Sample questions can include:

- *What will be the temperature of a city that is 250 miles away from the sea?*
- *What will be the grades of a student studying at a high school based on their primary school grades?*
- *What are the chances that a person will be obese based on the amount of their food intake?*

Regression and correlation have a number of important differences. Correlation does not imply causation. The change in the value of one variable may not be responsible for the change in the value of the second variable, although both may change at the same rate. This can occur due to an unknown third variable, known as the confounding factor. Correlation assumes that both variables are independent.

Regression, on the other hand, is applicable to variables that have previously been identified as dependent and independent variables and implies that there is a degree of causation between the variables. The causation may be direct or indirect.

Within Big Data, correlation can first be applied to discover if a relationship exists. Regression can then be applied to further explore the relationship and predict the values of the dependent variable, based on the known values of the independent variable.

Machine Learning

Humans are good at spotting patterns and relationships within data. Unfortunately, we cannot process large amounts of data very quickly. Machines, on the other hand, are very adept at processing large amounts of data quickly, but only if they know how.

If human knowledge can be combined with the processing speed of machines, machines will be able to process large amounts of data without requiring much human intervention. This is the basic concept of machine learning.

In this section, machine learning and its relationship to data mining are explored through coverage of the following types of machine learning techniques:

- Classification
- Clustering
- Outlier Detection
- Filtering

Classification (Supervised Machine Learning)

Classification is a supervised learning technique by which data is classified into relevant, previously learned categories. It consists of two steps:

1. The system is fed training data that is already categorized or labeled, so that it can develop an understanding of the different categories.
2. The system is fed unknown but similar data for classification and based on the understanding it developed from the training data, the algorithm will classify the

unlabeled data.

A common application of this technique is for the filtering of email spam. Note that classification can be performed for two or more categories. In a simplified classification process, the machine is fed labeled data during training that builds its understanding of the classification, as shown in Figure 8.11. The machine is then fed unlabeled data, which it classifies itself.

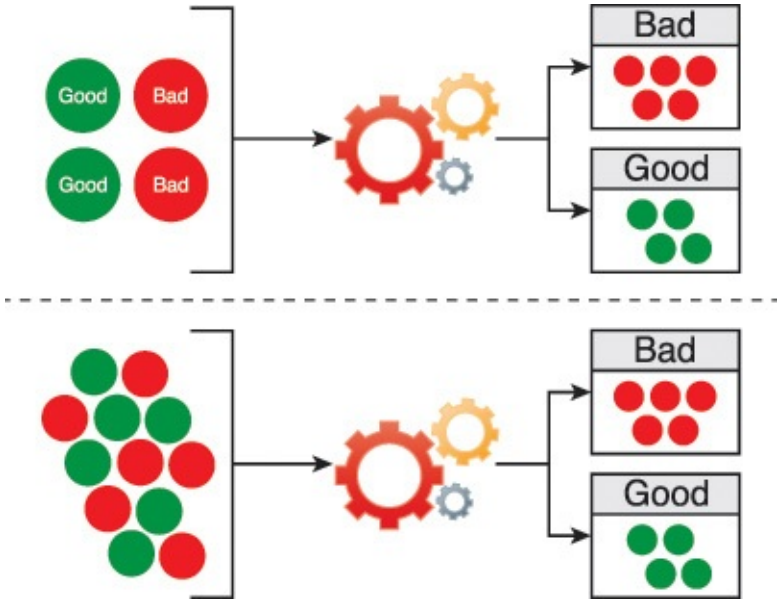

Figure 8.11 Machine learning can be used to automatically classify datasets.

For example, a bank wants to find out which of its customers is likely to default on loan payments. Based on historic data, a training dataset is compiled that contains labeled examples of customers that have or have not previously defaulted. This training data is fed to a classification algorithm that is used to develop an understanding of "good" and "bad" customers. Finally, new untagged customer data is fed in order to find out whether a given customer belongs to the defaulting category.

Sample questions can include:

- *Should an applicant's credit card application be accepted or rejected based on other accepted or rejected applications?*
- *Is a tomato a fruit or a vegetable based on the known examples of fruit and vegetables?*
- *Do the medical test results for the patient indicate a risk for a heart attack?*

Clustering (Unsupervised Machine Learning)

Clustering is an unsupervised learning technique by which data is divided into different groups so that the data in each group has similar properties. There is no prior learning of categories required. Instead, categories are implicitly generated based on the data groupings. How the data is grouped depends on the type of algorithm used. Each algorithm uses a different technique to identify clusters.

Clustering is generally used in data mining to get an understanding of the properties of a given dataset. After developing this understanding, classification can be used to make

better predictions about similar but new or unseen data.

Clustering can be applied to the categorization of unknown documents and to personalized marketing campaigns by grouping together customers with similar behavior. A scatter graph provides a visual representation of clusters in Figure 8.12.

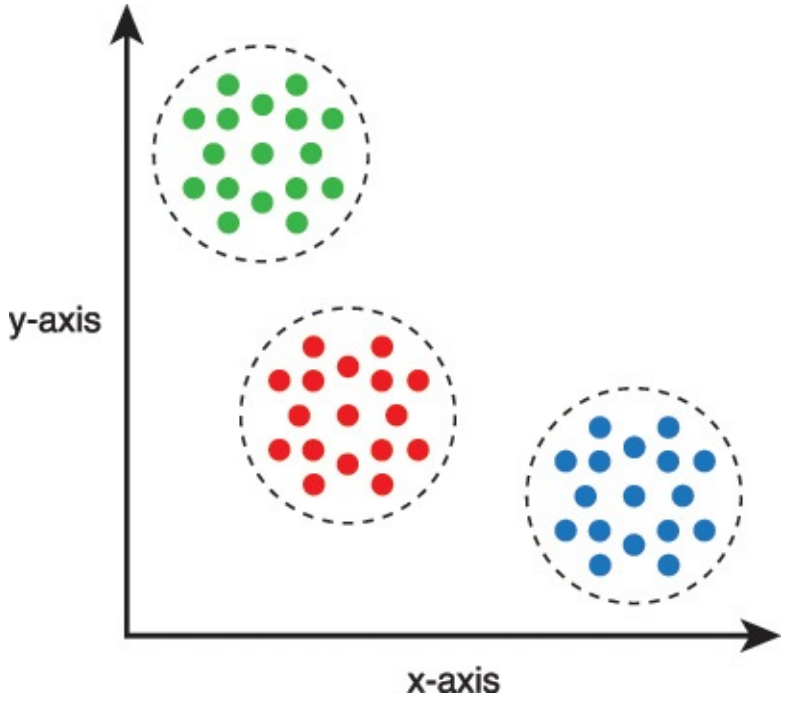

Figure 8.12 A scatter graph summarizes the results of clustering.

For example, a bank wants to introduce its existing customers to a range of new financial products based on the customer profiles it has on record. The analysts categorize customers into multiple groups using clustering. Each group is then introduced to one or more financial products most suitable to the characteristics of the overall profile of the group.

Sample questions can include:

- *How many different species of trees exist based on the similarity between trees?*
- *How many groups of customers exist based upon similar purchase history?*
- *What are the different groups of viruses based on their characteristics?*

Outlier Detection

Outlier detection is the process of finding data that is significantly different from or inconsistent with the rest of the data within a given dataset. This machine learning technique is used to identify anomalies, abnormalities and deviations that can be advantageous, such as opportunities, or unfavorable, such as risks.

Outlier detection is closely related to the concept of classification and clustering, although its algorithms focus on finding abnormal values. It can be based on either supervised or unsupervised learning. Applications for outlier detection include fraud detection, medical diagnosis, network data analysis and sensor data analysis. A scatter graph visually highlights data points that are outliers, as shown in Figure 8.13.

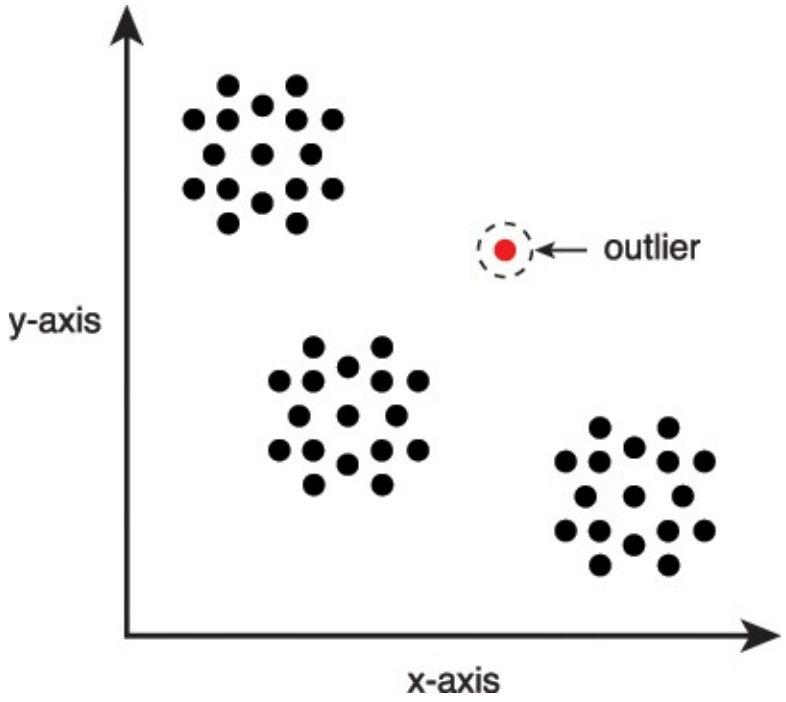

Figure 8.13 A scatter graph highlights an outlier.

For example, in order to find out whether or not a transaction is likely to be fraudulent, the bank's IT team builds a system employing an outlier detection technique that is based on supervised learning. A set of known fraudulent transactions is first fed into the outlier detection algorithm. After training the system, unknown transactions are then fed into the outlier detection algorithm to predict if they are fraudulent or not.

Sample questions can include:

- *Is an athlete using performance enhancing drugs?*
- *Are there any wrongly identified fruits and vegetables in the training dataset used for a classification task?*
- *Is there a particular strain of virus that does not respond to medication?*

Filtering

Filtering is the automated process of finding relevant items from a pool of items. Items can be filtered either based on a user's own behavior or by matching the behavior of multiple users. Filtering is generally applied via the following two approaches:

- collaborative filtering
- content-based filtering

A common medium by which filtering is implemented is via the use of a recommender system. Collaborative filtering is an item filtering technique based on the collaboration, or merging, of a user's past behavior with the behaviors of others. A target user's past behavior, including their likes, ratings, purchase history and more, is collaborated with the behavior of similar users. Based on the similarity of the users' behavior, items are filtered for the target user.

Collaborative filtering is solely based on the similarity between users' behavior. It requires a large amount of user behavior data in order to accurately filter items. It is an example of

the application of the law of large numbers.

Content-based filtering is an item filtering technique focused on the similarity between users and items. A user profile is created based on that user's past behavior, for example, their likes, ratings and purchase history. The similarities identified between the user profile and the attributes of various items lead to items being filtered for the user. Contrary to collaborative filtering, content-based filtering is solely dedicated to individual user preferences and does not require data about other users.

A recommender system predicts user preferences and generates suggestions for the user accordingly. Suggestions commonly pertain to recommending items, such as movies, books, Web pages and people. A recommender system typically uses either collaborative filtering or content-based filtering to generate suggestions. It may also be based on a hybrid of both collaborative filtering and content-based filtering to fine-tune the accuracy and effectiveness of generated suggestions.

For example, in order to realize cross-selling opportunities, the bank builds a recommender system that uses content-based filtering. Based on matches found between financial products purchased by customers and the properties of similar financial products, the recommender system automates suggestions for potential financial products that customers may also be interested in.

Sample questions can include:

- *How can only the news articles that a user is interested in be displayed?*
- *Which holiday destinations can be recommended based on the travel history of a vacationer?*
- *Which other new users can be suggested as friends based on the current profile of a person?*

Semantic Analysis

A fragment of text or speech data can carry different meanings in different contexts, whereas a complete sentence may retain its meaning, even if structured in different ways. In order for the machines to extract valuable information, text and speech data needs to be understood by the machines in the same way as humans do. Semantic analysis represents practices for extracting meaningful information from textual and speech data.

This section describes the following types of semantic analysis:

- Natural Language Processing
- Text Analytics
- Sentiment Analysis

Natural Language Processing

Natural language processing is a computer's ability to comprehend human speech and text as naturally understood by humans. This allows computers to perform a variety of useful tasks, such as full-text searches.