Machine learning principles and communications for material scientists

Lecture 4: ML experimental design

September 26, 2022

Sira Sriswasdi, PhD

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

Core of ML experimental design

- Clear objectives
  - What to predict? Why? Is ML the best answer? Human-in-the-loop

- Sufficient data collection
  - Aware of annotation/labeling cost
  - Beware of unintended biases

- Appropriate performance metrics
  - Match the objective and use case

- Be realistic + acknowledge limitation

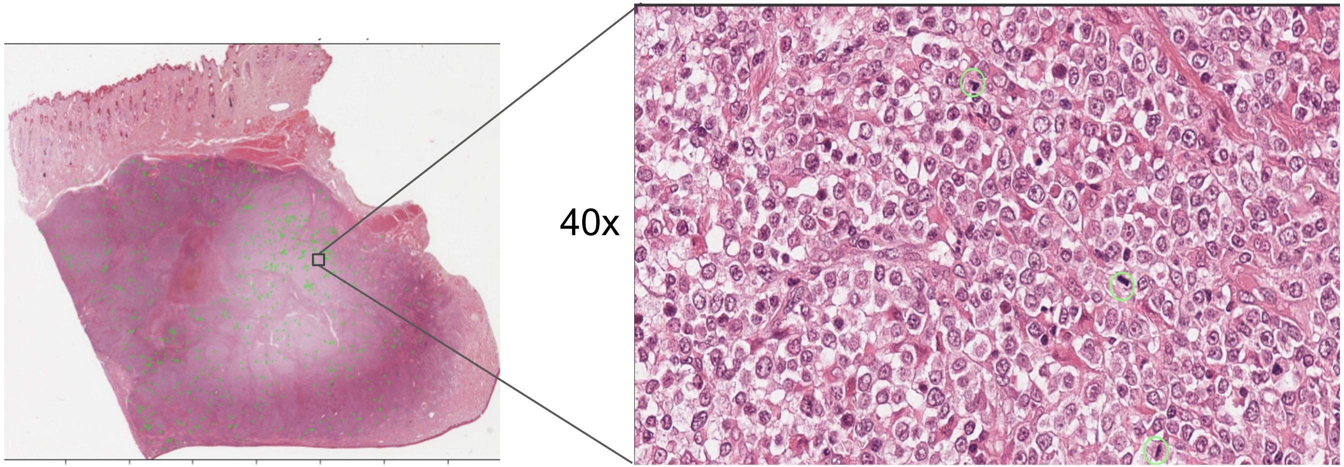
Objectives

Predict or not predict

Source: Lunit CXR webpage

- What is the pain point?
- Predictive model vs good visualization
  - Knowledge replaces sample size
- Level of performance required for the task
  - Can imperfect model still be useful?
- AI-only vs human-in-the-loop

Focus on the pain point

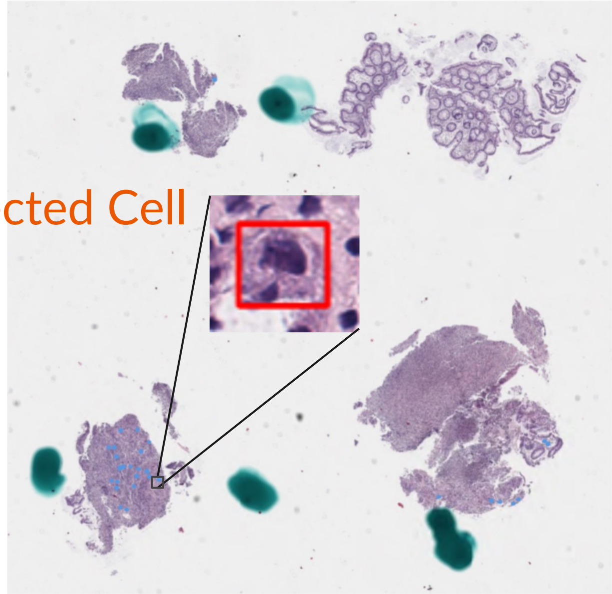


Whole Slide Image (WSI)
(150,000 x 150,000 pixels)

Individual mitotic figures

- Pain point = inspecting the whole image
- Imperfect object detector is good enough for estimating mitotic density

Human in the loop



Cell-level performance

F1	Precision	Recall
17.78	10.00	79.69

Low precision AI due to small training data

- Provide top 10 cells with highest p(infected) in each whole slide image
- 100% diagnosis when considering only proposed cells

Feasibility

- **Theory**: Is there relationship between input and output?
- **Literature**:
    - Has something similar been done?
    - What were the data and models used?
- **Pilot**:
    - Small-scale data
    - Simple benchmark: Linear → How much can I fit the training data?
    - Leave-one-out cross-validation
        - n – 1 samples for training, 1 for validation

Data collection

Data-centric approach

Conventional model-centric approach:

AI = Code + Data

(algorithm/model)

Work on this

Data-centric approach:

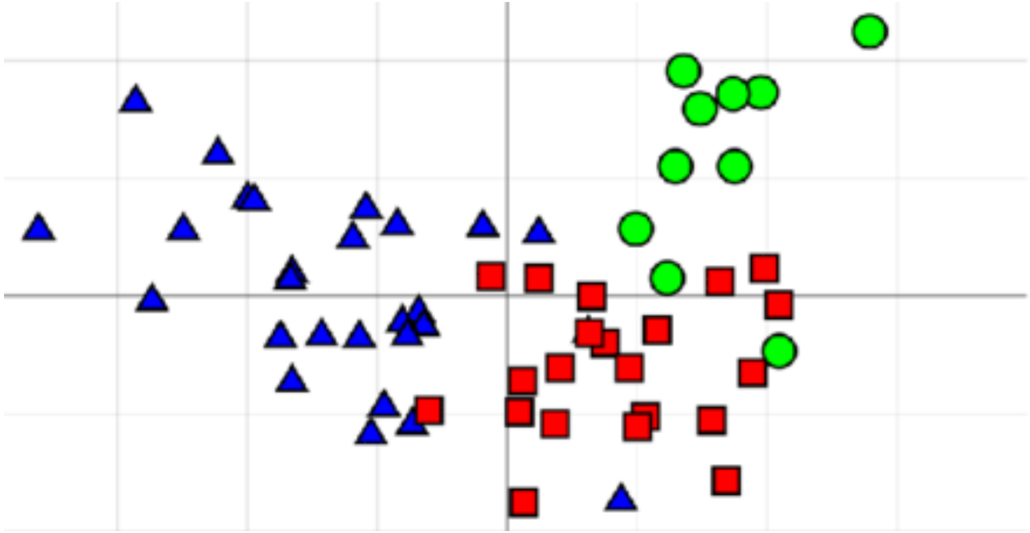
AI = Code + Data

(algorithm/model)

Work on this

https://landing.ai/tips-for-a-data-centric-ai-approach/

Sample size

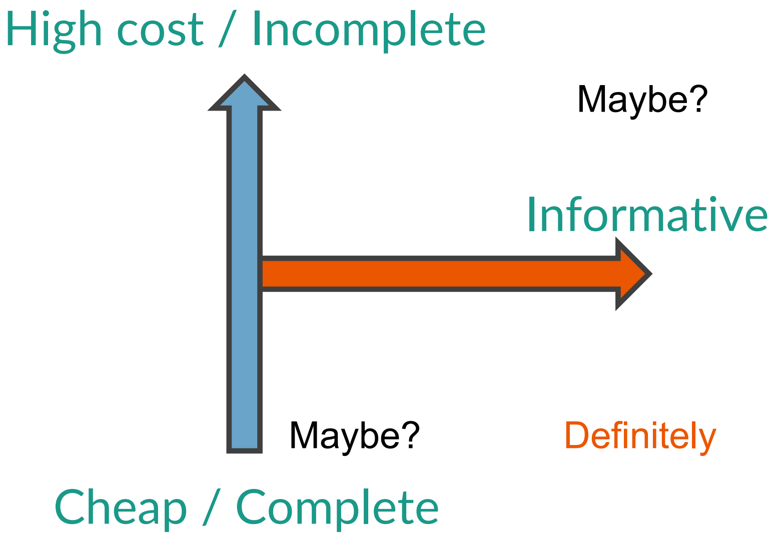


Stocchero, M. et al. Sci. Rep. 9(1):6151 (2019)

- **Objective**: Capture every mode & capture variance
  - Scale with data dimension
- **Best source**: Literature review

Cost-driven consideration

- Input features
  - Cheap and informative: Yes
  - Costly but informative: Maybe
  - Cheap but uninformative as input
    - Error analysis
    - Is it costly to re-collect?
- Annotation / label
  - Cost vs quality
- Unlabeled data can be useful



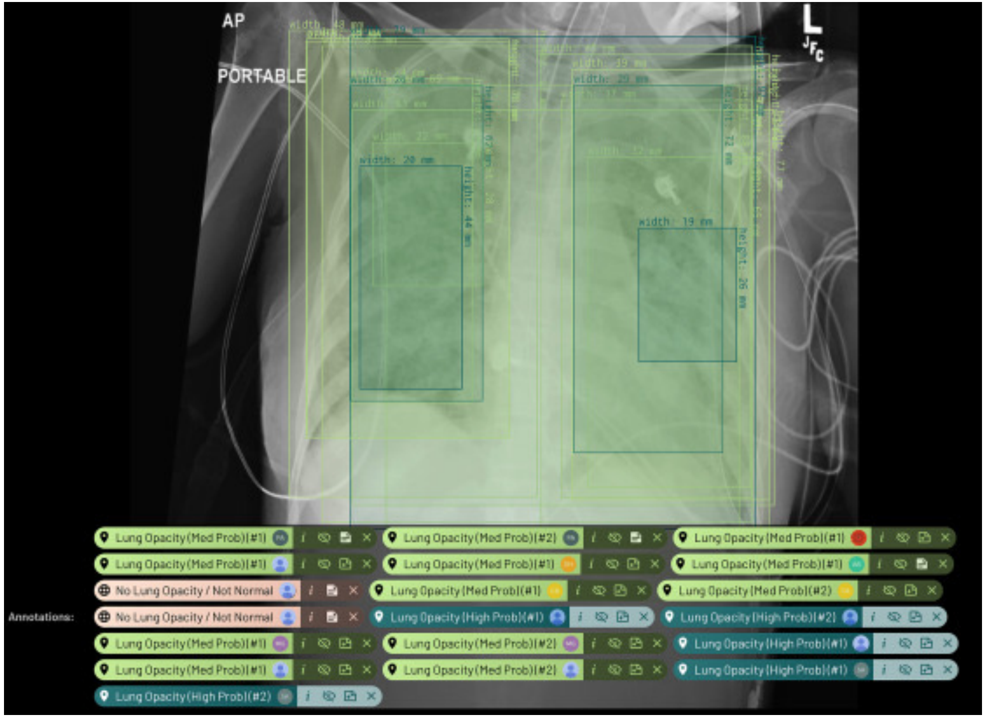
Objective-driven consideration

- Proof-of-concept
  - Show feasibility / achievable performance of model family

- Internal use
  - Get a working model
  - Internal validation (same measurement device, data collection process)

- Public deployment
  - External validation: Performance guarantee
  - Calibration: Interpretable output probability

Manual labeling for chest x-ray



Shih, G. et al. Radiol Artif Intell 1(1):e180041 (2019)

- 30,000 CXR images
  - From >200,000 total
- 18 radiologists
- 6 months

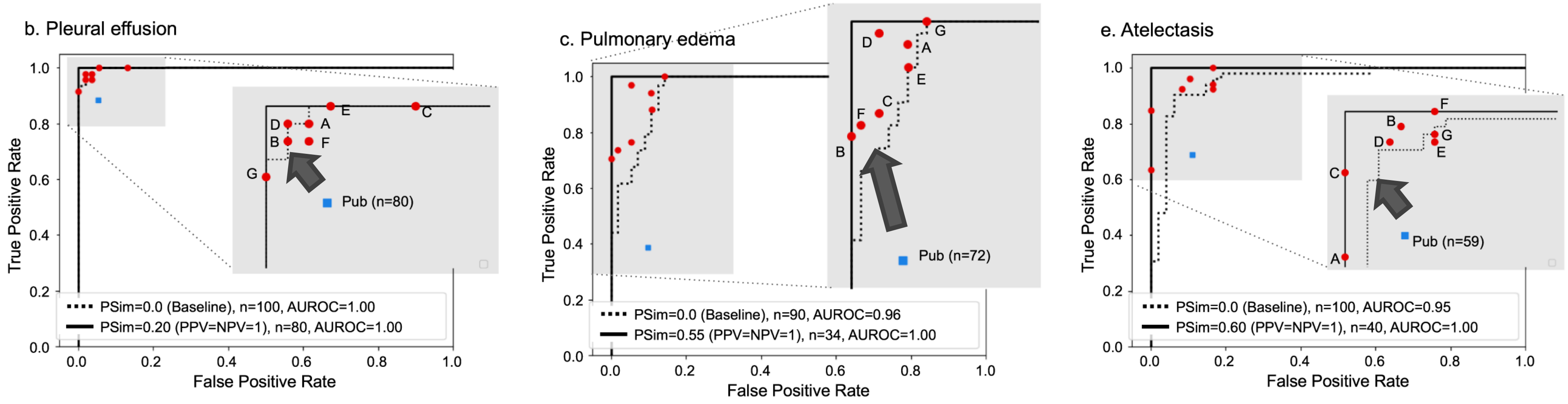
Automated labeling for chest x-ray

Report Segment and Labels	Reasoning
...two views of chest demonstrate *cariomegaly* with no focal consolidation... Cardiomegaly CheXpert: Blank ✗ T-auto: Positive ✓	T-auto, in contrast to CheXpert, recognizes conditions with misspellings in the report like "cariomegaly" in place of "cardiomegaly".
...*consistent with acute and/or chronic pulmonary edema*.... Edema CheXpert: Positive ✓ T-auto: Uncertain ✗	T-auto incorrectly detects uncertainty in the edema label, likely from the "and/or"; CheXpert correctly classifies this example as positive.
...*Normal heart size, mediastinal and hilar contours are unchanged in appearance*... Enlarged Cardiomediastinum CheXpert: Negative ✗ T-auto: Negative ✗ CheXbert: Uncertain ✓	T-auto and CheXpert both incorrectly label this example as negative for enlarged cardiomediastinum; CheXbert correctly classifies it as uncertain, likely recognizing that "unchanged" is associated with uncertainty of the condition. The condition cannot be labeled positive or negative without more information.

Smit, A. et al. https://arxiv.org/pdf/2004.09167

- **Radiologist's written report**: keywords + positive / negative / uncertain

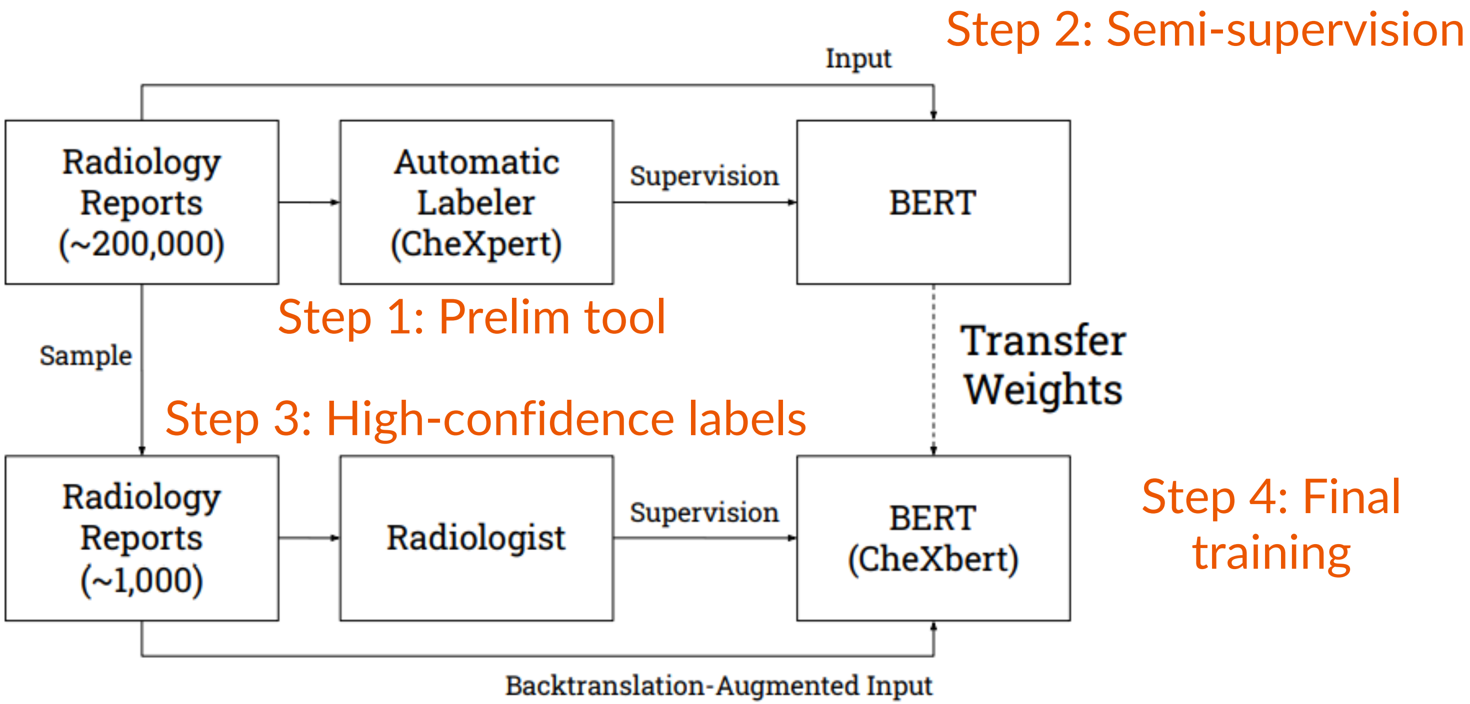
Impact of labeling quality



Kim, D. et al. Nature Comm 13:1867 (2022)

- Automated label extractions were previously used as ground truth
- Significant performance improvement by label cleaning

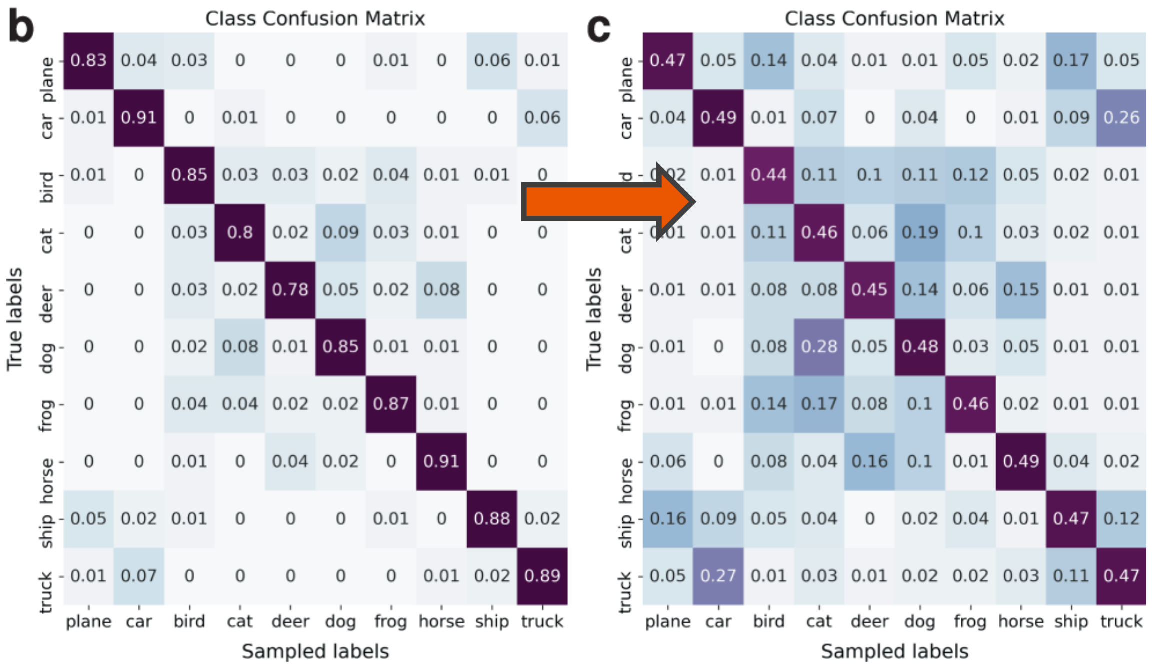
Iterative labeling process



Smit, A. et al. https://arxiv.org/pdf/2004.09167

- Automated labeling reduces time spent on easy samples

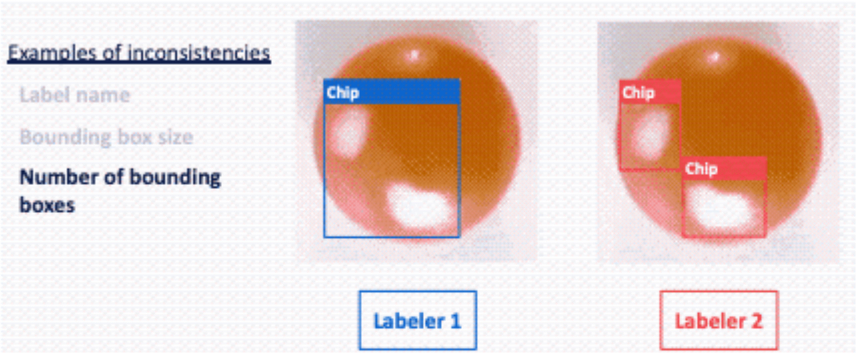
Beware of hard samples

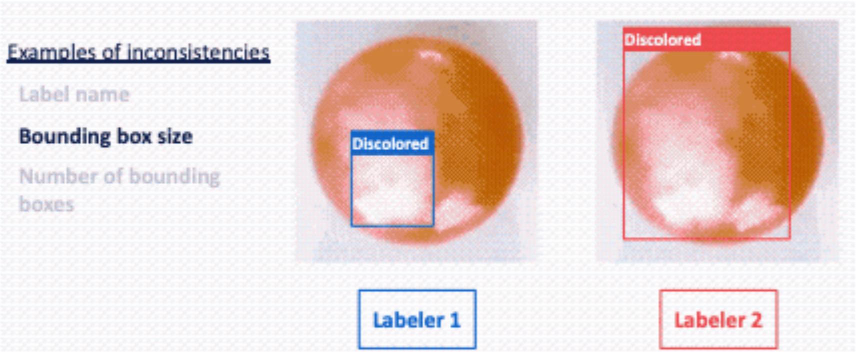


Bernhardt, M. et al. Nature Comm 13:1161 (2022)

- 3x-5x increase in label error in hard CIFAR10 samples (depending on task)

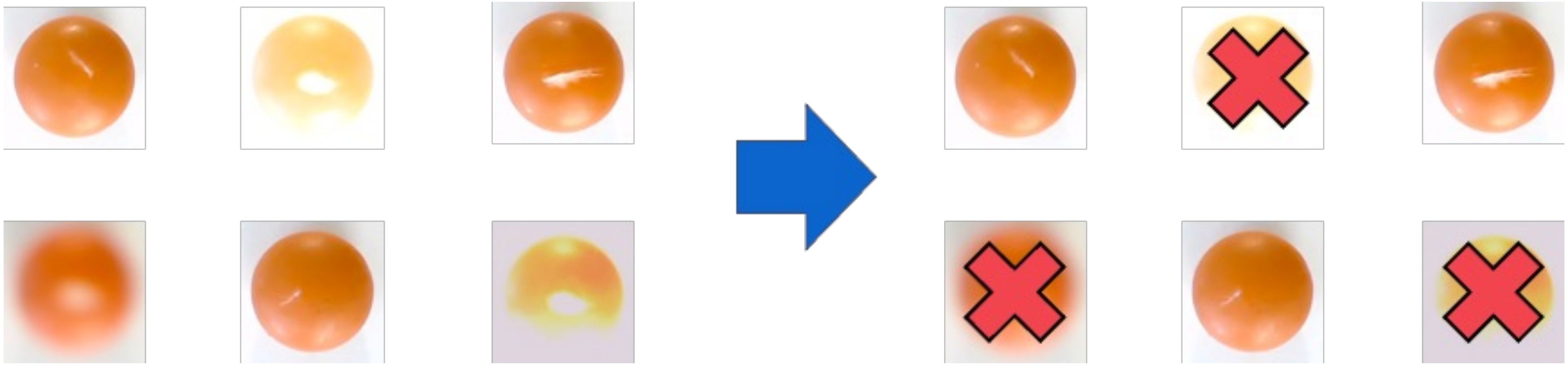
Ensure labeling standard





https://landing.ai/tips-for-a-data-centric-ai-approach/

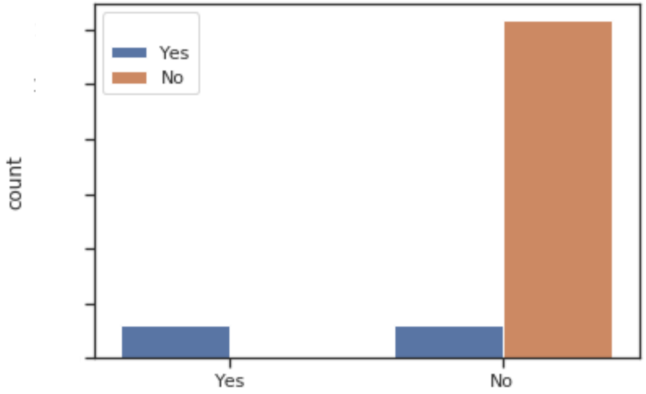
More is not always better



https://landing.ai/tips-for-a-data-centric-ai-approach/

- Bad, noisy, out-of-distribution data can fool any model

Beware of unintended biases

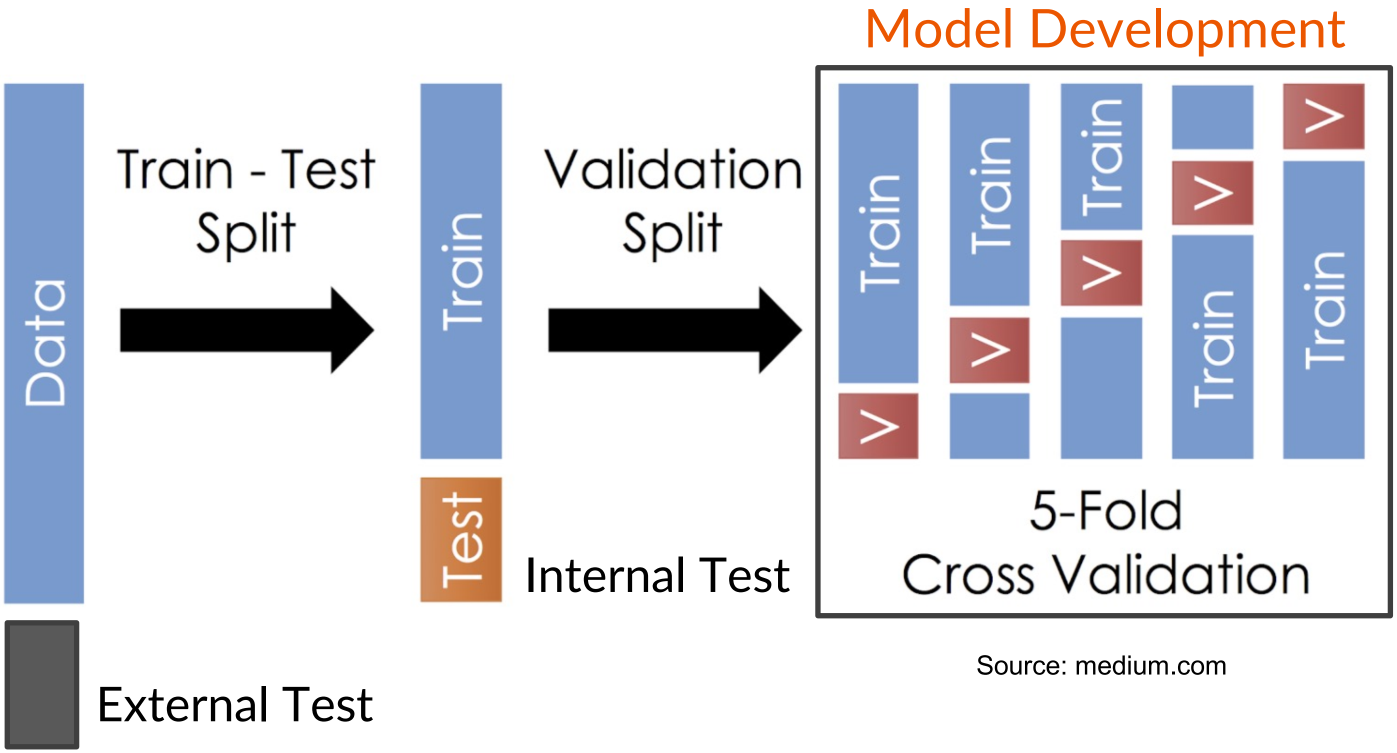


A binary feature

- Repeated measurements of the same samples do not fully count
- Ensure enough samples with different feature values

Validation scheme

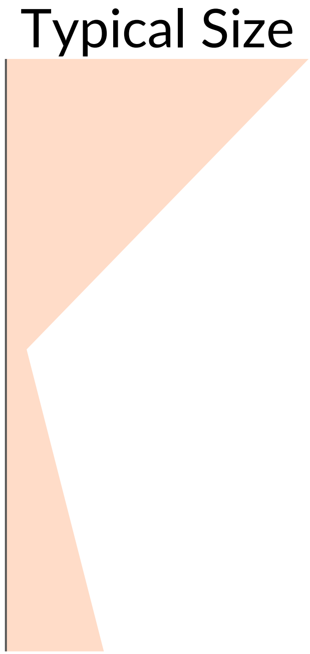
Train-Val-Test



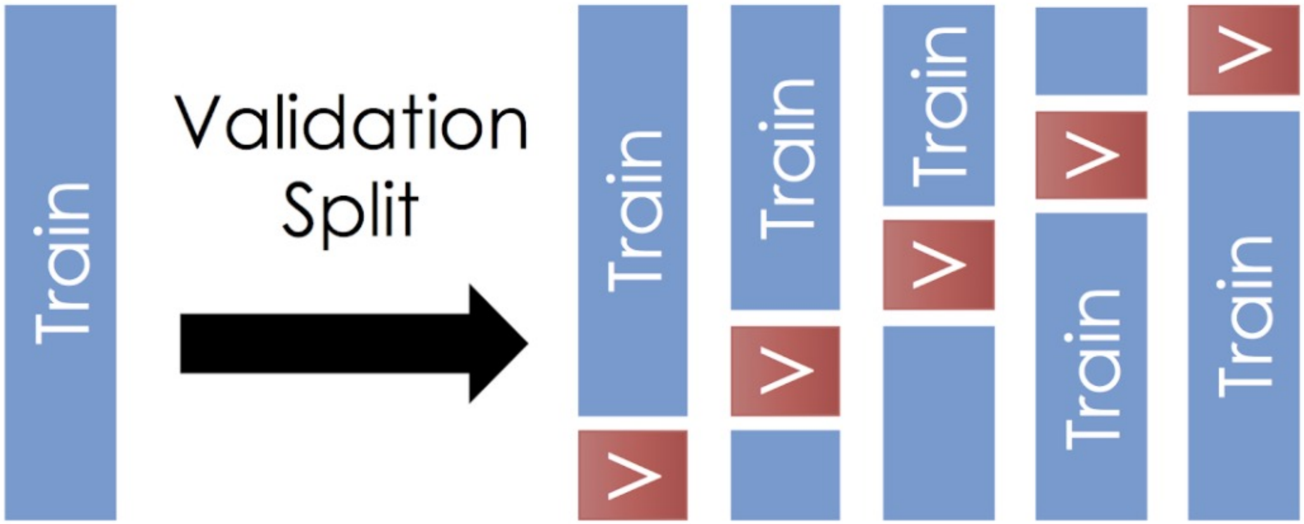
Source: medium.com

Roles of data split

- **Training:**
  - Represent data distribution
  - Find the best fit coefficients
- **Validation**: Find the best hyperparameters
- **Internal Test**: Performance evaluation
- **External Test**: Generalizability

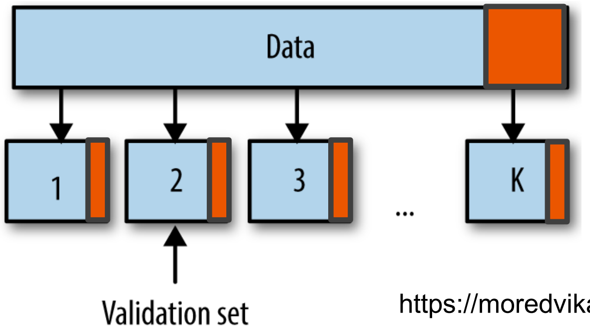


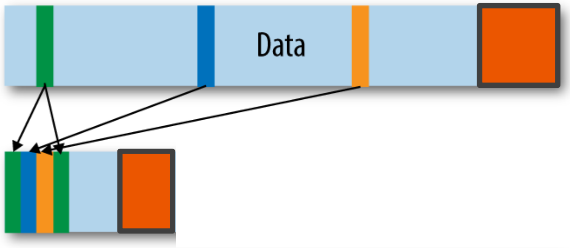
Test or no test



- For a proof-of-concept on small dataset, test sets can be dropped
    - Does not capture the variance of real data

Cross-validation and bootstrapping





https://moredvikas.wordpress.com/2018/10/10/machine-learning-model-validation-techniques/

- **Cross-validation** = equal split & used once
  - Minority class may be too few in each validation set
- **Bootstrapping** = repeated sampling
  - Customizable class ratios

Small dataset issues

- **Small test**
  - Estimated performance cannot be trusted
- **Small validation**
  - Select sub-optimal model
  - Select a biased model
- **Small training**
  - Poorly-fitted model
    - Less of a problem for linear model
    - Severe problem for tree model

Small dataset situations

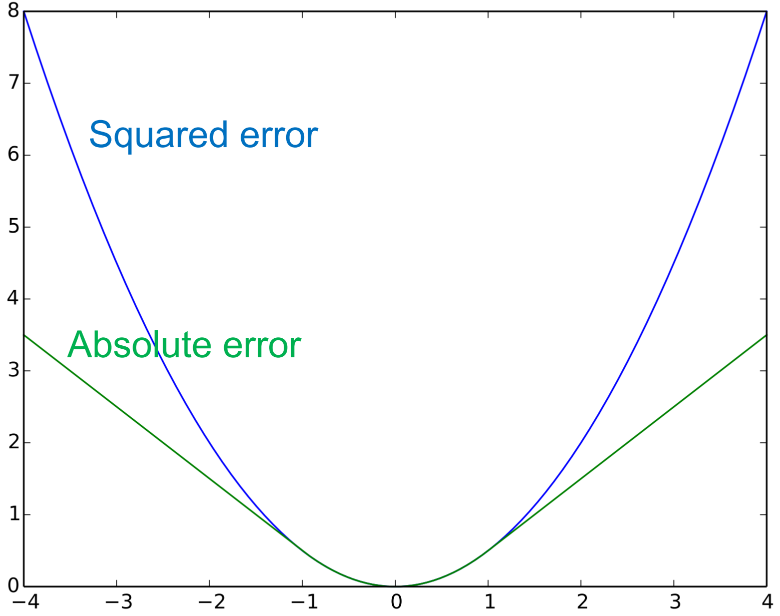
- **Example 1**: 223 negative, 77 positive
  - **Test**: 31 negative, 27 positive
  - **Validation**: 25 negative, 25 positive
  - **Training**: 167 negative, 25 positive

- **Example 2**: 48 negative, 23 positive
  - 2-fold cross-validation: 24 negative, 11 positive
  - Limited to logistic regression model
  - Limited to discussion of feature importance

Performance metrics

Regression metrics

- MSE, MAE, MAPE, R^2
- Select to match use case
  - Error <15%
  - Absolute error <1 unit



Classification metrics

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive
		Predicted < 0.5	Predicted > 0.5

- Accuracy = (TN + TP) / total
- Precision = TP / (TP + FP) = Positive predictive value

- Recall = TP / (TP + FN) = Sensitivity
- Specificity = TN / (TN + FP)

Classification use cases

- Screening for secondary inspection
  - Recall: Missed samples cannot be recovered
  - Improve precision during secondary inspection

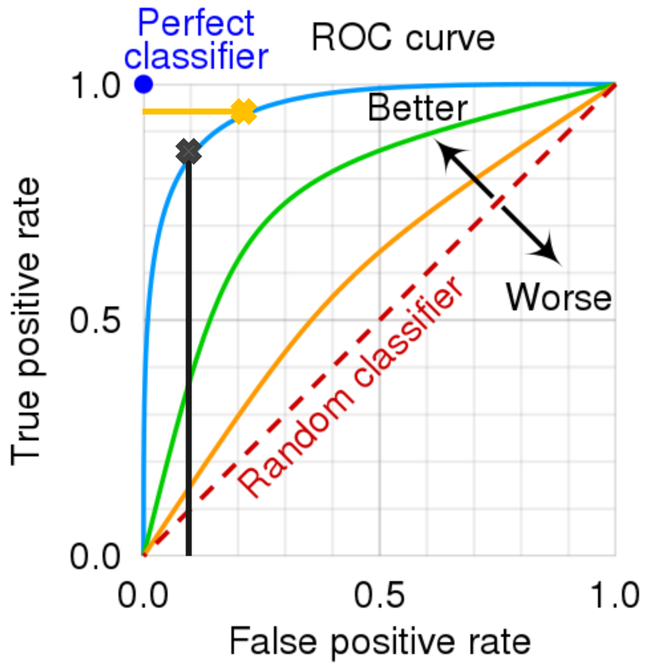
- Taking action based on prediction
  - Precision
    - Whether to perform surgery
  - Negative-class precision
    - Whether to send patient home
    - Whether the patient will be allergic to drug

Balanced classification metrics

- Accuracy

- $$F_1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$
- $$F_\beta = \frac{1 + \beta^2}{\frac{1}{\text{Precision}} + \frac{\beta^2}{\text{Recall}}}$$ give more weight to recall

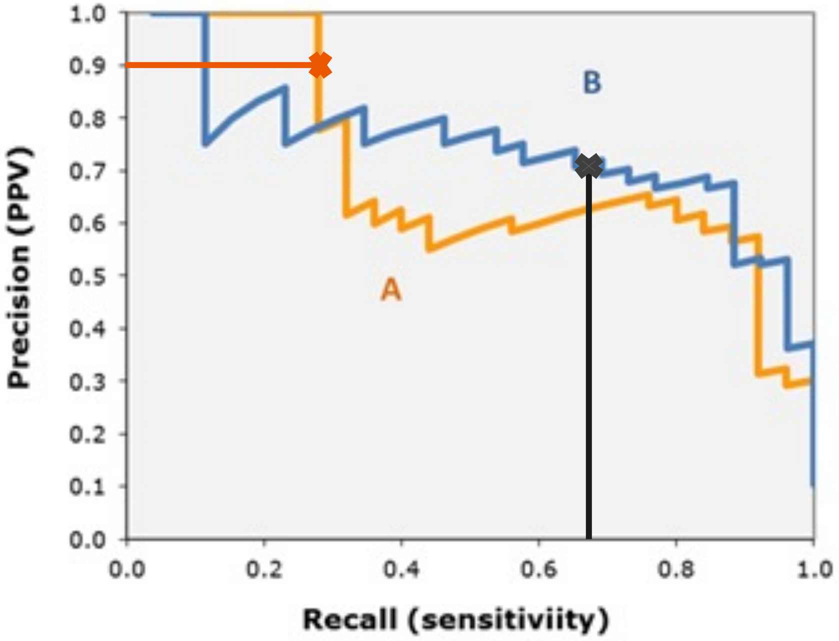
Threshold-free metrics



https://commons.wikimedia.org/wiki/File:Roc_curve.svg

- Sensitivity-specificity at every output threshold
- Area under the ROC curve (AUROC, AUC)
  - Random guess = 0.5
  - Perfect model = 1.0
- Pick threshold from use case
  - Specificity <0.1
  - Sensitivity >0.9

Precision-Recall curve



https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

- The best model can depend on use case

Do you need calibration?

Calibration curve

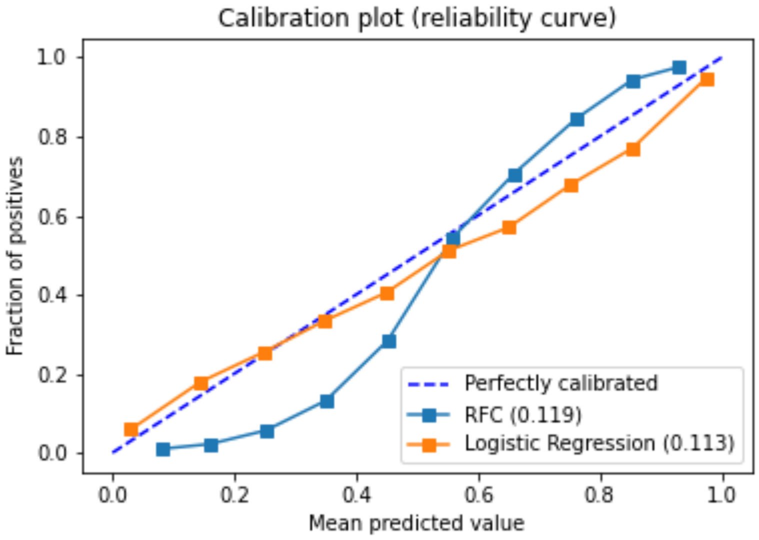
Classification Model

Predicted Probability

Calibrated Probability

Threshold

Positive or Negative



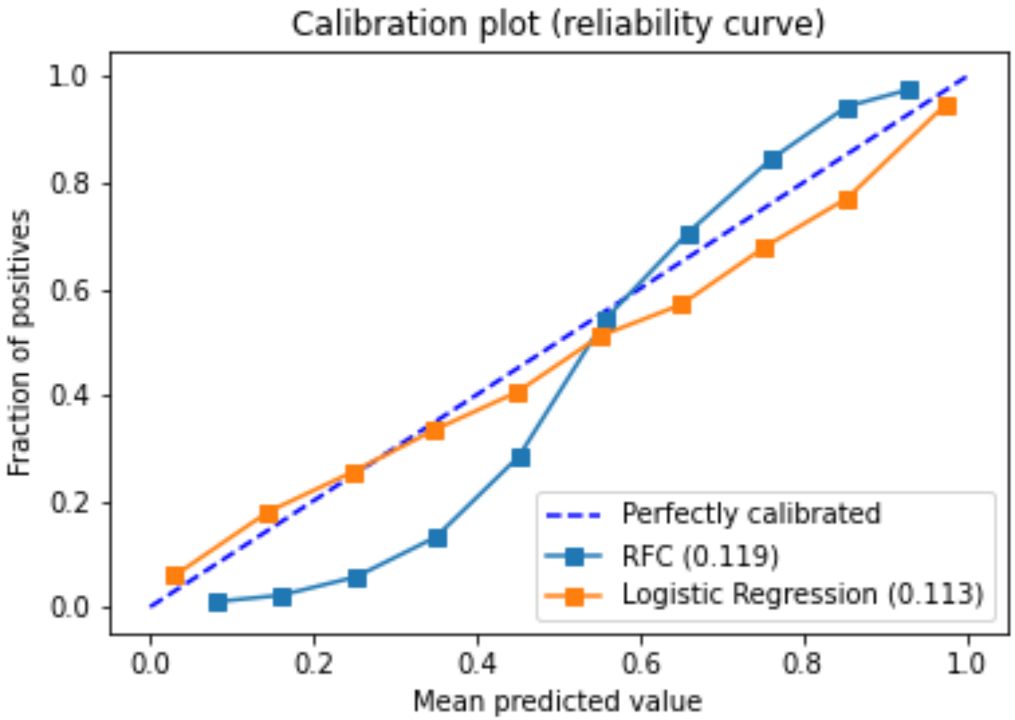
https://medium.com/analytics-vidhya/how-probability-calibration-works-a4ba3f73fd4d

- Calibration = correction of predicted probability
- Improve interpretability of the model

Data cost of calibration

- Estimate the true fraction of positive for EVERY OUTPUT RANGE

- 20 data points with predicted [0, 0.1]
- 20 data points with predicted [0.1, 0.2]
- ...



https://medium.com/analytics-vidhya/how-probability-calibration-works-a4ba3f73fd4d

Summary

- Clear objectives
  - What to predict? Why? Is ML the best answer? Human-in-the-loop

- Sufficient data collection
  - Aware of annotation/labeling cost
  - Beware of unintended biases

- Appropriate performance metrics
  - Match the objective and use case

- Be realistic + acknowledge limitation

Any question?

- See you on Wednesday 10-11am