CSC 2541: Machine Learning for Healthcare

Lecture 3: Clinical Time Series Modelling

Professor Marzyeh Ghassemi, PhD
University of Toronto, CS/Med
Vector Institute

VECTOR INSTITUTE | INSTITUT VECTEUR

UNIVERSITY OF TORONTO

Course Reminders!

- Submit the weekly reflection questions to MarkUs!
- Sign up for a paper presentation slot!
- Homework 1 due next week!
- Think about your projects!

Logistics

- Course website:
  https://cs2541-ml4h2020.github.io
- Piazza:
  https://piazza.com/utoronto.ca/winter2020/csc2541
- Grading:
  - 20% Homework (3 problem sets)
  - 10% Weekly reflections on Markus (5 questions)
  - 10% Paper presentation done in-class (sign-up after the first lecture)
  - 60% course project (an eight-page write up)

Schedule

Jan 9, 2020, Lecture 1: Why is healthcare unique?

Jan 16, 2020, Lecture 2: Supervised Learning for Classification, Risk Scores and Survival

Jan 23, 2020, Lecture 3: Clinical Time Series Modelling

Jan 30, 2020, Lecture 4: Causal inference with Health Data --- Dr. Shalmali Joshi (Vector)

Problem Set 1 (Jan 31 at 11:59pm)

Feb 6, 2020, Lecture 5: Fairness, Ethics, and Healthcare

Project proposals (Feb 6 at 5pm)

Feb 13, 2020, Lecture 6: Deep Learning in Medical Imaging -- Dr. Joseph Paul Cohen (MILA)

Problem Set 2 (Feb 14 at 11:59pm)

Feb 20, 2020, Lecture 7: Clinical NLP and Audio -- Dr. Tristan Naumann (MSR)

Feb 27, 2020, Lecture 8: Clinical Reinforcement Learning

Mar 5, 2020, Lecture 9: Interpretability / Humans-In-The-Loop --- Dr. Rajesh Ranganath (NYU)

Problem Set 3 (Mar 6 at 11:59pm)

Mar 12, 2020, Lecture 10: Disease Progression Modelling/Transfer Learning -- Irene Chen (MIT)

Mar 19, 2020, Project Sessions/Lecture

Mar 26, 2020, Course Presentations

April 4, 2020, Course Presentations

Project Report (Apr 3 at 11:59pm)

Outline

1. What's Time Got To Do With It?
    a. Missingness
    b. Representation

2. Case Study 1: MTGPs for Mortality Prediction and TBI

3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. Project Discussion

Outline

1. **What's Time Got To Do With It?**
   a. Missingness
   b. Representation

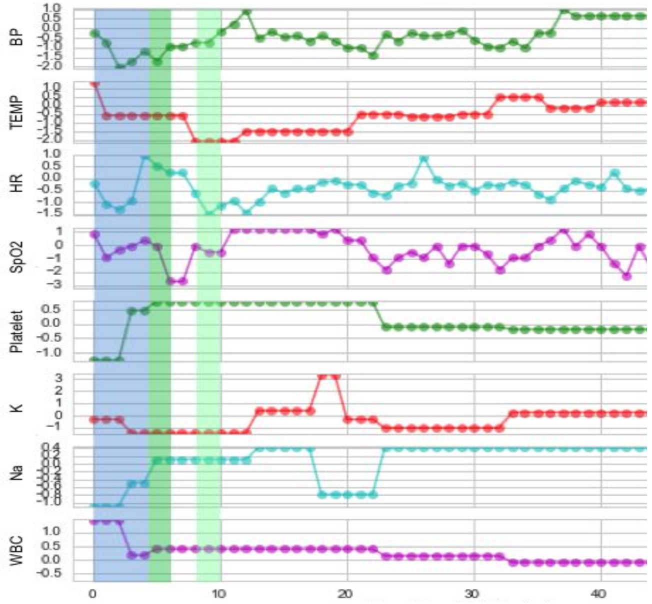
2. Case Study 1: MTGPs for Mortality Prediction and TBI

3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. Project Discussion

Problem: Hospital decision-making / care planning

Observe Patient Data

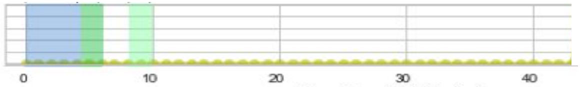


?

“Real-time” **Prediction**

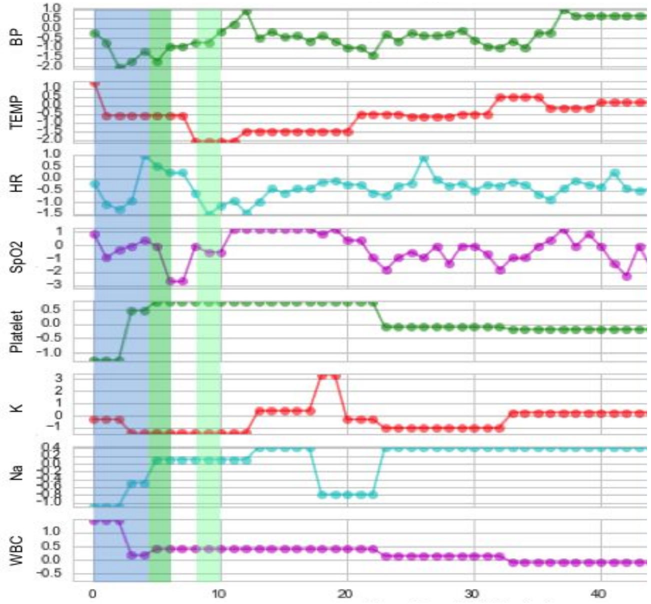
Of **{**Drug/Mortality/Condition**}**

By Gap Time



Problem: Hospital decision-making / care planning

Observe Patient Data

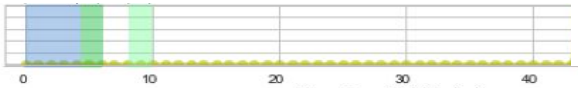


?

“Real-time” **Prediction**

Of **{**Drug/Mortality/Condition**}**

By Gap **Time**



How Do We Handle **Time**?

- An image gives a snapshot of an object, but a video dictates form!
- We want to model patient risks/treatments/outcomes as they **live**.
- Strategies:
  - Amortize - Make features out of mean, min, max, etc.
  - Stack - Inputs of fixed size, and concatenate.
  - Deal - Use a method that addresses dynamics.
- Focus on dealing in this lecture.

Outline

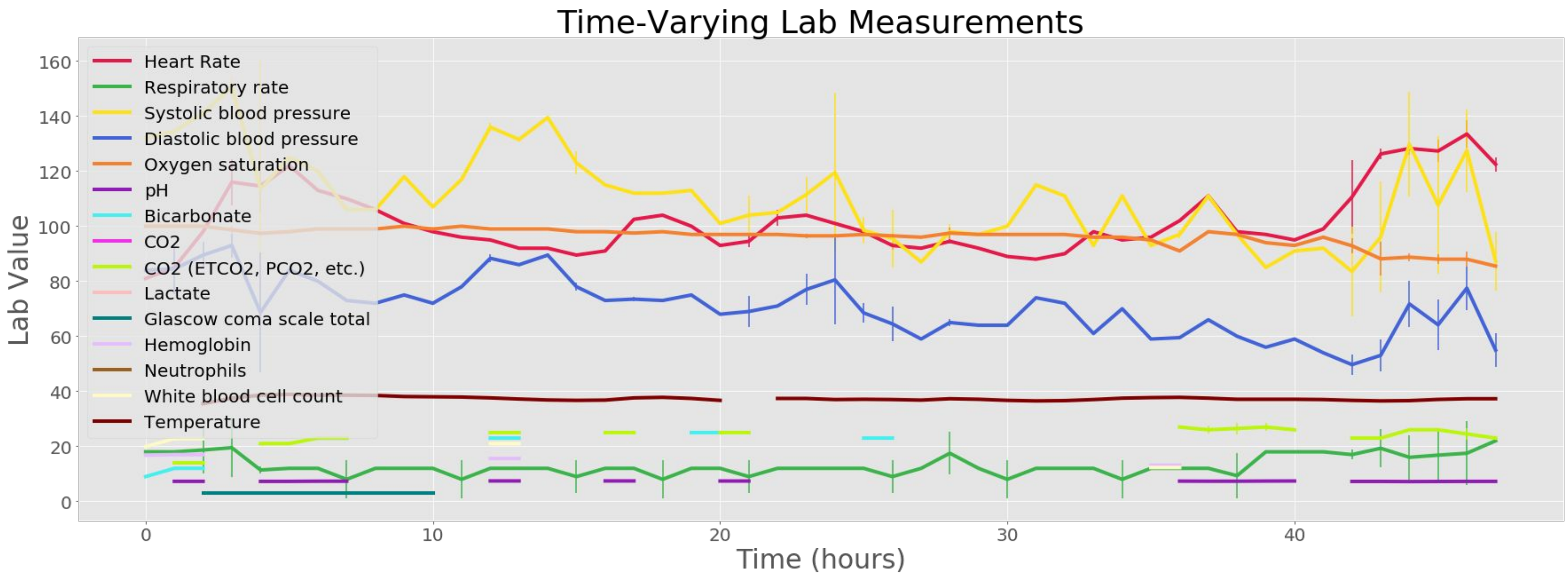
1. What's Time Got To Do With It?
   a. **Missingness**
   b. Representation

2. Case Study 1: MTGPs for Mortality Prediction and TBI

3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. Project Discussion

What is Missingness?



Missing Data Details

Data can be missing according to several regimes:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Missing Data Details

Data can be missing according to several regimes:

- Missing completely at random (MCAR)
  - The observed pattern of missingness is independent from the observed or missing values.
- Missing at random (MAR)
- Missing not at random (MNAR)

Missing Data Details

Data can be missing according to several regimes:

- Missing completely at random (MCAR)
  - The observed pattern of missingness is independent from the observed or missing values.
- Missing at random (MAR)
  - The observed pattern of missingness is independent from the missing values (but may depend on the observed values).
- Missing not at random (MNAR)

Missing Data Details

Data can be missing according to several regimes:

- Missing completely at random (MCAR)
  - The observed pattern of missingness is independent from the observed or missing values.
- Missing at random (MAR)
  - The observed pattern of missingness is independent from the missing values (but may depend on the observed values).
- Missing not at random (MNAR)
  - All bets are off.

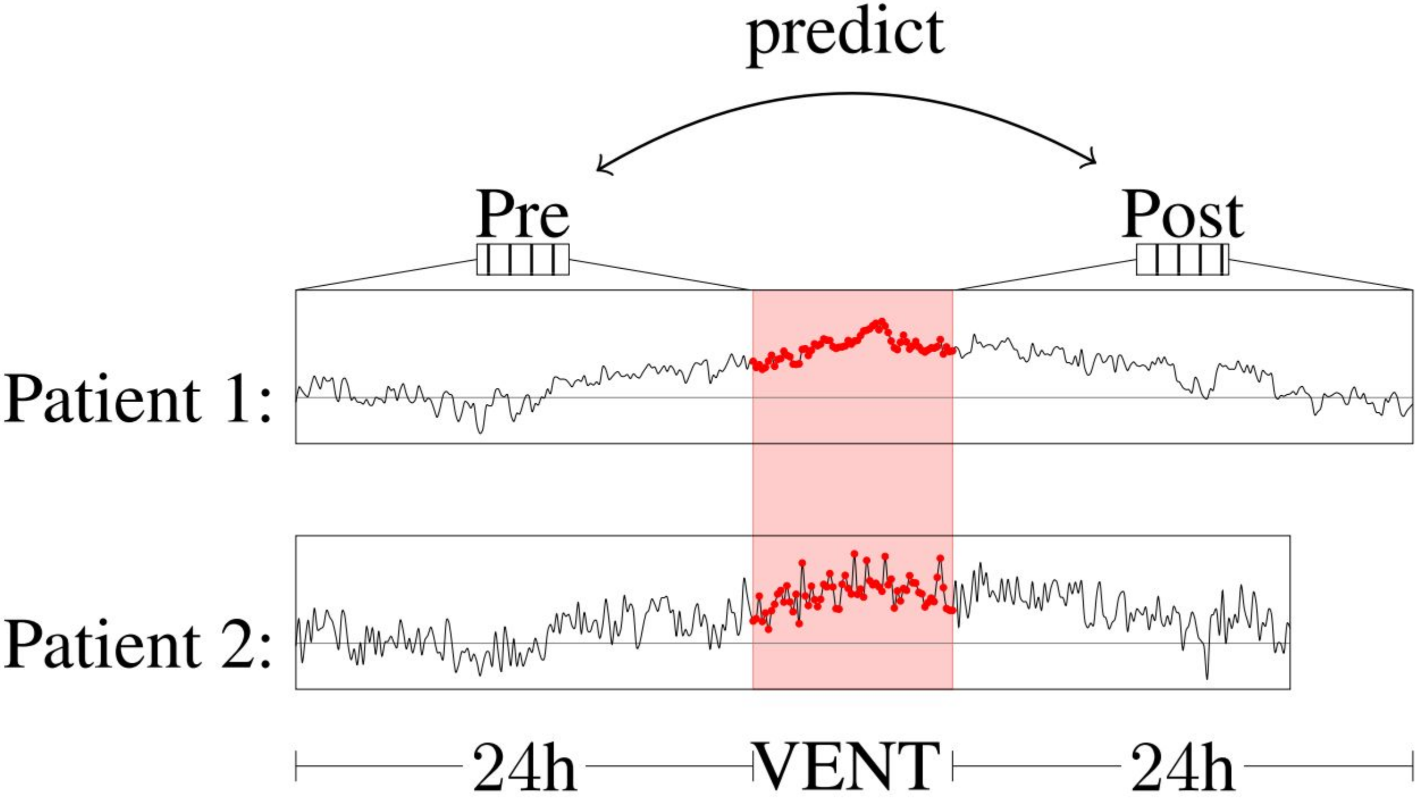
Missing Data Details

Data can be missing according to several regimes:

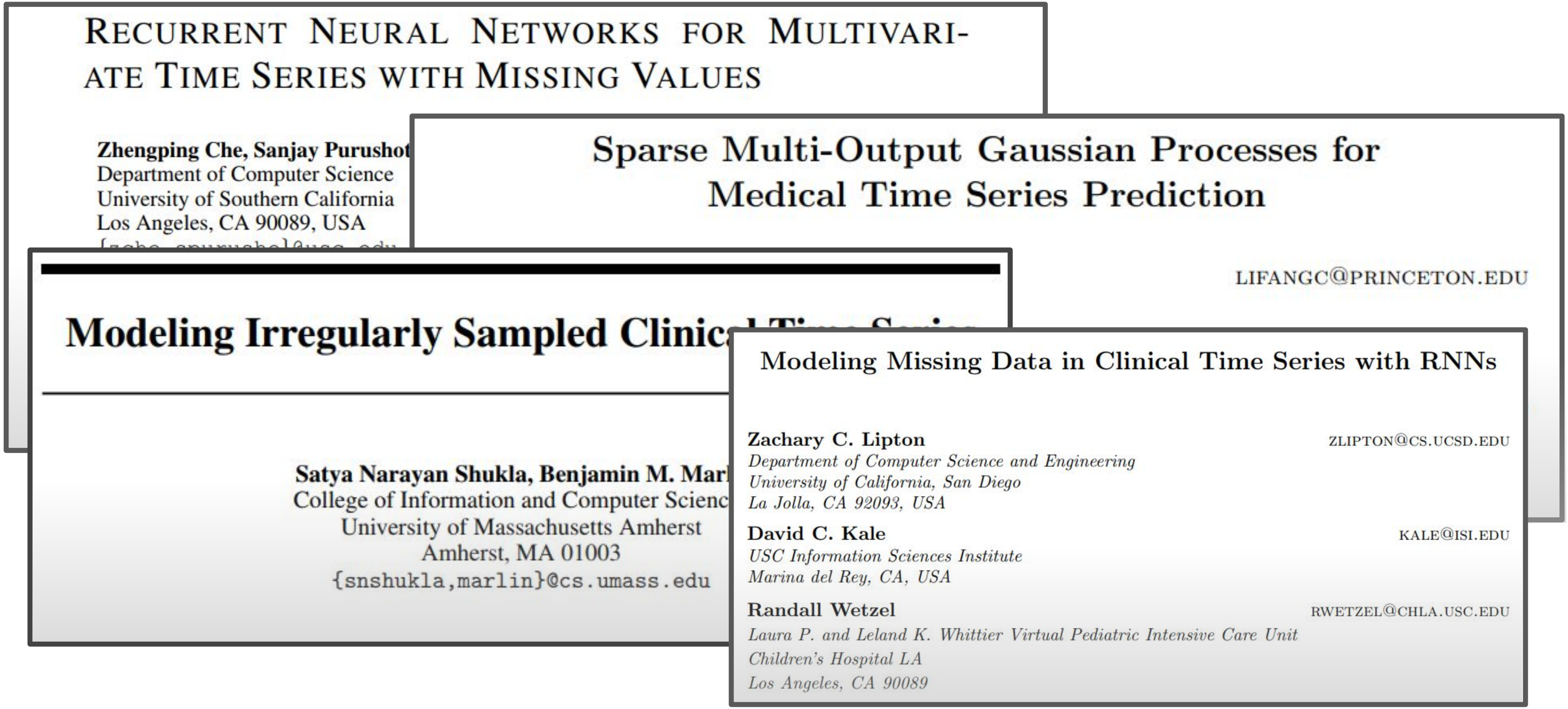
- Missing completely at random (MCAR)
    - The observed pattern of missingness is independent from the observed or missing values.
- Missing at random (MAR)
    - The observed pattern of missingness is independent from the missing values (but may depend on the observed values).
- Missing not at random (MNAR)
    - All bets are off.

Healthcare lives here.

Missing Data is Confounding



How do we handle missing data?



RECURRENT NEURAL NETWORKS FOR MULTIVARIATE TIME SERIES WITH MISSING VALUES

Zhengping Che, Sanjay Purushot
Department of Computer Science
University of Southern California
Los Angeles, CA 90089, USA

Sparse Multi-Output Gaussian Processes for Medical Time Series Prediction

LIFANGC@PRINCETON.EDU

Modeling Irregularly Sampled Clinic

Satya Narayan Shukla, Benjamin M. Marl
College of Information and Computer Scienc
University of Massachusetts Amherst
Amherst, MA 01003
{snshukla,marlin}@cs.umass.edu

Modeling Missing Data in Clinical Time Series with RNNs

Zachary C. Lipton ZLIPTON@CS.UCSD.EDU
Department of Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093, USA

David C. Kale KALE@ISI.EDU
USC Information Sciences Institute
Marina del Rey, CA, USA

Randall Wetzel RWETZEL@CHLA.USC.EDU
Laura P. and Leland K. Whittier Virtual Pediatric Intensive Care Unit
Children's Hospital LA
Los Angeles, CA 90089

Imputation

1. Statistical Timeseries Forecasting: ARMA/ARIMA/ARIMAX, etc.
2. Easy Baselines: Constant infilling, Sample & Hold (+ indicators), Interpolation
3. Traditional Imputation: MICE/3D-MICE, MissForest, Matrix/Tensor Completion
4. Gaussian Processes
5. Advanced neural methods (GRU-D, GANs, etc.)

Imputation

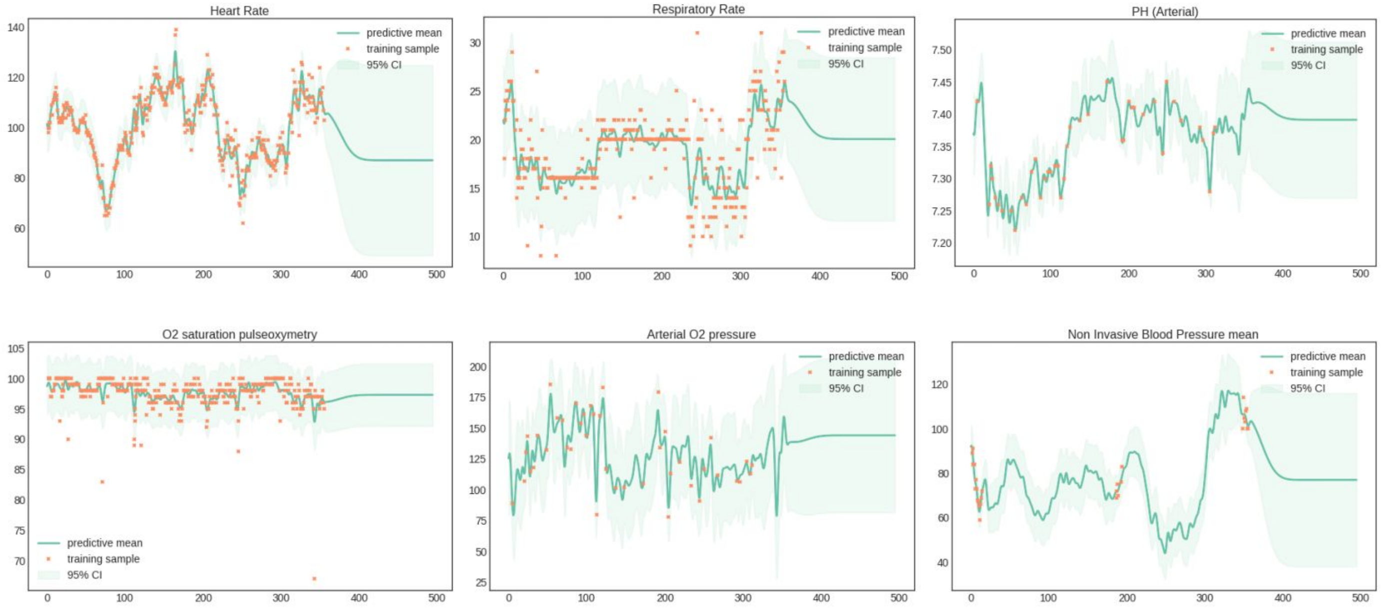


Figure 2: Example trajectories of six vital signs for a single admission, following imputation using Gaussian processes. Twelve vital signs are jointly modeled by the GP.

Prasad, Niranjani, et al. "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units." *arXiv preprint arXiv:1704.06300* (2017).

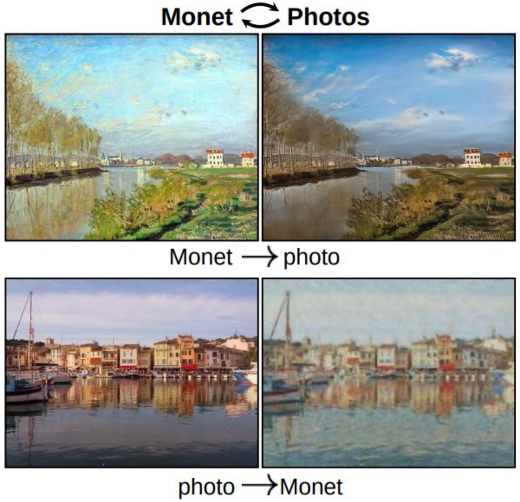
GANs for Imputation

GAIN: Missing Data Imputation using Generative Adversarial Nets

Jinsung Yoon [1 *] **James Jordon** [2 *] **Mihaela van der Schaar** [1 2 3]

GANs for Imputation

Figure 6. Qualitative comparisons with Deepfillv1 [18] on the CelebA-HQ validation sets.



Left: Jo, Youngjoo, and Jongyoul Park. "SC-FEGAN: Face Editing Generative Adversarial Network with User's Sketch and Color." arXiv preprint arXiv:1902.06838 (2019).
Middle: Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." Proceedings of the IEEE International Conference on Computer Vision. 2017.
Right: https://thispersondoesnotexist.com/

GAIN: Generative Adversarial Imputation

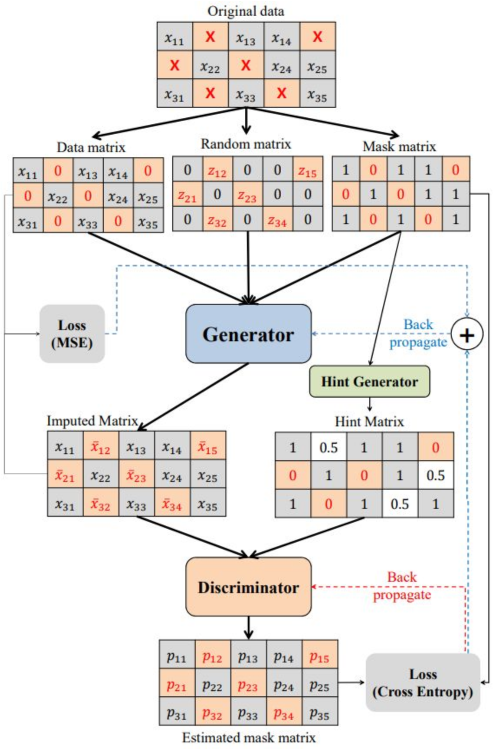


Figure 1. The architecture of GAIN

Imputation Papers

1. GAIN: https://arxiv.org/pdf/1806.02920.pdf
2. GRU-D: https://www.nature.com/articles/s41598-018-24271-9
3. GP Imputation: https://arxiv.org/pdf/1704.06300.pdf
4. Interpolation-prediction network: https://arxiv.org/pdf/1812.00531.pdf

Table 1: Performance on mortality and length of stay prediction tasks on MIMIC-III. Loss: Cross-Entropy Loss, MedAE: Median Absolute Error (in days), EV: Explained variance

Model	**Classification**			**Regression**	
	AUC	**AUPRC**	**Loss**	**MedAE**	**EV score**
Log/LinReg	0.772 ± 0.013	0.303 ± 0.018	0.240 ± 0.003	3.528 ± 0.072	0.043 ± 0.012
SVM	0.671 ± 0.005	0.300 ± 0.011	0.260 ± 0.002	3.523 ± 0.071	0.042 ± 0.011
AdaBoost	0.829 ± 0.007	0.345 ± 0.007	0.663 ± 0.000	4.517 ± 0.234	0.100 ± 0.012
RF	0.826 ± 0.008	0.356 ± 0.010	0.315 ± 0.025	3.113 ± 0.125	0.117 ± 0.035
GRU-M	0.831 ± 0.007	0.376 ± 0.022	0.220 ± 0.004	3.140 ± 0.196	0.131 ± 0.044
GRU-F	0.821 ± 0.007	0.360 ± 0.013	0.224 ± 0.003	3.064 ± 0.247	0.126 ± 0.025
GRU-S	0.843 ± 0.007	0.376 ± 0.014	0.218 ± 0.005	2.900 ± 0.129	0.161 ± 0.025
GRU-D	0.835 ± 0.013	0.359 ± 0.025	0.225 ± 0.009	$\mathbf{2.891 \pm 0.103}$	0.146 ± 0.051
Proposed	$\mathbf{0.853 \pm 0.007}$	$\mathbf{0.418 \pm 0.022}$	$\mathbf{0.210 \pm 0.004}$	$\mathbf{2.862 \pm 0.166}$	$\mathbf{0.245 \pm 0.019}$

Opportunities

1. Improved imputation methods. How do forecasting, GP, or adversarial methods compare to GRU-D/interpolation prediction network? Can we incorporate uncertainty offered by GPs usefully into downstream tasks? Can we make other models offer uncertainty?
2. Can we model the decision process by which clinicians choose what to measure and what to omit? How would this be helpful in downstream tasks? Can this help account for the MNAR nature of healthcare missingness?
3. Can we control for the confounding effects of missingness? Can we learn a model on underlying physiology from retrospective, care-byproduct data?

Outline

1. What's Time Got To Do With It?
   a. Missingness
   b. Representation

2. Case Study 1: MTGPs for Mortality Prediction and TBI

3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. Project Discussion

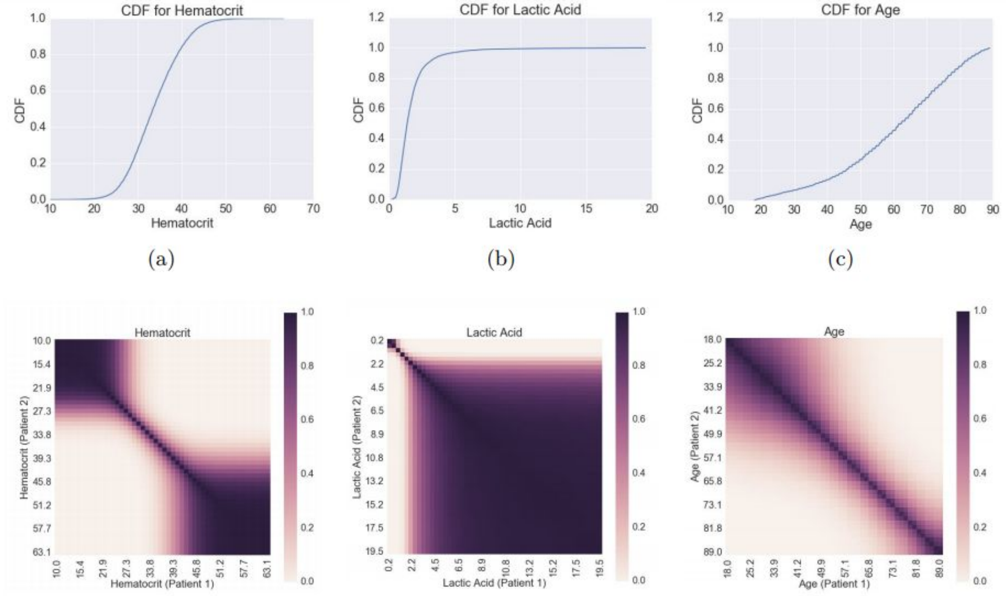
Representation: Why do we care?

Representations define a notion of “similarity”

Closer in “Conceptual Space”

Closer in “Pixel Space”

Representations learn a notion of similarity



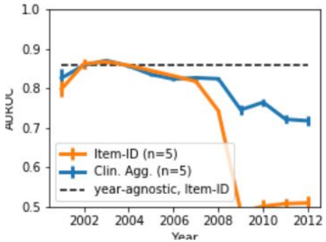
(a) (b) (c)

(d) Kernel on Hematocrit (e) Kernel on Lactic Acid (f) Kernel on Patient Age

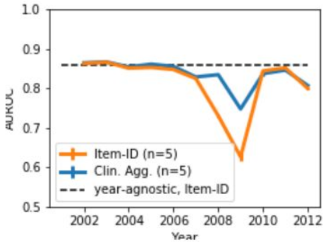
Figure 1: Examples of the kernel $k_{j,c}(x, z)$ in (1) with $c = 5$ on three features evaluated on adult ICU population: Hematocrit, Lactic Acid, and Patient Age

Conroy, Bryan, Minnan Xu-Wilson, and Asif Rahman. "Patient Similarity Using Population Statistics and Multiple Kernel Learning." Machine Learning for Healthcare Conference. 2017.

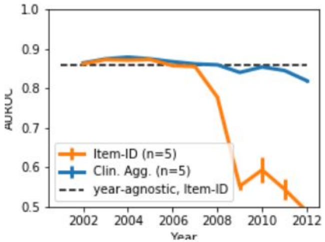
Representations can stabilize changing data



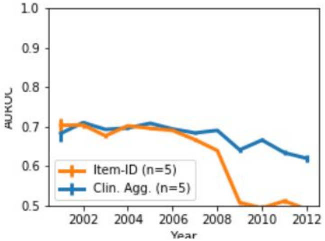
(a) Mortality AUC, models trained on 2001-2002 data.



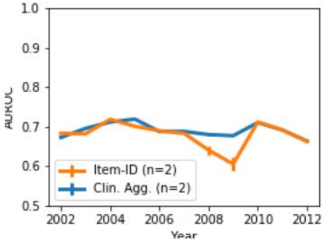
(b) Mortality AUC, models trained yearly on prior year only.



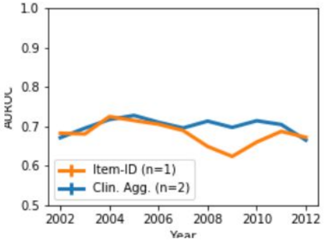
(c) Mortality AUC, models trained yearly on all prior data.



(d) LOS AUC, models trained on 2001-2002 data.



(e) LOS AUC, models trained yearly on prior year only.

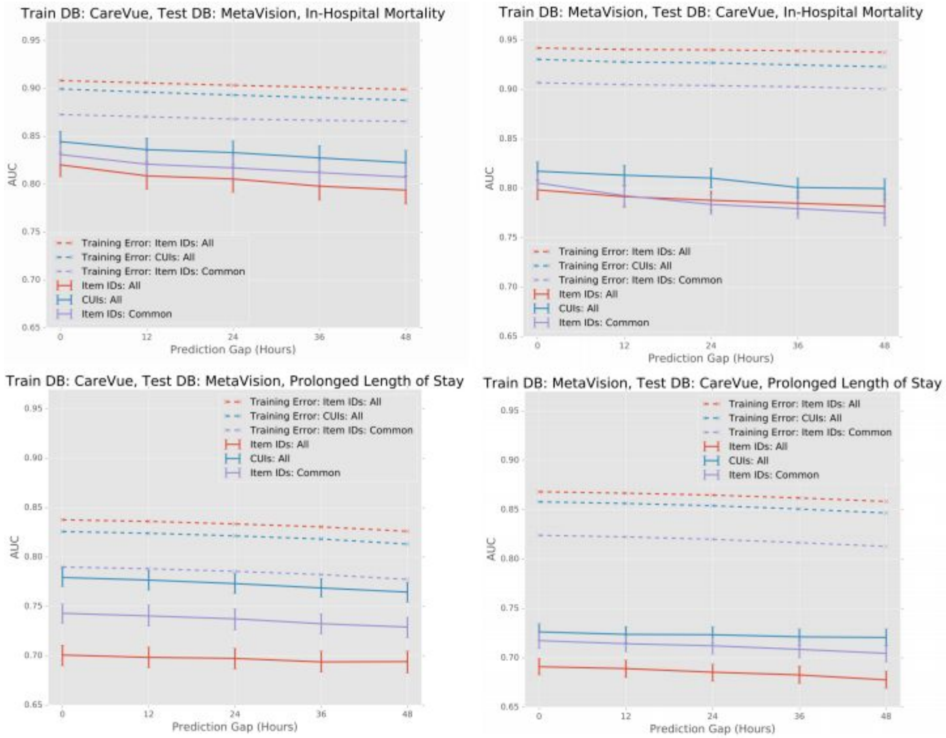


(f) LOS AUC, models trained yearly on all prior data.

Figure 1: Performance of RF classifiers using `Item-Id` and `Clinically Aggregated` representations on mortality (top) and LOS prediction (bottom). Error bars indicate ± standard error.

Nestor, Bret, et al. "Rethinking clinical prediction: Why machine learning must consider year of care and feature aggregation." Machine Learning for Healthcare Conference. 2019

Representations can stabilize changing data



Gong, Jen J., et al. "Predicting clinical outcomes across changing electronic health record systems." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

Representations can join disparate modalities

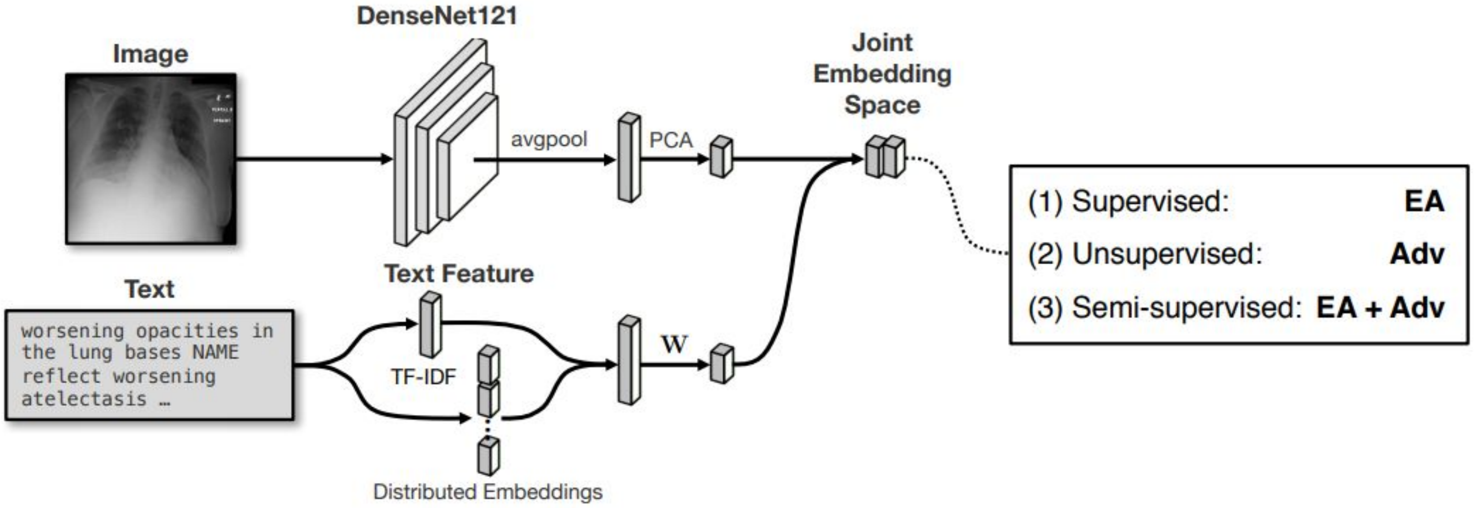


Figure 1: The overall experimental pipeline. EA: embedding alignment; Adv: adversarial training.

Hsu, Tzu-Ming Harry, et al. "Unsupervised multimodal representation learning across medical images and reports." *arXiv preprint arXiv:1811.08615* (2018).

DeepCluster: Why bother with labels?

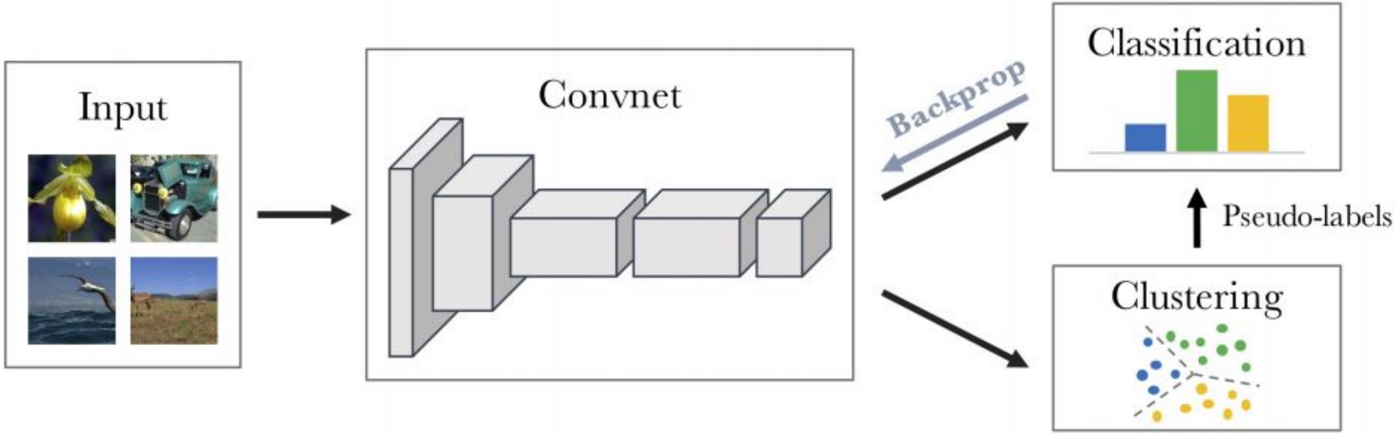


Fig. 1: Illustration of the proposed method: we iteratively cluster deep features and use the cluster assignments as pseudo-labels to learn the parameters of the convnet.

Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

Representation Learning in Action: Multitask Learning

Multi-task Prediction of Disease Onsets from Longitudinal Lab Tests

Narges Razavian, Jake Marcus, David Sontag
Courant Instit
{

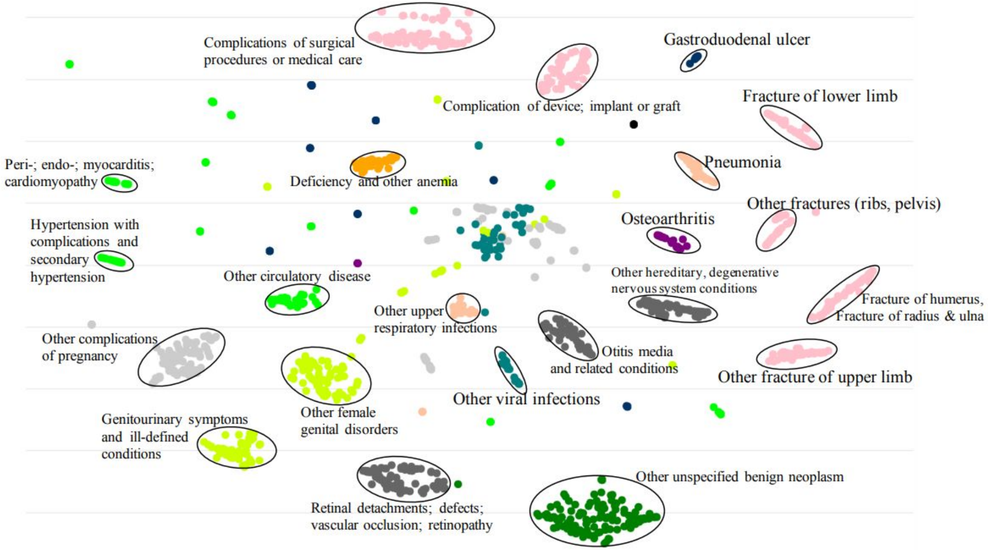
Multitask Learning and Benchmarking with Clinical Time Series Data

Hrayr Harutyunyan[1], Hrant Khachatrian[2,3], David C. Kale[1], Greg Ver Steeg[1], and Aram Galstyan[1]
nited States of America

MoleculeNet: a benchmark for molecular machine learning†

Zhenqin Wu,‡[a] Bharath Ramsundar,‡[b] Evan N. Feinberg,§[c] Joseph Gomes,§[a] Caleb Geniesse,[c] Aneesh S. Pappu,[b] Karl Leswing[d] and Vijay Pande*[a]

Representation Learning in Action: Clustering



(a) Scatterplot of the final representations $\mathbf{g}_i$'s of GRAM+

Choi, Edward, et al. "GRAM: graph-based attention model for healthcare representation learning." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.

Representation Learning in Action: Clustering

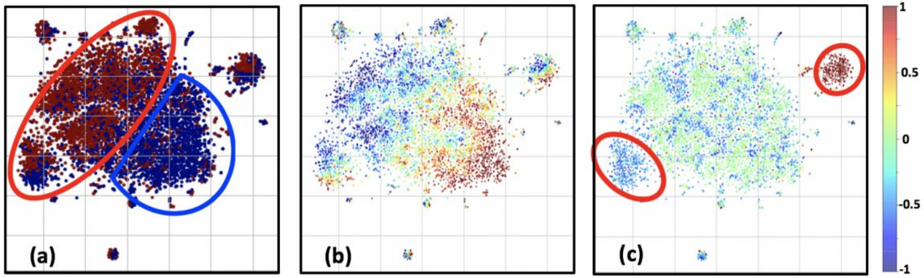


Figure 3: tSNE on context vectors of test dataset from BSS model colored by (a) red: positive examples and blue: negative examples, (b) average systemic diastolic blood pressure; and (c) average central venous pressure.

Dhamala, Jwala, et al. "Multivariate Time-Series Similarity Assessment via Unsupervised Representation Learning and Stratified Locality Sensitive Hashing: Application to Early Acute Hypotensive Episode Detection." *IEEE Sensors Letters* 3.1 (2019): 1-4.

Representation Learning in Action: Anomaly Detection

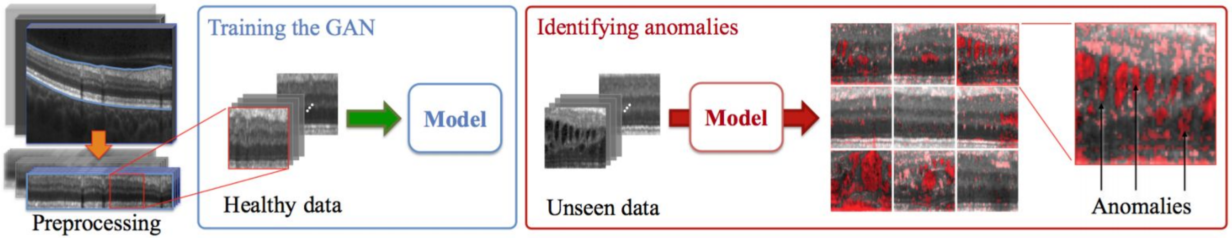


Fig. 1. Anomaly detection framework. The preprocessing step includes extraction and flattening of the retinal area, patch extraction and intensity normalization. Generative adversarial training is performed on healthy data and testing is performed on both, unseen healthy cases and anomalous data.

Schlegl, Thomas, et al. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery." *International Conference on Information Processing in Medical Imaging*. Springer, Cham, 2017.

Representation Learning in Action: Anomaly Detection

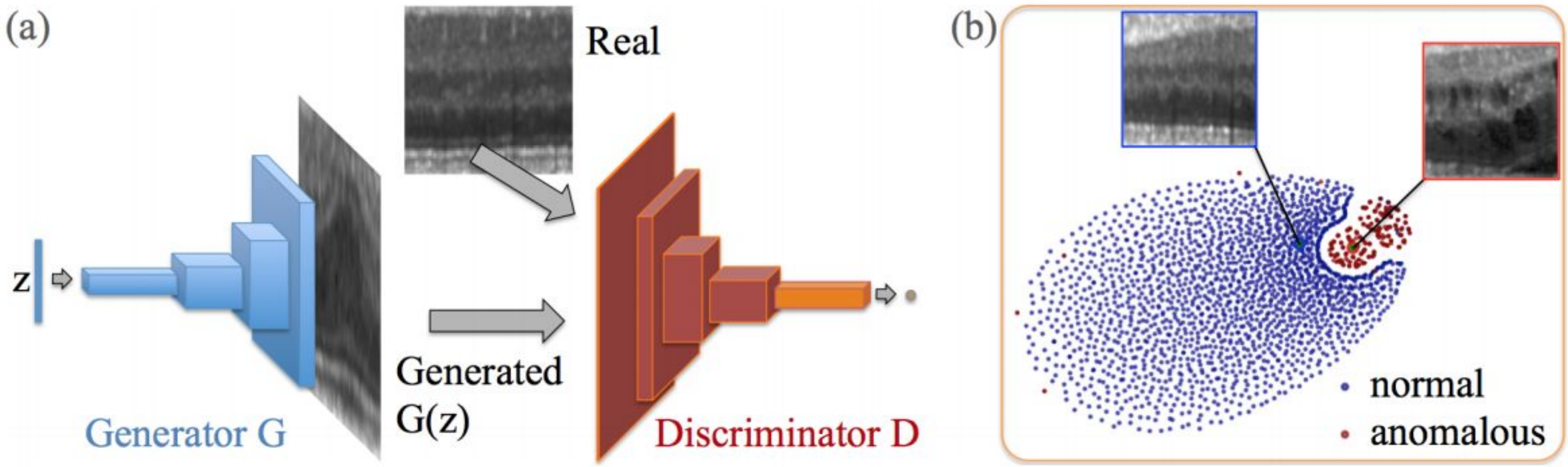


Fig. 2. (a) Deep convolutional generative adversarial network. (b) t-SNE embedding of normal (blue) and anomalous (red) images on the feature representation of the last convolution layer (orange in (a)) of the discriminator.

Key Points for Healthcare

- Representations can normalize.
- Generalization to unseen tasks is critical (e.g., patient subtyping).
- Representations can aid in interpretability.
- Representations can span many modalities.

Outline

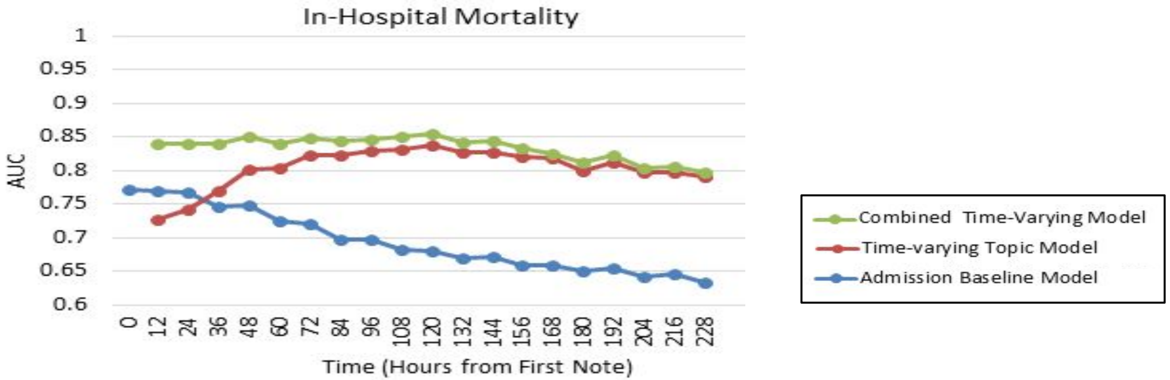
1. What's Time Got To Do With It?
    a. Missingness
    b. Representation

2. **Case Study 1: MTGPs for Mortality Prediction and TBI**

3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. Project Discussion

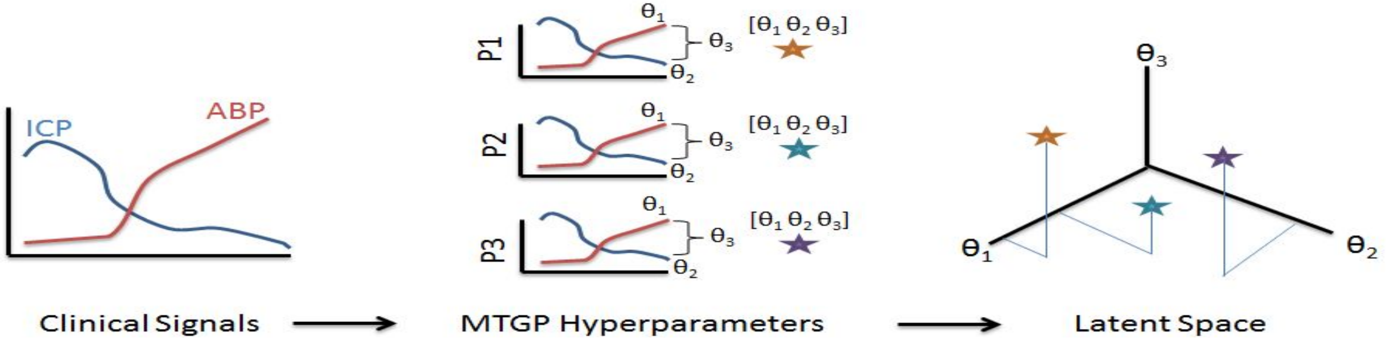
Remember This? Topics Improves Mortality Prediction



- **Forward-facing ICU mortality** prediction with notes.
- **Latent** representations **add** predictive power.
- Topics enable accurately **assess risk** from **notes.**

Add Information About Evolution of Signals

- Learn a new latent representation to evaluate multi-dimensional function similarity (θ).



MTGP models capture movements within and between signals.

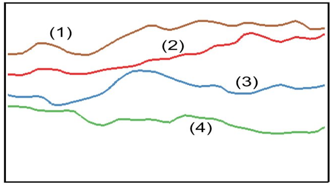
Transform signals into MTGP hyperparameter representation.

Compare patient similarly in the new representation.

Learning Single Task Gaussian Processes (STGP)

- Model each **signal** as a GP **task** with mean and covariance **functions**.

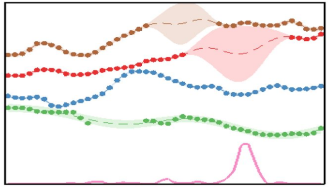
$$\tilde{\mathbf{y}}_{\mathbf{n}} = g(\vec{x}_n) \sim \mathcal{GP}\Big(m(\vec{x}_n), k(\vec{x}_n, \vec{x}'_n)\Big)$$



- GP's commonly used to predict at new indices.

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) \sim \mathcal{N}\Big(m(\mathbf{y}^*),\ \mathrm{var}(\mathbf{y}^*)\Big)$$

$$m(\mathbf{y}^*) = \mathbf{K}(\mathbf{x}, \mathbf{x}^*)^\top \mathbf{K}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{y}$$
$$\mathrm{var}(\mathbf{y}^*) = \mathbf{K}(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{K}(\mathbf{x}, \mathbf{x}^*)^\top \mathbf{K}(\mathbf{x}, \mathbf{x})^{-1}\mathbf{K}(\mathbf{x}, \mathbf{x}^*).$$



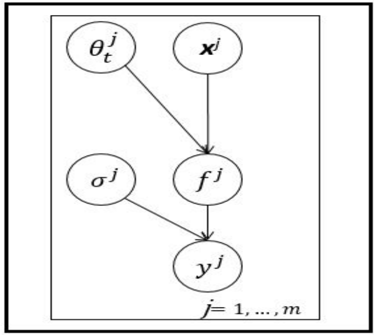
- Learn the parameters (θ) of the **kernel** from **data**.

$$\begin{aligned}\mathrm{NLML} &= -\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \\ &= \frac{1}{2}\log|\mathbf{K}| + \frac{1}{2}\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y} + \frac{n}{2}\log(2\pi)\end{aligned}$$

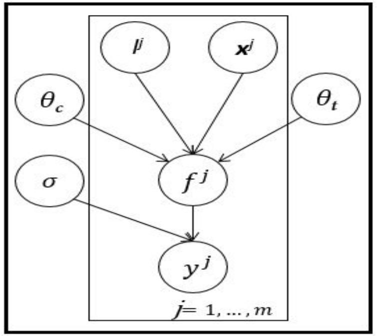
Single vs. Multi-task Gaussian Processes

- Assume we have *m* sets of:
  - Inputs X^i
  - Temporal covariance hyperparameters θ^i_t
  - Estimated functions f^i
  - Noise terms σ^i
  - Outcomes y^i

- We can train *m* single-task Gaussian process (STGP) (a) or a multi-task Gaussian process (MTGP) to relate the *m* tasks through all prior variables, with the tasks' labels *l* and similarity matrix θ_c (b).



(a)



(b)

Learning MTGPs As Representations

- Use an MTGP representation to relate *m* inputs through K_t and K_c.

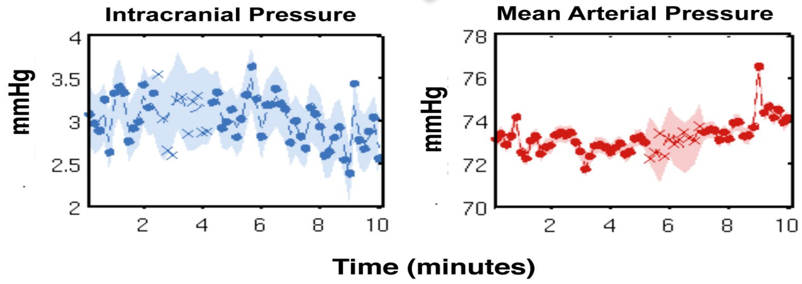
$$\mathbf{K}_{MT}(\mathbf{X}_n, \mathbf{l}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_t) = \mathbf{K}_c(\mathbf{l}, \boldsymbol{\theta}_c) \otimes \mathbf{K}_t(\mathbf{X}_n, \boldsymbol{\theta}_t)$$

Movement **between** signals

$$\mathbf{K}_c = \mathbf{L}\mathbf{L}^\top \qquad \mathbf{L} = \begin{bmatrix} \theta_{c,1} & 0 & \cdots & 0 \\ \theta_{c,2} & \theta_{c,3} & & 0 \\ \vdots & & \ddots & \vdots \\ \theta_{c,k-m+1} & \theta_{c,k-m+2} & \cdots & \theta_{c,k} \end{bmatrix}$$

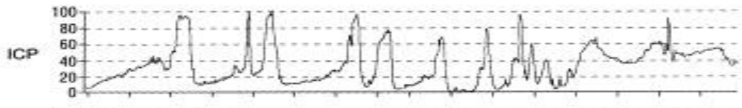
Movement **within** a signal

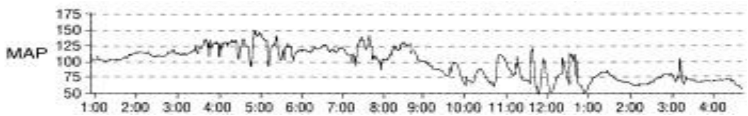
$$\mathbf{K}_t = \theta_A^2 \exp\left\{-\frac{\| x - x' \|^2}{2\theta_L^2}\right\}$$



[1] Bonilla, Edwin V., Kian M. Chai, and Christopher Williams. "Multi-task Gaussian process prediction." *Advances in neural information processing systems*. 2007.
[2] Carl Rasmussen's minimize.m was used for gradient-based optimization of the marginal likelihood.

Estimating Signal in Traumatic Brain Injury Patients





- Intracranial pressure (ICP) and mean arterial blood pressure (ABP) are important indicators of cerebrovascular autoregulation (CA) in traumatic Brain Injury (TBI) patients.

- CA sustains adequate cerebral blood flow[1] and impairment risks secondary brain damage and mortality.[2]

- CA is assessed using a sliding window Pearson's correlation between the ICP and ABP – the Pressure-Reactivity Index (PRx)[3].

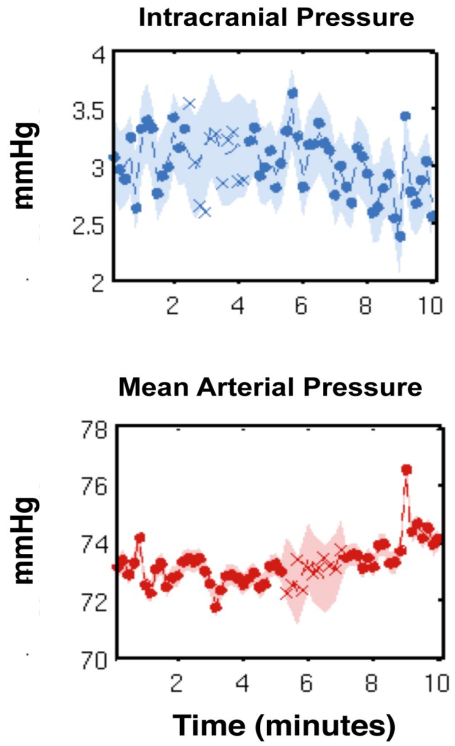
[1] Werner, C., and K. Engelhard. "Pathophysiology of traumatic brain injury." *British journal of anaesthesia* 99.1 (2007): 4-9.
[2] Hlatky, Roman, Alex B. Valadka, and Claudia S. Robertson. "Intracranial pressure response to induced hypertension: role of dynamic pressure autoregulation." *Neurosurgery* 57.5 (2005): 917-923.
[3] Czosnyka, Marek, et al. "Continuous assessment of the cerebral vasomotor reactivity in head injury." *Neurosurgery* 41.1 (1997): 11-19.

TBI Estimation Methodology

- PRx isn't calculated when either signal is contaminated - evaluate STGPs/MTGPs for interpolation, and MTGPs for PRx estimation.

- Collected data from 35 TBI patients with 24+ hours of ICP and ABP recordings sampled every 10 seconds.

- Selected 30 ten-minute windows where ICP/ABP were free from artifacts and missing values from each patient recording; randomly introduced artificial gaps in both signals (*x*'s).



MTGP Representations Improve Signal Forecasting and Outcome Prediction

Performance on Signal Forecasting

Signal	Measure	STGP	MTGP
ICP	RMSE	0.91	0.69
	MSLL	0.6	0.45
ABP	RMSE	2.77	1.98
	MSLL	0.65	0.55

- MTGPs outperform STGPs in signal reconstruction.
- Automatically estimate cerebrovascular autoregulation.

Performance on Mortality Prediction

Features	Hospital Mortality
Ave. Topics	0.759
SAPS-I + MTGP	0.775
Ave. Topics + MTGP	0.788
SAPS-I + Ave. Topics + MTGP	0.812

- MTGP hyperparameter representations improve short-term mortality prediction.

* Final cohort consisted of 10,202 patients, with 313,461 notes.

Outline

1. What's Time Got To Do With It?
    a. Missingness
    b. Representation

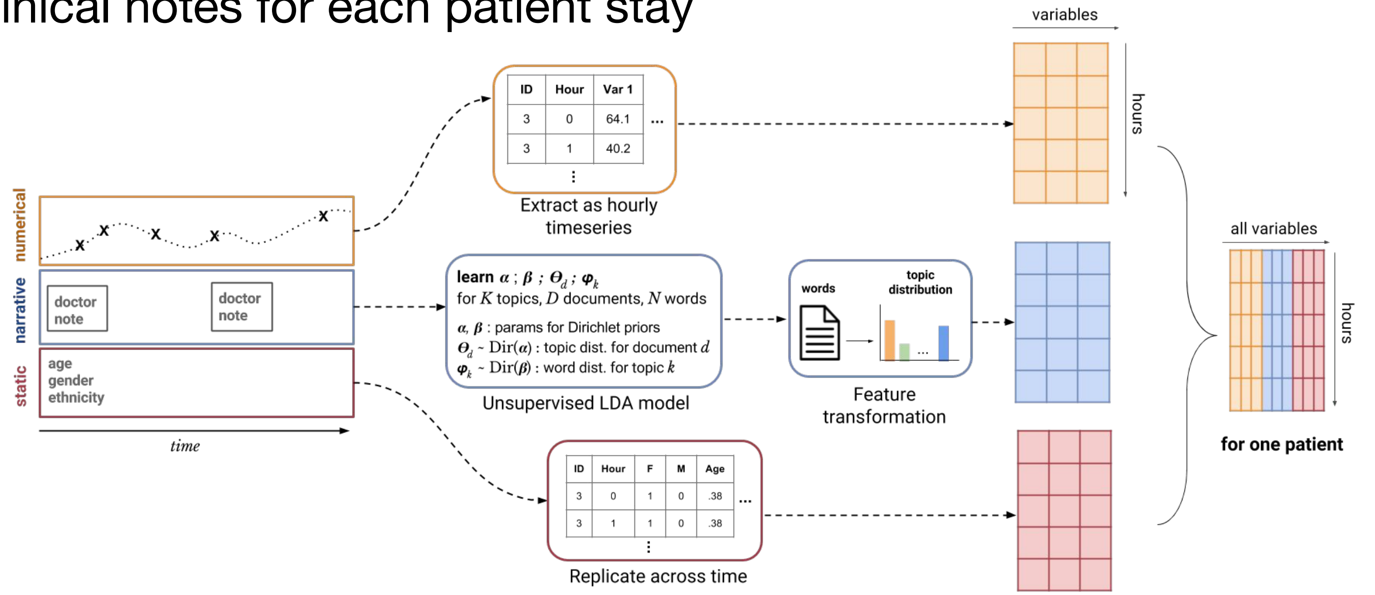
2. Case Study 1: MTGPs for Mortality Prediction and TBI

3. **Case Study 2: RNNs/CNNs for Intervention Onset Prediction**

4. Project Discussion

Can We Predict Interventions?

- 34,148 ICU patients from MIMIC-III
- 5 static variables (gender, age, etc.)
- 29 time-varying vitals and labs (oxygen saturation, lactate, etc.)
- All clinical notes for each patient stay



Raw Physiology vs “Words” Embedding

Numerical

patient	hours in	glucose
3	1	NaN
3	2	NaN
3	3	101.2344
⋮	⋮	⋮

Physiological Words

patient	hours in	glucose_-2	glucose_-1	glucose_0	glucose_1	glucose_2
3	1	0	0	0	0	0
3	2	0	0	0	0	0
3	3	0	1	0	0	0
	⋮	⋮	⋮ ⋮	⋮	⋮	

- Many values are missing!

Raw Physiology vs “Words” Embedding

Numerical

patient	hours in	glucose
3	1	NaN
3	2	NaN
3	3	101.2344
⋮	⋮	⋮

Physiological Words

patient	hours in	glucose_-2	glucose_-1	glucose_0	glucose_1	glucose_2
3	1	0	0	0	0	0
3	2	0	0	0	0	0
3	3	0	1	0	0	0
	⋮	⋮	⋮ ⋮	⋮	⋮	

-1

- Many values are missing!
- Z-score existing variables, rounding to the nearest int.

Raw Physiology vs “Words” Embedding

Numerical

patient	hours in	glucose
3	1	NaN
3	2	NaN
3	3	101.2344
⋮	⋮	⋮

Physiological Words

patient	hours in	glucose_-2	glucose_-1	glucose_0	glucose_1	glucose_2
3	1	0	0	0	0	0
3	2	0	0	0	0	0
3	3	0	1	0	0	0
	⋮	⋮	⋮ ⋮	⋮	⋮	

-1

- Many values are missing!
- Z-score existing variables, rounding to the nearest int.
- Convert each z-score into its own binary column.

Raw Physiology vs “Words” Embedding

Numerical

patient	hours in	glucose
3	1	NaN
3	2	NaN
3	3	101.2344
⋮	⋮	⋮

Physiological Words

patient	hours in	glucose_-2	glucose_-1	glucose_0	glucose_1	glucose_2
3	1	0	0	0	0	0
3	2	0	0	0	0	0
3	3	0	1	0	0	0
	⋮	⋮	⋮ ⋮	⋮	⋮	

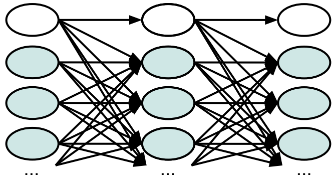
-1

A row of all zeros indicates
a missing value at that hour.

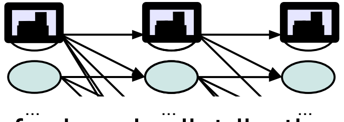
- Many values are missing!
- Z-score existing variables, rounding to the nearest int.
- Convert each z-score into its own binary column.

Many Ways to Model, What Do We Learn?

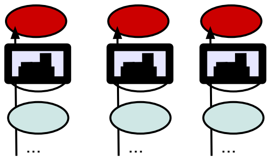
SSAM



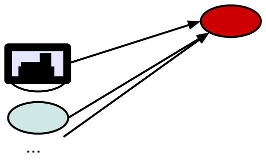
Learn model parameters over patients with variational EM.



Infer hourly distribution over hidden states with HMM DP (fwd alg.).

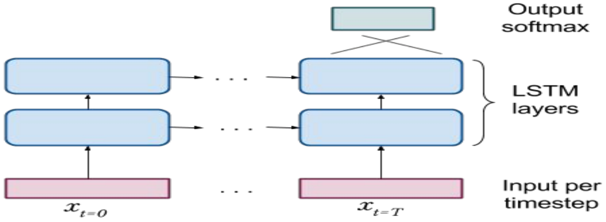


Logistic regression (with label-balanced cost function)



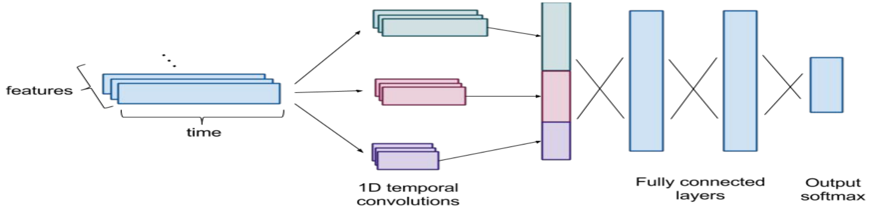
Predict onset in advance

LSTM



2 Layer/512 node LSTM with sequential hourly data; at end of window, use the final hidden state to predict output.

CNN



CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.

55

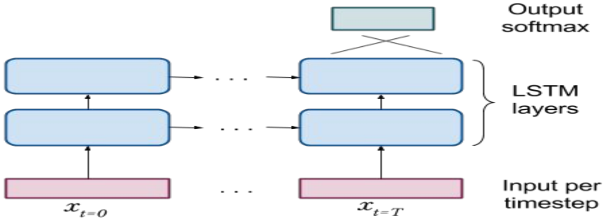
Many Ways to Model, What Do We Learn?

SSAM

Covered **last week**!

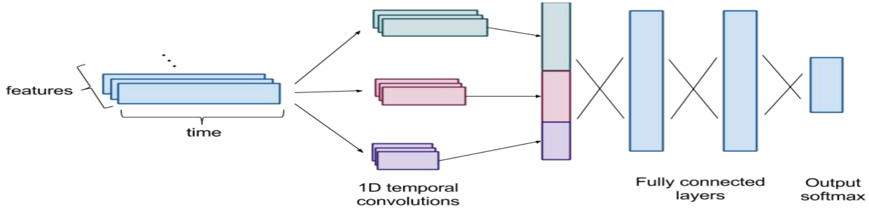
variational EM.

LSTM



2 Layer/512 node LSTM with sequential hourly data; at end of window, use the final hidden state to predict output.

CNN



CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.

56

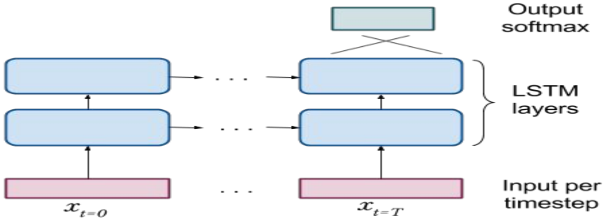
Many Ways to Model, What Do We Learn?

SSAM

Covered **last week**!

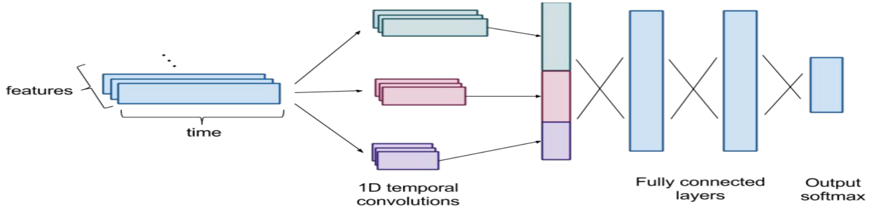
variational EM.

LSTM



2 Layer/512 node LSTM with sequential hourly data; at end of window, use the final hidden state to predict output.

CNN



CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.

57

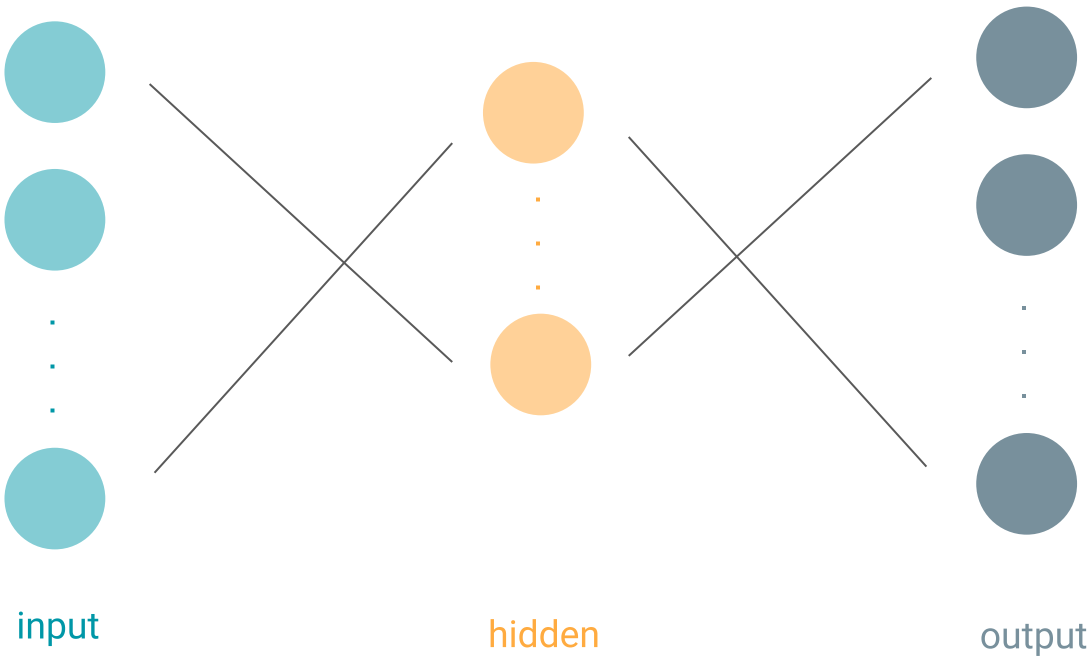
RNNs on Sequences

To model sequences, we need:

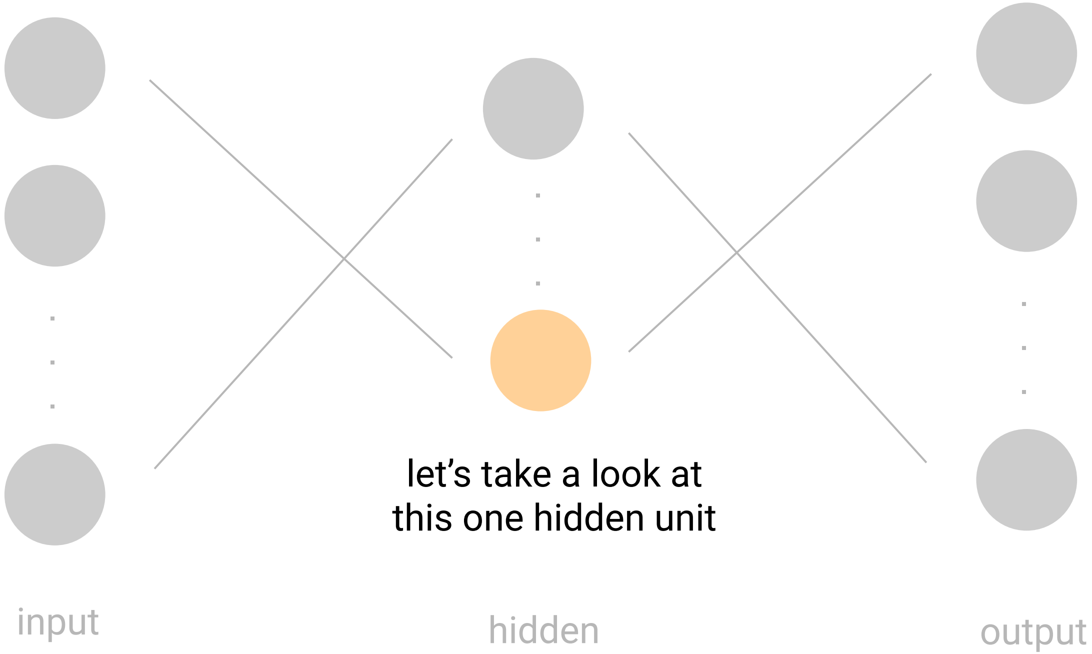
1. To deal with variable-length sequences
2. To maintain sequence order
3. To keep track of long-term dependencies
4. To share parameters across the sequence

Let's turn to **recurrent neural networks**.

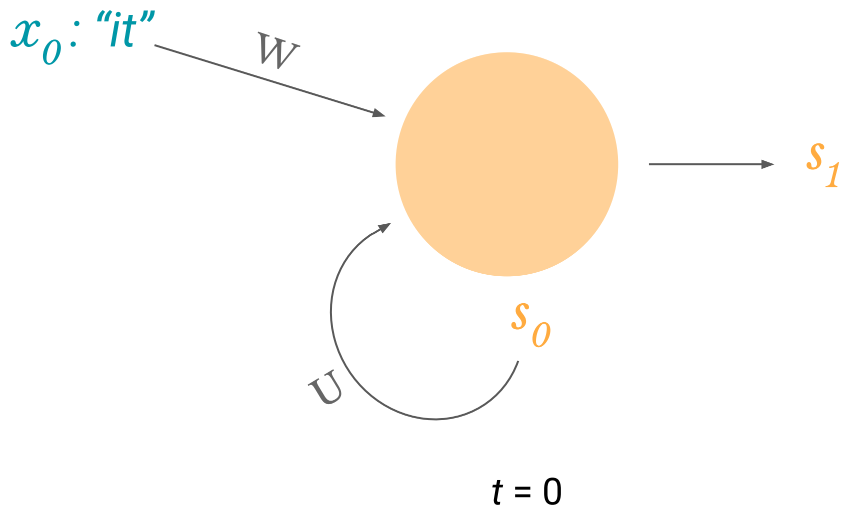
Example Network



Example Network



RNNS **remember** their previous state:

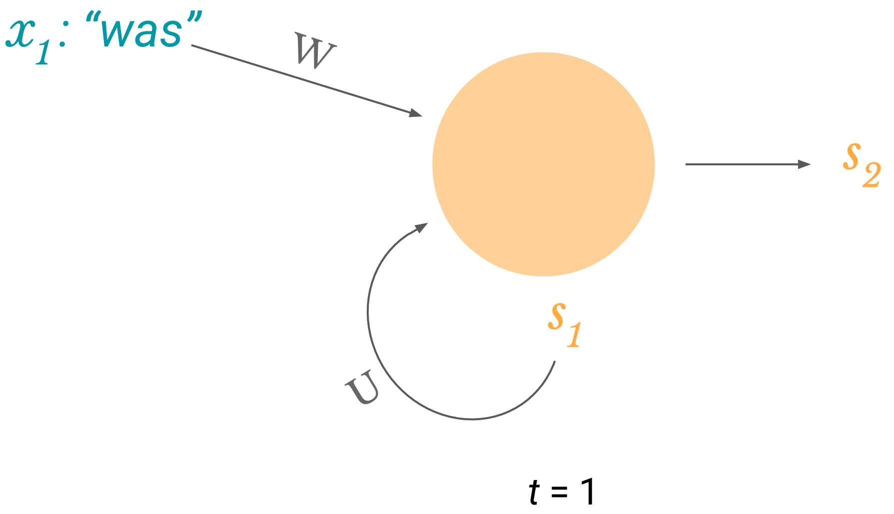


x_0 : vector representing first word
s_0 : cell state at $t = 0$ (some initialization)
s_1 : cell state at $t = 1$

$$s_1 = tanh(Wx_0 + Us_0)$$

$W,\ U$: weight matrices

RNNS **remember** their previous state:



x_1 : vector representing second word

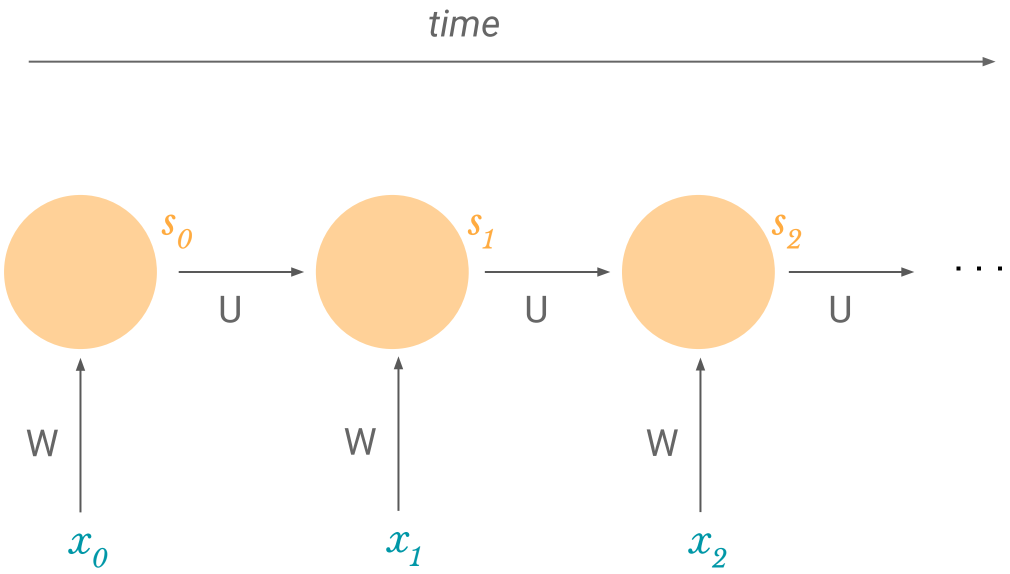
s_1 : cell state at $t = 1$

s_2 : cell state at $t = 2$

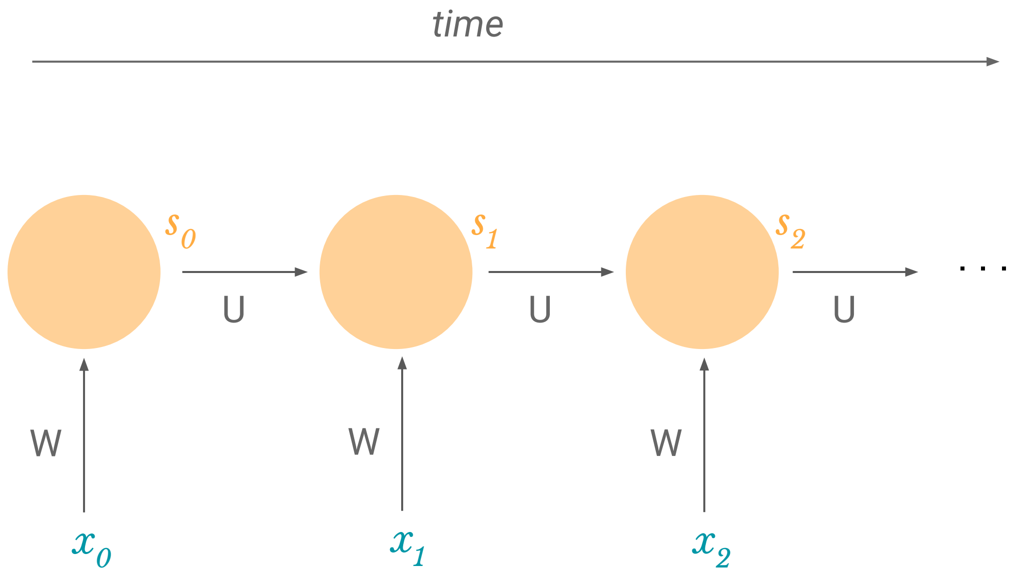
$$s_2 = tanh(Wx_1 + Us_1)$$

$W,\ U$: weight matrices

“Unfolding” the RNN across time:

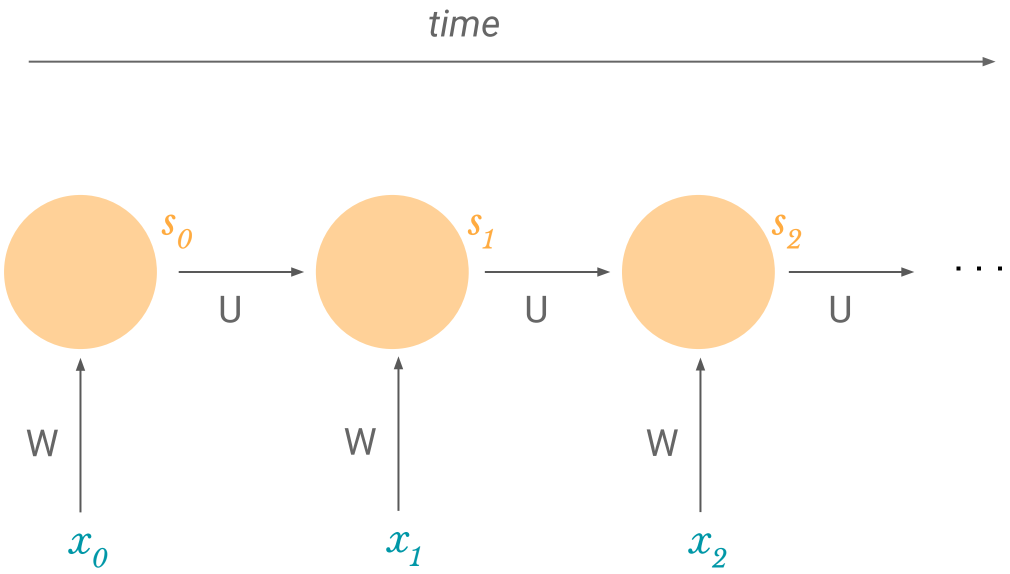


"Unfolding" the RNN across time:



notice that we use the same parameters, W and U

“Unfolding” the RNN across time:



s_n **can contain information from all past timesteps**

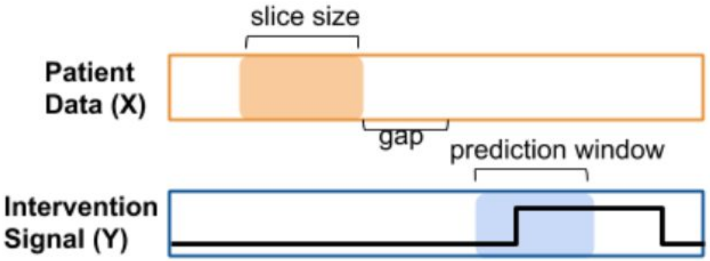
Why do LSTMs help?

1. Forget gate allows information to pass through unchanged
2. Cell state is separate from what's outputted
3. s_j depends on s_{j-1} through addition!

   → derivatives don't expand into a long product!

Predict Onsets of Interventions

- Delay prediction by 6-hour gap time.
- Attempt to predict onest, weaning, staying off, staying on.



	Onset	Weaning	Stay Off	Stay On
Ventilation	0.005	0.017	0.798	0.18
Vasopressor	0.008	0.016	0.862	0.114
NI-Ventilation	0.024	0.035	0.695	0.246
Colloid Bolus	0.003	-	-	-
Crystalloid Bol	0.022	-	-	-

NNs Do Well; Improved Representation Helps

Area-under-ROC

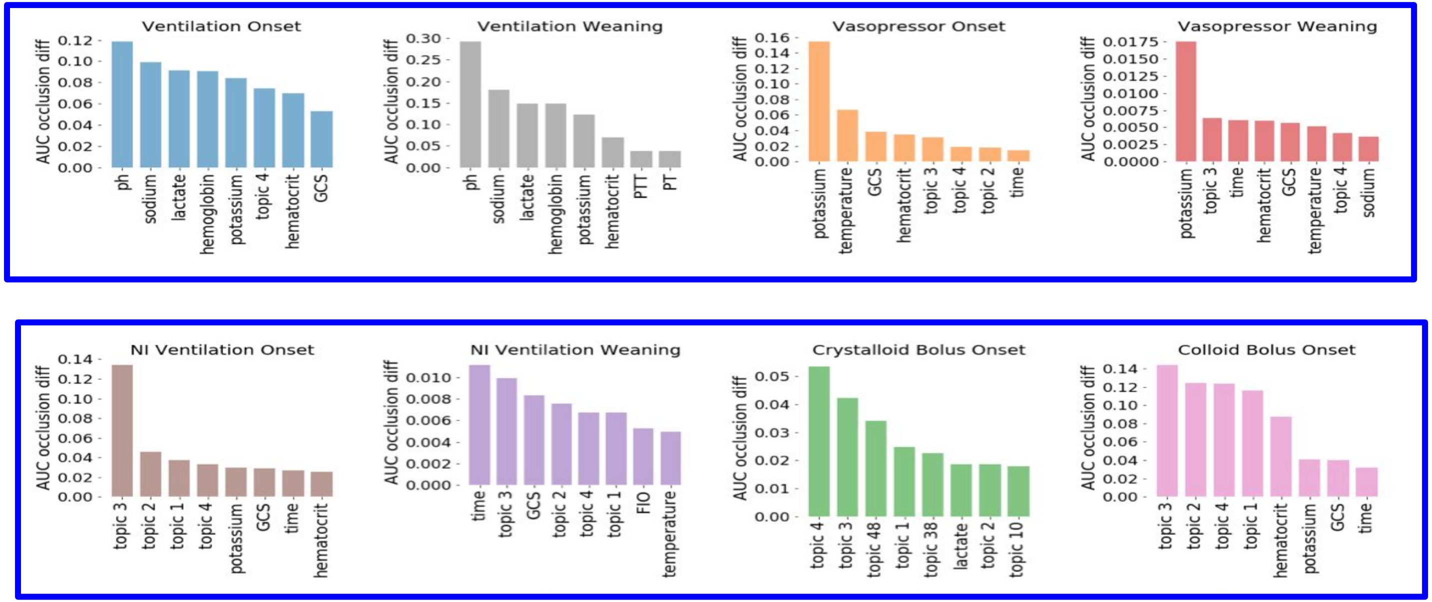
Task	Model	Intervention Type: VENT	NI-VENT	VASO	COL BOL	CRYS BOL
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	**0.76**	0.76	**0.72**	**0.71**
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	LSTM Raw	0.90	0.80	**0.91**	-	-
	LSTM Words	0.90	**0.81**	**0.91**	-	-
	CNN	**0.91**	0.80	**0.91**	-	-
Stay On AUC	Baseline	0.50	0.79	0.55	-	-
	LSTM Raw	0.96	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.96	**0.86**	**0.96**	-	-
Stay Off AUC	Baseline	0.94	0.71	0.93	-	-
	LSTM Raw	0.95	**0.86**	**0.96**	-	-
	LSTM Words	**0.97**	**0.86**	0.95	-	-
	CNN	0.95	**0.86**	**0.96**	-	-
Macro AUC	Baseline	0.72	0.72	0.66	-	-
	LSTM Raw	0.86	**0.82**	**0.90**	-	-
	LSTM Words	**0.90**	**0.82**	0.89	-	-
	CNN	0.86	0.81	**0.90**	-	-

Representations with “physiological words” for missingness significantly increased AUC for interventions with the lowest proportion of examples.

Deep models perform well in general, but words are important for ventilation tasks.

Feature-Level Occlusions Identify Per-Class Features

Decrease in AUC

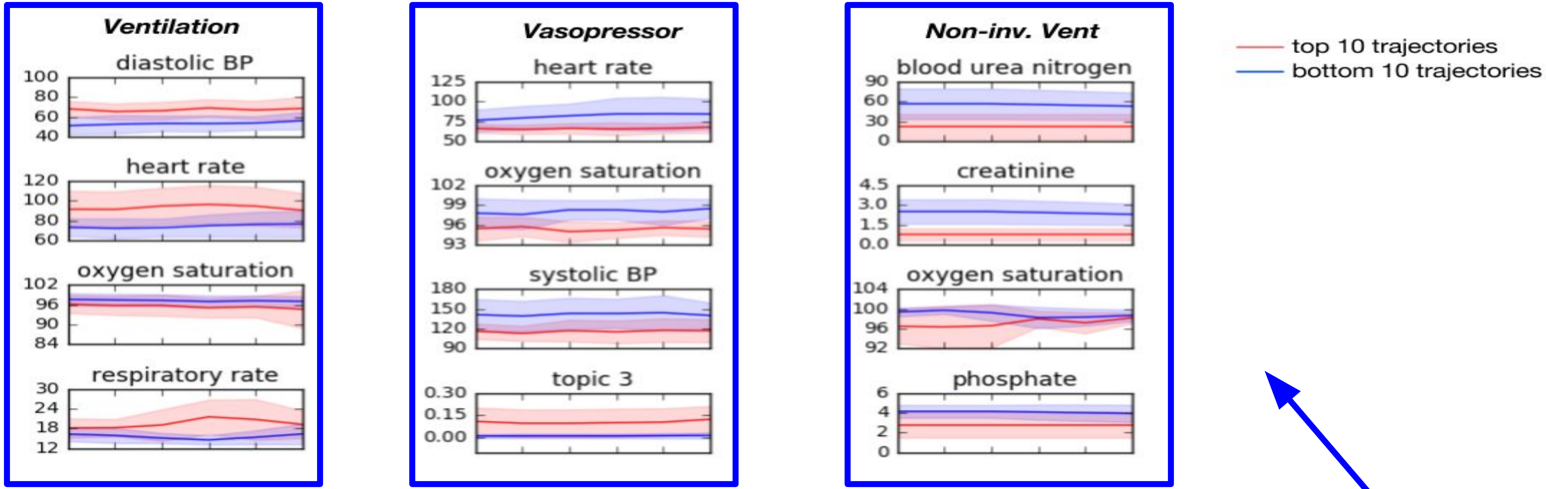


Physiological data were more important for the more **invasive** interventions.

Clinical **note topics** were more important for **less invasive** tasks.

Convolutional Filters Target Short-term Trajectories

Most differentiated features of 10 real patient trajectories that are highest/lowest activating for each task.



Higher diastolic blood pressure, respiratory rate, and heart rate, and lower oxygen saturation : **Hyperventilation**

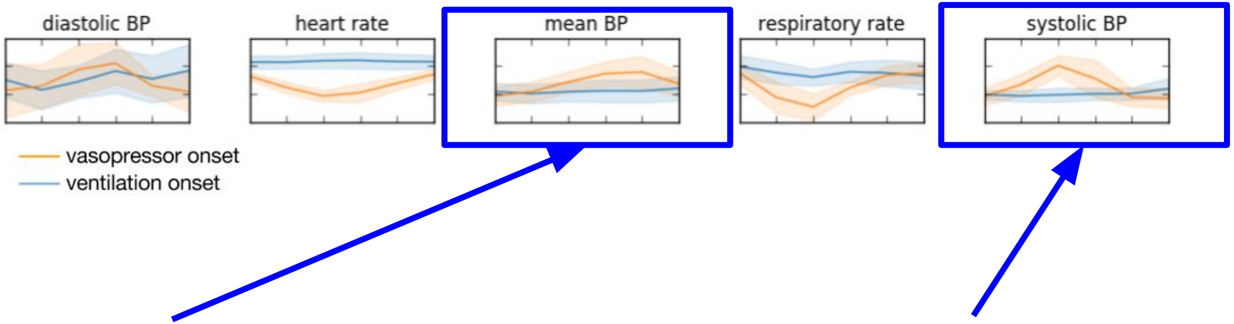
Decreased systolic blood pressure, heart rate and oxygen saturation rate : **Altered peripheral perfusion** or **stress hyperglycemia**

Decreased creatinine, phosphate, oxygen saturation and blood urea nitrogen : **Neuromuscular respiratory failure**

Convolutional Filters Target Short-term Trajectories

- “Hallucinations” give insight into underlying properties of the network.

- The trajectories are made to maximize the output of the model, (do not correspond to physiologically plausible trajectories).



Blood pressure drops are maximally activating for **vasopressor onset**.

Respiratory rate decreasing is maximally activating for **ventilation onset**.

Outline

1. What’s Time Got To Do With It?
    a. Missingness
    b. Representation

2. Case Study 1: MTGPs for Mortality Prediction and TBI

3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. **Project Discussion**