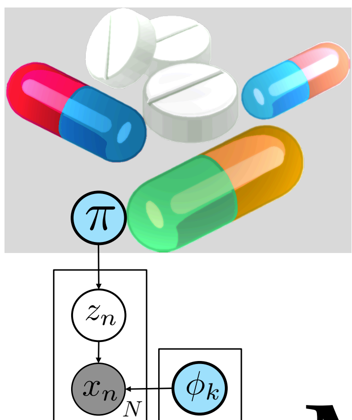
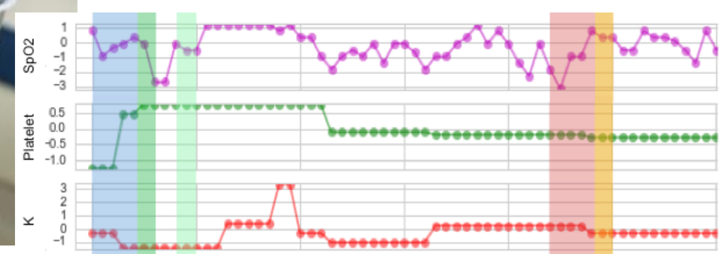
Overcoming Misspecification with Prediction Constrained Probabilistic Models





Mike Hughes

Assistant Professor of Computer Science, Tufts University

joint work with

Finale Doshi-Velez & Joe Futoma (Harvard)

Erik Sudderth & Gabe Hope (UC Irvine)

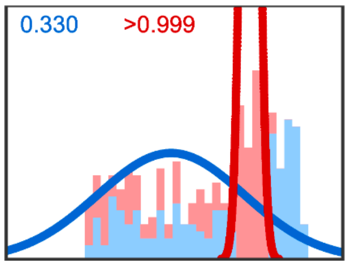
slides / papers / code

www.michaelchughes.com

Hughes Lab @ Tufts CS

Area: **statistical machine learning**; clinical informatics

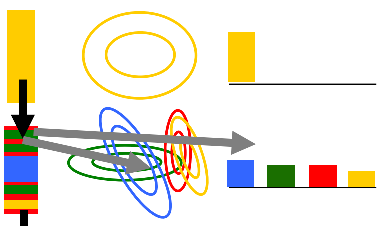
Lab goal: *Reliable training of interpretable models for real-world decisions*



New Training Goal: "Prediction-Constrained"

Avoids Model Misspecification for Decision Task via (Rare) Labels

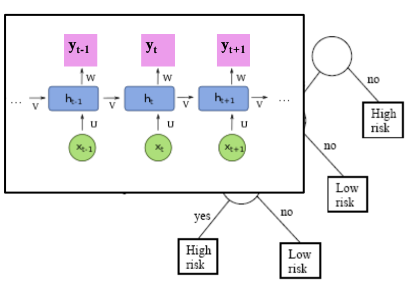
- Semi-supervised topic models — ***Hughes*** *et al. AISTATS 2018*
- End-to-end training of POMDPs for reinforcement learning — *Futoma, **Hughes**, et al. AISTATS 2020*



New Variational Algorithm: Scalable yet Reliable

Adapt Model Size to Data (Bayesian Nonparametrics)

- Add clusters during training — ***Hughes** & Sudderth NeurIPS 2013*
- Topic models for news articles — ***Hughes**, Kim & Sudderth AISTATS 2015*
- HMMs for mocap and genomics — ***Hughes** et al., NeurIPS 2015*
- Image composition models — *Ji, **Hughes**, & Sudderth ICML 2017*
- Speed-up model comparison — *Zhang & **Hughes**, AABI 2019*



New Training Objectives for Deep Neural Nets

Optimize for interpretability, don't just interpret afterwards

- Find diverse explanations
- Find tree-like neural nets

*Ross,, **Hughes**, Doshi-Velez IJCAI 2017*
*Wu, **Hughes**, Parbhoo, et al. AAAI 2018*
*Wu, Parbhoo, **Hughes** et al. AAAI 2020*

BNP Statistical Models : `github.com/bnpy/bnpy`
Time-series Prediction : `github.com/tufts-ml/time_series_predict`

Hughes Lab @ Tufts CS

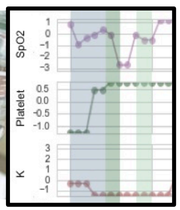
Area: statistical machine learning; **clinical informatics**

Lab goal: *Reliable training of interpretable models for real-world decisions*

Personalize treatments for major depression

- Discover subtypes and best treatments with topic models

***Hughes** et al. AISTATS 2018*
***Hughes** et al. in submission to JAMA Open*
Drs. McCoy and R. Perlis (MGH/Harvard)



Personalize treatments in the Intensive Care Unit

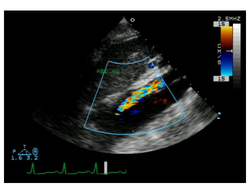
- Suggest interventions
- Address non-stationarity features

Ghassemi et al. AMIA CRI 2017
Nestor et al. MLHC 2019
Drs. R. Kindle and L. Celi (Beth Israel)

Predict mortality from chemotherapy for leukemia

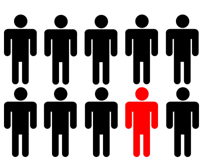
- Balance costs to decide who gets high-risk treatment

Siddiqui et al. Amer. Soc. Hematology 2019
Drs. N. Siddiqui, A. Klein, et al. (Tufts Med.)



Detect heart disease from few labeled images

In progress, Dr. Ben Wessler (Tufts Med.)



Predict individual treatment effects from drug trials

In progress, Dr. David Kent (Tufts Med.)

ICU Time-series Benchmarks: **`github.com/MLforHealth/MIMIC_Extract`**

Roadmap

- Motivation: Improve interventions in ICU
- Models for Clustering Structured Data
- Method: Prediction-Constrained Training

  Hughes et al. *AISTATS 2018*
- Prediction-Constrained HMMs

  Hope, ***Hughes****, Sudderth, et al. In Progress*
- Prediction-Constrained POMDPs

  Futoma, ***Hughes****, Doshi-Velez AISTATS 2020*

Problem: When will ICU patient **need intervention**?

*Ghassemi, Wu, **Hughes**, et al. AMIA CRI 2017*

Interventions:

- Ventilators to assist breathing
- Drugs to manage blood pressure

Early prediction helps:

prepare patient
plan staffing
try less aggressive options early

Data: ~30,000 ICU patients

mimic.physionet.org

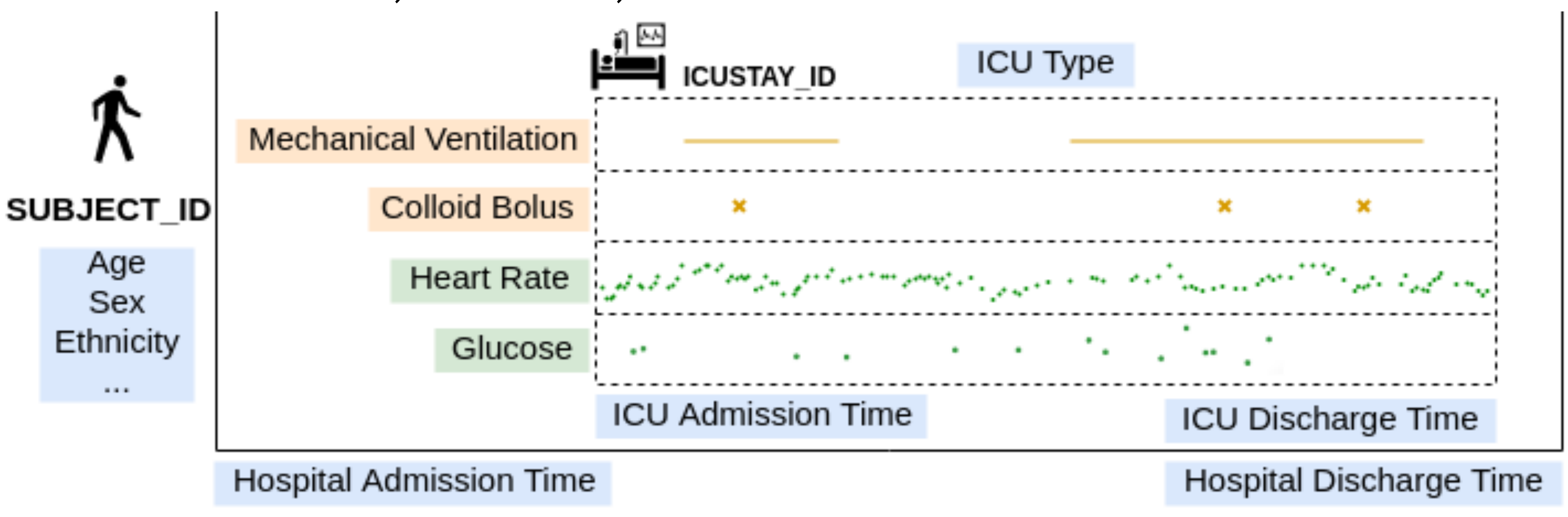
(Johnson et al. Sci. Data 2016)

Nurse-validated vital signs (irregular, hourly)

- heart rate, blood pressure, temp., SpO2, …

Lab measurements (irregular, every few hours)

- hematocrit, lactate, …



Key Goals for our Model

- How should we deal with missing data values?
  - We cannot draw blood every hour

- How to deal with missing labels?
  - Most patients never get some treatments of interest

- Punchline:

  Model is always wrong… **Is it sometimes useful?**

  Can we **adjust fitting procedure** to make more useful?

Approach:
Model data and labels with a **joint probabilistic graphical model**

$$\overset{\text{data}}{} \quad \overset{\text{labels}}{}$$

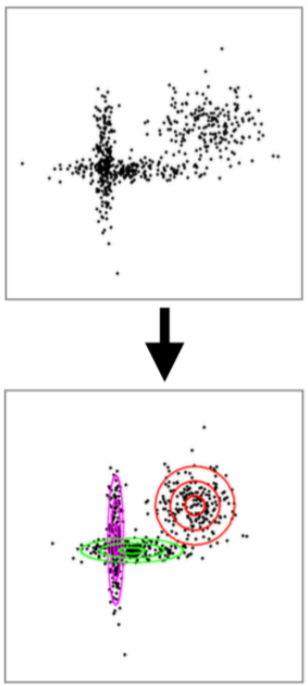
$$p(\underbrace{x}_{}^{\text{data}}, \overset{\text{labels}}{y}) = p(y|x)p(x)$$

Why a joint model?

- p(x) can help us reason about missing data
- p(y | x) can help us predict labels from data
  - even if some labels are missing from training set
- Tying these together makes
  - All parts work in unison
  - Simplifies training: solve one problem, not several disconnected pieces

Structured Clustering Models

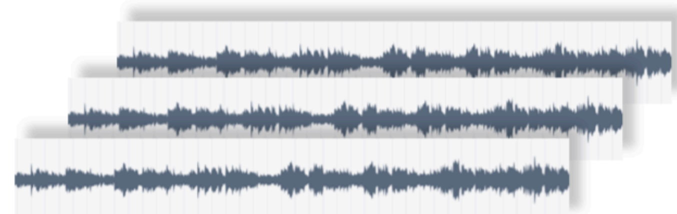
Mixture Models

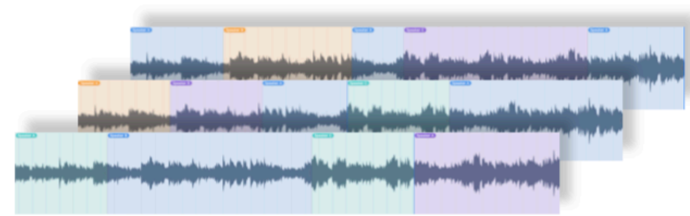


Topic Models

Oscar	Horror	Comedy	
Thrilling	Monster	Laughs	...
Screenplay	Dead	Low-brow	
Writing	Zombies	Hilarious	
Characters	Violence	Ferrell	
Dialogue	Suspense	Fun	
...	...	...	

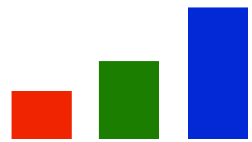
Hidden Markov Models



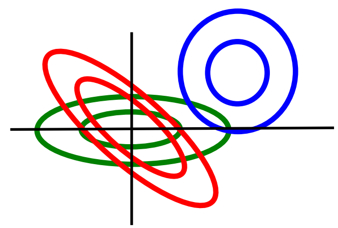


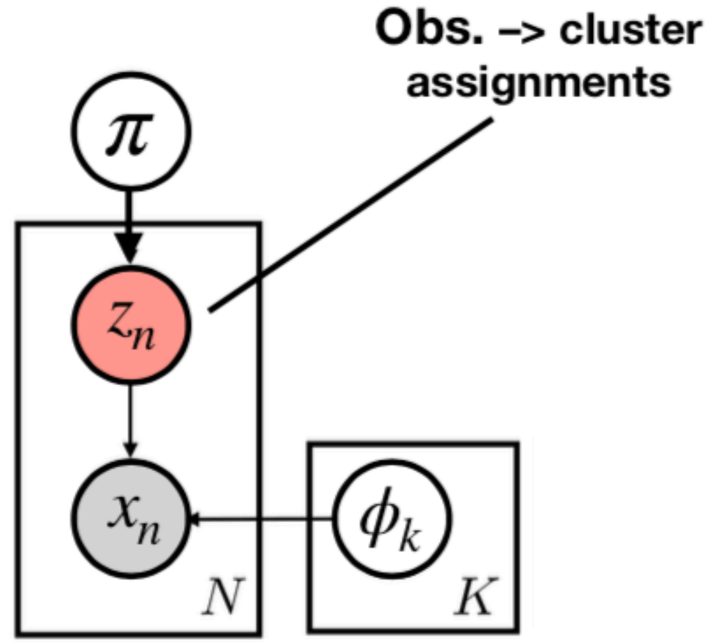
Simple case: Gaussian Mixture

cluster frequency π



cluster shape ϕ





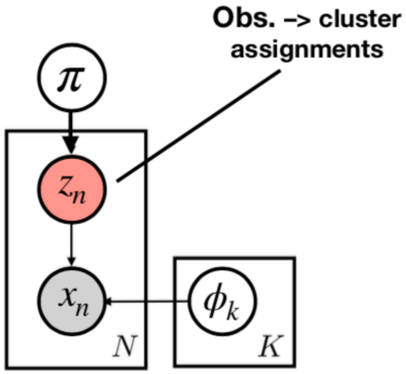
$$\phi_k = \{\mu_k, \sigma_k^2\}$$

obs.-to-cluster assignment $z_n \sim \mathrm{Discrete}(\pi_1, \ldots \pi_K)$

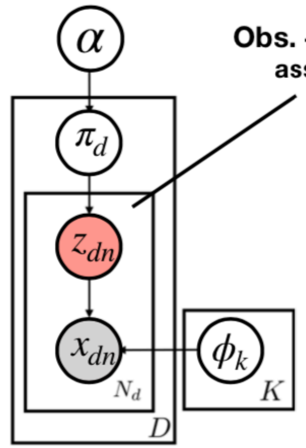
observed data $x_n | z_n{=}k \sim \mathrm{Normal}(\mu_k, \sigma_k^2)$

Z: Per-Example Membership

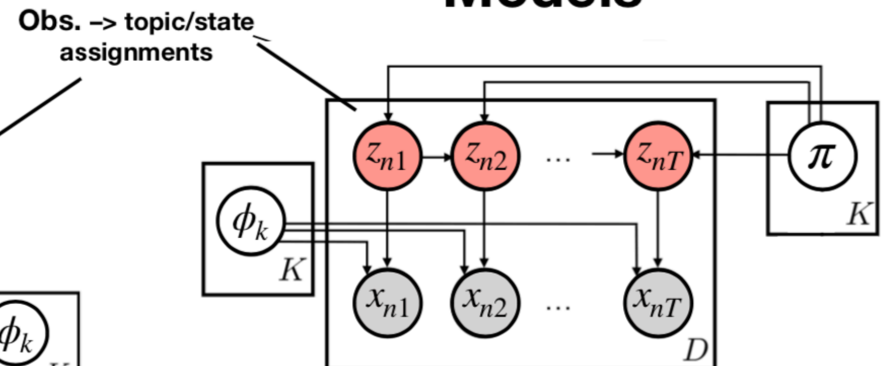
Mixture Models

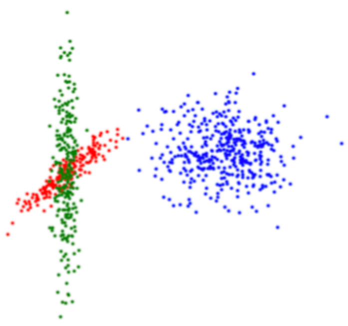


Topic Models

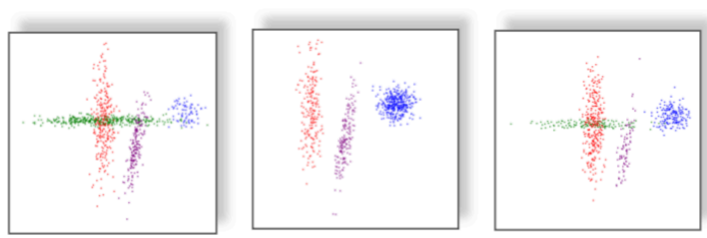


Hidden Markov Models

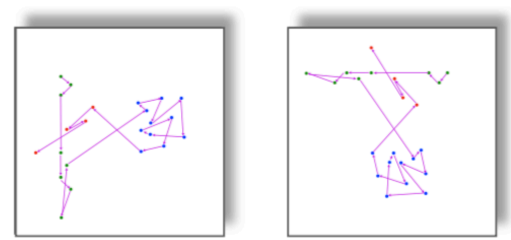




(Generic data)

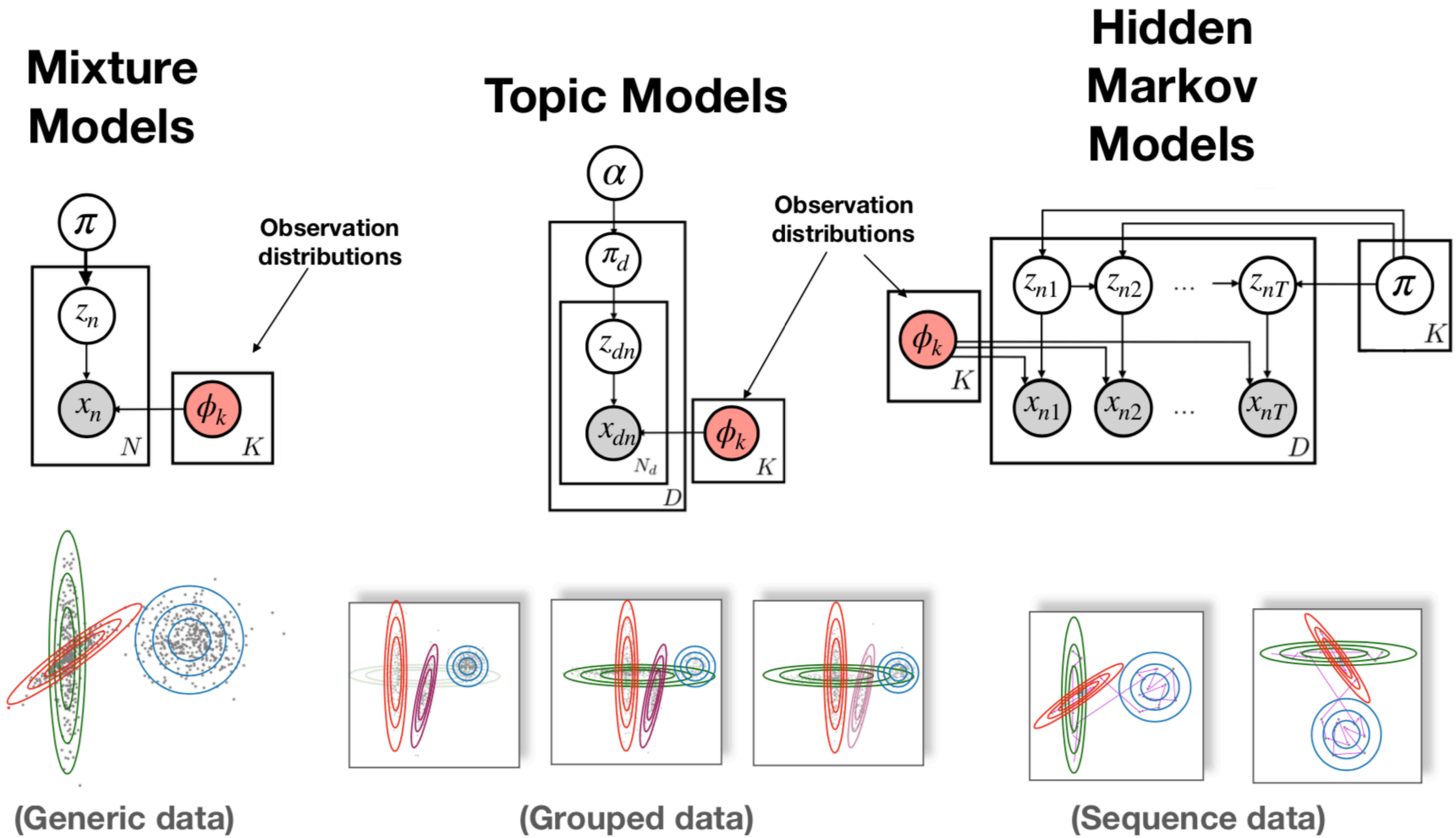


(Grouped data)

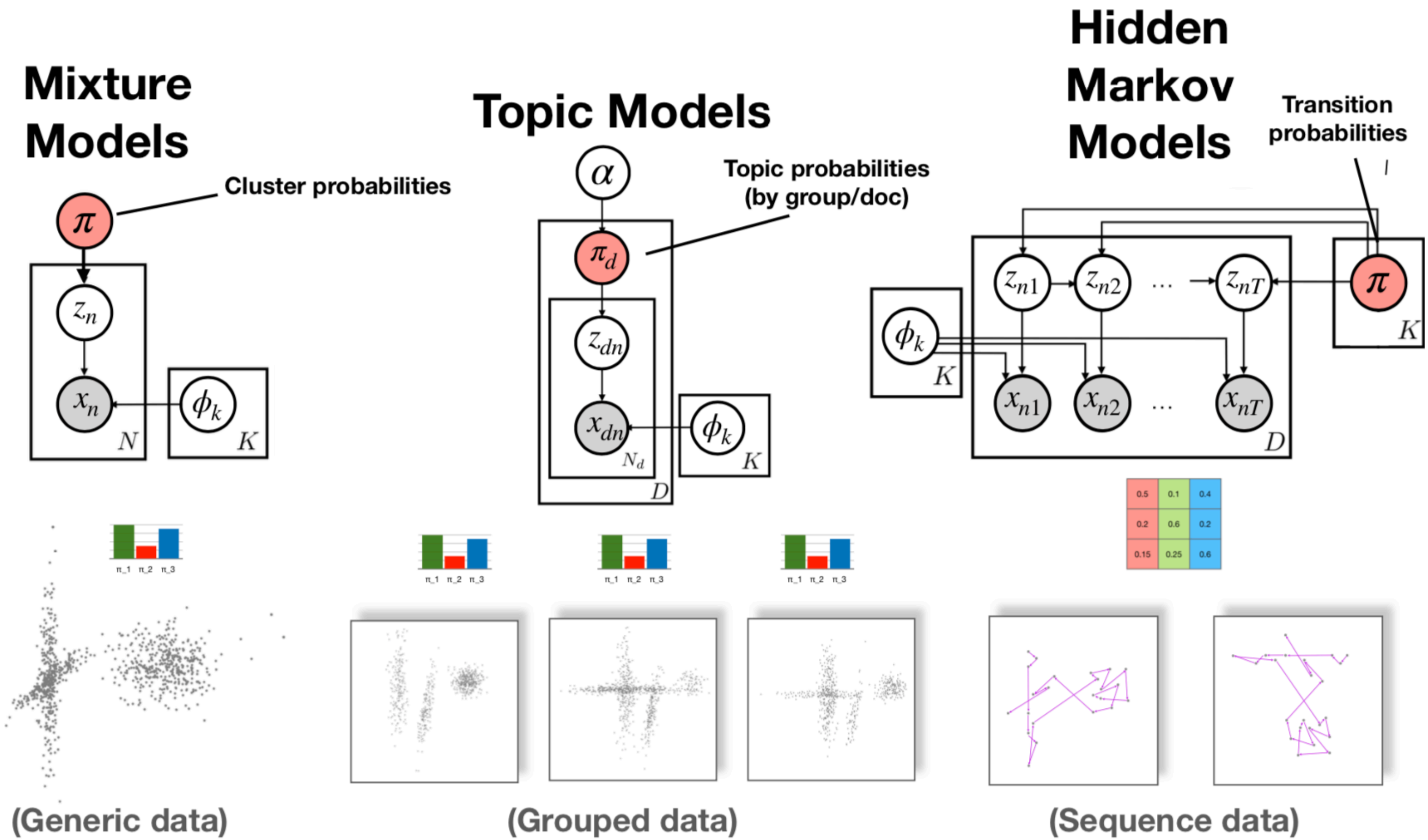


(Sequence data)

Phi: Cluster Emission Parameters

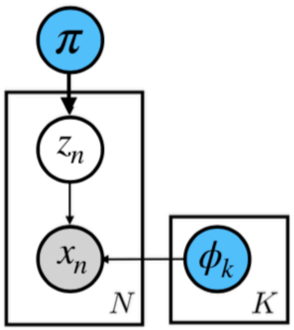


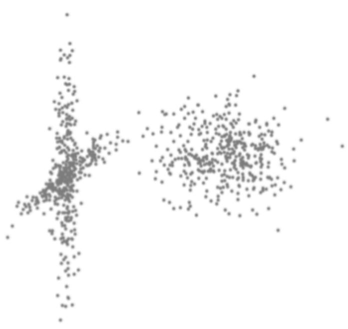
Pi: Cluster Appearance Frequency



How should we add **supervision**?

Mixture Models

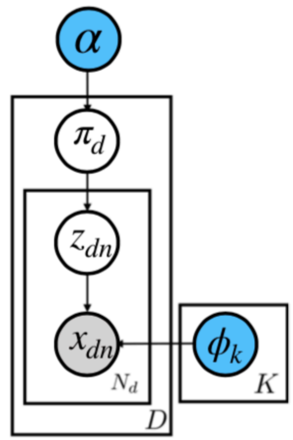




(Generic data)

+ Label per example

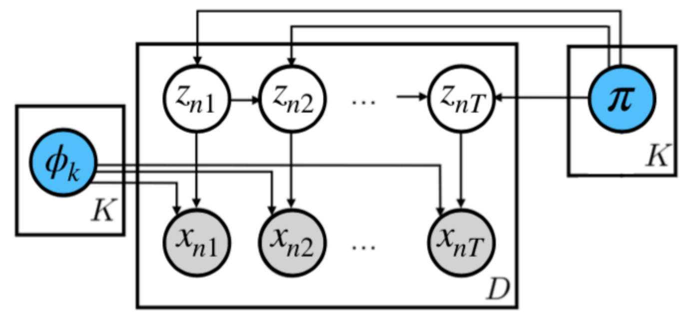
Topic Models

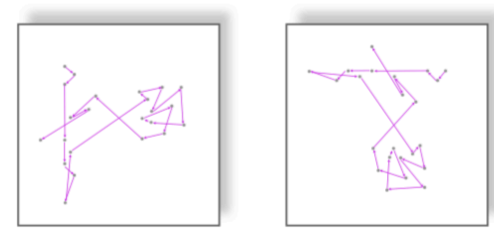


(Grouped data)

+ Label per group

Hidden Markov Models





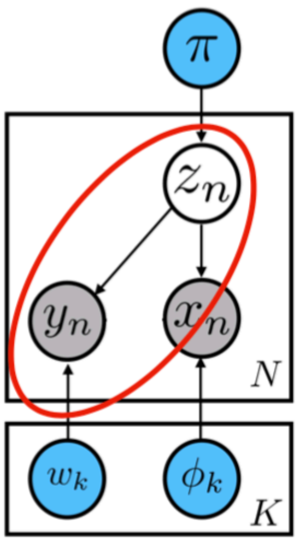
(Sequence data)

+ Label for sequence
+ Label per timestep

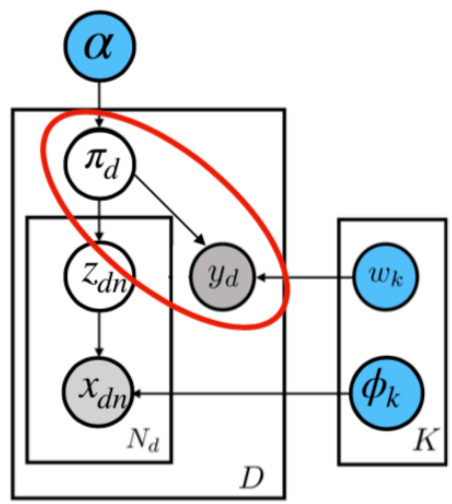
Supervised clustering models

Predict labels as a function of the cluster membership:

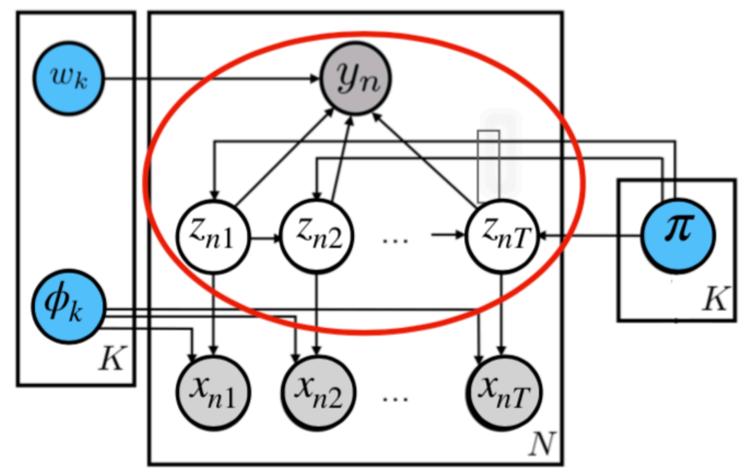
Mixture Models



Topic Models

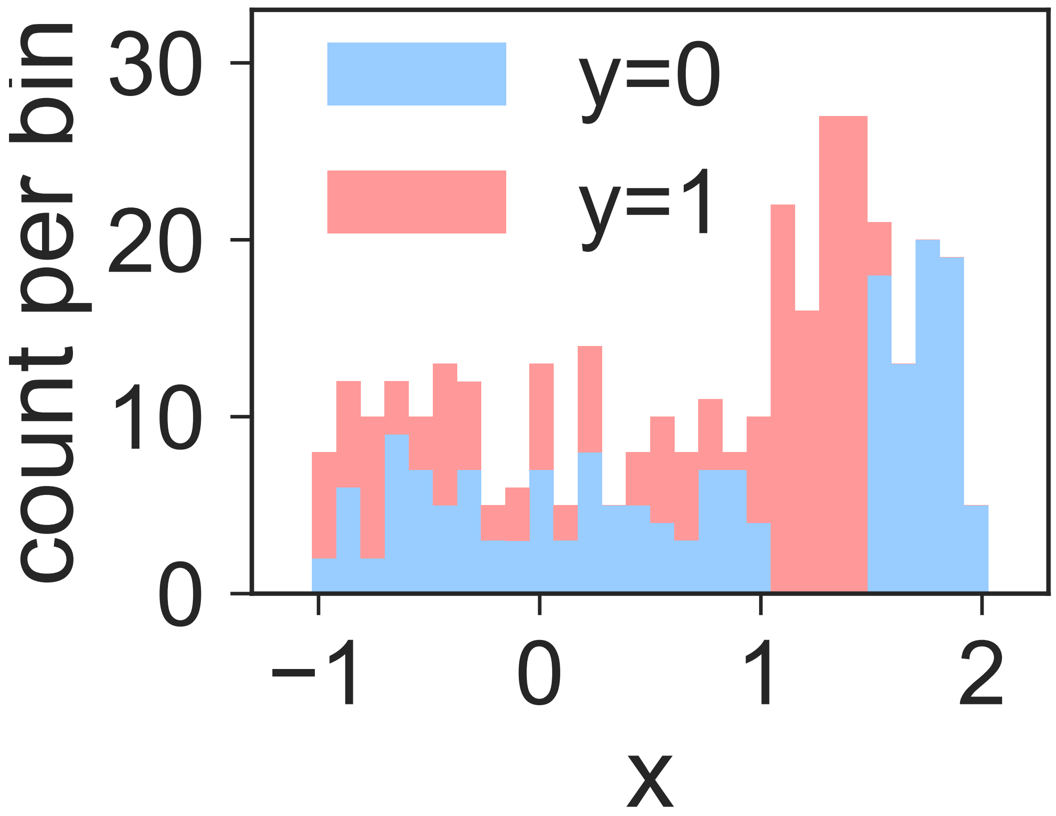


Hidden Markov Models



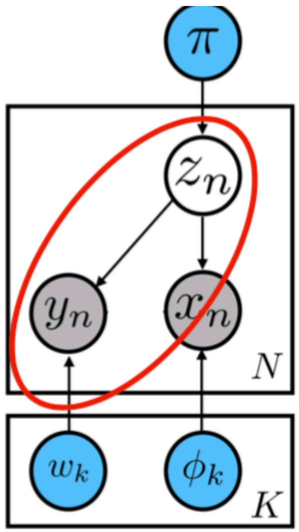
Directly modeling a label likelihood p(y | z) makes it easy when y is missing
Class conditional models like p(z | y) would require expensive marginalization

Simple Challenge: Model this Data!



Supervised Gaussian Mixture Model

Assume K possible clusters

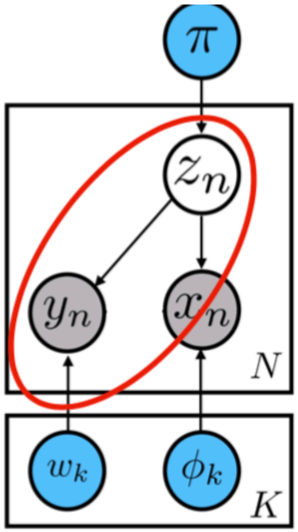


$$z_n \sim \text{Discrete}(\pi_1, \dots \pi_K)$$

$$x_n | z_n = k \sim \text{Normal}(\mu_k, \sigma_k^2)$$

$$y_n | z_n = k \sim \text{Bern}(w_k)$$

Supervised Gaussian Mixture Model



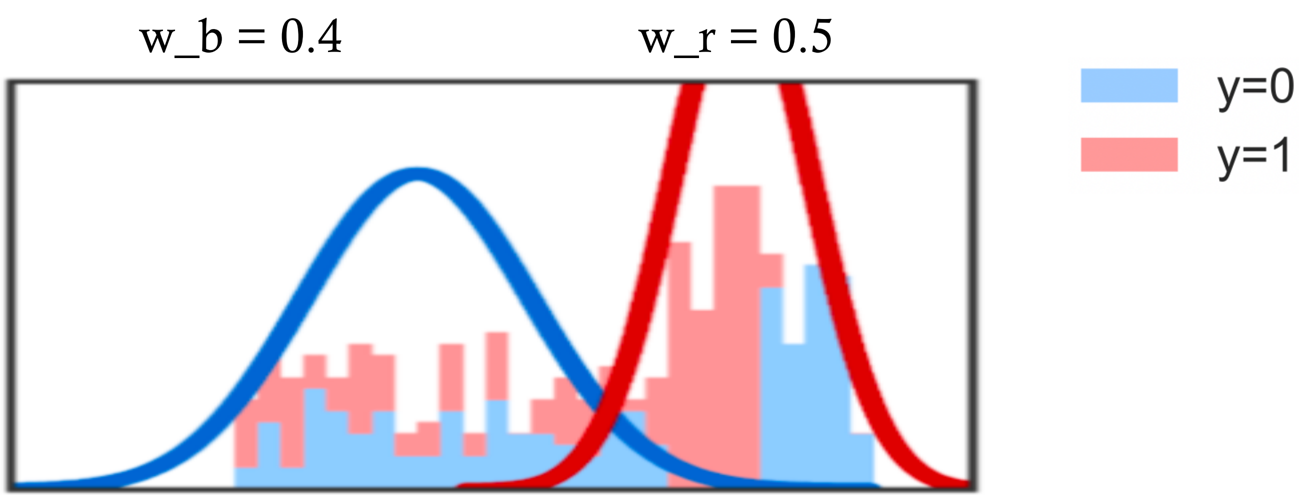
Assume K possible clusters

$$z_n \sim \mathrm{Discrete}(\pi_1, \ldots \pi_K)$$

$$x_n | z_n = k \sim \mathrm{Normal}(\mu_k, \sigma_k^2)$$

$$y_n | z_n = k \sim \mathrm{Bern}(w_k)$$

K = 2



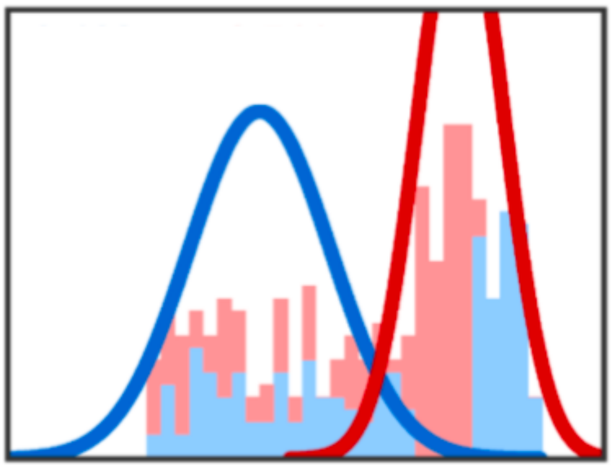
Haven't we known how to fit these models for >30 years?

$$\max_{\pi,\phi,w} \sum_{n=1}^{N} \log p(x_n, y_n | \pi, \phi, w)$$

Result:

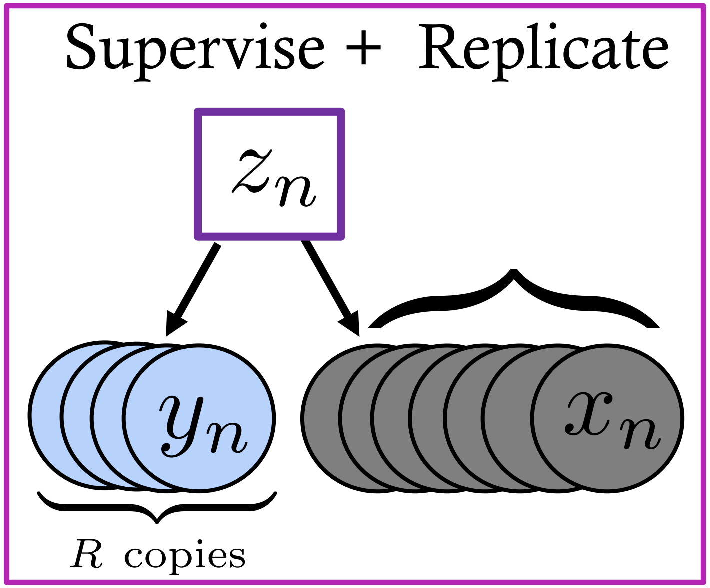
Terrible label prediction!

Forced to compromise p(y|x) to make p(x) look good



If my application need prioritizes p(y|x), maximizing joint likelihood may not yield useful results

Past Work Attempted Fix: Label Replication



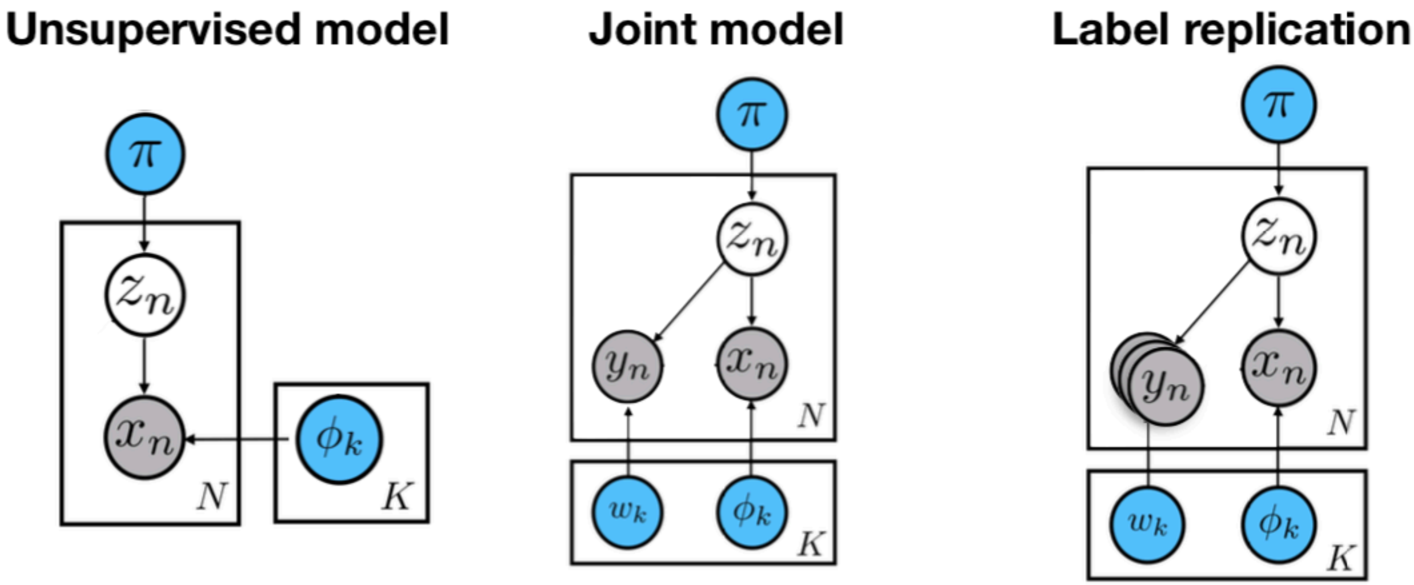
$$\max_{\phi, w} \quad \log p(x, y, y, \ldots y | \phi, w)$$

Proposed by

- Zhang & Kjellstrom (2014) as "Power sLDA"
- Zhu et al. (2012) as "Med-LDA"
- Ganchev et al. (2010) as "Posterior Regularization"

Hughes et al. AISTATS 2018 contribution:
Show many previous efforts equivalent to this basic idea.

No known alternatives work well



$\max_{\pi,\phi} p(x \mid \pi, \phi)$

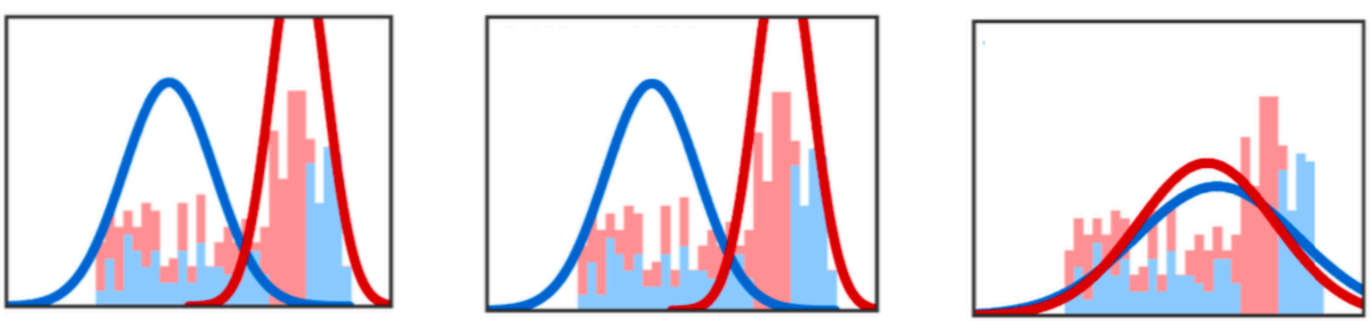
$\max_{\pi,\phi,w} p(x, y \mid \pi, \phi, w)$

$\max_{\pi,\phi,w} p(x, \beta y \mid \pi, \phi, w)$

$|\log p(x_n \mid z_n)| >> |\log p(y_n \mid z_n)|$

Equiv. to

$p(y_n \mid z_n, w) \rightarrow p(y_n \mid z_n, w)^{\beta}$



Prediction-Constrained Training

Key idea: Maximize likelihood of observations…

$$\max_{\pi,\phi} \log p(x \mid \pi, \phi)$$

Subject to: $-\log p(y \mid x, \pi, \phi, w) < \epsilon$

Subject to a ***constraint*** that we can achieve a given performance threshold for predicting labels ***given observations***

How to compute?

$$p(y_n|x_n, \pi, \phi, w) = \sum_{k=1}^{K} p(y_n, z_n = k|x_n, \pi, \phi, w)$$

How to optimize?

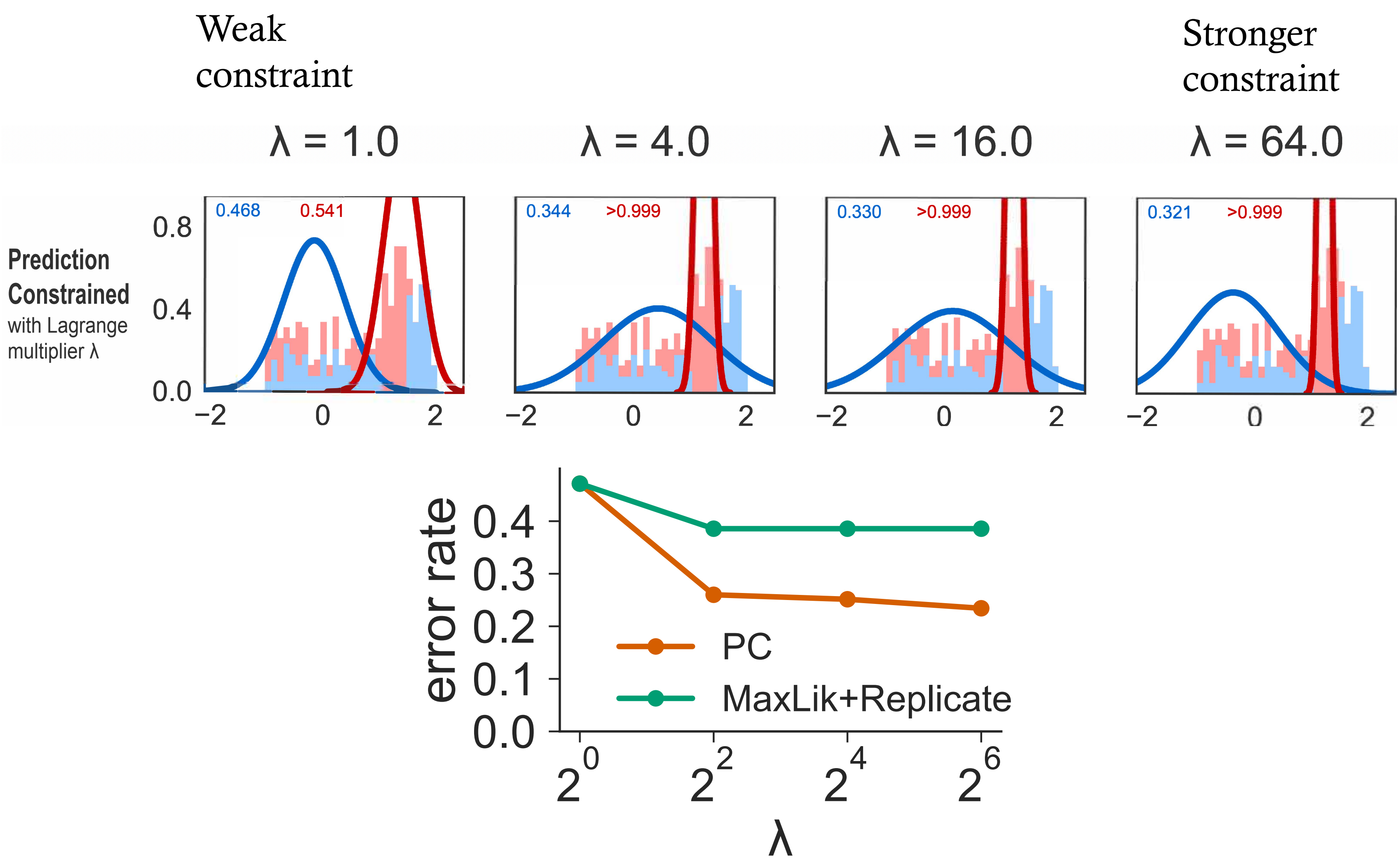
$$\max_{\phi,\eta} \quad \lambda \log p(y|x, \phi, \eta) + \log p(x|\phi)$$

Use Lagrange multiplier to form unconstrained objective

Optimize via stochastic gradient descent

- Write objective as Python code
- Automatic gradients from Tensorflow/Pytorch

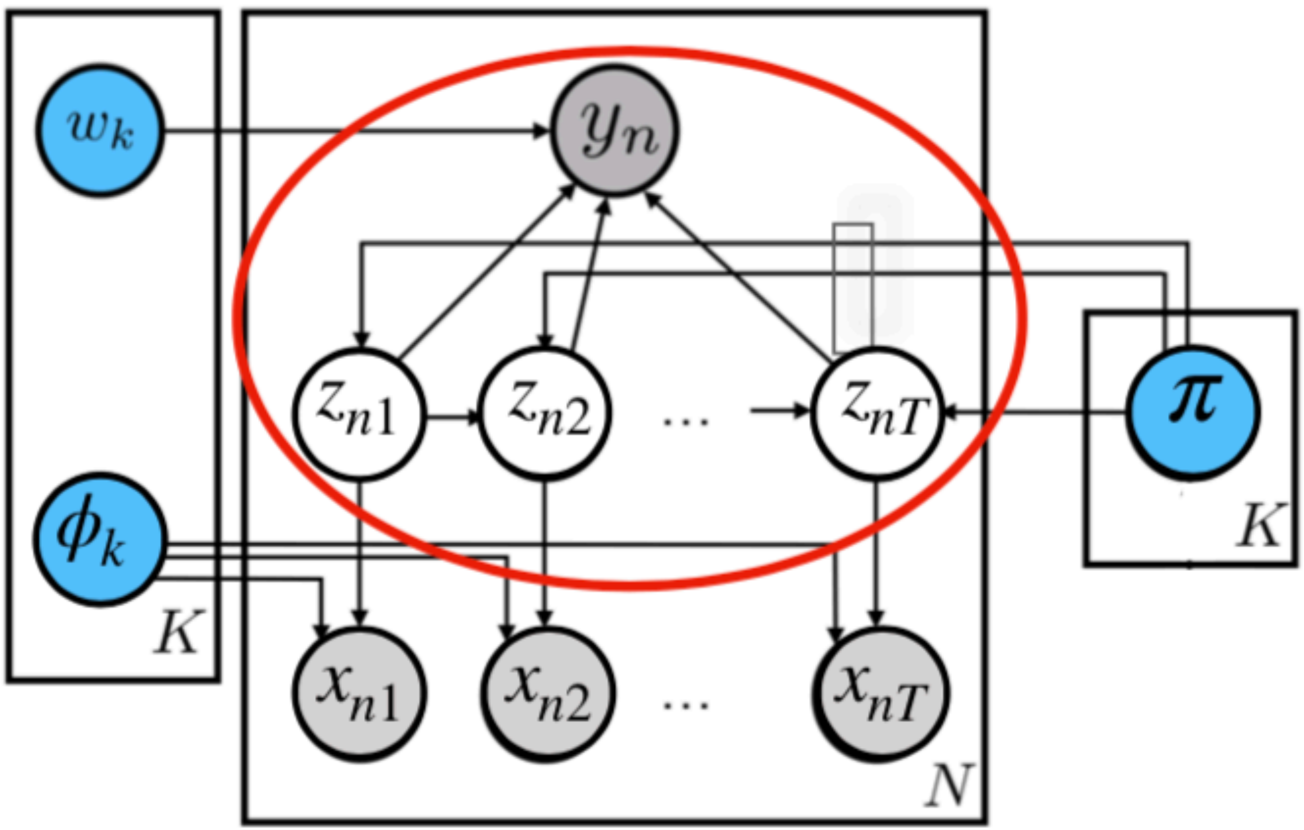
PC can **overcome** misspecification



Roadmap

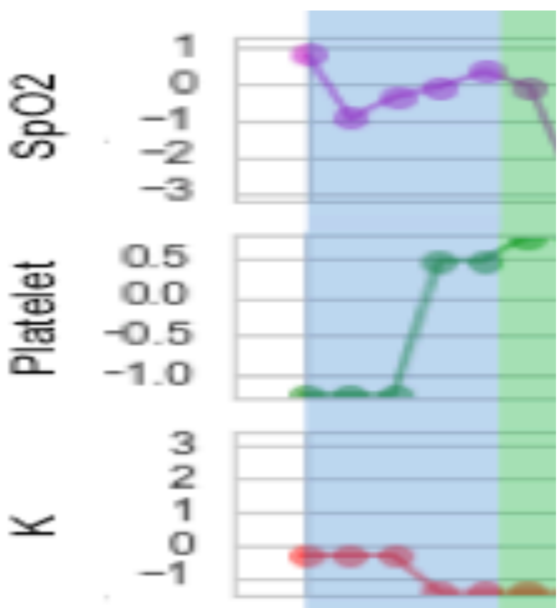
- Motivation: Improve interventions in ICU
- Model Family for Clustering Structured Data
- Method: Prediction-Constrained Training
- **Prediction-Constrained HMMs**
  *Hope, **Hughes**, Sudderth (in progress)*
- Prediction-Constrained POMDPs

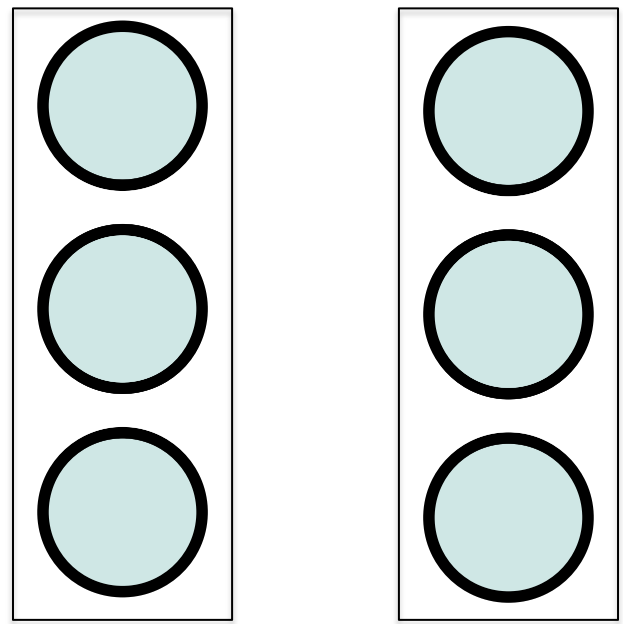
Prediction Constrained Hidden Markov Models



Probabilistic time-series model

Observed Feature Vector x_t





…. hour t hour t+1 …

Probabilistic time-series model

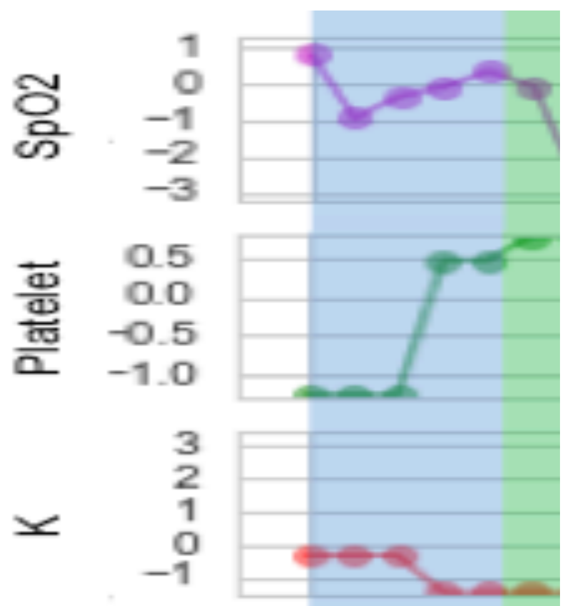
Hidden Patient State

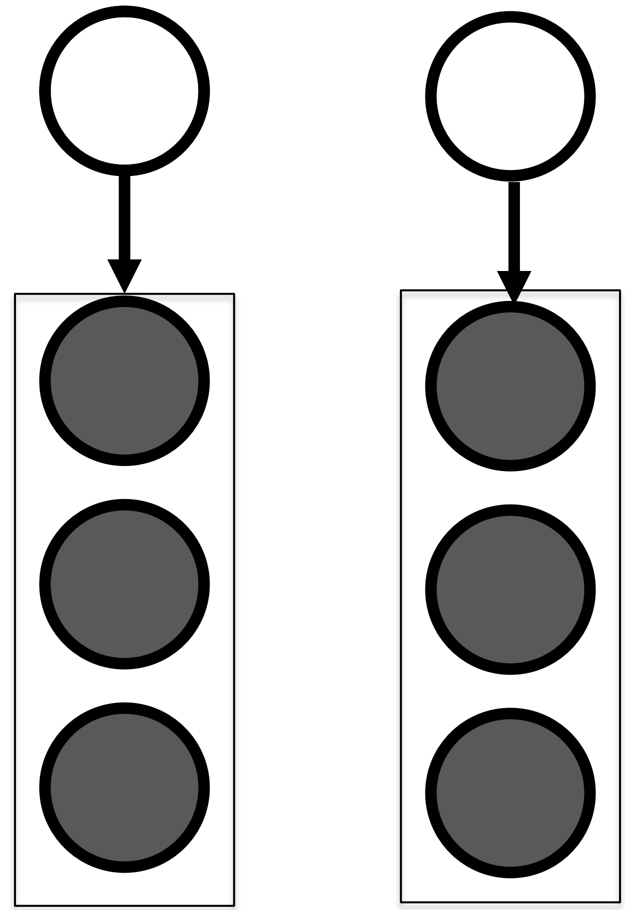
one of K possible values

z_t

Observed Feature Vector

x_t





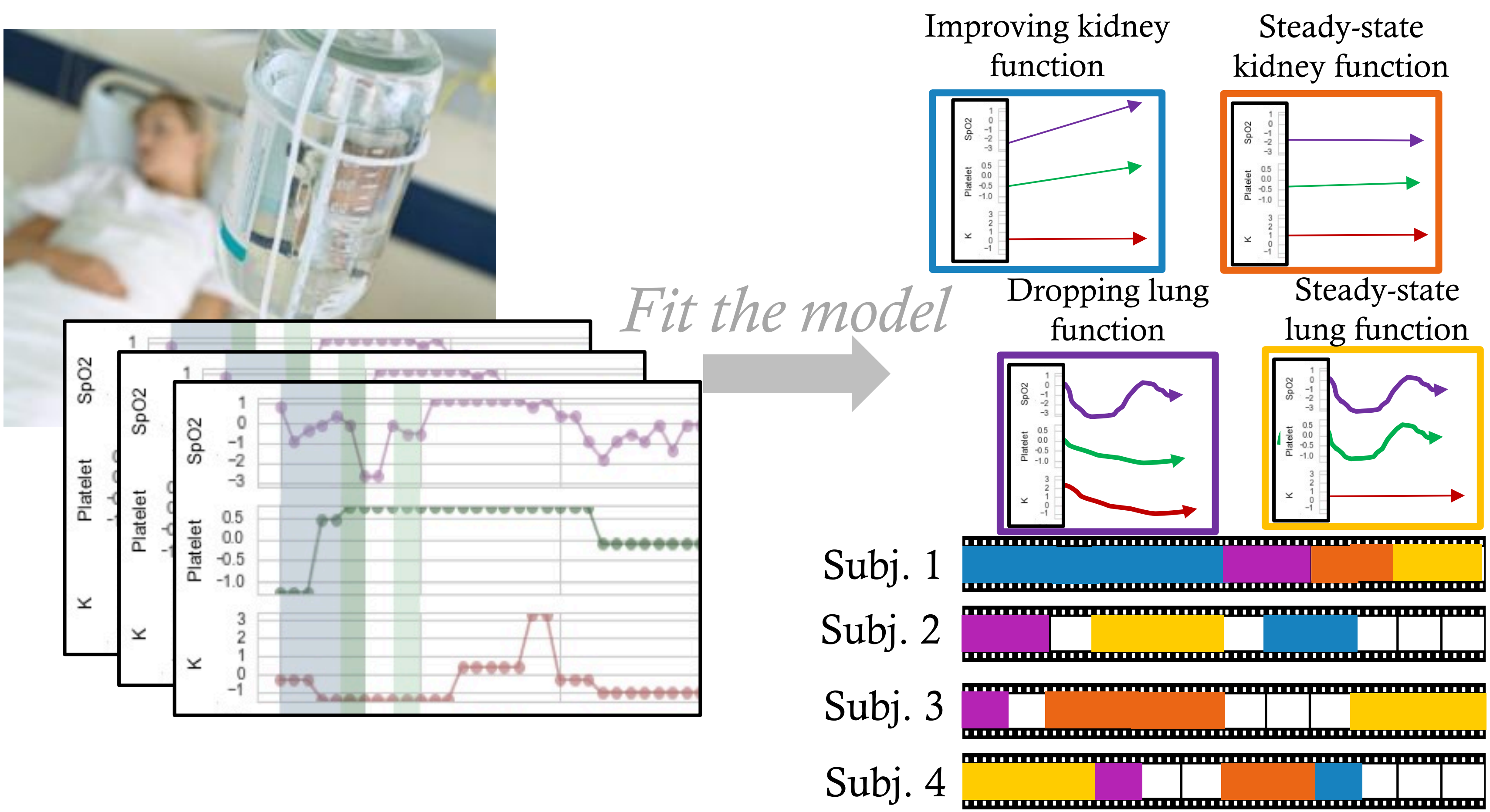
…. hour t hour t+1 …

Goal: Health States Trajectories

Ghassemi, Wu, Hughes, et al AMIA CRI '17

ICU signals from many patients

Health state trajectories



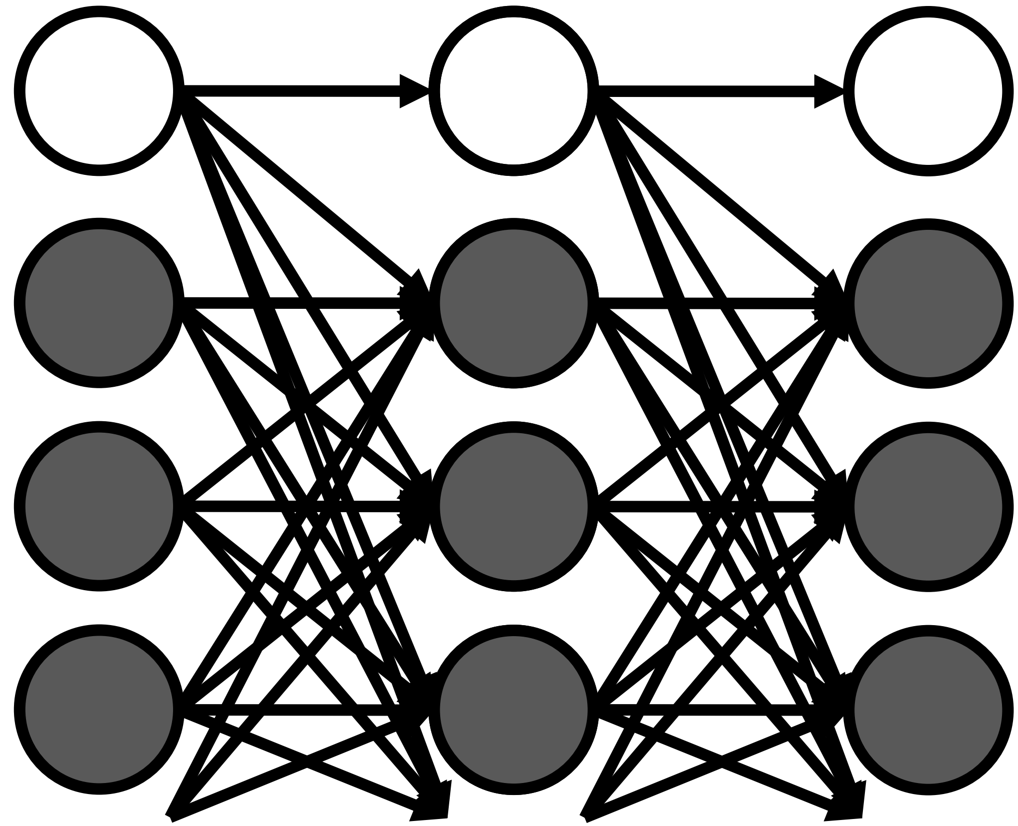
p(x,z) : Autoregressive HMM

Hidden
Patient State

z_t

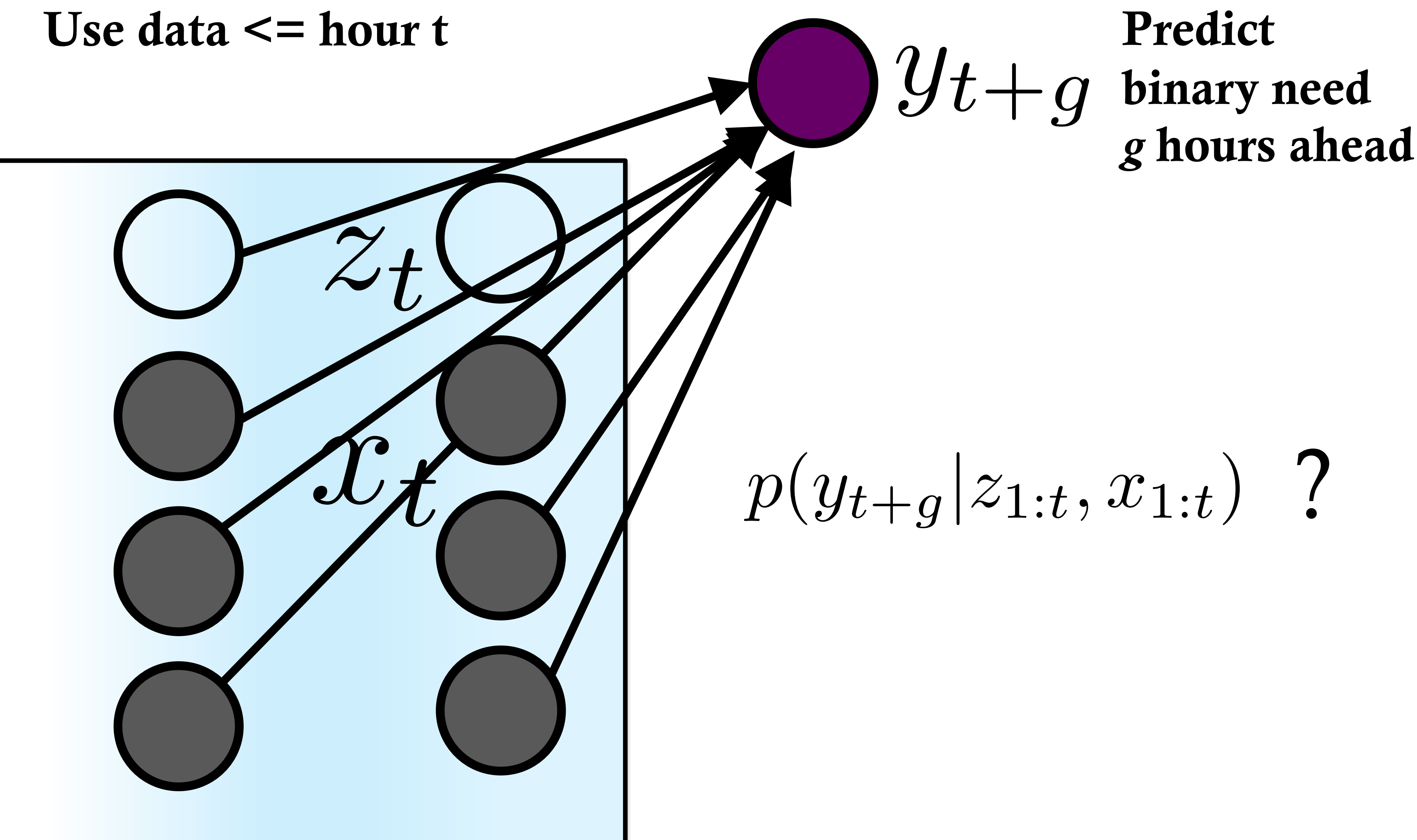
Observed Vitals

x_t



$$z_t | z_{t-1} = j \sim \text{Discrete}(\pi_{j1}, \ldots \pi_{jK})$$
$$x_t | z_t = k \sim \mathcal{N}(A_k x_{t-1} + \mu_k, \Sigma_k)$$

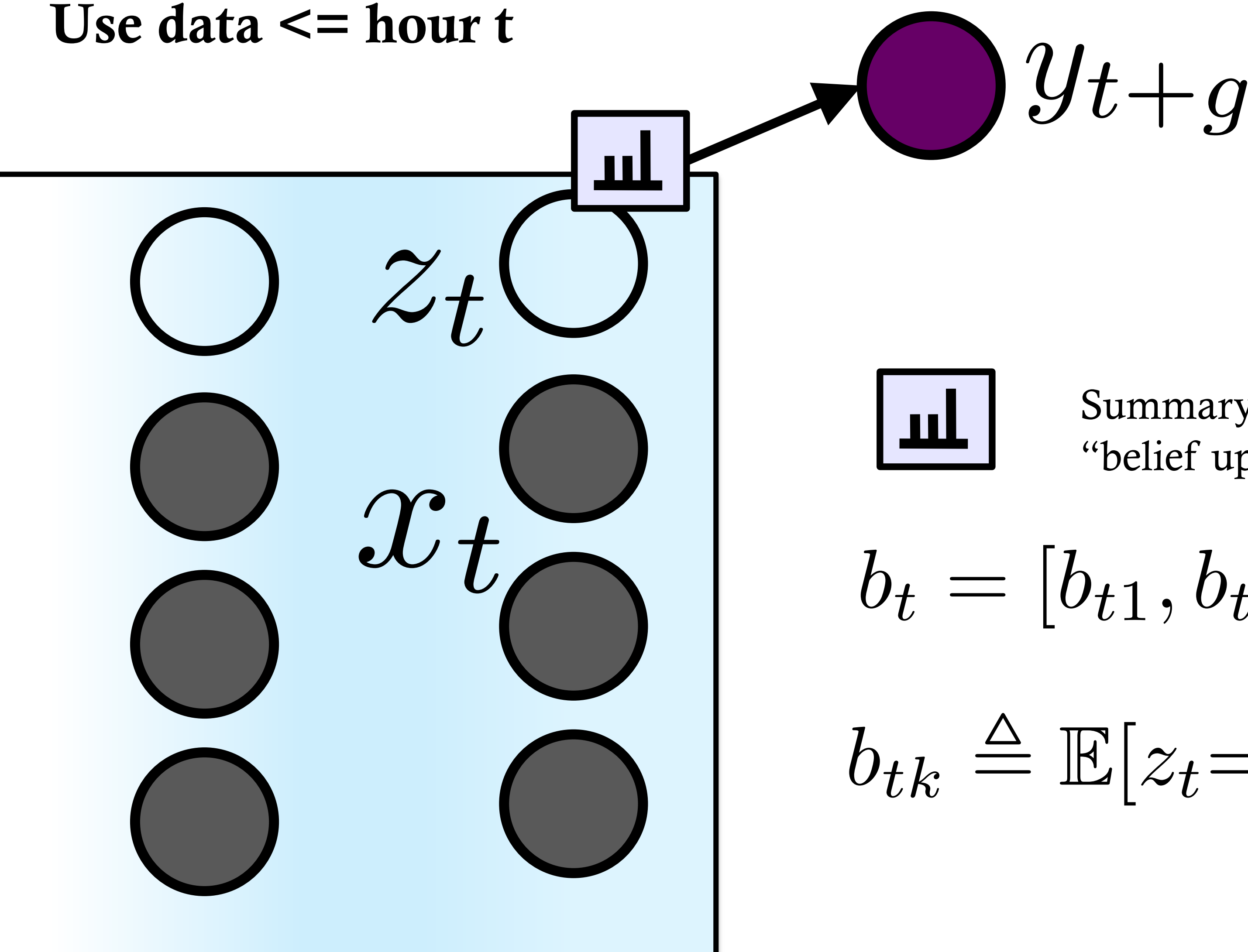
p(y|z,x) : Binary Label Prediction



$$p(y_{t+g} | z_{1:t}, x_{1:t}) \quad ?$$

p(y|z,x) : Binary Label Prediction

Use data <= hour t



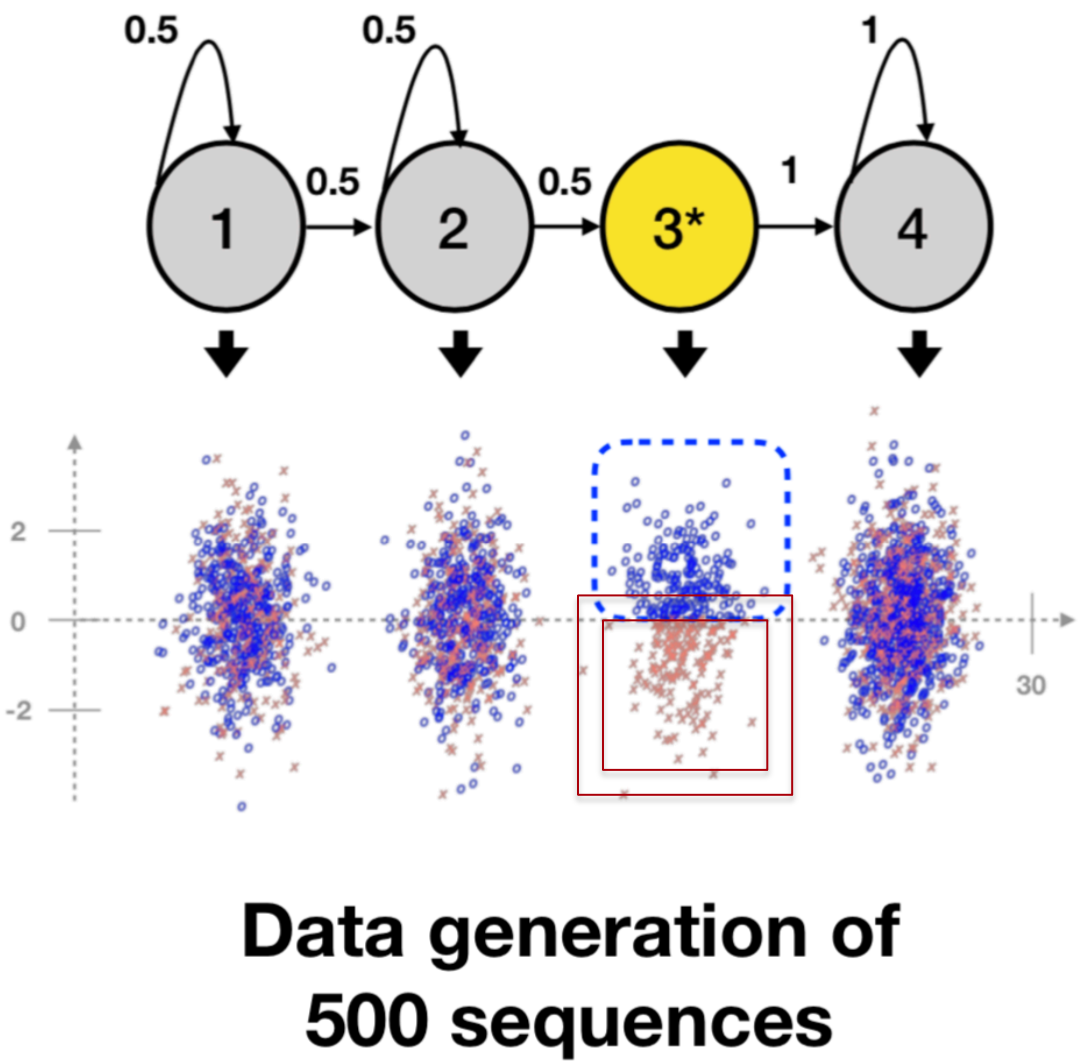
Predict binary need *g* hours ahead

Summary statistic:
"belief up to time *t*"

$$b_t = [b_{t1}, b_{t2}, \ldots b_{tK}]$$

$$b_{tk} \triangleq \mathbb{E}[z_t{=}k|x_{1:t}]$$

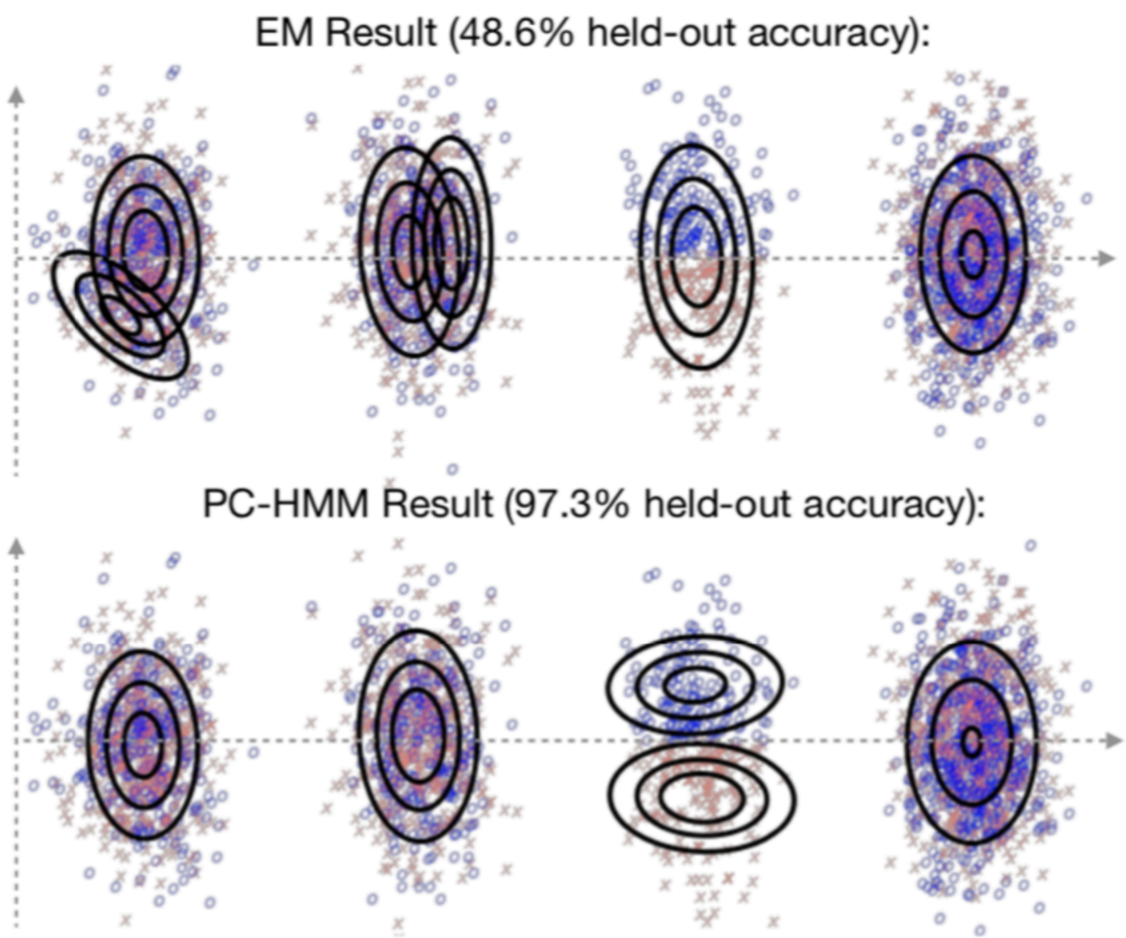
Example HMM



Each sequence gets binary label

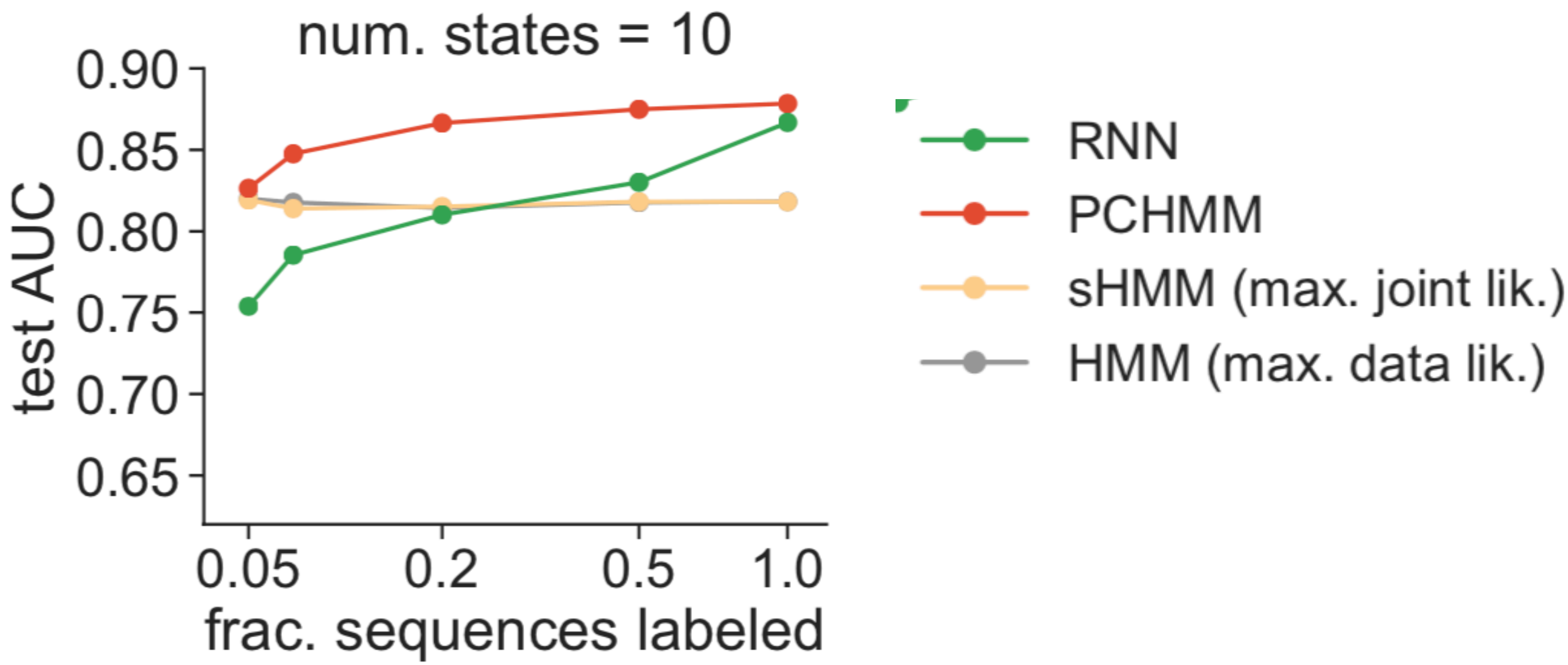
1 if above x-axis
0 if below x-axis

Fit with 100% sequences labeled



ICU Need for Ventilator Prediction

Using autoregressive HMM with 10 states



- PC is strictly better than maximum likelihood training
- When labels are rare, PC > deep learning on labels only

Roadmap

- Motivation: Improve interventions in ICU
- Model Family for Clustering Structured Data
- Method: Prediction-Constrained Training
- Prediction-Constrained HMMs
- **Prediction-Constrained POMDPs**

*Futoma, **Hughes**, Doshi-Velez AISTATS 2020*

RL IN GENERAL…

RL FOR ACUTE HYPOTENSION

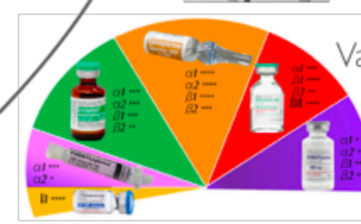
- High morbidity & mortality
- Heterogeneous: many causes
- Treatment not standardized

Environment

State

Action

Fluid
Bolus



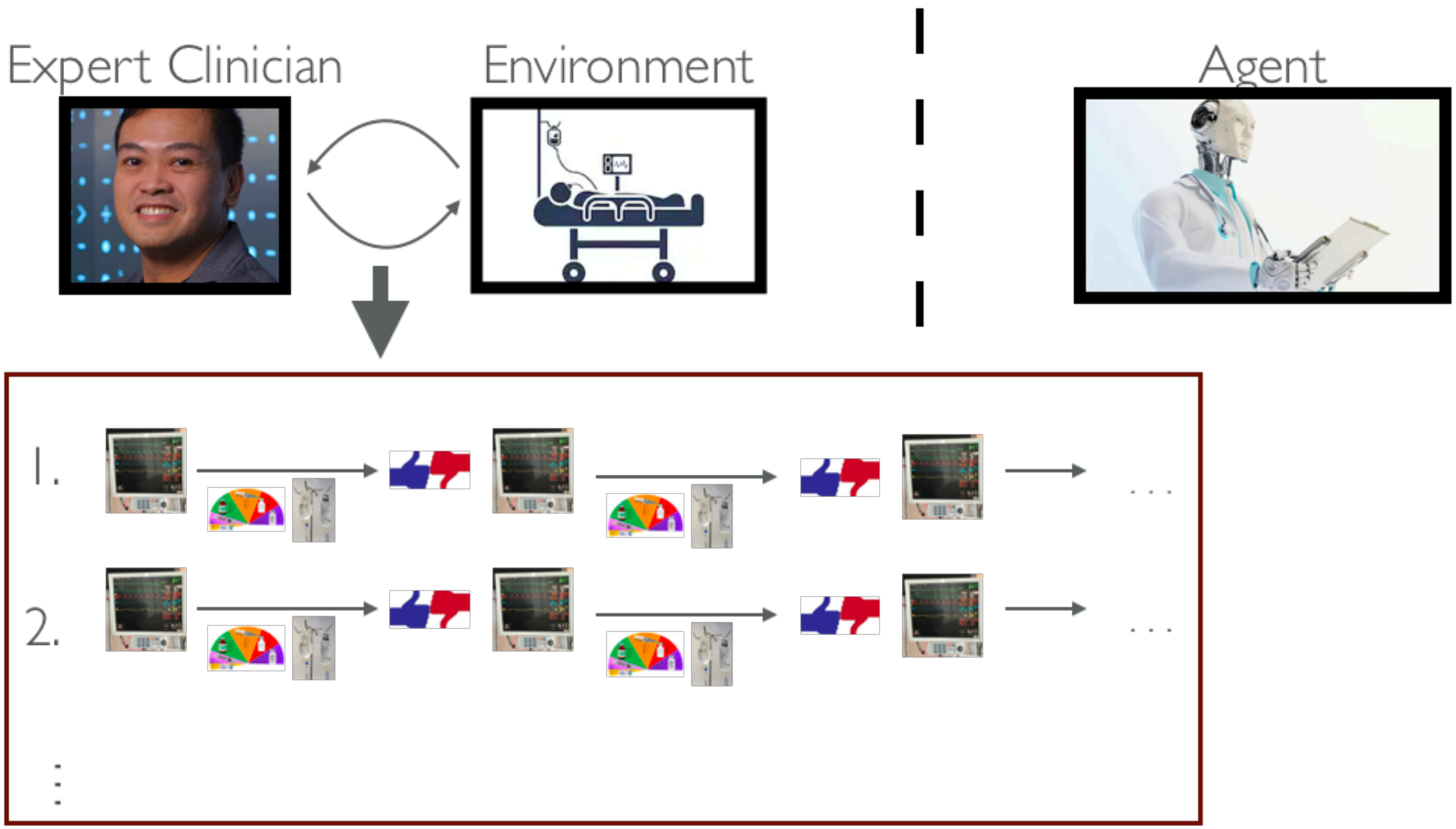
Vasopressors

Reward

Agent

RL FOR ACUTE HYPOTENSION

Retrospective data ONLY!

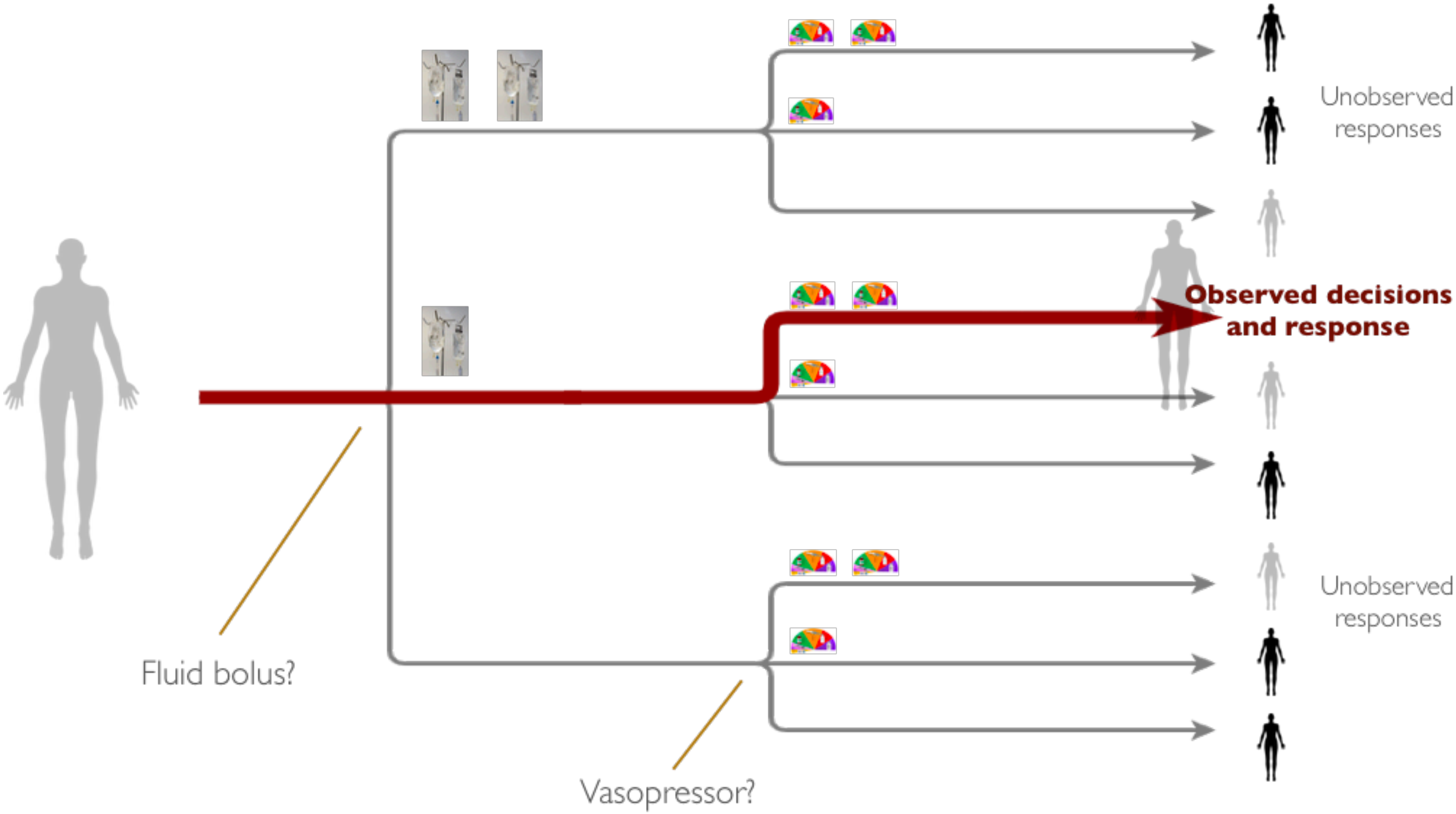


What is needed for Clinical RL

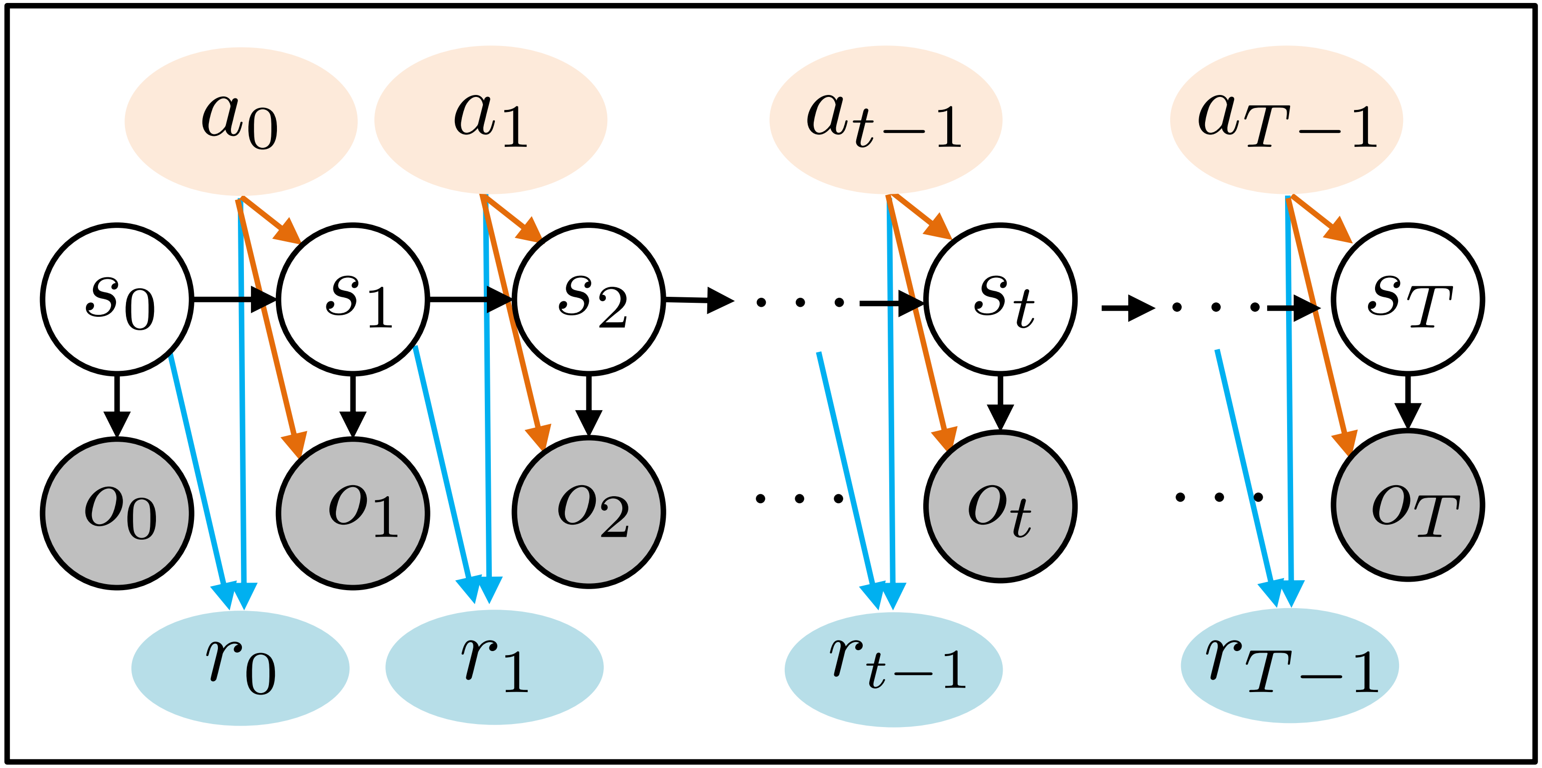
- Learn from retrospective histories only
  - Called "batch" setting of RL
- Use model-based RL
  - Can deal with **little data** and **missing data**
  - Can do **forecasting** and **simulation**
- Handle unknown state space
  - "POMDP": partially observed Markov decision process

Need: to **avoid misspecification** and get high reward

Actions: fluid & vaso at each hour



POMDP as structured clustering model: Input/Output-HMM



Estimating Value from Off Policy Data

$$V^{\text{CWPDIS}}(\pi_\theta) \triangleq \sum_{t=1}^{T} \gamma^t \frac{\sum_{n \in \mathcal{D}} r_{nt} \rho_{nt}(\pi_\theta)}{\sum_{n \in \mathcal{D}} \rho_{nt}(\pi_\theta)},$$

$$\rho_{nt}(\pi_\theta) \triangleq \prod_{s=0}^{t} \frac{\pi_\theta(a_{ns} | o_{n,0:s}, a_{n,0:s-1})}{\pi_{\text{beh}}(a_{ns} | o_{n,0:s}, a_{n,0:s-1})}.$$

Consistency Weighted Per-decision Importance Sampling
(CWPDIS, Thomas 2015)
Lower bias but high variance

Prediction Constrained POMDPs

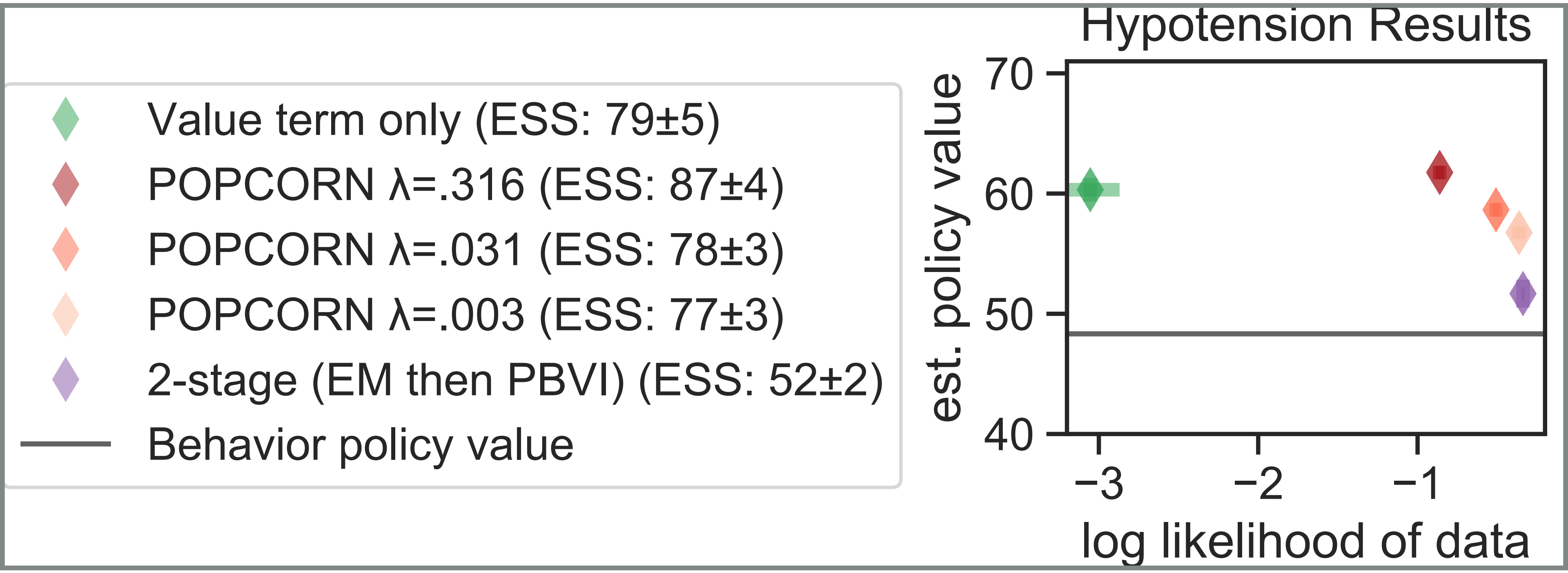
$$\max_\theta \; \mathcal{L}_{gen}(\theta) + \lambda V(\pi_\theta)$$

Generative likelihood of the observations given the model

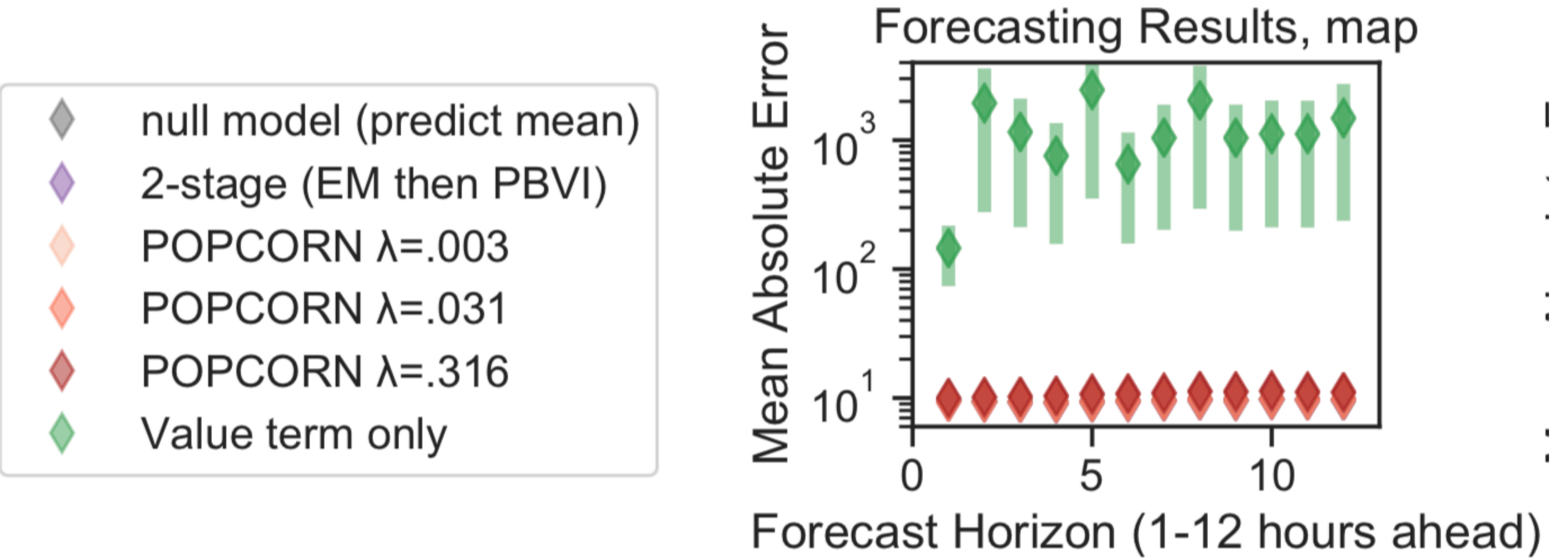
Value of policy Given the model

We call our method "POPCORN":
Partially Observed Prediction Constrained ReiNforcment Learning

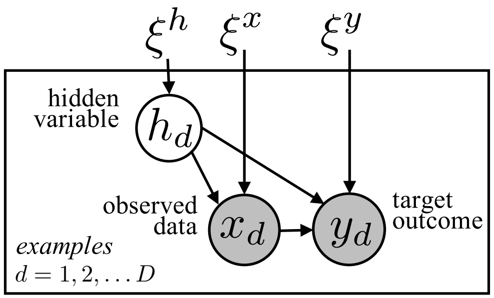
Results: PC-POMDP best for reaching sweet spot of value and likelihood



Can use model for forecasting!

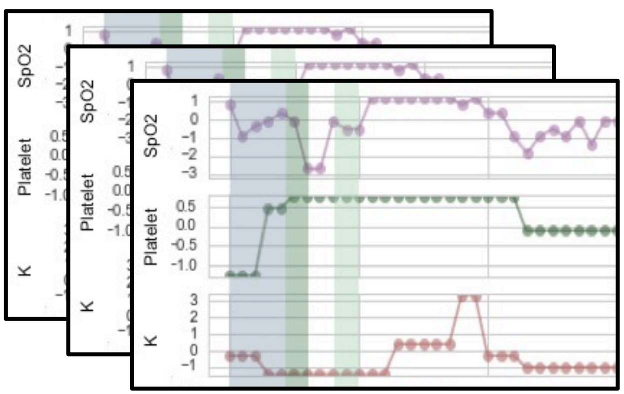


Future: PC training for rich family of **deep generative** models



- Mixture models
- Topic models
- Hidden Markov models
- Network models (MMSB)
- PCA or factor analysis
- Non-negative matrix factorization
- Probabilistic encoder/decoder (VAE)

- Disease progression over time



- Models of many data sources