Machine Learning for Healthcare: *What's next?*

David Sontag
Clinical Machine Learning Group
MIT
@david_sontag

INSTITUTE FOR MEDICAL ENGINEERING & SCIENCE


https://www.mlforhc.org
MACHINE LEARNING FOR HEALTHCARE
Machine Learning for Healthcare 2019
August 8-10
University of Michigan, Ann Arbor, MI
Registration: OPEN NOW!

https://ml4health.github.io/2018/

ML4H: Machine Learning for Health

Workshop at NeurIPS 2018

papers call for papers dates organizers program speakers sponsors

ML4H 2018: a workshop at NeurIPS 2018
Saturday December 08, 2018
Room 517 D, Palais des Congrès de Montréal, Montreal, Canada

This workshop will bring together machine learning researchers, clinicians, and healthcare data experts. The program consists of invited talks, contributed posters and panel discussions.

Direct questions to: ml4h.workshop.nips.2018@gmail.com

Hot topics in MLHC

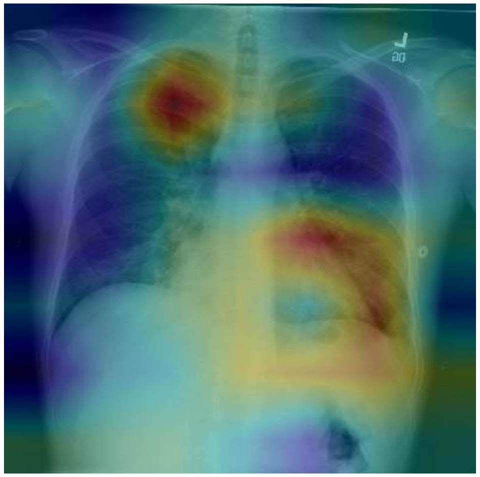
- Interpretability
- Robustness to adversaries, dataset shift
- Fairness
- Reinforcement learning

Hot topics in MLHC

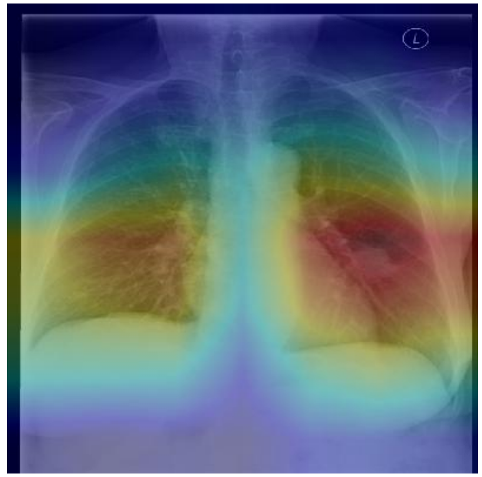
- **Interpretability**
- Robustness to adversaries, dataset shift
- Fairness
- Reinforcement learning

Interpretability

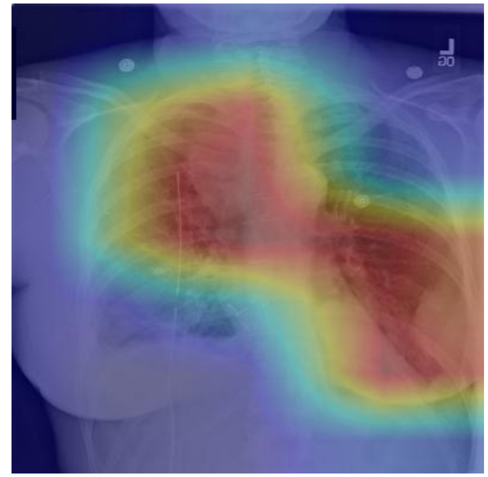
- Global interpretability – understand model as a whole
  - Will it work prospectively as intended?
  - What data was most useful?
- Local interpretability – understand predictions for individual patients
  - Build trust in predictions; recognize errors
  - Provide guidance to decision makers who may have additional information



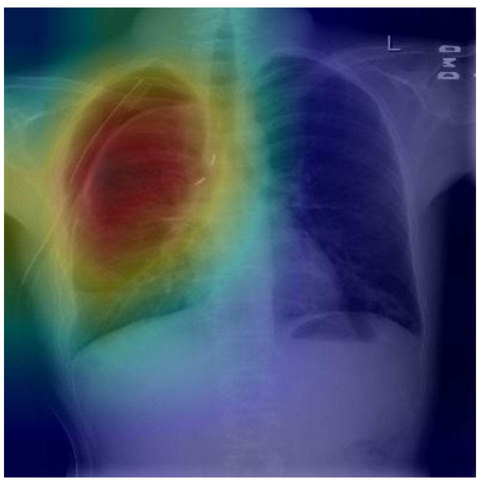
(a) Patient with multifocal community acquired pneumonia. The model correctly detects the airspace disease in the left lower and right upper lobes to arrive at the pneumonia diagnosis.



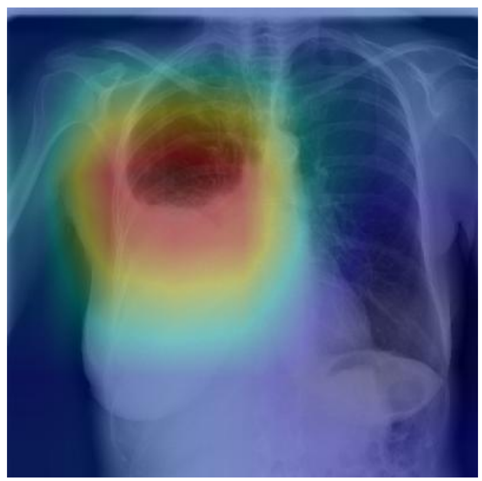
(b) Patient with a left lung nodule. The model identifies the left lower lobe lung nodule and correctly classifies the pathology.



(c) Patient with primary lung malignancy and two large masses, one in the left lower lobe and one in the right upper lobe adjacent to the mediastinum. The model correctly identifies both masses in the X-ray.



(d) Patient with a right-sided pneumothroax and chest tube. The model detects the abnormal lung to correctly predict the presence of pneumothorax (collapsed lung).



(e) Patient with a large right pleural effusion (fluid in the pleural space). The model correctly labels the effusion and focuses on the right lower chest.

(f) Patient with congestive heart failure and cardiomegaly (enlarged heart). The model correctly identifies the enlarged cardiac silhouette.

Hot topics in MLHC

- Interpretability
- **Robustness to adversaries, dataset shift**
- Fairness
- Reinforcement learning

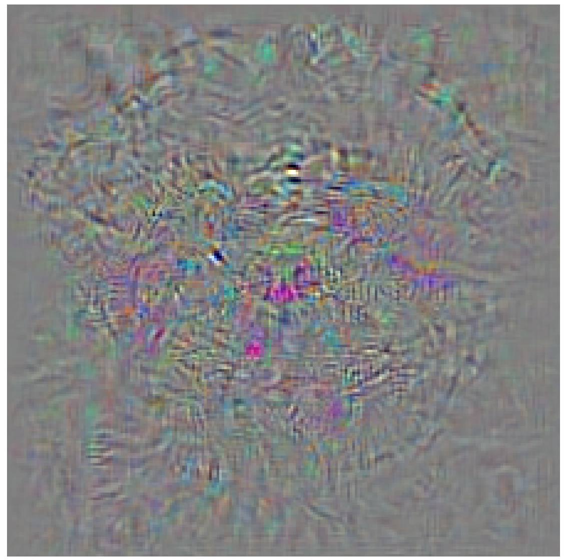
Machine learning is brittle: adversarial perturbations

Correctly classified as a Dog

[Szegedy et al., “Intriguing properties of neural networks”, ICLR 2014]

Machine learning is brittle: adversarial perturbations

+



Original image

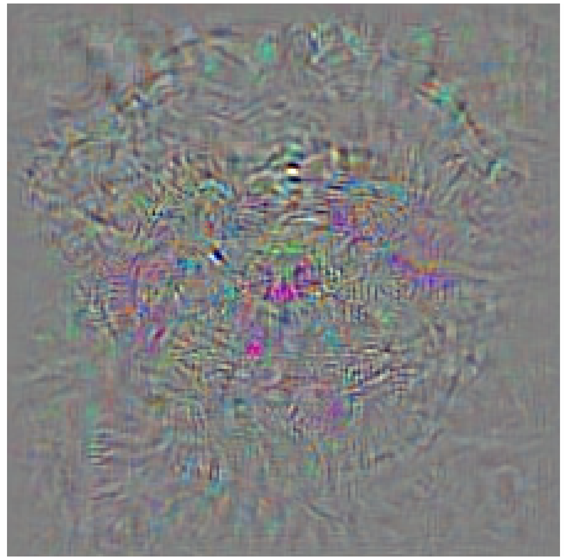
Noise (not random)

[Szegedy et al., "Intriguing properties of neural networks", ICLR 2014]

Machine learning is brittle: adversarial perturbations

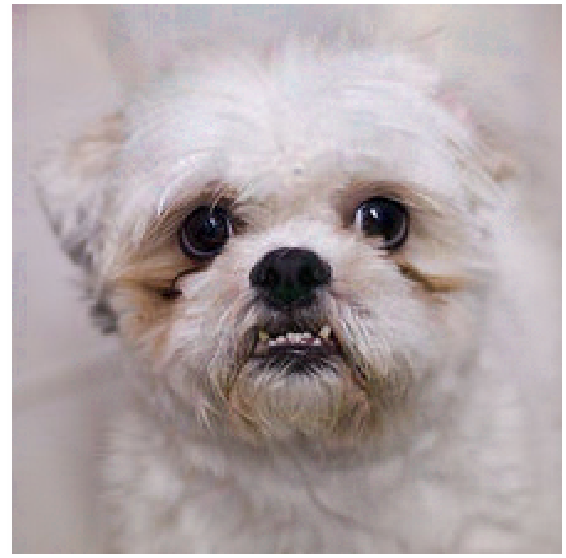
Original image

+



Noise (not random)

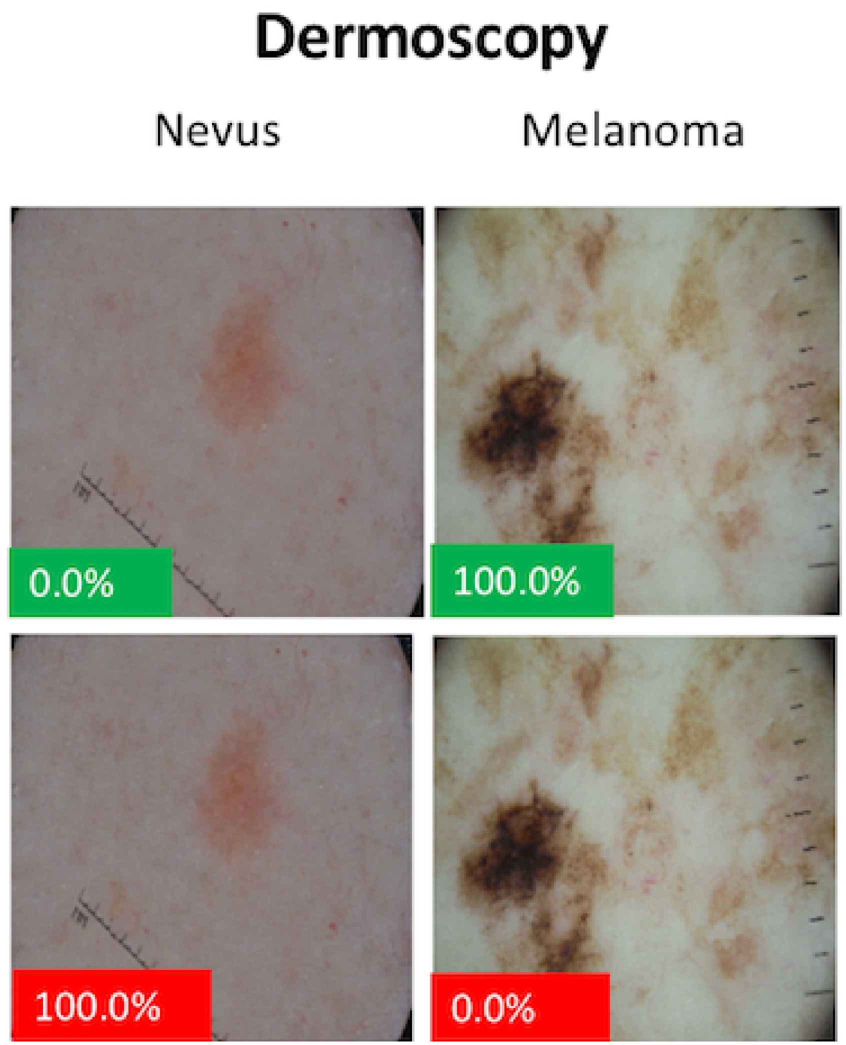
=



Classified as Ostrich!

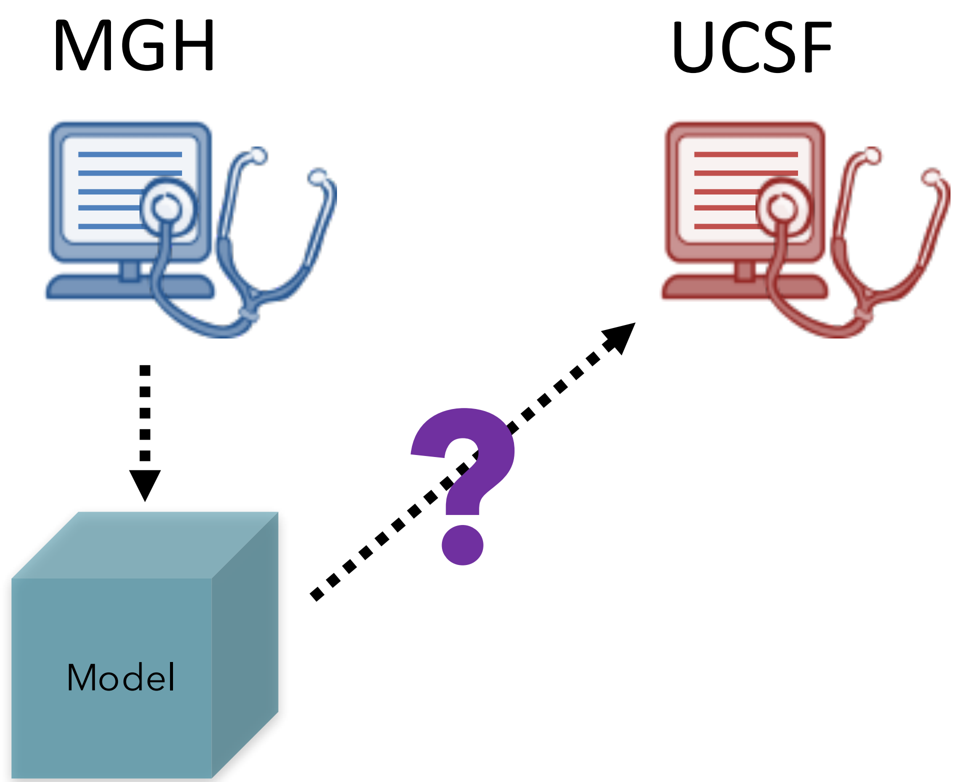
[Szegedy et al., “Intriguing properties of neural networks”, ICLR 2014]

Machine learning is brittle: adversarial perturbations



[Finlayson et al., “Adversarial Attacks Against Medical Deep Learning Systems”, Arxiv 1804.05296, 2018]

Machine learning is brittle: natural changes in the data



Build population-level checks into deployment/transfer

[Figure adopted from Jen Gong and Tristan Naumann]

Good practice: report “Table 1”

Table 1. Characteristics of 47 119 Hospitalized Patients

Characteristic	Finding[a]
Age, mean (SE), y	60.9 (18.15)
Female	23 952 (50.8)
Black/African American race	5258 (11.2)
Hispanic/Latino ethnicity	3667 (7.8)
Medicaid	8303 (17.6)
Heart failure in problem list	3630 (7.7)
Prior diagnosis of any heart failure	2985 (6.3)
Prior diagnosis of primary heart failure	615 (1.3)
Prior echocardiography	15 938 (33.8)
Loop diuretics	
Inpatient	6837 (14.5)
Outpatient	6427 (13.6)
ACE inhibitors or ARB	
Inpatient	13 166 (27.9)
Outpatient	14 797 (31.4)
β-Blockers	
Inpatient	19 748 (41.9)
Outpatient	14 870 (31.6)
Heart failure with β-blockers	
Inpatient	6310 (13.4)
Outpatient	8644 (18.4)
Blood pressure, mean (SE), mm Hg	
Systolic	123.3 (18.3)
Diastolic	67.8 (12.8)
Creatinine, mean (SE), mg/dL	1.01 (1.1)
Sodium, mean (SE), mEq/L	138.4 (3.7)
BNP, pg/mL	
<500	1721 (23.4)
500-999	878 (12.0)
1000-4999	2498 (34.0)
5000-9999	931 (12.7)
10 000-19 999	652 (8.9)
≥20 000	667 (9.1)
Blood pressure	
Any systolic	46 982 (99.7)
Any diastolic	46 982 (99.7)
Any creatinine	46 598 (98.9)
Any sodium	46 613 (98.9)
Any BNP	7347 (15.6)
Problem list	
Acute MI	952 (2.0)
Atherosclerosis	6147 (13.0)
Final discharge diagnosis of heart failure	
Any diagnosis	6549 (13.9)
Principal diagnosis	1214 (2.6)

[Blecker et al., Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data, JAMA Cardiology 2016]

Hot topics in MLHC

- Interpretability
- Robustness to adversaries, dataset shift
- **Fairness**
- Reinforcement learning

Machine Bias: There's Softwar ×

David

Secure https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Fair Regression for Health Care Spending

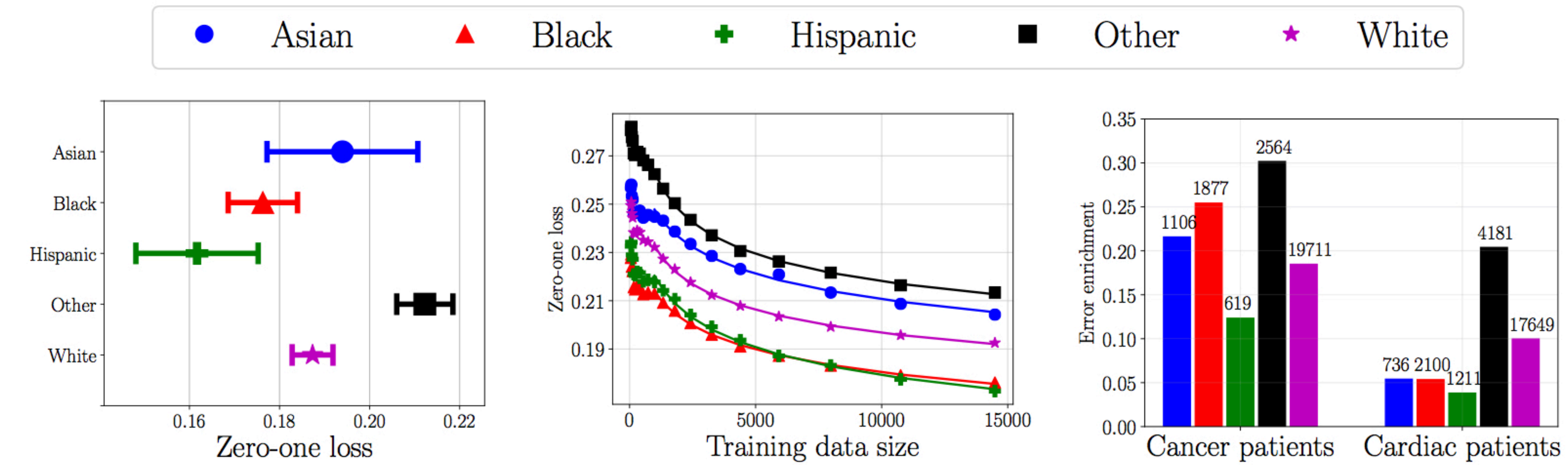
Anna Zink
Harvard University
and
Sherri Rose
Harvard Medical School*

January 31, 2019

Abstract

The distribution of health care payments to insurance plans has substantial consequences for social policy. Risk adjustment formulas predict spending in health insurance markets in order to provide fair benefits and health care coverage for all enrollees, regardless of their health status. Unfortunately, current risk adjustment formulas are known to undercompensate payments to health insurers for specific groups of enrollees (by underpredicting their spending). Much of the existing algorithmic fairness literature for group fairness to date has focused on classifiers and binary outcomes. To improve risk adjustment formulas for undercompensated groups, we expand on concepts from the statistics, computer science, and health economics literature to develop new fair regression methods for continuous outcomes by building fairness considerations directly into the objective function. We additionally propose a novel measure of fairness while asserting that a suite of metrics is necessary in order to evaluate risk adjustment formulas more fully. Our data application using the IBM MarketScan Research Databases and simulation studies demonstrate that these new fair regression methods may lead to massive improvements in group fairness with only small reductions in overall fit.

Keywords: Constrained regression, Penalized regression, Risk adjustment, Fairness



(a) Using Tukey's range test, we can find the 95%-significance level for the zero-one loss for each group over 5-fold cross validation.

(b) As training set size increases, zero-one loss over 50 trials decreases over all groups and appears to converge to an asymptote.

(c) Topic modeling reveals subpopulations with high differences in zero-one loss, for example cancer patients and cardiac patients.

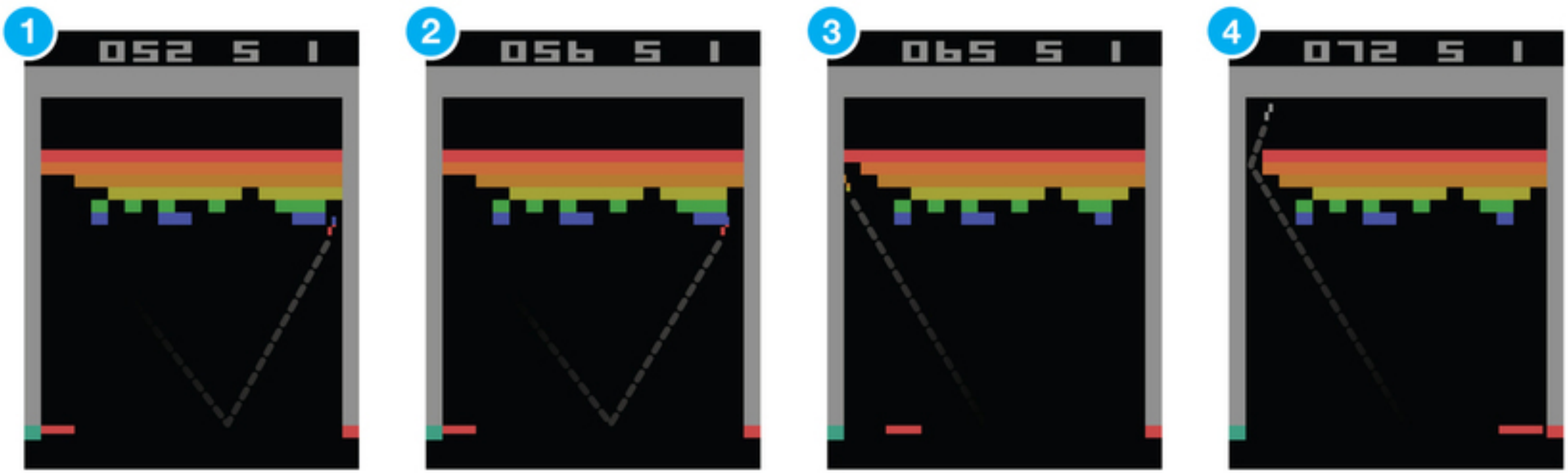
Figure 3: Mortality prediction from clinical notes using logistic regression. Best viewed in color.

[Chen, Johansson, Sontag, *Why is my classifier discriminatory?*, NeurIPS, 2018]

Hot topics in MLHC

- Interpretability
- Robustness to adversaries, dataset shift
- Fairness
- **Reinforcement learning**

Learning to play Atari games



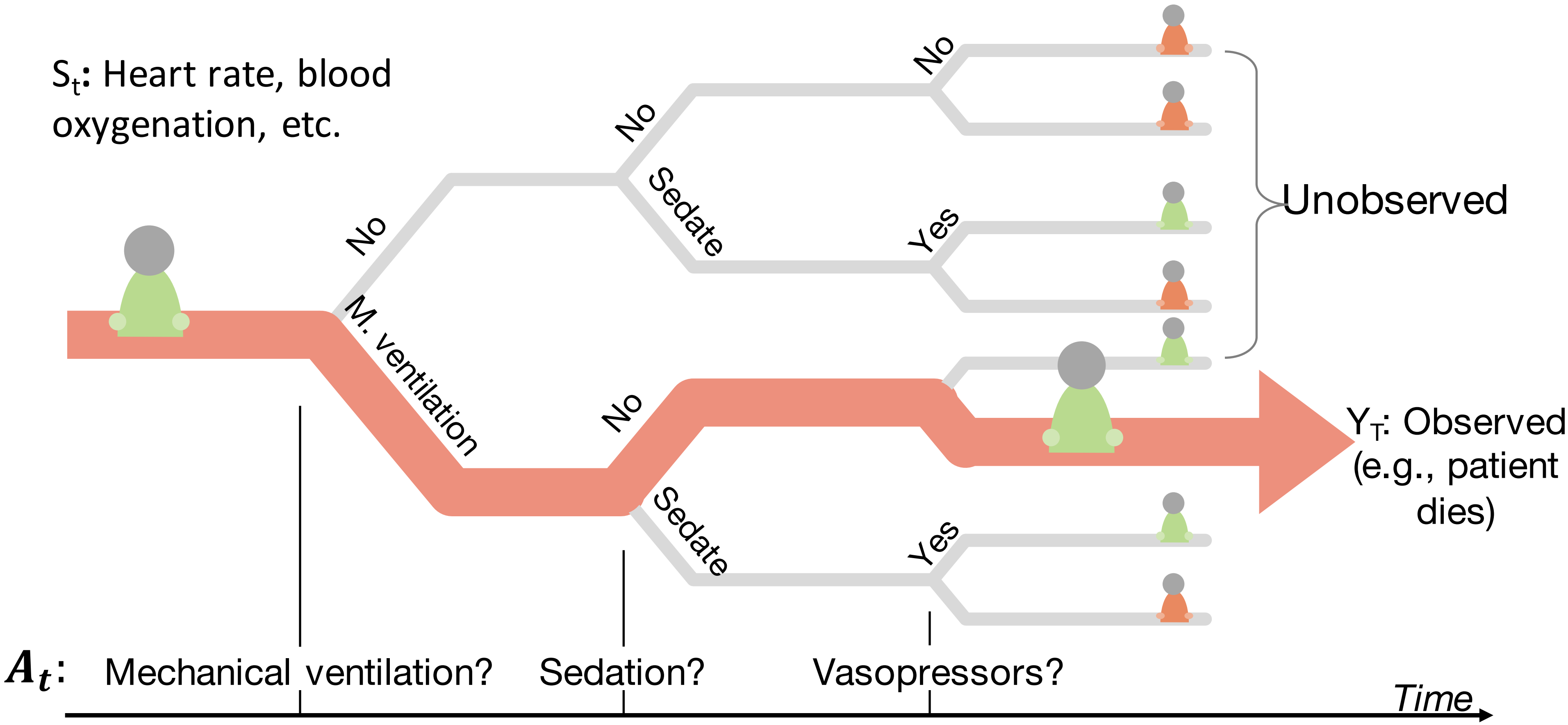
Watch video: https://www.youtube.com/watch?v=V1eYniJ0Rnk

Could we use such reinforcement learning algorithms in health care?

(Off-Policy) Reinforcement Learning

- **Goal:** Find a dynamic treatment regime (policy) $\pi(A_t \mid H_t)$
  - that selects **actions** $\boldsymbol{A_t}$
  - which optimize **outcomes** $\boldsymbol{Y_{t:T}}$ (i.e., future rewards)
  - given the history $H_t = \{(S_0, A_0, Y_0), \dots, (S_{t-1}, A_{t-1}, Y_{t-1}), S_t\}$ of **states** $\boldsymbol{S_t}$, actions and outcomes
- **Given**: samples of past histories (no exploration possible)
- **Algorithms**: e.g., deep Q-learning

Example: Managing sepsis in the ICU



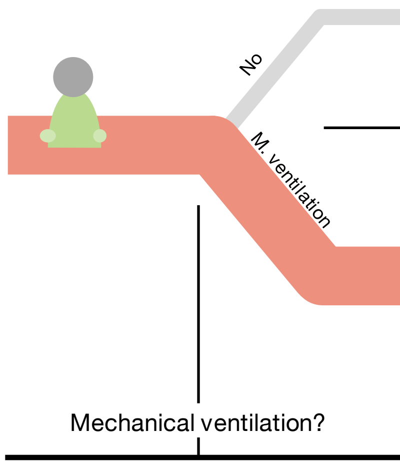
Komorowski et al.,"The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care", Nature Medicine 2018

Off-policy RL has to be done with care[1]

- In performing and evaluating observational studies of sequential decision making, we must ask:
  1. Do we have access to the information currently used in decision making?
  2. Are we optimizing the right reward/outcome?
  3. Is our data large enough to compare our proposed policy to existing ones?

[1]Guidelines for reinforcement learning in healthcare. Gottesman, O; Johansson, F; Komorowski, M; Faisal, A; Sontag, D; Doshi-Velez, F; and Celi, L. *Nature Medicine*, 25(1): 16–18. 2019

Off-policy RL guidelines: confounding

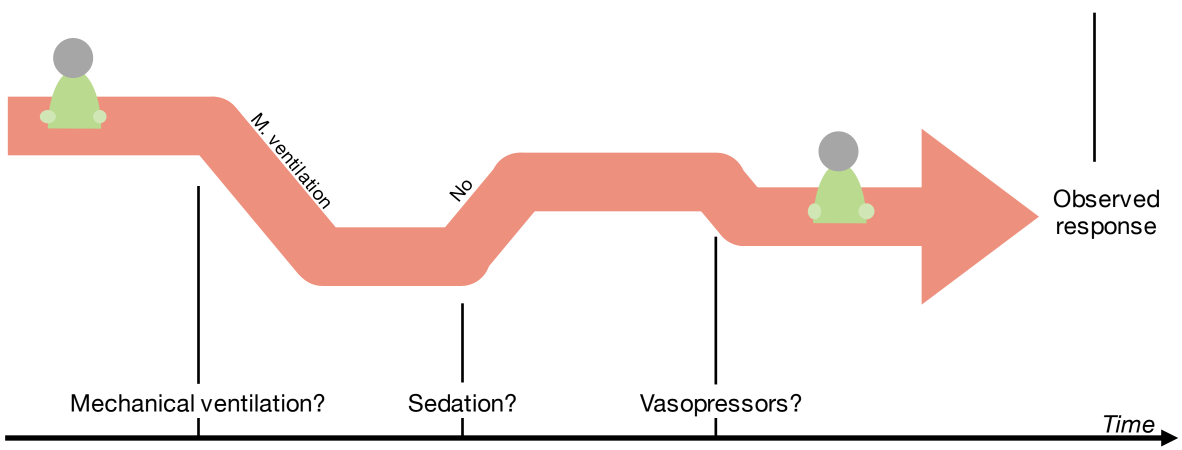


1. Do we have access to the information used by doctors in making this choice?

If not, our estimate will likely be **confounded**

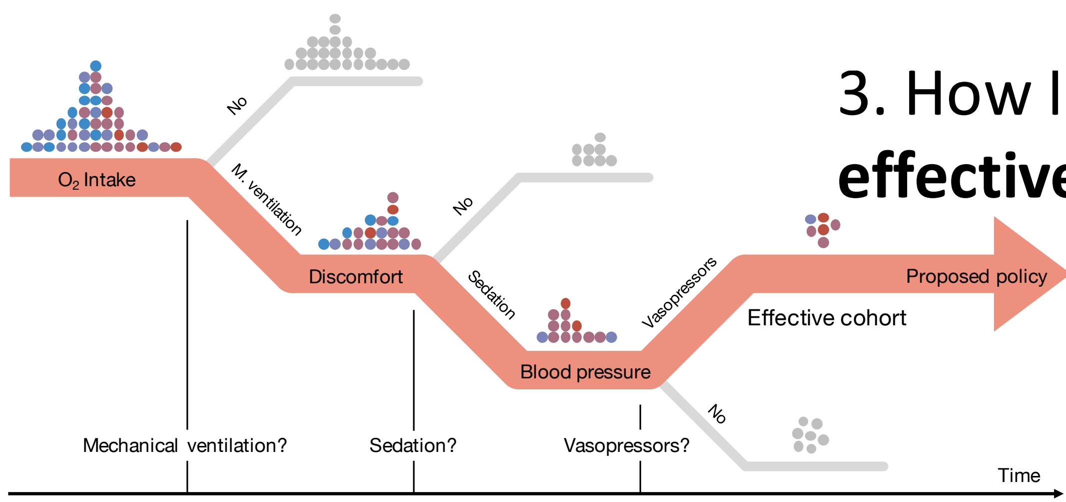
Off-policy RL guidelines: outcome label

2. What **reward** are we optimizing?
Does it capture long-term effects?



Off-policy RL guidelines: sample size

- Standard to make use only of patient trajectories that agree with the proposed policy—small *effective sample size*



3. How large is the **effective sample size**?

Hot topics in ML HC

- In
- Rc
- Fa
- Re

https://arxiv.org/pdf/1806.00388.pdf

Opportunities in Machine Learning for Healthcare

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

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

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

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

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

Abstract

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

And that’s a wrap!

- Thanks for a great two days
- Keep in touch:

E-mail: dsontag@csail.mit.edu
Twitter: david_sontag
LinkedIn: https://www.linkedin.com/in/david-sontag/

Readings

References for risk stratification

- Population-Level Prediction of Type 2 Diabetes using Claims Data and Analysis of Risk FactorsRazavian et al., Big Data 2015
  https://www.liebertpub.com/doi/pdf/10.1089/big.2015.0020
- Predicting the Risk and Trajectory of Intensive Care Patients Using Survival Models
  Caleb Hug, Master's thesis at MIT, 2006
  https://dspace.mit.edu/handle/1721.1/38326
- Passive Detection of Atrial Fibrillation Using a Commercially Available Smartwatch
  Tison et al., JAMA Cardiology 2018
  https://jamanetwork.com/journals/jamacardiology/article-abstract/2675364
- Moving From Big Data to Deep Learning— The Case of Atrial Fibrillation (Editorial)
  Turakhia, JAMA Cardiology 2018
  https://jamanetwork.com/journals/jamacardiology/article-abstract/2675362
- Scalable and accurate deep learning with electronic health records
  Rajkomar et al., Nature Digital Medicine, 2018
  https://www.nature.com/articles/s41746-018-0029-1
  Supplementary: https://static-content.springer.com/esm/art%3A10.1038%2Fs41746-018-0029-1/MediaObjects/41746_2018_29_MOESM1_ESM.pdf
- Horng, Sontag, et al. "Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning". PLOS ONE, 2017
  http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174708
- Electronic medical record phenotyping using the anchor and learn framework
  Halpern, Horng, Choi, Sontag, JAMIA '16
  https://academic.oup.com/jamia/article/23/4/731/2200279

References for causal inference

- Miguel Hernan's causal inference book
  https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/
- Paul Rosenbaum's *Design of Observational Studies*
  https://www.springer.com/us/book/9781441912121
- High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Schneeweiss et al., Epidemiology 2009
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3077219/
- Estimation and Inference of Heterogeneous Treatment Effects using Random Forests
  Stefan Wager, Susan Athey, JASA '18
  https://arxiv.org/abs/1510.04342
- Estimating individual treatment effect: generalization bounds and algorithms
  Shalit, Johansson, Sontag, ICML 2017.
  http://arxiv.org/pdf/1606.03976.pdf
- Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study
  Gabriel Brat et al., BMJ 2017
  https://www.bmj.com/content/360/bmj.j5790

References for time-series

- Factorial Switching Linear Dynamical Systems applied to Physiological Condition Monitoring
  Quinn et al., TPAMI 2008
  https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4586385
- Integration of Early Physiological Responses Predicts Later Illness Severity in Preterm Infants
  Saria et al., Science Translational Medicine 2010
  http://stm.sciencemag.org/content/2/48/48ra65
- Clifford et al. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge, Computing in Cardiology 2017
  https://www.physionet.org/challenge/2017/
- Abductive reasoning as the basis to reproduce expert criteria in ECG Atrial Fibrillation identification
  Teijeiro, Garcia, Castro, Felix. arXiv:1802.05998, 2018
  https://arxiv.org/abs/1802.05998
- Cardiologist-Level Arrhythmia Detection With Convolutional Neural Networks
  Rajpurkar et al. https://arxiv.org/abs/1707.01836
- Modeling Disease Progression via Fused Sparse Group Lasso
  Zhou et al., KDD ’12
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4191837/

References for causal inference

- Personalized Diabetes Management Using Electronic Medical Records
  Dimitris Bertsimas, Nathan Kallus, Alexander M. Weinstein, and Ying Daisy Zhuo
  Diabetes Care, 2016
  http://care.diabetesjournals.org/content/early/2016/12/01/dc16-0826.full-text.pdf
- Medical Homes and Cost and Utilization Among High-Risk Patients
  Susannah Higgins; Ravi Chawla; Christine Colombo; Richard Snyder; and Somesh Nigam
  American Journal of Managed Care, 2014
  https://www.ajmc.com/journals/issue/2014/2014-vol20-n3/medical-homes-and-cost-and-utilization-among-high-risk-patients?p=1

References for interpretability

- Implications of non-stationarity on predictive modeling using EHRs
  Kenneth Jung, Nigam Shah. JBI 2015
  https://www.sciencedirect.com/science/article/pii/S1532046415002282
- Intriguing properties of neural networks
  Szegedy et al. 2014
  https://arxiv.org/abs/1312.6199
- "Why Should I Trust You?" Explaining the Predictions of Any Classifier
  Ribeiro et al., KDD '16
  http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf
- Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission
  Caruana et al., KDD 2015
  http://people.dbmi.columbia.edu/noemie/papers/15kdd.pdf
- The Mythos of Model Interpretability
  Zachary C. Lipton
  https://arxiv.org/abs/1606.03490

References for disease subtyping

- Phenomapping for Novel Classification of Heart Failure with Preserved Ejection Fraction
  Shah et al., Circulation 2015
- Subtyping: What It Is and Its Role in Precision Medicine
  Saria & Goldberg, IEEE Intelligent Systems 2015
  https://www.dropbox.com/s/krofvs7da6u3r4k/Saria_IEEE2015_SubtypingAndPredicionMedicine.pdf
- Comorbidity Clusters in Autism Spectrum Disorders: An Electronic Health Record Time-Series Analysis
  Doshi-Velez, Ge, Kohane. Pediatrics, 2014.
  https://www.ncbi.nlm.nih.gov/pubmed/24323995
- Cluster Analysis and Clinical Asthma Phenotypes
  Haldar et al., Am J Respir Crit Care Med. 2008.
  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3992366/pdf/emss-29902.pdf

References on MLHC policy

- **Real-World Evidence In Support Of Precision Medicine: Clinico-Genomic Cancer Data As A Case Study**
  Vineeta Agarwala, Sean Khozin, Gaurav Singal, Claire O'Connell, Deborah Kuk, Gerald Li, Anala Gossai, Vincent Miller, and Amy P. Abernethy
  Health Affairs, 2018
  https://www.healthaffairs.org/doi/abs/10.1377/hlthaff.2017.1579
- **FDA Regulation of Mobile Medical Apps**
  Shuren, Patel, Gottlieb. JAMA, 2018
  https://jamanetwork.com/journals/jama/fullarticle/2687221
- **How Tech Can Turn Doctors Into Clerical Workers: The Threat That Electronic Health Records and Machine Learning Pose to Physicsians' Clinical Judgement -- and their Well-Being**
  Abraham Verghese
  https://www.nytimes.com/interactive/2018/05/16/magazine/health-issue-what-we-lose-with-data-driven-medicine.html
- **Predictive modeling of U.S. health care spending in late life**
  Einav et al., Science 2018
  http://science.sciencemag.org/content/360/6396/1462
- **Hacking Healthcare: A Guide to Standards, Workflows, and Meaningful Use**
  Trotter & Uhlman. O'Reilly Media, 2011
  https://www.amazon.com/Hacking-Healthcare-Standards-Workflows-Meaningful/dp/1449305024
- **Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients**
  https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2014.0041
- **Predicting the Future — Big Data, Machine Learning, and Clinical Medicine**
  Ziad Obermeyer, M.D., and Ezekiel J. Emanuel, M.D., Ph.D.
  https://www.nejm.org/doi/full/10.1056/NEJMp1606181

References for fairness

- **Why is My Classifier Discriminatory?**
  Irene Chen, Fredrik Johansson, David Sontag
  NeurIPS 2018
  https://papers.nips.cc/paper/7613-why-is-my-classifier-discriminatory.pdf