Business and data Understanding

Stuart Clarke - Consultant wit sap Service Org
spec Data science, machine learning, AI

Introduction to data science
Introduction to project methodologies
Business Understanding Phase - overview
Defining project success criteria
Data understanding Phase - overview
Initial data analysis and Exploratory
data analysis

Data Preparation

Data preparation Phase - overview
Predictive modeling methodology - overview
Data manipulation
Data encoding
Selecting data - variable and feature selection

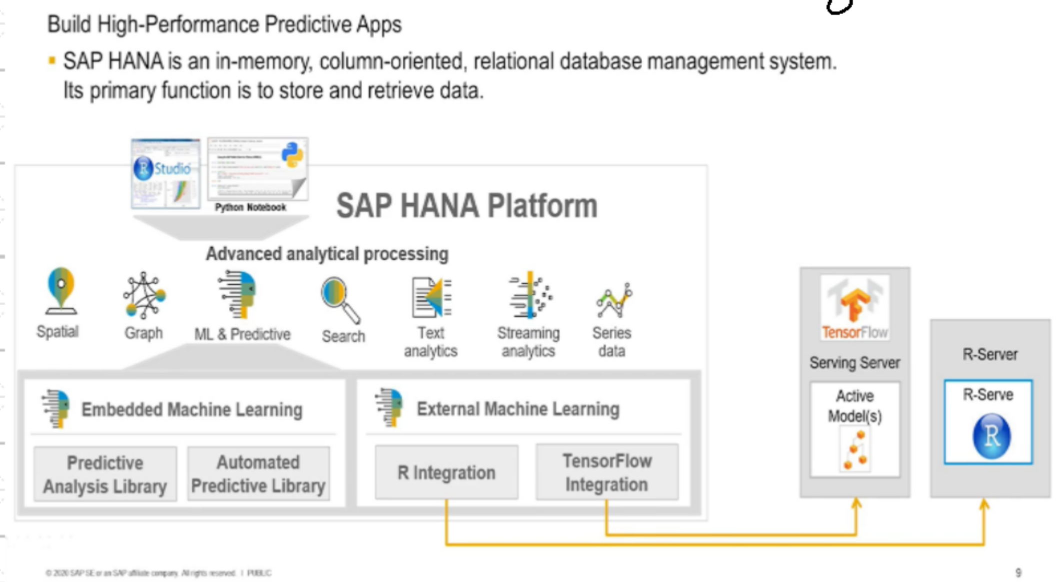
SAP Data intelligence enables across machine learning workloads From data preparation, to data science and model selection, validation and deployment. It tracks all elements of the machine learning process to understand model performance and helps ensure trustworthy results. Data inteligence enables customers to deploy machine learning without specialist data scientist expertise by using preconfigured features. It also provides open-source technology, such as Jupyter lab, so data scientists can use tools they're familiar with.

SAP Hana -

is an in-memory, column-oriented, relational database management system its primary function is to store and retrieve data.

the predictive analysis Library (PAL) is delivered wit SAP HANA. these machine learning functions can be called from within SAP HANA Sql script Procedures

PAL also provides several incremental machine learning algorithims that learn and update a model on the fly, so that predictions are based on a dynamic model with streaming data.
PAL is designed to perform in memory data mining and statistical calculations to provide high performance on large datasets and for real-time analytics.
There are other machine learning libraries avaliable aswell, including the Automated predictive Library (APL), and the External machine learning Library (EML).



PAL also includes over 100 classic and universal predictive analysis algorithims in the following categories:

- Association Analysis
- Classification Analysis
- Regression
- Cluster Analysis
- Time Series Analysis
- Probability Distribution
- Outlier Detection
- Link prediction
- Data preparation
- Statistic Functions (Univariate)
- Statistic Functions (Multivariate)

SAP HANA APL is the Automated predictive Library that let's you use SAP'S automated machine learning algorithims

On Your datasets stored in SAP HANA You can create a wide range of models to answer business questions and take advantage of the automated modeling capabilities, APL provides:

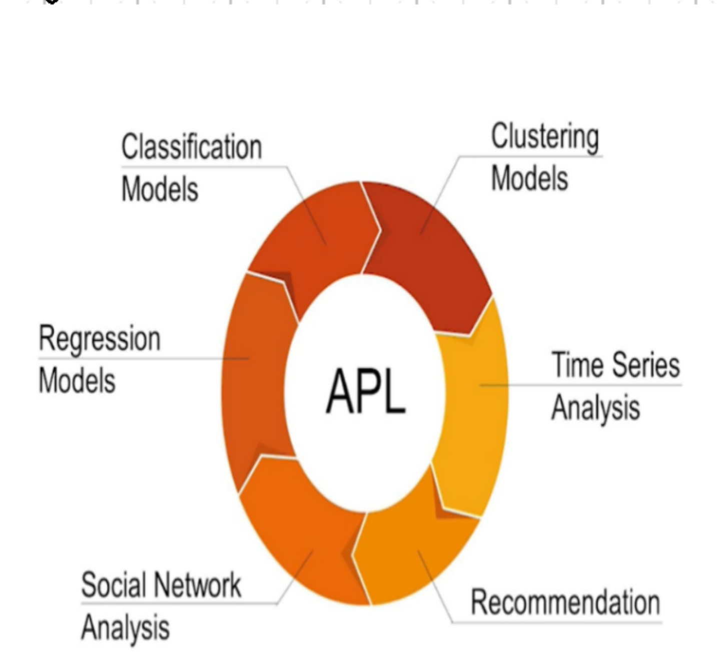
Classification and regression models

Clustering models

Time series analysis

Social network analysis

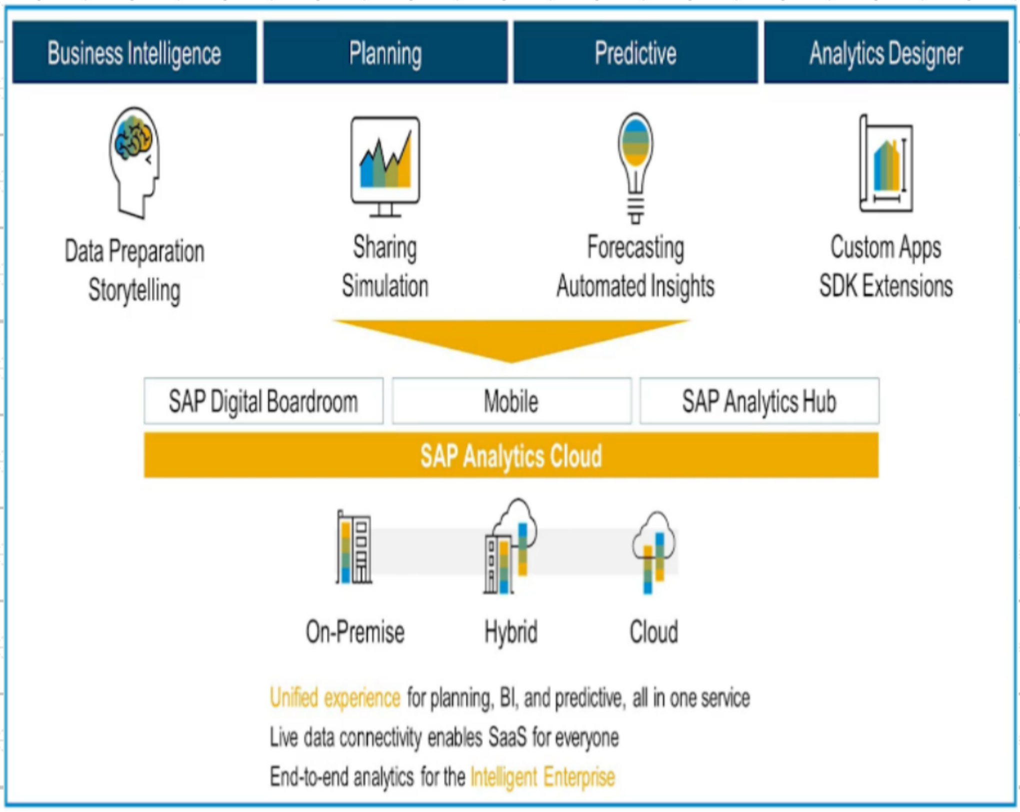
Recommendations



SAP HANA external machine learning library provides integration capabilities for R, Python and google tensorflow. The external machine learning library makes it possible to interact with the tensorflow models via HANA Sql script that is executed in SAP HANA

Tensor flow is an open-source framework for machine Learning, AI and deep Learning originally developed by google. it offers a wide range of possible applications, for example in the areas of image recognition, speech recognition and time series analysis and enables the creation of learning neural networks. Tensorflow is parti-cularly suitable for the application of neural networks, for example in deep learning, deep learning is a sub-set of machine learning where multi-Layered neu-ral networks modeled to Look Like the human brain learn from large amounts of data and make predictions, progressively Learning and gradually improving the accuracy of the outcome over time. R and python are both open-source programming languages commonly used for data science and machine learning.

SAP analytics cloud, SAC, is designed to meet the needs of data visualization in the cloud, it's delivered as an all-in-one, software-as-a-service based product. It covers the needs of data visualization, budget planning, and augmented predictive analytics, all integrated into one solution



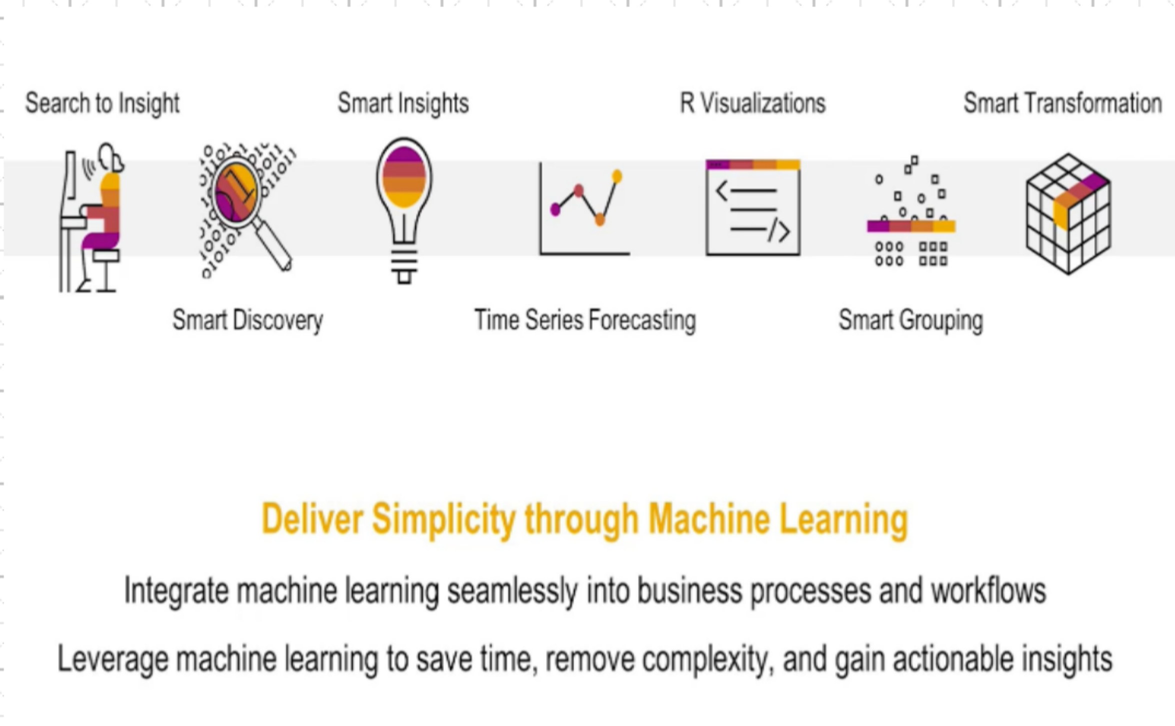
SAC provides Business users who have no data science training with a range of easy-to-use machine learning capabilities integrated seamlessly into business processes and workflows.

These augmented analytic capabities include:

- Search to insight - this provides conversational analysis capabilities where the user asks questions in real language terms and it automatically provides analysis that can be drilled into for more insights
- Smart insights - surfaces correlations in the data to help users better understand any relationships providing automatically generated charts and text.
- Forecasting capabilities - helps users make confident decision on time series data by predicting future values based on historical data.

• R visualization capabilities - users can write R scripts and insert the R visualization into storage.

• Smart grouping - automatically creates segments on different types of data

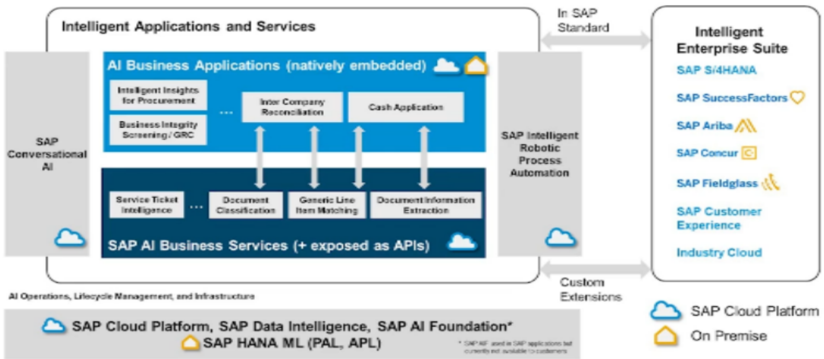


SAP exposes machine learning capabilities and AI capebilities in standart SAP applications, providing thes capabilities as enterprise-Specific solutions and Services to extend and enhance business processes.

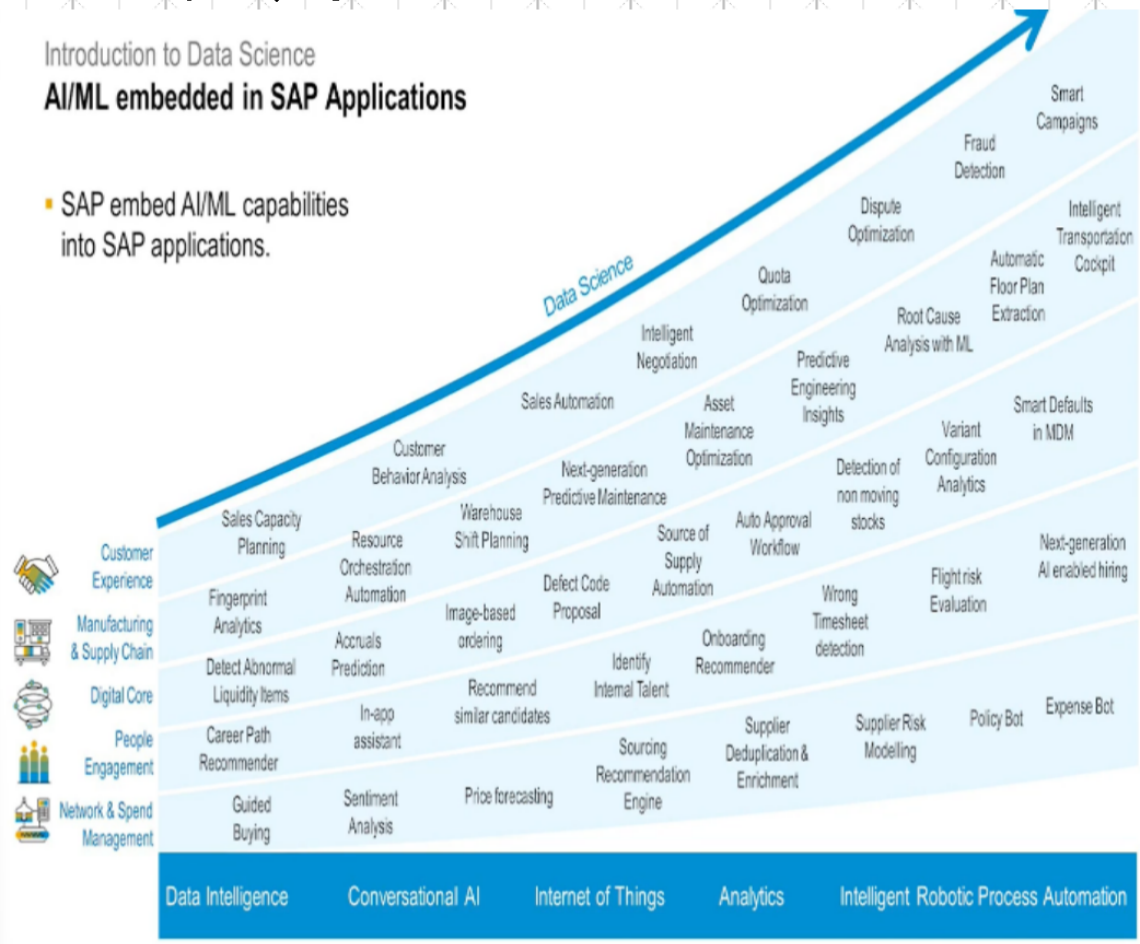
AI Foundation (AIF) is used internally in the development of SAP services and applications currently is not avaliable for customers. It enables SAP to embed intelligent data science machine learning and artificial intelligence applications into our SAP solutions, and it supports the Creation of the SAP Intelligent Enterprise

SAP AI Foundation (AIF)

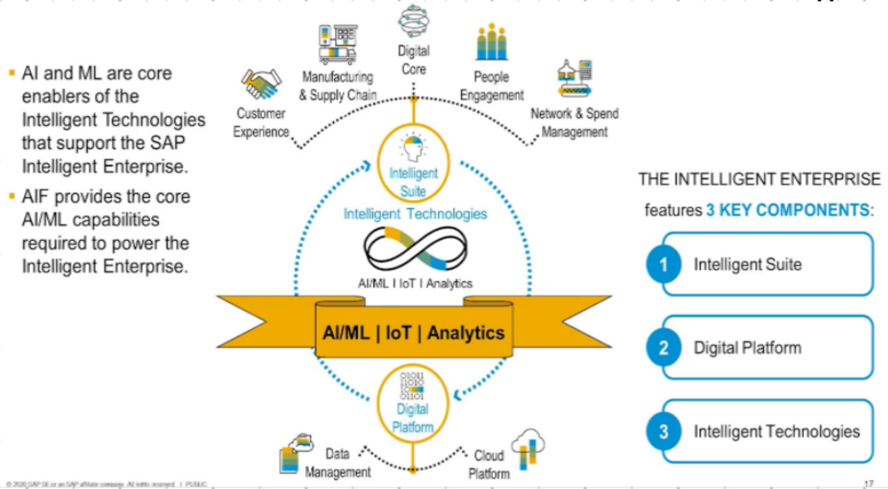
- SAP embeds ML/AI in standard SAP applications and exposes these capabilities as enterprise-specific solutions and services to extend business processes.
- AIF is used internally in SAP applications, but currently is not available for customers.
- AIF enables SAP to embed intelligent data science ML/AI applications into SAP solutions, and supports the creation of the SAP Intelligent Enterprise.



There are many examples where data science can enhance a range of business processes, from customer experience, manufacturing and supply chain, the digital core, people and engagement, to network and spend management. Embedding data science into an organization's business processes will encourage data-driven decision making, increase ROI on data investments, improve competitiveness, build customer satisfaction, and increase productivity and revenue.



Intelligent Enterprises apply advanced technologies and best practices within agile, integrated business processes. This helps them to be more resilient, profitable, and sustainable, and become best-run businesses. SAP offers an integrated suite of applications that support an organization's end-to-end business processes. The suite helps manage every part of the organization, employees, customers, products, spend, finance, and IT. With embedded analytics SAP provides a 360 degree view of the business. Machine learning and artificial intelligence are core enablers of the intelligent tecnologies that support the SAP Intelligent Enterprise. Data intelligence, HANA PAL, APL, and EML libraries, and SAC are key enablers for SAP customers. AI foundation enables SAP's development teams to build intelligence into the solutions that power the intelligent Enterprise.



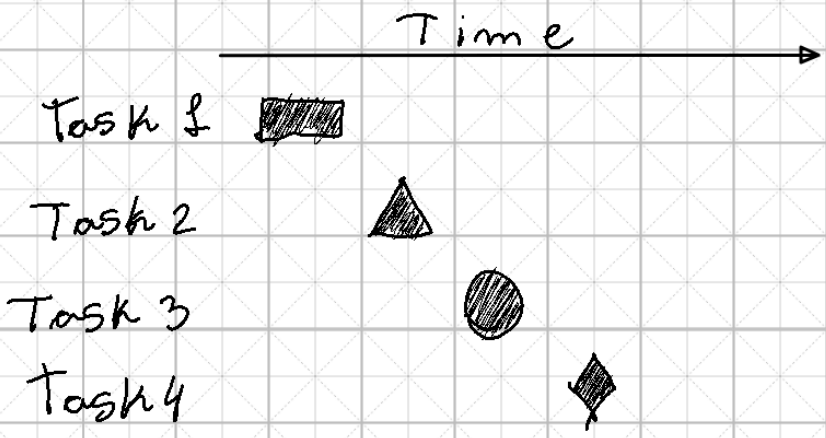
Why should there be a project methodology?

The data science process most be reliable and repeatable by people with little data science background.

A project methodology:

- Provides a framework for recording experience
- Allows projects to be replicated
- Provides an aid to project planning and management
- Is a "comfort factor" for new adopters
- Reduces dependencies on "stars"

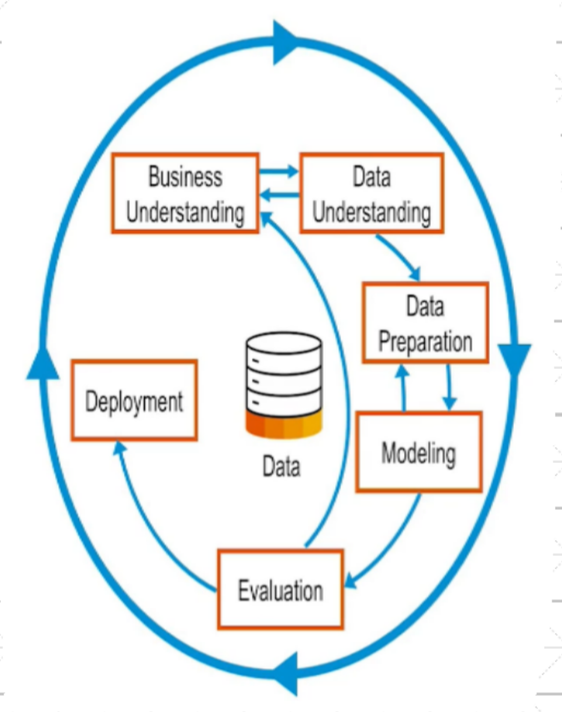
Ultimately, the methodology most support the effective integration of data science into the organization.



There is a wide range of data science prosect methodologies that have been developed over the years and in this unit we will be following the most popular of these which is called CRISP.

CRISP-DM - stands for -

CRoss Industry Standart Process for Data Mining



This was an initiative launched in 1996 Led by five companies, including SPSS, Teradata, DaimlerAG, and NCR corporation. Over 300 other organizations contributed to the process model.

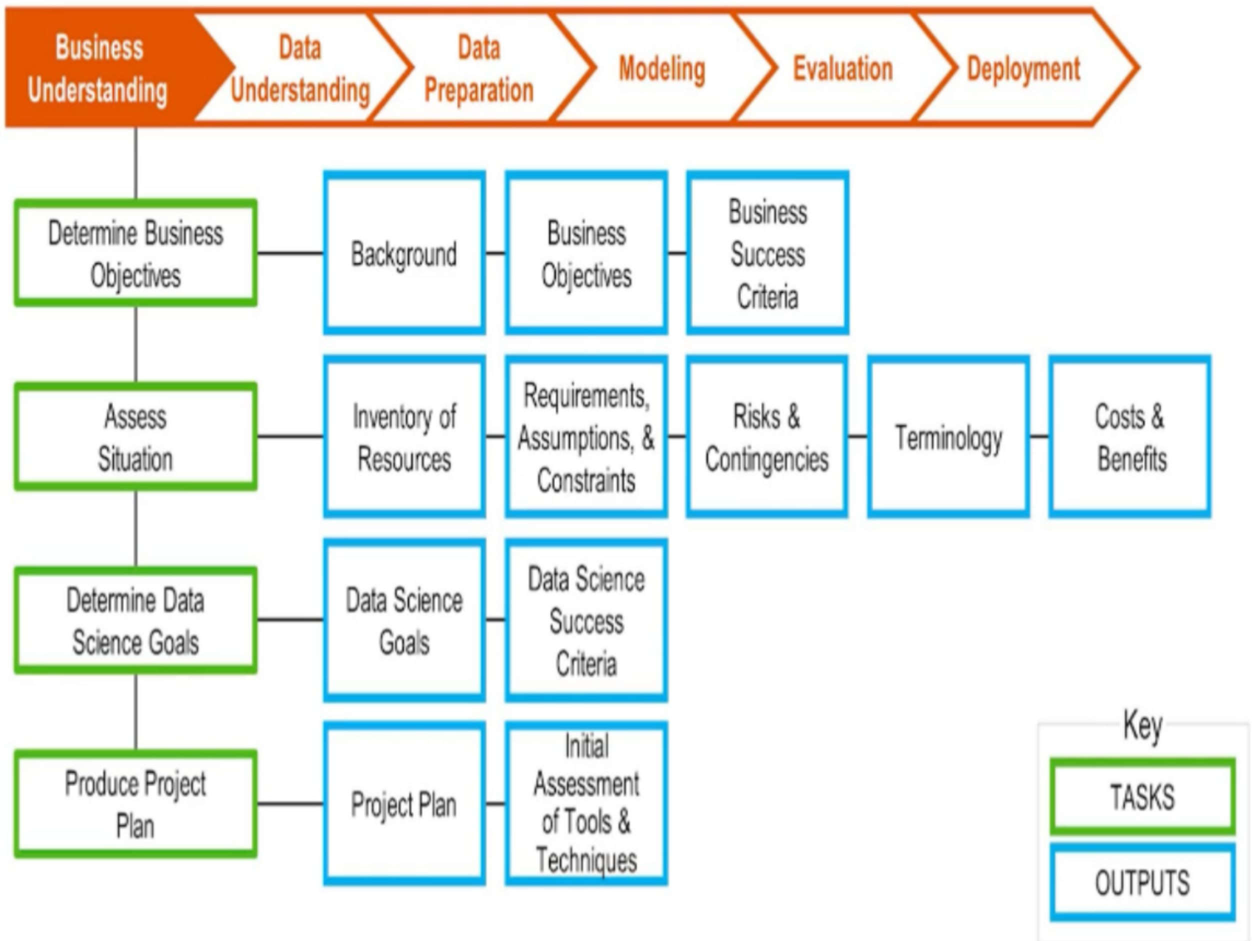
The goal was to create a data-centric proJect methodology that is non proprietary, application and industry neutral, tool neutral and focused on business issues as well as technical analysis.

The CRISP methodology is a hierarchical process model. 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 phases, compromising 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. The six generic phases are represented in this diagram.

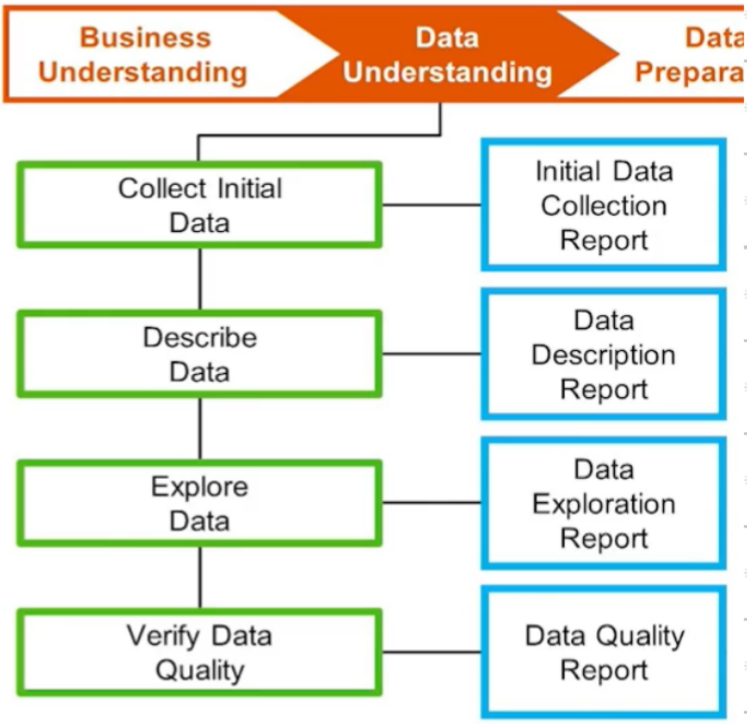
- Business understanding: confirms the project objectives and requirements from the business perspective it defines the data science approach that will answer these specific business objectives.
- Data Understanding: is the initial data collection and familiarization with the data. It identifies data quality problems.
- Data preparation: helps you to select the data tables records and attributes you will be using.

You will undertake any data transformations and any data cleaning that is required.
Then you move on to the modeling phase, where you select the modeling techniques that you'll be using and calibrate the model parameters and build the models.
The evaluation phase is where you confirm that the business objectives of the project have been achieved.
The deployment phase is where you deploy the models and productionize them if that's required. 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 allways 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. The process continues after a solution has been deployed. The lessons learned during the process can trigger new and often more →

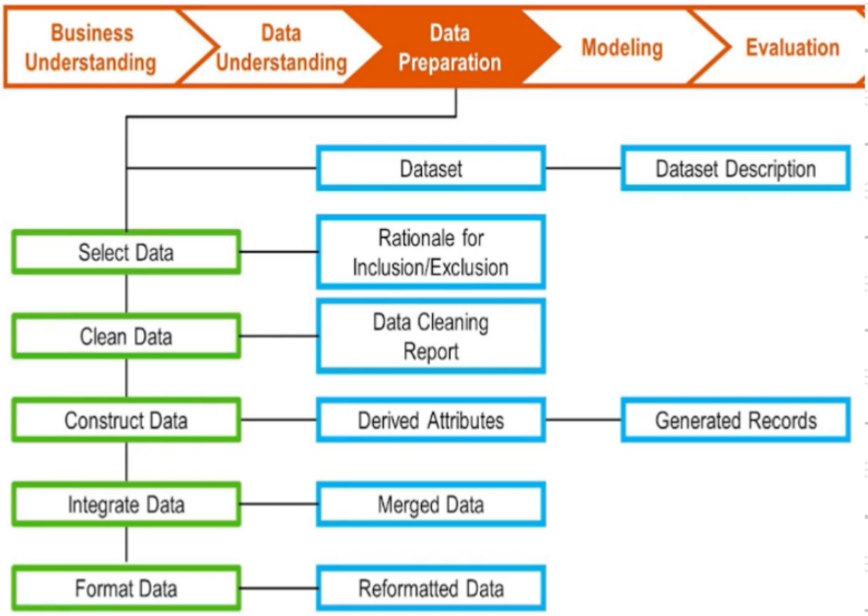
focused business questions, and subsequent data science processes will benefit from the experience of previous ones.



In the business understanding phase, it 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 is designed to achieve the objectives.



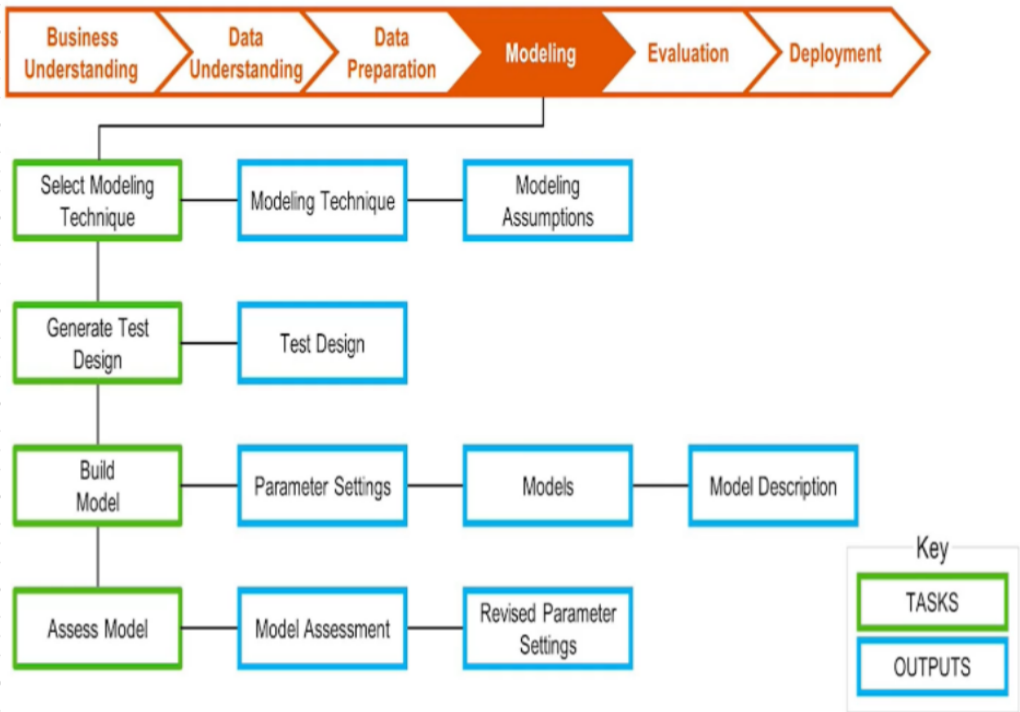
The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information.



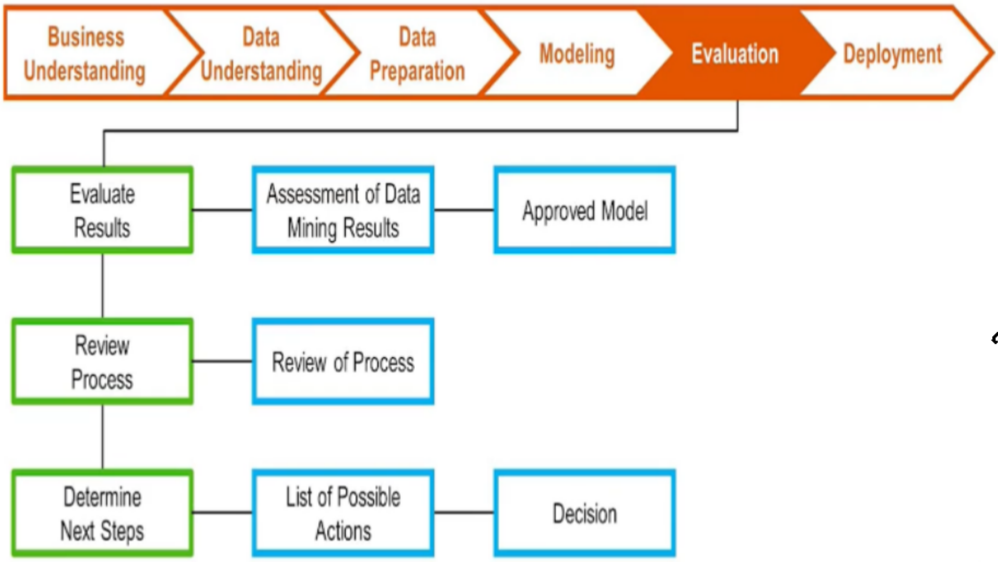
The data preparation phase covers all the activities to construct the final dataset from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order.

→

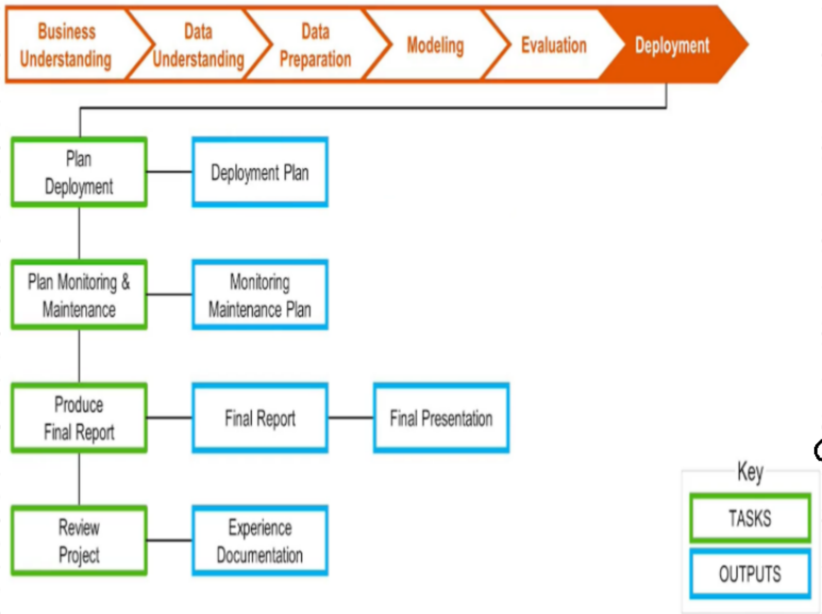
Tasks include table, record, and attribute selection as well as transformation and cleaning of data for the chosen algorithims.



In the modeling phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values. Some techniques have specific requirements for the form of data. Therefore, stepping back to the data preparation phase is often necessary.

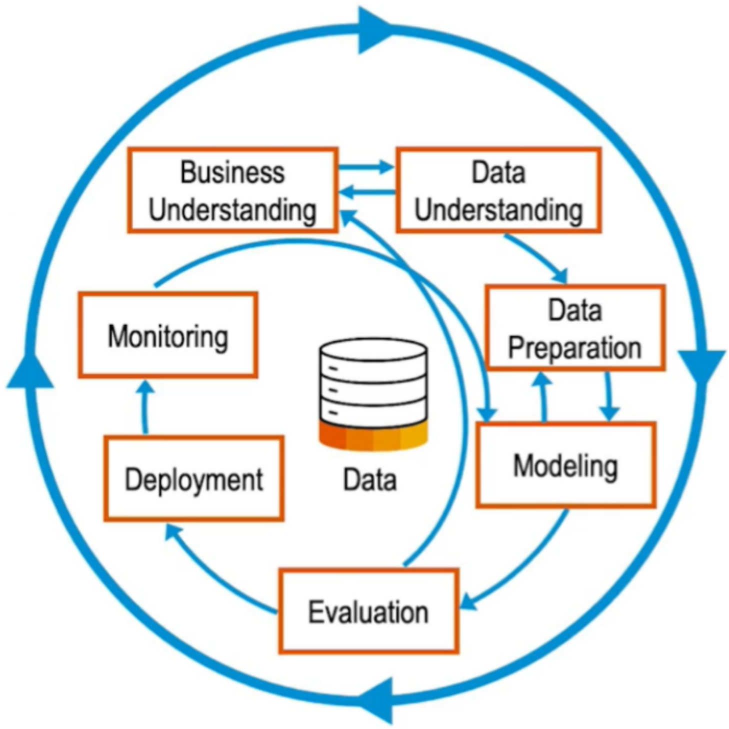


The evaluation phase thoroughly evaluates the model and reviews the model construction to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of these data science results will be reached.



The deployment phase is where the knowlege gained will need to be organized and presented in a way that the organization can use it. However depending on the requirements, →

the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data science process across the enterprise.



Most models' predictive performance will degrade over time, because the data that will be used to apply the model onto new applications will change. The data distributions might change as customer characteristics change. Competitors launch new campaigns and new products. And the general business enviroment will change The models must be updated when this happens. A monitoring phase can be added to the CRISP-DM methodology that specifically focuses on this very important aspect of any data science project.

business understanding phase overview -

Task one is to determine business objectives. We need to gain a thorough understanding, from a business perspective, of What the client really wants to accomplish, this is their business goal for the project. We must uncover any important factors at the beginning of the project that can influence the outcome of the project. Remember that missing this step could mean that we're wasting a great deal of time and effort producing the right answer to the wrong question.

Task two is to assess the situation we make more detailed fact-finding about all of the resources, the constraints, and the assumptions, and any other factors that we need to consider. And we need to flush out these details

Task three is to determine the data Science goals. Basically these state the objectives from a business perspective in business terminology. Here, we need to define a data science goal that clearly states the project objectives in technical terms.

Task four is to produce the project plan. We described the intended plan for achieving the data science goals and the business goals And remember that the plan should specify any anticipated steps that we need to perform during the rest of the project. And we need to include our understanding of the initial selection of the tools and techniques.

You also needs to understand how the customer wants to deploy the results from the analysis. So, for example, will a predictive model be deployed in a production system? Will a dashboard with real time updates of the model outputs be required? Or is the requirement for a report delivered in a text document? The outputs from this are an assessment of 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 business goal from a business perspective. And in addition to the primary business objectives, there are typically other related business questions that the organization might like us to address. So, for example, the primary business goal for a financial service business might be to keep current customers by predicting when they are most prone to move to a competitor. Examples of related business questions are " How does the primary channel a business customers use, so, for example, the ATM, or branch visits, or via the internet, affect whether they 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 it could be general and Subjective, such as "Can you give useful insights into a relationship?". And, of course, you will record this information that describes any deployment requirements for the analysis and the outputs from the analysis, so these are clearly defined and agreed between yourself and the customer.

Task two is to asses the business situation. In the previous task, the objective is to quickly get to the crux of the situation And here, we want to flesh out the details. This task involves more detailed fact-finding about all of the resources the constraints and assumptions, and any other factors that should be considered in determining the data science goal and the project plan. We need to list the resources that are avaliable to the project, including the personnel, so for example, the business experts, data experts, technical support that's required, and of course the data science personnel. The data that we need, are we going to be using fixed extracts, or do need access to a live warehouse or operational data. The computing resources we'll be using, the hardware for example, and software, the data science tools that we're going to be using. The outputs from this are the inventory of resources. So here what we do is where we list all the resources that are avaliable to the project, we list all the requirements of the project, including a schedule of completion. As part of this output, we need to make sure that we're allowed to get access and use the data. We list the assumptions made by the project. And these might be assumptions about the data that can be checked during the analysis process. It's particularly important to list these if they form conditions on the validity of the results. We list the constraints of the project. And these might be constraints on the availability of resources, but might also include technological constraints such as the size of data that is practical to use for the 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, and this might include two different components. You can have a glossary of relevant business terminology, which forms part of the business understanding available to the project, And also you can have a glossary of data science terminology that illustrates, ⇾

With examples, any relevant technological algorithms were using, so that it helps the business understand the approach that we're using. We often will also conduct a cost-benefit analysis for the project, which compares the project with the potential benefit to the business if it's sucessful.

Task three is to determine the data science goals. A business goal states objectives in business terminology, and we defined this in task one. A data science goal states project objectives in technical terms, so for example, the business goal might be "Increase 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, where they live, and the price of the item.", the output from this is a list of the data science goals. here we describe the intended output of the project that enables the achievement of the business objectives, but from the data science perspective. We also define the data science success criteria for a successful outcome of the project in technical terms. For example, a certain level

of predictive accuracy or a propensity to purchase profile with a given degree of model performance, such as lift.
Task four is to produce a 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 from this is, 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 really important to analyse the dependencies between the time schedule and the 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 really recommended. Specific review points for these updates must be made to

explicit within the project plan. At the end of the first phase, the project also performs an initial assessment of tools and techniques. So for example, You select a data science algorithm that supports the available data and what the required output will be. It's important to assess tools and techniques early in the process since this selection possibly could influence the entire project.

it's important for us to clearly define business and data science project success criteria. First we define the criteria for a successful or useful outcome to the project from from the business point-of-view. This might be quite specific and able to measure objectively, such as a reduction of customer churn to a certain Level, or it could be general and subjective, such as, give useful insights into a relationship. We also define the criteria for sucessful outcome to the project in data science technical terms. For example, a certain Level of predictive accuracy needs to be obtained.

→

Figure 5. Based on 110 users who have implemented predictive analytics initiatives that offer "very high" or "high" value. Respondents could select multiple choices.

PREDICTIVE ANALYTICS

Extending the Value of Your Data Warehousing Investment
By Wayne W. Eckerson

See: https://www.rexeranalytics.com/

In their Third Annual Data Miner Survey, Rexer Analytics, an analytics and renowned CRM consulting firm based in Winchester, Massachusetts, 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, R2, etc.) as the primary factor.

These industry surveys indicate standard methods of assessing data science project success. In both 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 important as you can see, and we'll be discussing how we calculate lift in more detail later in this course.

On the right-side, in their Third Annual Data Miner →

Survey, which is conducted by Karl Rexer Analytics organization, and this is a really renowned CRM consultancy firm that's 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, which is, for example, the Lift of a predictive model, as being the primary factor.

Model success criteria: descriptive or predictive models

The data science success criteria will differ depending on whether the models are predictive or descriptive type models and the type of algorithm chosen.

Descriptive Models

- Descriptive analysis describes or summarizes raw data and makes it more interpretable. It describes the past – i.e. any point in time that an event occurred, whether it was one minute ago or one year ago.
- Descriptive analytics are useful because they allow us to learn from past behaviors and understand how these might influence future outcomes.

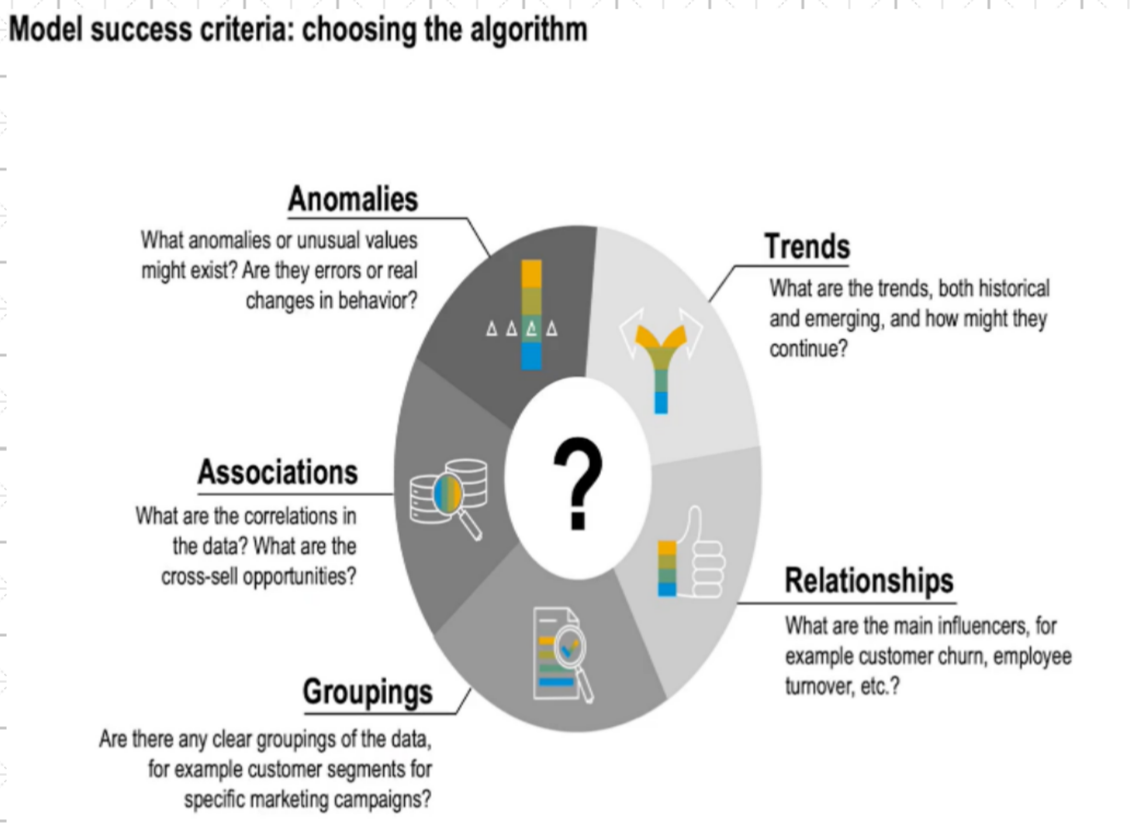
Predictive Models

- Predictive analysis predicts what might happen in the future – providing estimates 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 to determine the probability of customers making future credit payments on time.

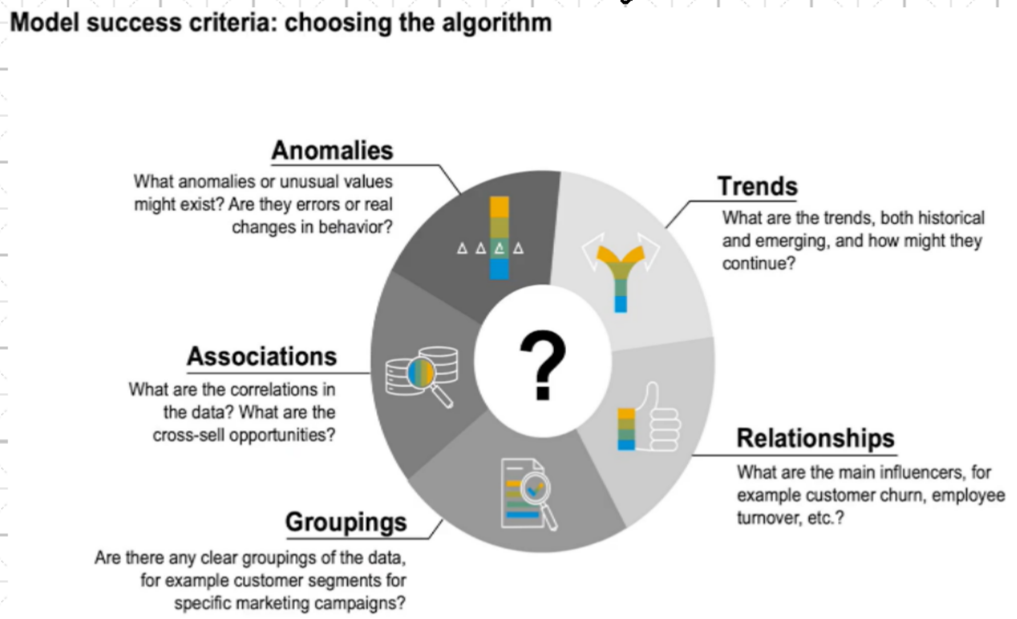
Descriptive analysis describes or summarizes data and makes it more easily interpretable. But it can only analyze historical performance, so, for example, what happened in the past week, month or year. Descriptive analytics are useful because they allow us to learn from past behaviours, and understand how they might influence future outcomes. Common examples of descriptive analytics are BI reports that provide historical insights about a company's production levels, or their financials, and customer transactions.

Descriptive analytical models will include cluster models, segmentations, association rules, and network analysis.

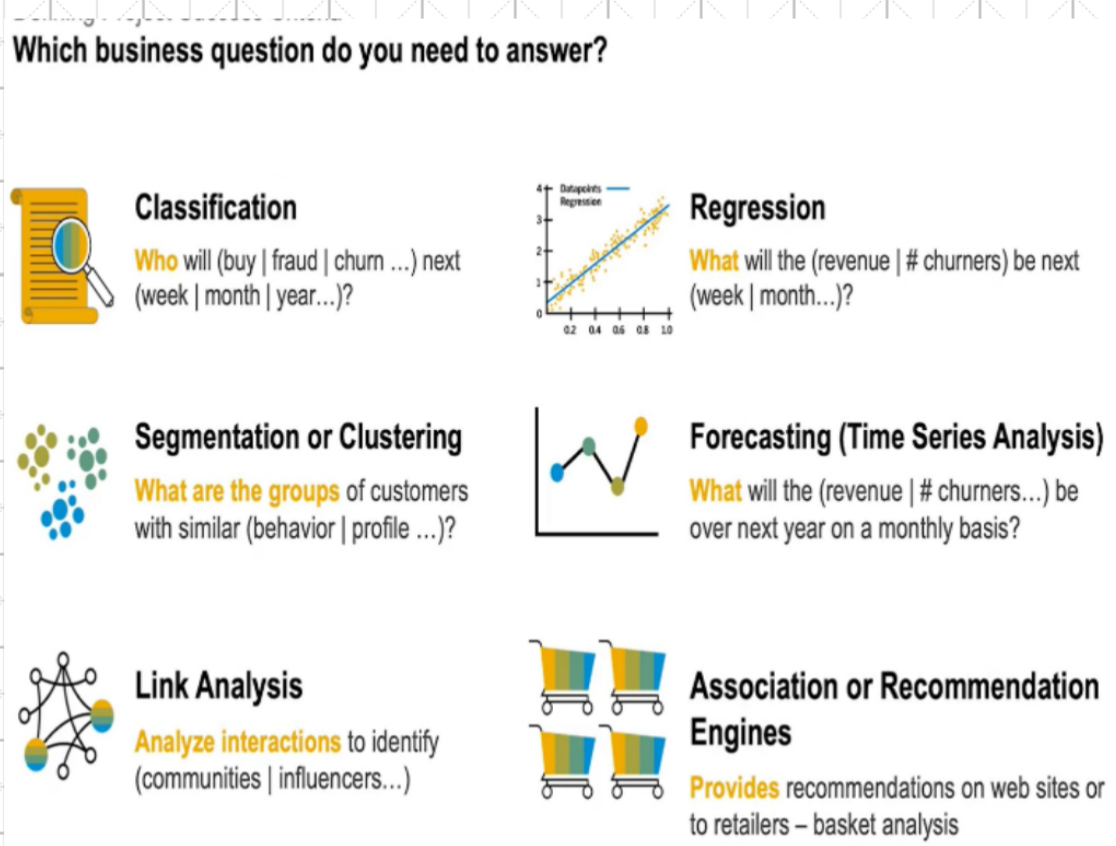
Predictive analytics is very different. For predictive analysis, you're predicting what's going to happen in the future. These models provide estimates or probabilities about the likelihood of a future outcome. One common example is the use of predictive analytics to produce a credit score. And these scores are used by financial service businesses to determine the probability of customers making future credit payments on time. Typical business uses will include forecasting how sales might close at the end of the year, or 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 analitical models include classification models, regression models and neural network models.



The business question helps to determine the most likely algorithm to use. We can choose algorithms to analyze trends in data and use this information for forecasting, we can identify the main influencers and relationships that could be driving customers to switch to another supplier, and we call this churn analysis. Or why certain customers are more likely to respond to a marketing campaign offer and buy specific products. Some algorithims group observations or customers together, so that all of the customers in a group have similar characteristics, and these are called cluster algorithims. We can use algorithims to understand which products to recommend to customers in a cross-sell 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 algorithims.



There is a wide range of algorithims to chose from depending on the type of question asked by the business, the output that is required, and the data that are avaliable. for association rules, or Sometimes we call this basket analysis, we have an Apriori algorithim that analyses the combinations of products purchased together in a basket or over time. For clustering, where we are creating groups of similar observations, SAP often uses a k-means algorithim. For classification analysis, where we are classifying observations into groups, we can use decision trees or neural networks. And we can use outlier analysis to identify which observations unusually high or low values. Regression algorithims enable us to forecast the values of continuous variables, such as customer spend, in the next 12 months. And time series analysis enables us to forecast future KPI values, and control stock and inventory levels.



Each of these categories of algorithims 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?". "How many customers will churn next year? For clustering and segmentations, 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 we can actually develop specific products or services for customers in each segment.
Forecasting allows us to estimate a KPI on a regular time interval, so, for example, we can forecast revenue for the next 12 months, accounting for trends, and seasonalities, and other external factors. Link analysis is used mainly in telecommunications to create communities of customers who are calling one another, or it can be used in retail analysis to analyse the links between customers and the products they have purchased, and then to support product recommendations, and association rules and recomendations are used for basket analysis, to produce product recommendations for customers.

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

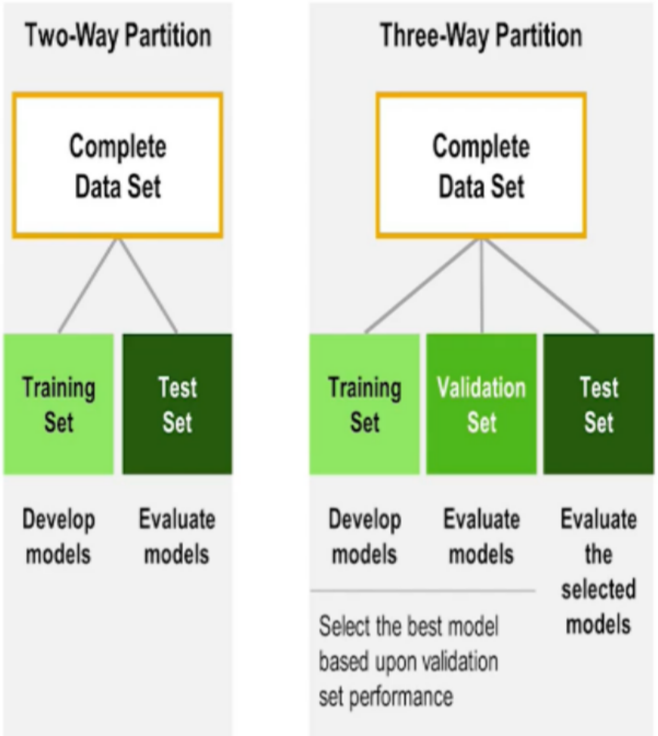
often the starting poit for analyzing the quality of a predictive model, as well as an obvious criterion for prediction. Accuracy measures the ratio of correct predictions to the total number of cases evaluated. There are a wide variety of metrics and methods to measure accuracy, such as lift charts and decide tables, which measure the performance of the model against random guessing, or what the results would be if you didn't use any model.

The robustness of a predictive model refers to how well the model works on alternate data. This might be hold-out data that the model is to be applied onto, and robustness is vitally important.

The predictive performance of a model must not deteriorate substantially when it is applied to data that were not used in model estimation.

Training and testing: data cutting strategies

- Central to developing predictive models and assessing if they are successful is a train-and-test regime.
- Data is partitioned into training and test subsets. There are a variety of cutting strategies (e.g. random/sequential/periodic).
- You build a model on the training subset (called the estimation subset) and evaluate its performance on the test subset (a hold-out sample called the validation subset).
- Simple two and three-way data partitioning is shown in the diagram.



Central to developing predictive models and assessing if they are successful is what we call a train-and-test regime. Data is partitioned into training and test subsets. There's a variety of different strategies to cut the data, so you can use random partitioning, sequential or periodic. We build our models on the training subset, that's called the estimation subset, and evaluate its performance on the test subset, this is a hold-out sample that we →

sometimes call validation subset.

Simple two-way and three-way data partitioning is shown in the diagram on the slide.

When a predictive model has been built on the estimation sub-sample, it's performance is tested on the validation and test subsamples. And we would expect that the model will have similar performance on the estimation, validation, and test subsets. The closer the performance of the models on all of the subsets, 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 wasn't 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 totally new campaigns to see how well it performs in a real environment. →

Initial data Analysis -

It's wise to begin any analysis with an exploratory examination of the data in order to get a feel for them.

This initial analysis includes processing the data into a suitable form for analysis and checking data quality

IDA = Initial data analysis -

coined by Chris chatfield in the book: Problem Solving, A Staticians Guide Christopher Chatfied

- Chatfield defines the various steps in IDA:
  - The structure of the Data
  - The quality of the Data
    - errors, outliers and missing observations
  - Descriptive Statistics
  - Graphs
- The data is modified according to the analysis:
  - Adjust extreme observations, estimate missing observations, transform variables, binning

data, forming new variables.

EDA - Exploratory Data Analysis -

Coined by John W. Tukey, used to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

The objectives of EDA are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which the analysis and statistical interference will be based.
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments.

quiz -

1- which of the following functions are integrated in SAP Analytics Cloud (3 answers).

- BI
- Planning
- Predictive

2- Why do you sometimes need to add a monitoring phase into the CRISP-DM Process?

- Because changes to the general business environment might mean the existing model needs updating
- Because the data that we apply onto the model has changed in some way
- Because the model's performance degrades in time

3- which phase of the CRISP-DM methodology immediately precedes the deployment phase?

- Evaluation

4- When you use a train-and-test regime, which subset is used to develop the model?

- Train

5- In which task of the Business Understanding phase of the CRISP-DM methodology do you list the risks and contingencies?

- Assess Situation

6- What are the capabilities provided by SAP HANA Automated predictive Library?

- Time series analysis
- Classification models
- Recommendations

7- In which report generated in the data understanding phase of CRISP-DM methodology do you list the quantity of the data?

- Data description report

→

8- which of the following are descriptive models?

- Cluster models
- Association Rules

9- Which of the following ar results of an Initial Data Analysis (IDA)?

- An analysis of the quality of the data
- The identification of missing observations
- The analysis of descriptive statistic

10- which statement regarding the CRISP-DM methodology is true?

- It focuses on business issues and technical analysis