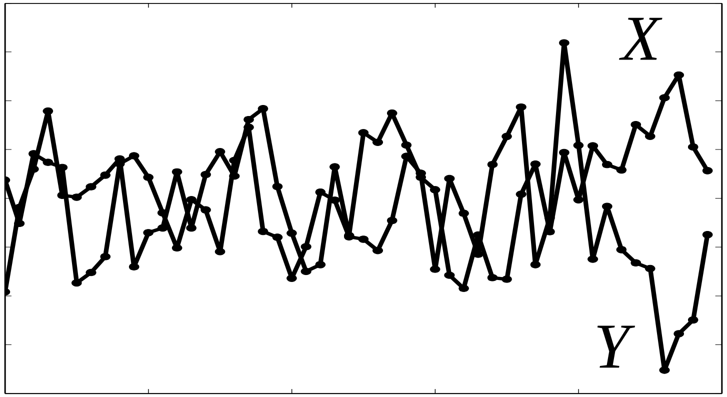
教師あり学習に基づく時系列の因果推論

A Supervised Learning Approach to Causal Inference in Time Series

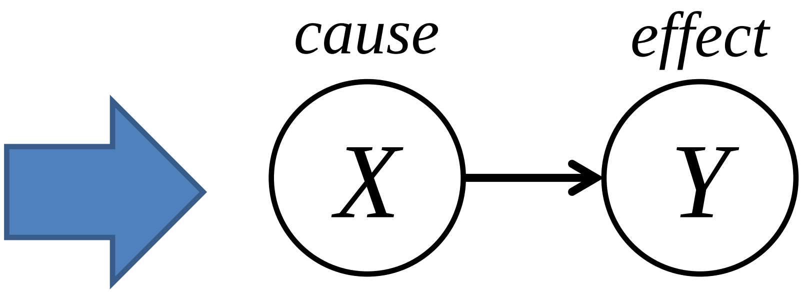
Yoichi Chikahara, Akinori Fujino
NTT Communication Science Laboratories
Kyoto, Japan

Causal inference in time series

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



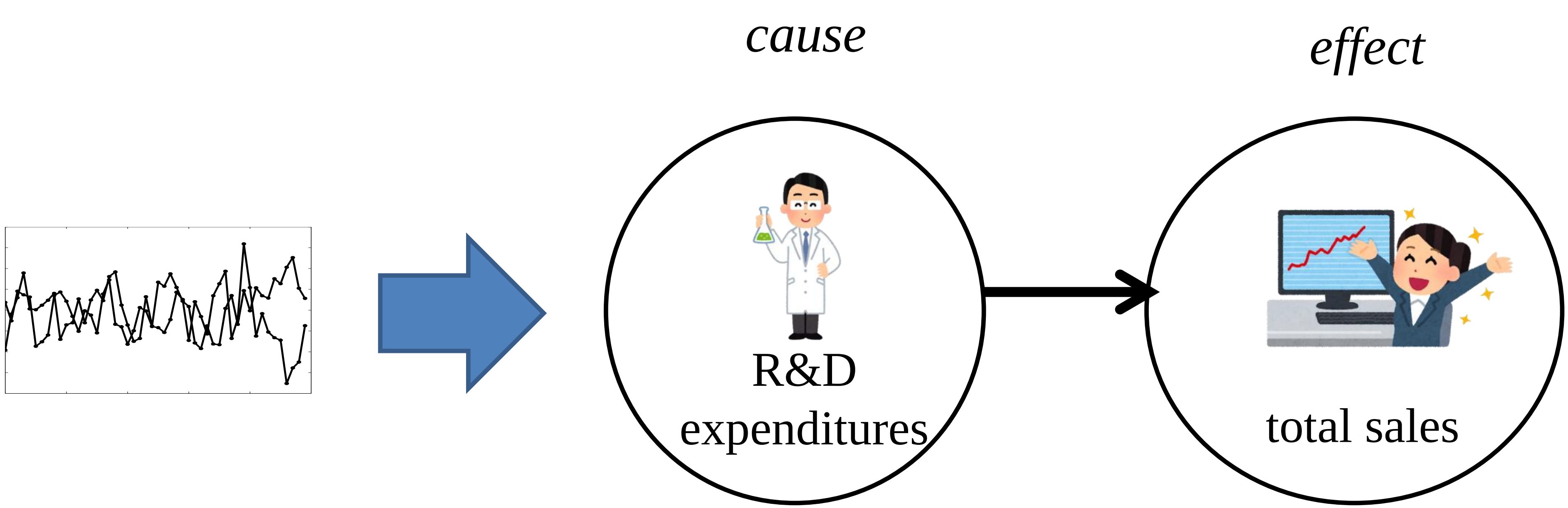
Input: Time Series Data



Output: *Causal Relationships*

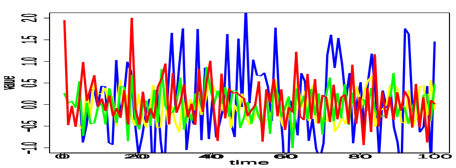
Application 1: Economics

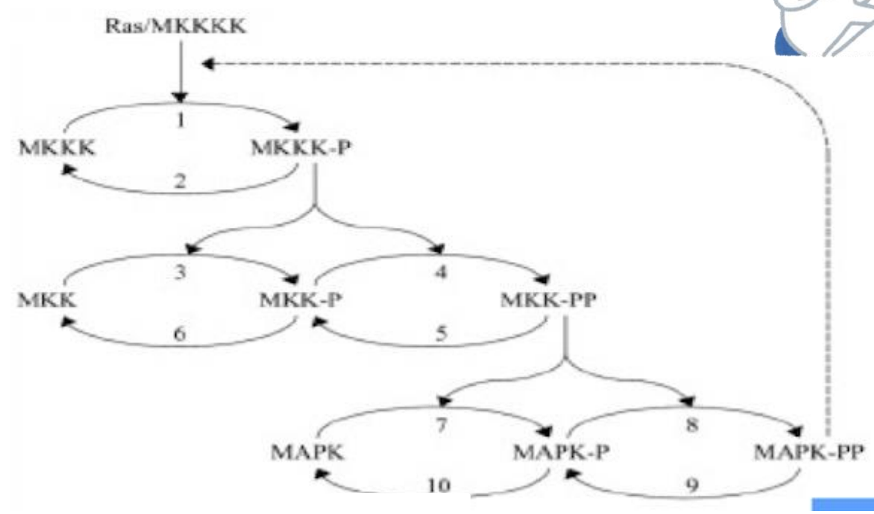
- Finding that R&D expenditures *influences* total sales is useful for companies



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

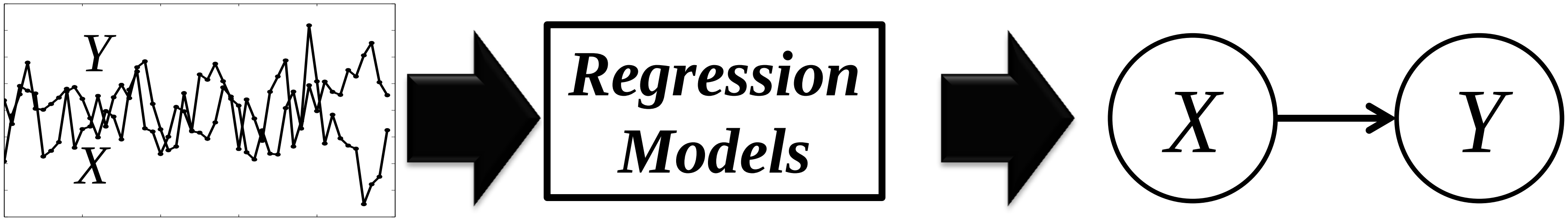
if the past values of X are **helpful in predicting** the future values of Y

Assumption

At **any** time point t

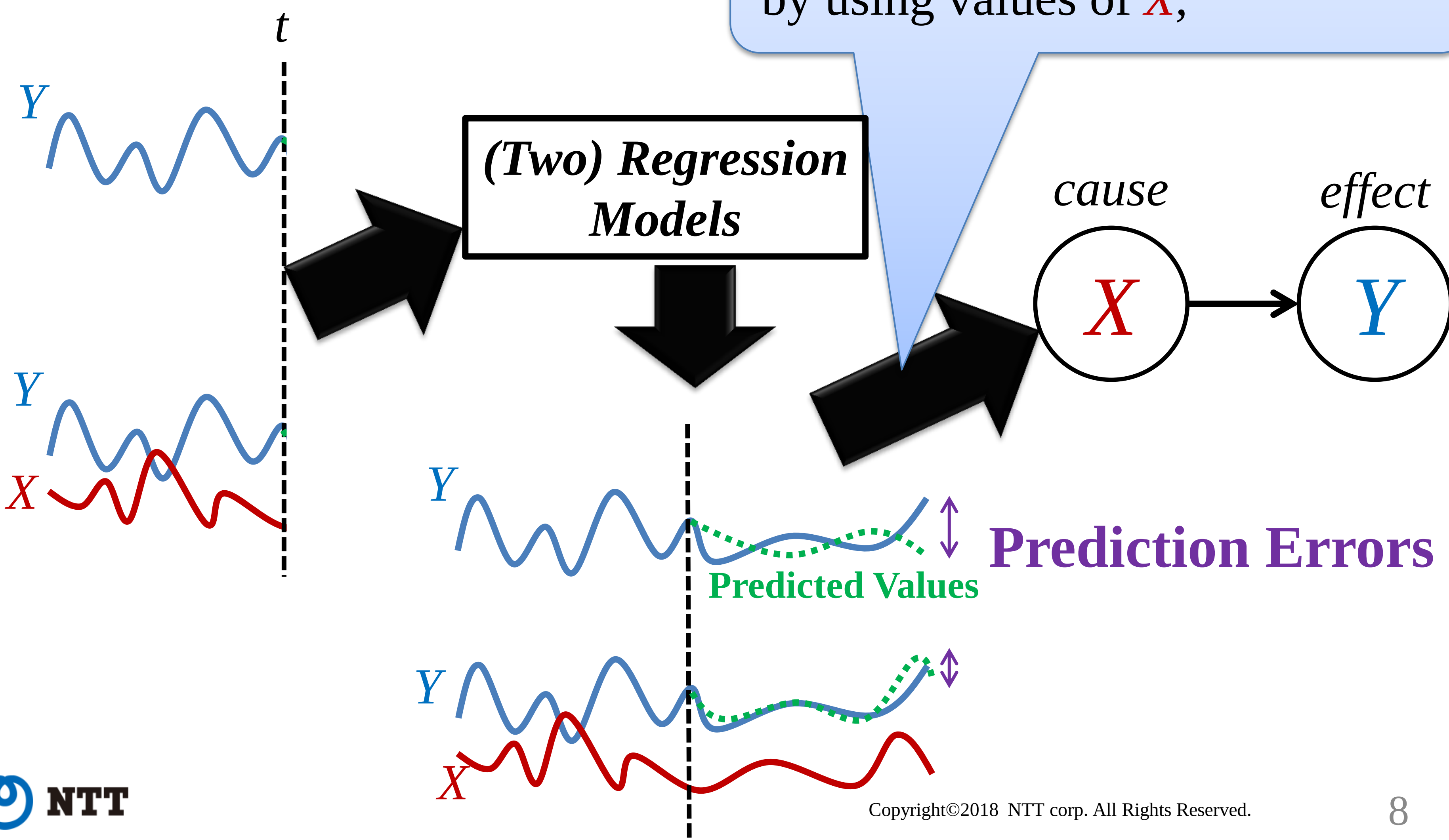
the causal direction is the same

Existing Approach: Using regression models

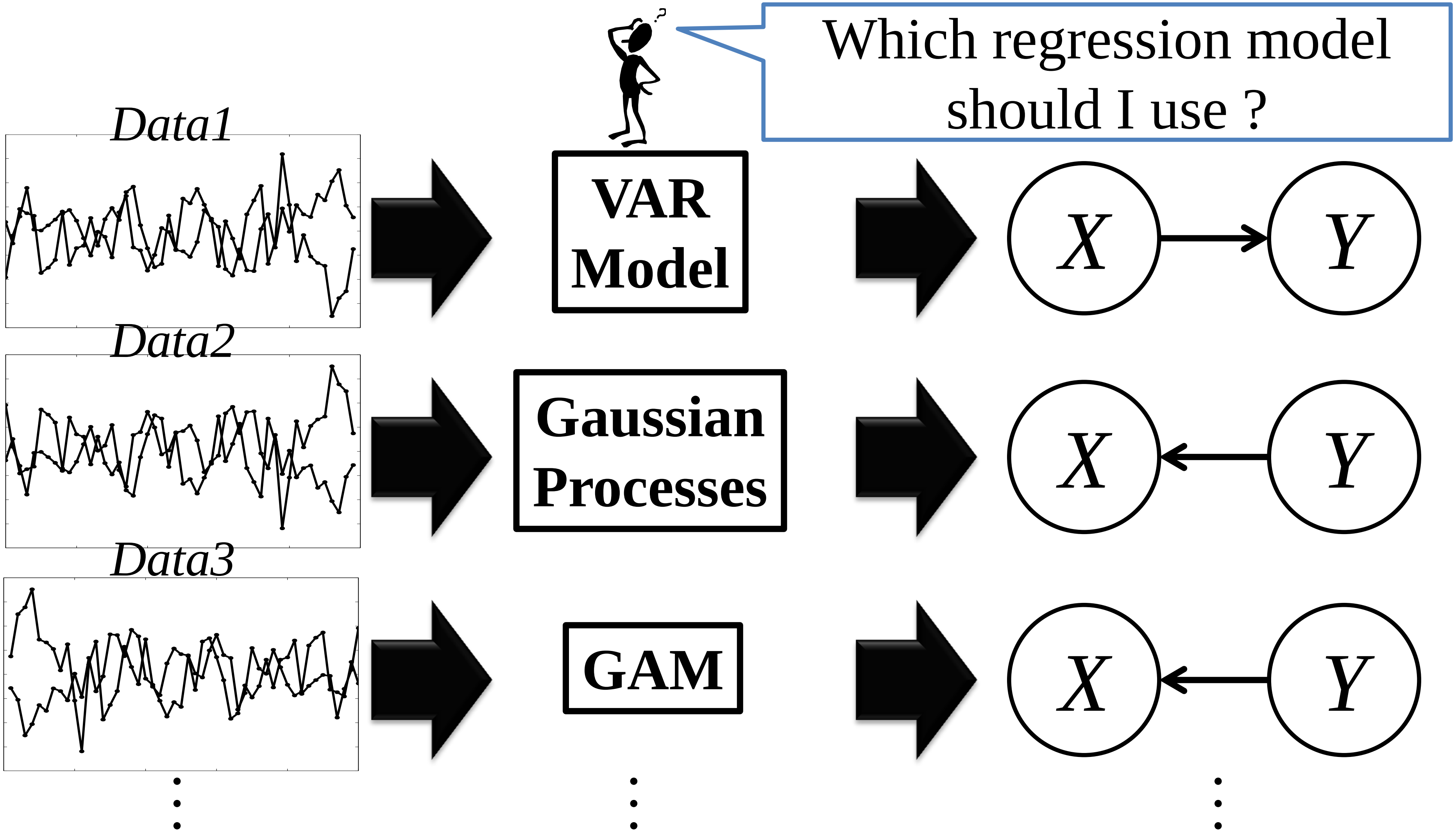


Existing approach:
Compare prediction errors with/without using values of *X*

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

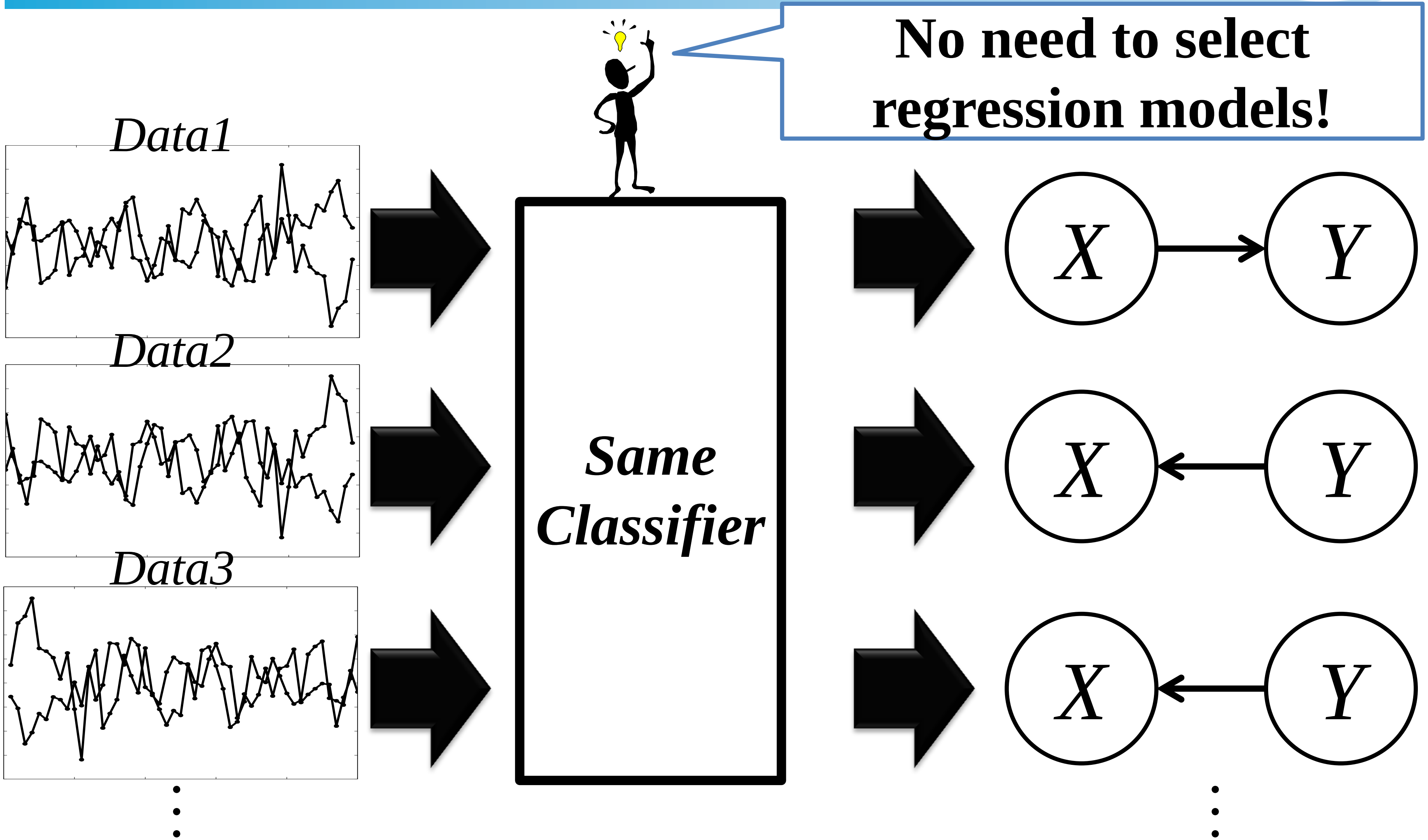


Weakness of existing approach

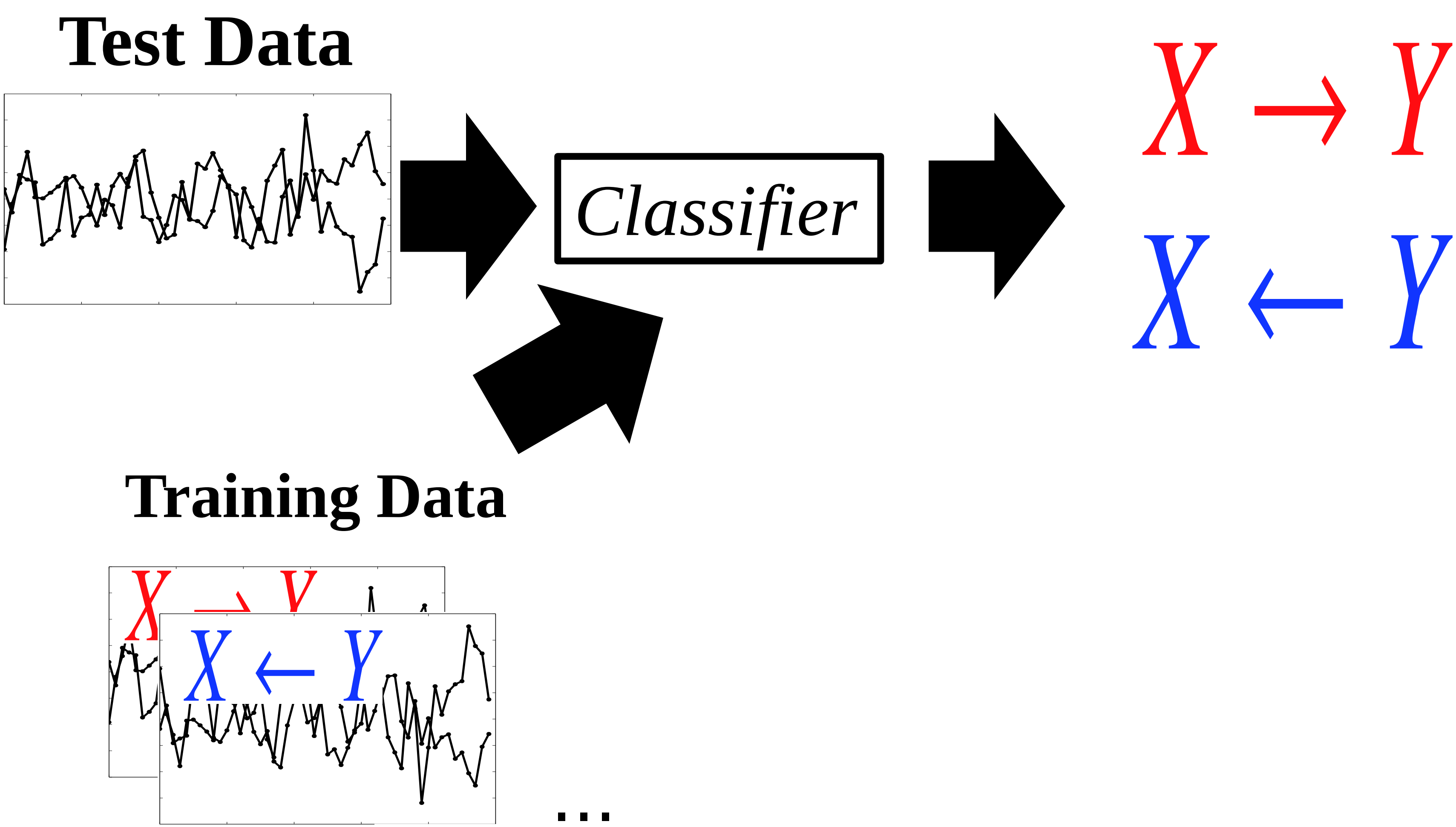


Model misspecification leads to **low inference accuracy**

Our approach:
Causal inference via classification

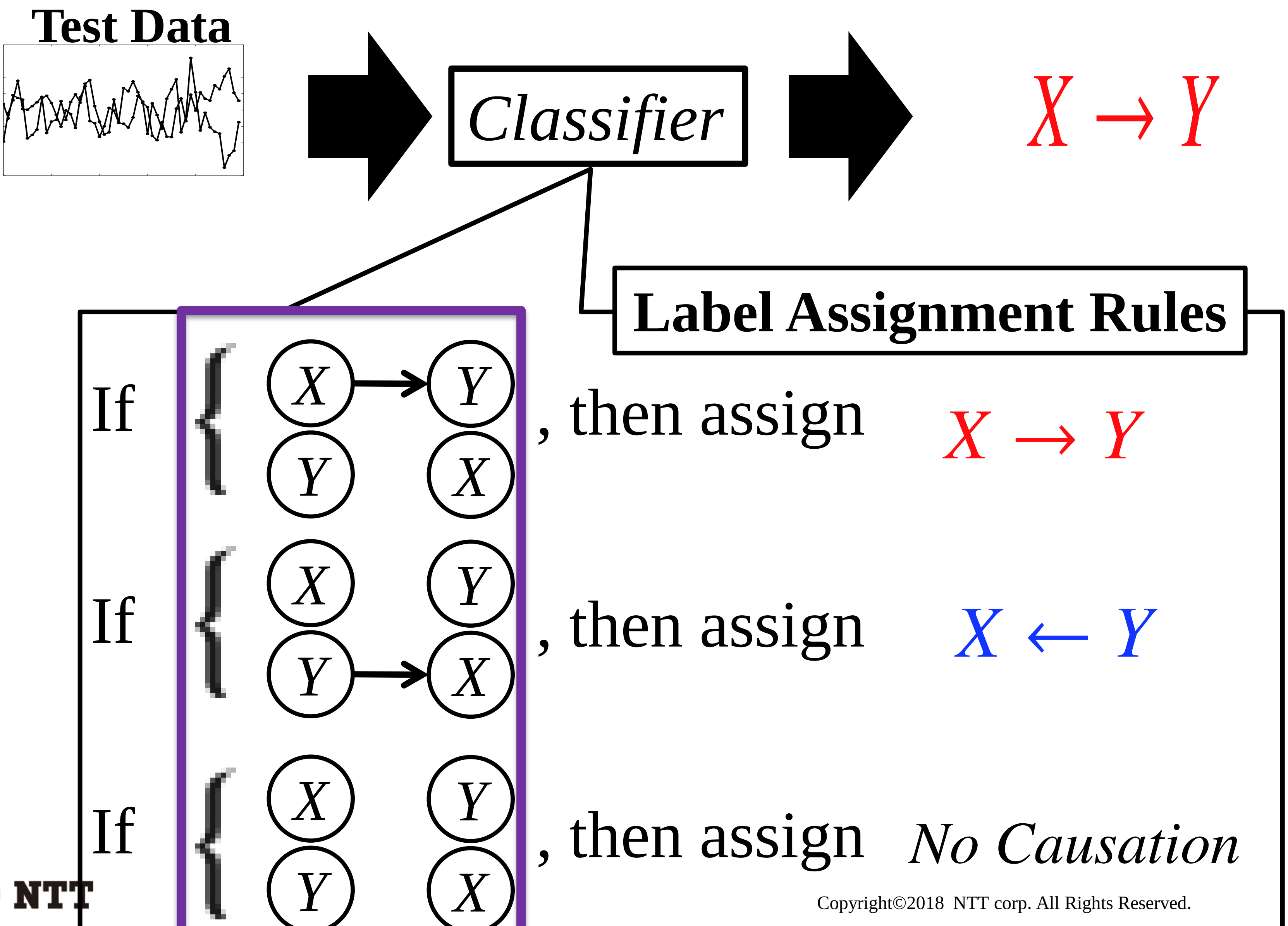


Our approach:
Causal inference <u>from time series data</u> via supervised learning

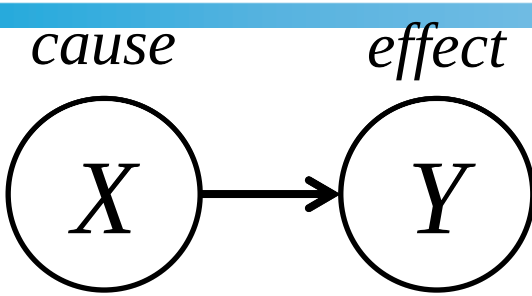


(Data where causal relationships are <u>known</u>)

Building a classifier for Granger causality identification



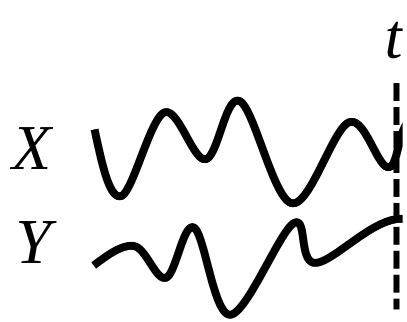
Revisiting definition of Granger causality



if the following holds:

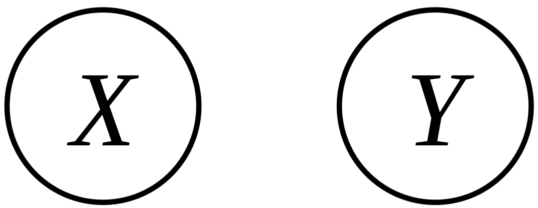
$$\underline{P(Y_{t+1}|S_X, S_Y)} \neq \underline{P(Y_{t+1}|S_Y)}$$

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



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

Similarly,



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

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

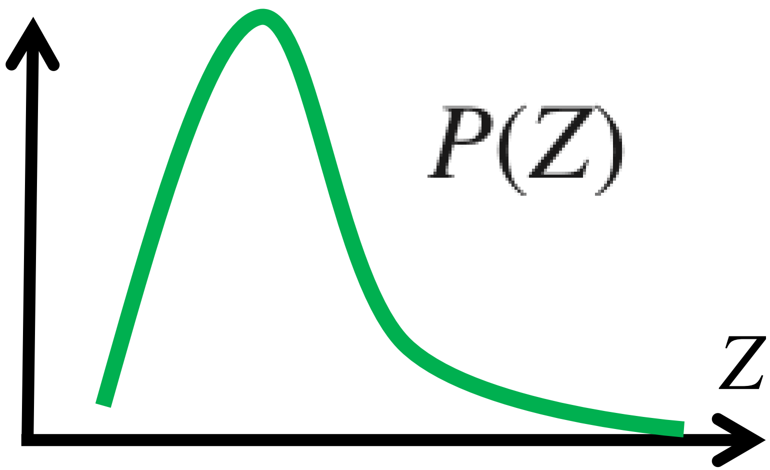
then $X \leftarrow Y$

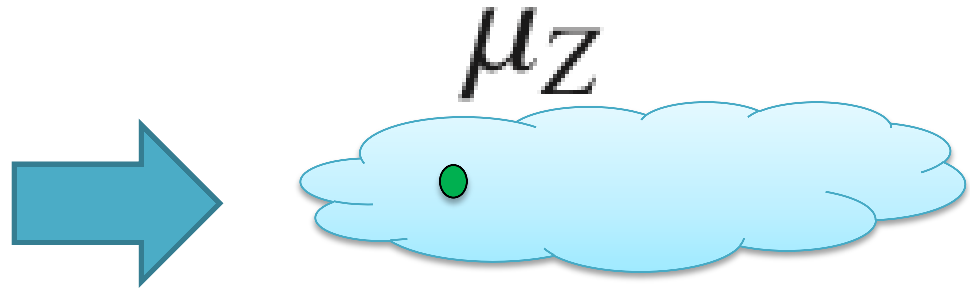
If $\begin{cases} P(Y_{t+1}|S_X, S_Y) = P(Y_{t+1}|S_Y) \\ P(X_{t+1}|S_X, S_Y) = P(X_{t+1}|S_X) \end{cases}$

then *No Causation*

Representing features of distributions

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





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

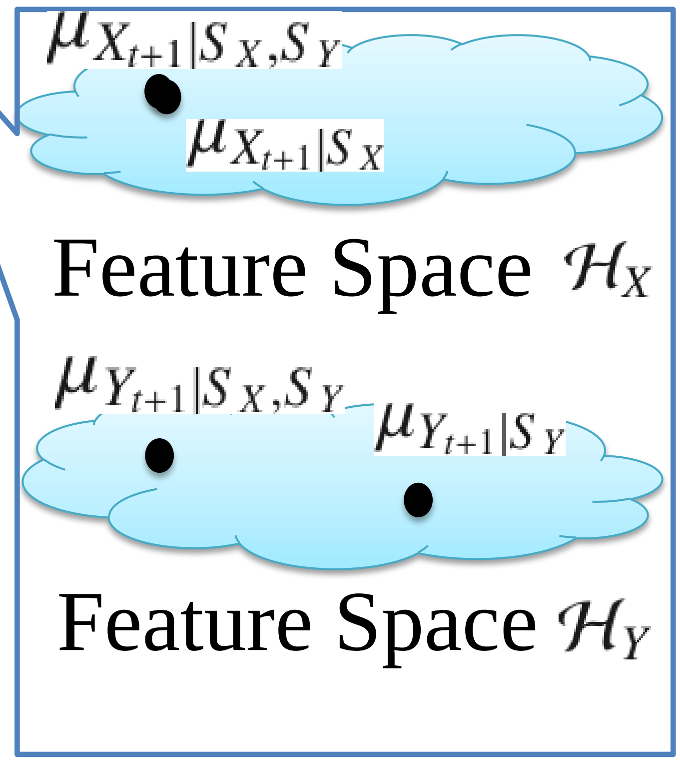
then $X \rightarrow Y$

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*

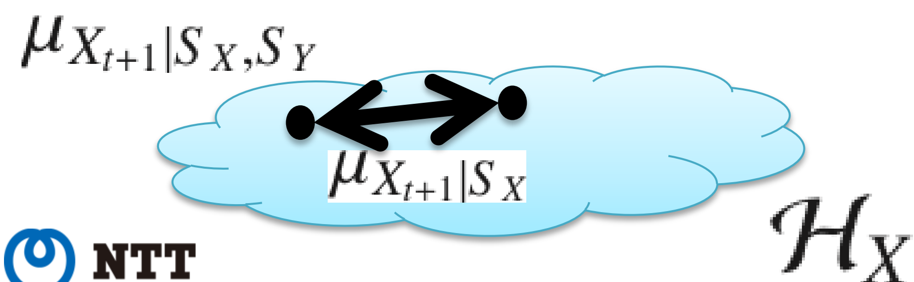


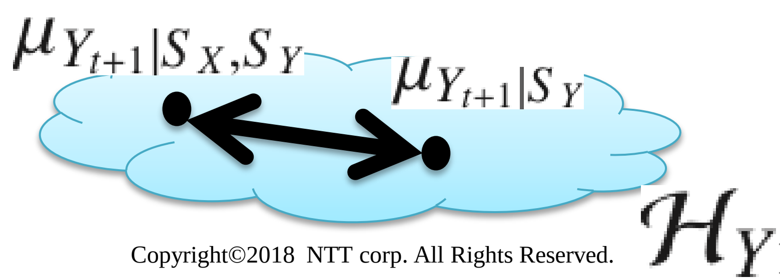
Feature representation

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

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

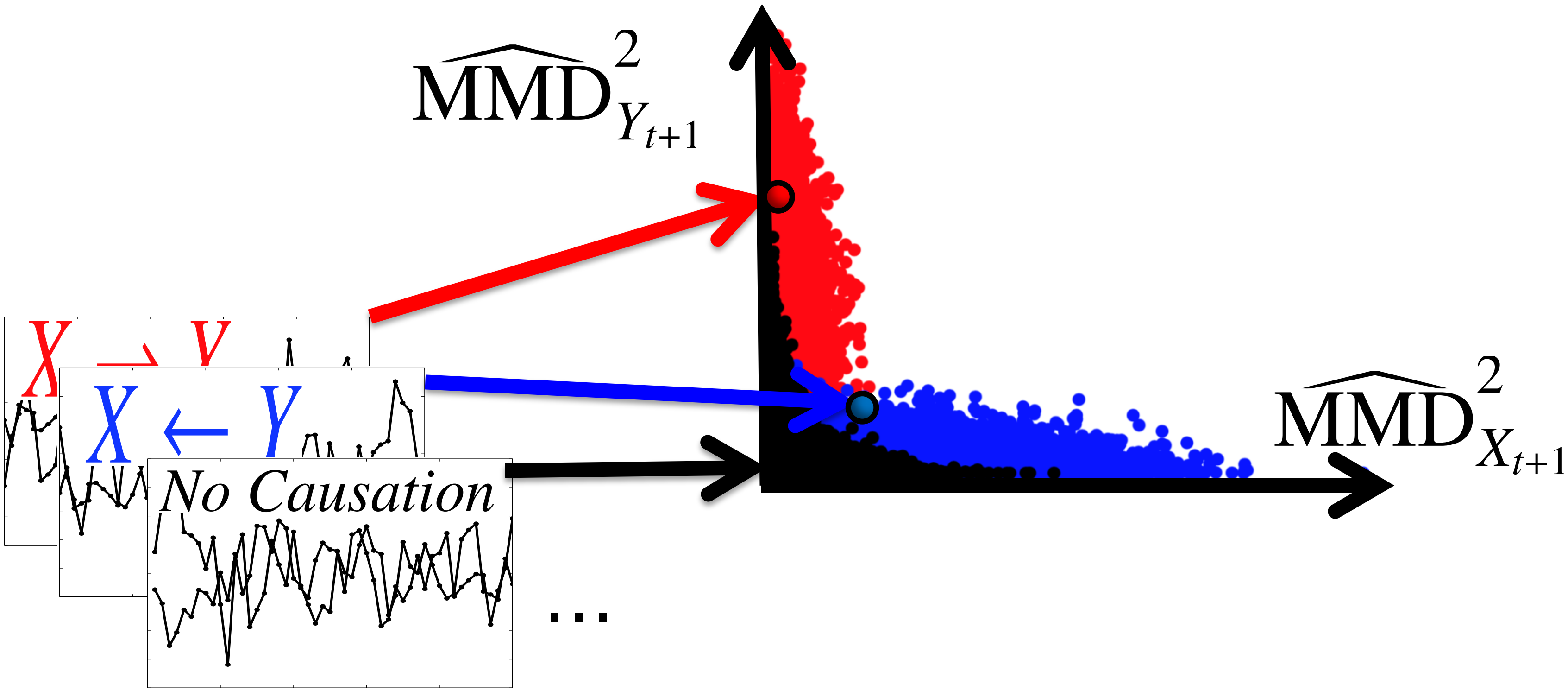
$MMD_{X_{t+1}}$ $\qquad\qquad$ $MMD_{Y_{t+1}}$





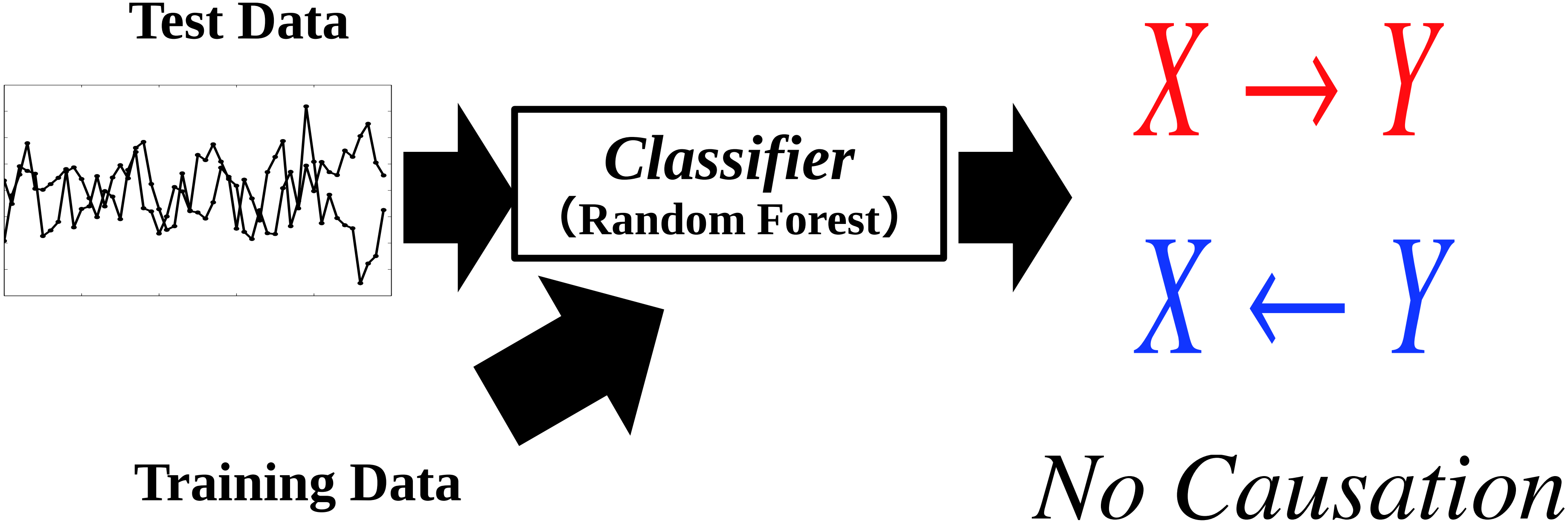
Feature representation

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

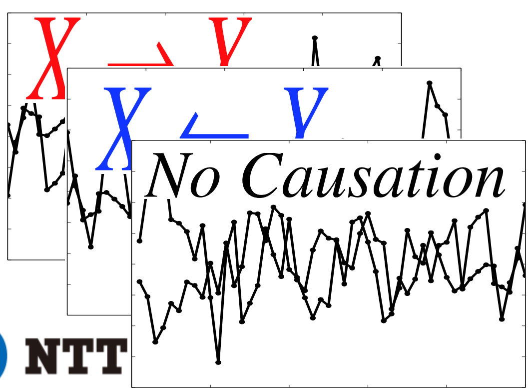


Experiments

Test Data



Training Data



…

- **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

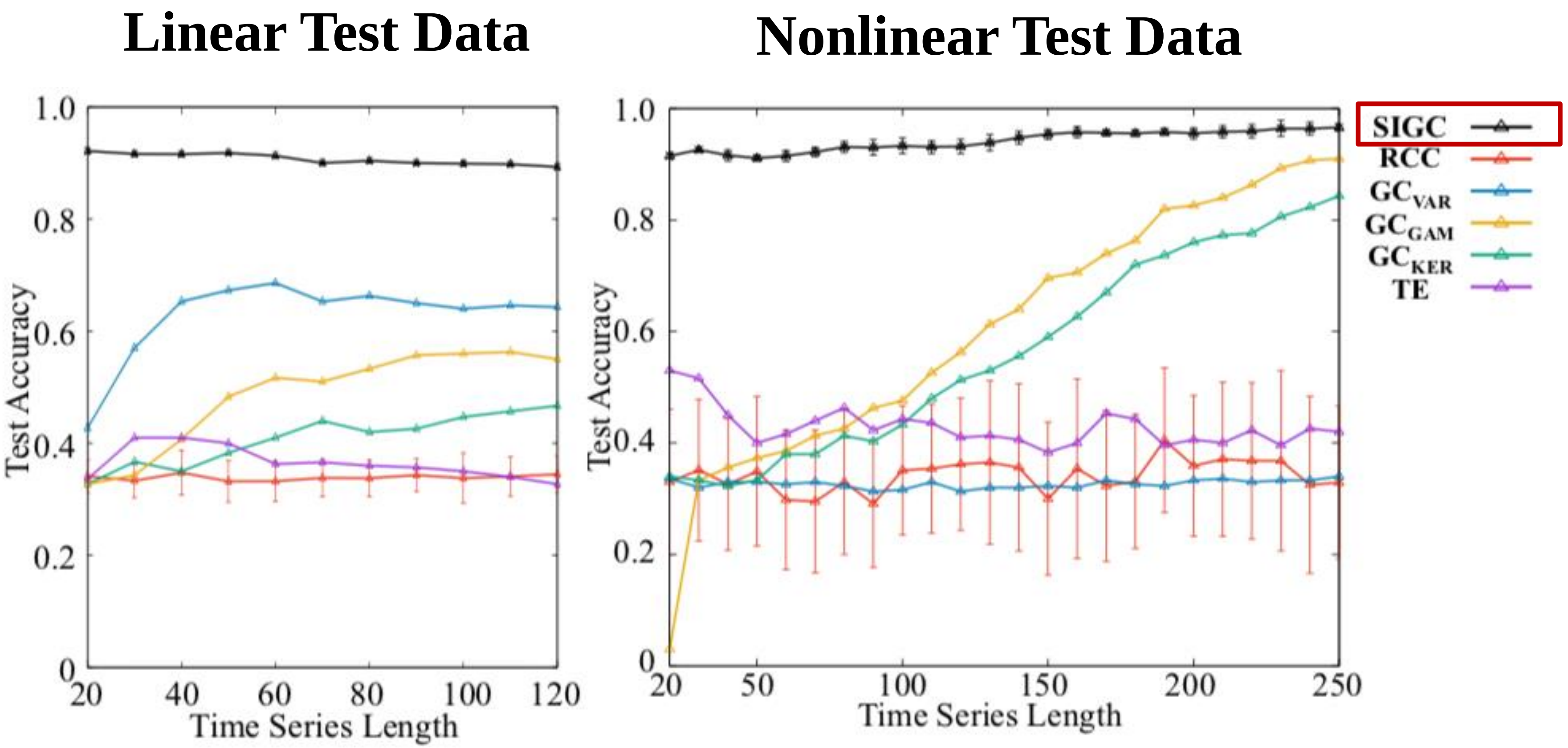
-- generated from

$$X_t = 0.2X_{t-1} + 0.9N_{X_t}$$

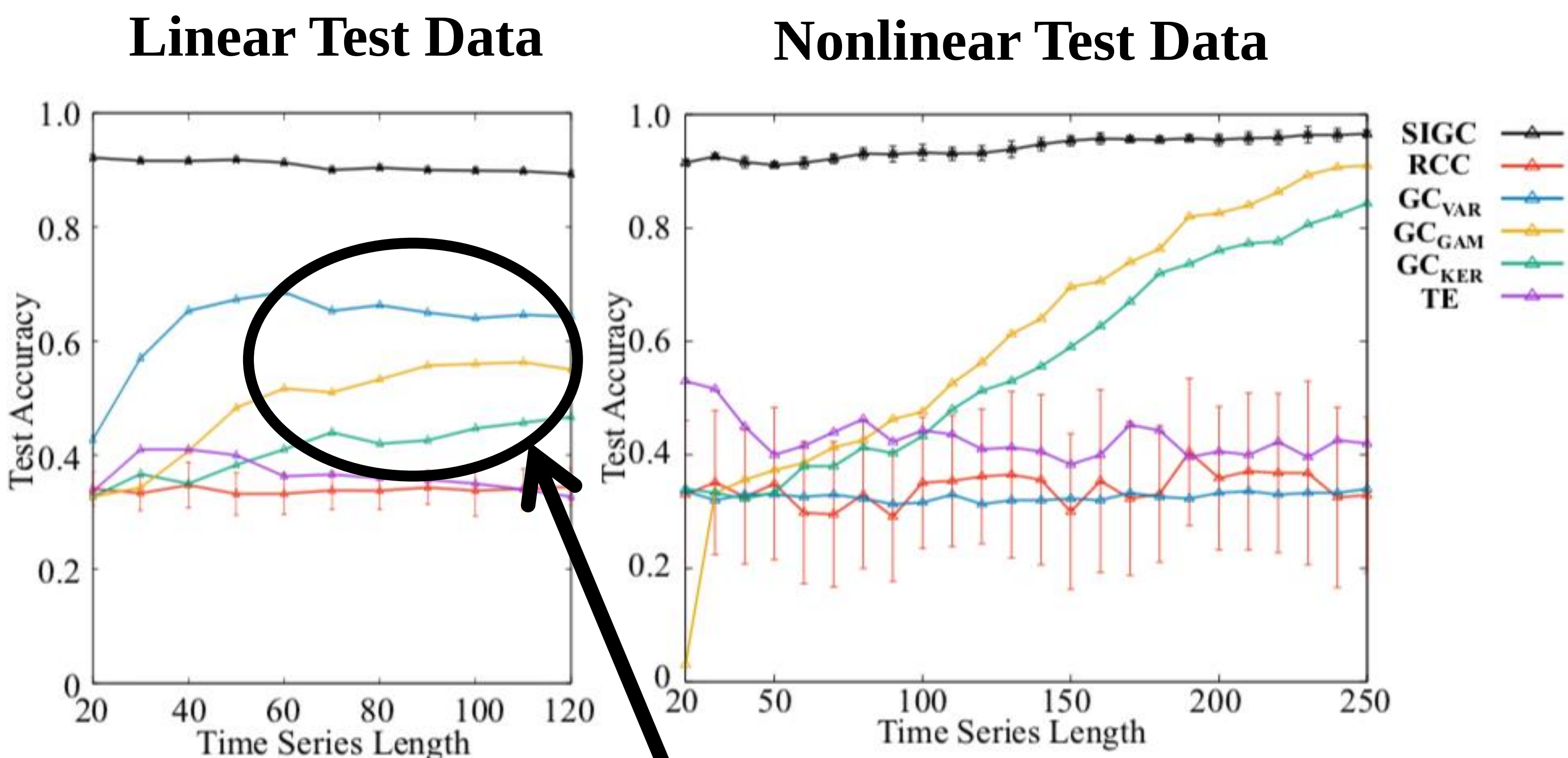
$$Y_t = -0.5 + \exp(-(X_{t-1} + X_{t-2})^2) + 0.7\cos(Y_{t-1}^2) + 0.3N_{Y_t}$$

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

Test accuracy
(横軸: 時系列の長さ, 縦軸: Test Accuracy)



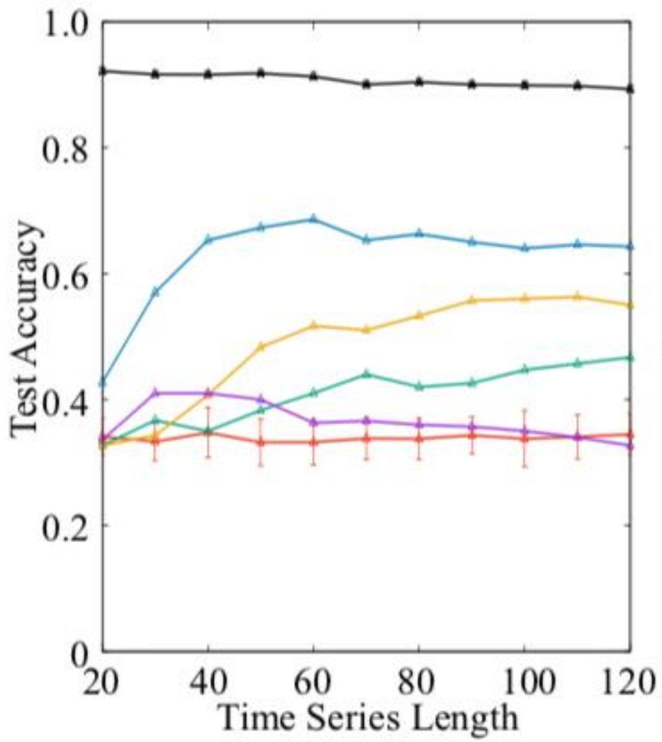
Test accuracy



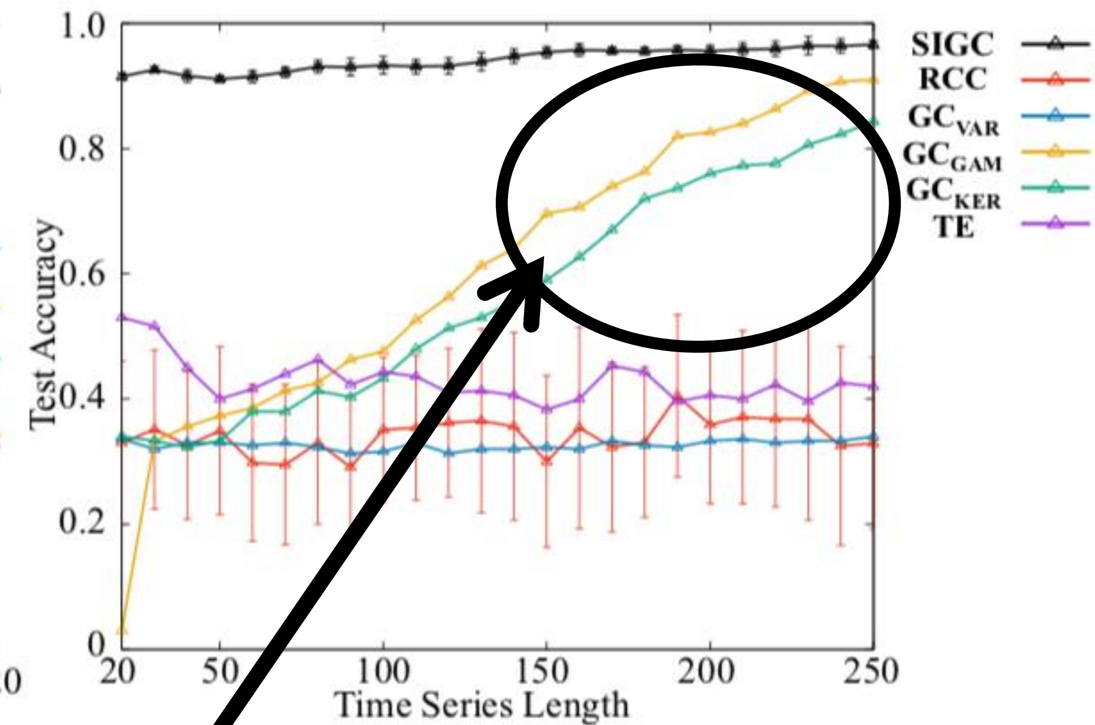
Existing Granger causality methods
Test accuracy strongly depends on the regression model

Test accuracy

Linear Test Data



Nonlinear Test Data

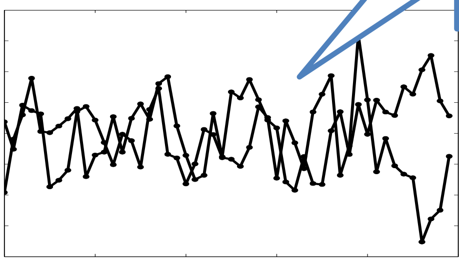


$GC_{KER} < GC_{GAM}$

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

Experiment 2: Real-world test data

Real-world
Test Data



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

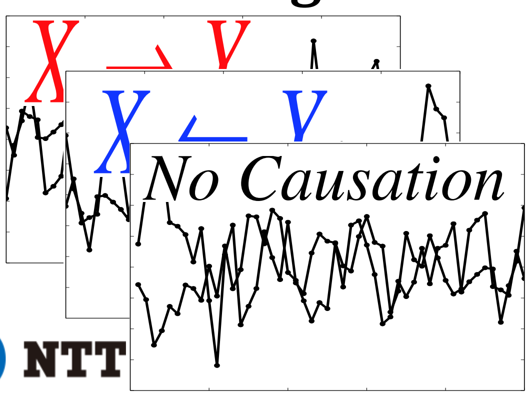
Classifier

$X \rightarrow Y$

$X \leftarrow Y$

No Causation

Synthetic
Training Data



…

True causal directions are given in literatures

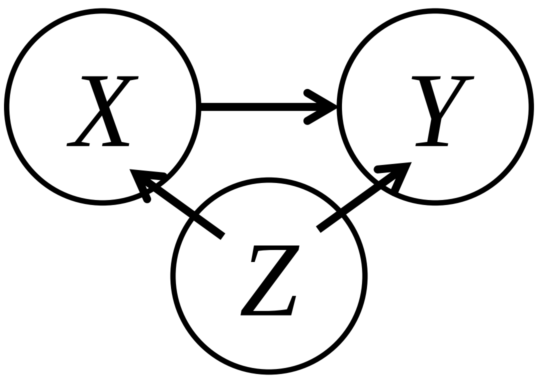
Test accuracy

	SIGC	RCC	GC_{VAR}	GC_{GAM}	GC_{KER}	TE
River Runoff ($T = 200$)	**0.958** (0.058)	0.399 (0.193)	0.684	0.406	0.155	0.485
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

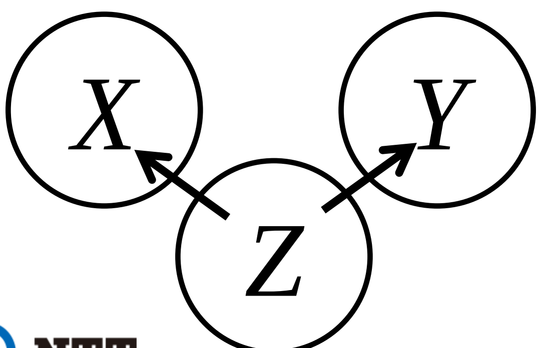
How can we extend proposed approach to multivariate time series?

Bivariate Methods do not work well

- With original Granger causality, we cannot distinguish the following trivariate case



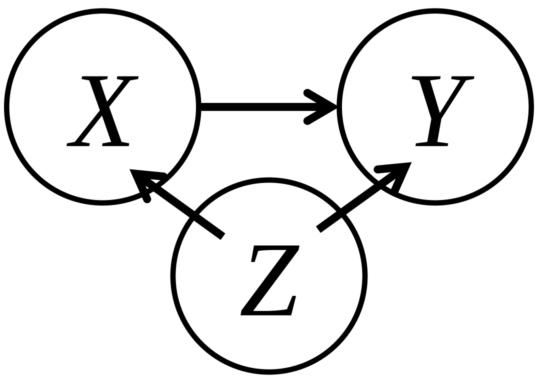
$$P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$$



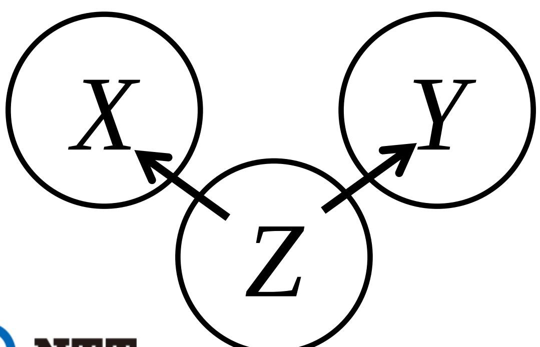
$$P(Y_{t+1}|S_X, S_Y) \neq P(Y_{t+1}|S_Y)$$

Granger causality definition for multivariate time series

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



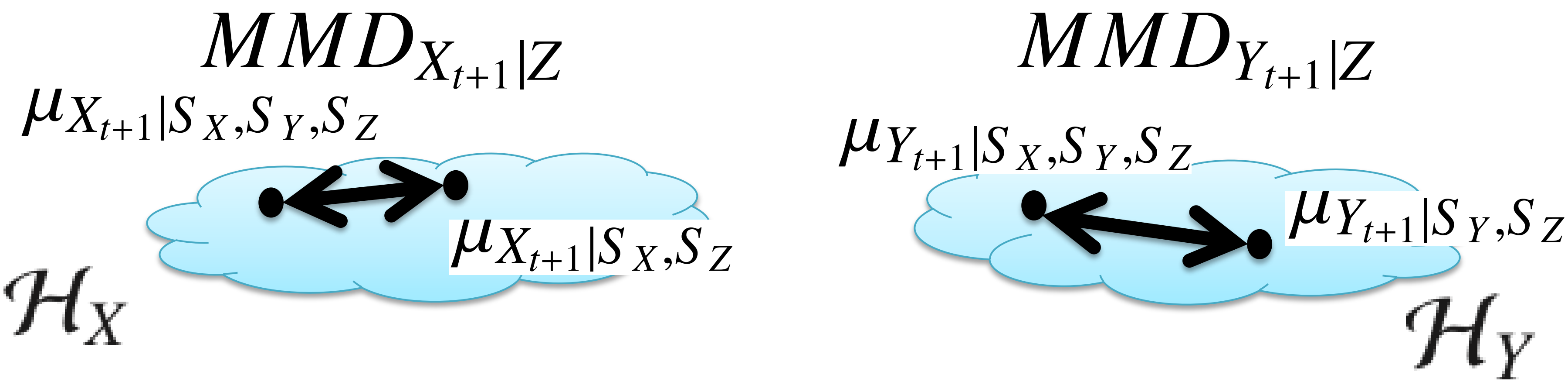
if $P(Y_{t+1} | S_X, S_Y, \underline{S_Z}) \neq P(Y_{t+1} | S_Y, \underline{S_Z})$



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

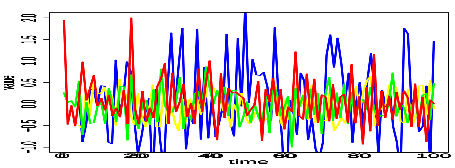


- By using additional MMDs, we formulate feature representation for multivariate time series

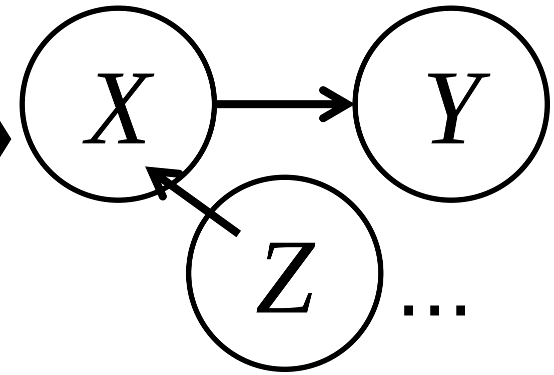
Experiment 3: Multivariate real-world data

<u>Real-world</u>
Test Data

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



Classifier



<u>Synthetic</u>
Training Data

<u>True causal directions are given in database</u>

Macro F1 score and micro F1 score

	$\mathbf{SIGC}_{tri}$	$\mathbf{GC}_{VAR}$	$\mathbf{GC}_{KER}$	$\mathbf{SIGC}_{bi}$	$\mathbf{GC}_{GAM}$	**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

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*