SEPARATING SLOW PIONS FROM OTHER PARTICLES USING PCA, LDA, T-SNE & DECISION TREES

STEPHANIE KÄS | JLU BELLE II GROUP

Results by: S. Käs, T. Schellhaas, J. Bilk, M. Peter

Bundesministerium für Bildung und Forschung

JENNIFER 2 EU grant n.822070

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Belle II

SLOW PIONS – RESEARCH@JLU

<u>There are three tasks:</u>

1. Separate slow pions from electrons
   See talks by Johannes and Timo. Munich and Giessen.
2. Separate slow pions from beam background
   See talk by Stephanie.
   So far only Giessen.
3. Separate highly ionizing particles from each other (slow pions, antideuterons, magnetic monopoles).
   Today no talk, but results were presented at DPG spring meeting 2021.

OVERVIEW

<u>Remarks:</u>

(a) numbers of signal/background separation are similar in all cases (70-85% accuracy)

(b) all of them will have no ROI (thus, PXD hits are removed on Onsen).

(c) cluster shape may be different: beam background are partially off-vertex, and monopoles have non-helix track.

JLU PXD RESULTS: 3 TALKS TODAY

Group Leader: J. Sören Lange

"NeuroGroup"

- Stephanie Käs
  - PCA, LDA, t-SNE, RF
    - Slow Pion | BGD
    - Slow Pion | Electron
- Timo Schellhaas
  - SVMs
    - Slow Pion | Electron
- Johannes Bilk
  - CNNs, MLPs
    - Slow Pion | Various particles

GOAL:

SEPARATE SLOW PIONS FROM BGD

Overview

1. Basic Statistics
2. PCA
3. LDA
4. T-SNE
5. Decision Trees / Random Forest

TERMINOLOGY

Accuracy

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

"amount of
correctly classified particles"

Efficiency (Precision)

$$PPV = \frac{TP}{TP + FP}$$

"amount of real pions among
all particles classified as pions"

Purity (Sensitivity)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = 1 - \mathrm{FNR}$$

"amount of correctly classified pions
among all pions"

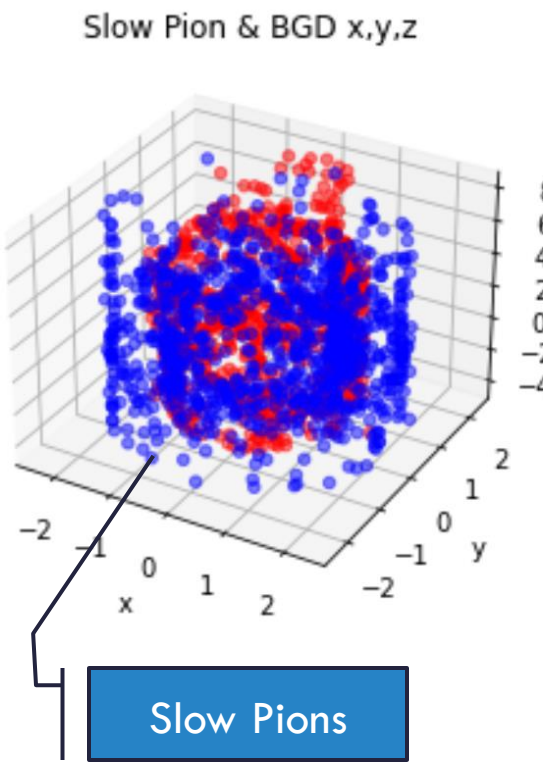
Rejection (Negative Prediction Value)

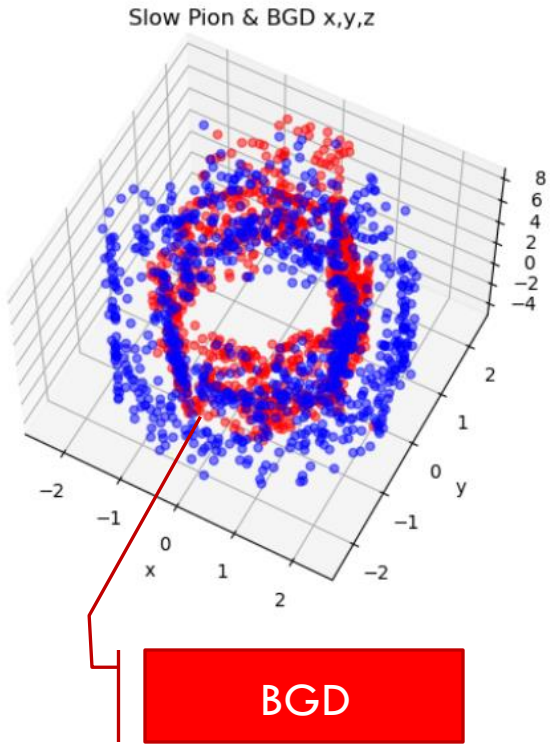
$$NPV = \frac{TN}{TN + FN}$$

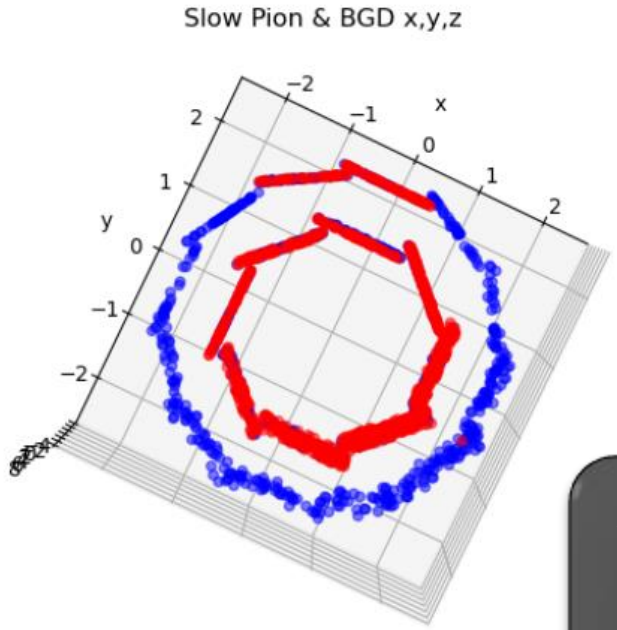
"amount of BGD/electrons among
particles classified as BGD/electrons"

Source: T. Schellhaas (2021) Using Support Vector Machines to find Slow Pions in the PXD. Belle II FSP Meeting "Slow Pion Tracking".

DISTRIBUTION OF SIMULATED DATA





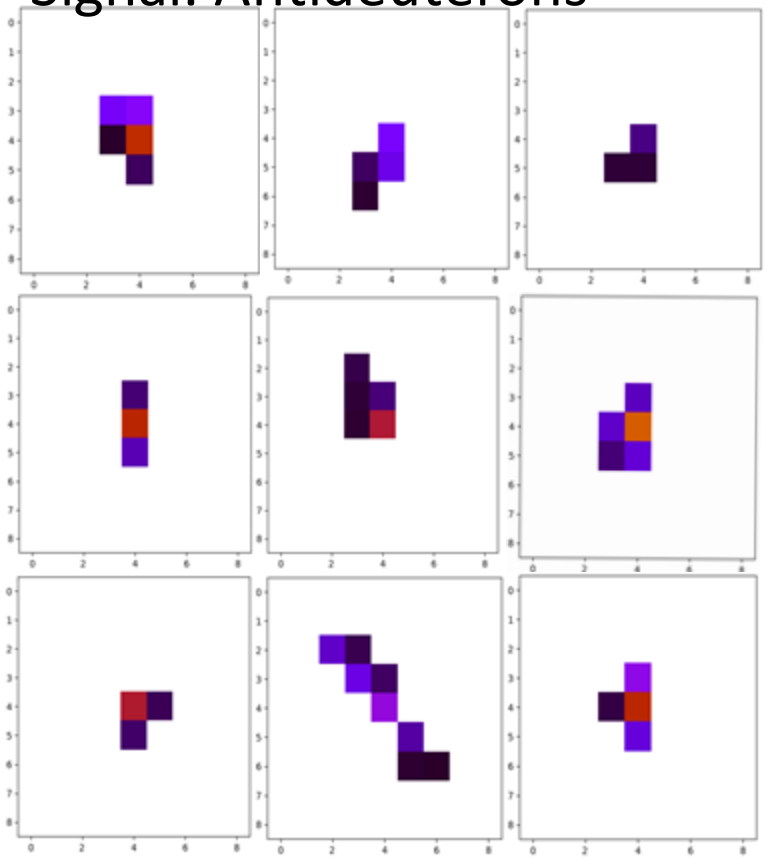


We can clearly see the barrel-like structure.

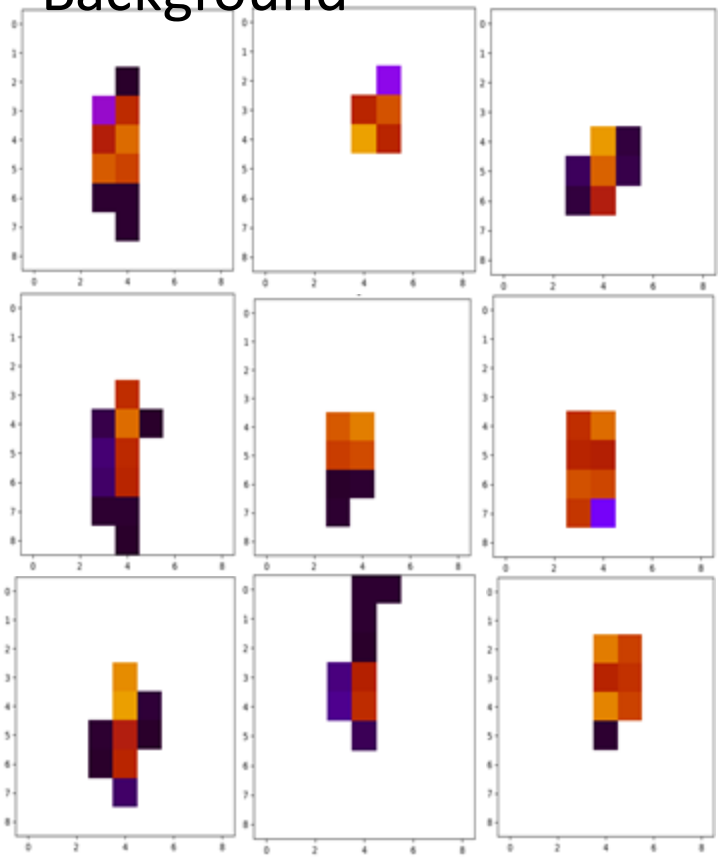
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

REMINDER: PXD DATA LOOKS LIKE THE GAME TETRIS

Signal: Antideuterons



Background



M. Peter, Unpublished

9x9 matrix
ADC values

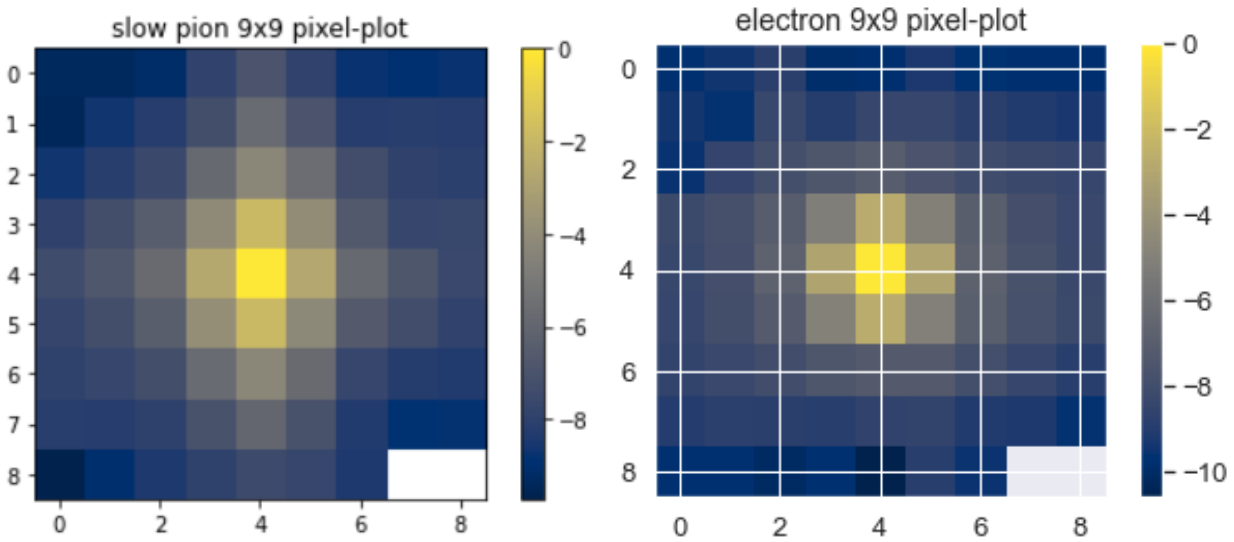
Low ADC values

High ADC values

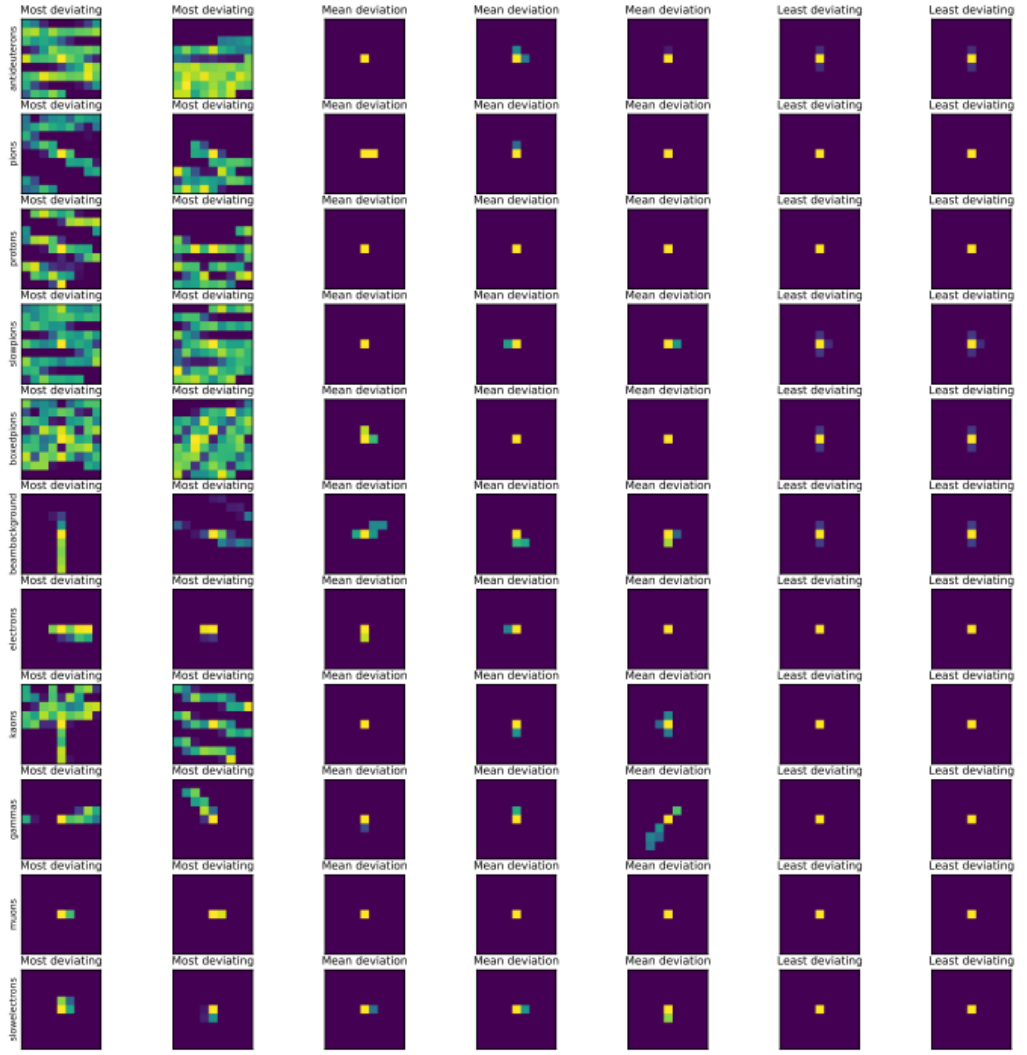
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

CLUSTER ANALYSIS

Logarithmic scaled sum of all charge values



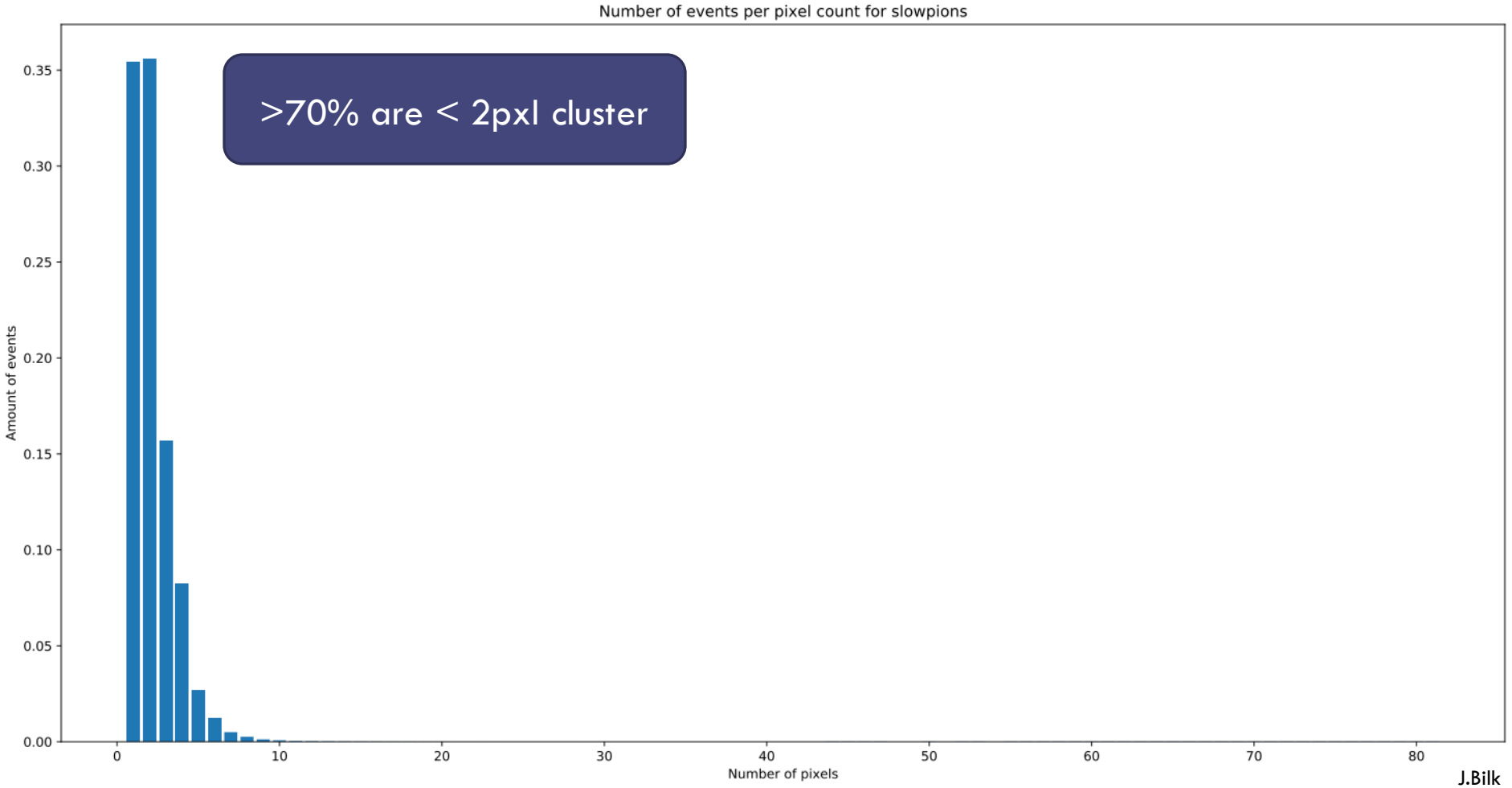
T. Schellhaas



J.Bilk

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

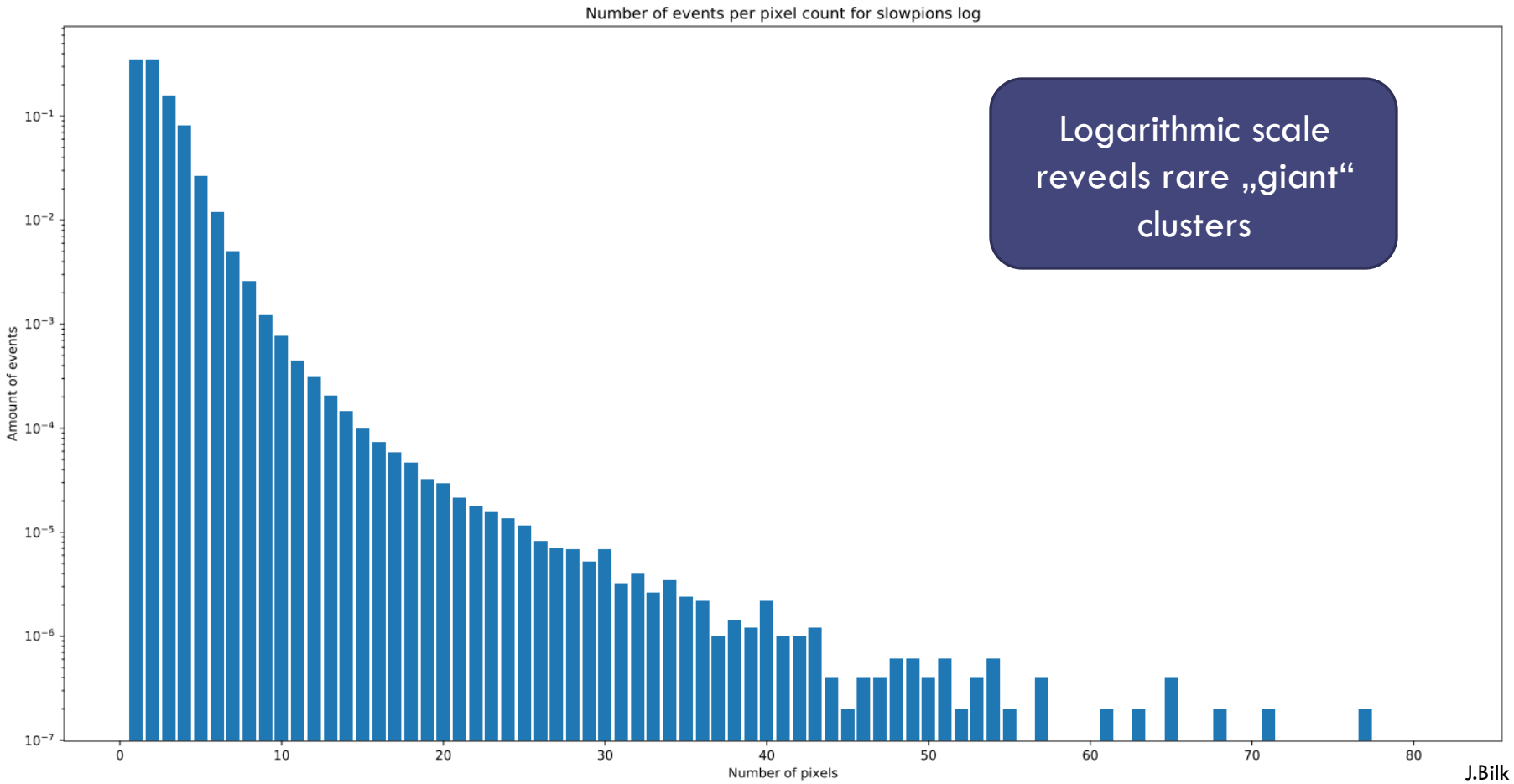
CLUSTER ANALYSIS



J.Bilk

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

CLUSTER ANALYSIS



Logarithmic scale reveals rare „giant“ clusters

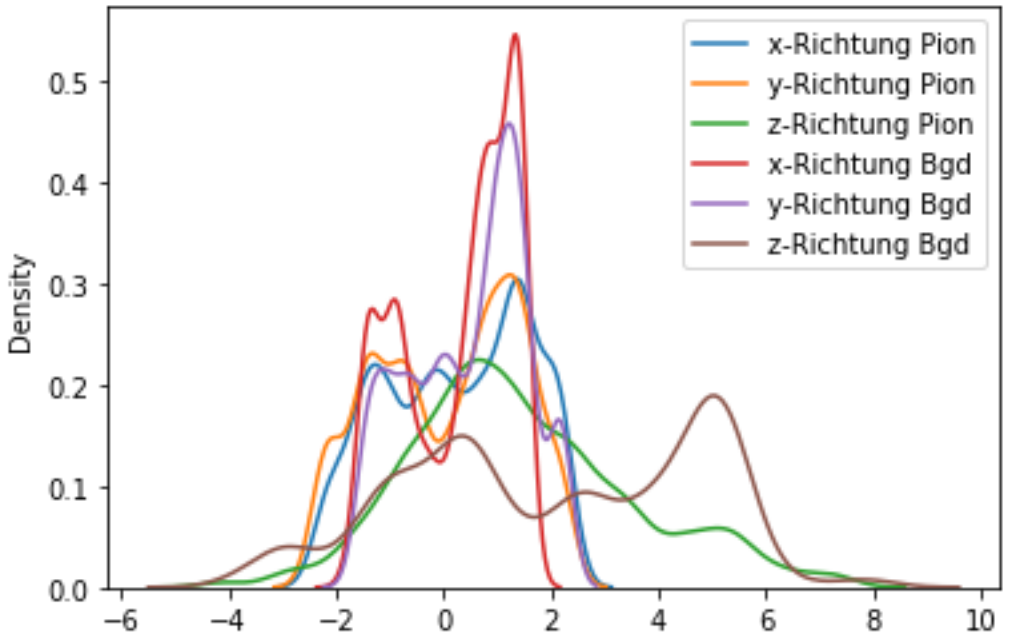
J.Bilk

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

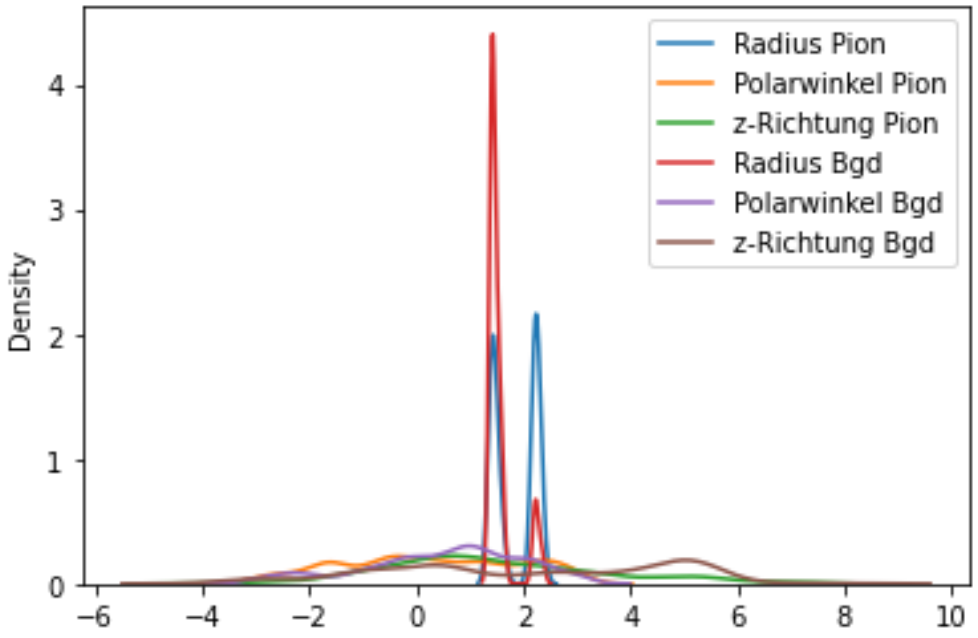
BASIC STATISTIC ANALYSIS

KERNEL DENSITY ESTIMATOR

Cartesian coordinates

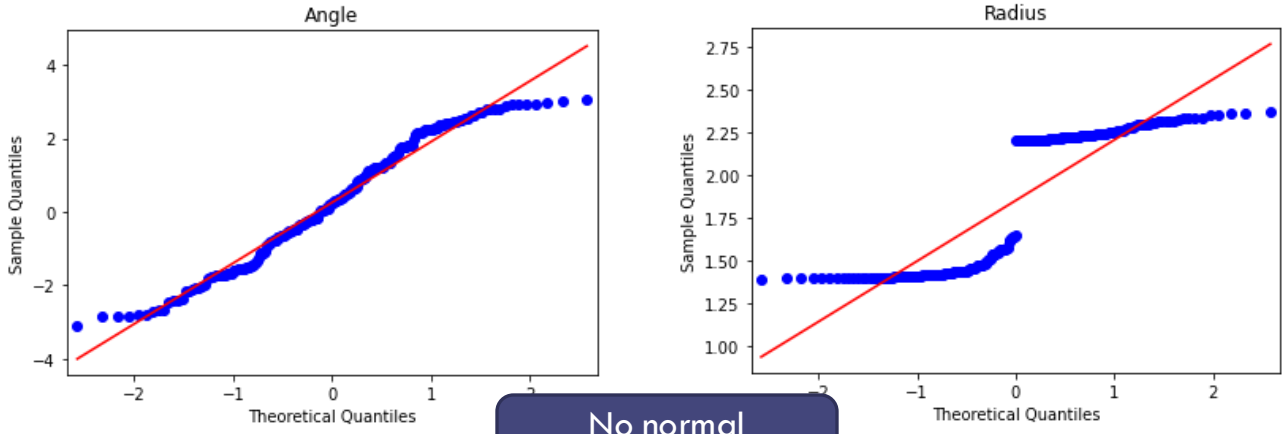


Cylindrical coordinates

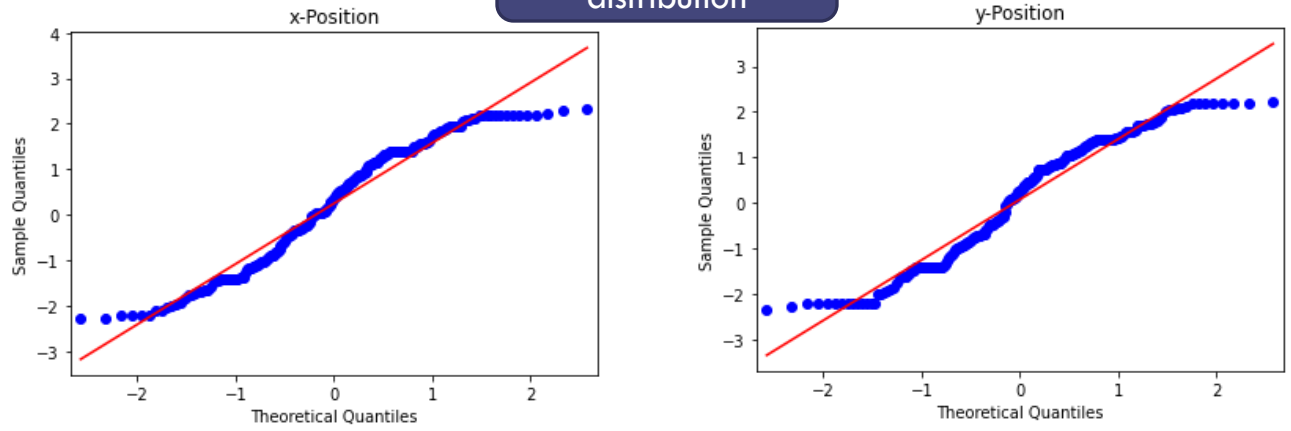


T. Schellhaas

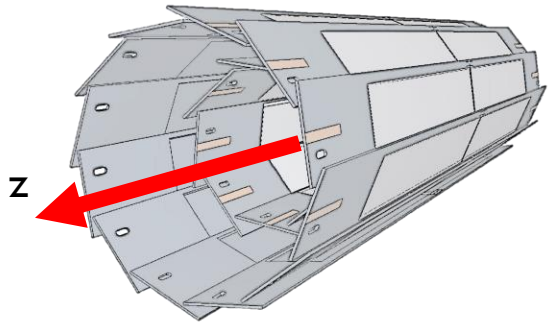
QQ-PLOT



No normal distribution

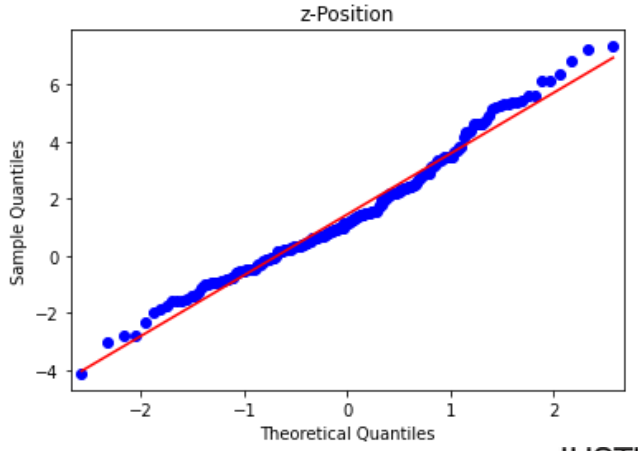


T. Schellhaas



Abe et al. Belle 2 Technical Design Report
arXiv:1011.0352 [physics.ins-det]

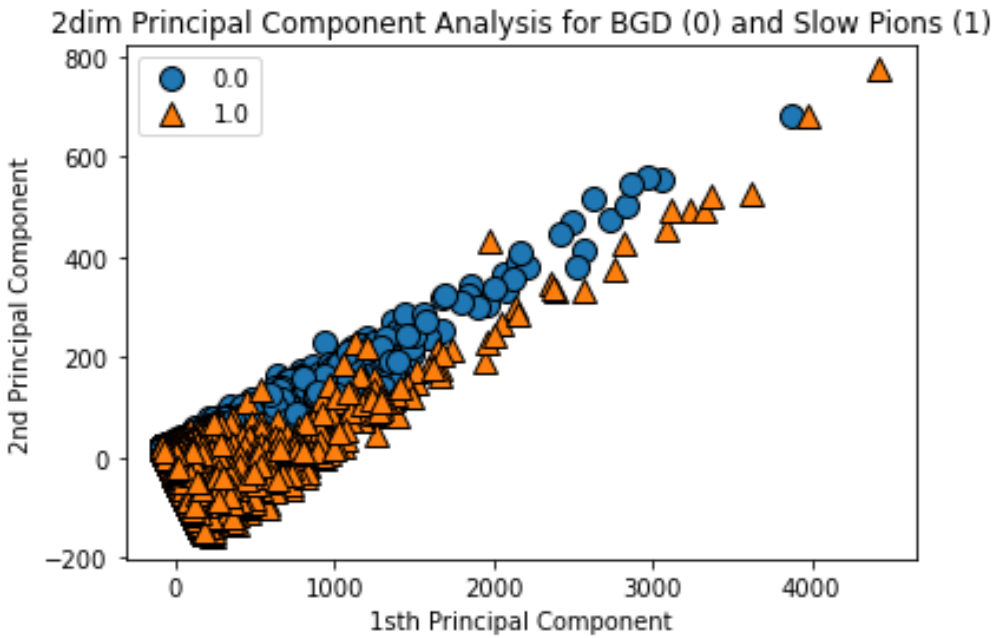
Normal distribution
In z-direction



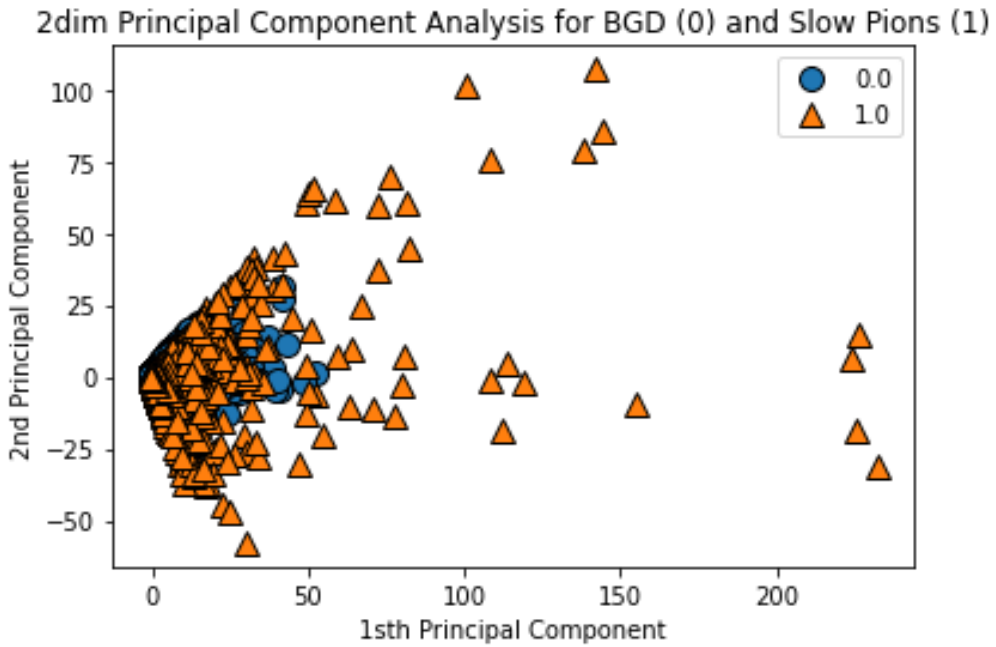
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PRINCIPAL COMPONENT ANALYSIS

PCA – RESULTS – SLOW PION | BGD



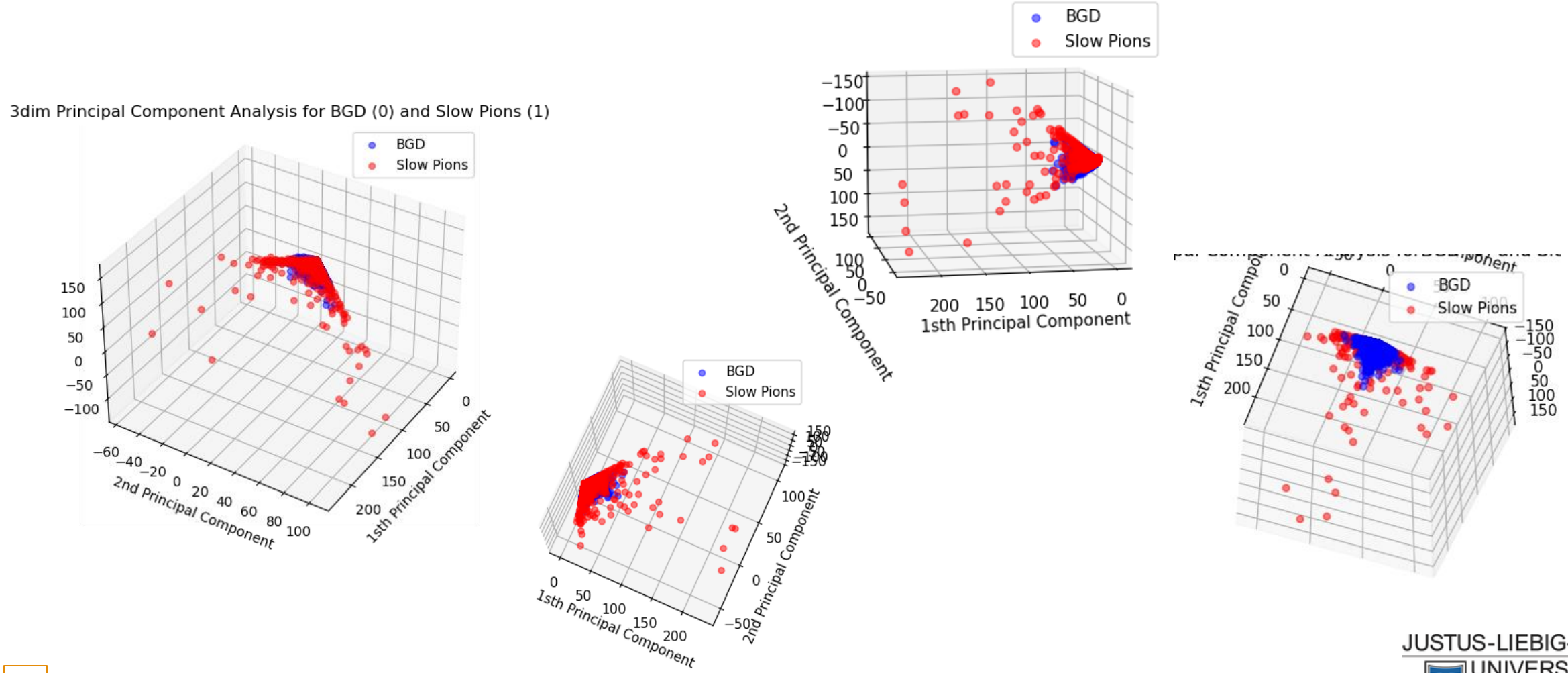
Without scaling



With scaling

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

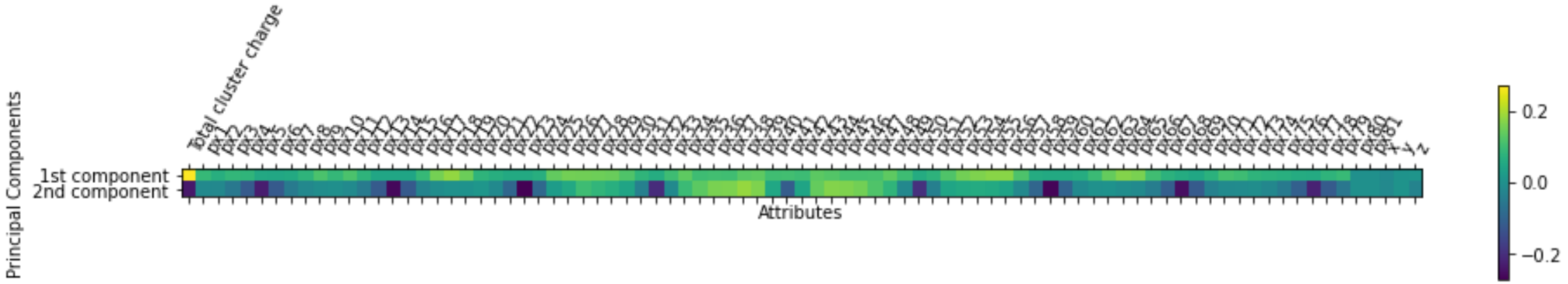
PCA – RESULTS – SLOW PION | BGD



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PCA – RESULTS – SLOW PION | BGD

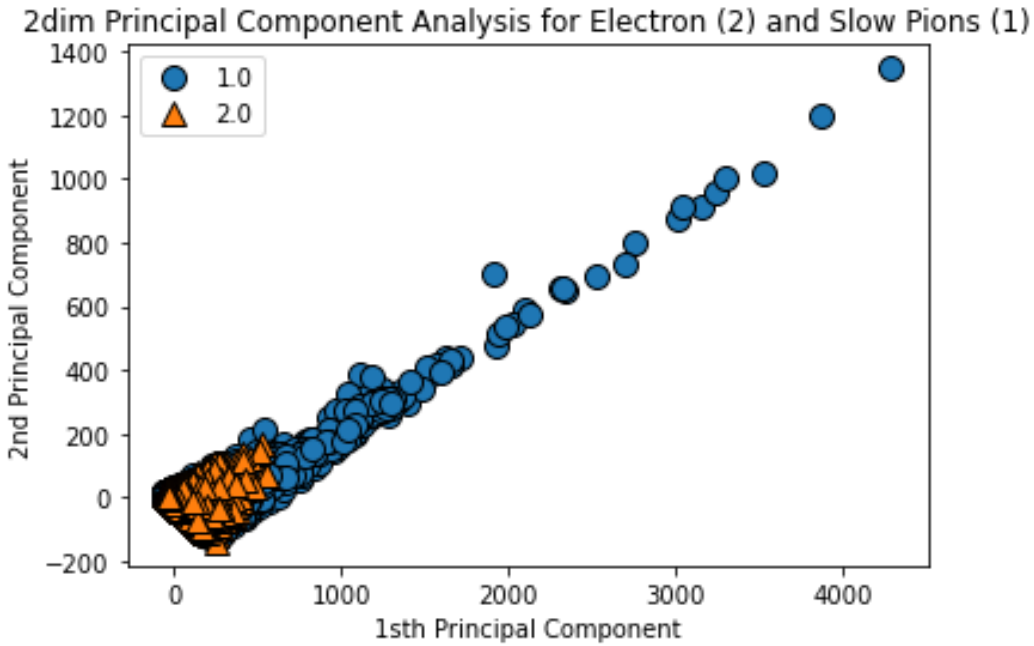
Contribution to the first two principal components (scaled dataset):



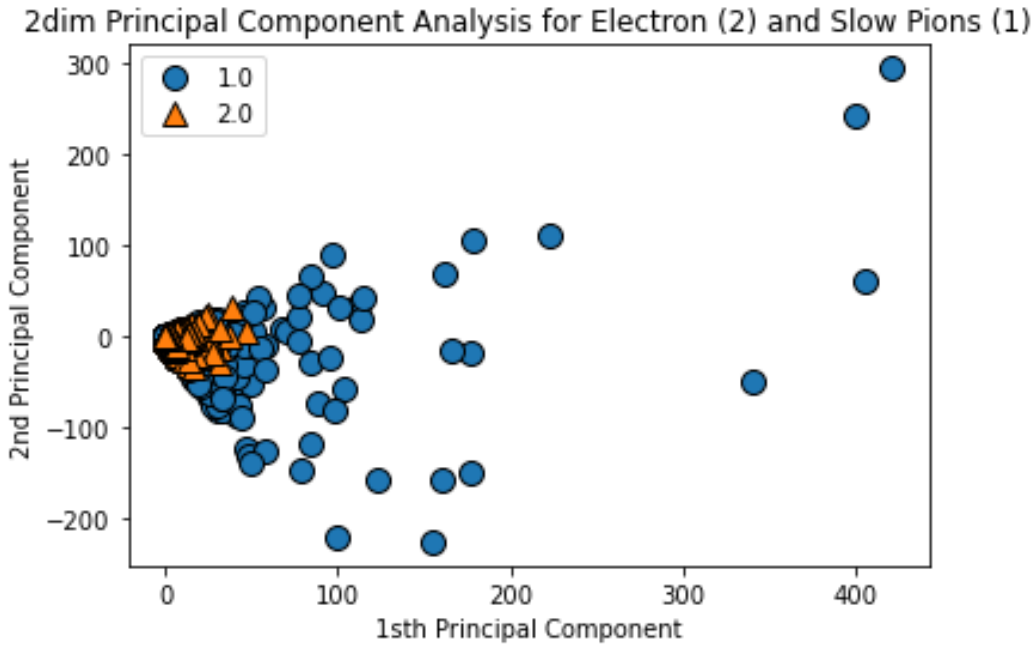
1sth component: Total cluster charge + higher charged pixels
2nd component: Total cluster charge + pxl 4, 13, 22, 31, 40, 49, 58, 67, 76

The influence of x,y,z seems to be almost neglectable among the first 2 components.

PCA – RESULTS – SLOW PION | ELECTRON



Without scaling

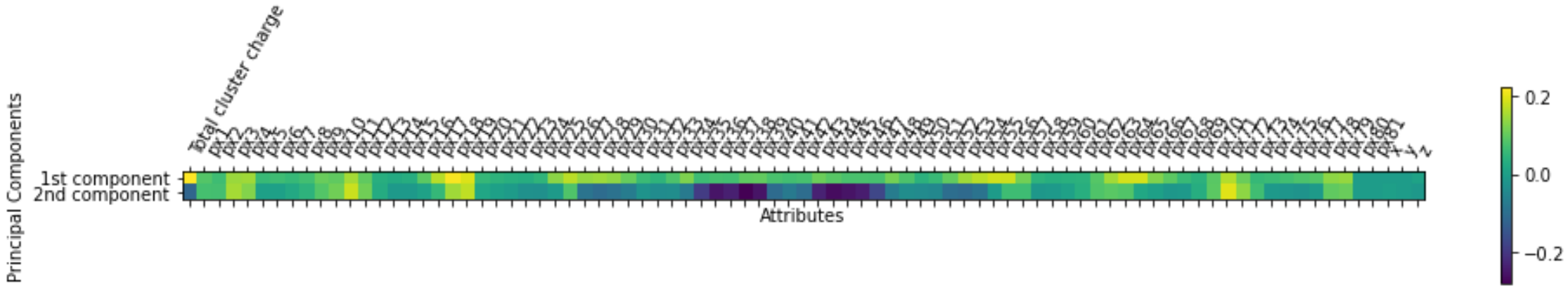


With scaling

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PCA – RESULTS – SLOW PION | ELECTRON

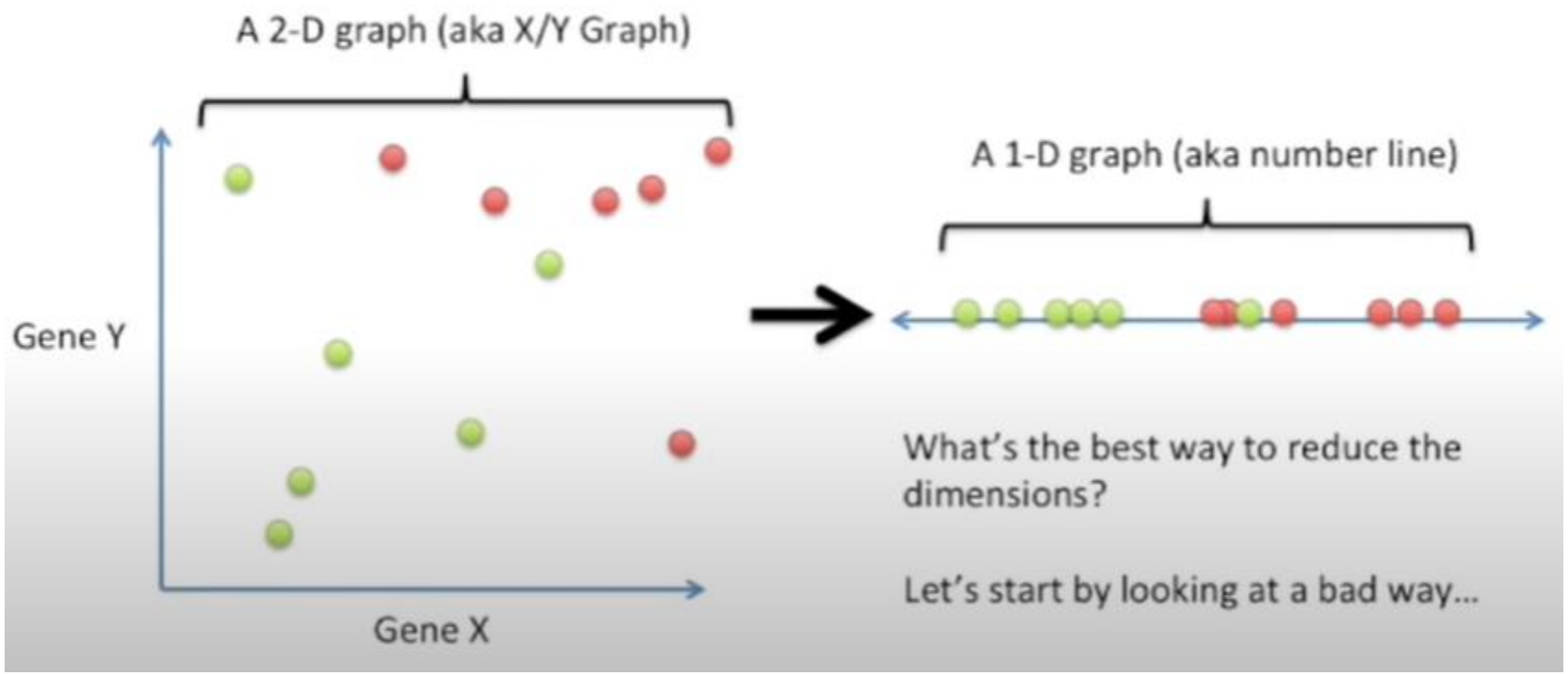
Contribution to the first two principal components (scaled dataset):



Different distribution of higher charged pixels.

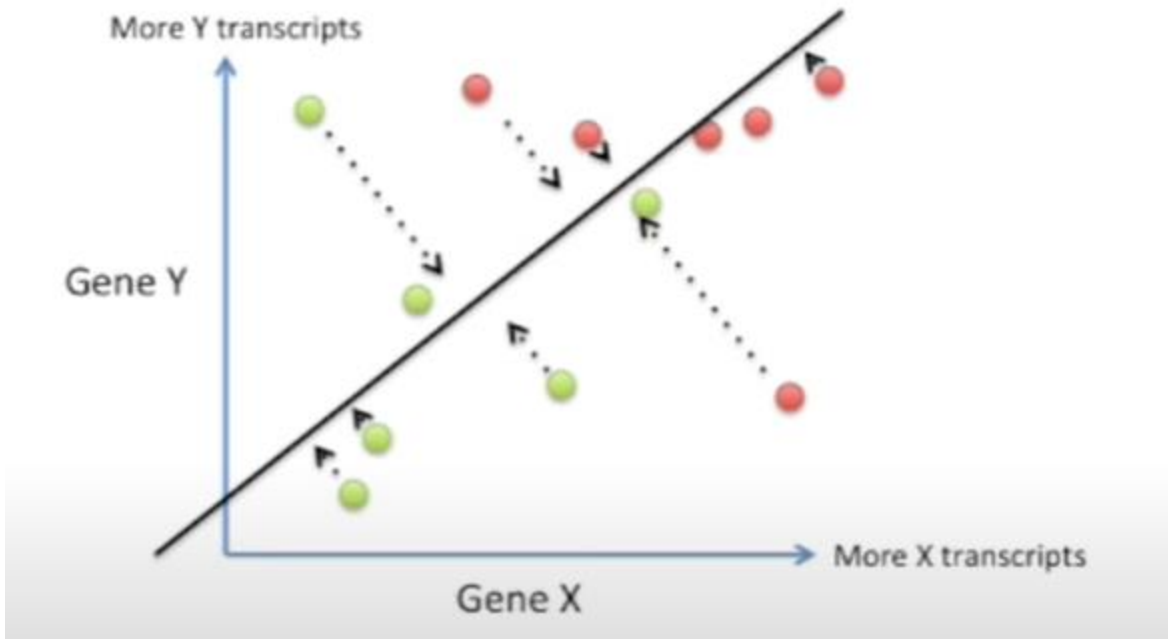
LINEAR DISCRIMINANT ANALYSIS

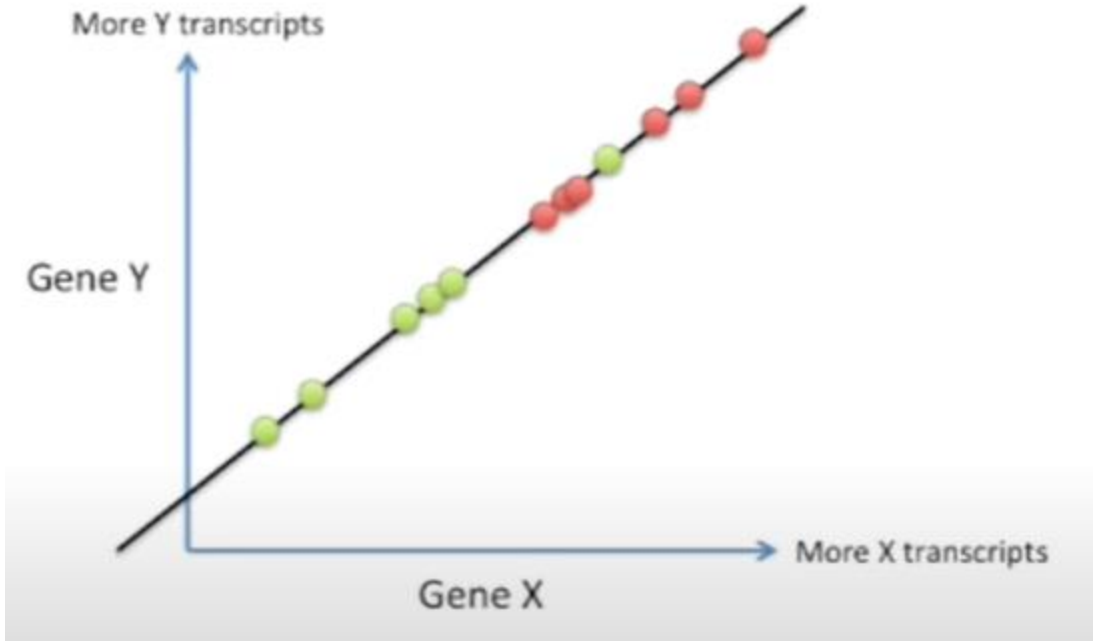
LDA - MAIN IDEA



J. Starmer. 2016. StatQuest: Linear Discriminant Analysis (LDA) clearly explained. YouTube.

LDA - MAIN IDEA





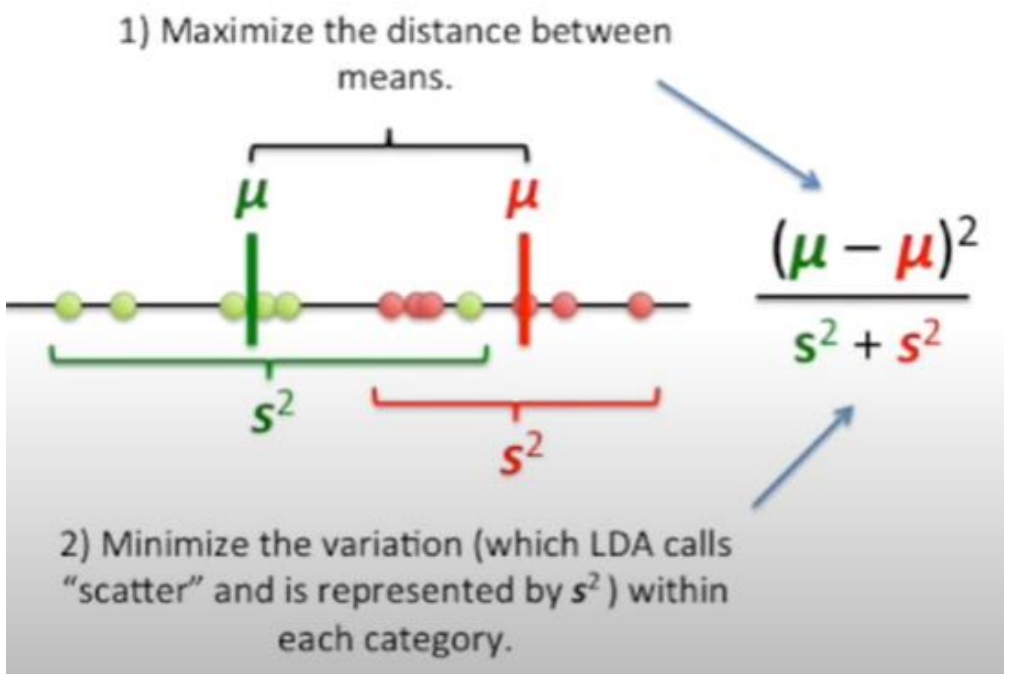
J. Starmer. 2016. StatQuest: Linear Discriminant Analysis (LDA) clearly explained. YouTube.

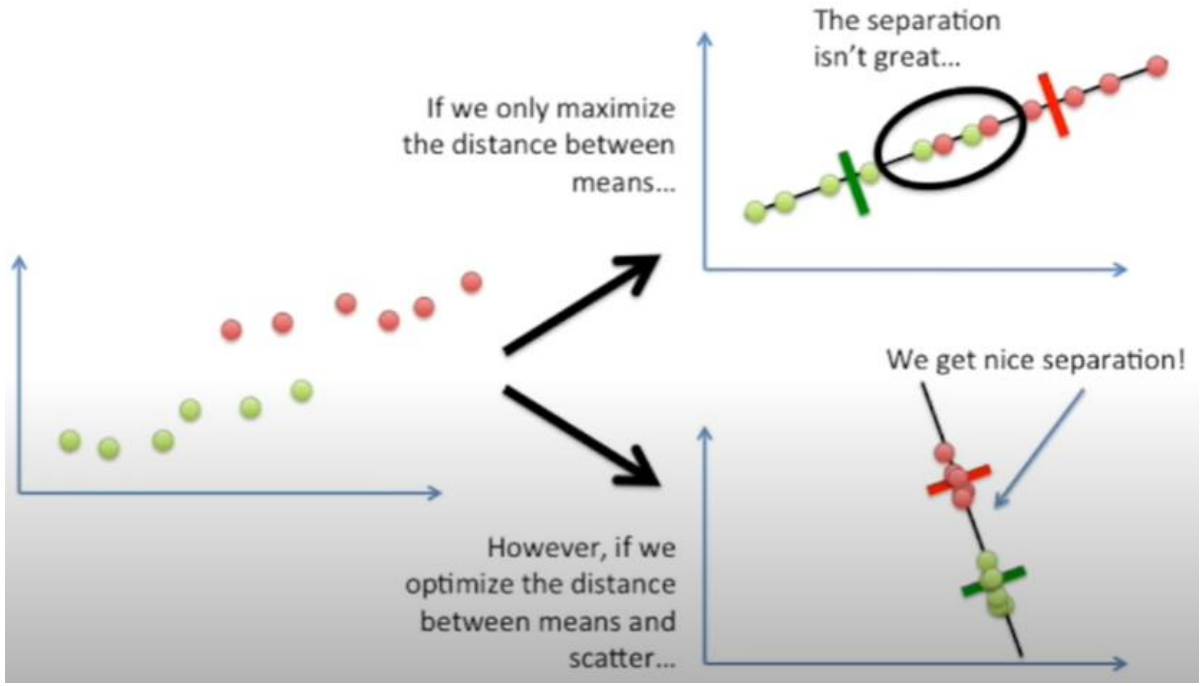
LDA creates a new axis by...

...maximizing the separability between the two classes.

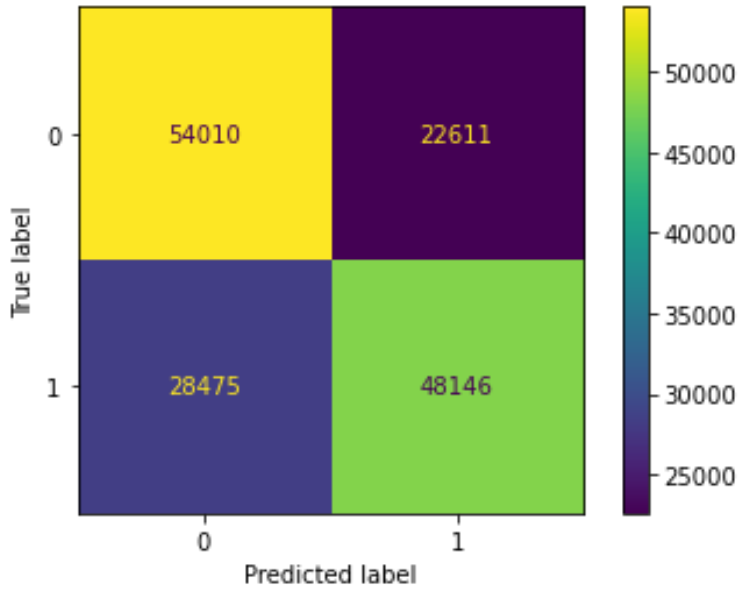
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

LDA - MAIN IDEA





J. Starmer. 2016. StatQuest: Linear Discriminant Analysis (LDA) clearly explained. YouTube.



LDA RESULTS

SLOW PION | BGD

LDA is not useful in this case!

67%

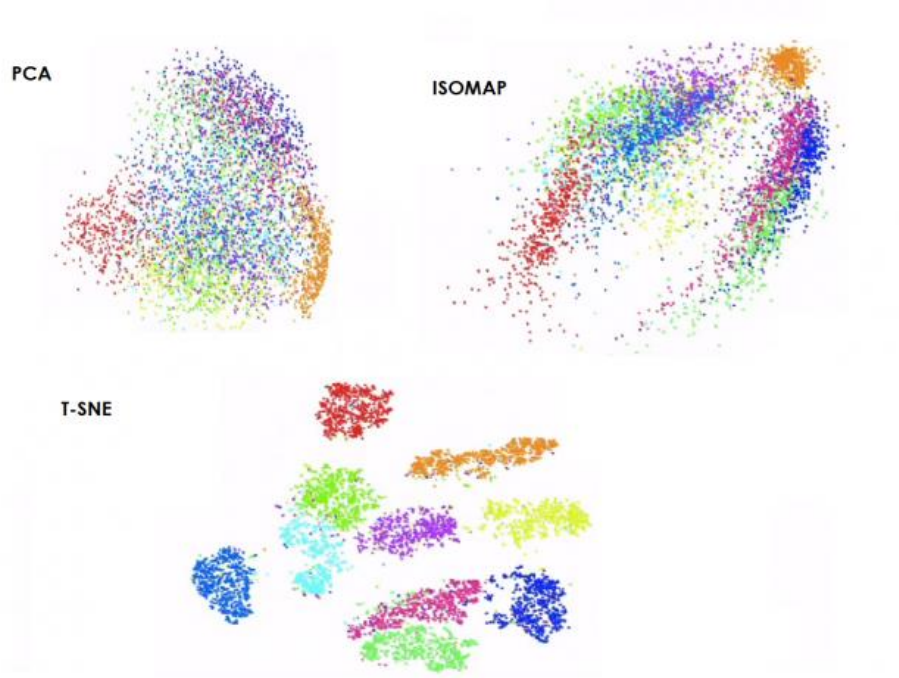
Accuracy

68%

Sensitivity

T-SNE

T-SNE: MAIN IDEA



1. Determine **similarity** of all the points in the scatter plot
2. Move similar points closer together

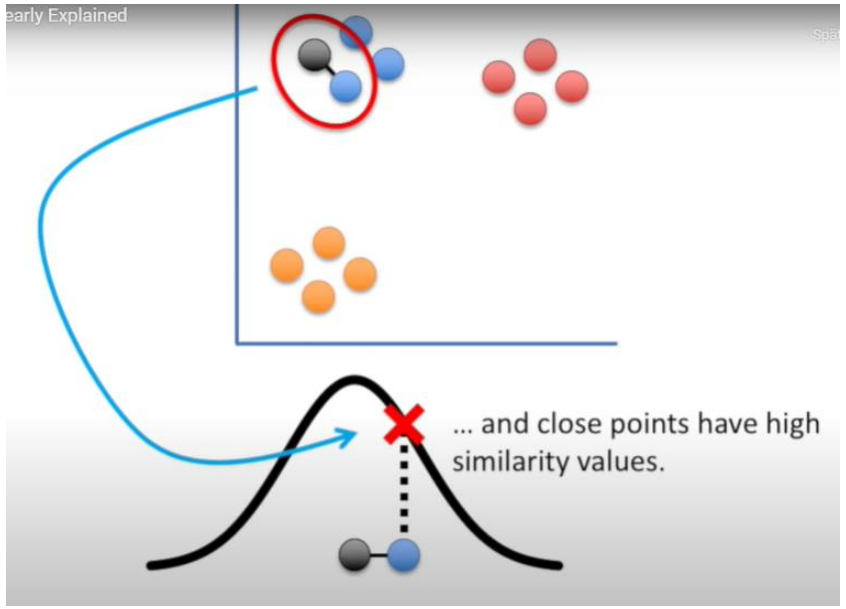
Project data into lower dim. space so that the clustering in higher spaces is preserved.

Unsupervsised

J. Starmer. 2017. StatQuest: t-SNE, Clearly explained. YouTube.

T-SNE: MAIN IDEA

How to determine the **similarity** of all the points in the scatter plot

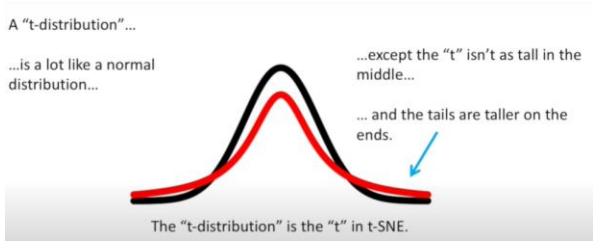


Measure the distance of a given point to all other points.

Plot all *point-point distances* and calculate the distance to a t-distribution centered around the point of interest.
-> „unscaled similarities“

Scale he similarities so that they add up to 1.

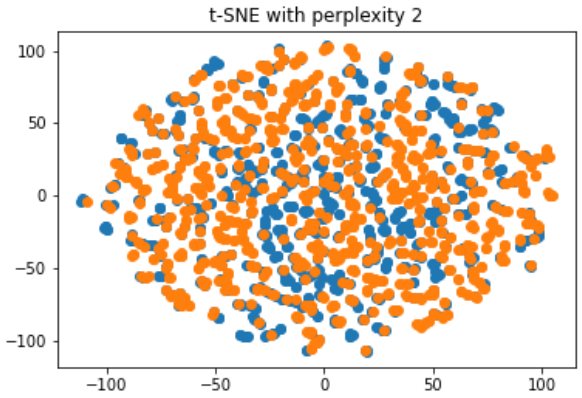
If points are normally distributed, close points will have high similarity values.

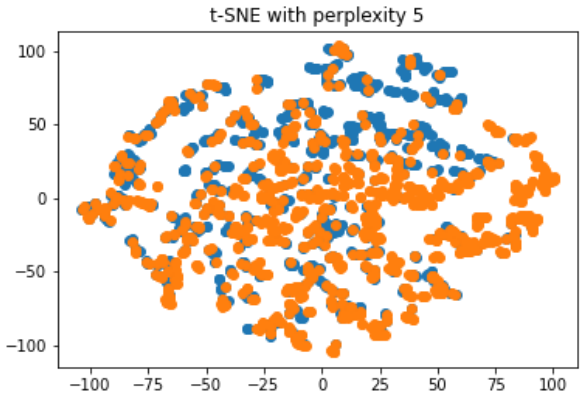


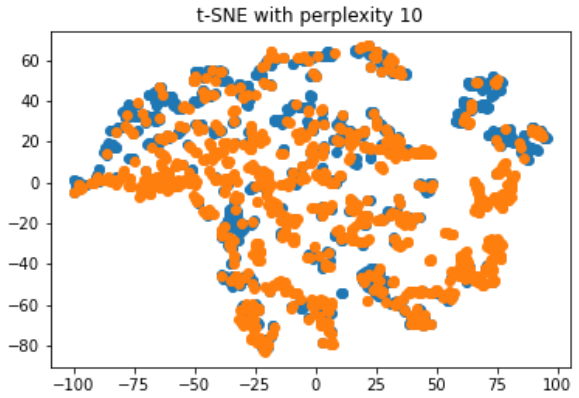
J. Starmer. 2017. StatQuest: t-SNE, Clearly explained. YouTube.

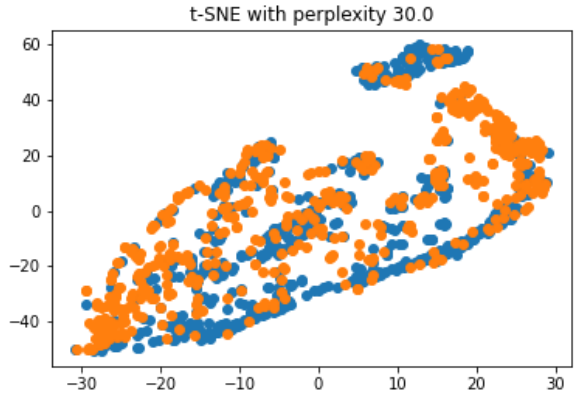
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

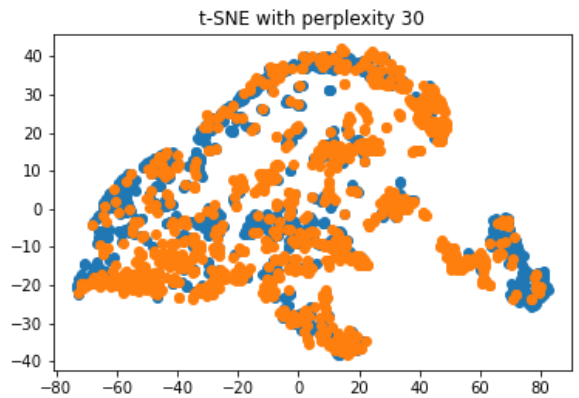
T-SNE: RESULTS

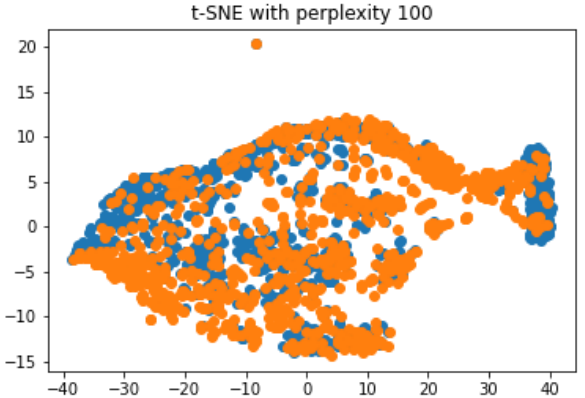




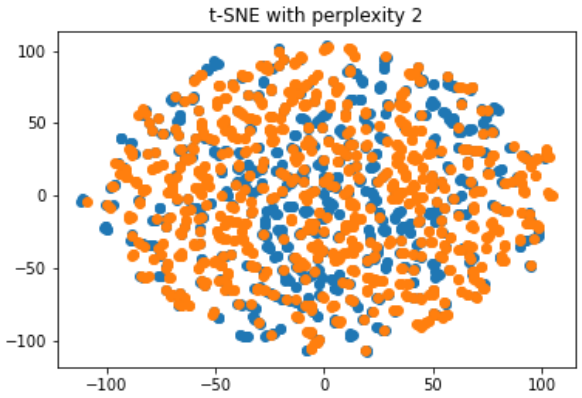


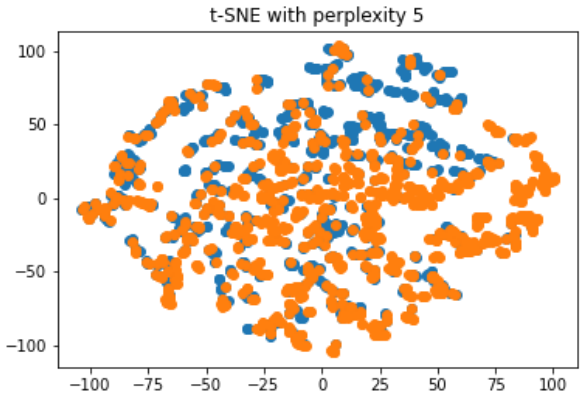


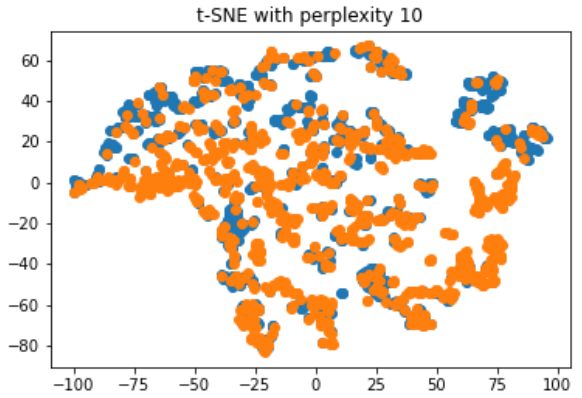


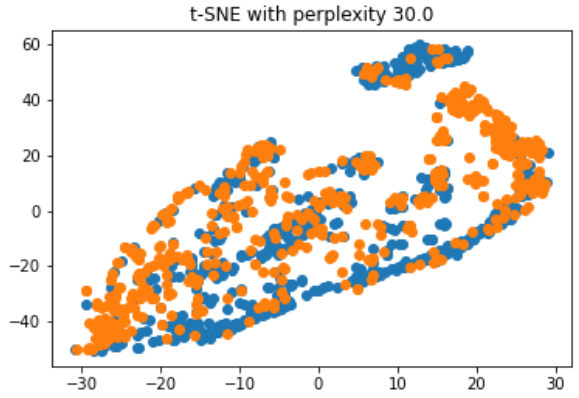


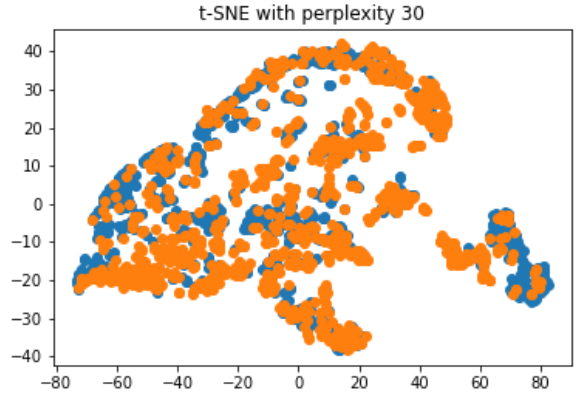
T-SNE: RESULTS

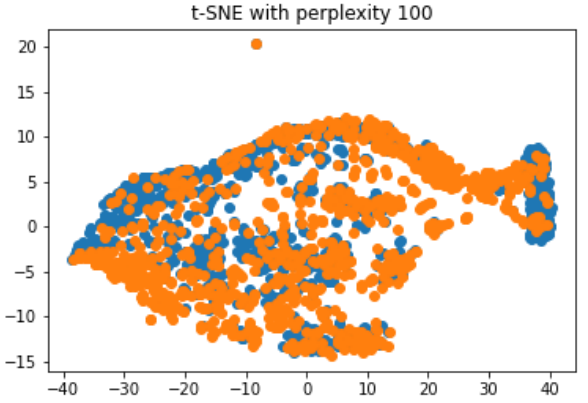








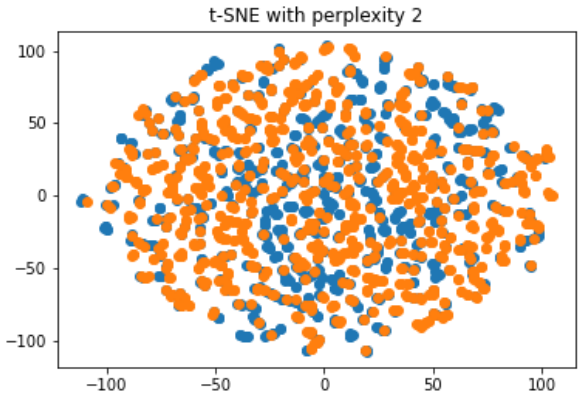


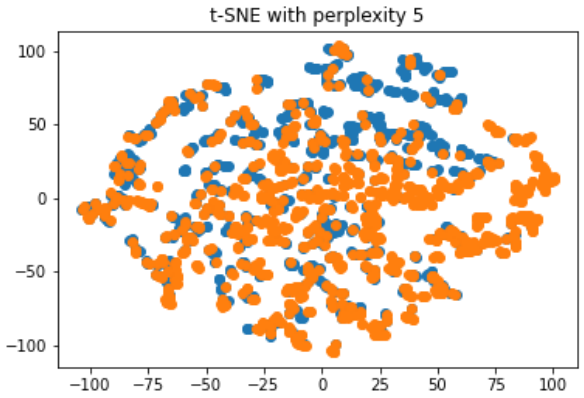


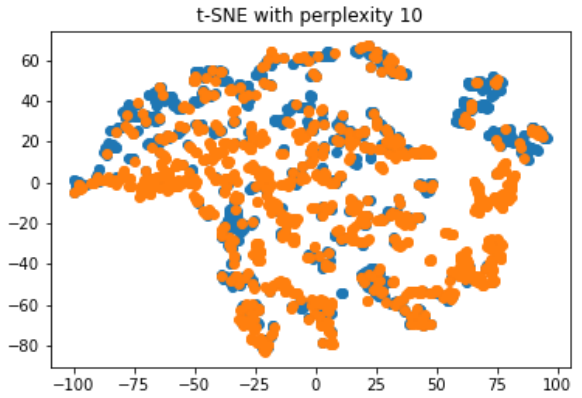
?

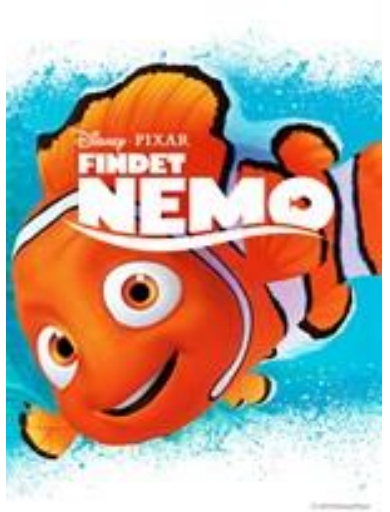
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

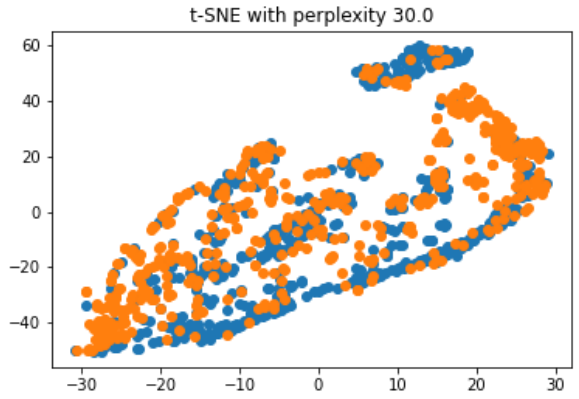
T-SNE: RESULTS

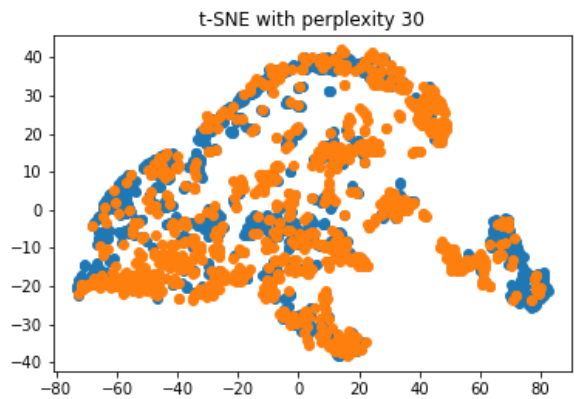


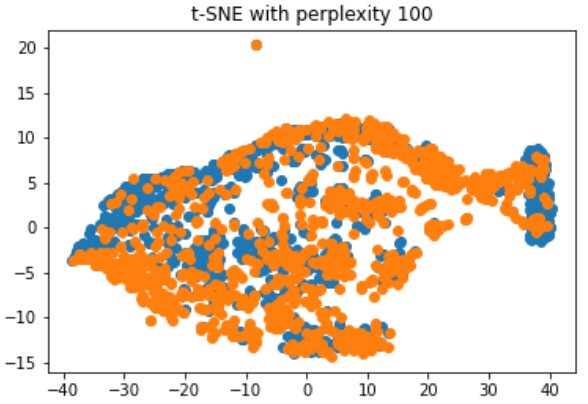










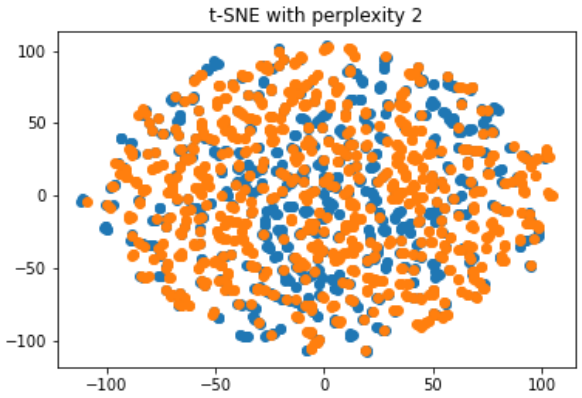


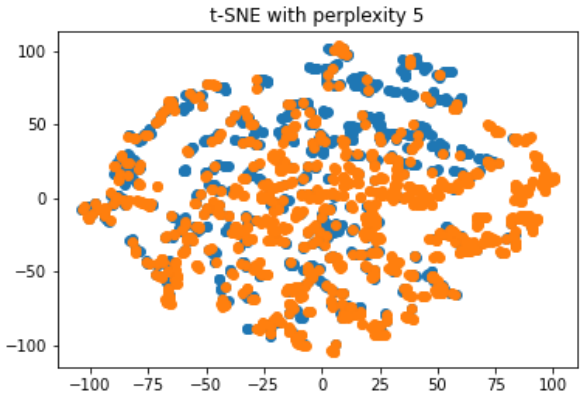
?

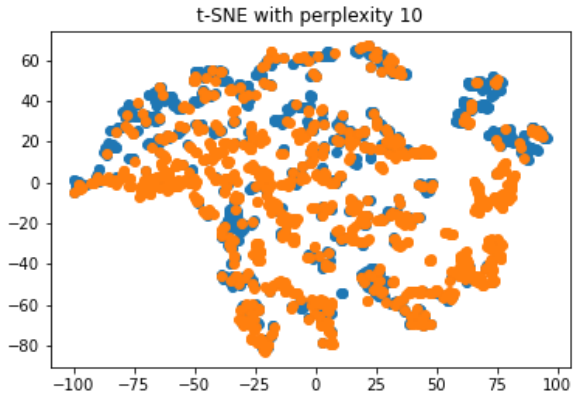
"Nemo" Image: https://www.microsoft.com/de-de/p/findet-nemo/8d6kgwxn0hk1?activetab=pivot%3aoverviewtab

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

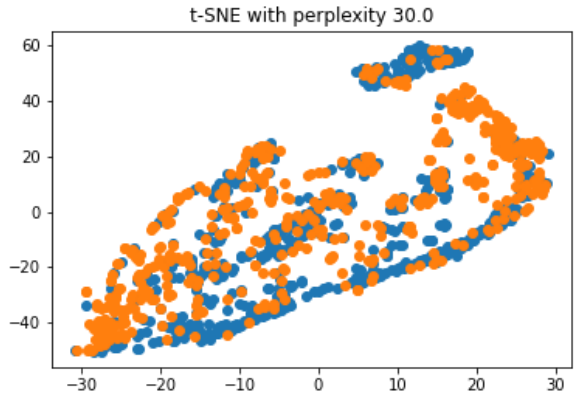
T-SNE: RESULTS

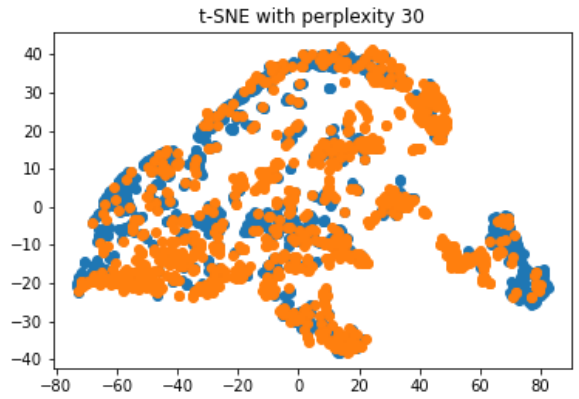


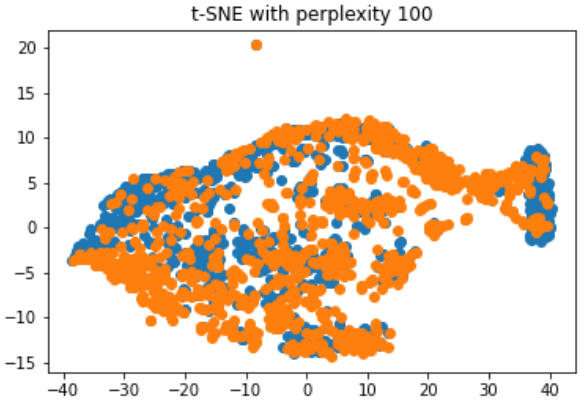




t-SNE is <u>not</u> useful in this case!







?

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

WHAT ABOUT A RANDOM FOREST?

DECSION TREES & RANDOM FOREST

DECISION TREES – MAIN IDEA

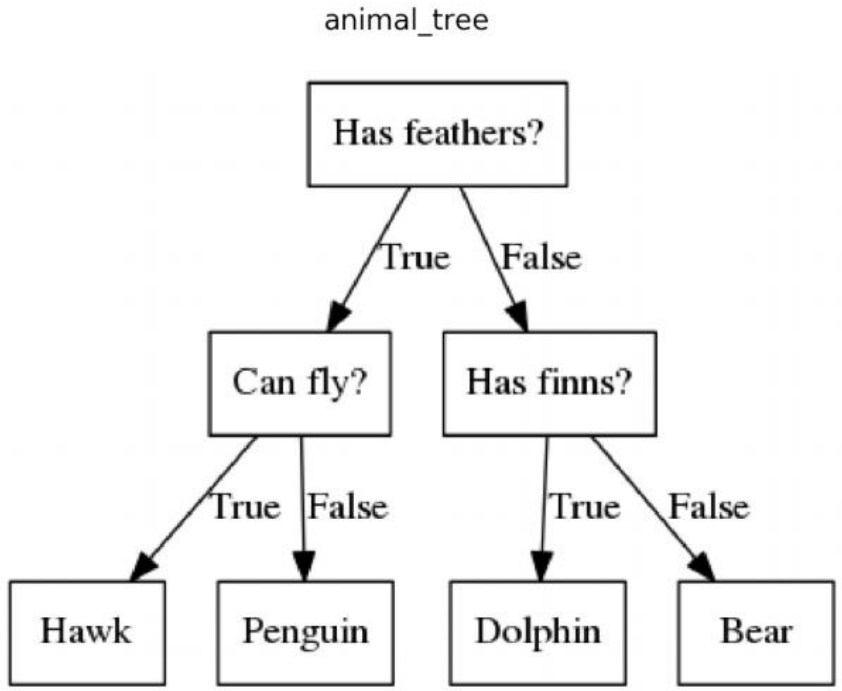
How are attributes selected?

Rating either via

"Gini index
The measure of the degree of probability of a particular variable being wrongly classified when it is randomly chosen."

or

"Information Gain
Entropy is the main concept of this algorithm, which helps [to] determine a feature or attribute that gives maximum information about a class. "



Source: towardsai.net

DECISION TREES – RESULTS – SLOW PION | ELECTRON

Without Pruning

20% Test data, 80% Training data
No scaling, Gini Entropy
Classes: Slow Pion & BG

Data	**Accuracy [%]**
Training	100
Test	81.8

Overfitted! Tree needs to be „cut“.

With Pruning

Max. depth	5	10	16	20
Data	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**
Training	72.4	82.8	89.6	94.0
Test	71.8	82.0	84.4	83.6

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

DECISION TREES – RESULTS – SLOW PION | ELECTRON

Without Pruning

20% Test data, 80% Training data
With scaling, Gini Entropy
Classes: Slow Pion & BG

Data	**Accuracy [%]**
Training	100
Test	82.2

Overfitted! Tree needs to be „cut“.

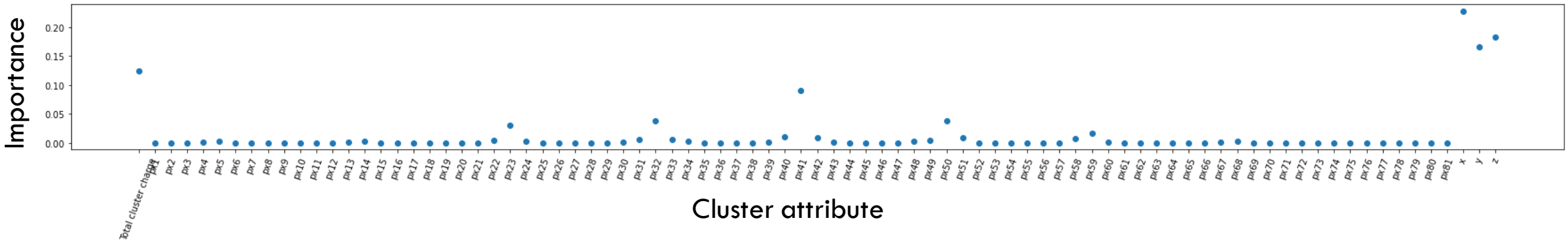
With Pruning

Max. depth	5	10	18	20
Data	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**
Training	72.4	83.0	91.9	93.5
Test	72.2	81.5	84.3	84.1

DECISION TREES – RESULTS – SLOW PION | ELECTRON

Relative feature importance

according to a decision tree with max_depth=18, scaled dataset



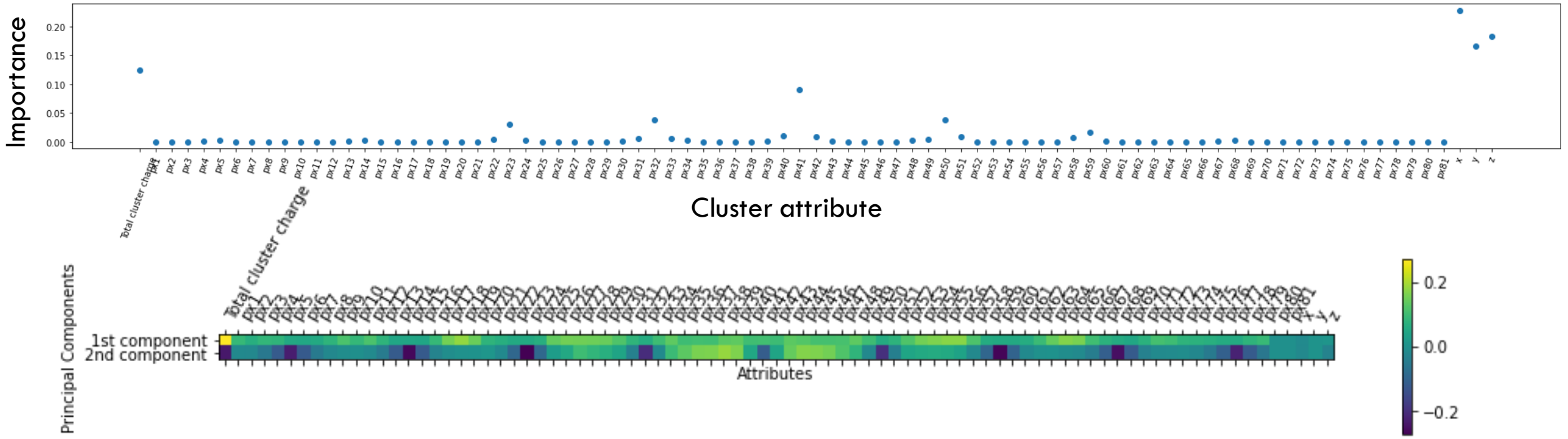
Feature	Total charge	Pxl #23	Pxl #32	Pxl #41	Pxl #50	x	y	z
Importance [%]	12.4	3.05	3.87	9.09	3.77	22.8	16.5	18.3

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

DECISION TREES – COMPARISON WITH PCA RESULTS

Relative feature importance

according to a decision tree with max_depth=18, scaled dataset



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

DECISION TREES – RESULTS – SLOW PION | ELECTRON

Without Pruning

20% Test data, 80% Training data
No scaling, Gini Entropy
Classes: Slow Pion & Electron

Data	**Accuracy [%]**
Training	100
Test	74.2

Overfitted! Tree needs to be „cut“.

With Pruning

Max. depth	5	10	15
Data	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**
Training	81.7	84.3	89.4
Test	81.5	80.9	78.8

DECISION TREES – RESULTS – SLOW PION | ELECTRON

Without Pruning

20% Test data, 80% Training data
With scaling, Gini Entropy
Classes: Slow Pion & Electron

Data	**Accuracy [%]**
Training	100
Test	74.0

Overfitted! Tree needs to be „cut“.

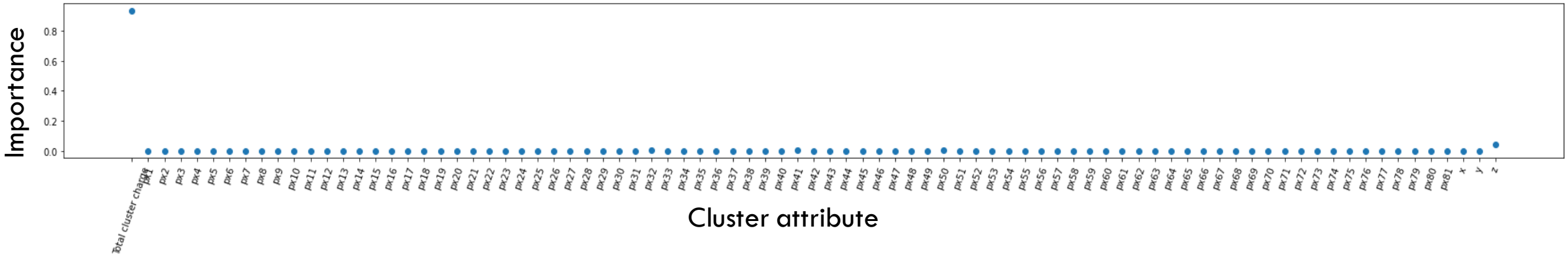
With Pruning

Max. depth	5	6	10	20
Data	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**	**Accuracy [%]**
Training	81.3	81.8	84.3	94.6
Test	81.3	81.5	80.7	76.8

DECISION TREES – RESULTS – SLOW PION | ELECTRON

Relative feature importance

according to a decision tree with max_depth=18, scaled dataset



Feature	Total charge	x	y	z
Importance [%]	93.3	<1%	1.02	4.16

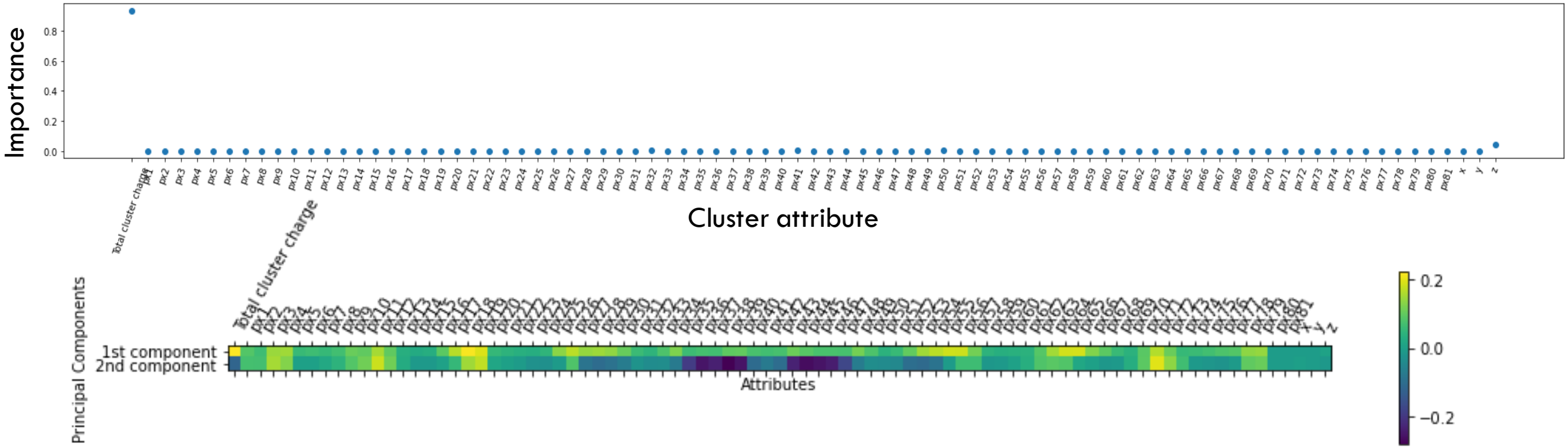
Major difference to Slow Pion | BGD !!!

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

DECISION TREES – COMPARISON WITH PCA RESULTS

Relative feature importance

according to a decision tree with max_depth=18, scaled dataset

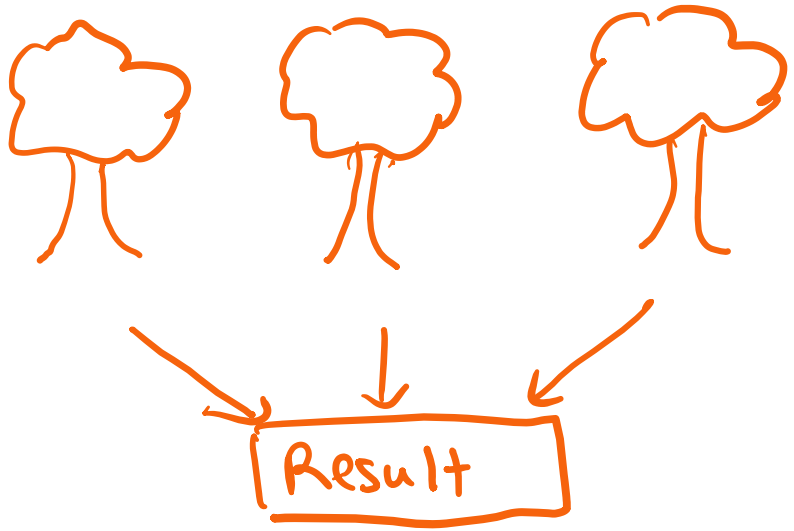


JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

RANDOM FOREST

= Combination of several decision trees

- All trees show different tendencies for overfitting
- Ensemble performance is better than the one of a single tree



RESULTS

Classes: Slow Pion | BGD
With scaling, n = 100

Data	**Accuracy** [%]
Training	100
Test	87.8

Classes: Slow Pion| BGDs
With scaling, n = 100

Data	**Accuracy** [%]
Training	100
Test	81.2

SUMMARY & OUTLOOK

GOAL:
SEPARATE SLOW PIONS FROM BGD

FUTURE PROJECTS:

Use MLP or SVM on PCA pretransformed data, Multiclassifier, GANs

Summary

- LDA, t-SNE failed
- PCA seems promising
- > 80% accuracy using Random Forest

THANK YOU FOR YOUR ATTENTION!

Stephanie.kaes@physik.uni-giessen.de