Science of Evaluation and Explainability

DATA SCIENCE INSTITUTE™
AMERICAN COLLEGE OF RADIOLOGY

Jayashree Kalpathy-Cramer, PhD
Co-Director, Center of Machine Learning, QTIM
Athinoula A. Martinos Center for Biomedical Imaging,
Scientific Director, MGH & BWH CCDS
Associate Professor, Dept. of Radiology,
MGH/Harvard Medical School

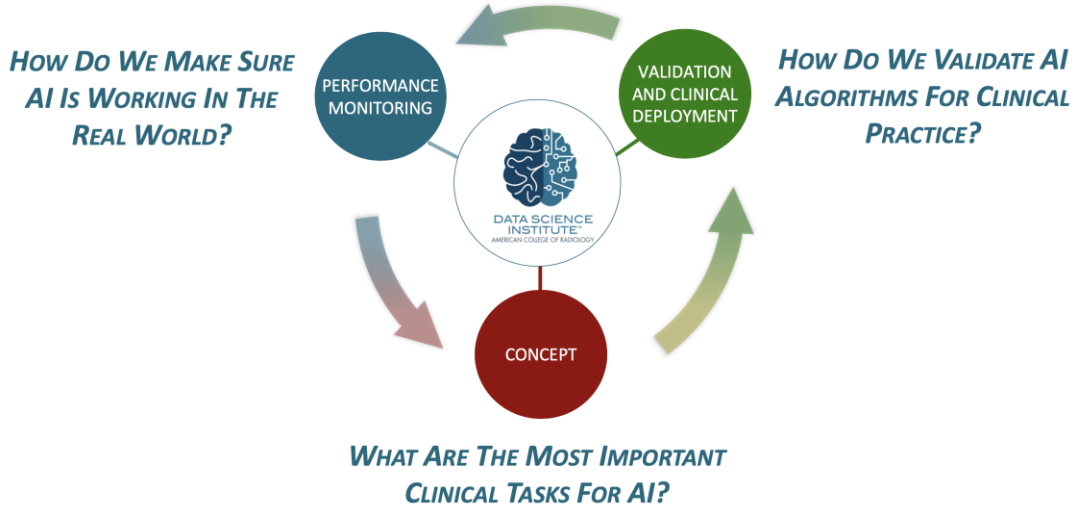
Advancing data science as core to clinically relevant, safe and effective radiologic care

"IF YOU CAN'T MEASURE IT, YOU CAN'T IMPROVE IT"

-- PETER DRUCKER/LORD KELVIN

In the context of AI in radiology, evaluation is critical

- To build trust
- Evaluate model brittleness
- Identify bias
- ***ENSURE SAFETY***



Dr. Allen

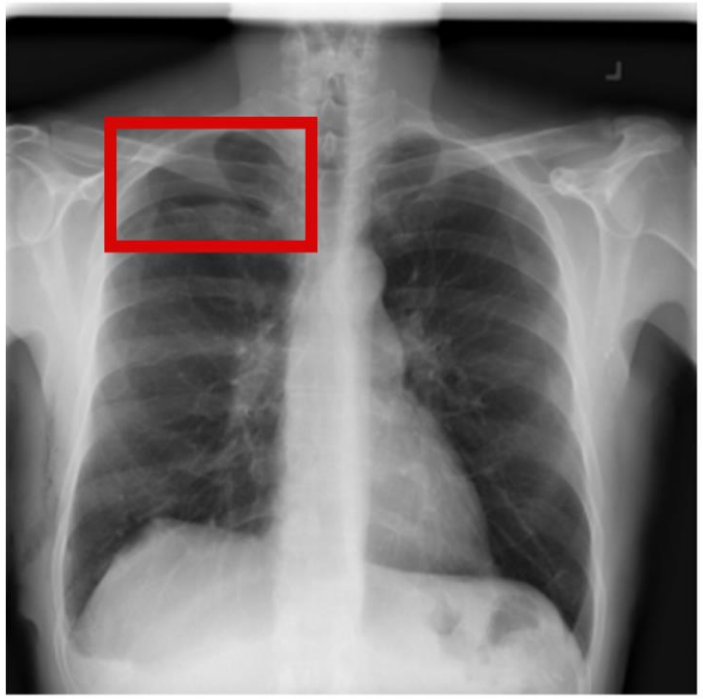
What does the model do?

Classification

Does this patient have pneumothorax?

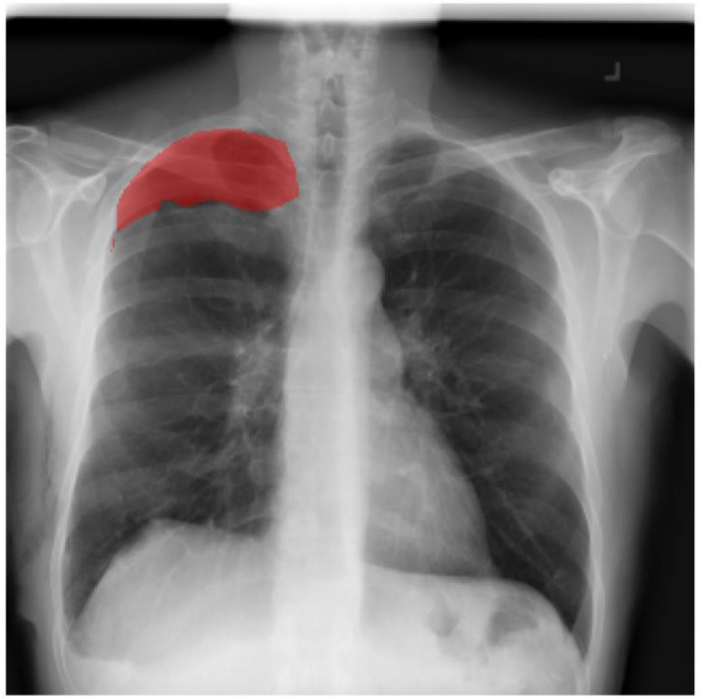
Detection

Which region is the pneumothorax in?



Segmentation

What are the boundaries of the pneumothorax?



Evaluation of classification algorithms

- Accuracy *
- Sensitivity
- Specificity
- Positive/negative predictive values
- Kappa (weighted or unweighted)
- AUROC *
- AUPRC

*Effect of class imbalance on performance metrics

Evaluation of detection algorithms

- Intersection over union (IoU), mAP
- Detection rate
- Sensitivity
- Specificity

Score : 0.995

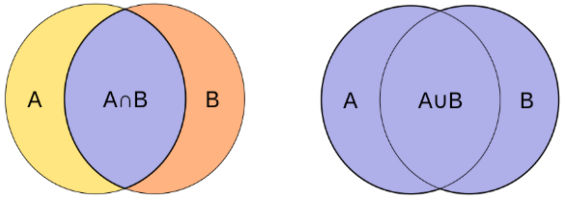
Score : 0.957 (FP) and 0.98

Score : 0.981

Visual Results on Forearm Fracture Detection
Green Box: Ground Truth | Blue Box: Predicted | Score: Classification confidence

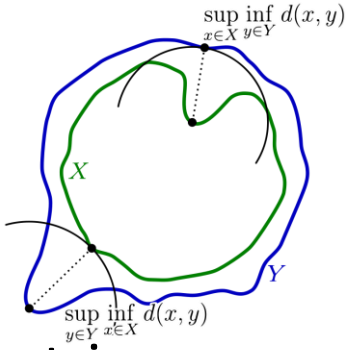
Evaluation of segmentation algorithms

Dice coefficient = $\frac{2 \times A \cap B}{A \cup B}$



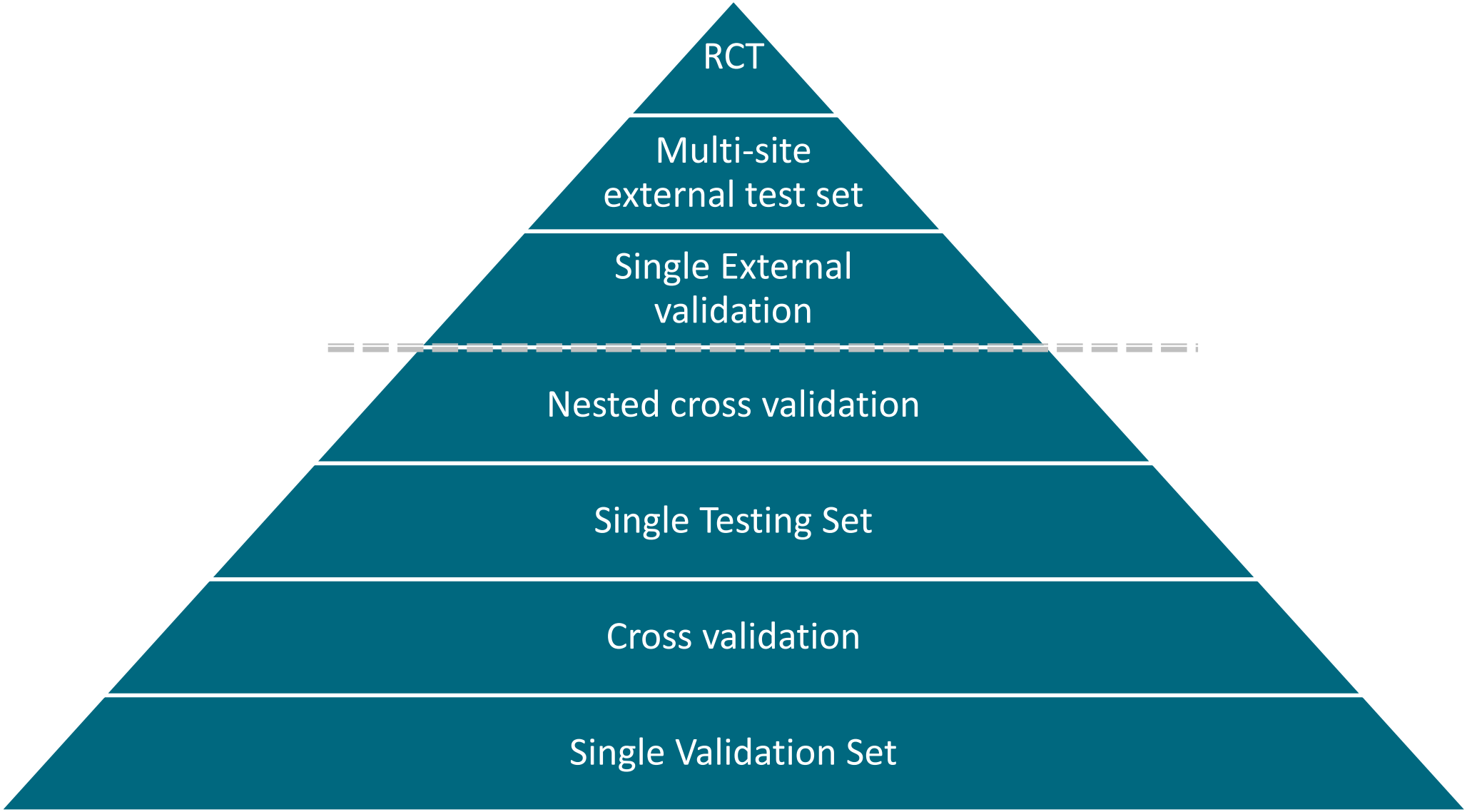
Intersection and union of two sets A and B

Hausdorff distance



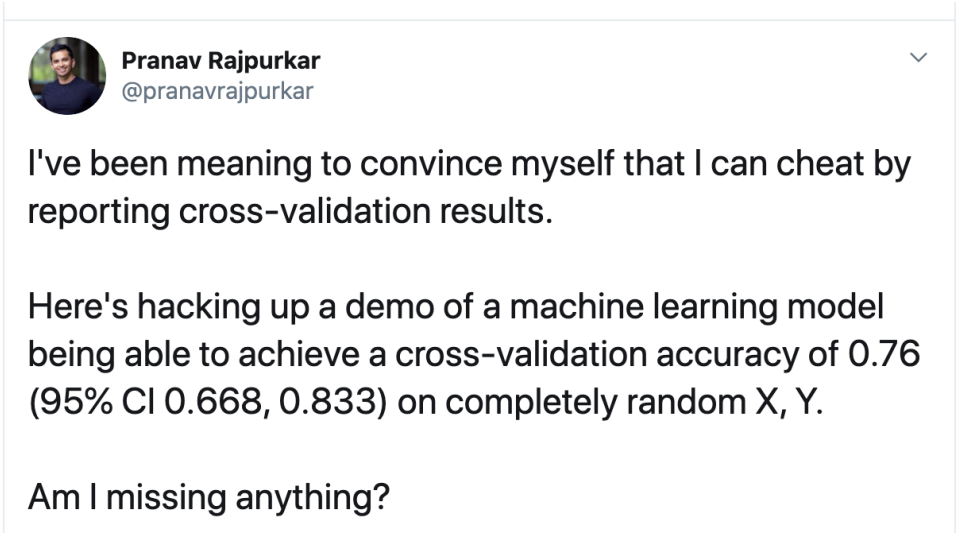
Effect of object size on performance metrics

Proposed Validation Hierarchy



Opportunities for “data leakage” with cross-validation

- Normalizing on all data
- Contamination through feature selection
- Contamination through parameter optimization



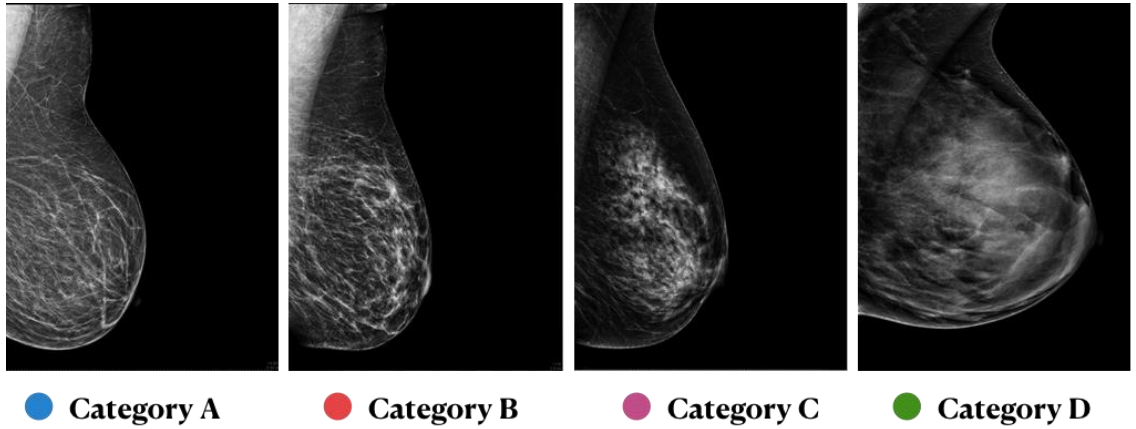
https://twitter.com/pranavrajpurkar/status/1173441639026573312

Breast density classification

Breast cancer is a leading cause of death among women in the US, with over 41,000 expected annual deaths.

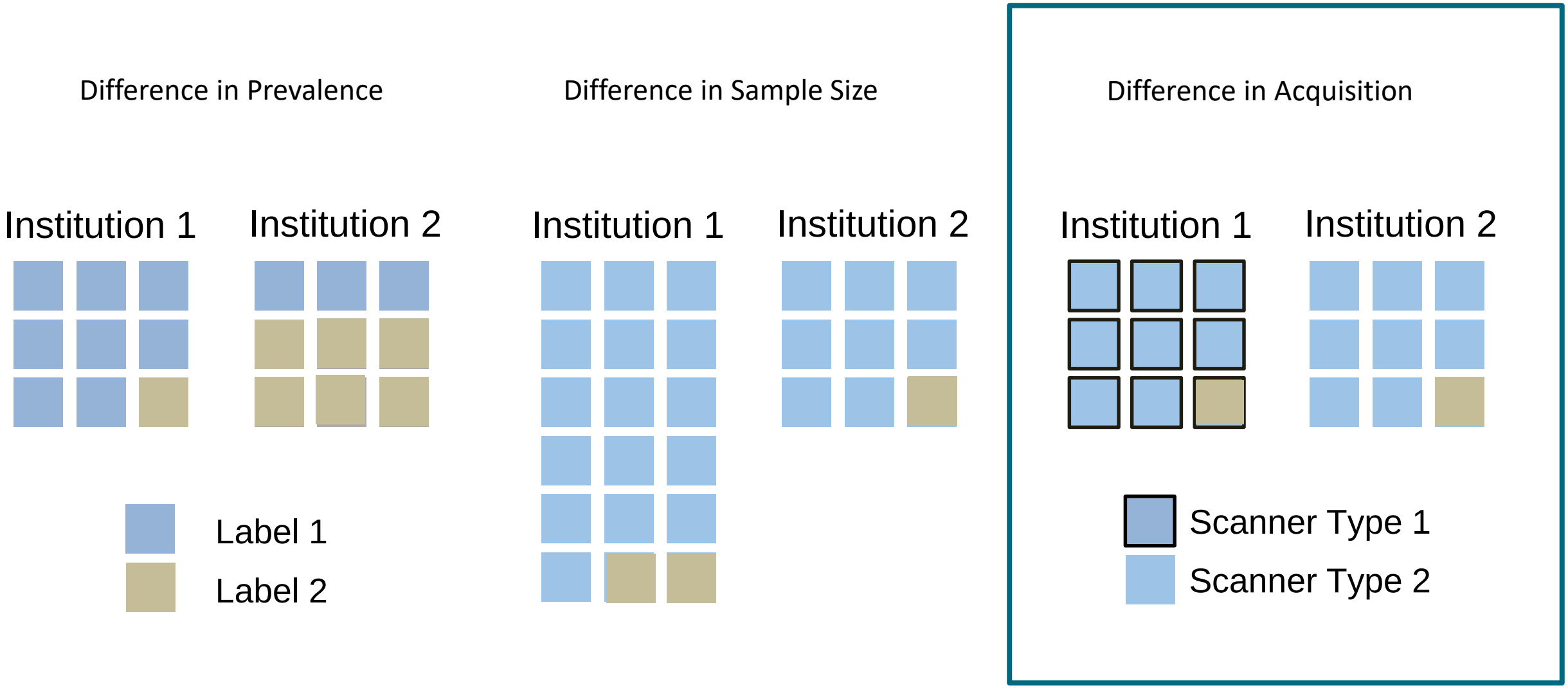
Previous DL works for breast density assessment have only focused on performing well on a single institution with a single digital mammography system.

Poor generalizability across different institutions owing to variability in patient demographics, disease prevalence, and imaging acquisition.

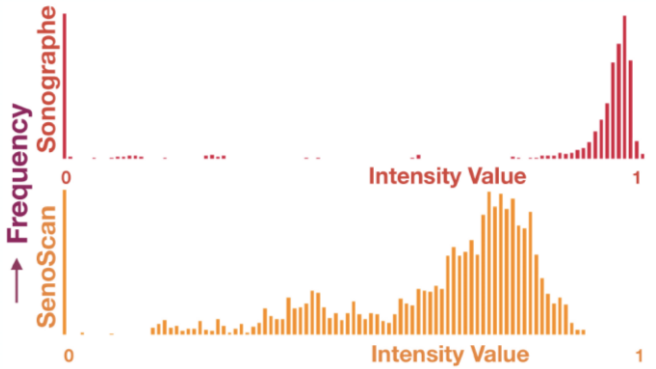


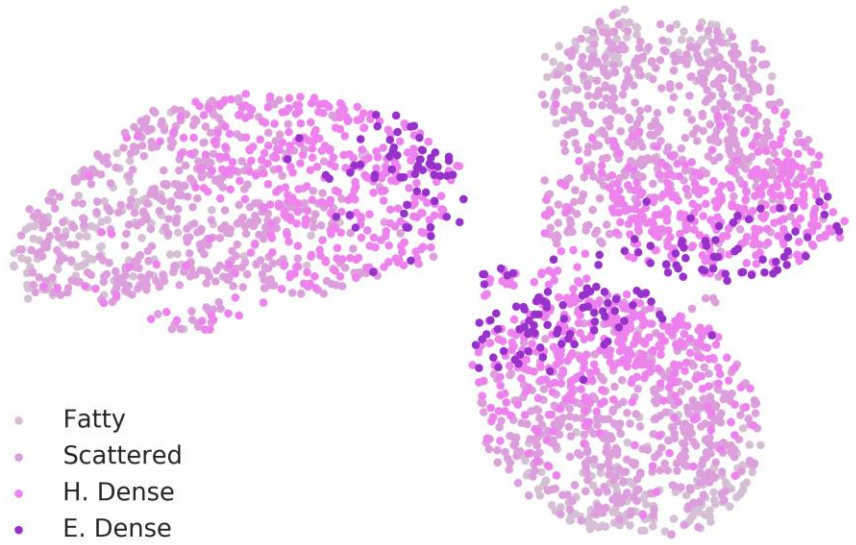
Chang et al, JACR, 2020

Dataset heterogeneity

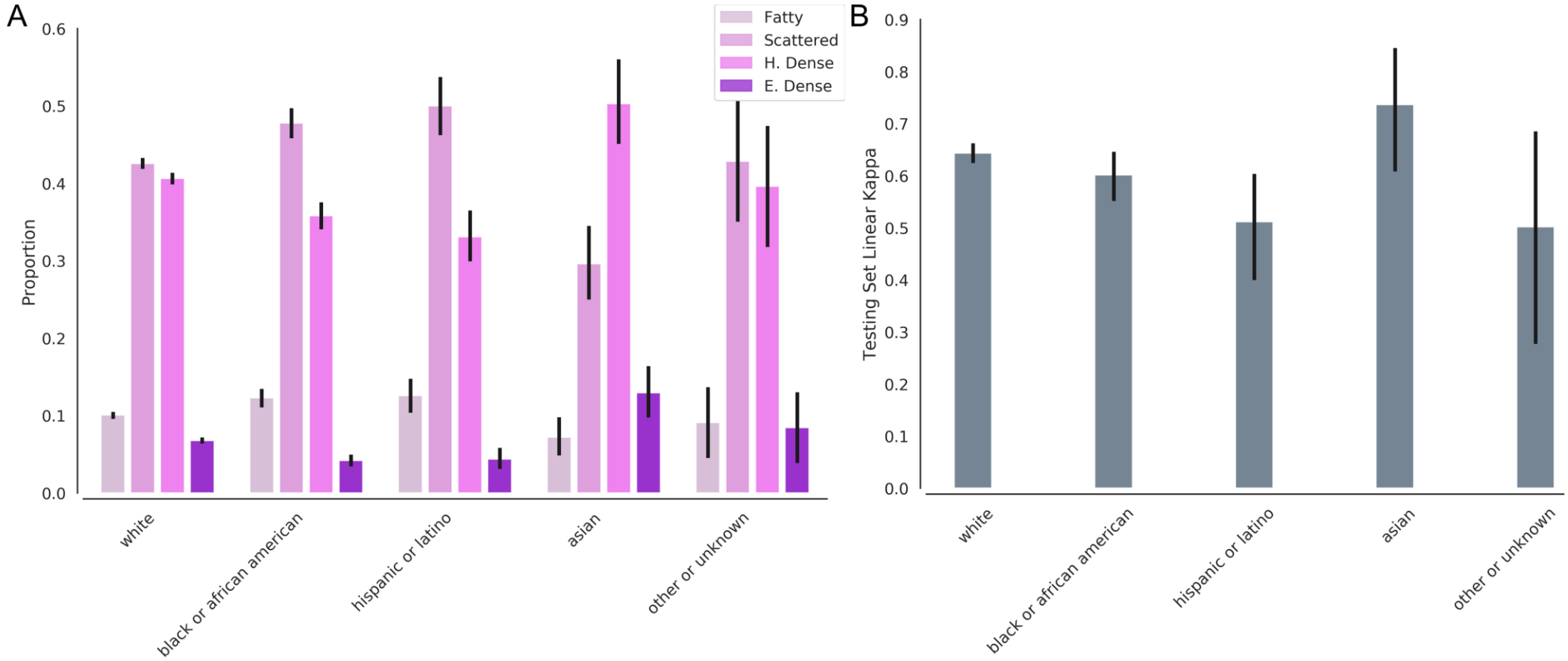


Chang et al, JACR, 2020





Chang et al, JACR, 2020



What is Explainability/Interpretability?

interpretable machine learning algorithm can be described as one in which the link between the features used by the machine learning system and the prediction itself can be understood by a human

Some simple models such as linear or logistic regression and shallow decision trees are more readily interpretable

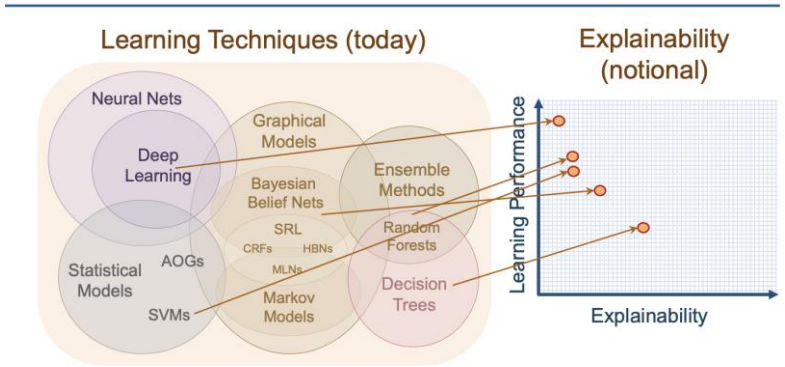
But arguably at the cost of predictive capability (?)

(Supervised) Deep learning based approaches have good performance but underlying reasoning is not easily accessible. (“Black boxes”)

Their acceptance may be limited by the inability to explain decisions and actions to users

Explainable AI (XAI) may help build trust in AI

May shed light on the underlying “reason” for the network’s



https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Reyes, et al, Radiology-AI, 2020

What are popular methods in the radiology literature?

Grad-CAM

XRAI

Smooth Grad

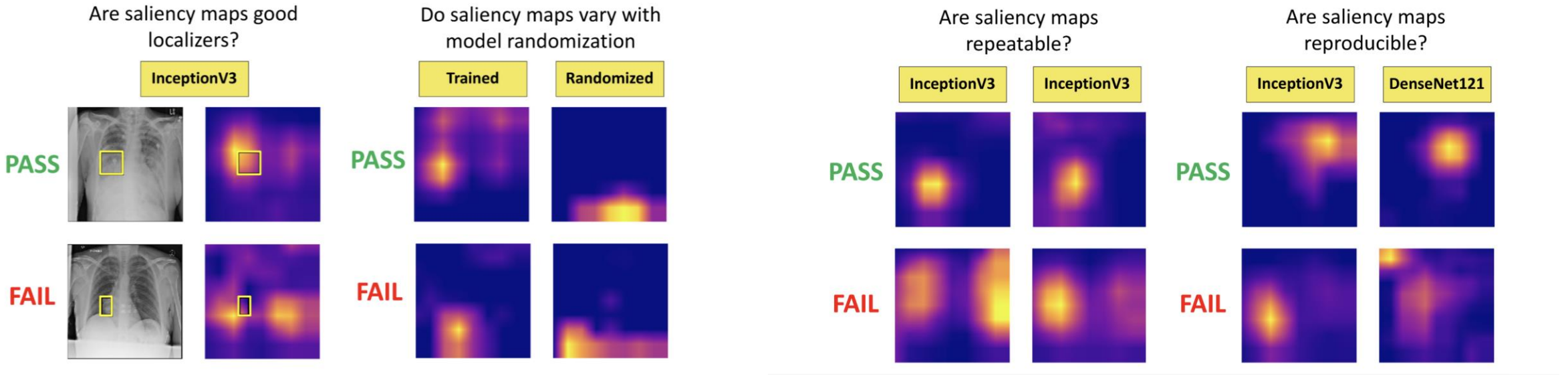
Integrated Grad

LIME

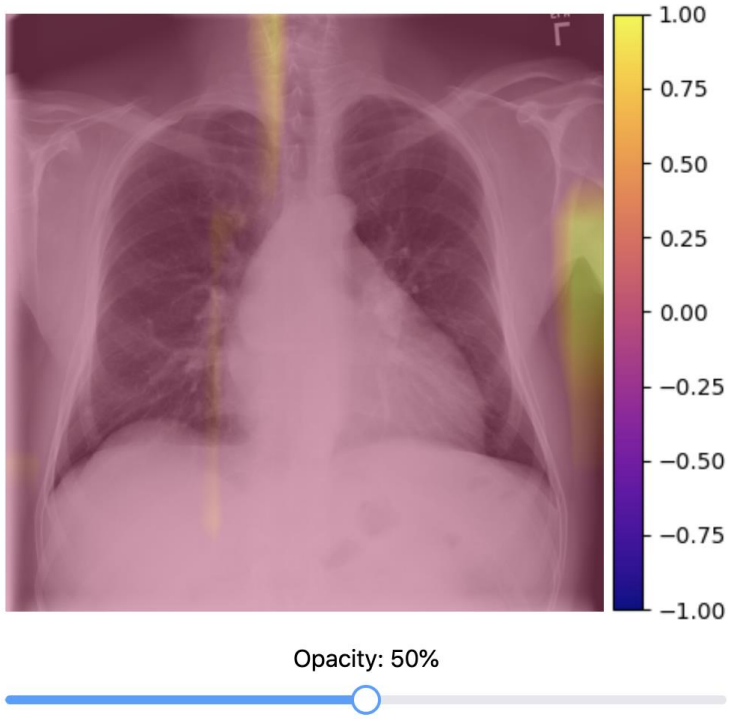
Occlusion maps

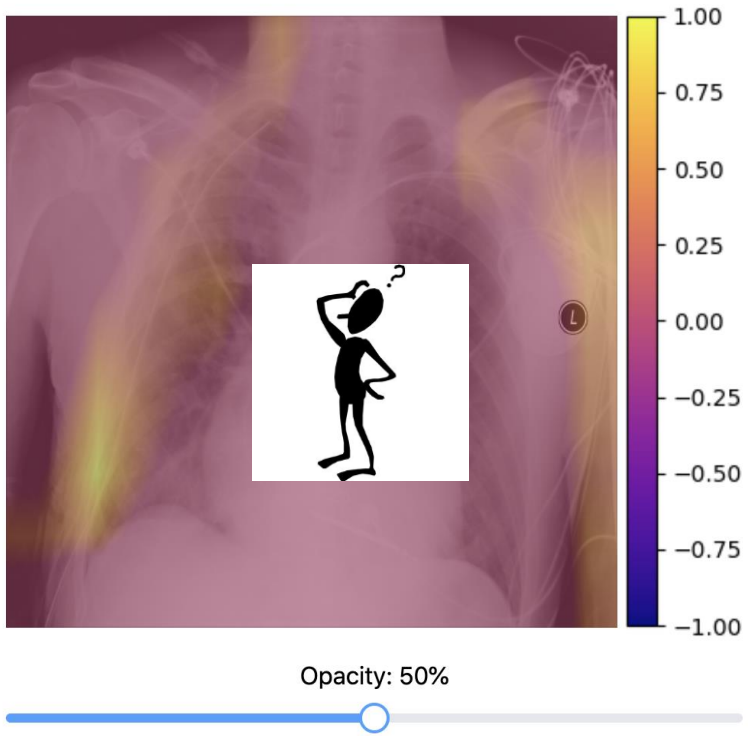
Reyes, et al, Radiology-AI, 2020

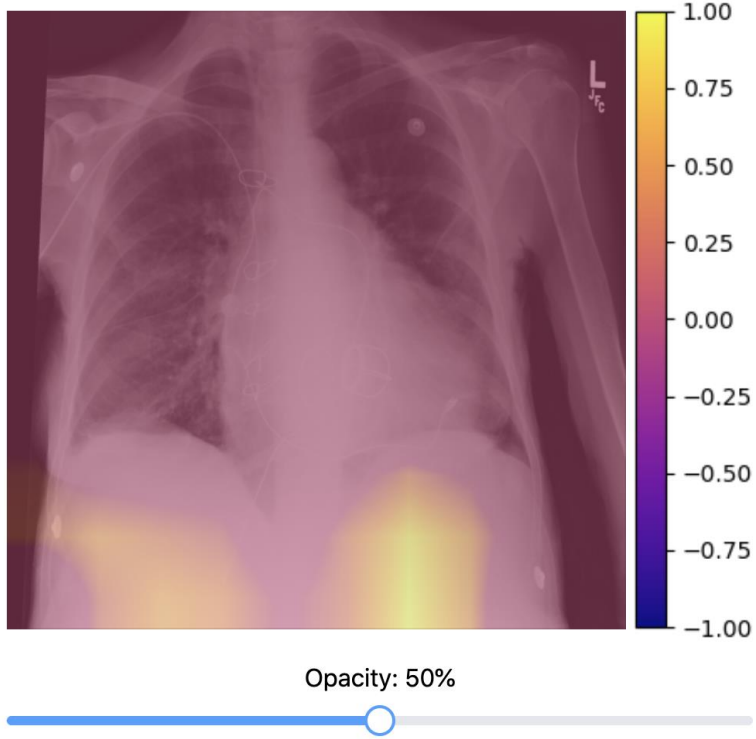
Framework to evaluate explainable methods



Cardiomegaly –GradCAM saliency maps

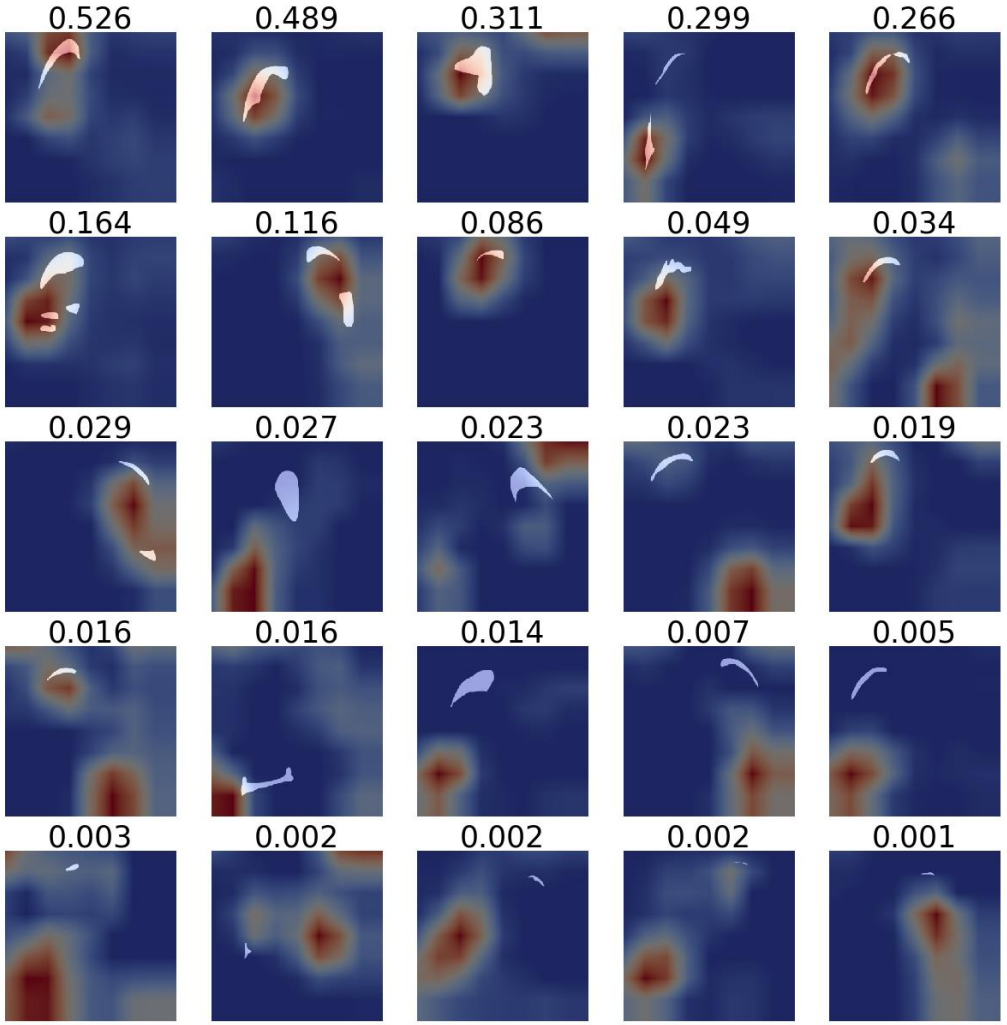






http://imimic-workshop.com/demo.html

Example maps for pneumothorax



Observations about post-hoc explainability methods

Do they highlight the area of interest? -not always!! (confirmation and publication bias)

Consider a quantitative analysis of the maps with ground truth.

Are they better than random? not always!! (confirmation and publication bias)

Compare to an ”average” or ”random” map

Are the repeatable? Reproducible? not always!!

Summary

Perform thorough quantitative evaluation of AI algorithm

- Appropriate metrics
- Sufficiently diverse dataset
- “Ground truth”
- External test set
- Algorithmic audits

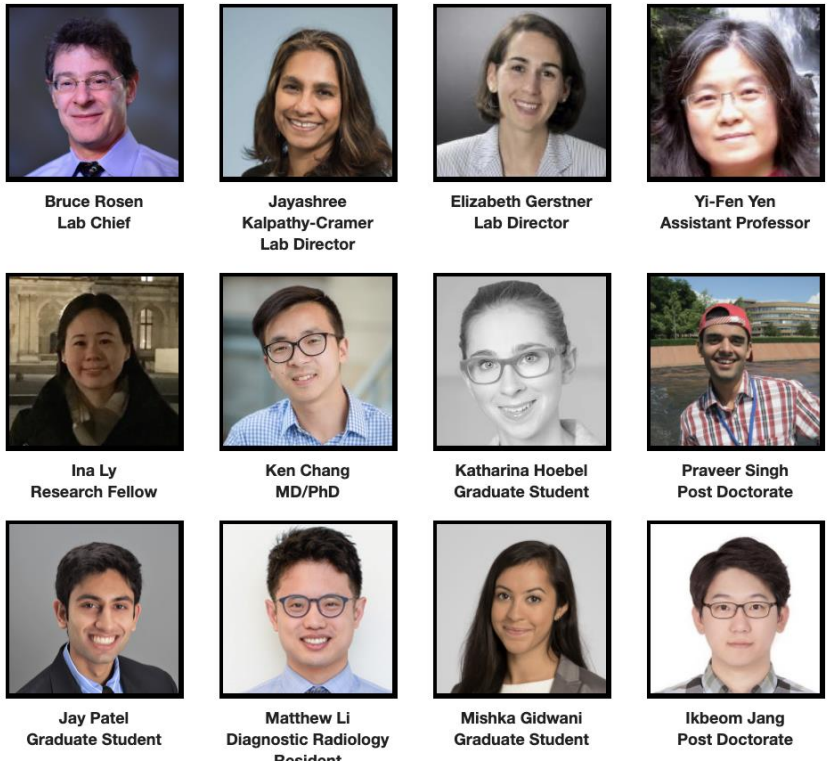
Consider if the technique is fit for the purpose

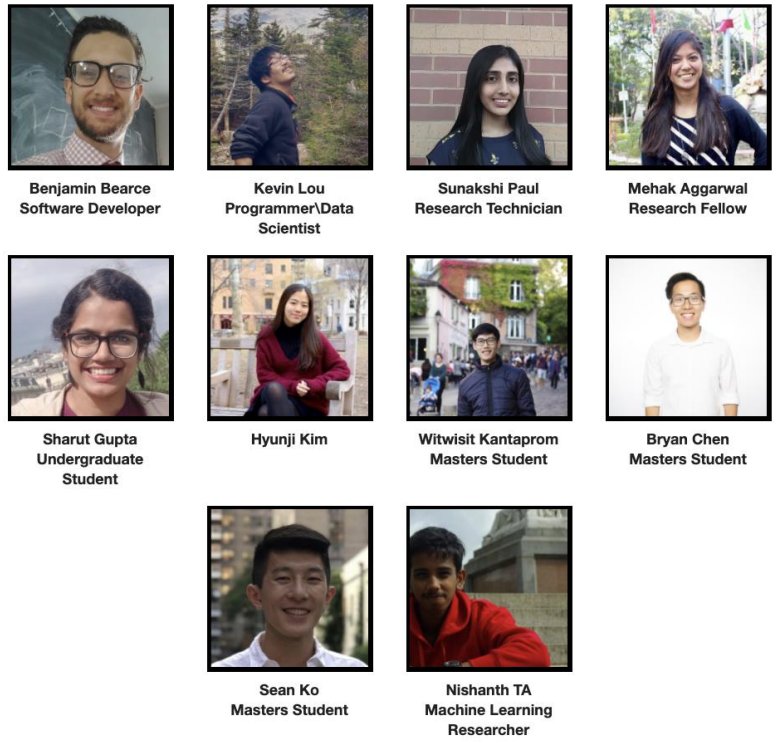
- If localization is desired, train an appropriate model (detection, segmentation)
- If using post-hoc visualization methods, confirm performance using an appropriate framework

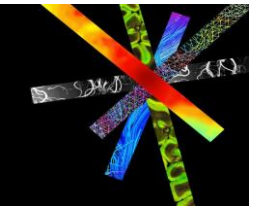
Acknowledgements

Funding & Support

- MGH/BWH Center for Clinical Data Science
- National Science Foundation
- National Institutes of Health







qtim-lab.github.io

- Grant funding from NIH, NSF
- Grant funding from Genentech Foundation
- Grant funding from GE
- Cloud credits from AWS and Microsoft