# Deep Kalman Filters


Rahul G. Krishnan, Uri Shalit, David Sontag

Courant Institute of Mathematical Sciences, New York University


## Motivation

- Time-series modeling is a ubiquitous task across many domains. We aim to create a powerful non-linear latent variable model of time-series.
- Patient records are a time series of diagnoses, lab tests, surgical procedures and drug prescriptions that represent observations of underlying conditions.
- What is the best treatment course for a patient? Which patients are similar to a given patient? Which policy is the most cost effective for a specific population? The wide availability of Electronic Health Records (EHR) gives machine learning the potential for addressing these questions.

## Background

**Kalman Filters** Generative time series models are characterized by their **action transition** functions and their **emission** functions. Classic Kalman filters use linear functions for both:

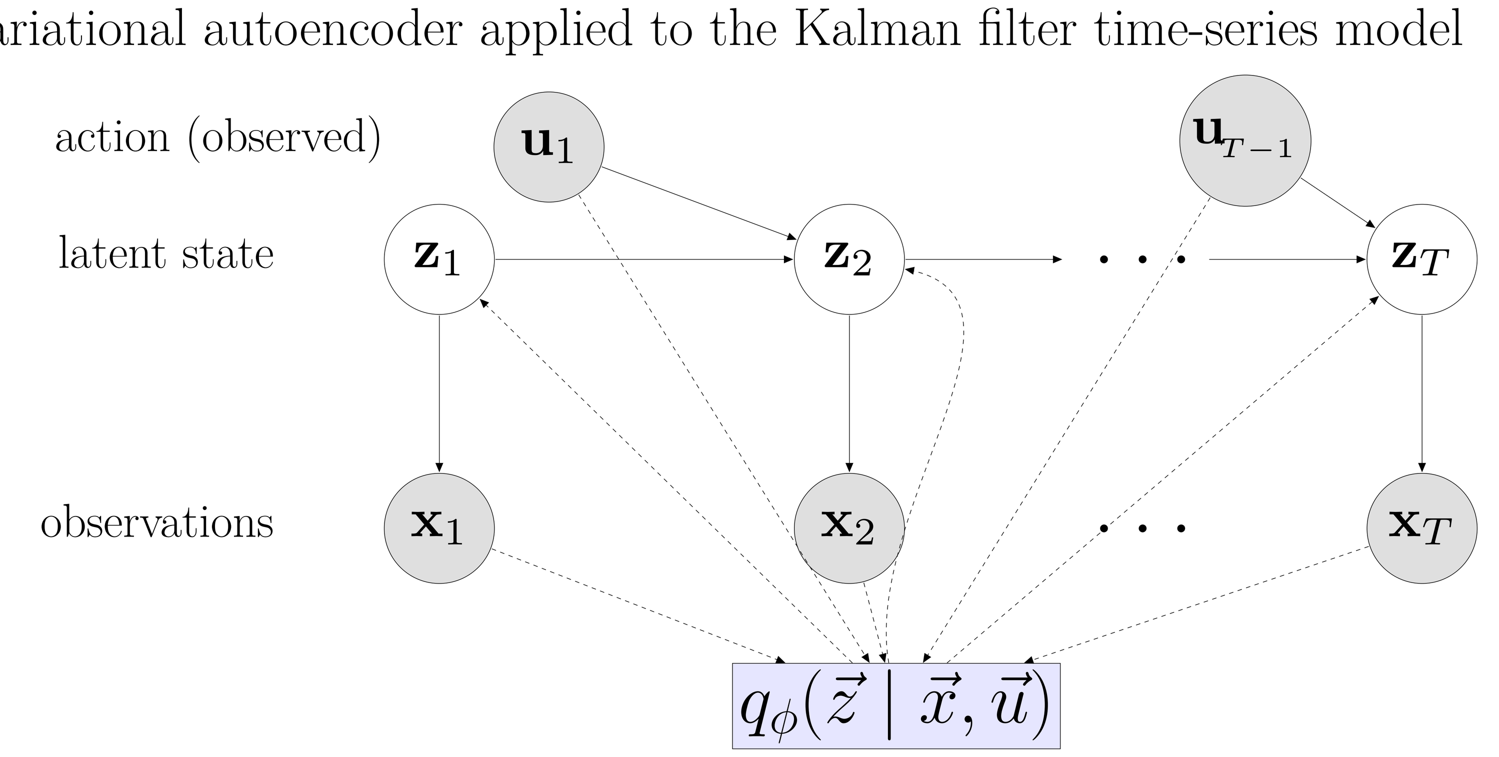
$$z_t \sim \mathcal{N}\left(G \cdot z_{t-1} + B \cdot u_{t-1}, \Sigma\right) \qquad \textit{(action transition)}$$
$$x_t \sim \mathcal{N}\left(F \cdot z_t, \Gamma\right) \qquad \textit{(emission)}$$

**Stochastic Backpropagation** Rezende *et al.* (2014) and Kingma & Welling (2013) proposed using a neural net as a variational autoencoder to optimize a lower-bound on the model log-likelihood:

**Counterfactual Inference** "*What would the patient's blood sugar level be had she taken a different medication?*"

## Probabilistic Model

Variational autoencoder applied to the Kalman filter time-series model




We apply **stochastic backpropagation** to learn a **non-linear Kalman filter**, and use the model to perform **counterfactual inference**.

$$z_t \sim \mathcal{N}\left(G_\alpha(z_{t-1}, u_{t-1}), S_\beta(z_{t-1}, u_{t-1})\right) \quad \textit{(action transition)}$$
$$x_t \sim p(x_t | z_t; F_\kappa(z_t)) \qquad \textit{(emission)}$$

$G_\alpha$, $S_\beta$ and $F_\kappa$ are functions parameterized by neural nets

### Proposition

For the graphical model we propose, the posterior factorizes as:

$$p(\vec{z}|\vec{x}, \vec{u}) = p(z_1|\vec{x}, \vec{u}) \prod_{t=2}^{T} p(z_t|z_{t-1}, x_t, \ldots, x_T, u_{t-1}, \ldots, u_{T-1})$$

## Approximating the Evidence Lower Bound

The recognition model prior $q_\phi$ is parameterized by a neural network $\phi$. We use the prior factorization:

$$q_\phi(\vec{z}|\vec{x}, \vec{u}) = \prod_{t=1}^{T} q_\phi(z_t|z_{t-1}, x_t, \ldots, x_T, \vec{u})$$

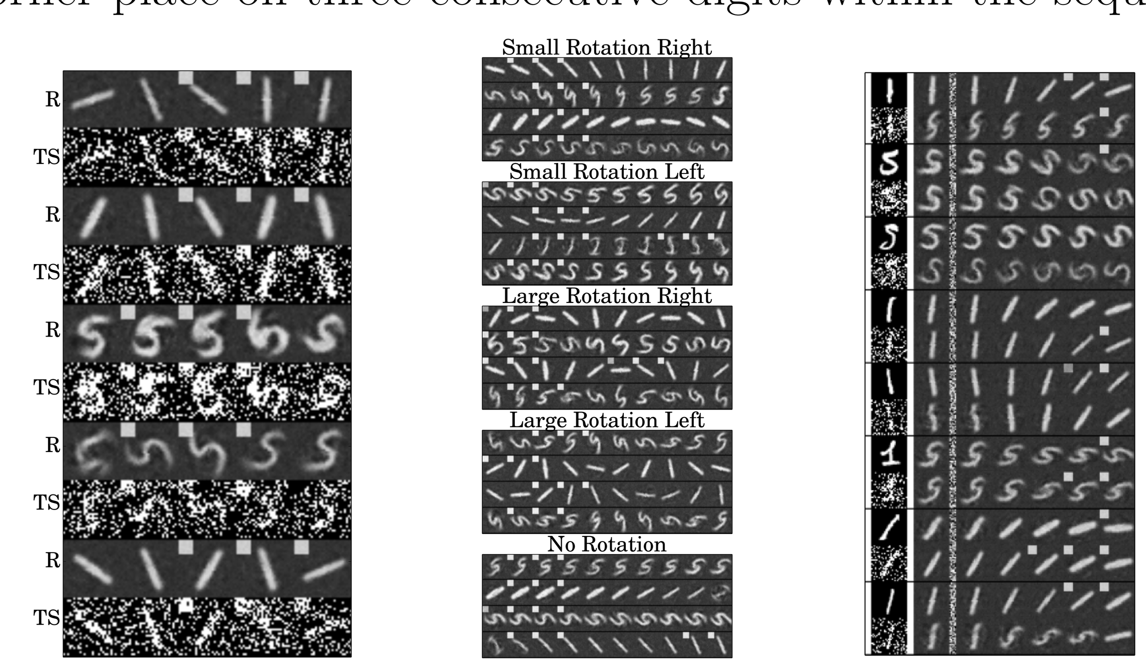
We maximize the following variational lower bound for training the generative and recognition models:

$$\log p_\theta(\vec{x}|\vec{u}) \geq \mathcal{L}(x; (\theta, \phi)) =$$
$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{q_\phi(z_t|\vec{x},\vec{u})} [\log p_\theta(x_t|z_t)] - \mathrm{KL}(q_\phi(z_1|\vec{x}, \vec{u}) || p_0(z_1))$$
$$- \sum_{t=2}^{T} \mathop{\mathbb{E}}_{q_\phi(z_{t-1}|\vec{x},\vec{u})} \left[\mathrm{KL}(q_\phi(z_t|z_{t-1}, \vec{x}, \vec{u}) || p_0(z_t|z_{t-1}, u_{t-1}))\right].$$

## Experiments

### Healing MNIST

Sequences of MNIST digits, where the action is a digit being rotated. We add random noise, and structured noise in the form of a block on the top-left corner place on three consecutive digits within the sequence.




(a) Reconstruction during training (b) Samples: different rotations (c) Inference on unseen digits

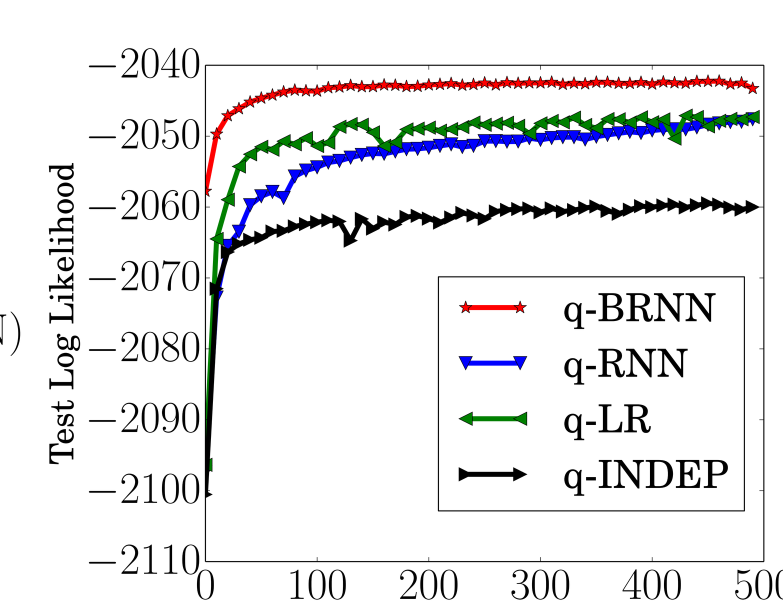
### Recognition Models

**q-INDEP**: (7 layer MLP)
$q(z_t|x_t, u_t)$

**q-LR**: (7 layer MLP)
$q(z_t|x_{t-1}, x_t, x_{t+1}, u_{t-1}, u_t, u_{t+1})$

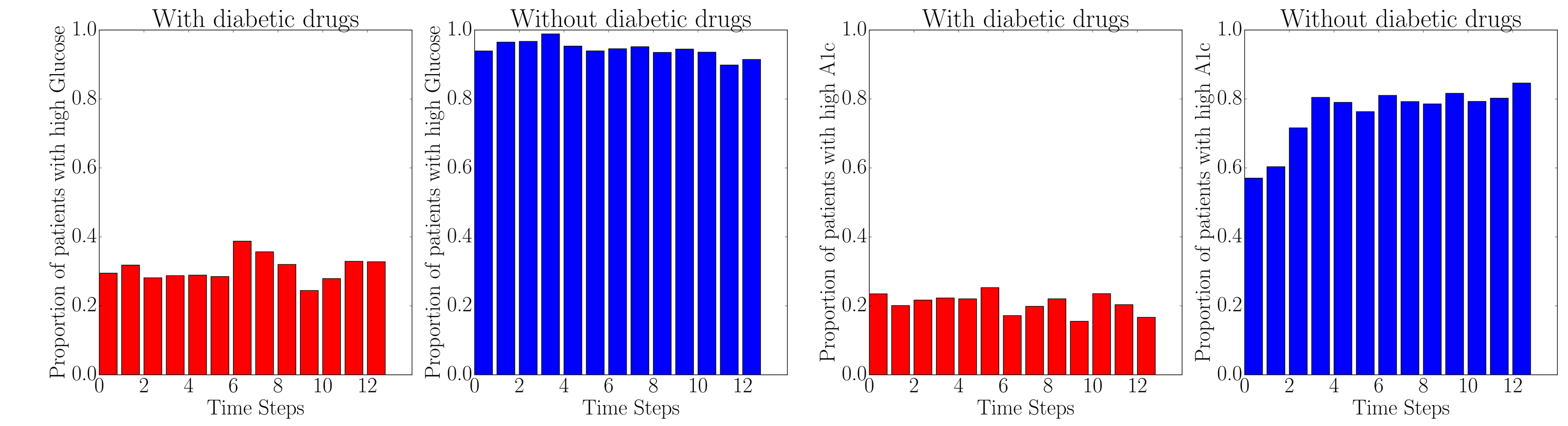
**q-RNN**: (2 layer MLP + 2 layer RNN)
$q(z_t|x_1, \ldots, x_t, u_1, \ldots u_t)$

**q-BRNN**:
(2 layer MLP + 2 layer bi-RNN)
$q(z_t|x_1, \ldots, x_T, u_1, \ldots, u_T)$



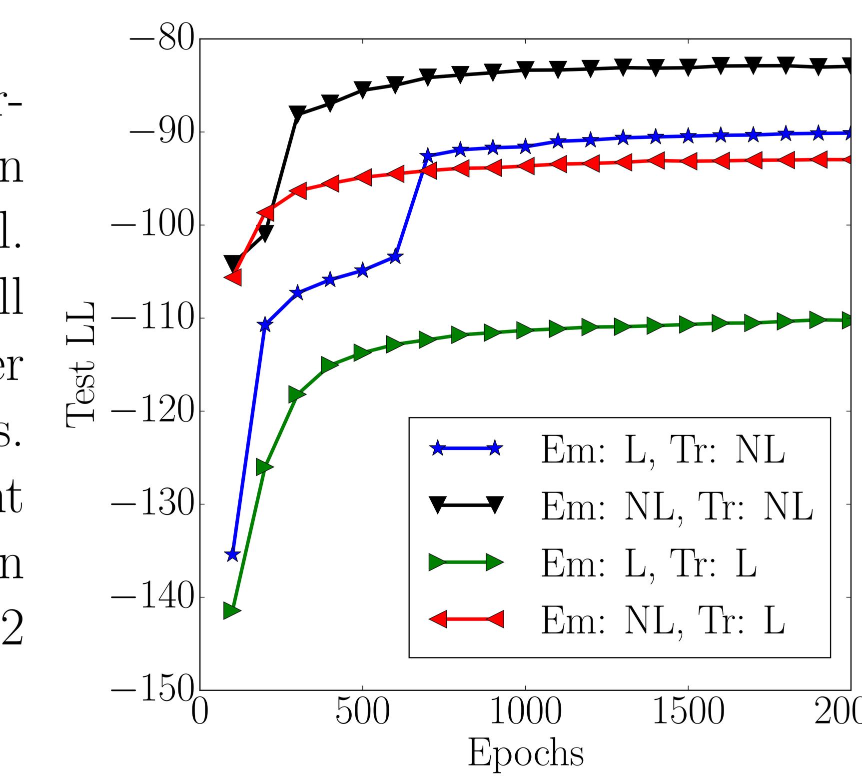

## Medical Data

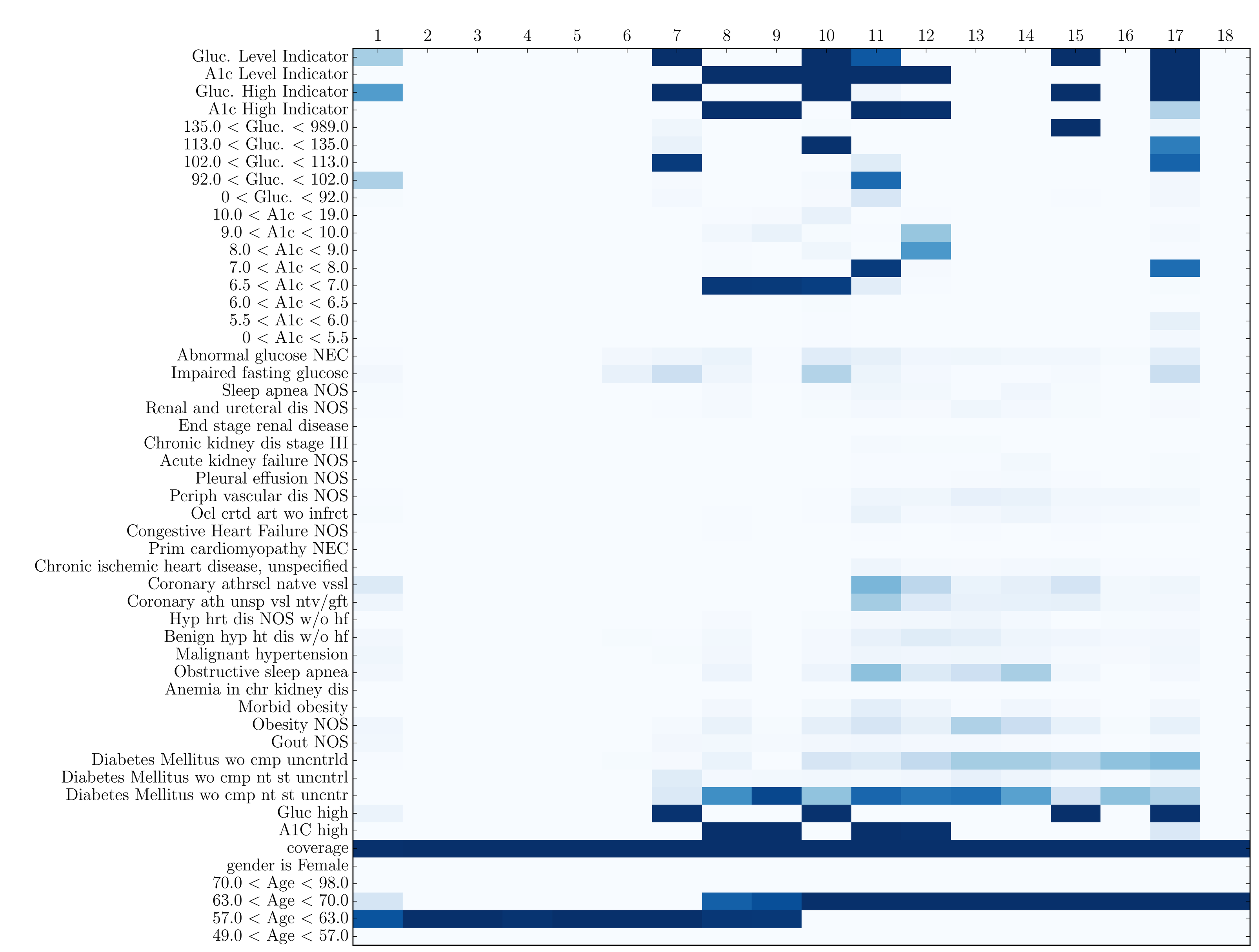
- Healthcare records data of 8000 diabetic and pre-diabetic patients.
- Infer future lab test values of A1c and glucose in counterfactual scenarios.
- Patient data: age, gender, and ICD-9 diagnoses code depicting comorbidities such as heart failure, kidney conditions or obesity.




Proportion of patients with high Glucose  Proportion of patients with high A1c

Test log-likelihood for different models. Em: emission model. Tr: transition model. L: linear. NL: non-linear. All non-linear models are 3-layer fully connected ReLU units. The dimension of the latent space $z_t$ is 30. Recognition model is a 2 layer MLP + 2 layer bi-directional RNN.




Sample Patient. The x-axis denotes time and the y-axis denotes the observations. The patient was sampled under no medication. The intensity of the colour denotes its value between zero and one.