JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

UNDERSTANDING THE 6-DIM DATA

RESULTS FROM PRINCIPAL COMPONENTS ANALYSIS AND SELF-ORGANIZING MAPS

STEPHANIE KÄS

Belle II

Experimental Physics II

MINI-WORKSHOP
ANOMALY DETECTION WITH NEURAL NETWORKS
GIESSEN 21.02.2020

ATTENTION

This event will be photographed.

Images may be used for JLU promotional material or publications.

If you do not want your image to be used, please tell me.

OVERVIEW

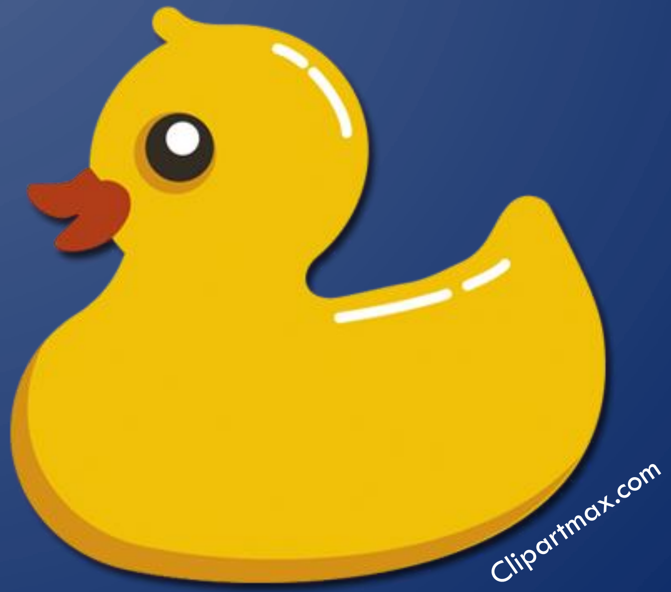
1. Principal components analysis
2. Antideuteron data set
3. Results of bachelor‘s thesis

PART 1

PRINCIPAL COMPONENTS ANALYSIS

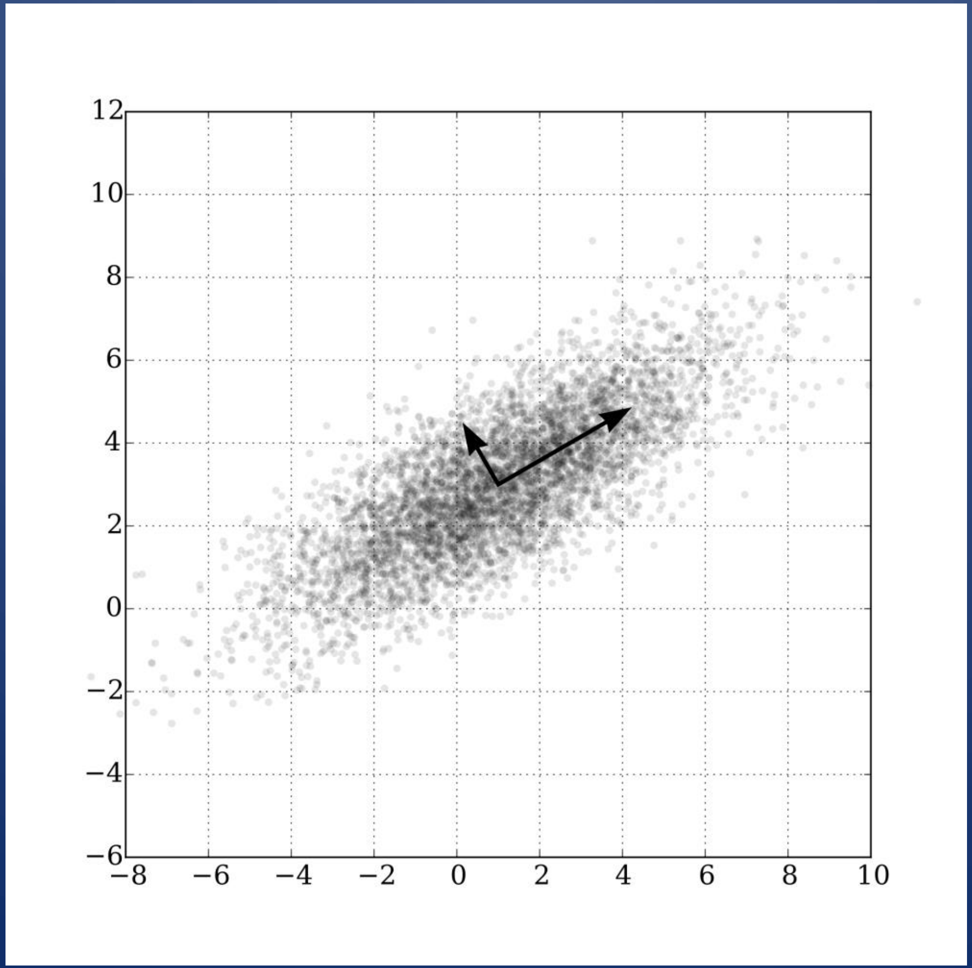
PRINCIPAL COMPONENTS ANALYSIS
IDEA

3D duck



2D duck

PRINCIPAL COMPONENTS ANALYSIS
DETAILS



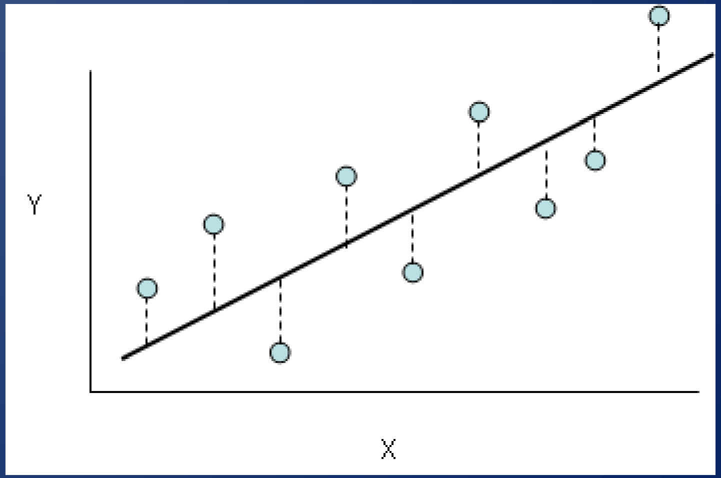
Wikipedia.org

PRINCIPAL COMPONENTS ANALYSIS
DETAILS

Given: data set of n attributes

Find n new axes by

- Minimization of error squares
- Maximization of variance



spcforexcel.com

PRINCIPAL COMPONENTS ANALYSIS
BASIC STATISTICS

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Covariance matrix

$$\Sigma = \begin{pmatrix} var(\vec{x}) & cov(\vec{x},\vec{y}) \\ cov(\vec{x},\vec{y}) & var(\vec{x}) \end{pmatrix}$$

For data set with two attributes $\vec{x}, \vec{y}$

Correlation matrix = normalized covariance matrix

PRINCIPAL COMPONENTS ANALYSIS
DETAILS

Find transformation matrix Γ so that Λ is diagonal.

$$\Lambda = \Gamma^T \Sigma \Gamma = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

Columns of Γ are the axes of the new coordinate system.

PRINCIPAL COMPONENTS ANALYSIS
DETAILS

Higher values of λ_i $\leftrightarrow$ higher information content

$$\Lambda = \Gamma^T \Sigma \Gamma = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{pmatrix}$$

$$\lambda_1 \geq \lambda_2 \ldots \geq \lambda_n$$

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

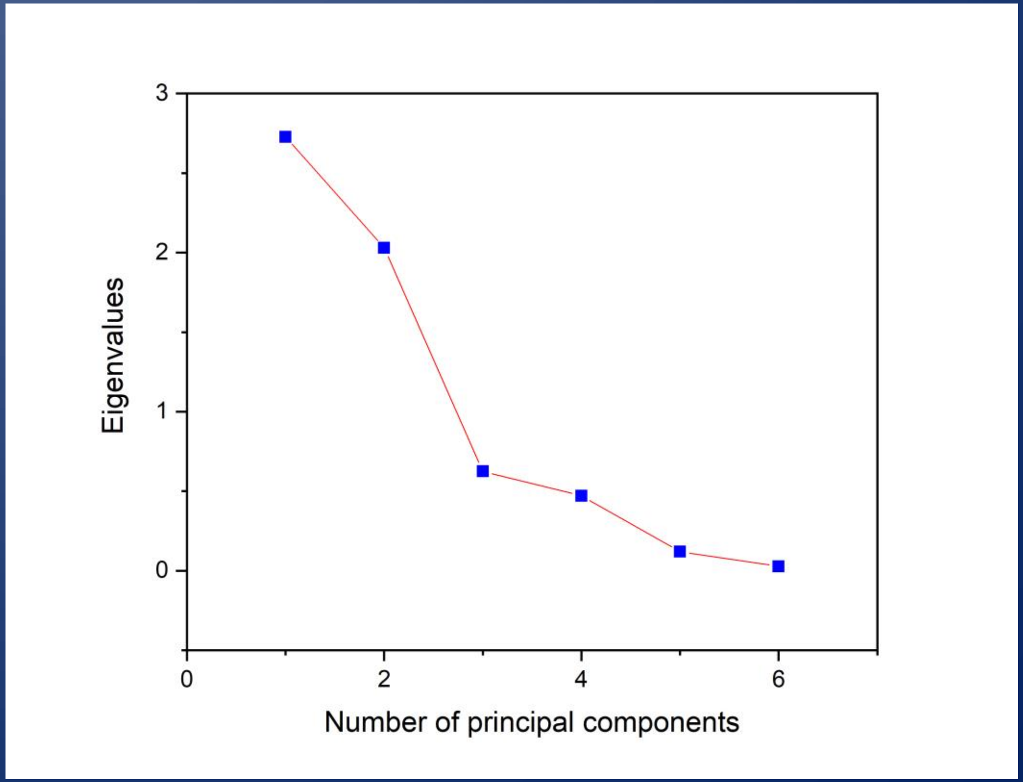
PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Eigenvalues

λ_i	Cum. Sum [%]
2.73	45.46
2.03	79.28
0.62	89.62
0.47	97.55
0.12	99.54
0.03	100

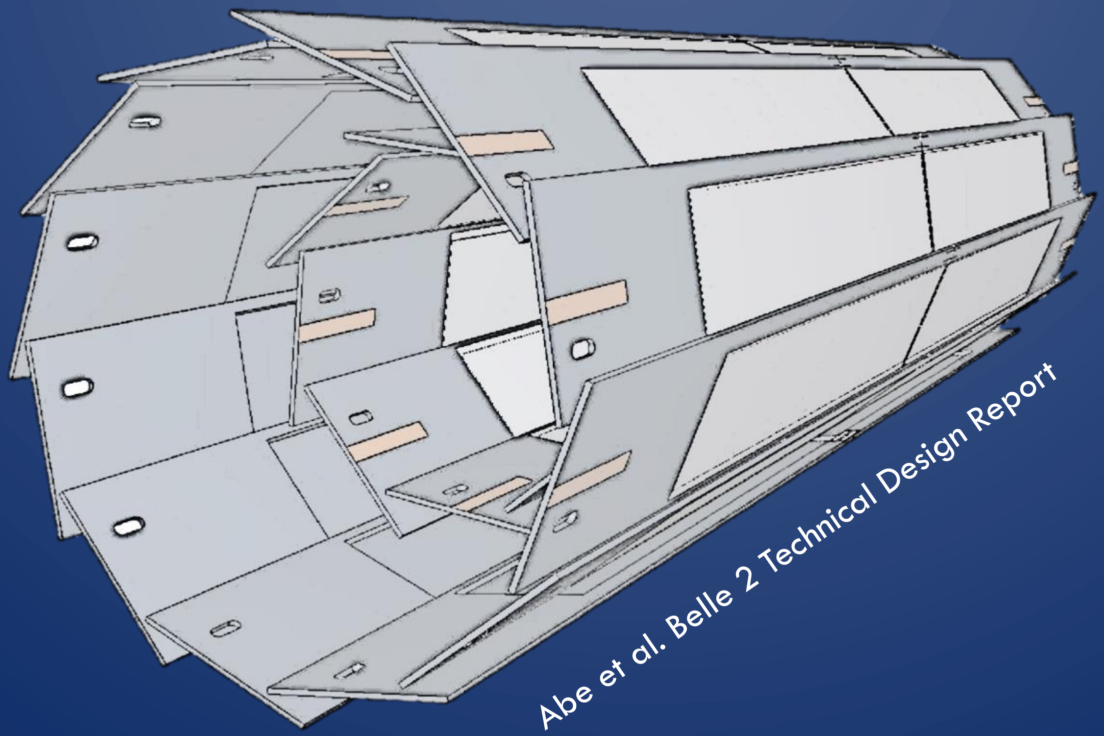
Scree graph (PXD)



PART 2

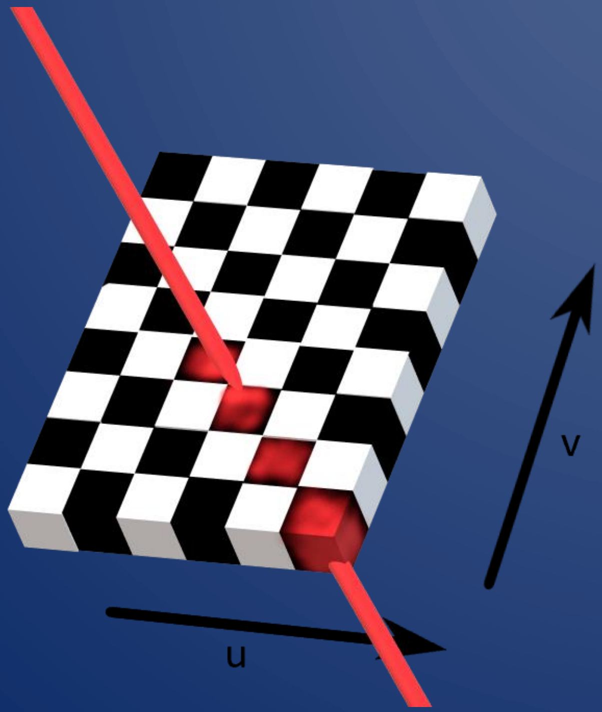
PIXEL DETECTOR'S ANTIDEUTERON DATA SET

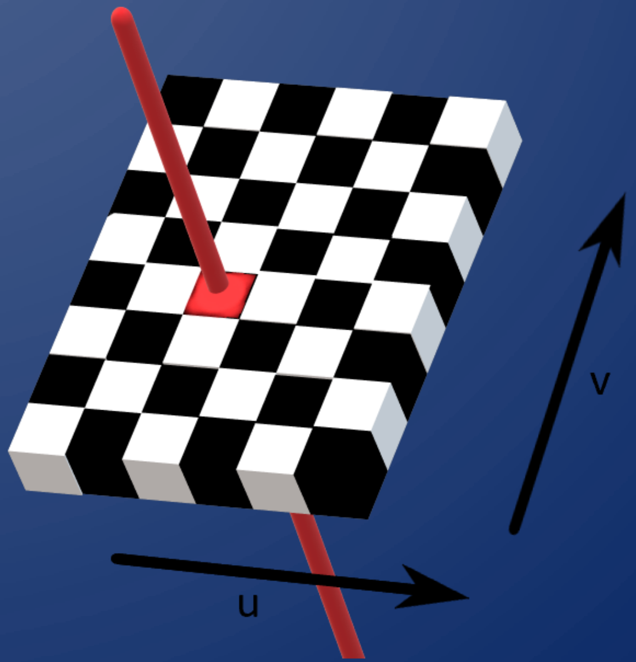
PIXEL DETECTOR



Abe et al. Belle 2 Technical Design Report

PIXEL DETECTOR



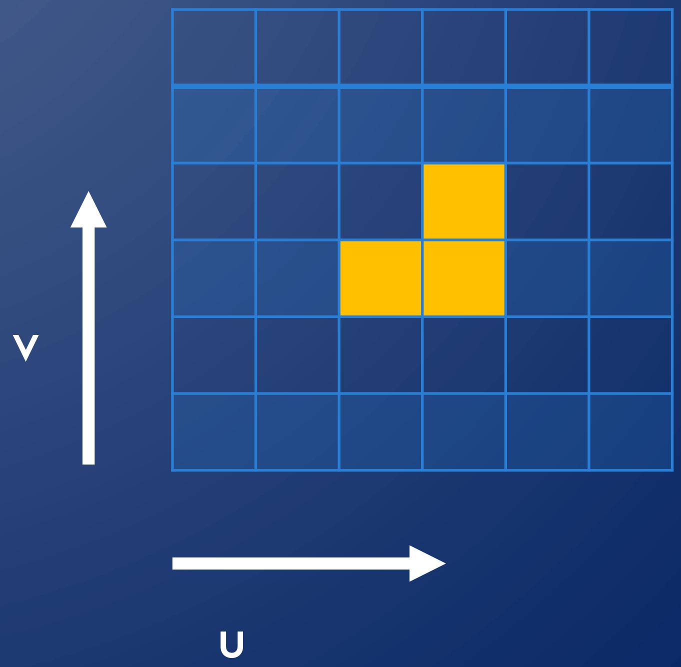


PIXEL DETECTOR
THE ANTIDEUTERON DATA SET

Group pixels into clusters

Cluster properties

- Total charge
- Seed charge
- Minimum charge



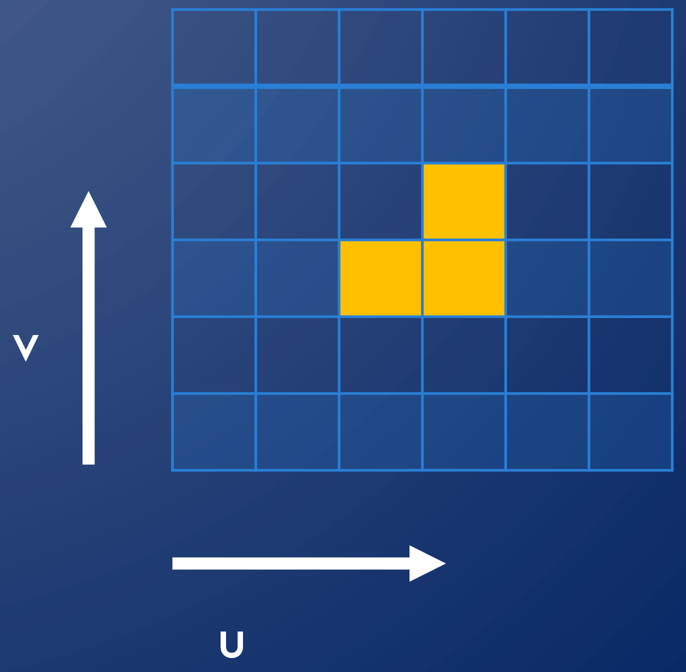
PIXEL DETECTOR
THE ANTIDEUTERON DATA SET

Group pixels to clusters

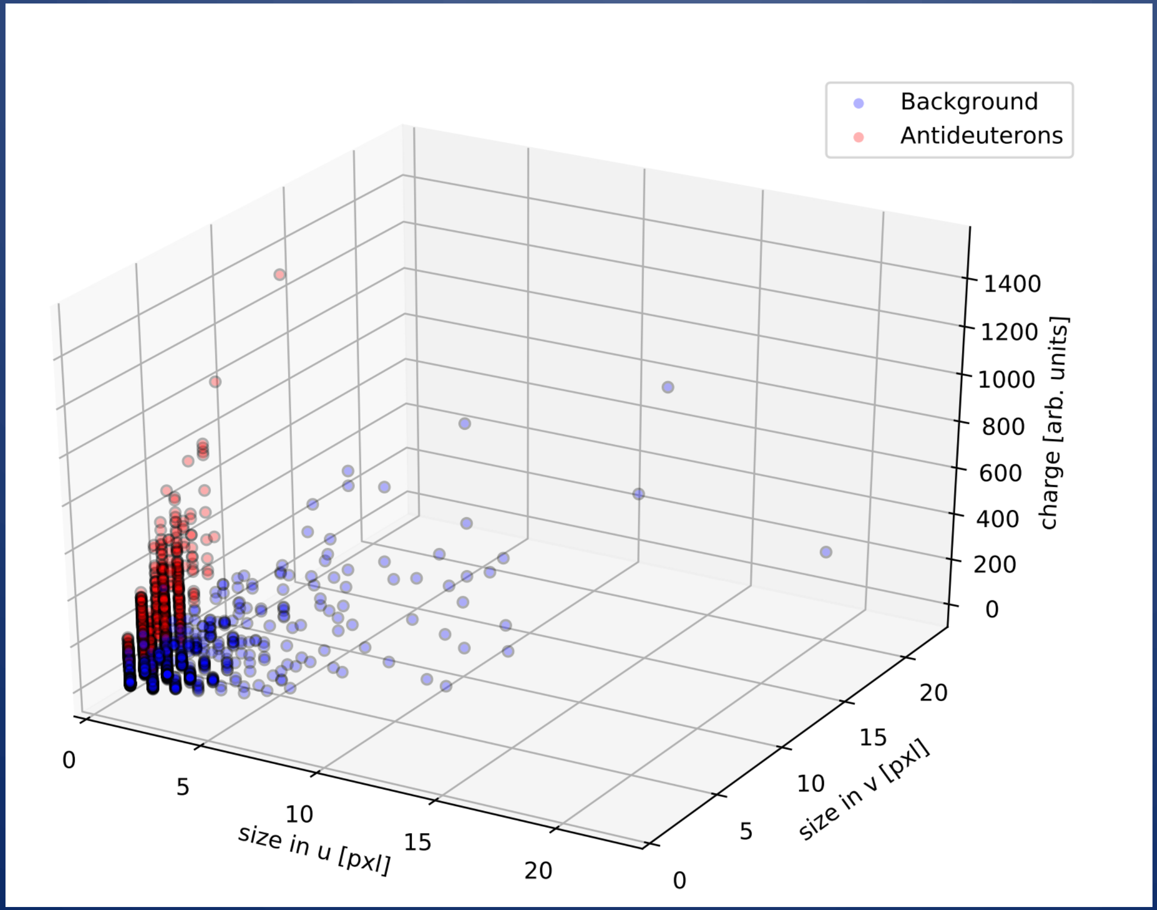
Cluster properties

- Total charge
- Seed charge
- Minimum charge

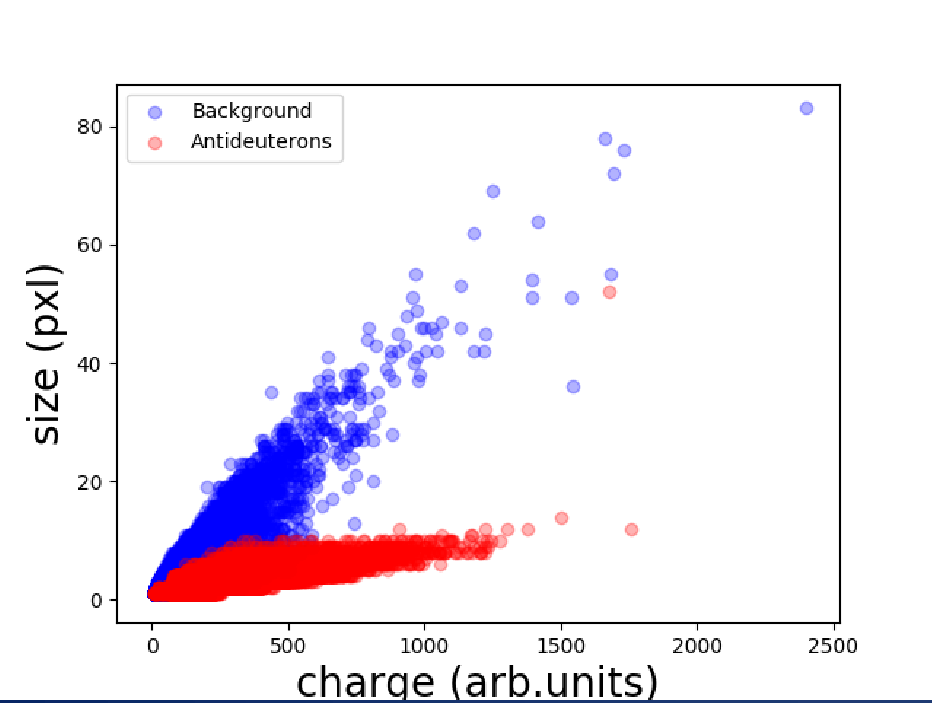
- Total size
- Size in u
- Size in v

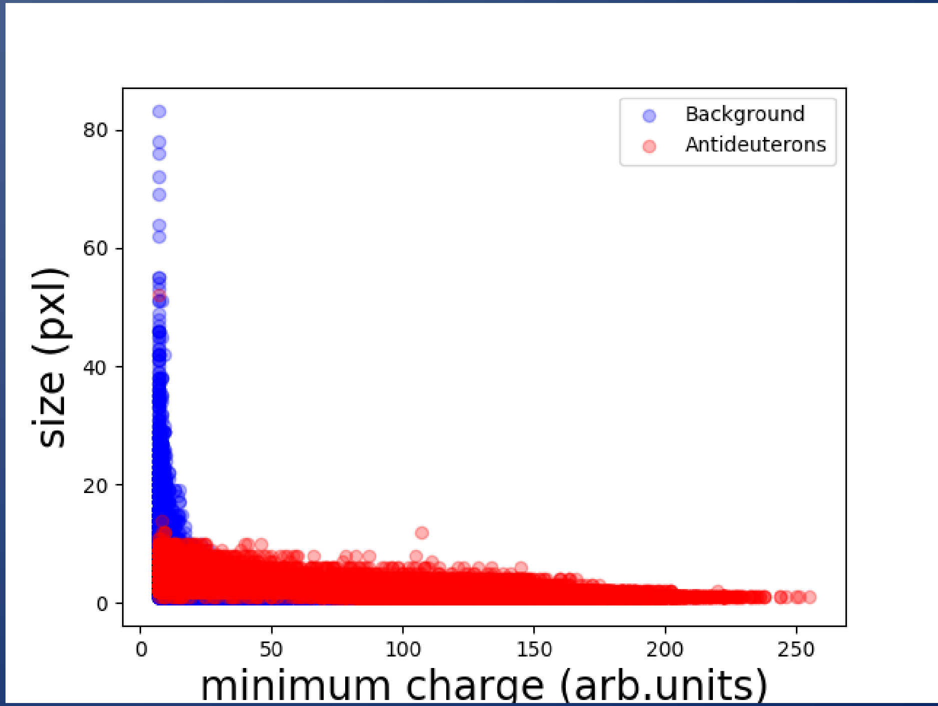


PIXEL DETECTOR
THE ANTIDEUTERON DATA SET



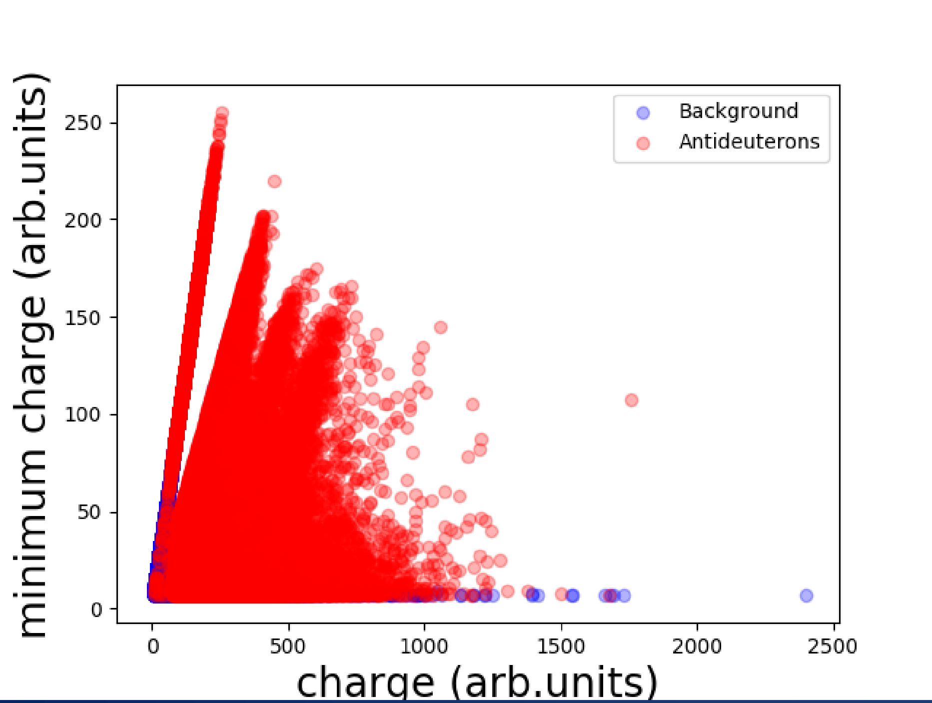
PIXEL DETECTOR
THE ANTIDEUTERON DATA SET

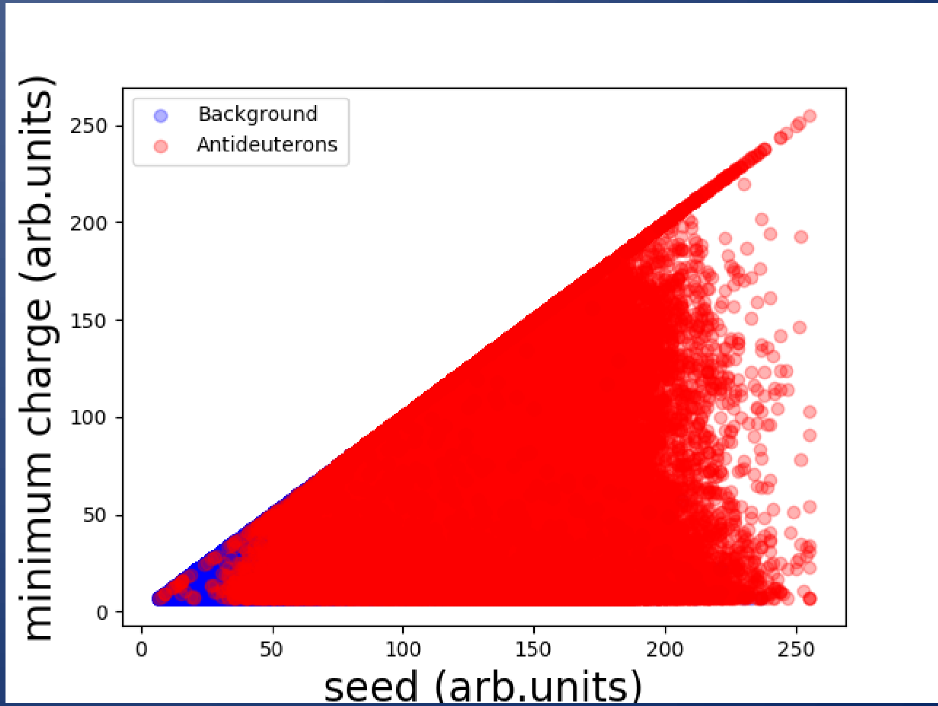




JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PIXEL DETECTOR
THE ANTIDEUTERON DATA SET





PART 3

BACHELOR'S THESIS

BACHELOR‘S THESIS - GOALS

1. Goal: Better understanding of data set
   - Find correlations in cluster properties -> PCA
   - Cluster shapes
2. Goal: Separate particles from background
   - Use SOMs: Separate more than 2 particles from background
   - Is pre-transformation into PCA-space helpful?

PROJECT 1: MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Which correlations between the 6 cluster properties exist ?

Try PCA!

PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Correlation matrix (lower half)

Cluster Property	Charge	Min. Charge	Seed	Size	Size in u	Size in v
Charge	1					
Min. Charge	0.2233	1				
Seed	0.7854	0.4771	1			
Size	0.4617	-0.2882	0.0392	1		
Size in u	0.1596	-0.2600	-0.1399	0.8044	1	
Size in v	0.4144	-0.2091	0.0546	0.8414	0.4627	1

Several high correlations!

PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Interpretation of principal components

PC1 = Measure for size

PC2 = Measure for charge

PC3 = ?

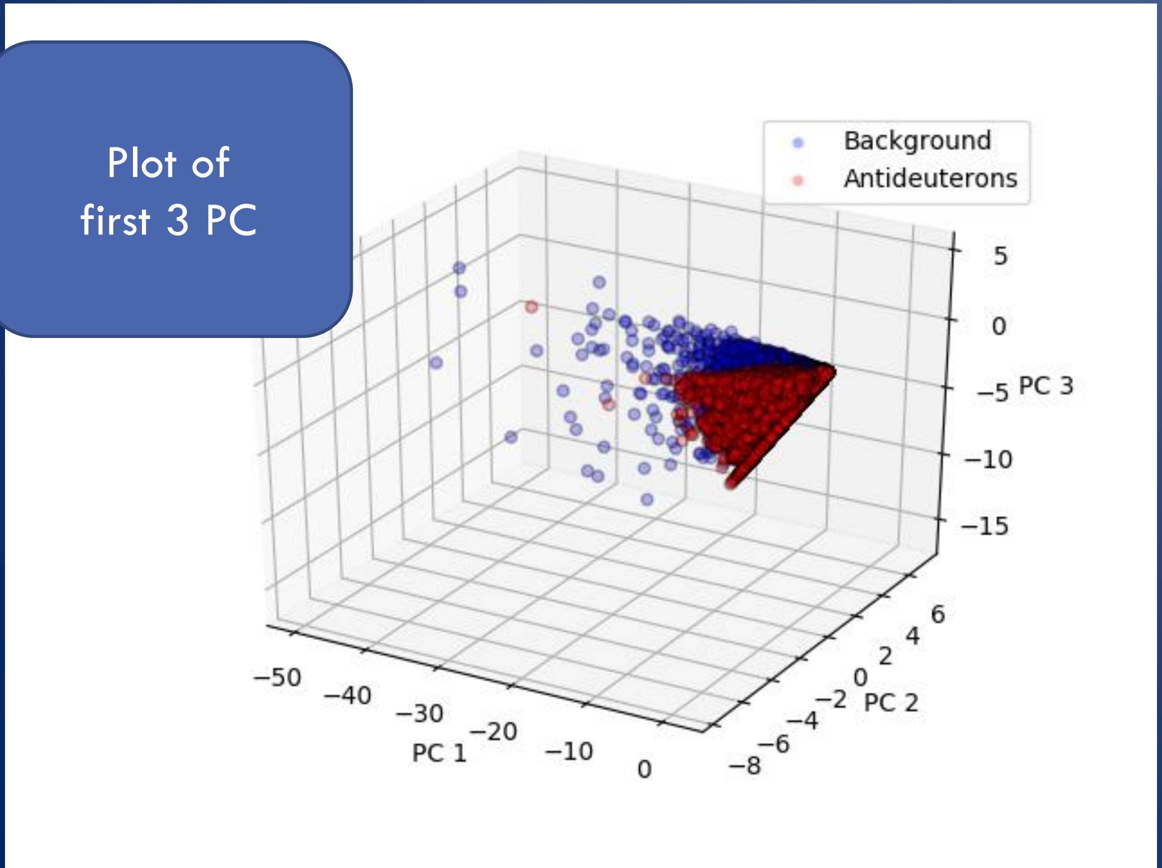
79.28 % of total information

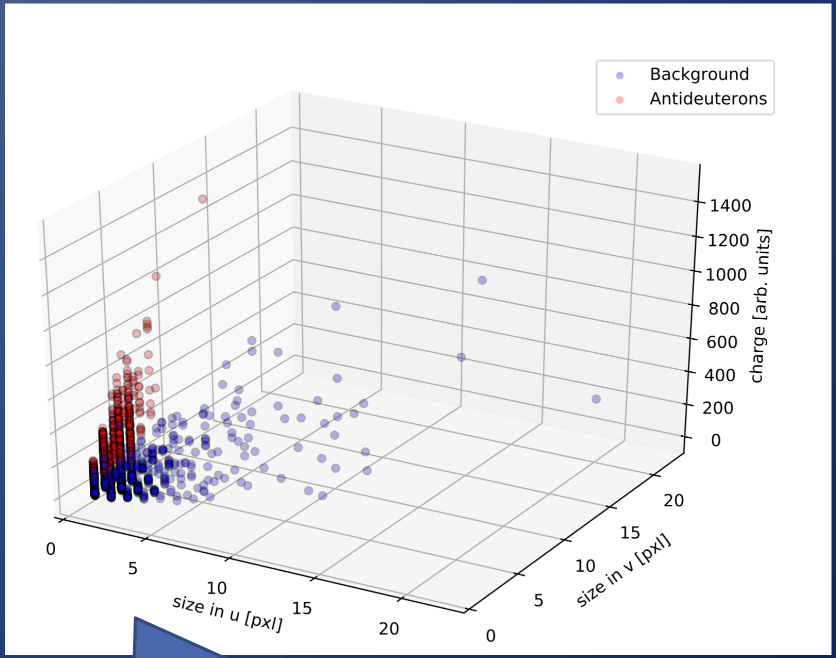
89.69 %

PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Plot of first 3 PC



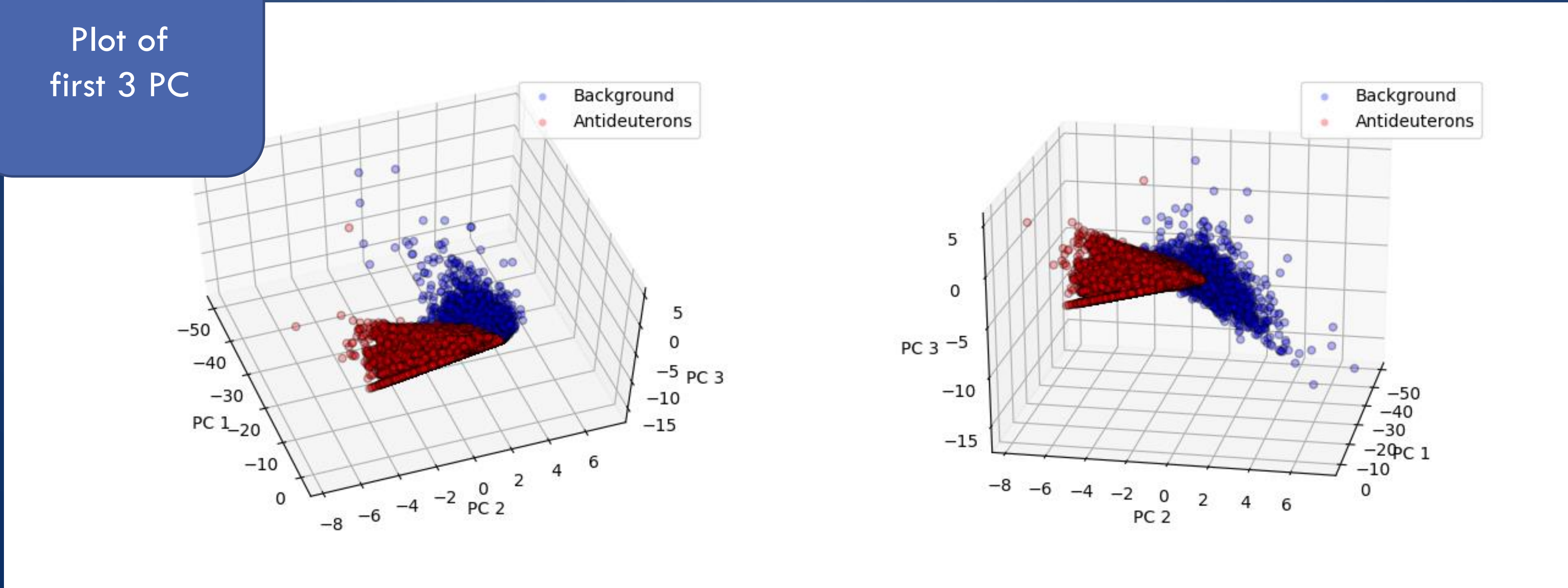


Original data set

PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Plot of first 3 PC



BACHELOR‘S THESIS - RESULTS

Dimension can be reduced to 3D

PROJECT 2: DATA SEPARATION USING SOMS

Use reduced data set as input for SOMs.

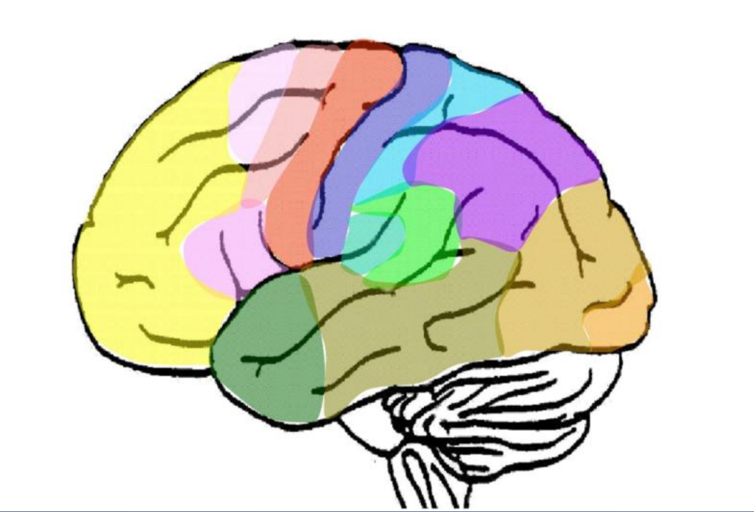
Try PCA & SOMs!

Does PCA-pre-transformation enhance SOM performance?

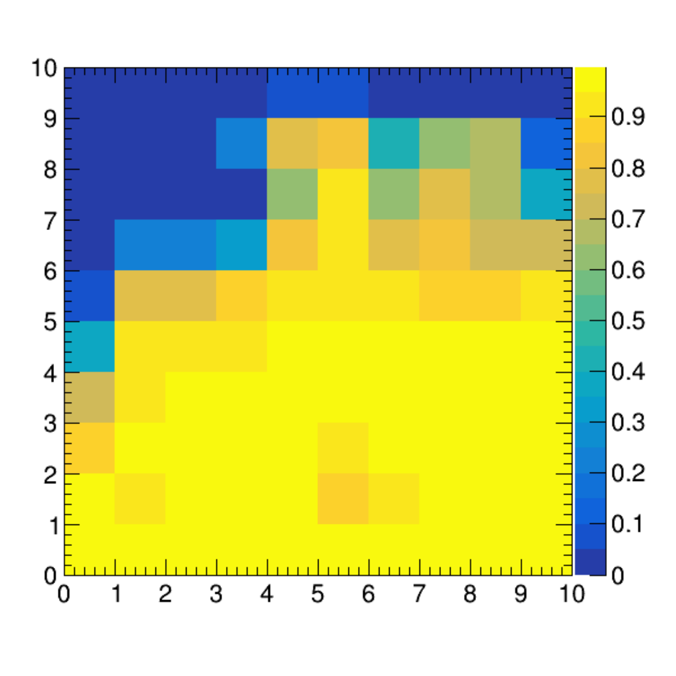
PROJECT 2: DATA SEPARATION USING SOMS

Idea: Separating particle signals from beam background

Method: Self-organizing maps



R. Westermann,
VL Neuroanatomie*



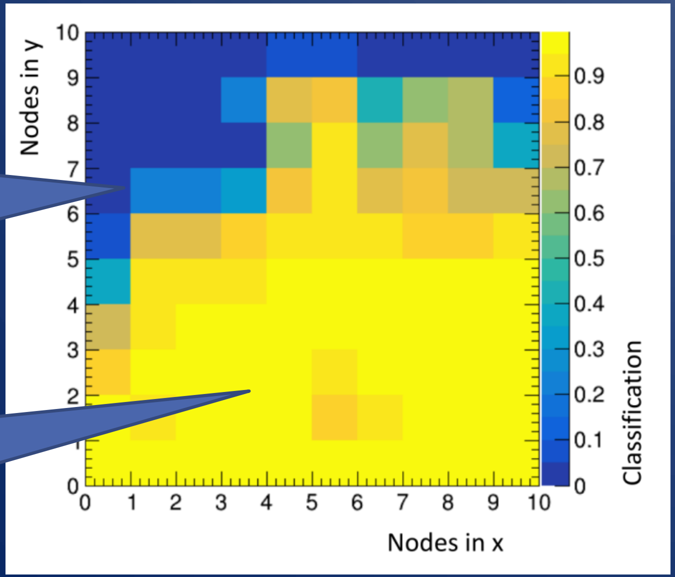
*Philipps Universität Marburg, SoSe 2016

PROJECT 2

DATA SEPARATION USING SOMS

ANTIDEUTERONS

ORIGINAL 6-DIM DATA SET



Background-like nodes

Antideuteron-like nodes

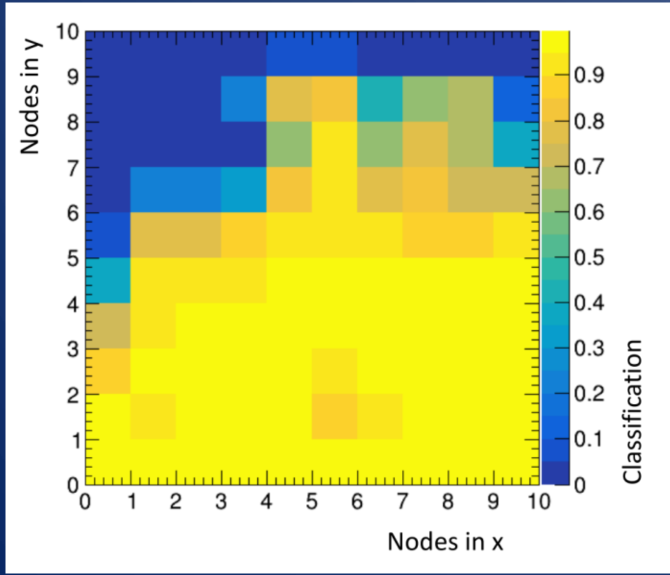
- Separation successful
- Results of Katharina's thesis confirmed

PROJECT 2

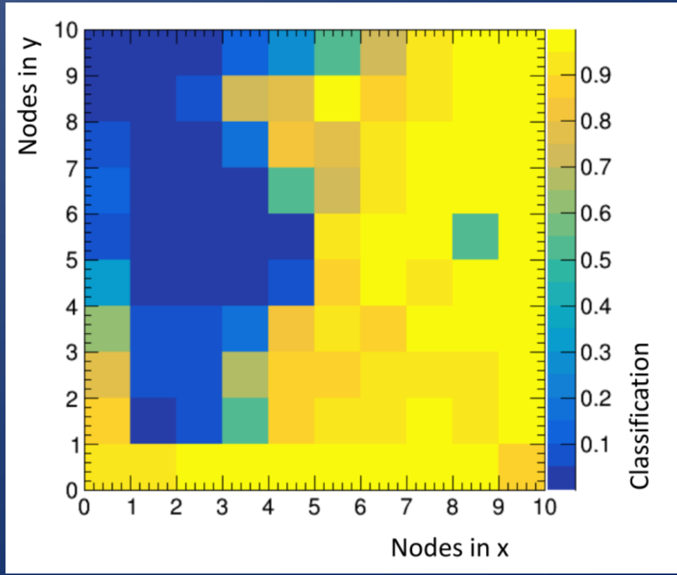
DATA SEPARATION USING SOMS

ANTIDEUTERONS

ORIGINAL 6-DIM DATA SET



PCA 6-DIM DATA SET



BACHELOR‘S THESIS - RESULTS

Dimension can be reduced to 3D

PCA does not enhance performance of SOMs

PROJECT 2

DATA SEPARATION USING SOMS

ANTIDEUTERONS, TETRAQUARKS AND PIONS

Can we separate 3 particles from background?

SOMs

Antideuterons, Tetraquarks and Pions

PROJECT 2

DATA SEPARATION USING SOMS

ANTIDEUTERONS, TETRAQUARKS AND PIONS

ROC-CURVES

"RECEIVER OPERATING CHARACTERISTIC"

Signal efficiency: True positive rate

$$P(classified\ as\ signal \mid signal)$$

Background rejection:

1- $P(classified\ as\ background \mid background)$

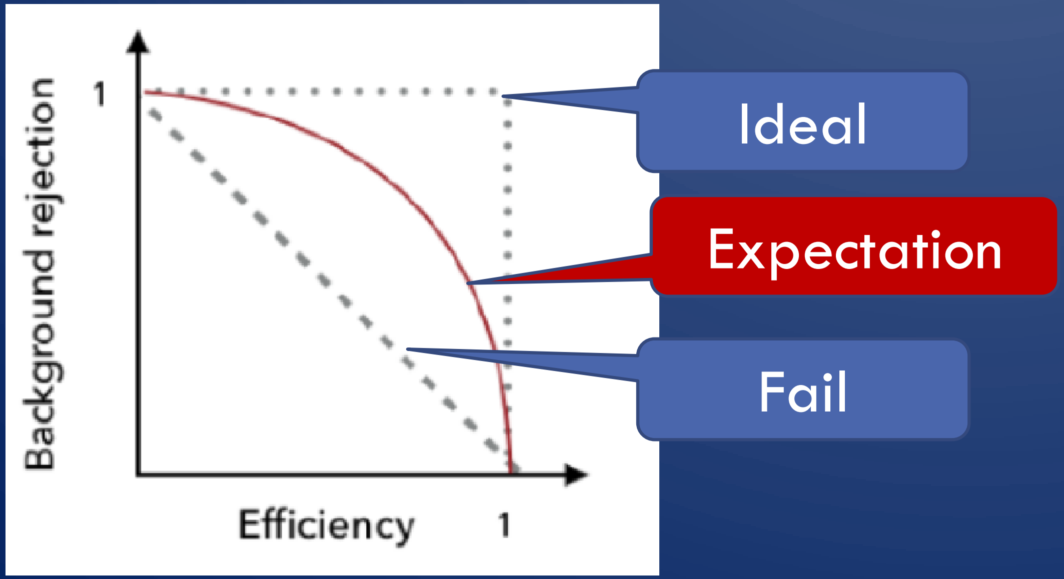
JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PROJECT 2

DATA SEPARATION USING SOMS

ANTIDEUTERONS, TETRAQUARKS AND PIONS

ROC-CURVES

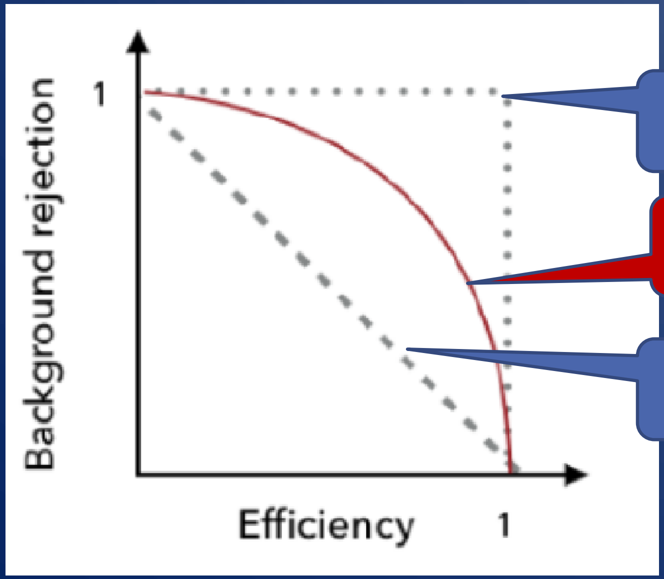


PROJECT 2

DATA SEPARATION USING SOMS

ANTIDEUTERONS, TETRAQUARKS AND PIONS

ROC-CURVES

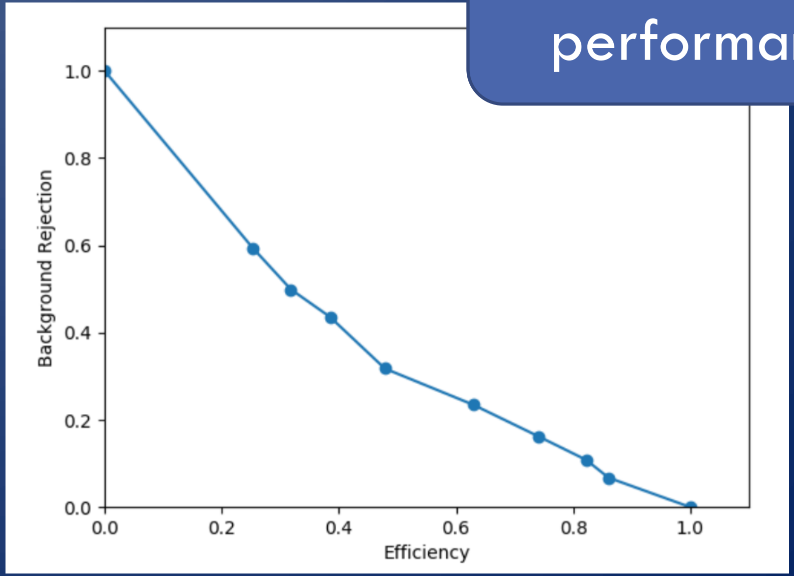


Ideal

Expectation

Fail

RESULT



Super bad performance!

BACHELOR'S THESIS - RESULTS

Dimension can be reduced to 3D

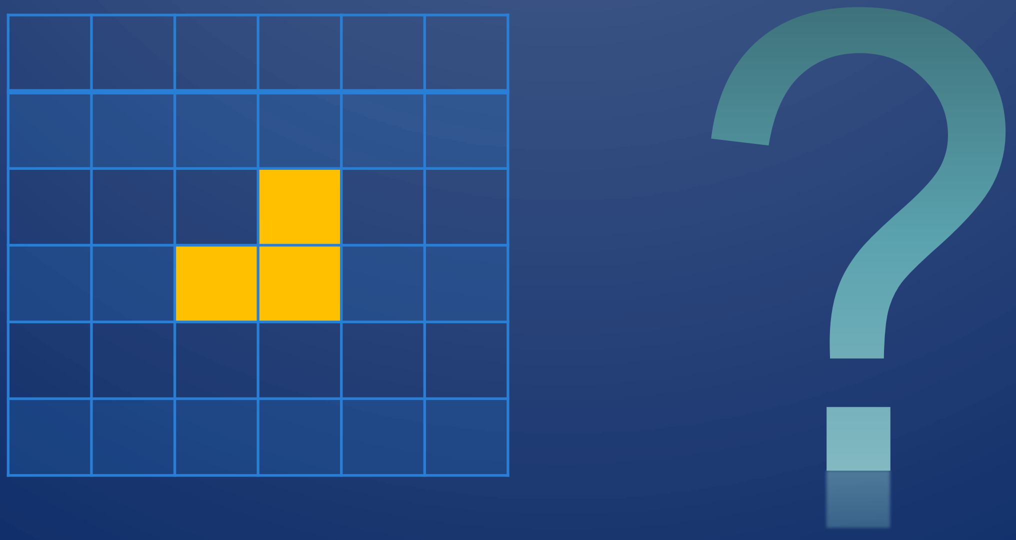
PCA does not enhance performance of SOMs

SOMs cannot separate more than 2 data types

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PROJECT 3: CLUSTER SHAPE ANALYSIS

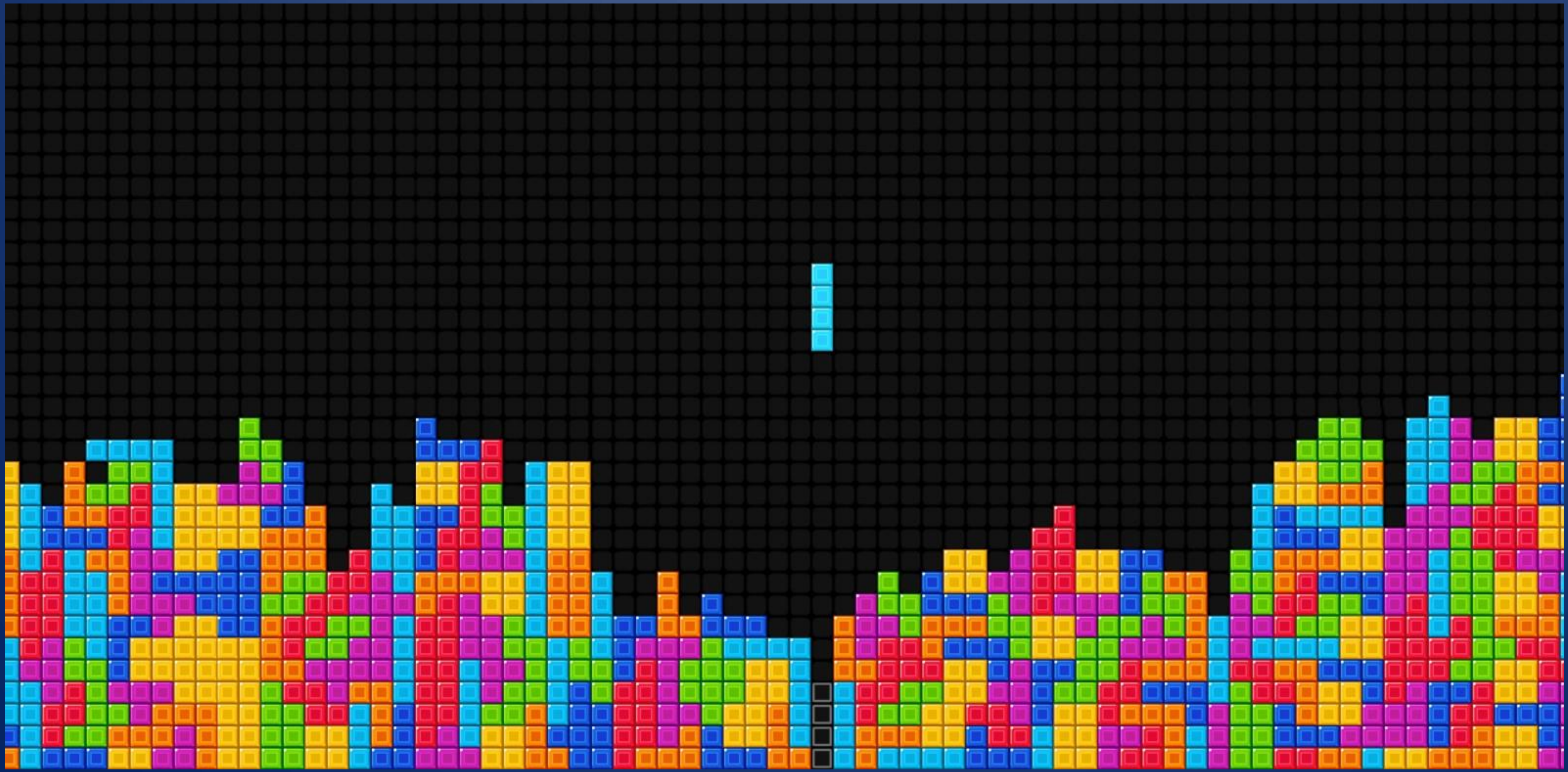
Which cluster shapes do appear?



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PROJECT 3

CLUSTER SHAPE ANALYSIS

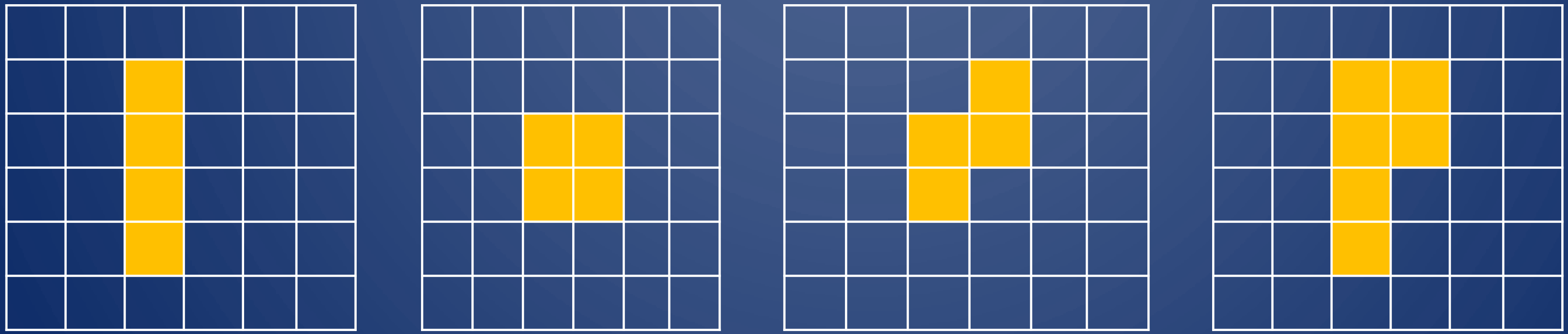


wallpaperaccess.com/retro-game

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

PROJECT 3

CLUSTER SHAPE ANALYSIS



Ca. 40% narrow rectangular

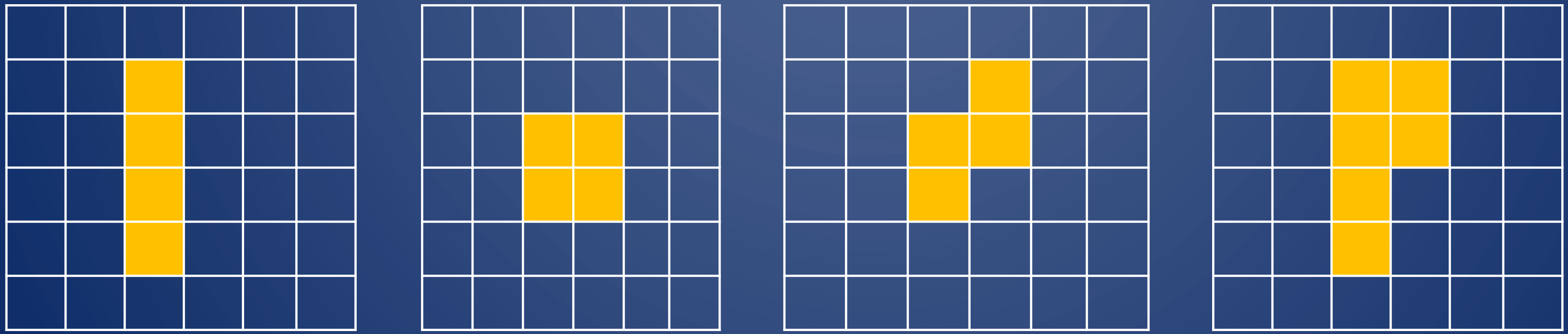
Ca. 17% squares

Ca. 1/3 2 pxl
Ca. 1/7 1 pxl

Ca. 10% > 6 pxl

PROJECT 3

CLUSTER SHAPE ANALYSIS



Pattern recognition?

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

BACHELOR‘S THESIS - RESULTS

Dimension can be reduced to 3D

PCA does not enhance performance of SOMs

SOMs cannot separate more than 2 data types

PXD clusters come in many different shapes

POSSIBLE FUTURE PROJECTS

- Compare different versions of PCA and SOMs
- Use pattern recognition on cluster shapes
  - Try reinforcement learning?
- Try support vector machines

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Thank you

for your attention!

N-te

Background: quantumfbi.com

PART 4

DETAILS ON PCA

PRINCIPAL COMPONENTS ANALYSIS

BASIC STATISTICS

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Variance

$$var(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

measure of spread of x_i

PRINCIPAL COMPONENTS ANALYSIS

BASIC STATISTICS

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

Variance

$$var(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Empirical covariance

$$cov(\vec{x}, \vec{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})$$

indication for correlations between $\vec{x}$ and $\vec{y}$

PRINCIPAL COMPONENTS ANALYSIS

DETAILS

Higher values of λ_i $\leftrightarrow$ higher information content

$$\lambda_1 \geq \lambda_2 \dots \geq \lambda_n$$

Information percentage of axis i (cumulative sum):

$$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \cdot 100\%$$

PRINCIPAL COMPONENTS ANALYSIS

CHOOSING NUMBER OF AXES

Critierion 1:

Choose only $\lambda_i \geq 1$

Criterion 2:

$$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \cdot 100\% \geq 90\%$$

Criterion 3:

Cut at first rapid change of slope in *Scree graph*.

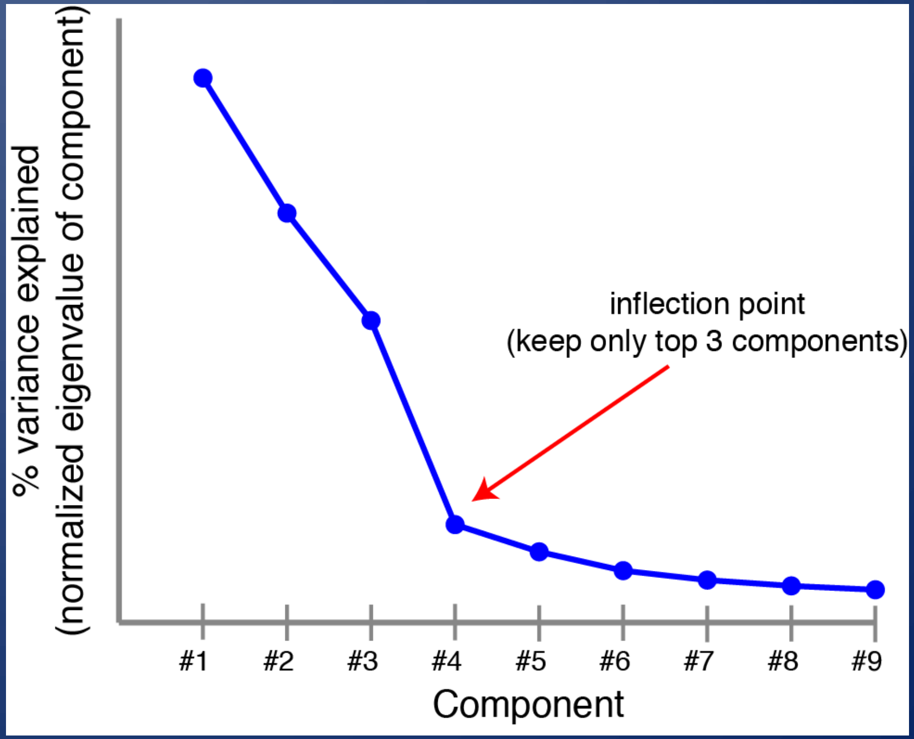
PRINCIPAL COMPONENTS ANALYSIS

CHOOSING NUMBER OF AXES

Criterion 3:

Cut at first rapid change of slope in *Scree graph.*

Scree graph (example)



alexhwilliams.info

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

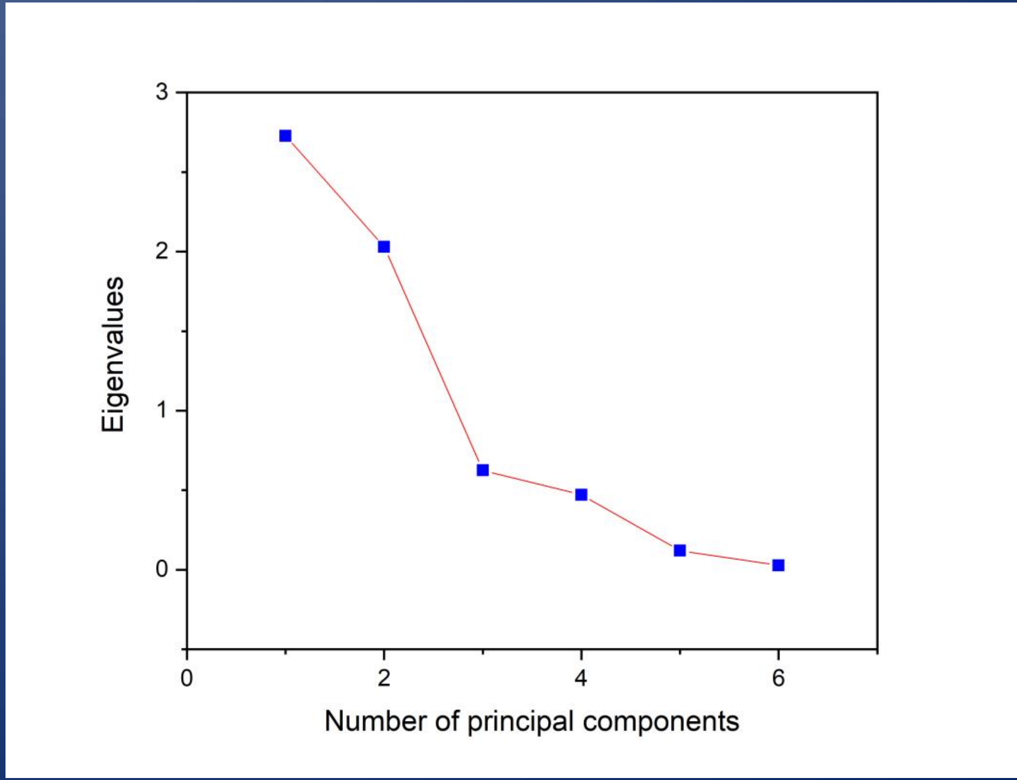
PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Eigenvalues

λ_i	Cum. Sum [%]
2.73	45.46
2.03	79.28
0.62	89.62
0.47	97.55
0.12	99.54
0.03	100

Scree graph (PXD)



JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

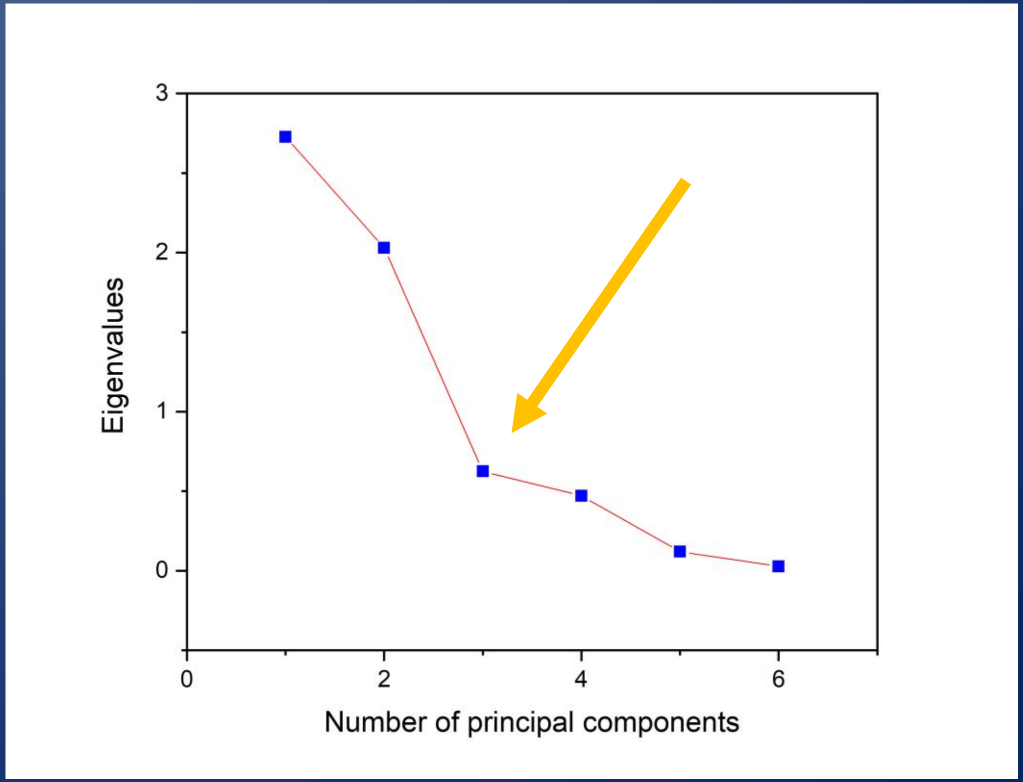
PROJECT 1

MULTIPARAMETER ANALYSIS OF ANTIDEUTERONS

Eigenvalues

λ_i	Cum. Sum [%]
2.73	45.46
2.03	79.28
0.62	89.62
0.47	97.55
0.12	99.54
0.03	100

Scree graph



PRINCIPAL COMPONENTS ANALYSIS
IMPORTANT REMARKS

Pre-processing is a must:

- Normalisation is required
- PCA presumes linear correlations
- if non-linear:
  use *log* or *actan*

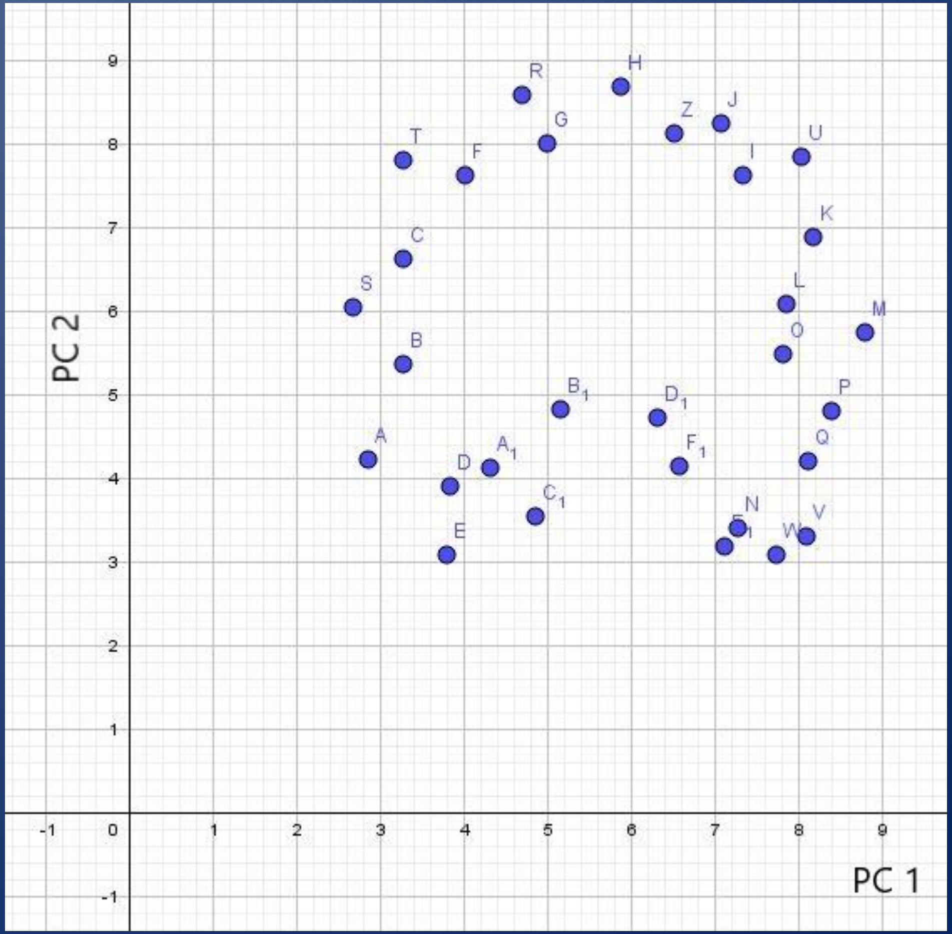
PCA is sensitive to outliers!

- Plot data set if possible

PRINCIPAL COMPONENTS ANALYSIS IMPORTANT REMARKS

<u>Horseshoe effect</u>

Example of „failure“ of PCA



PROJECT 3

CLUSTER SHAPE ANALYSIS