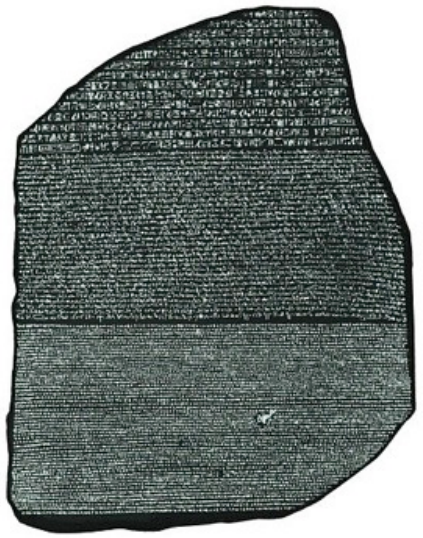
Foundation of Data Science
Lecture 10, Module 2
Fall 2022

Rumi Chunara, PhD

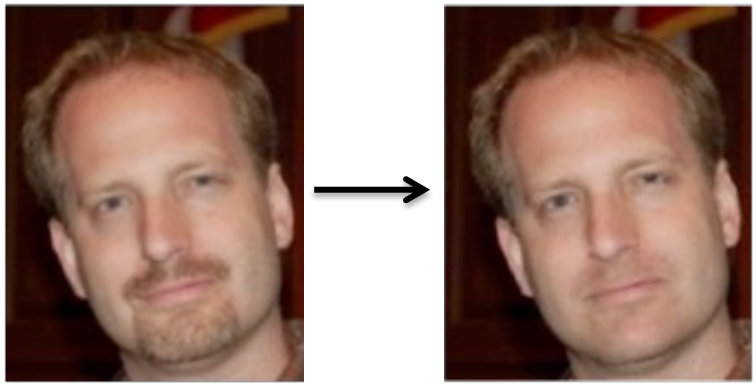
Machine Learning is good for:

Photo: Matthew W. Jackson

Facial recognition



Rosetta Stone



Nguyen et al., 2008

Real-world Topics

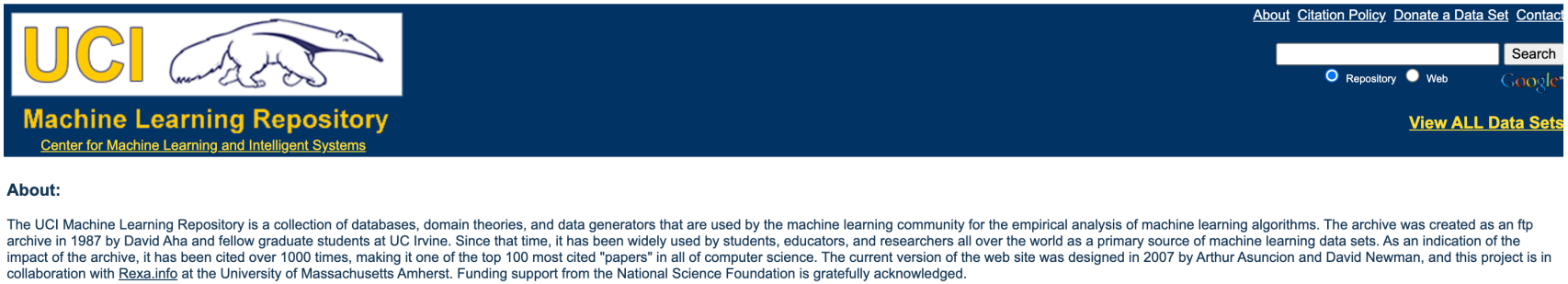
- Data and generalizability
- Interpretability
- Impact

Real-world Topics

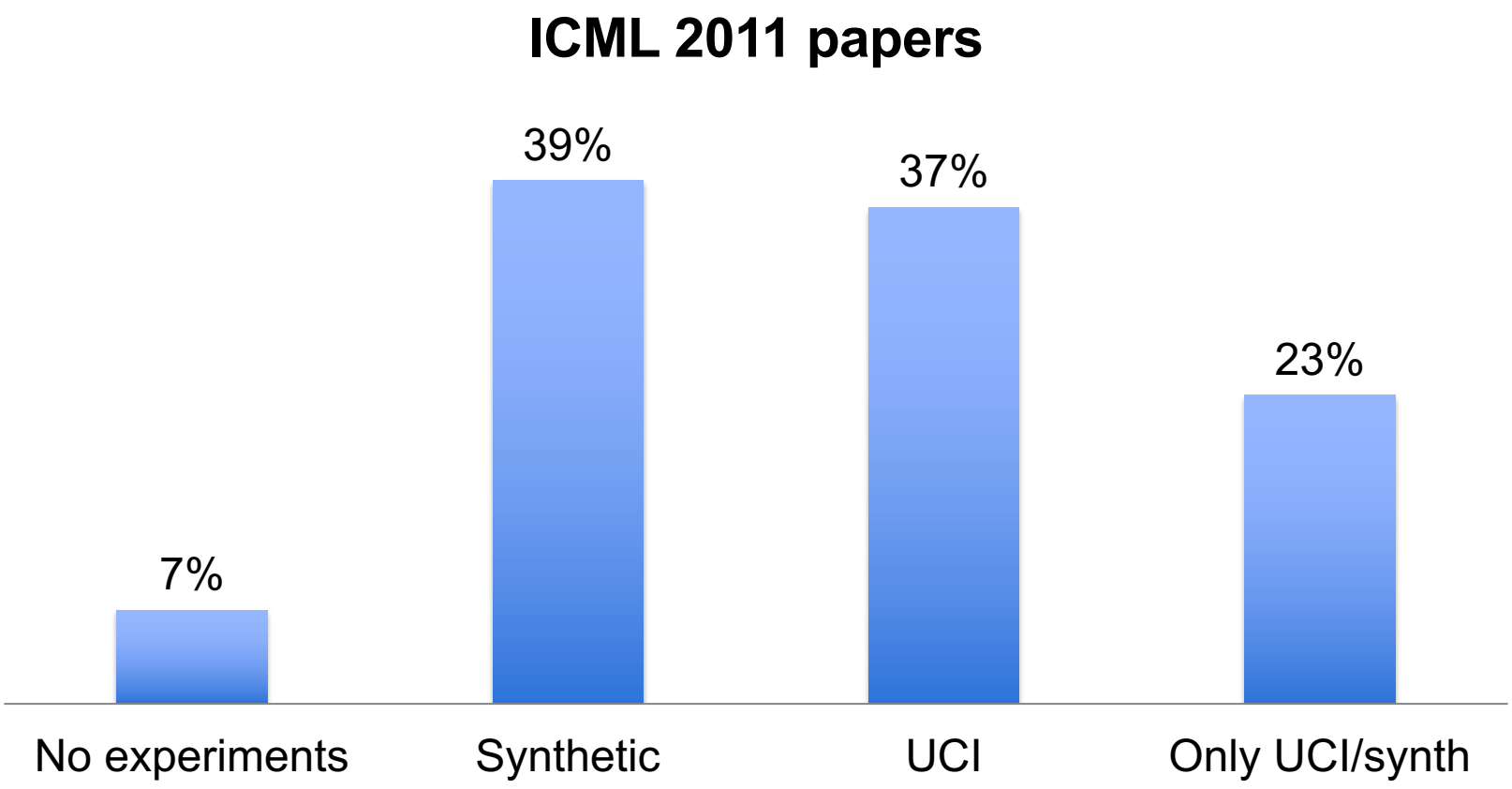
- Data and generalizability
- Interpretability
- Impact

Example: UCI data sets

"The standard Irvine data sets are used to determine percent accuracy of concept classification, without regard to performance on a larger external task."



UCI Data Sets ~~today~~ recently



Data Sets disconnected from meaning?

UCI initially

…

"Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one." – UCI Mushroom data set page

UCI today

3.2	1.5	2.9
2.6	1.8	3.1
2.9	1.4	3.3

1.2	-3.2	8.5
1.8	-2.7	7.9
0.9	1.3	8.2

0.1	0.8	4.7
0.3	0.7	4.9
-0.2	0.7	5.0

…

Did you know that the mushroom data set has 3 classes, not 2?
Have you ever used this knowledge to interpret your results on this data set?

Data Sets can be useful benchmarks

Too often, we fail at both goals

1. Enable direct empirical comparisons with other techniques
   - And reproducing others' results

No standard for reproducibility

2. Easier to interpret results since data set properties are well understood

We don't actually understand these data sets

The field doesn't require any interpretation

Real-world Topics

- Data and generalizability
- Interpretability
- Impact

Real-world Topics

- Data and generalizability
- Interpretability
- Impact

Interpretability: Academic Research

Interpretable Models for Decision-Making

[KDD'16, AISTATS'17, FAT ML'17, AIES'19]

Reliable Evaluation of Models for Decision-Making

[QJE'18, KDD'17, KDD'15]

Characterizing Biases in Human & Machine Decisions

[NIPS'16, SDM'15]

Diagnosing Failures of Predictive models

[AAAI'17]

Motivation for Interpretability

- ML systems are being deployed in complex high-stakes settings
- Accuracy alone is no longer enough
- Auxiliary criteria are important:
  - Safety
  - Nondiscrimination
  - Right to explanation

Motivation for Interpretability

- Auxiliary criteria are often hard to quantify (completely)
  - E.g.: Impossible to enumerate all scenarios violating safety of an autonomous car
- Fallback option: interpretability
  - *If the system can explain its reasoning, we can verify if that reasoning is sound w.r.t. auxiliary criteria*

Prior Work: Defining & Measuring Interpretability

- Little consensus on what interpretability is and how to evaluate it
- Interpretability evaluation typically falls into:
  - Evaluate in the context of an application
  - Evaluate via a quantifiable proxy

Prior Work: Defining & Measuring Interpretability

- Evaluate in the context of an application
  - If a system is useful in a practical application or a simplified version, it must be interpretable

- Evaluate via a quantifiable proxy
  - Claim some model class is interpretable and present algorithms to optimize within that class
  - E.g. rule lists

Lack of Rigor?

- Yes and No
  - Previous notions are reasonable

- However,
  - Are all models in all "interpretable" model classes equally interpretable?
    - Model sparsity allows for comparison
  - How to compare a model sparse in features to a model sparse in prototypes?
  - Do all applications have same interpretability needs?

Important to formalize these notions

What is Interpretability?

- **Definition**: Ability to explain or to present in understandable terms to a human

- No clear answers in psychology to:
  - What constitutes an explanation?
  - What makes some explanations better than the others?
  - When are explanations sought?

Leads to work on: data-driven ways to derive operational definitions and evaluations of explanations and interpretability

When and Why Interpretability?

- Not all ML systems require interpretability
  - E.g., ad servers, postal code sorting
  - No human intervention

- No explanation needed because:
  - No consequences for unacceptable results
  - Problem is well studied and validated well in real-world applications → trust system's decision

When do we need explanation then?

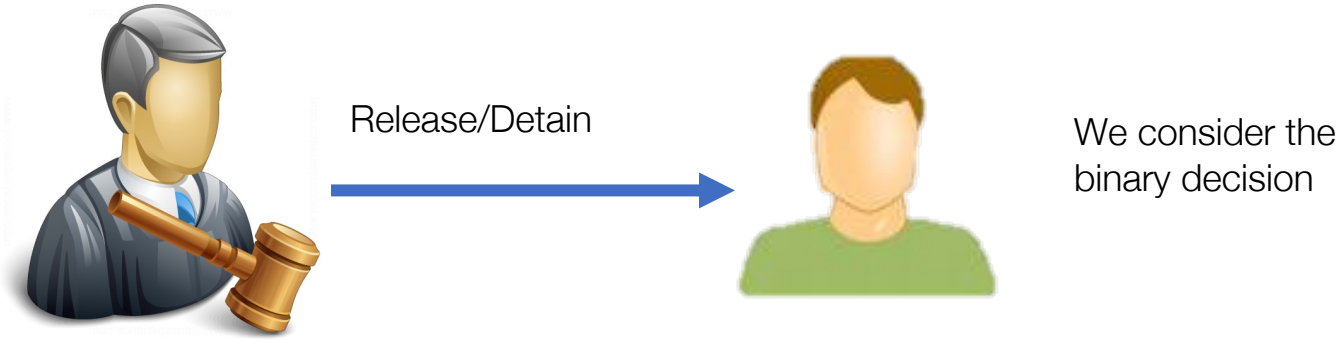
When and Why Interpretability?

- *Incompleteness* in problem formalization
  - Hinders optimization and evaluation

- Incompleteness ≠ Uncertainty
  - Uncertainty can be quantified
  - E.g., trying to learn from a small dataset (uncertainty)

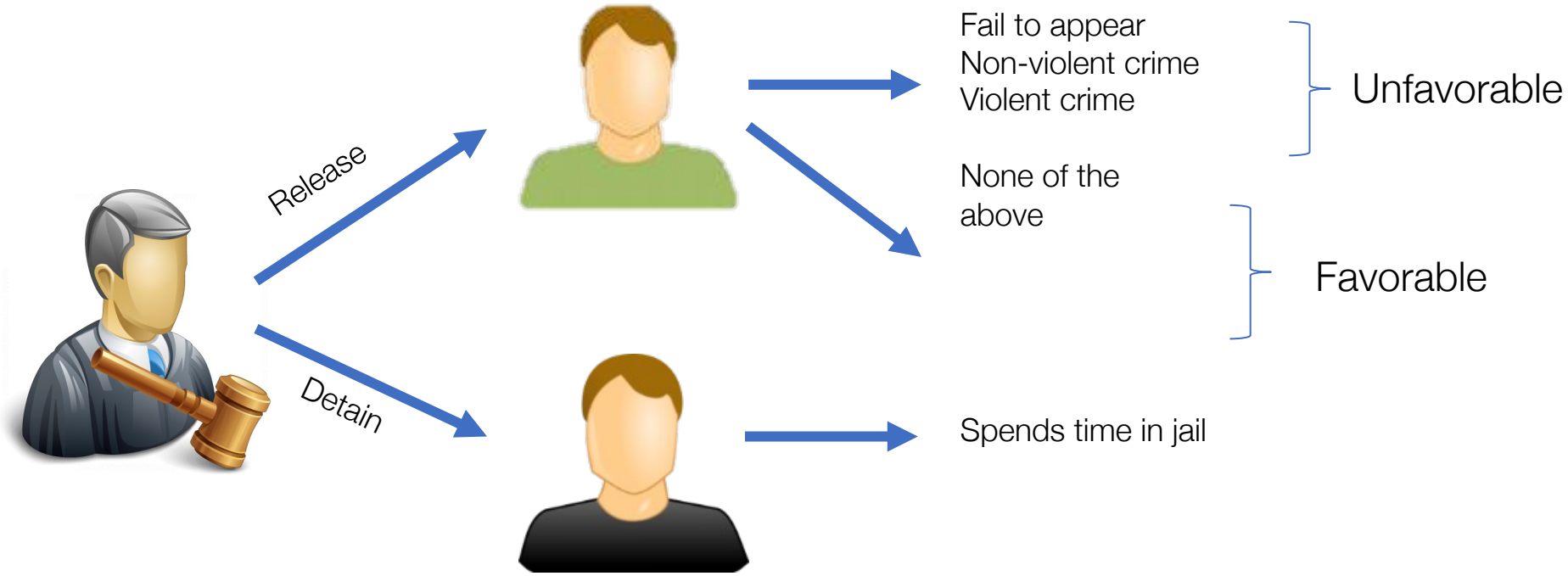
Real World Scenario: Bail Decision

- U.S. police make about 12M arrests each year



- Release vs. Detain is a high-stakes decision
  - Pre-trial detention can go up to 9 to 12 months
  - Consequential for jobs & families of defendants as well as crime

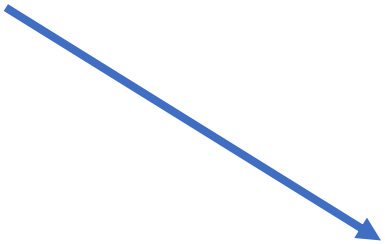
Bail Decision



Judge is making a prediction:
Will the defendant commit 'crime' if released on bail?

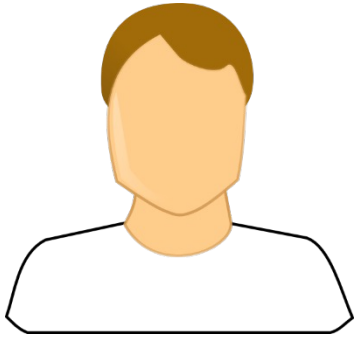
Bail Decision-Making as a Prediction Problem

Does making the model more understandable/transparent to the judge improve decision-making performance?
If so, how to do it?



Predictive Model

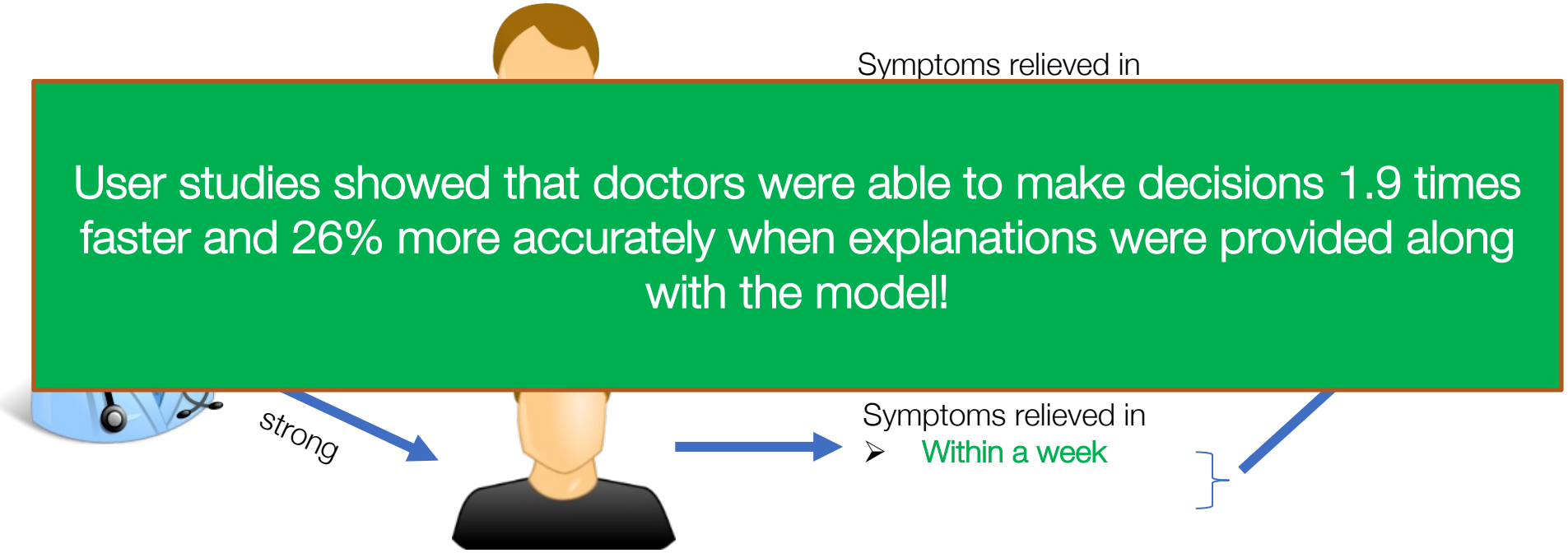
Real World Scenario: Treatment Recommendation



What treatment should be given?
Options: quick relief drugs (mild), controller drugs (strong)

Demographics:
Age
Gender
…..
Medical History:
Has asthma?
Other chronic issues?
……
Symptoms:
Severe Cough
Wheezing
……
Test Results:
Peak flow: Positive
Spirometry: Negative

Treatment Recommendation



User studies showed that doctors were able to make decisions 1.9 times faster and 26% more accurately when explanations were provided along with the model!

Doctor is making a prediction:
Will the patient get better with a milder drug?
Use ML to make a similar prediction

Post-hoc: Text Explanations

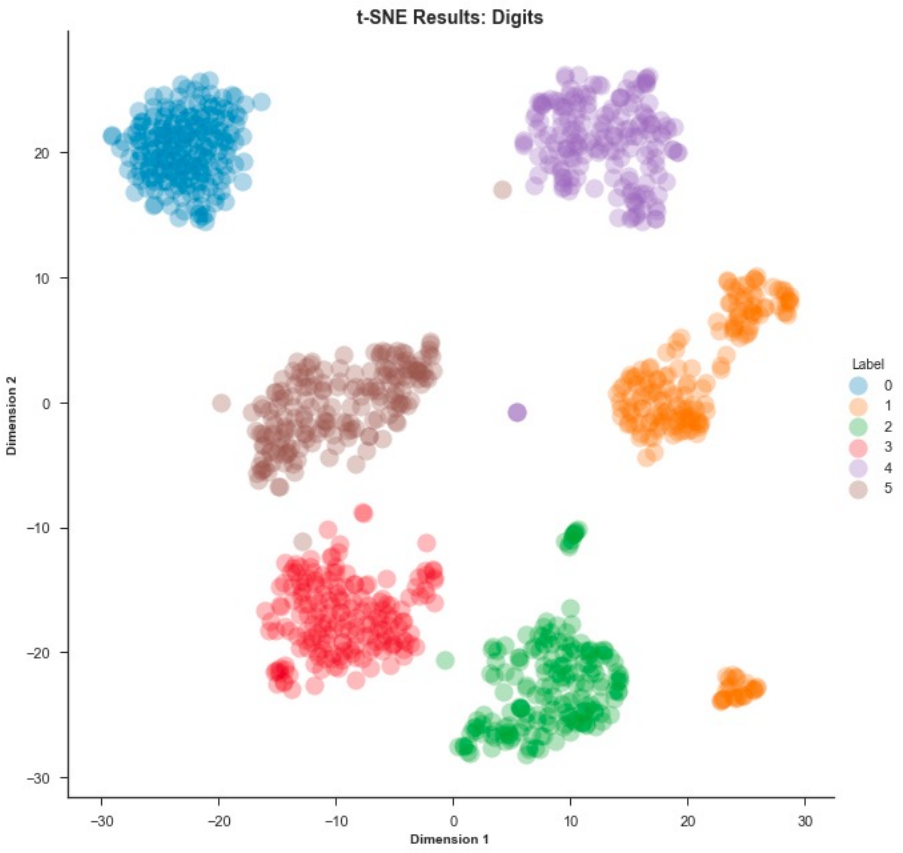
- Humans often justify decisions verbally (post-hoc)

- Krening et. al.:
  - One model is a reinforcement learner
  - Another model maps models states onto verbal explanations
  - Explanations are trained to maximize likelihood of ground truth explanations from human players
  - So, explanations do not faithfully describe agent decisions, but rather human intuition

Krening, Samantha, et al. "Learning from explanations using sentiment and advice in RL." *IEEE Transactions on Cognitive and Developmental Systems* 9.1 (2016): 44-55.

Post-hoc: Visualization

- Visualize high-dimensional data with t-SNE
  - 2D visualizations in which nearby data points appear close
- Perturb input data to enhance activations of certain nodes in neural nets (image classification)
  - Helps understand which nodes corresponds to what aspects of the image
  - E.g., certain nodes might correspond to dog faces



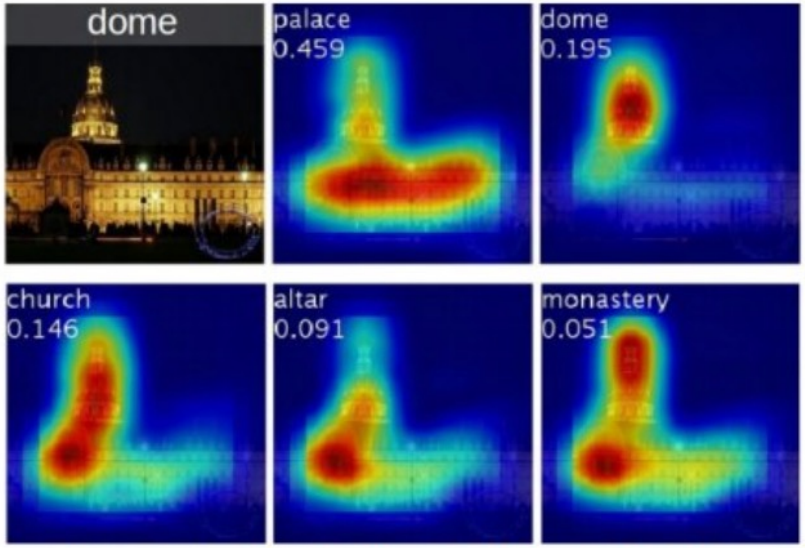
https://towardsdatascience.com/an-introduction-to-t-sne-with-python-example-5a3a293108d1

Post-hoc: Example Explanations

- Reasoning with examples
- E.g., Patient A has a tumor because he is similar to these k other data points with tumors
- k neighbors can be computed by using some distance metric on learned representations
  - E.g., word2vec

Post-hoc: Activation maps

Class activation maps are a simple technique to get the discriminative image regions used by a CNN to identify a specific class in the image. In other words, a class activation map (CAM) lets us see which regions in the image were relevant to this class.



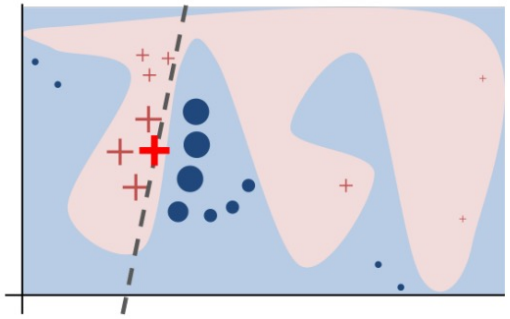
Class activation maps of top 5 predictions

Class activation maps for one object class

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Post-hoc: Local Explanations

- Hard to explain a complex model in its entirety
  - How about explaining smaller regions?



LIME (Ribeiro et. al.)

  - Explains decisions of any model in a local region around a particular point

  - Learns sparse linear model

Evaluating Interpretability

- Evaluating interpretability in the interpretable ML community:
  - Interpretability depends on human experience of the model
  - Disagreement about the best way to measure it
- These papers:
  - Evaluating factors related to interpretability through user studies

Other Relevant Fields

- Human-Computer Interaction (HCI):
  - Theories for how people interact with technology
- Psychology:
  - Theories for how people process information
- Both have thought carefully about experimental design

Manipulating interpretability

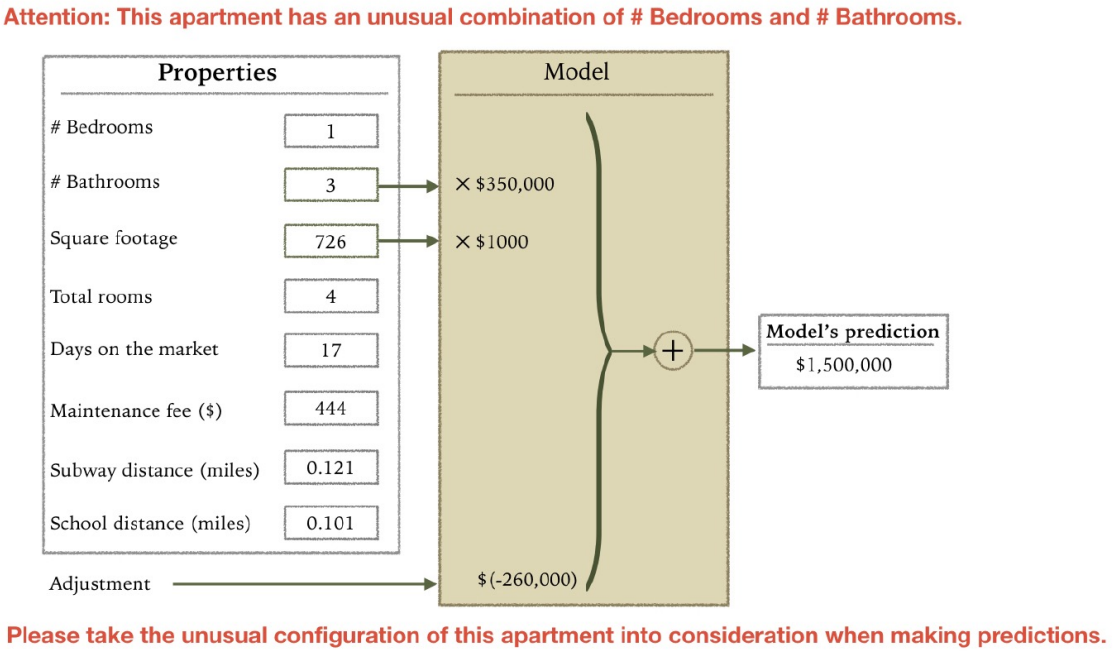
- Interpretability as a *latent property* that can be manipulated or measured indirectly
- What are the factors through which it can be manipulated effectively?
- Bring HCI methods to interpretable ML since interpretability is defined by user experience

Poursabzi-Sangdeh, Forough, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. "Manipulating and measuring model interpretability." In Proceedings of the 2021 CHI conference on human factors in computing systems, pp. 1-52. 2021.

Manipulating interpretability

- Research Questions:
  - How well can people estimate what a model will predict?
  - How much do people trust a model's predictions?
  - How well can people detect when a model has made a sizable mistake?

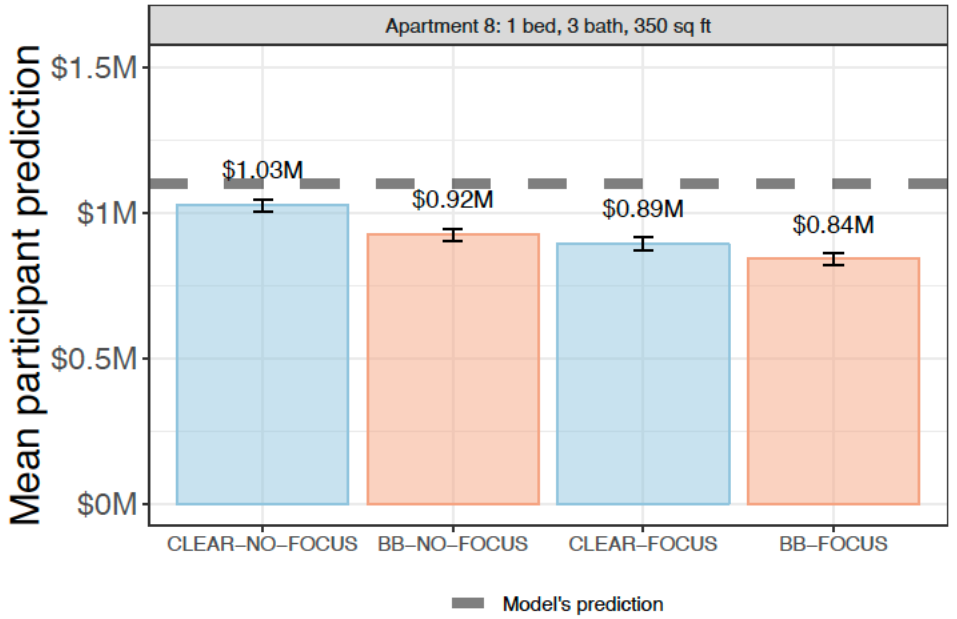
Approach:
Large-scale, pre-registered user studies to answer these questions in the context of linear regression models



Attention checks -> more deviation

Experiment 4: Attention check

Mean participant prediction
Lower is better

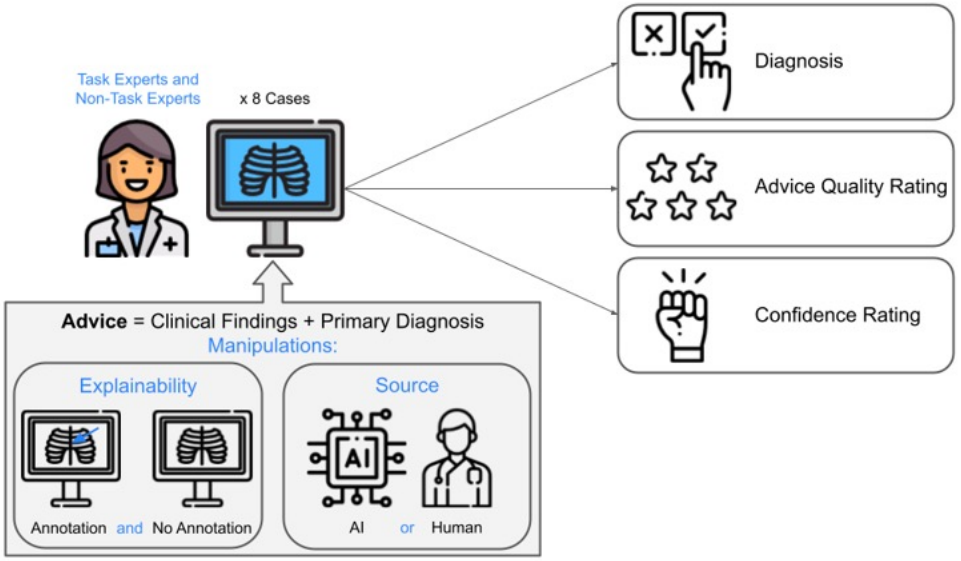


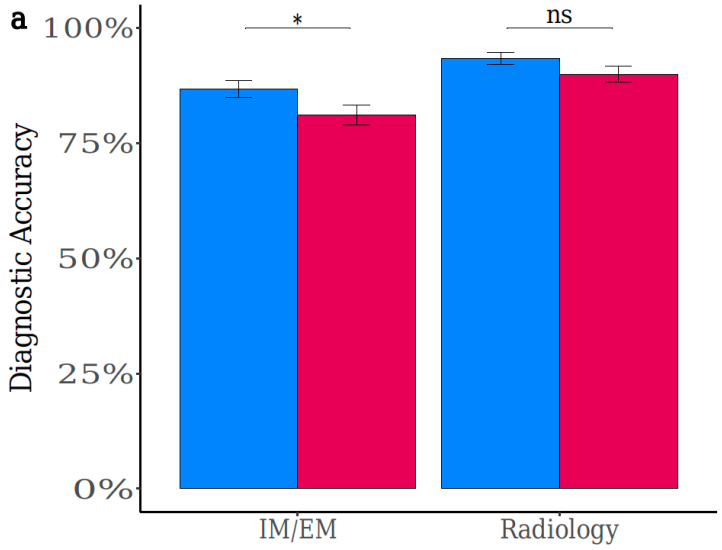
Explicit attention checks improve people's ability to catch errors

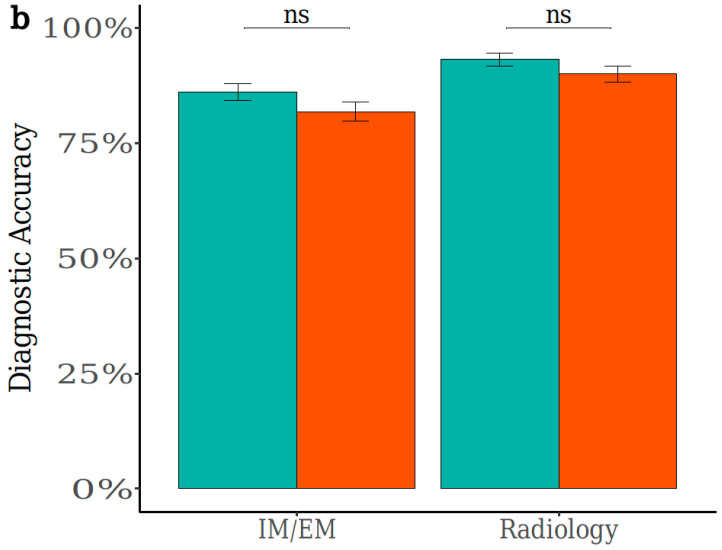
How do experts interpret AI decisions?

We manipulated whether the advice came with or without a visual annotation on the X-rays and whether it was labeled as coming from an AI or a human radiologist. Overall, receiving annotated advice from an AI resulted in the highest diagnostic accuracy. Physicians rated the quality of AI advice higher than human advice. Neither manipulation had strong effects on participants' confidence. Importantly, the results varied among task experts and non-task experts, with only the latter considerably benefiting from correct explainable AI advice.

Gaube, S., Suresh, H., Raue, M., Lermer, E., Koch, T., Hudecek, M., Ackery, A.D., Grover, S.C., Coughlin, J.F., Frey, D. and Kitamura, F., 2022. Who should do as AI say? Only non-task expert physicians benefit from correct explainable AI advice.







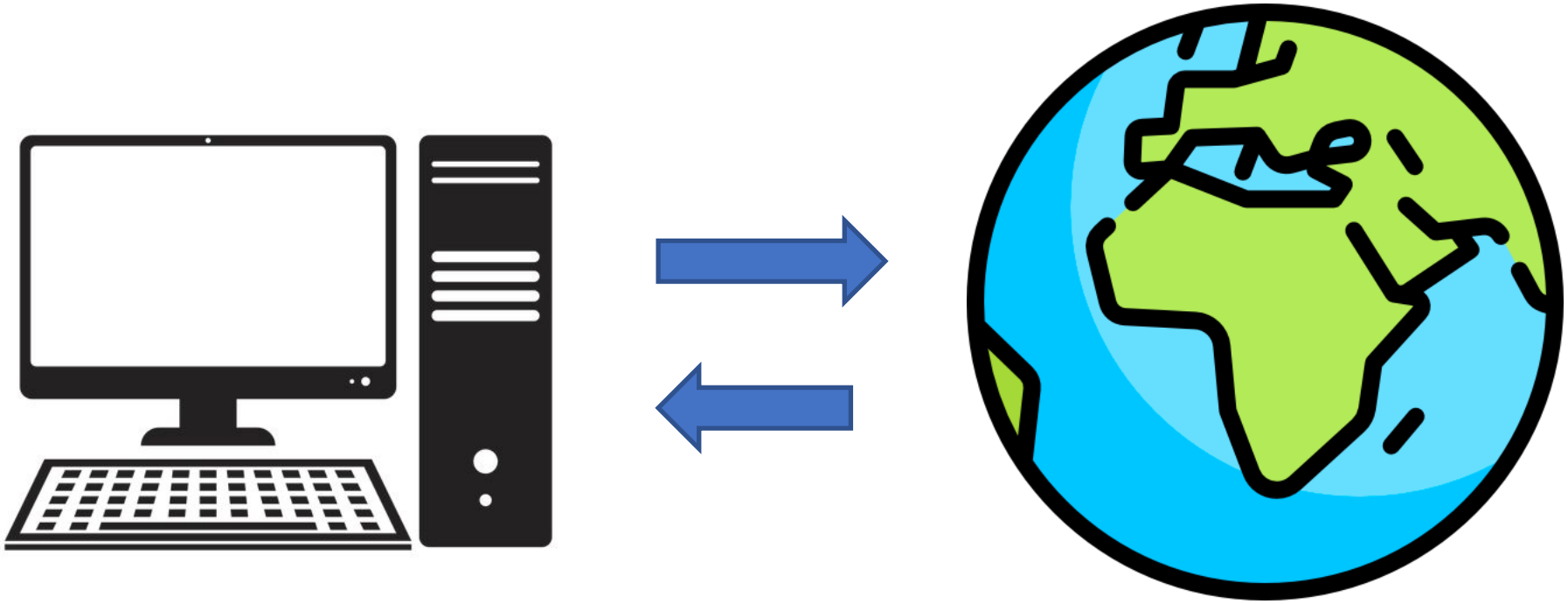
Real-world Topics

- Data and generalizability
- Interpretability
- Impact

Real-world Topics

- Data and generalizability
- Interpretability
- Impact

What is its impact?

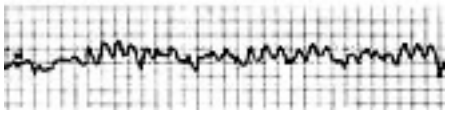


Evaluation metrics depend on task

- How your improvement matters to the originating field

96% accuracy in separating poisonous and edible mushrooms? Not good enough for me to trust it!

4.6% improvement in detecting cardiac arrhythmia? We could save lives!



ML Impact Limiting Factors

- Data sets disconnected from meaning
- Metrics disconnected from impact
- Lack of follow-through

Metrics Disconnected from Impact

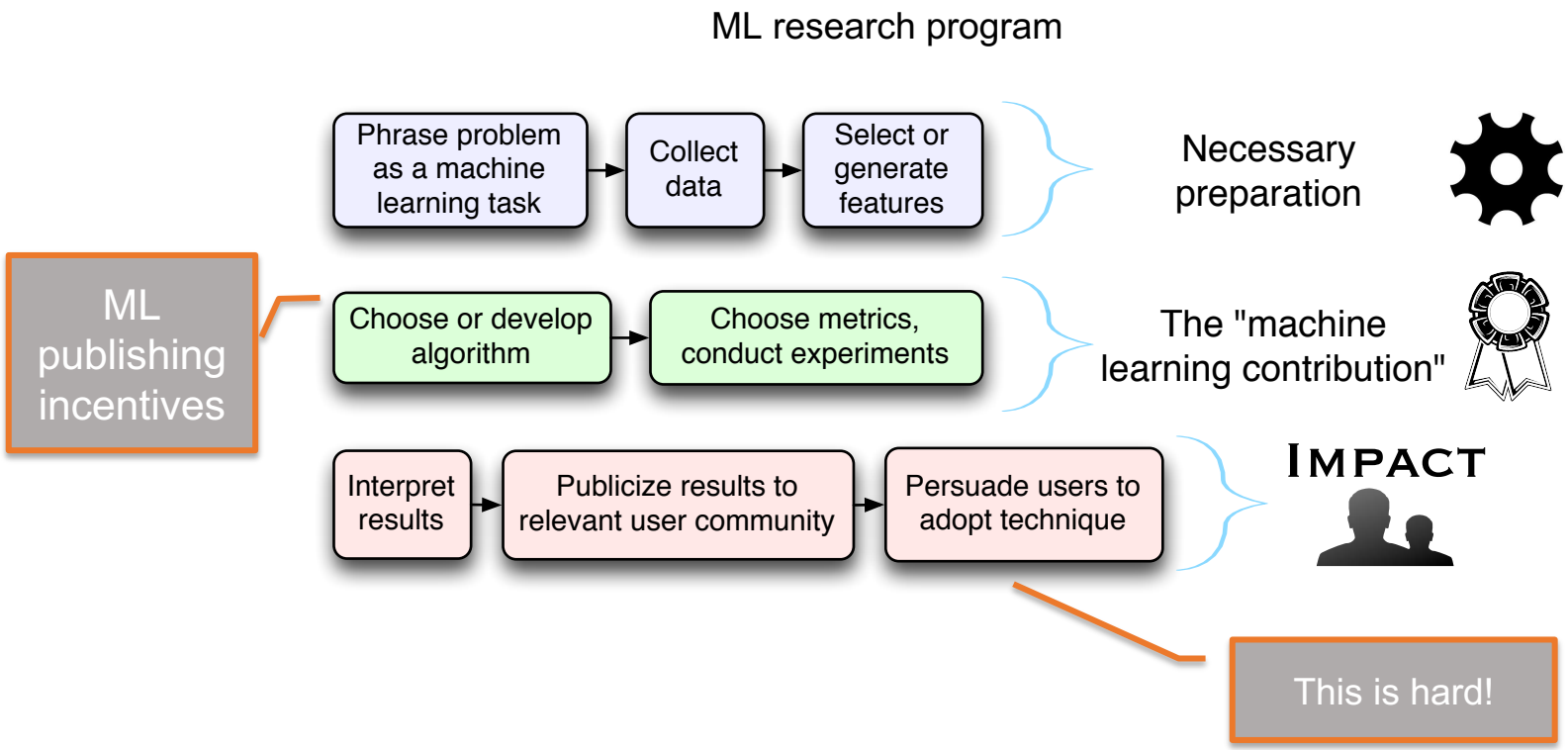
- Accuracy, RMSE, precision, recall, F-measure, AUC, …
- Deliberately ignore problem-specific details
- Cannot tell us
  - WHICH items were classified correctly or incorrectly?
  - What impact does a 1% change have? (What does it mean?)
  - How to compare across problem domains?

This doesn't mean accuracy, etc. are bad measures,
just that they should not remain abstractions

"A Machine Learning Approach
to the Detection of Fetal Hypoxia during Labor and Delivery"
by Warrick et al., 2010

"The approach we proposed in this paper
detected correctly half of the pathological cases,
with acceptable false positive rates (7.5%),
early enough to permit clinical intervention."

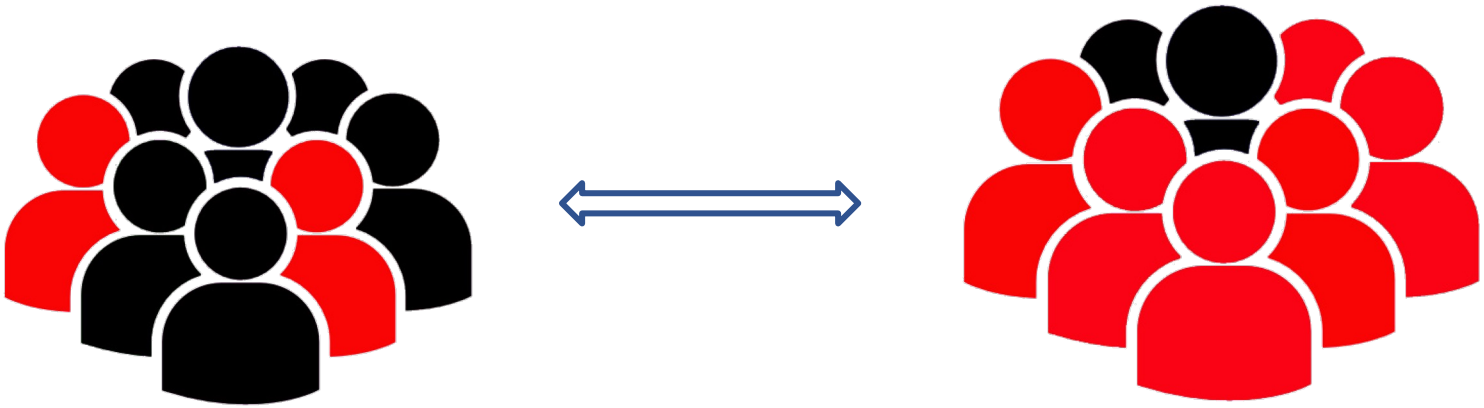
Workflow - follow through and incentives



Challenges for Increasing Impact

- Increase the impact of work
  1. Employ meaningful evaluation methods
     - Direct measurement of impact when possible
     - Translate abstract metrics into domain context
  2. Involve the world outside of ML
  3. Choose problems to tackle biased by expected impact

Generalizability, Robustness and Equity



HEALTH TECH

STAT+

Medical AI systems are disproportionately built with data from just three states, new research finds

By Rebecca Robbins Sept. 25, 2020

Reprints

Research Letter

September 22/29, 2020

Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms

Amit Kaushal, MD, PhD[1]; Russ Altman, MD, PhD[1]; Curt Langlotz, MD, PhD[2]

Distribution Shifts in Healthcare is a Challenge

Field trial of an algorithm to detect diabetic retinopathy from eye scans [Beede+ 2020]
Different lighting conditions across clinics

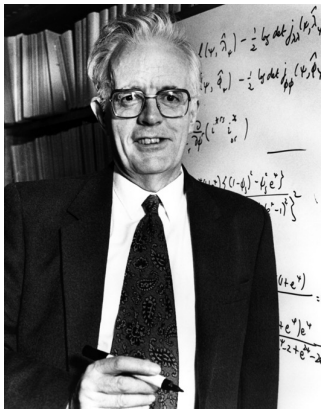
Different populations, “case mix” [Riley+ 2018]

Performance drop across time/datasets [Nestor+ 2019, Zech+ 2018]

Image: https://blog.google/technology/health/healthcare-ai-systems-put-people-center/

Causality is Important for Prediction Problems Too

Cox, D. R. (2001), Comment on "Two cultures"

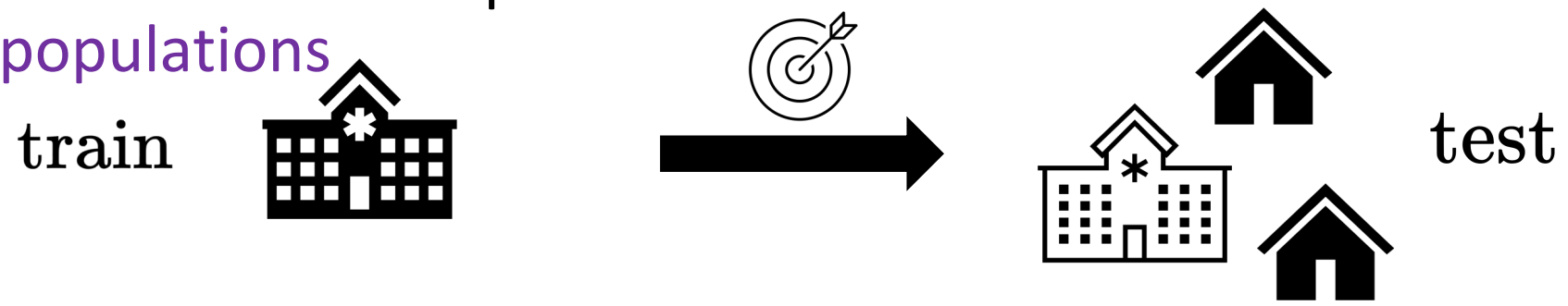


"... Often the prediction is under quite different conditions from the data; what is the likely progress of the incidence of the epidemic of v-CJD in the United Kingdom, …

As we move toward such more ambitious tasks, prediction, always hazardous, without some understanding of underlying process and linking with other sources of information, becomes more and more tentative…"

Methods to mitigate prediction with shifts

Consider a set of possible test distributions or sub-populations



University hospital in Manhattan

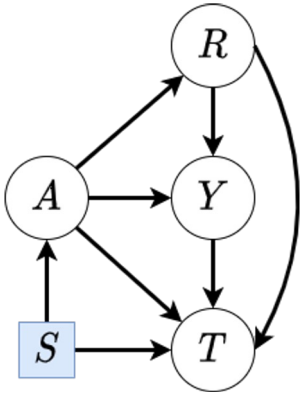
All (public, private) hospitals in NYC

Goal: Estimate a model that *performs well* across test distributions

Singh, H., Singh, R., Mhasawade, V. and Chunara, R., 2021, March. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and*

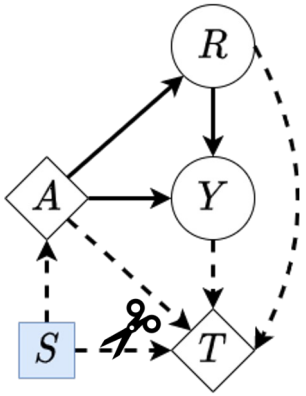
Methods to mitigate prediction with shifts

Step 1: Represent shifts



using an augmented causal graph

Step 2: Identify desirable models

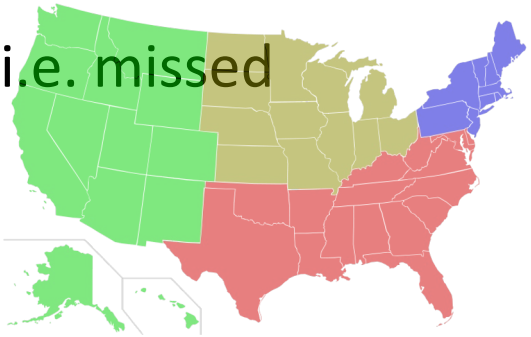


using interventions to ensure invariance

Singh, H., Singh, R., Mhasawade, V. and Chunara, R., 2021, March. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and*

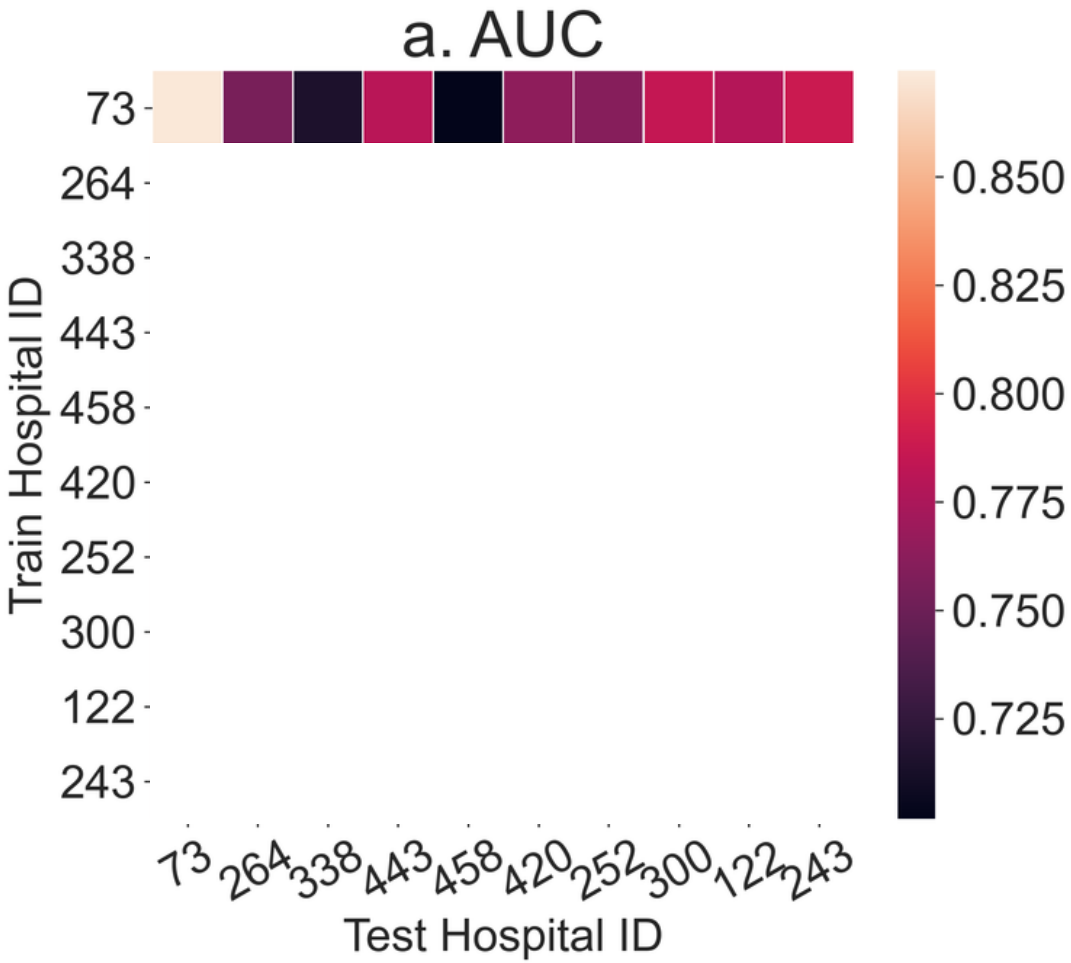
Real-world ICU data

- Electronic health records from **10 hospitals in US** from 2014 to 2015 extracted from eICU database
- Train on one hospital, test on another
- Model **performance** measures
  - AUROC and calibration slope
  - **Group differences** in false negative rates i.e. missed prognoses
  - Across hospitals and across 4 US regions



Singh, H., Mhasawade, V. and Chunara, R., 2022. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, *1*(4), p.e0000023.

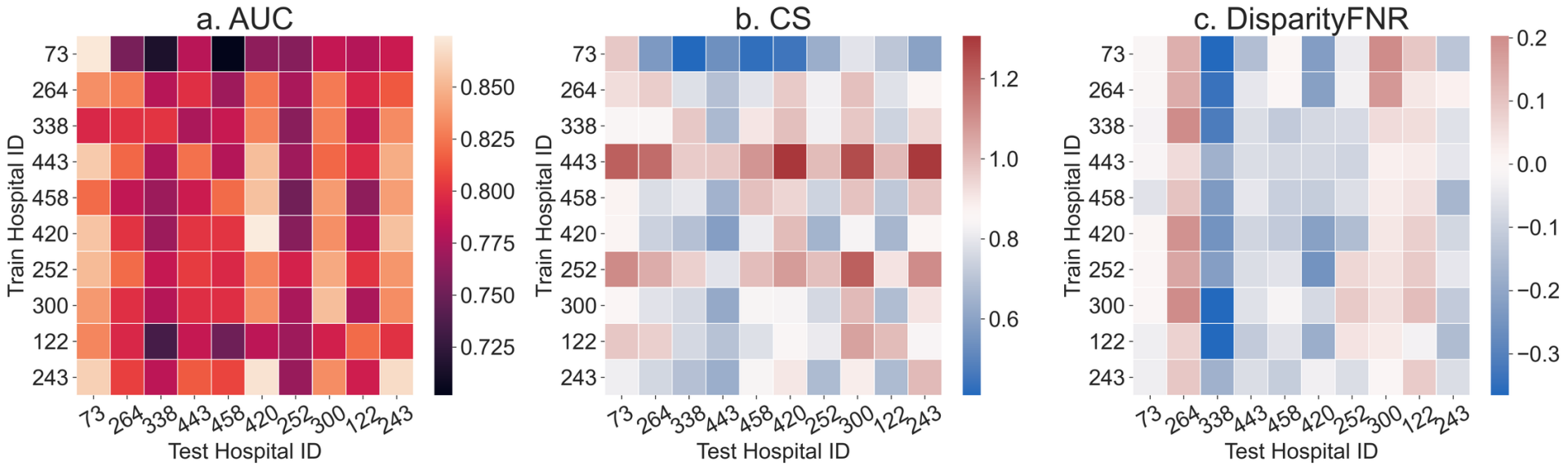
Real-world ICU data



Singh, H., Mhasawade, V. and Chunara, R., 2022. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, *1*(4), p.e0000023.

Real-world ICU data

1. **High variation** in accuracy, calibration across hospitals
2. Disparate performance across race variable **persists** even after aggregating hospitals into US regions
3. Changes in **all variable types** - demographics, labs, and vitals



Singh, H., Mhasawade, V. and Chunara, R., 2022. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, *1*(4), p.e0000023.

Real-world ICU data

A multi-center study which tests **geographic variation** in model performance in terms of both **standard and fairness** metrics

1. Group-level performance should be recommended in evaluation guidelines
2. Better documentation of provenance of data and health processes to identify and mitigate sources of variation

To think about…

- Where is the data from? Was it collected for the purpose you are using it? Are there any limitations to the data due to this?
- For your project, what are the appropriate evaluation metric(s)?
- Are there any real-world (interpretability, generalizability, impact) challenges in your proposal? How would you go about addressing them?
- Are there any important subgroups in the data? How does performance compare across subgroups?
- Who are your stakehodlers? What are important results to communicate?