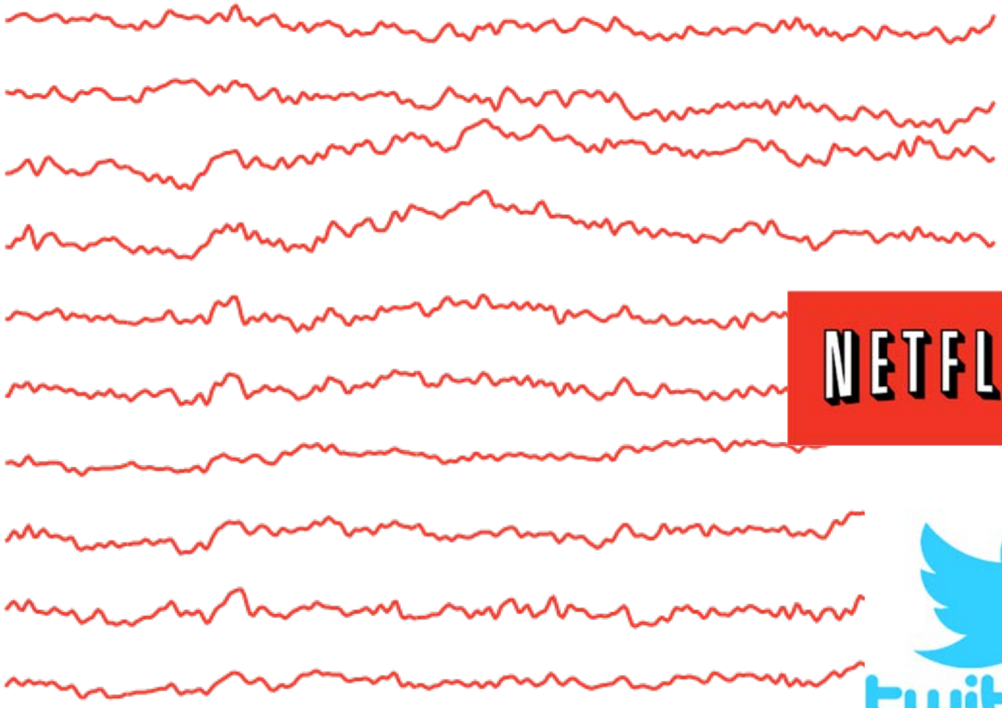
Flexibility, interpretability, and scalability in time series modeling

Emily Fox

University of Washington
Computer Science & Engineering (CSE) and Statistics

Modern sources of time series

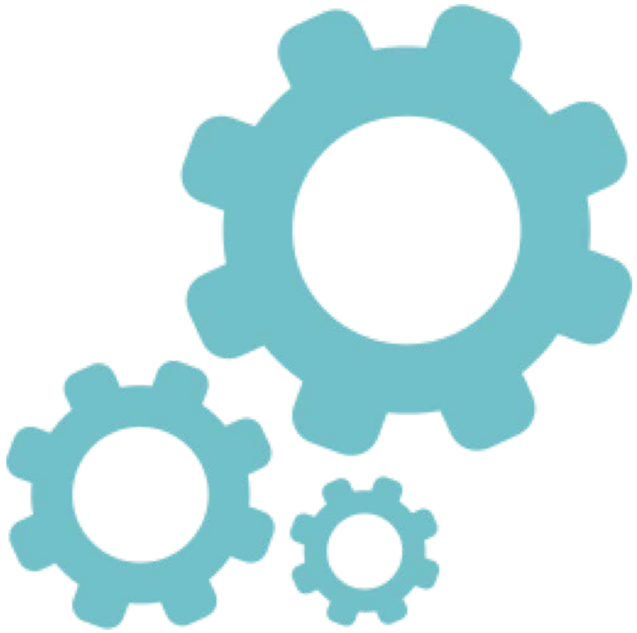


Until recently, ML (mostly) ignored time series

It's hard!

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

More data

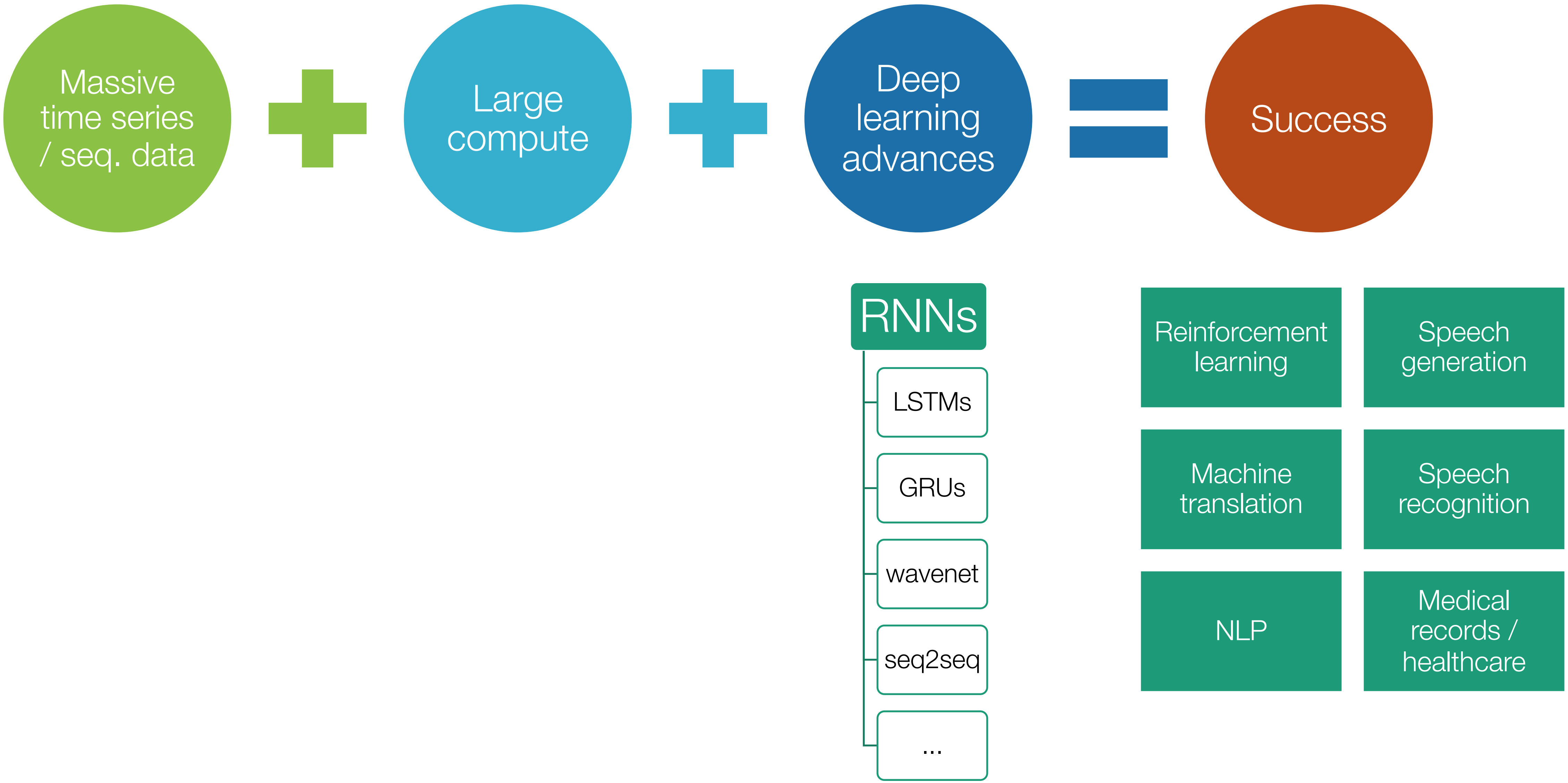


Algorithms more computationally intensive

More compute

Theory not applicable because typically assume no time dependencies

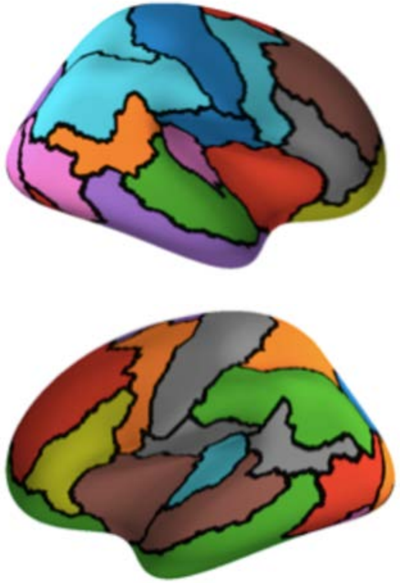
Now time series **are** “in”



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

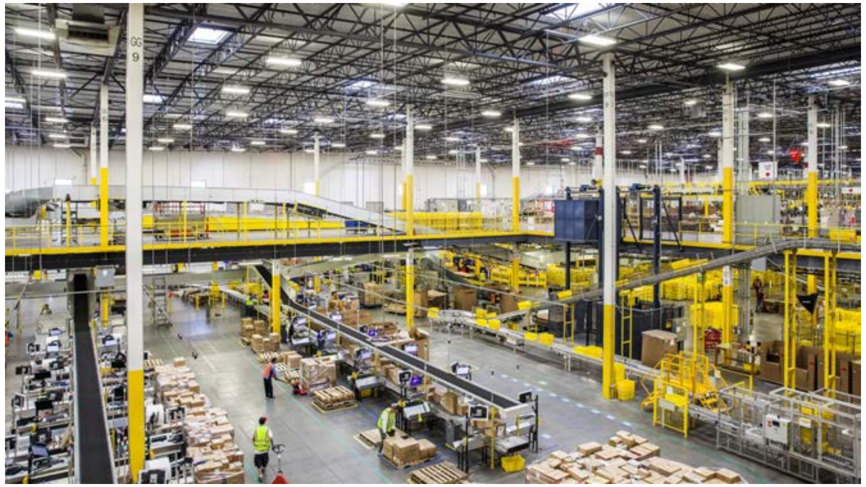


Inferring brain networks:
Costly data collection, significant subject-to-subject variability

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

Rare disease (or event) modeling:
Need to focus on tails of distribution

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

Changing context (non-stationarity):
Patient recovering or deteriorating, event-driven changes, etc.

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

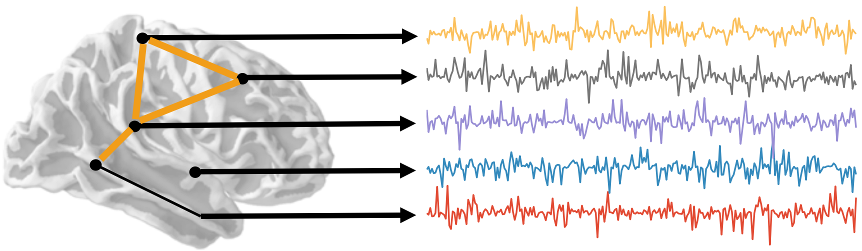
Clear prediction objective

- Word error rate for speech recognition
- BLEU score for machine translation
- Reward function in reinforcement learning

Few, low-trustworthy labels

No clear prediction metric

Structure learning, interpretability

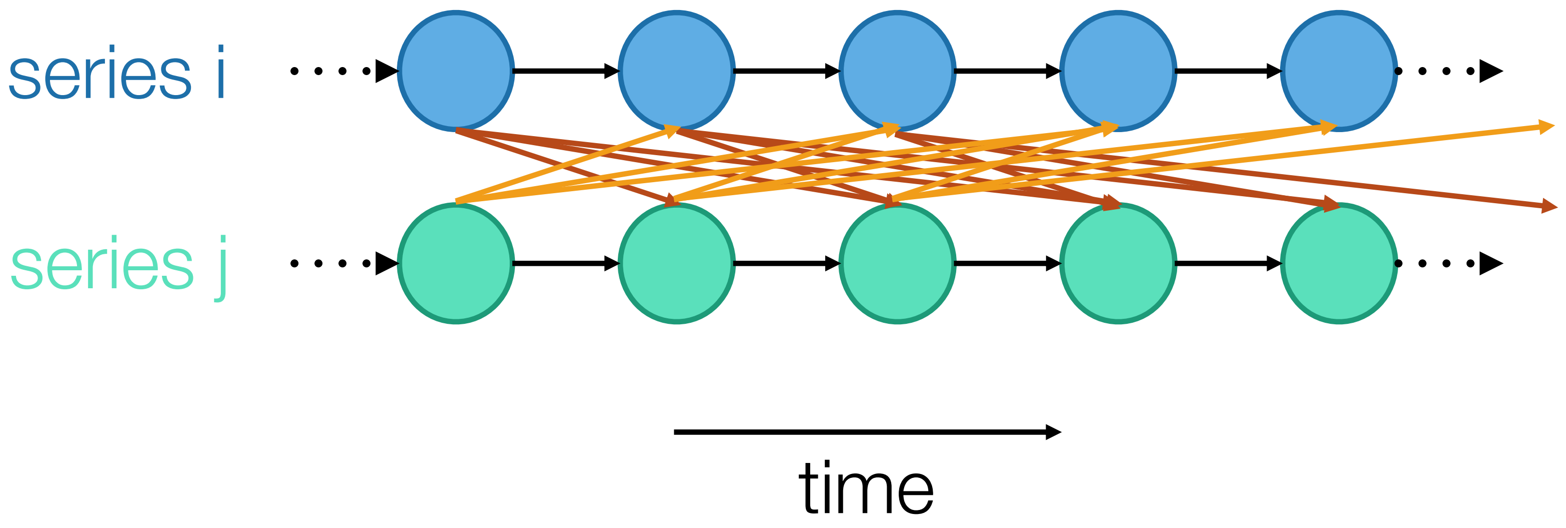


Interpretable interactions

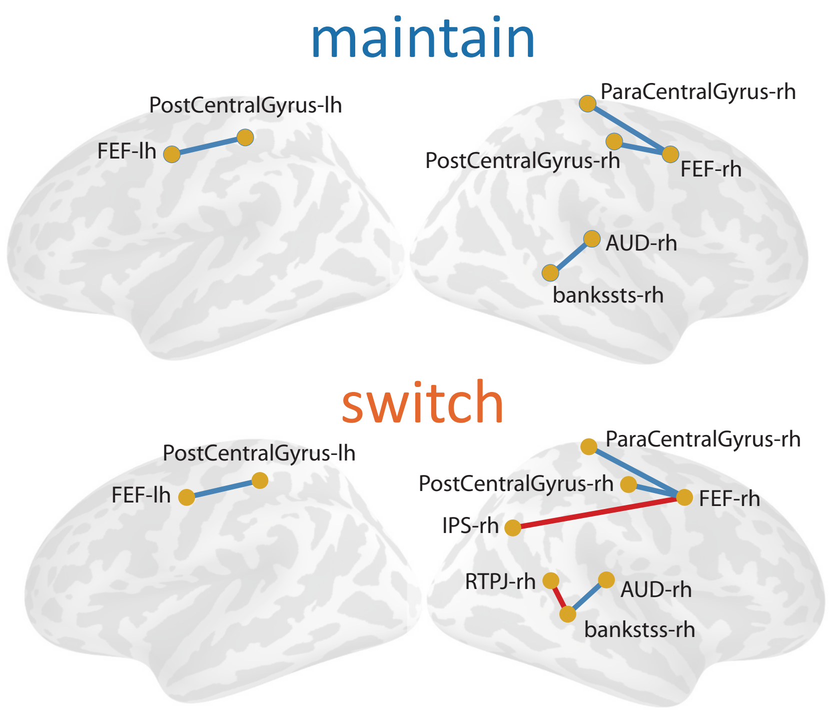
Modeling sparsely sampled, nonstationary time series

Handling bias in stochastic gradients of sequential data

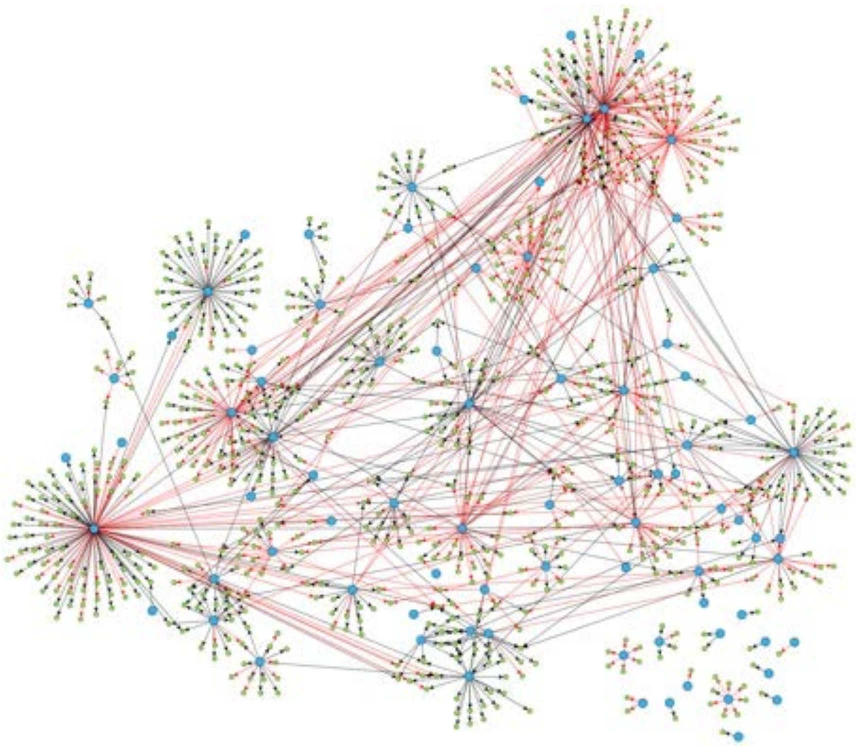
Granger causality:
Directed, lagged interactions in time series



Why are interactions important?

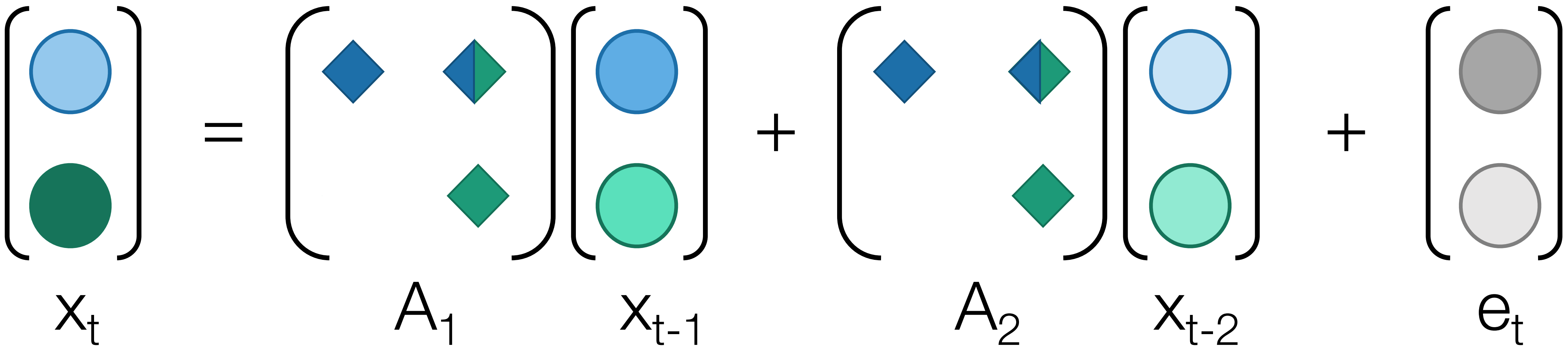


Functional networks in the brain



Gene regulatory networks

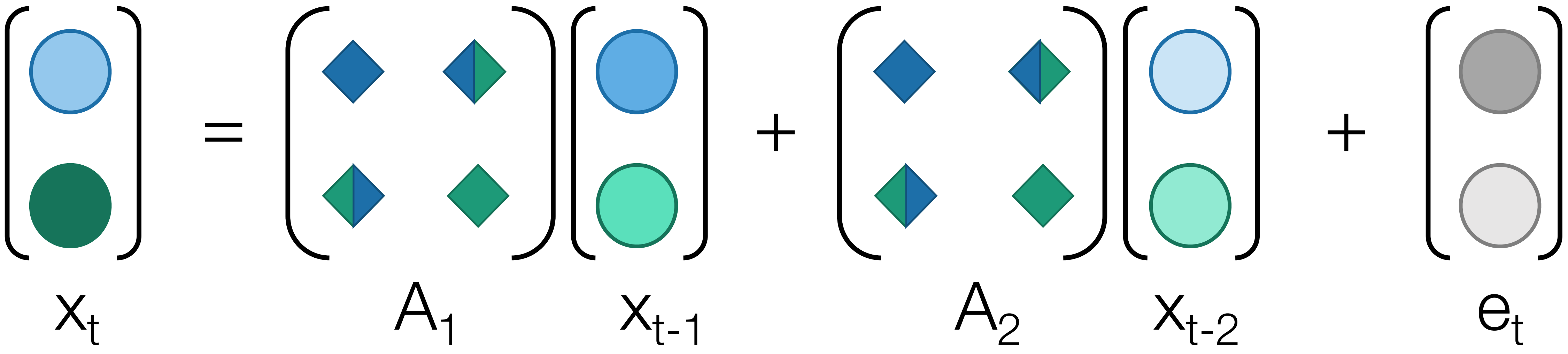
Granger causality selection – Linear model



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

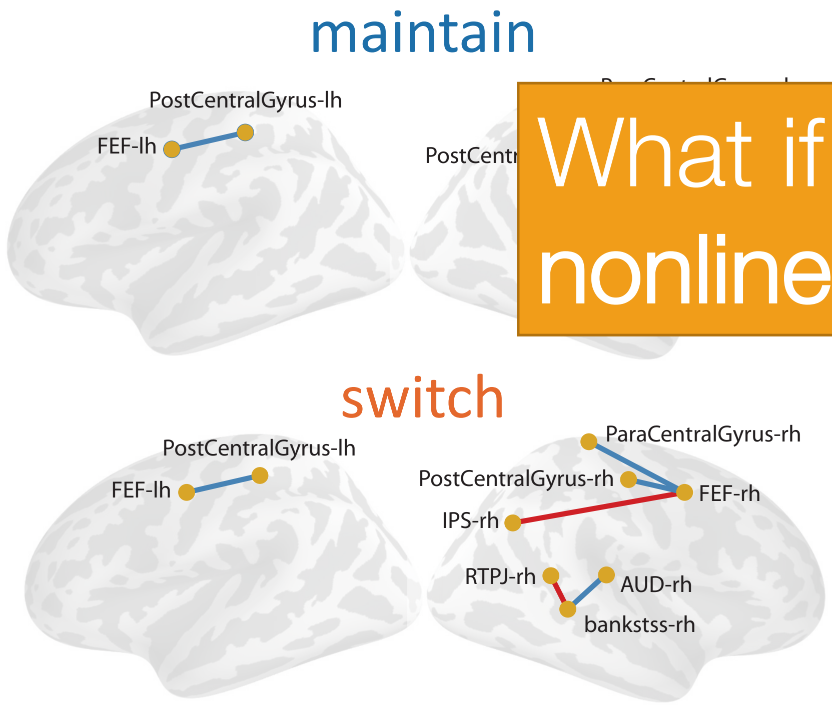
Lag k interaction

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

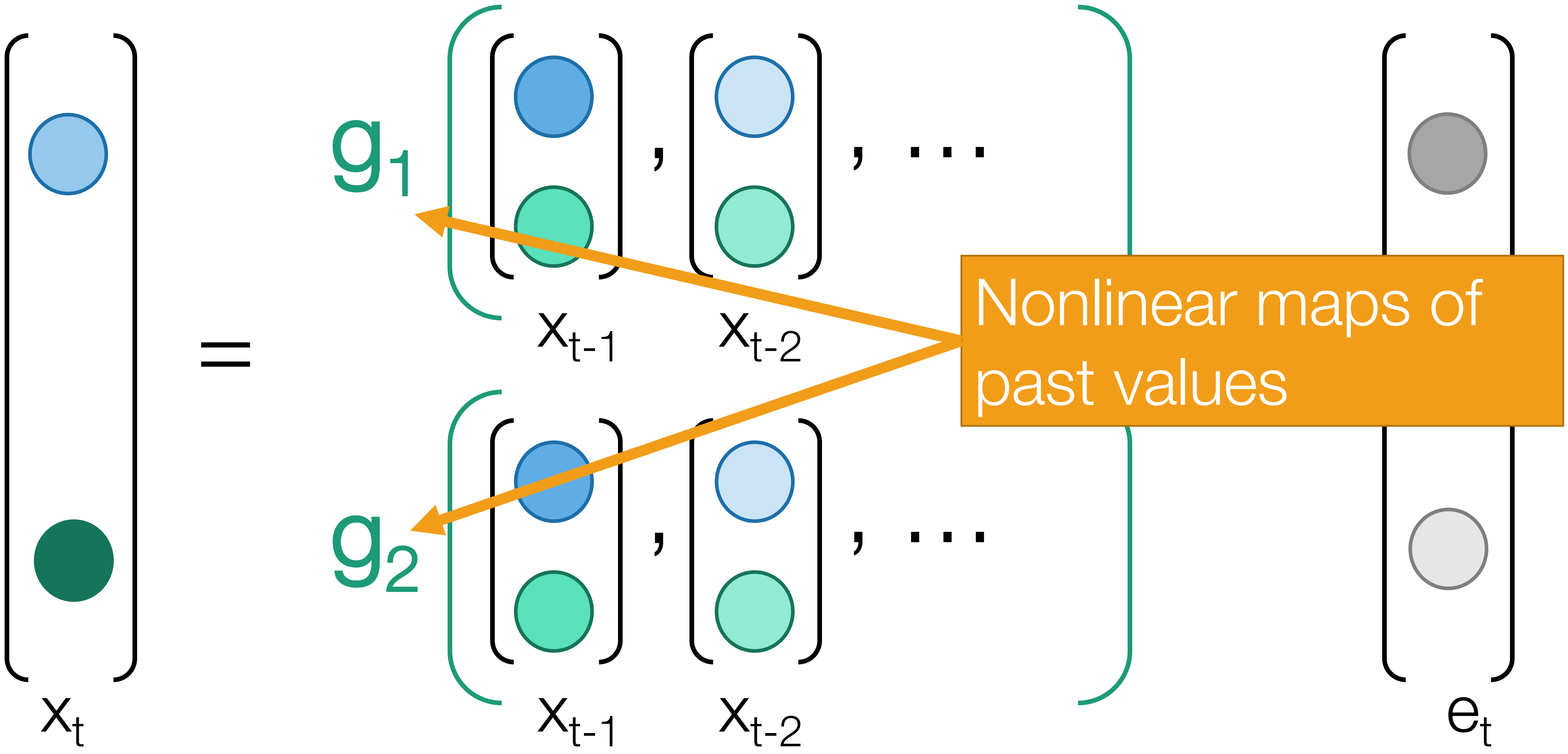


Functional networks in the brain

What if interactions are nonlinear?

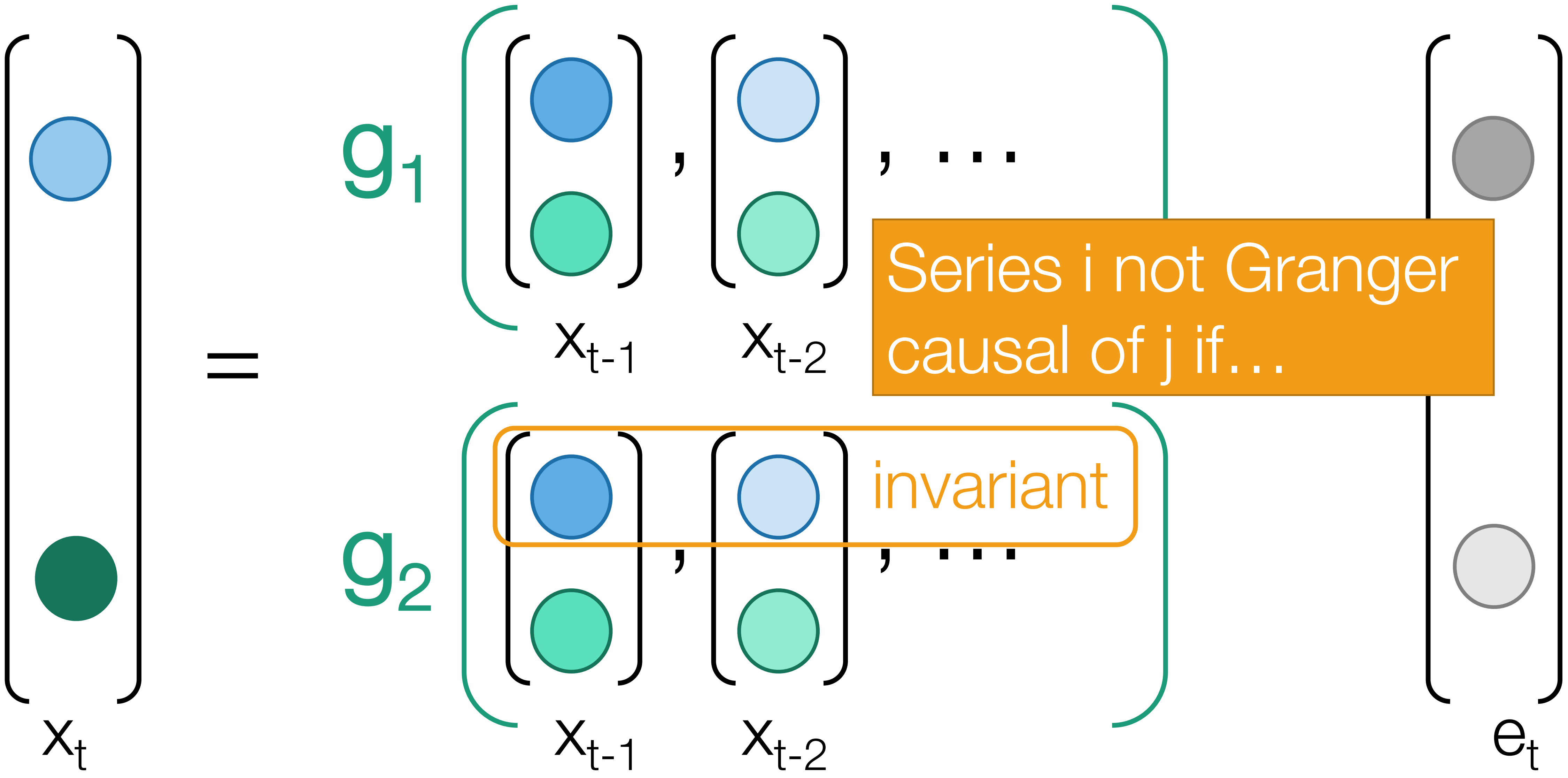
Gene regulatory networks

Modeling nonlinear dynamics



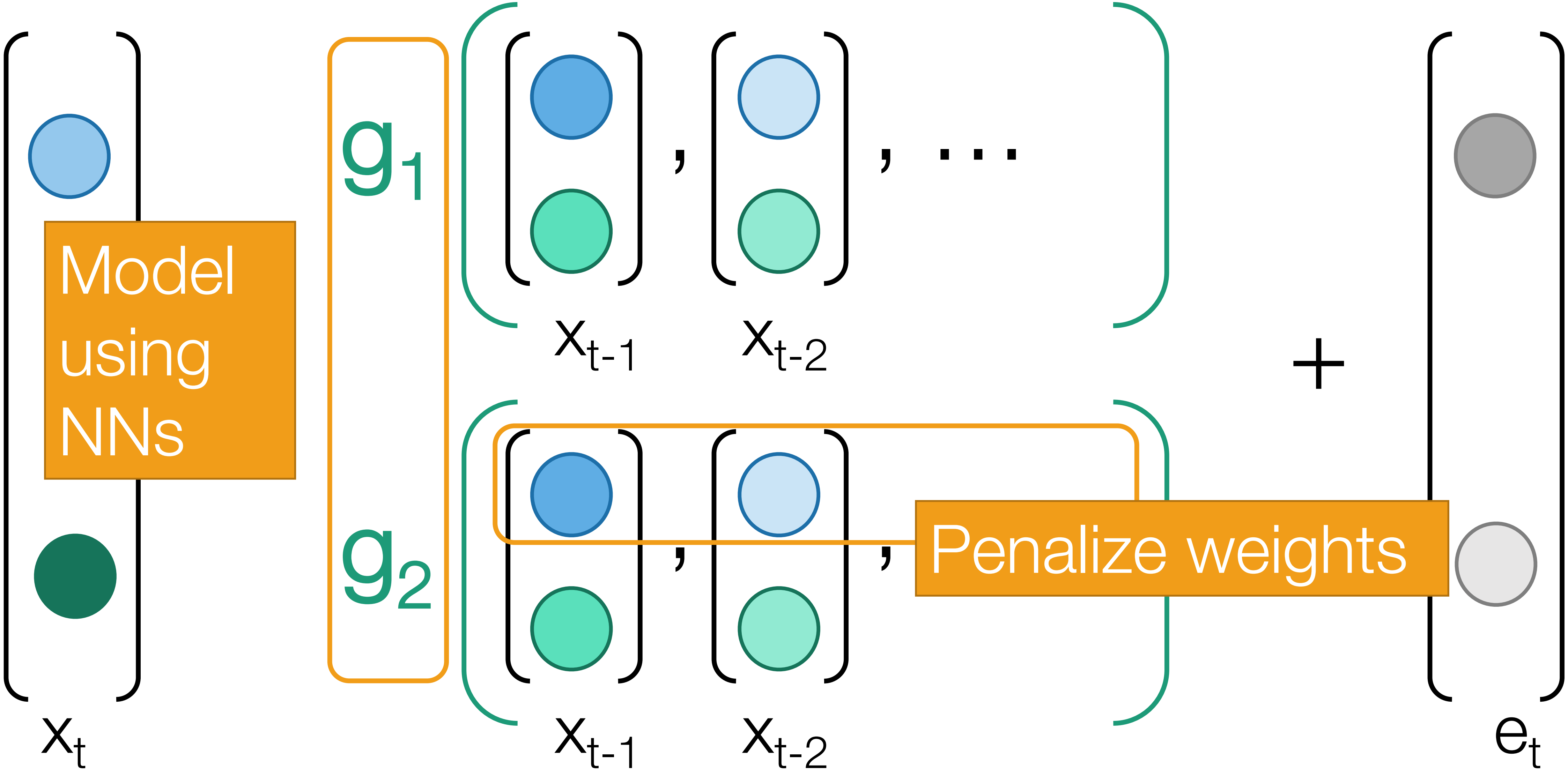
Tank, Covert, Foti, Shojaie, Fox, NIPS Time Series Workshop 2017, under review 2019

Identifying Granger causality



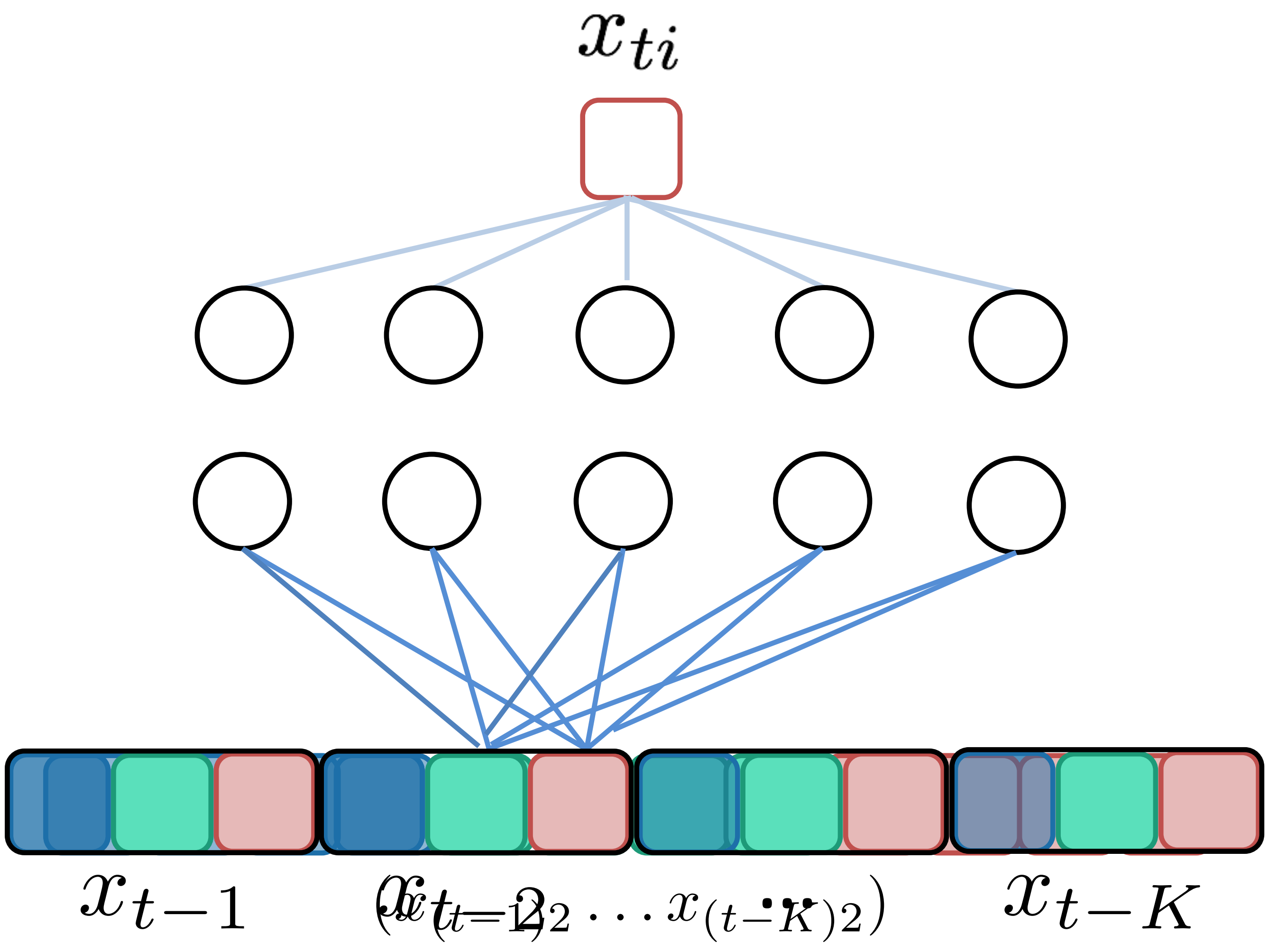
Tank, Covert, Foti, Shojaie, Fox, NIPS Time Series Workshop 2017, under review 2019

Using penalized neural networks



Tank, Covert, Foti, Shojaie, Fox, NIPS Time Series Workshop 2017, under review 2019

Penalized multilayer perceptron (MLP)

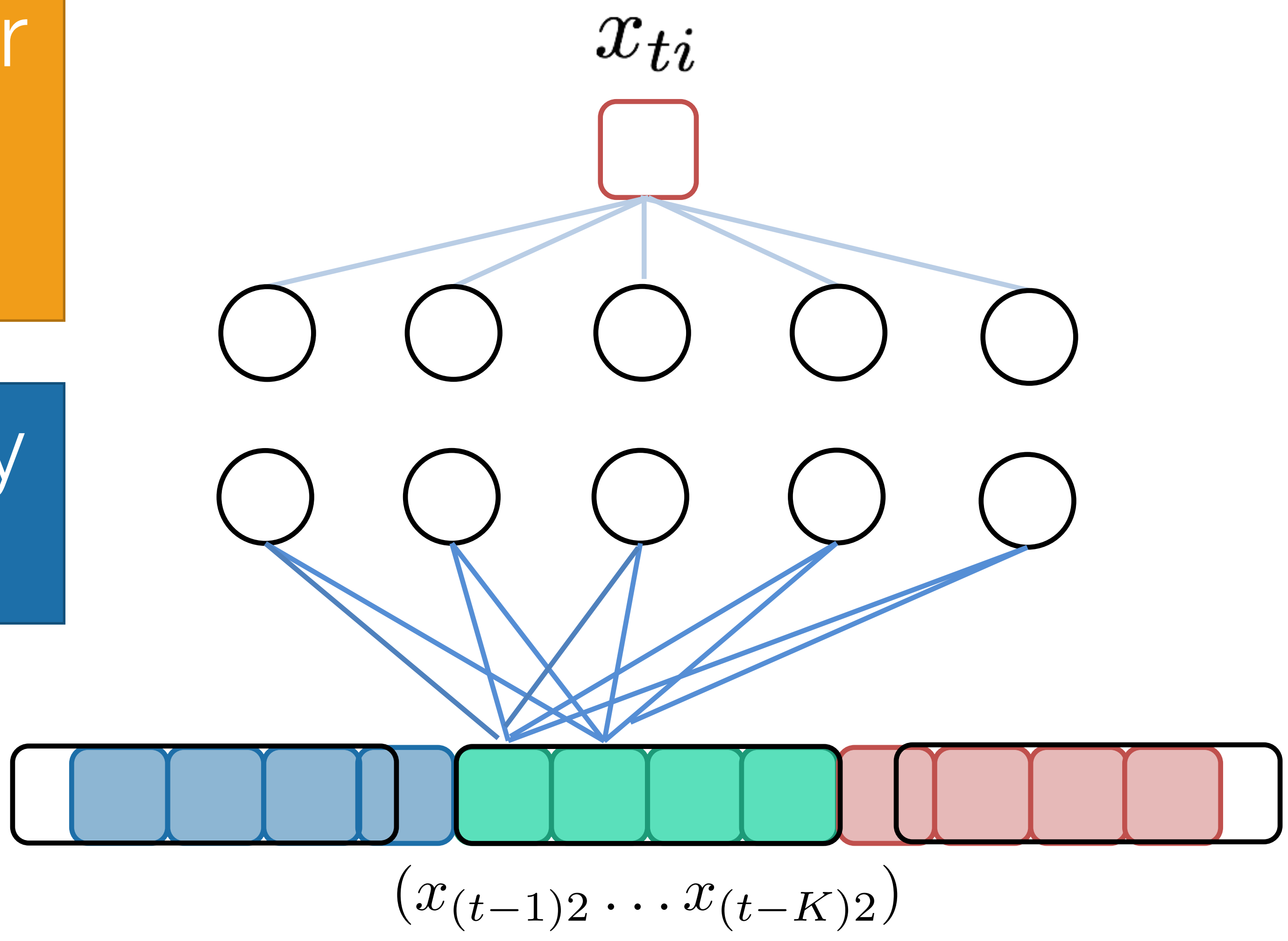


Penalized multilayer perceptron (MLP)

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

place group-wise penalty on layer 1 weights

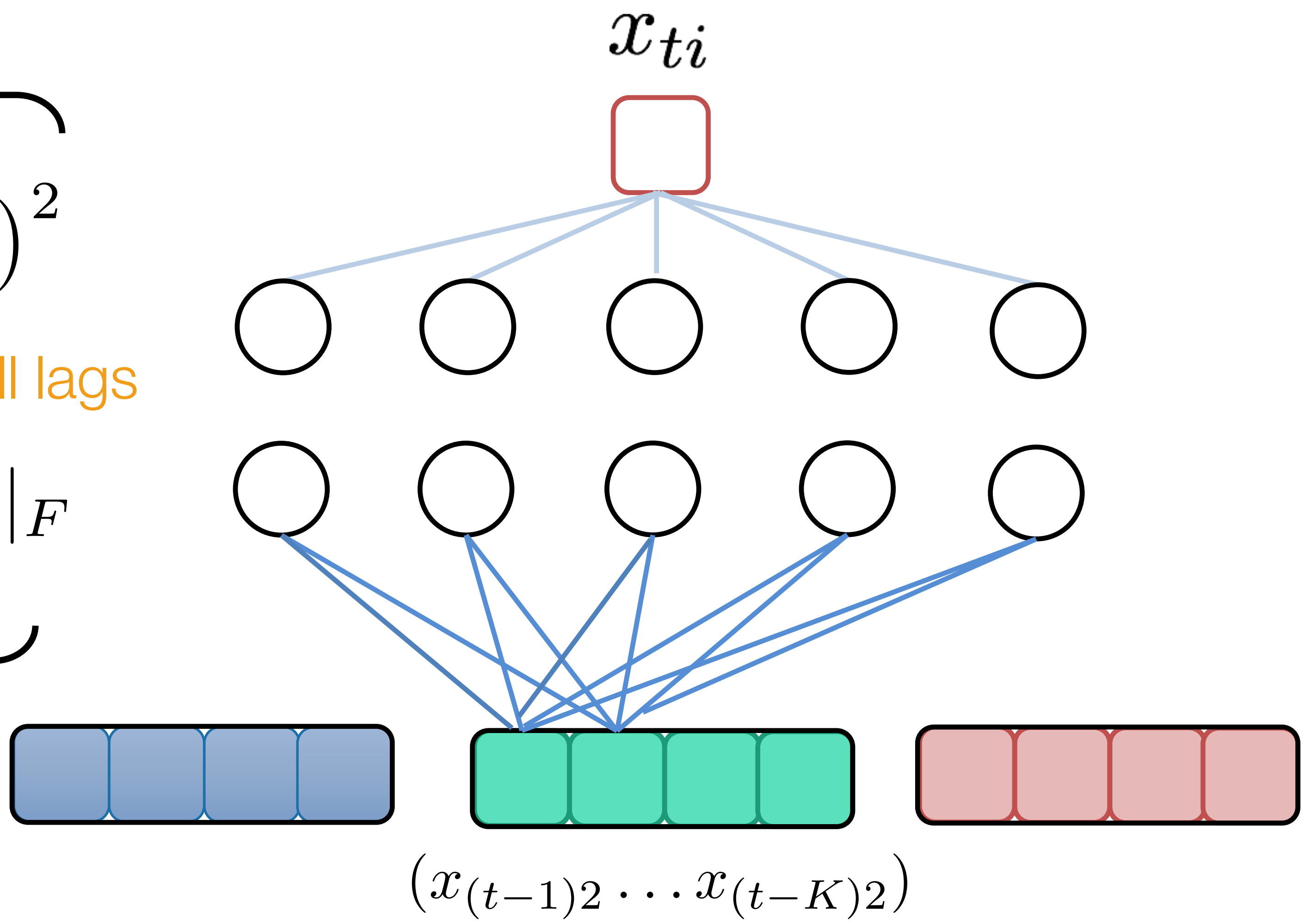
group inputs by:
(K lags of series j)



Penalized multilayer perceptron (MLP)

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

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

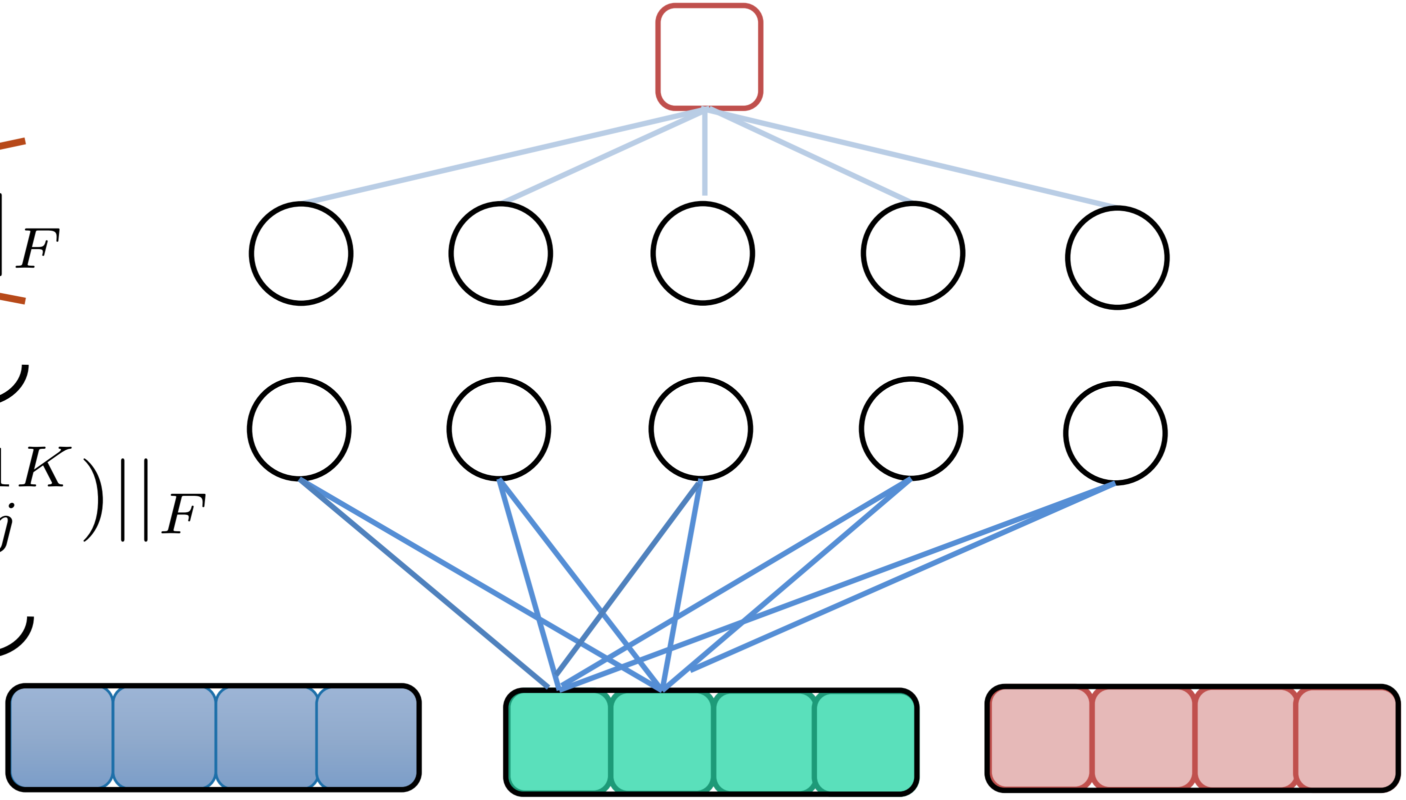


Lag selection via hierarchical group lasso

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

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

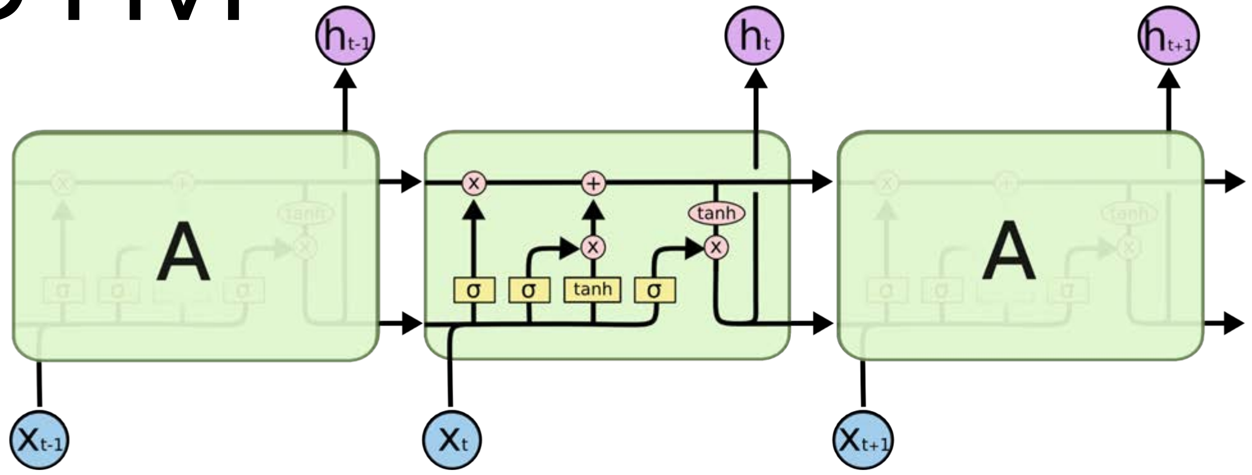
$$\underbrace{\lambda \sum_{j=1}^{p} \sum_{k=1}^{K} ||(W_{:j}^{1k}, \ldots, W_{:j}^{1K})||_F}_{\text{hierarchical group lasso penalty}}$$



$(x_{(t-1)2} \ldots x_{(t-K)2})$

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

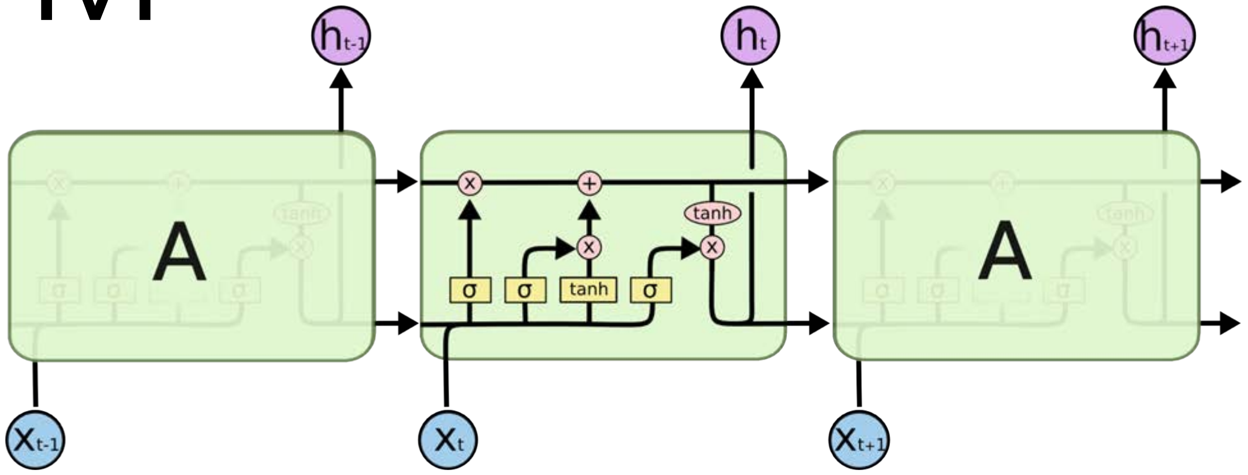
output gate $$o_t = \sigma\left(\boxed{W^o} x_t + U^o h_{(t-1)i}\right)$$

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

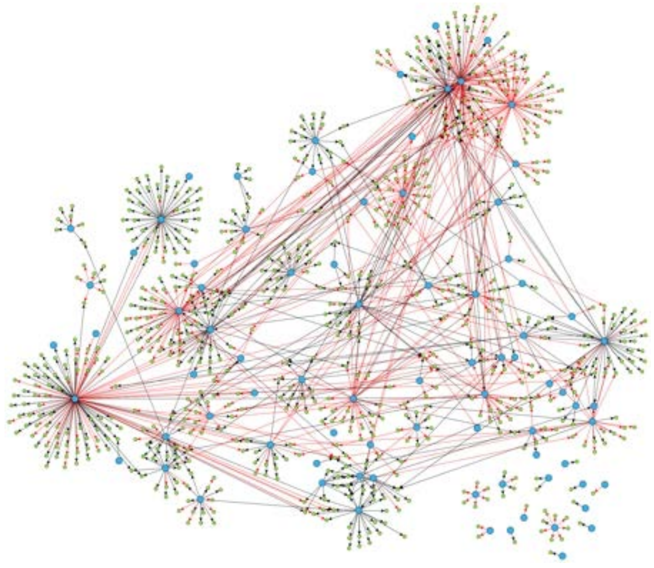
$$W = ((W^f)^T, (W^{in})^T, (W^o)^T, (W^c)^T)$$
define effect of input on prediction



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

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

DREAM3 challenge



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

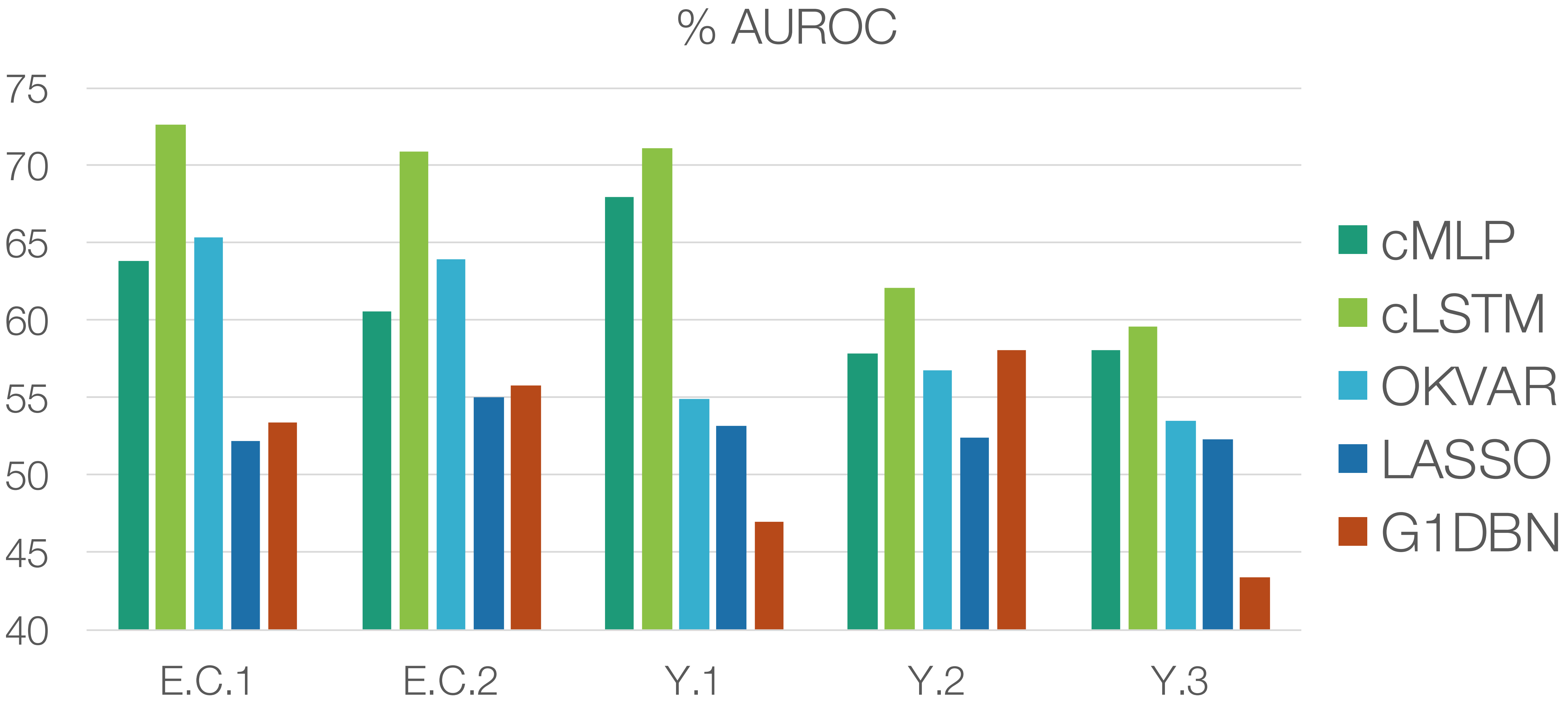
Simulated gene expression and regulation dynamics for:

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

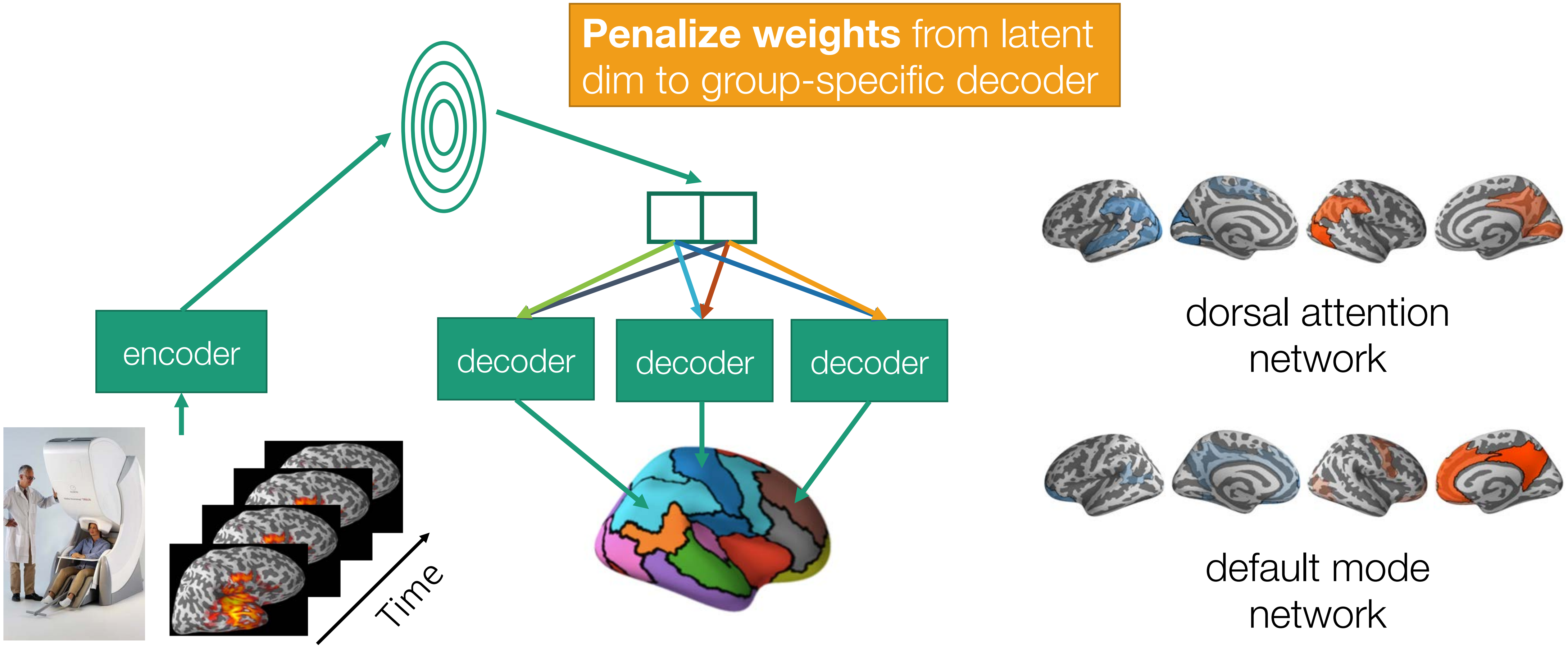
Very different structures

Structure extracted from currently established gene regulatory networks

DREAM3 results



Capturing contemporaneous interactions via structured deep generative models



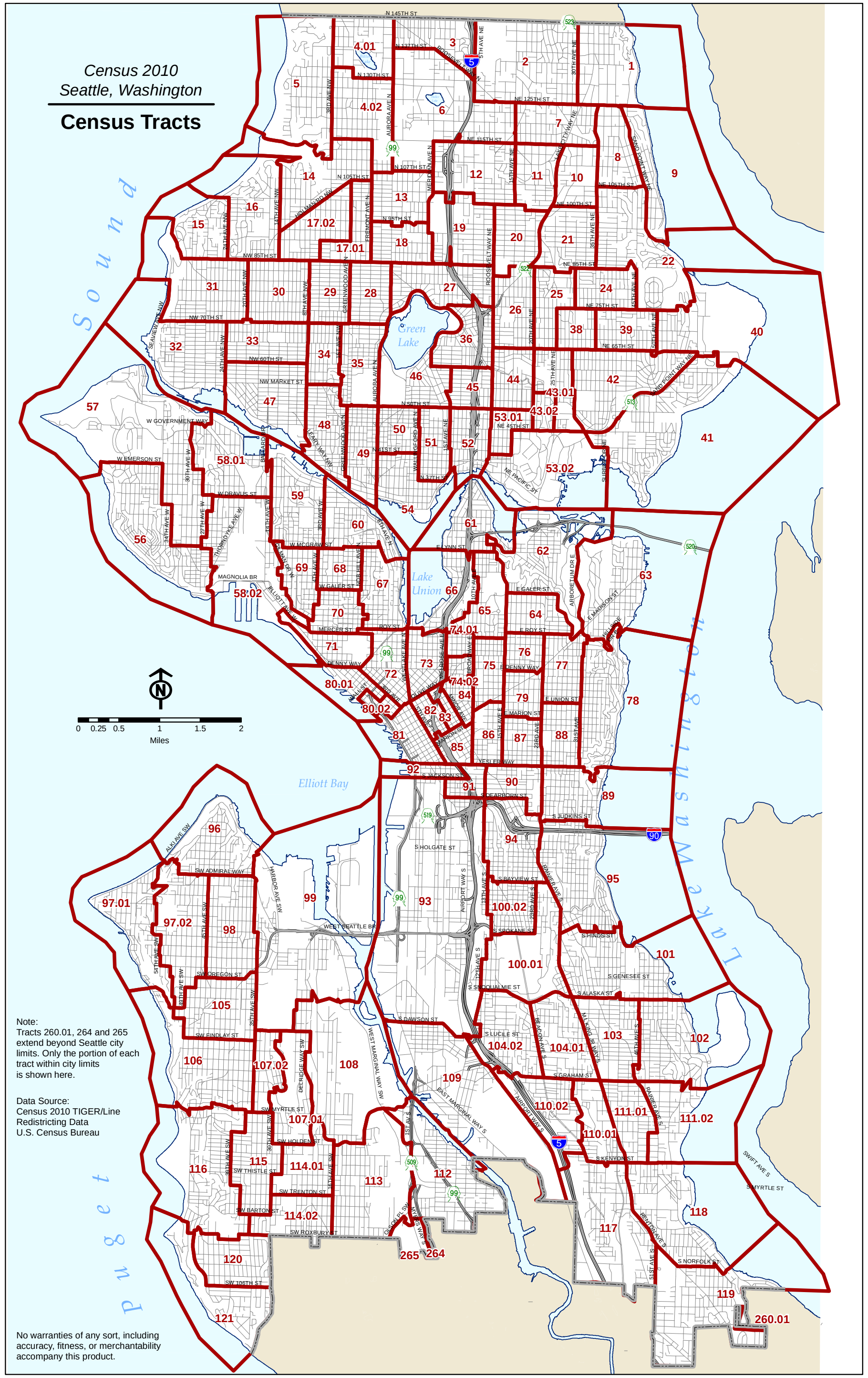
Ainsworth, Foti, Lee, Fox, ICML 2018

Interpretable interactions

Modeling sparsely sampled, nonstationary time series

Handling bias in stochastic gradients of sequential data

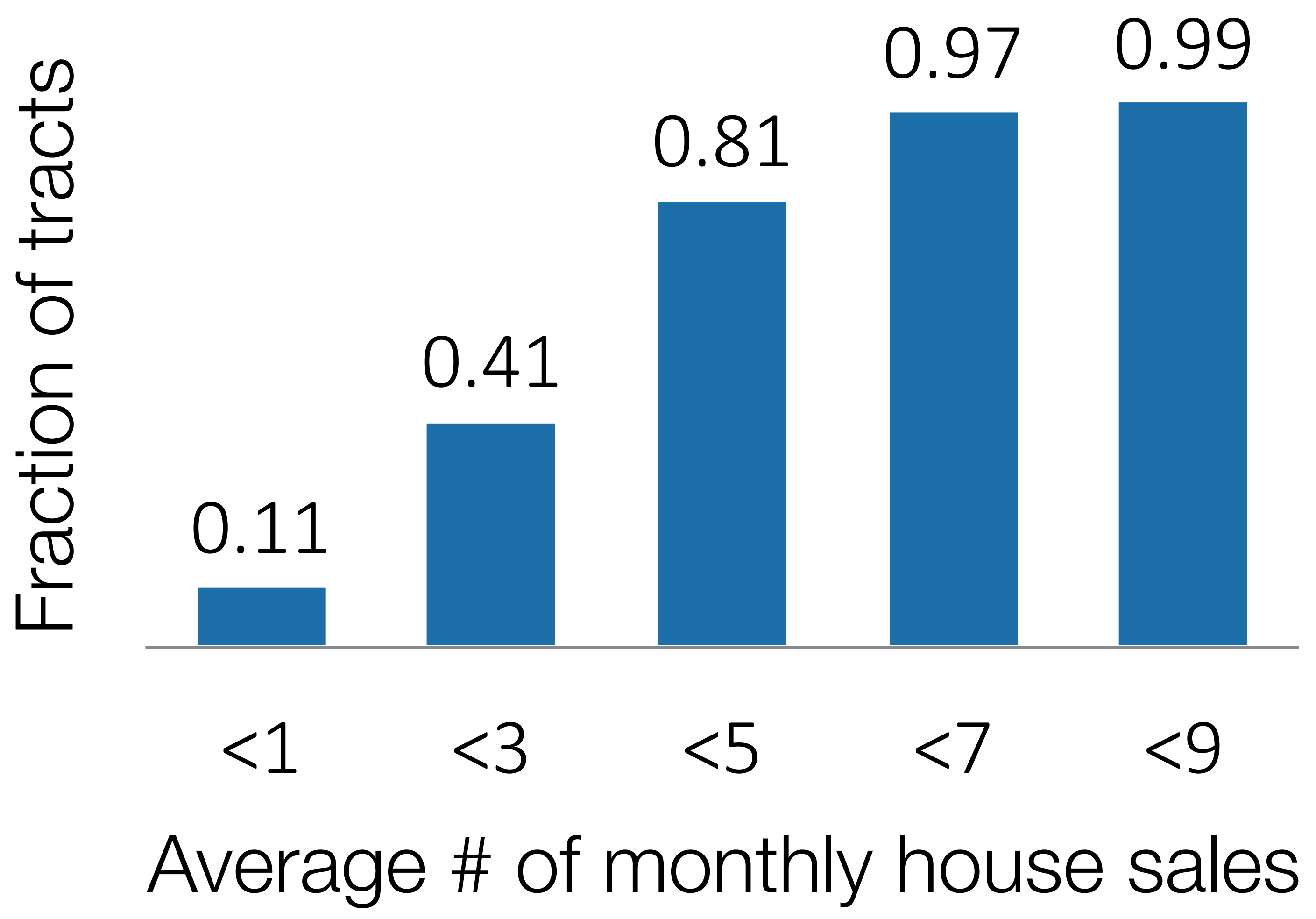
modeling a local housing index



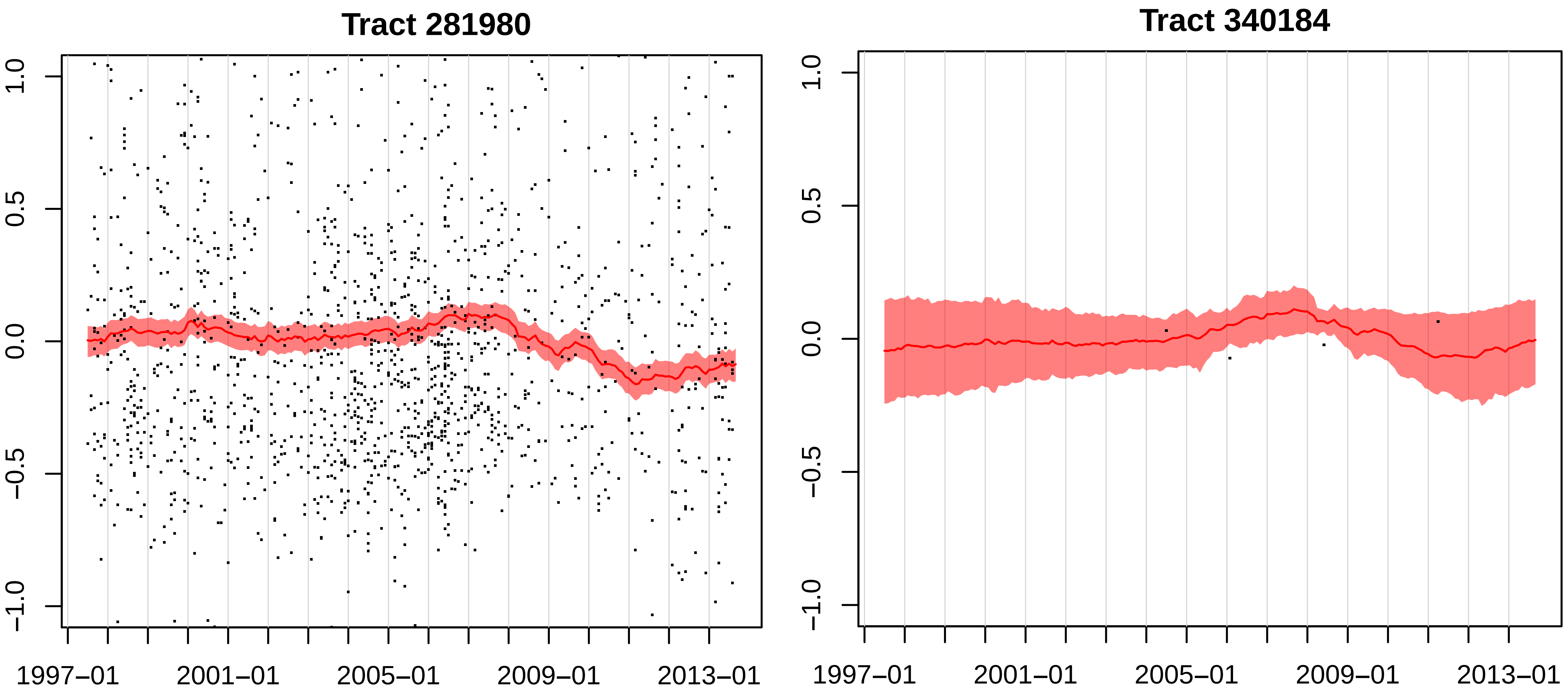
Census tracts in Seattle, WA

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

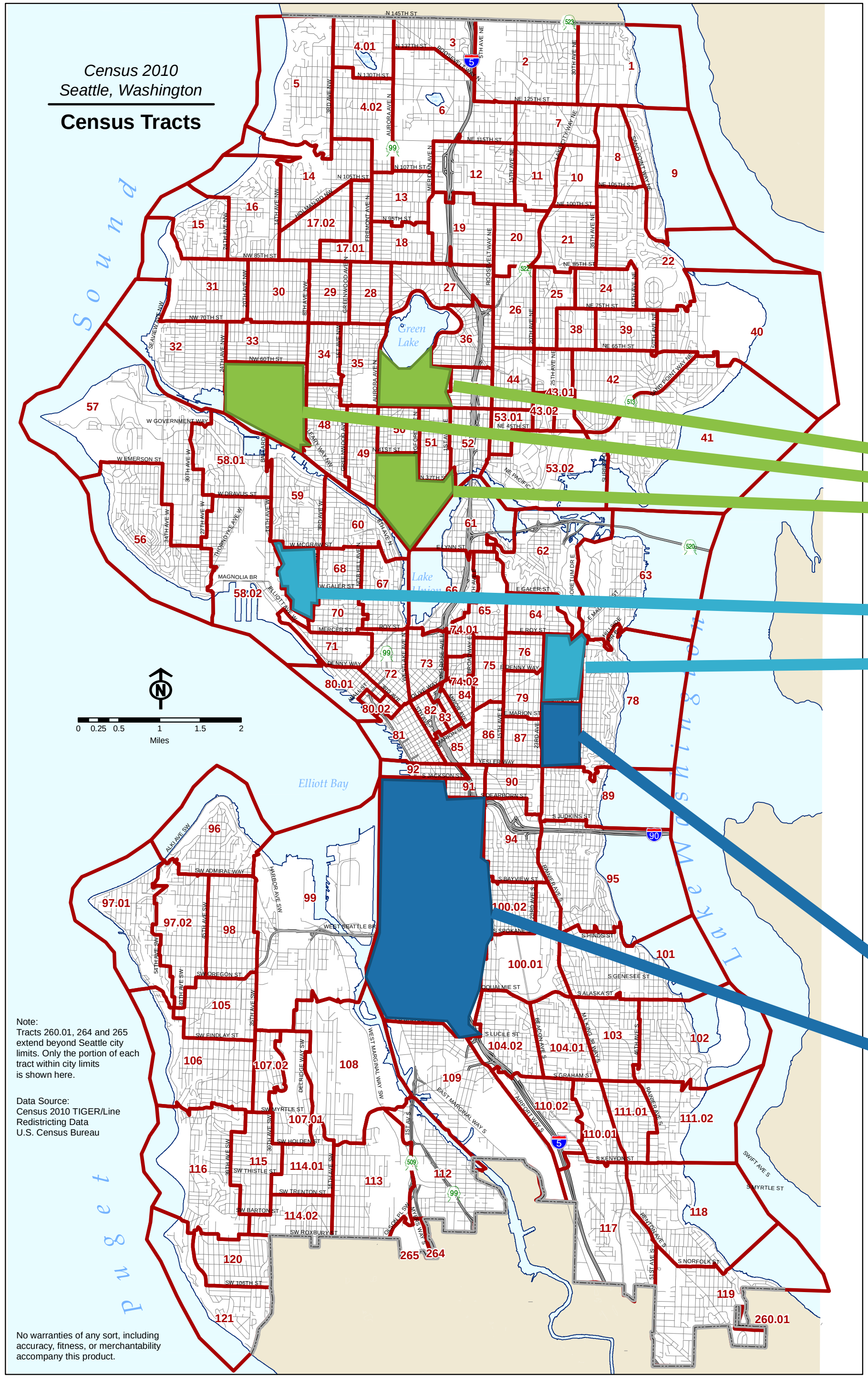
Challenge: Spatiotemporally sparse data

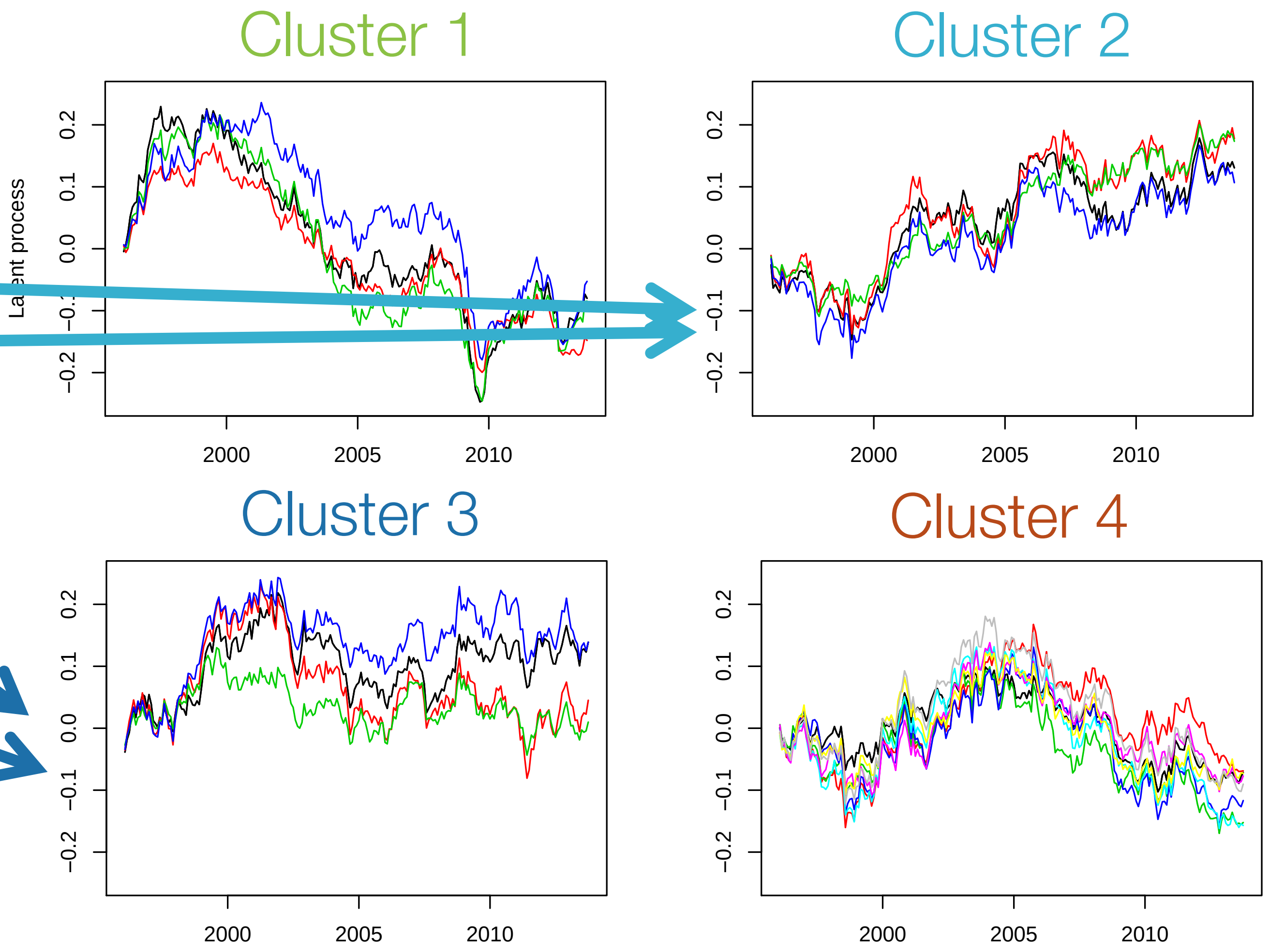


Challenge: Spatiotemporally sparse data



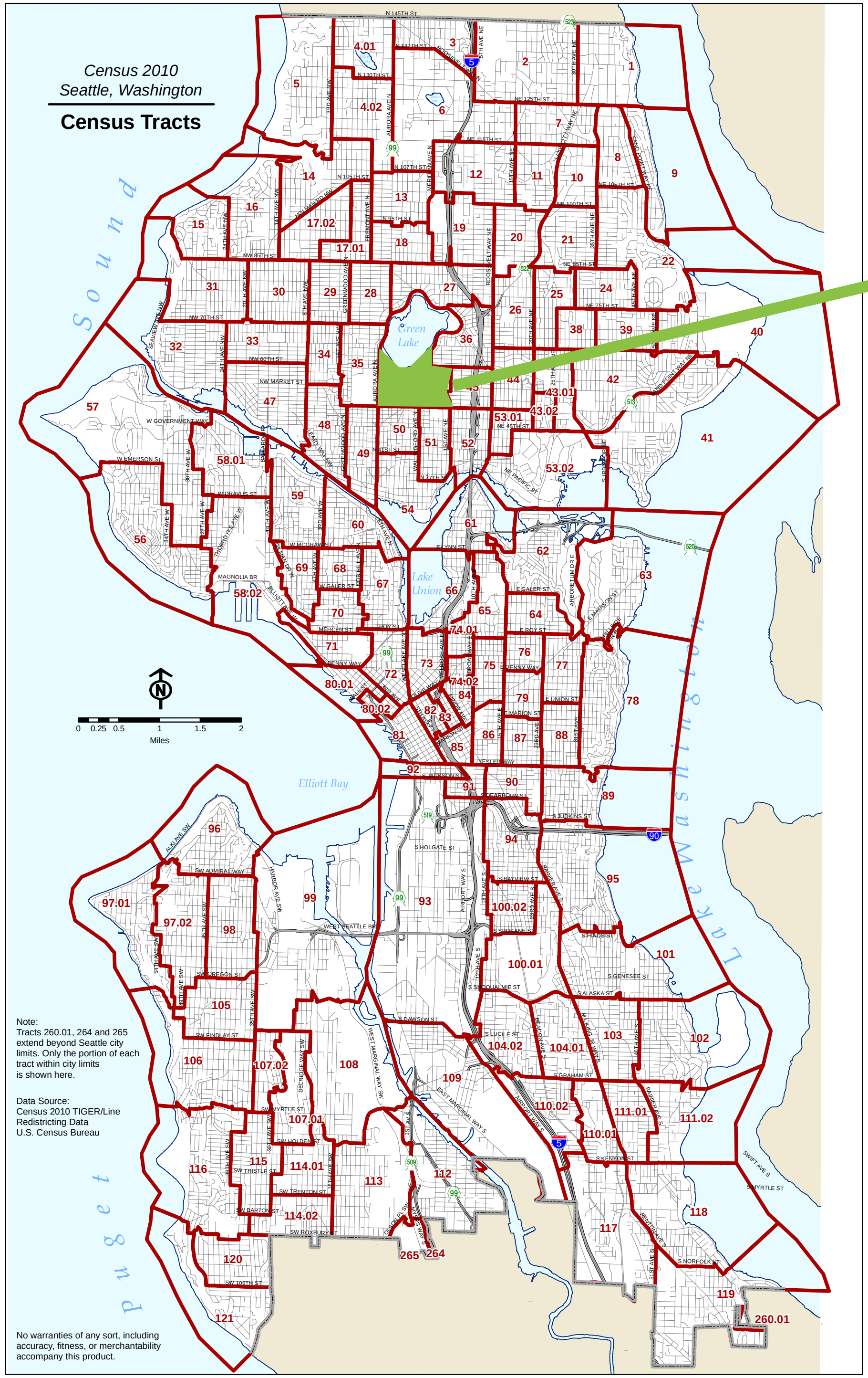
Solution: Discover clusters of latent price dynamics



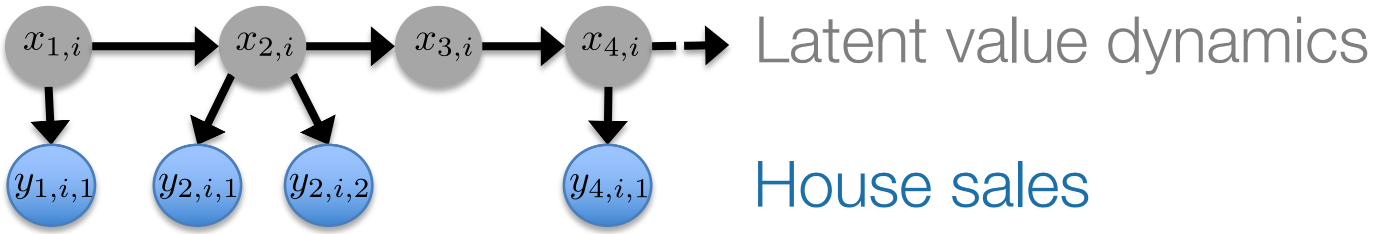


...

Ren, Fox, Bruce, Annals of Applied Statistics 2017



Single census tract model



tract *i*

Hidden: global trend + seasonality

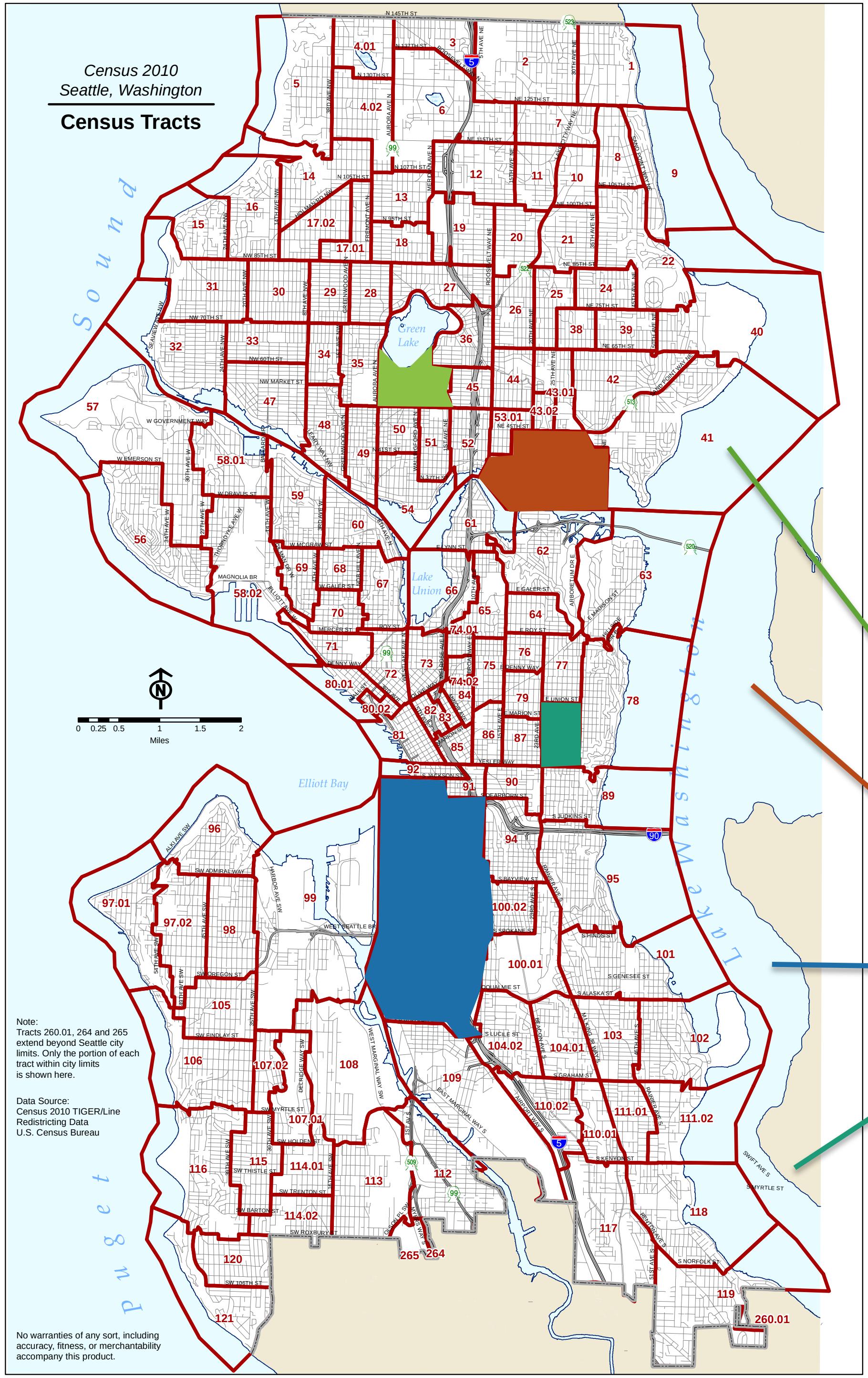
$$x_{t,i} = a_i x_{t-1,i} + \epsilon_{t,i} \quad \epsilon_{t,i} \sim \mathcal{N}(0, \sigma_i^2)$$

$$y_{t,i,l} = x_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,r} + v_{t,i,l} \quad v_{t,i,l} \sim \mathcal{N}(0, R_i)$$

***l*th sales**

house-level features

Ren, Fox, Bruce, Annals of Applied Statistics 2017



Multiple census tract model

tract *i*

Hidden: global trend + seasonality

"innovations"

$$x_{t,i} = a_i x_{t-1,i} + \boxed{\epsilon_{t,i}} \quad \epsilon_{t,i} \sim \mathcal{N}(0, \sigma_i^2)$$

$$y_{t,i,l} = x_{t,i} + \sum_{h=1}^{H} \beta_{i,h} U_{l,r} + v_{t,i,l} \quad v_{t,i,l} \sim \mathcal{N}(0, R_i)$$

l^{th} sales

house-level features

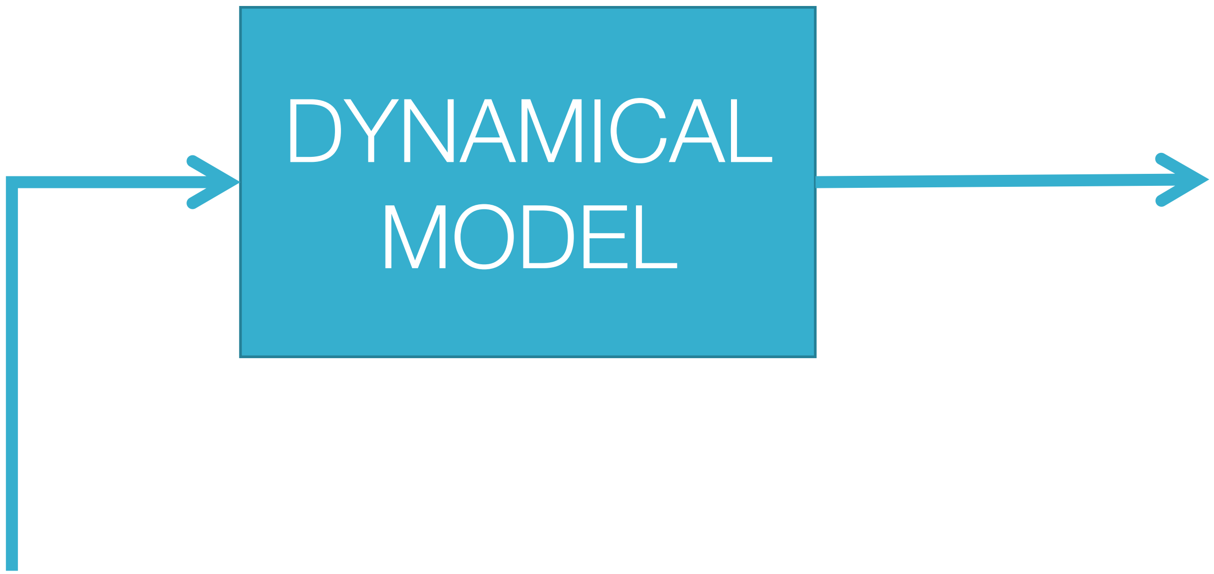
p-dim vector

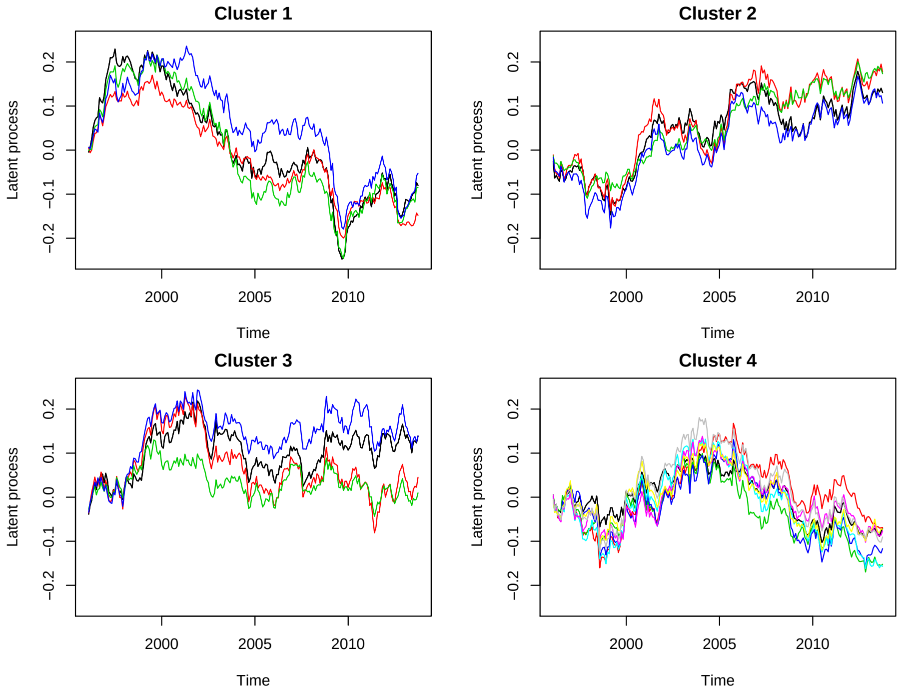
$$\epsilon_t \sim N(0, \Sigma)$$

Discover block structure

Ren, Fox, Bruce, Annals of Applied Statistics 2017

Cluster and correlate multiple time series



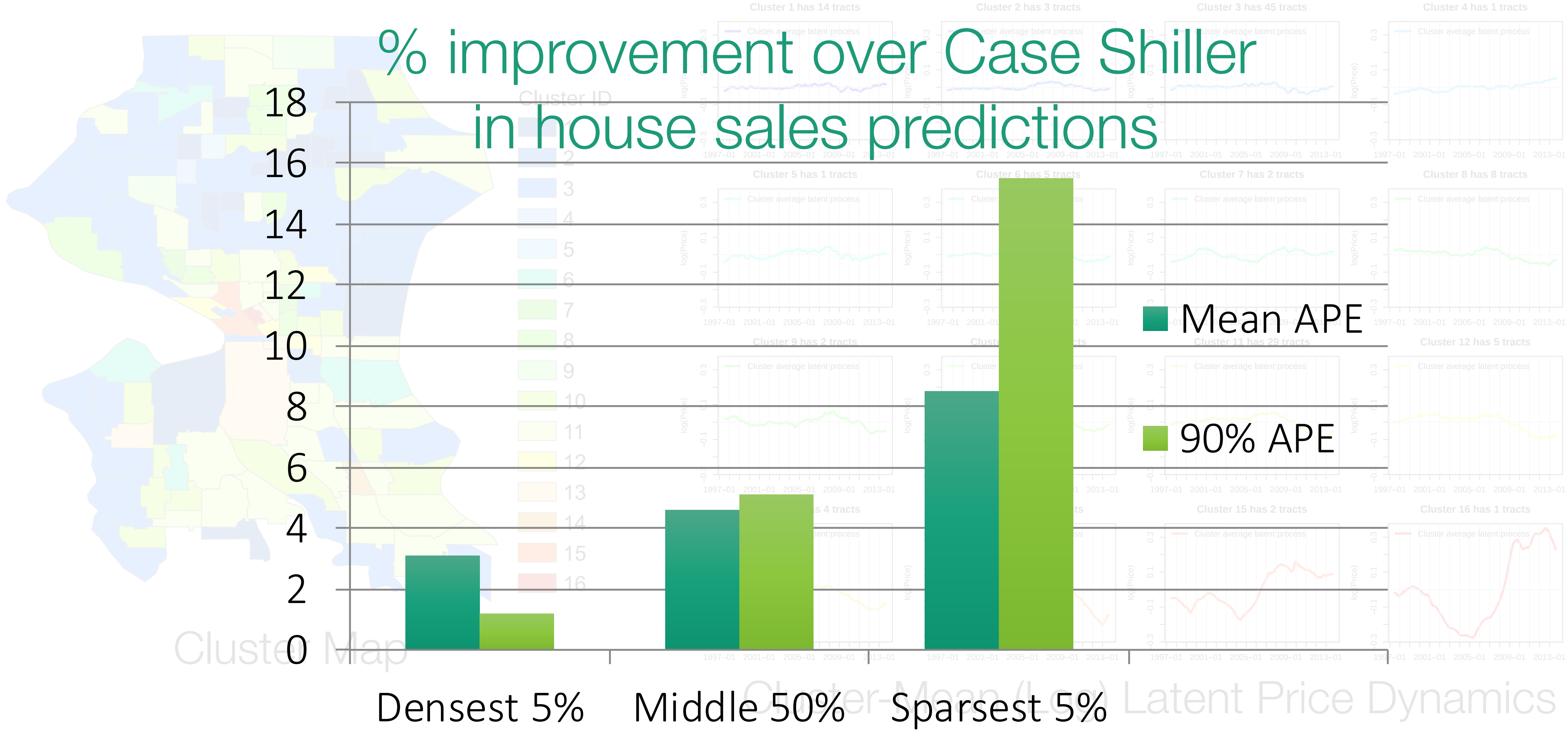


$$\begin{pmatrix} \epsilon_{t,1} \\ \epsilon_{t,2} \\ \epsilon_{t,3} \\ \epsilon_{t,4} \\ \vdots \\ \epsilon_{t,p} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \left(\quad \Sigma \quad \right) \right]$$

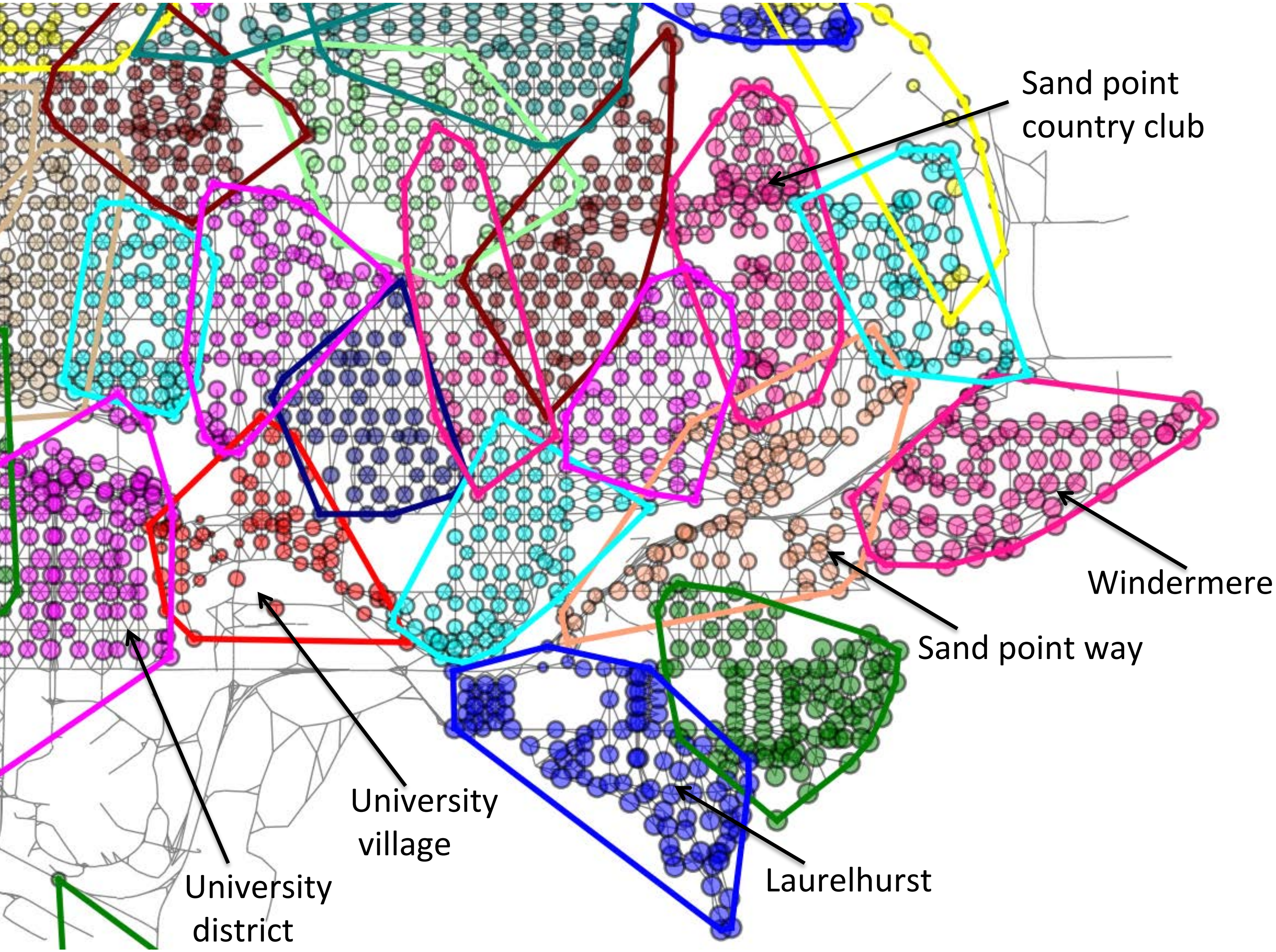
Latent factor model
+
Bayesian nonparametrics

Ren, Fox, Bruce, Annals of Applied Statistics 2017

Seattle City analysis



Robustness to even finer scales



Heuristically defined neighborhoods

Smaller than census tracts

5% improvement in predictive performance!

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

Goals:

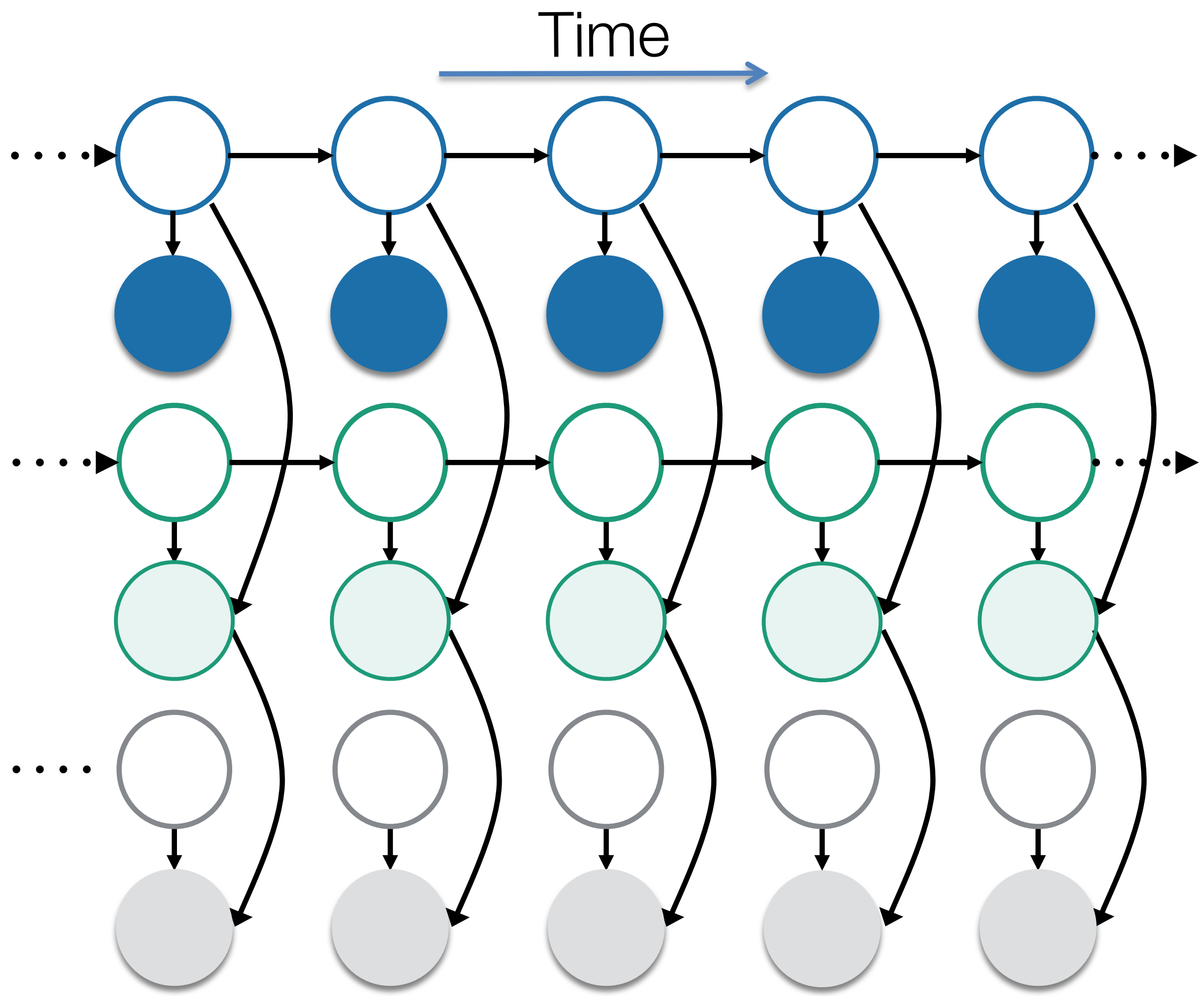
- Studying time-varying homeless populations **locally**
- Infer effect of **increases in rent** to homelessness rate
- **Forecast** future homeless population for decision-making
- Robustly quantify **uncertainty**

Data challenges:

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

measurement bias!

Per-metro count-based dynamical model



(Nonstationary) population dynamics

Noisy census counts **(observed)**

Log odds of homelessness
regressed on Zillow Rent Index (ZRI)

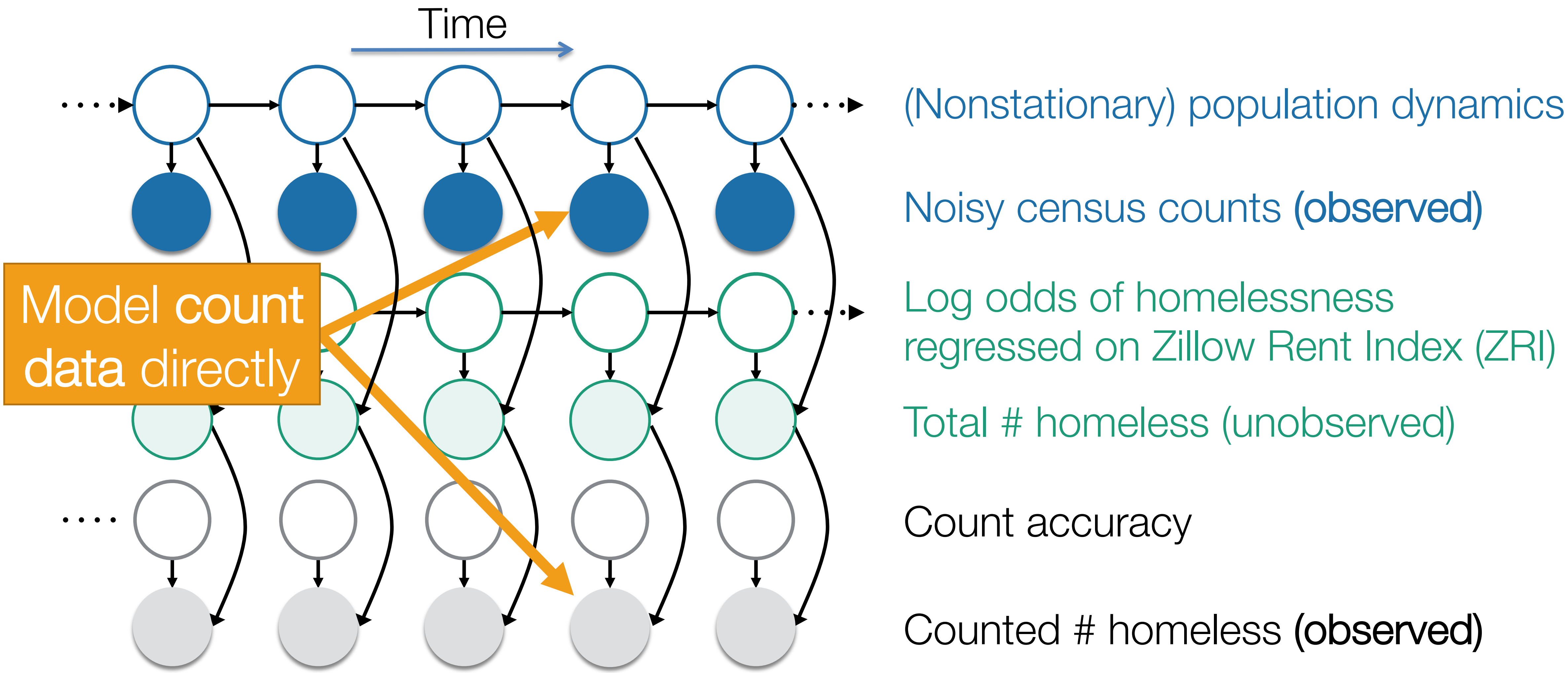
Total # homeless (unobserved)

Count accuracy

Counted # homeless **(observed)**

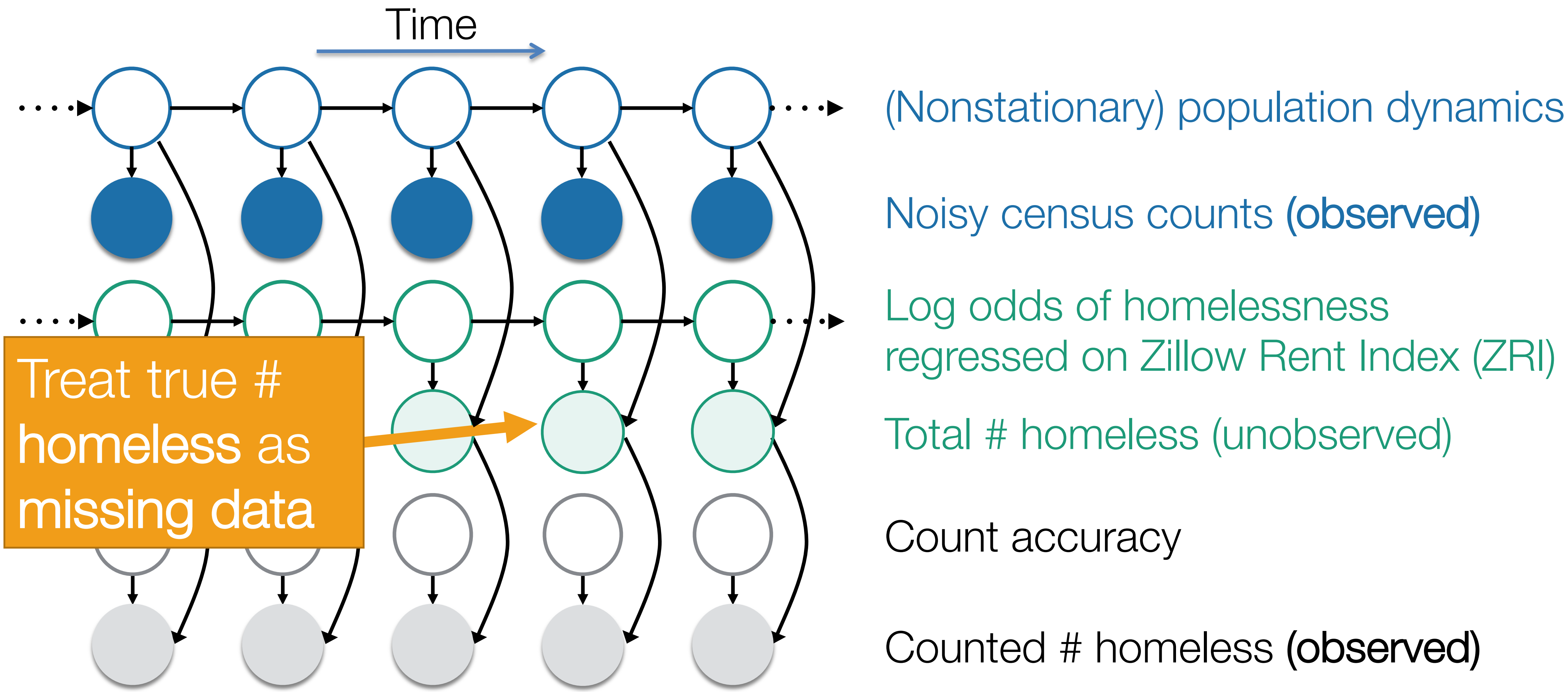
Glynn and Fox, Annals of Applied Statistics 2019

Benefits over past approaches…



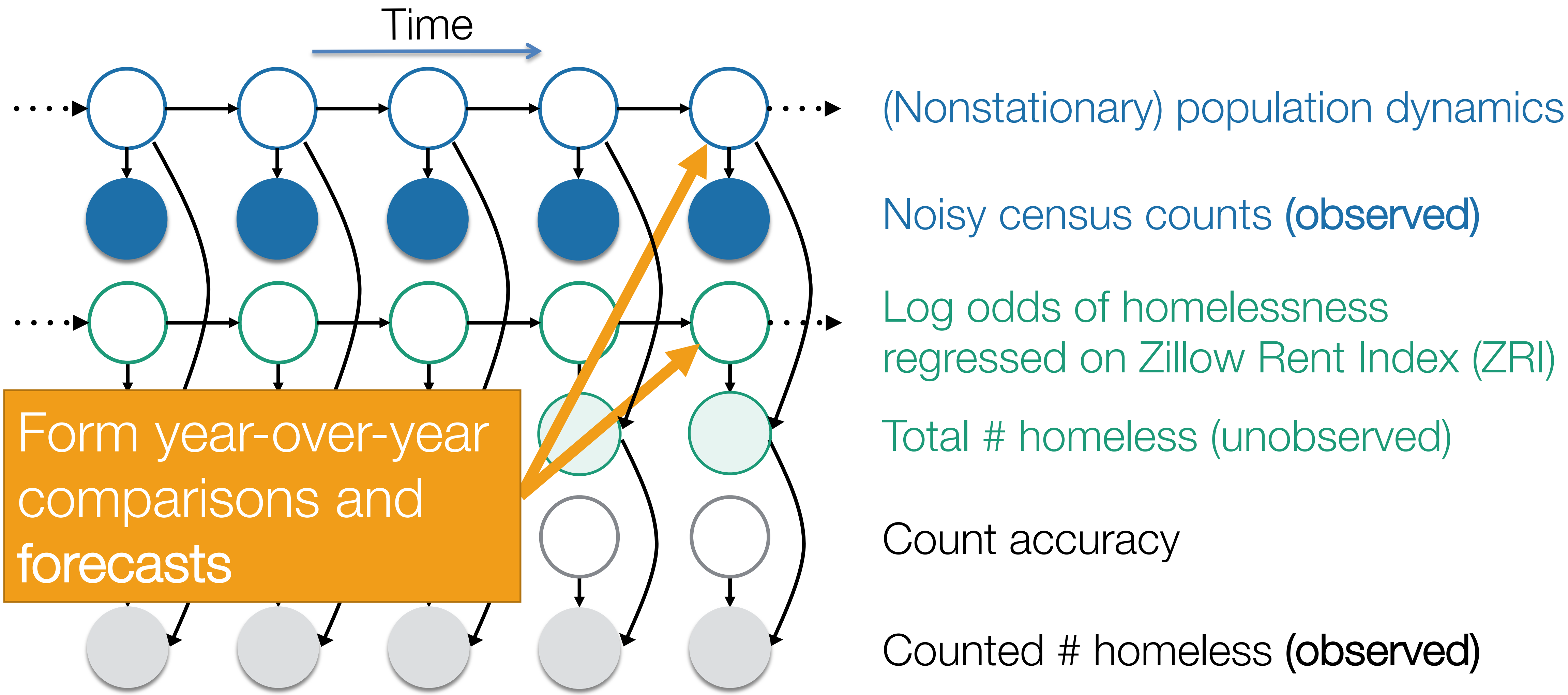
Glynn and Fox, Annals of Applied Statistics 2019

Benefits over past approaches…



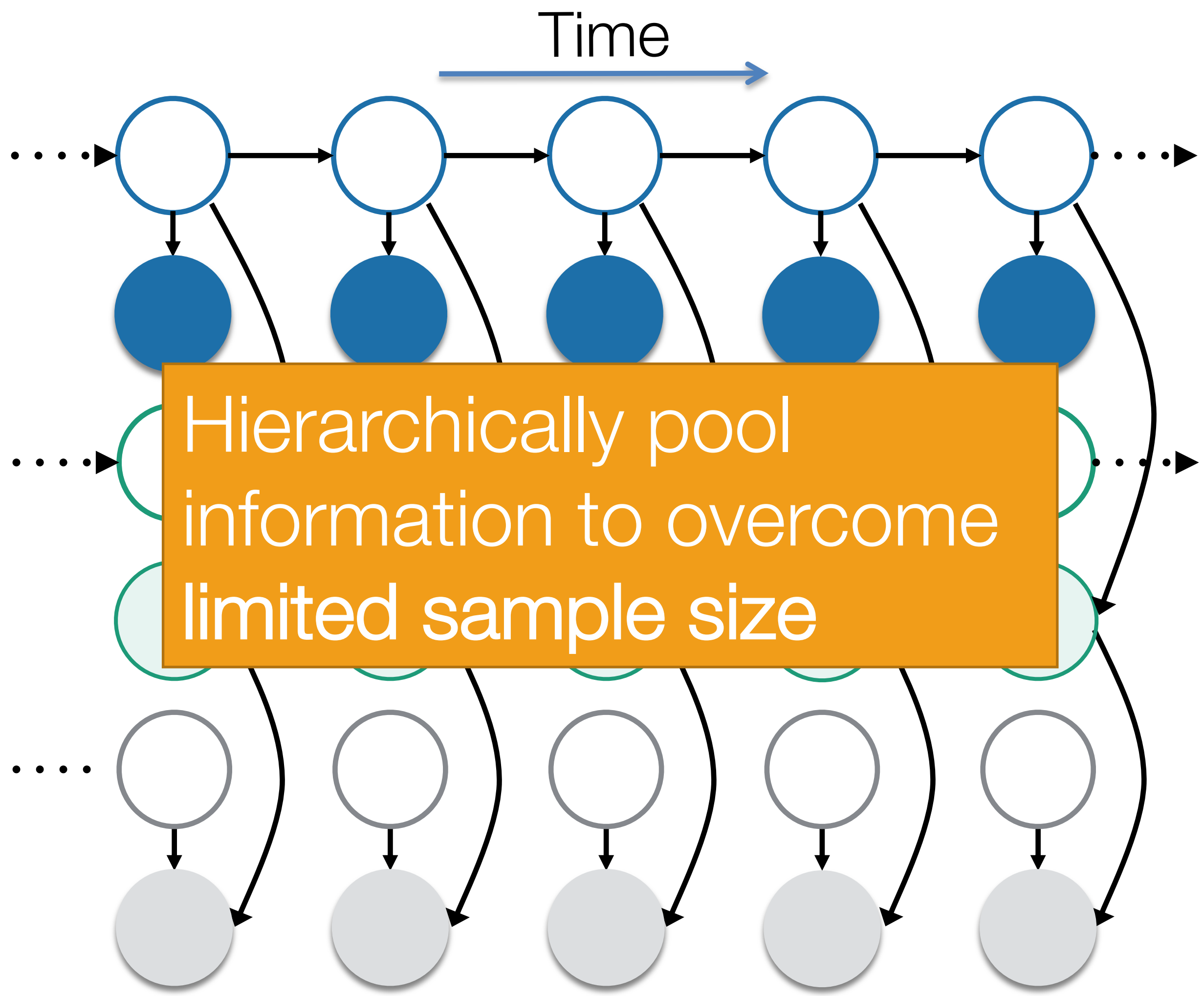
Glynn and Fox, Annals of Applied Statistics 2019

Benefits over past approaches…



Glynn and Fox, Annals of Applied Statistics 2019

Benefits over past approaches…



(Nonstationary) population dynamics

Noisy census counts **(observed)**

Log odds of homelessness regressed on Zillow Rent Index (ZRI)

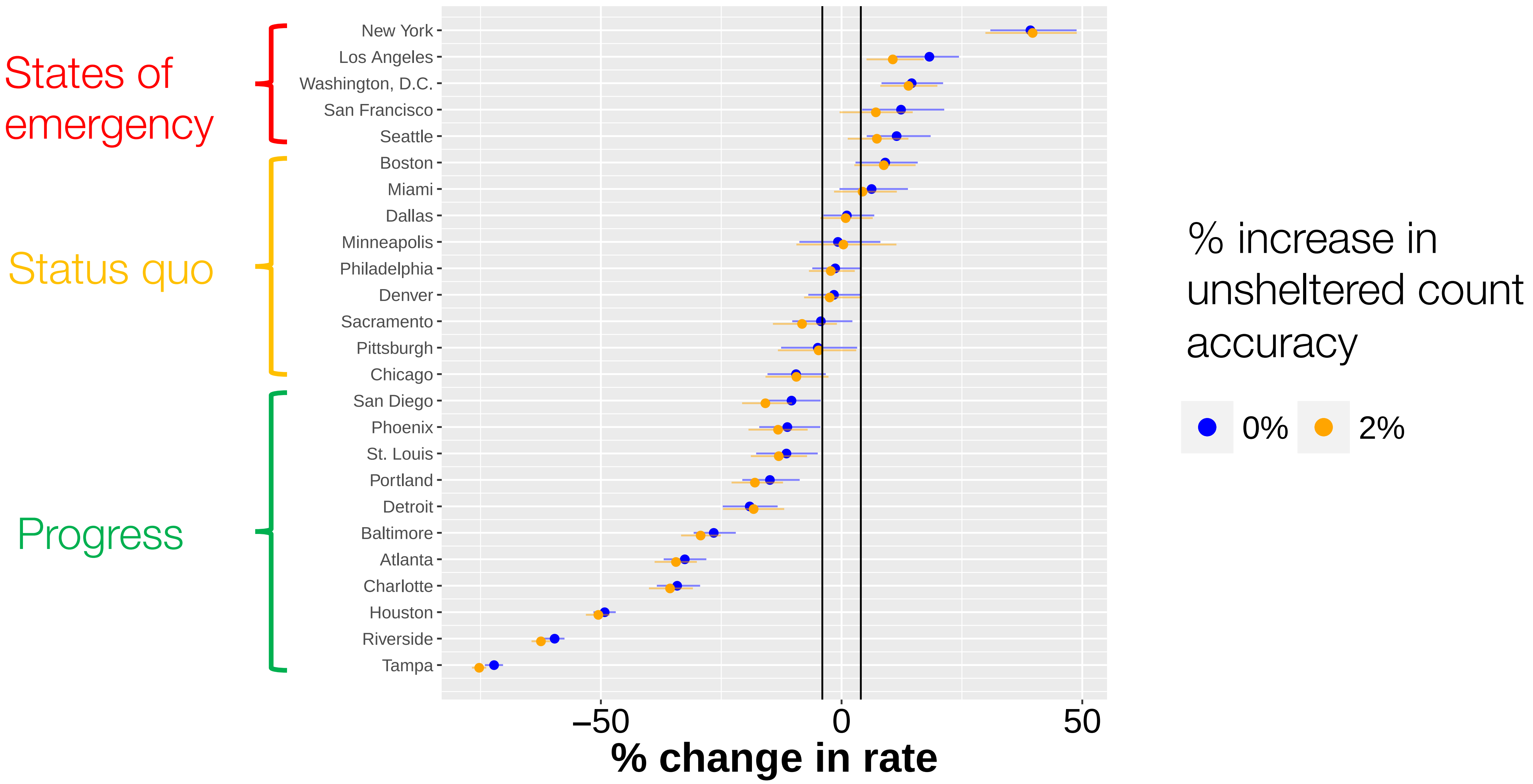
Total # homeless (unobserved)

Count accuracy

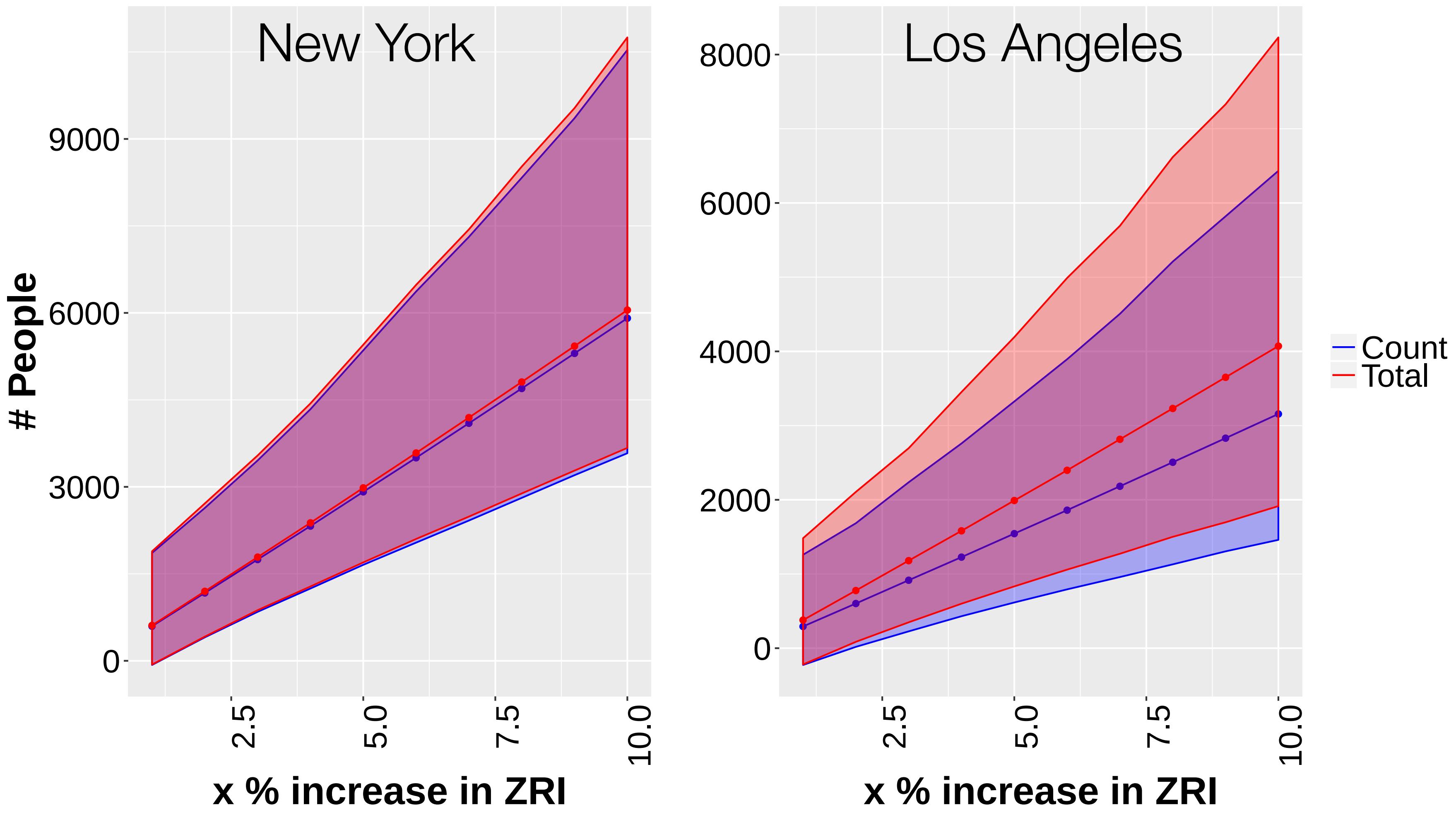
Counted # homeless **(observed)**

Glynn and Fox, Annals of Applied Statistics 2019

Adjusting for dynamics of count accuracy and total population, is homelessness rate increasing?



If rent increases x%, do # homeless increase?



Typically **weak relationship** + **wide uncertainty** intervals

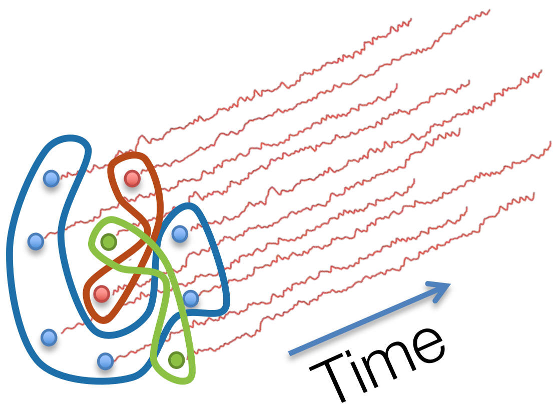
Past methods overly **confident**…ignore noise in homeless count and census data

Interpretable interactions

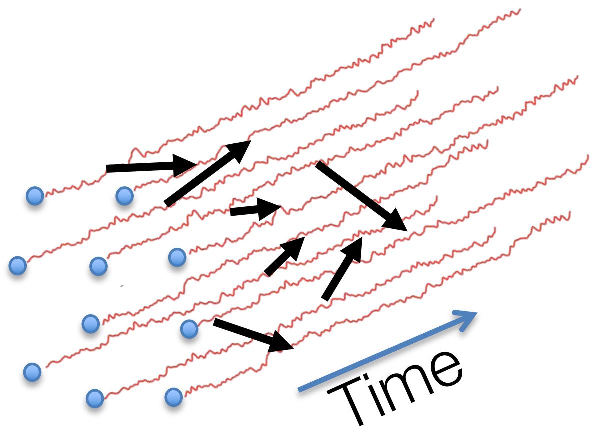
Modeling sparsely sampled, nonstationary time series

Handling bias in stochastic gradients of sequential data

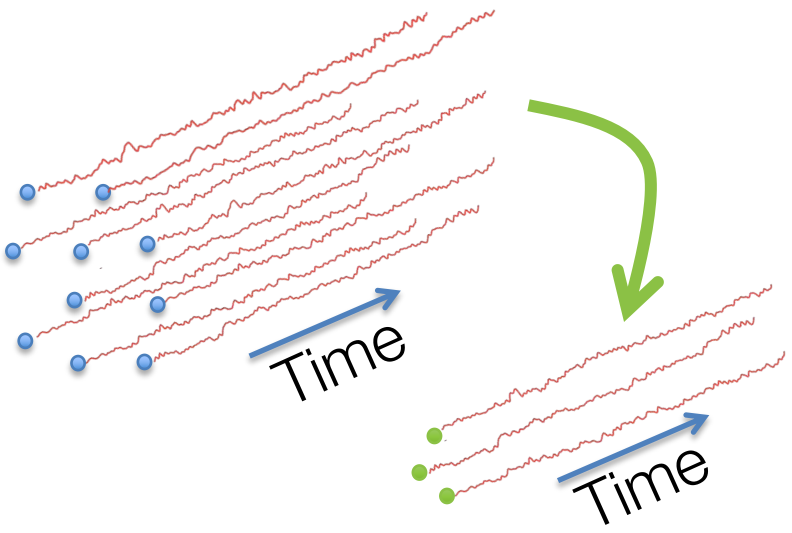
Recap: Mechanisms for coping with limited data



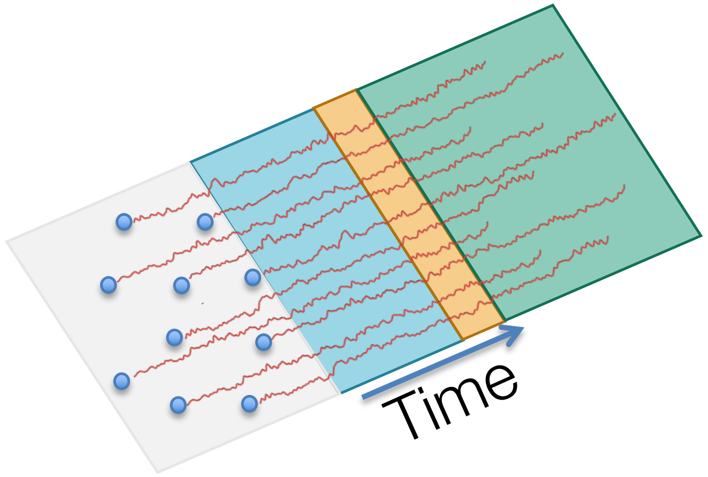
clusters and hierarchies



sparse directed interactions



low-dimensional embeddings

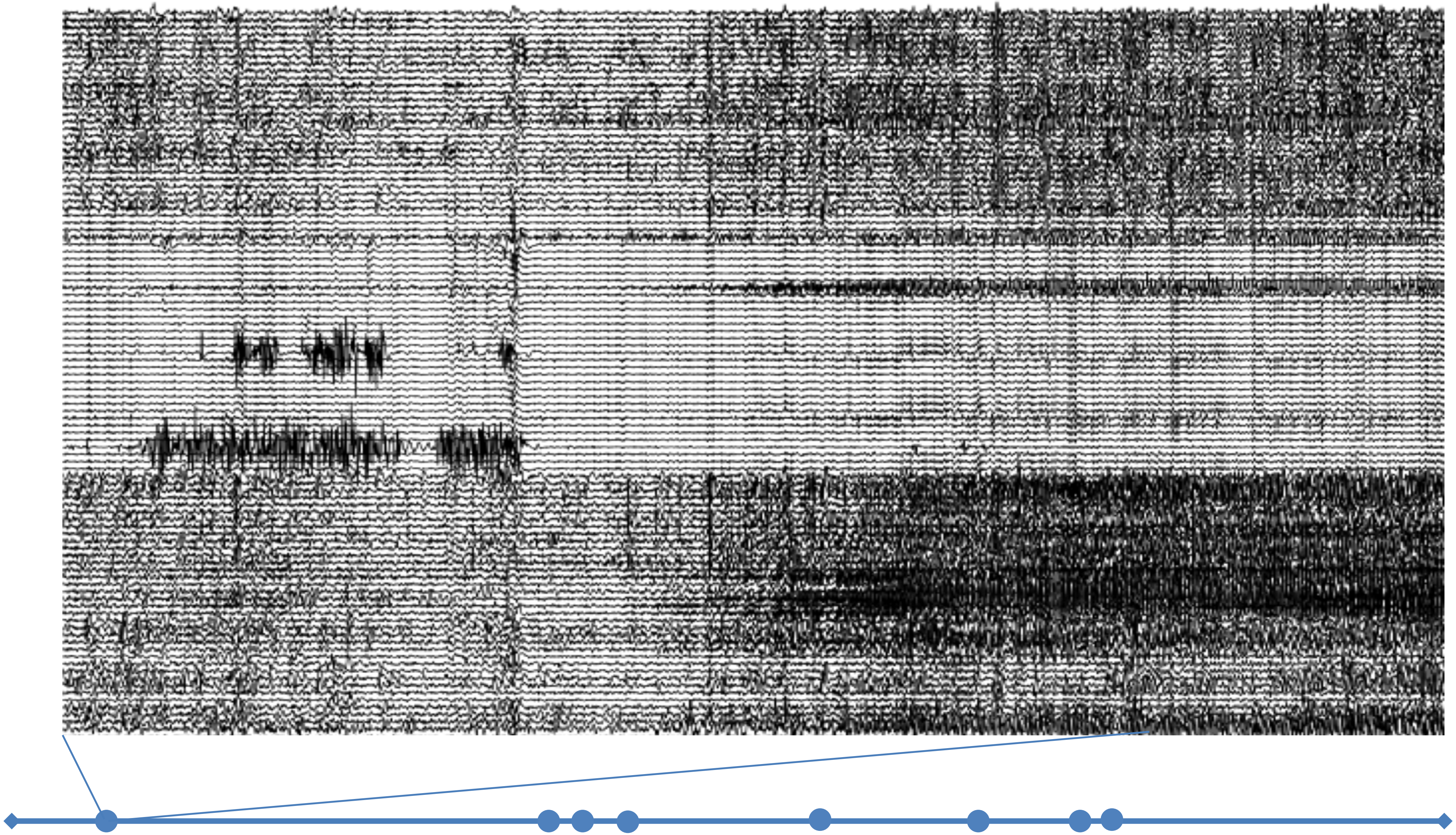


switching between simpler dynamics

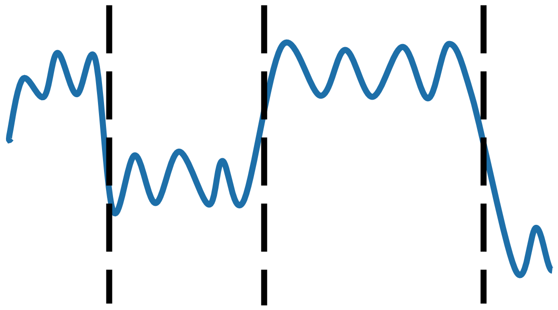
Interpretable interactions

Modeling sparsely sampled, nonstationary time series

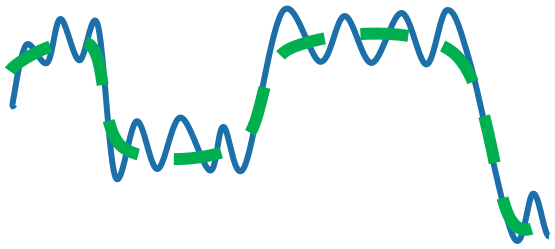
Handling bias in stochastic gradients of sequential data



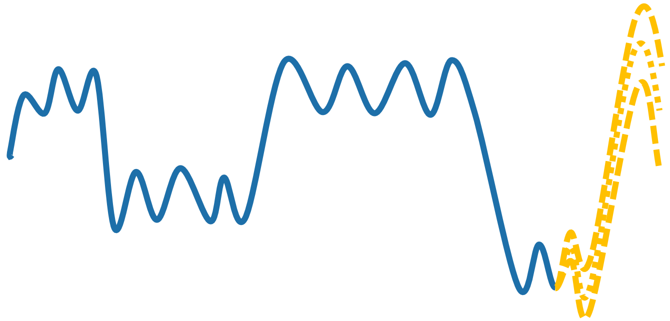
Discrete-time state space models



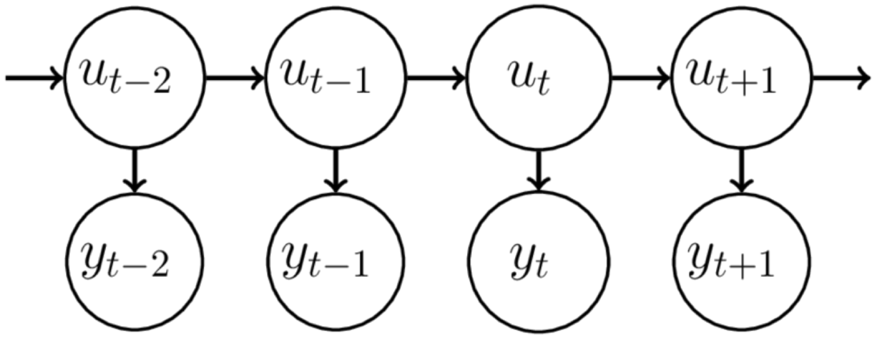
Segmentation



Smoothing/
Filtering



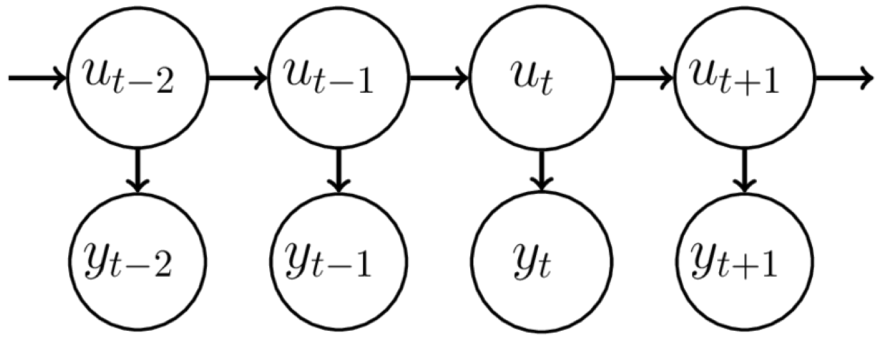
Forecasting



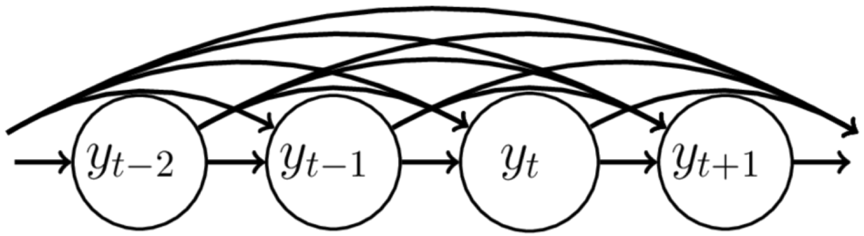
Examples: HMMs, AR-HMMs, linear Gaussian state space models, switching linear dynamical systems, nonlinear state space models, …

Learning challenge for SSMs

$$\log \Pr(y, u \mid \theta) = \sum_t \underbrace{\log \Pr(y_t \mid u_t, \theta)}_{\text{Emissions}} + \underbrace{\log \Pr(u_t \mid u_{t-1}, \theta)}_{\text{Transitions}}$$



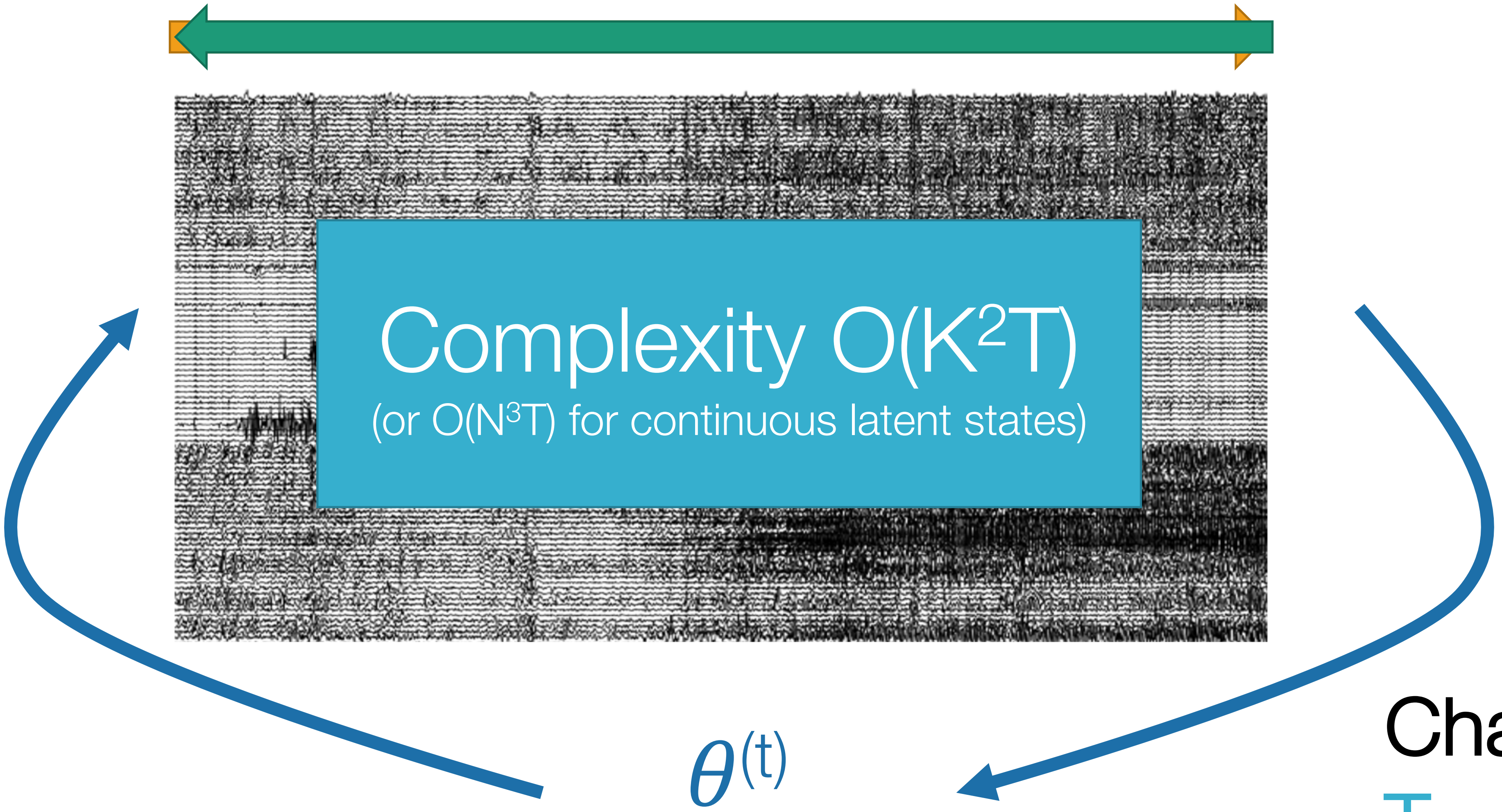
$$\log \Pr(y \mid \theta) = \sum_t \log \Pr(y_t \mid y_{<t}, \theta)$$



Fisher's Identity:

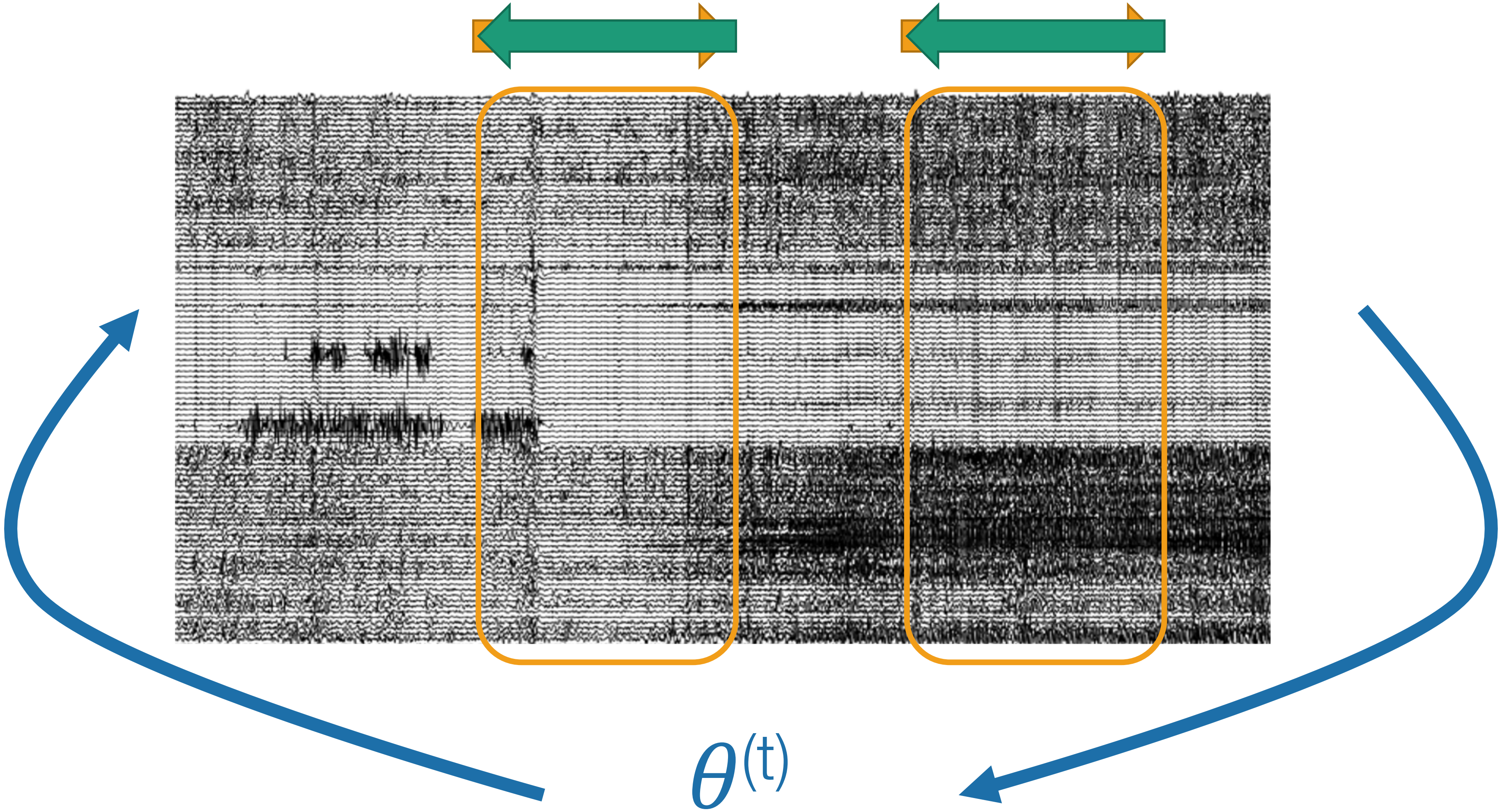
$$\nabla_\theta \log \Pr(y \mid \theta) = \underbrace{\mathbb{E}_{u \mid y, \theta}}_{\text{Expectation conditioned on full sequence}} \left[\nabla_\theta \log \Pr(y, u \mid \theta)\right]$$

Algorithms for SSMs



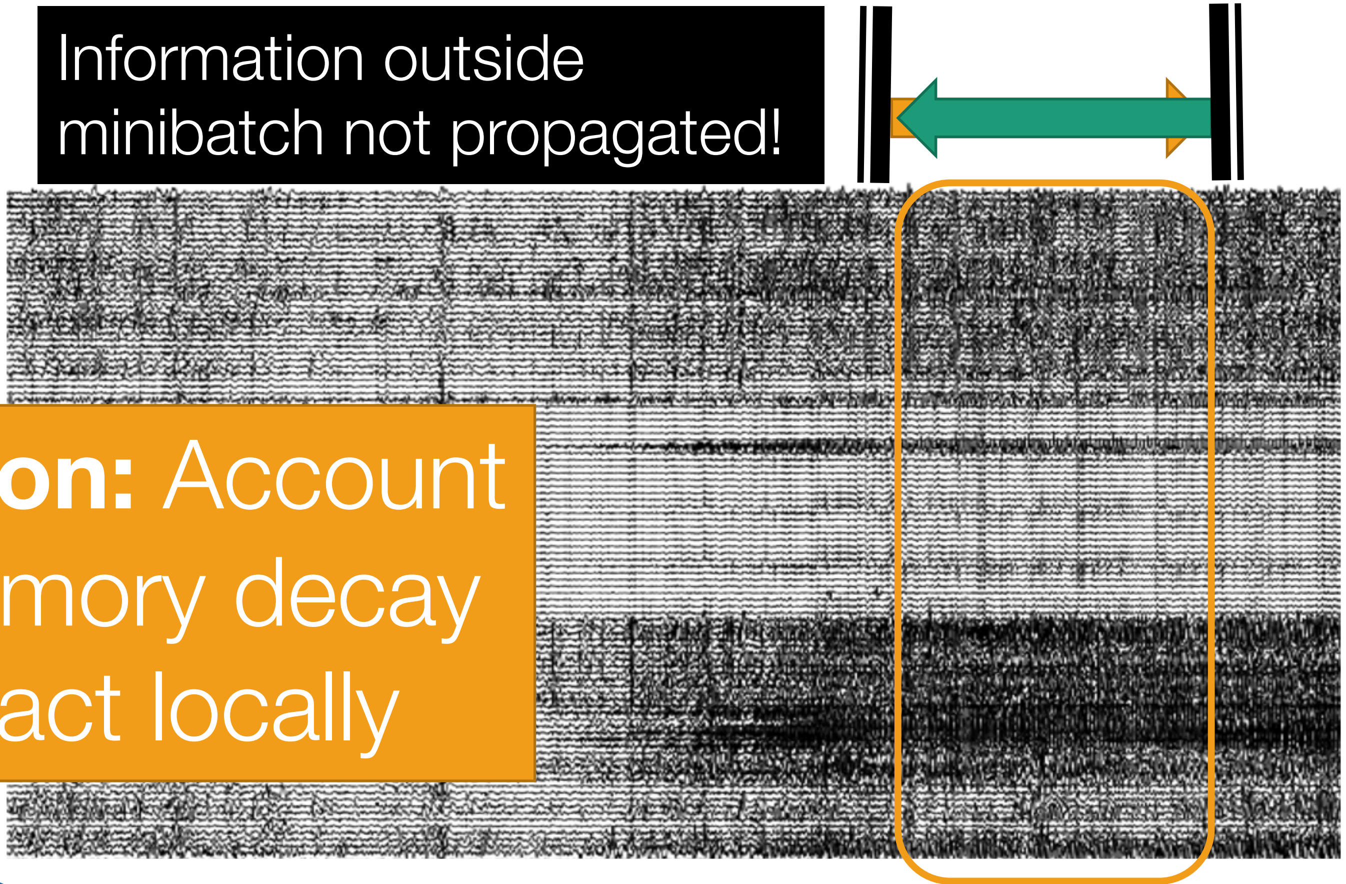
Challenge:
T = millions

Stochastic gradients + SSMs



Issue with naïve approach…

Information outside minibatch not propagated!



Solution: Account for memory decay to still act locally

$\theta^{(t)}$

A naïve stochastic gradient for SSMs

Fisher's Identity:

$$\nabla_\theta \log \Pr(y \mid \theta) = \mathbb{E}_{u|y,\theta}\left[\nabla_\theta \log \Pr(y, u \mid \theta)\right]$$

$$= \boxed{\sum_{t=1}^{T}} \underbrace{\mathbb{E}_{u|y,\theta}}_{\text{Expectation conditioned on full sequence}}\left[\nabla_\theta \log \Pr(y_t, u_t \mid u_{t-1}, \theta)\right]$$

Naive gradient estimator:

$$\nabla_\theta \widehat{\log \Pr}(y \mid \theta) = \Pr(\mathcal{S})^{-1} \cdot \boxed{\sum_{t \in \mathcal{S}}} \underbrace{\mathbb{E}_{u|y_{\mathcal{S}},\theta}}_{\text{Only take expectation conditioning on subsequence}}\left[\nabla_\theta \log \Pr(y_t, u_t \mid u_{t-1}, \theta)\right]$$

An unbiased, but impractical alternative

Fisher's Identity:

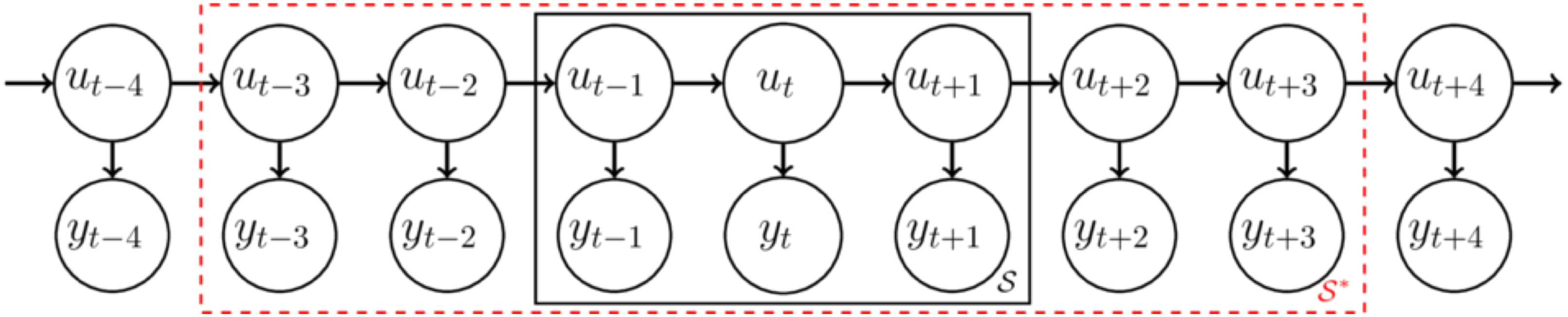
$$\nabla_\theta \log \Pr(y \,|\, \theta) = \mathbb{E}_{u|y,\theta}\left[\nabla_\theta \log \Pr(y, u \,|\, \theta)\right]$$

$$= \boxed{\sum_{t=1}^{T}} \mathbb{E}_{u|y,\theta}\left[\nabla_\theta \log \Pr(y_t, u_t \,|\, u_{t-1}, \theta)\right]$$

Unbiased gradient estimator:

$$\nabla_\theta \widehat{\log \Pr(y \,|\, \theta)} = \Pr(\mathcal{S})^{-1} \cdot \boxed{\sum_{t \in \mathcal{S}}} \underbrace{\mathbb{E}_{u|y,\theta}}_{\text{Requires message passing over full sequence O(|T|)}}\left[\nabla_\theta \log \Pr(y_t, u_t \,|\, u_{t-1}, \theta)\right]$$

Buffering for approximate unbiasedness



$$\mathcal{S} \subset \mathcal{S}^* \subset \mathcal{T}$$

"Buffered" gradient estimator:

$$\nabla_\theta \widetilde{\log \Pr}(y \mid \theta) = \Pr(\mathcal{S})^{-1} \cdot \sum_{t \in \mathcal{S}} \mathbb{E}_{\underbrace{u|y_{\mathcal{S}^*},\theta}_{\text{Computation } \mathcal{O}(|\mathcal{S}^*|) \text{ (and memory)}}} [\nabla_\theta \log \Pr(y_t, u_t \mid u_{t-1}, \theta)]$$

Aicher, Ma, Foti, Fox, *to appear in SIAM Journal on Mathematics of Data Science.*

Error analysis

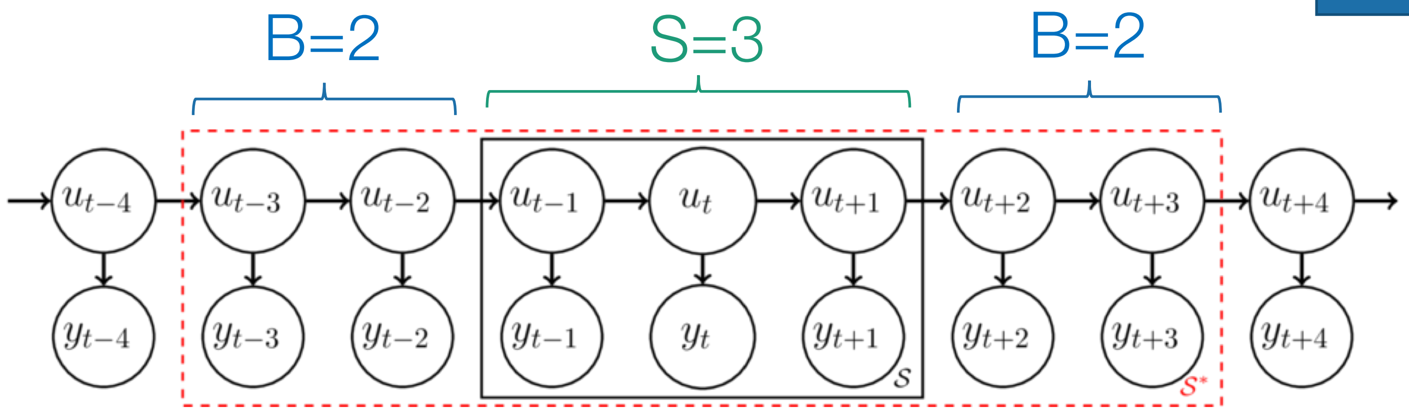
exact posterior $\gamma(u) = \Pr(u \,|\, y_{\mathcal{T}}, \theta)$
approx posterior $\tilde{\gamma}(u) = \Pr(u \,|\, y_{\mathcal{S}^*}, \theta)$

Theorem 1. *Let* $\epsilon_1 = \max\{\mathcal{W}_1(\gamma_{-B}, \tilde{\gamma}_{-B}), \mathcal{W}_1(\gamma_{S+B}, \tilde{\gamma}_{S+B})\}$. *If the gradient is Lipschitz in* u *with constant* L_U *and the forward and backward smoothing kernels are contractions with constant* $L < 1$, *then*

$$\left\|\mathbb{E}_{\gamma}\left[\nabla_\theta \log \Pr(y_{\mathcal{S}}, u_{\mathcal{S}}|\theta)\right] - \mathbb{E}_{\tilde{\gamma}}\left[\nabla_\theta \log \Pr(y_{\mathcal{S}}, u_{\mathcal{S}}|\theta)\right]\right\|_2 \leq$$

$$4L_U \cdot \frac{1-L^S}{1-L} \cdot \boxed{L^B} \cdot \epsilon_1$$

Geometrically in B



Aicher, Ma, Foti, Fox, *to appear in SIAM Journal on Mathematics of Data Science.*

LGSSM example:

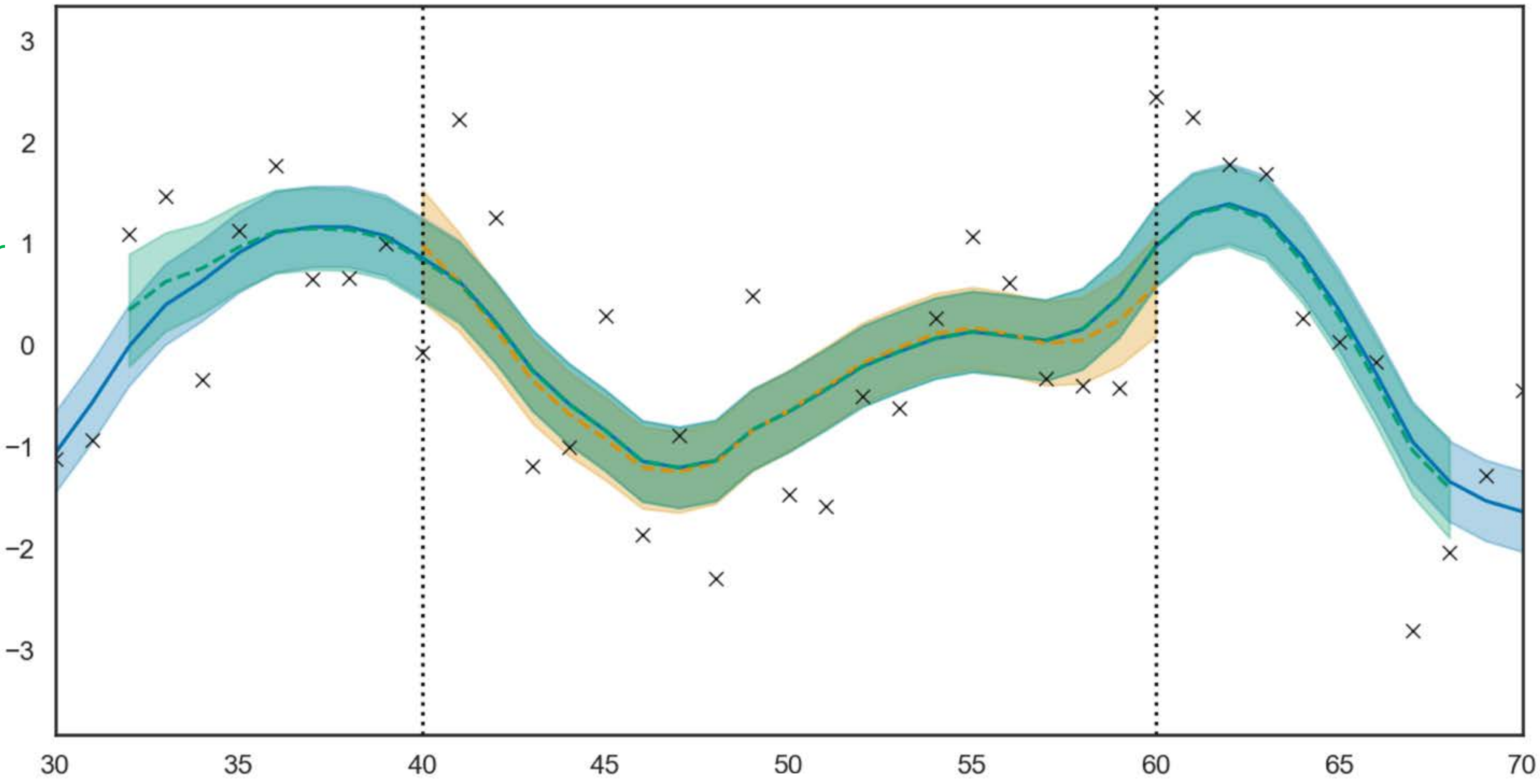
$$x_t = Ax_{t-1} + \mathcal{N}(0, Q)$$
$$y_t = x_t + \mathcal{N}(0, R)$$

$$\begin{cases} A = 0.9 \cdot \mathrm{Rot}(\pi/10) \\ Q = 0.1 \cdot \mathbb{I}_2 \\ R = \mathbb{I}_2 \end{cases}$$

Exact posterior

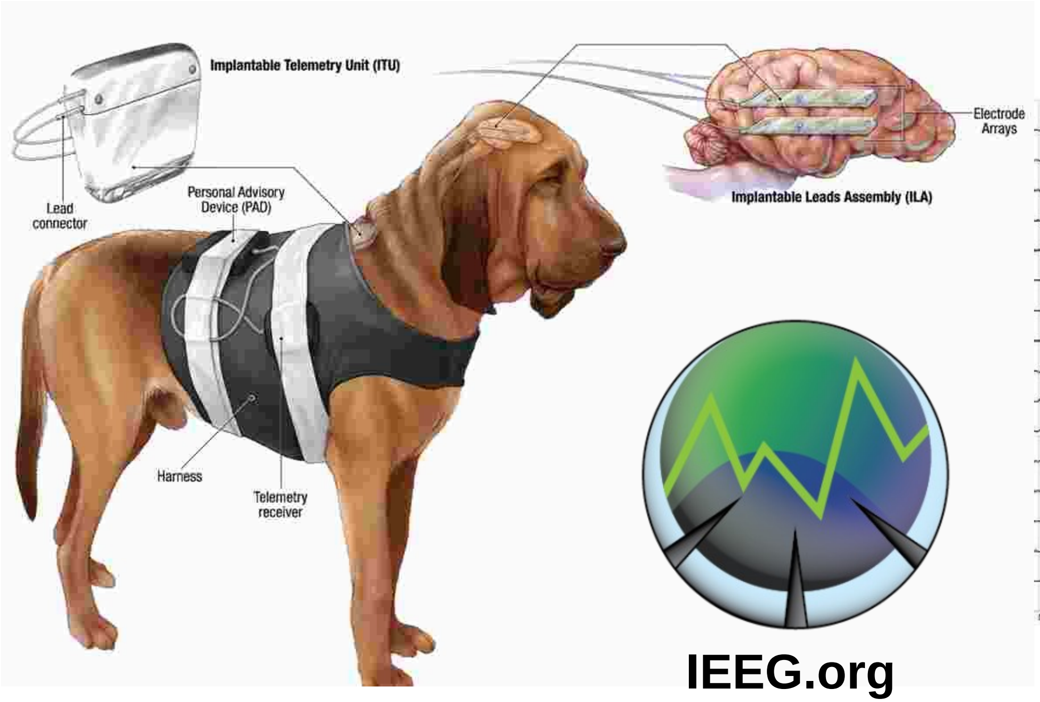
Naive posterior

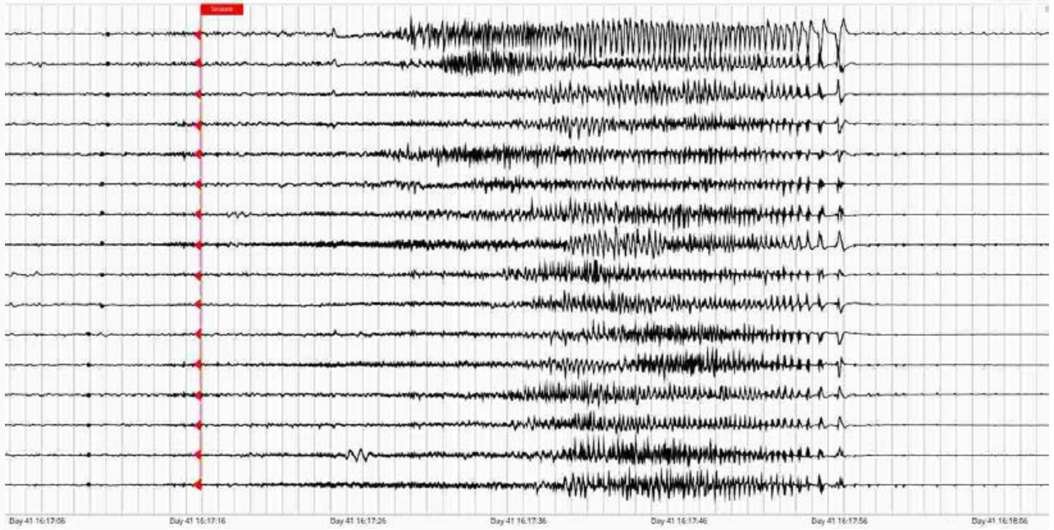
Buffer posterior



subsequence

Canine iEEG analysis

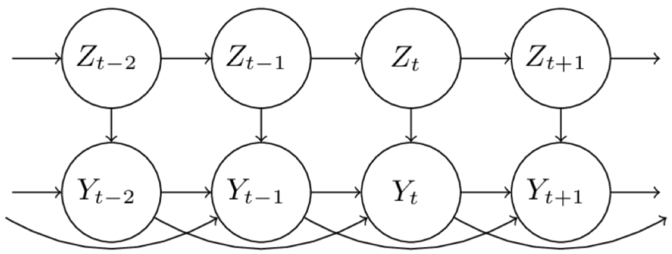


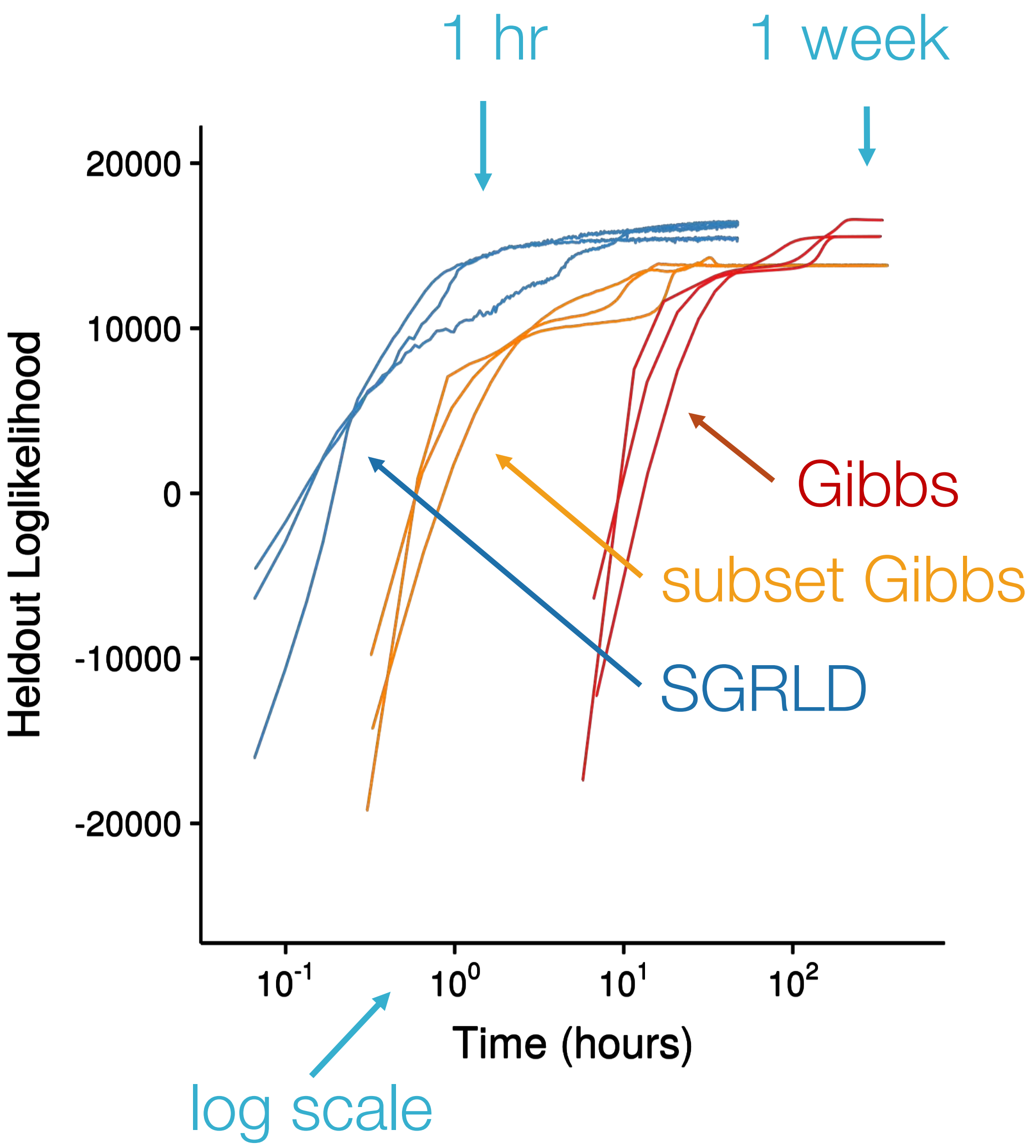


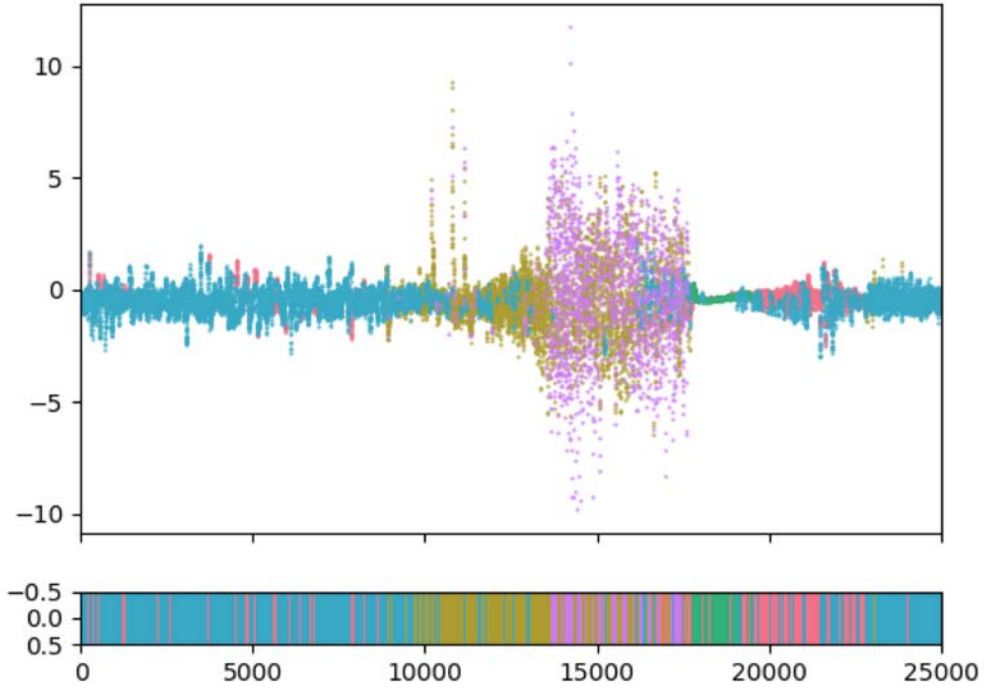
16 channels, 90 seizures

grab out 4 mins @ 200Hz per channel per seizure → 70 million time points

AR-HMM + MCMC

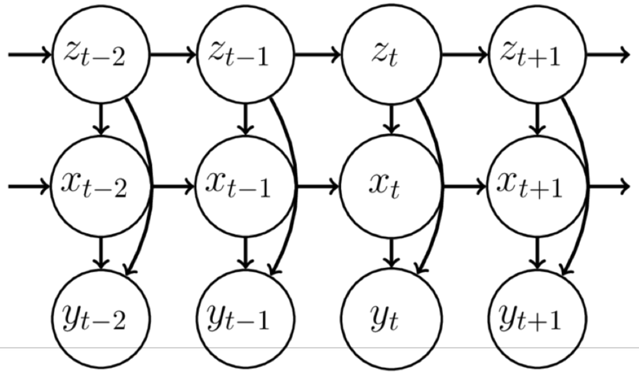


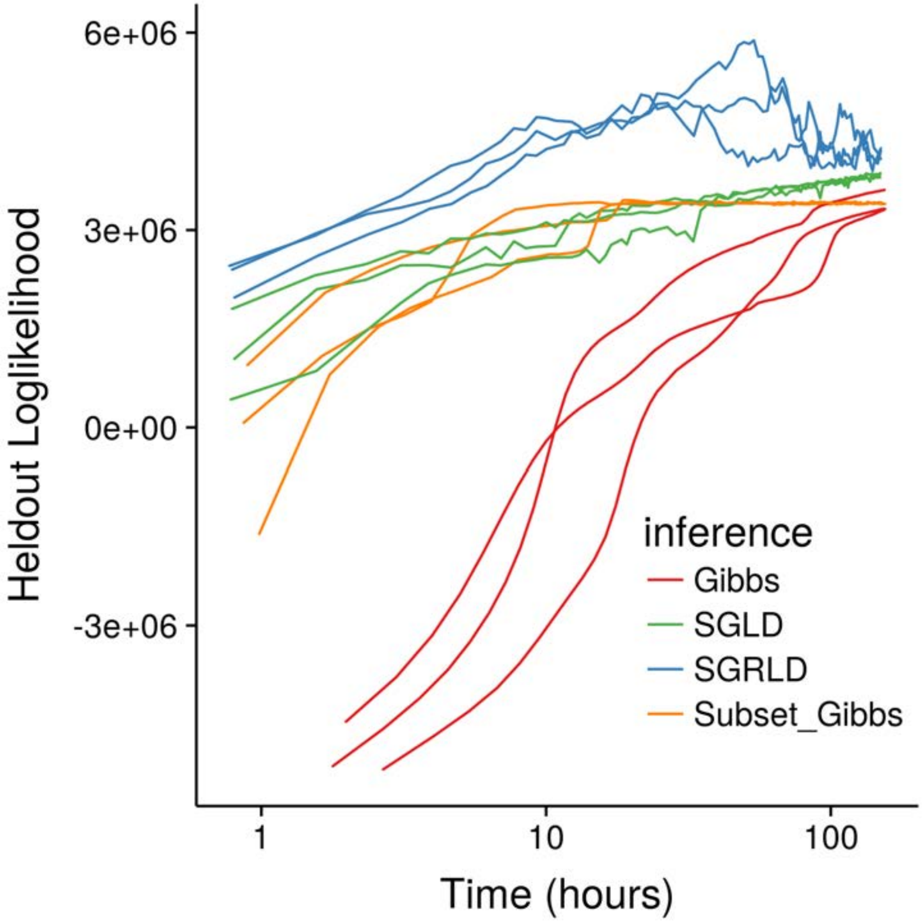


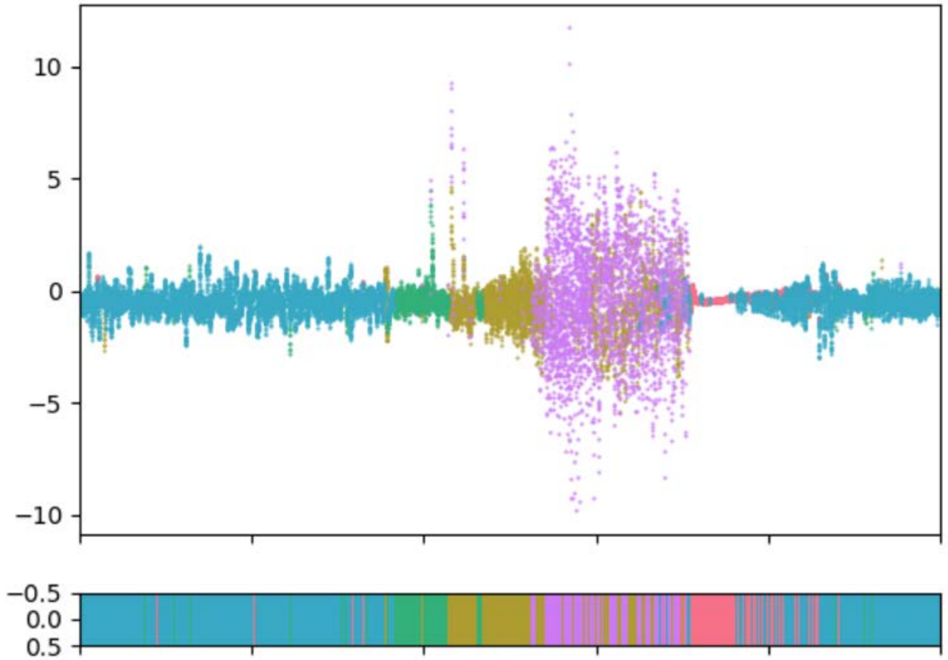


Example SGRLD segmentation
(zoomed in around a seizure)

SLDS + MCMC



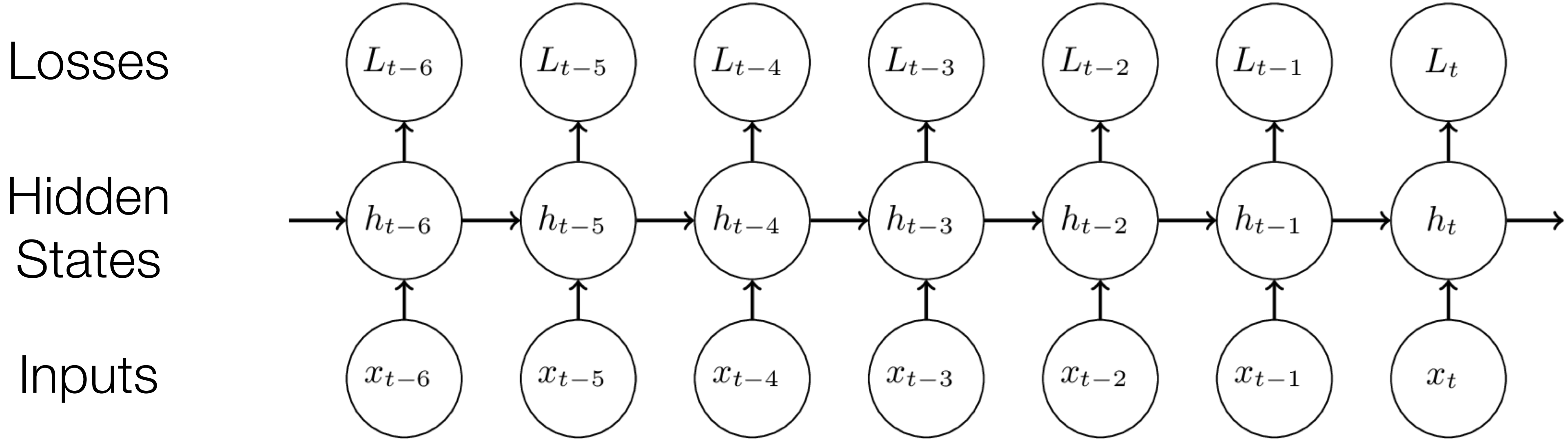




Example SGRLD segmentation
(zoomed in around a seizure)

Handling stochastic gradient bias in training RNNs

Goal: Low-bias training of RNNs

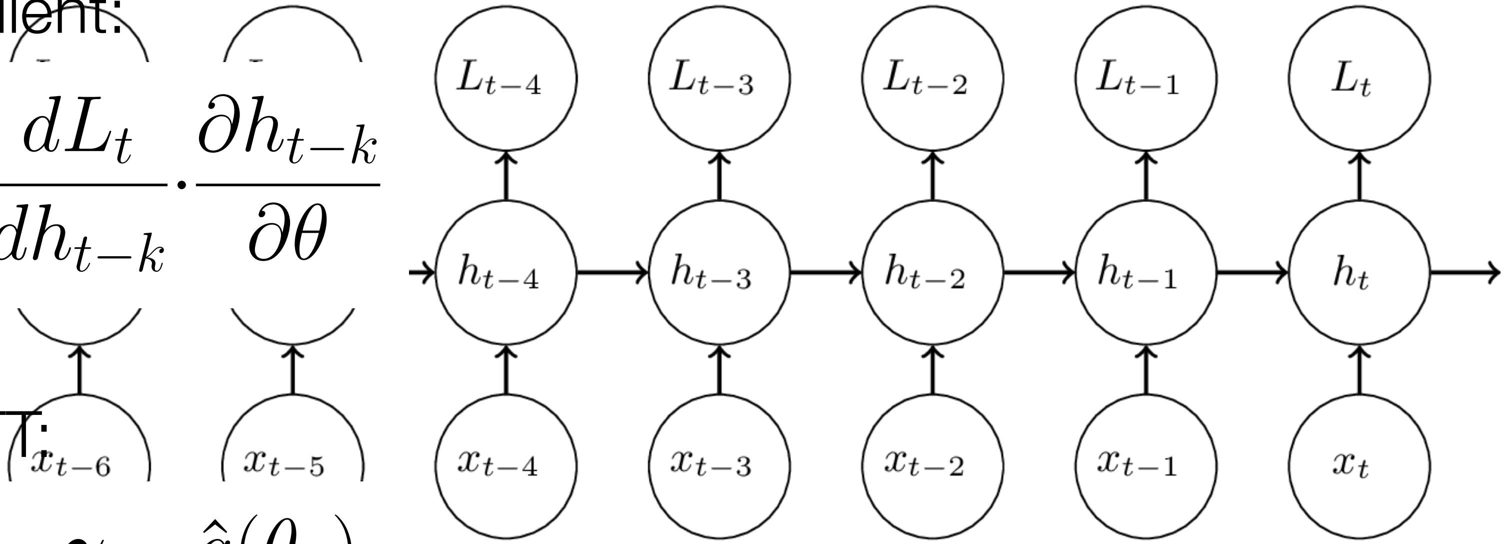


Unrolled recurrent neural network (RNN)

Backpropagation through time (BPTT)

Stochastic gradient:

$$\hat{g}(\theta) = \sum_{k=0}^{\infty} \frac{dL_t}{dh_{t-k}} \cdot \frac{\partial h_{t-k}}{\partial \theta}$$



SGD using BPTT:

$$\theta_{n+1} = \theta_n - \gamma_n \cdot \hat{g}(\theta_n)$$

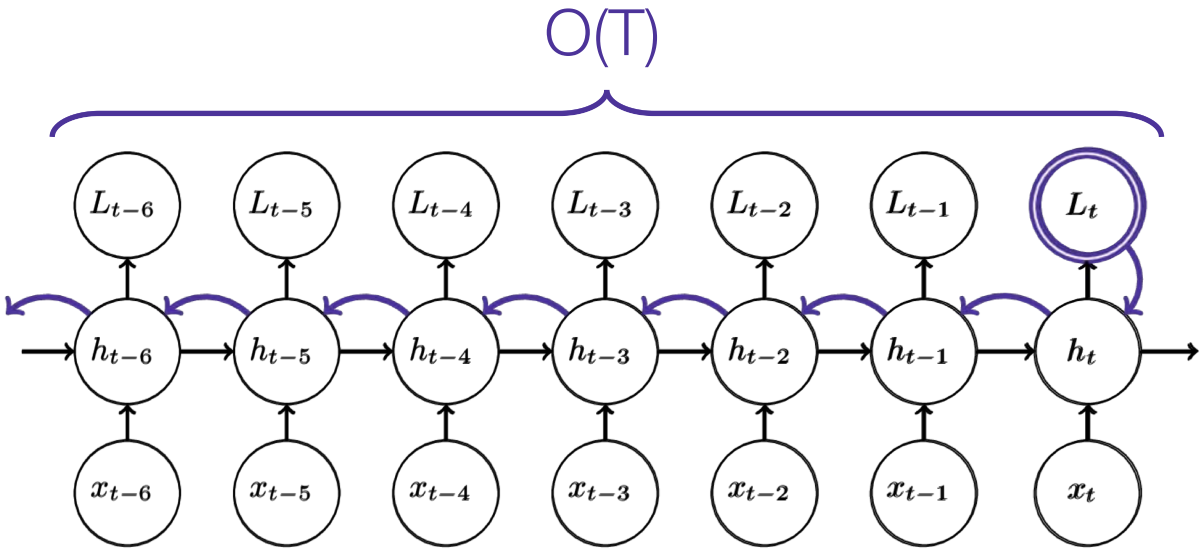
Backpropagation through time (BPTT)

Stochastic gradient:

$$\boxed{\hat{g}(\theta)} = \sum_{k=0}^{\infty} \frac{dL_t}{dh_{t-k}} \cdot \frac{\partial h_{t-k}}{\partial \theta}$$

SGD using BPTT:

$$\theta_{n+1} = \theta_n - \gamma_n \cdot \boxed{\hat{g}(\theta_n)}$$



O(T) computation time and memory

Expensive for long sequences

Truncated backpropagation through time (TBPTT)

Stochastic gradient:

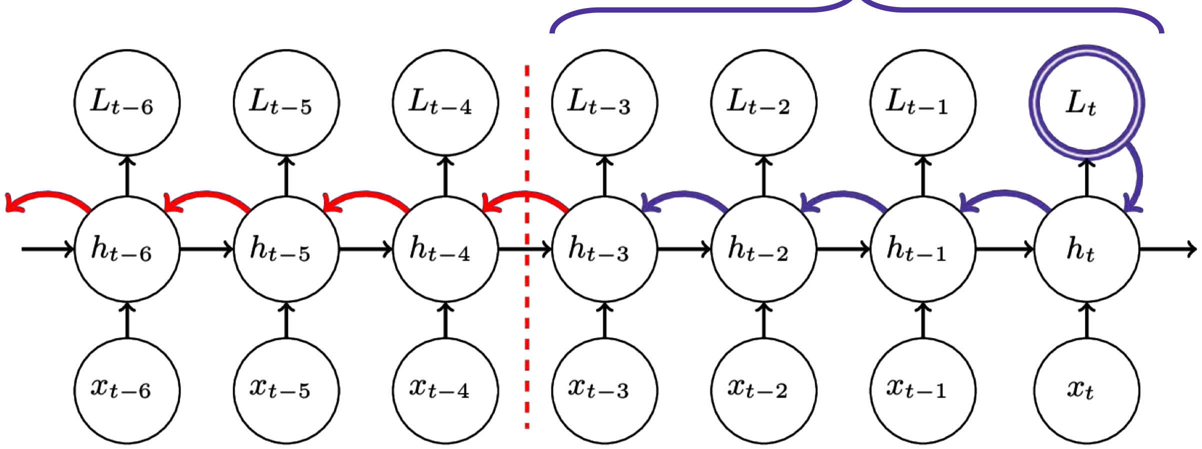
$$\boxed{\hat{g}_K(\theta)} = \sum_{k=0}^{K} \frac{dL_t}{dh_{t-k}} \cdot \frac{\partial h_{t-k}}{\partial \theta}$$

SGD using TBPTT:

$$\theta_{n+1} = \theta_n - \gamma_n \cdot \boxed{\hat{g}_K(\theta_n)}$$

Biased!

O(K)



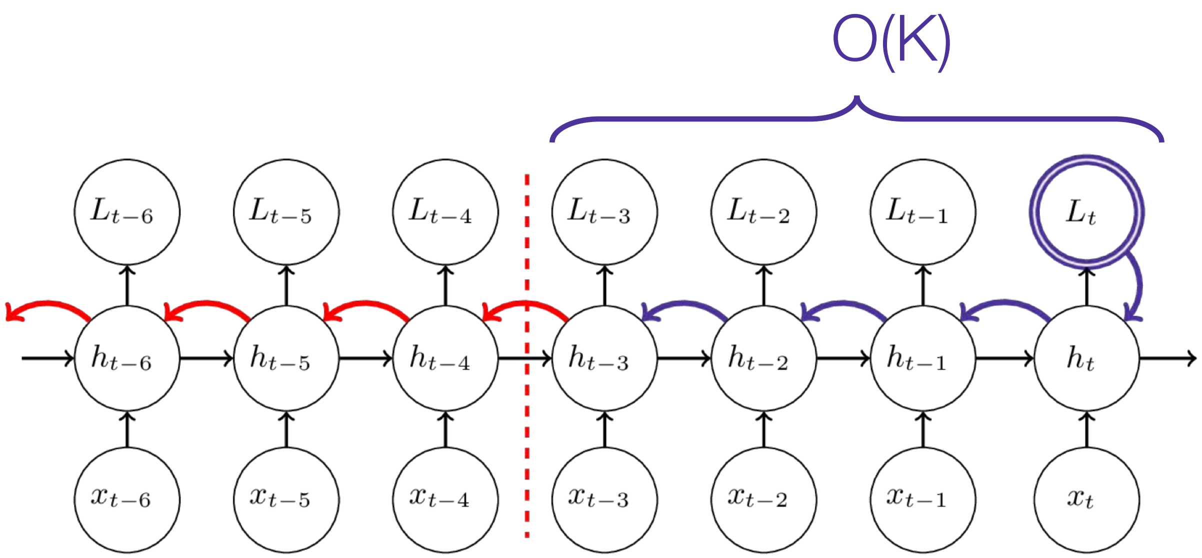
Truncate after K steps of BPTT

O(K) computation time and memory

What's the effect of this bias, and can we bound it?

How to choose K?

How does the bias affect learning?



Truncate after K steps of BPTT

Gradient decay assumptions

Stochastic gradient:

$$\hat{g}(\theta) = \sum_{k=0}^{\infty} \boxed{\frac{dL_t}{dh_{t-k}}} \cdot \frac{\partial h_{t-k}}{\partial \theta}$$

Chain Rule:

$$\boxed{\frac{\partial L_t}{\partial h_{t-k}}} = \frac{\partial L_t}{\partial h_t} \prod_{r=1}^{k} \boxed{\frac{\partial h_{t-r+1}}{\partial h_{t-r}}}$$

key term

Existing Work: Restrict RNN weights such that

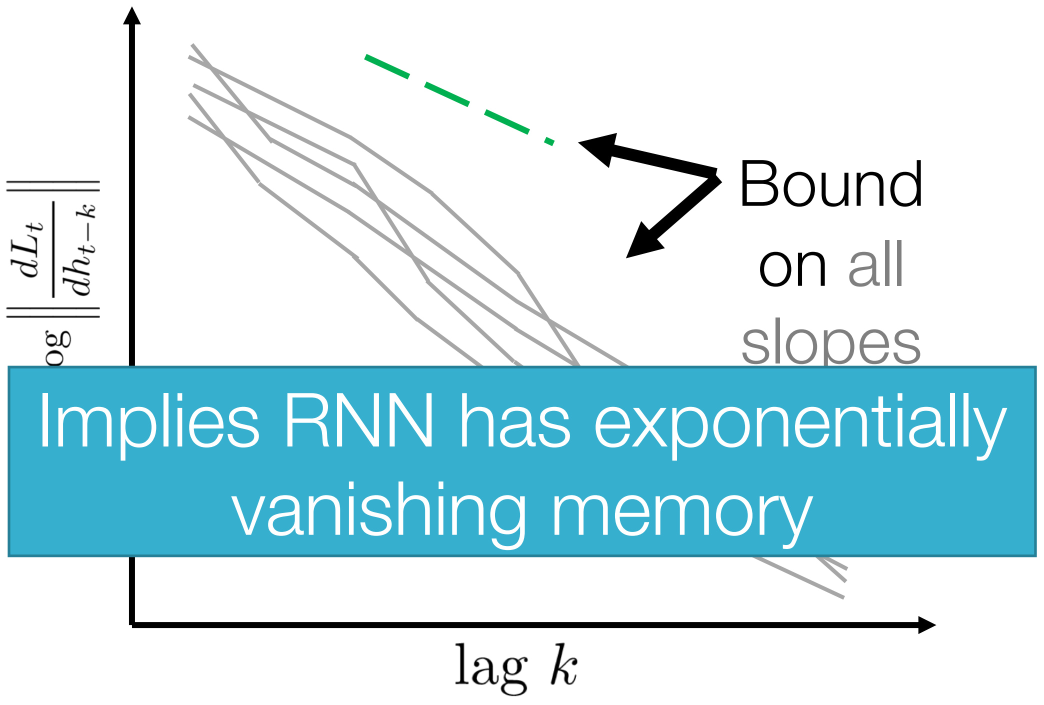
$$\left\| \frac{\partial h_{t-r+1}}{\partial h_{t-r}} \right\| \leq \lambda < 1 \longrightarrow \left\| \frac{dL_t}{dh_{t-k-1}} \right\| \leq \lambda \cdot \left\| \frac{dL_t}{dh_{t-k}} \right\|$$

"Stable" or "Chaos Free" RNN
[Laurent & von Brecht '16, Miller & Hardt '19]

Gradient decay assumptions

Previously:

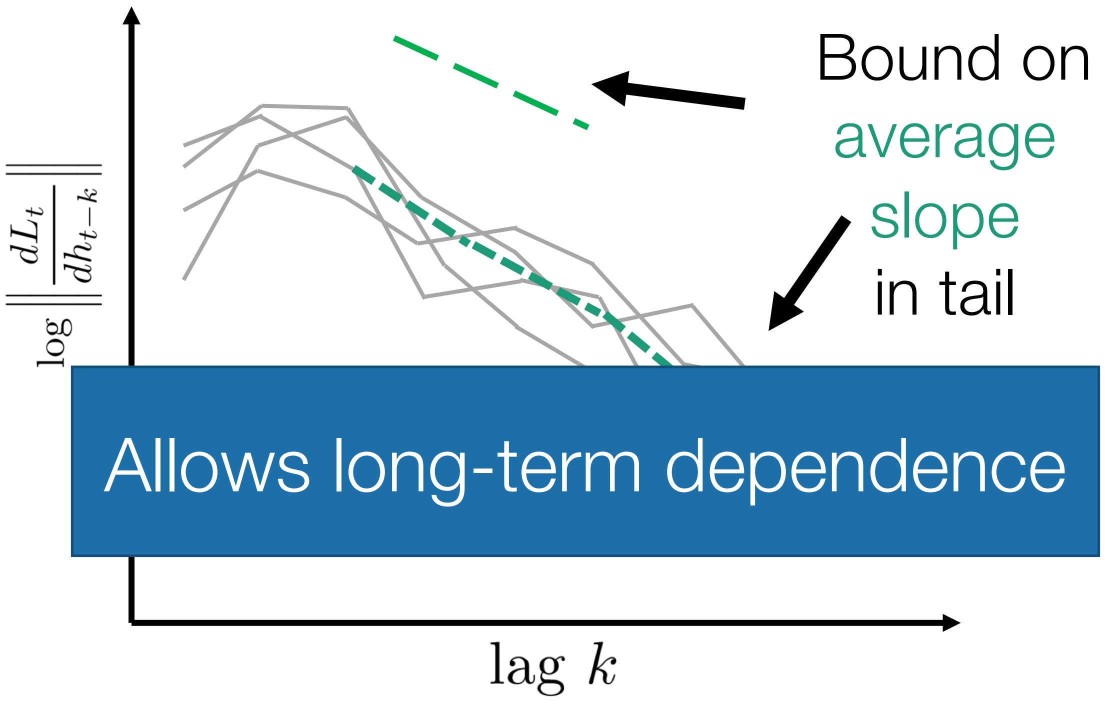
$$\left\| \frac{dL_t}{dh_{t-k-1}} \right\| \leq \lambda \cdot \left\| \frac{dL_t}{dh_{t-k}} \right\|$$



Implies uniform bound on $\hat{g} - \hat{g}_K$

Our relaxed assumption:

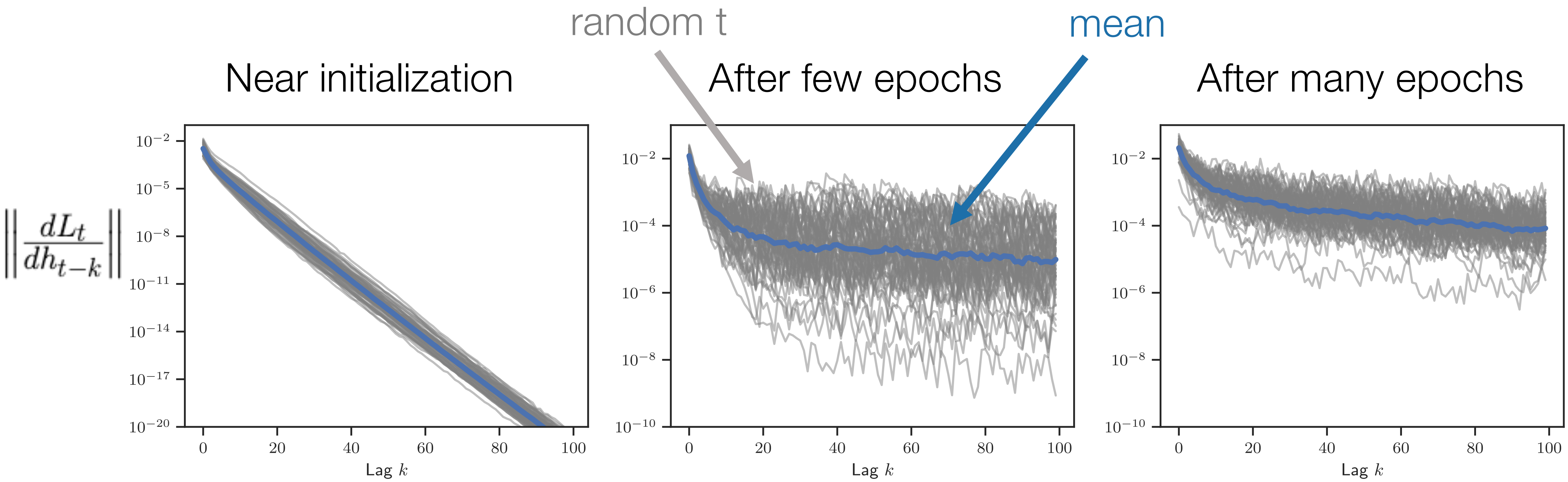
$$\mathbb{E}_t \left\| \frac{dL_t}{dh_{t-k-1}} \right\| \leq \beta \cdot \mathbb{E}_t \left\| \frac{dL_t}{dh_{t-k}} \right\| \quad \text{for all } k \geq \tau$$



Implies bound on gradient bias…will see

Example: LSTM on language modeling task

Penn Treebank dataset



Decay *on average*, but individual traces *do not*

Error analysis:
Bound on relative bias

RNN notation:

$$h_t = H(x_t, h_{t-1}; \theta)$$
$$y_t = F(h_t)$$

Assuming:

- Our gradient decay bound holds: $\mathbb{E}_t \left\| \frac{dL_t}{dh_{t-k-1}} \right\| \leq \beta \cdot \mathbb{E}_t \left\| \frac{dL_t}{dh_{t-k}} \right\| \text{ for all } k \geq \tau$
- $\partial H / \partial \theta$ is bounded

Then TBPTT has *bounded relative bias*:

$$\boxed{\delta} = \frac{\|\mathbb{E}[\hat{g}_K(\theta)] - g(\theta)\|}{\|g(\theta)\|} \leq \boxed{\mathcal{O}(\beta^{K-\tau})}$$

Relative bias

Geometric in K

Aicher, Foti, Fox, UAI 2019.

Error analysis:
Convergence rate of SGD with biased grads

Assuming:
- Relative bias at each step bounded by $\delta < 1$
- Loss is L-smooth and $\hat{g}$ has bounded variance

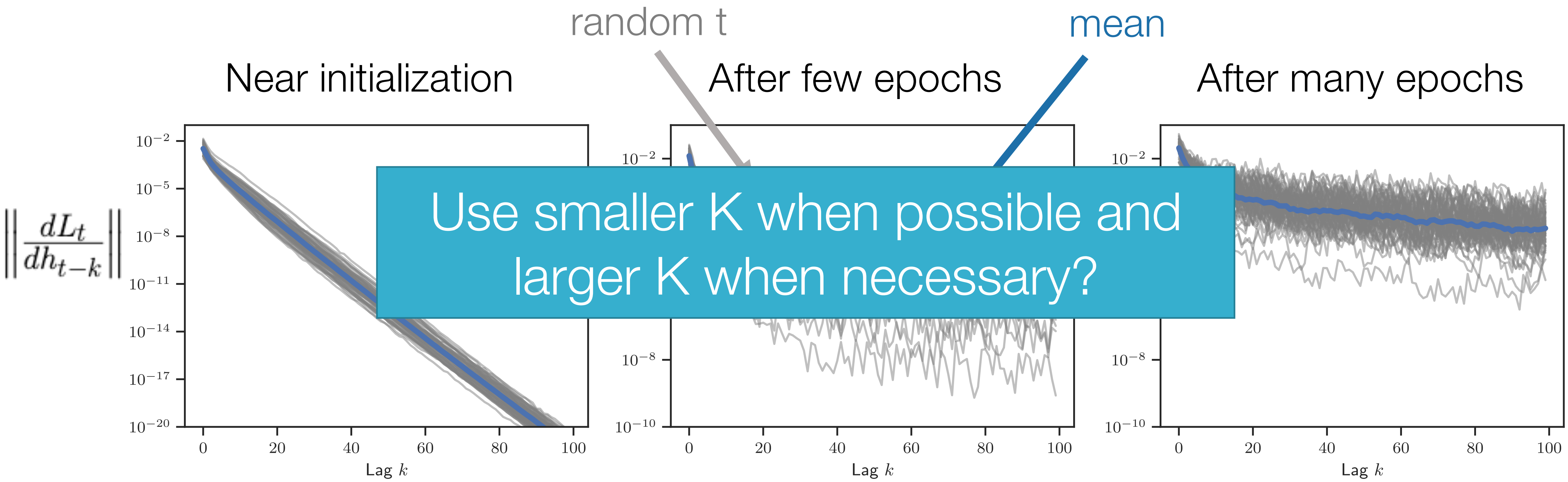
Then SGD with decaying stepsize $\gamma_n = \gamma \cdot n^{-1/2}$ *converges* at a rate:

$$\underbrace{\min_{n=1,\ldots,N} \|g(\theta_n)\|^2}_{\text{Convergence to stationary point}} = \mathcal{O}\left(\underbrace{(1-\delta)^{-1}}_{\text{Price of bias}} \cdot N^{-1/2} \log N\right)$$

Aicher, Foti, Fox, UAI 2019.

Example: LSTM on language modeling task

Penn Treebank dataset

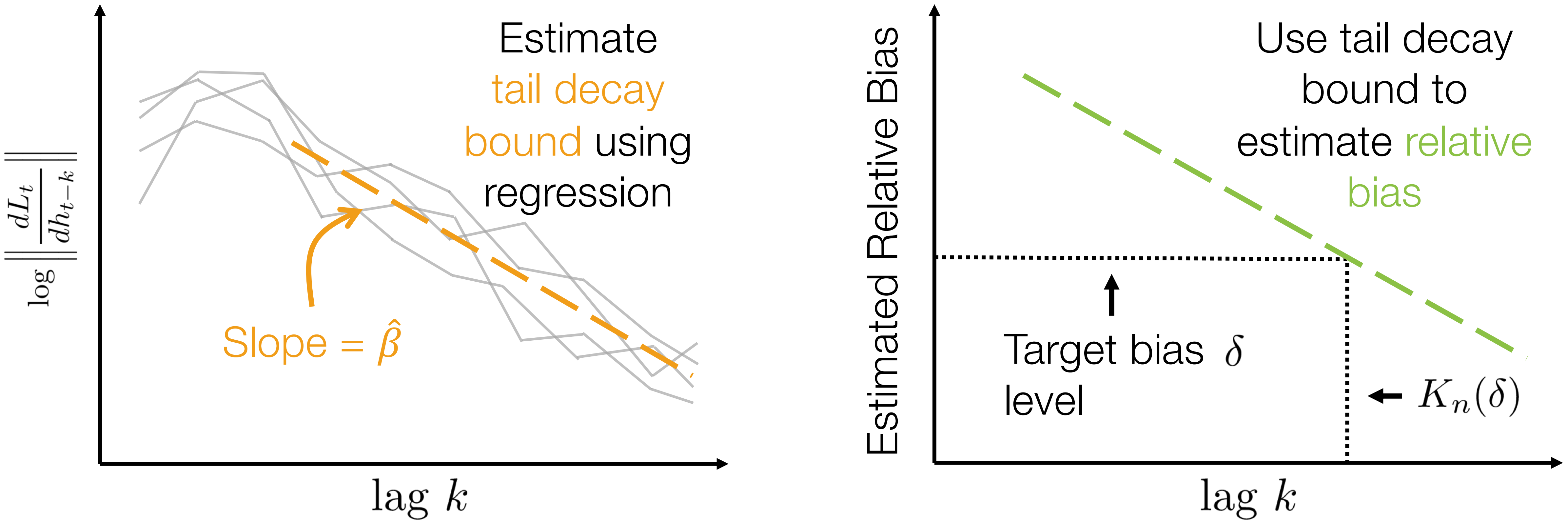


Use smaller K when possible and larger K when necessary?

For fixed K, relative bias increases during training *(in this example)*

Adaptive TBPTT algorithm

$$\mathbb{E}_t \left\| \frac{dL_t}{dh_{t-k-1}} \right\| \leq \beta \cdot \mathbb{E}_t \left\| \frac{dL_t}{dh_{t-k}} \right\| \text{ for all } k \geq \tau$$

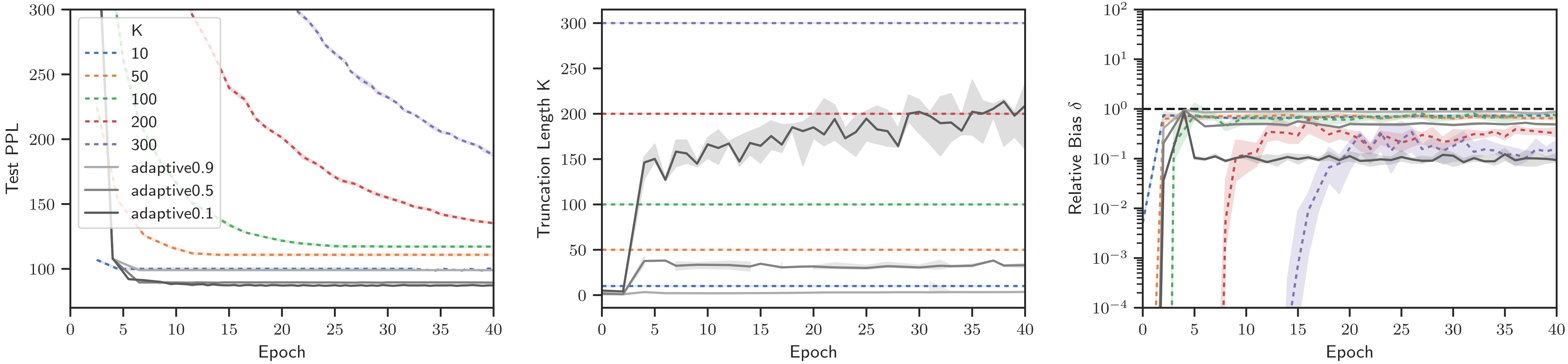


Aicher, Foti, Fox, UAI 2019.

TBPTT: Text Example

```
Data: “… no it was n't black monday 2
but while the new york stock exchange
did n't fall apart friday as the dow
jones industrial average plunged ...”
```

Penn Treebank, 1-Layer LSTM*



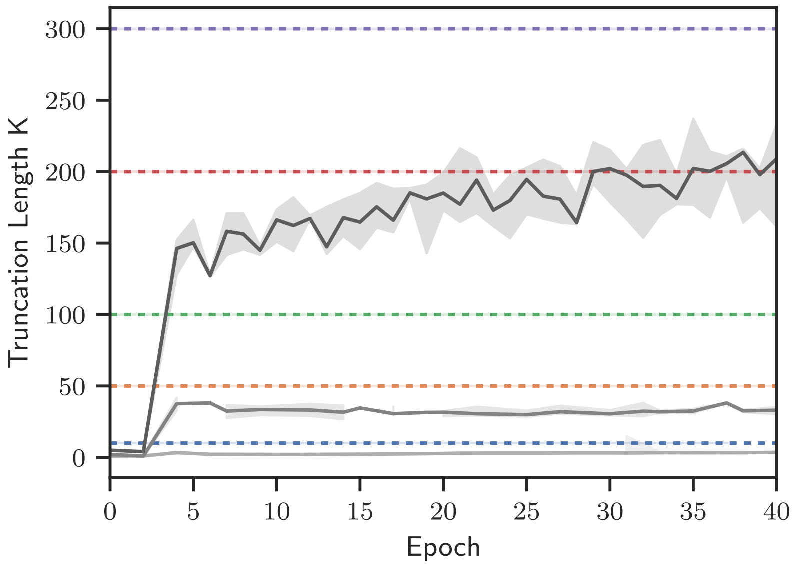
Perplexity vs. K – comparison

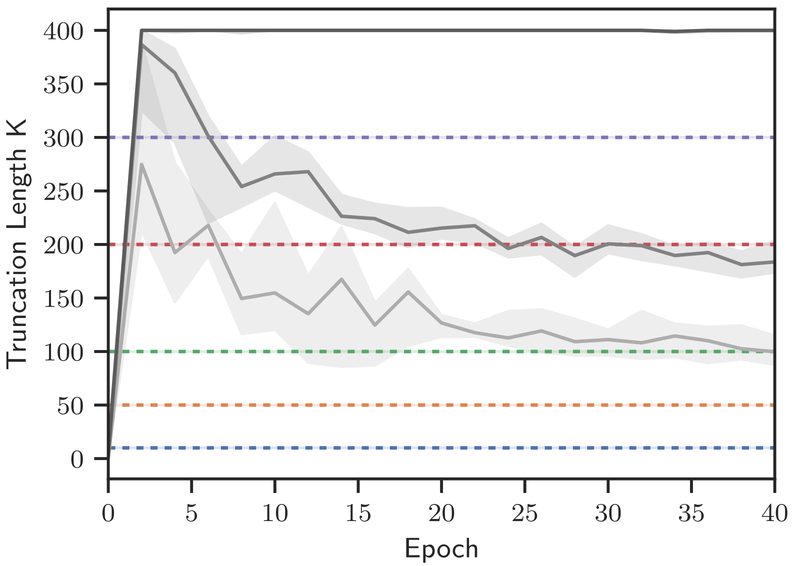
Penn treebank

K	Valid PPL	Test PPL
10	99.7 (0.6)	99.9 (0.8)
50	110.4 (0.4)	110.8 (0.8)
100	116.2 (0.5)	116.9 (0.5)
200	125.2 (1.2)	126.1 (0.9)
300	161.5 (0.5)	161.2 (0.3)
$\delta = 0.9$	100.1 (0.5)	99.0 (0.5)
$\delta = 0.5$	90.1 (0.4)	89.5 (0.3)
$\delta = 0.1$	**88.1 (0.2)**	**87.2 (0.2)**

Wikitext-2

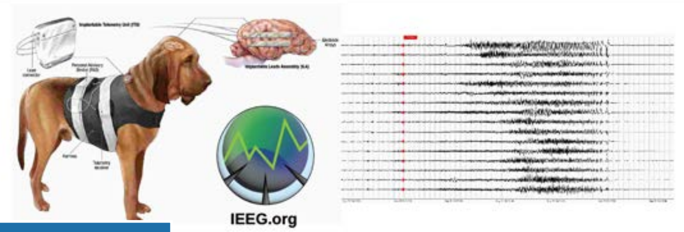
K	Valid PPL	Test PPL
10	144.2 (0.4)	136.5 (1.3)
50	133.4 (2.9)	127.2 (2.8)
100	134.4 (0.3)	127.8 (0.5)
200	130.3 (1.1)	124.6 (0.7)
300	**129.6 (1.4)**	**124.0 (2.2)**
$\delta = 0.9$	**130.0 (1.3)**	**124.1 (2.2)**
$\delta = 0.5$	**127.2 (0.7)**	**121.7 (0.6)**
$\delta = 0.1$	**127.5 (0.6)**	**121.9 (1.2)**





Summary

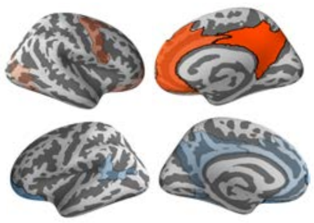
1. Deep learning offers **tremendous opportunities** for modeling complex dynamics, but problems much vaster than prediction + large corpora
2. Scaling learning is possible, but have to think carefully about broken dependencies (bias)

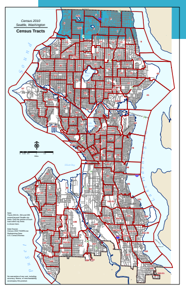


Interpretable interactions

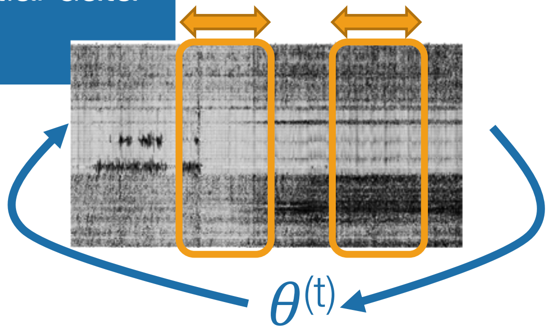
Modeling sparsely sampled, nonstationary time series

Handling bias in stochastic gradients of sequential data

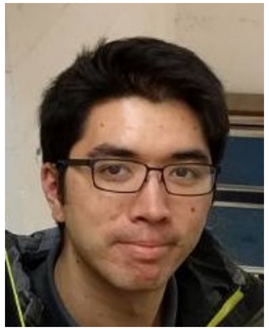




CHICAGO
2017
POINT-IN-TIME COUNT OF HOMELESS PERSONS



Credit for the hard work…


Chris Aicher
(Stat PhD)


Sam Ainsworth
(CSE PhD)

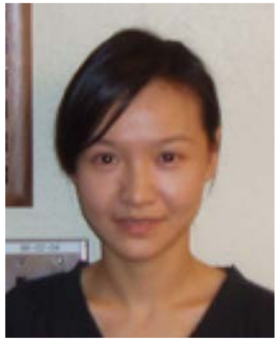

Ian Covert
(CSE PhD)


Nick Foti
(Research Scientist,
now at Apple)


Chris Glynn
(Postdoc,
Asst Prof at UNH)


Yian Ma
(AMath PhD,
postdoc at Berkeley)


Alex Tank
(Stat PhD,
now at Voleon)


Shirley You Ren
(Stat PhD,
now at Apple)