学習数理情報学研究室の紹介

東京大学　大学院情報理工学系研究科
創造情報学専攻/数理情報学専攻
山西健司
http://www.ibis.t.u-tokyo.ac.jp/
2014/6/7

学習数理情報学研究室

- 機械知能の本質を数理で捉える -

教授　山西健司 (H25より創造学専攻本務)

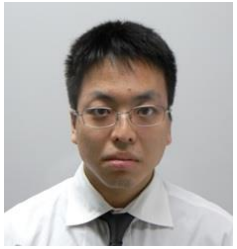
専門:　情報論的学習理論、Latent Dynamics
　　データマイニング応用、異常検知

講義:　確率数理工学(夏:学部)
　　情報論的学習理論(冬:大学院)
　　データマイニングによる異常検知(冬:大学院)

助教:松島慎

専門:　機械学習(大規模分類問題)

講義:　プログラミング演習



特任助教:　森野佳生

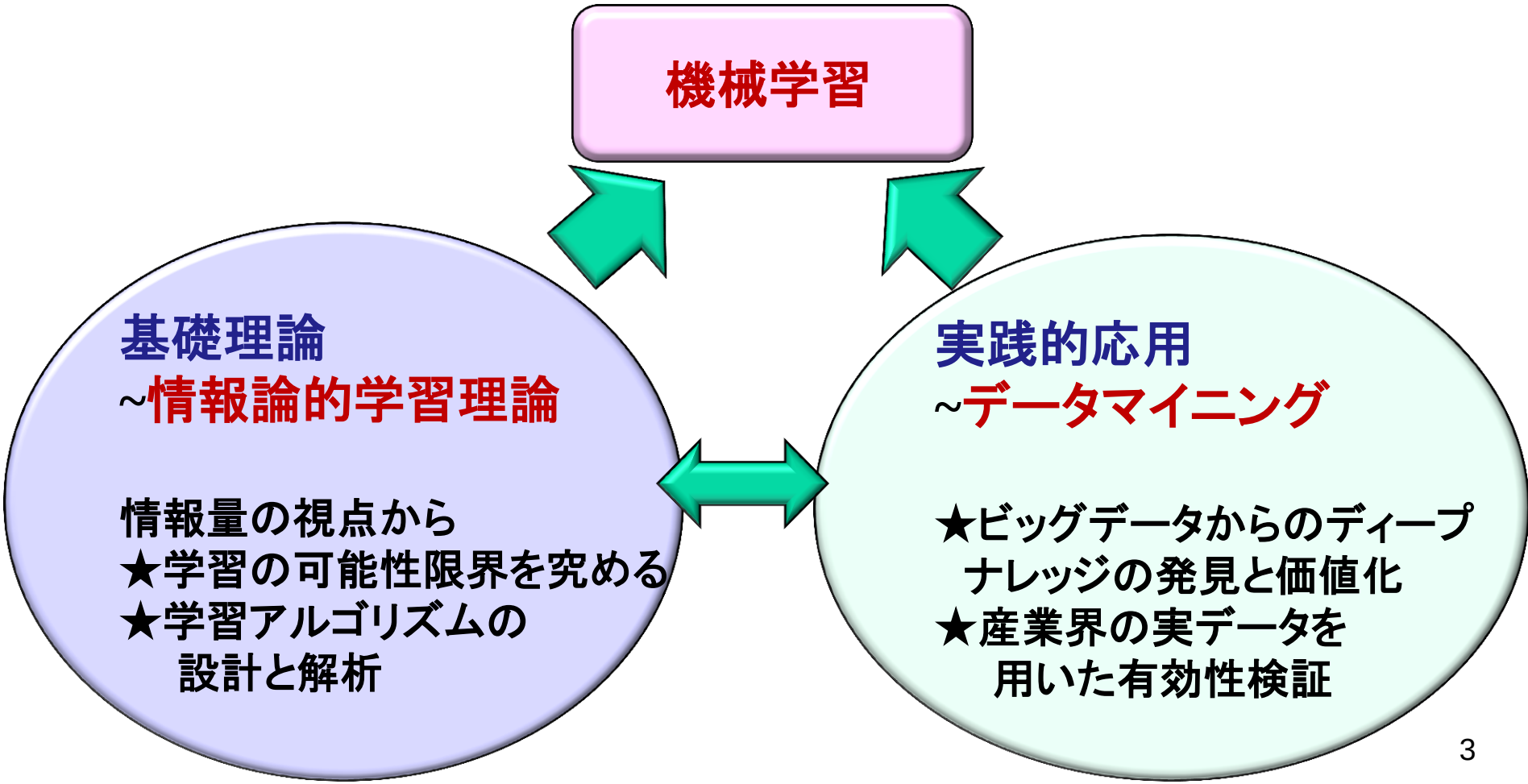
専門:　機械学習，非線型振動子

講義:　数理情報工学実験第二

大学院生:PD:　寺園，　田村
木村(D3)，　梶野(D2),
家入(M2)，伊藤(M2)，貝ヶ石(M2)，　真矢(M2)，渡邊(M2),
石井(M1), 江田(M1),尾亦(M1),友田(M1),中村(M1),
宮口(M1), 要名本(M1)

学習数理情報学における両輪

情報論的学習理論とデータマイニング実践を両輪とする



研究室の特徴

■しっかりとした基礎理論に立脚しつつ、 現場に生きる
機械学習・データマイニング研究をめざします

■ JST CREST ビッグデータ基盤「複雑データからのディープナレッジの発見と価値化」の研究拠点です

■幅広い産業界や公的研究機関との連携のもとで、研究を進めます

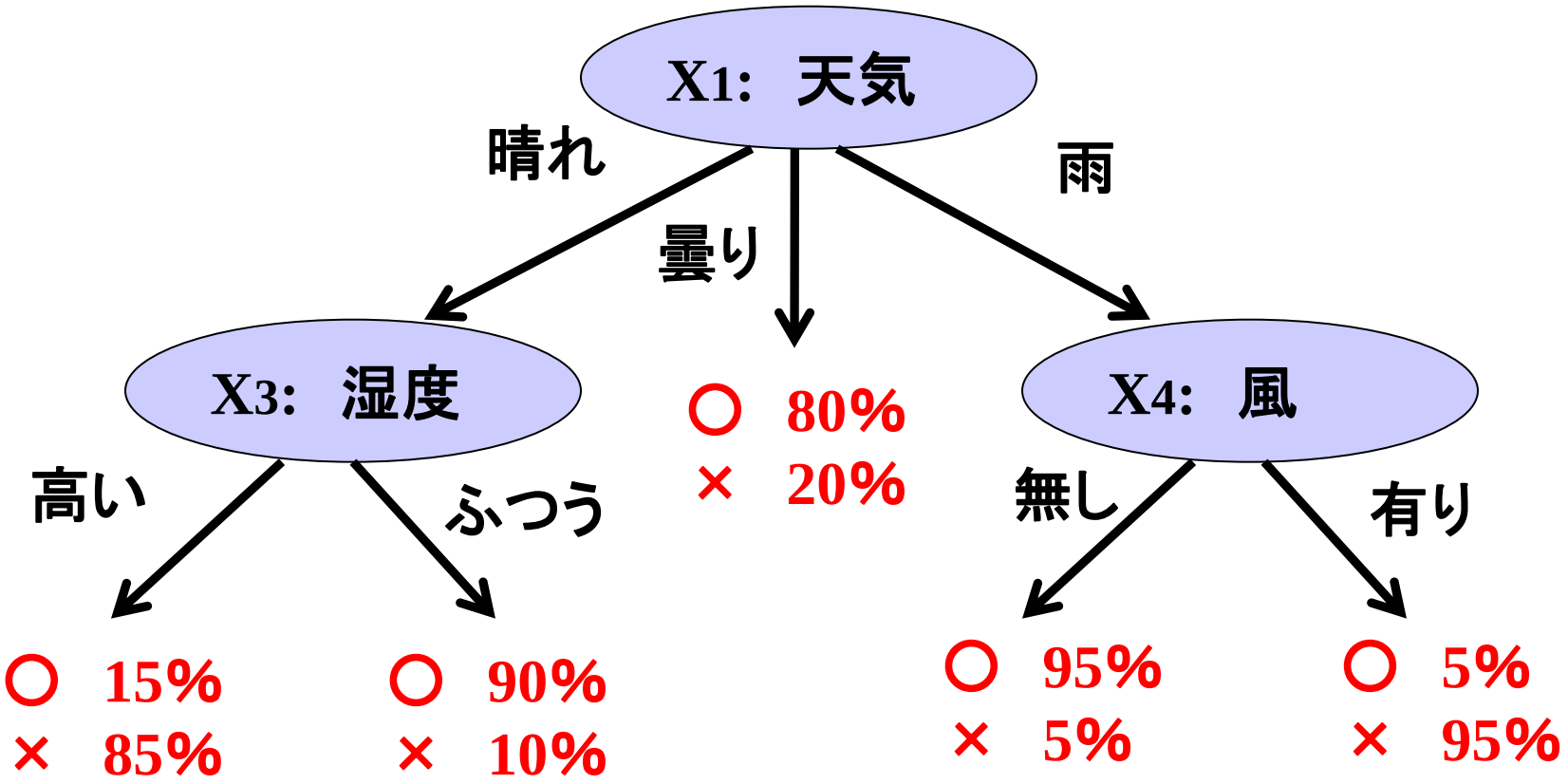
■顧客視点に立って独善的ではないデータの利活用を心掛けています。

情報論的学習とは

X1　天気	X2　温度	X3　湿度	X4　風	Y　プレイ
晴れ	暑い	高い	無し	×
晴れ	暑い	高い	有り	×
曇り	暑い	高い	無し	○
雨	暖かい	高い	無し	○
雨	涼しい	ふつう	無し	○
雨	涼しい	ふつう	有り	×
曇り	涼しい	ふつう	有り	○
晴れ	暖かい	高い	無し	×
晴れ	涼しい	ふつう	無し	○
雨	暖かい	ふつう	無し	○
晴れ	暖かい	ふつう	有り	○
曇り	暖かい	高い	有り	○
曇り	暑い	ふつう	無し	○
雨	暖かい	高い	有り	×
……	……..	…….	…….	

知識発見

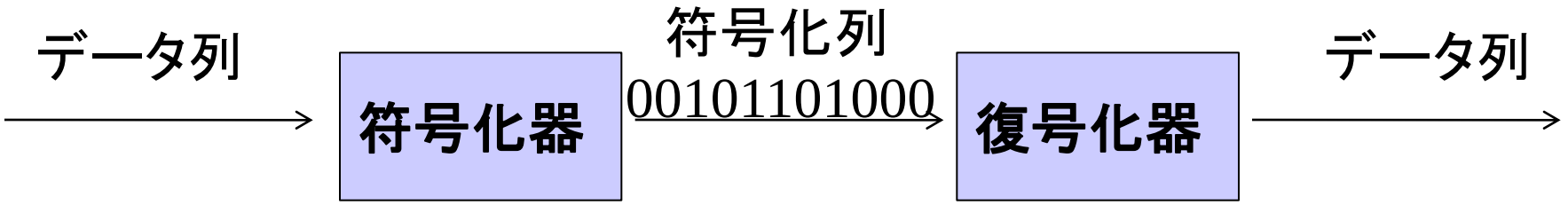
X=(X1, X2, X3, X4)　Y=○ or ×



P(Y|X)　……　知識

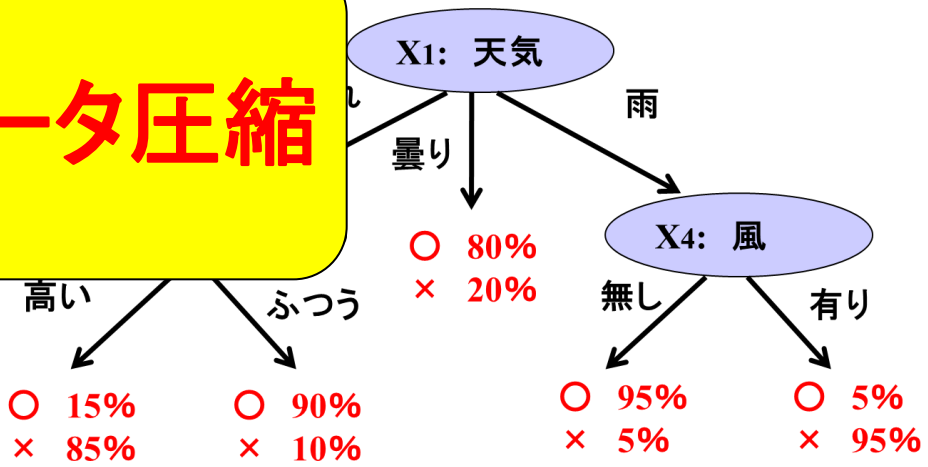
機械学習とは　データからP(Y|X)を導き出すこと

MDL原理と学習



X1 天気	X2 温度	X3 湿度	X4 風	Y プレイ
晴れ	暑い	高い	無し	×
晴れ	暑い	高い	有り	×
曇り	暑い	高い	無し	○
雨	暖かい	高い	無し	○
雨	涼しい	ふつう	無し	○
雨	涼しい	ふつう	有り	×
曇り	涼しい	ふつう	有り	○
晴れ	暖かい	高い	無し	×
晴れ	涼しい	ふつう	無し	○
雨	暖かい	ふつう	無し	○
晴れ	暖かい	ふつう	有り	○
曇り	暖かい	高い	有り	○
曇り	暑い	ふつう	無し	○
雨	暖かい	高い	有り	×
……	……	……	……	

データ圧縮



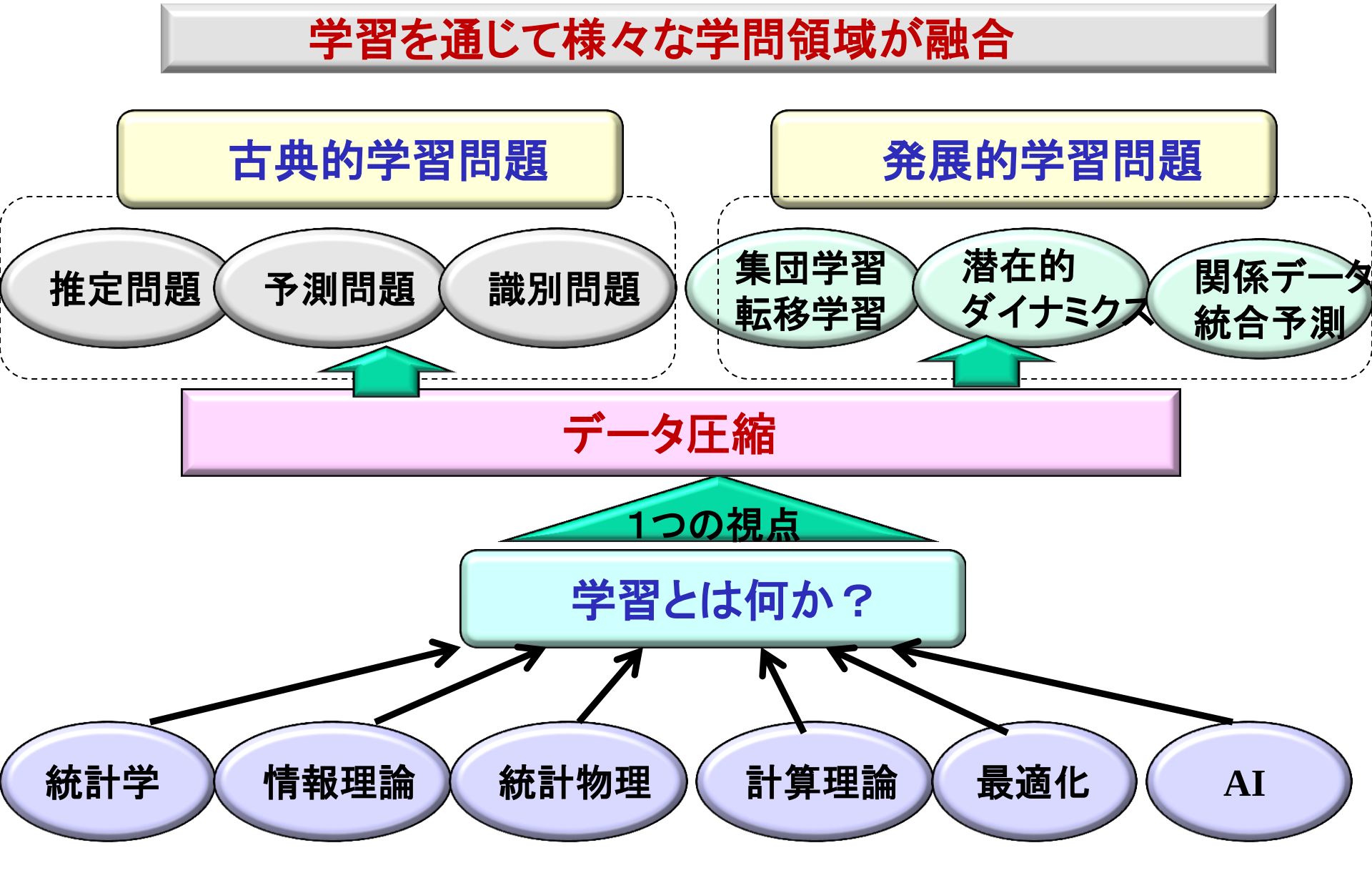
$$\mathcal{L}(D^n|M) + \mathcal{L}(M) \quad \Rightarrow \text{最小化}$$

[データ符号長] [モデル符号長] w.r.t. M

確率的コンプレキシティ

MDL(Minimum Description Length)原理

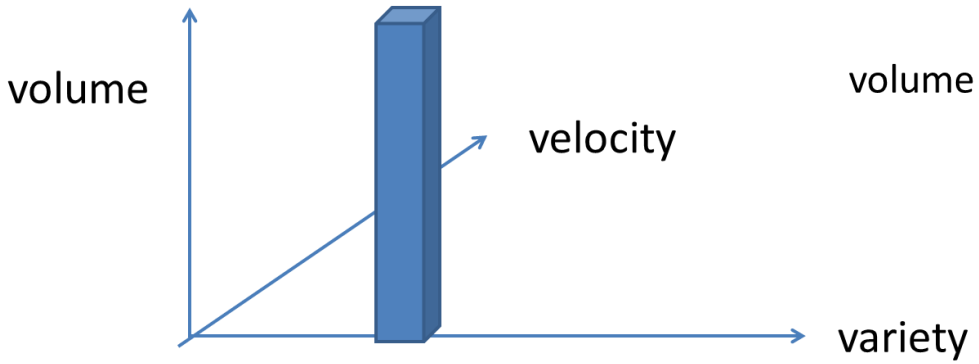
情報論的学習理論のシナリオ



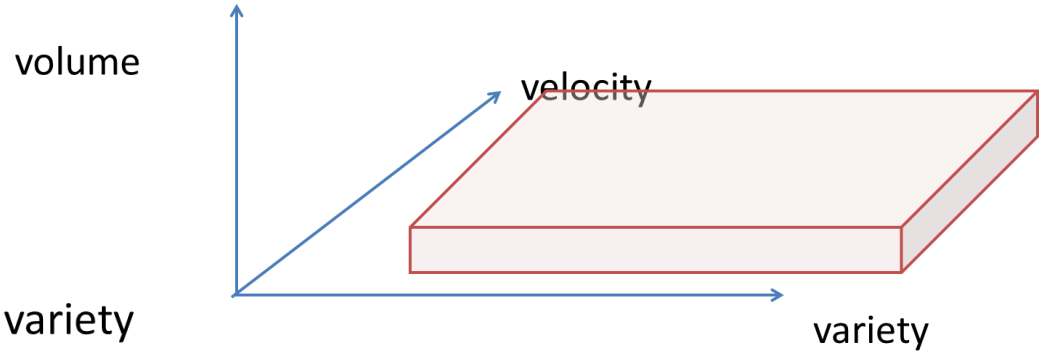
ビッグデータへの取り組み

5V

- Volume （大量性、大次元）
- Variety （多様性、複雑性（潜在性））
- Velocity （動的、非定常、変化）
- Value （付加価値）
- Veracity （真実性）



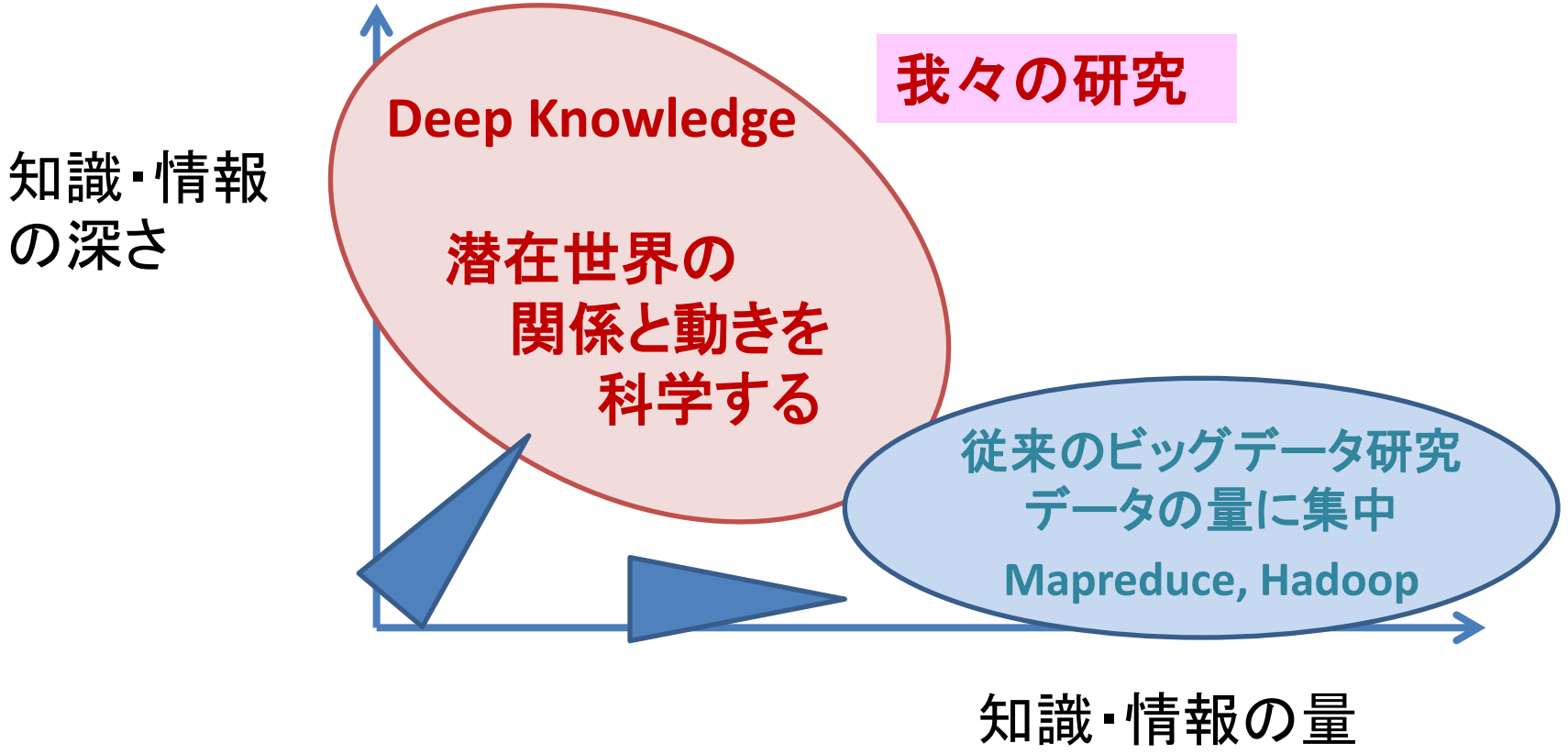
世間的イメージ



実際のビッグデータ

日本学術振興機構　CREST
「複雑データからのディープナレッジの発見」の研究拠点

知識の深さに重点を置き、データに潜在するDeep Knowledge を発見し、value を引き出す新しい方法論を開発し、体系化する



学習数理情報学研究室のキーワード

情報論的学習理論
IBIS(Information-Based Induction Sciences)

IBISML

ディープ
ナレッジ

潜在的
ダイナミクス
Latent
Dynamics

関係データ
統合予測

学習数理情報学研究室のテーマ

■ 情報論的学習理論

- ●潜在的ダイナミクス
- ●モデル選択、 MDL原理
- ●関係データ統合予測、潜在変数モデル
- ●正則化理論、テンソル型学習
- ●マルチタスク学習、転移学習

■ データマイニング （ディープナレッジの発見）

- ●異常検知・変化検知
- ●緑内障進行予測
- ●Social Networkマイニング
- ●教育データマイニング
- ●交通・マーケティング向けデータマイニング

潜在的ダイナミックス(1/4)

クラスタリング構造変化検知

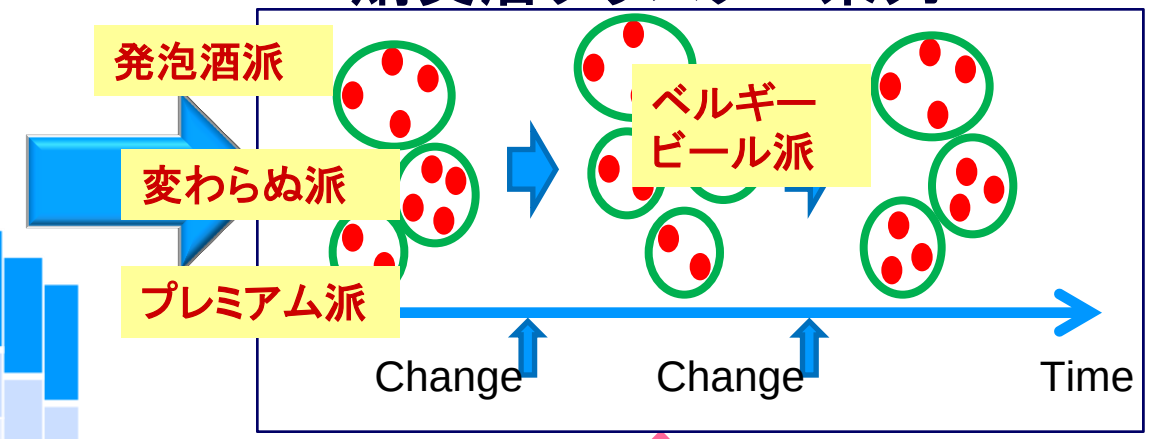
購買データからのユーザ層変化検知への応用

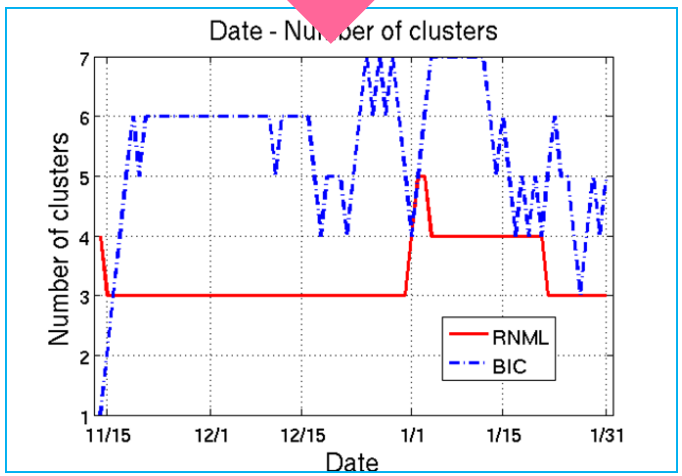
[Hirai Yamanishi KDD2012, ISIT2011, IEEE IT2013]

多変数時系列

	Beer 1	Beer 2	. . .
User 1	350	700	. . .
User 2	1050	350	. . .
. . .	. . .	. . .	. . .

購買層クラスター系列





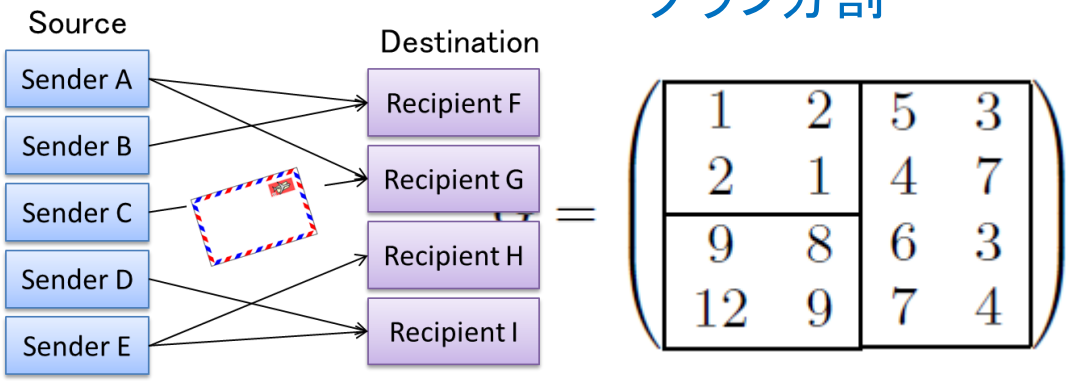
データ圧縮に基づく変化点解析

Inc.

潜在的ダイナミックス(2/4)

木構造に基づくグラフ分割構造変化検知

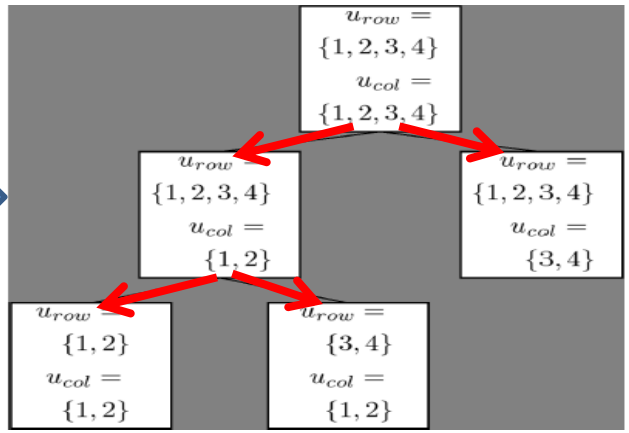
[Sato Yamanishi ICDM2013]



グラフ分割

$$= \left(\begin{array}{cc|cc} 1 & 2 & 5 & 3 \\ 2 & 1 & 4 & 7 \\ \hline 9 & 8 & 6 & 3 \\ 12 & 9 & 7 & 4 \end{array}\right)$$

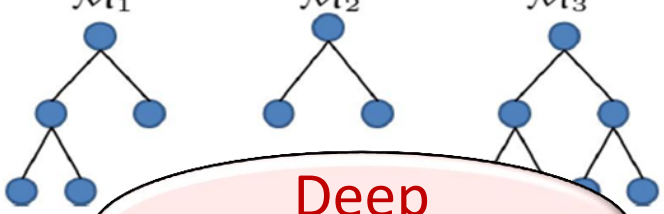
木構造



グラフ分割時系列

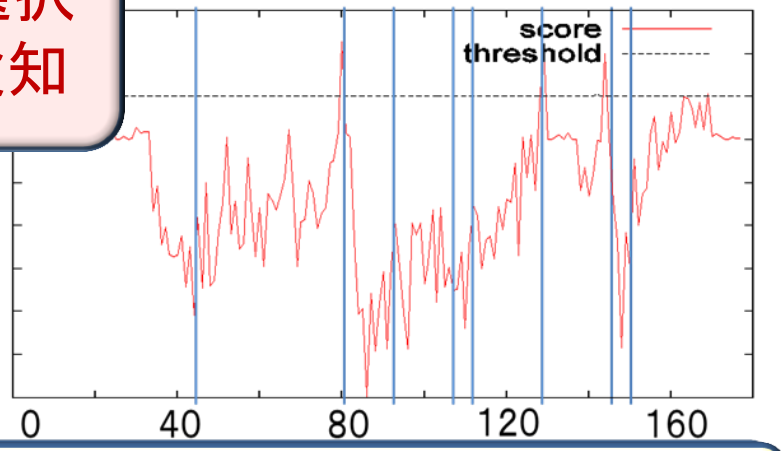
$$\overbrace{G_1, G_2, G_3}, \overbrace{G_4, G_5}, \overbrace{G_6, G_7,}$$

$$\underbrace{M_1, M_2, M_3}_{\mathcal{M}_1}, \underbrace{M_4, M_5}_{\mathcal{M}_2}, \underbrace{M_6, M_7, M_8}_{\mathcal{M}_3}, \ldots$$



Deep Knowledge

動的モデル選択による変化検知



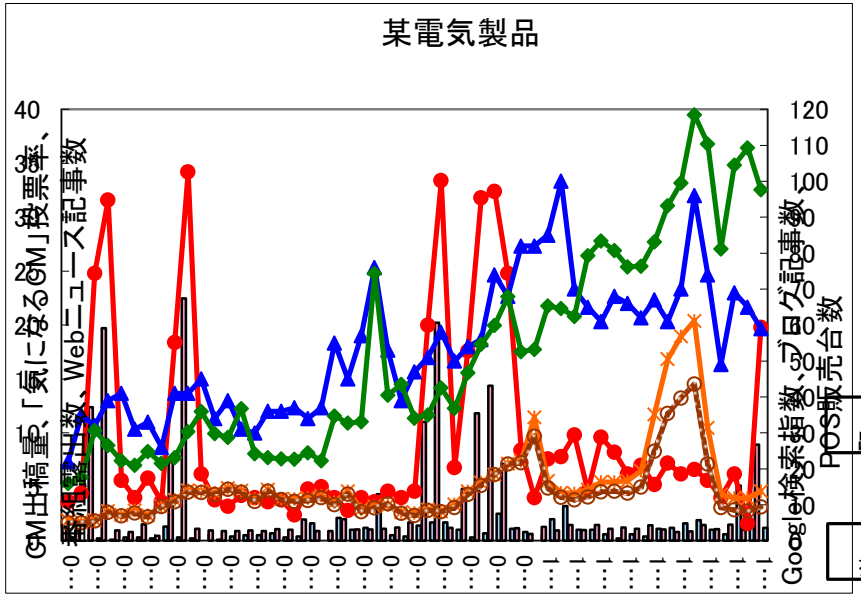
イベント予兆検知

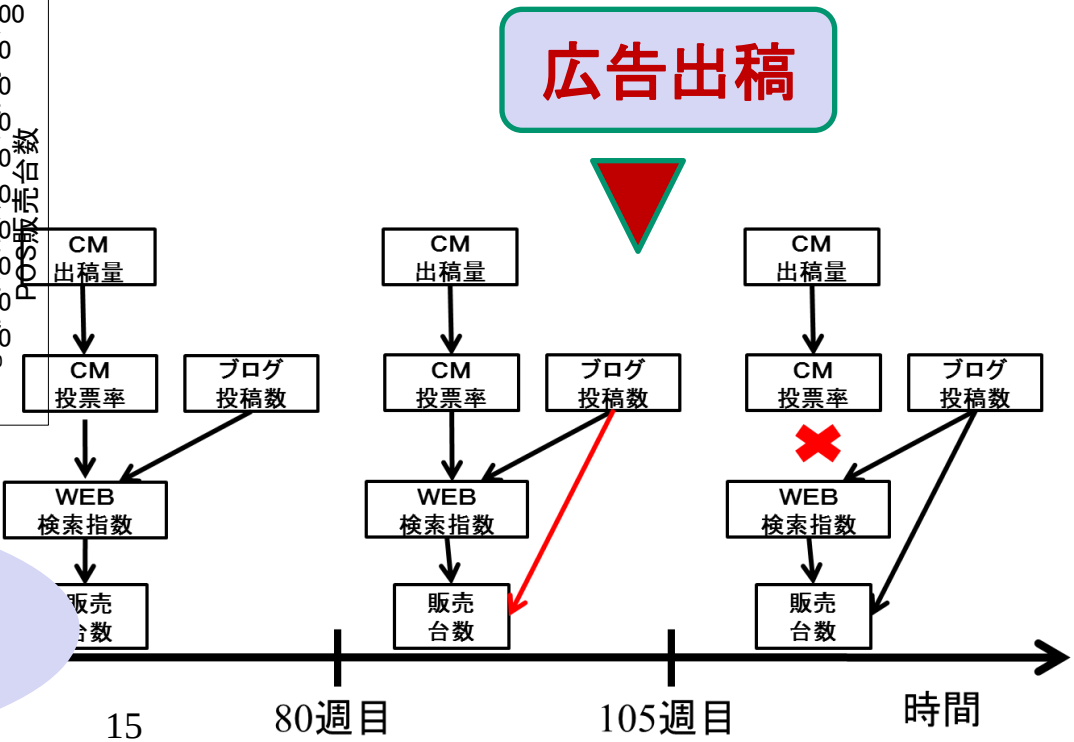
潜在的ダイナミックス(3/4)

ベイジアンネット構造変化検知に基づく広告効果測定

[Hayashi Yamanishi ICDM2013, DAMI2013]

多次元時系列データからベイジアンネットワークの構造変化検知により広告効果を測定





広告出稿により
ブログでの評判が高まり
売り上げに結び付いた

潜在的ダイナミクス(4/4)

教育データマイニング

[Oeda Yamanishi EDM2013]

試験時系列データからのスキル構造変化の検知

Many examination results

		Skills		
		s_1	s_2	s_3
Items	i_1	0	1	0
	i_2	0	0	1
	i_3	1	0	0
	i_4	0	0	1

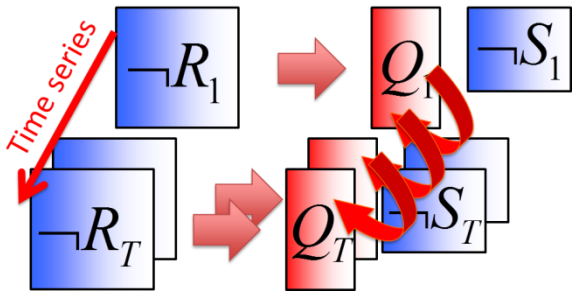
A made Q-matrix automatically

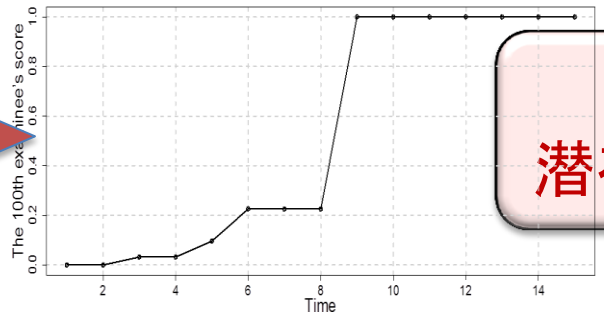
Analyzing new examination results by the Q-matrix

Non-negative matrix factorization(NMF)

$$\underset{\text{Items}}{\overset{\text{Examinees}}{\begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}}} = \underset{\text{Items}}{\overset{\text{Skills}}{\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}}} \underset{\text{Skills}}{\overset{\text{Examinees}}{\begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}}}$$

- **Online NMF:**





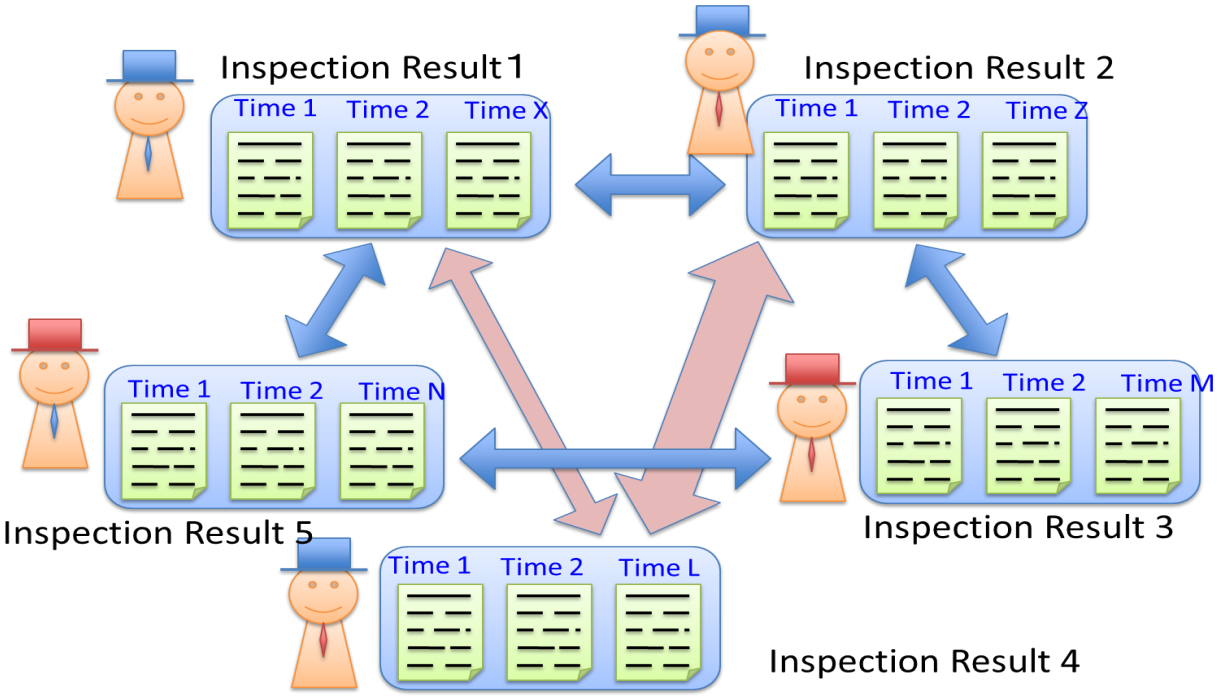
Deep Knowledge
潜在スキル構造の変化

関係データ統合予測(1/2)

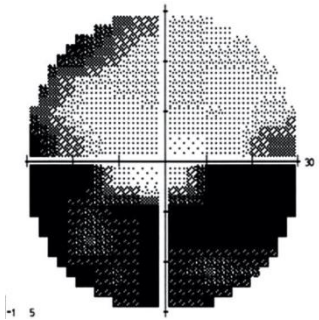
緑内障進行予測

[Liang, Tomioka, Murata, Asaoka, Yamanishi ICDM013]

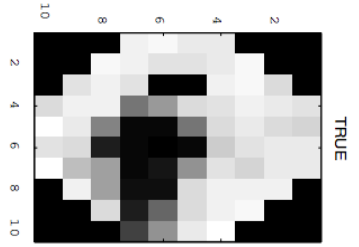
病状の時空間パタンの似ている患者の情報を利用して進行予測精度を増強



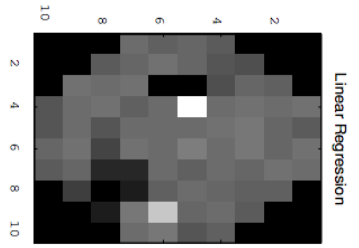
Deep Knowledge
＝患者ネットワーク
クラスタリング



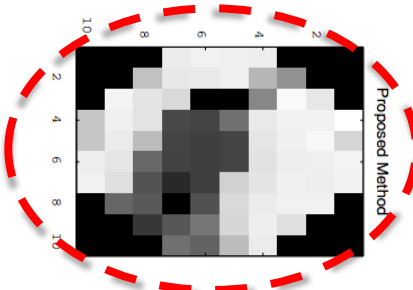
Correct



Prediction without using relational data



Prediction aggregating relational data

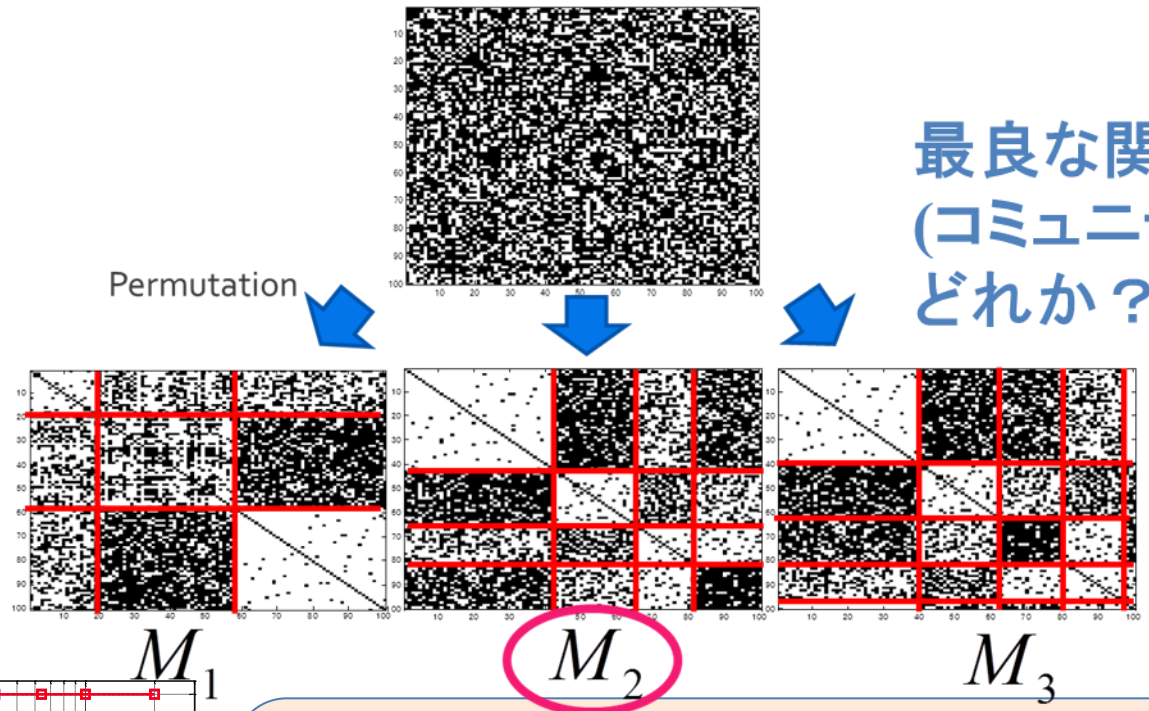


関係データ統合予測(2/2)

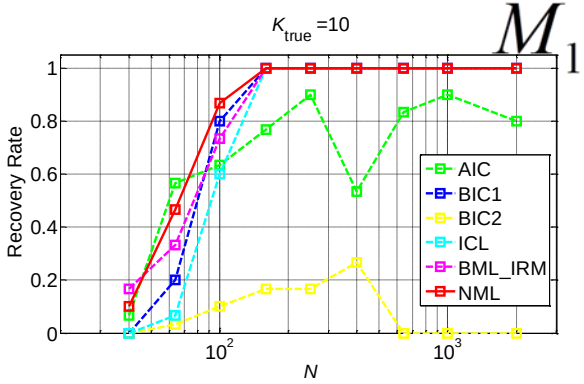
関係データモデル選択

[Sakai, Yamanishi BigData2013]

NML符号長によりヘテロな関係データの最適分割構造を推定



最良な関係性の分割(コミュニティ)構造はどれか？



一般関係モデルのモデル選択規準を開発

$$\ln C_{\mathrm{SBM}}(N,K) = \frac{K-1+K(K+1)}{2}\ln N - \frac{K-1+K(K+1)/2}{2}\ln 2\pi - \frac{K}{2}\ln 2$$

$$+K\ln\Gamma\big((K+2)/2\big) - \ln\Gamma\big(K(K+2)/2\big) + \frac{K(K+1)}{2}\ln\pi$$

SNSマイニング

リンク情報に基づくTwitterからの話題出現検知

[Takahashi, Tomioka, Yamanishi TKDE2014]

従来の話題検出法

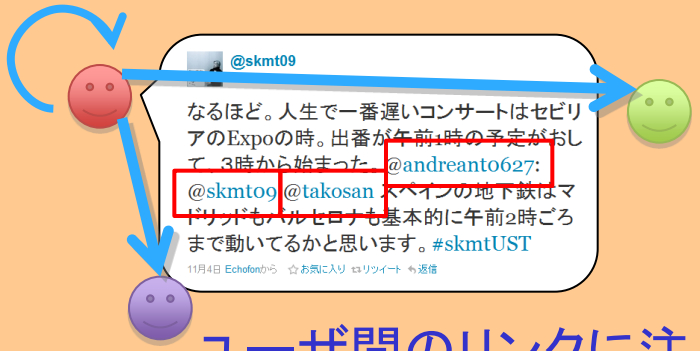
GW　原発

母の日

単語に注目

問題点：言い換えに弱い

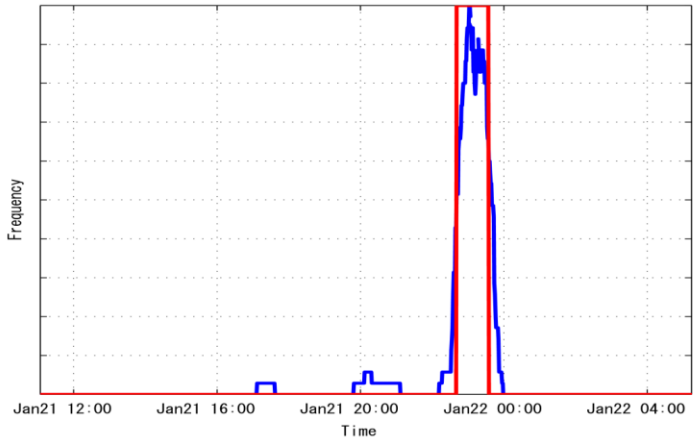
提案する話題検出法



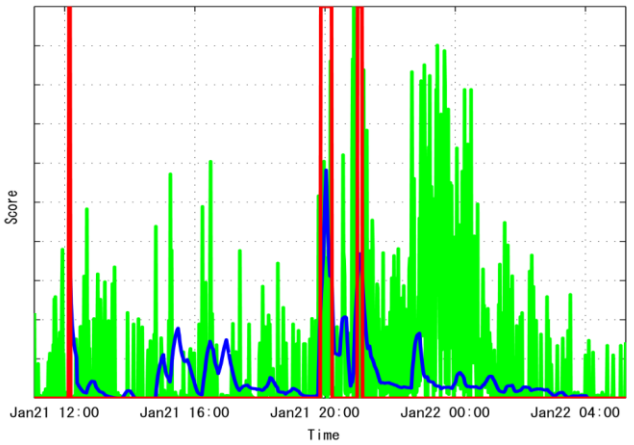
ユーザ間のリンクに注目

従来手法（キーワード頻度）より速い話題検出が可能

比較手法(キーワード頻度)



提案手法(変化点検出)



木構造データの高速カーネル関数の設計

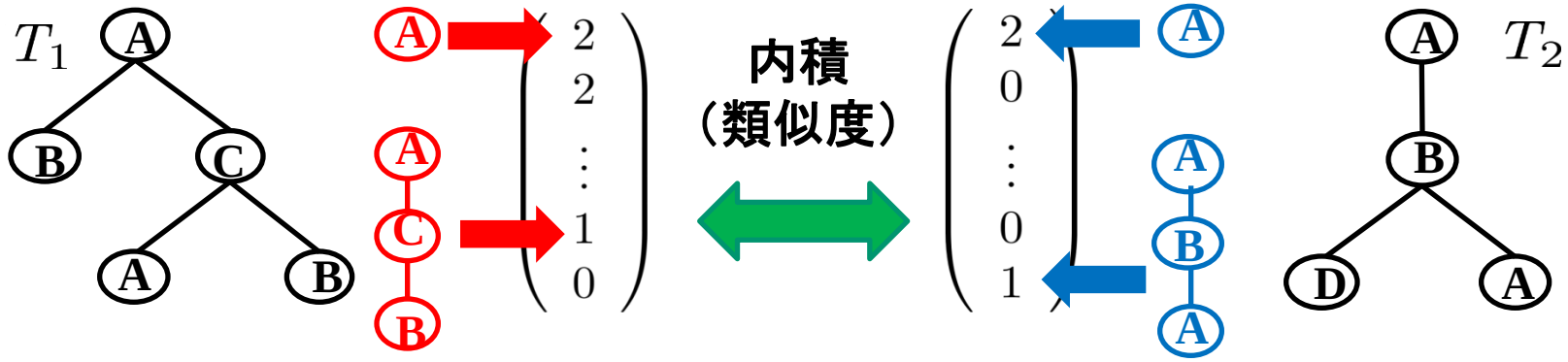
木構造をもつカーネルの高速計算で複雑な対象を分類

[Kimura Kashima ICML 2012]

$$k(T_1, T_2) = \sum_{p \in P} num(T_{1p}) num(T_{2p})$$

パスpの出現回数

パスに基づく木構造間の類似度を定義



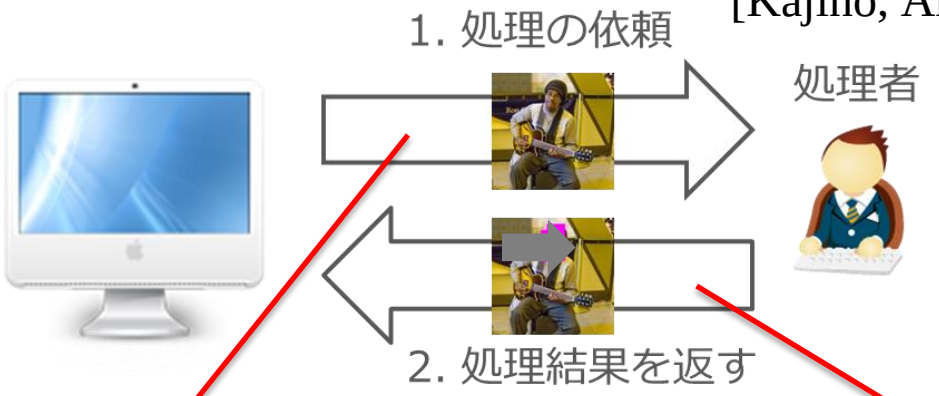
i	SA	LCP	*Si*
1	1	0	A
2	3	0	BA
3	2	1	CA
4	4	-1	CBA

拡張接尾辞配列を用いることで2つの木に共通するパス数を高速に計算可能

人-機械・協調計算のプライバシー保護

情報理論・計算理論に基づき情報漏洩、品質保証精度を定量化

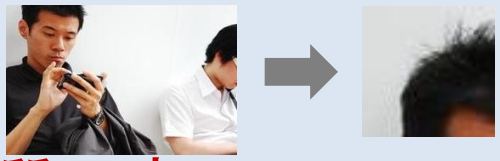
[Kajino, Arai, Kashima DAMI2014]



人と機械の協調計算

問題：処理対象の情報漏洩

解：必要以外の情報を排除



顔は塗りつぶせるが、誰かわからない

品質尺度：$\mathrm{E}_I[\mathrm{KL}(\underline{p(R \mid I)} \,\|\, \underline{p'(R \mid I)})]$

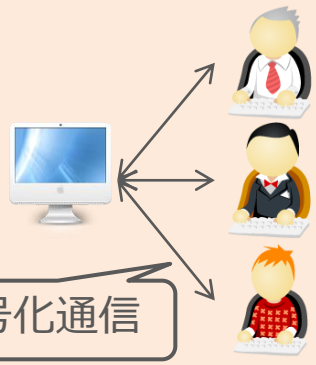
（R: 成果物, I: 対象物）　保護前　保護後

漏洩尺度：$\mathrm{E}_I[\mathrm{KL}(q(S \mid I) \,\|\, q(S))]$

（S: 秘匿情報, I: 対象物）　保護前　保護後

問題：処理結果の情報漏洩

解：統合化した統計量を返す
統合化には暗号化



単一マシンを用いた大規模最適化

学習最適化問題をDual Cashed Loopsを用いて高速処理

[Matsushima, Vishwanathan, Smola, KDD2012]

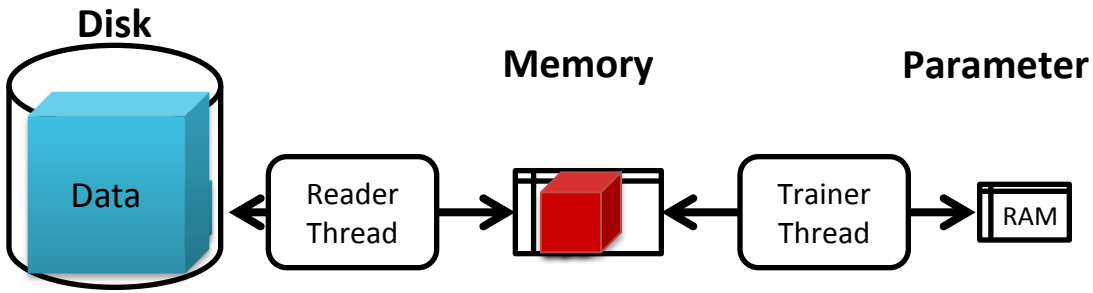
$$\min_{w \in \mathbf{R}^d} \underbrace{\sum_{j=1}^{d} \phi(w_j)}_{\text{regularizer}} + \underbrace{\sum_{i=1}^{n} \ell(w \cdot y_i x_i)}_{\text{loss}}$$

$$\Rightarrow \min_{\alpha \in \mathbf{R}^n} \sum_{j=1}^{d} \phi^*\left(\sum_{i=1}^{n} \alpha_i y_i x_{ij}\right) + \sum_{i=1}^{n} \ell^*(-\alpha_i)$$

Dual Problem:

Application: **線形回帰、ロジスティック回帰、サポートベクターマシン（SVM）など**

Solution: **Dual Cached Loops**



- 読込/訓練スレッドを同時に非同期的に動作
- 訓練スレッドは高速にアクセス可能なメモリにのみアクセス
- “uninformative”なデータを優先的にメモリから削除する

近年の修論テーマ

- 非定常情報源に対するパラメータ推定とモデル選択の研究
- 正規化最尤符号に基づくクラスタリングの研究
- ネットワーク構造変化検知の研究
- 木構造データに対する高速カーネル関数の設計
- Convex Formulation for Learning from Crowds
- 木構造に基づくグラフ構造変化検知の研究
- 区間的定常情報源の学習に関する研究
- 独立成分分析に基づく変化検知の研究
- 行列補完問題に対する転移学習アプローチ
- 非一様データからの統合的学習
- グループ最適化に基づくネットワーククラスタリングの研究
- Dueling Bandit問題に対する漸近最適アルゴリズム