Sentiment analysis
Evaluation

Evaluation (binary | Boolean)

“We’re addicted and annotated data is our heroine.” -?

Gold Labels

aka. Golden Truth, Golden Standard

Human-defined classes/labels for each document

Human judgment

Manually labeled

Manually annotated

~~Gold~~ Labels

aka. Golden Truth, Golden Standard

to err is human!

Gold Labels

aka. Golden Truth, Golden Standard

to err is human; to forgive, divine!

"Save This Word! All people commit sins and make mistakes. God forgives them, and people are acting in a godlike (divine) way when they forgive." - An Essay on Criticism, Alexander Pope.

Silver Labels

aka. Silver Truth, Silver Standard

Gold is very expensive!

Finding gold is needs a lot of effort!

automated-defined classes/labels for each document

Machine judgment

Machine labeled

Machine annotated

Transduction

Transductive Inference

Data has the labels already!

E.g., ?

Transduction

Transductive Inference

Data has the labels already!

Language Models!

Contingency Table
aka. Confusion Matrix

	Gold Positive	Gold Negative
Model Positive	True Positive	False Positive
Model Negative	False Negative	True Negative

Perfect Classifier

	Gold Positive	Gold Negative
Model Positive	N+	0
Model Negative	0	N-

Metrics

A descriptive quantity to show the quality

Single real number (higher better/lower better)
A set of numbers that shows a trend!

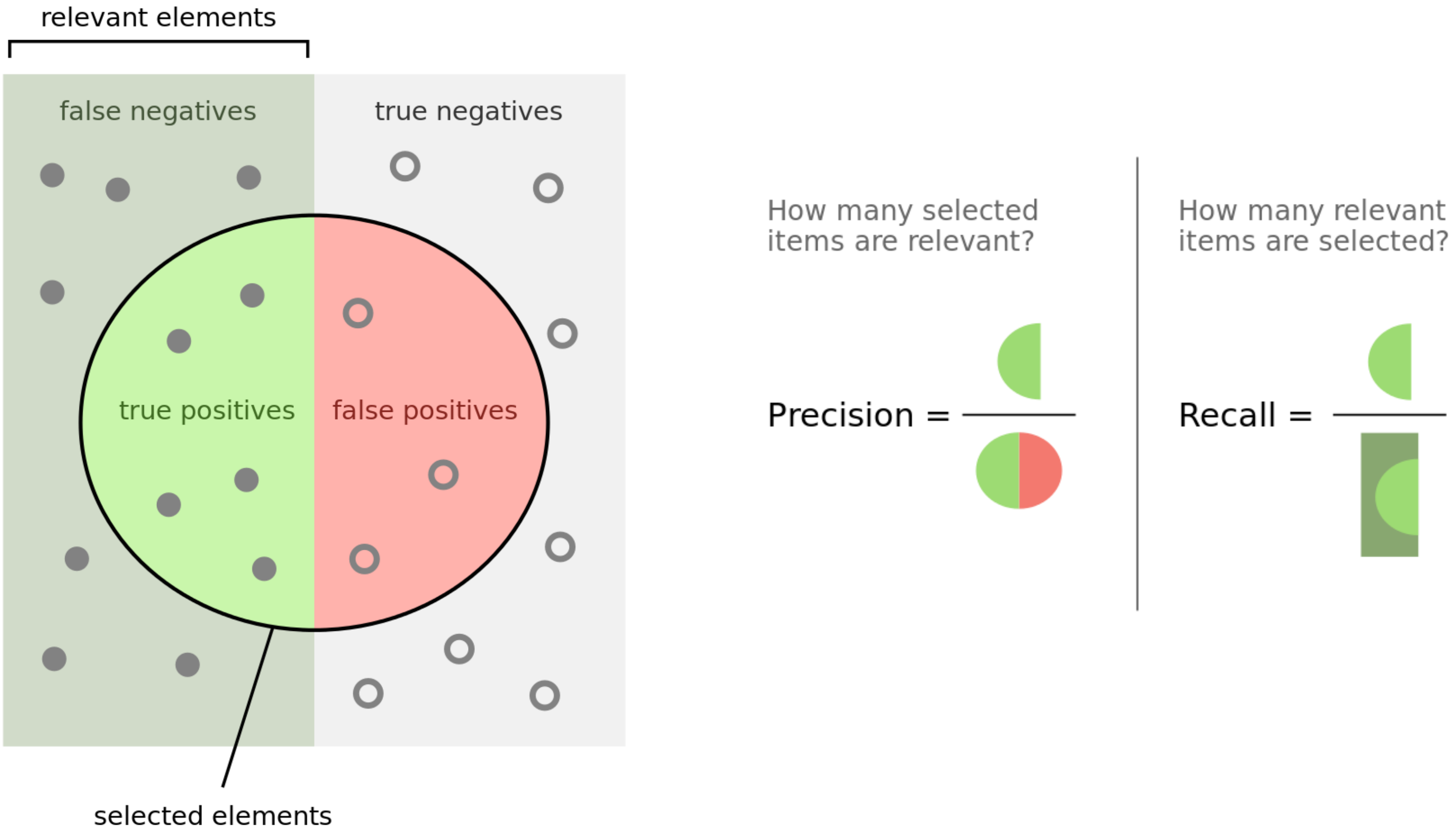
Contingency Table
aka. Confusion Matrix

		gold standard labels		
		gold positive	gold negative	
system output labels	system positive	**true positive**	**false positive**	**precision** $= \frac{\text{tp}}{\text{tp+fp}}$
	system negative	**false negative**	**true negative**	
		recall $= \frac{\text{tp}}{\text{tp+fn}}$		**accuracy** $= \frac{\text{tp+tn}}{\text{tp+fp+tn+fn}}$

Figure 4.4 Contingency table

Contingency Table

aka. Confusion Matrix



Precision

High vs. Low

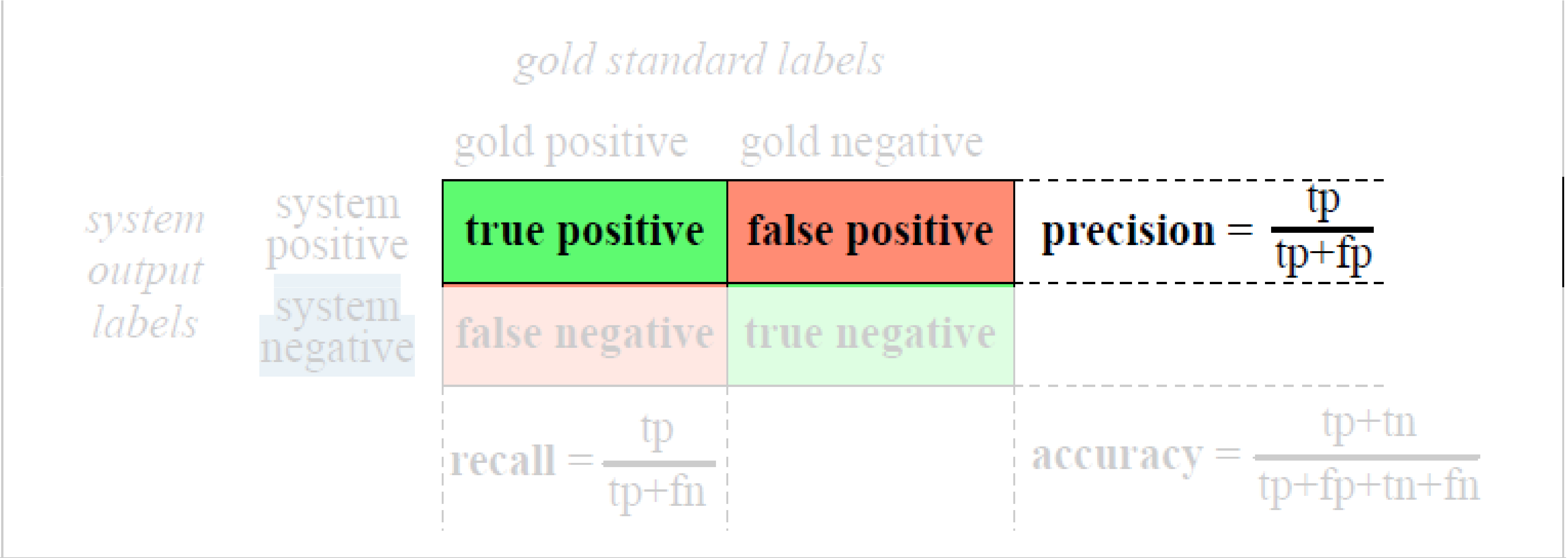


Figure 4.4 Contingency table

What scenarios require high precision?

Recall

High vs. Low

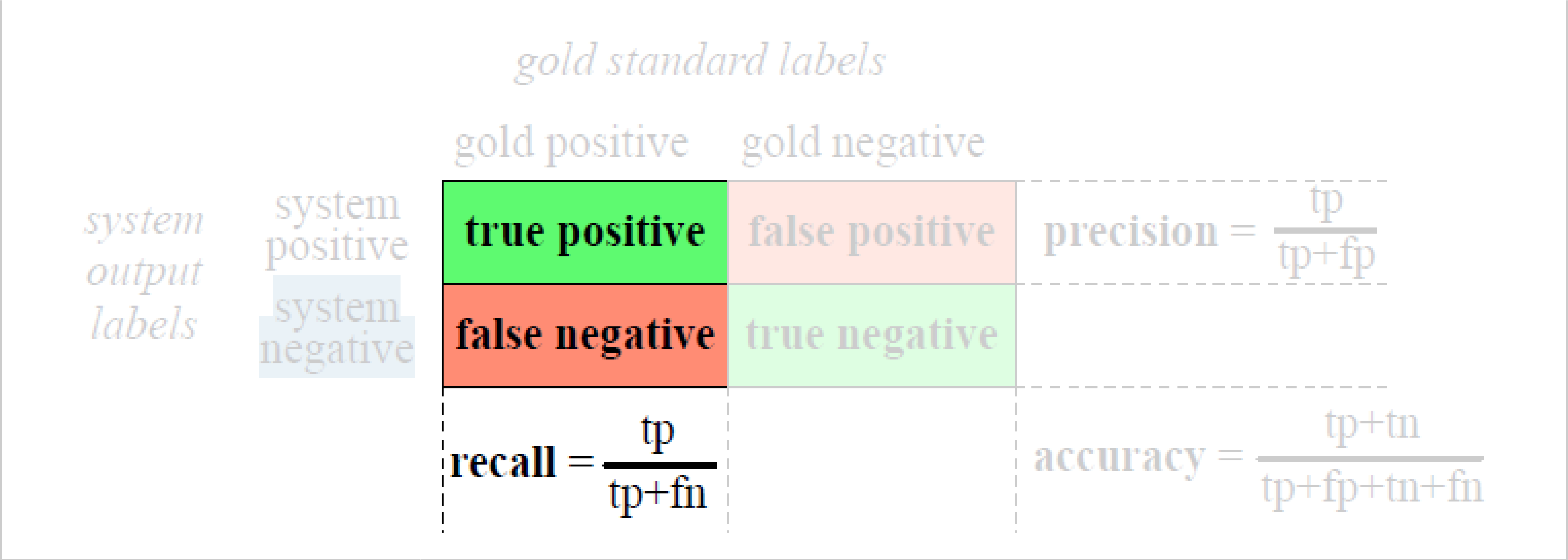
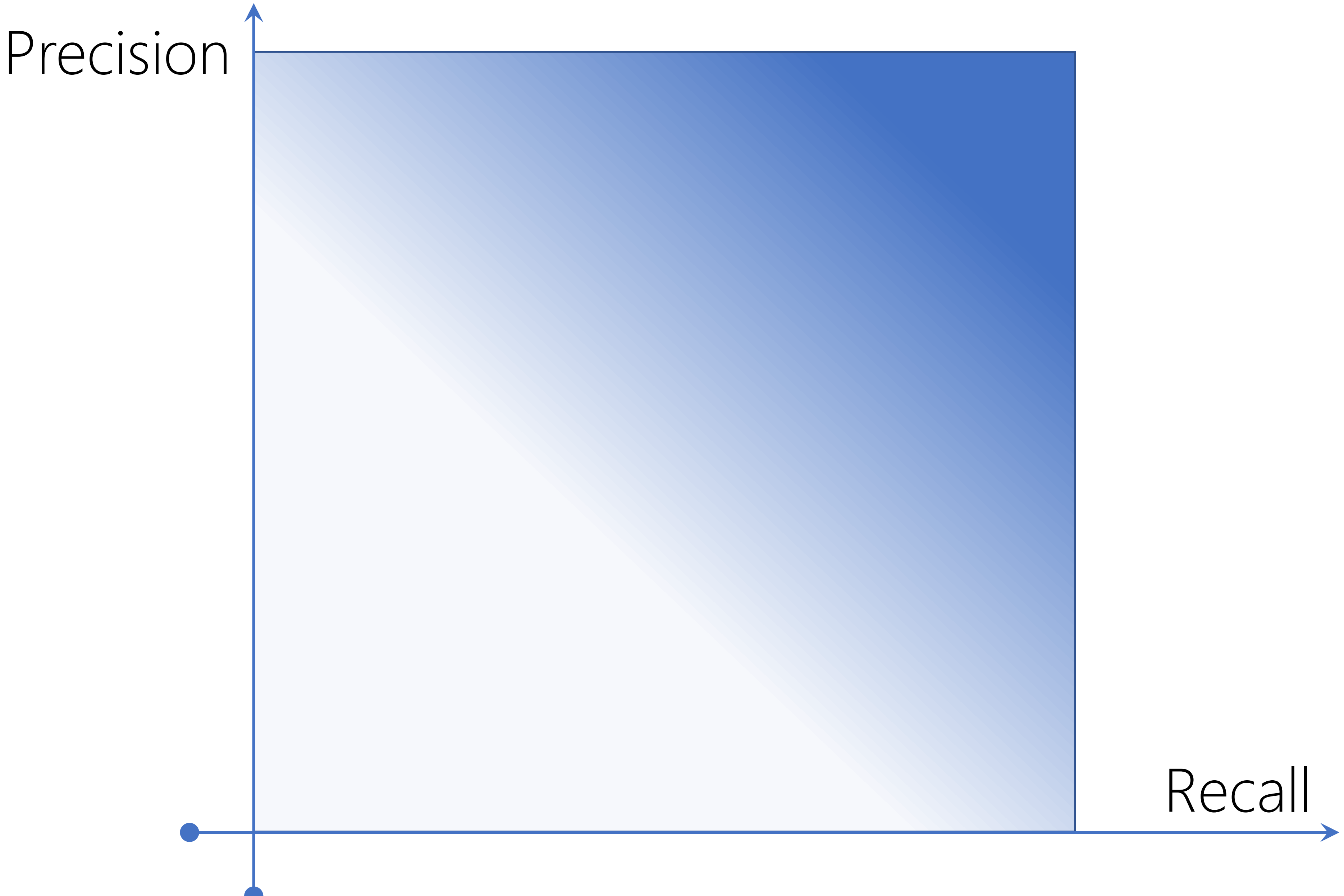


Figure 4.4 Contingency table

What scenarios require high recall?

Precision-Recall



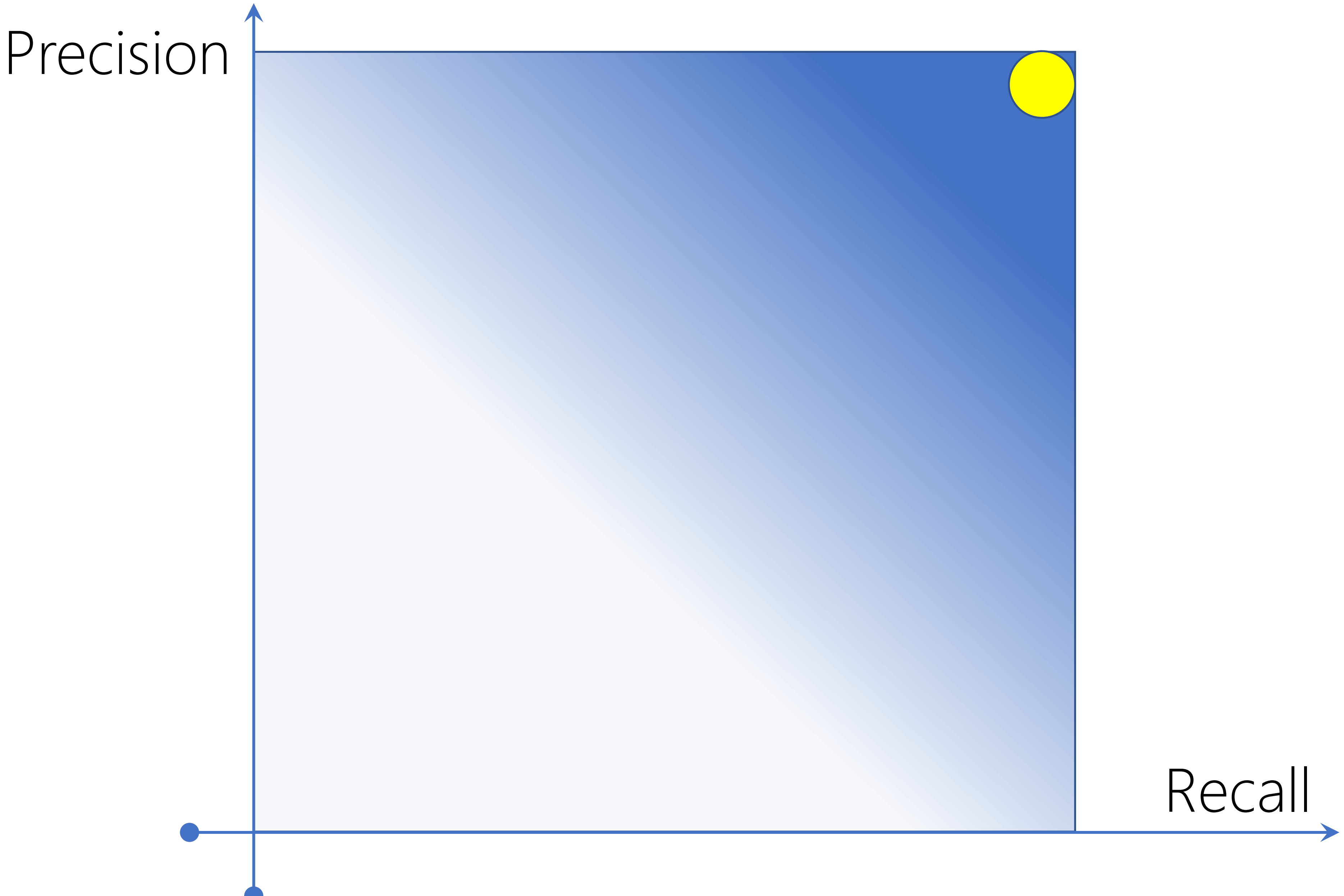
Perfect Classifier

	Gold Positive	Gold Negative
Model Positive	N+	0
Model Negative	0	N-

Precision $= \frac{\boldsymbol{N+}}{\boldsymbol{(N+)+0}}$ =1.0

Recall=$\frac{\boldsymbol{N+}}{\boldsymbol{(N+)+0}}$ =1.0

Precision-Recall



Balance Classes

~50% Positive, ~50% Negative

	Gold Positive (50)	Gold Negative (50)
Model Positive		
Model Negative		

Biased Model

All are Positive

	Gold Positive (50)	Gold Negative (50)
Model Positive	50	50
Model Negative	0	0

Biased Model

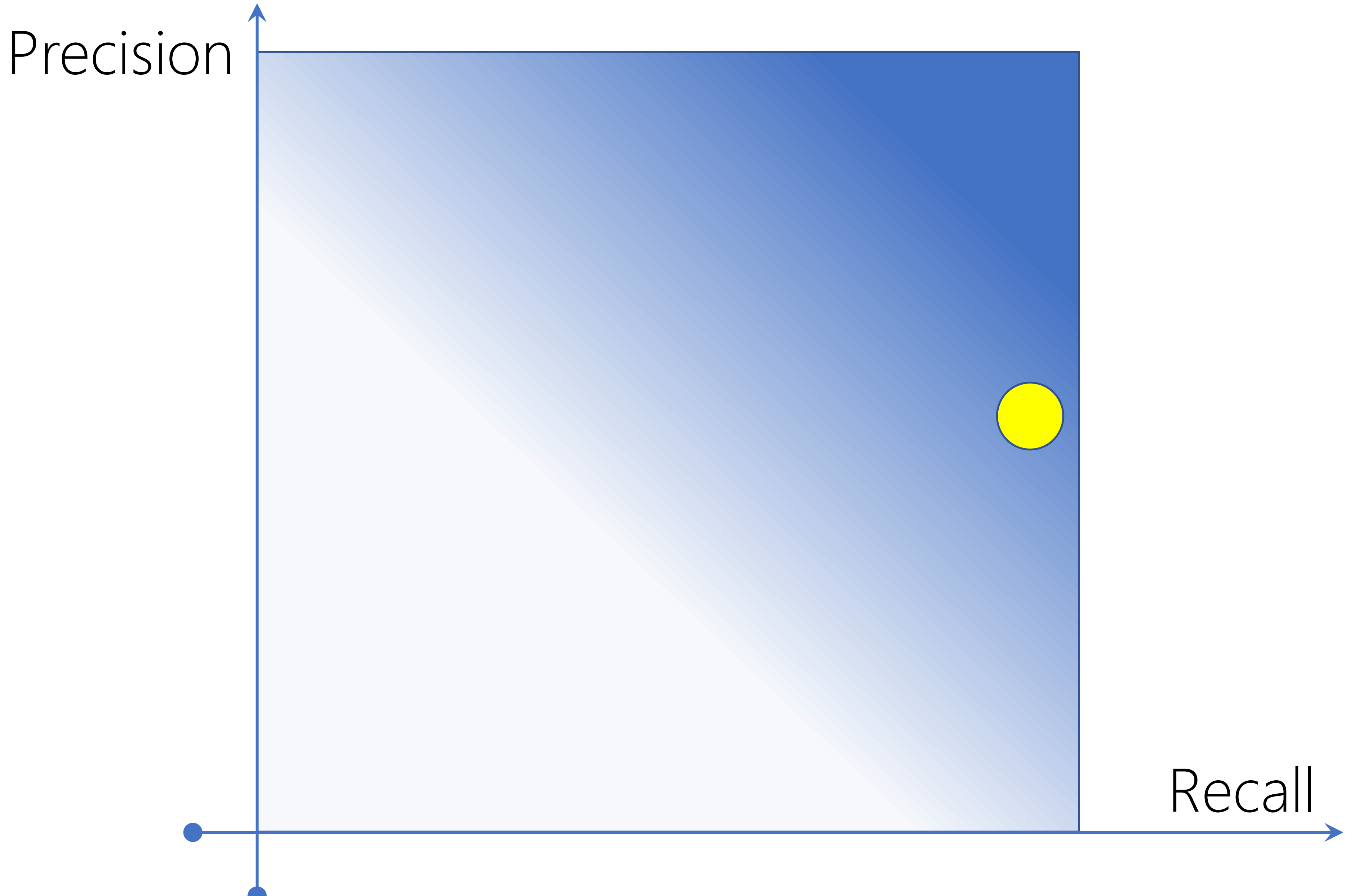
All are Positive

	Gold Positive (50)	Gold Negative (50)
Model Positive	50	50
Model Negative	0	0

Precision $= \frac{\mathbf{50}}{\mathbf{50+50}} = 0.5$

Recall$=\frac{\mathbf{50}}{\mathbf{50+0}} = 1.0$

Precision-Recall



Biased Model

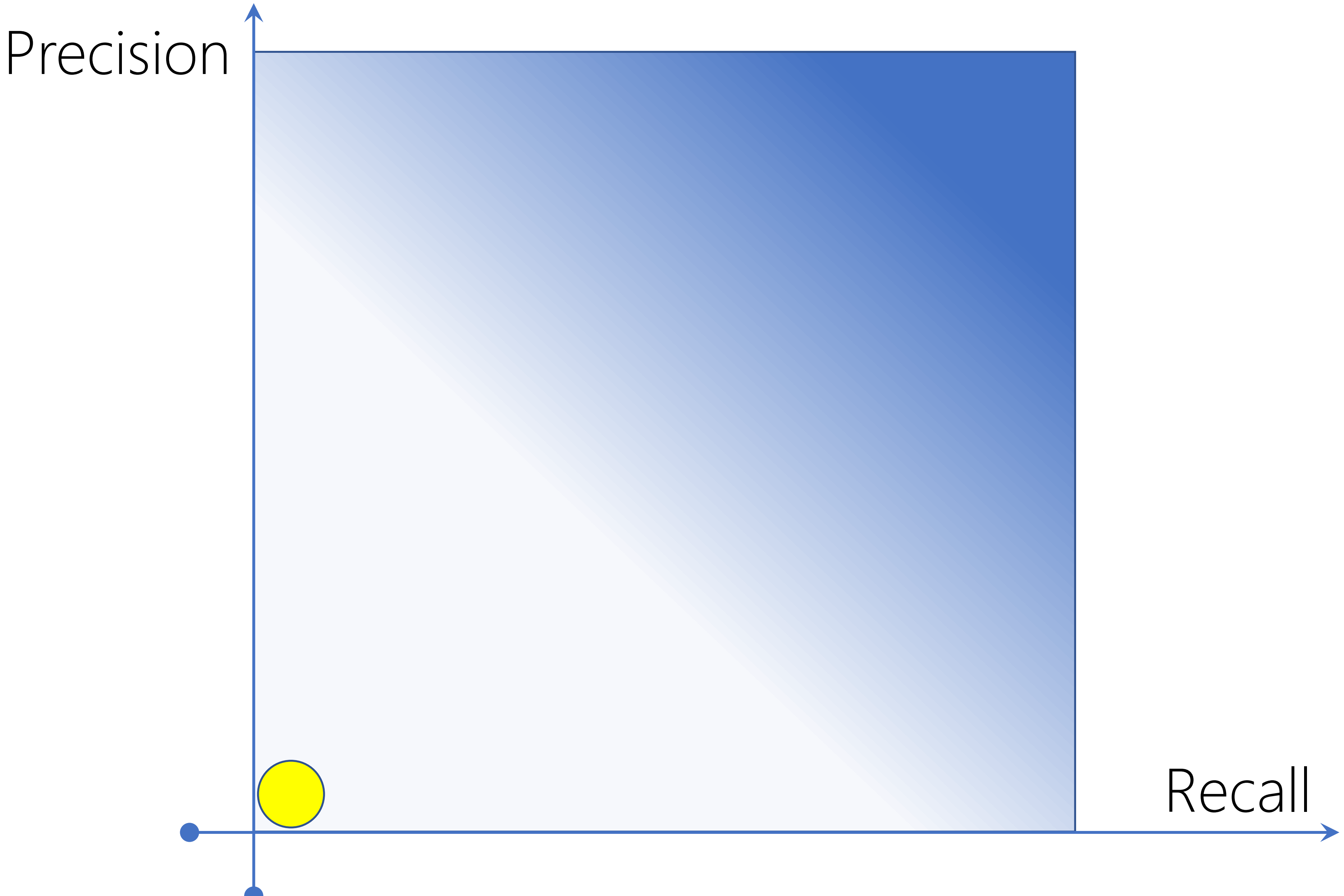
All are Negative

	Gold Positive (50)	Gold Negative (50)
Model Positive	0	0
Model Negative	50	50

$$\text{Precision} = \frac{\mathbf{0}}{\mathbf{0+0}} = 0.0$$

$$\text{Recall} = \frac{\mathbf{0}}{\mathbf{0+50}} = 0.0$$

Precision-Recall



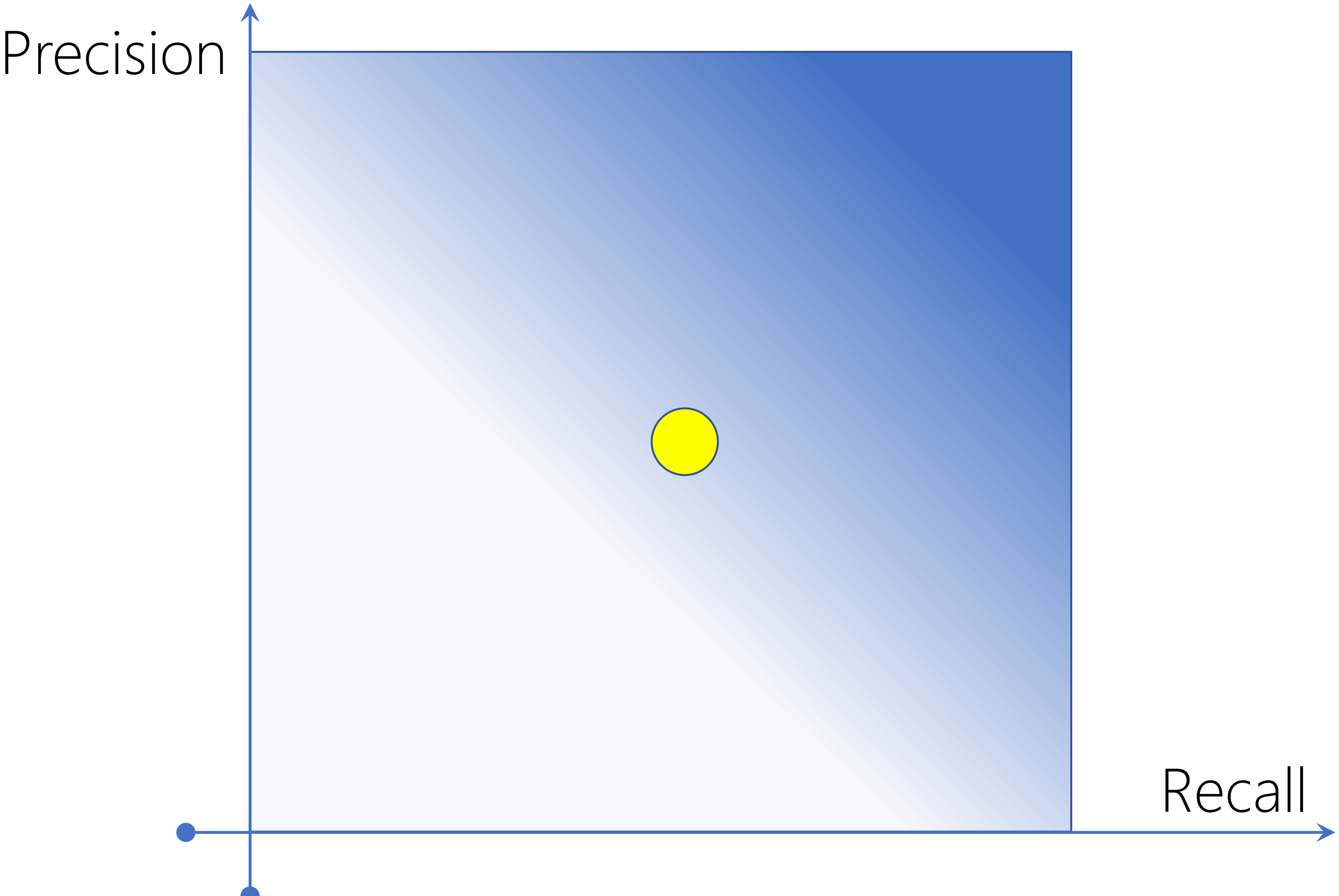
Uniformly Random Model

	Gold Positive (50)	Gold Negative (50)
Model Positive	25	25
Model Negative	25	25

$$\text{Precision} = \frac{25}{25+25} = 0.5$$

$$\text{Recall} = \frac{25}{25+25} = 0.5$$

Precision-Recall



Imbalance (Unbalanced) Classes

~10% Positive, ~90% Negative

	Gold Positive (10)	Gold Negative (90)
Model Positive		
Model Negative		

Biased Model

All are Positive

	Gold Positive (10)	Gold Negative (90)
Model Positive	10	90
Model Negative	0	0

Biased Model

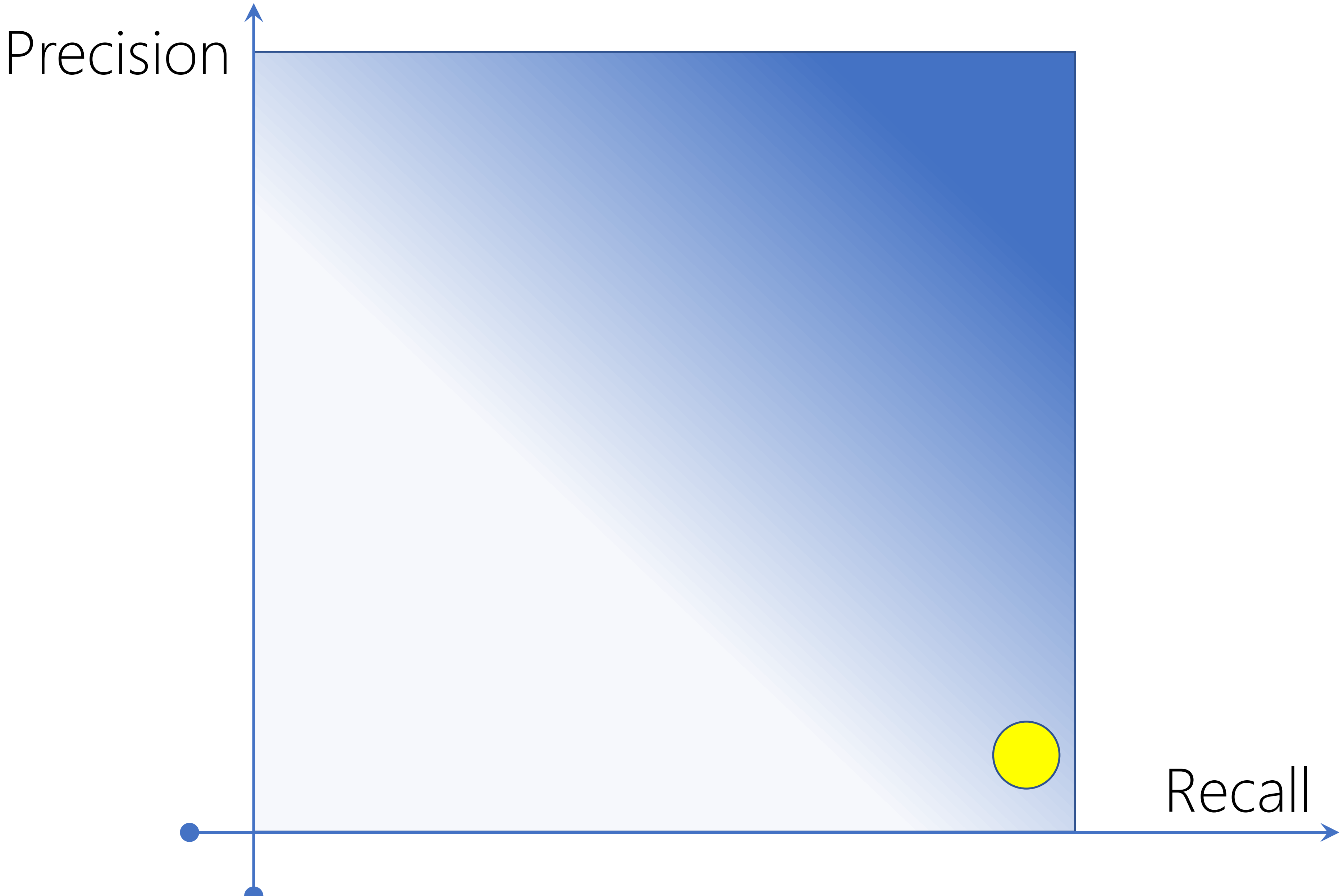
All are Positive

	Gold Positive (10)	Gold Negative (90)
Model Positive	10	90
Model Negative	0	0

Precision $= \frac{\mathbf{10}}{\mathbf{10+90}} = 0.1$

Recall$=\frac{\mathbf{10}}{\mathbf{10+0}} = 1.0$

Precision-Recall



Biased Model

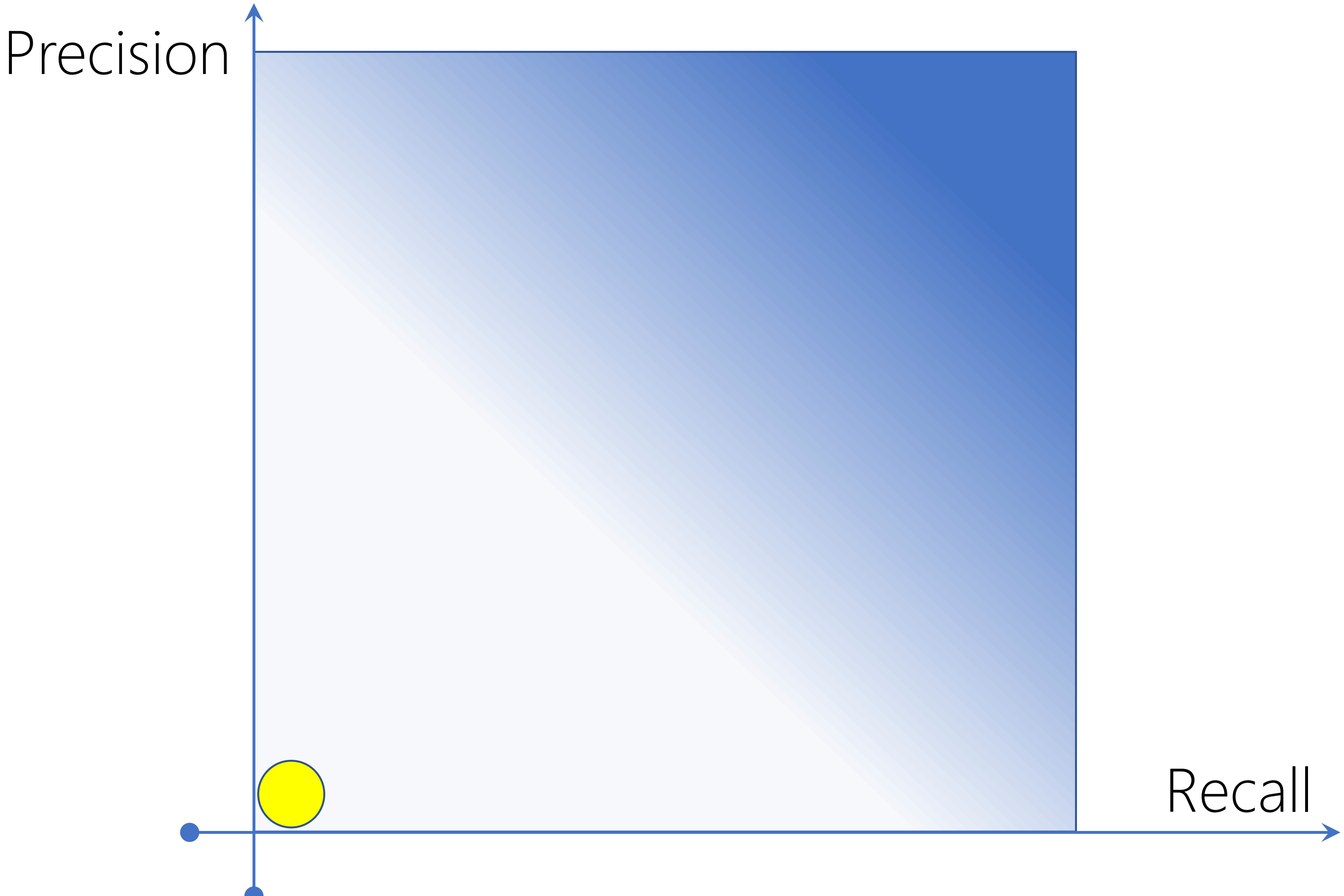
All are Negative

	Gold Positive (10)	Gold Negative (90)
Model Positive	0	0
Model Negative	10	90

$$\text{Precision} = \frac{\mathbf{0}}{\mathbf{0+0}} = 0.0$$

$$\text{Recall} = \frac{\mathbf{0}}{\mathbf{0+10}} = 0.0$$

Precision-Recall



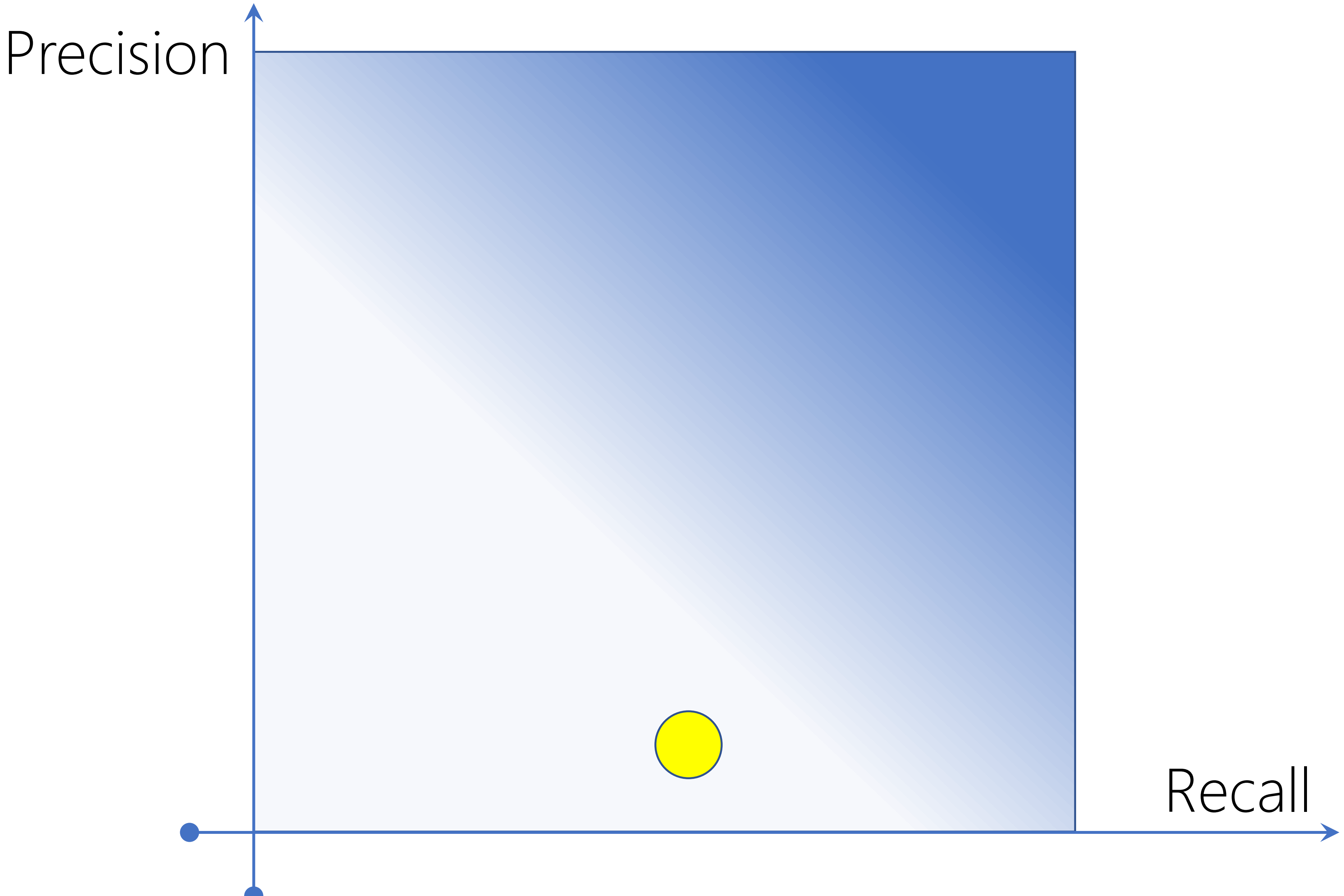
Uniformly Random Model

	Gold Positive (10)	Gold Negative (90)
Model Positive	5	45
Model Negative	5	45

$$\text{Precision} = \frac{5}{45+5} = 0.1$$

$$\text{Recall} = \frac{5}{5+5} = 0.5$$

Precision-Recall



Average of Precision and Recall: A Single Metric

$$\text{AVG-PR} = \frac{P+R}{2}$$

- Same weights
- Not fair! high precision may discount low recall or vice versa

Average of Precision and Recall: A Single Metric

$$\text{AVG-PR} = \frac{P+R}{2}$$

- Same weights
- Not fair! high precision may discount low recall or vice versa

$$\text{Harmonic AVG-PR} = \frac{2}{\frac{1}{P}+\frac{1}{R}} = 2\left(\frac{P \times R}{P+R}\right)$$

$$\mathrm{HarmonicMean}(a_1, a_2, a_3, a_4, \ldots, a_n) = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}$$

- Same weights
- Conservative! More toward the lower number.

Average of Precision and Recall: A Single Metric

$$\text{AVG-PR} = \frac{P+R}{2}$$

- Same weights
- Not fair! high precision may discount low recall or vice versa

$$\text{Harmonic AVG-PR} = \frac{2}{\frac{1}{P}+\frac{1}{R}} = 2\left(\frac{P \times R}{P+R}\right)$$

$$\mathrm{HarmonicMean}(a_1, a_2, a_3, a_4, \ldots, a_n) = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} + \ldots + \frac{1}{a_n}}$$

- Same weights
- Conservative! More toward the lower number.

$$\text{Weighted harmonic mean} = \quad F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad \text{or} \left(\text{with } \beta^2 = \frac{1-\alpha}{\alpha}\right) \quad F = \frac{(\beta^2+1)PR}{\beta^2 P + R}$$

$$F_\beta = \frac{(\beta^2+1)PR}{\beta^2 P + R} \begin{cases} \beta > 1: favors\ Recall \\ \beta = 1: 2\left(\frac{P \times R}{P+R}\right) \\ 0 \leq \beta < 1: favors\ Precision \end{cases}$$

Curves

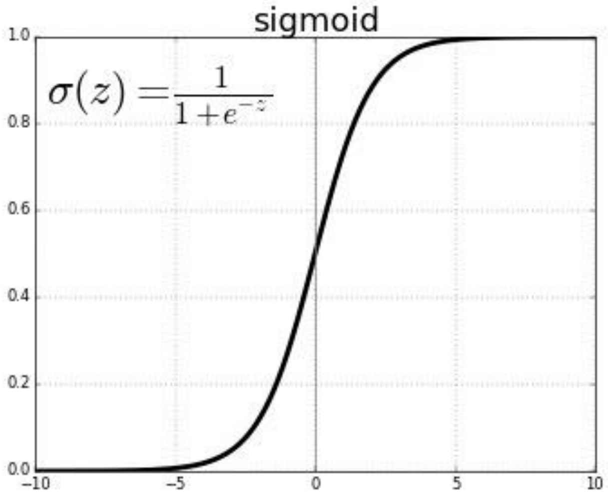
Threshold-based Model

$P(+|\, x) = 1 - P(-|\, x)$

$P(x) = \text{Sigmoid}(f(x))$

$P(x) \geq \delta \longrightarrow$ x is positive

$P(x) < \delta \longrightarrow$ x is negative



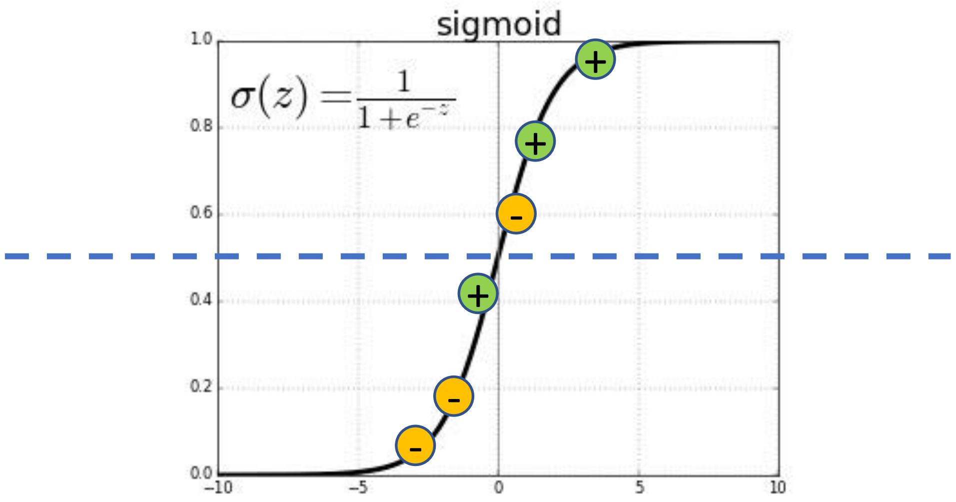
Threshold-based Model
$\delta = 0.5$

$P(+|\ x) = 1 - P(-|\ x)$

$P(x) = \mathrm{Sigmoid}(f(x))$

$P(x) \geq 0.5 \longrightarrow$ x is positive

$P(x) < 0.5 \longrightarrow$ x is negative



Threshold-based Model

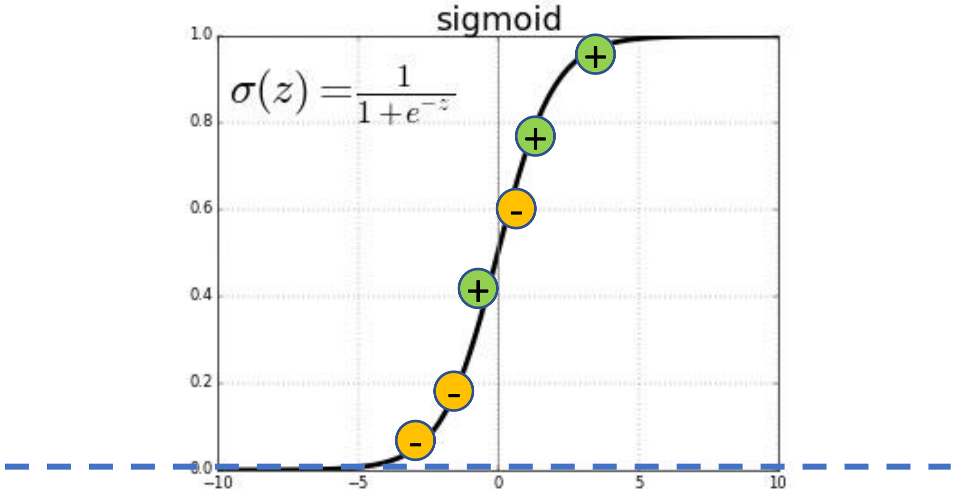
δ=0.0 ⟶ Biased Model ⟶ All are positives

P(+| x) = 1 - P(-| x)

P(x) = Sigmoid(f(x))

P(x) ≥ 0.0 ⟶ x is positive

P(x) < 0.0 ⟶ x is negative



Threshold-based Model

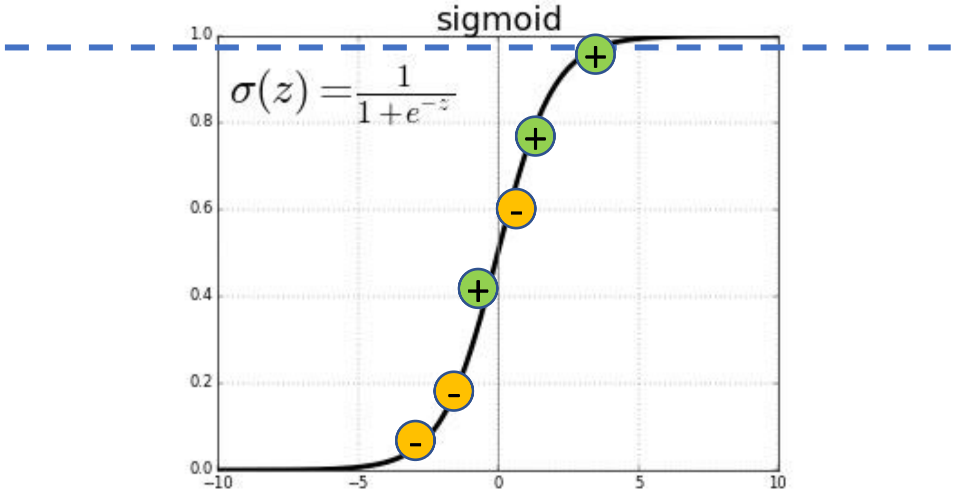
δ=1.0 ⟶ Biased Model ⟶ All are negatives

$P(+|\ x) = 1 - P(-|\ x)$

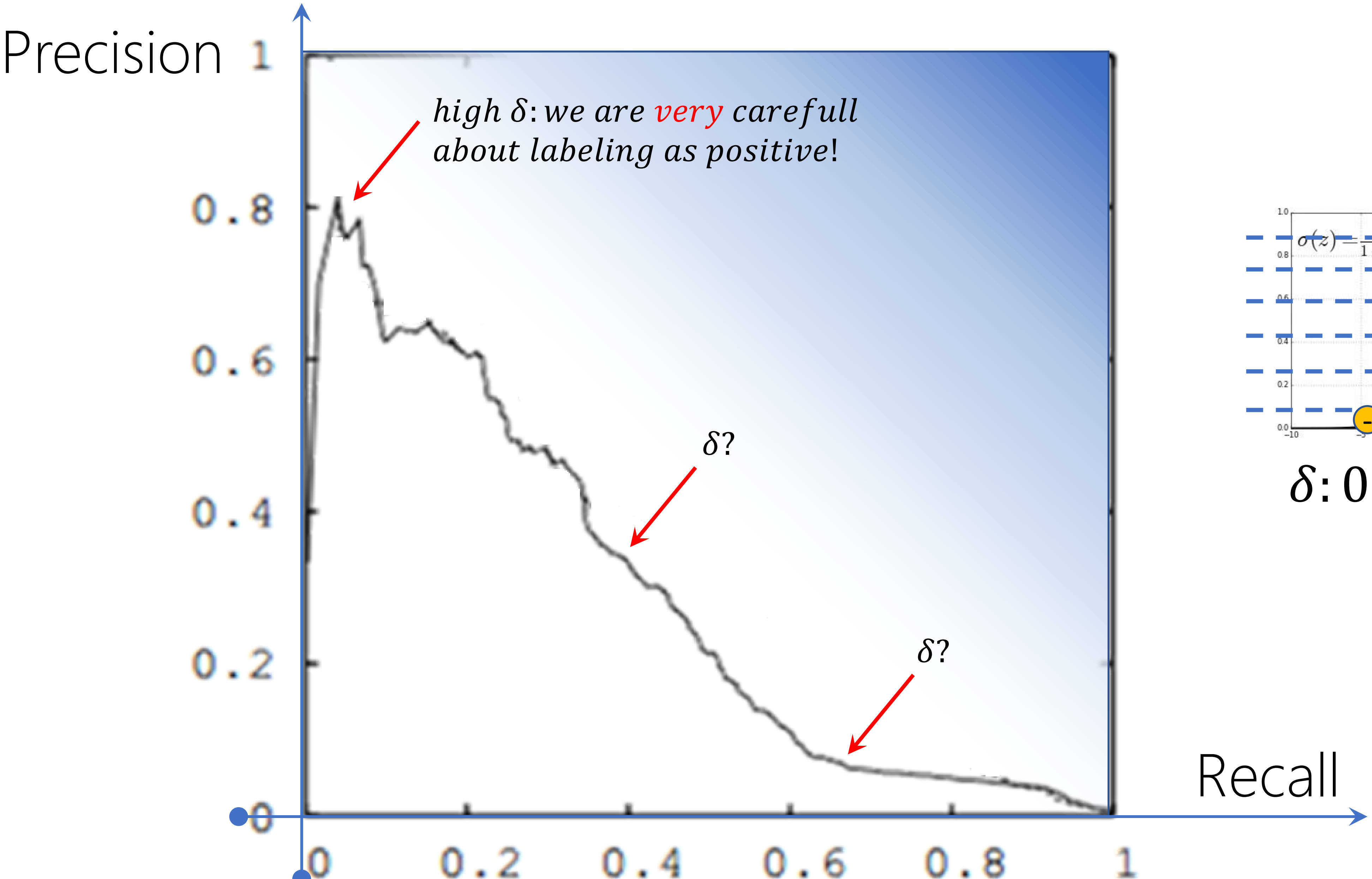
$P(x) = \text{Sigmoid}(f(x))$

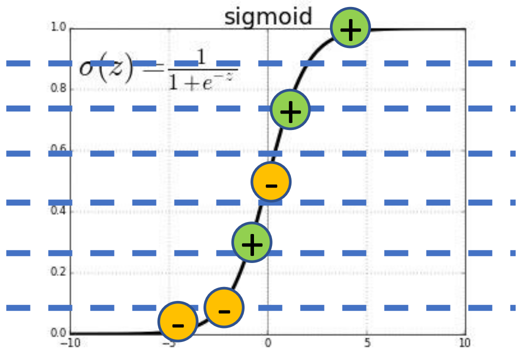
$P(x) \geq 1.0$ ⟶ x is positive

$P(x) < 1.0$ ⟶ x is negative



Precision-Recall Curve: Best δ





$\delta: 0.0 \rightarrow 1.0$

Precision-Recall Curve: Model Comparison