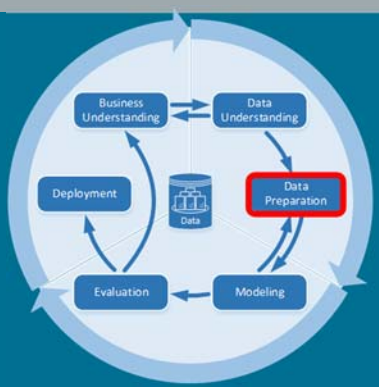
Fraunhofer
IESE

FRAUNHOFER INSTITUTE FOR EXPERIMENTAL SOFTWARE ENGINEERING IESE



DATA PREPARATION – TACKLE THE MOST EFFORT-PRONE PHASE IN DATA PROJECTS

The most effort-consuming phase in data science projects is data preparation. No standard procedure that covers all potential data preparation issues exist. In this seminar, you learn how to increase the efficiency of data preparation in order to gain faster insights into your data using data analytics.

From a process point of view, the CRoss-Industry Standard Process for Data Mining (CRISP-DM) describes six major steps for any data analysis project. After having gained Business Understanding, we need to identify and semantically understand required data (Data Understanding). This requires domain knowledge as well as data engineering and data analysis knowledge. Therefore, Data Understanding is the starting point for data ingestion and Data Preparation.

Fraunhofer-Institut für Experimentelles Software Engineering IESE

Fraunhofer-Platz 1

67663 Kaiserslautern

Kontakt

Dr. Andreas Jedlitschka

Tel. +49 631 6800-2260

andreas.jedlitschka@iese.fraunhofer.de

www.iese.fraunhofer.de

The task of the **Data Preparation** step is to extract and prepare required data from their sources through transformation, cleaning, filtering, missing value treatment, etc..

Each analysis technique pose concrete requirements on the data. During the course of a project experts iteratively adjust which data is important, how data needs to be prepared and which data lead to better results. In addition, it is essential to assess the quality of the data. It is among the most critical steps of any data-driven project, be it about classical data analytics or artificial intelligence (garbage in, garbage out).

Data analysts repeatedly perform the Data Preparation phase due to the explorative character of data analyses as well as to the strong influence of the Data Preparation on the results of the analysis. The analysis' goal, the planned analysis approach as well as technical aspects influence the technology stack used for data ingestion and Data Preparation.

Data scientist spend a lot of time solely with Data Preparation (up to 70% of the project effort).

Within this hands-on Seminar, you will learn how to prepare the data for your data analysis projects. Based on concrete examples and on our experience from many projects, we show caveats and possible solutions for Data Preparation. You learn how to realize data preparation steps using Jupyter Notebooks.

Language: English or German

Target Groups:

Data Engineers and Data Scientist from companies that use or develop data-driven approaches.

Contents (1 Day):

The seminar covers the following topics:

- ➔ Data Preparation
- ➔ Data Quality Assessment and Mitigation Strategies
- ➔ Hands-on with Jupyter Notebooks

Learning Objectives:

You will get an understanding of the necessity of the Data Preparation phase in CRISP-DM and get to know the methods and tools to assess data quality and learn how to mitigate commonly available issues.