山本研究室
（知能計算分野）　　知-5

教授 山本 章博　　特定准教授 市瀬 夏洋

研究室　：総合研究7号館 324(教授室), 323, 325, 327

Web　：　http://www.iip.ist.i.kyoto-u.ac.jp/

Email　：　akihiro@i.kyoto-u.ac.jp

当研究室では，機械学習理論を中心にして人間の高次推論機構の性質を解明し，またそれを用いて，与えられたデータから適切な情報を取出すための計算機構やソフトウェアを構築することを目標に研究を行っている．さらにこれらの研究を．生命情報学などにおけるデータ集合からの知識発見などへの応用し，数理論理学や計算数学との関係の解明へと展開している．

第２次AIブームの成果を第３次AIブームに生かす

研究の特徴

- 情報学基礎理論としての機械学習
  - 機械学習を基盤にし，論理や計算に立脚した新しい知的行為の基礎理論を求める
- 特に，データを読むための計算を対象にする．
  - プログラムを構築するには，構築のための“論理”が必要であるように，データを読むためにもそれ用の“論理”が必要
  - データ構造と代数的手法，数理論理学的手法特に“証明”を活用

こだわり所

第２次AIブーム：数理論理を基盤

知識表現と証明

第３次AIブーム：ニューラルネットを基盤

大規模データからの機械学習

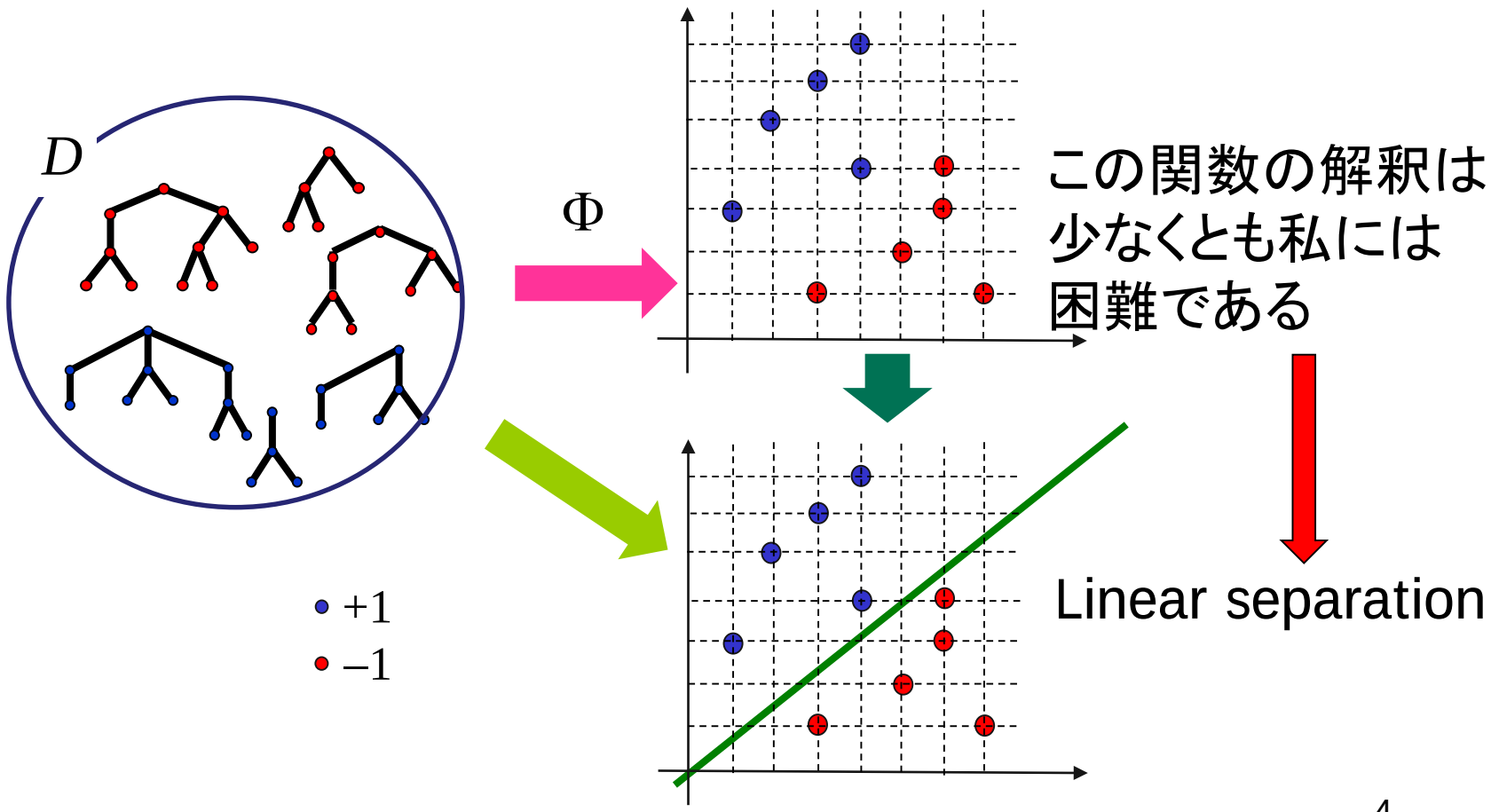
- データ構造に着目した機械学習
  - 木構造：ソースコードの構文解析木，HTMLのDOM-Tree
  - ２部グラフ: トランザクション(商品と顧客)
  - グラフ: 知識グラフ(Knowledge Graph)
- 説明可能な機械学習
  - 数学・数理論理は「推論の表現と正しさ」を数学的に保証
  - 機械学習でも「推論の表現と正しさ」を数学的に保証することが重要

    機械学習アルゴリズムの正しさとは何か？

    ・深層学習が有効なのは，学習の結果に説明を求められない場合

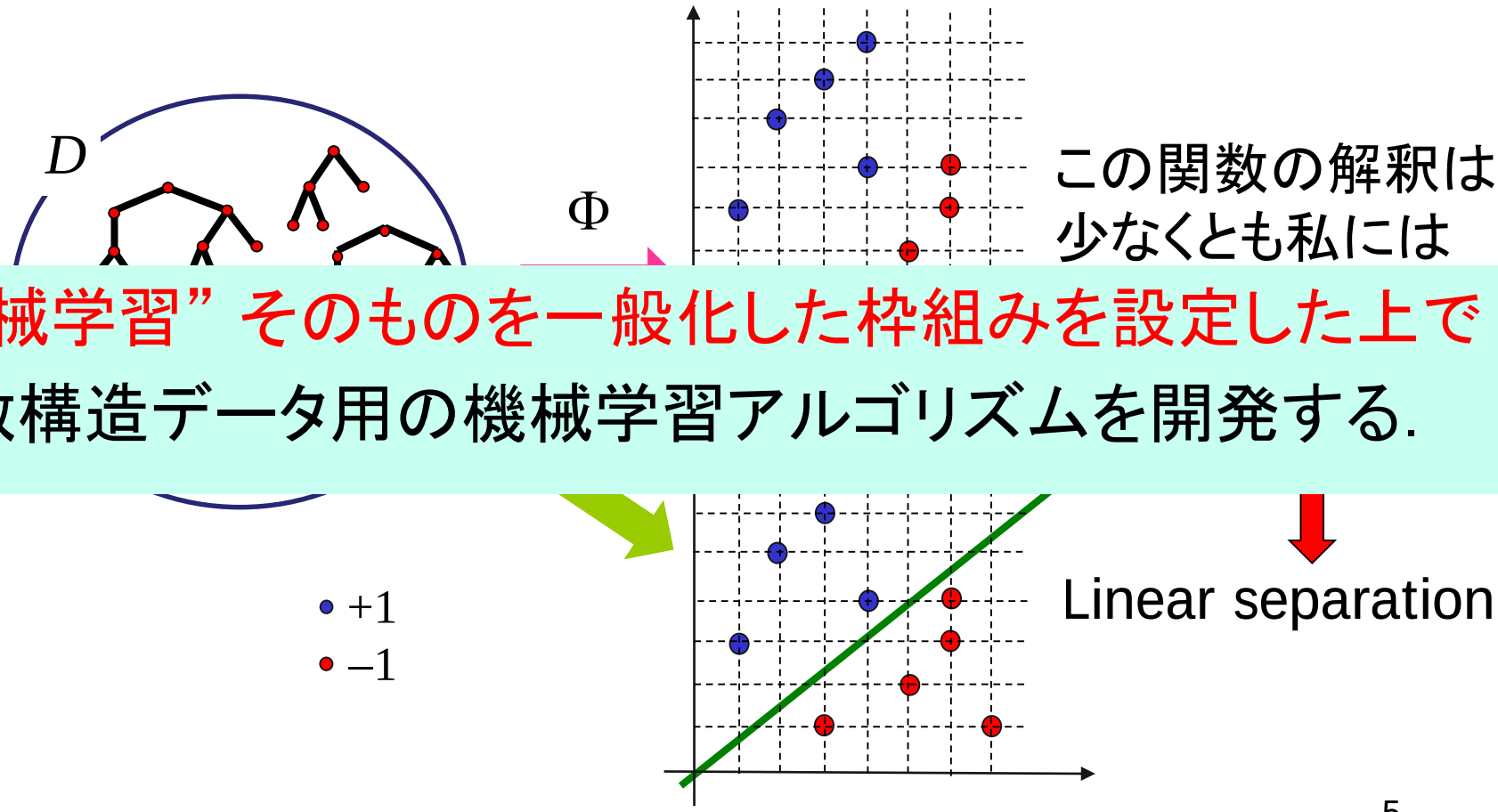
離散構造データからの機械学習

- 離散データを実数値ベクトルに変換して機械学習アルゴリズムを適用する手法も考えられる



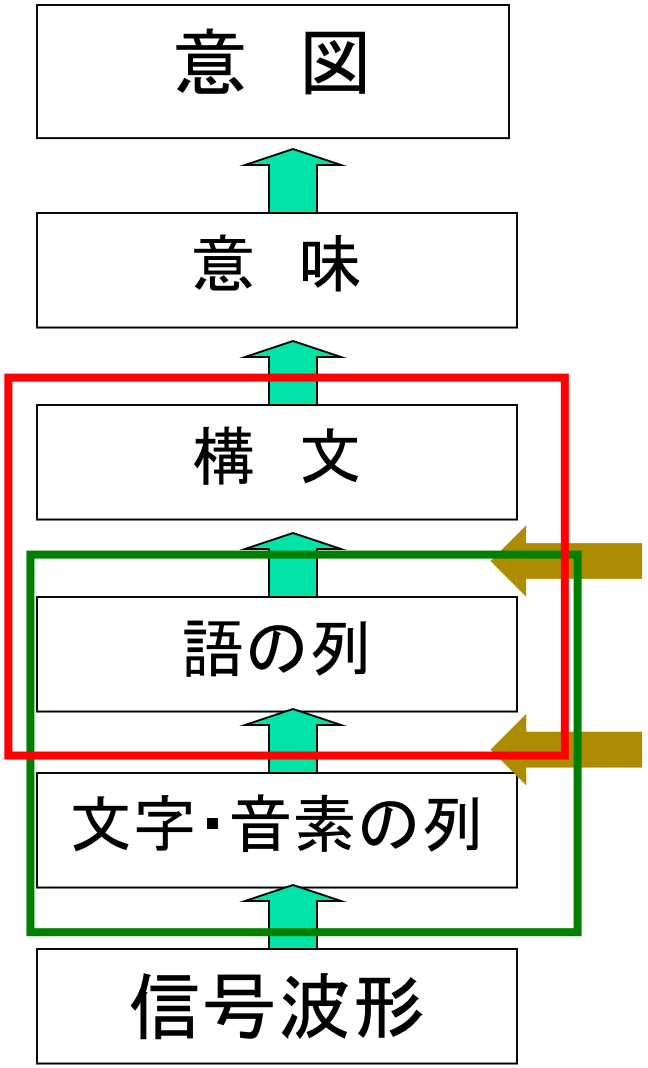
離散構造データからの機械学習

- 離散データを実数値ベクトルに変換して機械学習アルゴリズムを適用する手法も考えられる



“機械学習”そのものを一般化した枠組みを設定した上で
離散構造データ用の機械学習アルゴリズムを開発する.

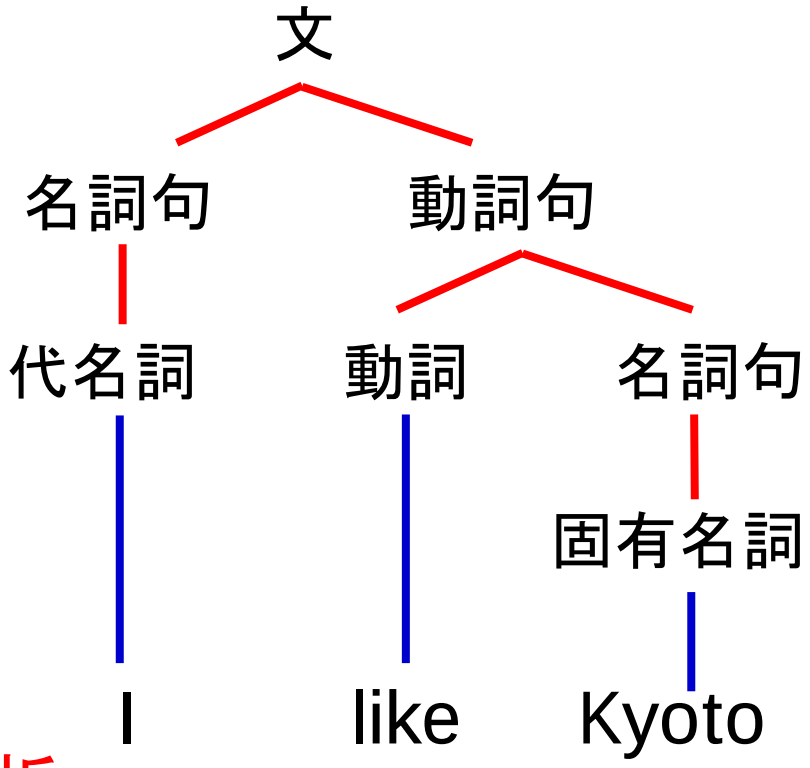
構造データの出どころ

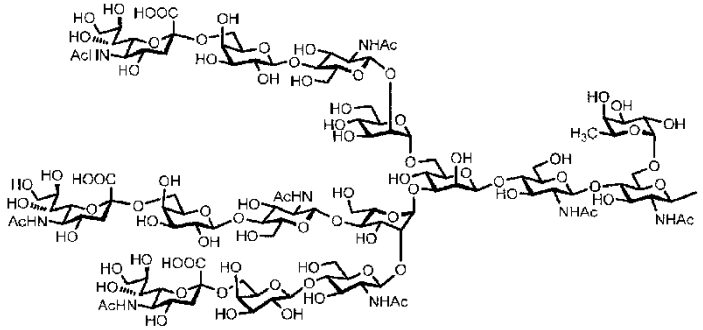


構文解析

形態素解析
(自然言語)

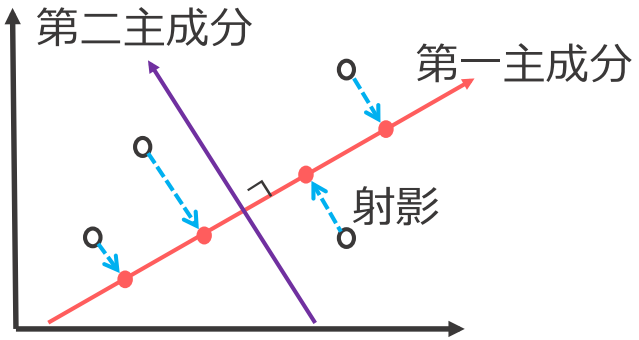
字句解析
(人工言語)



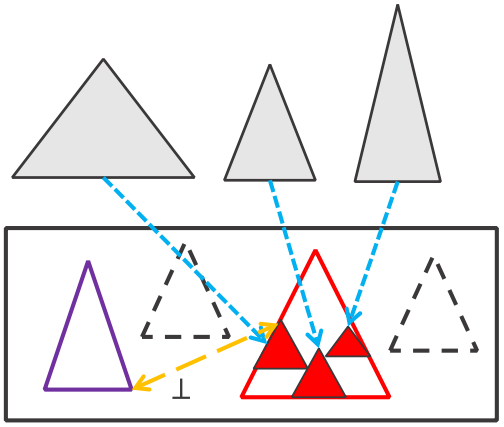


木構造と主成分分析の組み合せ

主成分分析(PCA): 情報量の損失が少ない低次元空間を求める.

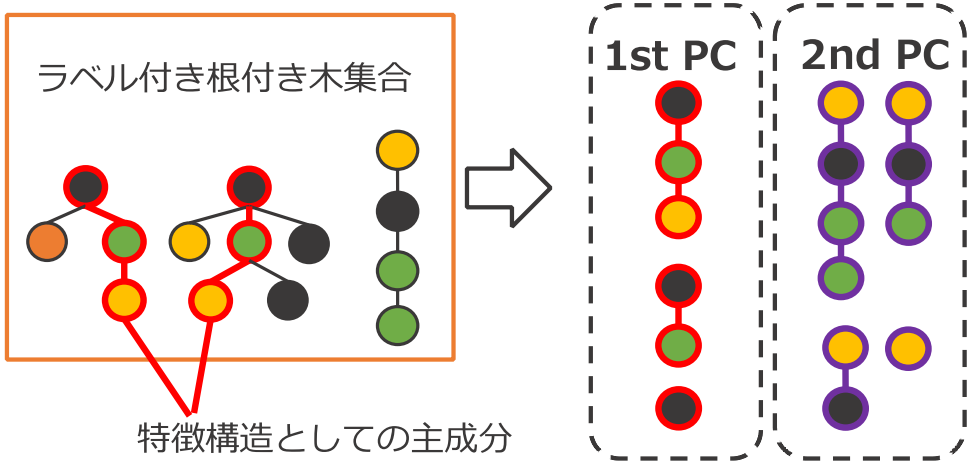


主成分分析



Tree PCA

- 木構造から特徴的な構造を抽出



特徴構造としての主成分

構造データ間の距離

文字列 s_0= abaaaab と近いのはどっち？

s_1= aababa　or s_2= abaaaaaab

- “機械学習の教科書”にあるN-gramを使ってベクトル化

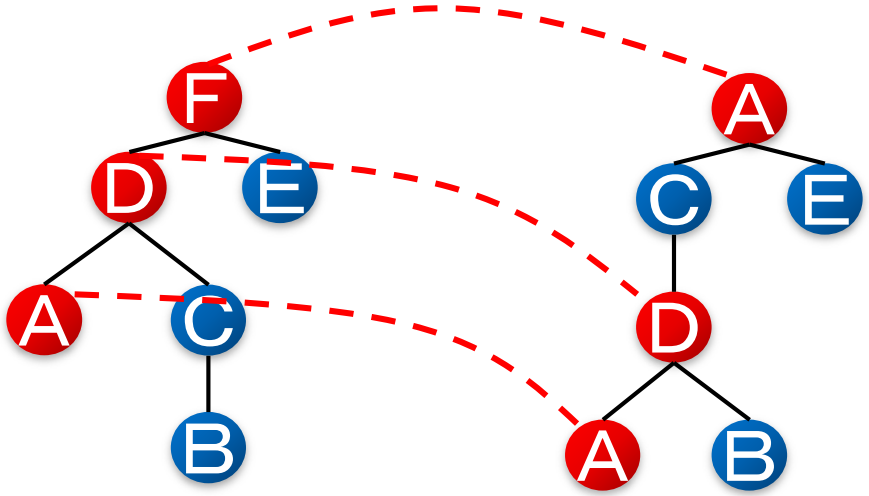
$$v_0 = \Phi(\text{abaaab}) = (1, 1, 1, 0, 1, 0, 0, 0),$$
$$v_1 = \Phi(\text{aababa}) = (0, 2, 1, 0, 0, 1, 0, 0),$$
$$v_2 = \Phi(\text{abaaaaaab}) = (4, 1, 1, 0, 1, 0, 0, 0),$$
$$\| v_1 - v_0 \|^2 = 4 < \| v_2 - v_0 \|^2 = 7$$

- 構造体特有の距離である編集距離を用いると

$$d(s_0, s_1) = 4 > d(s_0, s_2) = 3$$

木構造データ間の距離と直交性

- 木構造データ間の編集距離



  - 木構造は2次元的に広がるため，バリエーションが多い

- 木構造間の直交性：H30年度卒論で厳密に定式化

木構造間編集距離の計算量

	編集距離算出		局所構造算出	
	Tai マッピング	制約 マッピング	Tai マッピング	制約 マッピング
順序木	$O(n^3)$ [Demine 06]	$O(n^2)$ [Zhang 95]	$O(n^4)$ [Ouangraoua et al. 07]	$O(n^2 d \log d)$
無順序木	MAXSNP 困難	$O(n^2 d \log d)$ [Zhang 96]	MAXSNP 困難	$O(n^2 d \log d)$ [Ferraro et al. 05]
前順序木	MAXSNP 困難	$O(n^2 d \log d)$ [Ouangraoua et al. 09]	MAXSNP 困難	$O(n^2 d \log d)$

n:ノード数　d:最大次数

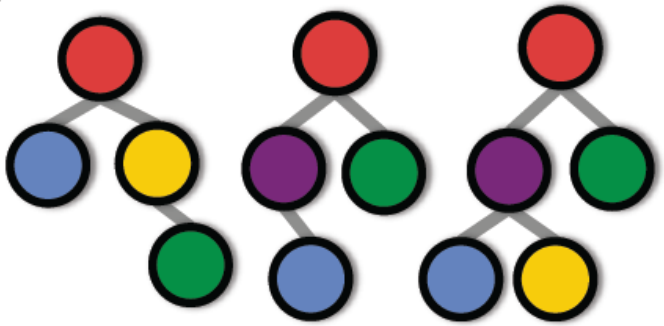
※国際会議 GABA2014@横浜で発表

IPソルバを用いた木構造間距離の高速計算

- 近年では高速な汎用ソルバ(SAT/IPなど)が利用可能

  ⇒ 高速なソルバをターゲットとする”コンパイラ”を開発

Input



IP Problem

Maximize: $\sum m_{i,j} - 2o_{|Ti|,|Ti+1|}$

Subject to:

$o_{1,1} \le o_{1,2}, \quad o_{1,1} \le o_{2,1}, \quad \dots$

IP Solver

Output

IP Solution

$m_{1,1} = 1,$

$m_{1,2} = 0, \dots$

IP = LP + integer restriction

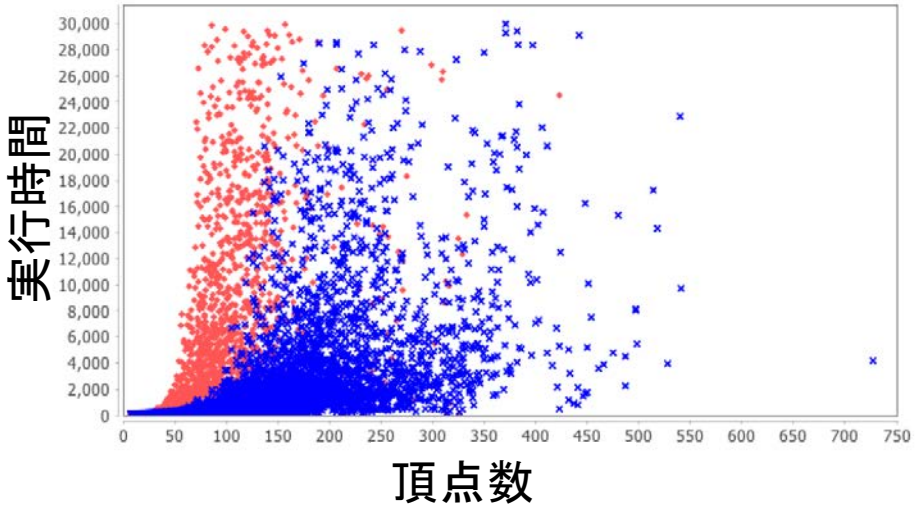
- Complexity is NP-hard
- But efficient solvers exist

さらに高速なIP定式化

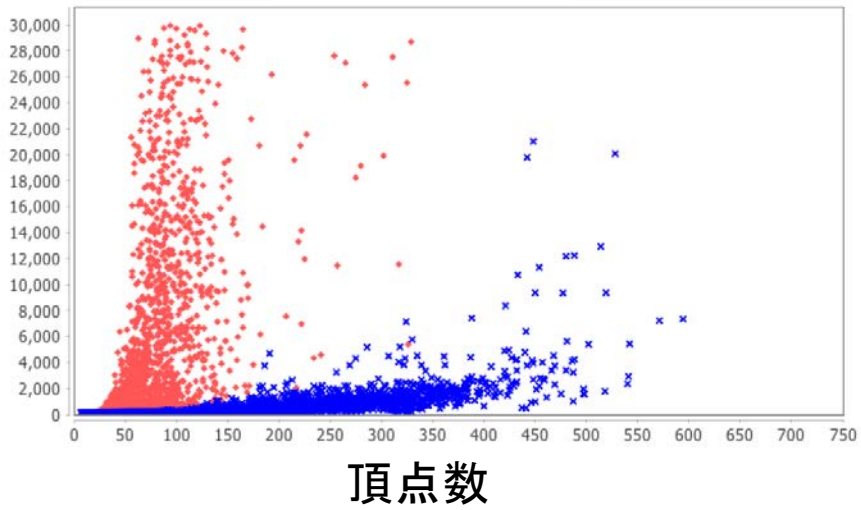
- 動的計画法を用いることで，コンパクトなIP定式化に成功

	編集距離		断片距離	
	既存手法	提案手法	既存手法	提案手法
部分問題数	1	n^2	1	1
制約式数	$O(n^4)$	$O(n)$	$O(n^4)$	$O(n)$

編集距離



断片距離



※国際会議COCOA2017@上海で発表

アラインメント距離のIP定式化

- アラインメント距離のIP定式化に成功
  - アラインメント距離：編集距離の変種

	既存手法	提案手法
方法	動的計画法 計算量: $O(6^d dn^2)$	IP 制約式数: $O(n^4)$
d が小さい	速い	遅い
d が大きい	遅い	速い

実行時間

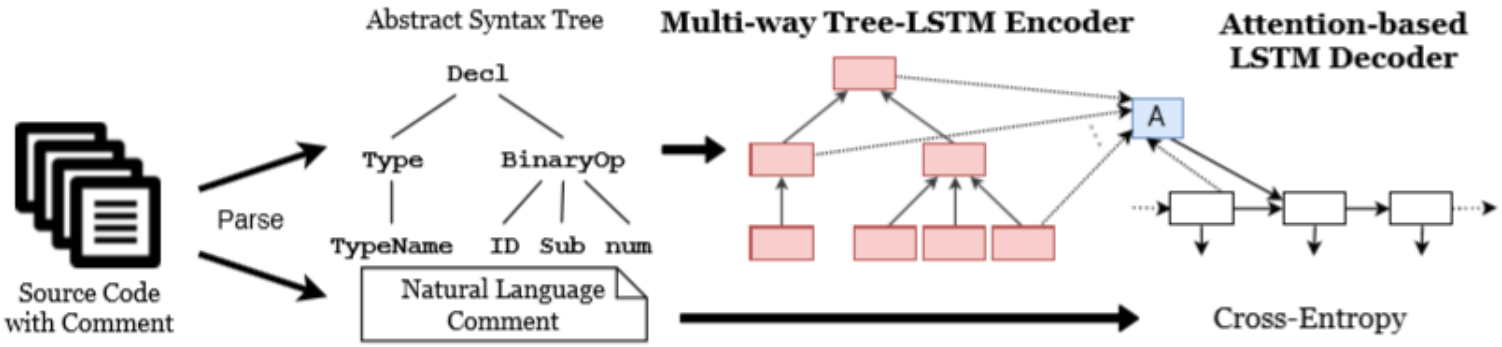
n	d	既存手法	提案手法
15-19	0-3	1ms以下	728ms
	8-11	7165ms	374ms

n: ノード数　　d: 最大次数

※第106回人工知能基本問題研究会@指宿で発表

NNによるソースコード要約(修士)

- Tree-LSTMを理論的整備してソースコード要約へ応用



Source code	`public boolean more() throws JSONException {`
	`  next();`
	`  if (end()) {`
	`  return false;`
	`  }`
	`  back();`
	`  return true;`
	`}`

Gold	Determine if the source string still contains characters that next() can consume
Generated	Determine if the source string still contains characters that next() can consume

Source code	`@JsonIgnore`
	`public boolean isDeleted(){`
	`  return state.equals(Experiment.State.DELETED);`
	`}`

Gold	Signals if this experiment is deleted
Generated	Returns true if this session has been deleted

※日立-京大ラボとの共同研究 ※国際会議 IJCNN@BudaPestで発表

Tree-LSTMを利用した情報統合

- 自然言語データに対する深層学習モデルであるLSTMをベースに木構造向けモデルを設計
- 同一目的だがサイトによって表現形式が異なる情報の統合に応用
  - 大学のシラバスなど

京都大学のシラバスサイト

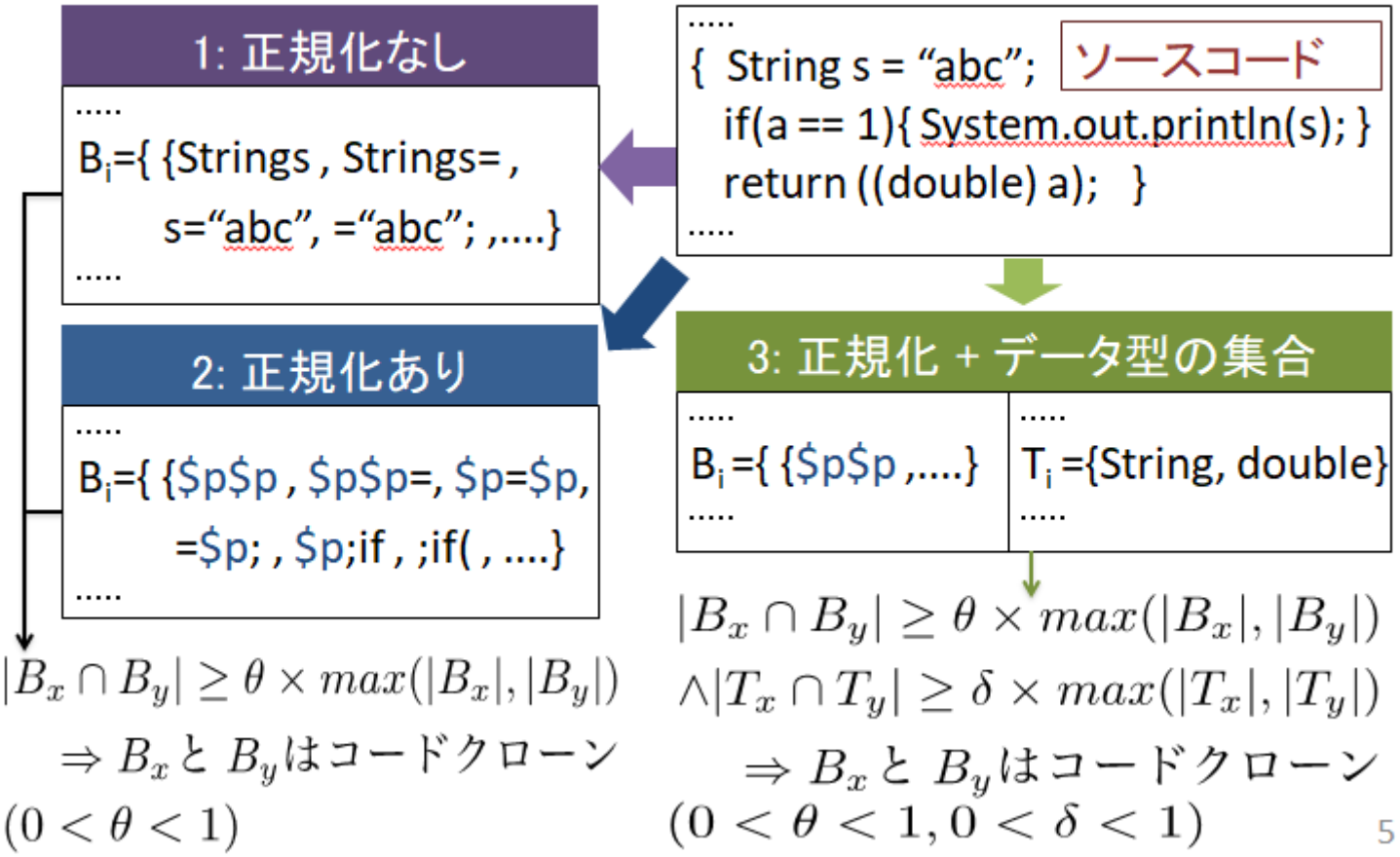
科目名	単位数	担当教員	曜時限
数学入門	2	山田	火1
英語	4	鈴木	水3
統計 I	2	佐藤	木4
方言学	2	田中	月2
グラフ理論	2	平山	水6
データ分析	1	井上	金1

東京大学のシラバスサイト

※第115回人工知能基本問題研究会で発表

ソースコード・クローンの検出

コードをトークンのN-gramで多重集合化 + CCFinderの正規化



※日立−京大ラボとの共同研究
※第199回ソフトウェア工学研究会@帯広で発表
※情報処理学会CS領域奨励賞受賞

定理自動証明の新しい側面

- 第2次AIブームの基盤であった定理自動証明を現代の視点から再考察する

可換マッチング問題

- $f(t_1,t_2) = f(t_2,t_1)$ とみなす(可換)

Input

s: 項

t: 変数を含まない項

Output

s を t に等しくするような変数への代入は存在するか？

Example

s = f(x, a)　　t = f(a, b)

代入 x/b

f(b, a)

答え Yes

5

Akutsu+の予想: sとtを等しくするような代入の個数は高々 2^{k-1} 個に対して，1つの場合を除いて成立することを証明した

※第112回(予稿)第113回(口頭)人工知能基本問題研究会で発表

BDDを用いたILPの解の列挙

ILP問題

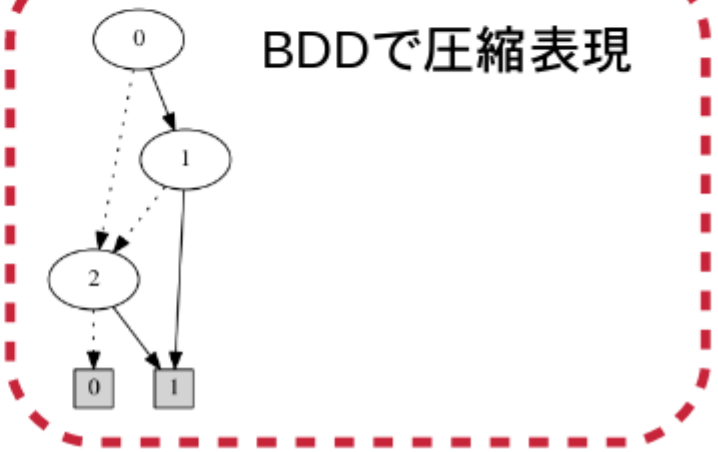
$\mathcal{E}^+ = \{e(0), e(s^2(0))\}$
$\mathcal{E}^- = \{e(s(0))\}$
$\mathcal{B} = \{\}$

$\Sigma \cup \mathcal{B} \models \mathcal{E}^+$
$\Sigma \cup \mathcal{B} \not\models \mathcal{E}^-$

大量に生成される仮説

$\Sigma = \{e(0), e(s^2(0))\}$
$\Sigma = \{e(0), e(s^2(x)) \leftarrow e(x)\}$
$\vdots$

命題変数を生成



- 帰納論理プログラミング(ILP)
  一階述語論理の技法で分類問題を解決
- 二分決定グラフ(BDD)
  ブール関数をコンパクトに表現

※国際会議 ILP 2018@イタリア で発表

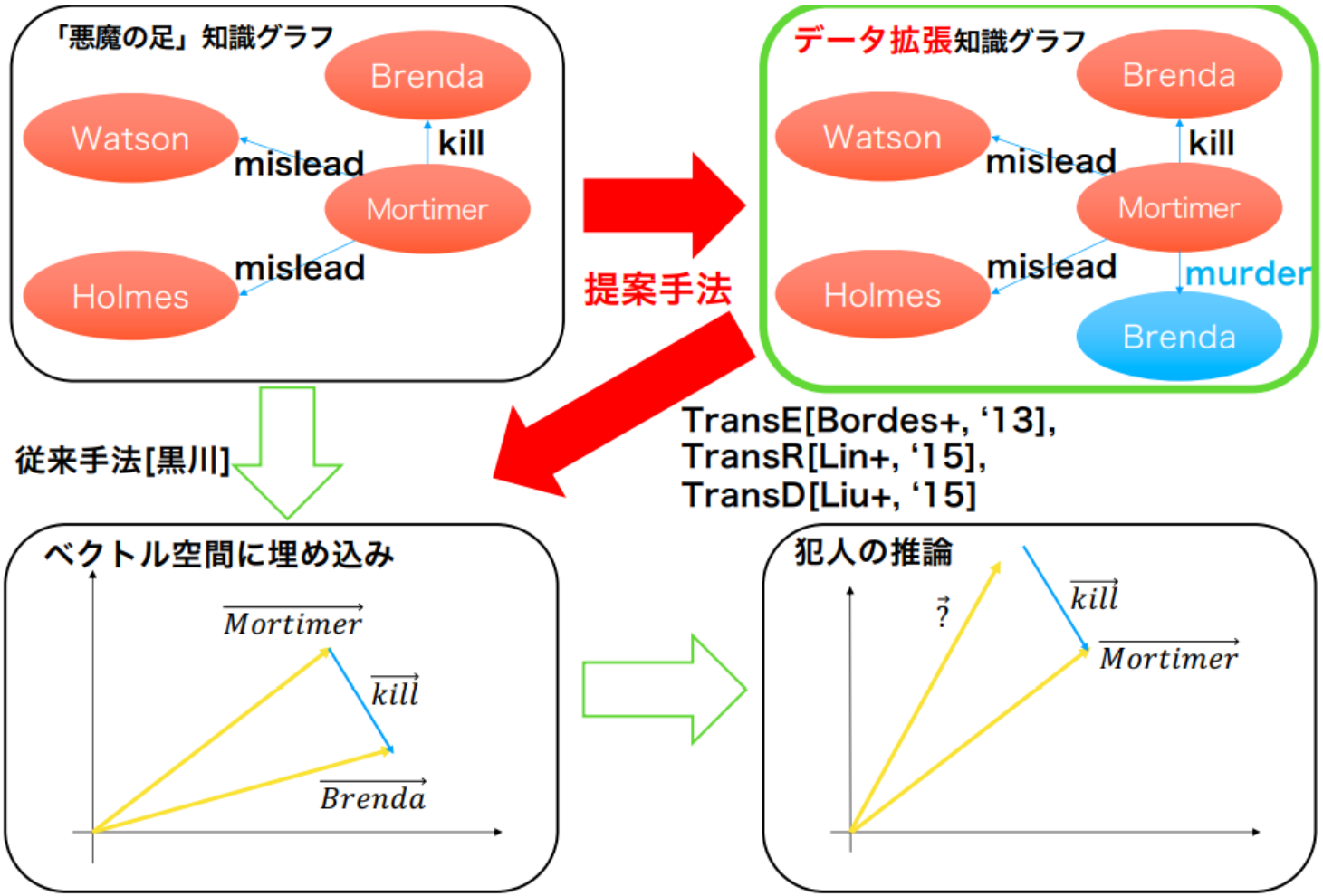
旅行者移動データからの頻出経路発見

- 外国からの旅行者の移動記録データに対して，位置情報を住所表記に変換した上で，頻出記号列発見手法を適用し，旅行者のクラスと経路のクラスの対応関係を閉集合として抽出することに成功

※国立情報学研究所，(株)ノースグリッドとの共同研究
※国際会議 DS 2020@ギリシャで発表

知識グラフ上の推論の精緻化

- 知識グラフ(knowledge graph): 劇的に発展した自然言語処理技術を用いて(大量の)自然言語データから構築された知識ベース



動的計画法のガベージコレクションによる省メモリ化

DNA・RNA配列同士の比較 → 動的計画法(DP)

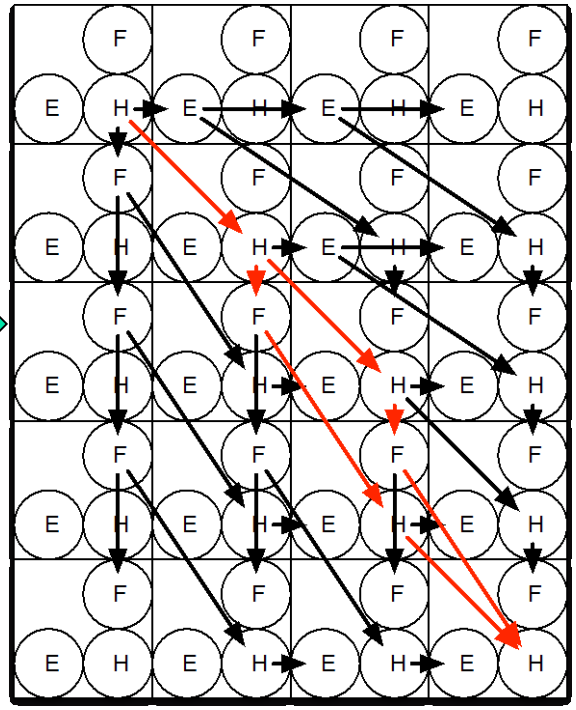
メモリ量は2乗オーダ：新型コロナウィルスゲノム 3万塩基 → 3.6GB

並列化は困難

配列比較のDP過程

		A		T		T	
	-INF		-INF		-INF		-INF
-INF	0	-6	-INF	-7	-INF	-8	-INF
	-6		-INF		-INF		-INF
A -INF	-INF	-INF	6	0	-8	-1	-9
	-7		0		-14		-15
G -INF	-INF	-INF	-8	-14	4	-2	-2
	-8		-1		-2		-8
C -INF	-INF	-INF	-9	-15	-2	-8	2
	-9		-2		-3		-4
T -INF	-INF	-INF	-10	-16	-3	-9	4

解グラフ化



各過程で最適解に寄与しないノードを削除

ガベージコレクション

実行時省メモリ化可能

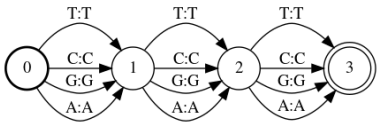
重み付き有限状態トランスデューサによる配列解析

重み付き有限状態トランスデューサ(WFST):
入出力シンボルおよび遷移重みを持つ有限オートマトン

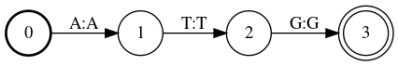
WFST間の様々な演算が存在→複雑なモデルが生成可能

例:coding DNA同士の比較器

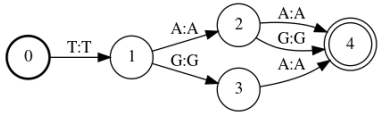
コドン C



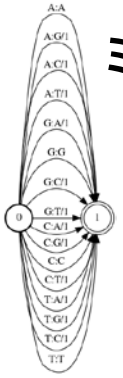
スタートコドン S



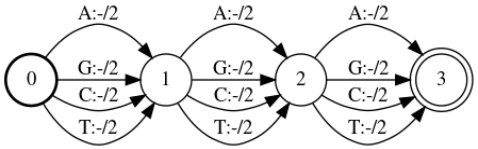
ストップコドン E



ミスマッチ M

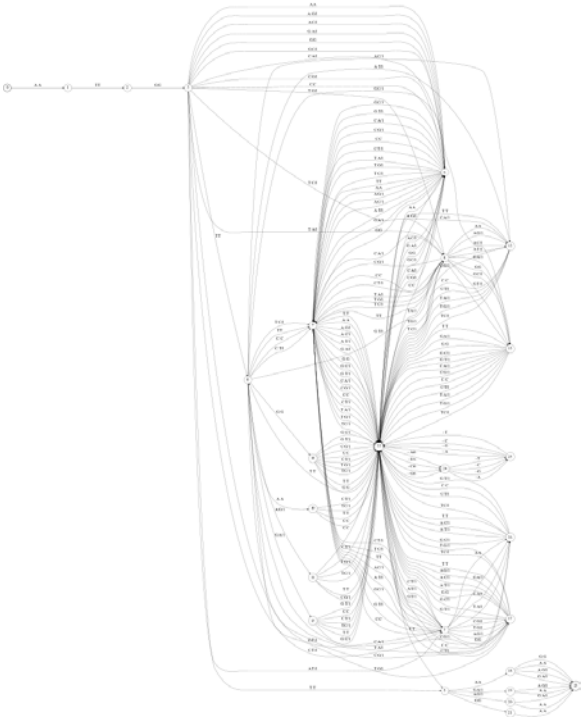


コドン挿入 I



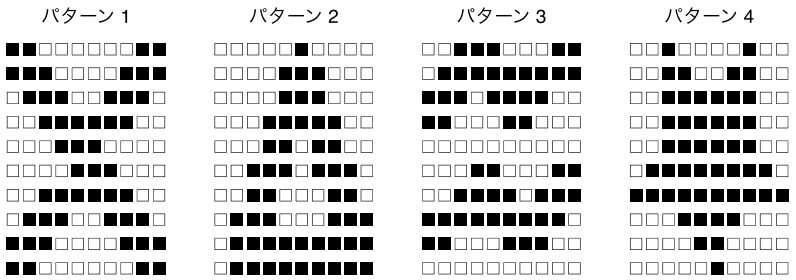
比較器の生成

$$S + \left((C - E) \circ M^* \circ (C - E) + I^* + I^{-1*}\right)^* + E \circ M^* \circ E$$



連想記憶モデルにおける正則化の影響

連想記憶モデル：リカレントNNにパターンを記憶させる手法



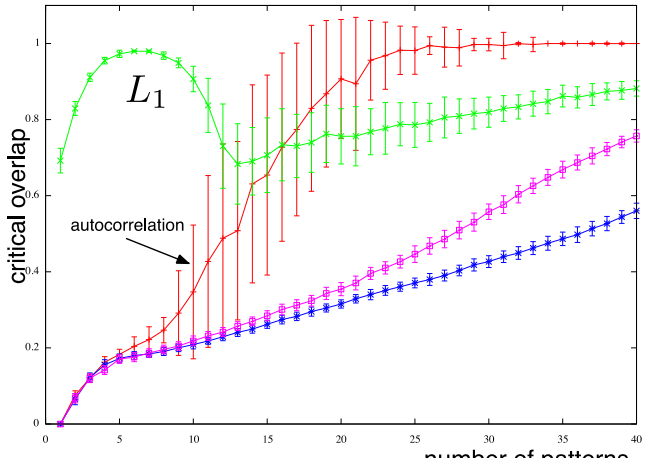
二次計画法、線形計画法で正則化を導入

$$\text{Minimize: } z_i = ||W_i||_p$$

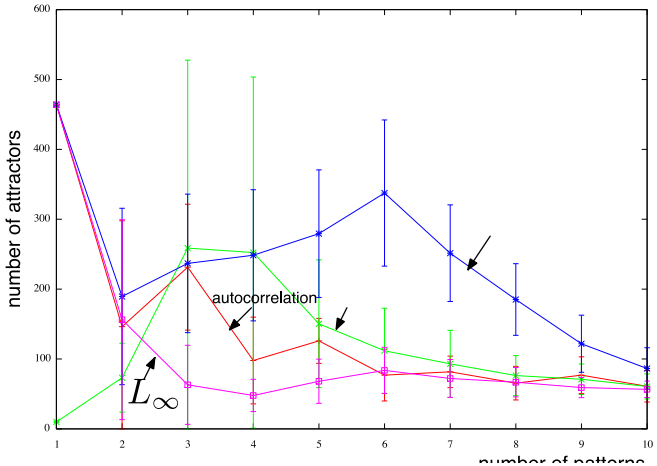
$$\text{Subject to: } \sum_{j=1}^{N} s_i^\mu w_{ij} s_j^\mu \geq 1$$

正則化の相違で引き込み領域の大きさおよび偽記憶の数に影響

引き込み領域：2-norm正規化がベスト



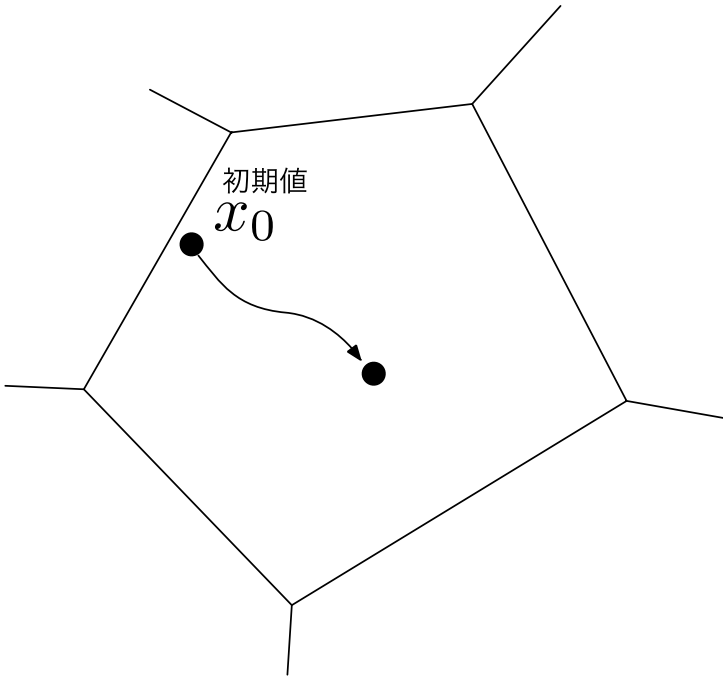
偽記憶：∞-norm正規化がベスト



機械学習による概念形成

NNはアナログ機械であり離散情報である「概念」を扱うことは不得意

中間層にリカレントNNを設け
力学系としての平衡点を学習



各平衡点が概念に相当

形成された概念

平衡点に引き込まれる事例の
平均画像(fashion MNIST使用)

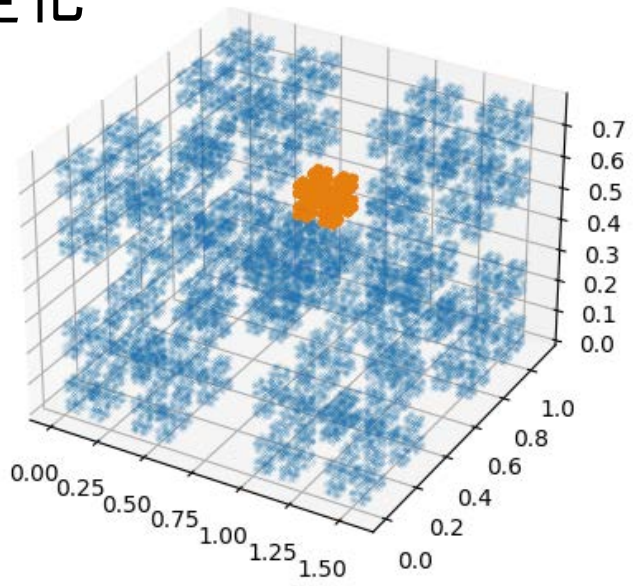
高次元カントールコーディング

カントールコーディング：フラクタル集合上に離散系列情報を埋め込む

脳におけるエピソード記憶モデルとして提案 (Tsuda & Kuroda 2001)

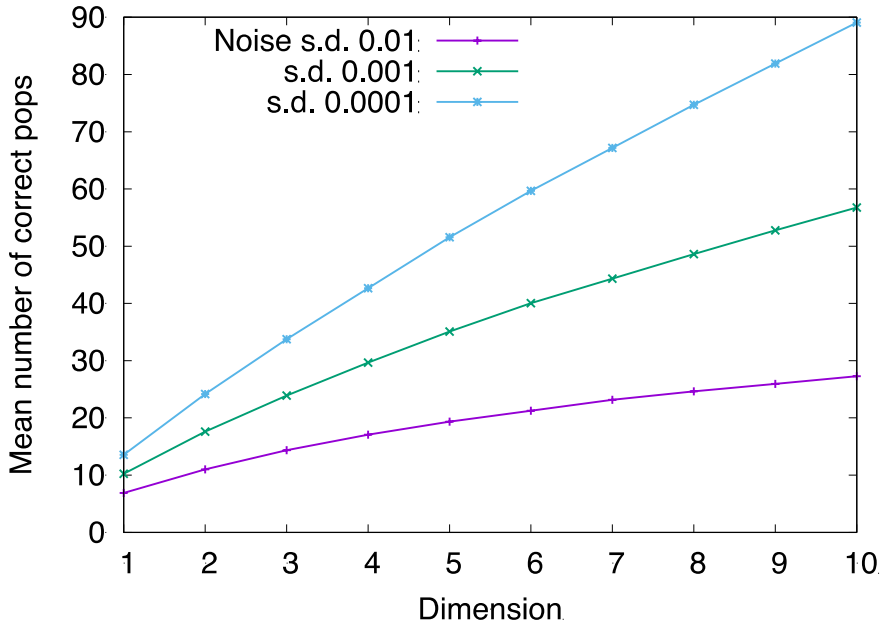
アナログ機械に実装されたスタックデータ構造に相当

高次元化によりスタック性能を安定化



3次元の例 (オレンジ部分は110101)

ノイズ存在下でのスタック性能



歓迎

- システム構築，基礎理論いずれに興味があっても歓迎
  - 議論を通じて自ら追究する
  - 簡単な事柄でも，動的計画法などの数学的手法に抵抗がない
  - 理論だけでなく実装による成果の確認を行える
  - 実装の数学的に意味を自分の言葉で説明する

研究室の環境

- 学生全員がiMac/Mac Pro使用できます！
  - Windows/Linuxも使えます！
  - 4Kディスプレイ
- (他の研究室と比べていませんが)机広いです！
- デロンギのエスプレッソメーカー使えます！
  - 豆は自分で買ってね
  - “A mathematician is a device for turning coffee into theorems.” – A. Rényi

質問や相談があればメールで

教授: 山本章博　akihiro@i.kyoto-u.ac.jp

特定准教授: 市瀬夏洋 ichinose.natsuhiro.7v@kyoto-u.ac.jp

※ メールで質問の際は件名を「大学院志望」とし，氏名を忘れずに！