CSC 2541: Machine Learning for Healthcare

Lecture 5: Clinical Time Series Modelling

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

Course Reminders!

- No weekly reflection questions to MarkUs this week!
- You finished the homework!
- Your project proposals are due next week!

Schedule

Jan 10, 2019, Lecture 1: Why is healthcare unique?

Jan 17, 2019, Lecture 2: Supervised Learning for Classification, Risk Scores and Survival

Jan 24, 2019, Lecture 3: Causal inference with observational data
Jan 31, 2019, Lecture 4: Fairness, Ethics, and Healthcare

Feb 7, 2019, Lecture 5: Clinical Time Series Modelling (Homework 1 due at 11:59 PM on MarkUs)
Feb 14, 2019, Lecture 6: Clinical Imaging (Project proposals due at 5PM on MarkUs)
Feb 21, 2019, Lecture 7: Clinical NLP and Audio

Feb 28, 2019, Lecture 8: Clinical Reinforcement Learning
Mar 7, 2019, Lecture 9: Missingness and Representations
Mar 14, 2019, Lecture 10: Generalization and transfer learning
Mar 21, 2019, Lecture 11: Interpretability / Humans-In-The-Loop / Policies and Politics

Mar 28, 2019, Course Presentations
April 4, 2019, Course Presentations (Project report due 11:59PM)

Outline

1. What’s Time Got To Do With It?
2. Case Study 1: MTGPs for Mortality Prediction and TBI
3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction
4. What’s Out There?
5. Project Discussion

Outline

1. **What's Time Got To Do With It?**

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

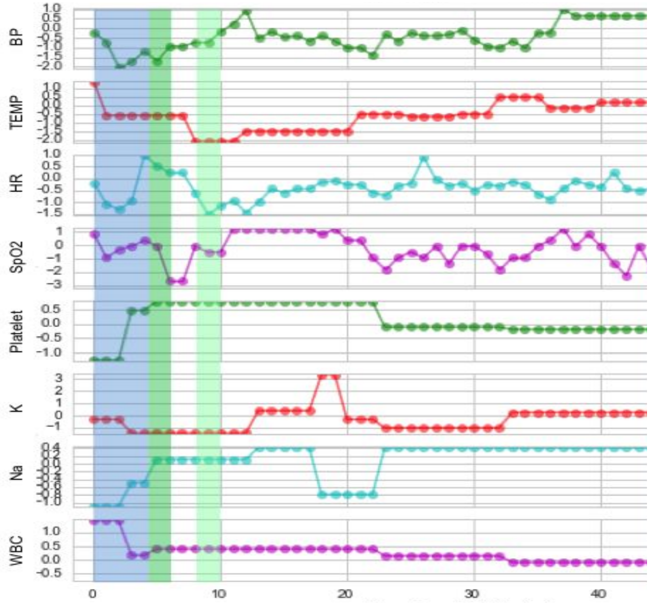
3. Case Study 2: RNNs/CNNs for Intervention Onset Prediction

4. What's Out There?

5. Project Discussion

Problem: Hospital decision-making / care planning

Observe Patient Data

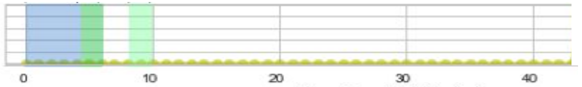


?

"Real-time" **Prediction**

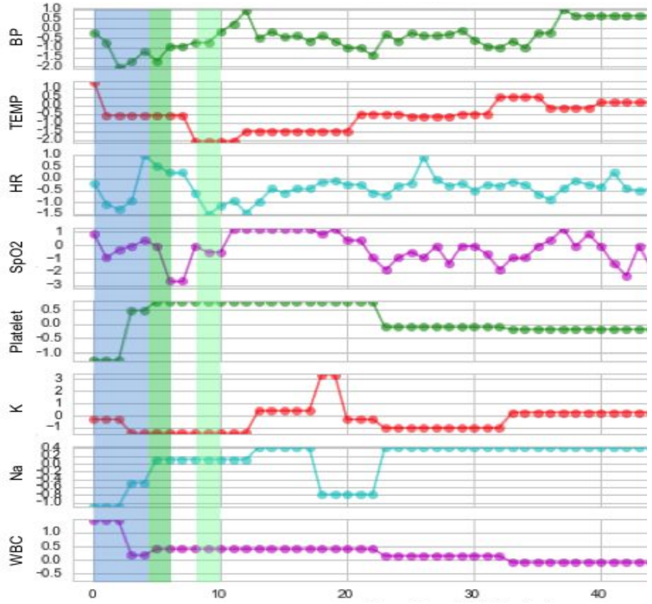
Of **{**Drug/Mortality/Condition**}**

By Gap Time



Problem: Hospital decision-making / care planning

Observe Patient Data

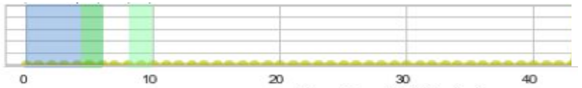


?

“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?

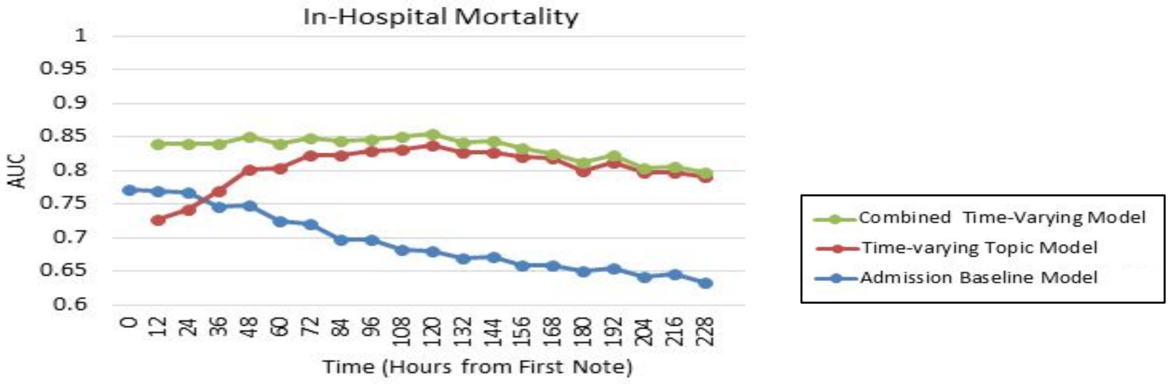
2. **Case Study 1: MTGPs for Mortality Prediction and TBI**

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

4. What’s Out There?

5. Project Discussion

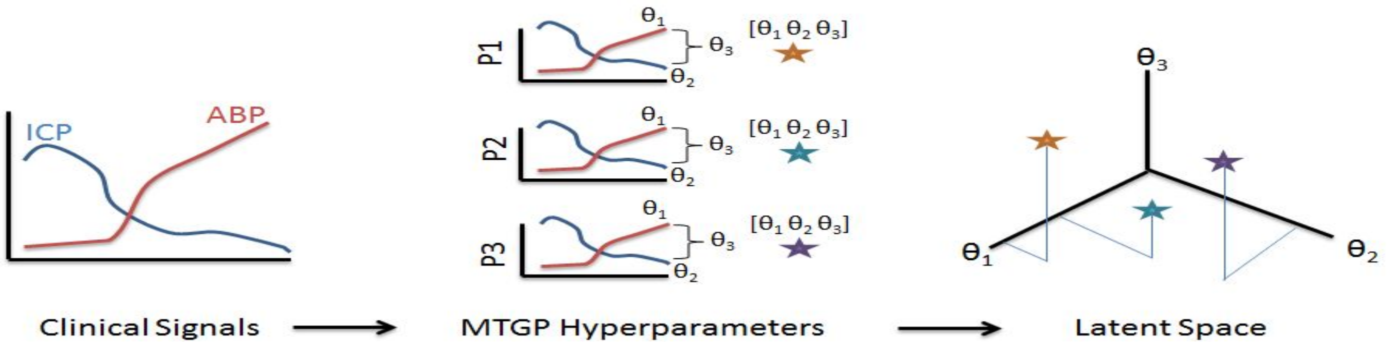
Remember This? Topics Improves Mortality Prediction



- **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.

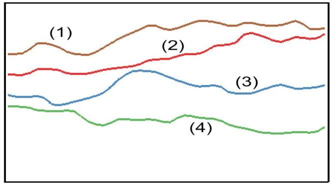
Transform signals into MTGP hyperparameter representation.

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**.

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

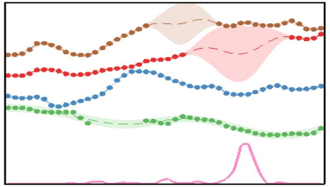


- GP's commonly used to predict at new indices.

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

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

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



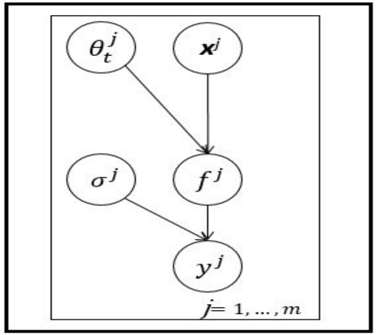
- Learn the parameters (θ) of the **kernel** from **data**.

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

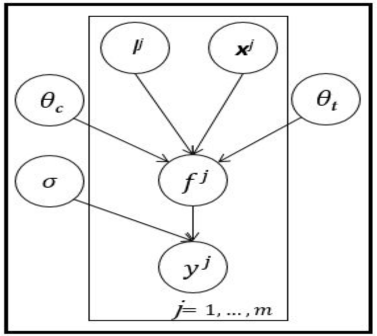
Single vs. Multi-task Gaussian Processes

- 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.

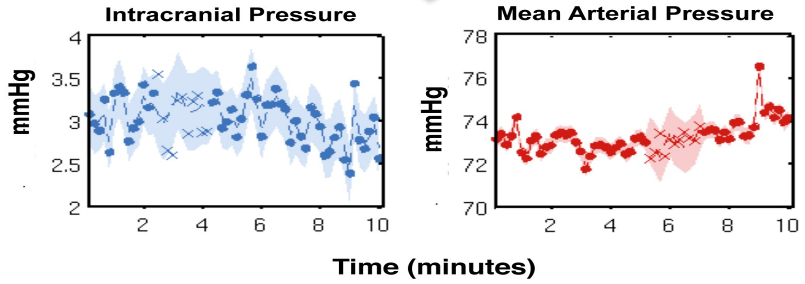
$$\mathbf{K}_{MT}(\mathbf{X}_n, \mathbf{l}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_t) = \mathbf{K}_c(\mathbf{l}, \boldsymbol{\theta}_c) \otimes \mathbf{K}_t(\mathbf{X}_n, \boldsymbol{\theta}_t)$$

Movement **between** signals

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

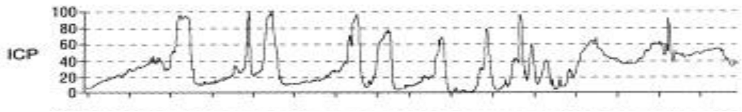
Movement **within** a signal

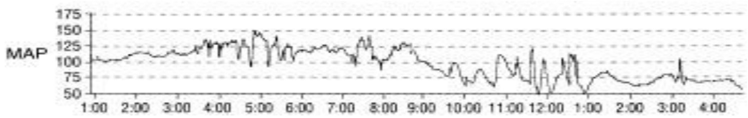
$$\mathbf{K}_t = \theta_A^2 \exp\left\{ -\frac{\parallel x - x' \parallel^2}{2\theta_L^2} \right\}$$



[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].

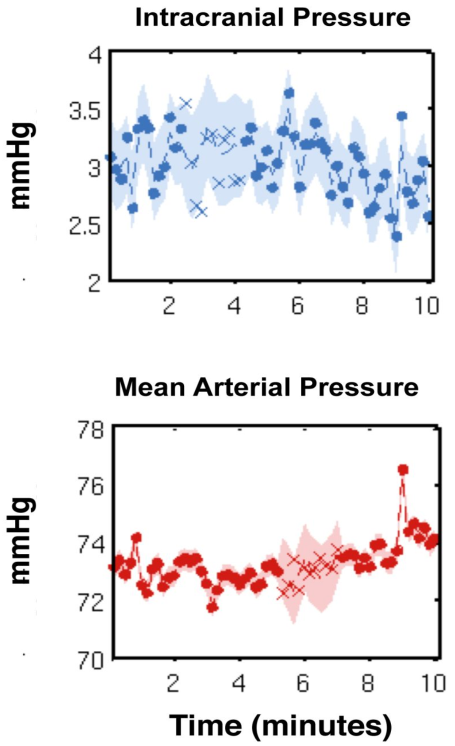
[1] Werner, C., and K. Engelhard. "Pathophysiology of traumatic brain injury." *British journal of anaesthesia* 99.1 (2007): 4-9.
[2] Hlatky, Roman, Alex B. Valadka, and Claudia S. Robertson. "Intracranial pressure response to induced hypertension: role of dynamic pressure autoregulation." *Neurosurgery* 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?

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

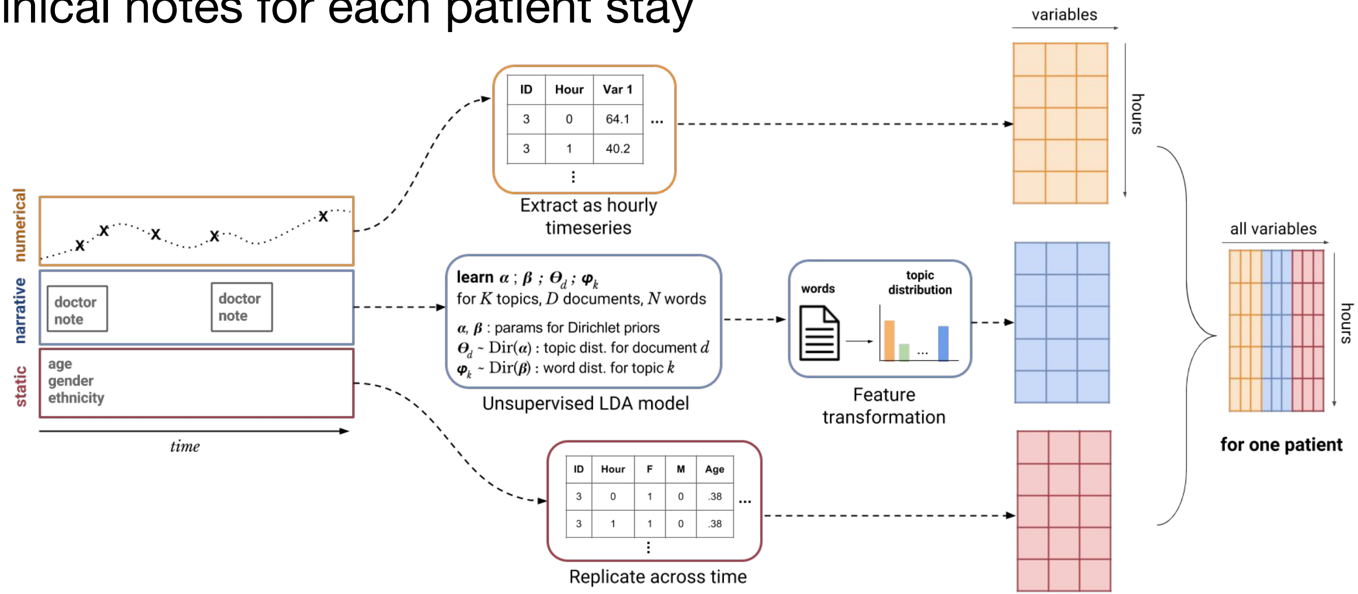
3. **Case Study 2: RNNs/CNNs for Intervention Onset Prediction**

4. What’s Out There?

5. 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
⋮	⋮	⋮

-1

Physiological Words

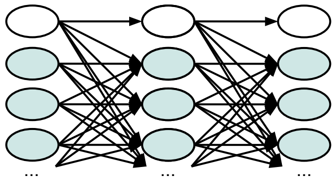
patient	hours in	glucose_-2	glucose_-1	glucose_0	glucose_1	glucose_2
3	1	0	0	0	0	0
3	2	0	0	0	0	0
3	3	0	1	0	0	0
	⋮	⋮	⋮ ⋮	⋮	⋮	

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

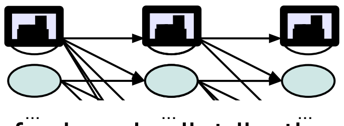
- Many values are missing!
- Z-score existing variables, rounding to the nearest int.
- Convert each z-score into its own binary column.

Many Ways to Model, What Do We Learn?

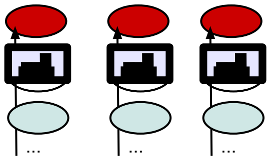
SSAM



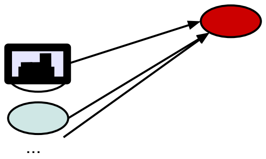
Learn model parameters over patients with variational EM.



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

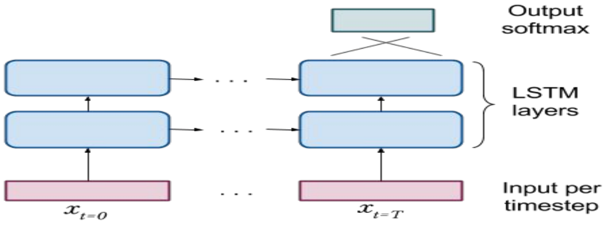


Logistic regression (with label-balanced cost function)



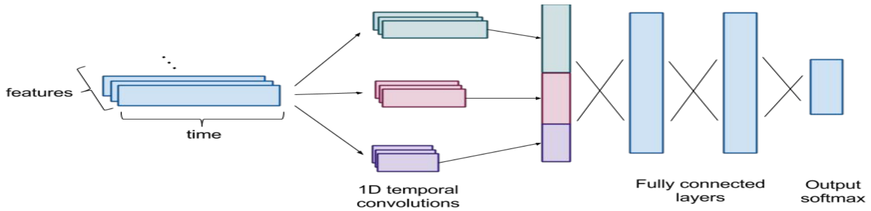
Predict onset in advance

LSTM



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

CNN

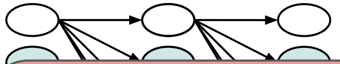


CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.

24

Many Ways to Model, What Do We Learn?

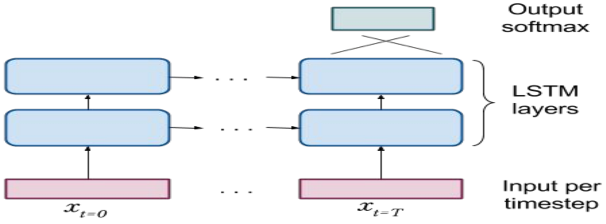
SSAM



Covered **weeks** ago!

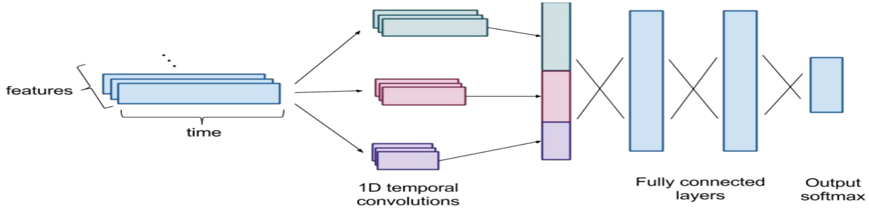
variational EM. (fwd alg.). cost function)

LSTM



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

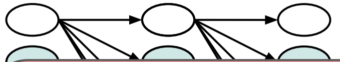
CNN



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

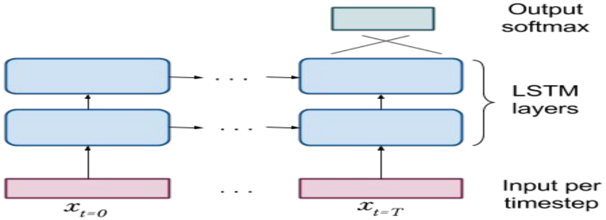
Many Ways to Model, What Do We Learn?

SSAM



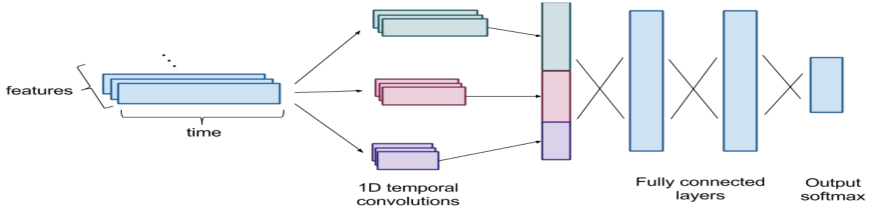
Covered **weeks** ago!

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.

26

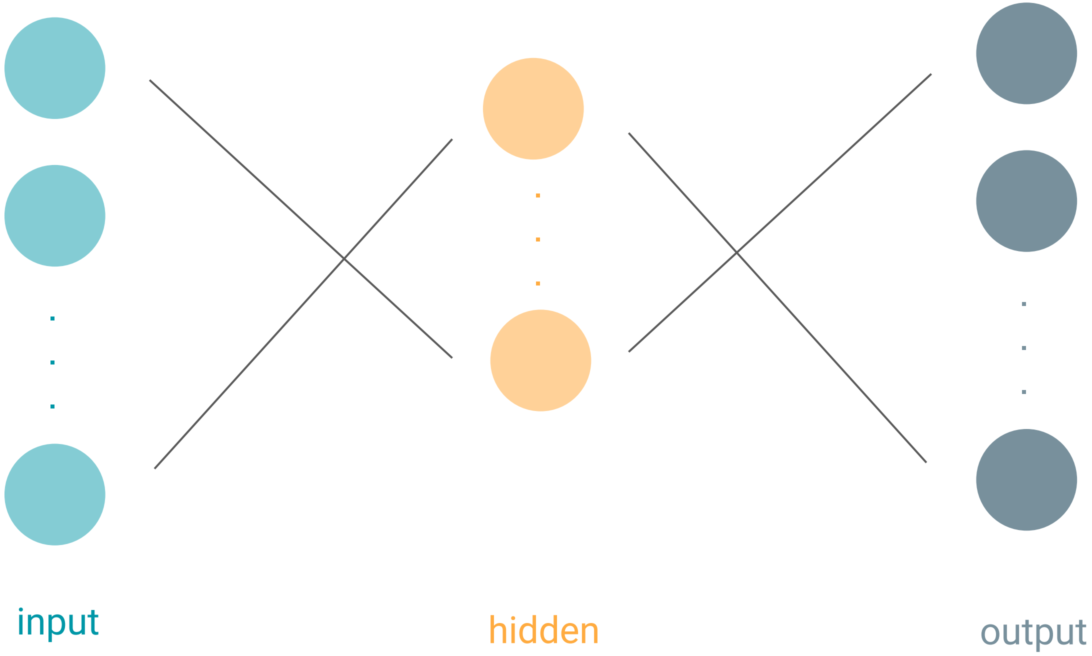
RNNs on Sequences

To model sequences, we need:

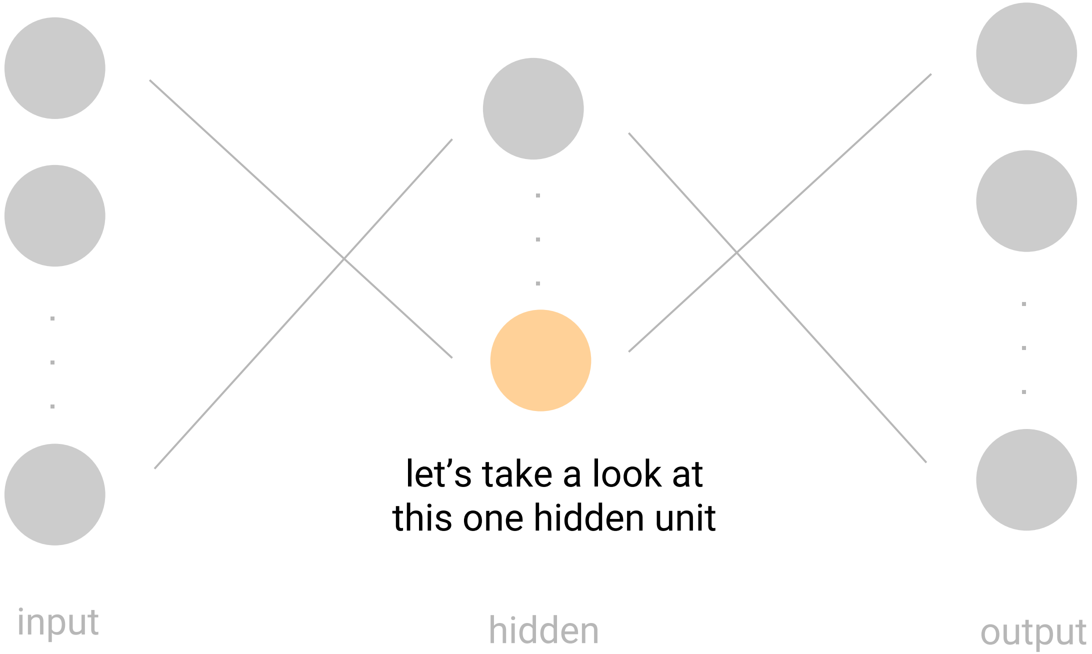
1. To deal with variable-length sequences
2. To maintain sequence order
3. To keep track of long-term dependencies
4. To share parameters across the sequence

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

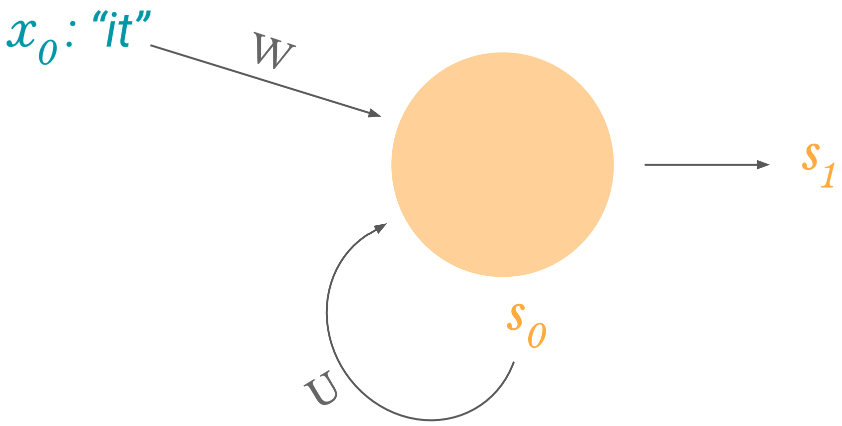
Example Network



Example Network



RNNS **remember** their previous state:



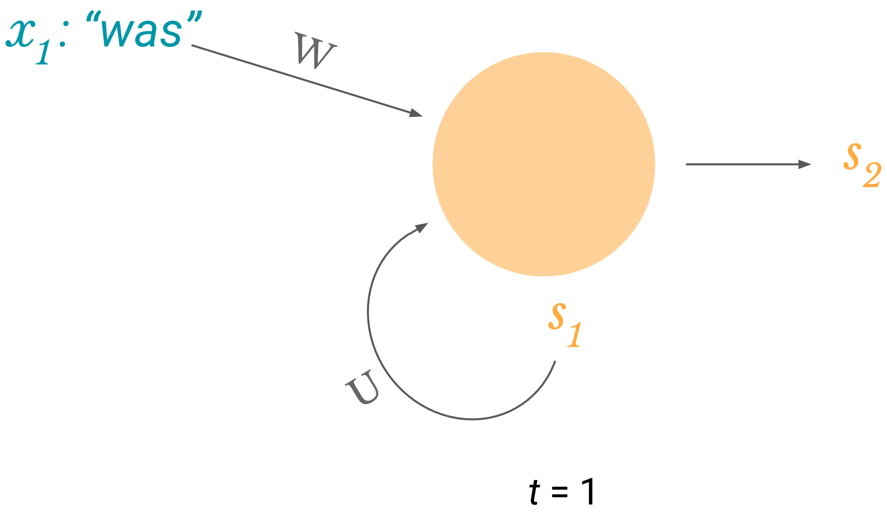
t = 0

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:

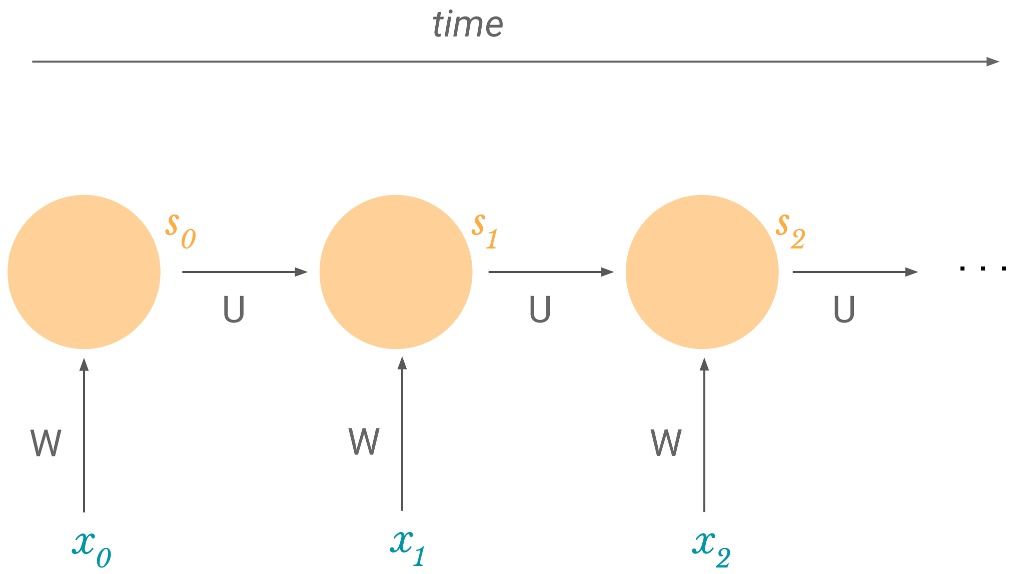


x_1 : vector representing second word
s_1 : cell state at $t = 1$
s_2 : cell state at $t = 2$

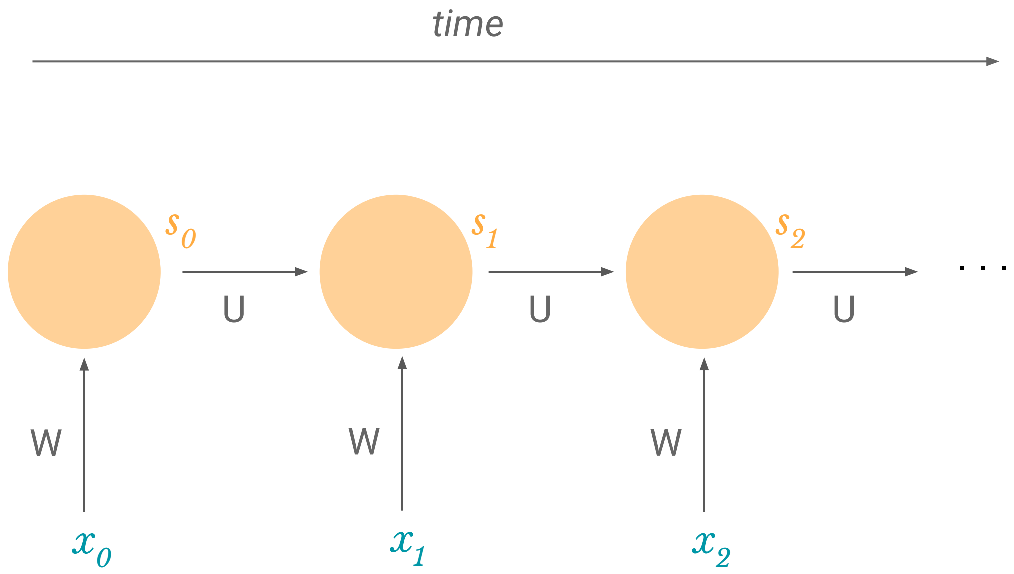
$$s_2 = tanh(Wx_1 + Us_1)$$

$W,\ U$: weight matrices

“Unfolding” the RNN across time:

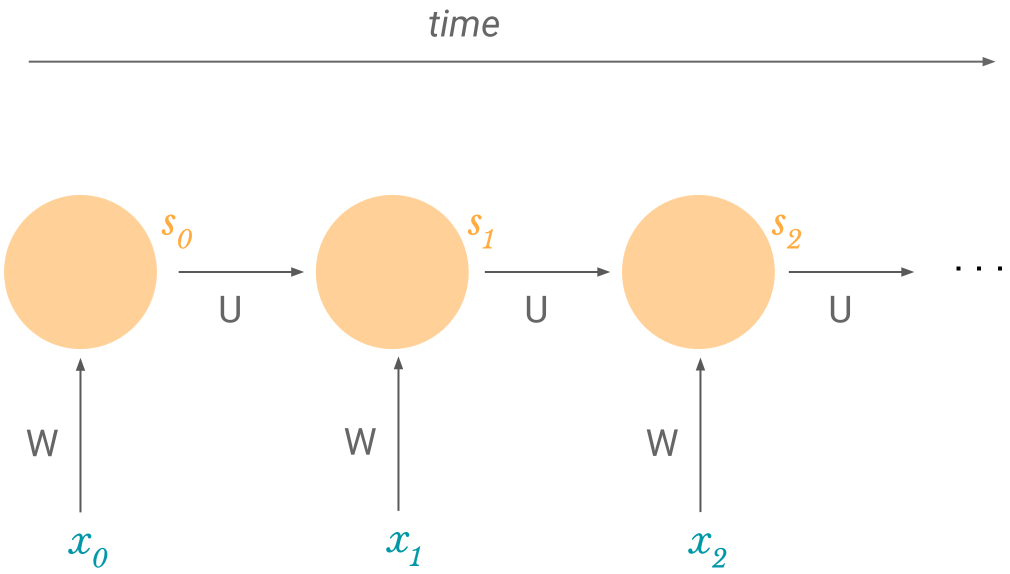


“Unfolding” the RNN across time:



notice that we use the same parameters, W and U

“Unfolding” the RNN across time:



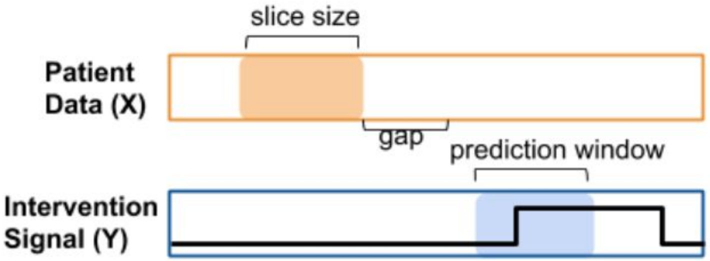
s_n can contain information from all past timesteps

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!
   $\rightarrow$ 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

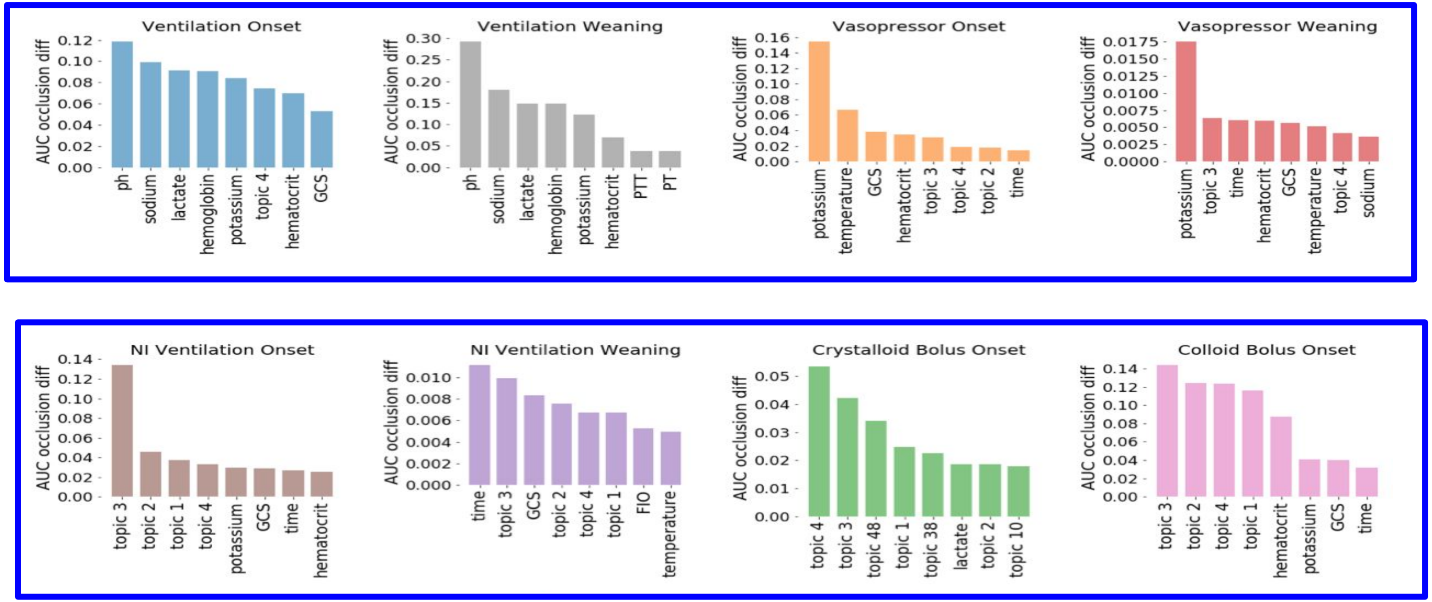
Task	Model	VENT	NI-VENT	VASO	COL BOL	CRYS BOL
		Intervention Type				
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	$\boxed{\mathbf{0.76}}$	0.76	**0.72**	$\boxed{\mathbf{0.71}}$
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	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**	$\boxed{\mathbf{0.82}}$	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

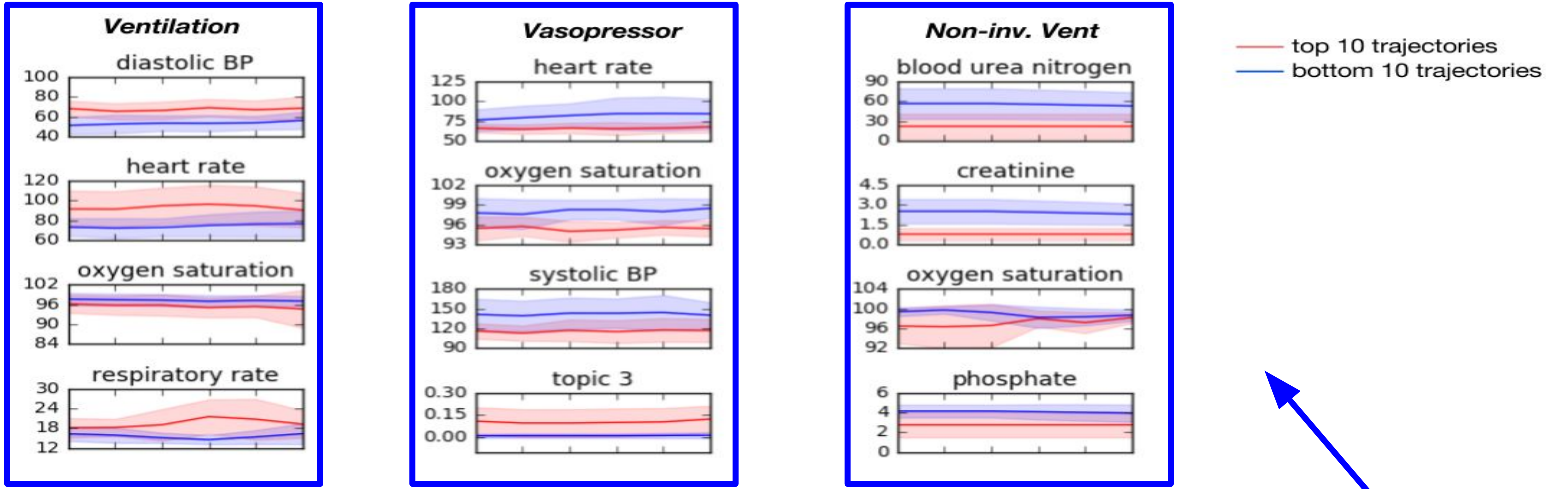


Physiological data were more important for the more **invasive** interventions.

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**

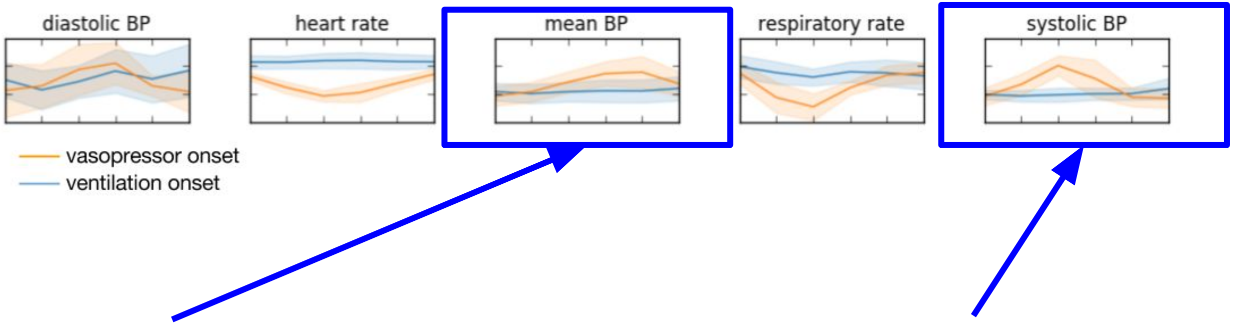
Decreased systolic blood pressure, heart rate and oxygen saturation rate : **Altered peripheral perfusion** or **stress hyperglycemia**

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?

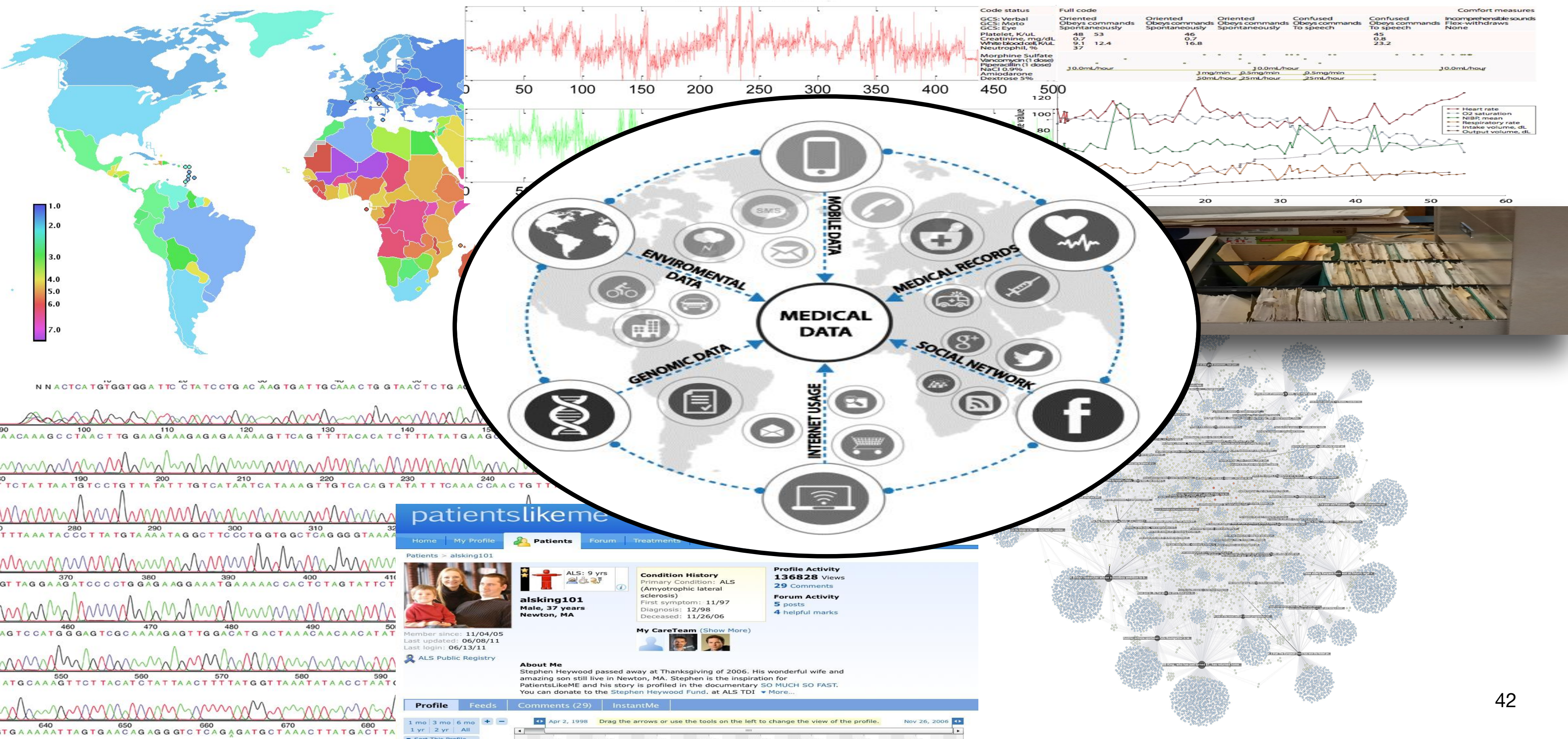
2. Case Study 1: MTGPs for Mortality Prediction and TBI

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

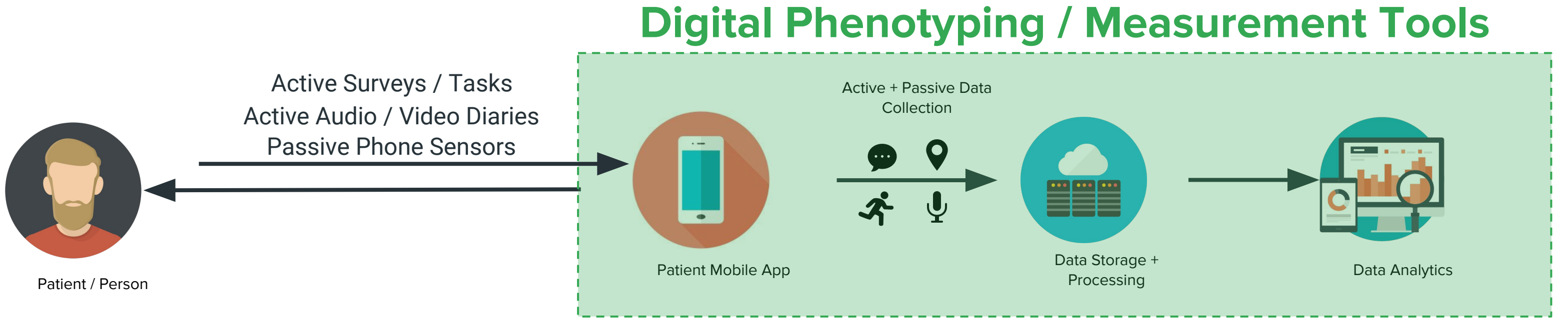
4. **What’s Out There?**

5. Project Discussion

Remember That Life Happens Outside the Clinic!



Example: Wearable Data from Project Myalo



- Phone sensors offer a potential low-cost, low-barrier method for **digital quantification of behavior** that can achieve scalability better than other wearable sensors

- Tracking over time: mood, sleep, physical activity, cognitive function, social activities!

Some Technology Required...

StudentLife

SENSING MENTAL HEALTH.
LIFEGRAPH

Purple Robot
CBITs Tools
Everyone
You don't have any devices

Patient Mobile App

Raw Data Sources

- Phone / App / Social Media Usage
- SMS / Call Logs
- Microphone / Audio
- GPS
- Accelerometer / Gyroscope
- Temperature / Light
- Active Surveys / Tasks

Data Storage / Processing

Processed Behavioral Labels

- Sleep
- Physical Activity
- Text / Speech Features
- Vocal Features
- Geolocation (home, work, ...)
- Social Media
- Mood / Affect

Machine Learning Algorithms

Derived Clinical Insights

- Voice & Speech Markers
- Location Variability
- Circadian Rhythm
- Social Activity
- Psychomotor Function
- Sleep Quality
- Cognitive Function

Predictive Modeling

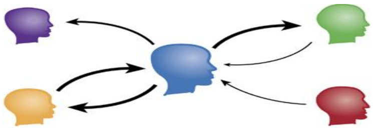
Saeb, PeerJ (2016)
Aung, Depression and Anxiety (2017)
Insel, JAMA (2017)
Wang, Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (2017)

44

What Can We Do With A Digital Phenotype?

"Moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices"

- **Expert-Reported Data** (EHR)
- **Self-Reported Data** (Surveys)
- **Passive Data** (Ambulatory)
  - Spatial trajectories (GPS)
  - Physical mobility (accelerometer)
  - Social networks/dynamics (call/text)
  - Voice samples (microphone)



JP Onnela and Scott Rauch, Neuropsychopharmacology (2016)

What Could We Ask With A Rich Phenotype?

- How do depressed patients (**Expert-Reported**) divide time between home and work (**Passive**)?

- Do the size and reciprocity of interaction networks (**Passive**) help with anxiety (**Self-Reported**)?

- Does activity (**Passive**) impact mood (**Self-Reported**) differently post-partum (**Expert-Reported**)?

Main Take Aways

- Combining data across modalities and time can be powerful.

- Kernel representations are intuitive comparisons for intra/inter signal modelling.

- Representations improve task performance.

Outline

1. What’s Time Got To Do With It?

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

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

4. What’s Out There?

5. **Project Discussion**