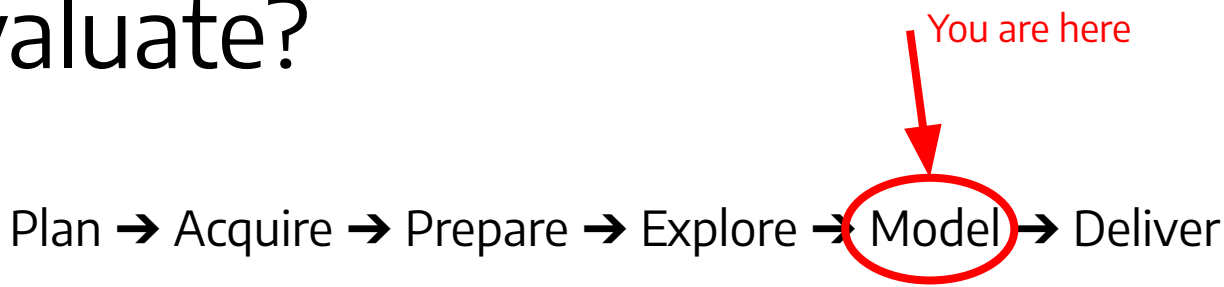
Classifier Evaluation

Zach Gulde

Why Evaluate?



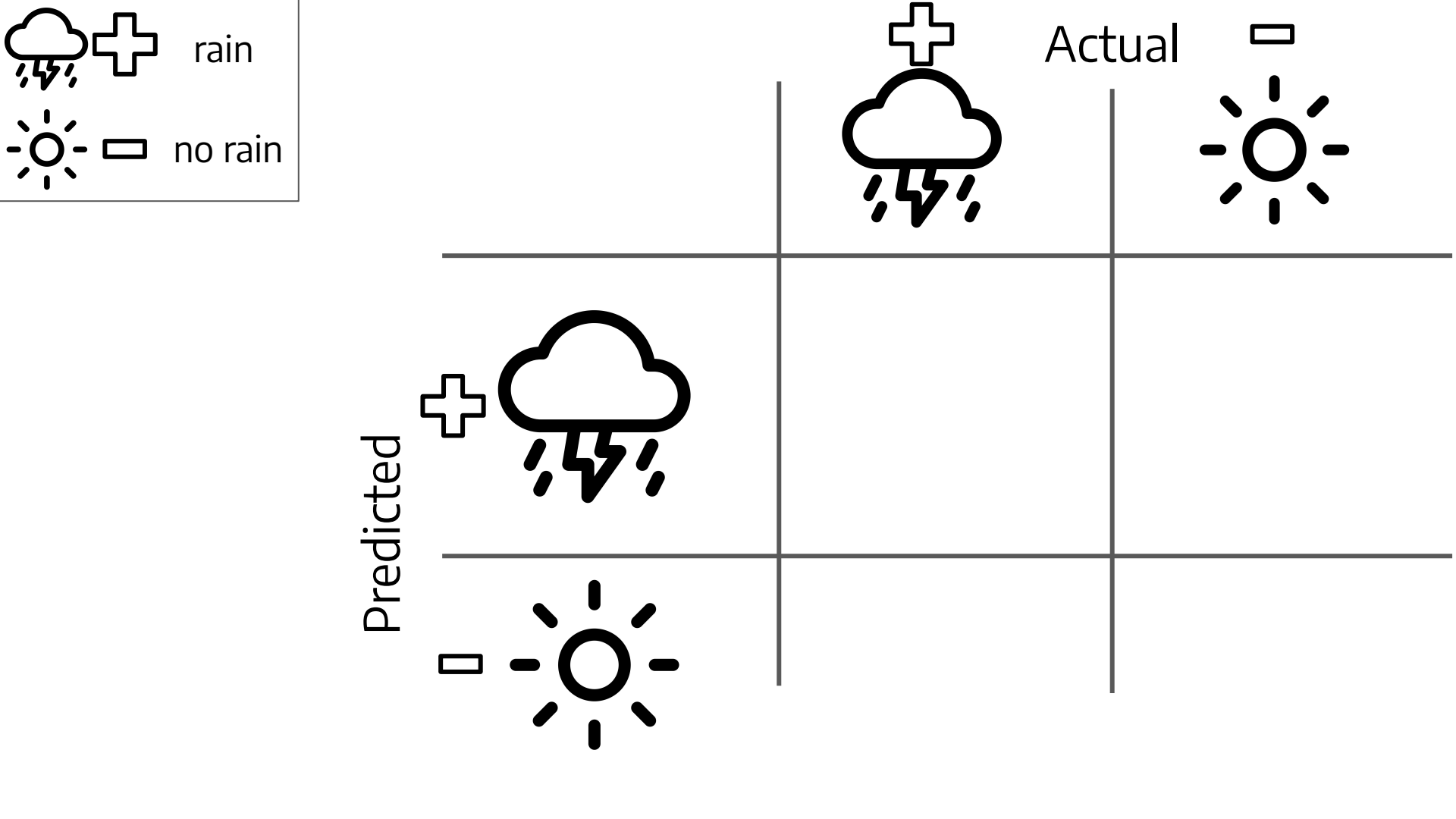
- Quantifying model performance allows to compare models (ML or otherwise!)
- How we quantify performance is key
- Many different ways to quantify depending on what we want to optimize for

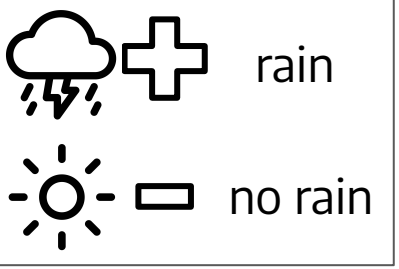
Vocab

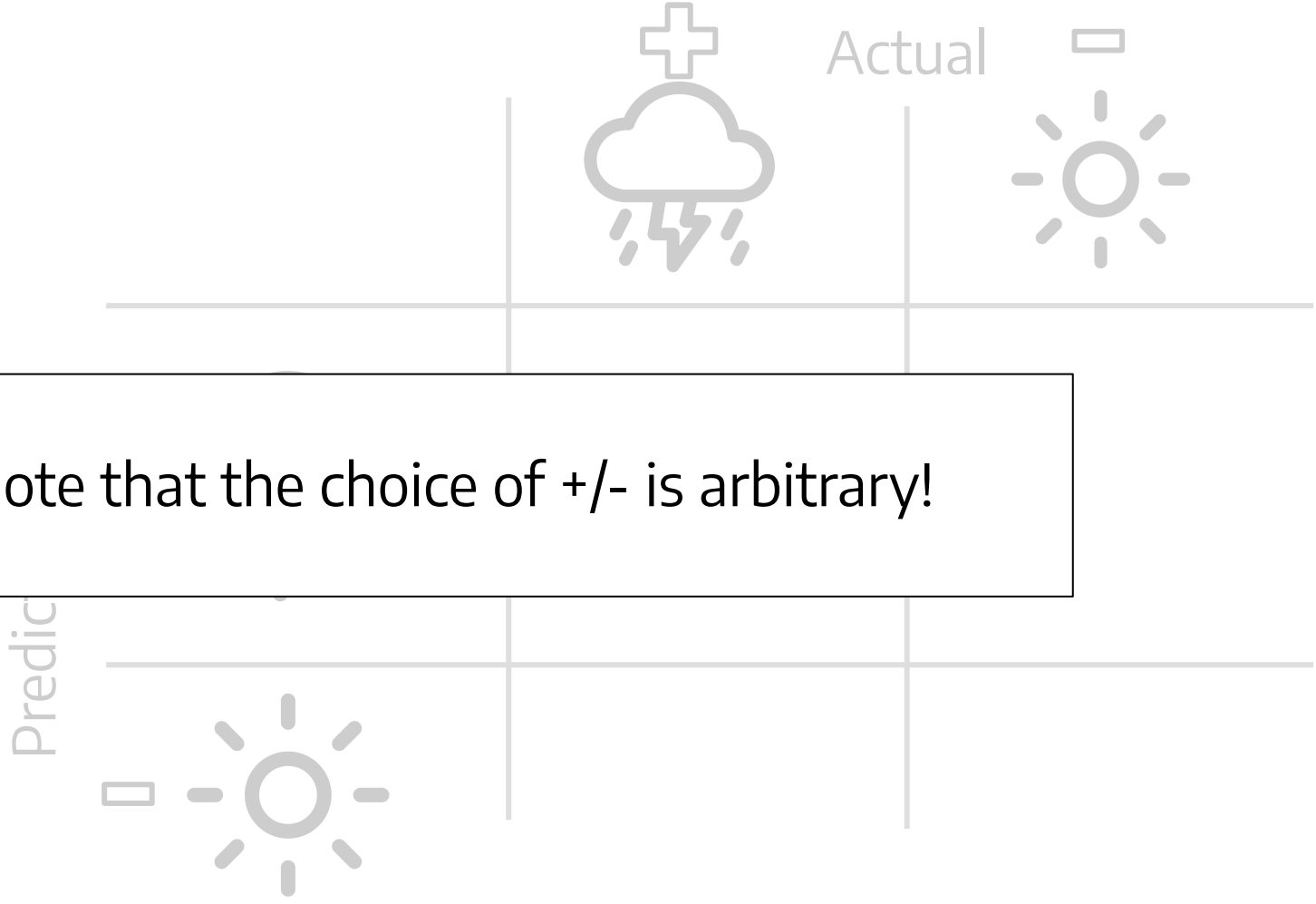
- Classifier
  - Binary
  - Multi-Class
- Evaluation Metric
- Label / target / outcome
- Actual and Predicted Values

Classification Outcomes

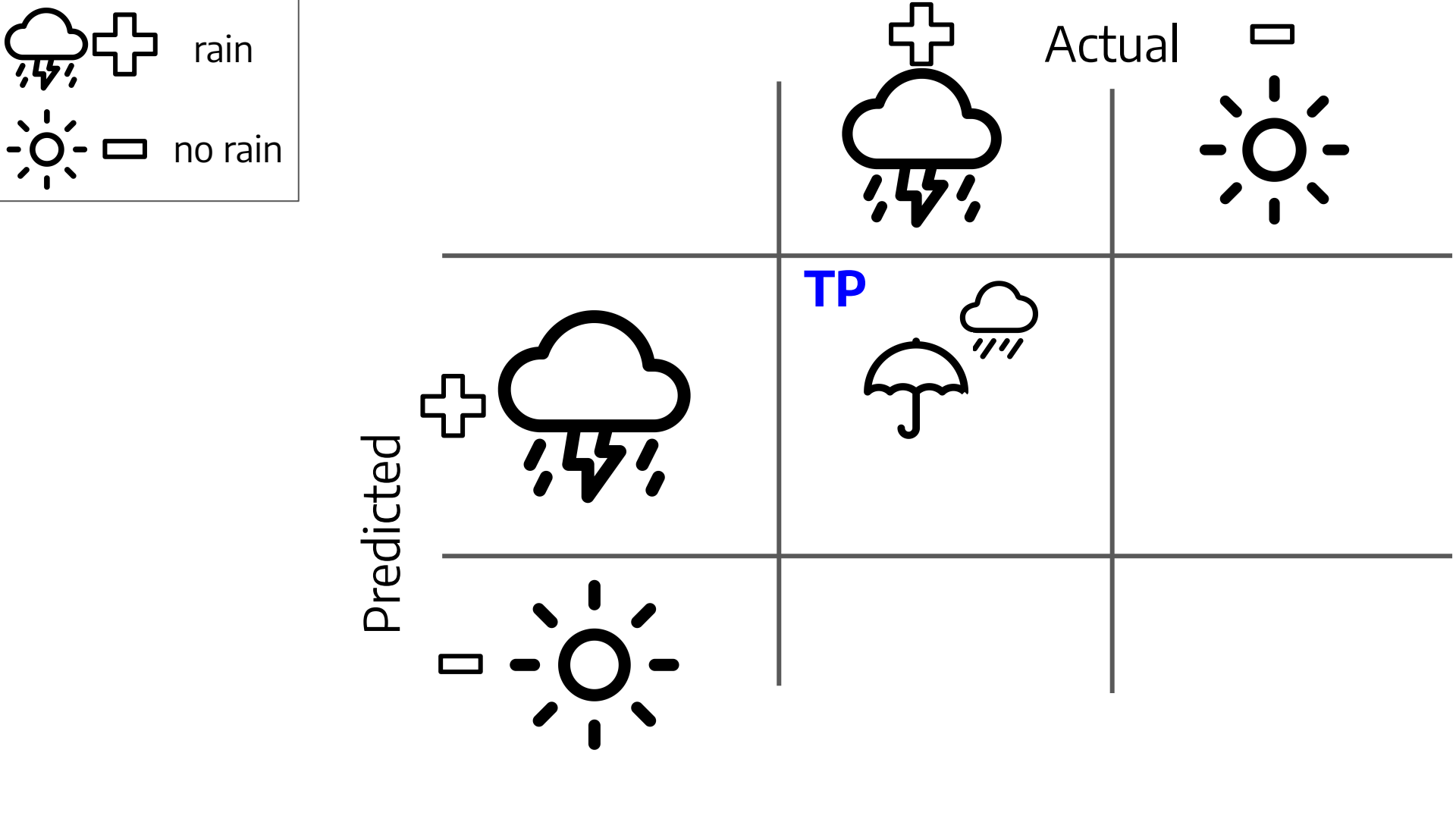
- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)

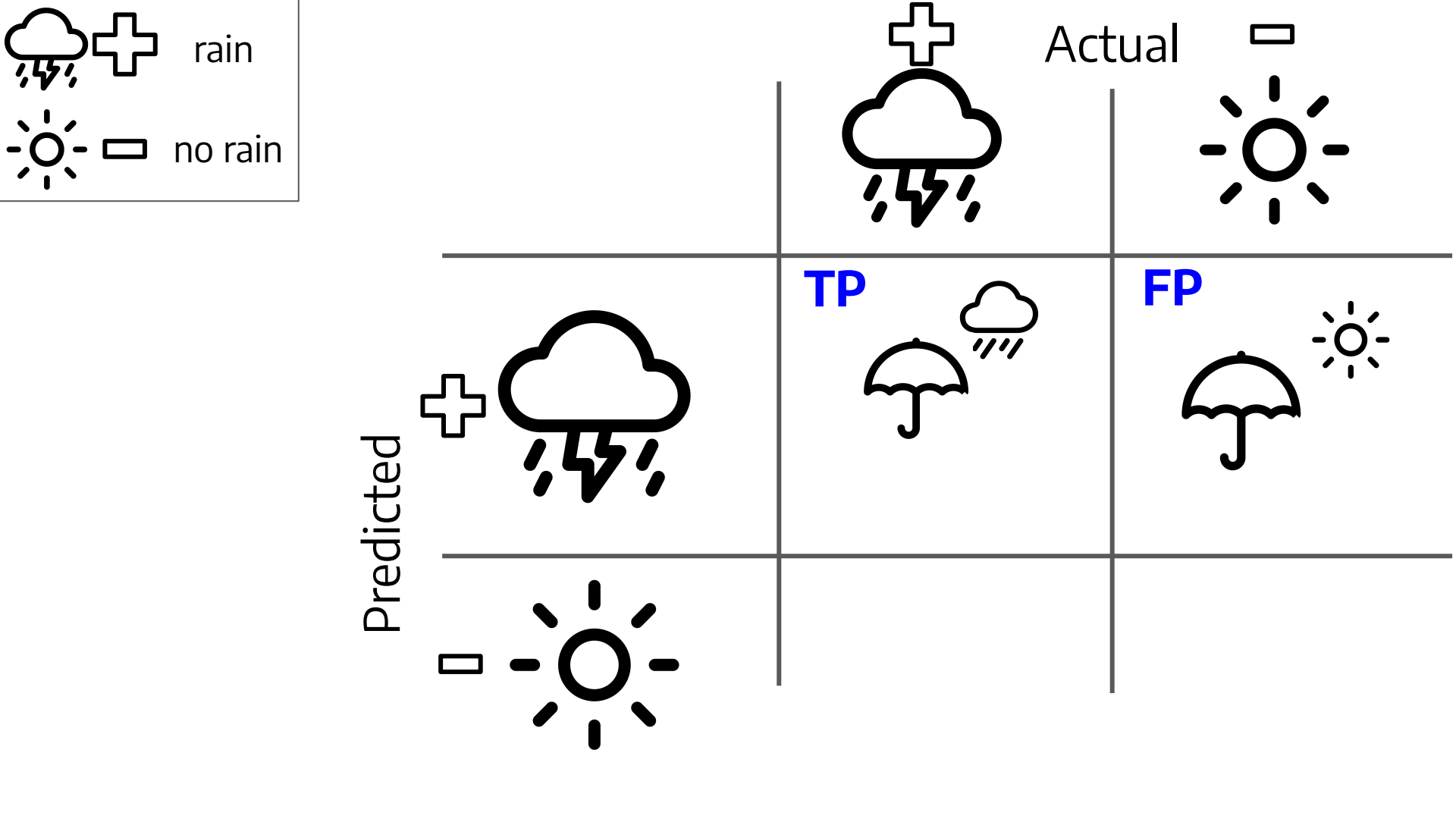

rain
no rain
Actual
Predicted

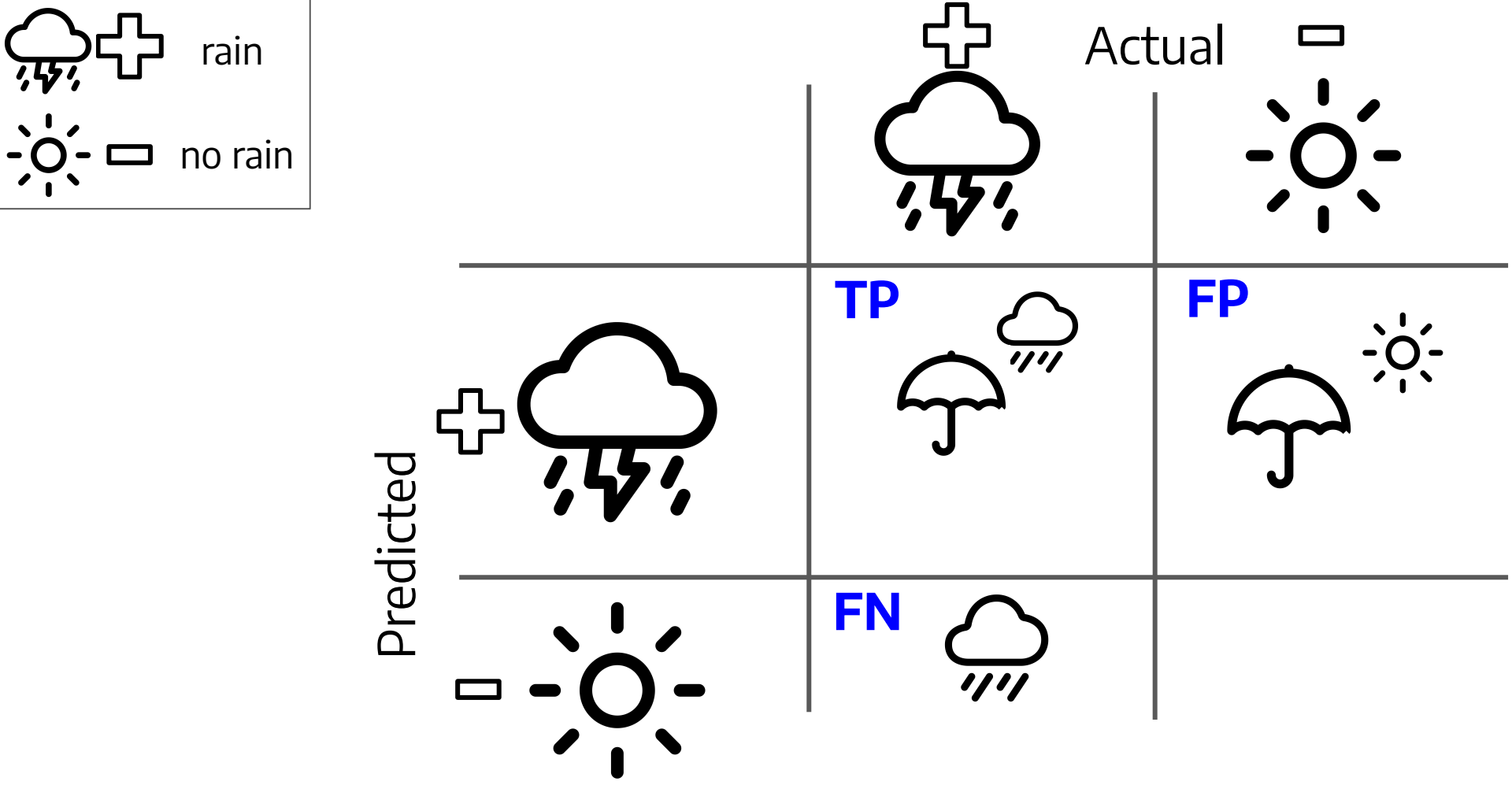


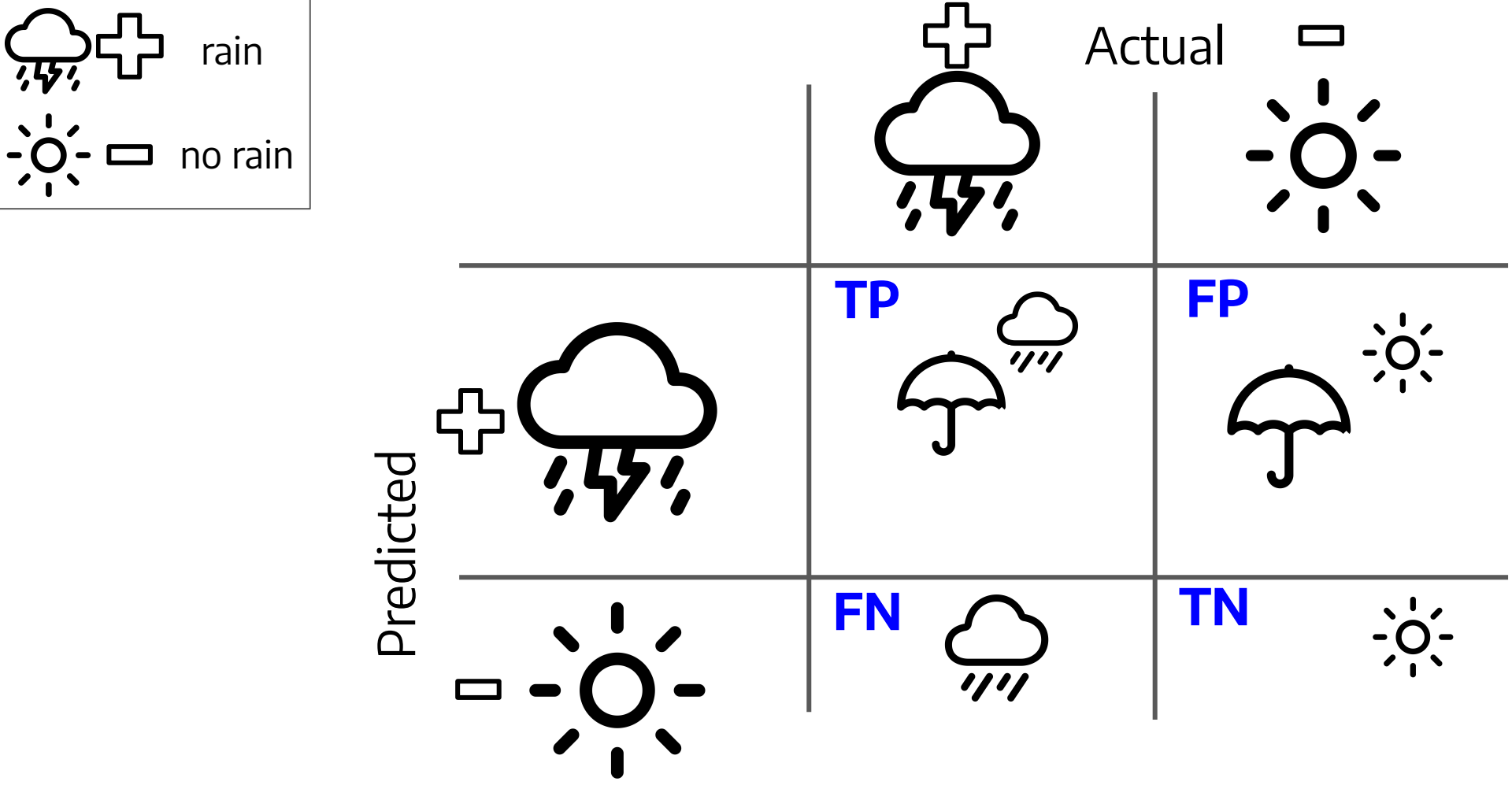


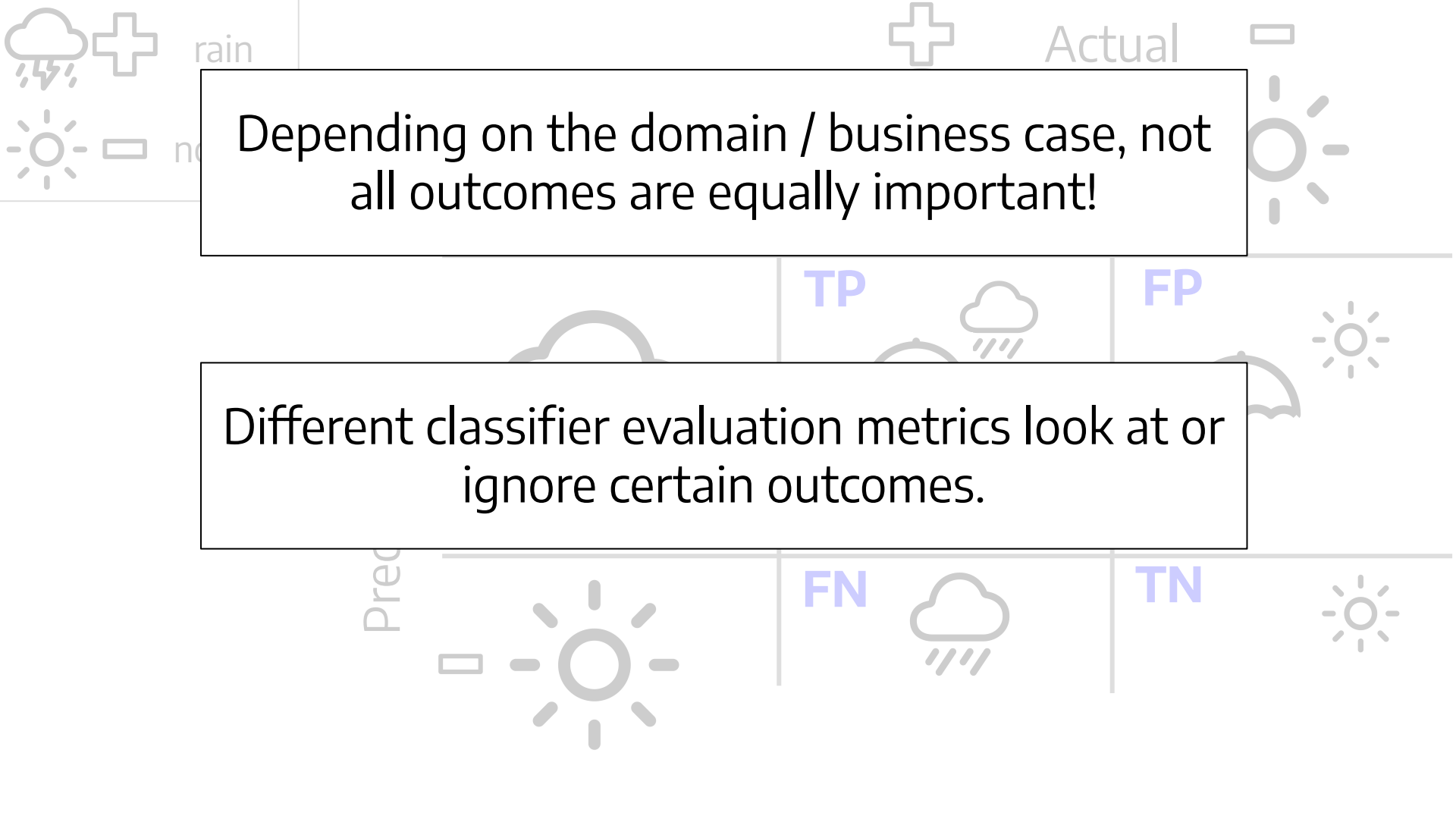
Note that the choice of +/- is arbitrary!


rain
no rain
Actual
Predicted
TP


rain
no rain
Actual
Predicted
TP
FP

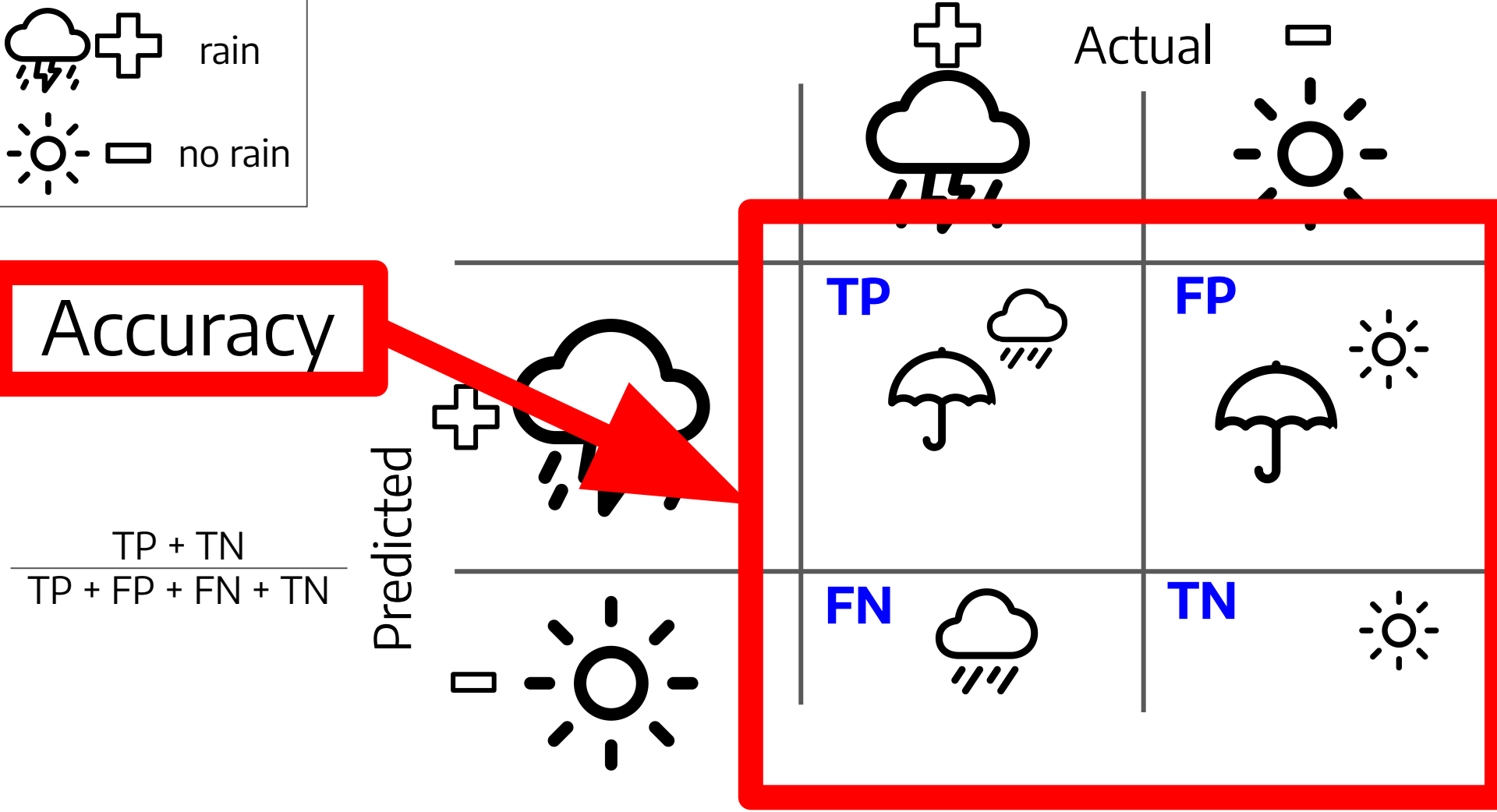

rain
no rain
Actual
Predicted
TP
FP
FN

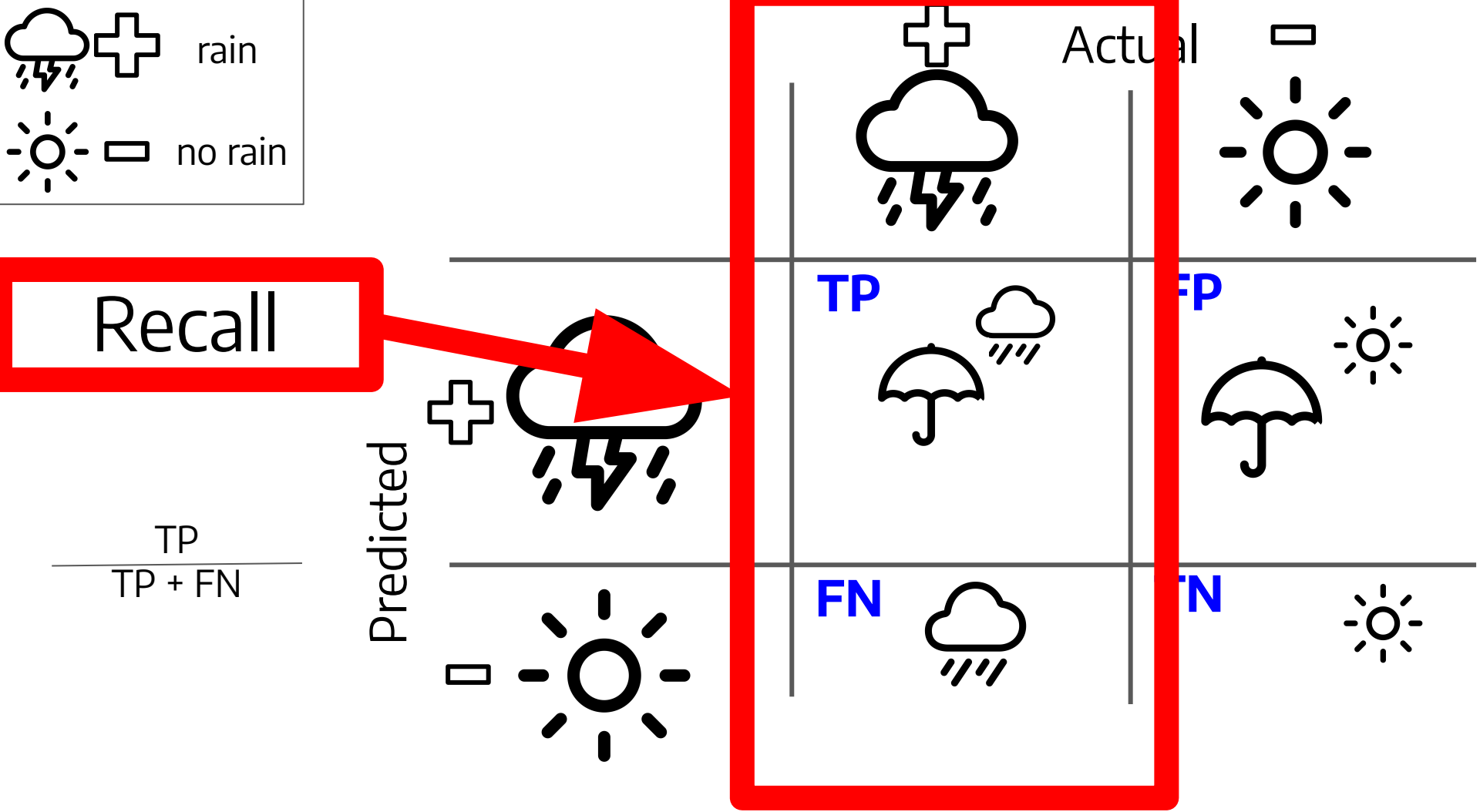

rain
no rain
Actual
Predicted
TP
FP
FN
TN

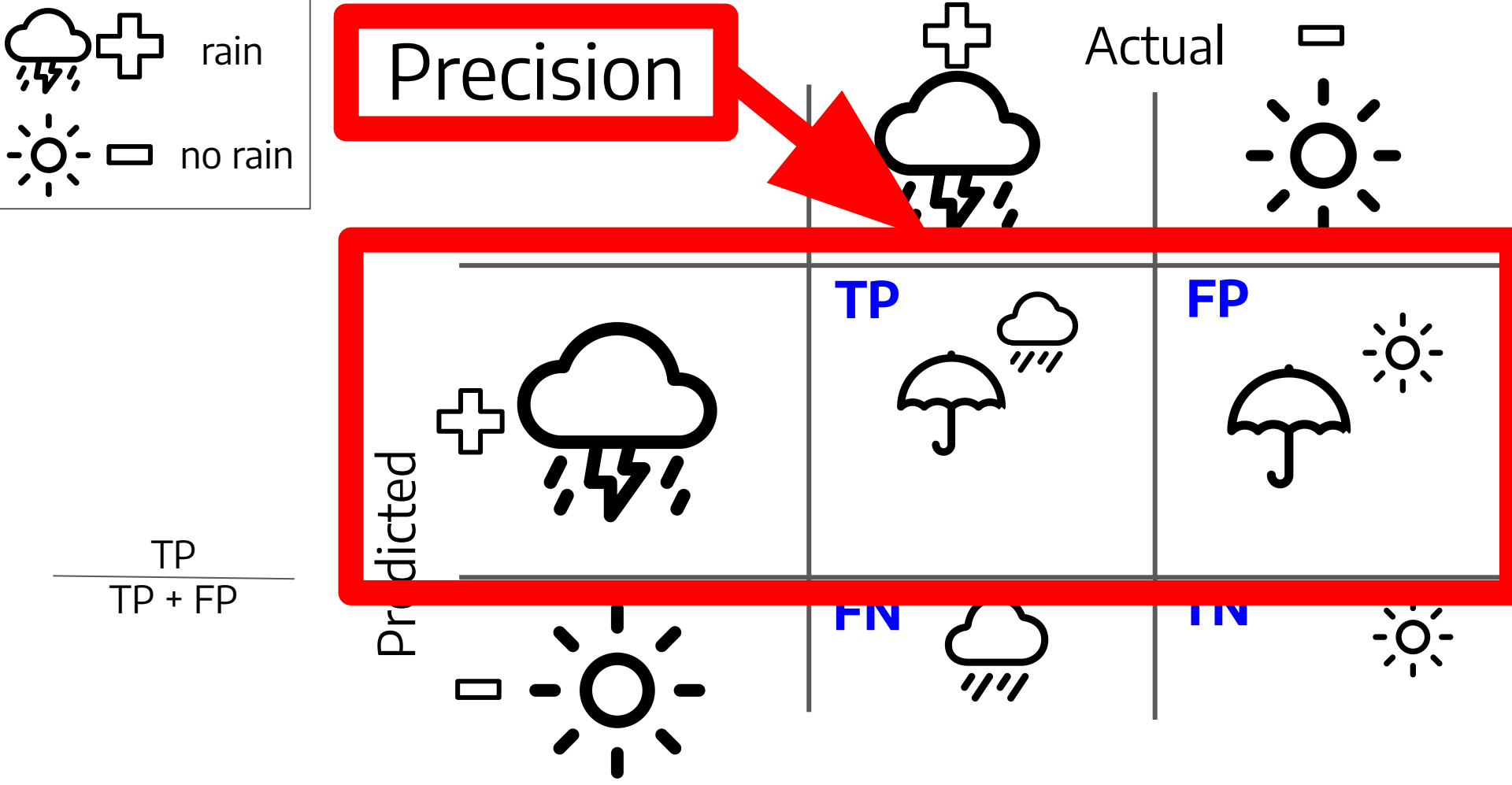


Depending on the domain / business case, not all outcomes are equally important!

Different classifier evaluation metrics look at or ignore certain outcomes.


rain
no rain
Actual
Predicted
TP
FP
FN
TN
Accuracy
TP + TN
TP + FP + FN + TN


rain
no rain
Recall
Actual
Predicted
TP
FP
FN
TN
TP
TP + FN

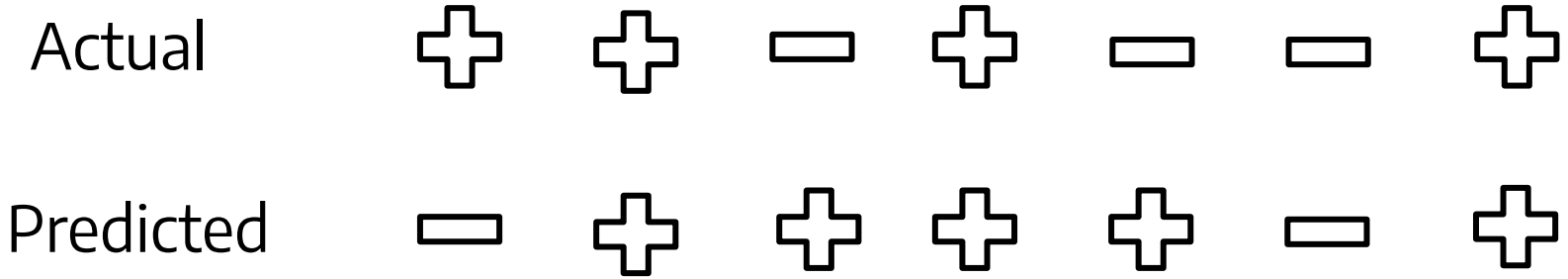


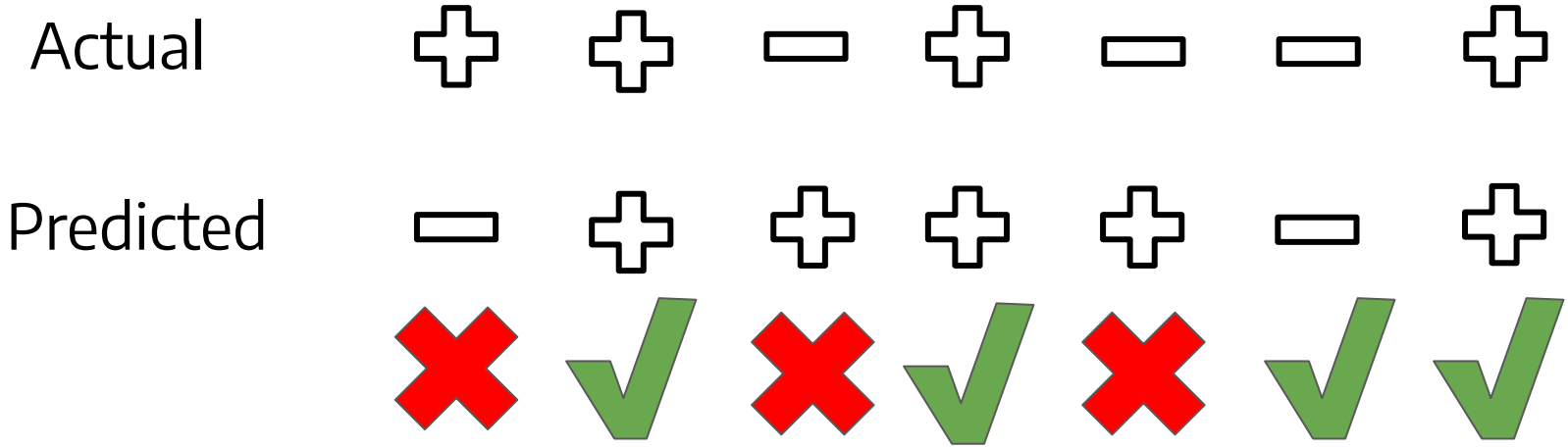
$$\frac{TP}{TP + FP}$$

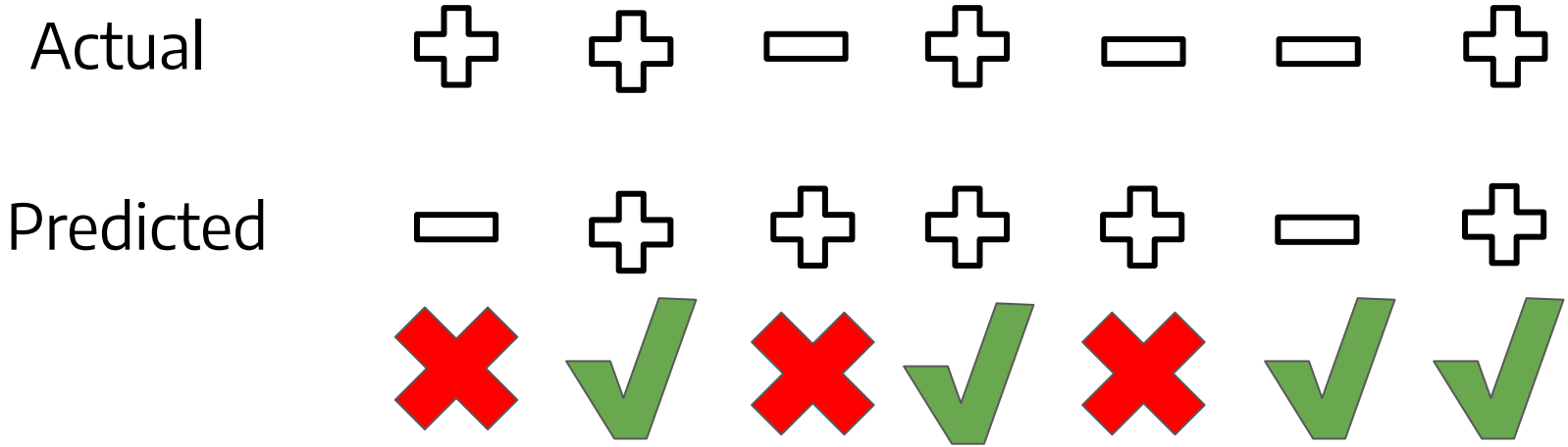
Actual

Predicted

Let's look at an example...


Actual
Predicted

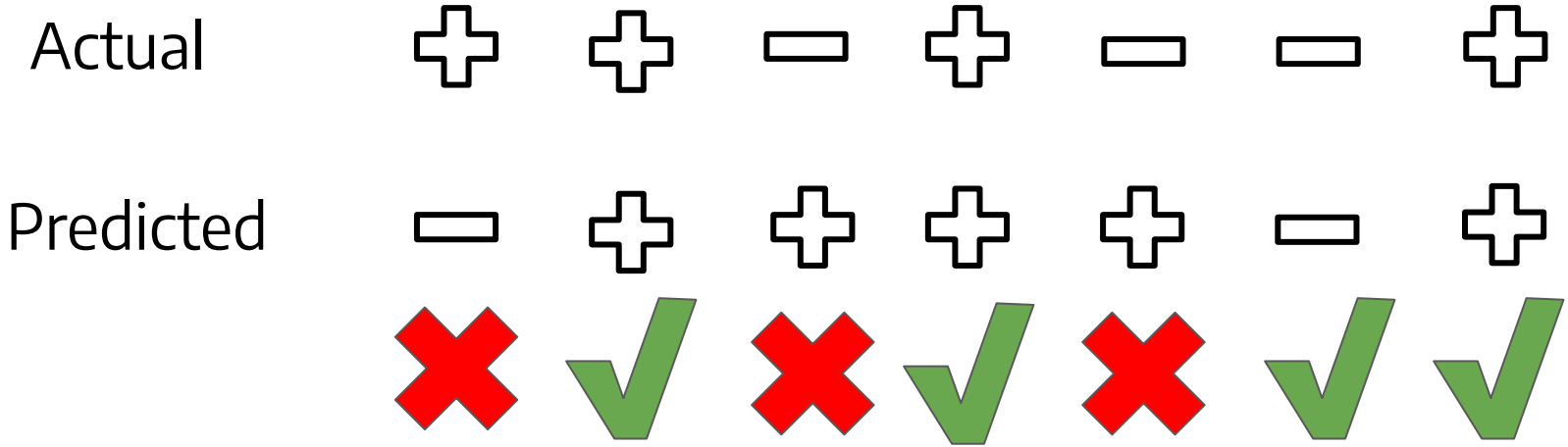

Actual
Predicted



Accuracy

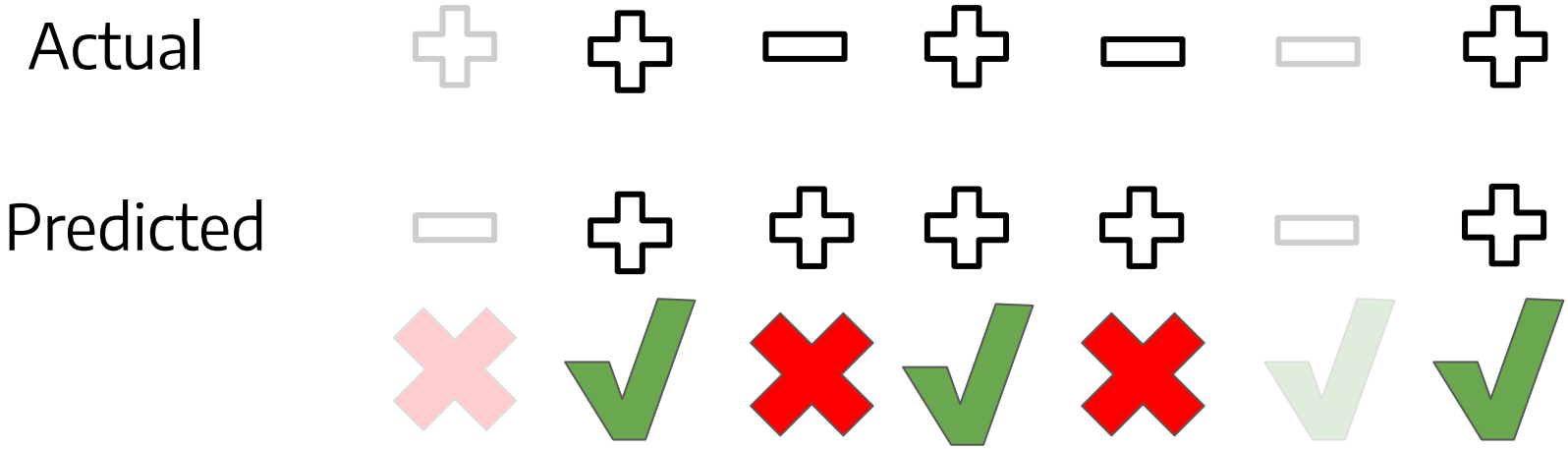
correct / total

= 4 / 7 ≈ 57%



Precision

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

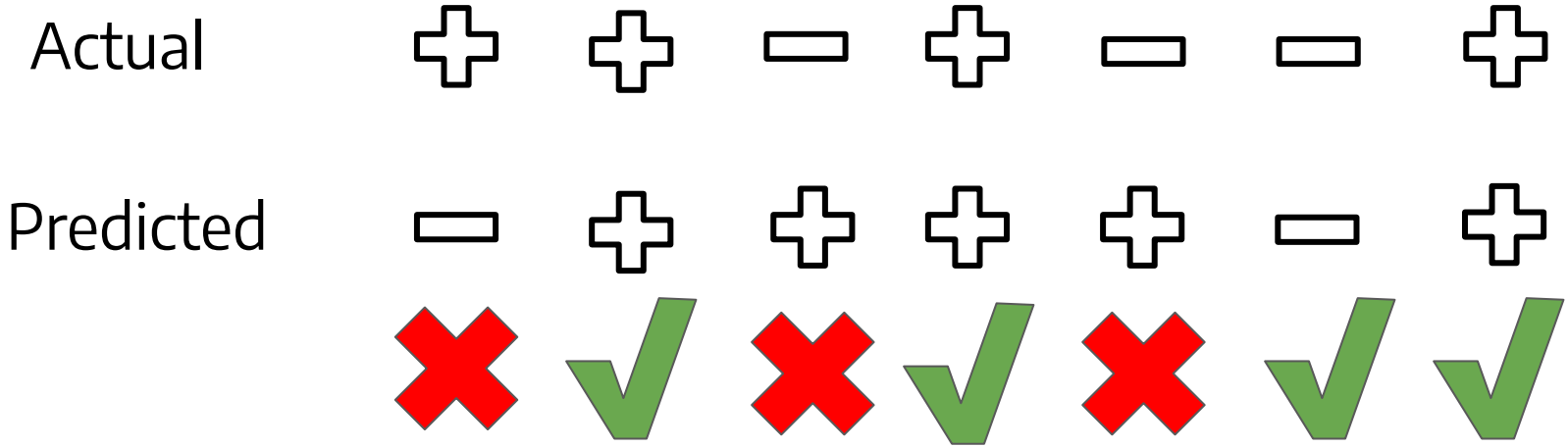


Precision

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

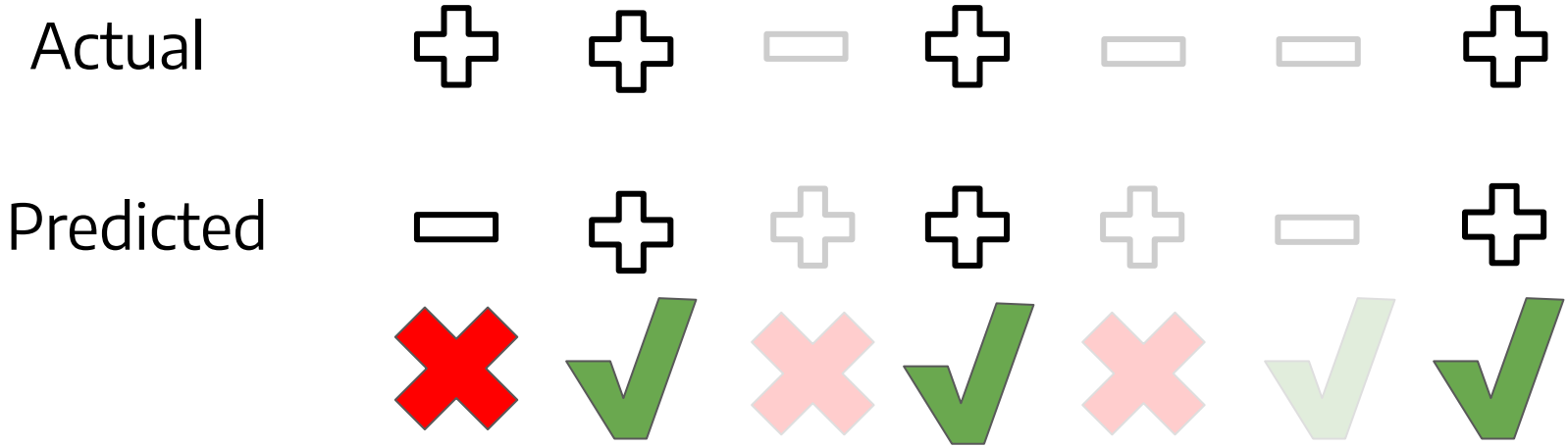
= 3 / 5 = 60%

How good are our positive predictions?



Recall

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

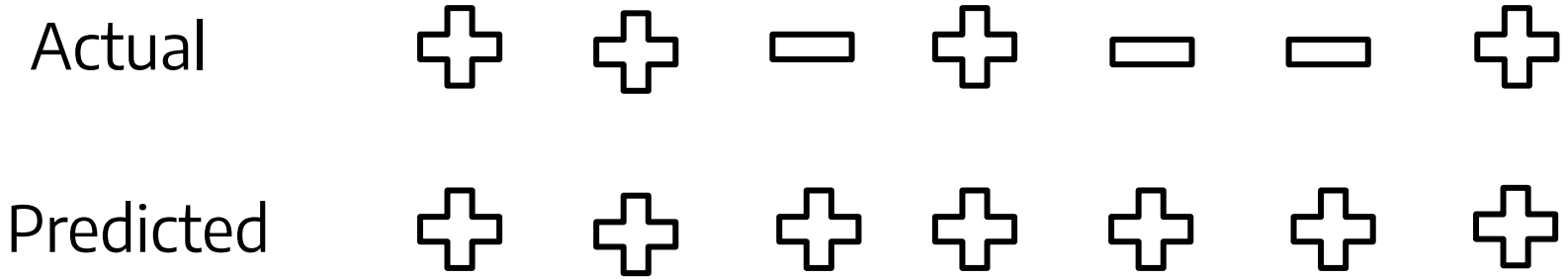


Recall

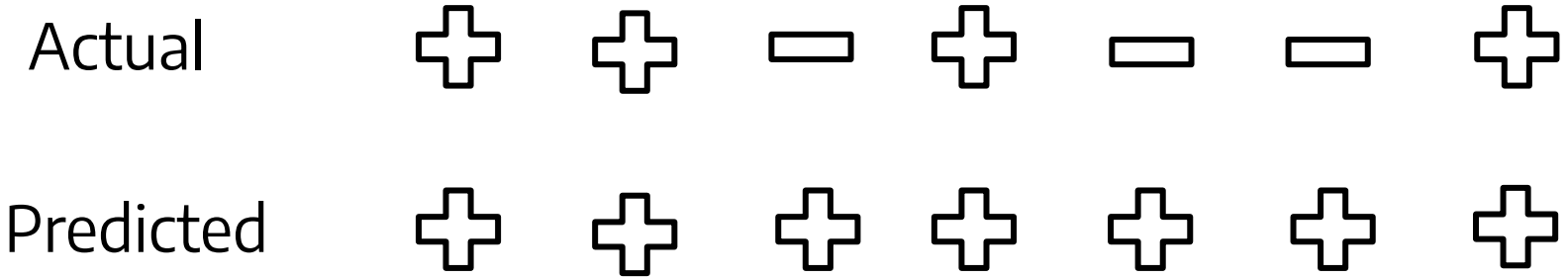
$$\frac{TP}{TP + FN}$$

= 3 / 4 = 75%

How many of the actually positive cases do we catch?



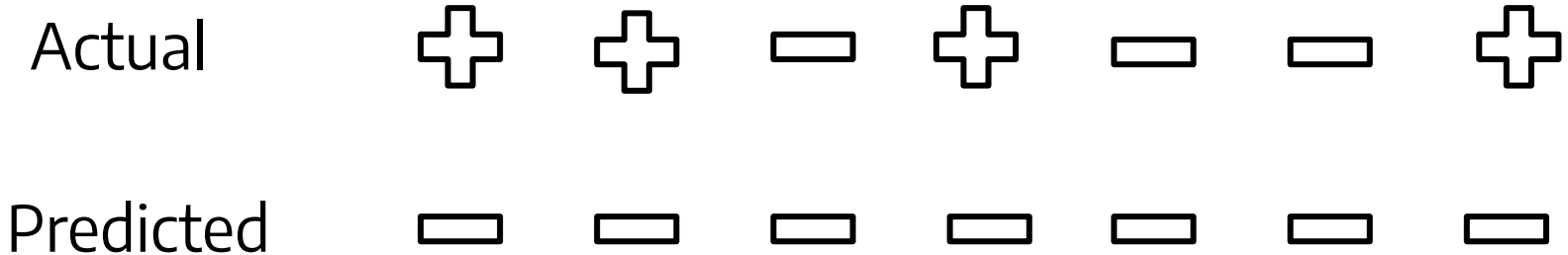
Consider a classifier that always predicts positive…



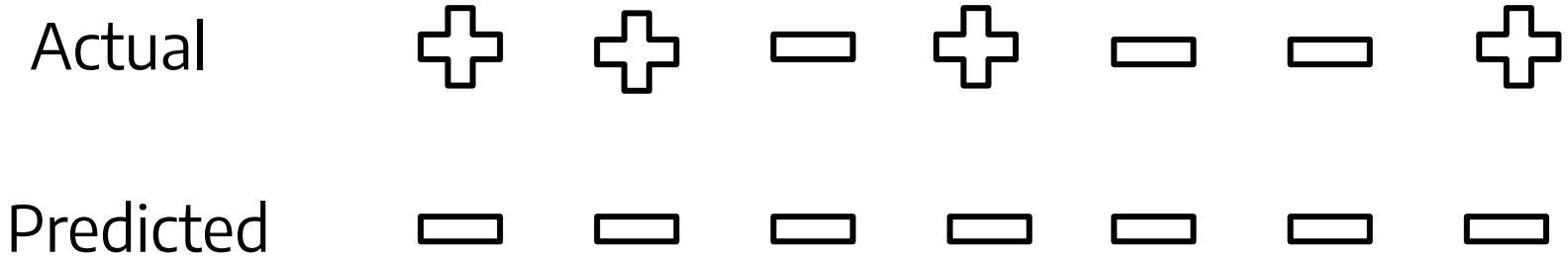
Accuracy = 4 / 7 = 57%

Recall = 4 / 4 = 100%

Precision = 4 / 7 = 47%



What if we always predict negative?



Accuracy = 3 / 7 = 43%

Recall = 0 / 4 = 0%

Precision = 0 / 0 = undef

Classifier Evaluation Metrics

Recap

- Accuracy: Overall Performance
- Recall: When we don't want to "miss out" on an actually positive case
- Precision: When a positive prediction is expensive

Other Metrics

- Sensitivity: aka recall
- Specificity: recall for the negative class
- F1 score: harmonic mean of precision and recall
- ROC / AUC