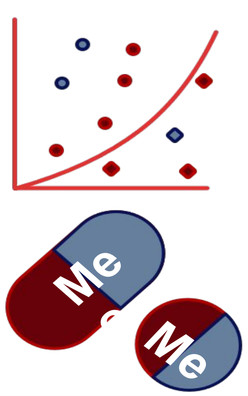
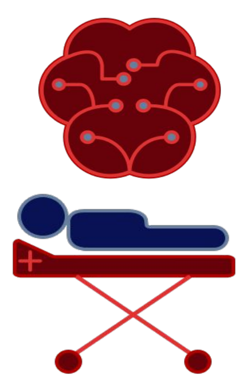
VECTOR INSTITUTE | INSTITUT VECTEUR

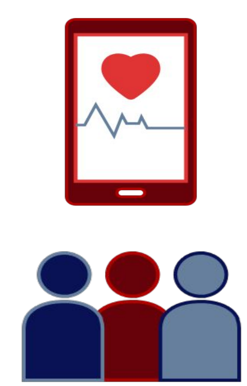
UNIVERSITY OF TORONTO

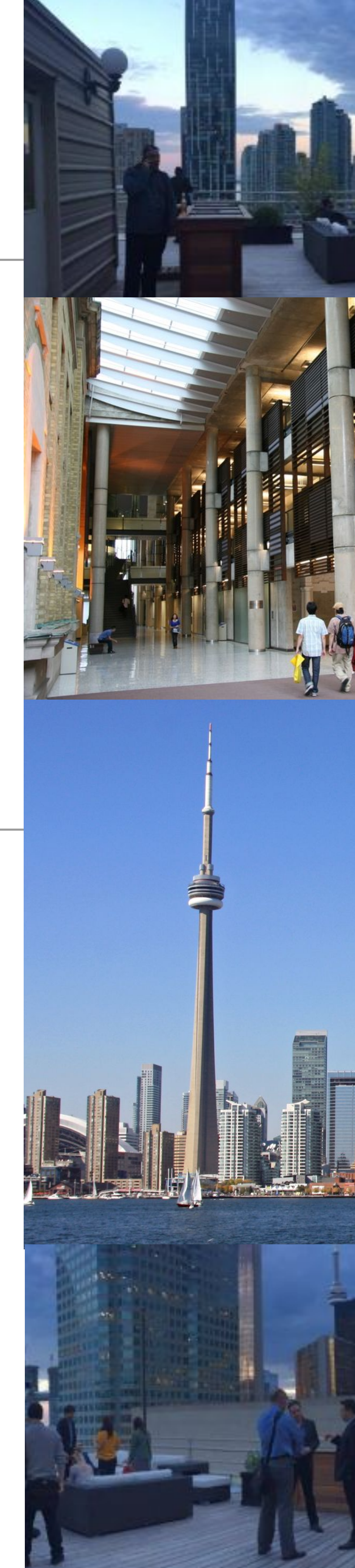
Learning "Healthy" Models for Healthcare

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









Why Try To Work in Health?

- Improvements in health **improve lives**.

- Same **patient** → different **treatments** across hospitals, clinicians.

- Improving care requires **evidence**.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Why Try To Work in Health?

- Improvements in health **improve lives**.
- Same **patient** → different **treatments** across hospitals, clinicians.
- Improving care requires **evidence**.

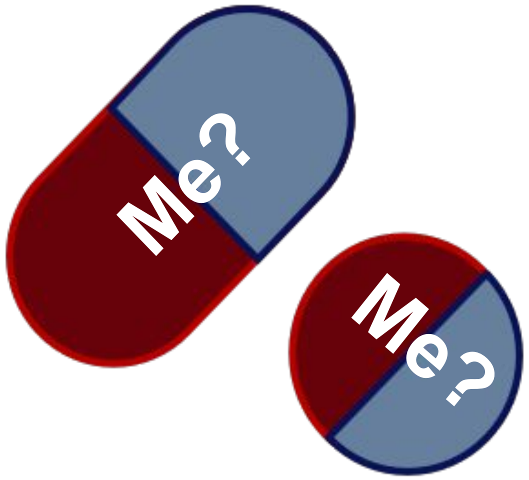
What does it mean **mean** to be **healthy**?

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Learning What Is Healthy?

Recruit a study population.

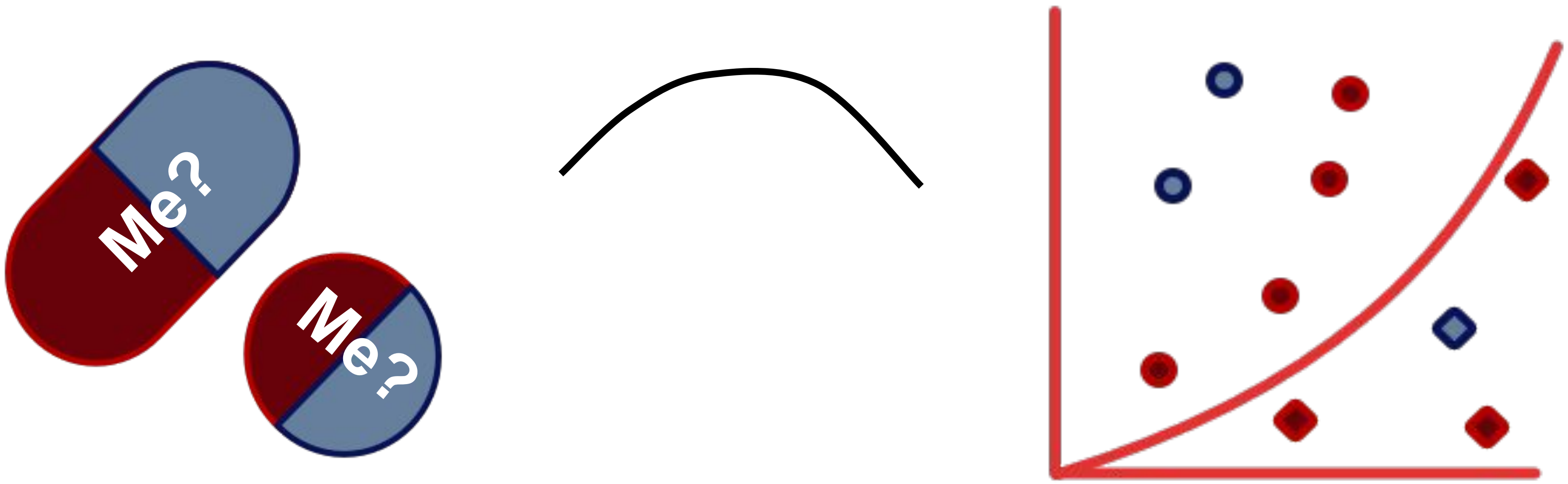


UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Learning What Is Healthy?

Learn a rule.

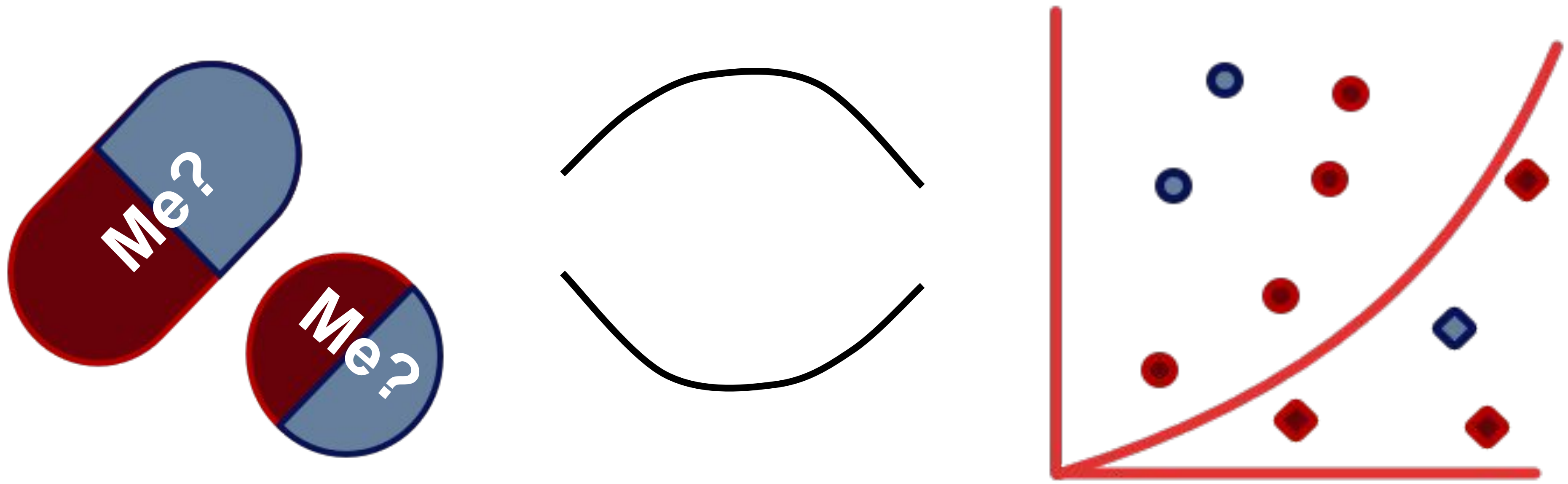


UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

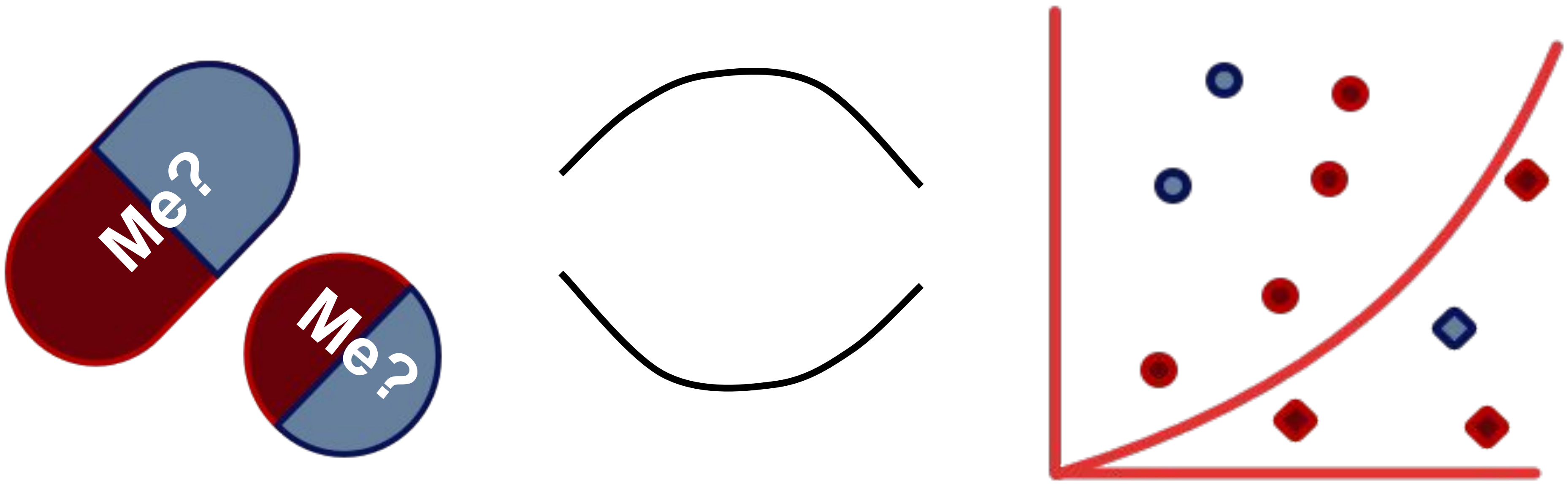
Learning What Is Healthy?

Does it generalize?



Learning What Is Healthy?

For whom does it generalize?



UNIVERSITY OF TORONTO

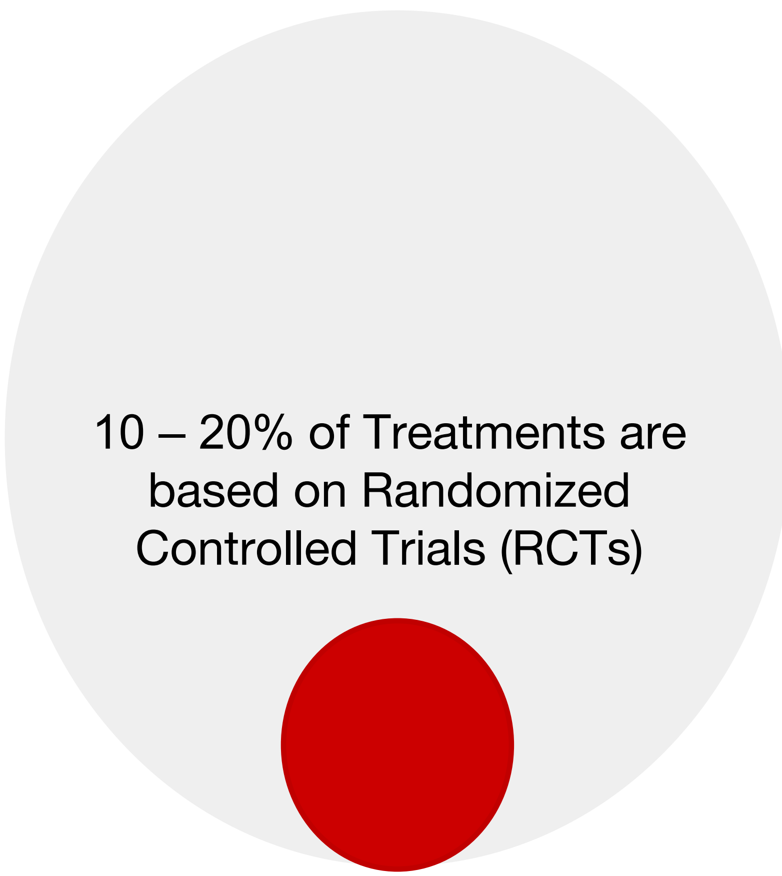
VECTOR INSTITUTE | INSTITUT VECTEUR

Evidence in Healthcare and Health?

Randomized Controlled Trials (RCTs) are

Evidence in Healthcare and Health?

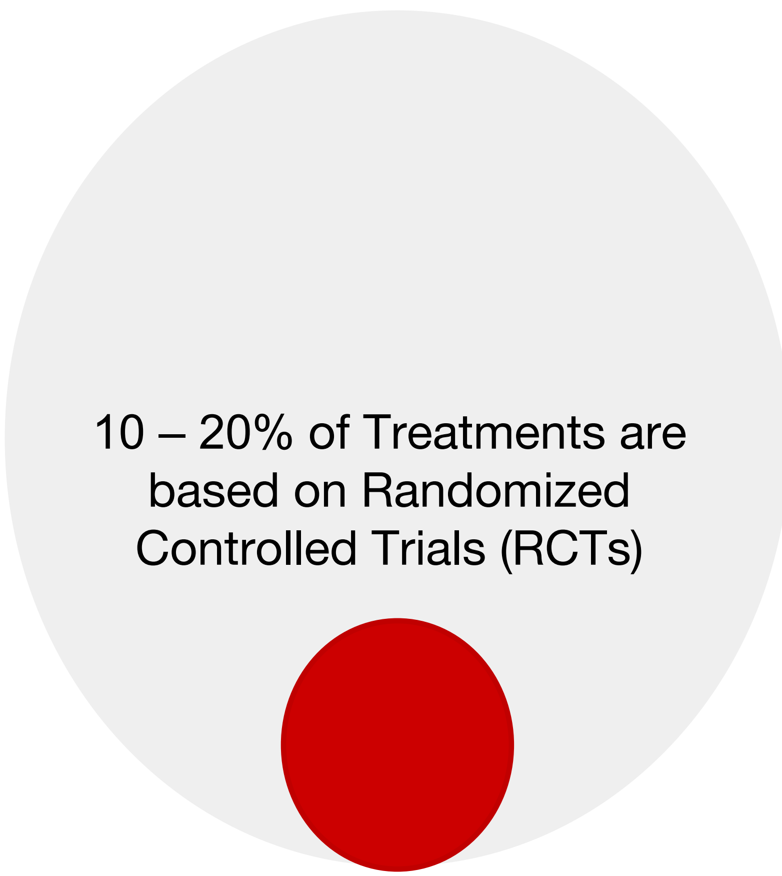
Randomized Controlled Trials (RCTs) are **rare and expensive**

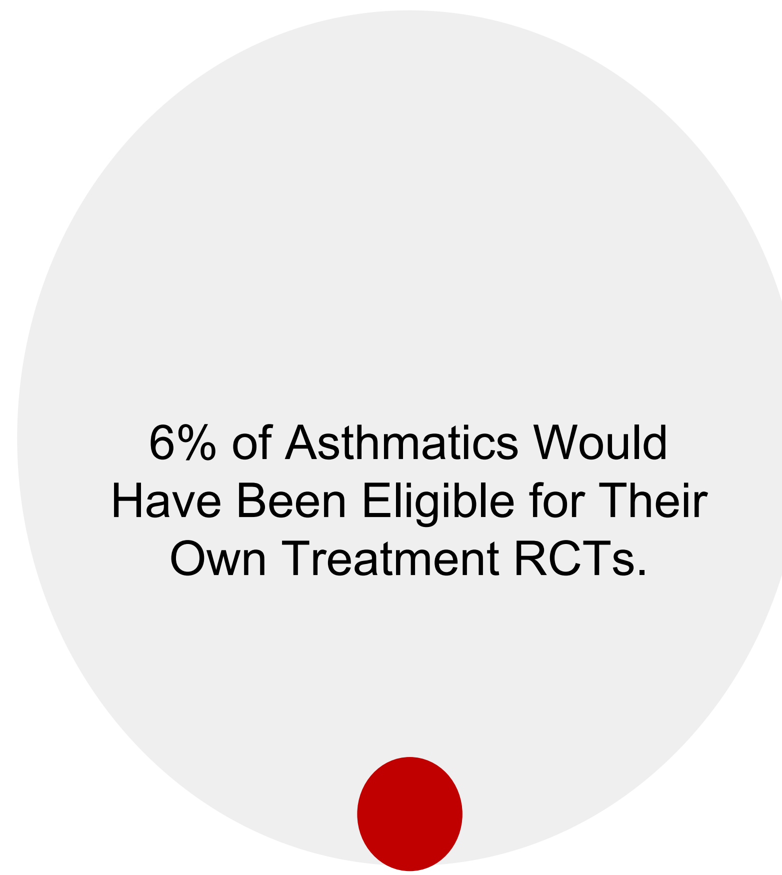


[1] *Smith M, Saunders R, Stuckhardt L, McGinnis JM, Committee on the Learning Health Care System in America, Institute of Medicine. Best Care At Lower Cost: The Path To Continuously Learning Health Care In America. Washington: National Academies Press; 2013..*

Evidence in Healthcare and Health?

Randomized Controlled Trials (RCTs) are **rare and expensive**, and can encode **structural biases** that apply to very few people.

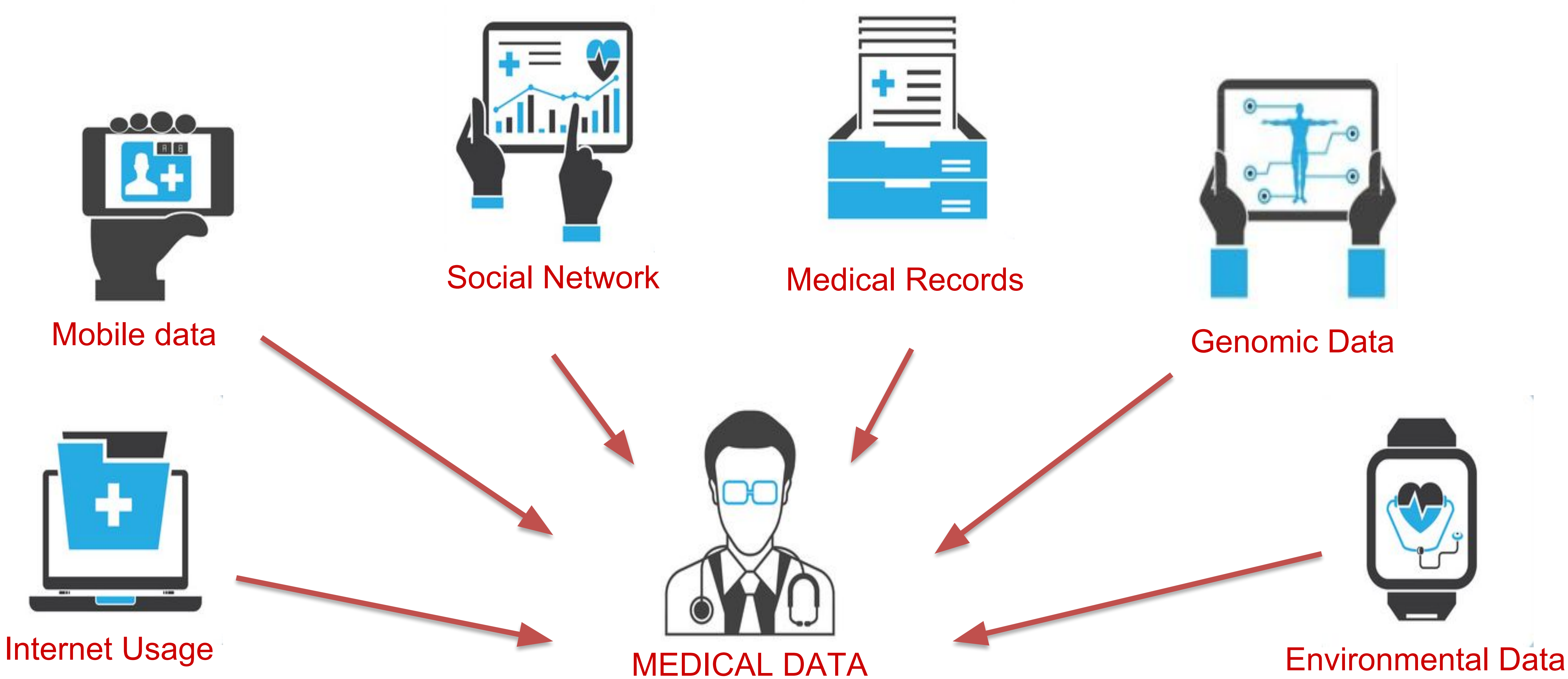




[1] *Smith M, Saunders R, Stuckhardt L, McGinnis JM, Committee on the Learning Health Care System in America, Institute of Medicine. Best Care At Lower Cost: The Path To Continuously Learning Health Care In America. Washington: National Academies Press; 2013.*
[2] Travers, Justin, et al. "External validity of randomised controlled trials in asthma: to whom do the results of the trials apply?." Thorax 62.3 (2007): 219-223.

Machine Learning What Is Healthy?

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



UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Extracting Knowledge is Hard in Health

- Data are **not gathered** to answer your hypothesis.
- **Primary** case is to provide **care**.
- Secondary data are **hard** to work with.

Heterogenous
Sampling
Data Type
Time Scale

Sparse
Unmeasured
Unreported
No Follow-up

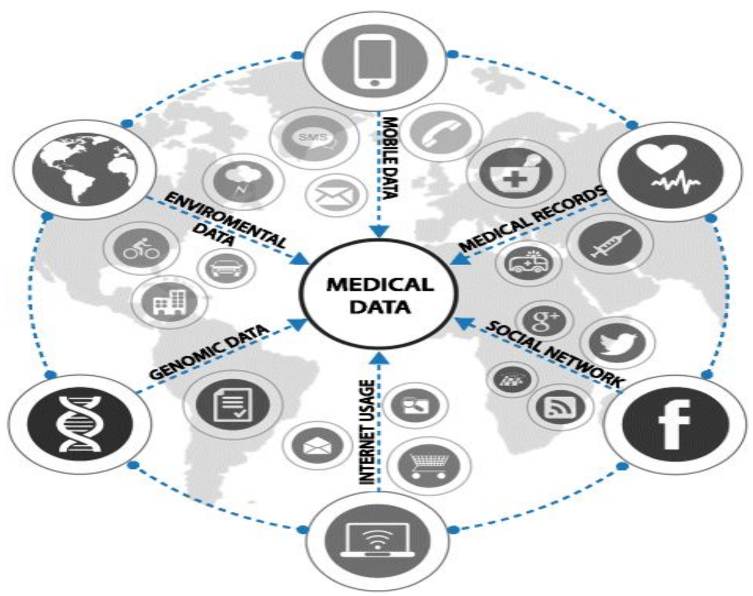
Uncertainty
Labels
Bias
Context

UNIVERSITY OF TORONTO

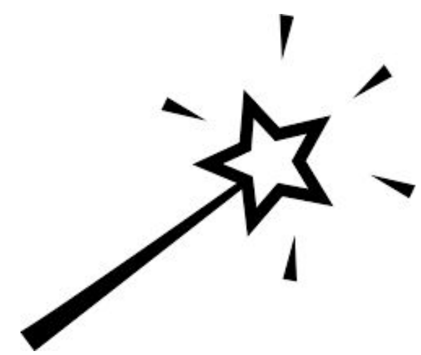
VECTOR INSTITUTE | INSTITUT VECTEUR

Lack of Expertise Is Challenging

- Media can create unrealistic expectations.



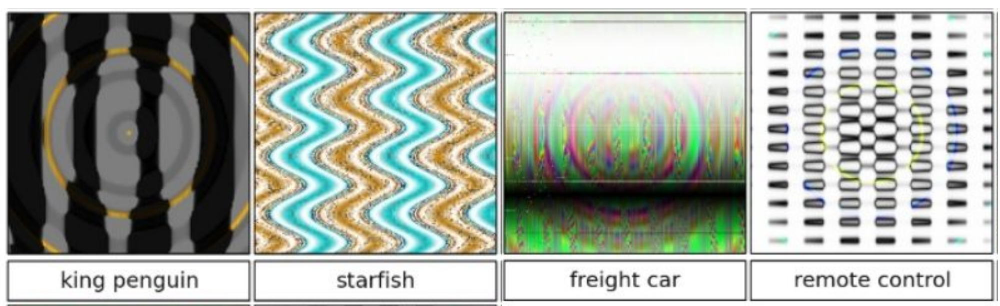
+



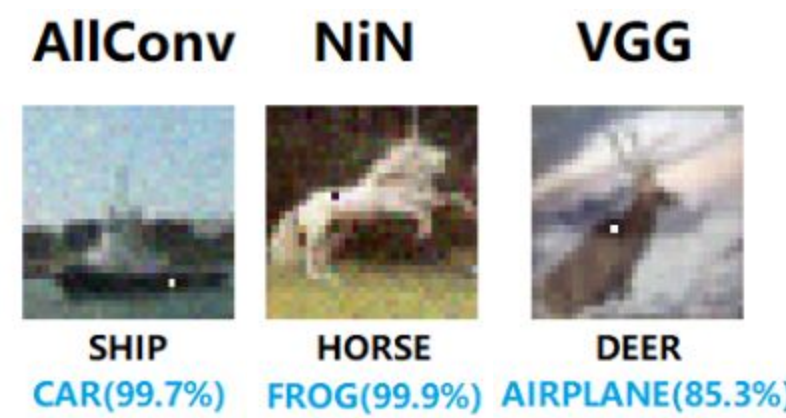
≠ Insight

Be Careful What You Optimize For

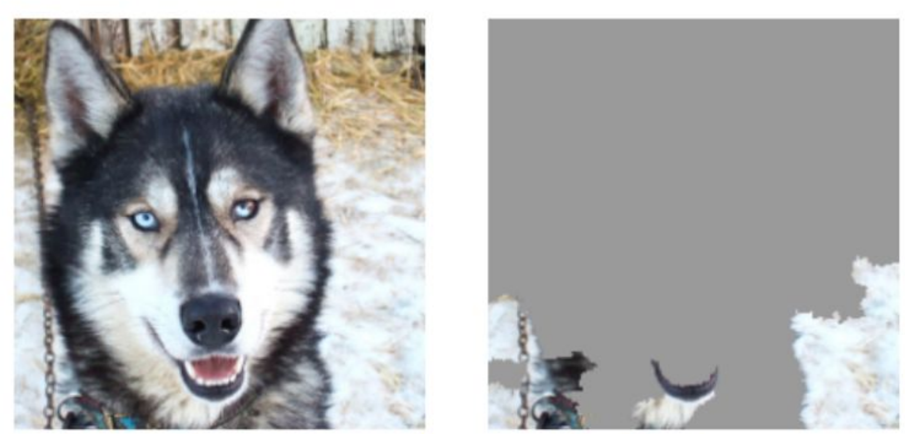
- ML can be confidently wrong.[1, 2]



or



- Humans are “natural” experts in NLP, ASR, Vision evaluation.[3]

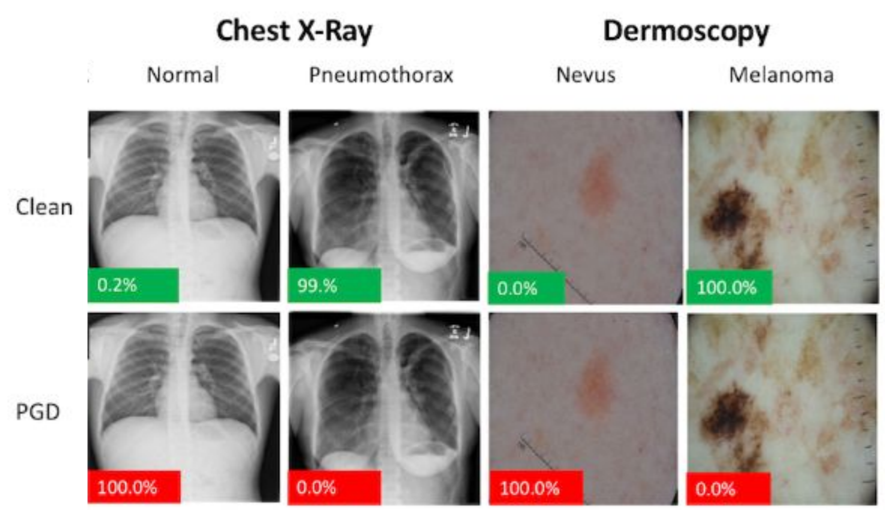


(a) Husky classified as wolf (b) Explanation

[1] Nguyen, Anh, Jason Yosinski, and Jeff Clune. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
[2] Su, Jiawei, Danilo Vasconcellos Vargas, and Sakurai Kouichi. "One pixel attack for fooling deep neural networks." *arXiv preprint arXiv:1710.08864* (2017).
[3] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

Healthy Models Require Domain Knowledge

- Hyper-expertise makes attacks in clinical data harder to spot.[1]



- Learning without understanding is dangerous.[2]

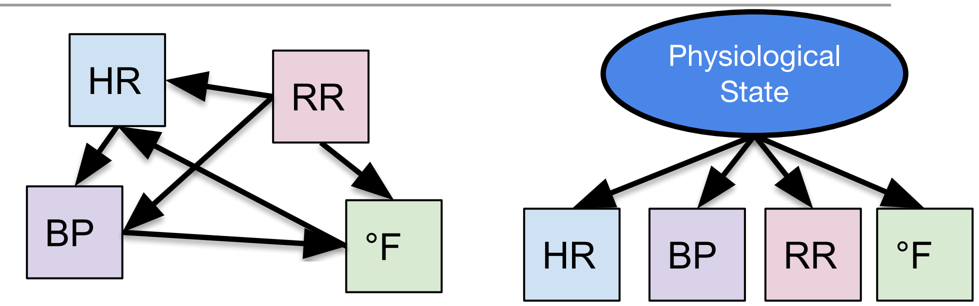
"...**aggressive care** received by asthmatic pneumonia patients (in the training set) was so effective that it **lowered their risk** of dying from pneumonia compared to the general population…"

→ "HasAsthma(x) ⇒ LowerRisk(x)"

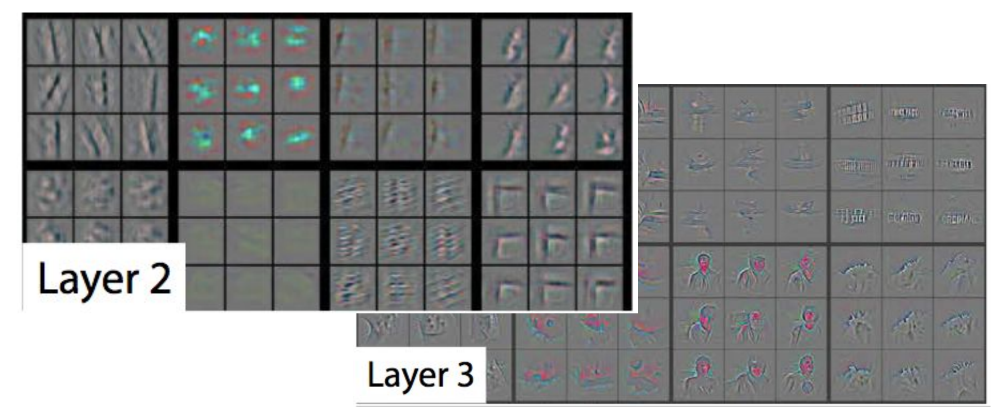
[1] Finlayson, Samuel G., Isaac S. Kohane, and Andrew L. Beam. "Adversarial Attacks Against Medical Deep Learning Systems." *arXiv preprint arXiv:1804.05296* (2018).
[2] Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.

Good Representations in ML for Health

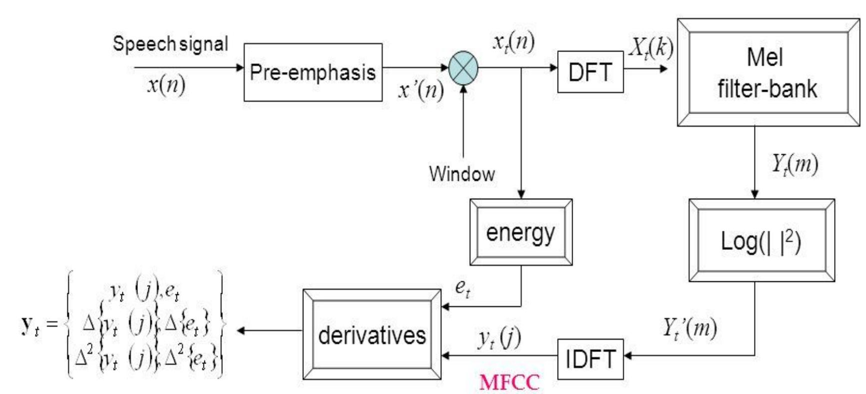
- Representations are **useful** abstractions of data *X* that **disentangle** underlying **factors**.



- Enables **semi-supervised learning**; factors explaining *P(X)* are useful for learning *P(Y|X)*.

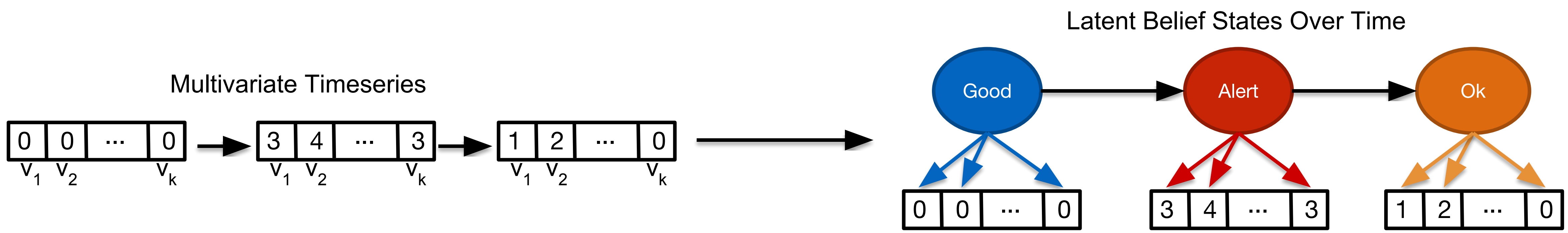


- Allows **shared factors** across **many** learning **tasks**.



Choosing the Right Representation For Each Problem

- Time Series → **Latent States**



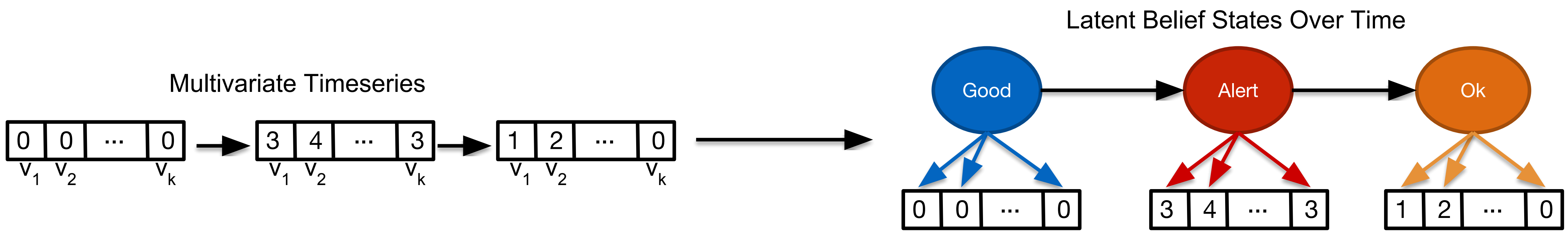
UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

16

Choosing the Right Representation For Each Problem

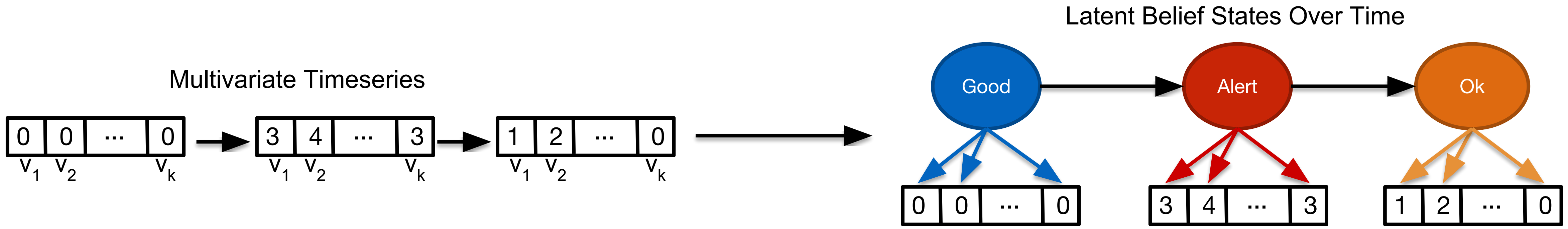
- Time Series → **Latent States**



- Text Data → **Topic Vectors**

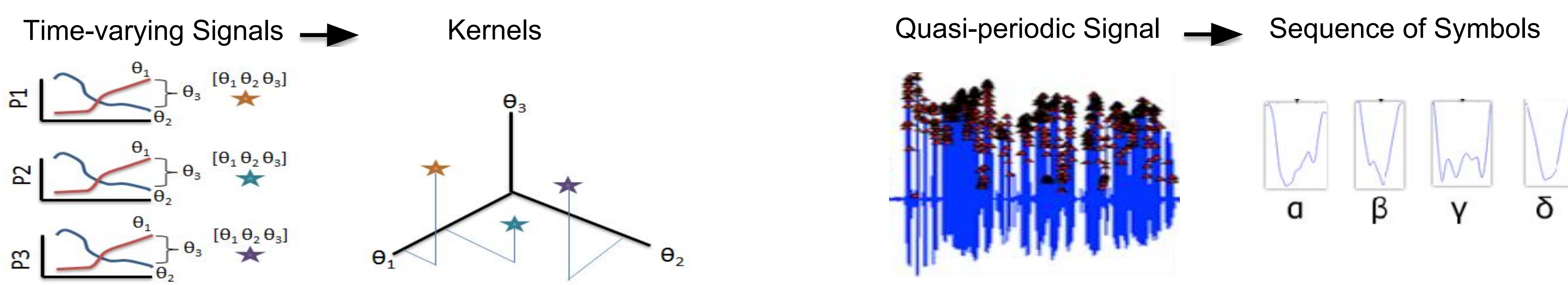
Choosing the Right Representation For Each Problem

- Time Series → **Latent States**



- Text Data → **Topic Vectors**

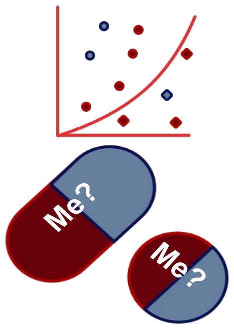
- Instrumentation Signals → **Symbols/Kernels**



UNIVERSITY OF TORONTO

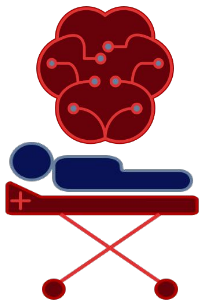
VECTOR INSTITUTE | INSTITUT VECTEUR

Machine Learning For Health (ML4H)



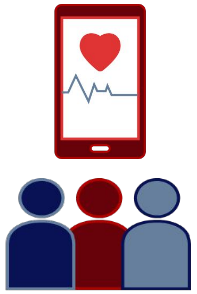
1. What Models are Healthy? Learning Good Representations.

Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014);   A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015);
Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016);
Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017);
Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68);
Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);



2. What Healthcare is Healthy? Stratifying Human Risks.

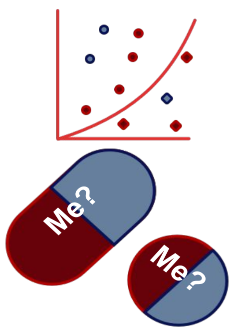
Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017);
Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop);
The Disparate Impacts of Medical and Mental Health with AI. (In submission);



3. What Behaviors are Healthy? Inferring Unseen Actions and States.

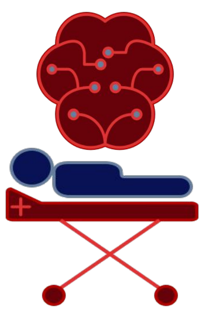
Learning to Detect Vocal Hyperfunction from Ambulatory Necksurface Acceleration Features (IEEE TBME 2014);
Uncovering Voice Misuse Using Symbolic Mismatch (MLHC 2016/JMLR W&C V56);
Project BASELINE Mood Study with Alphabet's Verily;
ClinicalVis Project with Google Brain. (*In submission);

Machine Learning For Health (ML4H)



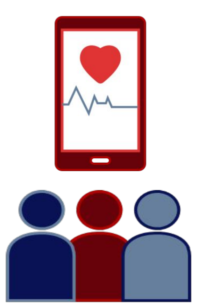
1. What Models are Healthy? Learning Good Representations.

Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014); A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015);
Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016);
Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017);
Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68);
Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);



2. What Healthcare is Healthy? Stratifying Human Risks.

Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017);
Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop);
The Disparate Impacts of Medical and Mental Health with AI. (In submission);

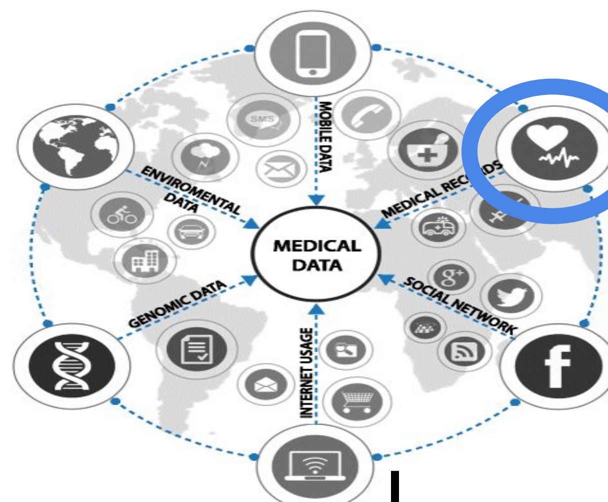


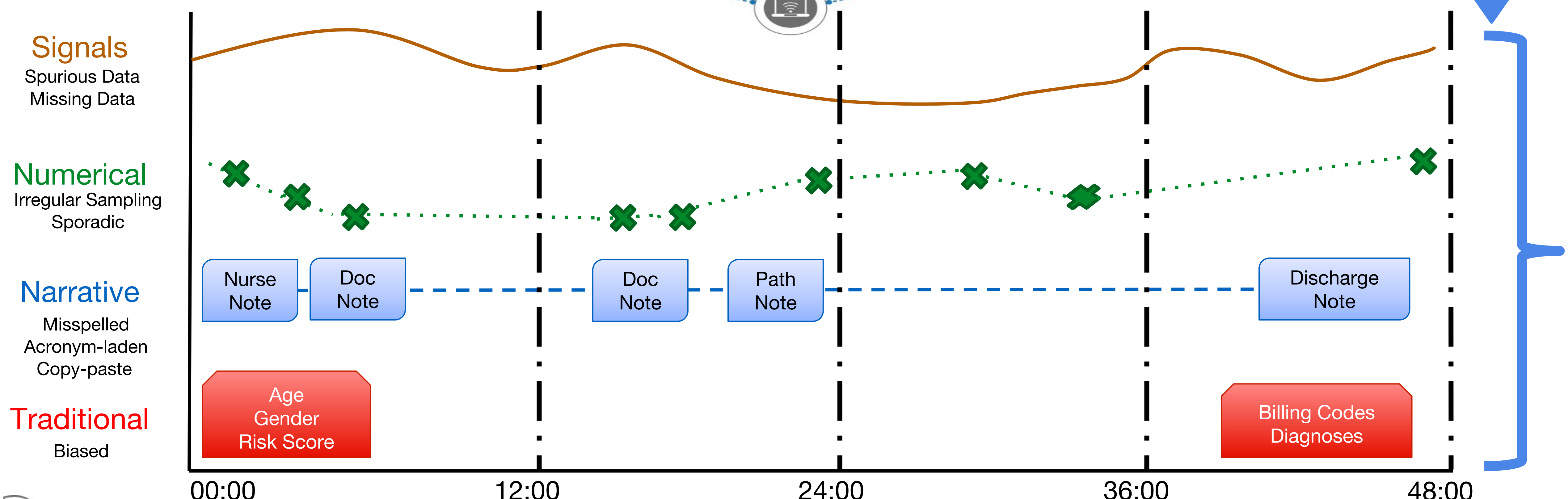
3. What Behaviors are Healthy? Inferring Unseen Actions and States.

Learning to Detect Vocal Hyperfunction from Ambulatory Necksurface Acceleration Features (IEEE TBME 2014);
Uncovering Voice Misuse Using Symbolic Mismatch (MLHC 2016/JMLR W&C V56);
Project BASELINE Mood Study with Alphabet's Verily;
ClinicalVis Project with Google Brain. (*In submission);

MIMIC III ICU Data

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

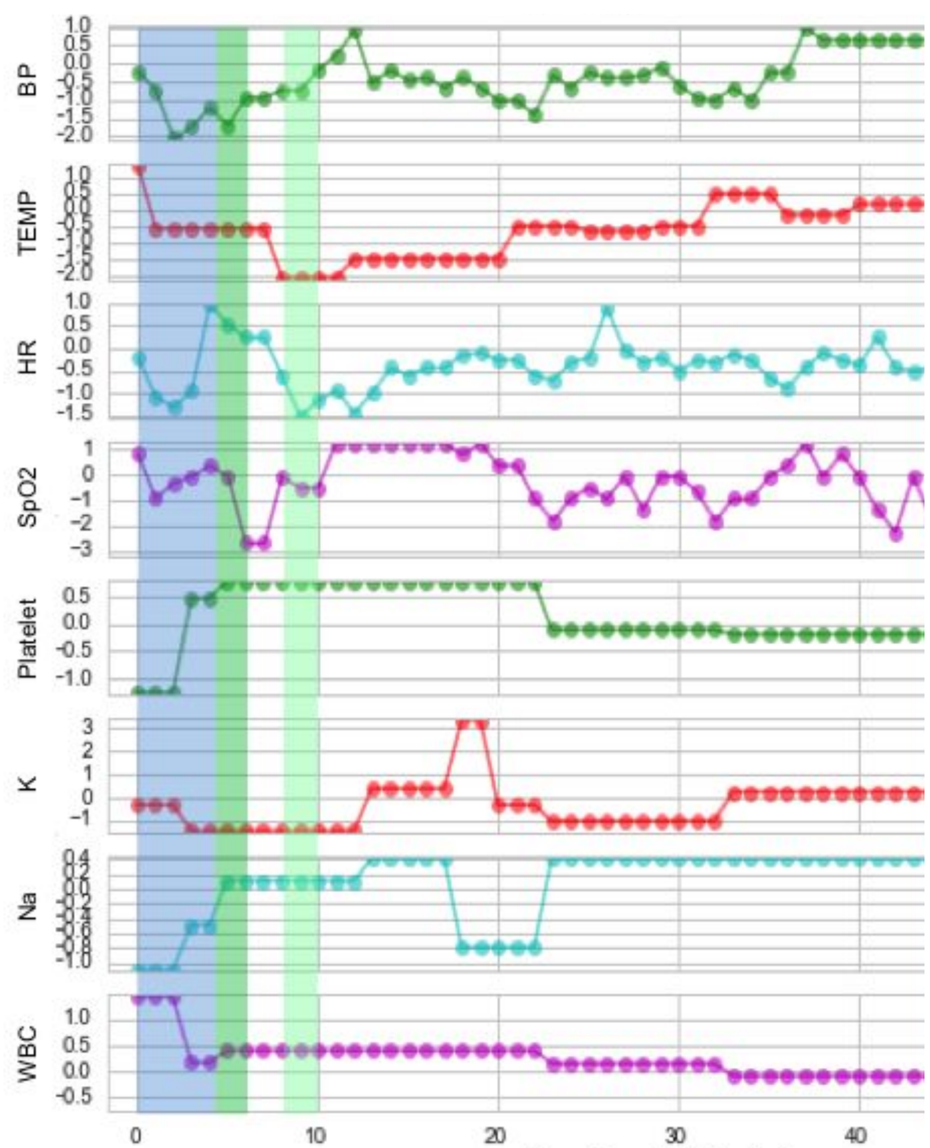




[1] Johnson, Alistair EW, et al. "MIMIC-III, a freely accessible critical care database." Scientific data 3 (2016).

Problem: Hospital Decision-Making / Care Planning

Observe Patient Data



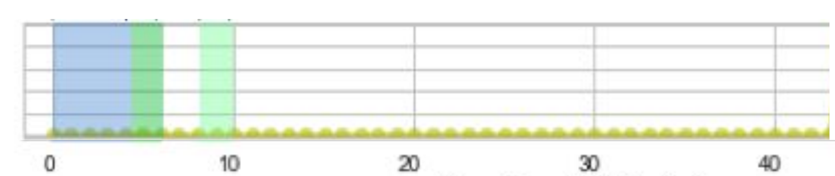
?

“Real-time” **Prediction**

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

By Gap Time

Before the Doctor Acted



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

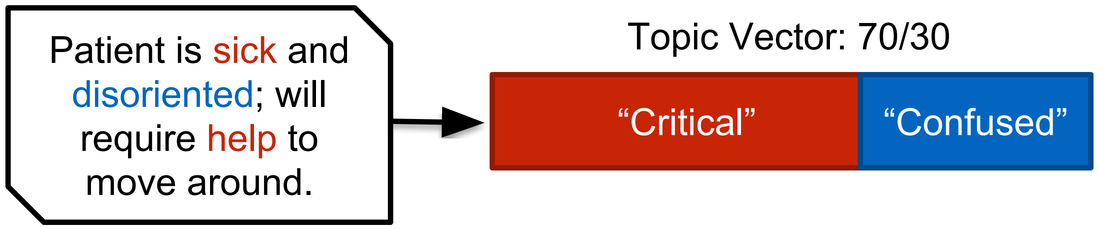
MISSPELLED

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

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.

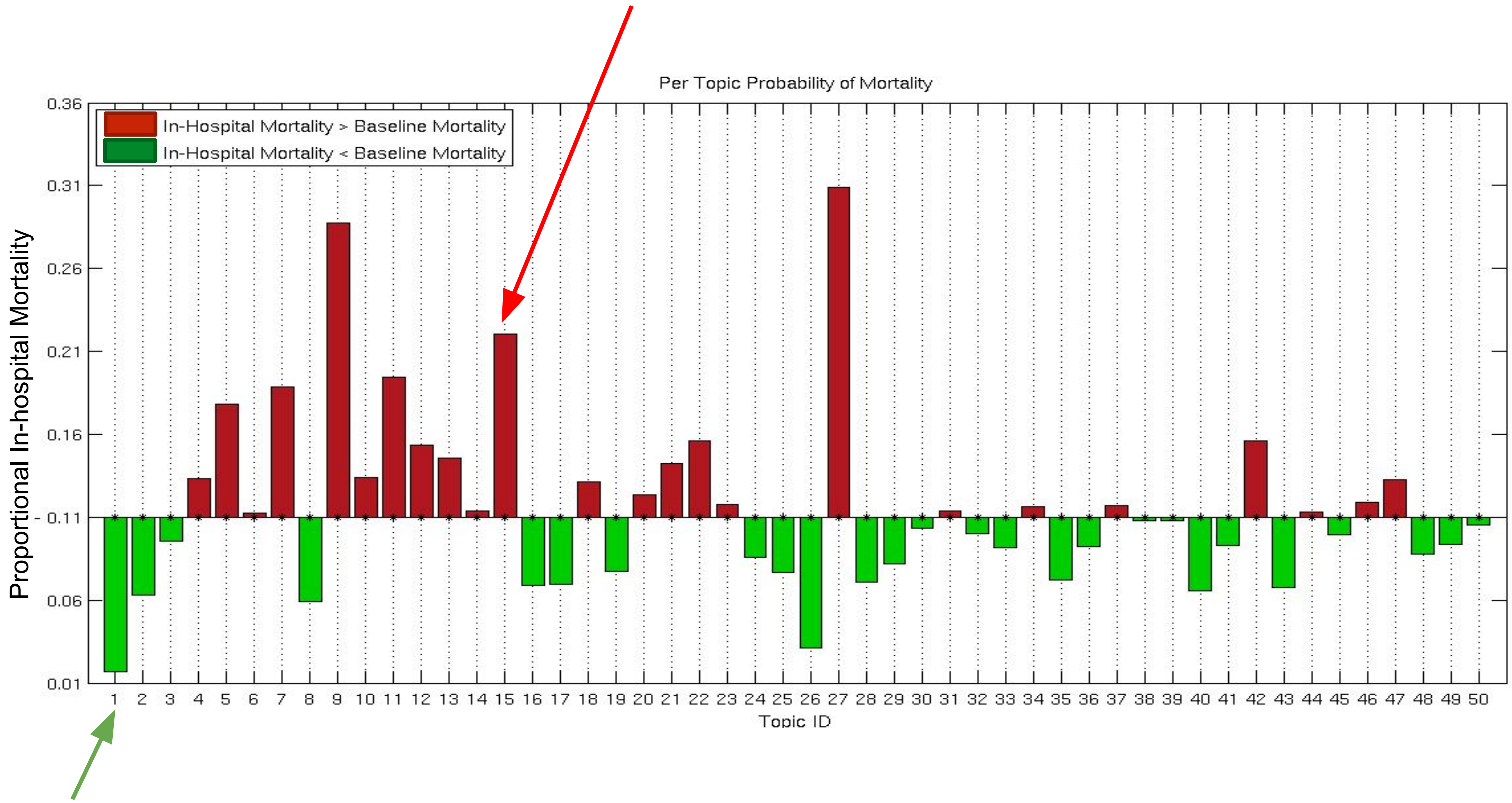


- Use Latent Dirichlet Allocation (LDA)[1] as an **unsupervised** way to **abstract** 473,000 notes from 19,000 patients into “topics”.[2]

[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

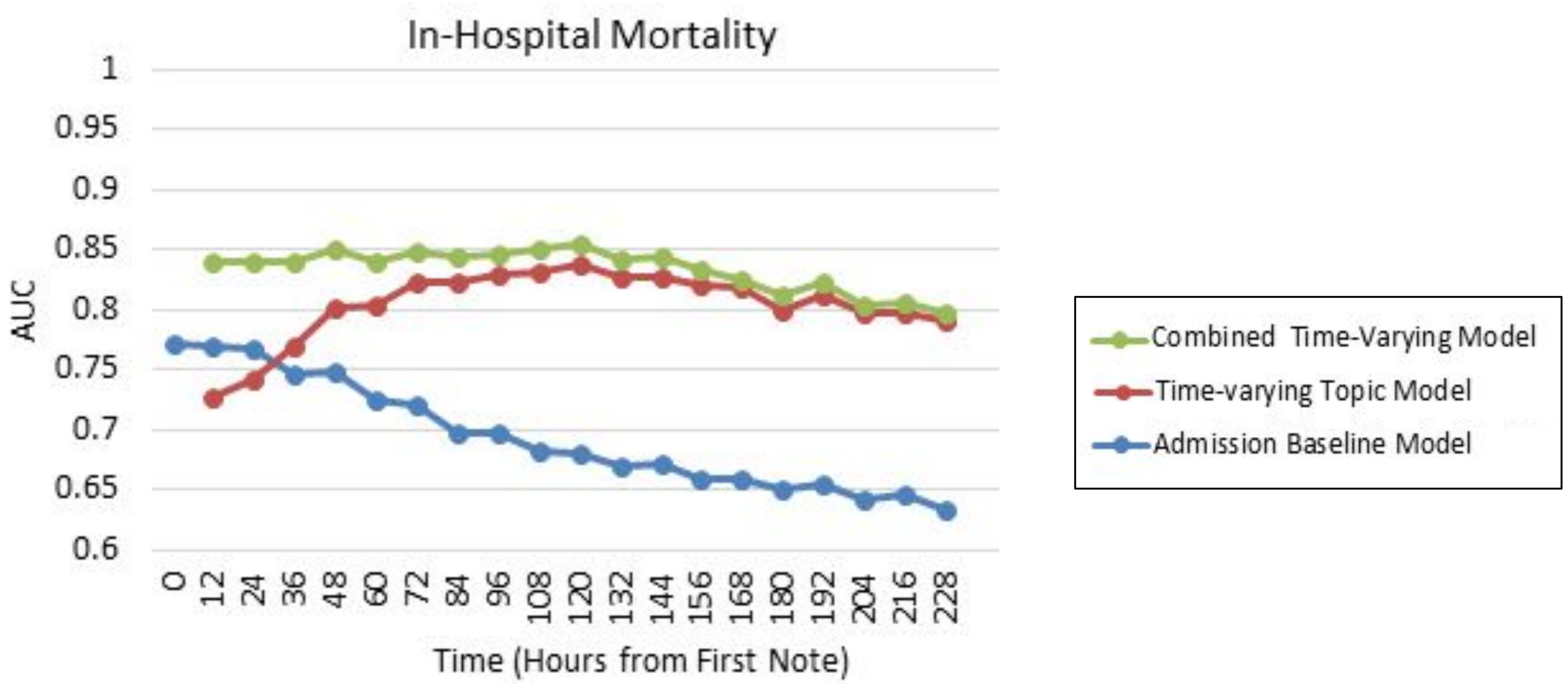
Correlation Between Average Topic Representation and Mortality

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



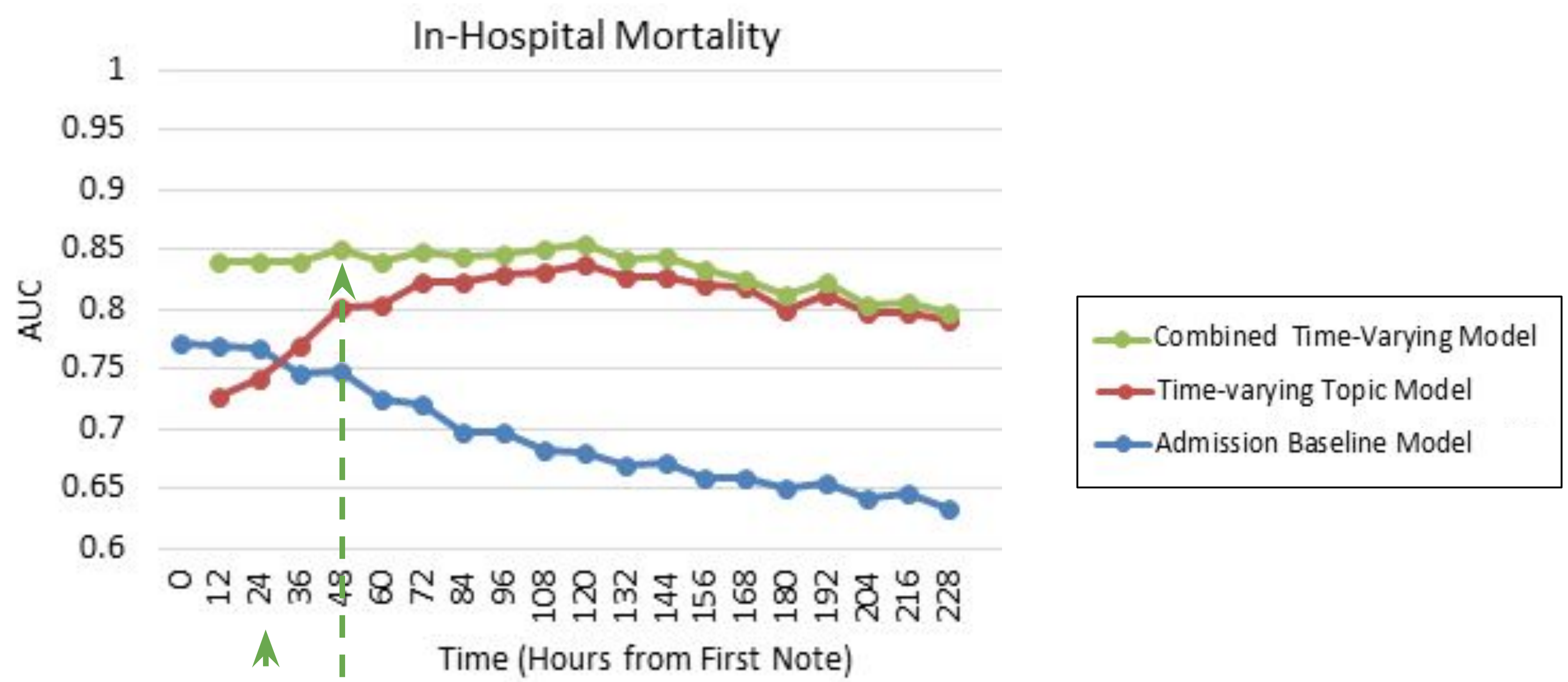
Topic #	Top Ten Words	Possible Topic
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp	Cardiovascular surgery

Topic Representation Improves 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 Haven’t Done 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

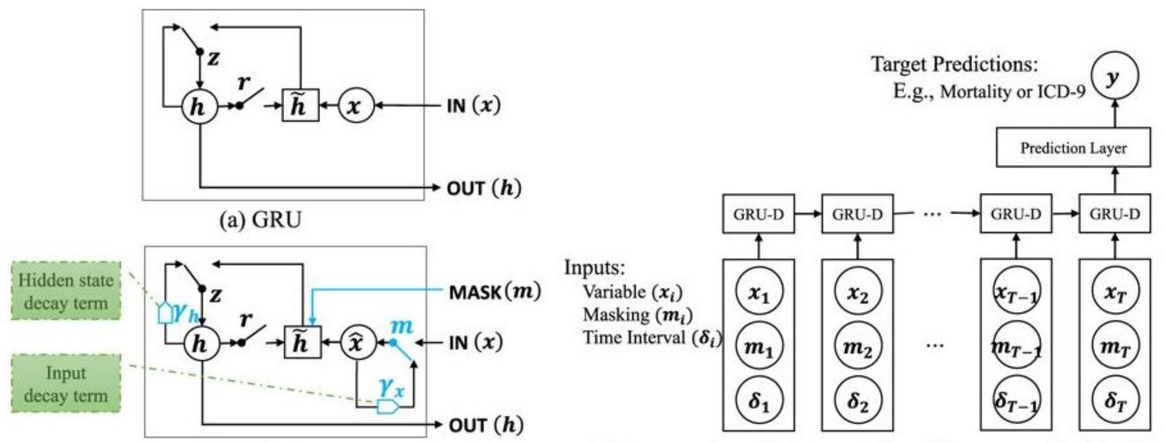
Caballero Barajas, Karla L., and Ram Akella. "Dynamically Modeling Patient's Health State from Electronic Medical Records: A Time Series Approach." *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015.
Che, Zhengping, et al. "Deep computational 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.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Even When Complex and Clever

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



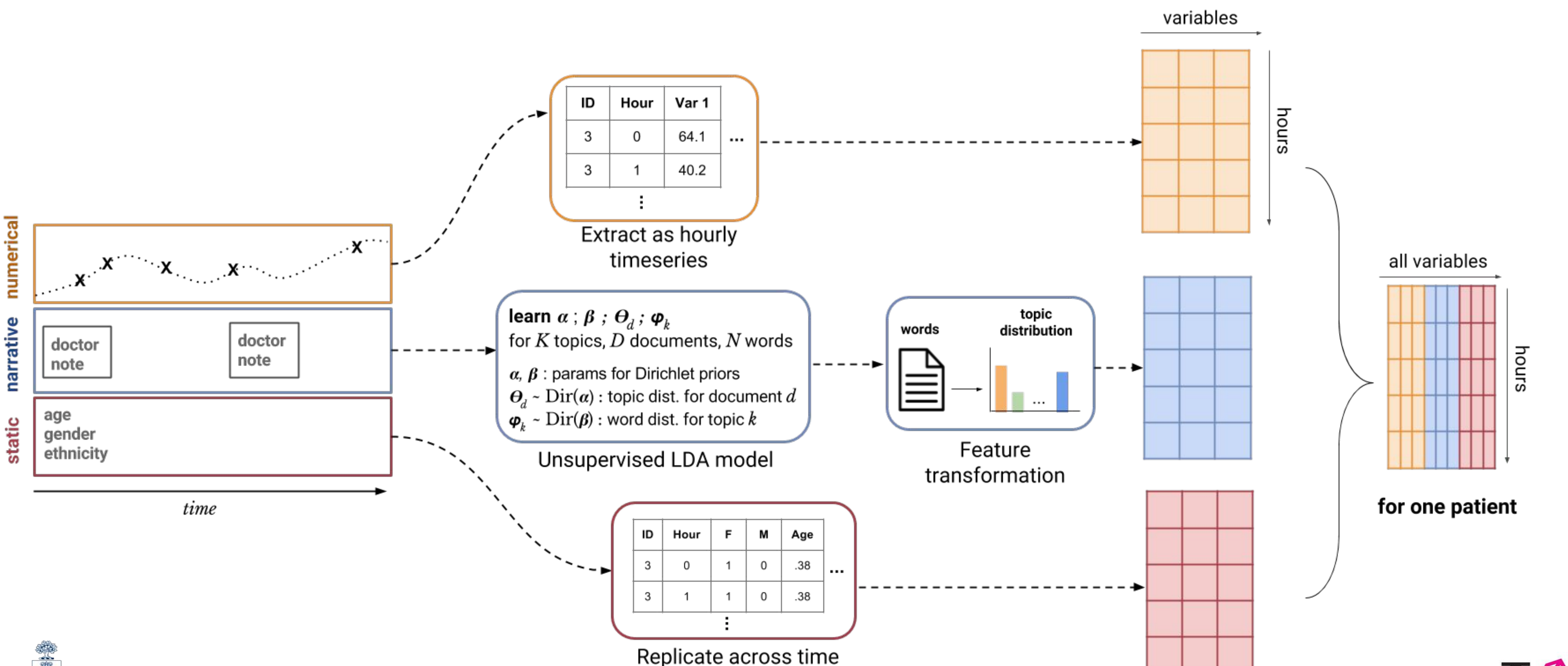
(b) GRU-D (Parts in cyan refer to the modifications.)
(c) Proposed prediction model architecture with GRU-D.

- 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**

Part 2: Predict **Interventions** With Clinical **Data**

- 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

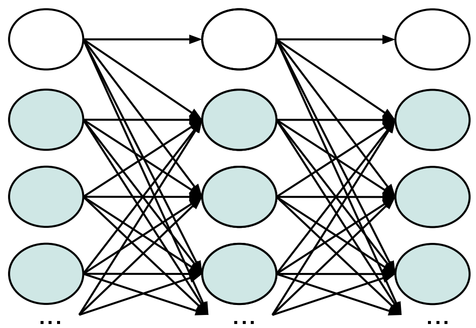


UNIVERSITY OF TORONTO

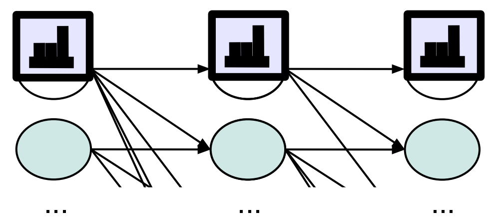
VECTOR INSTITUTE | INSTITUT VECTEUR

Many Ways to Model, What Do We Learn?

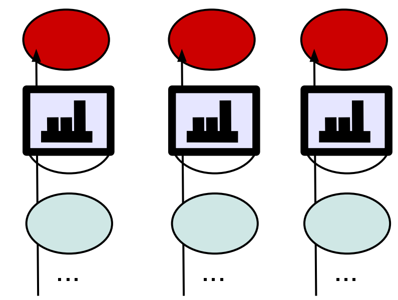
SSAM



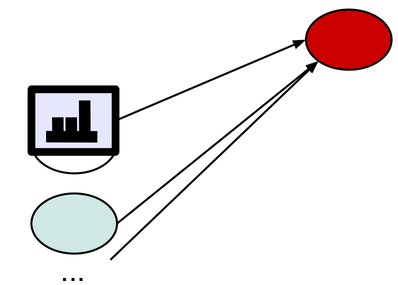
Learn model parameters over patients with variational EM.



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

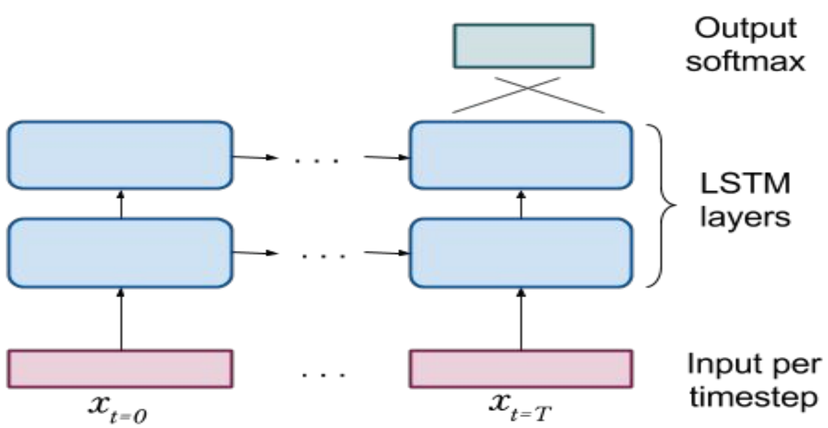


Logistic regression (with label-balanced cost function)



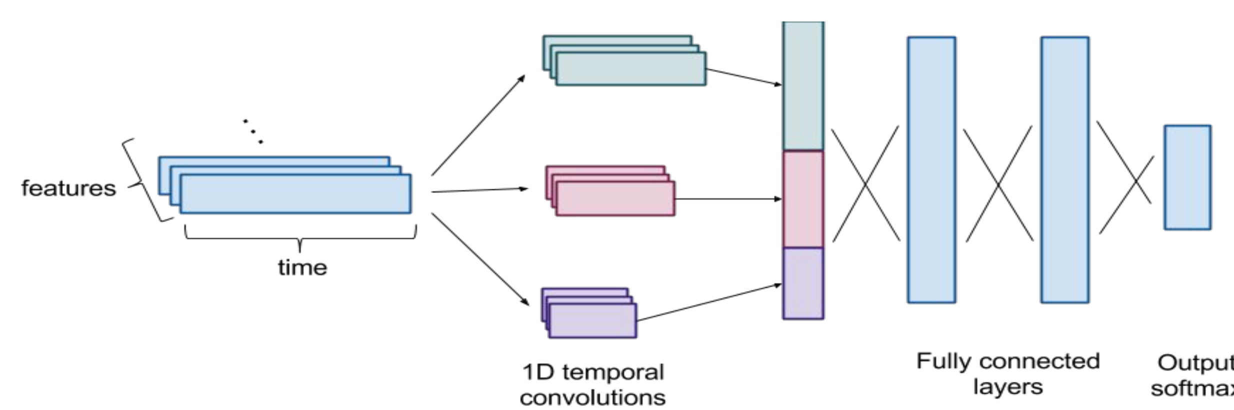
Predict onset in advance

LSTM



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

CNN



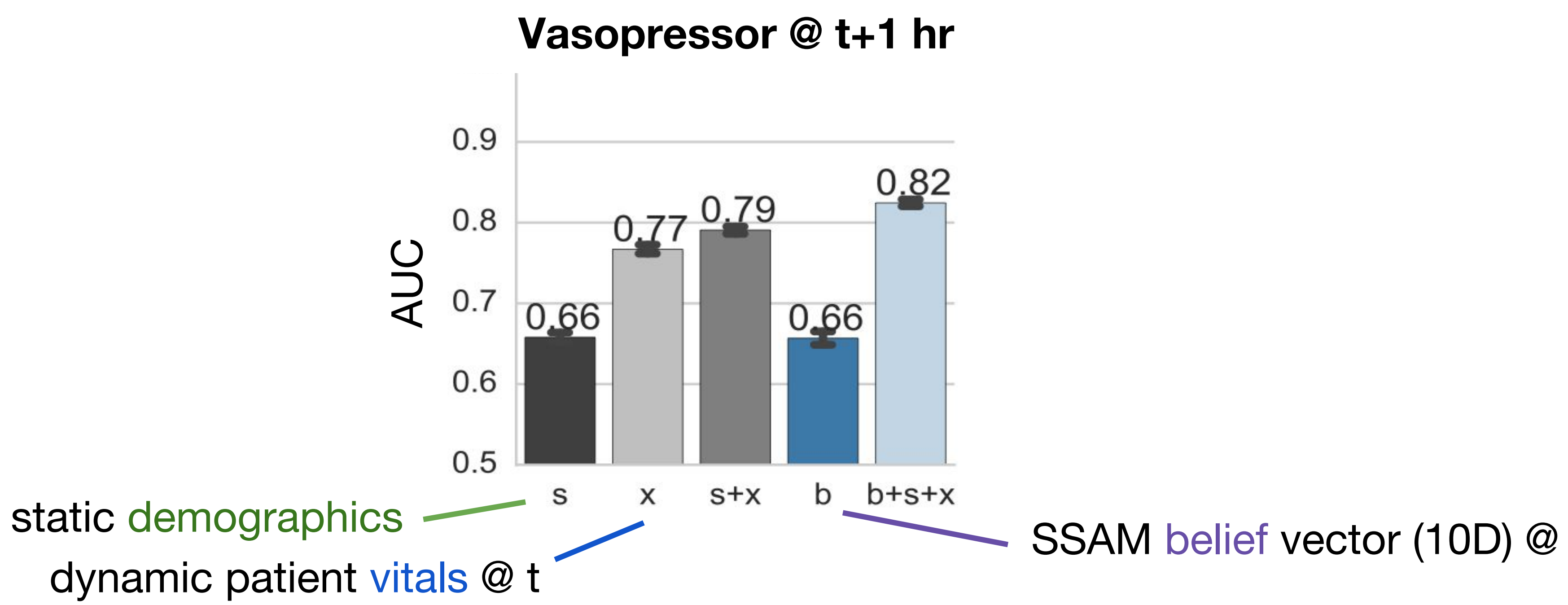
CNN for temporal convolutions at 3/4/5 hours, max-pool, combine the outputs, and run through 2 fully connected layers for prediction.

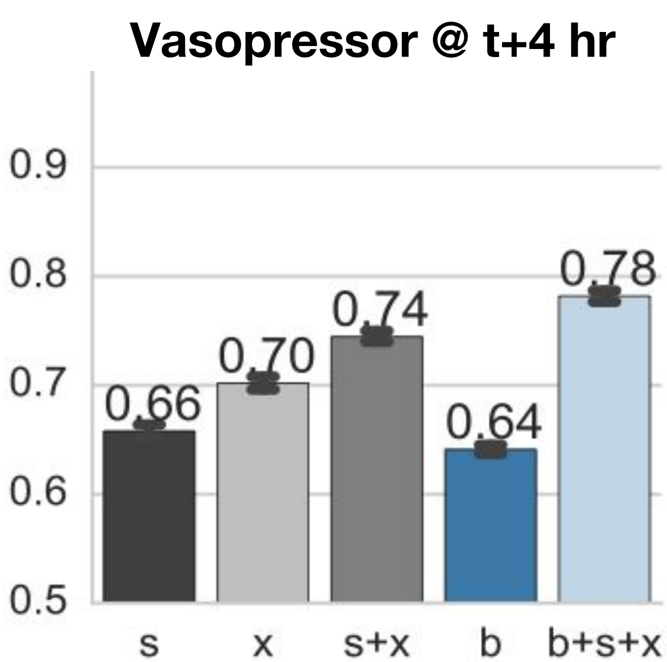
UNIVERSITY OF TORONTO

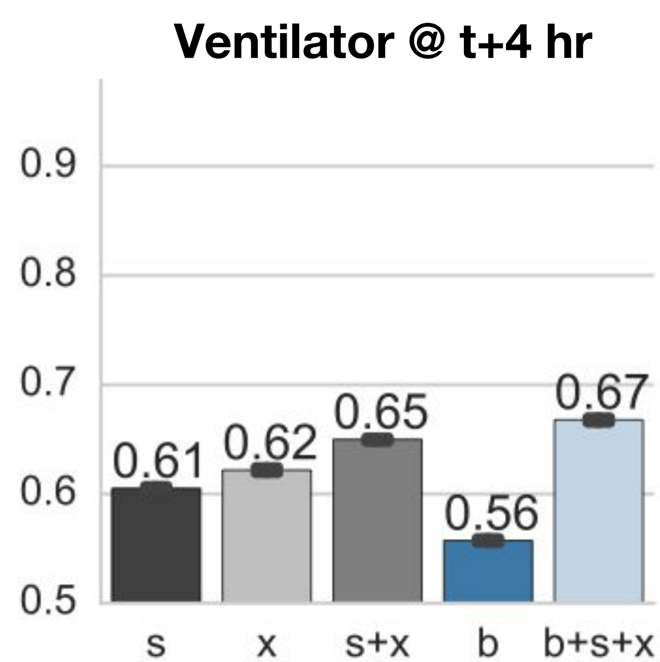
VECTOR INSTITUTE | INSTITUT VECTEUR

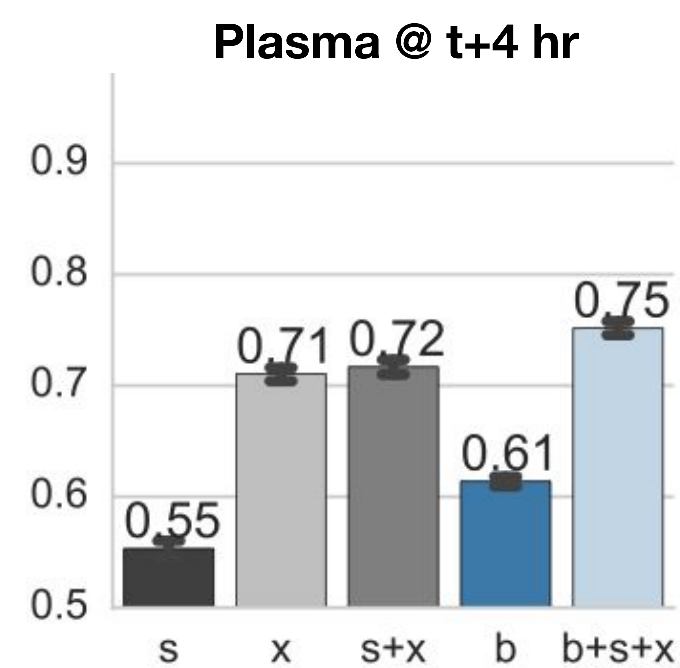
31

State Space Beliefs Improve Prediction







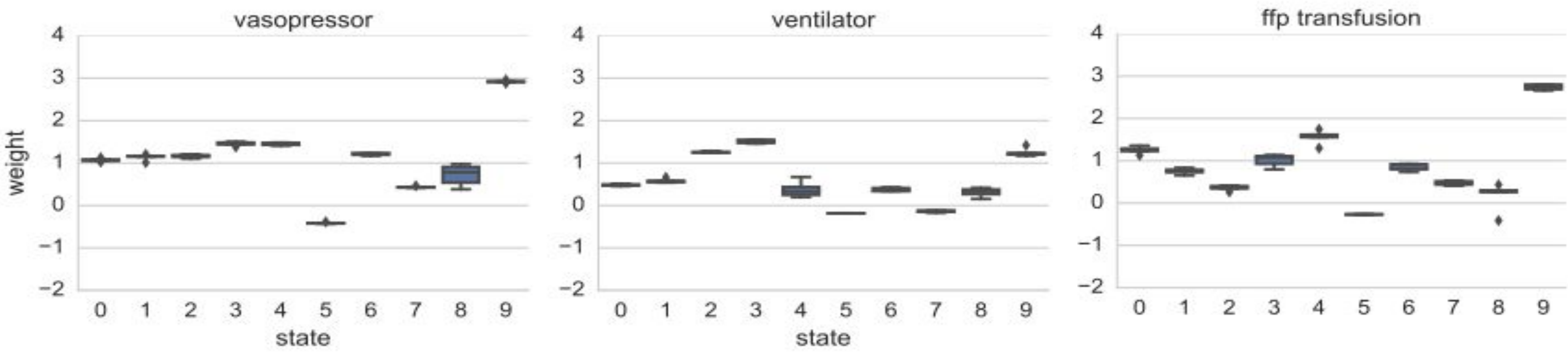


UNIVERSITY OF TORONTO

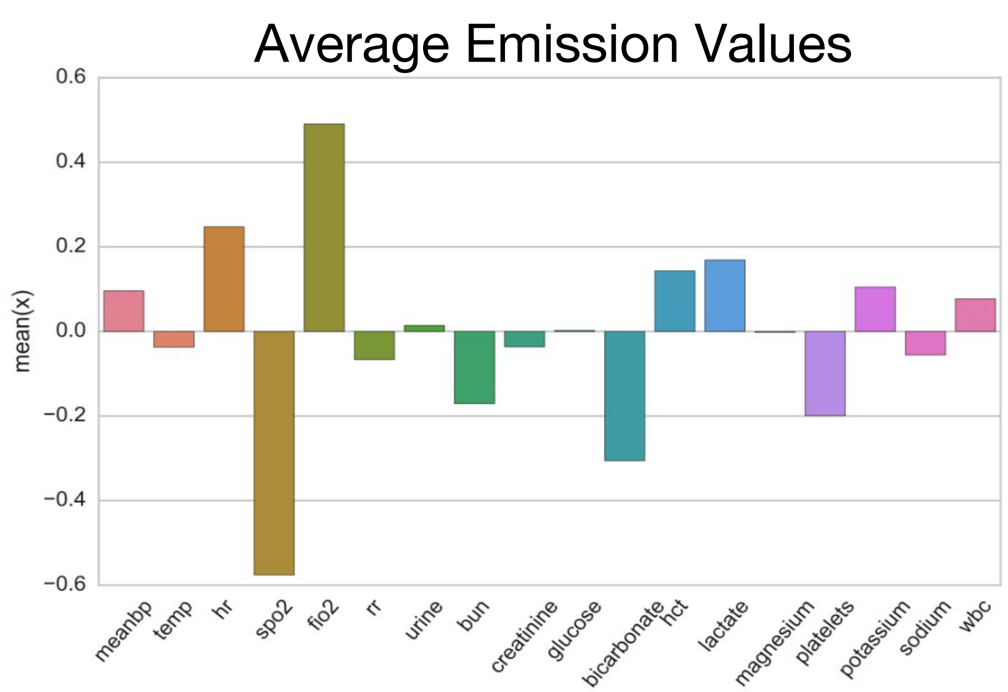
VECTOR INSTITUTE | INSTITUT VECTEUR

SSAM Post-hoc Interpretability

- Interpret classifier weights across interventions.



- Investigate data associated with vasopressor onset state (9).



NNs Do Well; Improved Representation Helps

Area-under-ROC

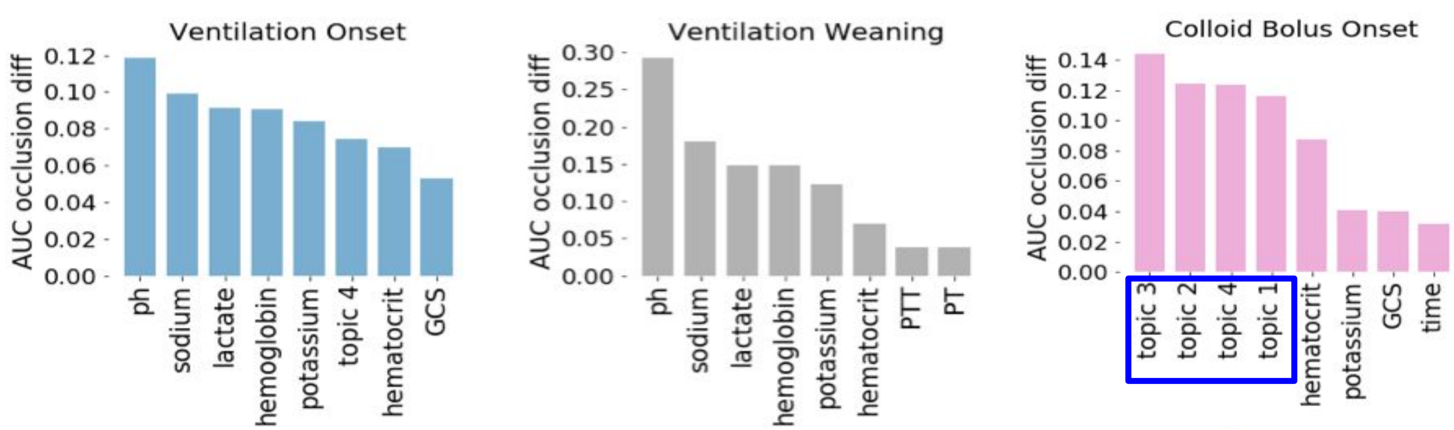
Task	Model	Intervention Type: VENT	NI-VENT	VASO	COL BOL	CRYS BOL
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	**0.76**	0.76	**0.72**	**0.71**
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	LSTM Raw	0.90	0.80	**0.91**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

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

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

NN Post-hoc Interpretability

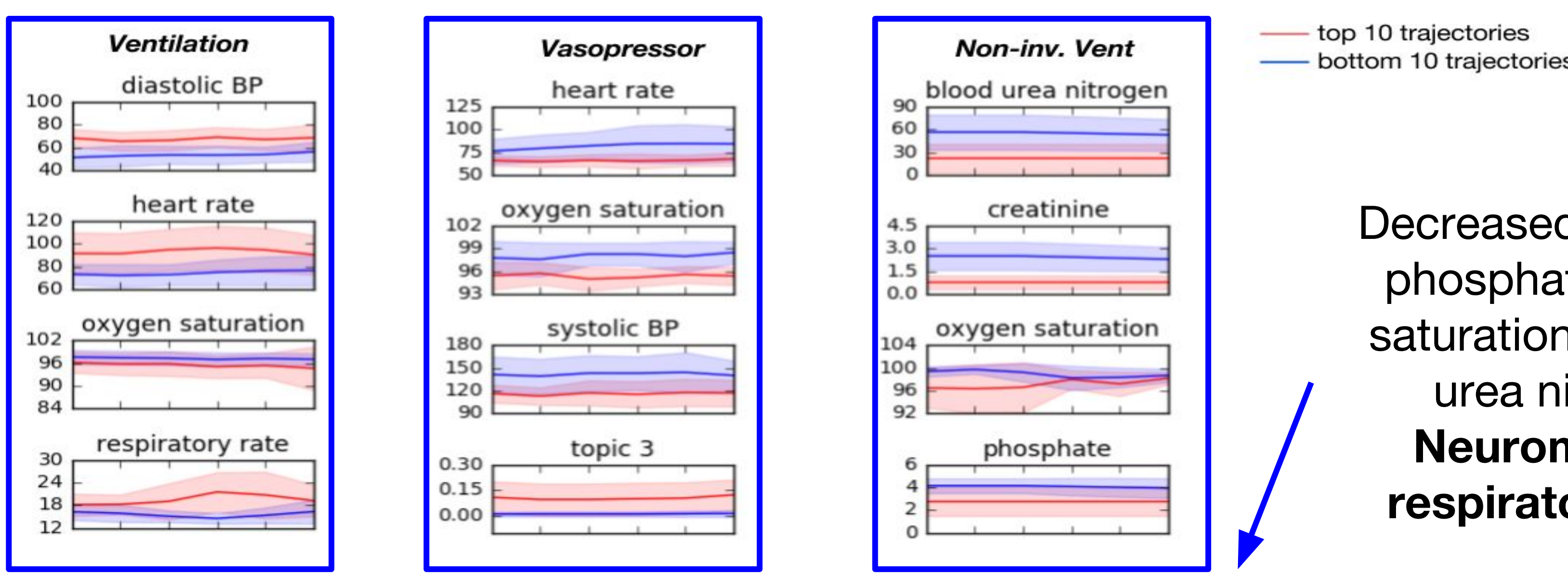
- Feature-level occlusions identify important per-class features.



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

- Convolutional filters target known short-term trajectories.

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



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

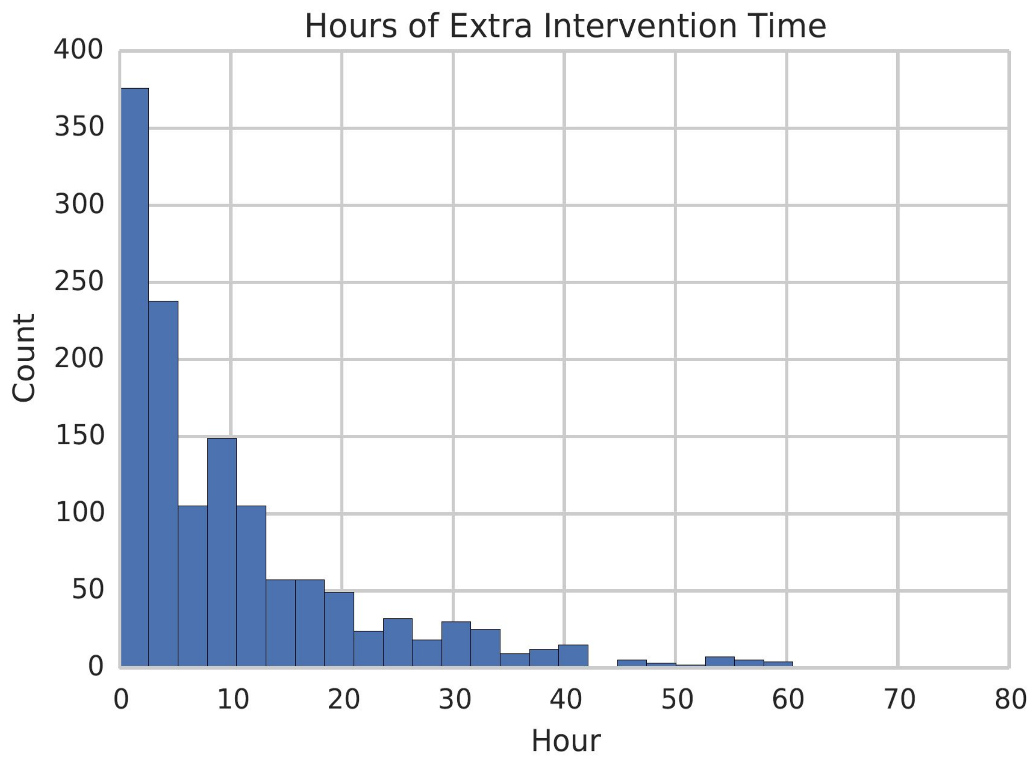
Decreased systolic blood pressure, heart rate and oxygen saturation rate : **Altered peripheral perfusion** or **stress hyperglycemia**

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

ML for Healthcare, or ML for Health?

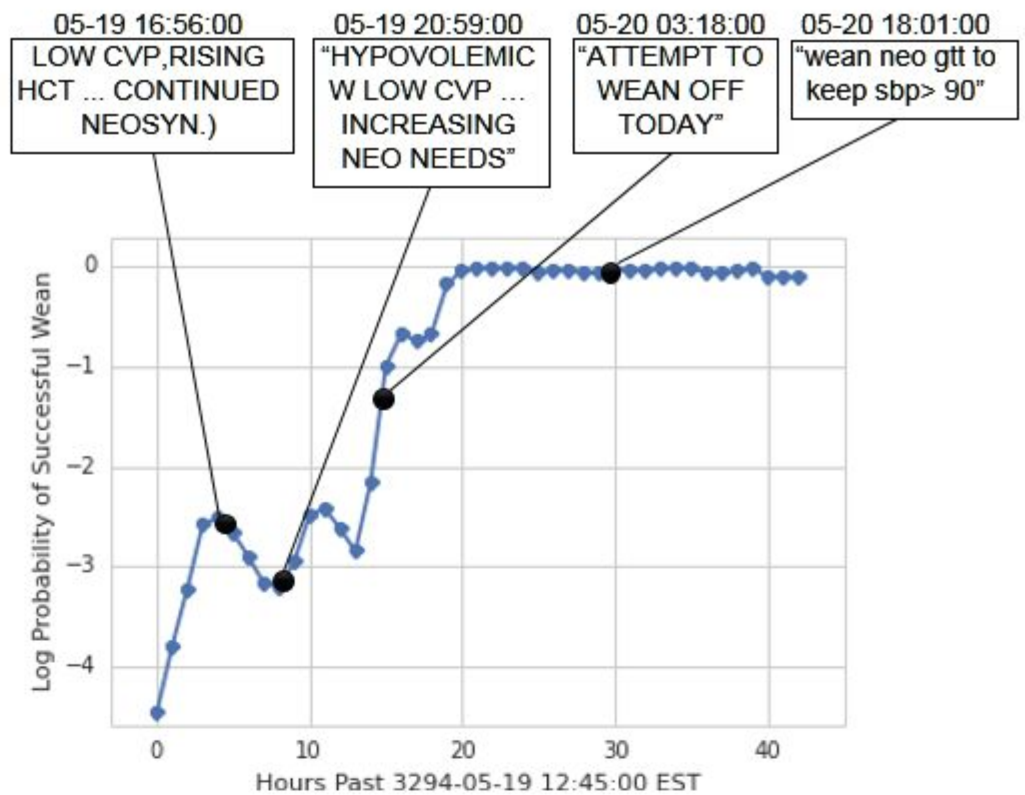
- Patients can be left on interventions longer than necessary.



- 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?



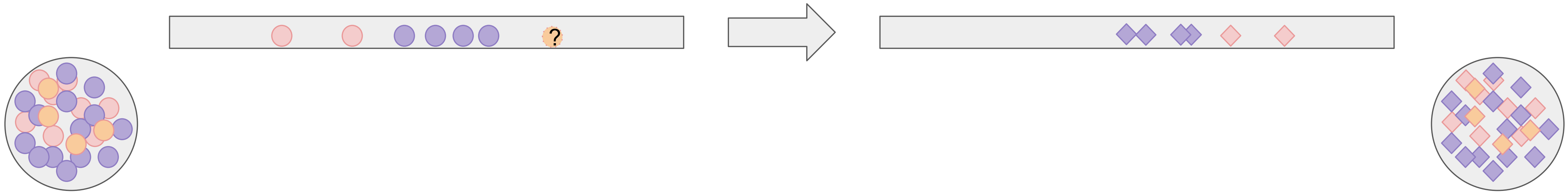
- One example of a 62-year-old male patient with a cardiac catheterization.
- More complexity/higher misclassification penalty don’t solve this!

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Part 3: Forecast **Response** to An **Intervention**

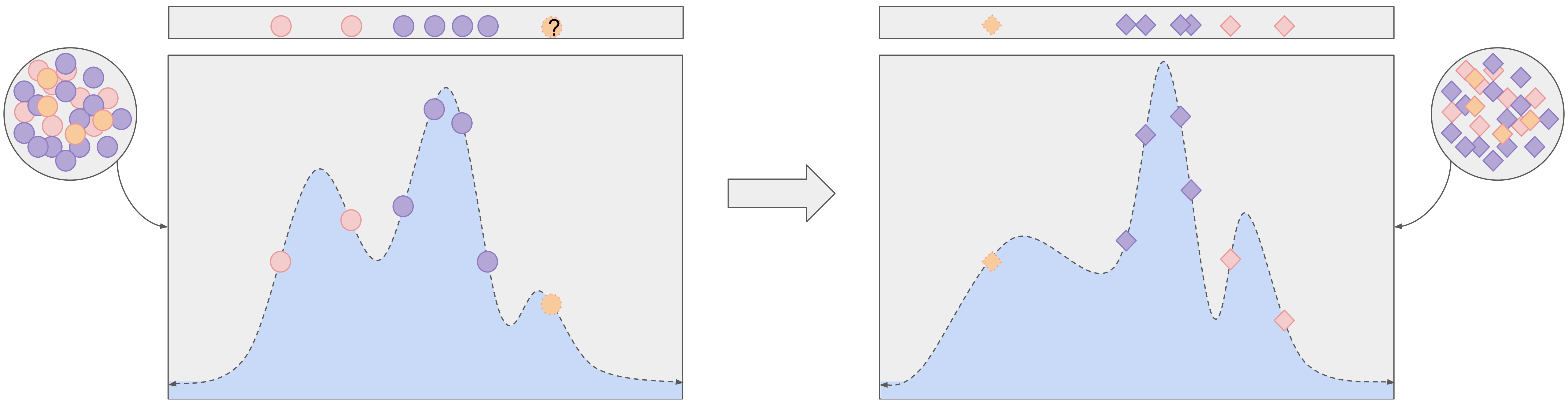
- Fully paired biomedical datasets are
  - Privacy sensitive
  - Expensive and difficult to collect
  - Often homogenous



- Sufficiently large, heterogeneous paired datasets are rare.

Using Adversarial Training To Overcome Missingness

- GANs are used for data augmentation[1], imputation[2].

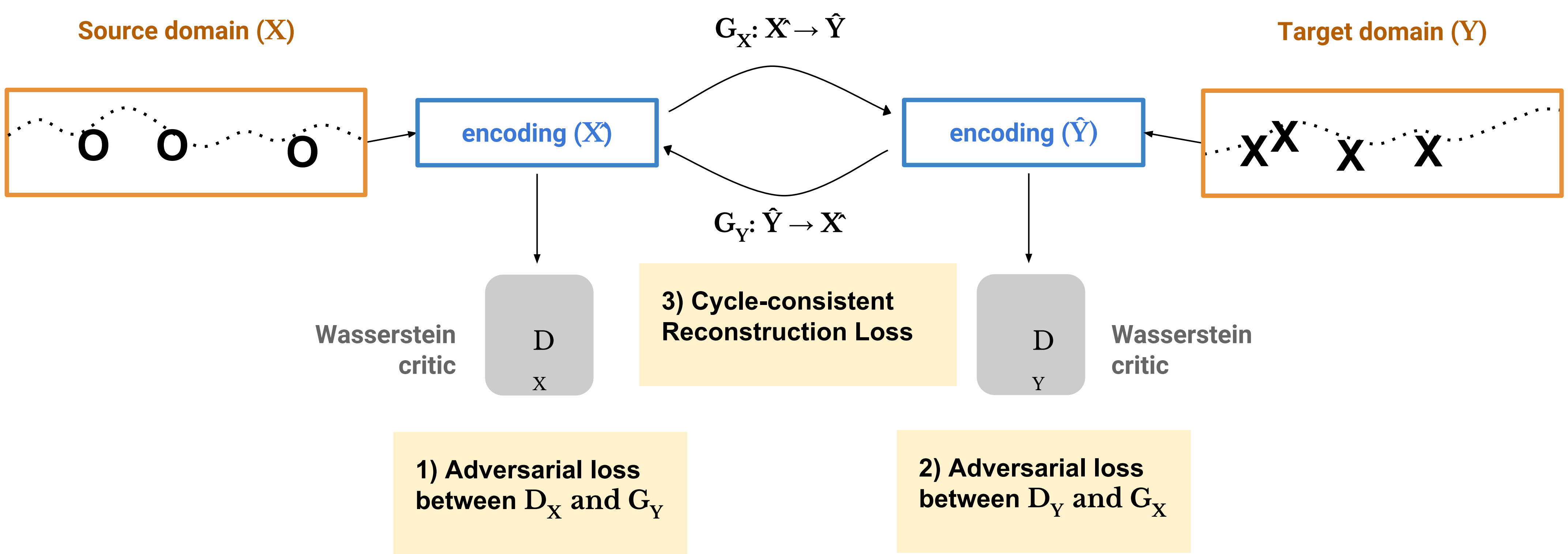


- We use adversarial learning techniques to learn distributional signals from additional, unpaired data to augment predictions on a limited training set.

[1] Armanious K, Yang C, Fischer M, Küstner T, Nikolaou K, Gatidis S, Yang B. MedGAN: Medical Image Translation using GANs. arXiv preprint arXiv:1806.06397. 2018 Jun 17.
[2] Yoon J, Jordon J, van der Schaar M. GAIN: Missing Data Imputation using Generative Adversarial Nets. arXiv preprint arXiv:1806.02920. 2018 Jun 7.

Model Learns on Unpaired Data, G_X Used to Eval

- Ensure generated samples are realistic, account for missing samples (not just missing features), and ensure cycle/self-consistency.[1]



[1] Ghasedi Dizaji K, Wang X, Huang H. Semi-Supervised Generative Adversarial Network for Gene Expression Inference. InProceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018 Jul 19 (pp. 1435-1444). ACM.

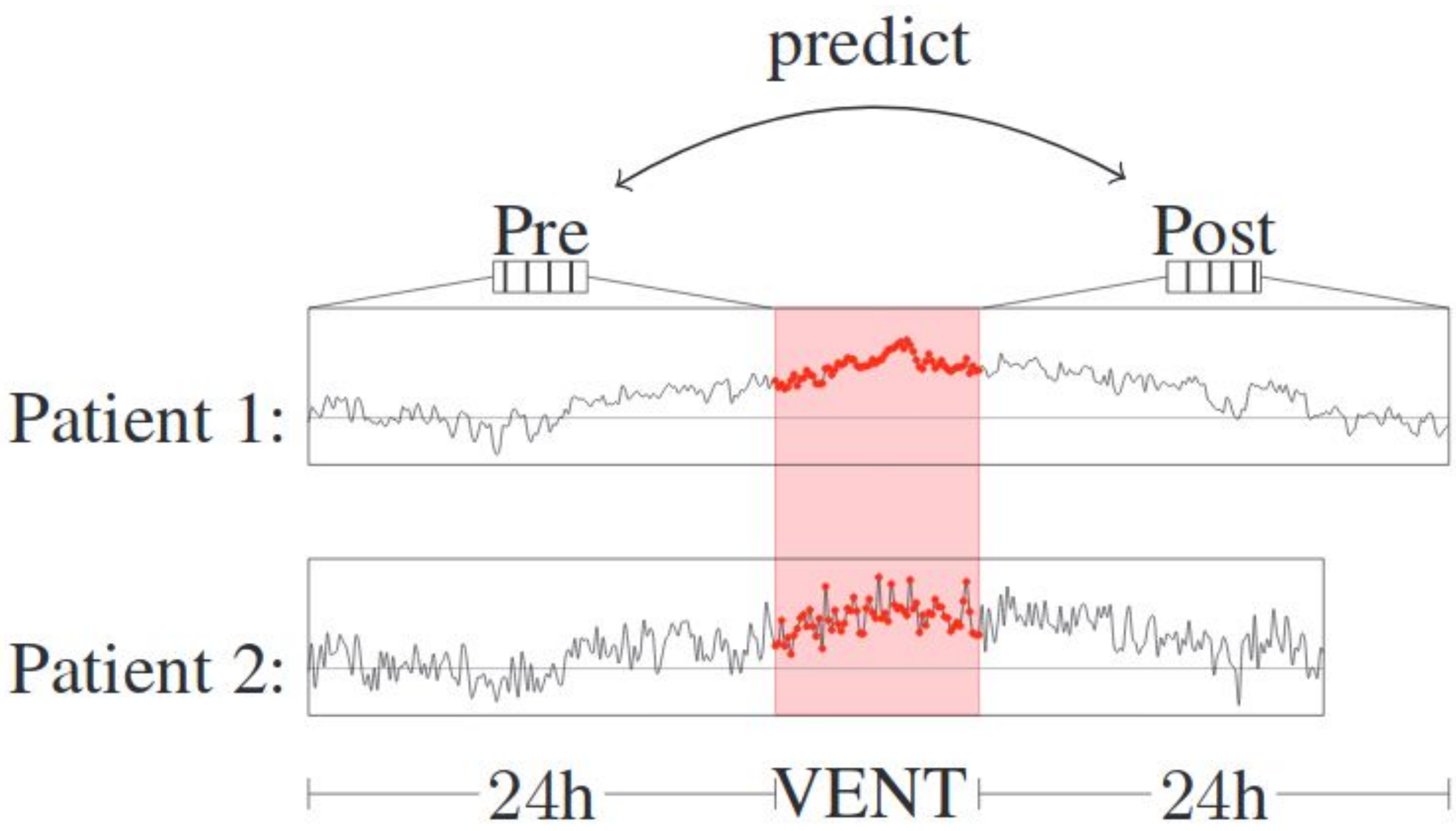
40

Improved Intervention Response Prediction

	Intervention Type			
Model MSE	VENT	NOREP	DOP	PHEN
Baseline MLP	3.780	2.829	2.719	3.186
CWR-GAN (% Delta)	-0.5%	-7.4%	+2.7%	-4.5%

- Mean-squared-error of a traditional MLP on only paired intervention data vs. the CWR-GAN augmented with data that failed to meet inclusion criteria on either the pre-intervention side or post-intervention side (~500 paired, ~3,000 unpaired patients).

The Problem With Models That Learn...



- Exciting work on to be done on learning what treatments are best for individuals based on environment and context!

- But there are other factors...

42

Health Questions Beyond The Obvious

Across these use cases, a number of ethical, social, and political challenges are raised and the 10 most important are:

01 What effect will AI have on human relationships in health and care?

02 How is the use, storage and sharing of medical data impacted by AI?

03 What are the implications of issues around algorithmic transparency/explainability on health?

04 Will these technologies help eradicate or exacerbate existing health inequalities?

05 What is the difference between an algorithmic decision and a human decision?

06 What do patients and members of the public want from AI and related technologies?

07 How should these technologies be regulated?

08 Just because these technologies could enable access to new information, should we always use it?

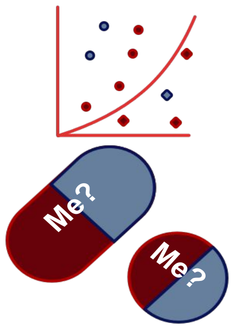
09 What makes algorithms, and the entities that create them, trustworthy?

10 What are the implications of collaboration between public and private sector organisations in the development of these tools?

[1] "Ethical, social, and political challenges of artificial intelligence in Health". *Wellcome Trust* April 2018. © Future Advocacy.

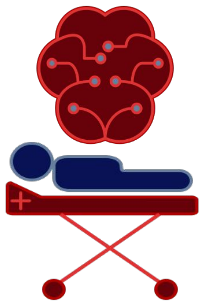
43

Machine Learning For Health (ML4H)



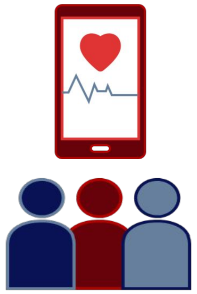
1. What Models are Healthy? Learning Good Representations.

Unfolding Physiological State: Mortality Modelling in Intensive Care Unit (KDD 2014); A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU … (AAAI 2015);
Predicting Early Psychiatric Readmission with Natural Language Processing of Narrative … (Nature Trans Psych 2016);
Predicting Intervention Onset in the ICU with Switching State Space Models (AMIA-CRI 2017);
Clinical Intervention Prediction and Understanding using Deep Networks (MLHC 2017/JMLR W&C V68);
Semi-supervised Biomedical Translation with Cycle Wasserstein Regression GANs (AAAI 2018);



2. What Healthcare is Healthy? Stratifying Human Risks.

Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017);
Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop);
The Disparate Impacts of Medical and Mental Health with AI. (In submission);
ClinicalVis Project with Google Brain. (*In submission);



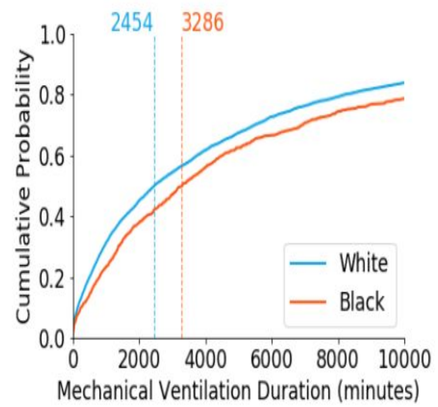
3. What Behaviors are Healthy? Inferring Unseen Actions and States.

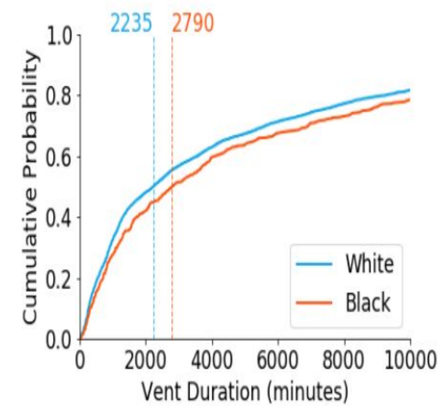
Learning to Detect Vocal Hyperfunction from Ambulatory Necksurface Acceleration Features (IEEE TBME 2014);
Uncovering Voice Misuse Using Symbolic Mismatch (MLHC 2016/JMLR W&C V56);
Project BASELINE Mood Study with Alphabet's Verily;

Modelling Mistrust in EOL Care

- Replicate documented racial disparities in open databases.

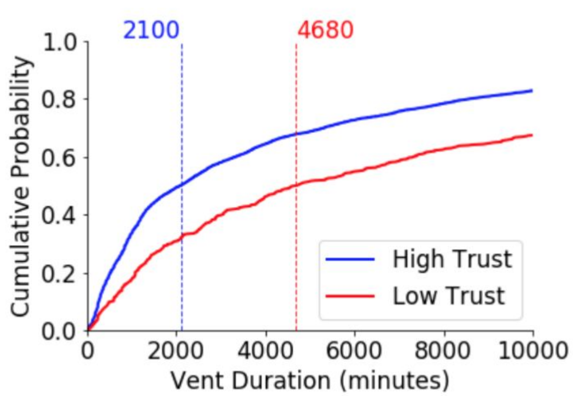
(a) *MIMIC Mechanical Ventilation*
White: 4810 patients
Black: 510 patients
p=0.005





(b) *eICU Mechanical Ventilation*
White: 4911 patients
Black: 655 patients
$p < 0.001$

- Algorithmically mistrust demonstrates treatment disparity > than race, even with acuity factored in.



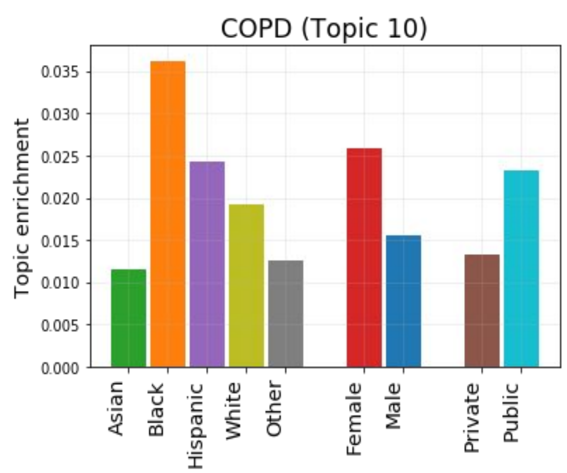
(a) **Mechanical Ventilation**
High Trust: 4810 patients
Low Trust: 510 patients
$p < 0.001$

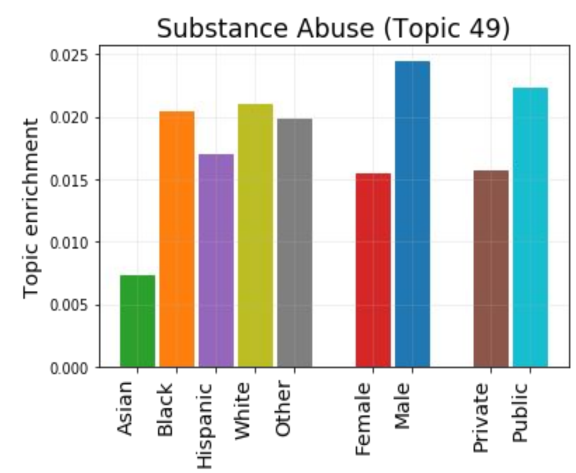
Table 4: Pairwise Pearson correlation coefficients between scores.

	OASIS	SAPS II	Noncompliance	Autopsy	Sentiment
OASIS	1.0	0.679	0.050	-0.012	0.075
SAPS II	0.679	1.0	0.013	-0.013	0.086
Noncompliance	0.050	0.013	1.0	0.262	0.058
Autopsy	-0.012	-0.013	0.262	1.0	0.044
Sentiment	0.075	0.086	0.058	0.044	1.0

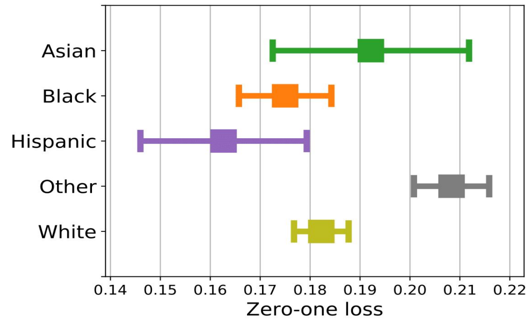
Disparate Impacts of Medical and Mental Health

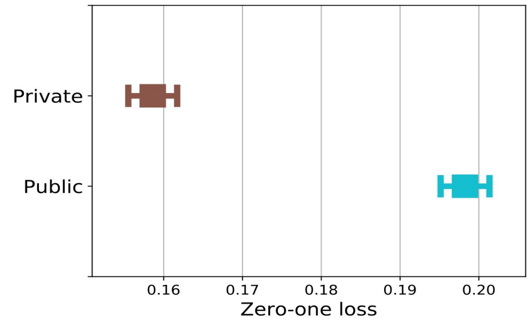
- We can predict **ICU** mortality and 30-day **psychiatric** readmission, but notes have group-specific heterogeneity.

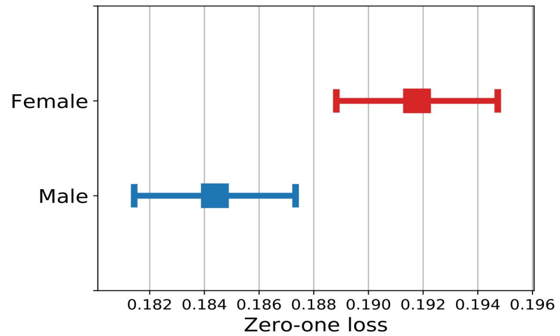


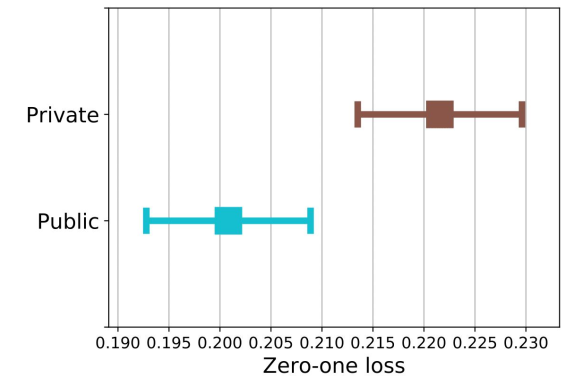


- Significant differences in model accuracy for race, sex, and insurance type in **ICU notes** and insurance type in **psychiatric notes**.









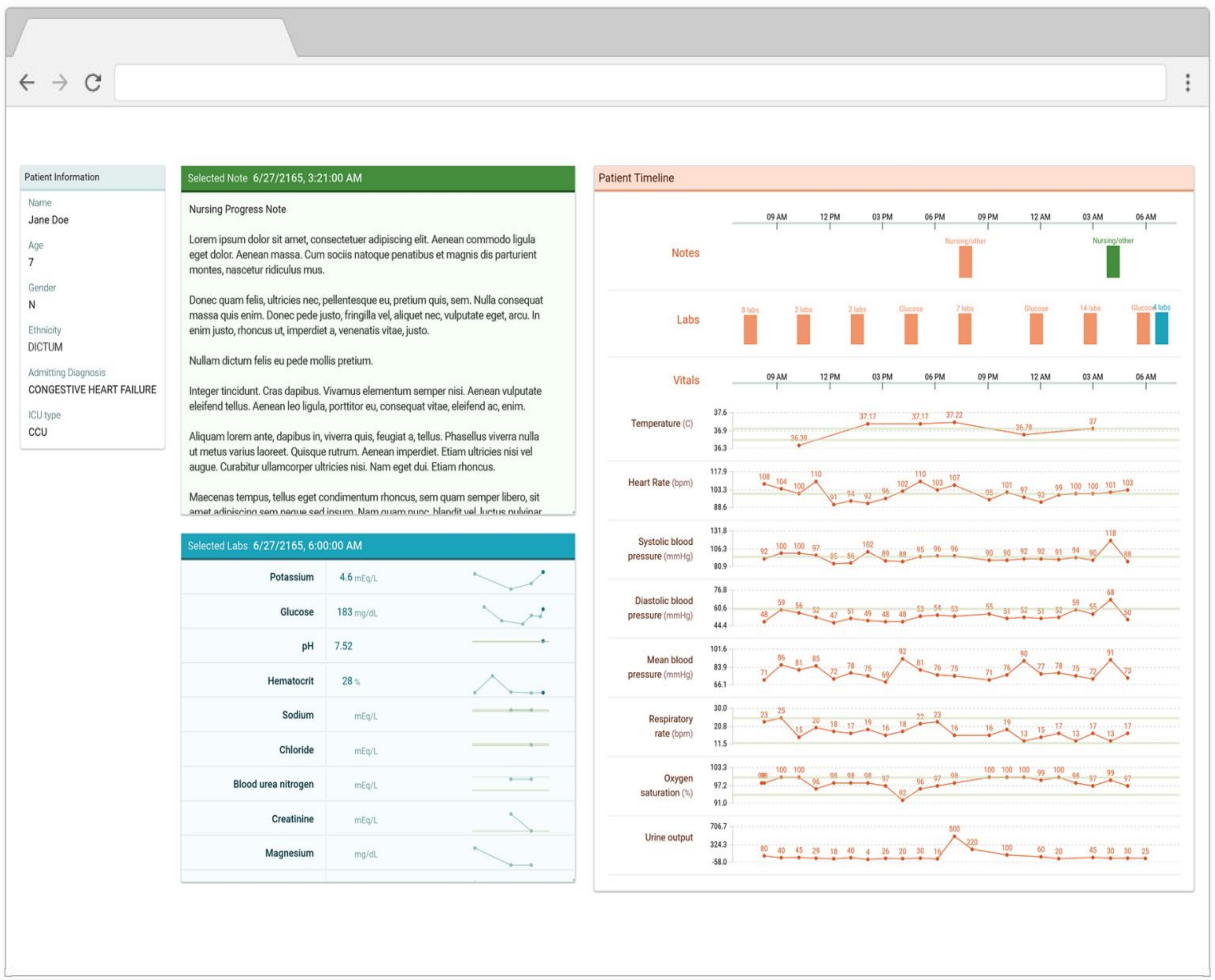
UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

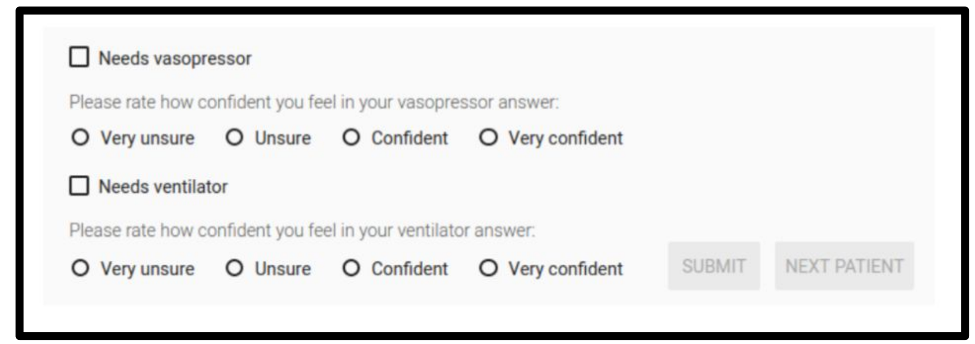
46

ClinicalVis: Supporting Clinical Task-Focused Design Evaluation

1. Present real patient data to HCPs using open-source prototype.



2. Ask HCPs to plan care for two interventions in an eICU simulation.



3. Evaluate the confidence, accuracy and time-to-task under different visual prototypes.

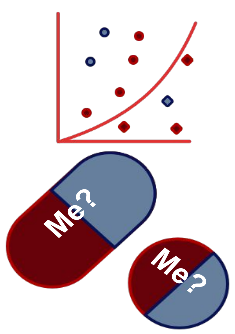
		Vasopressor Positive (VP+)	Ventilator Positive (VE+)
Accuracy (%)	Baseline	50.00 %	56.25 %
	ClinicalVis	68.83 %	62.79 %
Confidence Score	Baseline	0.68	0.87
	ClinicalVis	1.41	1.27
Average Time to Task (seconds)	Baseline	92.31 s	92.73 s
	ClinicalVis	84.43 s	86.86 s

UNIVERSITY OF TORONTO

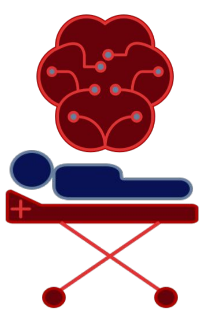
VECTOR INSTITUTE | INSTITUT VECTEUR

47

Future of Machine Learning For Health (ML4H)



1. What Models are Healthy? Learning Good Representations.
- Balancing multi-target output learning
- Finding useful abstractions
- “Explaining” decisions in case/controls



2. What Healthcare is Healthy? Stratifying Human Risks.
- Providing meaningful, calibrated notions of uncertainty
- Finding causes and establishing causality
- Defining and targeting fairness

3. What Behaviors are Healthy? Inferring Unseen Actions and States.
- Data quality and availability
- Real-time decision making
- Robustness in the face of unexpected data

CSAIL

ML4H @ UToronto Team!

Visit http://www.marzyehghassemi.com/ for more information.

University of Toronto Students

Bret Nestor

Denny Wu

MIT Students

Matthew McDermott

Irene Chen

Harini Suresh

Clinical Collaborators

Dr. Muhammad Mamdani

Dr. Leo Anthony Celi

Technical Collaborators

Tristan Naumann

Rajesh Ranganath

Anna Goldenburg

Andrew Beam

Peter Szolovits

49

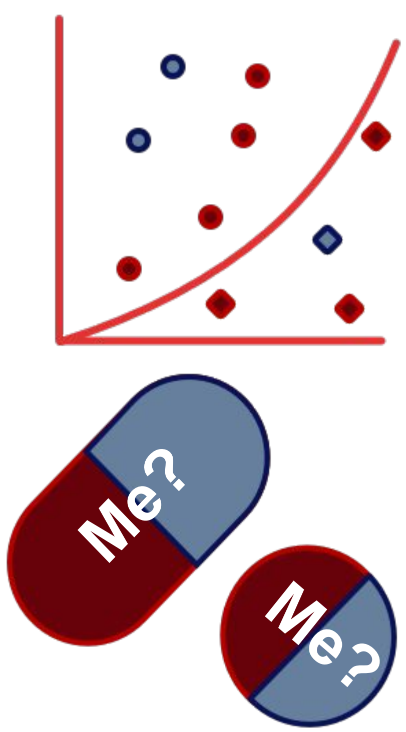
What Can You Do?

- **Help Identify Targets for Clinical Machine Learning That Matters!**
  Establish **clinical** opinions on existing ML targets, and suggest additional targets.
  https://goo.gl/forms/xEd9fcWcO80GuNJt1

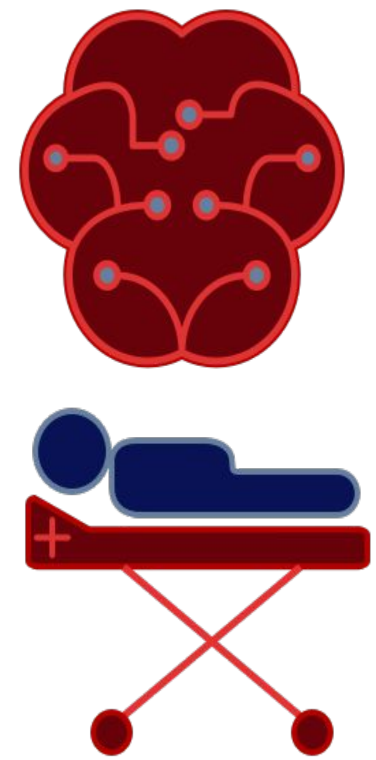
- **Mentor a Team in New Project-Based CS Grad Course for ML students!**
  Create collaborations between technical and non-technical researchers, and consider the implications of machine learning in health. If you have a potential project with a) data that students could access, b) a supervisor for the Winter term, and c) an interest in publishing the work with the student if it goes well!
  Topics in Machine Learning: Machine Learning for Health

- **Indicate interest in ML4H 2019 Unconference held in Toronto, Ontario!**
  Invitational "unconference" style meeting in May 2019 to facilitate junior ML researchers and doctors connecting. Many projects in ML4H suffer from a mismatch in data, tools, and skills. Our focus this year will be on What Problems Should ML4H Be Solving?
  https://goo.gl/forms/jzIvKaDpxfY0doYy2

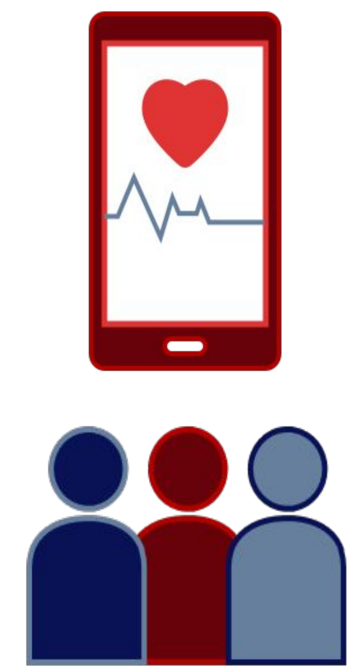
Machine Learning For Health (ML4H)



What models are healthy?



What healthcare is healthy?



What behaviors are healthy?

51