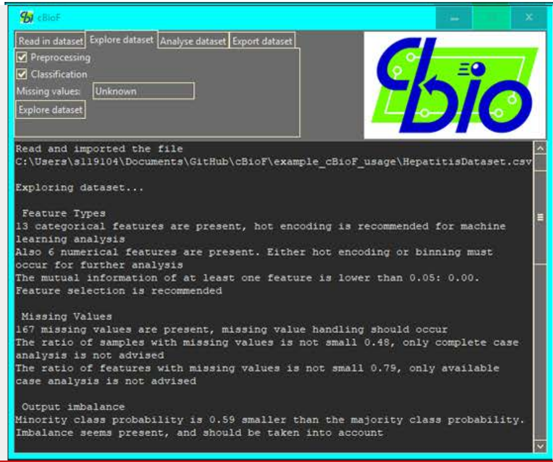
TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY



☑ Preprocessing

☑ Classification

Missing values: Unknown

```
Read and imported the file
C:\Users\s119104\Documents\GitHub\cBioF\example_cBioF_usage\HepatitisDataset.csv

Exploring dataset...

 Feature Types
13 categorical features are present, hot encoding is recommended for machine
learning analysis
Also 6 numerical features are present. Either hot encoding or binning must
occur for further analysis
The mutual information of at least one feature is lower than 0.05: 0.00.
Feature selection is recommended

 Missing Values
167 missing values are present, missing value handling should occur
The ratio of samples with missing values is not small 0.48, only complete case
analysis is not advised
The ratio of features with missing values is not small 0.79, only available
case analysis is not advised

 Output imbalance
Minority class probability is 0.59 smaller than the majority class probability.
Imbalance seems present, and should be taken into account
```

A Computational Biology Framework

A data analysis tool to support biomedical engineers in their research

Tim Beishuizen, October 23rd 2018

Supervisors: Dr. Dragan Bošnački
Dr. Ir. Joaquin Vanschoren
Prof. Dr. Peter Hilbers

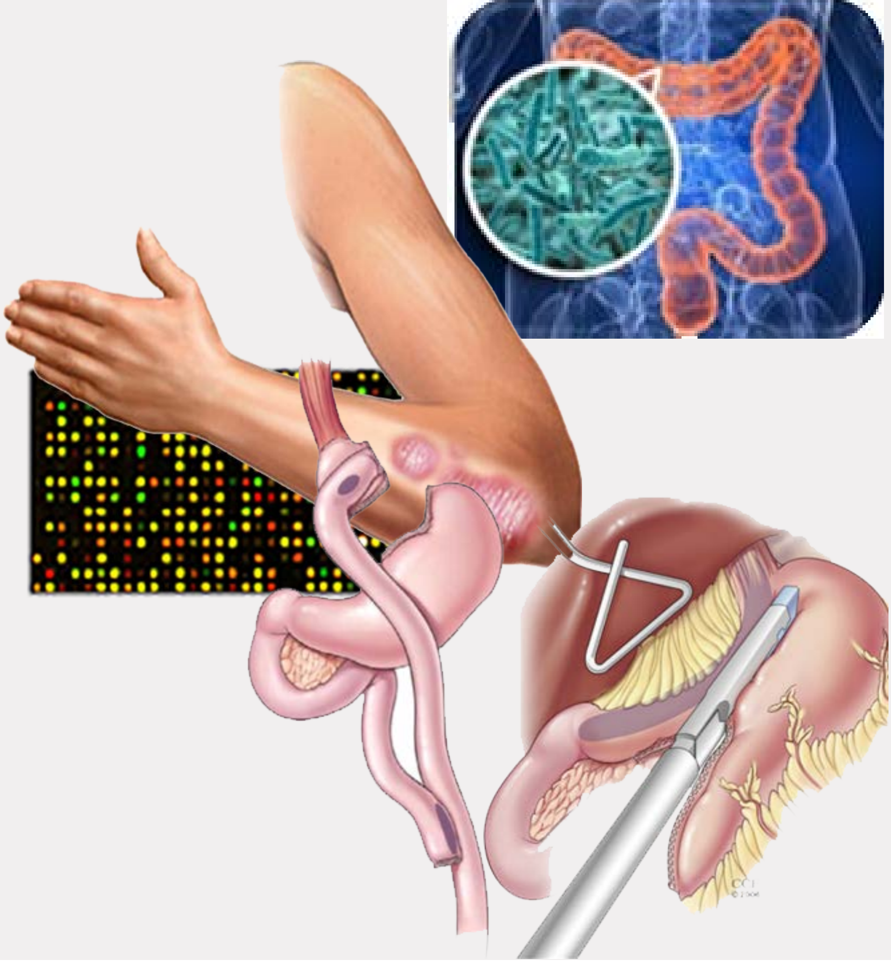
University of Technology Eindhoven, department Biomedical Engineering and department Computer Science and Engineering

Contents

- Introduction
  - Research questions
- Project description
- Research lines
  - Feature selection
  - Missing value handling
  - Dataset exploration
- Framework

Introduction

- Biomedical data
- Processing
- Analysis
- Framework



Main research question

What aspects are of importance to be included in a framework for Biomedical Engineers for more efficient data analysis?

Datasets	Preprocessing methods
Data issues	Analysis methods

Sub-questions

- What feature selection methods show the best performance and as such should be added to the framework?

- What missing value handling methods show the best performance and as such should be added to the framework?

- Which initial analyses help find suitable preprocessing and data analysis algorithms?

Project description

- Statistics
- Machine learning

- Data matrix
  - Samples (s)
  - Features (f)
  - Output (y)
  - Classification

- Python

$$\begin{array}{c c}
 & \begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \end{array} \\
\begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} & \left(\begin{array}{cccc} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{array}\right) \left(\begin{array}{c} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{array}\right)
\end{array}$$

Datasets

- Microarray datasets
  - Gene expression
  - Many features
  - Big datasets

- Mass spectrometry datasets
  - Protein mass
  - Many features
  - Differences between features

- Clinical datasets
  - Data out of a clinic
  - Missing values
  - Combining data issues

- Survey datasets
  - Questionnaire data
  - Missing values
  - Bias due to subjectivity

Feature selection

Feature selection

What feature selection methods show the best performance and as such should be added to the framework?

Feature selection methods

- Filter methods
- Wrapper methods
- Embedded methods

Datasets

- Microarray
- Mass spectrometry

Filter methods

- Separate features

Filter methods

- Separate features
- Evaluation function
- Selection
- Fast and simple

$$\begin{array}{cc} & \begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} & \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \\ x_{41} & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \end{array}$$

Wrapper methods

- Feature sets

Wrapper methods

- Feature sets
- Evaluation function
- Methods
  - Forward selection
  - Backward selection
  - Combinations
- Feature dependencies
- Longer computation time

$$\begin{array}{cc} & \begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} & \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \\ X_{51} & X_{52} & X_{53} & X_{54} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \end{array}$$

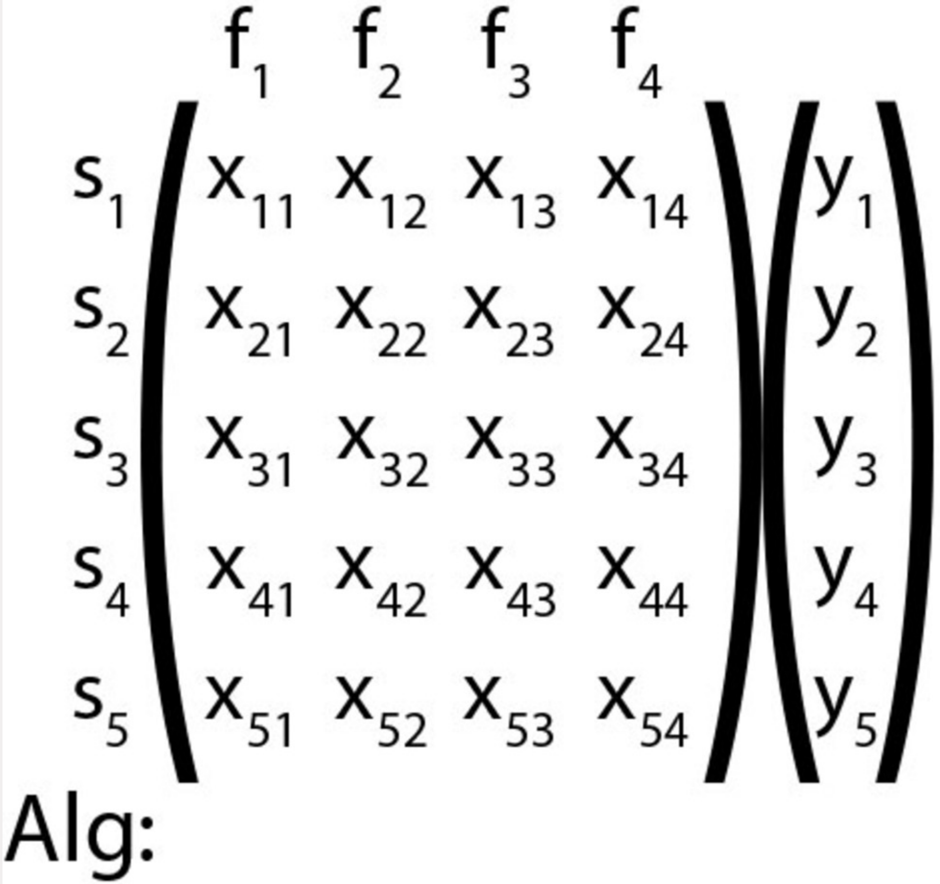
0.0

Embedded methods

- Machine learning algorithms

Embedded methods

- Machine learning algorithms
- Weights
- Selection
- Low computation time
- Some feature dependencies



$$\begin{array}{cccccc} & f_1 & f_2 & f_3 & f_4 & \\ s_1 & x_{11} & x_{12} & x_{13} & x_{14} & y_1 \\ s_2 & x_{21} & x_{22} & x_{23} & x_{24} & y_2 \\ s_3 & x_{31} & x_{32} & x_{33} & x_{34} & y_3 \\ s_4 & x_{41} & x_{42} & x_{43} & x_{44} & y_4 \\ s_5 & x_{51} & x_{52} & x_{53} & x_{54} & y_5 \end{array}$$

Alg:

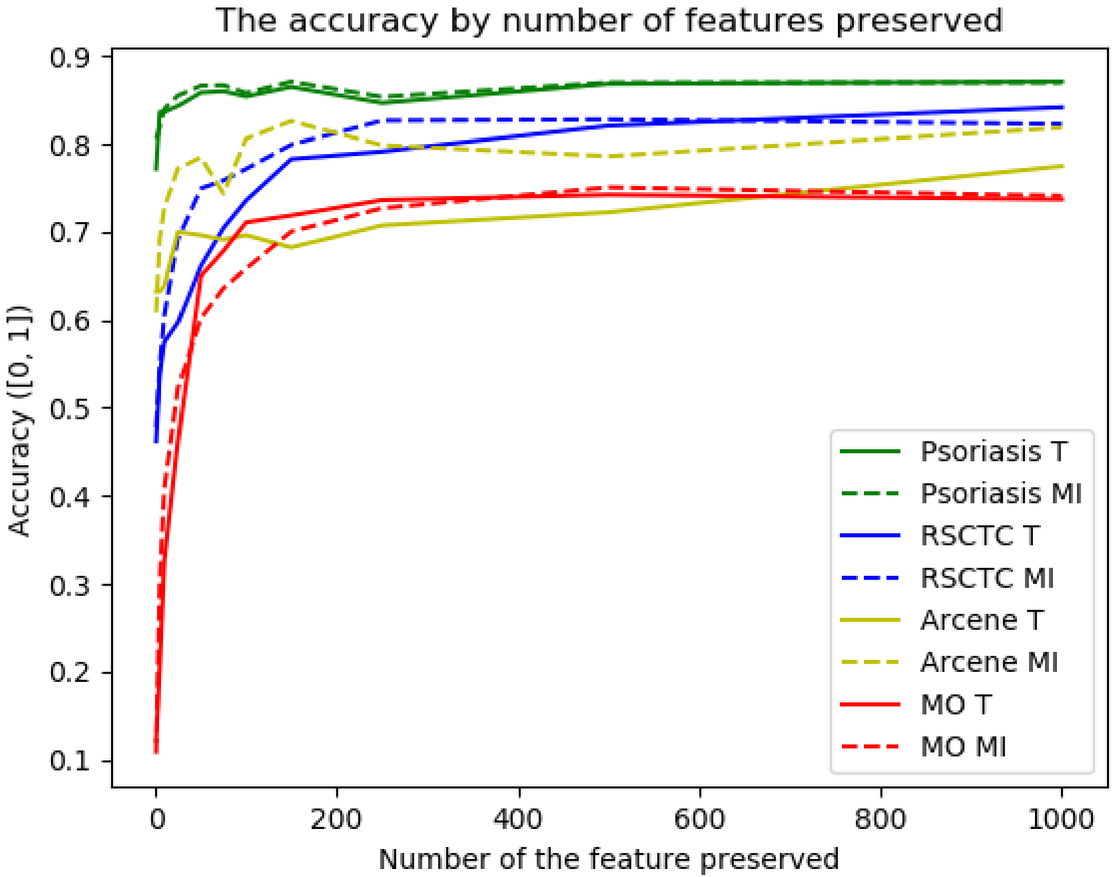
Experiment 1

Filter methods

- Evaluation functions
  - T-test/ANOVA
  - Mutual Information
- Feature selection
  - Minimum 1
  - Maximum 1000
- Quality testing
  - Logistic regression accuracy ([0, 1])

Experiment 1

- Evaluation functions
- 200 features

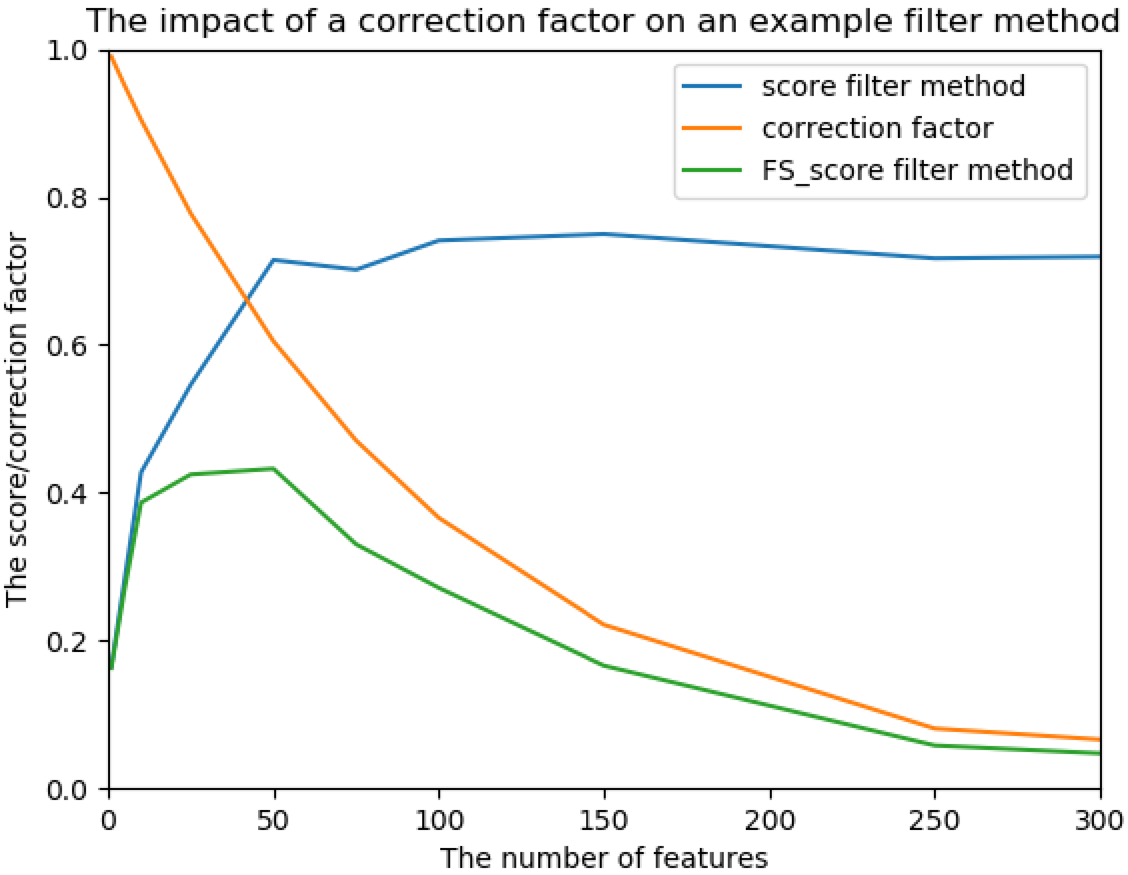


FS_score

- Combination
  - Prediction score
  - Features preserved
- Correction factor

$$FS\_score = score \ * \beta^{\#features}$$

  - β=0.99
- Optimum
- Prediction scores



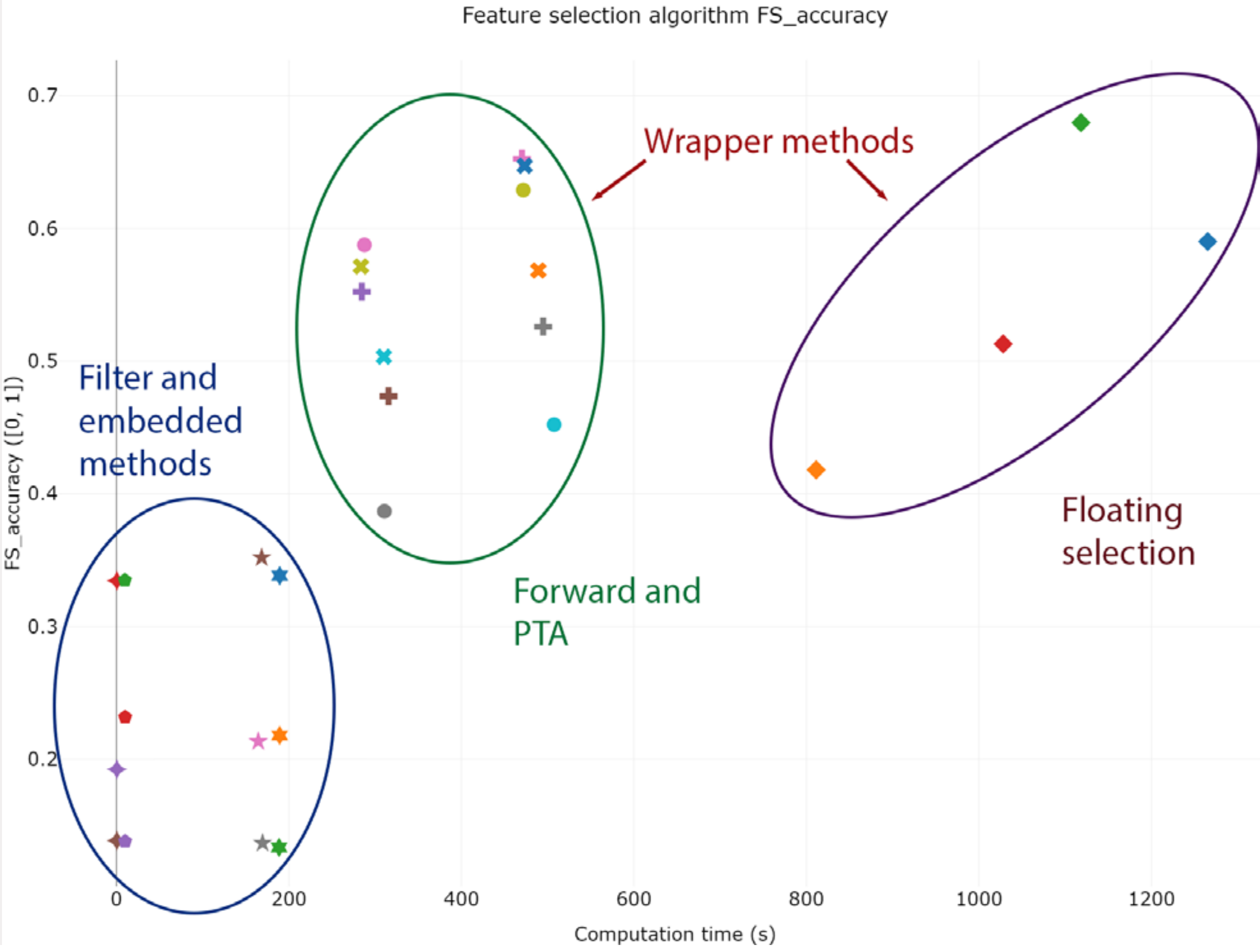
Experiment 2

Feature selection methods quality

- Quality testing
  - FS_score
  - Logistic regression accuracy ([0, 1])

Experiment 2

- Filter and embedded methods faster
- Wrapper methods better



Wrapper methods

Filter and embedded methods

Forward and PTA

Floating selection

Automated Machine learning

- Search for best analysis pipeline
- Preprocessing algorithms
- Machine learning algorithms
- Hyperparameters

Genetic Algorithms

- Solutions

- Evolving solutions
  - Mutations
  - Crossovers

- Selecting solutions

Tree-based Pipeline Optimization tool (TPOT)

- Genetic algorithms
- Searches for and stores best pipeline
- Issues with many features

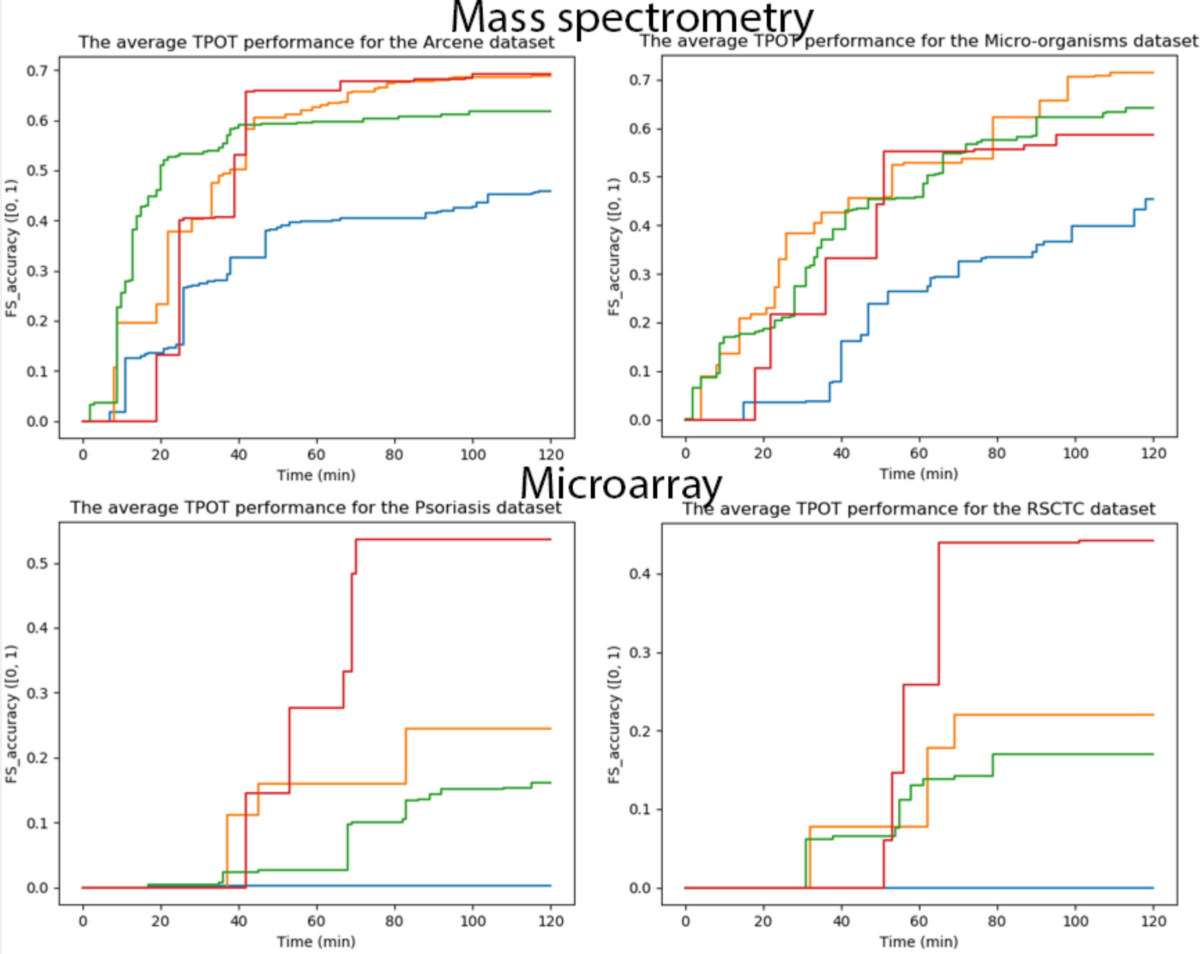
Experiment 3

TPOT for many features

- FS_score
  - Accuracy ([0, 1])
- 2 Additions
  - Bias towards feature selection
  - New set feature selection algorithms

Experiment 3

- Improved
- Microarray both best
- Mass spectrometry debatable



Legend

- Blue No addition
- Orange New algorithms
- Green Feature selection bias
- Red Both additions

Missing Value Handling

Missing Value Handling

What missing value handling methods show the best performance and as such should be added to the framework?

Missing value handling methods

- List Deletion
- Single imputation
- Multiple imputation

Datasets

- Clinical
- Survey

List deletion

- Available Case Analysis (ACA)
- Complete Case Analysis (CCA)
    - Weighted Case Analysis (WCA)

$$\begin{array}{cc} & \begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \end{array} \\ \begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} & \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & \cdots & x_{24} \\ x_{31} & x_{32} & \cdots & \cdots \\ \cdots & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \end{array}$$

Single imputation

Single imputation

- Mean imputation
  - Median
  - Mode
- Hot deck imputation

$$
\begin{array}{c}
 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5
\end{array}
\begin{array}{cccc}
f_1 & f_2 & f_3 & f_4
\end{array}
$$

$$
\begin{array}{c|cccc|c}
 & f_1 & f_2 & f_3 & f_4 & \\
\hline
s_1 & x_{11} & x_{12} & x_{13} & x_{14} & y_1 \\
s_2 & x_{21} & x_{22} & x_{13} & x_{24} & y_2 \\
s_3 & x_{31} & x_{32} & x_{53} & x_{24} & y_3 \\
s_4 & x_{31} & x_{42} & x_{43} & x_{44} & y_4 \\
s_5 & x_{51} & x_{52} & x_{53} & x_{54} & y_5
\end{array}
$$

Single imputation

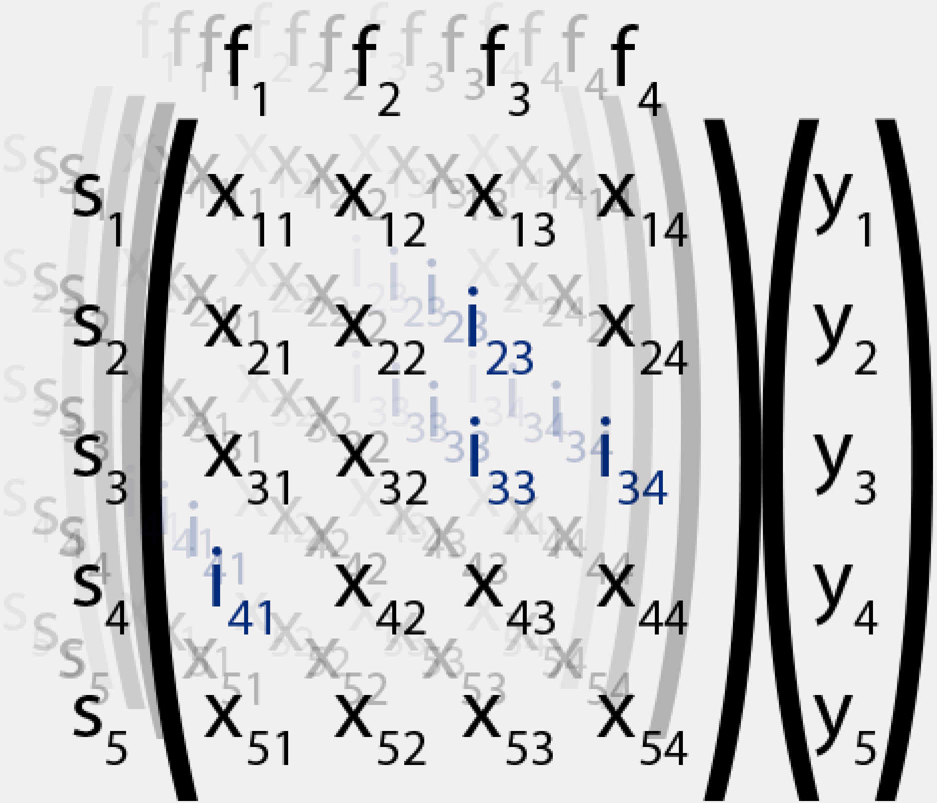
- Mean imputation
  - Median
  - Mode
- Hot deck imputation

- Advanced imputation
  - Regression imputation
  - Nearest neighbour imputation (kNN)

$$\begin{array}{cc}
 & \begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \end{array} \\
\begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} & \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & \cdots & x_{24} \\ x_{31} & x_{32} & \cdots & \cdots \\ \cdots & x_{42} & x_{43} & x_{44} \\ x_{51} & x_{52} & x_{53} & x_{54} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}
\end{array}$$

Multiple Imputation

- Multiple datasets
- Multiple Imputation Chained Equations (MICE)



$$\begin{array}{c|cccc} & f_1 & f_2 & f_3 & f_4 \\ \hline s_1 & x_{11} & x_{12} & x_{13} & x_{14} \\ s_2 & x_{21} & x_{22} & i_{23} & x_{24} \\ s_3 & x_{31} & x_{32} & i_{33} & i_{34} \\ s_4 & i_{41} & x_{42} & x_{43} & x_{44} \\ s_5 & x_{51} & x_{52} & x_{53} & x_{54} \end{array} \quad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

Experiment 1

Feature distributions

  - Mean
  - Standard deviation
- T-test similarity

Results

- > 15% missing values removed

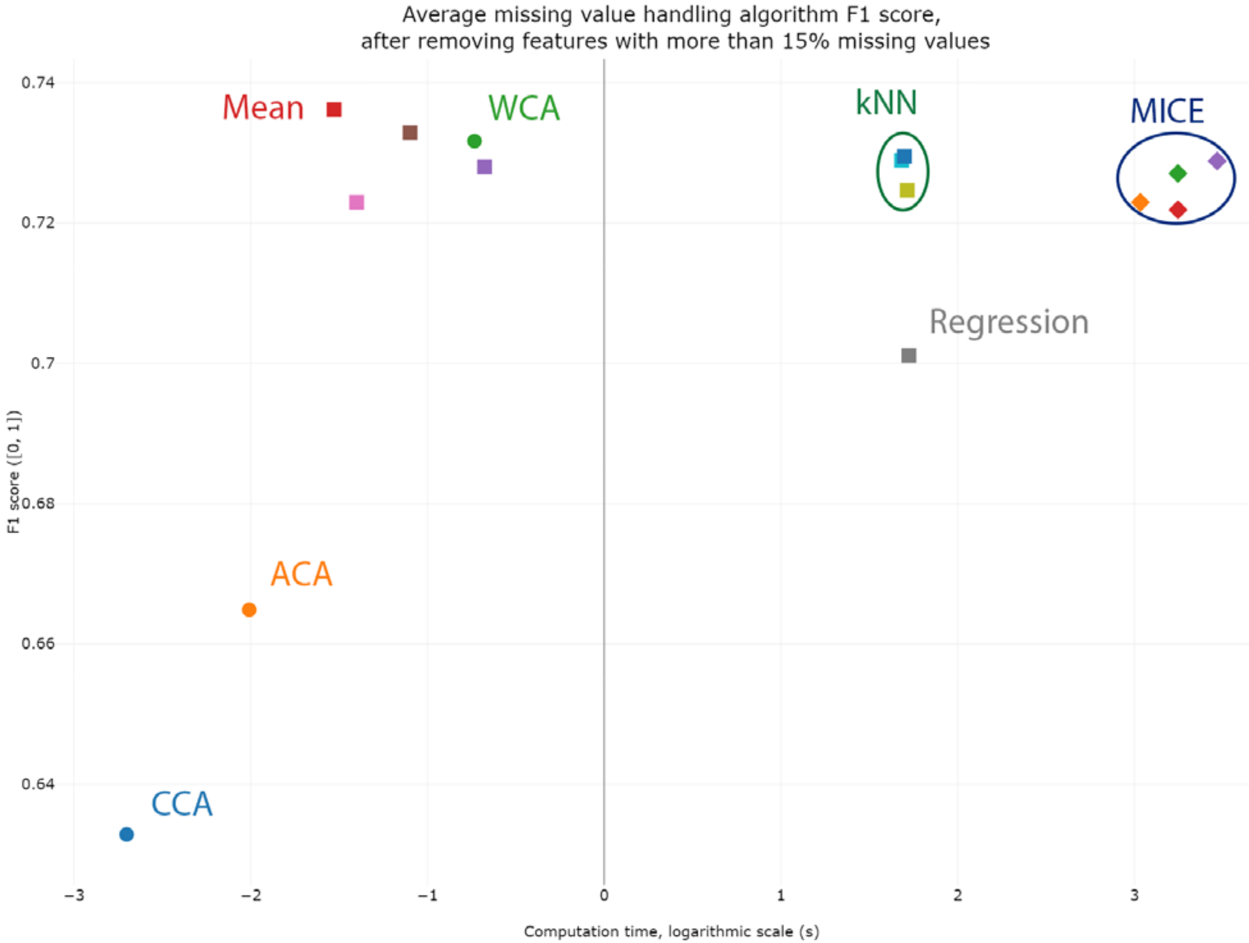
Experiment 2

Missing value handling quality

- >15% missing values removed
- Quality output
  - Logistic regression F1-score ([0, 1])

Experiment 2

- List Deletion
- Mean Imputation
- Regression, kNN and MICE



Dataset exploration

Dataset exploration

Which initial analyses help find suitable preprocessing and data analysis algorithms?

Exploration

- Dataset issues
- Meta-features
- Package *metalearn*

Datasets

- Clinical
- Mass spectrometry

Analysis 1

Dataset exploration with package *metalearn*

- Issues <> meta-features

Result

- Issues recognizable
  - Multicollinearity
  - Feature irrelevance
- Needs clarification

Analysis 2

Dataset exploration with package *metalearn*

- *metalearn* addition
  - meta-features
  - Outlier detection
  - Plots

Results

- All issues explained
- Better understanding

Framework cBioF

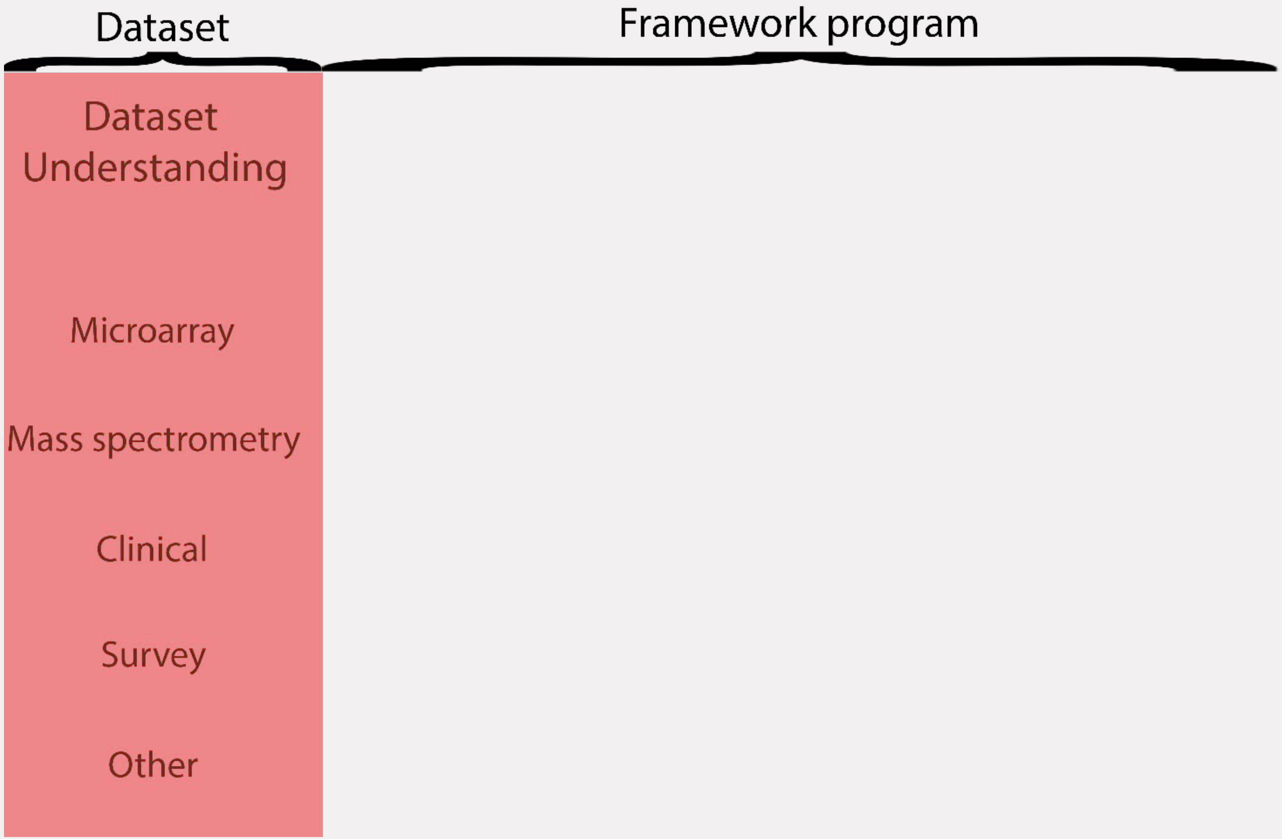
Framework cBioF

What aspects are of importance to be included in a framework for Biomedical Engineers for more efficient data analysis?

Phases

- Datasets
- Exploration
- Preprocessing
- Analysis

cBioF layout



Dataset

Framework program

Dataset Understanding

Microarray

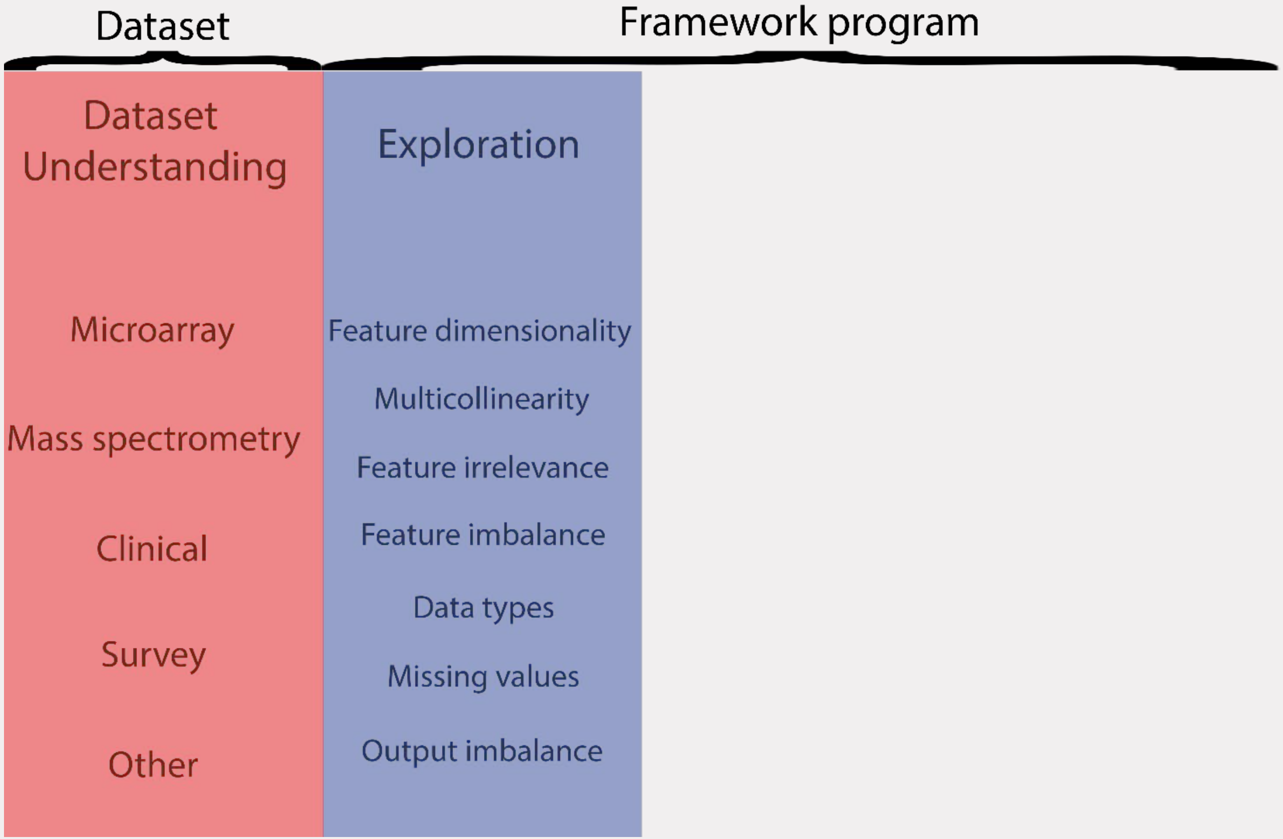
Mass spectrometry

Clinical

Survey

Other

cBioF layout



Dataset

Framework program

Dataset Understanding

Microarray

Mass spectrometry

Clinical

Survey

Other

Exploration

Feature dimensionality

Multicollinearity

Feature irrelevance

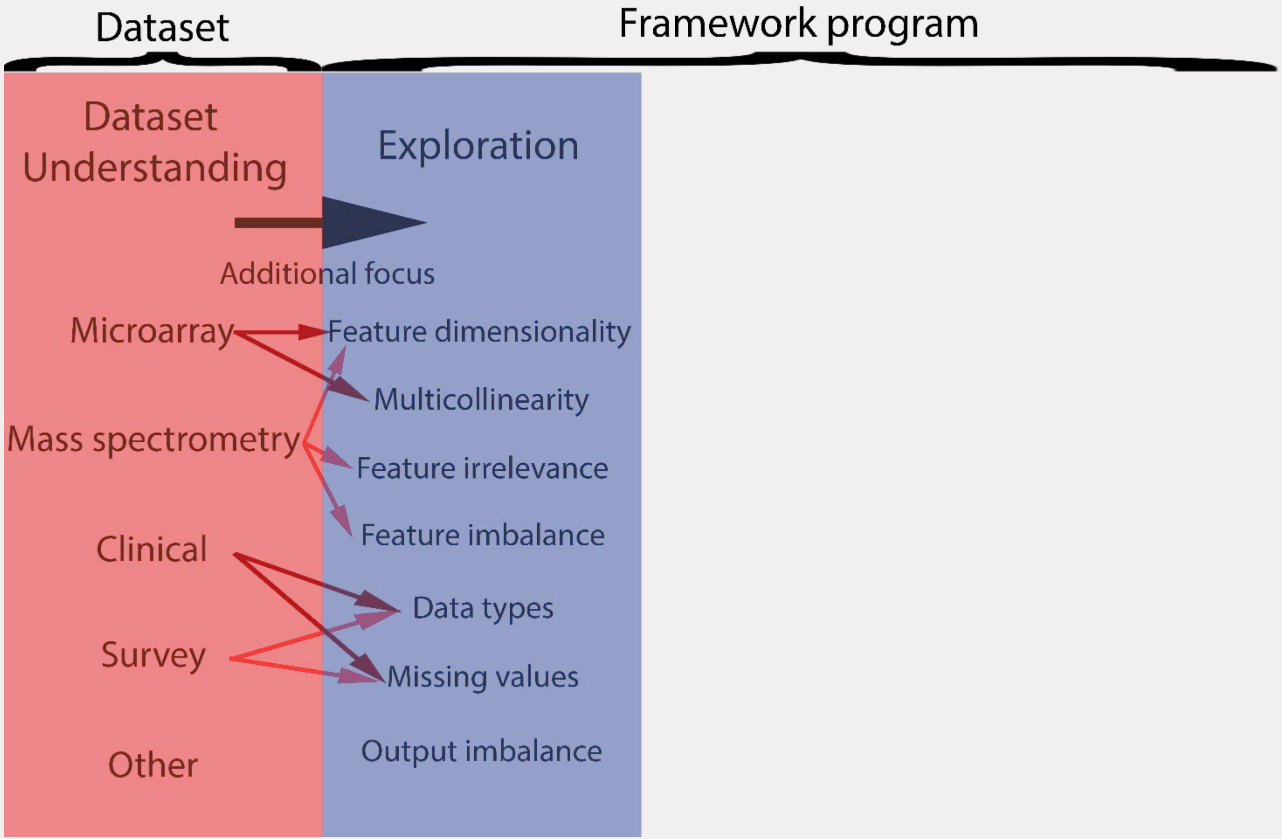
Feature imbalance

Data types

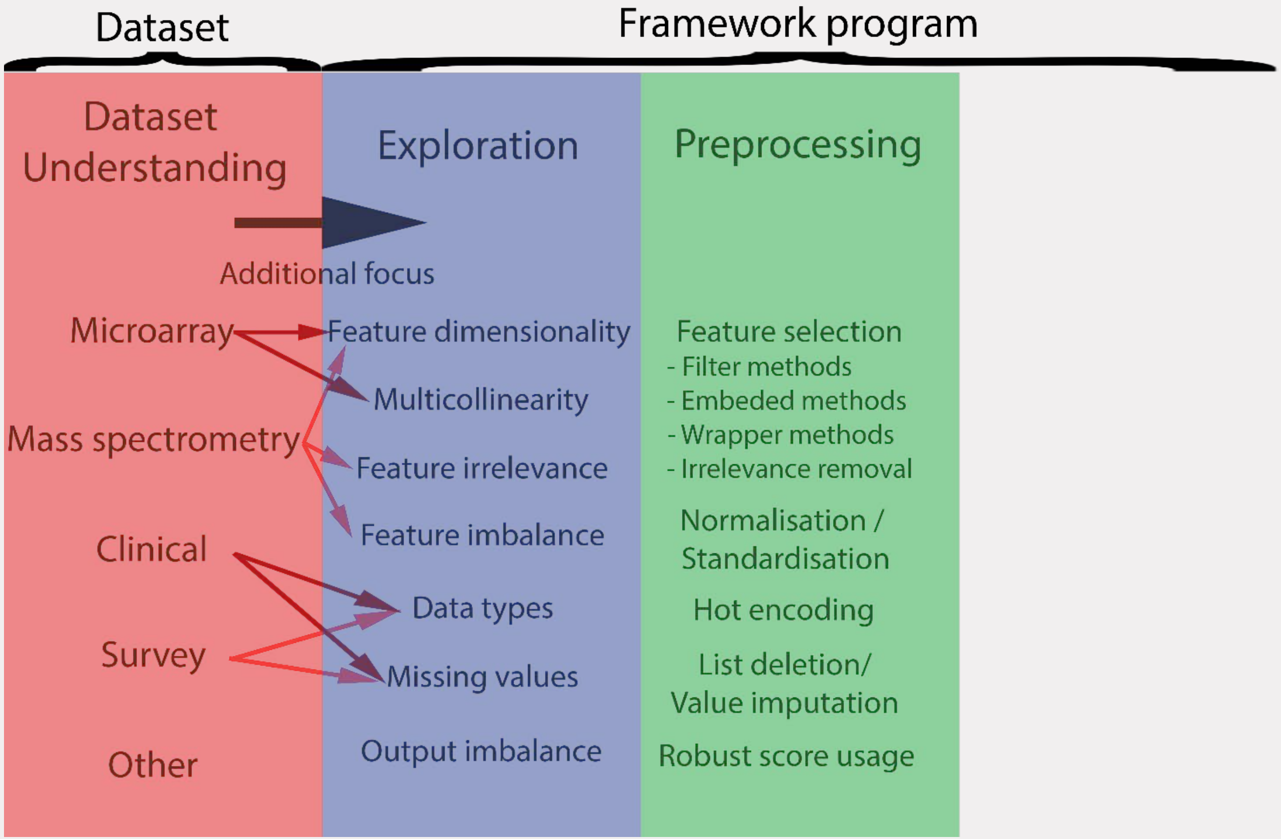
Missing values

Output imbalance

cBioF layout



cBioF layout



Dataset

Framework program

Dataset Understanding

Additional focus

Microarray

Mass spectrometry

Clinical

Survey

Other

Exploration

Feature dimensionality

Multicollinearity

Feature irrelevance

Feature imbalance

Data types

Missing values

Output imbalance

Preprocessing

Feature selection
- Filter methods
- Embeded methods
- Wrapper methods
- Irrelevance removal

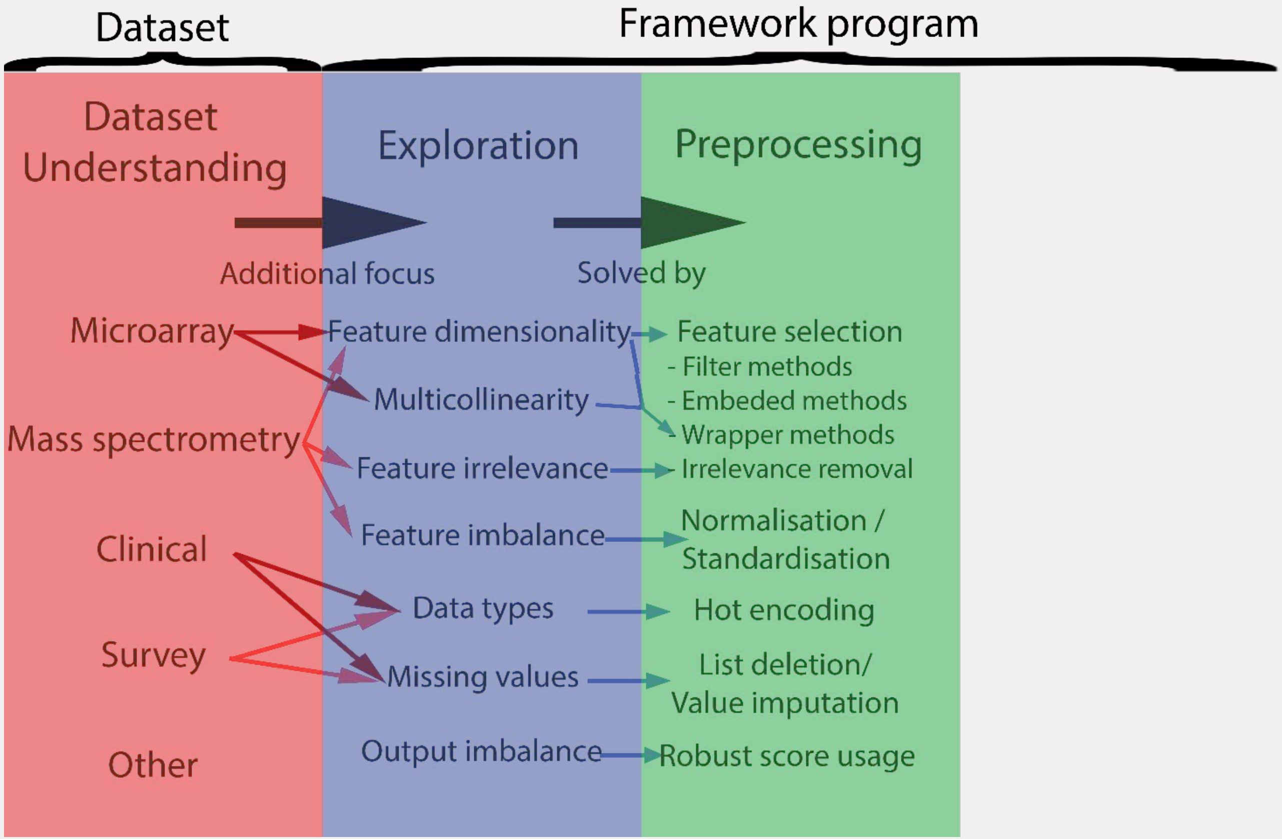
Normalisation / Standardisation

Hot encoding

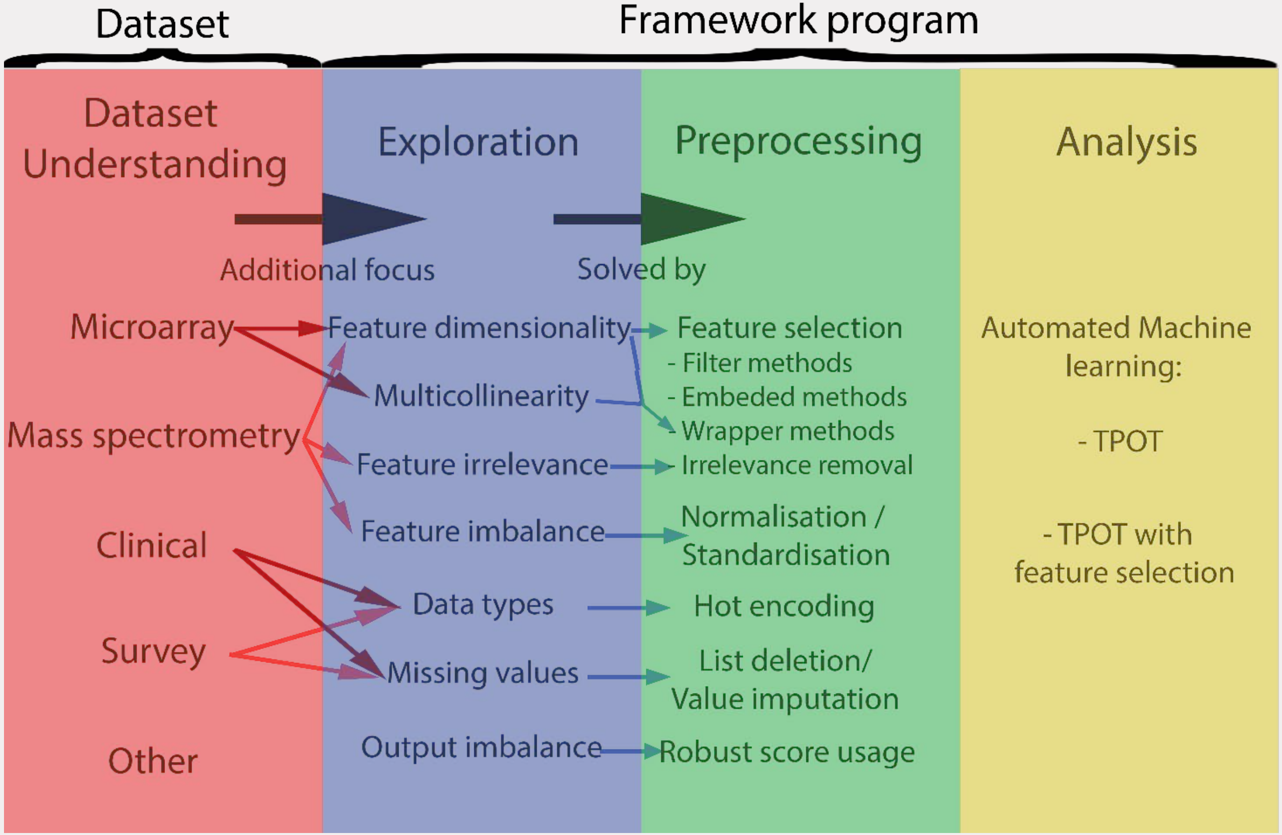
List deletion/ Value imputation

Robust score usage

cBioF layout

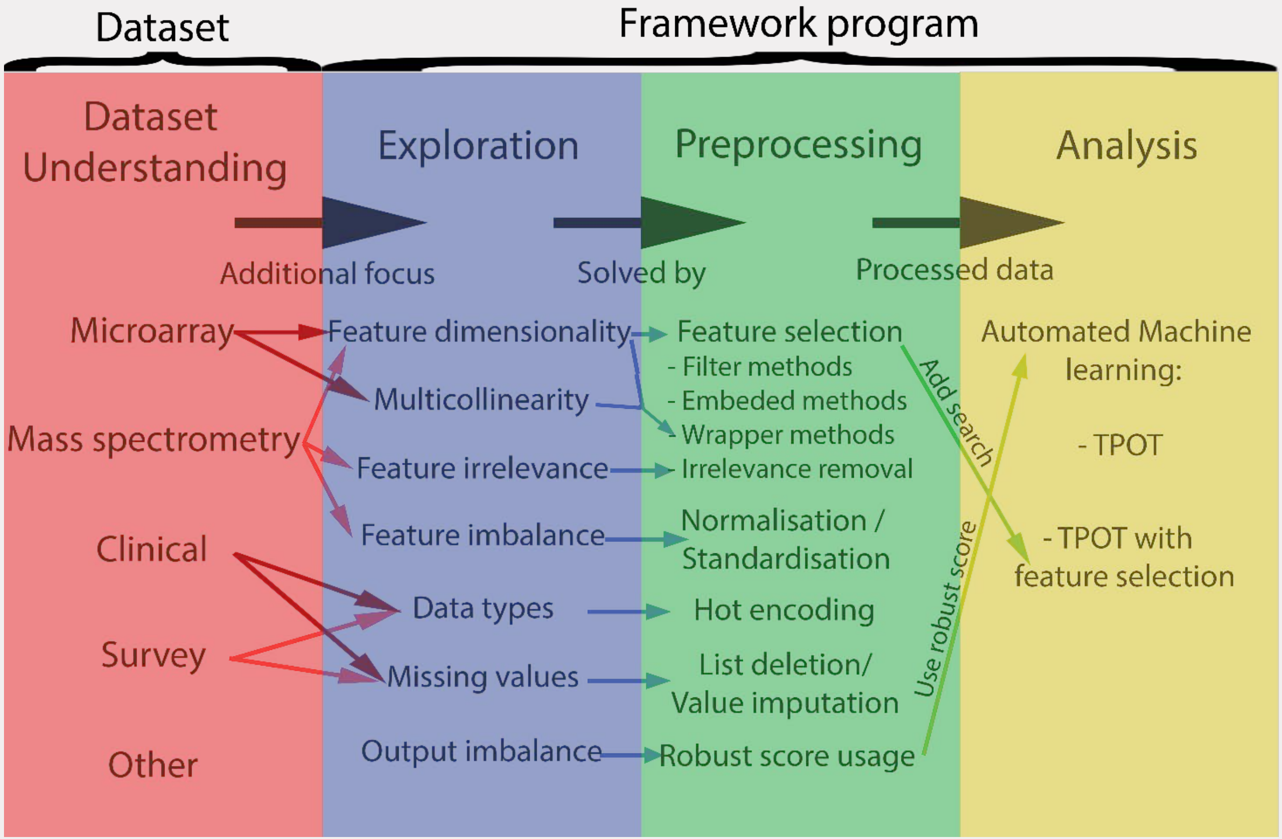


cBioF layout



TU/e

cBioF layout



Conclusions

Feature selection

- 200 features threshold
- Wrapper methods perform best

Missing value handling

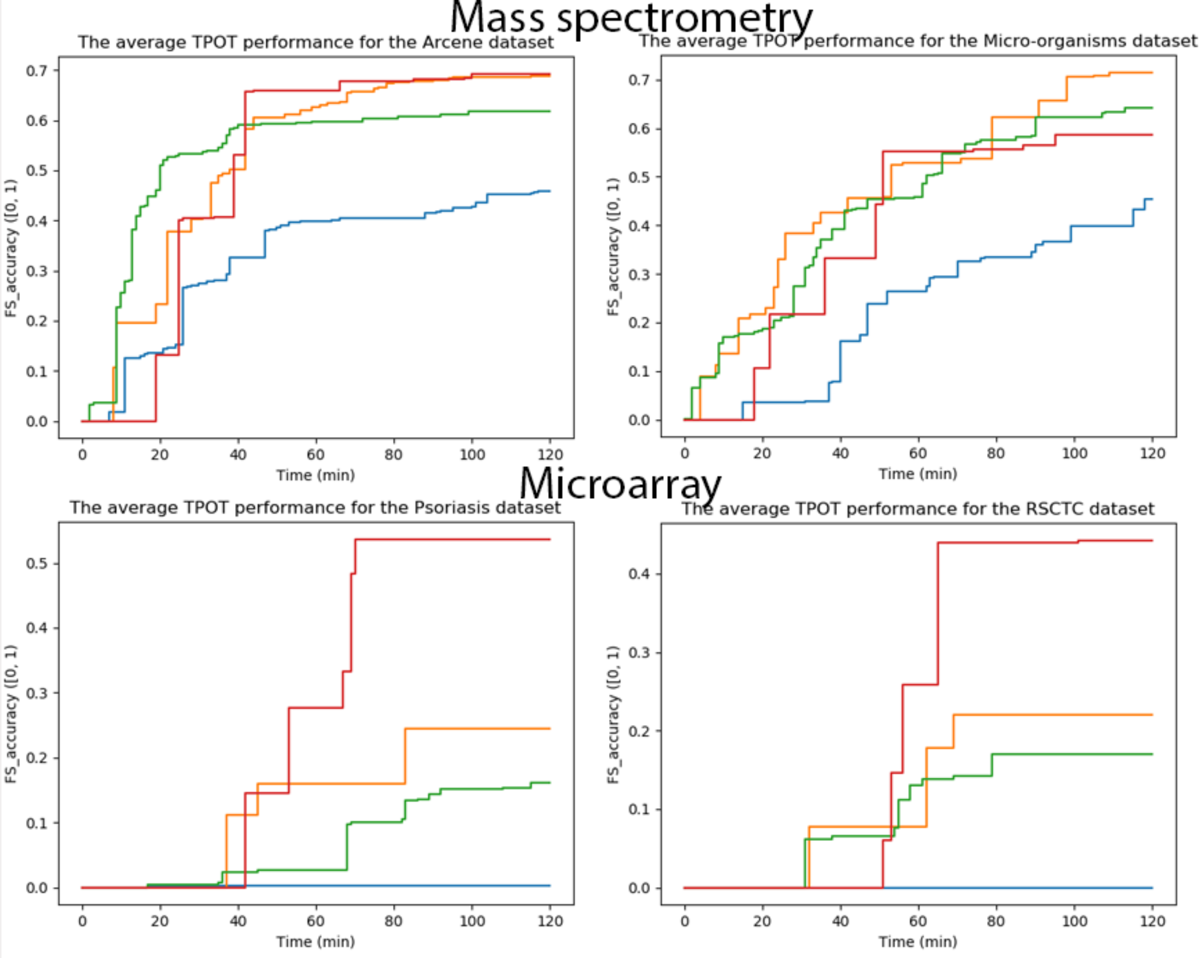
- > 15% missing values
- Mean imputation

Dataset exploration

- Possible with additions *metalearn*

Conclusion TPOT

- Problems many features
- FS_score
- Feature selection added:
  - Bias
  - Algorithms
- Better performance



Future work

Feature selection

- 200 features threshold
- Combination filter and wrapper

Missing value handling

- New datasets

Framework

- Multicollinearity
- Other dataset issues