CSC 2541: Machine Learning for Healthcare

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

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!
- Start the homework early (e.g., last week)!
- Sign up for a paper presentation slot!
- 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 can we do with supervised learning?**

2. Case study on intervention predictions:
    a. Frame the problem
    b. Evaluation
    c. Iterate

3. Survival Analysis

4. What else should we be thinking about?

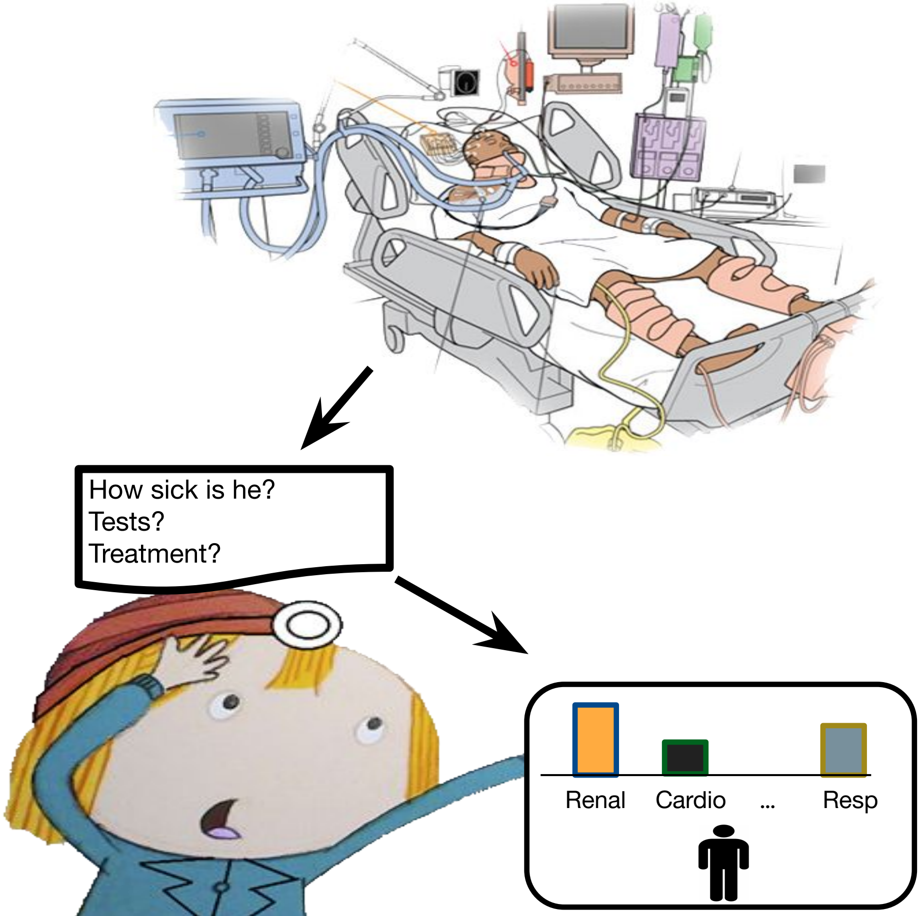
Clinicians Need to Estimate Patient State and Predict Outcome

- How do I figure out which patient needs my attention now?

- How will the patient's underlying cardiovascular system respond to my plan of care?

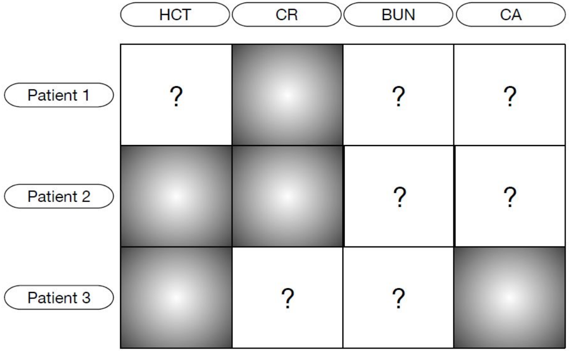
- If I discharge this patient, will they be readmitted?

- Are a patient's home behaviors impacting their health?

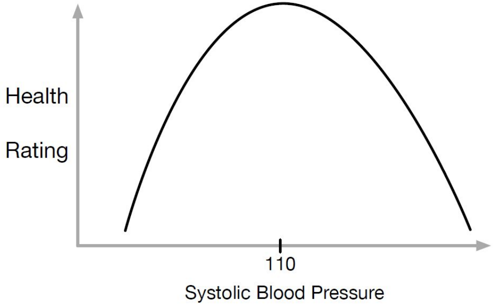


But Those Challenges...

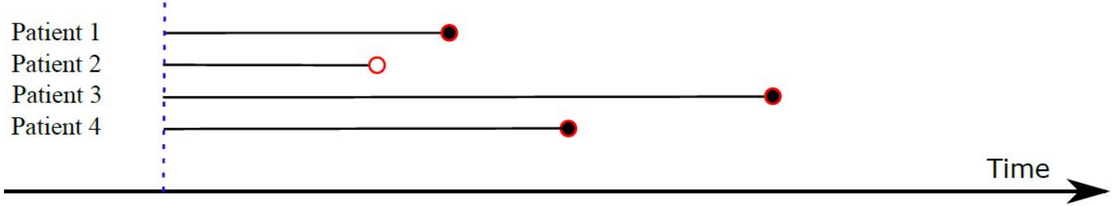
Incomplete Data



Non-linear Relationships

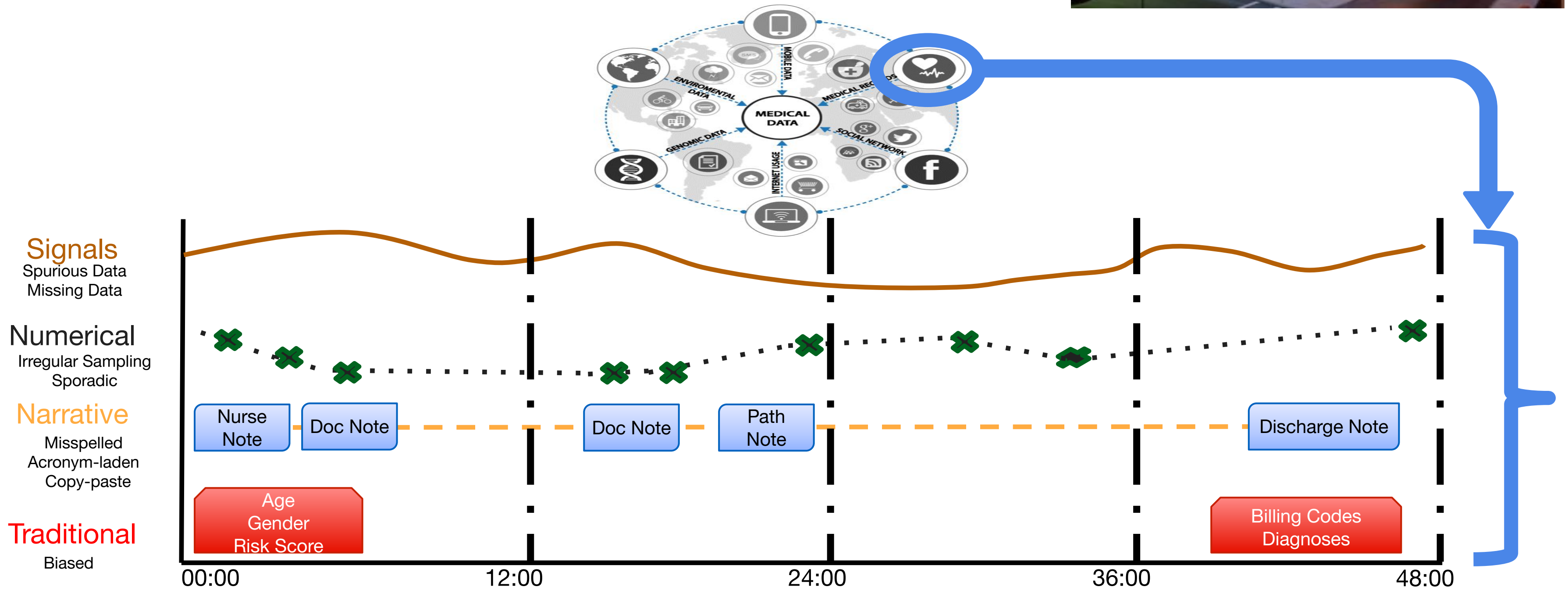


Censoring



MIMIC III ICU Data

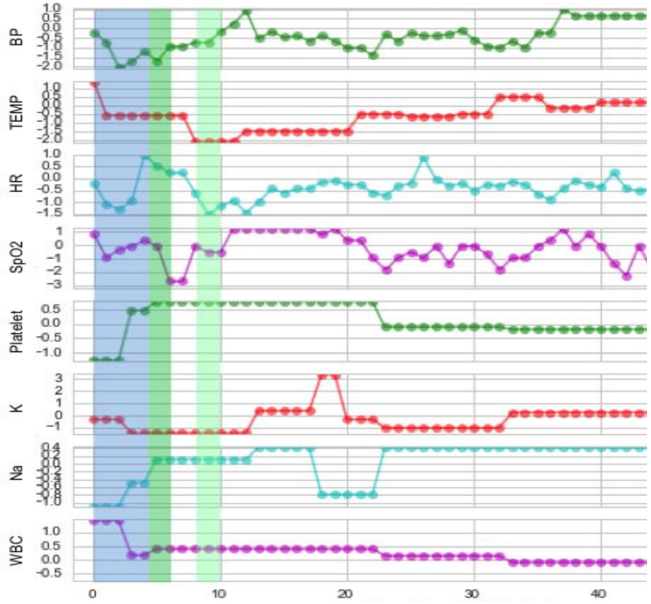
- Learning with real patient data from the Beth Israel Deaconess Medical Center ICU. [1]



[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3 (2016).

Problem: Hospital decision-making / care planning

Observe Patient Data

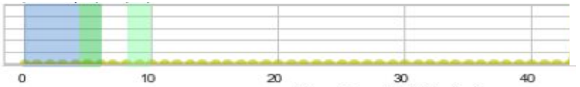


?

"Real-time" **Prediction**

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

By Gap Time



Part 1: Predict **mortality** with clinical **notes**

- **Acuity** (severity of illness) very important - use **mortality** as a **proxy** for **acuity.**[1]

- Prior state-of-the-art focused on feature engineering in **labs/vitals** for target populations.[2]

- But **clinicians** rely on **notes**.

[1] Siontis, George CM, Ioanna Tzoulaki, and John PA Ioannidis. "Predicting death: an empirical evaluation of predictive tools for mortality." *Archives of internal medicine* 171.19 (2011): 1721-1726.
[2] Grady, Deborah, and Seth A. Berkowitz. "Why is a good clinical prediction rule so hard to find?." *Archives of internal medicine* 171.19 (2011): 1701-1702.

Clinical notes are messy...

Patient Y, 12:45:00 EST

uneventful day. pt much improved. VS Stable nuero intact no compromise NSR BP stable Aline discontinued in afternoon. pt to transfer to floor awaiting bed. pt continues with nausea given anziment and started on reglan prn. small emesis in am. pt continues with ice chips. foley draining well adequate output. now replacing half cc for cc of urine. skin and surgical site unchanged, C/D/I. family (son and husband) at bedside for most of day. Plan: continue with current plan in progress, tranfer to floor.

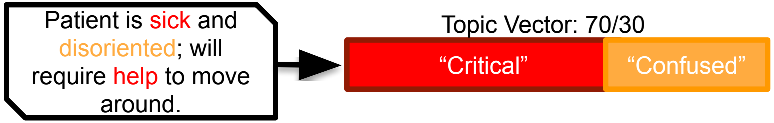
CONTEXT MATTERS

ACRONYM

MISSPELLED

Represent patients as topic vectors

- Model patient stays as an **aggregated set** of notes.
- Model notes as a **distribution** over topics.
- A “topic” is a **distribution** over words, that we learn.

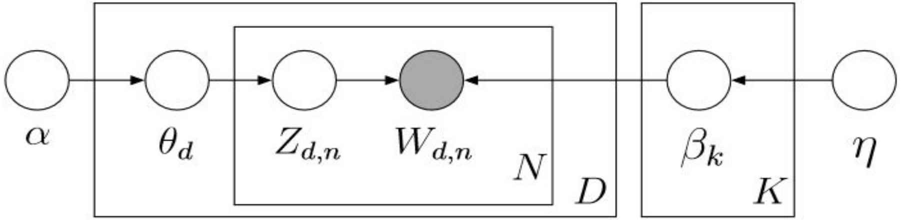


- Use Latent Dirichlet Allocation (LDA)[1] as an **unsupervised** way to **abstract** 473,000 notes from 19,000 patients into “topics”.[2]

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022
[2] T. Griffhs and M. Steyvers. Finding scientific topics.In PNAS, volume 101, pages 5228{5235, 2004

Learning topics

- Observe **words**, infer **Z**:



$$\prod_{i=1}^{K} p(\beta_i \mid \eta) \prod_{d=1}^{D} p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right)$$

Per-word topic assignment $Z_{d,n}$

Per-doc topic proportion θ_d

Corpus topic distribution β_k

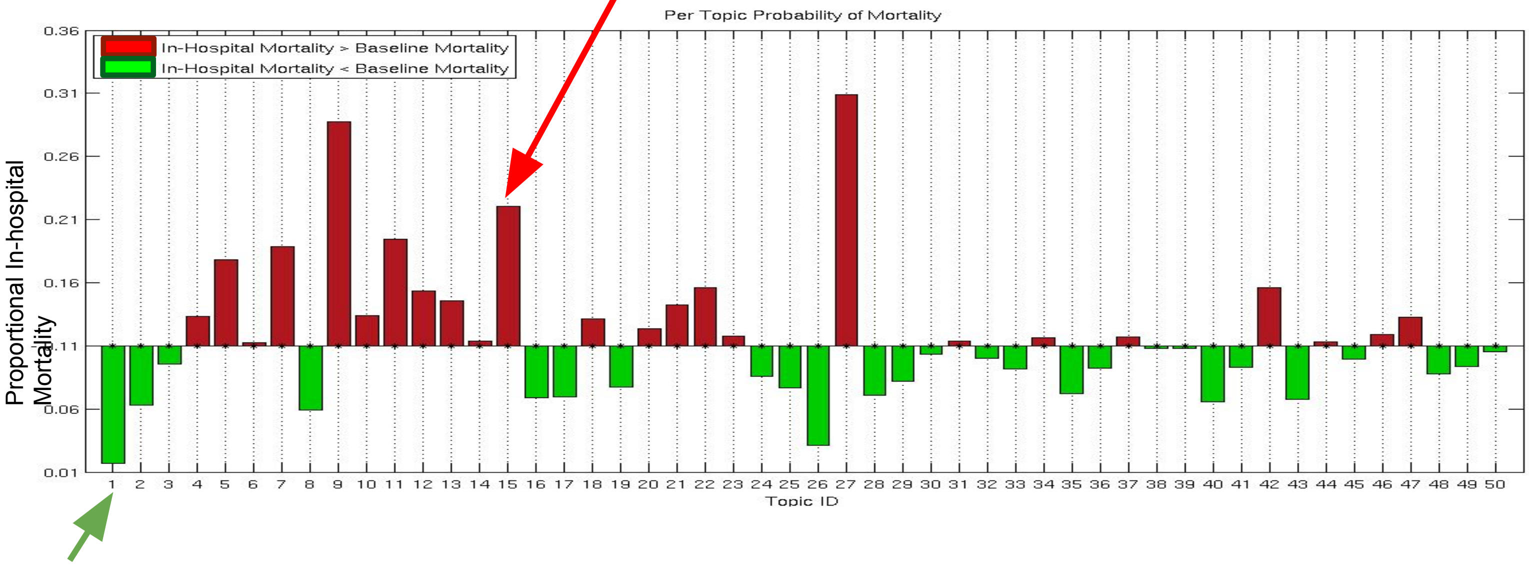
Sparsity α

Exclusivity η

[1] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *the Journal of machine Learning research* 3 (2003): 993-1022
[2] T. Griffhs and M. Steyvers. Finding scientific topics.In PNAS, volume 101, pages 5228{5235, 2004

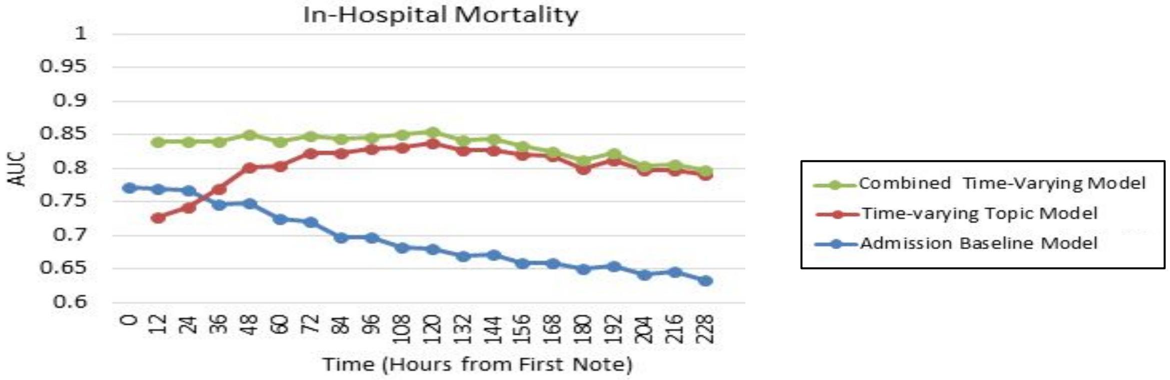
Correlation between average topic and mortality

Topic #	Top Ten Words	Possible Topic
15	intubated vent ett secretions propofol abg respiratory resp care sedated	Respiratory failure



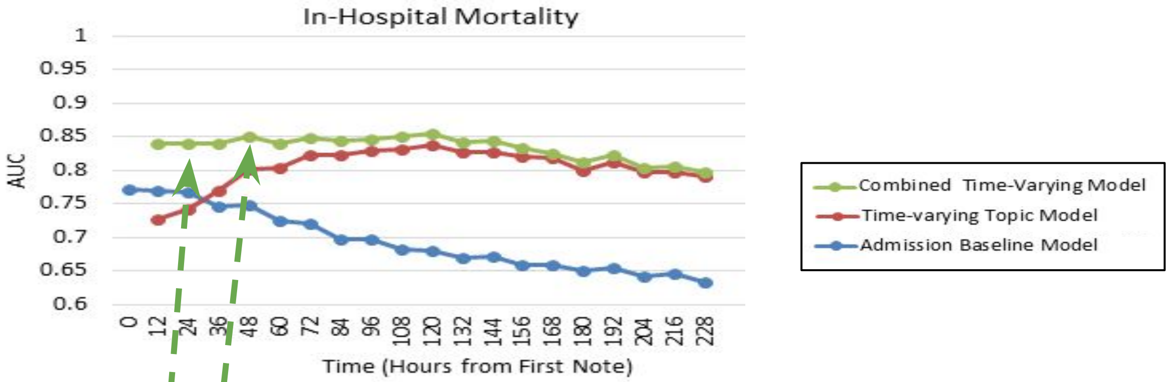
Topic #	Top Ten Words	Possible Topic
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp	Cardiovascular surgery

Topics improve in-hospital mortality prediction



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

More complex models are not always better



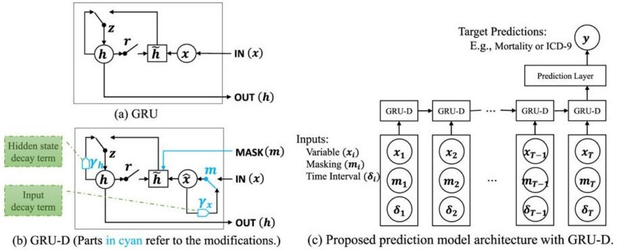
Author	AUC	Method	Episodes	Hours	Variables
Ghassemi, 2014	0.84/**0.85**	LDA	19,308	24/48	53 - notes
Caballero, 2015	0.86	Text processing + medication	15,000	24	? - notes/meds
Che, 2015	0.8-0.82	Deep Learning (LSTM)	3,940	48	30 - vitals
Che, 2016	0.7/0.85	Deep Learning (GRU)	19,714	12/48	99 – vitals/meds
Che, 2018	**0.85**	Deep Learning (GRU-D)	19,714	48	99 – vitals/meds

More Complex ≠ Better

Caballero Barajas, Karla L., and Ram Akella. "Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.
Che, Zhengping, et al. "Deep computational phenotyping." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.
Che, Zhengping, et al. "Recurrent Neural Networks for Multivariate Time Series with Missing Values." arXiv preprint arXiv:1606.01865 (2016).
Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Scientific reports. 2018 Apr 17;8(1):6085.

Even when complex and clever!

- Explicitly capture and use missing patterns in RNNs via systematically modified architectures.



(a) GRU

(b) GRU-D (Parts in cyan refer to the modifications.)

(c) Proposed prediction model architecture with GRU-D.

- Performance bump is small, is MIMIC mortality our MNIST?

Non-RNN Models						**RNN Models**	
Mortality Prediction On MIMIC-III Dataset						LSTM-Mean	0.8142 ± 0.014
LR-Mean	0.7589 ± 0.015	SVM-Mean	0.7908 ± 0.006	RF-Mean	0.8293 ± 0.004	GRU-Mean	0.8252 ± 0.011
LR-Forward	0.7792 ± 0.018	SVM-Forward	0.8010 ± 0.004	RF-Forward	0.8303 ± 0.003	GRU-Forward	0.8192 ± 0.013
LR-Simple	0.7715 ± 0.015	SVM-Simple	0.8146 ± 0.008	RF-Simple	0.8294 ± 0.007	GRU-Simple w/o δ[22]	0.8367 ± 0.009
LR-SoftImpute	0.7598 ± 0.017	SVM-SoftImpute	0.7540 ± 0.012	RF-SoftImpute	0.7855 ± 0.011	GRU-Simple w/o m[23,24]	0.8266 ± 0.009
LR-KNN	0.6877 ± 0.011	SVM-KNN	0.7200 ± 0.004	RF-KNN	0.7135 ± 0.015	GRU-Simple	0.8380 ± 0.008
LR-CubicSpline	0.7270 ± 0.005	SVM-CubicSpline	0.6376 ± 0.018	RF-CubicSpline	0.8339 ± 0.007	GRU-CubicSpline	0.8180 ± 0.011
LR-MICE	0.6965 ± 0.019	SVM-MICE	0.7169 ± 0.012	RF-MICE	0.7159 ± 0.005	GRU-MICE	0.7527 ± 0.015
LR-MF	0.7158 ± 0.018	SVM-MF	0.7266 ± 0.017	RF-MF	0.7234 ± 0.011	GRU-MF	0.7843 ± 0.012
LR-PCA	0.7246 ± 0.014	SVM-PCA	0.7235 ± 0.012	RF-PCA	0.7747 ± 0.009	GRU-PCA	0.8236 ± 0.007
LR-MissForest	0.7279 ± 0.016	SVM-MissForest	0.7482 ± 0.016	RF-MissForest	0.7858 ± 0.010	GRU-MissForest	0.8239 ± 0.006
						Proposed GRU-D	**0.8527 ± 0.003**

Outline

1. What can we do with supervised learning?

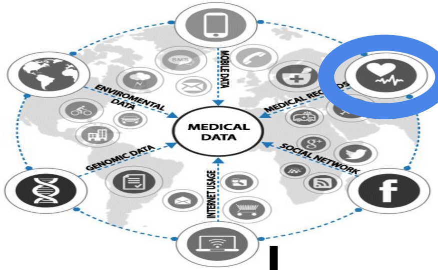
2. **Case study on intervention predictions:**
    a. **Frame the problem**
    b. Evaluation
    c. Iterate

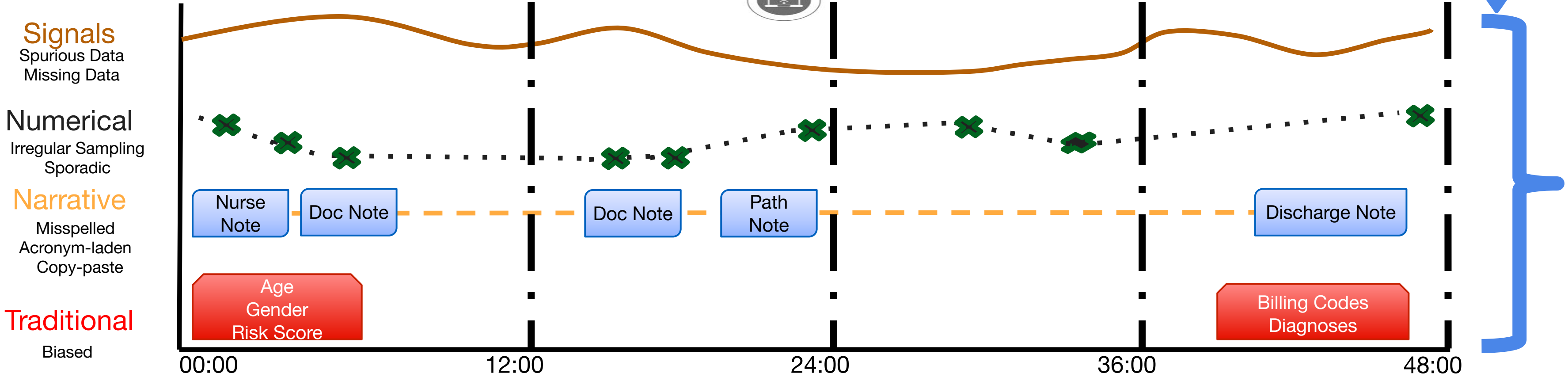
3. Survival Analysis

4. What else should we be thinking about?

MIMIC III ICU Data

- Learning with real patient data from the Beth Israel Deaconess Medical Center ICU. [1]

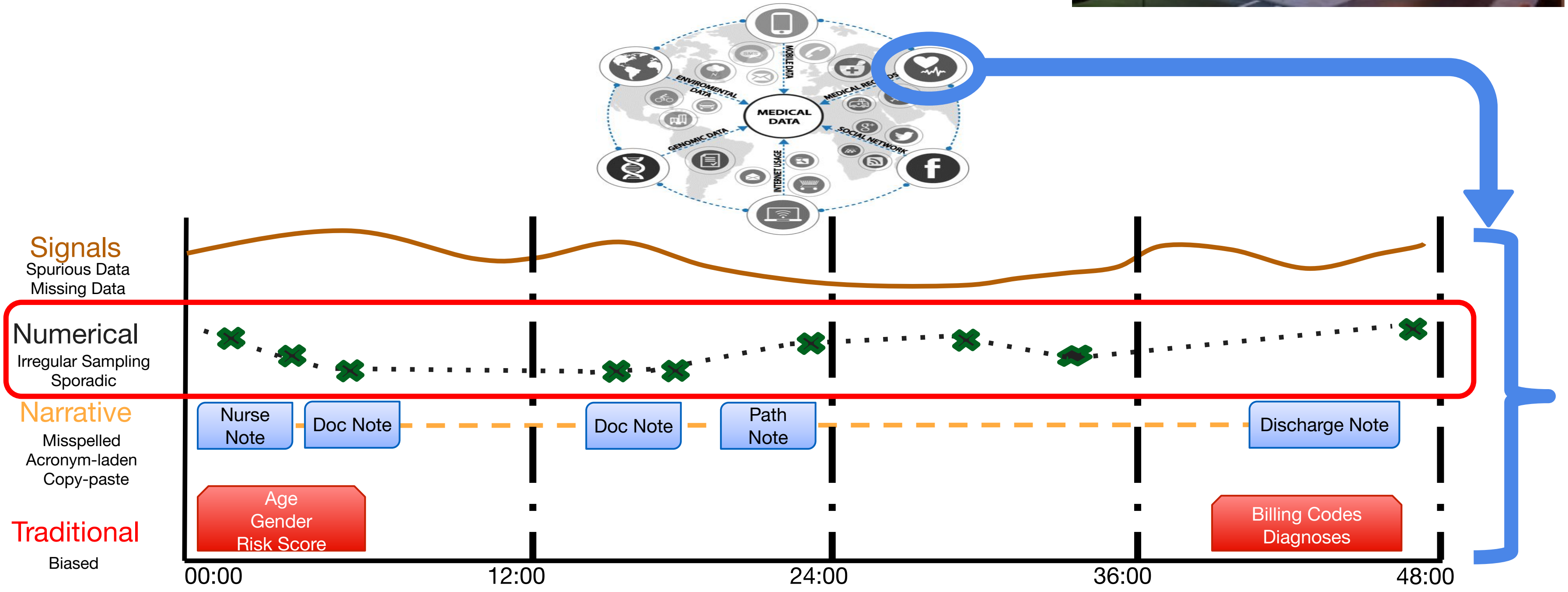




[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3 (2016).

MIMIC III ICU Data

- Learning with real patient data from the Beth Israel Deaconess Medical Center ICU. [1]



[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3 (2016).

Example: Early prediction of vasopressor interventions

- Vasopressors are a **common** drug to raise blood pressure.
- All drugs can be **harmful**, we'd like to avoid when possible.[1,2]
- Assume that real **clinical** actions are good learning **data**.
- Predict **upcoming interventions** based on evidence.[3,4]

[1] Müllner, Marcus, Bernhard Urbanek, Christof Havel, Heidrun Losert, Gunnar Gamper, and Harald Herkner. "Vasopressors for shock." *The Cochrane Library* (2004).
[2] D'Aragon, Frederick, Emilie P. Belley-Cote, Maureen O. Meade, François Lauzier, Neill KJ Adhikari, Matthias Briel, Manoj Lalu et al. "Blood Pressure Targets For Vasopressor Therapy: A Systematic Review." *Shock* 43, no. 6 (2015): 530-539.
[3] Vincent, Jean-Louis, and Mervyn Singer. "Critical care: advances and future perspectives." The Lancet 376.9749 (2010): 1354-1361.
[4] Ospina-Tascón, Gustavo A., Gustavo Luiz Büchele, and Jean-Louis Vincent. "Multicenter, randomized, controlled trials evaluating mortality in intensive care: Doomed to fail?." Critical care medicine 36.4 (2008): 1311-1322.

Define clinically actionable prediction tasks:

Tasks:

1. Short Term (5-10 hr) Need:
   Predicts before a clinician would have given.

2. Imminent (< 4 hr) Need:
   Predict when a clinician would have given.

3. Weaning (< 4 hr):
   Predict when a doctor would have stopped.

Define clinically actionable prediction tasks:

Tasks:

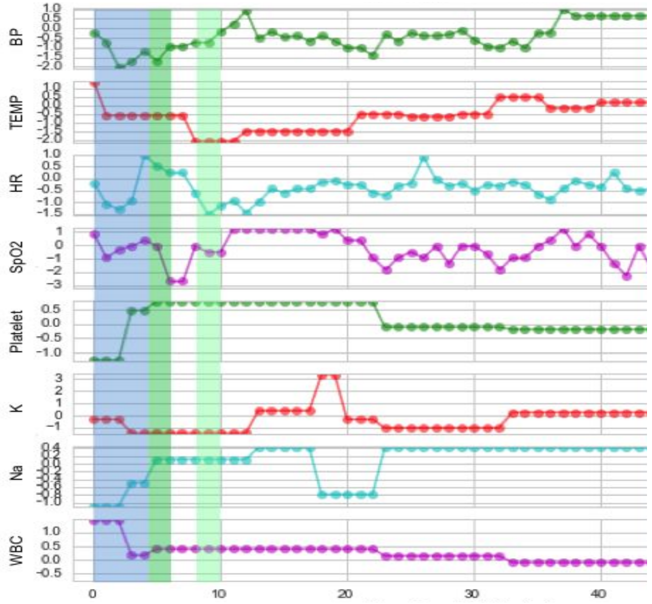
1. Short Term (5-10 hr) Need:
   Predicts before a clinician would have given.

2. Imminent (< 4 hr) Need:
   Predict when a clinician would have given.

3. Weaning (< 4 hr):
   Predict when a doctor would have stopped.

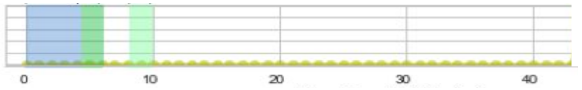
Define predictive task

Observe Physiological Signals



?

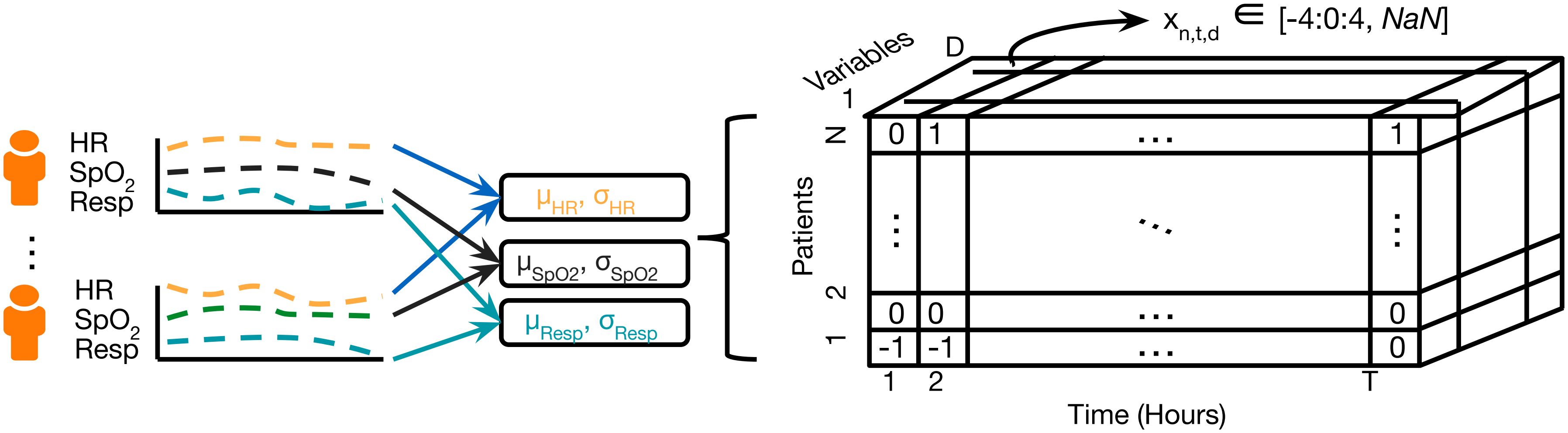
Every Hour
Predict Onset of Drug
Before the Doctor



24

Domain knowledge: Shared underlying physiological state

- **Z-score** (standardize) and **quantize** time series data.

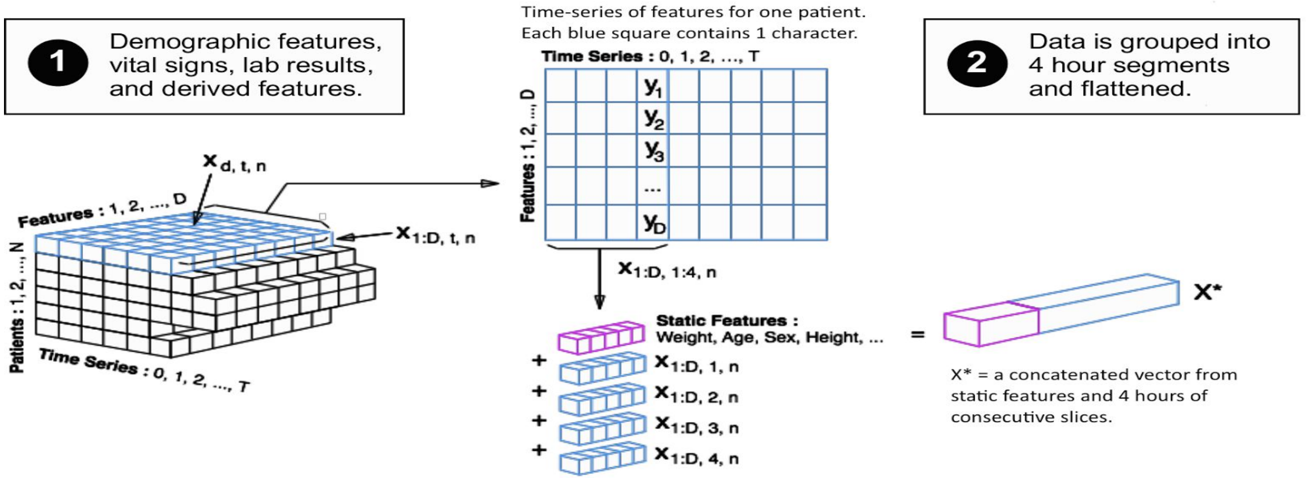


- Every $\mathbf{x}_{\mathbf{n,t,d}}$ is one of ten possible **characters**, -4:0:4 or *NaN*.

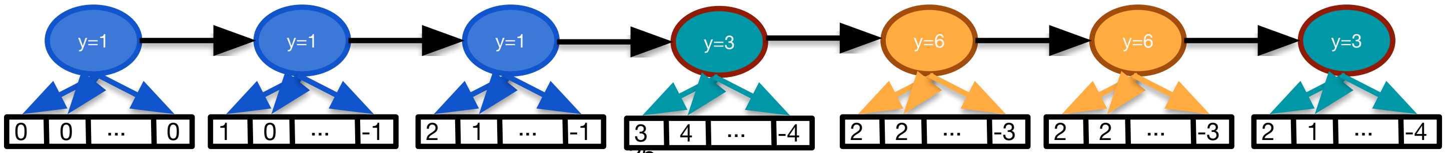
- Every $\mathbf{x}_{\mathbf{n,t}}$ is one of 10^D possible **words**.

Switching State Autoregressive Model Representation

- A patient *n* is a **sequence** of latent physiological **states** *y*.



- A physiological state *y* is a **distribution** over physiological words *x*.

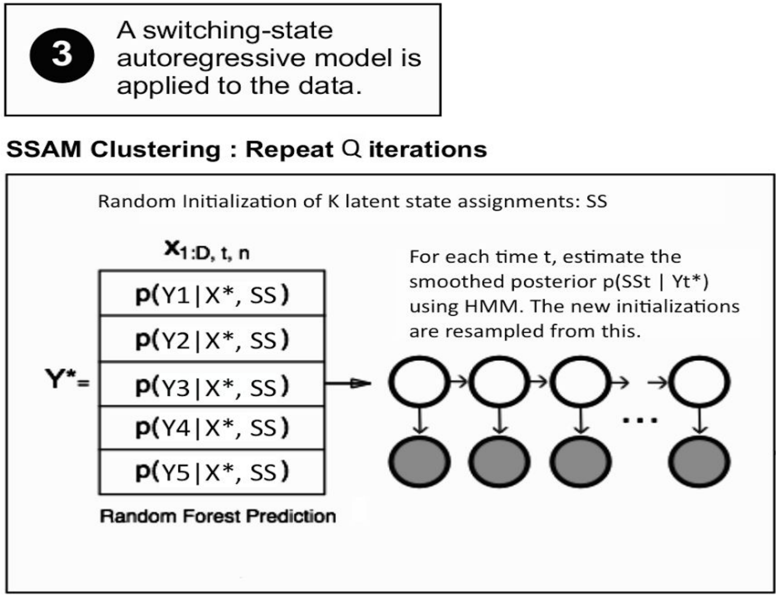


Extracting latent belief states from SSAM

- HMM sequence y^n_t on the signals x^n_t

$$\begin{aligned} y_t^n &\sim T_y(\cdot|y_{t-1}^n) \\ \boldsymbol{x}_t^n(p) &\sim T_x(\boldsymbol{x}_t^n(p)|\boldsymbol{x}_{t-1}^n, \theta_{p,y_{t-1}^n}) \end{aligned}$$

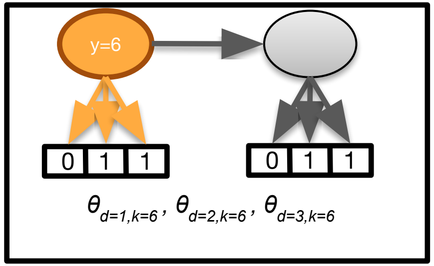
- x^n_t modeled by $T_x(\boldsymbol{x}'(p)|x,\theta)$; θ are governed by y^n_t.
- Each state 1… k has distinct set of parameters $\{\theta_{d,k}\}$, via K sets of tuples and D classifiers.
- Train $\theta_{d;k}$ to predict $x^n_t(d)|\ x^n_{t\text{-}4:t\text{-}1}$.
- Update state sequences y_t^n given $\{\theta_{d,k}\}$.



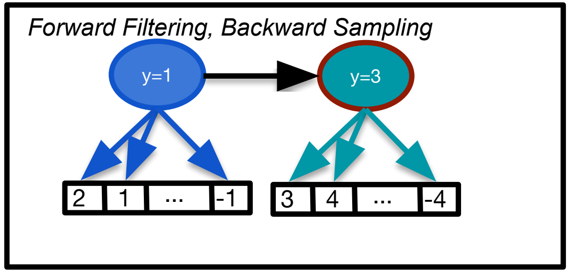
Discrete state space and per-variable missingness

- Use discrete state space.
- Model *NaN* (missing) as a valid emission.
- Cluster similar underlying states.
- For D variables and K latent states, perform inference iteratively:

1. Optimize parameters $\theta_{d,k}$

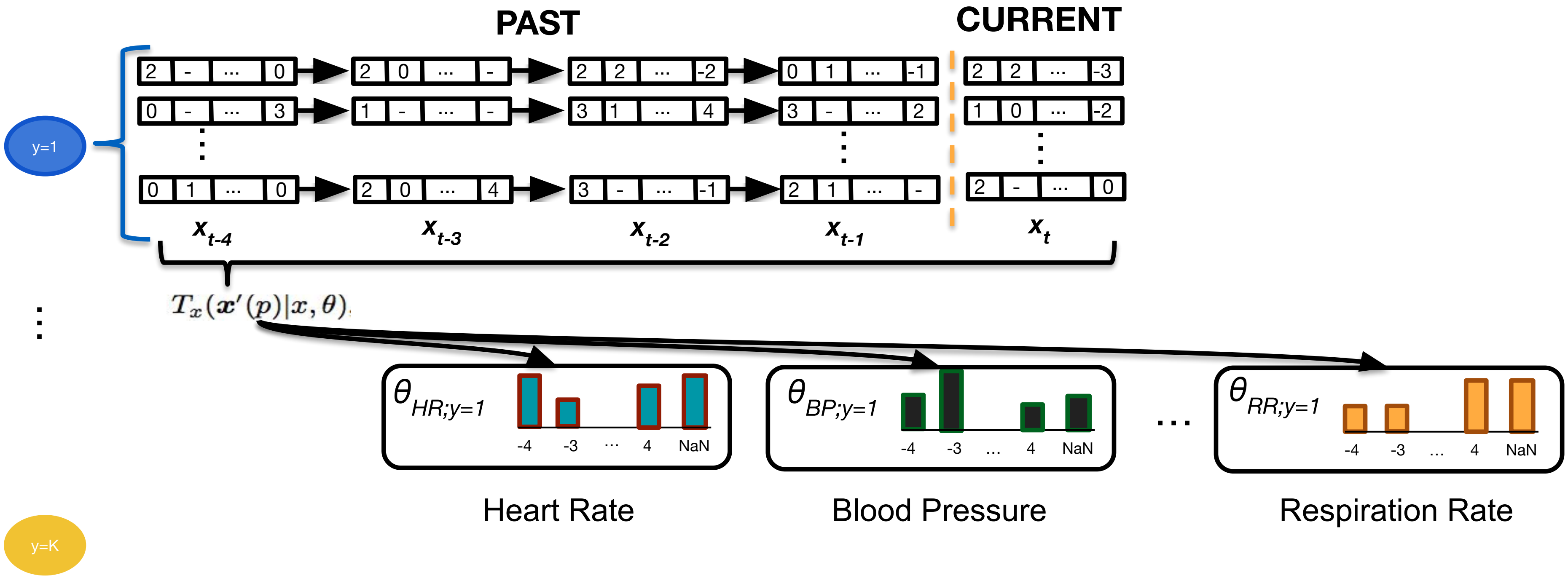


2. Sample states y^n_t



Distribution of values per-variable and latent

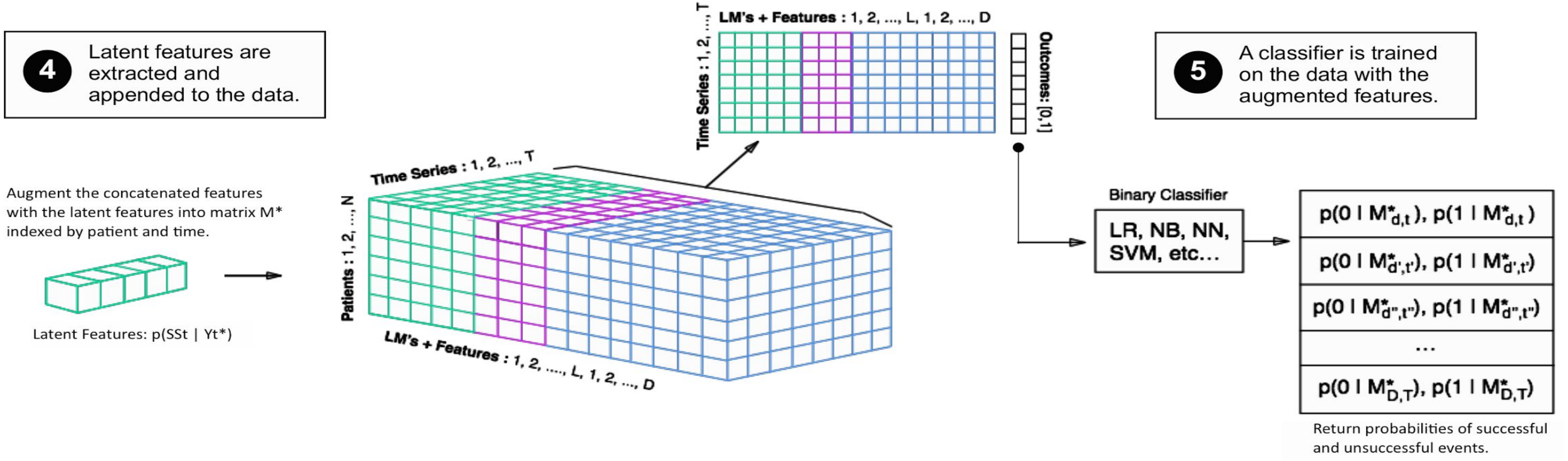
- Train parameters $\theta_{d;k}$ to predict $x^n{}_t(d)$ given $x^n{}_{t\text{-}4:t\text{-}1}$



Using SSAM for structured prediction

- SSAM states are **learned** in an **unsupervised** setting.

- **Evaluate** them in a **supervised** setting, on clinical tasks.



Outline

1. What can we do with supervised learning?

2. **Case study on intervention predictions:**
    a. Frame the problem
    b. **Evaluation**
    c. Iterate

3. Survival Analysis

4. What else should we be thinking about?

Previous work - use strong baselines

- **Baseline 1:** Prior work[1] predicted vasopressor onset in ICU patients with pre-treatment (fluids).
  - 2 hour gap
  - 3 demographics and 22 signals
  - AUC of 0.79

[1] Fialho, A. S., et al. "Disease-based modeling to predict fluid response in intensive care units." Methods Inf Med 52.6 (2013): 494-502.
* 2 hour gap, 22 derived/3 static features.

Vasopressor onset prediction beats SOTA results

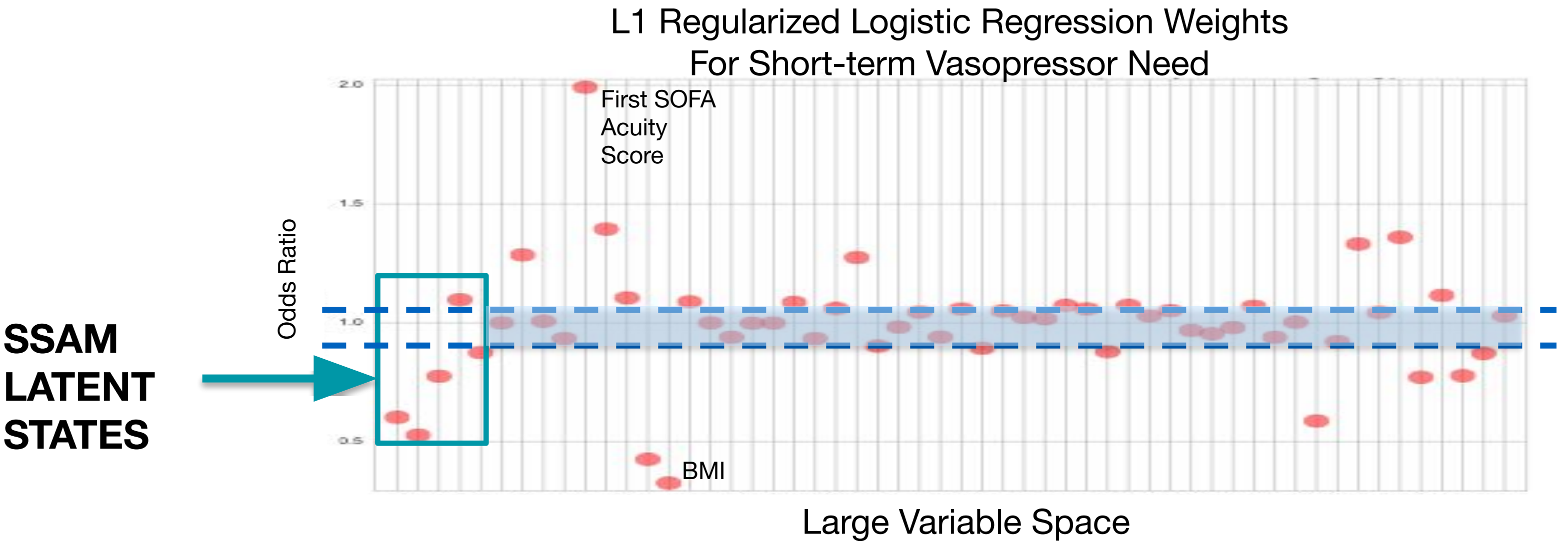
	AUC
Baseline 1 – Prior Work	0.79
Baseline 2 – Raw Data	0.83
SSAM Representations	0.83
Raw Data + SSAM Rep.	**0.88**

- **Latent** representations **add** predictive power.
- New state-of-the art prediction, 0.88 = thousands of **people treated early**!

[1] Fialho, A. S., et al. "Disease-based modeling to predict fluid response in intensive care units." Methods Inf Med 52.6 (2013): 494-502.
* 2 hour gap, 22 derived/3 static features.

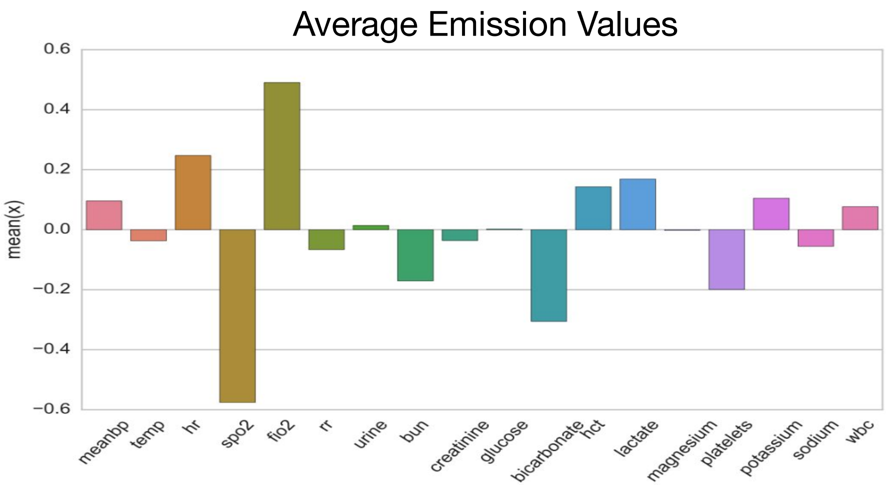
Regularized prediction emphasizes latent states

- **Latent states** are consistently **significant** across a large **variable space.**



Post-hoc justification

- Investigate state associated with vasopressor onset?



- Low average values of blood oxygenation and bicarbonate.

- Highest lactate levels of any state.

35

Similar trends in other predictive tasks

	Short-Term Need (Gapped AUC)	**Imminent Need (Ungapped AUC)**	**Weaning**
Baseline 1 – Prior Work	0.79	-	-
Baseline 2 – Raw Data	0.83	0.89	0.67
SSAM Representations	0.83	0.87	0.63
Raw Data + SSAM Rep.	**0.88**	**0.92**	**0.71**

- Our representations are **useful abstractions** for **multiple tasks.**

Outline

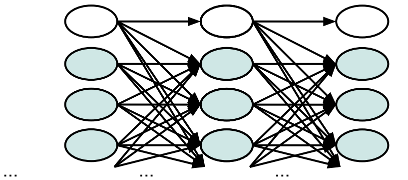
1. What can we do with supervised learning?

2. **Case study on intervention predictions:**
   a. Frame the problem
   b. Evaluation
   c. **Iterate**

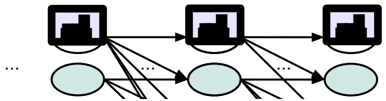
3. Survival Analysis

4. What else should we be thinking about?

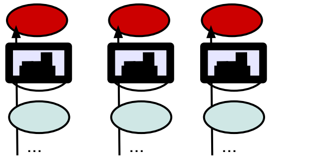
More outcomes and improved dynamics



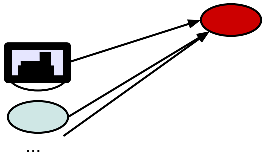
Learn model parameters over patients with variational EM.



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



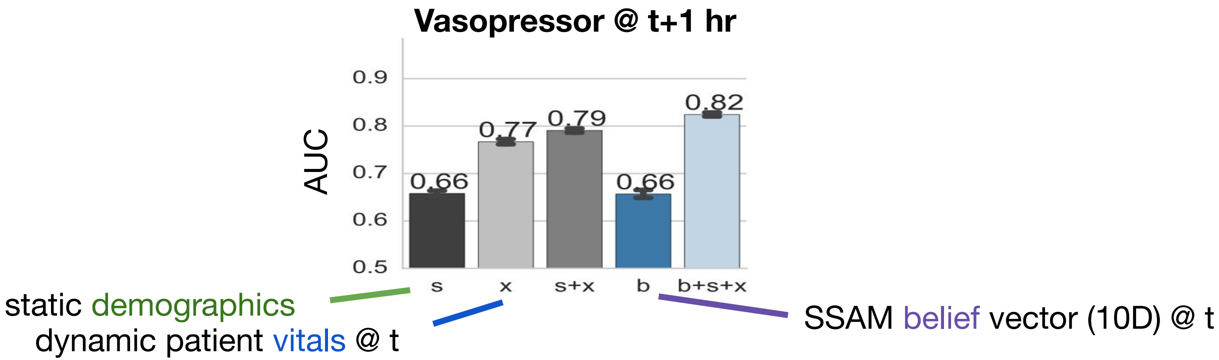
Logistic regression (with label-balanced cost function)

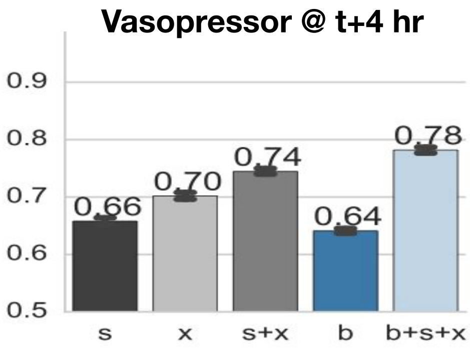


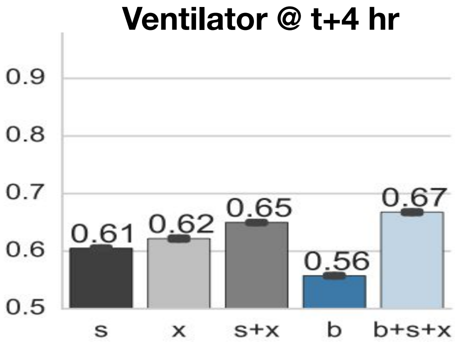
Predict onset in advance

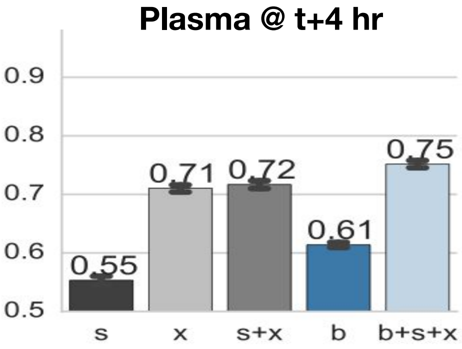
- More Interventions: fresh-frozen-plasma transfusion (ffp), platelet transfusion, red-blood-cell (rbc) transfusion, vasopressor administration, and ventilator intubation.
- Gaussian Emission Model for Dynamics:
  - Static observations s (10 dimensions using one-hot encoding),
  - Dynamic time-series observations x (18 dimensions)
  - Belief state vectors b (K=10 dimensions) from the switching state model forward belief state

State space beliefs improve prediction









Outline

1. What can we do with supervised learning?

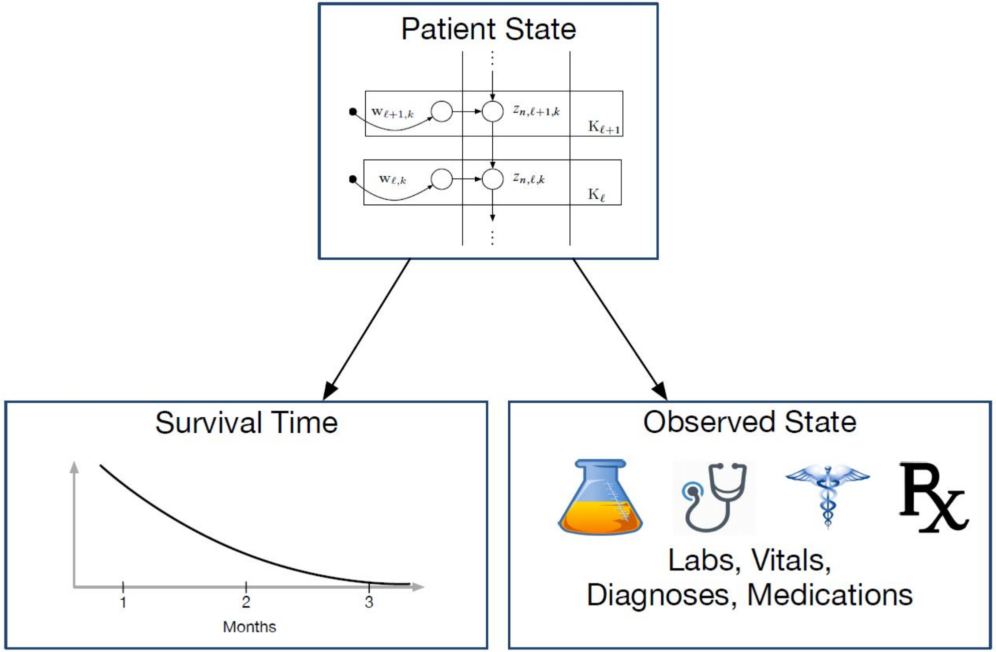
2. Case study on intervention predictions:
   a. Frame the problem
   b. Evaluation
   c. Iterate

3. **Survival Analysis**

4. What else should we be thinking about?

Survival Analysis

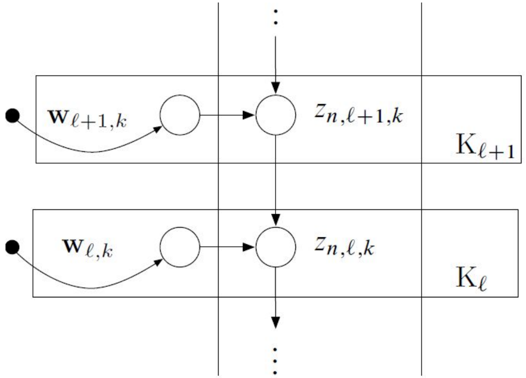
- Survival Analysis studies the time to an event.
- Commonly used in EHR for "time to" discharge/death/etc.
- We need **flexible hidden structures** to describe patient state.



Deep Exponential Families

- **x**, the set of covariates
- **β**, the parameters for the data with some prior **p(β)**
- **k**, a fixed scalar
- **n**, the index to an observation
- **z**, the latent variable
- **L**, the number of layers of latent variables each observation has

Deep Exponential Families



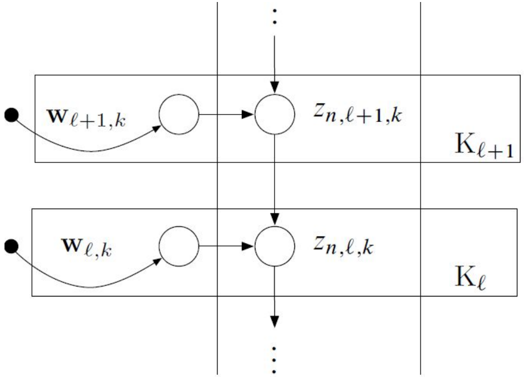
$$z_{n,\ell,k} \sim \text{EXP-FAM}(g(\mathbf{w}_{\ell,k}^{\top}\mathbf{z}_{n,\ell+1}))$$

- Use a DEF to represent state.
- All distributions are canonical in exponential family form

$$p(z_{n,\ell,k} \mid \mathbf{z}_{n,\ell+1}, \mathbf{w}_{\ell,k}) = \exp\{\eta(\cdot)^{\top} t(z_{n,\ell,k}) - a(\eta(\cdot))\}$$
$$\eta(\cdot) = g(\mathbf{z}_{n,\ell+1}^{\top} \mathbf{w}_{\ell,k})$$

- More general functions can also be used.

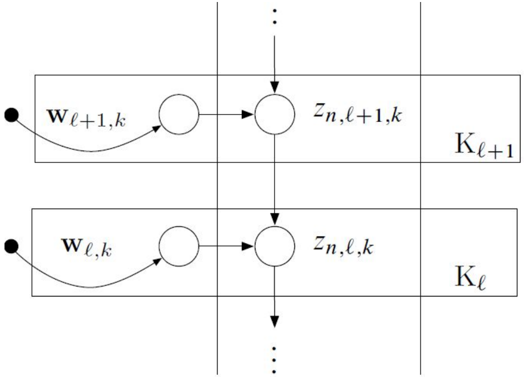
Deep Exponential Families



$$z_{n,\ell,k} \sim \text{Exp-Fam}(g(\mathbf{w}_{\ell,k}^{\top}\mathbf{z}_{n,\ell+1}))$$

- Possibilities for the hidden layers
    - Binary: Bernoulli
    - Count: Poisson
    - Non-negative (and sparse): Gamma
    - Real-valued: Gaussian

Deep Exponential Families



$$z_{n,\ell,k} \sim \text{EXP-FAM}(g(\mathbf{w}_{\ell,k}^\top \mathbf{z}_{n,\ell+1}))$$

- Many existing models are DEFs
    - Mixture models
    - Factorial mixture models [Ghahramani+ 1995]
    - Poisson factorization [Canny+ 2004]
    - Exponential family factor analysis [Mohamed+ 2008]
    - Correlated topic models [Blei+ 2007]

Deep Survival Analysis

$$
\begin{aligned}
b &\sim \text{Normal}(0, \sigma_b) \\
a &\sim \text{Normal}(0, \sigma_W) \\
z_n &\sim \text{DEF}(\mathbf{W}) \\
\mathbf{x}_n &\sim p(\cdot \mid \boldsymbol{\beta}, z_n) \\
t_n &\sim \text{Weibull}(\log(1 + \exp(z_n^\top a + b), k)
\end{aligned}
$$

- Use the Weibull distribution to model failure times as its cdf and pdf are both analytically tractable.

Deep Survival Analysis

$$
\begin{aligned}
b &\sim \text{Normal}(0, \sigma_b) \\
a &\sim \text{Normal}(0, \sigma_W) \\
z_n &\sim \text{DEF}(\mathbf{W}) \\
\mathbf{x}_n &\sim p(\cdot \mid \boldsymbol{\beta}, z_n) \\
t_n &\sim \text{Weibull}(\log(1 + \exp(z_n^\top a + b), k)
\end{aligned}
$$

- $\mathbf{x}_n$ can be missing ✔
- Relationships flexible through latent space ✔
- Censoring through tractable CDF ✔
- Make predictions via posterior inference
    - Works empirically! ✔

Predicting CHD from EHR

- 300K individuals from a large metropolitan hospital
- Adults with at least 5 interactions with the hospital's network
- Covariates:
    - 9 vital signs
    - 80 laboratory test measurements
    - 5K medication orders
    - 13K diagnosis
- Data aggregated at a month level
- CHD events were defined by the occurrence of
    - 413 (angina pectoris)
    - 410 (myocardial infarction)
    - 411 (coronary insufficiency)

Results

Model	Concordance (%)
Baseline Framingham Risk Score	**65.57**
Deep Survival Analysis; K=10	69.35
Deep Survival Analysis; K=5	70.45
Deep Survival Analysis; K=25	71.20
Deep Survival Analysis; K=75	71.65
Deep Survival Analysis; K=100	72.71
Deep Survival Analysis; K=50	**73.11**

Table 1: Concordance on a held-out set of 25,000 patients for different values of K and for the baseline risk score. All deep survival analysis dimensionalities outperform the baseline.

- It works, but remember!
  - Survival analysis is conditional distribution modeling
  - Imputation not useful for pure predictions
  - Reduces to deep-multiclass regression with missingness indicators

Outline

1. What can we do with supervised learning?

2. Case study on intervention predictions:
    a. Frame the problem
    b. Evaluation
    c. Iterate

3. Survival Analysis

4. **What else should we be thinking about?**

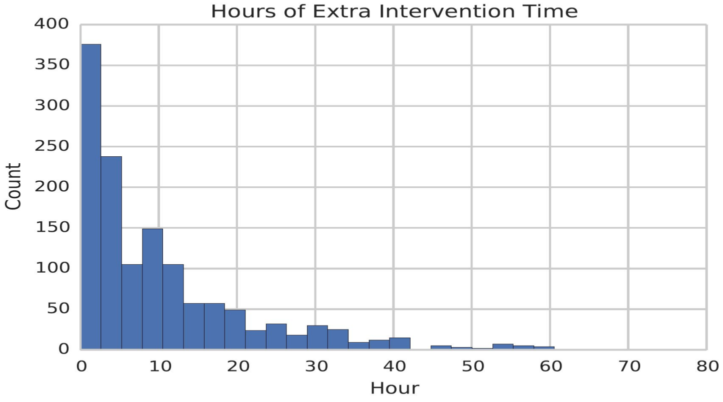
Similar trends in other tasks, except!

	Short-Term Need (Gapped AUC)	**Imminent Need (Ungapped AUC)**	**Weaning**
Baseline 1 – Prior Work	0.79	-	-
Baseline 2 – Raw Data	0.83	0.89	0.67
SSAM Representations	0.83	0.87	0.63
Raw Data + SSAM Rep.	**0.88**	**0.92**	**0.71**

- For the patients with vasopressors, we often predicted an early wean.

What exactly are we learning?

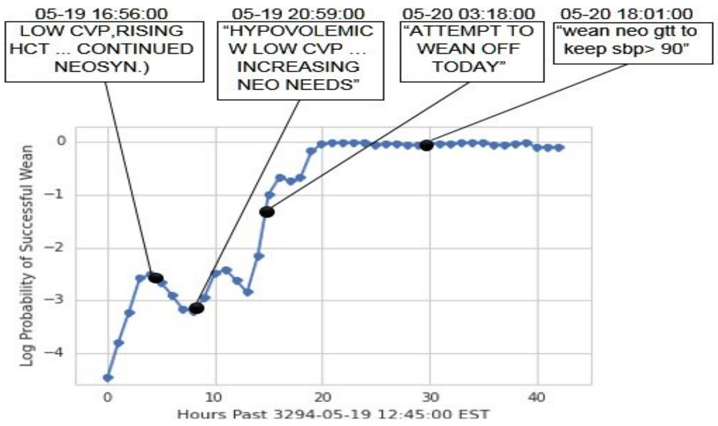
- Patients can be left on interventions longer than necessary.



- Extended interventions can be costly and detrimental to patient health.[1,2]

[1] Müllner, Marcus, Bernhard Urbanek, Christof Havel, Heidrun Losert, Gunnar Gamper, and Harald Herkner. "Vasopressors for shock." *The Cochrane Library* (2004).
[2] D'Aragon, Frederick, Emilie P. Belley-Cote, Maureen O. Meade, François Lauzier, Neill KJ Adhikari, Matthias Briel, Manoj Lalu et al. "Blood Pressure Targets For Vasopressor Therapy: A Systematic Review." *Shock* 43, no. 6 (2015): 530-539.

Finding where we “could” wean early?



- One example of a 62-year-old male patient with a cardiac catheterization.
- More complexity/higher misclassification penalty don’t solve this!

Missingness and representation

- How do we represent missing data?
- If we remove patients via a threshold, what groups are impacted?

Biases in electronic health record data due to processes within the healthcare system: retrospective observational study

Denis Agniel,[1] Isaac S Kohane,[1,2] Griffin M Weber[1,3]

ABSTRACT

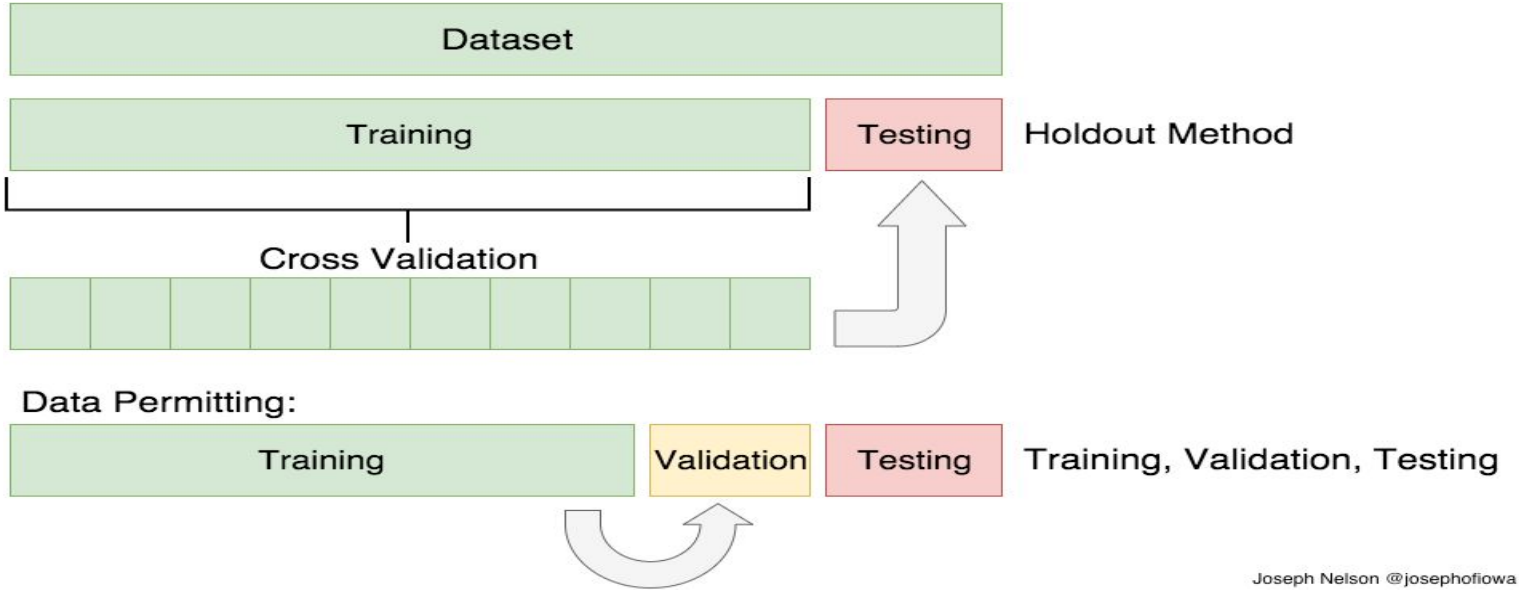
OBJECTIVE

To evaluate on a large scale, across 272 common types of laboratory tests, the impact of healthcare processes on the predictive value of electronic health record (EHR) data.

the routine delivery of healthcare.[1-3] This, in turn, is transforming biomedical research as investigators now have access to information on millions of patients through informatics tools that can query and analyze EHRs,[4-7] link to genomic and other types of biomedical data,[8 9] and scale to a national level and beyond.[10-14]

“Doctors typically do not **order a white blood cell** count test for a **patient on the weekend** or for a patient who **just had a white blood cell count** less than one day earlier, unless **they believe the patient is sick**.”

Details in training can be impactful



- Split by patient… generalize to new subjects?
- Split by hospital site… generalize to new doctors?
- Split by year… generalize to new policies?

Slide courtesy of Michael Hughes

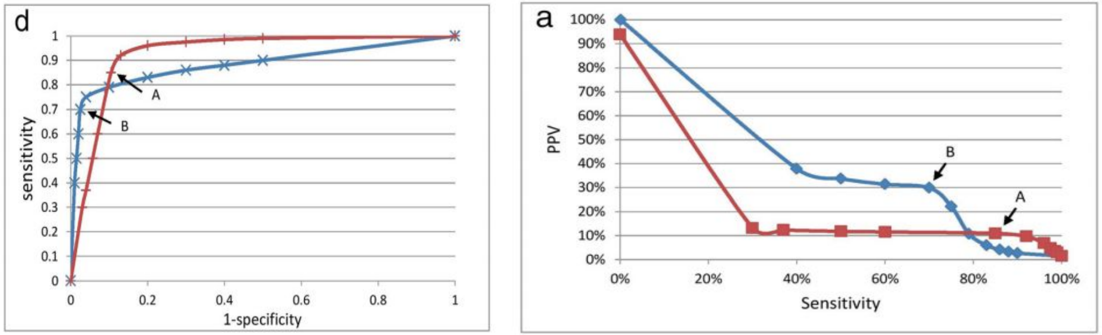
Careful evaluation is extremely important

- Spend as much time designing evaluation as with model prototyping.
- Make diagnostic plots, not just tables, and think about actual utility.

Why the *C*-statistic is not informative to evaluate early warning scores and what metrics to use

Santiago Romero-Brufau[1,2*], Jeanne M. Huddleston[1,2,3], Gabriel J. Escobar[4] and Mark Liebow[5]

By AUC... red is better



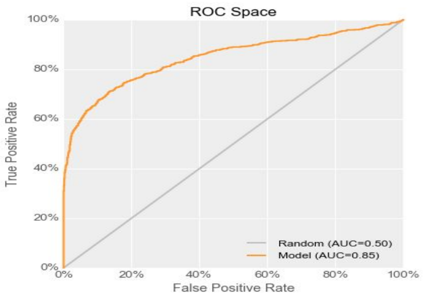
But blue is much better for alarm fatigue

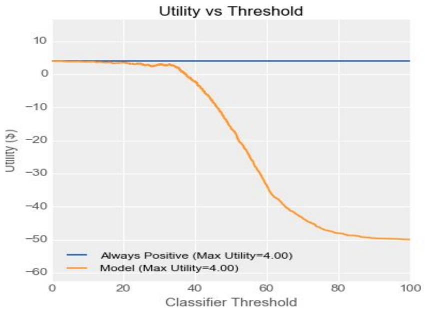
Calibration matters in practice

- What is the cost of an incorrect decision?

	Good	Bad
Positive	True Positive $utility = +\$20$ $rate(t) = TPR(t) \cdot 95\%$	False Positive $utility = -\$300$ $rate(t) = FPR(t) \cdot 5\%$
Negative	False Negative $utility = -\$50$ $rate(t) = (1 - TPR(t)) \cdot 95\%$	True Negative $utility = -\$50$ $rate(t) = (1 - FPR(t)) \cdot 5\%$

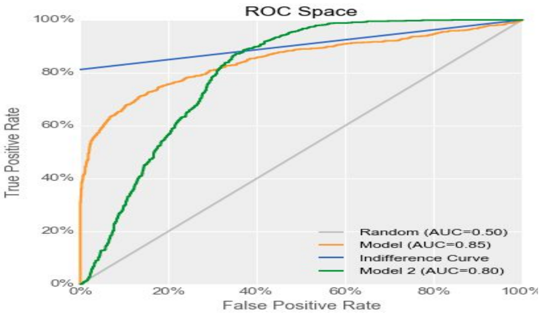
vs.

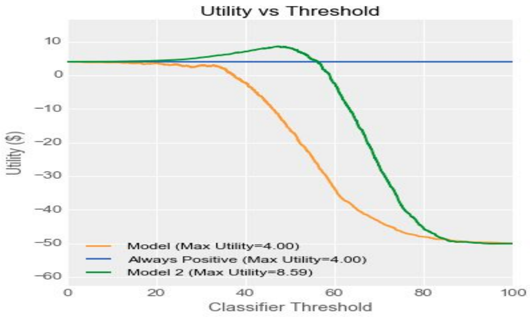




- Domain specific evaluation requires a goal.

Model 2 (green) has lower AUC





… but has operating points with much higher utility!

Slide courtesy of Michael Hughes

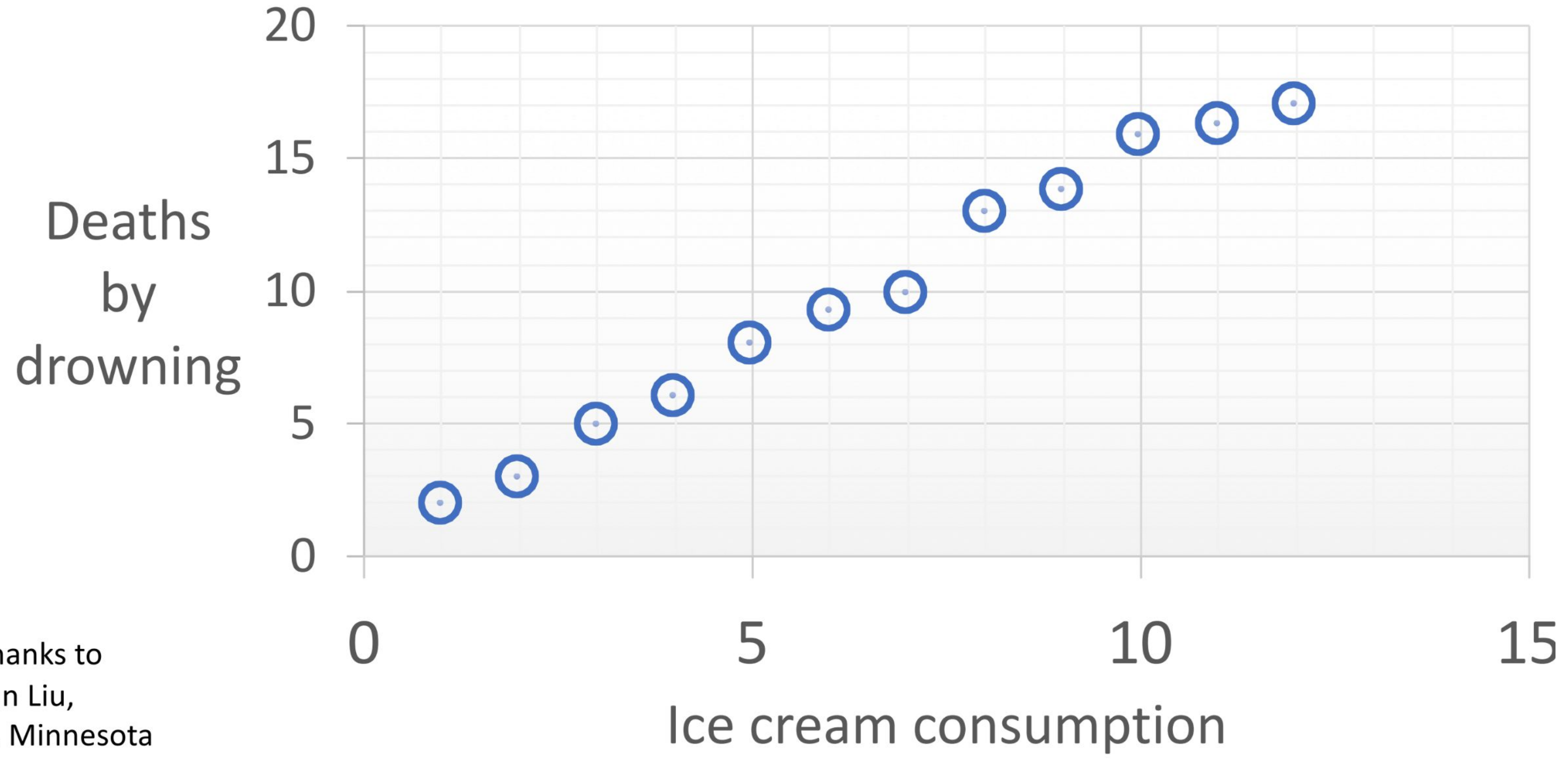
Credit: Nicolas Kruchten
http://blog.mldb.ai/blog/posts/2016/01/ml-meets-economics/

Causality is looming in healthcare

- Question: Who will be diabetic in 1 year?
- We build predictive model:
  features X = [lab_tests, diagnoses, medications]
  label y = [diabetic]
- We can predict y from X with AUC 0.8
- What **action** do we take with this knowledge?

Slide courtesy of Uri Shalit

Can you spot the confounding?



Thanks to
Lan Liu,
U. Minnesota

Slide courtesy of Uri Shalit

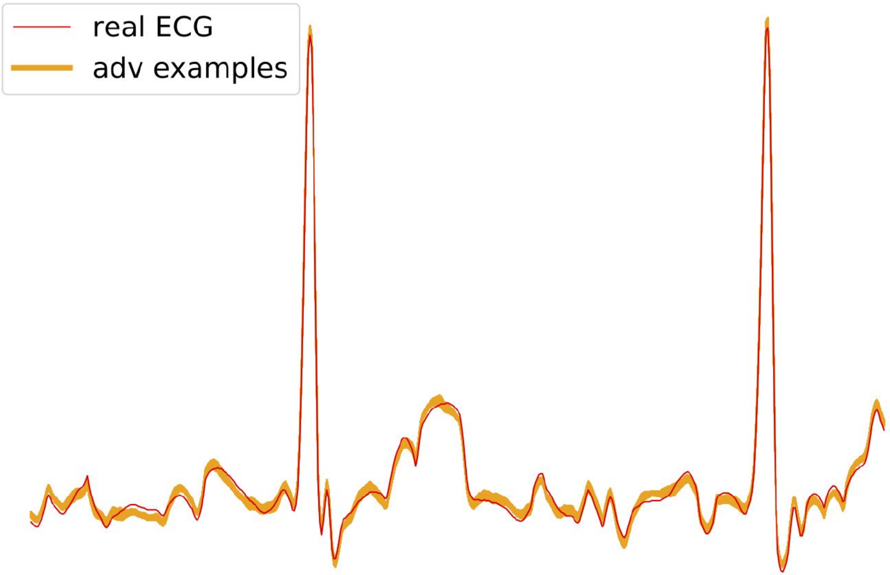
Remember Adversarial Examples?

- How hard is it to adversarially fool networks?

- Remember that bad loss means misclassification, and:
  1. Start with trained model
  2. Compute gradient with respect to loss function with respect to input
  3. Follow gradient to increase the loss
  4. Limit the movement to a norm

- Popular technique: Projected Gradient Descent [Madry+ 2017]

Adversarial Examples Are Not Rare



- Smooth adversarial perturbations that fool networks exist for over 85% of ECG tracings in the 2017 PhysioNet Challenge.