Safety Challenges with Black-Box Predictors and Novel Learning Approaches for Failure Proofing

Suchi Saria* and Adarsh Subbaswamy**

*John C. Malone Assistant Professor,
Computer Science, Stats, and Health Policy (ssaria@cs.jhu.edu)

** PhD candidate, Computer Science (asubbaswamy@jhu.edu)

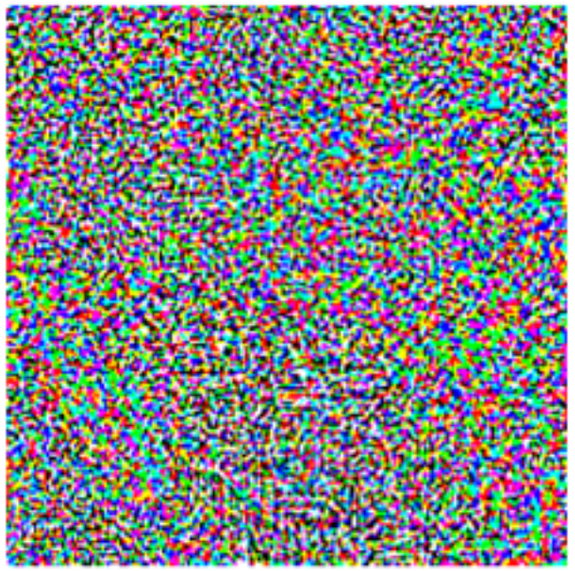
@suchisaria @_asubbaswamy

Adversarial Blindspots

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



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

"nematode"
8.2% confidence

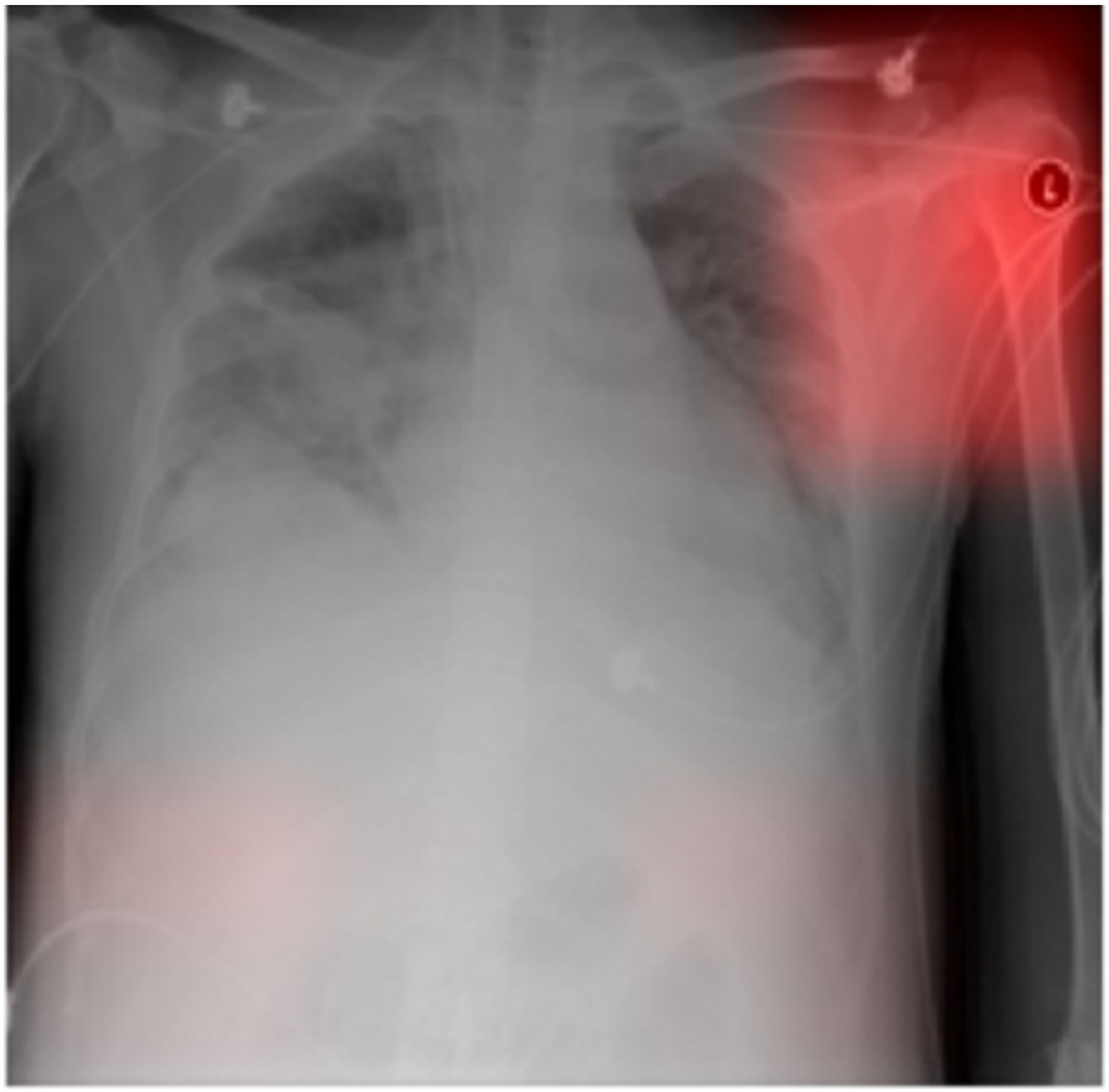
$=$

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

"gibbon"
99.3 % confidence

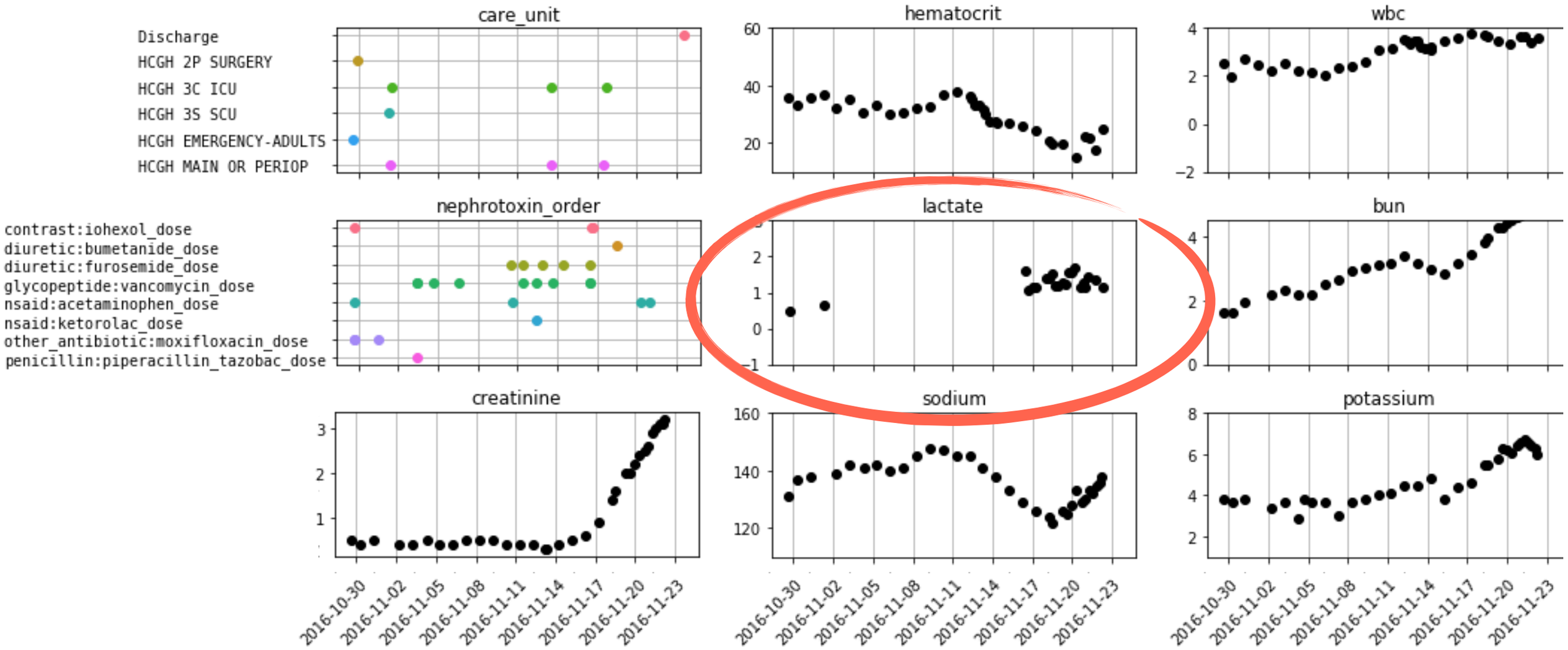
- Model becomes confidently wrong
- Failure complementary to *data shift: despite proactive correction, model can be susceptible to issue above.*

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



- 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.



- 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.

Inadequate Data

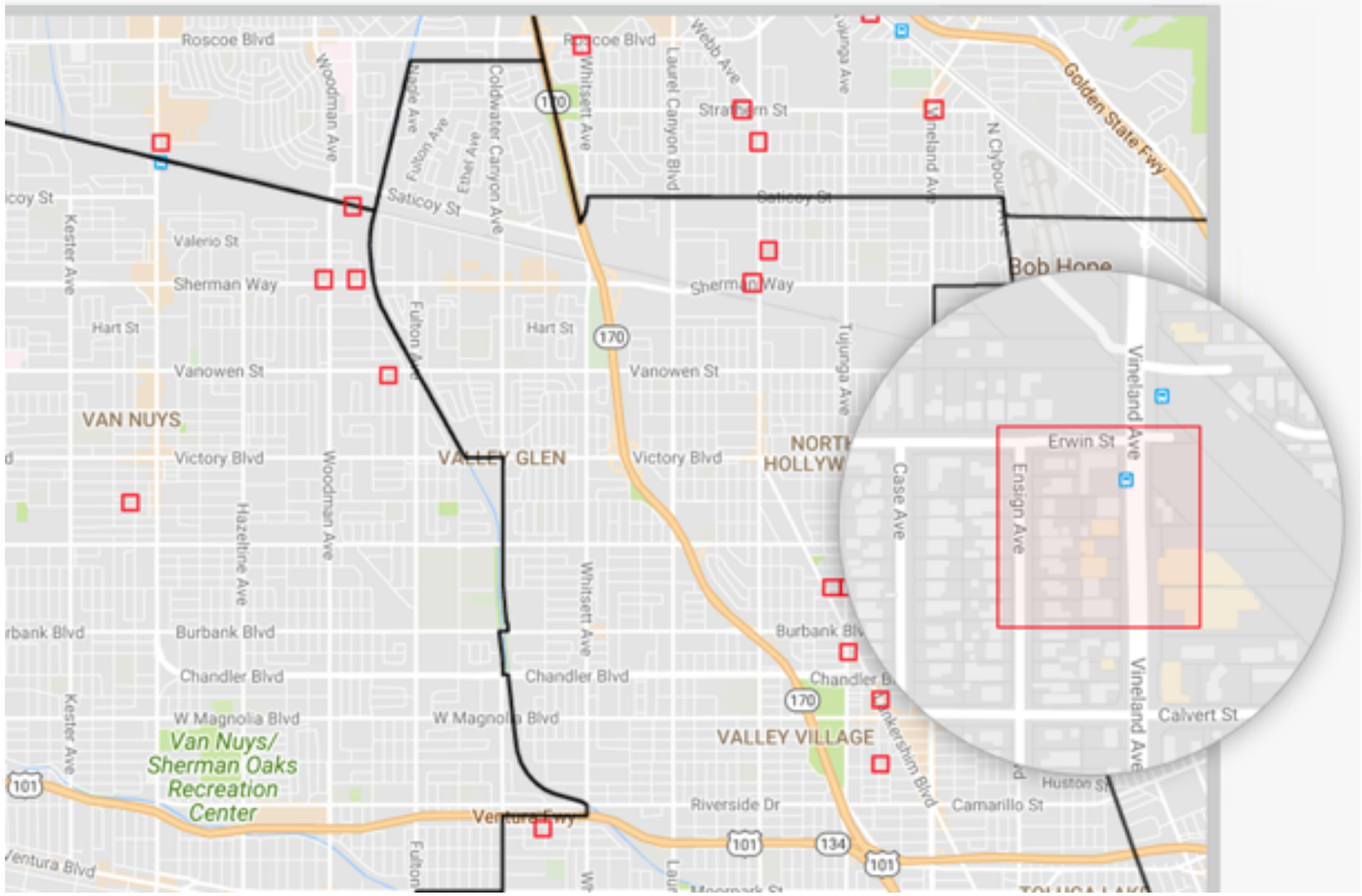
Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

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

Feedback loops

- Predictive policing: Predict when/where crime will occur

- Deploy system →sends police to same area regardless of true crime rates



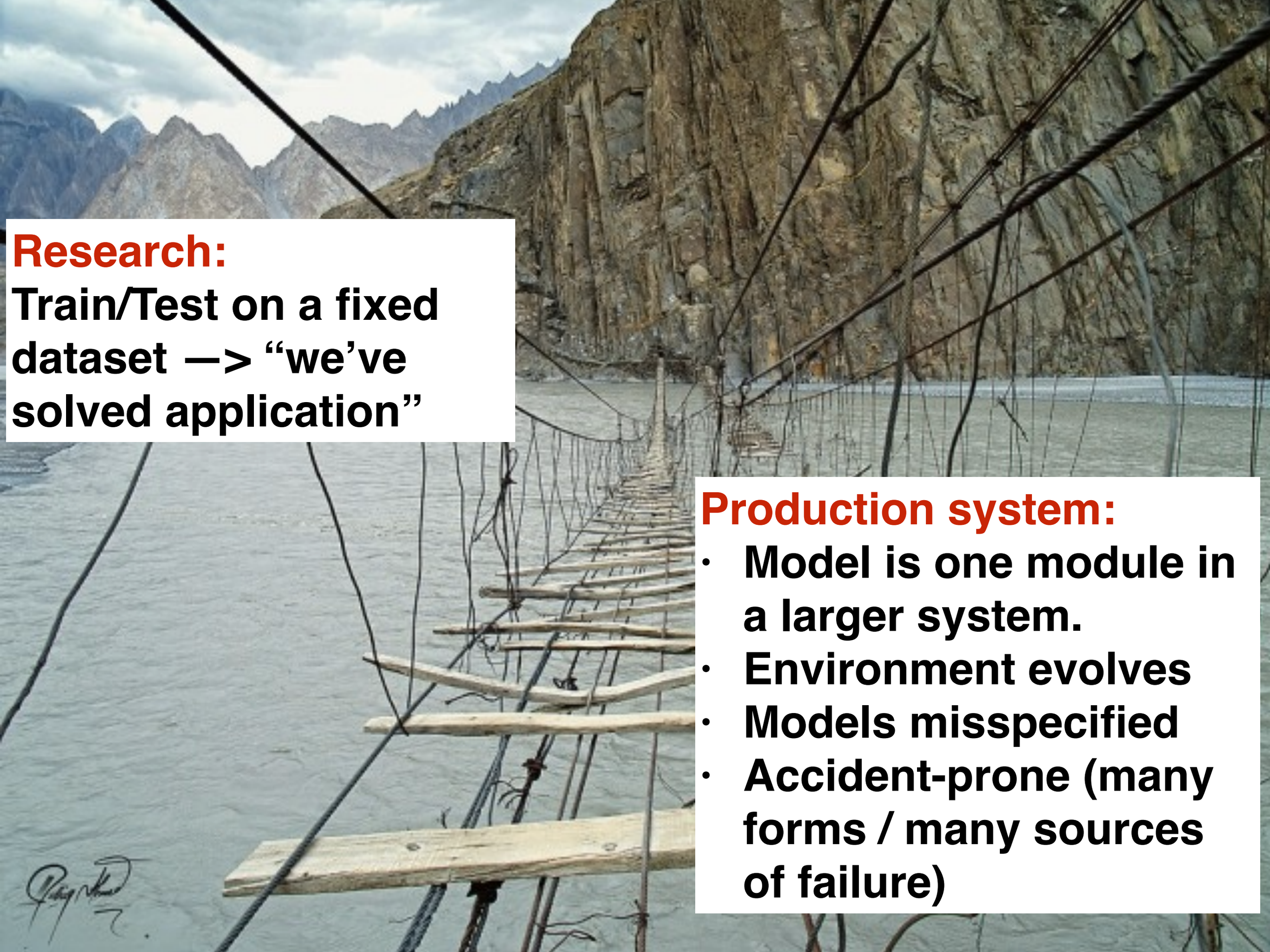
Lum, Kristian, and William Isaac. "To predict and serve?." *Significance* 13.5 (2016): 14-19.
Ensign, Danielle, et al. "Runaway Feedback Loops in Predictive Policing." *Conference on Fairness, Accountability and Transparency*. 2018.

View from the outside

ML: Bridge to the future

View from the inside

Current ML Pipelines



Research:

Train/Test on a fixed dataset —> “we’ve solved application”

Production system:

- **Model is one module in a larger system.**
- **Environment evolves**
- **Models misspecified**
- **Accident-prone (many forms / many sources of failure)**

ML

Reliability vs. Other Properties

- **Reliability**: Is the ML system behaving the way we expect it to behave?
  - **Robustness:** Technical approaches for tackling data "outliers"
  - **Bias:** Qualitatively characterizes what we want the system to NOT do.
- **Fairness**: How can we make sure ML systems don't discriminate? (needs understanding and clear specification for what is fair for a problem; ethics)
- **Privacy**: How can we ensure privacy when applying ML to sensitive data (privacy-preserving ML)
- **Security**: What can a malicious adversary do to a ML system? (systems security; attacks)
- **Abuse**: How do we prevent the misuse of ML systems to attack or harm people? (policy, regulation)
- **Transparency**: How can we understand what complicated ML systems are doing? (ML)
- **Policy**: How do we predict and respond to the economic and social consequences of ML? (policy)
- **System Failures**: Scaling issues, networking, cloud infrastructure (CS systems, networking)

Saria, S. and Subbaswamy, A. (2019). **Tutorial: Safe and Reliable Machine Learning**. ACM Conference on Fairness, Accountability, and Transparency, Atlanta, Ga.

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

STRUCTURAL HEALTH MONITORING OF LONG-SPAN SUSPENSION BRIDGES

Reliability of Safety-Critical Systems

Theory and Applications

We should treat algorithms like prescription drugs

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

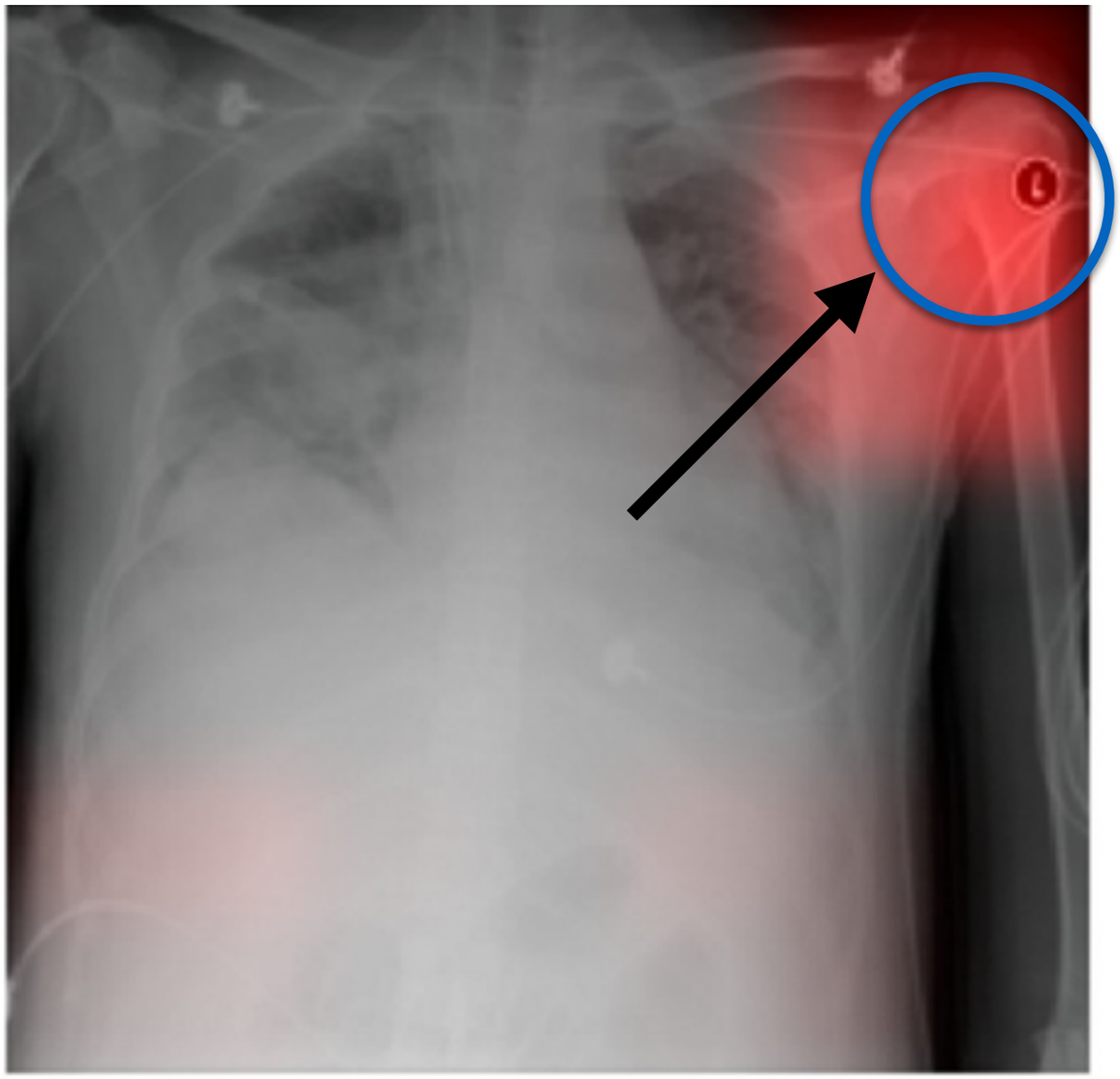
(1) **Failure Prevention:** Prevent or reduce the likelihood of failures.
(2) **Failure Identification & Reliability Monitoring:** Identify failures and their causes when they occur.
(3) **Maintenance:** Fix or address the failures when they occur.

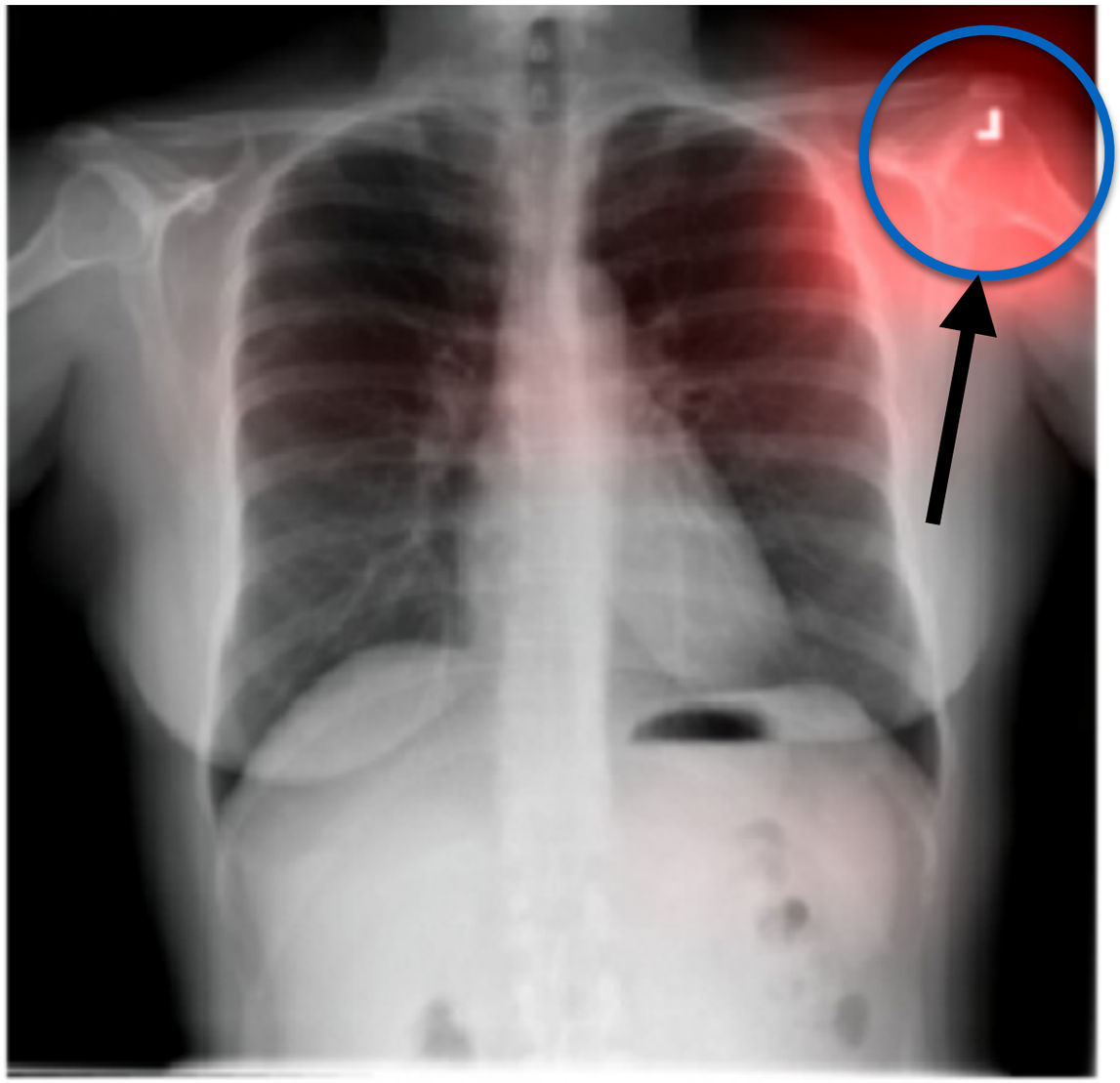
In what follows we will consider each of the principles of reliability in turn, summarizing key approaches when they exist and speculating about open problem areas. The focus of this tutorial is on supervised learning (i.e., classification and regression). For an overview of issues associated with reinforcement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

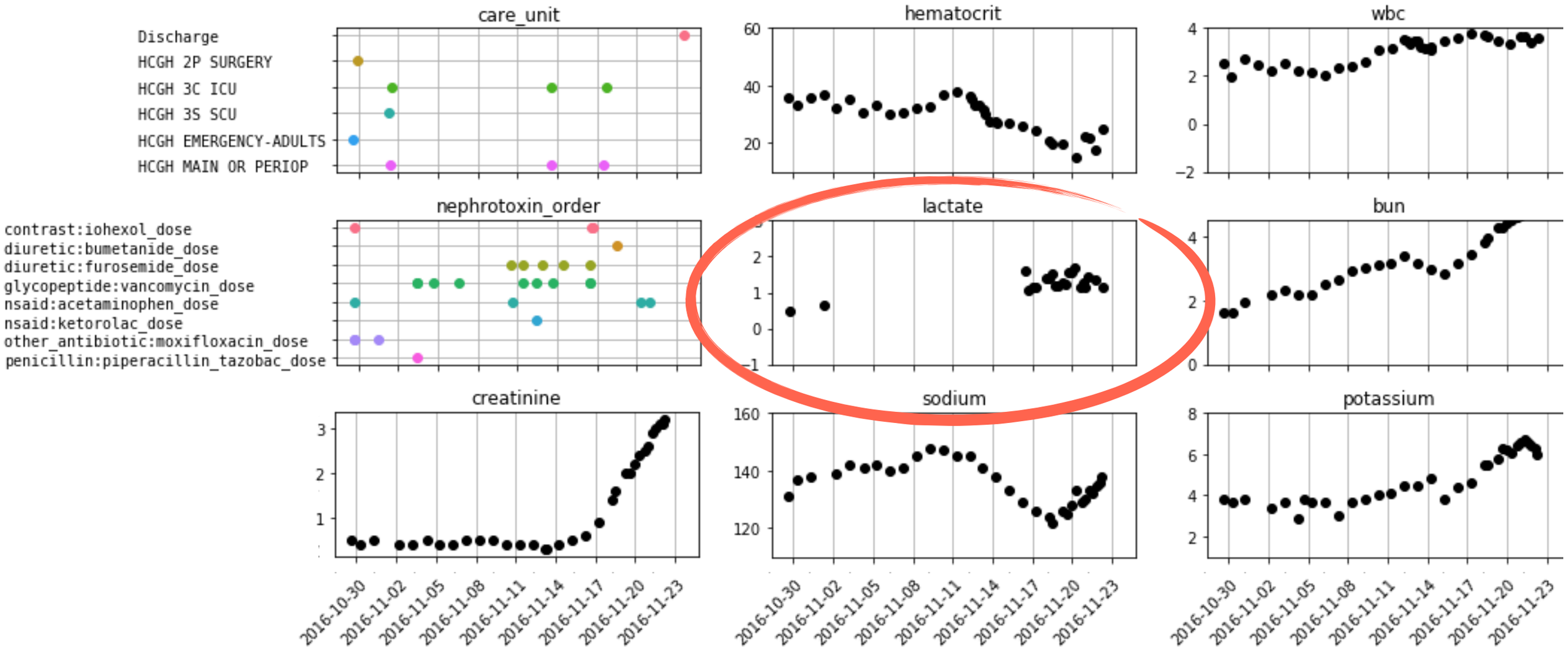
Saria, S. and Subbaswamy, A. (2019). **Tutorial: Safe and Reliable Machine Learning**. ACM Conference on Fairness, Accountability, and Transparency. **https://arxiv.org/abs/1904.07204**





- X-ray has style features (tokens or inlaid text)
- Encode geometry (orientation), color scheme, etc.
- Instance of learning relationships that do not generalize across hospitals
- Can we simply ignore style features? No, they carry information that explain image.

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.



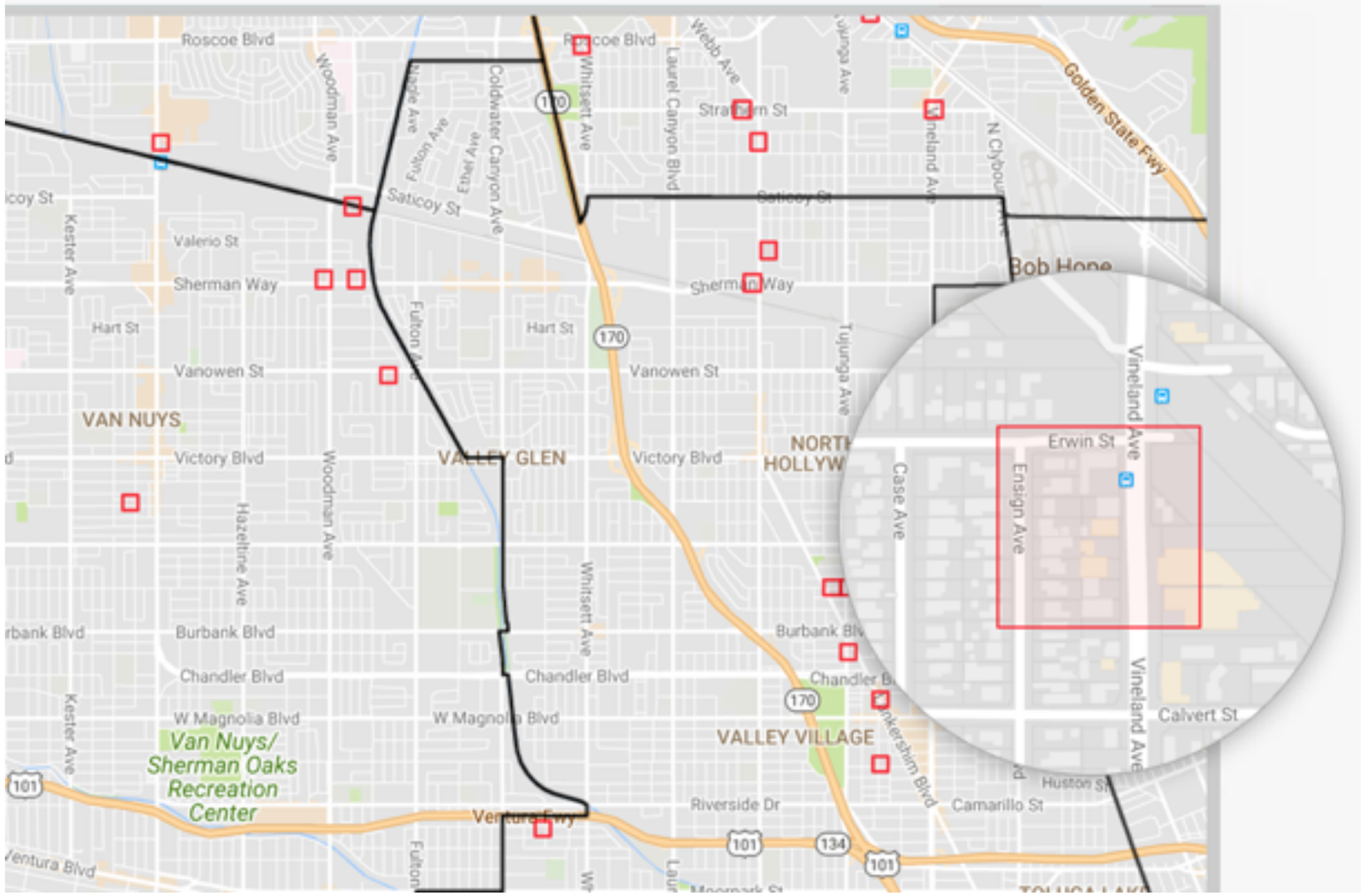
- Goal: Use labs to predict risk of an adverse event
- Trained on data from 2011-2013 and tested on 2014, it performed very well. When tested on 2015, performance deteriorated dramatically.
- Instance of learning a dependency that does not generalize across changes in provider ordering patterns.

Schulam, P and Saria, S. “Reliable Decision Support using Counterfactual Models”, Neural Information Processing Systems, 2017.

Feedback loops

- Predictive policing: Predict when/where crime will occur

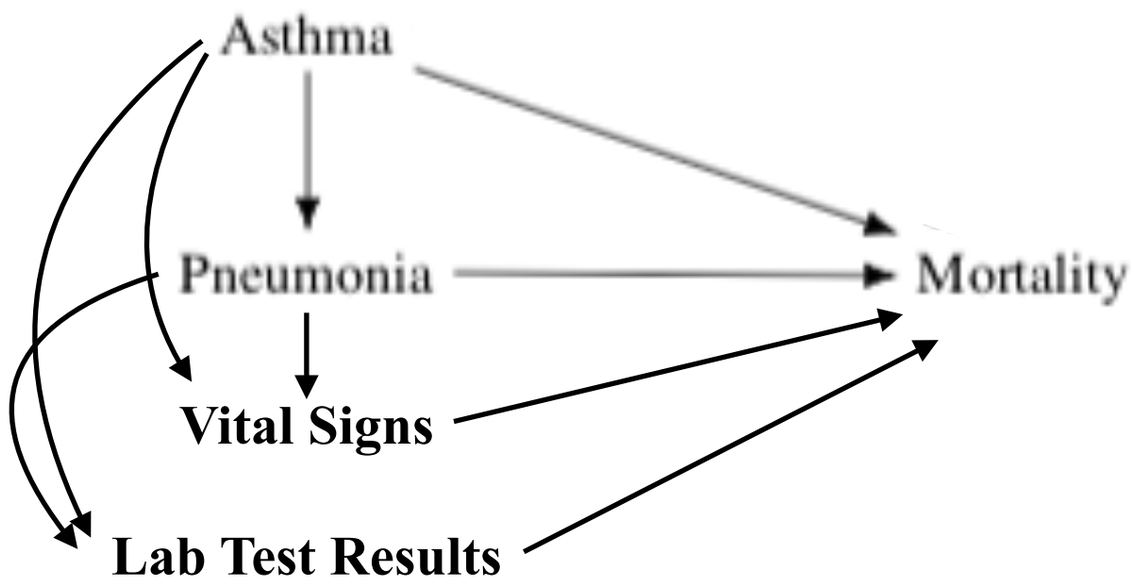
- Deploy system →sends police to same area regardless of true crime rates



Lum, Kristian, and William Isaac. "To predict and serve?." *Significance* 13.5 (2016): 14-19.
Ensign, Danielle, et al. "Runaway Feedback Loops in Predictive Policing." *Conference on Fairness, Accountability and Transparency*. 2018.

- Caruana et al. (2015) consider a model trained to predict mortality due to pneumonia (P) using data from hospitalized patients.

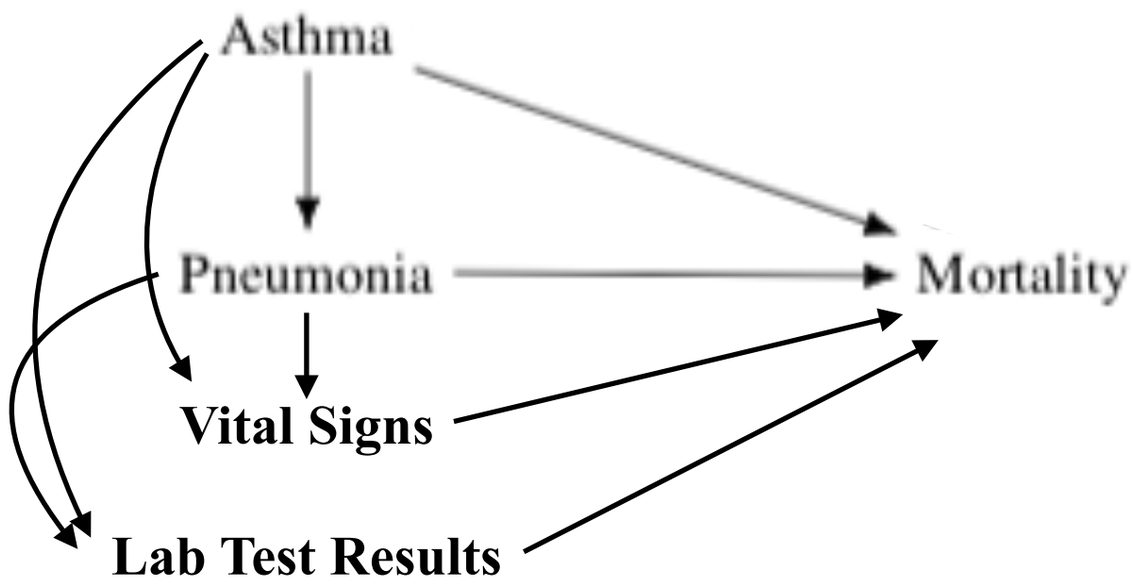
- Intended use of the model was for triage—to determine whether to admit patients or treat them at home.



- Caruana et al. (2015) consider a model trained to predict mortality due to pneumonia (P) using data from hospitalized patients.

- Intended use of the model was for triage—to determine whether to admit patients or treat them at home.
- The model learned that patients with pneumonia *AND* asthma (A) were less likely to die than those with only pneumonia.

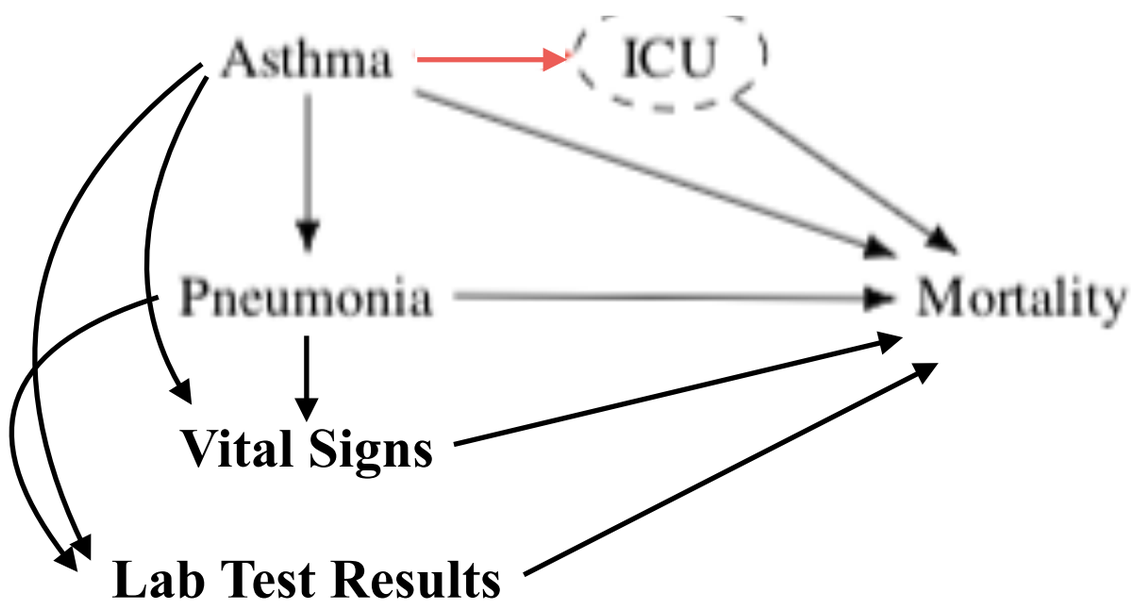
- Why?



- Caruana et al. (2015) consider a model trained to predict mortality due to pneumonia (P) using data from hospitalized patients.

- Intended use of the model was for triage—to determine whether to admit patients or treat them at home.
- The model learned that patients with pneumonia *AND* asthma (A) were less likely to die than those with only pneumonia.

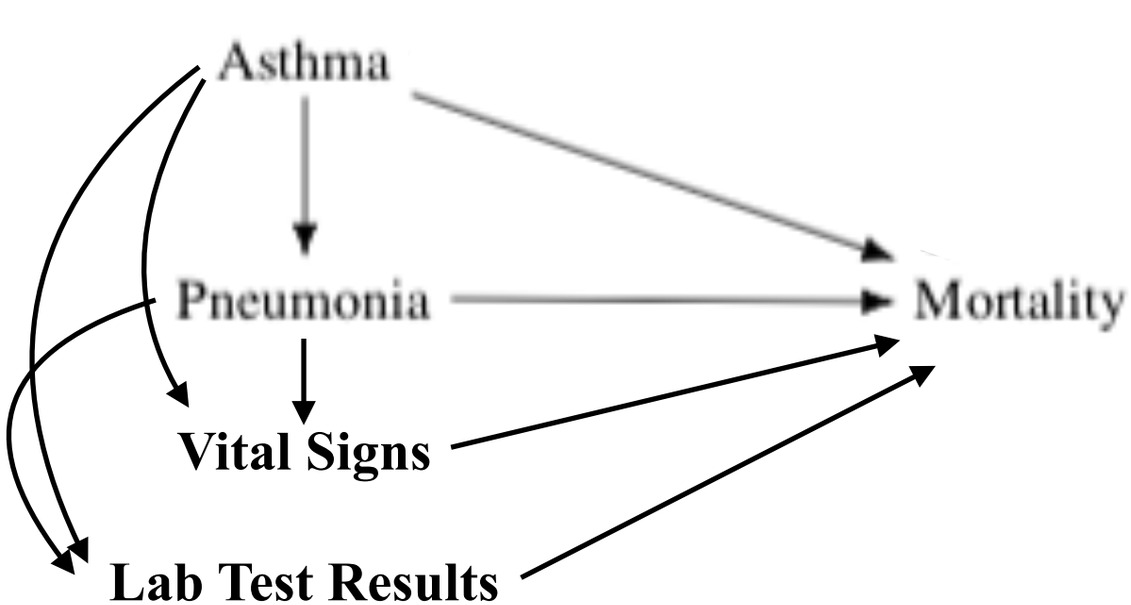
- Why? Patients were more likely to be directly admitted to the ICU (I) at the hospital.

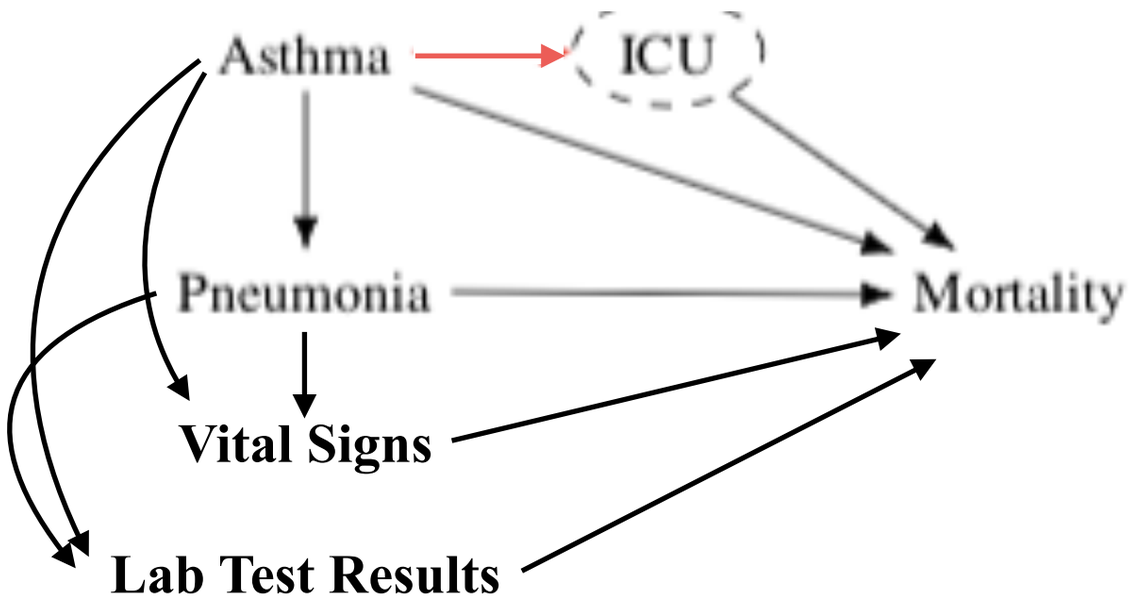


- Caruana et al. (2015) consider a model trained to predict mortality due to pneumonia (P) using data from hospitalized patients.

- Intended use of the model was for triage—to determine whether to admit patients or treat them at home.
- The model learned that patients with pneumonia *AND* asthma (A) were less likely to die than those with only pneumonia.

- Why? Patients were more likely to be directly admitted to the ICU (I) at the hospital.
- Asthma—> ICU dependency doesn't hold during test time; Learnt model lacks generalization.





- Caruana et al. (2015) consider a model trained to predict mortality due to pneumonia (P) using data from hospitalized patients.

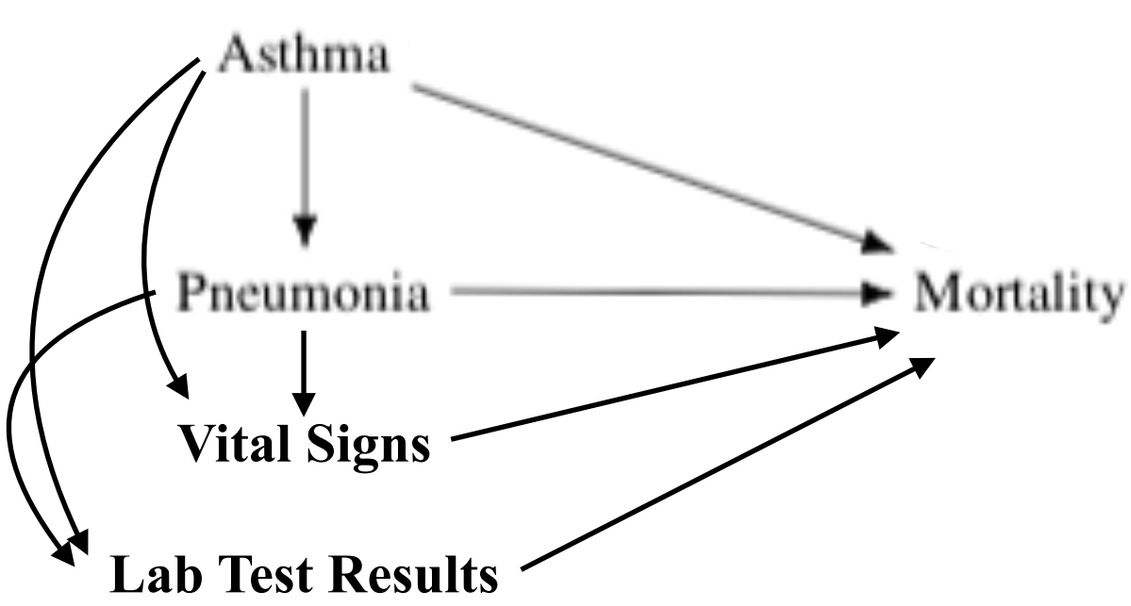
- Intended use of the model was for triage—to determine whether to admit patients or treat them at home.

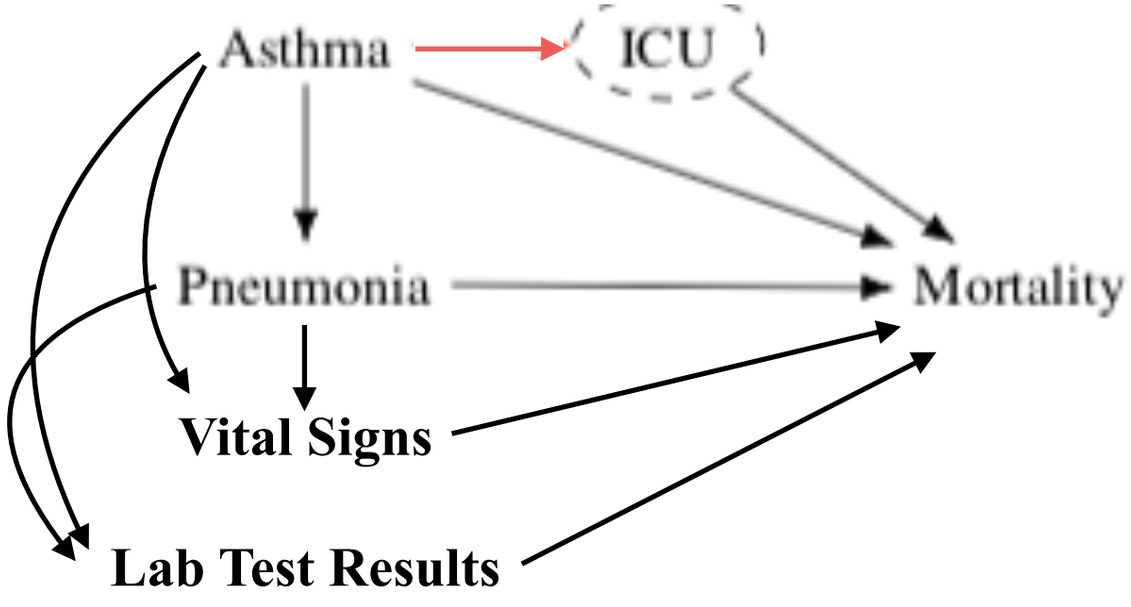
- The model learned that patients with pneumonia *AND* asthma (A) were less likely to die than those with only pneumonia.

- Why? P
  hosp

- Asthma
  generalization.

Data shift errors are subtle and occur easily as we go from a train to deploy. Can we learn models with guarantees to such failures?





Shifts in Data Hurt Generalization

- Why? Sampled data violates the iid assumption.
- Prior examples are different forms of shifts
- We need models that are *stable* (invariant) to shifts.

Transfer Learning is Reactive

- These examples are different forms of shifts
- We need models that are stable to shifts.
- Large literature is tackling data shifts.

Reactive

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

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).

Transfer Learning is Reactive

- These examples are different forms of shifts
- Can we learn models that are stable to shifts.
- Large literature is tackling data shifts.

Reactive

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

- But, all possible test environment data unrealistic to obtain training time.
- Want predictive model that generalizes to new, unseen environments

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).

Seek invariance proactively

- These examples are different forms of shifts
- Can we learn models that are stable to shifts.
- Large literature is tackling data 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:

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).

Subbaswamy, A. et al. “Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).

Schulam, Peter, and Suchi Saria. "**Reliable decision support using counterfactual models.**" *Advances in Neural Information Processing Systems*. 2017.

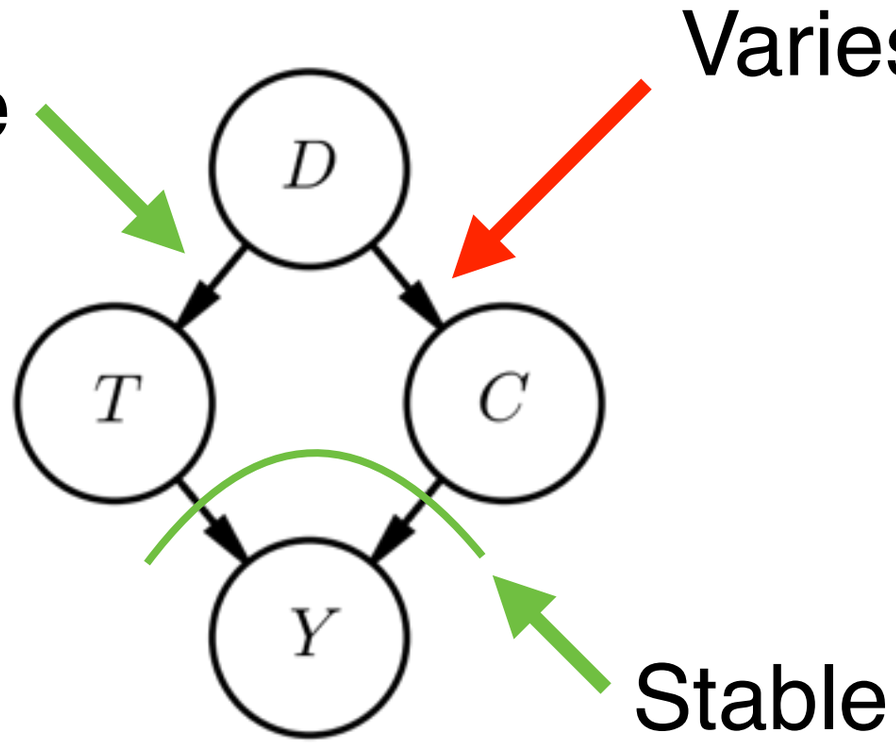
Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI),* (2018).

Key Questions

- We need a framework for representing different types of shifts
- We need a way to learn models that are guaranteed to be invariant to pre-specified shifts
- We don't want to limit ourselves to a simple class (e.g., linear) of models.

Simple Example: Unstable Paths and Unstable Models

- Consider naive discriminative model P(**T**|**C**,**Y**)
- Two active paths from **C** to **T** when conditioned on **Y**:
  - $C \leftarrow D \rightarrow T$
  - $C \rightarrow Y \leftarrow T$



Determine active paths using d-separation

- Is P(**T**|**C,Y)** stable?
  Is P(**T**|**D,C**,**Y**) stable?

Subbaswamy and Saria, UAI 2018

Key Takeaways

- Use a graphical formalism for expressing shifts in data.
    - Intuitive to reason with and provides a simple way to visualize different types of shifts
- Protecting from shifts requires trading-off accuracy
    - Learn everything —> includes spurious dependencies, no invariance
    - Learn conservatively —> ignore stable relationships that do generalize
    - Optimal: learn as much as possible without compromising on any desired invariances.
    - Different classes of methods can be analyzed under this lens of stability / accuracy tradeoff.

Subbaswamy, A, Chen, B, Saria, S. The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance. https://arxiv.org/abs/1905.11374, (2019).

Hierarchy of Stable Distributions

1. **Conditional Distributions**
   (e.g., graph pruning)

2. **Interventional Distributions**
   (e.g., surgery estimator)

3. **Counterfactual Distributions**
   (e.g., counterfactual normalization)

Increasing precision,
Increasing difficulty of
identifiability or estimation

Subbaswamy, A, Chen, B, Saria, S. **The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance**. https://arxiv.org/abs/1905.11374, (2019).

Deep Dive: Proactive Methods

Proactive

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

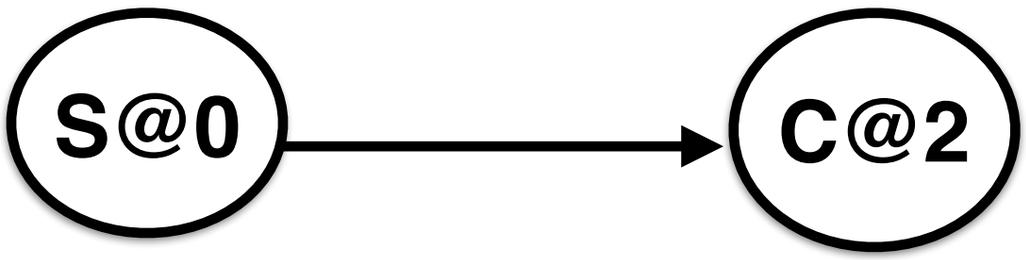
1. **Represent shifts using graphs; specs to identify which shifts to protect from**
2. Proactive Learning (graphs, specs) ==> preprocessing step that determines which parts of the distribution to fit
3. Use existing learning techniques to fit these components
4. Guarantees:
    1. Optimality
    2. 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. “Preventing Failures Due to Dataset Shift: 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).

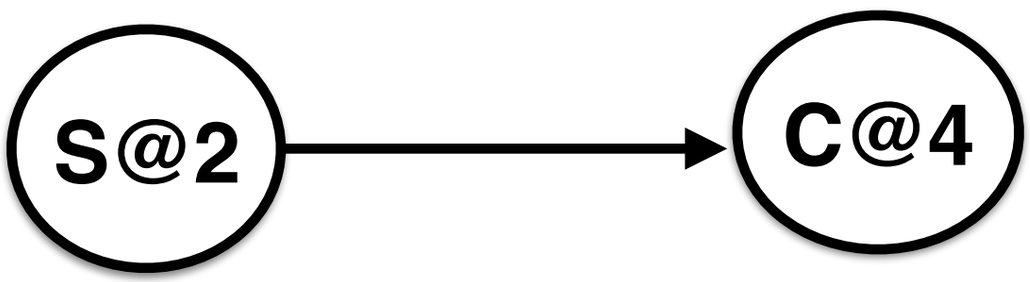
Revisiting Shifts using Graphs



Smoking habits now

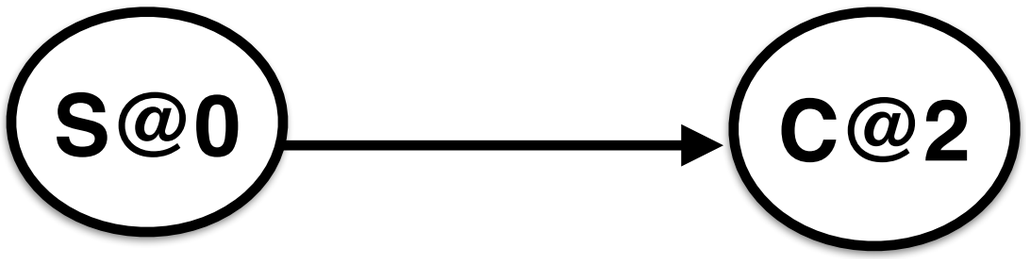
Freq. of Cancer in 5 years

Smoking ban .....



Storkey, Amos. "When training and test sets are different: characterizing learning transfer." *Dataset shift in machine learning* (2009): 3-28.
Schölkopf, Bernhard, et al. "On causal and anticausal learning." *Proceedings of the 29th International Conference on Machine Learning*, 2012.

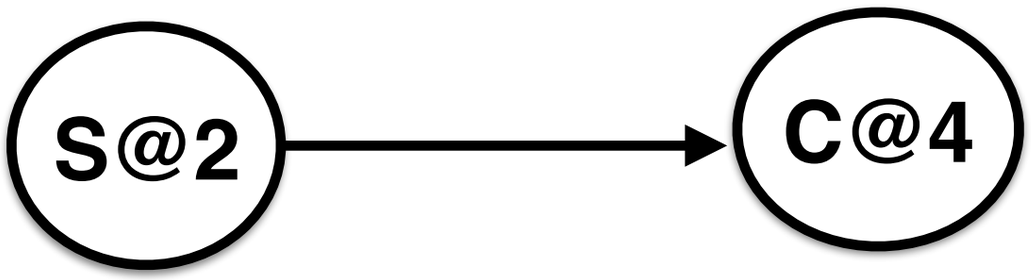
Revisiting Shifts using Graphs



Smoking habits now

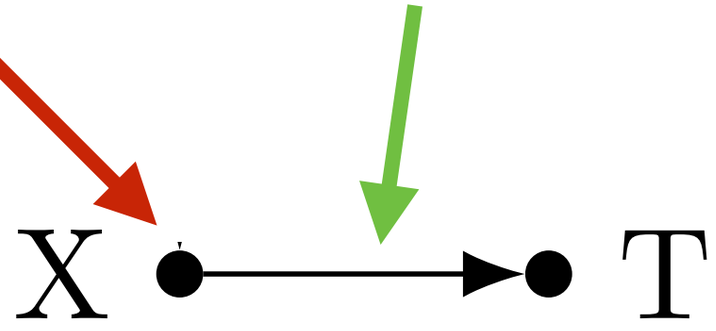
Freq. of Cancer in 5 years

Smoking ban .....



Unstable

Stable



P(**X**) changes

P(**T**|**X**) unchanged

Special case called *covariate shift*

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

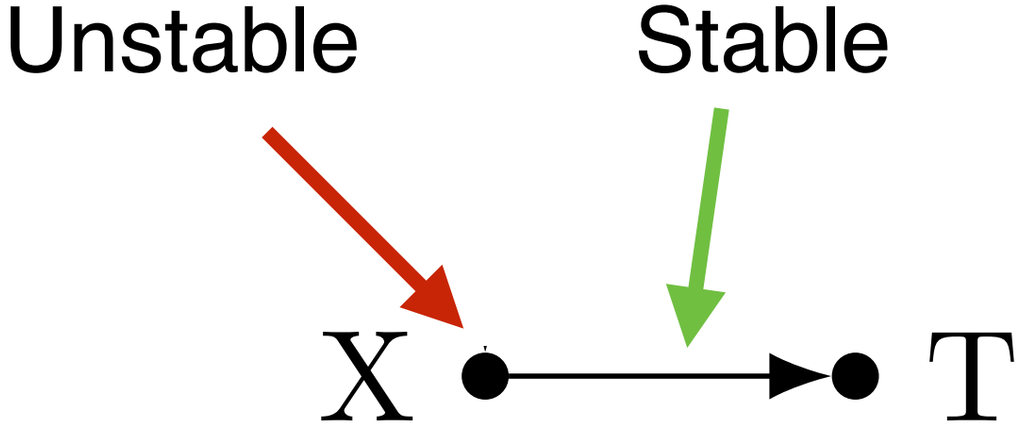
Schölkopf, Bernhard, et al. "On causal and anticausal learning." *Proceedings of the 29th International Conference on Machine Learning*, 2012.

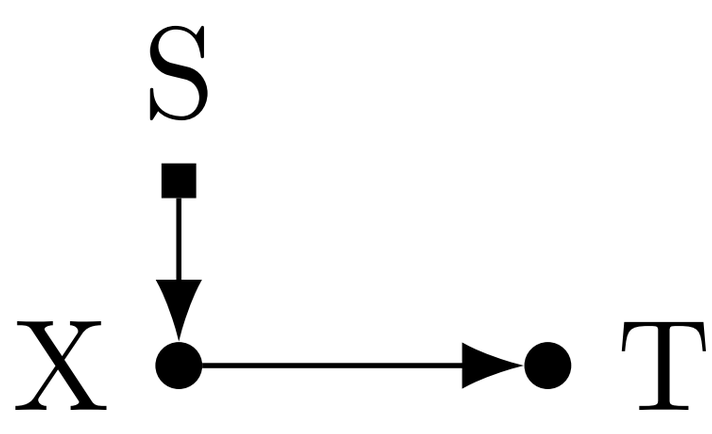
Shimodaira, Hidetoshi. "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of statistical planning and inference* 90.2 (2000): 227-244.

Gretton, Arthur, et al. "Covariate shift by kernel mean matching." (2009).

Covariate Shift

- Augment with selection variables to get selection diagram
- Selection vars point to variables who generation that can differ across environments

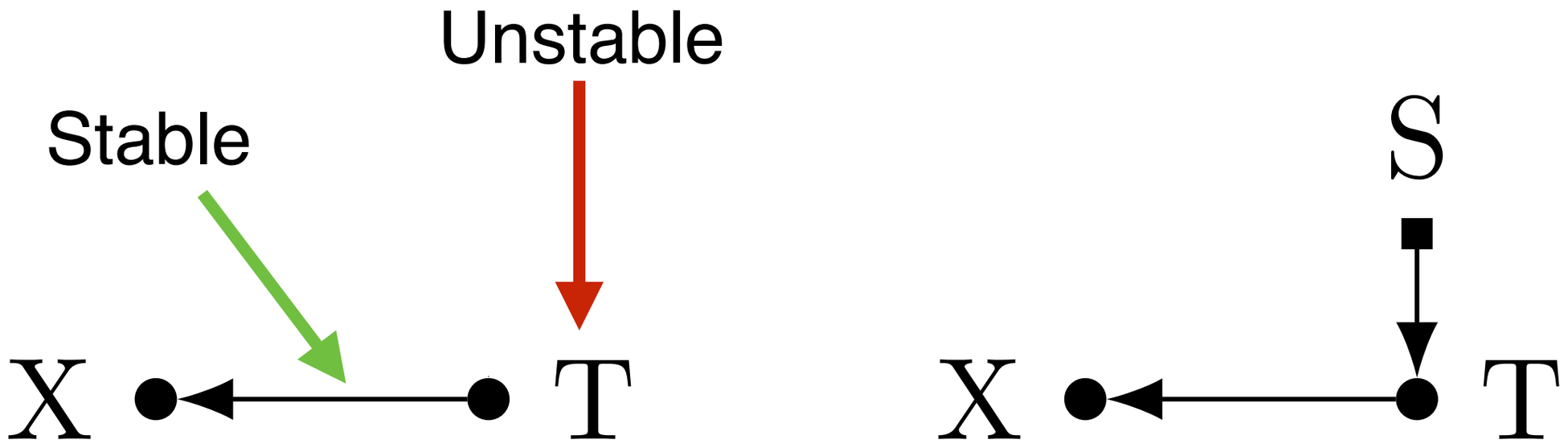




Pearl, Judea, and Elias Bareinboim. "Transportability of causal and statistical relations: a formal approach." AAAI, 2011.
Schölkopf, Bernhard, et al. "On causal and anticausal learning." *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Another special case: *Label Shift*

- Variables: input X, output T
- Distribution of target P(**T**) changes
- Conditional distribution of inputs is unchanged P(**X**|**T**)

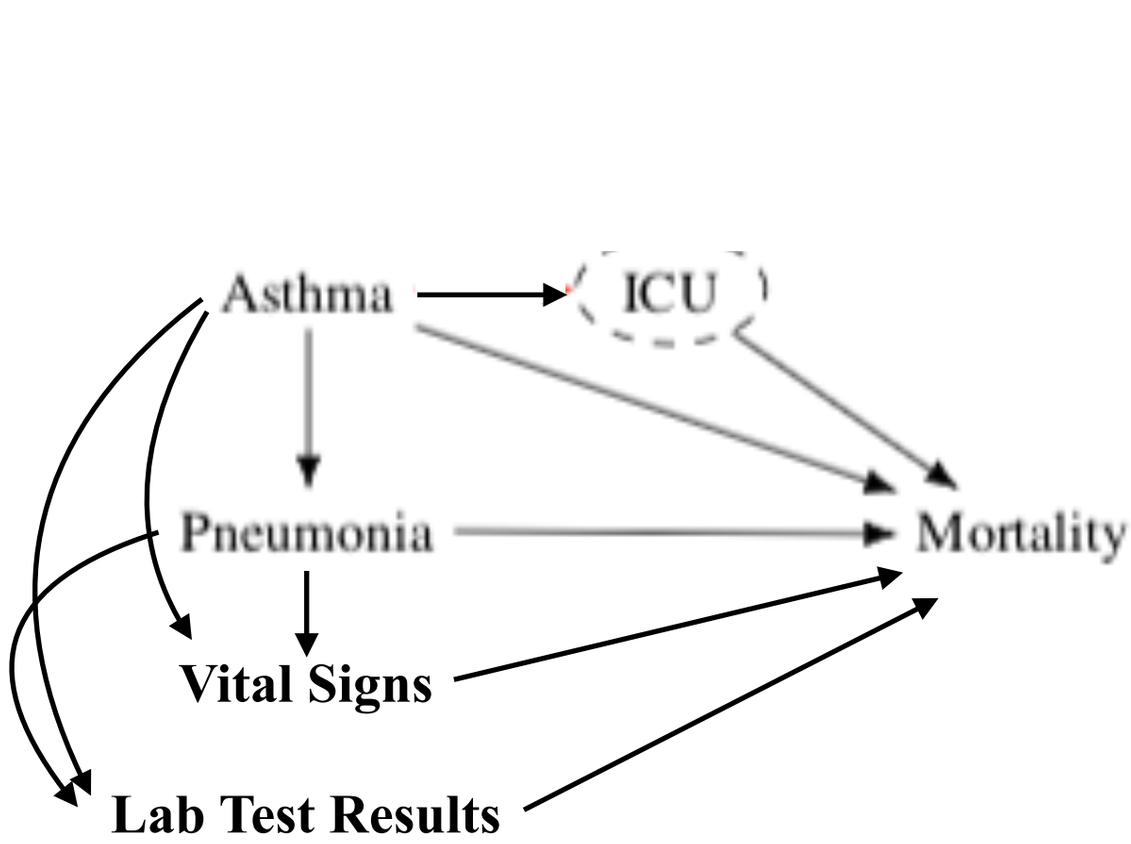


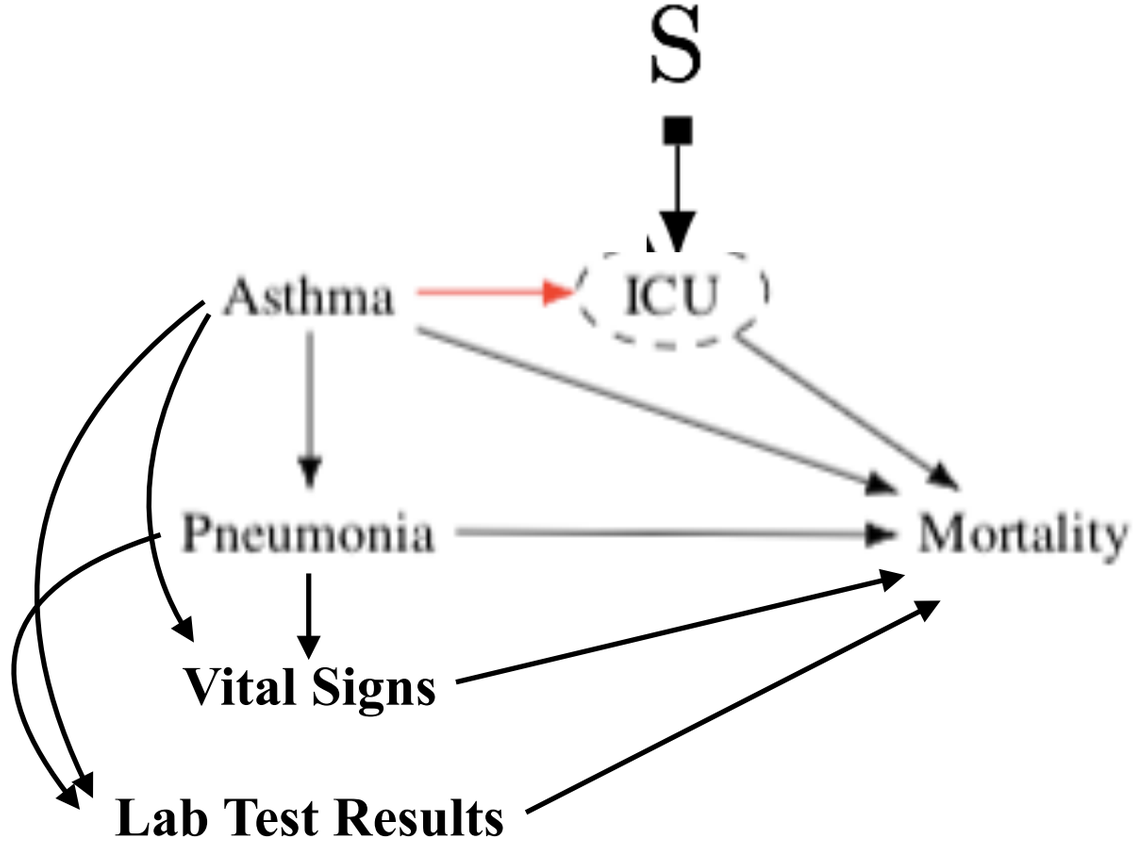
Chan, Yee Seng, and Hwee Tou Ng. "Word Sense Disambiguation with Distribution Estimation." *IJCAI*. Vol. 5. 2005.
Schölkopf, Bernhard, et al. "On causal and anticausal learning." *Proceedings of the 29th International Conference on Machine Learning*, 2012.
Zhang, Kun, et al. "Domain adaptation under target and conditional shift." *International Conference on Machine Learning*, 2013.
Lipton, Zachary C., et al. "Detecting and Correcting for Label Shift with Black Box Predictors." *International Conference on Machine Learning*, 2018.

Complex Shifts: Beyond Label and Covariate Shift

- Here, the policy determining when to admit to ICU will change from development to deployment environment.

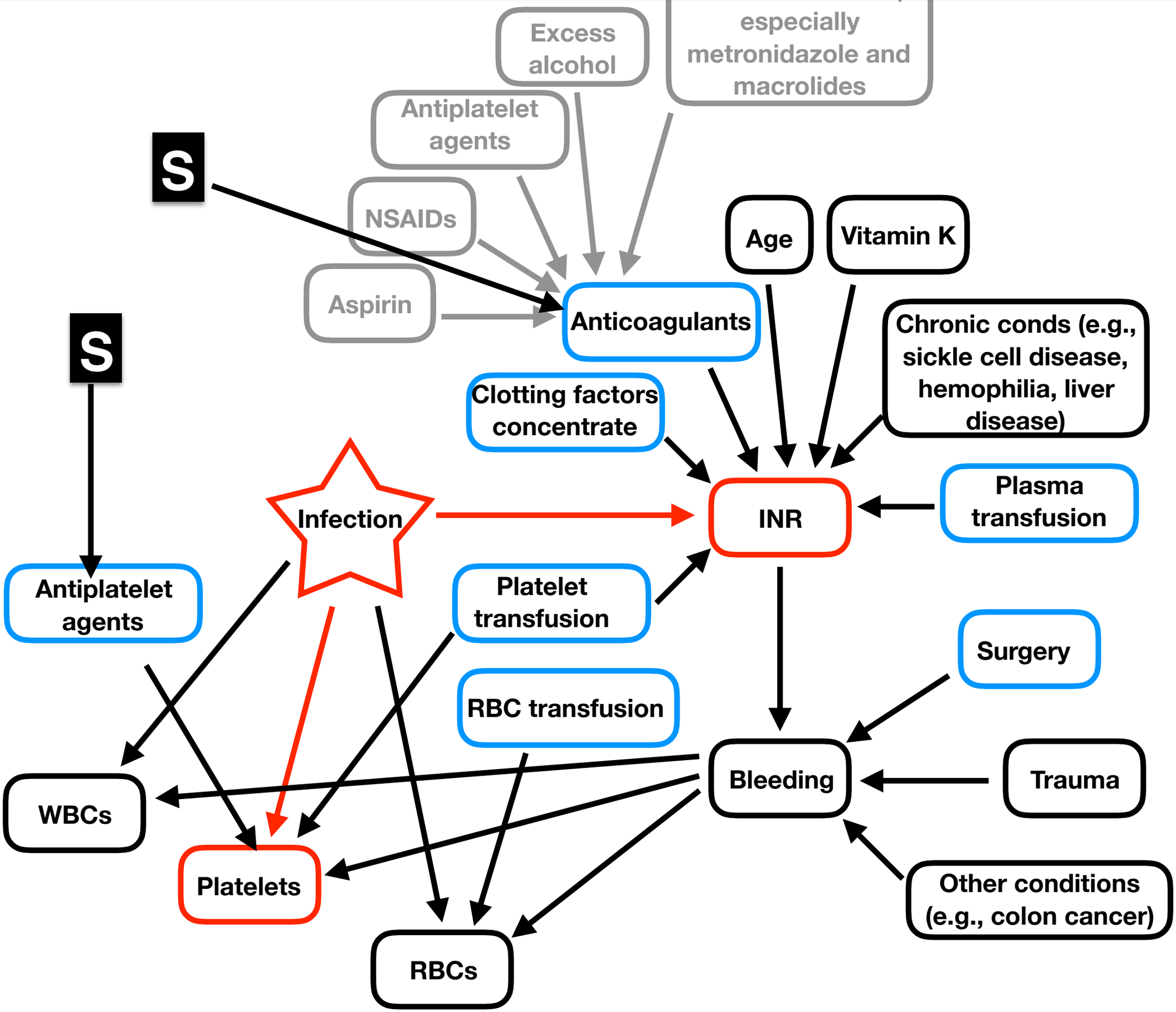
- This is an instance of *policy shift.*



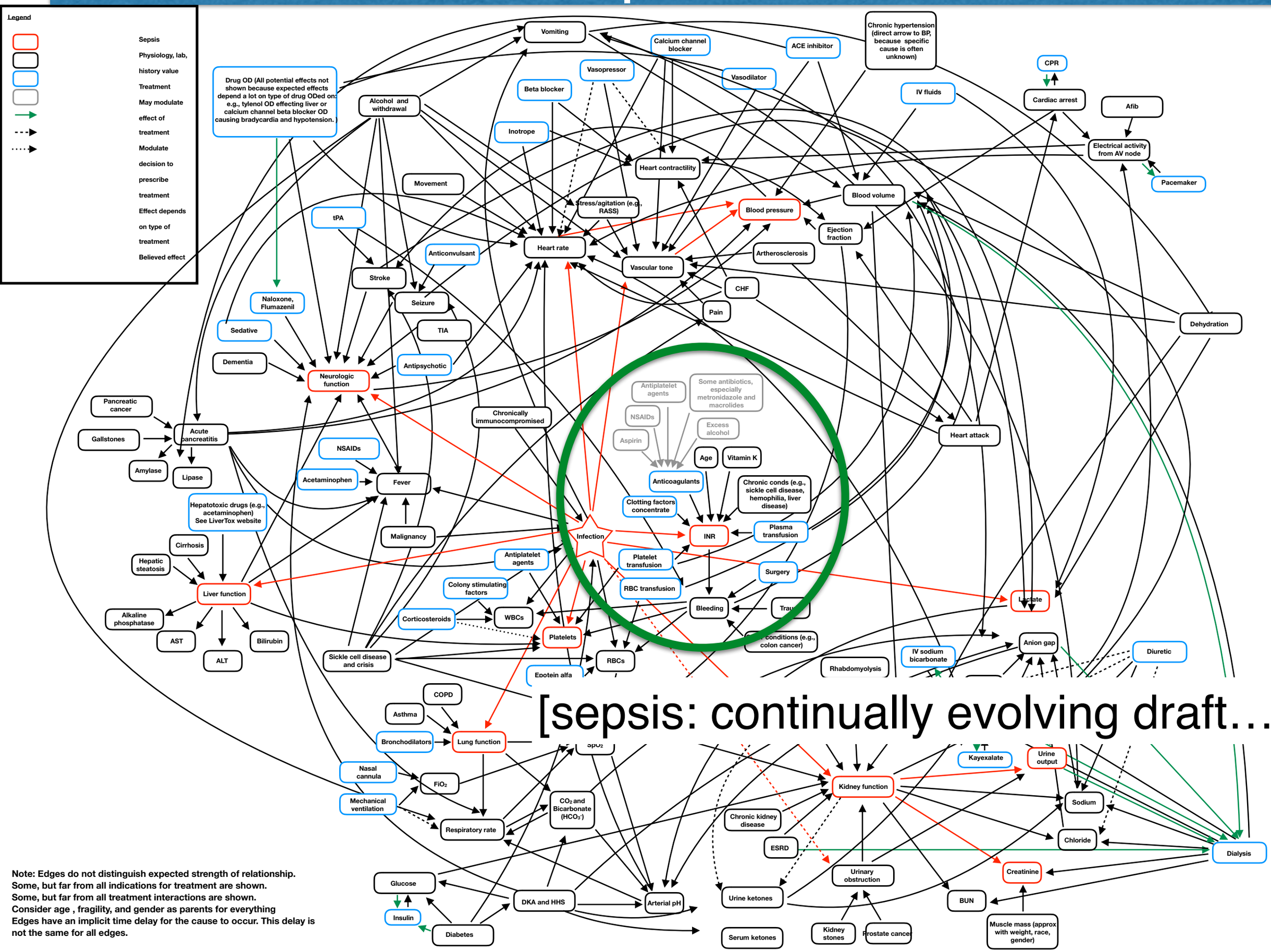


Schulam, Peter, and Suchi Saria. "**Reliable decision support using counterfactual models.**" *Advances in Neural Information Processing Systems*. 2017.

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.

Deep Dive: Proactive Methods

Proactive

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

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

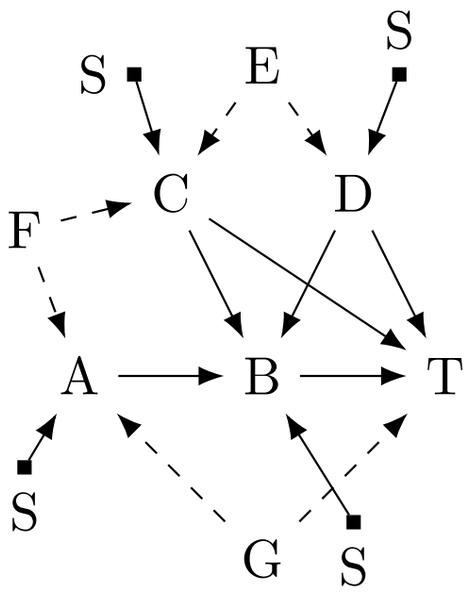
Shimodaira, Hidetoshi. "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of statistical planning and inference* 90.2 (2000): 227-244.

Gretton, Arthur, et al. "Covariate shift by kernel mean matching." (2009).

Algorithm: Surgery Estimator

Input: Graph w/ invariance specs (selection vars)

Output: Data conditionals to fit, how to combine



Algorithm: Surgery Estimator

output

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

Target: T

Observed vars: A, B, C, D

+

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

Subbaswamy, Adarsh et al. “Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).

Deep Dive: Proactive Methods

Proactive

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

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

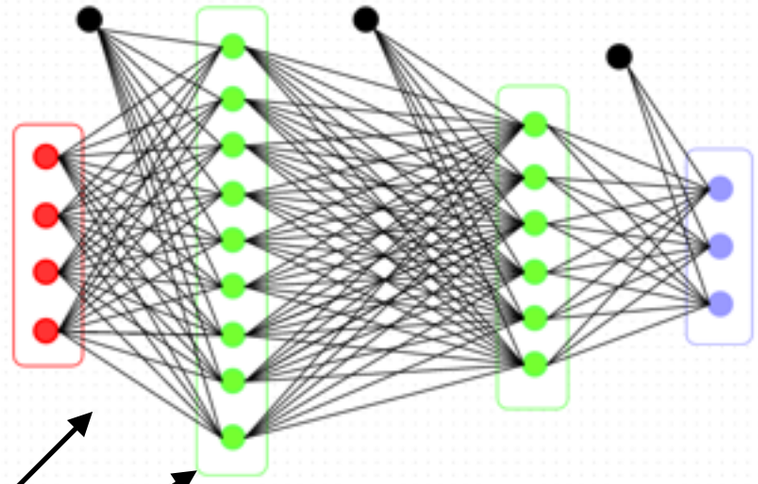
Shimodaira, Hidetoshi. "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of statistical planning and inference* 90.2 (2000): 227-244.

Gretton, Arthur, et al. "Covariate shift by kernel mean matching." (2009).

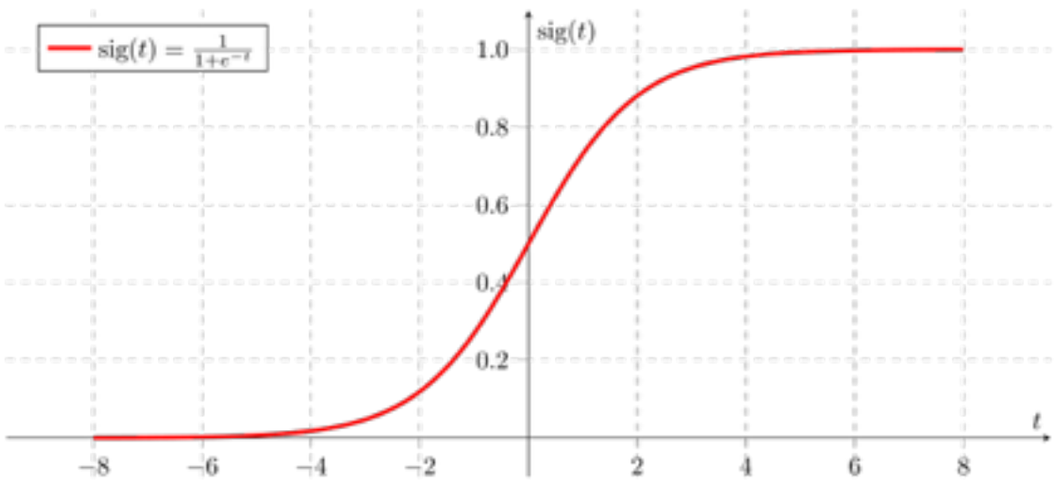
Fit conditionals with existing methods

Algorithm: Surgery Estimator

serves as pre-processing step



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



Deep Dive: Proactive Methods

Proactive

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

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

Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI),* (2018).
Shimodaira, Hidetoshi. "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of statistical planning and inference* 90.2 (2000): 227-244.
Gretton, Arthur, et al. "Covariate shift by kernel mean matching." (2009).

Guarantees: Surgery Estimator Algorithm

- Procedure is **sound**: returned distribution is invariant to anticipated shifts

- Procedure is **complete**: if it fails then no estimable invariant distribution exists in the class of conditional or interventional distributions

- Guide data collection: collect data on hidden variables or induce randomization to remove hidden confounding.

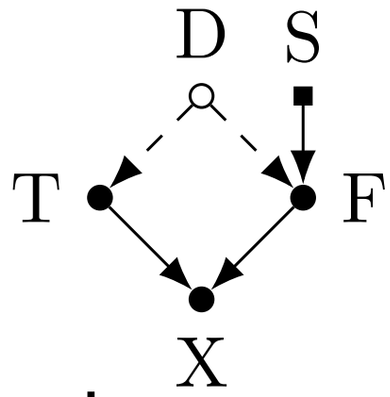
- Maybe able to capture more dependencies (to improve performance) but will require additional assumption.

Subbaswamy, Adarsh et al. “Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).

Guarantees: Distributional Robustness

- Distributional Robustness: Optimize for lowest worst-case (minimax) loss across a set of distributions

$$\inf_{B \in \Gamma} \sup_{Q_{\mathbf{s}} \in \Gamma} E_{Q_{\mathbf{s}}}[\ell(h_B^*, \mathbf{O})]$$



- Selection diagram defines a set of possible environments
  Ex: All departments that only differ in style preferences

- **Optimality**: Surgery estimator is distributionally robust across environments defined by selection diagram

Optimal stable estimator

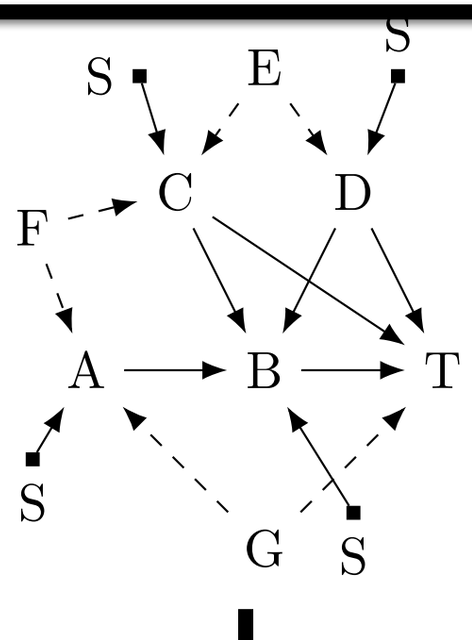
Subbaswamy, Adarsh et al. “Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).

Sinha, Aman, Hongseok Namkoong, and John Duchi. "Certifying some distributional robustness with principled adversarial training." *In International Conference on Learning Representations (ICLR)*. (2018).

Rothenhäusler, Dominik, et al. "Anchor regression: heterogeneous data meets causality." *arXiv preprint arXiv:1801.06229* (2018).

Dataset Shift vs Adversarial Robustness

(1) Check for reliability w.r.t. shifts



Target: T
Observed vars: A, B, C, D

Algorithm: Surgery Estimator

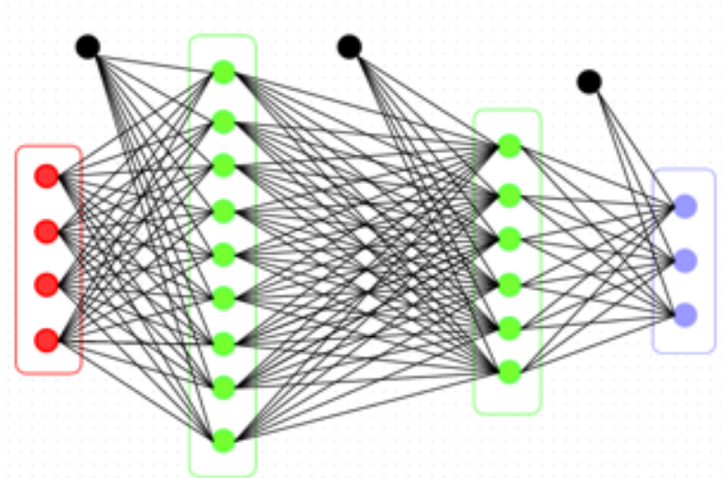
output

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

$$+$$

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

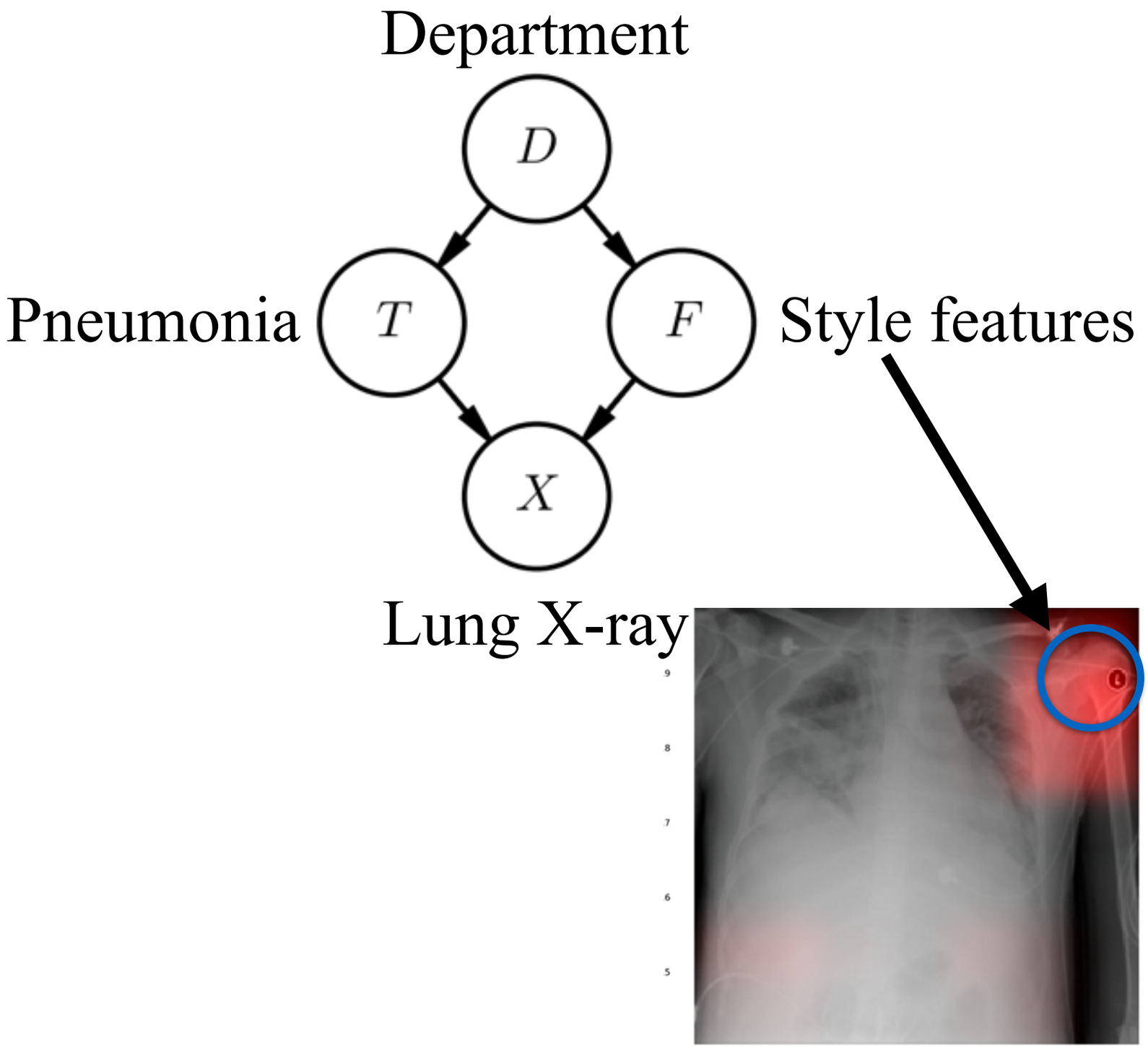
(2) Safety w.r.t. adversarial examples



Saria, Subbaswamy, **Tutorial: Safe and Reliable Machine Learning**.

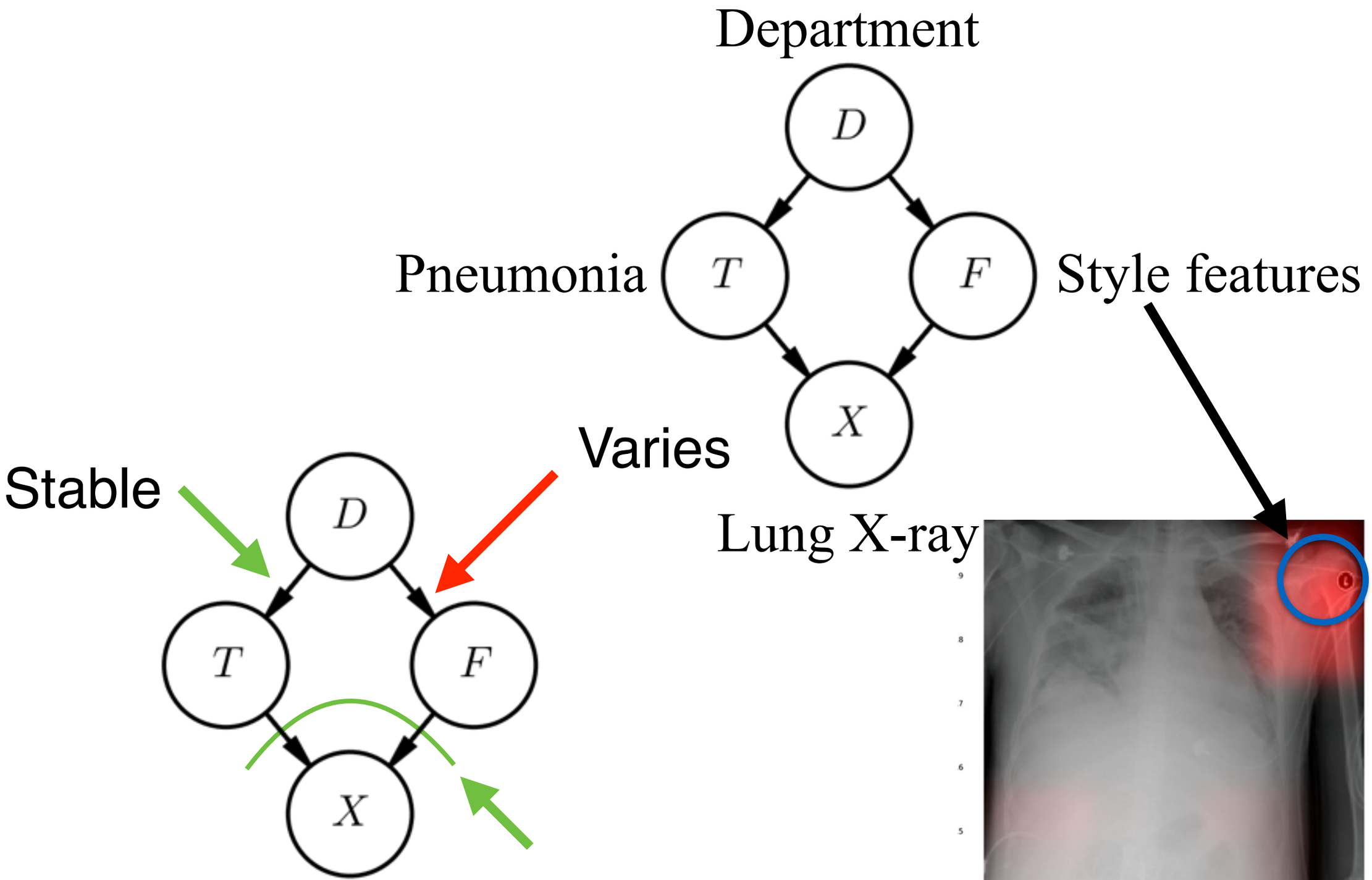
Returning to our pneumonia example

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



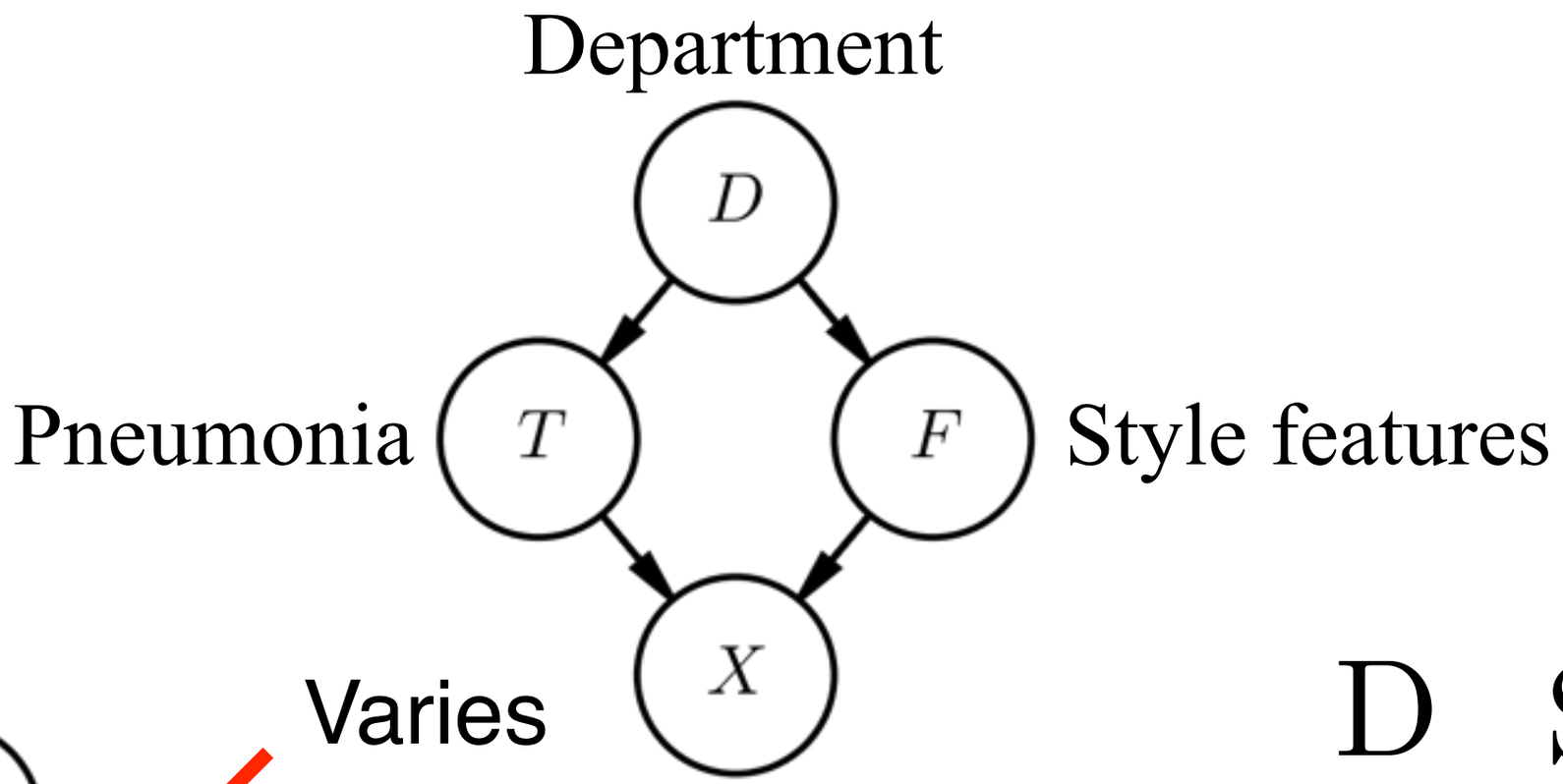
Returning to our pneumonia example

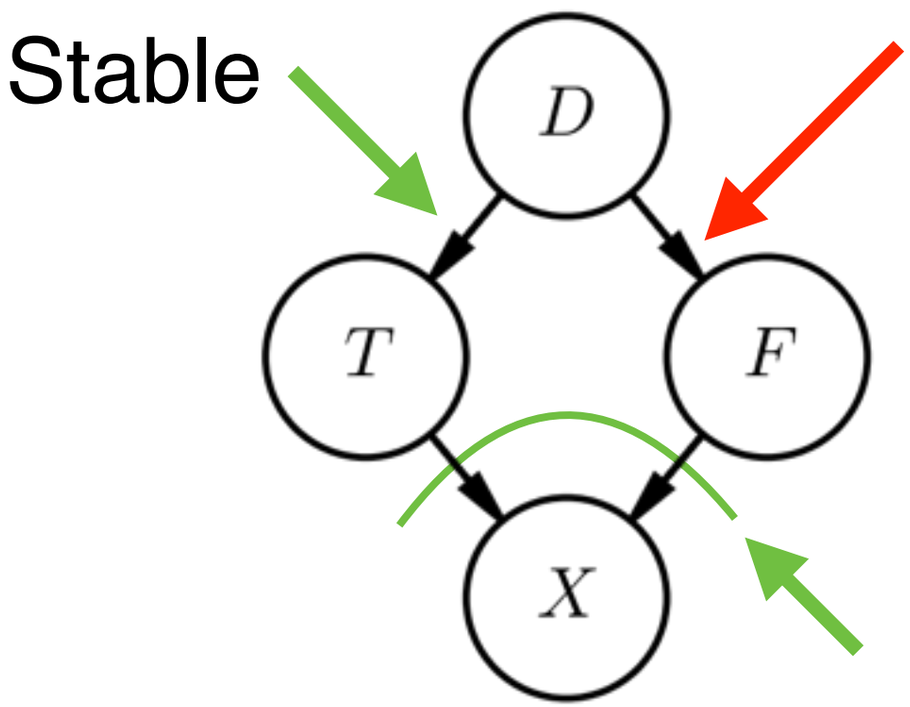
- Goal: Diagnose **T** from **F** and **X**

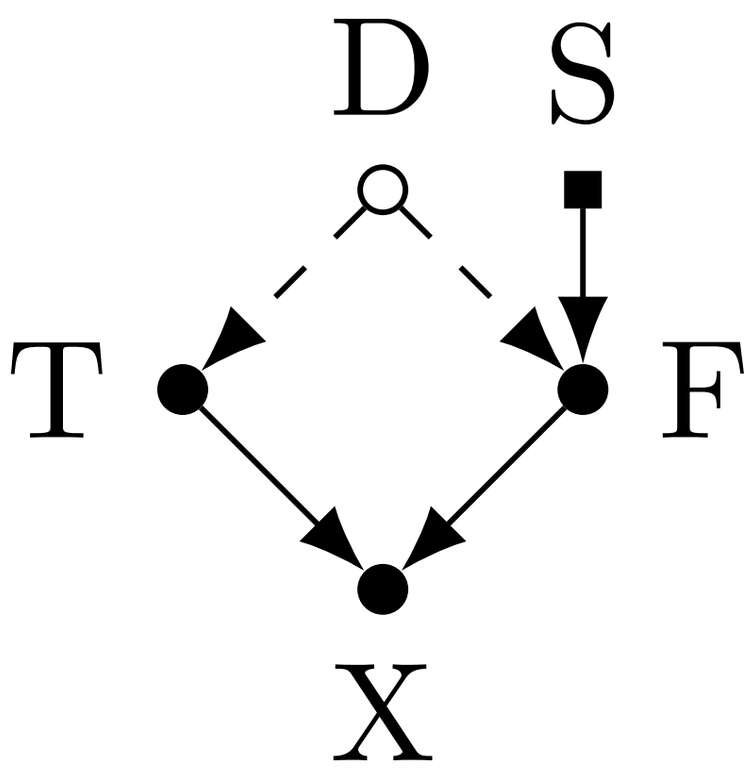


Returning to our pneumonia example

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

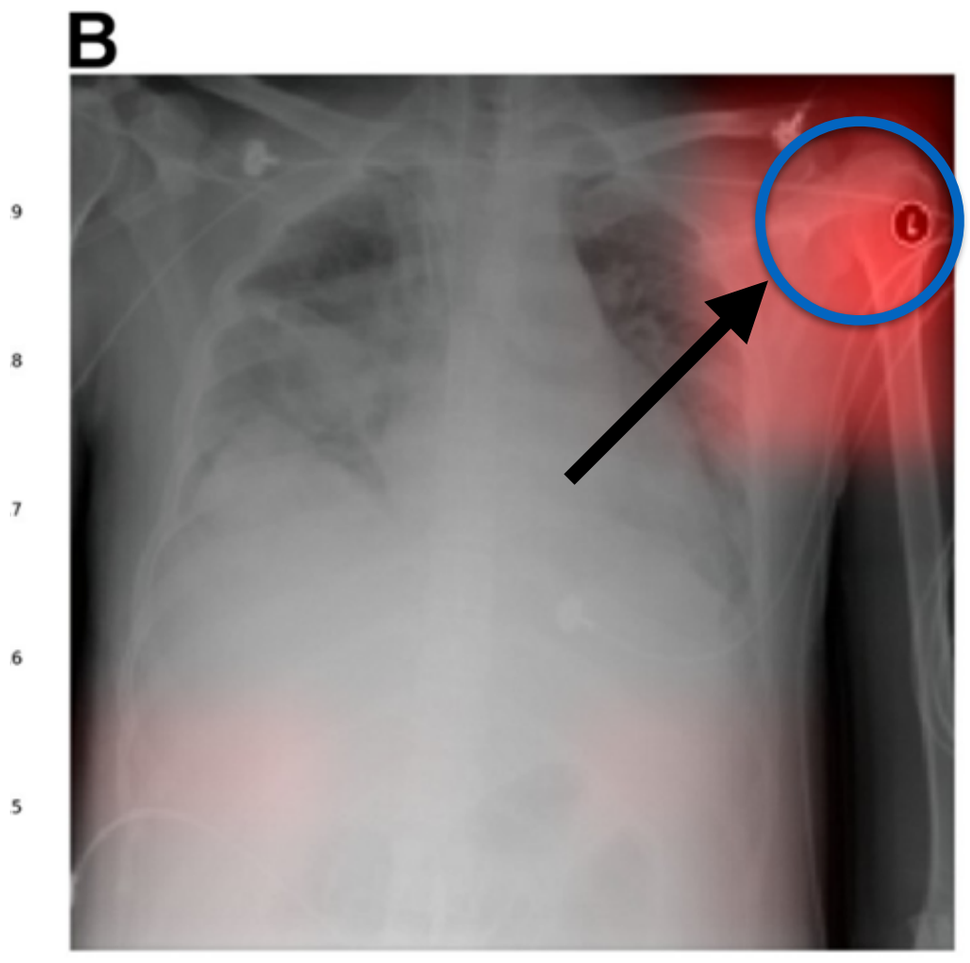


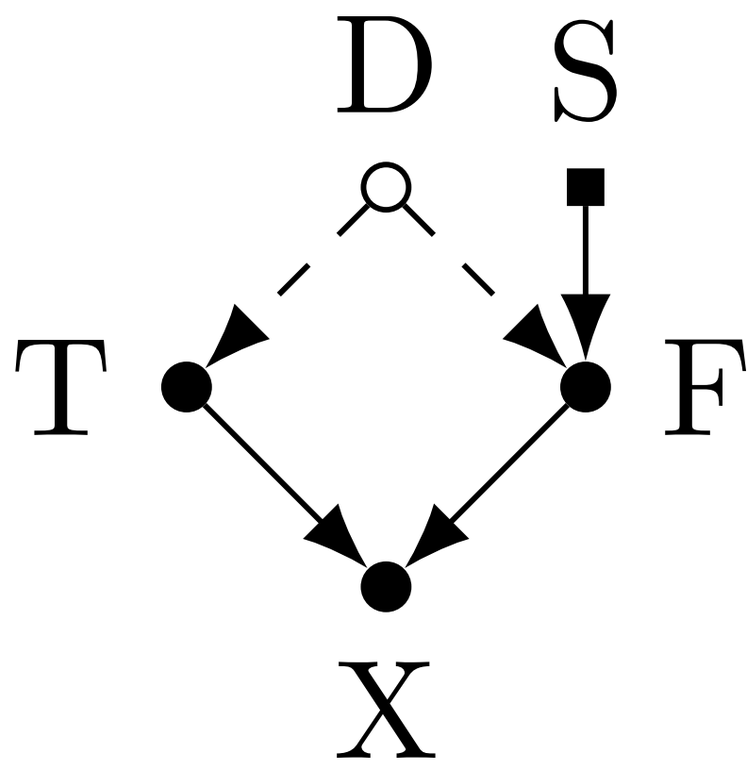




Returning to our pneumonia example

What does the surgery estimator look like?



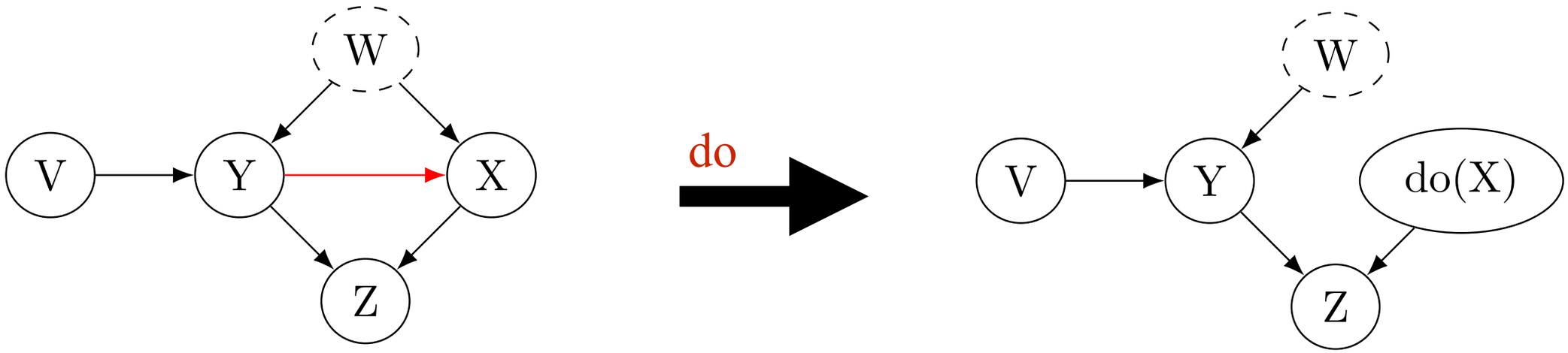


1. Estimate F for each image
2. Fit P(X | T, F) and P(T)
3. Integrate to obtain P(T | X, do(F))

Hierarchy of Stable Distributions

1. **Conditional Distributions**
2. **Interventional Distributions**

- Hypothetically intervene on variables with shifted mechanism
- Intervening deletes all edges into variable

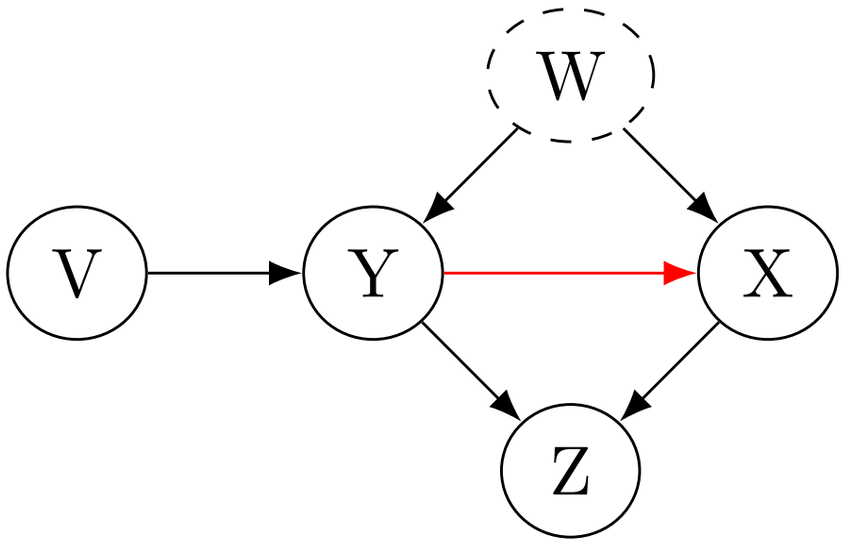


P(Y I V, Z, do(X)) is stable
P(Y I V, Z, X) is unstable
P(Y I V) is stable

Subbaswamy, A. et al. “Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).
Subbaswamy, A, Chen, B, Saria, S. The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance. https://arxiv.org/abs/1905.11374, (2019).

Hierarchy of Stable Distributions

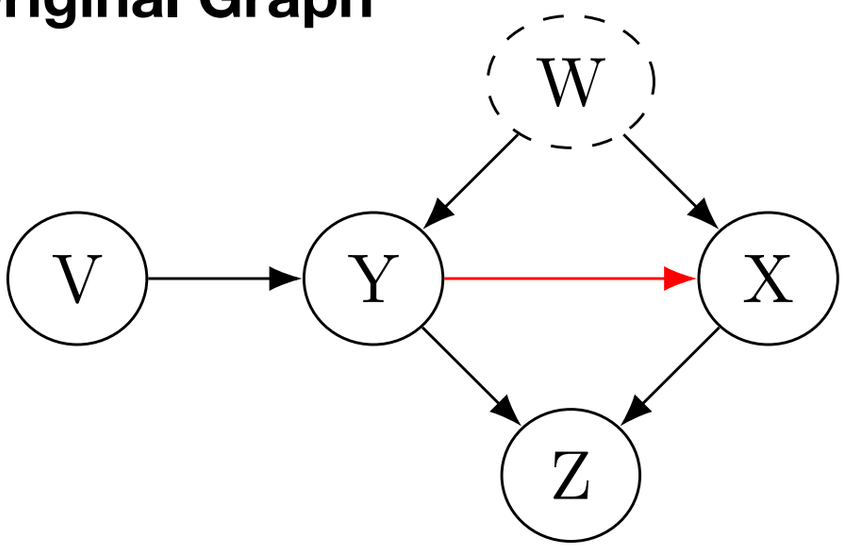
1. **Conditional Distributions**
2. **Interventional Distributions**
3. **Counterfactual Distributions**



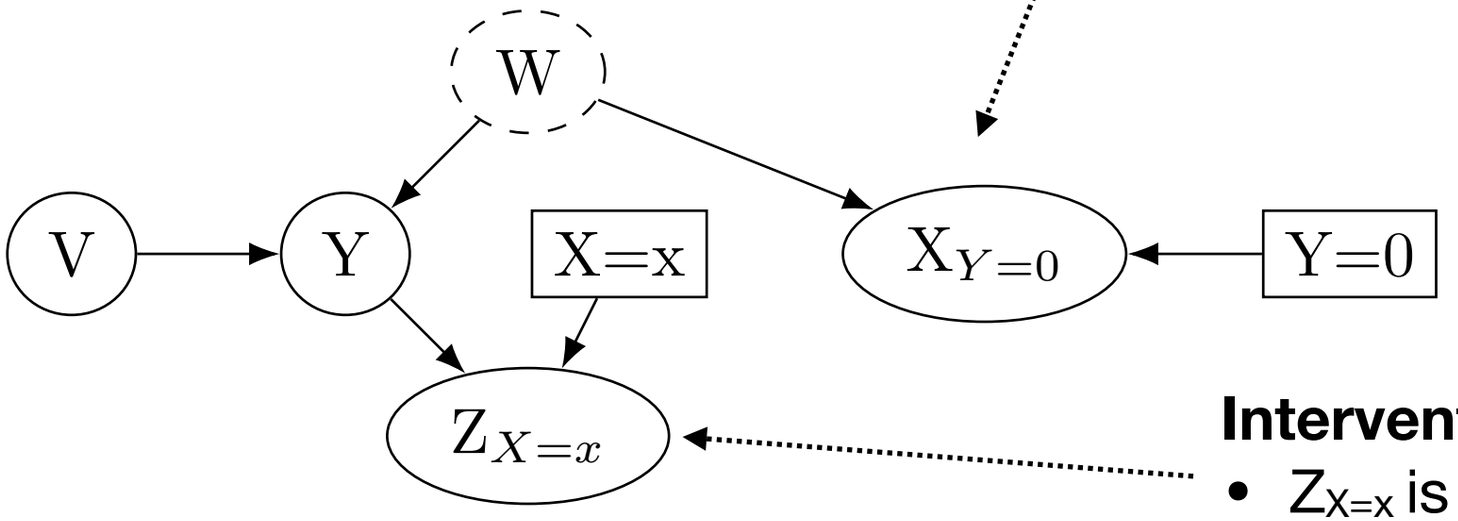
Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI)*, (2018).
Subbaswamy, A, Chen, B, Saria, S. The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance. https://arxiv.org/abs/1905.11374, (2019).

Counterfactuals

Original Graph



Counterfactual Graph



Counterfactual:

- Given we observed X to be x under Y=y, $X_{Y=0}$ is value of X had Y been set to 0.
- $X_{Y=0}$ is a counterfactual feature that is in general not equal to observed feature X.

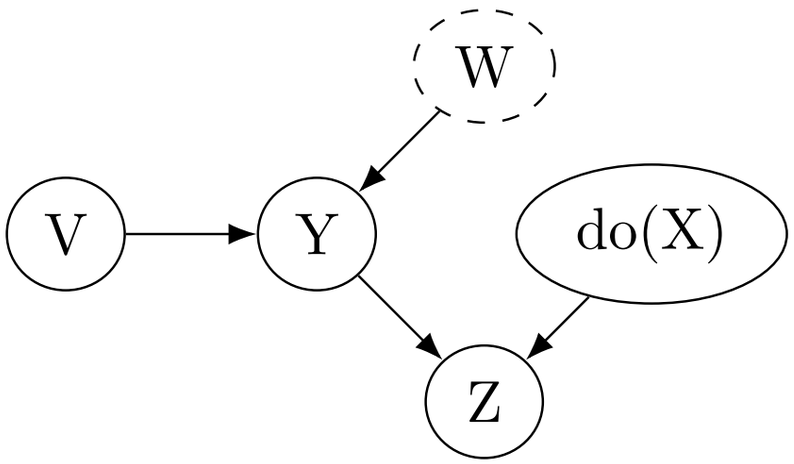
Intervention:

- $Z_{X=x}$ is Z had we set X to its observed value x. Corresponds to an intervention on X.

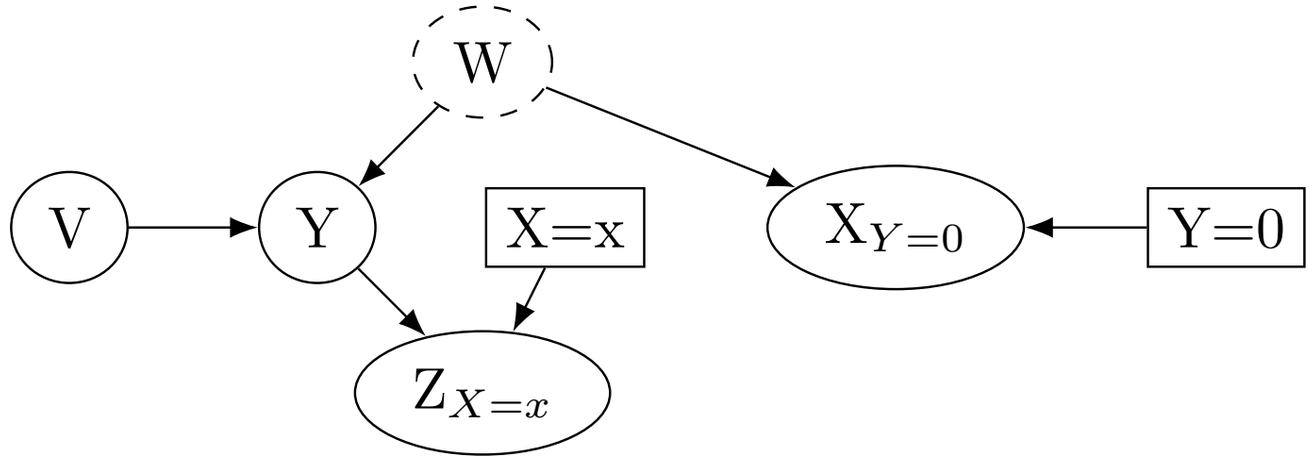
Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI),* (2018).

Hierarchy of Stable Distributions

1. **Conditional Distributions**
2. **Interventional Distributions**
3. **Counterfactual Distributions**



P(Y I V, Z, do(X)) is stable
Deletes W → X and Y → X



P(Y I V, $Z_{X=x}$, $X_{Y=0}$) is stable
Retains the stable edge W → X while deleting Y → X

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, Chen, B, Saria, S. The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance. https://arxiv.org/abs/1905.11374, (2019).

Hierarchy of Stable Distributions

1. **Conditional Distributions**
   (e.g., graph pruning)

2. **Interventional Distributions**
   (e.g., surgery estimator)

3. **Counterfactual Distributions**
   (e.g., counterfactual normalization)

Increasing precision,
Increasing difficulty of
identifiability or estimation

Subbaswamy, A, Chen, B, Saria, S. **The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance**. https://arxiv.org/abs/1905.11374, (2019).

Ensuring Reliablity via Test Time Monitoring

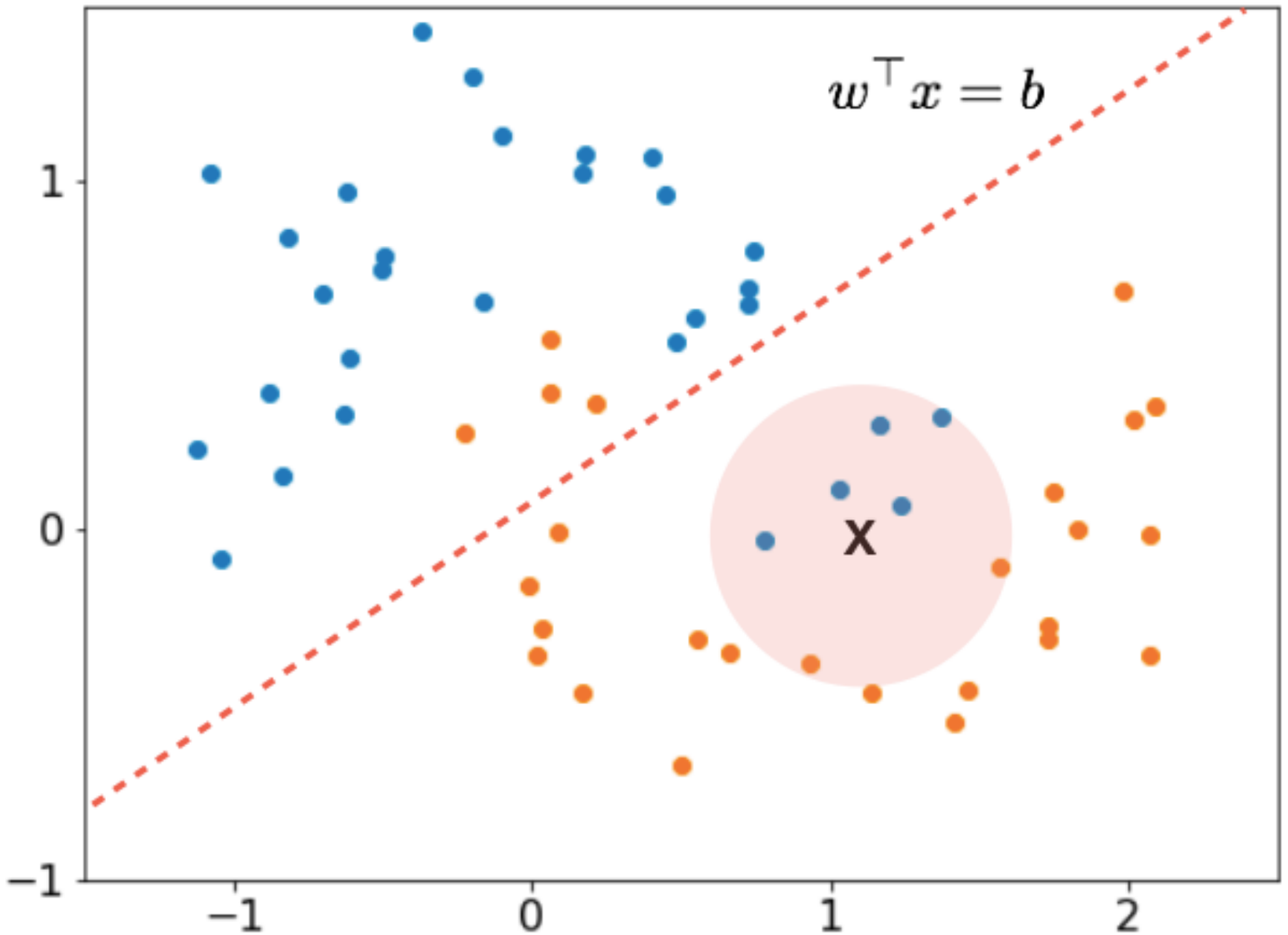
Algorithmic or Human-driven

- **General form**: arise from inadequate data, model poorly fits the data in a region

- **General approach:** Another algorithm that detects when the model output is unreliable and reject (we refer to these as assessing *point-wise reliability*).

- Special cases:
  - *Anomaly detection* **(**e.g., high score given by a one-class classifier**)**
  - *Open category detection* (when new OOD test points belong to a new concept class)

General Density Principle

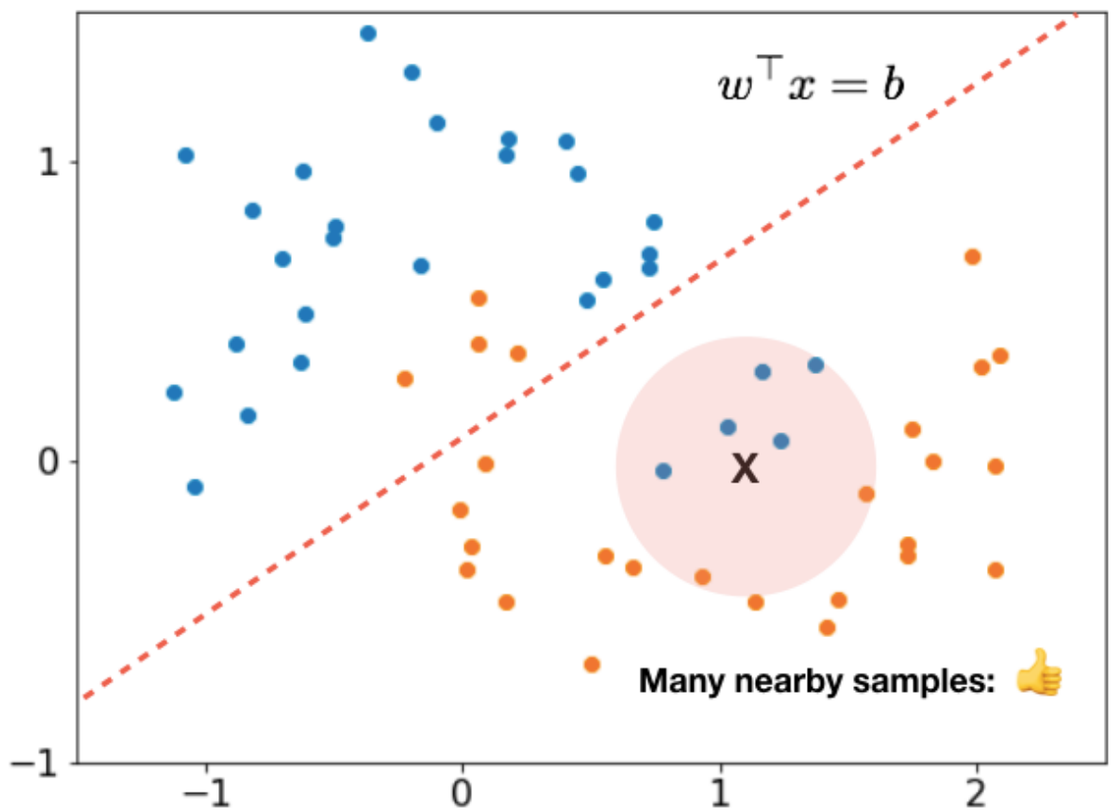


Use training samples near a test input (“X”) to help determine how reliable the prediction is

- Density principle to assess test point prediction reliability date backs to 1992:
  J.A. Leonard, M.A. Kramer, and L.H. Ungar. A neural network architecture that computes its own reliability. Computers & chemical engineering, 1992.

Resampling Uncertainty Estimation

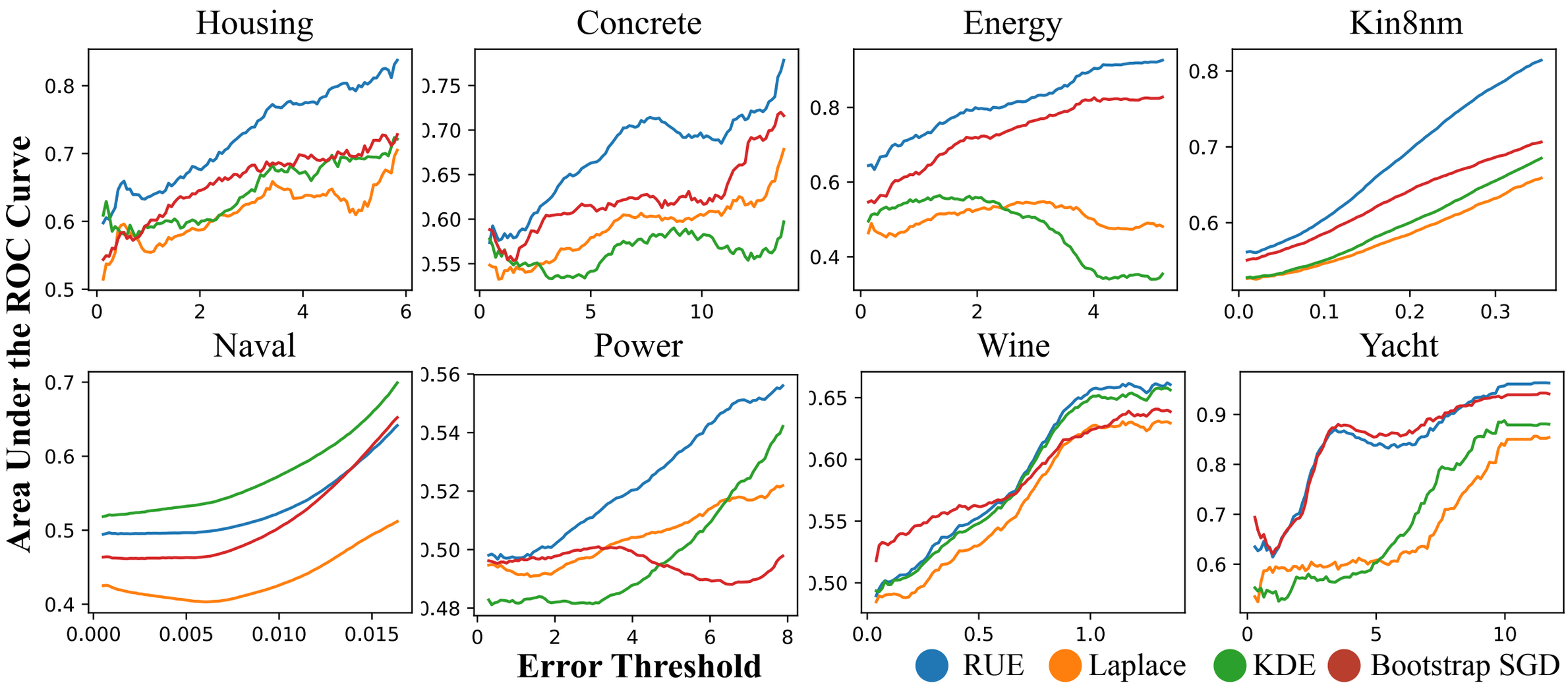
- Can show that it audits a model's predictions by checking
  - **Density**: Is the test case close to training samples?
  - **Local fit**: Did the model correctly fit nearby samples?
- Score derived from an approximation of the Bootstrap



P. Schulam, S. Saria. **Can You Trust This Prediction? Auditing Pointwise Reliability After Learning** . Artificial Intelligence and Statistics (AISTATS), 2019.

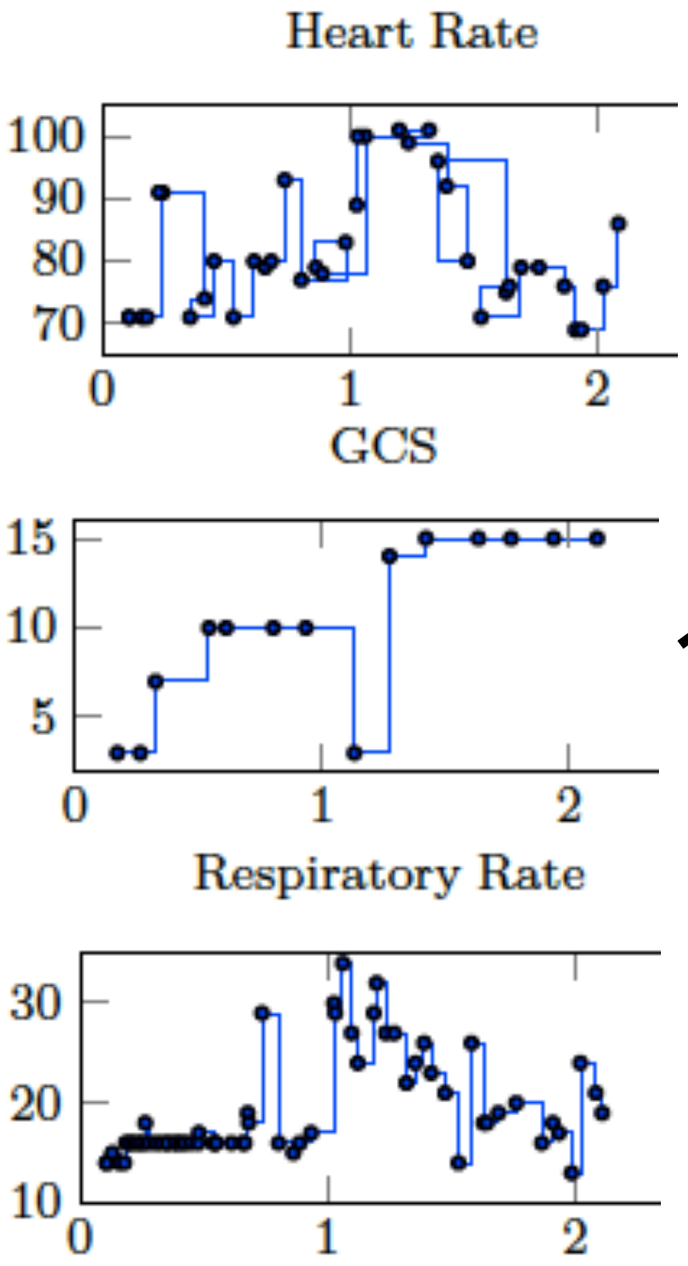
Jiang H, Kim B, Guan M, Gupta M. **To trust or not to trust a classifier**. In Advances in Neural Information Processing Systems 2018

Regression: Detect Inaccurate Predictions



P. Schulam, S. Saria. **Can You Trust This Prediction? Auditing Pointwise Reliability After Learning** . Artificial Intelligence and Statistics (AISTATS), 2019.

Ex: Healthcare (Adverse Event Detection)



Optimize alert reliability

Alert

Wait & watch

Improve alert reliability:

- At **Test-time**: Determine whether to alert or wait and collect more data to refine estimates
- Optimize sequential policy to determine alert vs. wait
- Reduces false alerts dramatically!

Soleimani H, Hensman J, Saria S. **Scalable joint models for reliable uncertainty-aware event prediction**. IEEE transactions on pattern analysis and machine intelligence. 2018 Aug 1;40(8):1948-63.

Cortes C, DeSalvo G, Mohri M. **Learning with rejection**. InInternational Conference on Algorithmic Learning Theory 2016 Oct 19 (pp. 67-82). Springer, Cham.

Need for mixed methods in ensuring reliability

- Basic theory, principles and **mixed methods for risk assessment and mitigation.**
- A suitable domain-specific **risk conceptualization** for the understanding, assessment and management of risk.
    - **(collective) mindfulness** as interpreted in the studies of High Reliability Organisations (HROs), capturing the following characteristics:
    - preoccupation with failure, **reluctance to simplify**, sensitivity to operations (e.g., how it's deployed), commitment to resilience, and **deference to expertise** (e.g., domain knowledge).
- Concepts and ideas highlighting the importance of **continuous improvement**.

Where do we go from here?

1. Need for mixed methods in understanding reliability — describing “good” and “bad” behaviors and checking if the system is behaving as expected.

2. Three key areas inspired from reliability engineering in safety critical domains

    1. **Failure Prevention:** Methods for specifying failure modes and learning methods that *proactively* protect from failures (e.g, graphically representing complex invariances and learning models that satisfy them)

    2. **Reliability Monitoring:** Monitor for failure sources at time of use; Making it easy for the user to debug systems for identifying failures

    3. **Maintenance:** Process for fixing failures when they occur; Identify when the system is stale and needs to be retrained

Thank you!

ssaria@cs.jhu.edu
asubbaswamy@jhu.edu

@suchisaria
@_asubbaswamy

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

Subbaswamy, A, Chen, B, Saria, S. **The Hierarchy of Stable Distributions and Operators to Trade Off Stability and Performance**. https://arxiv.org/abs/1905.11374, (2019).

Guarantees: Surgery Estimator Algorithm

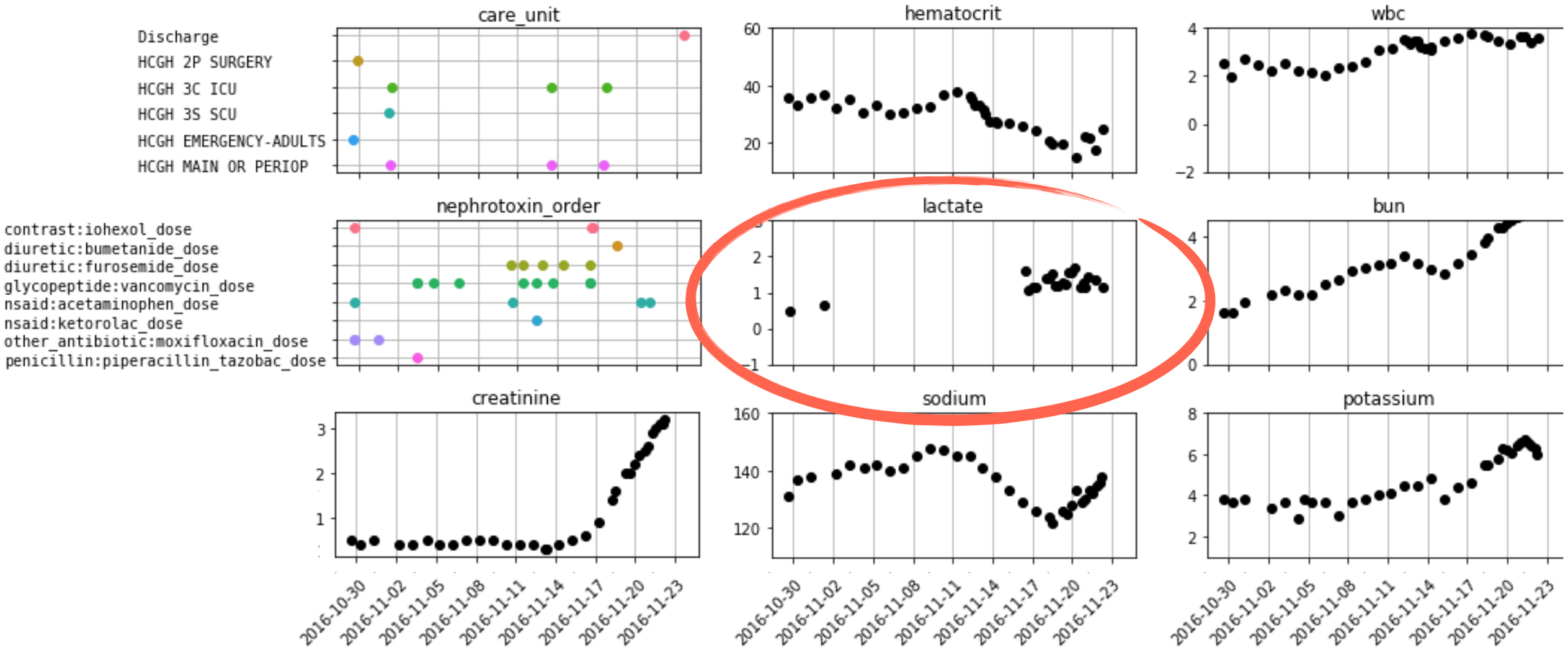
- Procedure is **sound**: returned distribution is invariant to anticipated shifts

- Procedure is **complete**: if it fails then no estimable invariant distribution exists

Practical issues & implications:

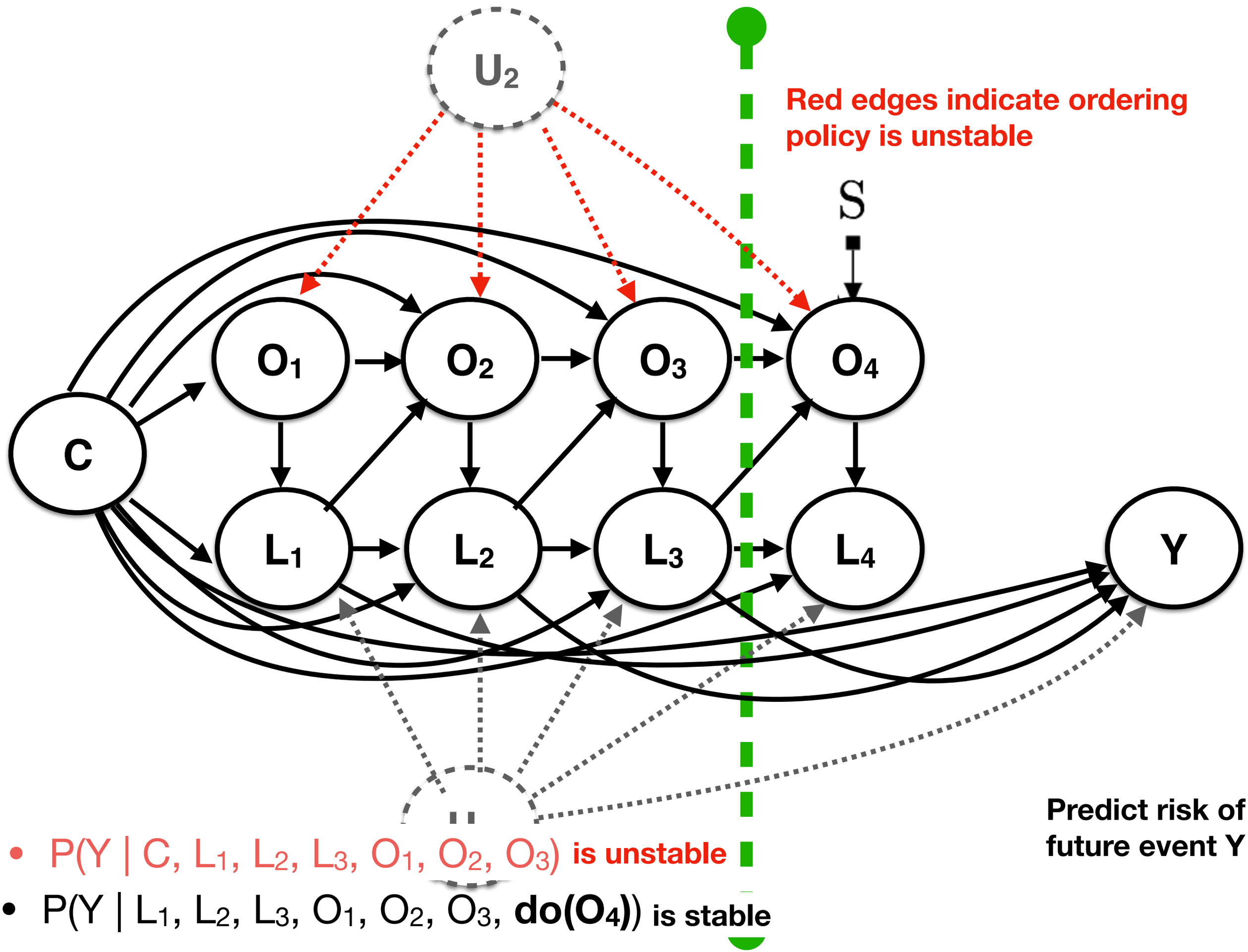
- What if you're uncertain about certain dependencies? Can do sensitivity analysis against multiple graphs and measure tradeoff.
- Is there a tradeoff between stability and performance? Yes. Prioritize invariance to "riskiest shifts".
- Graph representation allows intuitive understanding of loss in performance.
- Requires understanding domain!

Subbaswamy, Adarsh et al. "Preventing Failures Due to Dataset Shift: Learning Predictive Models That Transport." *International Conference on Artificial Intelligence and Statistics*. (2019).
Subbaswamy, A, and Saria, S. "Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms." *Uncertainty in Artificial Intelligence (UAI)*, (2018).



- 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.



- $P(Y \mid C, L_1, L_2, L_3, O_1, O_2, O_3)$ **is unstable**
- $P(Y \mid L_1, L_2, L_3, O_1, O_2, O_3, \mathbf{do(O_4)})$ **is stable**

Today's focus will be on the general case
For details see: Schulam and Saria 2017 (NeurIPS)

Reliable Decision Support using Counterfactual Models

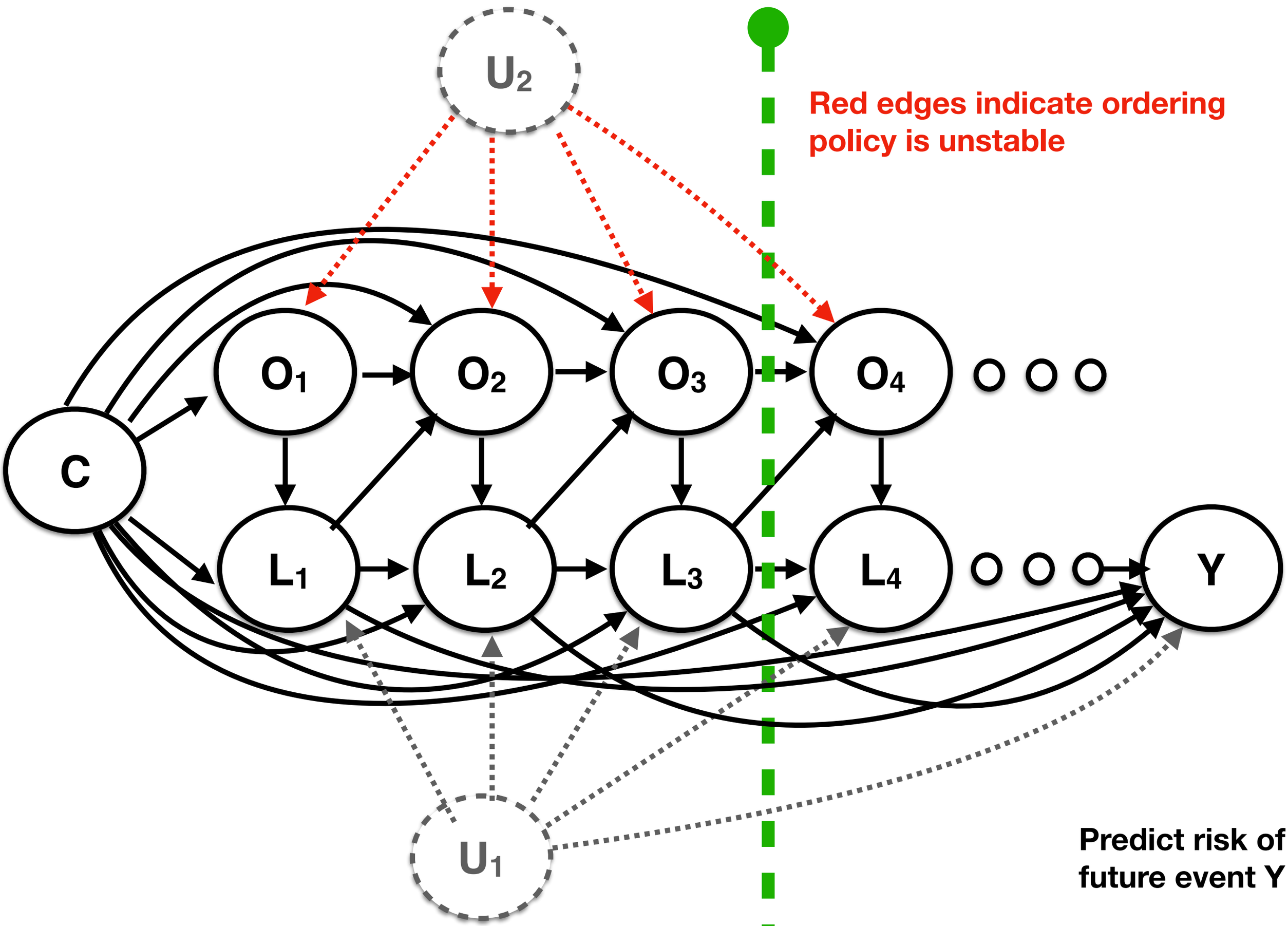
Counterfactual Gaussian Process

Peter Schulam
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
pschulam@cs.jhu.edu

Suchi Saria
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21211
ssaria@cs.jhu.edu

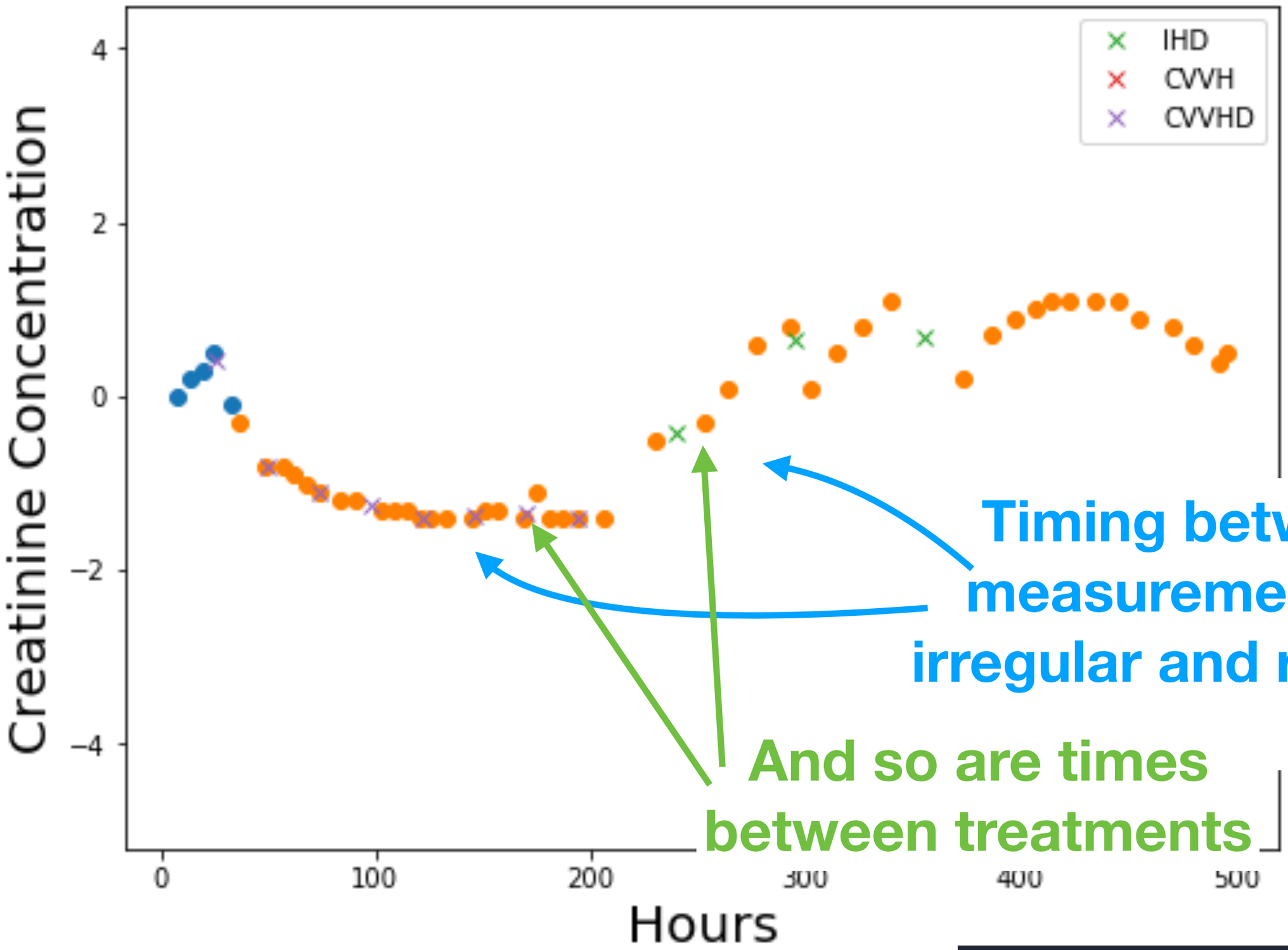
Abstract

Decision-makers are faced with the challenge of estimating what is likely to happen when they take an action. For instance, if I choose not to treat this patient, are they likely to die? Practitioners commonly use supervised learning algorithms to fit predictive models that help decision-makers reason about likely future outcomes, but we show that this approach is unreliable, and sometimes even dangerous. The key issue is that supervised learning algorithms are highly sensitive to the policy

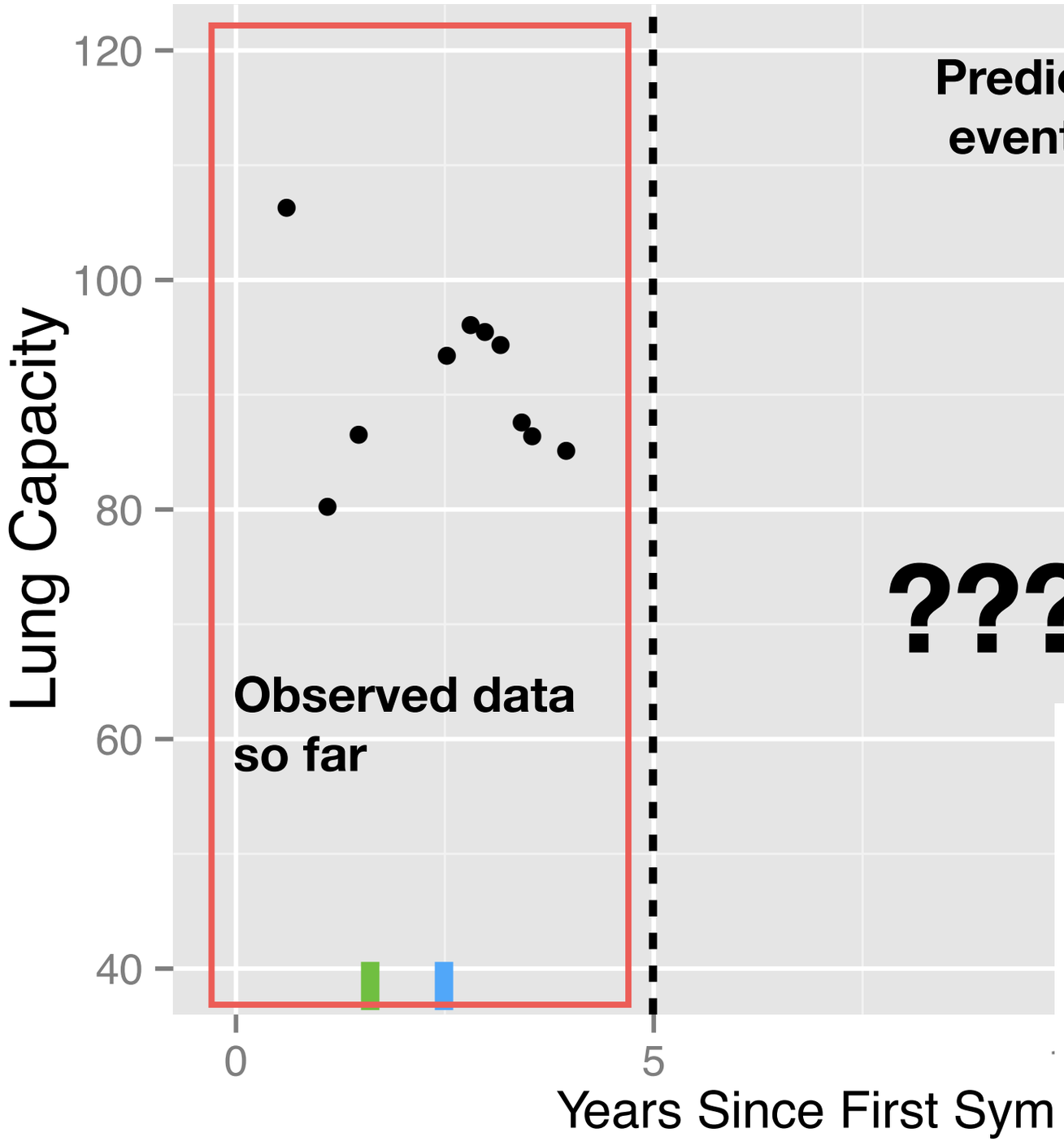


- $P(Y \mid C, L_1, L_2, L_3, O_1, O_2, O_3)$ **is unstable**

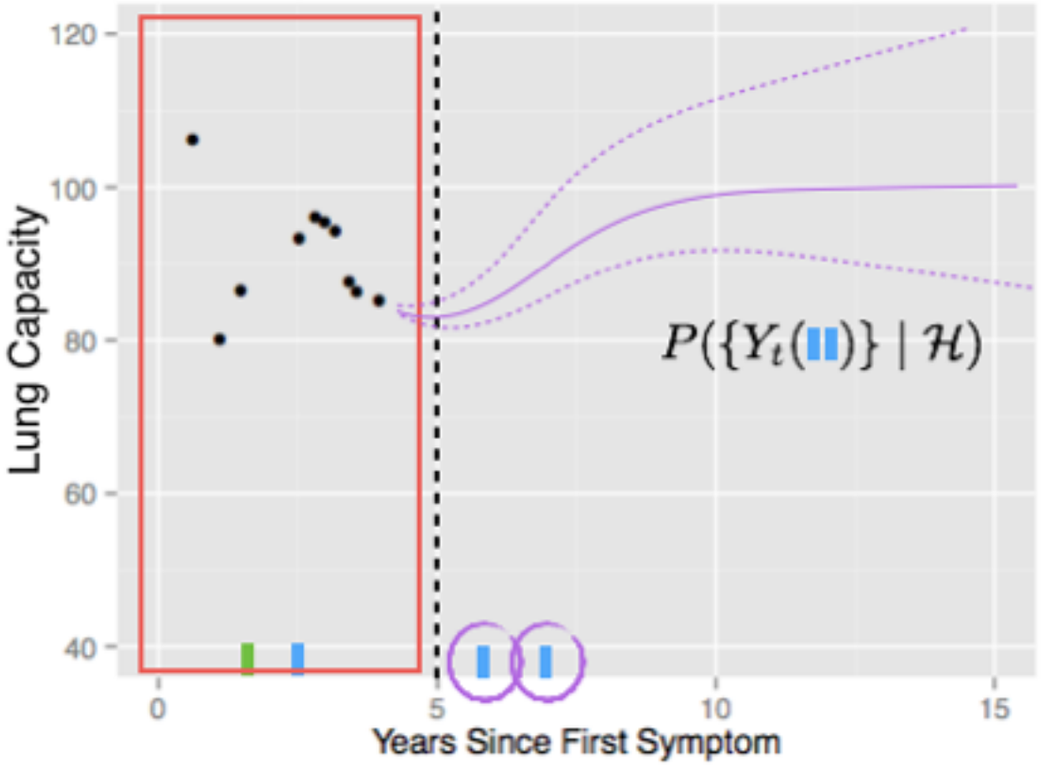
Observational Longitudinal Traces



Schulam and Saria, NIPS 2017

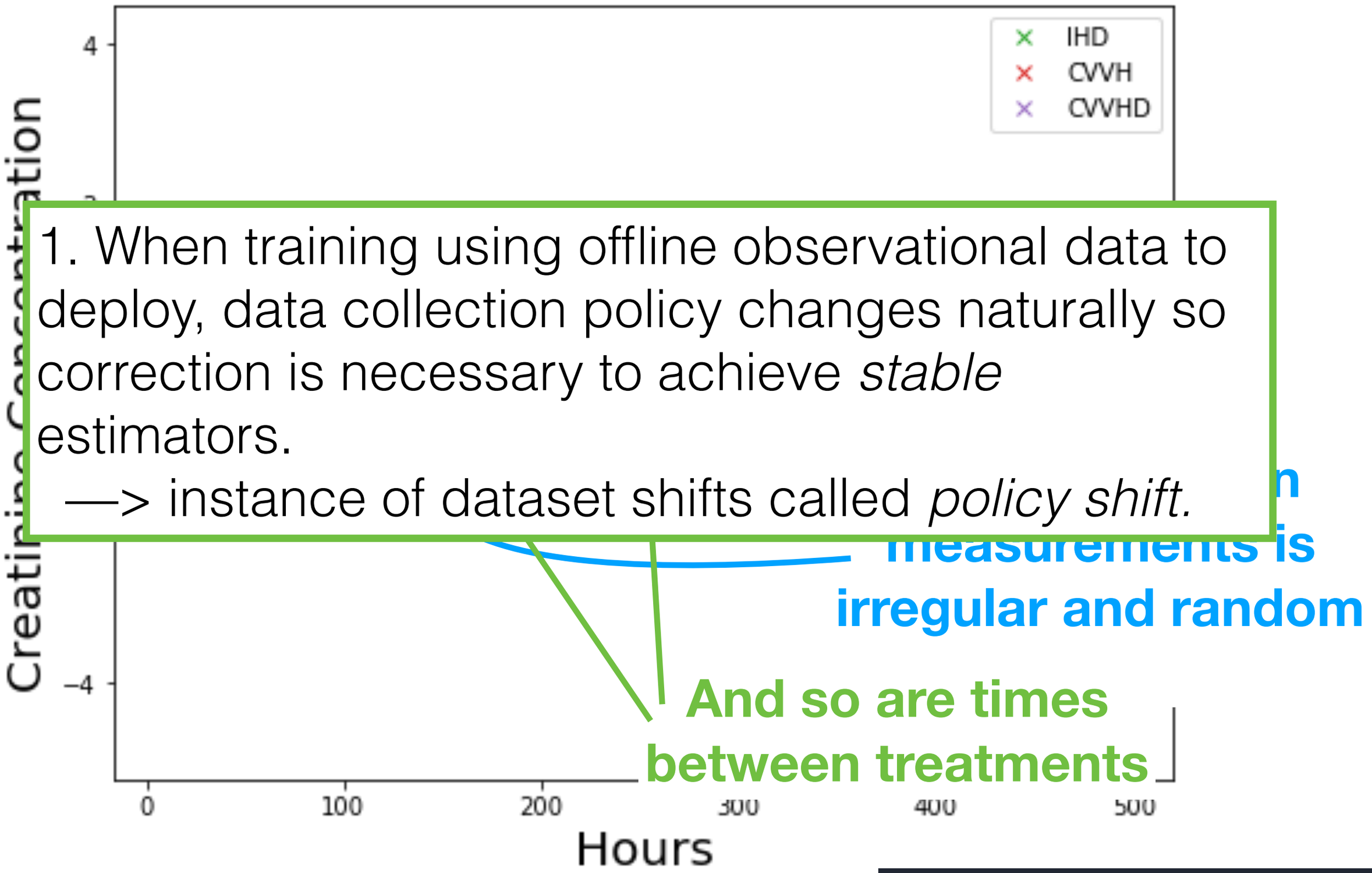


What happens if I give them a specific treatment regimen?



Schulam, P and Saria, S. “Reliable Decision Support using Counterfactual Models”, Neural Information Processing Systems, 2017.

Observational Longitudinal Traces



1. When training using offline observational data to deploy, data collection policy changes naturally so correction is necessary to achieve *stable* estimators.
   —> instance of dataset shifts called *policy shift.*

measurements is irregular and random

And so are times between treatments

Schulam and Saria, NIPS 2017