Topics in Machine Learning
Machine Learning for Healthcare

Rahul G. Krishnan
Assistant Professor
Computer science & Laboratory Medicine and Pathobiology

Announcements

- Friday – project proposals are due; you should all have teams and have begun making your reports; book TA office hours for help/feedback
- Friday: 2 presentations
  - Class participation grade depends on your attending and asking questions
- Poll:
  - Would you be more comfortable in a bigger classroom?

Outline

- Unsupervised disease progression modeling
  - Learning nonlinear state space models
  - Discussion of PPMI model presented by Kristen Severson (Microsoft)
- Alternative strategies for disease progression modeling:
  - Supervised learning
  - Learning from cross-sectional data

Patient data is often sequential

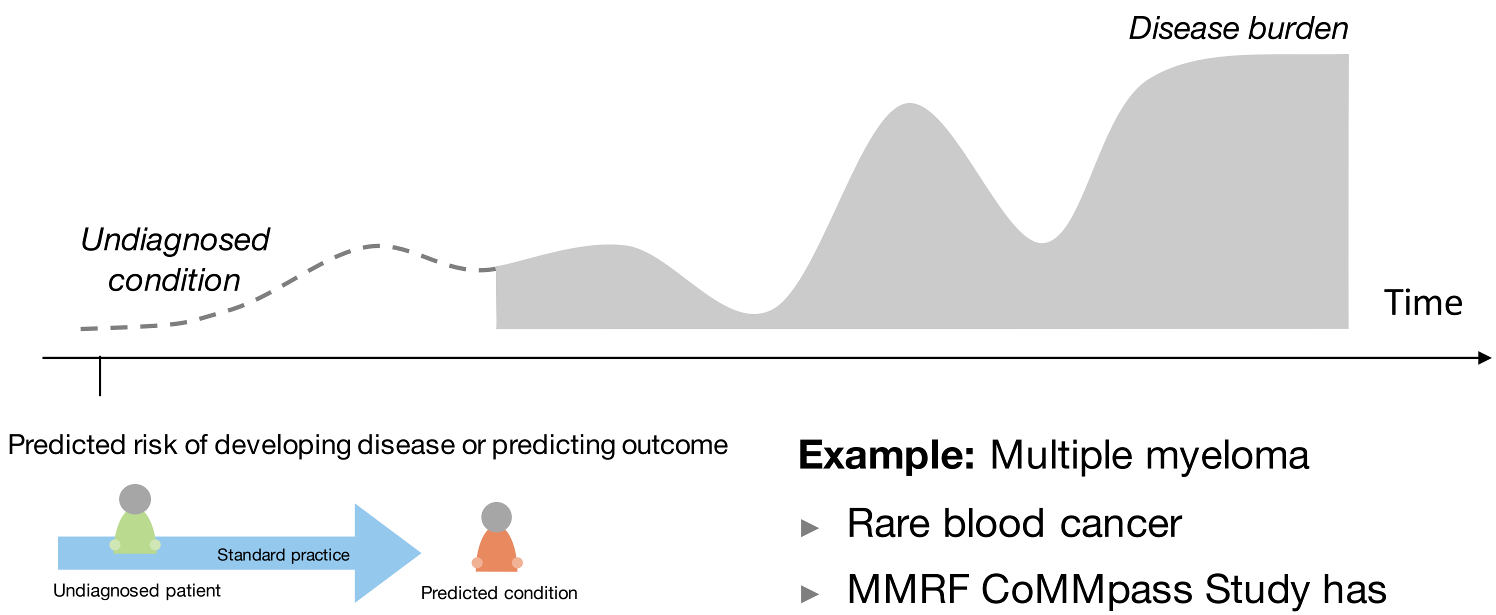
Disease registries track patient data over time

Smartwatch and app sensors collect daily activity data

MULTIPLE MYELOMA Research Foundation

THE MICHAEL J. FOX FOUNDATION FOR PARKINSON'S RESEARCH

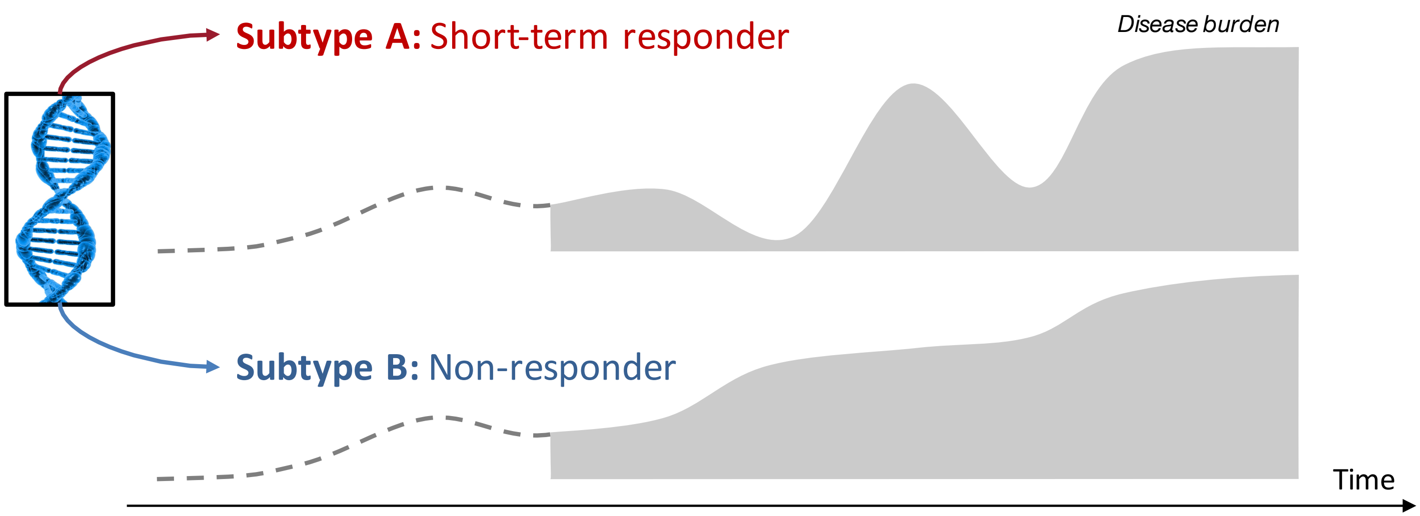
Disease progression – (1)



Example: Multiple myeloma

- Rare blood cancer
- MMRF CoMMpass Study has ~1000 patients

Disease Progression – (2)

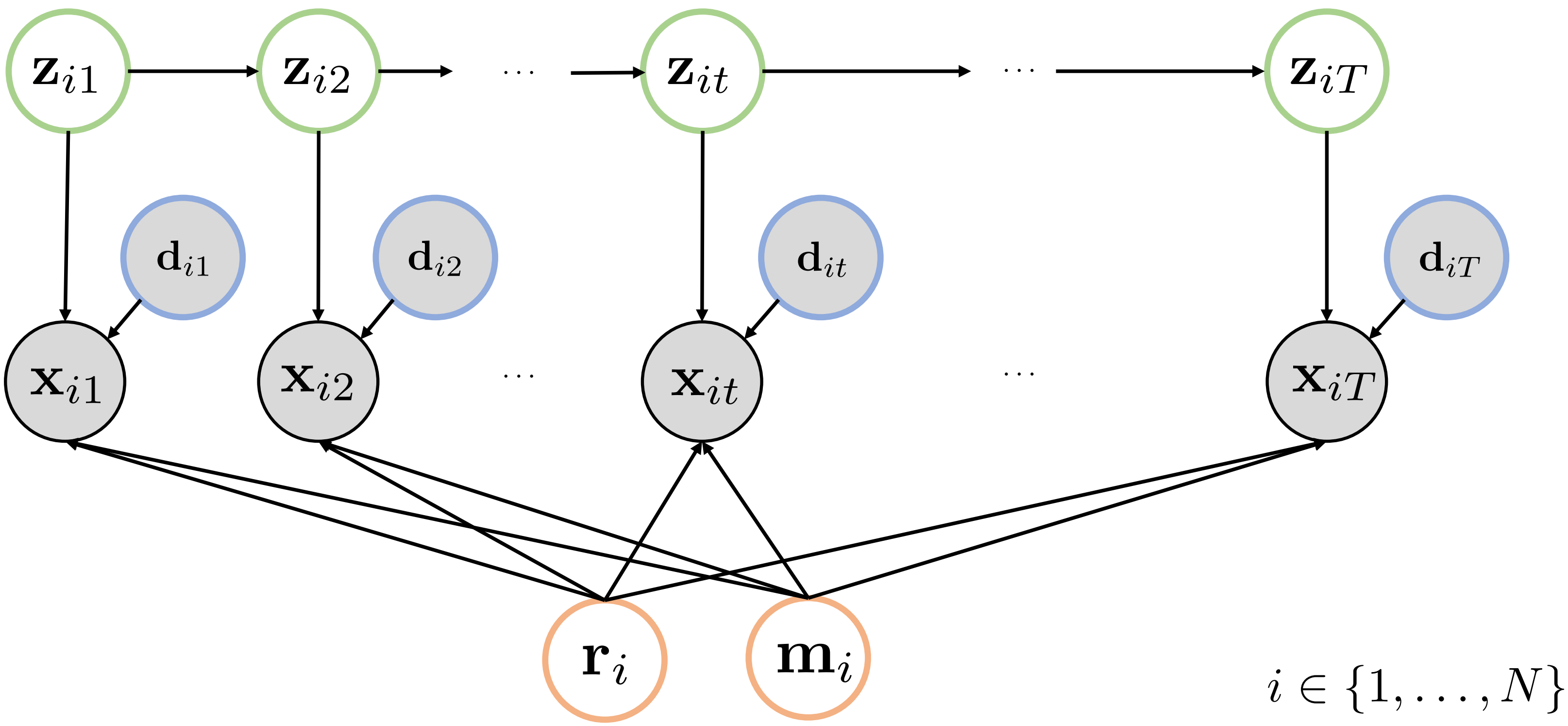


Why do we need good unsupervised models of sequential data?

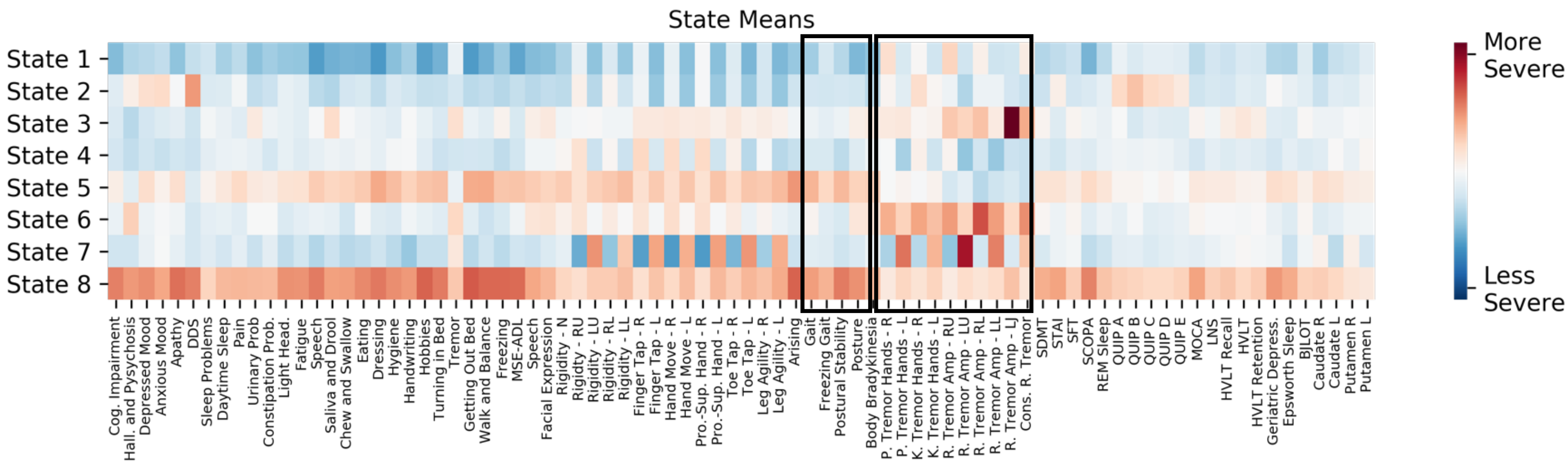
Dynamic Risk Prediction/Forecasting : Learn a representation of patient that is predictive of clinical outcomes in the future

Patient subtyping: Clustering patient trajectories to uncover subtypes corresponding to disease behaviors

Case study 1: Personalized I-O HMMs for disease progression modeling, Severson et al, MLHC 2020



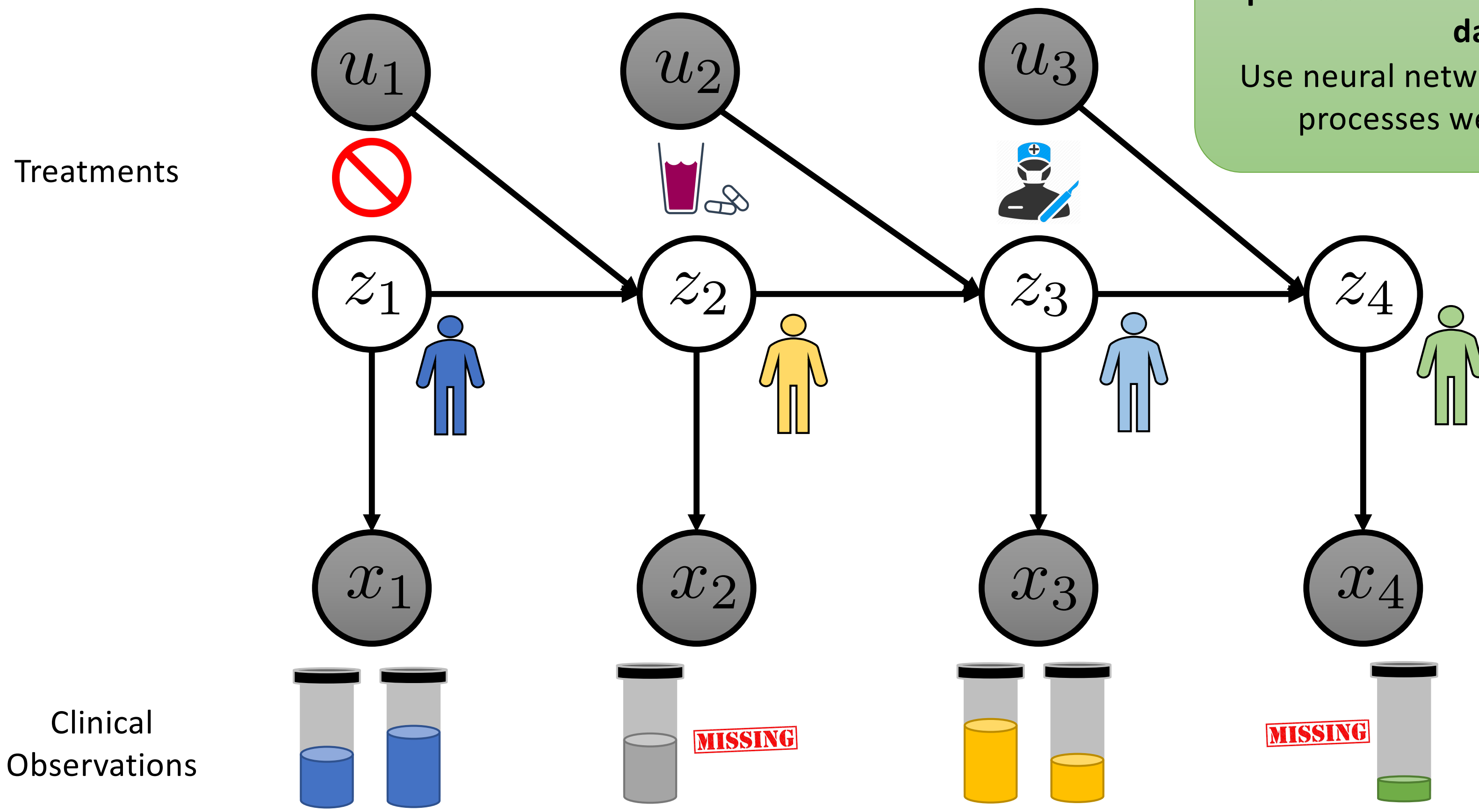
Inferred latent states across data dimensions



Unsupervised disease progression in a nutshell

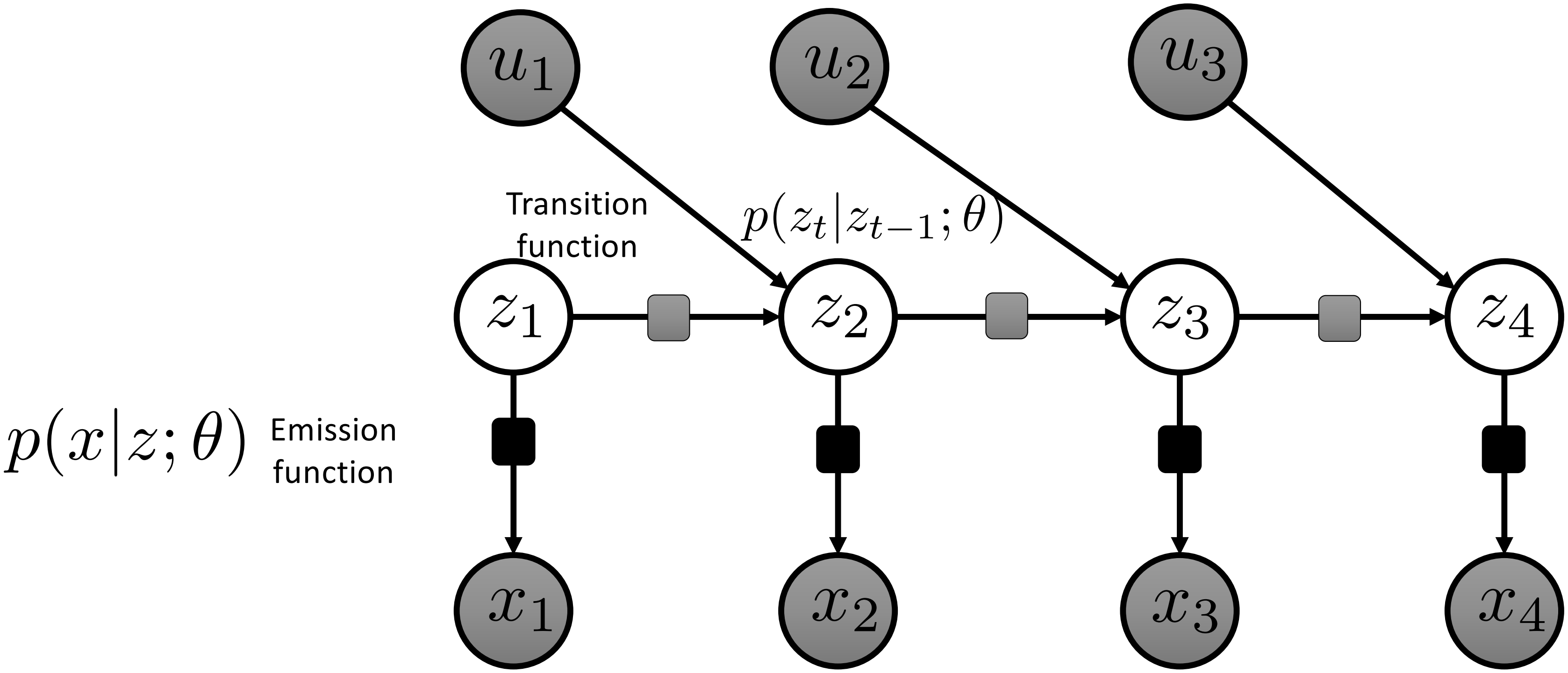
- Gather and collect all the time-varying data about patients
- Train a model to do unsupervised learning
- Using the model:
  - Introspect and attempt to interpret the model parameters
  - Use the model to forecast data into the future

State Space Models



Poor knowledge of data generating process for many kinds of clinical data:
Use neural networks as a proxy for processes we do not know

Deep Markov Models



Structured Inference Networks for Nonlinear State Space Models, **RGK,** US, DS, AAAI 2017

Unsupervised learning of nonlinear state space models

- *Previous work:*
  - Dual Extended Kalman Filters (Wan et a., 1996),
  - Particle filters (Schon et al., 2011),
  - Expectation Maximization (Briegel et al, 1999, Ghahramani et al, 1999),
  - Nonlinear dynamic factor analysis (Valpola, 2002)
- **Goals**:
  - Difficult to scale to high dimensional data, did not leverage modern hardware
  - Expensive test time inference

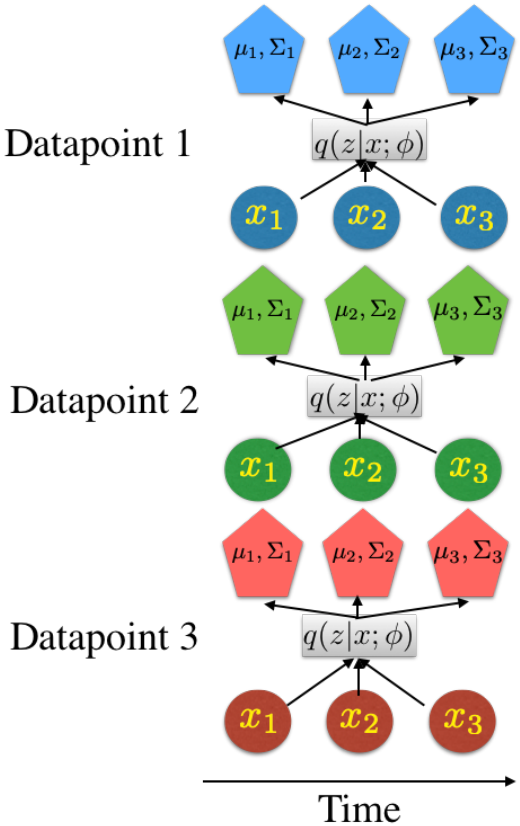
Technical challenge:
Variational learning via maximum likelihood

Loss function

$$\log p(\vec{x};\theta) = \log \int_z p(\vec{x},\vec{z};\theta) \geq \underbrace{\int_z q(\vec{z}|\vec{x};\phi)\log\frac{p(\vec{x},\vec{z};\theta)}{q(\vec{z}|\vec{x};\phi)}}_{\text{ELBO: } \mathcal{L}(\vec{x};\phi,\theta)}$$

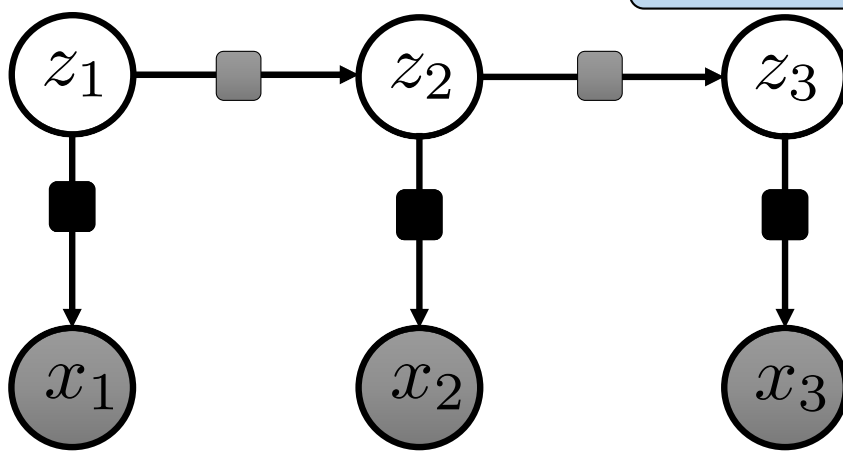
The variational distribution is over multiple different variables.

How should we design an inference network over multiple latent variables?



Key Idea
Mimic the factorization of the true posterior

$$p(\vec{z}|\vec{x}) = p(z_1, z_2, z_3|x_1, x_2, x_3) = p(z_1|x_{1:3})p(z_2|z_1, x_{1:3})p(z_3|z_1, z_2, x_{1:3})$$



$$z_2 \perp x_1|z_1$$

$$p(z_2|z_1, x_{1:3}) = p(z_2|z_1, x_{2:3})$$

$$z_3 \perp x_1, x_2, z_1|z_2$$

$$p(z_3|z_1, z_2, x_{1:3}) = p(z_3|z_2, x_3)$$

Factorization of the true posterior

$$p(\vec{z}|\vec{x}) = p(z_1, z_2, z_3|x_1, x_2, x_3) = p(z_1|x_{1:3})p(z_2|z_1, x_{1:3})p(z_3|z_1, z_2, x_{1:3})$$

$$p(\vec{z}|\vec{x}) = p(z_1|x_{1:3})p(z_2|z_1, x_{2:3})p(z_3|z_2, x_3)$$

Factorization of the variational distribution: $q(\vec{z}|\vec{x}) = q(z_1|x_{1:3})q(z_2|z_1, x_{2:3})q(z_3|z_2, x_3)$

According to the formula, at each time step we need:

a) previous latent state

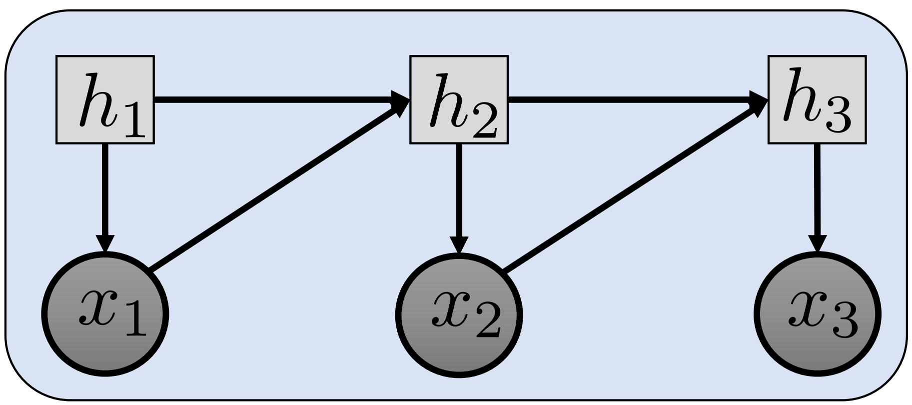
b) all future observations

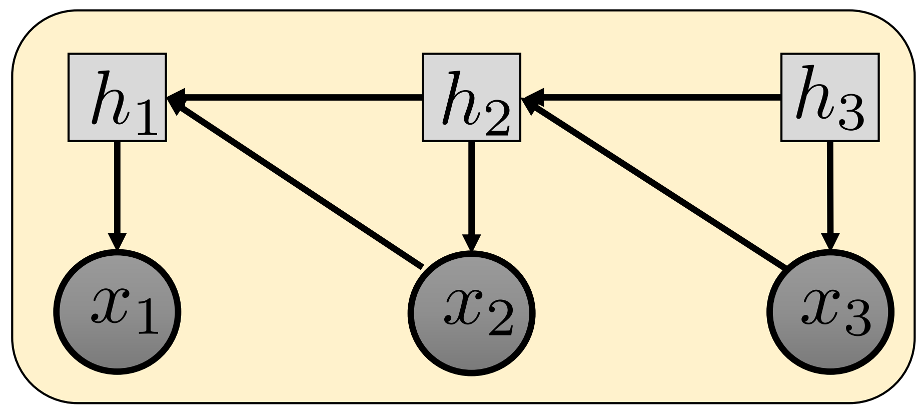
To build a representation of all future observations, we'll borrow a tool from Deep Learning
Recurrent Neural Networks

Recurrent Neural Networks

- Auto-regressive sequential models of data
- A forward-RNN models $p(x_1, x_2, x_3) = p(x_1|h_1)\hat{p}(h_2|h_1)p(x_2|h_2)\hat{p}(h_3|h_2)p(x_3|h_3)$
    - Each **hidden state** summarizes the variables in the **past**

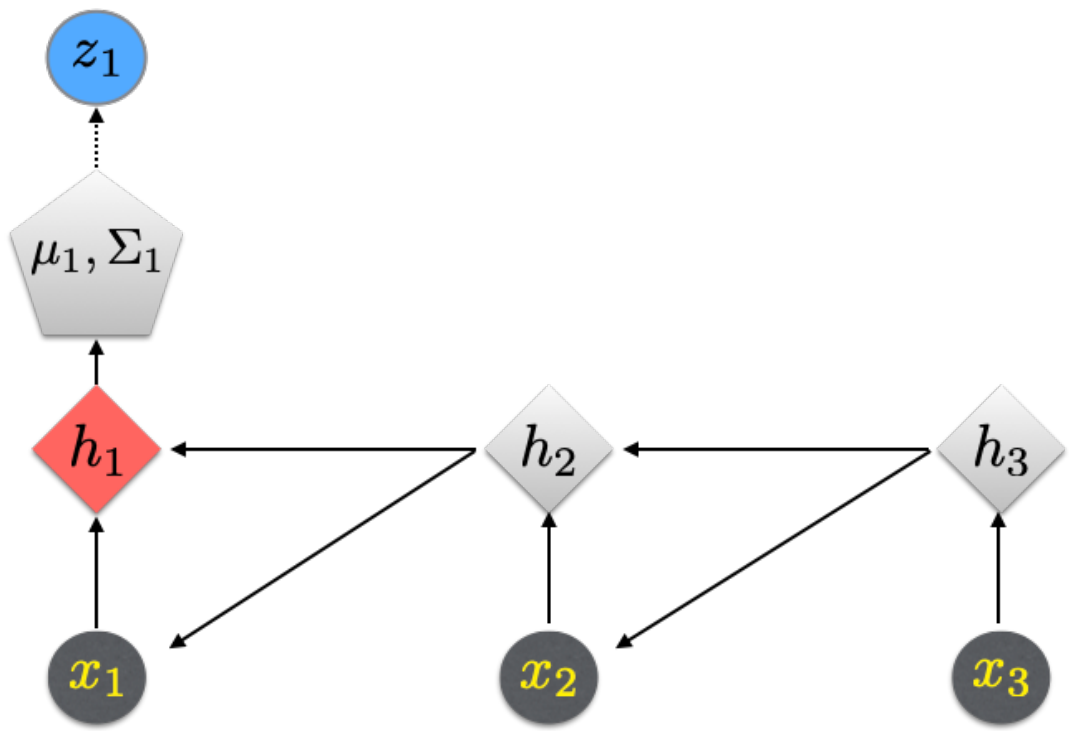
Key Idea: By running an RNN backward, we can use it to summarize the variables in the **future**





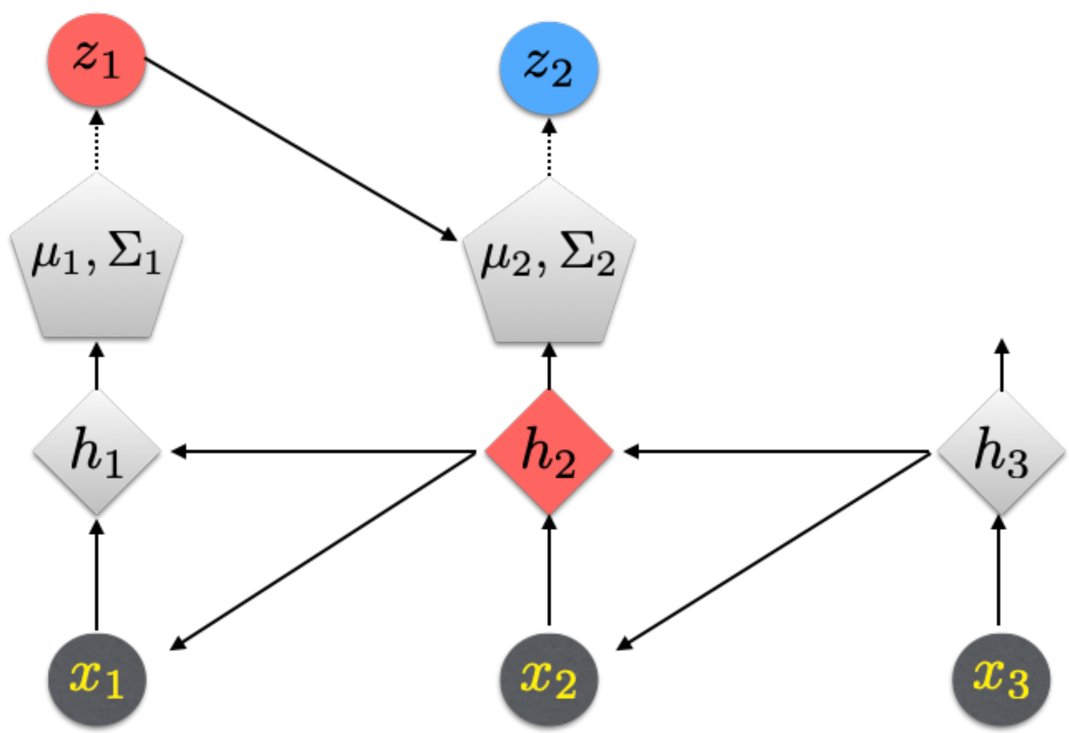
Structured Inference Network

$$q(\vec{z}|\vec{x}) = q(z_1|x_1, x_2, x_3)q(z_2|z_1, x_2, x_3)q(z_3|z_2, x_3)$$



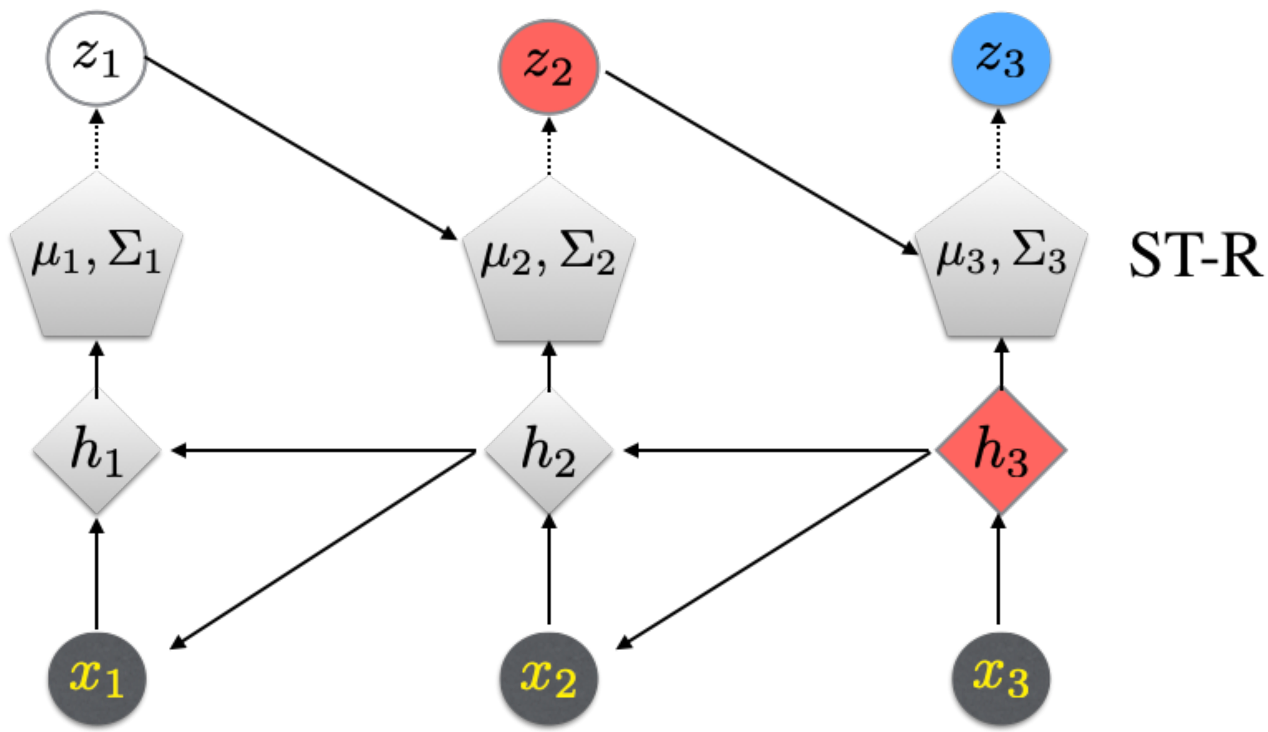
Structured Inference Network

$$q(\vec{z}|\vec{x}) = q(z_1|x_1, x_2, x_3)q(z_2|z_1, x_2, x_3)q(z_3|z_2, x_3)$$



Structured Inference Network

$$q(\vec{z}|\vec{x}) = q(z_1|x_1, x_2, x_3)q(z_2|z_1, x_2, x_3)q(z_3|z_2, x_3)$$



Mini-Recap of Structured Inference Networks

Question: How to select among a large set of factorizations for the variational distribution

Idea 1: Use the factorization of the true posterior

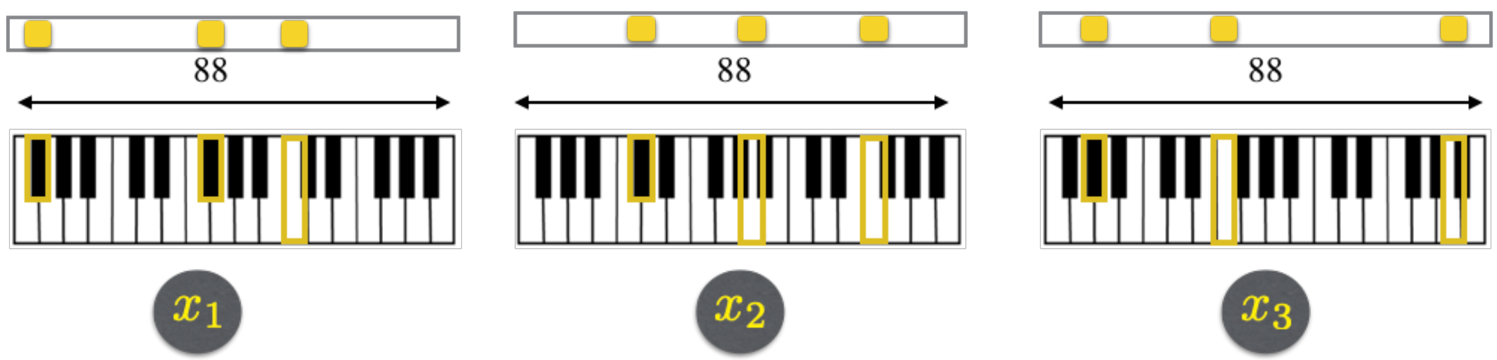
Idea 2: Use conditional independence statements in the model to simplify the factorization

Idea 3: Give a practical model by combining insights with advances in deep learning

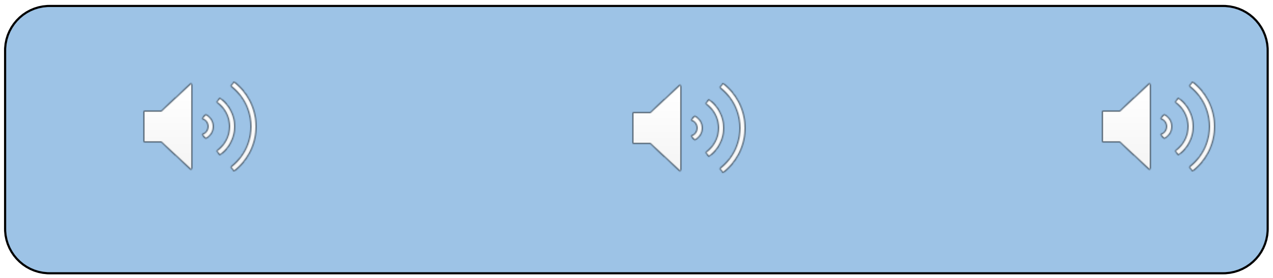
Evaluation of unsupervised time-series models

- **Metrics:**
  - (Upper bounds on) negative held-out log likelihood

Polyphonic Music Dataset (Boulanger-Lewandowski et al., 2012)



Use the model the generate music!



Captures some short- and long-term patterns.

Model the progression of disease

Forecast patient biomarkers

What can we do with Deep Markov Models?

Sequential treatment effects

Generate new examples of complex data

Case Study 1: Disease progression of diabetic patients

Dataset: Clinical data from a major insurance claims provider

Dataset size: 5000 diabetic patients. Each patient's data (over 4 years) is grouped into three month intervals, yielding a sequence of length 18.

Experiment: Vary the complexity of the transition and emission function in the Deep Markov Model

Observations

- 48 binary observations at each time step
- A1c level (a protein for which a high level indicates that the patient is diabetic)
- Glucose (blood sugar)
- Demographics: Age, Gender
- ICD-9 diagnosis codes for co-morbidities

Modeling diabetic patients

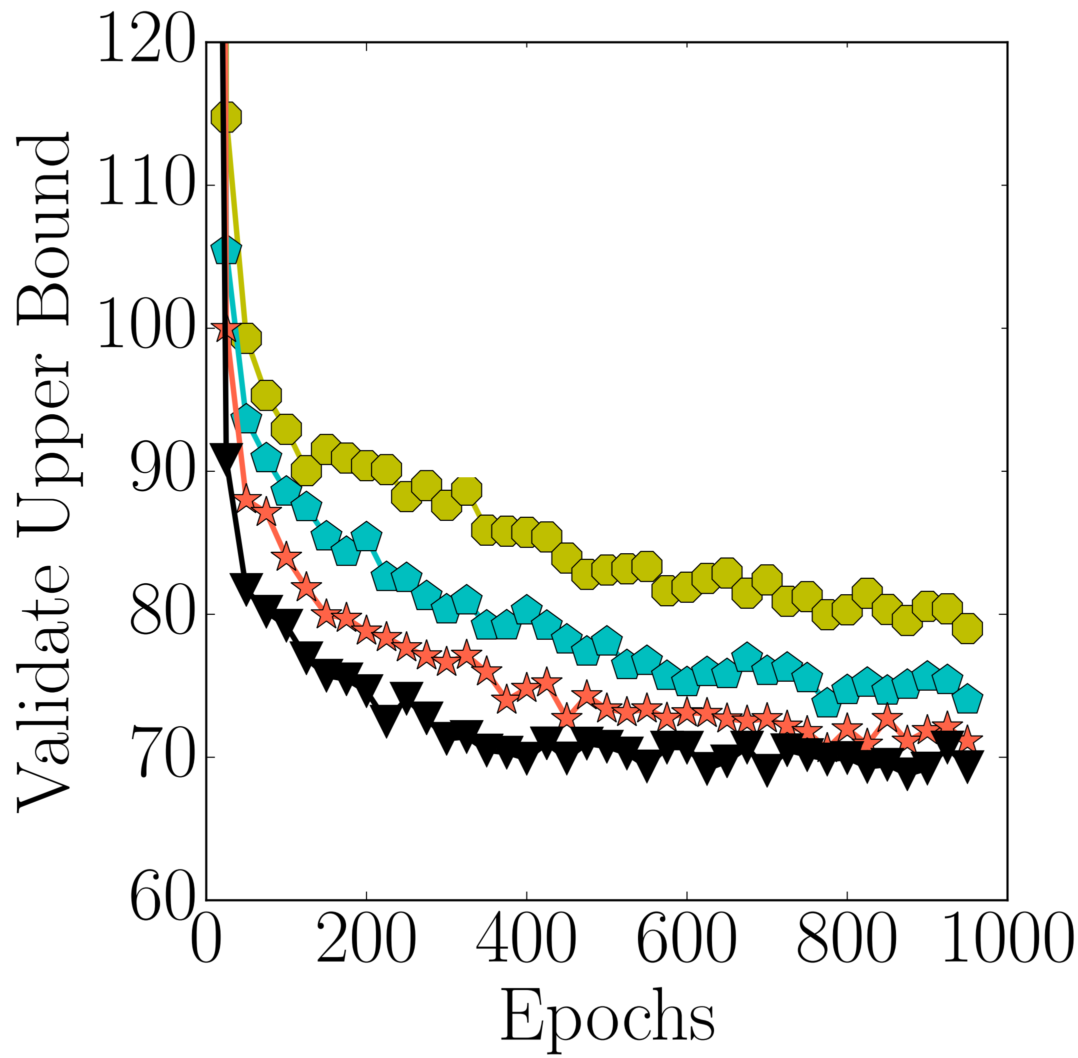
Metrics:

(Upper bounds on) negative held-out log likelihood

Linear State Space Mode

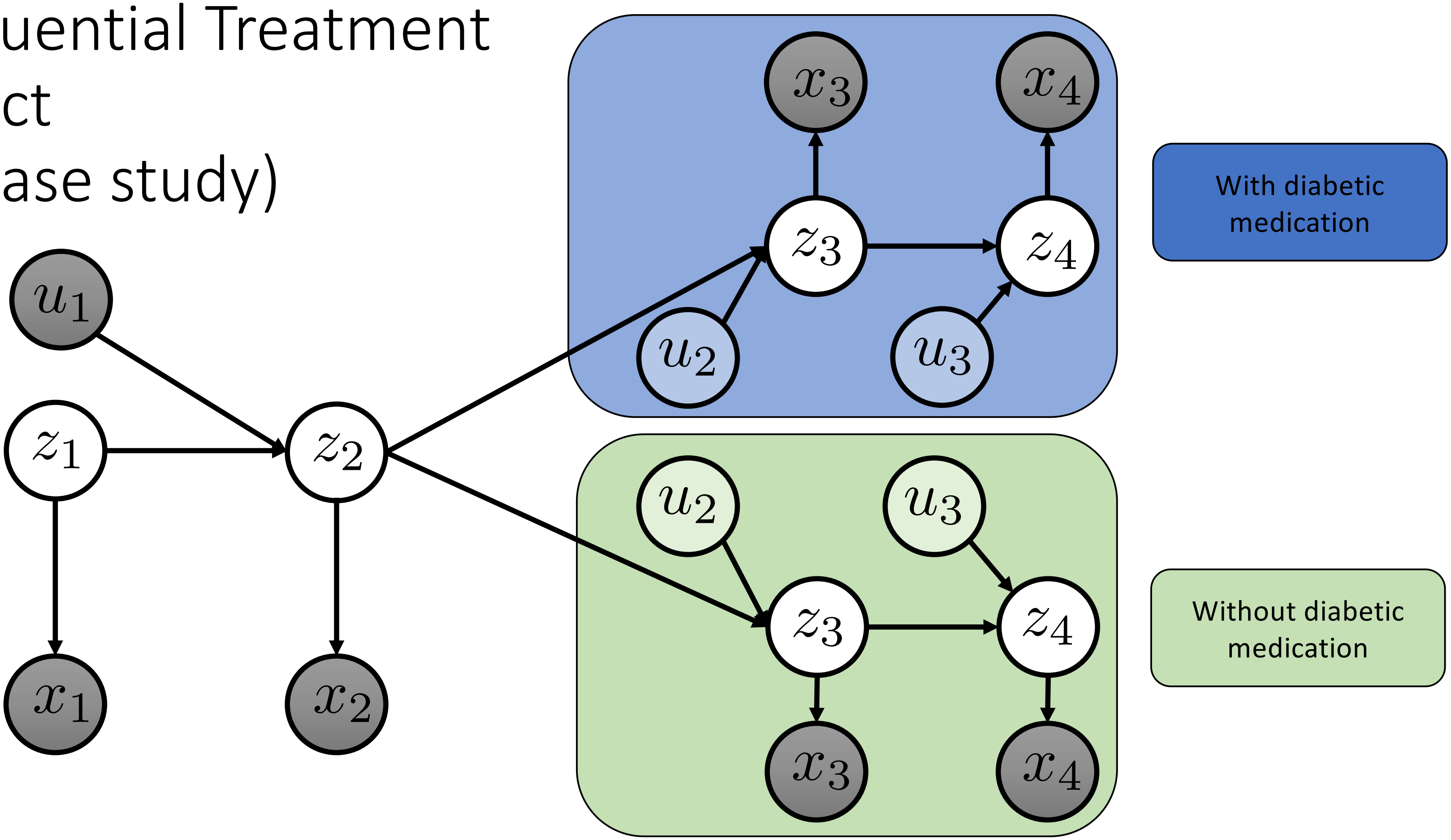
Deep Markov Model

There is benefit here, to using a nonlinear functions, i.e. Deep Markov Model, to model the sequence of clinical observations



Case Study 2: Treatment effect

Sequential Treatment Effect
(A case study)



Proof of concept
Sequential treatment effect

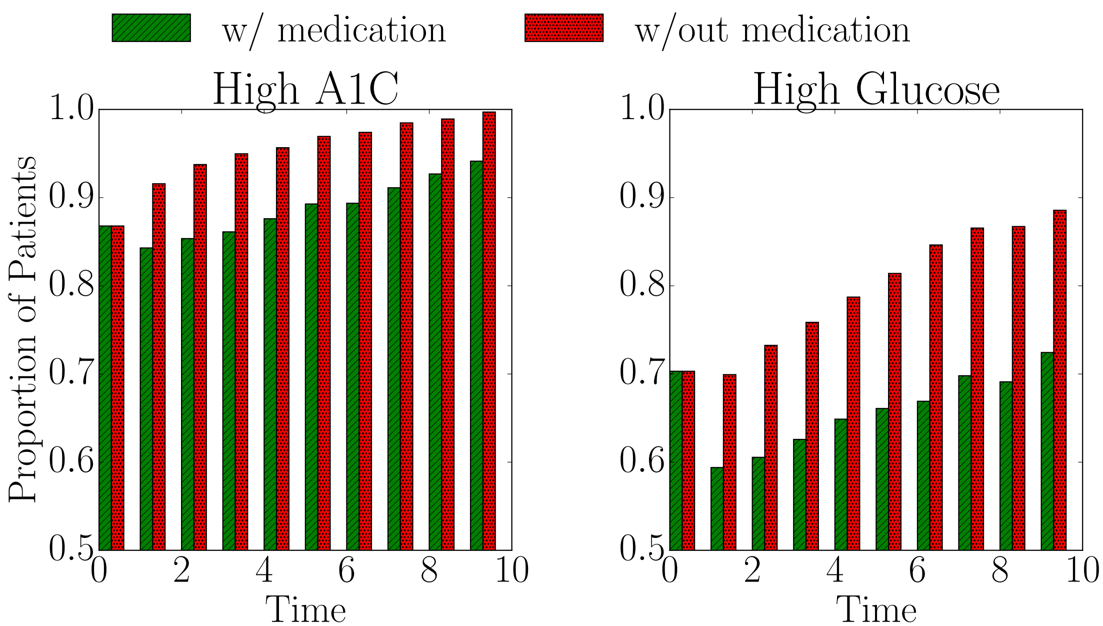
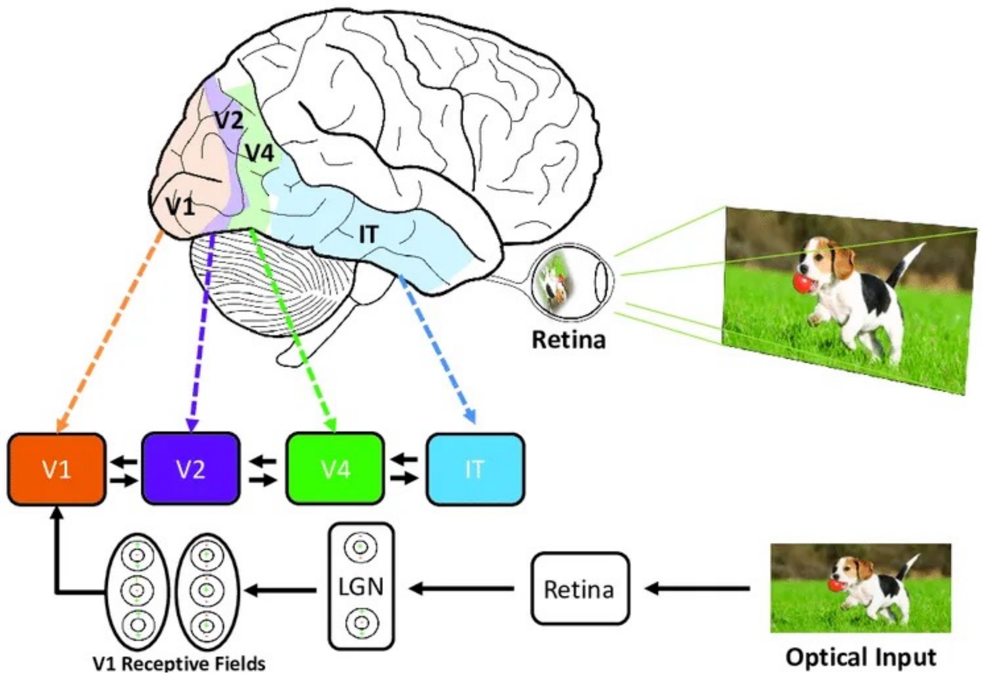


Figure: Comparing glucose levels from simulating with the model under the factual and the counterfactual

Deep Markov Models can be a powerful tool in estimators of sequential treatment effects

Case Study 3: Inductive Biases for Treatment effect



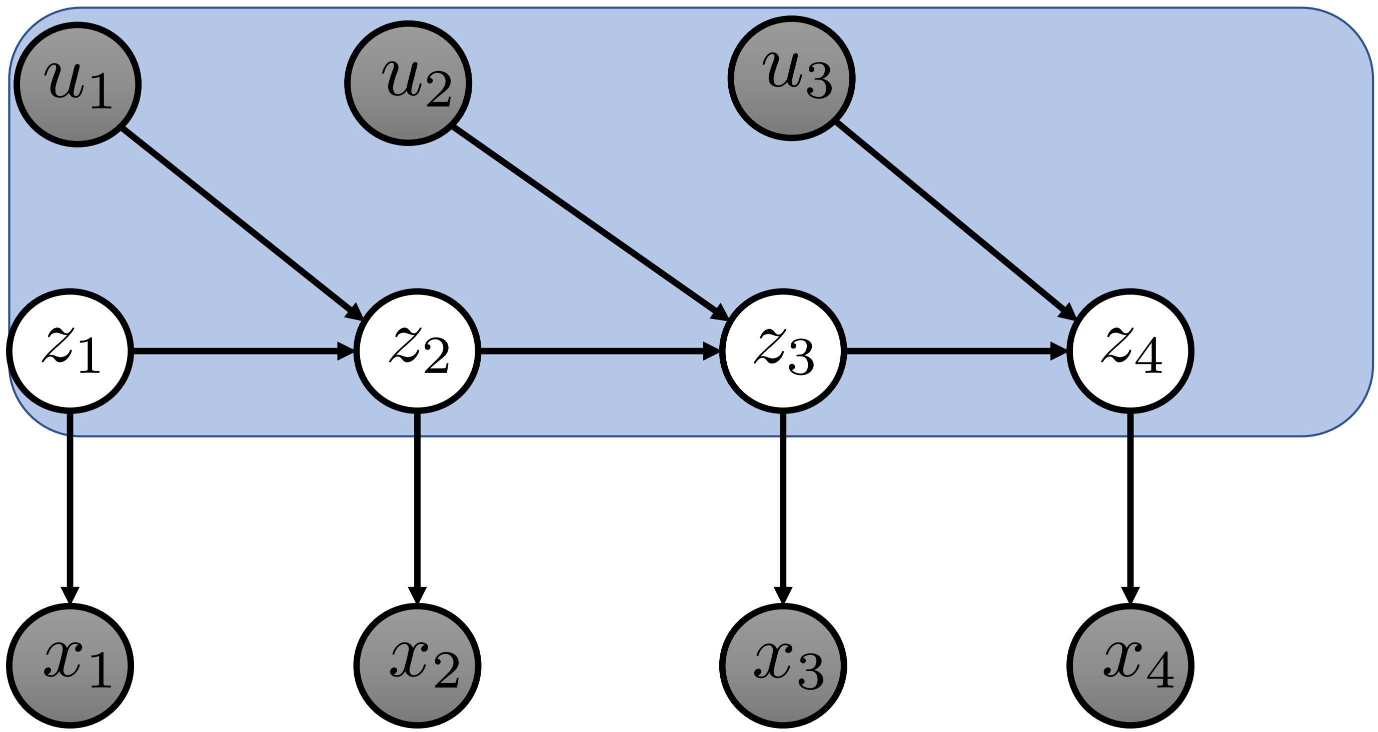
Where can we draw inspiration from when building Deep Markov Models?

Source: https://blog.knoldus.com/machinex-starts-with-why-ft-convolutional-neural-network/amp/

Inductive biases for treatment effect

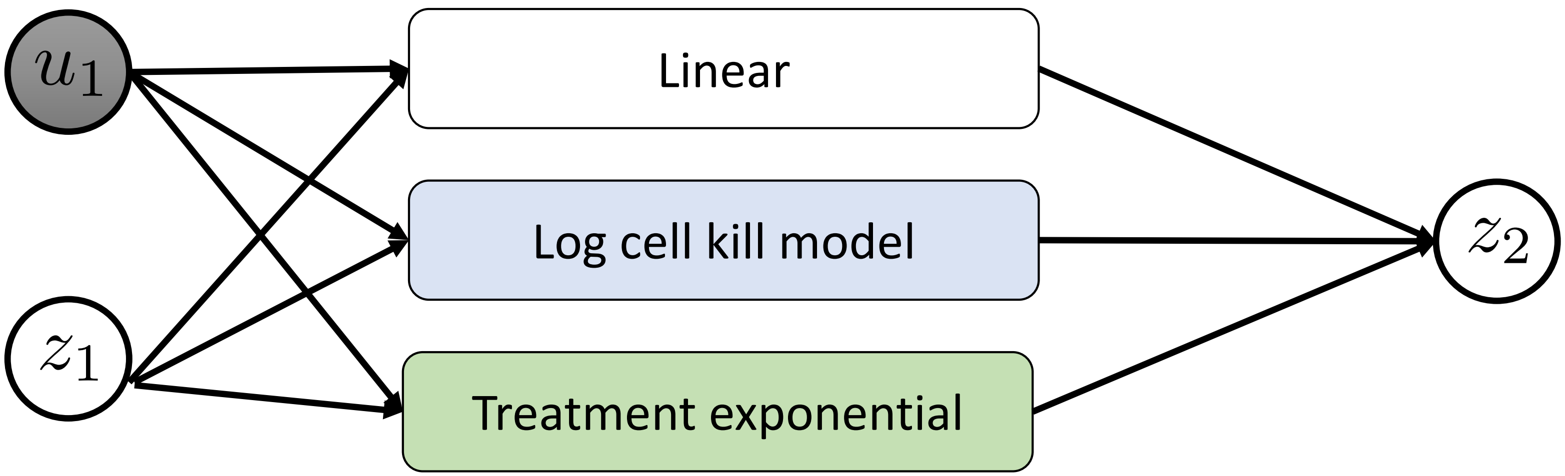
$$p(z_t | z_{t-1}, u_{t-1}; \theta)$$

Developed new neural network architectures inspired by the pharmacokinetic and pharmacodynamic modeling literature



Inductive Biases for the Transition Function

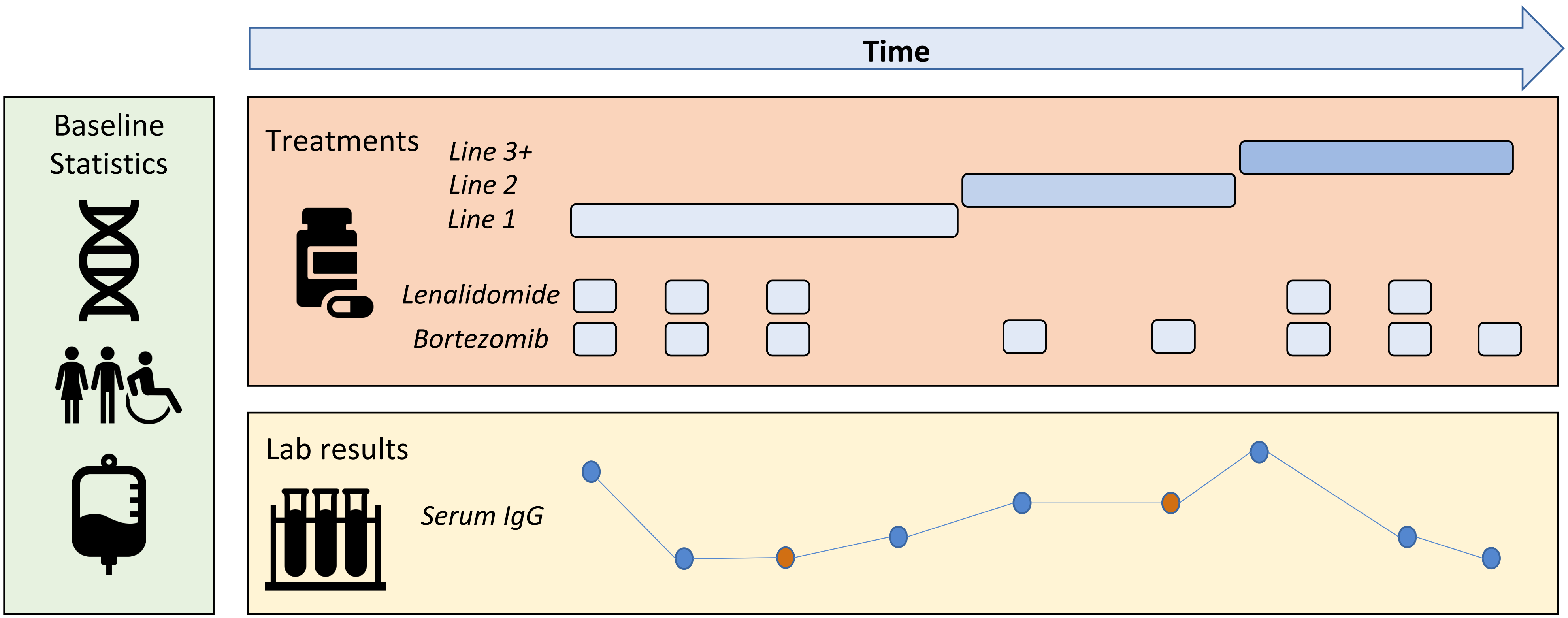
$$p(z_t | z_{t-1}, u_{t-1}; \theta)$$



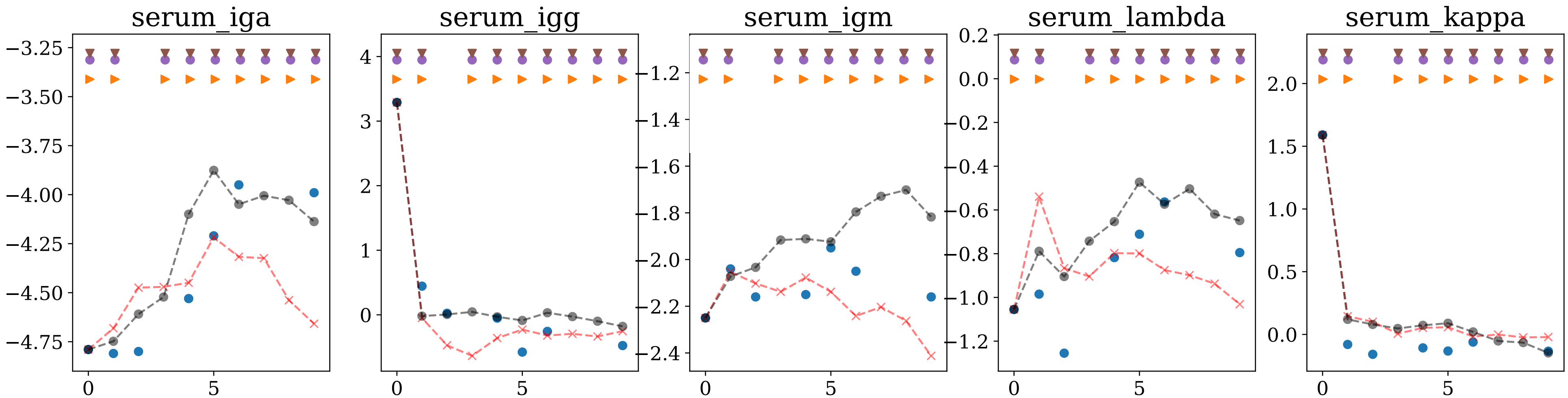
Cancer log-kill revisited, *Norton, 2014*

A Bayesian nonparametric approach for estimating individualized treatment-response curves, Xu et. al 2016

MULTIPLE MYELOMA
Research Foundation



Forecasting



Ground truth

PK/PD DMM

Linear State Space Model

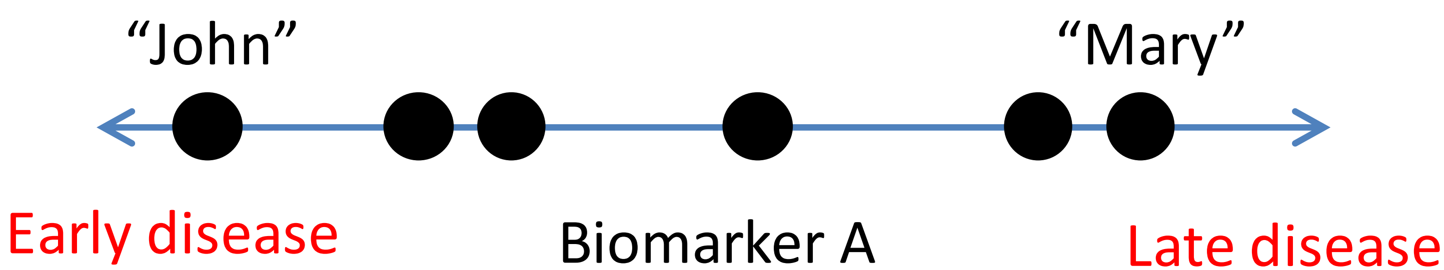
PK/PD DMM better at forecasting patient biomarkers

Supervised learning for disease progression

- Did not cover (but useful for further reading):
  - Supervised techniques for modeling the progression of diseases
  - **Modeling Disease Progression via Fused Sparse Group Lasso, Zhou et. Al, KDD 2012**
- **Key idea:**
  - Predict disease status in 6, 12, 24, 36 months with a single model (multi-task learning)
  - Have different weights for different time-horizons
  - The tasks are related so tie the weights together via a group-lasso penalty
  - Look at weights to assess the features most predictive of disease state

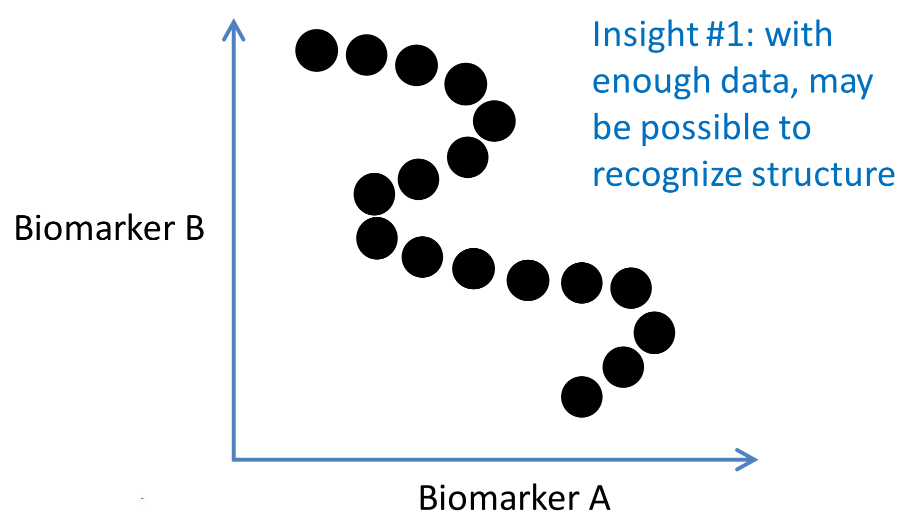
Cross-sectional data

- Thus far we've discussed models built on disease cohorts (many patients, many time-points)
- Only 1 time-point per patient (but potentially many patients)
- Goal is to construct a time-line that is shared by all or groups of patients

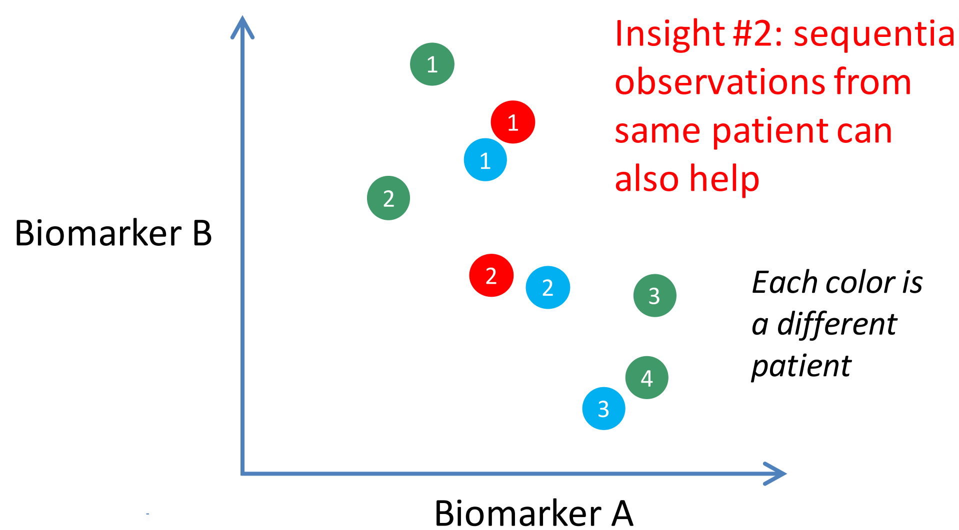


Slide credits: David Sontag

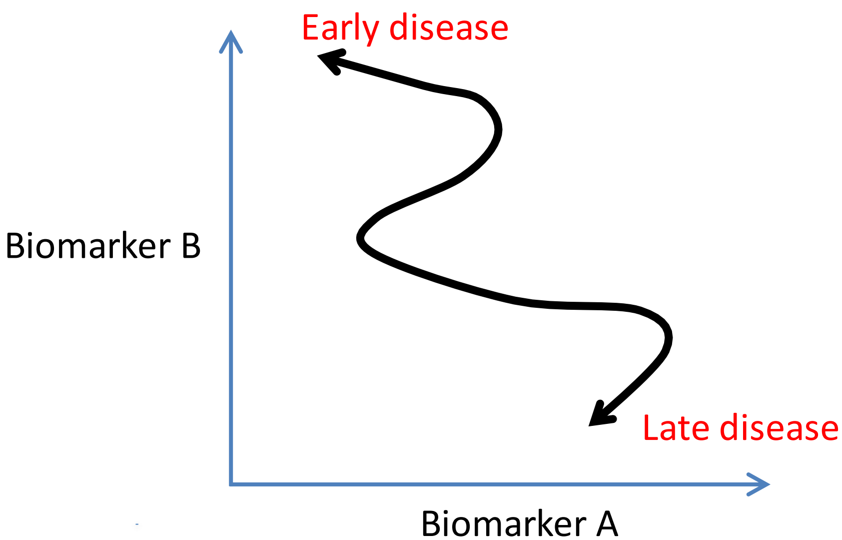
Insights to identify structure

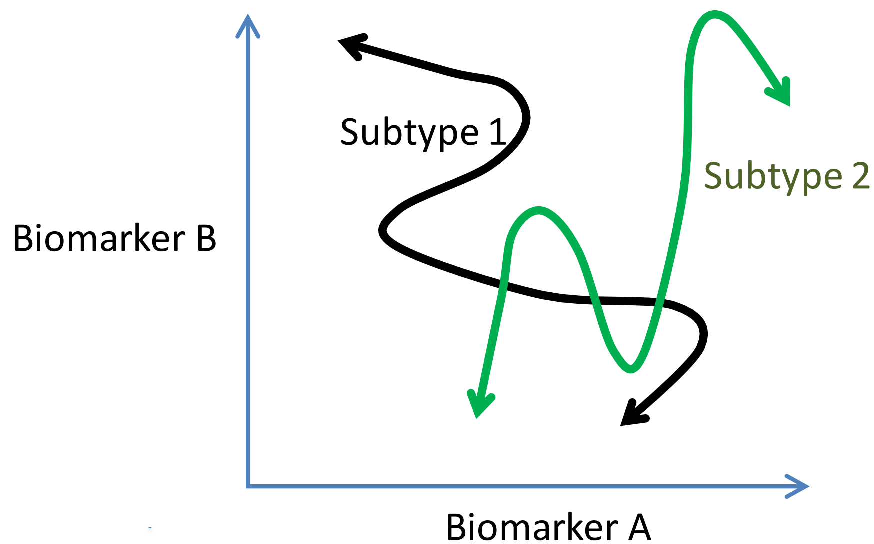


[Bendall et al., Cell 2014 (human B cell development)]



Goals with cross-sectional data





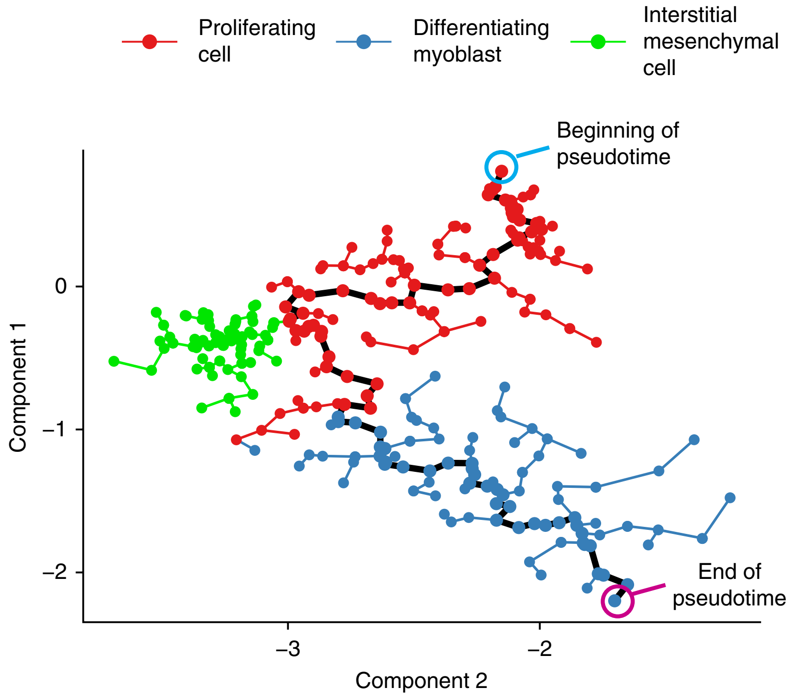
Creating trees in time

Reduce dimensionality of features via PCA/ICA

Build minimum spanning tree while treating lower dimensional representations as nodes

Use topological sort to identify time-axis

MST-based approach (Monocle)



[Trapnell et al., *Nature Biotechnology*, 2014]

Subtype and Stage Inference (SuStaIn)

- Generative model for a data point:
  - Sample subtype *c* ~ Categorical(f_1, …, f_C)
  - Sample stage *t* ~ Categorical(uniform)
  - For each biomarker *i*, sample $x_i \sim \mathcal{N}(g_{c,i}(t), \sigma_i)$
- Means are enforced to be monotonically increasing and piece-wise linear:

$$g(t) = \begin{cases} \frac{z_1}{t_{E_{z_1}}} t, 0 < t \leq t_{E_{z_1}} \\ z_1 + \frac{z_2 - z_1}{t_{E_{z_2}} - t_{E_{z_1}}}\left(t - t_{E_{z_1}}\right), t_{E_{z_1}} < t \leq t_{E_{z_2}} \\ \vdots \\ z_{R-1} + \frac{z_R - z_{R-1}}{t_{E_{z_R}} - t_{E_{z_{R-1}}}}\left(t - t_{E_{z_{R-1}}}\right), t_{E_{z_{R-1}}} < t \leq t_{E_{z_R}} \\ z_R + \frac{z_{max} - z_R}{1 - t_{E_{z_R}}}\left(t - t_{E_{z_R}}\right), t_{E_{z_R}} < t \leq 1 \end{cases}.$$

Explicitly incorporate variation due to sub-type and stage into a probabilistic model

Shown here for one choice of *c,i* – no parameter sharing across biomarkers or subtypes

[Young et al., *Brain* 2014; Young et al., *Nature Communications* 2018]

Questions?