# 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.

# 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.

# 2 Data structure

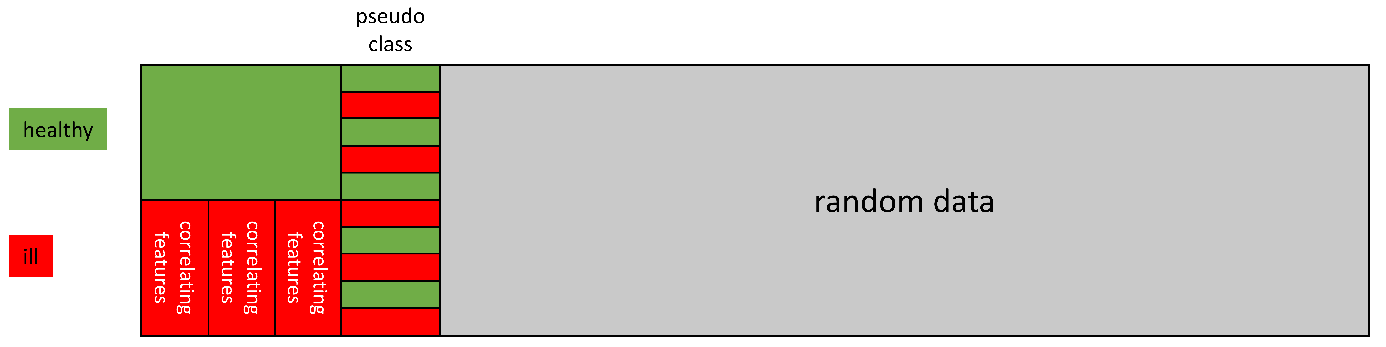
## 2.1 Different parts of the data set

The synthetic-data-generator produces data sets consisting of up to three main parts:

1. **Relevant features** belonging to an artificial class (for example artificial biomarkers)
2. [optional] **Pseudo-classes** (for example a patient's height or gender, which have no association with a particular disease)
3. [optional] **Random data** representing the features (for example biomarker candidates) that are not associated with any class.

The number of artificial classes is not limited. Each class is generated individually and then combined with the others. In order to simulate artificial biomarkers in total, all individual classes have the same number of features in total.

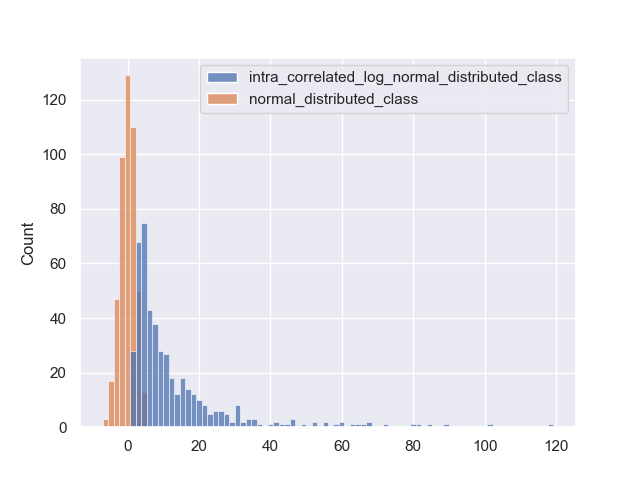
This is an example of simulated binary biological data including artificial biomarkers:

[](https://github.com/sigrun-may/artificial-data-generator/blob/main/docs/source/imgs/artificial_data.png)

## 2.2 Data distribution and effect sizes

For each class, either the **normal distribution or the log normal distribution** can be selected. The different **classes can be shifted** to regulate the effect sizes and to influence the difficulty of data analysis.

The normally distributed data could, for example, represent the range of values of healthy individuals. In the case of a disease, biological systems are in some way out of balance. Extreme changes in values as well as outliers can then be observed ([Concordet et al., 2009](https://doi.org/10.1016/j.cca.2009.03.057)). Therefore, the values of a diseased individual could be simulated with a lognormal distribution.

Example of log-normal and normal distributed classes:

Effect size is a statistical concept that measures the strength of the relationship between two variables in a study. Unlike significance tests, which tell us if a relationship exists, effect size tells us how strong that relationship is. It helps us understand the practical importance of research results beyond mere statistical significance. For example, a large effect size means a substantial relationship or difference between groups, whereas a small effect size indicates a minor relationship.

To better understand this, here is an exemplary graphical representation of the effect size and its impact.

Ein Bild, das Text, Reihe, Diagramm, Schrift enthält.

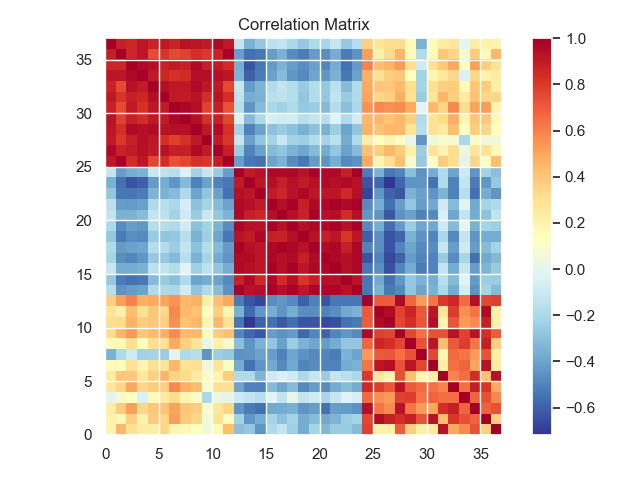
Automatisch generierte Beschreibung

Quelle: <https://loonylabs.org/2021/03/01/effect-size-in-statistics/> (Eigene Darstellung!)

## 2.3 Correlations

**Intra-class correlation can be generated for each artificial class**. Any number of groups containing correlated features can be combined with any given number of uncorrelated features.

However, a high correlation within a group does not necessarily lead to a high correlation to other groups or features of the same class. An example of a class with three highly correlated groups but without high correlations between all groups:

[](https://github.com/sigrun-may/artificial-data-generator/blob/main/docs/source/imgs/corr_3_groups.png)

It is probably likely that biomarkers of healthy individuals usually have a relatively low correlation. On average, their values are within a usual "normal" range. In this case, one biomarker tends to be in the upper normal range and another biomarker in the lower normal range. However, individually it can also be exactly the opposite, so that the correlation between healthy individuals would be rather low. Therefore, the **values of healthy people could be simulated without any special artificially generated correlations**.

In the case of a disease, however, a biological system is brought out of balance in a certain way and must react to it. For example, this reaction can then happen in a coordinated manner involving several biomarkers, or corresponding cascades (e.g. pathways) can be activated or blocked. This can result in a **rather stronger correlation of biomarkers in patients suffering from a disease**. To simulate these intra-class correlations, a class is divided into a given number of groups with high internal correlation (the respective strength can be defined).

# 3 Pseudo-classes

One option for an element of the generated data set is a pseudo-class. For example, this could be a patient's height or gender, which are not related to a specific disease.

The generated pseudo-class contains the same number of classes with identical distributions as the artificial biomarkers. But after the generation of the individual classes, all samples (rows) are randomly shuffled. Finally, combining the shuffled data with the original, unshuffled class labels, the pseudo-class no longer has a valid association with any class label. Consequently, no element of the pseudo-class should be recognized as relevant by a feature selection algorithm.

# 4 Random features

The artificial biomarkers and, if applicable, the optional pseudo-classes can be combined with any number of random features. Varying the number of random features can be used, for example, to analyze random effects that occur in small sample sizes with a very large number of features.