Lesson 4.3: Machine Learning Fundamentals

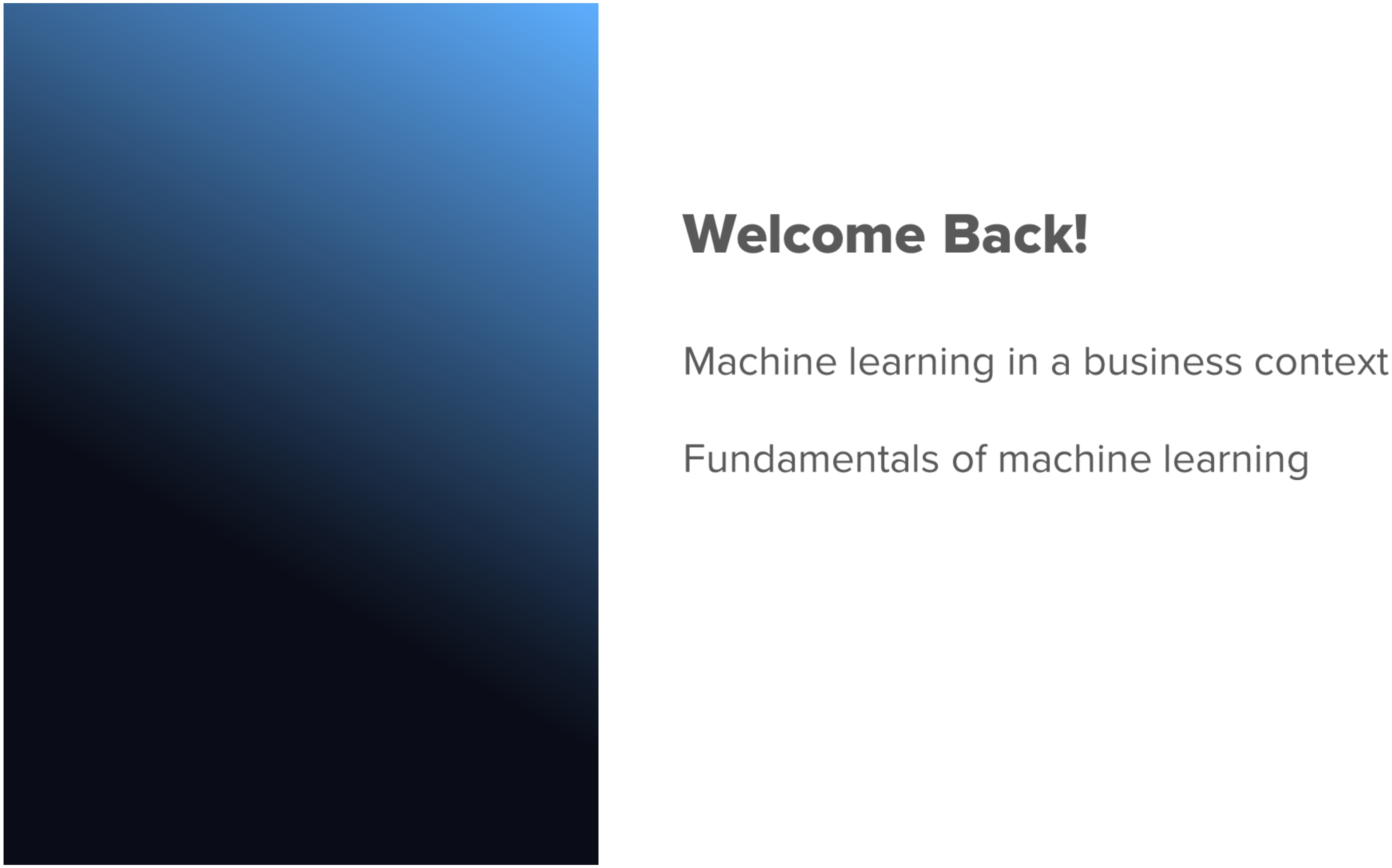
DISTRIBUTED COMPUTING WITH SPARK SQL

Machine Learning Fundamentals

Brooke Wenig/Conor Murphy
Machine Learning Practice Lead/
Data Scientist, Databricks

UCDAVIS
Continuing and Professional Education

Slide 2: Welcome Back!



Welcome Back!

Machine learning in a business context

Fundamentals of machine learning

Slide 3: Learning Objectives

Learning Objectives

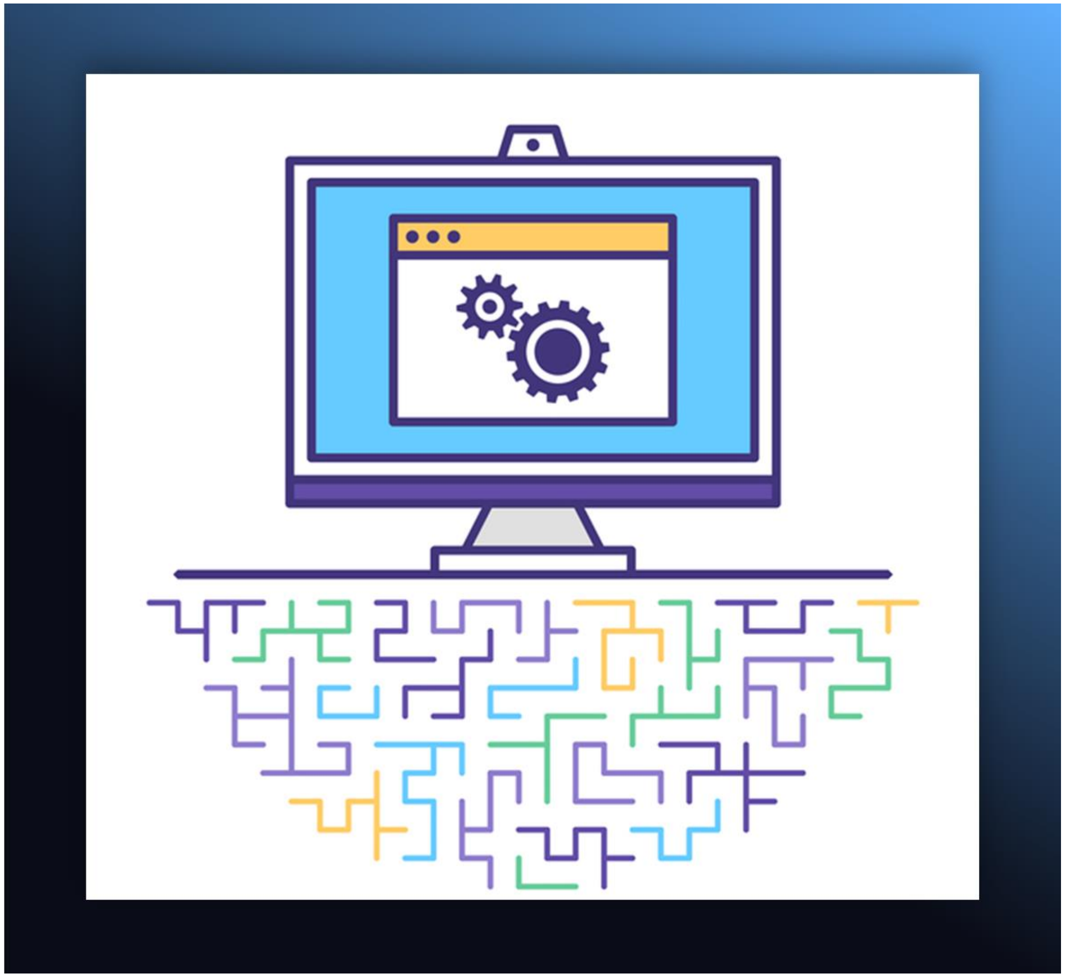
Differentiate between regression and classification use cases

Quantify what is model success

Slide 4: What is Machine Learning?

What Is Machine Learning?

A broad array of techniques that learns patterns and data, without being explicitly programmed



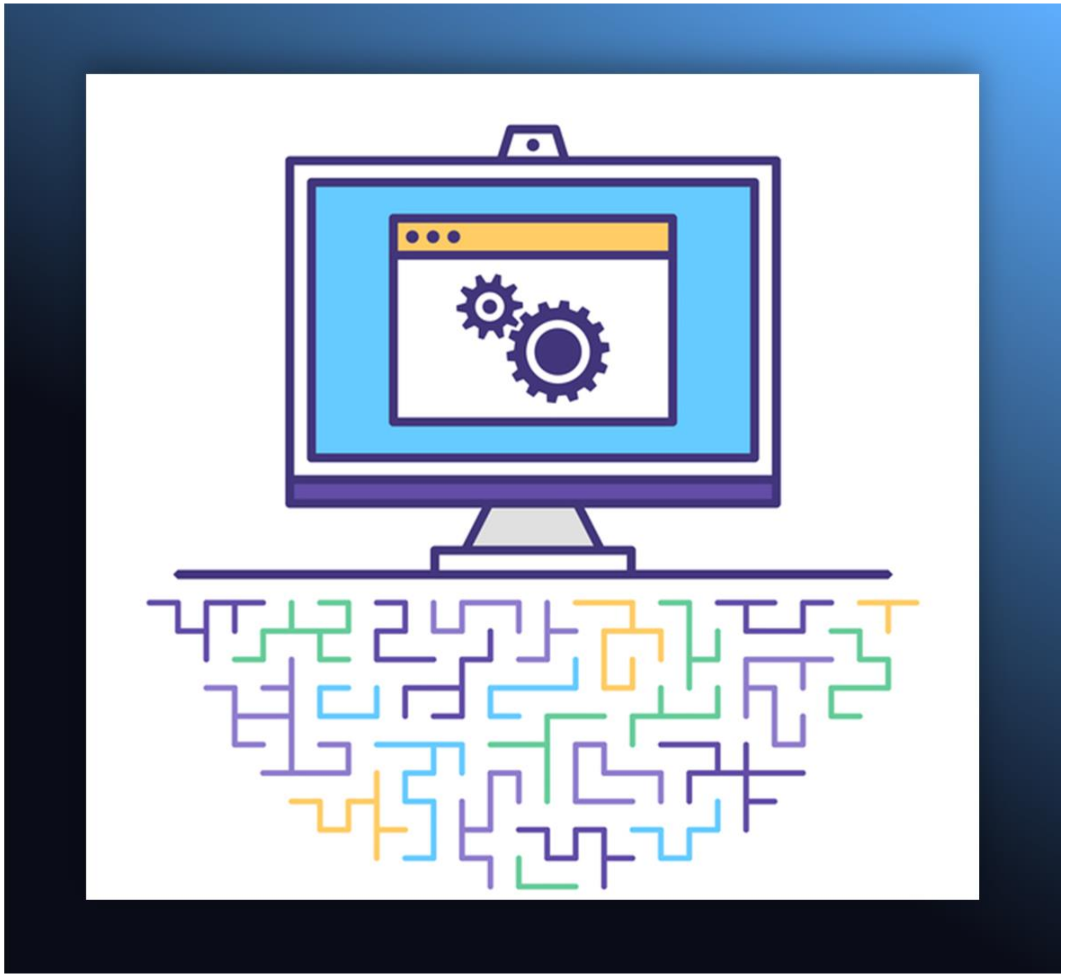
Slide 5: Churn Analysis

Churn Analysis

Machine learning would use past user data to find patterns between variables

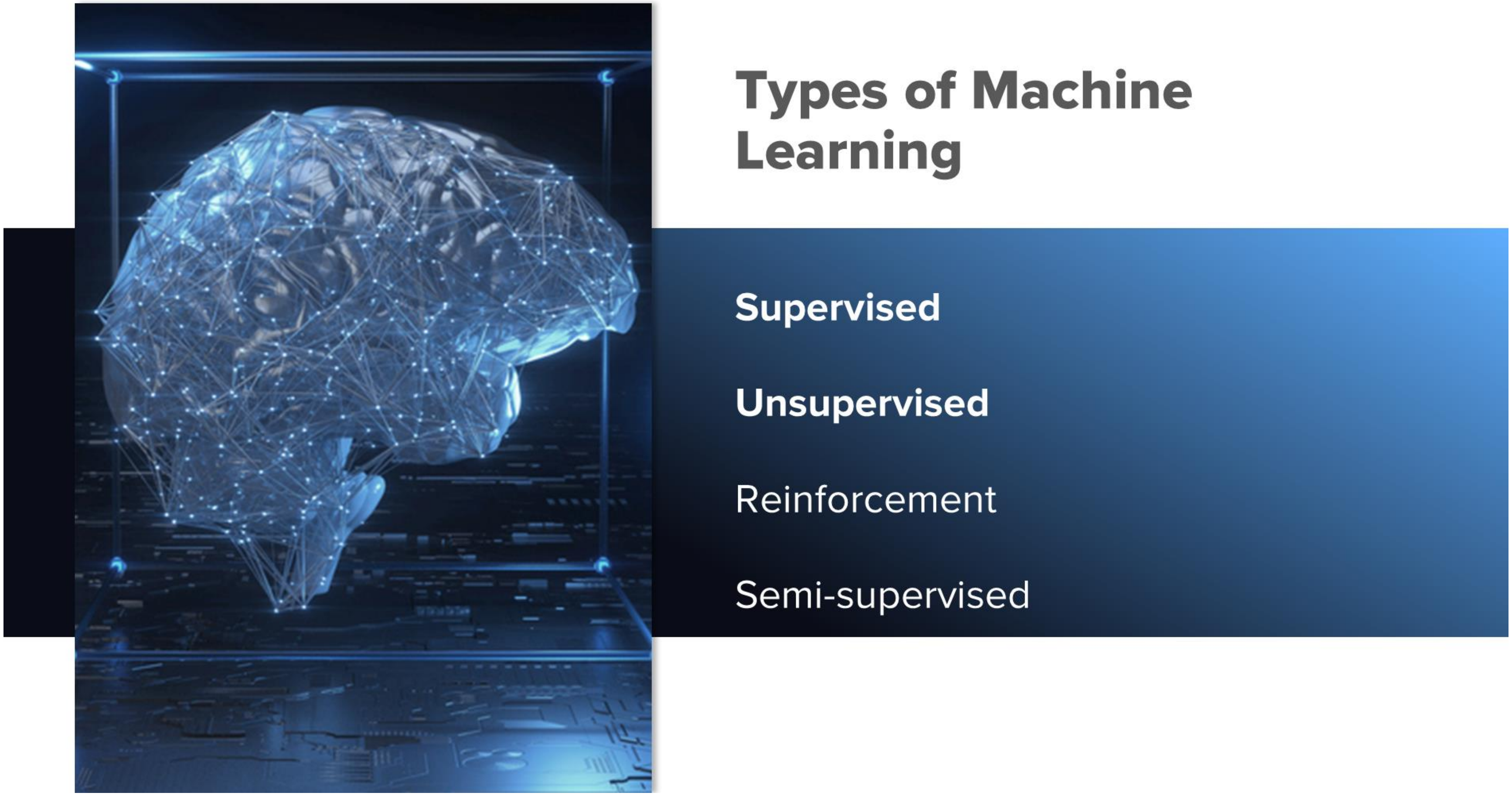
Slide 6: What is Machine Learning?

What is Machine Learning?



A function that maps features to an output

Slide 7: Types of Machine Learning



Types of Machine Learning

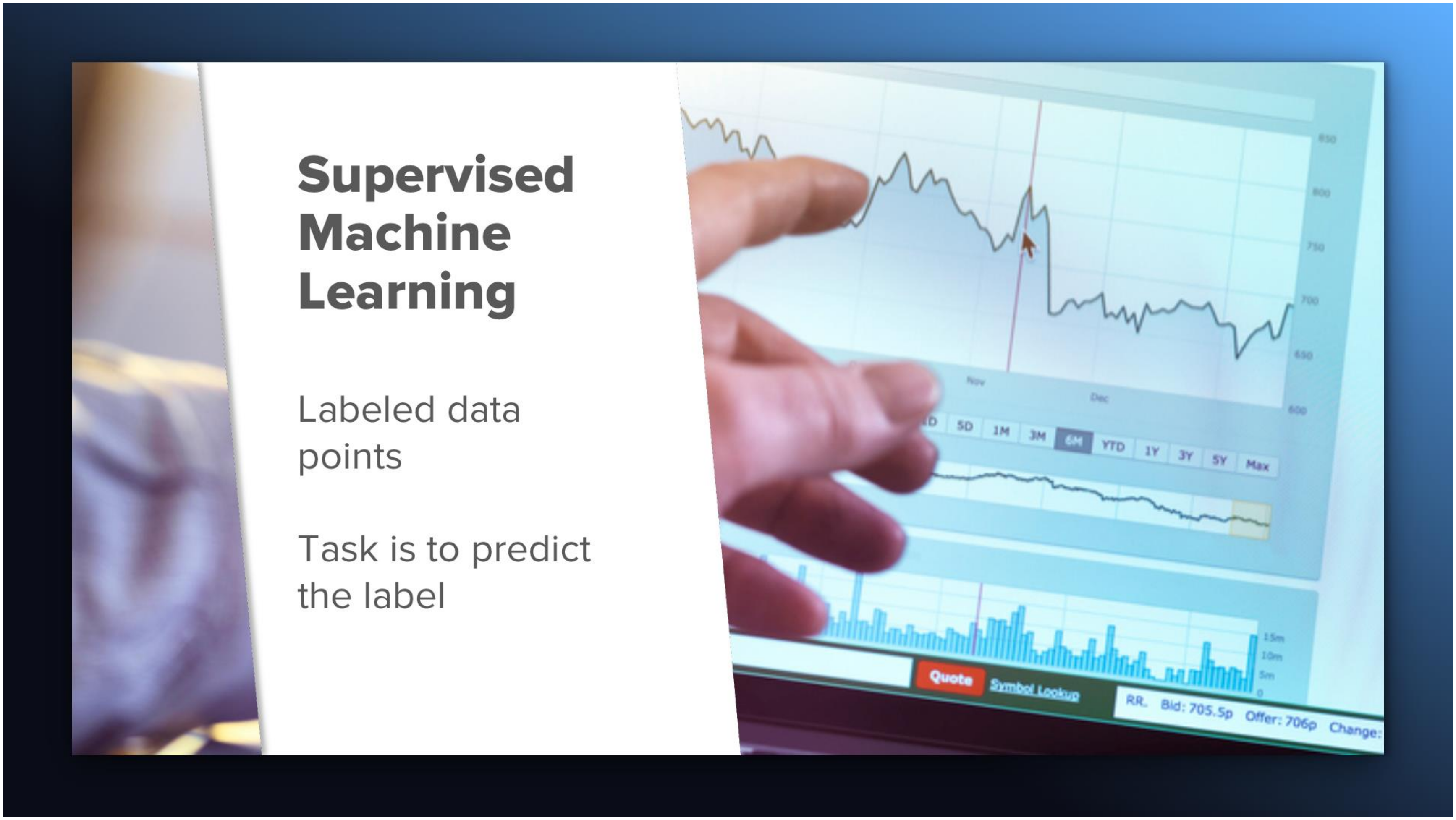
Supervised

Unsupervised

Reinforcement

Semi-supervised

Slide 8: Supervised Machine Learning



Supervised Machine Learning

Labeled data points

Task is to predict the label

Slide 9: Classification Tasks

Classification Tasks

Predicts a discrete set of categories

- Binary classification
- Multiclass classification

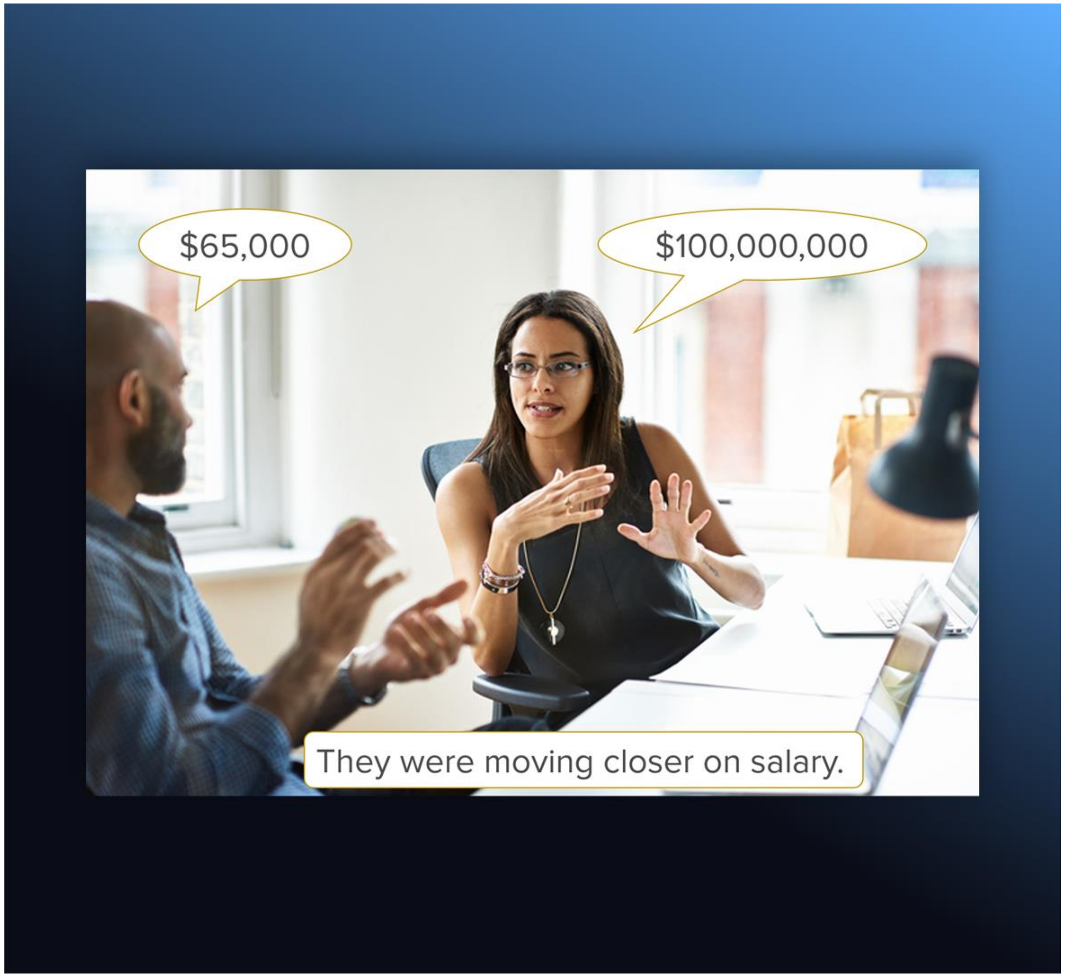
Slide 10: Regression Tasks

Regression Tasks

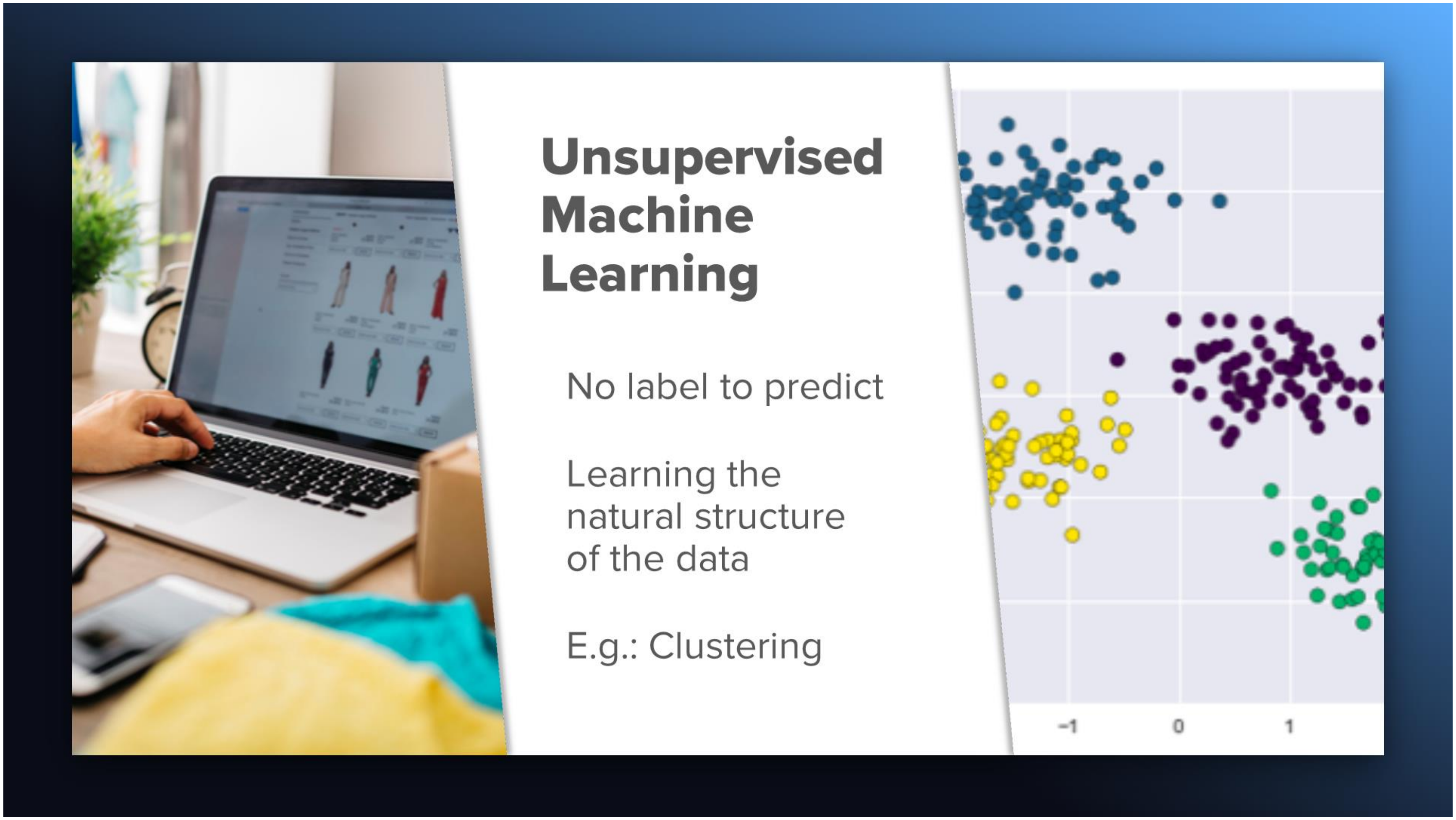
Predict a continuous value

Financial forecasting

Unbounded number rather than a category



Slide 11: Unsupervised Machine Learning



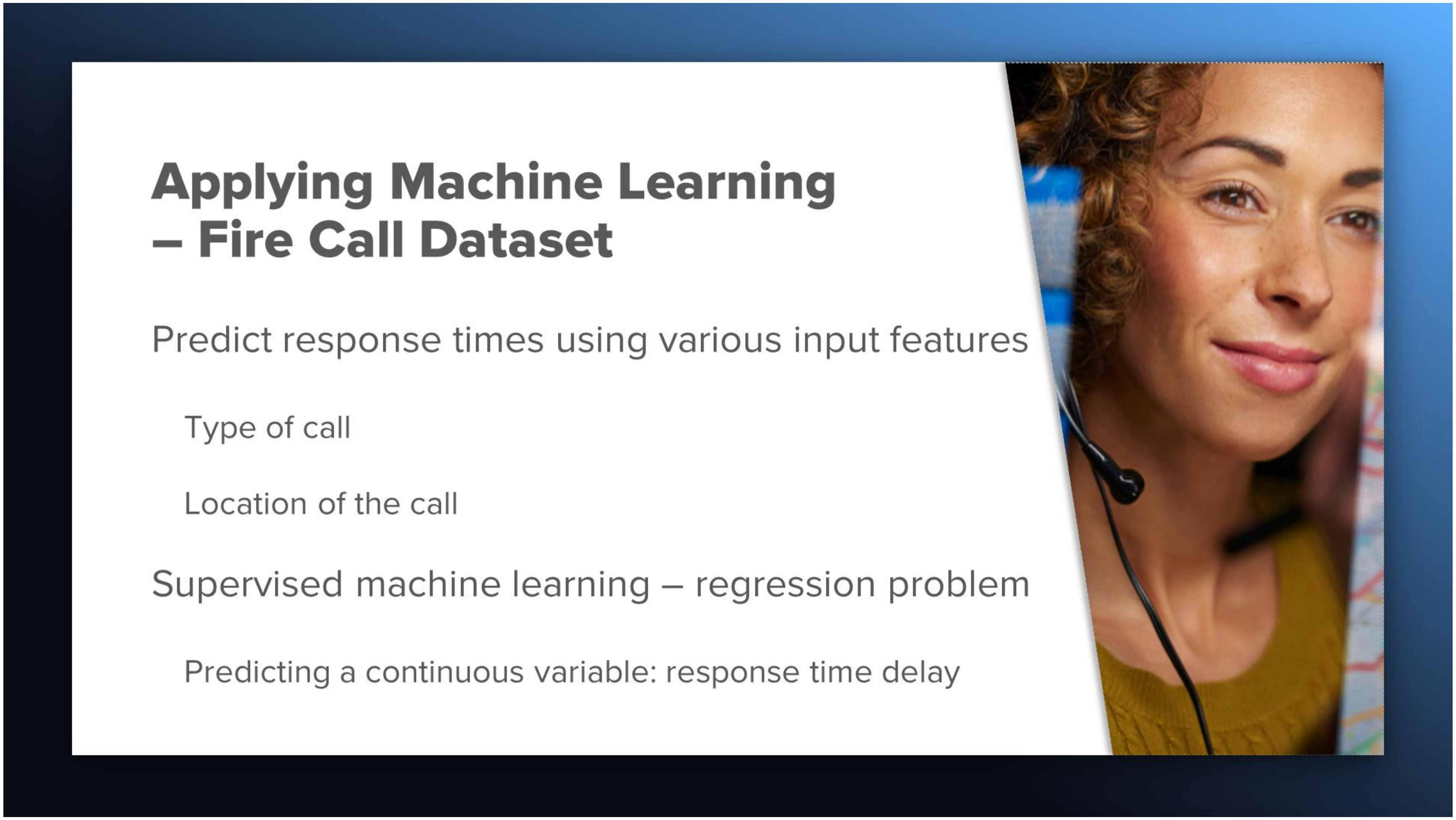
Unsupervised Machine Learning

No label to predict

Learning the natural structure of the data

E.g.: Clustering

Slide 12: Applying Machine Learning – Fire Call Dataset



Applying Machine Learning – Fire Call Dataset

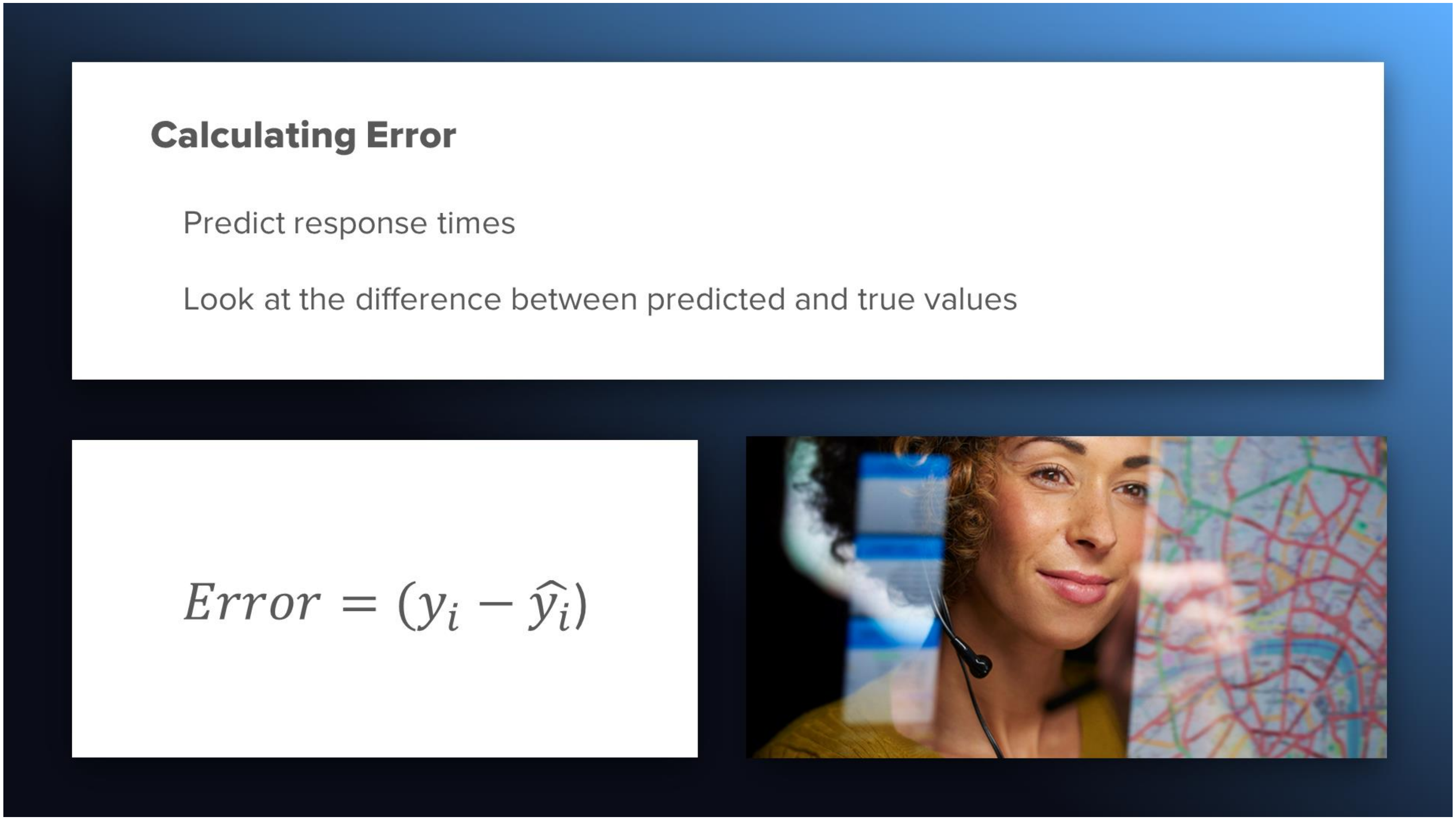
Predict response times using various input features

- Type of call
- Location of the call

Supervised machine learning – regression problem

- Predicting a continuous variable: response time delay

Slide 13: Calculating Error



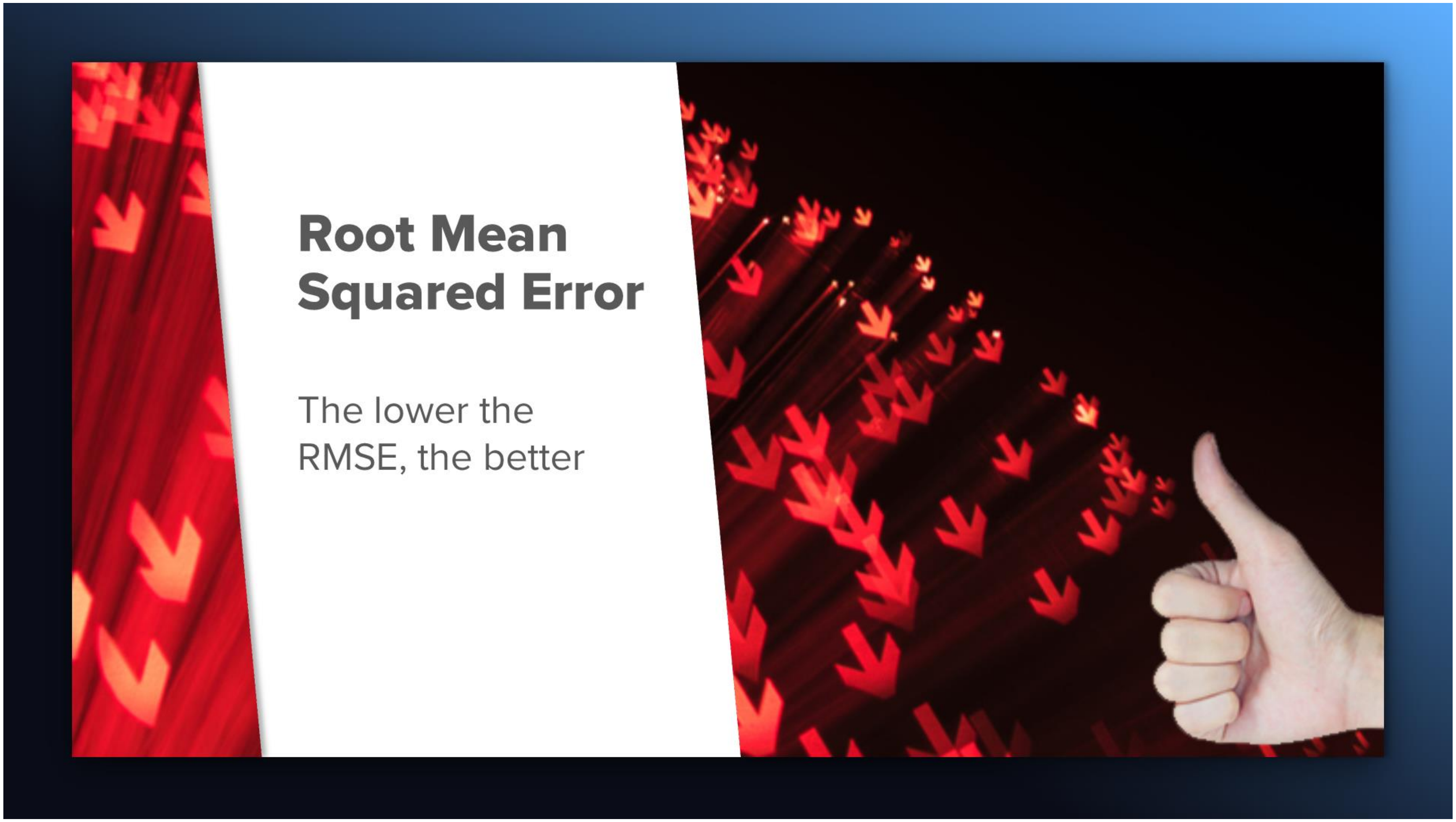
Calculating Error

Predict response times

Look at the difference between predicted and true values

$$Error = (y_i - \widehat{y_i})$$

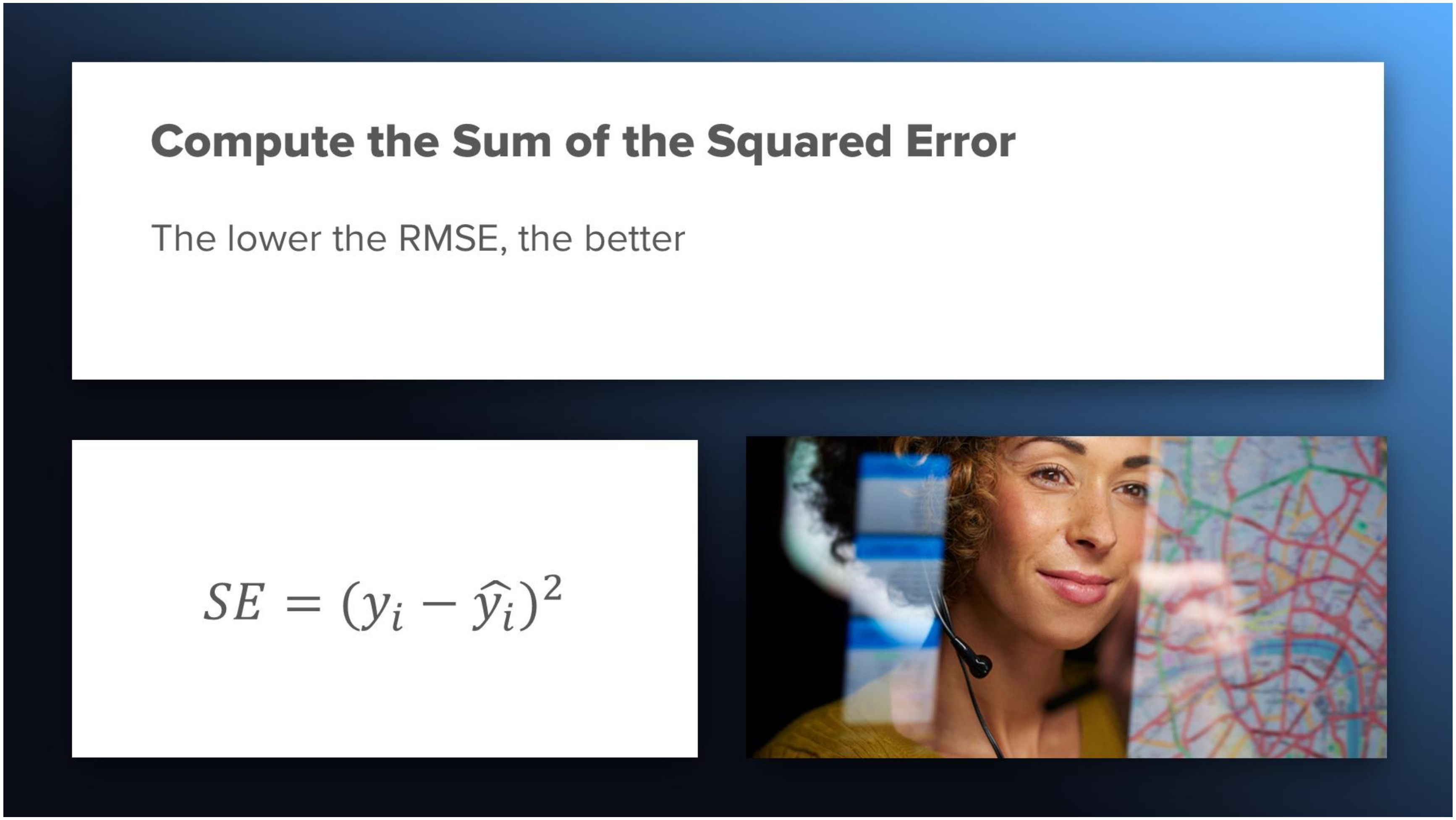
Slide 14: Root Means Squared Error



Root Mean Squared Error

The lower the RMSE, the better

Slide 15: Compute the Sum of the Squared Error

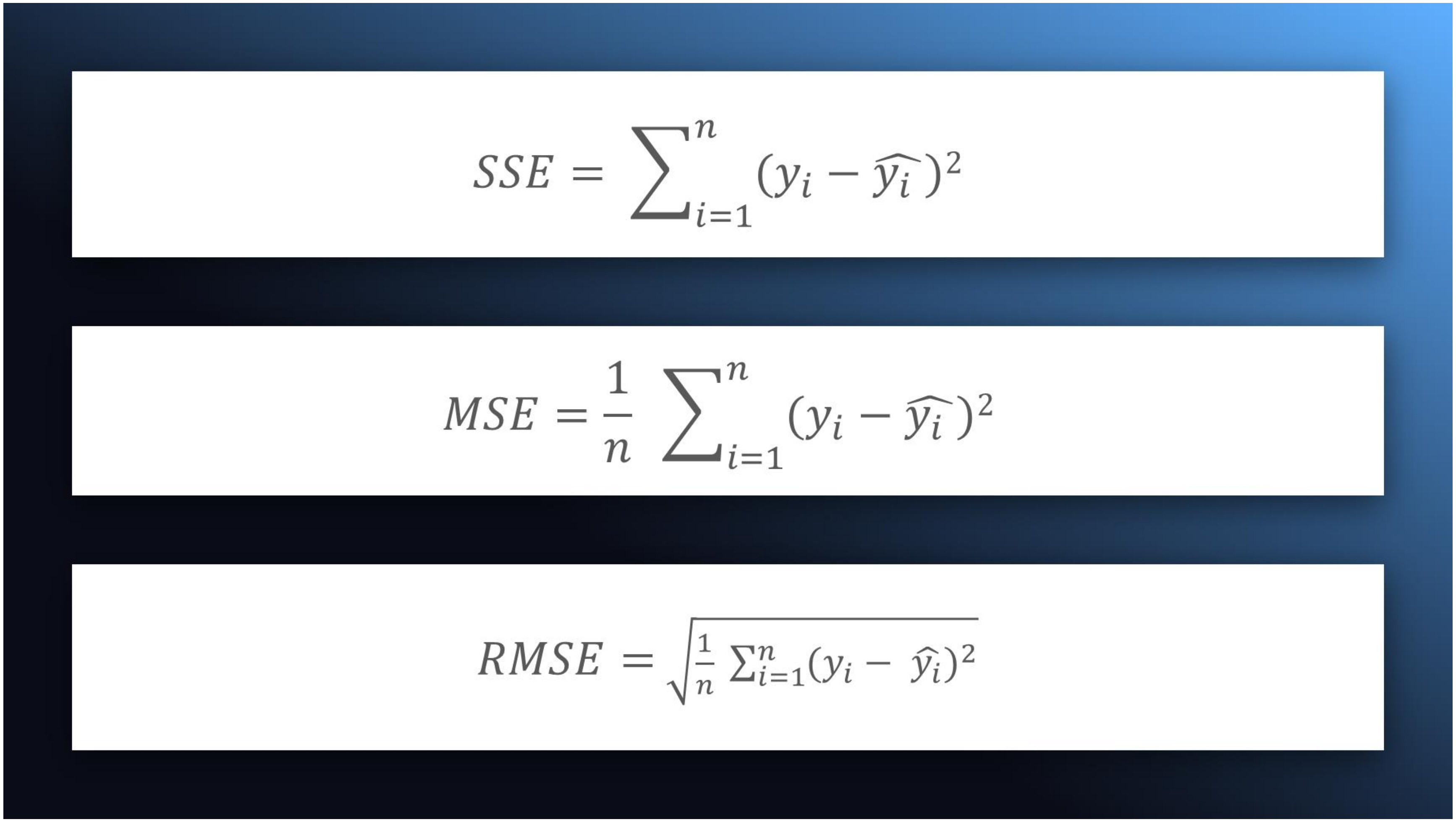


Compute the Sum of the Squared Error

The lower the RMSE, the better

$$SE = (y_i - \widehat{y_i})^2$$

Slide 16: Compute the Sum of the Squared Error

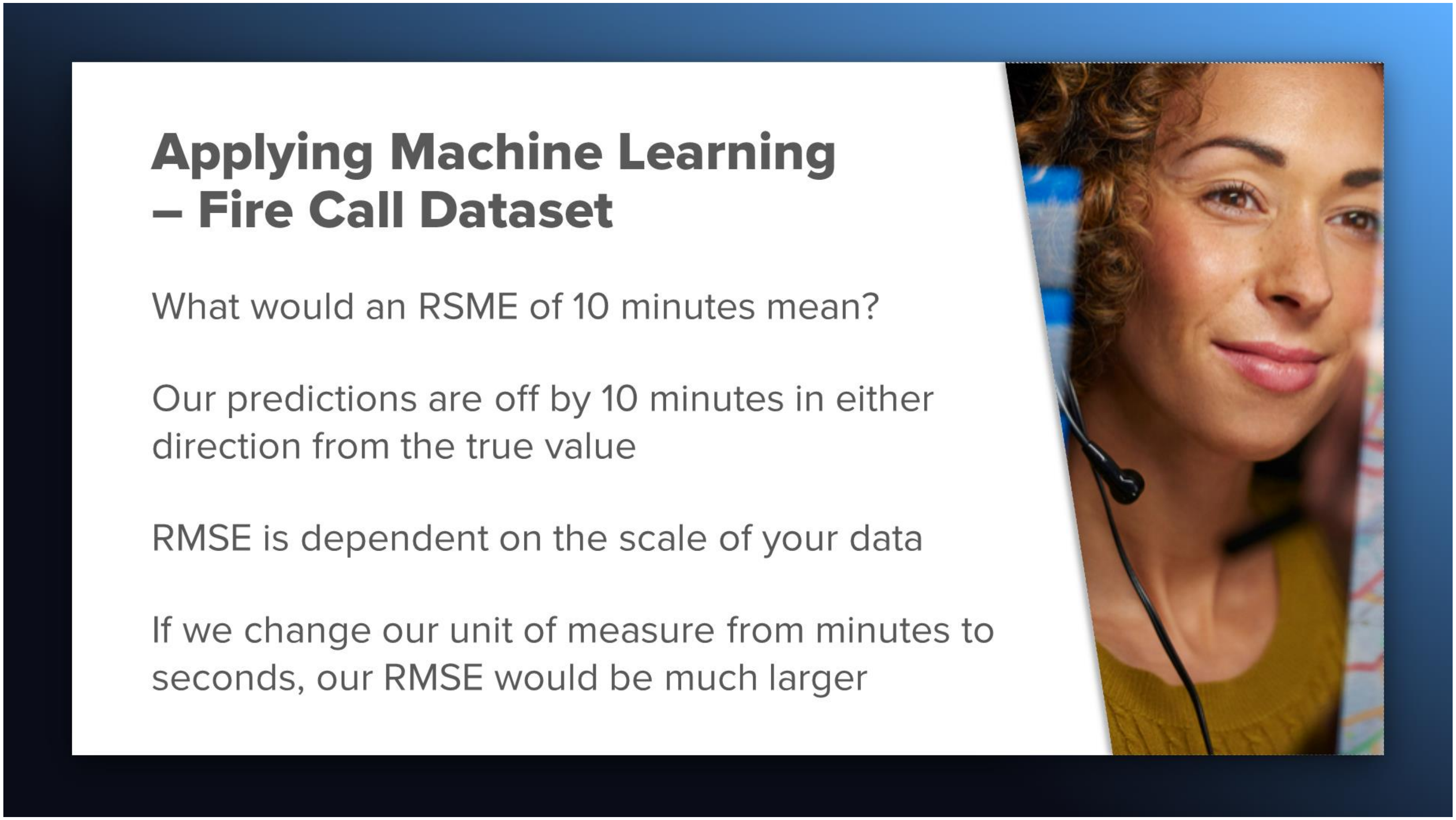


$$SSE = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y_i})^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Slide 17: Applying Machine Learning – Fire Call Dataset



Applying Machine Learning – Fire Call Dataset

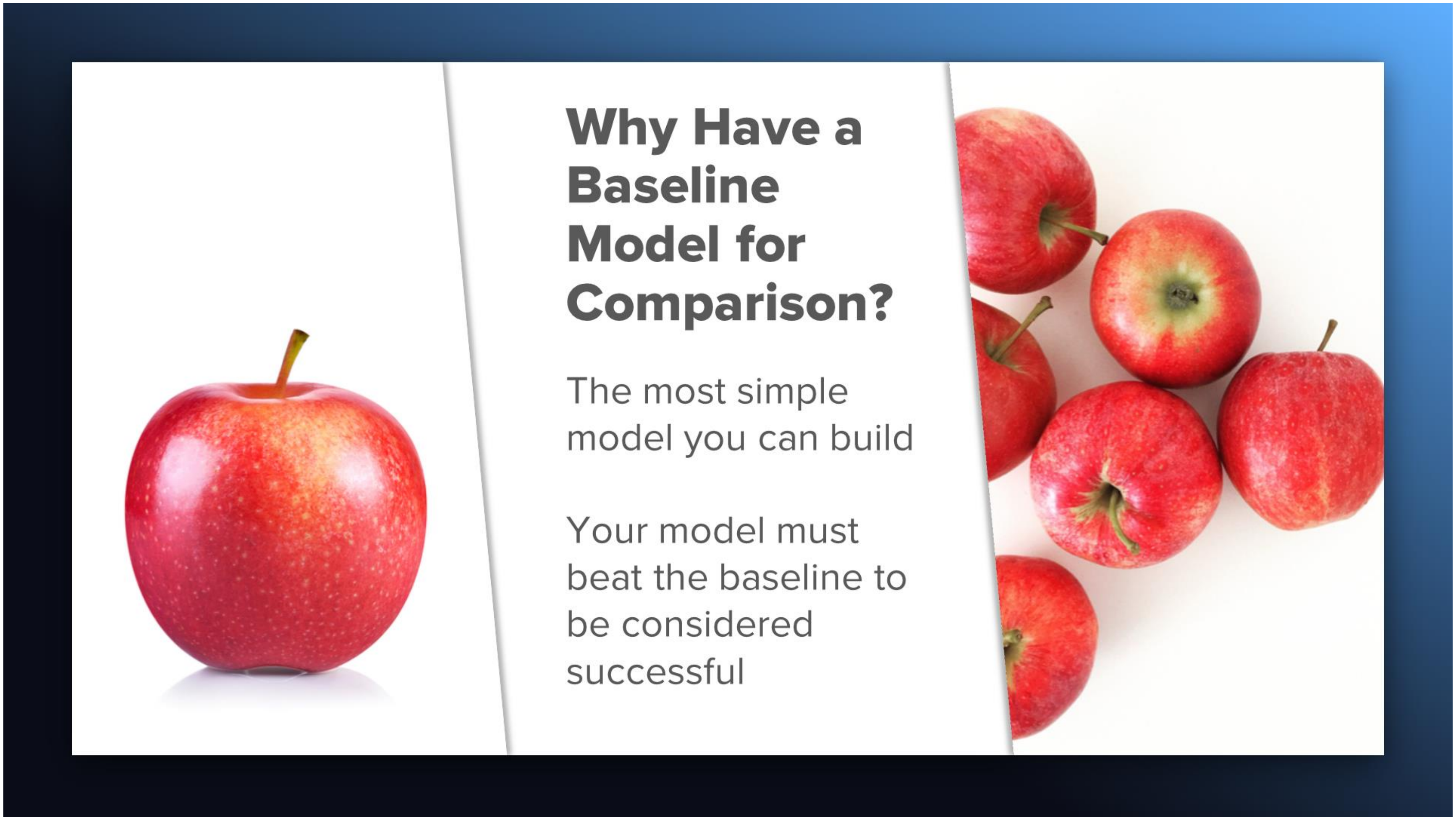
What would an RSME of 10 minutes mean?

Our predictions are off by 10 minutes in either direction from the true value

RMSE is dependent on the scale of your data

If we change our unit of measure from minutes to seconds, our RMSE would be much larger

Slide 18: Why Have a Baseline Model for Comparison?

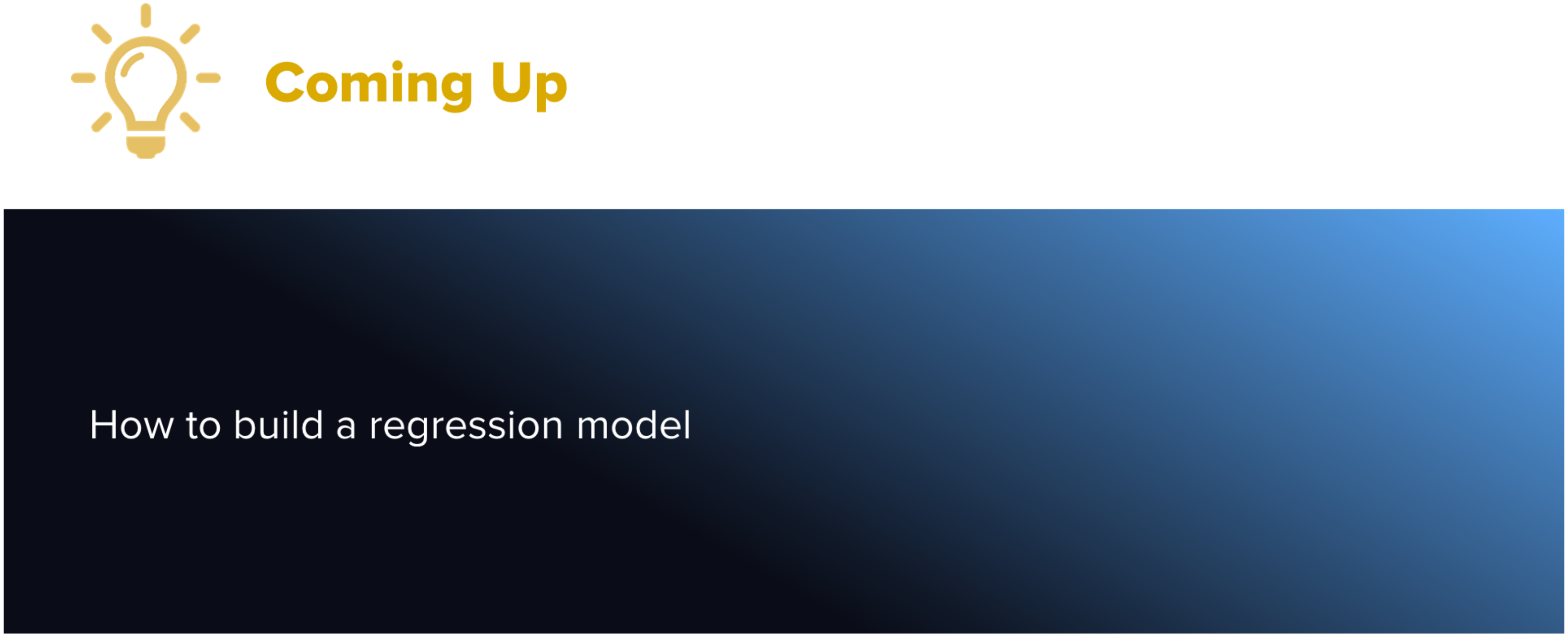


Why Have a Baseline Model for Comparison?

The most simple model you can build

Your model must beat the baseline to be considered successful

Slide 19: Coming Up



Coming Up

How to build a regression model