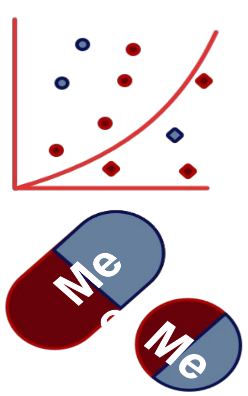
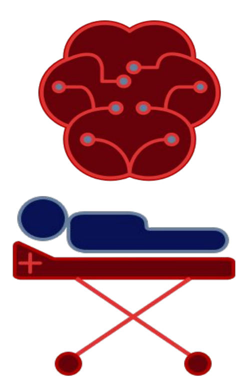
Learning Healthy Models for Healthcare

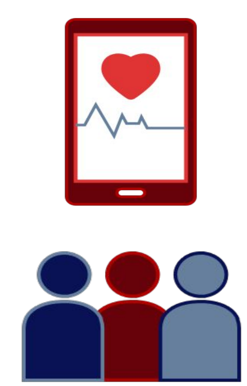
Dr. Marzyeh Ghassemi, PhD MIT CSAIL

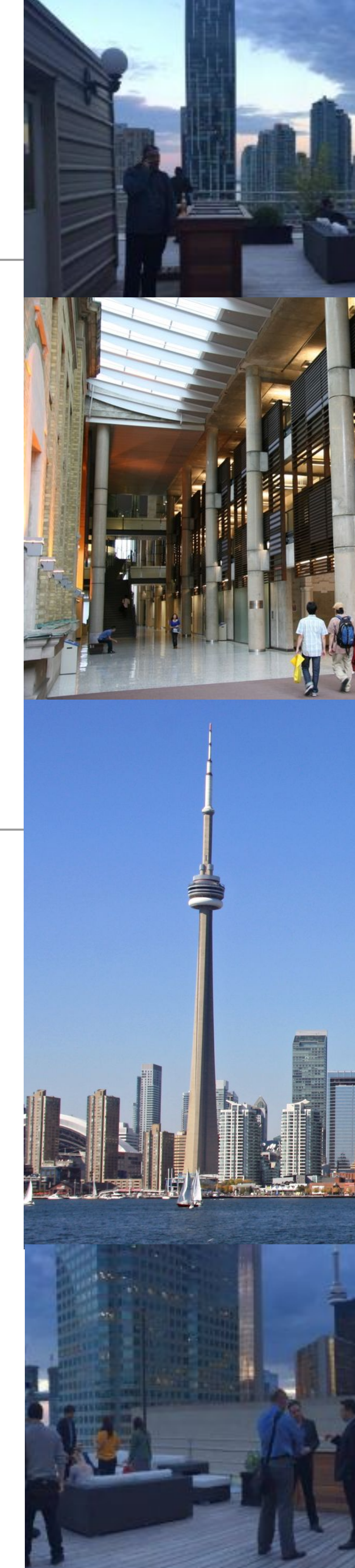
University of Toronto, Computer Science and Medicine
Canadian CIFAR AI Chair, Vector Institute
NeuroIPS Workshop Co-Chair

www.marzyehghassemi.com, @MarzyehGhassemi









Why Try To Work in Health?

- Improvements in health **improve lives**.

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

- Improving care requires **evidence**.

Why Try To Work in Health?

- Improvements in health **improve lives**.

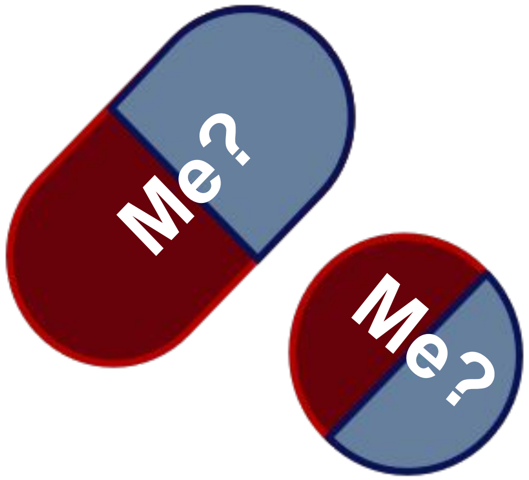
- Same **patient** → different **treatments** across hospitals, clinicians.

- Improving care requires **evidence**.

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

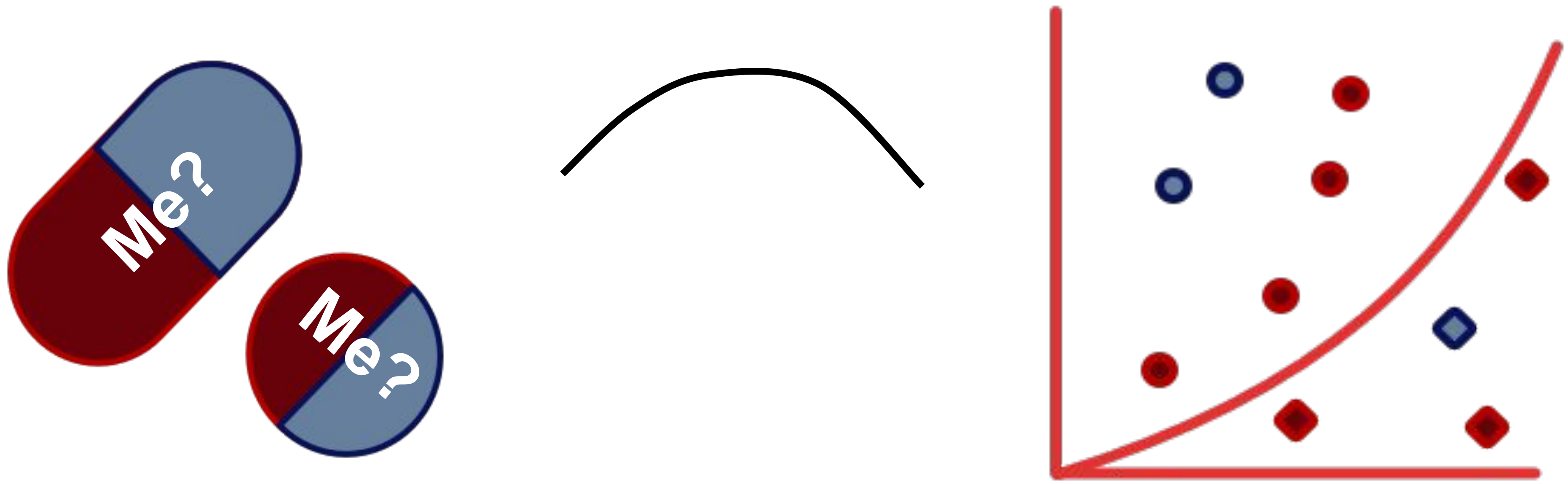
Learning What Is Healthy?

Recruit a study population.



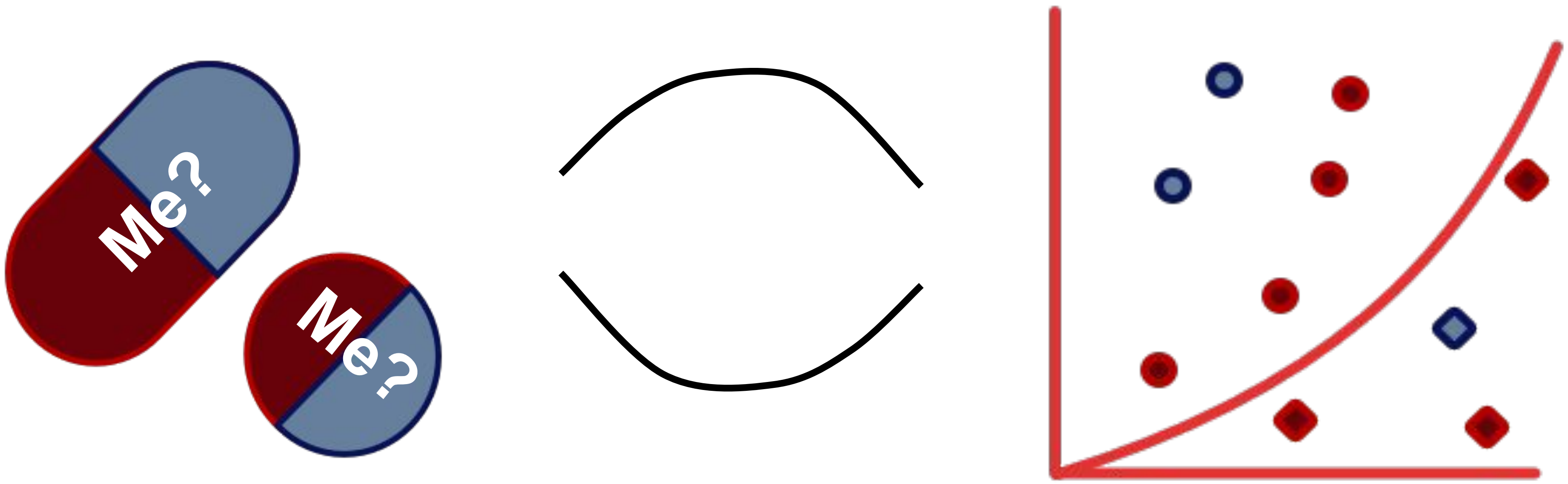
Learning What Is Healthy?

Learn a rule.



Learning What Is Healthy?

Does it generalize?

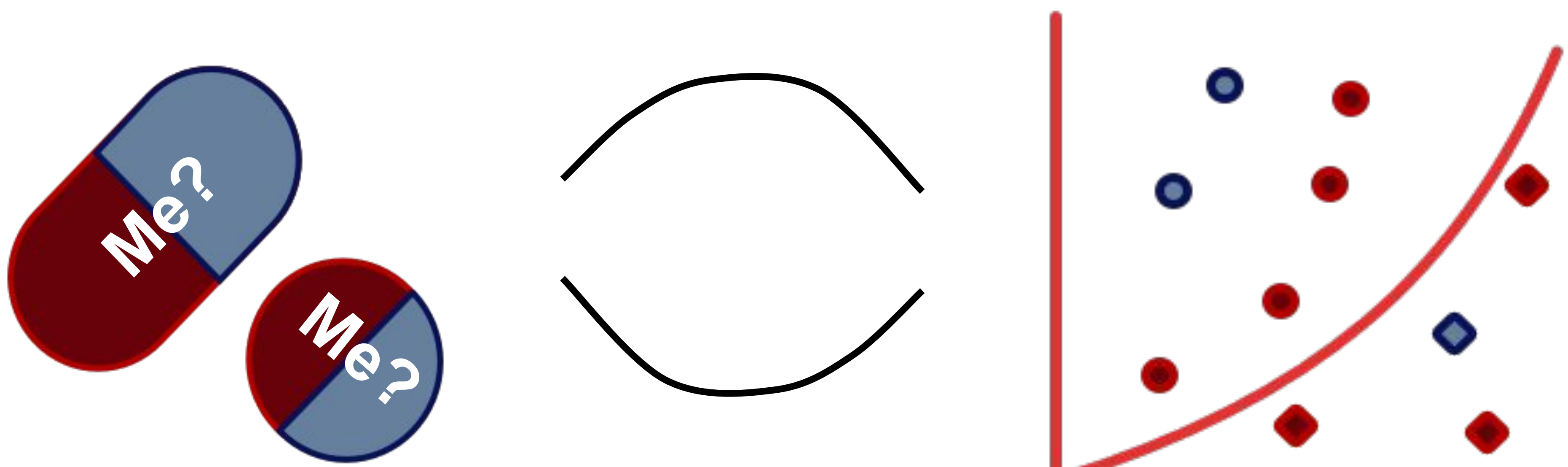


UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Learning What Is Healthy?

For whom does it generalize?

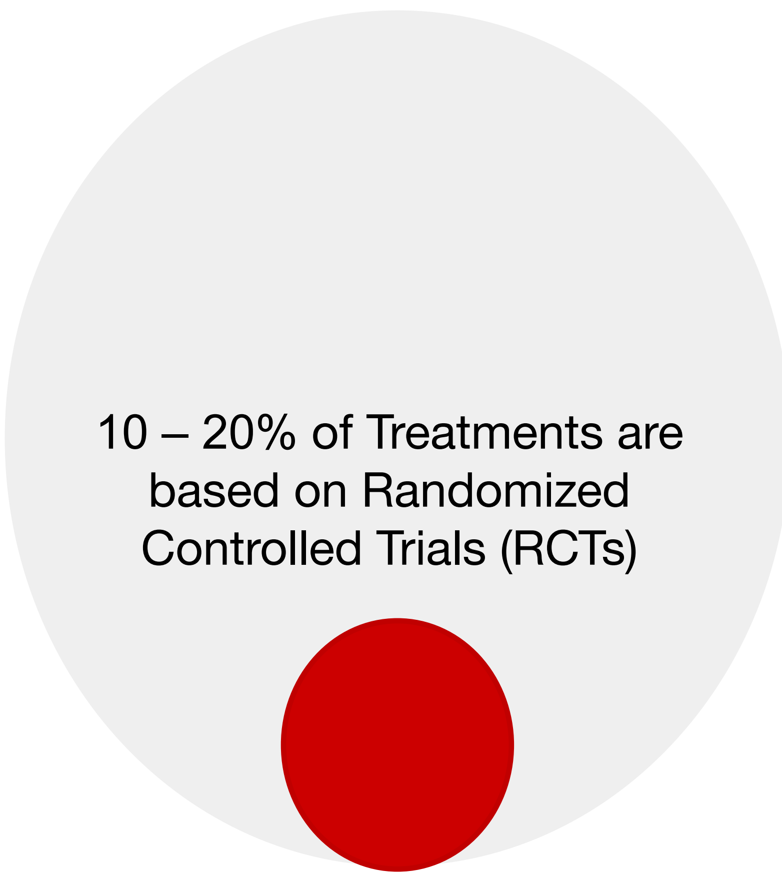


Evidence in Healthcare and Health?

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..*

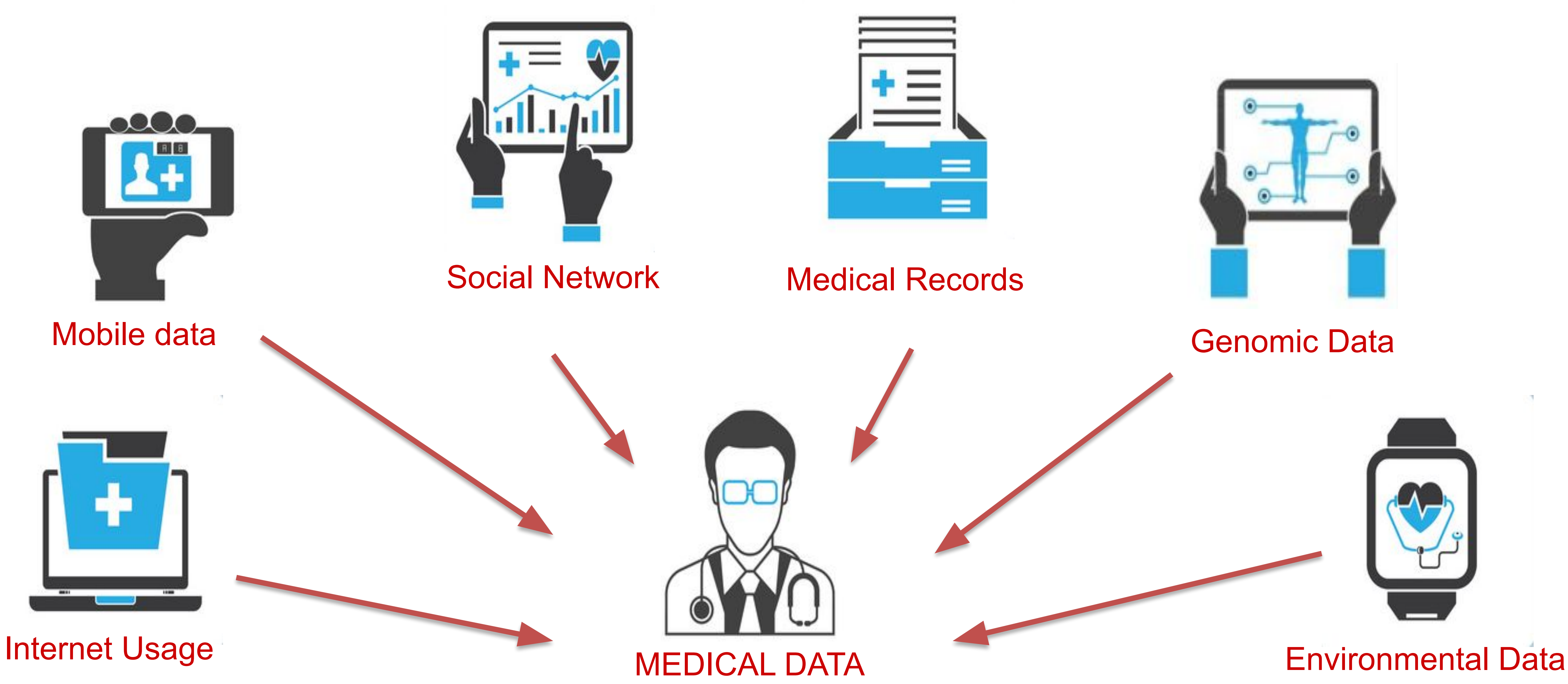
Evidence in Healthcare and Health?

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**?



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.

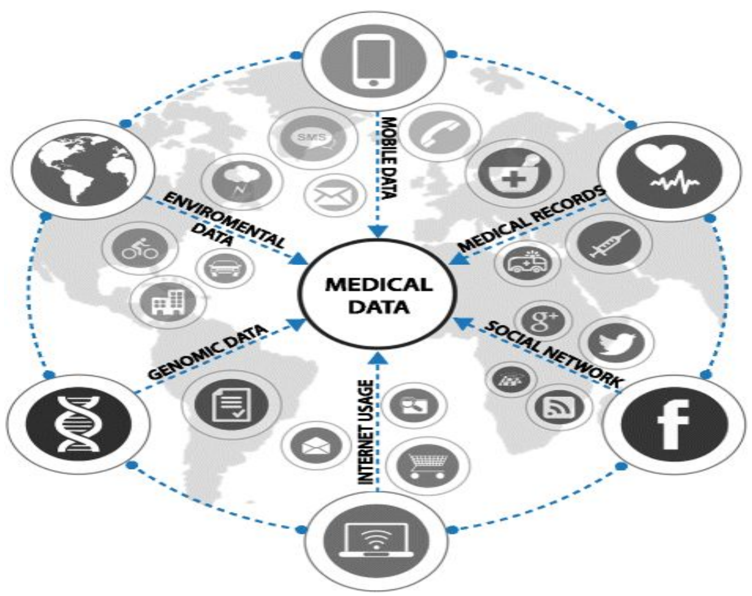
Heterogenous
Sampling
Data Type
Time Scale

Sparse
Unmeasured
Unreported
No Follow-up

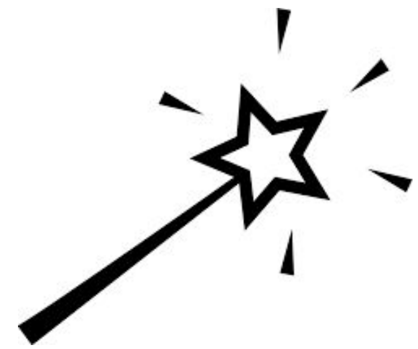
Uncertainty
Labels
Bias
Context

Lack of Expertise Is Challenging

- Media can create unrealistic expectations.



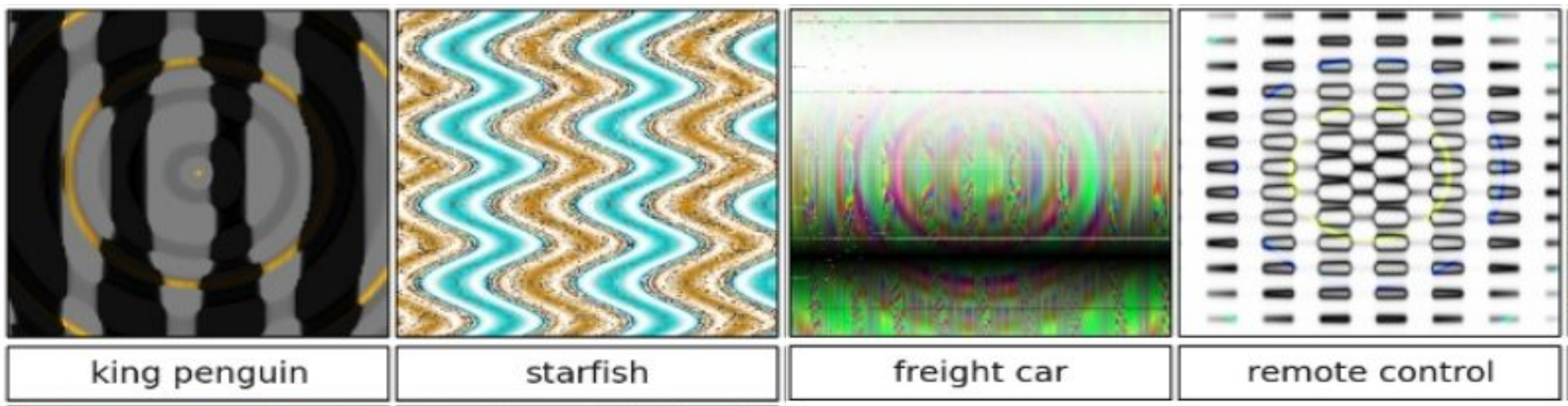
+



≠ Insight

Be Careful What You Optimize For

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



or

[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).

Natural Born Expertise Makes This Easier

- Humans are “natural” experts in NLP, ASR, Vision evaluation.[1]

(a) Husky classified as wolf (b) Explanation

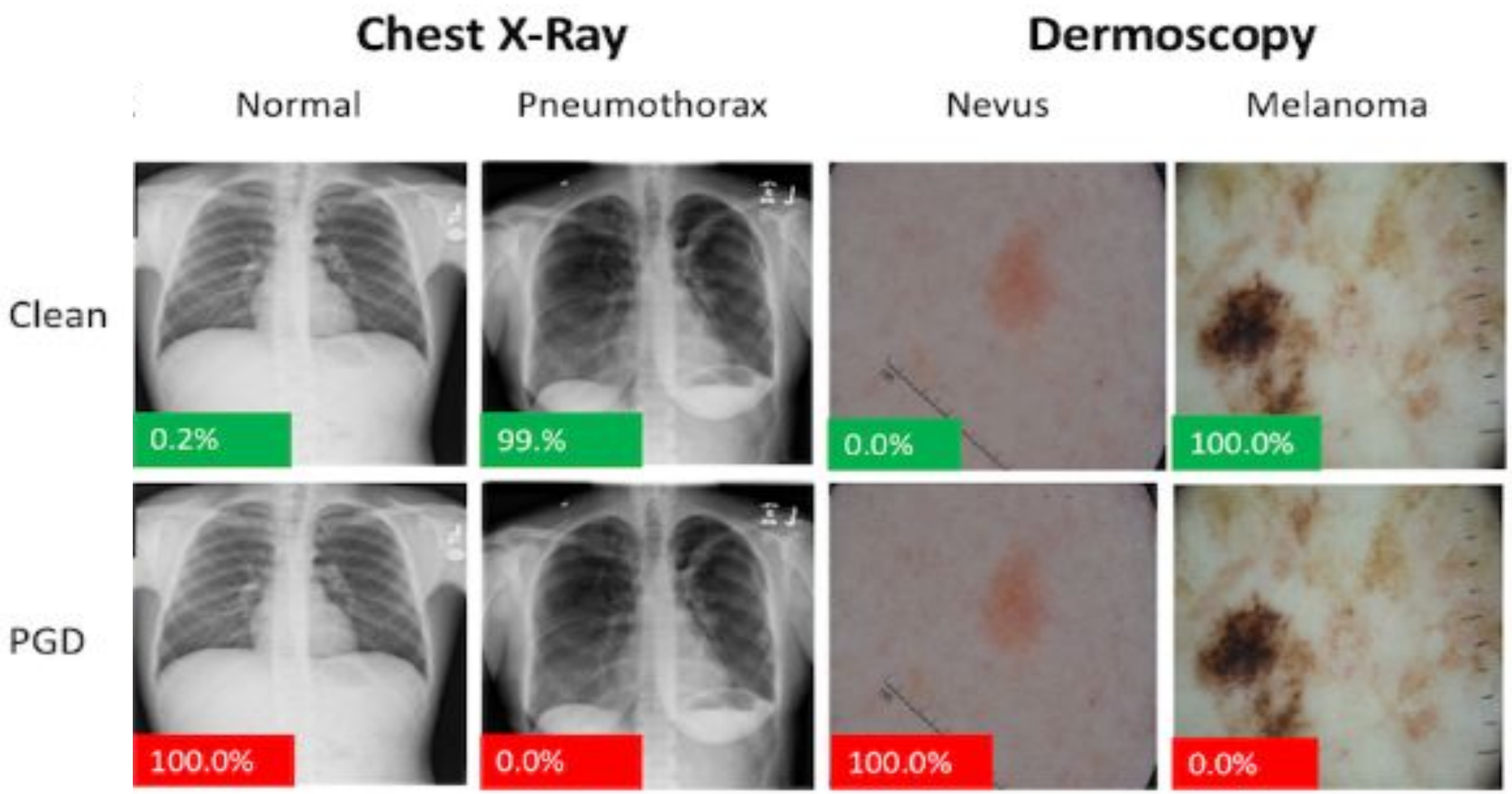
[1] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

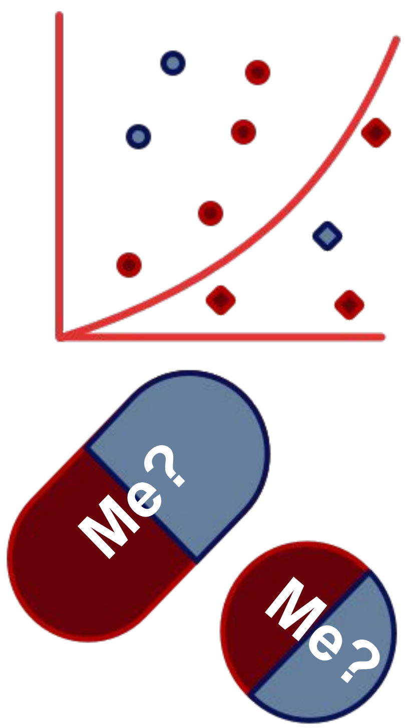
How Do We Know When We're Wrong?

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

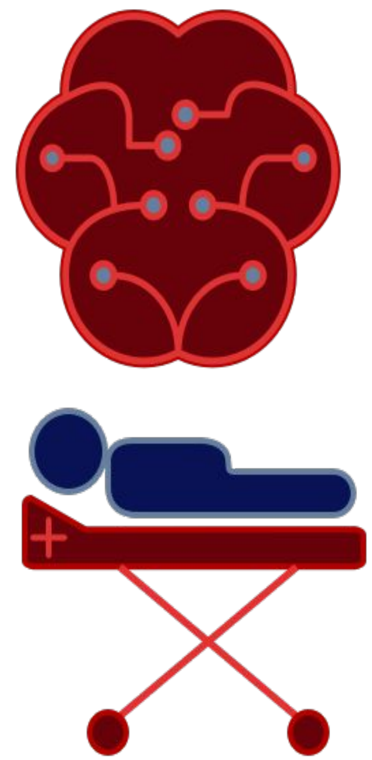


[1] Finlayson, Samuel G., Isaac S. Kohane, and Andrew L. Beam. "Adversarial Attacks Against Medical Deep Learning Systems." *arXiv preprint arXiv:1804.05296* (2018).

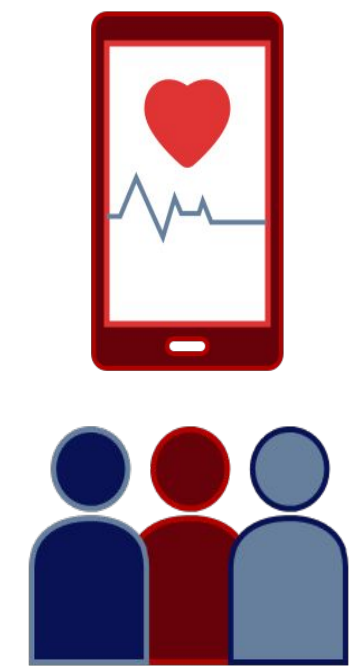
Machine Learning For Health (ML4H)



What models are healthy?

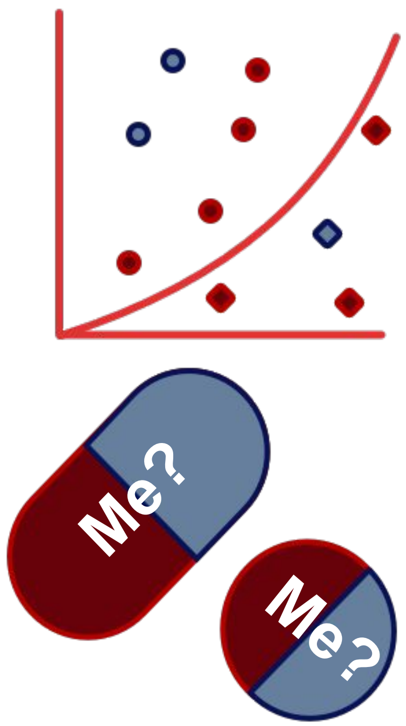


What healthcare is healthy?

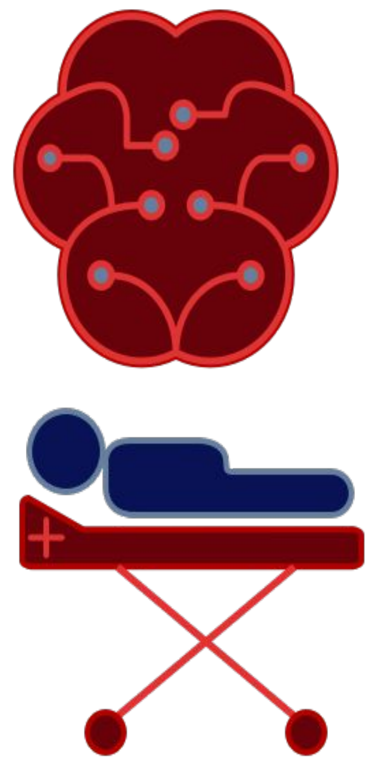


What behaviors are healthy?

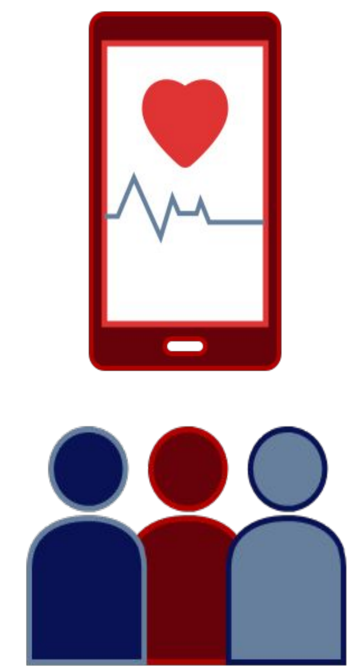
Machine Learning For Health (ML4H)



What models are healthy?



What healthcare is healthy?

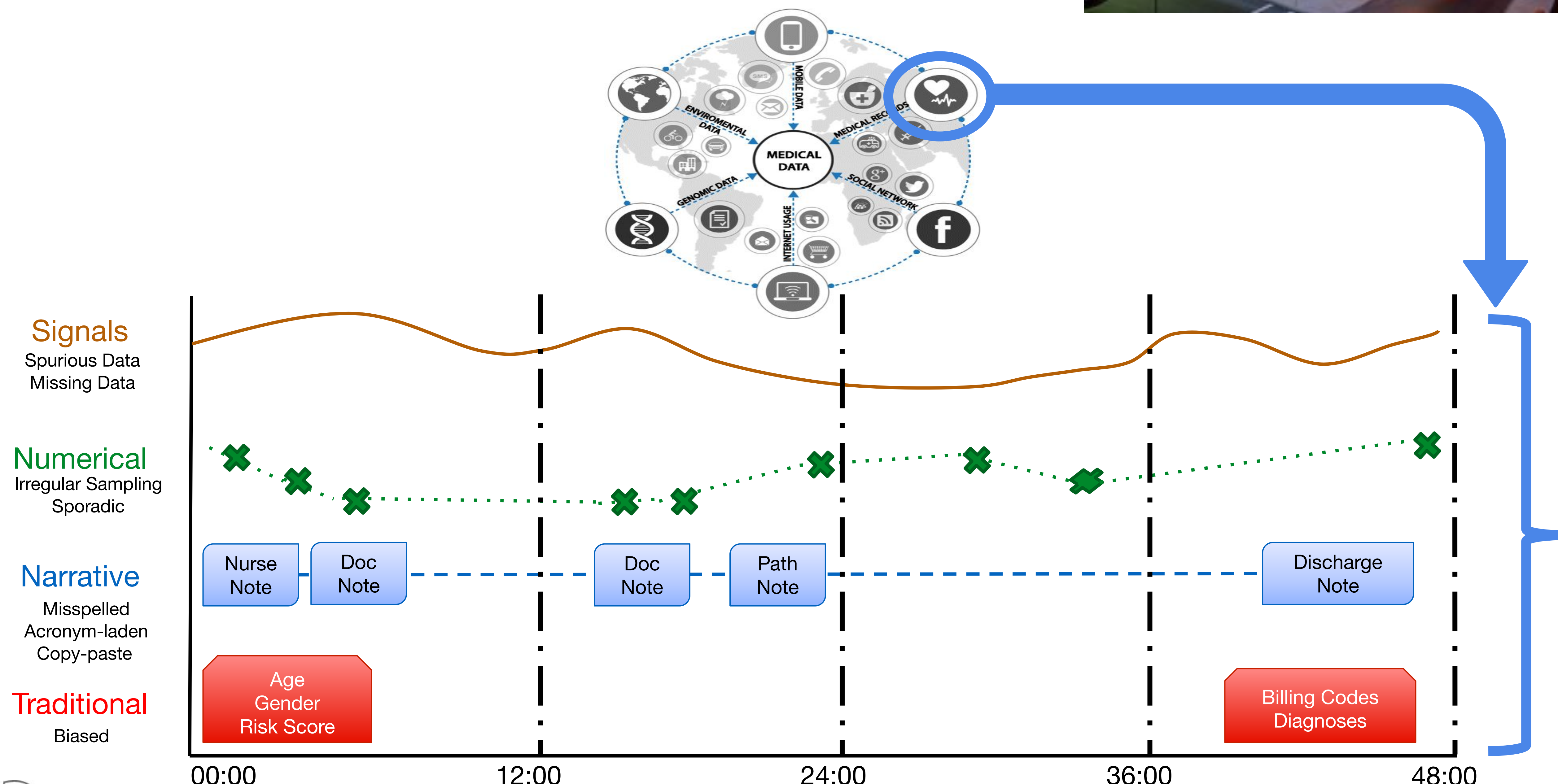


What behaviors are healthy?

17

MIMIC III ICU Data

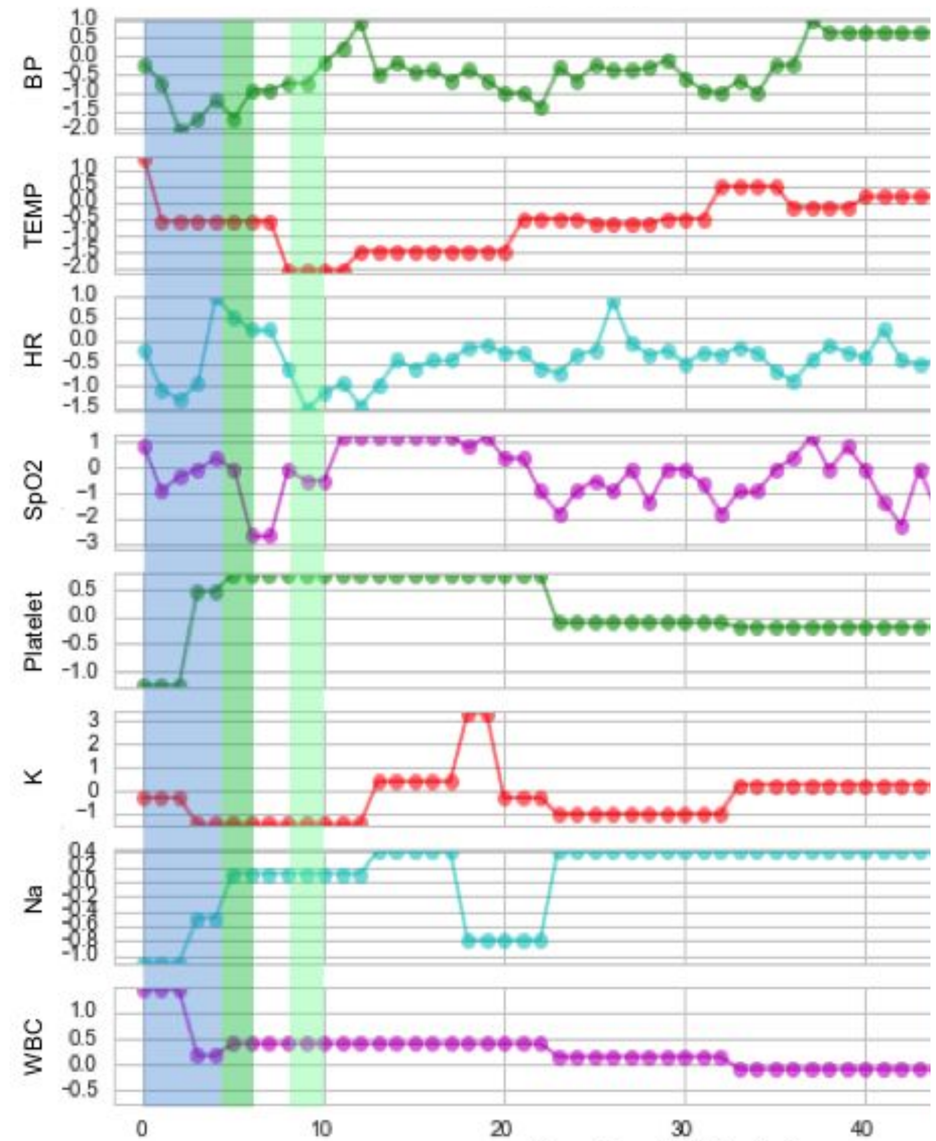
- Learning with real patient data from the Beth Israel Deaconess Medical Center ICU. [1]



[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



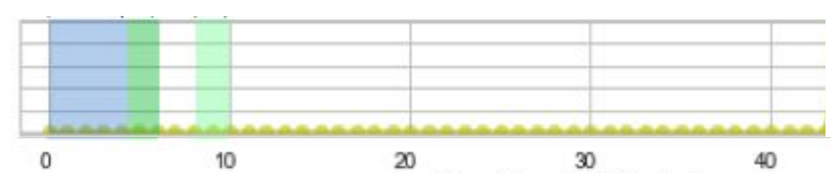
?

"Real-time" **Prediction**

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

By Gap Time

Before the Doctor Acted[1,2,3,4,5,6]



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

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Machine Learning For Health (ML4H)

Predict something **important** in **healthcare**.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

20

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.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

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.

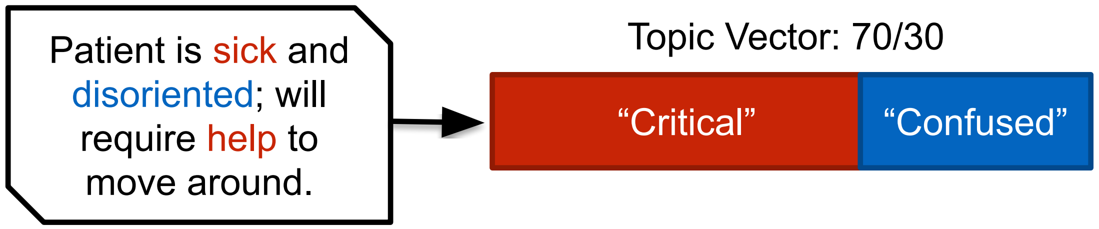
CONTEXT MATTERS

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.

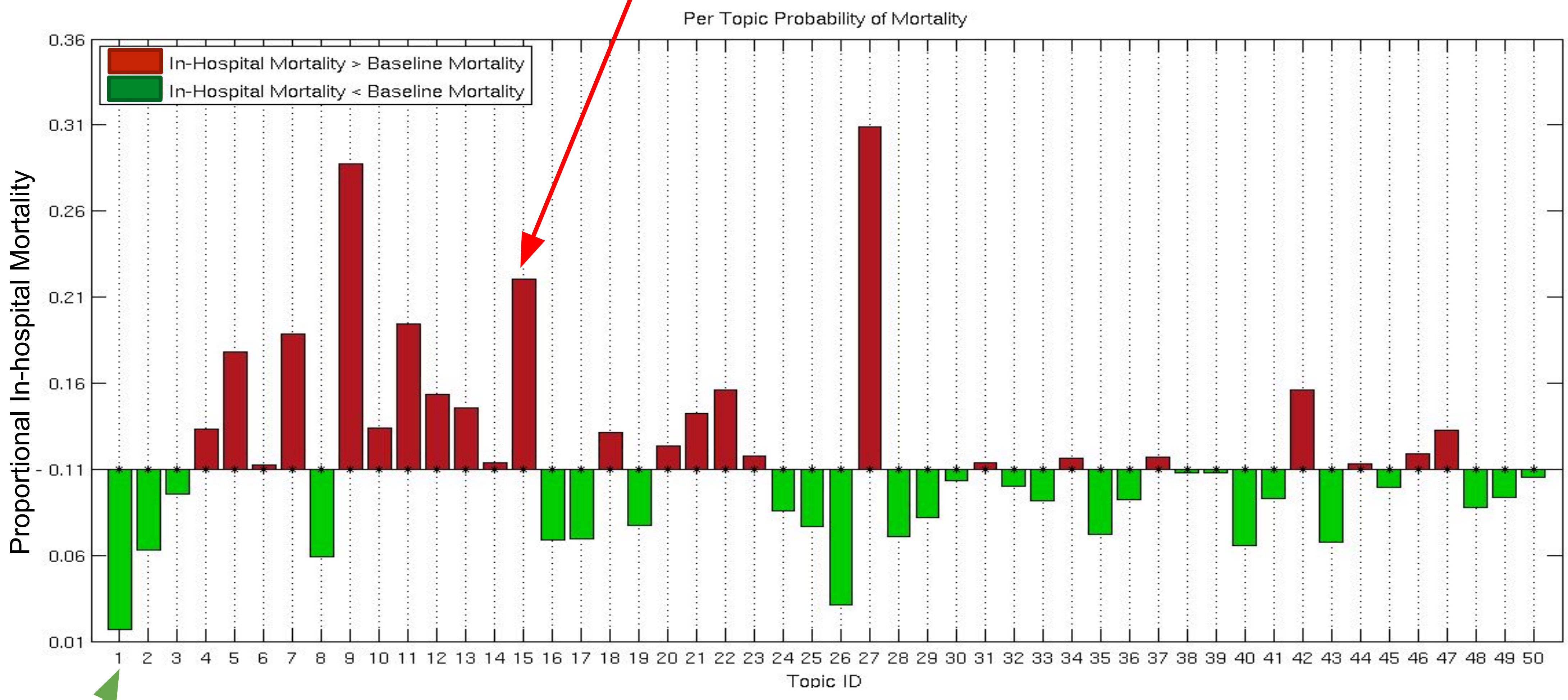


- Use Latent Dirichlet Allocation (LDA)[1] as an **unsupervised** way to **abstract** 473,000 notes from 19,000 patients into “topics”.[2]

[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

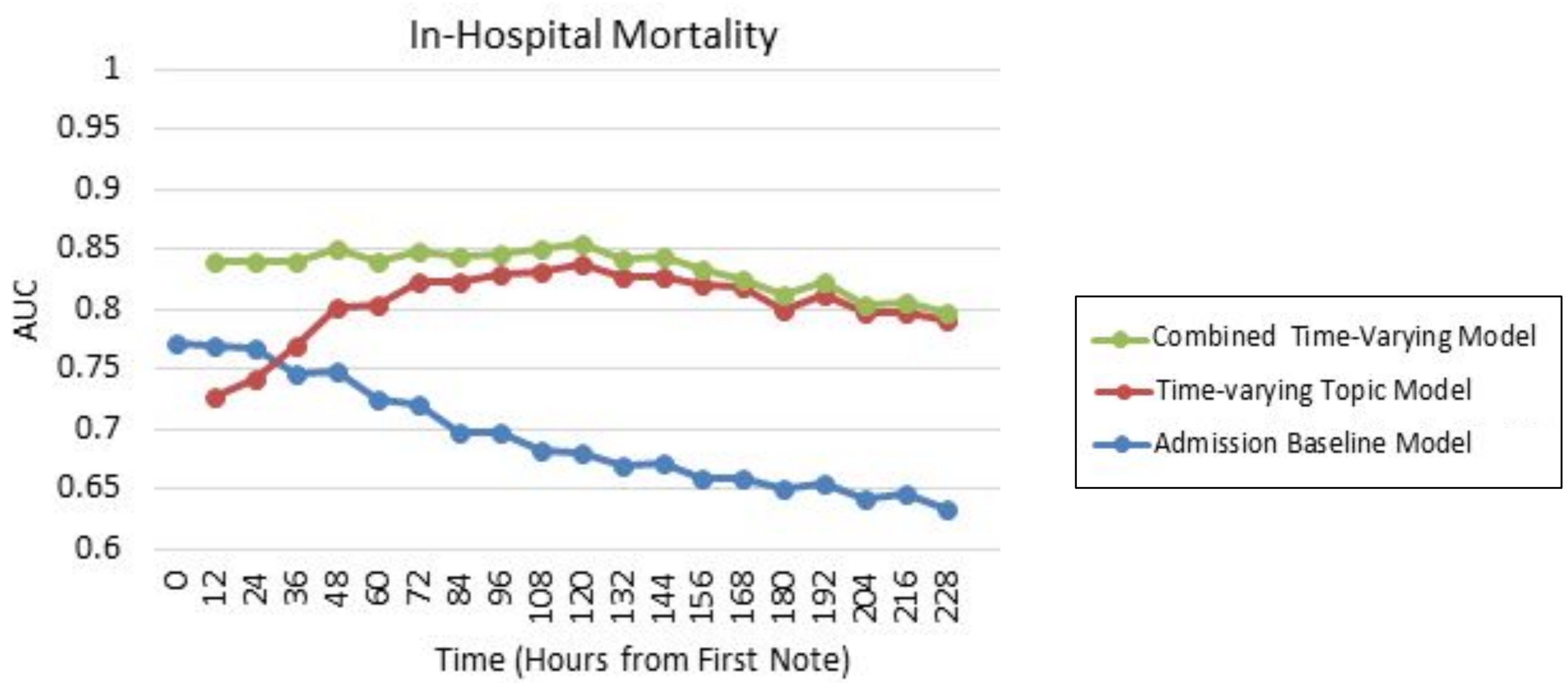
Correlation Between Average Topic Representation and Mortality

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



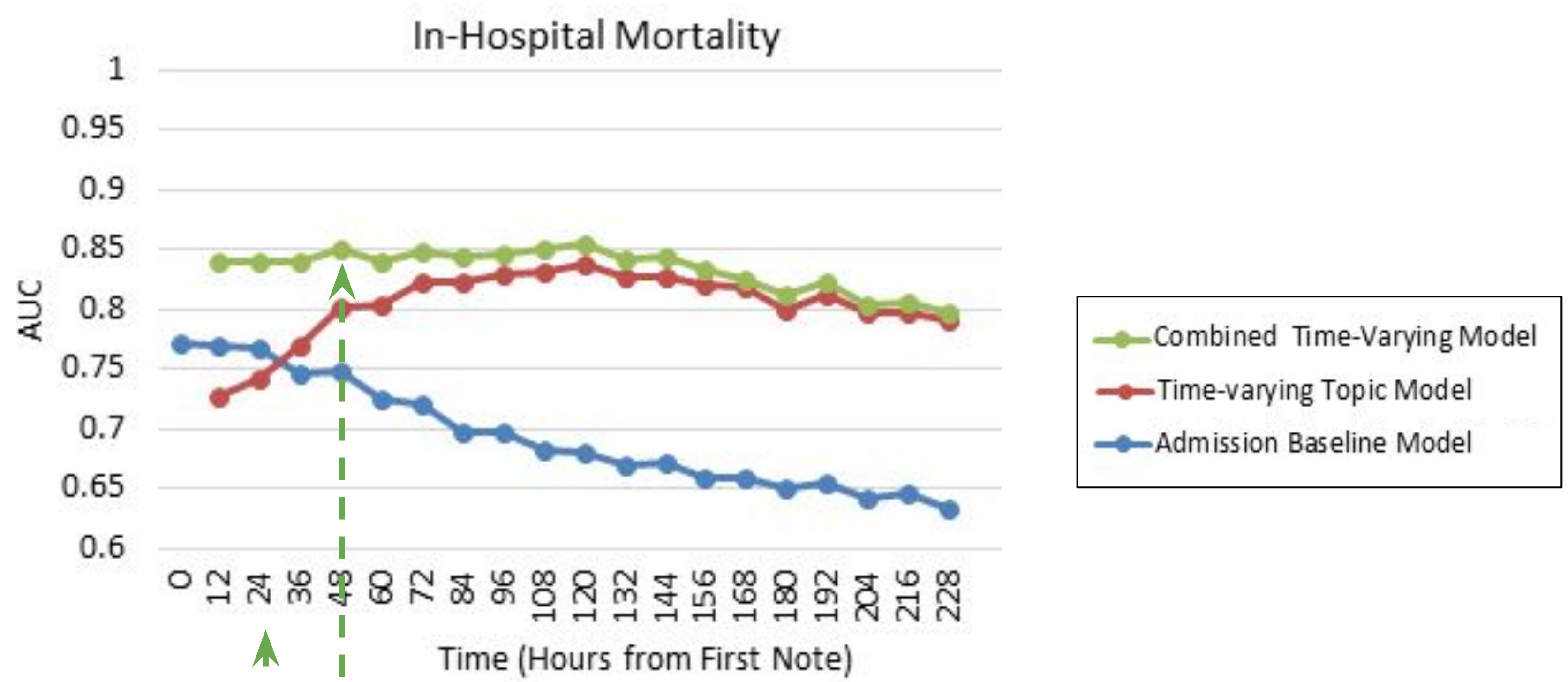
Topic #	Top Ten Words	Possible Topic
1	cabg, pain, ct, artery, coronary, valve, post, wires, chest, sp	Cardiovascular surgery

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**.

But Wait! More Complex Models Haven't Done Better…



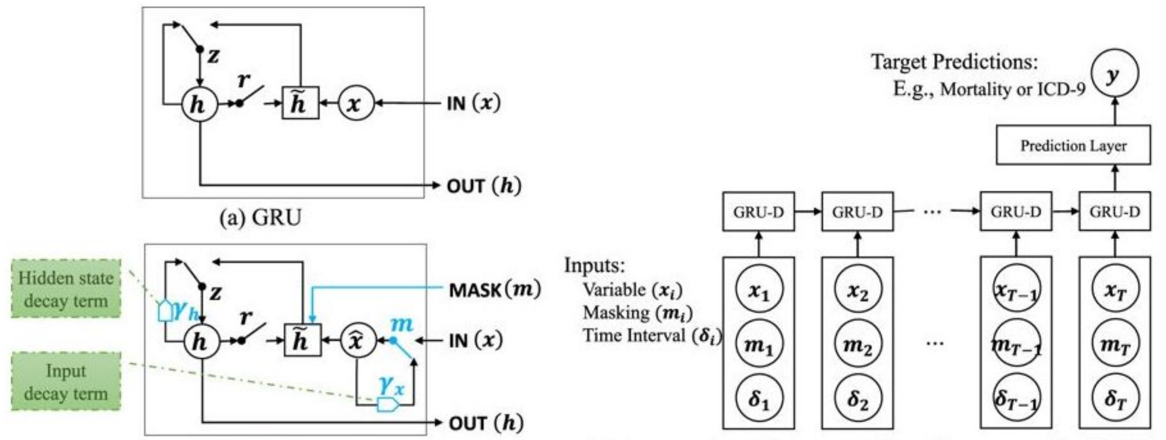
Author	AUC	Method	Episodes	Hours	Variables
Ghassemi, 2014	0.84/**0.85**	LDA	19,308	24/48	53 - notes
Caballero, 2015	0.86	Text processing + medication	15,000	24	? - notes/meds
Che, 2015	0.8-0.82	Deep Learning (LSTM)	3,940	48	30 - vitals
Che, 2016	0.7/0.85	Deep Learning (GRU)	19,714	12/48	99 – vitals/meds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

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**

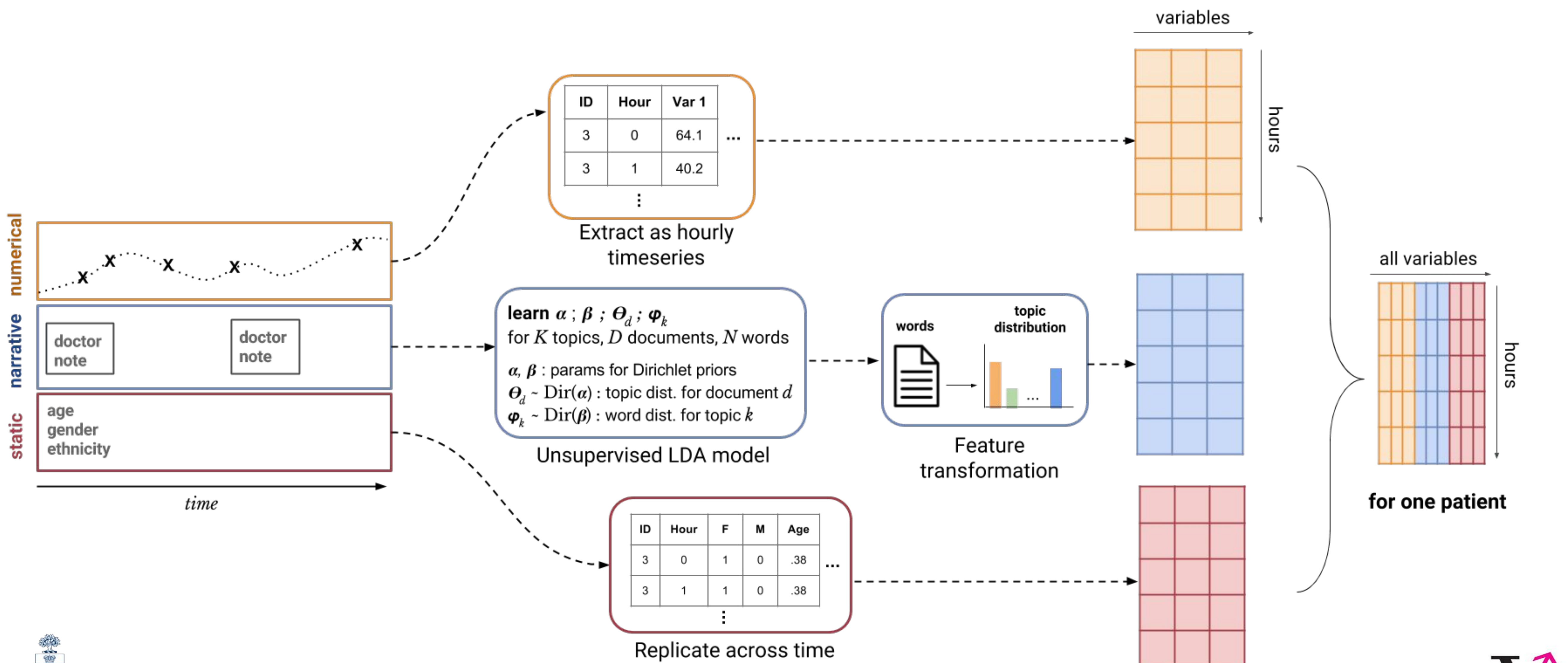
Machine Learning For Health (ML4H)

actionable

Predict something ~~important~~ in **healthcare**.

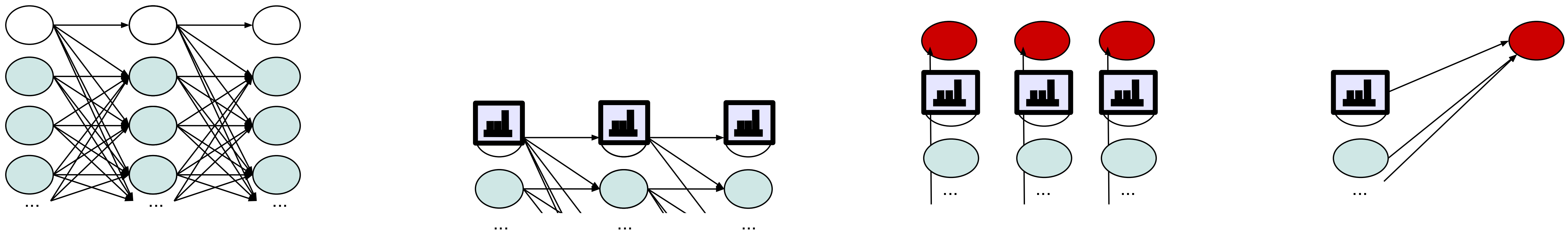
Part 2: Predict **Interventions** With Clinical **Data**

- 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



Many Ways to Model, What Do We Learn?

SSAM



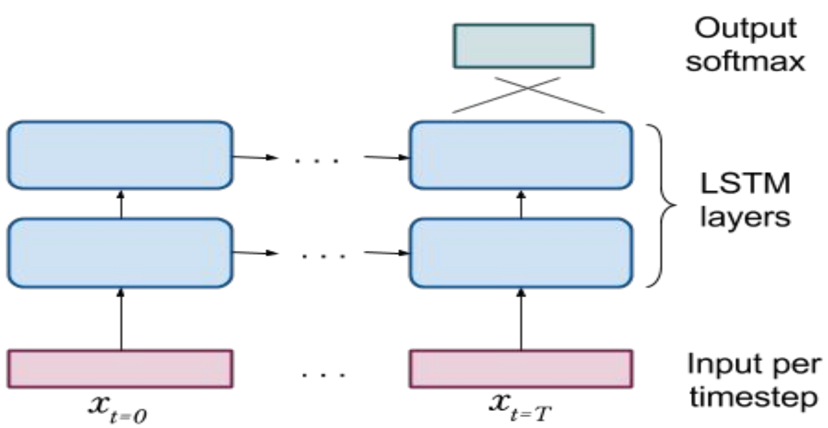
Learn model parameters over patients with variational EM.

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

Logistic regression (with label-balanced cost function)

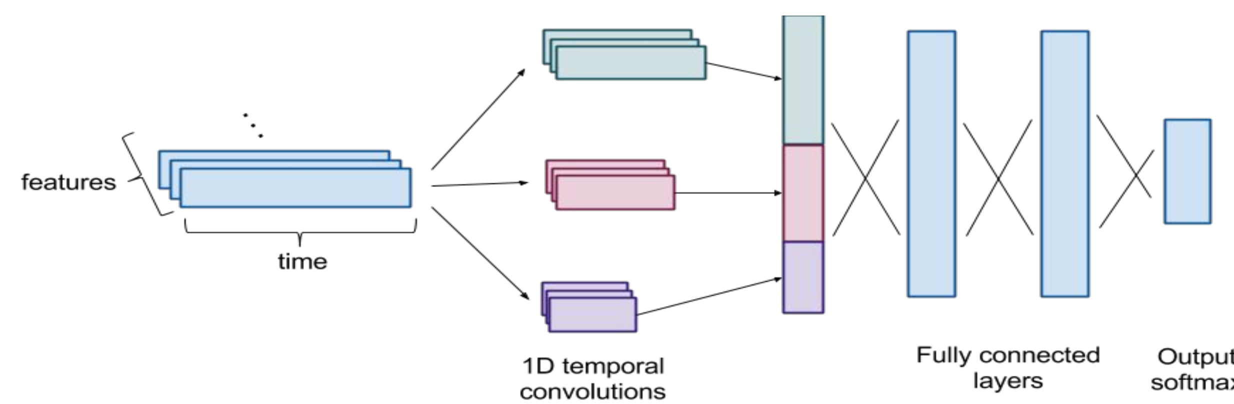
Predict onset in advance

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.

NNs Do Well; Improved Representation Helps

Area-under-ROC

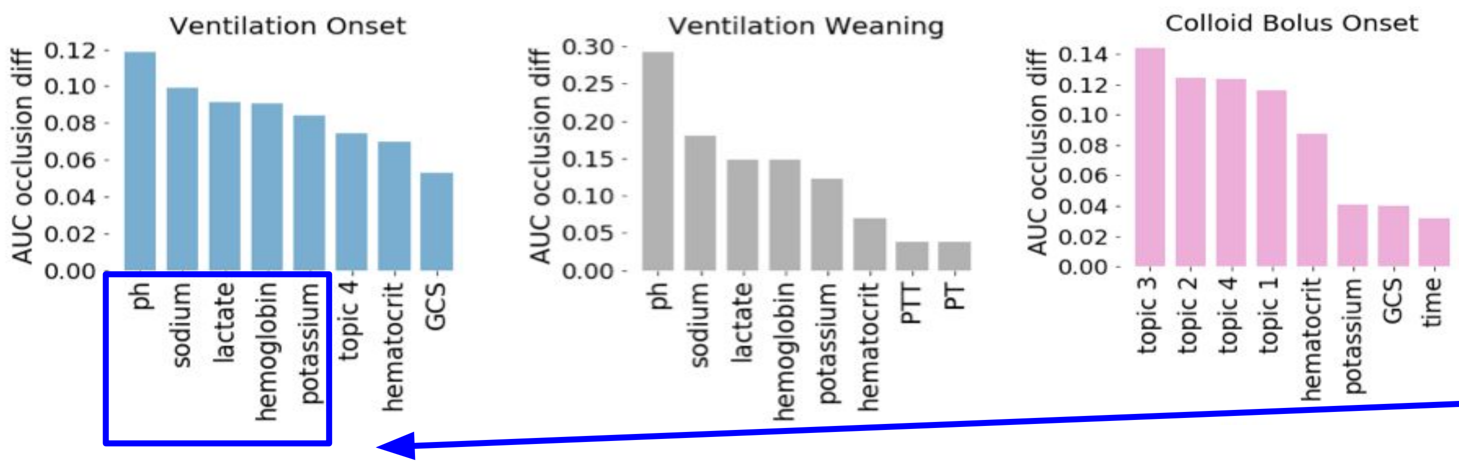
Task	Model	Intervention Type: VENT	NI-VENT	VASO	COL BOL	CRYS BOL
Onset AUC	Baseline	0.60	0.66	0.43	0.65	0.67
	LSTM Raw	0.61	0.75	**0.77**	0.52	0.70
	LSTM Words	**0.75**	**0.76**	0.76	**0.72**	**0.71**
	CNN	0.62	0.73	**0.77**	0.70	0.69
Wean AUC	Baseline	0.83	0.71	0.74	-	-
	LSTM Raw	0.90	0.80	**0.91**	-	-
	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.

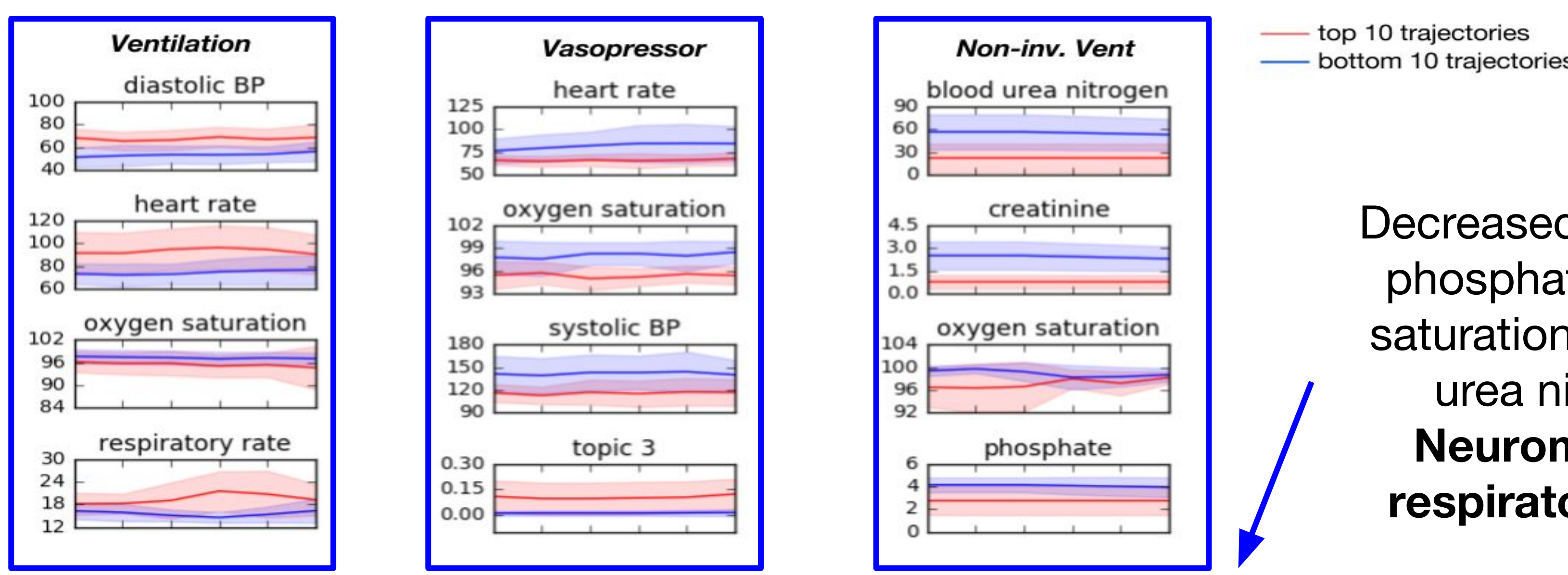
NN Post-hoc Interpretability

- 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**

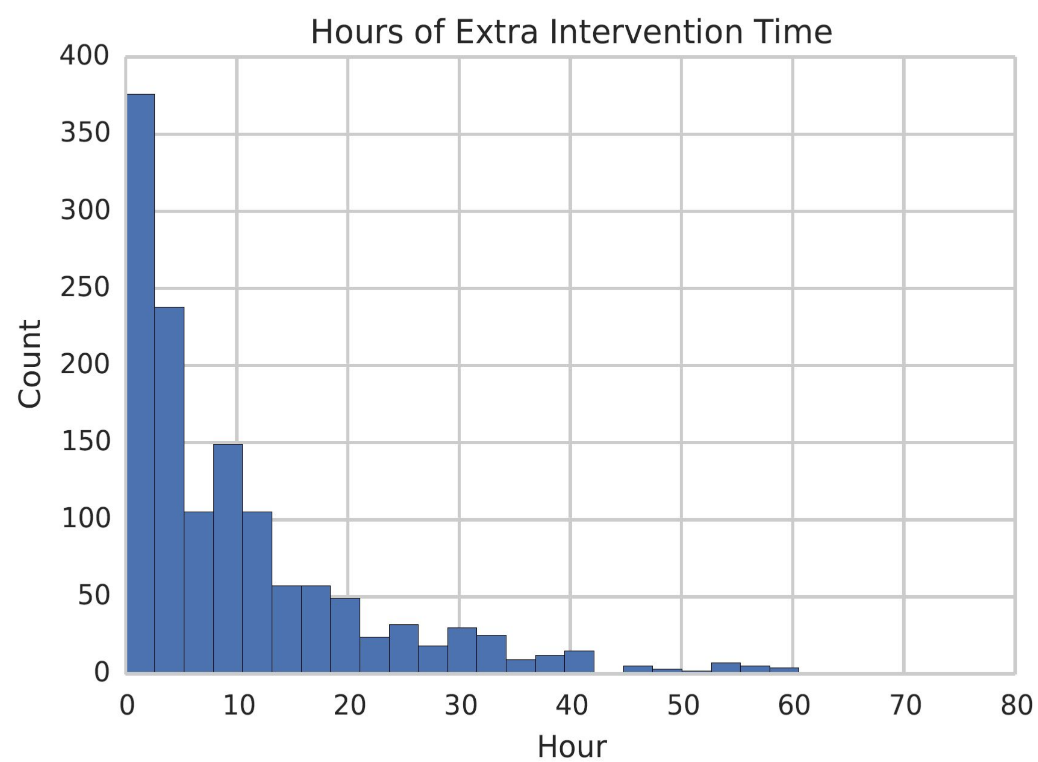
Decreased systolic blood pressure, heart rate and oxygen saturation rate : **Altered peripheral perfusion** or **stress hyperglycemia**

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

From Healthcare to Health

- Patients can be left on interventions longer than necessary.

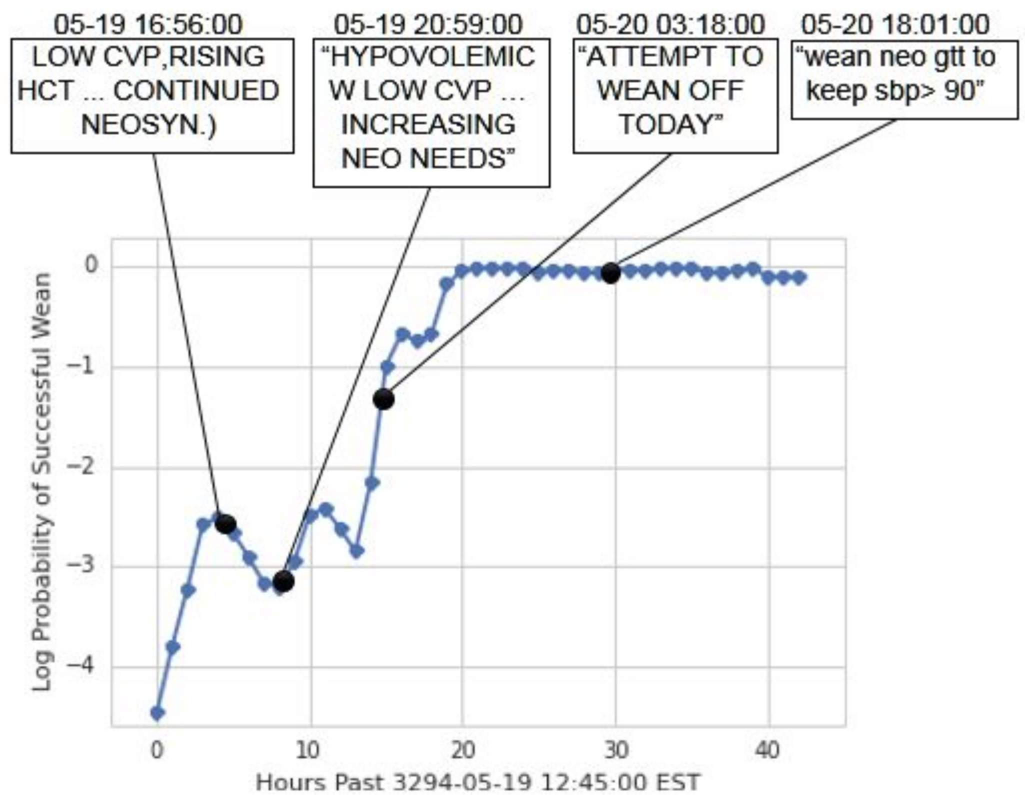


- Extended interventions can be costly and detrimental to patient health.[1,2]

[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.

33

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!

Machine Learning For Health (ML4H)

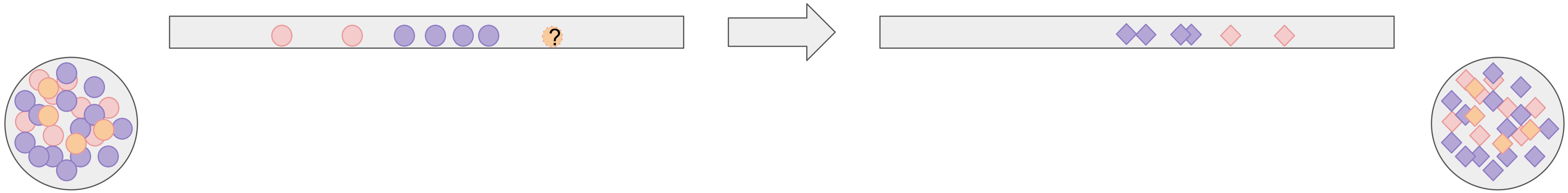
actionable insights

Predict ~~something~~ ~~**important**~~ in **healthcare**.

35

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

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

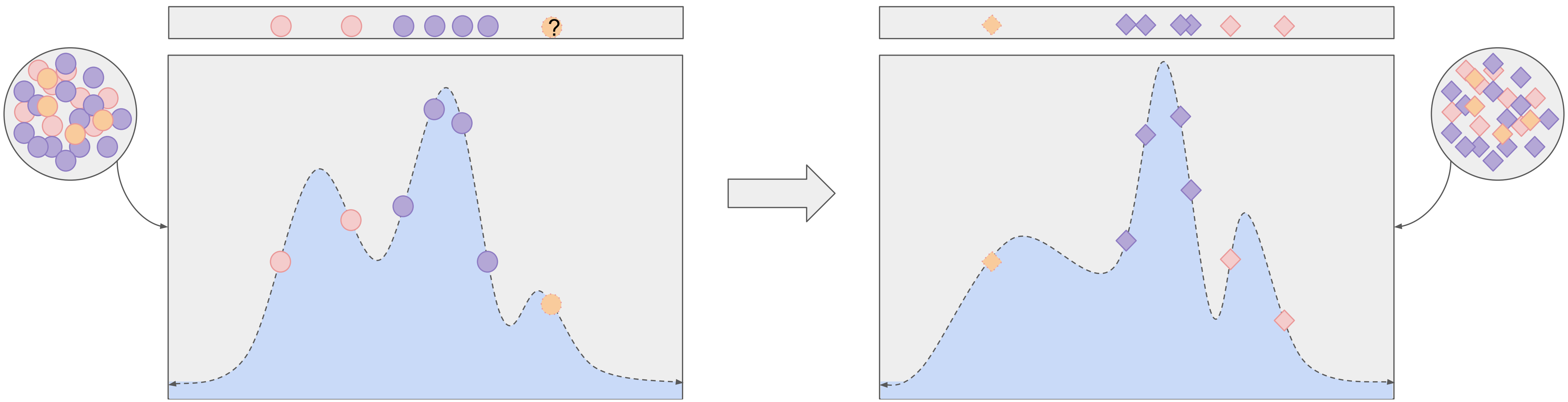


- Sufficiently large, heterogeneous paired datasets are rare.

36

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.

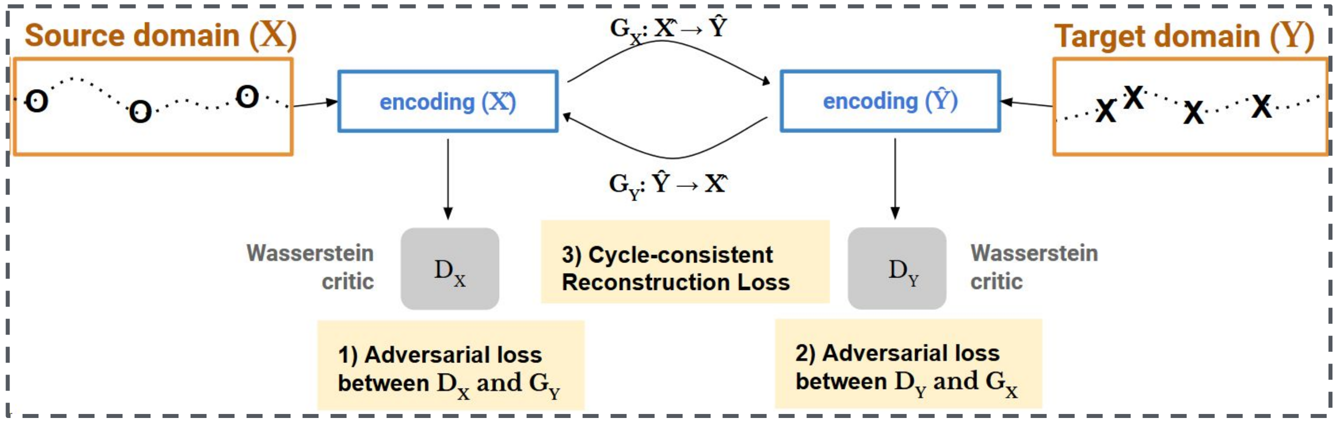
UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

37

Model Learns on Unpaired Data, G_X Used to Eval

- Generated samples are realistic
- Account for missing samples (not just missing features)
- Ensure cycle/self-consistency[1]



- Improved intervention response prediction
  - ~500 paired, ~3,000 unpaired patients

	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%

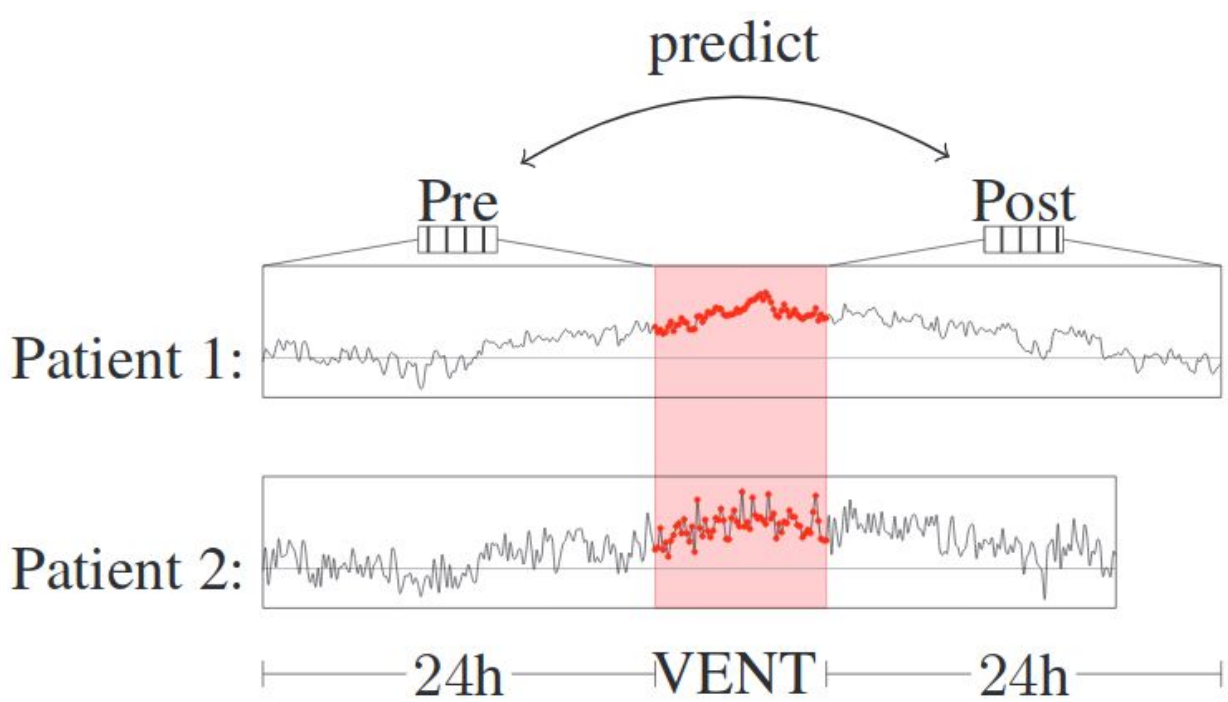
[1] Ghasedi Dizaji K, Wang X, Huang H. Semi-Supervised Generative Adversarial Network for Gene Expression Inference. InProceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining 2018 Jul 19 (pp. 1435-1444). ACM.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

38

Deploy Good Models To Forecast Response?



- Exciting work on to be done on learning what treatments are best for individuals based on environment and context!
- But there are other factors...

39

Machine Learning For Health (ML4H)

Create actionable insights

~~**Predict** something **important**~~ in **healthcare**.

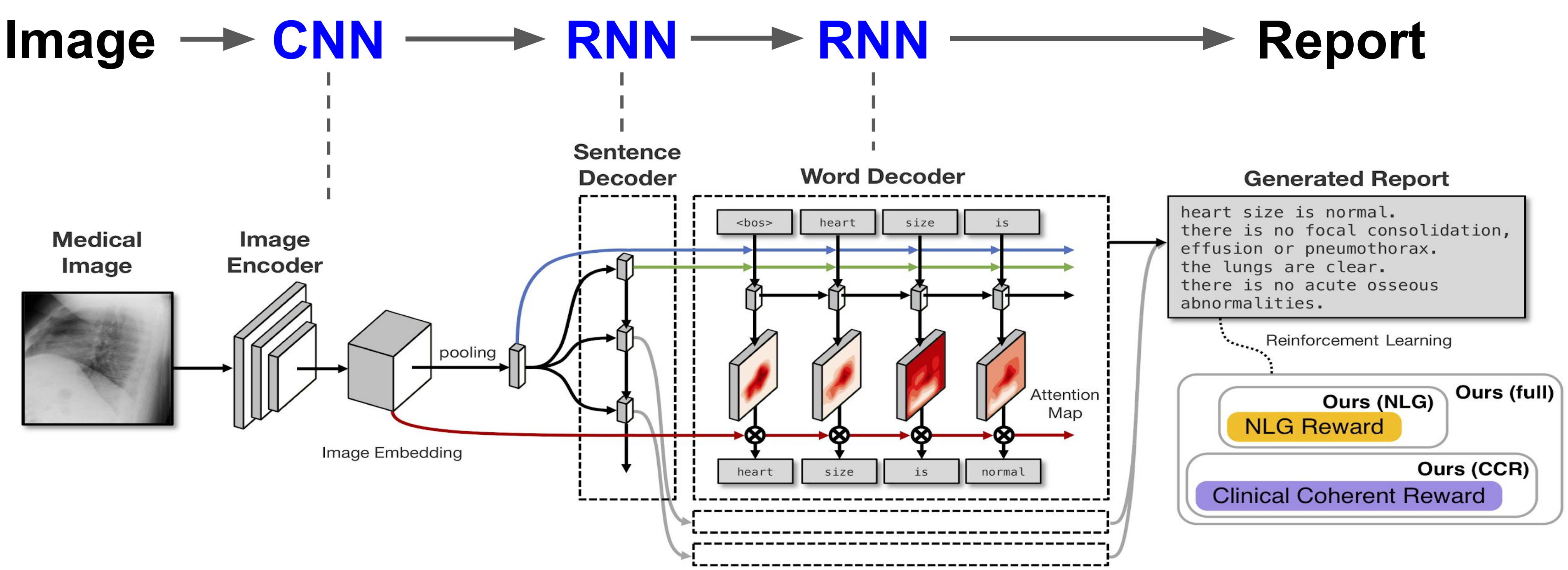
UNIVERSITY OF TORONTO

40

VECTOR INSTITUTE | INSTITUT VECTEUR

Part 4: Create **Reports** From Clinical **Images**

- Automatically **generate** radiology **reports** given **chest X-Rays**.
    - First predict **topics** in the report.
    - **Conditionally** generate **sentences** corresponding to topics.



- CNN-RNN-RNN structure gives model the ability to **use largely templated sentences** and **generate diverse text.**

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Evaluating Readability and Clinical Coherence

- Outperforms state-of-the-art methods in **readability** and **accuracy**.

Quantitative Results

Maintain high language fluency

	Model	Natural Language						Clinical
		CIDEr	ROUGE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Accuracy
MIMIC-CXR	*Major Class*	-	-	-	-	-	-	0.828
	Noise-RNN	0.716	0.272	0.269	0.172	0.113	0.074	0.803
	1-NN	0.755	0.244	0.305	0.171	0.098	0.057	0.818
	S&T	0.886	0.300	0.307	0.201	0.137	0.093	0.837
	SA&T	0.967	0.288	0.318	0.205	0.137	0.093	0.849
	TieNet	1.004	0.296	0.332	0.212	0.142	0.095	0.848
	Ours (NLG)	**1.153**	**0.307**	**0.352**	**0.223**	**0.153**	**0.104**	0.834
	Ours (CCR)	0.956	0.284	0.294	0.190	0.134	0.094	**0.868**
	Ours (full)	1.046	**0.306**	0.313	0.206	0.146	**0.103**	**0.867**

CCR generates higher accuracy

Qualitative Check

Unseen Image → **Generated Text** → **Actual Text**

Generated Text: as compared to the previous radiograph, there is no relevant change. tracheostomy tube is in place. there is a layering pleural effusions. NAME bilateral pleural effusion and compressive atelectasis at the right base. there is no pneumothorax.

Actual Text: as compared to the previous radiograph, the monitoring and support devices are unchanged. unchanged bilateral pleural effusions, with a tendency to increase, and resultant areas of atelectasis. the air collection in the bilateral soft tissues is slightly decreased. unchanged right picc line. no definite evidence of pneumothorax.

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

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?

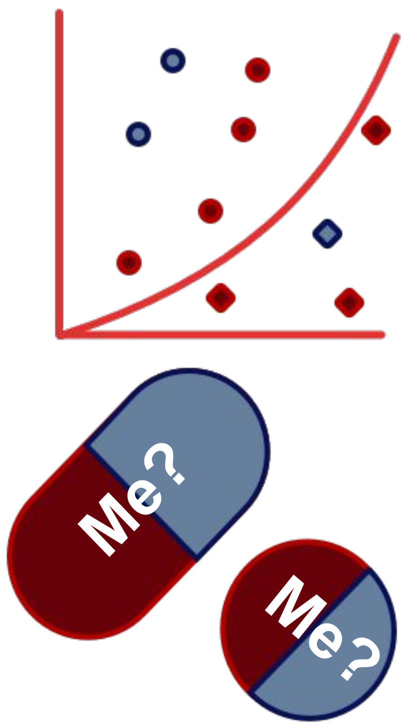
09 What makes algorithms, and the entities that create them, trustworthy?

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

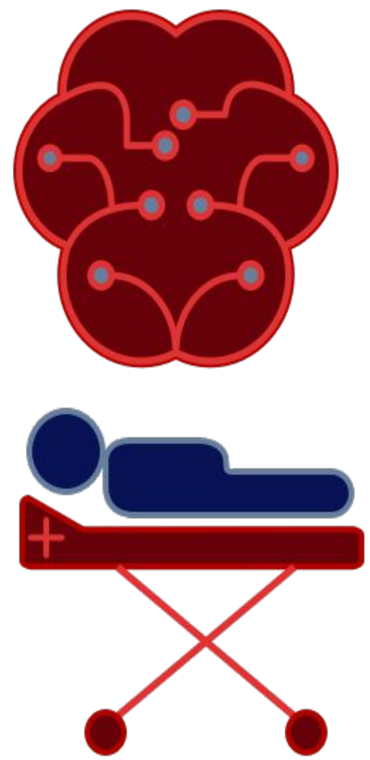
[1] “Ethical, social, and political challenges of artificial intelligence in Health”. *Wellcome Trust* April 2018. © Future Advocacy.

43

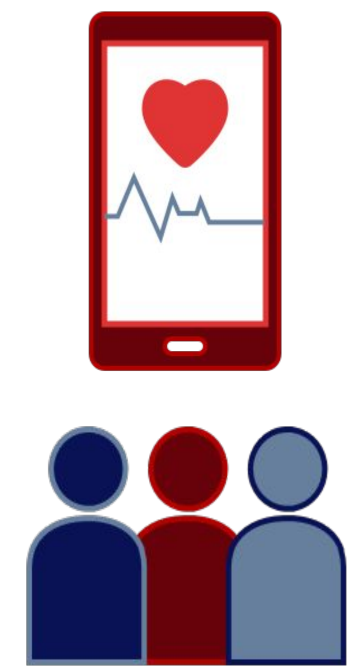
Machine Learning For Health (ML4H)



What **models** are **healthy**?



What **healthcare** is **healthy**?



What **behaviors** are **healthy**?

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

44

Bias Is Part of the Clinical Landscape Already

- How does/should ML interact with fairness/health[1,2,3,4,5]?

Viewpoint

August 11, 2015

Racial Bias in Health Care and Health
Challenges and Opportunities

David R. Williams, PhD, MPH[1,2]; Ronald Wyatt, MD, MHA[3]

Author Affiliations

JAMA. 2015;314(6):555-556. doi:10.1001/jama.2015.9260

J Palliat Med. 2013 Nov; 16(11): 1329–1334.
doi: 10.1089/jpm.2013.9468
PMCID: PMC3822363
PMID: 24073685

Racial and Ethnic Disparities in Palliative Care

Kimberly S. Johnson, MD, MHS[1,2]

Author information ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

(2)

The Girl Who Cried Pain: A Bias Against Women in the Treatment of Pain

Diane E. Hoffmann and Anita J. Tarzian

Am J Public Health. 2007 February; 97(2): 247–251.
doi: 10.2105/AJPH.2005.072975
PMCID: PMC1781382
PMID: 17194867

The Black–White Disparity in Pregnancy-Related Mortality From 5 Conditions: Differences in Prevalence and Case-Fatality Rates

Myra J. Tucker, BSN, MPH, Cynthia J. Berg, MD, MPH, William M. Callaghan, MD, MPH, and Jason Hsia, PhD

Author information ▶ Article notes ▶ Copyright and License information ▶ Disclaimer

Obes Rev. 2015 Apr;16(4):319-26. doi: 10.1111/obr.12266. Epub 2015 Mar 5.

Impact of weight bias and stigma on quality of care and outcomes for patients with obesity.

Phelan SM[1], Burgess DJ, Yeazel MW, Hellerstedt WL, Griffin JM, van Ryn M.

Author information

[1] Continuous State-Space Models for Optimal Sepsis Treatment - Deep Reinforcement Learning … (MLHC/JMLR 2017);
[2] Modeling Mistrust in End-of-Life Care (MLHC 2018/FATML 2018 Workshop);
[3] The Disparate Impacts of Medical and Mental Health with AI. (AMA Journal of Ethics 2019);
[4] ClinicalVis Project with Google Brain. (*In submission);

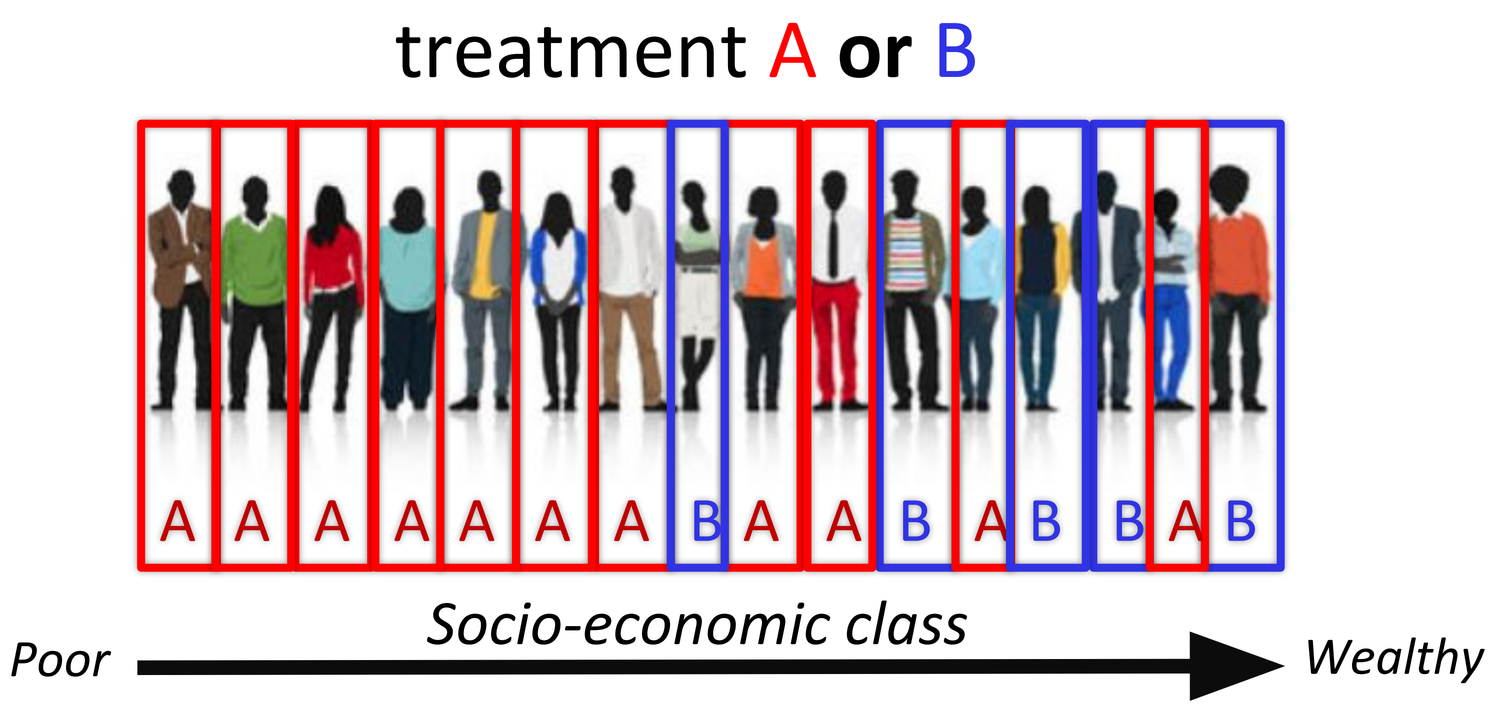
UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

45

How Can We Improve Health Care For **All**?

- Patient populations have differences in treatment by race, sex, and socioeconomic status



- Are there differences in prediction accuracy by group?

46

Machine Learning For Health (ML4H)

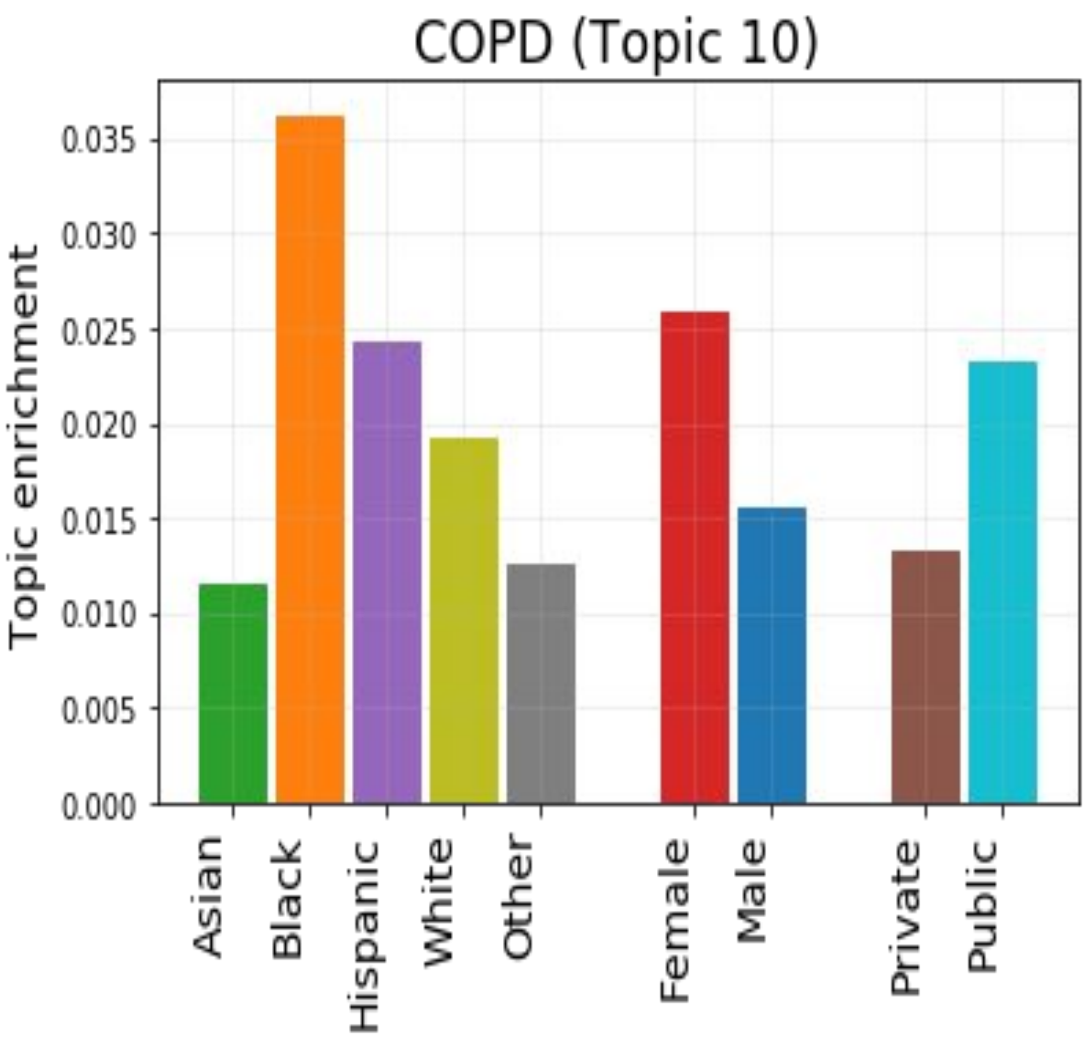
Create actionable insights in human health.

~~Predict something **important** in **healthcare**~~.

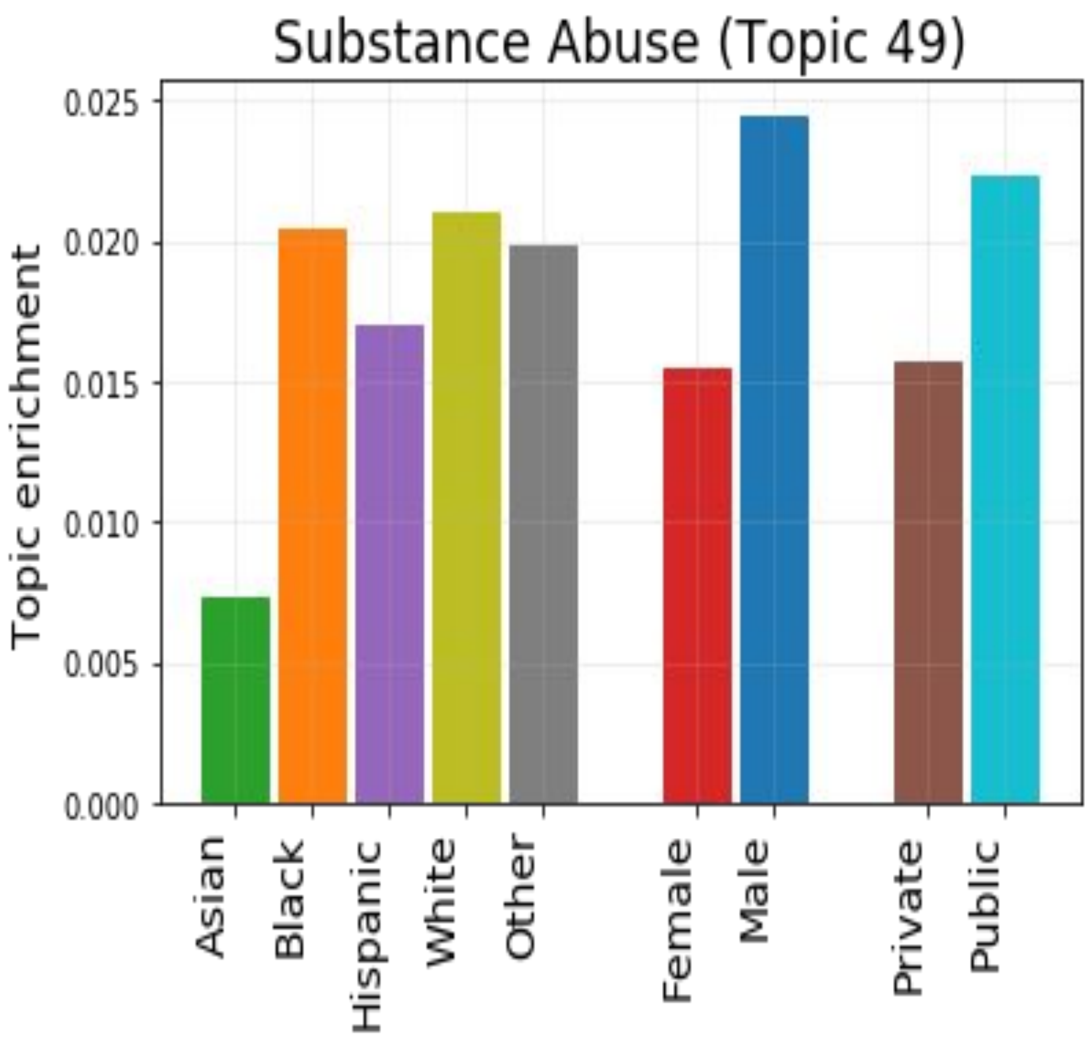
Topic Heterogeneity in Medical and Mental Health

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

Group-Specific ICU Topic 10



Group-Specific Psych Topic 49



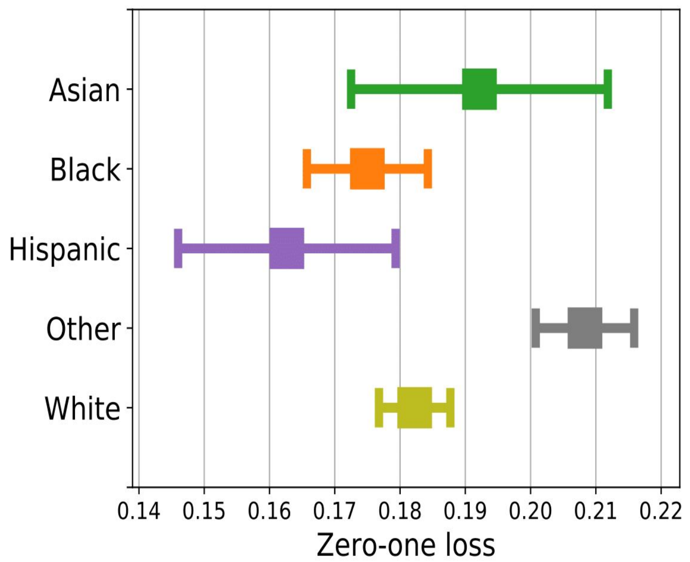
UNIVERSITY OF TORONTO

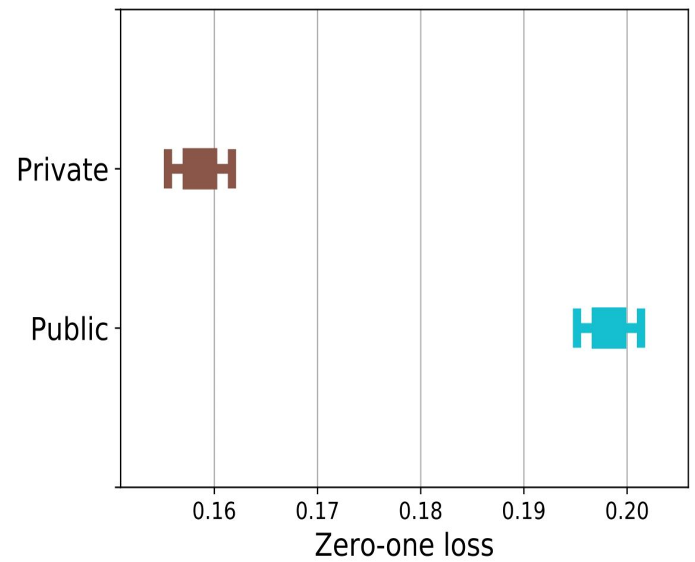
VECTOR INSTITUTE | INSTITUT VECTEUR

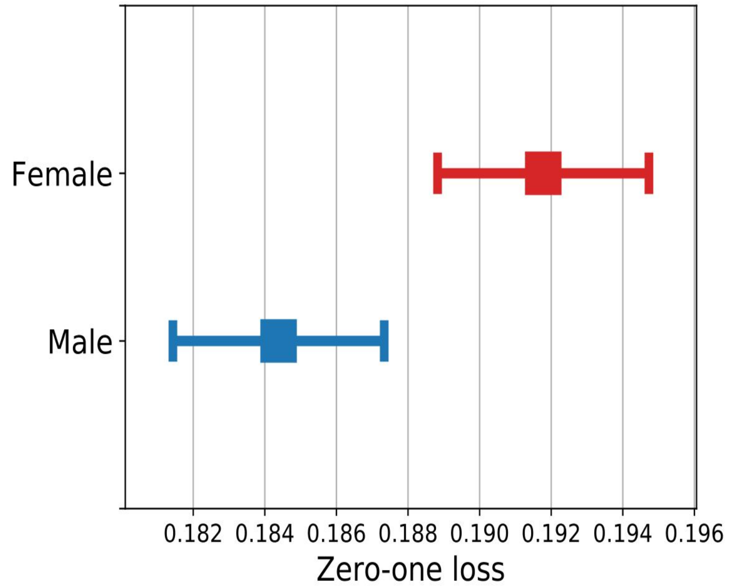
48

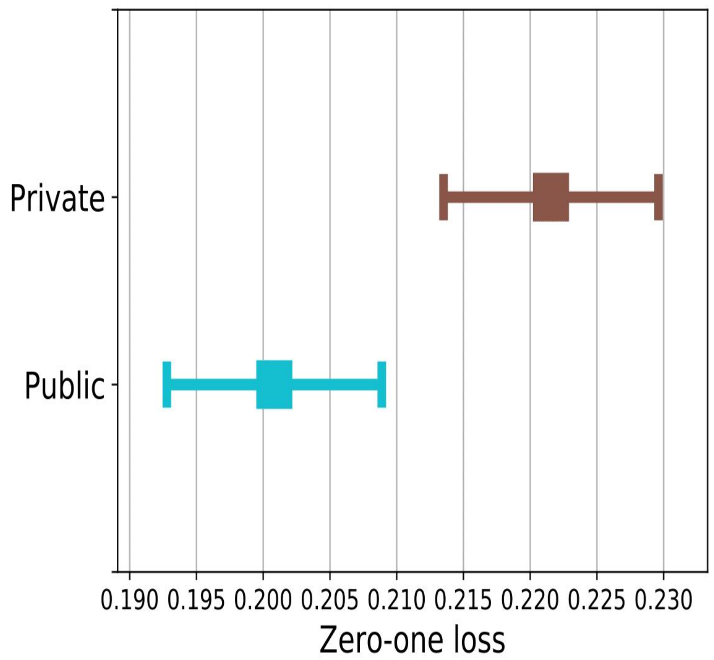
Unfair Accuracies in Medical and Mental Health

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









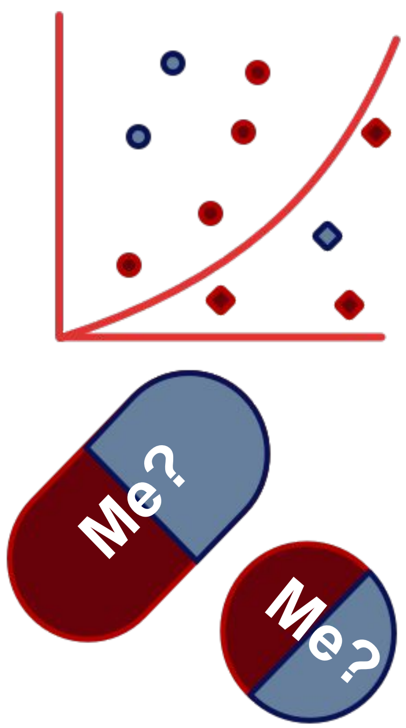
UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

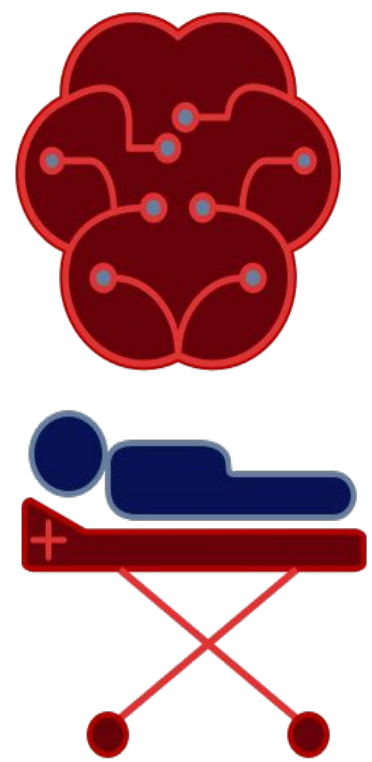
49

Machine Learning For Health (ML4H)

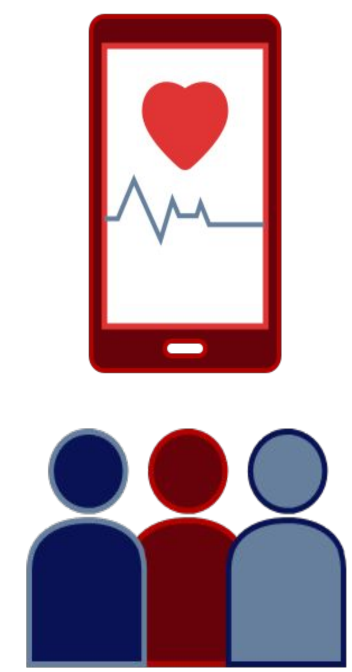
Creating actionable insights in human health.



What models are healthy?



What healthcare is healthy?



What behaviors are healthy?

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

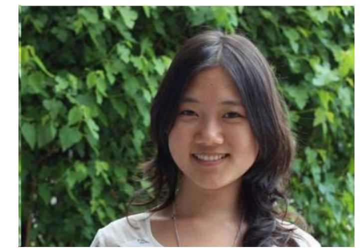
50

ML4H @ UofT / Vector Team

PhD Students

Bret Nestor

Denny Wu



Amy Lu

Matthew McDermott

Technical Collaborators

Dr. Anna Goldenburg

Dr. Shalmali Joshi

Clinical Collaborators

Dr. Amol Verma

Dr. Fahad Razak

Dr. Muhammad Mamdani

Challenges are Secret Opportunities!

Opportunities in Machine Learning for Healthcare

Marzyeh Ghassemi
University of Toronto, Vector Institute
Toronto, Canada
marzyeh@cs.toronto.edu

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

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

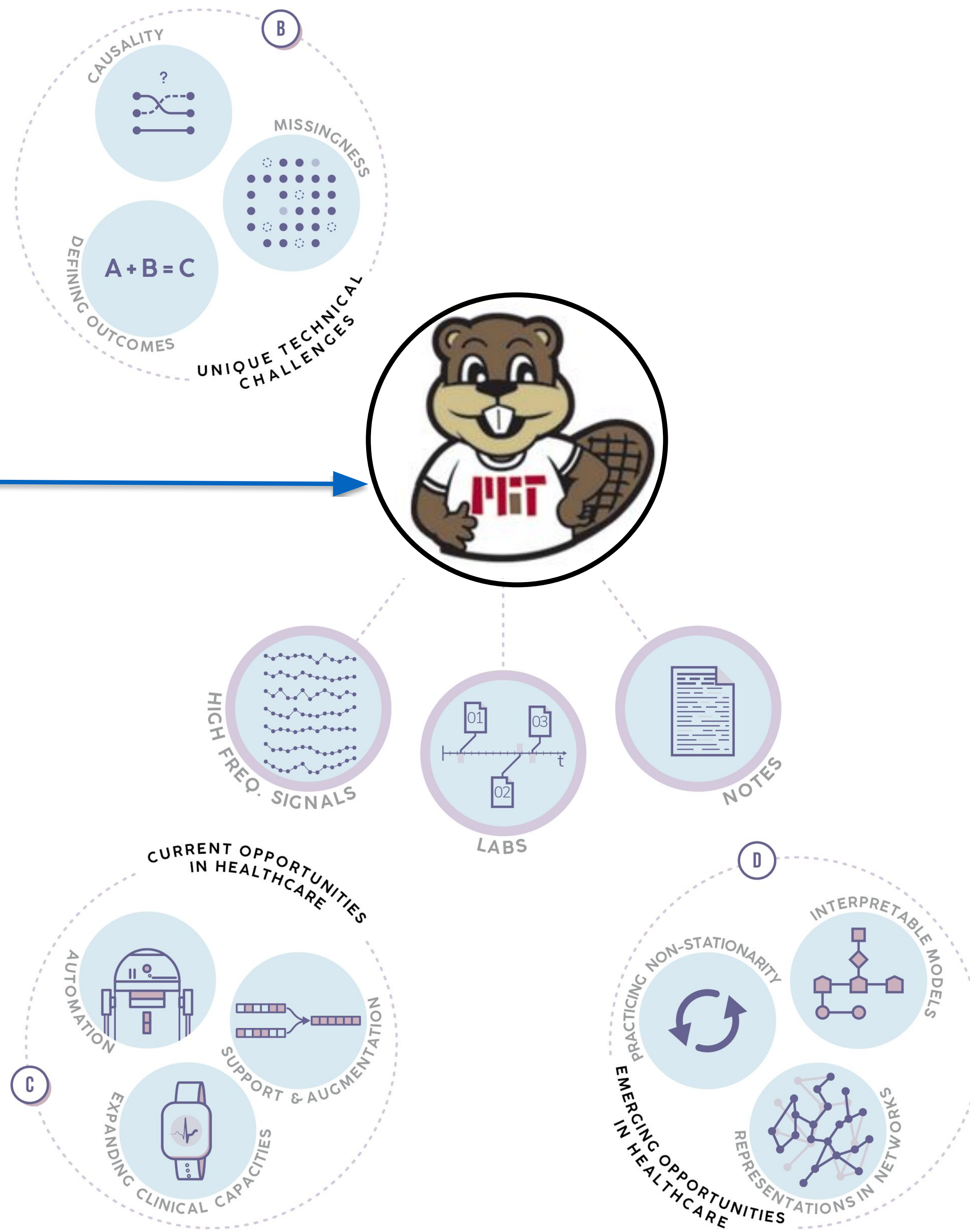
Andrew L. Beam
Harvard School of Public Health
Boston, MA 02115
andrew_beam@hms.harvard.edu

Irene Y. Chen
Massachusetts Institute of Technology
Cambridge, MA 02139
iychen@mit.edu

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

Abstract

Modern electronic health records (EHRs) provide data to answer clinically meaningful questions. The growing data in EHRs makes healthcare ripe for the use of machine learning. However, clinical data presents unique challenges that complicate the use of common machine learning methodologies. For example, these challenges include disease labels in EHRs, encompassing multiple underlying phenotypes, and the under representation of healthy individuals. This article serves as a primer to illuminate these challenges and highlights opportunities for members of the machine learning and data science communities to contribute to this domain.



UNIVERSITY OF TORONTO

52

VECTOR INSTITUTE | INSTITUT VECTEUR

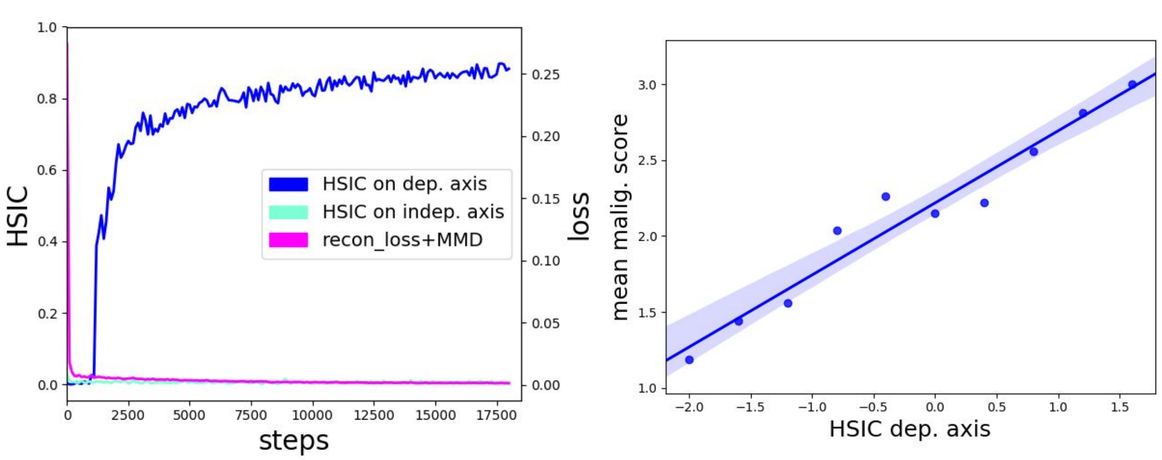
Unknown Knowns

- Fundamental research is needed in healthcare to understand **Difficult Disease Endotyping**, which may require that researchers work with clinicians to **Create Common Ground**.

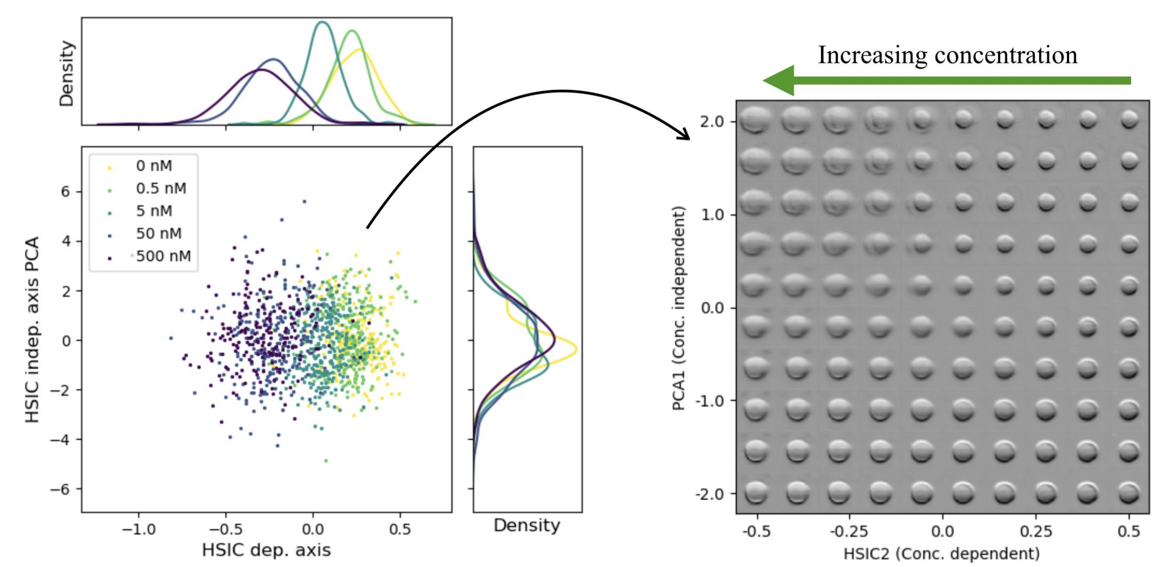
Modeling the Biological Pathology Continuum with HSIC-regularized Wasserstein Auto-encoders

NeuroIPS 2018 ML4H Workshop; Denny Wu, Hirofumi Kobayashi, Charles Ding, Lei Cheng, Keisuke Goda, Marzyeh Ghassemi

Create latent representations that reflect side information with WAE to model pathology continuum, and HSIC to enforce dependency between certain latent features and the provided side information



Training loss and HSIC loss vs. training steps + malignancy score of the nearest neighbors of generated samples vs. dependant axis; the trend of malignancy correlates with the dependent axis in Lung Image Data of thoracic scans from 1018 patient cases with 2670 images.



Scatter plot of test images on latent space of ~10,000 images from leukemia cell line K562 with dilutions of adriamycin. Class separation is obvious on x (dependant axis), but not on y (1st PC of independent axes. Generated images sampled from the dependent axis and the 1st PC of all other axes; generated cells vary in shape.

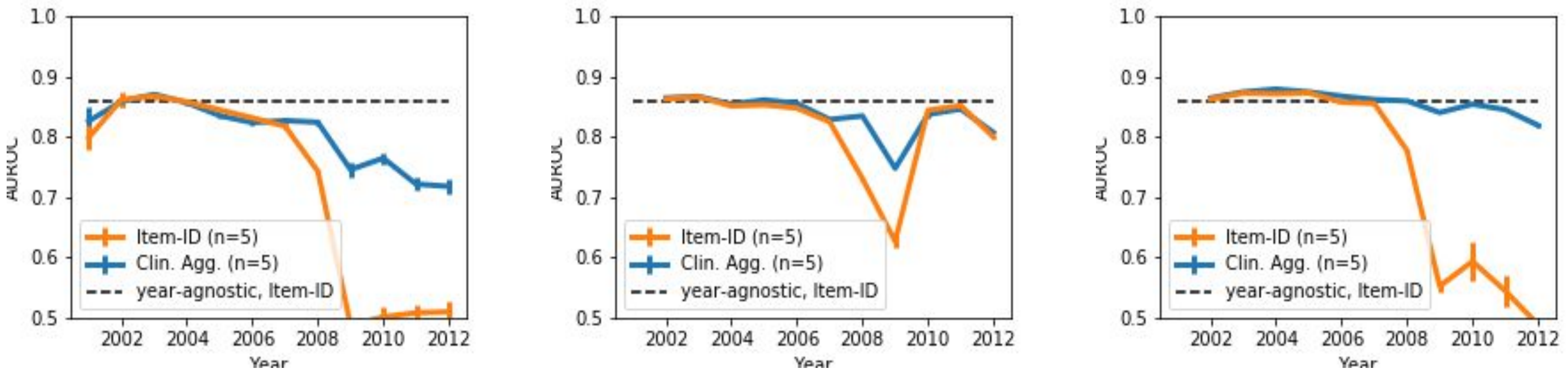
Complex Data Challenges

- We know that **Data Quality Matters**, but **Disease Data is Imbalanced**, and restrictive access makes **Data Only for Few** researchers.

Rethinking Clinical Prediction

NeuroIPS 2018 ML4H Workshop; Bret Nestor, Matthew B.A. McDermott, Geeticka Chauhan, Tristan Naumann, Michael C. Hughes, Anna Goldenberg, Marzyeh Ghassemi

Demonstrate that only models trained on all previous data using clinically aggregated features **generalise** across **hospital policy changes** and **year of care**.



Three training paradigms for mortality prediction in MIMIC III (~40,000 de-identified ICU patients from Beth Israel Deaconess Medical Center). Representations are trained on
A) 2001-2002 data only, B) previous year only, C) all previous years.

Dashed line is year-agnostic model performance - what most papers report for performance.

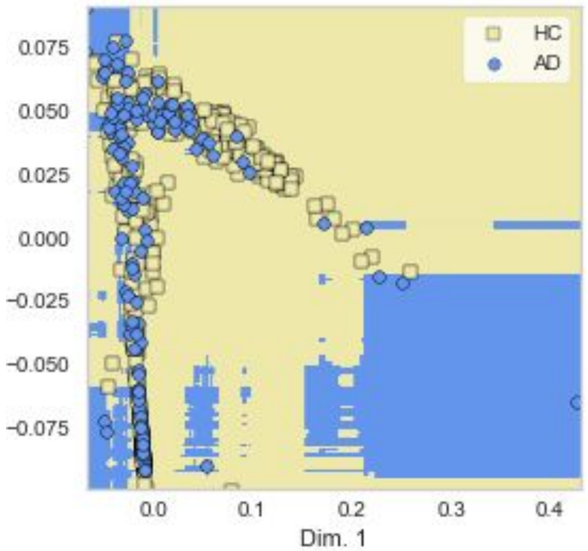
Robustness to The Unseen

- As devices and practices change the **Same Name maybe a Different Measure**, while novel *x*, *y, x*|*y* require **Anticipating New Data** and **Handling the Next Zika**.

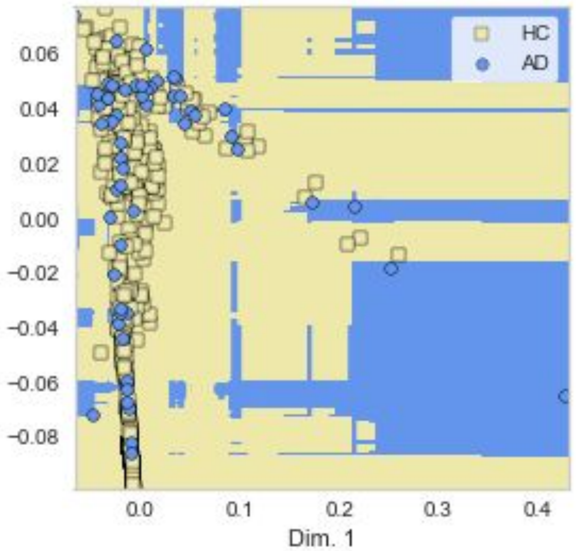
Effect of Heterogeneous Data for Alzheimer’s Disease Detection from Speech

NeurIPS 2018 ML4H Workshop Aparna Balagopalan, Jekaterina Novikova, Frank Rudzicz, Marzyeh Ghassemi

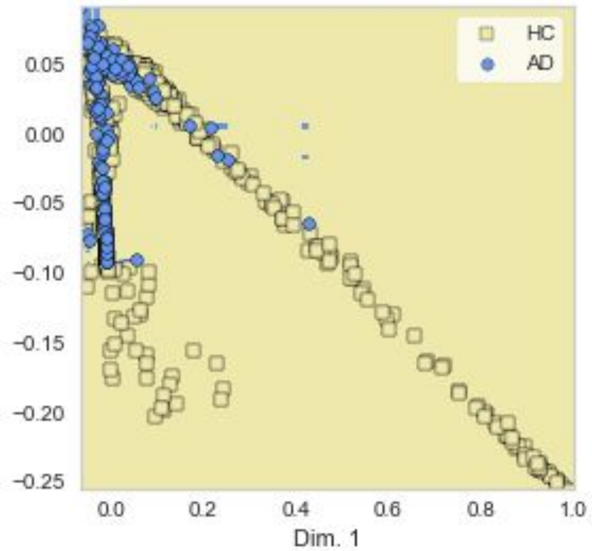
Augment AD with multi-task healthy data + analyze class boundaries. Adding in datasets with general, unstructured conversations improves models trained using structured tasks!



Adding in same task healthy data (122 samples). Pic. descriptions (PD); 28.6% out of task error



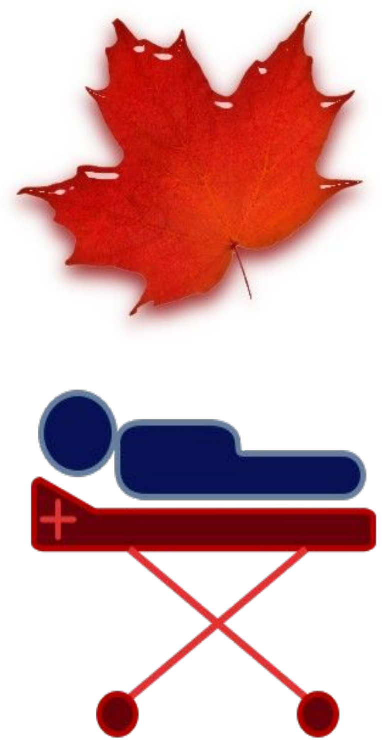
Adding in different structured task healthy data (327 samples) PD + structured tasks; 17.8% out of task error



Adding in general speech healthy data (231 samples) PD + general speech; 3.6% out of task error

Machine Learning For Health (ML4H)

What models are healthy?



What healthcare is healthy?

What behaviors are healthy?

What should Canada be doing?

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

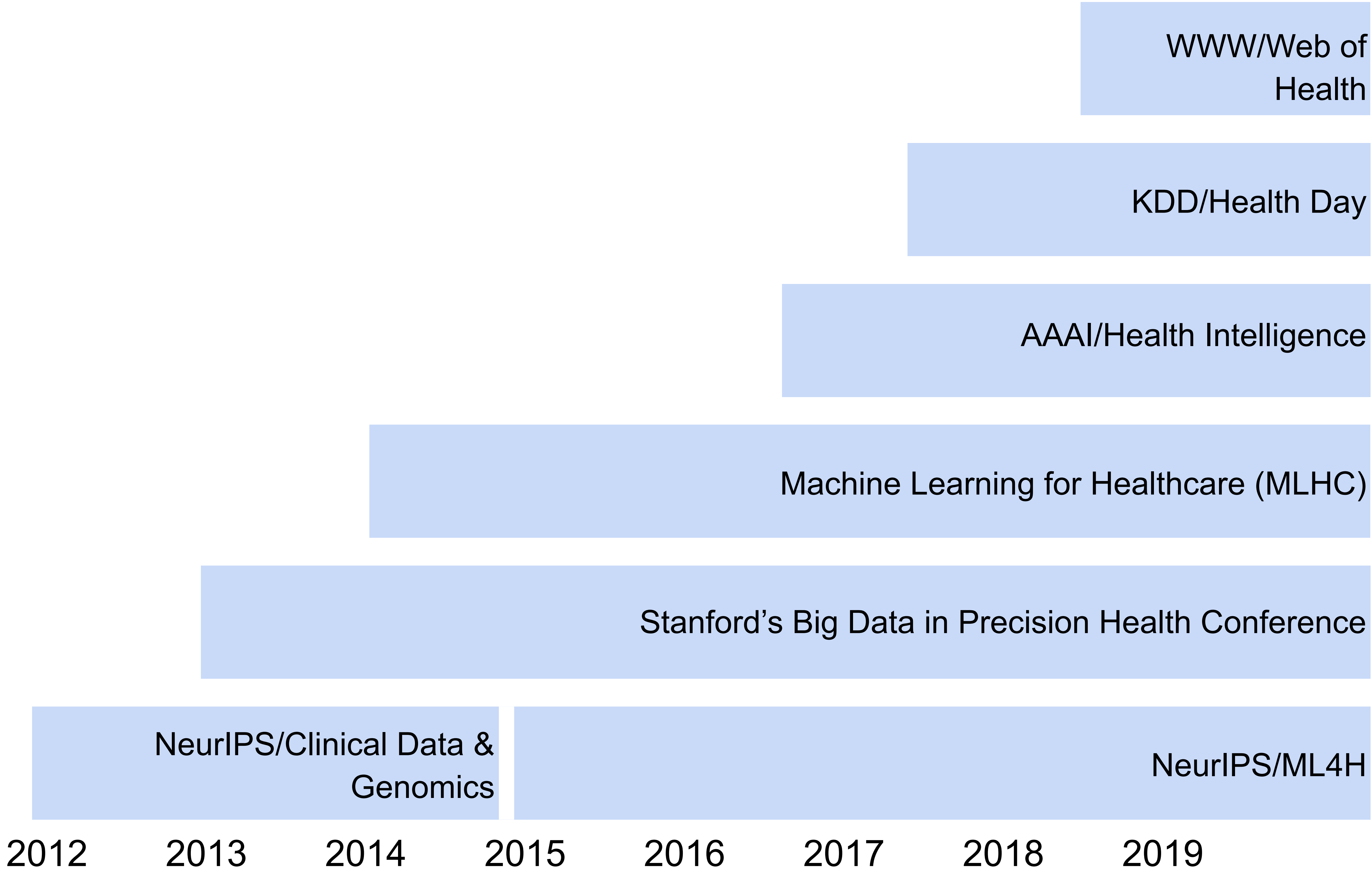
56

#1) Toronto Has a Limited Time To Lead ML4H

- Perfect mixture of technical and medical talent.
- Limited by vision, and resources.
- The field moves quickly…

57

ML is Growing Rapidly Into the Healthcare Space



UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

58

Applications Across the Human Lifespan

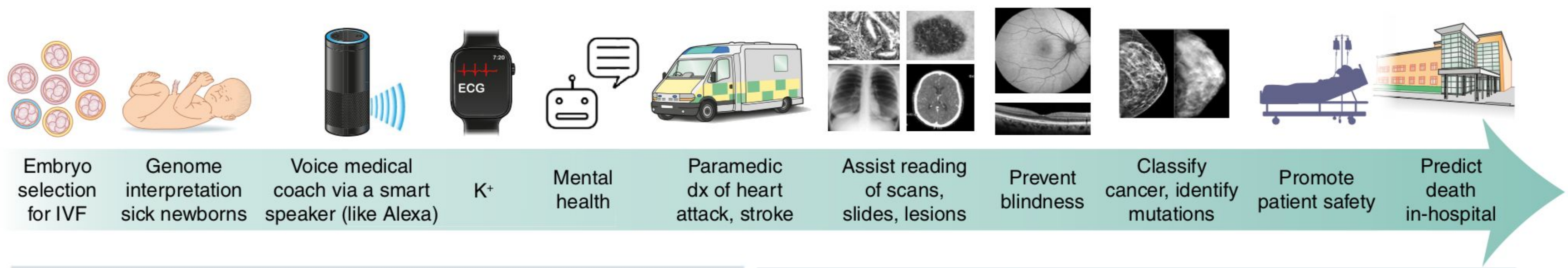


Table 3 | Selected reports of machine- and deep-learning algorithms to predict clinical outcomes and related parameters

Prediction	*n*	AUC	Publication (Reference number)
In-hospital mortality, unplanned readmission, prolonged LOS, final discharge diagnosis	216,221	0.93*0.75+0.85#	Rajkomar et al.[96]
All-cause 3-12 month mortality	221,284	0.93ˆ	Avati et al.[91]
Readmission	1,068	0.78	Shameer et al.[106]
Sepsis	230,936	0.67	Horng et al.[102]
Septic shock	16,234	0.83	Henry et al.[103]
Severe sepsis	203,000	0.85@	Culliton et al.[104]
Clostridium difficile infection	256,732	0.82++	Oh et al.[93]
Developing diseases	704,587	range	Miotto et al.[97]
Diagnosis	18,590	0.96	Yang et al.[90]
Dementia	76,367	0.91	Cleret de Langavant et al.[92]
Alzheimer's Disease (+ amyloid imaging)	273	0.91	Mathotaarachchi et al.[98]
Mortality after cancer chemotherapy	26,946	0.94	Elfiky et al.[95]
Disease onset for 133 conditions	298,000	range	Razavian et al.[105]
Suicide	5,543	0.84	Walsh et al.[86]
Delirium	18,223	0.68	Wong et al.[100]

LOS, length of stay; *n*, number of patients (training+validation datasets). For AUC values: *, in-hospital mortality; +, unplanned readmission; #, prolonged LOS; ˆ, all patients; @, structured+unstructured data; ++, for University of Michigan site.

Source: **High-performance medicine: the convergence of human and artificial intelligence** Eric Topol, Nature Medicine Jan 2019

UNIVERSITY OF TORONTO

Figure: Debbie Maizels / Springer Nature

VECTOR INSTITUTE | INSTITUT VECTEUR

59

ML As a Regulated Advice-Giver

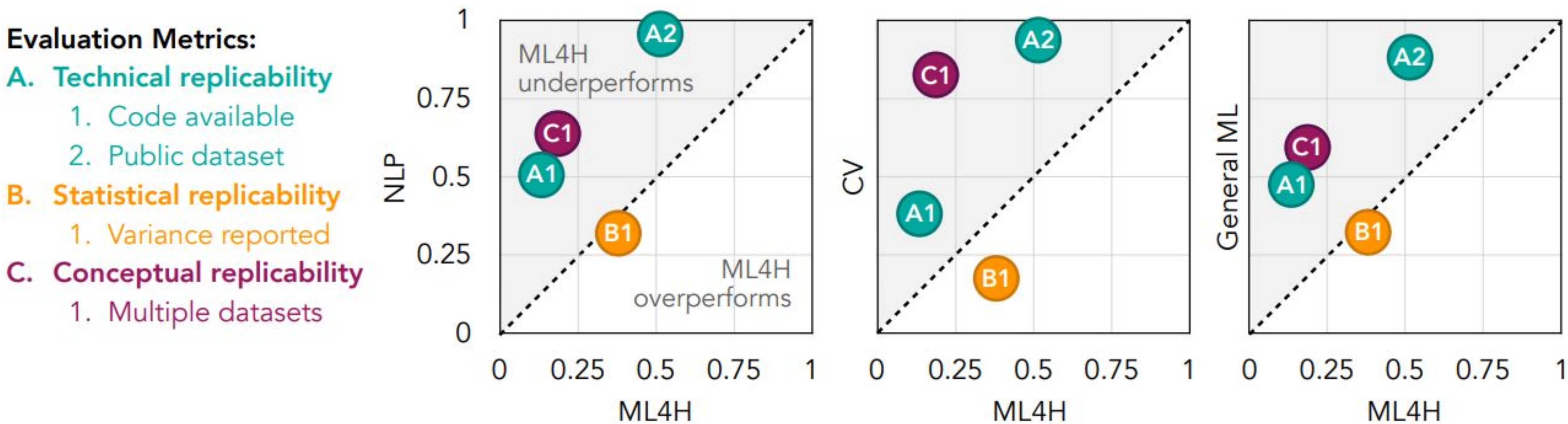
Table 2 | FDA AI approvals are accelerating

Company	FDA Approval	Indication
Apple	September 2018	Atrial fibrillation detection
Aidoc	August 2018	CT brain bleed diagnosis
iCAD	August 2018	Breast density via mammography
Zebra Medical	July 2018	Coronary calcium scoring
Bay Labs	June 2018	Echocardiogram EF determination
Neural Analytics	May 2018	Device for paramedic stroke diagnosis
IDx	April 2018	Diabetic retinopathy diagnosis
Icometrix	April 2018	MRI brain interpretation
Imagen	March 2018	X-ray wrist fracture diagnosis
Viz.ai	February 2018	CT stroke diagnosis
Arterys	February 2018	Liver and lung cancer (MRI, CT) diagnosis
MaxQ-AI	January 2018	CT brain bleed diagnosis
Alivecor	November 2017	Atrial fibrillation detection via Apple Watch
Arterys	January 2017	MRI heart interpretation

Source: **High-performance medicine: the convergence of human and artificial intelligence** Eric Topol, Nature Medicine Jan 2019

At least 12 additional AI applications have been cleared by FDA since the end of 2018, <u>a total of 26 to date.</u>

Unique Position To Promote Robust ML in Health



- Machine learning in healthcare requires robustness.
  - Technical replicability
  - Statistical replicability
  - Conceptual replicability

[1] Reproducibility in Machine Learning for Health; ICLR Reproducibility Workshop 2018 (under review); Matthew B. A. McDermott, Shirly Wang, Nikki Marinsek, Rajesh Ranganath, Marzyeh Ghassemi, Luca Foschini

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

Promote Better What is “Doctor-ing”

- 35% of doctors report burn-out + inability to make a personal patient connections.[1]
- 56% of doctors say they do not have time to be empathetic.[2]
- 40 seconds of compassion reduces patient anxiety.[3]
- Compassionate care can improve chronic low back pain, diabetes, the common cold, etc...[4]

[1] Shanafelt, Tait D., et al. "Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014." *Mayo Clinic Proceedings*. Vol. 90. No. 12. Elsevier, 2015.
[2] Riess, Helen, et al. "Empathy training for resident physicians: a randomized controlled trial of a neuroscience-informed curriculum." Journal of general internal medicine 27.10 (2012): 1280-1286.
[3] Fogarty, Linda A., et al. "Can 40 seconds of compassion reduce patient anxiety?." Journal of Clinical Oncology 17.1 (1999): 371-371.
[4] Trzeciak,Stephen and Mazzarelli, Anthony. “Compassionomics: The Revolutionary Scientific Evidence that Caring Makes a Difference.” 2019.

UNIVERSITY OF TORONTO

62

VECTOR INSTITUTE | INSTITUT VECTEUR

#2) Let’s Talk About Race

- **Lack** of ethnicity data in Canadian EHR is itself a **bias**.
- Our peers collect it to protect and **audit** care.
- Adding an extra RPDB column is **easy**.
- Not having ethnicity is a **liability** for our technical **leadership**.

DATA GAP

How Canada’s racial data gaps can be hazardous to your health

Canada lags far behind other countries in tracking how ethnicity affects the labour market, the justice system and health care. What are policy-makers missing?

TAVIA GRANT AND DENISE BALKISSOON
TORONTO
INCLUDES CORRECTION
PUBLISHED FEBRUARY 6, 2019
UPDATED FEBRUARY 11, 2019
23 COMMENTS

Olga Lambert of Ajax, Ont., has an aggressive form of breast cancer that she's battled three times in 11 years. Research in the U.S. and Britain has highlighted the elevated risks of cancer for black women, but Canada's information on race-based health issues is lacking.

TIJANA MARTIN/THE GLOBE AND MAIL

More • ‘Visible minority’ revisited • How you can help • Opinion: Andray Domise

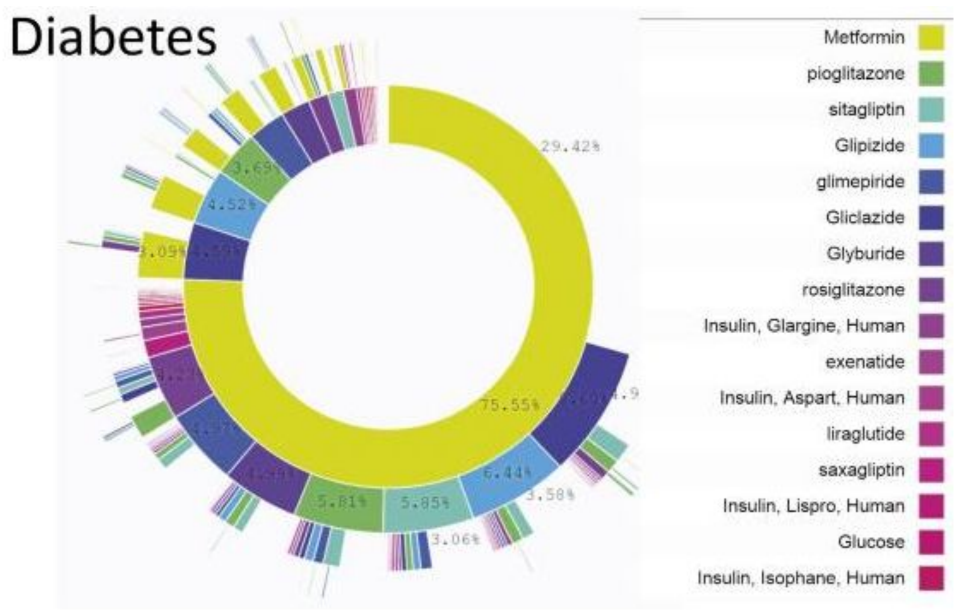
UNIVERSITY OF TORONTO

https://theconversation.com/how-anti-fat-bias-in-health-care-endangers-lives-115888
https://theconversation.com/the-fight-for-the-right-to-be-a-mother-9-ways-racism-impacts-maternal-health-111319
https://theconversation.com/racism-impacts-your-health-84112
https://torontoist.com/2016/04/african-canadian-prison-population/

63

VECTOR INSTITUTE | INSTITUT VECTEUR

Human Treatment Pathways Are Shockingly Unique



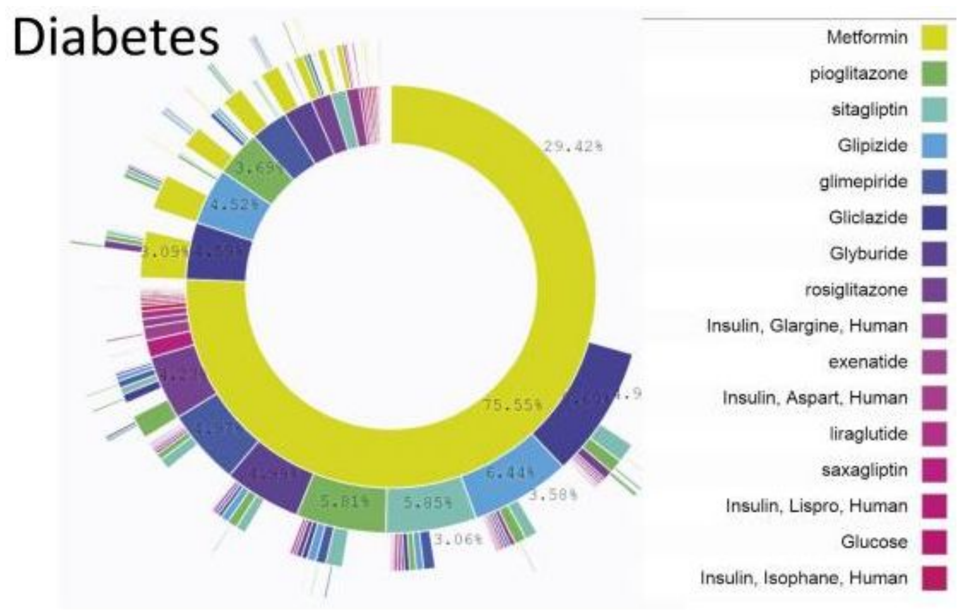
In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes:
- Depression:
- Hypertension:

[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

64

Human Treatment Pathways Are Shockingly Unique



In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes: **10%** of patients
- Depression:
- Hypertension:

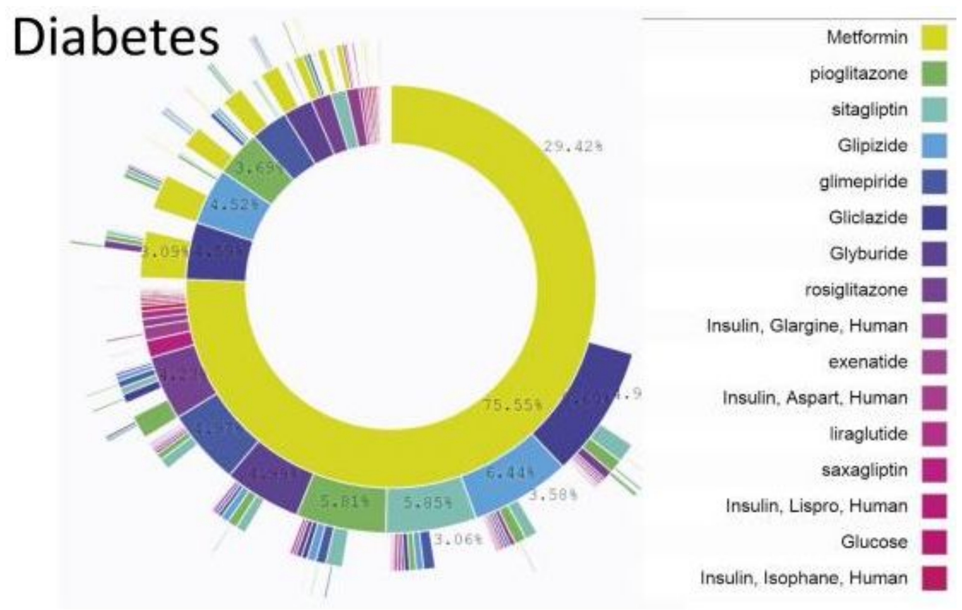
[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

UNIVERSITY OF TORONTO

65

VECTOR INSTITUTE | INSTITUT VECTEUR

Human Treatment Pathways Are Shockingly Unique



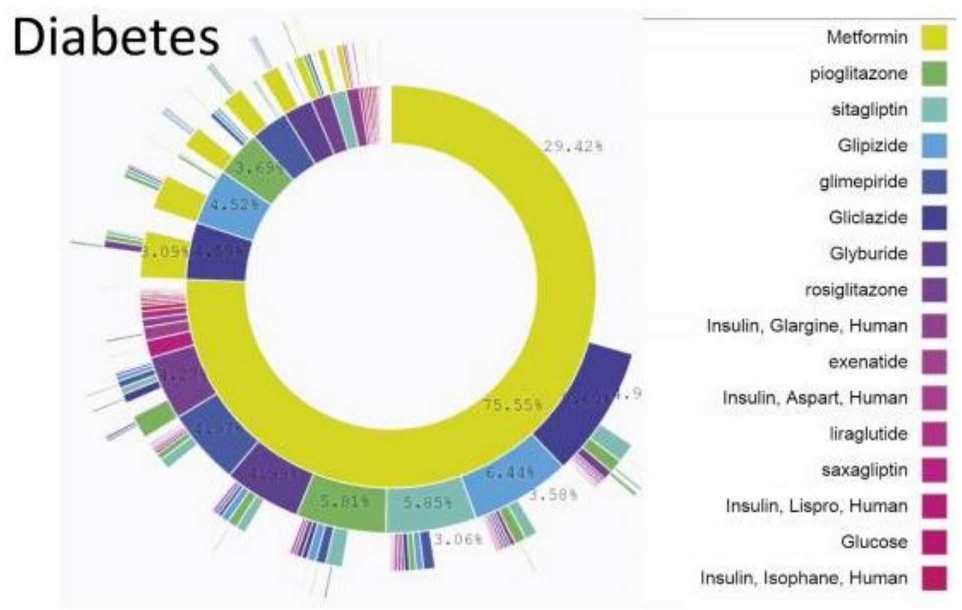
In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes: **10%** of patients
- Depression: **11%** of patients
- Hypertension:

[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

66

Human Treatment Pathways Are Shockingly Unique



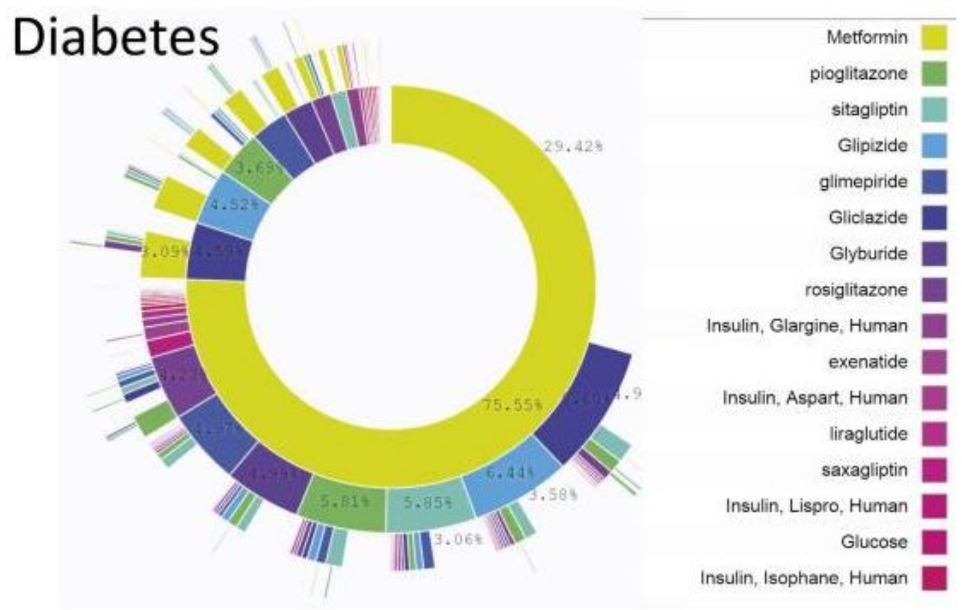
In a combined EHR/claims dataset from 11 sources/4 countries/250 million patients, how many followed a unique treatment pathway?

- Diabetes: **10%** of patients
- Depression: **11%** of patients
- Hypertension: **24%** of patients

[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

67

Human Treatment Pathways Are Shockingly Unique



"In an underlying population of 250 million, based on my 3-y treatment pathway, what patients are like me?"

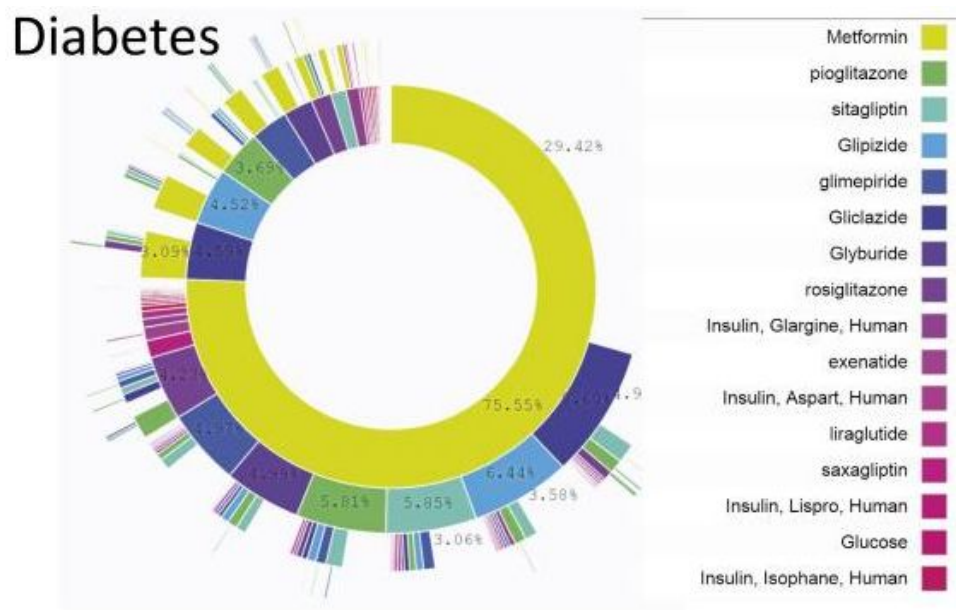
UNIVERSITY OF TORONTO

[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

68

VECTOR INSTITUTE | INSTITUT VECTEUR

Human Treatment Pathways Are Shockingly Unique



"In an underlying population of 250 million, based on my 3-y treatment pathway, what patients are like me?"

For 24% of hypertension patients, "**No one.**"

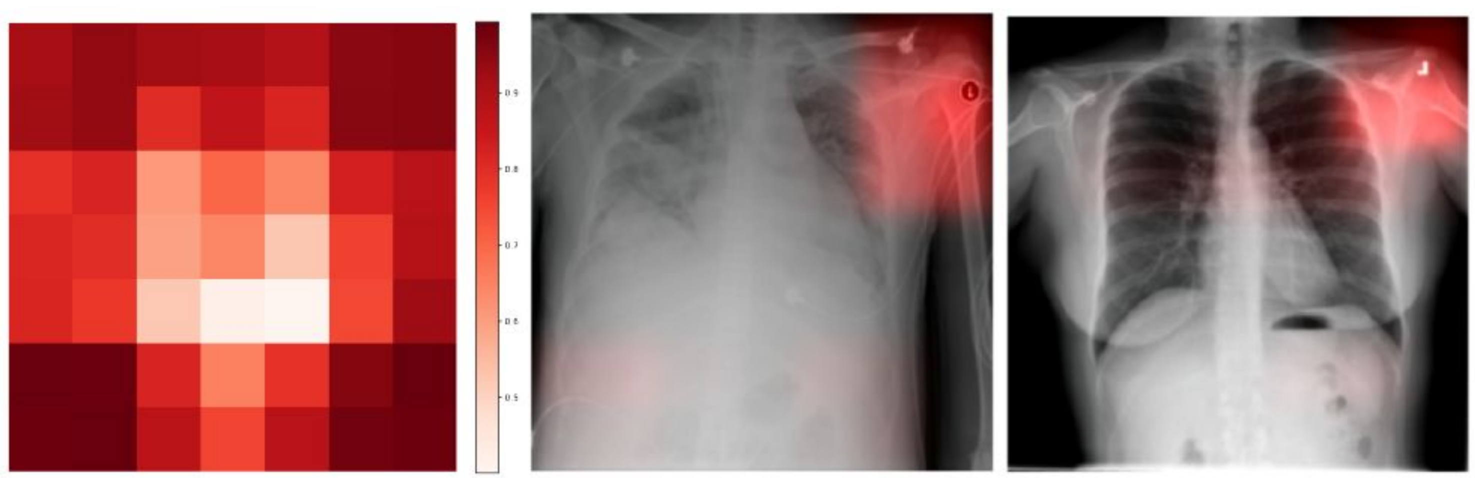
[1] Hripcsak, George, et al. "Characterizing treatment pathways at scale using the OHDSI network." Proceedings of the National Academy of Sciences 113.27 (2016): 7329-7336.

69

Learning Unintended Features Is Too Easy

- CNN models can determine the hospital that the patient was admitted to with 95% accuracy… from the X-ray.[1]

Fig 4. CNN to predict hospital system detected both general and specific image features. (a) We obtained activation heatmaps from our trained model and averaged over a sample of images to reveal which subregions tended to contribute to a hospital system classification decision. Many different subregions strongly predicted the correct hospital system, with especially strong contributions from image corners. (b-c) On individual images, which have been normalized to highlight only the most influential regions and not all those that contributed to a positive classification, we note that the CNN has learned to detect a metal token that radiology technicians place on the patient in the corner of the image field of view at the time they capture the image. When these strong features are correlated with disease prevalence, models can leverage them to indirectly predict disease.



(a) (b) (c)

[1] Zech, John R., et al. "Confounding variables can degrade generalization performance of radiological deep learning models." *arXiv preprint arXiv:1807.00431* (2018).

70

#3) Health Data As A Resource; Treat It That Way.

- All data is valuable; health data particularly so.
- Robust algorithms require large scale datasets for research use.

AWS Machine Learning Blog

Improving Patient Care with Machine Learning At Beth Israel Deaconess Medical Center

by Dr. Matt Wood | on 04 MAR 2019 | Permalink | Comments | Share

Beth Israel Deaconess Medical Center has launched a multi-year, innovative research program on how machine learning can improve patient care, supported by an academic research sponsorship grant from AWS. The Harvard Medical School-affiliated teaching hospital will use a broad array of AWS machine learning services to uncover new ways that machine learning technology can enhance clinical care, streamline operations, and eliminate waste, with the goal of improving patient care and quality of life.

Improving patient care with machine learning

Inefficiencies in hospital management and operations are not only extremely costly to providers, insurers, patients, and taxpayers, but they can result in precious resources being diverted away from patient care. These inefficiencies drive healthcare costs up and can contribute to life-threatening medical

Amazon Comprehend Medical

Extract information from unstructured medical text accurately and quickly

No machine learning experience required

Get started with Amazon Comprehend Medical

Amazon Comprehend Medical is a natural language processing service that makes it easy to use machine learning to extract relevant medical information from unstructured text. Using Amazon Comprehend Medical, you can quickly and accurately gather information, such as medical condition, medication, dosage, strength, and frequency from a variety of sources like doctors' notes, clinical trial reports, and patient health records.

Google Tries to Patent Healthcare Deep Learning, EHR Analytics

Google has applied for a sweeping patent including the fundamentals of deep learning and EHR analytics in the healthcare industry.

Source: Google

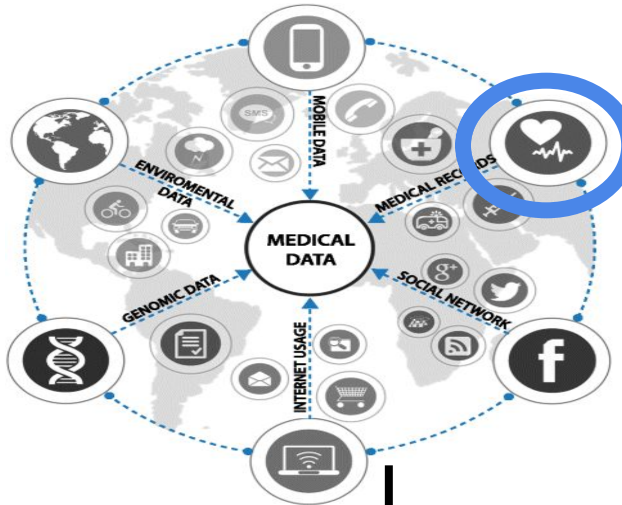
UNIVERSITY OF TORONTO

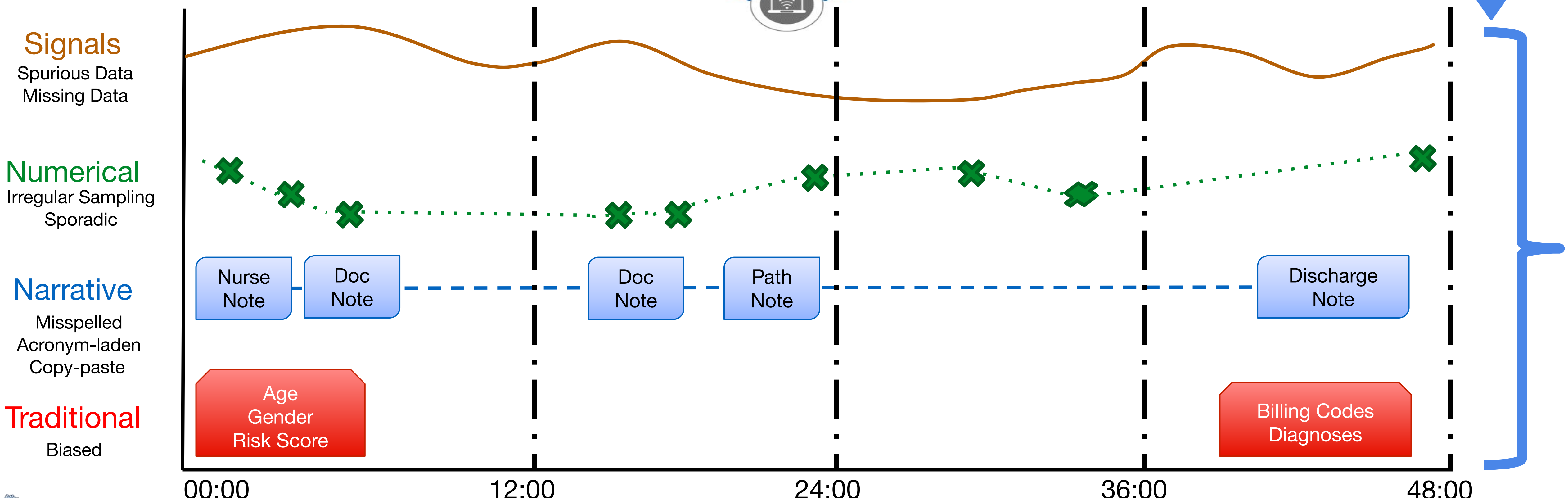
VECTOR INSTITUTE | INSTITUT VECTEUR

71

Create **Research** with a **Resource**

- ML4H is currently defined by ONE dataset - MIMIC 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).

UNIVERSITY OF TORONTO

VECTOR INSTITUTE | INSTITUT VECTEUR

72

A Decade of Vetted Access to De-identified Data

- MIMIC has been around for over a decade.
- No lawsuits or newspaper headlines regarding privacy failures.
- Vetted access to de-identified data demonstrably safe, even for a single source in a small city.

IRB Approval

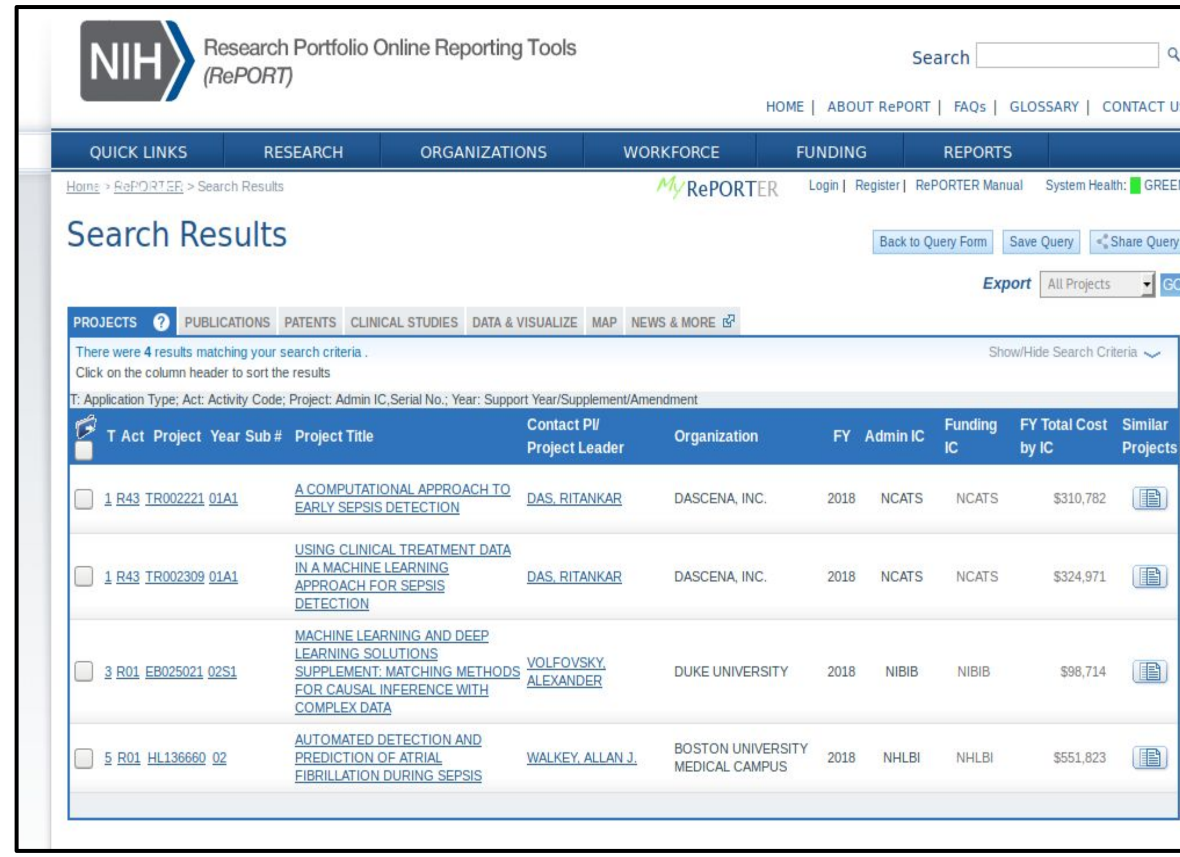
This study was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived as the study did not impact clinical care and all data were de-identified.

The MIMIC II database was collected as part of a Bioengineering Research Partnership (BRP) grant from the National Institute of Biomedical Imaging and Bioengineering entitled, "Integrating Data, Models and Reasoning in Intensive Care" (RO1-EB001659). The project was established in October 2003 and included an interdisciplinary team from academia (MIT), industry (Philips Medical Systems) and clinical medicine (Beth Israel Deaconess Medical Center). The objective of the BRP is to develop and evaluate advanced Intensive Care Unit (ICU) patient monitoring systems that will substantially improve the efficiency, accuracy and timeliness of clinical decision making in intensive care.

The MIMIC Model Works - ICES/GEMINI Options

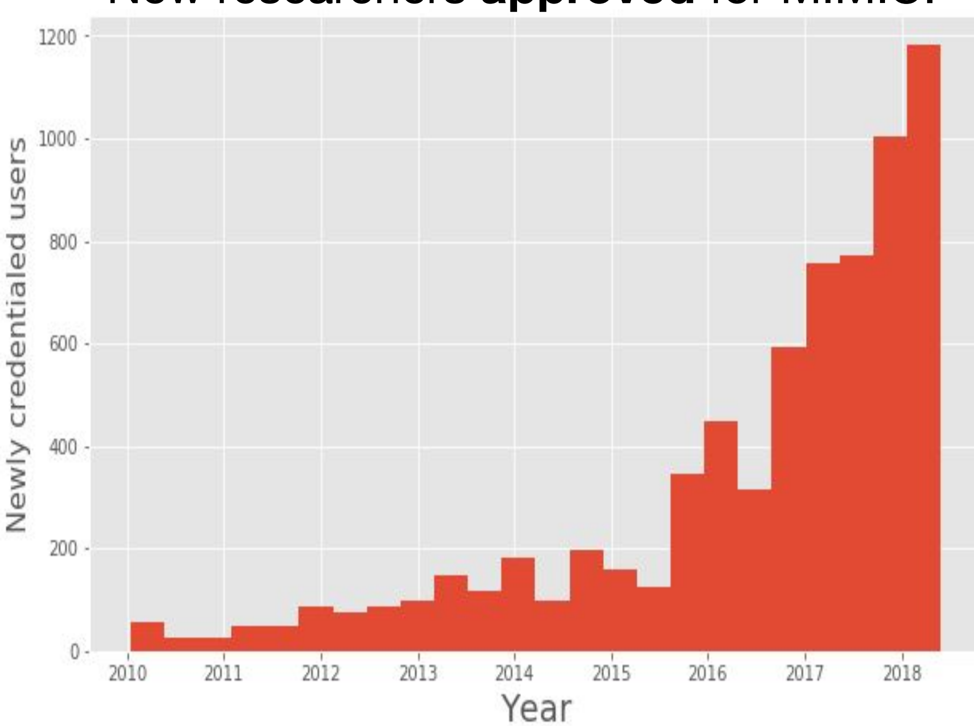
- Openly accessible, de-identified clinical dataset
- Privacy risks mitigated with vetted users under EULA
- Streamlined access to data
- Enabling collaboration, benchmarking, reproducibility

Funded NIH Grants **based** on MIMIC (~$1.3M in 2018):



T	Act	Project	Year	Sub #	Project Title	Contact PI/ Project Leader	Organization	FY	Admin IC	Funding IC	FY Total Cost by IC	Similar Projects
1	R43	TR002221	01A1		A COMPUTATIONAL APPROACH TO EARLY SEPSIS DETECTION	DAS, RITANKAR	DASCENA, INC.	2018	NCATS	NCATS	$310,782	
1	R43	TR002309	01A1		USING CLINICAL TREATMENT DATA IN A MACHINE LEARNING APPROACH FOR SEPSIS DETECTION	DAS, RITANKAR	DASCENA, INC.	2018	NCATS	NCATS	$324,971	
3	R01	EB025021	02S1		MACHINE LEARNING AND DEEP LEARNING SOLUTIONS SUPPLEMENT: MATCHING METHODS FOR CAUSAL INFERENCE WITH COMPLEX DATA	VOLFOVSKY, ALEXANDER	DUKE UNIVERSITY	2018	NIBIB	NIBIB	$98,714	
5	R01	HL136660	02		AUTOMATED DETECTION AND PREDICTION OF ATRIAL FIBRILLATION DURING SEPSIS	WALKEY, ALLAN J.	BOSTON UNIVERSITY MEDICAL CAMPUS	2018	NHLBI	NHLBI	$551,823	

New researchers **approved** for MIMIC:



Machine Learning in Health **overfits** models to MIMIC:

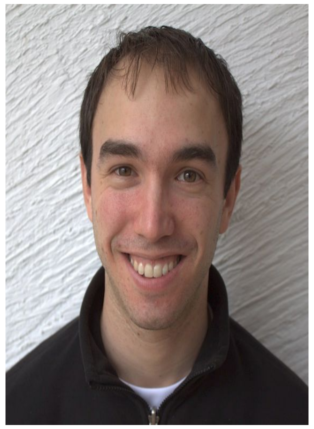
TORONTO

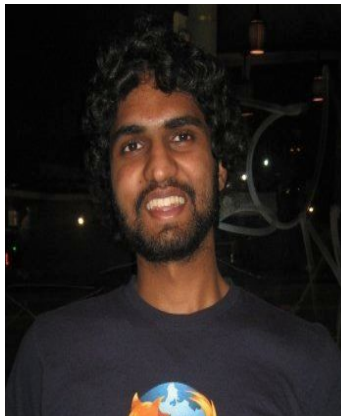
VECTOR INSTITUTE | INSTITUT VECTEUR

Speech or Vision?



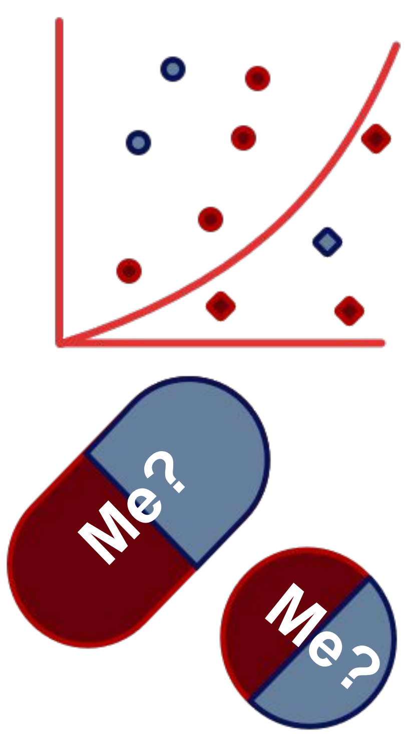




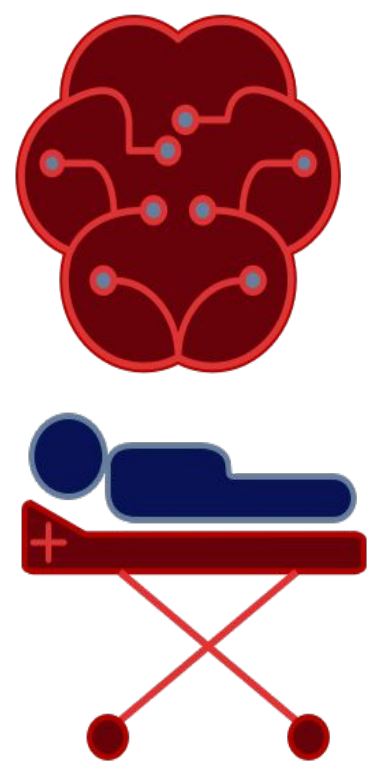


Machine Learning For Health (ML4H)

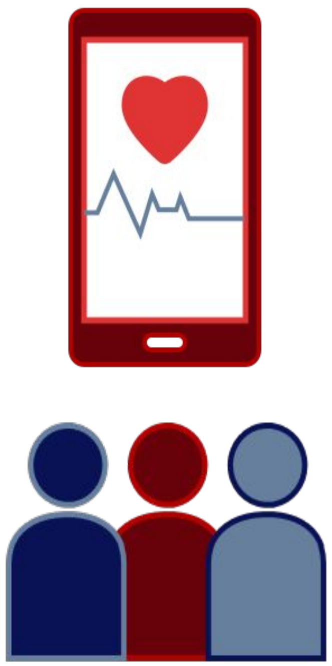
Creating actionable insights in human health.



What models are healthy?



What healthcare is healthy?



What behaviors are healthy?

76