MACHINE LEARNING FOR HEALTHCARE

6.S897, HST.S53

Lecture 7: Physiological and laboratory time-series

Prof. David Sontag

MIT EECS, CSAIL, IMES

Massachusetts Institute of Technology

Outline of today's class

1. **State space models for physiological condition modeling**
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion

Labs and physiological time-series

- Typical use cases:
  1. Risk stratification, e.g. predict clinical deterioration, or diagnosis
  2. Infer patient's past, current, or future health state from noisy observations, e.g. heart rate or glucose levels
- Approach taken varies depending on:
  - Is labeled data available?
  - Do we have a good mechanistic/statistical model?
  - How much training data is there?

Physiological time-series

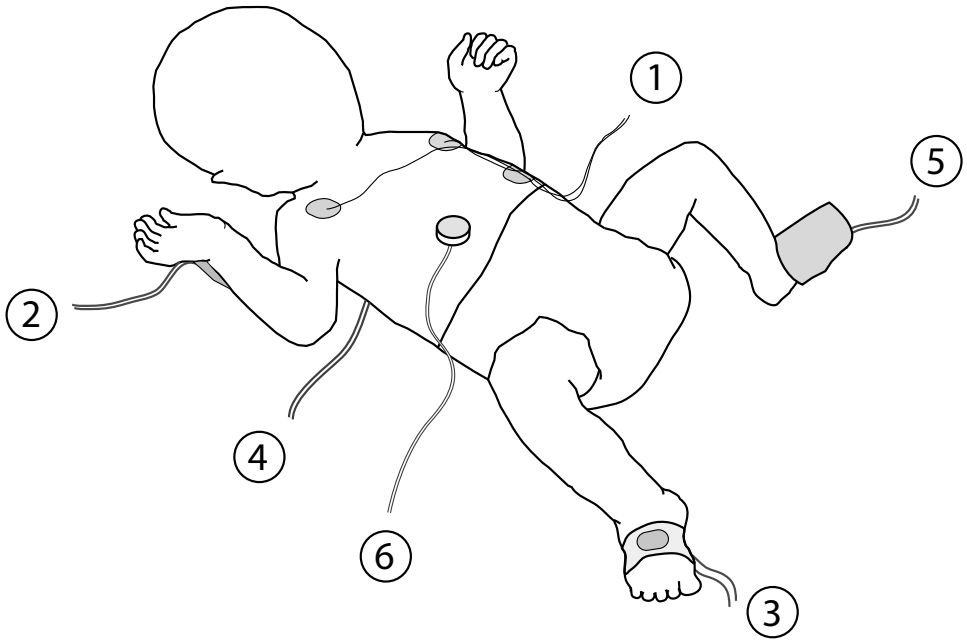
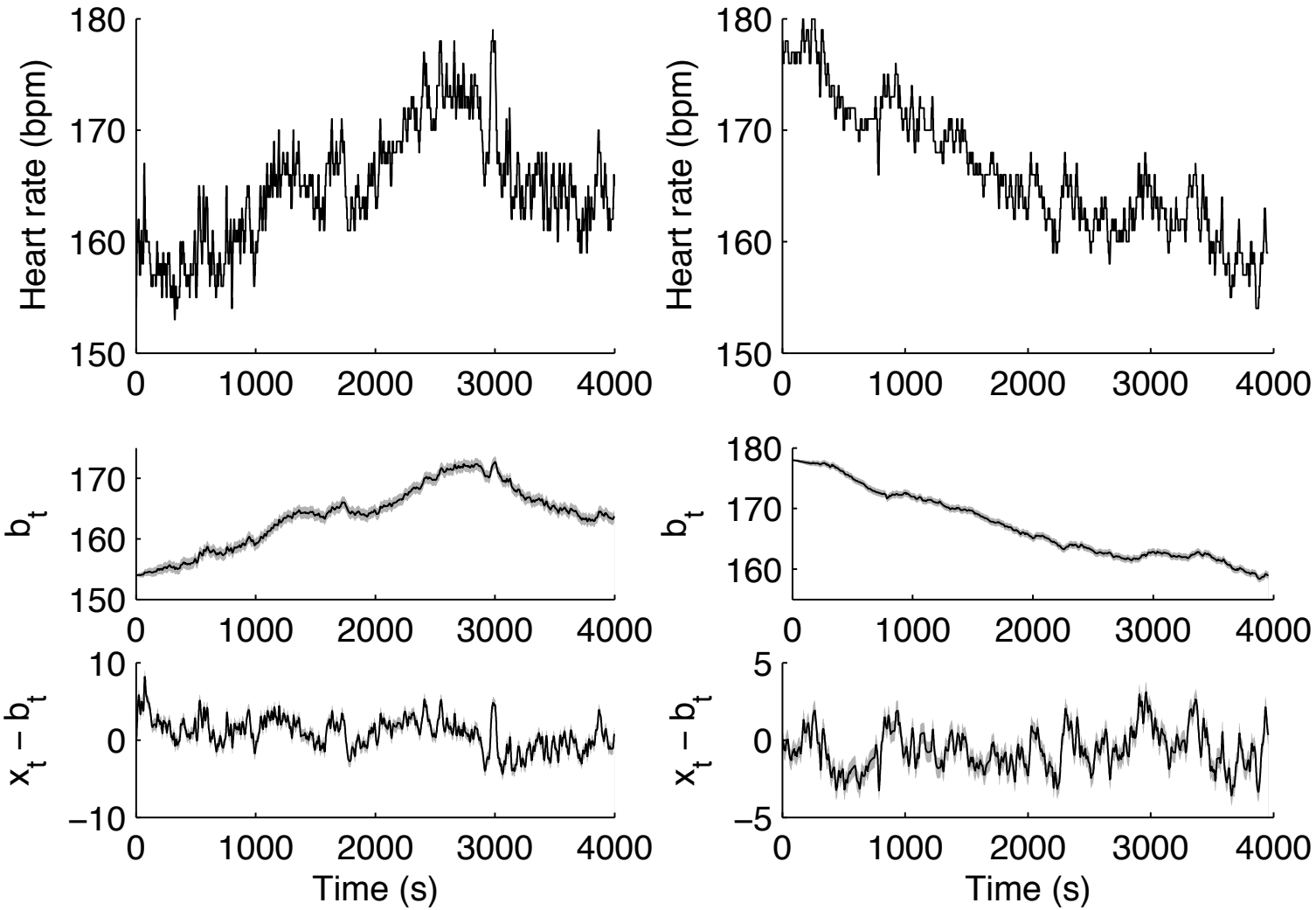


Fig. 4. Probes used to collect vital signs data from an infant in intensive care. 1) Three-lead ECG, 2) arterial line (connected to blood pressure transducer), 3) pulse oximeter, 4) core temperature probe (underneath shoulder blades), 5) peripheral temperature probe, 6) transcutaneous probe.

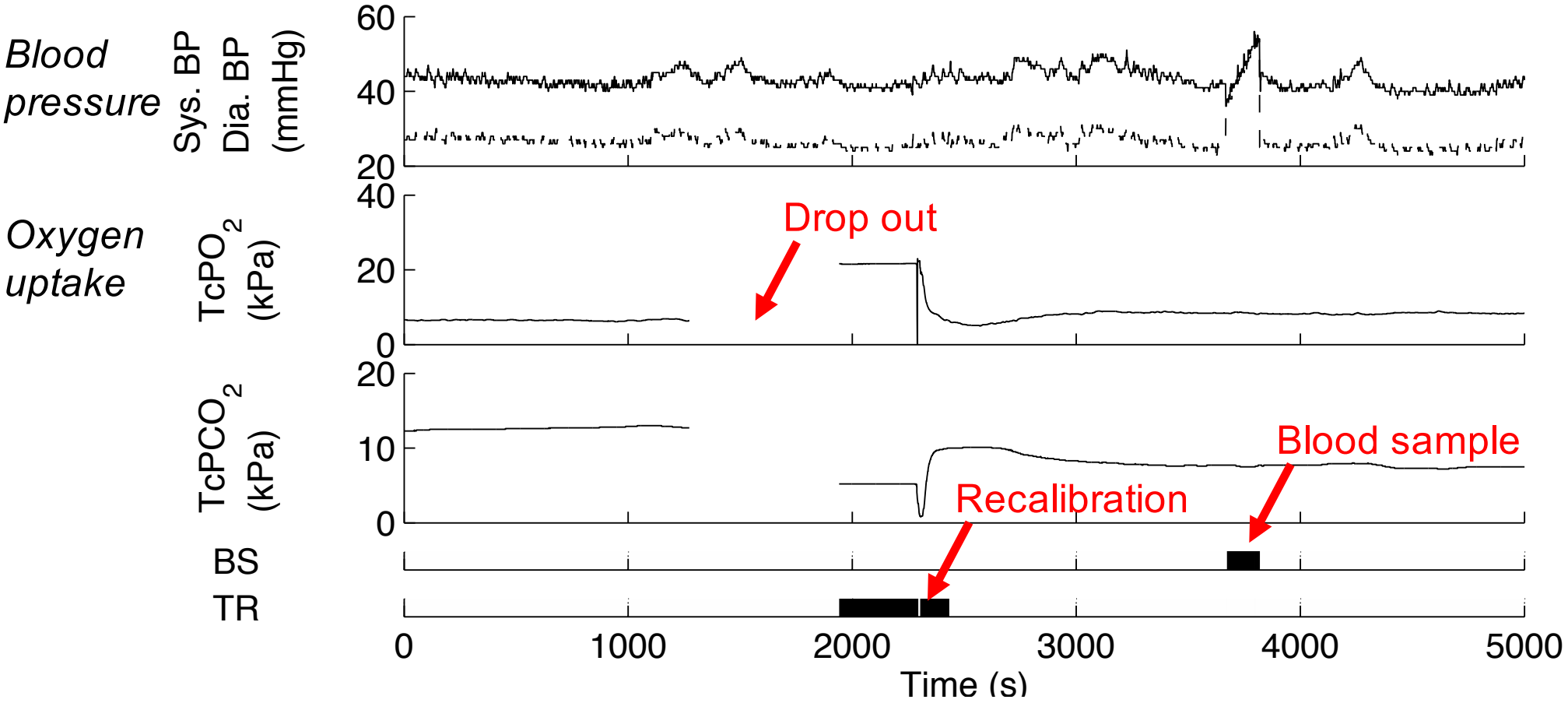
(Quinn et al., TPAMI 2008)

Heart rate dynamics



(Quinn et al., TPAMI 2008)

Confounded by interventions & measurement errors



(Quinn et al., TPAMI 2008)

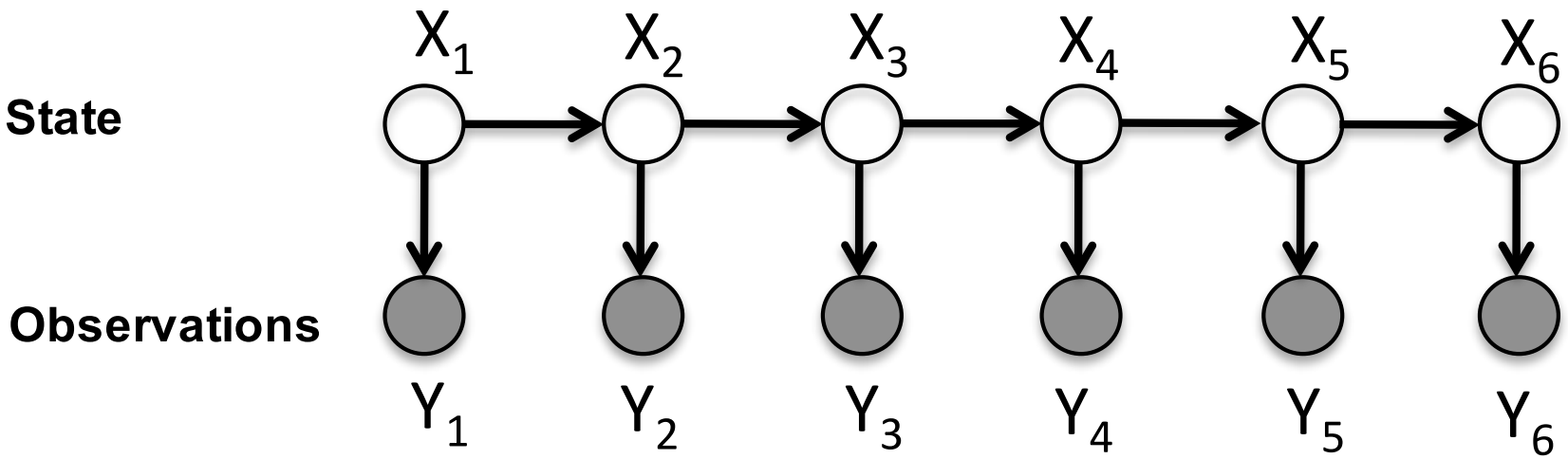
Can we identify the artifactual processes?

- Once identified, can remove for use in downstream predictive tasks (must deal with missing data)
- Can help mitigate **alarm fatigue** by not alerting the clinicians when unnecessary
- More broadly, can we maintain beliefs about the true physiological values of a patient?

(Switching) linear dynamical systems

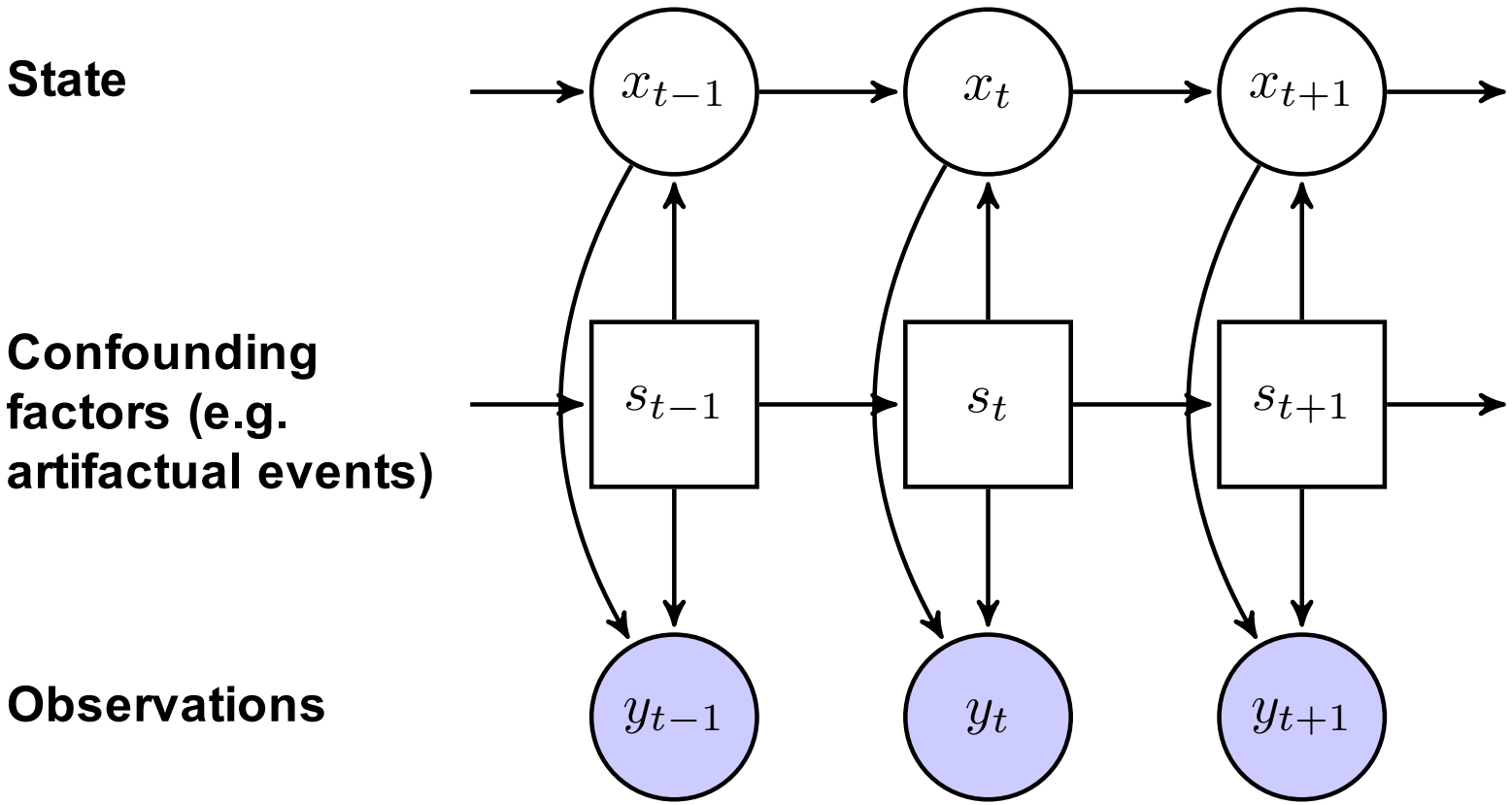
- Conditioned on s_t, linear Gaussian state-space models (Kalman filters):

$$\begin{aligned} \mathbf{x}_t &\sim \mathcal{N}\left(\mathbf{A}^{(s_t)}\mathbf{x}_{t-1} + \mathbf{d}^{(s_t)}, \mathbf{Q}^{(s_t)}\right) \\ \mathbf{y}_t &\sim \mathcal{N}\left(\mathbf{C}^{(s_t)}\mathbf{x}_t, \mathbf{R}^{(s_t)}\right) \end{aligned}$$

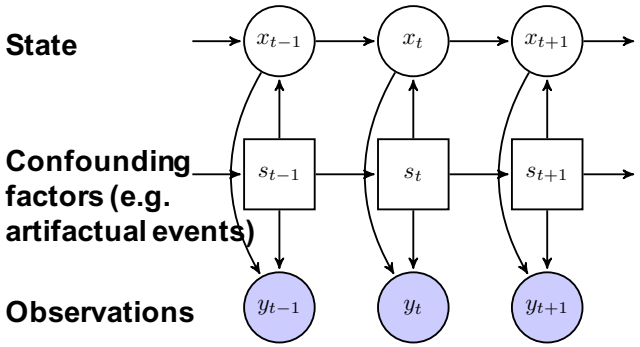


(Switching) linear dynamical systems

- Full model:



Learning SLDS models

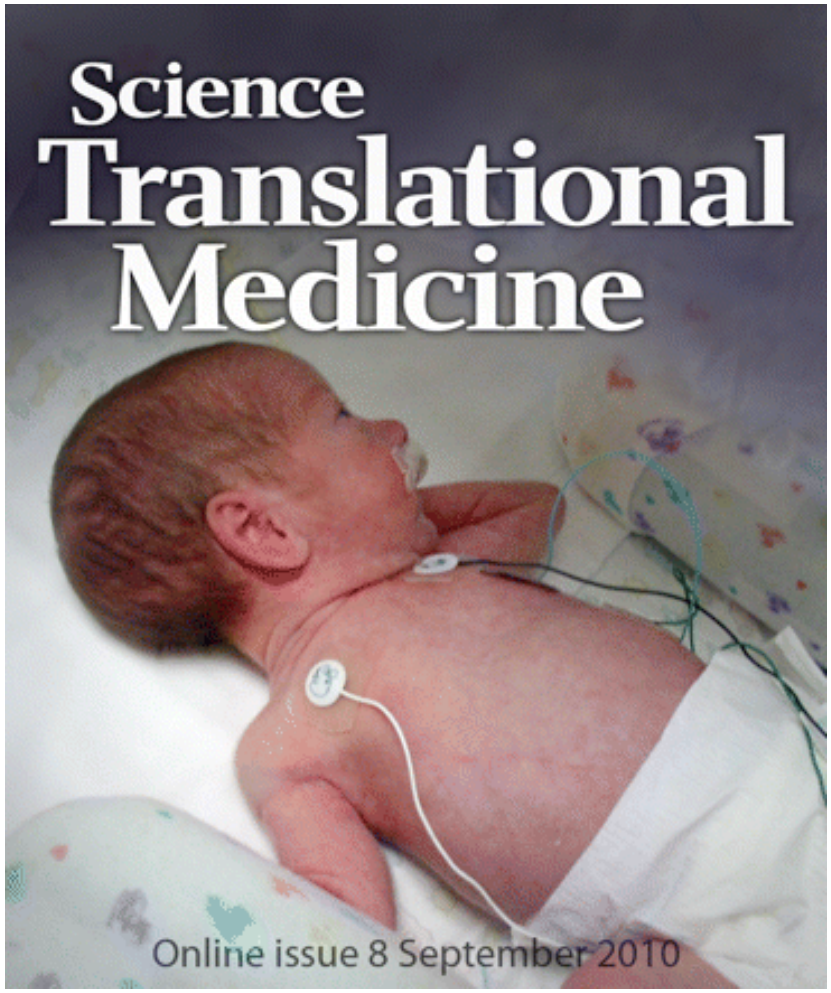


- Assume some labeled training data {**s**,**y**}
- *True state **x** assumed to never be observed*
- Parameterization for **x** depends on states **s**
- Learn using expectation maximization

Outline of today's class

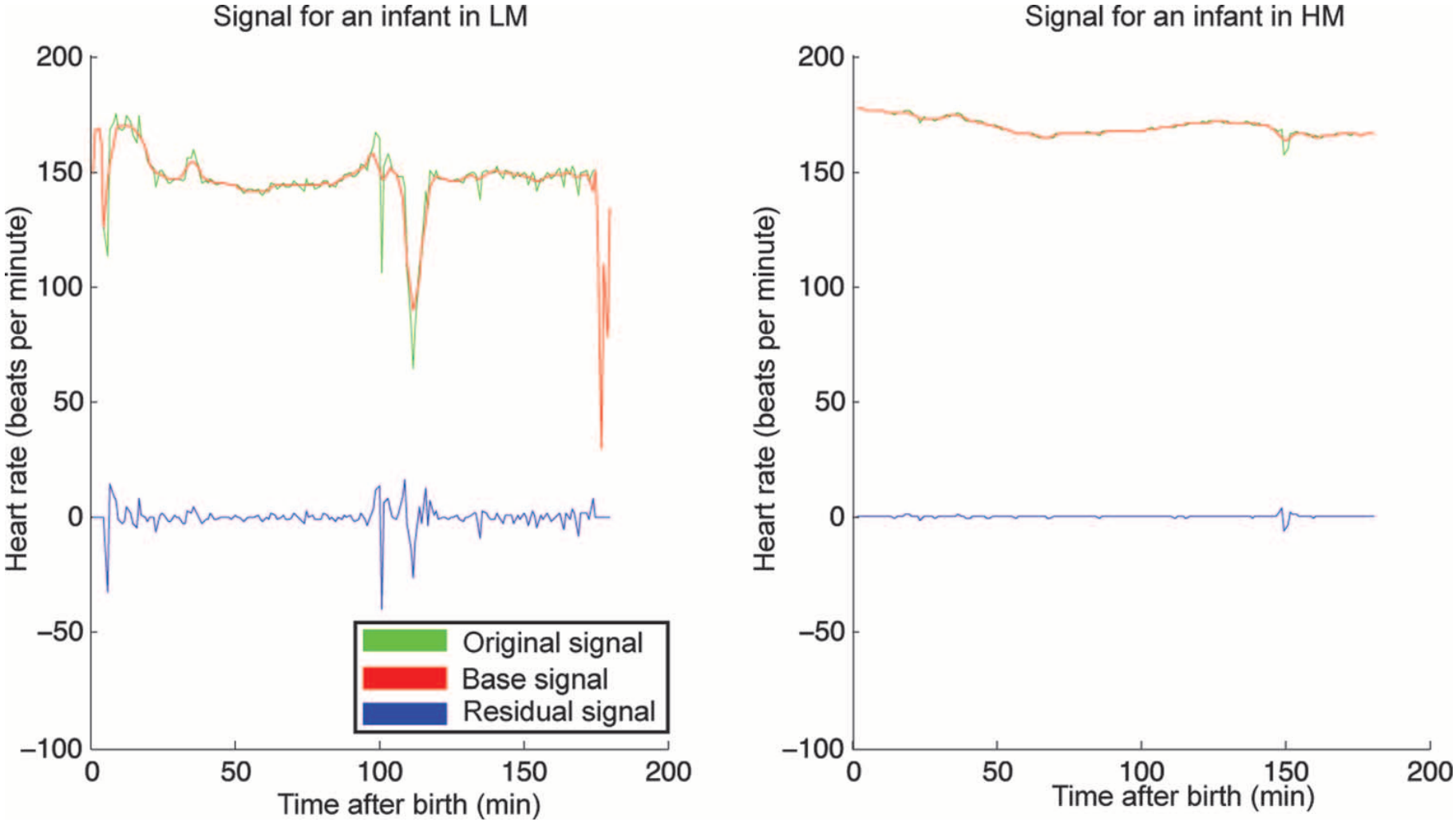
1. State space models for physiological condition modeling
2. **Physiological assessment score for preterm infants**
3. RNNs with missing values (on MIMIC)
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion

Example of risk stratification: predicting morbidity in preterm newborns



Saria et al.,
Science Translational
Medicine 2010

Measuring heart rate variability



(Saria et al., Science Translational Medicine 2010)

Learning algorithm / model

- Logistic regression used to predict whether baby will be "high morbidity" (HM):

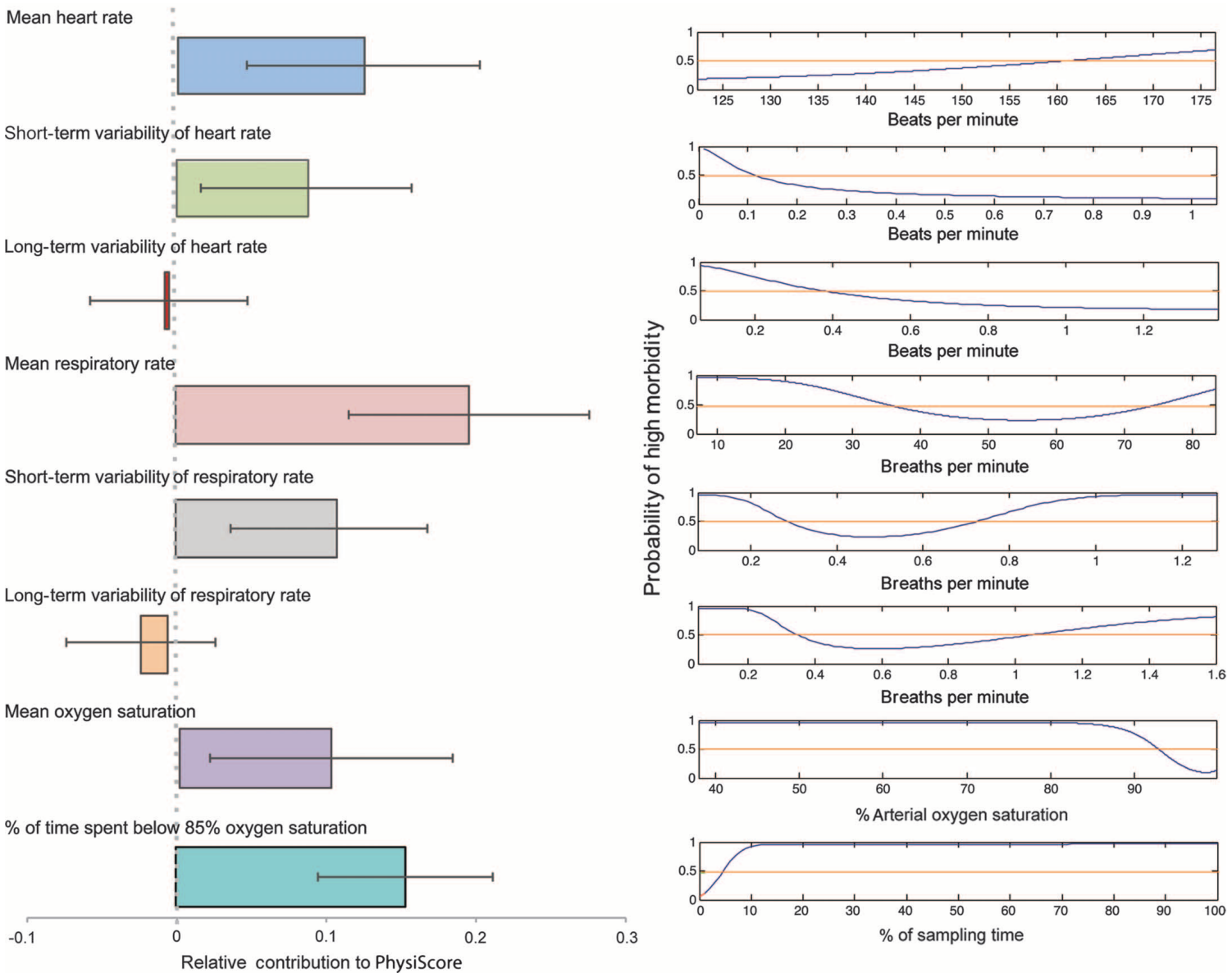
$$P(\mathrm{HM}|v_1,v_2,\ldots,v_n) = \left(1+\exp\left(b + w_0 * c + \sum_{i=1}^{n} w_i * f(v_i)\right)\right)^{-1}$$

- Features computed using 3 hours of data and nonlinear Bayesian model:
  - Estimated $P(v_i \mid C)$ for each class of patient C={HM or LM}) using parametric models such as exponential, Weibull, lognormal, gamma
  - Use log odds ratio of observed value as feature if observed, 0 otherwise:

$$\log P(v_i|\mathrm{HM})/P(v_i|\mathrm{LM})$$

  - *Assumes data missing at random*

Feature importance



Outline of today's class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. **RNNs with missing values (on MIMIC)**
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion

Modeling sequential data with neural networks

- Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the patient's data at time *t*
- By the chain rule, any distribution can be factorized as:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_T) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1})$$

- Train a neural net that composes history to predict next time step:

$g(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) \in \mathbb{R}^k$: composition function

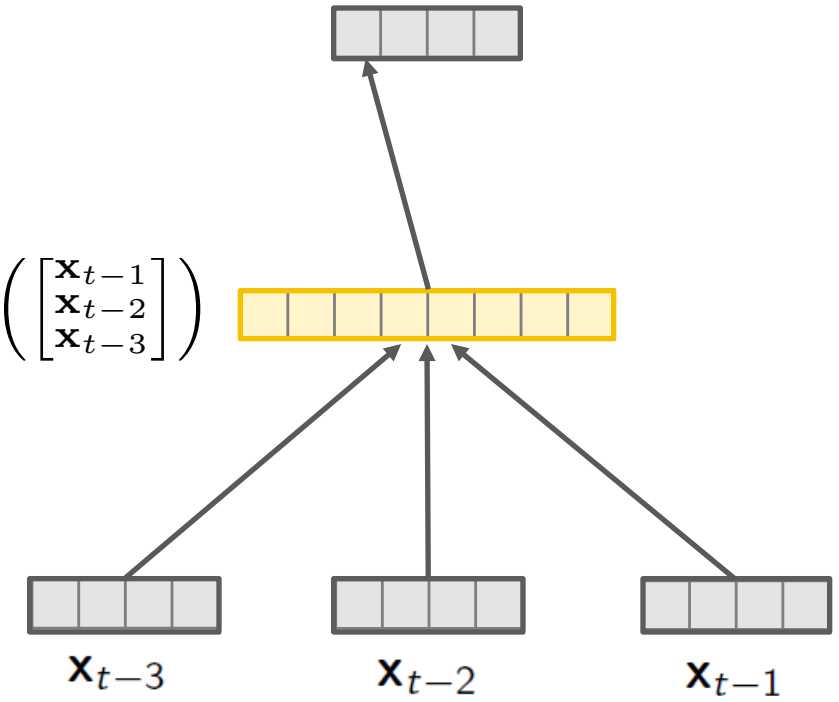
$\mathbf{w}_i \in \mathbb{R}^k, \quad b^i \in \mathbb{R}$: parameters $i = 1, \ldots, d$

$$\begin{aligned} p(x_t^i = 1 \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) &= \frac{e^{\mathbf{w}^i \cdot g(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) + b^i}}{1 + e^{\mathbf{w}^i \cdot g(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) + b^i}} \\ &= \text{logistic}(\mathbf{w}^i \cdot g(\mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) + b^i) \end{aligned}$$

Modeling sequential data with neural networks

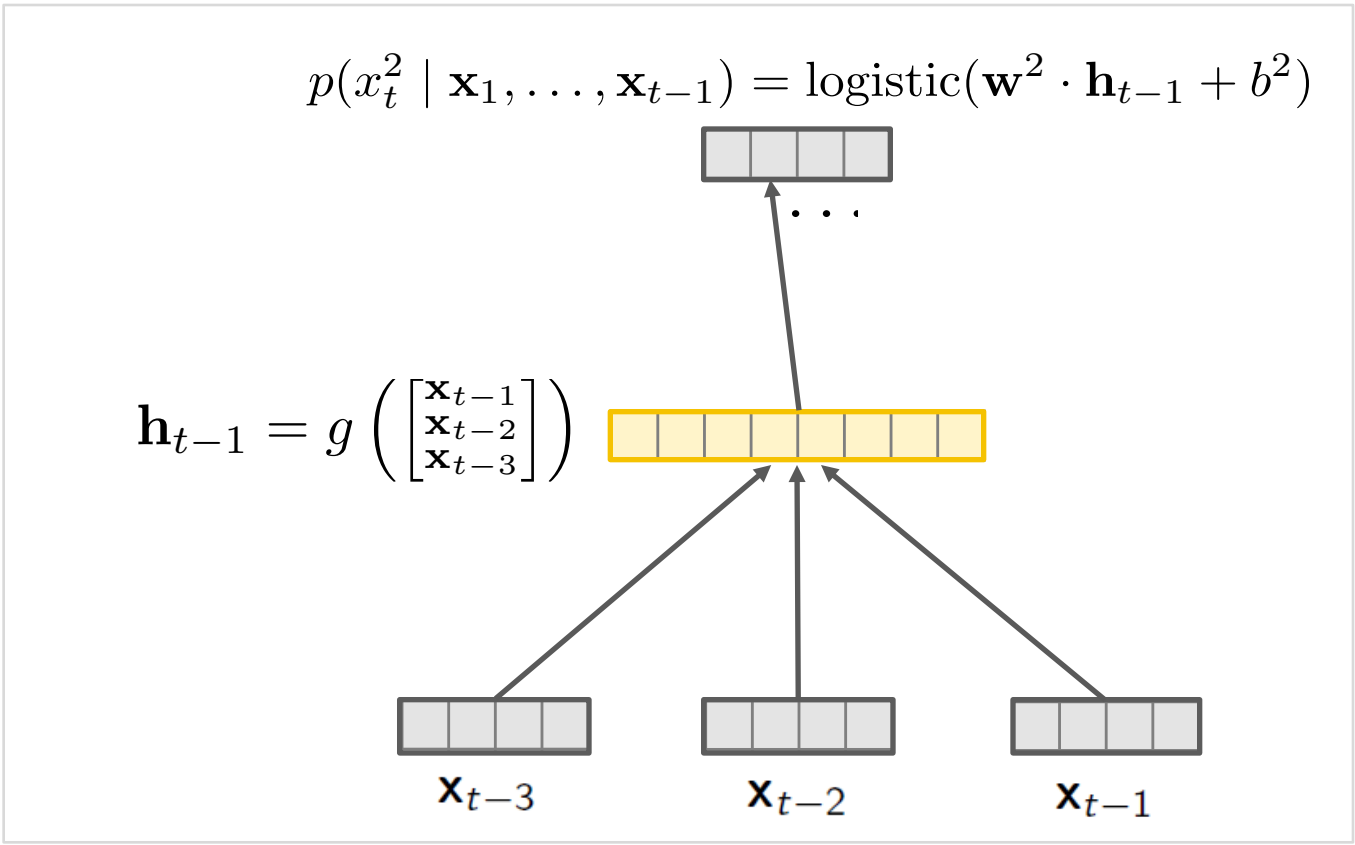
$$p(x_t^1 \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) = \text{logistic}(\mathbf{w}^1 \cdot \mathbf{h}_{t-1} + b^1)$$

$$\mathbf{h}_{t-1} = g\left(\begin{bmatrix}\mathbf{x}_{t-1}\\ \mathbf{x}_{t-2}\\ \mathbf{x}_{t-3}\end{bmatrix}\right)$$



Modeling sequential data with neural networks

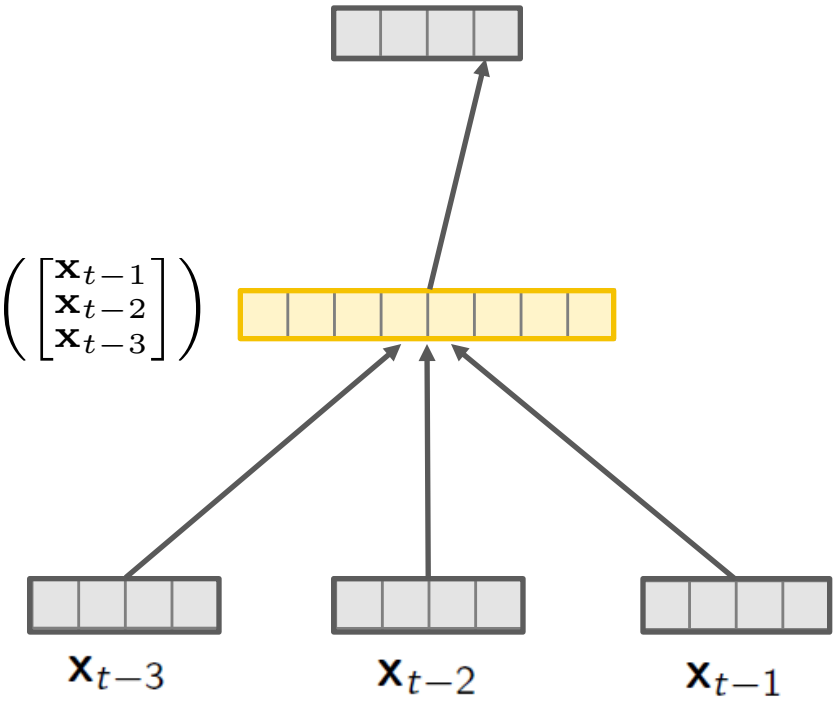
$$p(x_t^2 \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) = \text{logistic}(\mathbf{w}^2 \cdot \mathbf{h}_{t-1} + b^2)$$



Modeling sequential data with neural networks

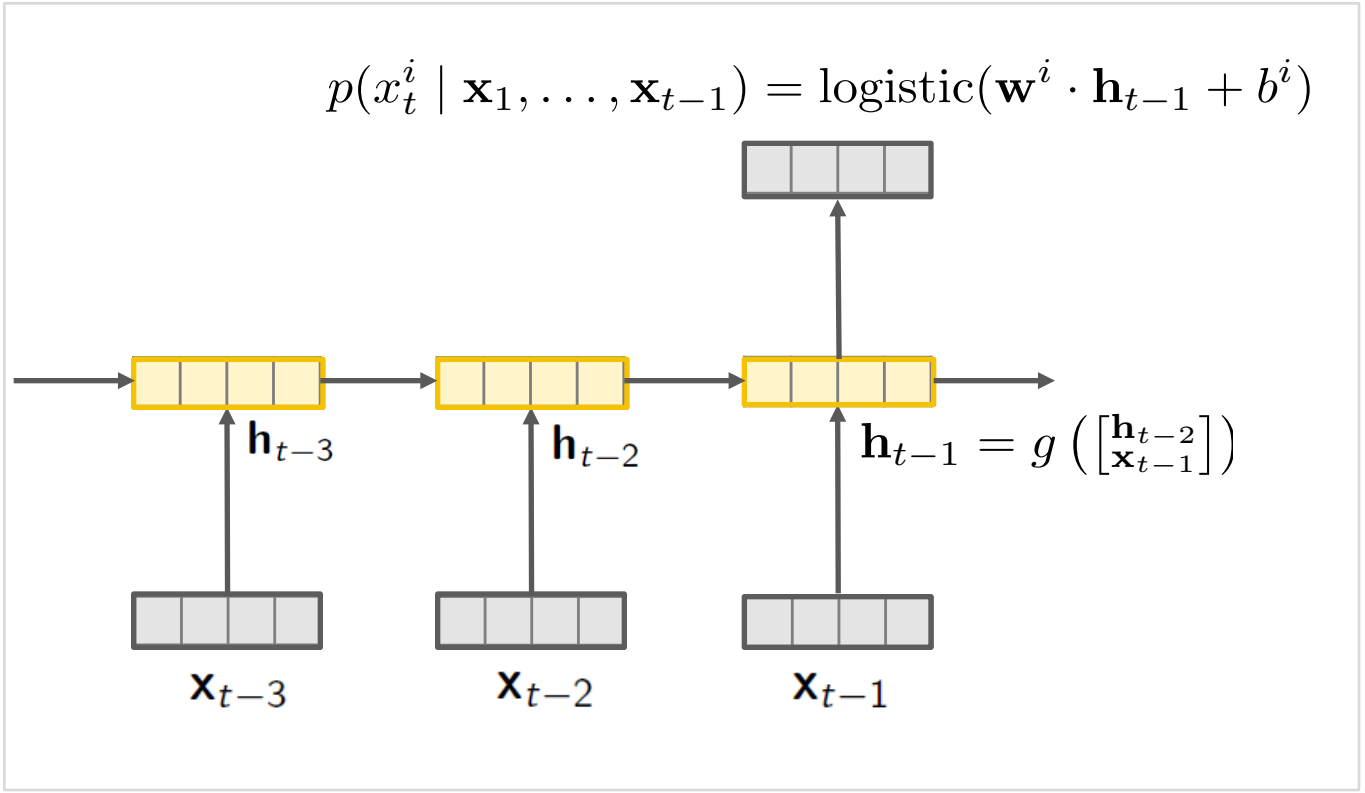
$$p(x_t^d \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) = \text{logistic}(\mathbf{w}^d \cdot \mathbf{h}_{t-1} + b^d)$$

$$\mathbf{h}_{t-1} = g\left(\begin{bmatrix}\mathbf{x}_{t-1}\\ \mathbf{x}_{t-2}\\ \mathbf{x}_{t-3}\end{bmatrix}\right)$$



Recurrent neural networks (RNNs)

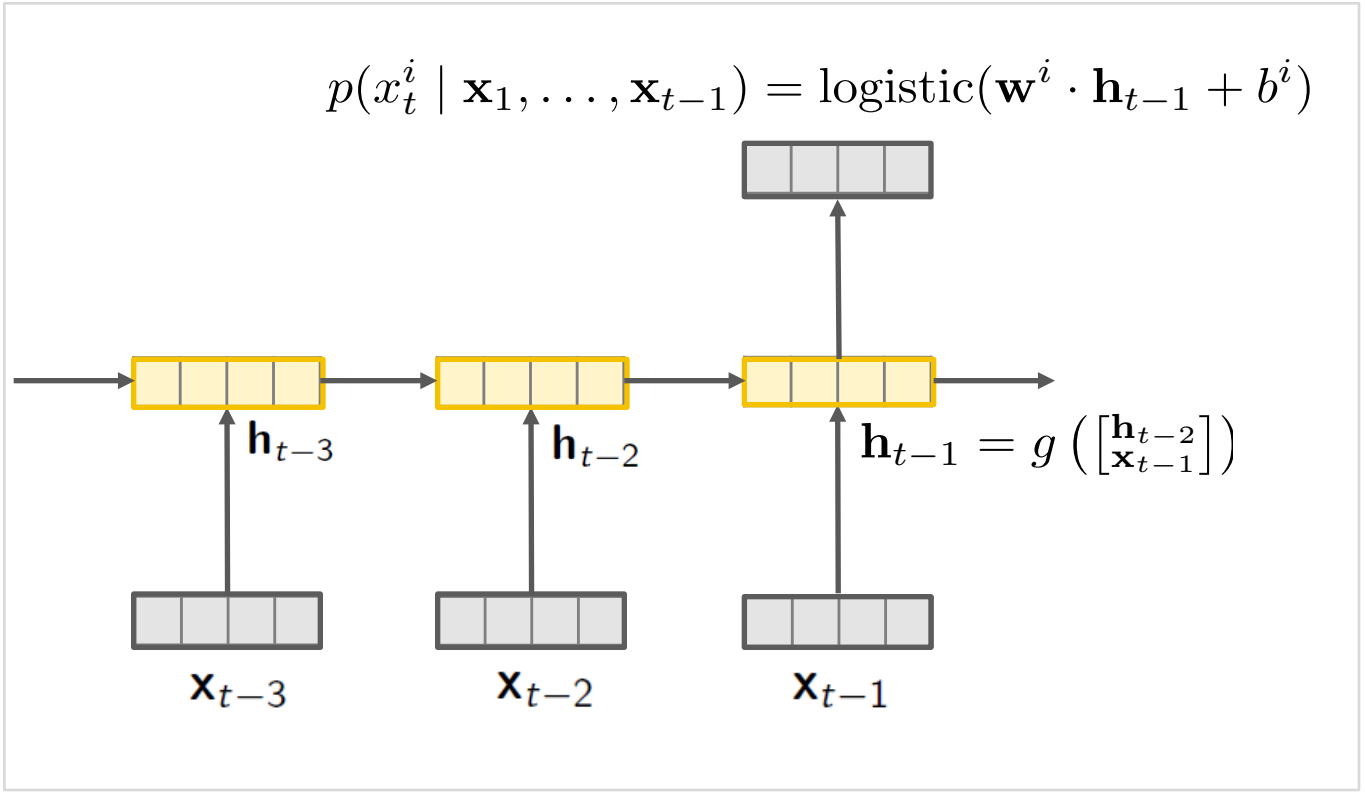
Maintain a hidden state vector $\mathbf{h}_t$ that is recursively calculated



RNN language models widely used in natural language processing: state-of-the-art performance for speech recognition and machine translation

Recurrent neural networks (RNNs)

Maintain a hidden state vector $\mathbf{h}_t$ that is recursively calculated

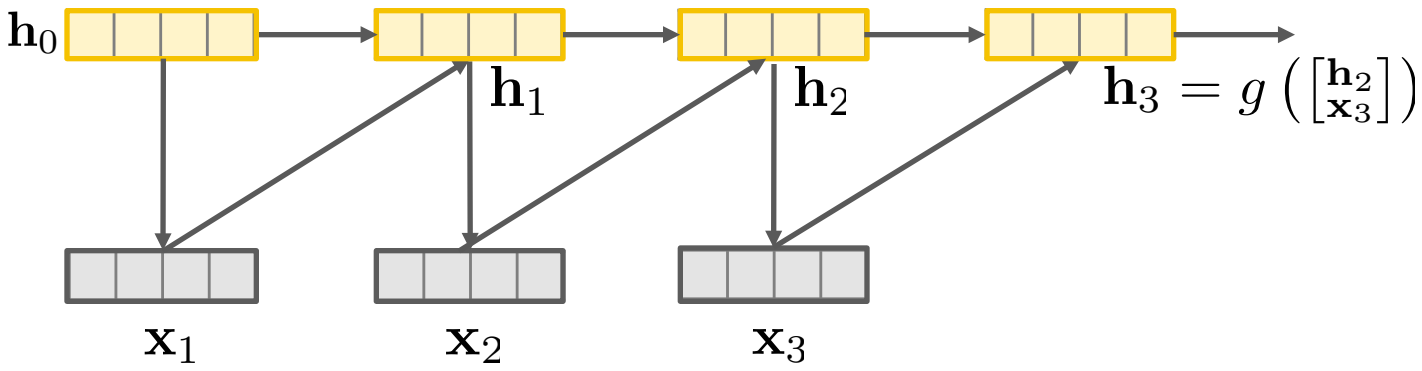


Significant interest in using RNNs for disease progression modeling:

- **Doctor AI**, Choi et al., *arXiv:1511.05942*, Nov. 2015.
- **DeepCare**, Pham et al., *arXiv:1602.00357*, Feb. 2016

RNNs versus HMMs

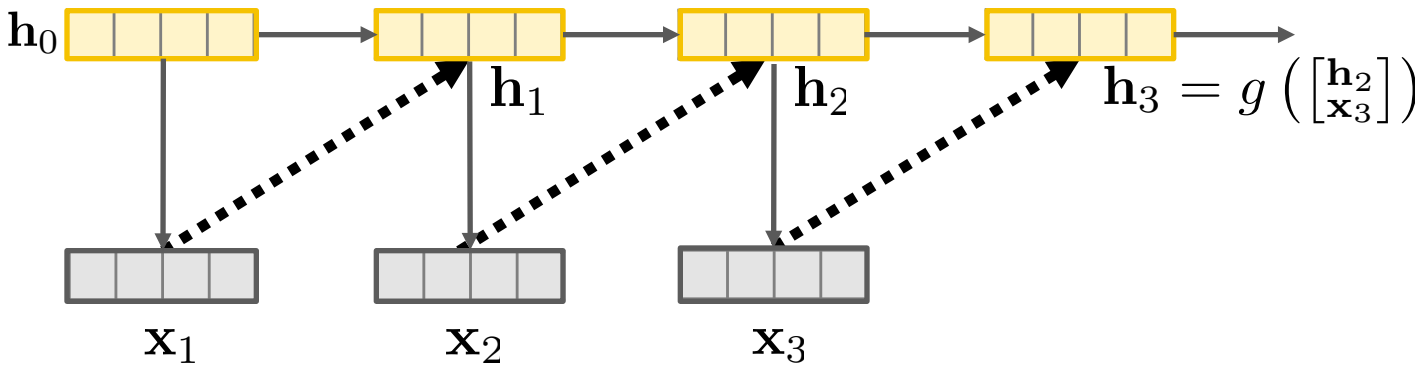
Equivalently, viewing the RNN as a Markov model:



- **Advantage:** Very powerful
- **Disadvantages:**
  - Not easy to deal with missing data in **x**
  - No ability to discover structure in **x** – can overfit if *d* is large
  - All randomness due to exogenous factors must be captured in **x** observations
  - Difficult (not impossible) to incorporate prior knowledge and to combine as part of a more complex model

RNNs versus HMMs

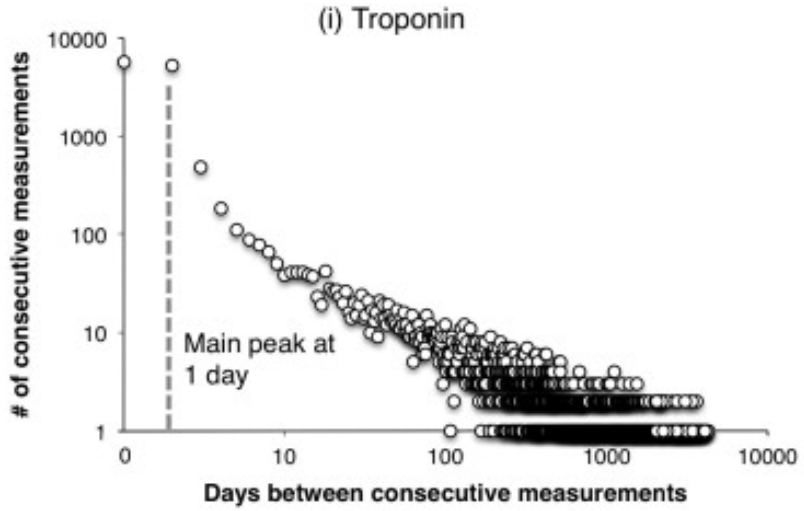
Equivalently, viewing the RNN as a Markov model:

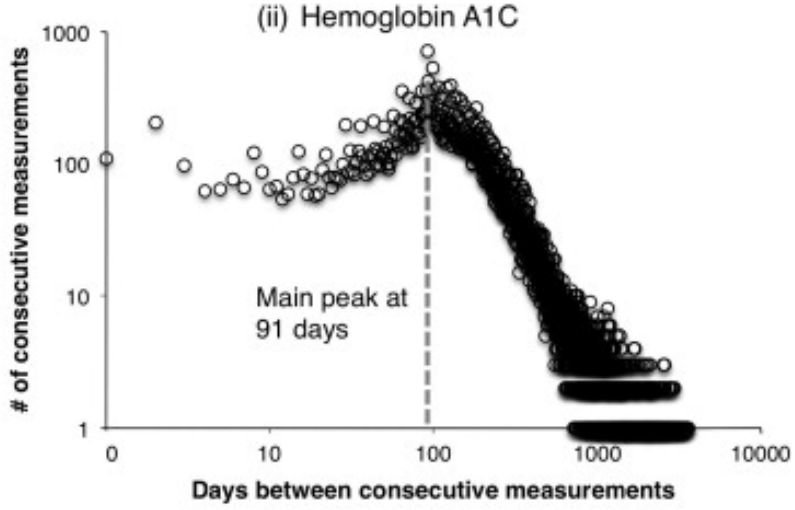


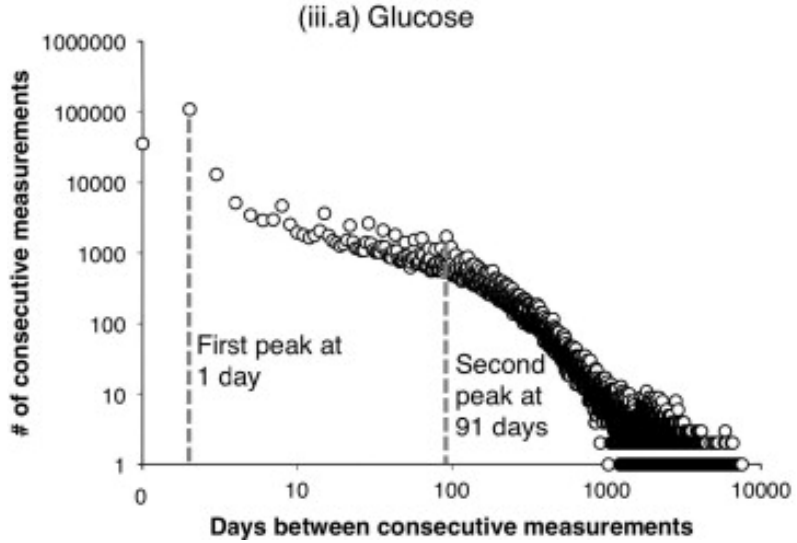
Can't remove the edges from **x** to **h** in this model, because it becomes useless (due to **h** transitions being deterministic):

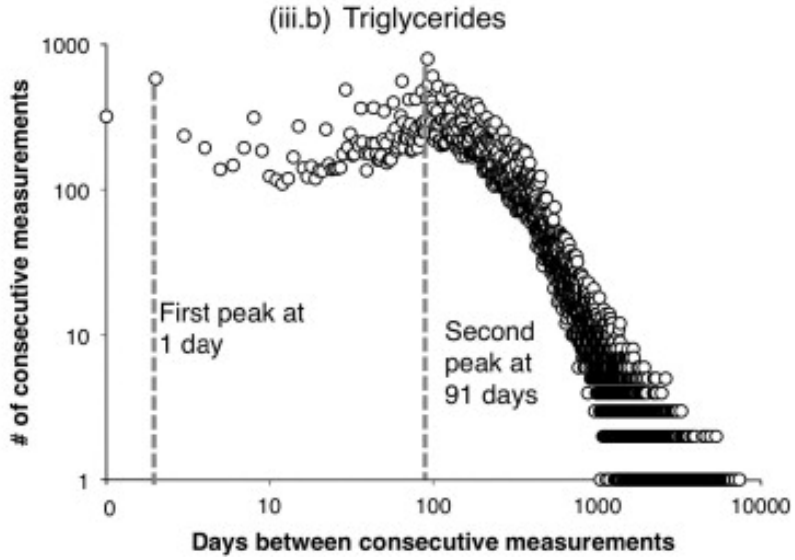
$$p(\mathbf{x}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_{t-1}) = p(\mathbf{x}_t)$$

Timing matters! Measurement motifs





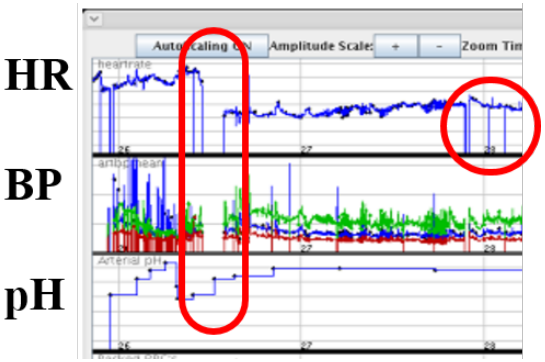




(Pivovarov et al., JBI 2014)

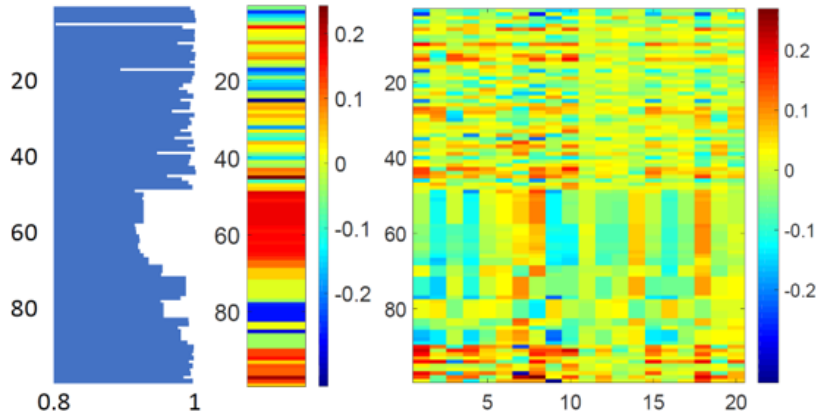
How can we exploit missingness in RNNs?

Missingness comes from various reasons.



AZ	BA	BB	BC	BD	BE	BF	BG	BH
CRS1	CRS2	CRS3	FIO21	FIO22	FiO23	HCO31	HCO32	HCO33
0.27649	0.23680	0.23079	0.45295	0.45729	0.44999	23.9965	23.4375	24.1134
0.61792	1.14405	0.73171	0.39041	0.35673	0.34999	19.05625	19	
0.60328	0.29352	0.29644	0.35100	0.37197	0.40717	19.2951	22.5520	28
0.72348	0.67720	0.59685	0.44999	0.44999	0.41788	20.1	29.6145	33.6753
0.40175			0.41777			18.6541	21.5583	22
0.27366	0.15783	0.24334	0.97458	0.69583	0.60762	28.1048	38.5090	38.4861
0.39656			0.35808			23.3631	26.9194	27
						18.87		
0.58429	0.44144	0.41550	0.44999	0.55625	0.37904			
0.39599	0.31453		0.49458	0.48620		21.46875	20.3508	
0.22629	0.20941	0.28634	0.40000	0.40000	0.40000	29.1194	28.0238	
0.34741	0.39616	0.39896	0.44554	0.43111	0.40000	26.1506	29.5548	33.5720
						33.0423	34.7375	38.8510
0.25392	0.30970	0.38195	0.46083	0.48041	0.49755	21.7972	24.9194	23.3015
0.79393	0.89380	0.59436	0.52899	0.33697	0.30000	22.9472	20.1298	20.1527

Missingness provides rich information about patients health condition.



(Che et al., “Recurrent Neural Networks for Multivariate Time Series with Missing Values”, arXiv:1606.01865, 2016)

Represent and Utilize Missing Values

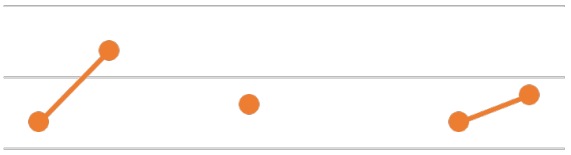
Two representations of missingness:

- *Masking* $\mathbf{M}$:

  Whether a variable is missing or not.

- *Time Interval* $\boldsymbol{\Delta}$:

  How long a variable has been missing.



Time Step	0	1	2	3	4	5	6	7
Value	47	55	N/A	49	N/A	N/A	47	50
Masking	1	1	0	1	0	0	1	1
Duration	1	1	1	2	1	2	3	1

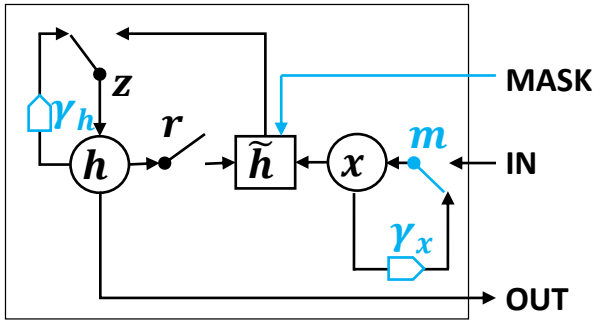
Decay Term γ: A flexible transformation on $\boldsymbol{\Delta}$ jointly learned with deep model.

$$\gamma_t = \exp\{-ReLU(\mathbf{W}_\gamma \delta_t + \mathbf{b}_\gamma)\}$$

$$x_t^d \leftarrow m_t^d x_t^d + (1 - m_t^d)\gamma_{x_t^d} x_{t'}^d + (1 - m_t^d)(1 - \gamma_{x_t^d})\tilde{x}^d$$

GRU-D model

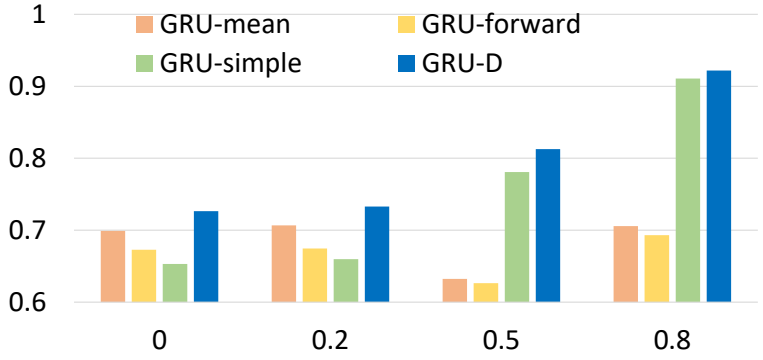
- Decay on the last observations.
- Decay on the hidden states.



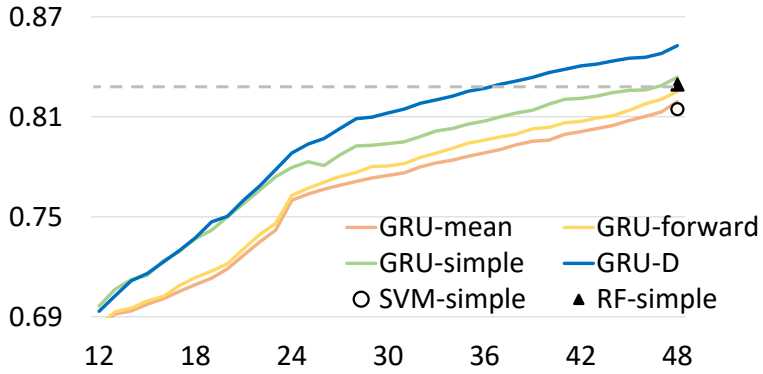
(Che et al., "Recurrent Neural Networks for Multivariate Time Series with Missing Values", arXiv:1606.01865, 2016)

Quantitative Evaluation

Evaluations on synthetic dataset with different missing rates



Evaluations for mortality early prediction

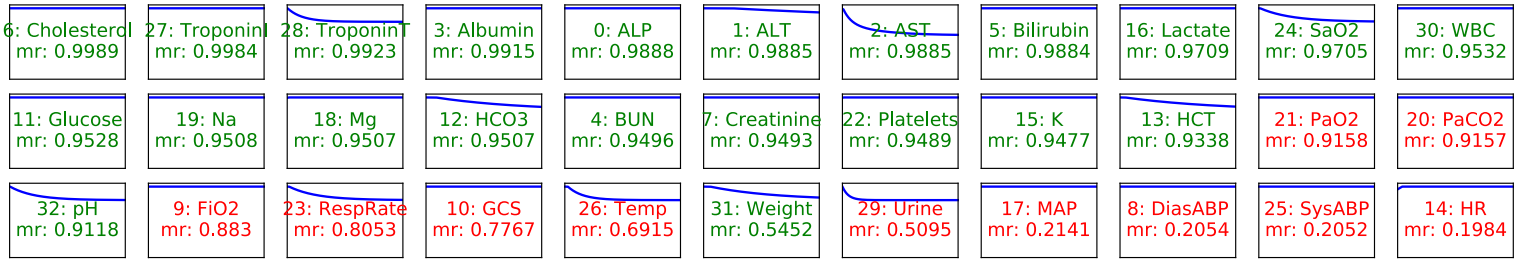


AUC score on mortality prediction

	Models	MIMIC-III	PhysioNet
Non-RNN	LR-forward	0.7589	0.7423
	SVM-forward	0.7908	0.8131
	RF-forward	0.8293	0.8183
	LR-simple	0.7715	0.7625
	SVM-simple	0.8146	0.8277
	RF-simple	0.8294	0.8157
RNN	LSTM-mean	0.8142	0.8025
	GRU-mean	0.8192	0.8195
	GRU-forward	0.8252	0.8162
	GRU-simple	0.8380	0.8155
Ours	GRU-D	**0.8527**	**0.8424**

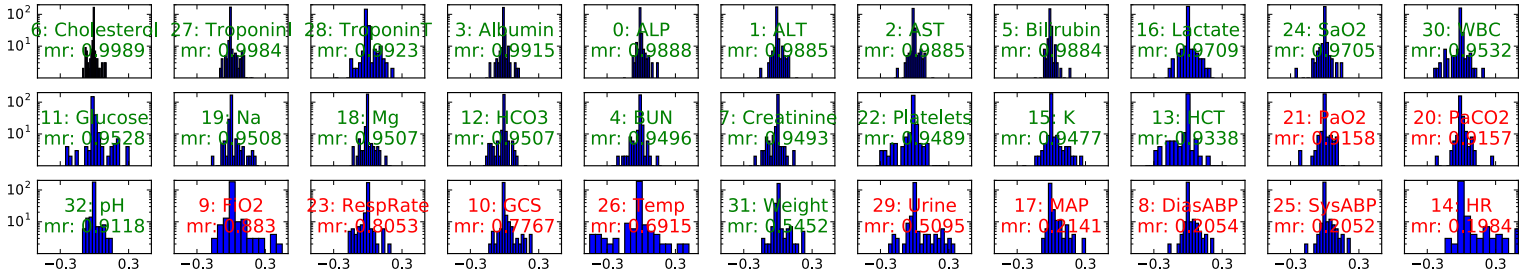
Qualitative Evaluation

Input decay plots of all 33 variables for mortality prediction on PhysioNet dataset



- Get a few important variables, e.g., weight, arterial pH, temperature, and respiration rate, etc.

Histograms of of hidden state decay for mortality prediction on PhysioNet dataset



- Parameters related to variables with smaller missing rate are more spread out.

Outline of today's class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. **CNNs for predicting disease onsets from longitudinal lab tests**
5. Project discussion

Multi-task prediction of disease onsets from longitudinal lab tests

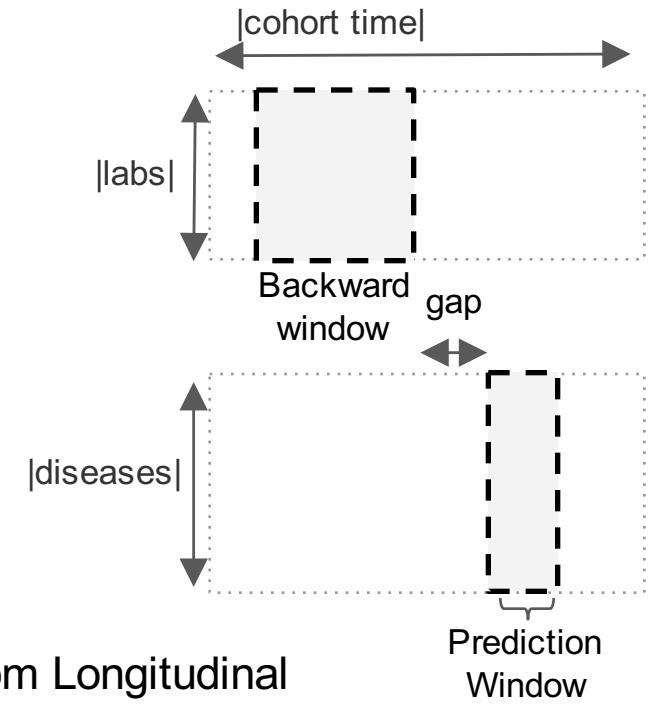
Goal:

- Early diagnosis of diseases for people who *do not* already have the disease.
- Going toward raw biological signals (i.e. lab measurements) and learning rich representations directly from the raw input

Framework:
Multi-task Supervised Prediction

Input
Biomarkers over time

Output
Disease onsets over time



(Razavian et al., "Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests". 1st Conference on Machine Learning and Health Care, 2016)

Cohort

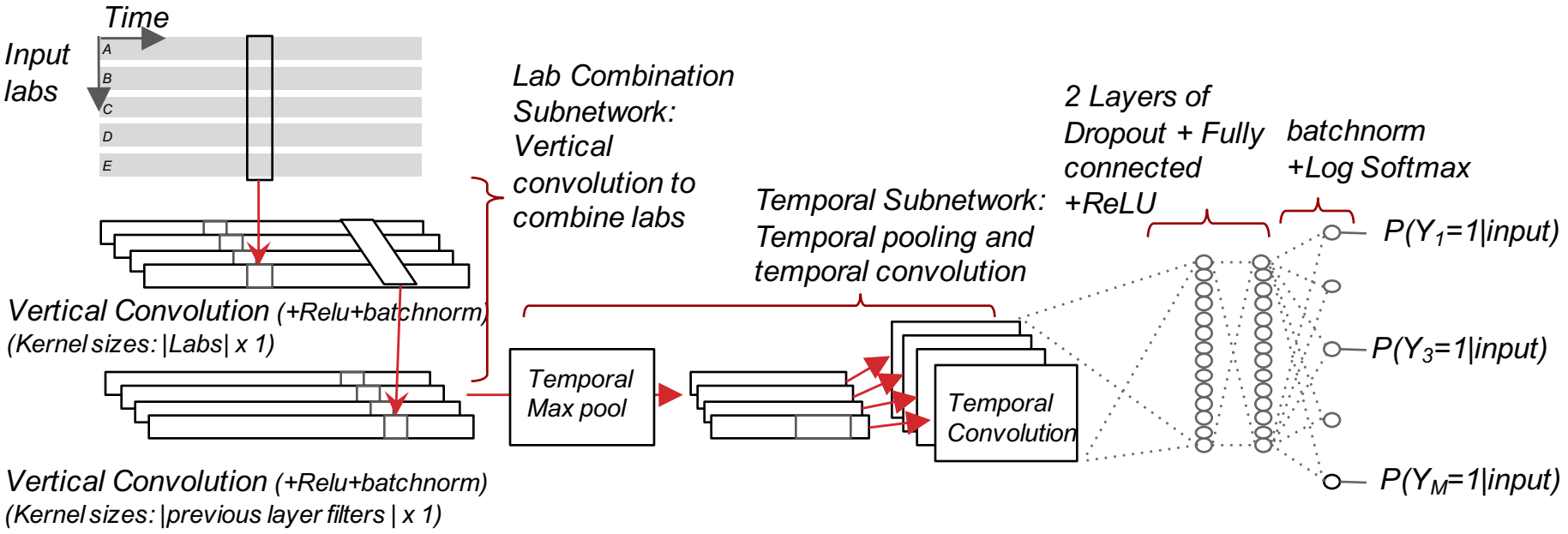
298,000 individuals, with at least once a year lab measurement for 3 consecutive years included

- **Input**: Comprehensive lab panel + cholesterol (18 labs)
- **Output**: 133 conditions.

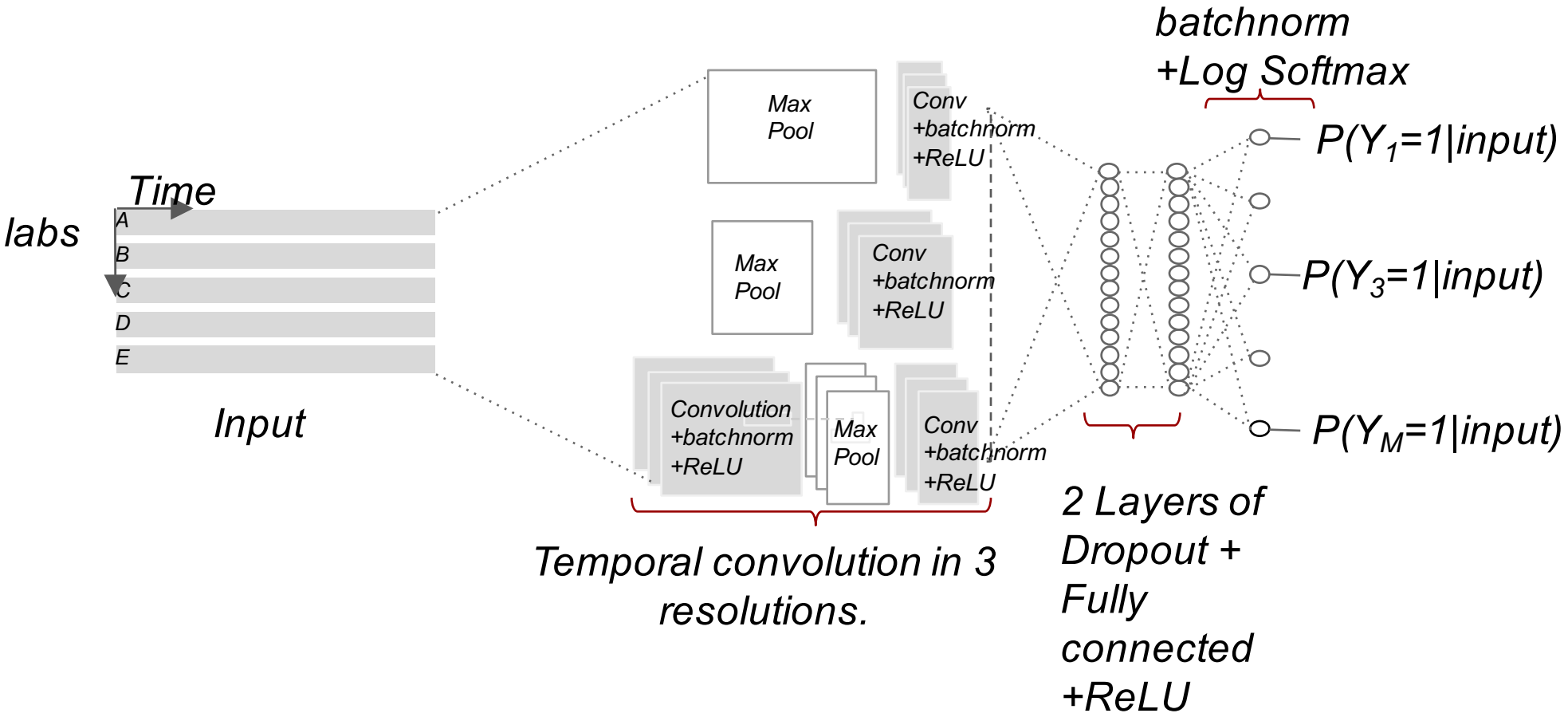
- **Exclusion Per Disease**: Anyone with even 1 measurement prior to start of prediction window
  - Done via masking the gradients in SGD process for excluded patients per task.

- Randomly Split to train(100K), validate(100K) and test(98k)set

CNN-1: Convolution over Labs then Time

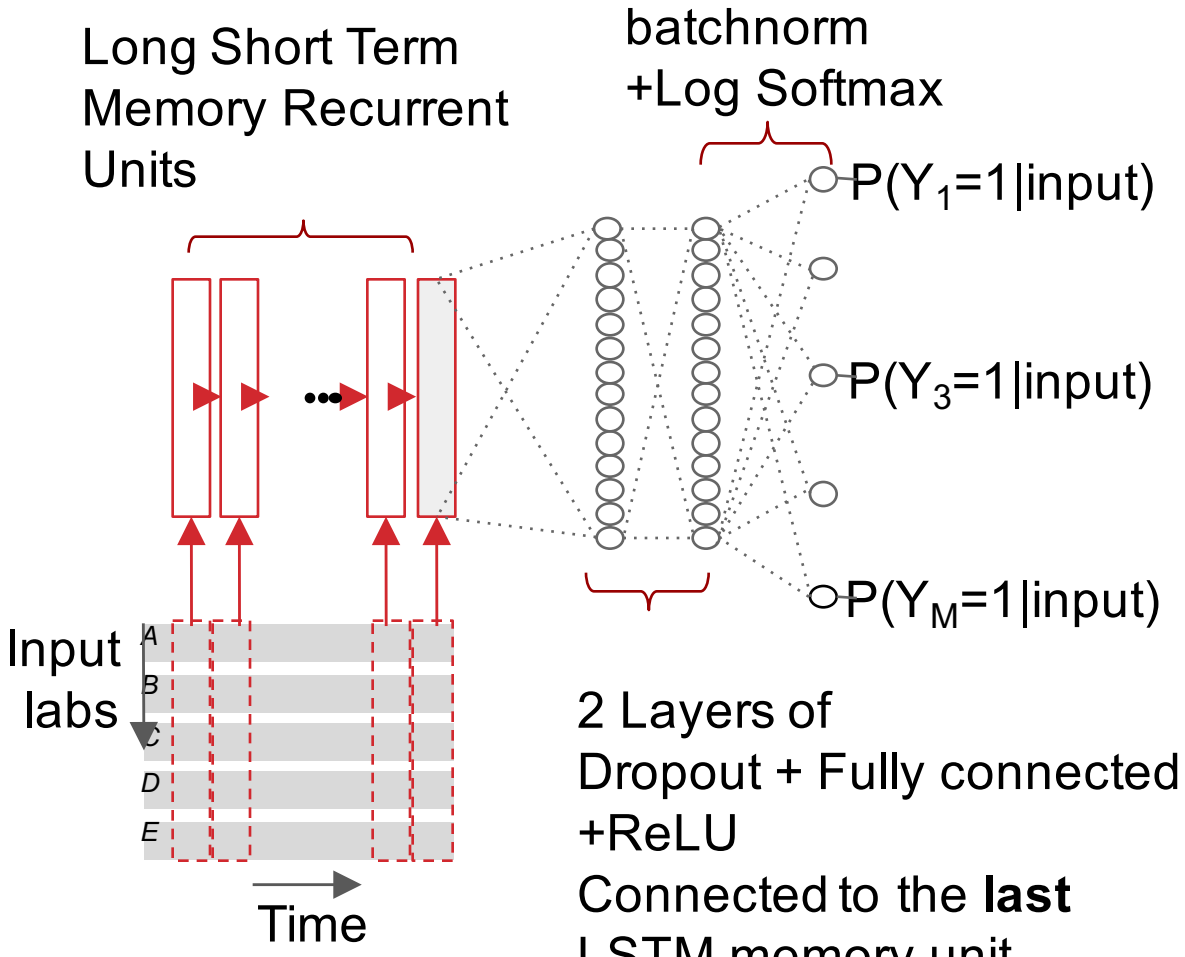


CNN-2: Multi-resolution Convolution over Time



LSTM for Sequence Embedding

Lab name
Creatinine
Urea nitrogen
Potassium
Glucose
Alanine aminotransferase
Aspartate aminotransferase
Protein
Albumin
Cholesterol
Triglyceride
Cholesterol.in LDL
Calcium
Sodium
Chloride
Carbon dioxide
Urea nitrogen/Creatinine
Bilirubin
Albumin/Globulin



Results

Goal: predict new onset of diseases 3 months in advance

ICD9 Code and disease description	LR	LSTM	CNN1	CNN2	Ens	Pos
585.6 End stage renal disease	0.886	0.917	0.910	0.916	0.920	837
285.21 Anemia in chr kidney dis	0.849	0.866	0.868	0.880	0.879	1598
585.3 Chr kidney dis stage III	0.846	0.851	0.857	0.858	0.864	2685
584.9 Acute kidney failure NOS	0.805	0.820	0.828	0.831	0.835	3039
250.01 DMI wo cmp nt st uncntrl	0.822	0.813	0.819	0.825	0.829	1522
250.02 DMII wo cmp uncntrld	0.814	0.819	0.814	0.821	0.828	3519
593.9 Renal and ureteral dis NOS	0.757	0.794	0.784	0.792	0.798	2111
428.0 CHF NOS	0.739	0.784	0.786	0.783	0.792	3479
V053 Need prphyl vc vrl hepat	0.731	0.762	0.752	0.780	0.777	862
790.93 Elvtd prstate spcf antgn	0.666	0.758	0.761	0.768	0.772	1477
185 Malign neopl prostate	0.627	0.757	0.751	0.761	0.768	761
274.9 Gout NOS	0.746	0.761	0.764	0.757	0.767	1529
362.52 Exudative macular degen	0.687	0.752	0.750	0.757	0.765	538

AUC sorted by maximum AUC achieved by any of the models

Observations

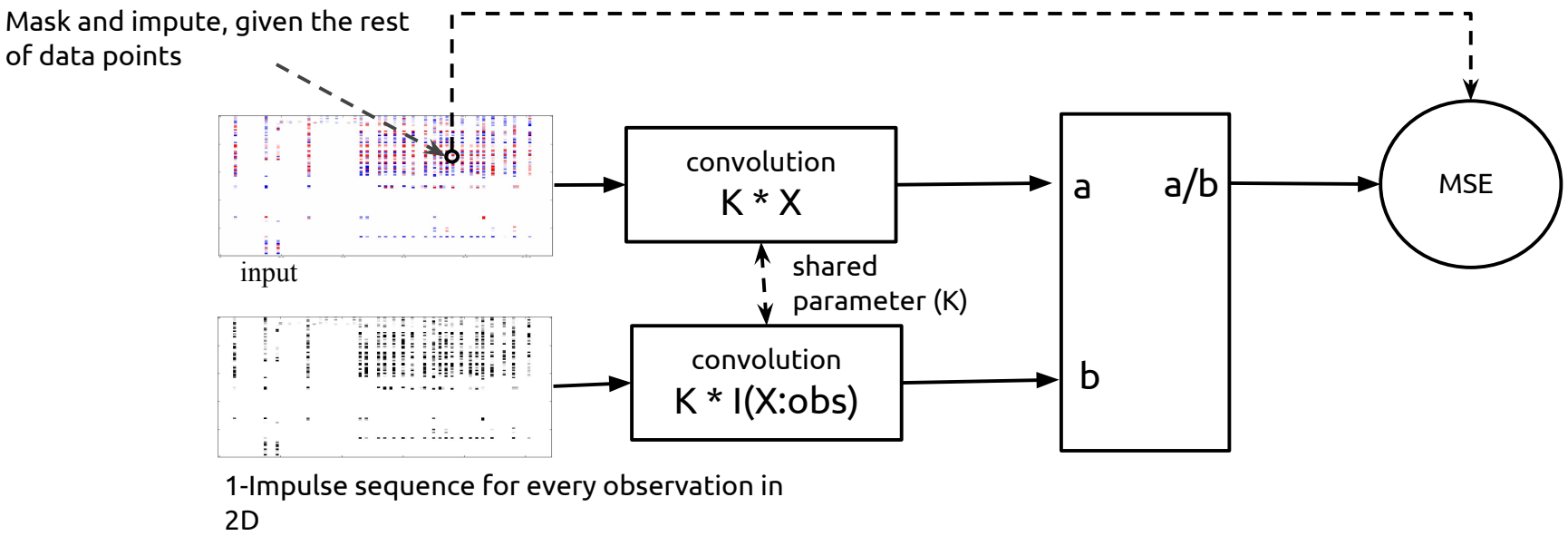
- Rich representation learning improves prediction quality of weaker tasks in the multi-task settings
- Most gains are on tasks where the predictive features are NOT directly included in the input already
  - Confirmed by the case study of chronic kidney disease progression, and our most-improved outcomes
- Different representation learning methods (CNN1, CNN2, LSTM) show similar improvements.
- Ensemble of best models *always* further improves results

https://github.com/clinicalml/deepDiagnosis

A Model for Imputation On Correlated Biomarkers

- Imputation model based on structured multivariate kernel regression/smoothing[1]
- Formulated as unsupervised learning method

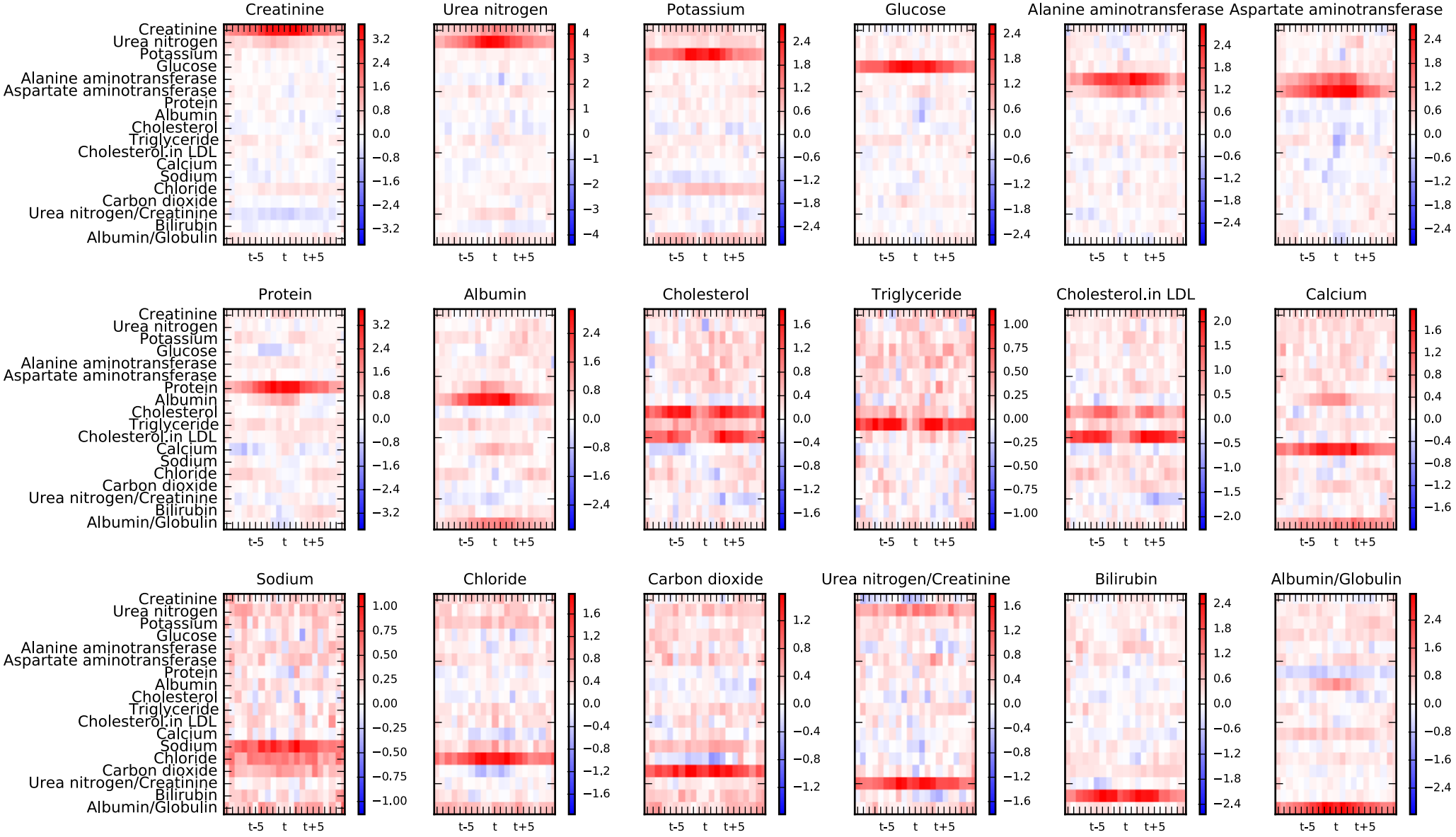
$$\mathbf{E}_{x \sim P(x|t=t_{new})}[x] = \frac{(K * \bar{X}_{train})(t_{new})}{(K * I(\bar{X}_{train} : observed))(t_{new})}$$



(Razavian & Sontag, arXiv:1511.07938, 2015)

[1] Nadaraya and Watson, 1964

Multivariate Kernels learned for each input dimension (total 18)



Data: 30K Individuals from the original training set.

Dataset split equally between train, test and validate set.

Loss: MSE. Train and evaluate only on (lab, person) with more than 1 observation.

Outline of today's class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. CNNs for predicting disease onsets from longitudinal lab tests
5. **Project discussion**

#5: Disease progression in multiple myeloma

Clinical mentor:

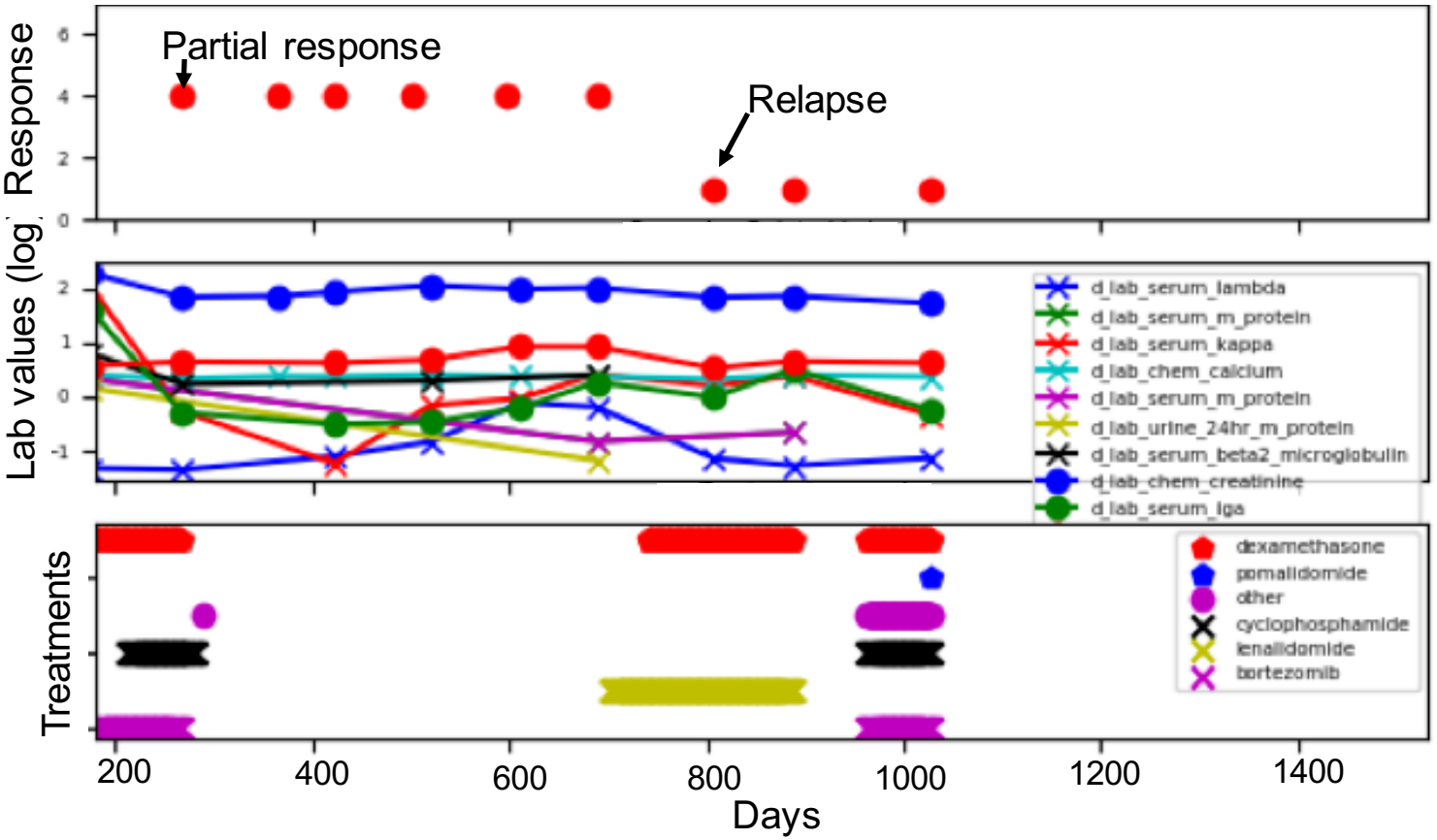
Nikhil Munshi, MD
Dana-Farber Cancer Institute
Professor of Medicine, Harvard Medical School

- Blood cancer, affecting 0.8% of US population at some point in their lifetime. 5 year survival rate is 49%
- Major advances in treatment, with 10+ new drugs on the market, more in clinical trials
- Project goal: predict patient survival and time to disease progression
- Data for ~1000 individuals:
  - Cytogenetics, mutations, gene expression
  - Biomarker levels across time (eg immunoglobulin levels)
  - Clinical outcomes including disease status, time to response, treatment response
  - Adverse events (eg anemia, bone pain, renal failure…)
  - Quality of life measures (e.g. appetite loss, fatigue) and symptoms
  - Treatment therapies including combination treatments

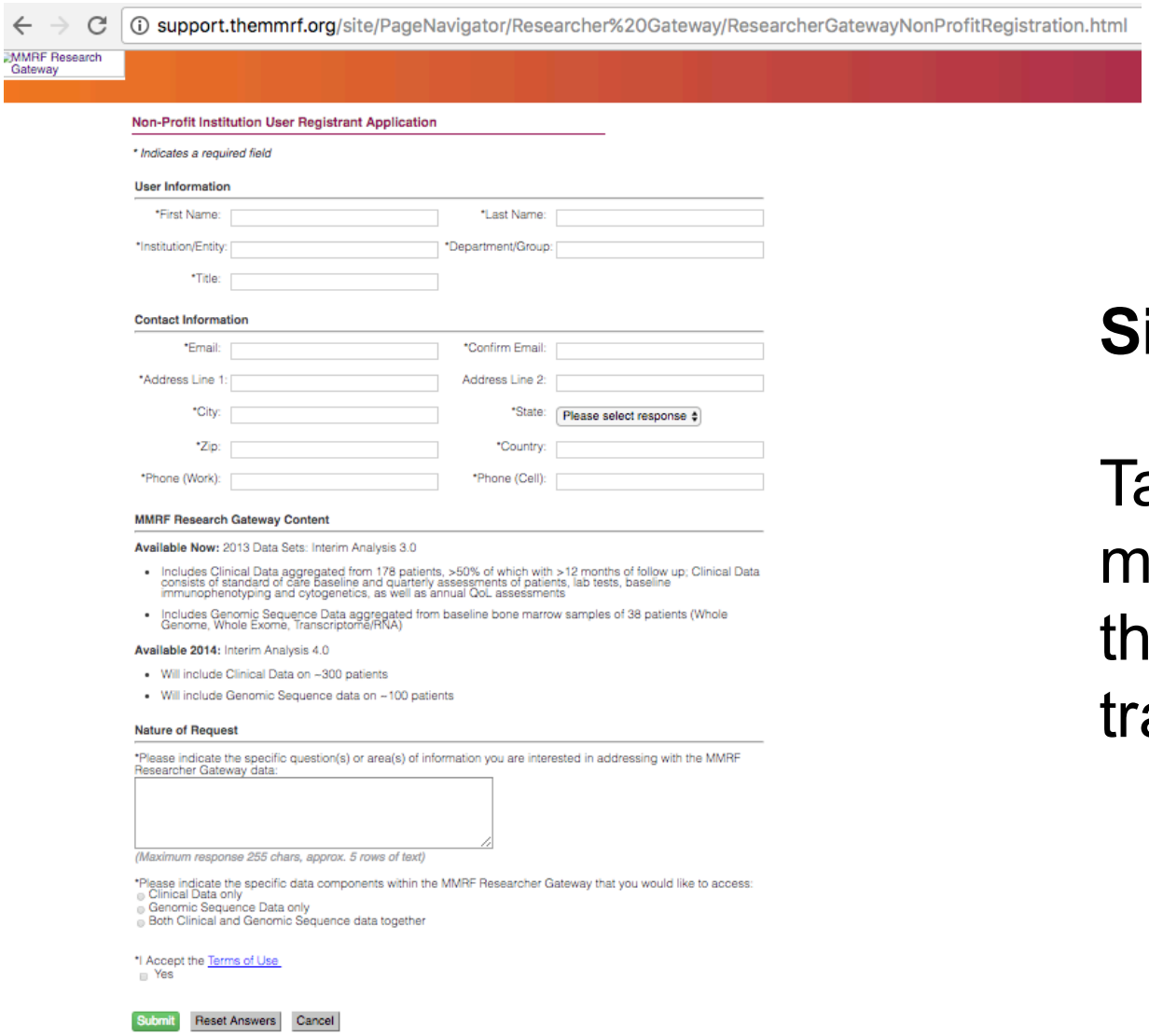
#5: Disease progression in multiple myeloma

Clinical mentor:

Nikhil Munshi, MD
Dana-Farber Cancer Institute
Professor of Medicine, Harvard Medical School



Data for multiple myeloma project



Simple form –

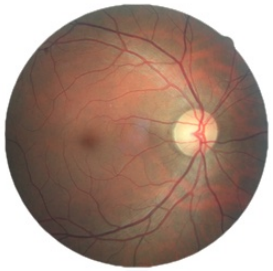
Takes just a few minutes to request the data, and no training needed

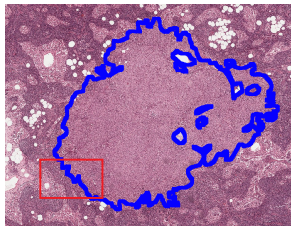
#6: Machine learning on medical images

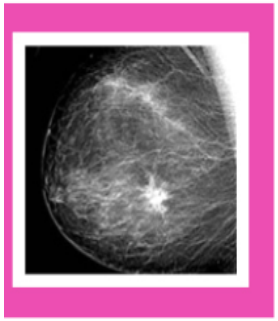
Clinical mentor:

Quanzheng Li, Ph.D.
Massachusetts General Hospital, Department of Radiology
Center for Clinical Data Science
Associate Professor, Harvard Medical School

- Led one of the top teams in Camelyon 2016 competition on cancer metastasis detection (pathology)
- Could use publicly available data and propose your own project in consultation with him







- **One project he proposed:**
  Study transfer learning using chest CT images from patients in two cohorts, emphysema and lung cancer