Hello and welcome to the openSAP course “Getting Started with Data Science”. My name is Stuart Clarke and I am a consultant with SAP, specializing in data science and predictive analytics.

So now, what should you expect over the next six weeks? Let’s take a look. The goal of this course is to provide you with an introduction to data science.

Data science covers a wide number of topics and many people find it difficult to understand. I will try to demystify data science for you so that you have a good, clear understanding of the major topics.

We will introduce Data Science concepts and tools, and we will cover these from a conceptual perspective and also include demos showing the key steps in the model development process.

By the end of the course, you should have a basic understanding that will enable you to take the next steps in your Data Science adoption journey.

We will teach you all you need to know about the fundamentals of Data Science. Let me share with you some of the more important details of what we’re going to cover in the next six weeks.

In Week 1, we’re going to look at business and data understanding. We’ll have an introduction to data science, we’ll look at project methodologies, and then we’ll look at an overview of the business understanding phase and also the data understanding phase. Then in Week 2 we’ll start to look at data preparation, and also we’ll look at the predictive modeling methodology and data manipulation. In Week 3, we’ll move on to the first week of modeling, and we’ll look at how we can detect anomalies within data, look at association analysis, cluster analysis, and classification analysis using regression.

In Week 4, we’ll move on to the second week of the modeling topic, where we’ll look at classification analysis using decision trees and neural networks.

We’ll also look at time series analysis and simulation and optimization. Then in Week 5 we’ll move on to the evaluation phase, and we’ll look at some of the model performance metrics and how we can test and improve models. And then on to Week 6, where we’ll start to look at the deployment of our models and how we can maintain these over time.

After having successfully completed the first six weeks, you’ll have one further week to prepare for and participate in the final exam to earn a record of your achievement.

Throughout the course your feedback, questions, and your ideas are very much appreciated in our discussion forum. So how to you get points and successfully complete the course?

Well, there are six graded assignments throughout the first six weeks of instructional content. Each assignment is worth 30 points for a total of 180 points, which is half of the total points available in the course.

The other half of the available points come from a final exam. And just like every openSAP course, you will need at least half of the maximum points available – in this case 180 points – to pass the course and receive your record of achievement. So let's start with topic 1, and let's have an introduction to “What is data science?”
Well, it's an interdisciplinary field about processes and systems to extract knowledge or insights from data. Data science employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, operations research, information science, and computer science, including signal processing, probability models, machine learning, statistical learning, data mining, database and data engineering, pattern recognition, data visualization, and predictive analytics.

SAP's Data Science products address the needs of a wide range of different users. There are very few trained data scientists available in many organizations, and we also know that many data analysts often require access to these very powerful data analysis tools. So, as well as providing a wide range of powerful tools for data scientists, we have also developed easy-to-use and understand data science prebuilt functionality that is accessible to data analysts and other business users.

Business users who need access to data science tools, possibly embedded into other SAP applications, such as Hybris Marketing or SAP’s fraud detection solutions, or to develop data visualizations, are now able to do so very easily.

Therefore, SAP’s Data Science solutions offer data science capabilities to users with low up to high data science skill sets. Over the next few weeks we will see how SAP’s data science tool kit can be used to leverage your data and deliver complex data science projects. Please remember that these data science capabilities are designed to work with data stored in text files or with a wide range of proprietary data systems, or embedded into SAP HANA, which means that the benefits of the SAP HANA platform (its Big Data and real-time capabilities) can also be used to enhance data science activities within an organization.

SAP offers a comprehensive range of data science products that... supports the full spectrum of creators and consumers of data science projects; supports the full spectrum of functionality for data science projects, including statistical analysis, machine learning, predictive modeling, optimization, and simulation; provides an in-database predictive analysis library in SAP HANA; provides SAP HANA integration for R; provides a Predictive Analytics workbench for data analysts and data scientists, as well as an Application Function Modeler in SAP HANA for application developers.

SAP HANA’s SQLScript is an extension of It includes enhanced control-flow capabilities and lets developers define complex application logic inside database procedures. However, it is difficult to describe predictive analysis logic within a procedure. For example, an application may need to perform a cluster analysis in a huge customer table with a billion records.

It is impossible to implement the analysis in a procedure using the simple classic K-means algorithm, or even with more complicated algorithms. Transferring large tables to the application server to perform the K-means calculation is also very costly. The SAP HANA Predictive Analysis Library (PAL) defines functions that can be called from within SQLScript procedures to perform analytic algorithms. PAL are designed to perform in-memory data mining and statistical calculations to provide high performance on large data sets and for real-time analytics. PAL includes classic and universal predictive analysis algorithms in nine data-mining categories: clustering, classification, regression, association, time series and forecasting, data preprocessing, statistics, social network analysis, and there's a range of miscellaneous algorithms.
SAP HANA APL is an Application Function Library that lets you use the data mining capabilities of the SAP Predictive Analytics automated analytics engine on your customer datasets stored in SAP HANA.

You can create a wide range of models to answer your business questions and take advantage of all of the automated modeling capabilities which we will look at in much more detail during this course. APL covers classification and regression models, clustering, time series analysis and forecasting.

This enables your favorite R scripts to be run as part of the analytics flow either using SAP HANA SQL or controlled by SAP Predictive Analytics.

The SAP HANA database uses the external R environment to execute this R code. This allows the application developer to embed R function definitions and calls within SQLScript and submit the entire code as part of a query to the database. Like any other operator, the R operator consumes a number of input objects, for example, intermediate tables retrieved from previously computed operations or other data sources like a column or a row store table, and then it returns a result table.

SAP Predictive Analytics is built for both data scientists as well as business and data analysts, making predictive analytics accessible to a broad spectrum of users.

It has automated and expert modes, which we will look at in more detail. It provides automated data preparation, predictive modeling, and model deployment and maintenance functionality.

It also provides access and control of the PAL and APL algorithms, as well as R language support. Because it is based on SAP Lumira, it has excellent advanced visualization features.

And it has native integration with SAP HANA. The Application Function Modeler is a graphical tool to build advanced applications in SAP HANA.

It provides a Web-based flow-graph editor as well as PAL function support and is designed for SAP HANA data science application developers.

Well, that's it for Unit 1. Thanks for joining.

In the next unit, we will talk about an introduction to the project methodologies. See you there.
Hello, and welcome back to Unit 2. In the last unit you heard an introduction to Data Science. In this unit I will give you an introduction to Project Methodologies. There is a wide range of data mining or data science project methodologies that have been developed over the years, some of which you may have heard of before. These include SPSS’s 5As, The Virtuous Circle, the KDD Process, SAS’s SEMMA, KXEN had a DMAIC process based on a Six Sigma methodology, and of course there is CRISP. In 2015, IBM Corporation released a new methodology which they have called Analytics Solutions Unified Method for Data Mining and Predictive Analytics, which is also known as the ASUM process, and this was designed to refine and extend CRISP. The most popular project methodology is the cross-industry standard process for data mining, which we call CRISP. This was an initiative that was launched in 1996 led by five companies: SPSS, Teradata, Daimler AG, NCR Corporation, and OHRA, which is an insurance company. Over 300 organizations also contributed to the process model. The goal was to create a data-centric project methodology that is non-proprietary, application and industry-neutral, tool-neutral, and focused on business issues as well as technical issues. Polls conducted by KDNuggets in 2002, 2004, 2007, and 2014 show that CRISP-DM was the leading methodology used by all industry data miners. The CRISP methodology is a hierarchical process. At the top level, the process is divided into six different generic phases, ranging from business understanding to the deployment of the project results. The next level elaborates each of these different phases, comprising several generic tasks. At this level, the description is generic enough to cover all data science scenarios. The third level specializes these tasks for specific situations. For example, the generic task might be cleaning data, and the specialized task could be cleaning of numeric or categorical values. The fourth level is the process – the record of actions, decisions, and results of an actual execution of a Data Science project. The six generic phases are represented in the diagram on this slide. “Business understanding” confirms the project objectives and requirements from the business perspective. It defines the data science approach that will answer the specific business objectives. Then “data understanding” looks at the initial data collection and familiarization. We identify any data quality problems. We look at data preparation and the selection of data tables, records, and attributes, and undertake any data transformations and cleaning that is required. Then we move on to the "modeling" phase, where we select the modeling techniques, we calibrate the model parameters and build the models. And then we move on to the "evaluation" phase, where we confirm that the business objectives have actually been achieved. We look at the "deployment" phase, where we actually deploy the models and then productization each of the models if that is required, and develop and implement a repeatable process that enables the organization to monitor and maintain each model’s performance. The sequence of the phases is not strict and moving back and forth between different phases is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of any data science project.
Of course, the process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions, and subsequent data science processes will benefit from the experiences of previous ones. Let's look at phase 1 – the business understanding process.

The objective is to focus on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data science problem definition and a preliminary plan designed to achieve the objectives.

Phase 2 is the data understanding phase. The objective is to start with an initial data collection and then proceed with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.

Phase 3 is the data preparation phase. Here, this covers all activities to construct the final dataset from the initial raw data.

Phase 4 is the modeling phase. Various modeling techniques are selected and applied, and their parameters are calibrated to the optimal values.

Some techniques have specific requirements for the form of data. Therefore, stepping back to the data preparation phase is often necessary.

Phase 5 is the evaluation phase. Here, we thoroughly evaluate the model and review the model construction to be certain it properly achieves the business objectives.

A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of these data science results should be reached.

Phase 6 is the deployment phase. The knowledge gained will need to be organized and presented in a way that the organization can actually use.

However, depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Most models' predictive performance will degrade over time, because the data that will be used to apply the model onto will change. The data distributions might change as customer characteristics change, competitors launch campaigns, and the general business environment changes. The models must be updated when this happens.

A monitoring phase can be added to the CRISP methodology that specifically focuses on this very important aspect of any data science project.

We will be looking at all of these phases in more detail over the next six weeks. In the next unit we will start with an overview of the business understanding phase.

See you there.
Hi and welcome to Unit 3 of the course “Getting Started with Data Science”. With this unit, we will start looking at the phases of the CRISP methodology in detail.

Let’s start with an overview of the business understanding phase. This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data science problem definition and a preliminary plan designed to achieve the objectives. Task 1 determines the business objectives.

We need to gain a thorough understanding – from a business perspective – of what the client really wants to accomplish. We must uncover any important factors at the beginning of the project that could influence the outcome of the project.

Remember that missing this step could mean we waste a great deal of effort producing the right answers to the wrong questions. Task 2 assesses the situation.

Task 3 determines the data science goals. A business goal states objectives in business terminology. A data science goal states project objectives in technical terms. Task 4 is to produce a project plan.

The first objective is to thoroughly understand – from a business perspective – what the client really wants to accomplish. Often the organization has many competing objectives and constraints that must be properly balanced.

The analyst’s goal is to uncover important factors at the beginning that can influence the outcome of the project. The outputs need to assess the background.

We record the information that is known about the organization’s business situation at the beginning of the project. We describe the customer’s primary objective from a business perspective.

In addition to the primary business objective, there are typically other related business questions that the organization would like to address.

For example, the primary business goal for a financial services business might be to keep current customers by predicting when they are prone to move to a competitor.

Examples of related business questions are: “How does the primary channel a bank customer uses, for example, the ATM, branch, or Internet, affect whether they are going to stay or go?”, or “Will lower ATM fees significantly reduce the number of high-value customers who leave?”

We describe the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and able to be measured objectively, such as a reduction of customer churn to a certain level, or general and subjective such as “give useful insights into the relationship”. The next task is to assess the business situation.

In the previous task, the objective is to quickly get to the crux of the situation. Here, we want to flesh out the details.

This task involves more detailed fact-finding about all of the resources, constraints, assumptions, and other factors that should be considered in determining the data science goal and project plan.

We need to list the resources available to the project, including: the personnel (for example, business experts, the data experts, the technical support, and data science personnel),
the data (for example, the fixed extracts we will be using, access to live warehouse or operational data), the computing resources (for example, the hardware platform we will be using).

and software we will use (for example, the data science tools and any other relevant software). The outputs are an inventory of resources.

We list all of the resources available to the project. We list all requirements of the project, including a schedule of completion.

As part of this output, we need to make sure that we are allowed to use the data. We list the assumptions made by the project,

and these assumptions are about the data that can be checked during the analysis process, but may also include non-checkable assumptions about the business.

It is particularly important to list these if they form conditions on the validity of the results. We list the constraints on the project.

These may be constraints on the availability of resources, but may also include technological constraints such as the size of the data that it is practical to use for modeling.

We list the risks and contingencies. Here we need to list the risks or events that might occur to delay the project or cause it to fail.

We list the corresponding contingency plans and identify what actions will be taken if the risks happen. We compile a glossary of terminology relevant to the project.

This may include two components: A glossary of relevant business terminology, which forms part of the business understanding available to the project,

and a glossary of data science terminology, illustrated with examples relevant to the business problem in question. We also construct a cost-benefit analysis for the project,

which compares the costs of the project with the potential benefit to the business if it is successful. The comparison should be as specific as possible, for example using monetary measures in a commercial situation.

The next task is to determine the data science goals. A business goal states objectives in business terminology.

A data science goal states project objectives in technical terms. For example, the business goal might be to increase catalogue sales to existing customers.

A data science goal might be “Predict how many widgets a customer will buy, given their purchases over the past three years, demographic information (such as their age, salary, city, and so on),

and the price of the item.” The output is listing the data science goals.

We describe the intended outputs of the project that enables the achievement of the business objectives. We also define the data science success criteria for a successful outcome to the project in technical terms.

For example, a certain level of predictive accuracy or a propensity to purchase profile with a given degree of “lift”. The next goal is to produce the project plan.

We describe the intended plan for achieving the data science goals and thereby achieving the business goals. The plan should specify the anticipated set of steps to be performed during the rest of the project,

including an initial selection of tools and techniques. The output is to actually produce the project plan.

We list the stages to be executed in the project, together with the duration, resources, inputs, outputs, and dependencies. Where possible we make explicit the large-scale iterations in the data science process,

for example, repetitions of the modeling and evaluation phases. As part of the project plan, it is also important to analyze dependencies between time schedule and risks.
The project plan is a dynamic document in the sense that at the end of each phase a review of progress and achievements is necessary and an update of the project plan is recommended.

Specific review points for these reviews are part of the project plan. At the end of the first phase, the project also performs an initial assessment of tools and techniques.

For example, you select a data science algorithm that supports the available data and required output. It is important to assess tools and techniques early in the process since this selection possibly influences the entire project.

That's it for now. In the next unit, we will look at the definition of project success criteria.

See you soon.
Welcome back to Unit 4: "Defining Project Success Criteria". As we have learned in previous units, it is important for us to clearly define business and data science project success criteria. We define the criteria for a successful or useful outcome to the project from the business point of view. This might be quite specific and able to be measured objectively, such as reduction of customer churn to a certain level, or it could be general and subjective, such as "give useful insights into the relationship". We also define the criteria for a successful outcome to the project in data science technical terms, for example, a certain level of predictive accuracy or a propensity to purchase profile with a given degree of "lift". These industry surveys indicate standard methods of assessing data science project success.

In both of these surveys, meeting business goals and model accuracy are the two most important factors. On the left-hand side, 57% of responders responded to the question "How do you measure success?" for a predictive analytics project as "meeting business goals", and 56% as "model accuracy". Lift is also an important factor. We will be discussing how we calculate lift in more detail later in this course. On the right side, in their Third Annual Data Miner Survey, conducted by Karl Rexer Analytics, a renowned CRM consultancy firm based in the US, they asked the BI community "How do you evaluate project success in Data Mining?" Out of 14 different criteria, a massive 58% ranked "Model Performance" (Lift and R²) as the primary factor.

The success criteria for the different data science models will differ depending on whether the models are predictive or descriptive type models and the type of algorithm chosen. Descriptive analysis describes or summarizes data and makes it more easily interpretable. It can only analyze historical performance – for example, what happened in the past week, month or year. Descriptive analytics are useful because they allow us to learn from past behaviors, and understand how they might influence future outcomes. Common examples of descriptive analytics are BI reports that provide historical insights regarding a company’s production, financials, operations, sales, finance, inventory, and customers. Descriptive analytical models include cluster models, segmentations, association rules, and network analysis. Predictive analysis predicts what might happen in the future. These models provide estimates or probabilities about the likelihood of a future outcome. One common application is the use of predictive analytics to produce a credit score. These scores are used by financial services businesses to determine the probability of customers making future credit payments on time. Typical business uses include: understanding how sales might close at the end of the year, predicting how much customers will spend in the next 12 months, or forecasting inventory levels based upon a range of variables such as weather conditions or new competitor campaigns. Predictive analytical models include classification models, regression models, and neural network models.

The business question helps to determine the most likely algorithms to use. We can choose algorithms to analyze trends in data and use this information for forecasting. We can identify the main influences and relationships that could be driving customers to switch to another supplier. We call this churn analysis. Or why certain customers are more likely to respond to a campaign offer and buy specific products. Some algorithms group observations or customers together, so that all of the customers in a group have similar characteristics.
These are cluster algorithms. We can use association type algorithms to understand which products to recommend to customers in a cross or up-sell marketing campaign, or to analyze the relationship between certain variables in a data set. And we can identify unusual values in a data set by using anomaly detection algorithms. There is a wide range of algorithms to choose from, depending on the type of question asked by the business.

The output that is required, and the data that is available. For association rules, or basket analysis, we have an Apriori algorithm that analyzes the combinations of products purchased together in a basket or over time. For clustering, where we are creating groups of similar observations, we often use a K-Means algorithm.

For classification analysis, where we are classifying observations into groups, we can use decision trees or neural networks. We can use outlier analysis to identify which observations have unusually high or low values.

Regression algorithms enable us to forecast the values of continuous variables, such as customer spend in the next 12 months. Time series analysis enables us to forecast future KPI values, and control stock and inventory levels.

All of these algorithms will be explored in detail during this course. Each of these broad categories of algorithm can answer different types of business question.

For classification, we can answer the "who" and "when" type questions: "Which customers will buy a product and when will they most likely make the purchase?" Or "Which machine will fail and when will it need preventative maintenance?" Or "Is that transaction fraudulent?"

For regression, we can answer the "what" type questions: "What will be the spend of each customer in the next 12 months?" or "How many customers will churn next year?"

For clustering and segmentation we are grouping together similar observations. This enables us to communicate to customers with similar needs and requirements who are grouped together in a cluster, or develop specific products or services for customers in a segment. Forecasting allows us to estimate a KPI on a regular time interval.

So for example, we can forecast revenue per month for the next 12 months, accounting for trends, seasonalities, and other external factors.

Link analysis is used mainly in telecommunications to create communities of customers who are calling one another, or in retail analysis to analyze the links between customers and the products they have purchased to support product recommendations. And association rules and recommendations are used for basket analysis and also to produce product recommendations for customers.

The accuracy and robustness of the model are two major factors to determine the quality of the prediction, which reflects how successful the model is.

Accuracy is often the starting point for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated.

There are a wide variety of metrics and methods to measure accuracy, such as lift charts and decile tables, which measure the performance of the model against random guessing, or what the results would be if you didn't use any model.

These will be discussed in more detail later in the course. The robustness of a predictive model refers to how well a model works on alternate data.

This might be hold-out data or new data that the model is to be applied onto. The predictive performance of a model must not deteriorate substantially when it is applied to data that was not used in model estimation.
Central to developing predictive models and assessing if they are successful is a train-and-test regime. Data is partitioned into training and test sub-sets.

There are a variety of strategies to cut this data, for example random, sequential, and periodic. We build our model on the training sub-set (called the estimation sub-set), and evaluate its performance on the test sub-set (a hold-out sample called the validation sub-set). Simple two-way and three-way data partitioning is shown in the diagram.

When a predictive model has been built on the estimation sub-sample, its performance is tested on the validation and test sub-samples. We would expect that the model will have similar performance on the estimation, validation, and test sub-sets.

The closer the performance of the model on the sub-sets, the more robust the model is. However, an even more rigorous test is to check how well the model performs on totally new data that was not used in the model estimation.

For example, if the model is to be used in a marketing campaign to identify which customers are most likely to respond to a discount offer, often the model’s performance is also tested to analyze how well it would have performed on historical campaign data.

Often a model is also tested on a new campaign to see how well it performs in a real environment. Appropriate “control groups” are defined, so the response to the “modeled” group can be compared to the response using other methods. There are extensions to and variations on the train-and-test theme.

For example, a random splitting of a sample into training and test sub-sets could be fortuitous, especially when working with small data sets, so we sometimes conduct statistical experiments by executing a number of random splits and averaging performance indices from the resulting test sets.

The following methods are common practice to assess the accuracy for a “classification” type model. Classification is a predictive type model.

The most frequent metric to assess model accuracy is Percent Correct Classification. This is a simple metric that compares the number of observations classified correctly to the overall number of observations.

The Confusion Matrix is also commonly used for classification and provides a summary of the different kinds of errors, called Type I and Type II errors.

This is a much preferred method of understanding model performance. If the requirement is to treat a sample of the population, then the approach taken is to rank order the population by model score and select only a portion of those entities in the selected group with the highest scores. Metrics such as Lift, Gain, ROC, and Area Under the Curve (AUC) are used.

SAP have developed metrics called the Predictive Power (KI) and the Prediction Confidence (KR). And these will be described in detail later in this course.

The robustness of a predictive model is very important. One of the popular ways of analyzing model robustness is to compare the performance of the model on a hold-out sample.

If the performance is similar on these sub-samples, then the model is deemed to be “robust”. Different algorithms can use different metrics to measure accuracy.

These accuracy and robustness measures will be discussed in much more detail later in this course. The following methods are common practice to assess a cluster model – which is a descriptive type of model.

The cluster model groups observations into clusters. This is a “labeling” that assigns each observation to a cluster.

External criteria are used to measure the extent to which the cluster model’s labels match pre-defined labels. This analysis compares how well the cluster model assigns observations according to some pre-defined groupings.
This is often achieved through correlation analysis between the model clusters and the pre-defined clusters. Internal criteria are used to measure the goodness of clustering structure without respect to any pre-defined groupings.

There is a wide range of techniques used, for example: Cluster cohesion, which measures how closely the observations in a cluster are related.

This uses the metric Sum of Squared Error (SSE), which is the sum of the squared differences between each observation and its group's mean.

It can be used as a measure of variation within a cluster. It is also a useful metric to estimate the number of clusters.

Cluster separation measures how distinct or well separated a cluster is from other clusters. This uses the metric Between Cluster Sum of Squares (BSS).

The silhouette coefficient. This combines ideas of both cohesion and separation, but for individual points as well as clusters.

Relative criteria are used to compare different cluster models. Often an external or internal index is used for this function, for example the Sum of Squared Errors (SSE) can be used.

Different algorithms can use different metrics to measure performance. These examples here are for cluster models.

These accuracy and robustness measures will be discussed in more detail later in this course. Thanks for joining.

See you in the next unit, where we will look at the data understanding phase in detail. See you there.
Hi, nice to see you again. In this unit, we will take a look at the data understanding phase.

So let's start. Phase 2 of CRISP is the data understanding phase.

So it starts with an initial data collection and proceeds with activities to get familiar with the data, to identify data quality problems, to discover first insights into the data, and to detect interesting subsets to form hypotheses for hidden information.

Task 1 is to collect initial data. We acquire the data listed in the project resources.

Then we load data if necessary to enhance our data understanding. This might lead to initial data preparation steps.

If we are acquiring multiple data sources, integration is an additional issue that can be tackled either here or in the later data preparation phase.

Then we describe the data – examine the “gross” or “surface” properties of the acquired data and report on the results. We need to explore the data – this tackles the data questions that can be addressed using querying, visualization, and reporting,

including analysis of the distributions of key attributes, results of simple aggregations, relations between pairs or small numbers of attributes,

properties of significant sub-populations, simple statistical analyses.

This may contribute to or refine the data description and quality reports, and it may feed into the transformation and other data preparation that's needed.

Next we verify data quality, addressing questions such as: “Is the data complete?”, “Are there missing values in the data?” Here, we acquire the data – or access to the data – listed in the project resources.

This initial collection includes data loading into the data exploration tool and data integration if multiple data sources are acquired. We write the Initial Data Collection Report that lists the dataset (or datasets) acquired,

the dataset locations, the methods used to acquire the datasets, and any problems encountered. We also record the problems encountered and any solutions achieved to aid with future replication of this project

or with the execution of similar future projects. This task examines the “gross” or “surface” properties of the acquired data and reports on the results.

We create a Data Description Report that describes the data which has been acquired, including the format of the data, the quantity of data (for example, the number of records and fields in each table),

the identities of the fields, and any other surface features of the data which have been discovered. This task tackles the data mining questions, which can be addressed using querying, visualization, and reporting.

These include distributions of key attributes, for example the target attribute of a prediction task, relations between pairs or small numbers of attributes, results of simple aggregations, properties of significant sub-populations, and simple statistical analyses. These analyses may address directly the data science project goals.

They may also contribute to or refine the data description and quality reports and feed into the transformation and other data preparation needed for further analyses.

More details of this type of analysis are given in the next unit. The Data Exploration Report describes the results of this task,

including our first findings or initial hypotheses and their impact on the remainder of the project. It can also include graphs and plots,

which indicate data characteristics or lead to interesting data subsets for further examination. This task examines the quality of the data, addressing questions such as:

Is the data complete? Does it cover all the cases required? Is it correct or does it contain errors? And if there are errors, how common are they?
Are there missing values in the data? If so, how are they represented, where do they occur, and how common are they? The Data Quality Report lists the results of the data quality verification.

If quality problems exist, it lists possible solutions, and it lists solutions to data quality problems in general. So now we get close to the end of this week.

The next unit will be the last one of Week 1, where we will look at initial data analysis and exploratory data analysis.
Hi, and welcome back. In the last unit of the first week we will now take a look at initial data analysis and exploratory data analysis.

It is usually wise to begin any analysis with an exploratory examination of the given data in order to get a feel for it. The general aim is to clarify the general structure of the data and to obtain simple descriptive summaries.

This may help to suggest a suitable model, which will in turn suggest an appropriate algorithm. This initial analysis includes processing the data into a suitable form for analysis, and checking data quality.

Are there errors, outliers, missing values, or other peculiarities? Does the data need to be modified in any way?

This first phase of the analysis is called the initial data analysis (IDA), and it was suggested by Chatfield in 1985. He defines the various steps in IDA.

It includes analysis of the structure of the data, the quality of the data for errors, outliers, and missing values. It looks at observations with descriptive statistics, graphs, modifying the data, adjusting extreme observations,
estimating missing observations, transforming variables, binning data, and forming new variables. IDA has many things in common with exploratory data analysis (EDA).

EDA is an approach to analyzing data for the purpose of formulating hypotheses that are worth testing. We often use data visualization techniques.

Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data and possibly formulate hypotheses that could lead to new data collection and experiments.

Tukey wrote the book "Exploratory Data Analysis" in 1977. The objectives of EDA are to suggest hypotheses about the causes of observed phenomena,

assess assumptions on which the analysis and statistical inference will be based, support the selection of appropriate statistical tools and techniques,

and provide a basis for further data collection through surveys or experiments. EDA is different from initial data analysis (IDA),

which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed.

However, in many reference books, EDA now seems to encompass IDA. I'm now going to show you a demonstration using SAP Predictive Analytics expert system.

We will use some US stores' retail data to walk you through this topic. The dataset contains the following variables: store location, turnover, margin, staff, and store size.

So this is the interface to the Predictive Analytics expert system, and as the first step I need to acquire the data.

So here I am adding a new dataset. This is the data, and I'm pointing to a CSV file here, which has got this data.

And I create it... In the visualization layer here, the first thing that I might want to do is to start to look at the data that I've actually acquired.

And to do that, I can create a table. So here I can include these measures,

"Margin", "Size", "Staff", and "Turnover", and I can then look at that information for each of the different stores.

And that produces this table. So for each different store, I can clearly see now what the margin, size, staff, and turnover are.

Now that I've got that data, what I might want to do is to start to visualize what it actually looks like. And so one of the first things I might want to do is to look at a scatter plot.
So this is a scatter plot of margin versus size for each of the different stores. The stores are shown in different colors here.

And you can see quite clearly here that there might actually be two separate groups based on the store, looking at the size and the margin. This might mean that when I come to do some predictive analysis on this data,

that in fact instead of building one model across the whole of the dataset, I might want to split this data initially and develop models on each of the different groups.

Then I might want to do a scatter matrix looking at all of the different variables. So this is for each of the different four numeric variables,

we can see the cross-wise plots for each variable. And again, this reinforces my hypothesis that there could actually be two different groups of data here.

Next, what I might want to do is produce a bubble chart. Now the bubble chart allows me to plot four dimensions of the data,

so here I'm looking at size and margin for each of the different stores, and also the diameter of the bubble indicates the staff size.

So this is giving me a lot of information about the data itself. Then I might want to look at a parallel coordinates chart.

This plots each of the different variables on a vertical axis, and I can see here that there are groups of stores which are bunching together.

And again, this reinforces my initial hypothesis that there could actually be two different groups. But there might be some outliers as well,

and one way to assess the outliers is to use something that is called a "box plot". We'll be returning to box plots in more detail in this course.

However, you can see that there are outliers, which are represented by these circles here for these variables, and this is a very easy-to-understand visualization of how we can analyze the data and identify outliers.

With this I'd like to close the first week. I hope you enjoyed the first units of this course, and I'm happy to get in touch with you in our Discussion Forum if you have any content-related questions. Now I wish you all the best for the weekly assignment,

and see you next week, where we will cover the topic of "data preparation".