Quality Guideline

Training Program

Data Mining Advanced Course

European
Six Sigma Club Deutschland e.V.
www.sixsigmaclub.de

Version: 1.0

Status: 23/09/2021

Contents

Foreword

Background

Digitalization is in full swing, providing us with an unprecedented availability of data. More data, both structured and unstructured, is available from an increasing number of sources in varying quality.

This is changing the way we approach data-driven problem solving and process improvement, including Six Sigma. Previously, the focus was on collecting the right data to solve a problem on a project-by-project basis. Today, however, the focus is increasingly on efficiently selecting the relevant data from large amounts of existing data to achieve the best possible result.

To address this development, it is necessary to extend the Six Sigma toolbox towards data sciences, in particular big data, data mining and machine learning, and to complement the existing Six Sigma training.

The training content described below describes the skills that are increasingly being used in quality and process improvement.

Origin

This guide was developed by the Six Sigma Thinking Ahead working group. Based on the experience gained from numerous projects and training courses, the working group discussed the impact of "digitalisation" and "big data" on the proven Six Sigma process improvement method.

This was based on different data availability scenarios (from "no data available" to "all data available in context"). At an early stage, the working group linked the DMAIC cycle to recognised process models for data analysis. The aim was to define the current and target situation in relation to the phase and scenario, and thus identify the necessary changes and additions to the Six Sigma toolbox.

The results were presented and discussed at the Six Sigma Club conferences. In addition, several relevant publications and presentations were produced. Finally, the results of the work and the training content identified as new were presented to brand owners and members for evaluation in separate surveys. These results, together with the results of ongoing research projects at Koblenz University of Applied Sciences and FAU Erlangen, have been used as the basis for this training guide.

The diagram below illustrates the training requirements in the identified competence areas. Theoretical principles and practical applications are intended to fill these gaps. The network diagram is based on the sources mentioned above. Low numbers indicate no/low expertise or little need for additional training. High numbers indicate a high level of training or high requirements. In other words: 0 means no/low competence and 3 means high competence.

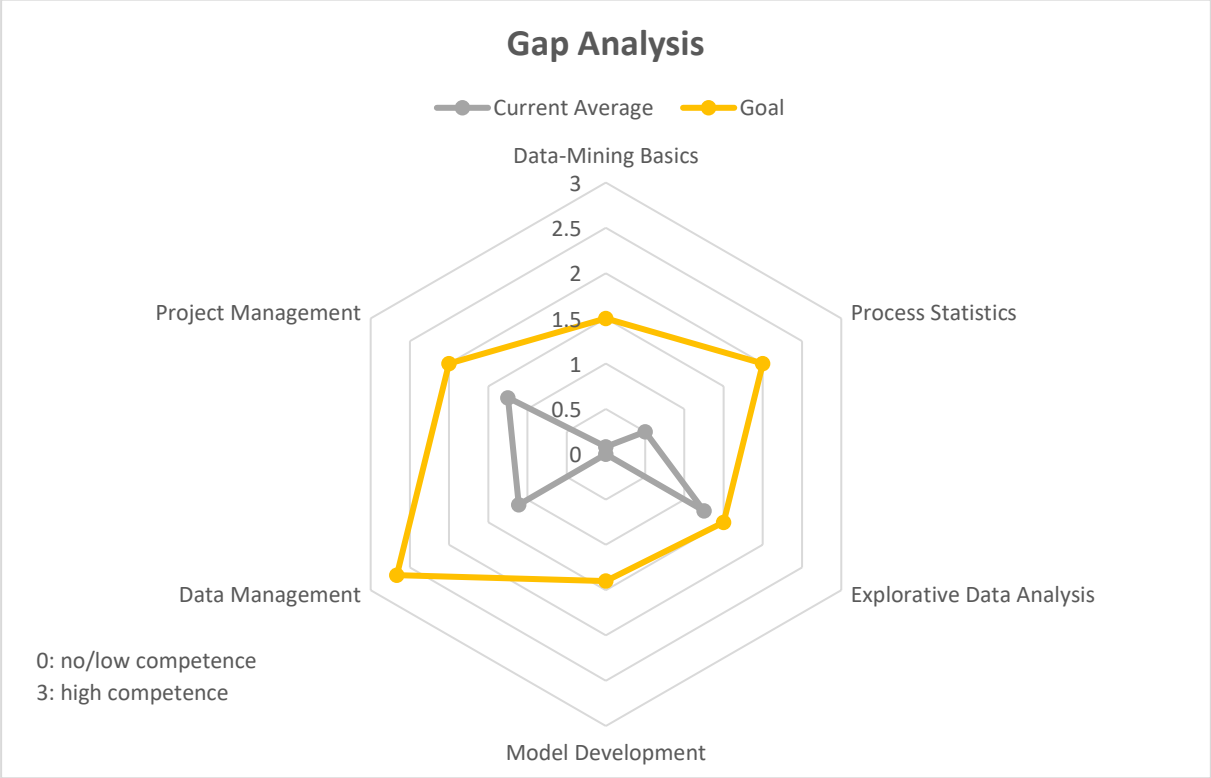

Illustration 1: Gap analysis methodological competences

Goal and target group

The aim of the Advanced Data Mining course is to provide method experts with the tools to analyse more complex data structures. At the end of this course, participants will know the requirements for analysable data structures and be able to create them, including across different data sources. They will be familiar with key data mining (DM) terms and their meaning. You will be able to apply common data mining methods and algorithms to identify correlations in data and to verify developed models. Both regression and classification models will be developed, benchmarked and evaluated. You will also be able to use machine learning (ML) approaches to develop, deploy and monitor correlation and prediction models in applications.

This course is specifically designed for method experts (not just Six Sigma Belts) who are repeatedly faced with questions about relationships in data in their work or projects. Participants in this course are familiar with their area of expertise and already perform statistical data analysis on a regular basis.

CRISP-DM and similarities to Six Sigma DMAIC

With the changing availability of data, new process models often come into focus. One of the most prominent process models in the big data and I4.0 environment is CRISP-DM. In fact, there are countless comparable process models with different names and only minimal adjustments. The focus of CRISP-DM is on the various activities and objectives of the respective phases. In contrast, CRISP-DM provides little or no guidance regarding the tools that can be used.

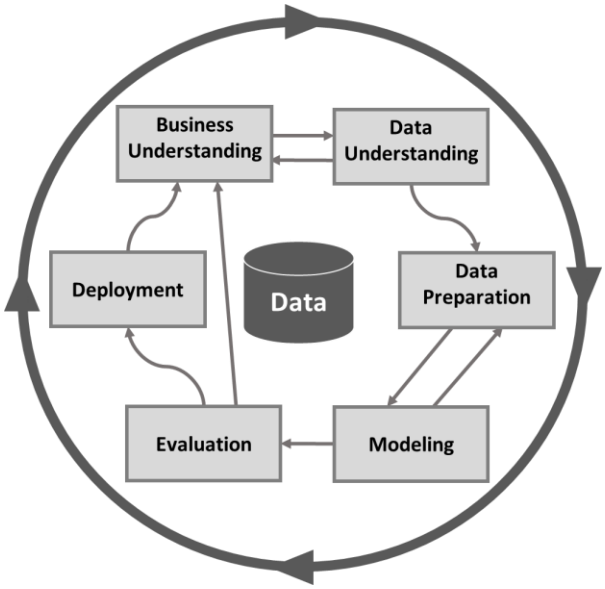

Illustration 2: Process model CRISP-DM

A closer look reveals that the DMAIC and CRISP-DM phase models complement each other perfectly. In many cases, DMAIC provides tools and methods for the corresponding phases and phase objectives of CRISP-DM. By overlaying or integrating CRISP-DM into DMAIC, the circle of experts who need to be consulted depending on the phase of a project is expanded. Interdisciplinary collaboration is more important than ever for project teams in the big data environment.

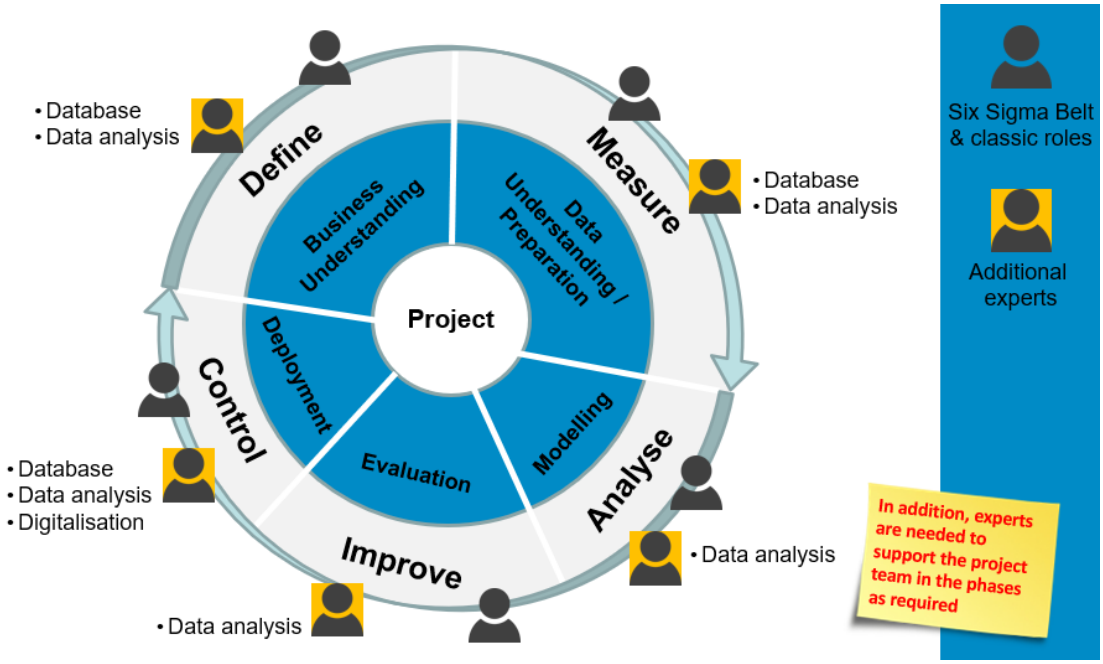

Illustration 3: Integration of CRISP-DM and DMAIC

1 Introduction

This guide is divided into different sections.

First, a brief definition of the scope of the advanced data mining course is given, followed by an explanation of the duration of the course.

The training content is then categorised into the different project phases according to DMAIC.

In the next section, the training content is categorised in terms of its scope (delivery) and its objectives as part of the advanced course, using the ESSC-D's own customary evaluation scheme.

It is important to note that the training of the tools is not specified in each phase. Rather, the list is intended as a suggestion. As many tools can be used in several phases, it is up to the trainer to deliver the content at the appropriate time according to the didactics used.

The individual training topics are then described in more detail to show the close connection between digitalisation and big data on the one hand and Six Sigma on the other. The current requirements for the changed qualifications of Six Sigma Belts are clarified.

Finally, an example of a possible training programme is provided as a guide for trainers.

2 Areas of application

The guideline describes the recommended additions to the training for Six Sigma Green Belts, Six Sigma Black Belts and for method experts in general. The scope and depth of the training are described in a classified manner and are used for comparison with existing or newly developed training courses. As this is a basic course, there are no differences in scope or depth of coverage. Participants will learn to apply basic data analysis techniques independently. Building on the fundamentals of data mining and machine learning, it is easy to build on what has been learned. However, to be able to understand the language of specialists in projects, it is not recommended to differentiate between the depths of learning.

3 Training duration

For the additional training program Data Mining for Six Sigma Green Belts, Six Sigma Black Belts and methodology experts, a minimum of 3 days of instruction with a minimum of 30 teaching units of 45 minutes each plus breaks must be completed to teach the content described below and to achieve the required level of instruction.

In practice, it may make sense to extend the training duration to 4 teaching days with a total of 40 teaching units of 45 minutes each, plus breaks.

Universities may achieve the required level of instruction by dividing the teaching units between attendance (lecture) and a proportion of self-study. The maximum proportion of self-study allowed is 25% of the total number of teaching units, based on the minimum scope of the standard course described above (30 teaching units). Self-study units will be multiplied by a factor of three. For the additional Data Mining training, with a maximum use of self-study of 25% and a minimum number of

teaching units, this results in 24 teaching units of classroom time and an additional 24 teaching units (8x3) of self-study.

To receive a certificate of attendance, the student must have attended at least 85% of the total number of hours scheduled for this course.

4 Training content, project phase orientated

In the following chapter, the individual training contents and their respective benefits are allocated to the DMAIC phases.

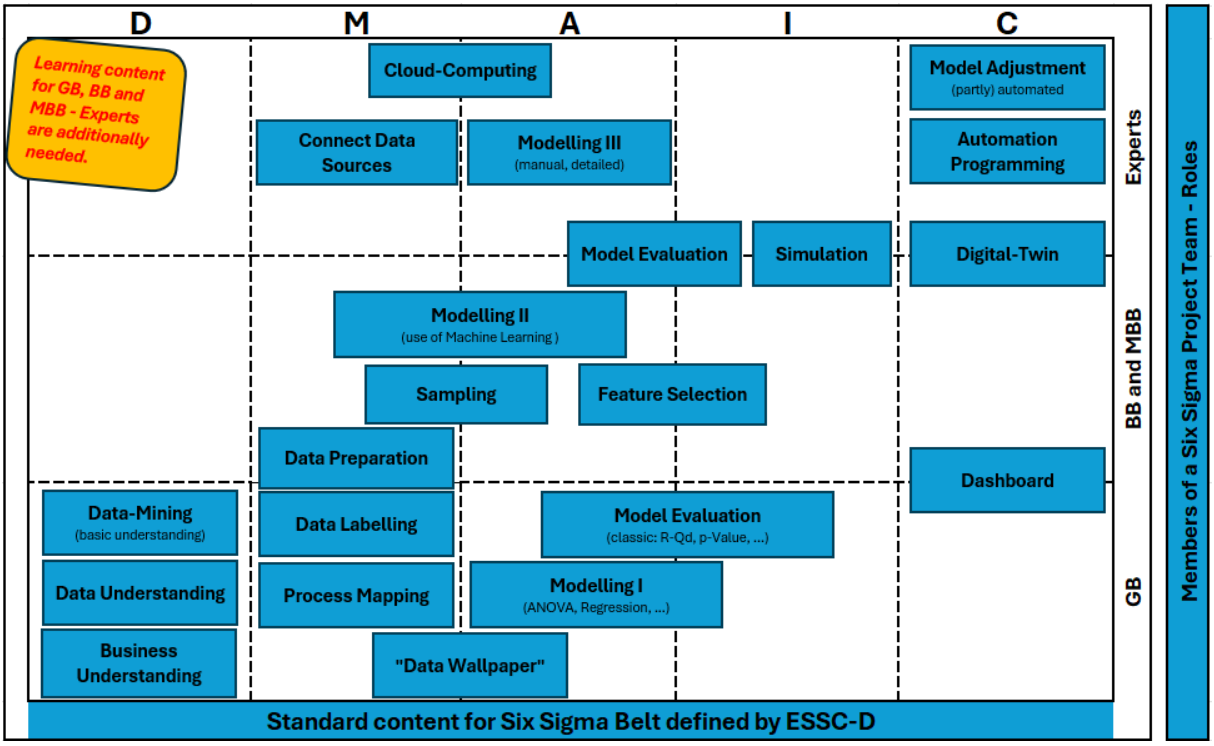

Illustration 4 :Overview - New training content and phase allocation

4.1 Kick-off (K)

- Purpose and need of the advanced data mining course
  - Focus on: Descriptive analytics and correlation models in more complex data structures, predictive analytics and prescriptive analytics
- Basic concepts of digitalisation (AI, AI, ML, AR, I4.0, IOT)
  - Targeted use of keywords and their meaning and the associated approaches
- Modern project management concepts in the digital transformation
  - DMAIC and modern project management approaches, such as Lean, Scrum, Agile, Design Thinking
  - Integration of modern management methods into classic DMAIC

4.2 DEFINE (D) - Business Understanding

- Stakeholder analysis (possibly extended and/or combined with RACI)
  - Extended process focus through modern computer-based analysis methods

- Growing importance of comprehensive stakeholder management
- Data SIPOC / Process Data Map
  - Delineation of areas of the value chain from which data is analysed
  - Identification and description of the parameters and data sources affecting processes and products

4.3 MEASURE (M) - Data Understanding and Preparation

- Extended data collection plan
  - Overview of project-relevant parameters with data origin and examples with data quality requirements
  - Identification and processing of the rapidly growing amount of structured and unstructured data
- Basics of ETL and common tools
  - Extracting the relevant data from various sources
  - Transforming the data into the schema and format of the target database
  - Loading the data into the target database
- Feature Selection
  - Quick and easy identification of potential factors from a large number of parameters
- Extended C&E matrix
  - Comparison of expert and data views as a basis for a discussion on the distinction between correlation and causality
- Tools for data preparation (Data Preparation)
  - Aggregating, pivoting, transposing, binning
  - Dealing with missing values
  - Dealing with multicollinearity and outliers

4.4 ANALYZE (A) - Modelling

- Differences between supervised and unsupervised learning (using the example of anomaly detection and prediction model)
  - Supervised learning: developing a predictive model based on input and output data
  - Unsupervised learning: grouping and interpreting data based on input data
- Frequently used data mining methods
  - Bootstrapping (better and more robust estimates for the spread of parameters)
  - Labelling (data categorisation)
  - Ensemble (run multiple models in parallel and let them decide)
  - Feature selection (automatic detection of influencing variables)
- Exploratory data analysis (EDA): graphics and descriptive statistics
  - Uncovering regularities, dependencies or special correlations between seemingly completely dissimilar data
  - Multidimensional graphical representations
- Dealing with unbalanced data
  - Preparing data when there are few errors and a lot of "good" data, using methods such as over-sampling or under-sampling

- Common data mining tasks and related procedures (algorithms)
    - Classification and regression (linear and logistic regression, decision trees, artificial neural networks)
    - Clustering and segmentation (K-Means, artificial neural networks, clustering methods) for grouping based on predefined criteria
    - Dependency analysis (association analysis)
    - Discriminant analysis (distinguishing between two or more groups described by several characteristics (variables))
- Typical data mining tools
    - Python, R-Studio, KNIME, Rapid Miner, SPM, ...
    - Concrete example with one of the tools presented
- Strategies for the selection of training and test data
    - Presentation of different approaches depending on the application
- Parameters for evaluating the model quality for training and test data
    - R-Qd, ROC, BIC, AIC, Confusion Matrix: Meaning and advantages and disadvantages of the various parameters
- Cloud computing
    - Advantages and disadvantages of cloud computing, the external infrastructure for computationally intensive modelling, typically including storage space, computing power, application software as a service, etc.
- Overview of standard solutions based on big data and AI
    - Available libraries, platforms, image and language processing as well as word processing (mostly available as open source)
- Possible uses of more complex data mining models
    - Setting recommendations for models
    - Practical examples of projects with predictive analytics and prescriptive analytics

4.5 IMPROVE (I) - Verification

- Basic principles of model deployment
    - CICD - Continuous Integration Continuous Deployment as a possible basic principle

4.6 CONTROL (C) - Deployment

- Performance monitoring and re-training of the models
    - Monitoring the performance and forecasting capability of the model used for technically relevant changing processes
    - Checking the training and test data and, if necessary, initiating a new training of the model
- SPC 4.0
    - Extended process control using modern methods up to full automation
    - Using a statistical model for process monitoring

5 Scope and objectives of the individual topics

The topics, methods and tools defined above describe the minimum content required for the training. This section specifies the scope and objectives of these topics based on classifications. The outcome of the training shall be at or above the specified level to comply with the guideline.

5.1 Classification legend

5.1.1 Classification for the scope (delivery)

Class	Meaning
A	Method was explained
B	Method was shared
C	Method was practiced alone or in a group
D	Method was practiced including feedback on the exercise

5.1.2 Classification of goals

Class	Meaning
1	The participant has understood the principle of the application
2	"1" and participant can select & use tool
3	"2" and participant can interpret important results
4	"3" and participant knows the calculation background in detail
5	"4" and participant can also calculate the result manually

5.2 Classification for the advanced data mining course

	Topic	Phase	Scope (Delivery)	Goal
1	Smart Knowledge Picking	K	A	2
2	DMAIC and modern project management approaches (Lean, Scrum, Agile, Design Thinking, ...)	K	B	1
3	Basic concepts of digitalisation (AI, AI, ML, AR, I4.0, IOT)	K	B	1
4	Possibilities and limits of inductive statistics and modern data mining	K	B	1
5	Stakeholder analysis (possibly extended and/or combined with RACI)	D	B	3
6	Data SIPOC / Process Data Map	D	C	3
7	Extended data collection plan (for project-relevant parameters with data origin, data quality and examples)	M	B	3
8	Data quality requirements (measured values, curves, images, text) for data mining	M	C	4
9	Feature selection for complexity reduction	M	B	3
10	Extended C&E matrix (feature selection: plausibility check of selected features with experts)	M	B	3
11	Tools for data preparation (aggregating, pivoting, transposing, binning; dealing with missing values, multicollinearity and outliers)	M	C	3
12	ETL/ELT basics and common tools	M	B	1
13	Differences between supervised and unsupervised learning	M	B	1
14	Frequently used data mining methods (bootstrapping, labelling, committee, feature selection, ...)	M	A	1
15	Exploratory data analysis (graphics and descriptive statistics)	M	C	2
16	Dealing with unbalanced data (only a few errors, but lots of "good" data)	M	A	1
17	Frequently used data mining algorithms (decision tree, neural networks, Naive Bayes, SVM, K-Nearest, K-Means, principal component analysis, Lasso, ...)	A	A	1
18	Advanced multivariate analyses (discriminant analysis, cluster analysis, ...)	A	A	1

19	Typical DM tools (brief overview of tools: Python, R-Studio, KNIME, Rapid Miner, SPM...)	A	B	3
20	Multidimensional graphical representations	A	A	1
21	Strategies for the selection of training and test data	A	C	3
22	Parameters for evaluating model quality (R-Qd, ROC, BIC, AIC, Confusion Matrix) for training and test data	A	B	3
23	Advantages and disadvantages of cloud computing	A	A	1
24	Overview of standard solutions based on big data and AI (available libraries, platforms, image processing, language processing, ...)	I	A	1
25	Possible uses of more complex DM models (setting recommendations, predictive analytics, prescriptive analytics) with practical examples	I	A	1
26	Basic principles of model deployment and re-training (performance monitoring, AI models, CICD - Continuous Integration Continuous Deployment)	C	C	3

6 Benefits and typical areas of application

1. **Smart Knowledge Picking (SKP)**
   The concept of learning has undergone fundamental change in recent years. The Six Sigma Thinking Ahead working group within the ESSC-D has addressed this topic and included numerous projects, surveys and technical papers in its analysis. The experience of all involved parties, and the presentation and publication of the results provide a clear picture. For acute tasks, it is often necessary to acquire the knowledge/specialised knowledge first. SCP refers to the targeted and task-related collation and utilisation of knowledge.

2. **DMAIC and modern project management approaches**
   Since its introduction, the Six Sigma methodology has lost none of its relevance and flexibility. A wide variety of project management approaches and methods are used, depending on the company's philosophy. Many companies develop their own preferred approach from the available methods. Six Sigma Belts with their methodological knowledge also often work in teams that pursue other philosophies/methods. For the Belt of the future to be able to contribute safely and supportively in this environment, he or she should be familiar with the basic principles of common project management methods.

3. **Basic concepts of digitalisation**
   Project teams, sponsors and senior management in organisations are increasingly faced with the challenges of digitalisation. Ideally, they will also play an active role in shaping these challenges. In both cases, it is necessary for the Belt of the future to be able to correctly classify, evaluate and apply digitalisation terms and approaches.

4. **Possibilities and limitations of inductive statistics and modern data mining**
   In quality and digitalisation projects, as things stand today (2021), many data analysis tasks can be addressed using traditional statistical methods. However, certain tasks, not only in the context of big data, can be handled much faster and more reliably using modern data mining methods. For certain data situations and tasks, data mining algorithms and methods are often the only way to analyse data.

5. **Stakeholders (extended)**
   As a result of digitalisation, optimisations and correlation analyses are carried out more frequently across the entire value chain. This automatically increases the number of parties and people to consider.

6. **Data SIPOC / Process Data Map**
   An important part of data analysis is knowing what specific data is recorded at what point in the process. This allows the process flow to be synchronised with the various data sources ("data understanding").

7. **Extended data collection plan/data catalogue**

   All characteristics are visualised in the process data map and assigned to the process steps. On this basis, the domain knowledge is compiled and categorised in the data collection plan. Relevant here is information such as collection frequency, scale level, specifications, noise and control variables and all other topics that may be relevant to the project. The resulting table can be used as a basis for selecting the right characteristics for modelling and for FMEAs or other classic tools.

8. **Data quality requirements for data mining**

   The discussion of data quality also includes considerations regarding the suitability of the measurement system. Data quality and representativeness also play a decisive role in the context of data mining.

   With digitalization and the associated data volumes, digital transmission errors and compatibility issues are also becoming more common.

   Various tools and methods for data orchestration are required to create an analysable data structure. The target structure is usually readable.

9. **Feature selection for complexity reduction**

   By analysing entire value chains and being able to store large amounts of information, you can quickly obtain a large number of features (columns of variables). This often results in tables with several hundred columns that can be used for correlation models. Machine learning techniques can be used to identify the potentially most and least important features to answer the question.

10. **Extended C&E matrix**

    Once the features have been selected, the results should be validated with experts. On the one hand, the aim is to check whether the most important features are technically plausible. On the other hand, it is important to check that technically expected features have not been unintentionally removed. An extended C&E matrix can be used to have individual features weighted by the experts and, in parallel, by a statistical model. In this way, inconsistencies can be identified, discussed and resolved.

11. **Tools for data preparation (Data Preparation)**

    Creating an analysable data structure takes up most of the time used for data analysis. Transferring analysis data with many to very many characteristics into an analysable structure with "on-board tools" is not recommended. The handling of missing values, different information densities, the calculation of proxies, the aggregation of data and highly correlated characteristics should be practiced. In addition, the way in which data are prepared must be documented in a comprehensible manner and is often also part of the discussion of the results.

    Typical tools include aggregation, pivoting, transposing and binning, as well as methods for dealing with missing values, multicollinearity and outliers.

12. **Basics of ETL/ELT and common tools**

    The ETL (Extract-Transform-Load) process is a series of individual steps that integrate data from different data sources (structured or unstructured) into a data warehouse (DW). The

process is often used to process large volumes of data in the context of big data and business intelligence.

13. **Differences between supervised and unsupervised learning**
In machine learning, there are mainly (but not exclusively) two types of learning: Supervised and Unsupervised Learning. While supervised learning is about "recognising correlations between X and Y" or "predicting the future", unsupervised learning is about discovering and understanding patterns in the available data.

14. **Frequently used data mining methods**
When analysing complex data structures, some data mining methods and procedures are also part of the basic tools of the trade for Six Sigma Belts. These include bootstrapping, labelling, ensemble, and feature selection, for example.

15. **Exploratory data analysis (graphics and descriptive statistics)**
Visualisation of input and output variables and results, including statistical measures, is an essential part of data analysis. For more complex data structures and results, appropriate software and graph types are required to visualise the facts (keywords: Heatmap, Correlation Matrix, Parallel Coordinates and other multidimensional visualisations).

16. **Dealing with unbalanced data (only a few errors, but lots of "good" data)**
When developing correlation models, it is generally advantageous to have balanced data available to train a model. If models are developed on the basis of historical data, these are usually not balanced.
Based on the example "Only a few errors, but a lot of "good" data", the following methods can provide support:
    - Remove "good" data to obtain balanced data.
      (same number of good data as errors)
    - Create artificial errors by copying or slight variation. (As many errors as good)
    - Choose an algorithm that has been optimised for unbalanced data.

Belts should know how to deal with this and develop a solid training data set.

17. **Frequently used data mining algorithms**
Belts should know the most common algorithms and typical examples of their use in practice, to be able to name them and possibly apply them to specific problems. These include decision trees, neural networks, Naive Bayes, SVM, K-Nearest, K-Means, Principal Component Analysis and Lasso.

18. **Advanced multivariate analysis**
The classic Six Sigma Green Belt or Six Sigma Black Belt training ends with model development using simple and multiple regression. To benefit from more complex data structures or to simplify them, multivariate methods should be known and applied. These include cluster analysis, dendrograms, principal component analysis, discriminant analysis and others.

19. **Typical data mining tools**
Typical software packages or free scripting languages are introduced and used in the course. These include, for example, Python, R-Studio, KNIME, Rapid Miner and SPM.

20. **Multidimensional graphical representations**
For more complex data structures and results, appropriate software and graph types are required to analyse the facts (keywords: Heat maps, 3D plots, contour plots, correlation matrices, parallel coordinates, and others).

21. **Strategies for the selection of training and test data**
More complex data structures often also contain many observations (rows). Classical statistical methods, such as t-tests, but also variance analysis, mainly indicate statistical significance. The "p-values" often do not provide any useful information here. It is therefore common practice to form a subset of the available data. Statistical analysis is performed on one subset (training data) and cross-validation is performed on the other subset (test data). The quality of the cross-validation detects possible "overfitting" and provides another criterion for the usability of the statistical model. There are different philosophies for the division into training and test data, which a belt should know and be able to apply.

22. **Parameters for evaluating the model quality for training and test data**
Belts should be able to evaluate and critically question the quality and therefore applicability of statistical models. Important parameters include R-Qd, ROC, BIC, AIC and Confusion Matrix, the advantages and disadvantages of which are explained.

23. **Advantages and disadvantages of cloud computing**
Different systems have different strengths, depending on the application and the data availability required. For example, when it comes to high availability and managing sensitive data, cloud systems are often ruled out for technical reasons. Whenever local or cloud systems are used, the legal aspects must always be considered. However, when it comes to scalability of computing power, cloud systems are much more flexible and cheaper, as you only pay for the computing time you use.

24. **Overview of standard solutions based on big data and AI**
In the field of big data and AI, there are many open-source solutions that can be used effectively and practically for a belt. In general, it is recommended to use standard solutions that have already proven themselves in the Six Sigma environment. These include libraries for data analysis, image and speech recognition.

25. **Possible uses of more complex data mining models**
Participants will be presented with real-world examples of predictive and prescriptive analytics models. Identify success factors in data preparation and model parameterisation.

26. **Basic principles of model deployment and re-training**
When a model is used regularly in operations, rules and principles need to be defined to ensure that the model is up-to-date and therefore performs well. A continuous improvement process is established for the models, often following the CICD (Continuous Integration Continuous Deployment) pipeline.

7 Possible training program

The following suggestion can be used for inspiration when developing a training program on these topics.

Day 1:

Morning (focus on *refresher*):

- Exchange of experiences, expectations, and goals
- Exploratory data analysis and data preparation
- Developing models as part of inductive statistics (correlation, regression and ANOVA
- Evaluating model and forecast quality with confidence intervals

Afternoon (focus on *terms, possibilities, and limits*):

- Industry 4.0 in production processes - transforming data into benefits
- Typical data structures and the resulting challenges
- Typical methods, approaches and terms used in data mining
- A process model for structured data analysis (e.g. CRISP-DM)
- Model Lifecycle Management - "SPC of the future"

Day 2:

Morning (focus on *data preparation and data exploration*):

- Presentation of case study(s)
- Exploratory data analysis and data preparation
- Ensure that the task and data are understood (business and data understanding)
- Strategies for developing training and test data sets

Afternoon (focus on *supervised learning → regression models*):

- Developing regression models and decision trees
- Validation of models with the help of cross-validation
- Model selection procedure
- Feature selection, bootstrapping and others
- Transferring the results/findings into a stringent statement logic (clear recommendation for action/clear and simple presentation of results)
- Model lifecycle management for the jointly developed model

Day 3:

Morning (focus on *supervised learning → classification models*):

- Presentation of case studies
- Exploratory data analysis and data preparation
- Ensure that the task and data are understood (business and data understanding)
- Strategies for developing training and test data sets, especially for unbalanced targets

Afternoon (focus on *unsupervised learning*):

- Insight into multivariate data analysis to identify clusters
- Classic data mining approaches for the identification of clusters
- Validation of findings with the help of cross-validation
- Transferring the results/findings into a stringent statement logic

"*Learning is like rowing against the current. If you stop, you drift backwards.*"
(Laozi, Chinese philosopher, 6th century BC)