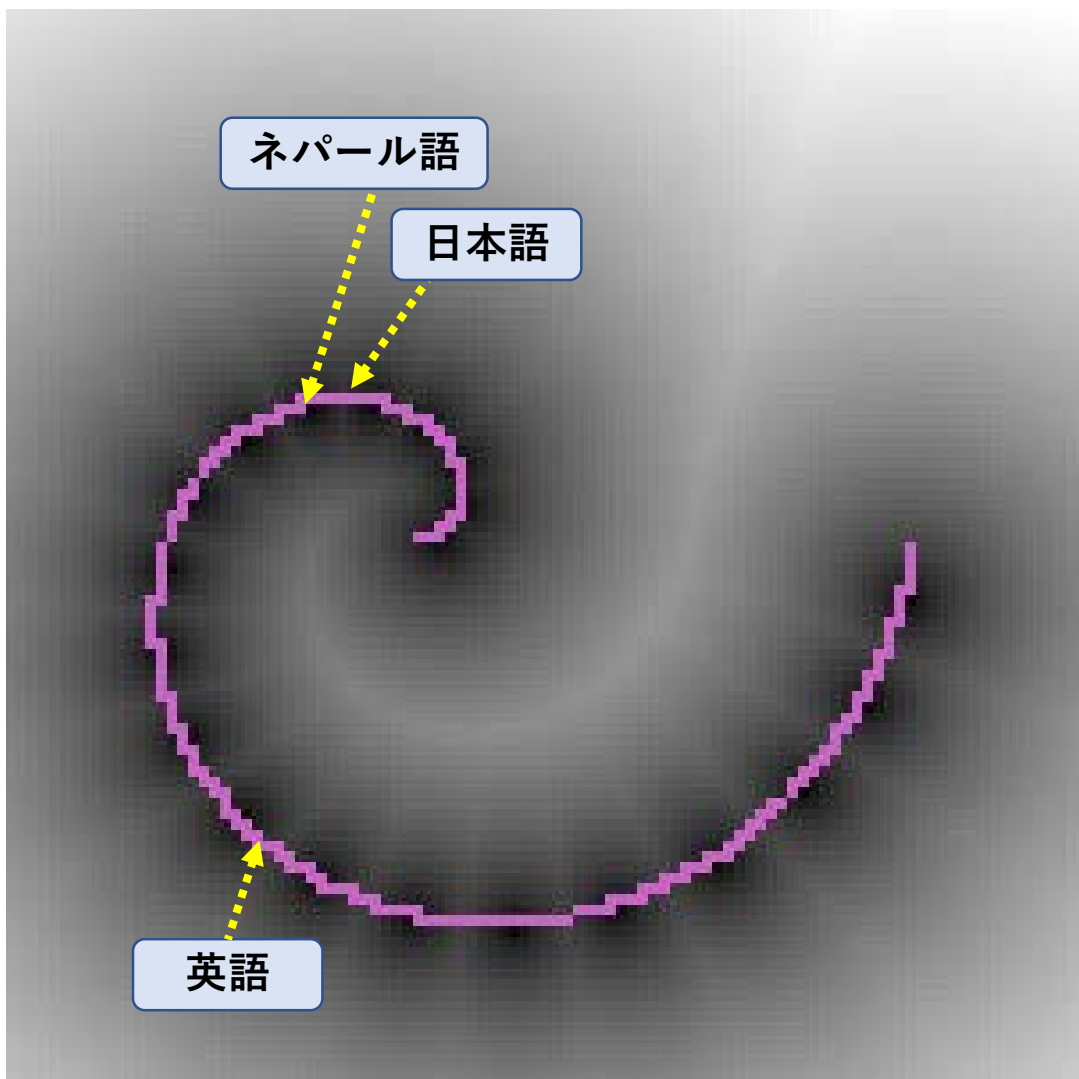
Possible and Probable Languages

数理的な一般化: データ多様体



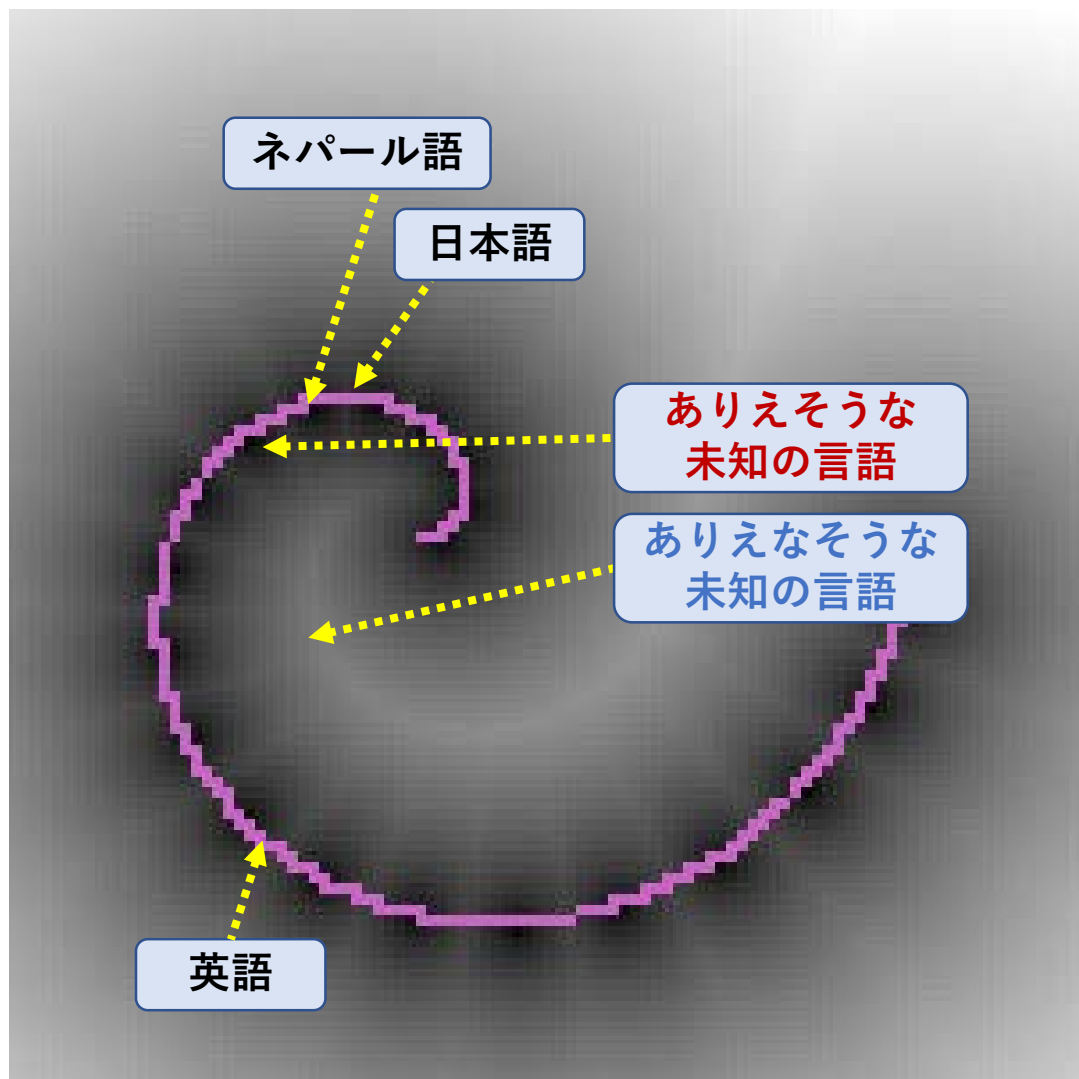
- 2次元空間中でらせんの上でデータが観測されている
- 観測データを一般化することで、任意のデータ点がどれくらいありえそうかを推測したい

数理的な一般化: データ多様体



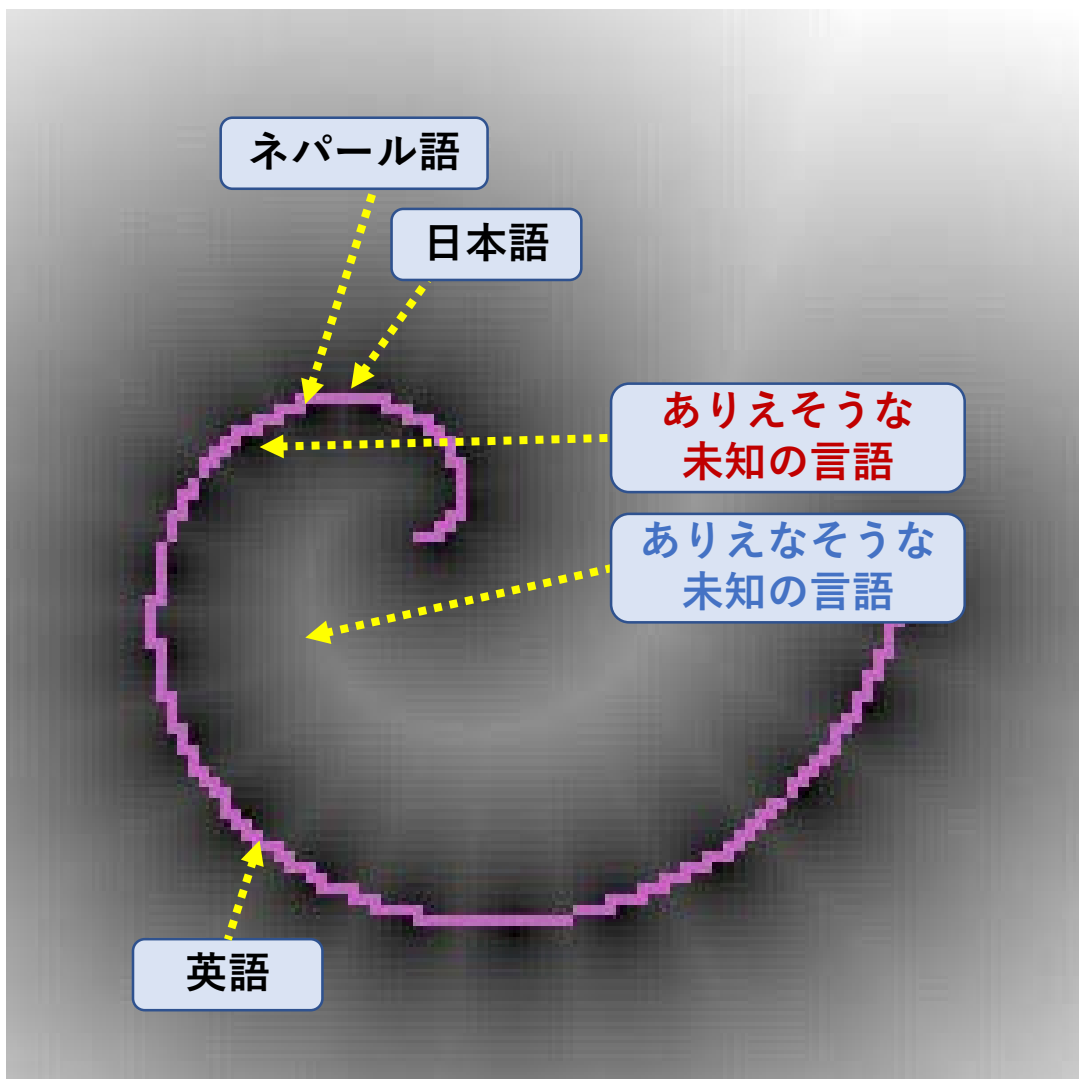
- 2次元空間中で**らせんの上**でデータが観測されている
- 観測データを一般化することで、任意のデータ点がどれくらいありえそうかを推測したい

数理的な一般化: データ多様体



- 2次元空間中で**らせんの上**でデータが観測されている
- 観測データを一般化することで、任意のデータ点がどれくらいありえそうかを推測したい

数理的な一般化: データ多様体



- 2次元空間中で**らせんの上**でデータが観測されている
- 観測データを一般化することで、任意のデータ点がどれくらいありえそうかを推測したい



多様体学習の定石は表現学習

[Murawaki, 2019]

特徴列

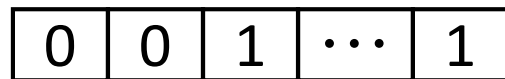
SOV	ReIN	性なし	受動有	...
-----	------	-----	-----	-----

多様体学習の定石は表現学習

[Murawaki, 2019]

- 特徴列は潜在表現から確率的に生成されたと仮定

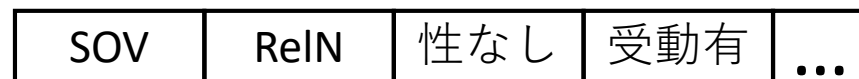
潜在表現



確率的
生成 ↓

↑ 確率的
推論

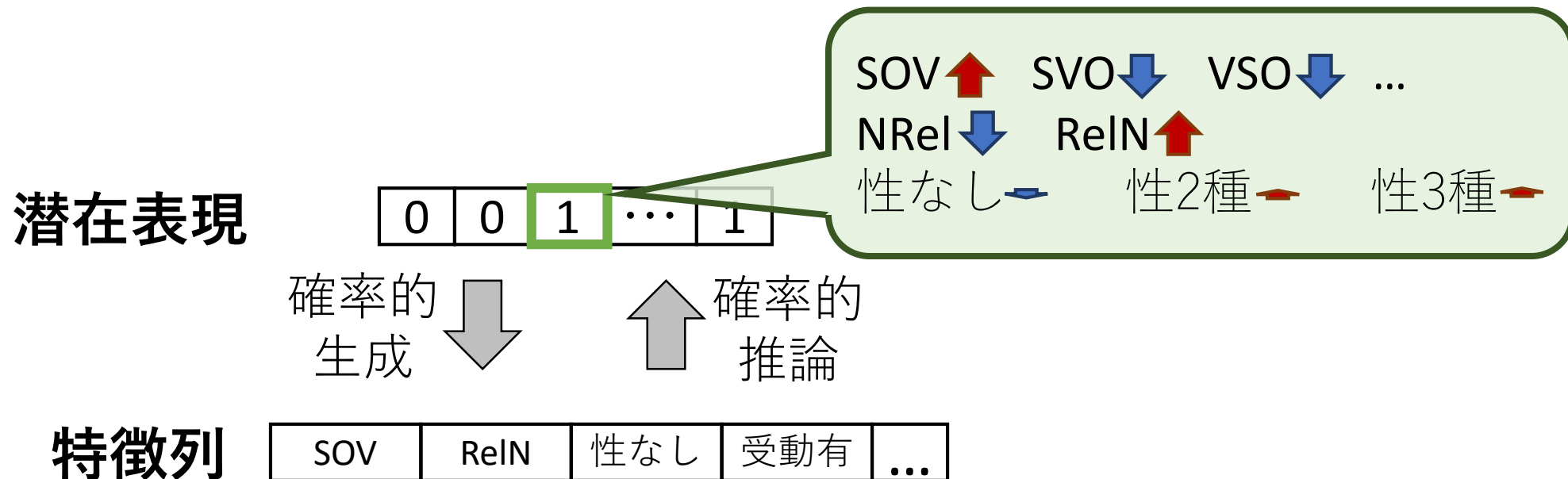
特徴列



多様体学習の定石は表現学習

[Murawaki, 2019]

- 特徴列は潜在表現から確率的に生成されたと仮定
- ありえそうな言語を高い確率で、ありえなそうな言語を低い確率で生成するような生成モデルを学習

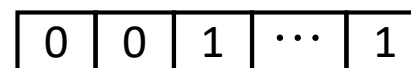


系統学的比較法 + 数理モデル + 潜在表現

[Murawaki, 2019]

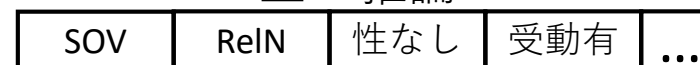
ステップ1:
各言語を潜在表現に変換

潜在表現



↑ 確率的
推論

特徴列



系統学的比較法 + 数理モデル + 潜在表現

[Murawaki, 2019]

ステップ1:
各言語を潜在表現に変換

潜在表現

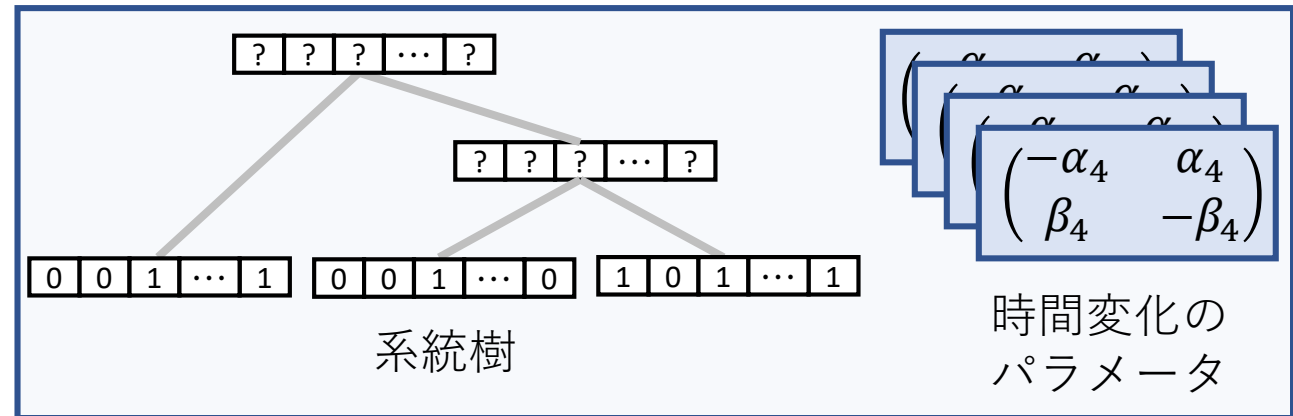
0 0 1 ... 1

↑ 確率的
推論

特徴列

SOV ReIN 性なし 受動有 ...

ステップ2:
系統樹を使ったシミュレーション



系統学的比較法 + 数理モデル + 潜在表現

[Murawaki, 2019]

ステップ1:
各言語を潜在表現に変換

潜在表現

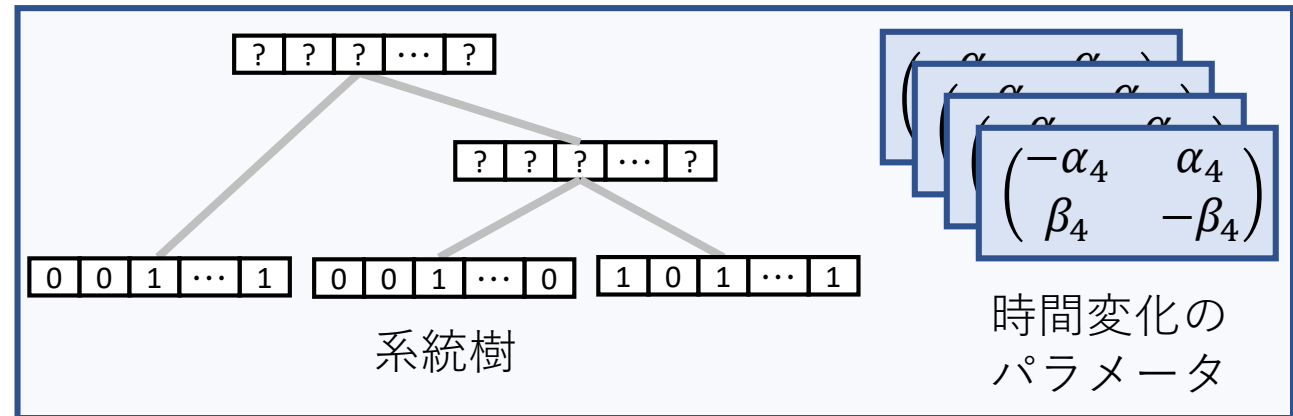
0 0 1 ... 1

確率的
推論

特徴列

SOV ReIN 性なし 受動有 ...

ステップ2:
系統樹を使ったシミュレーション



ステップ3:
時間変化のパラメータを使って
推定された言語を特徴列に変換

潜在表現

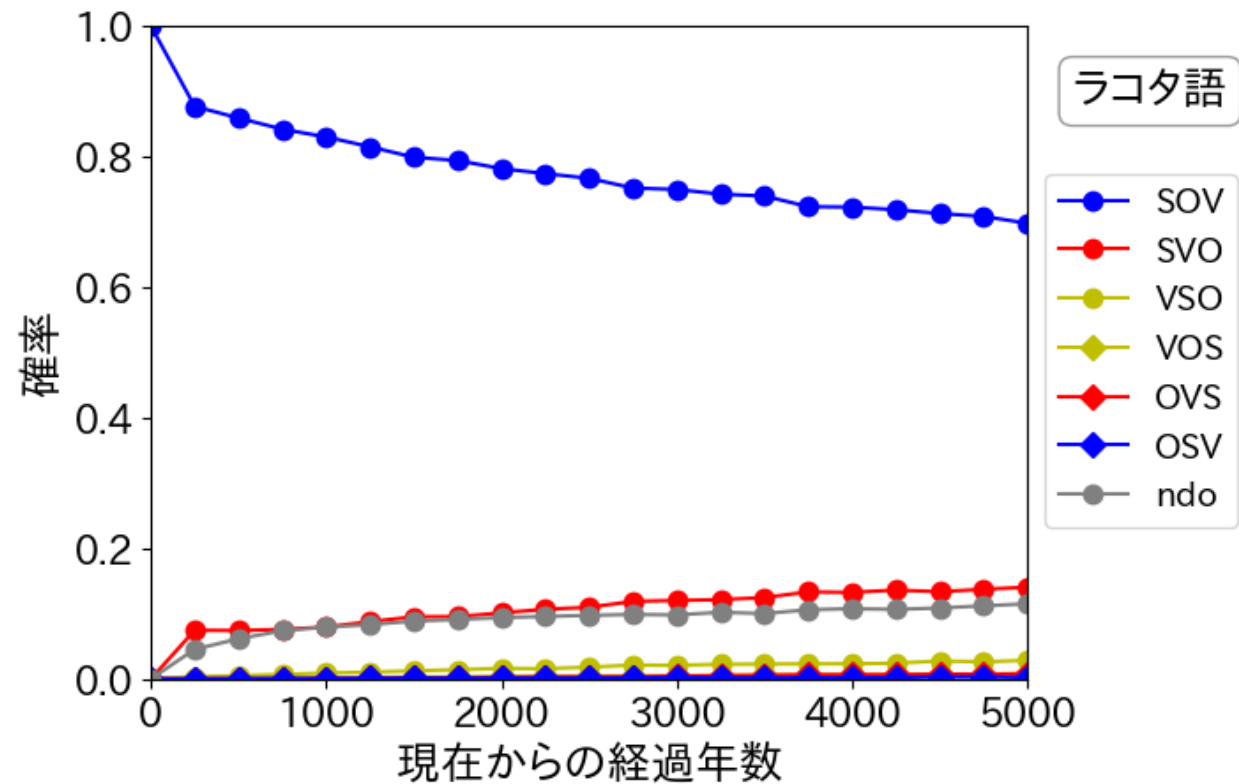
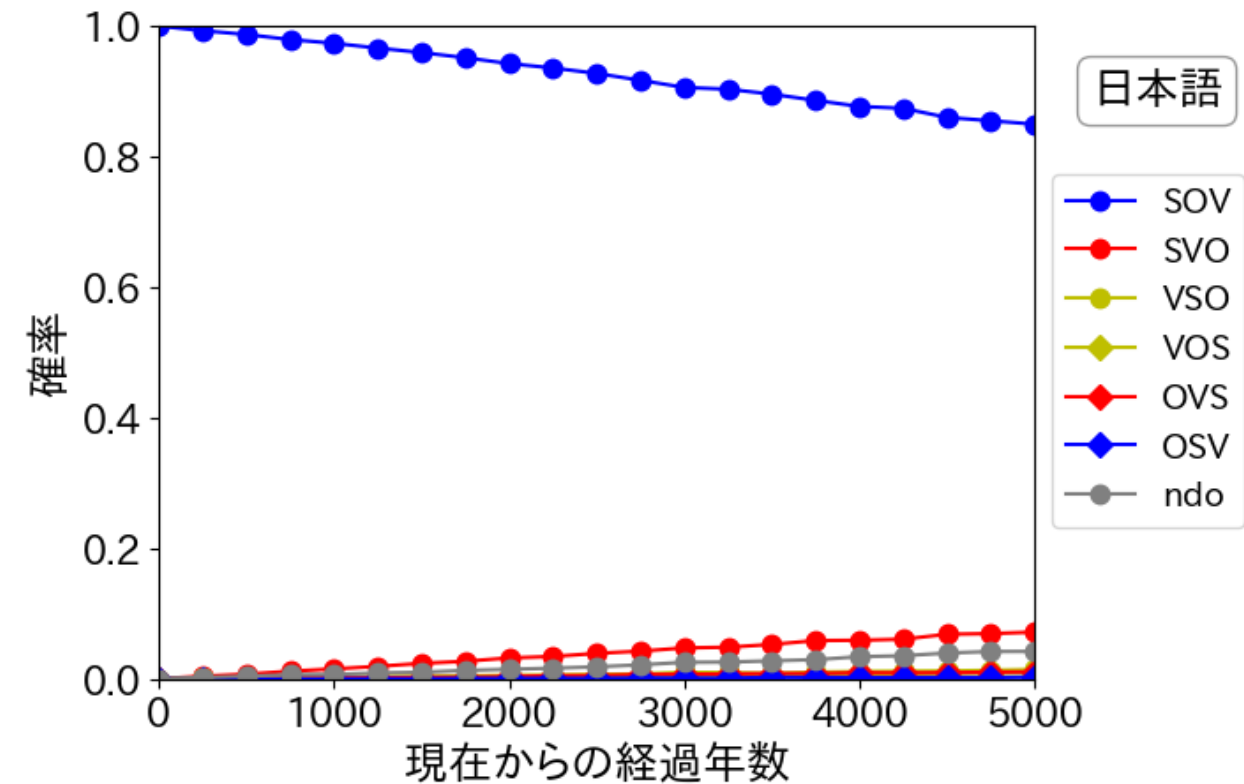
1 0 1 ... 0

確率的
生成

特徴列

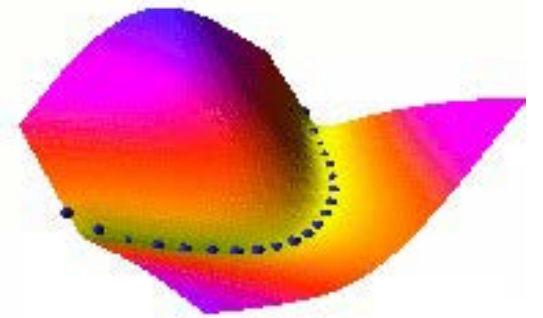
SOV ReIN 性2種 受動無 ...

SOV言語の将来をシミュレーション



今後の展望

- これまで使っていたデータ (WALS) は質的に限界があるので、新しいデータ (Grambank) が公開されるまで様子見
- ありえそうな言語群が構成する特徴列部分空間はどんな構造を持っているか？
 - 祖先から子孫に至るまでにどのような経路を通るか？
- 日本語の系統の解明に使えるか？
 - 構造的特徴は系統解明の手がかりとしては信頼できないとみなされてきた
 - ありえそうな言語に着目した精緻な分析ならあるいは？
- 基本語順以外の、よりマニアックな特徴の解析



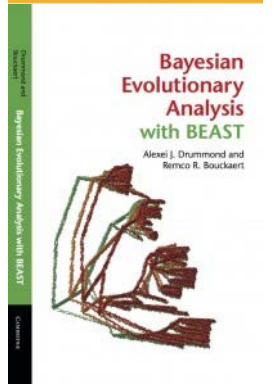
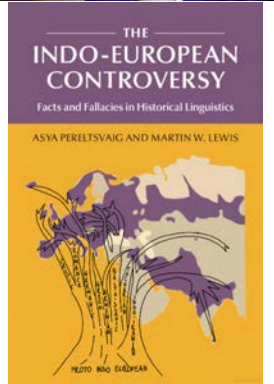
文理融合？



- 問題の発見と性質の記述: 人文系
- 問題を解くためのデータ構築: 人文系
- 問題を解くための数理的道具の提供: 情報系
- 人間とコンピュータが補完的
 - 人間による論証が難しい問題をコンピュータに解かせる

数理モデルの導入に 言語学者は必ずしも好意的ではない

- 進化生物学者のラッセル・グレイが2000年代にDNAの分析手法を言語に転用したのが流行のきっかけ
- グレイは言語学者の間できわめて評判が悪い仮説(インド・ヨーロッパ語族のアナトリア起源説)を推している
- 進化生物学分野で開発された複雑な統計モデルは言語学者にはまったく理解できない
- データを雑に扱っていることは言語学者にも理解できる



新旧の対立は古くからの課題だが…



新しい科学的真理が勝利をおさめるのは、反対者を納得させ彼等の蒙を啓くことによってではなく、反対者が徐々に死に絶え、新しい世代が初めから真理に慣れ親しむことによってである。——マックス・プランク (1948)

新旧の対立は古くからの課題だが…



新しい科学的真理が勝利をおさめるのは、反対者を納得させ彼等の蒙を啓くことによってではなく、反対者が徐々に死に絶え、新しい世代が初めから真理に慣れ親しむことによってである。——マックス・プランク (1948)

- “反対者” が予算・人事等を握っているのが世の常
- 技術革新が加速する現在、世代交代を悠長に待ってられない
- 人文系不要論が蔓延するなか、組織拡大は期待薄

方法論としての情報学と 組織としての情報学



情報学はすべての学問領域において不可欠になり、情報学自体いずれ消滅



情報学は旧来の文系・理系とも異なる第3の領域

- この2つの説明は時間軸を導入すれば矛盾しない
- 情報学がすべてを覆うまでの過渡期には、独立組織としての情報学が変化を加速させる！

