Causal Inference in Time Series via Supervised Learning (IJCAI2018, to appear)

<u>Yoichi Chikahara</u>, Akinori Fujino
NTT Communication Science Laboratories
Kyoto, Japan

A bit about myself

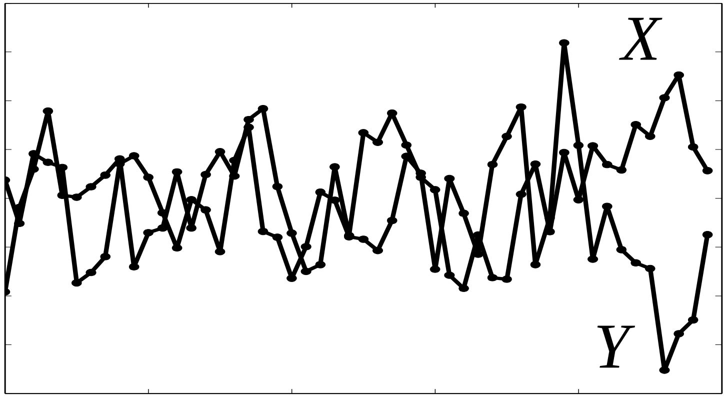
- Name: Yoichi Chikahara (近原 鷹一)
- Contact: chikahara.yoichi@lab.ntt.co.jp
- Education:
  - ➢ 2013.03: B. Sc. from **Keio University**
  - ➢ 2015.03: M. Info. Sci. & Tech. from **University of Tokyo**
    DNA information analysis lab. (Miyano lab.) in Dept. of Computer Science
  - ➢ 2015.04 – Now: Researcher @ NTT Communication Science Laboratories

- Research:
  - ➢ Machine Learning, Bioinformatics / Systems Biology

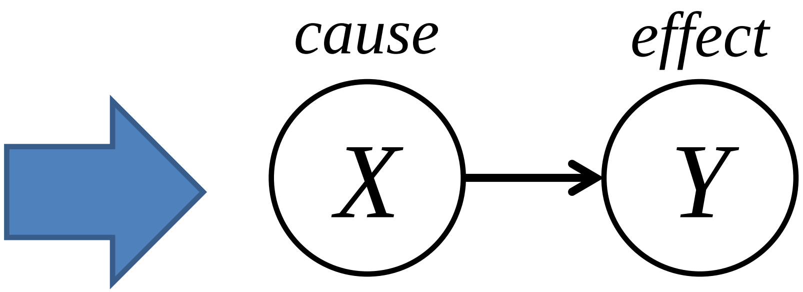
Causal Inference in Time Series

Causal inference in time series

- Given time series data
- Infer *causal relationships* between variables



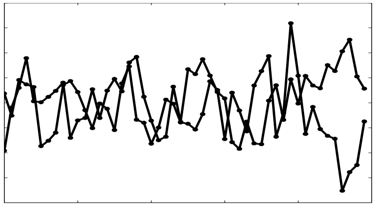
Input: Time Series Data

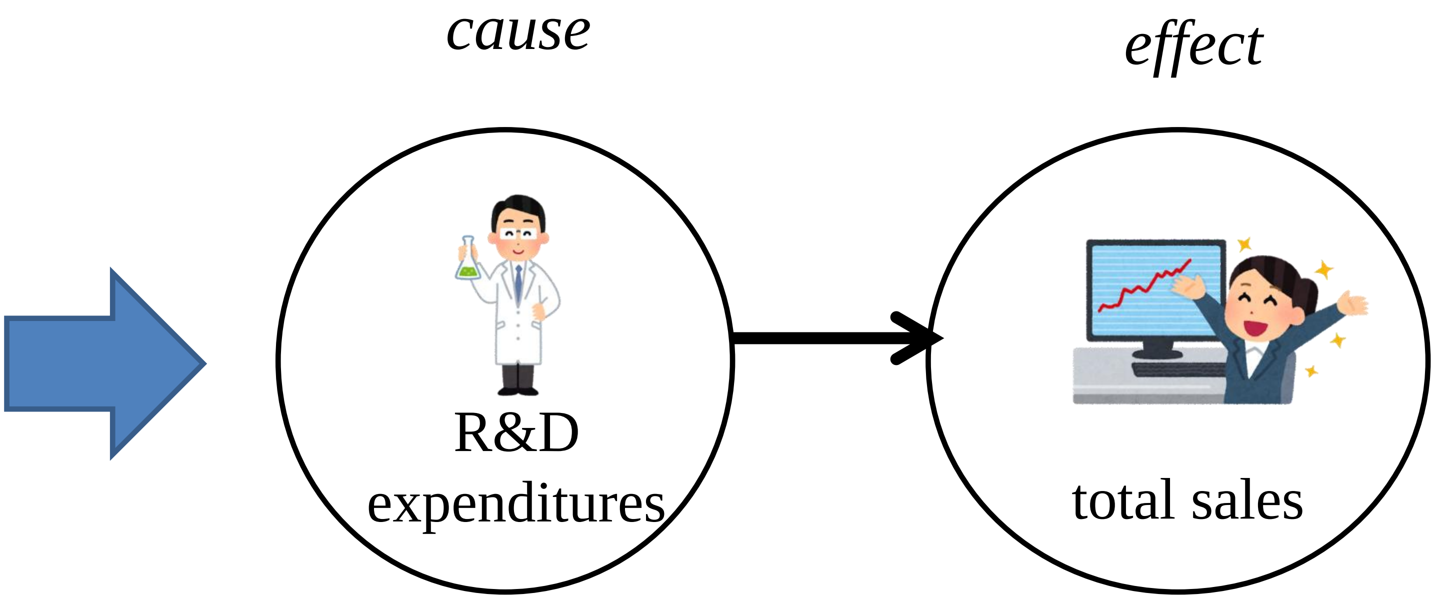


Output: *Causal Relationships*

Application 1: Economics

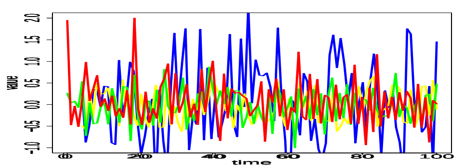
- Finding that R&D expenditures *influences* total sales is useful for companies

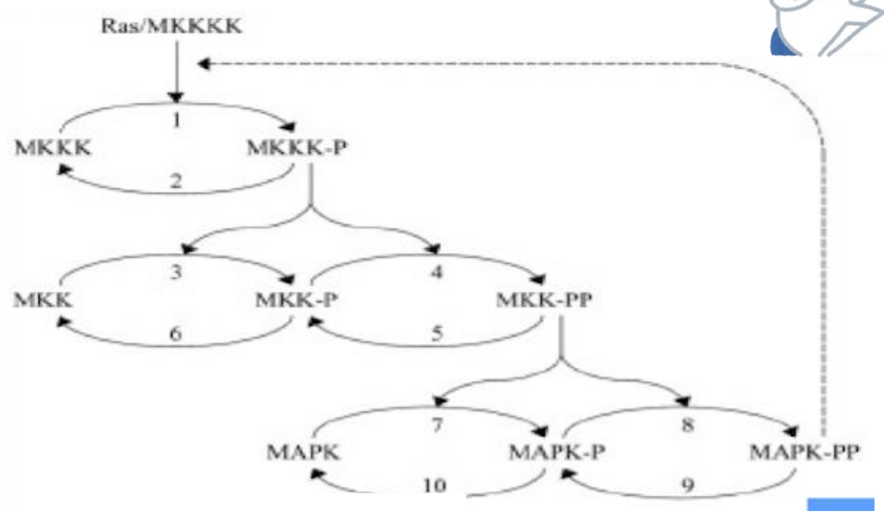




Application 2: Bioinformatics

- Discovering gene regulatory relationships is useful for drug discovery





What is “causal relationship”?

How can we define *causal relationships* between variables?

A definition of temporal causality

Granger causality [Granger1969]

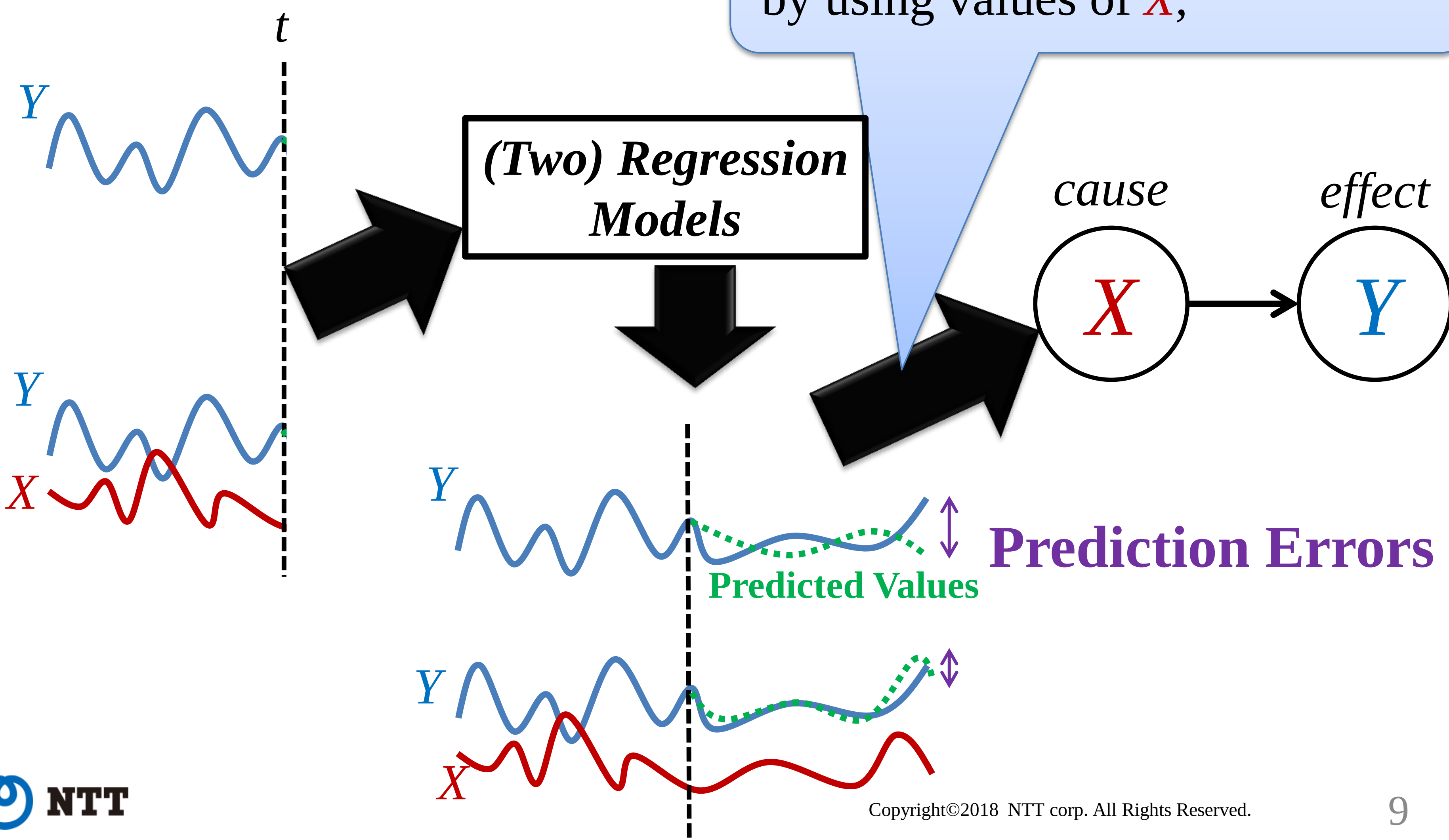
X is the cause of Y

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

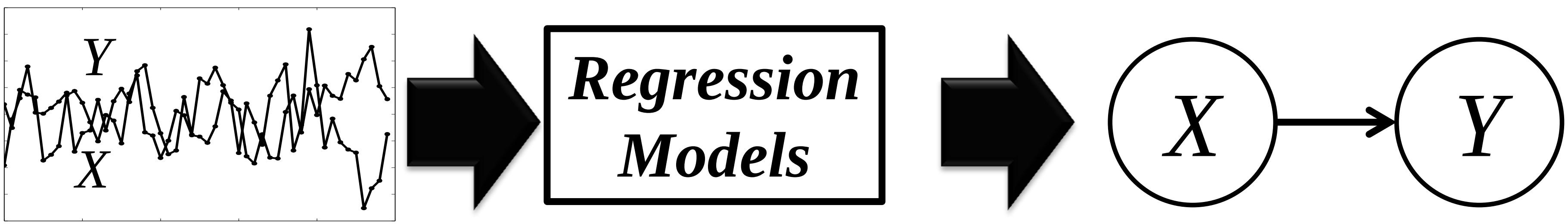
Clive W. J. Granger (1934-2009)

Existing approach:
Compare prediction errors with/without using values of *X*

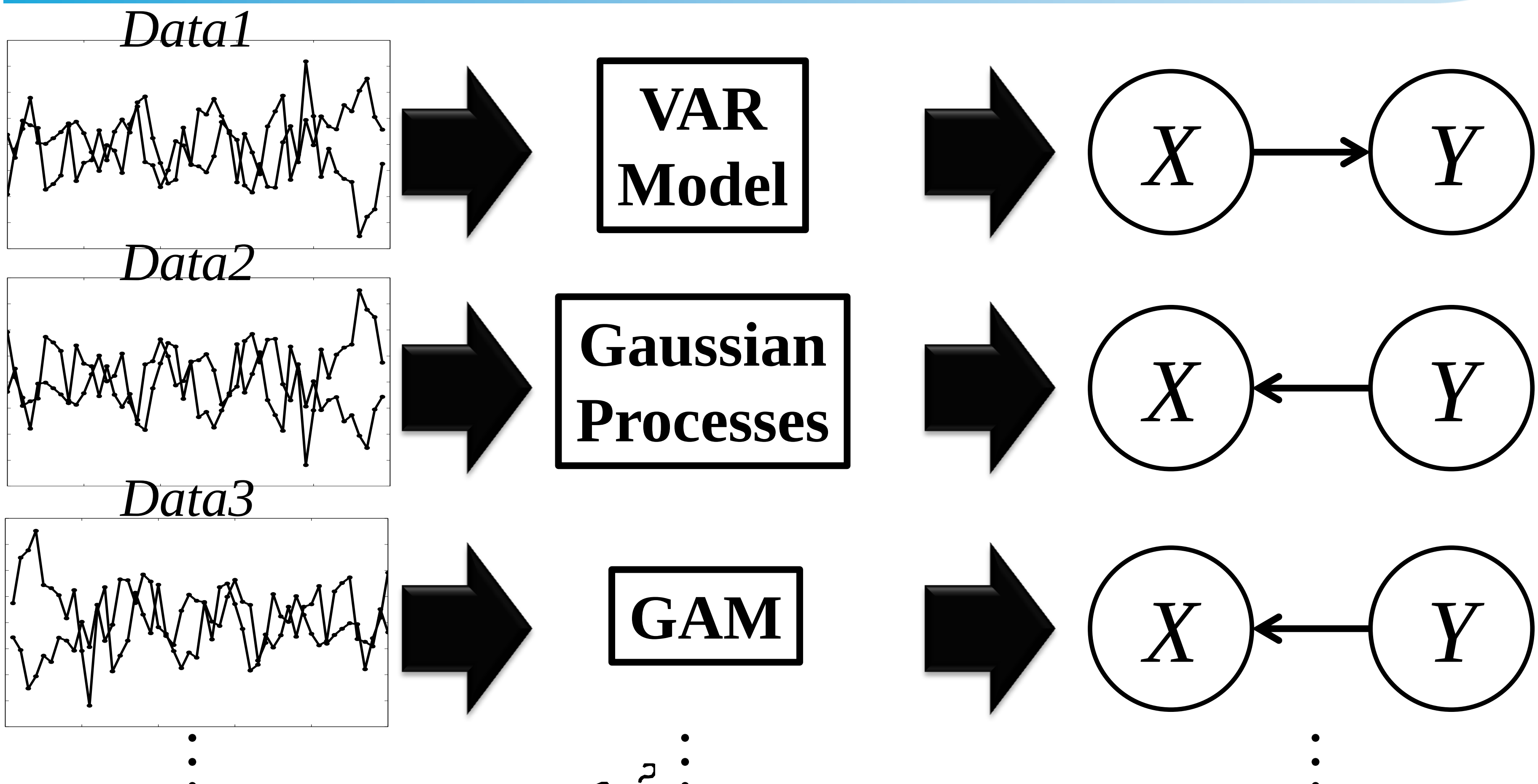
If errors are significantly reduced by using values of *X*,

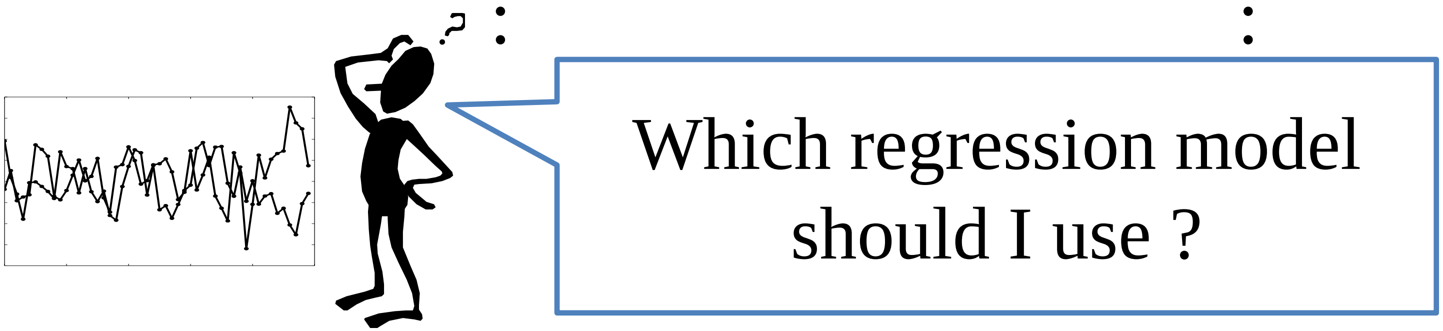


In Summary,



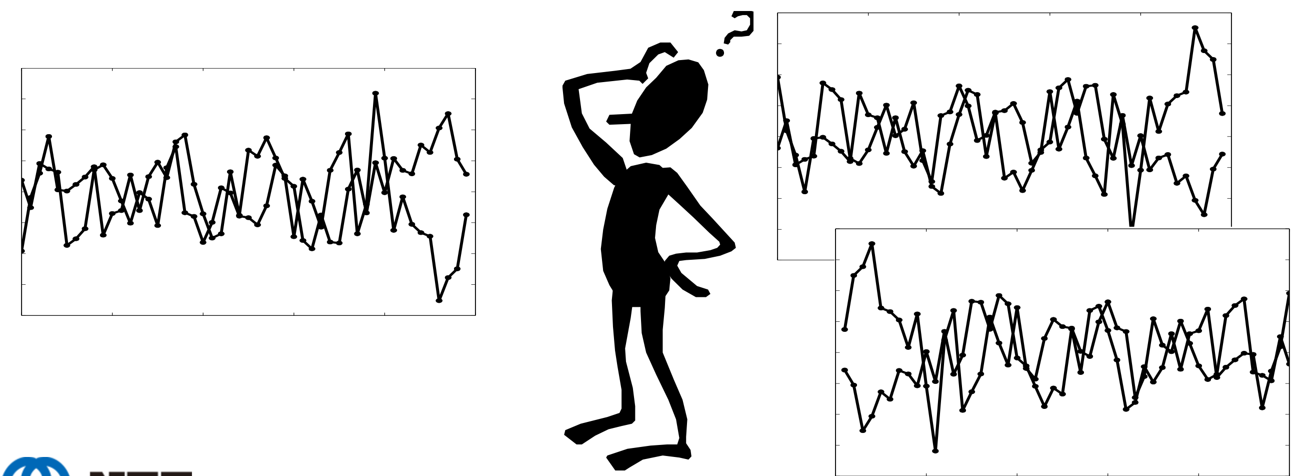
Weakness: Model selection problem



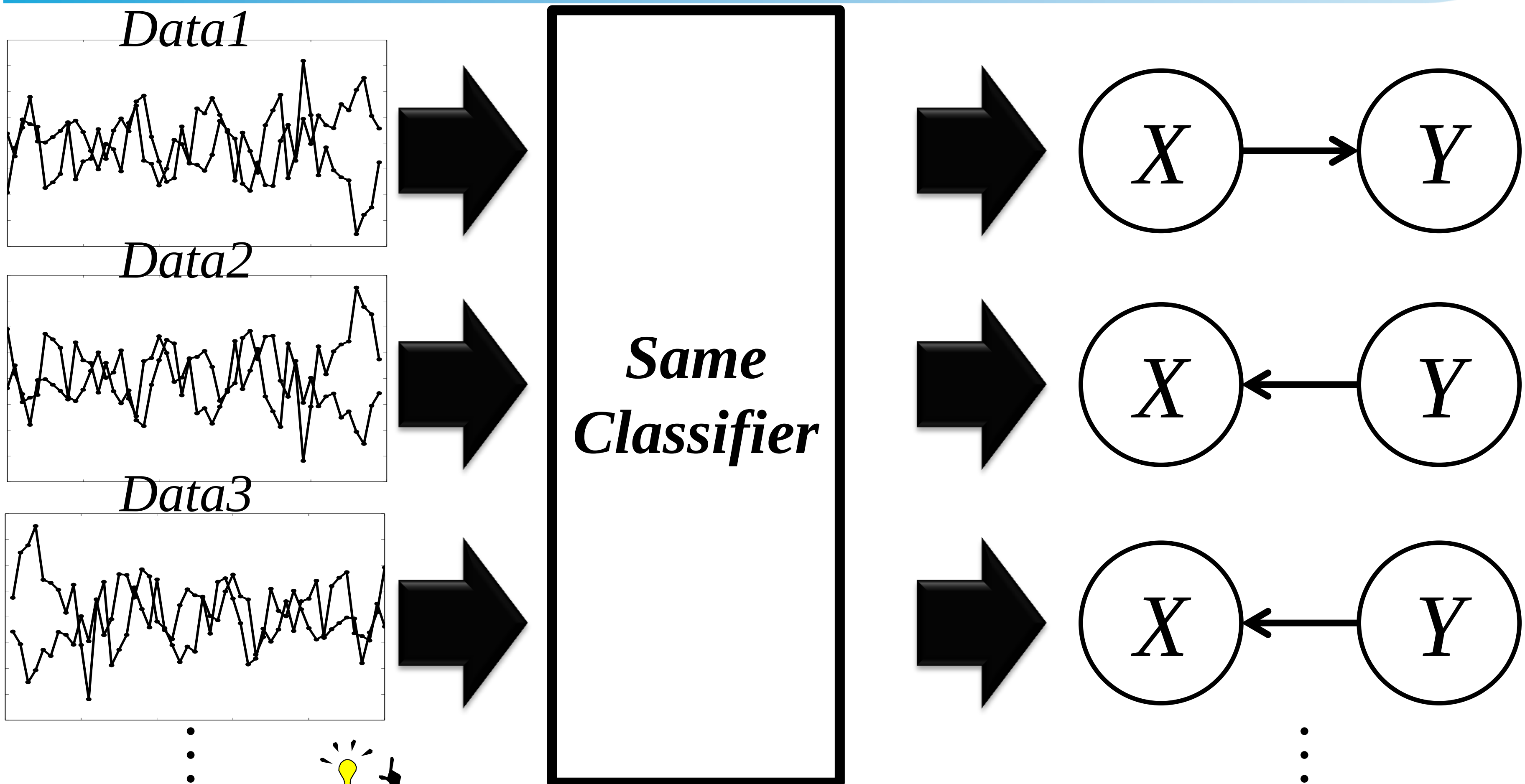


Weakness: Model selection problem

- **Problem**
  - ✓ Selecting appropriate regression models **is difficult** (needs a deep understanding of data analysis)
  - ✓ It is known that existing approach **does not work** when regression models cannot be well fitted to data



Our approach: Causal inference via classification

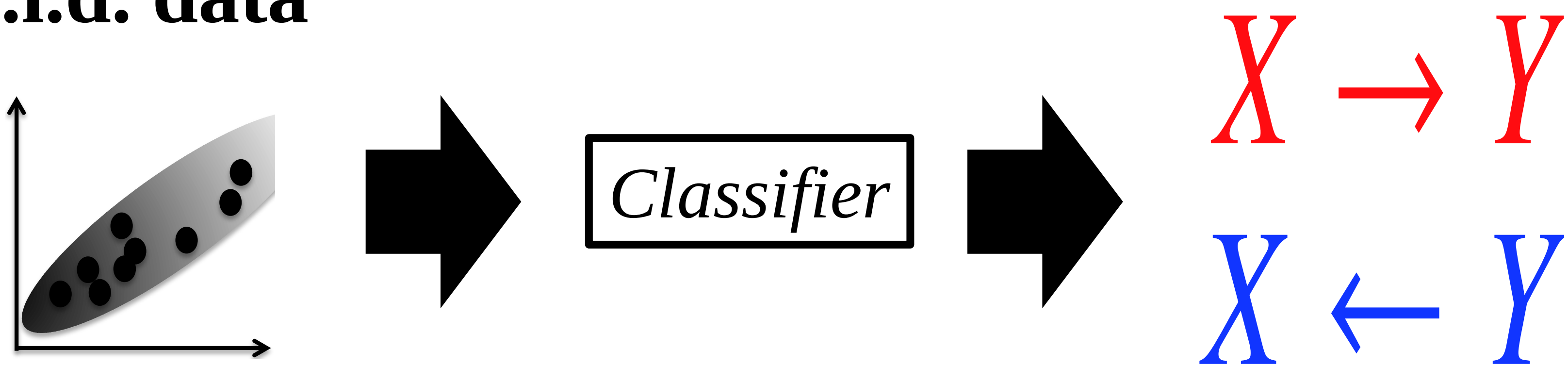


No need to select regression models!

Related work [ICML15, JMLR15, CVPR17]: Causal inference from i.i.d. data via classification

- In fact, in case of i.i.d. data, there are several existing methods based on classification

i.i.d. data

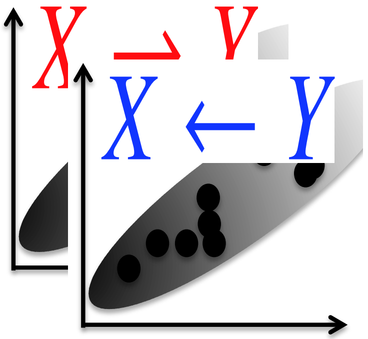


Related work [ICML15, JMLR15, CVPR17]:
1) Train a classifier

Classifier

Training data

(Data where causal relationships are <u>known</u>)



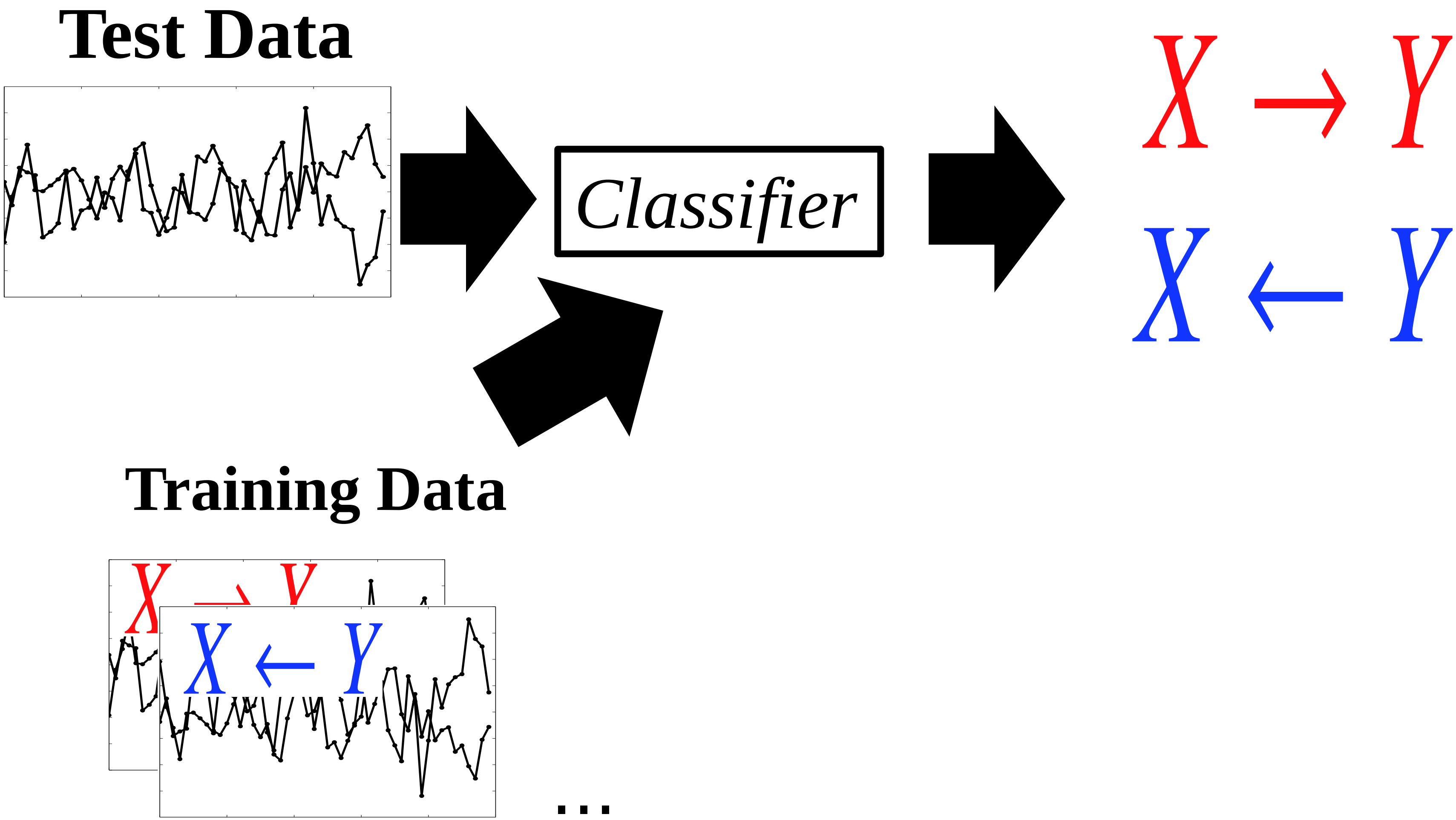
…

Related work [ICML15, JMLR15, CVPR17]:
2) Infer causal relationship by using trained classifier

Test Data

(Data where causal relationships are unknown)

Our approach:
Causal inference from time series data via supervised learning



Classification approach seems good,

but how can we solve
Granger causality identification problem
via classification?

Classification appr ... good,

bu ... solve

Grange ... ntification problem

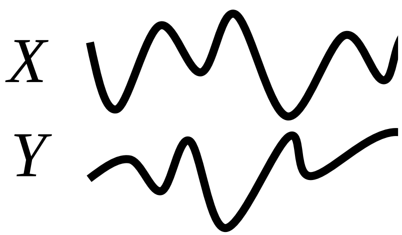
... ssification?

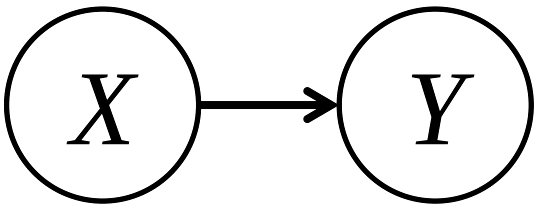
Key ideas lie in definition of Granger causality!

Revisiting assumption of Granger causality: Causal direction **never** changes over time

- Granger causality assumes that

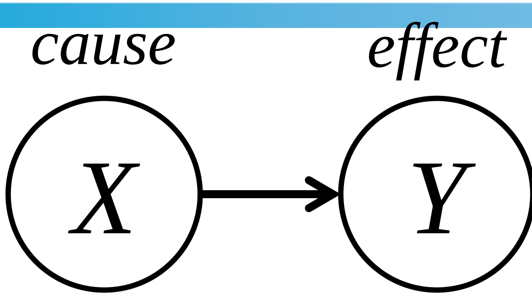
At **any** time point t, the causal direction is the same





(Our method also uses the assumption)

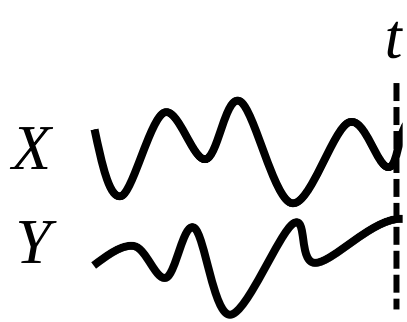
Revisiting definition of Granger causality



if the following holds:

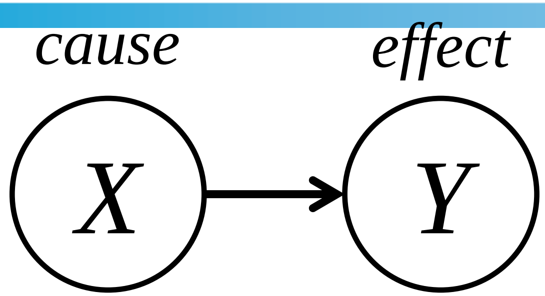
$$P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$$

at any time point t



$$S_X = \{x_1, \cdots, x_t\}$$
$$S_Y = \{y_1, \cdots, y_t\}$$

Revisiting definition of Granger causality

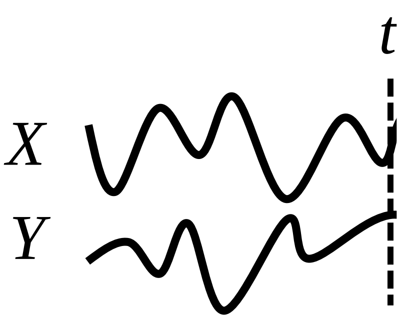


S_X is useful in prediction!

if the following holds:

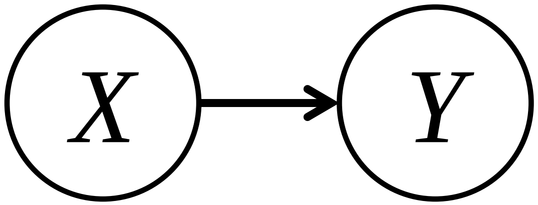
$$P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$$

Distribution of Y_{t+1} given past values of Y <u>and X</u> $\neq$ Distribution of Y_{t+1} given past values of Y

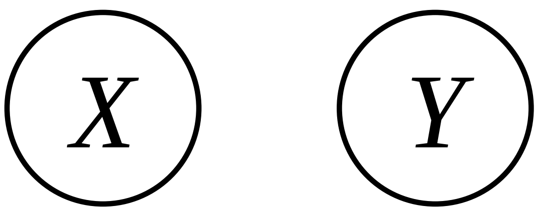


$$S_X = \{x_1, \cdots, x_t\}$$
$$S_Y = \{y_1, \cdots, y_t\}$$

Revisiting definition of Granger causality

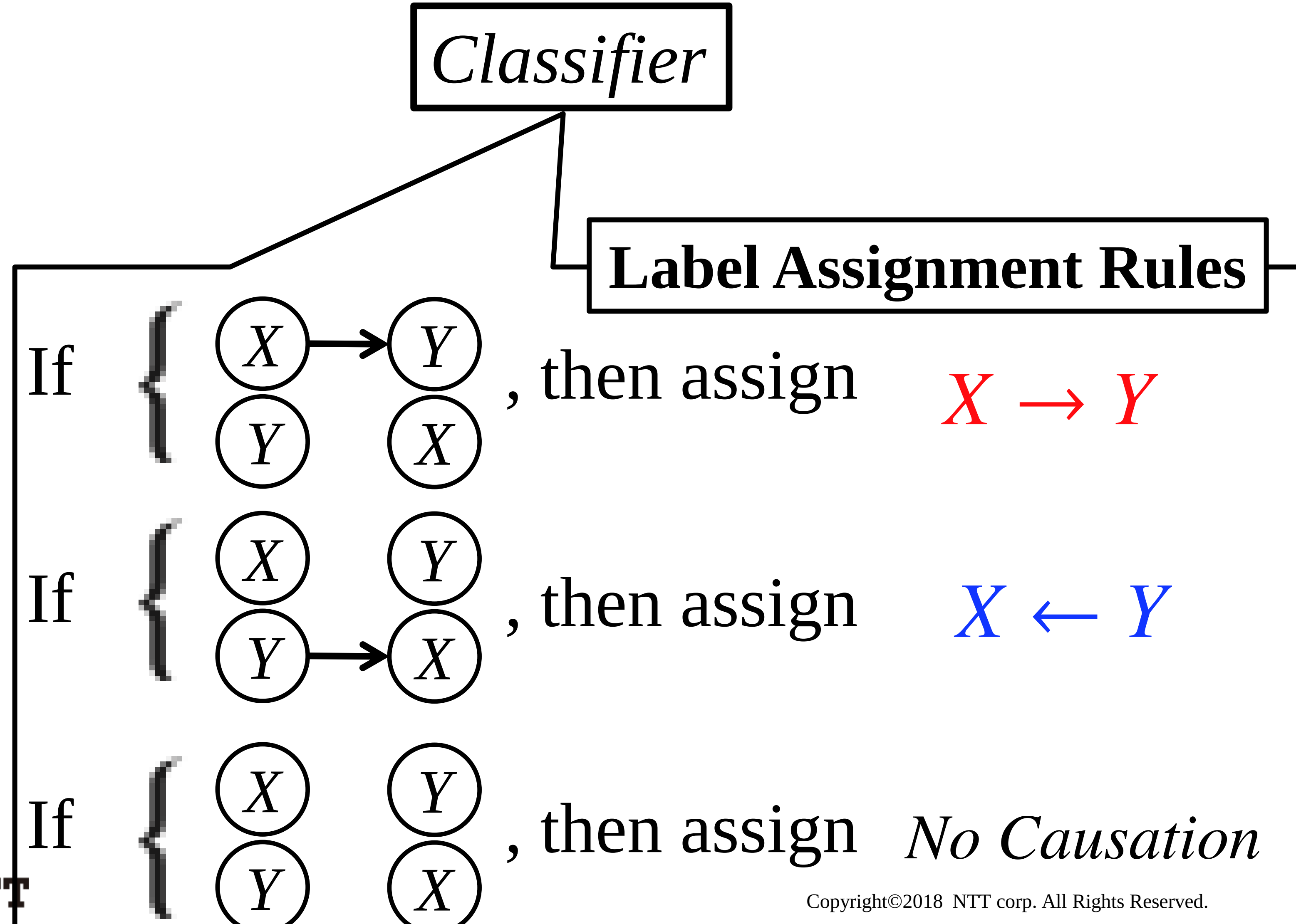


if $$P(Y_{t+1} | S_X, S_Y) \neq P(Y_{t+1} | S_Y)$$

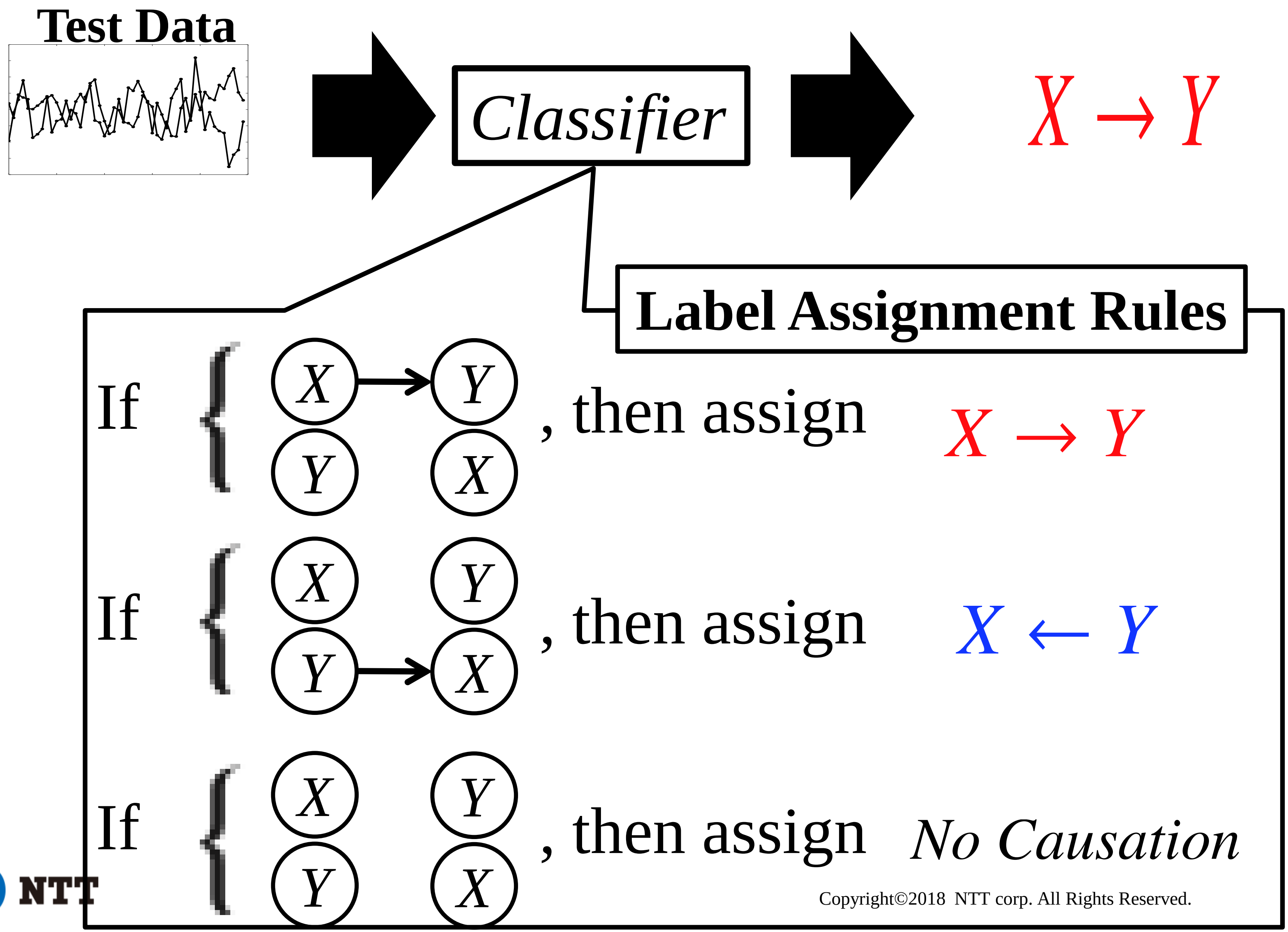


if $$P(Y_{t+1} | S_X, S_Y) = P(Y_{t+1} | S_Y)$$

Building a classifier for Granger causality identification



Building a classifier for Granger causality identification



Building a classifier for Granger causality identification

Label Assignment Rules

If
$$\begin{cases} P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y) \\ P(X_{t+1}|S_X, S_Y) = P(X_{t+1}|S_X) \end{cases}$$
then $X \rightarrow Y$

If
$$\begin{cases} P(Y_{t+1}|S_X, S_Y) = P(Y_{t+1}|S_Y) \\ P(X_{t+1}|S_X, S_Y) \neq P(X_{t+1}|S_X) \end{cases}$$
then $X \leftarrow Y$

If
$$\begin{cases} P(Y_{t+1}|S_X, S_Y) = P(Y_{t+1}|S_Y) \\ P(X_{t+1}|S_X, S_Y) = P(X_{t+1}|S_X) \end{cases}$$
then *No Causation*

Key information lies in distributions

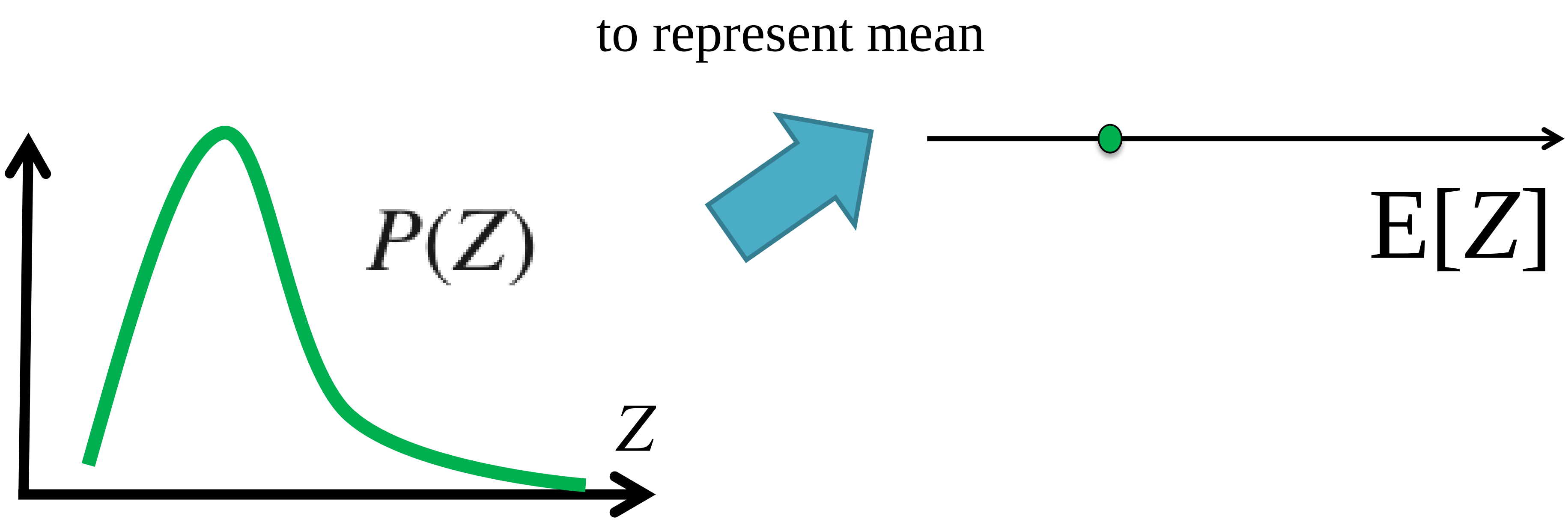
-> To determine whether or not the two distributions are identical, how do we obtain feature vectors for classification?

Key information lies i … ns

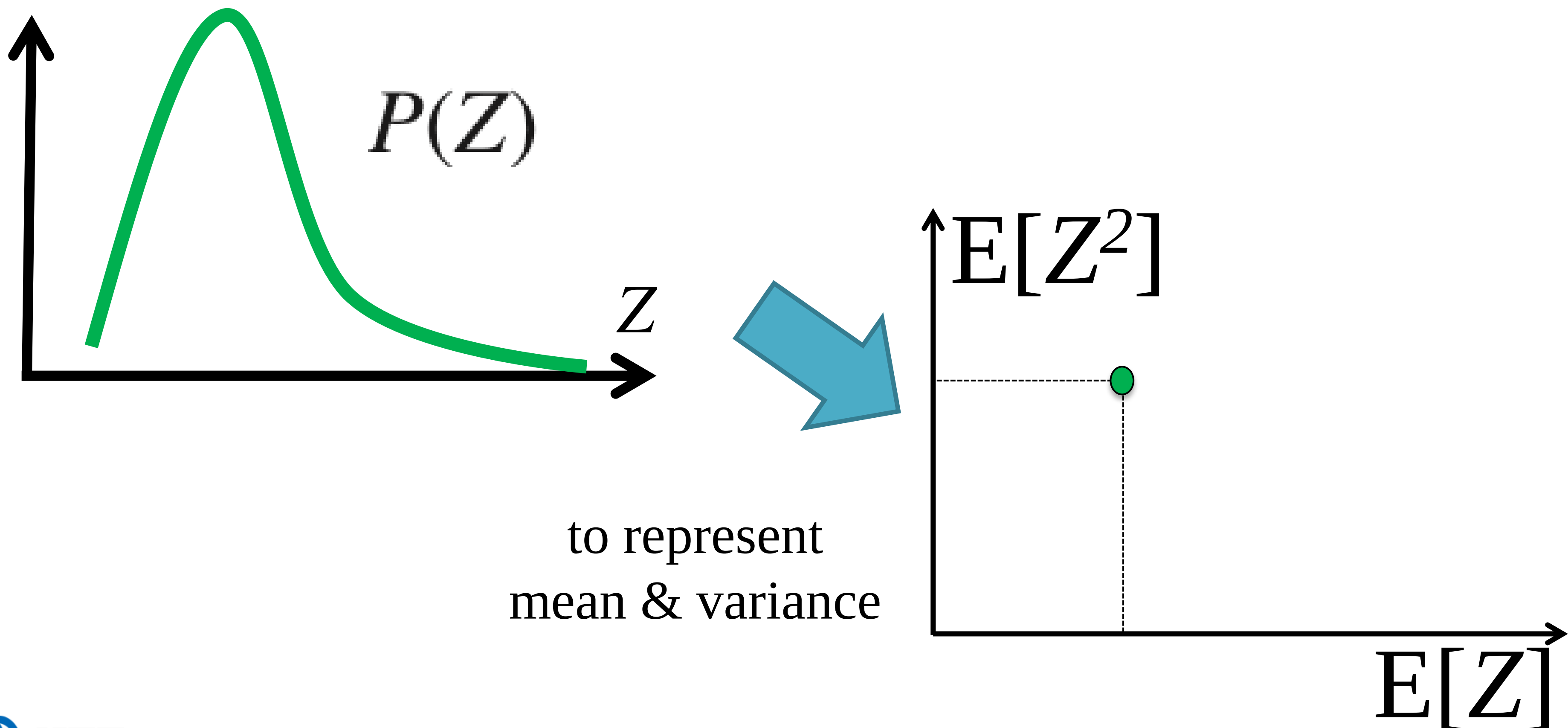
-> To deter … or not
the two d … e identical,
how … feature vectors
… ssification?

Kernel mean embedding!!!

Representing features of distributions

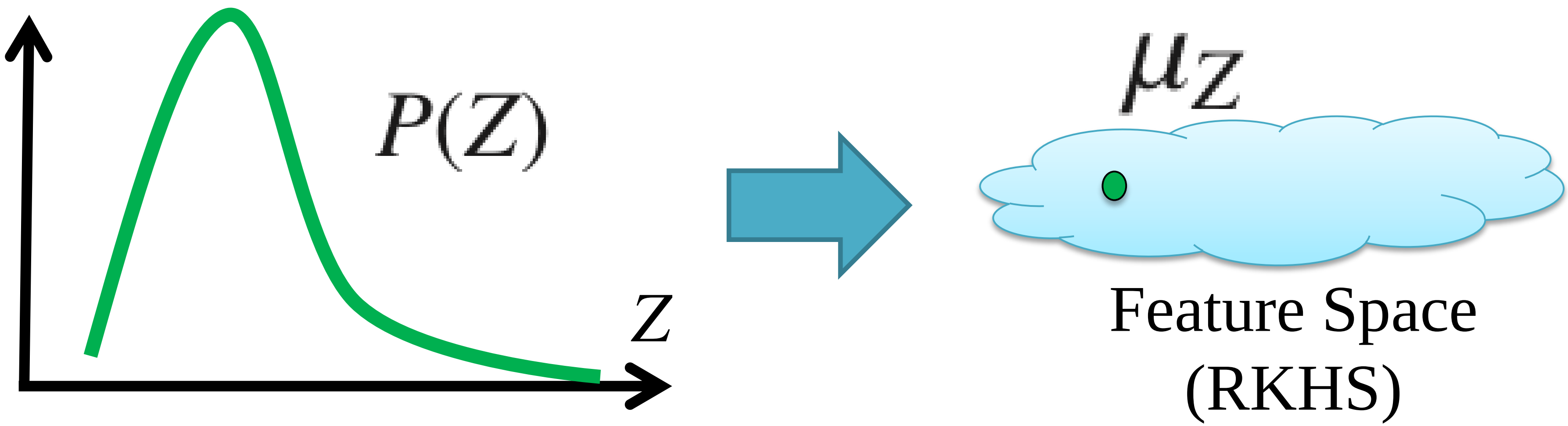


Representing features of distributions



Representing features of distributions

- **Kernel mean embedding**: map a distribution to a point in feature space called RKHS



When using Gaussian kernel,

$$\mu_Z \equiv \begin{bmatrix} \mathrm{E}[Z] \\ \mathrm{E}[Z^2] \\ \mathrm{E}[Z^3] \\ \vdots \end{bmatrix}$$

Reformulating label assignment rules

- By mapping distributions to points, label assignment rules can be rephrased as

If $\begin{cases} \mu_{X_{t+1}|S_X,S_Y} = \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} \neq \mu_{Y_{t+1}|S_Y} \end{cases}$

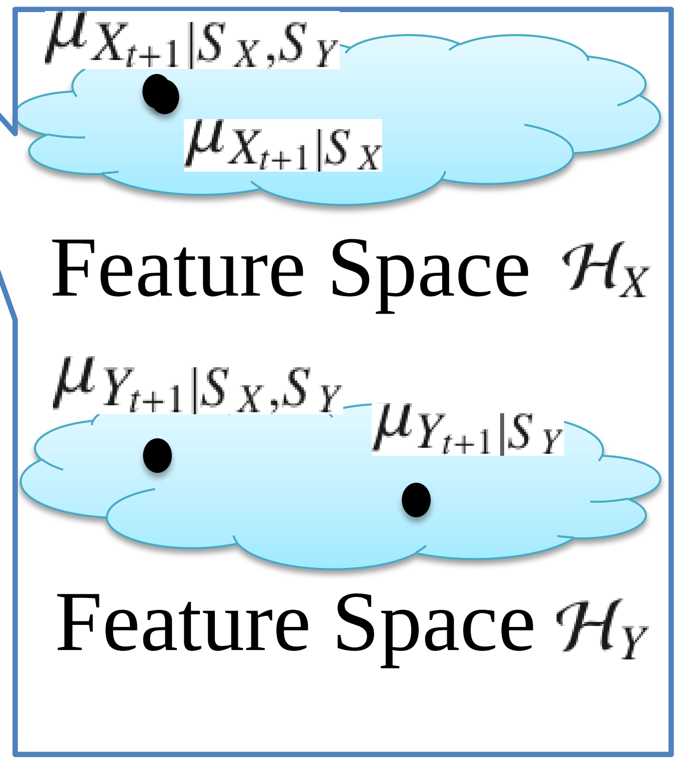
then $X \rightarrow Y$

If $\begin{cases} \mu_{X_{t+1}|S_X,S_Y} \neq \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} = \mu_{Y_{t+1}|S_Y} \end{cases}$

then $X \leftarrow Y$

If $\begin{cases} \mu_{X_{t+1}|S_X,S_Y} = \mu_{X_{t+1}|S_X} \\ \mu_{Y_{t+1}|S_X,S_Y} = \mu_{Y_{t+1}|S_Y} \end{cases}$

then *No Causation*

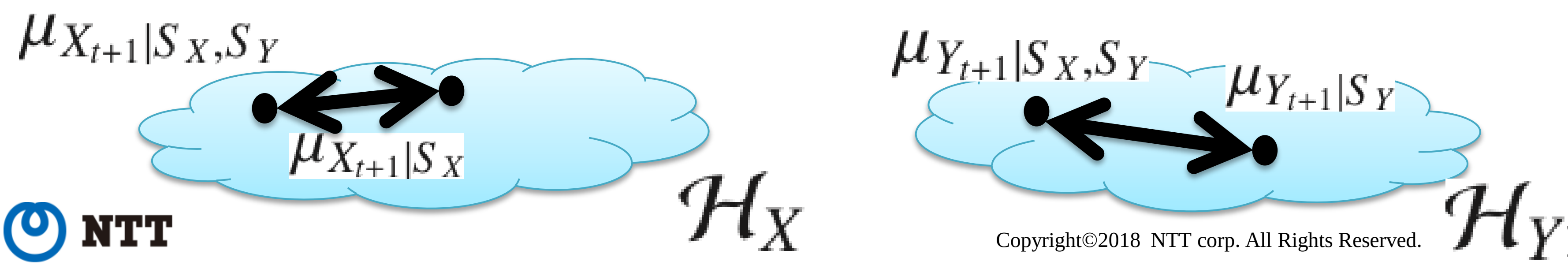


Feature representation

- We only have to determine <u>whether or not the two points are equal over time *t*</u>

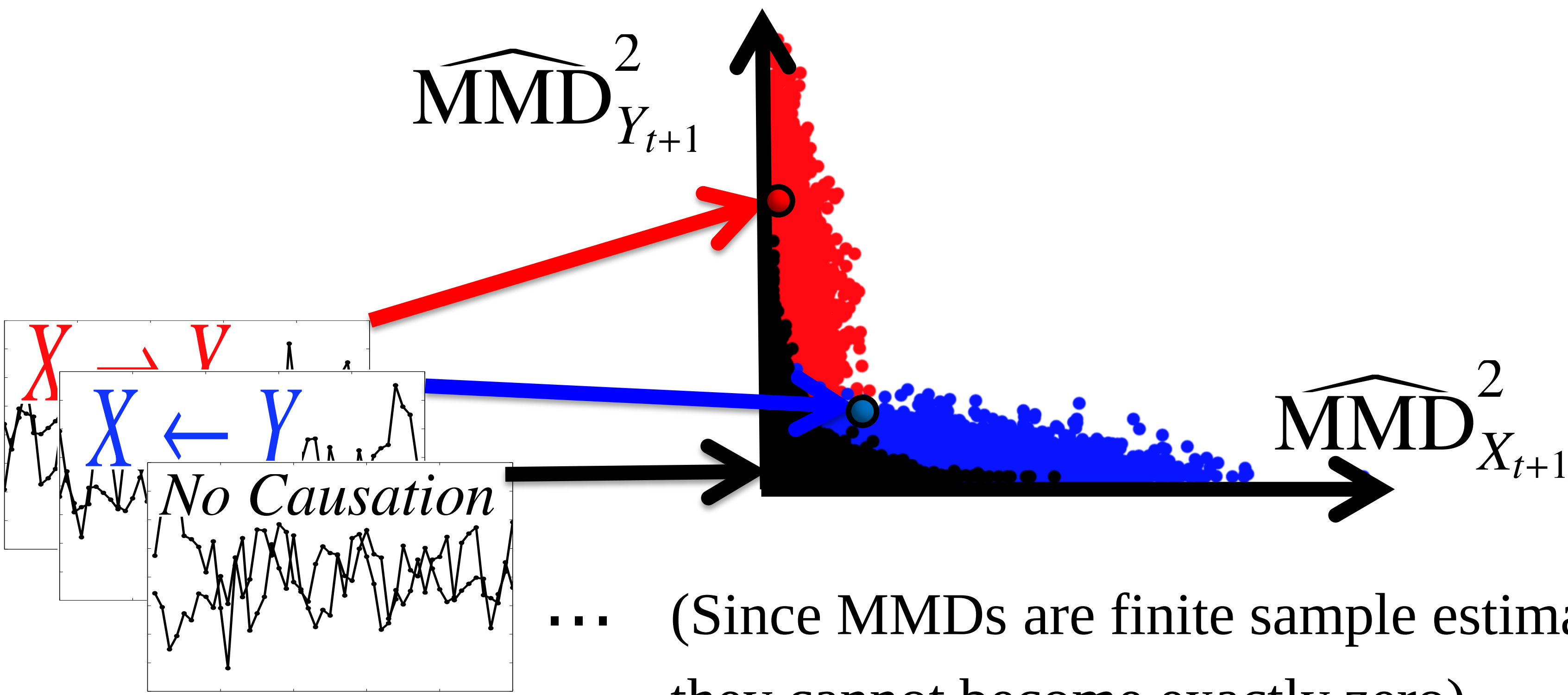
- We obtain feature vectors by using the distance between the points (called maximum mean discrepancy (MMD) [Gretton+ NIPS2007] in kernel method community)

$$MMD_{X_{t+1}} \qquad MMD_{Y_{t+1}}$$



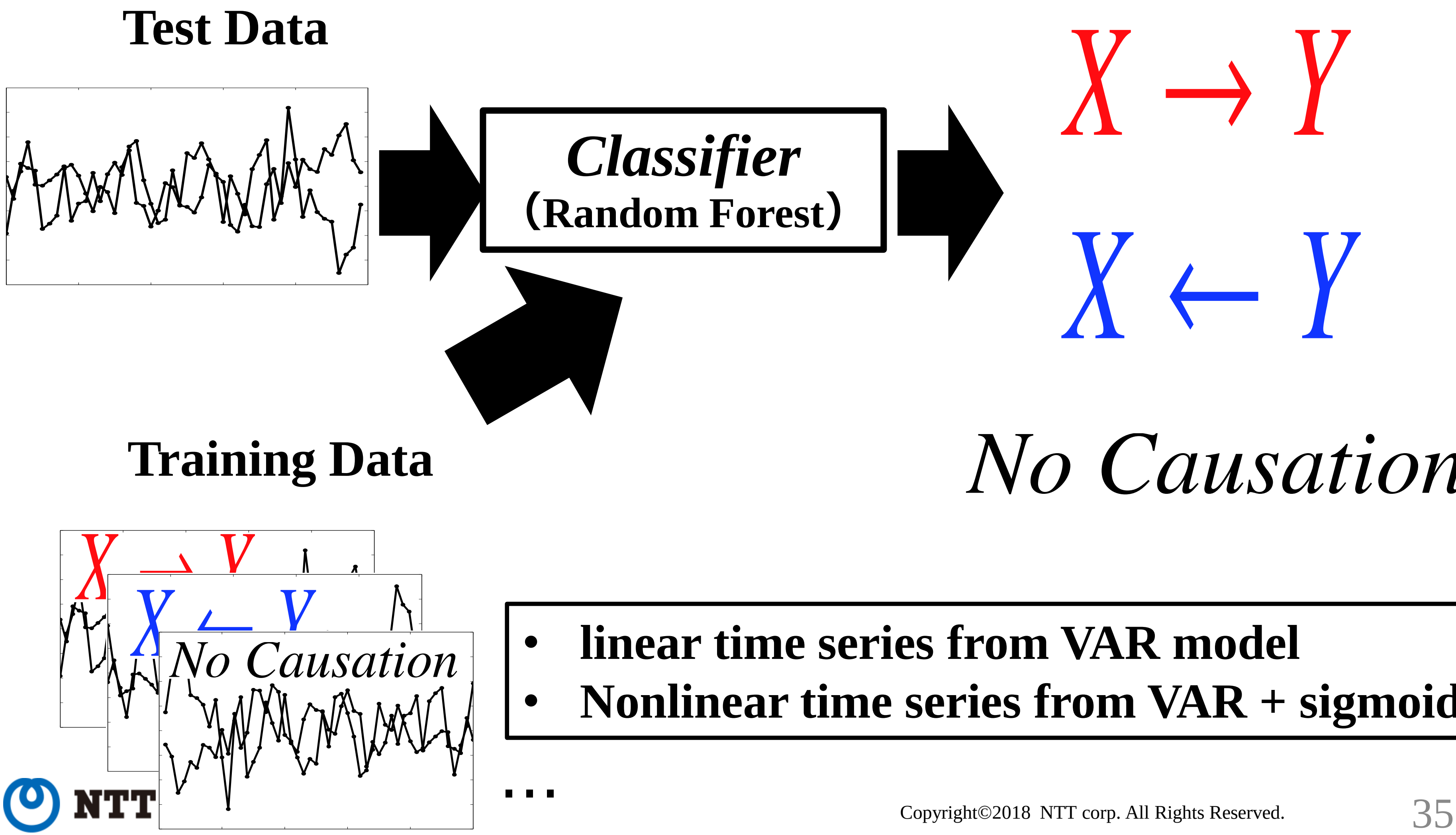
Feature representation

- By utilizing MMDs, we can obtain feature vectors that are sufficiently different depending on Granger causality



(Since MMDs are finite sample estimates, they cannot become exactly zero)

Experiments



- linear time series from VAR model
- Nonlinear time series from VAR + sigmoid

Experiment 1: Synthetic test data

Linear Test Data
-- generated from VAR model

$$\begin{bmatrix} X_{t+1} \\ Y_{t+1} \end{bmatrix} = \sum_{\tau=0}^{P-1} A_\tau \begin{bmatrix} X_{t-\tau} \\ Y_{t-\tau} \end{bmatrix} + E_\tau$$

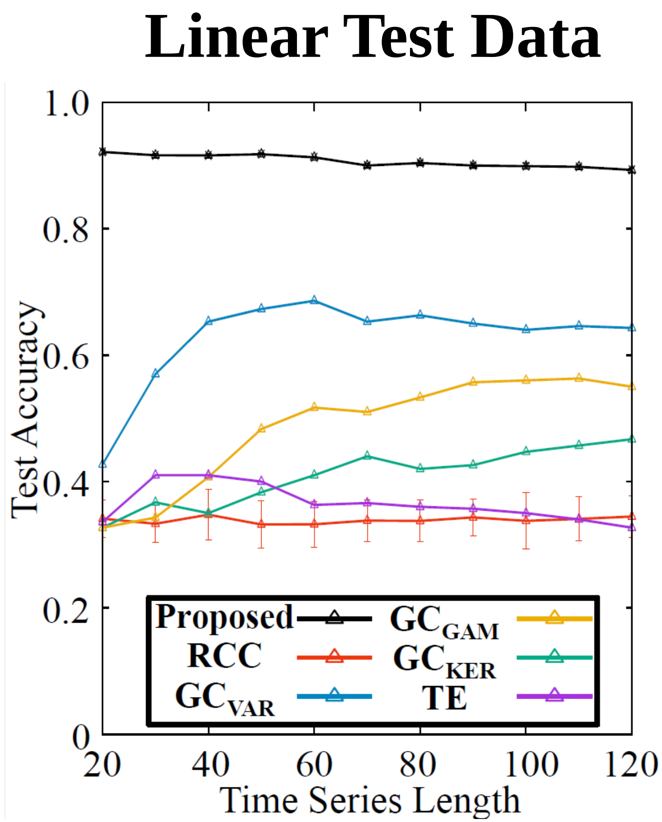
Nonlinear Test Data
-- generated from

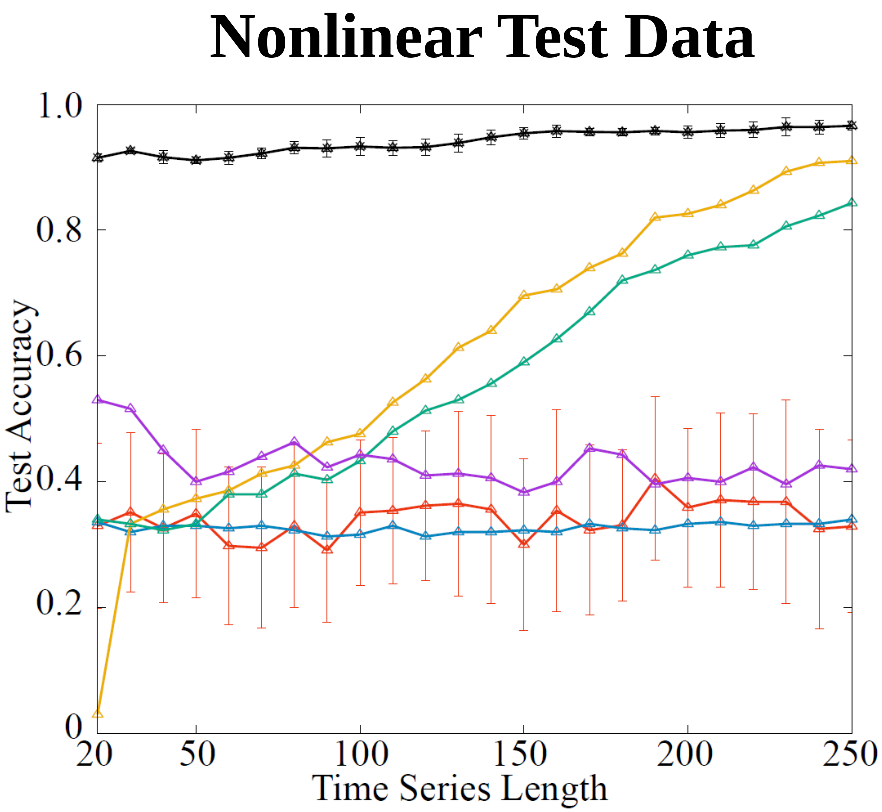
$$X_t = 0.2X_{t-1} + 0.9N_{X_t}$$
$$Y_t = -0.5 + \exp(-(X_{t-1} + X_{t-2})^2) + 0.7\cos(Y_{t-1}^2) + 0.3N_{Y_t}$$

- Prepare 300 pairs of bivariate time series
- Evaluate the number of time series whose causal relationships are correctly inferred (i.e., Test Accuracy)

NTT

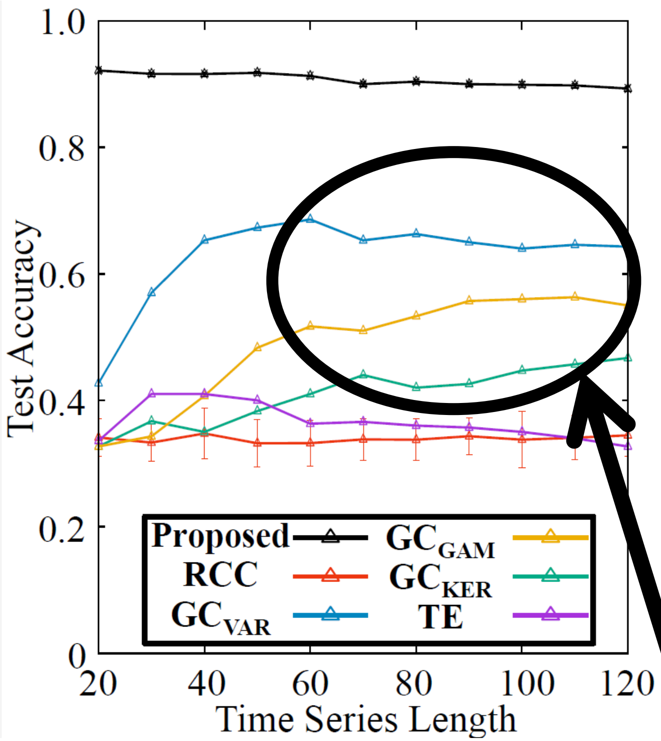
Test accuracy



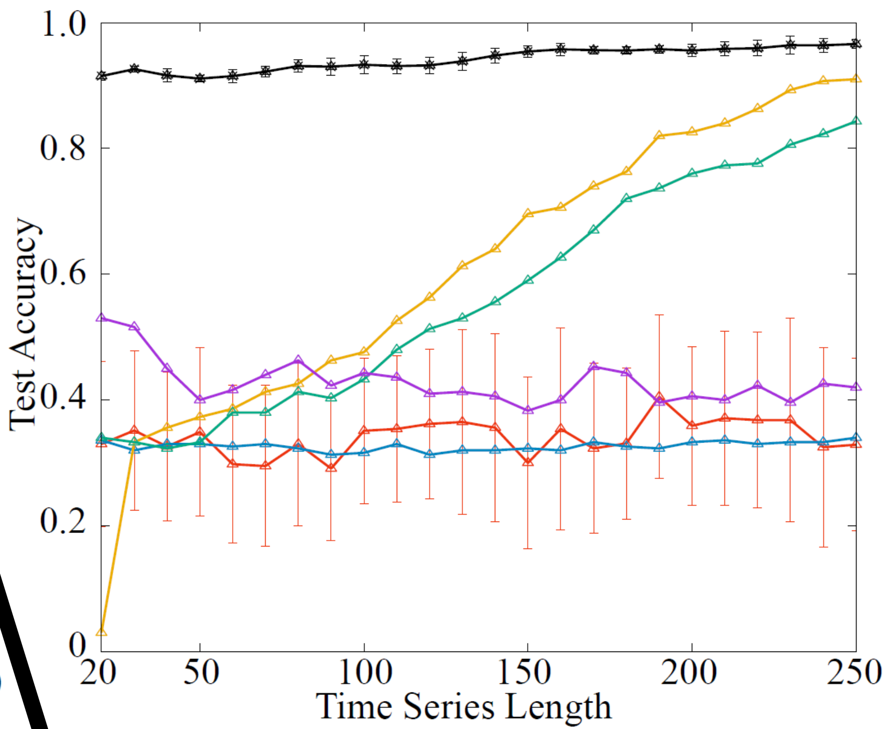


Test accuracy

Linear Test Data



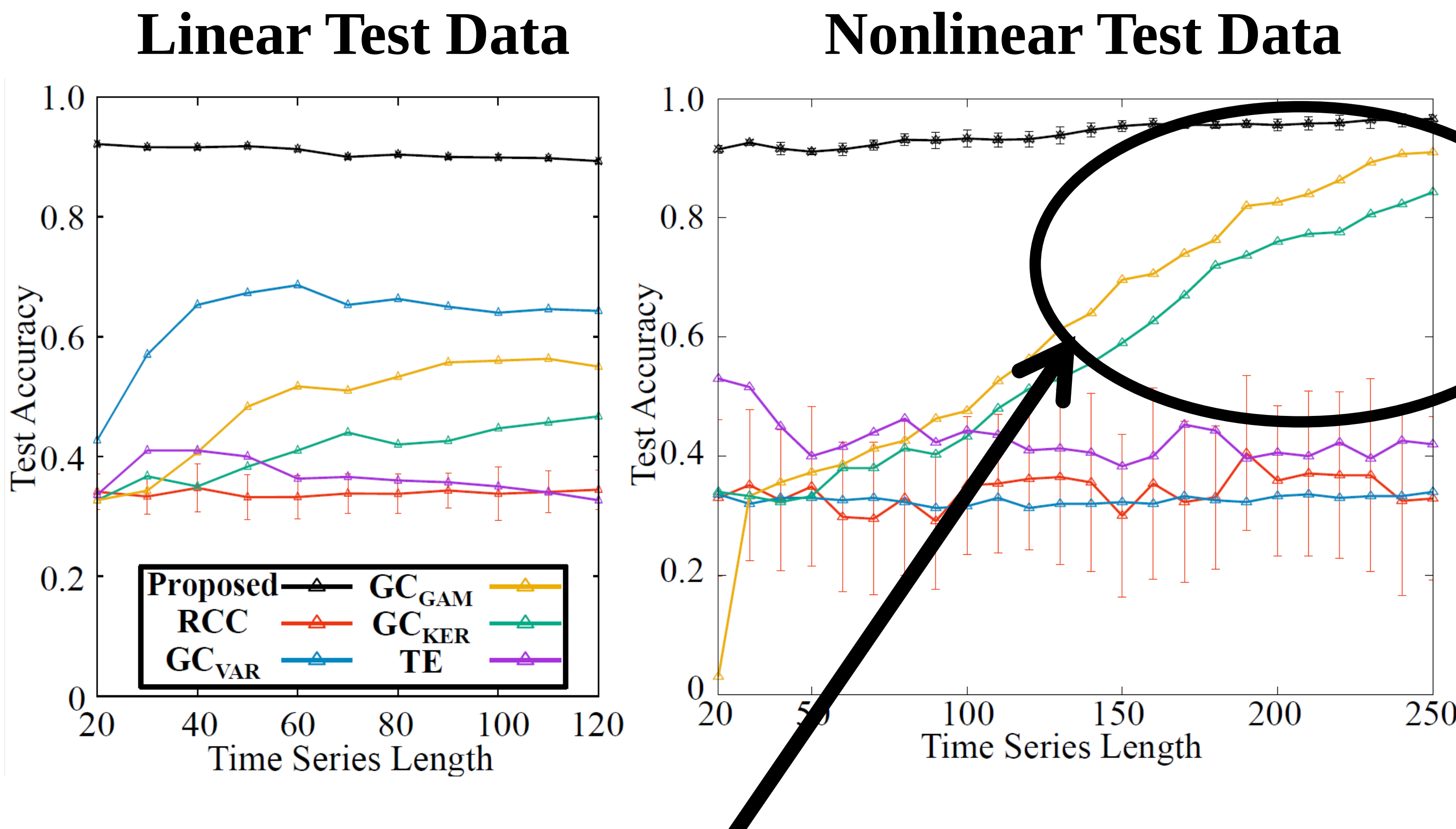
Nonlinear Test Data



Existing Granger causality methods

Test accuracy strongly depends on the regression model

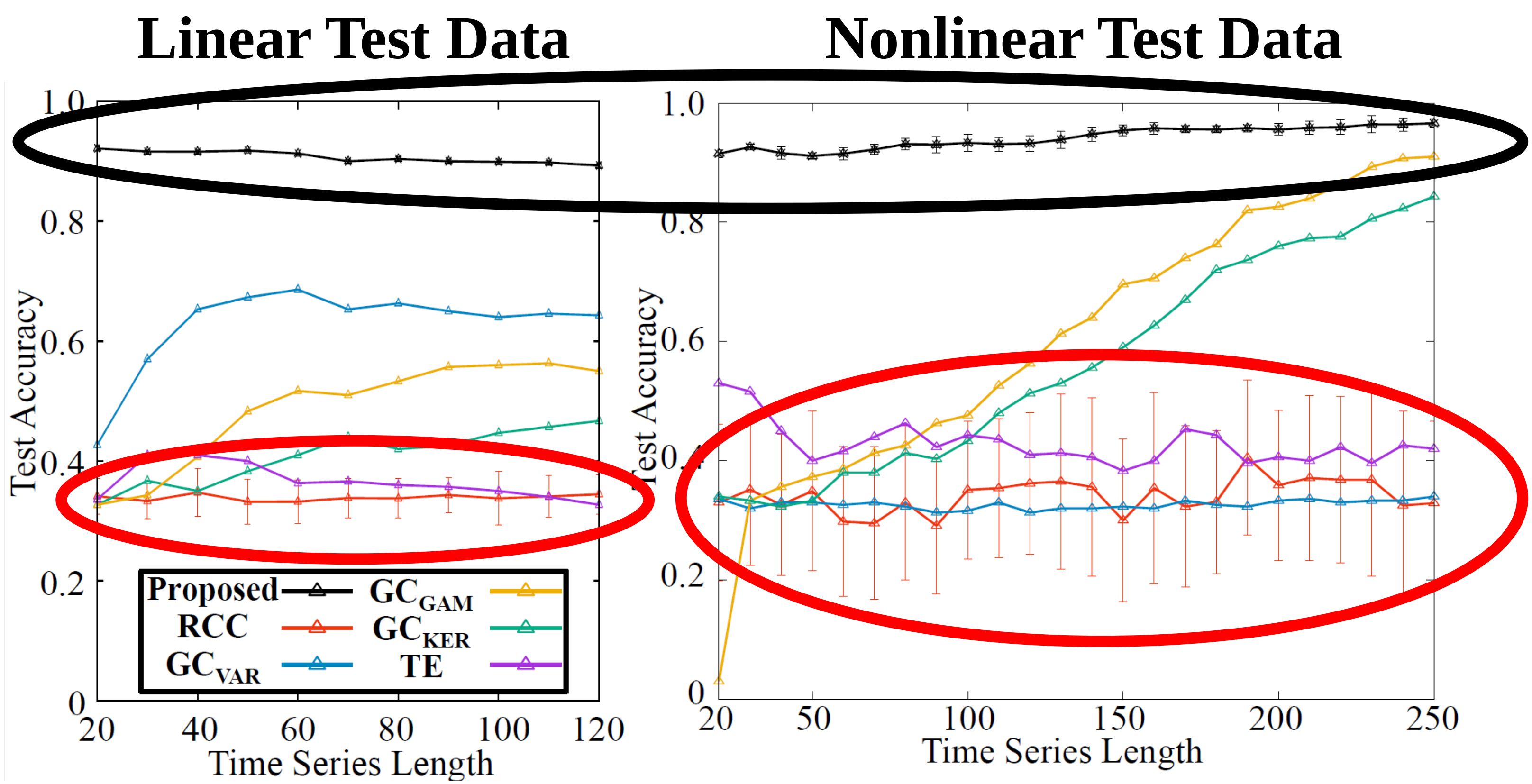
Test accuracy



GC_{KER} < GC_{GAM}

Kernel regression cannot be well fitted since time series are too short

Test accuracy



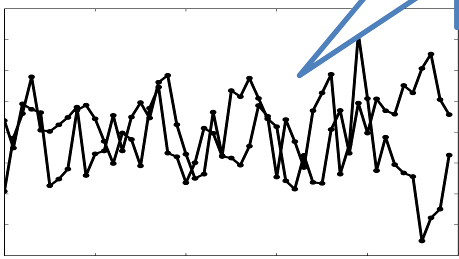
Proposed > Existing classification approach for i.i.d. data

Our feature representation is effective

Experiment 2: Real-world test data

e.g., *River Runoff*
X: Precipitation
Y: River runoff
(※truth: $X \rightarrow Y$)

Real-world
Test Data



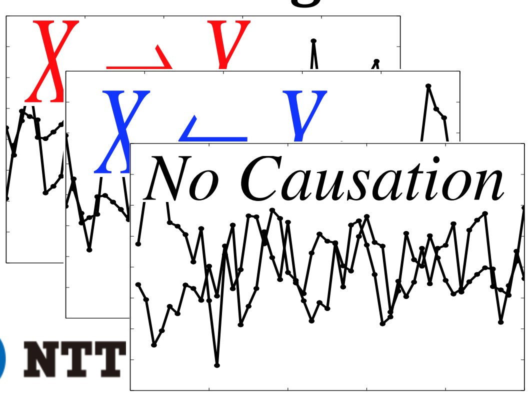
Classifier

$X \rightarrow Y$

$X \leftarrow Y$

No Causation

Synthetic
Training Data



…

True causal directions are given in literatures

Test accuracy

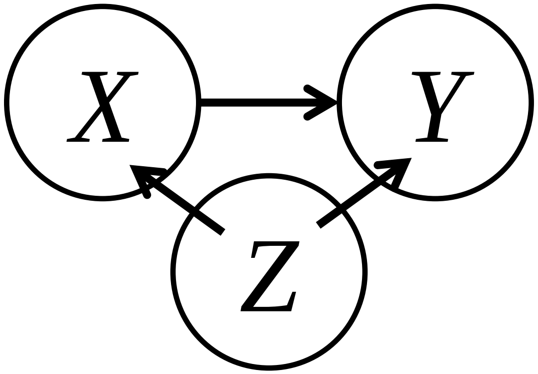
	Proposed	RCC	GC_{VAR}	GC_{GAM}	GC_{KER}	TE
Temperature ($T = 200$)	**0.961** (0.011)	0.432 (0.242)	0.950	0.848	0.234	0.492
Radiation ($T = 200$)	**0.987** (0.053)	0.515 (0.345)	0.156	0.0	0.782	0.394
Internet ($T = 200$)	**1.0** (0.0)	0.478 (0.222)	0.157	0.387	0.261	0.498
Sun Spots ($T = 200$)	**1.0** (0.0)	0.435 (0.182)	0.908	0.704	0.076	0.522
River Runoff ($T = 200$)	**0.958** (0.058)	0.399 (0.193)	0.684	0.406	0.155	0.485

Our Proposed sufficiently worked better than other methods

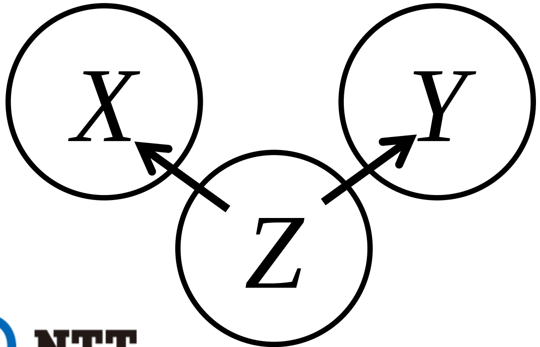
How can we extend proposed approach to multivariate time series?

Granger causality definition for multivariate time series

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



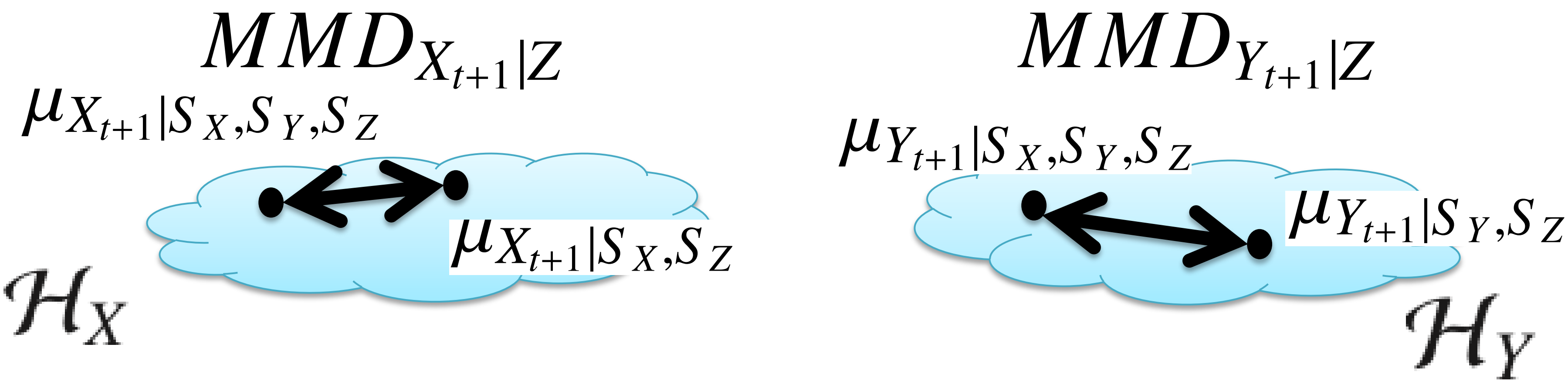
if $P(Y_{t+1}|S_X, S_Y, \underline{S_Z}) \neq P(Y_{t+1}|S_Y, \underline{S_Z})$



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

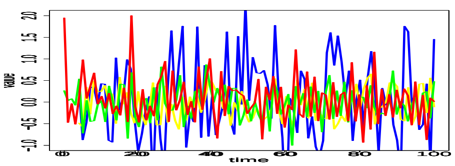


- By using additional MMDs, we formulate feature representation for multivariate time series

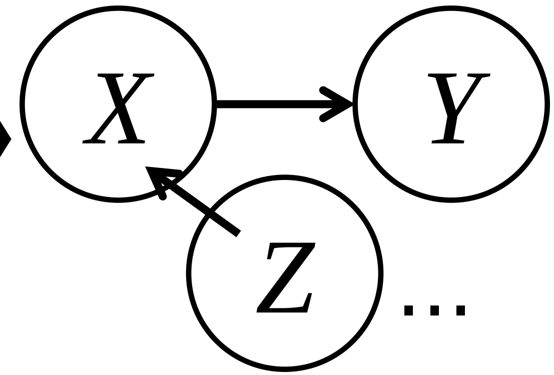
Experiment 3: Multivariate real-world data

Real-world
Test Data

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



Classifier



Synthetic
Training Data

True causal directions are given in database

Macro F1 score and micro F1 score

	$\textbf{Proposed}_{tri}$	$\textbf{Proposed}_{bi}$	$\textbf{GC}_{VAR}$	$\textbf{GC}_{GAM}$	$\textbf{GC}_{KER}$
macro F1 score	**0.483**	0.415	0.457	0.437	0.351
micro F1 score	**0.637**	0.549	0.567	0.513	0.436

※Higher is better

Macro F1 score and micro F1 score

	Proposed$_{tri}$	**Proposed**$_{bi}$	**GC**$_{VAR}$	**GC**$_{GAM}$	**GC**$_{KER}$
macro F1 score	**0.483**	0.415	0.457	0.437	0.351
micro F1 score	**0.637**	0.549	0.567	0.513	0.436

※Higher is better

Proposed with extended feature representation worked better

Conclusion

- **Classification approach to Granger causality identification**
  - ✓ Requires no selection of regression models
  - ✓ Performs sufficiently better than existing model-based approach
  - ✓ Can be applied to multivariate time series

- Future work:
  - ✓ Addressing more complicated setting
    - ➢ e.g., causal direction changes over time t

Questions ?