Identification of Causal Dependencies in Multivariate Time Series

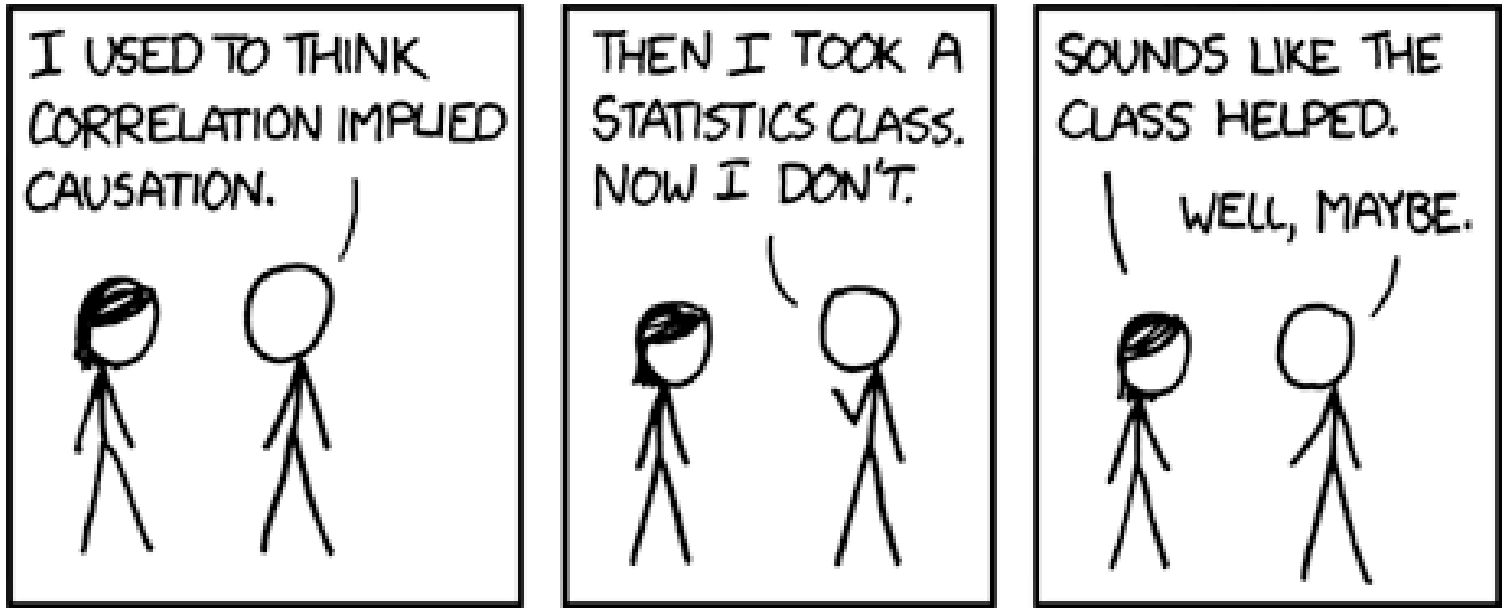
Sujoy Roy Chowdhury (Principal Data Scientist),
Serene Banerjee (Master Researcher),
Ranjani H G (Principal Data Scientist),
Chaitanya Kapoor (Research Intern)

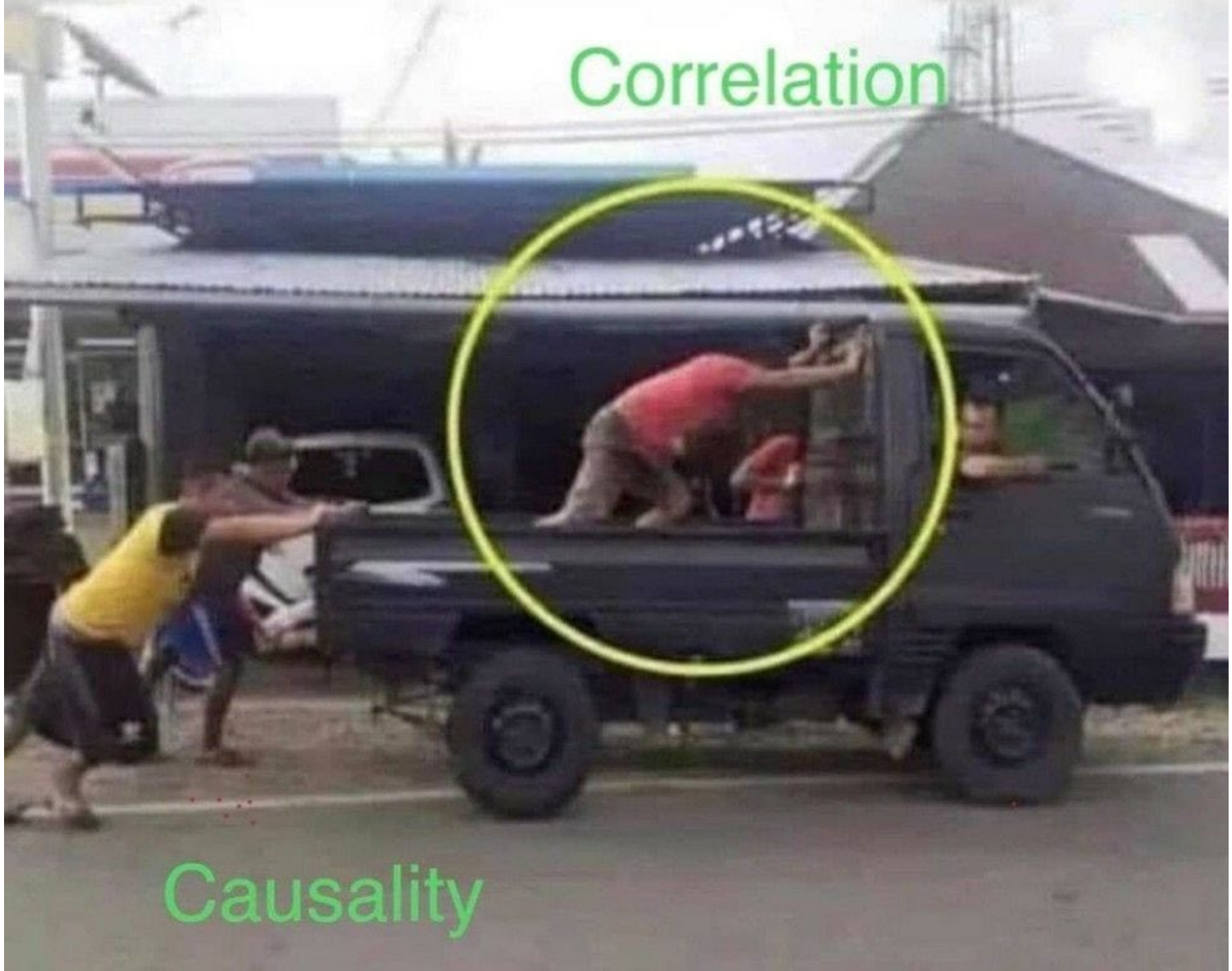
Ericsson, Bangalore

Agenda

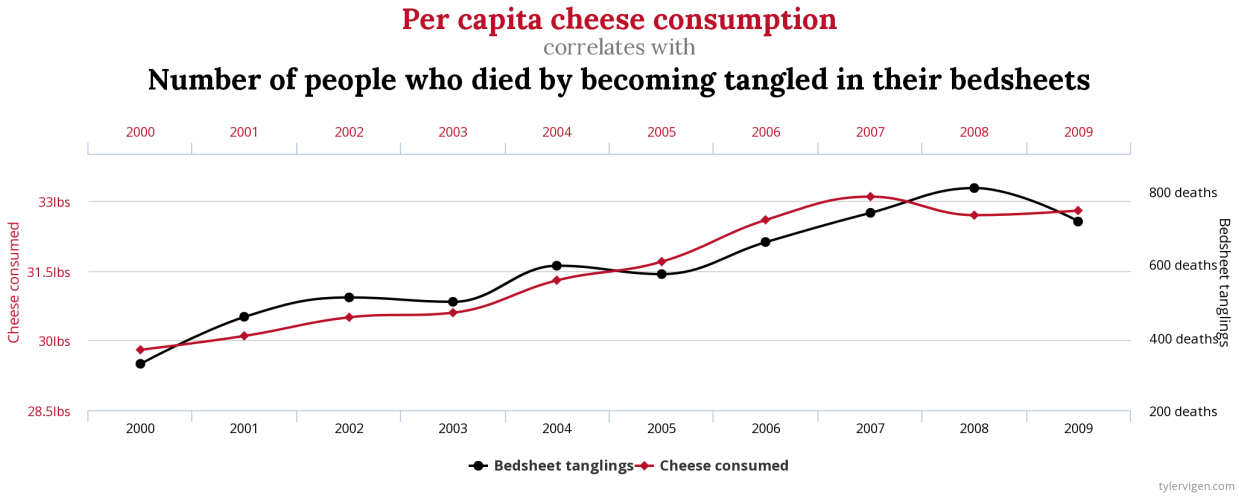
- Motivation for Causal Analysis
- Examples of Causal Analysis Use Cases in Telecom
- Double Pendulum Dataset
- Causal Methodologies
  - Granger Causality
  - Non-linear Granger Causality
  - PCMCI
  - Convergent Cross Mapping
  - Temporal Causal Discovery Framework
- Notebook walkthrough
- Comments
- Q&A

Motivation for Causal Analysis

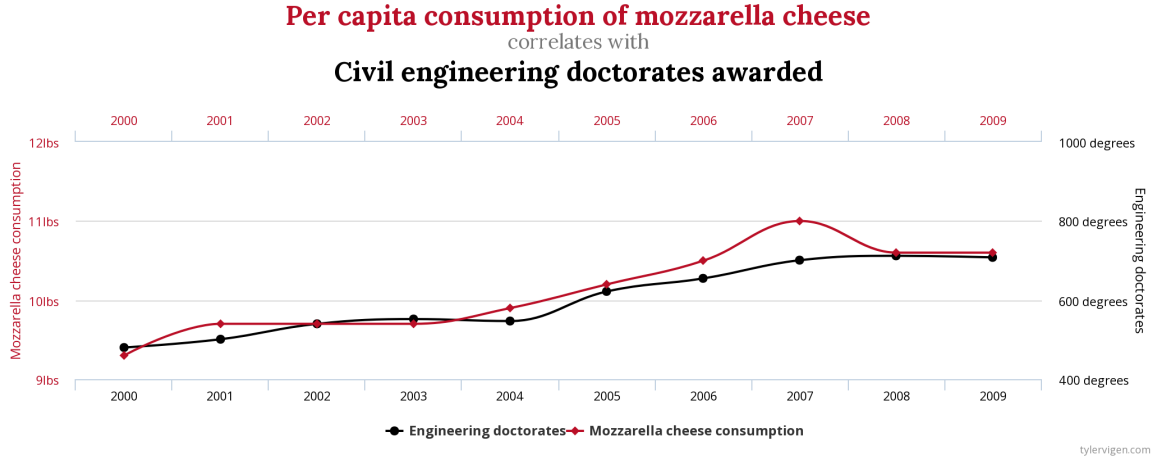

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.


Correlation
Causality

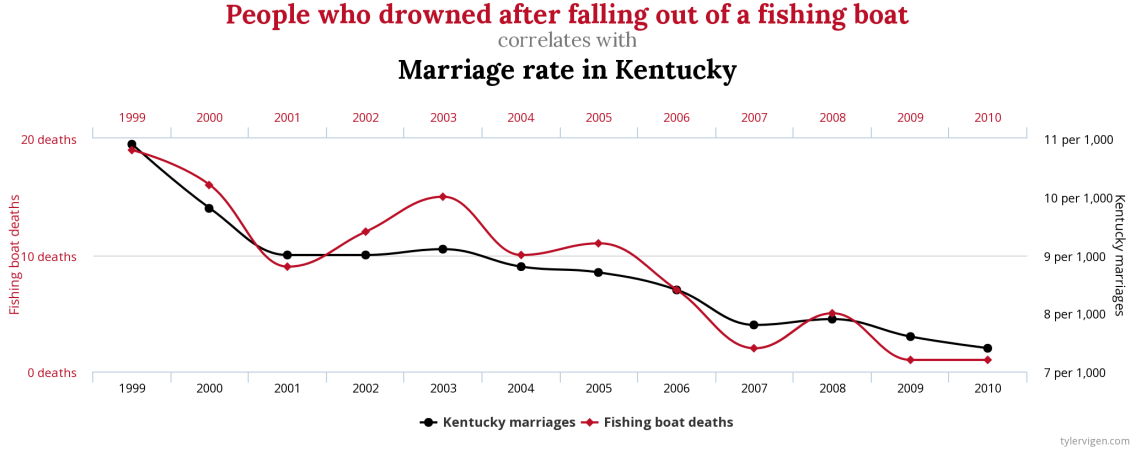
Correlation can be spurious



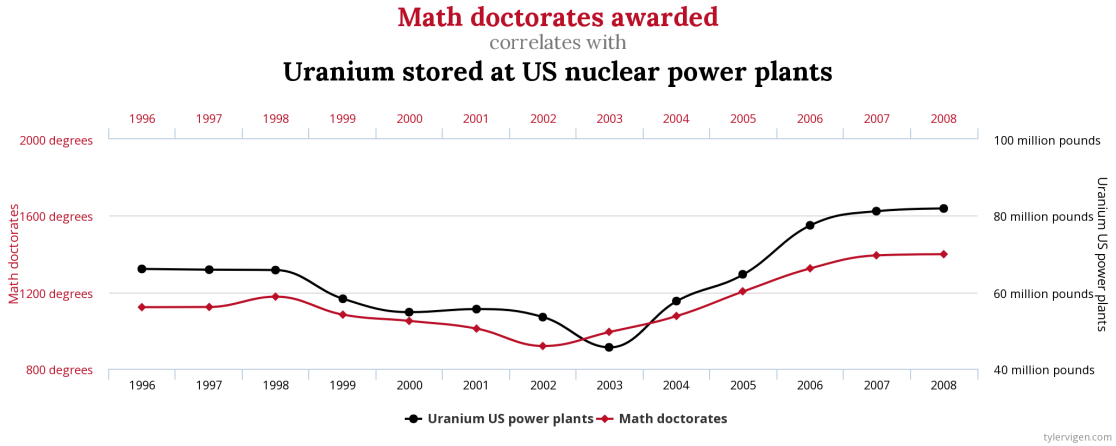
$\rho = 0.95$



$\rho = 0.96$



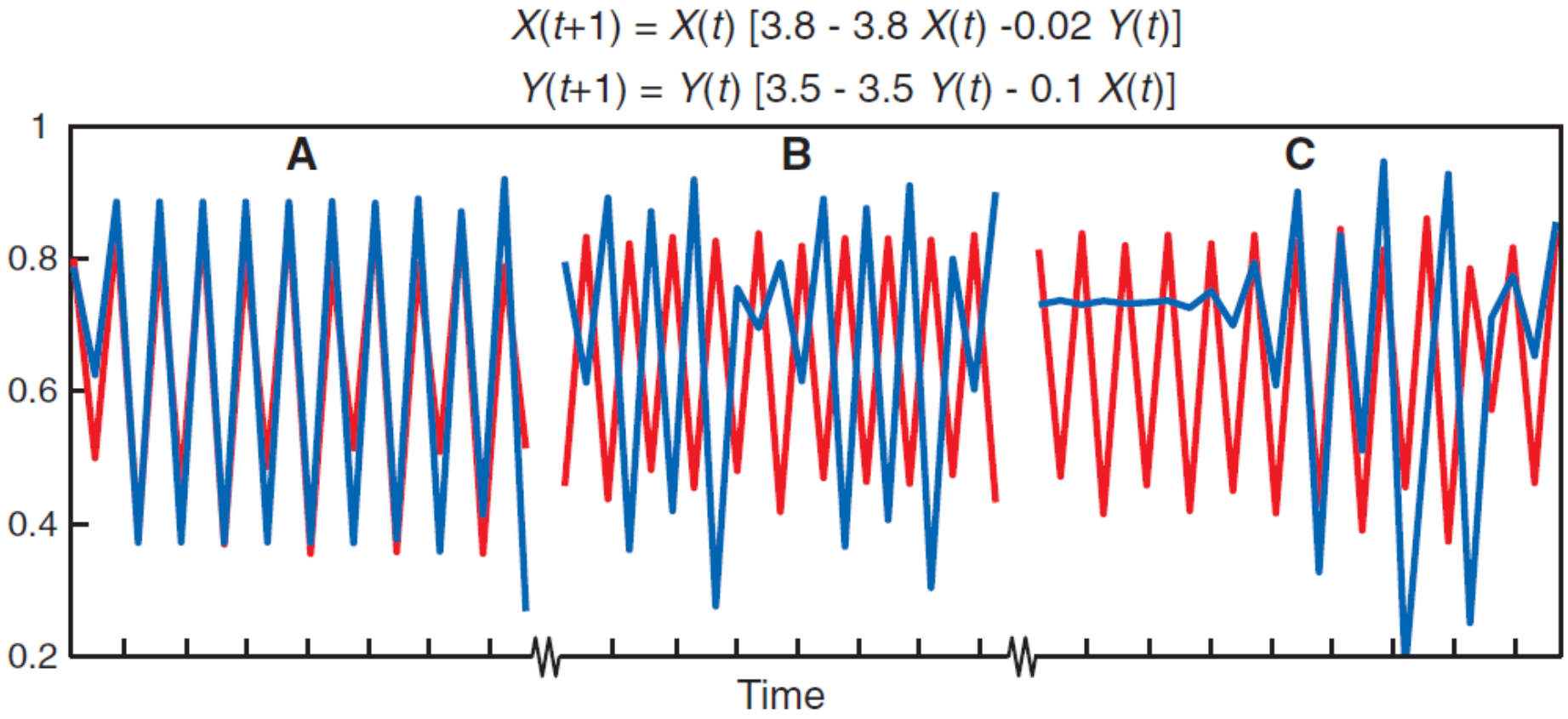
$\rho = 0.95$



$\rho = 0.95$

Source:-[Spurious Correlations (tylervigen.com)]

Correlation does not imply Causation but also ....



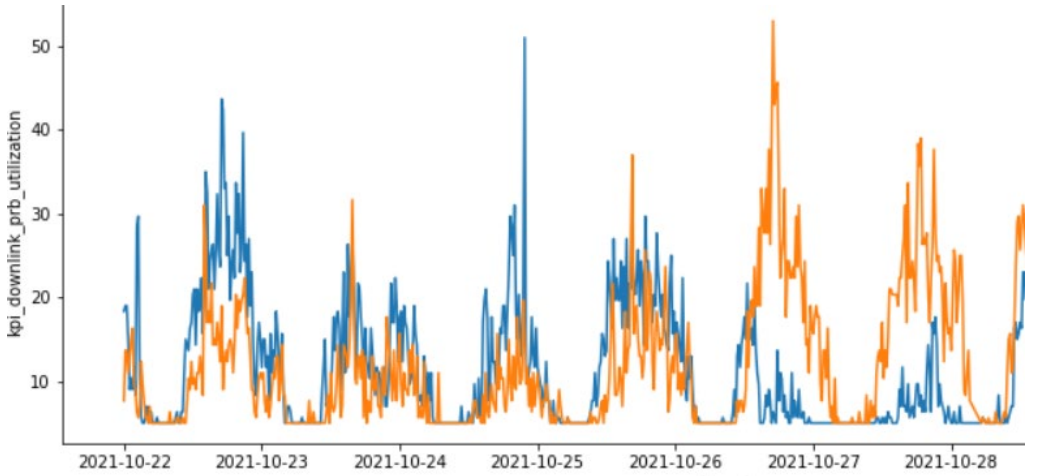
Source: Sugihara, George, et al. "Detecting causality in complex ecosystems." *science* 338.6106 (2012): 496-500.

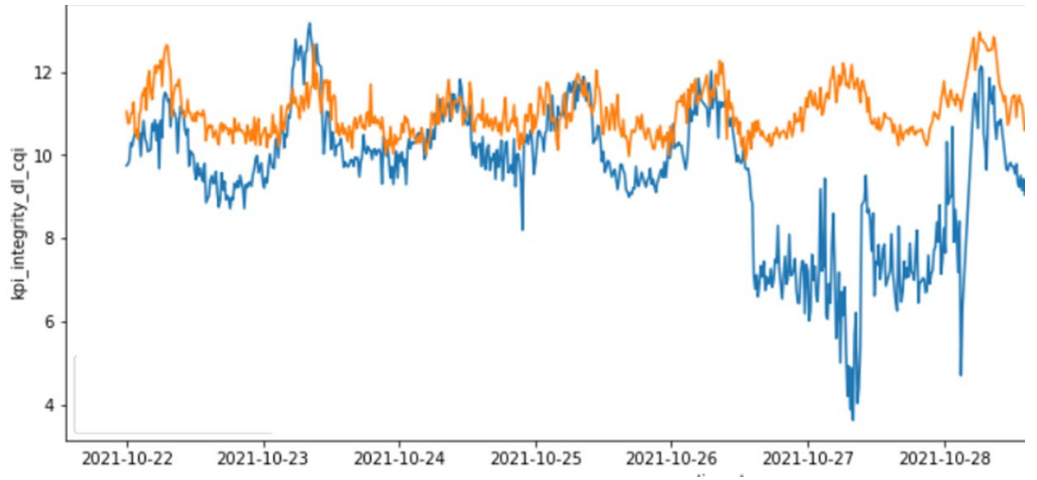
Why do we need causal

- Most AI / ML models are correlational and do not have any causal direction
- Causal is needed to attribute reason and identify actions which can result in change
- However causal and explainable models are not necessarily the same
  - Causal is driven based on underlying system
  - Explainable is driven based on finding what aspects helped a model choose the decision it took
- Challenges of causal
  - Ground truth is often not available

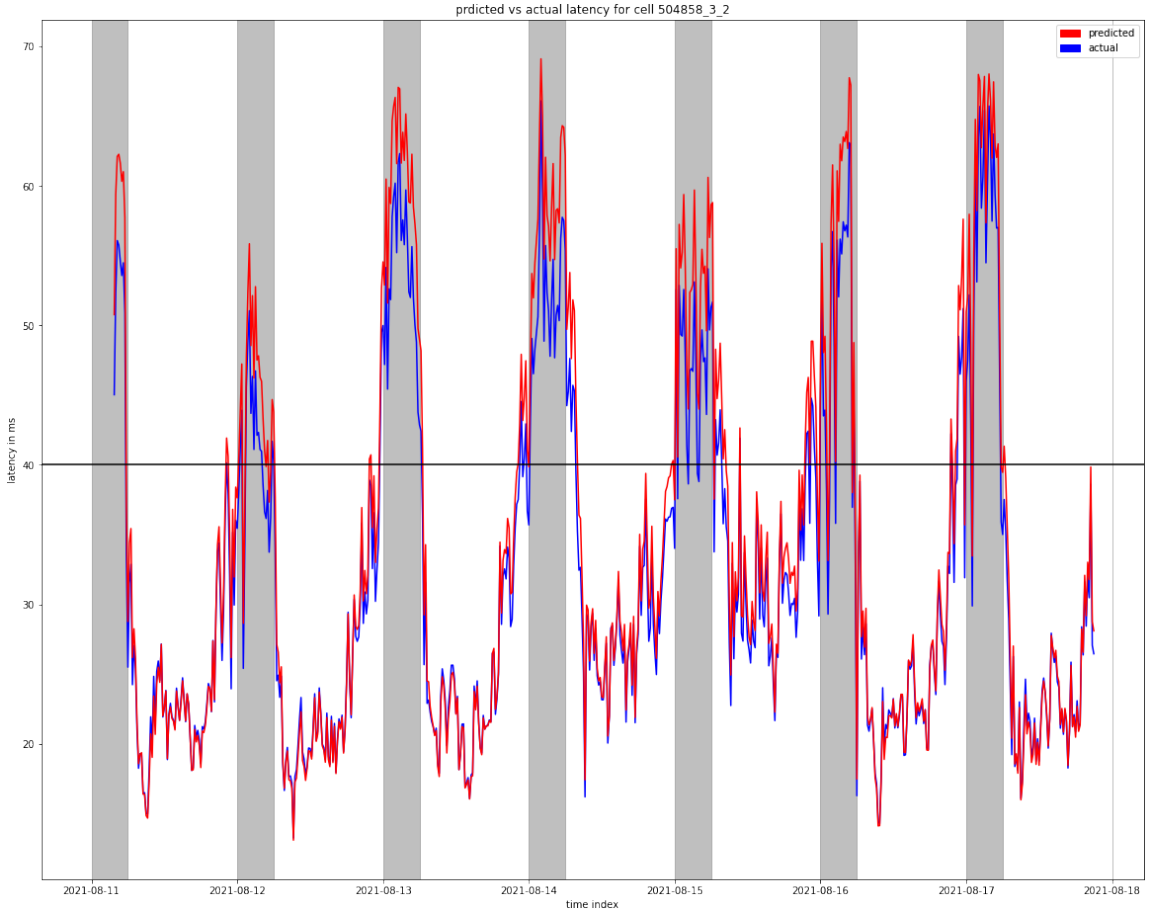
Telecom Networks KPI prediction

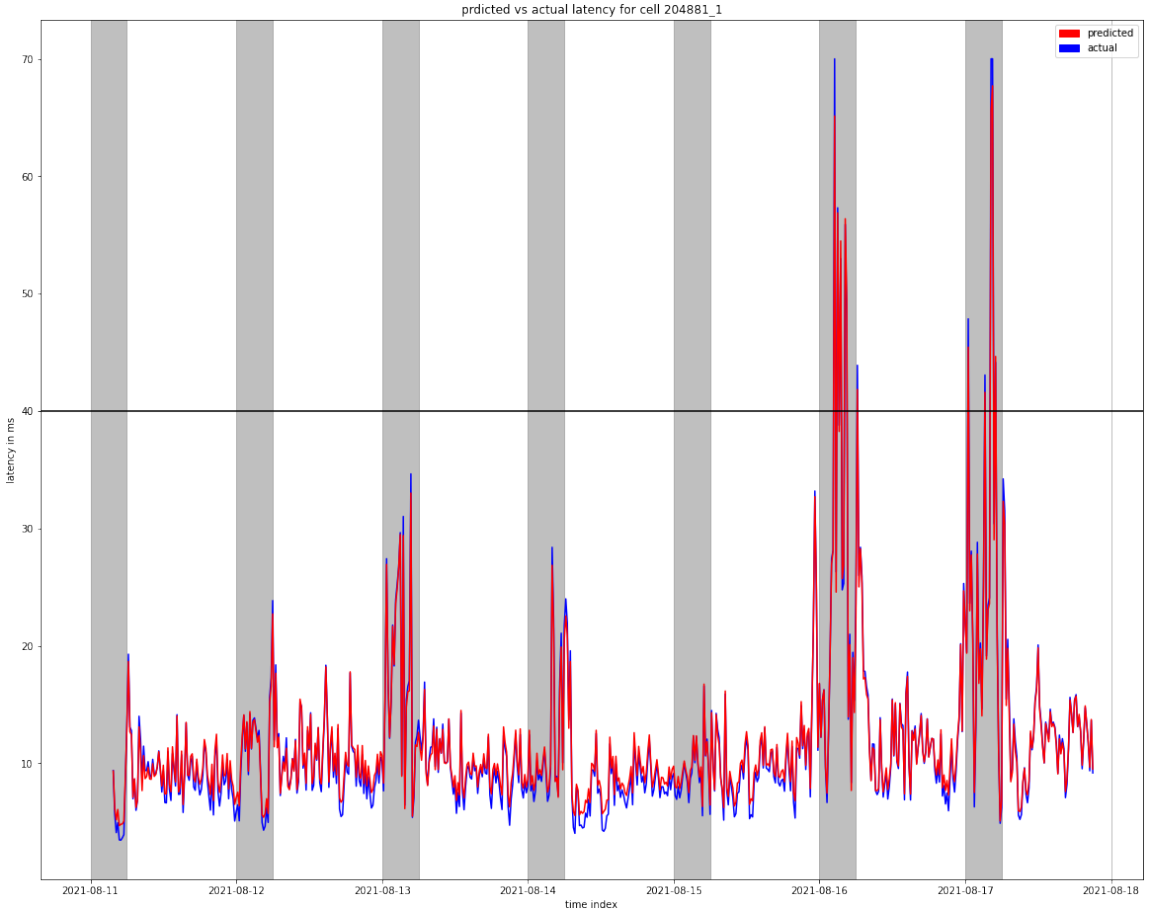
- Telecom networks collect multiple counters
- Counters are used to measure different KPIs
- Right below are examples of two counters (simulated)
- Specialists in NOCs monitor the KPIs to take corrective action
- There is lot of research on prediction of network KPIs from other KPIs
- However none of these models take an approach towards identifying causal direction



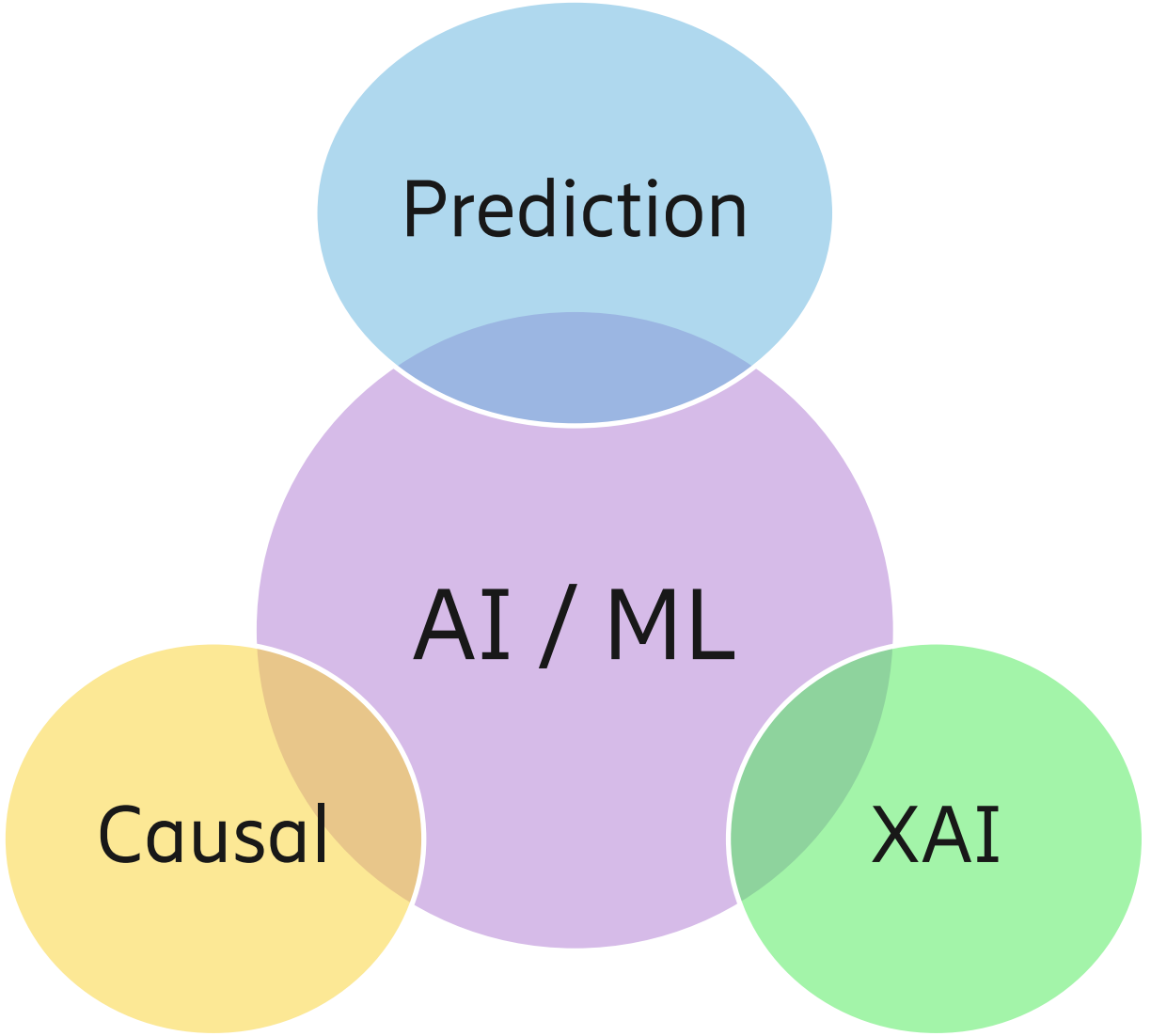


Example Prediction





Complimentary Views



The Causal Landscape

Causal Landscape

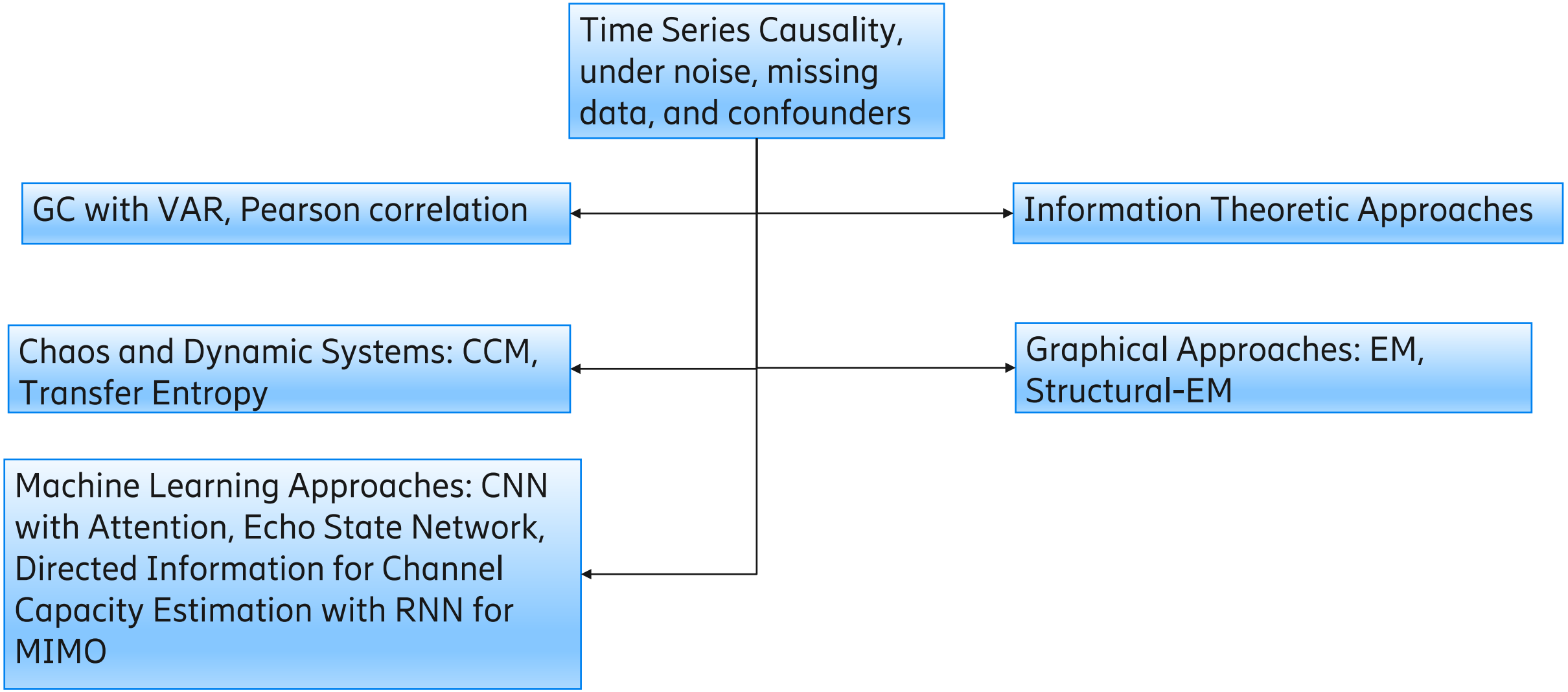
— Do-why Analysis

— Causal RL

— Time Series Causality ...

Causal Methodologies

Taxonomy of Time-Series Causal Discovery



https://github.com/ckassaad/causal_discovery_for_time_series

Example Use Cases in Telecom for Causal Modelling

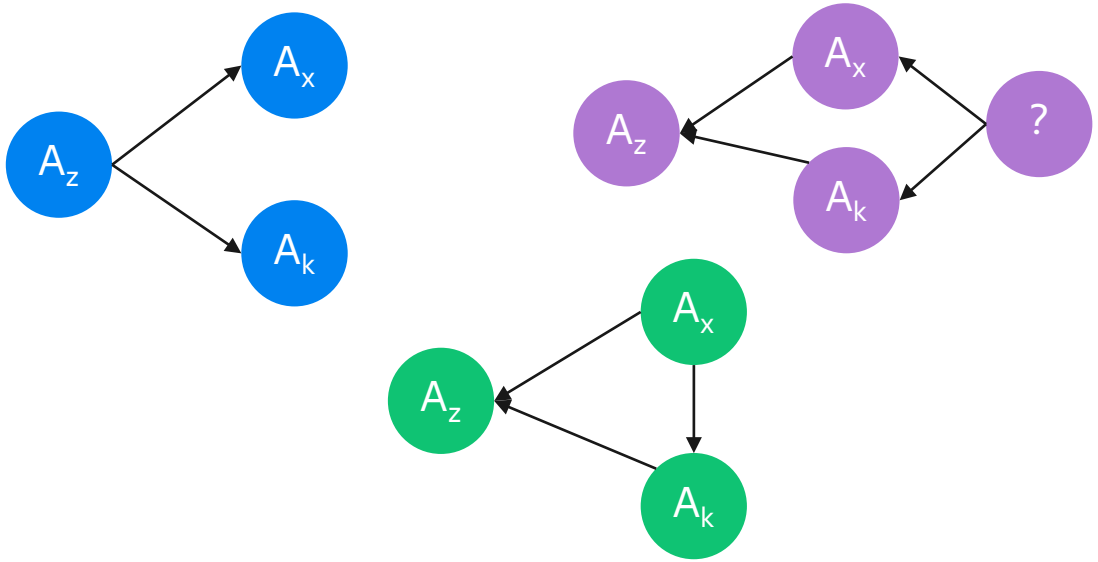
Example Causal Problem

- Identify causal relationships for anomalous events
- Given a set of time series A, how to obtain a graph representing the degree of influence between elements of A ?
  - Attribute weights to different edges;
  - Weights can be calculated as conditional probabilities.

Example:

A_z represents QoE for YouTube sessions in Bangalore.
A_x represents throughput for users at Bangalore.
A_k represents accessibility for users at Bangalore.

Which graph represents the relationship between the different time series ?



Hardware Faults Detection

Dedicated Networks

Passive Intermodulation (PIM) detection

Causally aware RL Sample Efficient Training

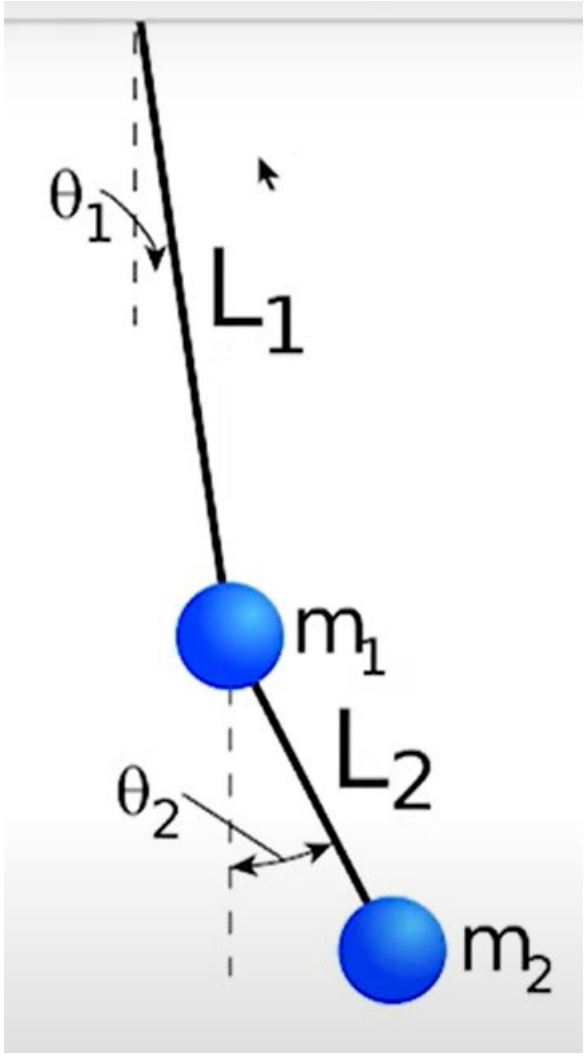
Causally-aware domain adaptation for scenes, Radio, and multimodal signals

Causally-aware channel state estimation compression

Double Pendulum Dataset

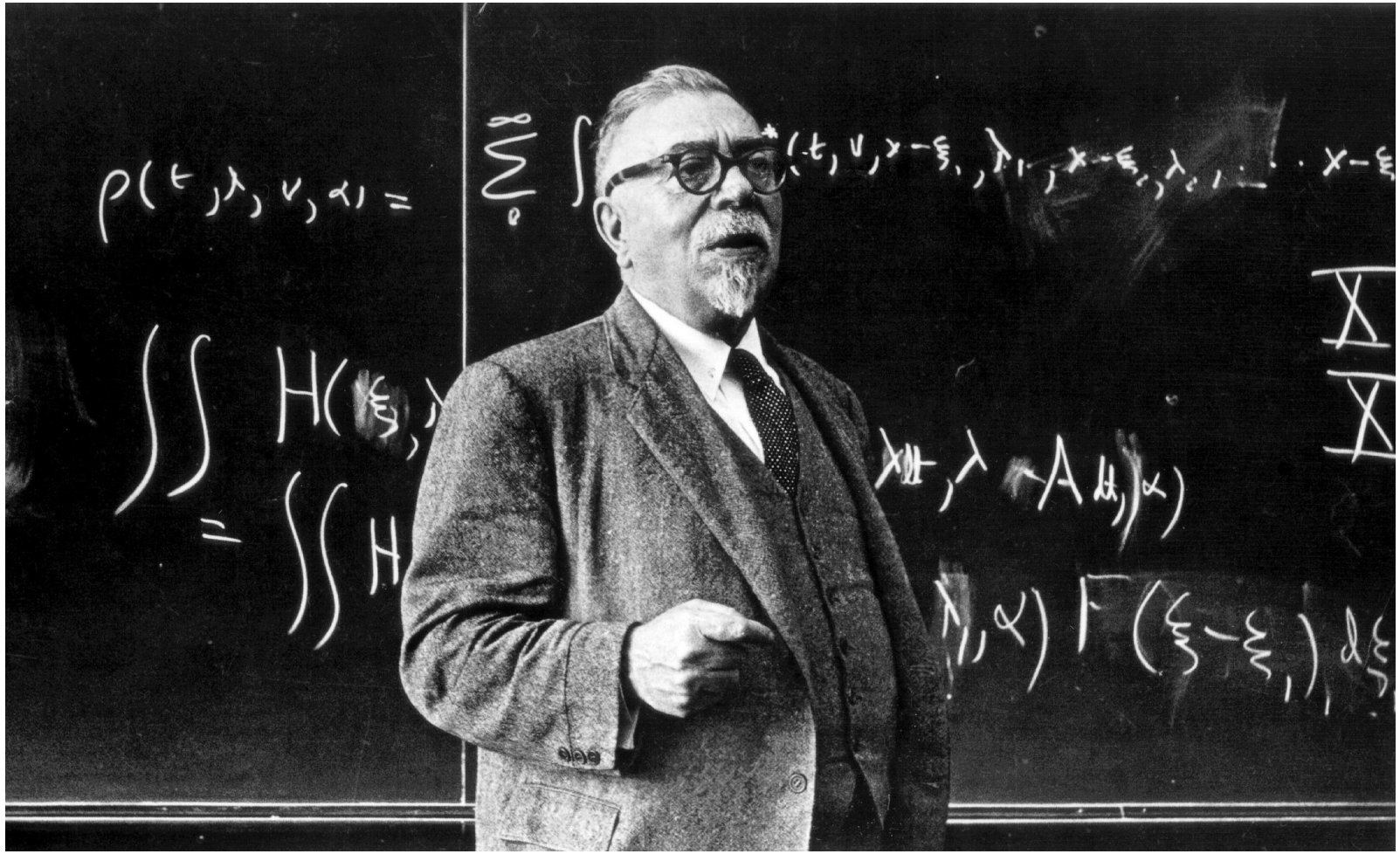
Double Pendulum

- The angles depend on time
- The lengths and mass are fixed
- (x^1, y^1) and (x^2, y^2) are the cartesian coordinates of the two pendulums
- Bob position and velocities are considered as 8 time series in total



Granger Causality

Norbert Weiner (Nov. 26, 1894 – Mar. 18, 1964)



Insights

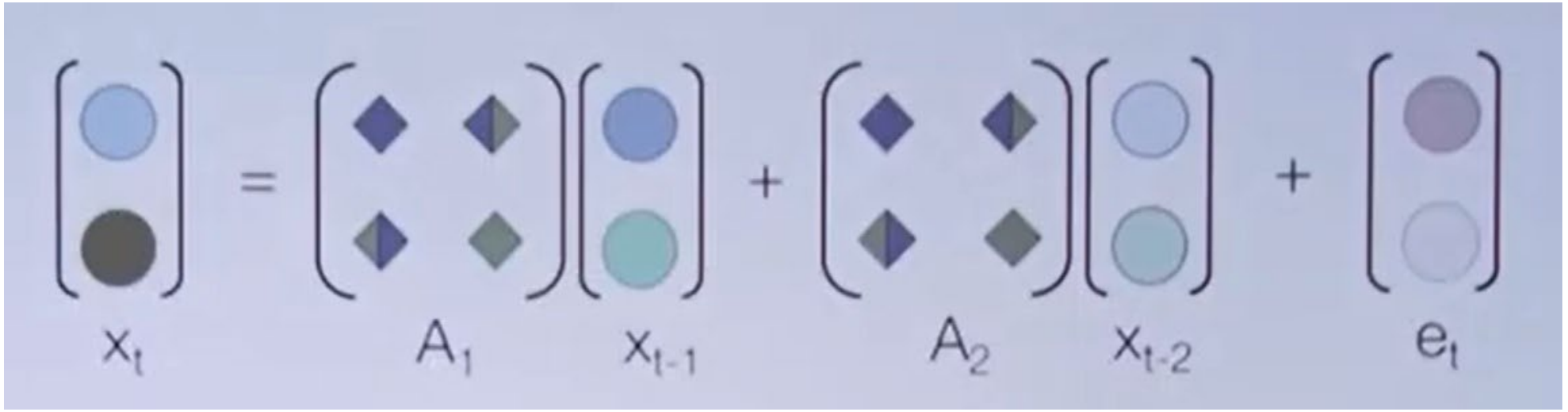
- Cause ***precedes*** the effect
- Multiple causation
- (Hidden) confounders
- Causes carry unique information to ***predict*** the effect

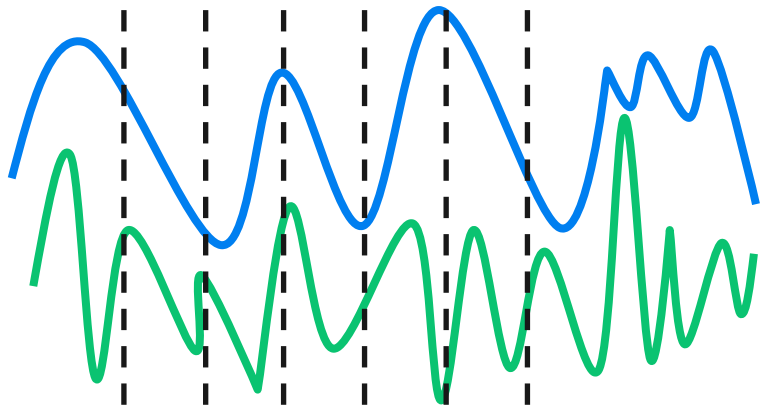
Granger – Implement Wiener's Causality Insights

1. Collect timeseries data
2. Use all the data to predict the future of one variable (X)
3. Repeat Step 2, but remove one variable (Y)
4. Compare error terms from Step 2 and 3
5. If error term is bigger in Step 3, then we conclude Y Granger causes X

Prediction by lagged vector auto regressive model (VAR)

Granger Causality – Linear Model

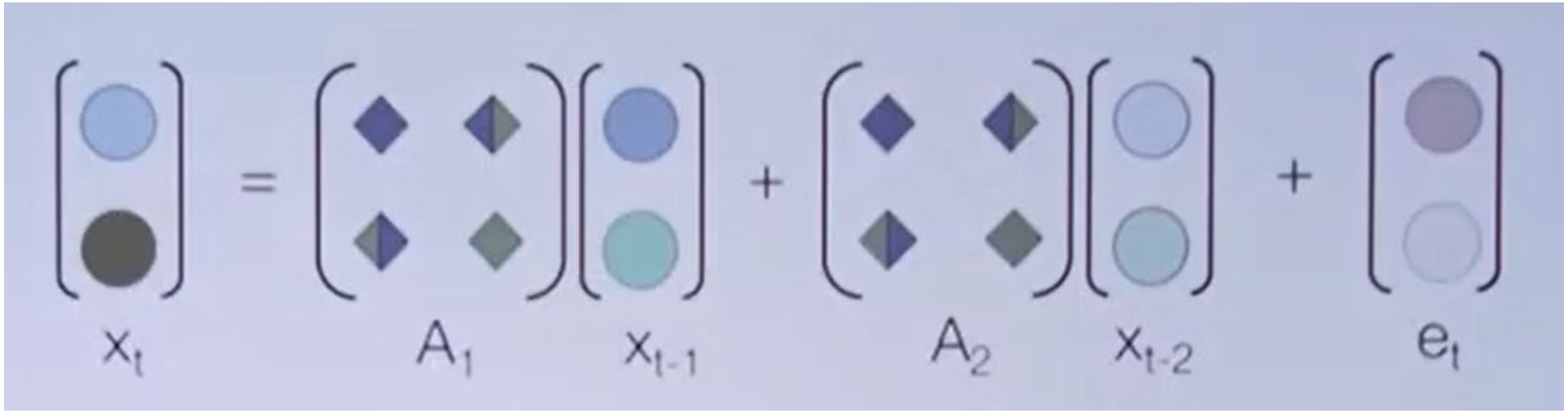




$$x^t = \sum_{k=1}^{K} A^k x^{t-k} + e^t$$

Series *i* does not cause Series *j*, iff $A^{ij,\,k}=0,\ \forall k$

Granger Causality – Linear Model



Explain data and encourage (structured) 0s:

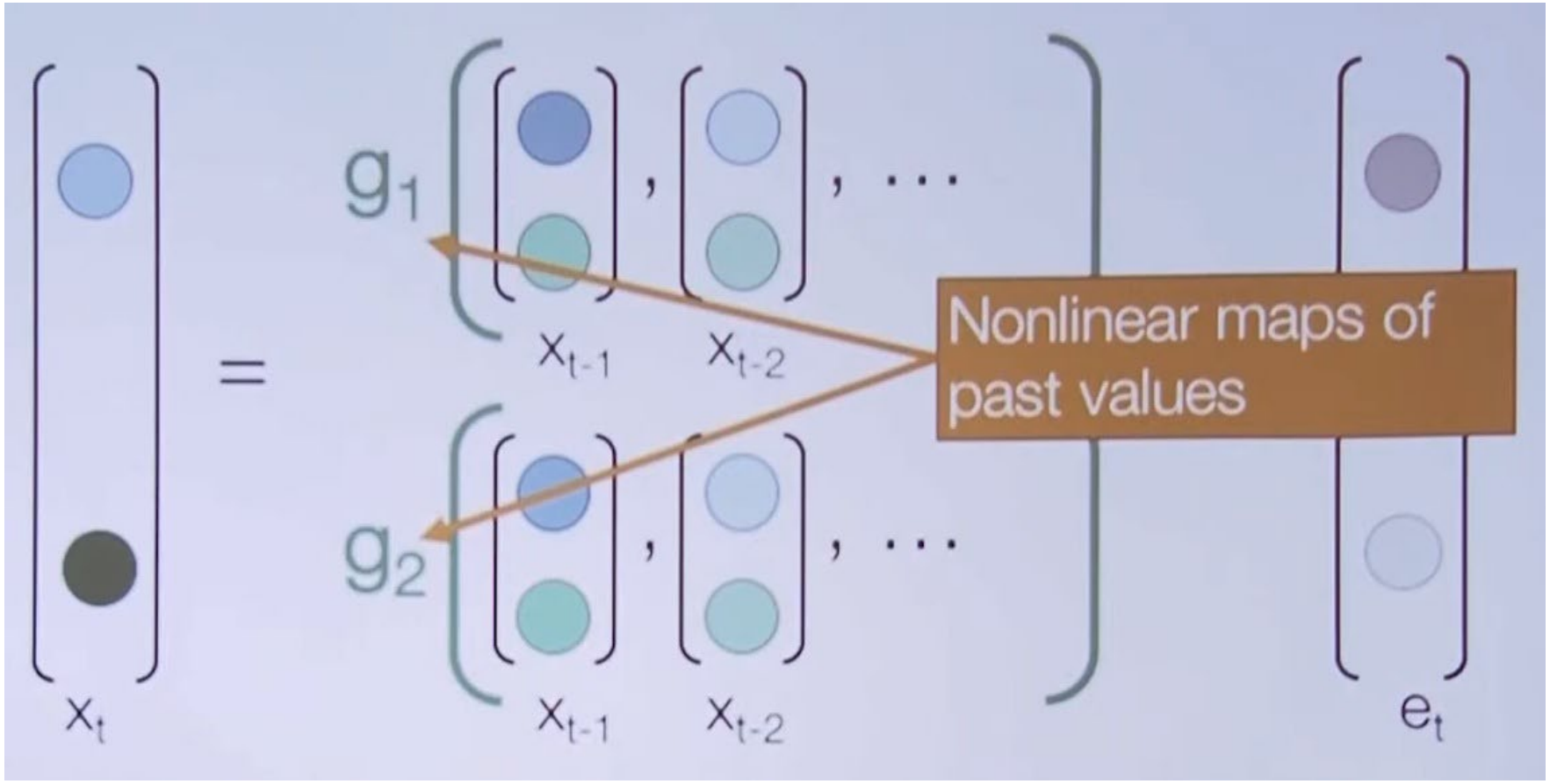
max $\{A^1, \ldots, A^K\}$ loglike$(x^1, \ldots, x^T; A^1, \ldots, A^K)$ - $\lambda \sum^{ij}$ penatly$(A^{ji,1}, \ldots, A^{ji,K} = 0)$

Minimize reconstruction error and apply group lasso penalty:

$$\min_{A^1 \ldots A^k} \sum_{t=K}^{T} (x^t - \sum_{k=1}^{K} A^k x^{t-k})^2 + \lambda \sum_{ij} \left|\left|(A^{ji,1}, \ldots . A^{ji,K})\right|\right|_2$$

Non-linear Granger Causality

Modeling Non-linearity



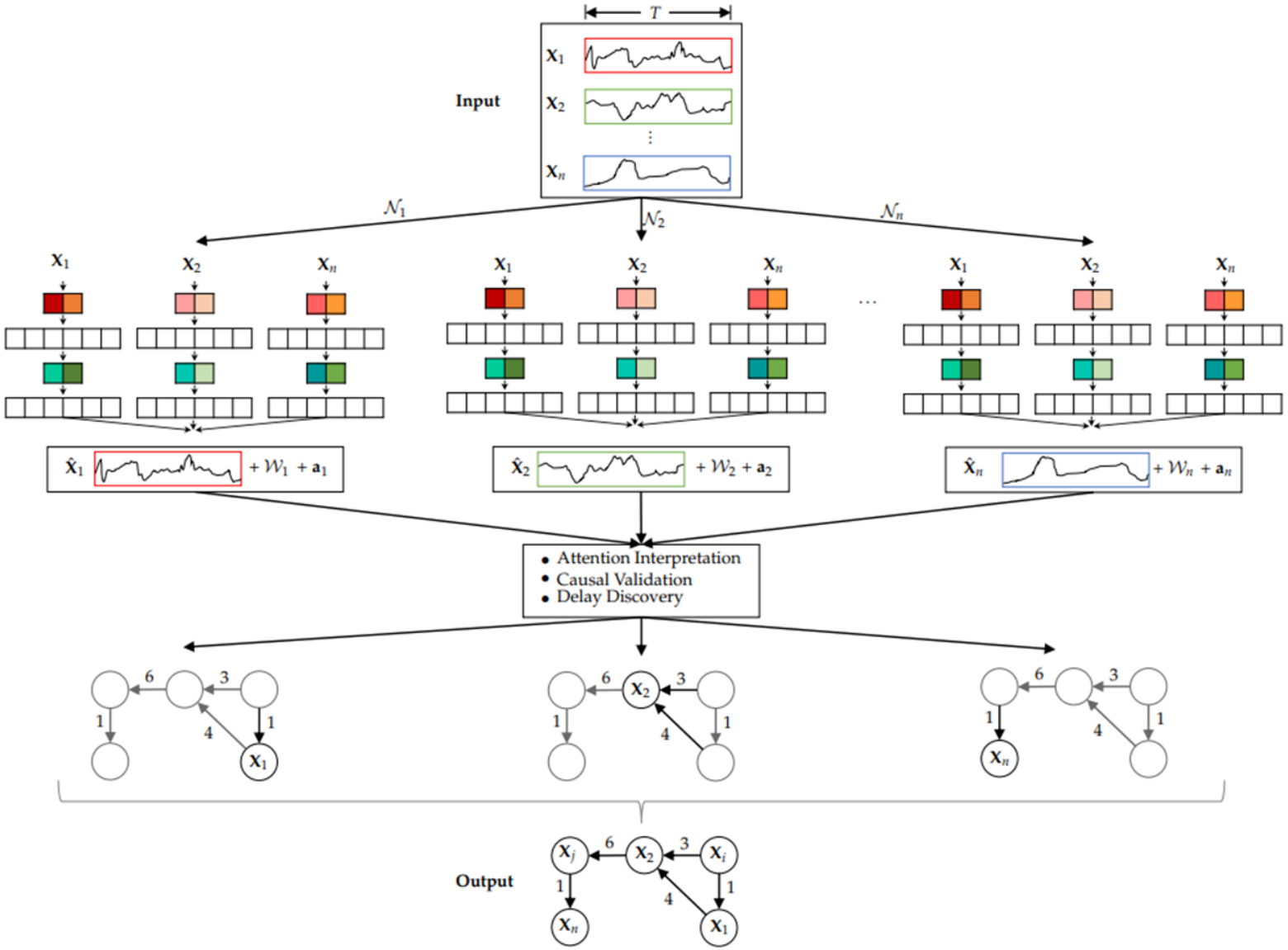
- Series *i* does not cause Series *j*, iff the non-linear value is invariant to past values of Series *i*
- Define non-linear mapping using neural networks, and penalize weights to identify the invariances

Adaptations to Straightforward NN-approach

- Shortcomings of a straightforward Neural Network based approach
  - With just an MLP, RNN or LSTM, it is difficult to find the invariances
  - All time series need to have the same lag for modelling
- Adaptations
  - Associate individual lags for every time series
  - Group inputs by: (K lags of series)
  - Place group-wise penalty on layer 1 weights
  - Can be extended to hierarchical group lasso penalty
  - RNNs and LSTMs can further capture the non-linearity via the hidden states that captures the historical context
- Series *i* does not cause Series *j* if the group *i* weights are 0

Temporal Causal Discovery Framework

Temporal causal discovery framework



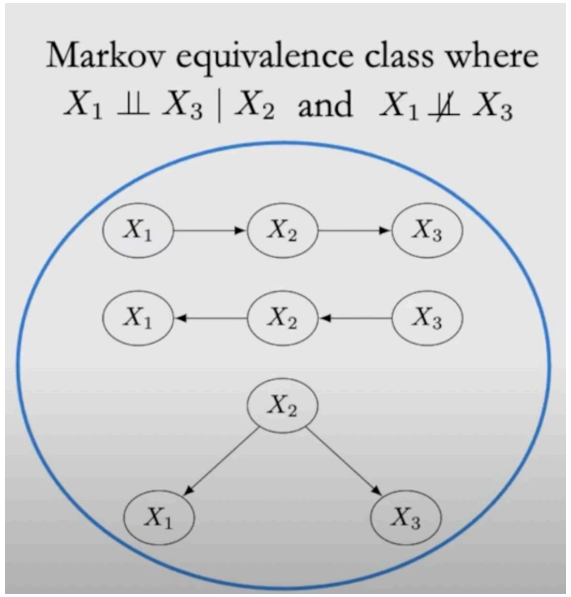
PCMCI

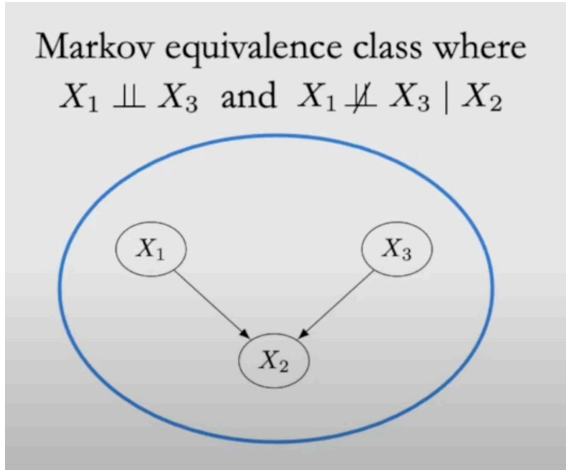
An approach for causal graphical model

Preliminaries

- Markov assumption (Directed graph to data)
  - $X \perp\!\!\!\perp_G Y \mid Z \implies X \perp\!\!\!\perp_P Y \mid Z$; where $\perp\!\!\!\perp_G$ is d-sep
- Goal: Data to causal graph – Faithfulness assumption
  - $X \perp\!\!\!\perp_P Y \mid Z \implies X \perp\!\!\!\perp_G Y \mid Z$
- Causal sufficiency
  - No unobserved confounders in any variables in graph
- Acyclic
  - No cycles in the graph
- Markov equivalence classes[1]
  - Chains and forks give same dependencies
  - Immoralities
  - Distinguish graphs
    - Skeleton & immoralities

  Theorem: Two graphs are Markov equivalent iff same skeleton and immoralities





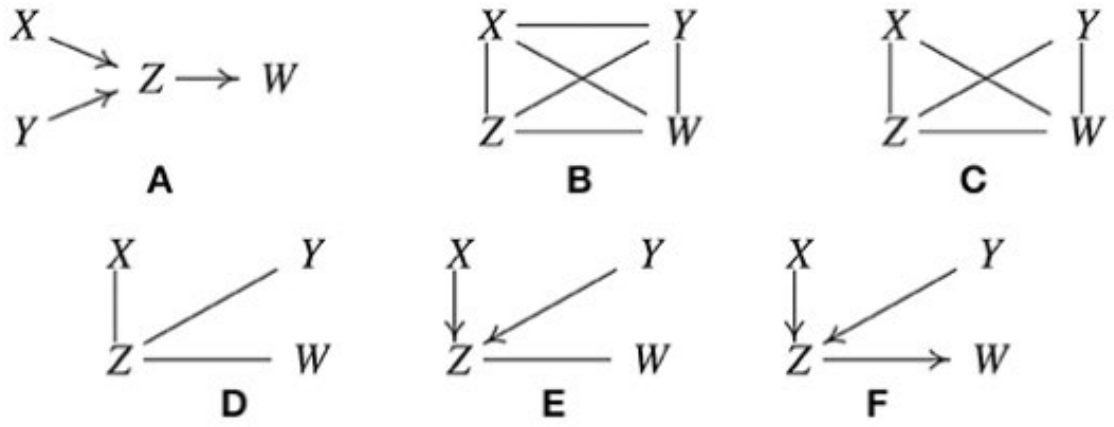
Source: Brady Neal, Causal inference course

[1] D. Koller & N.Friedman, Probabilistic Graphical Models, 2009

PC algorithm [1]

Assumptions:

- Markov property
- Faithfulness
- Causal sufficiency
- Acyclicity



Source: Review of Causal Discovery Methods Based on Graphical Models [2]

Algorithm:

- Form a complete undirected graph for the given nodes
- Pairwise independence by eliminating edges between variables
  - That are unconditionally independent
  - Check for conditional independence by adding node to separation set S, i.e., check $a \perp\!\!\!\perp b \mid S$
- Finding V-structures in triplet variables
- Orientation propagation, i.e., for triplet variables of form $a \rightarrow b - c$, orient the edge $b - c$ as $b \rightarrow c$

[1] P. Sprites and C. Glymour, An algorithm for fast recovery of Sparse Causal Graphs, 1991

[2] C. Glymour, K Zhang and P. Sprites, Review of Causal Discovery Methods Based on Graphical Models, 2019

PC-MCI [1]

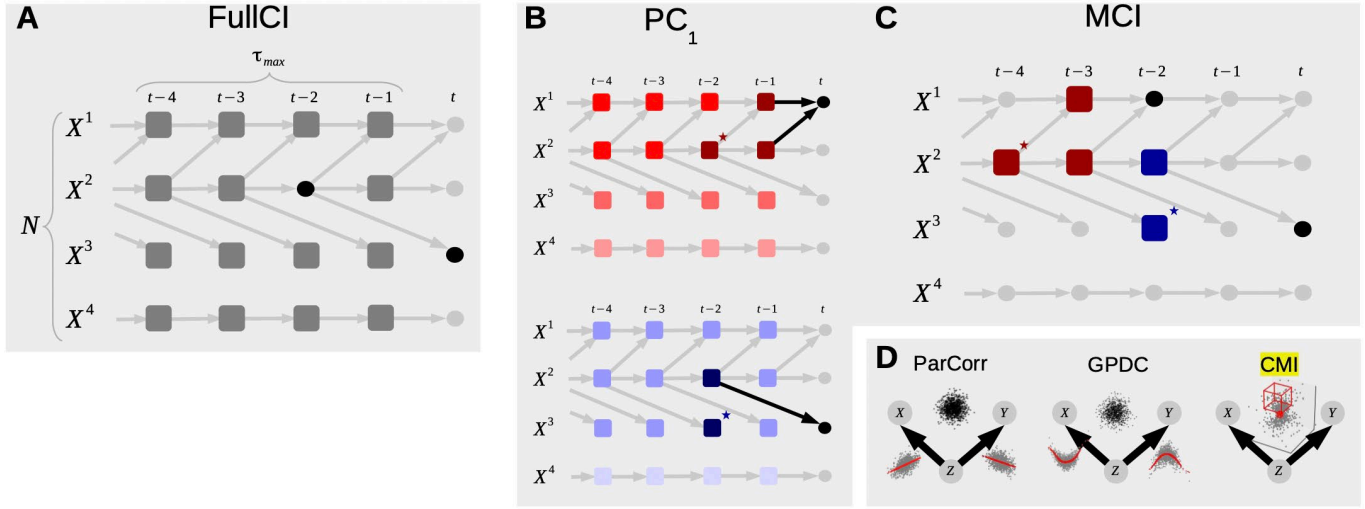
Consider a stationary time series $\boldsymbol{X}_t = \{X_t^1, X_t^2, \dots, X_t^N\}$. Goal is to identify causal relations $X_{t-\tau}^i \rightarrow X_t^j$ if $X_{t-\tau}^i$ is not conditionally independent of X_t^j given the past of all variables i.e., $X_{t-\tau}^i \not\perp\!\!\!\perp X_t^j \mid \boldsymbol{X}_t^- \setminus \{X_{t-\tau}^i\}$

Algorithm

- PC step: identify relevant conditions $\hat{P}\left(X_t^j\right)\ \forall j \in \{1,2,\dots,N\}$ through conditional tests

  Parameters: significance threshold, max time lag
- MCI step: test $X_{t-\tau}^i \rightarrow X_t^j$ i.e., $X_{t-\tau}^i \perp\!\!\!\perp X_t^j \left| \hat{P}\left(X_t^j\right) \setminus \{X_{t-\tau}^i\}, \hat{P}\left(X_{t-\tau}^i\right) \right.$

  Parameters: max parents of X_t^j

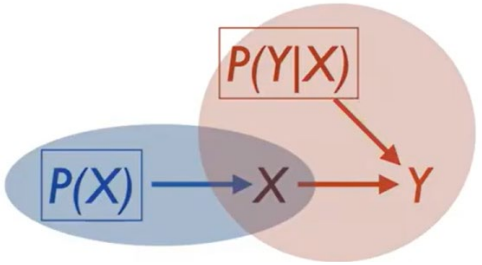


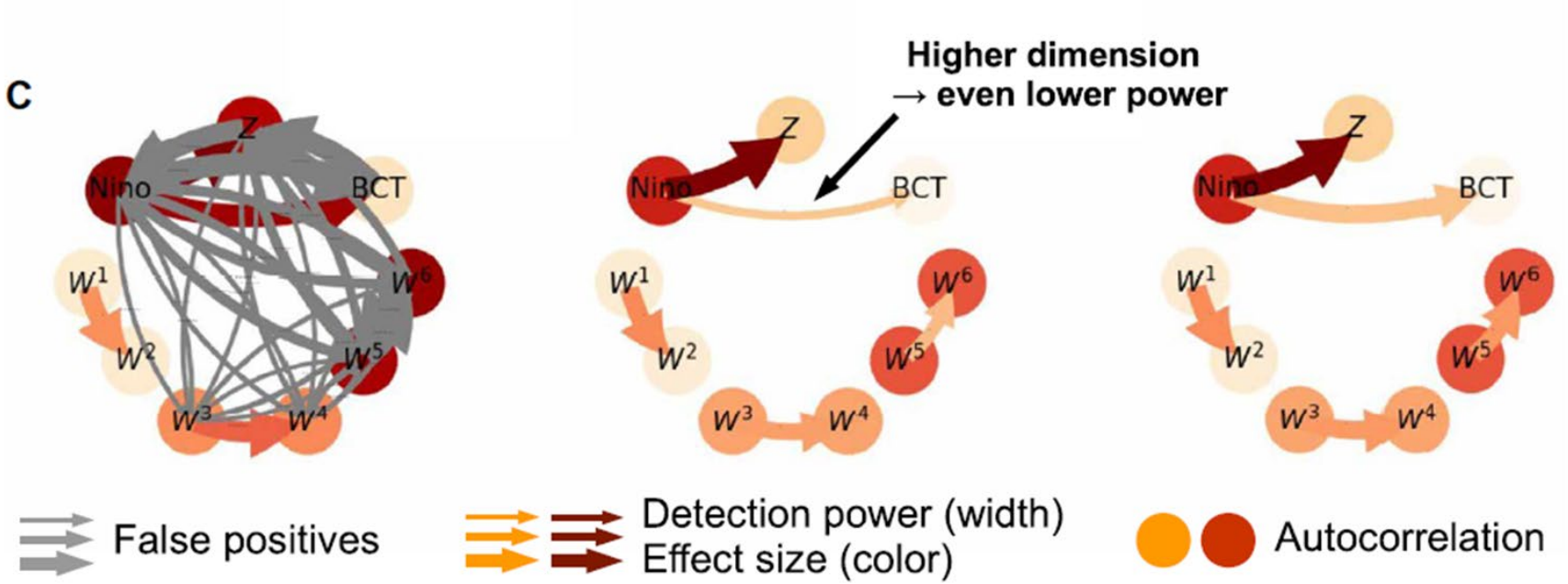
Source: Detecting causal associations in large non-linear time series datasets [1]

[1] J. Runge et al, Detecting causal associations in large non-linear time series datasets, 2019

PC-MCI

- FullCI directly tests the link defining conditional independence
- PCMCI checks (a) conditional selections and (b) momentary conditional independence tests
- Benefits: scales to higher dimensions
- More robust to non-stationarity compared to Lasso regression and conditional selections





Source: Detecting causal associations in large non-linear time series datasets [1]

[1] J. Runge et al, Detecting causal associations in large non-linear time series datasets, 2019

Other approaches

Relax causal sufficiency

- Gerhardus, A. & Runge, J. High-recall causal discovery for autocorrelated time series with latent confounders, 2020
- Mastakouri, et al, Necessary and sufficient conditions for causal feature selection in time series with latent common causes, 2021

Non-stationary

- Li, and Bühlmann, Estimating heterogeneous treatment effects in nonstationary time series with state-space models, 2018
- Huang, B., Zhang, K., Gong, M. and Glymour, C., Causal discovery and forecasting in nonstationary environments with state-space models, 2019

Contemporaneous causal effects

- Peters, J., Janzing, D. and Schölkopf, B., Elements of causal inference: foundations and learning algorithms, 2017

Convergent Cross Mapping

A dynamical systems approach

Non-linear dynamical systems (1/2)

— A dynamical system is modelled by a system of differential equations / partial differential equations

$$m\frac{d^2x}{dt^2} + b\frac{dx}{dt} + kx = 0$$

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

- A dynamical system is often modelled by a state-space representation

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -\tfrac{b}{m}x_2 - \tfrac{k}{m}x_1 .$$

- A linear system is one which is represented by a system of equations whose state variables appear in first power only

Non-linear dynamical systems (2/2)

— A non-linear system however has state space representation with non-linear terms

$$\ddot{x} + \frac{g}{L}\sin x = 0$$

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = -\frac{g}{L}\sin x_1 .$$

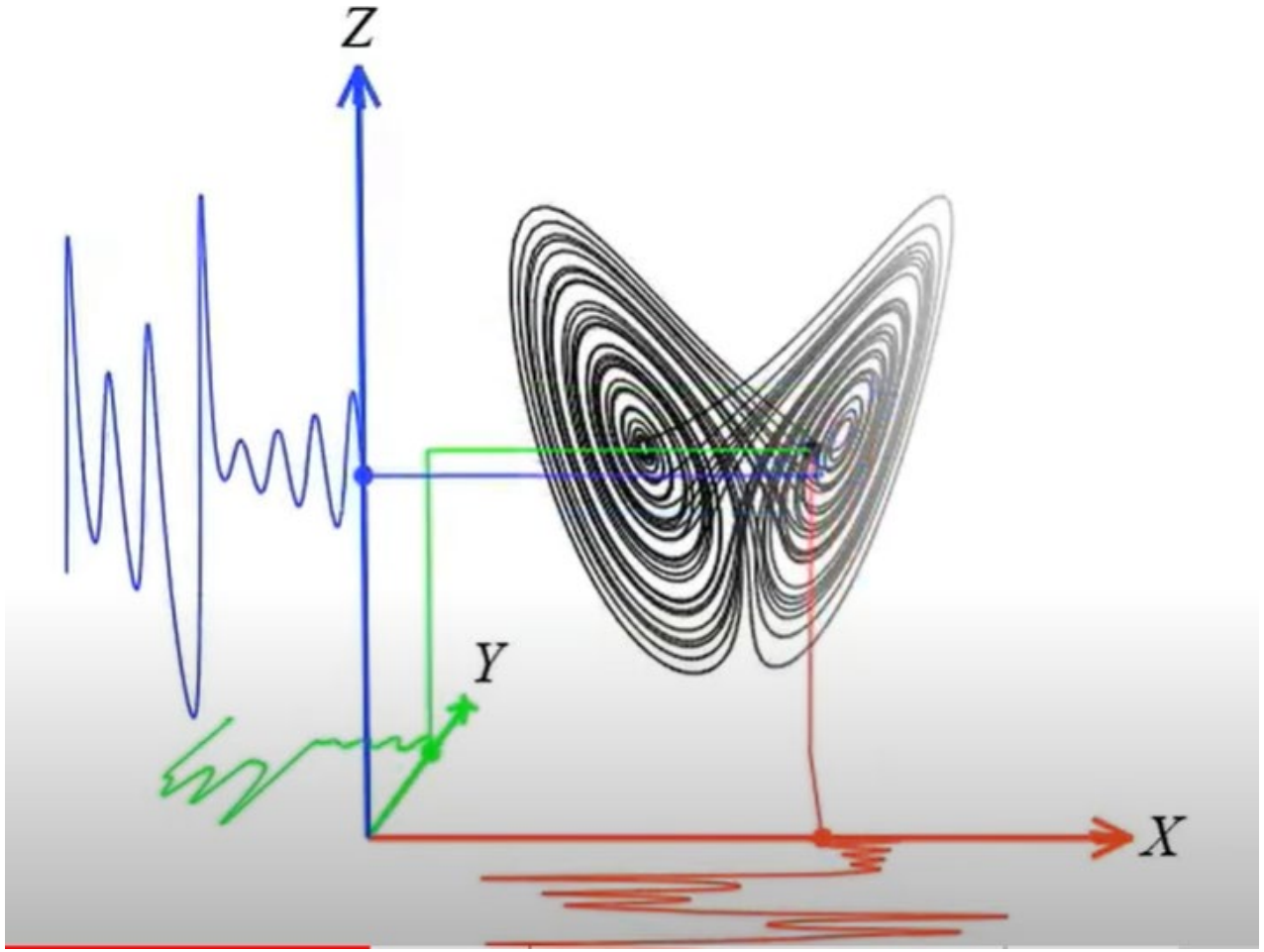
- Non-linear systems response often depends to a large extent on the initial state. Hence these systems are sometimes said to exhibit chaotic behaviour although they are modelled completely by a deterministic system

Example of a non-linear system

Some definitions

- Orbit / trajectory – is a sequence of variables of the state variables of a dynamical system
- Fixed point – states from which system will not move from there by the dynamics. Can be stable (attracting) / unstable
- Attractor is a set of states toward which a system tends to evolve as time goes to infinity
- Set of initial conditions which go to the same attractor is called a **basin of attraction**
- **Manifolds** are surfaces on the state space such that if a state starts on one manifold stays on it.

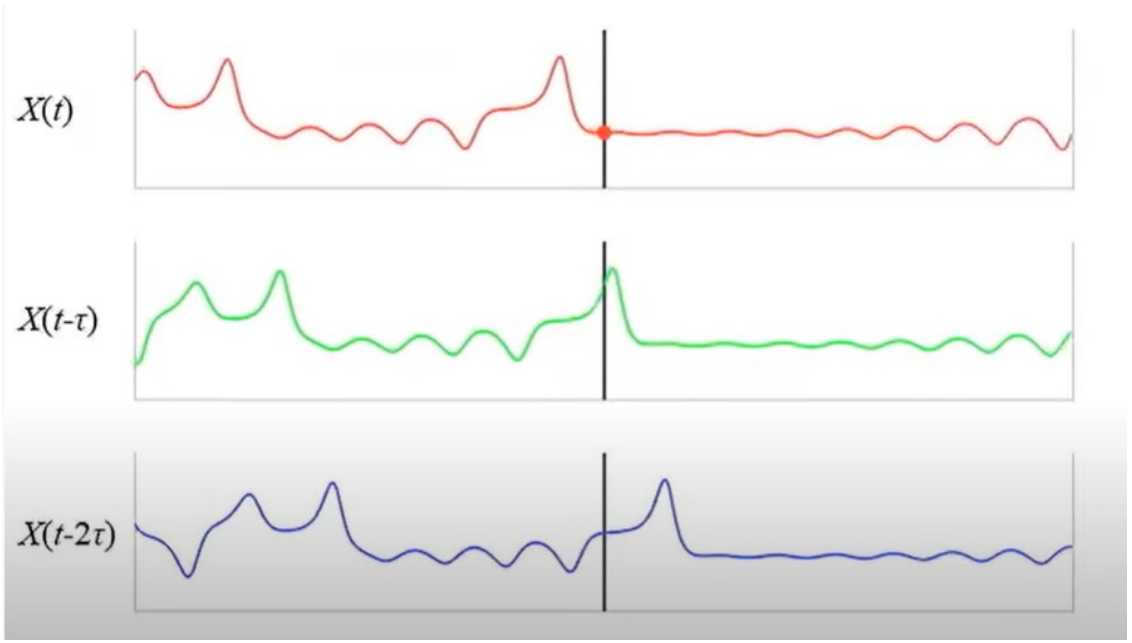
Manifolds and time series

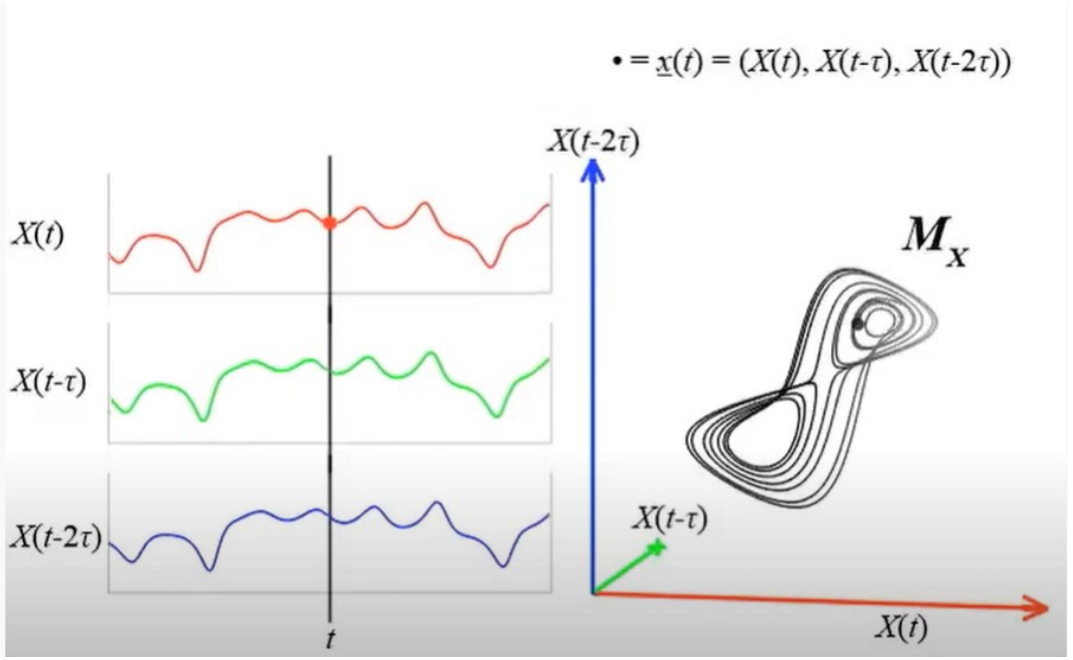


Source:- [Takens' theorem in action for the Lorenz chaotic attractor - YouTube]

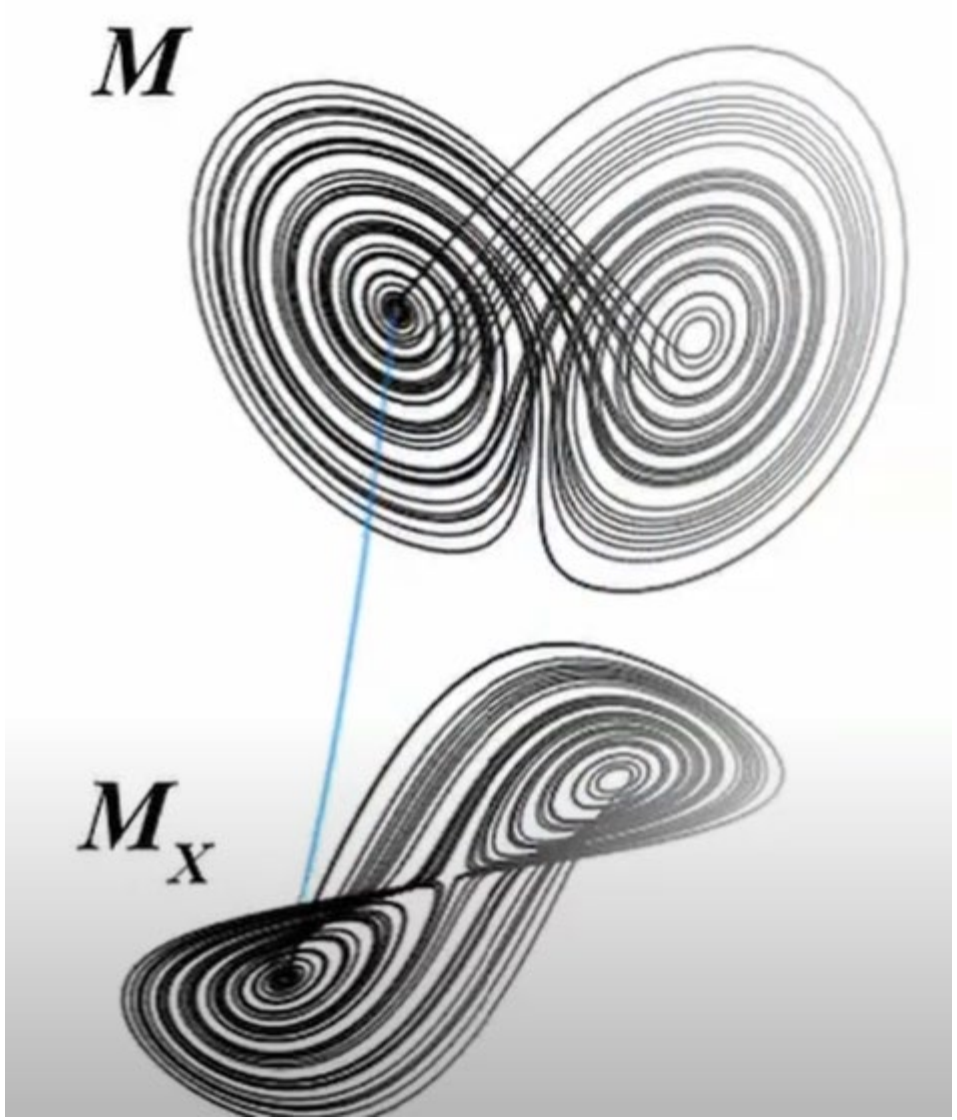
Taken's Theorem

— Loosely speaking, Taken's theorem states that one can reconstruct a shadow version (topologically invariant) of its manifold by looking at the delay embeddings of any of the time series.



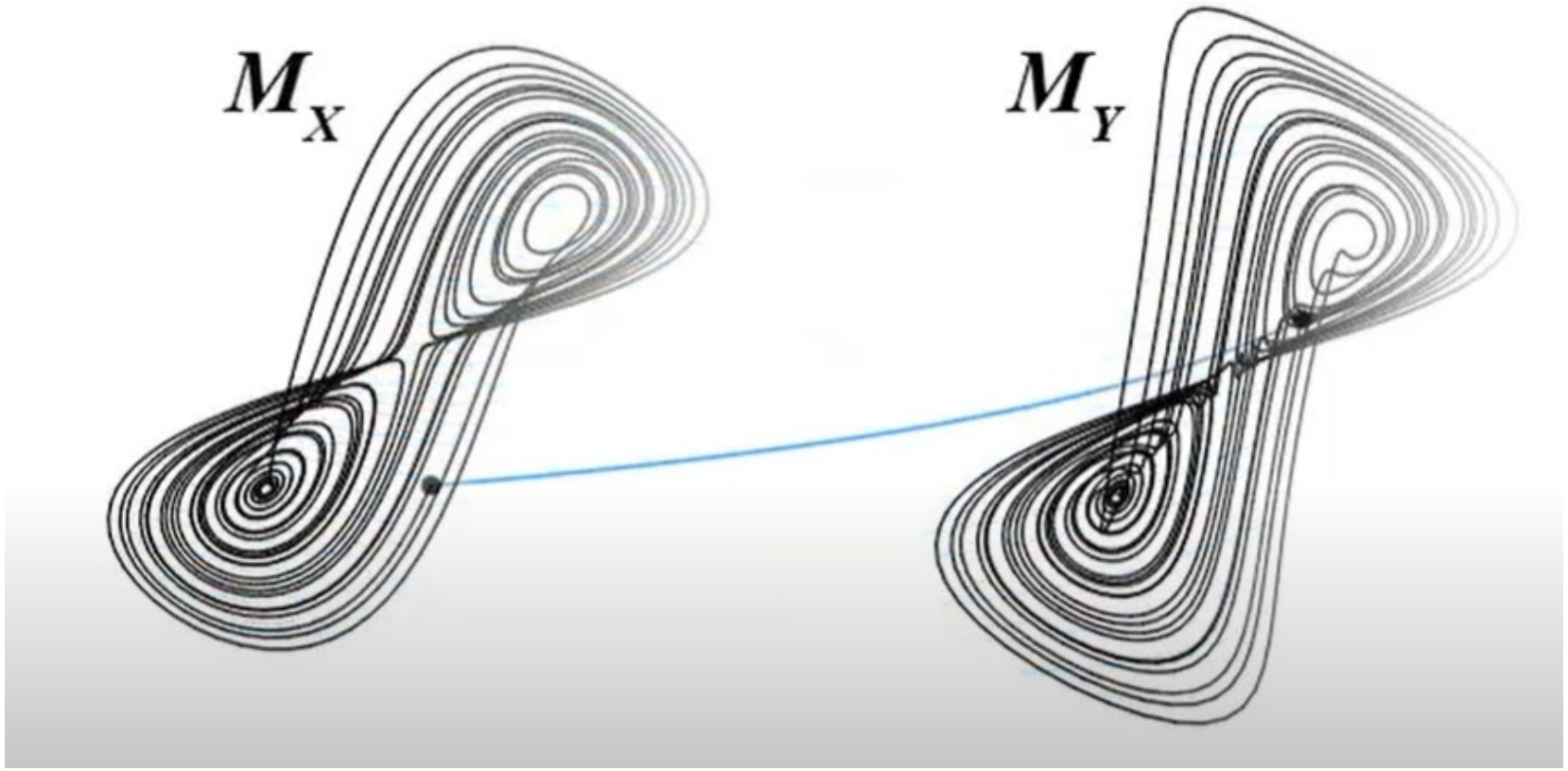


1-2-1 correspondence



There is a one-to-one correspondence between the original manifold in the state space and a shadow manifold formed by the time lagged versions of any time series

Identifying causal relationship between variables



Two variables are said to be drawn from the state dynamical system and hence have a causal link if the nearest neighbours on one shadow manifold are time indexed nearest neighbours on another shadow manifold

Use of cross mapping

- We build an estimate

$$\hat{x}_t | M_y = \sum_{i=1}^{E+1} w_i x_{t_i}$$

- A library of L points from M_y is used to provide estimates of L points in x.
- The Pearson correlation ρ_{xx} of the L true values and the L cross-mapped values is an indicator of how much the dynamics of one system is influences the other
- If x→y then the estimate of x obtained from M_y should improve as the number of points L become larger – since they are a better representation of the nearest neighbour points.
- The convergence of this ρ_{xx} is a critical aspect to understand causal relationship
- The values of ρ_{xx} and ρ_{yy} and their convergence can show the direction of causality

Convergence of Cross mapping

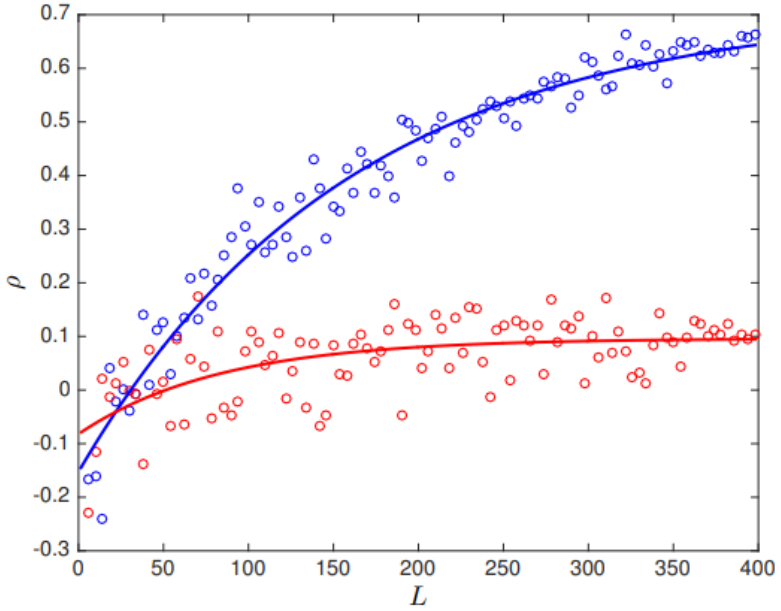


Figure 5: The correlation coefficient $\rho_{x,\hat{x}}$ (blue circles) and $\rho_{y,\hat{y}}$ (red circles) as a function of library length L. Fits to the function in Eq. 5 are shown as solid lines. For x the converged value is $\rho_\infty = 0.71$, and for y it is $\rho_\infty = 0.096$. Parameter values: $r_x = 3.65$, $r_y = 3.77$, $\beta_{xy} = 0$, $\beta_{yx} = 0.05$.

$$x_{t+1} = x_t(r_x(1 - x_t) - \beta_{xy}y_t) + \varepsilon_{x,t}$$
$$y_{t+1} = y_t(r_y(1 - y_t) - \beta_{yx}x_t) + \varepsilon_{y,t} \quad (2)$$

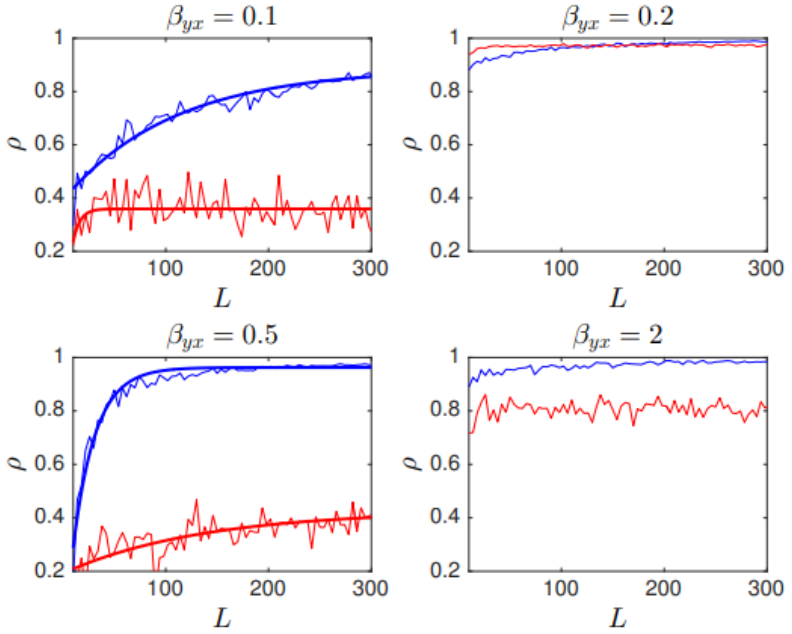


Figure 6: Correlations between observed and cross-mapped estimates of x (blue) and y (red) for four different values of β_{yx}. Thin lines are the computed correlations and thick lines are fits to Eq. 5. For $\beta_{yx} = 0.2$ and $\beta_{yx} = 2$ a good fit is not possible, and no fits are shown.

Source:- Inferring causality from noisy time series data Monster et al

Latent Convergent Cross Mapping

- One of the major issues around CCM is that it will be very sensitive to missing data , outliers and other commonly observed issues in real world data
- The same issues would be present with short / sporadic time series – where the time series is observed for small bursts of time only

Neural ODEs

- In general not all dimensions of a dynamical system is observed
- So we model this as a continuous latent process H(t) conditioned on which the observations are generated

- Learning the dynamics $X[t] = g(H[t])$ with $\frac{dH(t)}{dt} = f_\theta(H(t), t)$ the a finite of set of potentially noisy obser
- Neural ODEs parametrise this process via a Neural network

GRU-ODE-Bayes model

— A GRU-ODE-Bayes model uses a Gated Recurrent Network (GRU) and uses a Bayesian update process to estimate the system

To derive the GRU-based ODE, we first show that the GRU proposed by Cho et al. (2014) can be written as a difference equation. First, let $\mathbf{r}_t$, $\mathbf{z}_t$, and $\mathbf{g}_t$ be the reset gate, update gate, and update vector of the GRU:

$$\begin{aligned}\mathbf{r}_t &= \sigma(W_r\mathbf{x}_t + U_r\mathbf{h}_{t-1} + \mathbf{b}_r)\\ \mathbf{z}_t &= \sigma(W_z\mathbf{x}_t + U_z\mathbf{h}_{t\ \ 1} + \mathbf{b}_z)\\ \mathbf{g}_t &= \tanh(W_h\mathbf{x}_t + U_h(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h),\end{aligned} \tag{2}$$

where $\odot$ is the elementwise product. Then the standard update for the hidden state $\mathbf{h}$ of the GRU is

$$\mathbf{h}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \mathbf{g}_t.$$

We can also write this as $\mathbf{h}_t = \mathbf{GRU}(\mathbf{h}_{t-1}, \mathbf{x}_t)$. By subtracting $\mathbf{h}_{t-1}$ from this state update equation and factoring out $(1 - \mathbf{z}_t)$, we obtain a difference equation

$$\begin{aligned}\Delta\mathbf{h}_t = \mathbf{h}_t - \mathbf{h}_{t-1} &= \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \mathbf{g}_t - \mathbf{h}_{t-1}\\ &= (1 - \mathbf{z}_t) \odot (\mathbf{g}_t - \mathbf{h}_{t-1}).\end{aligned}$$

This difference equation naturally leads to the following ODE for $\mathbf{h}(t)$:

$$\frac{d\mathbf{h}(t)}{dt} = (1 - \mathbf{z}(t)) \odot (\mathbf{g}(t) - \mathbf{h}(t)), \tag{3}$$

where $\mathbf{z}$, $\mathbf{g}$, $\mathbf{r}$ and $\mathbf{x}$ are the continuous counterpart of Eq. 2. See Appendix A for the explicit form.

We name the resulting system *GRU-ODE*. Similarly, we derive the *minimal GRU-ODE*, a variant based on the minimal GRU (Zhou et al., 2016), described in appendix G.

Algorithm 1 GRU-ODE-Bayes

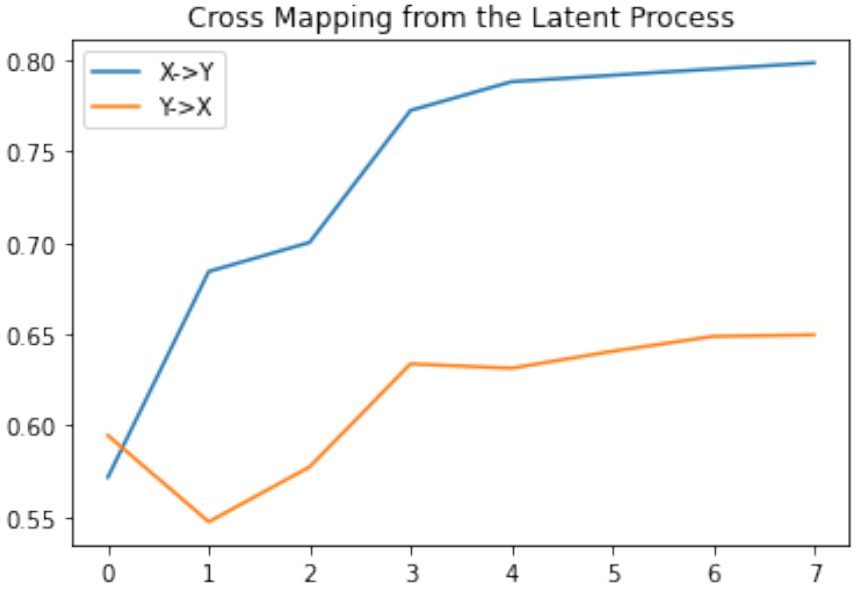
Input: Initial state $\mathbf{h}_0$,
observations $\mathbf{y}$, mask $\mathbf{m}$,
observation times $\mathbf{t}$, final time T.
Initialize time $= 0$, loss $= 0$, $\mathbf{h} = \mathbf{h}_0$.
for $k = 1$ **to** K **do**
{ODE evolution to $\mathbf{t}[k]$}
$\mathbf{h} = \mathbf{GRU\text{-}ODE}(\mathbf{h}, \text{time}, \mathbf{t}[k])$
time $= t[k]$
{Pre-jump loss}
loss $+= \text{Loss}_{\text{pre}}(\mathbf{y}[k], \mathbf{m}[k], \mathbf{h})$
{Update}
$\mathbf{h} = \mathbf{GRU\text{-}Bayes}(\mathbf{y}[k], \mathbf{m}[k], \mathbf{h})$
{Post-jump loss}
loss $+= \lambda.\ \text{Loss}_{\text{post}}(\mathbf{y}[k], \mathbf{m}[k], \mathbf{h})$
end for
{ODE evolution to T}
$\mathbf{h} = \mathbf{GRU\text{-}ODE}(\mathbf{h}, \mathbf{t}[N_K], T)$
return $(\mathbf{h}, \text{loss})$

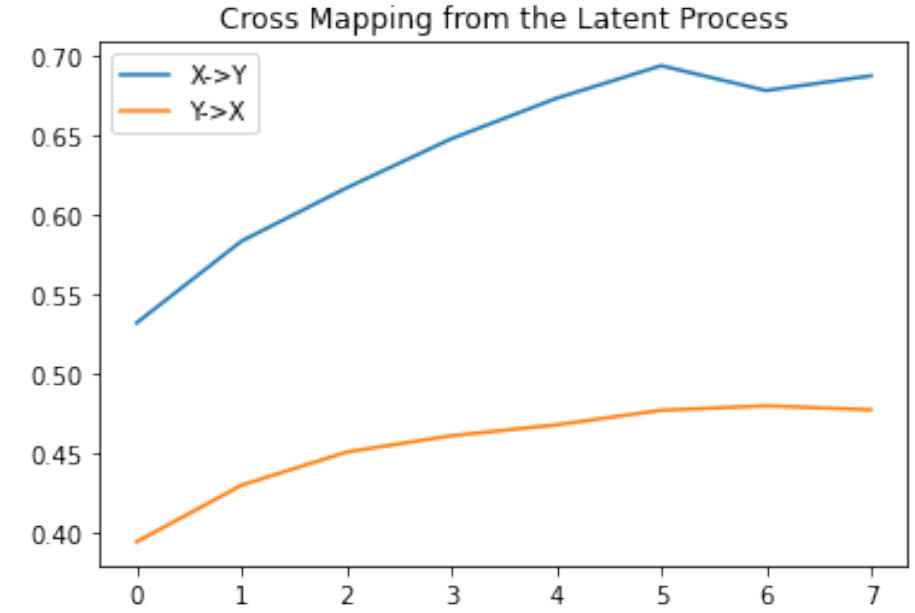
Source - Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. Gru-ode-bayes: Continuous modeling of sporadically-observed time series. In Advances in Neural Information Processing Systems, pp. 7377–7388, 2019

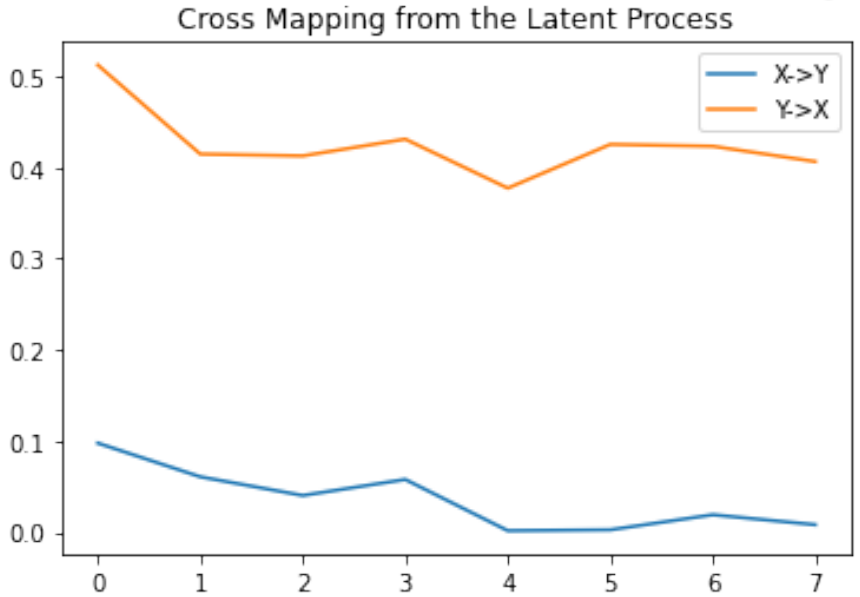
Latent CCM

- Train a GRU-ODE-Bayes model to learn H(t)
- Use this hidden representation as a complete representation of the state space shadow manifold and use it to estimate cross mapping and convergence

Results


Cross Mapping from the Latent Process
X->Y
Y->X
0.80
0.75
0.70
0.65
0.60
0.55
0
1
2
3
4
5
6
7


Cross Mapping from the Latent Process
X->Y
Y->X
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0
1
2
3
4
5
6
7

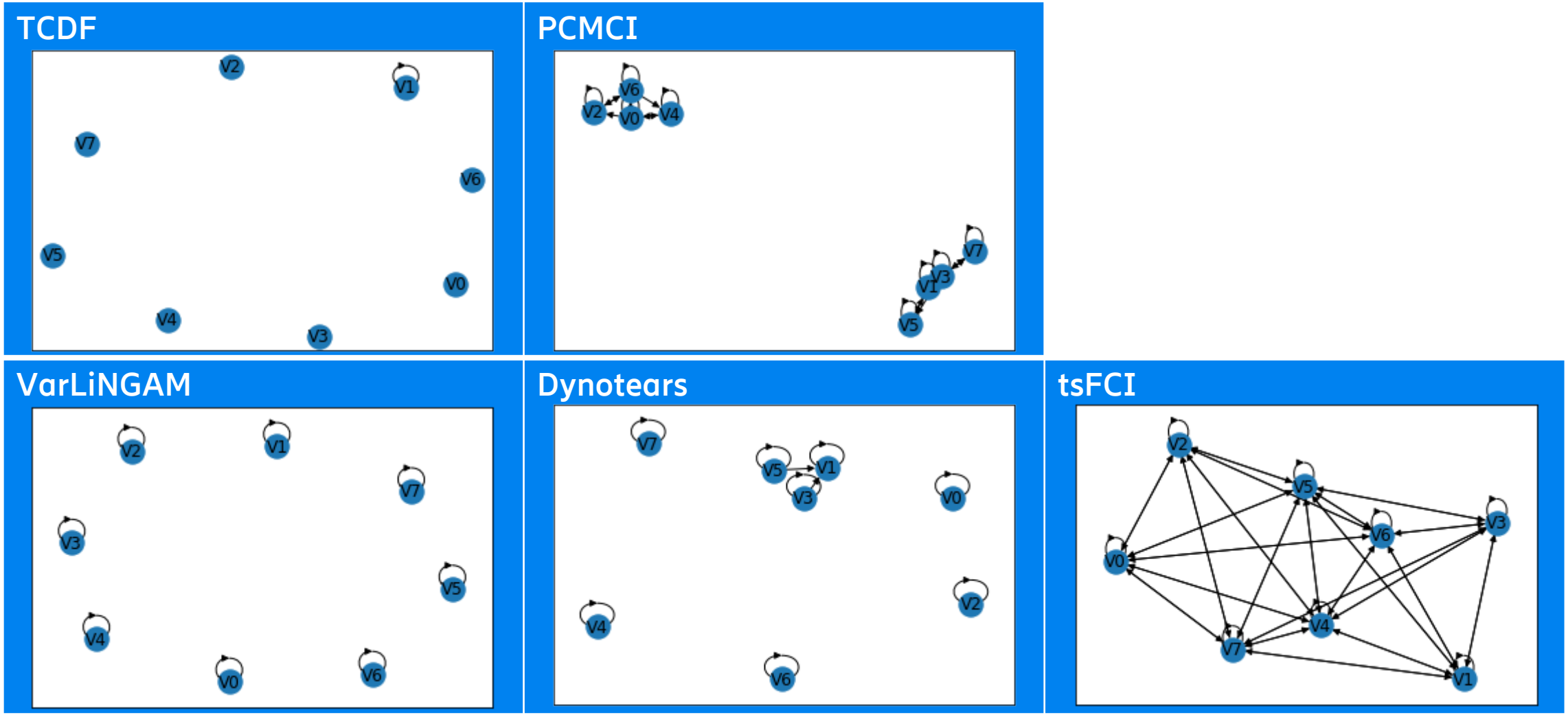

Cross Mapping from the Latent Process
X->Y
Y->X
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7

Advantages of CCM over GC

— Granger Causality (GC) assumes stochastic relation between time series and not deterministic dynamical relations

— GC assumes separability between causal and caused variables which is not needed for CCM

Comparative Results



Notebook Walkthrough

Conclusions and comments

Q&A