Apache Spark Machine Learning Tutorial

Contributed by

Carol McDonald (/blog/author/carol-mcdonald/)

(/blog/author/carol-mcdonald/)

17 min read

This blog post was updated February 20, 2019.

Editor's Note: Download this Free eBook: Getting Started with Apache Spark 2.x – from Inception to Production (/ebook/getting-started-with-apache-spark-v2/)

In this blog post, we will give an introduction to machine learning and deep learning, and we will go over the main Spark machine learning algorithms and techniques with some real-world use cases. The goal is to give you a better understanding of what you can do

with machine learning. Machine learning is becoming more accessible to developers, and data scientists work with domain experts, architects, developers, and data engineers, so it is important for everyone to have a better understanding of the possibilities. Every piece of information that your business generates has potential to add value. This overview is meant to provoke a review of your own data to identify new opportunities.

With Apache Spark 2.0 (/products/apache-spark/) and later versions, big improvements were implemented to make Spark easier to program and execute faster:

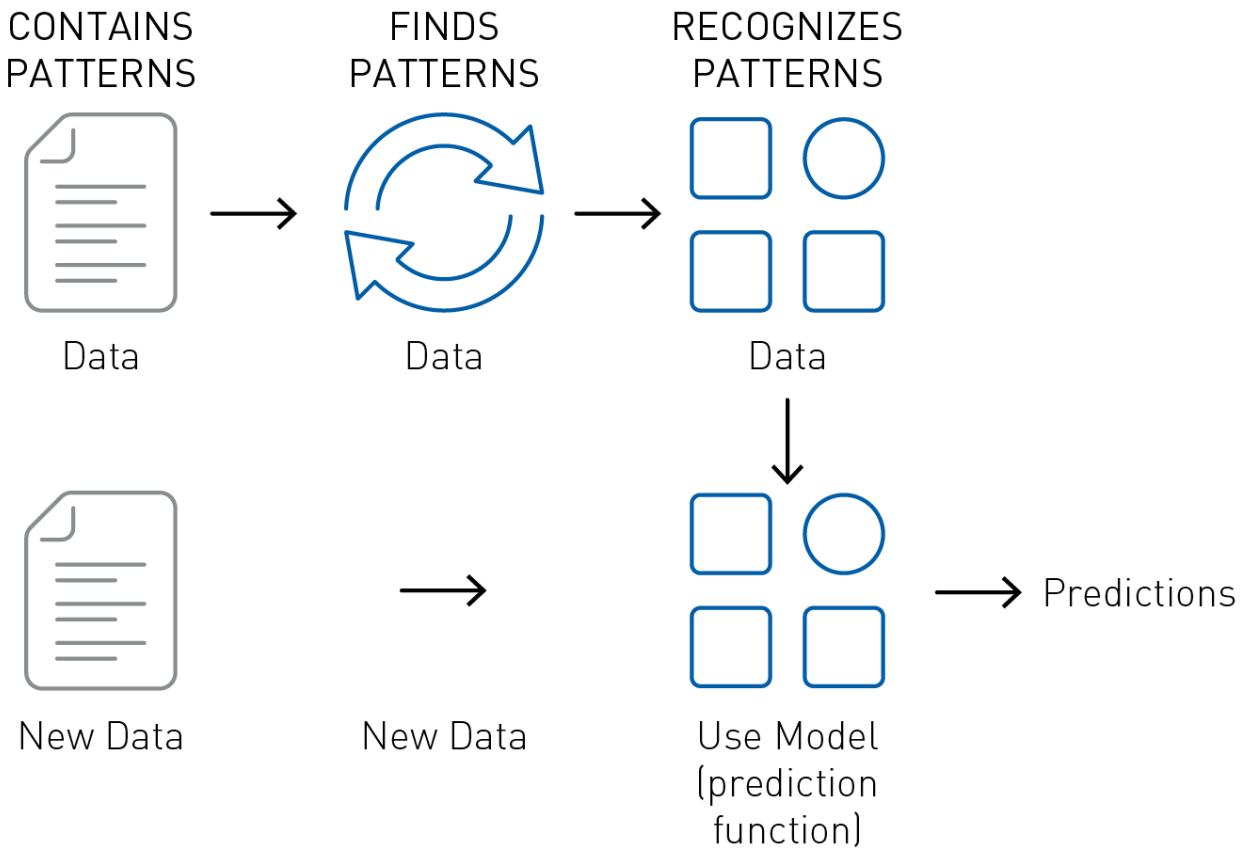
- the Spark SQL and the Dataset/DataFrame APIs provide ease of use, space efficiency, and performance gains with Spark SQL's optimized execution engine.
- Spark ML provides a uniform set of high-level APIs, built on top of DataFrames with the goal of making machine learning scalable and easy.

You can learn more about programming with Spark 2.x machine learning in the ebook Getting Started with Spark 2.x: From Inception to Production (/ebook/getting-started-with-apache-spark-v2/).

Retail	Marketing	Healthcare	Telco	Finance
Demand forecasting	Recommendation engines and targeting	Predicting patient disease risk	Customer churn	Risk analytics
Supply chain optimization	Customer 360	Diagnostics and alerts	System log analysis	Customer 360
Pricing optimization	Click-stream analysis	Fraud	Anomaly detection	Fraud
Market segmentation and targeting	Social media analysis		Preventive maintenance	Credit scoring
Recommendations	Ad optimization		Smart meter analysis	

What Is Machine Learning?

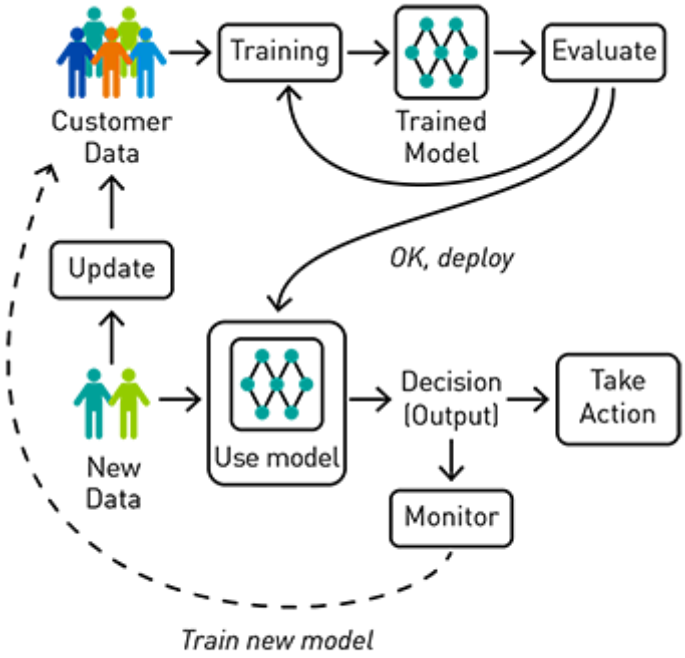
Machine learning uses algorithms to find patterns in data and then uses a model that recognizes those patterns to make predictions on new data.



There are typically two phases in machine learning:

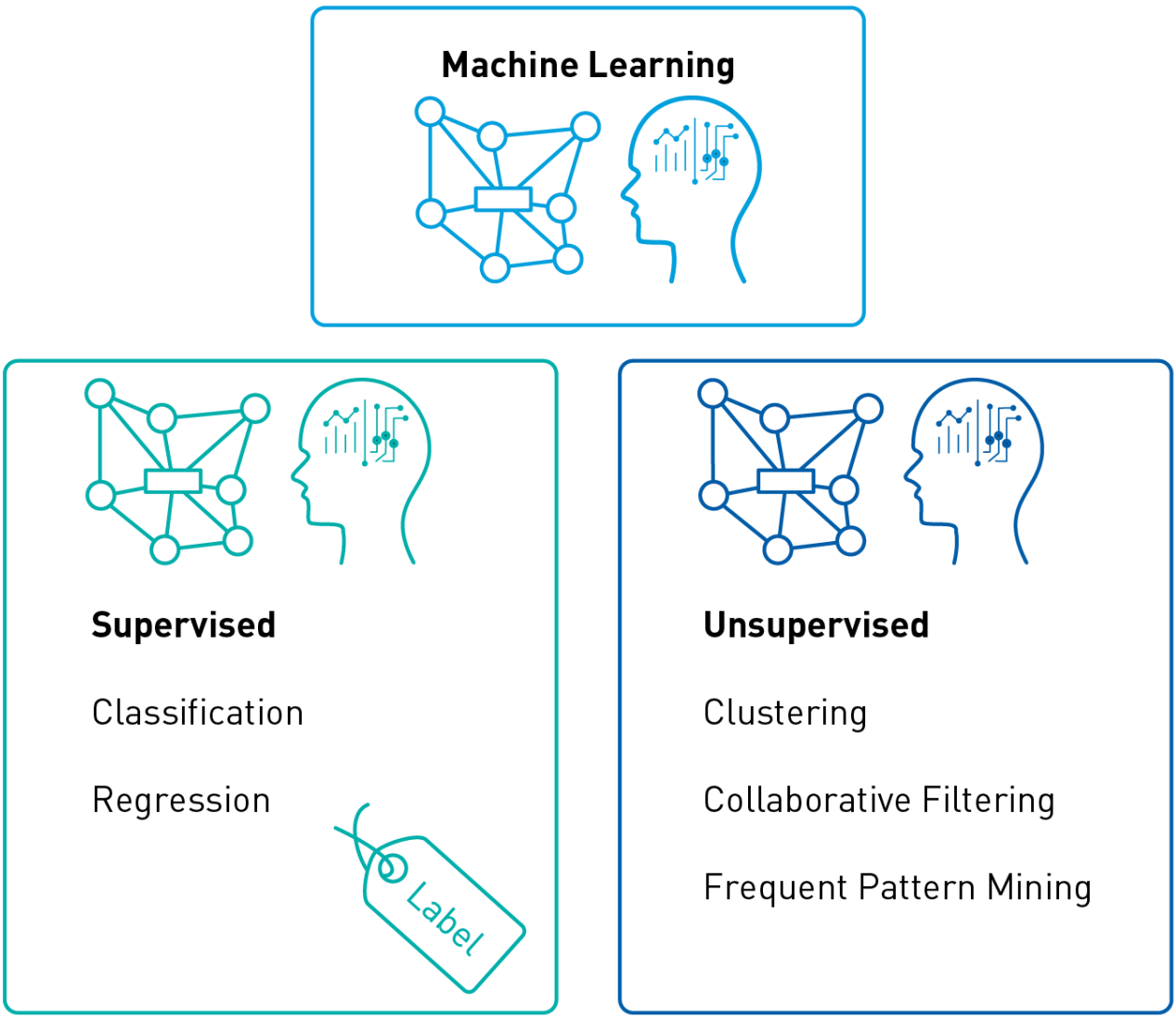
- Data Discovery: The first phase involves analysis on historical data to build and train the machine learning model.
- Analytics Using the Model: The second phase uses the model in production on new data.

In production, models need to be continuously monitored and updated with new models when needed.



(image reference https://mapr.com/ebook/ai-and-analytics-in-production/ (/ebook/ai-and-analytics-in-production/))

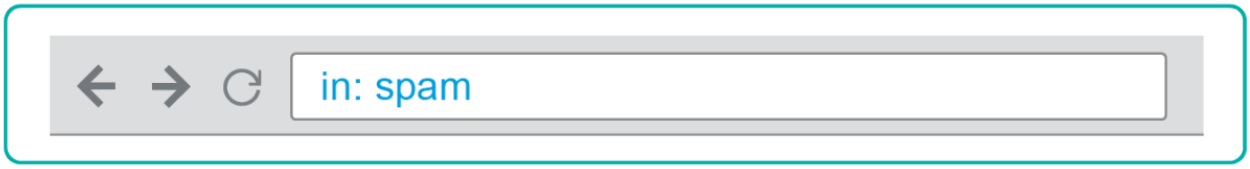
In general, machine learning may be broken down into two types: supervised, unsupervised, and in between those two. Supervised learning algorithms use labeled data; unsupervised learning algorithms find patterns in unlabeled data. Semi-supervised learning uses a mixture of labeled and unlabeled data. Reinforcement learning trains algorithms to maximize rewards based on feedback.



Three Common Categories of Techniques for Machine Learning

Three common categories of machine learning techniques are classification, clustering, and collaborative filtering.

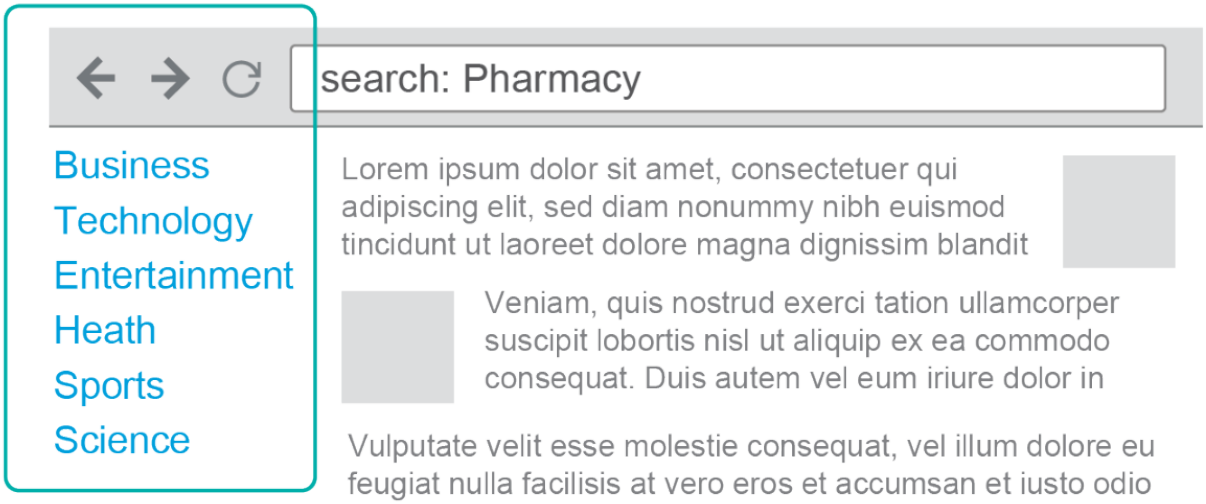
Classification



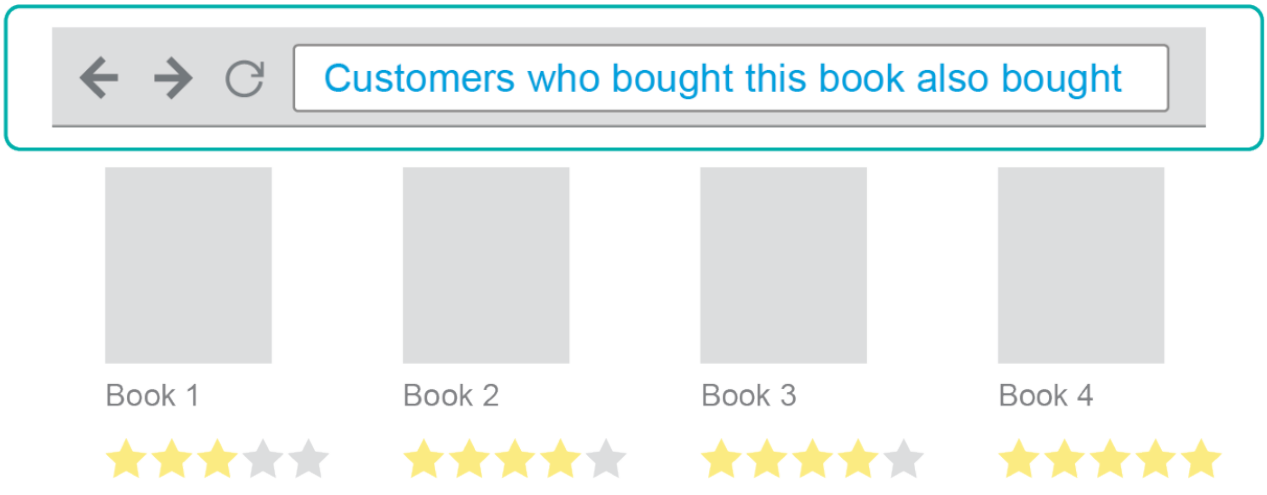
Email

☐	Mr. Norman	Accept My Donation
☐	Lending	Simple Loans
☐	election time	Please Help My Campaign
☐	Hi friend	Limited time offer
☐	confirm	Confirmation Needed Now

Clustering



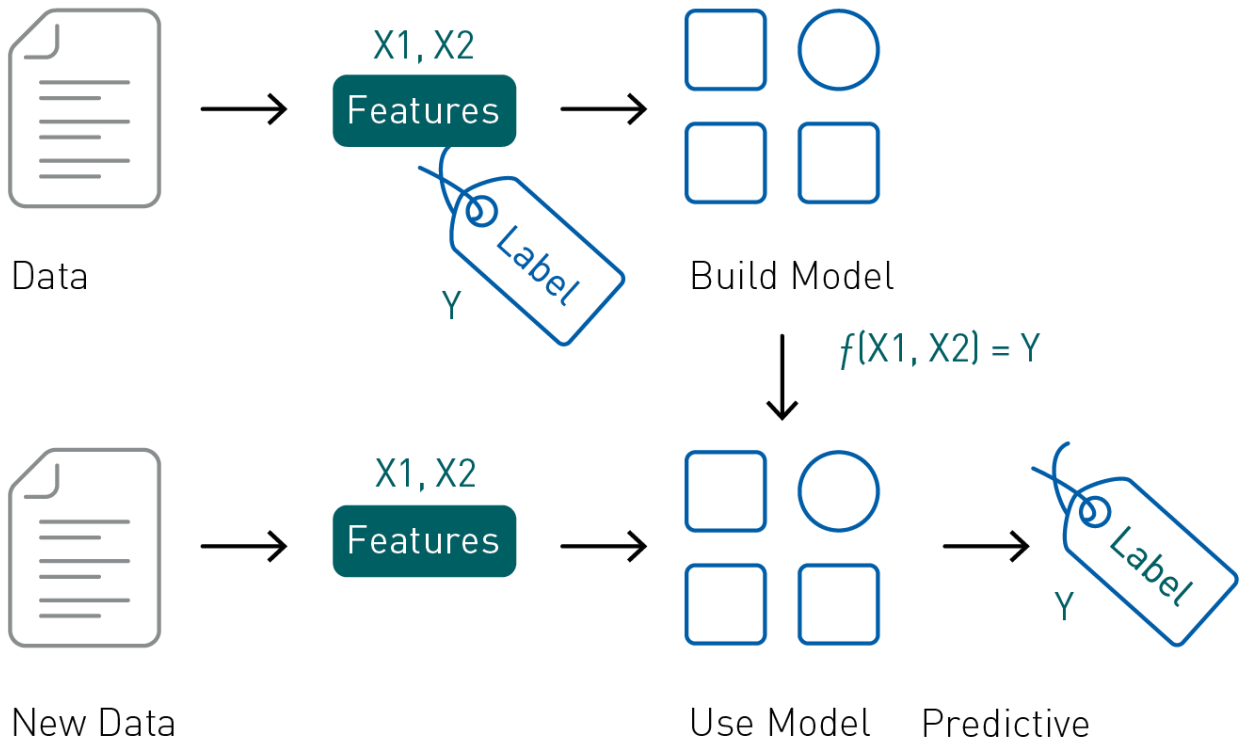
Collaborative Filtering
(Recommendation)



- Classification: Gmail uses a machine learning technique called classification to designate if an email is spam or not, based on the data of an email: the sender, recipients, subject, and message body. Classification takes a set of data with known labels and learns how to label new records based on that information.
- Clustering: Google News uses a technique called clustering to group news articles into different categories, based on title and content. Clustering algorithms discover groupings that occur in collections of data.
- Collaborative Filtering: Amazon uses a machine learning technique called collaborative filtering (commonly referred to as recommendation) to determine which products users will like, based on their history and similarity to other users.

Supervised Learning: Classification and Regression

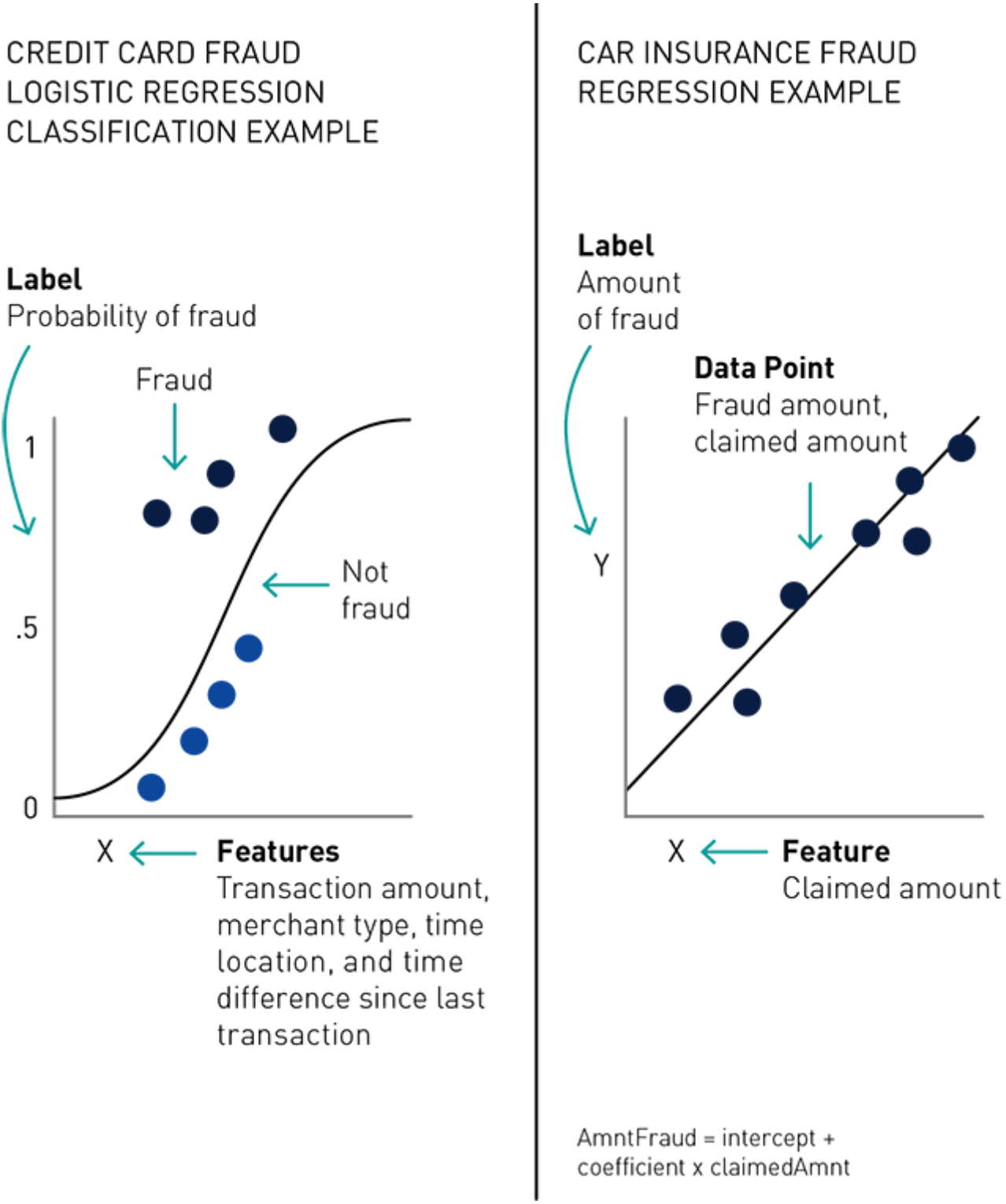
Supervised algorithms use labeled data in which both the input and target outcome, or label, are provided to the algorithm.



Supervised learning is also called predictive modeling or predictive analytics, because you build a model that is capable of making predictions.

Some examples of predictive modeling are classification and regression. Classification identifies which category an item belongs to (e.g., whether a transaction is fraud or not fraud), based on labeled examples of known items (e.g., transactions known to be fraud or

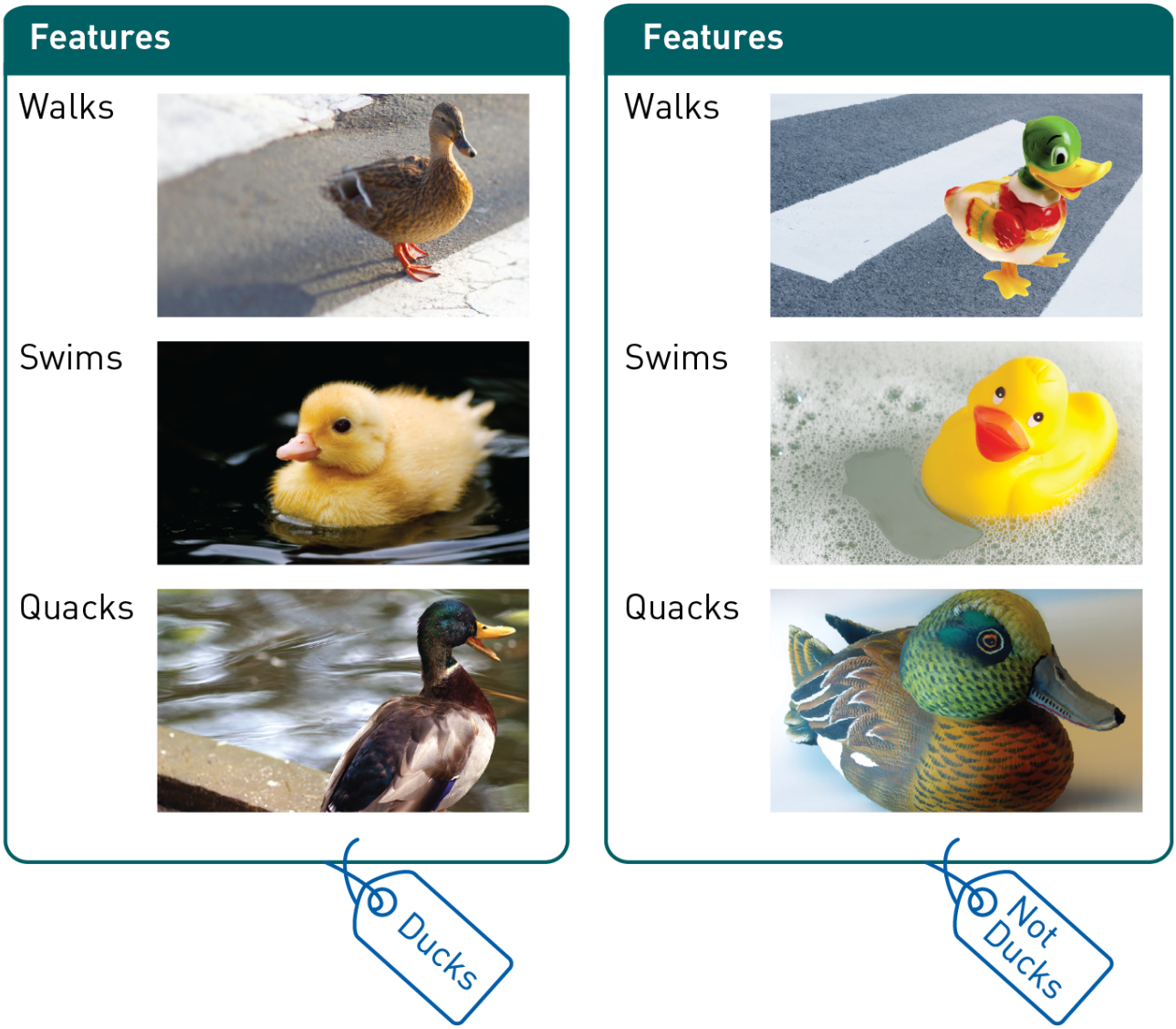
not). Logistic regression predicts a probability (e.g., the probability of fraud). Linear regression predicts a numeric value (e.g., the amount of fraud).



Classification and Regression Example

Classification and regression take a set of data with known labels and predetermined features and learns how to label new records based on that information. Features are the "if questions" that you ask. The label is the answer to those questions.

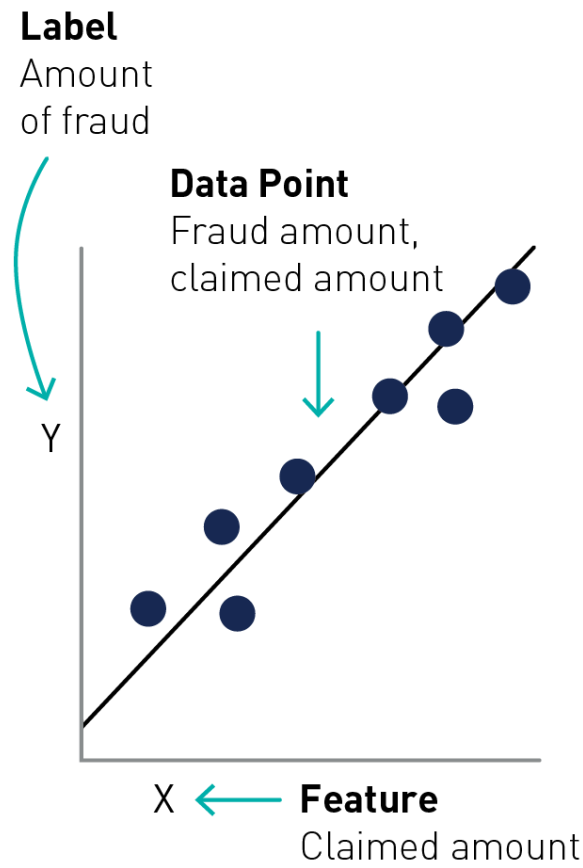
If it walks/swims/quacks like a duck ... then it must be a duck.



Regression Example

Let's go through an example of car insurance fraud:

- What are we trying to predict?
  - This is the label: the amount of fraud
- What are the "if questions" or properties that you can use to predict?
  - These are the features: to build a classifier model, you extract the features of interest that most contribute to the classification.
  - In this simple example, we will use the claimed amount.



AmntFraud = intercept + coefficient x claimedAmnt

Linear regression models the relationship between the Y "Label" and the X "Feature," in this case the relationship between the amount of fraud and the claimed amount. The coefficient measures the impact of the feature, the claimed amount, and on the label, the fraud amount.

Multiple linear regression models the relationship between two or more "Features" and a response "Label." For example, if we wanted to model the relationship between the amount of fraud and the age of the claimant, the claimed amount, and the severity of the accident, the multiple linear regression function would look like this:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$

Amount Fraud = intercept + (coefficient1 *age) + (coefficient2* claimed Amount) + (coefficient3 * severity) + error.

The coefficients measure the impact on the fraud amount of each of the features.

Some examples of linear regression include:

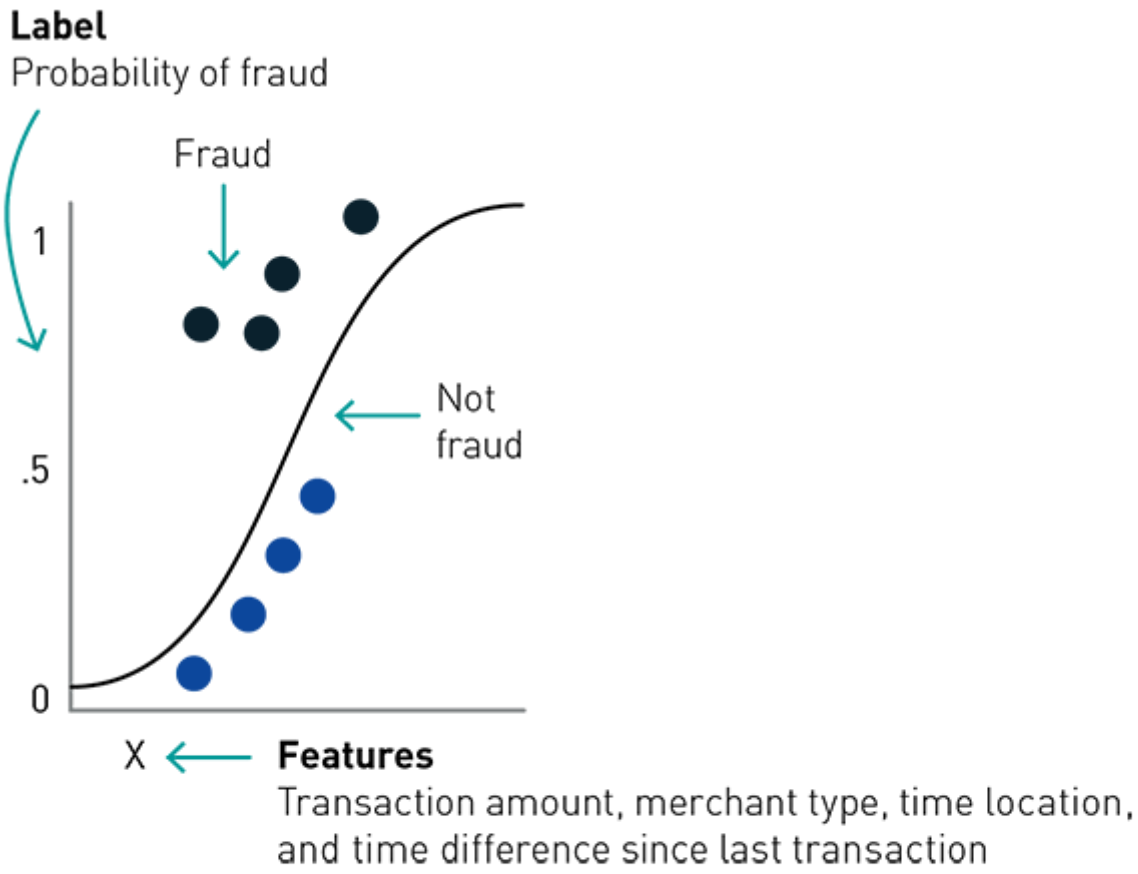
- Given historical car insurance fraudulent claims and features of the claims, such as age of the claimant, claimed amount, and severity of the accident, predict the amount of fraud.
- Given historical real estate sales prices and features of houses (square feet, number of bedrooms, location, etc.), predict a house's price.
- Given historical neighborhood crime statistics, predict crime rate.

Classification Example

Let's go through an example of debit card fraud:

- What are we trying to predict?
    - This is the label: probability of fraud
- What are the "if questions" or properties that you can use to make predictions?
    - Is the amount spent today > historical average?
    - Are there transactions in multiple countries today?
    - Are the number of transactions today > historical average?
    - Are the number of new merchant types today high compared to the last 3 months?
    - Are there multiple purchases today from merchants with a category code of risk?
    - Is there unusual signing activity today, compared to historically using pin?
    - Are there new state purchases compared to the last 3 months?
    - Are there foreign purchases today compared to the last 3 months?

To build a classifier model, you extract the features of interest that most contribute to the classification.



Logistic regression measures the relationship between the Y "Label" and the X "Features" by estimating probabilities using a logistic function (https://en.wikipedia.org/wiki/Logistic_function). The model predicts a probability, which is used to predict the label class.

Some examples of classification include:

- Given historical car insurance fraudulent claims and features of the claims, such as age of the claimant, claimed amount, and severity of the accident, predict the probability of fraud.
- Given patient characteristics, predict the probability of congestive heart failure.
- Credit card fraud detection (fraud, not fraud)
- Credit card application (good credit, bad credit)
- Email spam detection (spam, not spam)
- Text sentiment analysis (happy, not happy)
- Predicting patient risk (high risk patient, low risk patient)
- Classifying a tumor (malignant, not malignant)

Spark Supervised Algorithms Summary

Classification

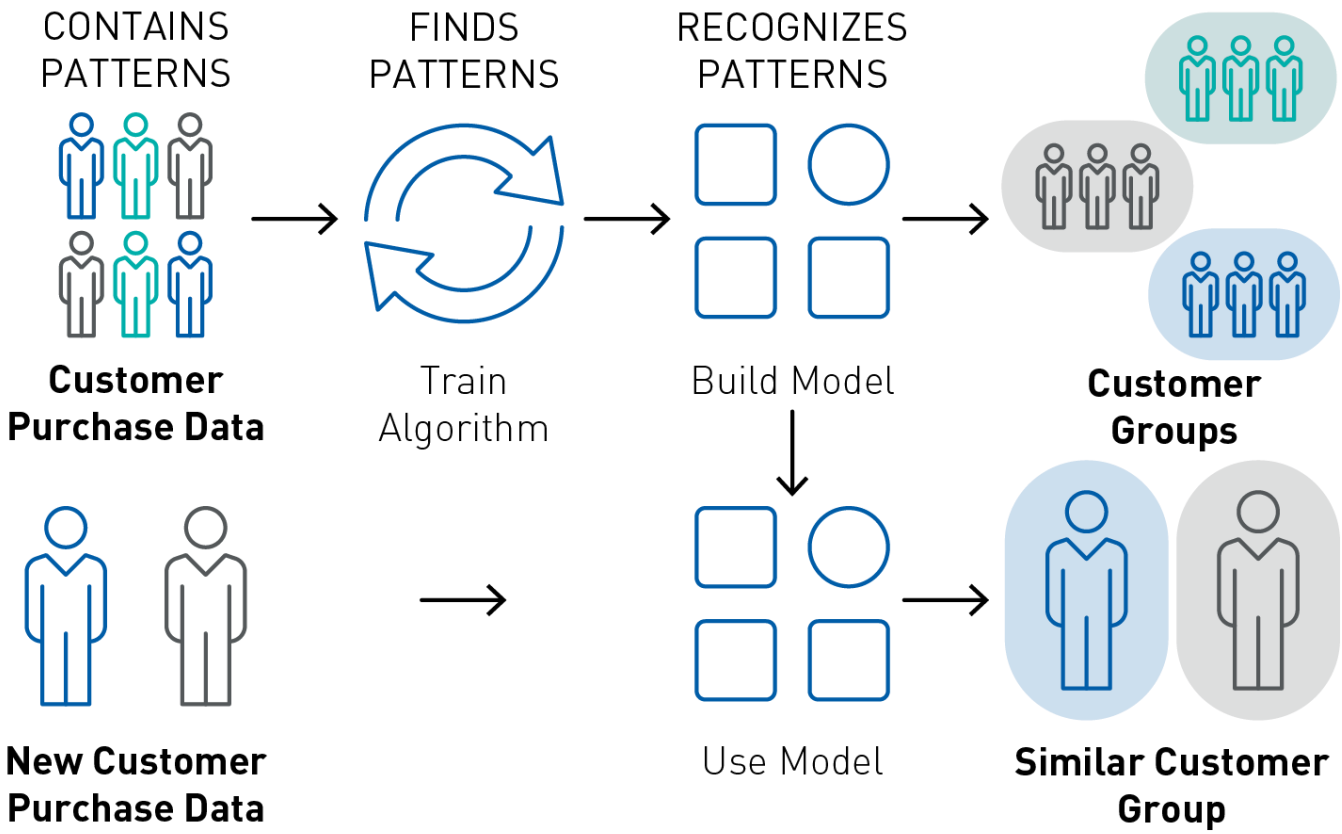
- Logistic regression
- Decision tree classifier
- Random forest classifier
- Gradient-boosted tree classifier
- Multilayer perceptron classifier
- Linear Support Vector Machine
- Naive Bayes

Regression

- Linear regression
- Generalized linear regression
- Decision tree regression
- Random forest regression
- Gradient-boosted tree regression
- Survival regression
- Isotonic regression

Unsupervised Learning

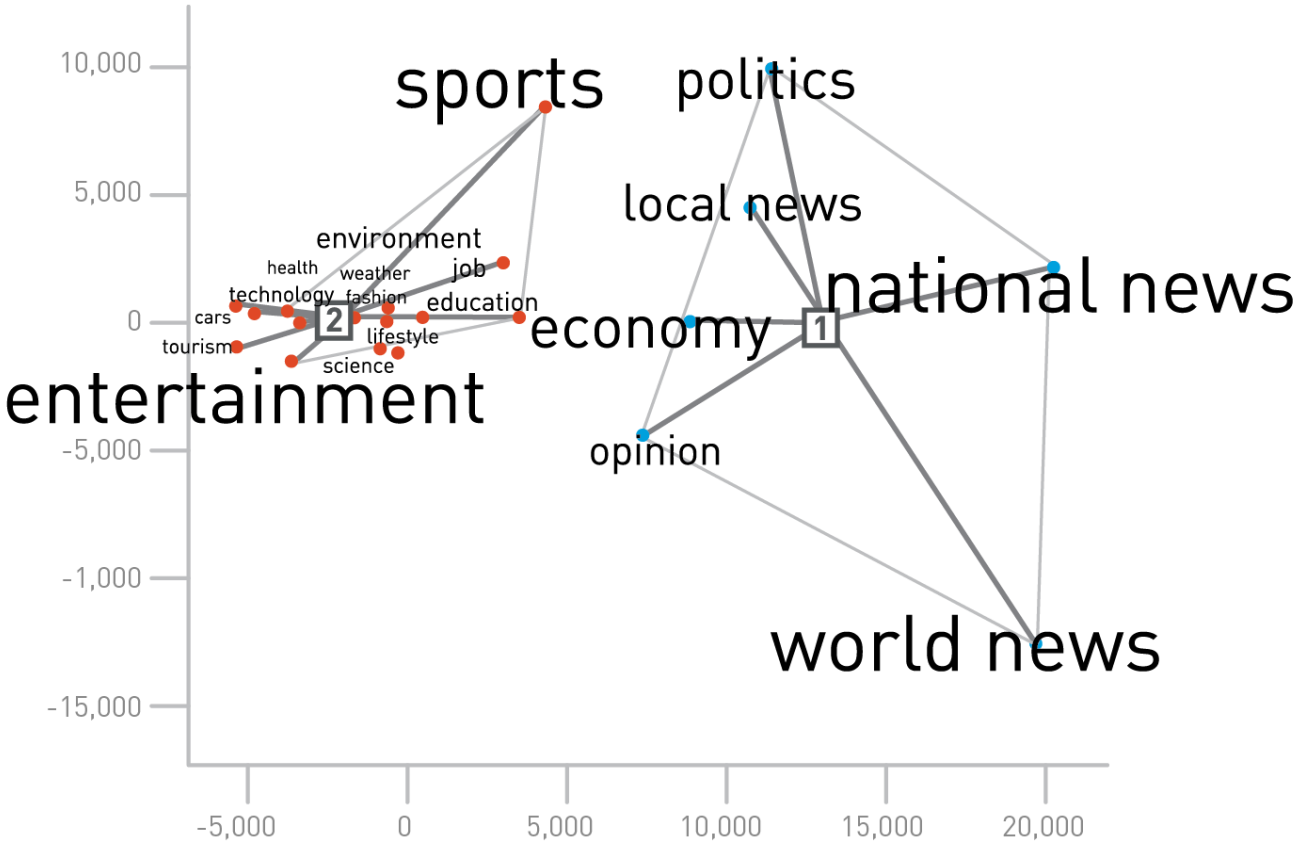
Unsupervised learning, also sometimes called descriptive analytics, does not have labeled data provided in advance. These algorithms discover similarities, or regularities, in the input data. An example of unsupervised learning is grouping similar customers, based on purchase data.



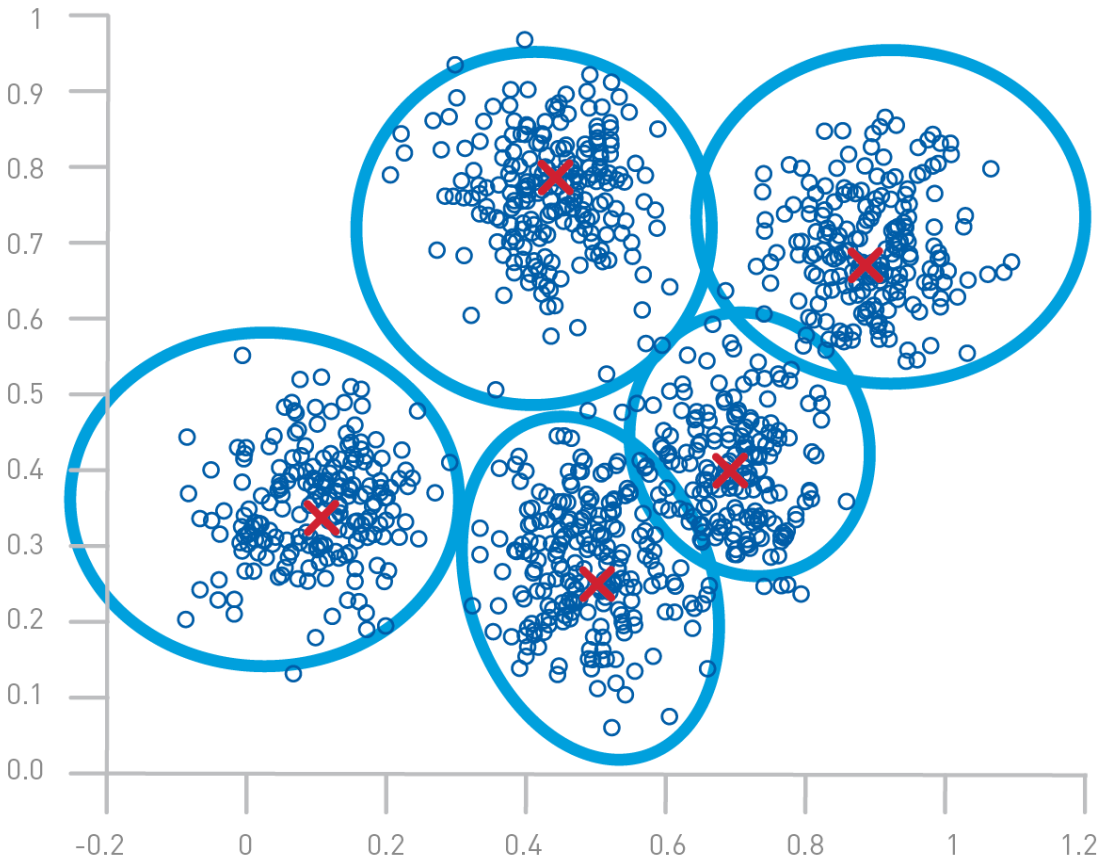
Clustering

In clustering, an algorithm classifies inputs into categories by analyzing similarities between input examples. Some clustering use cases include:

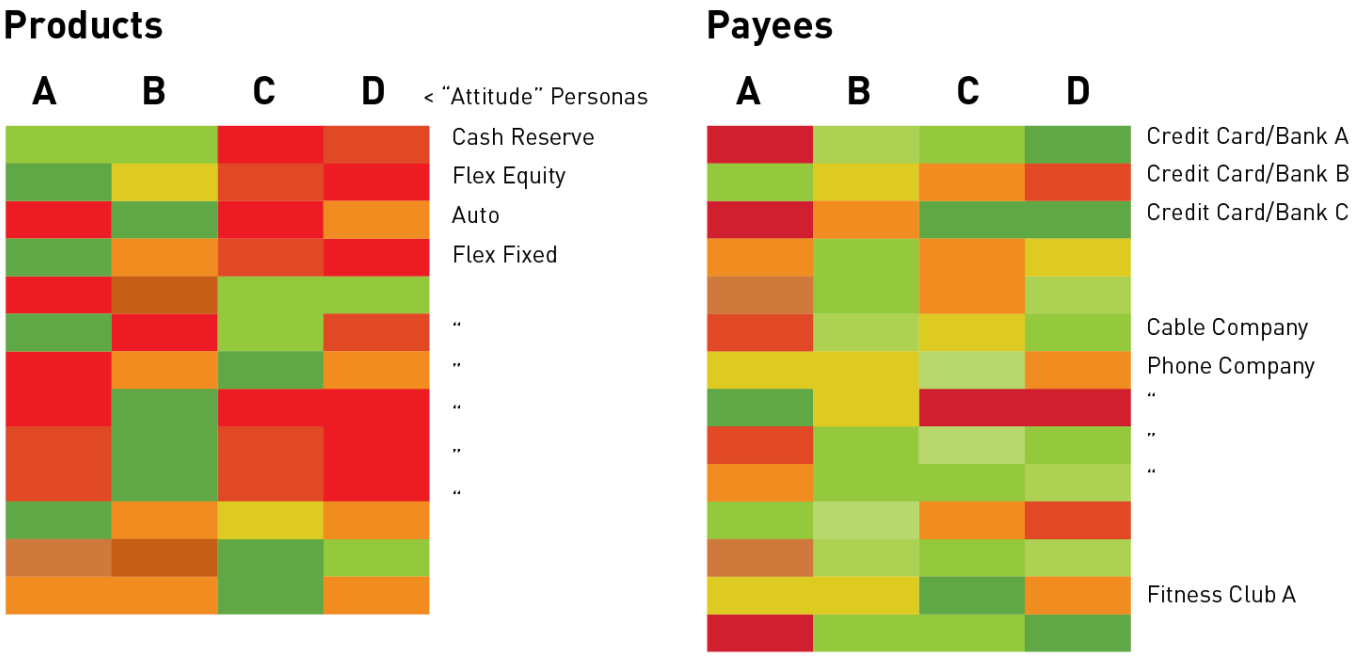
- Search results grouping
- Grouping similar customers
- Grouping similar patients
- Text categorization
- Network Security Anomaly detection (anomalies find what is **not** similar, which means the outliers from clusters)



The *k*-means algorithm groups observations into *k* clusters in which each observation belongs to the cluster with the nearest mean from its cluster center.



An example of clustering is a company that wants to segment its customers in order to better tailor products and offerings. Customers could be grouped on features such as demographics and purchase histories. Clustering with unsupervised learning is often combined with supervised learning in order to get more valuable results. For example, in this banking customer 360 (/blog/how-use-data-science-and-machine-learning-revolutionize-360-customer-views-part-2/) use case, customers were first clustered based on answers to a survey. The customer groups were analyzed and then labeled with customer personas. Next, the persona labels were linked by customer ID with customer features, such as types of accounts and purchases. Finally, supervised machine learning was applied and tested with the labeled customers, allowing it to link the survey customer personas with their banking actions and provide insights.



Frequent Pattern Mining, Association, Co-Occurrence, Market Basket Recommendations

Frequent pattern or association rule mining finds frequent co-occurring associations among a collection of items, such as products often purchased together. A famous story about association rule mining is the "beer and diaper" story. An analysis of behavior of grocery shoppers discovered that men who buy diapers often also buy beer.

Walmart mined their massive retail transaction database (https://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html) to see what their customers really wanted to buy prior to the arrival of a hurricane. They found one particular item which had an increase in sales by a factor of 7 over normal shopping days, a huge lift factor for a real-world case. The item was not bottled water, batteries, beer, flashlights, generators, or any of the usual things that you might imagine: it was strawberry pop tarts (/blog/association-rule-mining-not-your-typical-data-science-algorithm/)!

Another example is from Target, which analyzed that when a woman starts buying scent-free lotion, vitamin supplements, and a combination of some other items, it signals she could be pregnant. Unfortunately, Target sent a coupon for baby items to a teenager whose father questioned why she was receiving such coupons.

Co-occurrence analysis is useful for:

- Store layouts
- Determining which products to put on specials, promotions, coupons, etc.
- Identifying healthcare patients, like mine cohorts

Collaborative Filtering

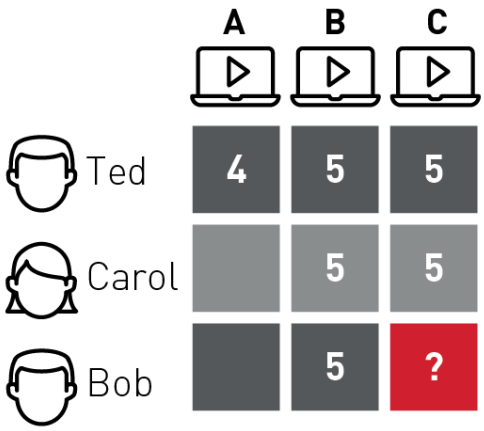
Collaborative filtering algorithms recommend items (this is the filtering part) based on preference information from many users (this is the collaborative part). The collaborative filtering approach is based on similarity; people who liked similar items in the past will like similar items in the future. The goal of a collaborative filtering algorithm is to take preferences data from users and create a model that can be used for recommendations or predictions. Ted likes movies A, B, and C. Carol likes movies B and C. We take this data and run it through an algorithm to build a model. Then, when we have new data, such as Bob likes movie B, we use the model to predict that C is a possible recommendation for Bob.

Ted and Carol like movies B and C.

Bob likes movie B; what else might he like?

Bob likes movie B, so **predict movie C**.

User item rating matrix



Spark Unsupervised Algorithms Summary

Clustering

- *k*-means
- Latent Dirichlet allocation (LDA)
- Gaussian mixture model (GMM)

Collaborative Filtering

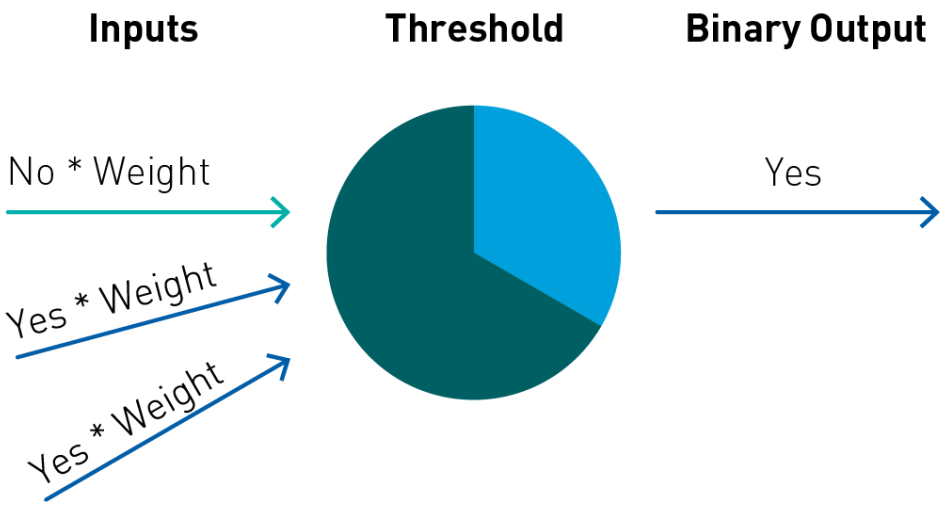
- Alternating least squares (ALS)

Frequent Pattern Mining

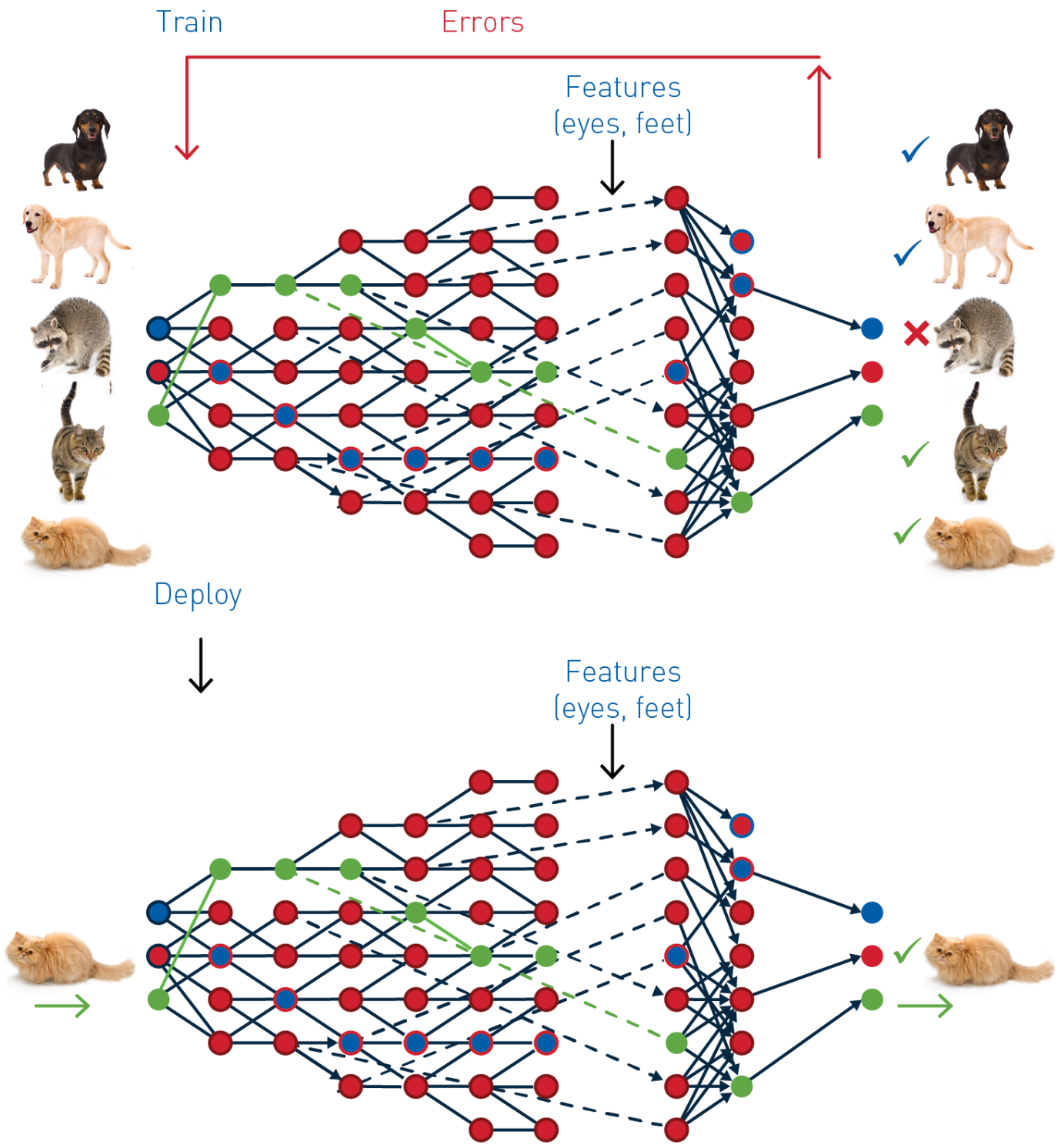
- FP-Growth Algorithm

Deep Learning

Deep learning is the name for multilayered neural networks, which are networks composed of several "hidden layers" of nodes between the input and output. There are many variations of neural networks, which you can learn more about on this neural network cheat sheet. (http://www.asimovinstitute.org/neural-network-zoo/) Improved algorithms, GPUs, and massively parallel processing (MPP) have given rise to networks with thousands of layers. Each node takes input data and a weight and outputs a confidence score to the nodes in the next layer, until the output layer is reached, where the error of the score is calculated.



With backpropagation (https://en.wikipedia.org/wiki/Backpropagation) inside of a process called gradient descent (https://en.wikipedia.org/wiki/Gradient_descent), the errors are sent back through the network again and the weights are adjusted, improving the model. This process is repeated thousands of times, adjusting a model's weights in response to the error it produces, until the error can't be reduced anymore.



During this process, the layers learn the optimal features for the model, which has the advantage that features do not need to be predetermined. However, this has the disadvantage that the model's decisions are not explainable. Because explaining the decisions can be important, researchers are developing new ways to understand the black box of deep learning (http://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning).

There are different variations of deep learning algorithms, which can be used with the Distributed Deep Learning Quick Start Solution from MapR (/solutions/quickstart/deep-learning-quick-start/) to build data-driven applications, such as the following:

- Deep Neural Networks for improved traditional algorithms
    - Finance: enhanced fraud detection through identification of more complex patterns
    - Manufacturing: enhanced identification of defects, based on deeper anomaly detection
- Convolutional Neural Networks for images
    - Retail: in-store activity analysis of video to measure traffic
    - Satellite images: labeling terrain, classifying objects
    - Automotive: recognition of roadways and obstacles
    - Healthcare: diagnostic opportunities from x-rays, scans, etc.
    - Insurance: estimating claim severity, based on photographs
- Recurrent Neural Networks for sequenced data
    - Customer satisfaction: transcription of voice data to text for NLP analysis
    - Social media: real-time translation of social and product forum posts
    - Photo captioning: search archives of images for new insights
    - Finance: Predicting behavior based via time series analysis (also enhanced recommendation systems)

Deep Learning with Spark

Deep learning libraries or frameworks that can be leveraged with Spark include:

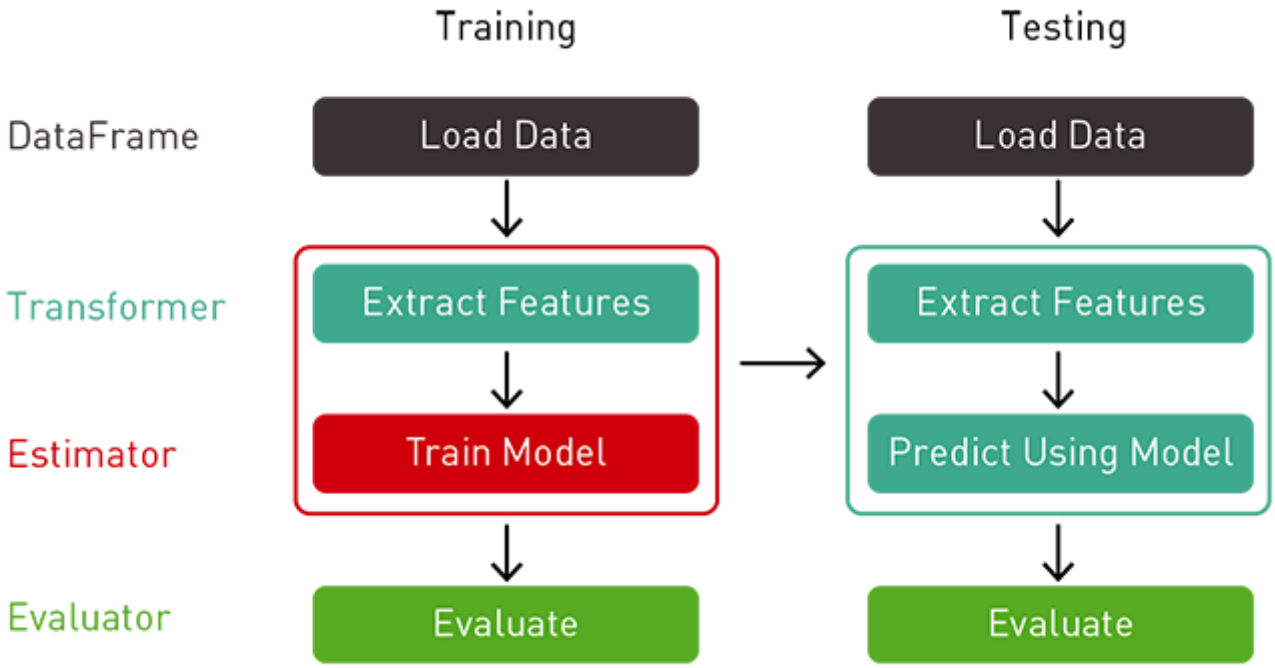
- BigDL
- Spark Deep Learning Pipelines
- TensorFlowOnSpark
- dist-keras
- H2O Sparkling Water
- PyTorch
- Caffe
- MXNet

USING SPARK ML

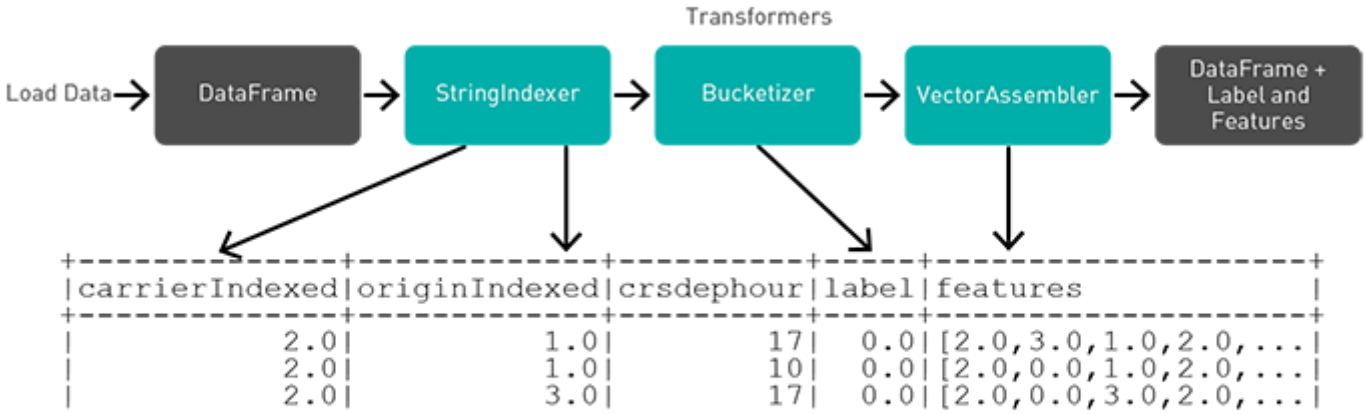
With Apache Spark 2.0 and later versions, big improvements were implemented to make Spark easier to program and execute faster:

- the Spark SQL and the Dataset/DataFrame APIs provide ease of use, space efficiency, and performance gains.

- Spark ML provides a uniform set of high-level APIs, built on top of DataFrames with the goal of making machine learning scalable and easy. Having ML APIs built on top of DataFrames provides the scalability of partitioned data processing with the ease of SQL for data manipulation.



You can use a Spark ML Pipeline to pass your data through transformers in order to extract the features, an estimator to produce a model, and an evaluator to evaluate the model.



In chapters 5-9 of the MapR eBook Getting Started with Spark 2.x: From Inception to Production (/ebook/getting-started-with-apache-spark-v2/), you can explore some Spark machine learning examples and download the corresponding code.

Summary

A confluence of several different technology shifts have dramatically changed machine learning applications. The combination of distributed computing, streaming analytics, and machine learning is accelerating the development of next-generation intelligent applications (/blog/what-is-next-gen-app/), which are taking advantage of modern computational paradigms, powered by modern computational infrastructure. The MapR Data Platform (/products/) integrates global event streaming (/products/mapr-eventstore/), real-time database capabilities (/products/mapr-database/), and scalable enterprise storage (/products/mapr-xd/) with a collection of data processing and analytical engines to power this new generation of data processing pipelines and intelligent applications.

References and More Information

MapR ebook: Getting Started with Apache Spark 2.x (/ebook/getting-started-with-apache-spark-v2/)

MapR blog: Apache Spark (/blog/apache-spark/)

MapR blog: Machine Learning (/blog/machine-learning/)

Apache Spark on MapR (/products/apache-spark/)

Spark on Demand Training (https://learn.mapr.com/series/sparkv2)

MapR ebook: Buyer's Guide to AI and Machine Learning (/ebook/buyers-guide-to-ai-and-machine-learning/)

MapR ebook: AI and Analytics in Production (/ebook/ai-and-analytics-in-production/)

All MapR ebooks (/ebooks/)

This blog post was published February 22, 2016.

This blog post was updated February 20, 2019.

Categories

All (/blog/)

Apache Drill (/blog/apache-drill/)

Apache Hadoop (/blog/apache-hadoop/)

Apache Hive (/blog/apache-hive/)

Apache Mesos (/blog/apache-mesos/)

Apache Myriad (/blog/apache-myriad/)

Apache Spark (/blog/apache-spark/)

Cloud Computing (/blog/cloud-computing/)

Enterprise Data Hub (/blog/enterprise-data-hub/)

Machine Learning (/blog/machine-learning/)

MapR Platform (/blog/mapr-platform/)

MapReduce (/blog/mapreduce/)

NoSQL (/blog/nosql/)

Open Source Software (/blog/open-source-software/)

Partners (/blog/partners/)

Streaming (/blog/streaming/)

Use Cases (/blog/use-cases/)

Whiteboard Walkthrough Videos (/blog/whiteboard-walkthrough-videos/)

Email Us (/company/contact-mapr/#contact-us)

+1 855-NOW-MAPR (tel:8556696277)

Download MapR for Free (/try-mapr/)

Request a Demo (/demo/)

Why MapR?
(/why-mapr/)

Customers (/customers/)

Solutions (/solutions/)

Products (/products/)

Services (/services/)

Training (/training/)

Company

(/company/)

Press (/company/press-releases/) | News (/company/news/)

Leadership (/company/leadership/)

Investors (/company/investors/)

Resellers (/resellers/)

Partners (/partners/)

Careers (/careers/)

Awards (/company/awards/)

Contact Us
(/company/contact-mapr/)

Contact Sales

(mailto:sales@mapr.com)

United States: +1 408-914-2390

(tel:4089142390)

Outside the US: +1 855-NOW-MAPR

(tel:8556696277)

Legal
(/legal/)


(https://www.linkedin.com/company/mapr-technologies)


(https://www.facebook.com/maprtech/)


(https://twitter.com/mapr)


(https://www.youtube.com/user/maprtech)


(/company/contact-mapr/)