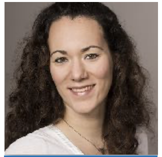
HELMHOLTZ AI Artificial Intelligence Cooperation Unit

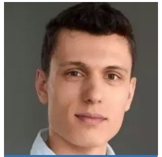
A PRACTICAL GUIDE TO DIMENSIONALITY REDUCTION

Helmholtz AI consultant team

YOUR MENTORS



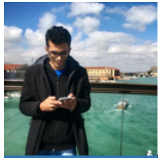
Lisa Barros de Andrade e Sousa



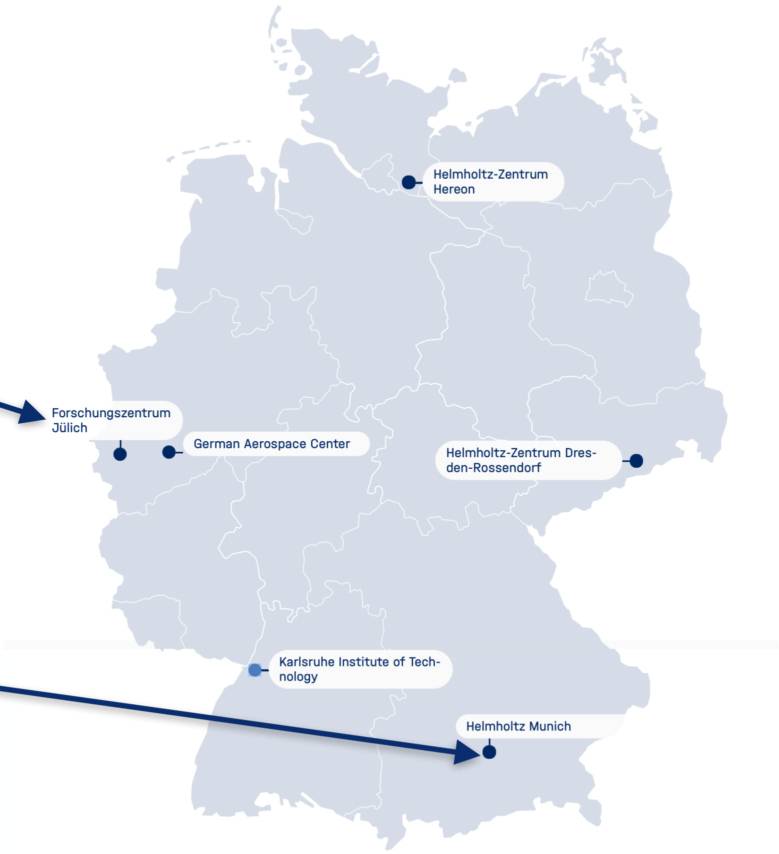
Francesco Campi

Elisabeth Georgii

Isra Mekki



Anoop K. Chandran



SCHEDULE FOR THE COURSE

- Day 1

10:00 - 10:15	Main room: introduction
10:20 - 11:20	Breakout rooms: feature transformation notebook
11:20 - 11:30	Break
11:30 - 12:00	Breakout rooms: autoencoder notebook

- Day 2

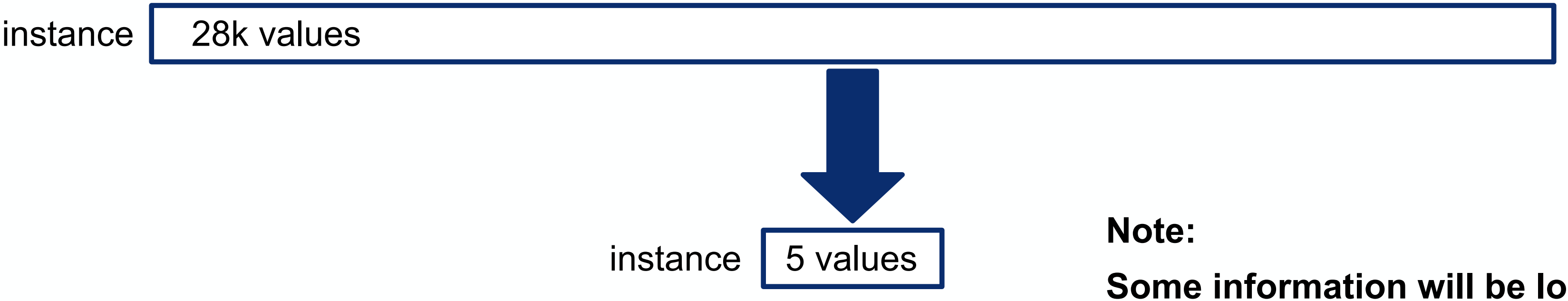
10:00 - 10:20	Breakout rooms: feature aggregation notebook
10:20 - 11:20	Breakout rooms: feature selection notebook
11:30 - 11:50	Breakout rooms: stability optimization notebook
11:50 - 12:00	Main room: wrap-up and conclusions

DEFINITION OF DIMENSIONALITY REDUCTION

- Given a dataset of n instances with p-dimensional measurement profiles, use a transformation to represent the n instances by l-dimensional feature profiles, with l<<p

- Example: gene expression measurements
  - n=24 biological samples
  - p=28k genes

  - Can we reduce the 28k-dimensional profile of each instance to a 5-dimensional profile?



instance 28k values

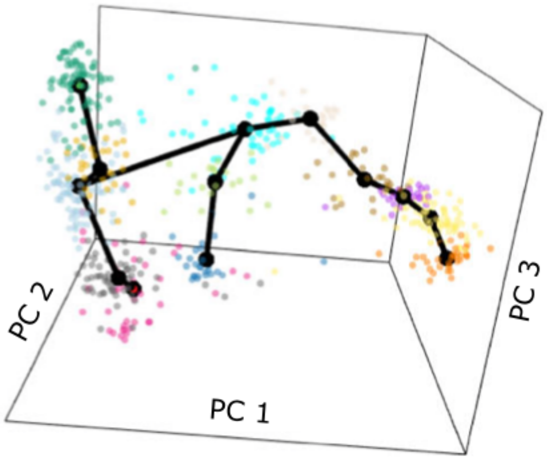
instance 5 values

Note:

Some information will be lost!

PURPOSE OF DIMENSIONALITY REDUCTION

- **Visualize the main dissimilarities between instances** (e.g. a subgroup structure)
  - Allowing humans to visually grasp the data
  - Not only useful for scatterplots of instances but also for heatmaps (number of columns)
  - Also quality control of data (e.g. batch effects)
- **Facilitate machine learning analysis of instances**
  - Clustering for detection of subgroups
  - Robust classification, e.g. healthy vs. cancer
  - Trajectory identification, e.g. cell development

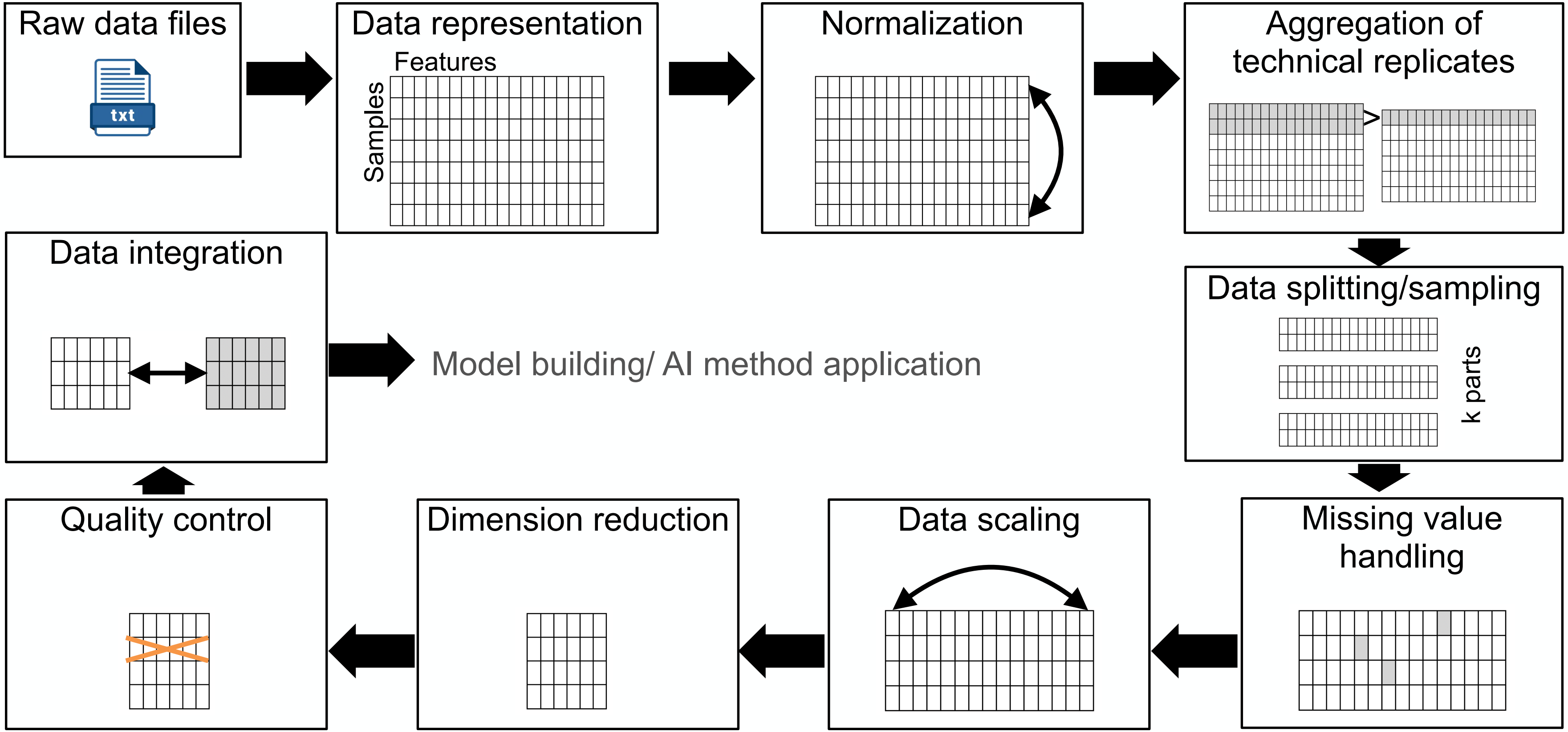


Street *et al. BMC Genomics* (2018) 19:477
https://doi.org/10.1186/s12864-018-4772-0

PURPOSE OF DIMENSIONALITY REDUCTION: DETAILS

- Dimensionality reduction **facilitates data visualization** but not necessarily scientific interpretation: its purpose is **different from explainable AI** (XAI, see other course)
  - E.g. transformed features are composed of contributions from all original variables and transformations might even be nonlinear, so the influence of single variables may not be easy to trace
- Dimensionality reduction **counteracts the curse of dimensionality** in machine learning
  - With increasing number of dimensions, the available data become sparser in the space and instance similarities and groupings get harder and harder to detect
- Dimensionality reduction **leads to computational advantages**: less storage space, more efficient training of models

PART OF DATA PREPROCESSING



Raw data files

Data representation

Normalization

Aggregation of technical replicates

Data splitting/sampling

Missing value handling

Data scaling

Dimension reduction

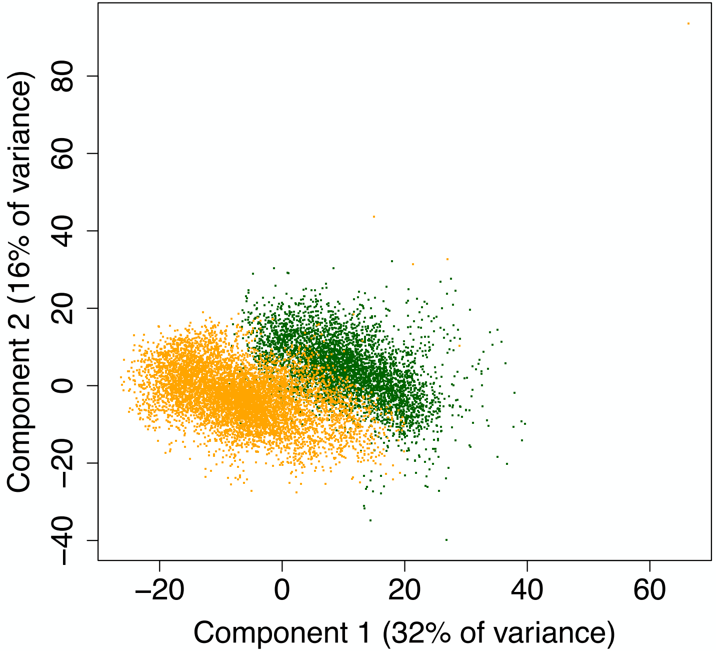
Quality control

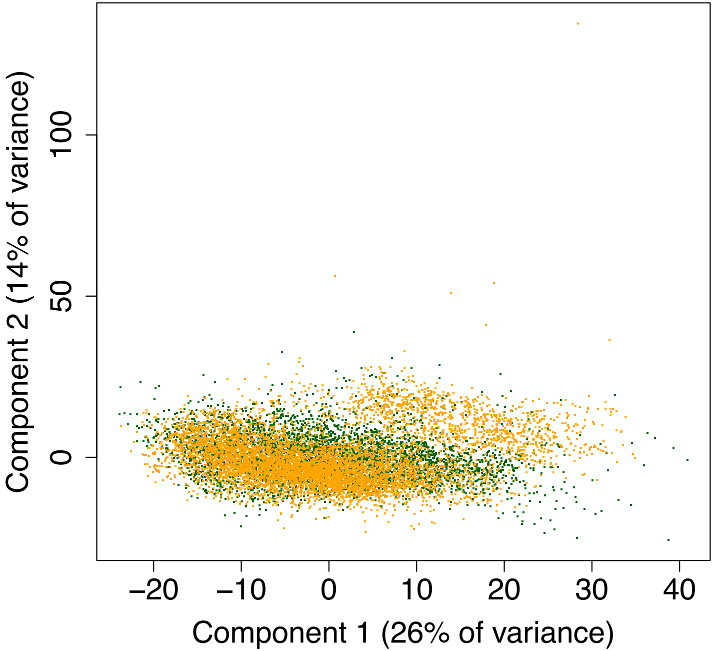
Data integration

Model building/ AI method application

QUALITY CONTROL

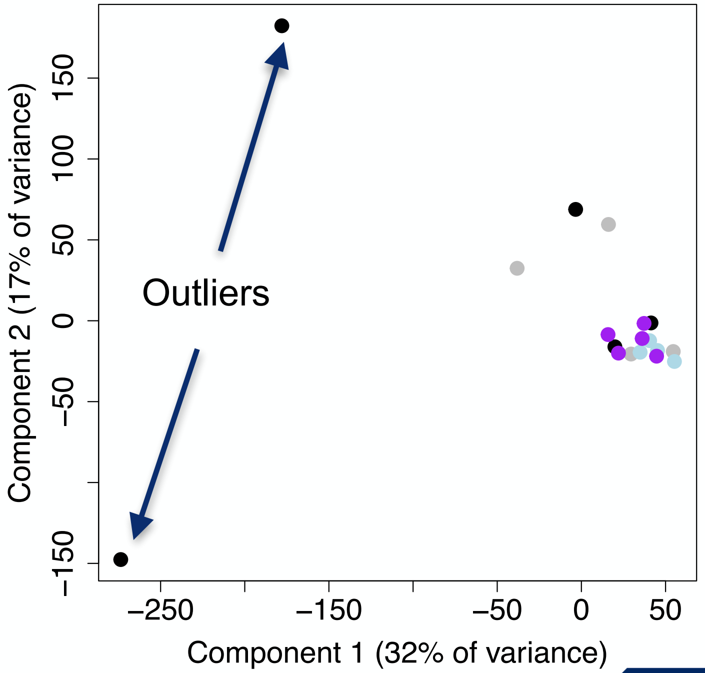
- Dimension reduction assists in revealing batch effects (left)
- Dimension reduction assists in checking success of batch effect correction (right)





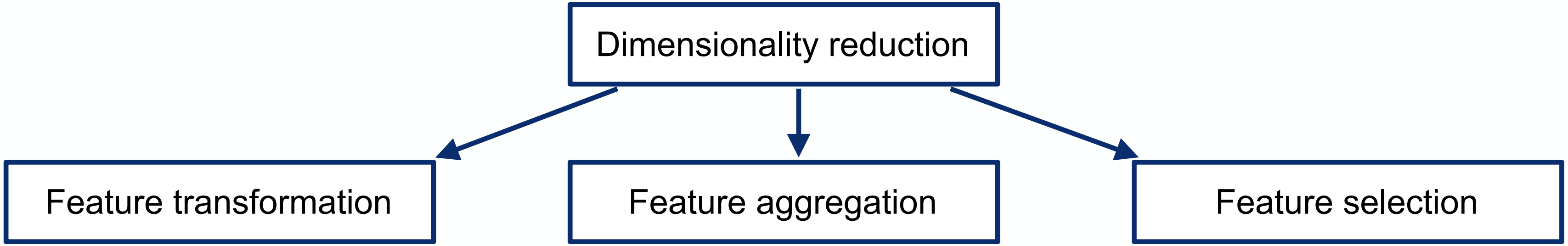
QUALITY CONTROL

- Dimension reduction assists in revealing outlier samples
- Removing outliers might be beneficial for downstream AI analysis



Example visualization
of RNA-seq samples

DIMENSIONALITY REDUCTION APPROACHES



- Linear methods
  - PCA
  - ICA
  - CCA (multimodal)
- Nonlinear methods
  - MDS
  - t-SNE
  - UMAP
  - Autoencoders

- Data-driven methods
  - kMedoid clustering
  - Graph partitioning
- Knowledge-based methods
  - GO slim categories
  - Molecular pathways

- Filter methods
  - ReliefF
  - Correlation
- Wrapper methods
  - Forward selection
  - Backward elimination
- Embedded methods
  - Elastic net

Mostly unsupervised: no target variable taken into account

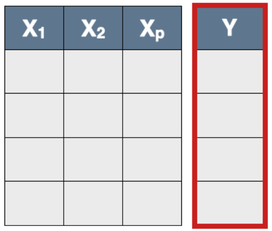
Mostly supervised prediction

SUPERVISED VS. UNSUPERVISED APPROACHES

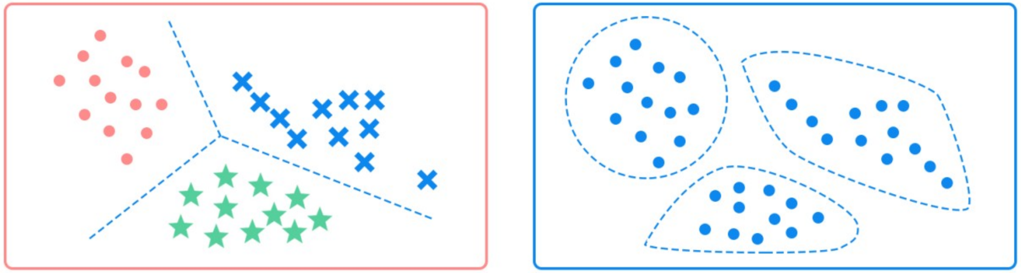
- Most often, *feature transformation and feature aggregation are* performed in an *unsupervised* manner, i.e. no target variable for the instances (e.g. class label, continuous output) is taken into account during dimensionality reduction

- However, *there also exist supervised feature transformation methods*, e.g. linear discriminant analysis, and feature aggregation may be targeted toward supervised tasks

- *Feature selection methods are typically supervised, but there are also unsupervised methods*, e.g. variance-based filters



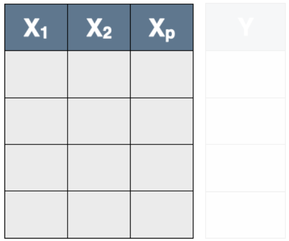
Target



Supervised learning

Unsupervised learning

https://www.linkedin.com/pulse/supervised-vs-unsupervised-learning-whats-difference-smriti-saini/



No Target

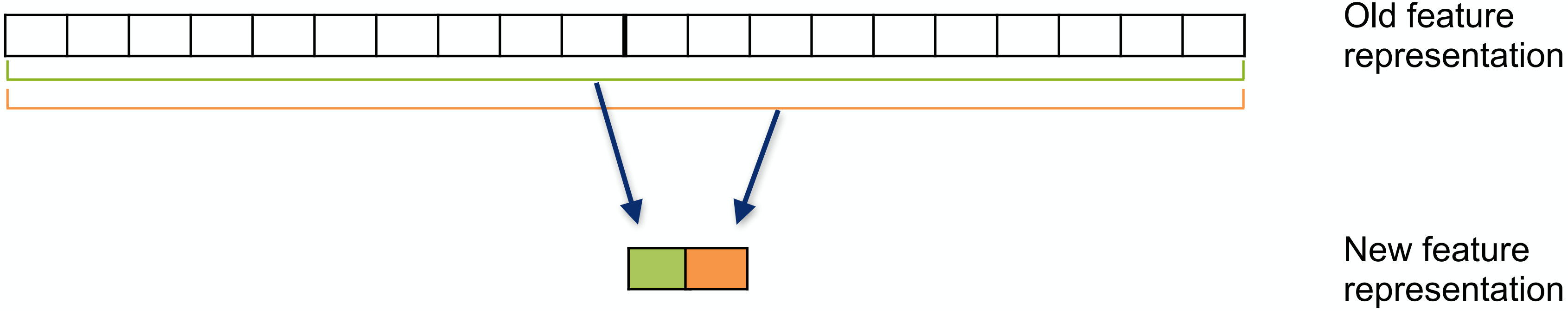
https://www.sharpsightlabs.com/blog/supervised-vs-unsupervised-learning/

FEATURE TRANSFORMATION

- **Idea:** reduce dimensionality by computing new features based on all original features to condense the most relevant information

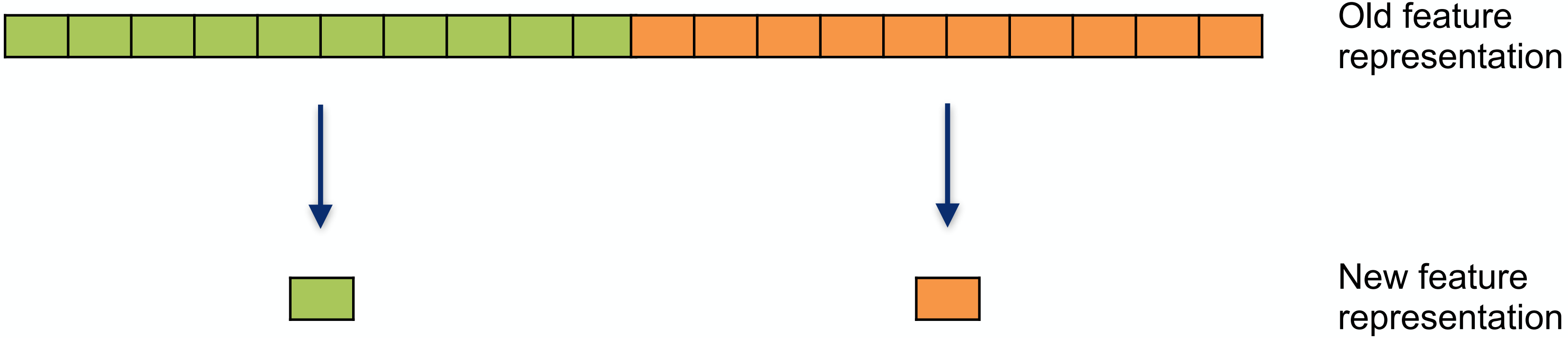
- **Linear methods:** new features are linear combinations of original features

- **Nonlinear methods:** new features are nonlinear transformations of original features



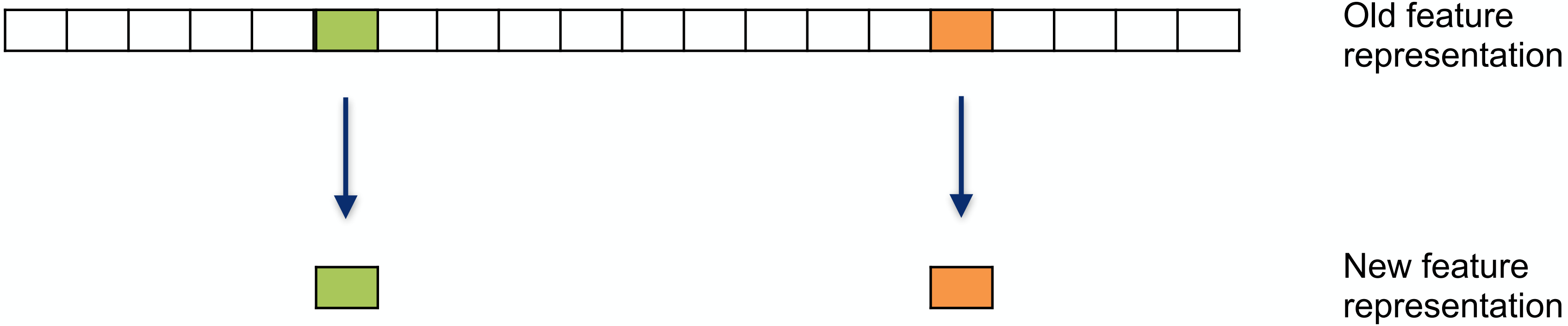
FEATURE AGGREGATION

- **Idea:** reduce dimensionality by summarizing each group of original features into one aggregated feature
- **Data-driven methods:** use the dataset at hand to group similar features
- **Knowledge-based methods:** use annotation databases to define feature groups



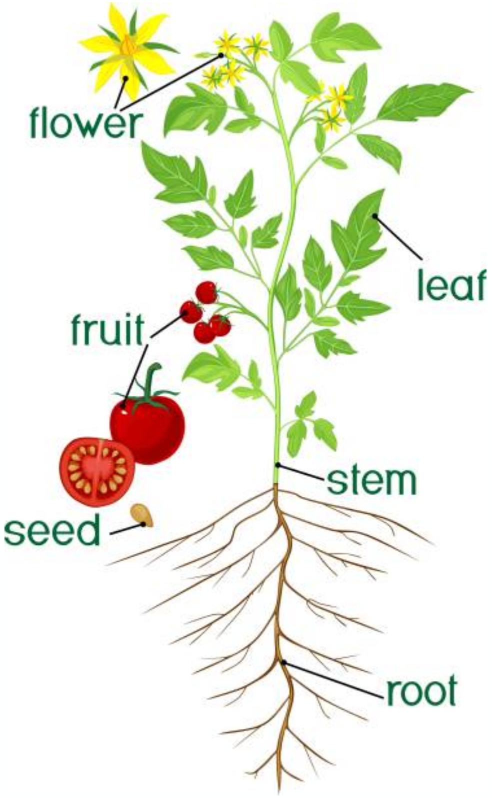
FEATURE SELECTION

- **Idea:** reduce dimensionality by picking a subset of original features
- **Filter methods:** score each individual feature
- **Wrapper methods:** test different sets of features
- **Embedded methods:** learn the importance of features



EXAMPLE DATASET

- Gene expression dataset (RNA-seq) of tomato plants
- For each plant sample, there are expression measurements for more than 28k genes (Koenig et al., PNAS)
- There are 24 samples, which cover all combinations of three experimental factors:
    - Plant tissue: floral tissue, leaf, root, seedling, stem, vegetative tissue
    - Tomato species: Solanum lycopersicum M82 (domesticated), Solanum pennellii (wild, desert-adapted)
    - Growing location: sun, shade



https://www.istockphoto.com/de/vektor/teile-der-pflanze-morphologie-der-blüte-tomatenpflanze-mit-titel-gm989916016-268335890

REPO

https://github.com/HelmholtzAI-Consultants-Munich/DimRed-Course

How to get started:

- Go to the notebooks folder
- Open 1_feature_transformation.ipynb
- Use Google colab (recommended, Google account needed)

Alternative:

- Clone or download the repo
- Run the following commands in a terminal

```
conda create -n dimred python=3.10
conda activate dimred
pip install -r requirements.txt
```

- Open 1_feature_transformation.ipynb from the notebooks folder and select dimred as kernel

BREAKOUT ROOMS: 0, F, 9

Please distribute now evenly among these 3 rooms, you are allowed to freely choose one of the rooms. The following is a rough suggestion to get started.

Last name A-G	Room 0
Last name H-R	Room F
Last name S-Z	Room 9