Apache Spark for Machine Learning

By: Dan Sullivan

Notes from: Vidal Sisneros

Begin Date: 1/30/2018

Introduction:

***Goal of the course:*** Learn how to use the Apache Spark Platform for ML.

**Structure of the course:**

* installing Spark
* reviewing the basics of the data frame data structure.
* how to pre-process both numeric and text data so that's ready to use with Spark's MLlib machine learning library.
* Describe multiple algorithms for clustering, classification, and regression.
* Briefly describe a recommendation system as well.

**Common pattern in Machine Learning:**

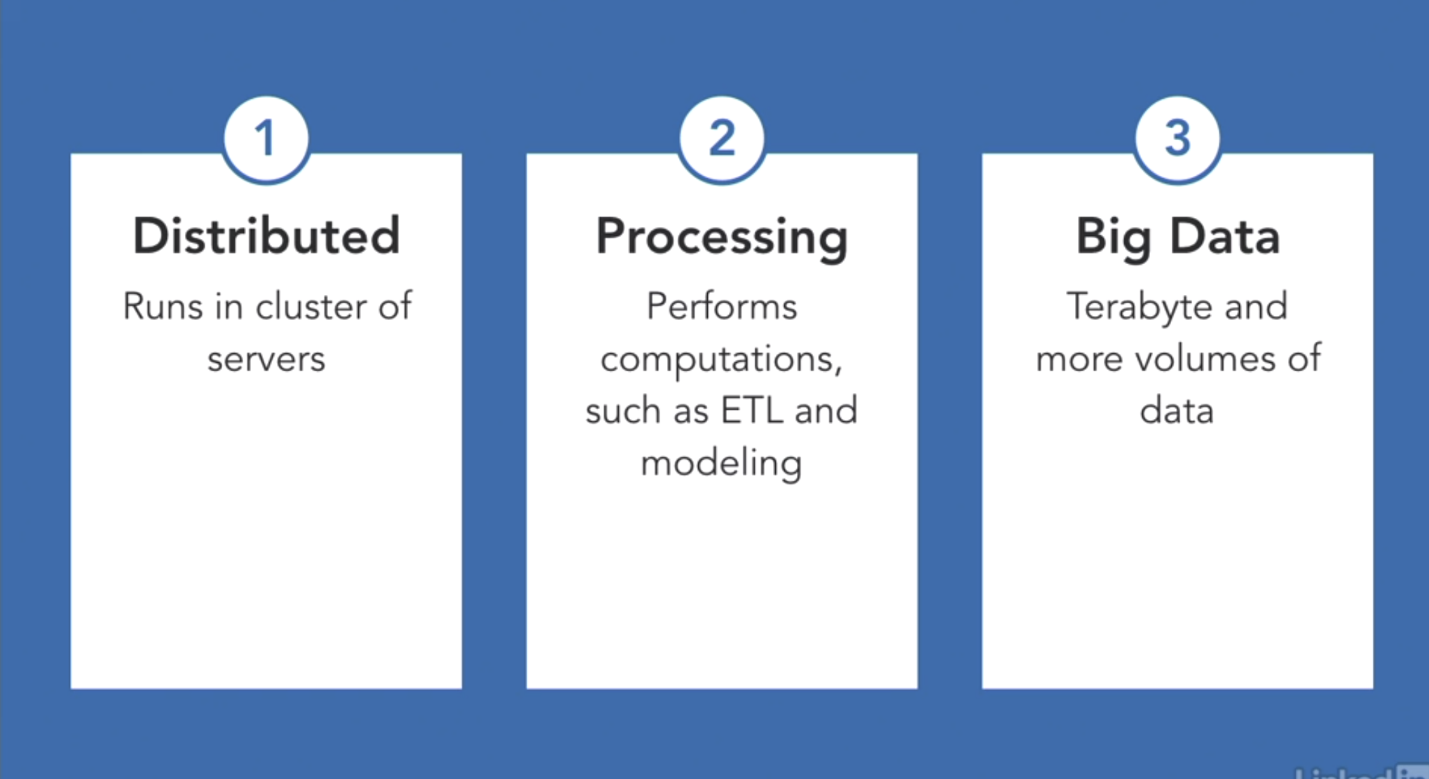
**pre-processing > model training > evaluation**

*that helps streamline the building of machine learning pipelines*

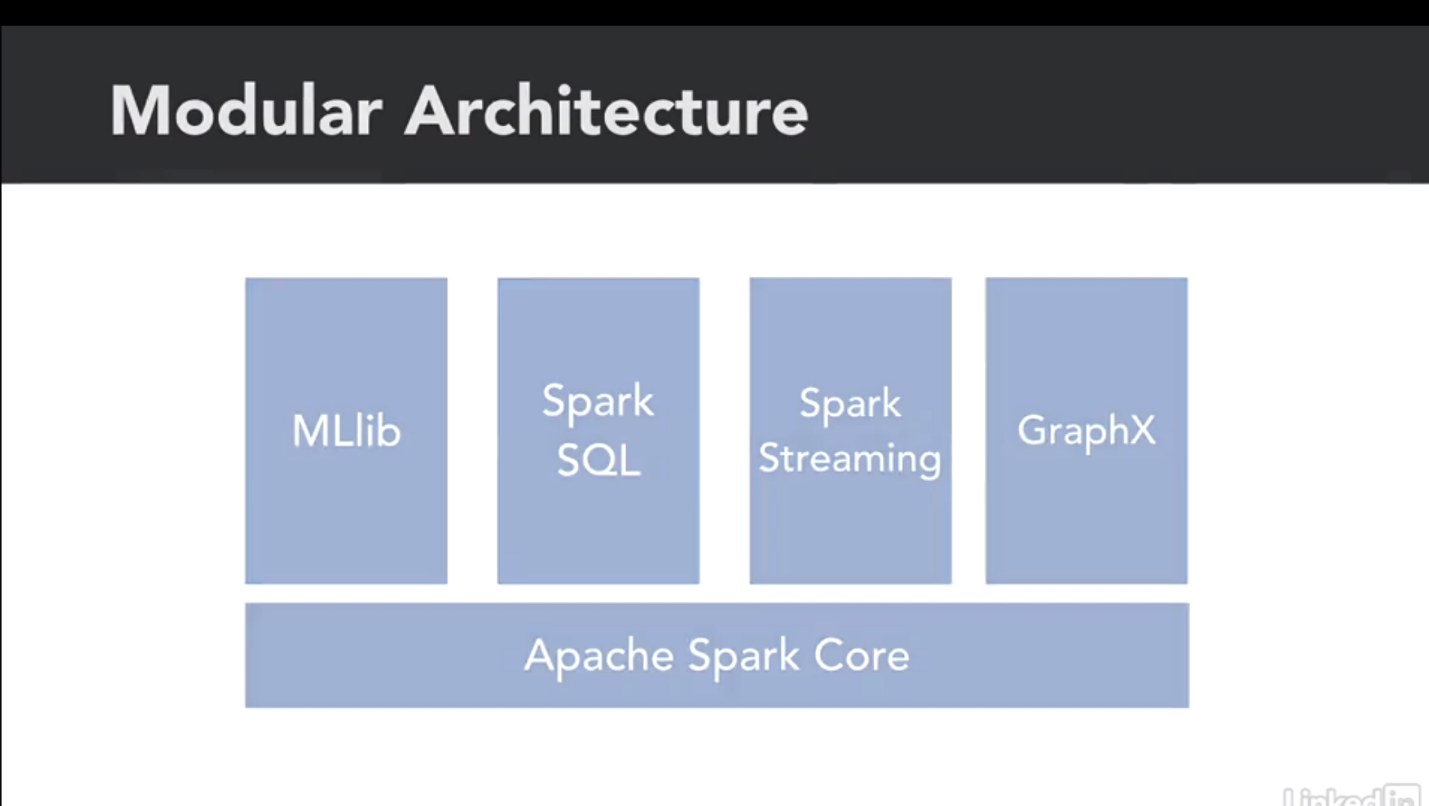
1. **Introduction to Spark and MLlib**

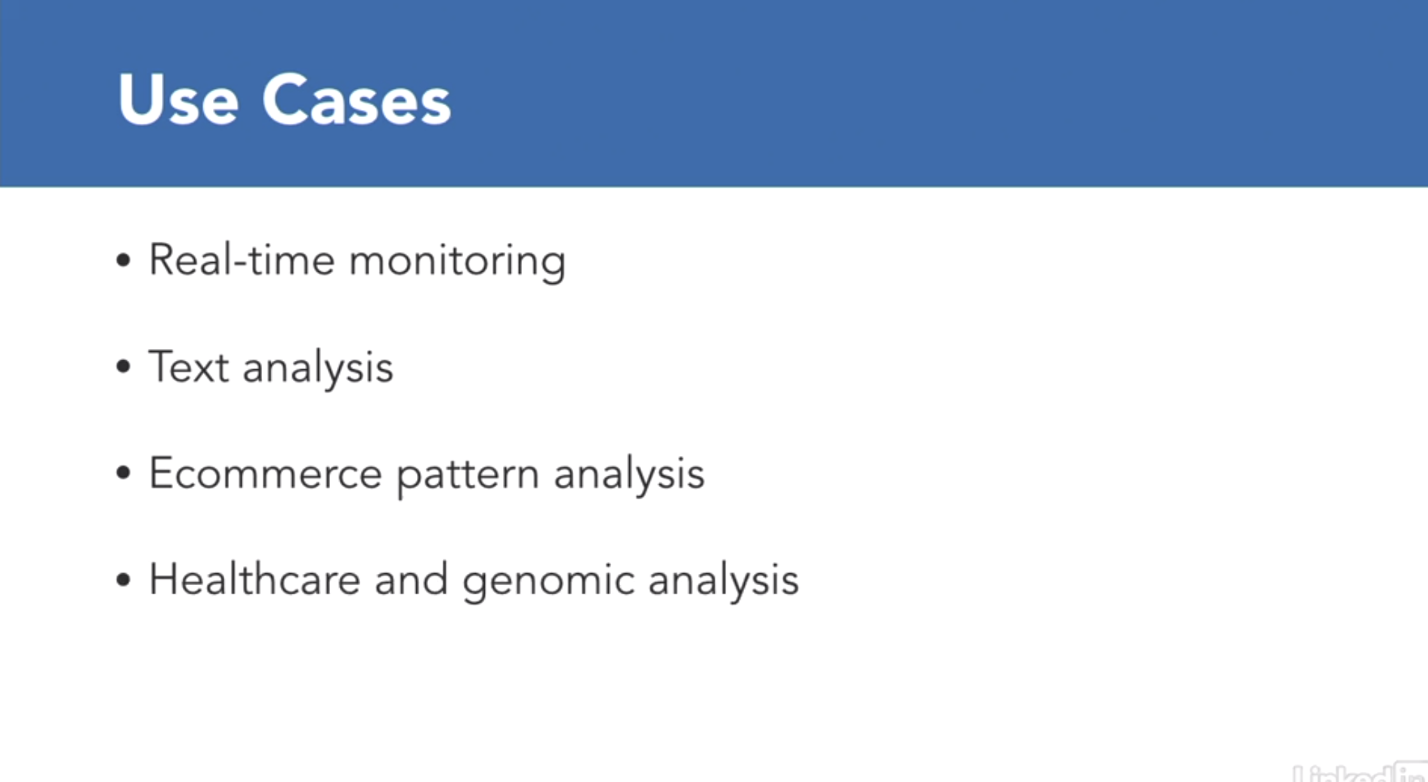


* Distributed means Spark runs on a ***cluster of servers***. Now, it runs equally well on a single server and that's what we'll use in this course. However, in a production environment, you typically run many servers to work with large data sets.
* Data processing means it ***performs computation***. And, in the case of Spark, some of the most interesting computations are related to machine learning and data analysis.
* Big data is a term broadly applied to ***data sets that are not easily analyzed on a single server or using older data management systems that were designed to run on a single server.***





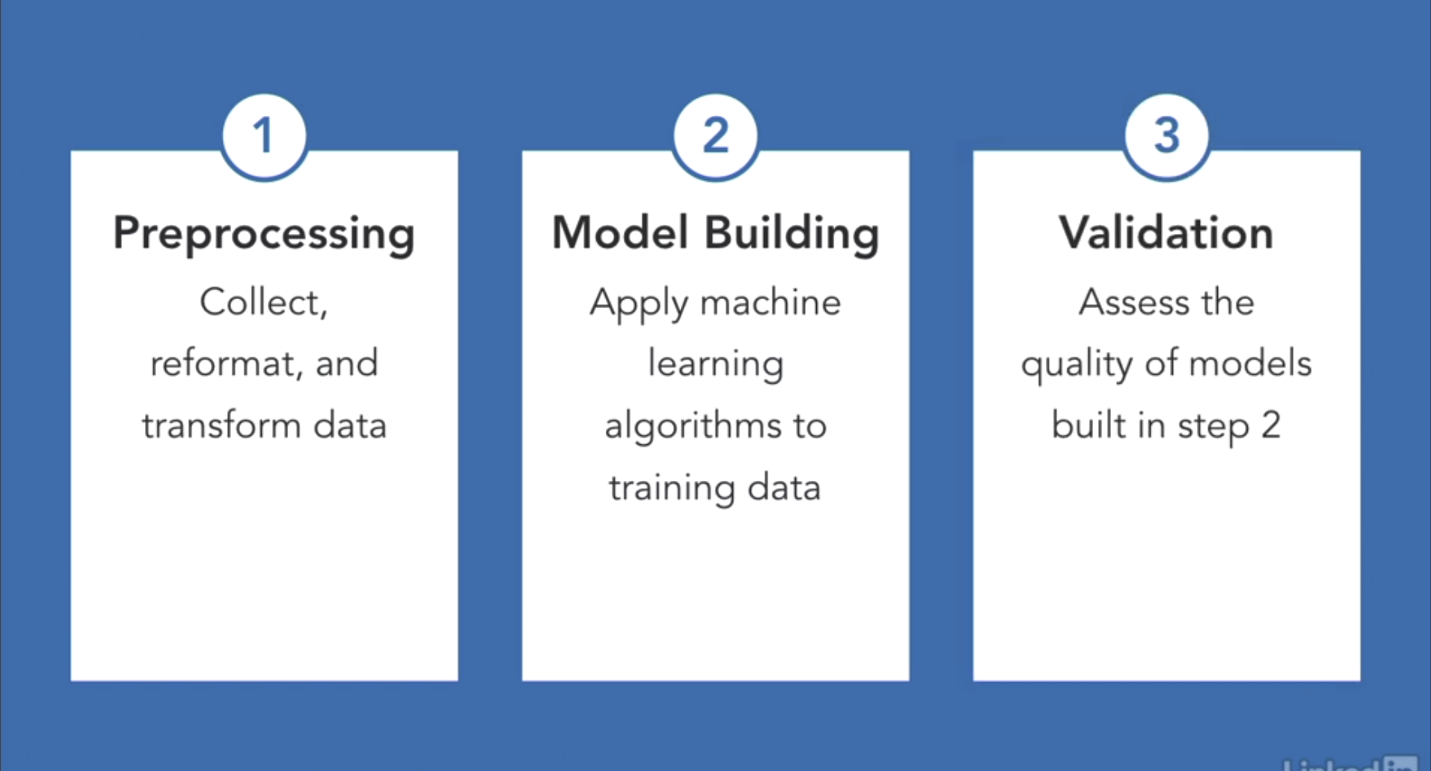


**Spark is a *generalized computation platform* designed to manage large data sets.**

It's found use in a wide number of industries and applications, including:

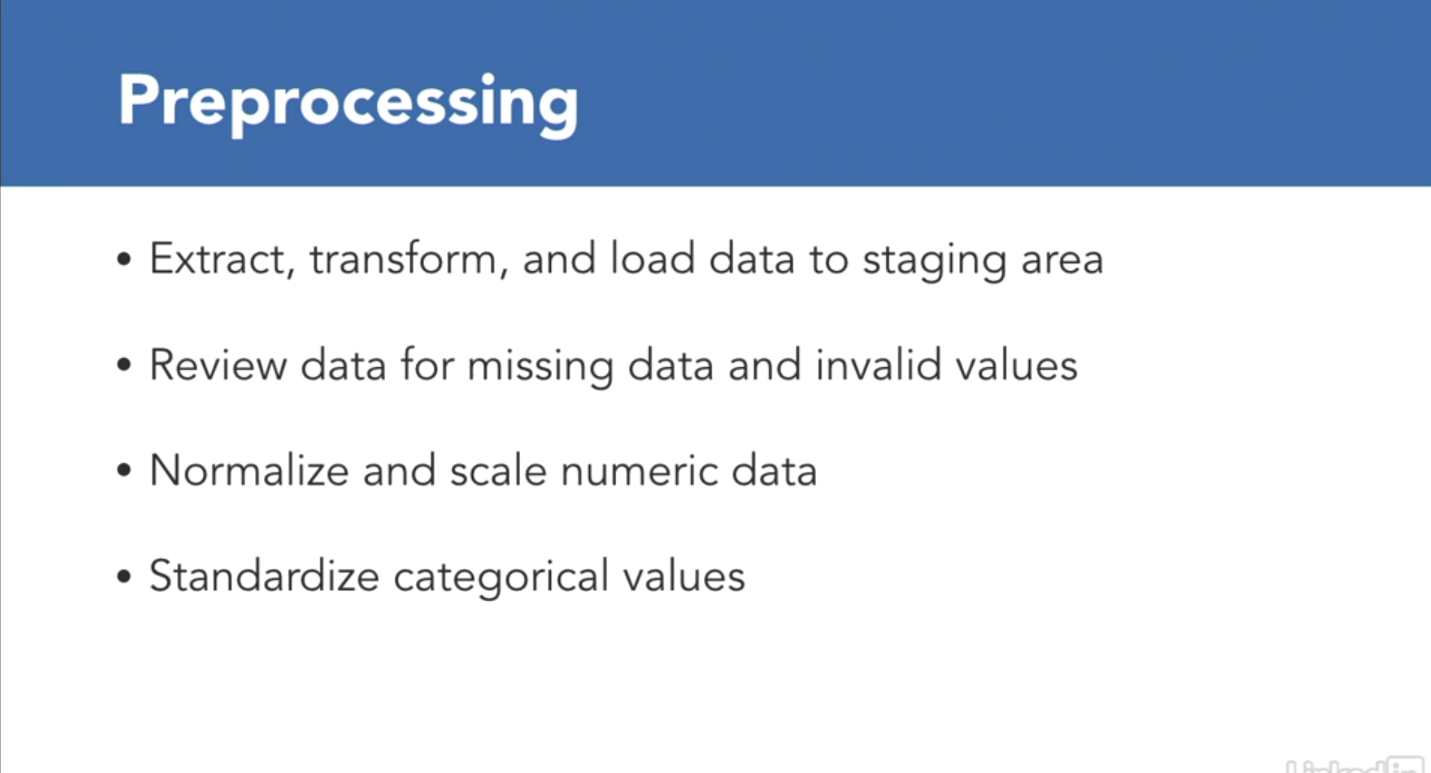
* real-time monitoring of financial data,
* text analysis related to competitive intelligence and compliance,
* analyzing how customers use eCommerce sites, and
* healthcare applications, such as analyzing genomes.

**Steps in the machine learning process:**



There are three broad steps in the machine learning process.

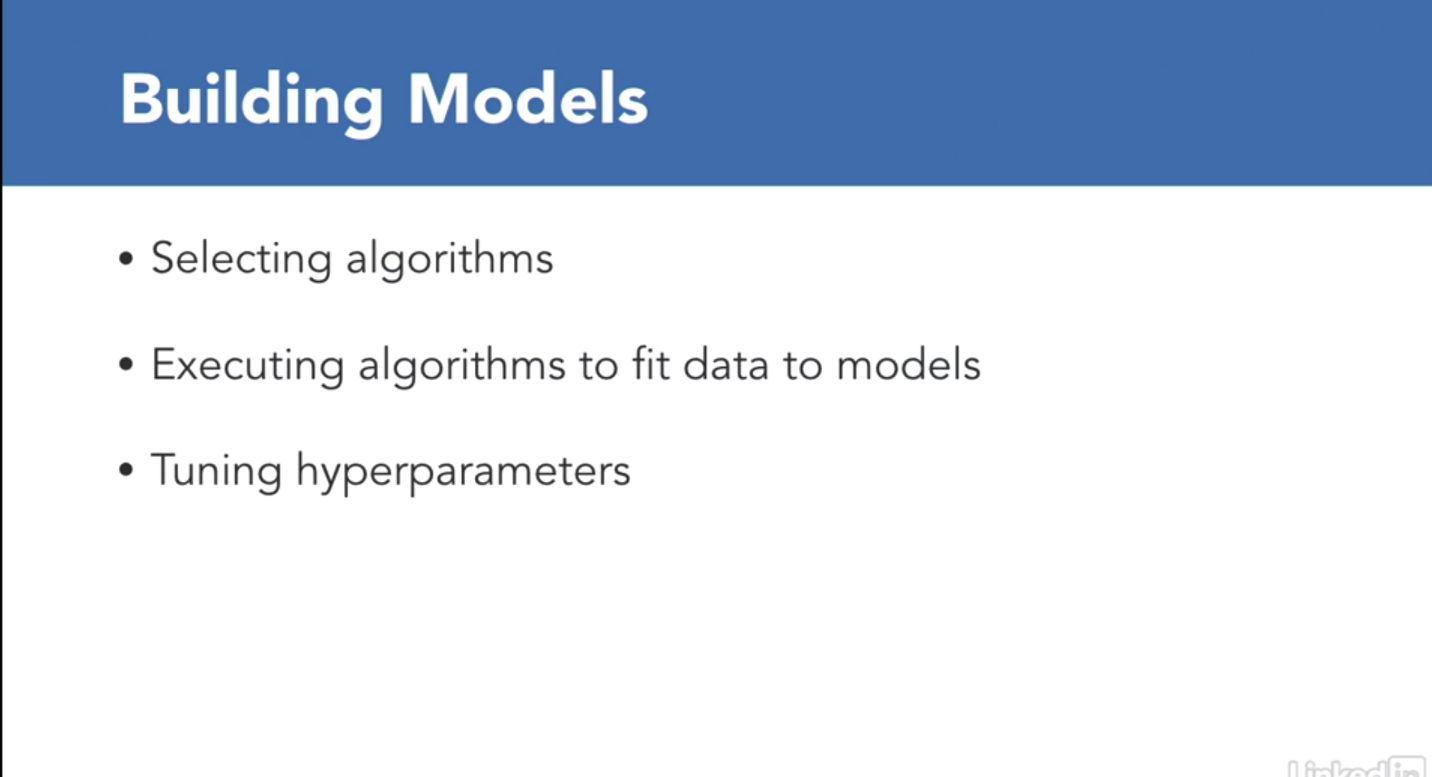
1. The first is preprocessing, which includes collecting, reformatting, and transforming data, so that it's readily used by machine learning algorithms.
2. The second step is model building, in which we apply machine learning algorithms to training data to build models. Models are pieces of code that capture the information implicit in training data.
3. The last step is validation, in which we measure how well models are performing. There are multiple ways to measure performance.



The preprocessing phase includes extracting, transforming, and loading data.

This is similar to the ETL process used in business intelligence and data warehousing.

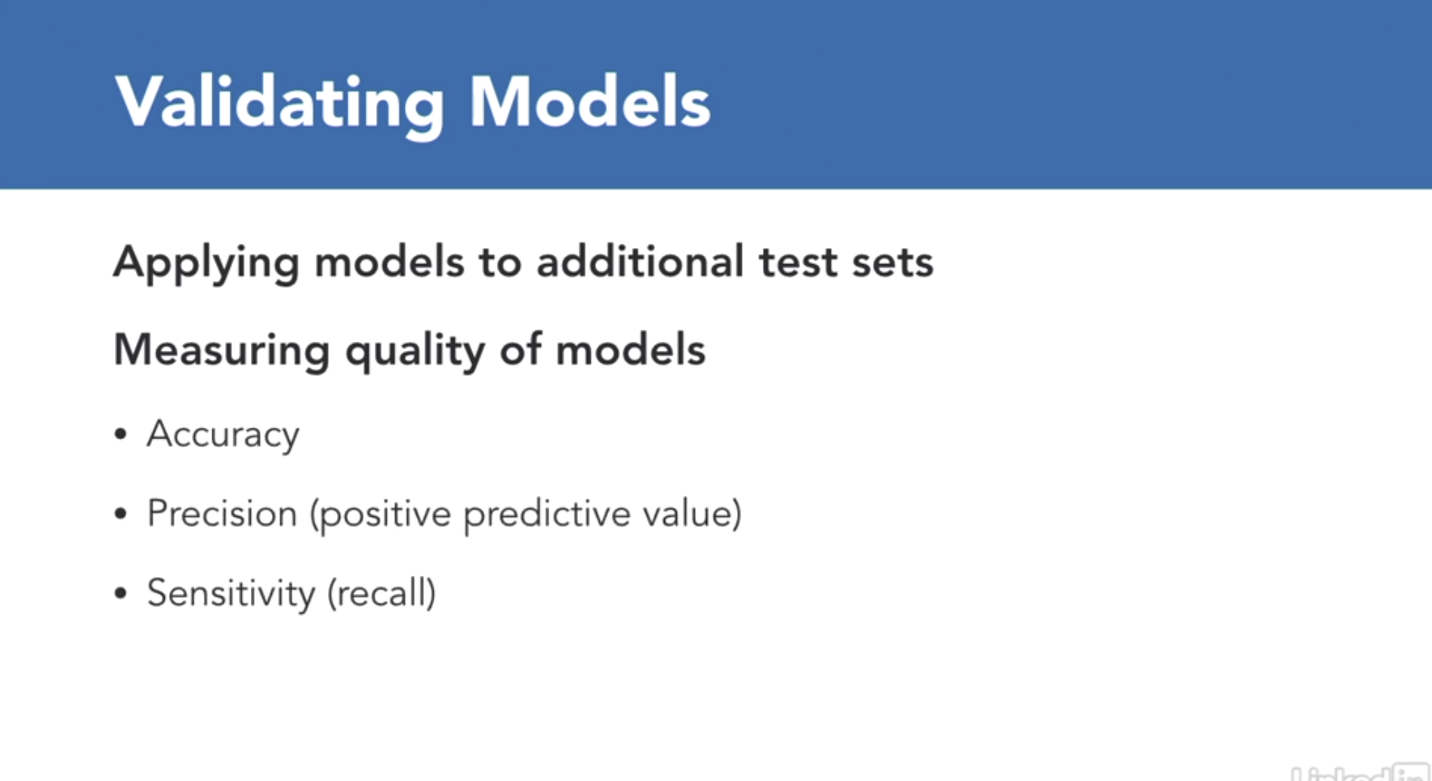
* It's a good practice to review data at this stage to determine if there are any missing or invalid values. If values are missing or invalid, you may want to set those values to some default value or ignore those records. The best way to deal with missing or invalid data will depend upon your use case.
* Normalizing and scaling data is the process of changing the range of values in a feature or attribute. We'll discuss normalizing and scaling in an upcoming video.
* Another good preprocessing practice is to standardize categorical variables. For example, you could ensure that all country names in your data set are designated by three-letter ISO codes.



In the model building stage, we may:

* experiment with different algorithms to see which works well with our data and use case.
* Applying an algorithm to a data set is called fitting a model to the data.
* Some algorithms require us to specify parameters, such as how many levels to have in a decision tree.
* These are called hyperparameters, and we often need to experiment to find optimal hyperparameters.

Now, just a note about terminology, when we set a parameter to the machine learning algorithm, we call those parameters **hyperparameters**. When the machine learning algorithm learns the value of the parameter from training data, then we call those simply **parameters.**

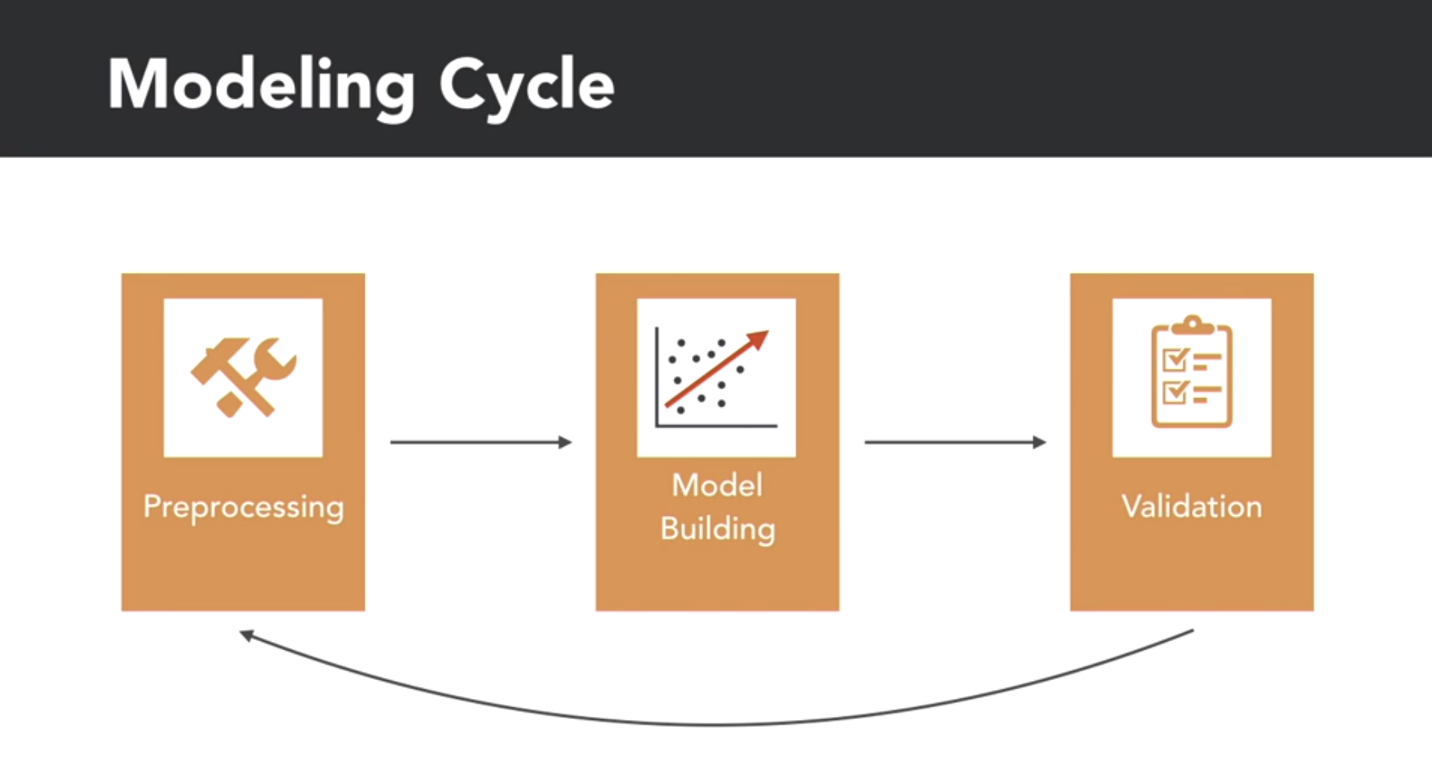


The last step in the machine learning process is validating models.

*In this step, we're trying to understand****how well our models work with data they have not seen before.***

We can use metrics like:

* accuracy,
* precision, and
* sensitivity.



**These three steps constitute the basic steps in building machine learning models.** In practice, we typically do these steps repeatedly. Each iteration often gives us new information about our data and our assumptions and help us hone our models.

1. Data Preparation and Transformation
2. Clustering
3. Classification
4. Regression
5. Recommendations
6. Conclusion