A Road Map for Robust AI

Gilmer Valdes PhD, DABR
Department of Radiation Oncology
Department of Epidemiology and Biostatistics
University of California, San Francisco.

Outline

1. Causal Inference.

2. Risks of Models built using correlation.

3. Interpretability.

4. Expert Augmented Machine Learning

Causal Inference Framework

$Y_A \rightarrow survival\ time\ when\ treatment\ t = A$

$Y_B \rightarrow survival\ time\ when\ treatment\ t = B$

$x\ \rightarrow patient's\ characteristics.$

Treatment Effect

$$Effect = E_x\ [(Y_A - Y_B)|\mathrm{x}]$$

$$Effect = E_x\ [Y_A|\ \mathrm{x}]\ -\ E_x\ [Y_B\ |\ \mathrm{x}]$$

Where the expectation is taken ***over all patients***

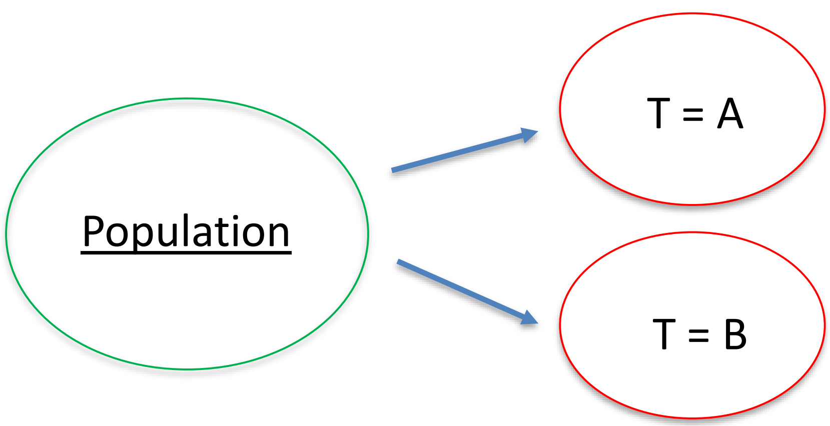
Causal Inference from Training Data

$$Effect = E_x\,[\,Y_A \mid x] \;-\; E_x\,[\,Y_B \mid x]$$

Problem: We never observe Y_A and Y_B for a patient because they either receive treatment A or B.

Approximation:

$$E_x\,[\,Y_A \mid x] \;\approx\; E_x\,[Y \mid x, t = A]$$



$$E_x\,[\,Y_B \mid x] \;\approx\; E_x\,[Y \mid x, t = B]$$

No hidden confounders

$$(Y_A, Y_B) \perp t \mid x$$

No Hidden Confounder

$$(Y_A, Y_B) \perp t \mid x$$

When does it break?

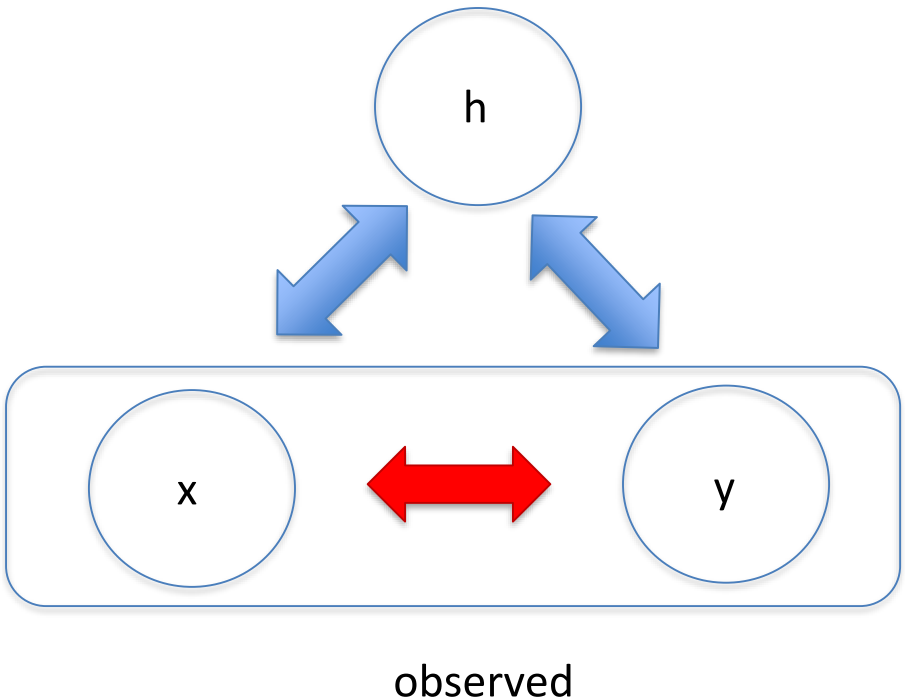
t = A is only given to very sick patients. Both Y_A, Y_B are very small.

t = B is only given to healthy patients. Both Y_A, Y_B are big.

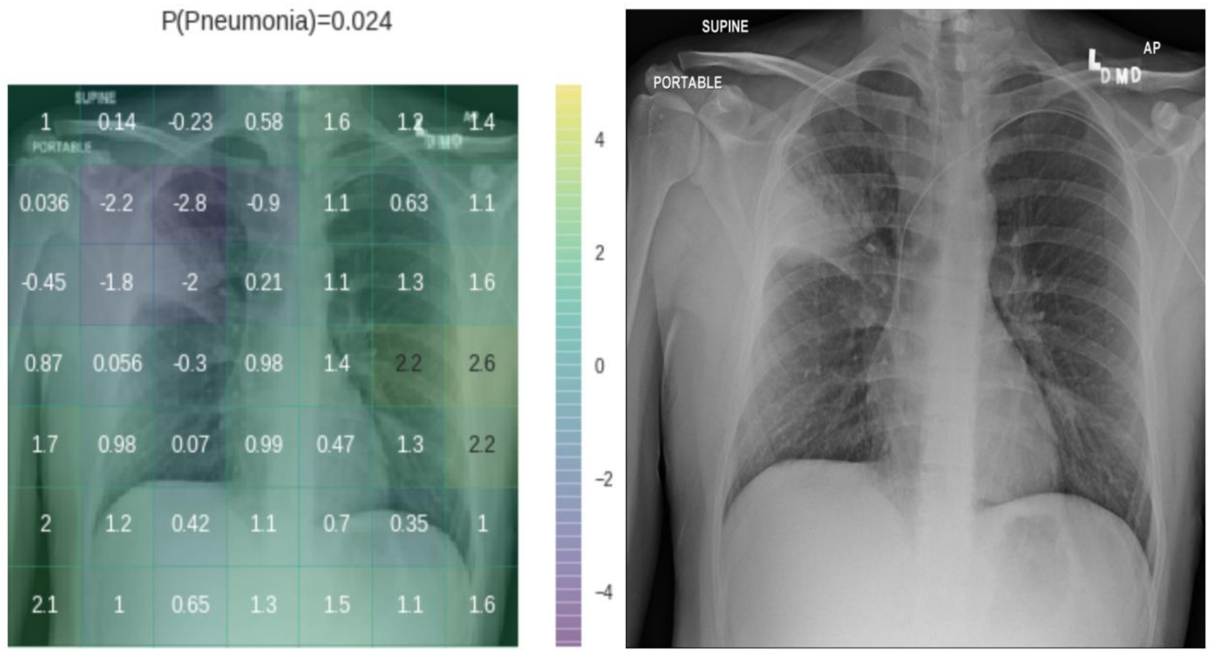
Hidden Confounder: patient selection mechanism

Hidden Confounder in Prediction Settings

$$Prediction = E\,[Y\,|\,\mathrm{x}]$$



Risks of Models built using correlation



Patient with pneumonia: heatmap of CNN on left, original image on right

88% ranking. The CNN was learning the hospital type.

https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Risks of Models built using correlation

Example: *Predicting Risk of dying of Pneumonia for In-hospital patients*

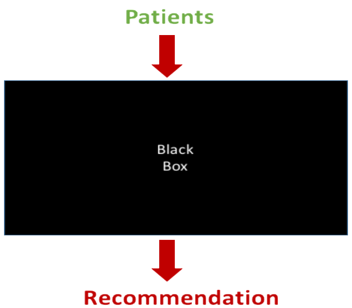
Most accurate model trained: Multi-purpose neural net….

Rule Based Model

Asthmatic → Lower Risk

- **Harmful to patients**
- **High Risk of Liability**

=



https://www.ncbi.nlm.nih.gov/pubmed/9040894

Risks of Models built using correlation

Example: *Predicting Risk of stroke for Emergency Department patients*

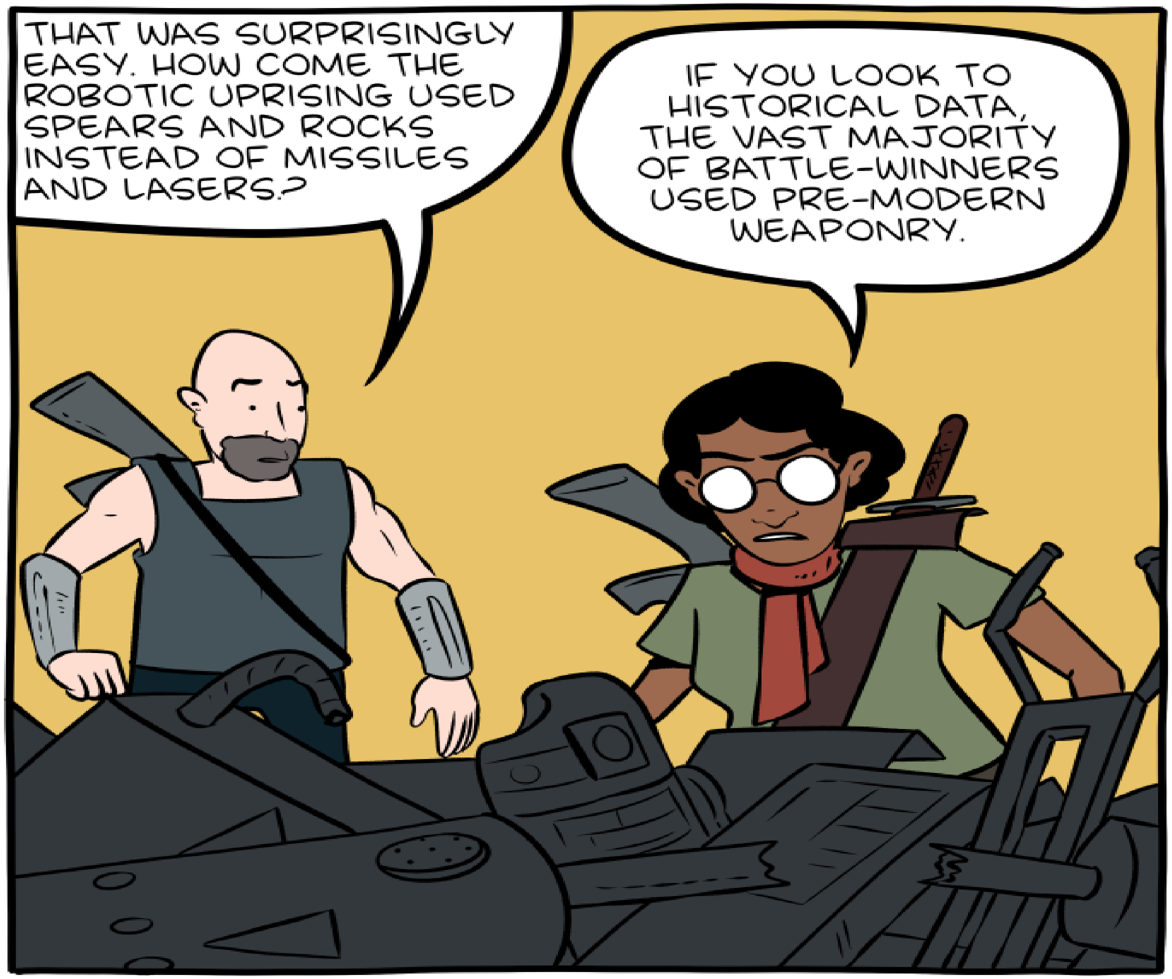
TABLE 1—PREDICTING AND MISPREDICTING

	Stroke	30-day mortality
Prior stroke	0.302	0.041
	(0.012)	(0.014)
Prior accidental injury	0.285	0.007
	(0.095)	(0.101)
Abnormal breast finding	0.224	0.162
	(0.092)	(0.110)
Cardiovascular disease history	0.218	−0.017
	(0.029)	(0.034)
Colon cancer screening	0.242	−0.475
	(0.178)	(0.222)
Acute sinusitis	0.220	0.056
	(0.155)	(0.166)

Notes: Logistic regression on demographics and prior diagnoses in EHR data. Sample: 177,825 ED visits in 2010–2012 to a large academic hospital.

"Does Machine Learning Automate Moral Hazard and Error?" .American Economic Review: Papers & Proceedings 2017, 107(5): 476–480

Context is everything



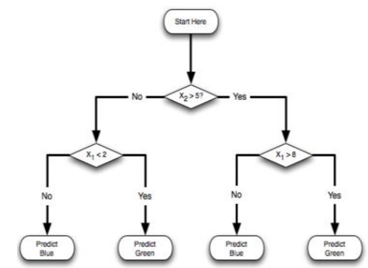
Thanks to machine-learning algorithms,
the robot apocalypse was short-lived.

Possible Solutions

$$Prediction = E\,[Y|\mathrm{x}]$$

Interpretability

1. The model is interpretable in a global sense



2. The model is interpretable locally.
Post-hoc justifications or explanations.

Variable Importance (salient map), Use a simpler model to explain a more complex one, visualizations, etc

Possible Solutions

Post-hoc interpretations are rarely faith full

Salient Map

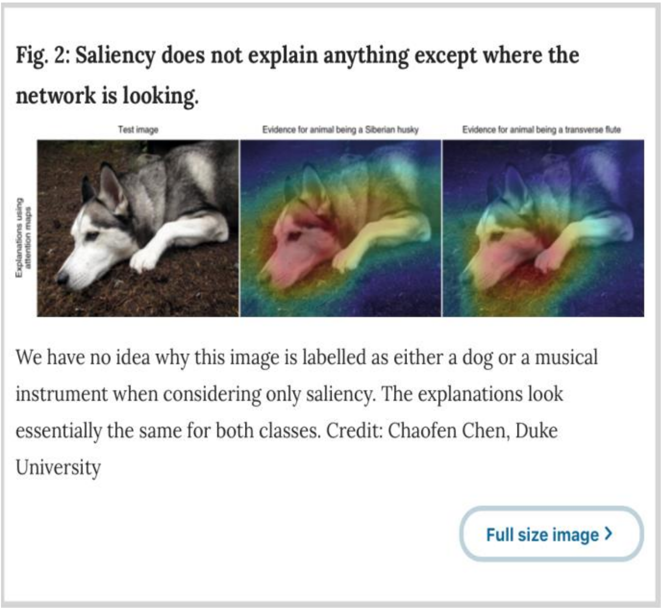


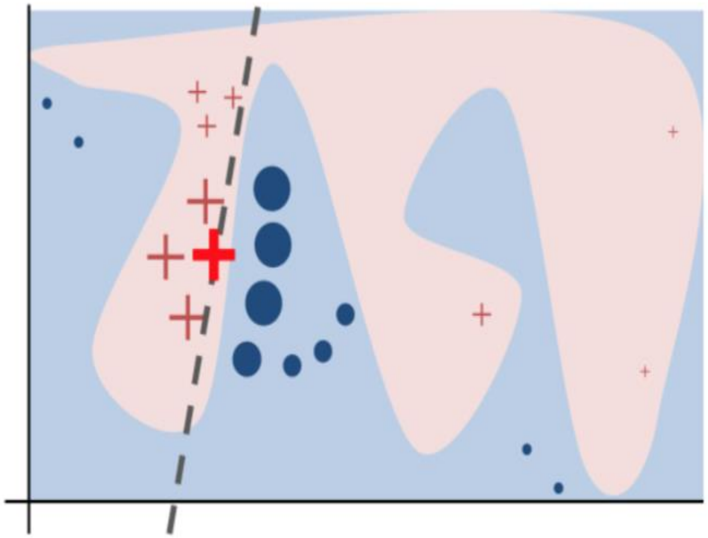
Fig. 2: Saliency does not explain anything except where the network is looking.

We have no idea why this image is labelled as either a dog or a musical instrument when considering only saliency. The explanations look essentially the same for both classes. Credit: Chaofen Chen, Duke University

Full size image ›

https://www.nature.com/articles/s42256-019-0048-x

LIME



https://arxiv.org/pdf/1602.04938.pdf

Possible Solutions

Expert-Augmented Machine Learning

Train a state-of-the-art predictive model using **RuleFit**
This represents the best machine-learned model to predict the outcome of interest given the training data

+

Extract human expert knowledge from panel of domain experts using **MediForest.com**
This provides an automated way to extract problem-specific human expert priors

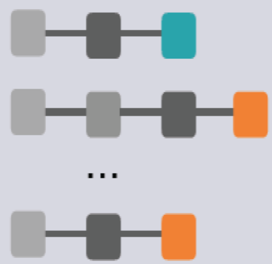
Combine ML model with expert priors to build a robust, efficient, and interpretable **EAML model**
This represents the merging of human expert knowledge with a machine-learned model for best-of-both performance

Expert-Augmented Machine Learning incorporates human expertise into ML models

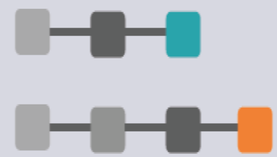
RuleFit

1. Build Trees using Gradient Boosting

2. Convert Trees to Rules



3. Select Rules using LASSO

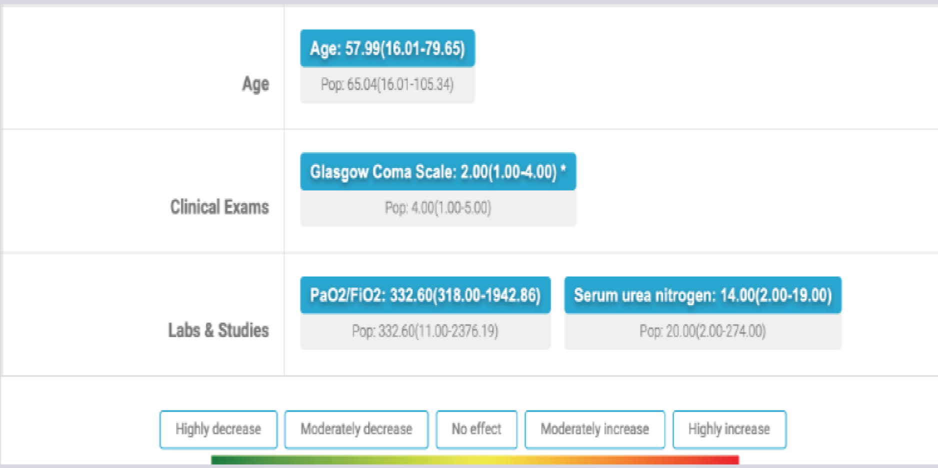


5. Compare empirical and expert rule ranking

Rule	$Rank_e$	$Rank_c$	$\Delta Rank$

MediForest Web app

4. Extract experts' assessment of each rule on MediForest.com and rank



6. Build EAML model by combining rules & expert assessments

$$\hat{C} = argmin_C \sum_{i=1}^{N} [y_i - CR(i,:)]^2$$

$$+\lambda \sum_{k=1}^{K} f(\Delta Ranking_k, STDEV_k) \, \|C_k\|_2$$

n cases; *k* rules

Gennatas et al 2019

arxiv.org/abs/1903.09731

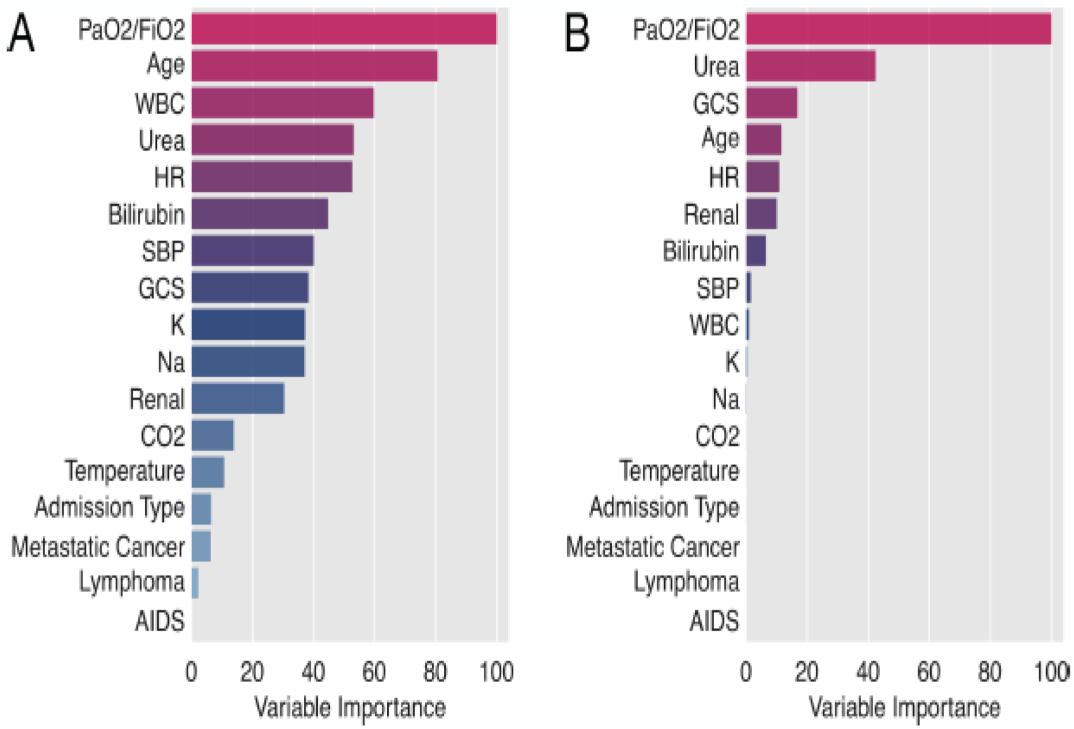
Experts assess a few simple rules instead of a vast number of individual cases

EAML

Model- vs. Expert-derived variable importance

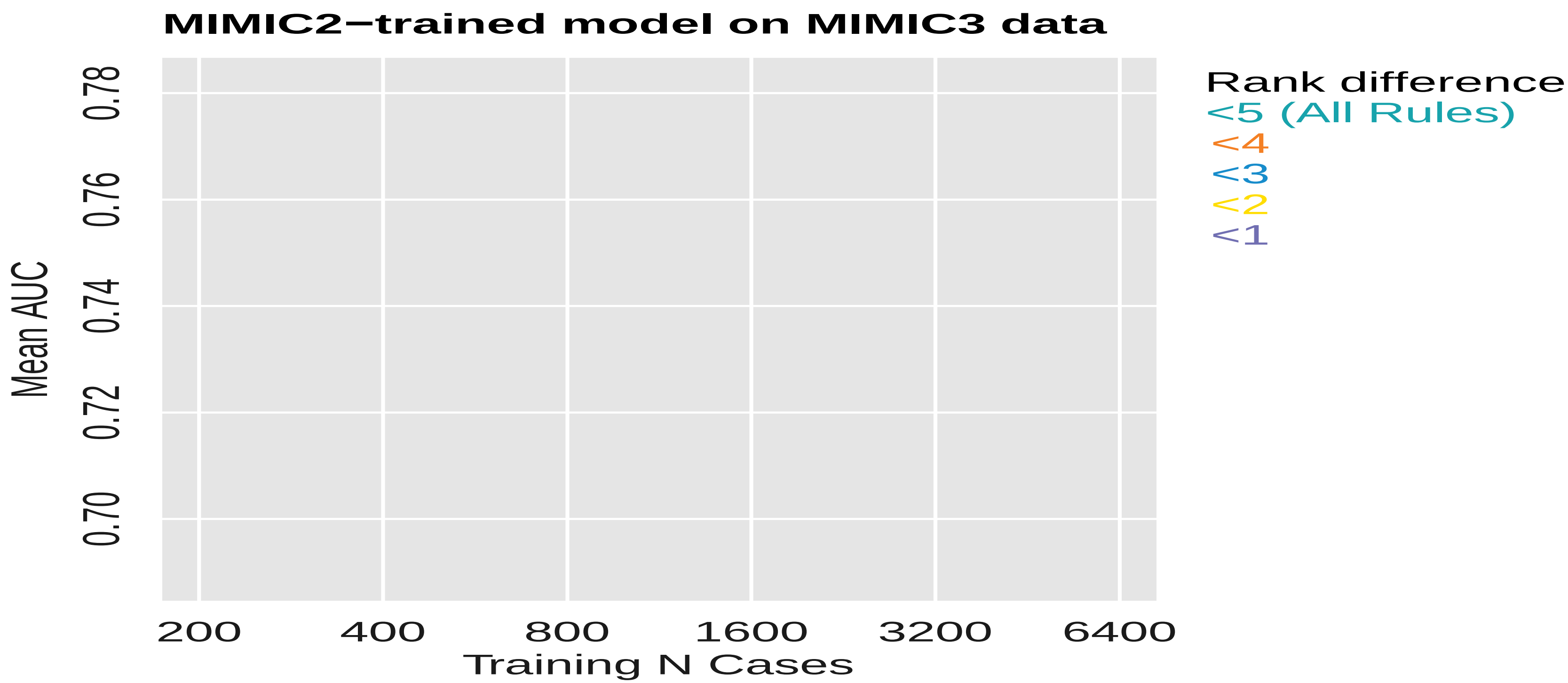
Model variable importance based on correlational structure of data

Variable importance estimated from clinicians' responses based on their causal & correlational knowledge



PaO_2/FiO_2 is the most important feature for both the model and clinicians but for different reasons

EAML allows to train with less data



$$\hat{C} = argmin_C \sum_{i=1}^{N} [y_i - C * R(i,:)]^2 + \lambda * \sum_{k=1}^{K} f(\Delta Ranking_k, STDV_k) * C_k^2$$

The Machine Learning Dogma

CHRIS ANDERSON SCIENCE 06.23.08 12:00 PM

THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE

"…There is now a better way. Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot…"

https://www.wired.com/2008/06/pb-theory/

EAML Project: The team

We are a multidisciplinary team with extensive clinical and quantitative expertise, and a shared goal of developing advanced Machine Learning algorithms for accurate, interpretable, safe and fair clinical predictive modeling

In alphabetical order:

Andrew Auerbach, MD, MPH
University of California, San Francisco, *Hospital Medicine*

Elier Delgado, MSc
Innova Montreal, *Web Application Engineering*

Eric Eaton, PhD
University of Pennsylvania, *Machine Learning, Deep Learning*

Jerome H. Friedman, PhD
Stanford University, *Statistics, Machine Learning*

Efstathios D. Gennatas, MBBS, PhD
University of Pennsylvania, *Neuroscience, Machine Learning, Biomedical Data Science*

Yannet Interian, PhD
University of San Francisco, *Data Science*

Mark J. van der Laan, PhD
University of California, Berkeley, *Biostatistics*

Jose Marcio Luna, PhD
University of Pennsylvania, *Radiation Oncology, Machine Learning*

Romain Pirracchio, MD, PhD
University of California, San Francisco, *Anesthesia, Biostatistics*

Lara G. Reichmann, PhD
University of San Francisco, *Data Science*

Charles B. Simone II, MD
NY Proton, *Radiation Oncology*

Timothy D. Solberg, PhD
University of California, San Francisco, *Radiation Oncology*

Lyle H. Ungar, PhD
University of Pennsylvania, *Machine Learning, Biomedical Data Science, Natural Language Processing*

Gilmer Valdes, PhD
University of California, San Francisco, *Radiation Oncology, Machine Learning*