Classical Classification

FinTech

Lesson 11.1

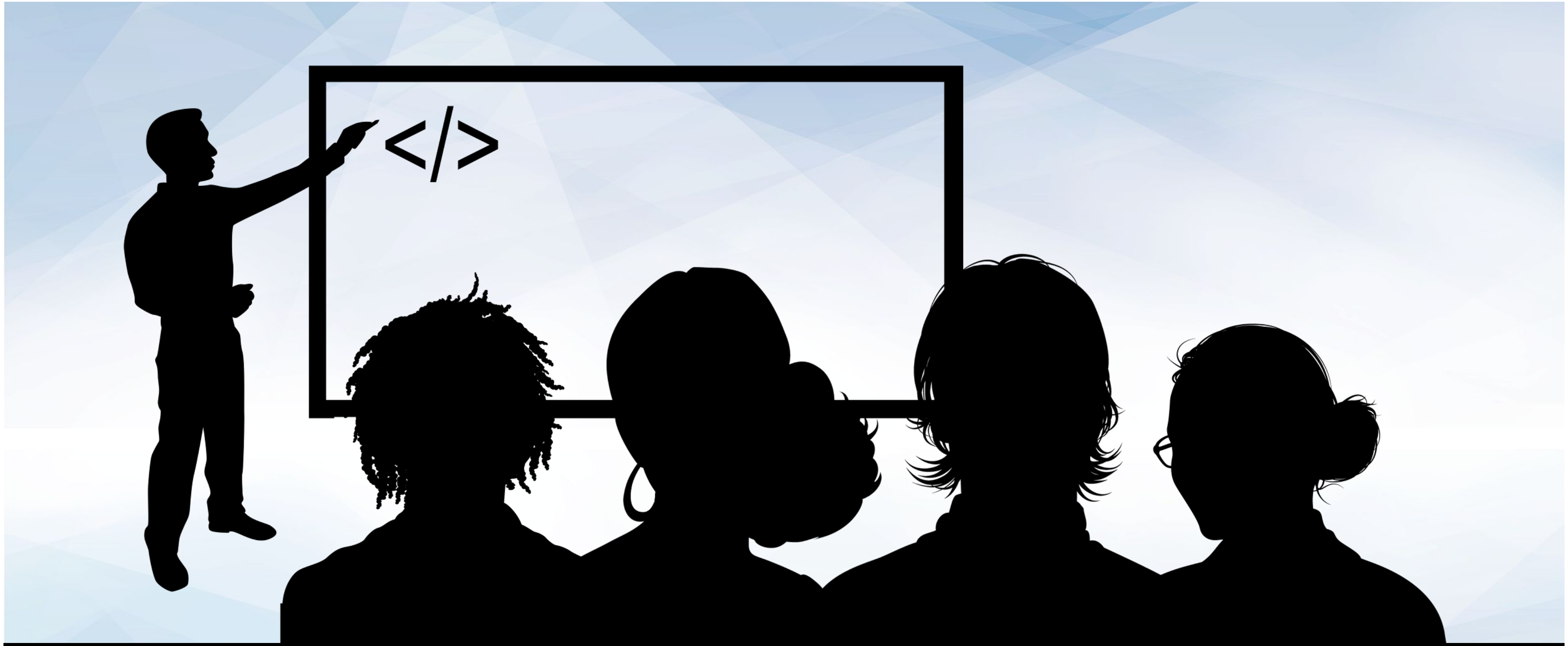
Class Objectives

In today's class we'll learn about classification algorithms

- Logistic regression
- Support vector machines

Use Cases:

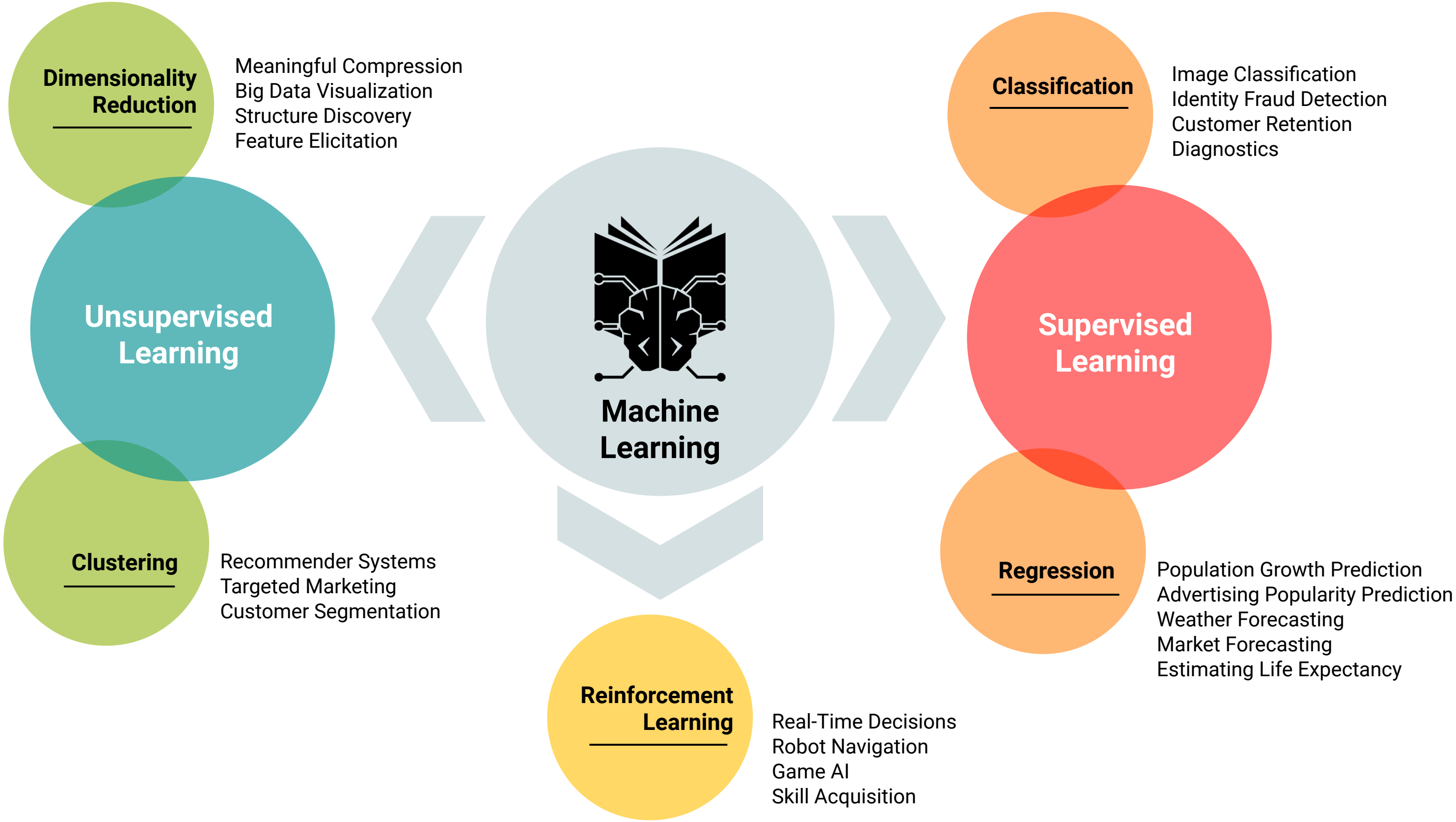
- Fraud detection
- Lending decisions
- Medical disease/condition diagnosis



Instructor Demonstration

Demo Homework

This is the second week of machine learning!



Dimensionality Reduction

Meaningful Compression
Big Data Visualization
Structure Discovery
Feature Elicitation

Unsupervised Learning

Clustering

Recommender Systems
Targeted Marketing
Customer Segmentation

Machine Learning

Reinforcement Learning

Real-Time Decisions
Robot Navigation
Game AI
Skill Acquisition

Supervised Learning

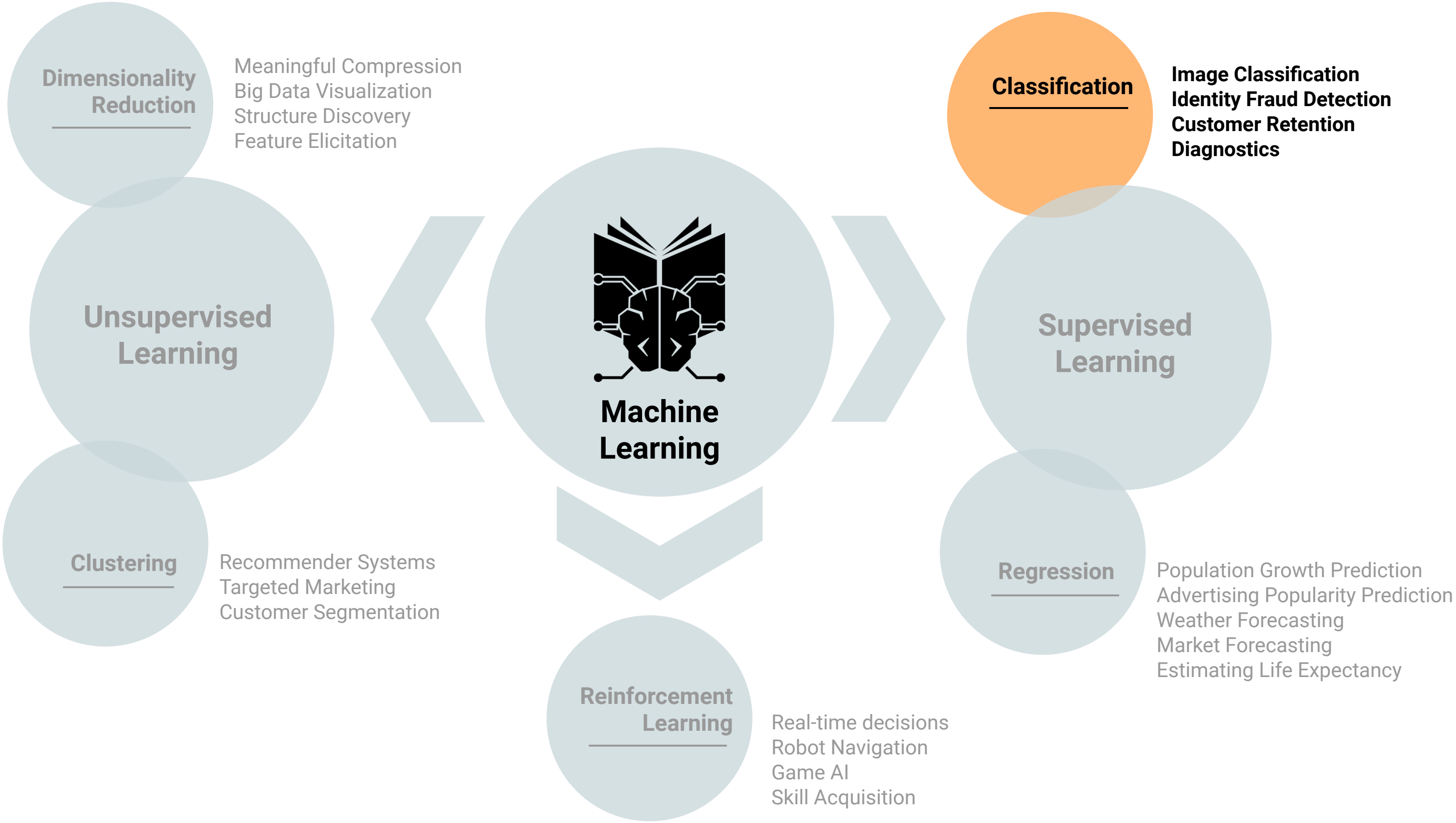
Classification

Image Classification
Identity Fraud Detection
Customer Retention
Diagnostics

Regression

Population Growth Prediction
Advertising Popularity Prediction
Weather Forecasting
Market Forecasting
Estimating Life Expectancy

Intro to Classification



Classification is the action or process of categorizing something according to shared qualities or characteristics.

Classification

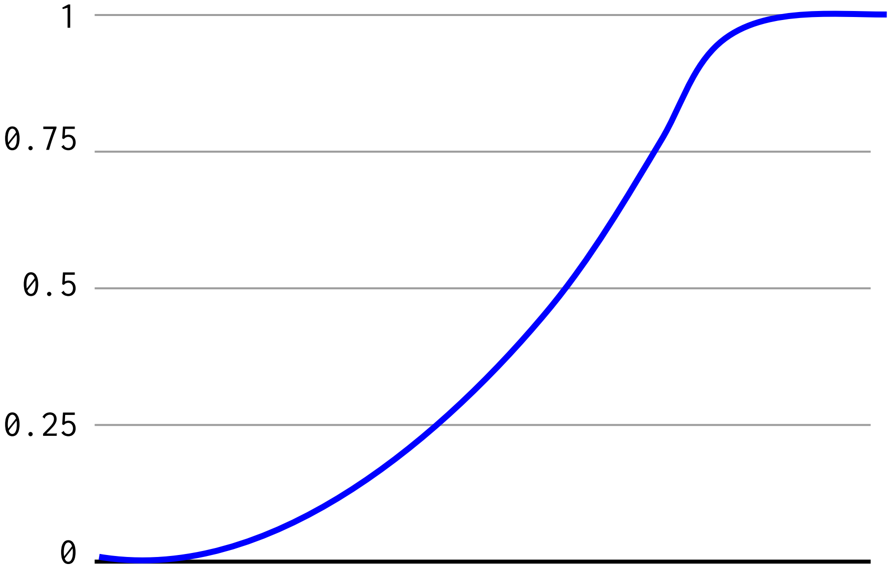
Classification is the prediction of discrete outcomes. Outcomes are identified as labels/discrete outputs, which serve to categorize bi-class and multi-class features.

Spam		not spam
Risk	vs.	no risk
0		1

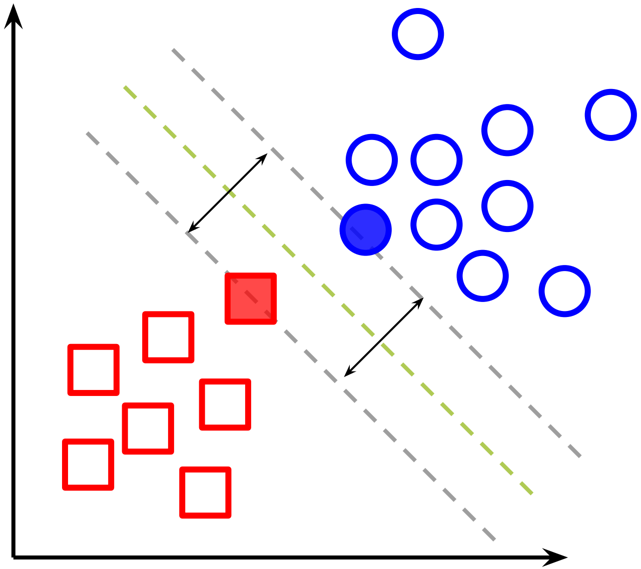
Classification

There are multiple approaches to classification. These include:

Logistic Regression



Support Vector Machines



Classification

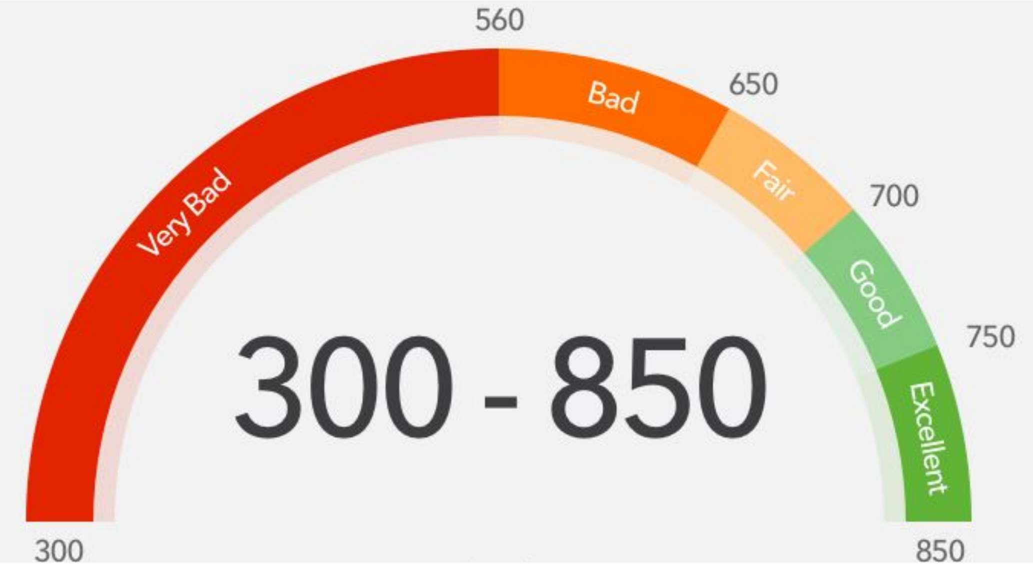
Classification is used to forecast and predict financial outcomes, automate underwriting and insurance premiums, detect and categorize health issues and overall health.

Classification

Classification models have drastically improved financial efforts to properly categorize applicants, predict market decline, and categorize fraudulent transactions or suspicious activity.

Classification

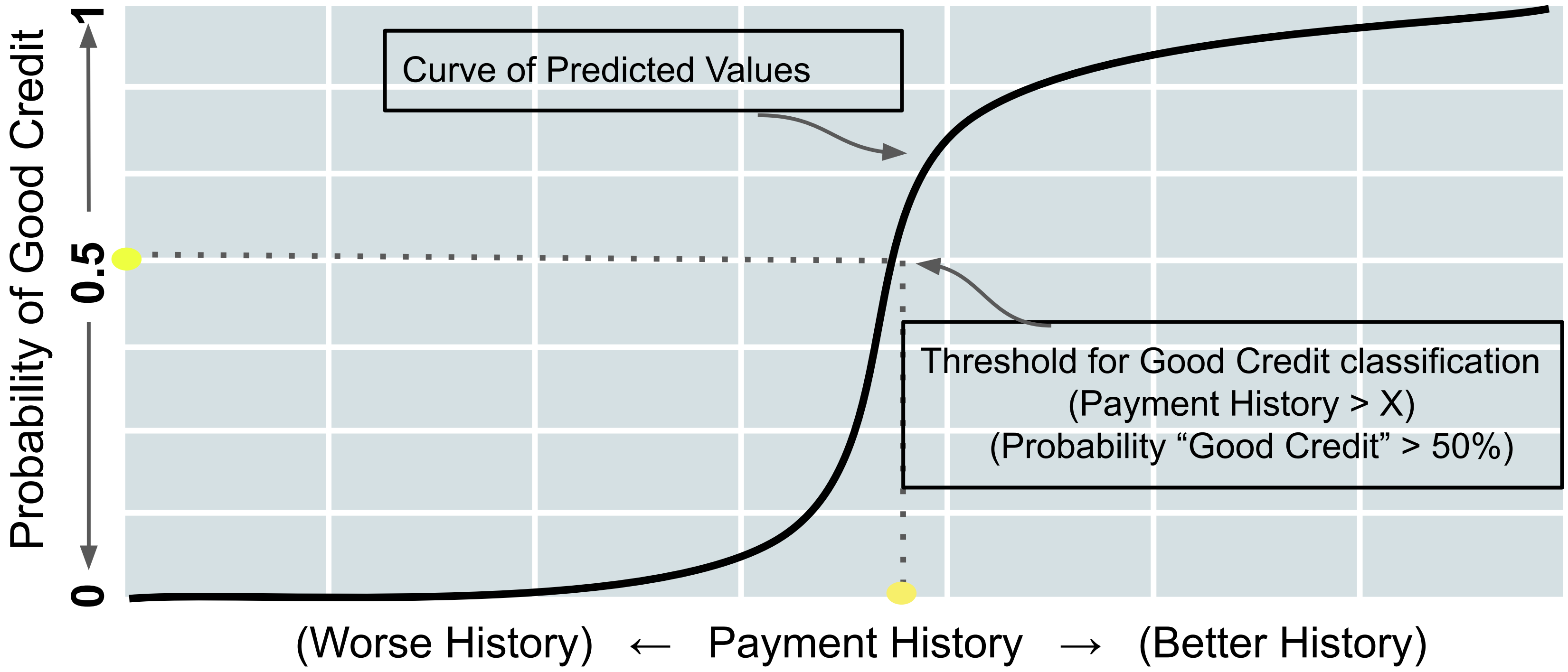
FICO credit scoring uses a classification model for its cognitive fraud analytics platform. Classification engines have allowed the financial industry to become more effective and efficient at mitigating risk.



Making Predictions with Logistic Regression

Making Predictions with Logistic Regression

A probability is assigned for each data point, of being in the **1** category (e.g. “good credit”)

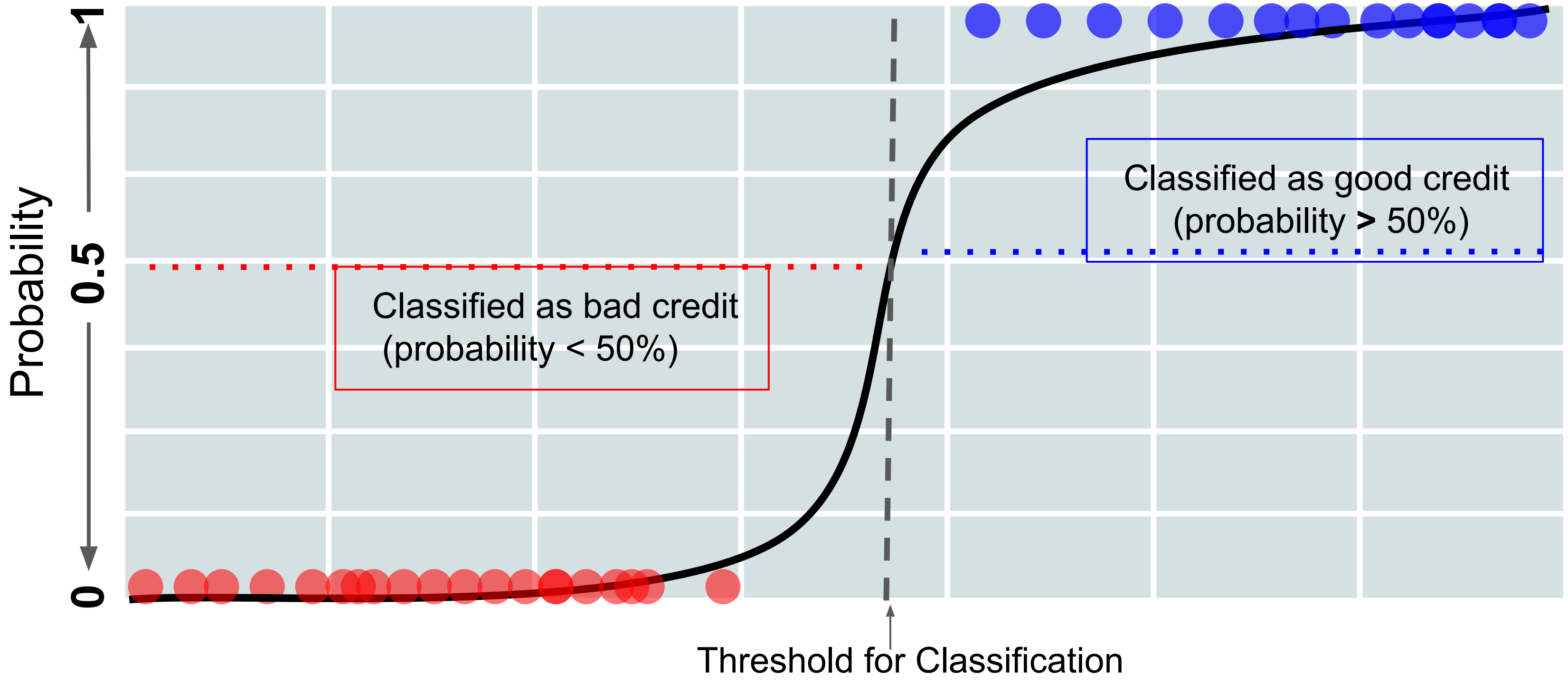


Curve of Predicted Values

Threshold for Good Credit classification
(Payment History > X)
(Probability “Good Credit” > 50%)

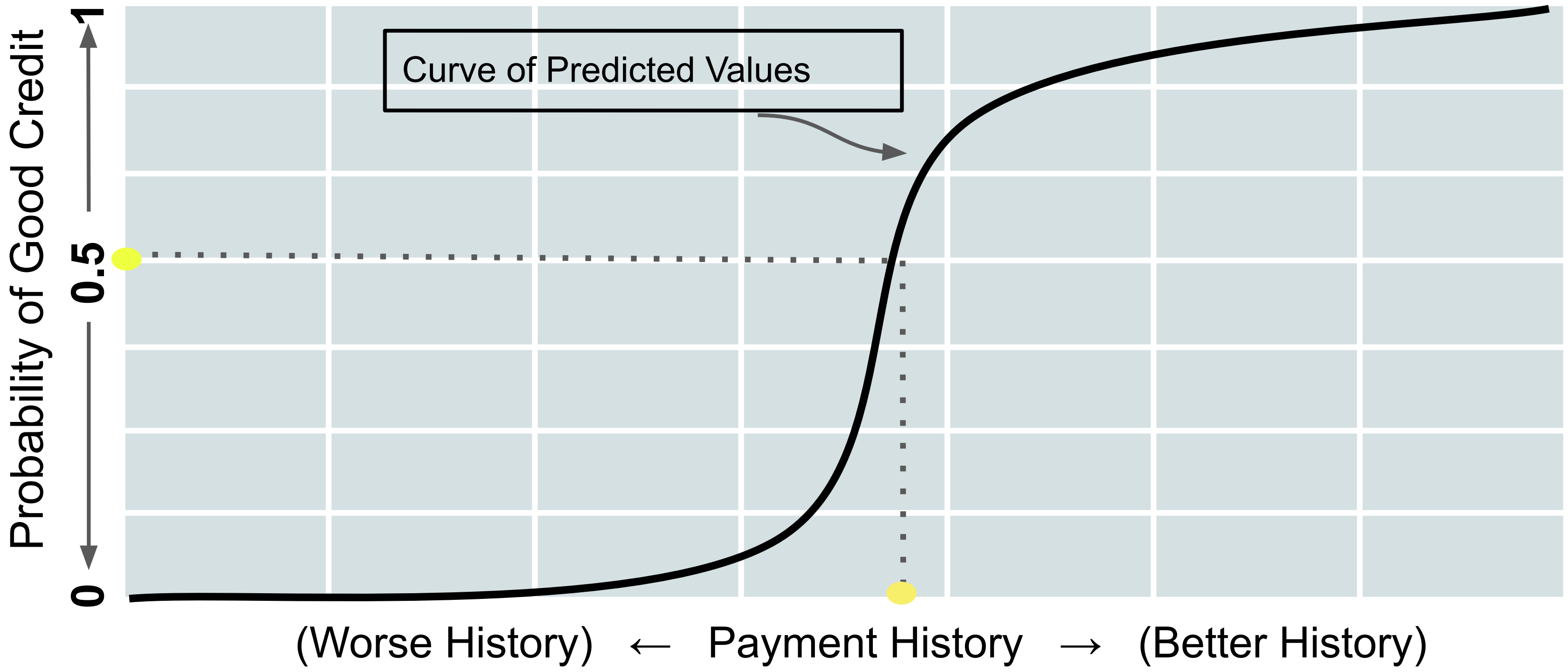
Making Predictions with Logistic Regression

If the probability is above a certain threshold, that data point is estimated to be a **1** ("good credit"), and a (**0**) otherwise.



The Sigmoid Function

How do we convert **Payment History** data (measured in months) to a **Probability** of Good Credit **(**which ranges from 0 to 1)?



The Sigmoid Function

How do we convert **Payment History** data (measured in months) to a **Probability** of Good Credit (which ranges from 0 to 1)?

Logistic Regression converts using a **sigmoid**, or **squashing** function:

$$Probability\,of\,Good\,Credit = \frac{1}{1 + e^{-Payment\ History}}$$

Done behind the scenes, this function converts continuous data on the borrower (e.g., number of months without a delinquent payment) to a percentage probability of being a “good credit” borrower.

A good logistic regression model will use more information than just “payment history”, but the **sigmoid** function can still convert all this information into a probability.

Logistic Regression Model

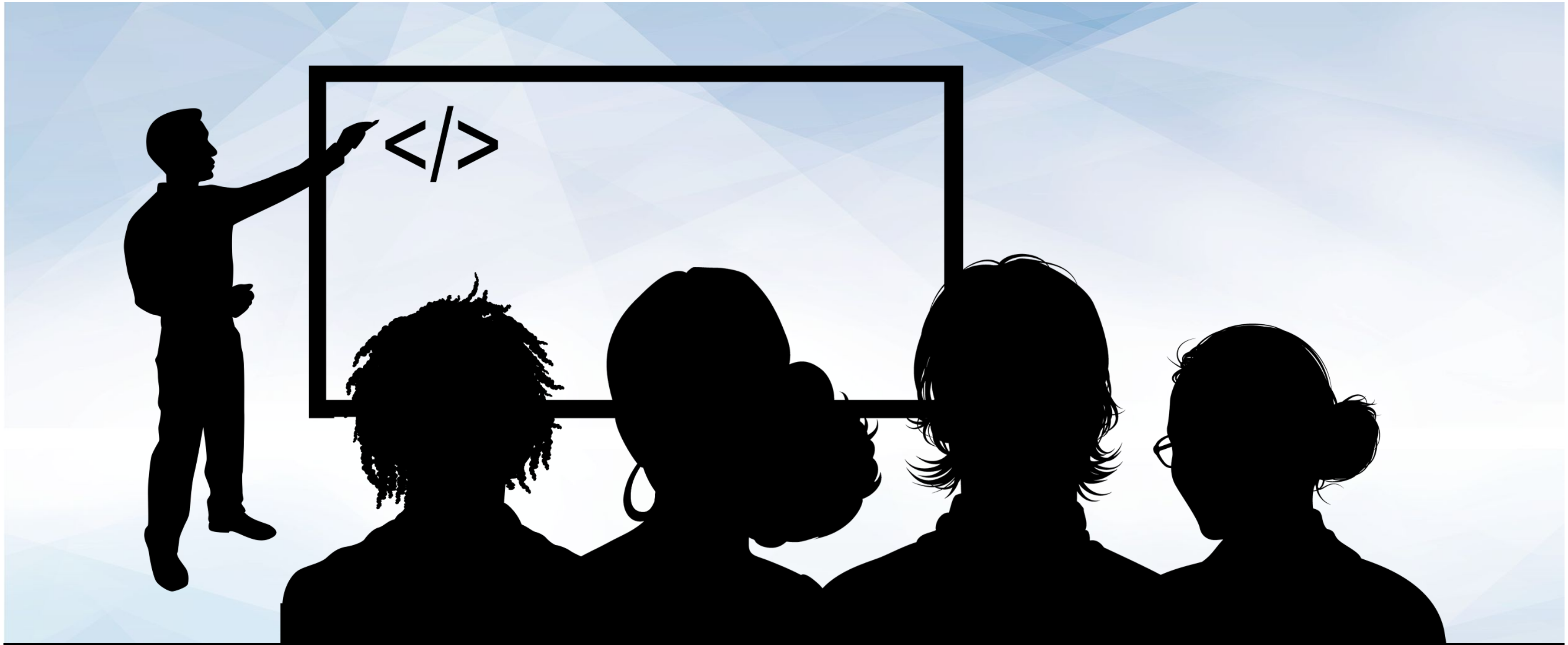
Running a logistic regression model involves 4 steps, which can be applied when running any machine-learning model:

01 Preprocess

02 Train

03 Validate

04 Predict



Instructor Demonstration

Logistic Regression using Scikit-Learn

Activity: Predicting Diabetes

In this activity, you will use the sklearn library to execute logistic regression models in order to predict whether or not an individual has diabetes.

Suggested Time:
15 minutes

Time’s Up! Let’s Review.

Review: Predicting Diabetes

How well did your model perform?

How do you know? Did you count the results?

If you were asked to diagnose a patient, how confident would you be in your model's prediction?

Evaluating Logistic Regression Predictions

How sure are you that models can actually predict diabetes?

Answer

75%

sure, as described by the scored accuracy.

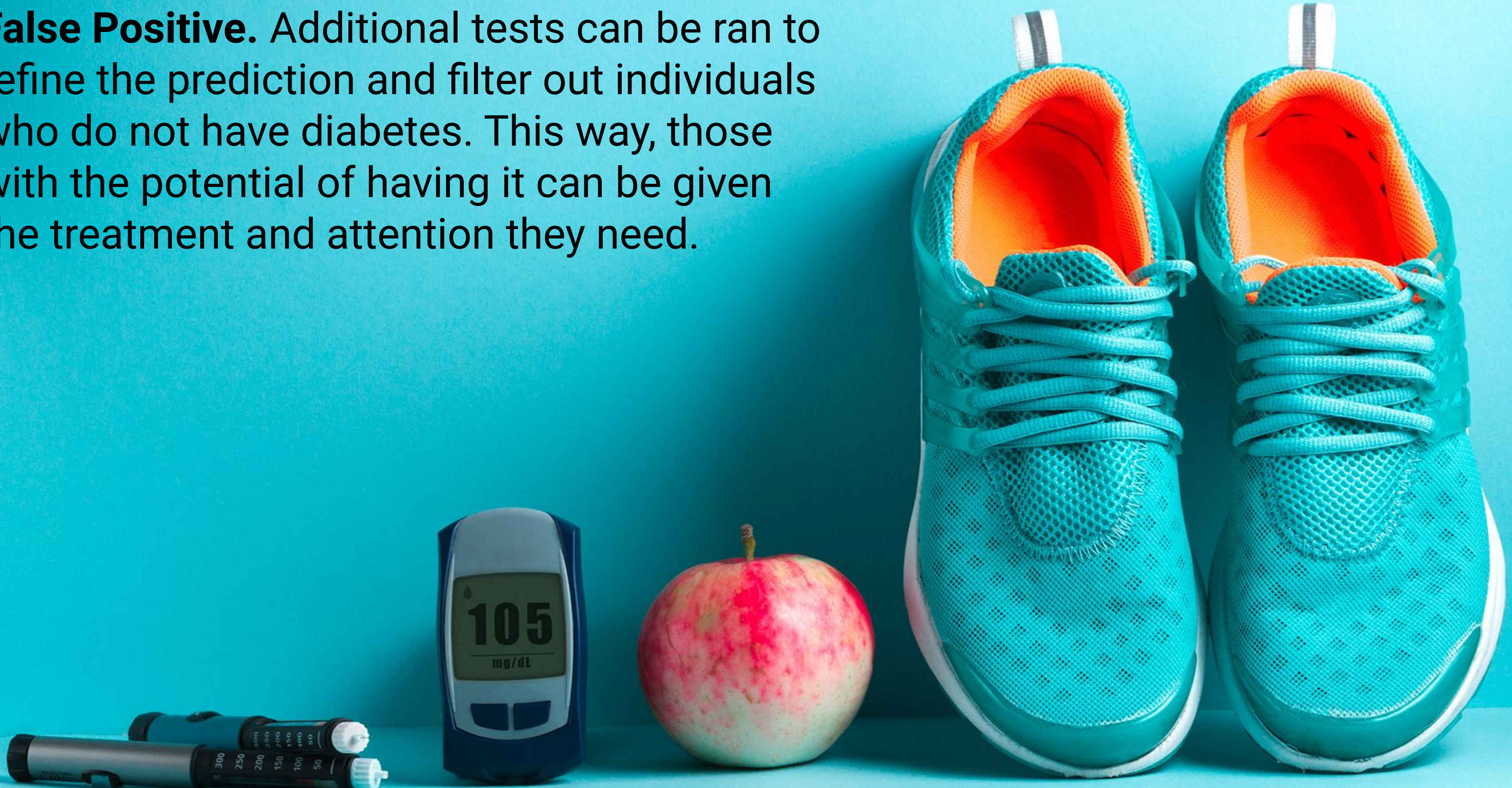
Would you feel comfortable giving the diagnosis of diabetes based off the predictions of the model.

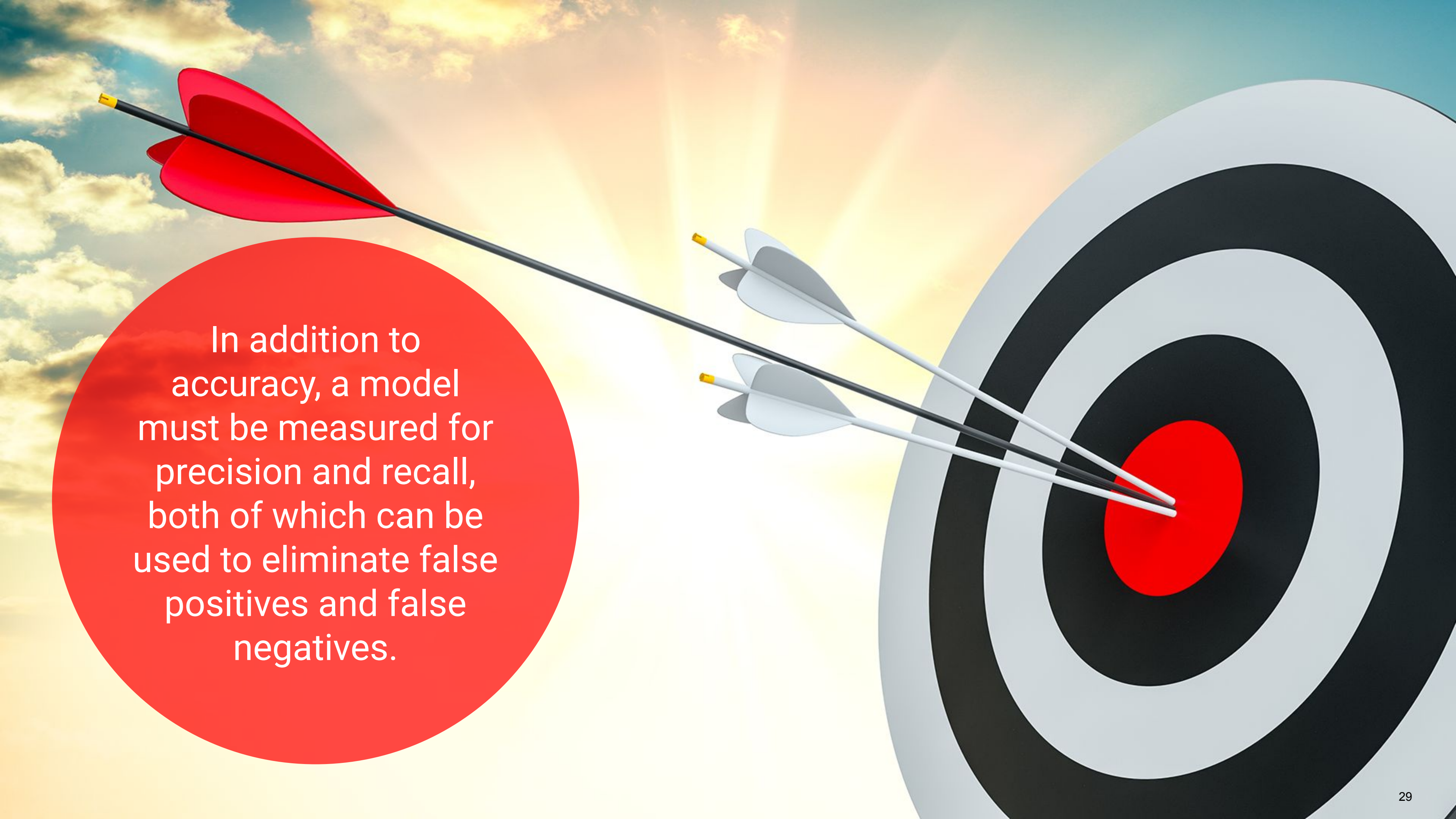
No. The prediction is not 100% accurate. There is room for error, as well as false positives.

What is better: the false positive or false negative?

Answer

False Positive. Additional tests can be ran to refine the prediction and filter out individuals who do not have diabetes. This way, those with the potential of having it can be given the treatment and attention they need.



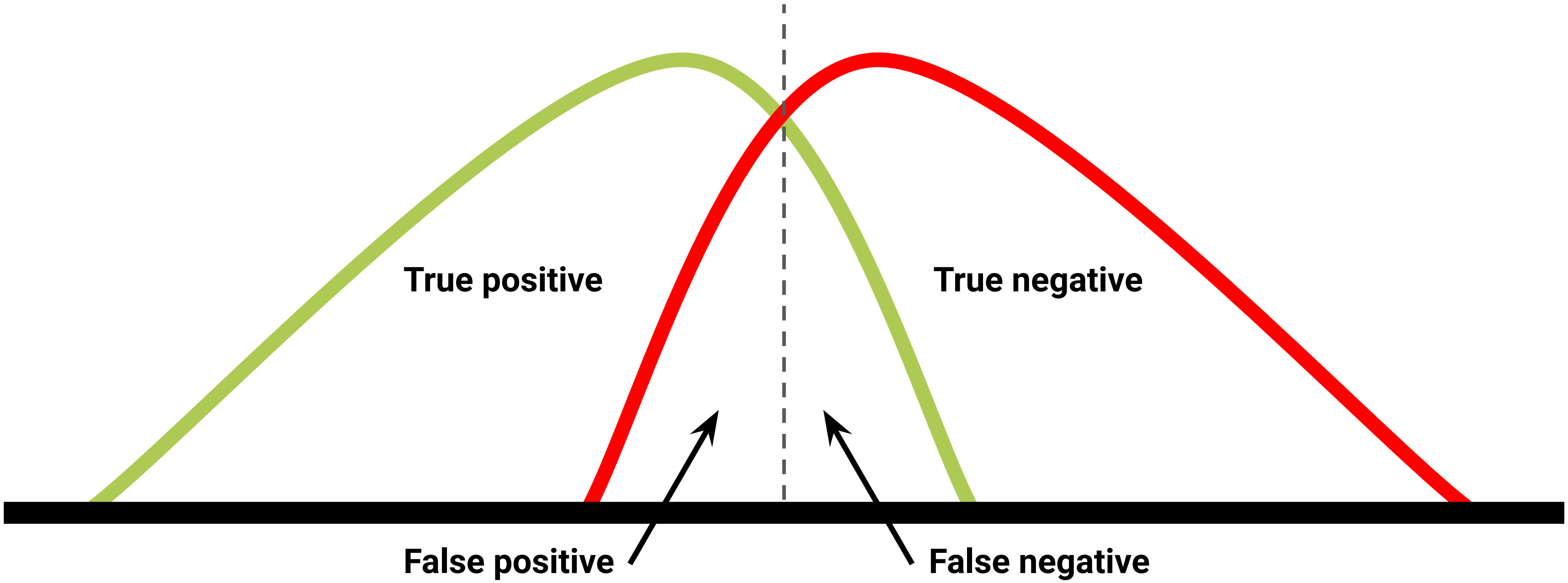


In addition to accuracy, a model must be measured for precision and recall, both of which can be used to eliminate false positives and false negatives.

Accuracy, Precision, Recall

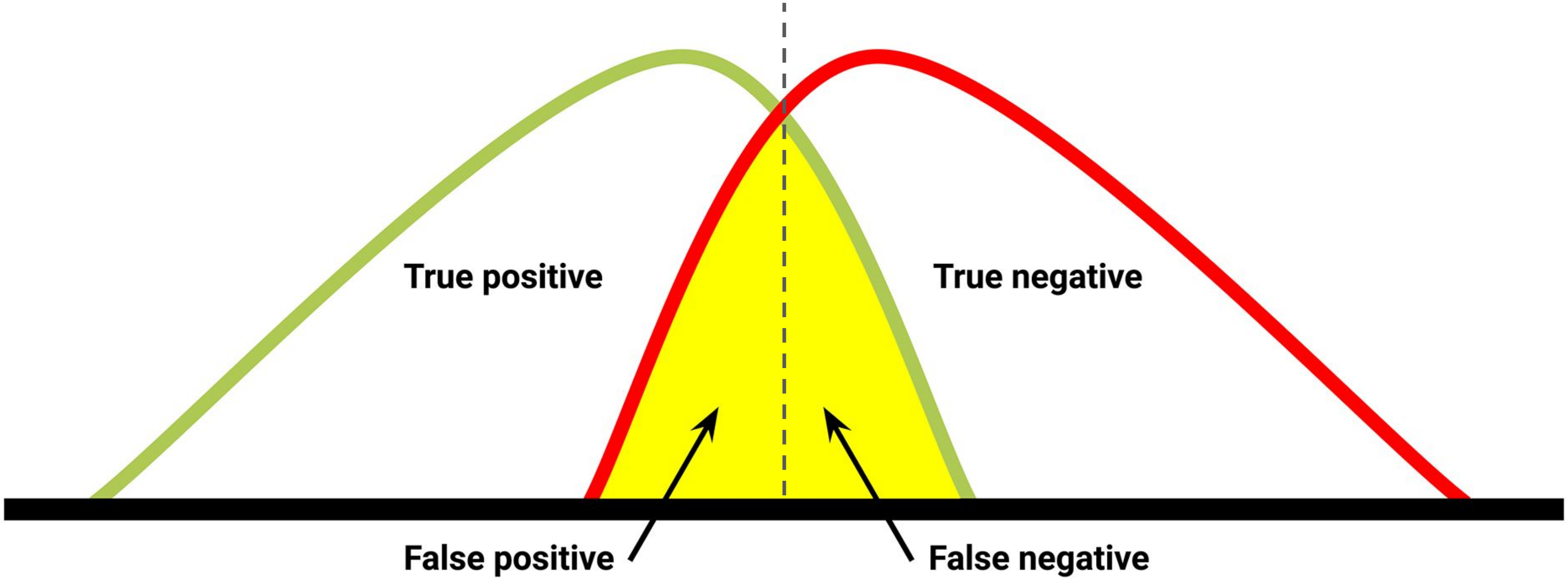
Accuracy, Precision, Recall

Accuracy, precision, and recall are especially important for classification model that involve a binary decision problem. Binary decision problems have two possible correct answers: **True Positive** and **True Negative**.



Accuracy, Precision, Recall

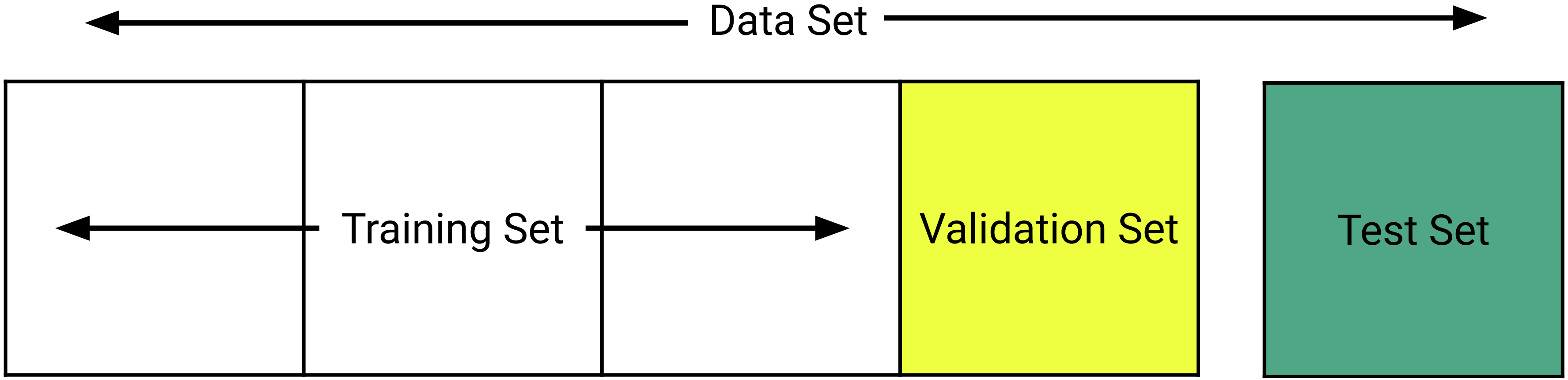
Inaccurate and imprecise models result in models returning false positives and false negatives.



Accuracy is how often the model is correct—the ratio of correctly predicted observations to the total number of observations.

Accuracy

Scoring will reveal how accurate the model. However, it does not communicate how precise it is.



Accuracy

Accuracy can be very susceptible to imbalanced classes. In the case of the homework assignment, the number of good loans greatly outweighs the number of at-risk loans. In this case, it can be really easy for the model to only care about the good loans because that has the biggest impact on accuracy. However, we also care about the at-risk loans, so we need a metric that can help us evaluate each class prediction.

Calculation:

$$(TP + TN) / (TP + TN + FP + FN)$$

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

Precision

Another example of precision is of all of the individuals that were classified by the model as being a credit risk, how many actually were a credit risk?

The question at hand: Did we classify comprehensively and correctly?

470
VERY POOR
300-580
POOR
580-640
FAIR
640-720
GOOD
720-780
EXCELLENT
780-850

Precision

High precision relates to a low false positive rate.

Calculation:

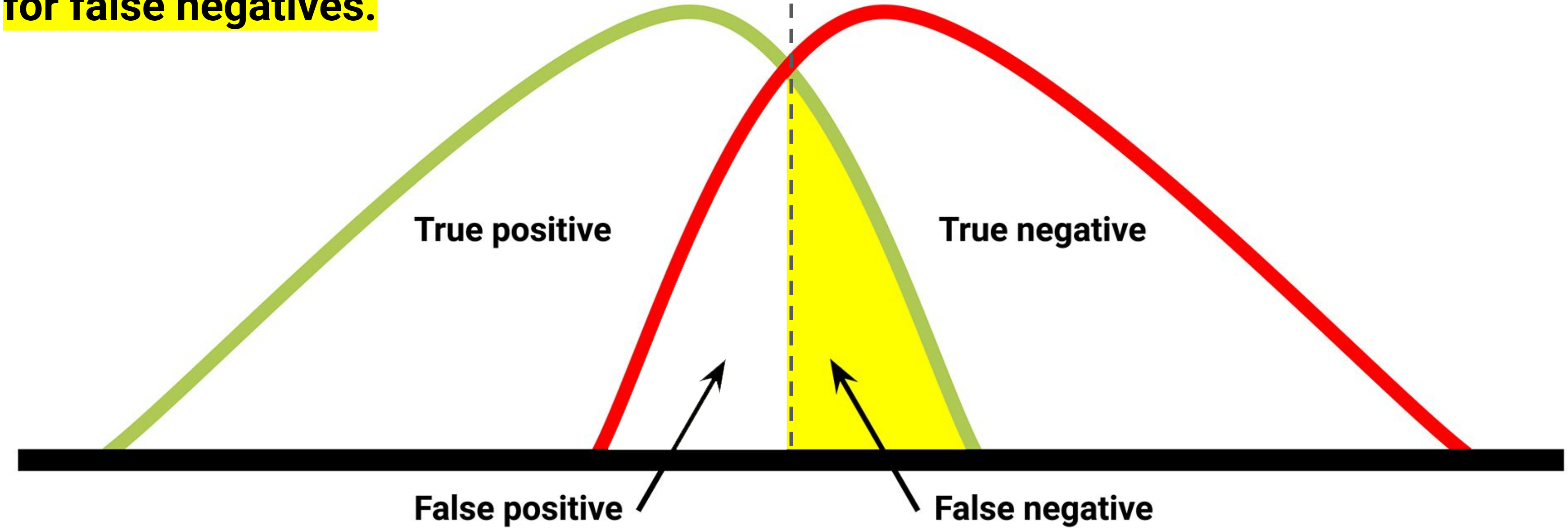
$$TP / (TP + FP)$$

Recall is the ratio of correctly predicted positive observations to all predicted observations for that class

Recall

Of all of the actual diabetes/credit risk samples, how many were correctly classified as having diabetes/being a credit risk.

The question at hand: Did we classify all samples correctly, leaving little room for false negatives.



Recall

High recall relates to a more comprehensive output and a low false negative rate.

Calculation:

$$TP \;/\; (TP + FN)$$

Confusion Matrix & Classification Report

A **confusion matrix** is used to measure and gauge the success of a model.

Confusion Matrix

Confusion matrices reveal the number of true negatives and true positives (actuals) for each categorical class and compares it to the number of predicted values for each class.

n=165	**Predicted: No**	**Predicted: Yes**
Actual=No	50	10
Actual=Yes	5	100

Confusion Matrix

These values are then individually summed by column and row. The aggregate sums are then compared to gauge accuracy and precision. If the aggregates match, the model can be considered accurate and precise.

n=165	**Predicted: No**	**Predicted: Yes**	
Actual=No	50	10	**=60**
Actual=Yes	5	100	**=105**
	=55	**=110**	

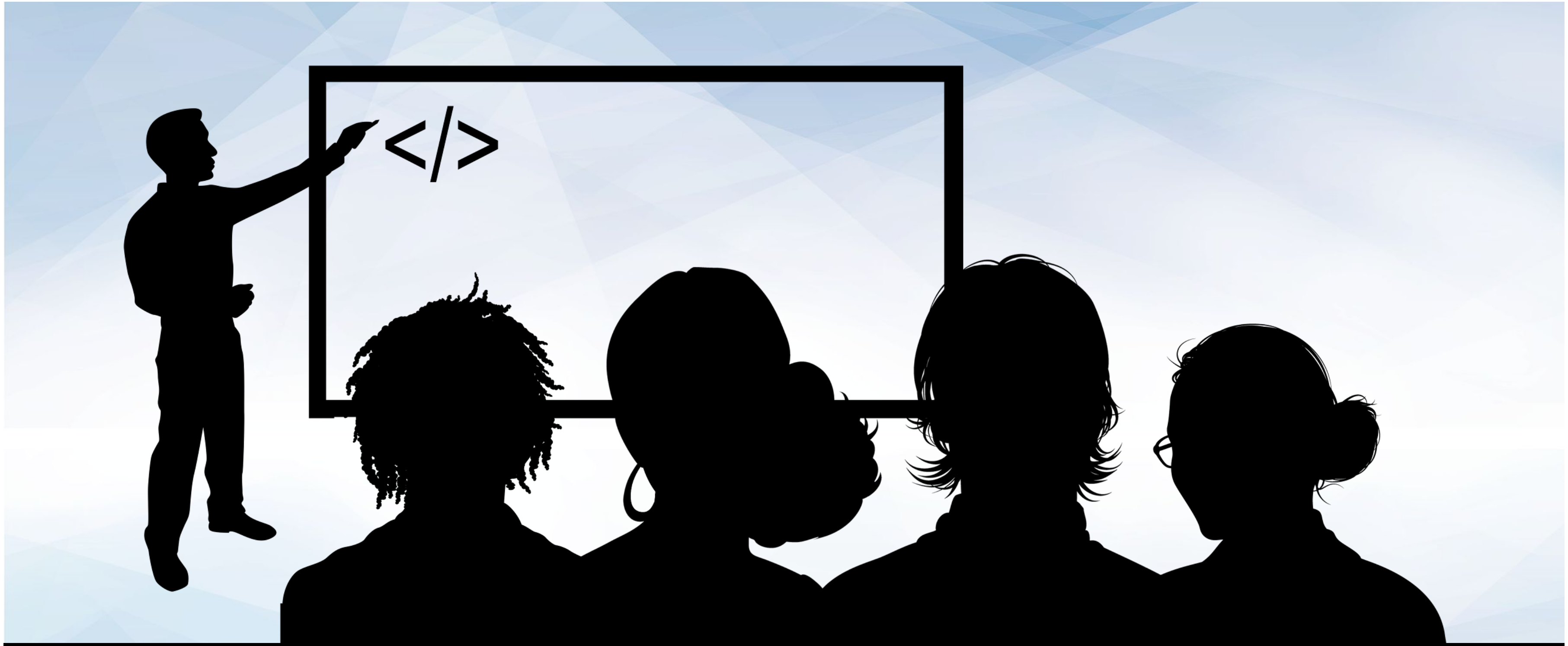
Classification Report

Classification report identifies the **precision**, **recall**, and **accuracy** of a model for each given class.

```
              precision    recall  f1-score   support

 No Diabetes       0.77      0.90      0.83       125
    Diabetes       0.72      0.49      0.58        67

    accuracy                           0.76       192
   macro avg       0.74      0.69      0.71       192
weighted avg       0.75      0.76      0.74       192
```



Instructor Demonstration

Confusion Matrix & Classification Report

Activity: Diagnosing the Model

In this activity, you will return to the model you created to predict diabetes and will use a confusion matrix and classification report to evaluate and diagnose the model.

Suggested Time:
10 minutes

Time’s Up! Let’s Review.

Activity: Build Loan Approver

In this activity you will apply the machine learning concepts and technical skills learned thus far to create a model for approving loans.

Suggested Time:
15 minutes

Time’s Up! Let’s Review.

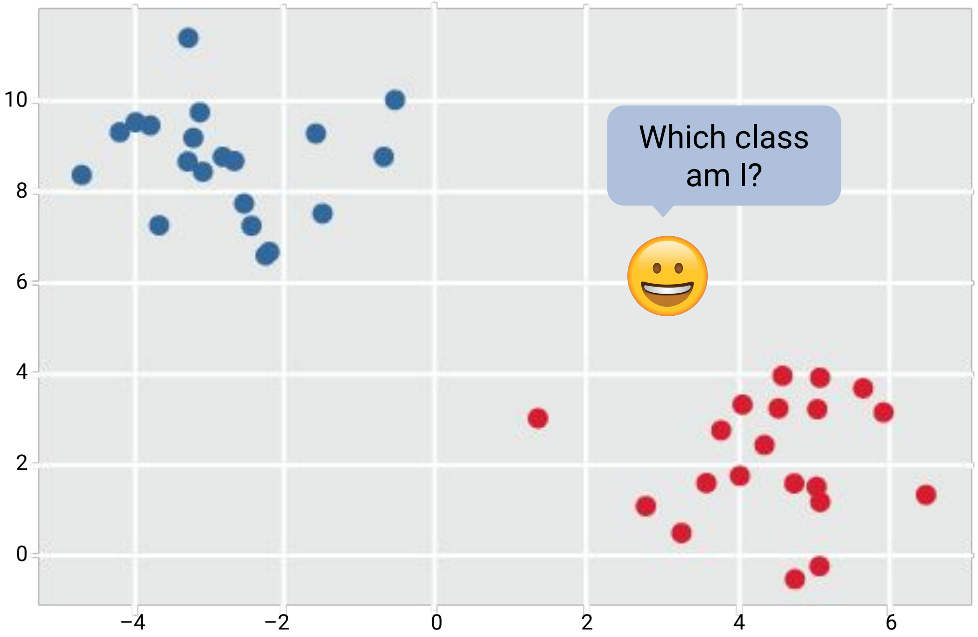
Break

Support Vector Machines

Support Vector Machines (SVM) is a supervised learning model that can be used for classification and regression analysis. SVM separates classes of data points into multidimensional space.

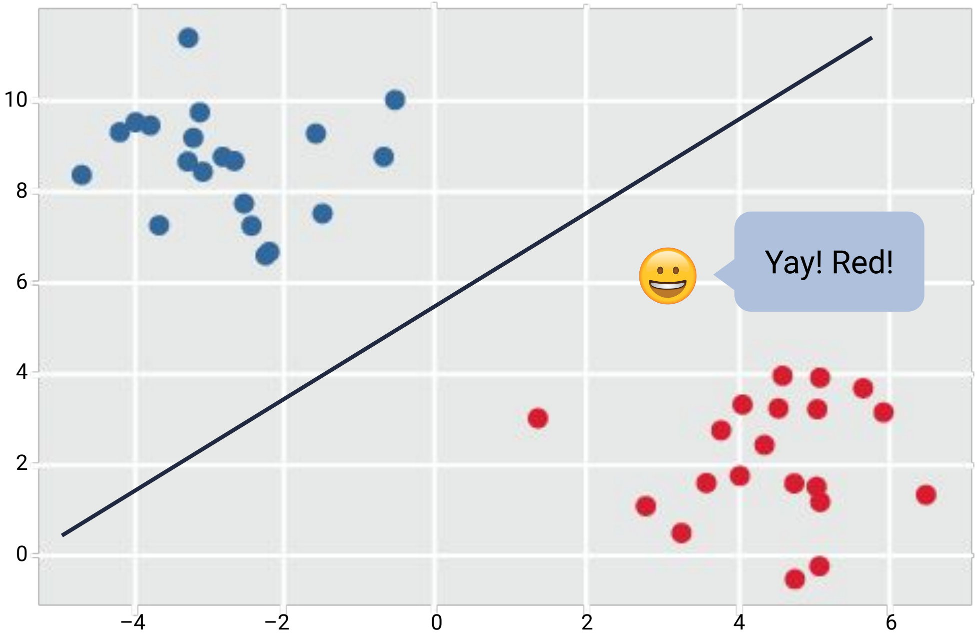
Linear Classifiers

Linear classifiers attempt to draw a line that separates the data, but which line best separates the groups?



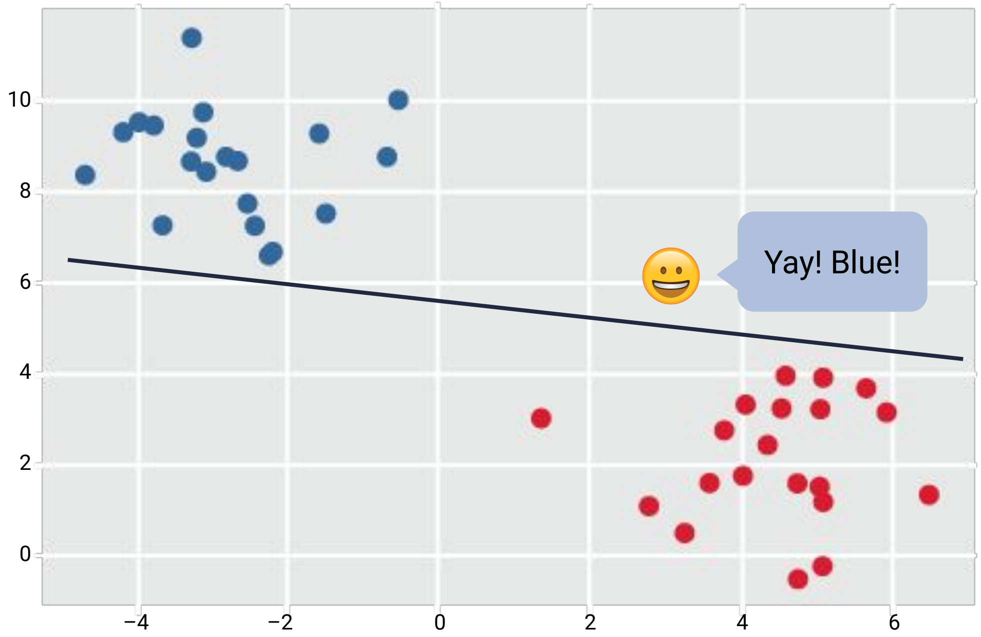
Linear Classifiers

Linear classifiers attempt to draw a line that separates the data, but which line best separates the groups?



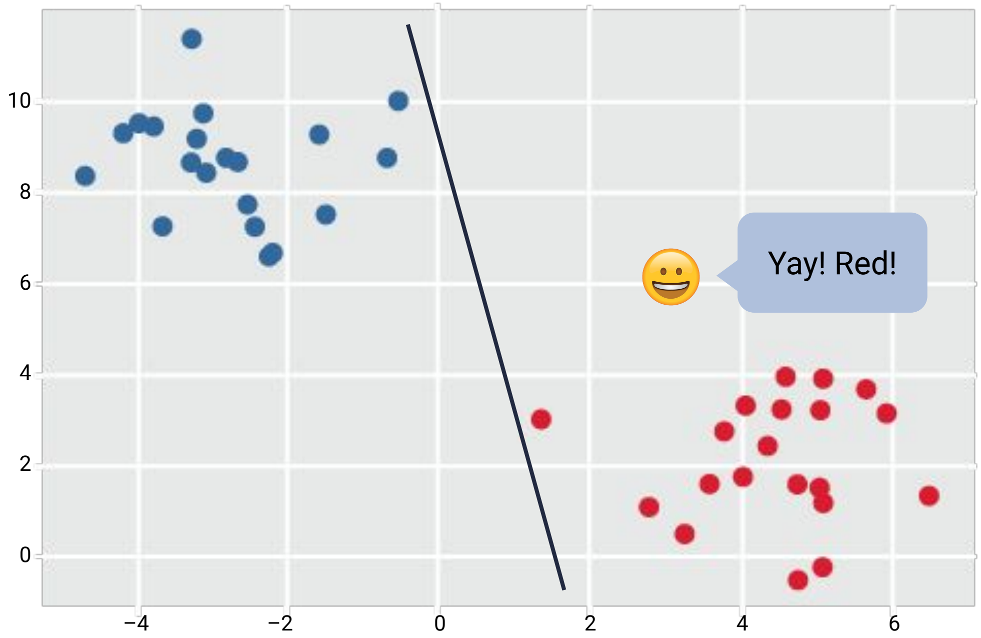
Linear Classifiers

Linear classifiers attempt to draw a line that separates the data, but which line best separates the groups?



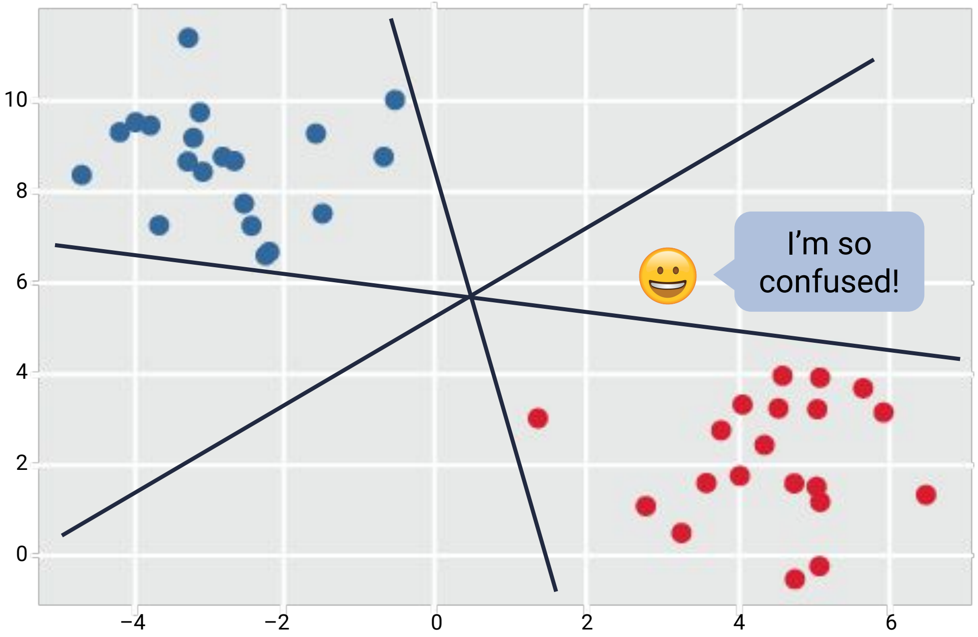
Linear Classifiers

Linear classifiers attempt to draw a line that separates the data, but which line best separates the groups?



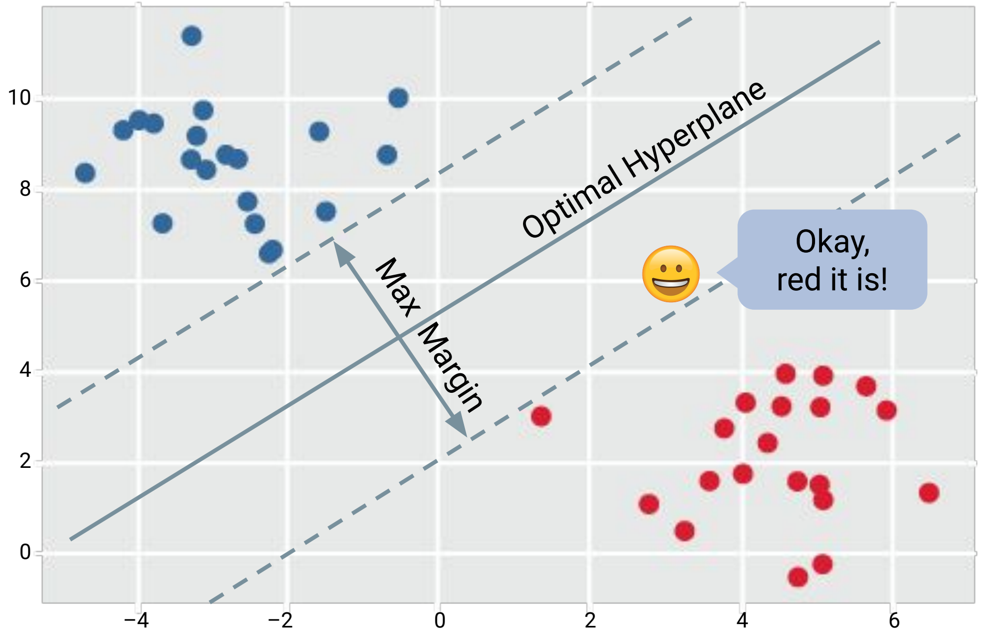
Linear Classifiers

Linear classifiers attempt to draw a line that separates the data, but which line best separates the groups?



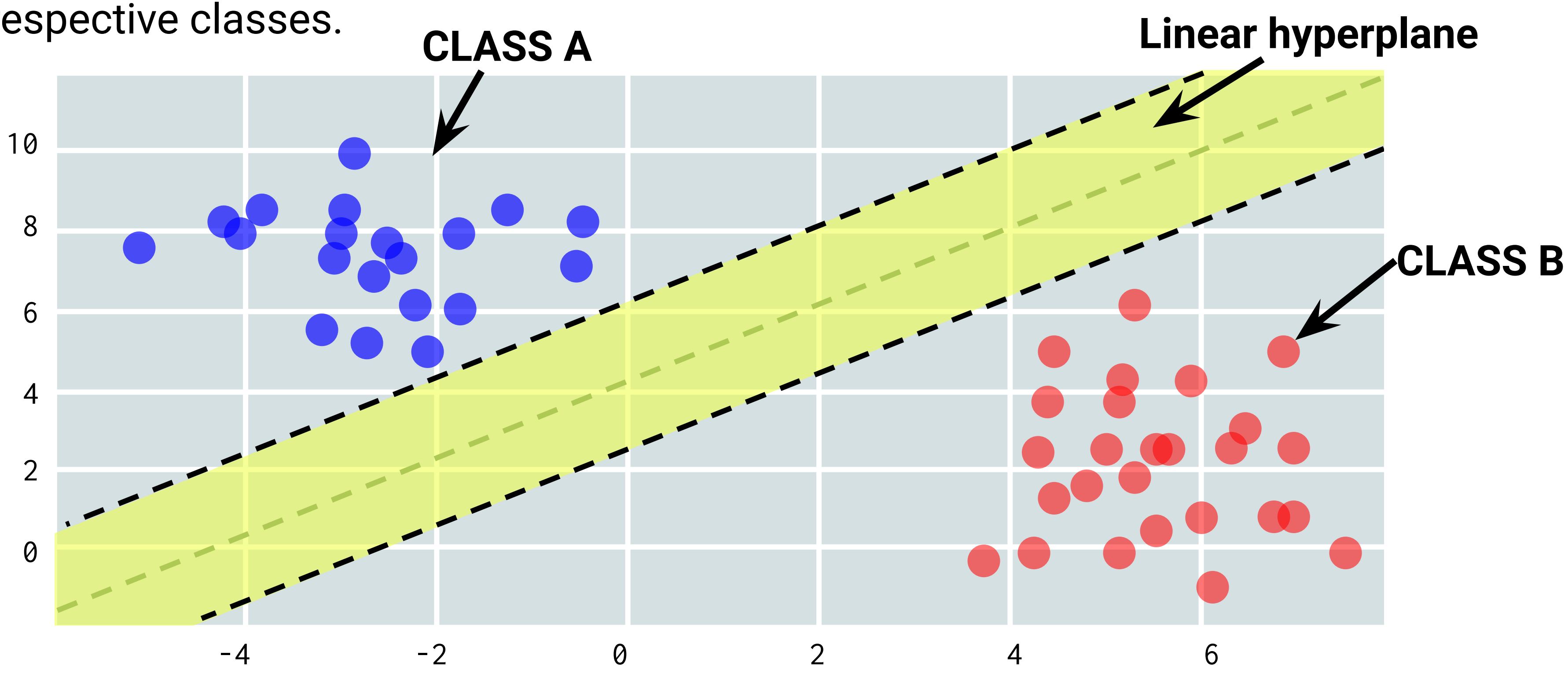
Support Vector Machines

The Support Vector Machines (SVM) algorithm finds the optimal hyperplane that separates the data points with the largest margin possible.



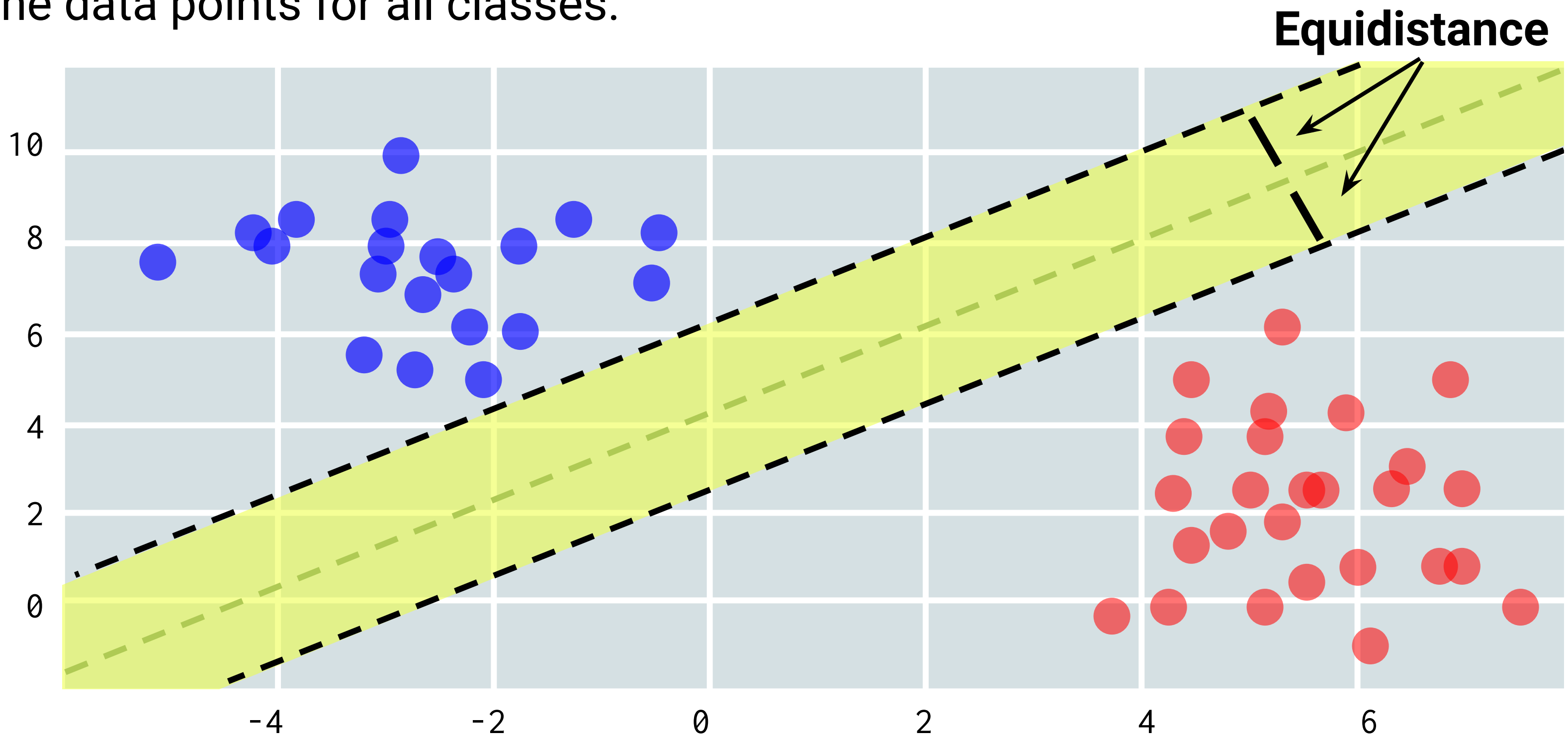
Support Vector Machines

The space is segmented by a line or plane that groups data points into their respective classes.



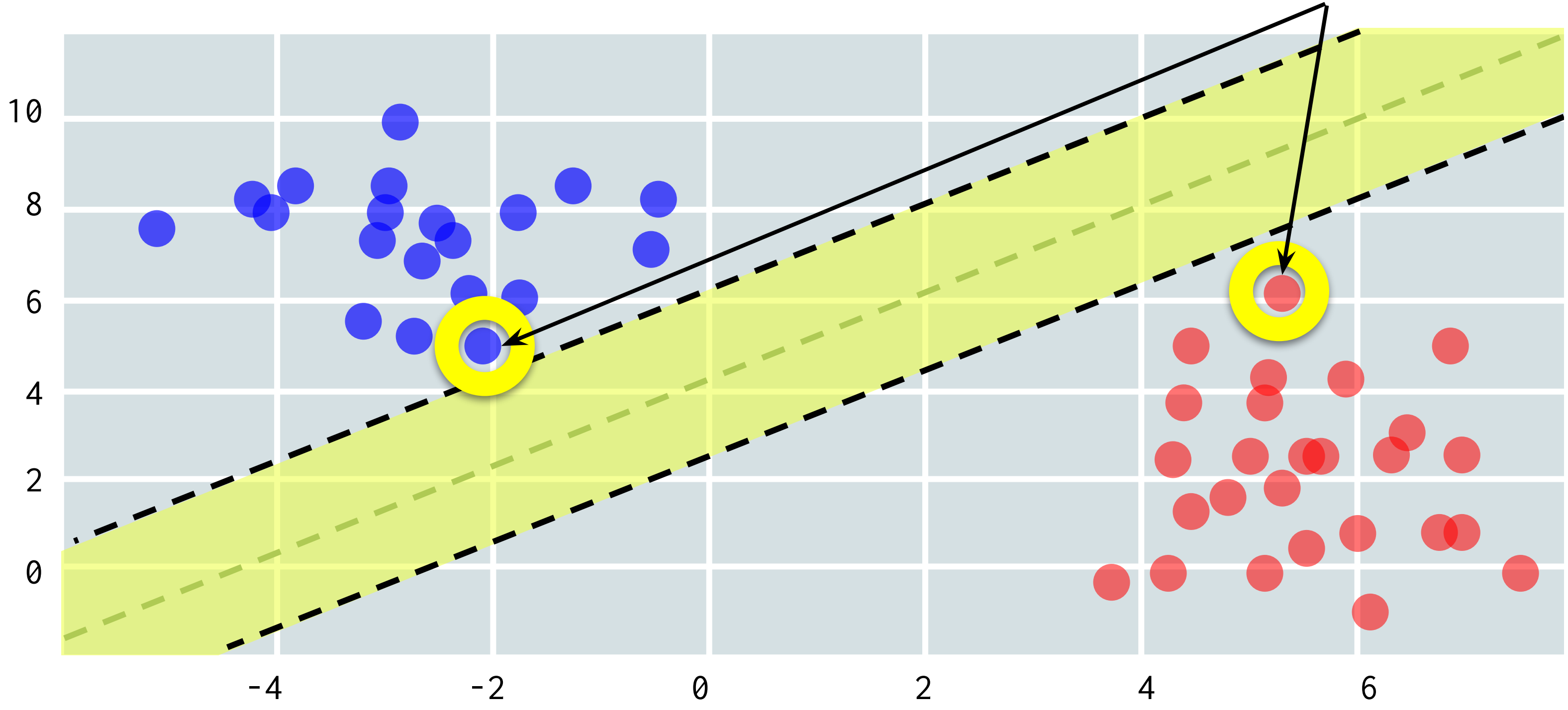
Support Vector Machines

The goal with hyperplanes is to get the margin of the hyperplane equidistant to the data points for all classes.



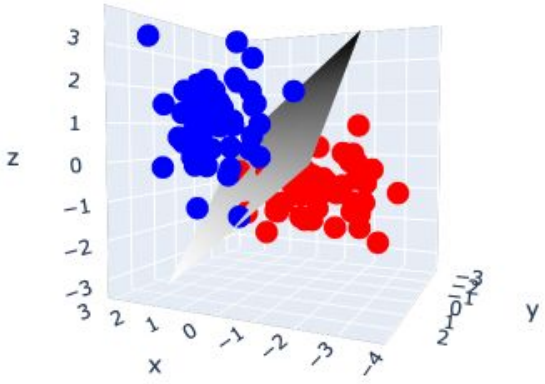
Support Vector Machines

The data closest to/within the margin of the hyperplane are called support vectors, and they are used to define boundaries of the hyperplane.



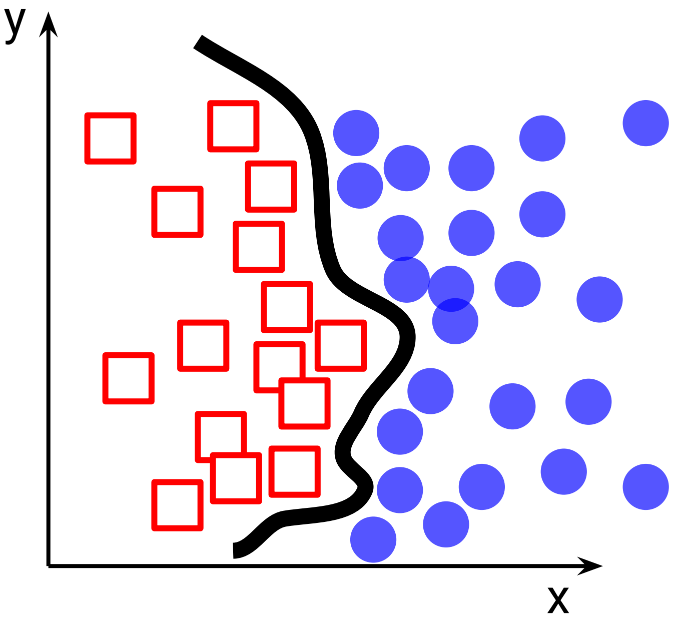
Hyperplanes

Hyperplanes can be used clearly delineate classes in multiple dimensions.



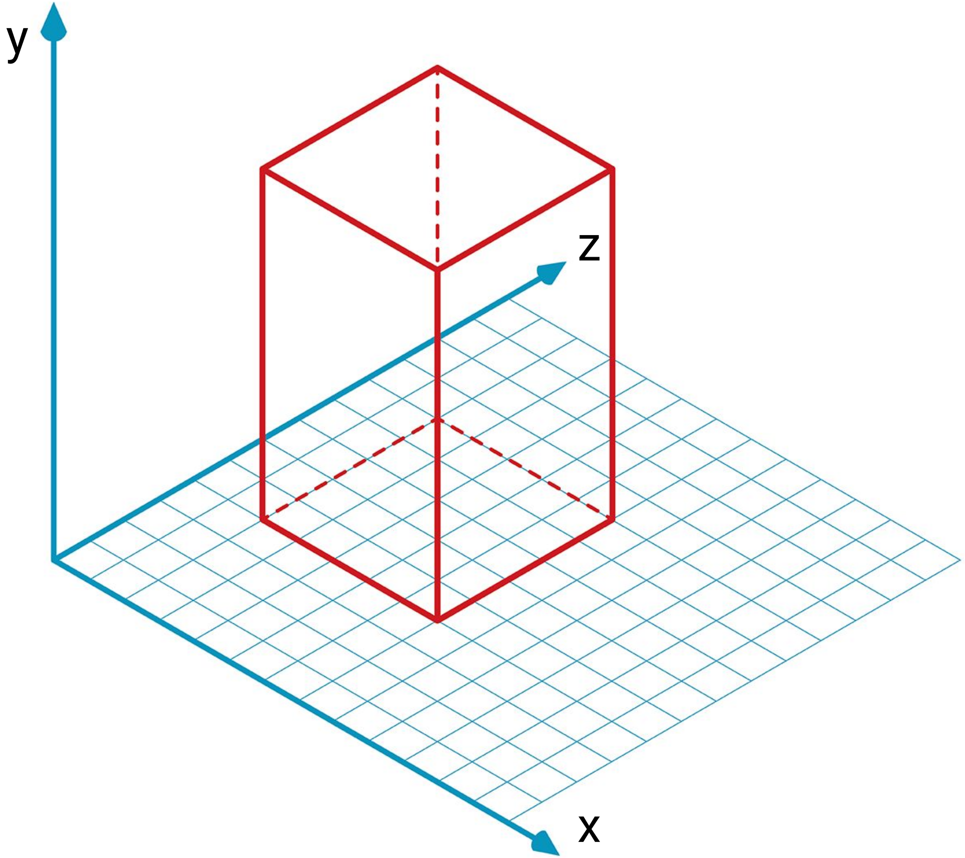
Zero tolerance with perfect partition

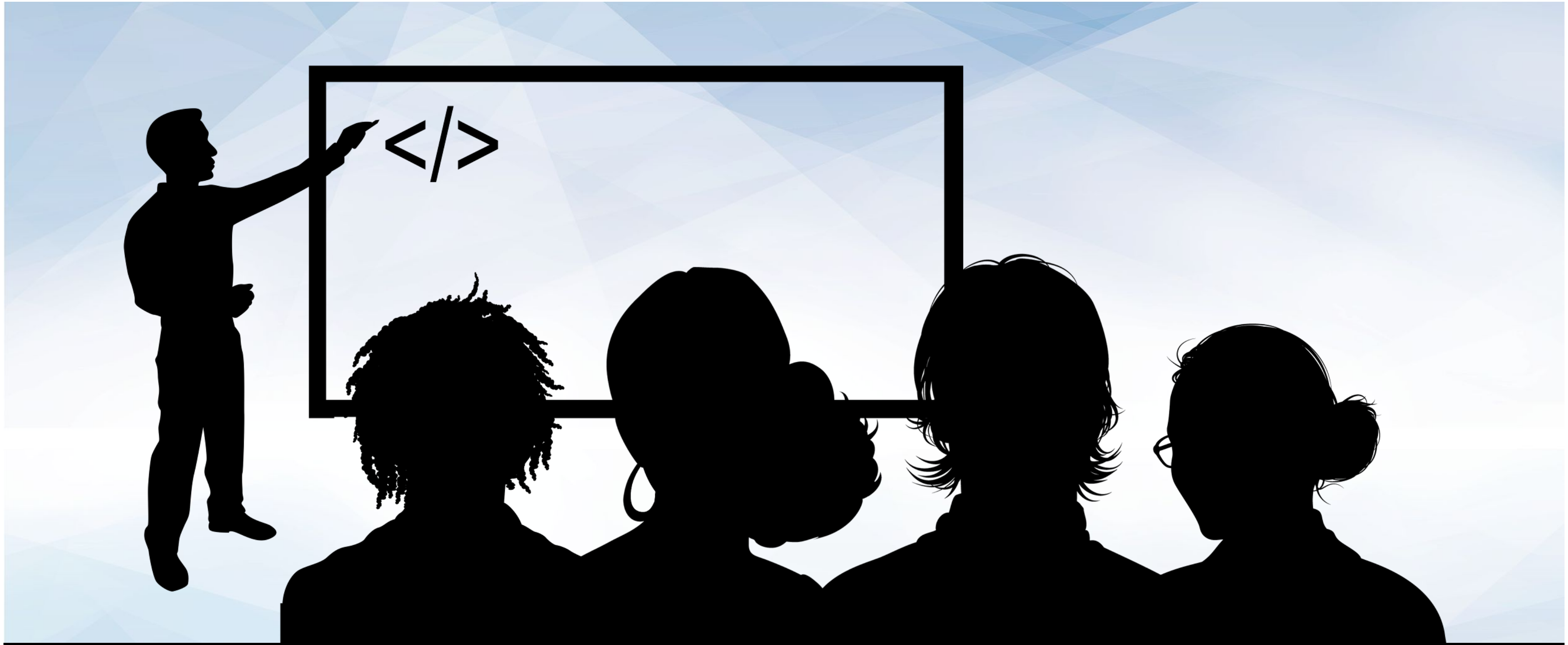
Hyperplane also supports what is considered zero tolerance with perfect partition, which is a nonlinear hyperplane that will position and orient the hyperplane to correctly classify overlapping or outlying data points.



Zero tolerance with perfect partition

In order to establish zero tolerance with perfect partition, the SVM model may introduce a new `z-axis` dimension for nonlinear hyperplanes.





Instructor Demonstration

SVM model with sklearn

SVM model

Steps to implement an SVM model include:

01 Create the model with appropriate `kernel` parameters

02 Fit the model

03 Extract min and max decision boundaries and store in a mesh grid

04 Execute the `decision_function` to get classifier scores for pre-existing data points

05 Run the `predict` function to classify new data points

Activity: SVM Loan Approver Activity Review

In this activity you will update your loan approver with an SVM model and rerun the evaluation metrics.

Suggested Time:
15 minutes

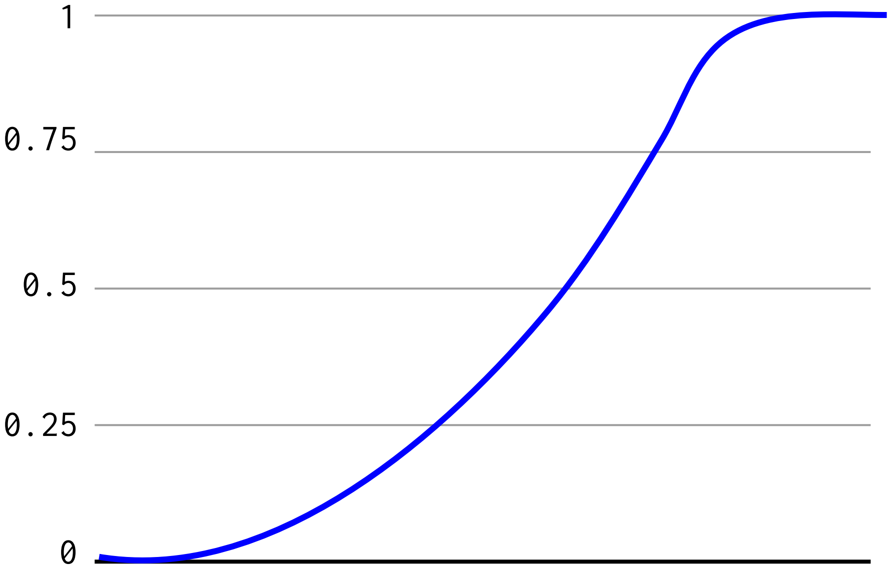
Time’s Up! Let’s Review.

Which Model is the Best?

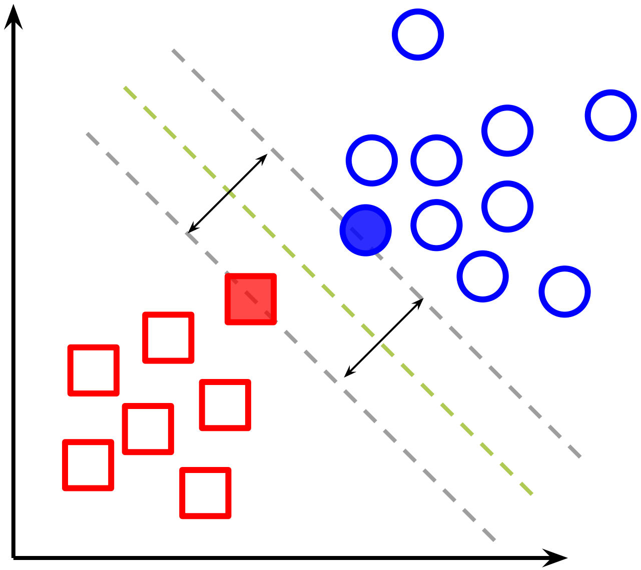
Which Model is the Best?

Both the Logistic Regression and SVM models were both able to predict outcomes; however, the important question is **which model performed best?**

Logistic Regression



Support Vector Machines



Which is the best approach to evaluate both models.

Answer:

Compare the confusion matrices and classification reports.

Confusion Matrices:

n=165	**Predicted: No**	**Predicted: Yes**	
Actual= No	50	10	**=60**
Actual= Yes	5	100	**=105**
	=55	**=110**	

Classification Reports:

```
              precision    recall   f1-score   support

No Diabetes        0.77      0.90       0.83       125
   Diabetes        0.72      0.49       0.58        67

   accuracy                             0.76       192
  macro avg        0.74      0.69       0.71       192
weighted avg       0.75      0.76       0.74       192
```

What is the best approach to evaluate both models?

Logistic Regression Loan Approver Classification Report

```
              precision    recall  fl-score   support

     approve       0.44      0.33      0.38        12
        deny       0.50      0.62      0.55        13

   micro avg       0.48      0.48      0.48        25
   macro avg       0.47      0.47      0.47        25
weighted avg       0.47      0.48      0.47        25
```

SVM Loan Approver Classification Report

```
              precision    recall  fl-score   support

     approve       0.58      0.58      0.58        12
        deny       0.62      0.62      0.62        13

    accuracy                           0.60        25
   macro avg       0.60      0.60      0.60        25
weighted avg       0.60      0.60      0.60        25
```

What is the best approach to evaluate both models?

The SVM model performed best. **Precision**, **recall**, and **accuracy** were all higher for the SVM loan approver.

```
              precision    recall  f1-score   support

     approve       0.44      0.33      0.38        12
        deny       0.50      0.62      0.55        13

   micro avg       0.48      0.48      0.48        25
   macro avg       0.47      0.47      0.47        25
weighted avg       0.47      0.48      0.47        25
```

```
              precision    recall  f1-score   support

     approve       0.58      0.58      0.58        12
        deny       0.62      0.62      0.62        13

    accuracy                           0.60        25
   macro avg       0.60      0.60      0.60        25
weighted avg       0.60      0.60      0.60        25
```

What is the best approach to evaluate both models?

Recall percentage for deny is the same for the SVM and logistic regression loan approver, meaning both algorithms correctly predicted the same number of true positive denies.

```
              precision    recall  fl-score   support

     approve       0.44      0.33      0.38        12
        deny       0.50      0.62      0.55        13

   micro avg       0.48      0.48      0.48        25
   macro avg       0.47      0.47      0.47        25
weighted avg       0.47      0.48      0.47        25
```

```
              precision    recall  fl-score   support

     approve       0.58      0.58      0.58        12
        deny       0.62      0.62      0.62        13

    accuracy                           0.60        25
   macro avg       0.60      0.60      0.60        25
weighted avg       0.60      0.60      0.60        25
```

Questions?