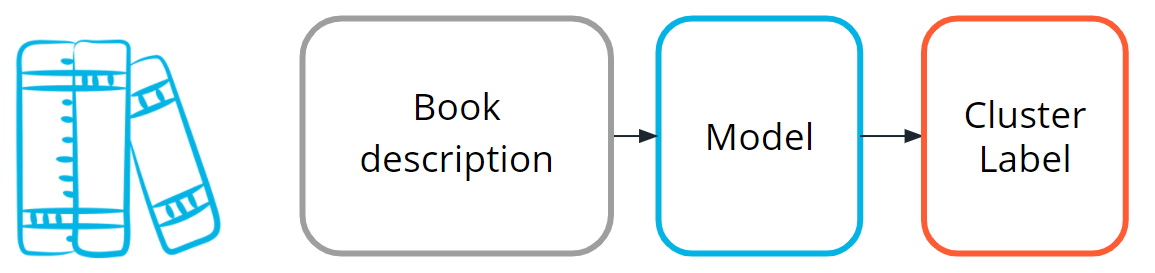
Case Study 2 - Book Genre Exploration

# Step One: Define the Problem

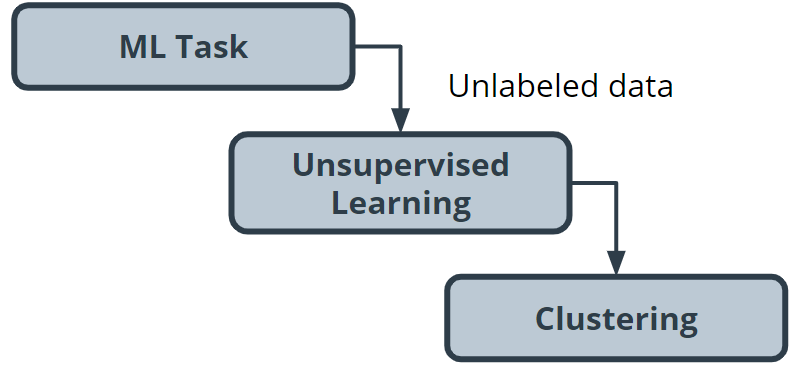


Find clusters of similar books based on the presence of common words in the book descriptions.

You do editorial work for a book recommendation company, and you want to write an article on the largest book trends of the year. You believe that a trend called "micro-genres" exists, and you have confidence that you can use the book description text to identify these micro-genres.

By using an unsupervised machine learning technique called clustering, you can test your hypothesis that the book description text can be used to identify these "hidden" micro-genres.

Earlier in this lesson, you were introduced to the idea of unsupervised learning. This machine learning task is especially useful when your data is not labeled.



# Step Two: Build your Dataset

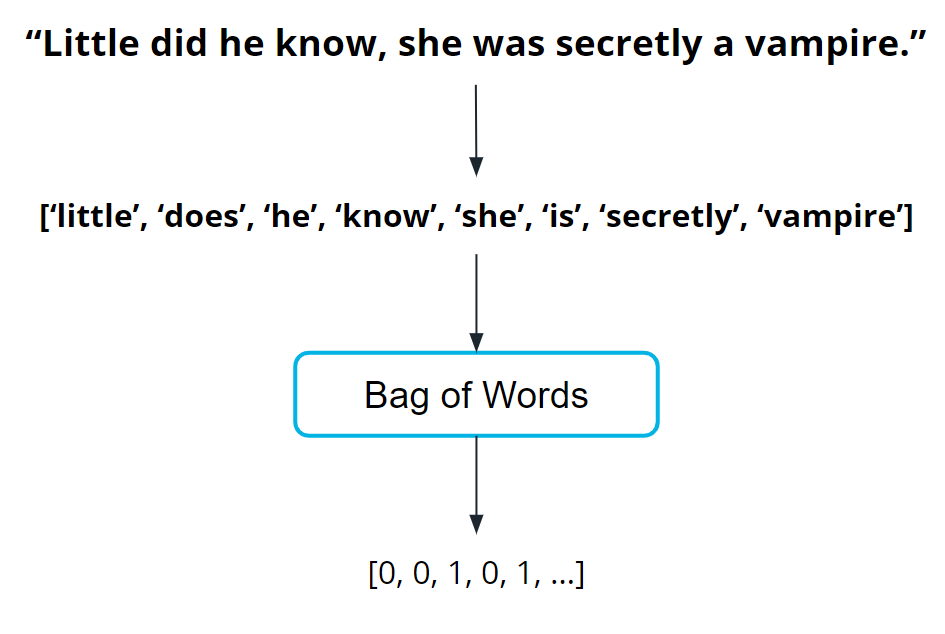
To test the hypothesis, you gather book description text for 800 romance books published in the current year.

## Data exploration, cleaning and preprocessing

For this project, you believe capitalization and verb tense will not matter, and therefore you remove capitals and convert all verbs to the same tense using a Python library built for processing human language. You also remove punctuation and words you don’t think have useful meaning, like 'a' and 'the'. The machine learning community refers to these words as stop words.

Before you can train the model, you need to do some data preprocessing, called data vectorization, to convert text into numbers.

You transform this book description text into what is called a bag of words representation shown in the following image so that it is understandable by machine learning models.



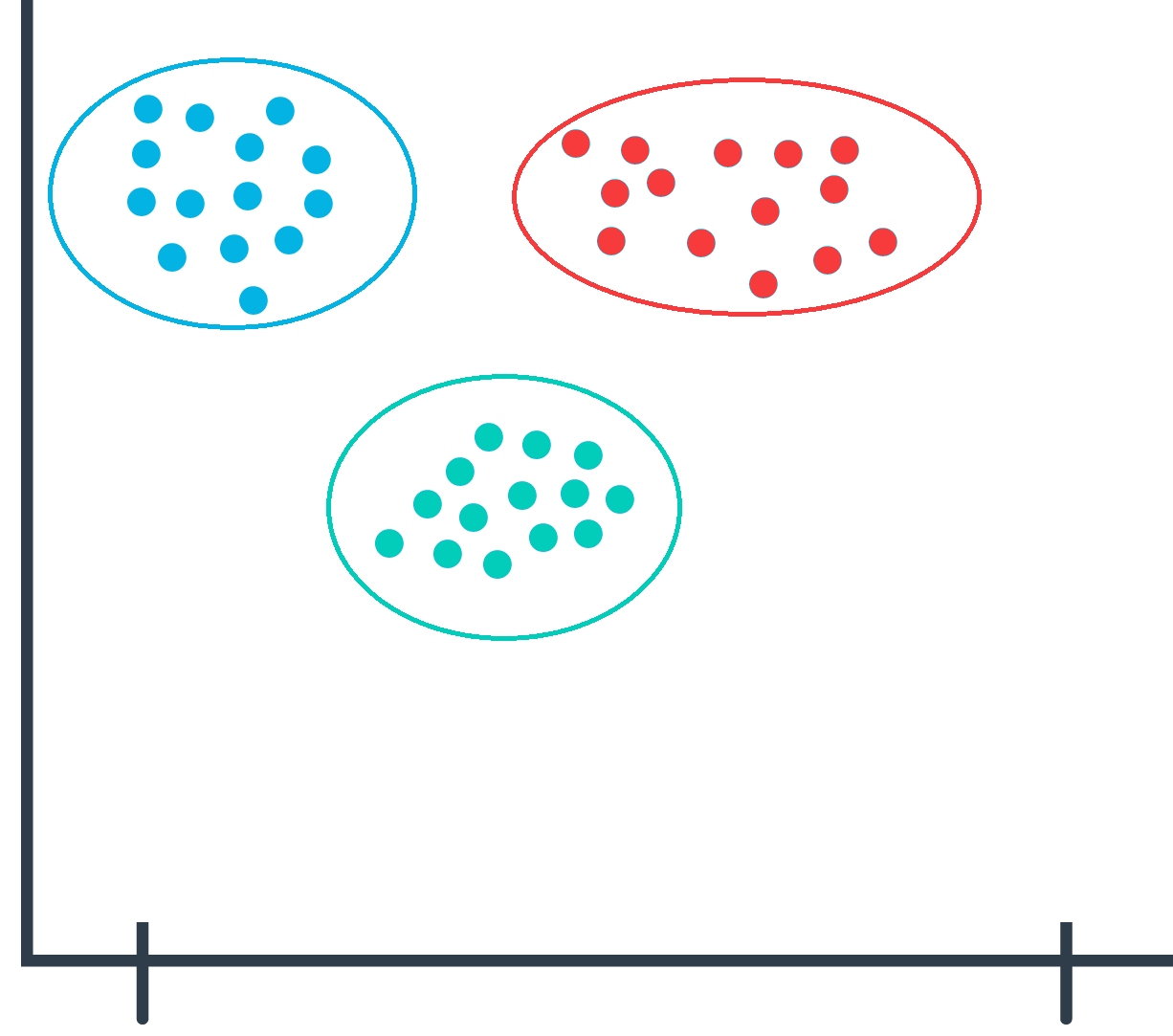
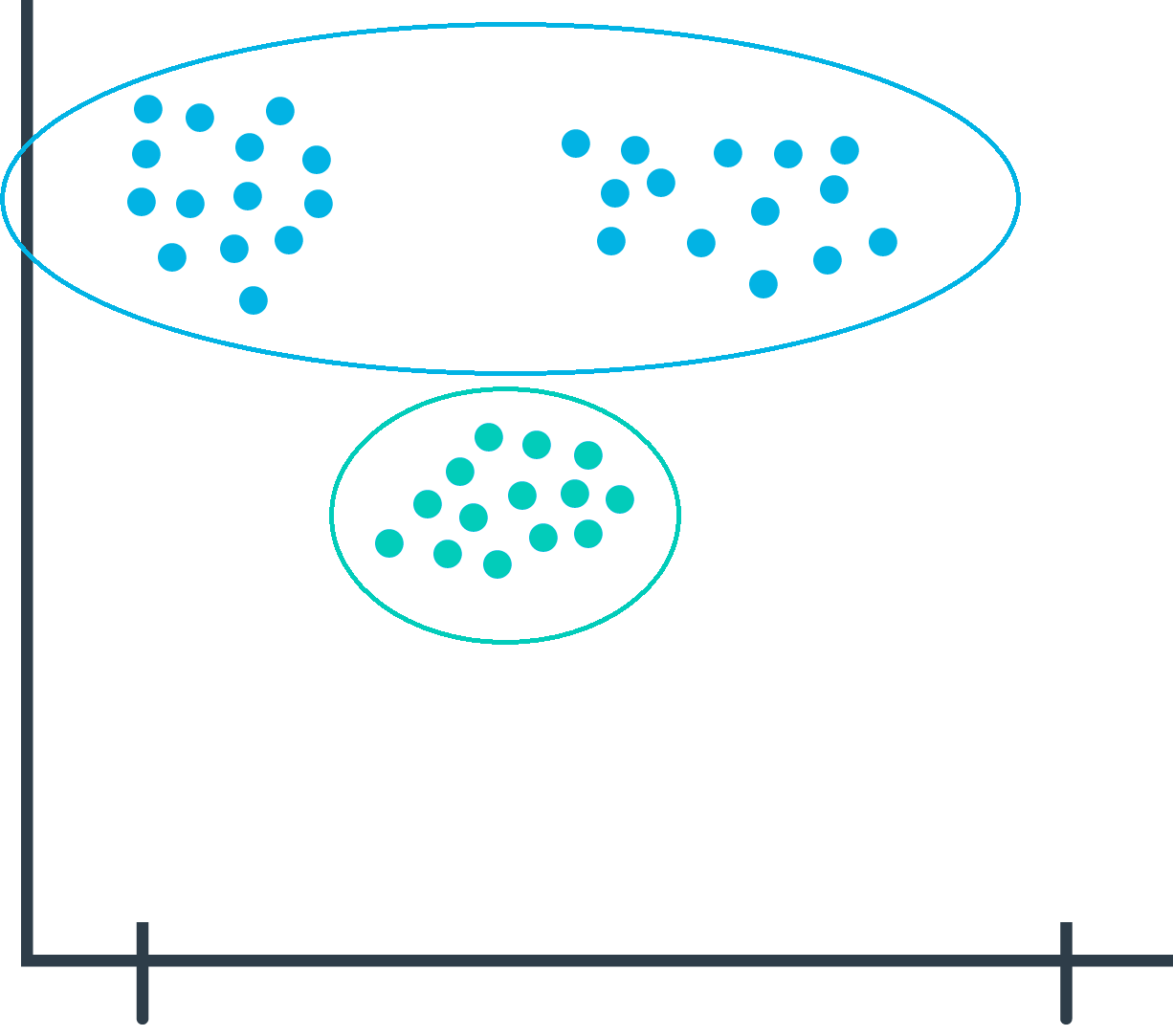
# Step Three: Train the Model

Now you are ready to train your model.

You pick a common cluster-finding model called k-means. In this model, you can change a model parameter, k, to be equal to how many clusters the model will try to find in your dataset.

Your data is unlabeled: you don't how many microgenres might exist. So, you train your model multiple times using different values for k each time.

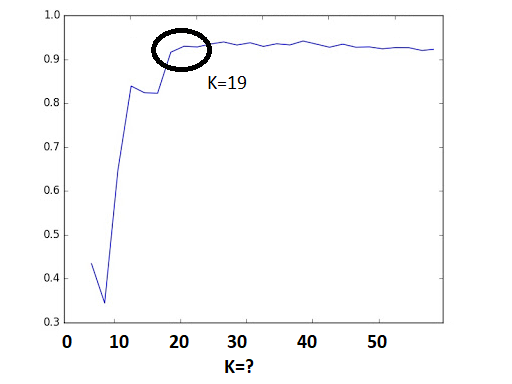
What does this even mean? In the following graphs, you can see examples of when k=2 and when k=3.



During the model evaluation phase, you plan on using a metric to find which value for **k** is most appropriate.

# Step Four: Model Evaluation

In machine learning, numerous statistical metrics or methods are available to evaluate a model. In this use case, the silhouette coefficient is a good choice. This metric describes how well your data was clustered by the model. To find the optimal number of clusters, you plot the silhouette coefficient as shown in the following image below. You find the optimal value is when k=19.

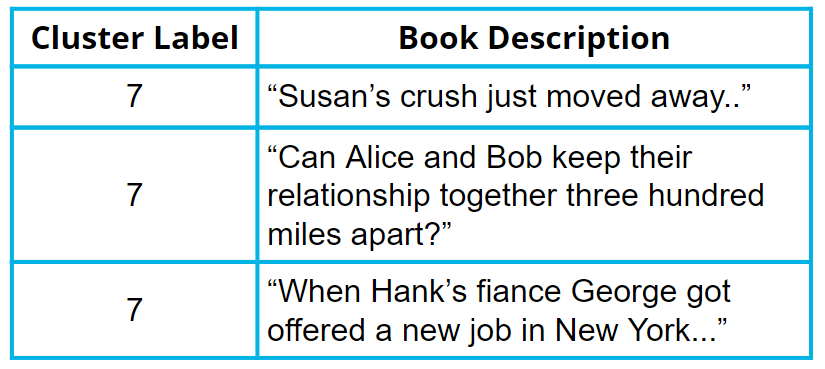


Often, machine learning practitioners do a manual evaluation of the model's findings.

You find one cluster that contains a large collection of books you can categorize as “paranormal teen romance.” This trend is known in your industry, and therefore you feel somewhat confident in your machine learning approach. You don’t know if every cluster is going to be as cohesive as this, but you decide to use this model to see if you can find anything interesting about which to write an article.

# Step Five: Inference (Use the Model)

As you inspect the different clusters found when k=19, you find a surprisingly large cluster of books. Here's an example from fictionalized cluster #7.



As you inspect the preceding table, you can see that most of these text snippets are indicating that the characters are in some kind of long-distance relationship. You see a few other self-consistent clusters and feel you now have enough useful data to begin writing an article on unexpected modern romance microgenres.