- Now that we prepared our data,

to capture associations between items or events.

The rules are used to determine

when items or events occur together.

A common application of association analysis

is known as market basket analysis

which is used to understand customer purchasing behavior.

For example,

association analysis can reveal

that banking customers who have

certificate of deposit accounts, shortly CDs,

also tend to be interested in other investment vehicles

such as money market accounts.

This information can be used for cross selling.

If you advertise money market accounts to your customers

with CDs, they're likely to open such an account.

According to data mining folklore,

a supermarket chain used association analysis

to discover a connection

between two seemingly unrelated products.

They discovered that many customers

who go to the supermarket late on Sunday night

to buy diapers also tend to buy beer.

This information was then used

to place beer and diapers close together on Sundays

and they saw a jump in sales of both items.

This is the famous diaper-beer connection.

Then your data can be transformed

into a graph representation with nodes and links.

Then you want to use graph analytics

to analyze your data.

This kind of data comes about

when you have a lot of entities

and connections between those entities

like social networks.

Some examples where graph analytics can be useful

are exploring the spread of a disease

or epidemic by analyzing hospitals

and doctors' records,

identification of security threats

by monitoring social media, email and text data,

and optimization of mobile communication network traffic

to ensure data quality

and reduce dropped calls.

Modeling starts with selecting one of these techniques

we listed as the appropriate analysis technique

depending on the type of problem you have.

Then,

you construct the model using the data

that you've prepared.

To validate the model,

you apply it to new data samples.

This is to evaluate how well the model does

on data that was used to construct it.

The common practice is to divide the prepared data

into a set of data for constructing the model

and reserving some of the data

for evaluating the model

after it has been constructed.

You can also use new data

prepared the same way

as the data that was used to construct the model.

Evaluating the model

depends on the type of analysis technique you used.

we finally can work on the analysis.

After this video,

you will be able to describe

what is involved in applying an analysis technique

to your data,

and list three basic analysis techniques.

Now that you have your data nicely prepared,

the next step is to analyze the data.

Data analysis involves building a model from your data

which is called input data.

The input data

is used by the analysis technique to build a model.

What your model generates is the output data.

There are different types of problems

and so there are different types of analysis techniques.

The main categories of analysis techniques

are classification, regression, clustering,

association analysis and graph analysis.

We will now describe each one.

In classification,

the goal is to predict the category of the input data.

An example of this is predicting the weather

as being sunny, rainy, windy or cloudy.

The categories to be predicted

are sunny, rainy, windy and cloudy in this case.

Another example is to classify a tumor

as either benign or malignant.

In this case,

the classification is referred to as binary classifications

as they are only two categories

but you can have many categories as well,

as the weather prediction problem shown here

with four different categories.

Another example is to identify

handwritten digits as being one of 10 categories,

that is zero to nine.

When your model has to predict a numeric value

instead of a category,

then the task becomes a regression problem.

An example of regression

is to predict the price of a stock over time.

The stock price is a numeric value

not a category,

so this is a regression task

instead of a classification task.

Other examples of regression

are estimating the weekly sales of a new product

and predicting the score on a test.

Next is clustering.

In clustering

the goal is to organize similar items into groups.

An example is grouping a company's customer base

into distinct segments for more effective targeted marketing

like seniors, adults and teenagers as we see here.

Other examples include

identifying areas of similar topography,

for example, mountains, desert, plains

for a land user application,

and determining different groups of weather patterns

like rainy, cold or snowy.

The goal in association analysis

is to come up with a set of rules

Let's briefly look at how to evaluate each technique.

and we need to know about these techniques

to make sure we apply the right technique

to our dataset and problem.

For classification and regression

you will have the correct output

for each sample in your input data.

Comparing the correct output

and the output predicted by the model

provides a way to evaluate the model.

For clustering,

the groups resulting from clustering

should be examined to see if they make sense

for your application.

For example,

do the customer segments reflect your customer base?

Are they helpful for use

in your targeted marketing campaigns?

For association analysis and graph analysis

some investigation will be needed

to see if the results are correct.

For example,

network traffic delays need to be investigated

to see if what your model predicts

is actually happening,

and whether the sources of the delays

are where they are predicted to be in the real system.

After you have evaluated your model

to get a sense of its performance on your data,

you will be able to determine the next steps.

Some questions to consider are

should the analysis be performed

with more data in order to get better model performance?

Would using different data help?

For example, in your clustering results,

is it difficult to distinguish customers

from distinct regions?

Or would adding zip codes to your input data

to help generate final grain customer segments is needed?

Do the analysis results suggest a more detailed look

at some aspect of the problem?

For example, predicting sunny weather

gives very good results

but rainy weather predictions are just so so.

This means you should take a closer look

at your examples for rainy weather.

Perhaps there are some anomalies in those samples

or perhaps there are some missing data

that needs to be included

in order to completely capture rainy weather.

The ideal situation would be

that your model performs very well

with respect to success criteria

that were determined

when you defined a problem at the beginning of the project.

In that case,

you're ready to move on to communicating

and acting on the results that you obtained

from your analysis.

As a summary,

data analysis involves selecting the appropriate technique

for your problem,

building the model,

then evaluating the results.

As there are different types of problems,

there also different types of analysis techniques,