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!

- CS2541 capped to students with an appropriate background (quiz/email)!
- 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!

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

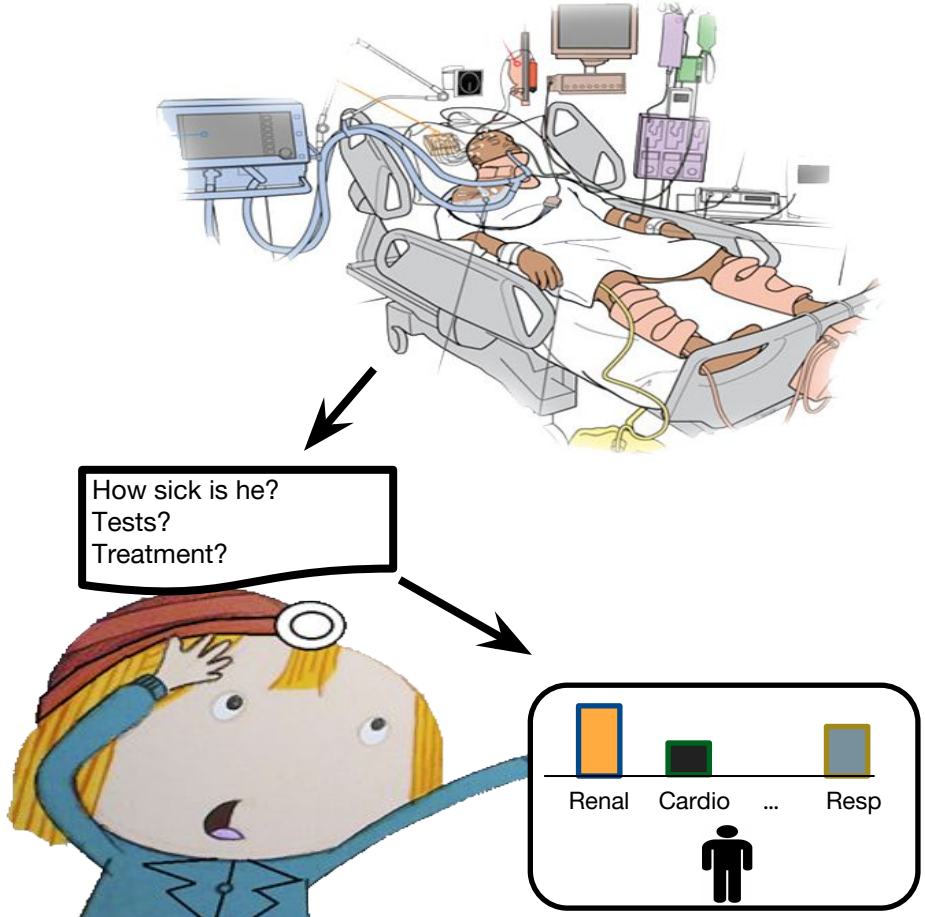
Feb 7, 2019, Lecture 5: Clinical Time Series Modelling (Homework 1 due at 11:59 PM on MarkUs)
Feb 14, 2019, Lecture 6: Clinical Imaging (Project proposals due at 5PM on MarkUs)
Feb 21, 2019, Lecture 7: Clinical NLP and Audio

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)

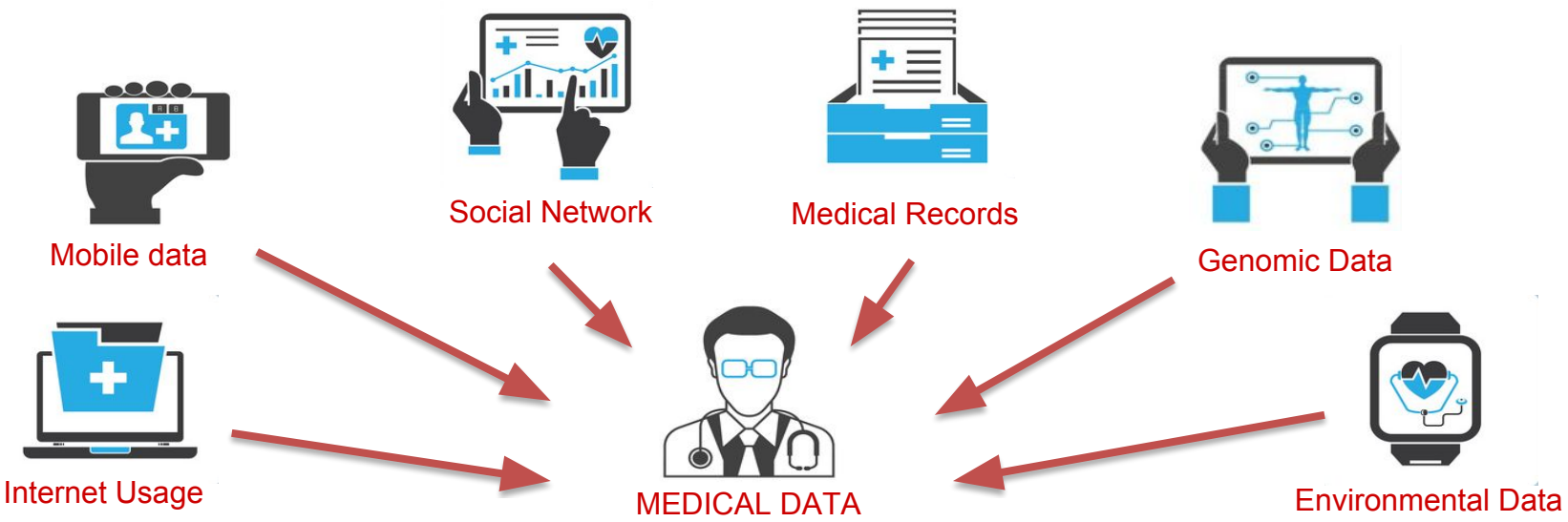
Clinicians Need to Estimate Patient State and Predict Outcome

- 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?



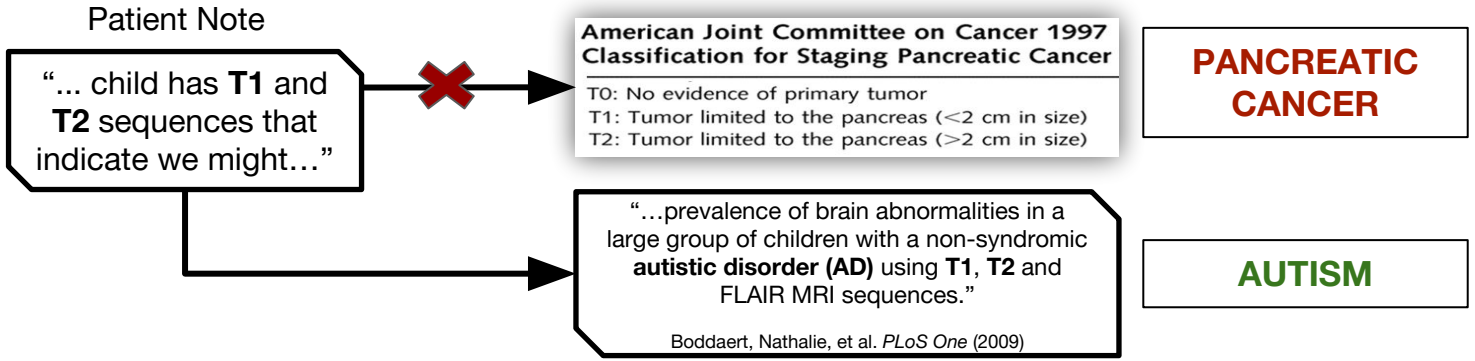
Our goal is machine learning what is healthy

Can we use **data** to **learn** what is **healthy**?



Caution required… Autism isn't correlated with cancer

- **Claim:** Autism is correlated with cancer!
- **Methods:** State-of-the-art NLP techniques out of the box.
- **Results:** Every autistic patient has cancer.



Health requires **domain integration**.

Outline

1. What can we do with supervised learning?

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

3. What else should we be thinking about?

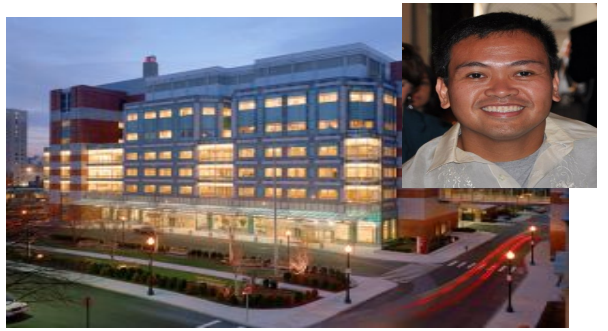
Outline

1. **What can we do with supervised learning?**

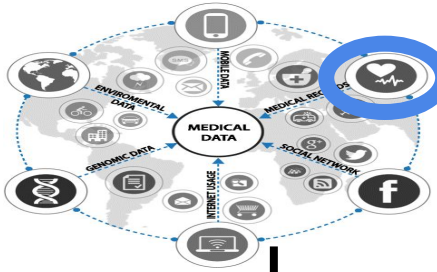
2. Case study on intervention predictions:
   a. Frame the problem
   b. Evaluation
   c. Iterate

3. 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]



Signals
Spurious Data
Missing Data

Numerical
Irregular Sampling
Sporadic

Narrative
Misspelled
Acronym-laden
Copy-paste

Biased

Traditional

Nurse Note | Doc Note | Doc Note | Path Note | Discharge Note

Age
Gender
Risk Score

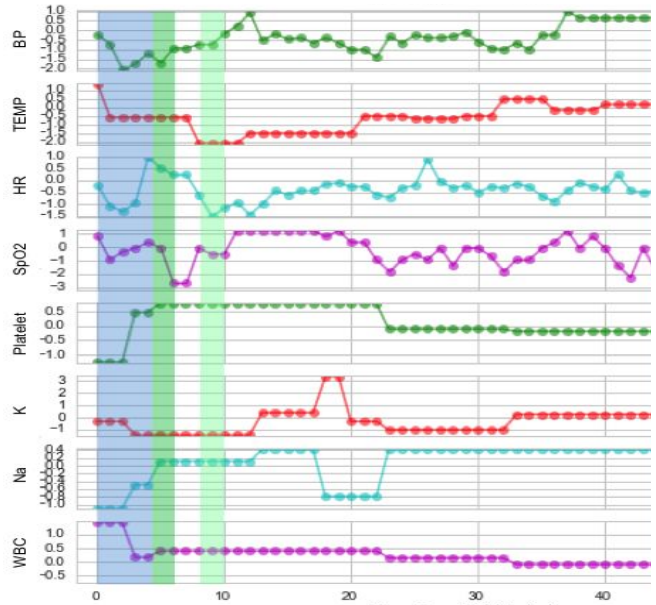
Billing Codes
Diagnoses

00:00 | 12:00 | 24:00 | 36:00 | 48:00

[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

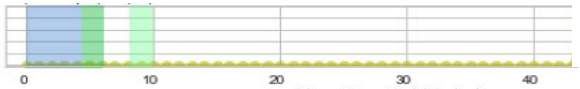


?

“Real-time” **Prediction**

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

By Gap Time

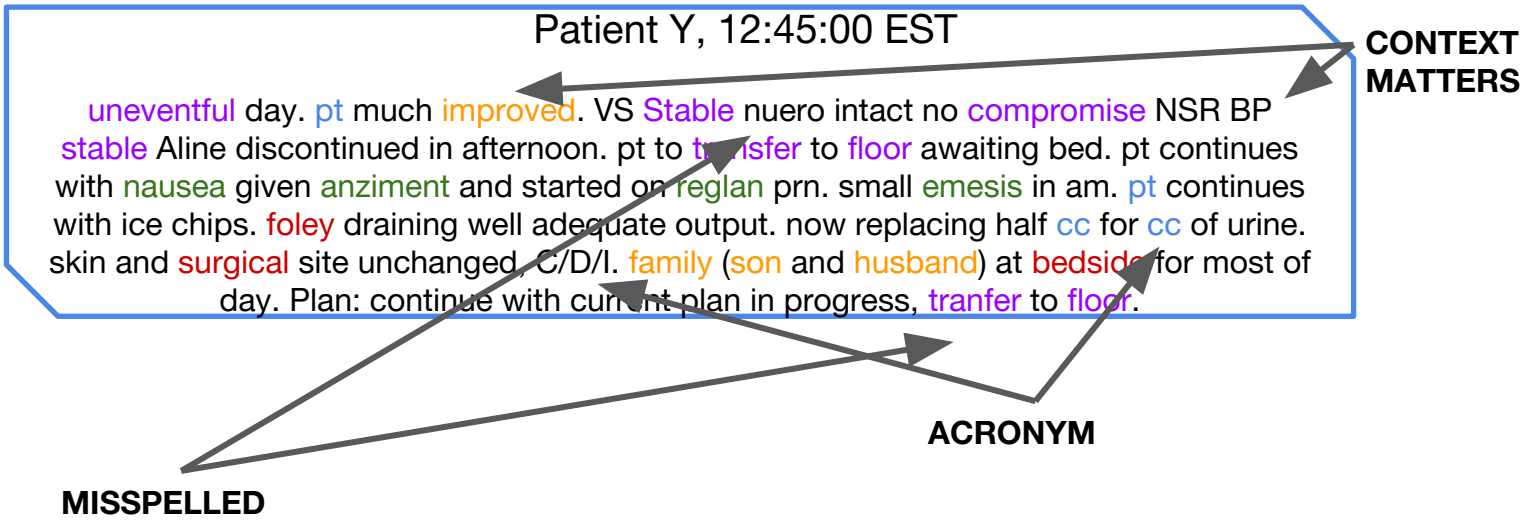


Part 1: Predict **mortality** with clinical **notes**

- **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...



Represent patients as topic vectors

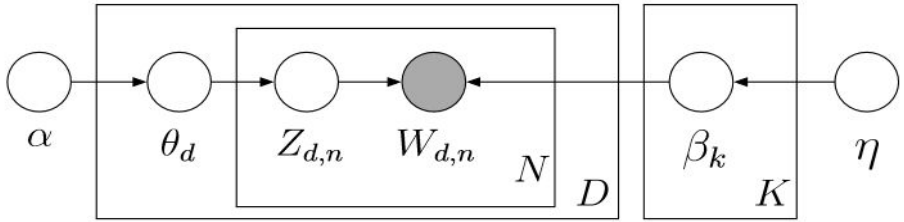
- Model patient stays as an **aggregated set** of notes.
- Model notes as a **distribution** over topics.
- A “topic” is a **distribution** over words, that we learn.

- 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

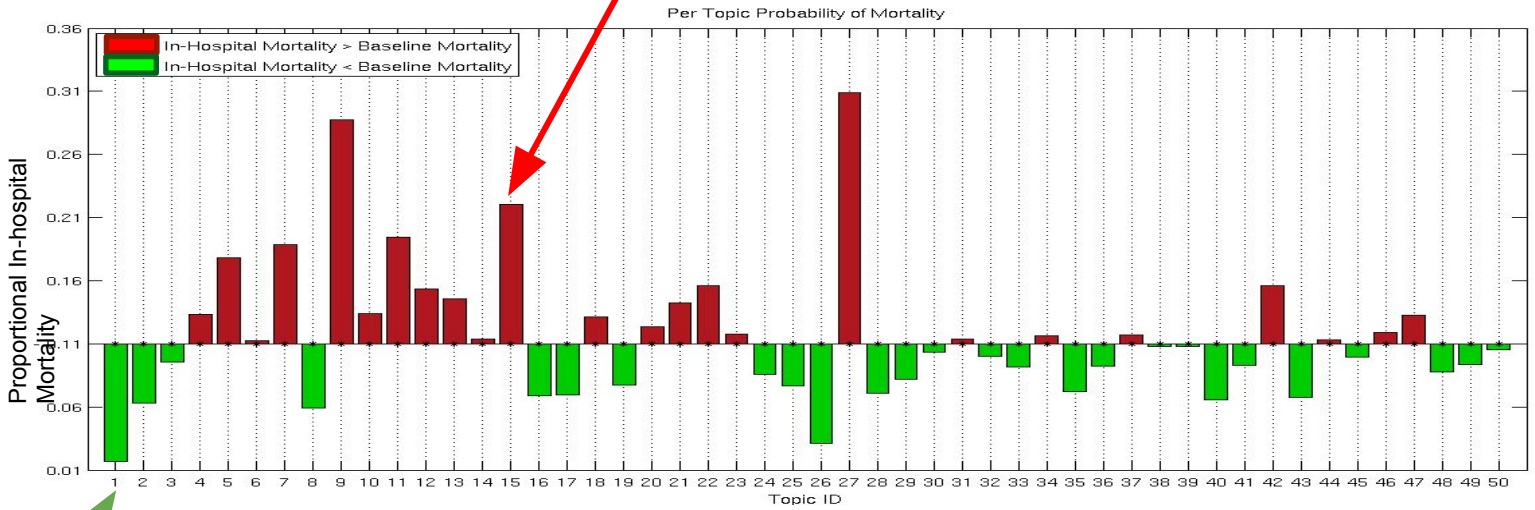
Sparsity α

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

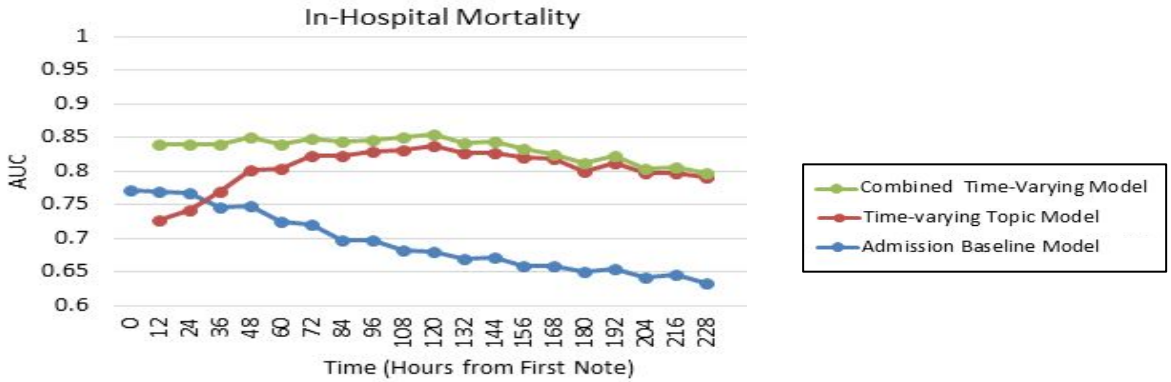
Correlation between average topic and mortality

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



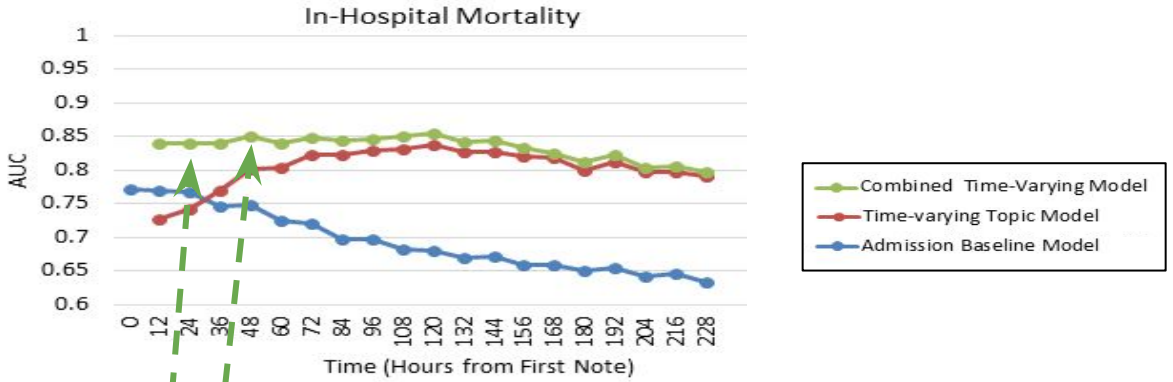
Topic #	Top Ten Words	Possible Topic
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp	Cardiovascular surgery

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



More Complex ≠ 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

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.

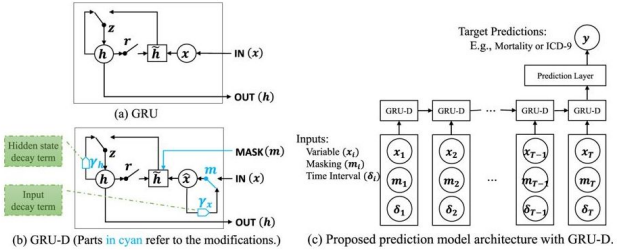
Che, Zhengping, et al. "Deep computational phenotyping." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

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.



- 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**

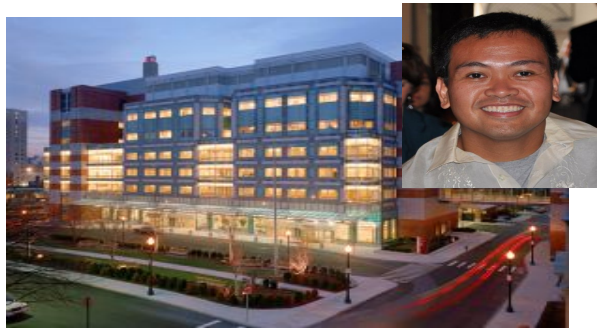
Outline

1. What can we do with supervised learning?

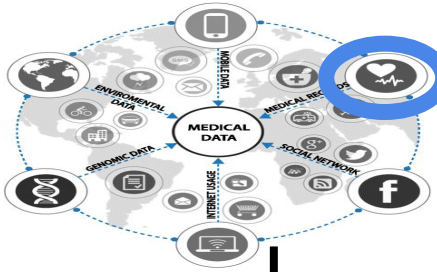
2. **Case study on intervention predictions:**
   a. **Frame the problem**
   b. Evaluation
   c. Iterate

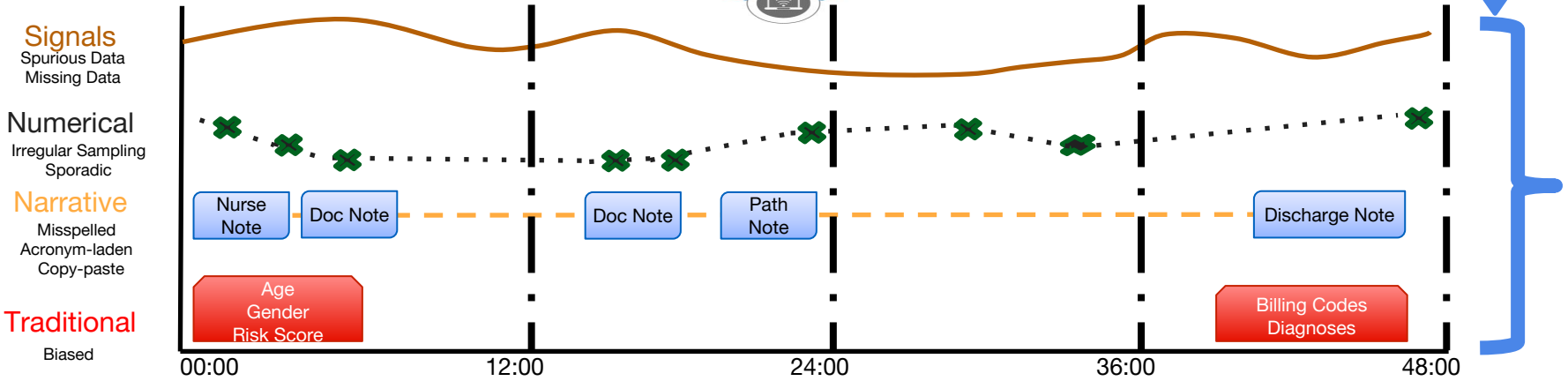
3. What else should we be thinking about?

MIMIC III ICU Data



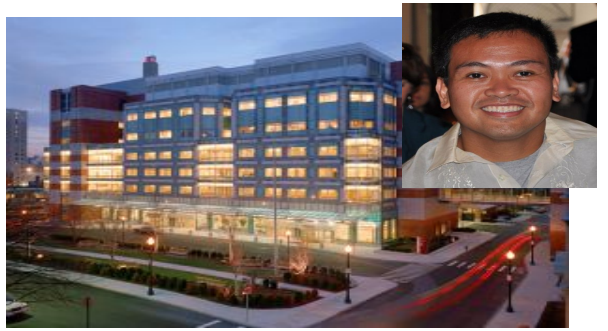
- Learning with real patient data from the Beth Israel Deaconess Medical Center ICU. [1]



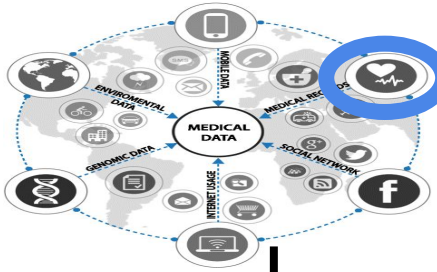


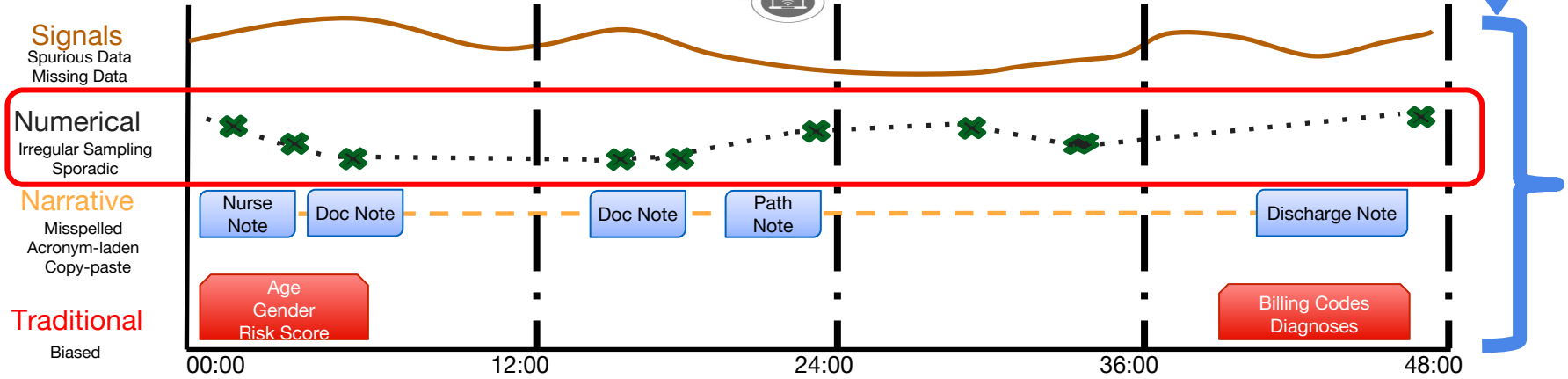
[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3 (2016).

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:

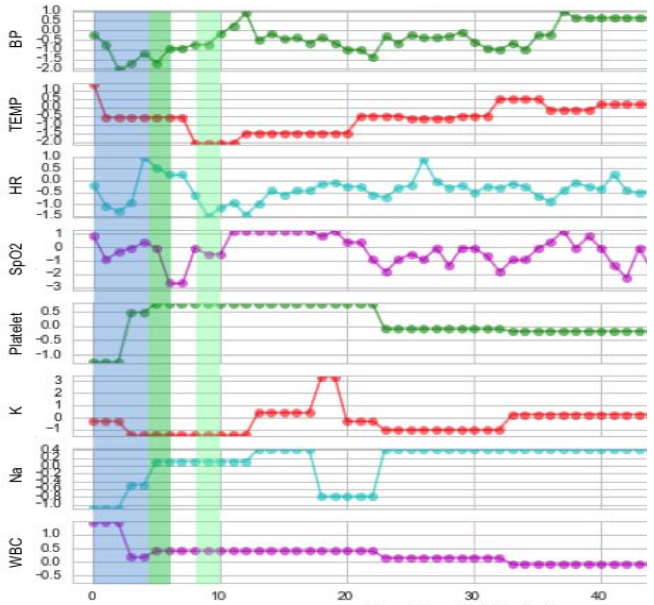
1. Short Term (5-10 hr) Need:
   Predicts before a clinician would have given.

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

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

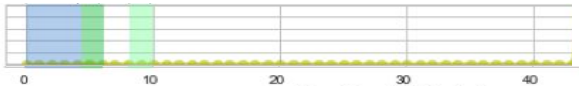
Define predictive task

Observe Physiological Signals



?

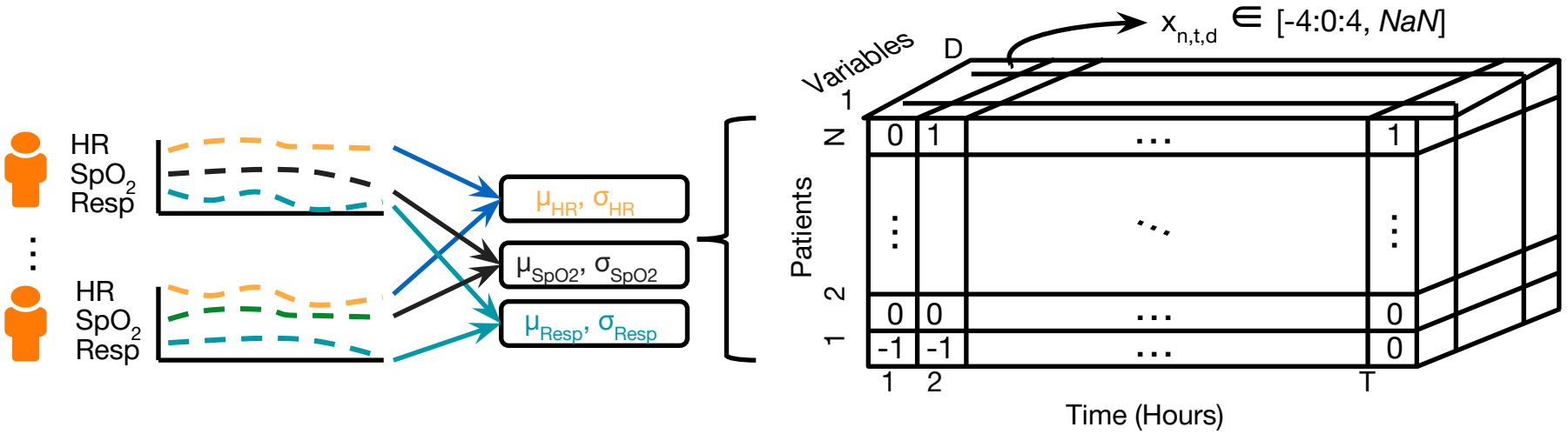
Every Hour
Predict Onset of Drug
Before the Doctor



25

Domain knowledge: Shared underlying physiological state

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

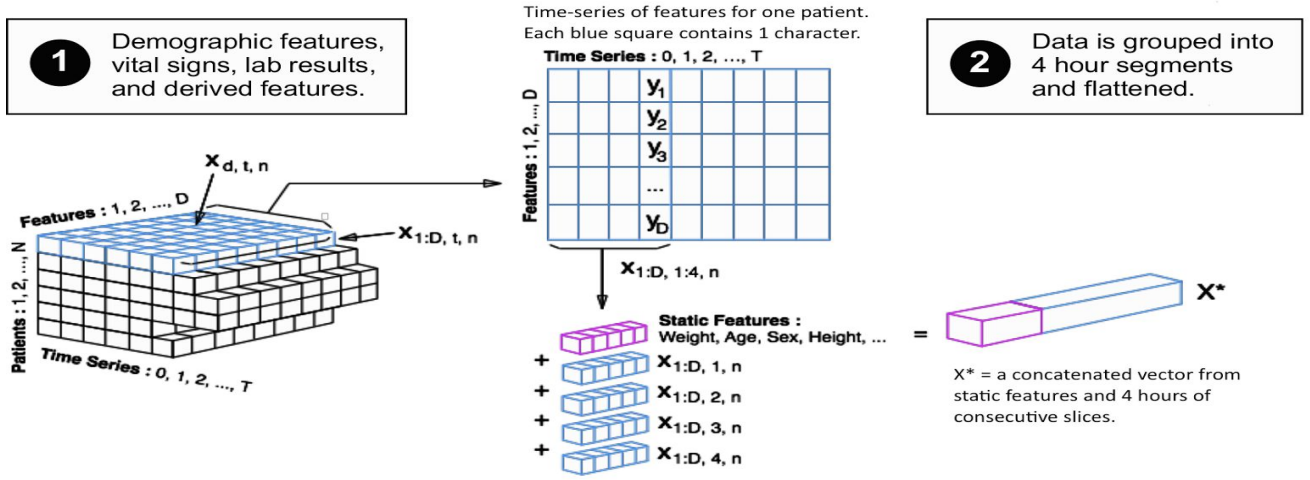


- Every $\mathbf{x_{n,t,d}}$ is one of ten possible **characters**, -4:0:4 or *NaN*.

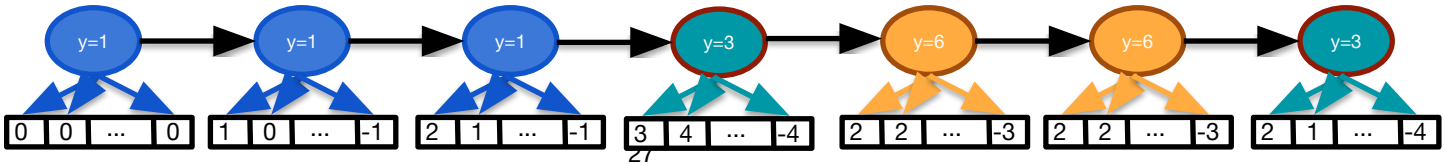
- Every $\mathbf{x_{n,t}}$ is one of 10^D possible **words**.

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

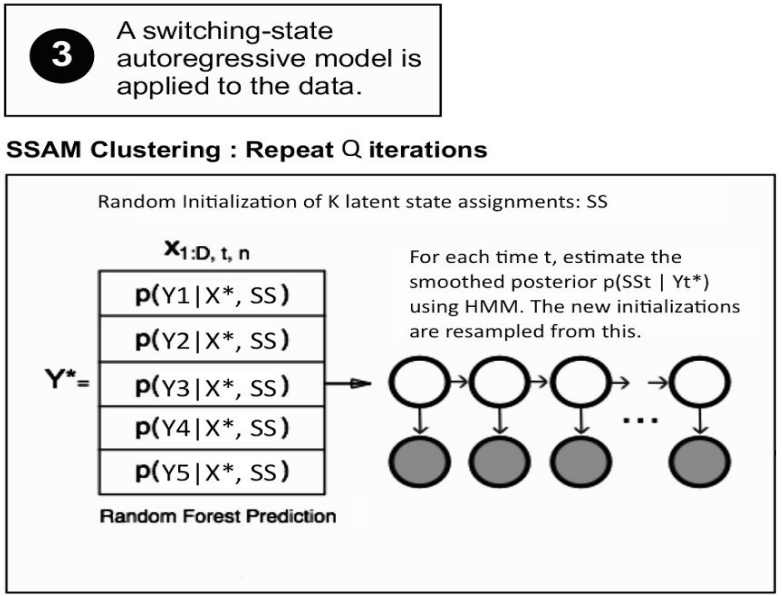
$$\begin{aligned} y_t^n &\sim T_y(\cdot|y_{t-1}^n) \\ \boldsymbol{x}_t^n(p) &\sim T_x(\boldsymbol{x}_t^n(p)|\boldsymbol{x}_{t-1}^n, \theta_{p,y_{t-1}^n}) \end{aligned}$$

- 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 $x^n_t(d)|\, x^n_{t-4:t-1}$.

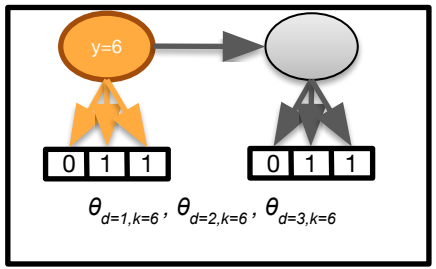
- Update state sequences y^n_t given $\{\theta_{d,k}\}$.



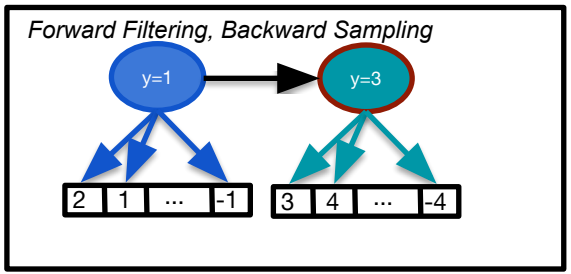
Discrete state space and per-variable missingness

- 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 $\theta_{d,k}$

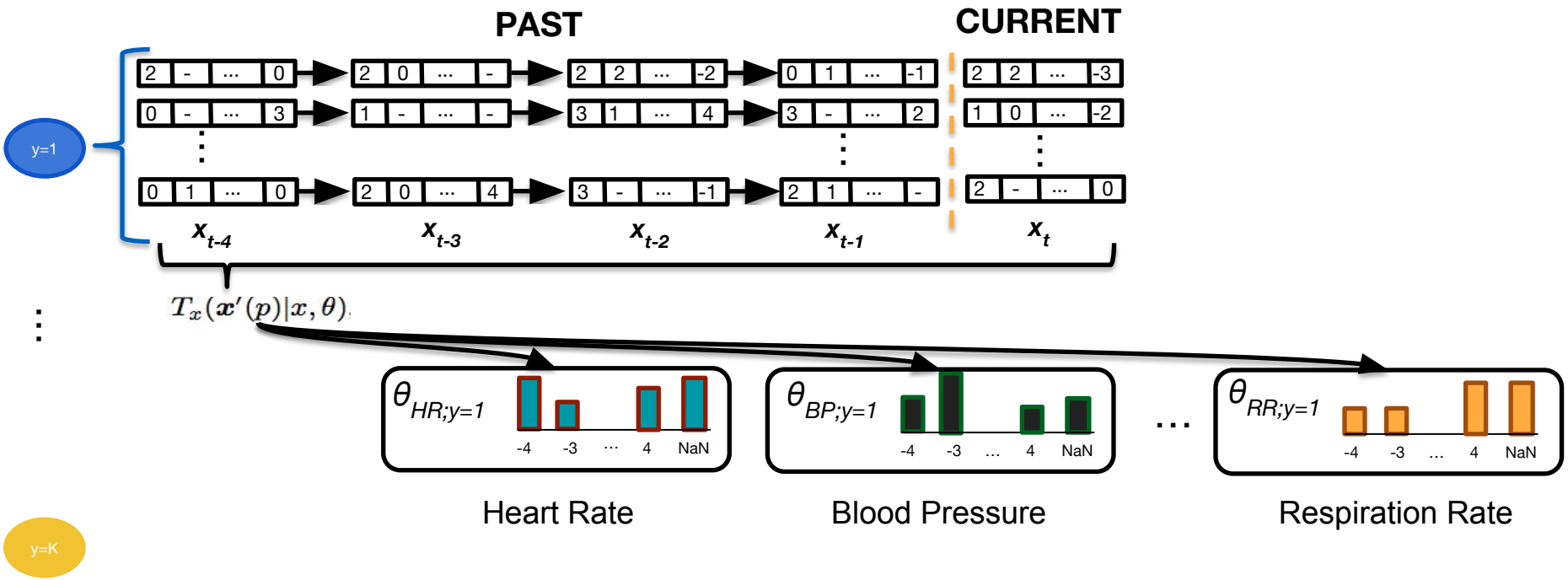


2. Sample states y^n_t



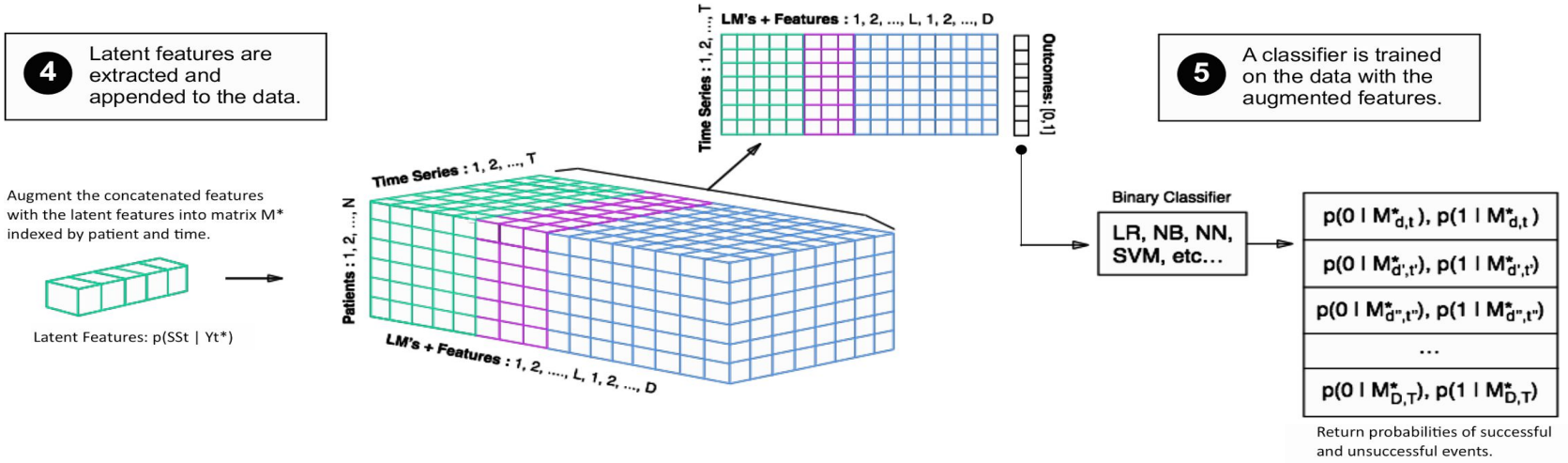
Distribution of values per-variable and latent

- Train parameters $\theta_{d;k}$ to predict $x^n_t(d)$ given $x^n_{t-4:t-1}$



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. 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

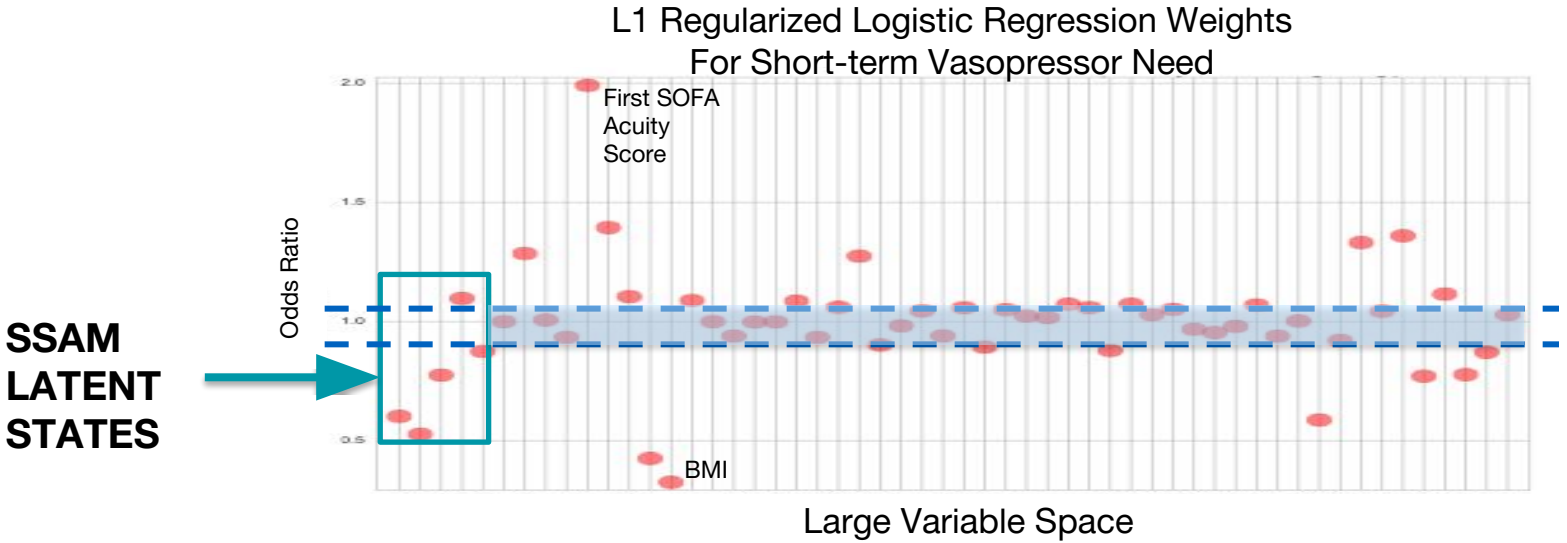
	AUC
Baseline 1 – Prior Work	0.79
Baseline 2 – Raw Data	0.83
SSAM Representations	0.83
Raw Data + SSAM Rep.	**0.88**

- **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.

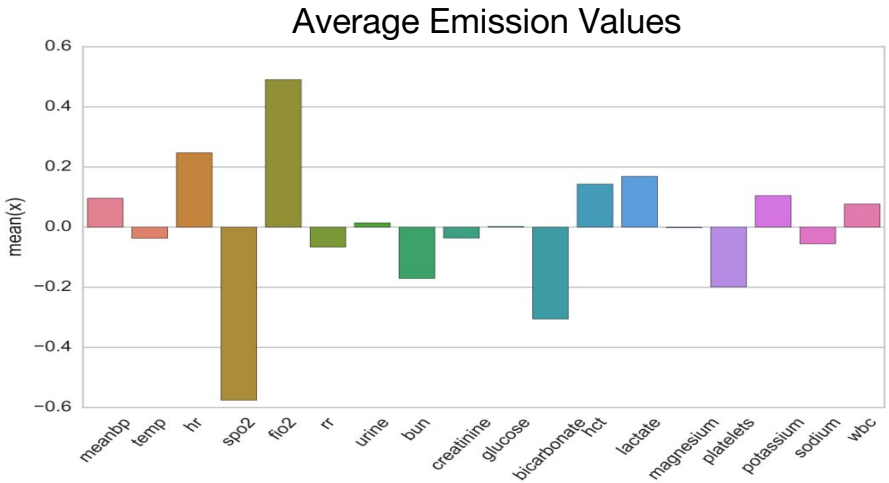
Regularized prediction emphasizes latent states

- **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.

36

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.**

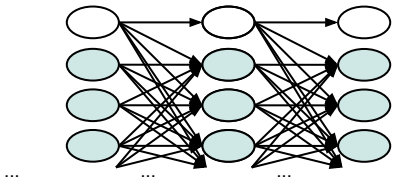
Outline

1. What can we do with supervised learning?

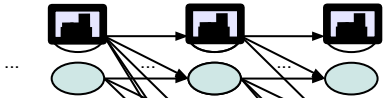
2. **Case study on intervention predictions:**
   a. Frame the problem
   b. Evaluation
   c. Iterate

3. What else should we be thinking about?

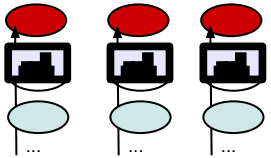
More outcomes and improved dynamics



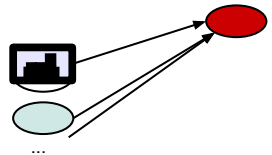
Learn model parameters over patients with variational EM.



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



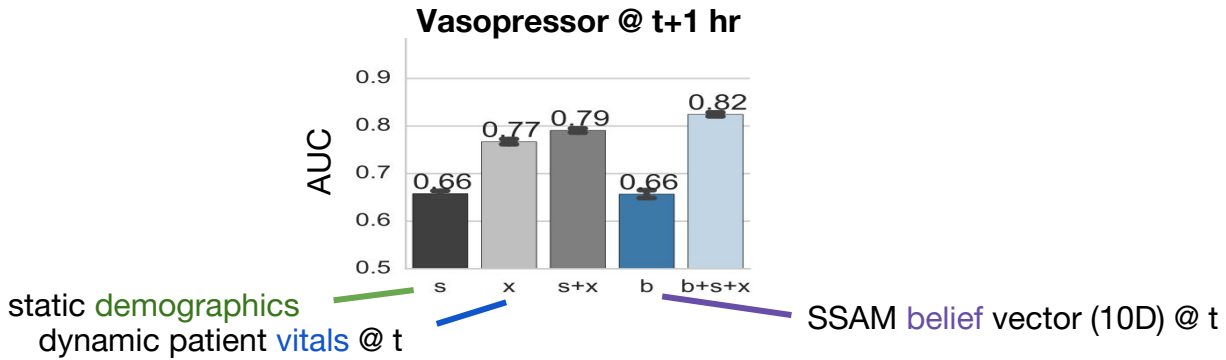
Logistic regression (with label-balanced cost function)

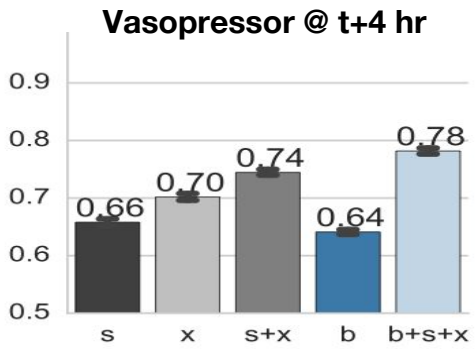


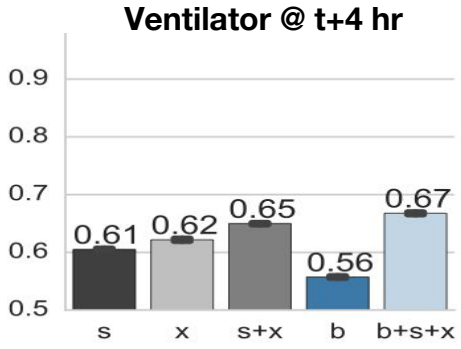
Predict onset in advance

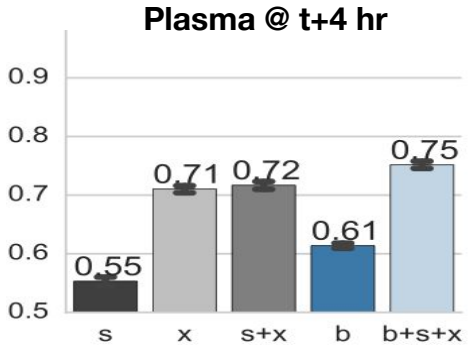
- More Interventions: fresh-frozen-plasma transfusion (ffp), platelet transfusion, red-blood-cell (rbc) transfusion, vasopressor administration, and ventilator intubation.
- Gaussian Emission Model for Dynamics:
  - Static observations s (10 dimensions using one-hot encoding),
  - Dynamic time-series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

State space beliefs improve prediction









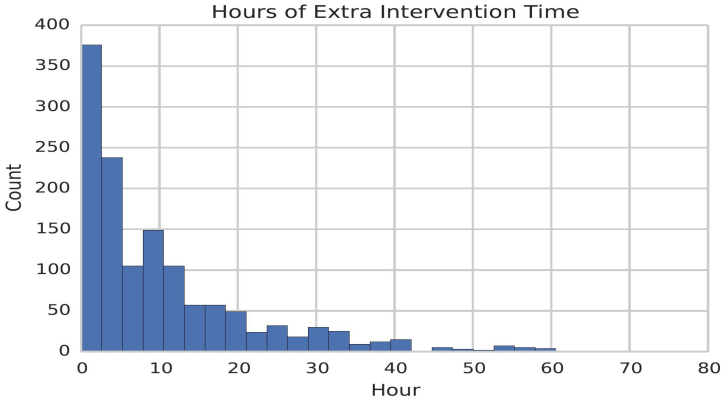
Similar trends in other tasks, except!

	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?

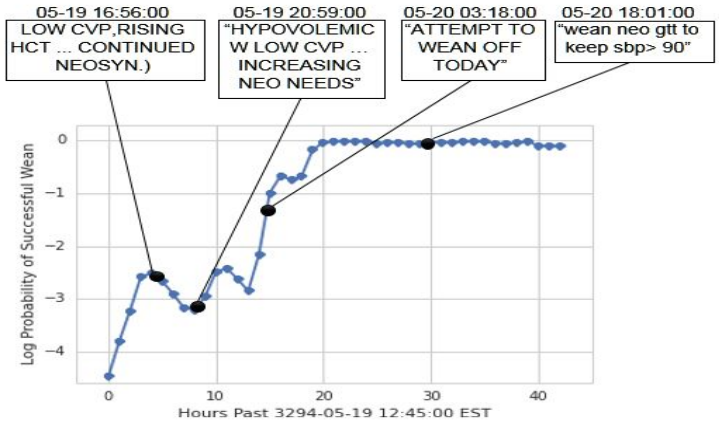
- Patients can be left on interventions longer than necessary.



- 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!

Outline

1. What can we do with supervised learning?

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

3. What else should we be thinking about?

Thinking carefully about what we learn

Opportunities in Machine Learning for Healthcare

Marzyeh Ghassemi
Massachusetts Institute of Technology, Verily
Cambridge, MA 02139
mghassem@mit.edu, marzyeh@google.com

Tristan Naumann
Massachusetts Institute of Technology
Cambridge, MA 02139
tjn@mit.edu

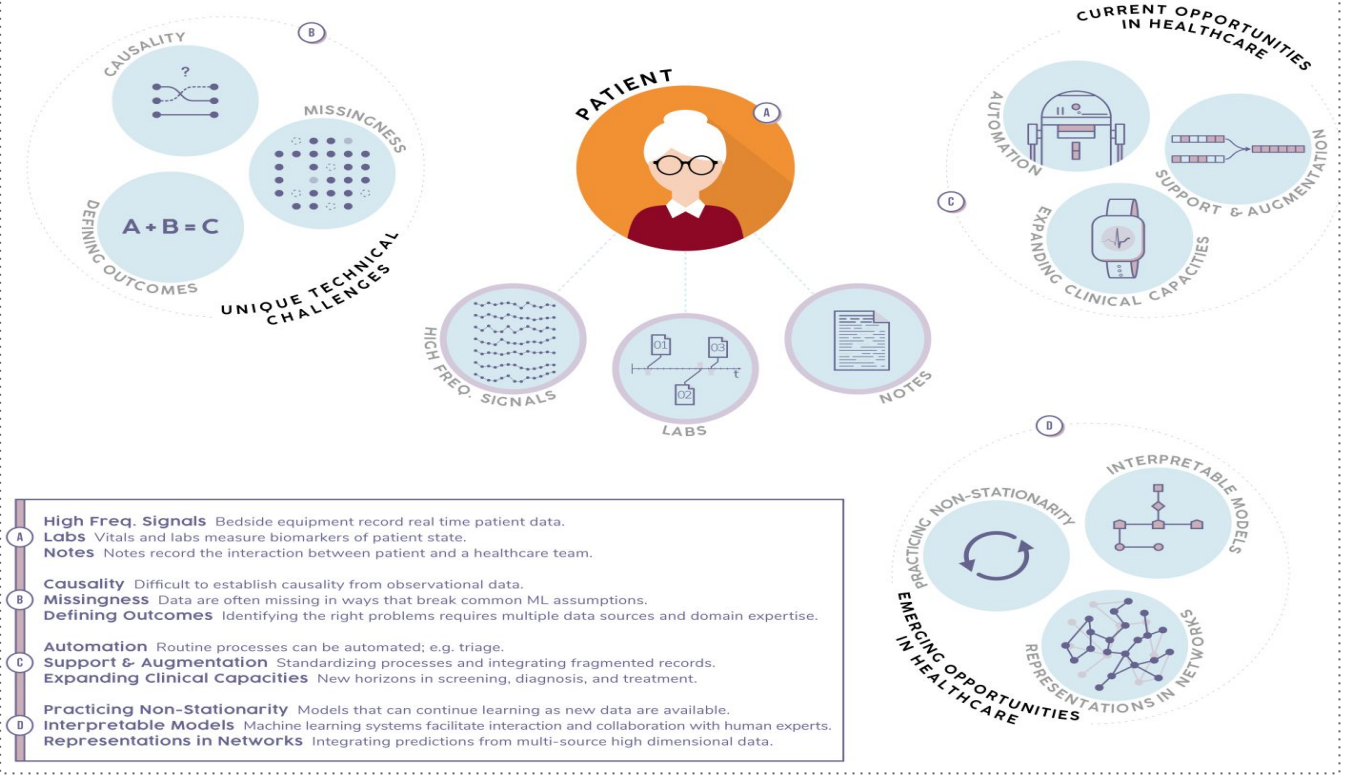
Peter Schulam
Johns Hopkins University
Baltimore, MD 21218
pschulam@cs.jhu.edu

Andrew L. Beam
Harvard Medical School
Boston, MA 02115
andrew_beam@hms.harvard.edu

Rajesh Ranganath
New York University
New York, NY 10011
rajeshr@cims.nyu.edu

Abstract

Healthcare is a natural arena for the application of machine learning, especially as modern electronic health records (EHRs) provide increasingly large amounts of data to answer clinically meaningful questions. However, clinical data and practice present unique challenges that complicate the use of common methodologies. This article serves as a primer on addressing these challenges and highlights opportunities for members of the machine learning and data science communities to contribute to this growing domain.



A — **High Freq. Signals** Bedside equipment record real time patient data.
Labs Vitals and labs measure biomarkers of patient state.
Notes Notes record the interaction between patient and a healthcare team.

B — **Causality** Difficult to establish causality from observational data.
Missingness Data are often missing in ways that break common ML assumptions.
Defining Outcomes Identifying the right problems requires multiple data sources and domain expertise.

C — **Automation** Routine processes can be automated; e.g. triage.
Support & Augmentation Standardizing processes and integrating fragmented records.
Expanding Clinical Capacities New horizons in screening, diagnosis, and treatment.

D — **Practicing Non-Stationarity** Models that can continue learning as new data are available.
Interpretable Models Machine learning systems facilitate interaction and collaboration with human experts.
Representations in Networks Integrating predictions from multi-source high dimensional data.

Technical Challenges!

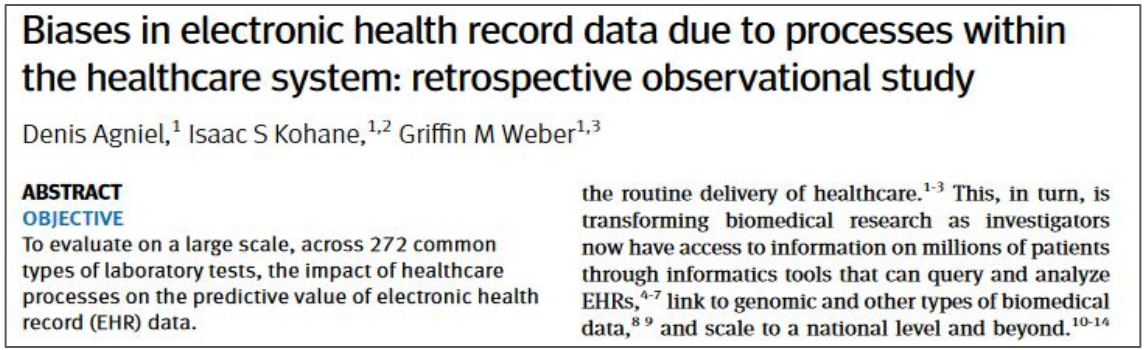
Health Opportunities!

ML Work Needed!

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

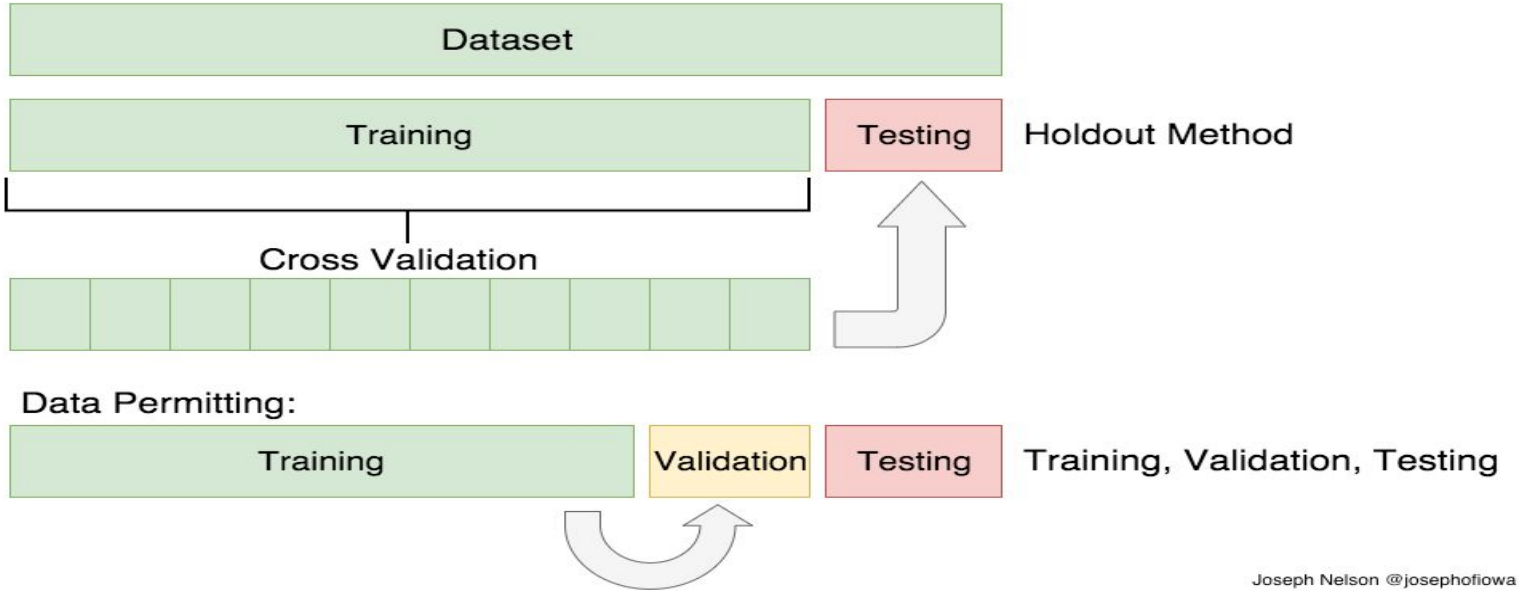
OBJECTIVE

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

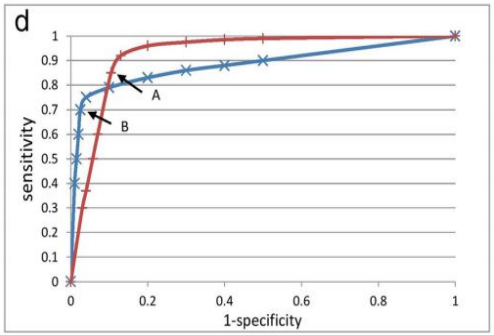
Careful evaluation is extremely important

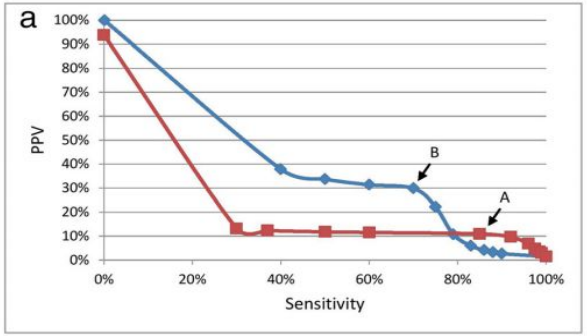
- Spend as much time designing evaluation as with model prototyping.
- Make diagnostic plots, not just tables, and think about actual utility.

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

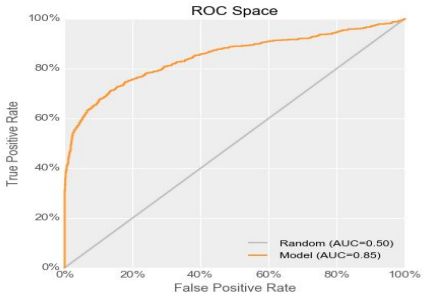
Slide courtesy of Michael Hughes

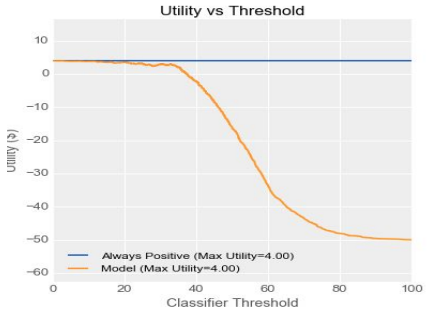
Calibration matters in practice

- 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 $utility = -\$50$ $rate(t) = (1 - FPR(t)) \cdot 5\%$

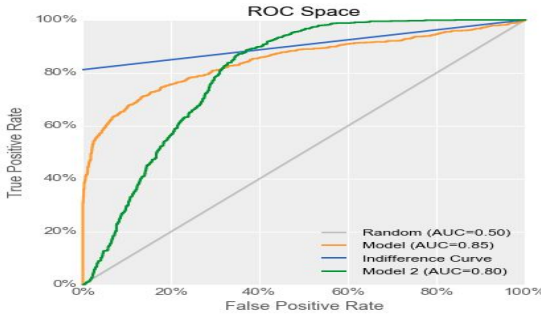
vs.

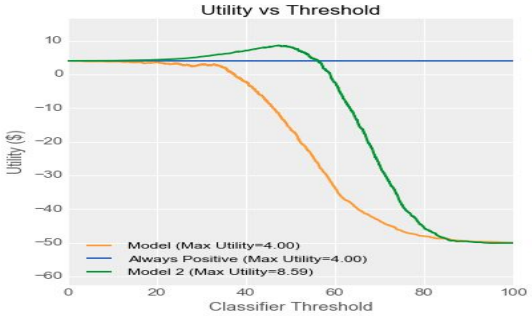




- Domain specific evaluation requires a goal.

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?

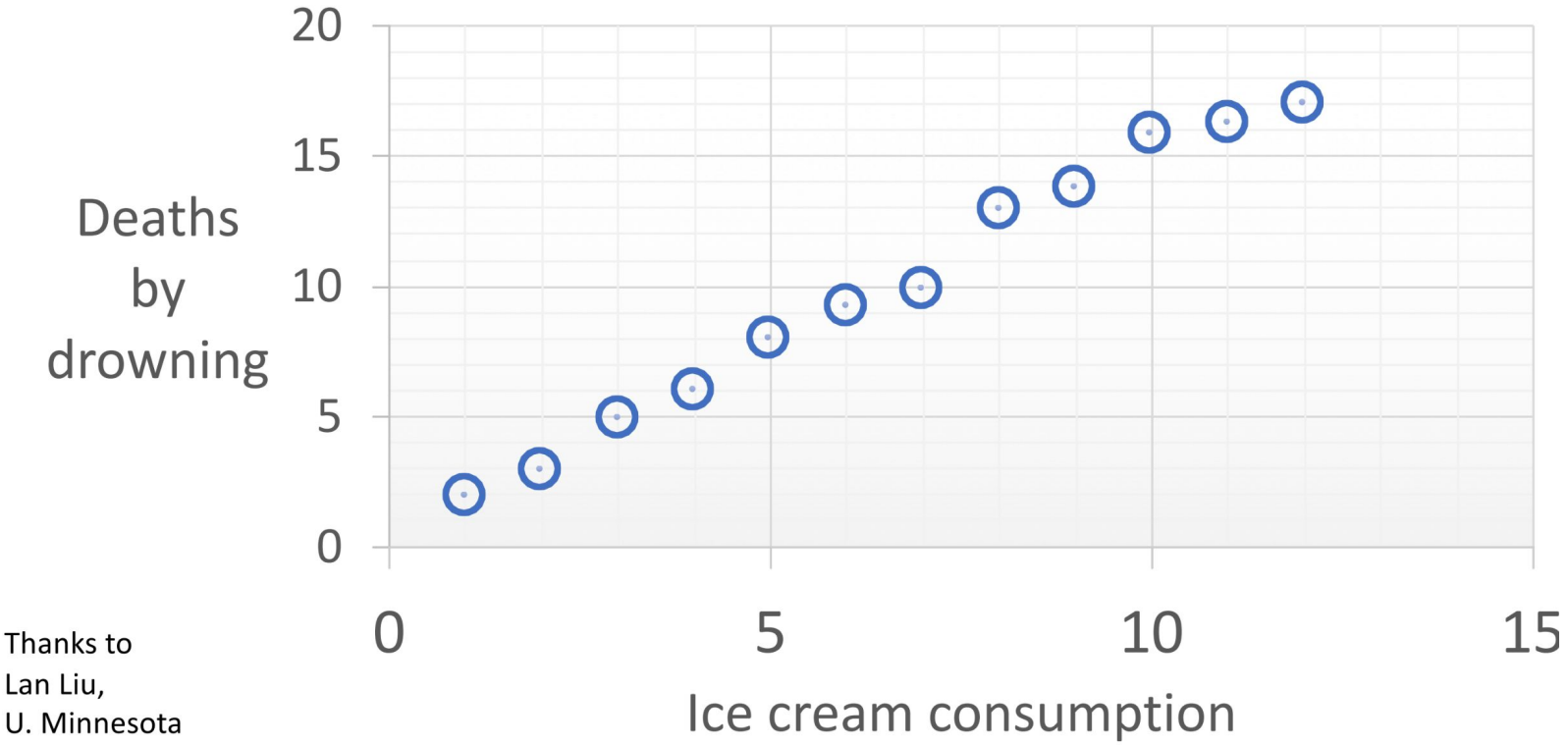
- We build predictive model:
  features X = [lab_tests, diagnoses, medications]
  label y = [diabetic]

- 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