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/

2022/5/21

大学院入試説明会

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）

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

特任講師　木脇太一

専門：ニューラルネット

講義：データマニング実践演習

助教　池祐一

専門：位相的データ解析（TDA）

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
学習数理情報学をめざします

■幅広い産業界との連携のもとで、研究を進めます
以下の企業とはすでに協力を開始しています。

Microsoft
IBM
東大病院 The University of Tokyo Hospital
三菱商事
NTT
HAKUHODO
LAC Little eArth Corporation
sas THE POWER TO KNOW.
YAHOO! JAPAN
ネットリサーチなら、マクロミル macromill
VALUES Consulting & Creation Group
TORAY Innovation by Chemistry
MIZUHO
みずほ銀行
HONDA The Power of Dreams
MODEC
SOMPO 損保ジャパン
NIPPON STEEL
社会福祉法人 聖隷福祉事業団 総合病院 聖隷浜松病院 SEIREI HAMAMATSU GENERAL HOSPITAL

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

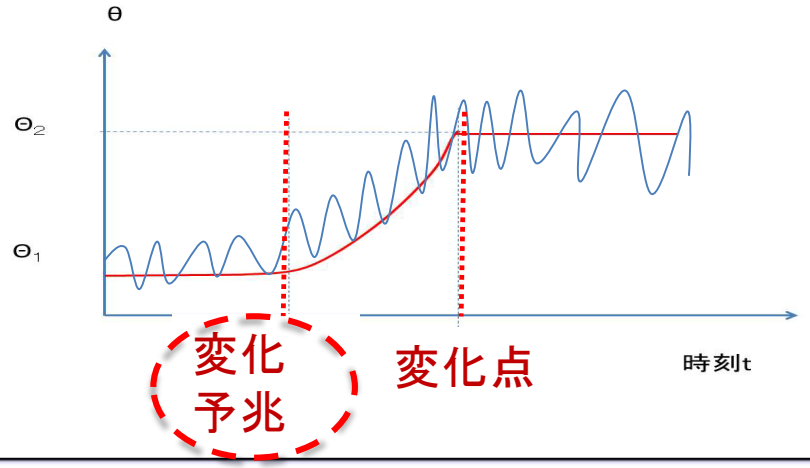
量子機械学習

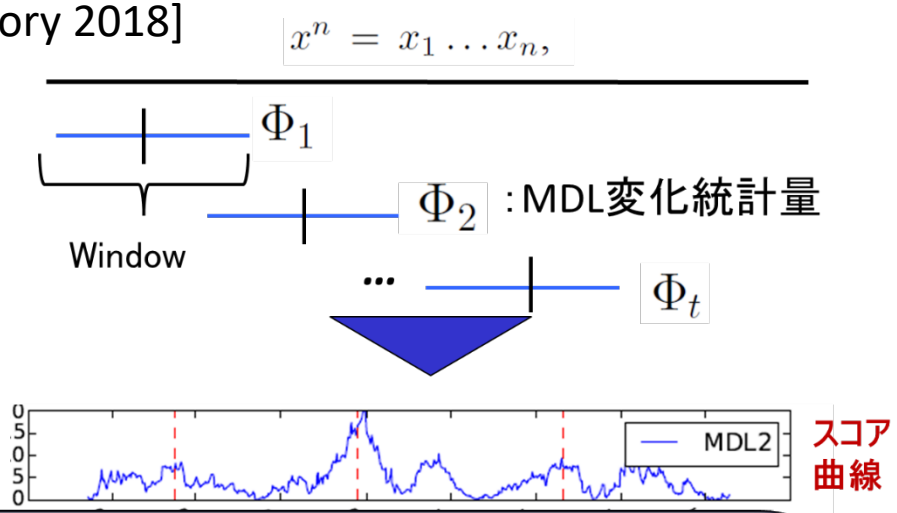
潜在的ダイナミクス1

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

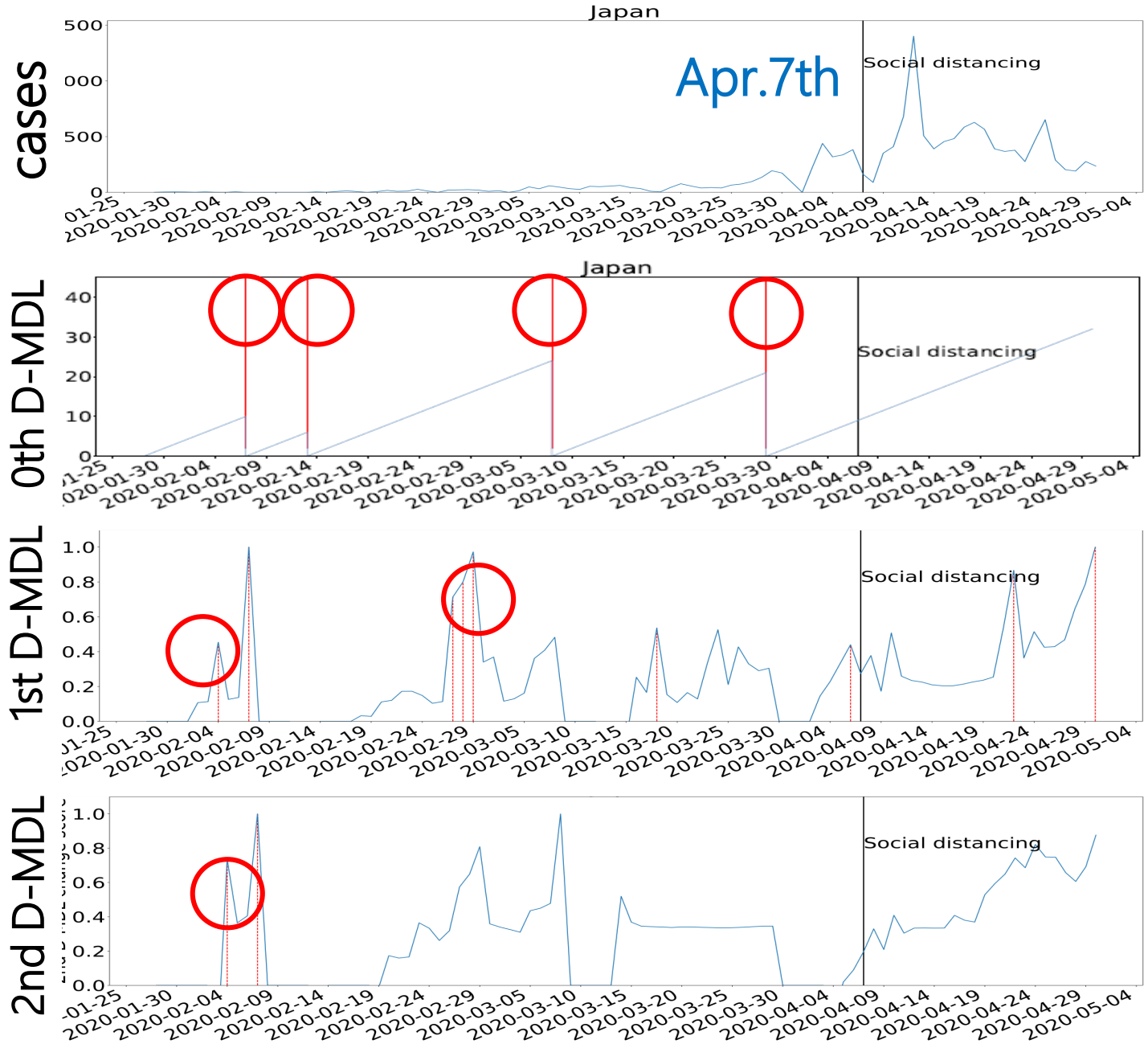
MDL：Minimum Description Length

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

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COVID-19感染爆発予兆検知



[Yamanishi et al. Scientific Reports 2021]

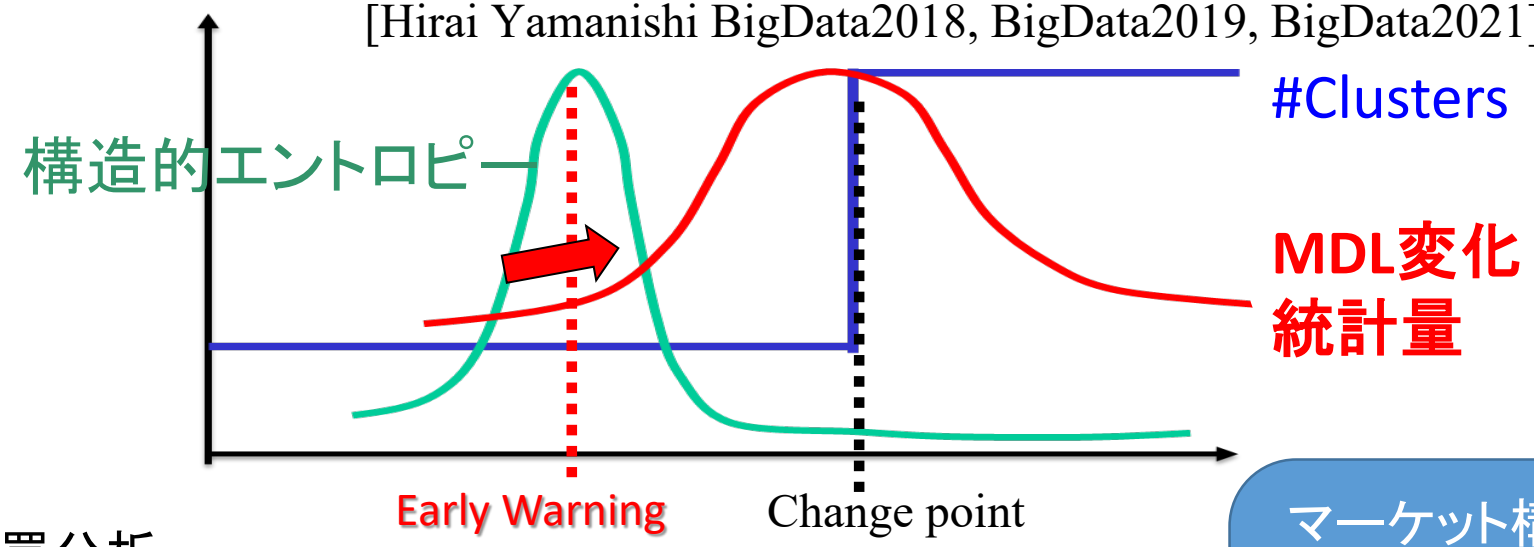
- 感染爆発
  緊急事態宣言(4/7)前に発現
  (R0 >1)
- MDL変化統計量アラート
  感染爆発時点に対応 (R0>1)
- 1次のD-MDLアラート
  変化の速度
  感染爆発前に上がっている
- 2次のD-MDLアラート
  変化の加速度
  感染爆発前に上がっている

潜在的ダイナミクス2

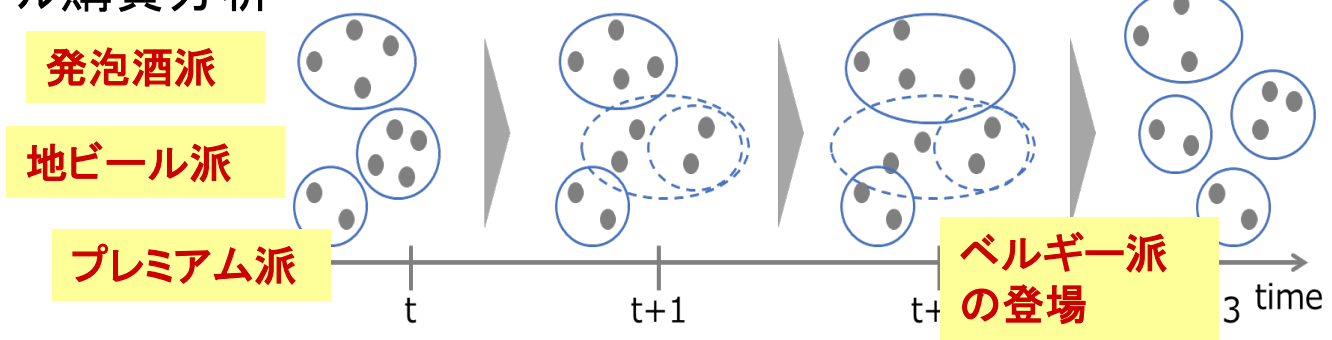
潜在構造変化予兆検知

連続量（構造エントロピー、MDL変化統計量）で変化過渡期を検知

[Hirai Yamanishi BigData2018, BigData2019, BigData2021]



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

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

潜在的ダイナミクス3

階層的変化検知

変化はどこから来るのか？　要因を分析し、変化の感度分析を実現

[Fukushima Yamanishi ICDM2020, ComplexSystem 2021]

潜在変数モデル

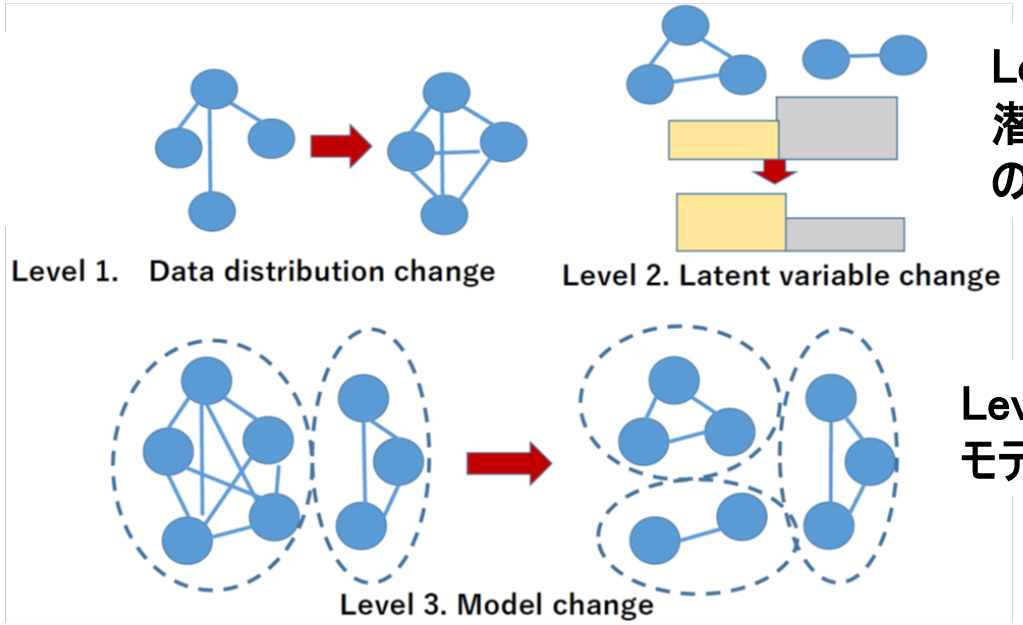
$$(x_t, z_t) \sim f(X, Z; \theta, M),$$

$$f(X, Z; \theta, M) = f(X|Z; \theta_1, M) f(Z; \theta_2, M)$$

X: observed variable
Z: latent variable
$\theta = (\theta_1, \theta_2)$

Level1 パラメータの変化

θ_1



Level2 潜在変数の変化

θ_2

Level3 モデルの変化

M

人口移動データ分析、センサーデータ解析で有効性実証

分解型変化スコア（MDL変化統計量＋DNML基準）

$$\Phi_t = \Phi_t^{X|Z} + \Phi_t^{Z} + \Delta L_t(\hat{M}, \hat{M}_1, \hat{M}_2)$$

パラメータのMDL変化統計量

潜在変数のMDL変化統計量

モデルのMDL変化統計量（全体）

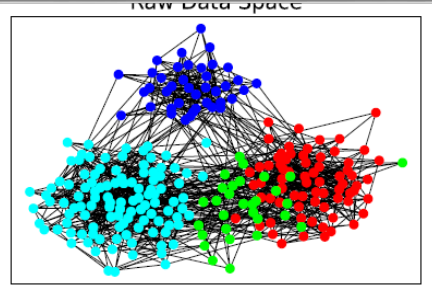
潜在的ダイナミクス4

ネットワーク変化検知

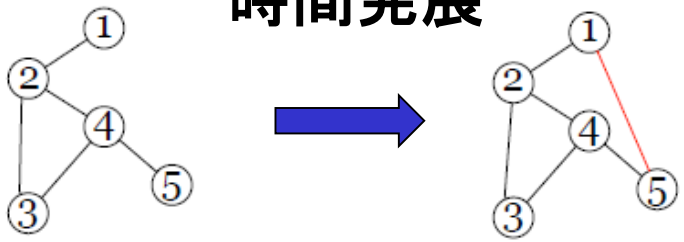
「Lin, Xu, Yamanishi *IEEE TKDE* 2021]

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

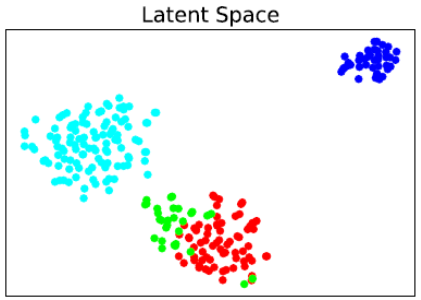
実空間

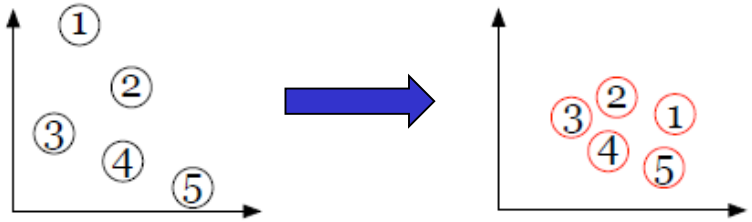


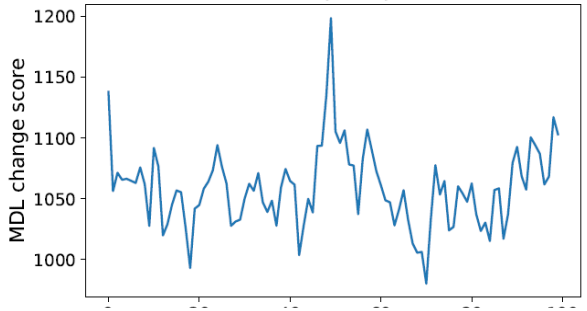
時間発展



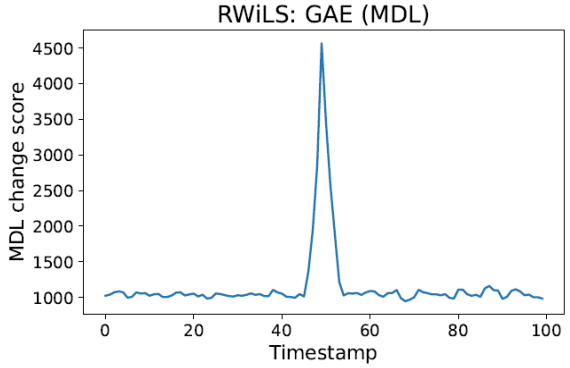
潜在空間





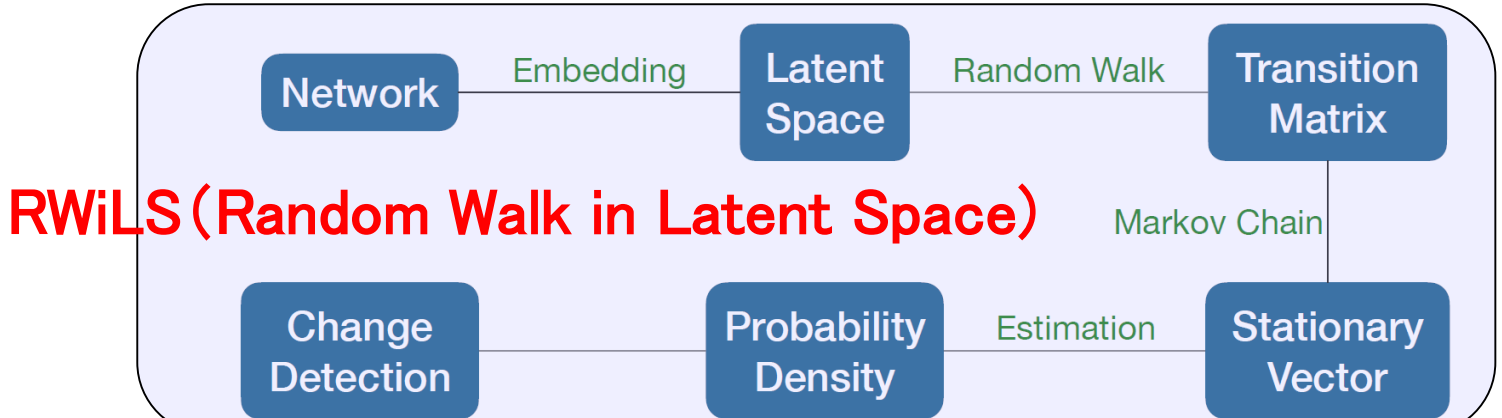


実空間の変化検知



潜在空間の変化検知

よりロバスト

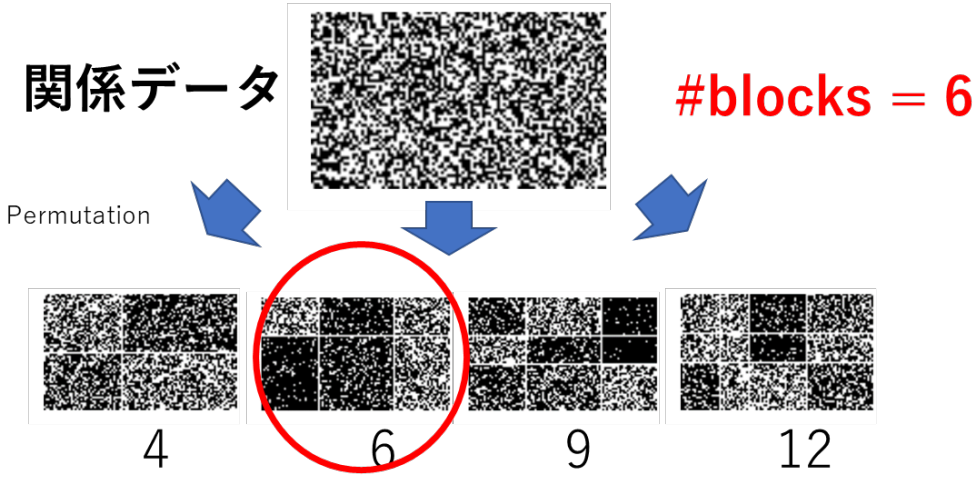


Δd_{exp}	0.01	0.1	1	10
DynGEM	0.4900	0.3940	0.6007	0.6200
EEC	0.5053	0.5096	0.6165	0.7755
DeltaCon	0.4616	0.6143	0.4313	0.7636
SACPD	0.5084	0.5740	0.7280	0.7276
LAD	0.5206	0.5631	0.5208	0.5948
RWiLS(MDL) with DeepWalk	0.8005	0.9433	0.9443	0.9600
RWiLS(KL) with DeepWalk	0.8002	0.9395	0.9047	0.8212
RWiLS(MDL)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

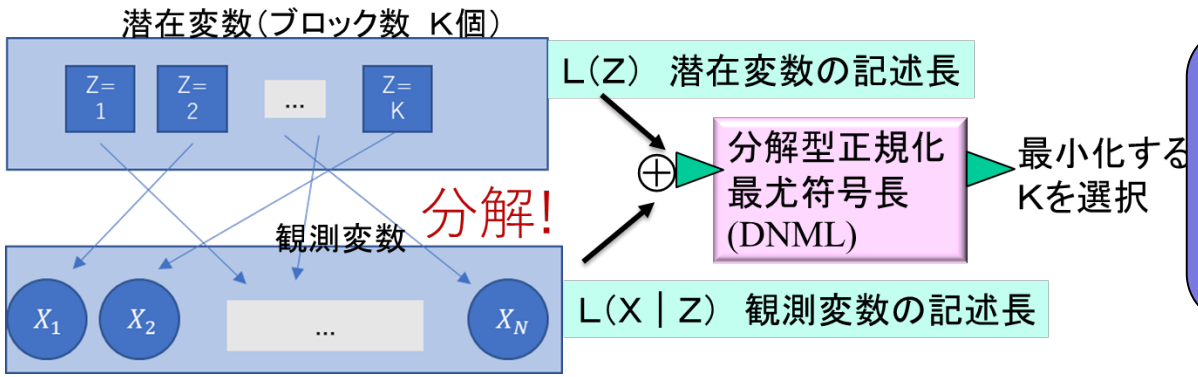
潜在構造最適化 1

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

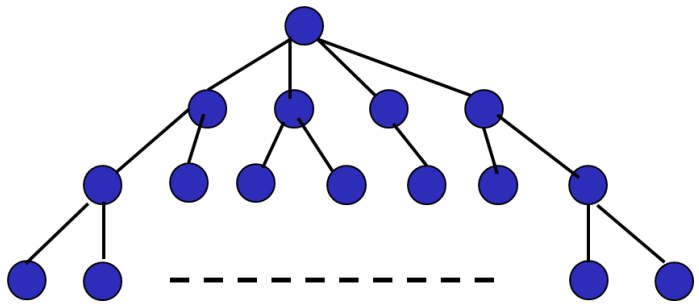
$$\min_{K} \{L_{\mathrm{NML}}(X|Z;K) + L_{\mathrm{NML}}(Z;K)\}$$

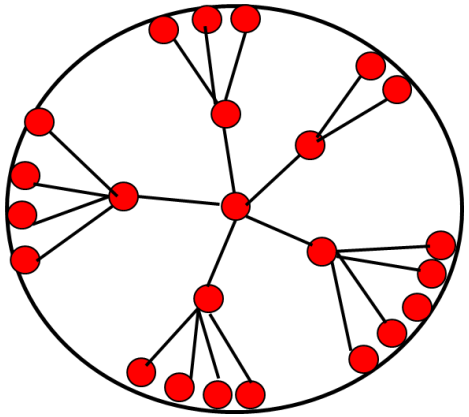
潜在構造最適化 2
双極順序付き埋め込み

順序関係を保ちつつ双極空間に埋め込むには？

[Suzuki, Wang, Tian, Nitanda, Yamanishi ACMLI19]
[Wang, Suzuki, Xu, Fei, Yamanishi IJCAI2019]

- Embedding: map of objects to a metric space





高次元ユークリッド空間

低次元非ユークリッド空間（双曲空間）

- Hyperbolic Ordinary Embedding

$i \leftrightarrow j$ is more similar than $i \leftrightarrow k \Rightarrow d_{\mathbb{H}^D}(x_i, x_j) \leq d_{\mathbb{H}^D}(x_i, x_k)$

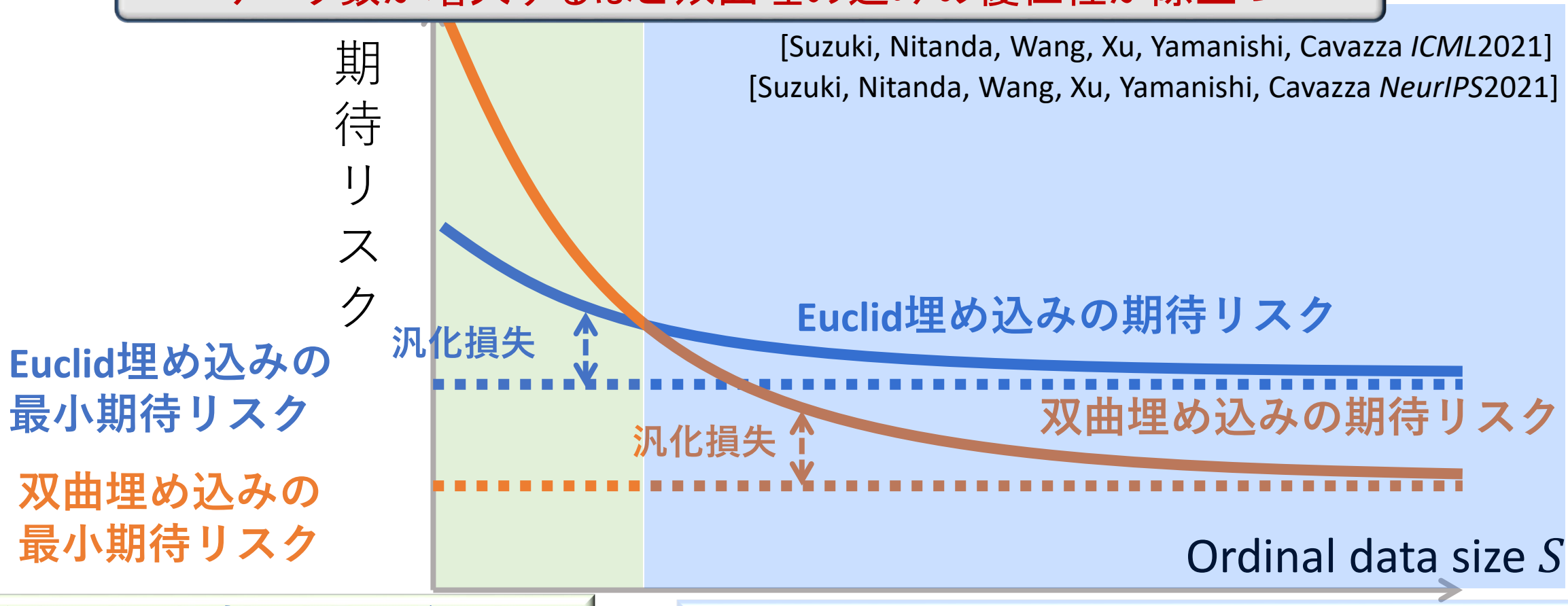
- Distance function $d_{\mathbb{H}^D}(\boldsymbol{x}, \boldsymbol{y}) = \operatorname{arcosh}\left(1 + \frac{2(\boldsymbol{x}-\boldsymbol{y})^\top(\boldsymbol{x}-\boldsymbol{y})}{(1-\boldsymbol{x}^\top\boldsymbol{x})(1-\boldsymbol{y}^\top\boldsymbol{y})}\right)$

効果的な可視化とユークリッド空間よりも高精度なリンク予測を実現

Euclid埋め込みVS双曲埋め込み

―汎化誤差の観点から―

データ数が増大するほど双曲埋め込みの優位性が際立つ



S が小➡ **Euclid埋め込みがよい**
Euclidの汎化誤差が双曲のそれよりも小さい

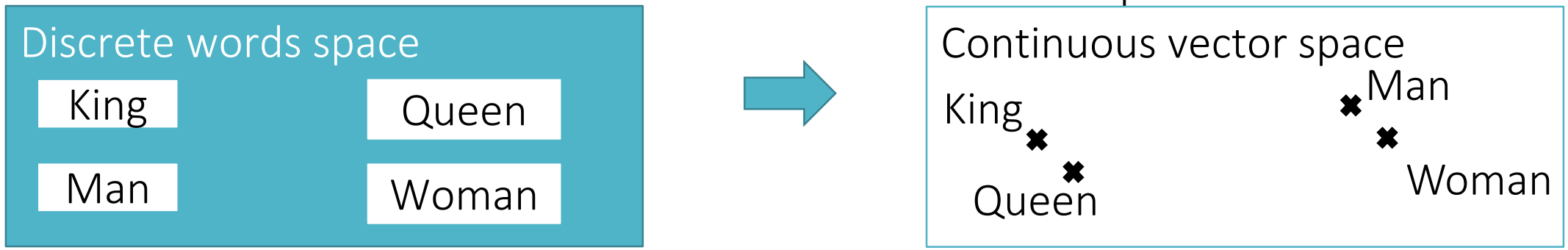
S が大➡ **双曲埋め込みがよい**
∵ 双曲埋め込みのEmpirical Risk minimizerの期待リスクが、Euclidのそれよりも小さい

潜在構造最適化 3

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]



Model performance

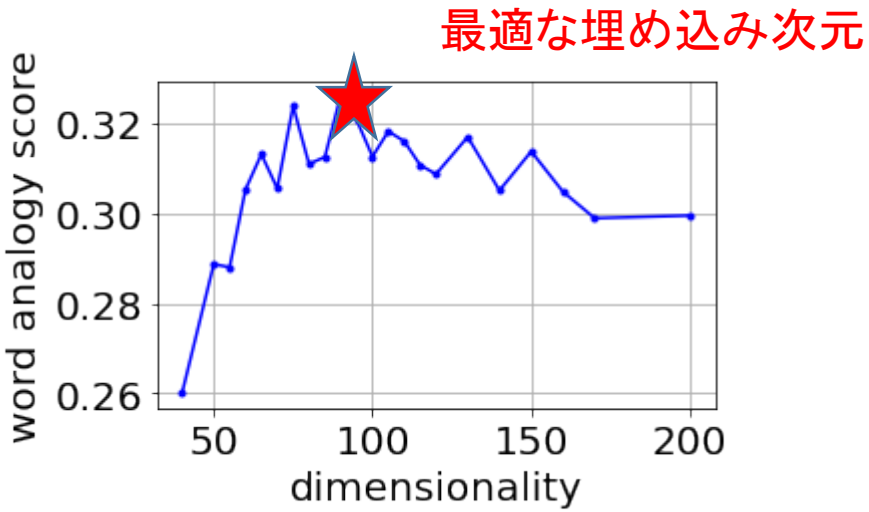
$$\overrightarrow{King} - \overrightarrow{Queen} = \overrightarrow{Man} - \overrightarrow{Woman}$$

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

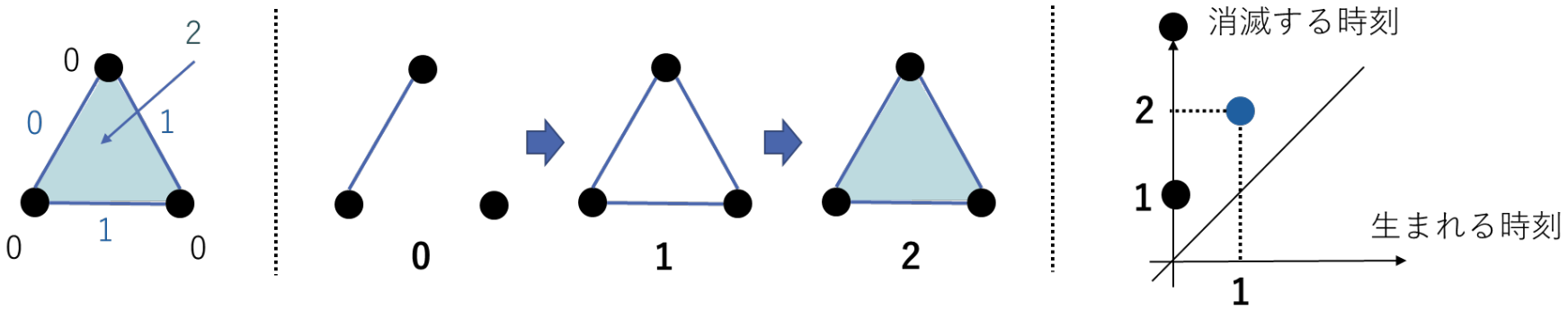


位相的データ解析
（Topological Data Analysis, TDA）

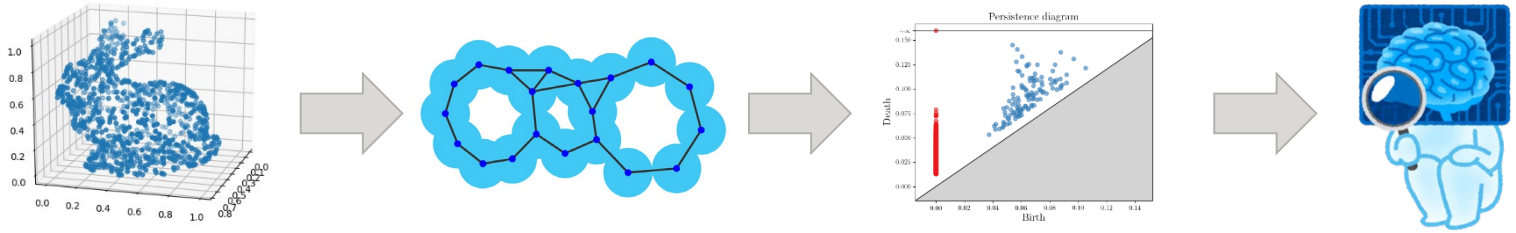
■データの**「形（トポロジー）」を抽出して解析**する手法

■「形」を捉えるためフィルトレーションという増大列構造を作る

■トポロジー変化を「穴」の生成・消滅時刻としてパーシステンス図に出力



■パーシステンス図を機械学習等と組み合わせてデータを解析する



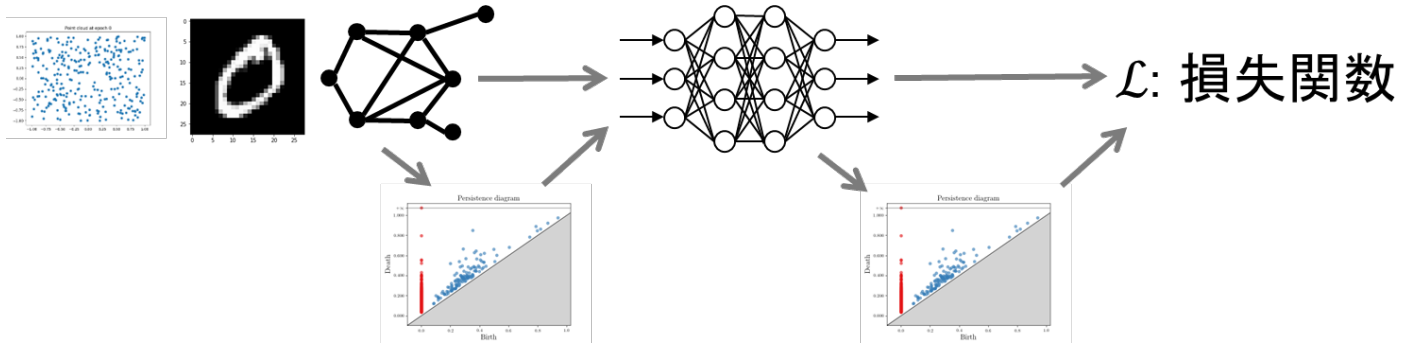
データ　**フィルトレーション**　**パーシステンス図**　**機械学習**

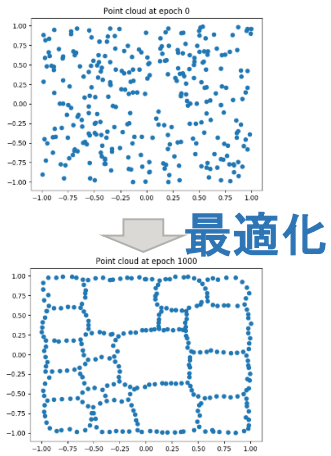
■位相的データ解析はおおまかな「形」に着目するので**ノイズ耐性**がある

位相的データ解析（TDA）研究内容の例

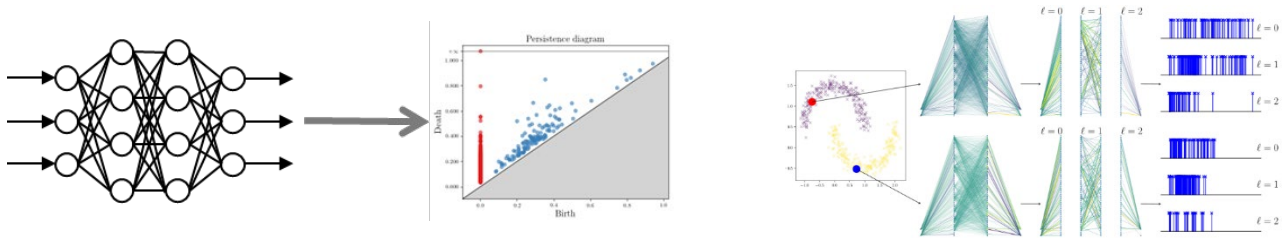
■**点群・画像・グラフ**データ解析への応用

■**トポロジー的損失関数**の設計と最適化

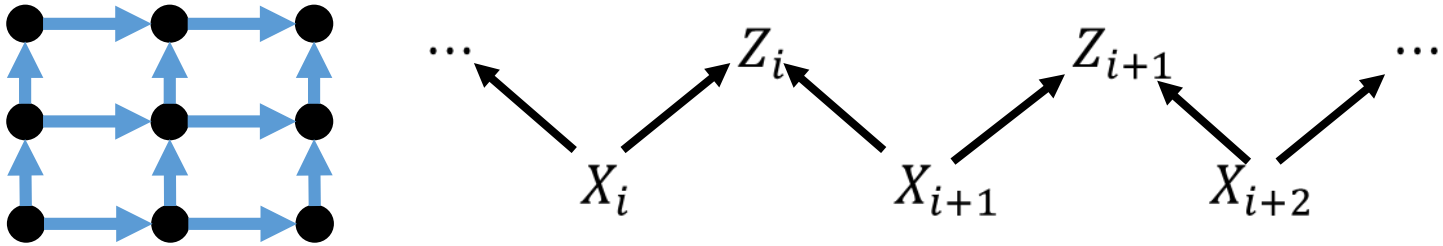




■**学習器の状態**を**トポロジー的に記述**



■**多変数のパーシステンス加群**からの情報抽出

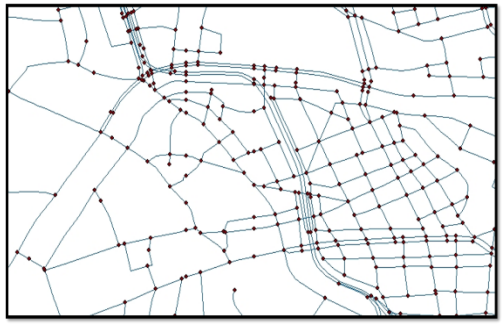


交通リスクマイニング

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

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

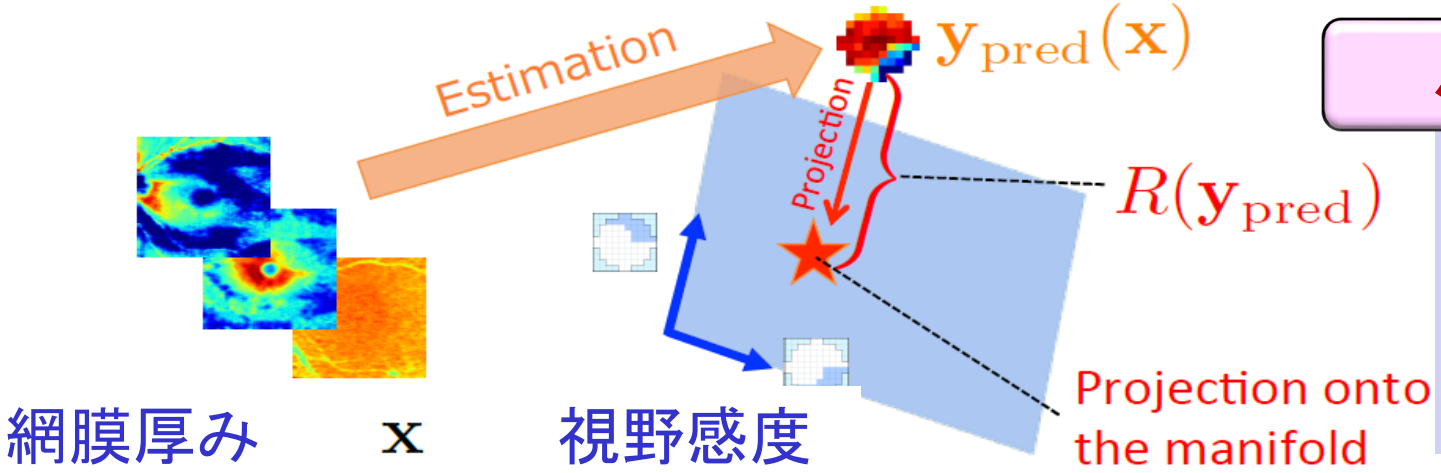
緑内障進行予測 1

パターン正則化

パターン正則化により網膜厚みから視野感度を高精度に推定

[Sugiura,Kiwaki, Siamak. Murata,Asaoka, Yamanishi KDD2018]

[Xu, et al. American Jr. Opht 2020][Hashimoto et al. British Jr.Opht.2020]



パターン正則化

教師なしで学習した視野パタンからずれないように正則化

平均二乗誤差

	Median	Mean	Worst
CNN [30]	6.68	6.76	21.7
CNN + DA	6.70	6.52	21.7
CNN + DA + PBR1(NMF)	6.41	6.30	23.5
CNN + DA + PBR1(PCA)	6.31	6.26	24.1
CNN + DA + PBR1(AE)	6.39	6.26	24.2
CNN + DA + PBR1(PCA) + PBR2	**6.00**	**6.16**	21.4
CNN + DA + PBR1(AE) + PBR2	6.02	**6.16**	21.5

提案手法の最新結果

$$\frac{1}{N}\sum_{n=1}^{N}\|\boldsymbol{y}_{\text{pred}}(\boldsymbol{x}_n)-\boldsymbol{y}_n\|^2+\frac{\beta}{2}R(\boldsymbol{y}_{\text{pred}}(\boldsymbol{x}_n)),$$

$$\boldsymbol{y}_{\text{proj}} \stackrel{\text{def}}{=} \Theta^\top(\Theta\Theta^\top)^{-1}\Theta(\boldsymbol{y}_{\text{pred}}-\boldsymbol{b})+\boldsymbol{b}.$$

東大附属病院眼科教室との共同研究

緑内障進行予測 2

DLLR(deeply regularized latent space linear regression)

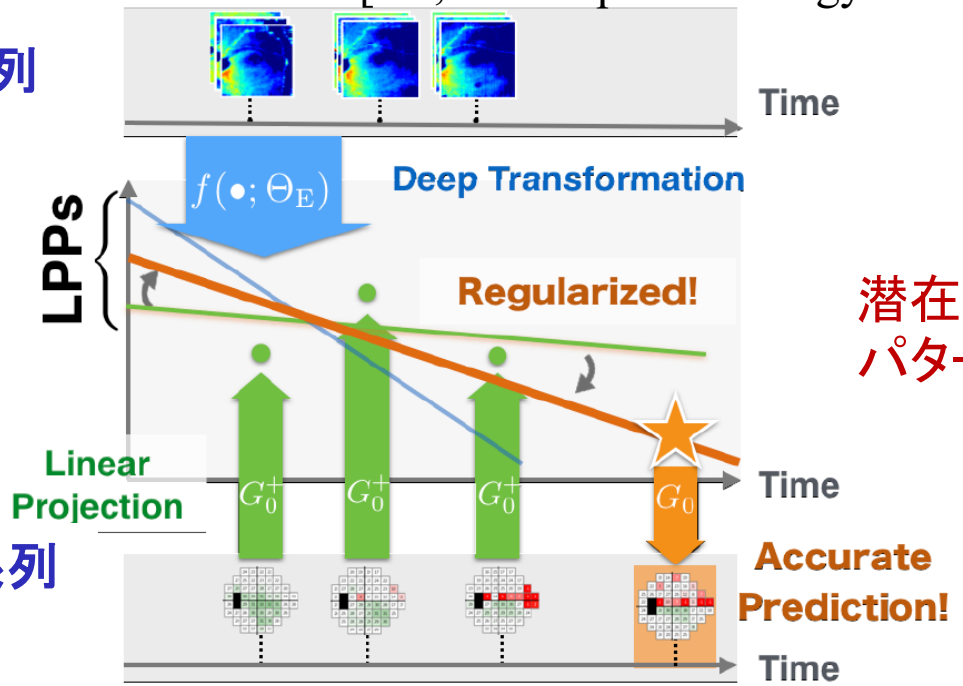
網膜厚と視野感度の時系列の双方を潜在空間で統合、予測

[Tei, Xu, Kiwaki, Wang, Murata, Asaoka, Yamanishi KDD2019, KDD2021]
]
[Xu, et al. Ophthalmology Glaucoma 2020]

網膜厚時系列

潜在空間

視野感度時系列



潜在進行パターン

予測誤差(RMSE)比較

Methods	S						
	2	3	4	5	6	7	8
LR	50.66	21.11	13.21	9.76	7.98	6.52	6.06
TSLR	6.27	6.01	5.65	5.54	5.52	5.09	5.02
MTMC	5.65	5.47	4.99	4.82	4.83	4.67	4.62
DLLR	**4.96**	**4.76**	**4.61**	**4.49**	**4.48**	**4.46**	**4.43**

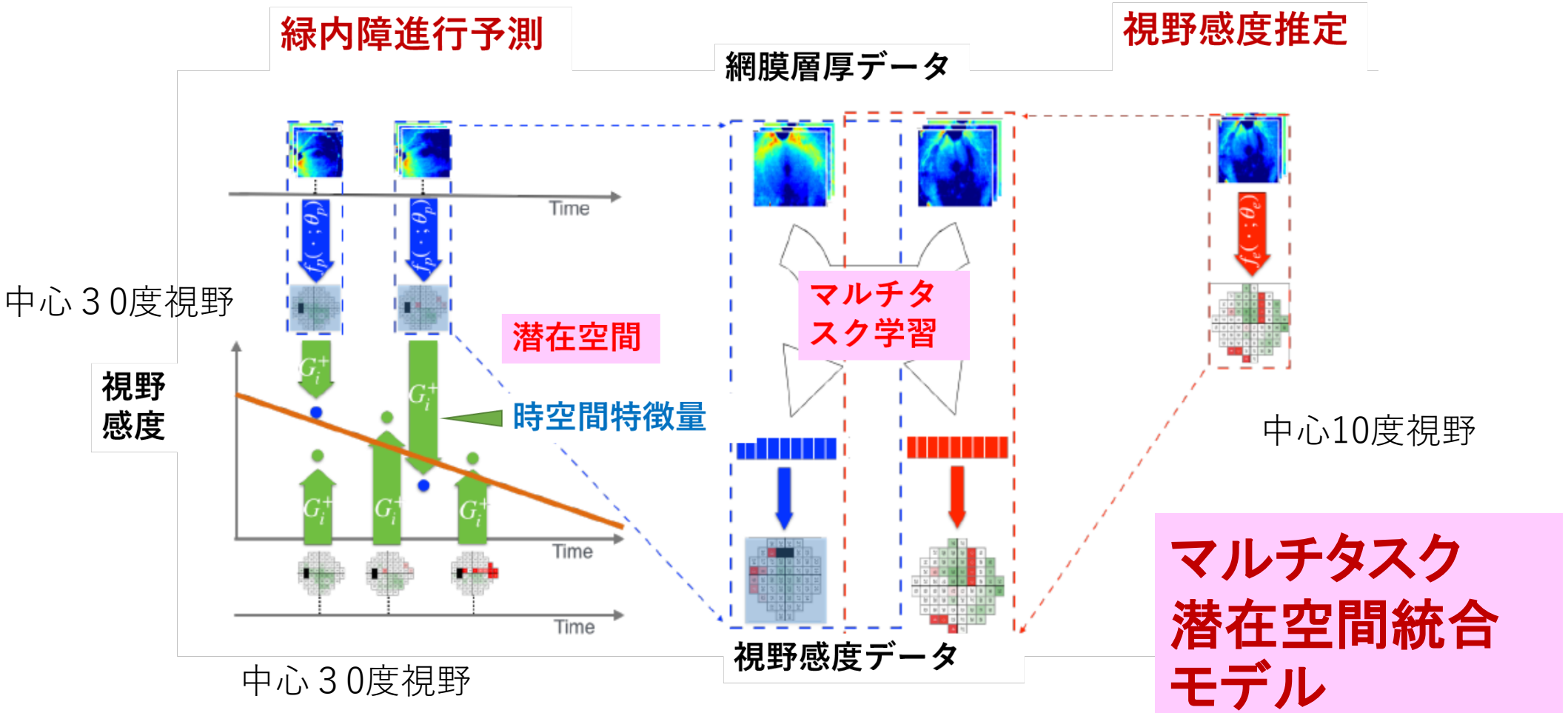
視野感度と網膜厚の双方の活用で最高精度達成

緑内障進行予測 3

マルチタスク学習で視野感度予測と推定を統合

マルチタスク潜在回帰モデルにより推定と予測の相乗効果で精度増強

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



東京大学大学院情報理工学系研究科
データサイエンティスト養成講座
(Data Science School: DSS)

数理6研はDSSを支援する主力研究室です。

データサイエンスでは
いかに分析するか(How)
だけではなく
何を分析するか(What)
が重要です

講座の模様。　画像提供：山西研究室

2015－2017の修論テーマ

- ネットワーク中心性を用いたテンポラルネットワークの変化検知
- ヘテロデータに基づく緑内障進行予測のための個々の患者に特化した病状のモデル化
- 記述長最小化原理に基づく関係データ学習の研究
- Support Vector Machine 及びトピックモデルを用いた位置情報による個人識別の研究
- 完全変数化NML 符号長に基づく正準相関分析のランク推定の研究
- Study on Learning from Nonstationary Time Series
- サポートベクトルマシンの最適化アルゴリズムの研究
- 野球データ解析に対する機械学習アプローチ
- 交通データからのリスク因子発見に関する研究
- 緑内障眼における網膜層厚視野感度間マルチビュー学習
- A Study on Model Selection for Hierarchical Mixture Models
- 畳み込み行列分解のパラメータ推定と構造選択
- Model Selection for Non-negative Tensor Factorization with MDL

DSAA16採択

Healthinf16採択

2015研究科長賞（創造）

SDM16採択

DSAA17採択

DSAA15採択

DSAA16採択

2016研究科長賞（創造）

KDD17採択

2016研究科長賞（数理）

KDD17採択

ICDM16採択

Entropy掲載

2017-2021の修論テーマ

- パターン正則化学習を用いた網膜層厚からの視野感度推定と知識発見 KDD18採択
- 分解型正規化最尤符号長を用いた連続値潜在変数モデルの選択 DAMI掲載 RSOS採択
- 統計的モデル選択を用いた勾配ブースティング木の早期停止ルールの研究
- A machine learning approach to glaucoma progression prediction KDD19採択
- 双曲空間における測地線更新アルゴリズムの研究 2016研究科長賞（創造） AIPラボ長賞
- グラフ埋め込みにおける正則化の研究
- 敵対的学習における目的関数の設計及び解析に関する研究
- Dynamic Nondiagonal Mixture of Dirichlet Distributions for Economic Network Comp Net採択
- Network Change Detection Based on Random Walk in Latent Space IEEE TKDE掲載
- Cluster Structure Analysis in Finite Mixture Models With Component Overlap Entropy掲載
- Change Sign Detection with Two-Stage MDL Change Statistics
- Word2vec Skip-gram Dimensionality Selection via Information Criteria Entropy掲載
- クリギング法を用いたゼロ次最適化アルゴリズムによる量子状態作成
- ネットワーク上を伝播するイベント時系列データのクラスタリング

2017-2020の博論テーマ

- Learning High-dimensional Models with the Minimum Description Length Principle

SDM18採択　ML掲載　AISTATS19採択

- Machine Learning over Space Forms

ISIT18採択　IEEE IT掲載　ACML18採択

- Detecting Model Changes and their Early Warning Signals with the Minimum Description Length Principle

BigData18採択　BigData19採択　BigData21採択

- An Information-Theoretic Approach to Hierarchical Change Detection

Entropy掲載　ICDM20採択　Comp Net採択

2016-2020 の卒論テーマ

- 可変窓上の記述長最小化に基づく時系列データの変化検知 BigData17採択
- ネットワーク情報を用いた道路の潜在的危険予測 BigData17採択
- オンライン非定常ロバスト学習とその異常検知への応用 2017工学部長賞（計数）
- 複数の関係パターンを持つマルコフ確率場のMDL 原理に基づく構造学習
- 緑内障診断のための網膜厚に対する視野感度の微分値の異常検知
- 混合分布を用いる パターン正則化学習と その眼科学への応用
- 潜在空間埋め込みを用いる ベイズ実験計画の研究
- 有限混合モデルの変化予兆検知とCOVID-19 の感染爆発検知への応用
- 位相的データ解析を用いた時系列データの変化点検知
- 予測モデル組み込みによる変化検知の早期化と COVID-19 データへの適用