Course Guide

Introduction to IBM SPSS Modeler and Data Science (v18.1.1)

Course code 0A008 ERC 1.0

IBM Training

Course overview

Preface overview

This course provides the fundamentals of using IBM SPSS Modeler and introduces the participant to data science. The principles and practice of data science are illustrated using the CRISP-DM methodology. The course provides training in the basics of how to import, explore, and prepare data with IBM SPSS Modeler v18.1.1, and introduces the student to modeling.

Intended audience

This course is recommended for:

- Business analysts
- Data scientists
- Clients who are new to IBM SPSS Modeler or want to find out more about using it

Topics covered

Topics covered in this course include:

- Introduction to data science
- Introduction to IBM SPSS Modeler
- Introduction to data science using IBM SPSS Modeler
- Collecting initial data
- Understanding the data
- Setting the unit of analysis
- Integrating data
- Deriving and reclassifying fields
- Identifying relationships
- Introduction to modeling

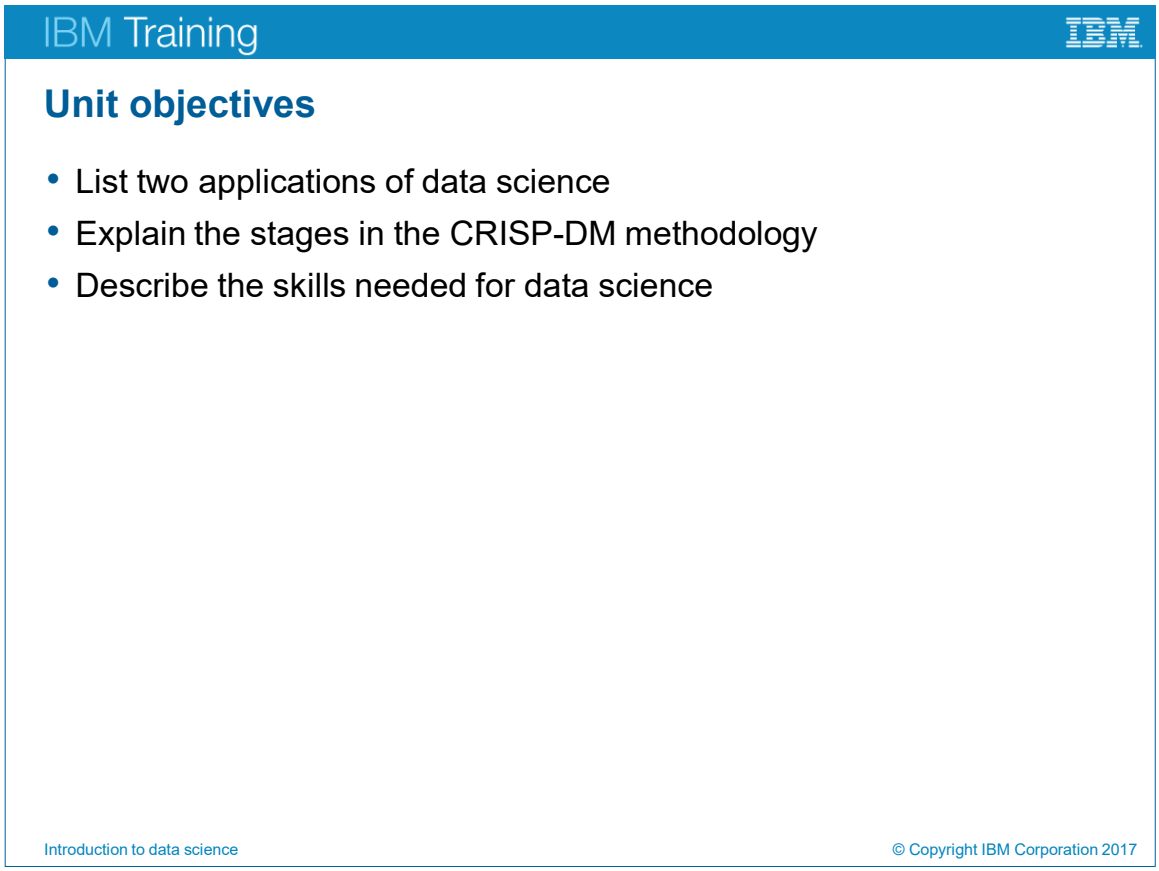
Course prerequisites

- It is recommended that you have an understanding of your business data

IBM Training

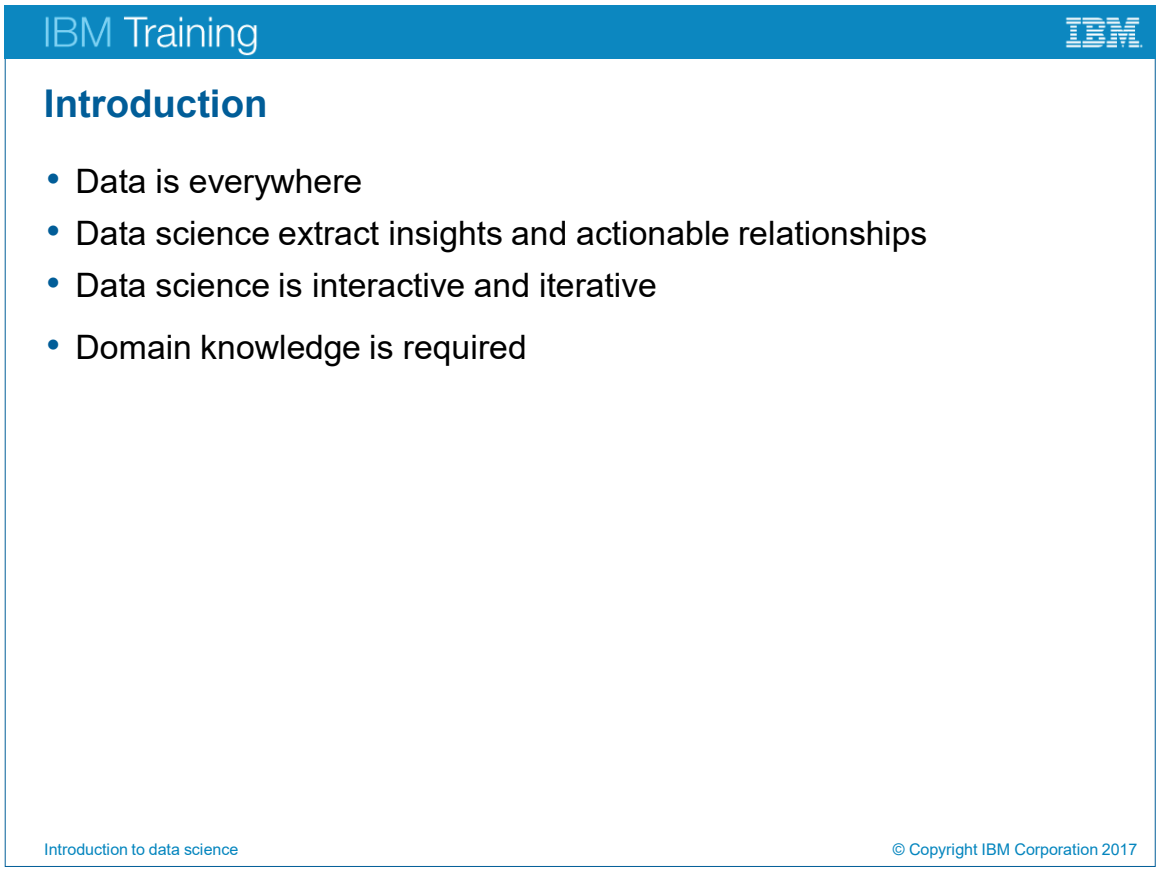
Introduction to data science

IBM SPSS Modeler (v18.1.1)



Unit objectives

No prior knowledge is required for this unit.

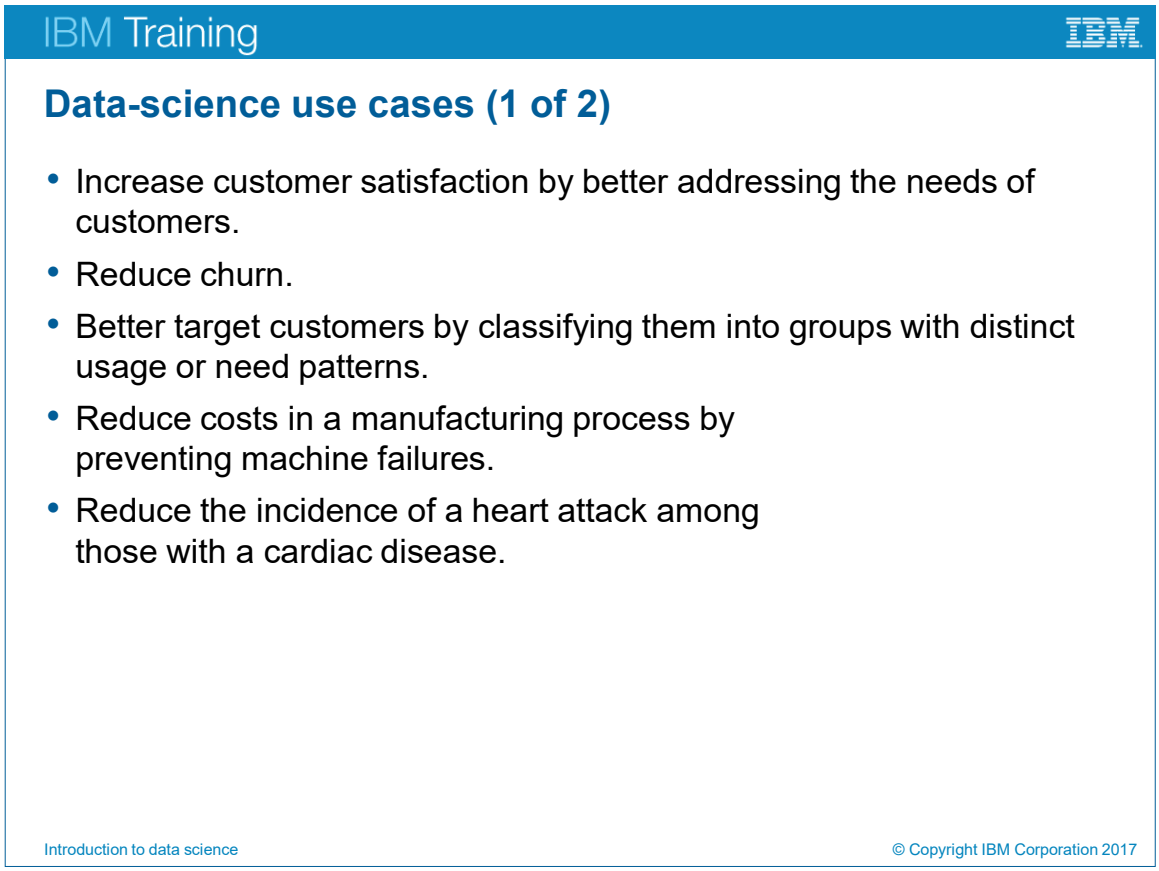


Introduction

With increasingly competitive markets and the vast capabilities of computers, many businesses find themselves faced with data and a need to identify useful patterns and actionable relationships.

Data science is an interdisciplinary field that combines machine learning, statistics, advanced analysis, and programming. It is a new form of art that draws out hidden insights and puts data to work in the cognitive era.

A common misconception is that data science involves passing huge amounts of data through intelligent technologies that alone find patterns and give magical solutions to business problems. Data science is an interactive and iterative process. Technology must be used jointly with business expertise to identify underlying relationships and features in the data.



Data-science use cases

As an example of a data-science project, think of a telecommunications firm that is confronted with churn, customers who cancel their policies, subscriptions, or accounts. The firm can use its historical data to build models to identify groups of customers with high churn rates. The firm can then apply these models to the current customer database to identify customers at risk. These customers can be made an interesting offer, so they can hopefully be retained.

A common application in many areas is segmentation, also referred to as profiling. For example, a telecommunications firm can segment its customers into distinct groups such as "leaders" and "followers" by analyzing service usage data (phone calls, text messages, internet usage). Each group can then be approached in its own way.

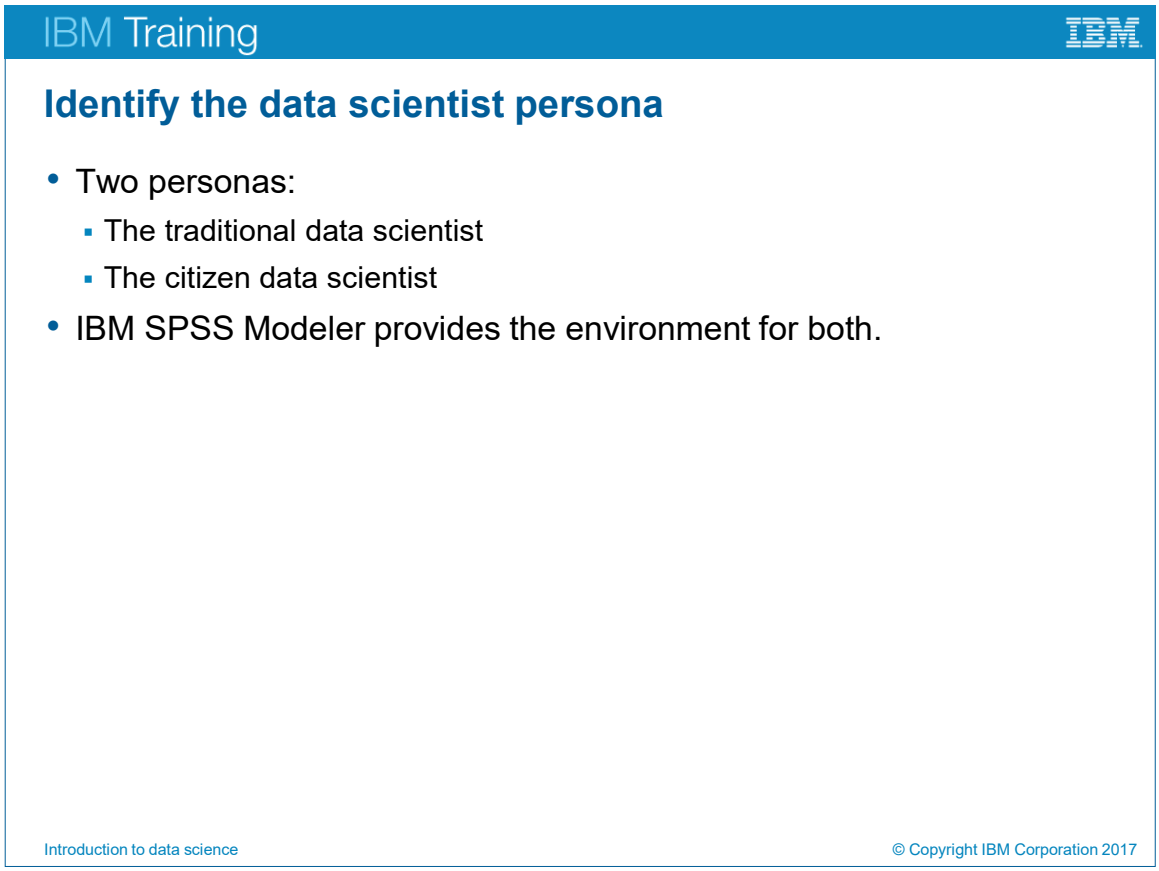
IBM Training

Data-science use cases (2 of 2)

- Reduce costs by better targeting customers in direct mail campaigns.
- Reduce costs by preventing fraudulent credit-card activity, or detecting it in an earlier stage.
- Increase revenues by increasing the number of products sold by up- or cross-selling.
- Increase revenues by showing a visitor the best-next- page on a website.

Another use case is found in database marketing, where huge volumes of mail are sent out to customers or prospects. Typically, response rates lie around 2%. To cut costs in sending out mail, the database marketing department can use historical data to build models that identify groups with high response rates, so that only these customers will be approached in future campaigns. This will cut mailing costs, while the number of responders (people purchasing the product) will not change significantly. All in all, costs will go down and revenues will stay the same, so the ROI (Return On Investment) will improve.

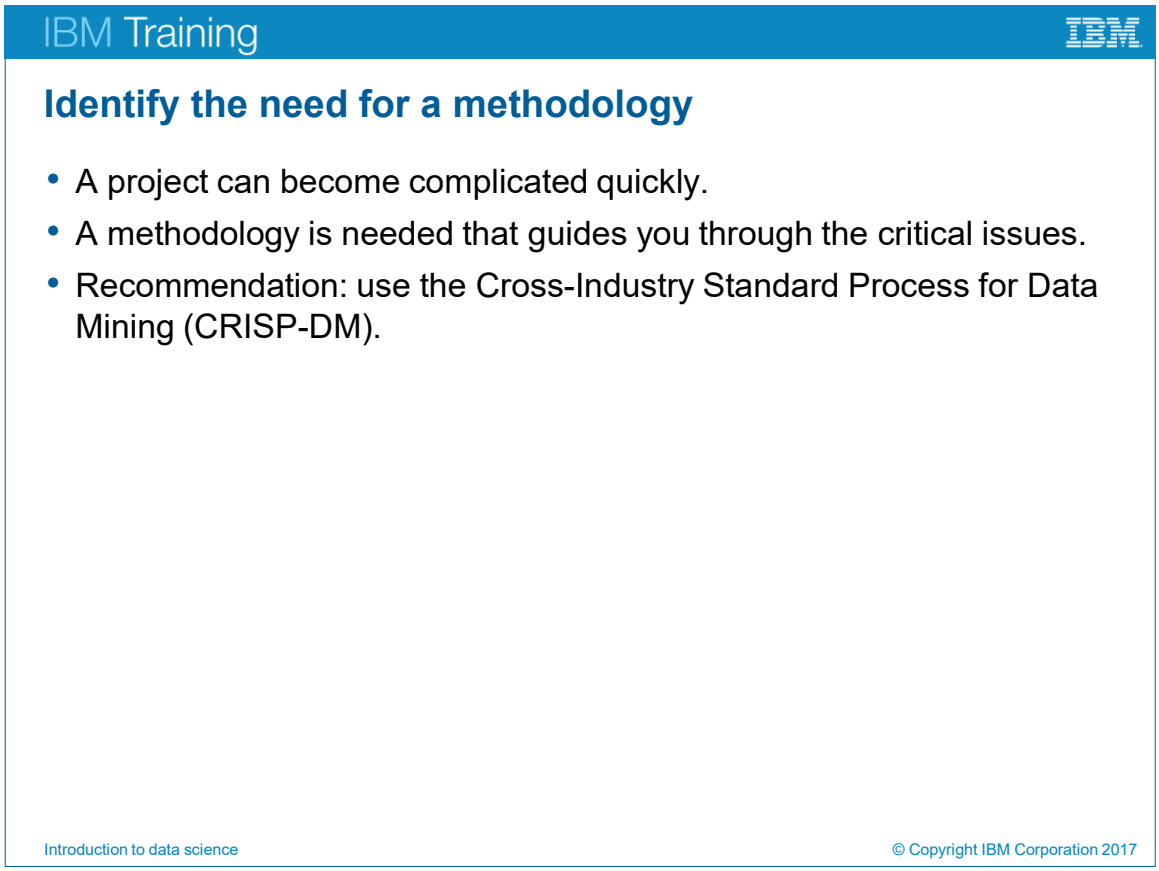
For more applications, refer to a site that hosts data science competitions, such as Kaggle.



Identify the data scientist persona

The requirements for a data scientist are rigorous. Ideally, the data scientist should have a background in IT, domain knowledge, experience with open source tools such as Python, Scala, R, and Spark, profound knowledge of statistics and machine learning, and excellent communication skills, among others. For this persona, IBM provides a single workspace with IBM Data Science Experience. In addition to that, the data scientist will most likely use a stand-alone tool such as IBM SPSS Modeler, especially if he needs to connect to Big Data using IBM SPSS Analytic Server, or when using IBM SPSS Collaboration and Deployment Services to manage the work or to deploy models.

More recently, the requirements for a data scientist were relaxed, since tools and technology have advanced to a place where a business analyst can now perform analytic tasks that traditionally have required the expertise of a data scientist. This new role is commonly referred to as citizen data scientist. This persona is also referred to as business analyst. IBM SPSS Modeler provides the workspace for this persona too, where data science capabilities can leverage domain knowledge.



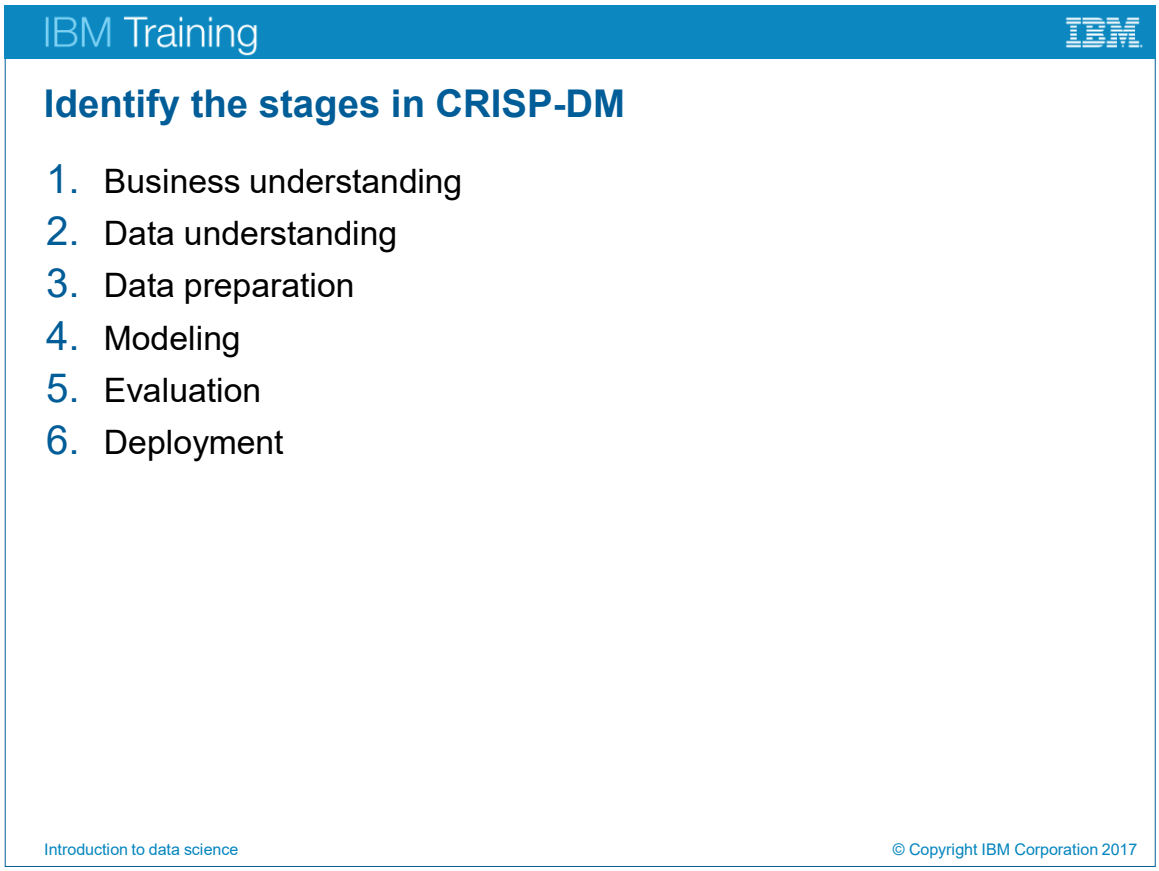
Identify the need for a methodology

A data-science project can become complicated very quickly. There is a lot to keep track of: complex business problems, multiple data sources, varying data quality across data sources, an array of modeling techniques, different ways of measuring success, and so on.

To stay on track, it helps to have an explicitly defined methodology in place. The methodology guides you through the critical issues and makes sure that the important points are addressed. It serves as a road map so that you will not lose your way as you dig into the complexities of the data.

There are multiple methodologies that you can adhere to. The most frequently used in analytics, data mining, and data science is the Cross-Industry Standard Process for Data Mining (CRISP-DM), conceived in 1996. In 2015, IBM Corporation released a new methodology called Analytics Solutions Unified Method for Data Mining/Predictive Analytics (also known as ASUM-DM), which refines and extends CRISP-DM. More recently, the Standard Methodology for Analytical Models (SMAM) has been suggested as yet another alternative for CRISP-DM.

CRISP-DM is the starting point for the other methodologies, and is designed as a general methodology that can be applied to a wide variety of industries and business problems.



Identify the stages in CRISP-DM

CRISP-DM includes six stages, described in more detail on the next slides.

A comprehensive discussion of CRISP-DM is beyond the scope of this course. Please refer to IBM SPSS Modeler's Help for an overview of all tasks and sub tasks, or refer online for that and more information about ASUM-DM and SMAM, which both augment CRISP-DM.

IBM Training

Explore stage 1: Business understanding

Task	Sub task 1	Sub task 2	Sub task 3
Determine business objectives	Background	Business objectives	Business success criteria
Assess situation	Inventory of resources	Risks and contingencies	Terminology
Determine modeling objectives	Modeling success criteria		
Produce project plan	Write a project plan	Initial assessment of tools and techniques	

Introduction to data science

Explore stage 1: Business understanding

Business Understanding is perhaps the most important phase in a project. Business objectives and success criteria, resources, constraints, assumptions, risks, costs, and benefits are identified in this stage. Also, specific data-science goals are set, a project plan is written, and agreed upon.

Think of a telecommunications firm that is confronted with high volumes of churn (customers cancelling their subscription). The firm could start a project with the objective to reduce churn, and the project could be declared successful if churn is reduced by at least 10%.

After having formulated the business question, the situation needs to be assessed. The required resources need to be listed and you have to ensure that they will be made available. For example, ensure that you have the appropriate software tools and staff to operate them. Also, check whether the project involves colleagues from other departments or requires external consultants, and whether they are available.

A critical issue is the availability of data. Important questions are:

- Who are the key persons in accessing the data?
- Will the data be enriched by purchasing demographic data? And if so, how will you measure the added value?
- Are there legal restrictions on the use of data?

When you work together in a project, ensure that the terminology is clear to everyone. For example, define what "churn" is. A customer whose subscription has ended because he did not pay the bill, would you regard that as churn? You might want to distinguish between voluntary and involuntary churn.

Translating the business objectives into specific modeling goals is another task in this stage. A modeling goal derived from a business objective such as "reduce churn" could be to have a model in place that returns the likelihood that a customer will cancel his subscription. Ask yourself if you want to apply the model to every customer, or to high-value customers only; preventing churn for low-value customers may cost the company more than letting them go.

You may decide to use only a certain class of models. For example, the modeling goal could be to understand the reasons why customers cancel their subscription. This means that you would only consider models that provide insight and you would discard black-box models (refer to the *Introduction to Modeling* unit in this course for an example of both types of models). The type of model you choose will also affect actions to be taken later. For example, when the model tells you that customers with a certain handset show high churn rates, you may offer a new handset to customers with that handset. If you use a black-box model instead, no such specific offer can be made, only a more generic one, such as giving a discount.

Ensure that there is a project plan that lists the tasks and responsibilities. The project plan may be written along the lines of the six stages in the CRISP-DM methodology. What is the time needed to complete each stage? And, which persons are responsible in each stage? What are the risks in each stage, and is there a contingency plan?

IBM Training

Explore stage 2: Data understanding

Task	Sub task 1
Collect initial data	Data-collection report
Describe data	Data-description report
Explore data	Data-exploration report
Verify data quality	Data-quality report

Introduction to data science

Explore stage 2: Data understanding

Data provides the raw materials of data science. The stage of Data Understanding addresses the need to understand what the data resources are and the characteristics of those resources. It includes collecting initial data, describing data, exploring data, and establishing the quality of the data.

For example, a telecommunications firm that wants to reduce churn could use the information they have on their customers, such as gender, age, and region. Another data source might be comprised of call detail records. Also, the firm could use data from the call center. Although the latter type of data is text and needs special tools and skills, it could well be related to churn, especially if customers call in with a complaint. Visits to the web site could also be tracked, as well as activities on social media. All these data sources together provide a 360 degrees view on the customer's behavior.

In a data-collection report, describe each dataset in terms of the number of rows and columns. Also, note the number of records per customer. A dataset that stores information about customers, such as gender and age, will have one record per customer, whereas a dataset that has one record for each purchase that the customer made will have multiple records for a customer. If you need a dataset with one record per customer to answer the business question, think about how you can bring a dataset with multiple records per customer back to a dataset with one record per customer, without losing valuable information.

In a data-description report, explain what the fields and values in the various datasets represent. For example, it may not be clear to everyone in the project what a field such as HS means. Consider renaming such fields, for example, rename HS into HANDSET. Also, list the abbreviations that appear in the data, and what they mean. It may be obvious what values F and M represent for a field such as GENDER, but if you have values 1, 2, 3 for a field named SEGMENT, explain what each code means to better interpret the modeling results later.

In the data-description report, also map field names in different datasets to each other. In databases, for example, a field such as KEY_ID in the customer table could be named CUSTOMER_ID in the product table.

In the data-exploration report, report inconsistencies or errors in the data. For example, a field GENDER might have values F, FEMALE, Female, or a field such as AGE might show a value of -1. These instances should be reported and you should investigate why these values appear in the data. For example, it could be that the database administrator plugs in a value -1 when a value is unknown.

In the data-quality report, report the amount of missing data and decide what to do with records or fields that have many missing values. Again, investigate the reason why data is missing. For example, values for END_DATE_OF_SUBSCRIPTION are missing by definition for all current customers. You cannot discard this field, because you will need it in order to derive a field such as HAS_CHURNED from it later.

Two units in this course relate to this stage:

- *Collecting initial data* shows how you can import data into IBM SPSS Modeler.
- *Understanding the data* introduces you to methods to explore the data, to establish its quality, and to take action upon it.

IBM Training

Explore stage 3: Data preparation

Task	Sub task 1	Sub task 2
Select data	Rationale for inclusion and exclusion	
Clean data	Data-cleaning report	
Construct data	Derived attributes	
Format data and combine datasets	Set the unit of analysis	Integrate data

Introduction to data science

Explore stage 3: Data preparation

After cataloging the data resources data needs to be prepared for modeling. This includes selecting, cleaning, constructing, formatting and integrating data. These tasks can be very time consuming but are critical for the success of the project.

No model will be able to compensate for large amounts of error in the data. In the worst case a potentially good set of predictors may fail because of error that masks their effect. Take the time to thoroughly prepare and clean the data and continue to check the data as it is modified during the analysis.

Constructing the dataset for modeling requires considerable effort and thought. Assume, in the example of a telecommunications firm addressing churn, that campaigns are run monthly, targeted at customers who are likely to cancel their subscription. Assume that the marketing department needs to have a list of customers at risk on the first day of March. The dataset used for model building must reflect that design. The dataset used for model building has be created at the end of January, and that dataset has to be enriched with churn information for the month February. You can think of it as freezing the end-of-January data and adding a field to it that flags whether the customer left in February. When a model is built at the end of February (using the frozen January data and churn data collected in February), that model can be used to predict who are likely to churn in the next month.

In building the dataset for modeling, however, it is not uncommon that that dataset stores information of the latest month rather than of the moment the data should be frozen. Continuing with the example, if NUMBER OF PRODUCTS PURCHASED is thought of as being a predictor for churn, it should be the number recorded at the end of January, not at the end of February.

In this stage, think about whether all records and fields are needed to answer the business question. Again, think of a telecommunications firm that has initiated a project to reduce churn. Fields such as GENDER and AGE are valuable because they hopefully can predict a field named CHURN (Yes for customers who cancelled their subscription, No for current customers). There may also be fields in the dataset that are derived from the CHURN field itself. For example, those who cancelled their subscription may have received a letter of thanks, and so a field such as HAS_RECEIVED_LETTER might be included in the dataset. If this field would be used to predict CHURN, rules such as the following could be found:

If HAS_RECEIVED_LETTER = True, then CHURN = True
If HAS_RECEIVED_LETTER = False, then CHURN = False

Thus, HAS_RECEIVED_LETTER will predict CHURN perfectly. However, HAS_RECEIVED_LETTER should not have been used as predictor, because it is a consequence of ending the subscription.

This example may sound trivial, but in a database with, for example, 34 tables and 481 fields, it may not be that easy to distinguish between fields that are relevant predictors and fields that are a consequence of the field that needs to be predicted. It is in this stage of data preparation that relevant fields must be selected and irrelevant ones must be identified.

More in general, the research design is important. A carefully formulated study will consider whether there is a cause-and-effect relationship between the predictors and target field. For example, customer satisfaction research often uses attitudes about product or service to predict overall satisfaction, willingness to buy again, or willingness to recommend a product or service. In terms of cause and effect, all these attitudes and satisfaction occur at one point in time, that is, when the survey is conducted. It can then be argued that while these attitudes may be correlated, claiming that one attitude causes another is not necessarily correct; instead, the attitudes may be mutually reinforcing. When this is true, the predictions from a model about how changes in attitudes affect the target field may be invalid. The basic point is that for true predictions, you must be certain that predictors in a model occur before the target field.

Apart from discarding fields, records could be removed. For example, if you are only interested in preventing churn of high-value customers, low-value customers should not be included in the analysis and thus they should be removed from the dataset.

In the data-cleaning report, list all the actions that were taken to cleanse the data. For example, you may have reclassified F, FEMALE, and Female into a single Female category. Or, maybe missing data were replaced.

New fields will be derived in this stage. A field such as DATE_OF_BIRTH needs to be transformed into AGE, in order to interpret the modeling results later. Also, consider deriving fields by taking differences or ratios. Think of two customers of a telecommunications company. One customer might have phoned for 5 minutes and sent 10 text messages, whereas a second has phoned for 50 minutes and has sent 100 text messages. Although different in absolute value, their patterns of phoning and texting are the same, and it might be that the pattern is related to churn. In general, based on domain knowledge, new fields are derived in order to include them in model building later. In data science, this is referred to as feature engineering. Most likely, you will gain much more by feature engineering than by trying different models. This emphasizes the importance of domain knowledge.

In formatting data, think of restructuring data into a form that the analysis requires. A telecommunications firm might have a dataset of call detail records, where every record represents a call. So, if one customer made one call, and another customer made 15, the first customer will have one record in the dataset, and the second customer 15. Now suppose that another dataset stores customers, with their gender, age, and a field that flags whether the customer cancelled his subscription. When building a model for churn, you should count a customer's gender, age, and churn only once, and not as many times as he has call detail records. Thus, each record should represent a unique customer and the call detail dataset should be transformed into a dataset where you have one record per customer.

The entity that you want to build models for is referred to as unit of analysis. When datasets do not have the required unit of analysis, they should be transformed so they will have the required unit of analysis. When the unit of analysis is correct for all datasets, they can be integrated, that is, combined into a single dataset.

Three units in this course relate to this stage in a project:

- *Setting the unit of analysis* presents various ways to transform a dataset with the incorrect unit of analysis into a dataset with the required unit of analysis.
- *Integrating data* presents methods to combine datasets.
- *Deriving and reclassifying fields* focuses on two methods to cleanse and enrich data.

IBM Training

Explore stage 4: Modeling

Task	Sub task 1	Sub task 2
Select modeling techniques	Modeling assumptions	
Generate test design	Test design	
Build model	Set model parameters	Model descriptions
Assess model	Model assessment	Revise model parameters

Introduction to data science

Explore stage 4: Modeling

Modeling is the part of data science where sophisticated analysis methods are used to extract information from the data. This stage involves selecting modeling techniques, generating test designs, and building and assessing models.

In the modeling stage you will probably try multiple models. To evaluate them, it is common practice to apply the models to a testing dataset, a dataset that was not used for model building. To do so, the entire dataset needs to be partitioned into a training set (on which models are built) and a testing set (on which models are tested). It is common to use a 70/30 split for training-testing, but this may leave too few records to build the model. In that case, boosting methods, such as duplicating records, could be tried. This is not standard in traditional statistical modeling, but it is widely accepted in data science.

When building the model you can start with a model's default settings and then fine tune the parameters. Report which parameters you have changed, and how that affected the results, not only in terms of accuracy, but also in execution time. For example, model A may complete within seconds, whereas model B may take hours. Would you then prefer model A, or will you run model B in batch, at night? And if you choose the latter, what are the consequences, for example do you need a colleague to schedule the job for you?

You should also describe the model results in terms of how the model deals with missing data. Some models do not have issues with missing data, whereas other models delete records with missing data. Decide on which model you prefer. If you use a model that discards records with missing data, how does that affect deploying the model later? If the dataset to which the model needs to be applied has a high percentage of missing data, the model cannot be applied to a significant part of that data.

Given the results, rank the models according to criteria such as model accuracy, ease of use, interpretation of the results, and ease of deployment. You may also want to rerun models, adjusting model parameters. Or you could try a model that was not thought of before. For example, you did not run black-box models but given the disappointing results so far you may want to give such models a try. The question then is how this relates to the data-science goals, and if those should be revised.

Two units in this course relate to this stage of a project:

- *Identifying relationships* is a first step in exploring relationships in the data.
- *Introduction to modeling* unit provides an overview of models and presents two of them in more detail.

IBM Training

Explore stage 5: Evaluation

Task	Sub task 1	Sub task 2
Evaluate results	Assessment of data-science results with respect to business success criteria	Approve models
Review process	Review of process	
Determine next steps	List of possible actions	Decision

Introduction to data science

Explore stage 5: Evaluation

In this stage you have built one or more models that appear to have high quality from a data analysis perspective. You now should evaluate how the results can help to achieve the business objectives.

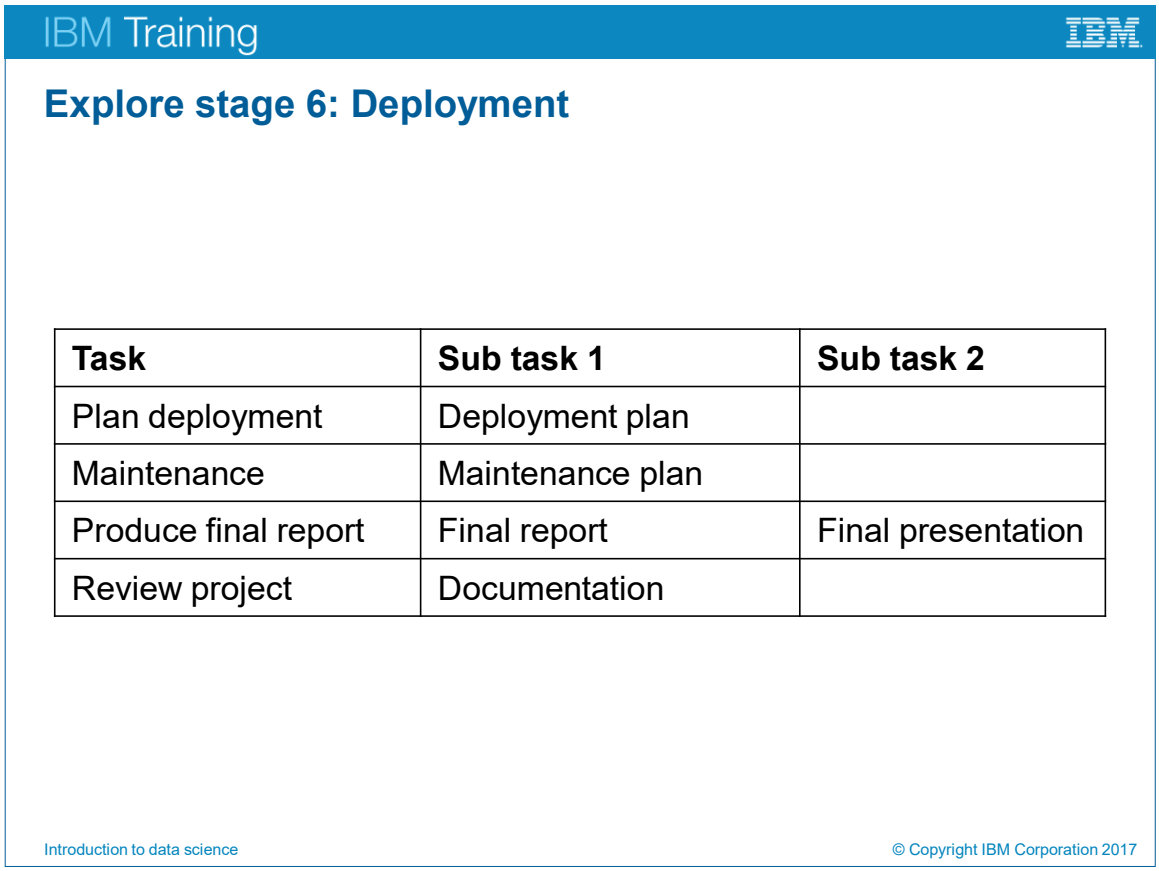
Continuing with the example of the telecommunications firm, suppose that two candidate models were found. The first model tells you that adolescent men with handset A show the highest churn percentage, and a second model returns the likelihood to churn, but does not give further insight. Suppose that the first model has a lower accuracy than the second model. When you apply the first model to the current customers, you can target adolescent men with handset A, and make them an interesting offer for a new handset. When you apply the second model to the current customers, you will use a general strategy, such as offering a discount for those at risk. This is the moment to choose one of the models, or maybe to go ahead with both.

Also, review the project so far. For example, was the project plan, with the tasks, responsibilities and deadlines for each stage met? If not, what were the reasons for the delay?

Finally, determine the next steps. In a worst-case scenario you may have to conclude that the results are unsatisfactory, and you need to circle back to an earlier stage. For example, it might be that the accuracy of the model is too low, and thus you might consider bringing in text data from the call center (data that was not used yet in the project). If you decide to use text data, you will probably need specialized software (such as Text Analytics, available in IBM SPSS Modeler Premium) and you will need someone who has the skills, including business knowledge, to run the analyses.

All in all, you can iterate through the previous stages, and come to a point where you are confident enough to deploy one or more models.

Refer to the *Introduction to modeling* unit for more information about how to compare models in terms of accuracy.



Explore stage 6: Deployment

Now that you have invested all of this effort, it is time to reap the benefits. Depending on the requirements, the deployment stage can be as simple as generating a report or as complex as implementing a repeatable process.

Deployment is a critical issue in the entire project. A Health Maintenance Organization (HMO) investigated ways to reduce costs by looking at patterns of treatment and care, and found that there was an optimal length of stay in the hospital for several types of major surgeries. While not requiring doctors to rigidly follow the statistical results (which would be inappropriate for any specific patient), the HMO encouraged doctors to take this information into account. But after a few months it was clear that length of stay decisions were not changing. The physicians were sticking to their current practices.

When organizational resistance occurs, the best strategy is usually further education on the potential benefits of the solution, or perhaps, implementation in only a portion of the organization. In the case of this HMO this would mean convincing a few doctors initially to change their release decisions, hoping that eventually more will follow this lead.

Sometimes a model cannot be deployed for factors other than organizational opposition. The most common reason is because factors found to be important are out of the control of the organization, or cannot legally be used in marketing or in making decisions. A consumer products company discovered that certain types of promotions were successful and led to repeat business, but could only offer these promotions to customers it could readily identify, which in practice were those who returned a registration card or bought a service contract. Some obstacles can be anticipated, and the can be adjusted accordingly. If a model can be only partially implemented, as with the consumer products firm, it may still be worthwhile to do the analysis when sufficiently good results would justify the effort (which is always a judgment call).

A plan should be developed that, given the models that will be deployed, lists the actions to take. Continuing with the telecommunications example, suppose that the data scientist generates a list of customers who are likely to leave. Perhaps you want to deploy the model in the call center, so that when a customer at risk calls in, the call center agent can act upon it. Or maybe a customer at risk should get a pop-up with an interesting offer when he visits the web site. Thus, be it a call center or the web, the model needs to be deployed real time and must be fully integrated with other systems that are in place. The deployment plan includes such scenarios.

Another task in this stage is to create a maintenance plan. Eventually the model will expire and you will need to start a new project. In the telecommunications example, suppose that the model tells you that adolescent men with handset A are likely to leave. Now assume that these customers were approached with an interesting offer for a new handset, and that they all accepted the offer. The model will then be the victim of its own success and the model will no longer be applicable. You will then have to return to an earlier stage in the project.

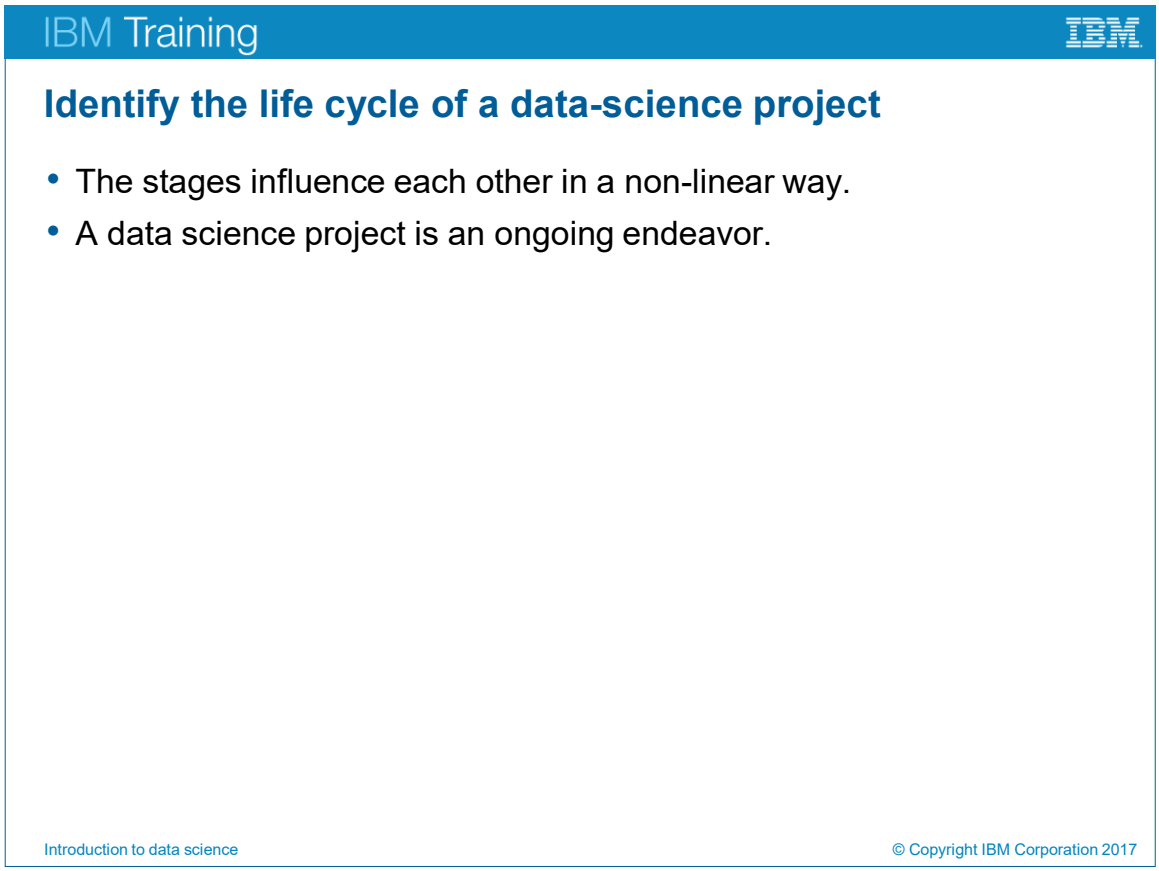
The maintenance plan should also include directives about how a model's success can be monitored. For example, in database marketing it is a common strategy to apply the model to only a part of the customer database. The response rate of this targeted group will then be compared with the response rate of a group of randomly selected customers. Or, when a real-time model is being used to supply sales representatives with offers for customers, both the suggested offer and the customer's decision, among other factors, must be retained in a database for future analysis.

Also, try to assess the costs of making errors. For example, mis-predicting which insurance claims are fraudulent may be expensive because of the effort involved to investigate the claim further. Some data-science tools let you take costs into account when the model is built. Use this feature if it is possible to make even a rough estimate of the costs. When costs cannot be incorporated in the modeling stage, be sure to think carefully about the costs of errors before deployment. If no reliable cost estimates are possible beforehand, try to gather this information after the fact for use in future projects and as ad hoc evaluation criteria.

Deployment of the model is not the end of the project. Even if the purpose of the model is to only increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the organization can use for decision making. Essentially in all projects a final report will need to be produced and distributed.

Finally, review the entire project and document the lessons learned. Report the results in terms of the business objectives. In the example of the telecommunications firm, was churn reduced by 10%? Also, calculate the ROI (Return On Investment) of the project, by estimating the project costs (staff, software) and the extra revenues that you can attribute to the retained customers.

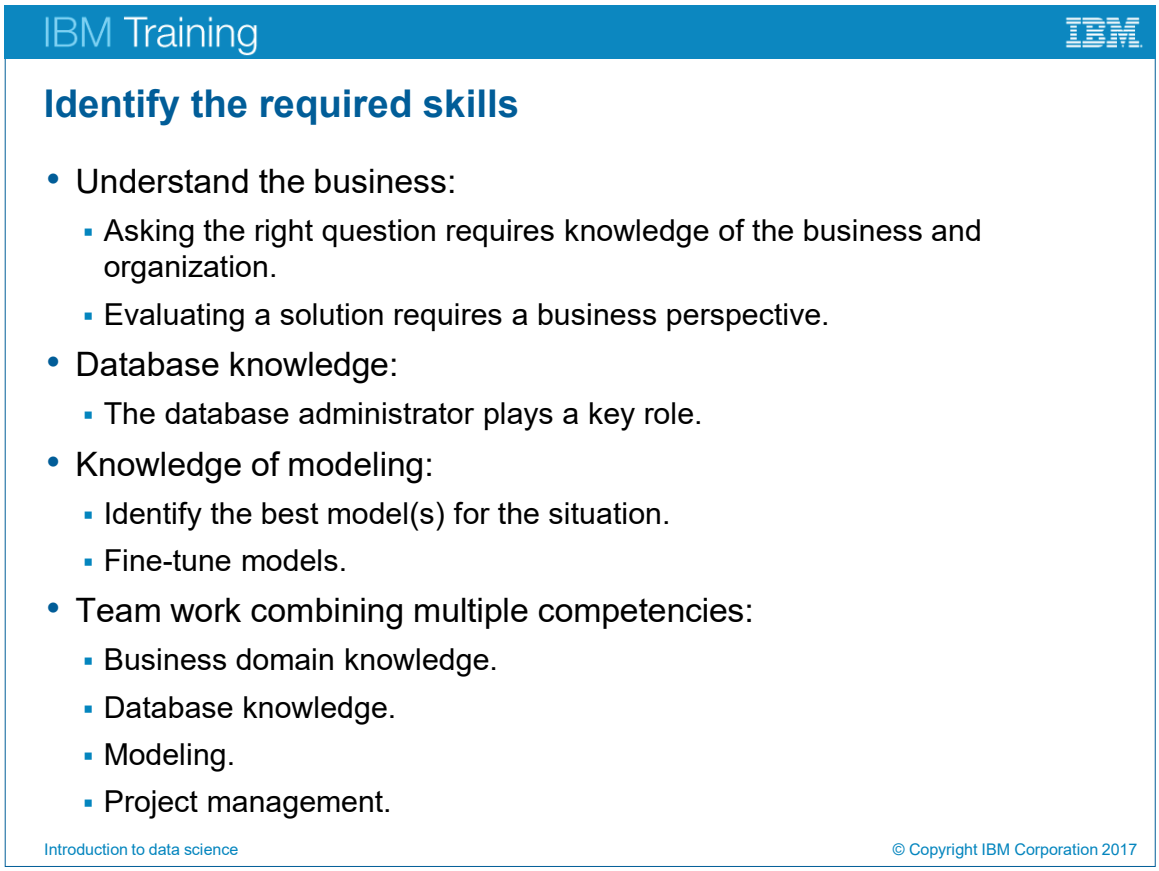
In the *An introduction to data science using IBM SPSS Modeler* unit of this course you will learn how to apply a model to new cases. Real-time deployment is beyond the scope of this course, because it requires additional tools, such as IBM SPSS Collaboration and Deployment Services (C&DS), and it depends on the infrastructure of the organization.



Identify the life cycle of a data-science project

While there is a general tendency for a project to flow through the steps in the order outlined before, there are also a number of places where the stages influence each other.

You will rarely, if ever, simply plan a data-science project, execute it and then pack up the results and go home. Addressing customers' demands is an ongoing iterative endeavor. The knowledge gained from one cycle will almost invariably lead to new questions, new issues, and new opportunities to identify and meet customers' needs. Those new questions, issues, and opportunities can usually be addressed in a new project. This process of identifying new opportunities should become part of the way you think of the business and a cornerstone of the overall business strategy.



Identify the required skills

For a successful data-science project, several disparate skills are required, and they rarely reside in a single individual.

Framing the business question, evaluating the results in terms of business objectives, and presenting the recommendations all require knowledge of the specific business area and organization. Thus someone who knows the critical issues facing the organization is well suited to pose questions that data science might address. He can also evaluate a solution in terms of business objectives and whether it makes sense. Experienced data scientists who focus within an industry could also develop a good knowledge of these issues. Without this component, a project runs the risk of producing a good technical solution to a question unimportant to the business.

A project cannot succeed without good data. Typically, neither the business expert nor the analyst has a sufficiently deep knowledge of the data available on the company's systems to do this. What data tables or files are available? How are they linked? What do the fields really mean? Only someone familiar with the corporate data systems, typically the database administrator (DBA), can answer these and other questions. Therefore, the DBA is usually a key member of the team.

Although data-science tools are available that allow pushbutton ease of running an analysis, as you would expect, knowledge of models is needed. Deciding on the best tools to use for a specific question, knowing how to tweak a technique to its optimum, being able to assess the effects of odd data values or missing data, and recognizing that something does not look right, can all contribute to the success of the project. A data scientist skilled in these techniques is needed. Without this component, you may fail to answer or may incorrectly answer an important question, even with the benefit of good data.

The deployment of a model on new data may be done outside of IBM SPSS Modeler in the database, or you might embed a model in an application in the call center or on the web. Specific skills are needed to implement these types of deployments, and this may call for other team members with programming skills that a data scientist does not possess.

For these reasons, most projects require teams of individuals who contribute differently to the various steps in the project. It would be ideal if all the needed skills were to reside in one person, but this is rarely the case. Occasionally, a team member can serve multiple functions (business and database knowledge, or database and modeling knowledge), but it is relatively rare that all these skills reside in one individual. Of necessity, this confluence of skills in an individual is more likely to occur in small companies and small projects (those that are resource challenged), and that can be limited in the various types of software employed.

IBM Training

Unit summary

- List two applications of data science
- Explain the stages in the CRISP-DM methodology
- Describe the skills needed for data science

Unit summary