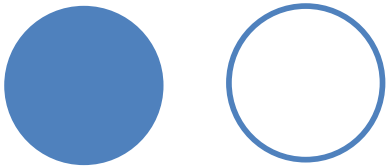
Classification Evaluation: the 2-by-2 contingency table

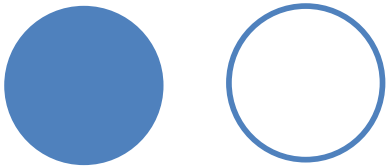
	Actually Correct	Actually Incorrect
Selected/ Guessed		
Not selected/ not guessed		



Classes/Choices

Classification Evaluation: the 2-by-2 contingency table

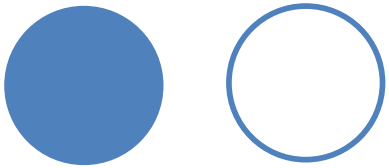
	Actually Correct	**Actually Incorrect**
Selected/ Guessed	True Positive (TP) 	
Not selected/ not guessed		



Classes/Choices

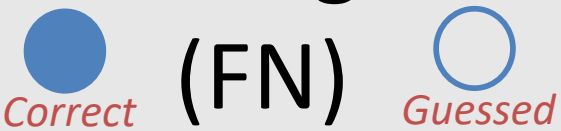
Classification Evaluation: the 2-by-2 contingency table

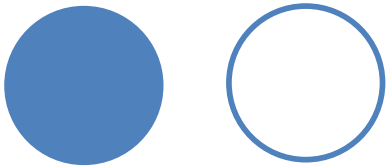
	Actually Correct	**Actually Incorrect**
Selected/ Guessed	True Positive 	False Positive 
Not selected/ not guessed		



Classes/Choices

Classification Evaluation: the 2-by-2 contingency table

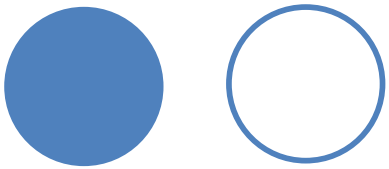
	Actually Correct	**Actually Incorrect**
Selected/ Guessed	True Positive 	False Positive 
Not selected/ not guessed	False Negative 	



Classes/Choices

Classification Evaluation: the 2-by-2 contingency table

	Actually Correct	**Actually Incorrect**
Selected/ Guessed	True Positive 	False Positive 
Not selected/ not guessed	False Negative 	True Negative 



Classes/Choices

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

	Actually Correct	Actually Incorrect
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

Precision: % of selected items that are correct

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

	Actually Correct	Actually Incorrect
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

Precision: % of selected items that are correct

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

Recall: % of correct items that are selected

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

	Actually Correct	Actually Incorrect
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Classification Evaluation: Accuracy, Precision, and Recall

Accuracy: % of items correct

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

Precision: % of selected items that are correct

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

Min: 0 ☹️
Max: 1 😀

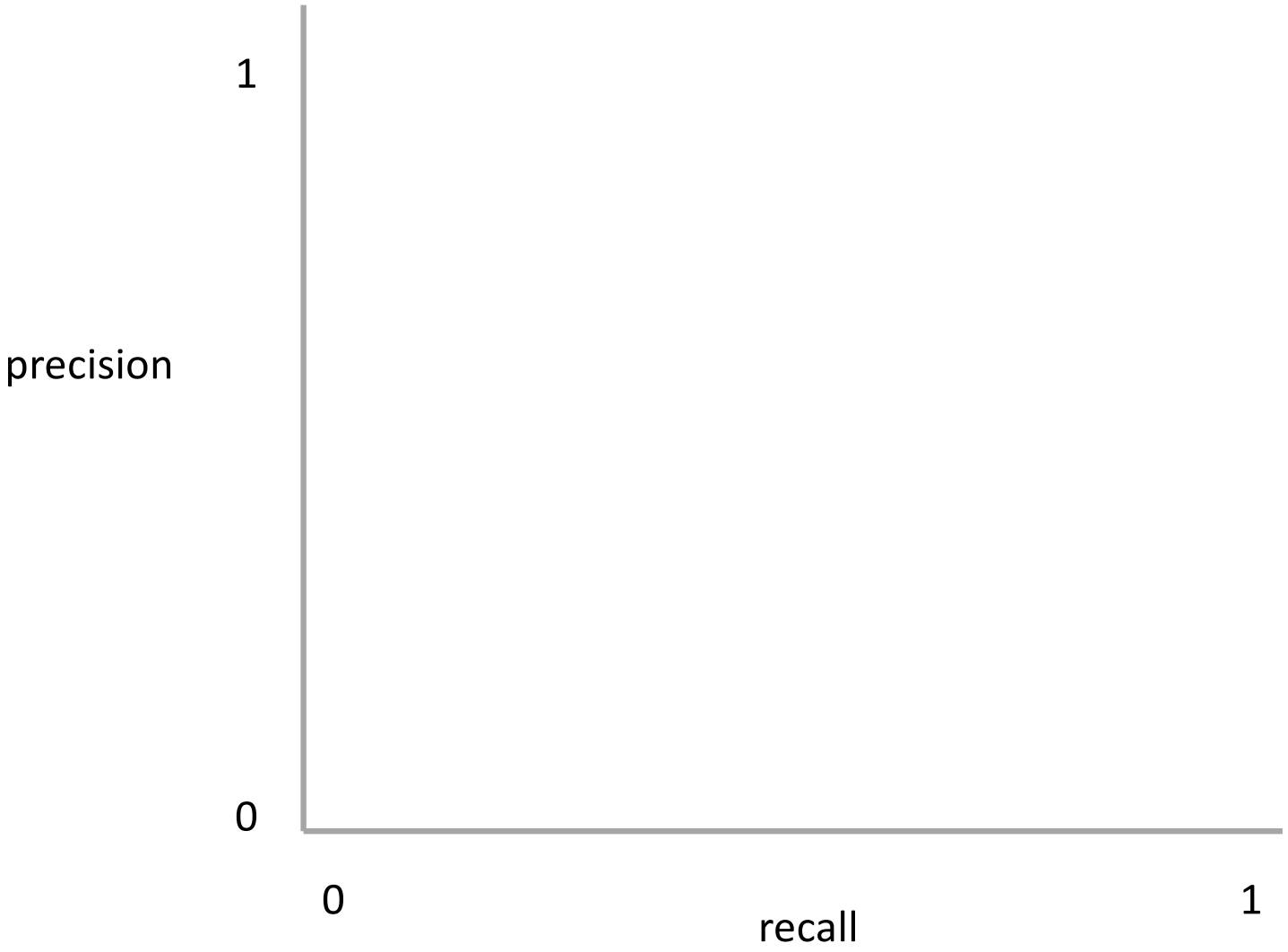
Recall: % of correct items that are selected

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

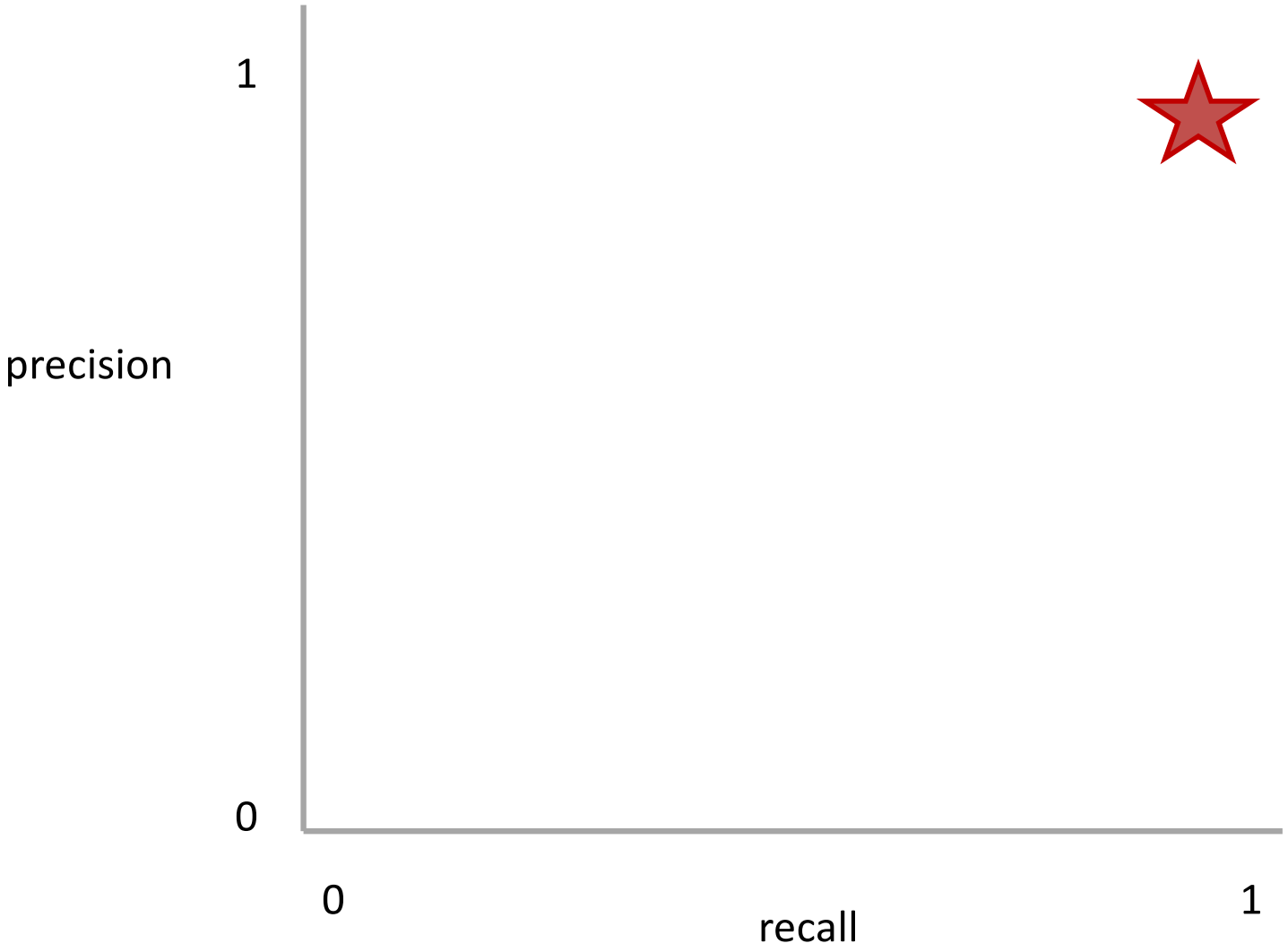
	Actually Correct	Actually Incorrect
Selected/Guessed	True Positive (TP)	False Positive (FP)
Not select/not guessed	False Negative (FN)	True Negative (TN)

Precision and Recall Present a Tradeoff

Q: Where do you want your ideal model ?



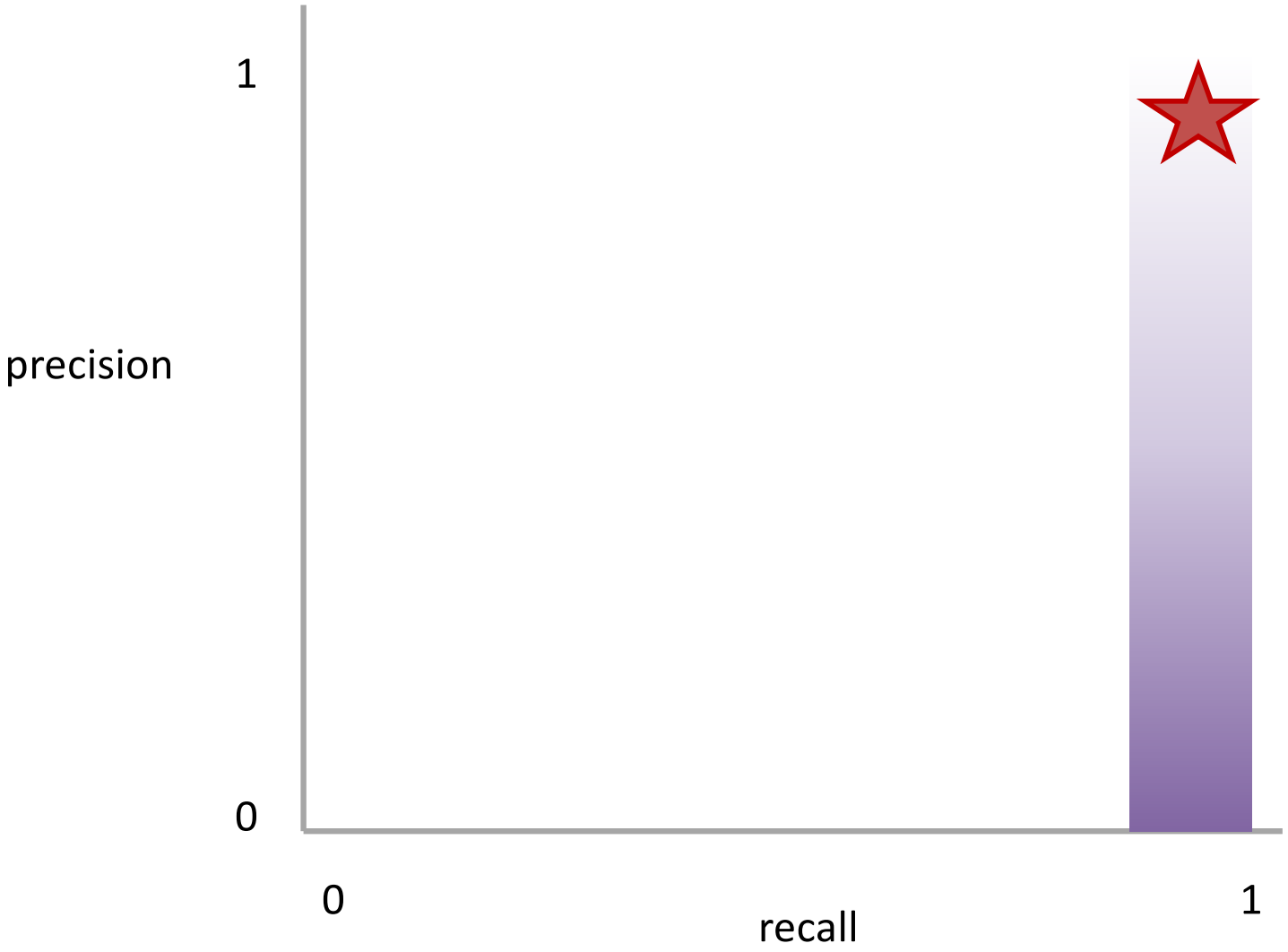
Precision and Recall Present a Tradeoff



Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Precision and Recall Present a Tradeoff

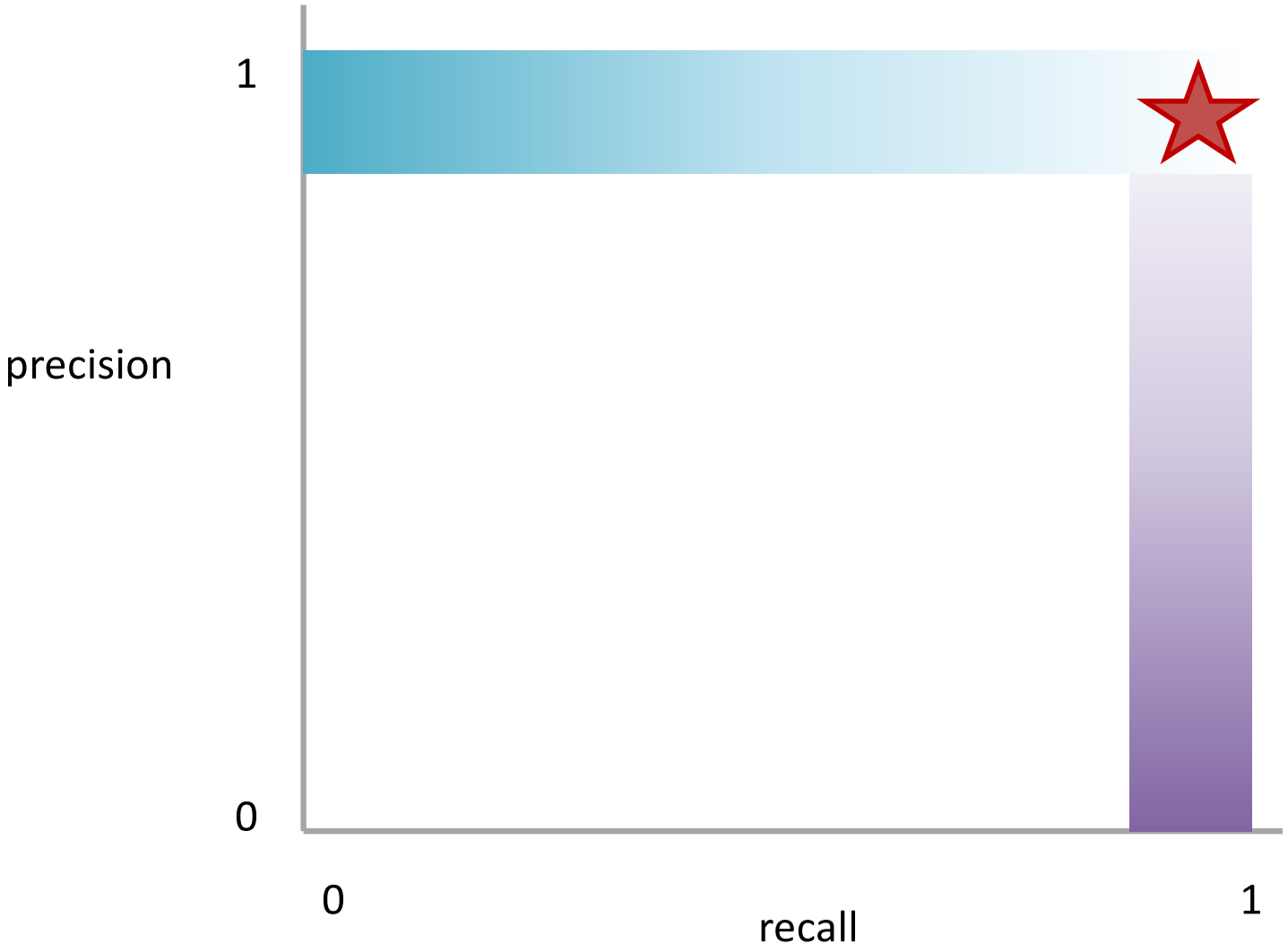


Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Precision and Recall Present a Tradeoff

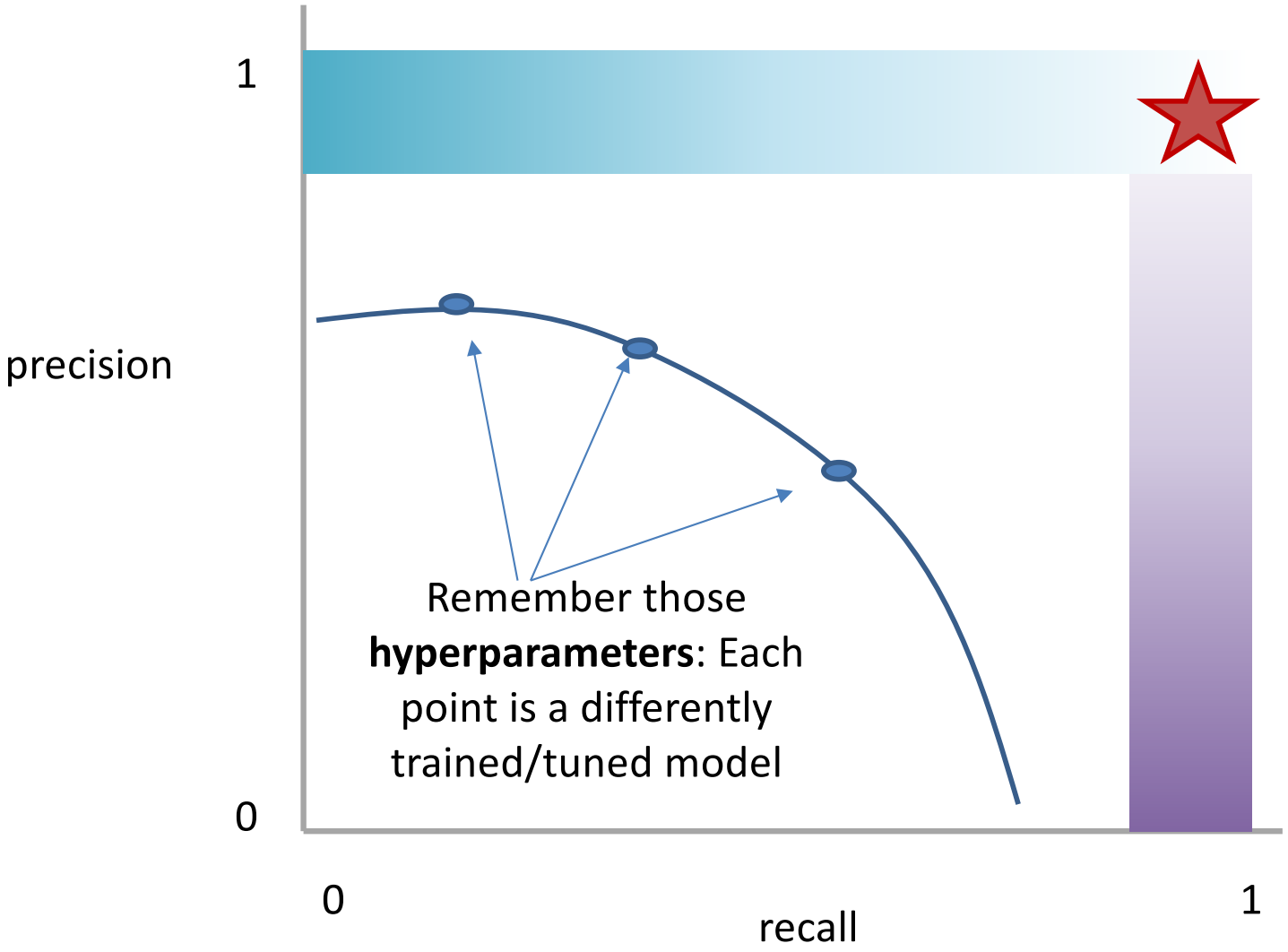


Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Precision and Recall Present a Tradeoff



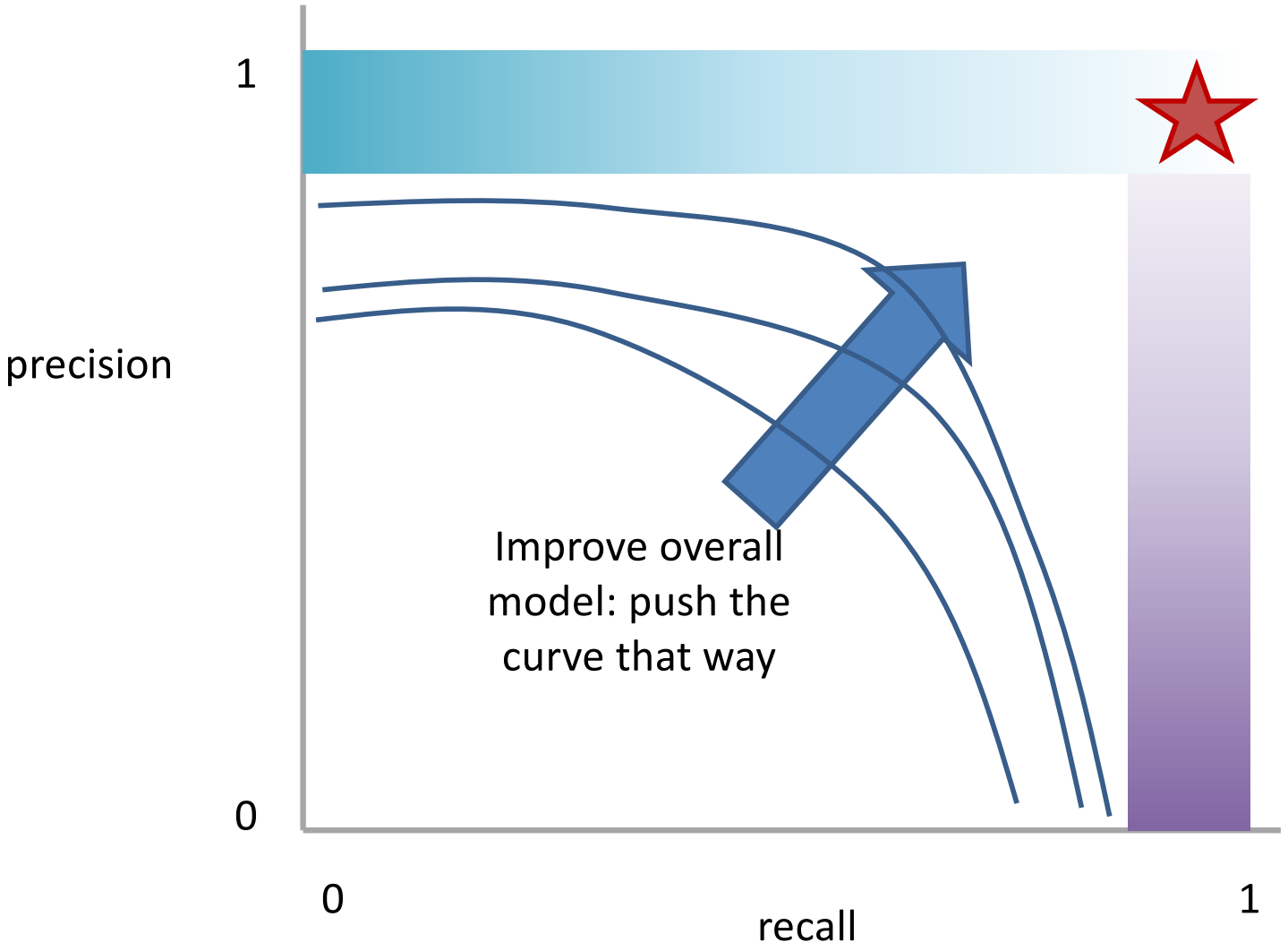
Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Idea: measure the tradeoff between precision and recall

Precision and Recall Present a Tradeoff



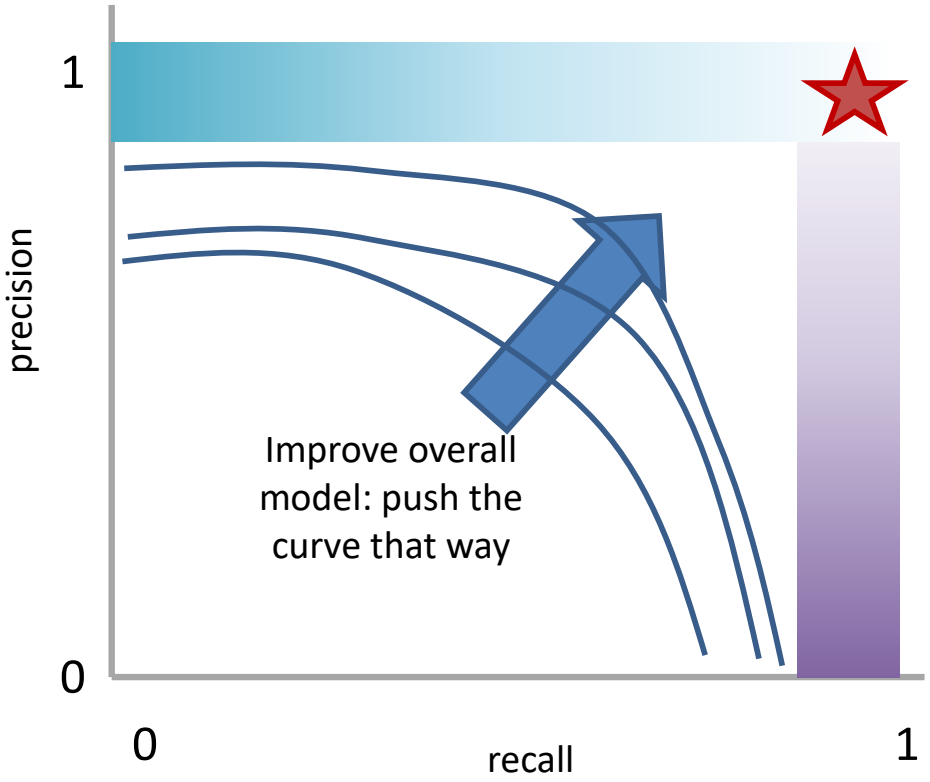
Q: Where do you want your ideal model ?

Q: You have a model that always identifies correct instances. Where on this graph is it?

Q: You have a model that only make correct predictions. Where on this graph is it?

Idea: measure the tradeoff between precision and recall

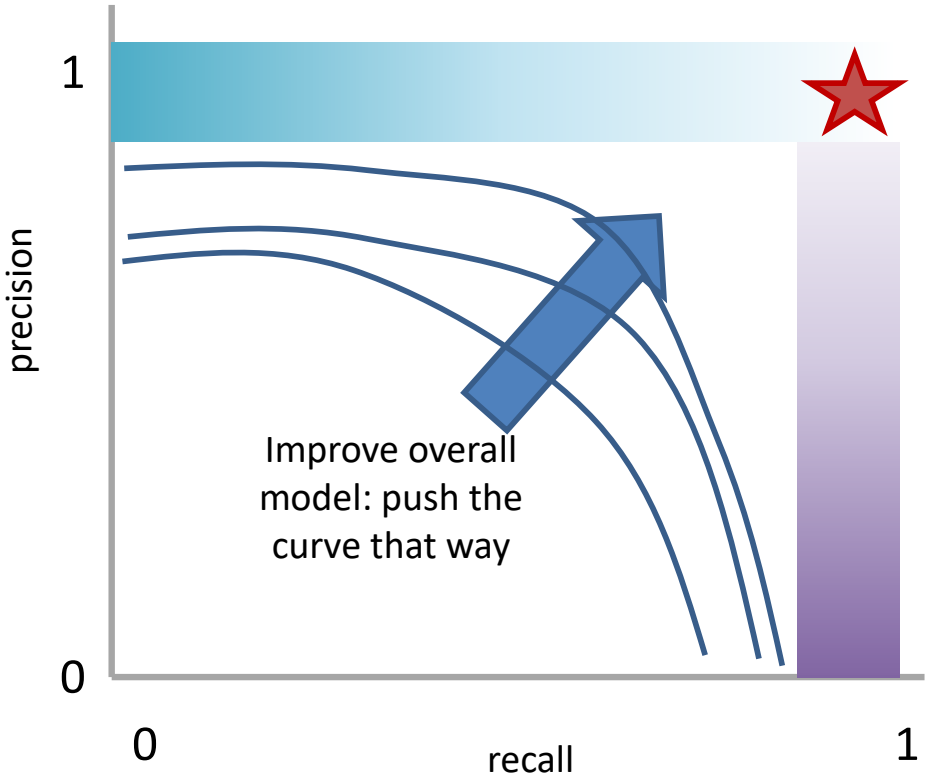
Measure this Tradeoff: Area Under the Curve (AUC)



AUC measures the area under this tradeoff curve

Min AUC: 0 ☹️
Max AUC: 1 😀

Measure this Tradeoff: Area Under the Curve (AUC)



Min AUC: 0 ☹

Max AUC: 1 😀

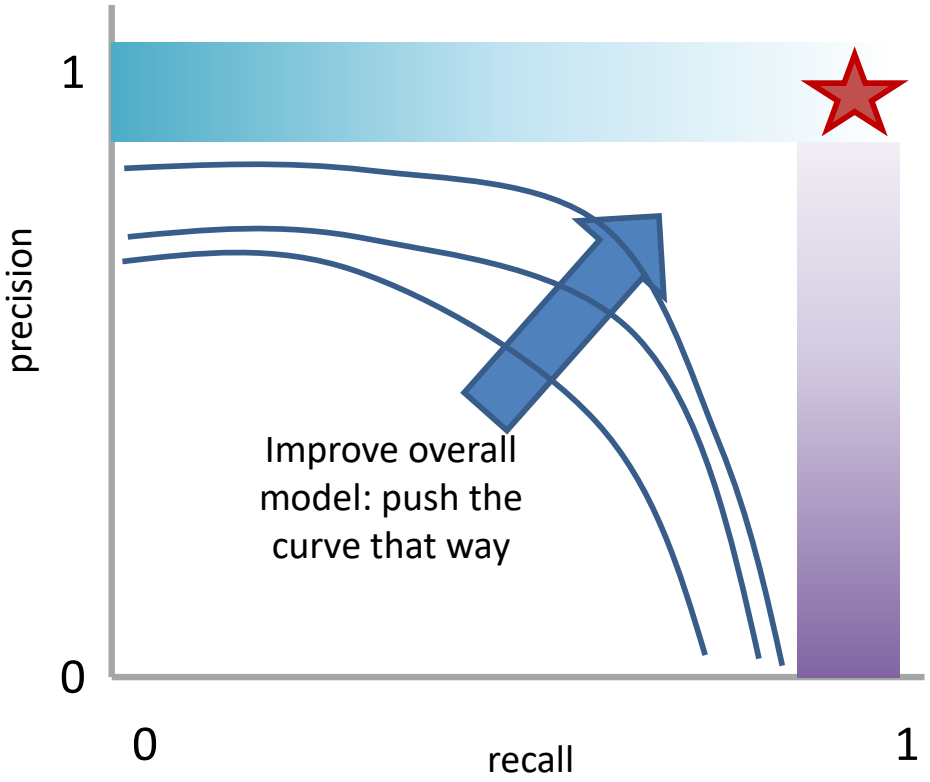
AUC measures the area under this tradeoff curve

1. Computing the curve

   You need true labels & predicted labels with some score/confidence estimate

   Threshold the scores and for each threshold compute precision and recall

Measure this Tradeoff: Area Under the Curve (AUC)



Min AUC: 0 ☹

Max AUC: 1 😀

AUC measures the area under this tradeoff curve

1. Computing the curve

   You need true labels & predicted labels with some score/confidence estimate

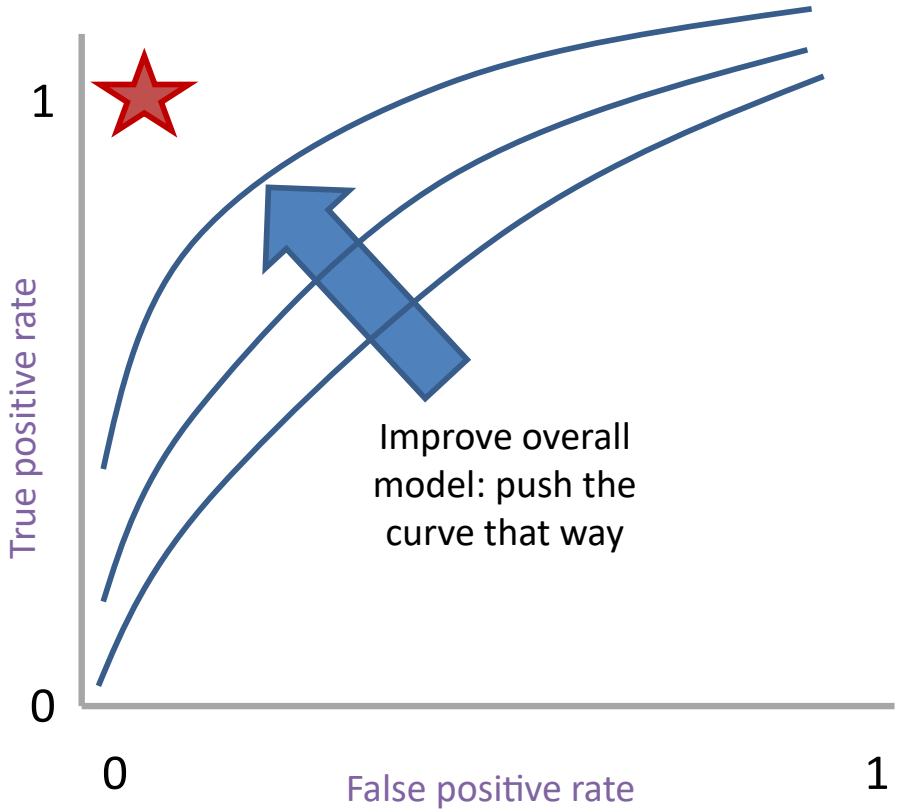
   Threshold the scores and for each threshold compute precision and recall

2. Finding the area

   How to implement: trapezoidal rule (& others)

In practice: external library like the sklearn.metrics module

Measure A Slightly Different Tradeoff: ROC-AUC



Min ROC-AUC: 0.5 ☹

Max ROC-AUC: 1 😀

AUC measures the area under this tradeoff curve

1. Computing the curve
   - You need true labels & predicted labels with some score/confidence estimate
   - Threshold the scores and for each threshold compute metrics
2. Finding the area
   - How to implement: trapezoidal rule (& others)

In practice: external library like the sklearn.metrics module

Main variant: ROC-AUC

Same idea as before but with some flipped metrics

A combined measure: F

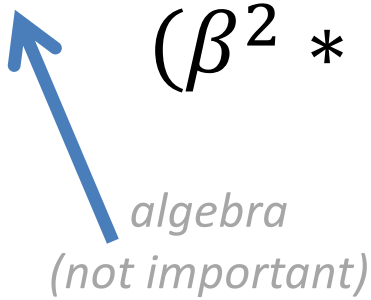
Weighted (harmonic) average of **P**recision & **R**ecall

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

A combined measure: F

Weighted (harmonic) average of **P**recision & **R**ecall

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(1+\beta^2) * P * R}{(\beta^2 * P) + R}$$



A combined measure: F

Weighted (harmonic) average of **P**recision & **R**ecall

$$F = \frac{(1 + \beta^2) * P * R}{(\beta^2 * P) + R}$$

Balanced F1 measure: β=1

$$F_1 = \frac{2 * P * R}{P + R}$$

P/R/F in a Multi-class Setting: Micro- vs. Macro-Averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

Macroaveraging: Compute performance for each class, then average.

Microaveraging: Collect decisions for all classes, compute contingency table, evaluate.