3050571 Practical Clin Data Sci

Session 16: Explainability

March 5, 2024

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Caution when using AI

AI (silently) makes mistakes and biases

But can you spot them?

Alkaissi, H. et al. Cureus 15:e35179 (2023)

> Late onset Pompe disease (LOPD) is a rare genetic disorder characterized by the deficiency of acid alpha-glucosidase (GAA), an enzyme responsible for the breakdown of glycogen in lysosomes. The accumulation of glycogen in various tissues leads to progressive muscle weakness, primarily affecting the skeletal and respiratory muscles. However, recent studies have also reported liver involvement in LOPD, which is thought to occur as a result of the accumulation of glycogen in liver cells.

- There was no prior publication about liver involvement with LOPD
- However, the authors of this paper have an unpublished manuscript showing a link between liver disease and LOPD
  - *Did ChatGPT just synthesized new knowledge? Or simply hallucinated?*

Huge gap between development and actual use

Healthcare, Law, Regulation, and Policy, Machine Learning

“Flying in the Dark”: Hospital AI Tools Aren’t Well Documented

MODEL REPORTING GUIDELINES	EPIC MODEL BRIEFS											
	Deterioration Index	Early Detection of Sepsis	Risk of Unplanned Readmission	Risk of Patient No-Show	Pediatric Risk of Hospital Admission or ED Visit	Risk of Hospital Admission or ED Visit	Inpatient Risk of Falls	Projected Block Utilization	Remaining Length of Stay	Risk of Admission of Heart Failure	Risk of Hospital Admission or ED Visit for Asthma	Risk of Hypertension
TRIPOD	63%	63%	61%	48%	42%	61%	47%	36%	55%	48%	44%	51%
CONSORT-AI	63%	43%	63%	60%	33%	67%	53%	47%	47%	49%	42%	51%
SPIRIT-AI	61%	55%	54%	54%	38%	61%	44%	49%	51%	41%	39%	46%
Trust and Value	46%	33%	39%	50%	29%	42%	38%	46%	46%	25%	33%	46%
ML Test Score	27%	15%	33%	24%	9%	33%	15%	6%	18%	12%	9%	15%

Evaluation of sepsis diagnosis AI

Results We identified 27 697 patients who had 38 455 hospitalizations (21 904 women [57%]; median age, 56 years [interquartile range, 35-69 years]) meeting inclusion criteria, of whom sepsis occurred in 2552 (7%). The ESM had a hospitalization-level area under the receiver operating characteristic curve of 0.63 (95% CI, 0.62-0.64). The ESM identified 183 of 2552 patients with sepsis (7%) who did not receive timely administration of antibiotics, highlighting the low sensitivity of the ESM in comparison with contemporary clinical practice. The ESM also did not identify 1709 patients with sepsis (67%) despite generating alerts for an ESM score of 6 or higher for 6971 of all 38 455 hospitalized patients (18%), thus creating a large burden of alert fatigue.

- AUC of 0.63 in practice
- Missed 67% of sepsis

(Un)expected behaviors

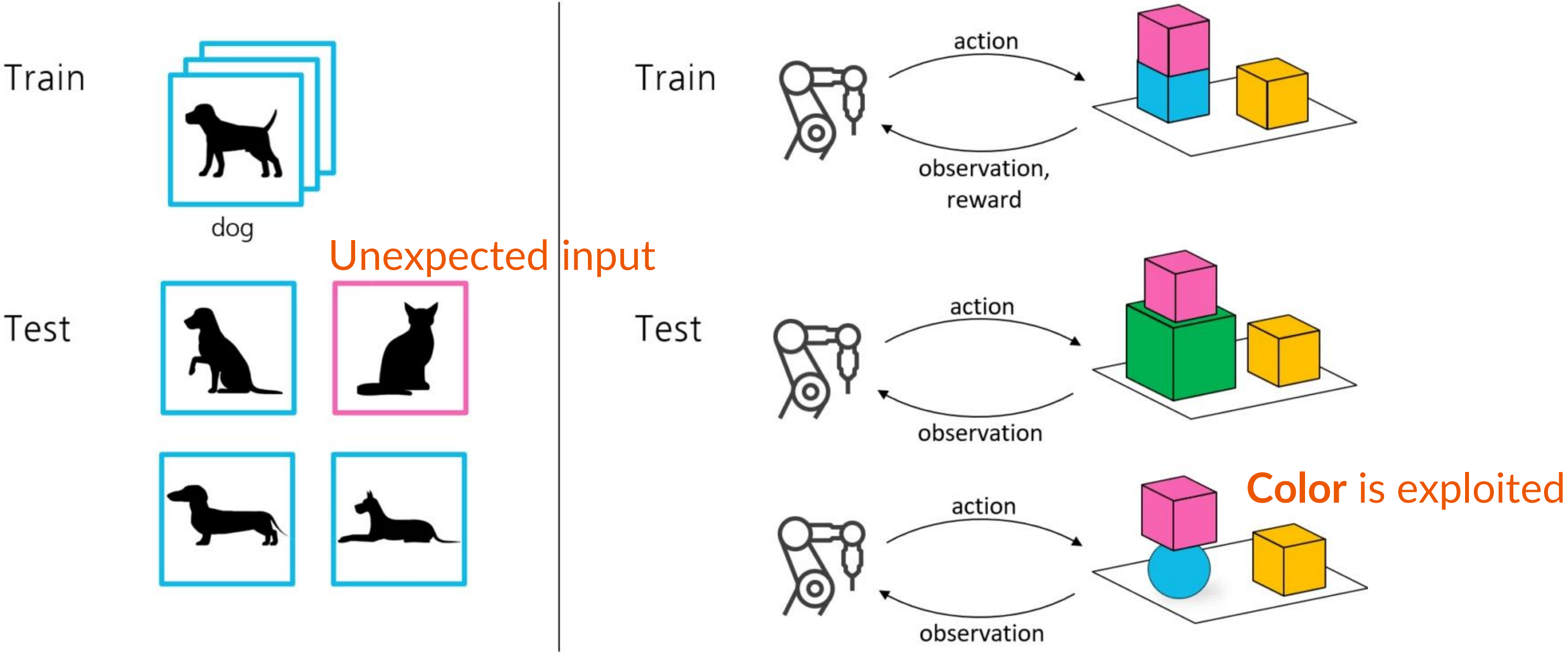
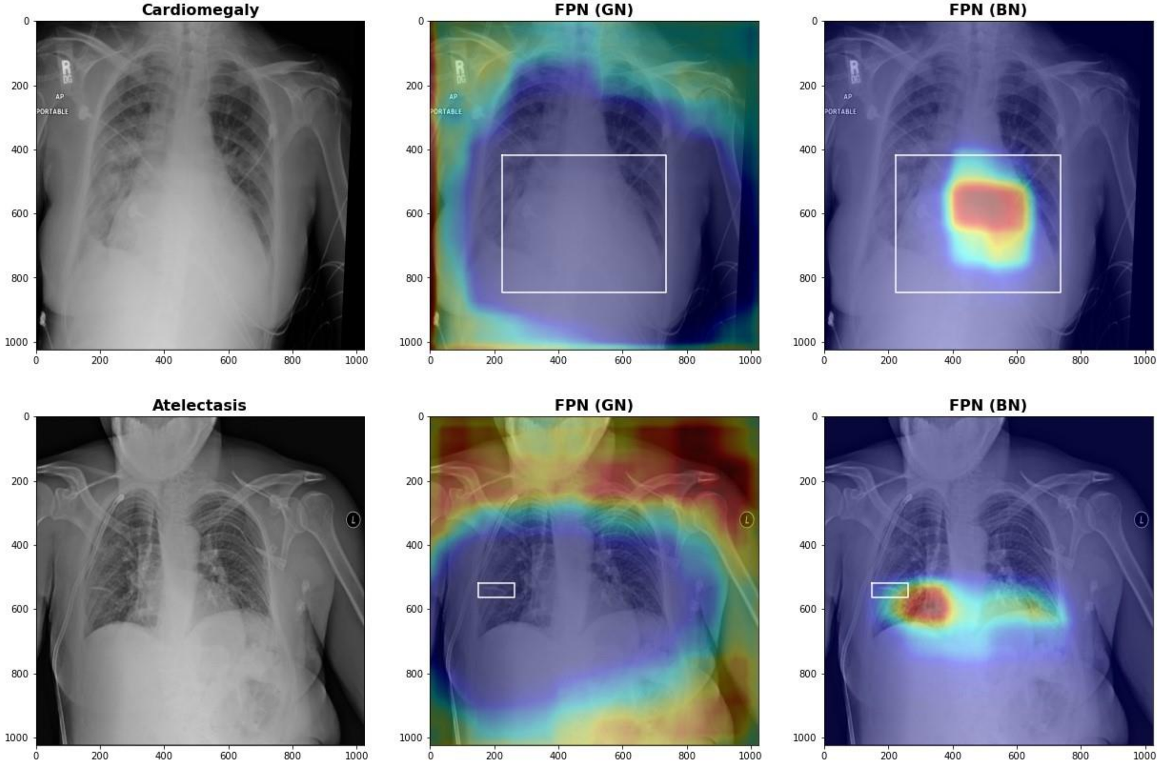


Image from https://safe-intelligence.fraunhofer.de/en/articles/out-of-distribution-detection-for-reinforcement-learning

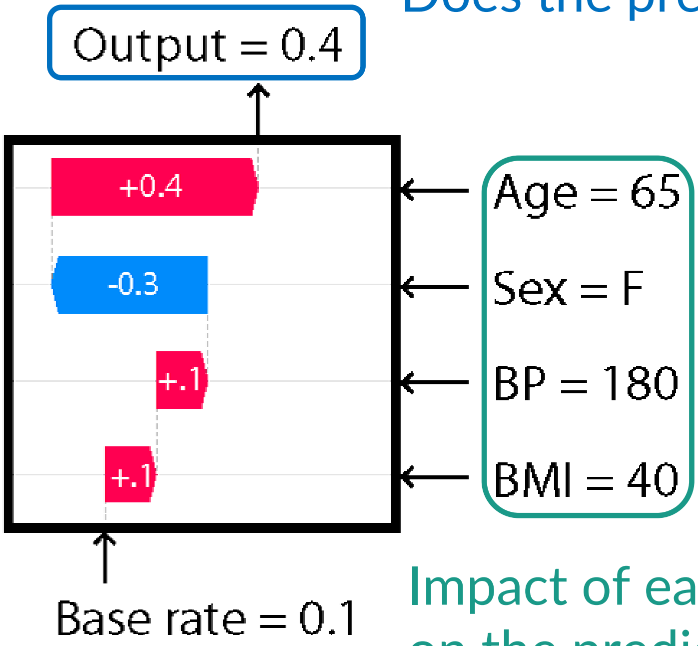
Correct prediction is not enough



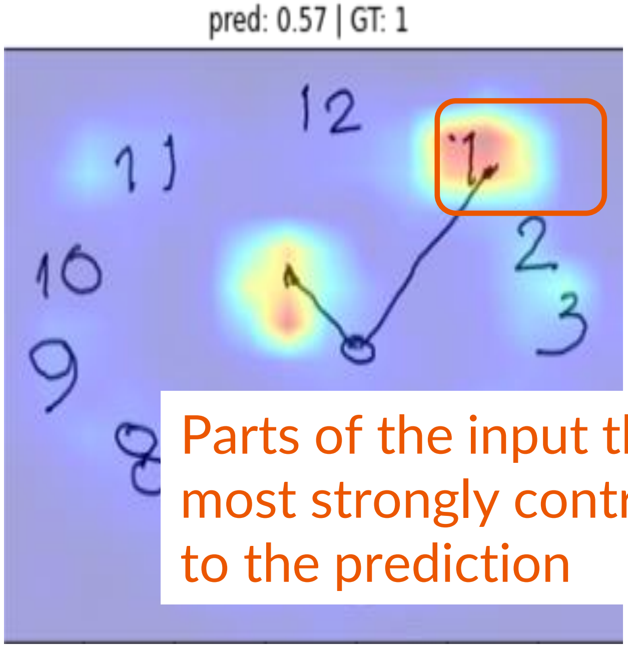
- Two models with the **same classification performance**
- Both images were **correctly classified**
- But the **explanations** complete differ

Explainability is key

Does the predicted confidence match your expectation?



Impact of each input on the prediction

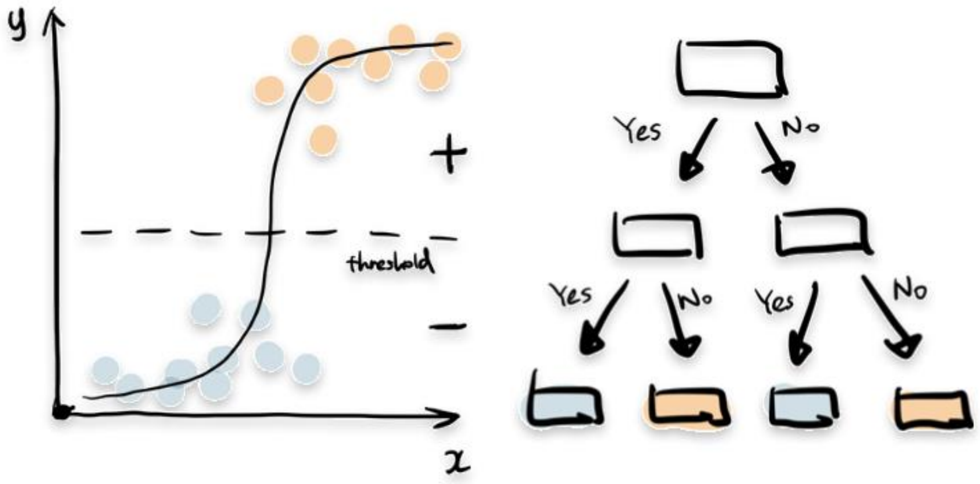


Parts of the input that most strongly contribute to the prediction

Inherently interpretable models

Linear and tree models

Model

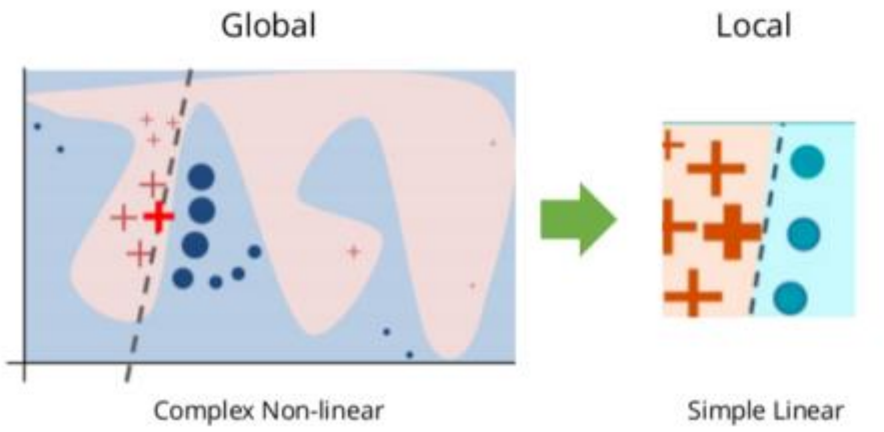


https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501

- Model decisions are immediately understandable
  - Examine coefficients
  - Trace the decision in a tree

- We can use this to approximate a more complex model

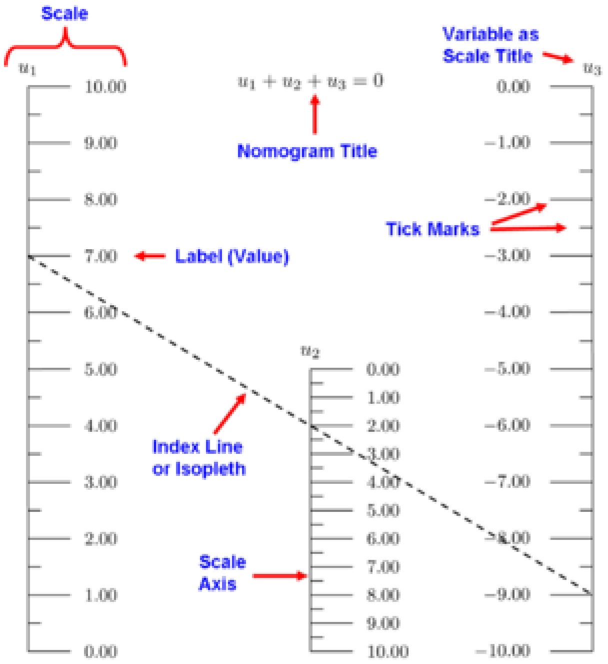
LIME



https://c3.ai/glossary/data-science/lime-local-interpretable-model-agnostic-explanations/

- Approximate decision boundary surround a data point
- Slightly perturb original data and get predictions from the model

Nomogram



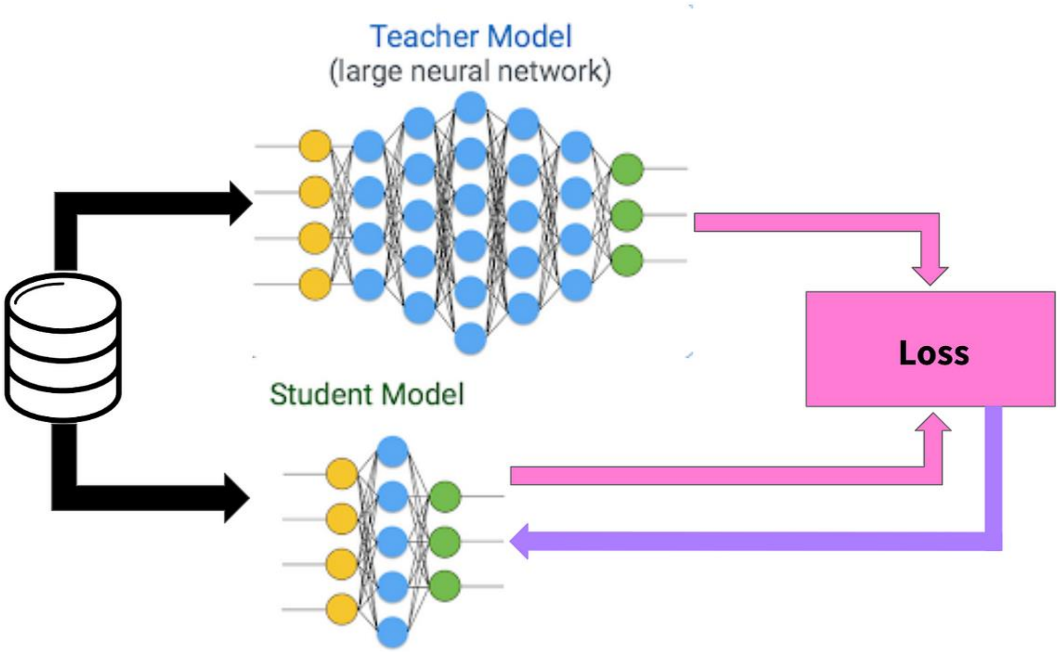
https://en.wikipedia.org/wiki/Nomogram

- After a complex model (C) is developed, a simpler model (S) can be fitted on the input data and the prediction made by C

- Nomogram can be fitted to mimic a random forest model

- Easy to use on site and interpretable

[Related] Distillation



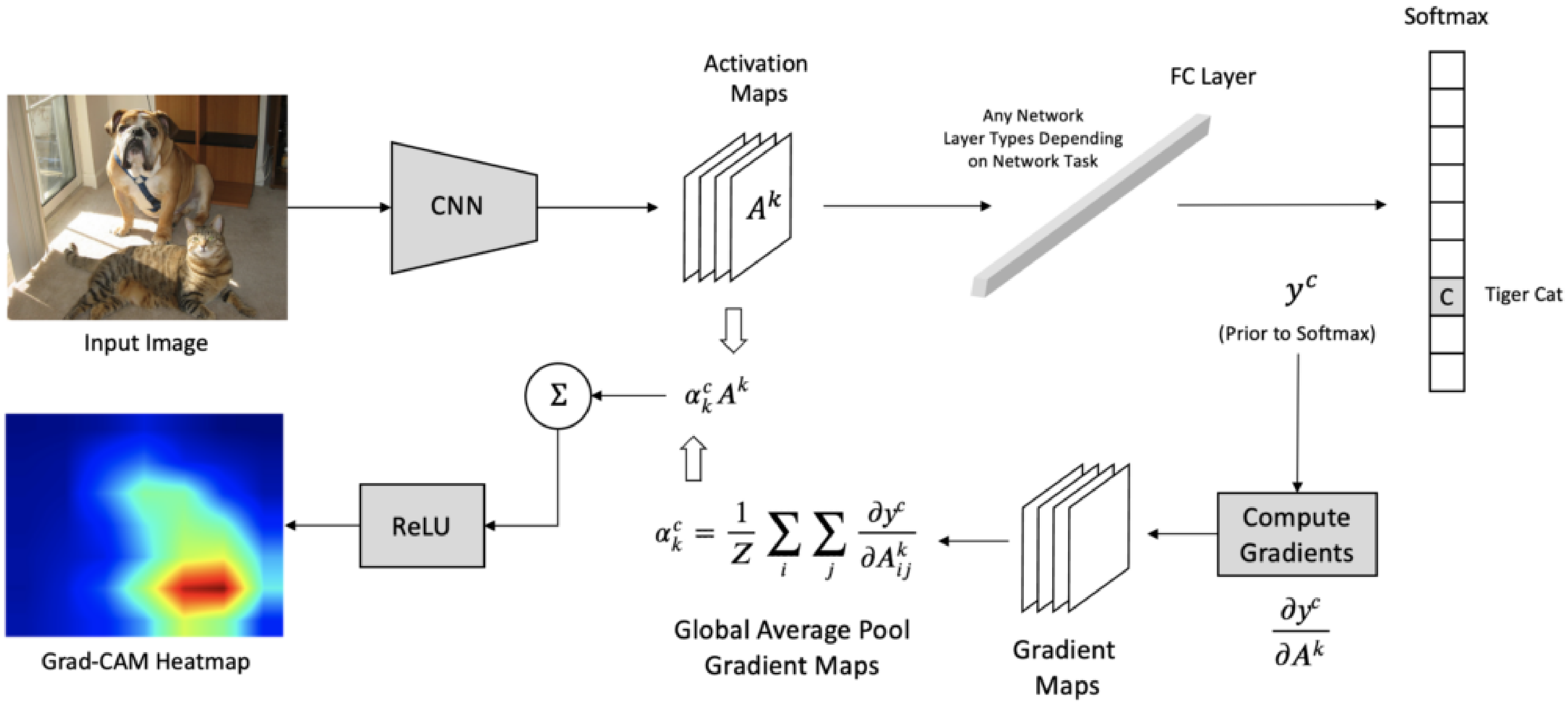
- Larger model encodes knowledge from the training data

- 0/1 label into probability and embedding

https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

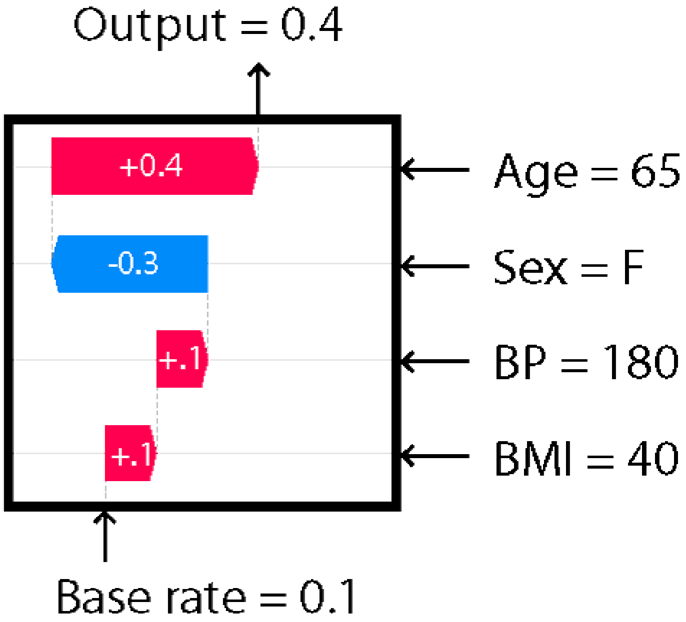
Explainability techniques

Gradient tracing & saliency map

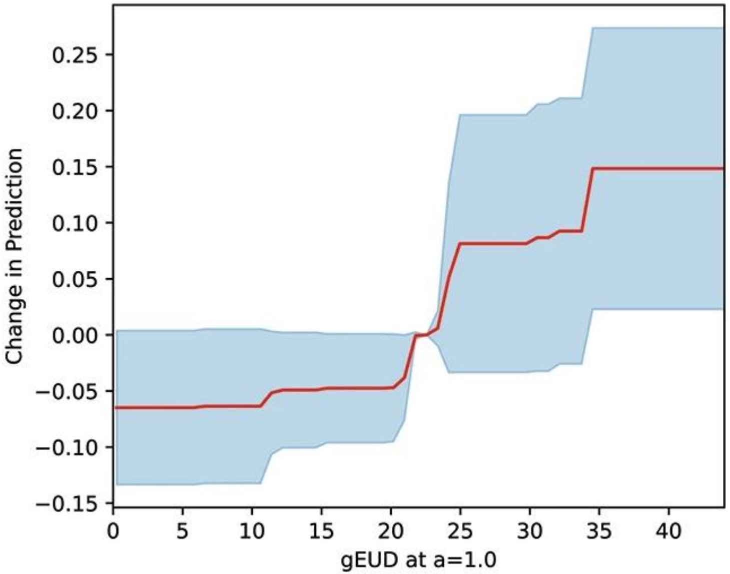


https://learnopencv.com/intro-to-gradcam/

Input value perturbation

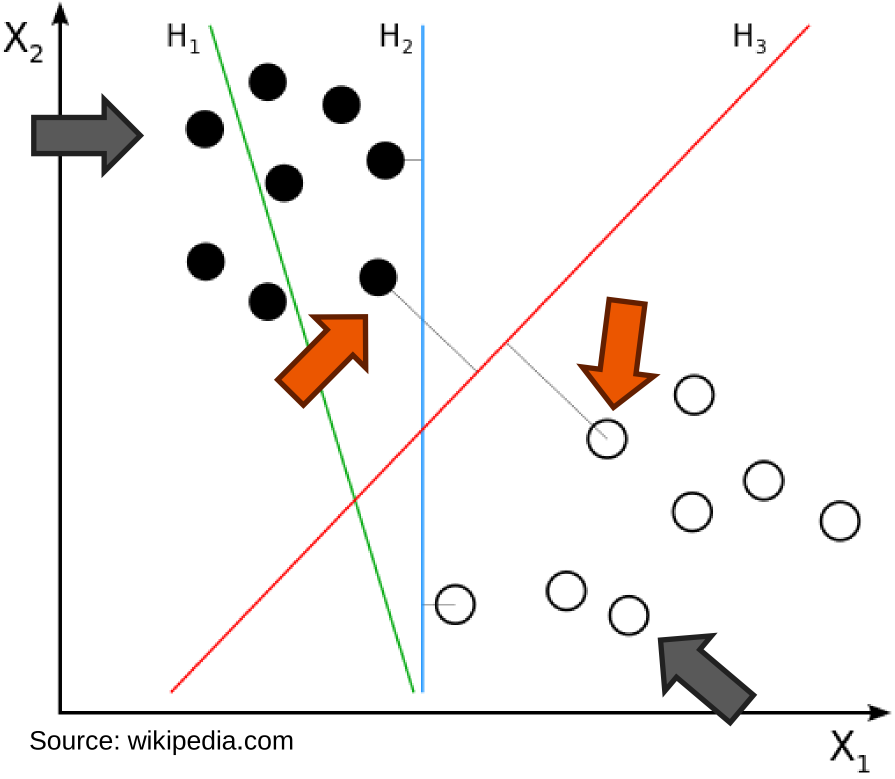


Prayongrat, A. *et al*. Radiation Oncology 2022



- Calculate prediction changes when target feature values are perturbed

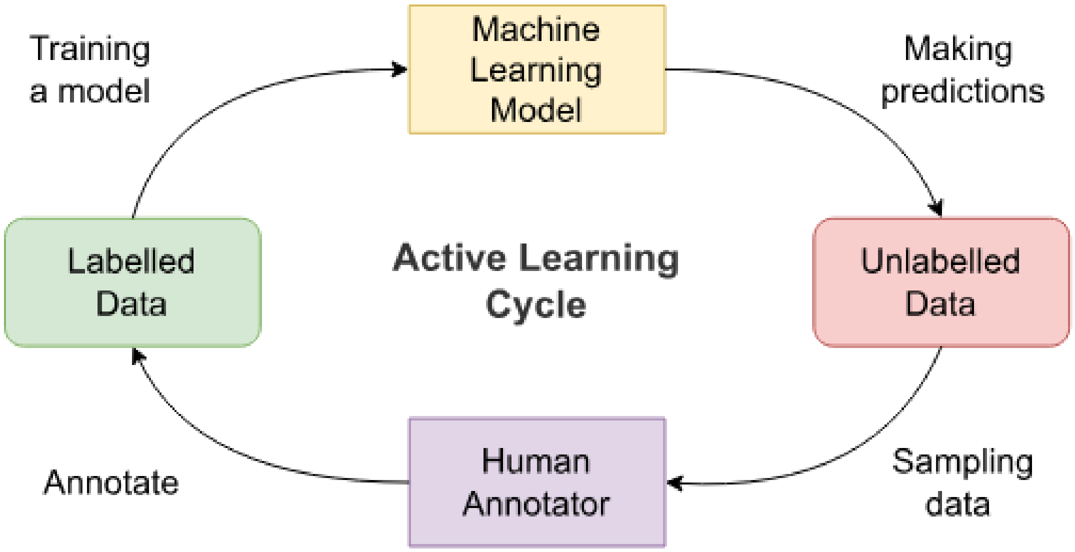
Impact of each training sample



Source: wikipedia.com

- Some define the decision boundary
- Some are located among many other similar data points
- Drive additional data collection

[Related] Active learning



northlineschool.org

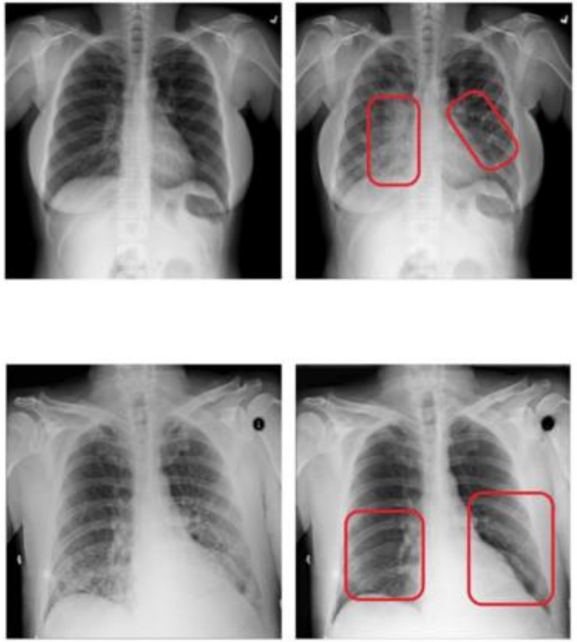
- Use model predictions to determine which data to collect and which experiments to perform

Error analysis

- Identify systematic errors
- Bias in data, spurious association, mismatch between training and test set, or model limitation
- Hard examples
- Drive additional data acquisition and model improvement

Counterfactual argument

“Thinking about what did not happen but could have happened”

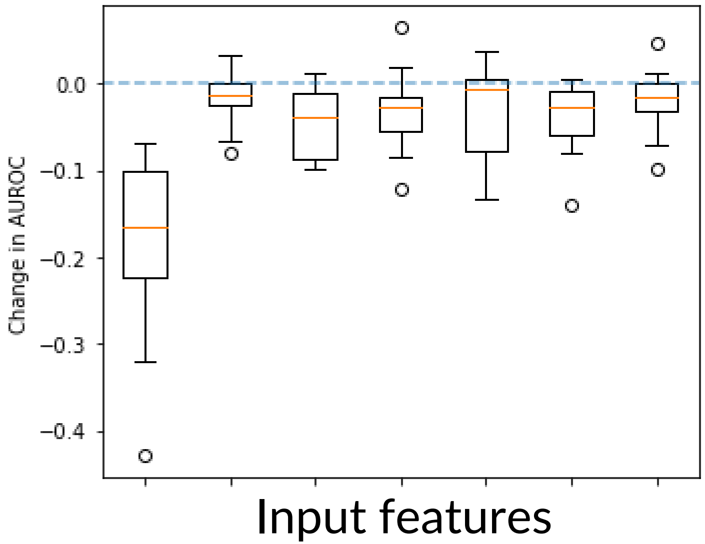


Mertes, S. *et al*. Frontiers in AI 2022

- Alter regions identified to be important to the predicted class / value

- Observe whether the **prediction changes**

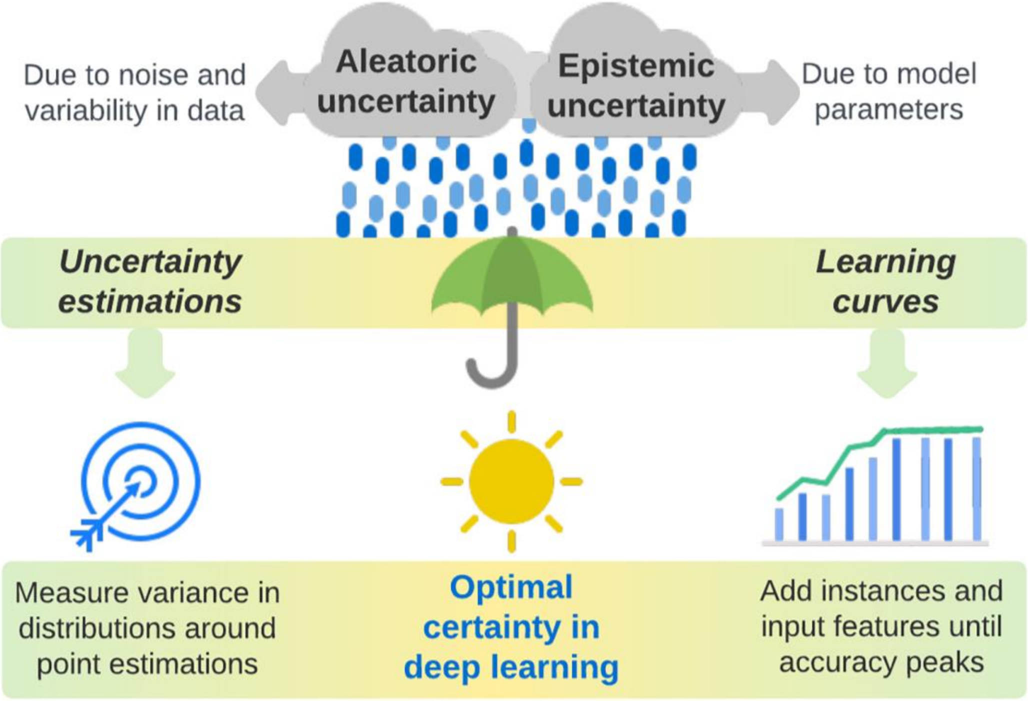
Change in performance after dropping a feature



- Compare performance with and without each input feature
- Big drop = important

Beyond accuracy

Uncertainty

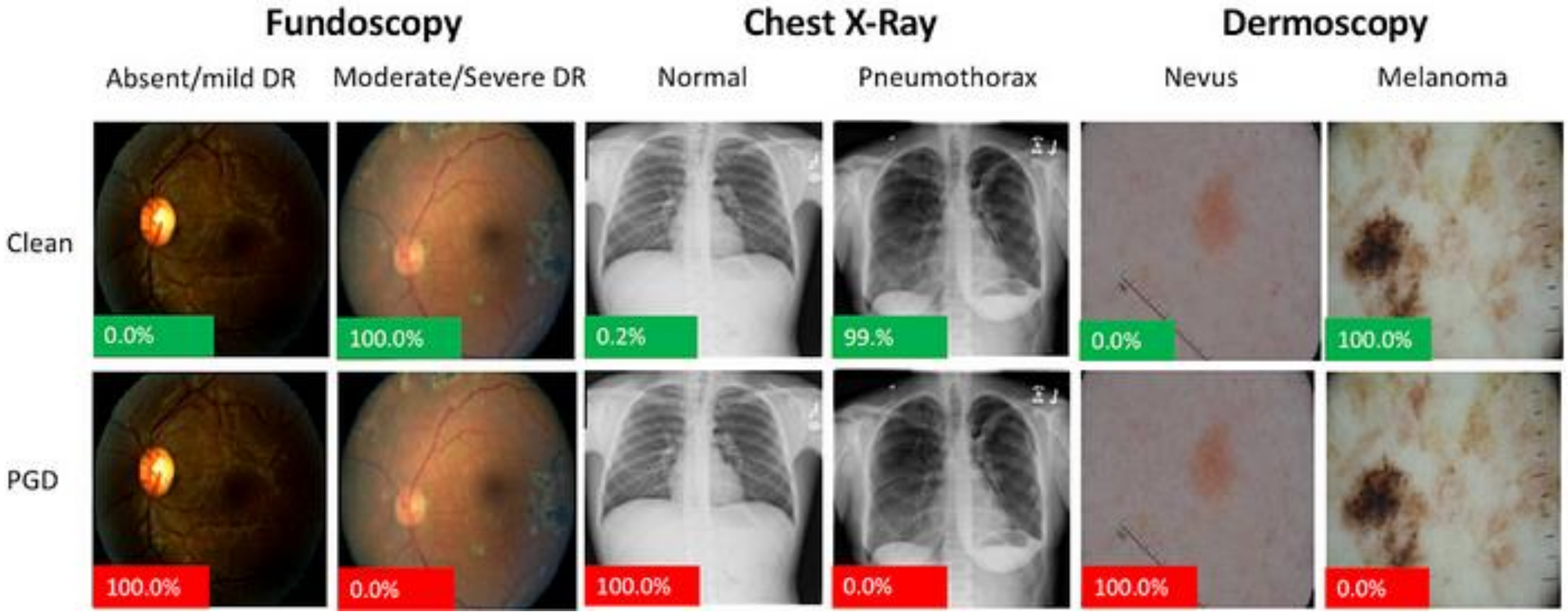


Loftus, T.J. *et al*. PLOS Digital Health 2022

- Innate variability in data
  - Drug 3D structure
  - May be predictable
- Variability in trained model
  - Bootstrapping
  - Impact of data size
  - Ensemble approach

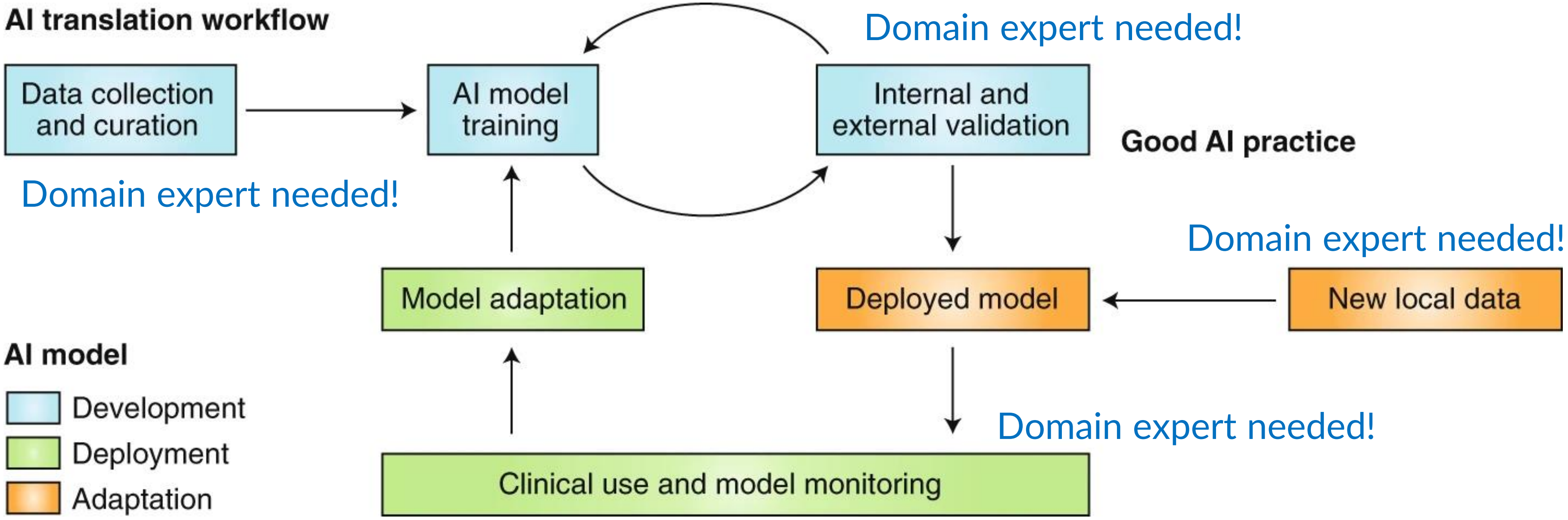
Stability

https://medium.com/pymedix/machine-learning-adversarial-attacks-its-all-fun-and-games-until-someone-gets-hurt-3f726113134d



- Small input perturbation should not drastically change the prediction

Sustainability and other concerns



- Anyone can feed data through ML library
- Only domain experts can spot model weakness and find data to fix it

Any questions?

See you on March 6th