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
数理情報学専攻
山西研究室

http://www.ibis.t.u-tokyo.ac.jp/

2023/5/20

大学院入試説明会

参加者はチャットにお名前と所属をお書きください

Please write your name and affiliation to chat.

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

専門:　情報論的学習理論、データマイニング、異常検知

講義:　機械学習の数理（夏：学部）

数理情報学特別講義I　データマイニングによる異常検知(冬：大学院）

情報論的学習理論（冬：大学院）

准教授　鈴木大慈

専門:　機械学習，数理統計学，確率的最適化

講義:　確率数理工学（夏：学部）

確率数理要論（冬：大学院），

講師　久野遼平　(MIセンター本務、数理兼務）

専門:　経済学、ビッグデータ解析

講義:　データマイニング実践演習

特任助教　近藤亮磨　（MIセンター本務、数理兼務）

専門:　ウェブデータベースシステムとデータサイエンス応用

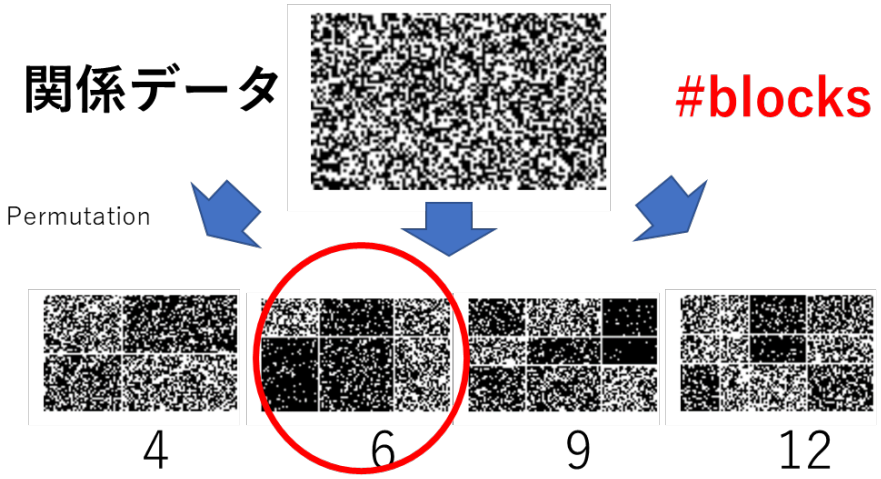
講義:　データマニング実践演習

山西研究室の概要

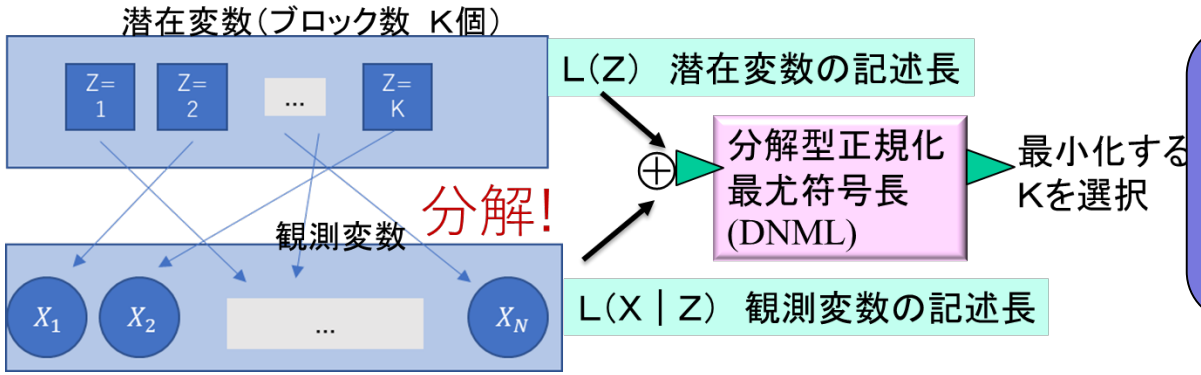

コア技術：情報論的学習理論
目標： ディープナレッジの発見
基礎理論
潜在的ダイナミクス
潜在構造最適化
予兆情報学
Sign Informatics
主要応用
医学応用
経済応用
マーケット解析
交通リスク解析

潜在構造最適化 1
DNMLによる潜在変数モデル選択

分解型正規化最尤符号長規準(DNML)で潜在変数の数を最適化



[Yamanishi,Wu,Sugawara Okada DAMI 2019]
[Wu,Sugawara,Yamanishi KDD2017]
[Okada, Yamanishi, Masuda RSOS2019]



1)効率計算可能
2)広い適用可能性
3)高精度かつロバスト
なモデル選択

$$P(X;K) = \sum_{Z} P(X,Z;K)$$

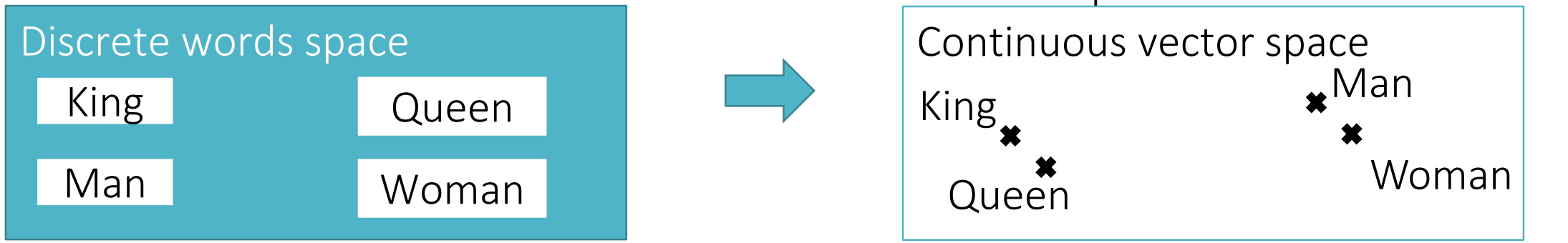
顕在変数　潜在変数

$$\min_{K} \{ L_{\mathrm{NML}}(X|Z;K) + L_{\mathrm{NML}}(Z;K) \}$$

潜在構造最適化 2
Word2vec 単語埋め込みの次元推定

逐次的NML符号規準により埋め込み次元推定の決定を世界で初めて実現

[Hung, Yamanishi *Entropy* 2021]



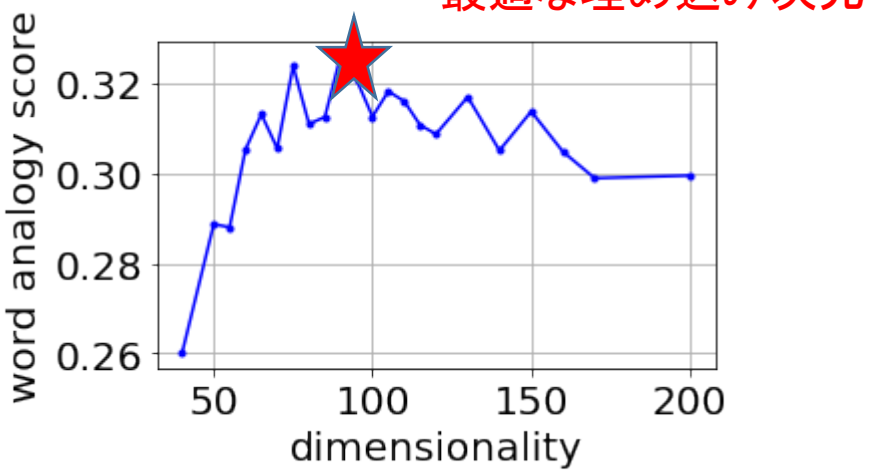
$$\overrightarrow{King} - \overrightarrow{Queen} = \overrightarrow{Man} - \overrightarrow{Woman}$$

SNML規準

$$\mathcal{L}_{SNML}(\mathbf{c}|\mathbf{w};\mathcal{M}_d) = \sum_{i=1}^{n} \mathcal{L}_{SNML}(c_i|w^i, c^{i-1};\mathcal{M}_d) \Rightarrow \min \text{ w.r.t. d}$$

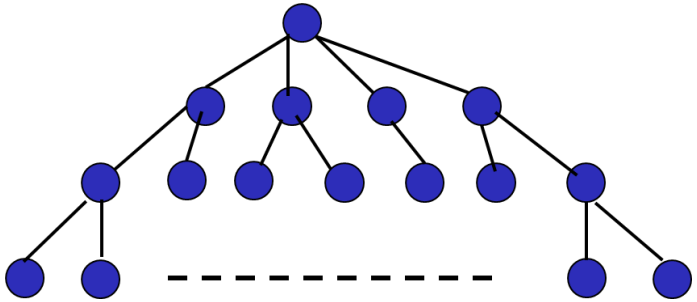
$$\mathcal{L}_{SNML}(c_i|w^i, c^{i-1};\mathcal{M}_d) = -\log P(c_i|w^i, c^{i-1};\hat{\theta}(w^i, c^i)) + \log \sum_{c\in V_c} P(c|w^i, c^{i-1};\hat{\theta}(w^i, c^{i-1}, c))$$

where $c^i = c_1 c_2 \ldots c_i$ and $w^i = w_1 w_2 \ldots w_i$

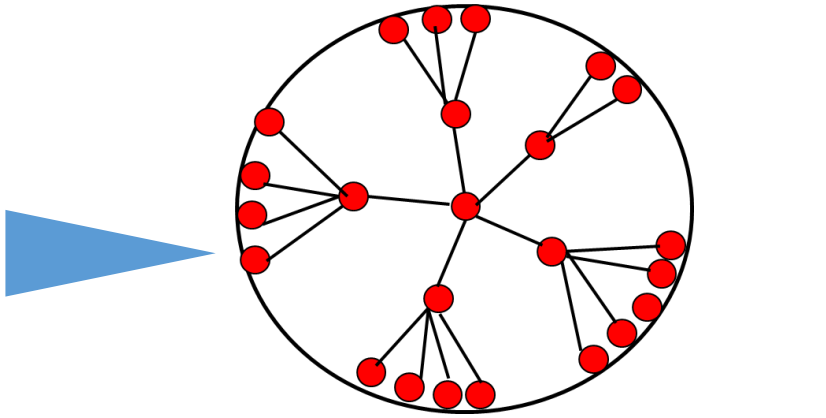


潜在構造最適化 3
グラフの双曲空間埋め込み次元推定と性能解析

階層グラフを双曲空間埋め込みする際の次元をDNMLで最適推定



階層的グラフ



双曲空間埋め込み

[Suzuki, Nitanda, Wang, Xu, Yamanishi Cvazza, Neurlips 2022]

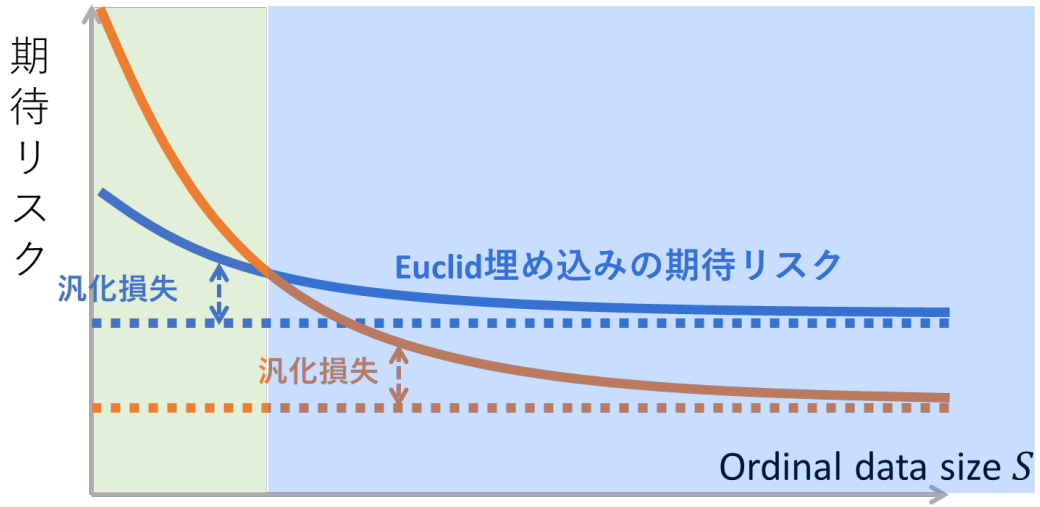
[Suzuki, Nitanda, Suzuki, Wang, Tian, Yamanishi ICML2023]

次元選択

DNML 符号長規準 [Yuki, Ike, Yamanishi ICDM2022]

$$\min_d \mathcal{L}_{\mathrm{DNML}}(y,z) = \min_d \{\mathcal{L}_{\mathrm{NML}}(y|z;d) + \mathcal{L}_{\mathrm{NML}}(z;d)\}$$

y: グラフ, z:双曲埋め込み, d: 次元

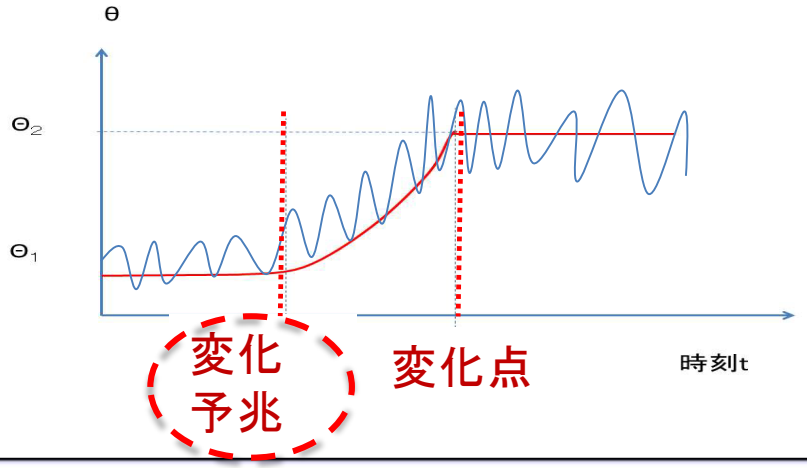


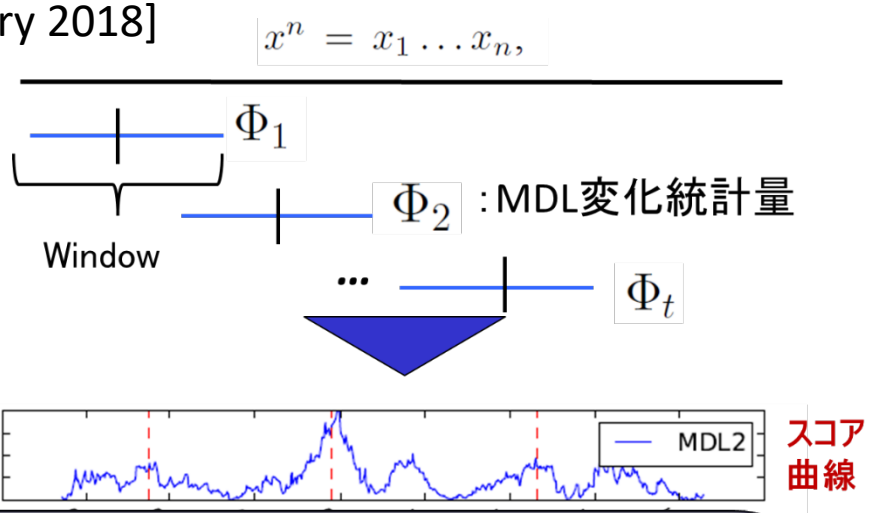
潜在的ダイナミクス1

MDL変化統計量による変化予兆検知

データ圧縮度合いで漸進的変化の開始点（予兆）を検知

[Yamanishi, Miyaguchi BigData2016] [Kaneko, Miyaguchi. Yamanishi BigData2017]
[Yamanishi, Fukushima IEEE Trans Inform Theory 2018]





技術コア： MDL変化統計量によるオンライン変化検知を提案

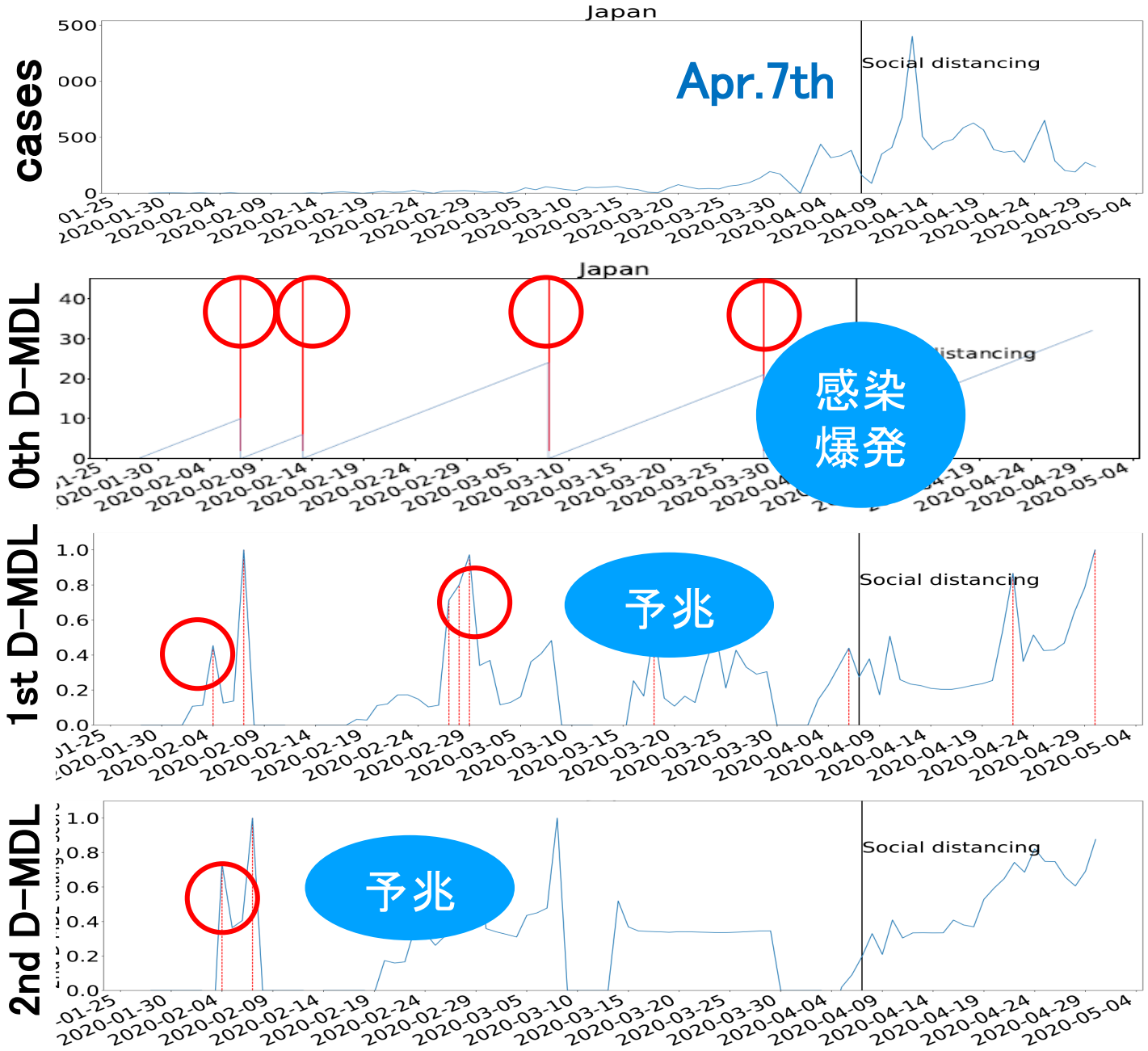
MDL：Minimum Description Length

$$\Phi_t = \min_{M} L_{\mathrm{NML}}(x_1^n; M) - \min_{M_1, M_2} \{L_{\mathrm{NML}}(x_1^t; M_1) + L_{\mathrm{NML}}(x_{t+1}^n; M_2)\}$$

変化した場合としない場合の記述長の差

⇒
- 固定窓で移動計算、漸進的変化検知 [BigData 2016]
- 可変窓で移動計算、検知精度を大幅改善 [BigData 2017]
- 工場のボイラー事故の予兆を検知（東レとの実証実験）

微分的MDL変化統計量による感染爆発検知



[Yamanishi et al. Scientific Reports 2021]

- **感染爆発**
  緊急事態宣言(4/7)前に発現
  (R0 >1)

- **MDL変化統計量アラート**
  感染爆発時点に対応 (R0>1)

- **1次のD-MDLアラート**
  変化の速度
  感染爆発前に上がっている

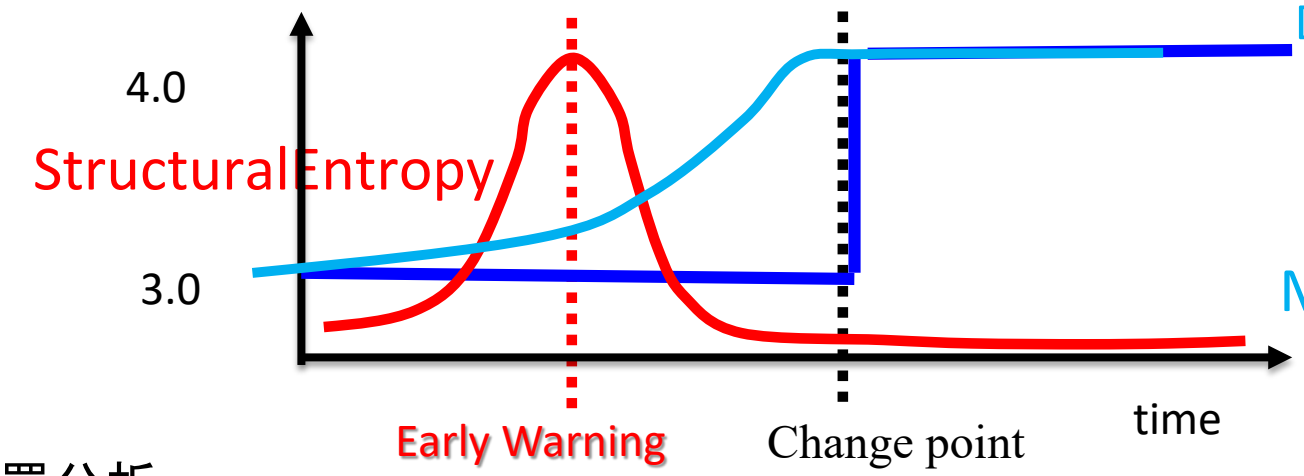
- **2次のD-MDLアラート**
  変化の加速度
  感染爆発前に上がっている

潜在的ダイナミクス2

潜在構造変化予兆検知

連続量（構造エントロピー、Ddim, MC）でクラスタリング構造変化の予兆を検知

[Hirai Yamanishi BigData2018, BigData2019, BigData2021]



Descriptive dimension（Ddim)

$$\mathrm{Ddim}_t = \sum_k p(k|\boldsymbol{x})k$$

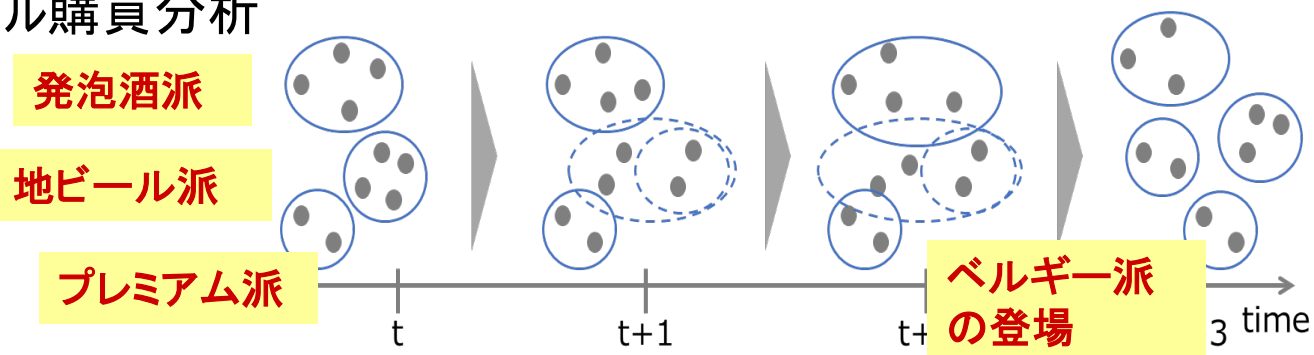
[Yamanishi et al.arxiv 23]

Mixture Complexity（MC）

$$\mathrm{MC}_t = \exp(\hat{I}(Z) - \hat{I}(X|Z))$$

[Kyoya Yamanishi Entropy22]

ビール購買分析



マーケット構造変化予兆検知

電力消費パタン変化予兆検知で有効性検証

StructuralEntropy

[Hirai Yamanishi Bigdata18]

$$H_t = -\sum_k p(k|\boldsymbol{x}) \log p(k|\boldsymbol{x}_t) \qquad p(k|\boldsymbol{x}_t) = \frac{\exp(-\beta L_{\mathrm{NML}}(\boldsymbol{x}_t; k))}{\sum_{k'} \exp(-\beta L_{\mathrm{NML}}(\boldsymbol{x}_t; k'))}$$

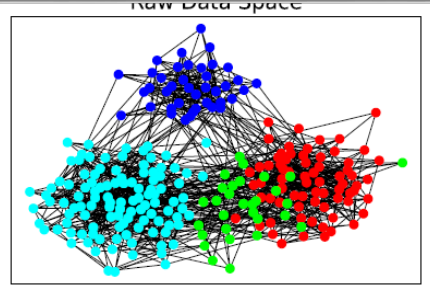
潜在的ダイナミクス3

ネットワーク変化検知

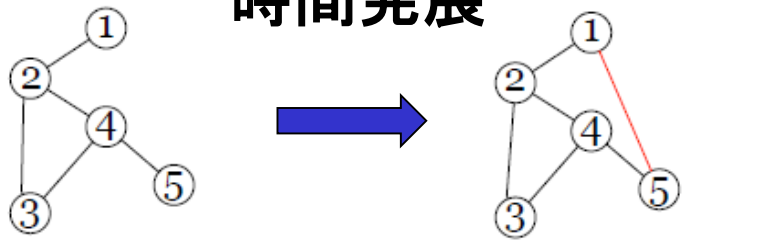
「Lin, Xu, Yamanishi *IEEE TKDE* 2021]

RWiLS: グラフを潜在空間に埋め込み、ランダムウォークの定常ベクトルを変化検知

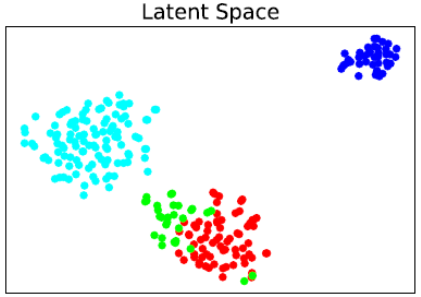
実空間

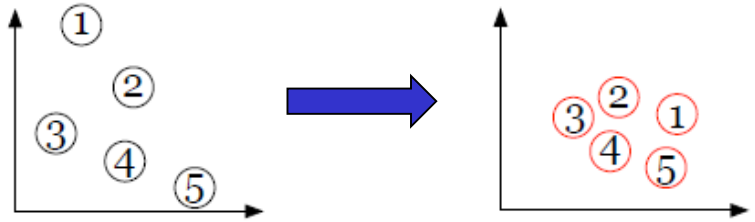


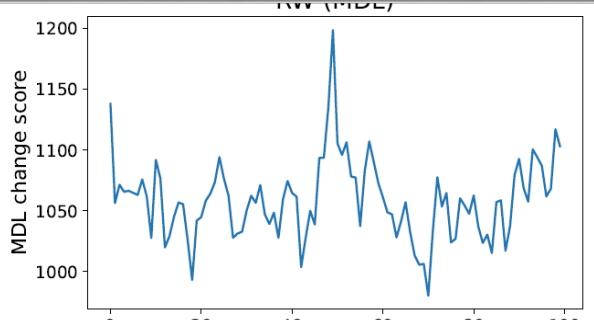
時間発展



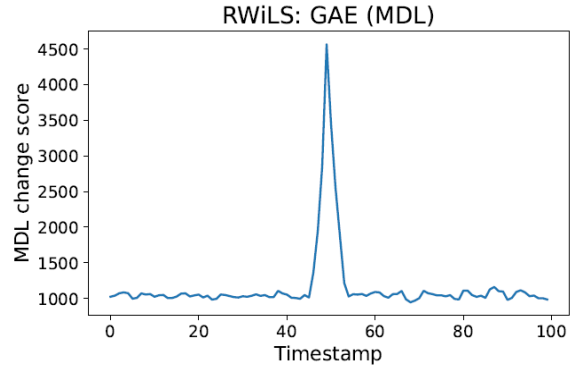
潜在空間





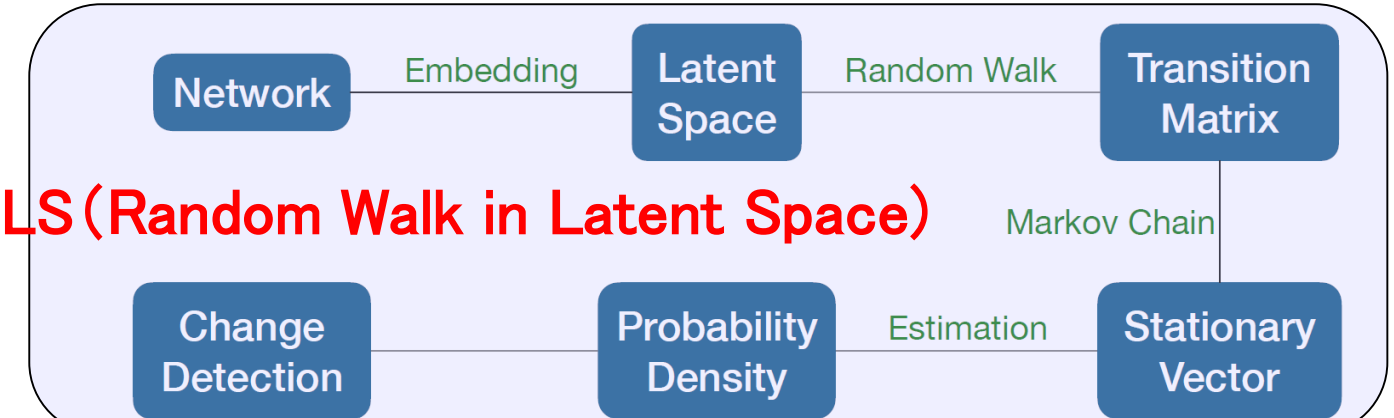


実空間の変化検知



潜在空間の変化検知

よりロバスト

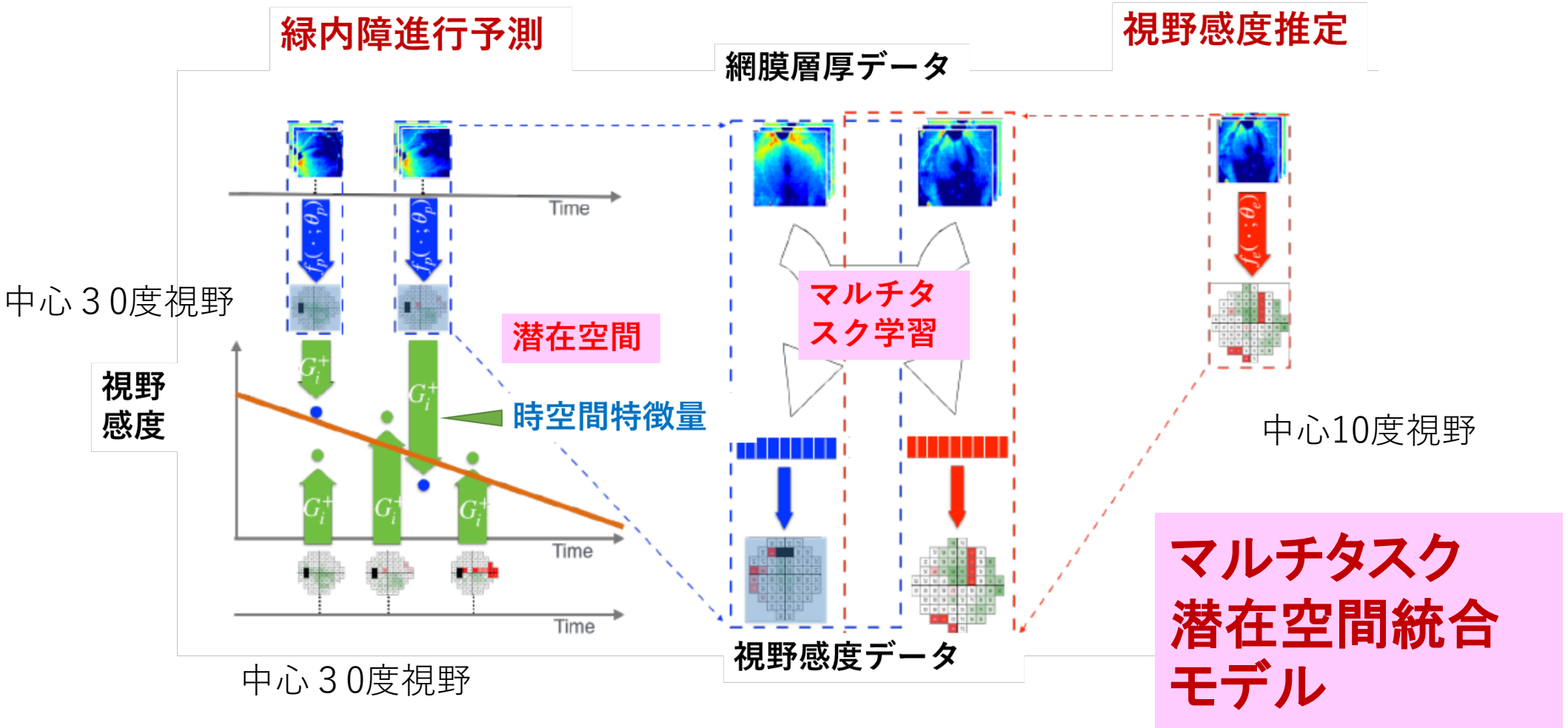


Δd_{exp}	0.01	0.1	1	10
DynGEM	0.4900	0.3940	0.6007	0.6200
EEC	0.5053	0.5096	0.6165	0.7755
DeltaCon	0.4616	0.6143	0.4313	0.7636
SACPD	0.5084	0.5740	0.7280	0.7276
LAD	0.5206	0.5631	0.5208	0.5948
RWiLS(MDL) with DeepWalk	0.8005	0.9433	0.9443	0.9600
RWiLS(KL) with DeepWalk	0.8002	0.9395	0.9047	0.8212
RWiLS(MDL) with GAE	0.8660	**0.9709**	**0.9700**	**0.9687**
RWiLS(KL) with GAE	**0.9373**	0.9462	0.9584	0.8716

緑内障進行予測

マルチタスク潜在回帰モデルにより視野感度推定と予測の両者で精度増強

[Xu, Asaoka, Kiwaki, Murata, Fujino, Yamanishi KDD2021]

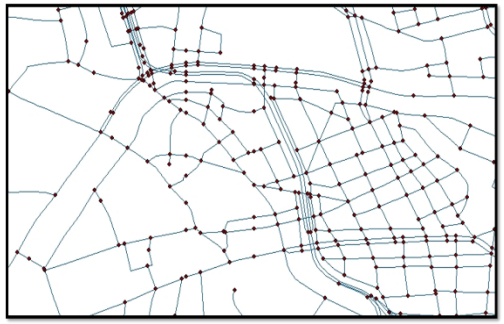


交通リスクマイニング

地形データをネットワーク中心性で数値化、危険パタンを特徴づけ

[Kobayashi, Lee, Matsushima, Yamanishi BigData 2017]
[Lee, Matsushima, Yamanishi DAMI 2019]

地形データ



高速パタンマイニングに基づく分類

$$\min_{\boldsymbol{w}} C \sum_{i=1}^{m} L\left(\sum_{\phi \in \Phi^{(k)}} w_\phi \phi(\boldsymbol{x}_i), y_i\right) + \|\boldsymbol{w}\|_1$$

ただし, $L(f(\boldsymbol{x}), y) = \log(1 + \exp(-yf(\boldsymbol{x}))$

Network中心性で数値表現
（次数、近傍ノード数、ページランク、近傍中心性、平均リンク数 etc)
⇓
低コストデータ

- 危険パタンの抽出
- 潜在的危険個所の特定

ビッグデータからコンパクトなパタンを高速に抽出