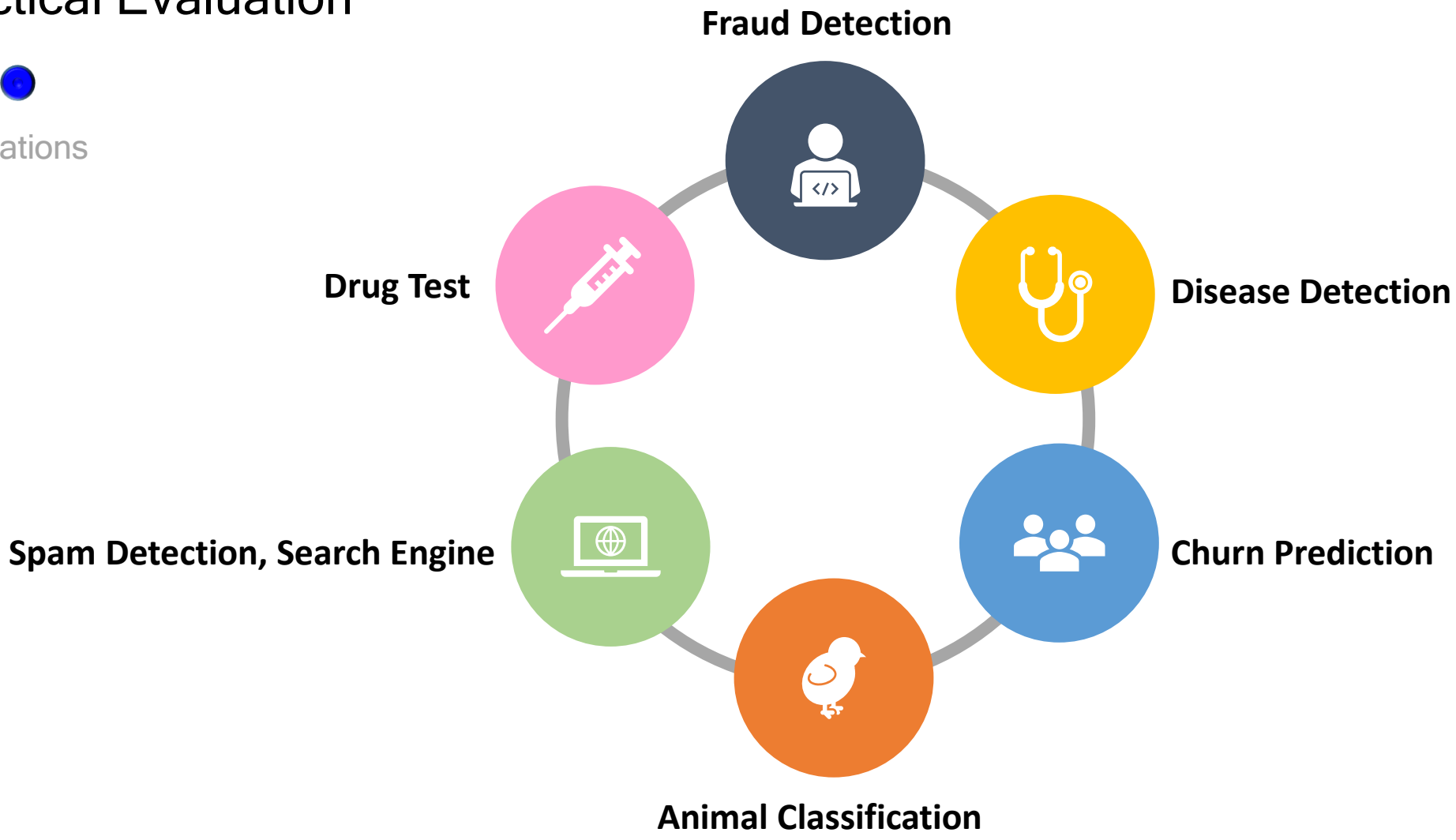
Mahidol University
Faculty of Medicine Ramathibodi Hospital
Section for Clinical Epidemiology and Biostatistics

Model Evaluation

Ratchainant Thammasudjarit, Ph.D.

Practical Evaluation

Applications



Practical Evaluation

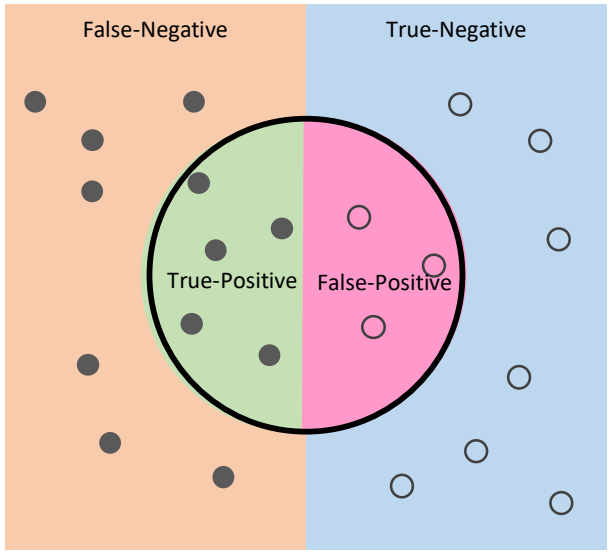
FAQ

Accuracy: 0.90

Q: Is my model good enough?

A: Depends on application, impact, and expectations

Choose the right measures



Accuracy answers the following question:

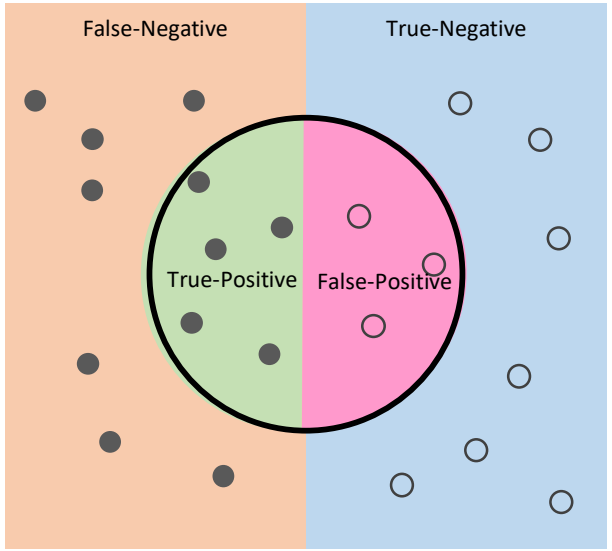
How many samples are correctly labeled out of all samples?

Accuracy is a good measure when impact of FP and FN are similar and balanced class distribution

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Ex. Distinguishing Male and Female Chick

Choose the right measures



Precision answers the following question:

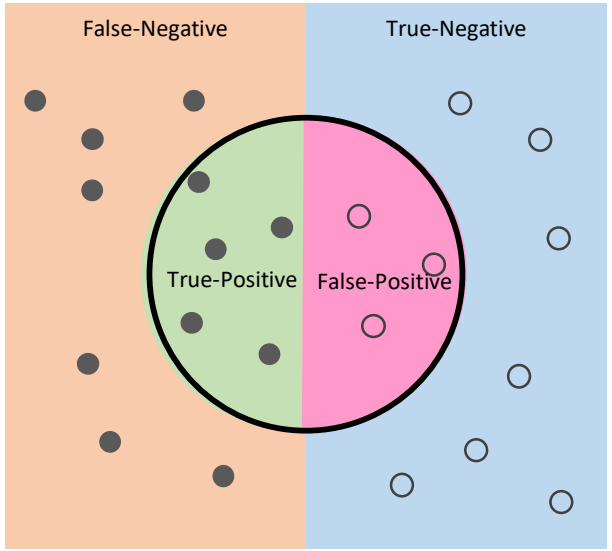
How many samples labeled as positive are actually positive samples?

Precision is a good measure when impact of FP must be minimized

$$precision = \frac{TP}{TP + FP}$$

Ex. Spam Mail Detection

Choose the right measures



Recall (a.k.a. Sensitivity) answers the following question:

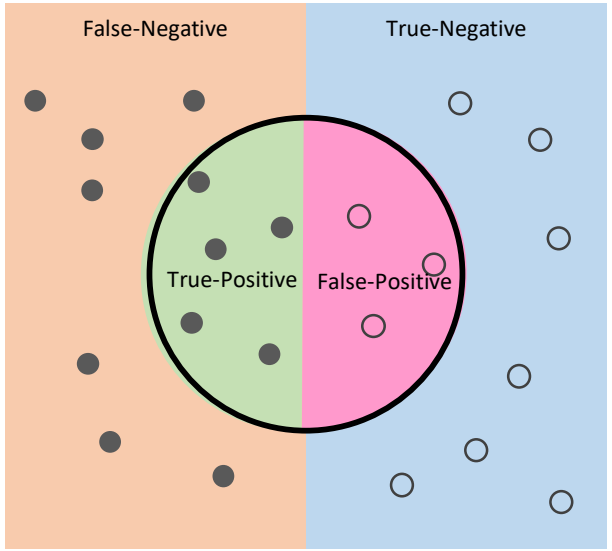
How many samples from all positive samples are correctly predicted?

Recall is a good measure when impact of FN must be minimized

$$recall = \frac{TP}{TP + FN}$$

Ex. Disease detection, Fraud Detection, Churn Prediction

Choose the right measures



Specificity answers the following question:

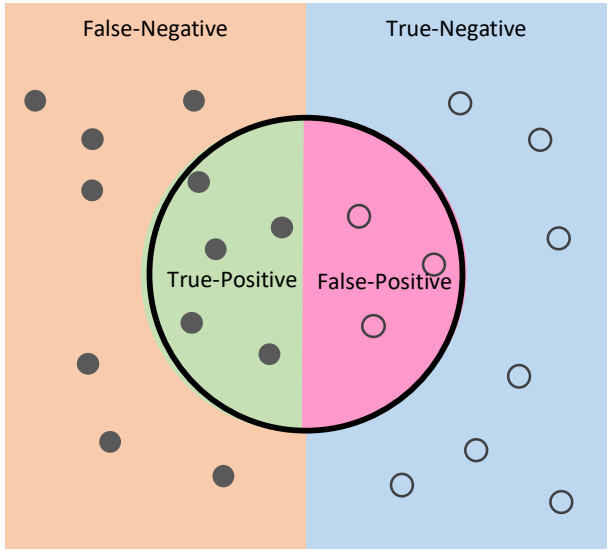
How many samples from all negative samples are correctly predicted?

Specificity is a good measure when we want to cover all TN and do not want FP

$$specificity = \frac{TN}{TN + FP}$$

Ex. Drug Test, Alcohol Test

Choose the right measures



F1-Score (a.k.a. F-Score) balances between precision and recall:

F1-Score is a good measure when impact of FP and FN are different and imbalanced class distribution

$$F1 = \frac{2PR}{P + R}$$

Ex. Search Engine

Model selection

Classifiers	Accuracy	Runtime (ms)
A_1	90%	80
A_2	92%	90
A_3	95%	1,500

Which classifier is the best?

- Given the following models

  Linear Weight Combination (Bad Idea)

  Optimizing subjected to satisfying (Better Idea)

  Ex. Accuracy represents optimizing while runtime represents satisfying

  Selection: Maximize accuracy subjected to runtime < 100 ms

Model selection

Classifiers	Metric 1	…	Metric k
A_1			
…			
A_N			

- Summary

Maximize one objective subjected to at least one constraint