COMP3516: Data Analytics for IoT

Lecture 10: Systems Evaluation

Chenshu Wu
Department of Computer Science
2025 Spring
April 15, 2025

香 港 大 學
THE UNIVERSITY OF HONG KONG

HKU AIoT LAB
香港大學人工智能物联网實驗室

Contents

- Learning Outcome: How to evaluate a system/method?

- Metrics and Charts
  - Accuracy, Precision, Recall, F1-Score
  - Confusion Matrix
  - Detection Rate and False Alarm Rate
  - ROC Curve
  - Boxplots and Violin plots
  - CDF
- Methodology

Why to evaluate?

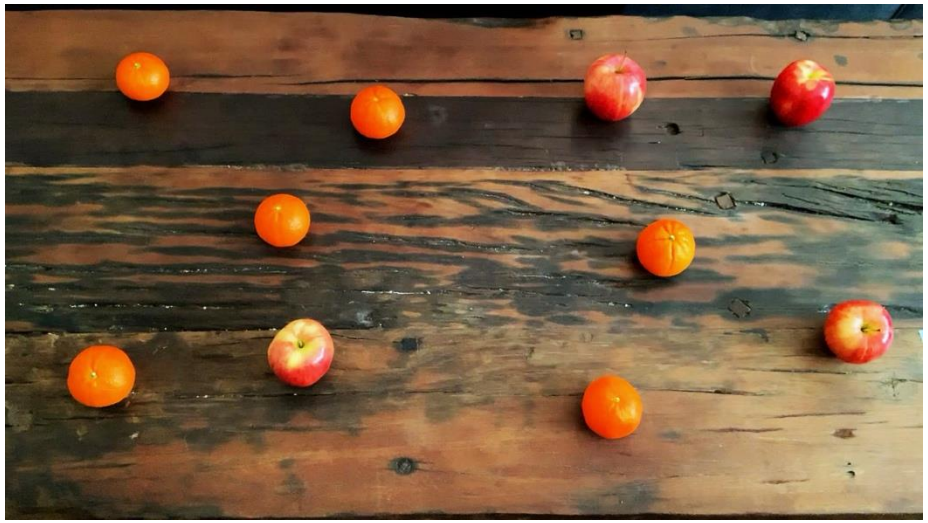
- Ensure system requirements are met
- Understand the boundaries (including the limitations)
- Compare with benchmarks or SOTA
- Explore trade-offs

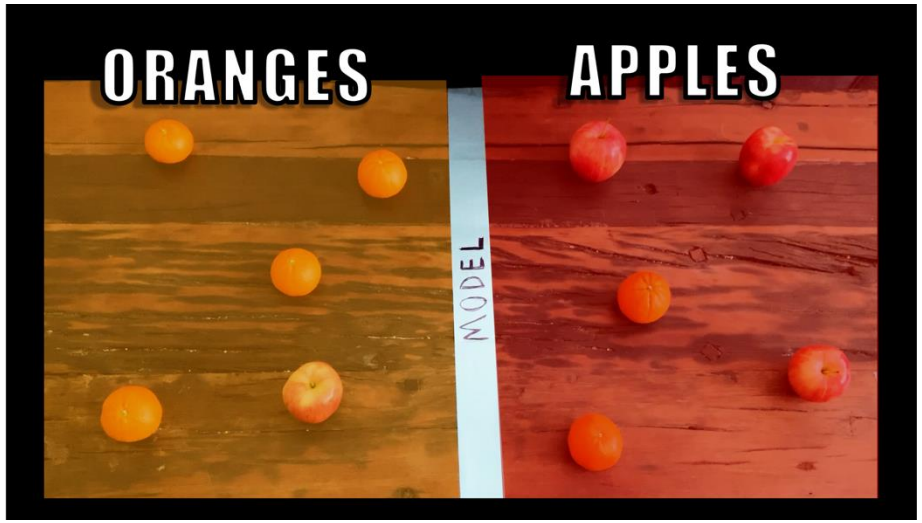
How did you evaluate your results?

- In labs and projects

Classification Metrics

- consider building a model to classify apples and oranges on a flat surface

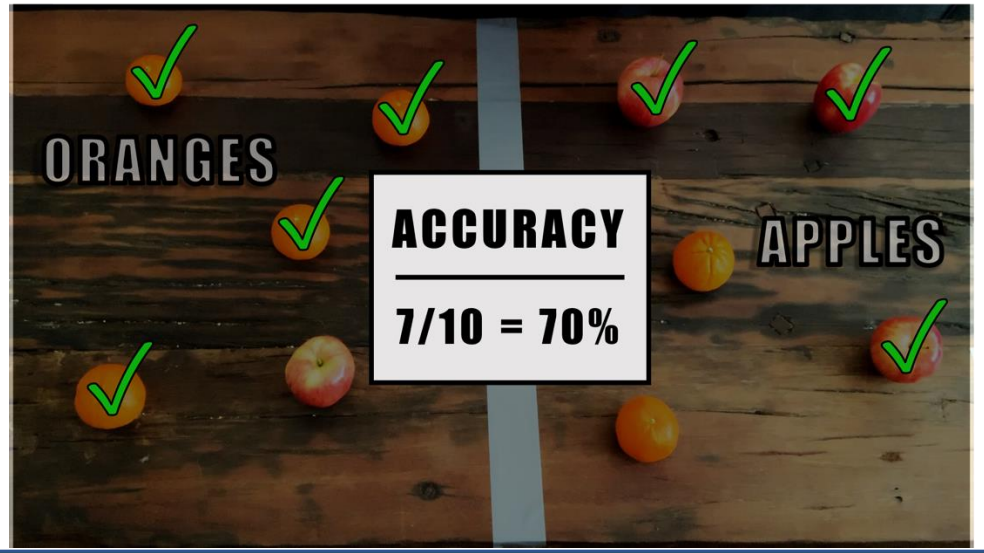




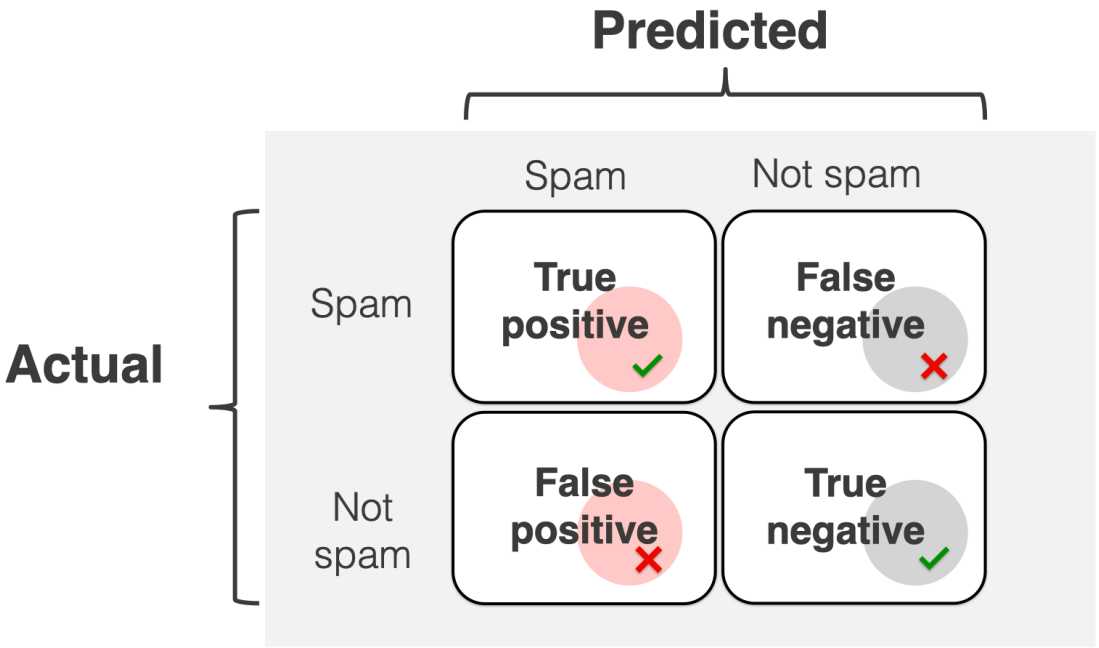
Source: https://kimberlyfessel.com/mathematics/data/accuracy-precision-recall/

Accuracy

- Accuracy: all of the correctly classified observations and divide by the total number of observations
- one of the most popular classification metrics
- Imbalanced dataset; that is, what if we have 990 oranges and only 10 apples?



Confusion matrix (binary classification)



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

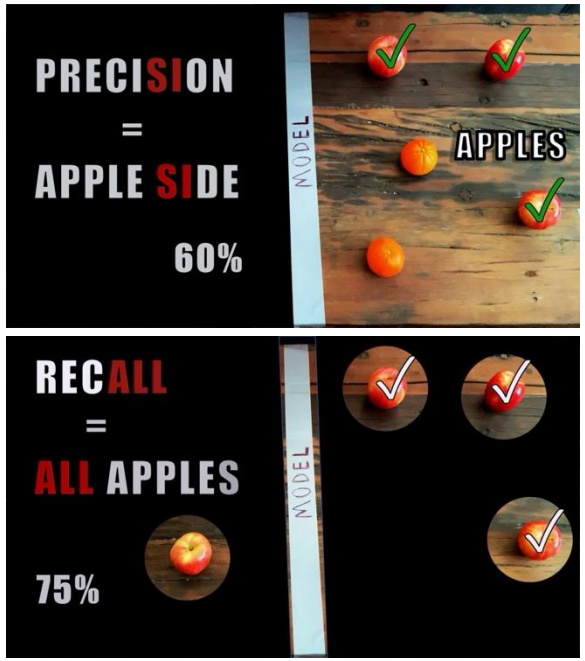
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

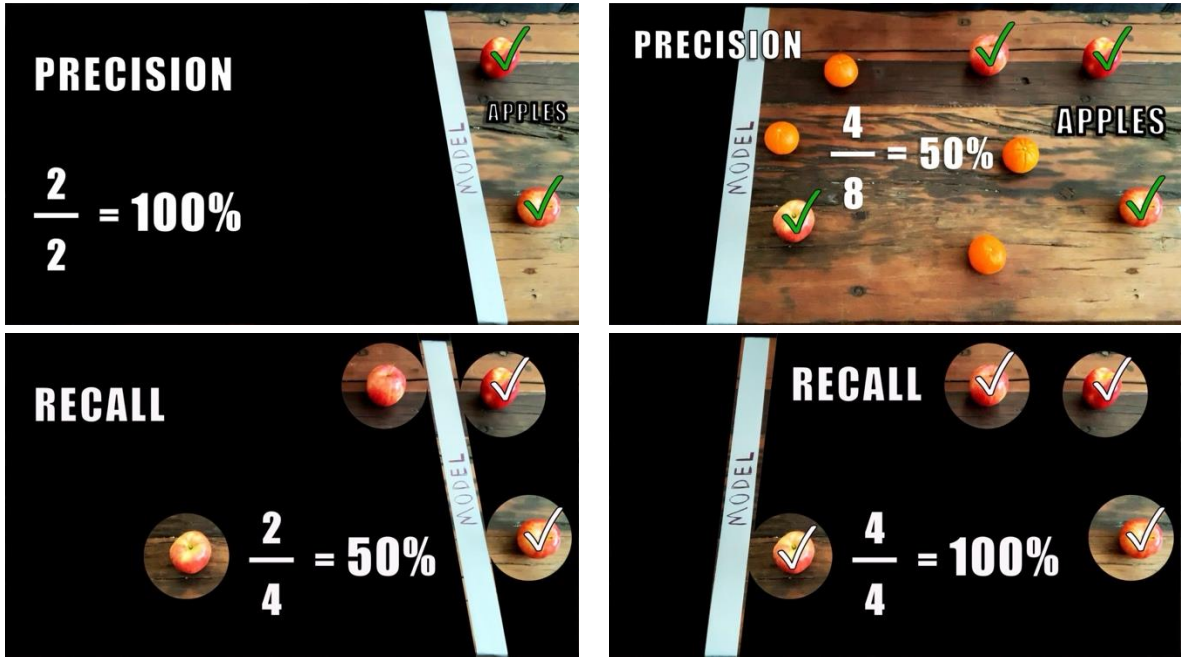
- precision measures the accuracy of positive predictions
- recall measures the ability to find all relevant positive instances

Precision and Recall

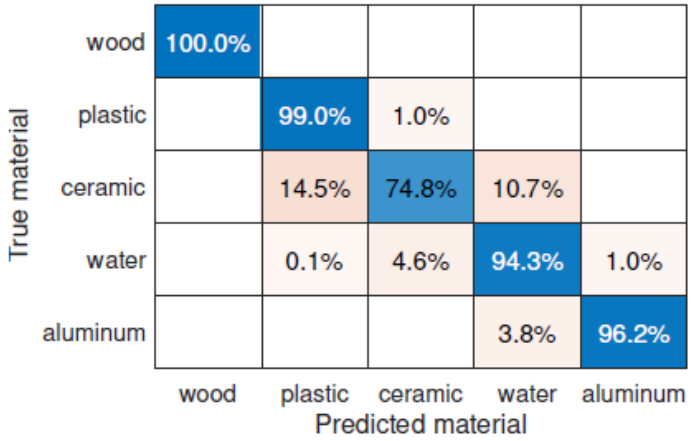
- Both precision and recall are defined in terms of just one class, oftentimes the positive—or minority—class.
- Precision-Recall Tradeoff

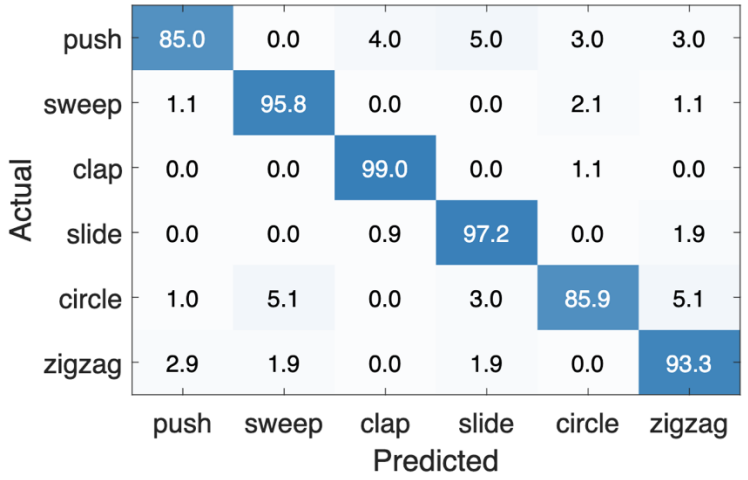
$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

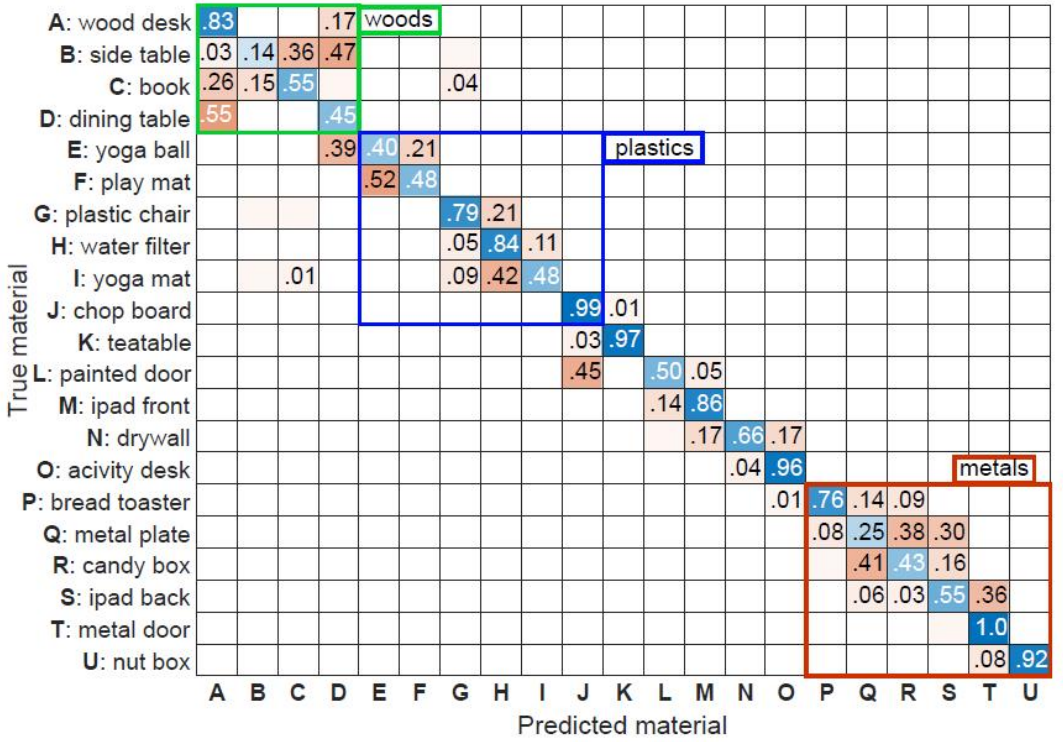




Confusion matrix







When to use which?

- Precision: Use when false positives are more costly than false negatives.
- Recall: Use when false negatives are more costly than false positives.
- F1-score: Use when you want to balance precision and recall.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

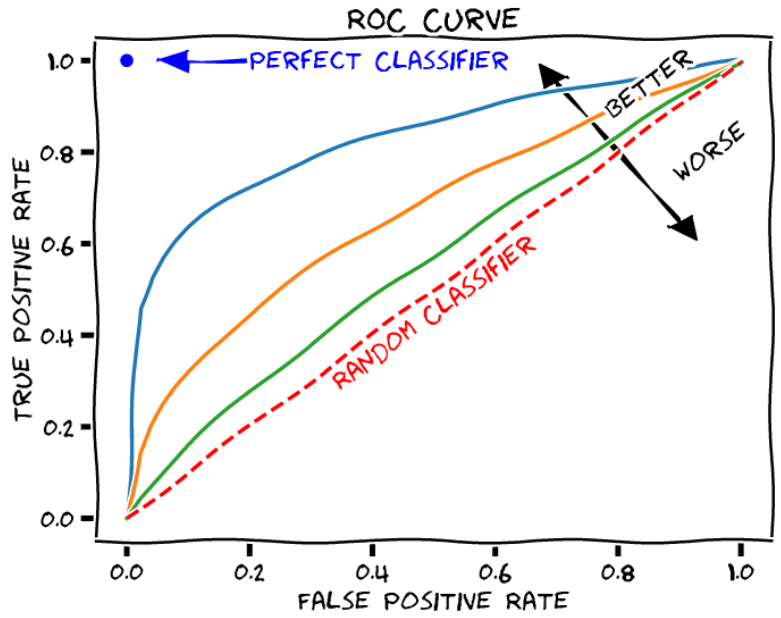
$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

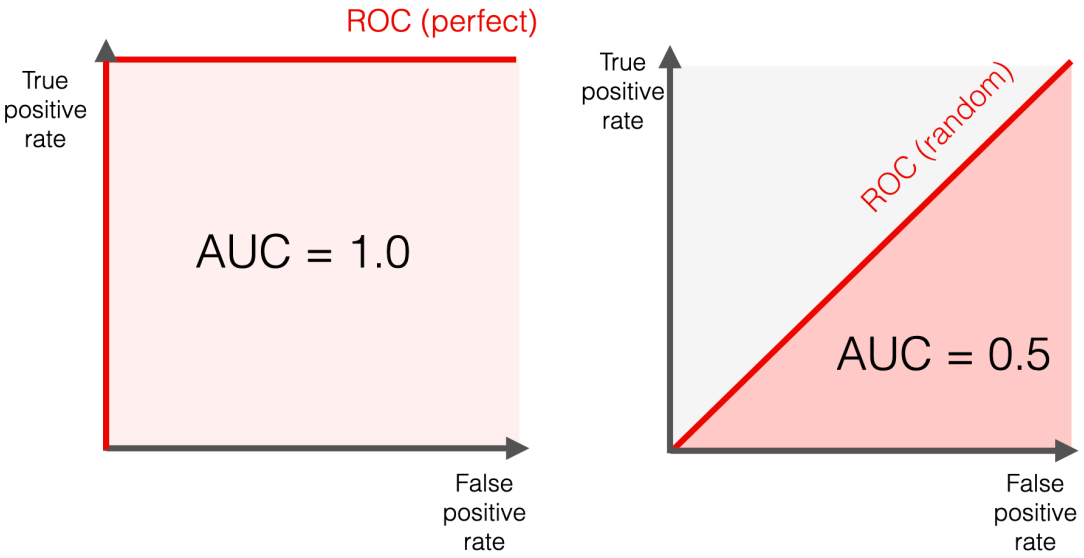
Detection

- Detection Rate
  - How many events (X) are correctly detected out of a total of N of true events? DR = X/N
- False Alarm Rate
  - How many events are detected when there are actually no events of interests?
  - Can depend on the definition of the period/number of samples for “no events”.
- Trade-off between DR and FAR
  - Applications can be more DR-sensitive or FAR-sensitive
  - ROC curve

ROC curve

- Receiver-Operating Characteristics (ROC)
- Useful to show sensitivity to threshold, or search for the best threshold





AUC-ROC

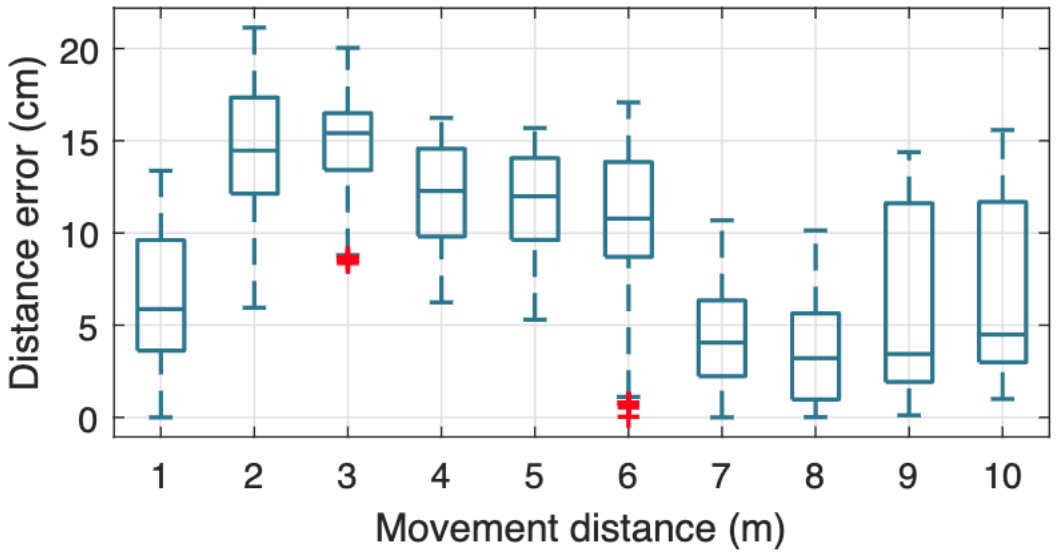
Regression Tasks

- Example tasks:
  - Localization
  - Heart/Breathing rate estimation
- Calculation
  - L1/L2 norm: e = ||estimate – ground_truth||
  - Consider n samples: 1/n * sum(e_i)
  - mean squared error (MSE)
  - root mean squared error (RMSE)
- Statistics
  - Mean, median, percentiles
  - More informative formats?

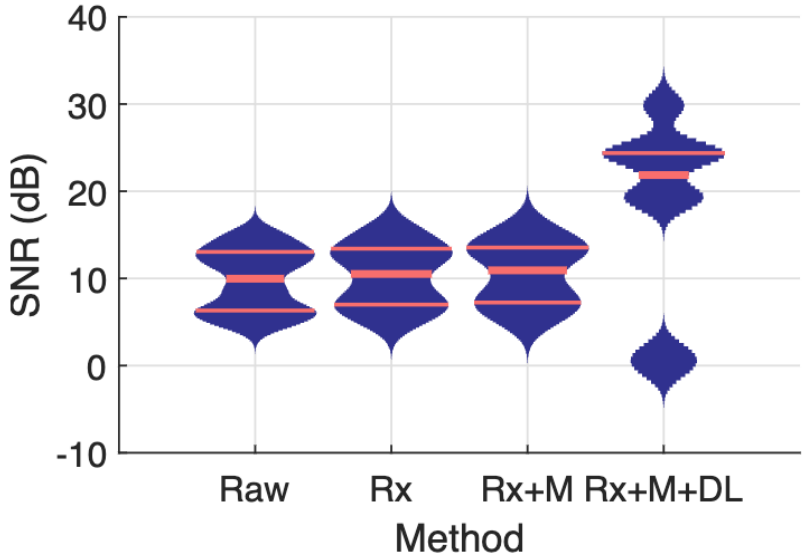
Mean Error Squared

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Boxplots & Violin plots



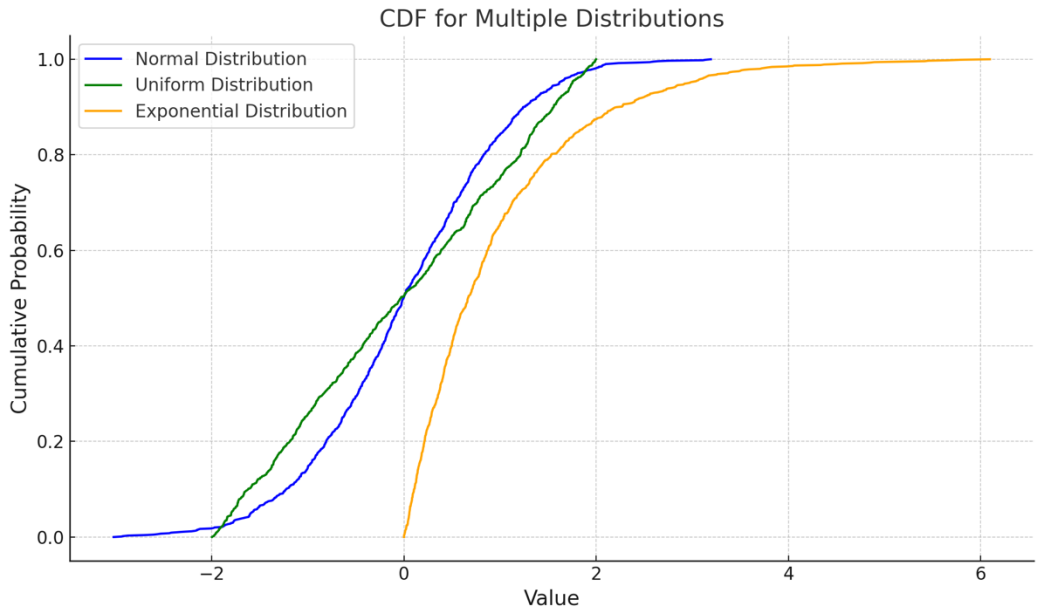
showing the median, quartiles, and outliers

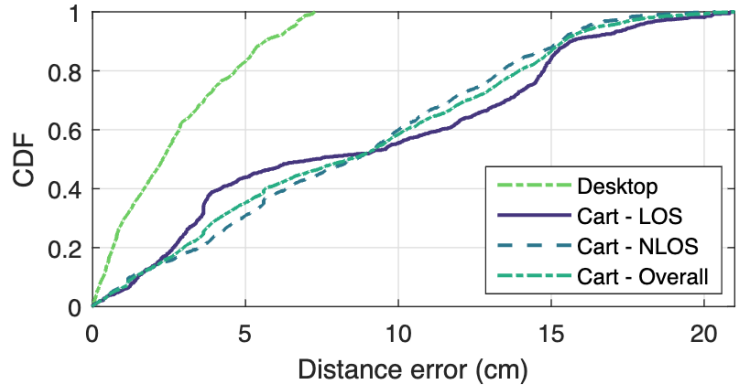


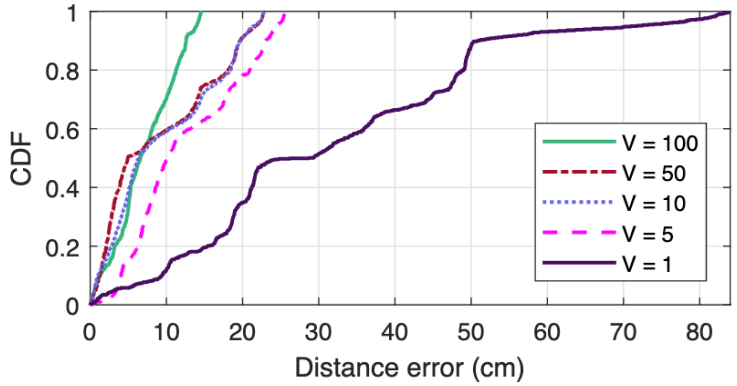
show the full distribution, including the median, quartiles, and the overall distribution shape

Cumulative Distribution Function (CDF)

- Show the distribution of a variable over its entire/possible range
- Give a sense of percentile performance
- Default metric for localization







Beyond Accuracy

- Computation Complexity
- Latency
- Throughput
- Memory/cpu usage
- Energy/Power consumption

- Scalability
- Security & Privacy metrics

- Usually more important for a (real-time) systems
- Essential to clearly state what platforms are used for evaluation

What to evaluate

- Overall performance
  - One or two key metrics
- Ablation study
  - How different components contribute to the overall performance?
- Parameter study
  - How does the system work under all impacting factors?
- Comparison study
  - How does your method perform compared with baseline approaches?
- Cross-validation
- Field trials/Real-world evaluation

Case Study: Breathing Rate Estimation

Summary

- There are many more metrics to explore and use.
- Some are specific to certain applications/fields (eg, PESQ, STOI for speech)
- Need to decide the most appropriate metric(s) depending on your applications and the performance you care about.

- The key is that, it is important to conduct extensive evaluation in a logical and scientific manner, and report the results in a clear way.

Questions?

- Thank you!