3050571 Practical Clin Data Sci

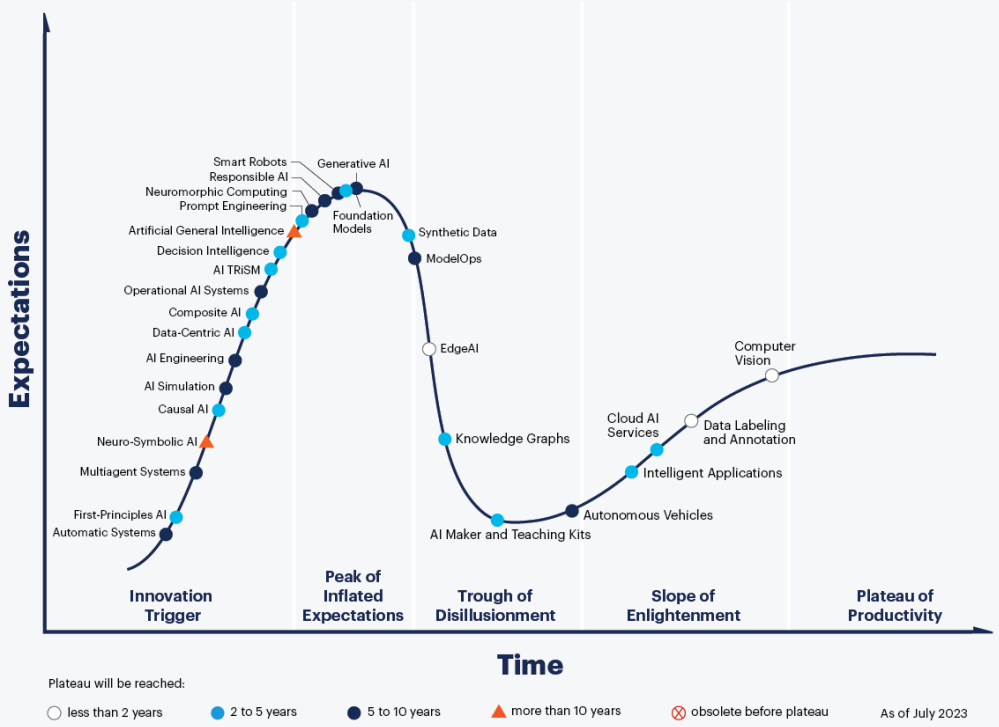
Session 17: Machine learning project design

March 7, 2024

Sira Sriswasdi, PhD

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

Hype Cycle for Artificial Intelligence, 2023



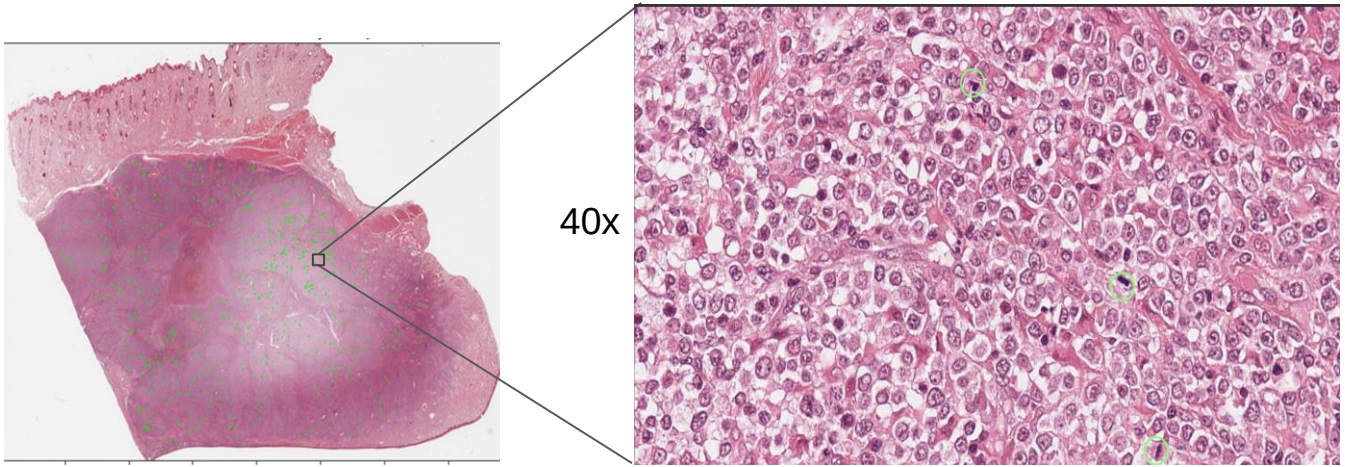
gartner.com

Gartner

- Solve the **most important bottleneck** with the **simplest, proven solution**
- Deep learning model designs may be easy to understand conceptually but can be very difficult to tune

Define the tasks and use cases

Focus on the pain point

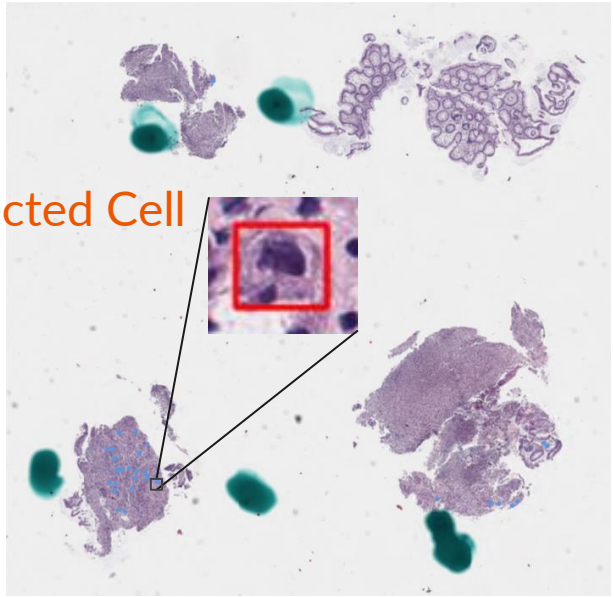


Whole Slide Image (WSI)
(150,000 x 150,000 pixels)

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

Use cases define performance metrics

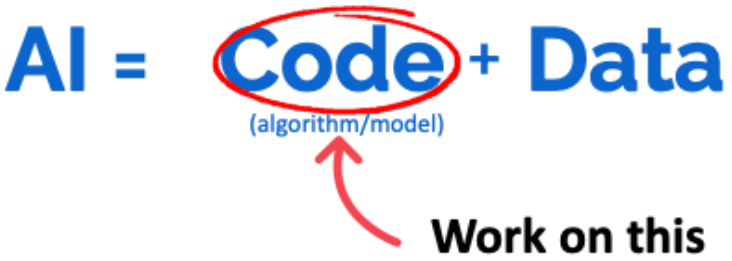
- Screening patients for secondary inspection
  - Recall: Missed samples cannot be recovered
  - Improve precision during secondary inspection

- Taking high-risk action based on prediction
  - Precision (positive predictive value)
    - Whether to perform surgery
  - Negative-class precision (negative predictive value)
    - Whether to send patient home
    - Whether the patient will be allergic to drug

Data-centric approach

Data-centric approach

Conventional model-centric approach:



Data-centric approach:

Require understanding of the data generation and collection process

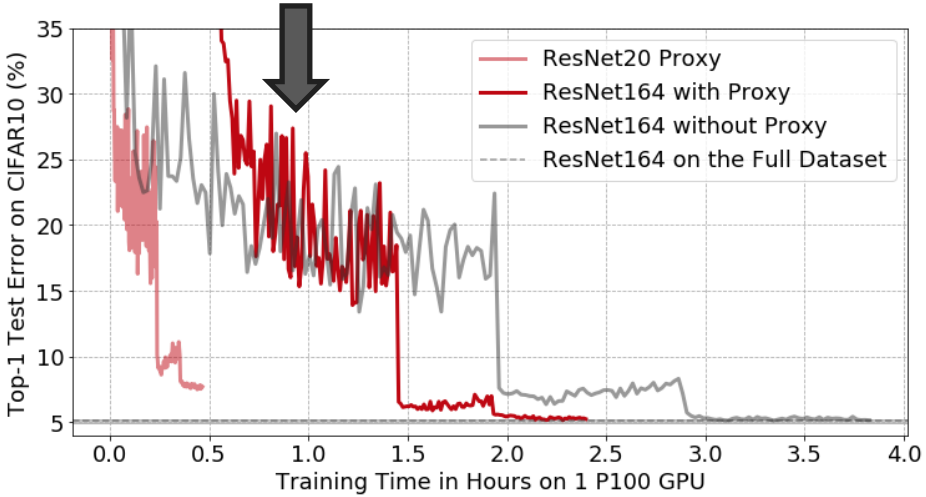
https://landing.ai/tips-for-a-data-centric-ai-approach/

Smart data selection

Passive: err ~ n^{-1}
Active: err ~ 2^{-n}

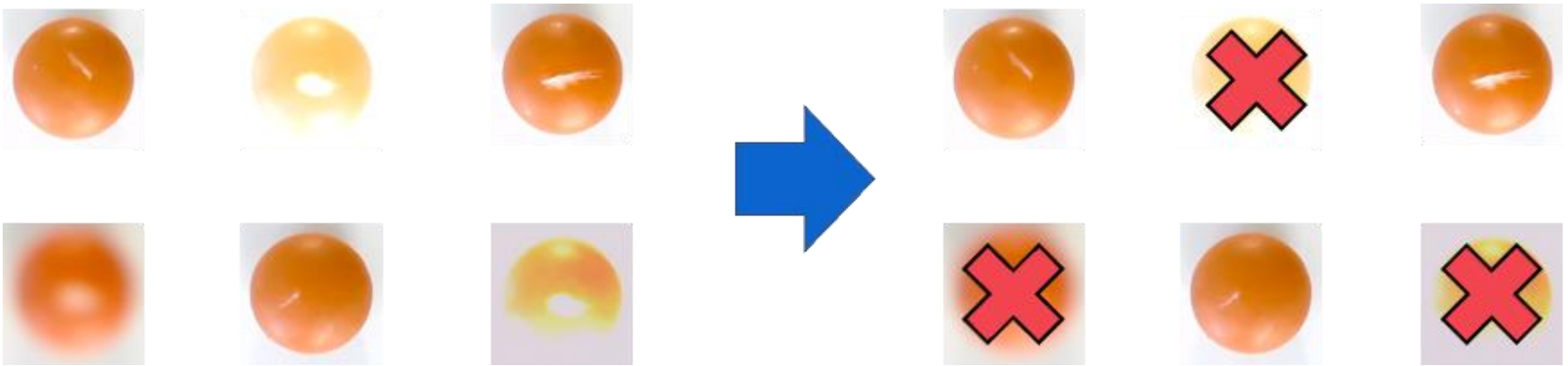
https://dcai.csail.mit.edu/2023/growing-compressing-datasets/

Fewer data needed to achieve the same performance



- Problems come from powerful model + incomplete data

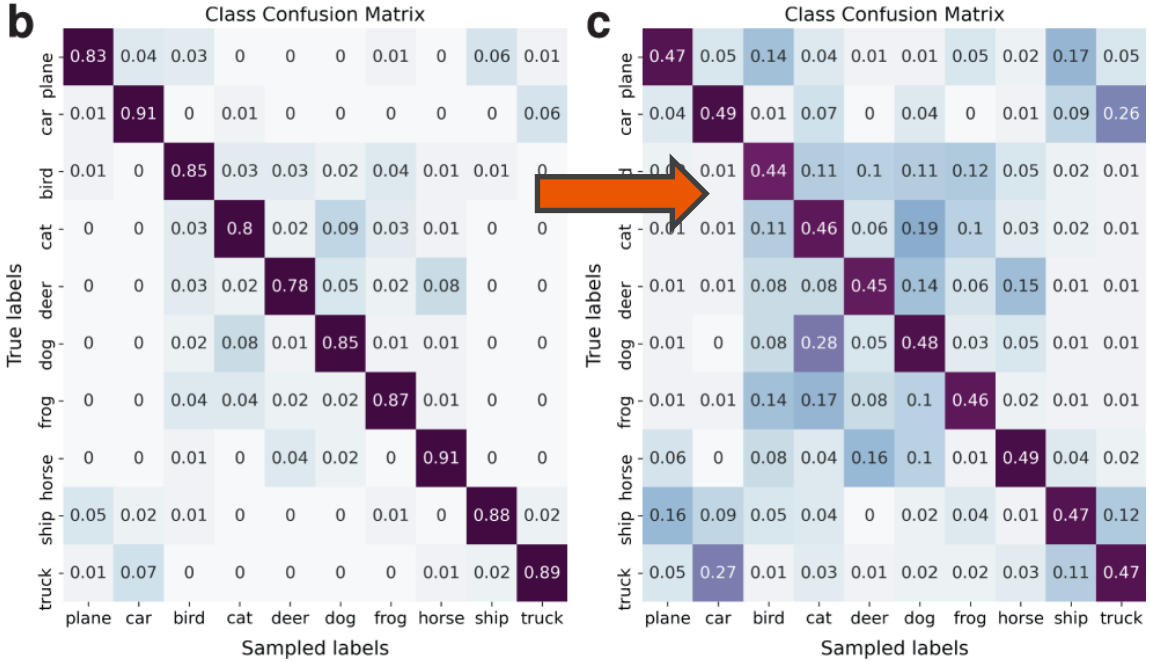
More is not always better



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

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

Beware of hard samples



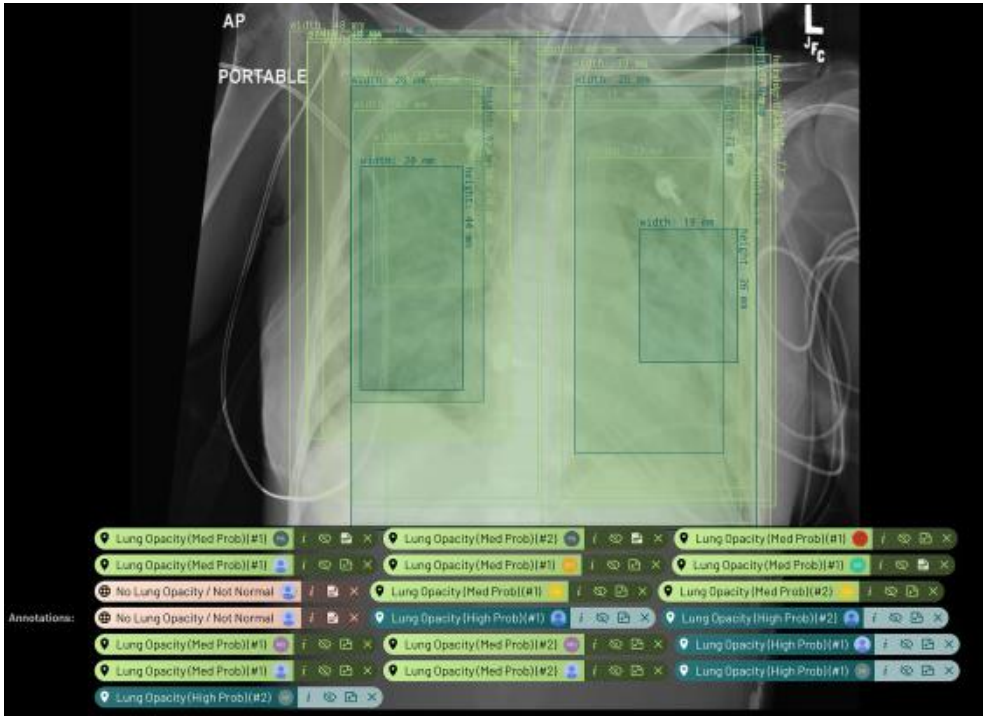
Bernhardt, M. et al. Nature Comm 13:1161 (2022)

- 3x-5x increase in label error in hard samples

Synthetic data from generative AI

- Reduce data requirement for new hospital
- Minimize data privacy issue
- Self-supervised or annotated
  - Generated with guided prompt

Manual labeling for chest x-ray



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

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

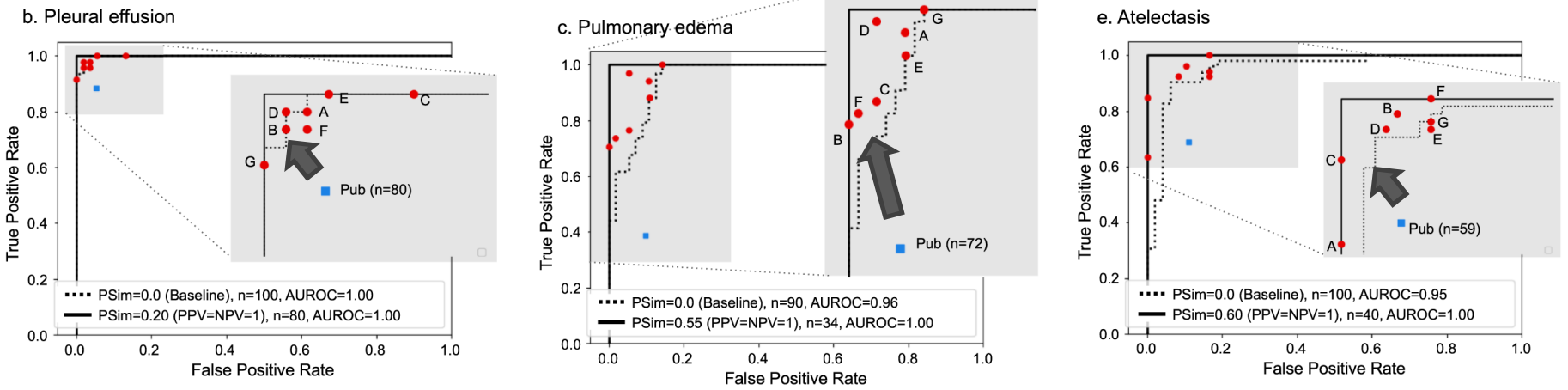
Automated labeling for chest x-ray

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

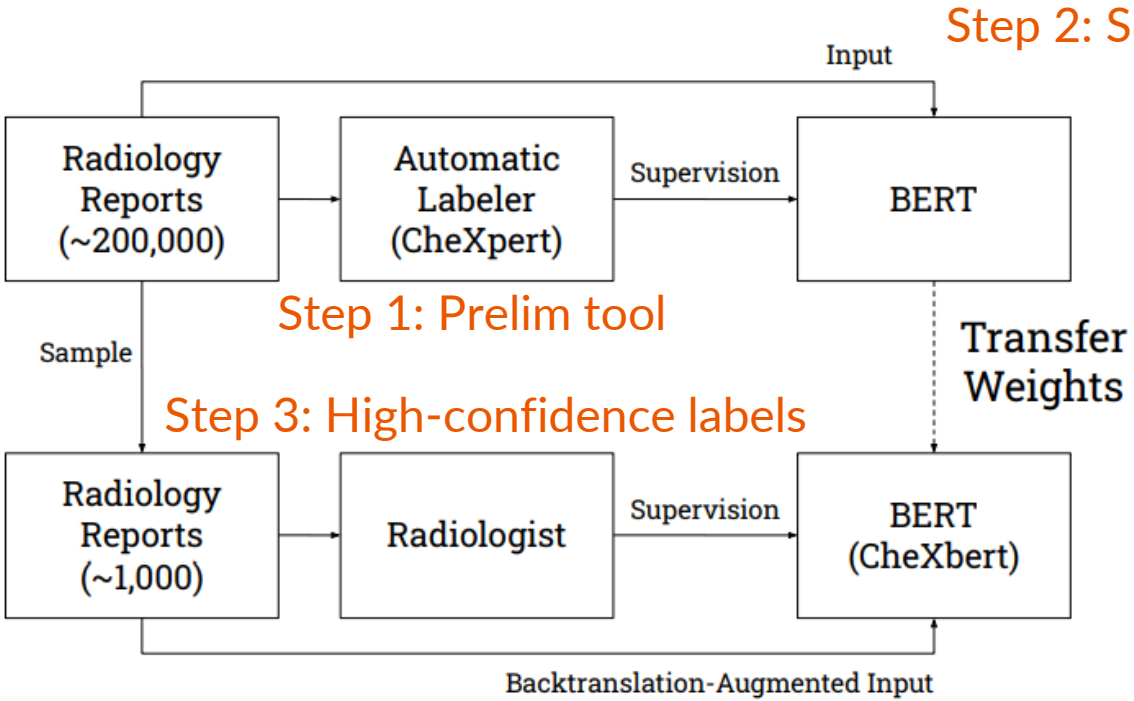
Poor quality from automated labeling



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

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

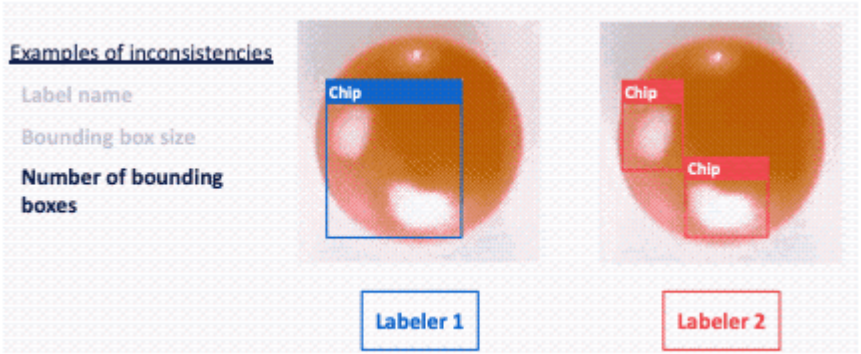
Iterative labeling process

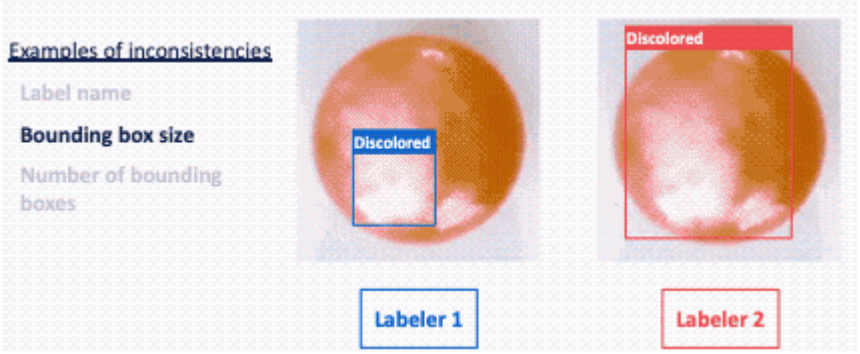


Smit, A. et al. https://arxiv.org/pdf/2004.09167

- Automated labeling reduces time spent on easy samples

Labeling standard



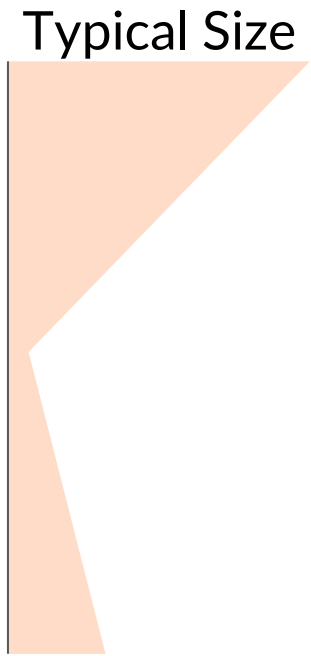


https://landing.ai/tips-for-a-data-centric-ai-approach/

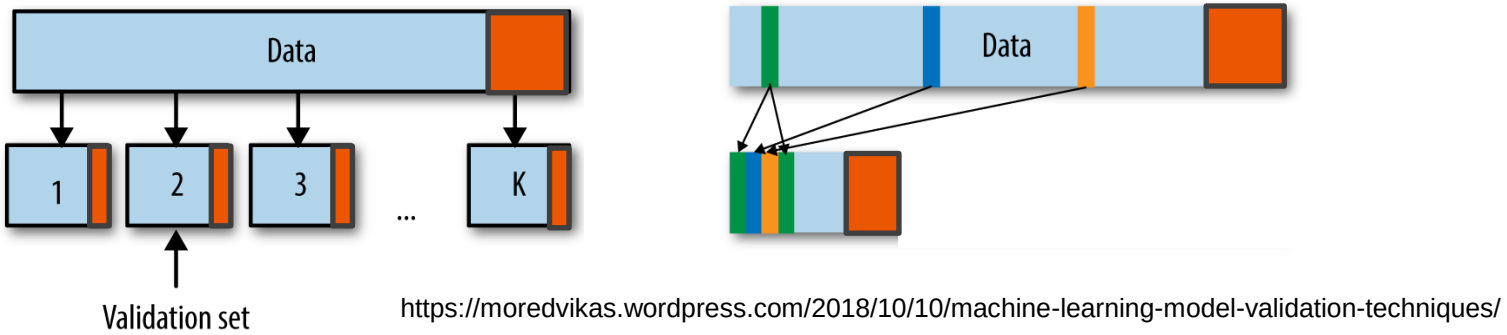
Data spliting tips

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



Cross-validation vs bootstrapping



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

- **Cross-validation** = equal split & used once
- **Bootstrapping** = repeated sampling
  - Full control over the proportion of every class

Problems from small dataset

- **Small test set**
  - Estimated performance cannot be trusted ← silent problem!
  - Can use validation instead initially
- **Small validation set**
  - Select suboptimal, biased model ← silent problem!
- **Small training set**
  - Poorly-fitted model ← clearly observed
  - Limit the usage of complex model

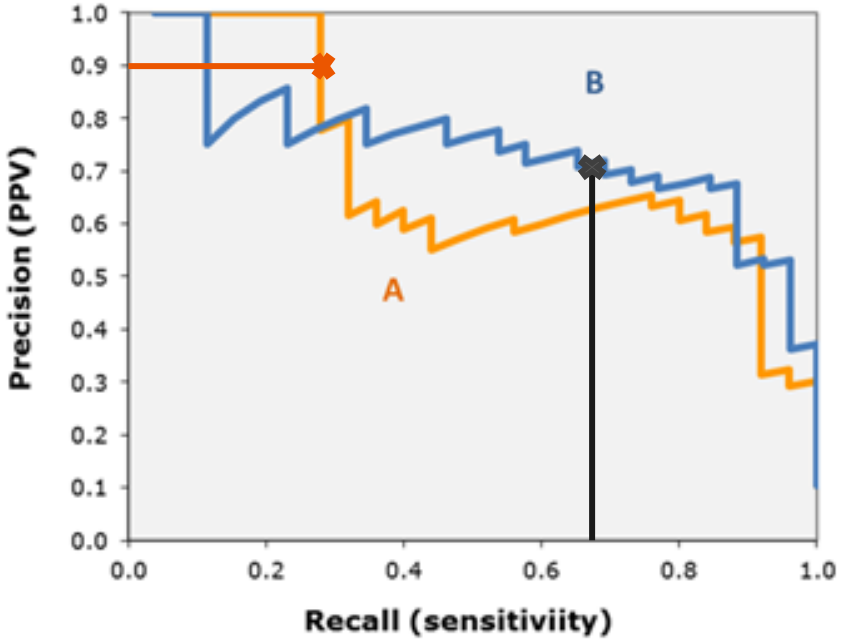
Some example situations

- **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

Model calibration / cutoff tuning

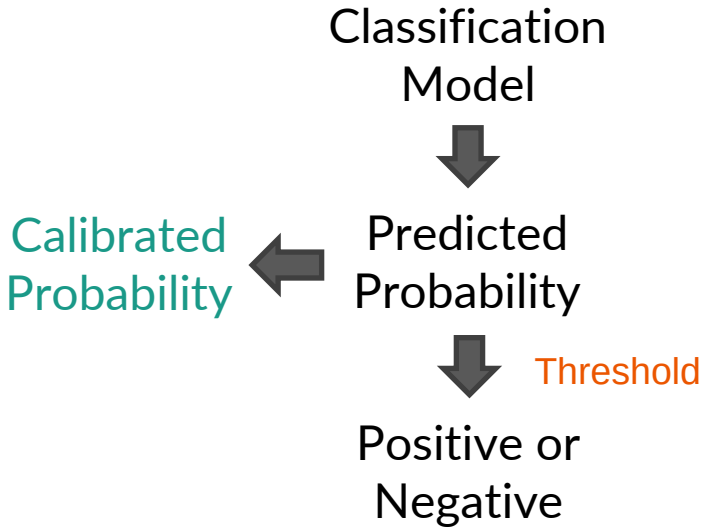
Finding the right cutoff

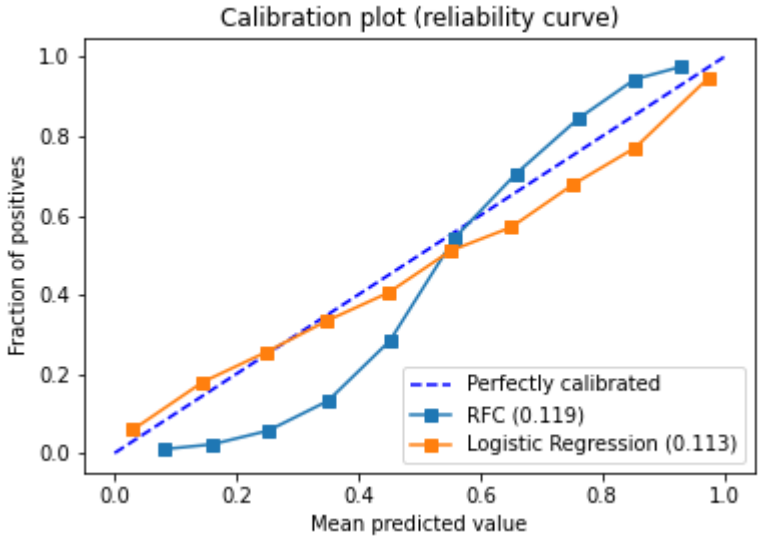


https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

- Requirement user testing in realistic environment

Calibration curve





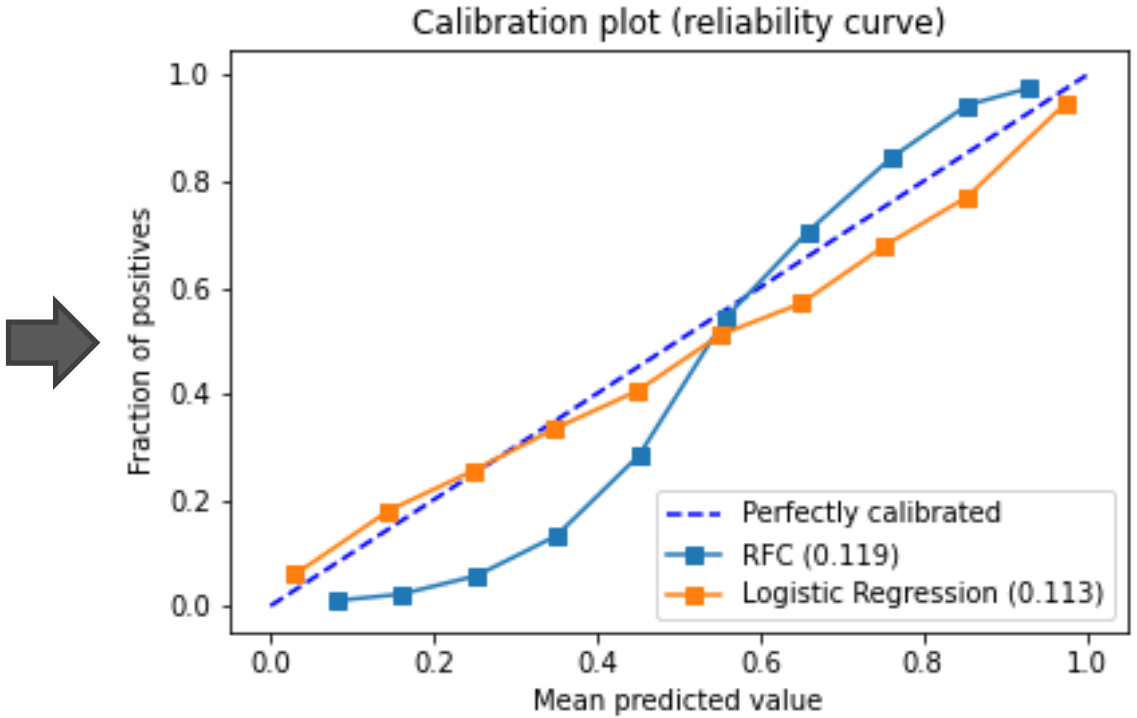
https://medium.com/analytics-vidhya/how-probability-calibration-works-a4ba3f73fd4d

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

Data cost of calibration

- Estimate the true fraction of positive for EVERY OUTPUT RANGE

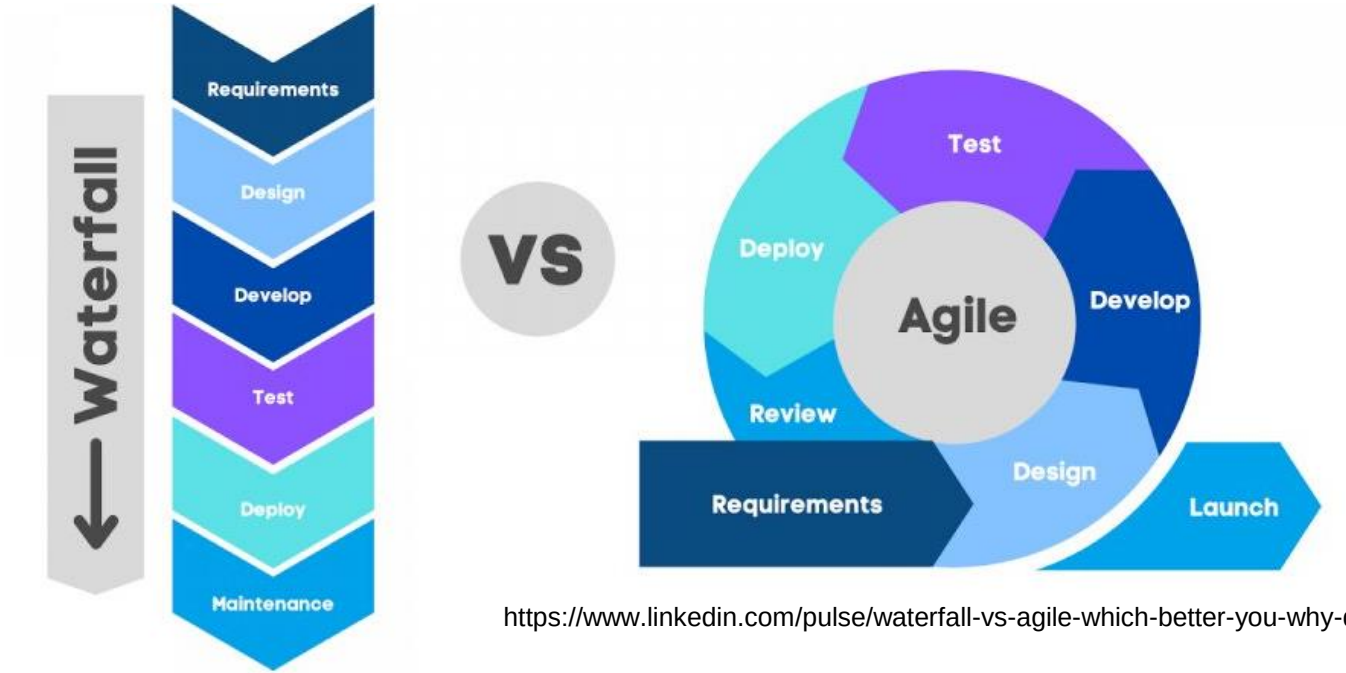
- 20 data points with predicted [0, 0.1]
- 20 data points with predicted [0.1, 0.2]
- …



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

Agile management

Agile development cycle



https://www.linkedin.com/pulse/waterfall-vs-agile-which-better-you-why-datacademy-cloud/

Back to computational thinking principles

- **Decomposition & Abstraction**
  - Identify pain points and use cases
- **Pattern Recognition**
  - Proof-of-concept with public / small datasets
  - Start with simple model
- **Algorithm**
  - Design workflow with your solution integrated

Any questions?

See you on March 8th