# Target audience

Students, professionals and the general public who are interested in obtaining a … (Sigrun)

# Prerequisites

The reader should have a foundational understanding of basic mathematical concepts. Additionally, familiarity with the basics of Python is crucial (see [Microcredit Python Introduction](https://git.rz.tu-bs.de/ifn-public/ki4all/python-introduction)) and Python libraries like NumPy and Pandas for data analysis, along with an introductory knowledge of machine learning principles (as outlined in the [Microcredit Machine Learning Basics](https://git.rz.tu-bs.de/ifn-public/ki4all/machine-learning-introduction). It is essential to understand the composition and significance of training data, alongside fundamental statistical concepts such as classification, features (which correspond to data columns), samples (which correspond to data rows), labels (which form the vector indicating the category of each sample), as well as statistical distributions, including the normal and log-normal distributions. Knowledge of effect size and clustering techniques is also vital for data analysis and interpretation. We recommend consulting [THIS LITERATURE] for comprehensive coverage of these topics.

Moreover, an understanding of feature engineering, selection, extraction, and the importance of these processes is crucial for the optimization of machine learning models. Additionally, a basic knowledge of biomarkers, the application areas of artificial data, and proficiency in managing data formats such as CSV will significantly benefit practitioners in the field of machine learning and data science.

----------------------------------------------

The reader should have some understanding of basic mathematical topics. Furthermore, the reader should also be familiar with the basics of Python (see [Microcredit Python Introduction](https://git.rz.tu-bs.de/ifn-public/ki4all/python-introduction)), machine learning (see [Microcredit Machine Learning Basics](https://git.rz.tu-bs.de/ifn-public/ki4all/machine-learning-introduction)), training data and basic knowledge of statistics like classification, features (columns), samples (rows), labels (vector), statistical distributions, normal and lognormal distribution, effect size and clustering. We recommend THIS LITERATURE. Furthermore, you should have some knowledge of feature engineering, selection, extraction and importance. Additionally, you should know about biomarkers, application areas of artificial data and CSV.

--------------------------------------------

Python (<https://git.rz.tu-bs.de/ifn-public/ki4all/python-introduction>)  
Machine Learning (<https://git.rz.tu-bs.de/ifn-public/ki4all/machine-learning-introduction>)  
Classification (<https://en.wikipedia.org/wiki/Statistical_classification>)

labels (vector) (<https://www.ibm.com/topics/data-labeling>)  
Samples (rows) [(https://en.wikipedia.org/wiki/Sampling\_(statistics)]((https:/en.wikipedia.org/wiki/Sampling_(statistics)))  
Features (columns) (<https://www.datarobot.com/wiki/feature/)>  
Statistical distributions [(https://en.wikipedia.org/wiki/Probability\_distribution]((https:/en.wikipedia.org/wiki/Probability_distribution))   
Normal distribution (<https://en.wikipedia.org/wiki/Normal_distribution>)   
Log-normal distribution (<https://en.wikipedia.org/wiki/Log-normal_distribution>)  
Effect size (<https://loonylabs.org/2021/03/01/effect-size-in-statistics/>)

Clustering/ Cluster

Feature extraction (<https://www.mathworks.com/discovery/feature-extraction.html>)  
Feature engineering [(https://www.techtarget.com/searchdatamanagement/definition/feature-engineering]((https:/www.techtarget.com/searchdatamanagement/definition/feature-engineering)) (difficult to find)  
Feature importance (<https://builtin.com/data-science/feature-importance>)

Feature selection (<https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/>)

training data (matrix) (<https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets>)

Biomarkers (biological background) (<https://en.wikipedia.org/wiki/Biomarker>)

Application areas of artificial data  
  
CSV (https://en.wikipedia.org/wiki/Comma-separated\_values)

# Learning Goals

After reading this document, the reader should be able to:

* Understand the purpose of synthetic data (benchmarking of machine learning algorithms).
* Understand the structure of synthetic data.
* Be able to generate (needed) data oneself.

# A note to the reader

Explaining at full length some terms and concepts regarding machine learning is beyond the scope of this document. They will, however, in the majority be explained briefly. The reader is, however, referred to further textbooks or other material in case a more detailed understanding is desired.

# 1 Introduction

In order to develop new methods or to compare existing methods for feature selection, reference data with known dependencies and importance of the individual features are needed. This data generator can be used to simulate biological data for example artificial high throughput data including artificial biomarkers. Since commonly not all true biomarkers and internal dependencies of high-dimensional biological datasets are known with certainty, artificial data **enables to know the expected outcome in advance**. In synthetic data, the feature importances and the distribution of each class are known. Irrelevant features can be purely random or belong to a pseudo-class. Such data can be used, for example, to make random effects observable.