Interpretable Neural Network Models for Granger Causality Discovery

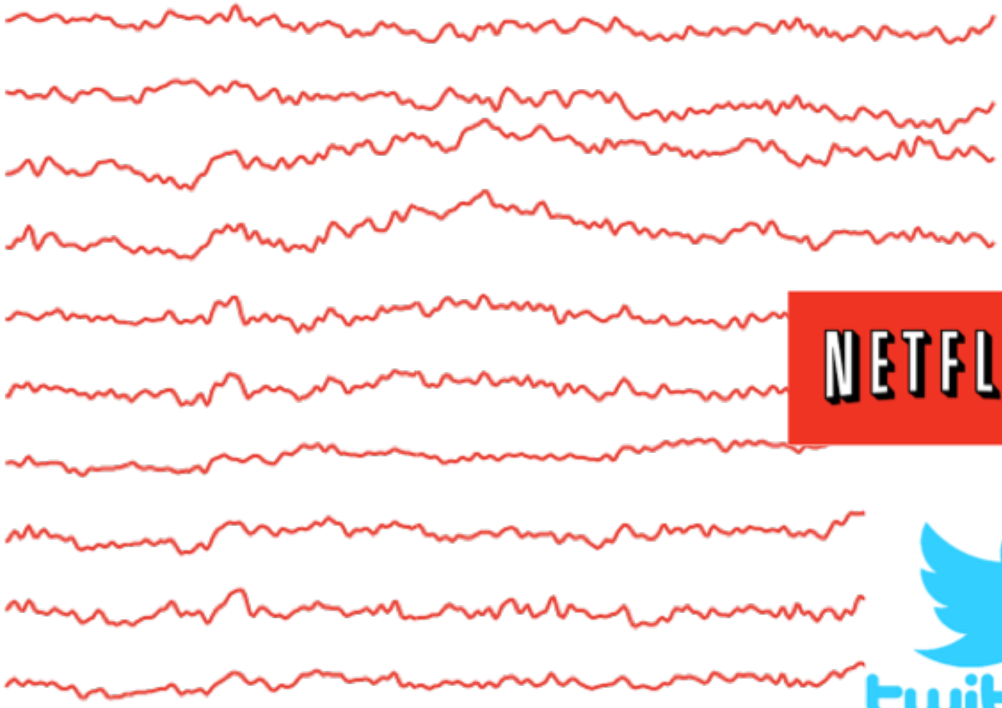
Emily Fox

University of Washington

Amazon Professor of Machine Learning

CSE and Statistics

Modern sources of time series



NETFLIX

facebook

twitter

flickr

Insta

PANDORA

ebay

amazon

Until recently, ML (mostly) ignored time series

It's hard!

parameters (naively) grows rapidly with
- # of series
- complexity of dynamics captured

More data

Algorithms more computationally intensive

More compute

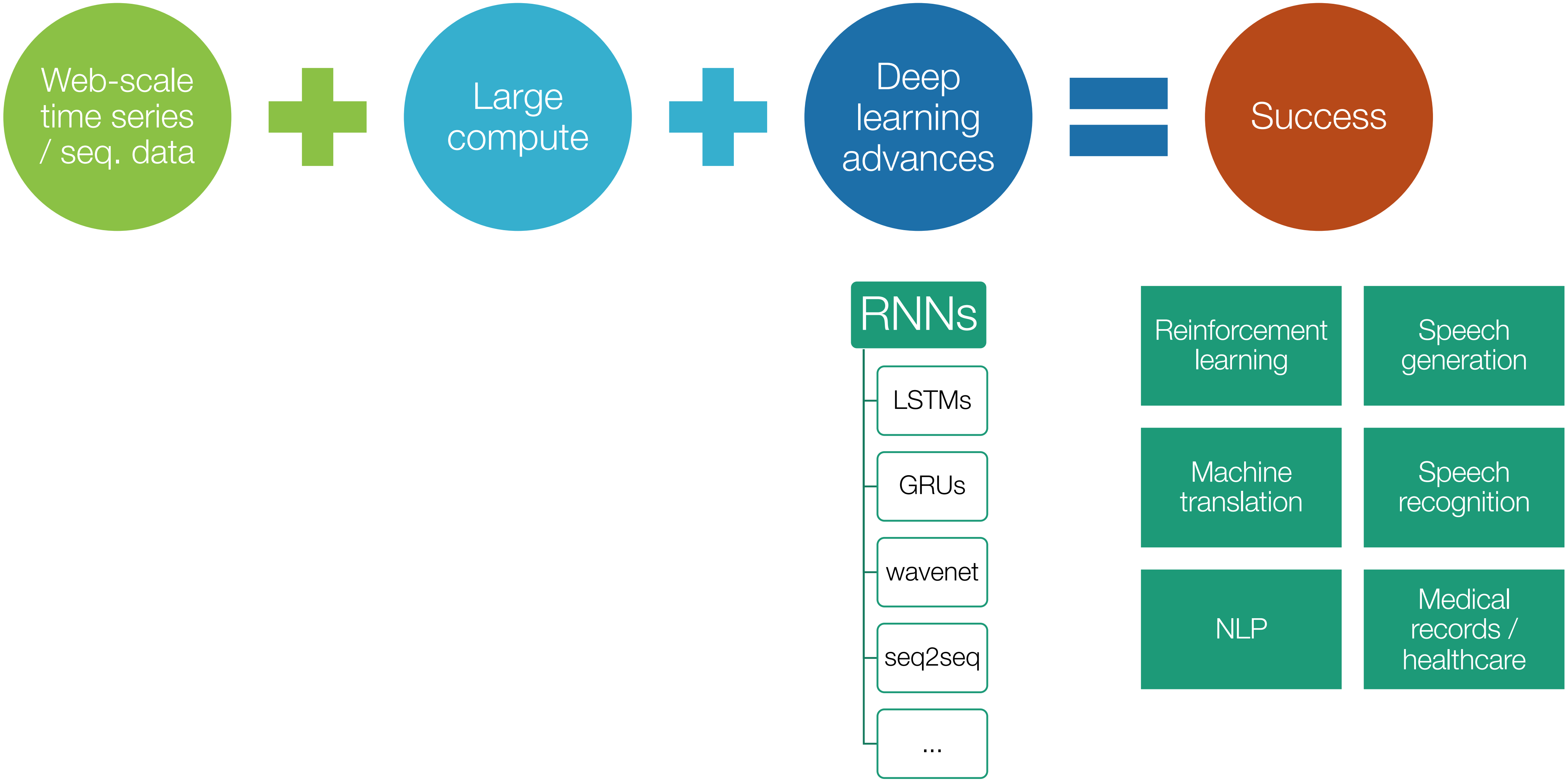
Theory not applicable because typically assume no time dependencies

Importance of modeling dynamics

	Independent observations
Accuracy	50%

With dynamic model, can get **improved prediction accuracy**

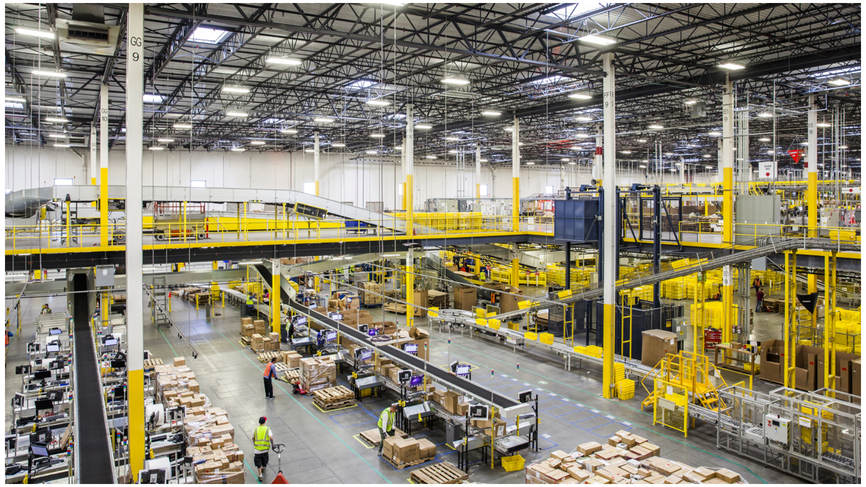
Now time series **are** “in”



But, success also relies on…

Lots of replicated series

- Lots of correspondence data
- Lots of trials of a robot navigating every part of the maze
- Lots of transcribed audio



Demand forecasting of new item:
Tons of data, but not for question of interest

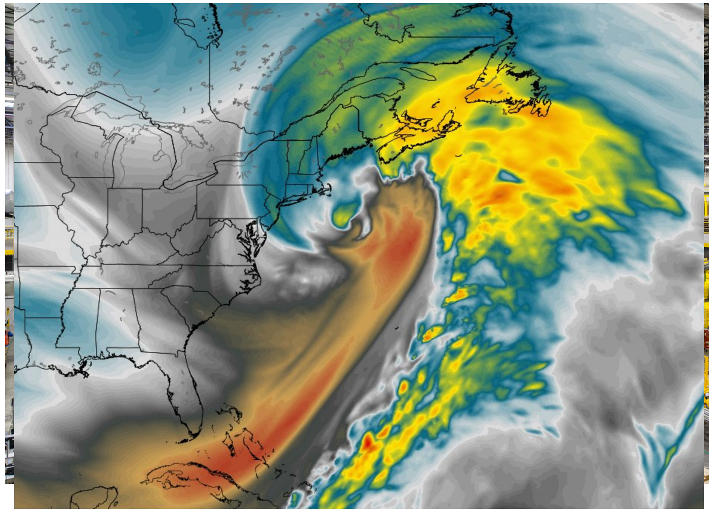
But, success also relies on…

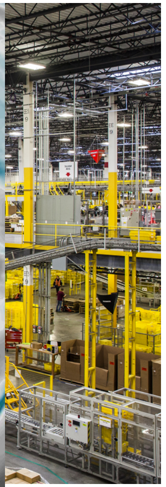
Lots of replicated series

- Lots of correspondence data
- Lots of trials of a robot navigating every part of the maze
- Lots of transcribed audio

Manageable contextual memory

- Seen this structure in a maze before
- Seen these words in this context before
- Seen patient with these symptoms and test results before





Extremely complicated context:

air temperature, dew point, relative humidity, wind direction, wind speed, altimeter, sea level pressure, precipitation, visibility, wind gust, cloud coverage, cloud height, present weather code

Demand forecasting of new item:

Tons of data, but not for question of interest

But, success also relies on…

Lots of replicated series	• Lots of correspondence data • Lots of trials of a robot navigating every part of the maze • Lots of transcribed audio
Manageable contextual memory	• Seen this structure in a maze before • Seen these words in this context before • Seen patient with these symptoms and test results before
Clear prediction objective	• Word error rate for speech recognition • BLEU score for machine translation • Reward function in reinforcement learning

Beyond prediction on big data

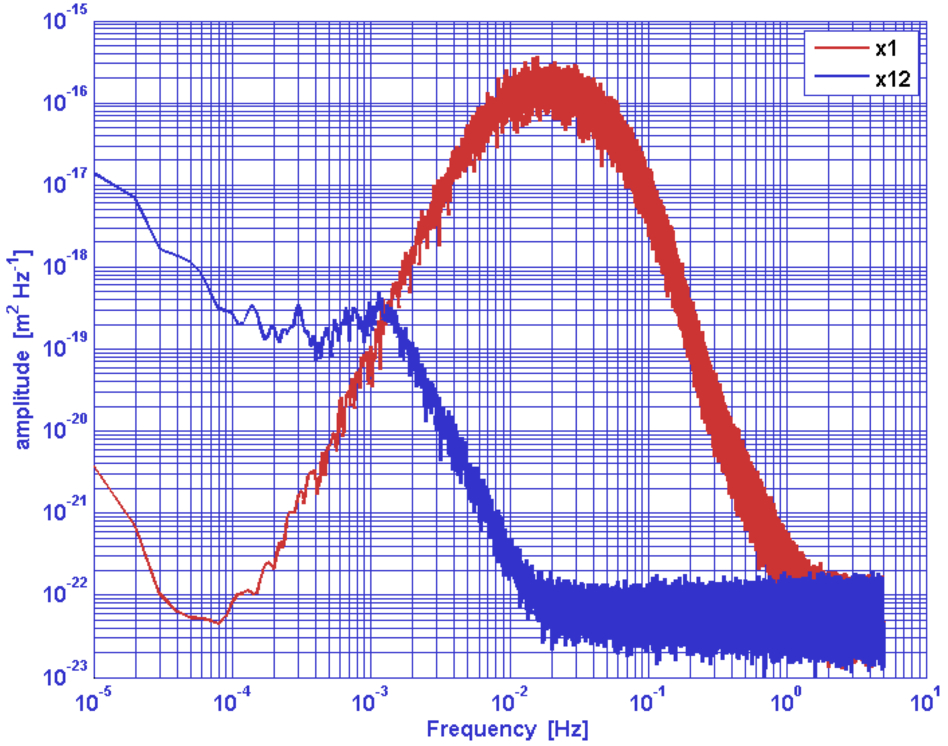
Characterizing dynamics

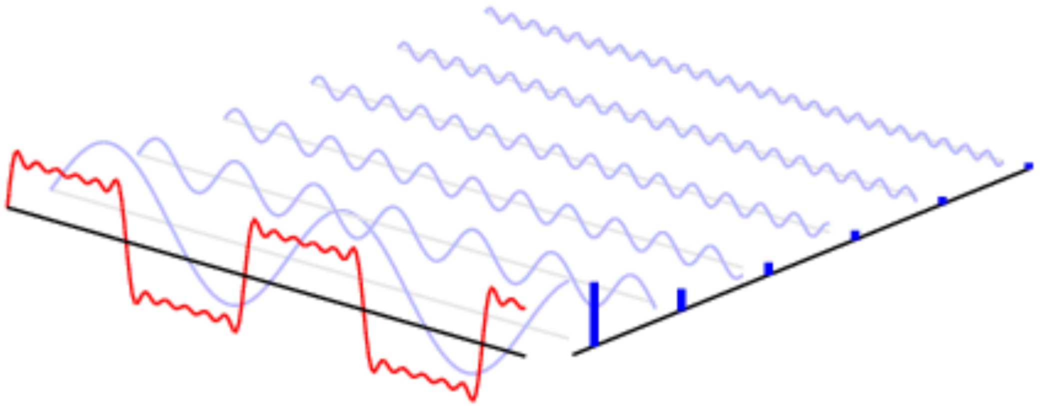
Efficiently sharing information

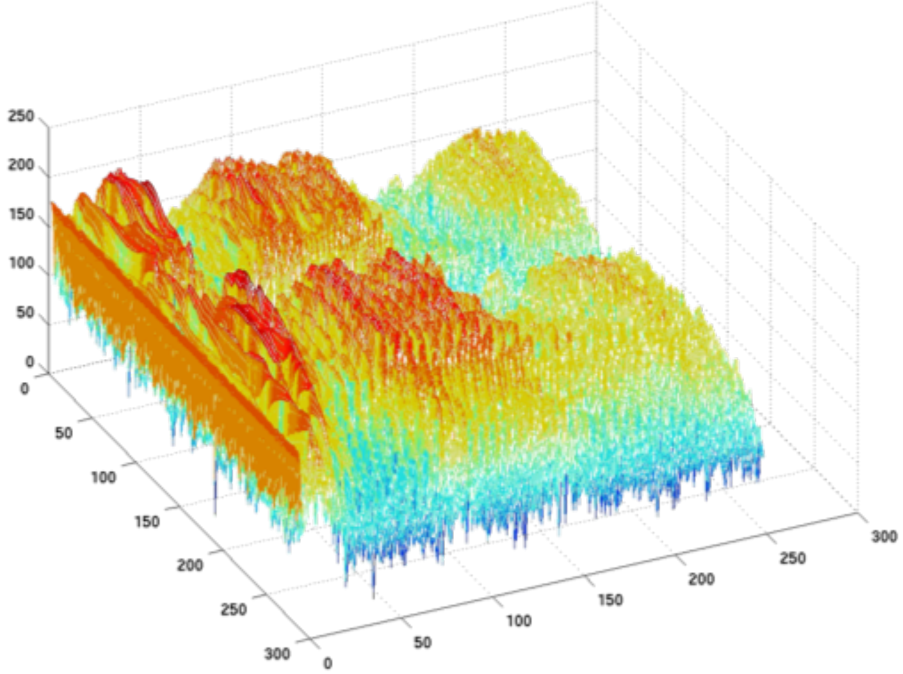
Interpretable interactions

Non-stationarity & measurement bias

Spectral analysis



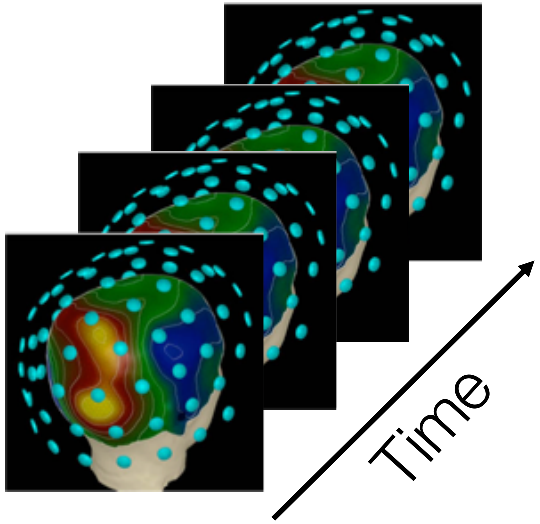


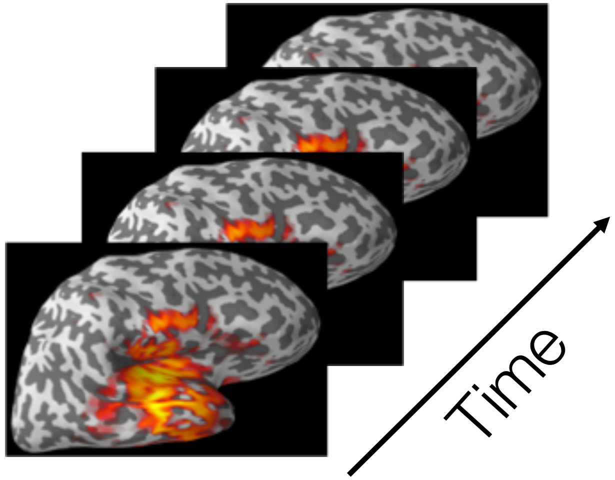


- **Frequency domain** analysis
- Local stationarity
- Time-frequency analysis

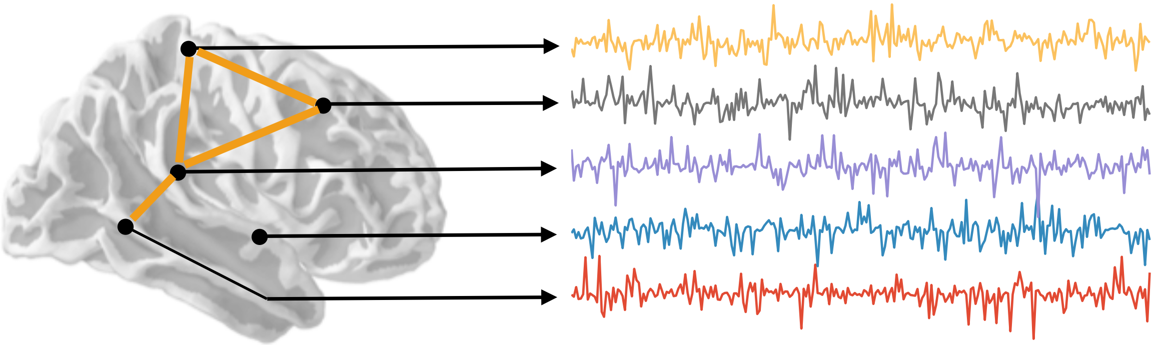
Spectral analysis of neuroimaging data

Magnetoencephalography (MEG) data of brain activation over time





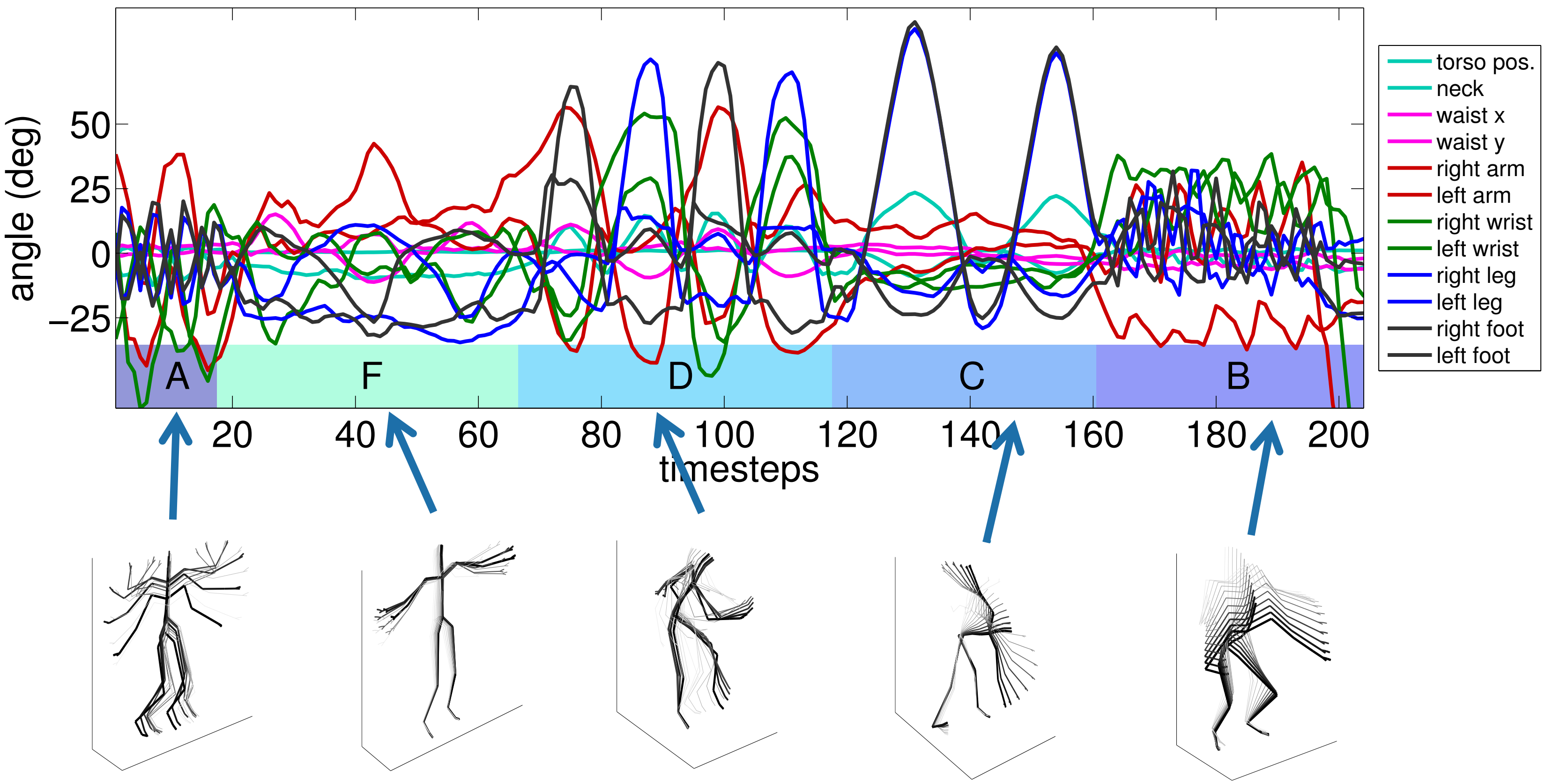
Goal:
Infer functional connectivity



Discovering human motion behaviors

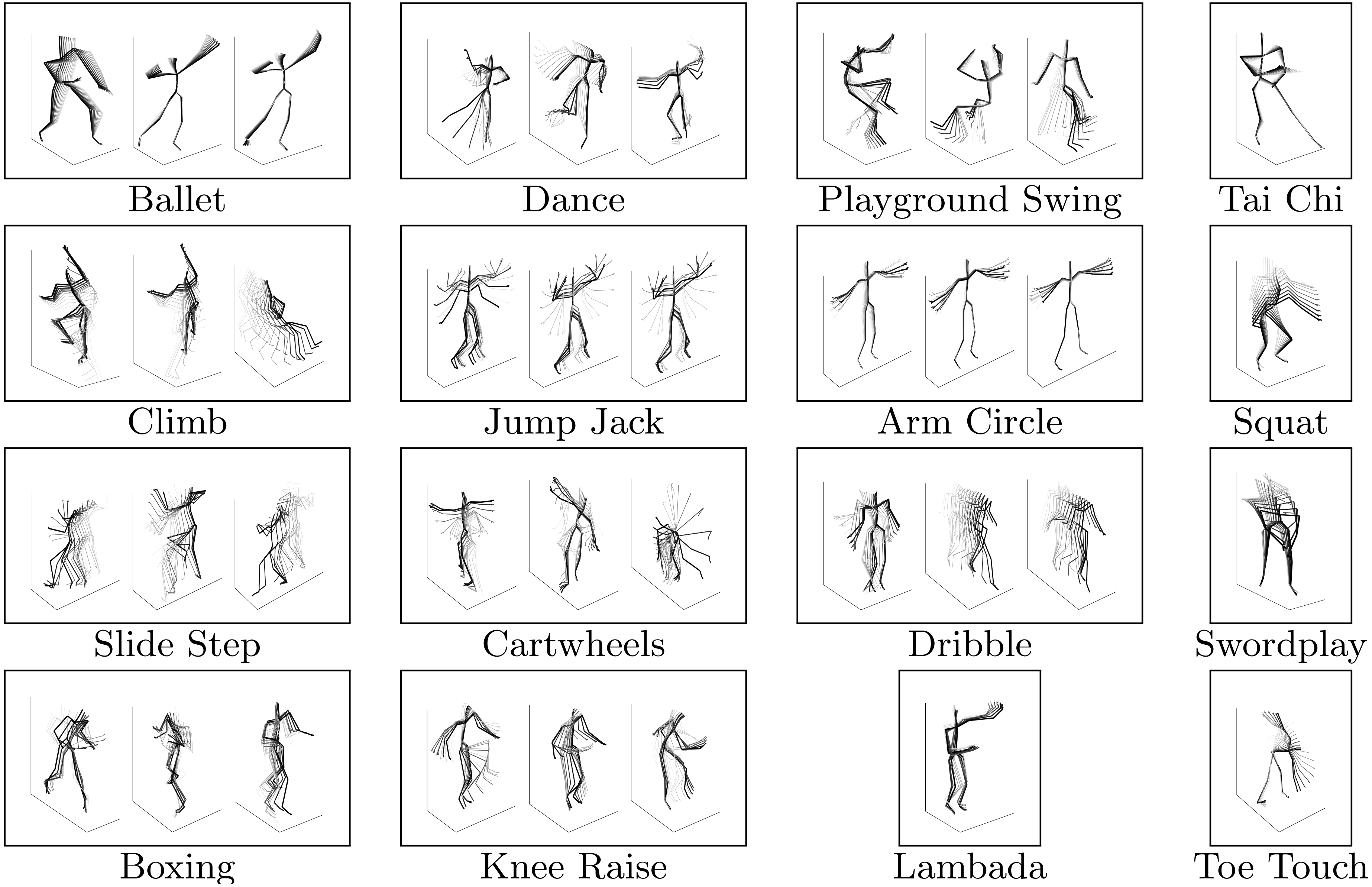
Parse videos into underlying behaviors **without training labels**

Recording modeled as **switches between simple behaviors**



Fox, Hughes, Sudderth, and Jordan, AoAS 2014

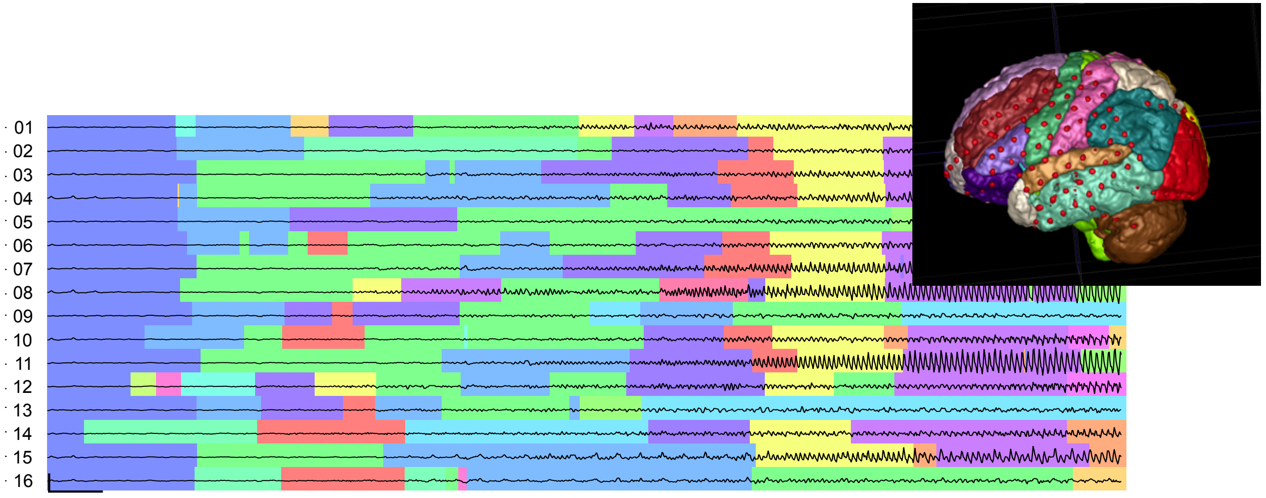
Automatically parsing large collections

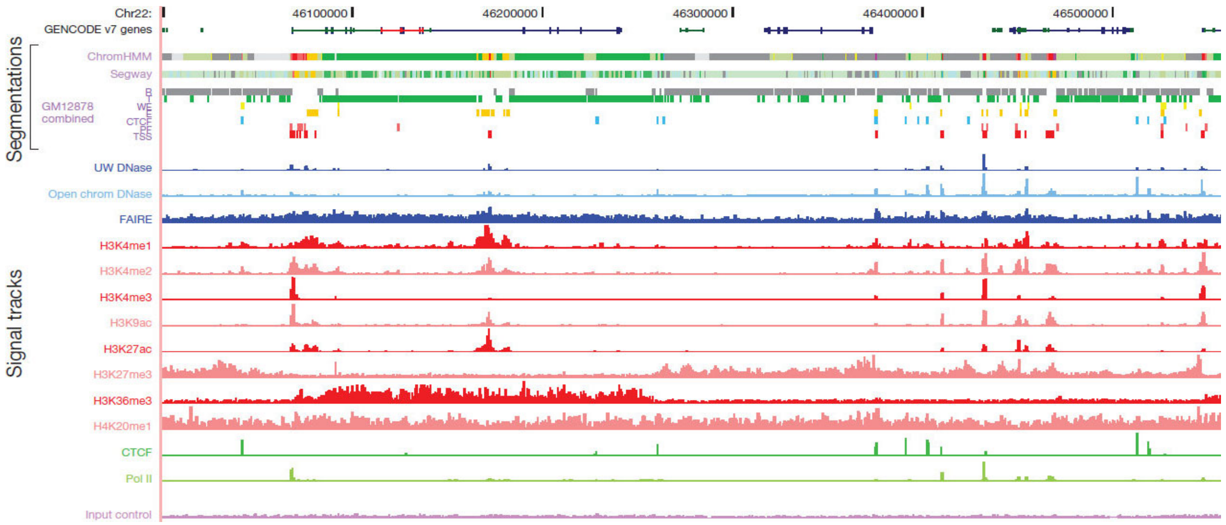


Ideas appear in many domains…

Example applications:

- Parsing EEG recordings
- Speech segmentation
- Volatility regimes in financial time series
- Genomics
- …





Beyond prediction on big data

Characterizing dynamics

Efficiently sharing information

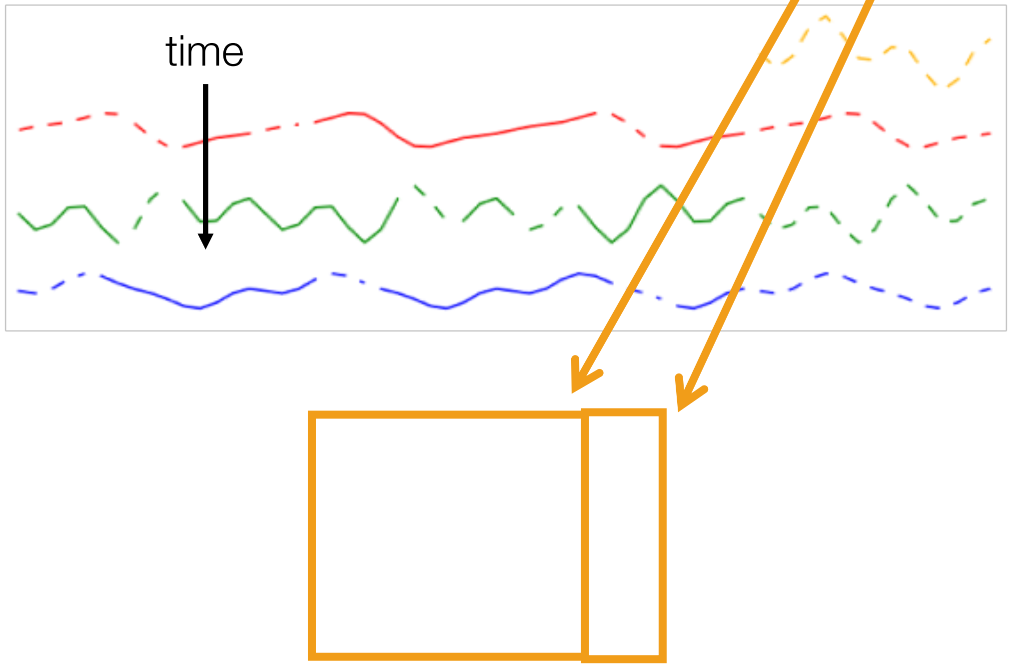
Interpretable interactions

Non-stationarity & measurement bias

Predicting demand for products

Long-range and cold-start forecasting

Long-range forecasting
Cold start (new product)



Prediction task with lots of data, but not much for question of interest

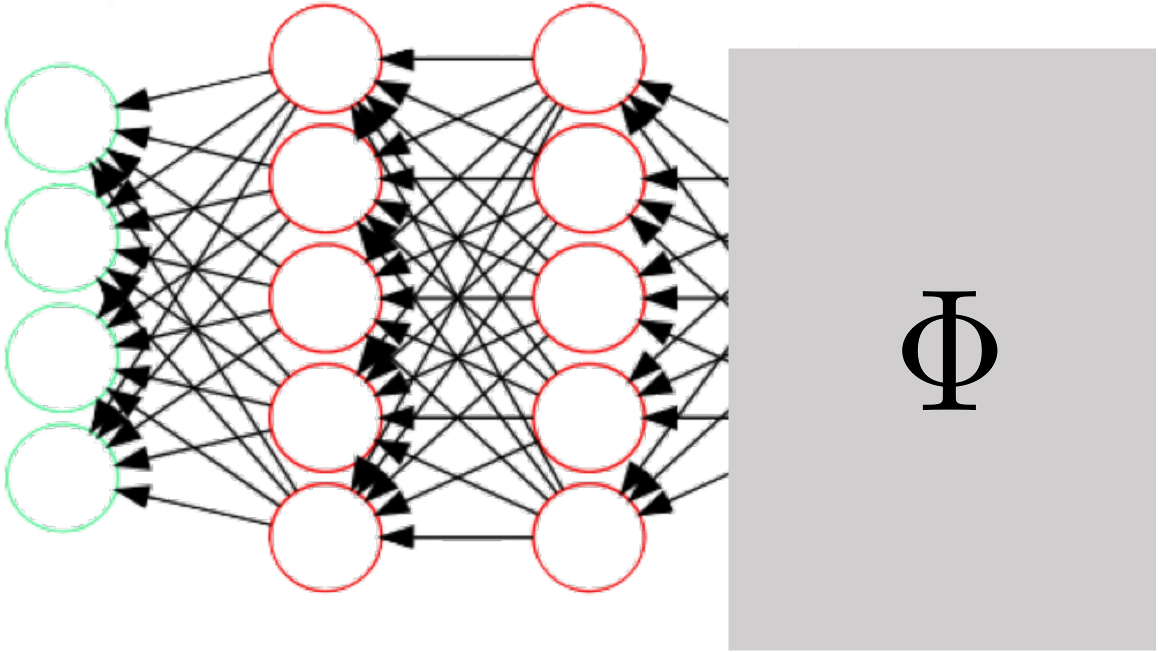
Leveraging low-dim structure + side info

Low-rank description of observed years of available products

Function approx. via neural network

Product-specific meta data

+



Xie, Tank, and Fox, *NIPS Time Series Workshop* 2016.

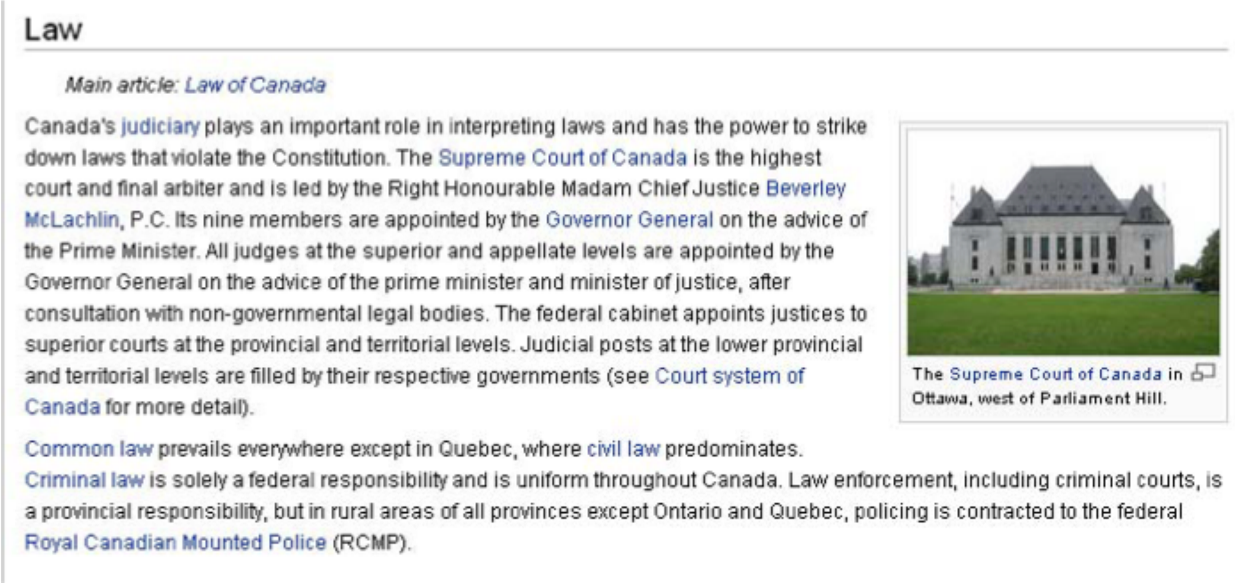
Analysis of Wikipedia data

4500 Wikipedia articles

Daily **page traffic** counts 2008-2014

Per article, 1 to 6 years of data
→ 29,000 columns

Features = tf-idf of article summary
→ 22,000 dimensions, but **sparse**



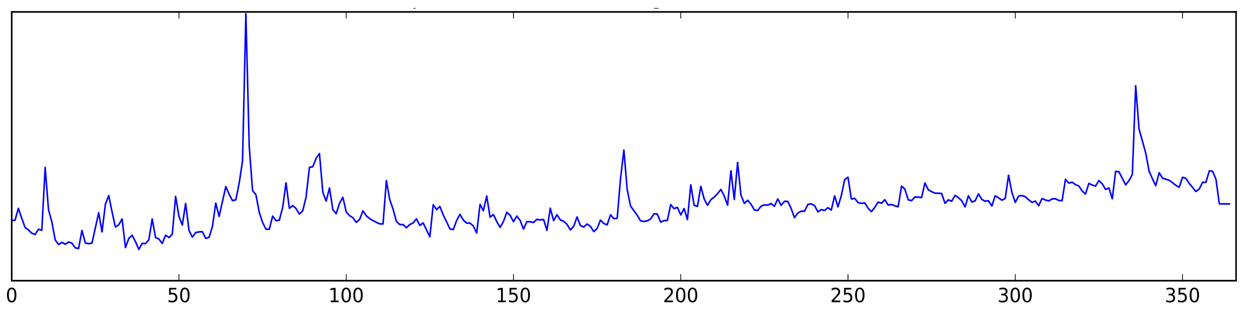
Law

Main article: Law of Canada

Canada's judiciary plays an important role in interpreting laws and has the power to strike down laws that violate the Constitution. The Supreme Court of Canada is the highest court and final arbiter and is led by the Right Honourable Madam Chief Justice Beverley McLachlin, P.C. Its nine members are appointed by the Governor General on the advice of the Prime Minister. All judges at the superior and appellate levels are appointed by the Governor General on the advice of the prime minister and minister of justice, after consultation with non-governmental legal bodies. The federal cabinet appoints justices to superior courts at the provincial and territorial levels. Judicial posts at the lower provincial and territorial levels are filled by their respective governments (see Court system of Canada for more detail).

The Supreme Court of Canada in Ottawa, west of Parliament Hill.

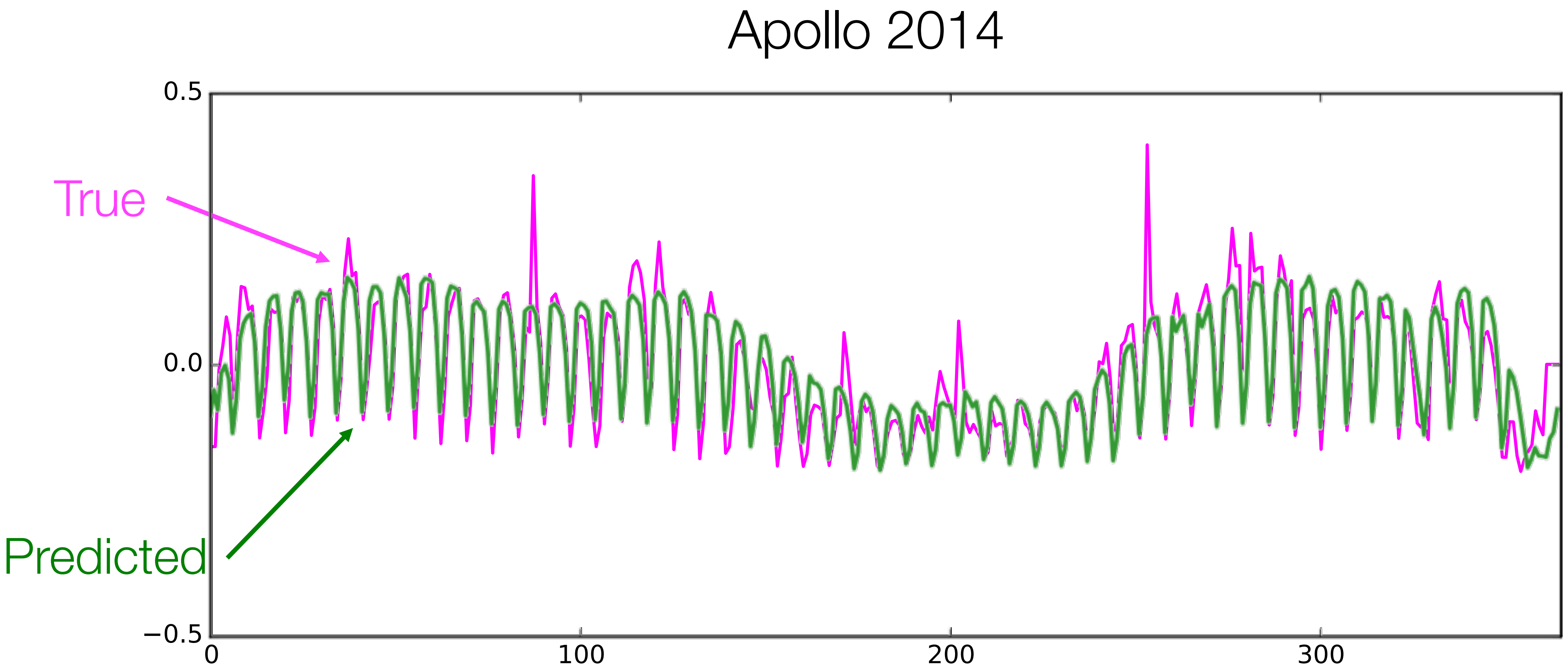
Common law prevails everywhere except in Quebec, where civil law predominates. Criminal law is solely a federal responsibility and is uniform throughout Canada. Law enforcement, including criminal courts, is a provincial responsibility, but in rural areas of all provinces except Ontario and Quebec, policing is contracted to the federal Royal Canadian Mounted Police (RCMP).



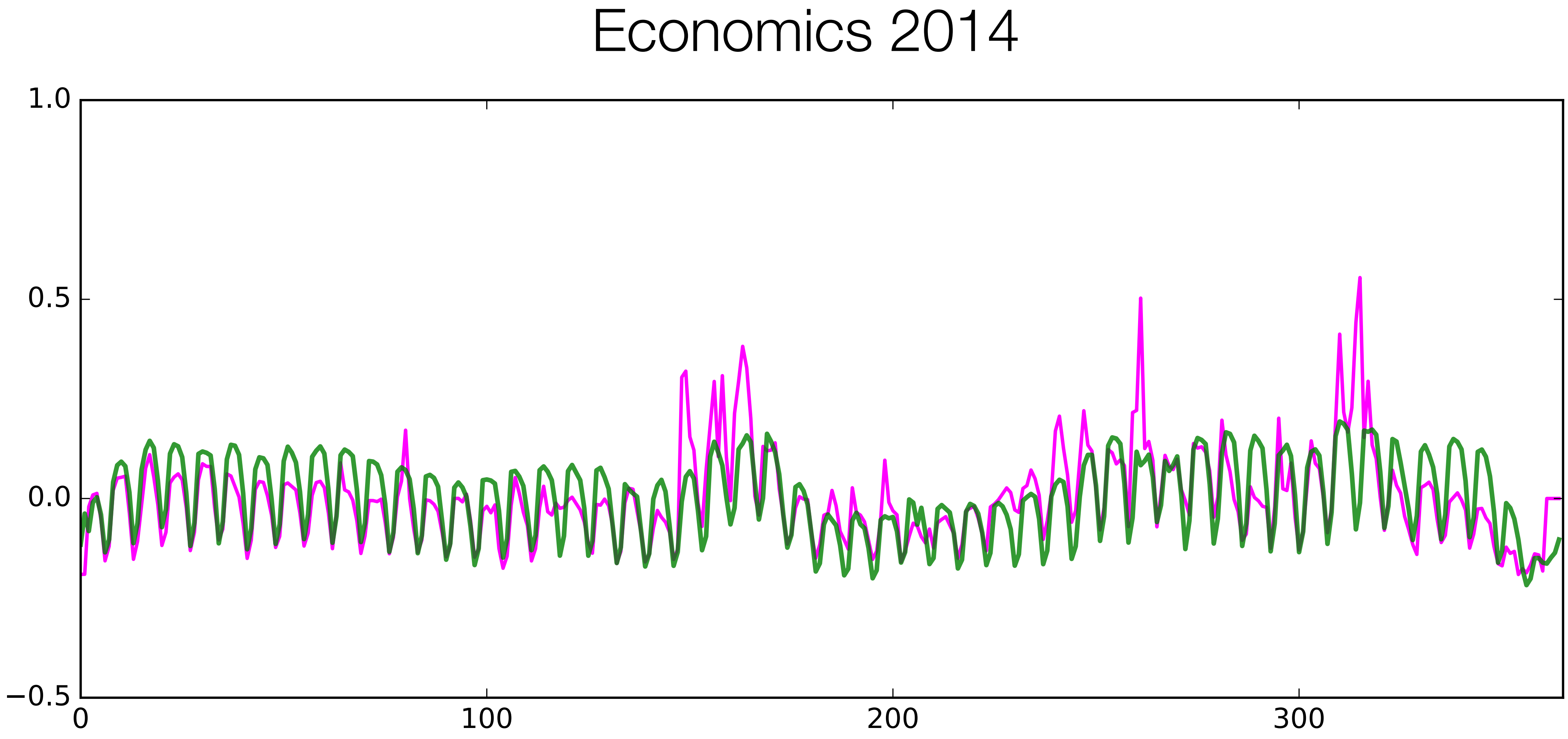
(Wiki Trends: Georges St-Pierre)

Xie, Tank, and Fox, *NIPS Time Series Workshop* 2016.

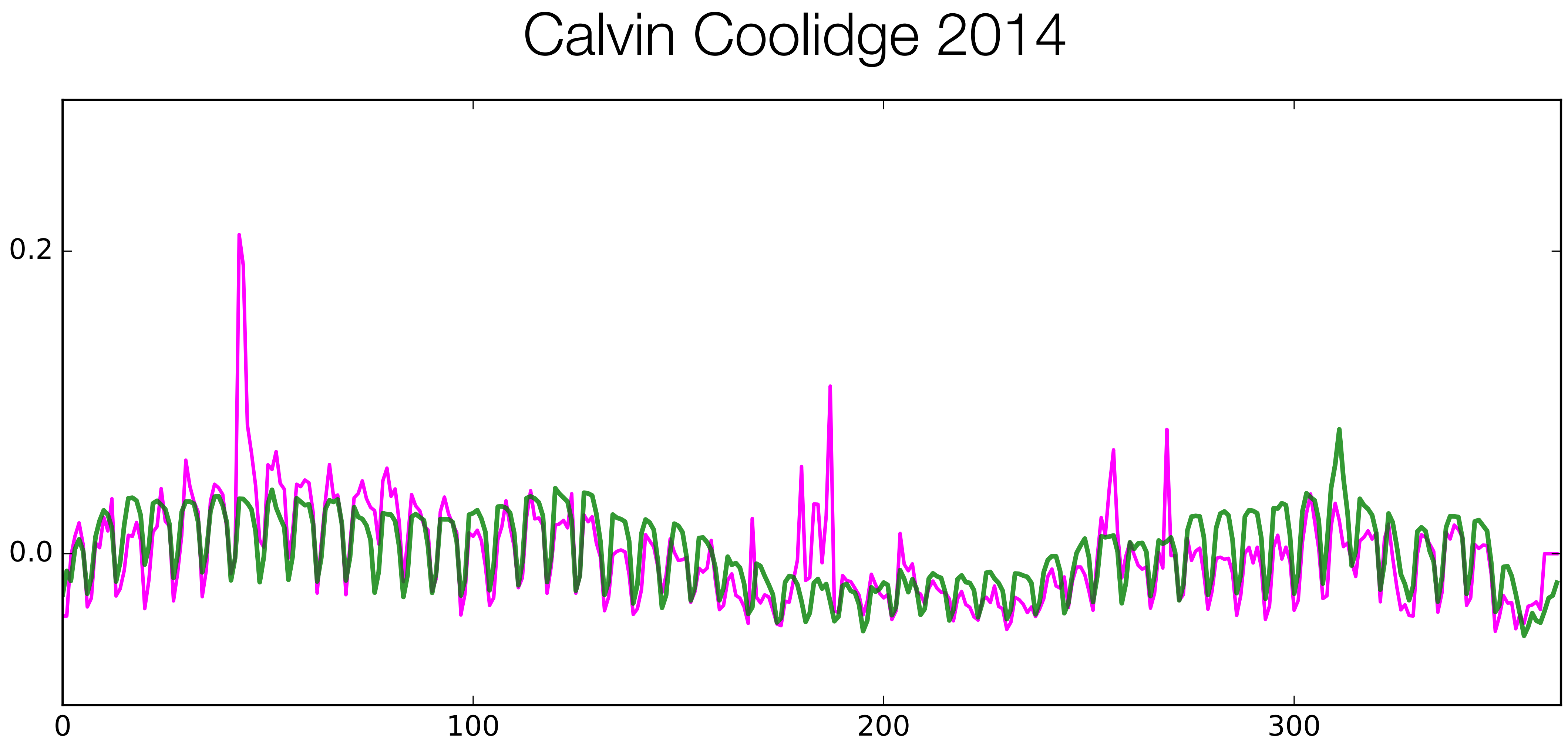
Long-range forecasts



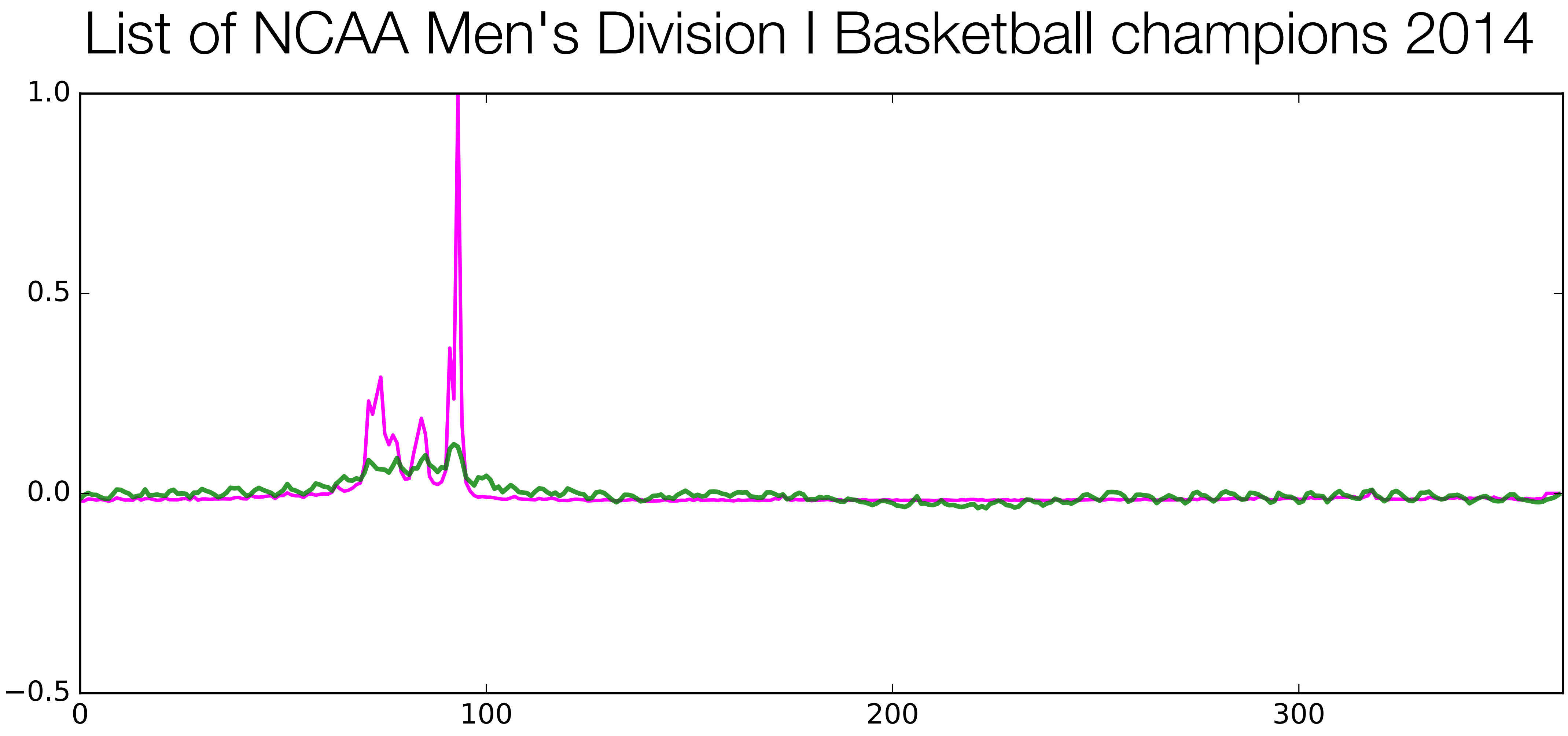
Long-range forecasts



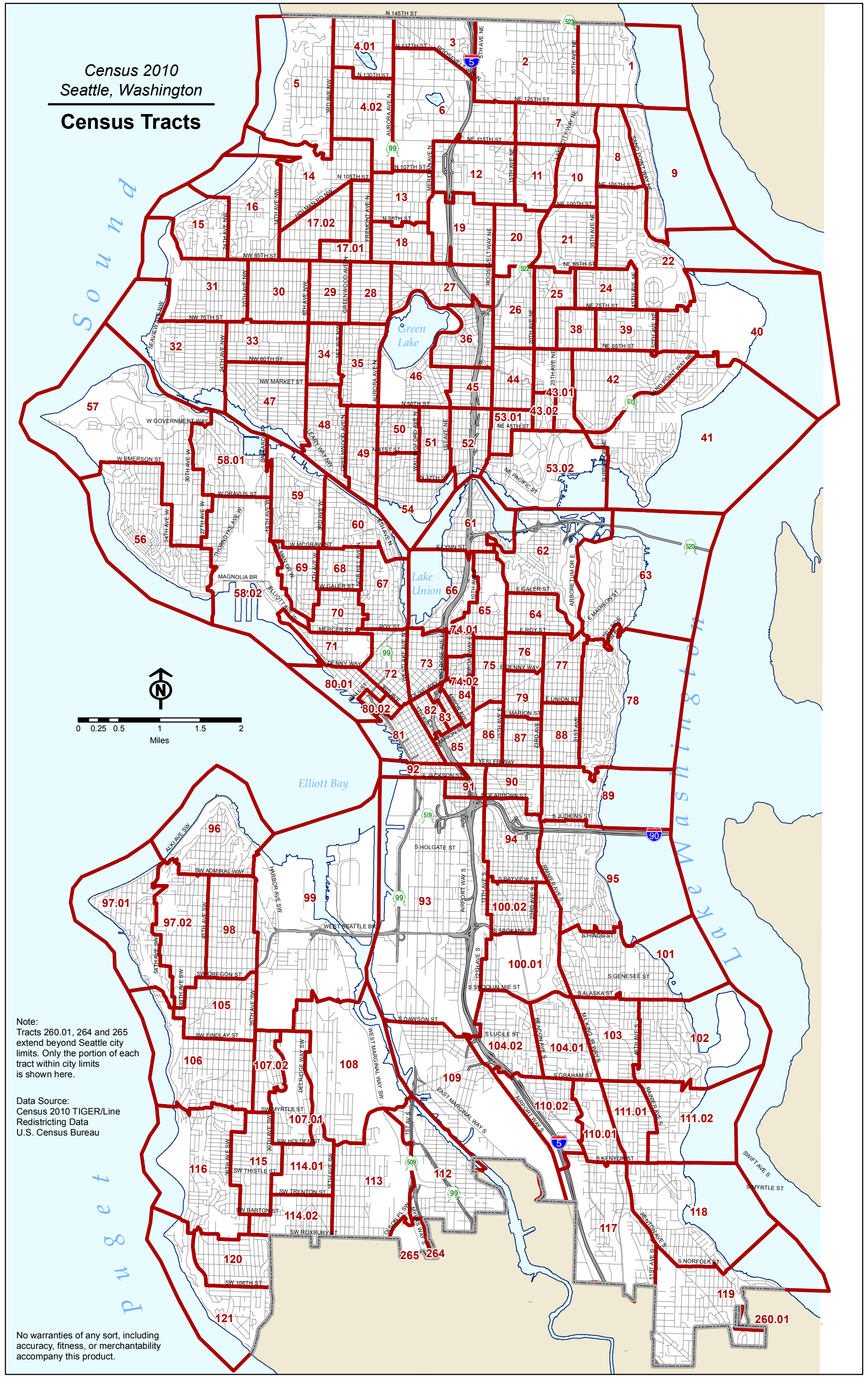
Cold start forecasts



Cold start forecasts



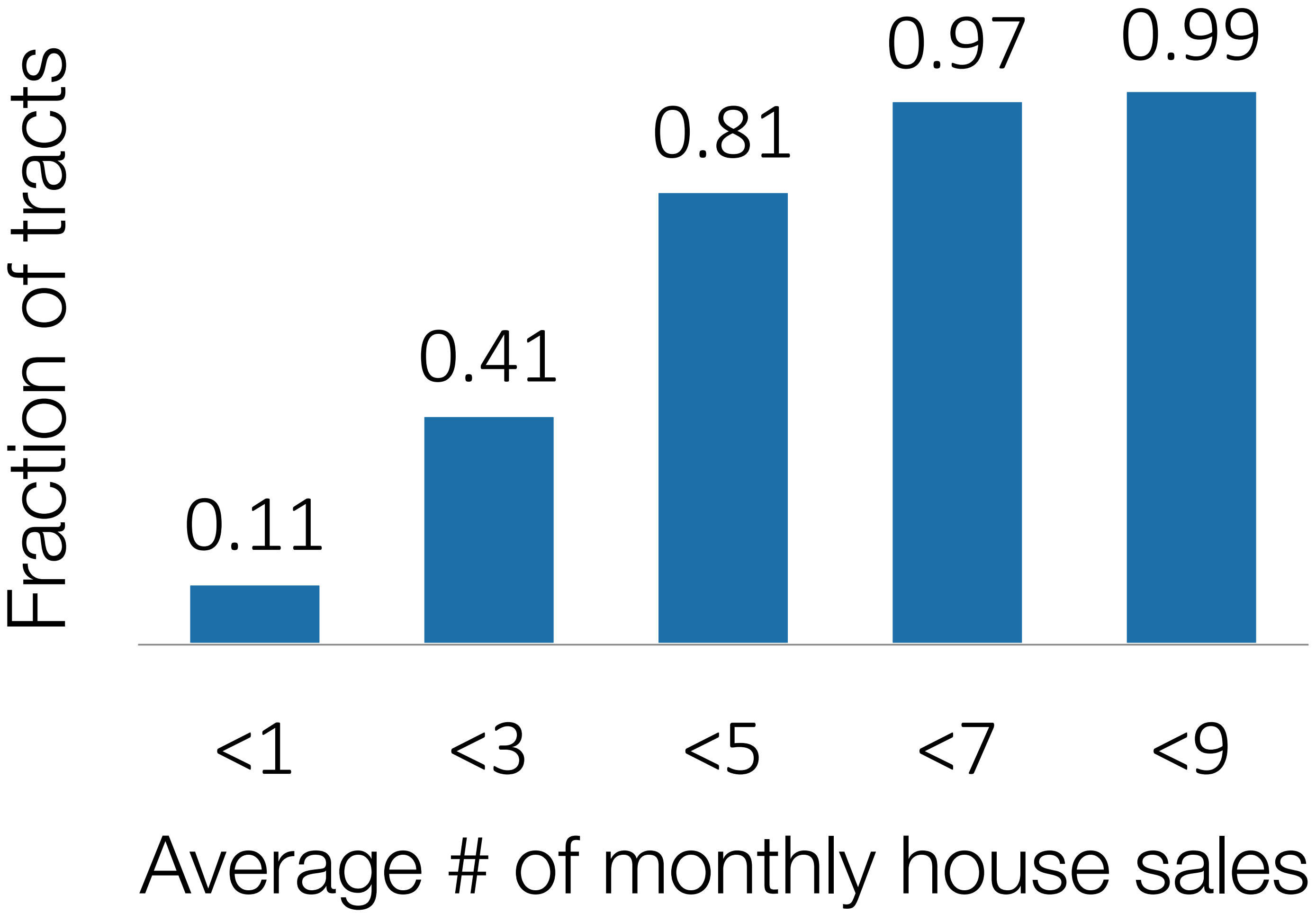
modeling a local housing index



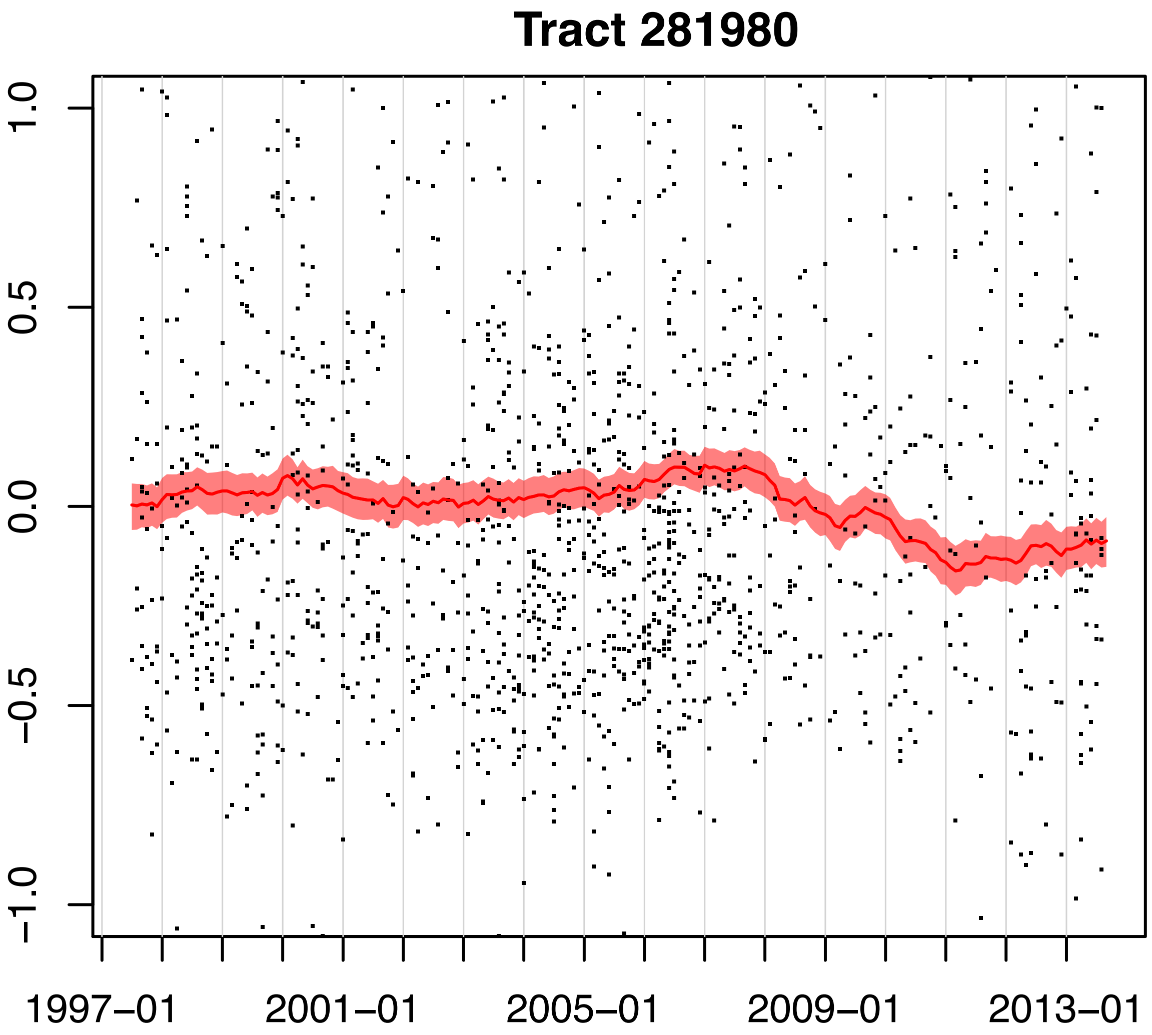
Census tracts in Seattle, WA

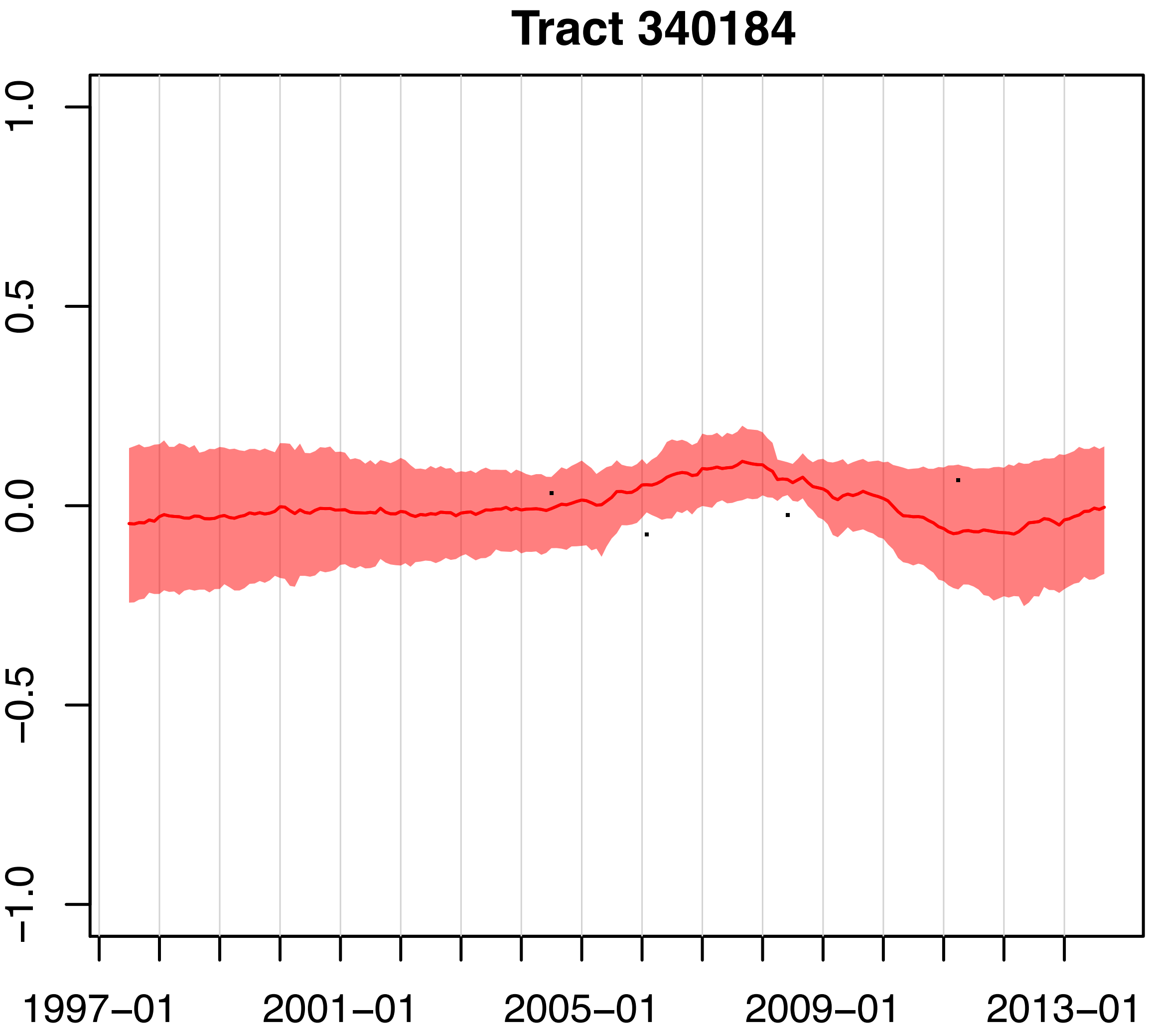
What is the value of housing in each region over time?

Challenge: Spatiotemporally sparse data



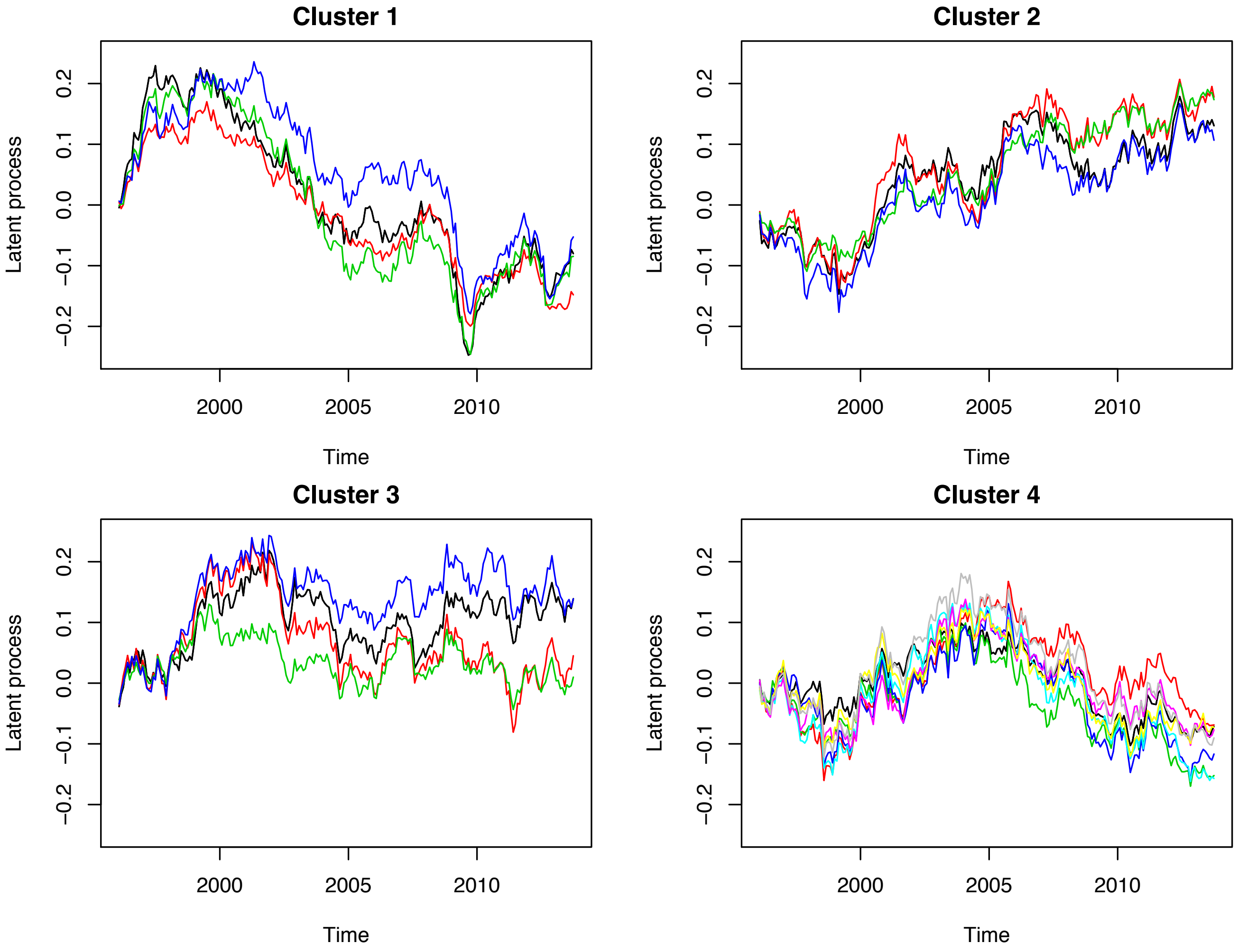
Challenge: Spatiotemporally sparse data





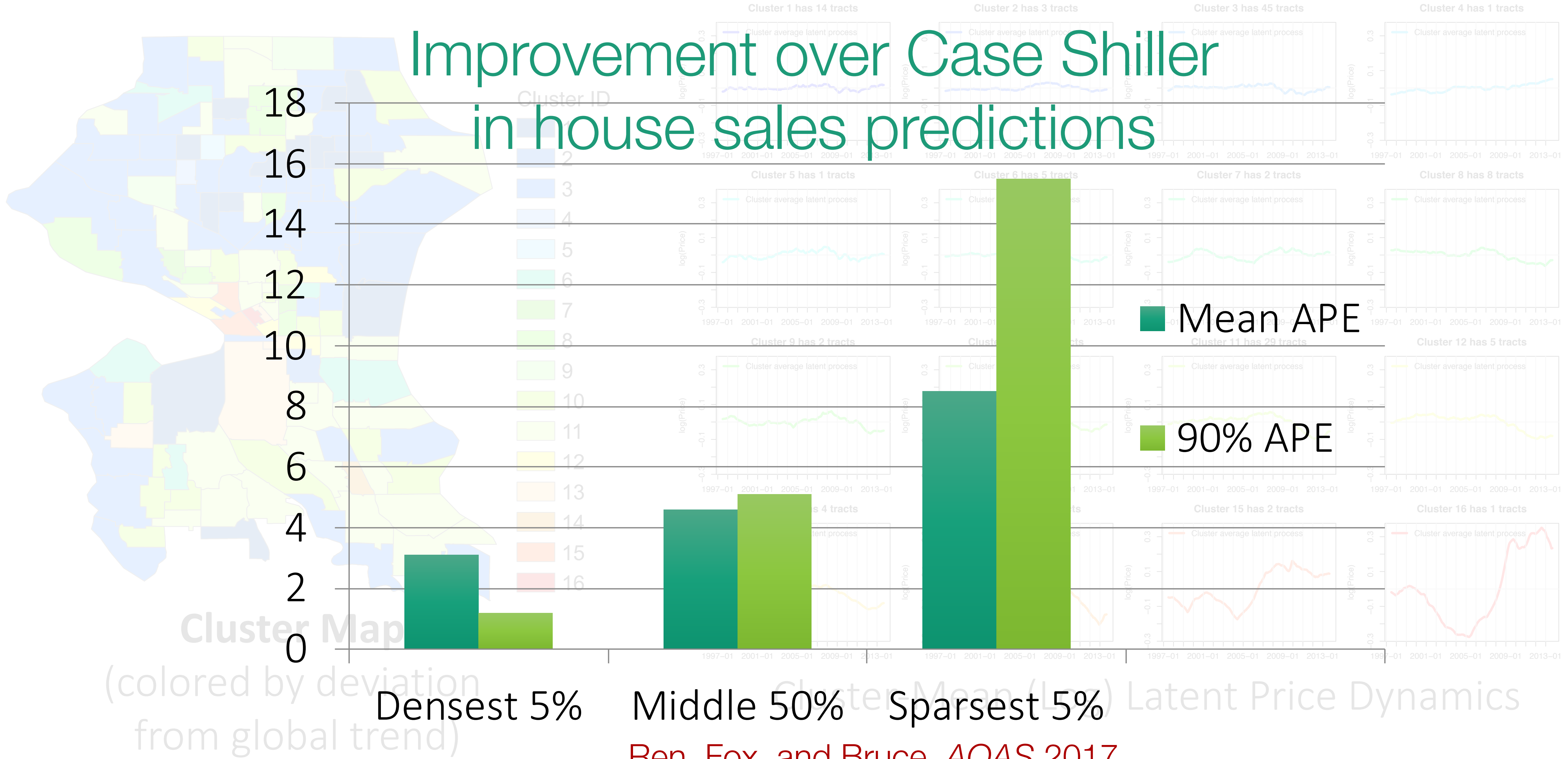
Solution: Cluster regions based on underlying price dynamics

Discover **groups** of tracts with **correlated dynamics**



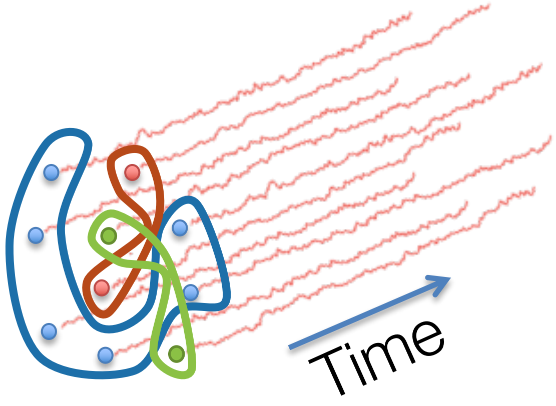
Ren, Fox, and Bruce, *AOAS* 2017

Seattle City analysis
(17 years, 140 tracts, 125k transactions)

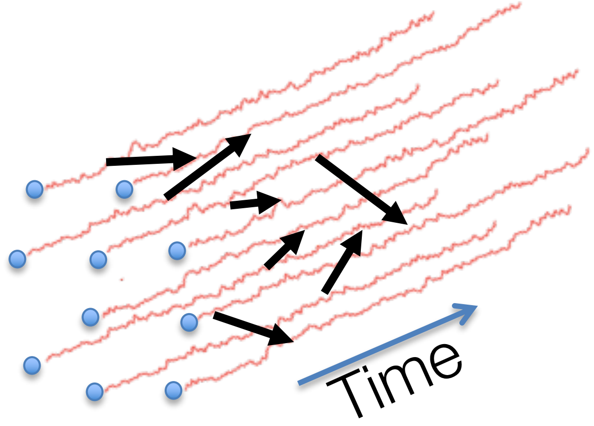


Ren, Fox, and Bruce, *AOAS* 2017

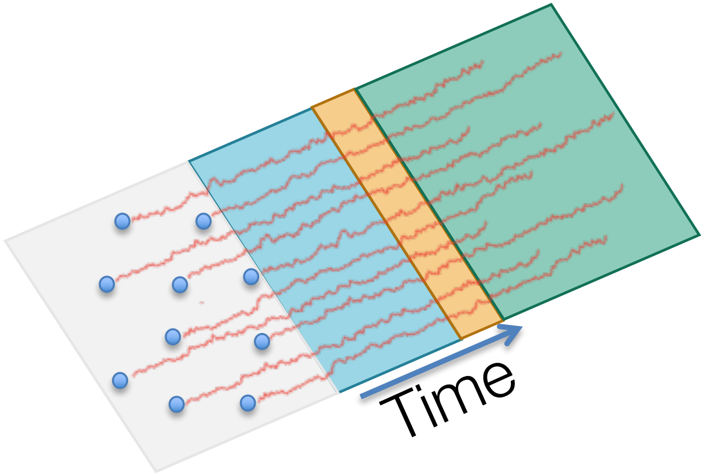
Recap: Mechanisms for coping with limited data



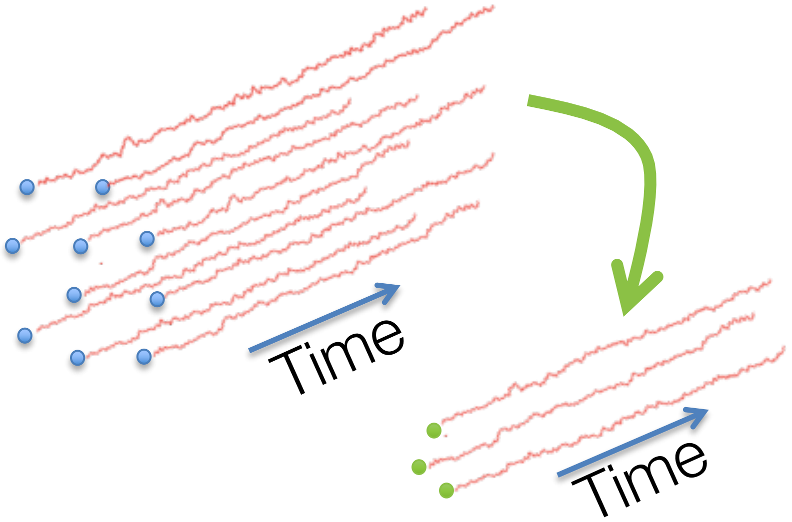
clusters and hierarchies



sparse directed interactions



switching between simpler dynamics



low-dimensional embeddings

Beyond prediction on big data

Characterizing dynamics

Efficiently sharing information

Interpretable interactions

Non-stationarity & measurement bias

Another **data-scarce** study: Dynamics of homelessness

- Counts occur on **single night** in January
- **Count method varies** from metro to metro and across time
- Observe most of those in shelters and **only a fraction of those on the streets**
- % sheltered varies largely between metros

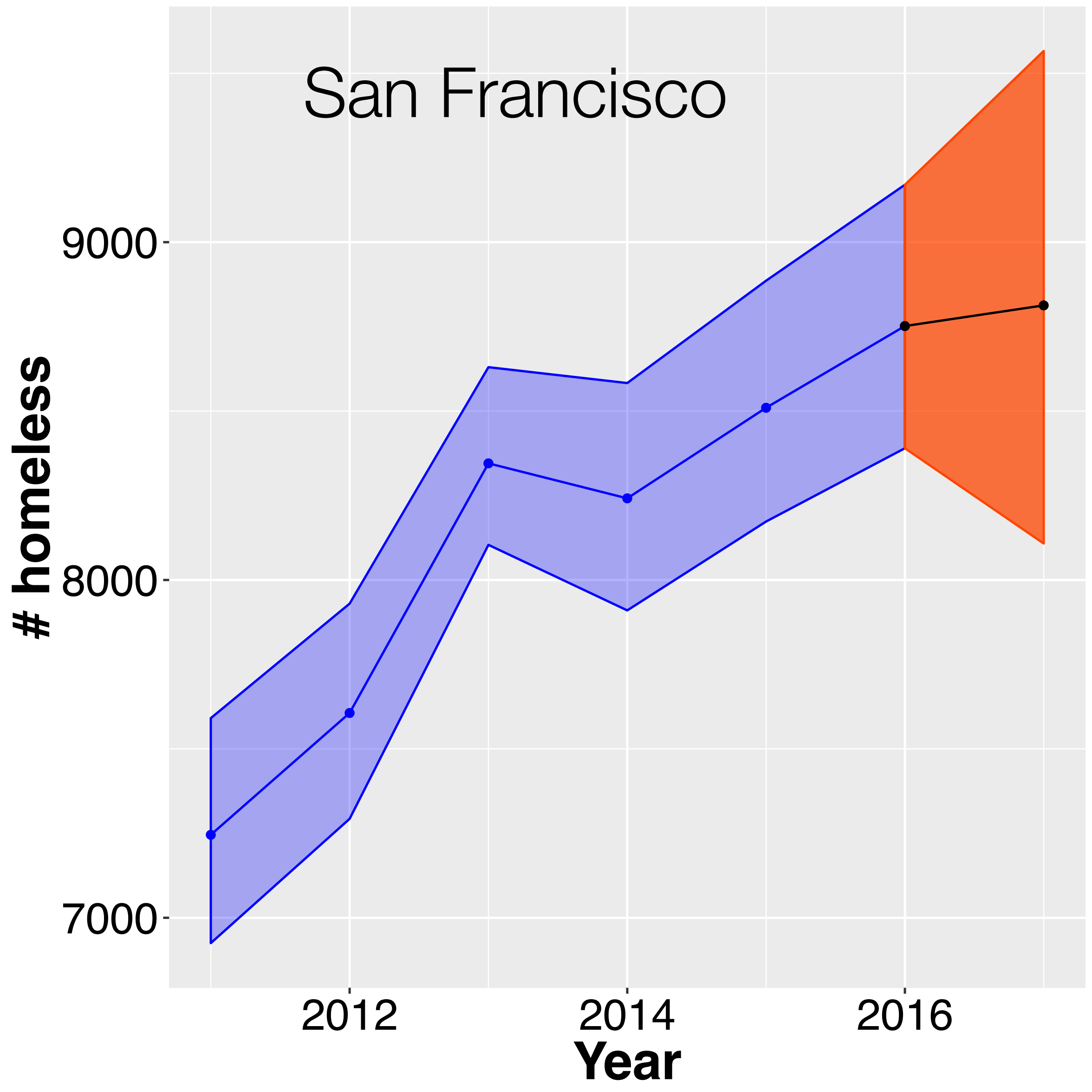
CHICAGO

2017

POINT-IN-TIME COUNT OF HOMELESS PERSONS

measurement bias!

What is the 1-yr-ahead forecast of homeless population?



Bayesian model-based approach accounts for:

- **Imperfect measurement mechanism** and changes in count quality
- Predicted increase in total population (**nonstationary process**)

Glynn and Fox, *arXiv* 2017

Beyond prediction on big data

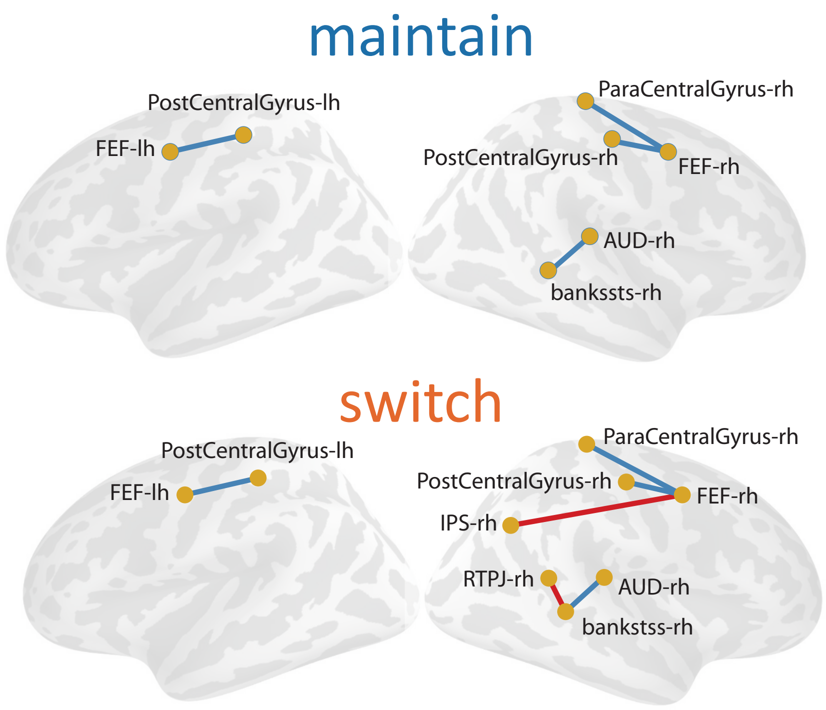
Characterizing dynamics

Efficiently sharing information

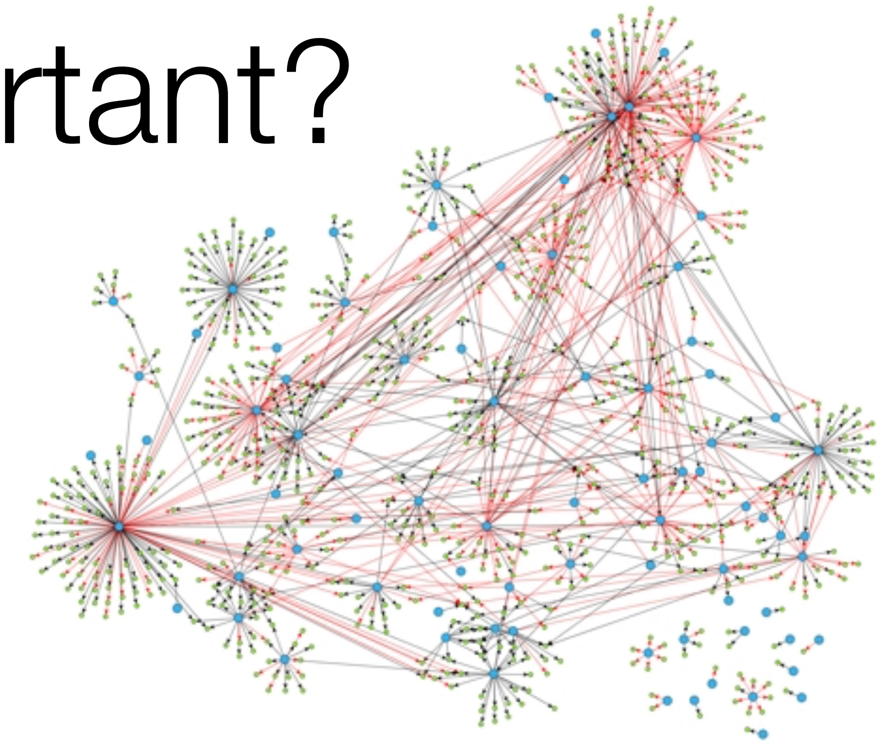
Interpretable interactions

Non-stationarity & measurement bias

Why are interactions important?

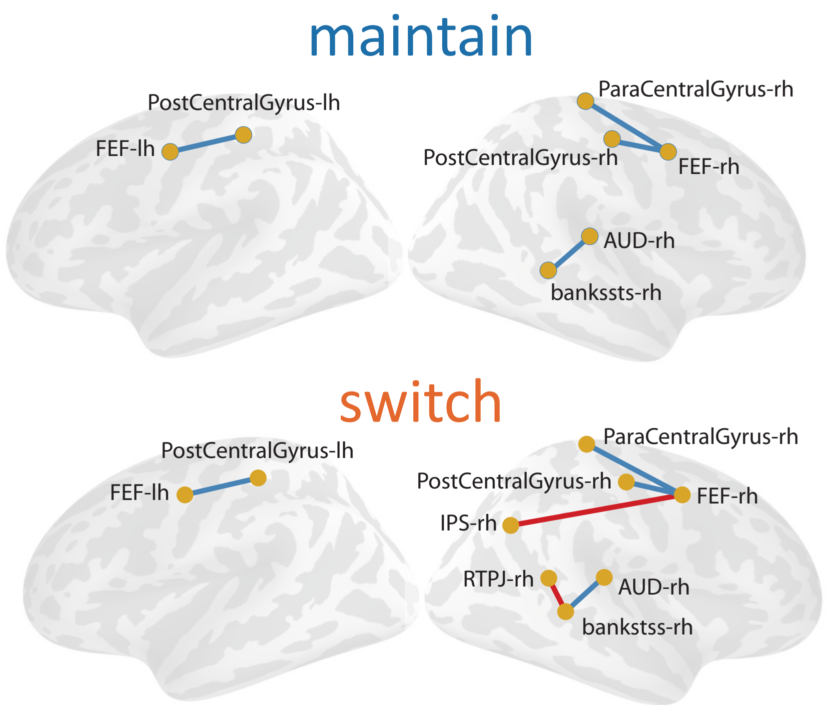


Functional networks in the brain

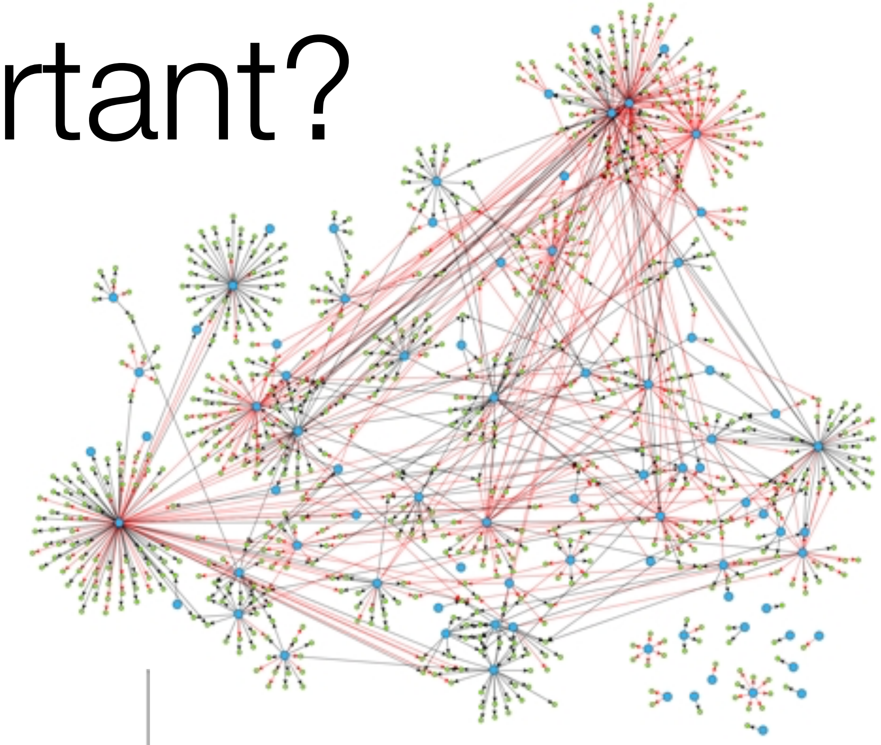


Gene regulatory networks

Why are interactions important?



Functional networks in the brain



Gene regulatory networks

Interactions between players on the court

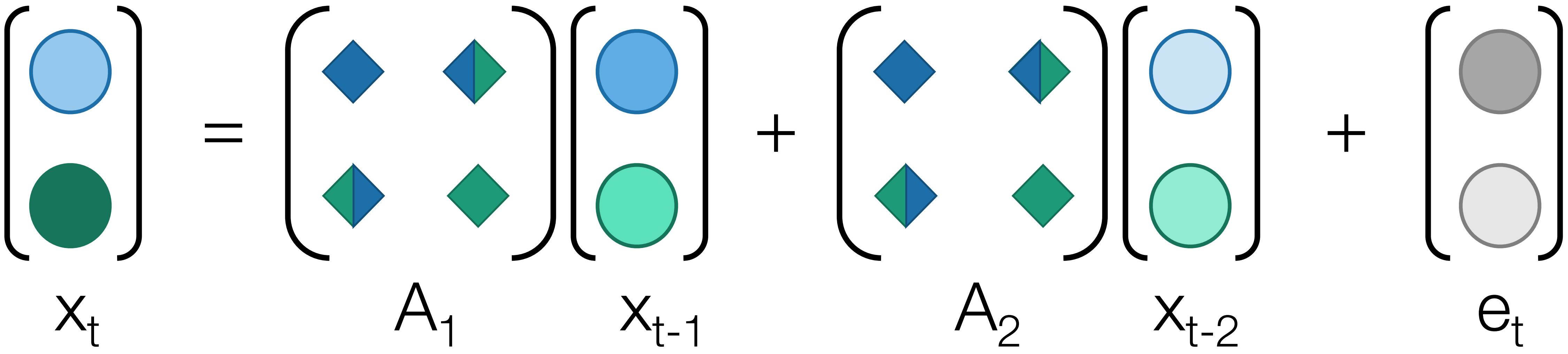
(Video of results from BenShitrit et al. *ICCV* 2011)

Discovering interactions between players

Identify directed interactions between players and ball

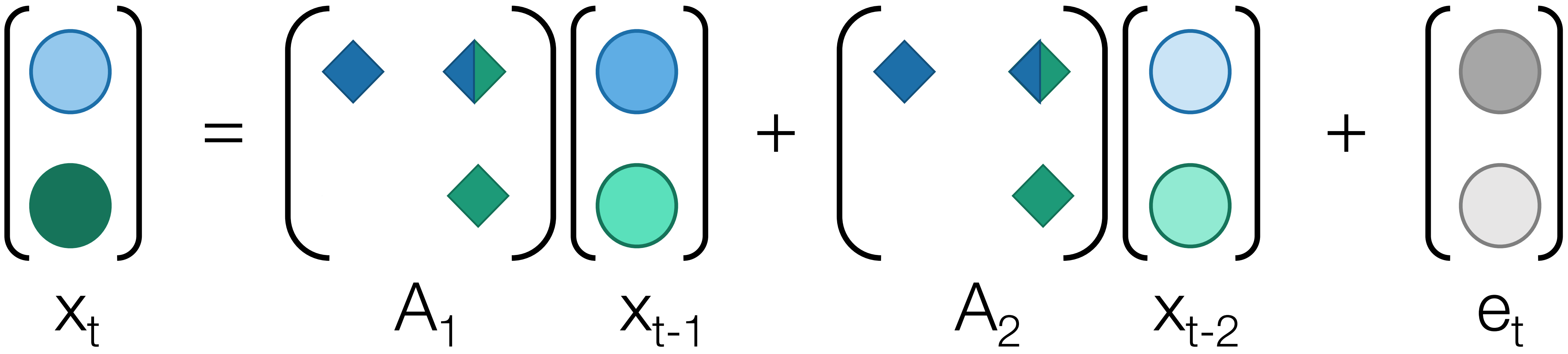
E.g., Position of point guard at time t influences ball at time t+1

Granger causality selection – Linear model



$$x_t = \sum_{k=1}^{K} A_k x_{t-k} + e_t$$

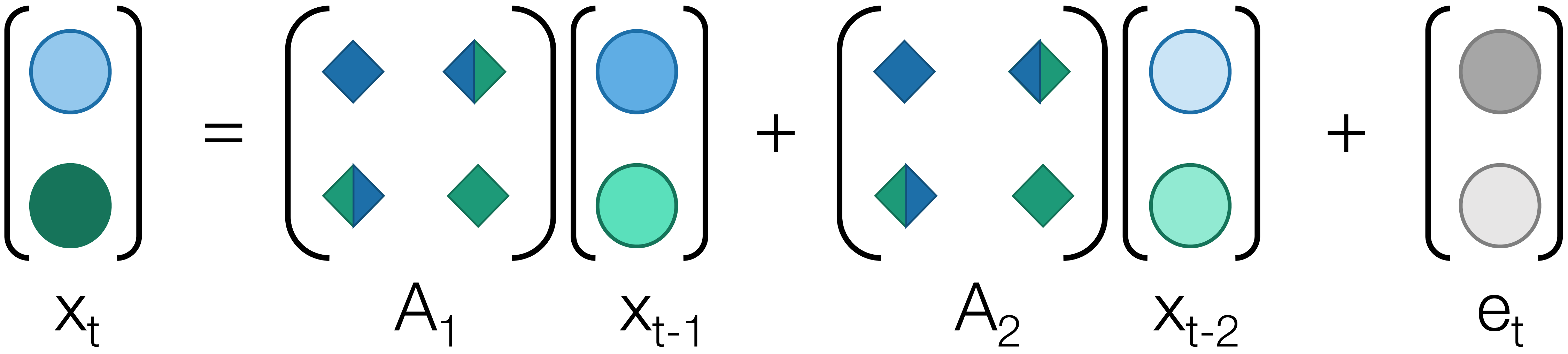
Granger causality selection – Linear model

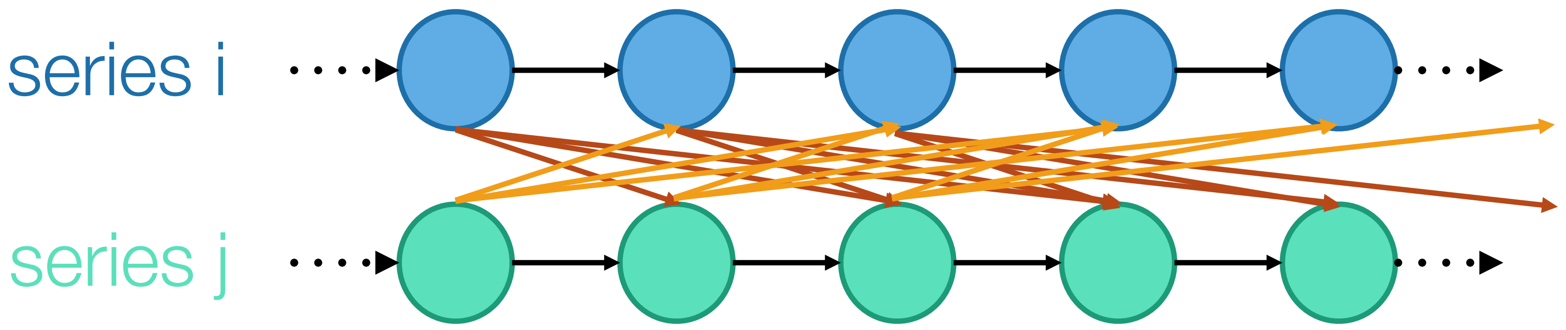


Series i **does not** Granger cause series j iff $A_{ji,k} = 0 \quad \forall k$

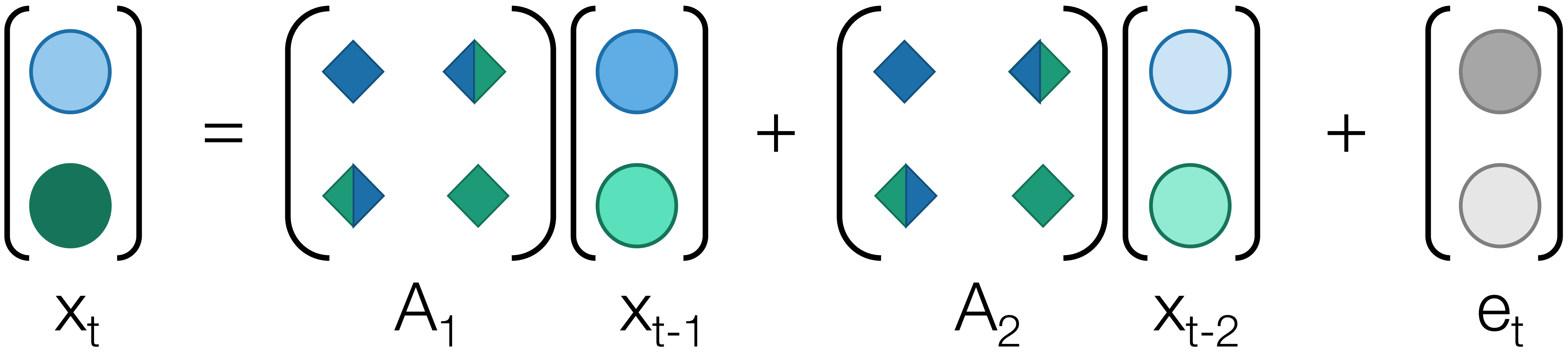
Lag k interaction

Granger causality selection – Linear model



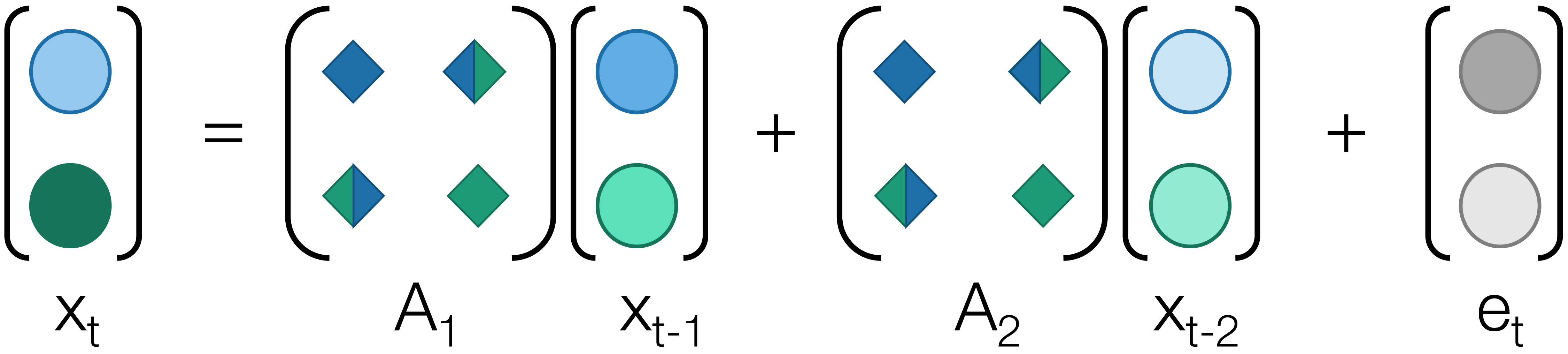


Granger causality selection – Linear model



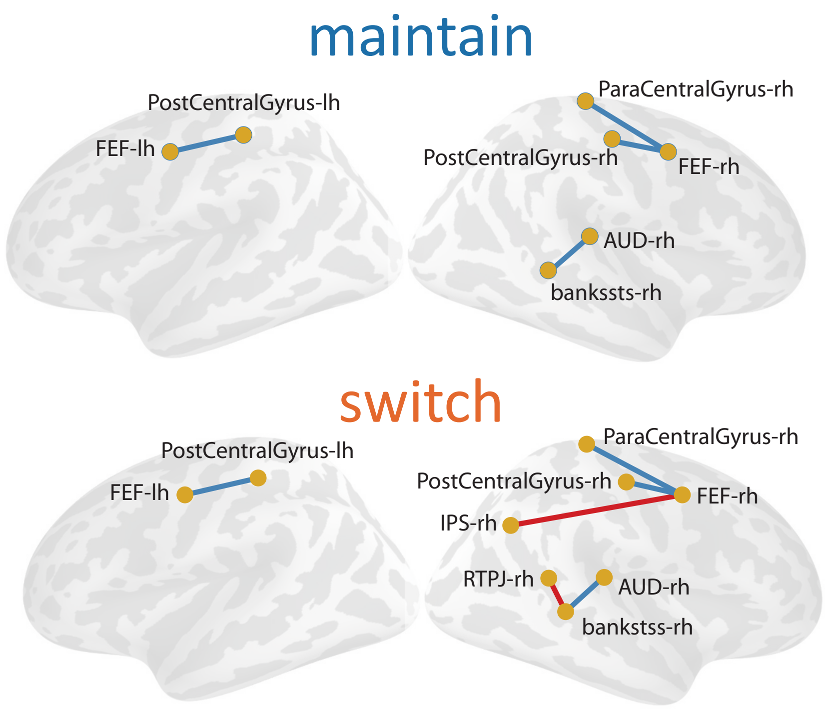
$$\max_{A_1,\ldots,A_K} \underbrace{\text{loglike}(x_1,\ldots,x_T\,;\,A_1,\ldots,A_K)}_{\text{explain data well}} - \lambda \sum_{j,i} \underbrace{\text{penalty}(A_{ji,1},\ldots,A_{ji,K} \neq 0)}_{\text{encourage (structured) 0's}}$$

Granger causality selection – Linear model



$$\min_{A_1,\ldots,A_K} \underbrace{\sum_{t=K}^{T} \left(x_t - \sum_{k=1}^{K} A_k x_{t-k} \right)^2}_{\text{reconstruction error}} + \lambda \underbrace{\sum_{ij} ||(A_{ji,1},\ldots,A_{ji,K})||_2}_{\text{group lasso penalty}},$$

The issue with a linear approach

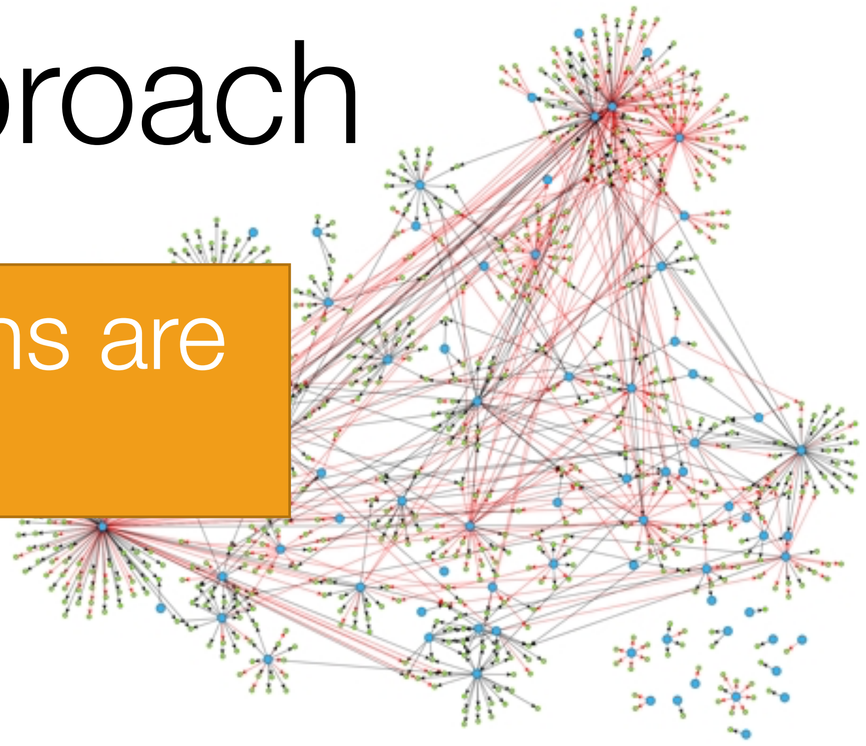


Functional networks in the brain

What if interactions are nonlinear?

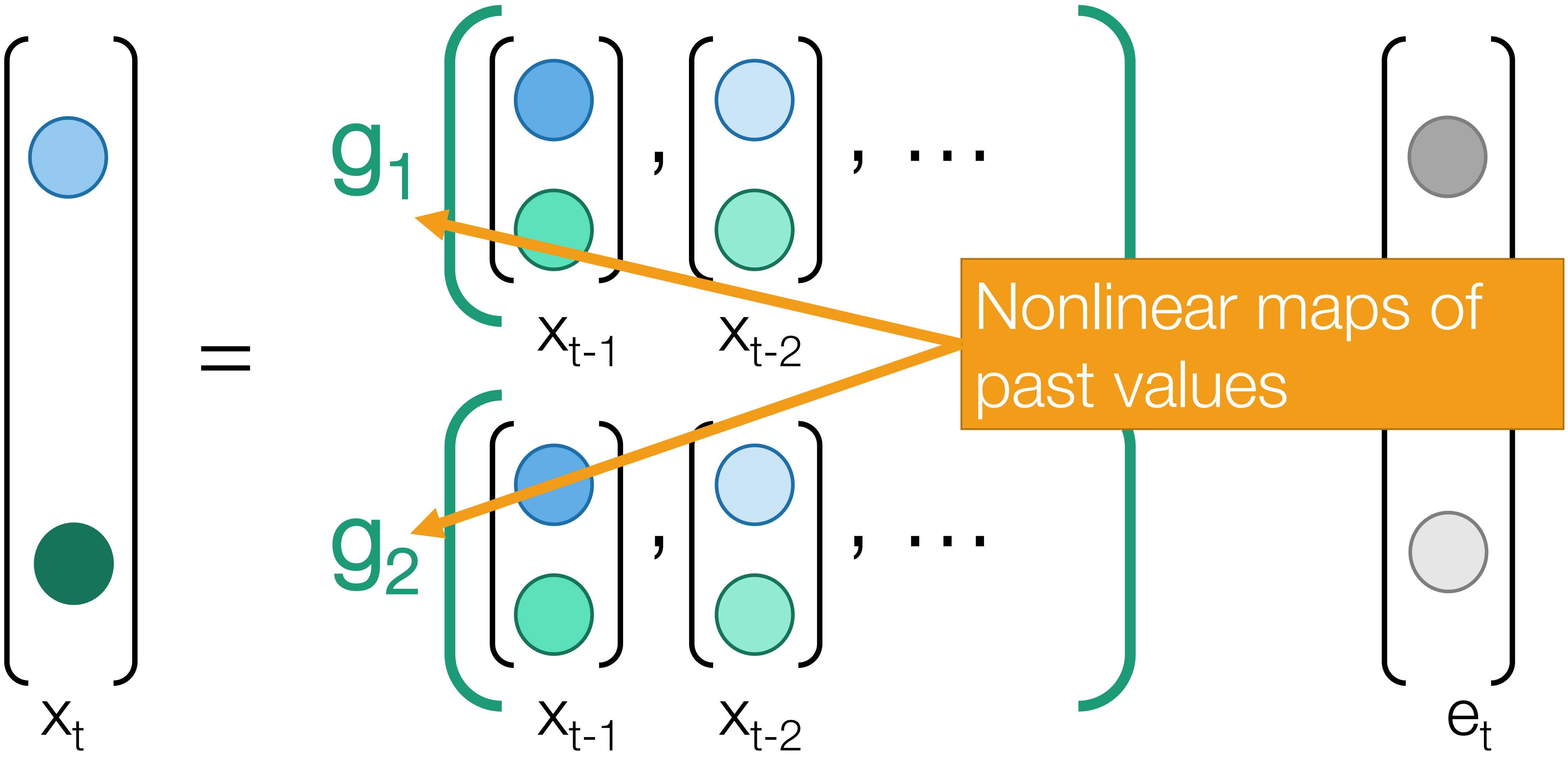
Interactions between players on the court

(Video of results from BenShitrit et al. *ICCV* 2011)

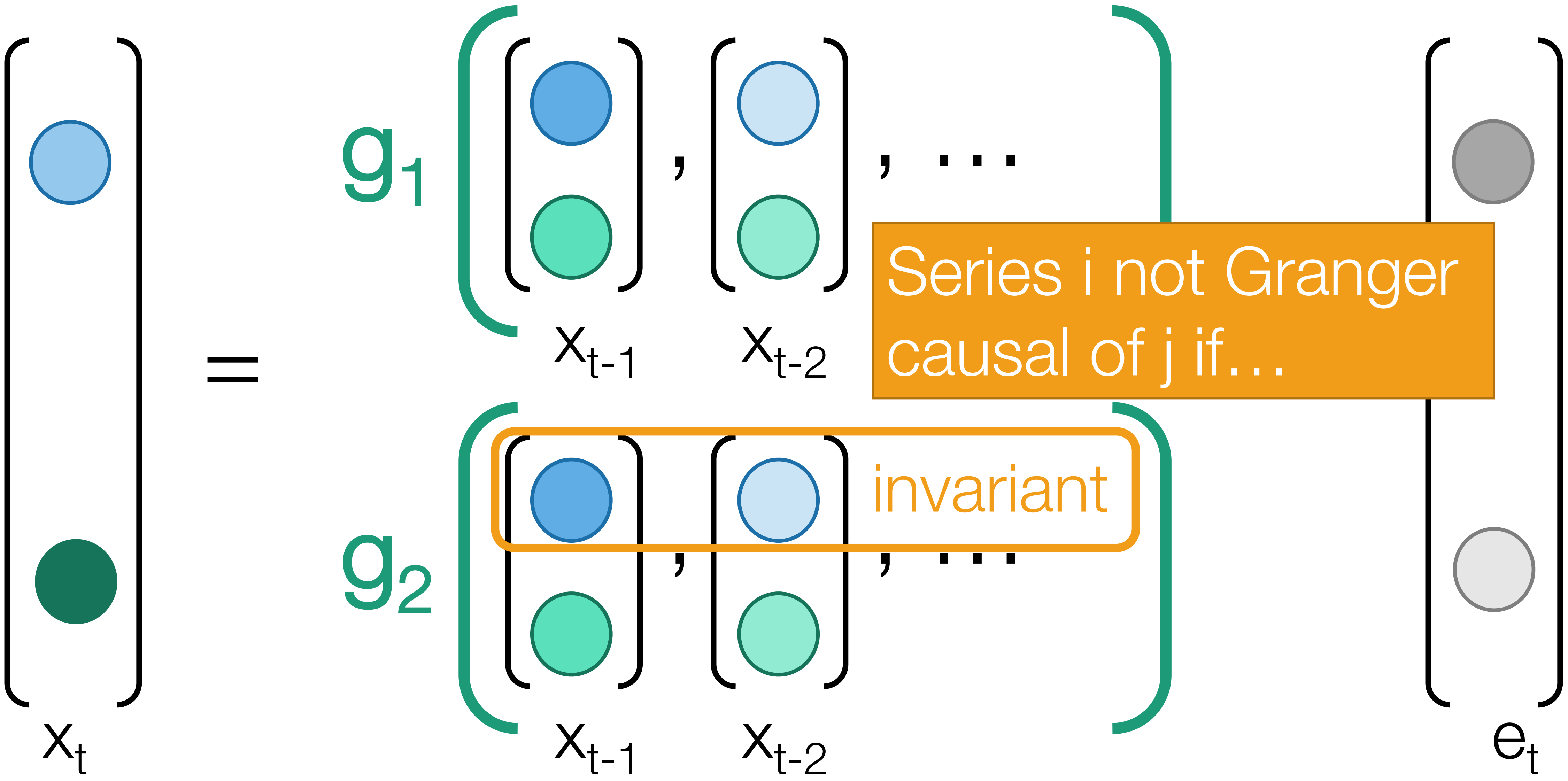


Gene regulatory networks

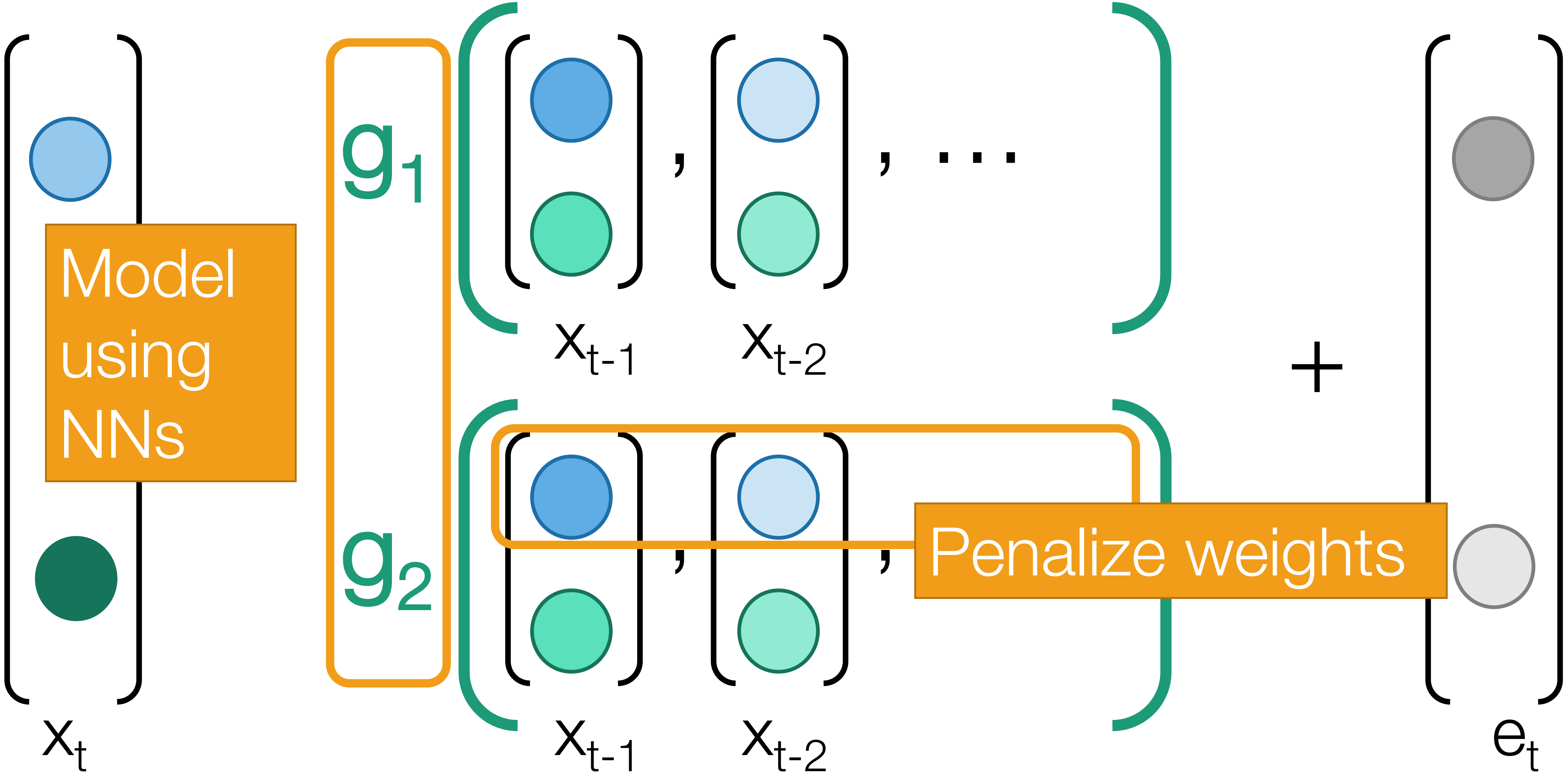
Modeling nonlinear dynamics



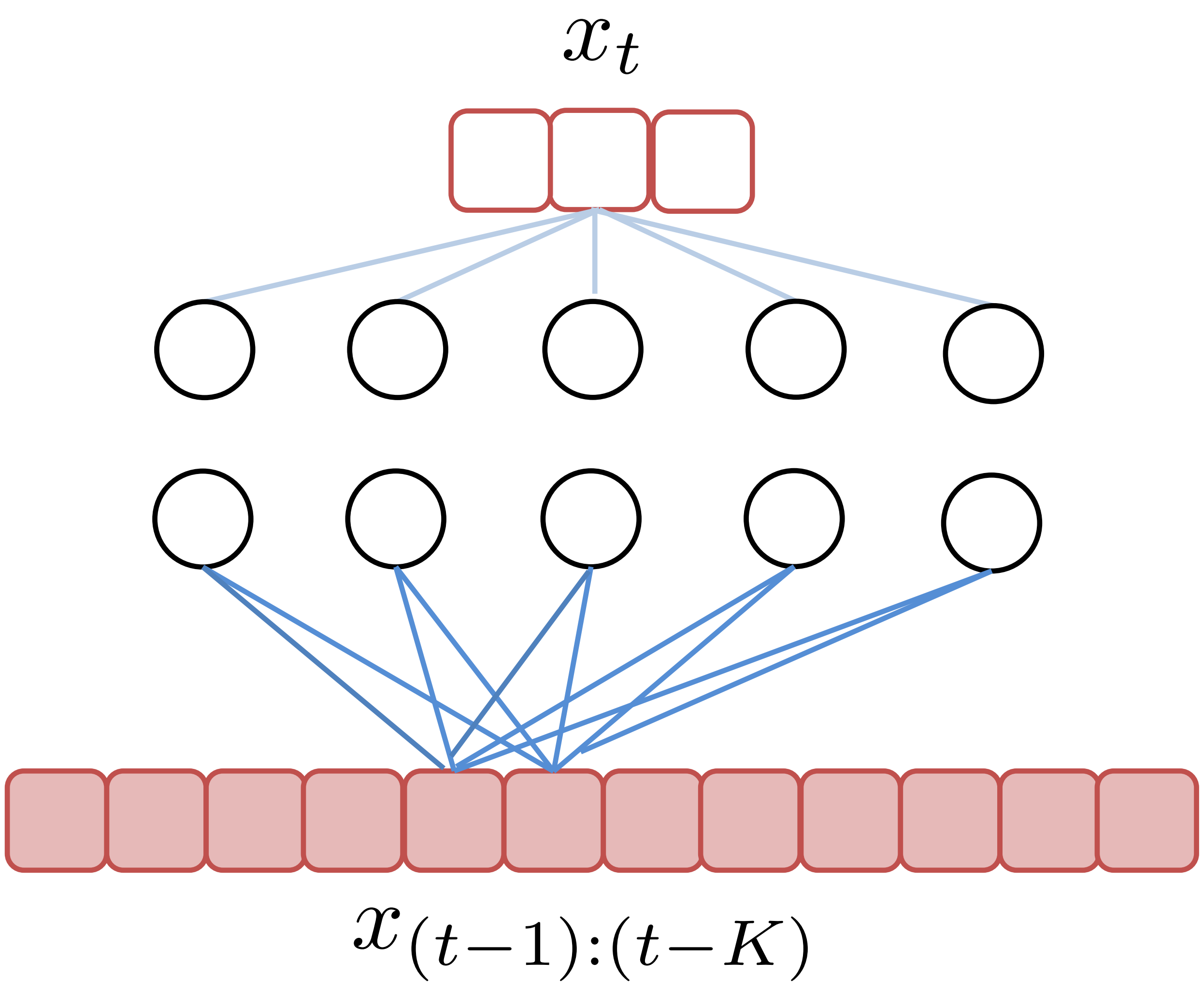
Identifying Granger causality



Using penalized neural networks



A multilayer perceptron (MLP) approach

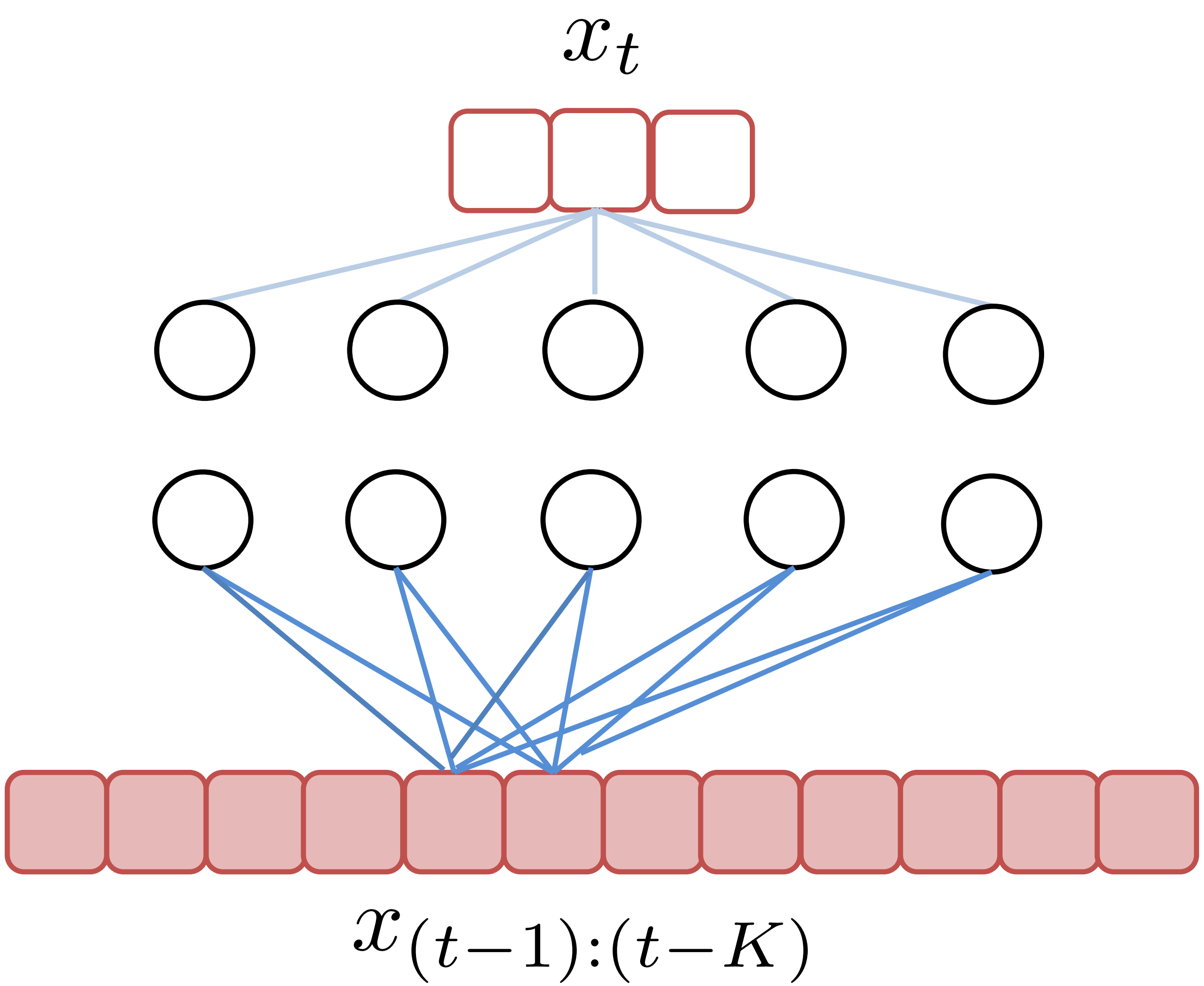


full set of outputs

MLP

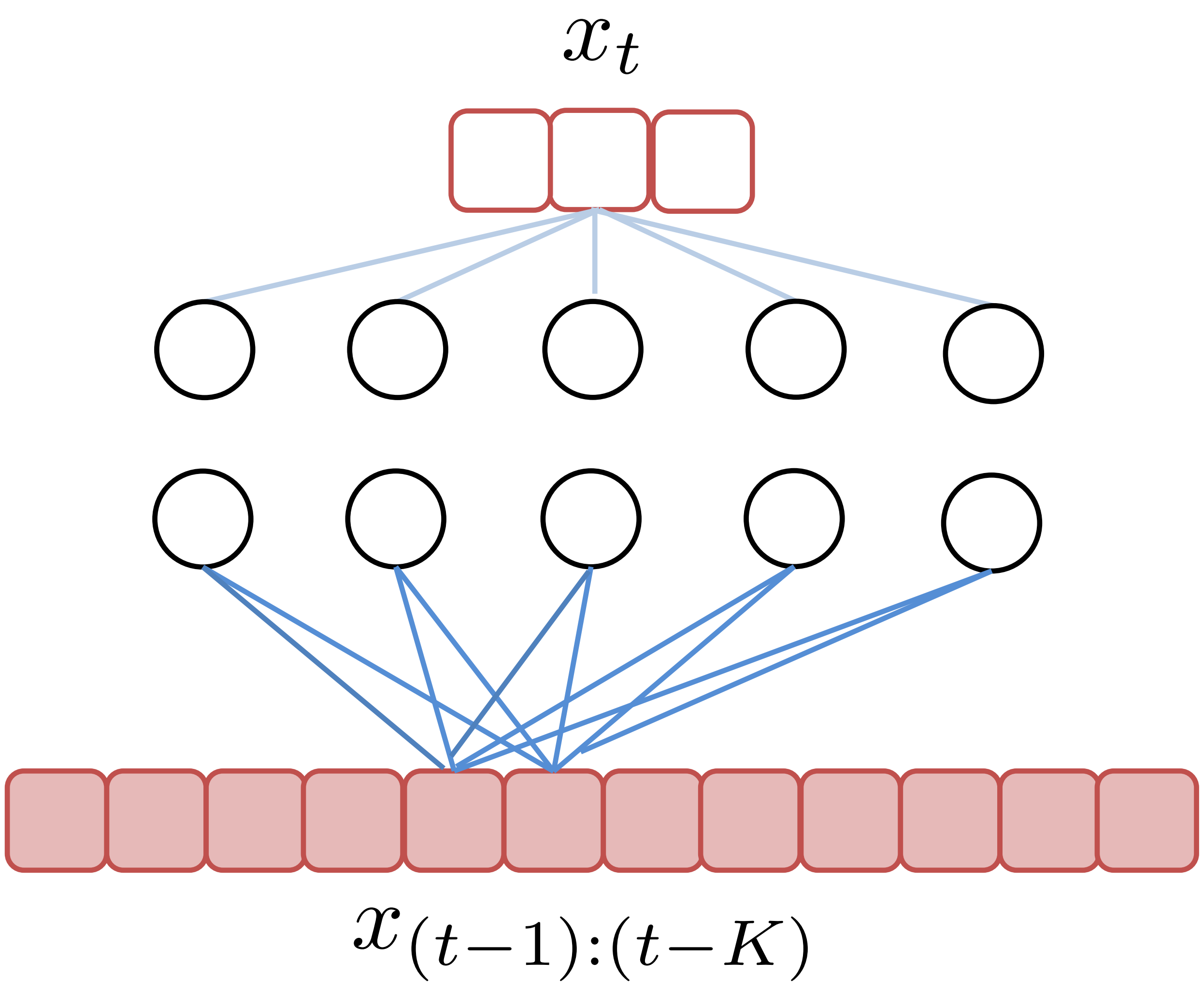
lag K past values as inputs

A multilayer perceptron (MLP) approach



difficult to identify Granger causality with shared hidden units

A multilayer perceptron (MLP) approach



all g_i must rely on same set of lags

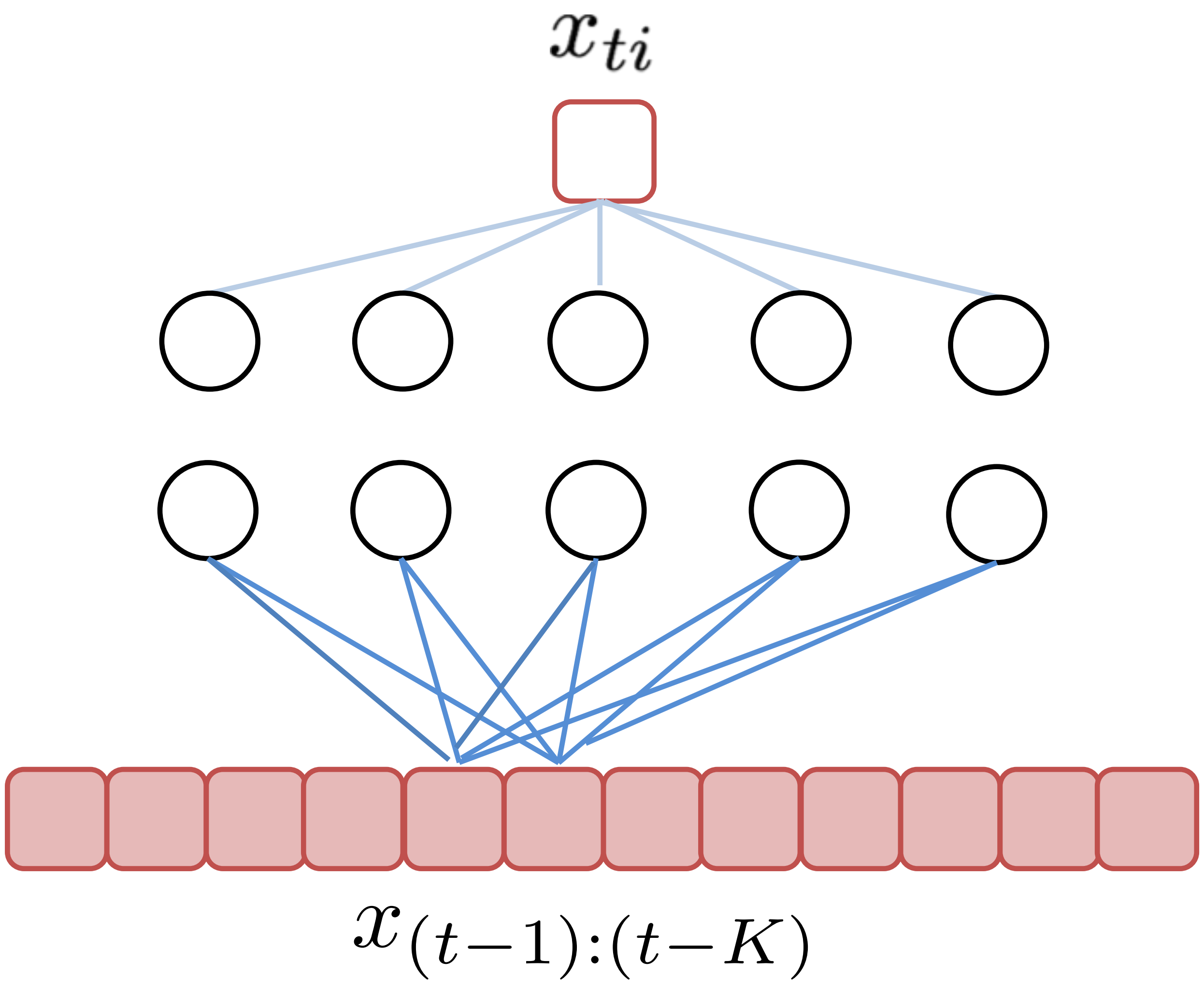
Penalized **multilayer perceptron** (MLP)

For function g_i

max

$\text{loglike}(x_{1i},\ldots,x_{Ti}\ ;\ W^1,\ldots,W^L)$

$-\ \lambda \sum_j \text{penalty}(W^j \neq 0)$



Specifying the MLP – function g_i

Linear output decoder:

$$x_{ti} = w_O^T h_t^L + \epsilon_t$$

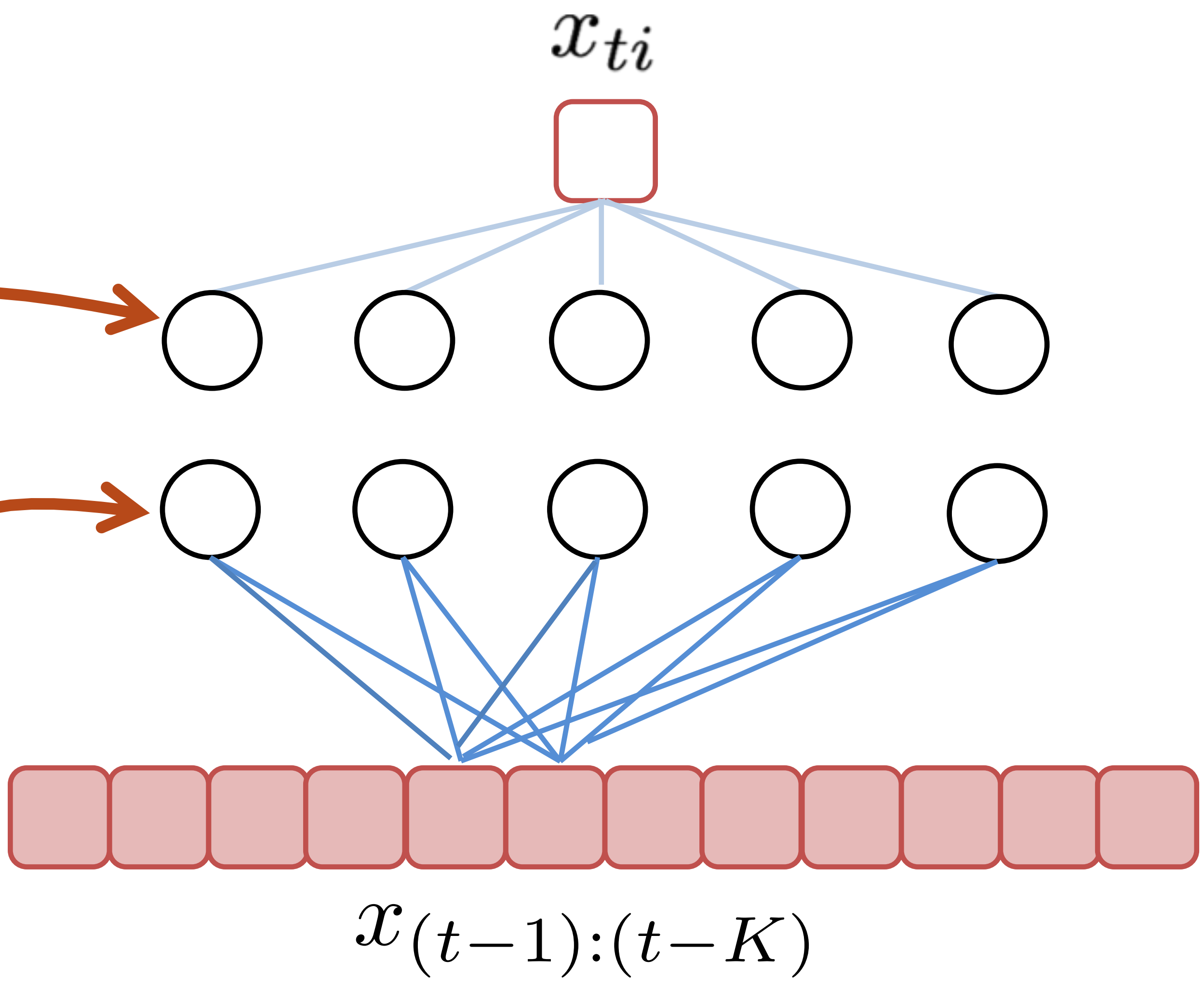
Layer ℓ hidden values:

$$h_t^\ell = \sigma\left(W^\ell h_t^{\ell-1} + b^\ell\right)$$

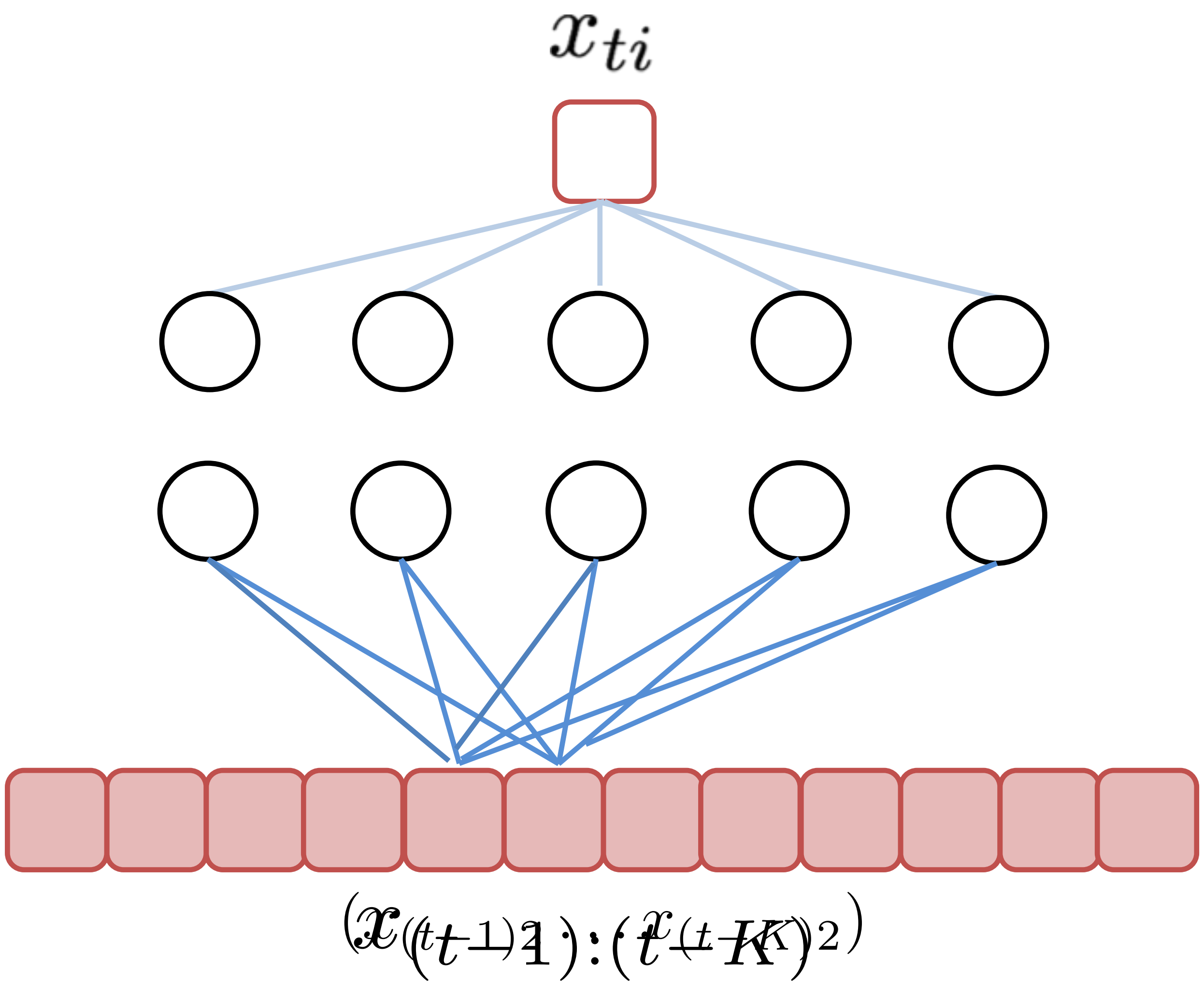
Layer 1 hidden values:

$$h_t^1 = \sigma\left(\sum_{k=1}^{K} \boxed{W^{1k}} x_{t-k} + b^1\right)$$

lag-specific weights



Disentangling input to output effects

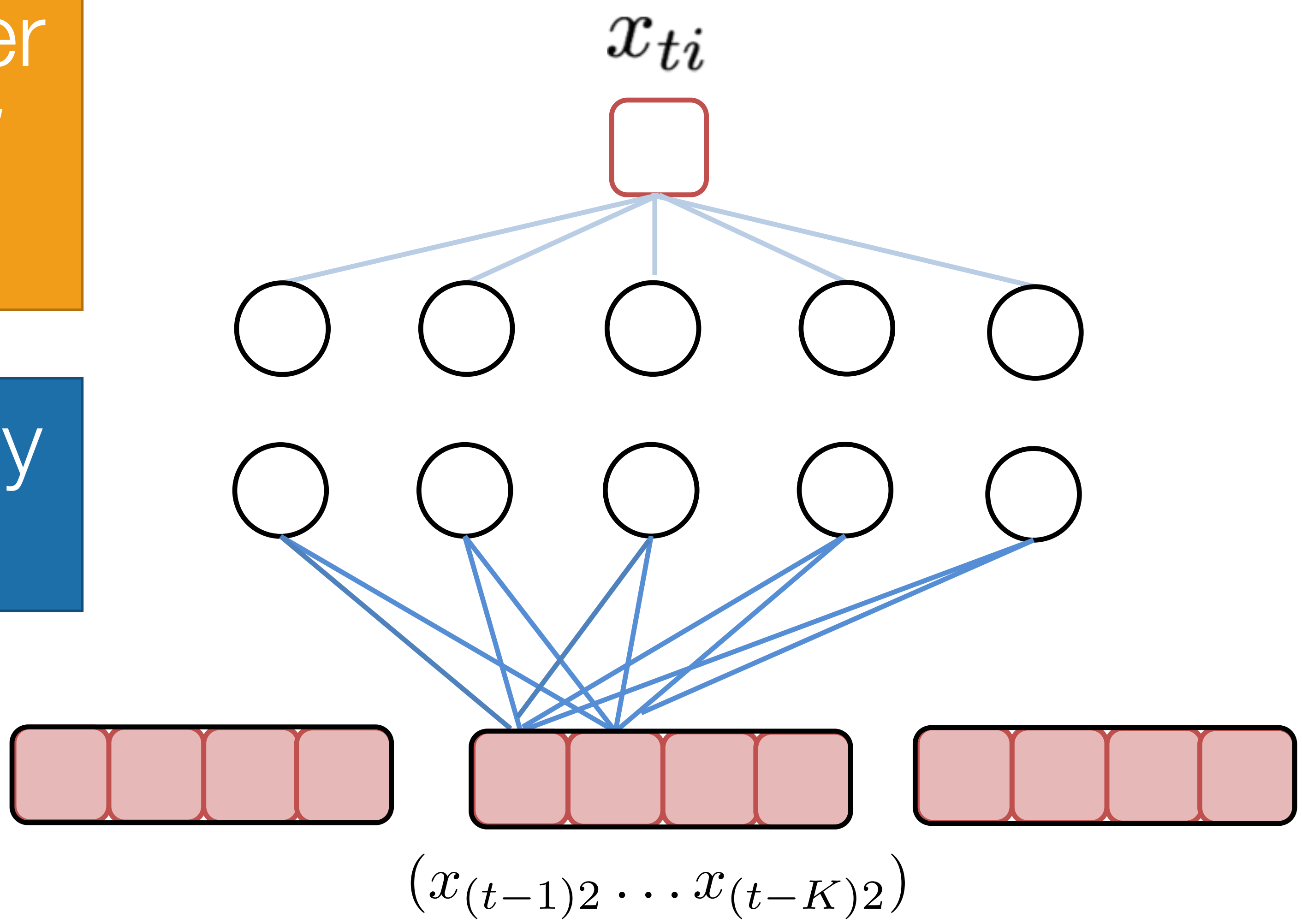


Disentangling input to output effects

series j does not Granger cause **series i** if *group j weights are 0*

place group-wise penalty on layer 1 weights

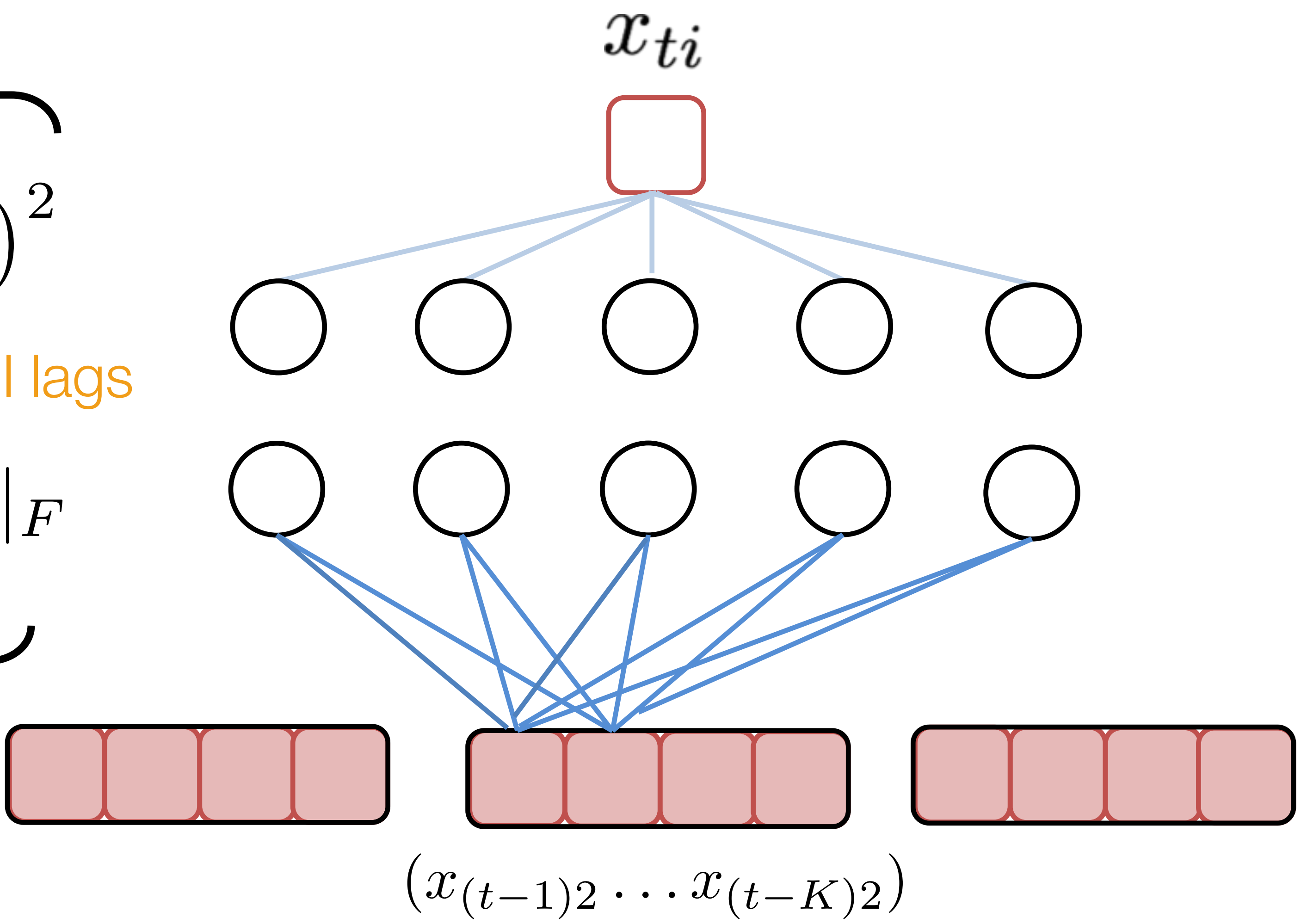
group inputs by:
(K lags of series j)



Penalized **multilayer perceptron** (MLP)

$$\overbrace{\min_{\mathbf{W}} \sum_{t=K}^{T} \left(x_{it} - g_i(x_{(t-1):(t-K)})\right)^2}^{\text{reconstruction error}}$$

$$\underbrace{+\,\lambda \sum_{j=1}^{p} \|\overbrace{\boxed{(W_{:j}^{11}, \ldots, W_{:j}^{1K})}}^{\text{weights from series j at all lags}}\|_F}_{\text{group lasso penalty}}$$



Algorithmic notes…

Often, focus of deep learning evaluation is on prediction error…

Can get away with **optimizing approximately**

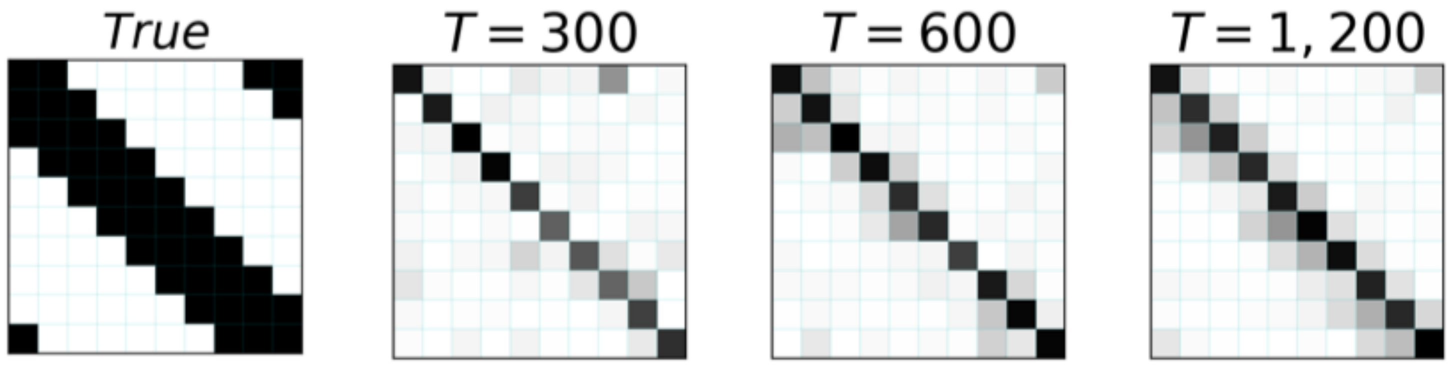
SGD

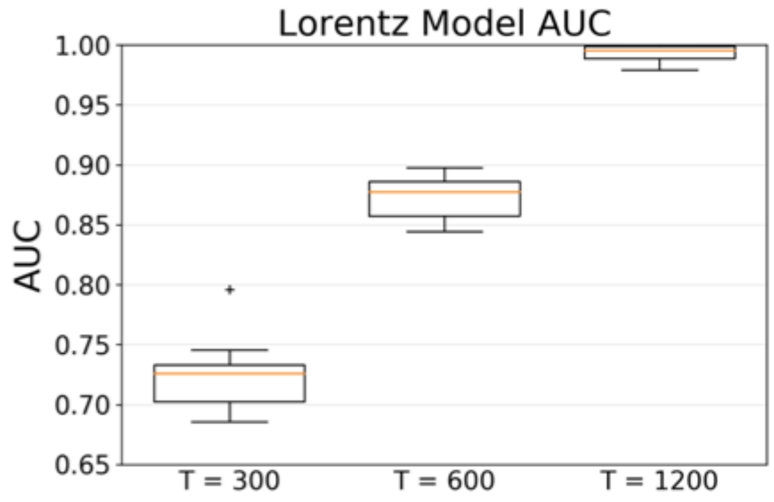
We care about zeros of solution, so important to get very close to a **stationary point**

Proximal gradient descent with line search

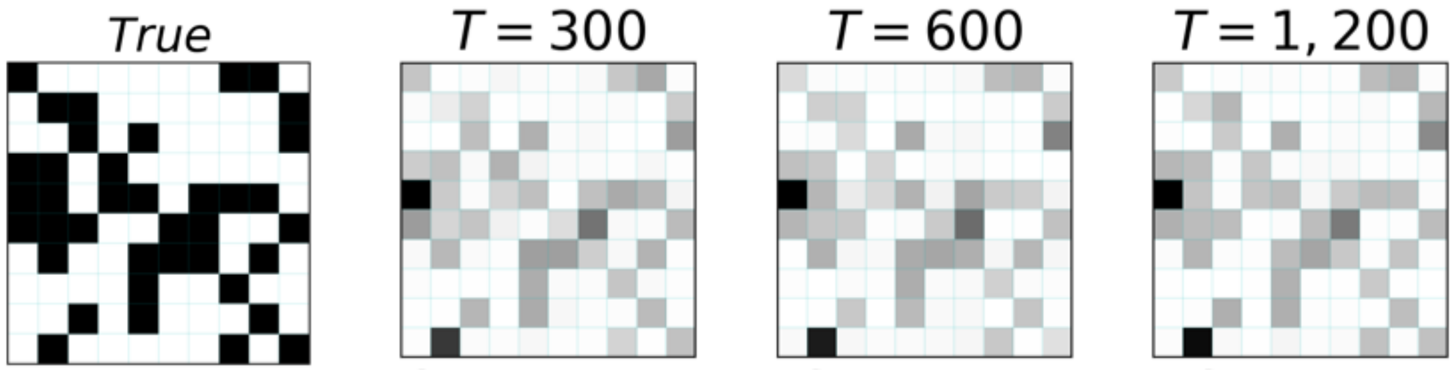
Simulated results – MLP

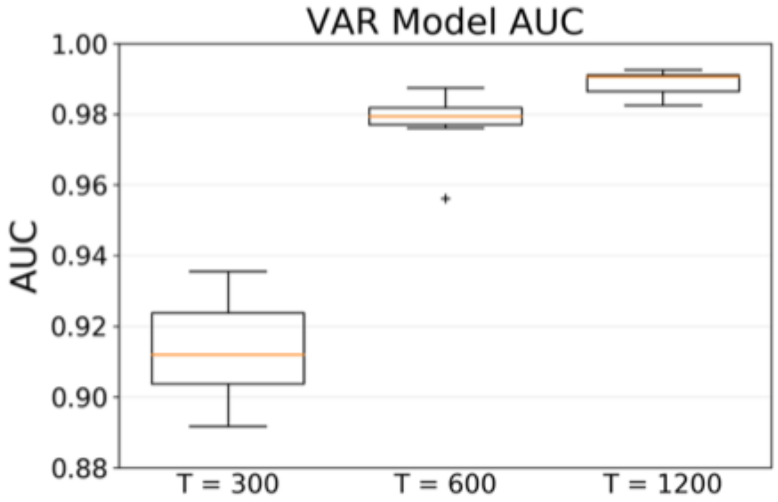
Lorenz-96 (nonlinear) data





VAR (linear) data



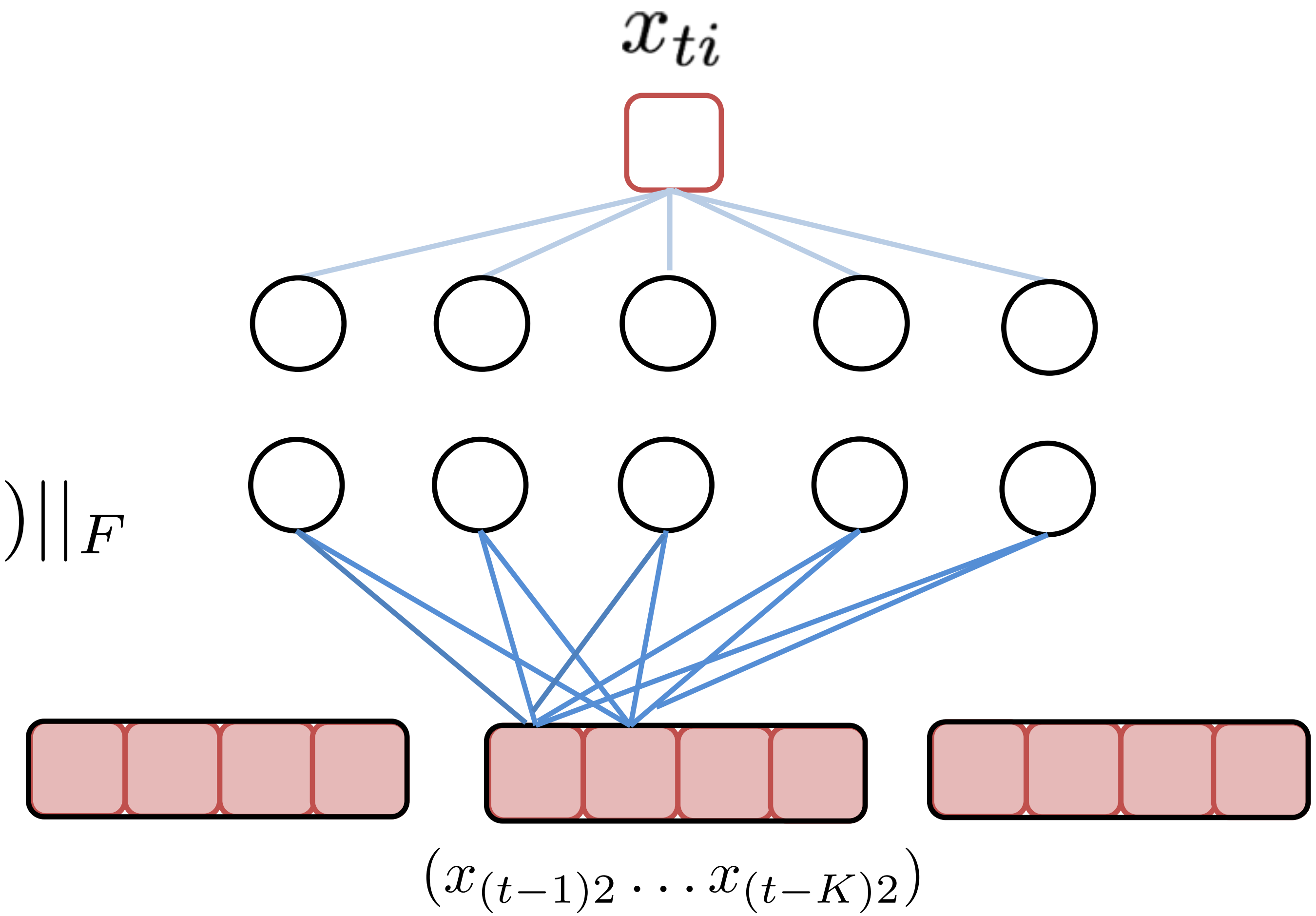


Lag selection via hierarchical group lasso

$$\min_{\mathbf{W}} \sum_{t=K}^{T} \left(x_{it} - g_i(x_{(t-1):(t-K)})\right)^2$$

$$\underbrace{\cancel{+ \lambda \sum_{j=1}^{p} ||(W_{:j}^{11}, \ldots, W_{:j}^{1K})||_F}}_{\text{group lasso penalty}}$$

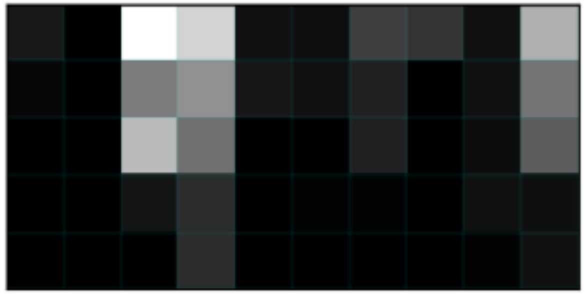
$$\underbrace{\lambda \sum_{j=1}^{p} \sum_{k=1}^{K} ||(W_{:j}^{1k}, \ldots, W_{:j}^{1K})||_F}_{\substack{\textbf{hierarchical} \\ \text{group lasso penalty}}}$$



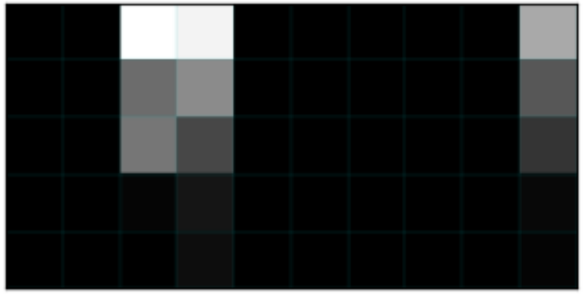
Lag selection results

True

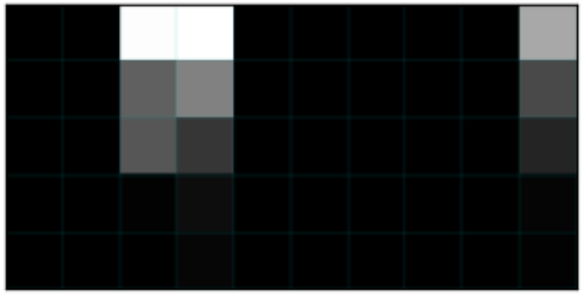
$\lambda = 0.010$



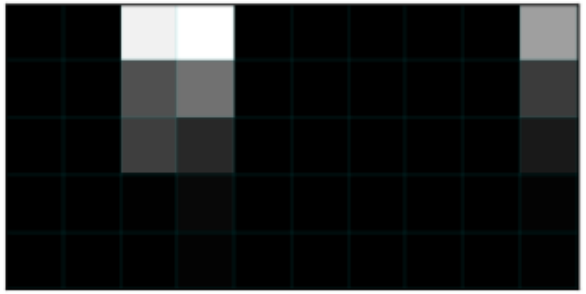
$\lambda = 0.037$



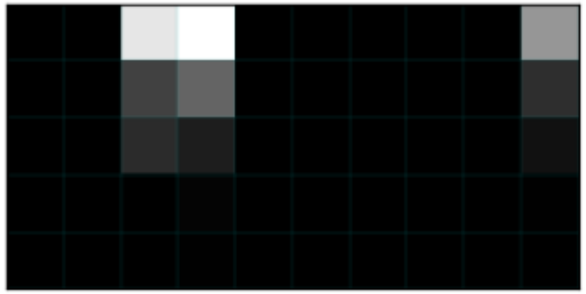
$\lambda = 0.063$



$\lambda = 0.090$

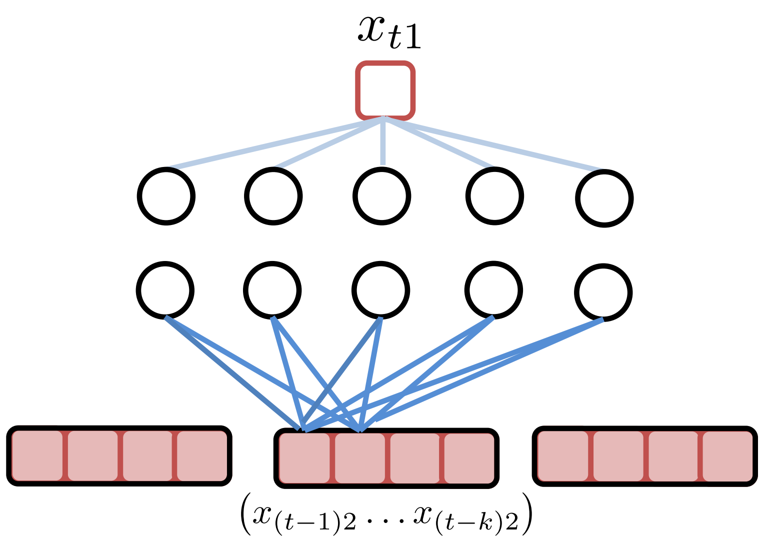


$\lambda = 0.117$

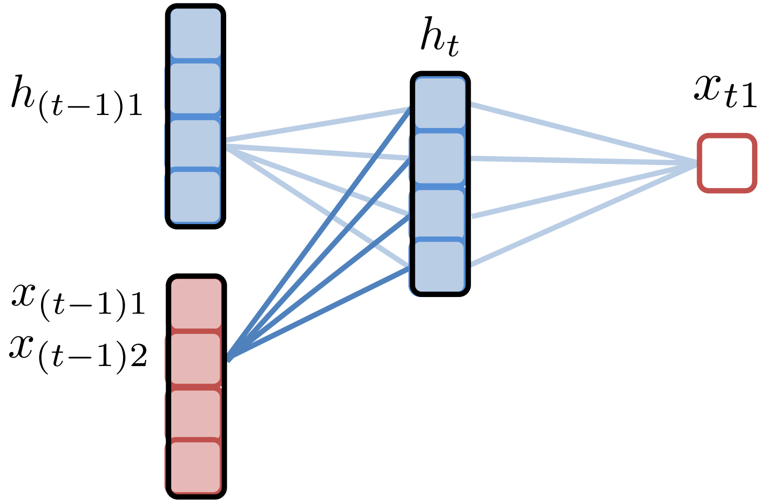


increasing sparsity penalty λ

Multilayer Perceptron



Recurrent Network



Long-range dependencies between series via **nonlinear hidden variables**

Generic RNN formulation

Hidden state evolution:

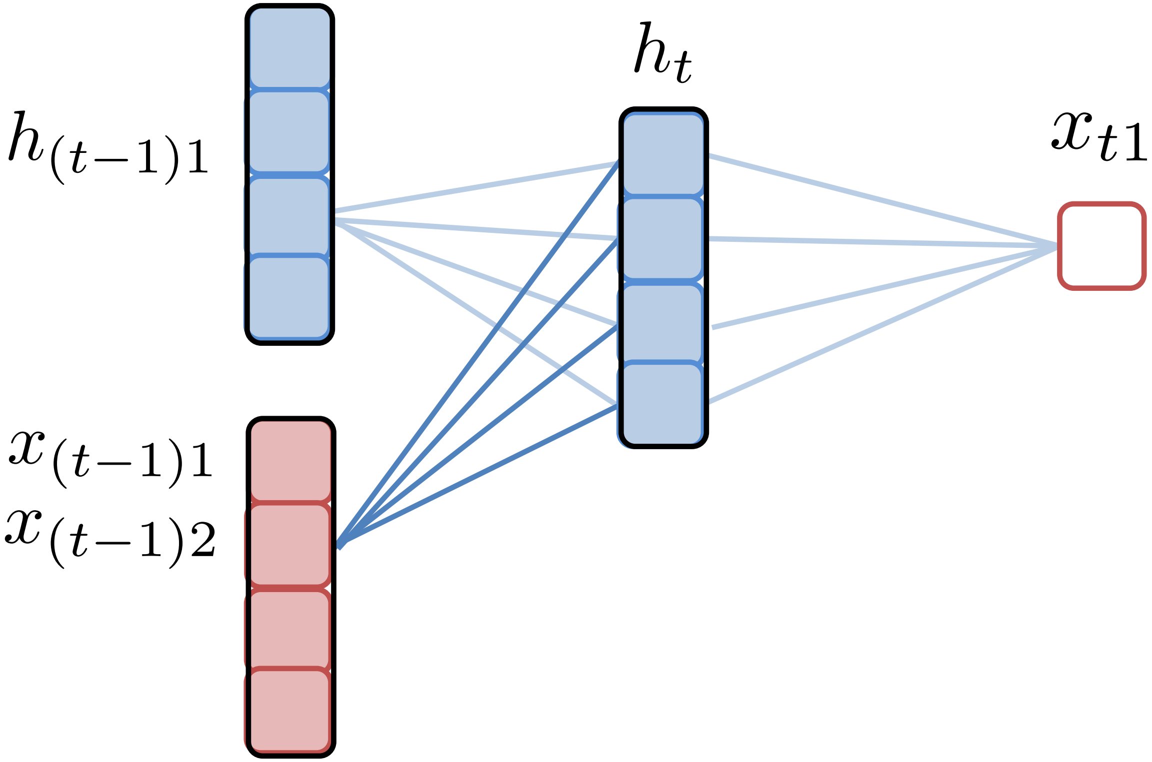
nonlinear fcn depending on architecture

$$h_t = f(x_t, h_{t-1})$$

hidden state capturing historical context

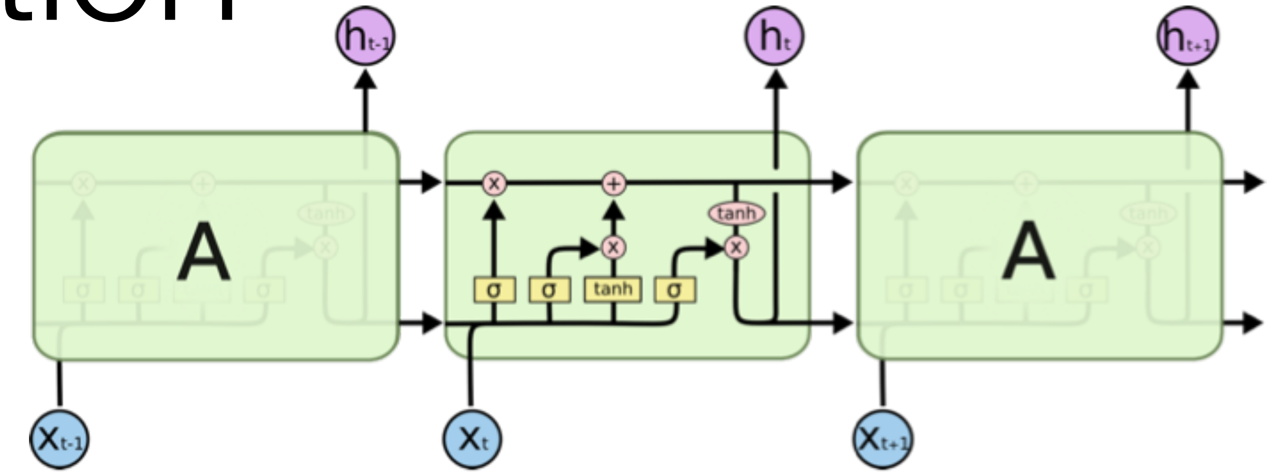
Linear output layer (for simplicity):

$$x_{it} = w_O^T h_t + \epsilon_t$$



LSTM specification

Introduce **cell state** c_t
in addition to h_t



forget gate $$f_t = \sigma\left(W^f x_t + U^f h_{(t-1)}\right)$$

input gate $$i_t = \sigma\left(W^{in} x_t + U^{in} h_{(t-1)}\right)$$

output gate $$o_t = \sigma\left(W^o x_t + U^o h_{(t-1)i}\right)$$

Control how cell state is updated and transferred to predicted hidden state

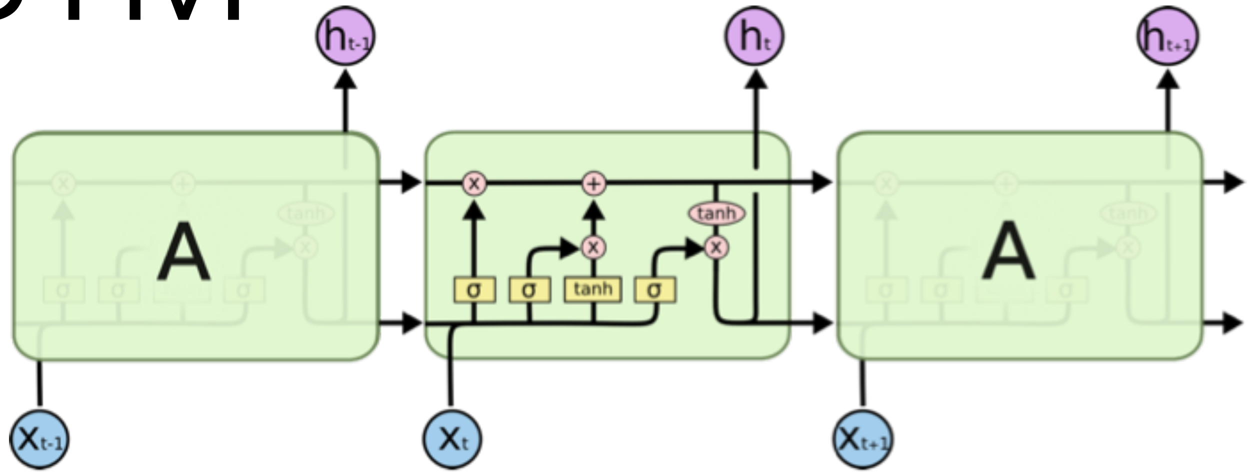
cell state evolution $$c_t = f_t \odot c_{t-1} + i_t \odot \sigma\left(W^c x_t + U^c h_{(t-1)}\right)$$

hidden state evolution $$h_t = o_t \odot \sigma(c_t)$$

Weights of the LSTM

$W = ((W^f)^T, (W^{in})^T, (W^o)^T, (W^c)^T)$

define effect of input on prediction



forget gate $f_t = \sigma\left(\boxed{W^f} x_t + U^f h_{(t-1)}\right)$

input gate $i_t = \sigma\left(\boxed{W^{in}} x_t + U^{in} h_{(t-1)}\right)$

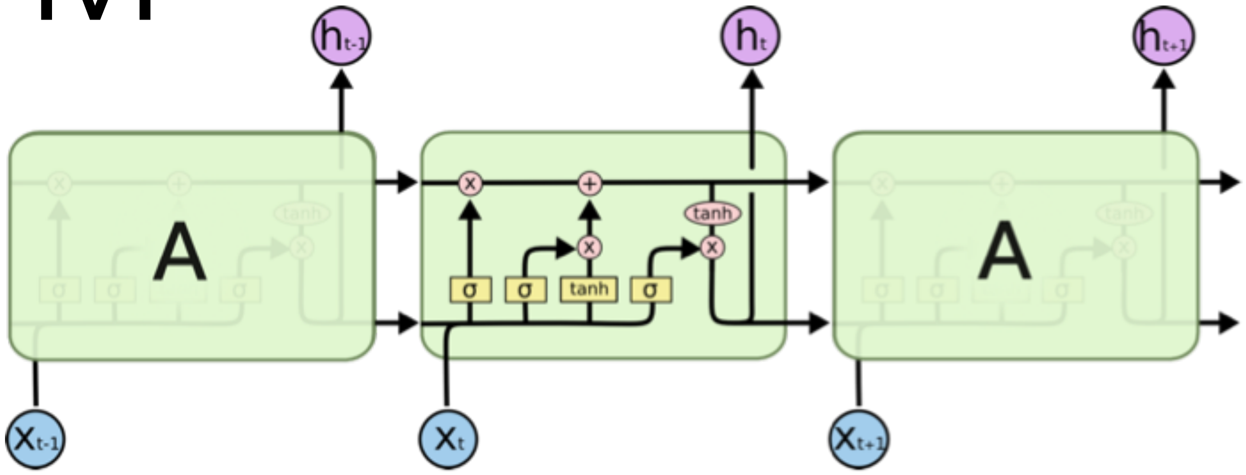
output gate $o_t = \sigma\left(\boxed{W^o} x_t + U^o h_{(t-1)i}\right)$

cell state evolution $c_t = f_t \odot c_{t-1} + i_t \odot \sigma\left(\boxed{W^c} x_t + U^c h_{(t-1)}\right)$

hidden state evolution $h_t = o_t \odot \sigma(c_t)$

A penalized LSTM

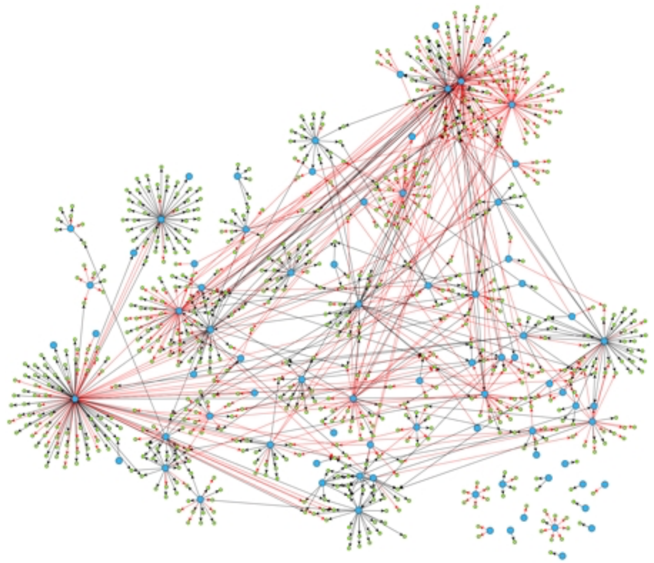
$W = ((W^f)^T, (W^{in})^T, (W^o)^T, (W^c)^T)$
define effect of input on prediction



series j does not Granger cause **series i** if
jth column of weights W is 0

$$\min_{W,U,w^O} \underbrace{\sum_{t=2}^{T} (x_{it} - \underset{w_O^T h_t}{g_i}(x_{<t}))^2}_{\text{reconstruction error}} + \lambda \underbrace{\sum_{j=1}^{p} ||W_{:j}||_2}_{\text{group lasso penalty}}$$

DREAM3 challenge



Difficult **non-linear dataset** used to benchmark Granger causality detection

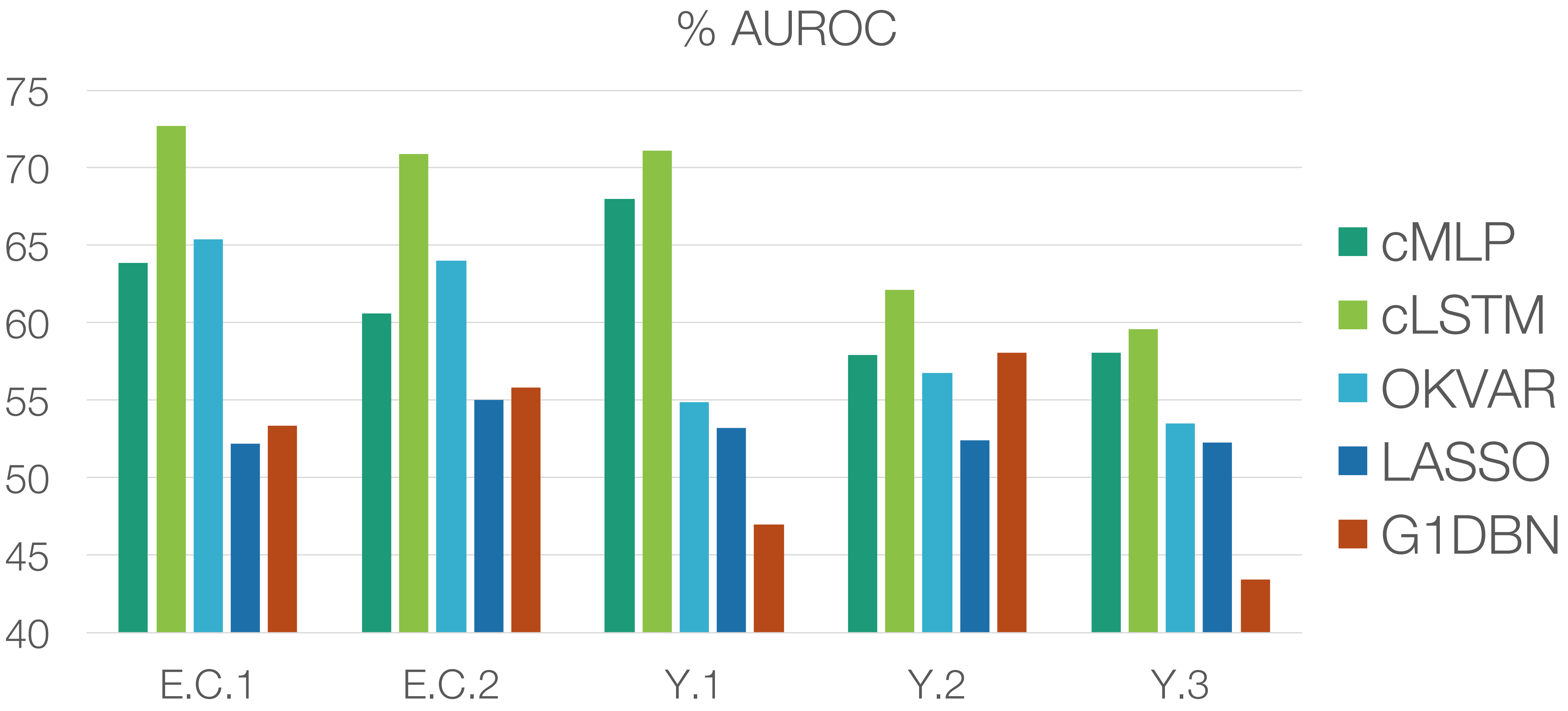
Simulated gene expression and regulation dynamics for:

- 2 E.Coli and 3 Yeast
- 100 series (network size)
- 46 replicates
- 21 time points

Very different structures

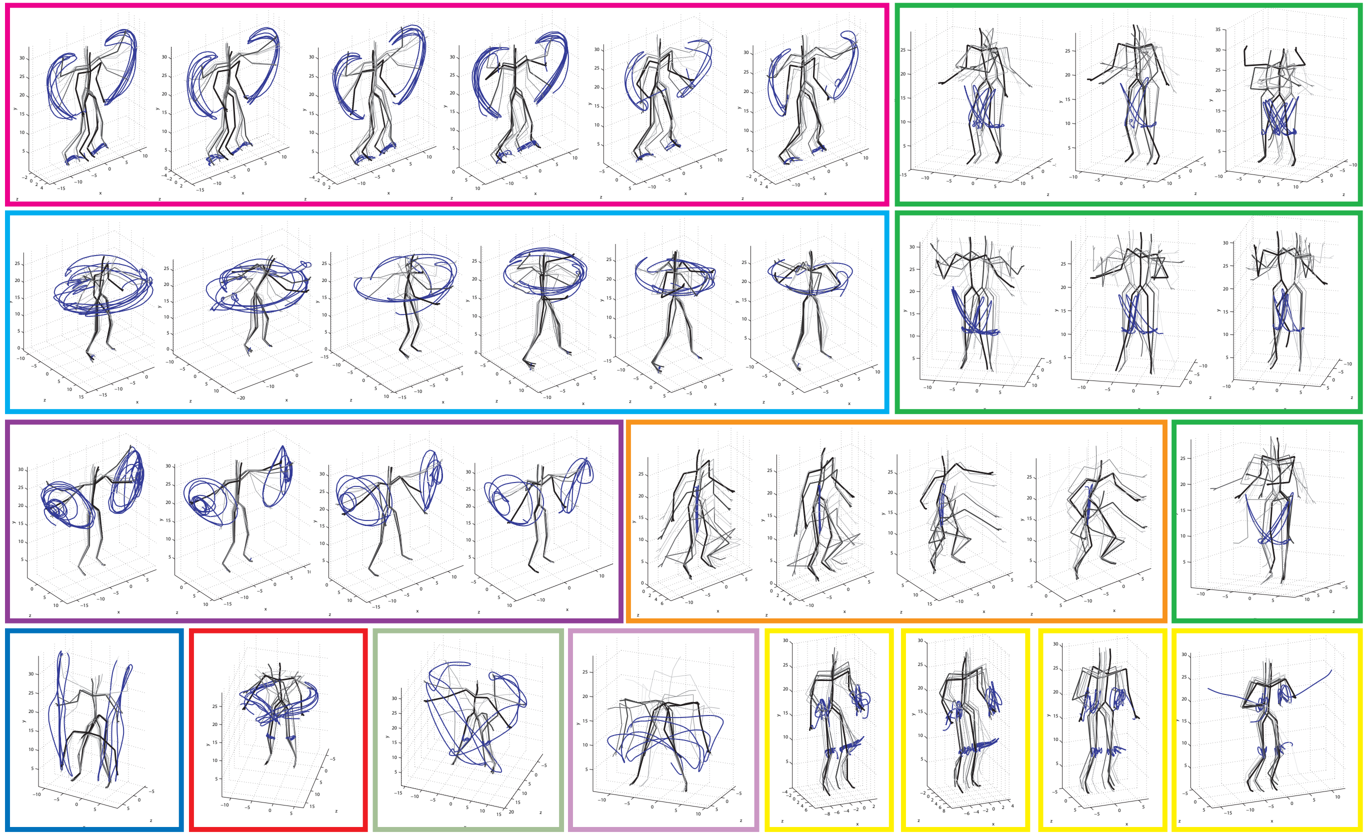
Structure extracted from currently established gene regulatory networks

DREAM3 results

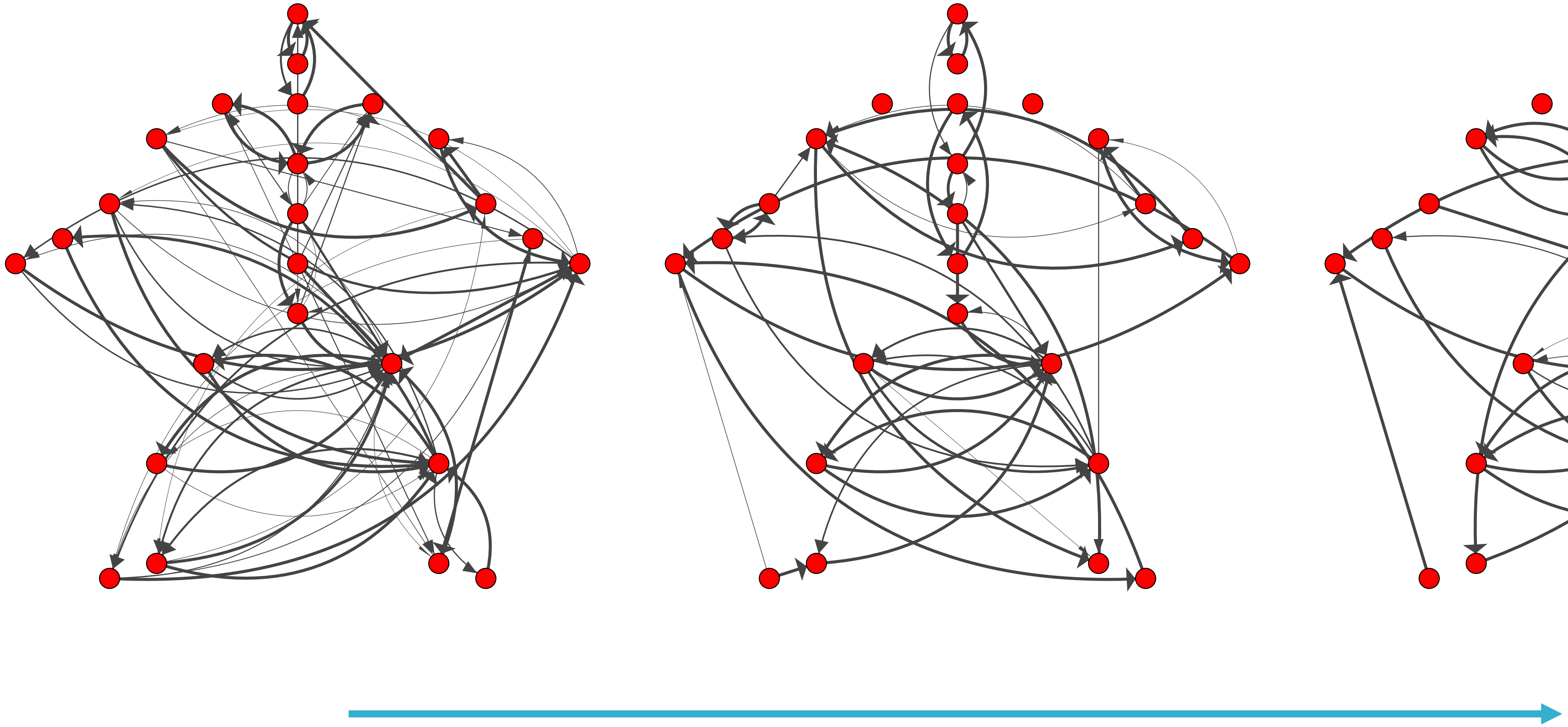


Interactions of the human body

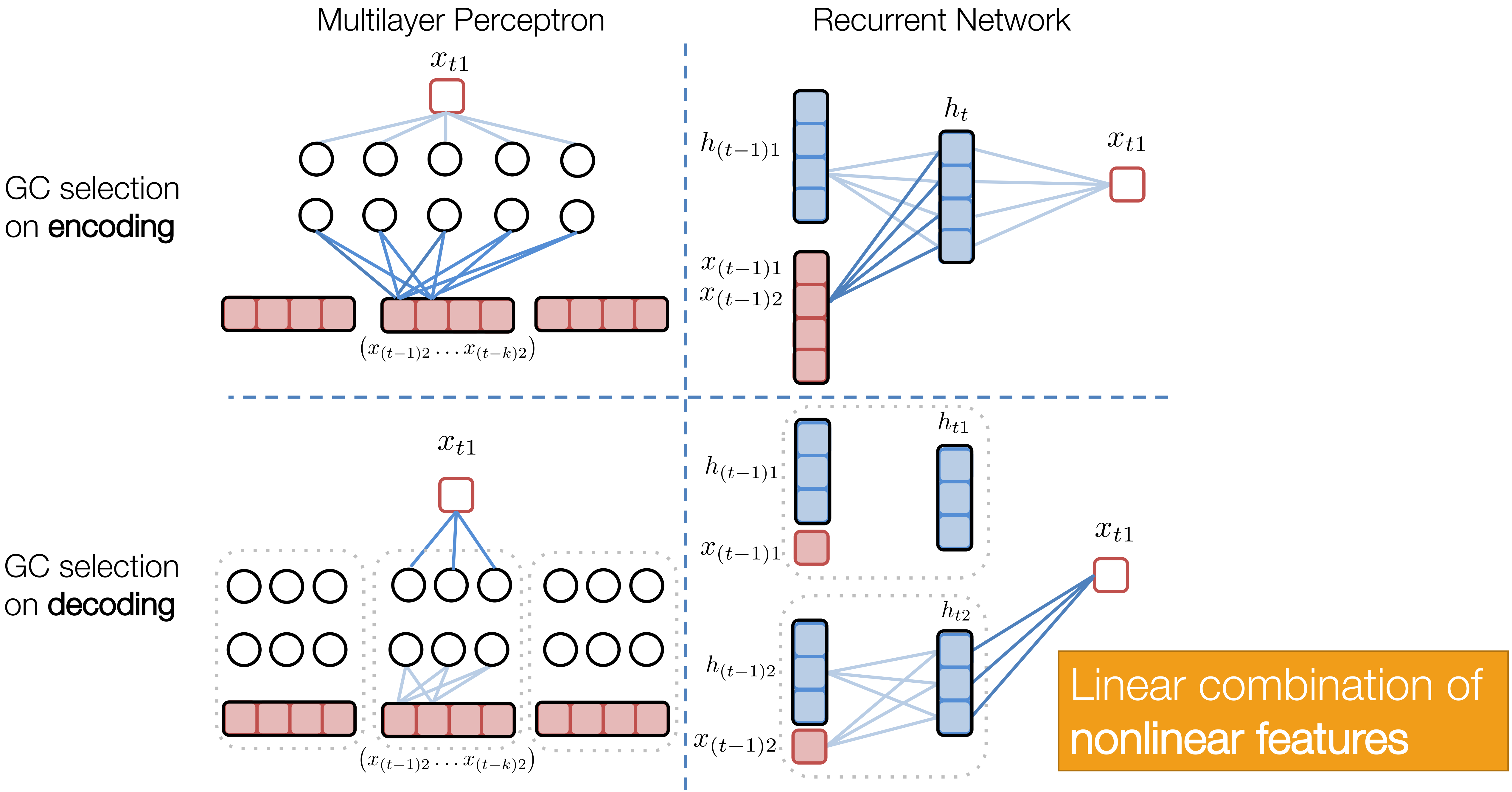
6 videos, 56 dims
2000 total time points



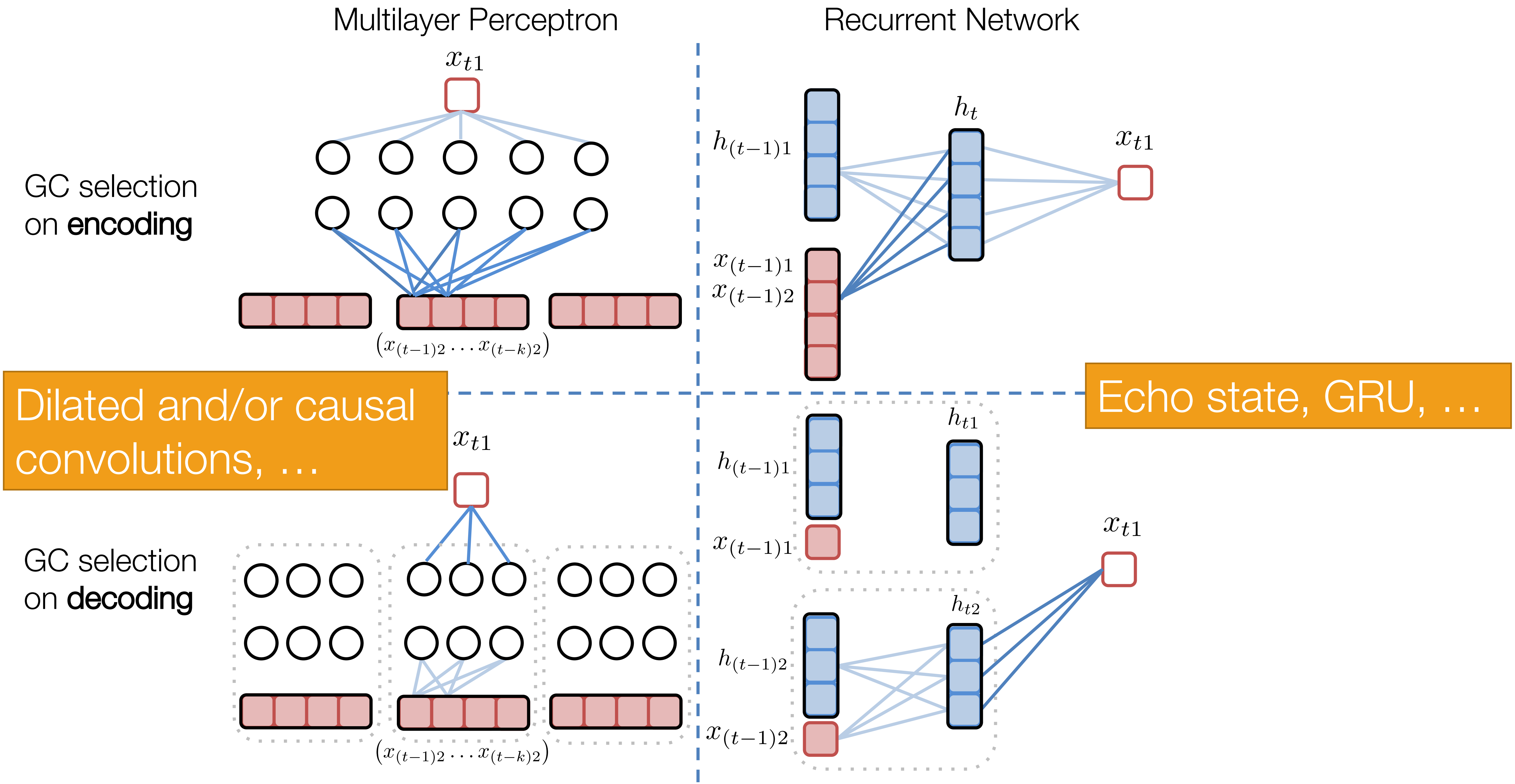
Learned MoCap interactions

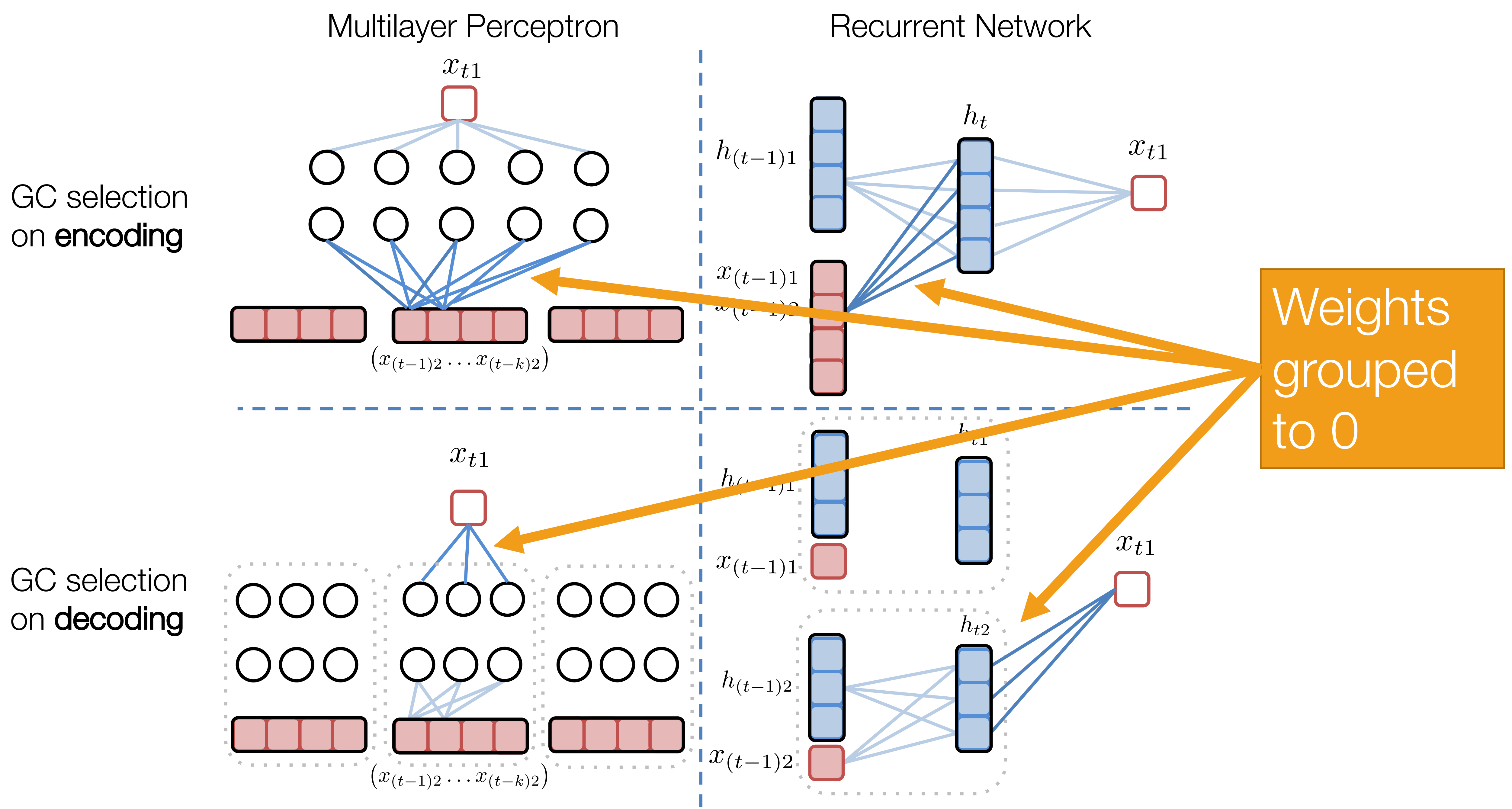


increasing sparsity penalty λ



Multilayer Perceptron
Recurrent Network
GC selection on **encoding**
GC selection on **decoding**
x_{t1}
$(x_{(t-1)2} \dots x_{(t-k)2})$
$h_{(t-1)1}$
$x_{(t-1)1}$
$x_{(t-1)2}$
h_t
h_{t1}
h_{t2}
$h_{(t-1)2}$
Linear combination of **nonlinear features**


Multilayer Perceptron
Recurrent Network
GC selection
on **encoding**
GC selection
on **decoding**
Dilated and/or causal convolutions, …
Echo state, GRU, …
x_{t1}
$(x_{(t-1)2} \dots x_{(t-k)2})$
$h_{(t-1)1}$
h_t
$x_{(t-1)1}$
$x_{(t-1)2}$
h_{t1}
$h_{(t-1)2}$
h_{t2}


Multilayer Perceptron
Recurrent Network
GC selection on **encoding**
GC selection on **decoding**
Weights grouped to 0
x_{t1}
$(x_{(t-1)2} \dots x_{(t-k)2})$
$h_{(t-1)1}$
$x_{(t-1)1}$
h_t
h_{t2}
$h_{(t-1)2}$
$x_{(t-1)2}$

Summary

Deep learning offers **tremendous opportunities** for modeling complex dynamics

- Traditional approaches often limited to linear, Gaussian, stationary, …

But, time series problems are much vaster than just prediction with large corpora

Characterizing dynamics	Efficiently sharing information	Interpretable interactions	Non-stationarity & measurement bias

Credit for the hard work…


Ian Covert
(CSE PhD)


Nick Foti
(Research Scientist)

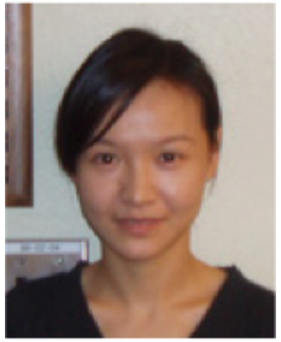

Chris Glynn
(Postdoc,
Asst Prof at UNH)


Alec Greaves-Tunnell
(Stat PhD)


Mike Hughes
(Brown CS PhD,
postdoc at Harvard)


Alex Tank
(Stat PhD)


Chris Xie
(CSE PhD)


Shirley You Ren
(Stat PhD,
Data Scientist at
Microsoft)