



Causal Inference in Time Series via Supervised Learning [FIT2019 Top Conference Session]

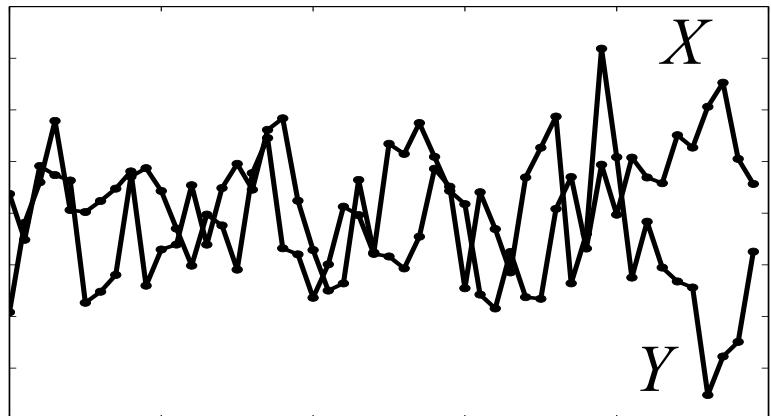
Yoichi Chikahara, Akinori Fujino

NTT Communication Science Laboratories
Kyoto, Japan

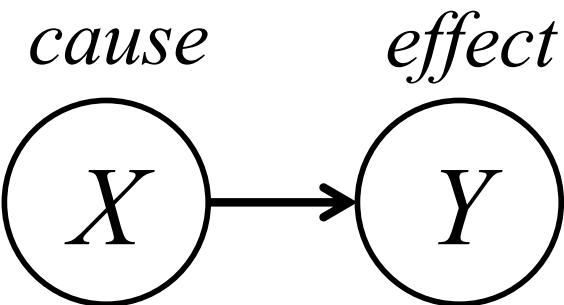
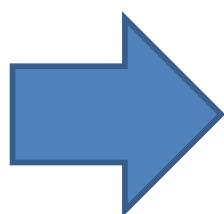
時系列の因果推論



- 与えられた時系列データから
変数間の因果関係を推定するタスク



入力: 時系列データ



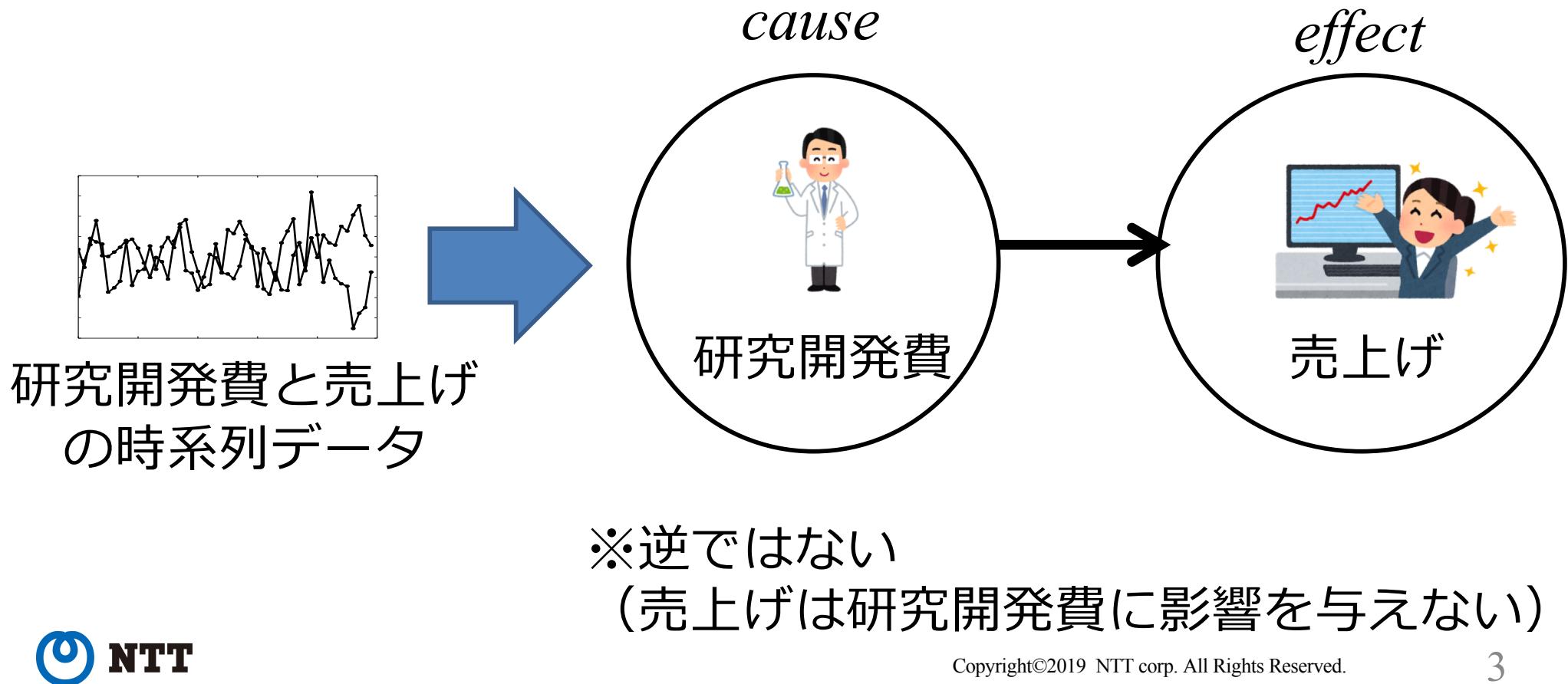
出力: 因果関係

目的: 因果関係に関する知識の発見

Application 1: Economics



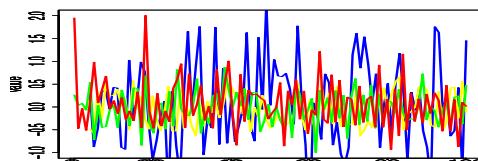
- 「研究開発費は売上金に影響を与える」
という知識は、企業にとって有用



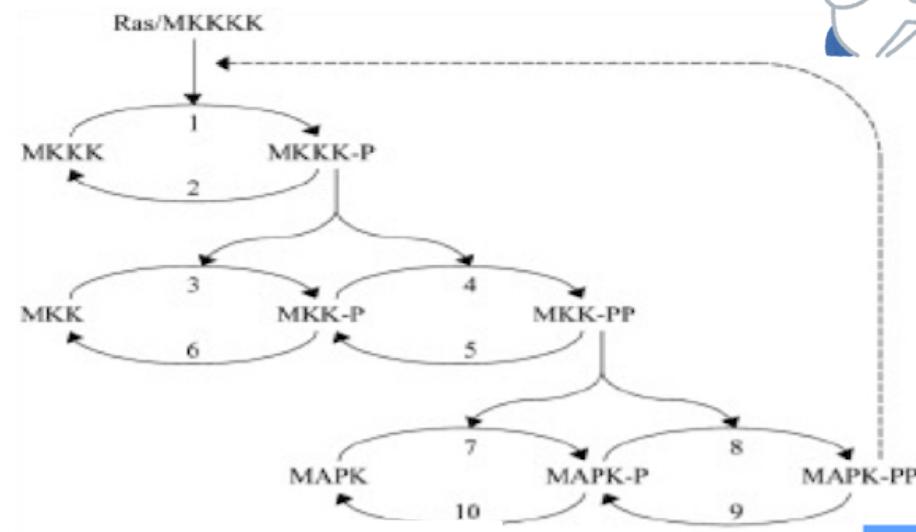
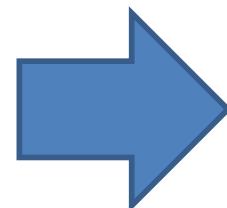
Application 2: Bioinformatics



- 遺伝子の発現量の時系列データから「遺伝子の制御関係」がわかると新規薬剤の探索に役立つ



遺伝子発現量
に関するデータ



遺伝子の制御関係

“因果関係”って何？

どう定義されるの？

A definition of temporal causality

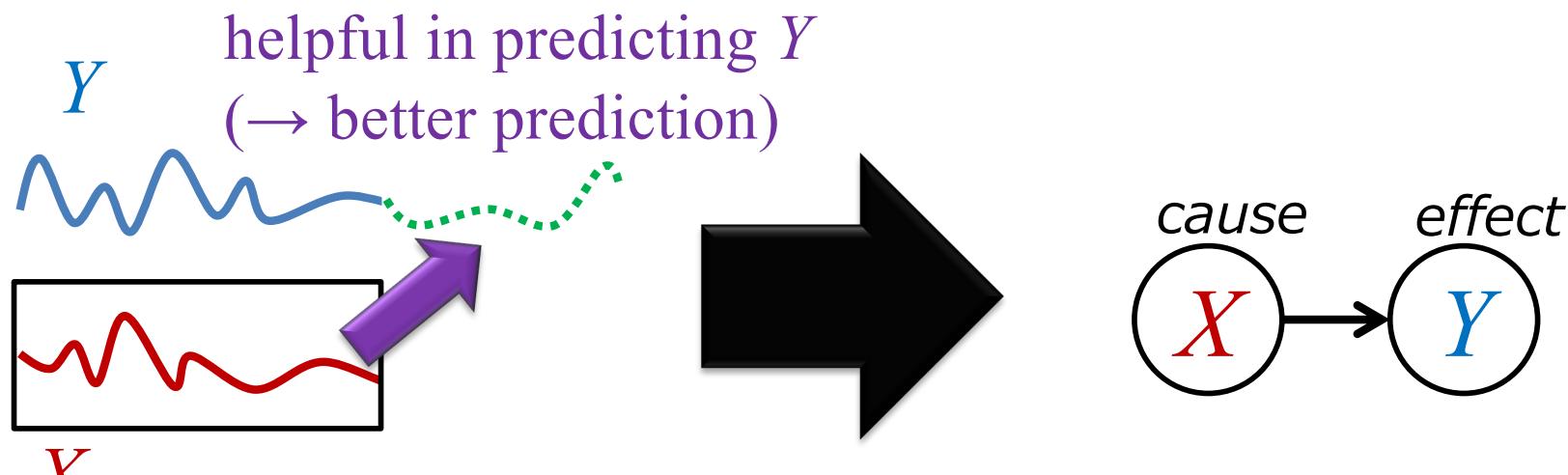


Granger causality [Granger1969]

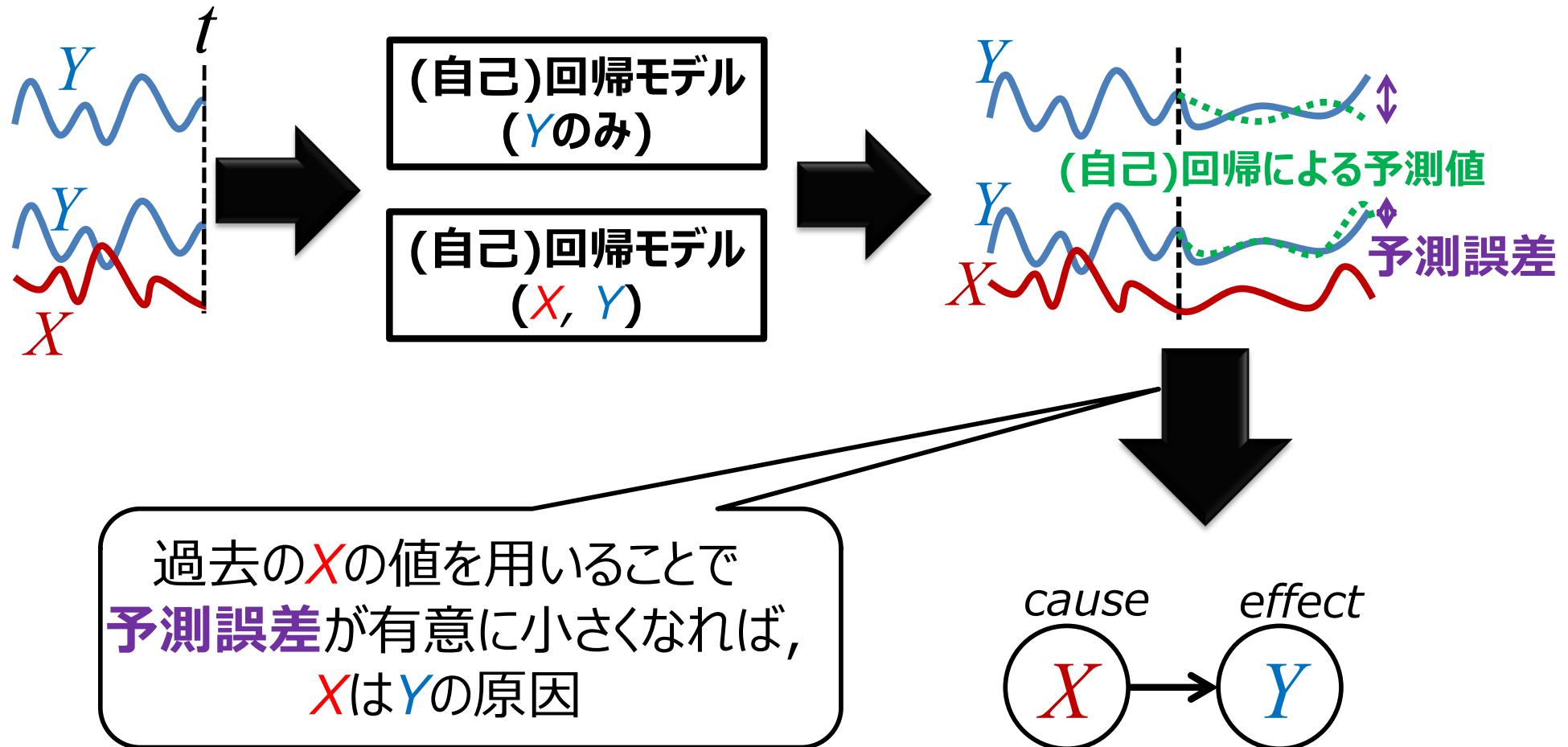
X is the cause of Y



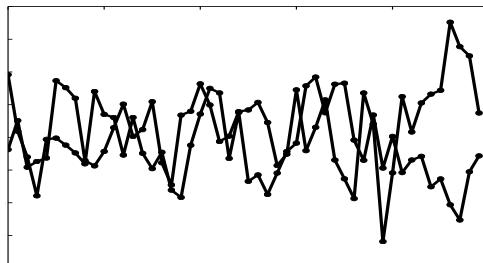
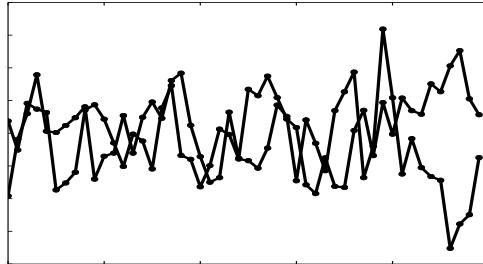
if the past values of X are **helpful in predicting**
the future values of Y



既存手法によるGranger causalityの推定



既存手法の問題点



⋮

(自己)回帰モデル



どの回帰モデルを
使えばいいの？

VAR
モデル

GAM

ガウス過程

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \frac{1}{P} \sum_{\tau=1}^P A_\tau \begin{bmatrix} X_{t-\tau} \\ Y_{t-\tau} \end{bmatrix} + \begin{bmatrix} E_{X_t} \\ E_{Y_t} \end{bmatrix}$$

(自己)回帰モデルを各データに対して適切に選ばなければ
正しくGranger causalityの方向・有無を推定できない



本研究のゴール：回帰モデルの選択が不要なアプローチの提案

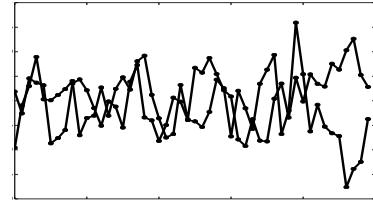
提案手法：

教師あり学習に基づくGranger causalityの推定

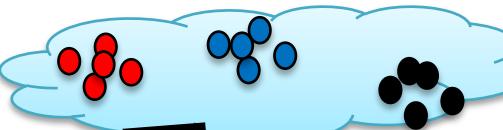


テストデータ

(因果関係を知りたいデータ)



特徴ベクトル



分類器

$X \rightarrow Y$

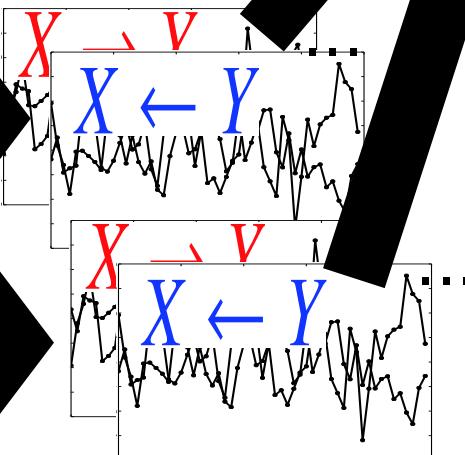
$X \leftarrow Y$

因果なし

サンプリング

線形
モデル
(VAR)

非線形
モデル



訓練データ

(因果関係が既知のデータ)

さまざまな(人工)データ
を活用

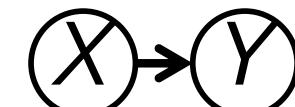
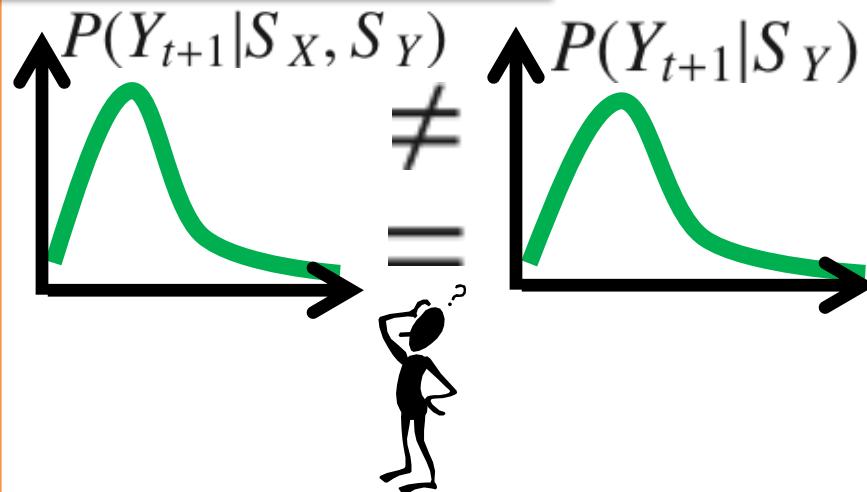
さまざまな
確率モデル

How can we obtain feature vectors?

The answer lies in Granger causality definition

Granger因果性の有無・方向は
2つの(条件付き)分布が等しいか否かで決まる

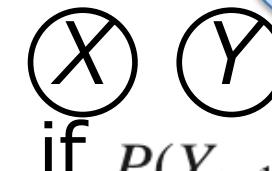
Granger因果性の
フォーマルな定義



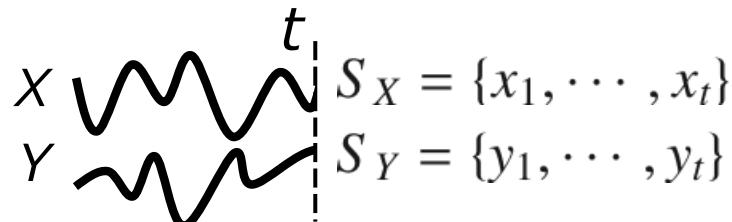
if $P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$

X を用いると
 Y の予測が異なる

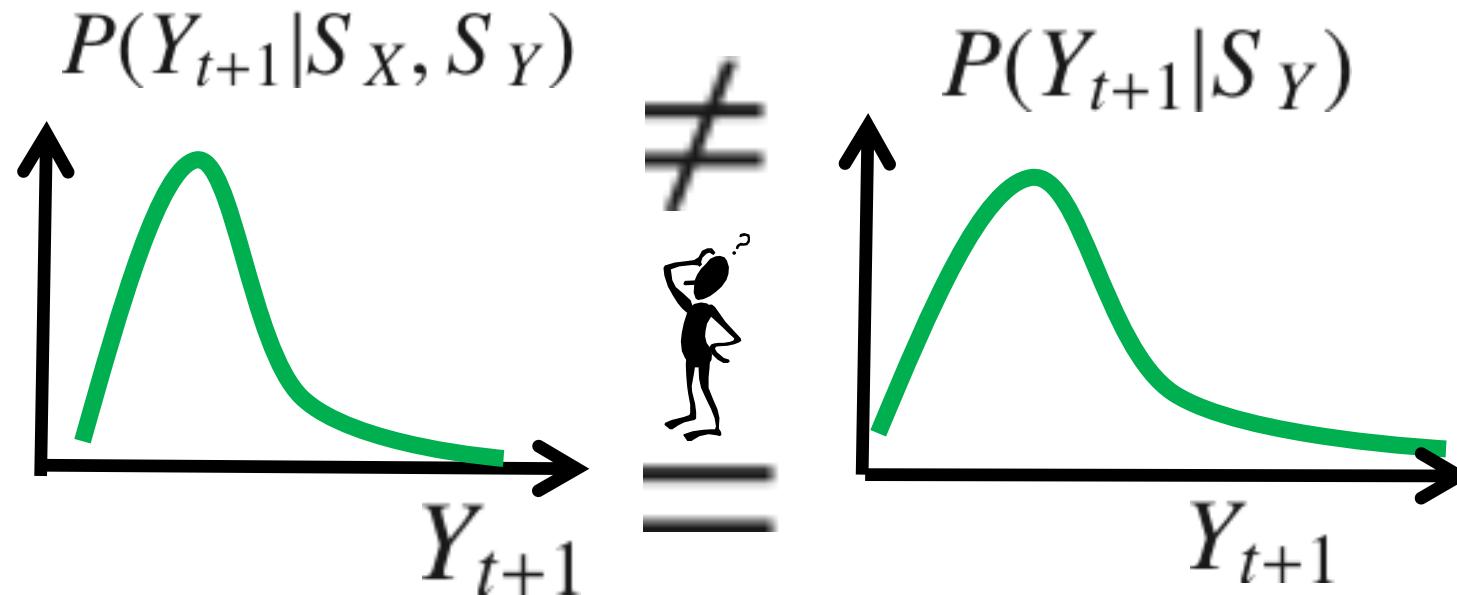
X あってもなくても
 Y の予測は変わらない



if $P(Y_{t+1}|S_X, S_Y) = P(Y_{t+1}|S_Y)$



Whether or not two distributions are equal is important



How can we determine whether or not
two distributions are equal?

We used MMD (distance between kernel means)

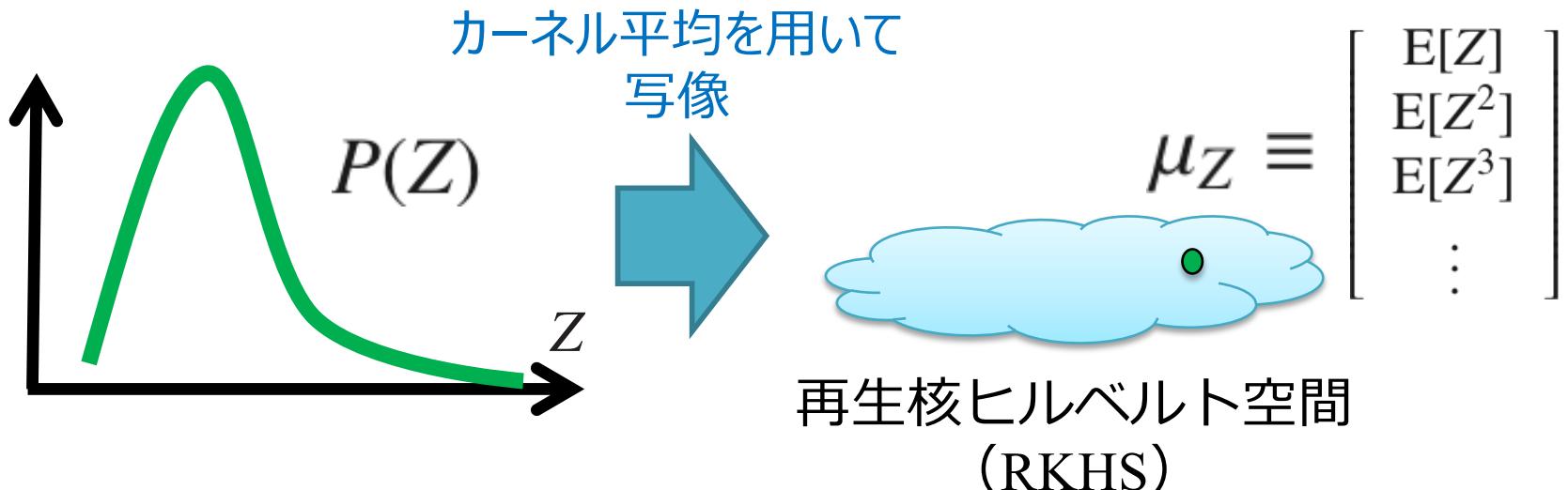
Kernel meanを用いて確率分布の特徴を抽出



Innovative R&D by NTT

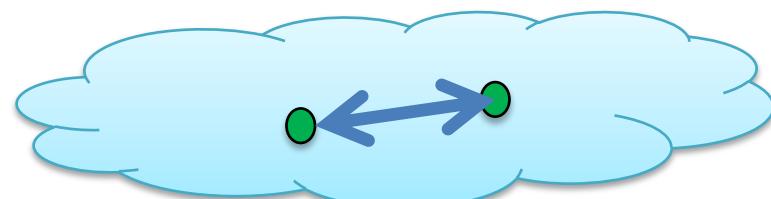
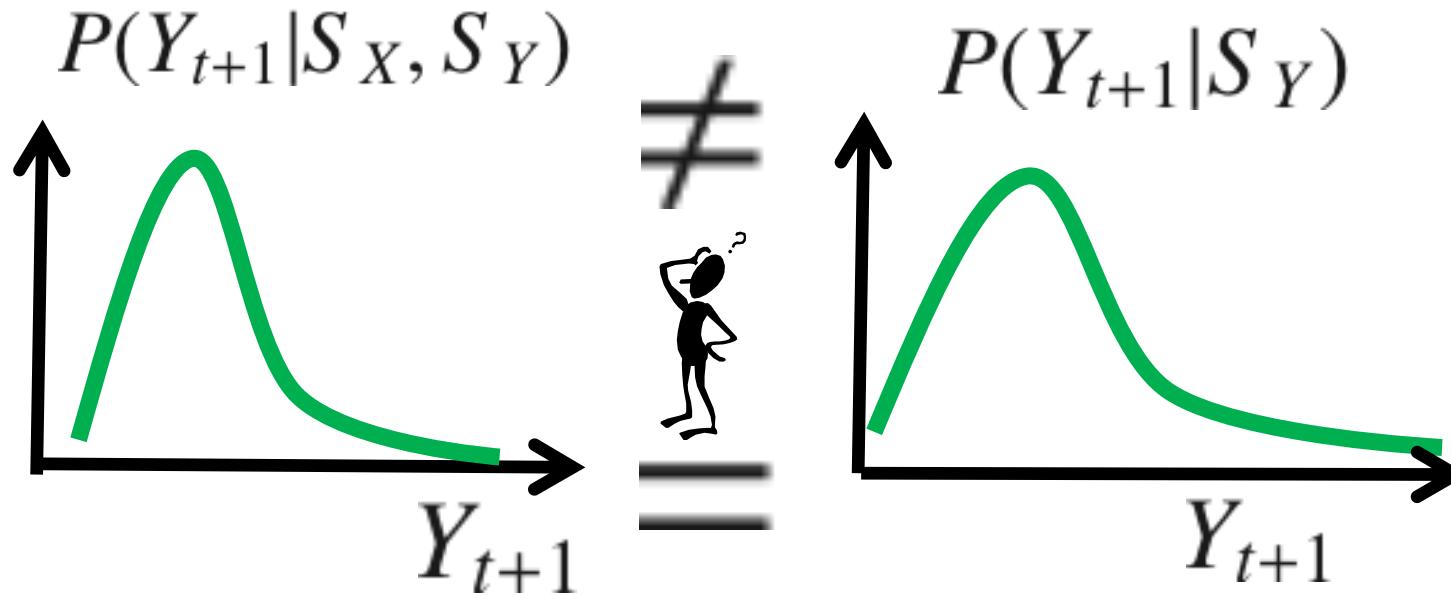
MMD: 条件付き確率分布の違いをkernel mean間の距離として定量

- kernel meanとは：分布の平均，分散，…を返す関数



※特殊なカーネル(Gaussian kernel等)を用いれば
異なる2つの分布は同一の点に写像されない(写像による情報損失がない)

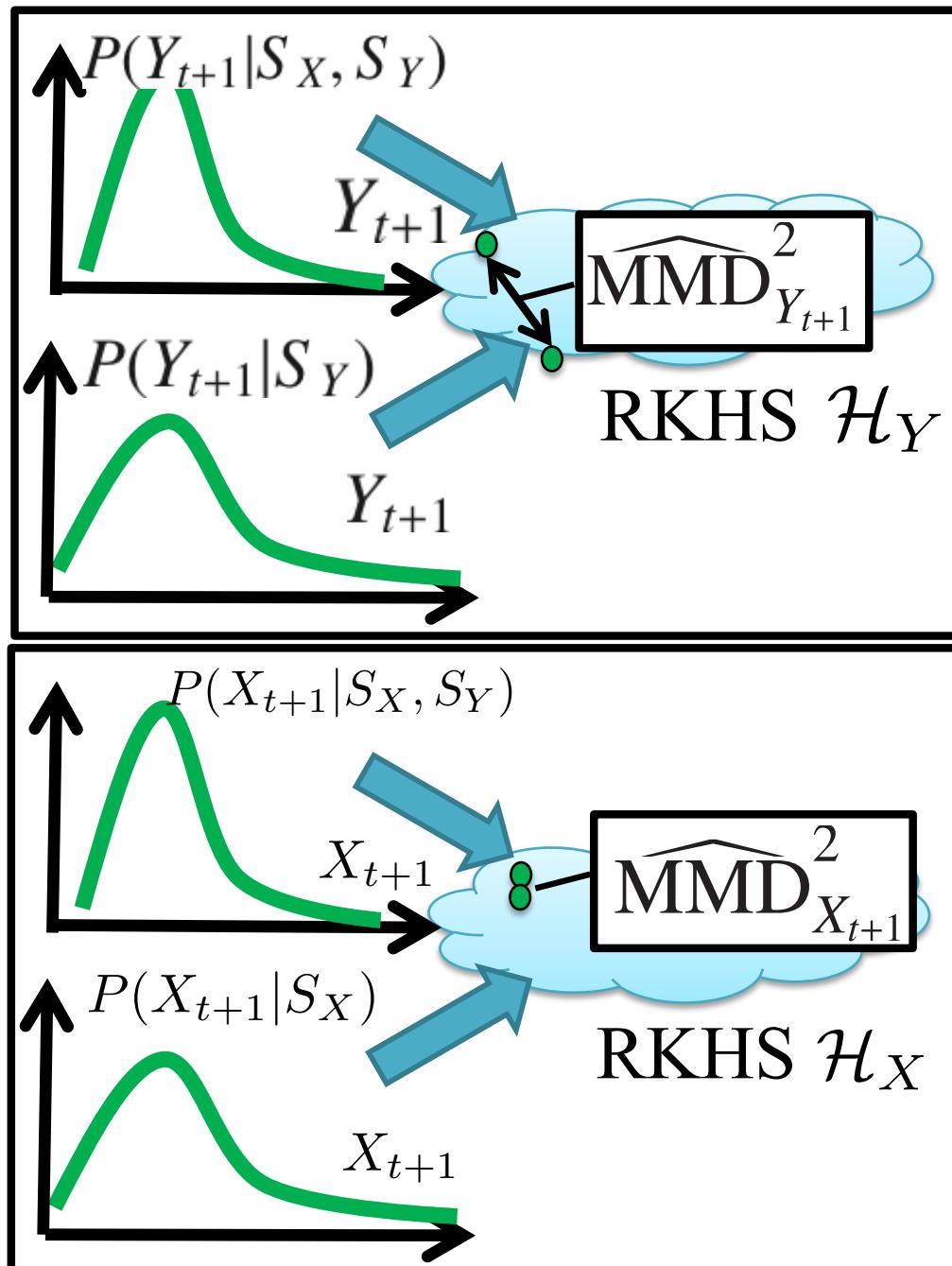
Whether or not distance between points is zero is important



Distance (MMD)
(maximum mean discrepancy
[Gretton+; NIPS2007])

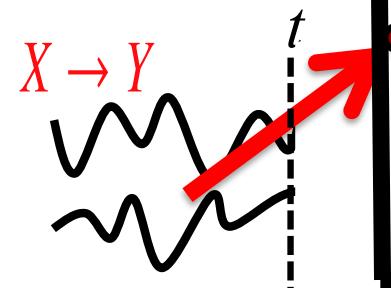
A blue downward-pointing arrow points from the text "Distance (MMD)" to a comparison diagram. The diagram shows two black stick figures with question marks, positioned above a horizontal double-equals sign ($=$). To the right of the first figure is a blue not-equals sign (\neq) followed by the number "0". To the right of the second figure is another blue not-equals sign (\neq) followed by the number "0".

Kernel meanを用いて確率分布の特徴を抽出

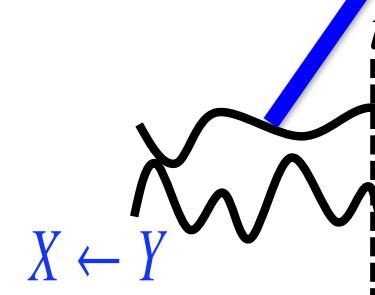


特徴ベクトル

$$\widehat{\text{MMD}}_{Y_{t+1}}^2$$



$$\widehat{\text{MMD}}_{X_{t+1}}^2$$

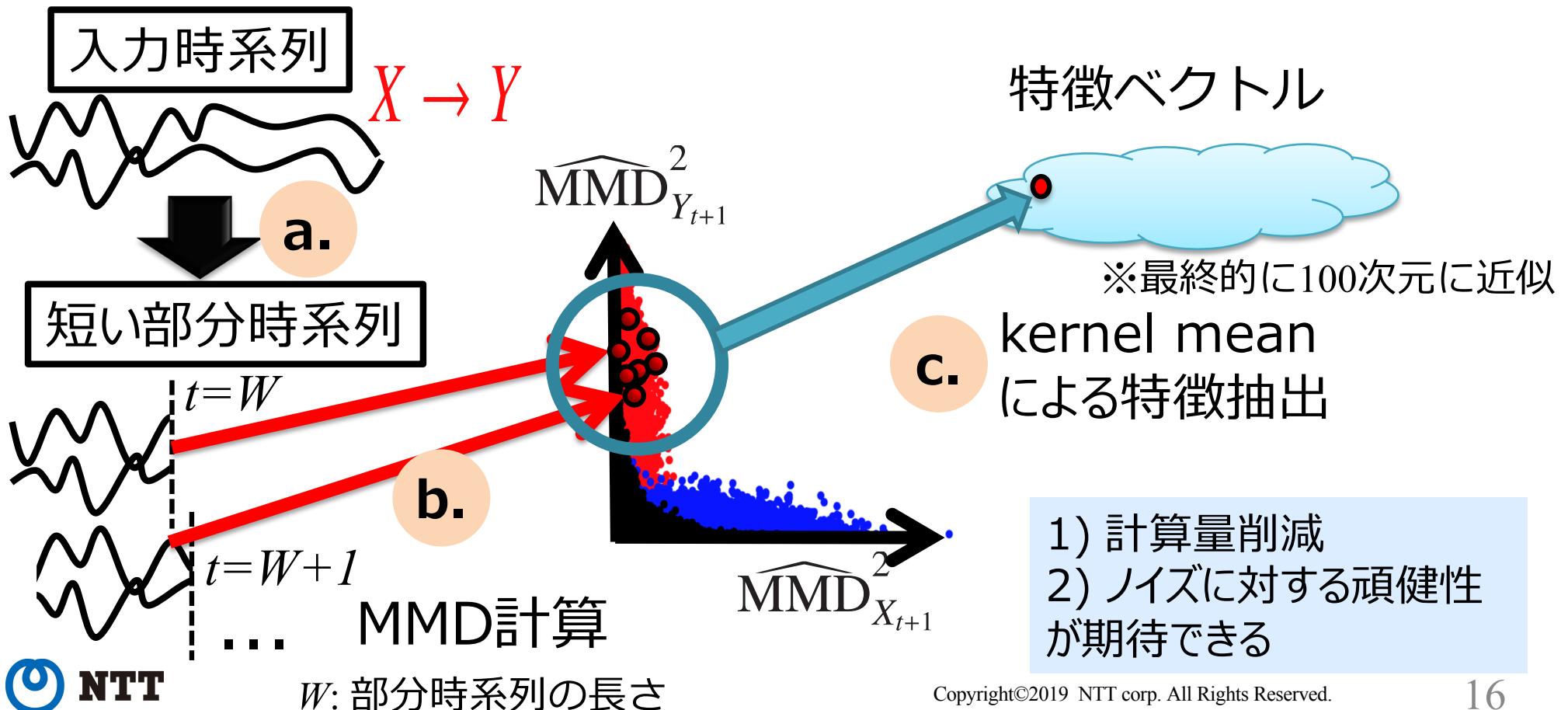


実際の特徴量の計算方法

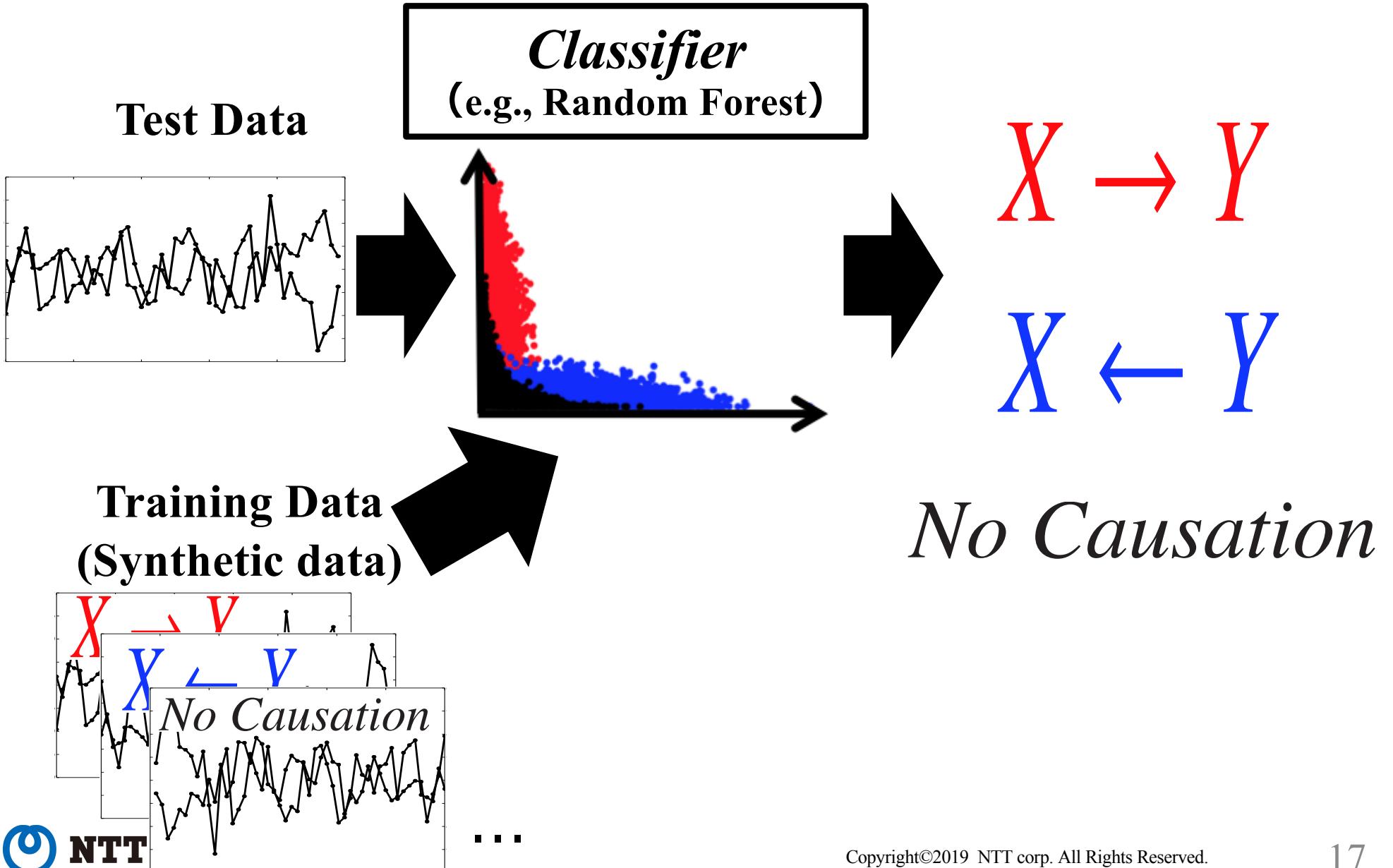


以下の手順で特徴量を計算

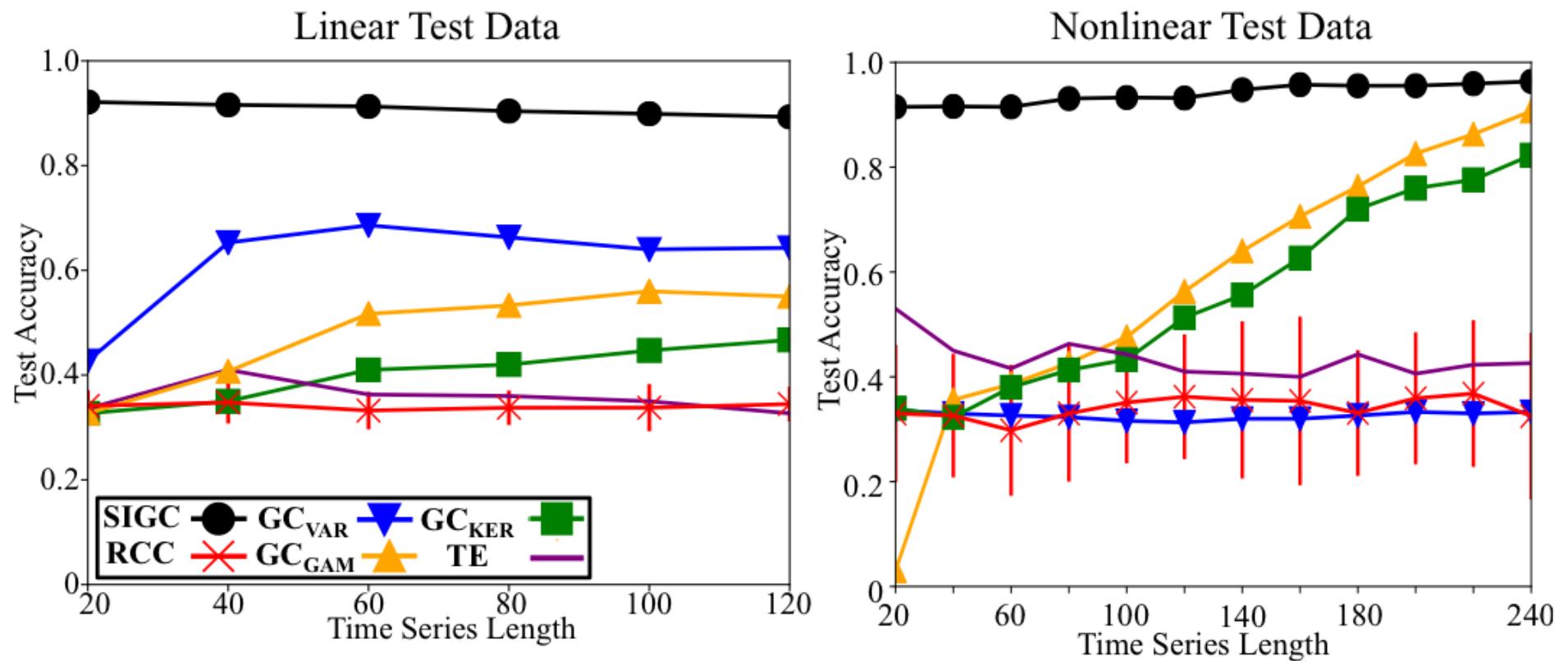
- 長い時系列を短い部分時系列に区切る
- 短い部分時系列に対しMMDをそれぞれ計算
- MMDの集合を(経験)分布とみなし, kernel meanを用いて分布の特徴を抽出し, 分類の特徴量とする



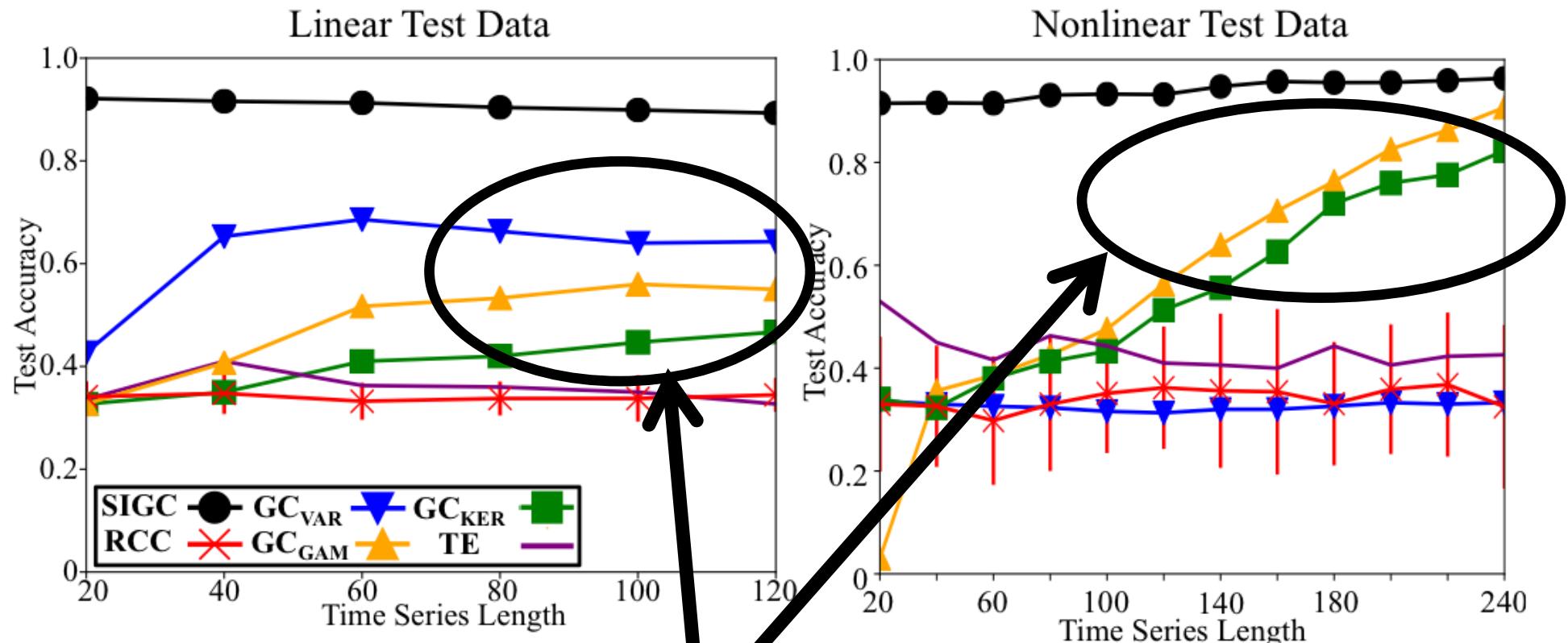
Experiments



Results on synthetic test data



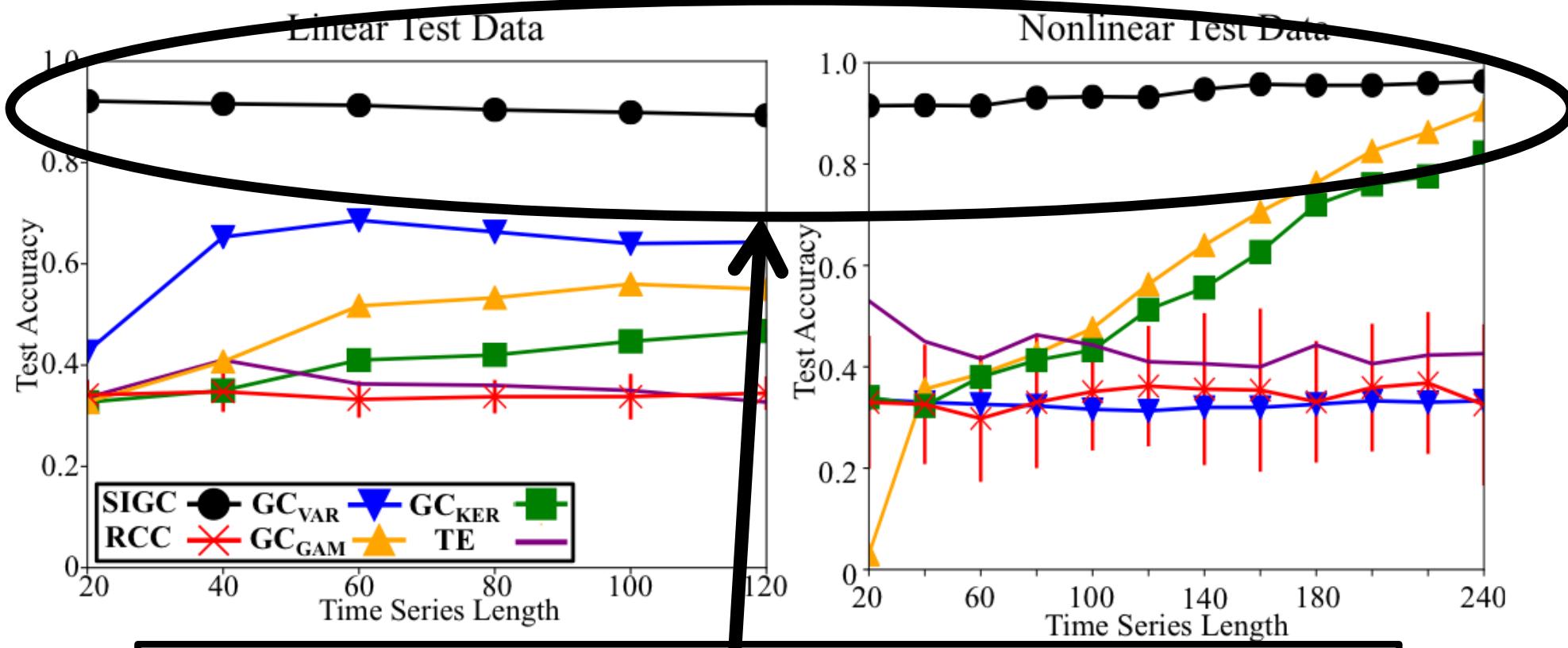
Results on synthetic test data



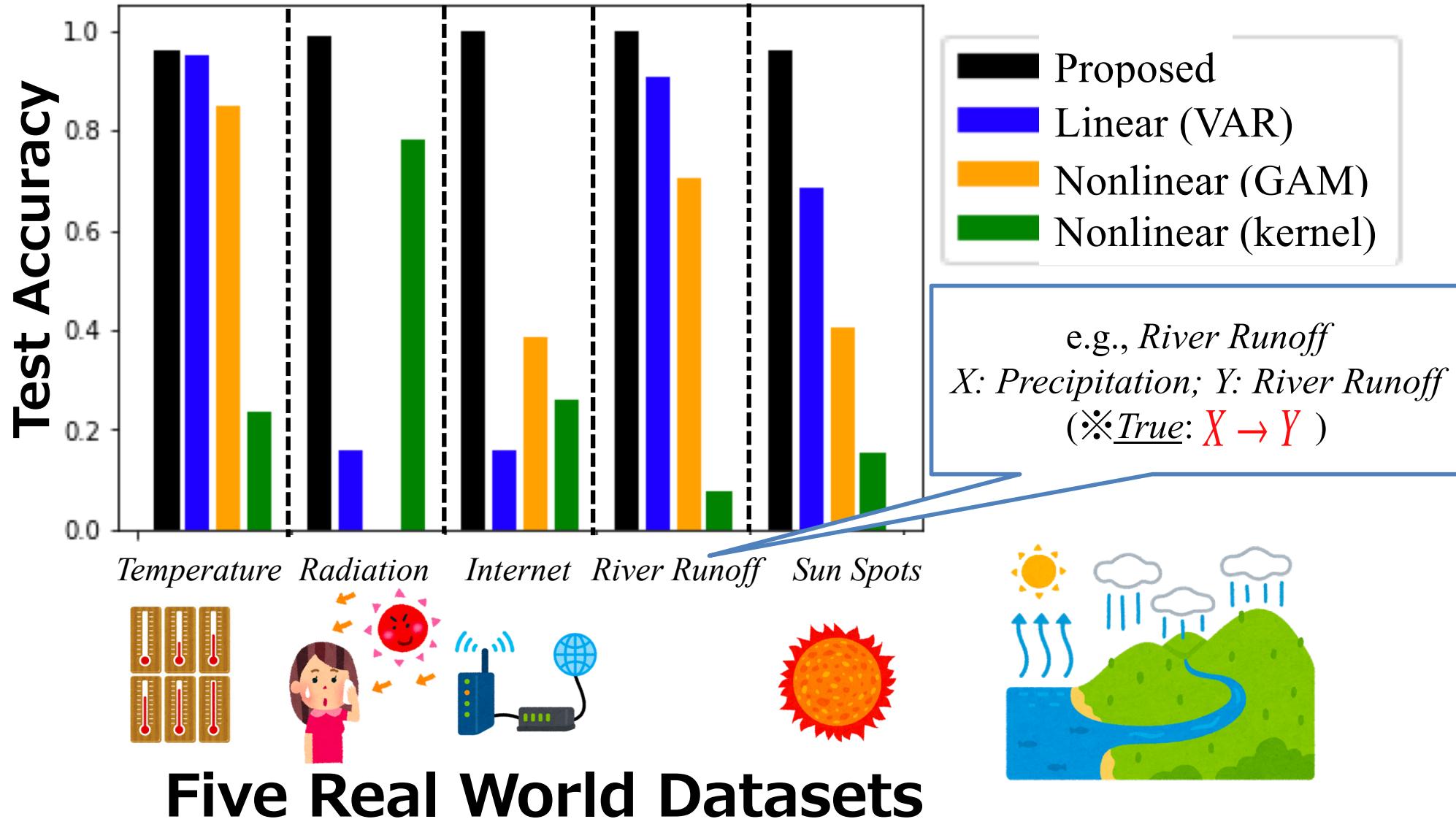
Existing Granger causality methods

Test accuracy strongly depended on the regression model

Results on synthetic test data



Results on real world data





- Granger causalityの有無・方向を推定する問題を教師あり学習の問題として解くアプローチを提案
 - ✓ 従来手法：回帰モデルがどれだけデータにフィットするかに強く依存
 - ✓ 提案手法：訓練データの情報を活用できる(data augmentation)





以下、FAQ



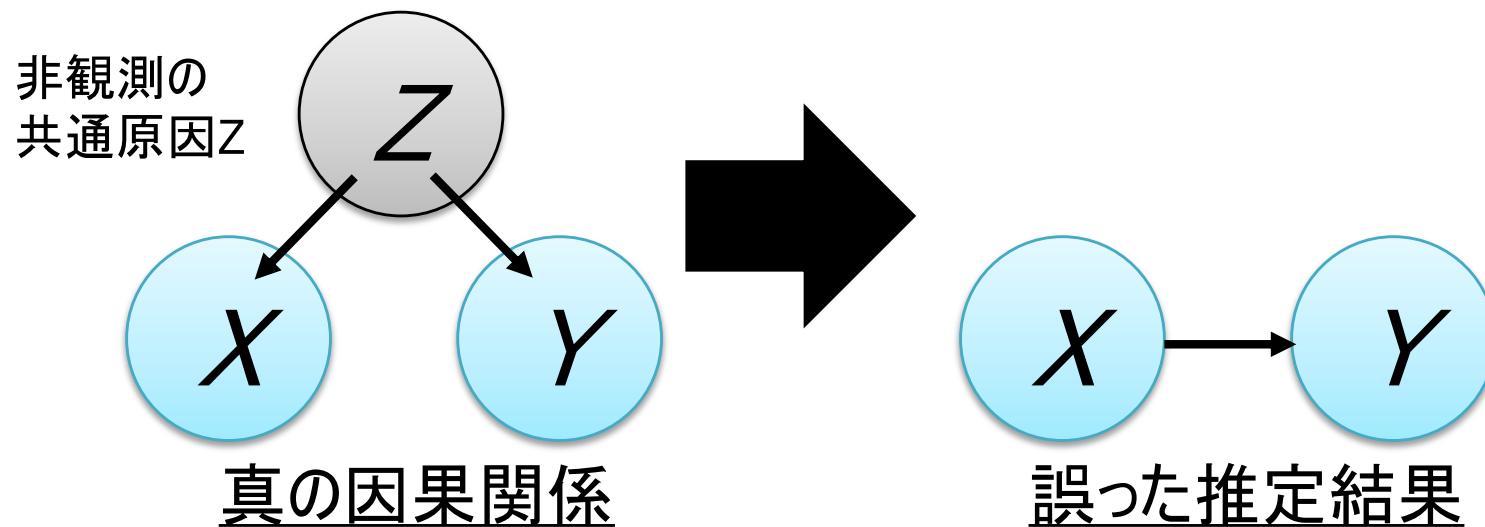
1. 予測が良いで因果と言つていいの？

- 非観測の変数がない（関連する全ての変数のデータが与えられている）
- 十分細かい時間スケールで時系列の値が観測されている

のであれば、Granger因果性は因果関係を表す（後者の仮定を緩和する因果の定義はある）

2. 非観測変数があるときは難しい？

- すごく難しい。因果関係がない変数のペアに対しても因果関係があると推定してしまう



- 因果関係の推定はまだまだ基礎研究なので、正確に推定するのは難しい



3. なぜ“MMD”を使うのか？

以下 2つより魅力的だと判断した

- KL divergence：
一般に密度関数（もしくは密度比）の推定が必要。一般に密度関数を推定するよりカーネル平均を推定するほうが簡単
（必要なデータ数が少なくてすむ）
- Kormogorov Smirnov統計量(2分布間に違いがあるかないかを検定する(KS検定)際の統計量)：
時系列の場合、ARモデルなど**何らかの回帰モデル**を用いる必要がある

人工データ実験の場合の正解ラベル



例 1 : 以下の線系モデル(AR)から生成されたデータ

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} 0.2 & \underline{1.0} \\ \underline{0.0} & 0.2 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} E_{X_t} \\ E_{Y_t} \end{bmatrix}$$

※ E_{xt}, E_{yt} ノイズ変数

→ 正解ラベル(真の因果関係)は $\mathbf{Y} \rightarrow \mathbf{X}$

例 2 : 以下の線系モデルの場合は

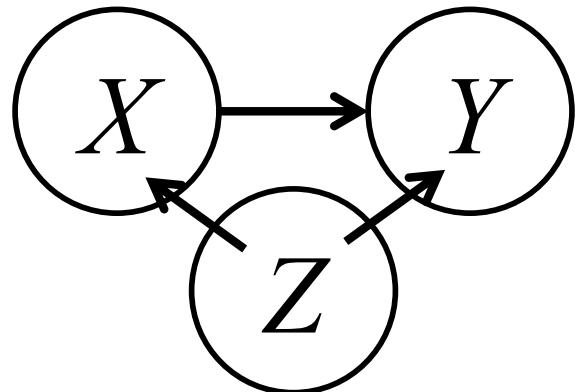
$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \begin{bmatrix} 0.2 & \underline{0.0} \\ \underline{0.0} & 0.2 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \begin{bmatrix} E_{X_t} \\ E_{Y_t} \end{bmatrix}$$

正解ラベル(真の因果関係)は 因果なし

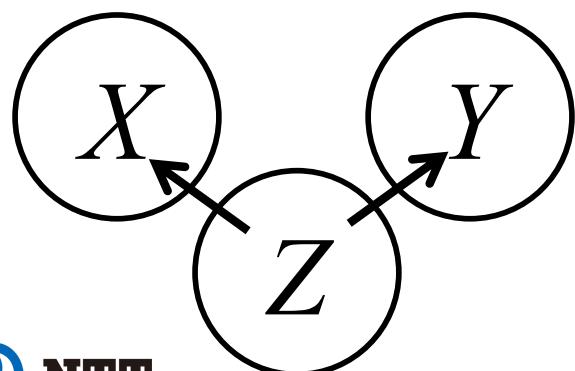
Granger causality definition for multivariate time series



- **Conditional Granger causality** [Geweke JASA1984]: compare two conditional distributions given past values of the third variable Z



if $P(Y_{t+1}|S_X, S_Y, \underline{S_Z}) \neq P(Y_{t+1}|S_Y, \underline{S_Z})$

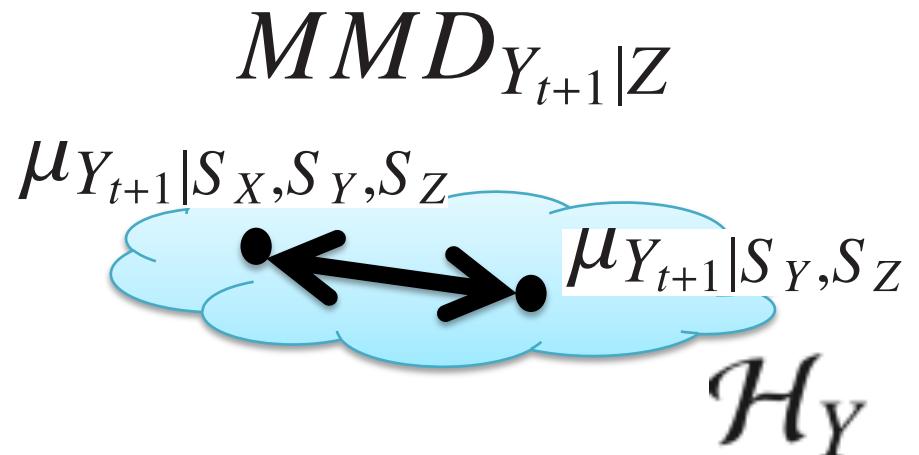
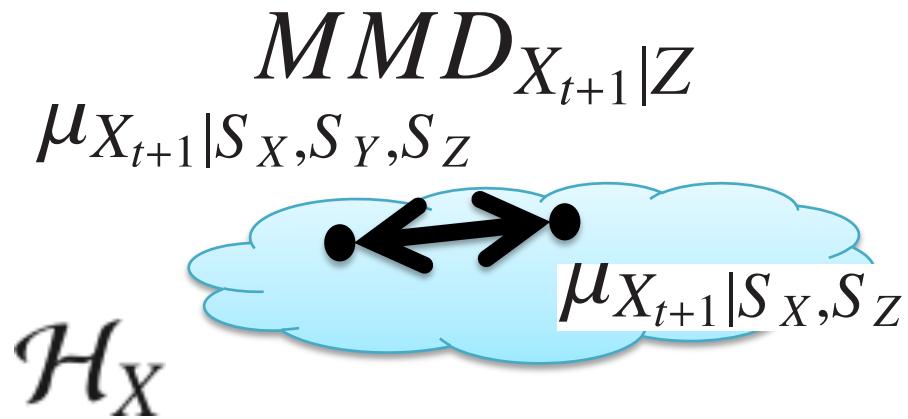


if $P(Y_{t+1}|S_X, S_Y, \underline{S_Z}) = P(Y_{t+1}|S_Y, \underline{S_Z})$

Feature representation



- Similarly, we map conditional distributions to points in feature spaces and measure the distance



- We use four MMDs to obtain feature vectors

$$\widehat{\text{MMD}}_{X_{t+1}}^2 \quad \widehat{\text{MMD}}_{Y_{t+1}}^2$$

$$\widehat{\text{MMD}}_{X_{t+1}|Z}$$

$$\widehat{\text{MMD}}_{Y_{t+1}|Z}$$

Assigning causal labels to n -variate time series



- Assumption: number of parent of each variable is **at most 1**

(Due to exponentially growing # of possible combinations of causal graphs)

- To each of ${}_nC_2$ pairs from n variables, we assign one of the three causal labels:

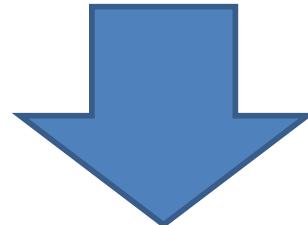
$X \rightarrow Y$ $X \leftarrow Y$ *No Causation*

Assigning causal labels to n -variate time series



For each triplet (X, Y, Z_v) ($v \in \{1, \dots, n-2\}$),

1. Prepare feature vector
2. Compute probability of three labels
by using trained classifier

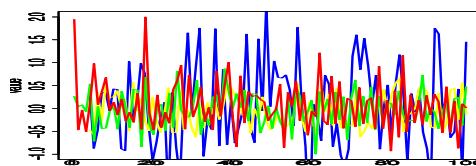


Then, we assign the label
with highest average probability

Experiment 3: Multivariate real-world data

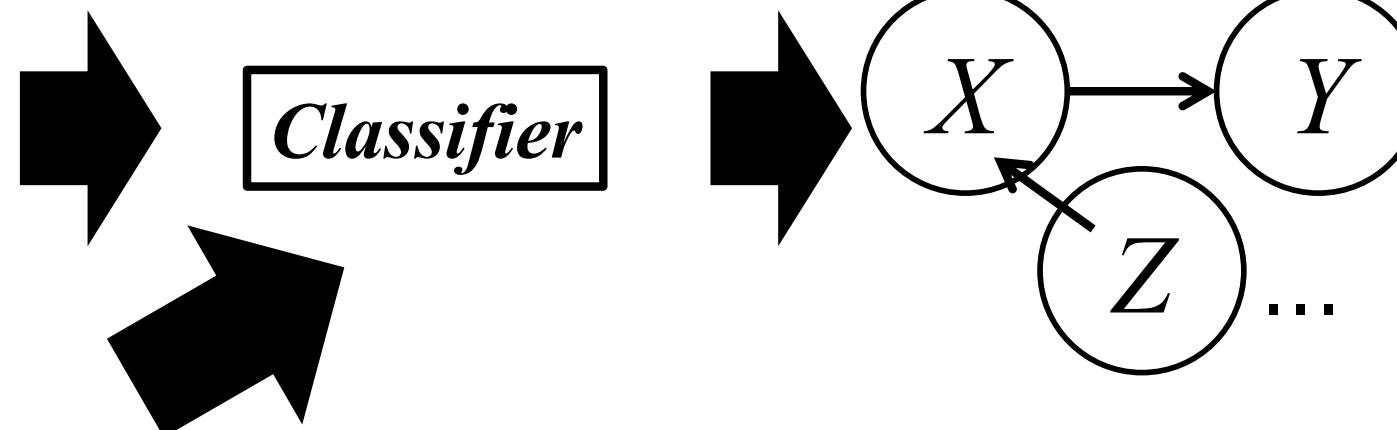


Real-world
Test Data



Yeast cell cycle gene expression data
[Spellman+ 1998]
14 variables (genes)

Synthetic
Training Data



True causal directions are given in database

Macro F1 score and micro F1 score



	SIGC _{tri}	GC _{VAR}	GC _{KER}	SIGC _{bi}	GC _{GAM}	TE	RCC
macro-averaged F1	0.483 (0.0)	0.351	0.437	0.431 (0.007)	0.457	0.430	0.407 (0.096)
micro-averaged F1	0.637 (0.0)	0.436	0.513	0.578 (0.011)	0.567	0.449	0.567 (0.161)

※Higher is better

