



# Machine Learning for Clinicians:

Advances for Multi-Modal Health Data

Michael C. Hughes

A Tutorial at MLHC 2018, August 16, 2018

## **PART 3:** Methods that Address Challenges in Health Data

Missing data, incomplete data, multimodal data, interpretability, causality, sequential decision-making

Slides / Resources / Bibliography:

[https://michaelchughes.com/mlhc2018\\_tutorial.html](https://michaelchughes.com/mlhc2018_tutorial.html)

# Part 3 outline: Challenges

M : “missing data”

I : “incomplete labels” (semisupervised learning)

M : “multimodal data” (text + images + EHR codes)

I : “interpretability”

C : “causality”

S : “sequential decision making” (reinforcement learning)

# MLHC Challenge 1

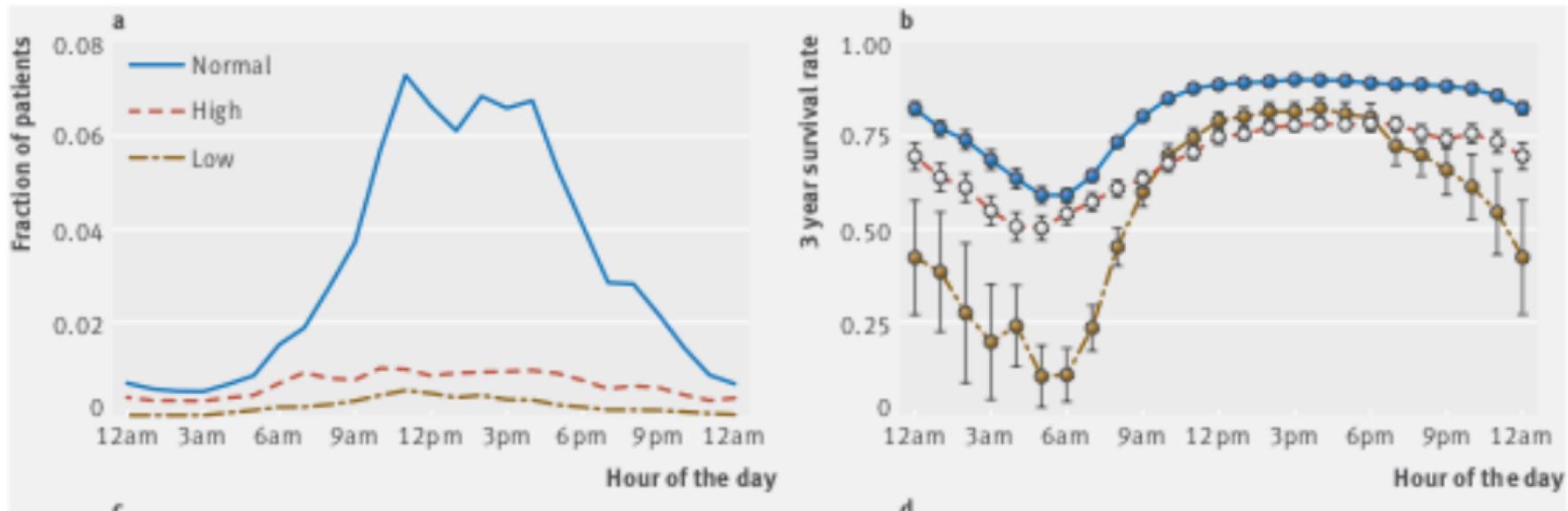
# Missing Data

All supervised prediction methods we've discussed require each example's features to be **fully observed**.

Problem: Medical data often missing,  
and almost always not at random

- What are strategies?
- What models/methods are available?

# Time-of-day for ordering blood test predicts survival



Low value + Ordered 3-6PM: >75% survival rate

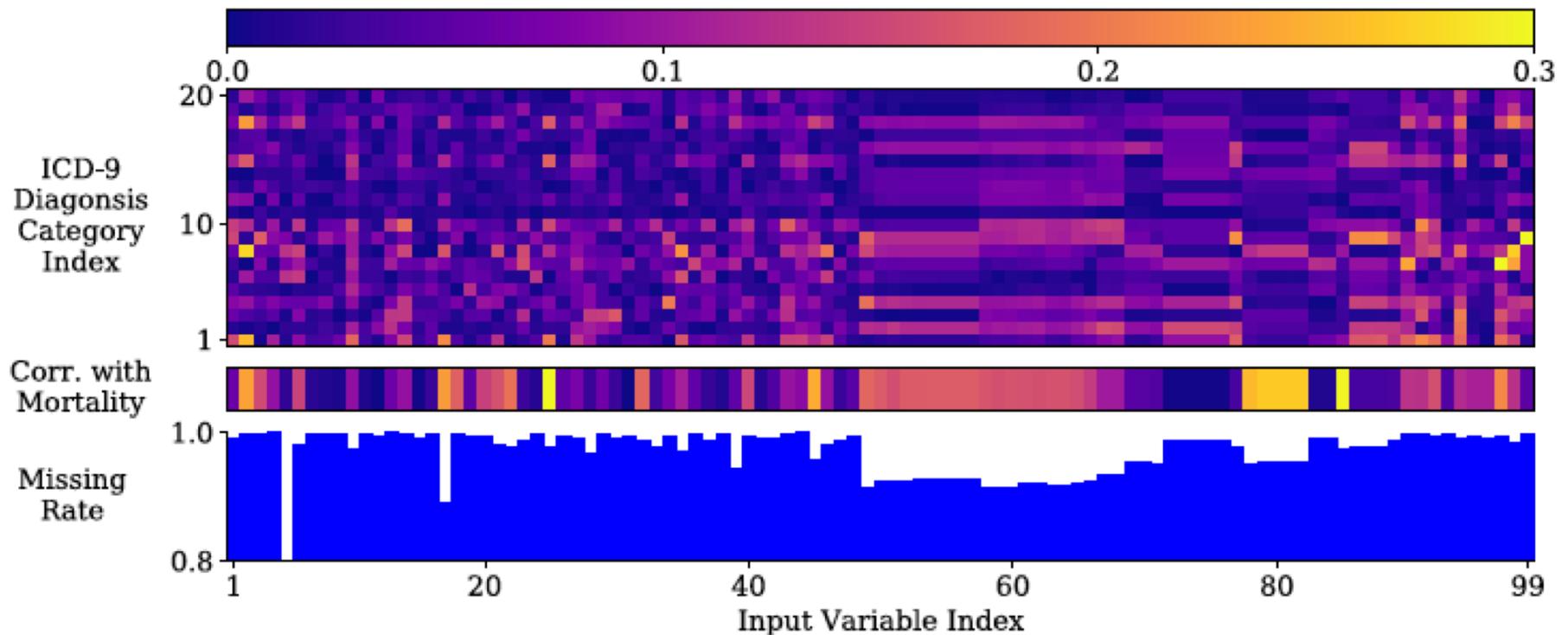
Low value + Ordered 3-6AM: <25% survival rate

*Need to capture human processes behind decisions to collect data!*

Credit: Agniel, Kohane, & Weber  
BMJ 2018

# Missingness predicts mortality

Absolute Values of Pearson Correlations between Variable Missing Rates and Labels  
(Mortality and ICD-9 Diagnosis Categories on MIMIC-III Dataset)



*Credit: Che et al. 2018 Scientific Reports*

# Imputation Strategies

- Fill with Population Mean
- Forward-carry
  - Fill with nearest value from patient's past
- Model-based
  - Discriminative:
    - Build predictor that imputes missing values given others
    - Build embedding that is amenable to missing input
  - Generative
    - Draw samples of missing data

# Example: impute by predicting

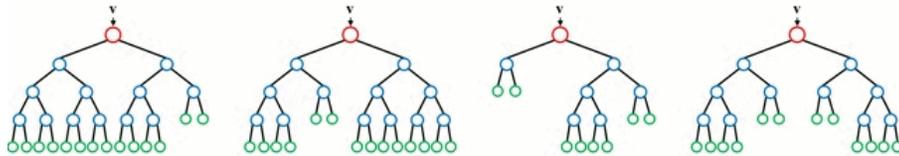
## MissForest—non-parametric missing value imputation for mixed-type data FREE

Daniel J. Stekhoven ✉, Peter Bühlmann

*Bioinformatics*, Volume 28, Issue 1, 1 January 2012, Pages 112–118,

<https://doi.org/10.1093/bioinformatics/btr597>

**Published:** 28 October 2011 **Article history** ▼

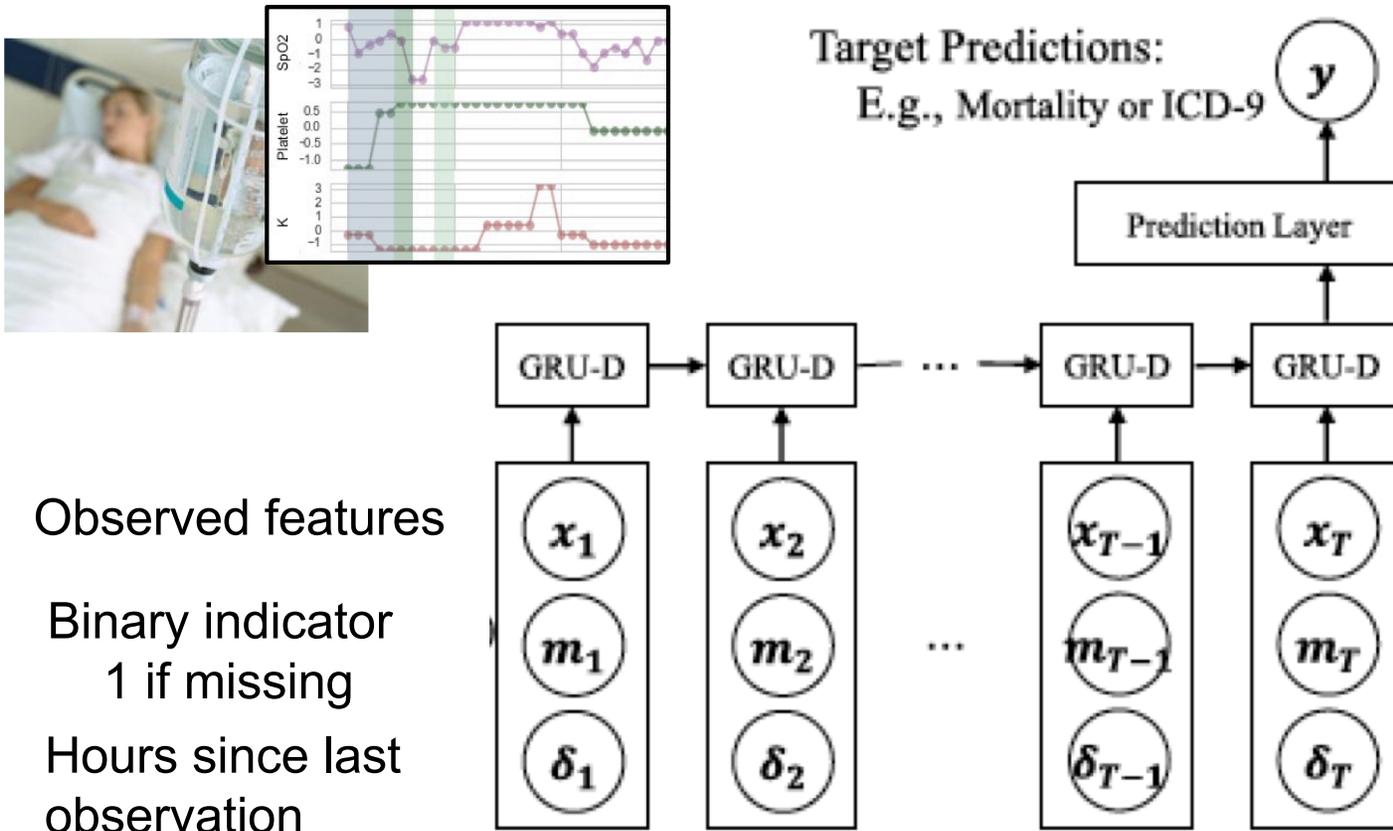


RF naturally handles multiple data types (real, categorical\*, binary)

# GRU-D (Che et al. 2018)

GRU unit is a simple alternative to LSTM unit

Use case: ICU time series



RNN that deliberately handles *missingness*

# GRU-D: Improvements over baseline imputation strategies

Non-RNN Models				RNN Models	
<i>Mortality Prediction On MIMIC-III Dataset</i>				LSTM-Mean	0.8142 ± 0.014
LR-Mean	0.7589 ± 0.015	RF-Mean	0.8293 ± 0.004	GRU-Mean	0.8252 ± 0.011
LR-Forward	0.7792 ± 0.018	RF-Forward	0.8303 ± 0.003	GRU-Forward	0.8192 ± 0.013
LR-Simple	0.7715 ± 0.015	RF-Simple	0.8294 ± 0.007	GRU-Simple w/o $\delta^{22}$	0.8367 ± 0.009
LR-SoftImpute	0.7598 ± 0.017	RF-SoftImpute	0.7855 ± 0.011	GRU-Simple w/o $m^{23,24}$	0.8266 ± 0.009
LR-KNN	0.6877 ± 0.011	RF-KNN	0.7135 ± 0.015	GRU-Simple	0.8380 ± 0.008
LR-CubicSpline	0.7270 ± 0.005	RF-CubicSpline	0.8339 ± 0.007	GRU-CubicSpline	0.8180 ± 0.011
LR-MICE	0.6965 ± 0.019	RF-MICE	0.7159 ± 0.005	GRU-MICE	0.7527 ± 0.015
LR-MF	0.7158 ± 0.018	RF-MF	0.7234 ± 0.011	GRU-MF	0.7843 ± 0.012
LR-PCA	0.7246 ± 0.014	RF-PCA	0.7747 ± 0.009	GRU-PCA	0.8236 ± 0.007
LR-MissForest	0.7279 ± 0.016	RF-MissForest	0.7858 ± 0.010	GRU-MissForest	0.8239 ± 0.006
				<b>Proposed GRU-D</b>	<b>0.8527 ± 0.003</b>

Table 1. Model performances measured by AUC score (*mean ± std*) for mortality prediction.

*Credit: Che et al. 2018*

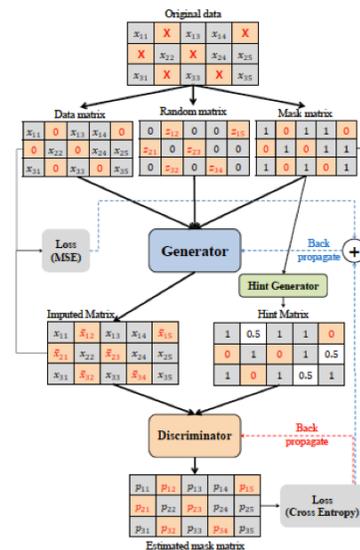
RandForest (RF) very competitive at 0.83.

# GAIN: Generative Adversarial Imputation Network

**GAIN: Missing Data Imputation using Generative Adversarial Nets**

Jinsung Yoon<sup>1\*</sup> James Jordon<sup>2\*</sup> Mihaela van der Schaar<sup>1,2,3</sup>

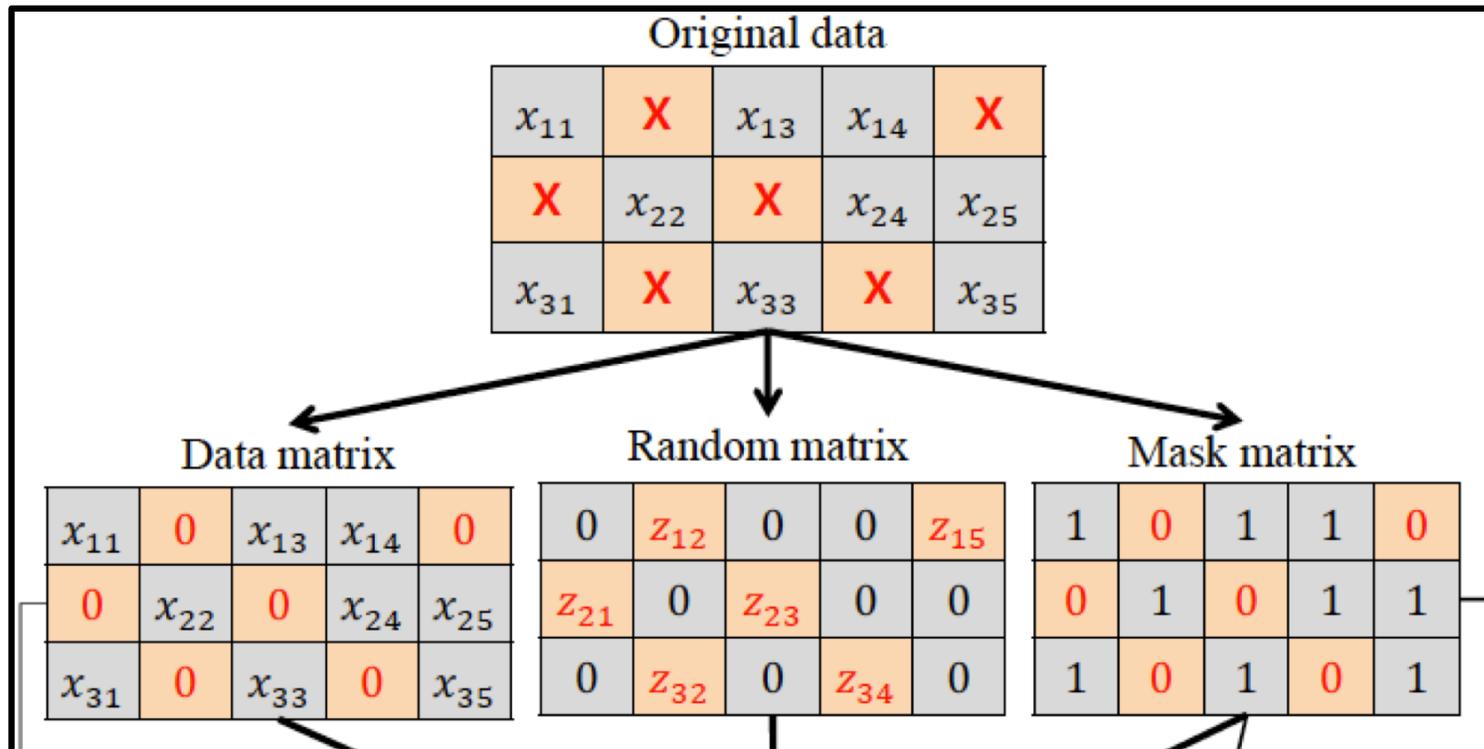
*Yoon et al. ICML 2018*



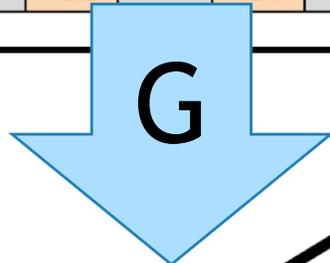
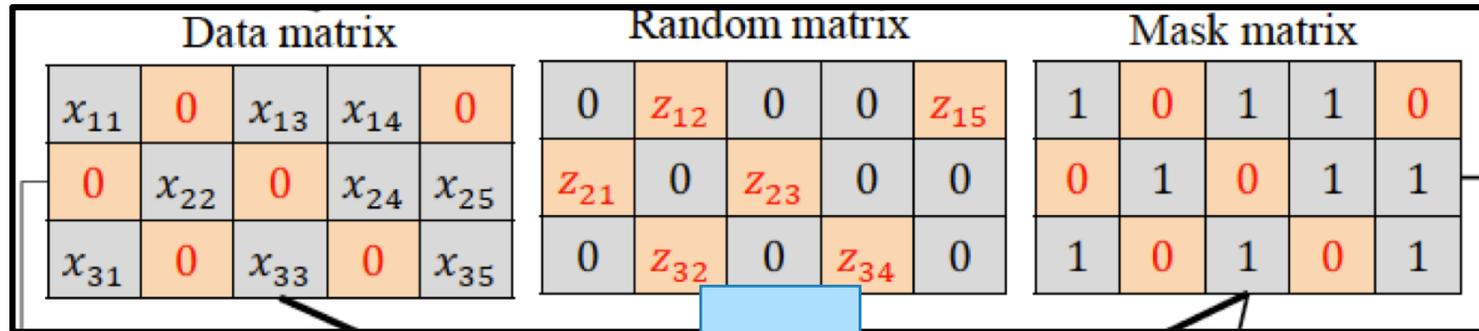
GAN that deliberately represents *missingness*

*Can draw samples!*

# GAIN: Data Representation



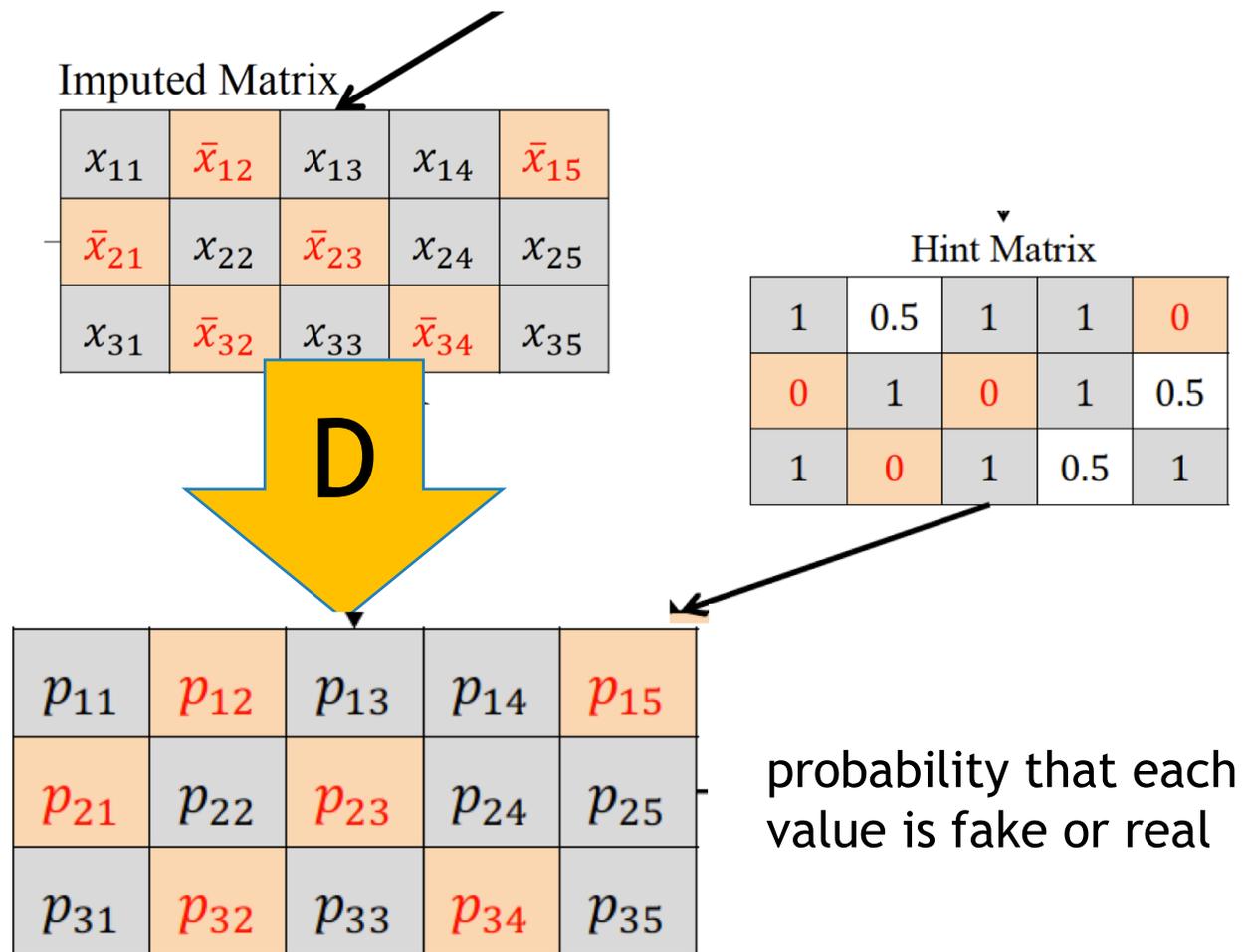
# GAIN: Generator



Imputed Matrix

$x_{11}$	$\bar{x}_{12}$	$x_{13}$	$x_{14}$	$\bar{x}_{15}$
$\bar{x}_{21}$	$x_{22}$	$\bar{x}_{23}$	$x_{24}$	$x_{25}$
$x_{31}$	$\bar{x}_{32}$	$x_{33}$	$\bar{x}_{34}$	$x_{35}$

# GAIN: Discriminator

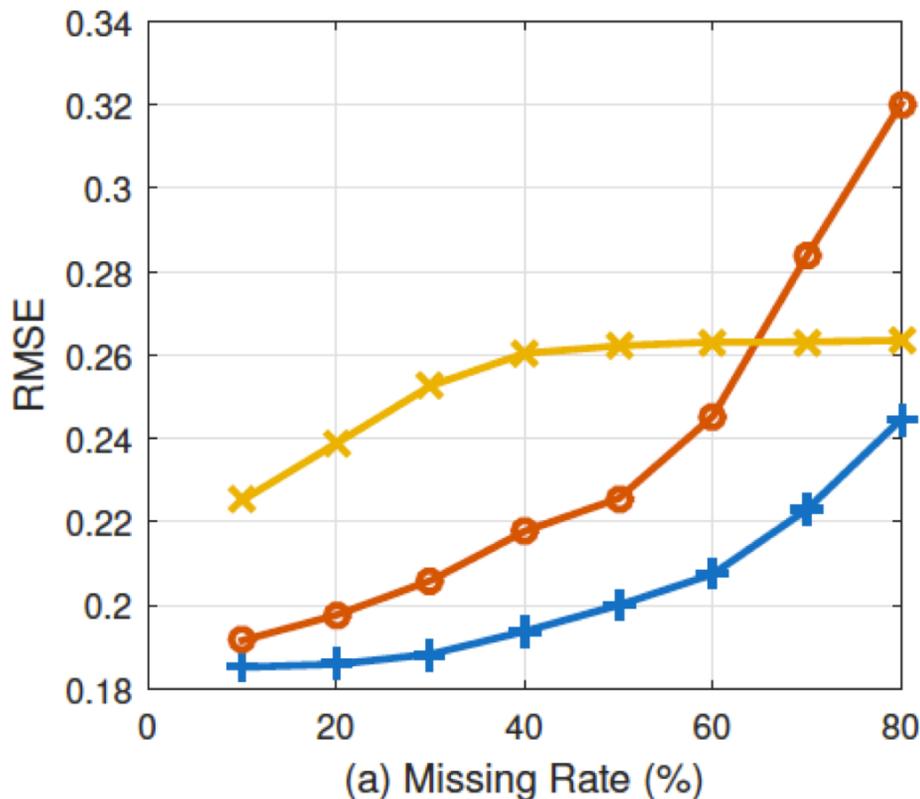


Discriminator predicts which values are fake (but needs some hints)

# GAIN: Improvements over Baseline Imputation Strategies

UCI Credit dataset. 690 examples. 15 features (mix of real, cat, binary)

How good are imputations?  
(lower = better)



How good are predictions with imputed data?  
(higher = better)

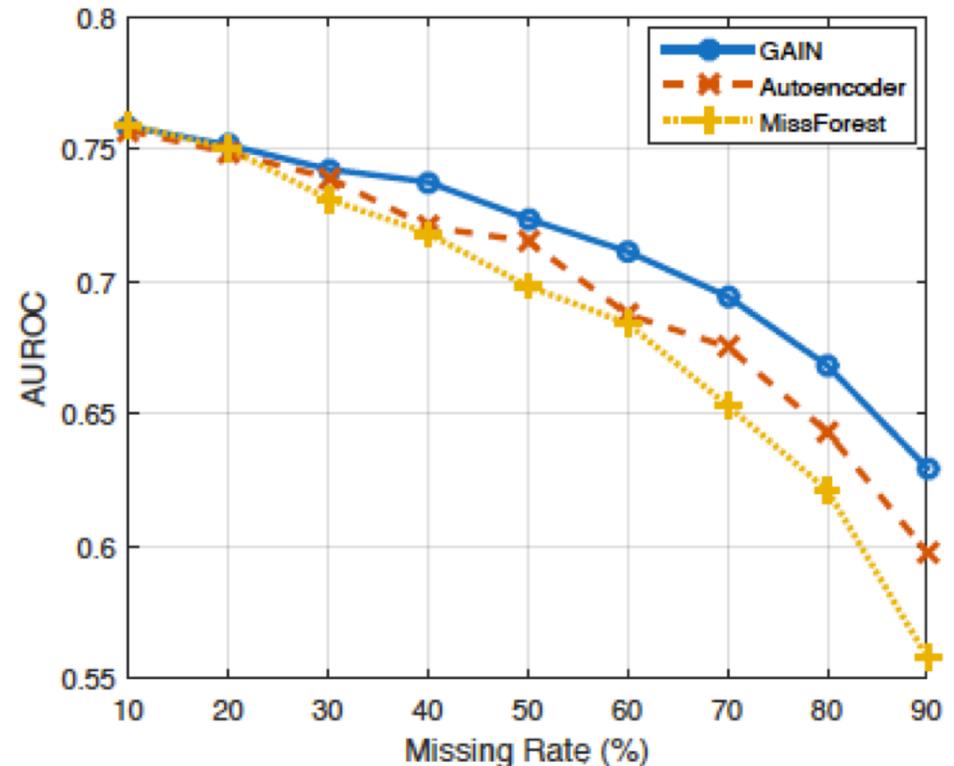


Fig. Credit: Yoon et al. 2018

# Takeaways: Missing Data

- Know where your data comes from!
  - Help your data scientist friends understand what missing values mean in the clinic.
- Some flexible methods exist; more work is needed.
- Think about sanity checks for imputed values from any predictive system.

# MLHC Challenge 2

## Incomplete Labels

Some examples have associated labels

Many more examples available, but have no labels

Expensive

Time-consuming

Dangerous (give drug to new patient)

*Supervised learning* can only use labeled set

*Semi-supervised learning* tries to learn from both!

# Possible Approaches

- Self-training
- Co-training
- Two stage: Pretrained features + classifier
- Generative models

# Self-training

- 1) Train predictor on labeled set
- 2) Predict outcomes for unlabeled data
- 3) Add “high confidence” predictions to labeled set
- 4) Return to (1)

Very easy to do with any classifier.  
BUT probably a bad approach. Do we trust the predictor?

# Co-training

## REQUIRES

- Two “views” or modalities
  - Image and text
- Each view predicts well on its own
- Each view is “independent” given label
  
- Add V1’s most confident predictions to V2’s training set, and vice versa

Very easy to do with any classifier.  
BUT probably a bad approach. Do we trust the predictor?

# Co-training + Active Sensing

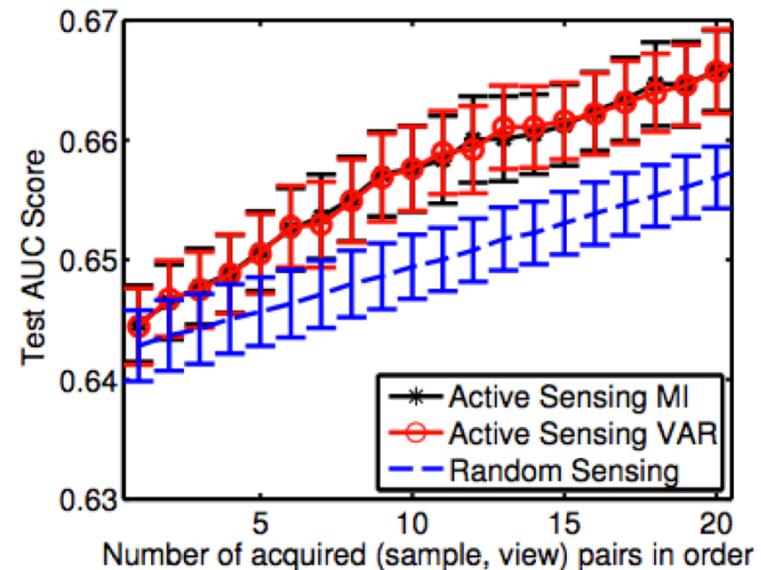
Yu et al. "Bayesian co-training" JMLR 2011

Features for NSCLC 2-years Survival Prediction

Feature	Description	View
GENDER	1-Male, 2-Female	1st
WHO	WHO performance status	1st
FEV1	Forced expiratory volume in 1 second	1st
GTV	Gross tumor volume	2nd
NPLN	Number of positive lymph node stations	2nd

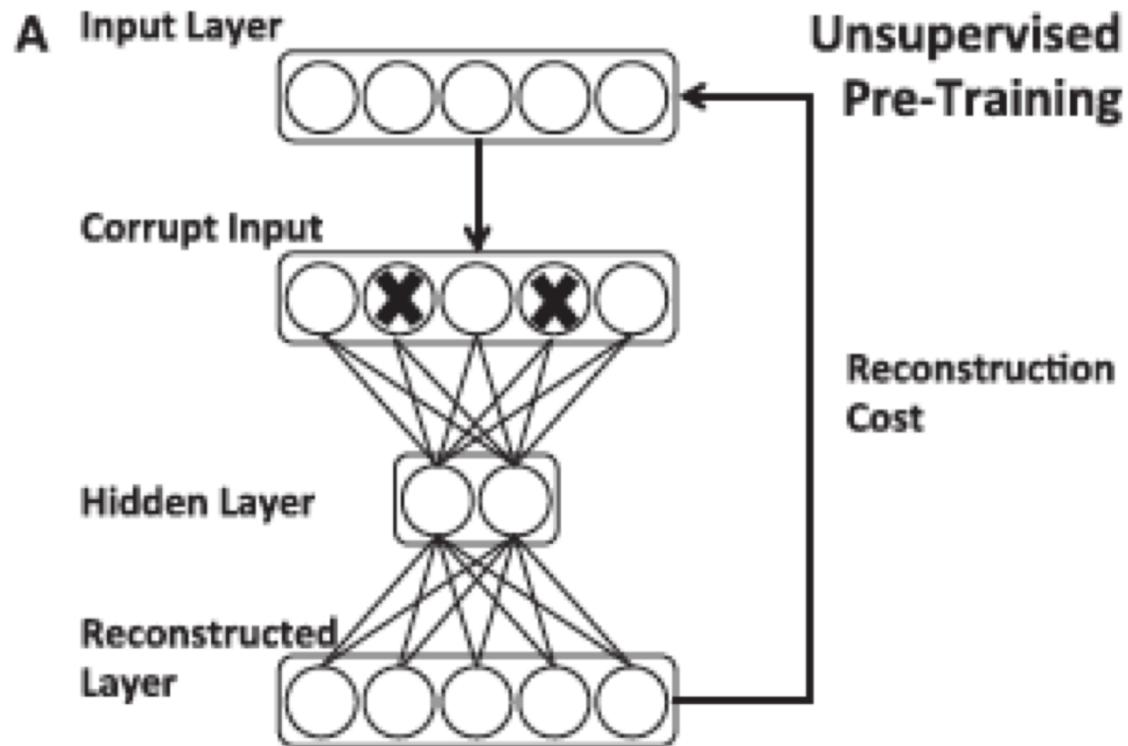
0.73 = all labels available

0.62 = AUC when just use "fill with mean"



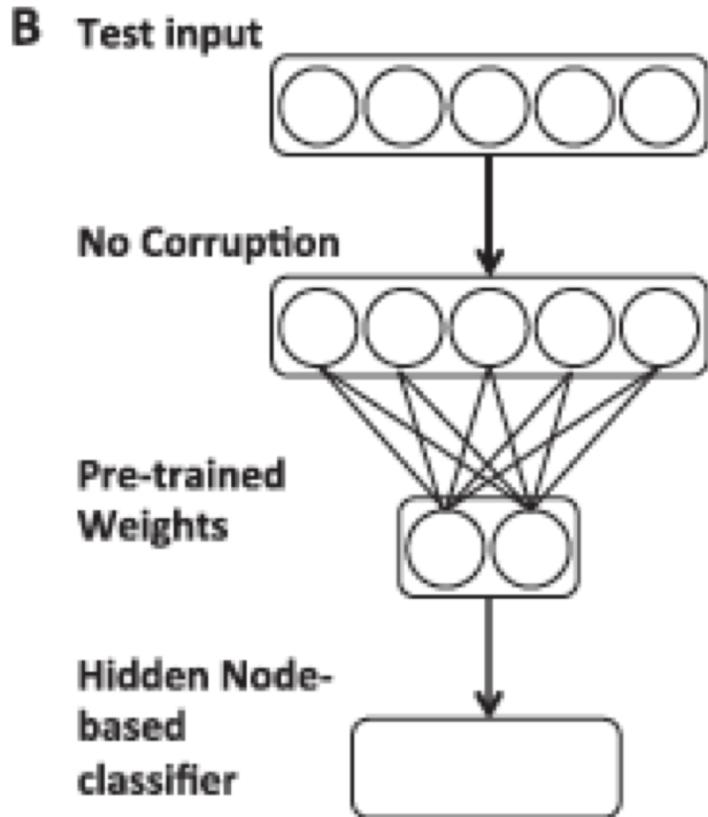
Can we use predictions from demographic data to guess which patients we should image?

# Denoising Autoencoder



*Credit: Beaulieu-Jones, Greene, et al. J. Biomed. Informatics 2016*

# 2-stage Classifier using DA Features

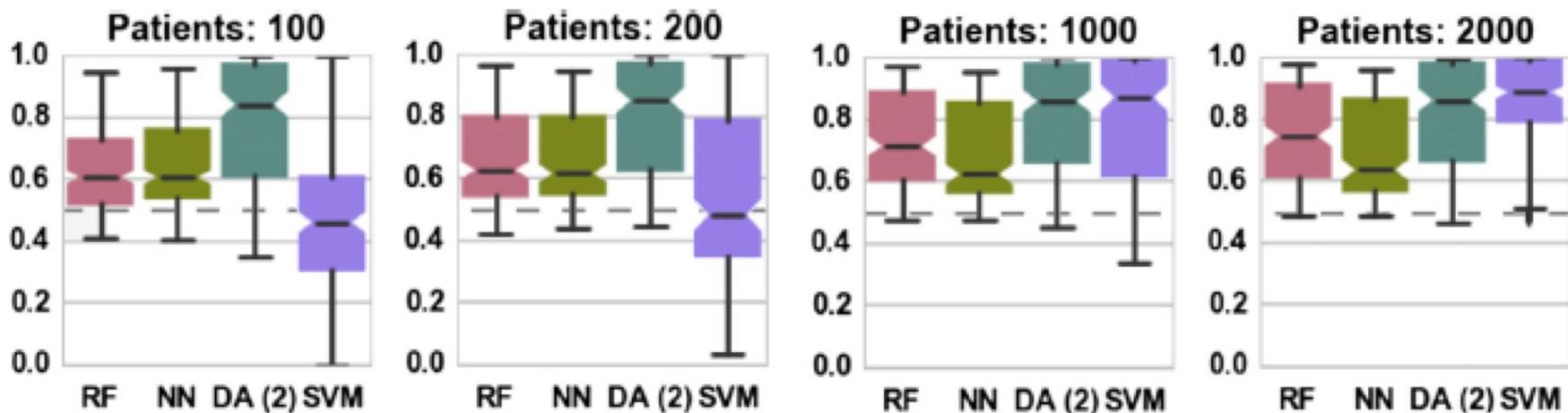


# 2-stage DA: Results on Simulated Data

**Table 2**

Mean receiver operating curve area under curve by method under simulation model 1.

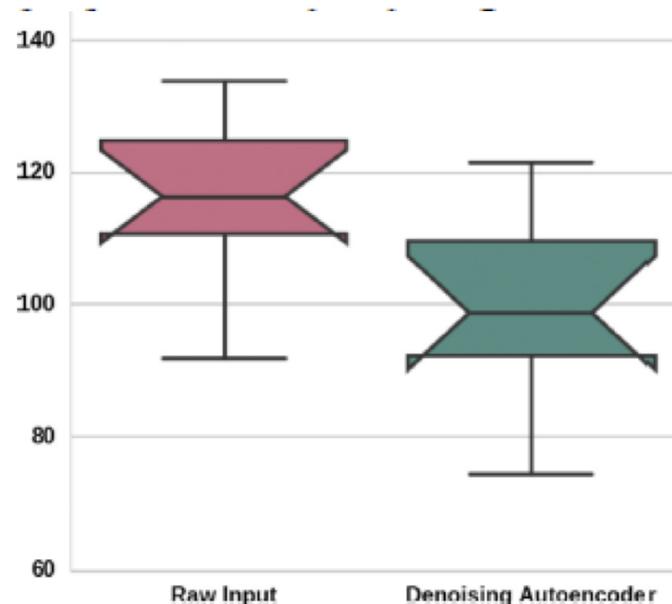
Patients	DA + RF	Random forest	Support vector machine
100	0.618	0.653	0.504 with RBF kernel
200	0.637	0.610	0.449
500	0.677	0.690	0.663
1000	0.774	0.717	0.776
2000	0.755	0.736	0.862



# 2-stage DA: Results on ALS Data

The PRO-ACT dataset includes 23 clinical trials covering 10,723 patients. We limit our survival analysis to the 3398 patients with known death information, but perform unsupervised pre-training of the DA with all 10,723 patients.

Mean Avg. Error  
(lower is better)



Raw Input  
size 6812

DA  
size 256

# Cool Idea: Disentangled Semi-supervised VAE

*Siddharth et al. NIPS 2017*

Visual Analogies



Can model **styles** of handwriting (lots of unlabeled data).  
Transfer those to different labels (using small labeled data).

# Reproducibility for SSL?

---

## Realistic Evaluation of Semi-Supervised Learning Algorithms

---

Avital Oliver<sup>\*1,2</sup> Augustus Odena<sup>\*1</sup> Colin Raffel<sup>\*1</sup>  
Ekin D. Cubuk<sup>1</sup> Ian J. Goodfellow<sup>1</sup>

Method	CIFAR-10 4k Labels
II-M (Sajjadi et al., 2016b)	11.29%
II-M (Laine & Aila, 2017)	12.36%
MT (Tarvainen & Valpola, 2017)	12.31%
VAT (Miyato et al., 2017)	11.36%
VAT + EM (Miyato et al., 2017)	10.55%
<hr/> <b>Results above this line cannot be directly compared to those below</b>	
Supervised	20.26 ± 0.38%
II-Model	16.37 ± 0.63%
Mean Teacher	15.87 ± 0.28%
VAT	13.86 ± 0.27%
VAT + EM	13.13 ± 0.39%
Pseudo-Label	17.78 ± 0.57%

Using only 4k labeled examples, but lots more effort on regularization, data augmentation

**13.4%**

Results in literature might be overly optimistic.  
Careful reproduction shows baselines much stronger than claimed.

# When is SSL the right choice?

Our discoveries also hint towards settings where SSL is most likely the right choice for practitioners:

- When there are no high-quality labeled datasets from similar domains to use for fine-tuning.
- When the labeled data is collected by sampling i.i.d. from the pool of the unlabeled data, rather than coming from a (slightly) different distribution.
- When the labeled dataset is large enough to accurately estimate validation accuracy, which is necessary when doing model selection and tuning hyperparameters.

Credit: Oliver et al. 2018

# Takeaways: Incomplete Labels

- When can you do better with unlabeled data?
  - Unlabeled & labeled from same distribution
  - Large-enough validation set
- Should you spend months of research effort on:
  - Applying tricks of trade (data augmentation)
  - Trying different SSL methods with data you have
  - Labeling more data?

## MLHC Challenge 3

# Multiple Data Sources

Health records contain many types of data:

- images
- genetics
- survey response
- mobile health
- note text
- diagnostic codes
- lab tests
- outcomes

*Multimodal machine learning*

tries to bring these together

# Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency

**Abstract**—Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors. *Modality* refers to the way in which something happens or is experienced and a research problem is characterized as *multimodal* when it includes multiple such modalities. In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. *Multimodal machine learning* aims to build models that can process and relate information from multiple modalities. It is a vibrant multi-disciplinary field of increasing importance and with extraordinary potential. Instead of focusing on specific multimodal applications, this paper surveys the recent advances in multimodal machine learning itself and presents them in a common taxonomy. We go beyond the typical early and late fusion categorization and identify broader challenges that are faced by multimodal machine learning, namely: representation, translation, alignment, fusion, and co-learning. This new taxonomy will enable researchers to better understand the state of the field and identify directions for future research.

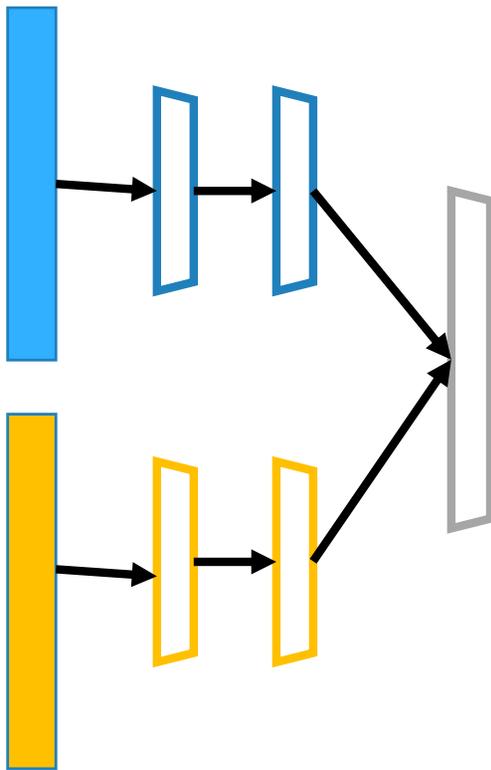
**Index Terms**—Multimodal, machine learning, introductory, survey.

---

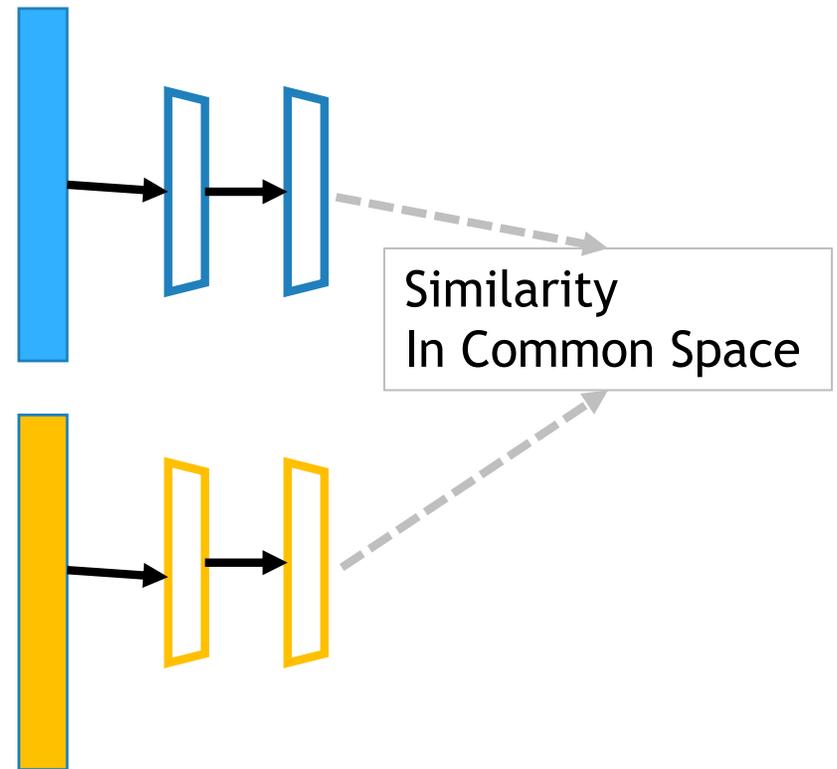
<https://arxiv.org/pdf/1705.09406.pdf>

# Multimodal ML: How to Represent?

Joint/Shared Representation



Coordinated but Separate Representation



# Multimodal ML: How to predict?

- Early fusion
  - concatenate features > feed to standard classifier
  
- Late fusion
  - build separate classifiers > combine with meta classifier

# Example Joint Representation

## EHR Analysis via Deep Poisson Factor Models

### Electronic Health Record Analysis via Deep Poisson Factor Models

**Ricardo Henao**

*Electrical and Computer Engineering Department  
Duke University  
Durham, NC 27708, USA*

R.HENAO@DUKE.EDU

**James T. Lu**

*School of Medicine  
Electrical and Computer Engineering Department  
Duke University  
Durham, NC 27708, USA*

JAMES.LU@DUKE.EDU

**Joseph E. Lucas**

*Electrical and Computer Engineering Department  
Duke University  
Durham, NC 27708, USA*

JOE@STAT.DUKE.EDU

**Jeffrey Ferranti**

*School of Medicine  
Duke University  
Durham, NC 27708, USA*

JEFFREY.FERRANTI@DM.DUKE.EDU

**Lawrence Carin**

*Electrical and Computer Engineering Department  
Duke University  
Durham, NC 27708, USA*

LCARIN@DUKE.EDU

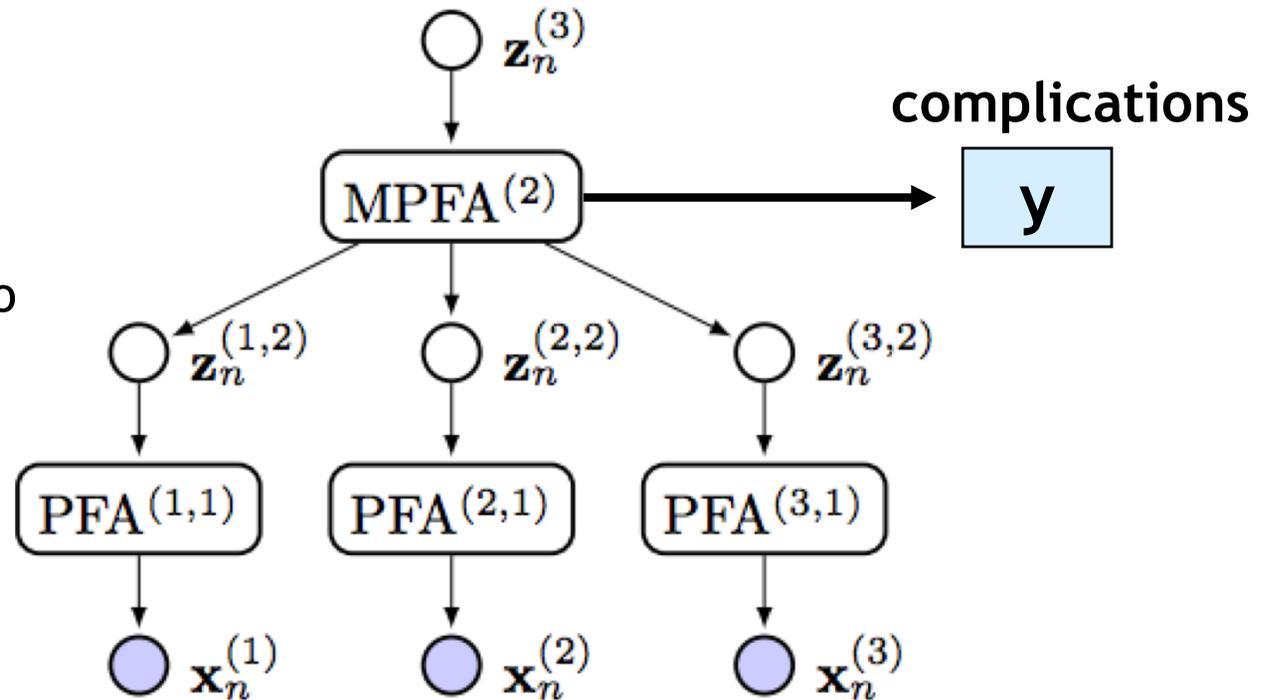
# DMPFA: a Hierarchical Topic Model

16,756 patients  
With diabetes

Can we capture joint structure that leads to different EHR data?

Can we predict complications like:

- cardiovascular disease?



- ...
- (13 total)

meds  
RXNORM  
size= 1694  
253

lab tests  
LOINC  
size=4391  
606

diagnoses/procedures  
ICD/CPT  
size=21,305  
4,222

# Joint Representation Leads to Better Generative Model

	treat each modality separate			concatenate to one big vector			DMPFA shared top-level representation		
Size	Med	Lab	Code	Med	Lab	Code	Med	Lab	Code
64–32	1.930	76.724	210.690	1.930	76.575	208.785	1.865	72.919	194.260
96–48	1.851	76.736	192.851	1.825	76.787	193.782	1.788	72.662	176.737
128–64	1.803	76.538	182.803	1.759	76.495	182.049	1.748	72.415	167.423
64–32–16	1.918	76.648	207.932	1.911	76.400	209.652	1.861	72.773	191.854
96–48–24	1.822	76.967	192.530	1.816	76.660	192.505	1.759	72.531	176.451
128–64–32	1.787	76.556	182.365	1.764	76.528	180.806	<b>1.730</b>	<b>72.364</b>	<b>166.759</b>

Negative heldout likelihood (lower is better)

# Joint Repr. Gives Better Predictions

	treat each modality separate			concatenate to one big vector	DMPFA shared top-level representation
Size	Med	Lab	Code	All	All
64-32	0.592±0.05	0.594±0.05	0.745±0.06	0.751±0.06	0.771±0.07
96-48	0.596±0.04	0.583±0.05	0.727±0.06	0.750±0.06	0.781±0.06
128-64	0.590±0.04	0.590±0.05	0.725±0.06	0.751±0.06	0.779±0.06
64-32-16	0.601±0.05	0.594±0.05	0.726±0.05	0.742±0.06	0.771±0.06
96-48-24	0.587±0.04	0.588±0.06	0.735±0.06	0.758±0.07	<b>0.785±0.07</b>
128-64-32	0.590±0.04	0.588±0.05	0.732±0.05	0.757±0.06	<b>0.784±0.07</b>

Outcome

---

Acute Myocardial Infarction  
 Amputation  
 Cardiac Catheterization  
 Coronary Artery Disease  
 Depression  
 Heart Failure  
 Kidney Disease  
 Neurological Diseases  
 Obesity  
 Ophthalmic Disease  
 Stroke  
 Unstable Angina

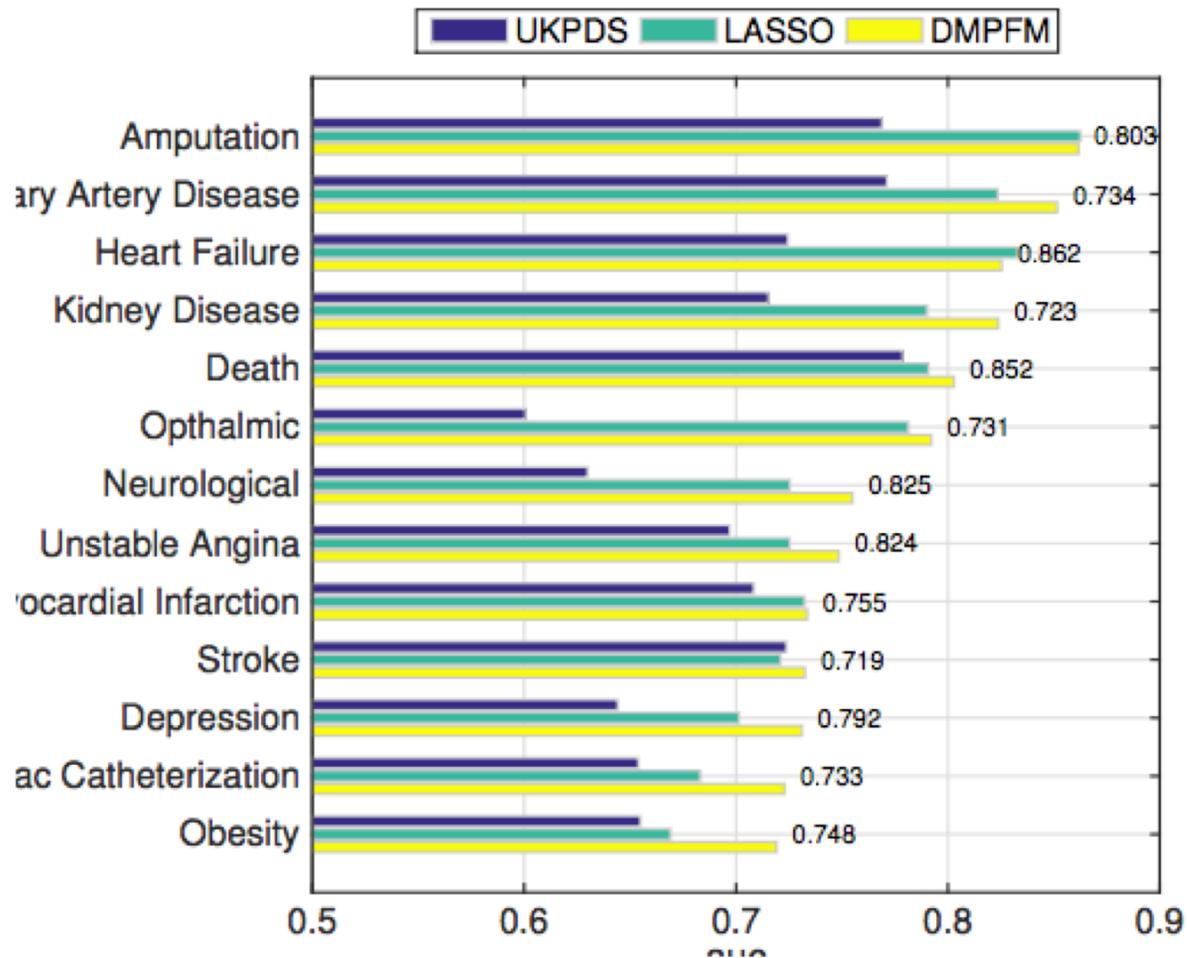
---

Death

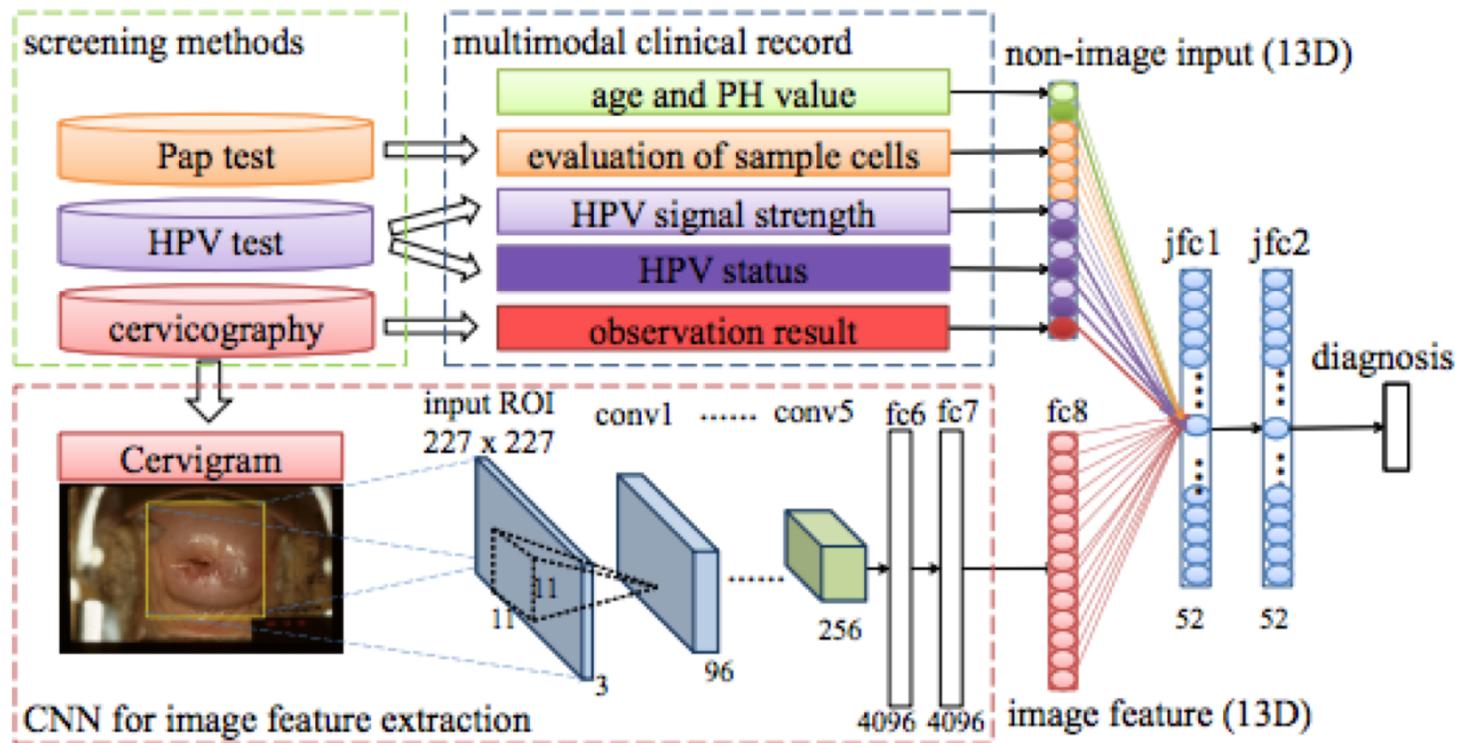
Avg. AUC across 13 outcomes  
 (higher is better)

*Model trained “discriminatively”  
 likelihoods for both data and labels*

# Better Predictions than Baselines



# Another Joint Repr. Example: Multimodal learning for Cervical Cancer Diagnosis



*Credit: Xu et al. MICCAI 2016*

# Joint Representations w/ Time series

Recurrent Attentive and Intensive Model  
*Xu et al. KDD 2018*

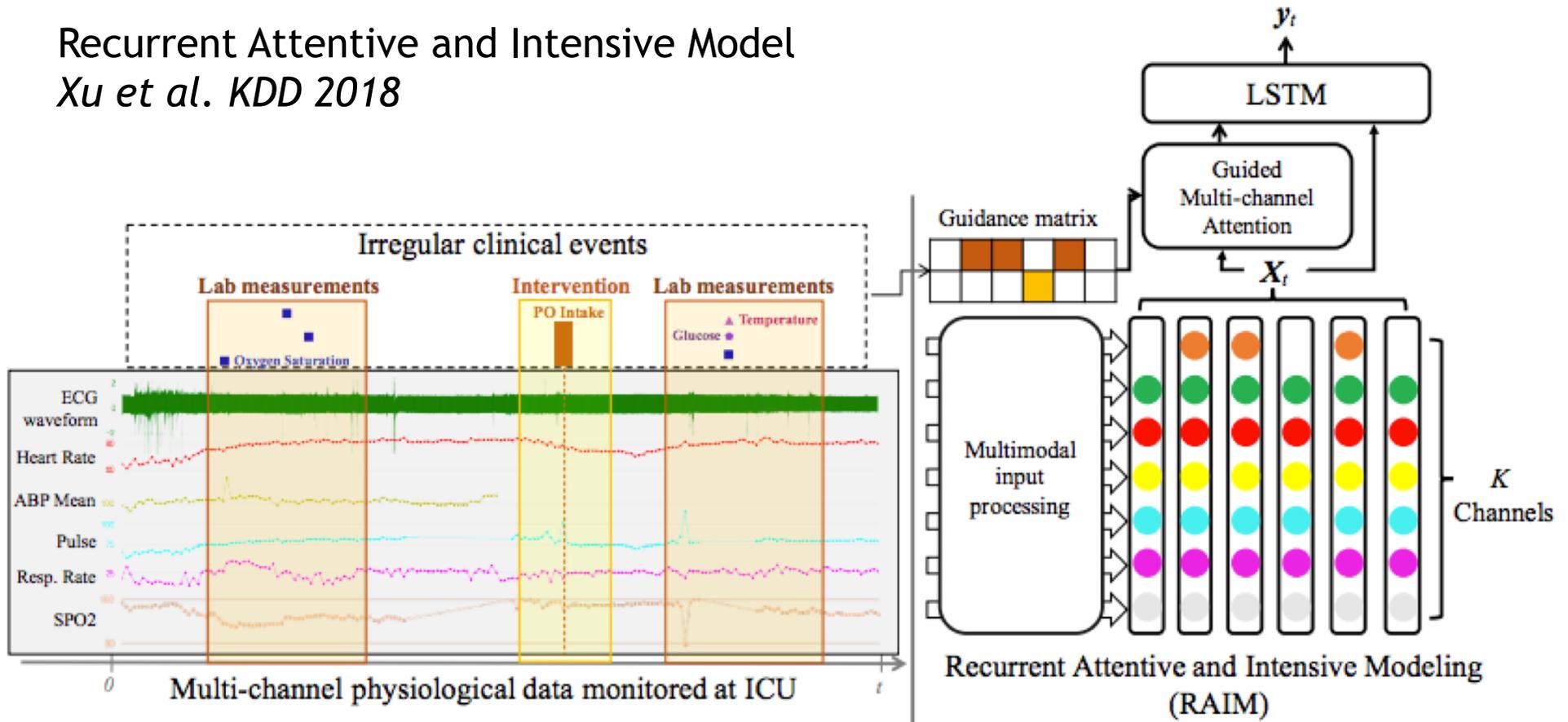
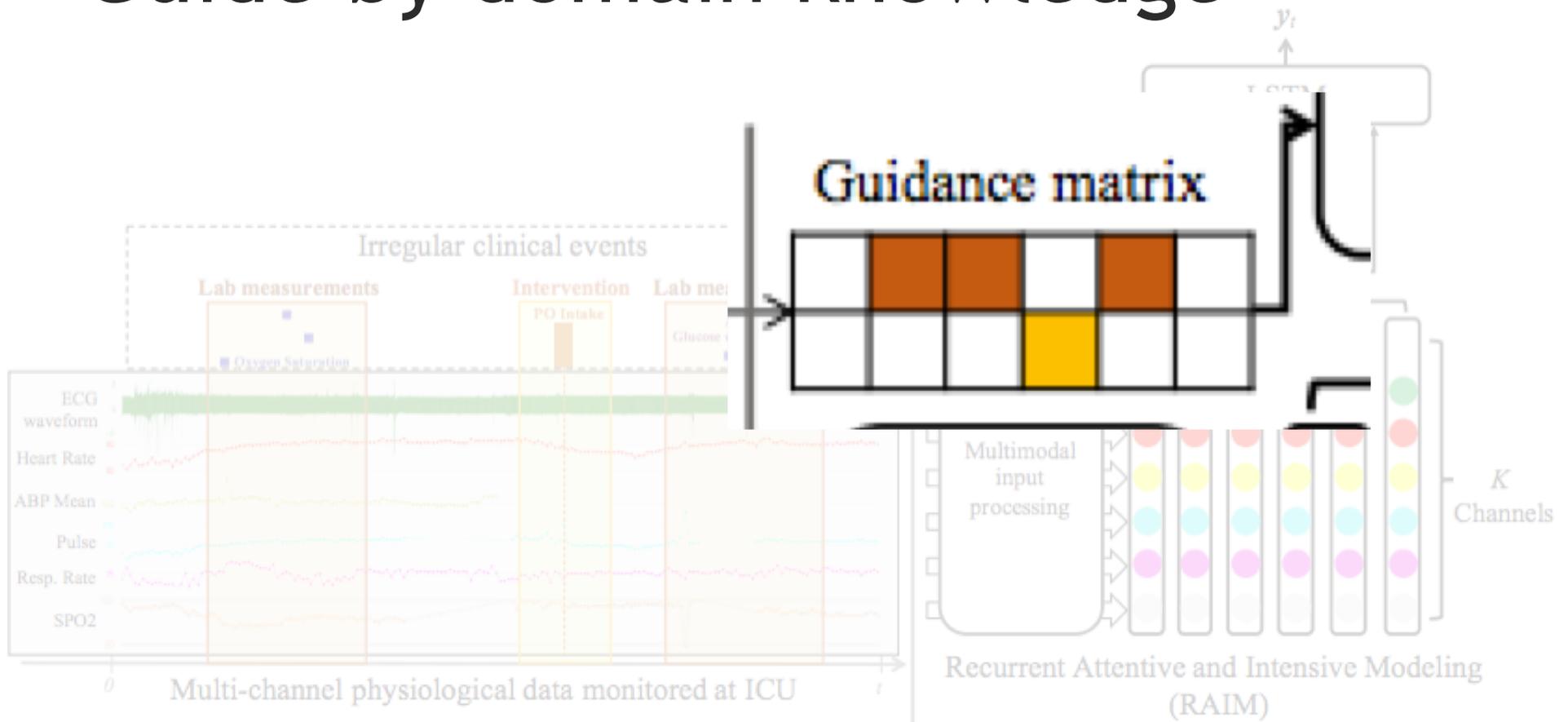
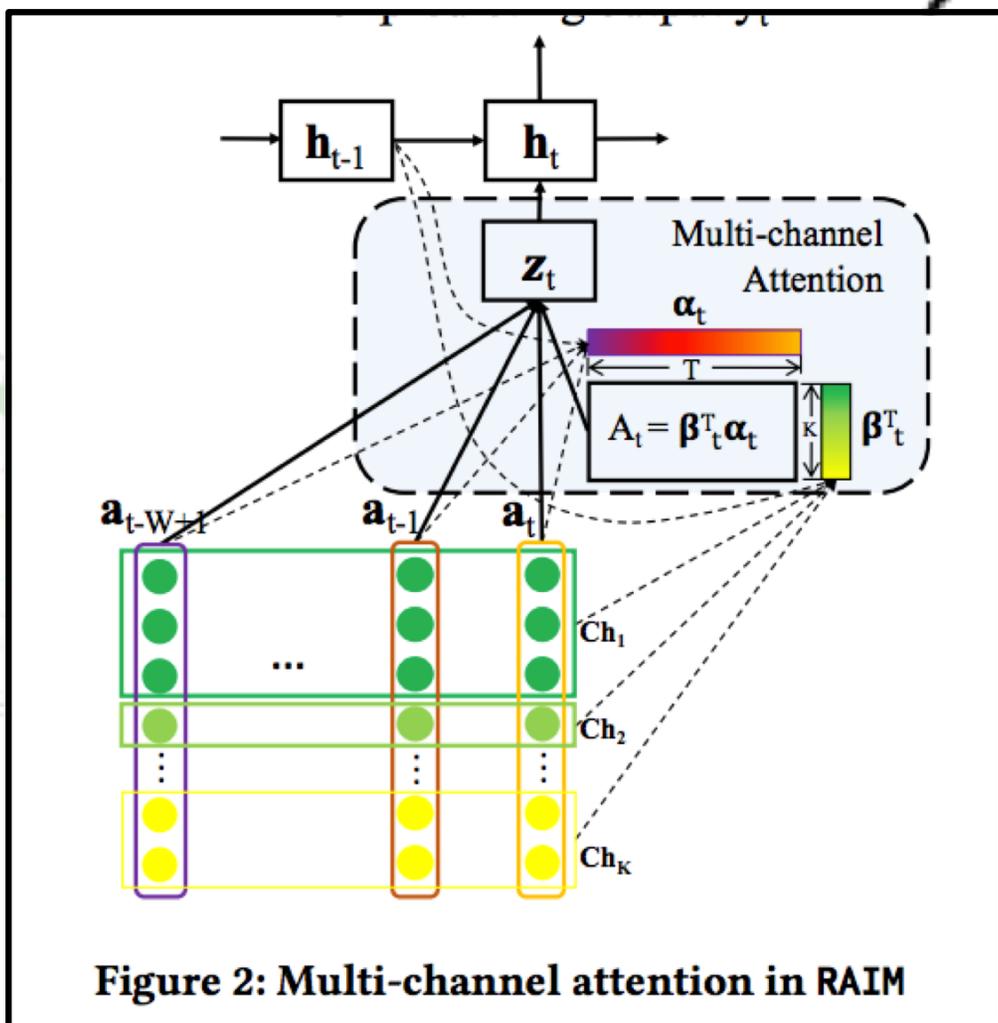
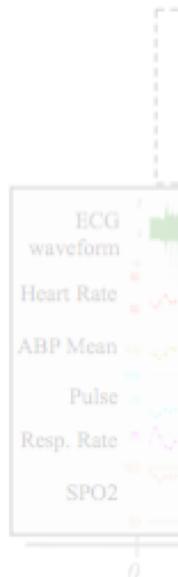


Figure 1: An overview of RAIM on multimodal continuous patient monitoring data.

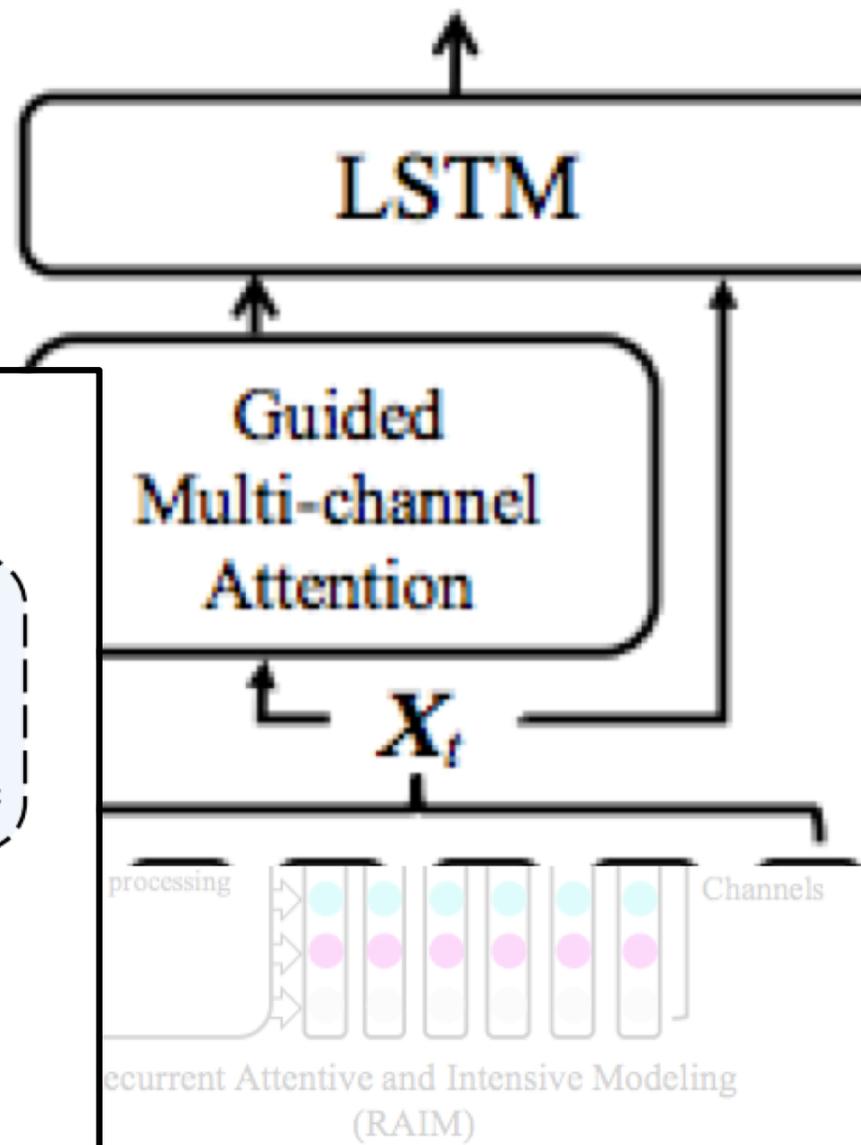
# Guide by domain knowledge



Top row: binary indicator of when labs ordered  
Bottom: binary indicator of when interventions ordered



**Figure 2: Multi-channel attention in RAIM**



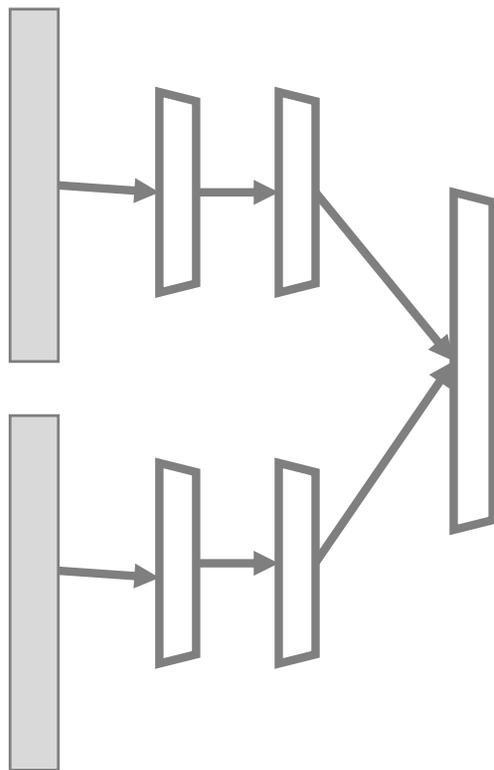
- At each timestep:
- Which channels matter?
  - Which previous times matter?

# RAIM Predictive Performance

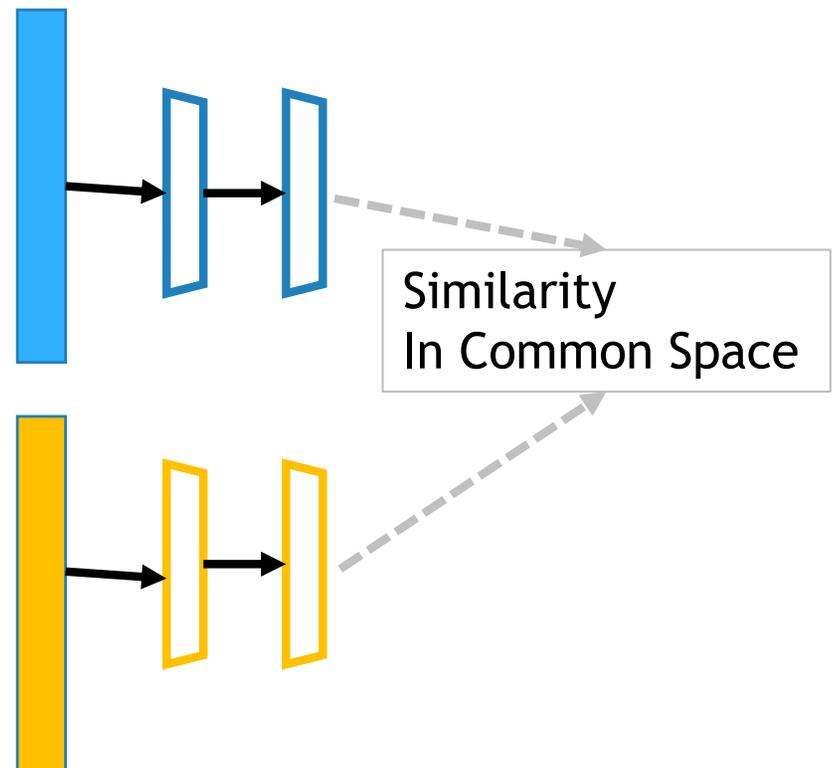
	Decompensation			Length of Stay	
	AUC-ROC	AUC-PR	Accuracy	Kappa	Accuracy
CNN (ECG)	87.84%	21.56%	88.38%	0.7681	82.16%
CNN-RNN	87.45%	23.19%	88.25%	0.8027	85.34%
CNN-AttRNN	88.19%	25.81%	89.28%	0.8186	84.89%
RAIM-0	87.81%	25.56%	88.96%	0.8125	85.84%
RAIM-1	88.25%	25.61%	88.91%	0.8215	86.74%
RAIM-2	88.77%	26.85%	90.27%	0.8217	85.21%
RAIM-3	<b>90.18%</b>	<b>27.93%</b>	<b>90.89%</b>	<b>0.8291</b>	<b>86.82%</b>

# Multimodal ML: How to Represent?

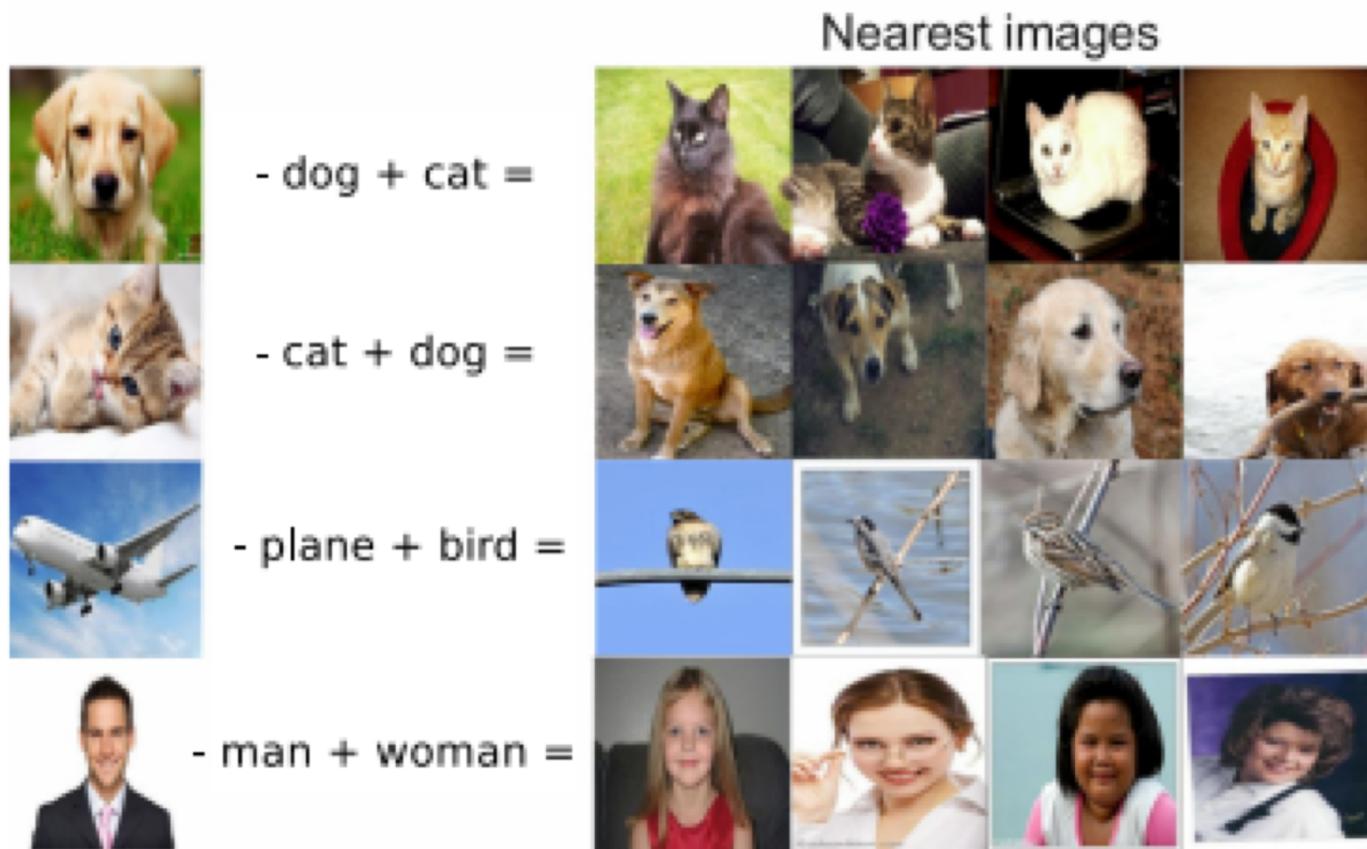
Joint/Shared Representation



Coordinated but Separate Representation

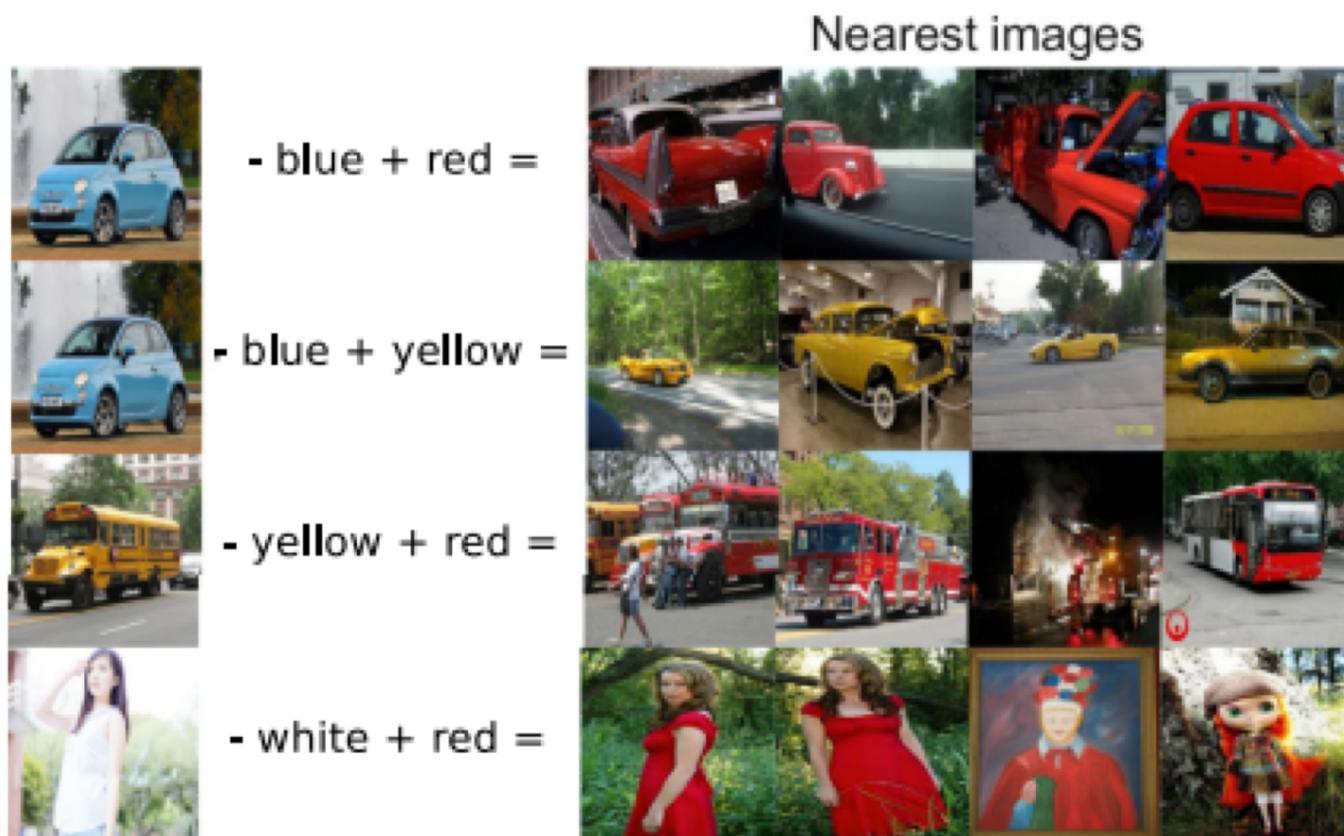


# Example Coordinated Embedding



*Credit: Kiros, Saludinakov, Zemel DLWorkshop@NIPS 2014*

# Example Coordinated Embedding



*Credit: Kiros, Saludinakov, Zemel DLWorkshop@NIPS 2014*



# Takeaway: Multimodal Representations

- When to use joint/shared repr.?
  - When you have many modalities
    - Easy to scale: linear with number of modalities
  - When you want a generative model
- When to use coordinated-but-separate repr.?
  - For two key modalities. More is hard.
    - Too many pairs to coordinate!
  - Inspecting how modalities are related
    - analogies

# Takeaway: Multimodal predictions

- When to use early fusion?
  - Low-level interactions between modalities useful
  - Training data scarce
- When to use late fusion?
  - Hard to access all raw modalities
  - Low-level interactions between modalities
- When to use end-to-end learned representation?
  - Large training set and validation set available

## MLHC Challenge 4

# Interpretability

*Interpretable machine learning*  
helps humans understand model predictions

Also called “Explainable AI”

# Position Papers

Towards A Rigorous Science of Interpretable Machine Learning

Finale Doshi-Velez\* and Been Kim\*

---

**The Mythos of Model Interpretability**

---

Zachary C. Lipton <sup>1</sup>

# Challenges

- What does "interpretability" mean?
  - How do I measure it?

# Approaches

Use models with understandable internals

Use complicated model, interpret post-hoc

*Train* deep model to get max “interpretability”

# SLIM: Super-sparse Linear Integer Models

$$\min_{[w_1 \dots w_D]} \frac{1}{N} \sum_{n=1}^N \text{loss}(y_n, \sum_{d=1}^D w_d x_{nd}) + \sum_d \text{is-non-zero}(w_d)$$

subject to :  $w_d \in \{-10, -9, -8, \dots, 0, \dots, 8, 9, 10\}$

Operational constraints

- SIZE: Use at most 5 features
- SIGN: Obey established relationships for individual features
- LOW FPR: Do not produce too many false positives

# SLIM for Sleep Apnea

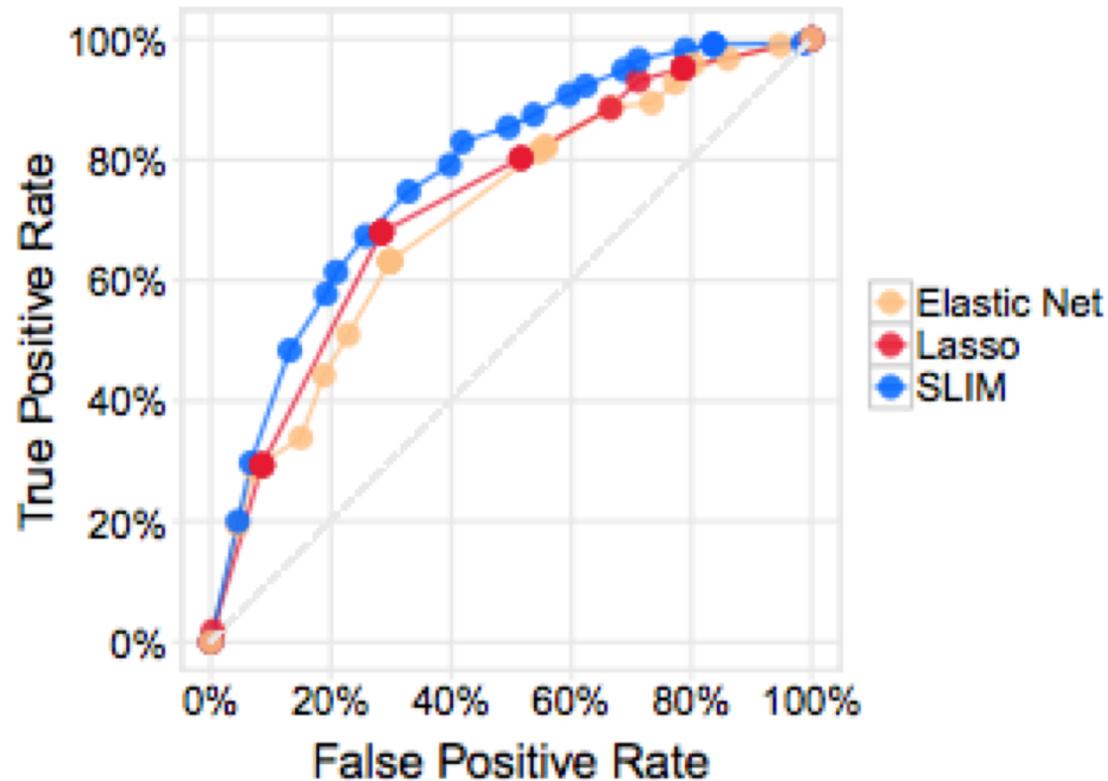
PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1.	<i>age</i> $\geq$ 60	4 points	.....
2.	<i>hypertension</i>	4 points	+ .....
3.	<i>body mass index</i> $\geq$ 30	2 points	+ .....
4.	<i>body mass index</i> $\geq$ 40	2 points	+ .....
5.	<i>female</i>	-6 points	+ .....
<b>ADD POINTS FROM ROWS 1 – 5</b>		<b>SCORE</b>	<b>= .....</b>

**Fig. 8:** SLIM scoring system for sleep apnea screening. This model achieves a 10-CV mean test TPR/FPR of 61.4/20.9%, obeys all operational constraints, and was trained without parameter tuning. It also generalizes well due to the simplicity of the hypothesis space: here the training TPR/FPR of the final model is 62.0/19.6%.

*Credit: Ustun & Rudin Machine Learn. 2016*

# SLIM for Sleep Apnea



*Credit: Ustun & Rudin Machine Learn. 2016*

Use models with understandable internals

**Use complicated model, interpret post-hoc**

Train deep model to get max “interpretability”

# How to interpret a fixed deep model?

*Olah et al. 2017*  
*Google Brain*



## Dataset

Examples show us what neurons respond to in practice



## Optimization

isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.



Baseball—or stripes?  
*mixed4a, Unit 6*



Animal faces—or snouts?  
*mixed4a, Unit 240*



Clouds—or fluffiness?  
*mixed4a, Unit 453*

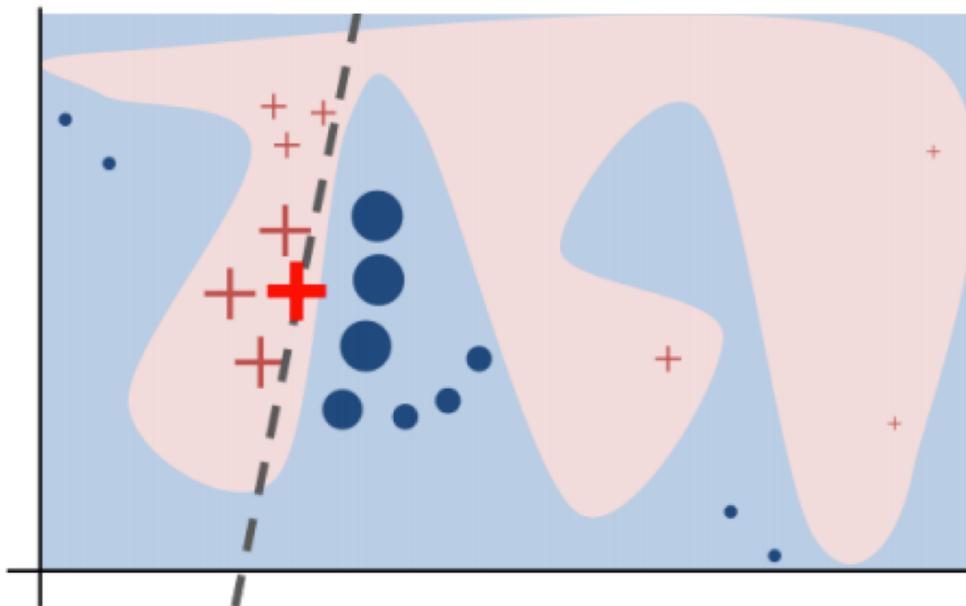
# LIME: Local fit of linear model

## “Why Should I Trust You?” Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro  
University of Washington  
Seattle, WA 98105, USA  
marcotcr@cs.uw.edu

Sameer Singh  
University of Washington  
Seattle, WA 98105, USA  
sameer@cs.uw.edu

Carlos Guestrin  
University of Washington  
Seattle, WA 98105, USA  
guestrin@cs.uw.edu

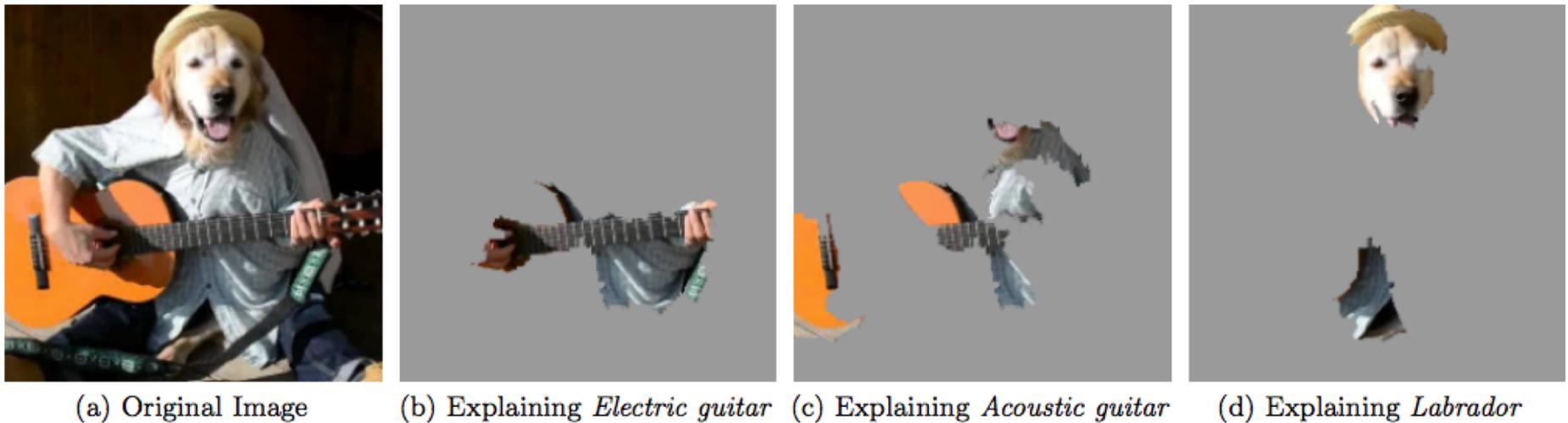


Model may have complex boundaries, but for specific example the boundary looks locally linear.

Local linear fit can be sparse (L1-reg. regression)

*Credit: Ribeiro et al. KDD 2016*

# Sparsity of post-hoc linear model indicates relevant features



**Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ( $p = 0.32$ ), "Acoustic guitar" ( $p = 0.24$ ) and "Labrador" ( $p = 0.21$ )**

*Credit: Ribeiro et al. KDD 2016*

Use models with understandable internals

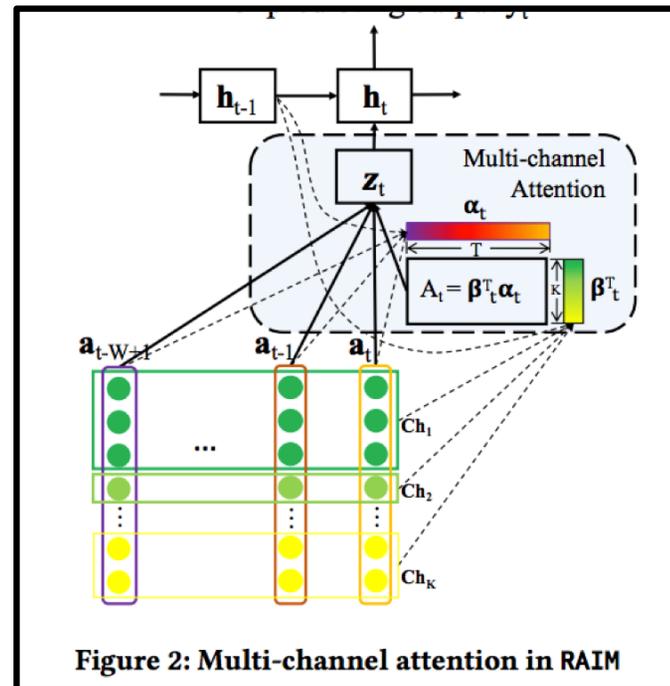
Use complicated model and interpret post-hoc

***Train*** deep model to get max “interpretability”

# Attention mechanisms

Tell you what part of input model “looks at”

NOT the same thing as “why”



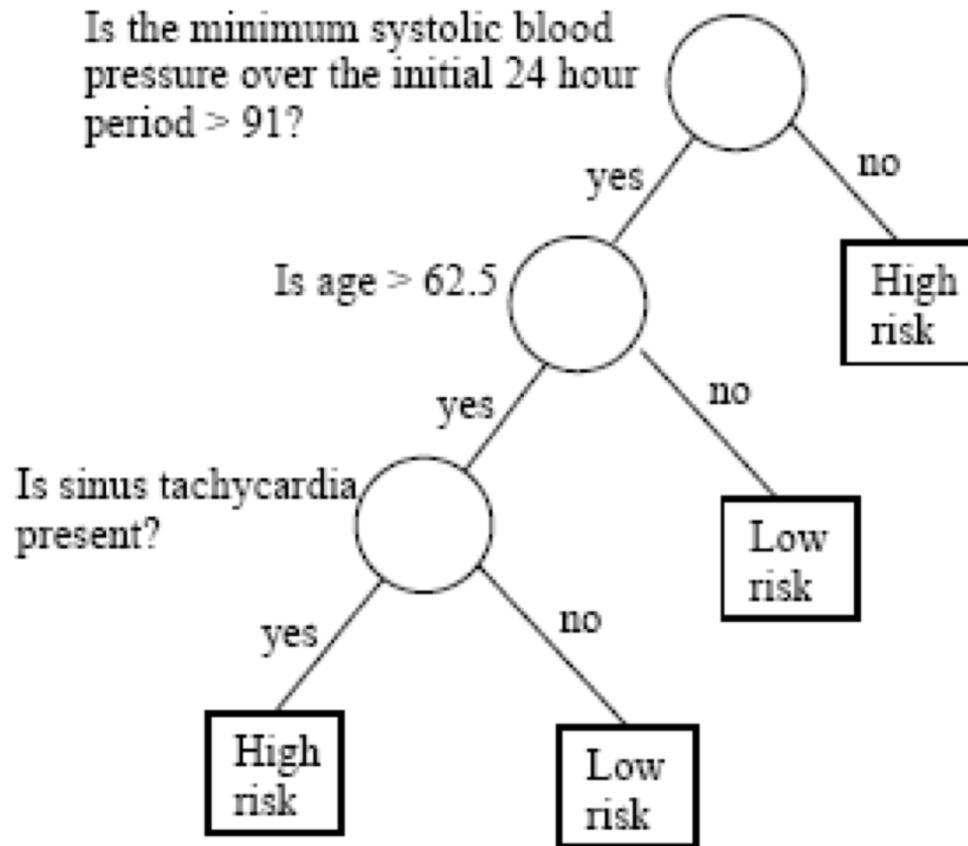
# Interpretability → Simulatability

**Def:** a model is **simulatable** if a human can “take input data together with the parameters of the model and in *reasonable* time step through *every* calculation required to produce a prediction” - *Lipton 2016*

## Advantages of simulation

- check each step against expert knowledge
- check predictions at counter-factual inputs
  - *what if the blood pressure was lower?*
- identify dataset biases / causal leakage / etc

# Decision Trees are Simulatable



decent predictions, but definitely inferior to modern deep methods

## Beyond Sparsity: Tree Regularization of Deep Models for Interpretability

**Mike Wu<sup>1</sup>, Michael C. Hughes<sup>2</sup>, Sonali Parbhoo<sup>3</sup>,  
Maurizio Zazzi<sup>4</sup>, Volker Roth<sup>3</sup>, and Finale Doshi-Velez<sup>2</sup>**

<sup>1</sup>Stanford University, wumike@cs.stanford.edu

<sup>2</sup>Harvard University SEAS, mike@michaelchughes.com, finale@seas.harvard.edu

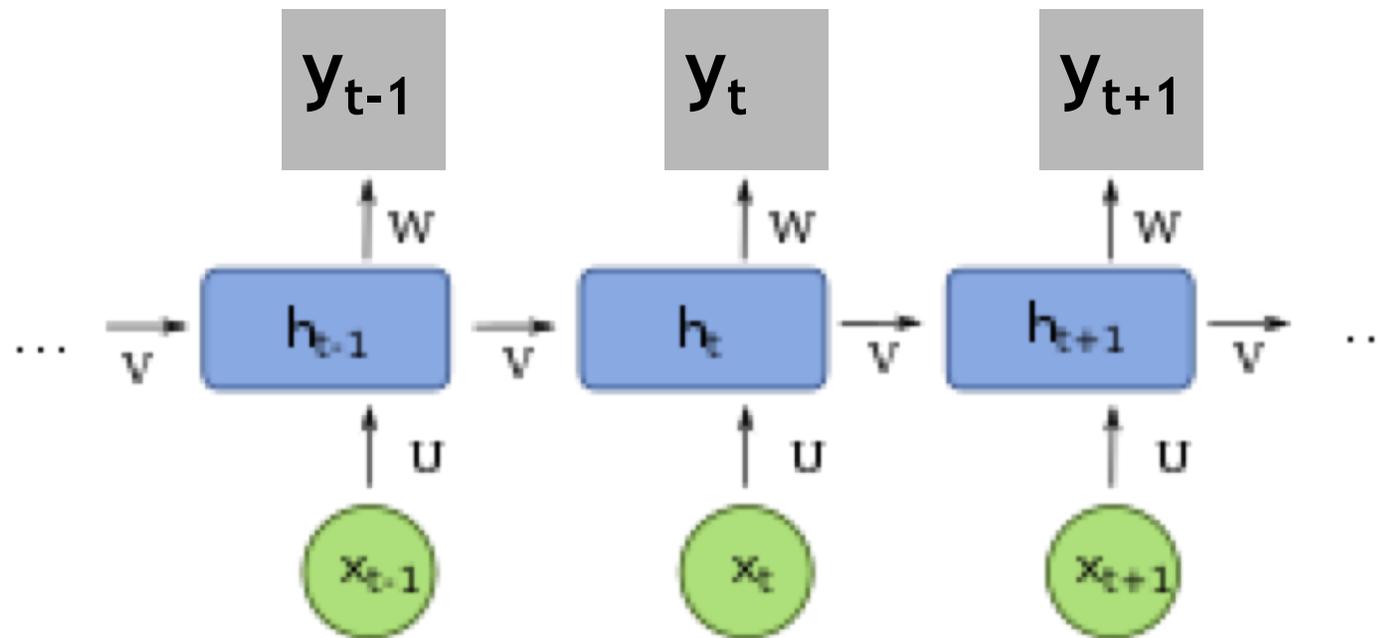
<sup>3</sup>University of Basel, {sonali.parbhoo,volker.roth}@unibas.ch

<sup>4</sup>University of Siena, maurizio.zazzi@unisi.it

*Wu et al. AAAI 2018*

Can we *optimize* RNNs so that their decision boundaries are easily explained by small decision trees?

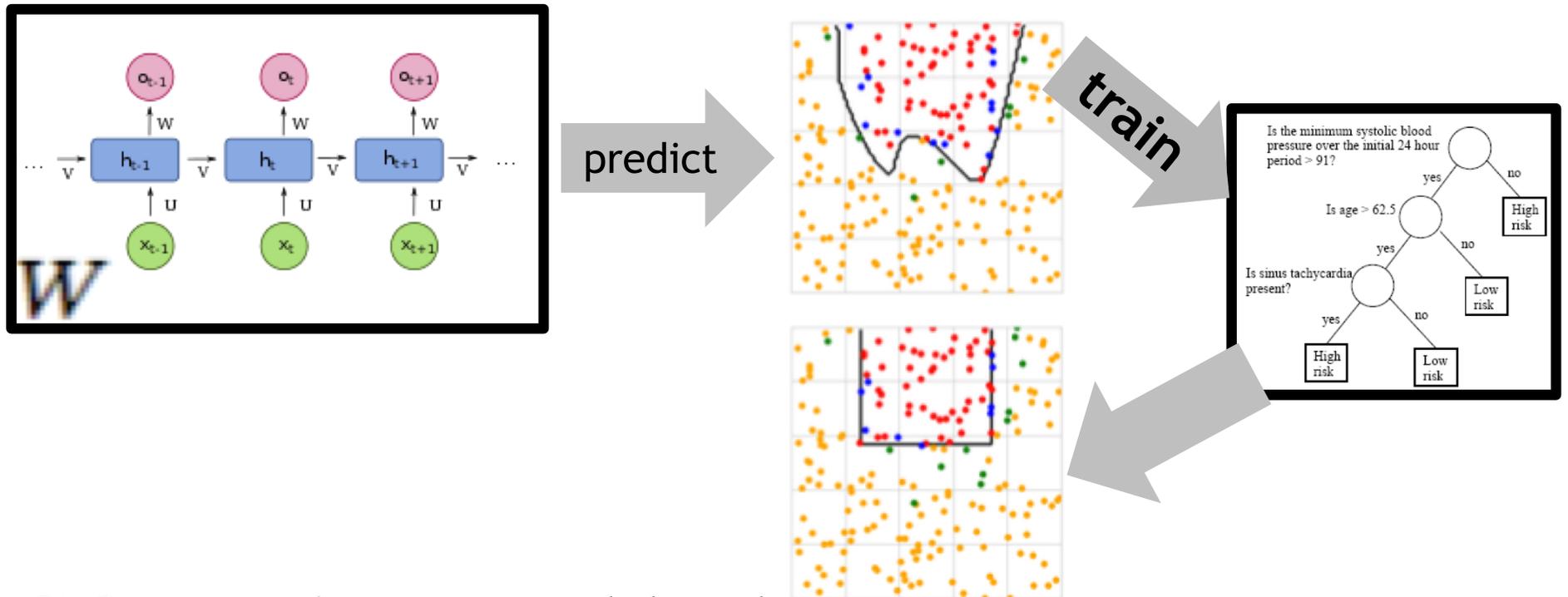
# Model: Recurrent Neural Net



$$\min_W \lambda \Psi(W) + \sum_{n=1}^N \text{loss}(y_n, \hat{y}_n(x_n, W))$$

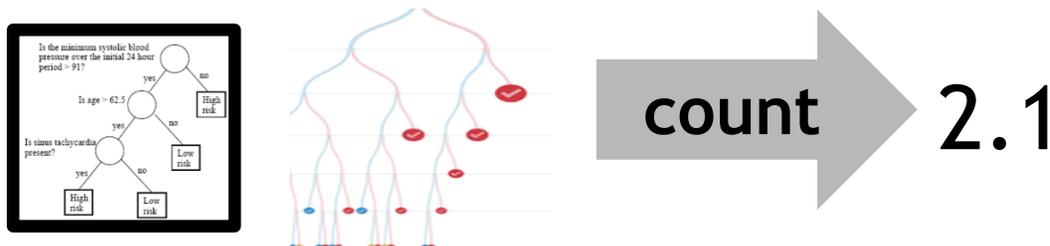
# How to measure simulatability of deep models?

1) Train tree to match the predictions of a deep model

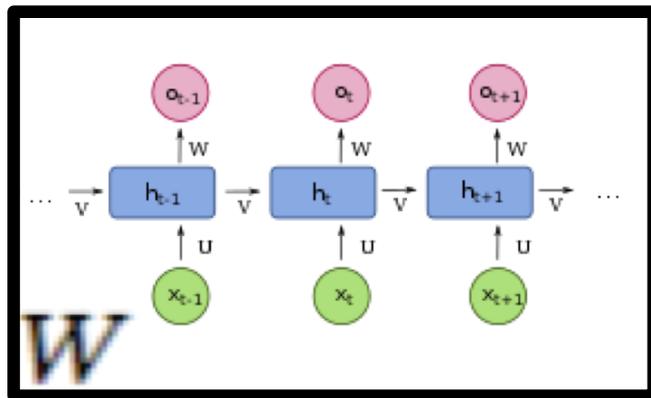


2) Count tree's average path length

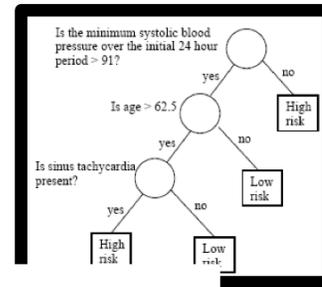
= cost of simulating the average input example



# Tree Regularization: Penalize deep model's (lack of) simulatability



train



count

2.1

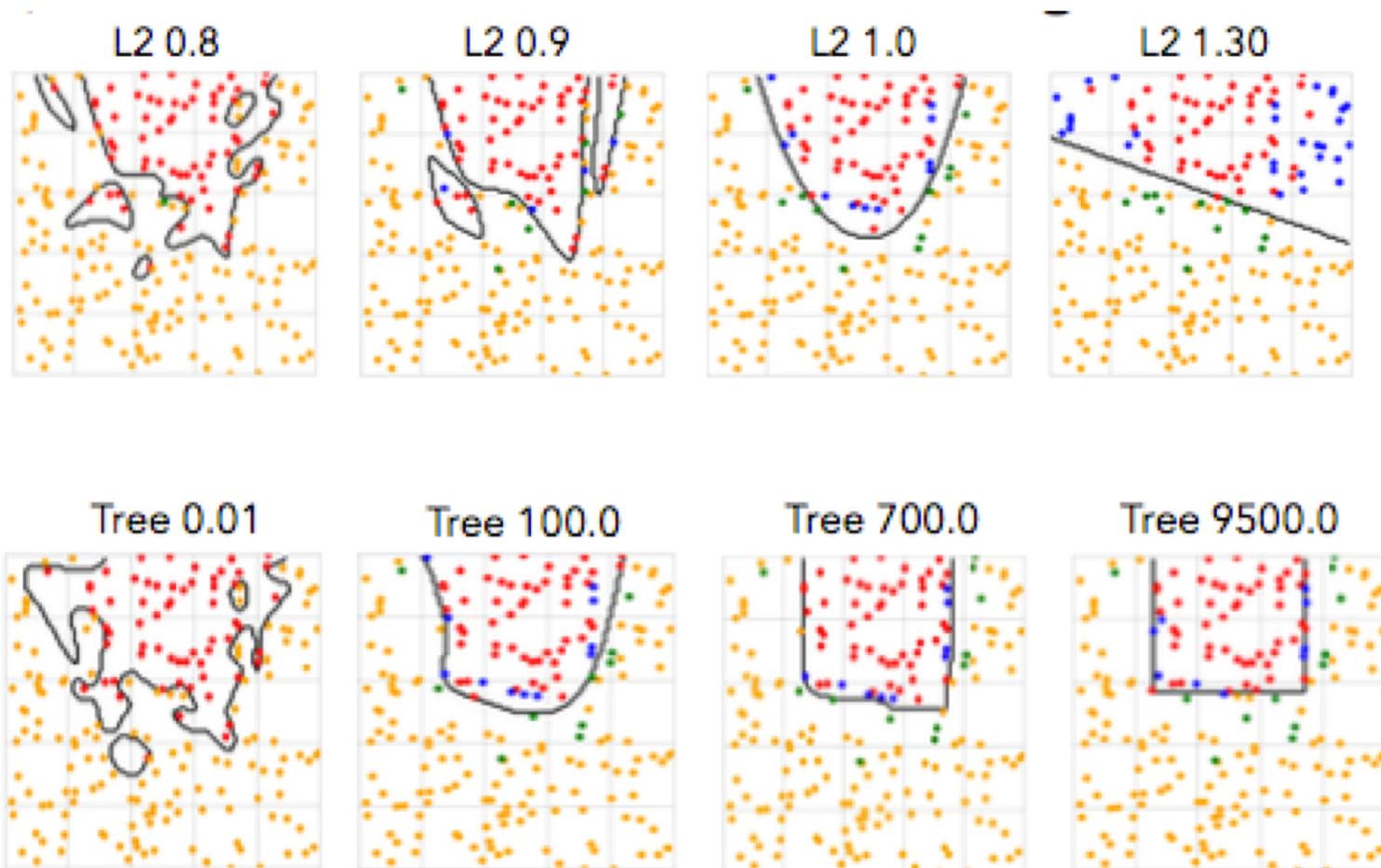
surrogate MLP

2.2

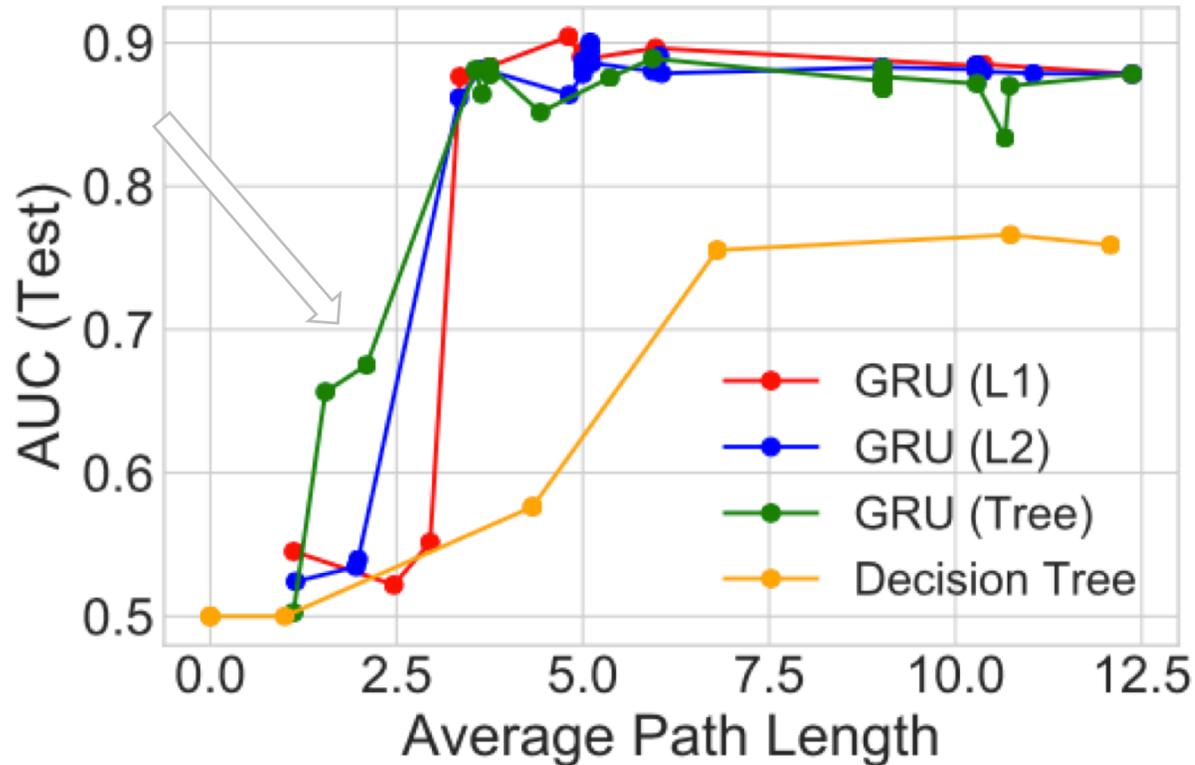
How to train:

- Model step: given fixed **surrogate MLP**, update  $W$  via gradient
- Surrogate step: given fixed  $W$ , retrain the **surrogate MLP**

# Tree Regularization: tree-like decision boundaries for deep models



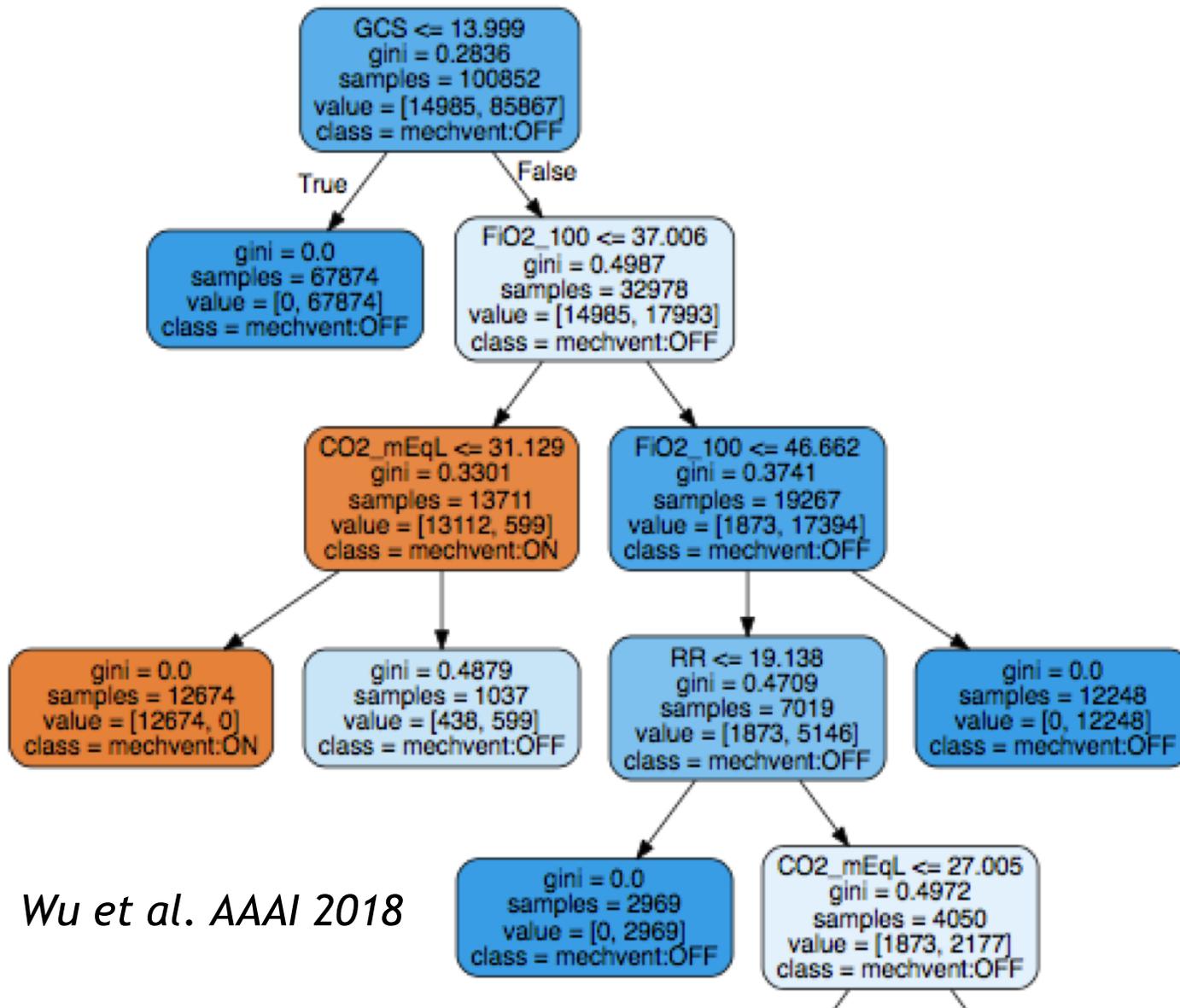
Tree-reg. finds sweet spot  
high AUC & low path length



### (c) Mechanical Ventilation

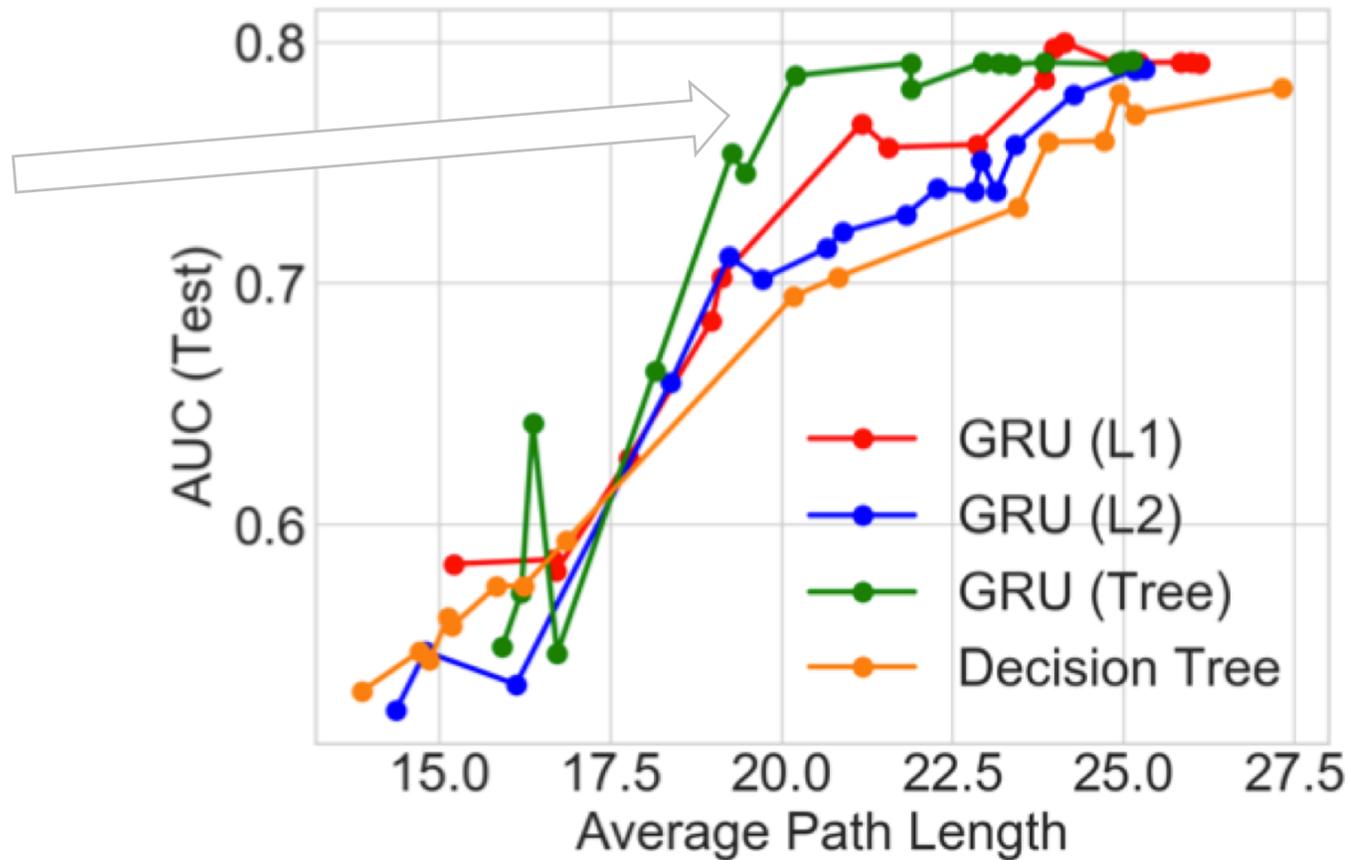
Wu et al. AAI 2018

# Tree Proxy for Mech.Vent.



Wu et al. AAAI 2018

# Tree reg. finds sweet spot



(c) HIV Therapy Adherence

Wu et al. AAI 2018

# Takeaways: Interpretability

Really ask why you need interpretability

Define precise notion:

- Sparse model with few coefficients?
- Human simulatable?

Find precise, domain-specific evaluation

# MLHC Challenge 5

# Causality

Possible goals of a personalized medicine strategy:

Individual treatment effect

*Would the patient's symptoms be reduced by drug A?*

Average treatment effect

*Would the average patient benefit if we prescribed drug A?*

These “What if?” questions aren't possible with supervised learning

# A Provocative Challenge

Theoretical Impediments to Machine Learning  
With Seven Sparks from the Causal Revolution

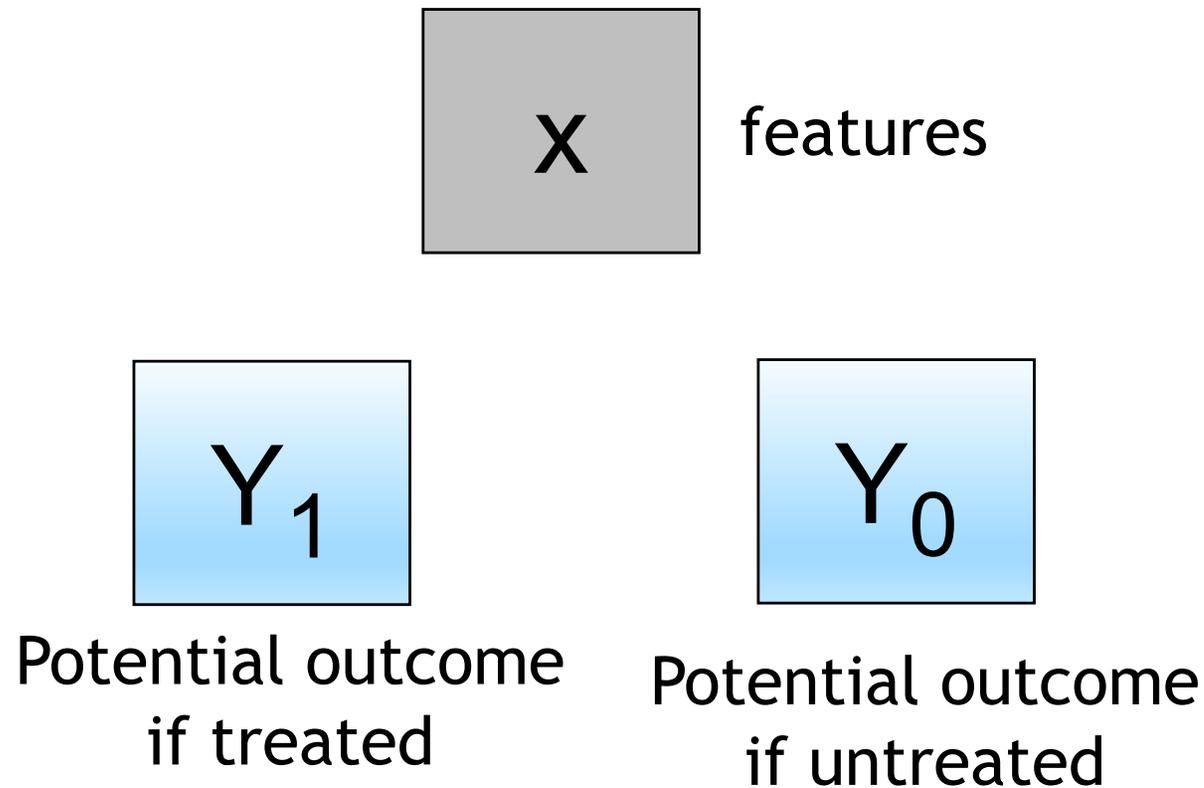
Judea Pearl  
University of California, Los Angeles  
Computer Science Department  
Los Angeles, CA, 90095-1596, USA  
judea@cs.ucla.edu

July 11, 2018

Big Question: Can we do *anything* with observational data?

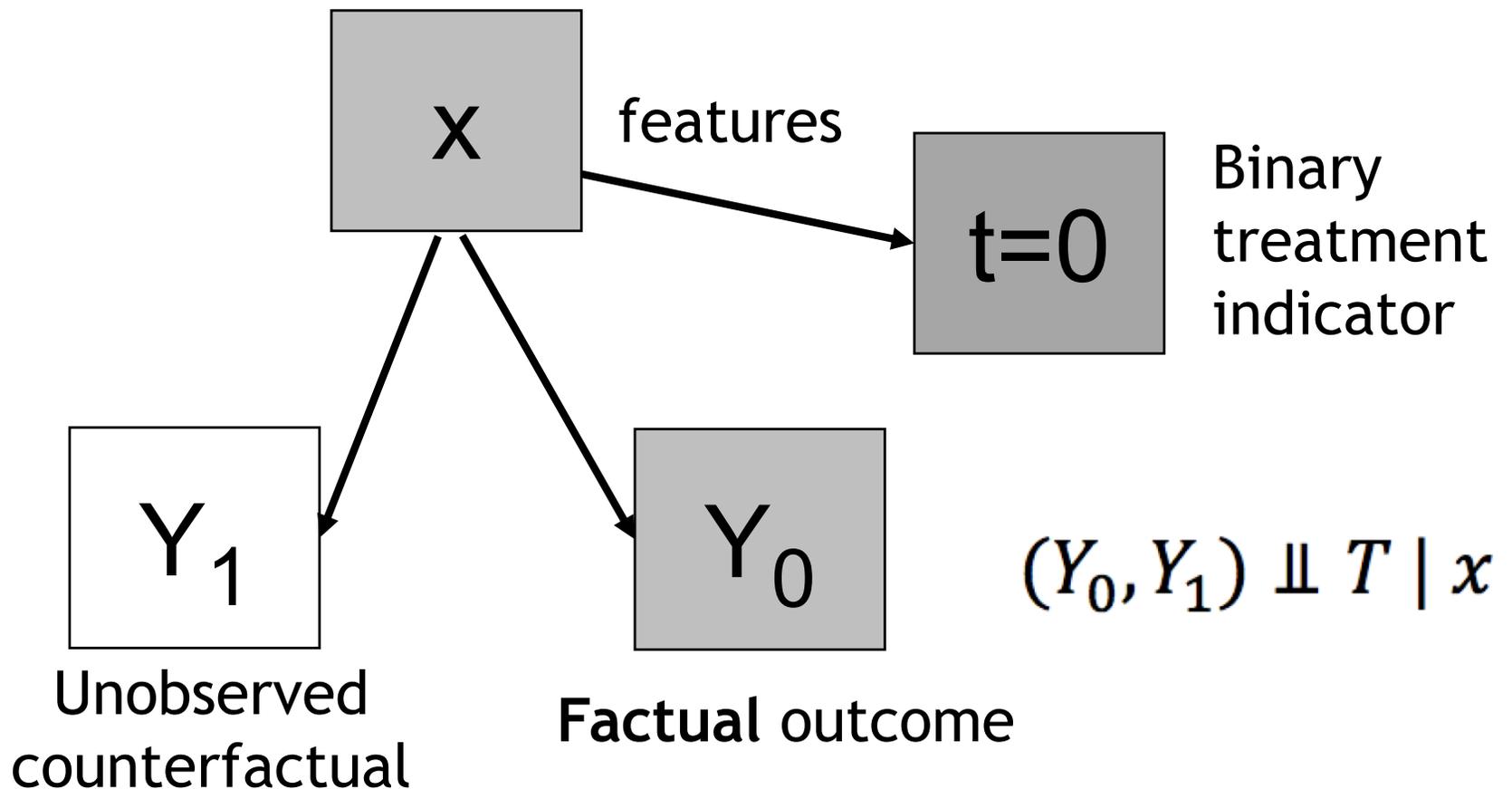
# Potential Outcomes Framework

Neyman-Rubin model (Rubin 2011, Neyman 1923)



# Potential Outcomes Framework

Neyman-Rubin model (Rubin 2011, Neyman 1923)

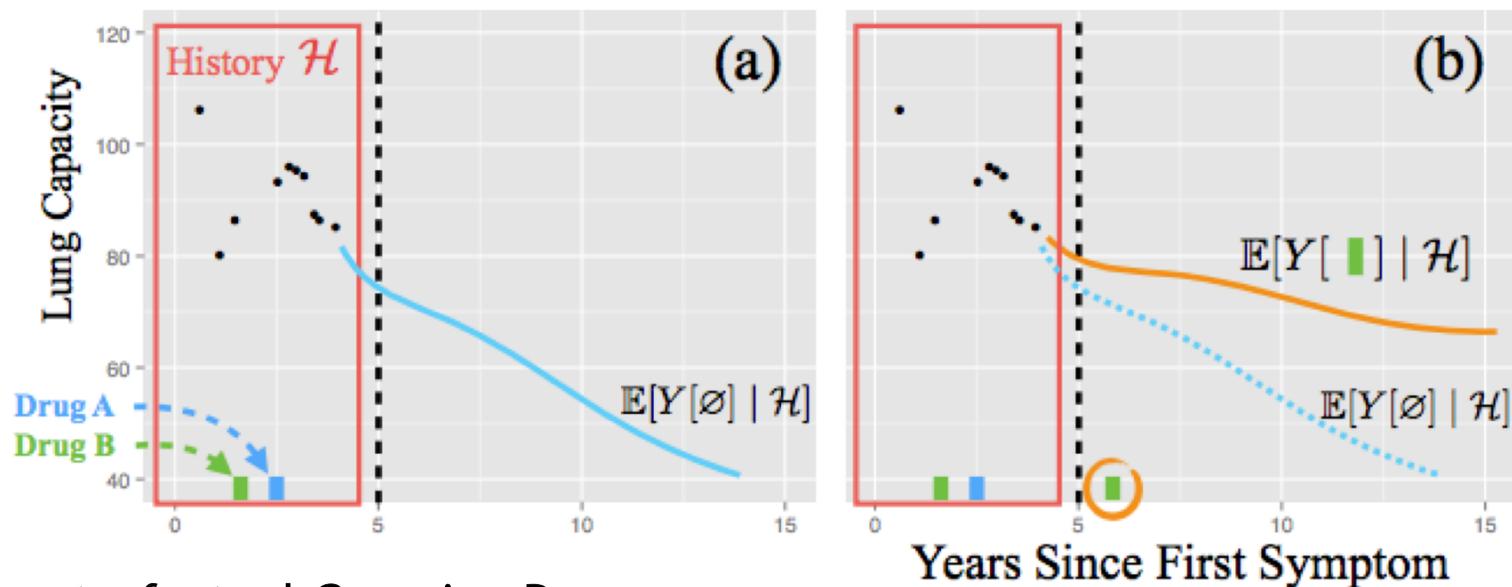


# Assumptions for Neyman-Rubin Framework

- Assumption 1: Common support
  - No set of patient features leads to ZERO probability of treatment (or non-treatment)
- Assumption 2: No unmeasured confounders
  - Also called *conditional ignorability*
  - Treatments and potential outcomes are conditionally independent given features  $X$

These results suggest a number of new questions and directions for future work. First, the validity of the CGP is conditioned upon a set of assumptions (this is true for all counterfactual models). In general, **these assumptions are not testable**. The reliability of approaches using counterfactual models therefore critically depends on the plausibility of those assumptions in light of domain knowledge.

# Paper: Schulam & Saria NIPS 2017



Counterfactual Gaussian Process

Assumes measurable risk score, can be tracked over time

Goal: Model future trajectory of risk score given history

Outcomes are measured and actions are taken at irregular, discrete points in continuous-time

# Compare CGP to supervised learning

	Collect data under treatment policy A		Collect data under treatment policy B	
	Regime A		Regime B	
	Baseline GP	CGP	Baseline GP	CGP
AUC	0.853	0.872	0.832	0.872

Predictions from CGP are same regardless of the policy used to collect data

Baselines are unduly influenced by the observed treatments

# Compare CGP to supervised learning

Collect data under treatment policy C which violates assumptions

	Regime <i>A</i>		Regime <i>B</i>		Regime <i>C</i>	
	Baseline GP	CGP	Baseline GP	CGP	Baseline GP	CGP
AUC	0.853	0.872	0.832	0.872	0.806	0.829

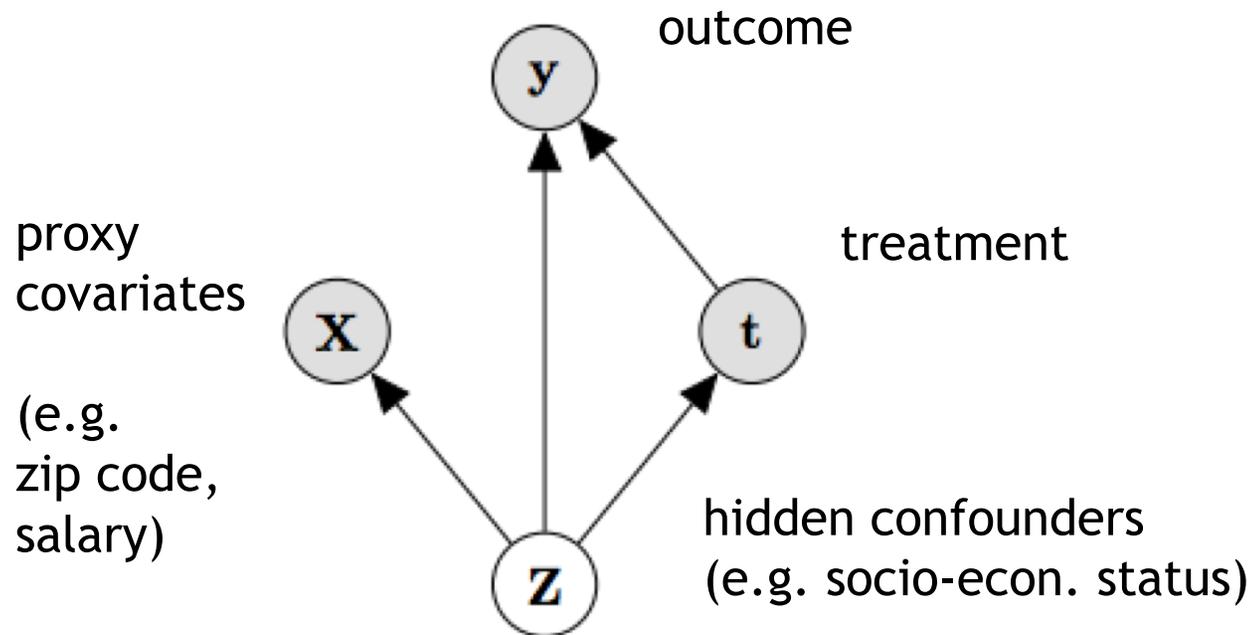
When assumptions are violated, predictions become unreliable!

# Paper:

## Causal Effect Inference with Deep Latent-Variable Models

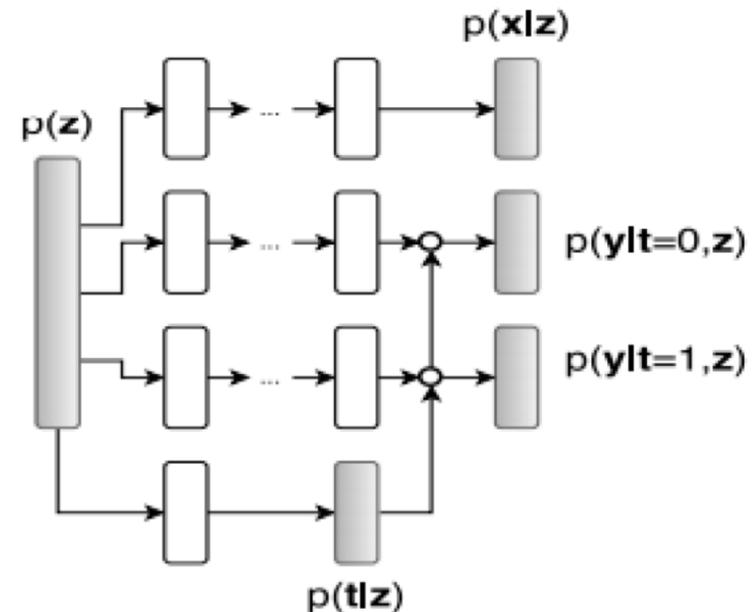
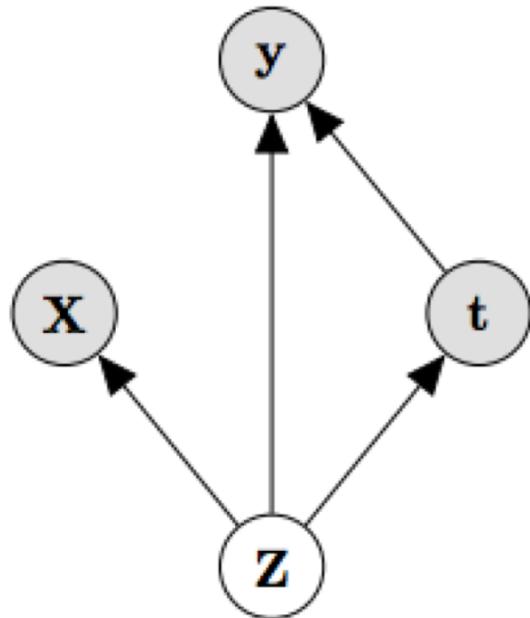
Credit: Louizos et al. NIPS '17

- How to do causal inference when we have imperfect views (proxies) of confounders?



# Approach: Model joint $p(z, x, t, y)$

Credit: Louizos et al. NIPS '17

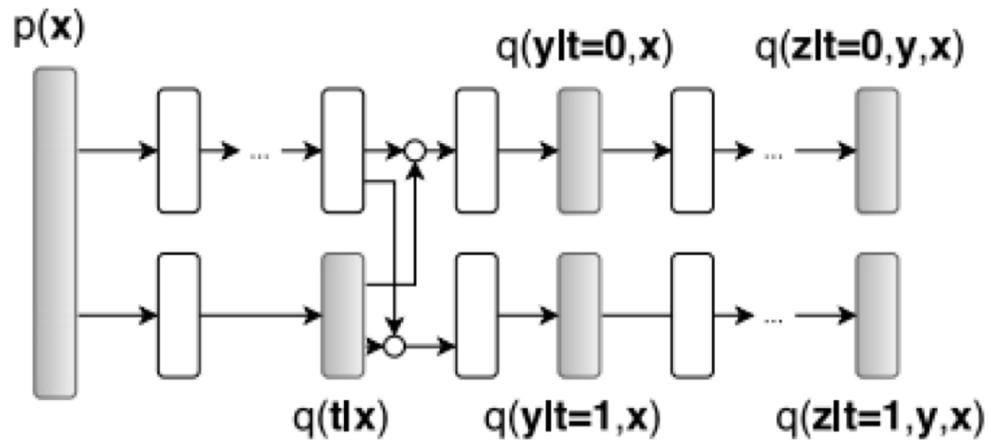


(b) Model network,  $p(x, z, t, y)$ .

Theorem: if we can estimate the joint, we can estimate causal effects

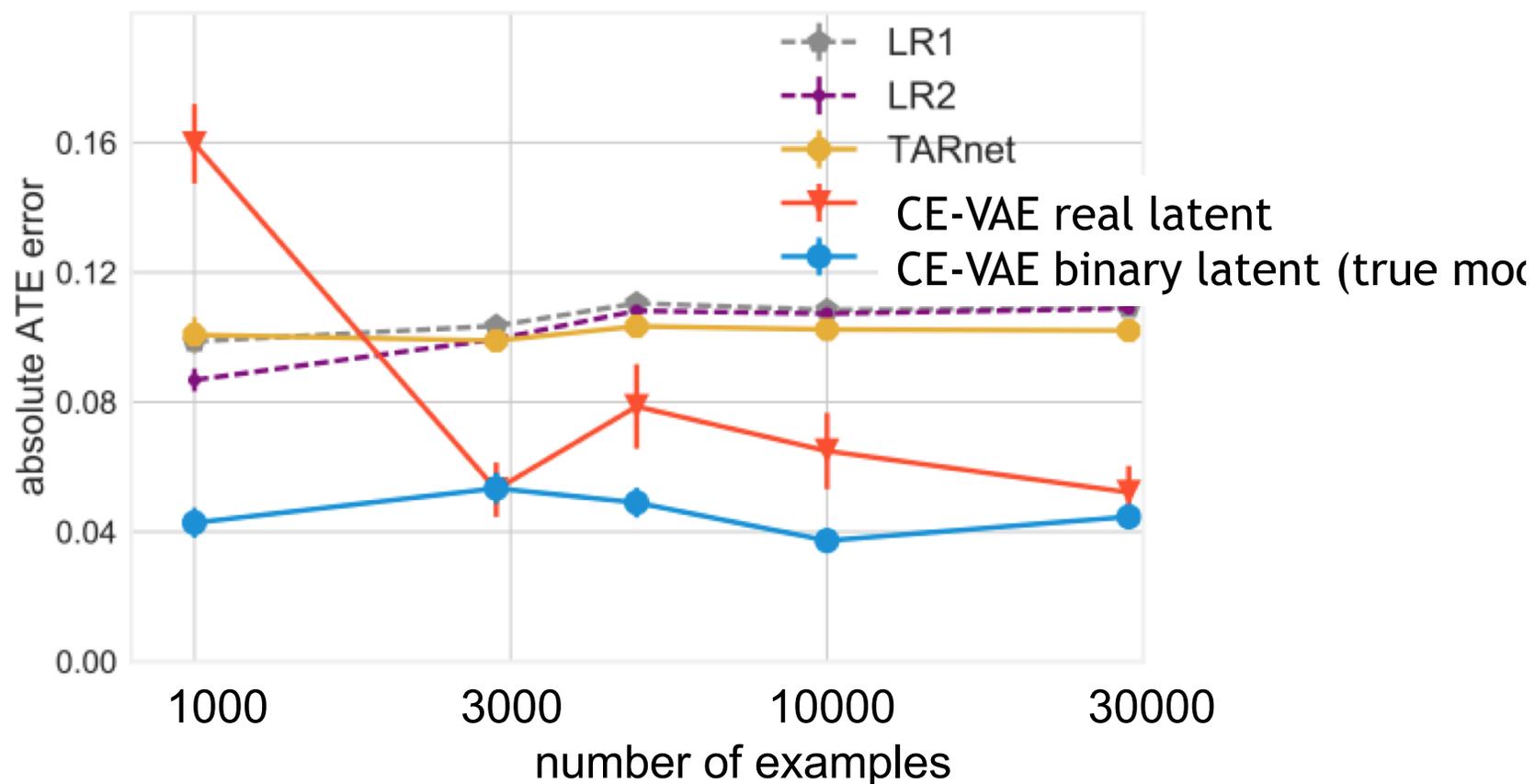
Need flexible model: Use a specialized deep generative model

# Causal VAE for fast inference



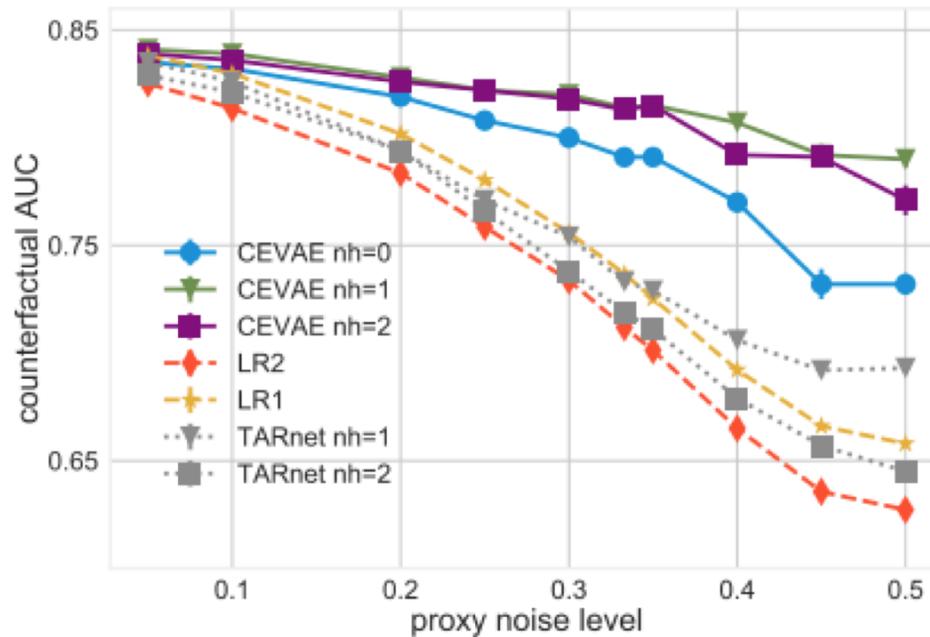
(a) Inference network,  $q(\mathbf{z}, t, \mathbf{y} | \mathbf{x})$ .

# Toy experiments show CE-VAE better at predicting avg. effect than baselines



True even when the latent space is “misspecified”

# Experiments on Twin birth data



(a) Area under the curve (AUC) for predicting the mortality of the unobserved twin in a hidden confounding experiment; higher is better.

# Takeaways

Hard problem!

Assumptions are everything

Try to capture any confounders you can

Multiple views of confounders even better  
e.g. zip code & salary & job title

## MLHC Challenge 6

# Reinforcement Learning

### Reinforcement Learning:

Train agent to repeatedly observe state and take action.  
Goal of high reward after many steps.

### Recent successes:

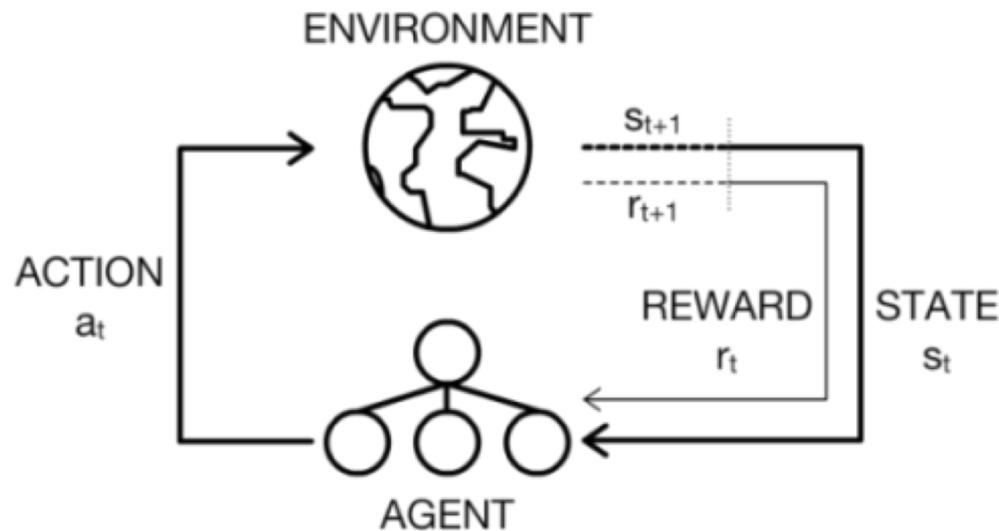
AlphaGo, Atari



### BIG QUESTION

Can we use RL for  
sequential treatment  
decisions in healthcare?

# Taking actions to seek reward



Why is this hard?

- Positive reward may not be easily reached from starting state
- **Exploration** and exploitation needed

*Not safe to explore in healthcare!*

# Early work in RL for clinical treatment

## SIMULATION

- Ernst et al. (2006) : HIV drugs
- Escandell-Montero et al. (2014): Anemia

## OBSERVATIONAL DATA

- Nemati et al. (2016):
  - Heparin for Coagulation
- Shortreed et al. (*Mach Learn* 2011):
  - Schizophrenia: clinical trial to select among 5 drugs
- Prasad et al. UAI 2017:
  - Mech. Ventilator and Sedation, uses MIMIC
- Raghu et al. MLHC 2017:
  - Sepsis treatment with fluids and vasopressors, uses MIMIC

# Emerging “Best Practices”

## Evaluating Reinforcement Learning Algorithms in Observational Health Settings

Omer Gottesman<sup>1</sup>, Fredrik Johansson<sup>2</sup>, Joshua Meier<sup>1</sup>, Jack Dent<sup>1</sup>,  
Donghun Lee<sup>1</sup>, Srivatsan Srinivasan<sup>1</sup>, Linying Zhang<sup>3</sup>, Yi Ding<sup>3</sup>, David  
Wihl<sup>1</sup>, Xuefeng Peng<sup>1</sup>, Jiayu Yao<sup>1</sup>, Isaac Lage<sup>1</sup>, Christopher Mosch<sup>4</sup>, Li-wei  
H. Lehman<sup>2</sup>, Matthieu Komorowski<sup>5,6</sup>, Aldo Faisal<sup>7</sup>, Leo Anthony Celi<sup>5,8,9</sup>,  
David Sontag<sup>2</sup>, and Finale Doshi-Velez<sup>1</sup>

*Gottesman et al. arXiv 2018*

# RL for Sepsis

- Observed data
  - Trajectories of 19k patients in ICU (MIMIC-III)
    - All meet the Sepsis-3 Criteria
  - 47 observed features (lab test values, vitals, demog.)
  - Recorded every 4 hours
- Action space (*Usual assumptions require discretizing*)
  - 5 discrete levels of IV fluids
  - 5 discrete levels of vasopressors
    - No treatment, medians of [0-25, 25-50, 50-7, 75-100] dosage quartiles
- Reward: Mortality

time	1	2	3	4	...	T-1	T
reward	0	0	0	0		0	+/-100

# Off Policy Evaluation

Can we compute value (expected reward) of a new target policy A, using trajectories collected under different policy B?

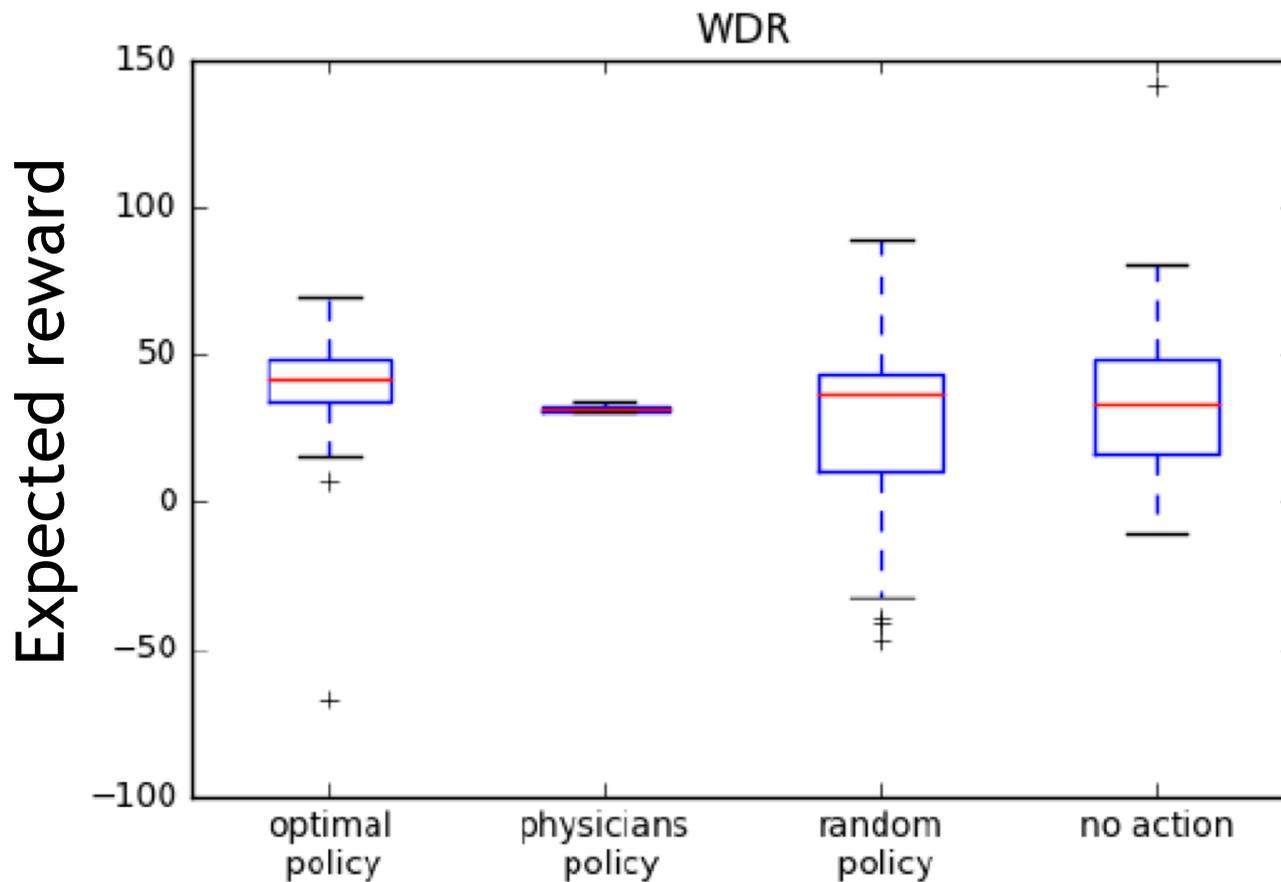
Strategy: weighted average of all rewards

- Upweight patients who had similar trajectories to those suggested by policy A

Many estimators exist. Are they good?

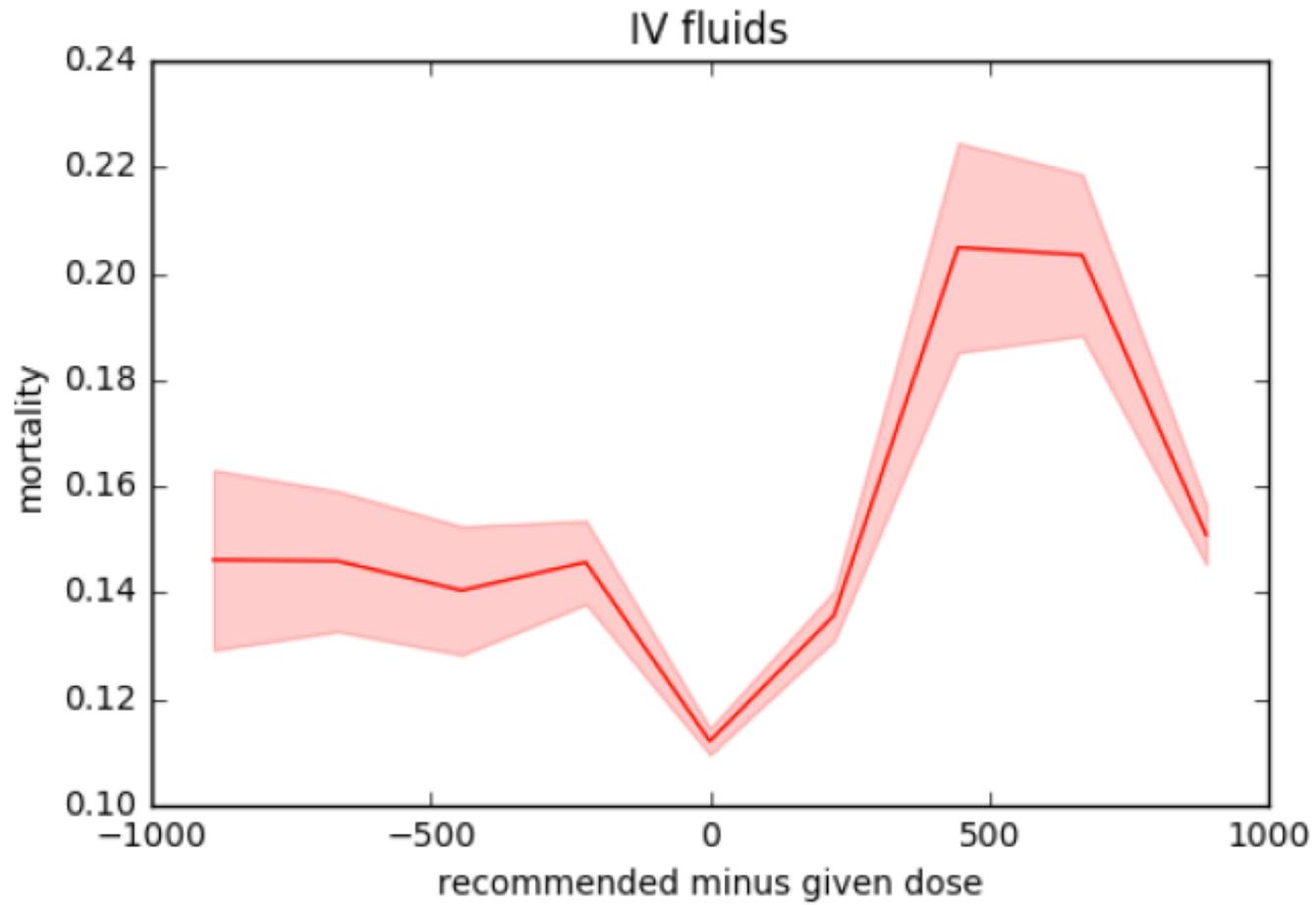
- Precup 2000
- Thomas & Brunskill 2016

# Compare RL policy to baselines

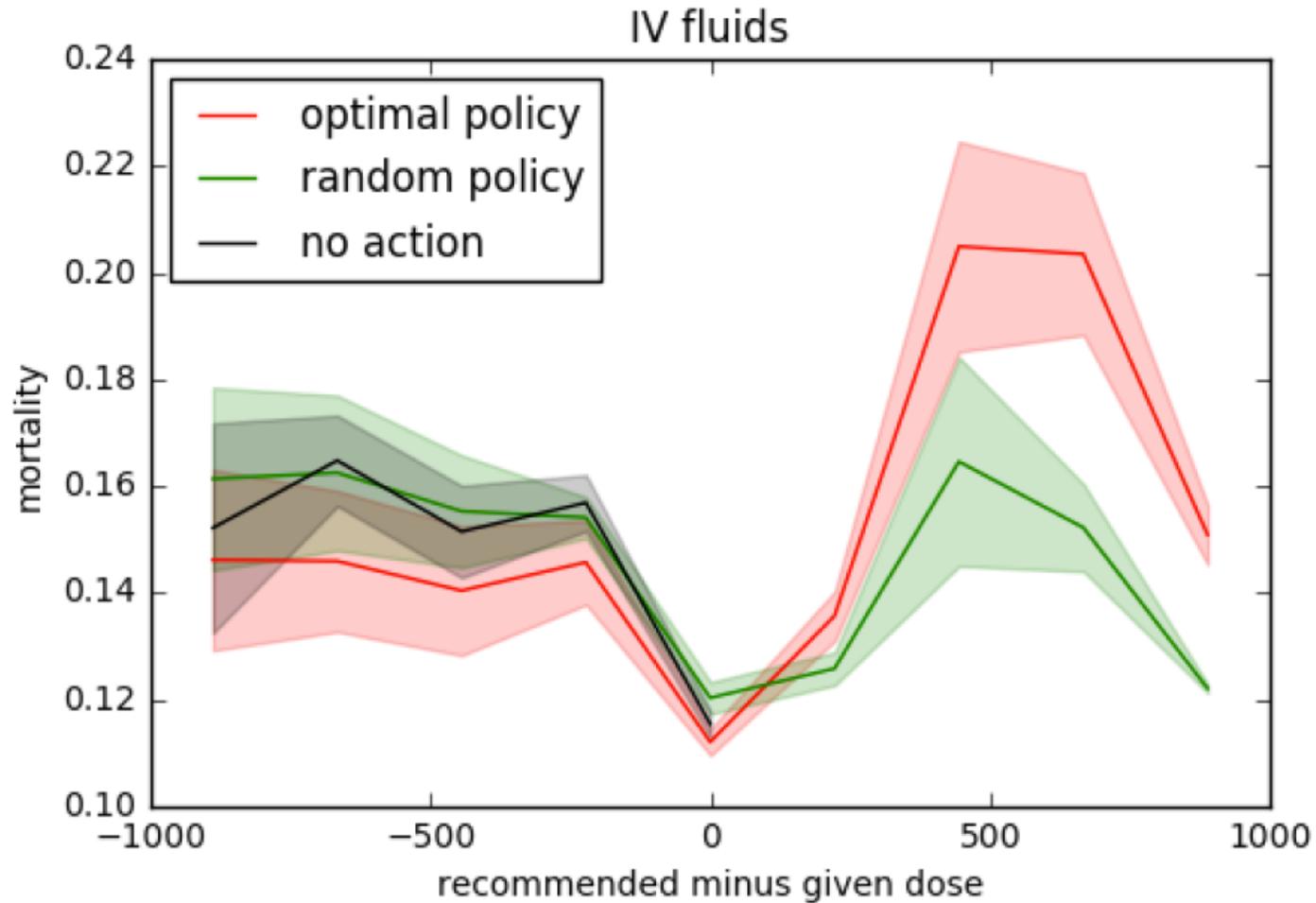


Can't distinguish random policy from no action

# U-curve comparisons



# U curve with naïve baselines



- 1) Sicker patients get higher dosages!
- 2) Discretizing dosages by quantile bad.

# Takeaways

- Choose good representations and actions
  - Err on side of capturing all potential confounders
- Work closely (clinicians & MLers) throughout process, especially when making simplifying assumptions
- With observational data, we can't expect RL to magically find an ideal policy; can maybe mimic the best clinicians at their best moments