Advanced Analytics

Apache Spark

Advanced Analytics

- Advanced analytics refers to a variety of techniques aimed at solving the core problem of deriving insights and making predictions or recommendations based on data.
- Supervised learning, including classification and regression, where the goal is to predict a label for each data point based on various features.
- Recommendation engines to suggest products to users based on behavior.
- Unsupervised learning, including clustering, anomaly detection, and topic modeling, where the goal is to discover structure in the data.
- Graph analytics tasks such as searching for patterns in a social network.

Supervised Learning:

- Supervised learning is probably the most common type of machine learning.
- The goal is simple: Using historical data that already has labels (often called the dependent variables), train a model to predict the values of those labels based on various features of the data points.
- One example would be to predict a person's income (the dependent variable) based on age (a feature).
- This training process usually proceeds through an iterative optimization algorithm such as gradient descent. The training algorithm starts with a basic model and gradually improves it by adjusting

Supervised Learning:

- Various internal parameters (coefficients) during each training iteration. The result of this process is a trained model that you can use to make predictions on new data.

- The training algorithm starts with a basic model and gradually improves it by adjusting various internal parameters (coefficients) during each training iteration.

- The result of this process is a trained model that you can use to make predictions on new data.

Classification

- One common type of supervised learning is classification. Classification is the act of training an algorithm to predict a dependent variable that is categorical (belonging to a discrete, finite set of values).

- The most common case is *binary classification*, where our resulting model will make a prediction that a given item belongs to one of two groups.

- The canonical example is classifying email spam.

Classification

- Using a set of historical emails that are organized into groups of spam emails and not spam emails, we train an algorithm to analyze the words in, and any number of properties of, the historical emails and make predictions about them.
- Once we are satisfied with the algorithm's performance, we use that model to make predictions about future emails the model has never seen before.
- When we classify items into more than just two categories, we call this multiclass classification.
- For example, we may have four different categories of email (as opposed to the two categories in the previous paragraph): spam, personal, work-related, and other.

Classification

- Predicting disease
- Classifying images
- Predicting customer churn
- Buy or won’t buy

Regression

- In classification, our dependent variable is a set of discrete values. In regression, we instead try to predict a continuous variable (a real number). In simplest terms, rather than predicting a category, we want to predict a value on a number line.

- Predicting sales: A store may want to predict total product sales on given data using historical sales data. There are a number of potential input variables, but a simple example might be using last week's sales data to predict the next day's data.

Recommendation

- Recommendation is one of the most intuitive applications of advanced analytics.
- By studying people's explicit preferences (through ratings) or implicit ones (through observed behavior) for various products or items, an algorithm can make recommendations on what a user may like by drawing similarities between the users or items.
- By looking at these similarities, the algorithm makes recommendations to users based on what similar users liked, or what other products resemble the ones the user already purchased. The recommendation is a common use case for Spark and well suited to big data.

Recommendation

- **Movie recommendations:**
  - Netflix uses Spark, although not necessarily its built-in libraries, to make large-scale movie recommendations to its users. It does this by studying what movies users watch and do not watch in the Netflix application. In addition, Netflix likely takes into consideration how similar a given user's ratings are to other users.
- **Product recommendations :** Amazon uses product recommendations as one of its main tools to increase sales.
- For instance, based on the items in our shopping cart, Amazon may recommend other items that were added to similar shopping carts in the past. Likewise, on every product page, Amazon shows similar products purchased by other users.

Unsupervised Learning

- *Unsupervised learning* is the act of trying to find patterns or discover the underlying structure in a given set of data. This differs from supervised learning because there is no dependent variable (label) to predict.

- **Anomaly detection**
- Given some standard event type often occuring over time, we might want to report when a nonstandard type of event occurs. For example, a security officer might want to receive notifications when a strange object (think vehicle, skater, or bicyclist) is observed on a pathway.

Unsupervised Learning

- **User segmentation**

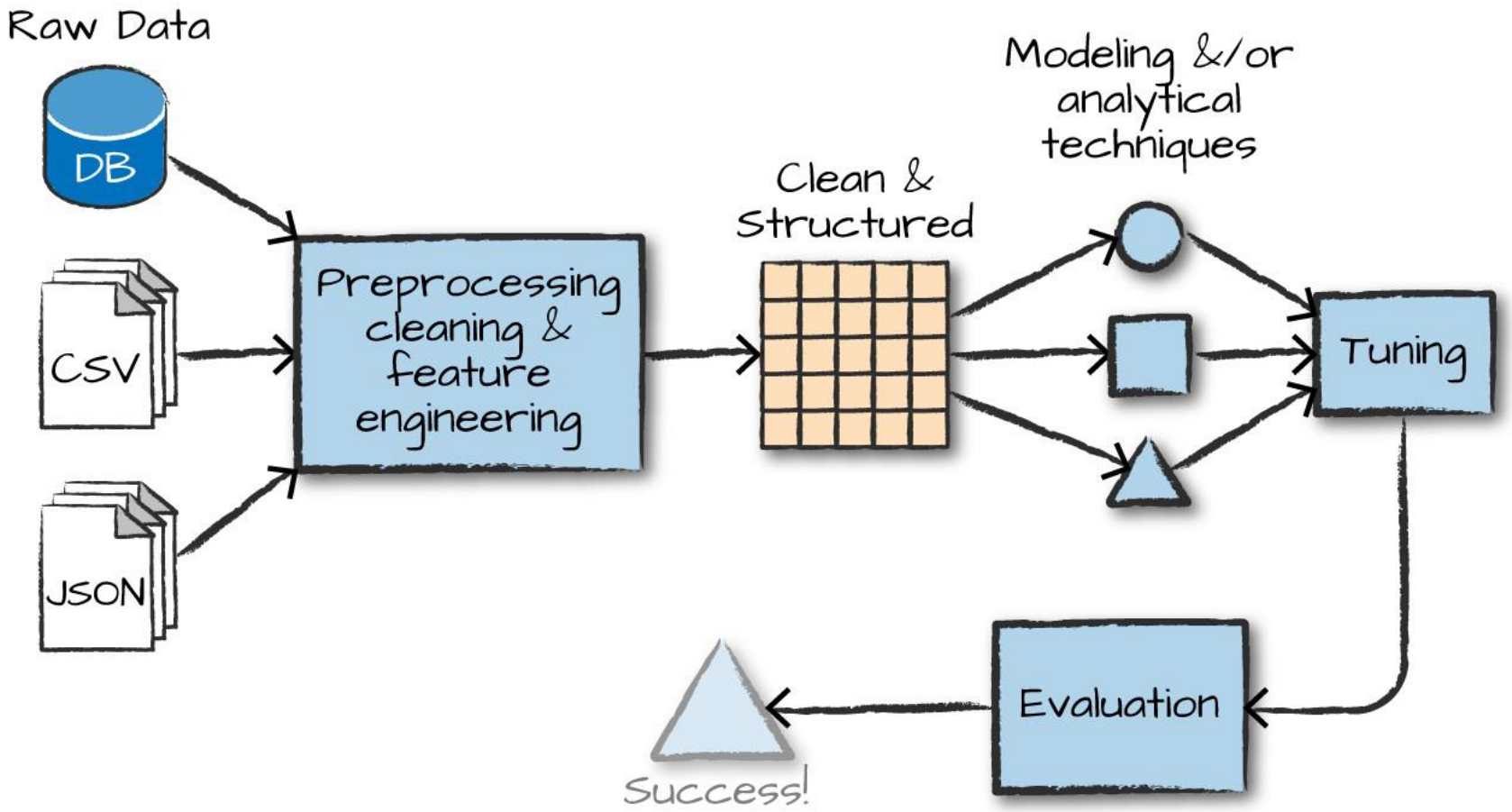
- Given a set of user behaviors, we might want to better understand what attributes certain users share with other users. For instance, a gaming company might cluster users based on properties like the number of hours played in a given game.
- The algorithm might reveal that casual players have very different behavior than hardcore gamers, for example, and allow the company to offer different recommendations or rewards to each player.

Unsupervised Learning

- **Topic modelling**

- Given a set of documents, we might analyze the different words contained therein to see if there is some underlying relation between them.
- For example, given a number of web pages on data analytics, a topic modeling algorithm can cluster them into pages about machine learning, SQL, streaming, and so on based on groups of words that are more common in one topic than in others.

Machine learning workflow



Machine learning workflow

The overall process involves, the following steps (with some variation):

- Gathering and collecting the relevant data for your task.
- Cleaning and inspecting the data to better understand it.
- Performing feature engineering to allow the algorithm to leverage the data in a suitable form (e.g., converting the data to numerical vectors).

Machine learning workflow

- Using a portion of this data as a training set to train one or more algorithms to generate some candidate models.
- Evaluating and comparing models against your success criteria by objectively measuring results on a subset of the same data that was not used for training.

This allows you to better understand how your model may perform in the wild.

- Leveraging the insights from the above process and/or using the model to make predictions, detect anomalies, or solve more general business challenges.