JOHNS HOPKINS
UNIVERSITY

Proactively Addressing Failures to Engineer Reliable Machine Learning Systems

Adarsh Subbaswamy*
work with Suchi Saria**

* PhD candidate, Computer Science (asubbaswamy@jhu.edu)
**Associate Professor, Computer Science (ssaria@cs.jhu.edu)

@suchisaria @_asubbaswamy

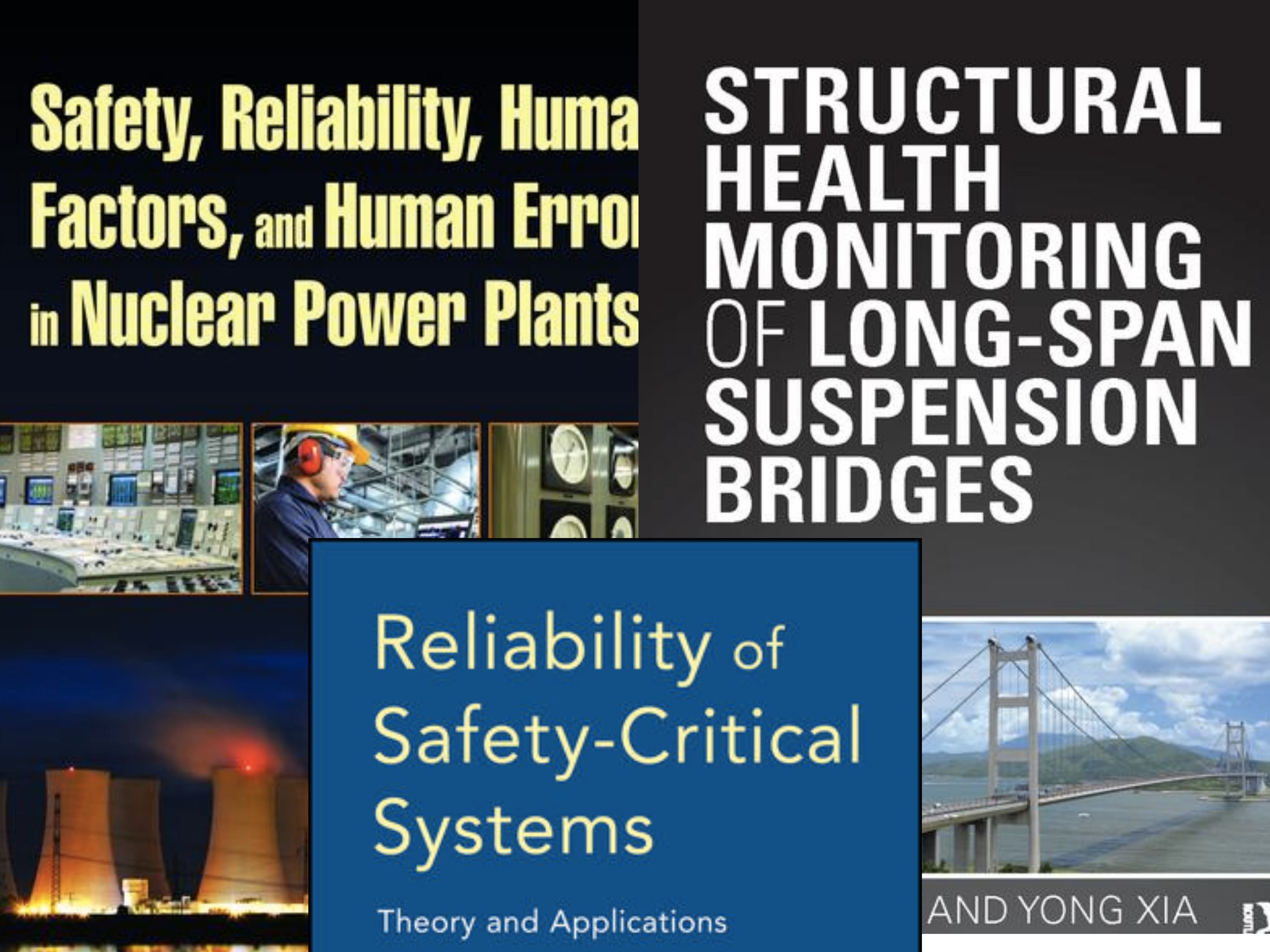
EDUCATION

CRIMINAL JUSTICE

ML

HEALTHCARE

TRANSPORTATION



Safety, Reliability, Huma
Factors, and Human Erro
in Nuclear Power Plants

STRUCTURAL
HEALTH
MONITORING
OF LONG-SPAN
SUSPENSION
BRIDGES

AND YONG XIA

Reliability of
Safety-Critical
Systems

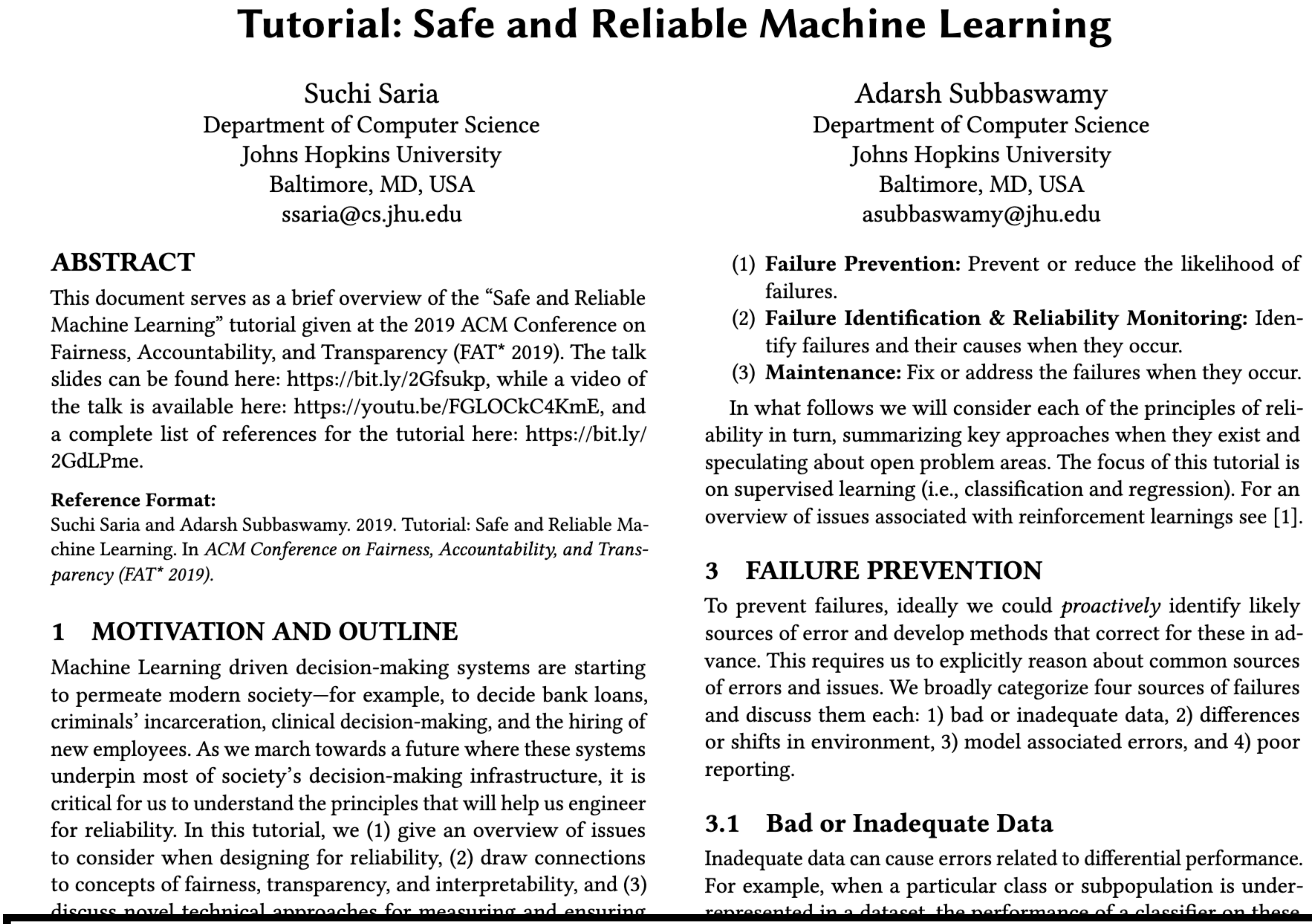
Theory and Applications

Principle Areas of Reliability Engineering

1. Prevent or reduce the likelihood of failures [**Failure Prevention**]

2. Identify failures and their causes when they occur [**Failure identification**] and [**Reliability Monitoring**]

3. Fix the failures when they occur [**Maintenance**]



Tutorial: Safe and Reliable Machine Learning

Suchi Saria
Department of Computer Science
Johns Hopkins University
Baltimore, MD, USA
ssaria@cs.jhu.edu

Adarsh Subbaswamy
Department of Computer Science
Johns Hopkins University
Baltimore, MD, USA
asubbaswamy@jhu.edu

ABSTRACT

This document serves as a brief overview of the "Safe and Reliable Machine Learning" tutorial given at the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019). The talk slides can be found here: https://bit.ly/2Gfsukp, while a video of the talk is available here: https://youtu.be/FGLOCkC4KmE, and a complete list of references for the tutorial here: https://bit.ly/2GdLPme.

Reference Format:
Suchi Saria and Adarsh Subbaswamy. 2019. Tutorial: Safe and Reliable Machine Learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*.

1 MOTIVATION AND OUTLINE

Machine Learning driven decision-making systems are starting to permeate modern society—for example, to decide bank loans, criminals' incarceration, clinical decision-making, and the hiring of new employees. As we march towards a future where these systems underpin most of society's decision-making infrastructure, it is critical for us to understand the principles that will help us engineer for reliability. In this tutorial, we (1) give an overview of issues to consider when designing for reliability, (2) draw connections to concepts of fairness, transparency, and interpretability, and (3) discuss novel technical approaches for measuring and ensuring

(1) **Failure Prevention:** Prevent or reduce the likelihood of failures.
(2) **Failure Identification & Reliability Monitoring:** Identify failures and their causes when they occur.
(3) **Maintenance:** Fix or address the failures when they occur.

In what follows we will consider each of the principles of reliability in turn, summarizing key approaches when they exist and speculating about open problem areas. The focus of this tutorial is on supervised learning (i.e., classification and regression). For an overview of issues associated with reinforcement learnings see [1].

3 FAILURE PREVENTION

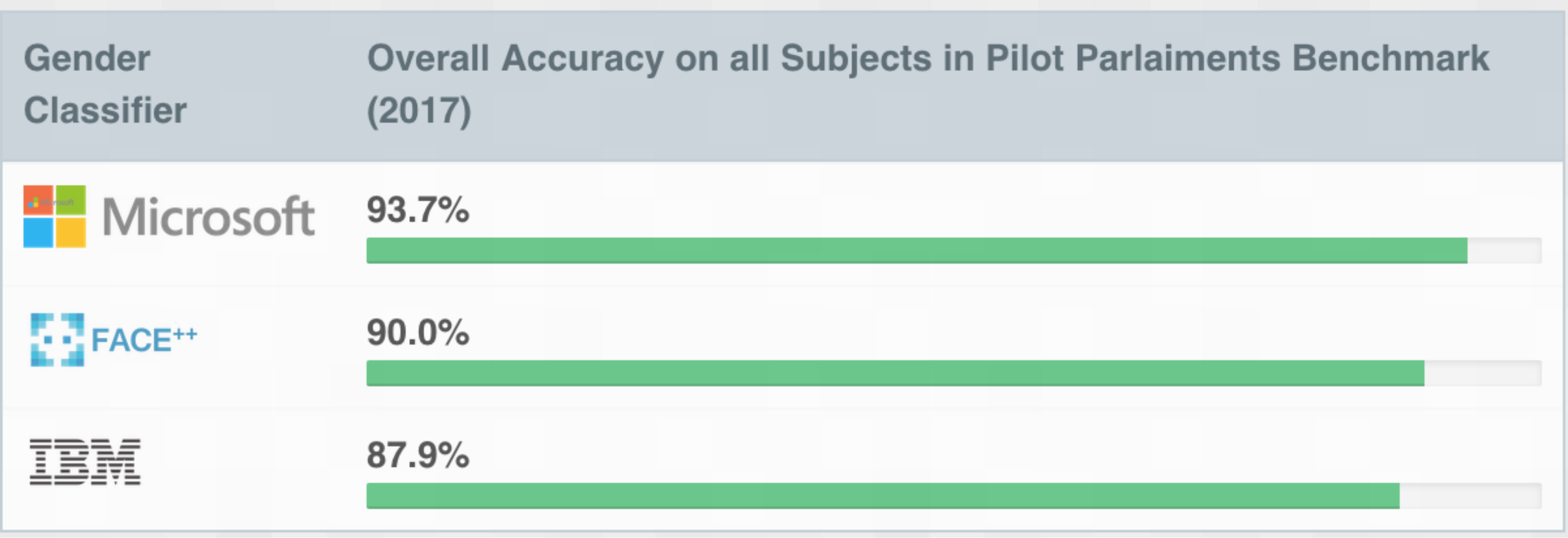
To prevent failures, ideally we could *proactively* identify likely sources of error and develop methods that correct for these in advance. This requires us to explicitly reason about common sources of errors and issues. We broadly categorize four sources of failures and discuss them each: 1) bad or inadequate data, 2) differences or shifts in environment, 3) model associated errors, and 4) poor reporting.

3.1 Bad or Inadequate Data

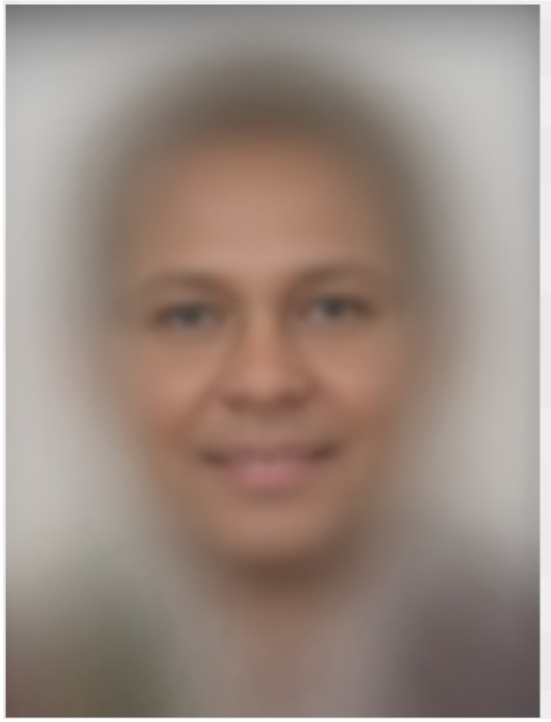
Inadequate data can cause errors related to differential performance. For example, when a particular class or subpopulation is underrepresented in a dataset, the performance of a classifier on these

Saria, S. and Subbaswamy, A. (2019). **Tutorial: Safe and Reliable Machine Learning**. ACM Conference on Fairness, Accountability, and Transparency. **https://arxiv.org/abs/1904.07204**

Inadequate Data

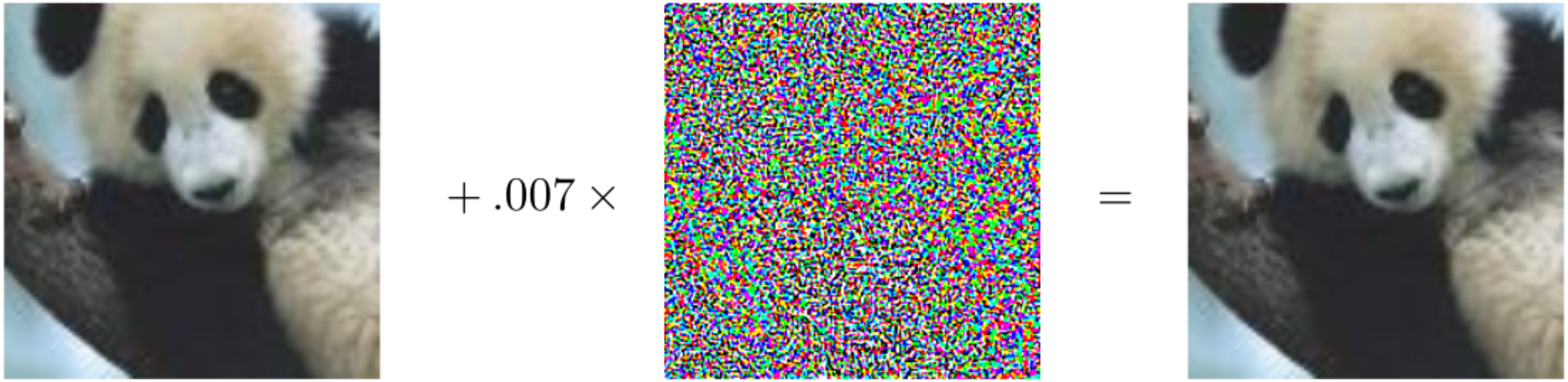


Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%



Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." *Conference on Fairness, Accountability and Transparency*. 2018. http://gendershades.org/

Adversarial Blindspots



$\boldsymbol{x}$

"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

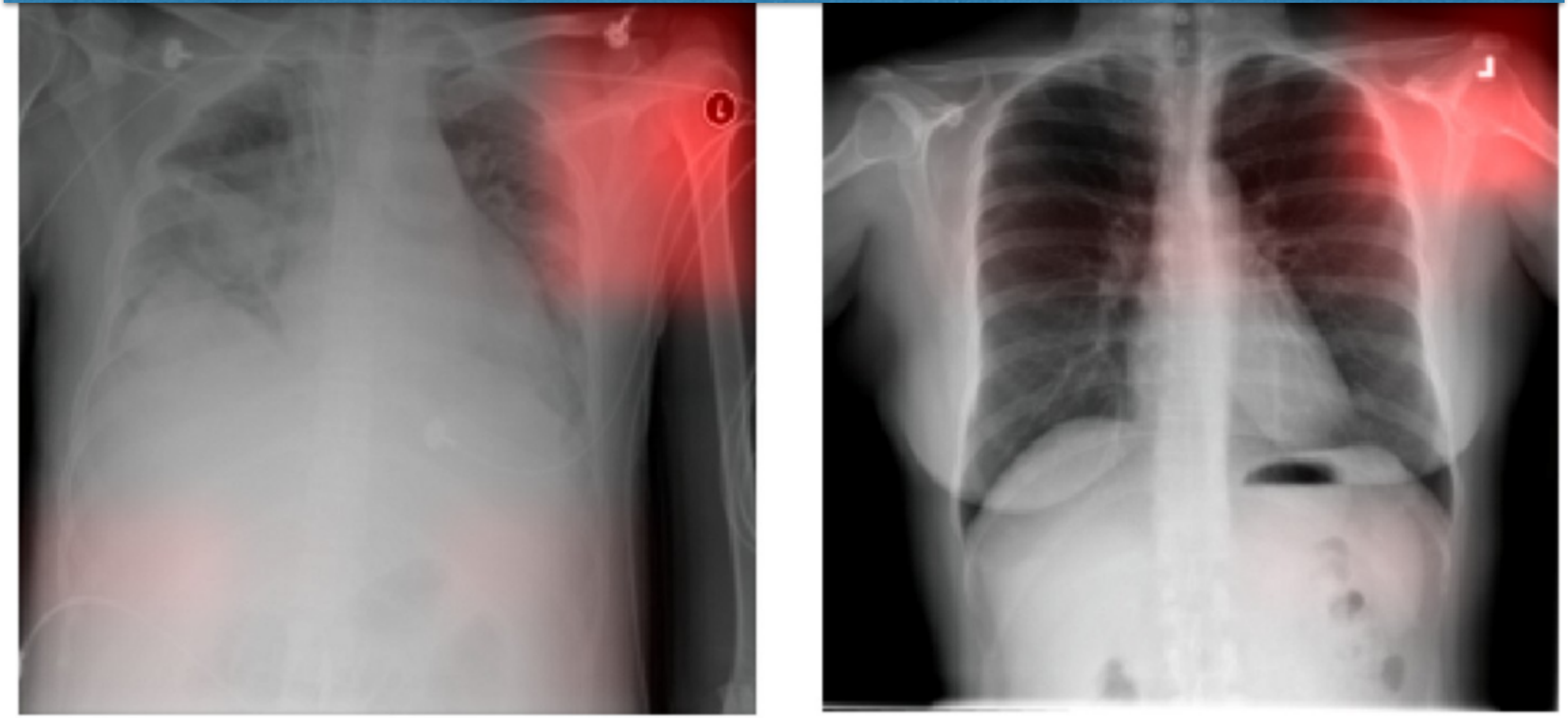
$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

- Model becomes confidently wrong
- Failure complementary to *dataset shift: despite proactive correction, model can be susceptible to this.*

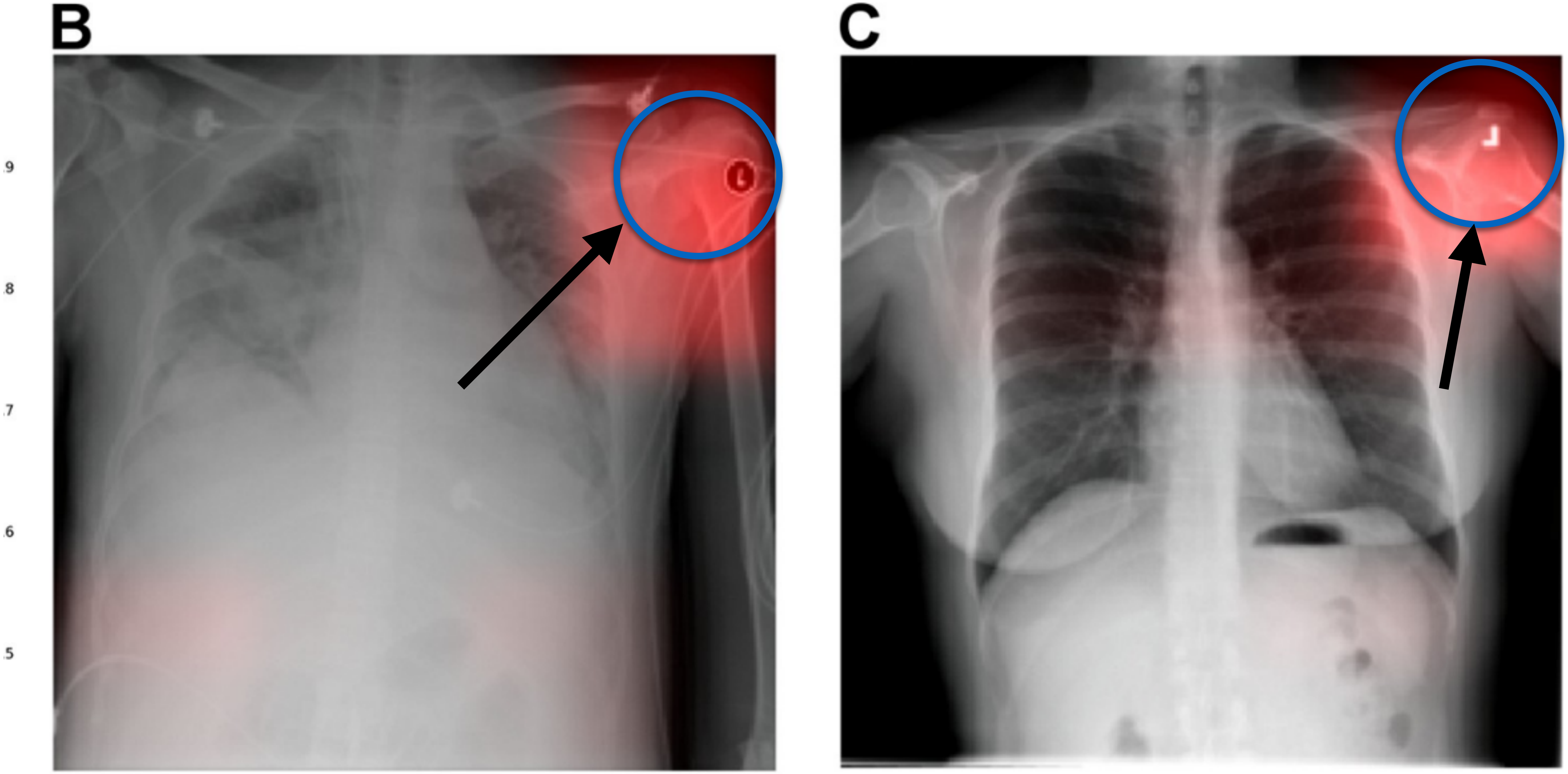
I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR), 2015.

Dataset Shift



- Goal: Use lung X-rays to diagnose pneumonia
- Developed a model using a large training dataset. Measured performance on this data. Deemed high-quality using evaluation on held-out dataset.
- When the model is evaluated **beyond** that dataset, your model performance degrades.

Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study." *PLoS medicine* 15.11 (2018): e1002683.



- X-ray has style features (tokens or inlaid text)
- Encode geometry (orientation), color scheme, etc.

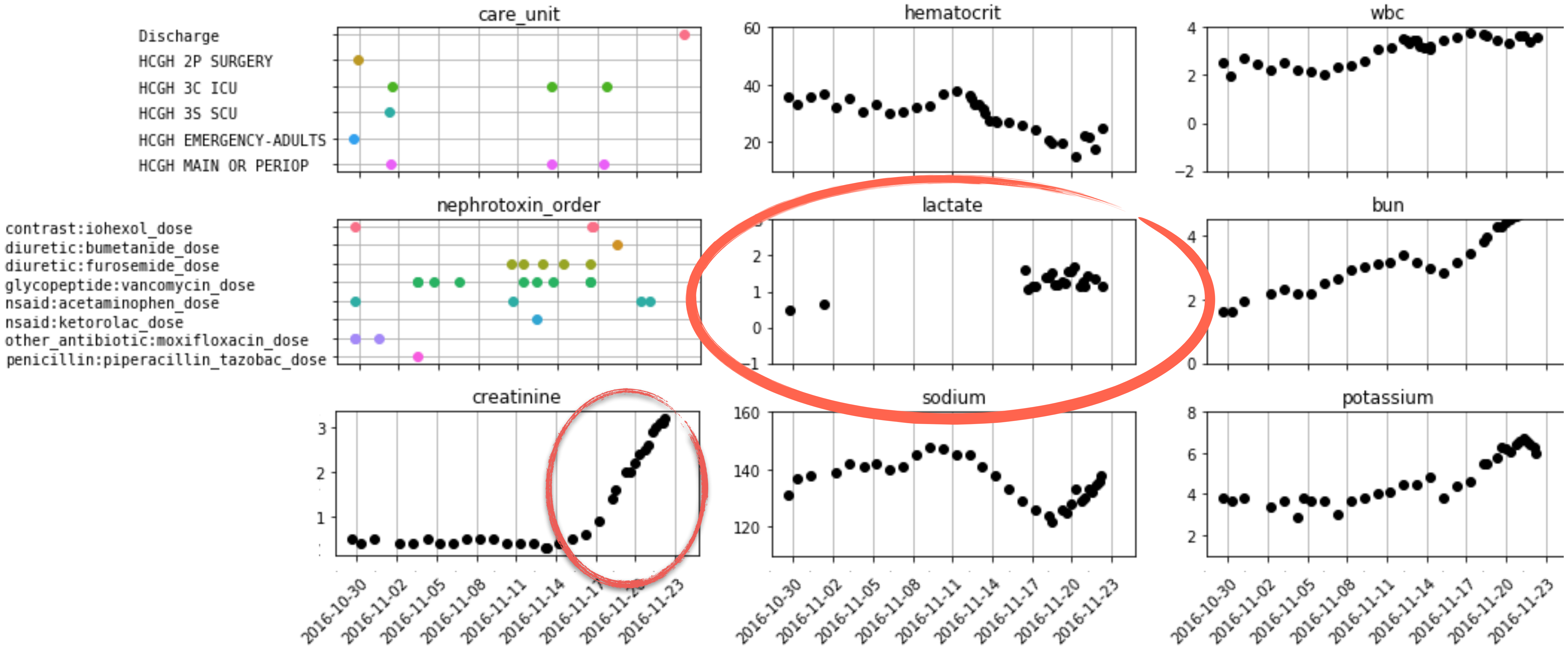
Zech, John R., et al. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study." *PLoS medicine* 15.11 (2018): e1002683.

Types of Dataset Shift

- A primary challenge: the accuracy and reliability of an ML model is dependent on the training context and the deployment context

- Context: Aspects of dataset that ML model uses to make predictions
    - **Equipment**: CT scanner manufacturer, image settings, …
    - **Population**: Patient demographics, disease prevalence, …
    - **Behavior**: Timing/frequency of lab tests and treatments

Dataset Shift Robustness is Critical: Healthcare Practice Evolves Over Time



- Goal: Use labs to predict risk of an adverse event

- Trained on data from 2011-2013 and tested on 2014, it performed very well. When tested on 2015, performance deteriorated dramatically.

- Instance of learning a dependency that does not generalize across changes in provider ordering patterns.

Schulam, P and Saria, S. "Reliable Decision Support using Counterfactual Models", Neural Information Processing Systems, 2017.
Subbaswamy, A and Saria S. "I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models." https://arxiv.org/pdf/2002.08948.pdf.

Shifts in Data Hurt Generalization

- In order to prevent failures, can we learn models that are stable to shifts?

Reactive

Use unlabeled samples from target distribution to optimize model for target environment

Proactive

Failure prevention paradigm: learn model to protect from likely problematic shifts

To start see:

Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI),* (2018).

Storkey, Amos. "When training and test sets are different: characterizing learning transfer." *Dataset shift in machine learning* (2009): 3-28.

Quionero-Candela, Joaquin, et al. "Dataset Shift in Machine Learning." (2009).

Overview: Proactive Methods

Proactive

Failure prevention paradigm: learn model to protect from likely problematic shifts

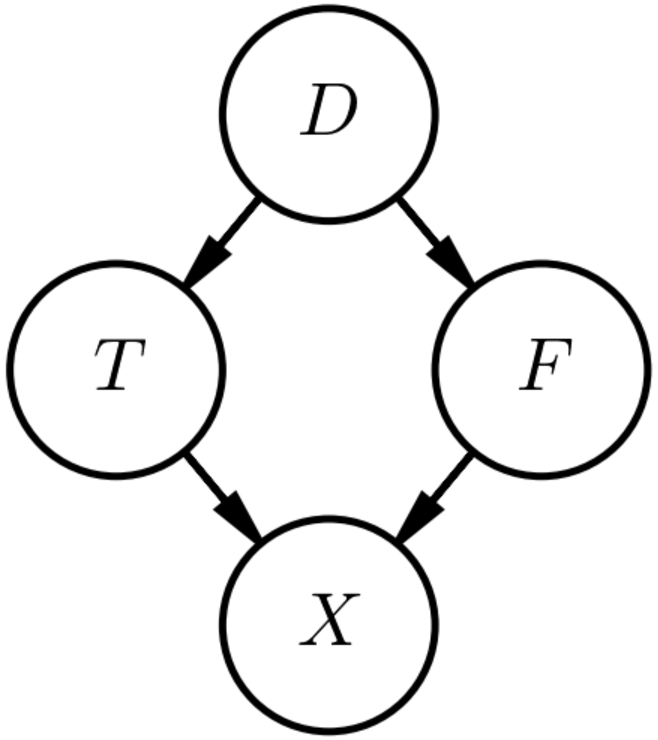
1. Represent shifts using graphs
2. Specs to identify which shifts to protect from
3. Proactive Learning (graphs, specs) ==> preprocessing step that determines which parts of the distribution to fit
4. Use existing learning techniques to fit these components
5. Guarantees:
    1. Optimality
    2. Soundness/Completeness

Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI),* (2018).
Subbaswamy, A. et al. “Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).
Schulam, P, and Suchi S. "Reliable decision support using counterfactual models." *Advances in Neural Information Processing Systems* (2017).

Pneumonia Example

- Goal: Diagnose **T** from **F** and **X**

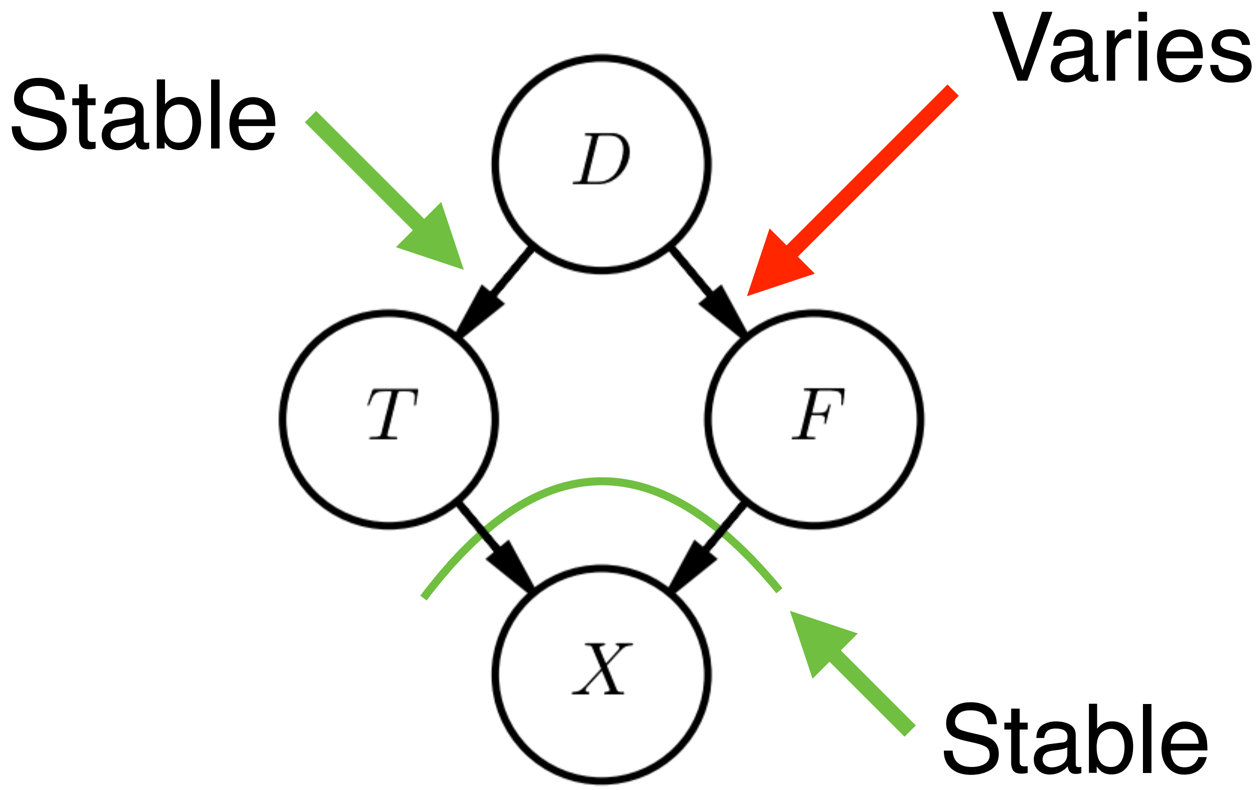
T: Pneumonia
D: Department
F: Style features
X: Lung X-ray



Data Generating Process

- Goal: Diagnose **T** from **F** and **X**

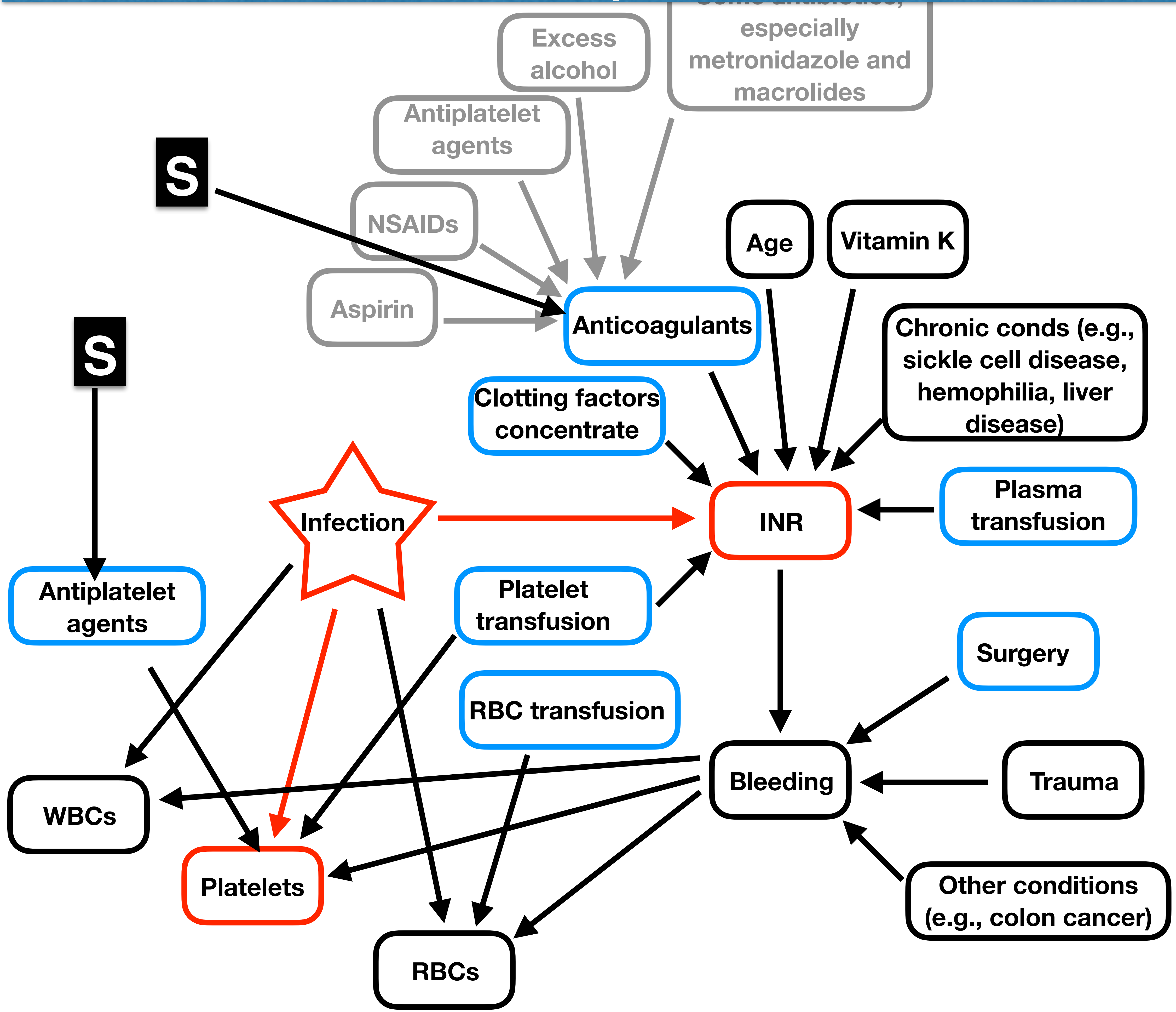
T: Pneumonia
D: Department
F: Style features
X: Lung X-ray



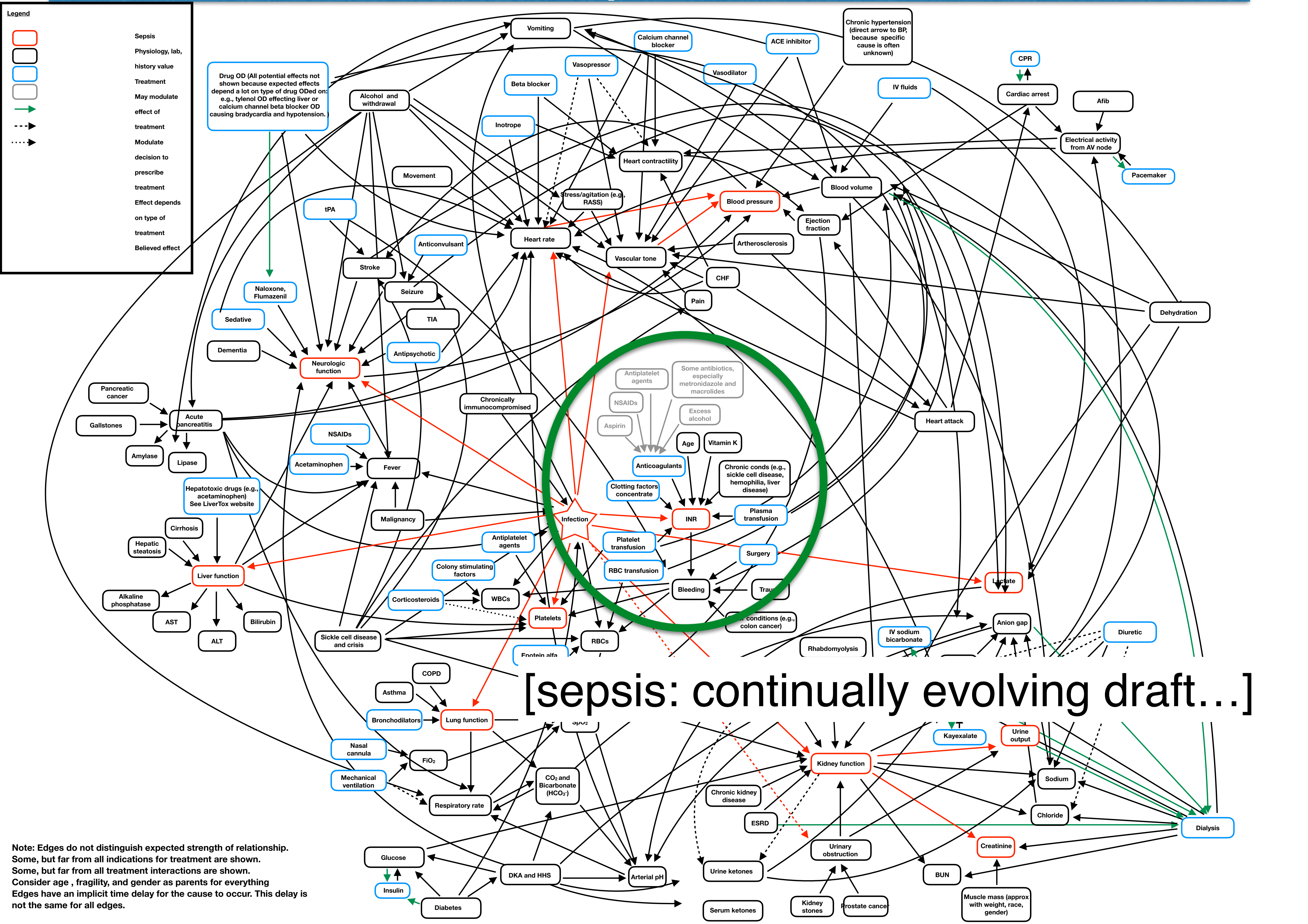
- Some of these mechanisms will be stable across environments, others are unstable and more likely to change

    - Ex: Effect of pneumonia and style on X-ray image does not change. Ex: Protocols/preferences for style features differ from department to department or even technician to technician

More complex domain…



More complex domain…



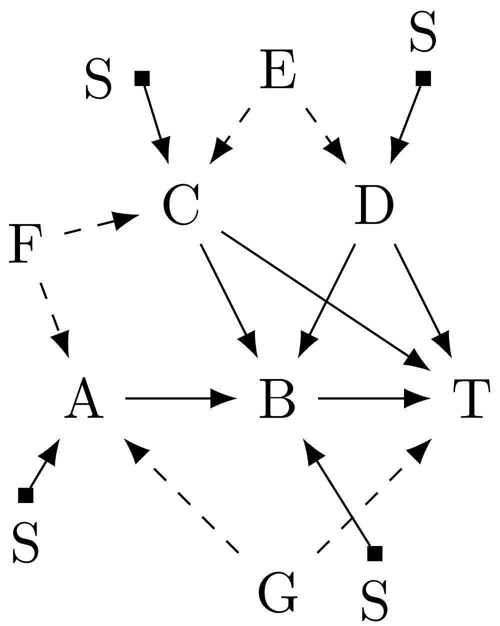
[sepsis: continually evolving draft…]

Note: Edges do not distinguish expected strength of relationship.
Some, but far from all indications for treatment are shown.
Some, but far from all treatment interactions are shown.
Consider age , fragility, and gender as parents for everything
Edges have an implicit time delay for the cause to occur. This delay is not the same for all edges.

Algorithm: Graph Surgery Estimator

Input: Graph w/ invariance specs (selection vars)

Output: Data conditionals to fit, how to combine



Target: T

Observed vars: A, B, C, D

Algorithm: Surgery Estimator

output

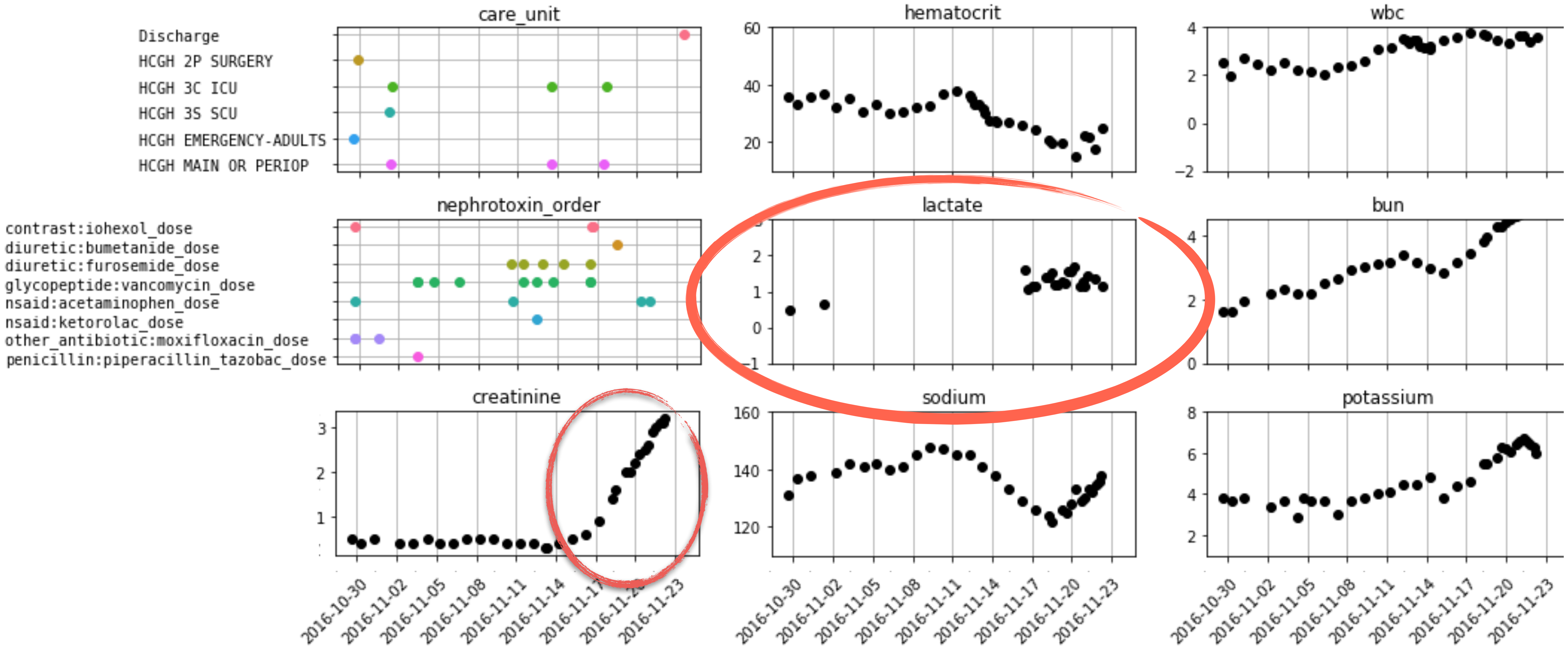
$$\left\{ \begin{matrix} P(T|A,B,C,D) \\ P(A|C,D) \\ P(D|C) \\ P(C) \end{matrix} \right\}$$

+

$$\text{Combination} \propto \sum_{A} P(T|A,B,C,D)P(A|C,D)P(D|C)P(C)$$

Subbaswamy, A et al. “Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).

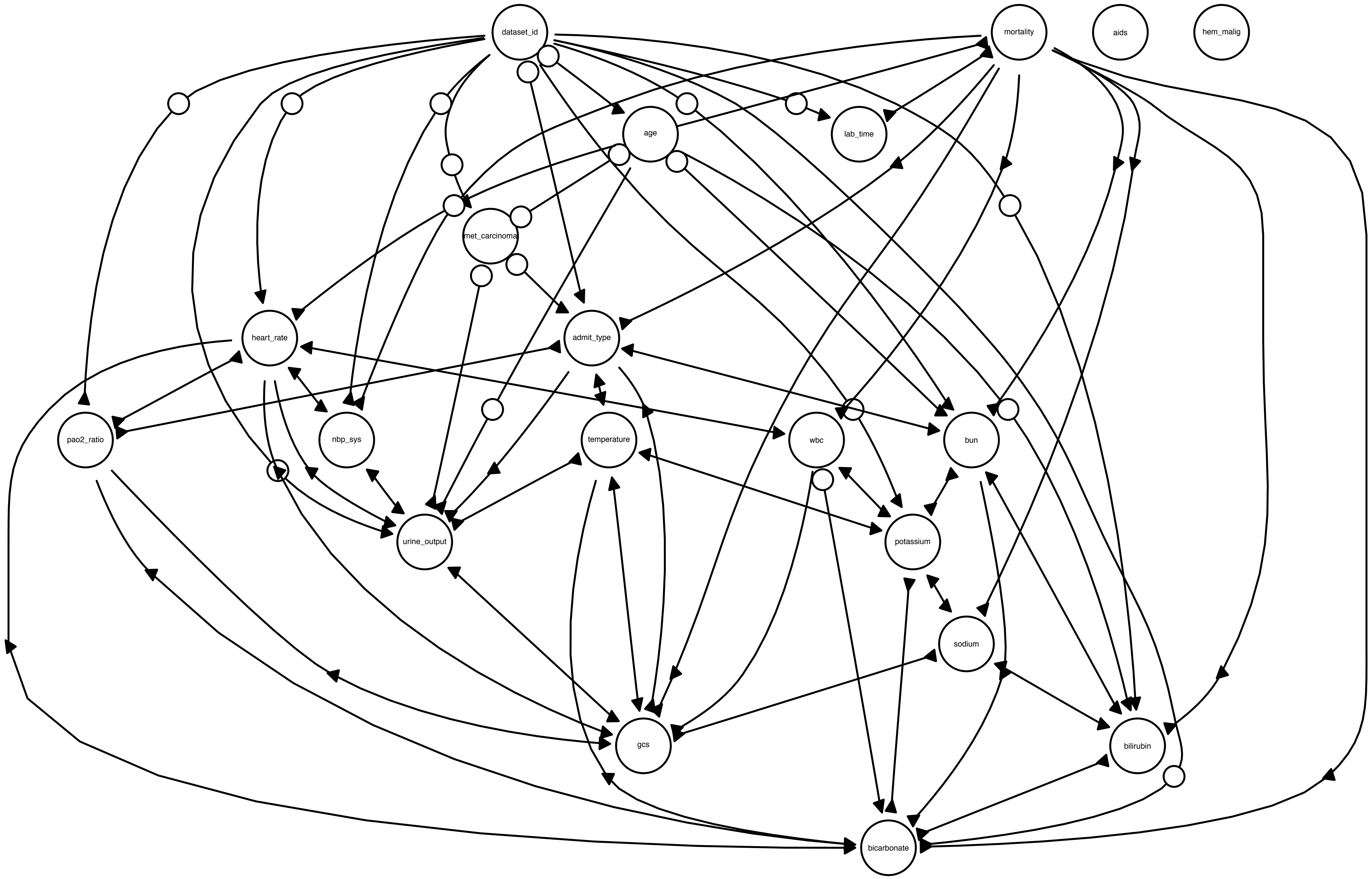
I-SPEC: Invariance Specifications



- Goal: Use labs to predict risk of an adverse event
- Trained on data from 2011-2013 and tested on 2014, it performed very well. When tested on 2015, performance deteriorated dramatically.
- Instance of learning a dependency that does not generalize across changes in provider ordering patterns.

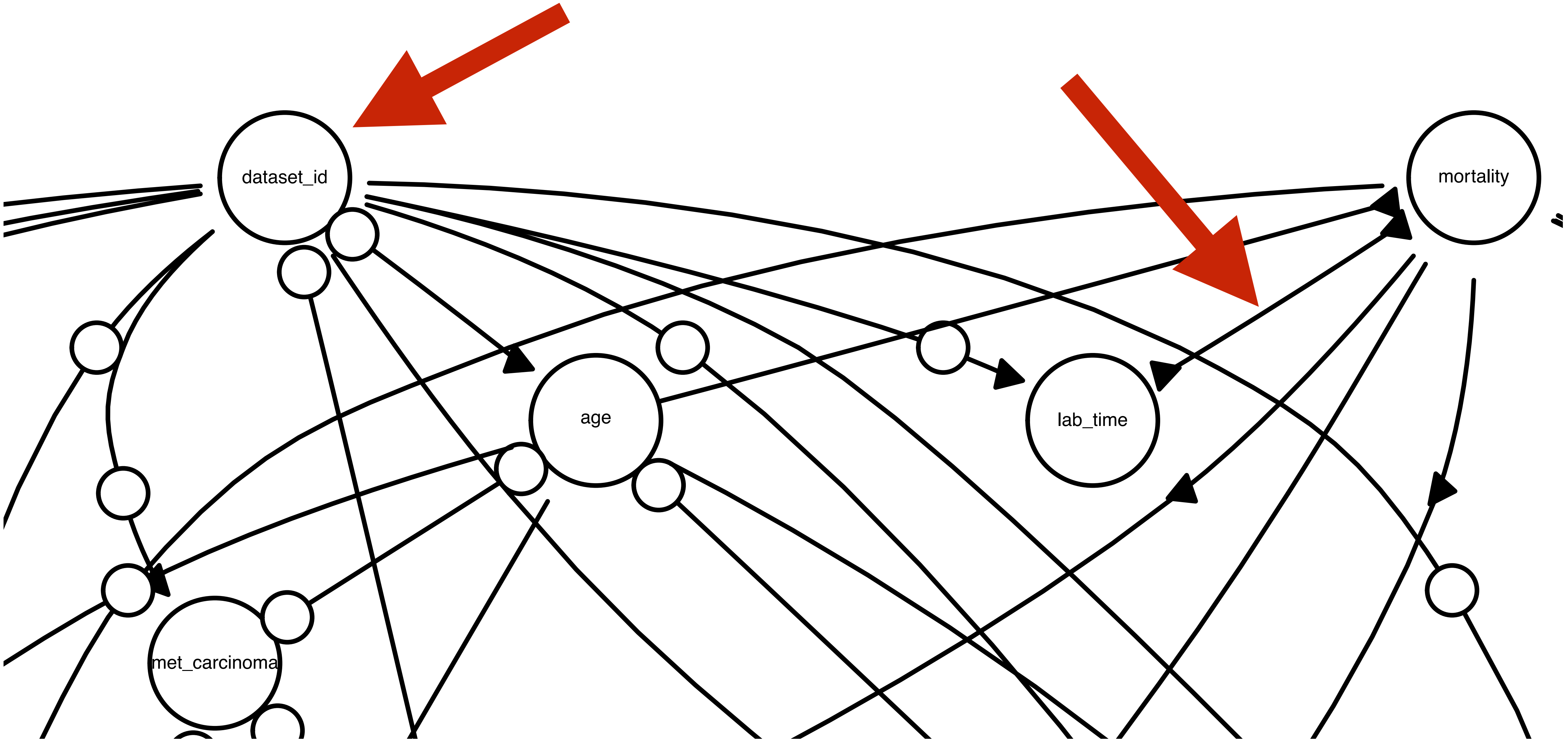
Schulam, P and Saria, S. “Reliable Decision Support using Counterfactual Models”, Neural Information Processing Systems, 2017.
Subbaswamy, A and Saria S. “I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models.” https://arxiv.org/pdf/2002.08948.pdf.

Learn Structure of Invariance Spec



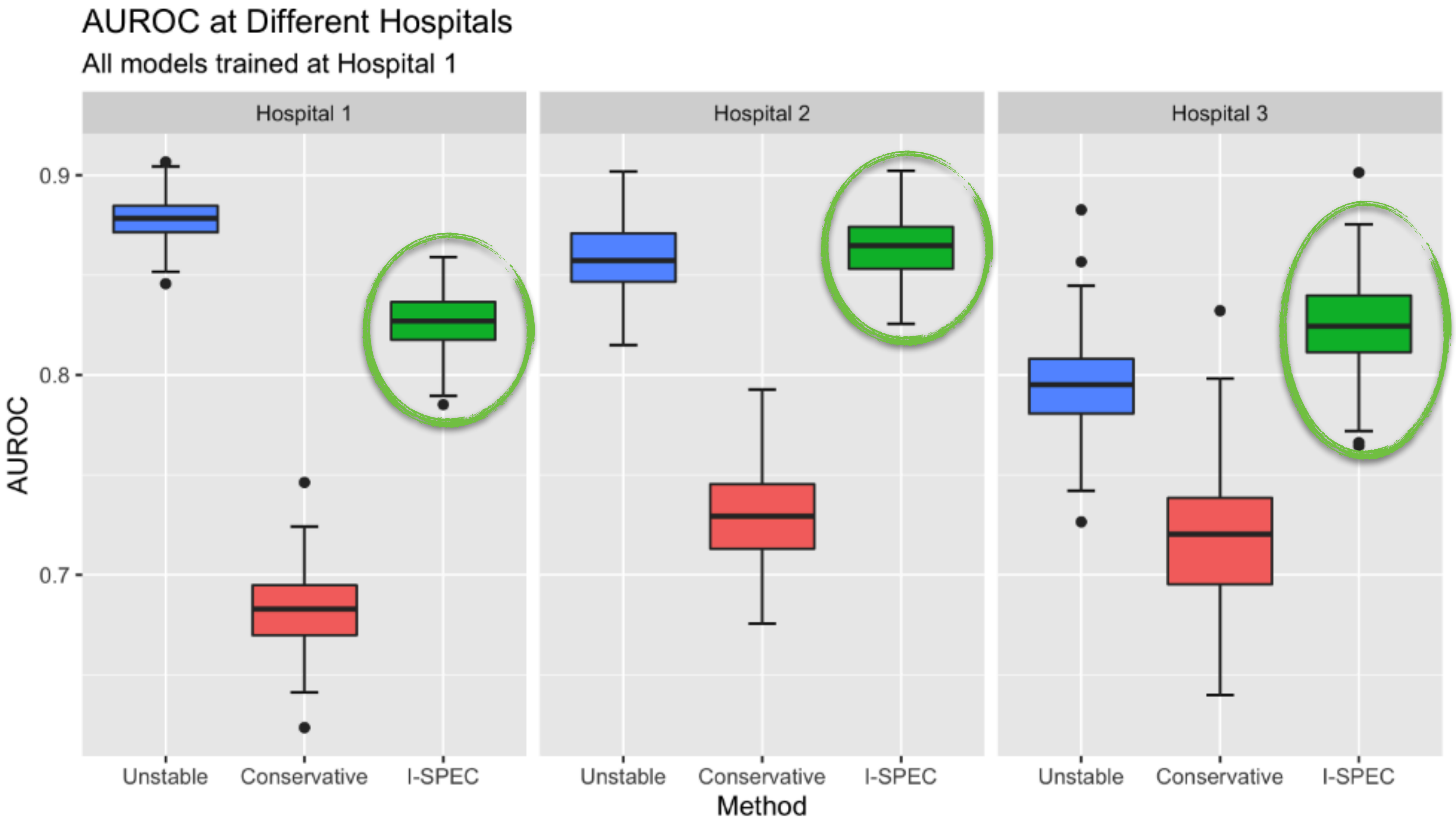
Subbaswamy, A and Saria S. "I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models." https://arxiv.org/pdf/2002.08948.pdf.

Declare Desired Invariances



- Graph suggests shifts that occurred across datasets
- Model developers complete invariance spec by declaring desired invariances
- E.g., declare we want stability to shifts lab ordering patterns

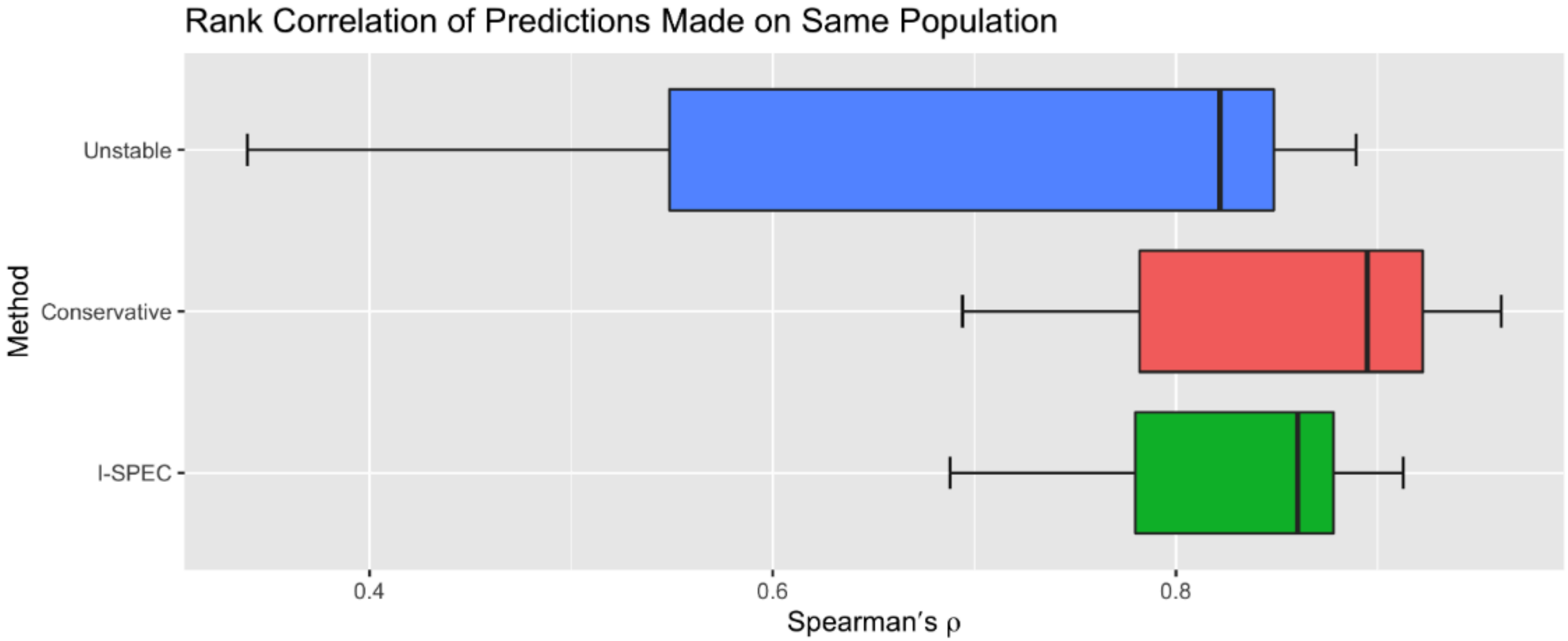
I-SPEC Implications



- Stable models have more consistent performance across sites
- Naive evaluation on a single dataset is overly optimistic
- Evaluation on multiple datasets doesn't tell you how the datasets differ

Subbaswamy, A and Saria S. "I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models." https://arxiv.org/pdf/2002.08948.pdf.

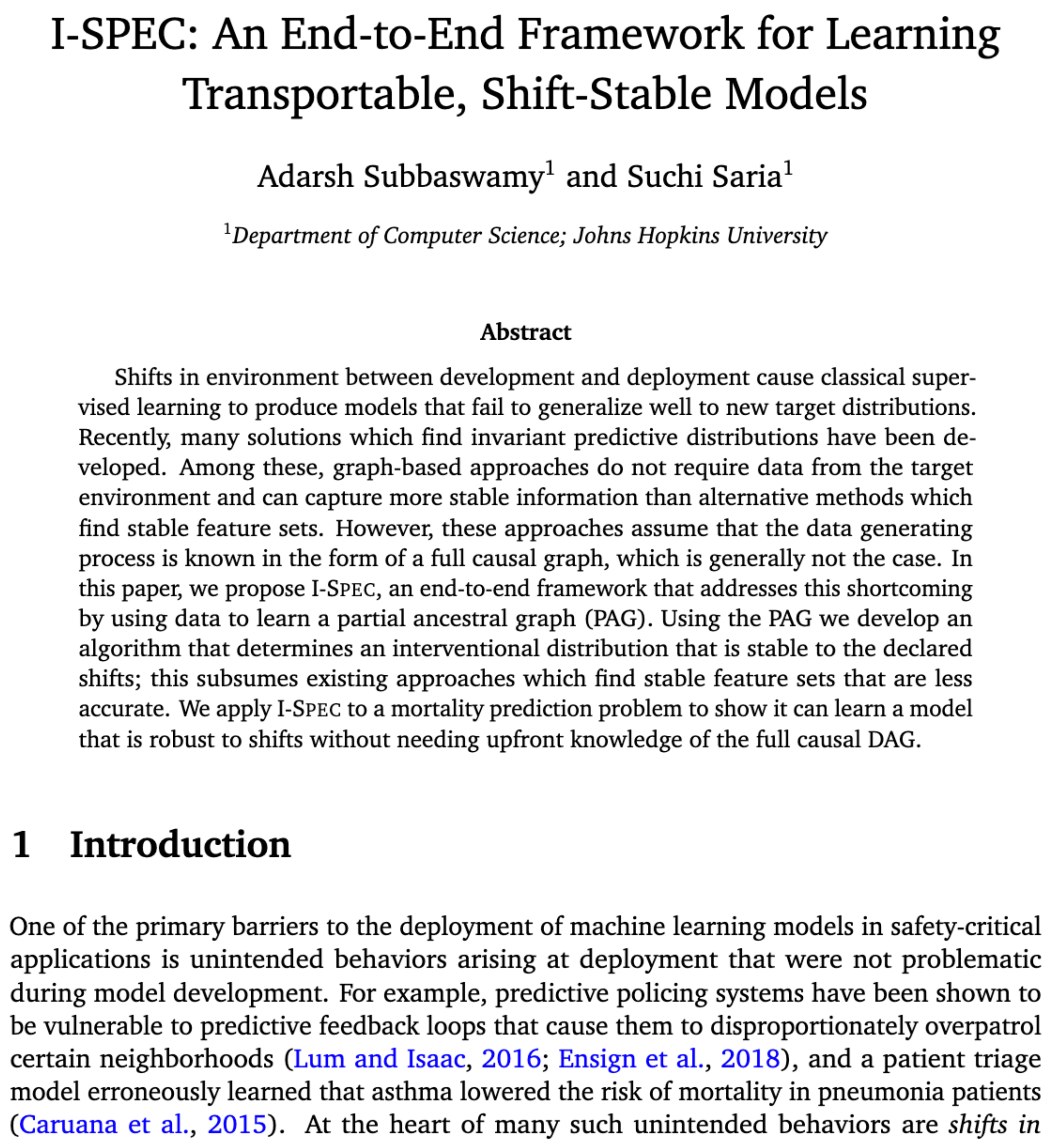
I-SPEC Implications



- More consistent risk predictions regardless of training population.
- Stable predictions yield stable decisions.

Subbaswamy, A and Saria S. “I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models.” https://arxiv.org/pdf/2002.08948.pdf.

Full Technical Paper on Framework



I-SPEC: An End-to-End Framework for Learning Transportable, Shift-Stable Models

Adarsh Subbaswamy[1] and Suchi Saria[1]

[1]*Department of Computer Science; Johns Hopkins University*

Abstract

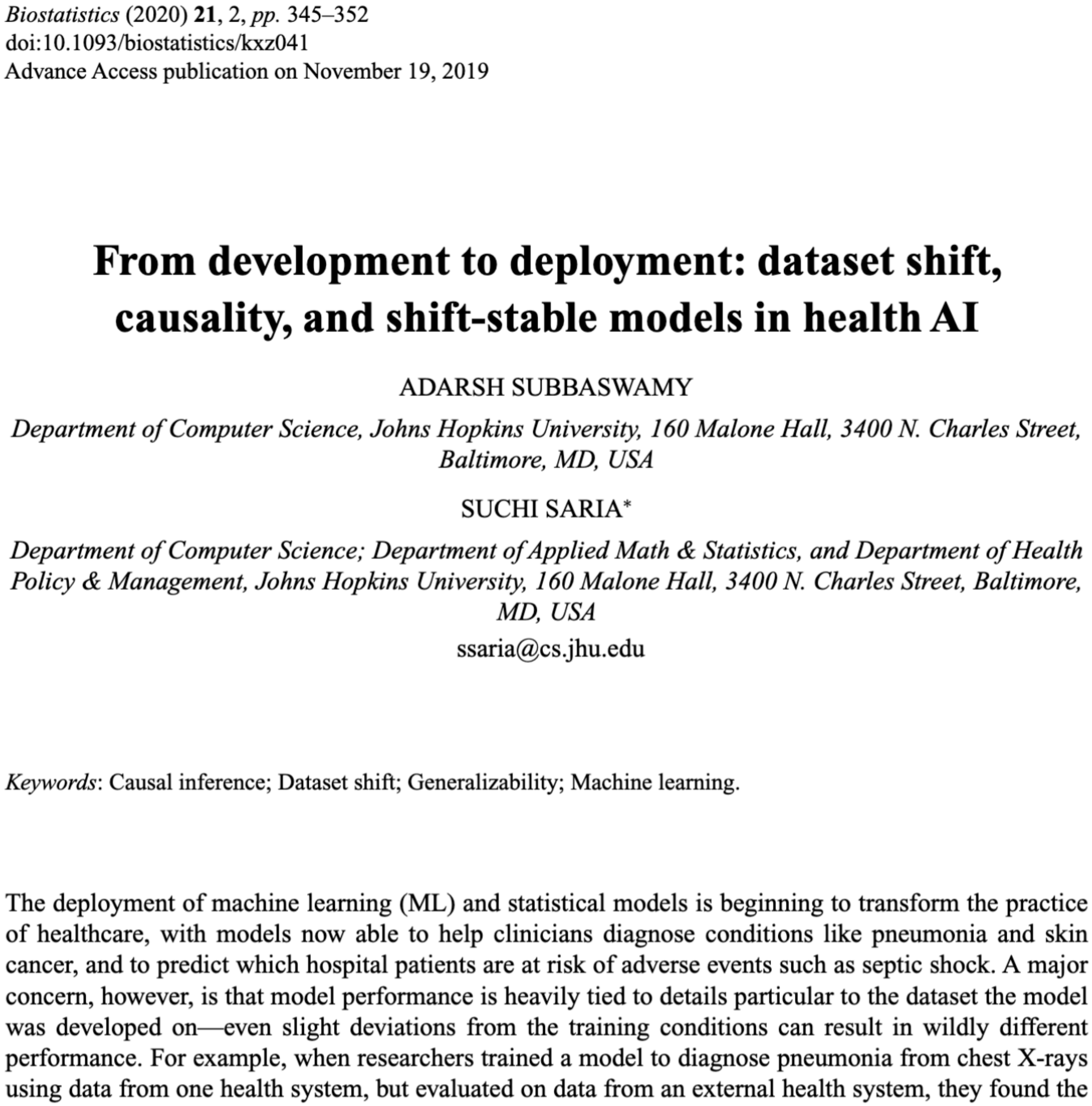
Shifts in environment between development and deployment cause classical supervised learning to produce models that fail to generalize well to new target distributions. Recently, many solutions which find invariant predictive distributions have been developed. Among these, graph-based approaches do not require data from the target environment and can capture more stable information than alternative methods which find stable feature sets. However, these approaches assume that the data generating process is known in the form of a full causal graph, which is generally not the case. In this paper, we propose I-SPEC, an end-to-end framework that addresses this shortcoming by using data to learn a partial ancestral graph (PAG). Using the PAG we develop an algorithm that determines an interventional distribution that is stable to the declared shifts; this subsumes existing approaches which find stable feature sets that are less accurate. We apply I-SPEC to a mortality prediction problem to show it can learn a model that is robust to shifts without needing upfront knowledge of the full causal DAG.

1 Introduction

One of the primary barriers to the deployment of machine learning models in safety-critical applications is unintended behaviors arising at deployment that were not problematic during model development. For example, predictive policing systems have been shown to be vulnerable to predictive feedback loops that cause them to disproportionately overpatrol certain neighborhoods (Lum and Isaac, 2016; Ensign et al., 2018), and a patient triage model erroneously learned that asthma lowered the risk of mortality in pneumonia patients (Caruana et al., 2015). At the heart of many such unintended behaviors are *shifts in*

https://arxiv.org/pdf/2002.08948.pdf.

General Overview Centered Around Healthcare Applications



Biostatistics (2020) **21**, 2, *pp.* 345–352
doi:10.1093/biostatistics/kxz041
Advance Access publication on November 19, 2019

From development to deployment: dataset shift, causality, and shift-stable models in health AI

ADARSH SUBBASWAMY

Department of Computer Science, Johns Hopkins University, 160 Malone Hall, 3400 N. Charles Street, Baltimore, MD, USA

SUCHI SARIA*

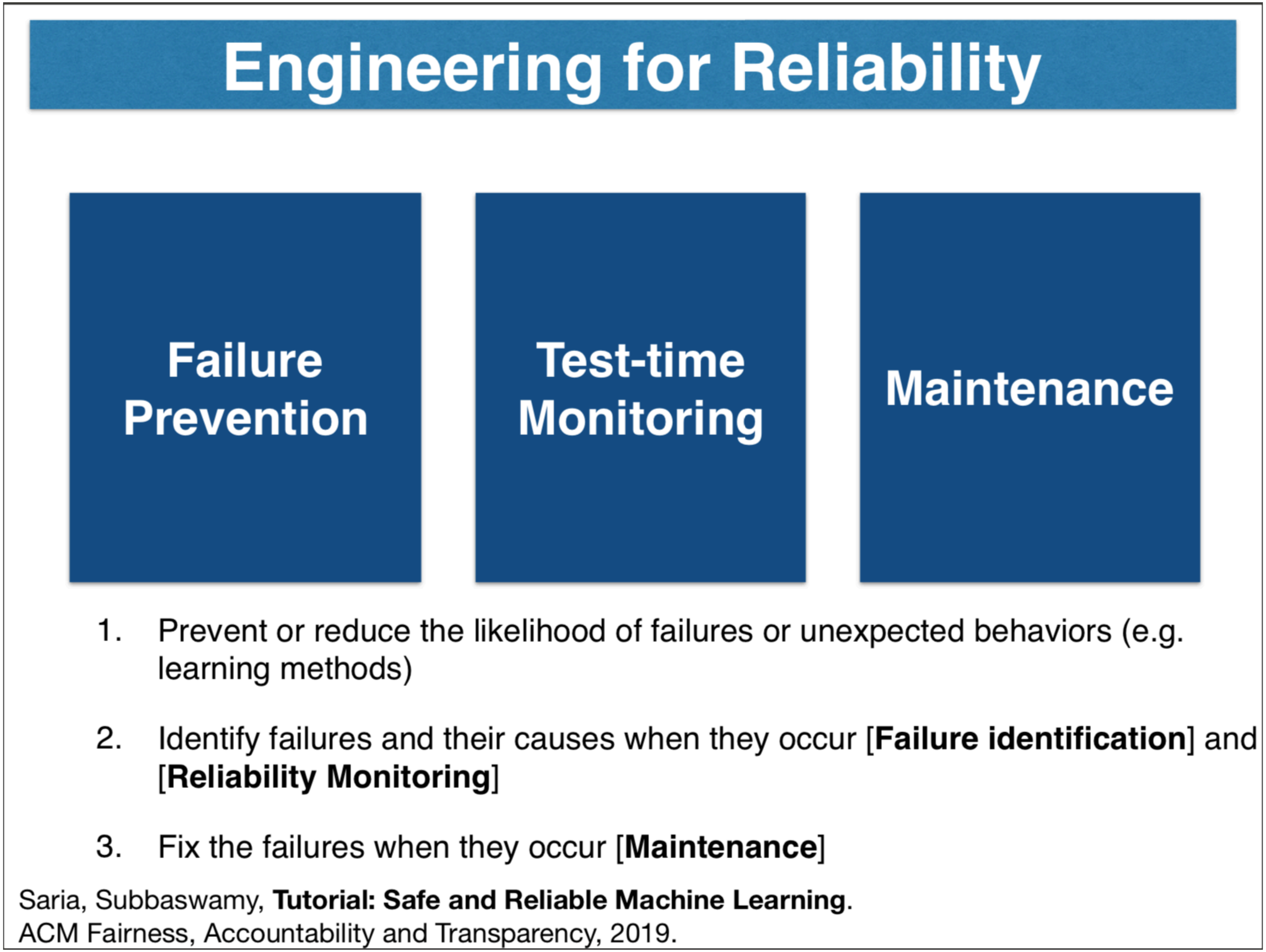
Department of Computer Science; Department of Applied Math & Statistics, and Department of Health Policy & Management, Johns Hopkins University, 160 Malone Hall, 3400 N. Charles Street, Baltimore, MD, USA

ssaria@cs.jhu.edu

Keywords: Causal inference; Dataset shift; Generalizability; Machine learning.

The deployment of machine learning (ML) and statistical models is beginning to transform the practice of healthcare, with models now able to help clinicians diagnose conditions like pneumonia and skin cancer, and to predict which hospital patients are at risk of adverse events such as septic shock. A major concern, however, is that model performance is heavily tied to details particular to the dataset the model was developed on—even slight deviations from the training conditions can result in wildly different performance. For example, when researchers trained a model to diagnose pneumonia from chest X-rays using data from one health system, but evaluated on data from an external health system, they found the

https://academic.oup.com/biostatistics/article/21/2/345/5631850



Engineering for Reliability

Failure Prevention

Test-time Monitoring

Maintenance

1. Prevent or reduce the likelihood of failures or unexpected behaviors (e.g. learning methods)

2. Identify failures and their causes when they occur [**Failure identification**] and [**Reliability Monitoring**]

3. Fix the failures when they occur [**Maintenance**]

Saria, Subbaswamy, **Tutorial: Safe and Reliable Machine Learning**.
ACM Fairness, Accountability and Transparency, 2019.

Thank you!