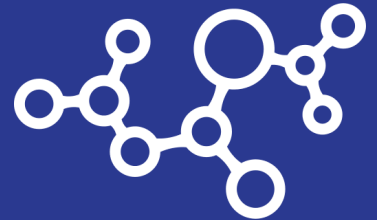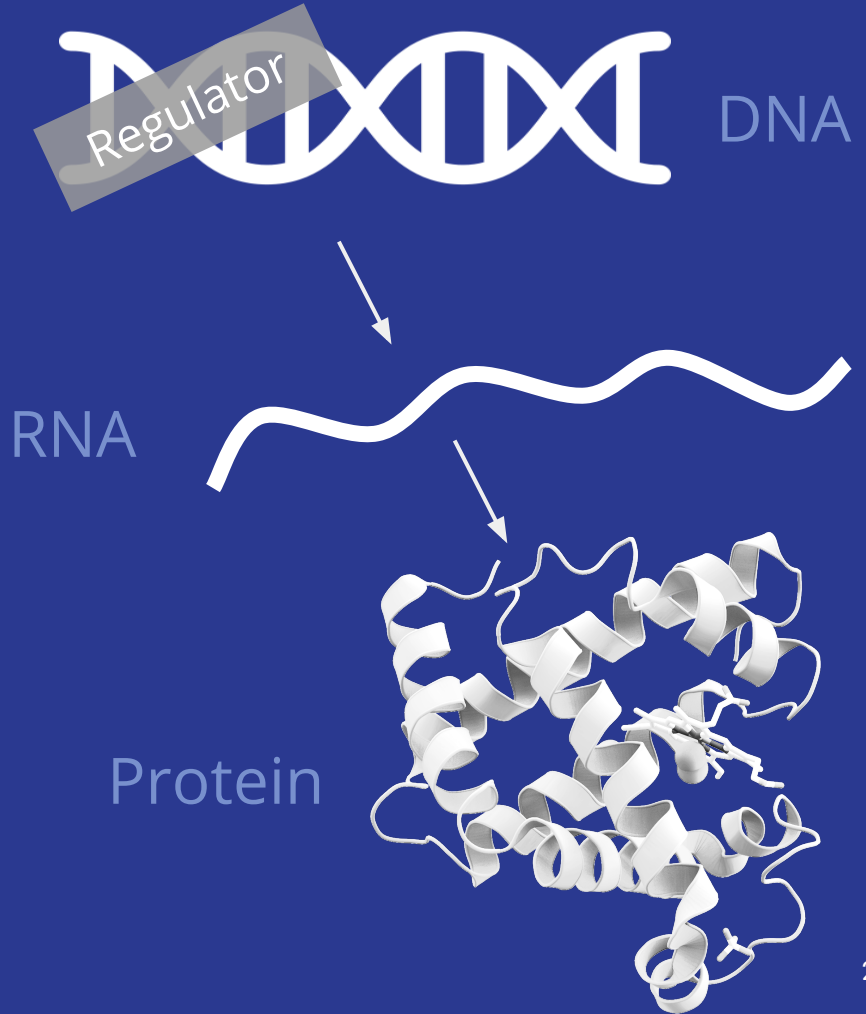# Causal Computational Models for Gene Regulatory Networks

Parul Jain
Sahil Loomba
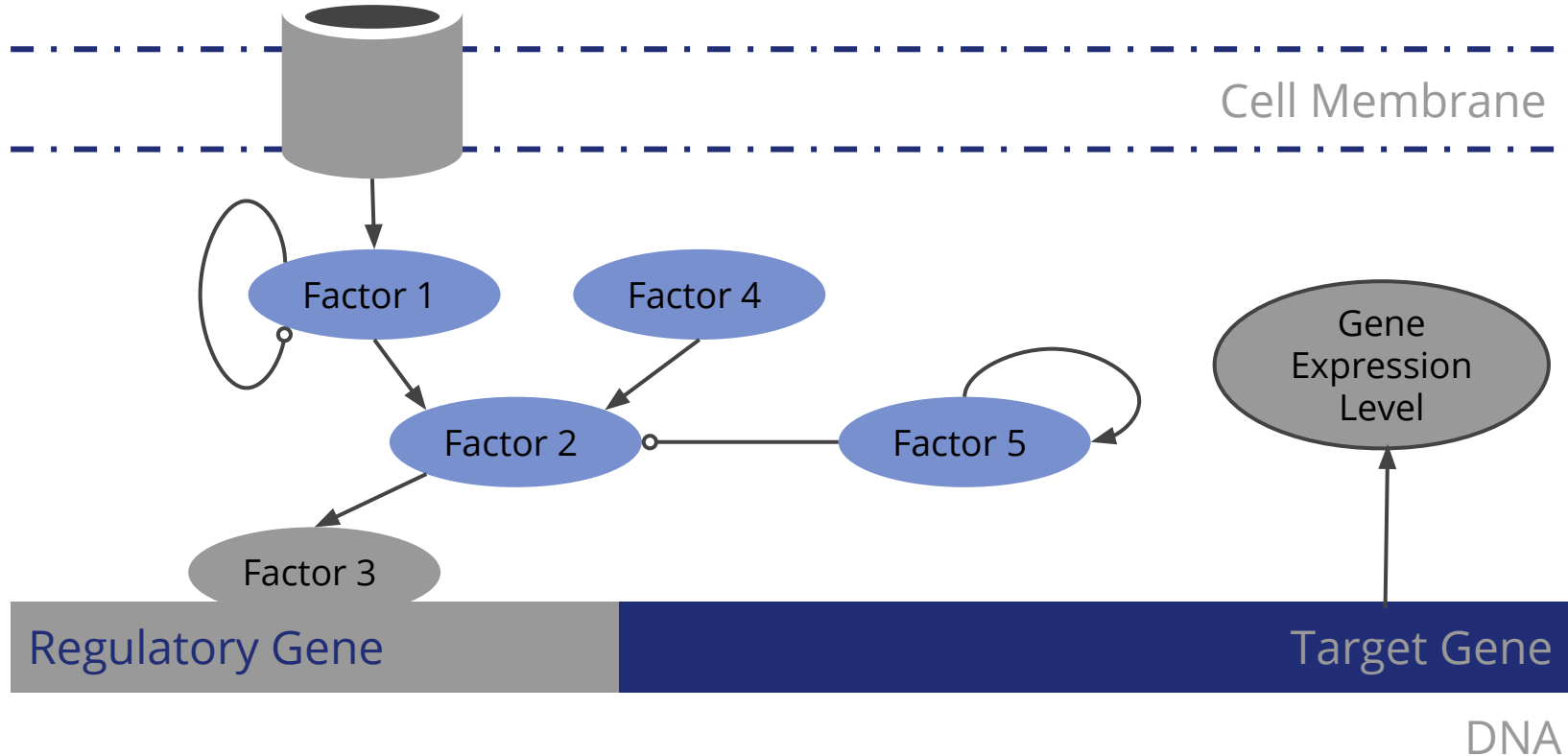
Advisors
Dr. Sumeet Agarwal
Dr. Parag Singla

The Central Dogma of Biology
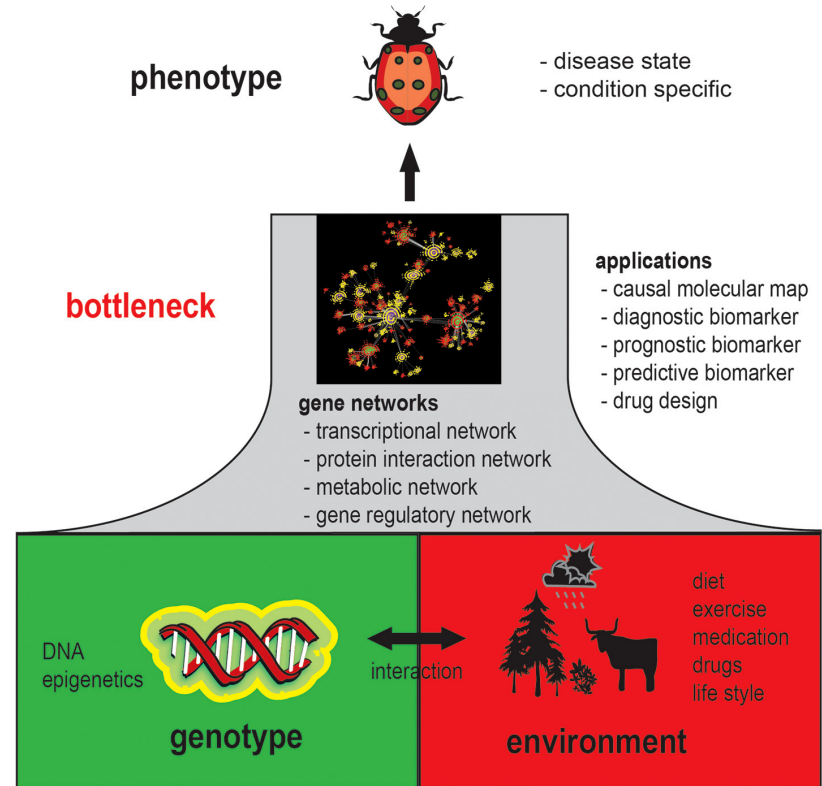
Gene Regulatory Networks - An Overview

Cell Membrane

Factor 1

Factor 4

Factor 2

Factor 5

Factor 3

Gene Expression Level

Regulatory Gene

Target Gene

DNA

3

Importance of GRNs

- Model causal interactions

- Large dimensionality, lower it

  - ~20,000 genes in Humans

- Network Medicine

  - Drug design

phenotype
- disease state
- condition specific

bottleneck

applications
- causal molecular map
- diagnostic biomarker
- prognostic biomarker
- predictive biomarker
- drug design

gene networks
- transcriptional network
- protein interaction network
- metabolic network
- gene regulatory network

DNA
epigenetics

interaction

diet
exercise
medication
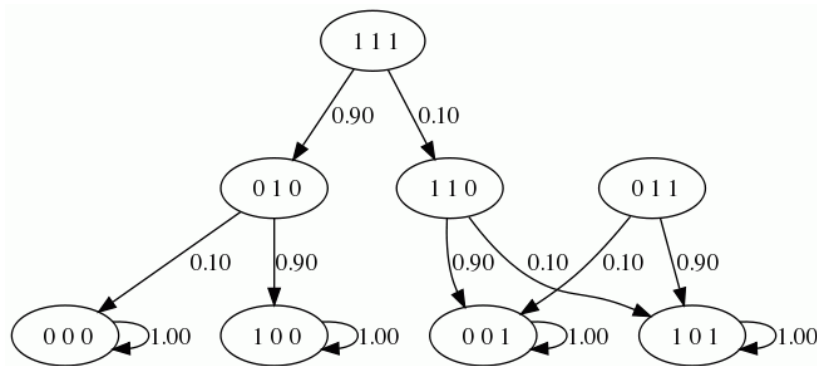drugs
life style

genotype

environment

Modelling GRNs

$$\frac{dy_1}{dx} = f_1(x, y_1, y_2, \ldots, y_n)$$

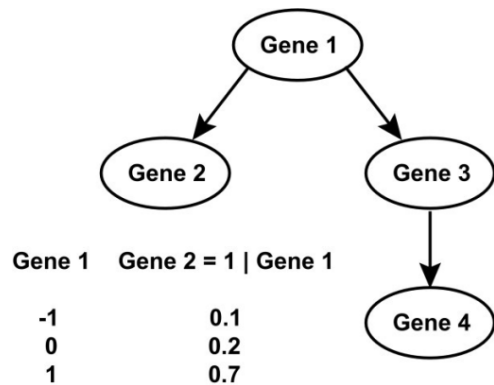$$\frac{dy_2}{dx} = f_2(x, y_1, y_2, \ldots, y_n)$$

......

$$\frac{dy_n}{dx} = f_n(x, y_1, y_2, \ldots, y_n)$$

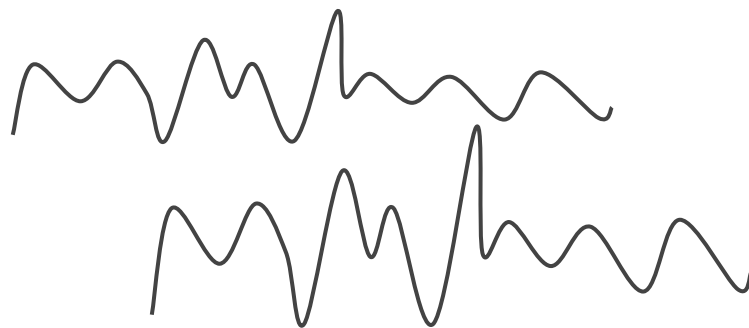Ordinary Differential
Equations

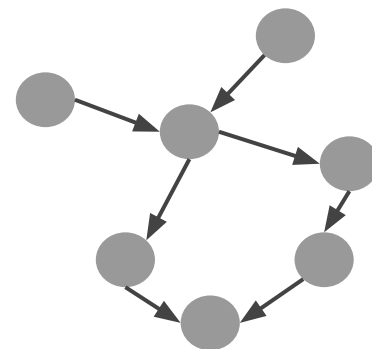Boolean Networks

Bayesian Networks

Discovering GRNs

In Vitro
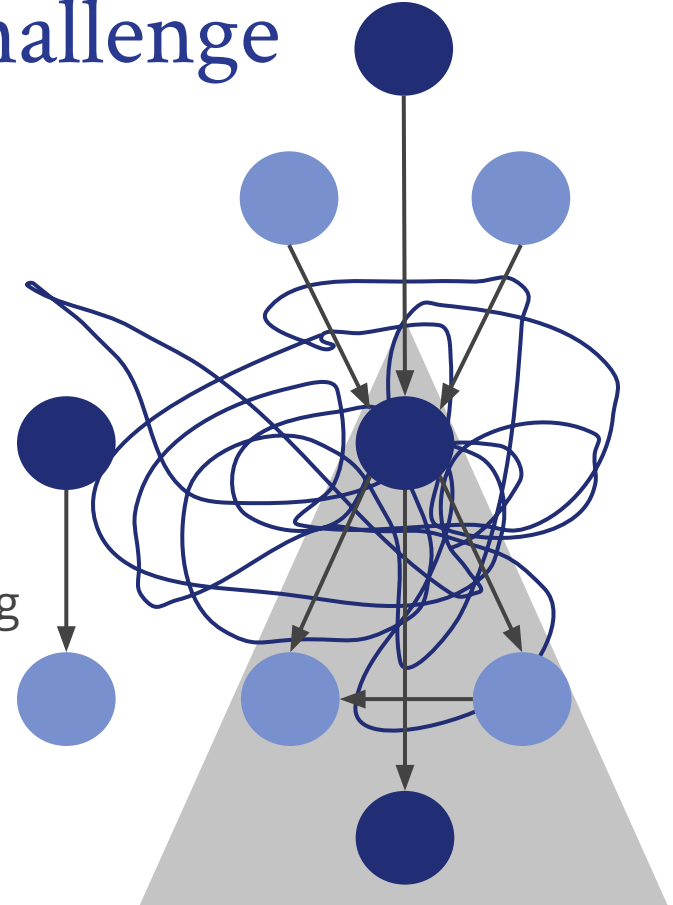
- Gene Silencing
- mRNA expression levels

In Silico

Input: Gene Expression Levels → Time series signal

Output: GRN → Causal interaction between nodes

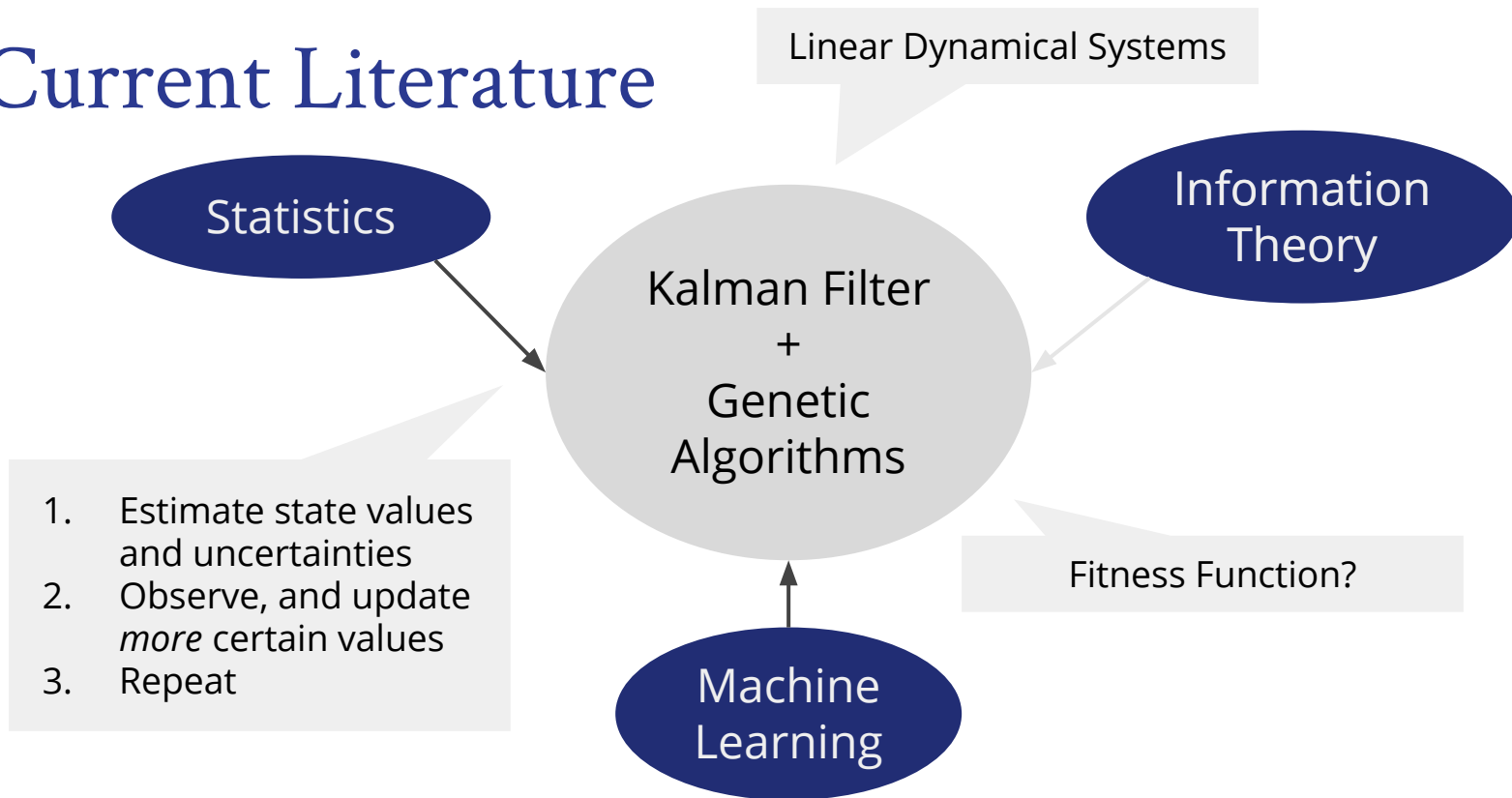# Understanding The GRN Challenge

- Non-linearity of Biological Systems

- Causation versus Correlation

- Structure: Ambiguity and Specificity

- Whole is more than the Sum: Signal Masking
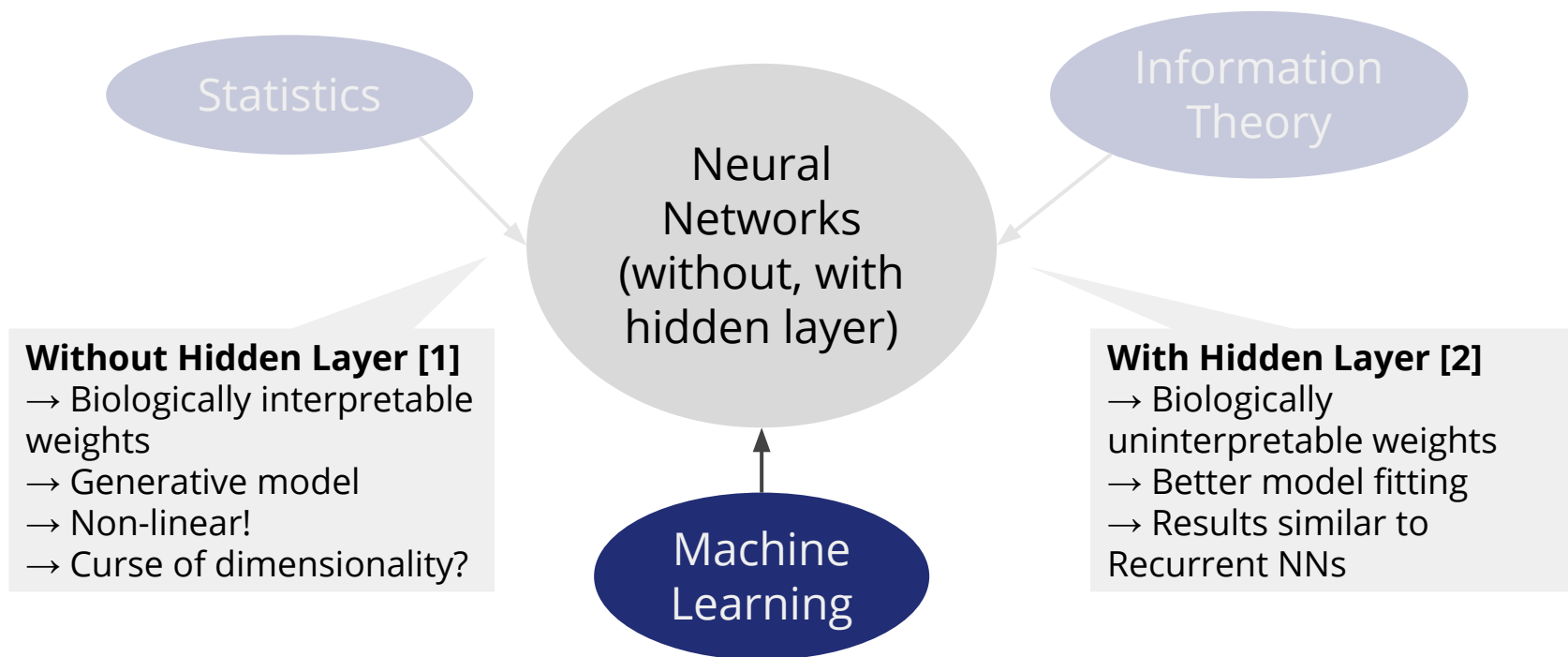
- Under constrained modelling

The Problem Statement

To create **computational** models of GRNs,

  which capture **causal interactions** between the genes involved,

  while placing special emphasis on reducing **dimensionality** of the problem,

  to allow wet lab work for **in/validation** for pressing disease networks

Current Literature

Linear Dynamical Systems

Statistics

Information Theory

Kalman Filter + Genetic Algorithms

1. Estimate state values and uncertainties
2. Observe, and update *more* certain values
3. Repeat

Fitness Function?

Machine Learning

Current Literature

Statistics

Information Theory

Neural Networks (without, with hidden layer)

Machine Learning

Without Hidden Layer [1]
→ Biologically interpretable weights
→ Generative model
→ Non-linear!
→ Curse of dimensionality?

With Hidden Layer [2]
→ Biologically uninterpretable weights
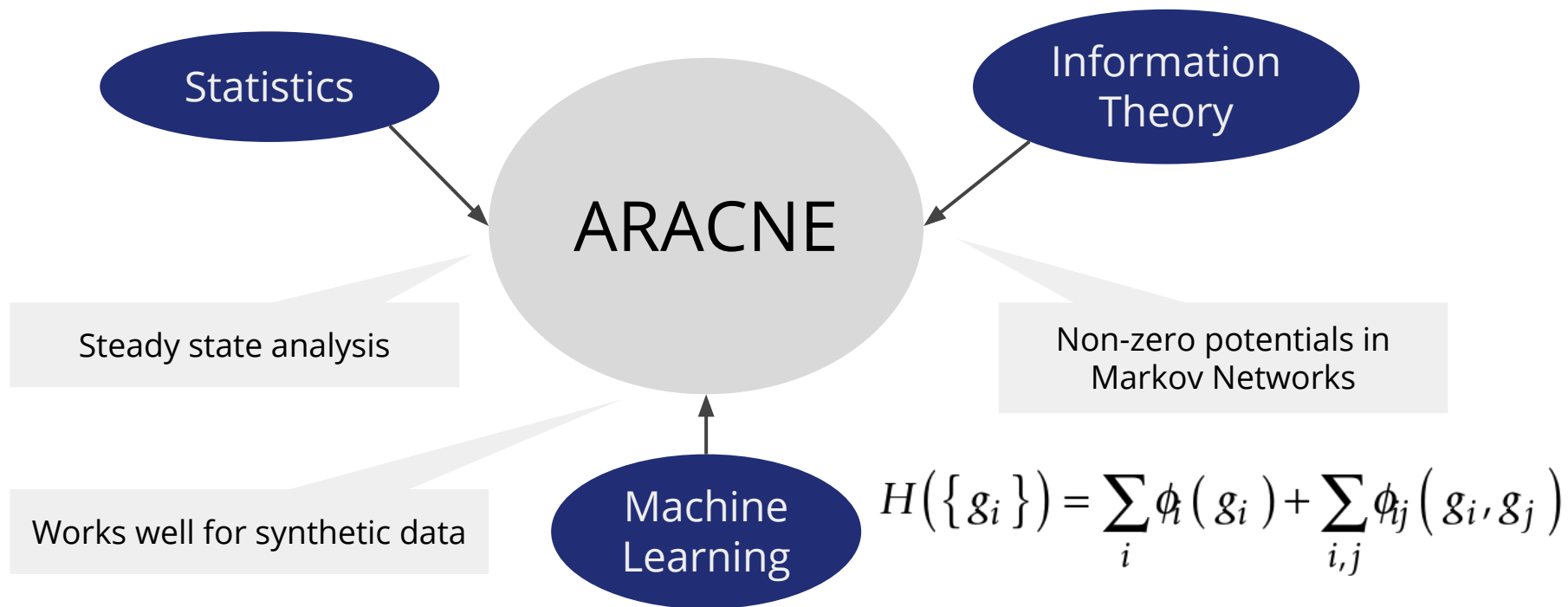→ Better model fitting
→ Results similar to Recurrent NNs

Source [1]: Patrik D'haeseleer, Xiaobin Wen, *Linear modeling of mRNA expression levels during CNS development and injury*, Pacific Symposium on Biocomputing 4:41-52 (1999)
Source [2]: Smith et al. *Time Series Gene Expression Prediction using Neural Networks with Hidden Layers*

Current Literature

$$H(\{g_i\}) = \sum_i \phi_i(g_i) + \sum_{i,j} \phi_{ij}(g_i, g_j)$$

Statistics

Information Theory

ARACNE

Machine Learning

Steady state analysis

Non-zero potentials in Markov Networks

Works well for synthetic data

Datasets

- SysGenSIM: a simulation for gene expression levels, under given network constraints
  - For small number of nodes (10-50)

- DREAM Challenges pose fundamental questions about systems biology and translational medicine
  - DREAM-5 data; Variable time series length

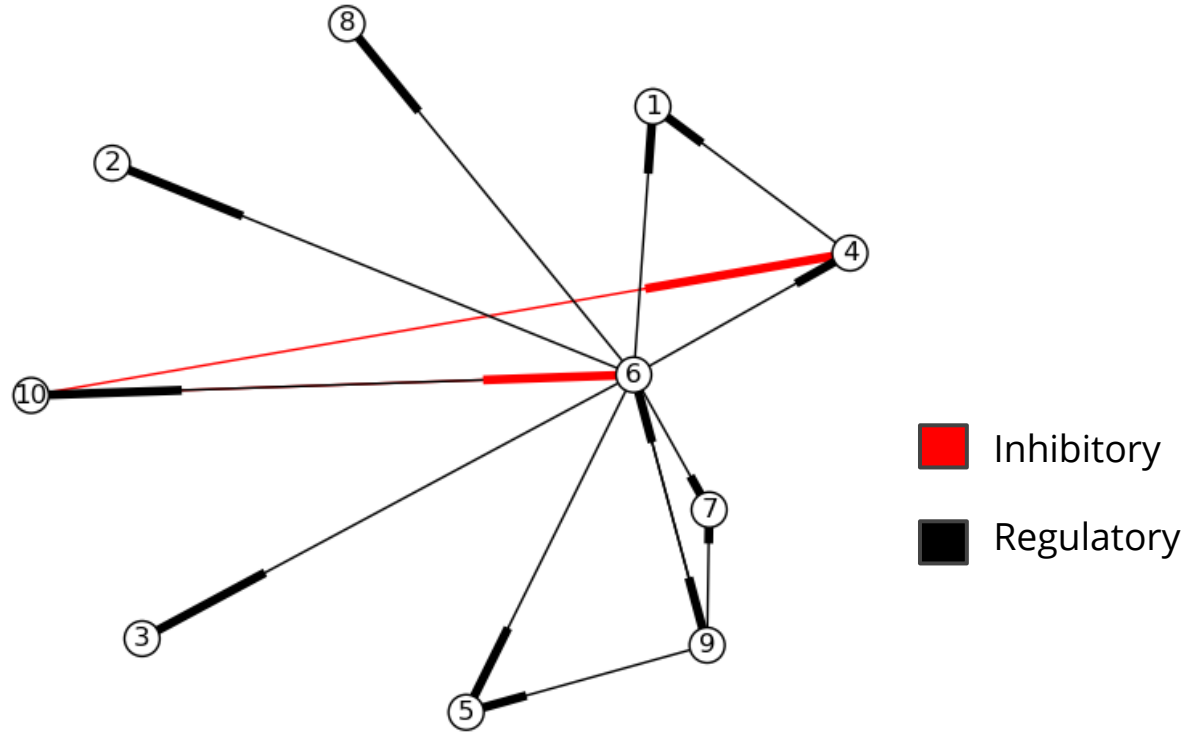- Real-life data for diseases such as cancer, neurodegenerative disorders

Granger Causality

If a signal X_1 "Granger-causes" (or "G-causes") a signal X_2, then past values of X_1 should contain information that helps predict X_2 above and beyond the information contained in past values of X_2 alone.
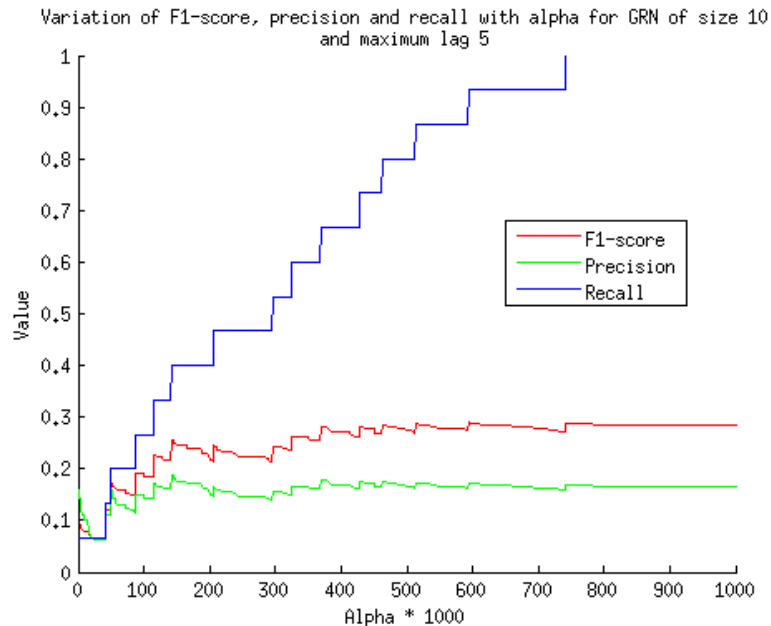
Parameters:

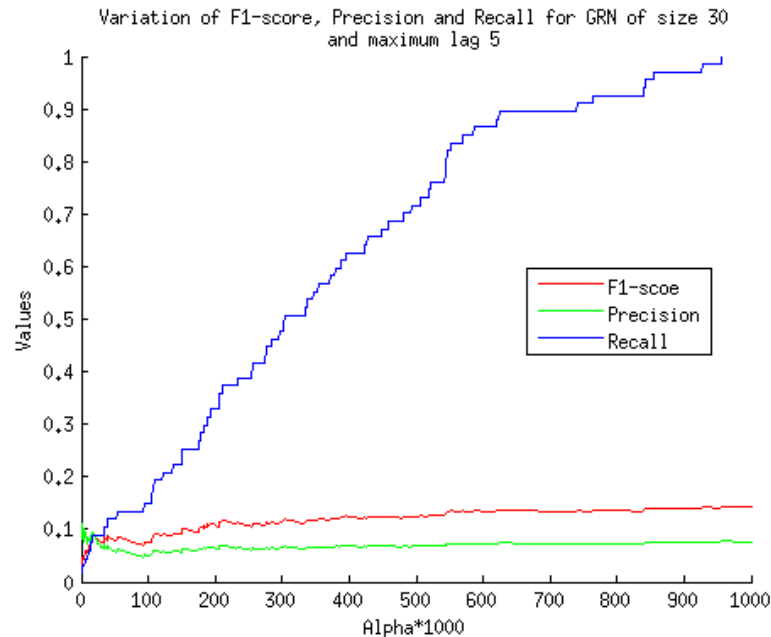- Alpha (Bonferroni Factor)
- Maximum lag

Network: Ground Truth

GRN of size 10

Variation in Alpha (Bonferroni Correction)

Variation of F1-score, precision and recall with alpha for GRN of size 10 and maximum lag 5

Variation of F1-score, Precision and Recall for GRN of size 30 and maximum lag 5
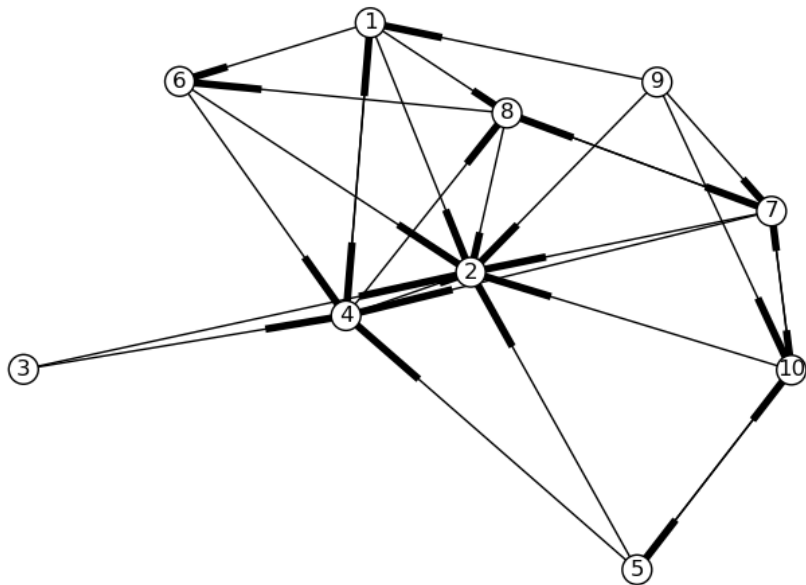
F1-score, Precision and Recall for GRN of size 10 with max lag 5
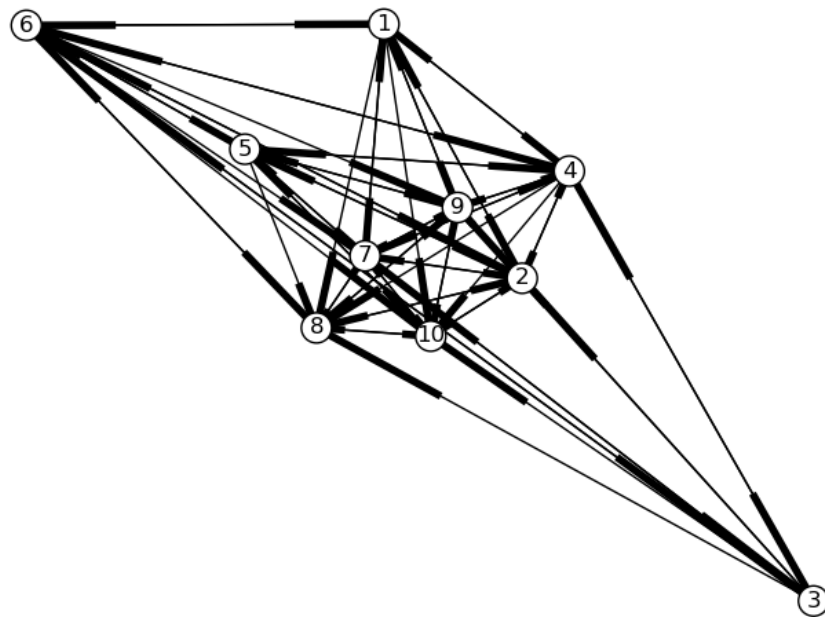
F1-score, Precision and Recall for GRN of size 30 with max lag 5

15

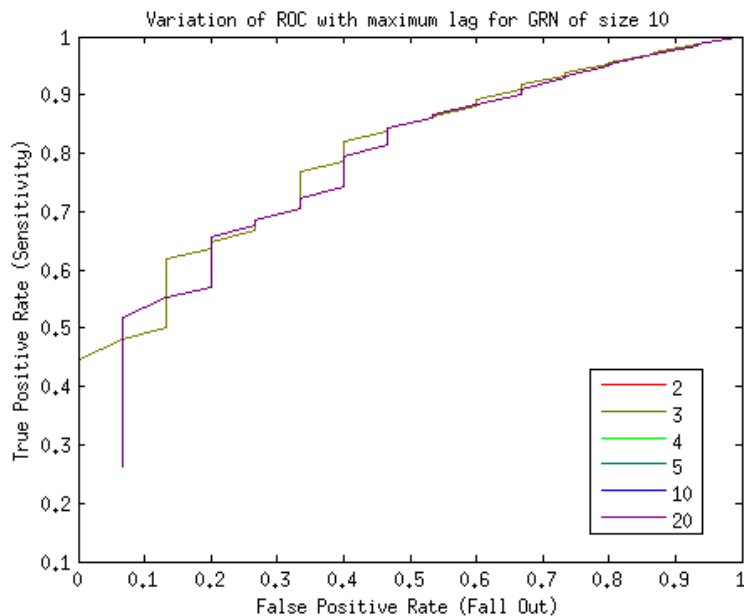# Variation in Alpha (Bonferroni Correction)

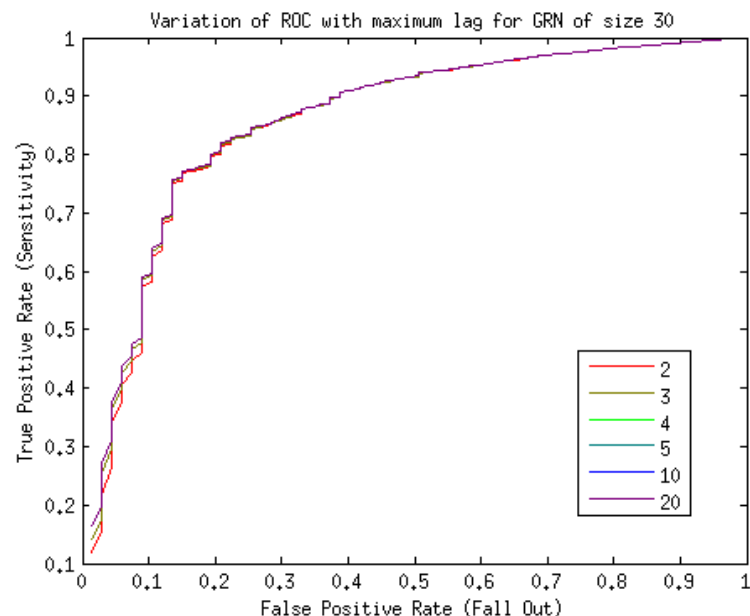GRN of size 10 with max lag 5 and alpha = 0.1

GRN of size 10 with max lag 5 and alpha = 0.5

Variation in Maximum Lag

ROC curve for GRN of size 10 with varying alpha, for different values of maximum lag

ROC curve for GRN of size 30 with varying alpha, for different values of maximum lag

17

Key Findings - Pairwise GC Analysis

- Pros

    - Simple approach

    - Validation can improve precision

- Cons

    - High time complexity

    - Assumes linearity

    - False positives

    - Fails to provide a holistic view

The Way Ahead - Transfer Entropy

$$T_{X \rightarrow Y} = H(Y_t | Y_{t-1:t-L}) - H(Y_t | Y_{t-1:t-L}, X_{t-1:t-L})$$
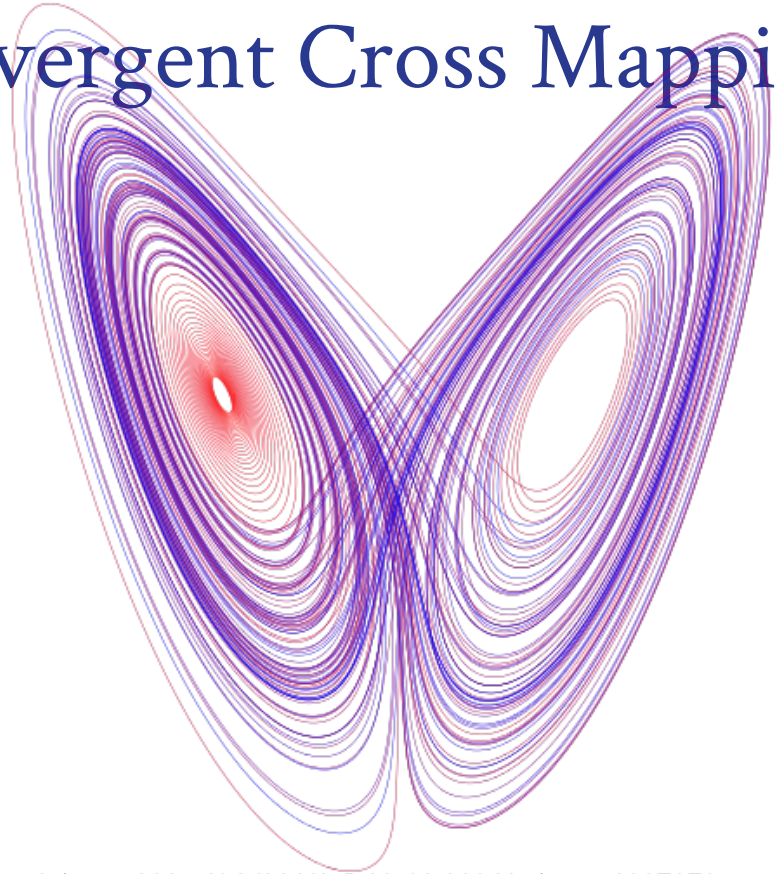
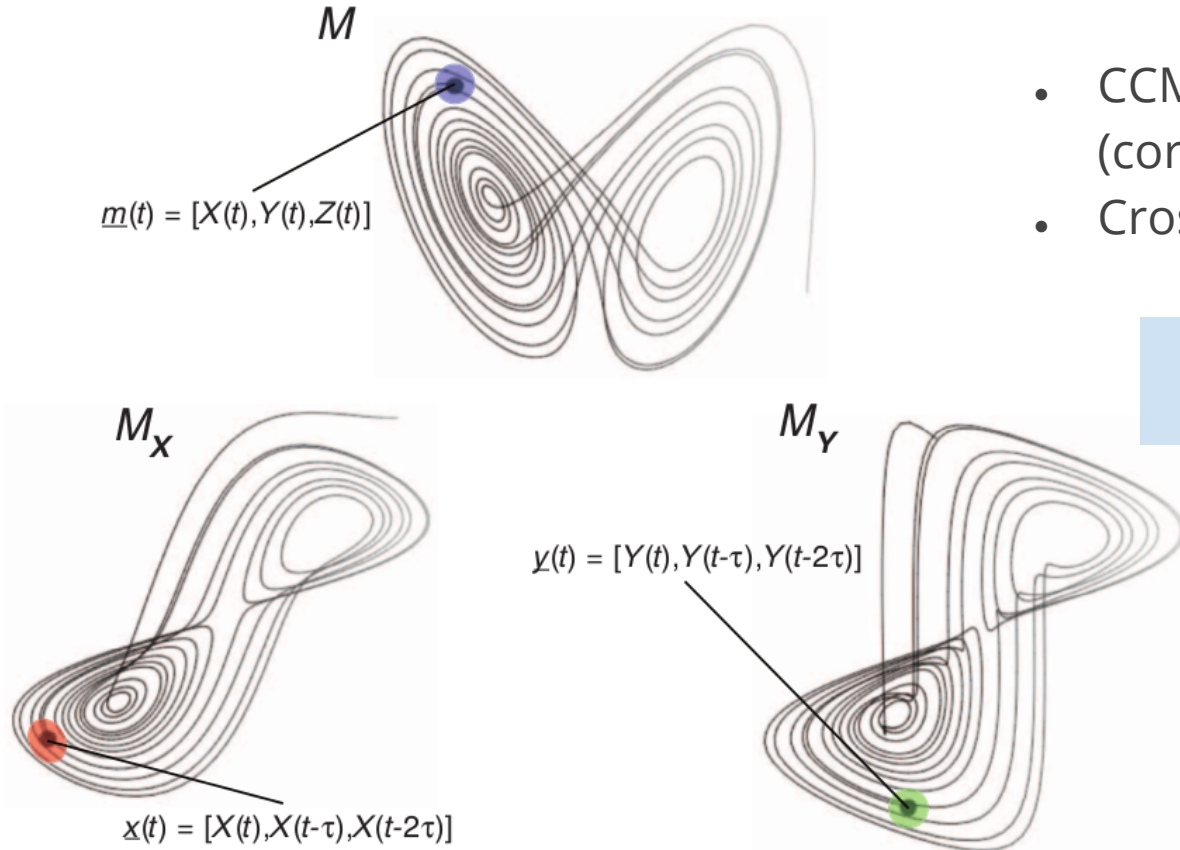| Conditional Entropy (over itself) | - | Conditional Entropy (over another signal) |

= Conditional Mutual Information

The Way Ahead - Convergent Cross Mapping

- Granger Causality doesn't work in

  - deterministic dynamic systems,

  - with weak to moderate coupling,

  - amongst inseparable factors.

- Not competing with GC

- New notion of causality →
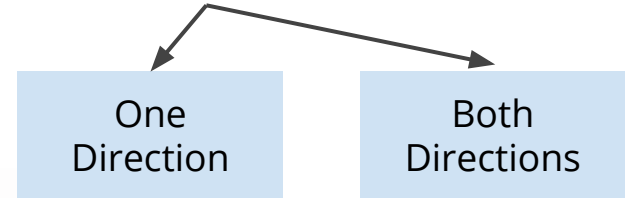 Belong to same dynamical system

The Way Ahead - Convergent Cross Mapping

M

$\underline{m}(t) = [X(t),Y(t),Z(t)]$

M$_X$

$\underline{x}(t) = [X(t),X(t-\tau),X(t-2\tau)]$

M$_Y$

$\underline{y}(t) = [Y(t),Y(t-\tau),Y(t-2\tau)]$

- CCM involves convergence (correlation =/= causation)
- Cross-map in:

| One Direction | Both Directions |

$X \rightarrow Y$
We can predict X from Y

The Hybrid Approach

Information Theory

Theory of Manifolds

✗ For stochastic, non-linear systems

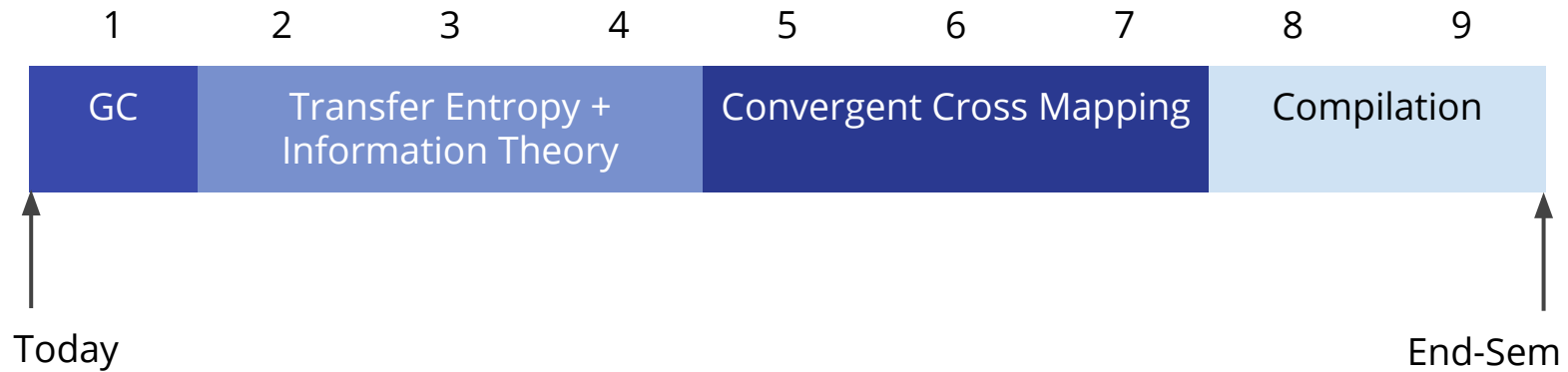✓ Can differentiate, through thresholding

✓ For possibly synergistic, deterministic systems

✗ No difference between first order and transitive causality

Project Timeline

Thank You

Questions?