# Stream Mining Time-evolving Causality in Time Series

Naoki Chihara
SANKEN, Osaka University
naoki88@sanken.osaka-u.ac.jp

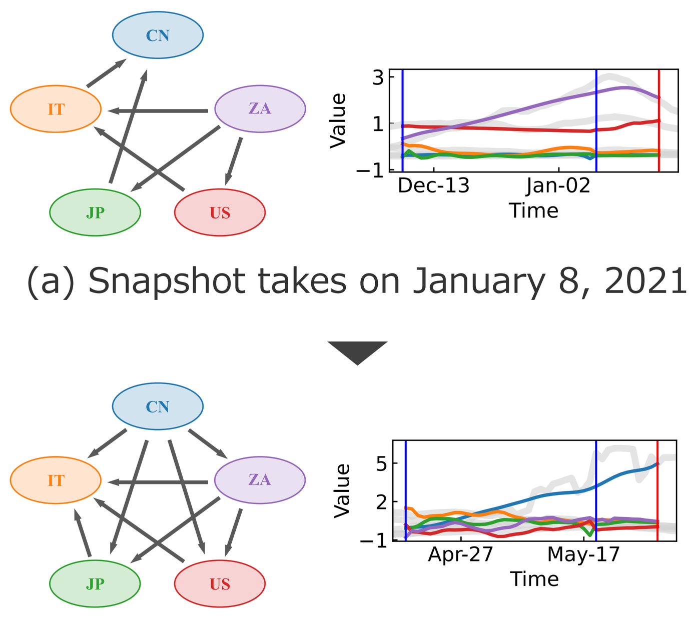
Yasuko Matsubara
SANKEN, Osaka University
yasuko@sanken.osaka-u.ac.jp

Ren Fujiwara
SANKEN, Osaka University
r-fujiwr88@sanken.osaka-u.ac.jp

Yasushi Sakurai
SANKEN, Osaka University
yasushi@sanken.osaka-u.ac.jp

## Motivation - Given: Multivariate data streams

e.g., Spread of infectious diseases, coronavirus (COVID-19)



(a) Snapshot takes on January 8, 2021

(b) Snapshot taken on May 19, 2022

### Challenges

How can we discover time-changing causal relationships?

**Given:** Multivariate Data stream i.e., $\boldsymbol{X} = \{\boldsymbol{x}(1), \ldots, \boldsymbol{x}(t_c), \ldots\}$

**Goal:** Achieve all of the followings

- **Find** distinct dynamical patterns / **regimes**
- **Discover** causal relationships, which changes over time / **time-evolving causality**
- **Forecast** an $l_s$-steps ahead future values

✨ **ModePlait: novel streamimg method**

## Proposed Model - ModePlait

**Key Concepts** - Our model is designed based on SEM

Exogenous variables evolve over time / **inherent signals**

$$\boxed{\boldsymbol{X}_{\text{sem}}} = \boxed{\boldsymbol{B}_{\text{sem}}}\boldsymbol{X}_{\text{sem}} + \boxed{\boldsymbol{E}_{\text{sem}}}$$

Observed variables — Causal adjacency matrix — Exogenous variables

### Main idea (P1): Latent temporal dynamics

😢 Each inherent signal $e_{(i)}(t)$ is only a single dimension

⇒ superposition of computed basis vectors (i.e., **modes**)

Latent activities — Eigenvalues matrix

$$\boxed{\boldsymbol{s}_{(i)}(t+1)} = \boxed{\boldsymbol{\Lambda}_{(i)}}\boldsymbol{s}_{(i)}(t) \;:\; k_i\text{-dimensional space}$$

Augmentaion

$$\boxed{e_{(i)}(t)} = \boxed{g^{-1}}(\boxed{\boldsymbol{\Phi}_{(i)}}\boldsymbol{s}_{(i)}(t)) \;:\; \text{Projection } (\mathbb{C}^{k_i} \to \mathbb{R})$$

$i$-th univariate inherent signal — Time-delay embedding — Modes

$g(e_{(i)}(t)) := (e_{(i)}(t), e_{(i)}(t-1), \ldots, e_{(i)}(t-h+1))$, $k_i$: # of modes

$\mathcal{D}_{(i)} = \{\boldsymbol{\Phi}_{(i)}, \boldsymbol{\Lambda}_{(i)}\}$ / **self-dynamics factor set**

### Main idea (P2): Dynamical patterns

Describe distinct dynamical pattern (i.e., **regime**)

$$\boxed{\begin{aligned} \boldsymbol{s}_{(i)}(t+1) &= \boldsymbol{\Lambda}_{(i)}\boldsymbol{s}_{(i)}(t) \quad (1 \le i \le d) \\ e_{(i)}(t) &= g^{-1}(\boldsymbol{\Phi}_{(i)}\boldsymbol{s}_{(i)}(t)) \quad (1 \le i \le d) \end{aligned}}$$

$$\boxed{\boldsymbol{v}(t)} = \boxed{\boldsymbol{W}^{-1}}\boldsymbol{e}(t)$$

Estimated value — Mixing matrix — (P1) Collection of $d$ self-dynamic factor sets

$\theta = \{\boldsymbol{W}, \mathcal{D}_{(1)}, \ldots, \mathcal{D}_{(d)}\}$ / **regime**, $\Theta = \{\theta^1, \ldots, \theta^R\}$ / **regime set**

$\mathcal{B} = \{\boldsymbol{B}^1, \ldots, \boldsymbol{B}^R\}$ / **time-evolving causality**

https://www.linkedin.com/in/naoki-chihara-0a35a827a/

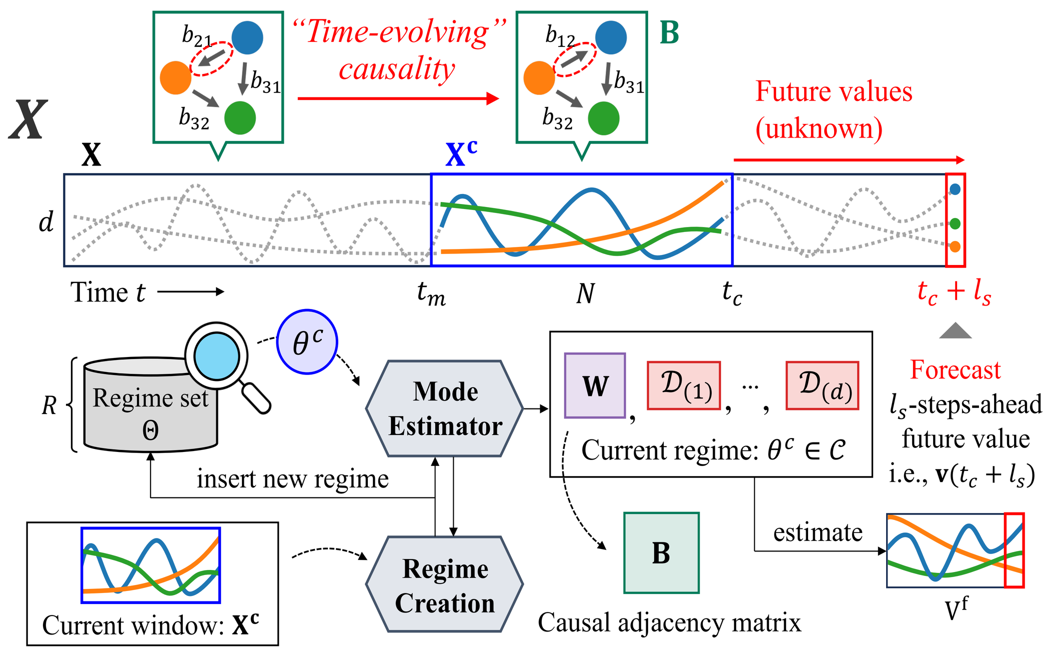
## Optimization algorithm

**Given:**

- Multivariate data Stream $\boldsymbol{X}$

**Estimate:**

- Full parameter set $\mathcal{F} = \{\Theta, \Omega\}$, $\Omega$: update param
- Model candidate $\mathcal{C} = \{\theta^c, \omega^c, S_{en}^c\}$
- Time evolving causality $\mathcal{B} = \{\boldsymbol{B}^1, \ldots, \boldsymbol{B}^R\}$, $R$: # of regimes
- $l_s$-steps ahead future value $\boldsymbol{v}(t_c + l_s)$, $t_c$: current time point

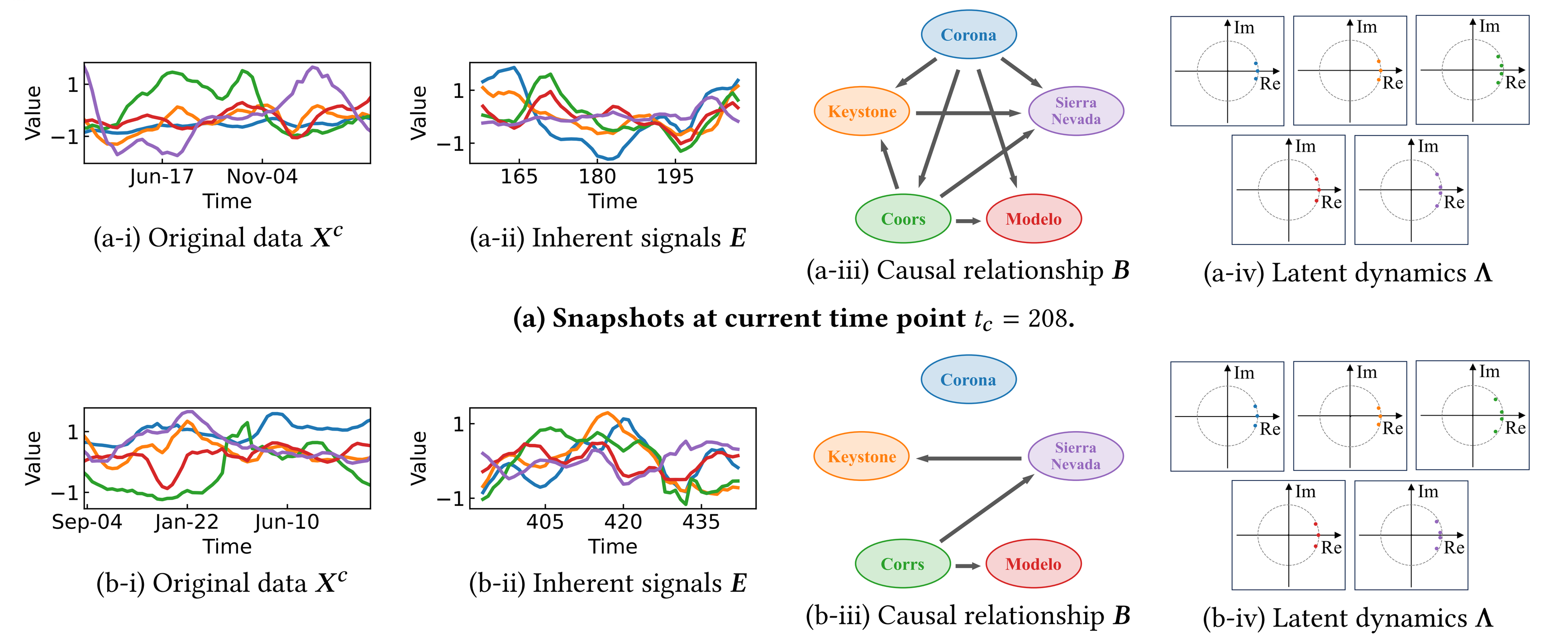



## Experiments - Answer the essential questions

**Datasets**: we used the following four real datasets

*(#1) covid19*: was obtained from Google COVID-19 Open Data [9].
*(#2) web-search*: consists of web-search counts on Google [10].
*(#3) chicken-dance*, *(#4) exercise*: were obtained from the CMU motion capture database [4].

### Q1. Effectiveness - web-click activity stream



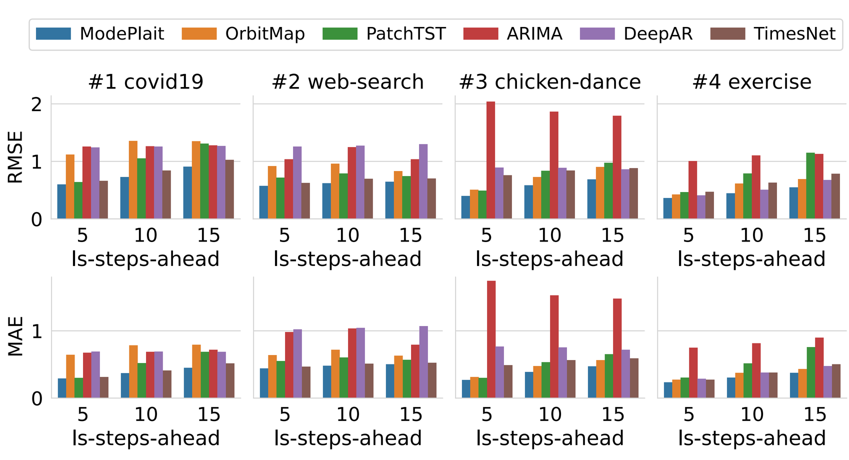
(a-i) Original data $\boldsymbol{X}^c$ (a-ii) Inherent signals $\boldsymbol{E}$ (a-iii) Causal relationship $\boldsymbol{B}$ (a-iv) Latent dynamics $\boldsymbol{\Lambda}$

**(a) Snapshots at current time point $t_c = 208$.**

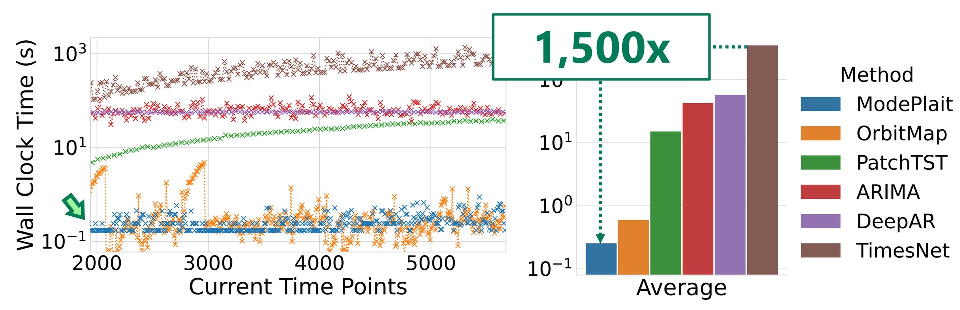
(b-i) Original data $\boldsymbol{X}^c$ (b-ii) Inherent signals $\boldsymbol{E}$ (b-iii) Causal relationship $\boldsymbol{B}$ (b-iv) Latent dynamics $\boldsymbol{\Lambda}$

**(b) Snapshots at current time point $t_c = 443$.**

### Q2. Accuracy




### Q3. Scalability




Time vs. data length $t_c$ — Average time

🔥 **ModePlait outperforms its competitors**

## Conclusion - ModePlait has following properties:

**Effective**: it discovers time-evolving causality
**General**: it is adaptable to various real-world datasets
**Scalable**: it does not depend on stream length

## Future work - Enhancing our proposed model:

**Causal discovery evaluation:** we plan to use synthetic datasets generated from Erdös-Rényi (ER) model [Erdös and Rényi 1960] for quantitative evaluation.

**Multi-task optimization:** we optimize multiple tasks (i.e., causal discovery and time series forecasting) while considering the mutual dependencies between them.

**Other downstream tasks:** we detect anomalous activities based on time-evolving causality in data streams.