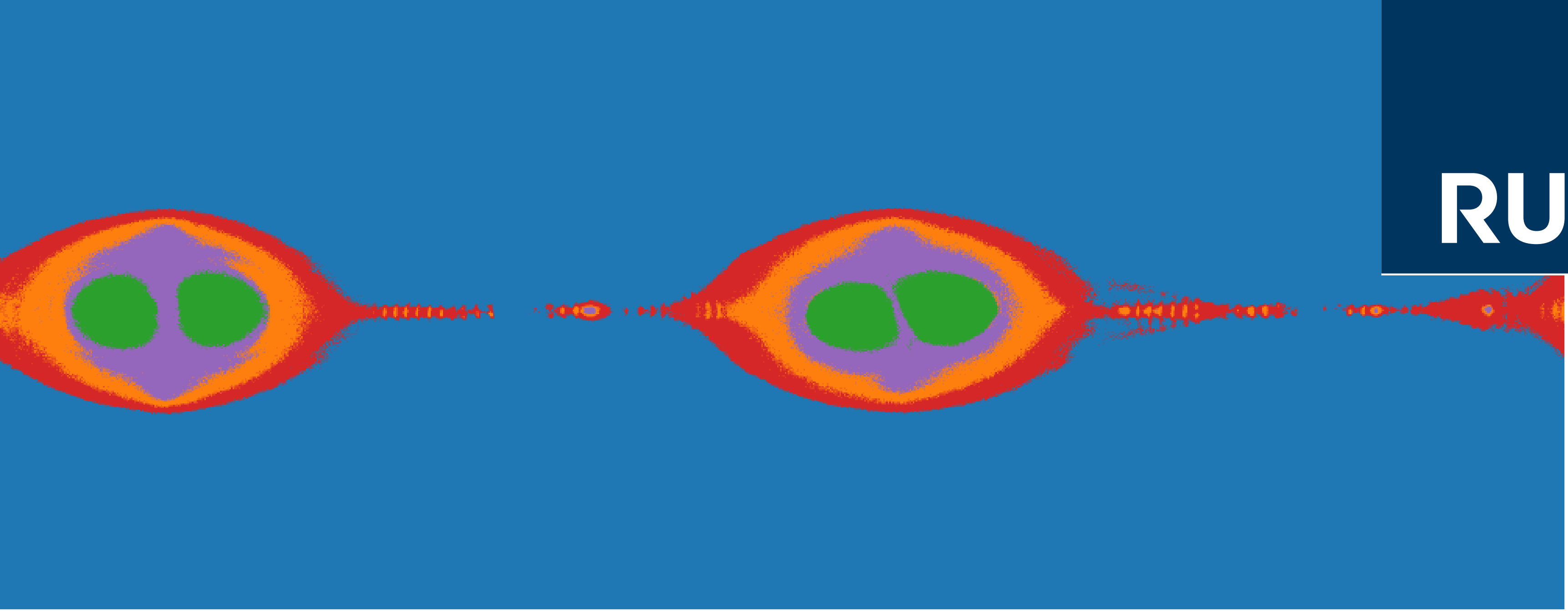


RUHR-UNIVERSITÄT BOCHUM

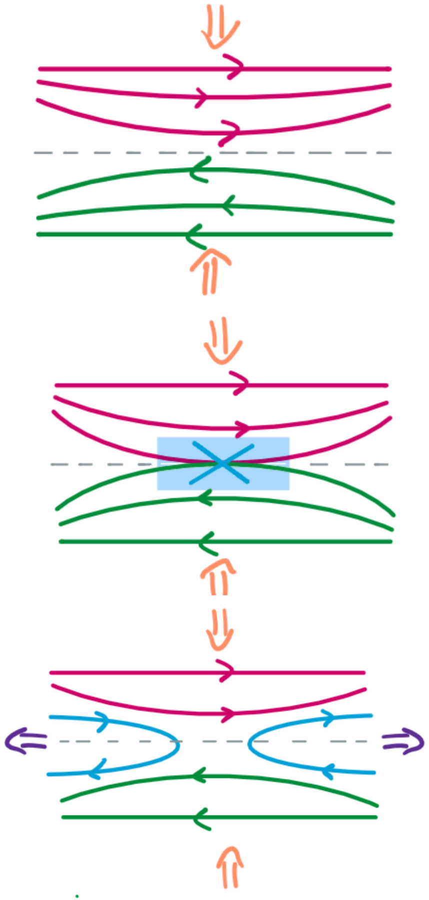
STREAMLINING THE ANALYSIS OF MAGNETIC RECONNECTION SIMULATIONS USING MACHINE LEARNING METHODS

Sophia Köhne – sophia.koehne@rub.de

General Assembly SFB 1491 – Theoretische Physik 1 (Jun.-Prof. Dr. Maria Elena Innocenti)

Motivation | Magnetic Reconnection

- How are particles **preaccelerated** to reach high energies observed in **astrophysical contexts**? – possible answer: magnetic reconnection
- Magnetic field lines ‚snap open‘ and reconnect in new topology
- **magnetic energy is converted** to heat and non-thermal acceleration



Fundamental for understanding:

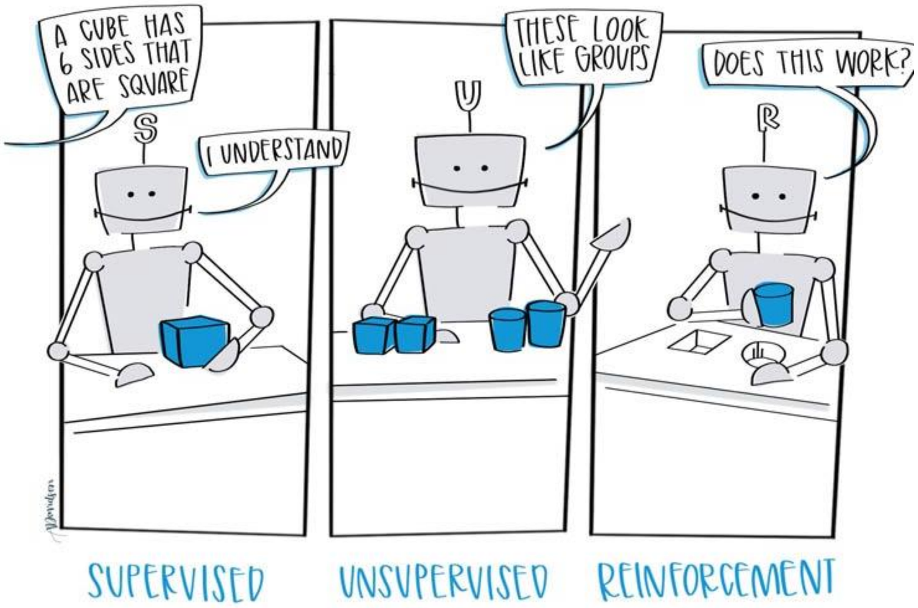
- analyse and **classify** data from spacecrafts and simulations

Motivation | Machine Learning as Solution

Goal:
- Streamline identification of physically distinct regions of magnetic reconnection

Problems:
- huge amounts of data
- unconscious bias

Solution (?):
- Unsupervised machine learning



https://www.ceralytics.com/3-types-of-machine-learning/

Background

Clustering and Self Organizing Maps (SOMs)

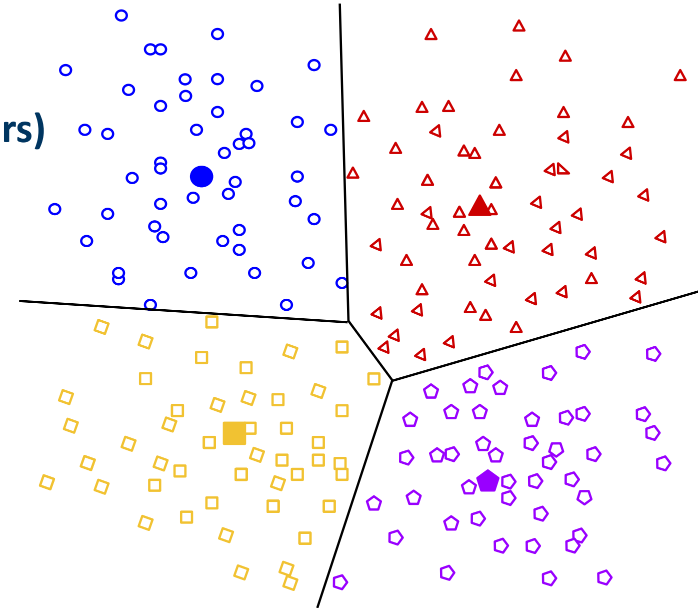
Background | Unsupervised learning methods

Clustering:

- **Process of segmenting unlabeled data into subgroups (clusters) of similar input** (Jo, 2021)
- Centroid clustering: Finds the prototypes of clusters
- **K-means** is the most widely used algorithm for centroid clustering

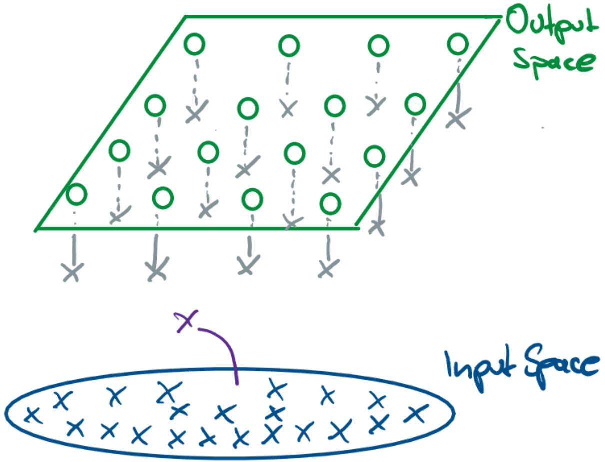
Self-organizing maps (Kohonen, 1982):

- Form two-dimensional maps of high-dimensional data
- **Preserve the topology** of input data
- Can be used for clustering but also offer powerful visualizations



Background | SOM training

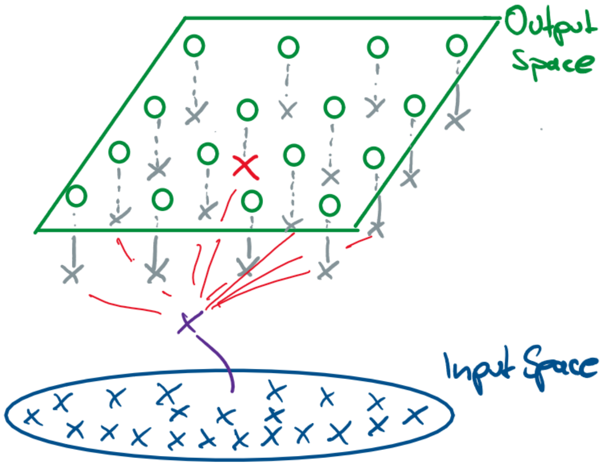
Goal: Find weight values, such that adjacent units have similar values



Background | SOM training

Goal: Find weight values, such that adjacent units have similar values

- **Inputs are assigned to unit (-weights) that are most similar to them – Best Matching Units (BMU)**
- The weights of the most similar unit and its neighbours get activated for weight update
- The activated units‘s weights get updated



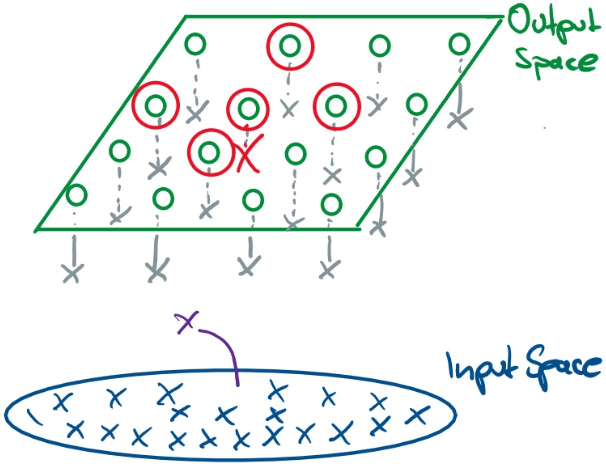
The SOM training is seperable in three phases (van Hulle, 2012):

Competition **Collaboration** **Weight Updates**

Background | SOM training

Goal: Find weight values, such that adjacent units have similar values

- Inputs are assigned to unit (-weights) that are most similar to them
- **The weights of the most similar unit and its neighbours get activated for weight update**
- The activated units‘s weights get updated



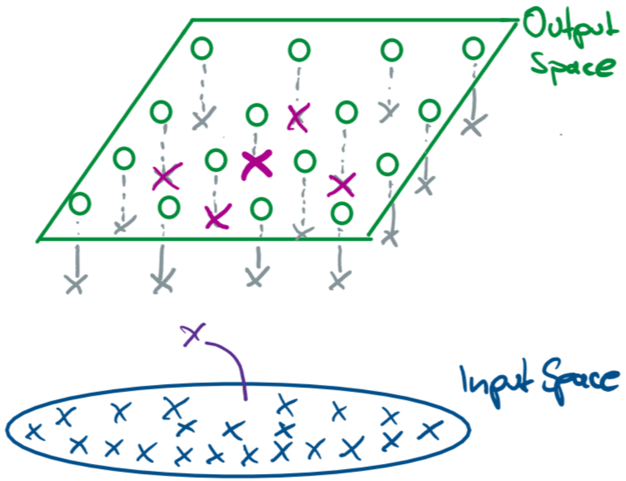
The SOM training is seperable in three phases (van Hulle, 2012):

Competition **Collaboration** Weight Updates

Background | SOM training

Goal: Find weight values, such that adjacent units have similar values

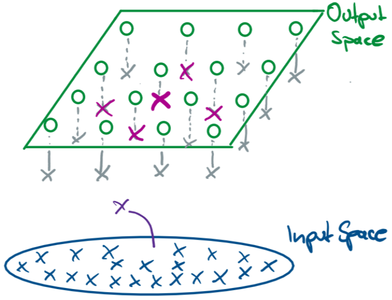
- Inputs are assigned to unit (-weights) that are most similar to them
- The weights of the most similar unit and its neighbours get activated for weight update
- **The activated units‘ weights get updated**

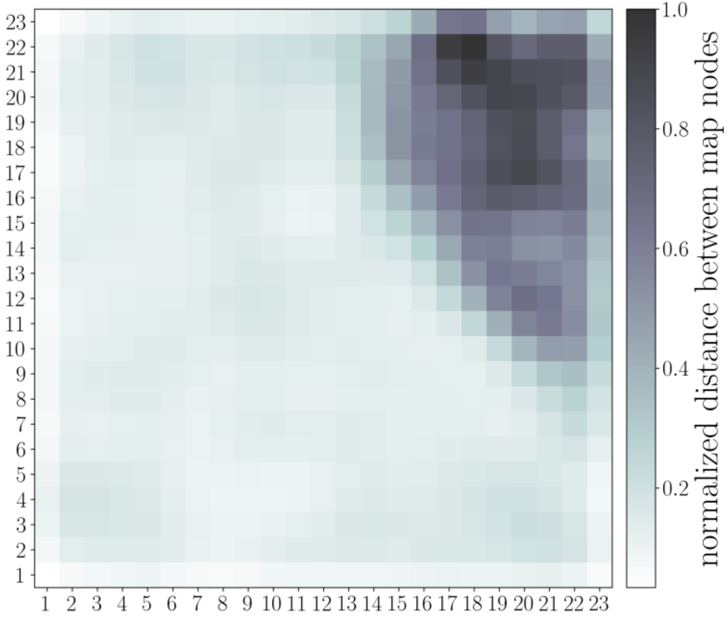


The SOM training is separable in three phases (van Hulle, 2012):

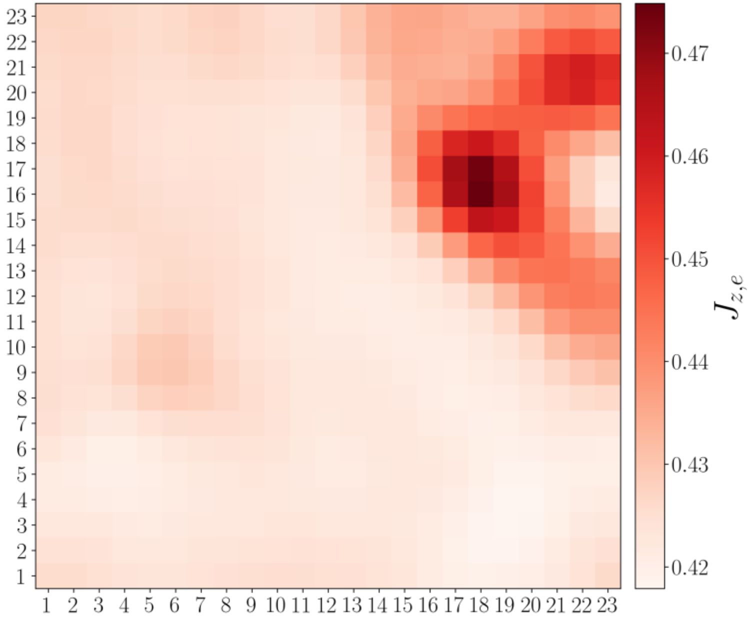
Competition | Collaboration | **Weight Updates**

Background | Visualizations of the SOM





Unified Distance Matrix
(UDM)



Feature Map

RUHR UNIVERSITÄT BOCHUM

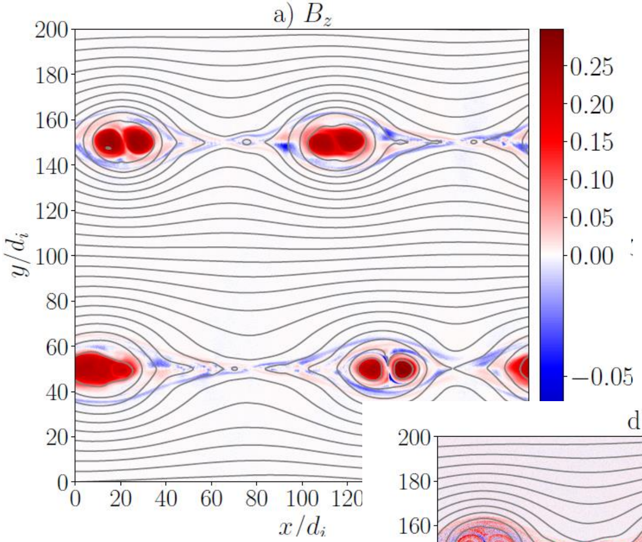
Data and Methodology

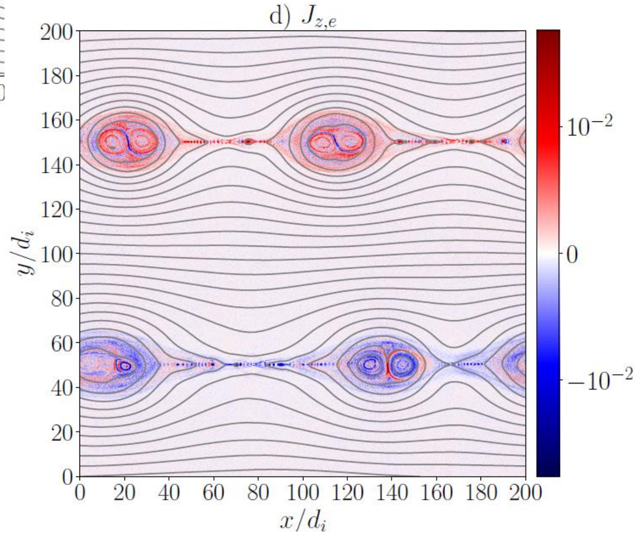
Data |Simulation

Data points produced by fully kinetic simulation of **plasmoid instability**

- semi-implicit, energy conserving PIC code ECsim (Lapenta et al., 2017)
- Force free initial conditions, periodic boundary conditions, collisionless regime
- Reduced mass ratio $m_r = 25$

→ **4528384 samples with 26 features each**





(Köhne et al., 2023, doi:10.1017/S0022377823000454)

Method | How results were obtained

1. Preprocessing
2. Initialise SOM
3. SOM training on preprocessed data
4. k-Means clustering of SOM
5. Visualize results

RUHR UNIVERSITÄT BOCHUM RUB

Method

1. **Preprocessing**
2. Initialise SOM
3. SOM training on preprocessed data
4. k-Means clustering of SOM
5. Visualize results

Scaling the data

- Scale feature values according to defined rule
- Assures that all features have same level of influence on model
- Most common: scale to interval, e.g. [0,1]

Method

1. Preprocessing
2. **Initialise SOM**
3. **SOM training on preprocessed data**
4. k-Means clustering of SOM
5. Visualize results

SOM implementation used:
https://github.com/JustGlowing/minisom
(serial implementation)
https://github.com/mistrello96/CUDA-SOM
(parallel implementation)

(Hyper-) Parameters (Kohonen, 2014):

- Number of neurons $n \approx 5\sqrt{N}$ (J. Tian et al., 2014)
- Aspect ratio of x,y dimensions is the same as the ratio of the two first principal components (J. Tian et al., 2014)
- Initial neighborhood radius $\sigma_0 = 0.2 * \max\{x, y\}$
- Initial learning rate $\eta_0 = 0.5$
- # of iterations: $n_{iter} = 5N$
- Neighborhoodfunction: gaussian
- Neighborhood distance function: euclidean
- Decay function: asymptotic
- Weights initialization: random samples

RUHR UNIVERSITÄT BOCHUM

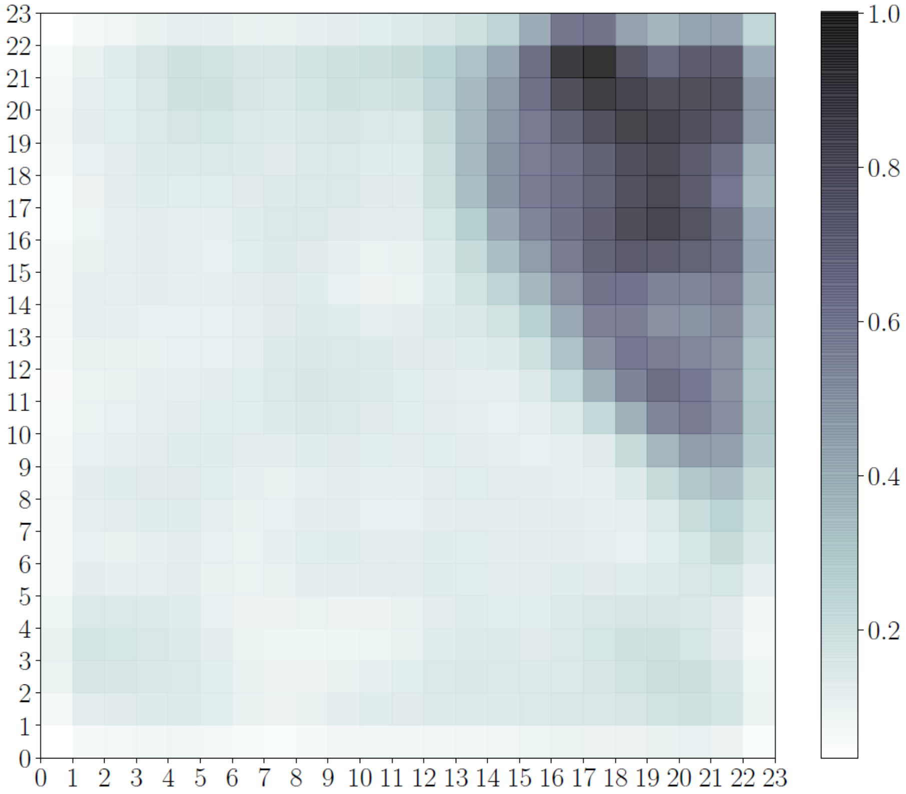
Method

1. Preprocessing
2. Initialise SOM
3. SOM training on preprocessed data
4. **k-Means clustering of SOM**
5. Visualize results

- $k \in [2,8]$
- Optimal cluster number was determined using Satopaa kneedle method (Satopaa et al., 2011)

Method

1. Preprocessing
2. Initialise SOM
3. SOM training on preprocessed data
4. k-Means clustering of SOM
5. **Visualize results**

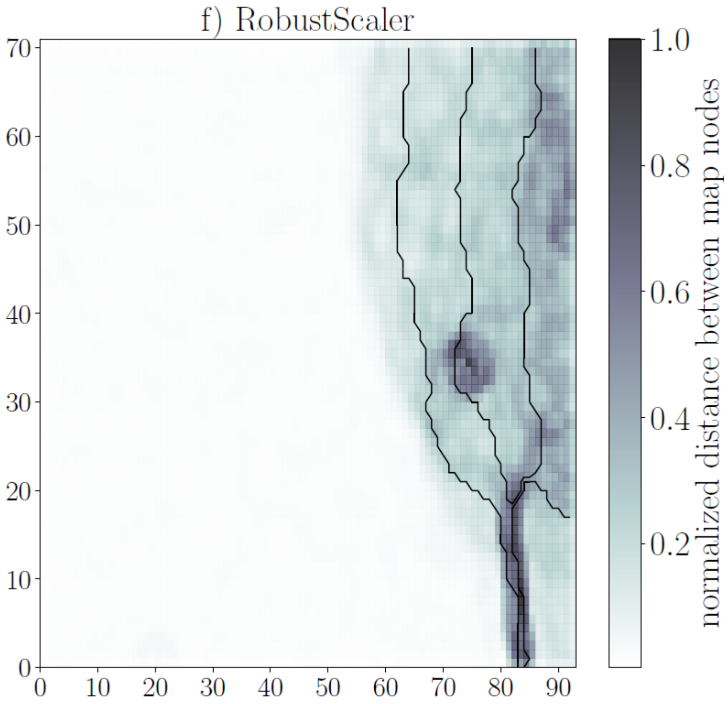


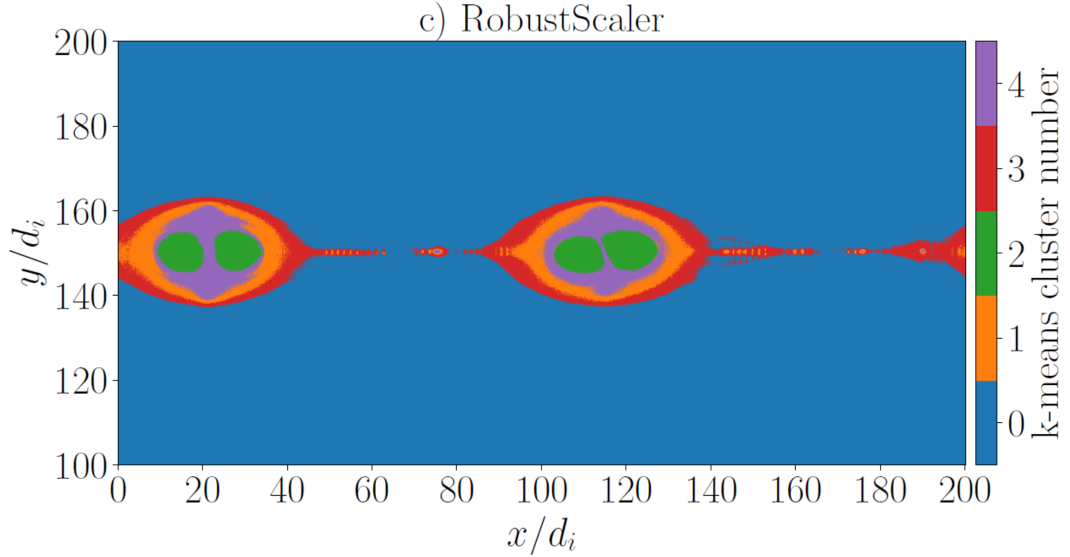
Unified Distance Matrix (UDM)

Clustering results

(Köhne et al., 2023, doi:10.1017/S0022377823000454)

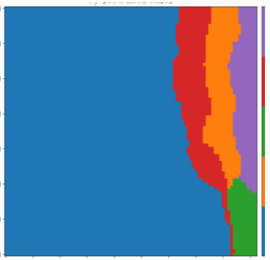
Results | Classification

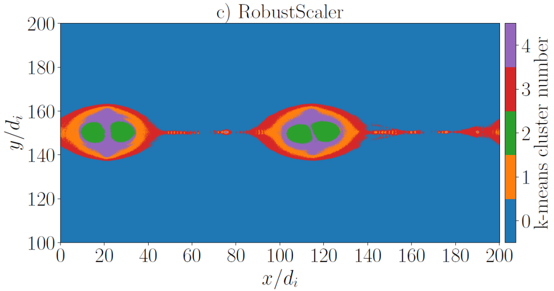




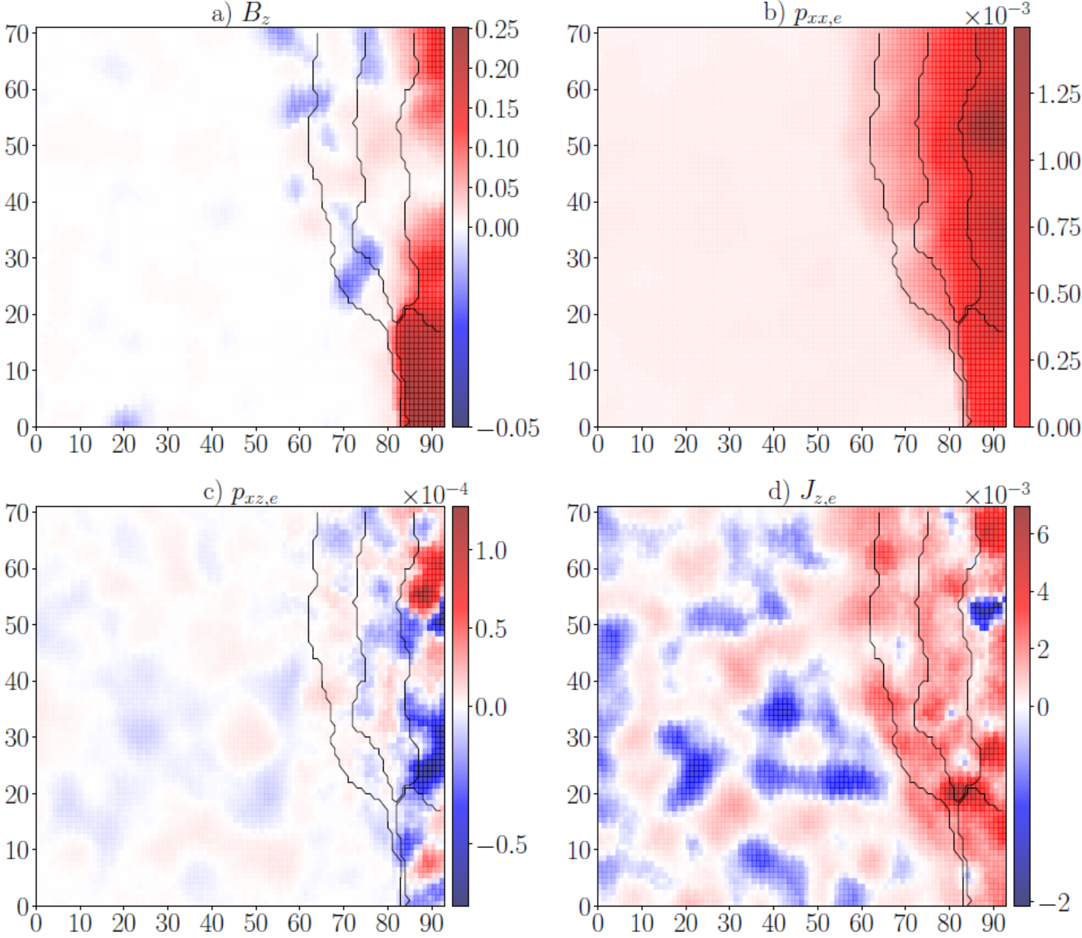
Results | Feature Maps

- show values of selected features in the weights of each node
- possibility to investigate peculiarities





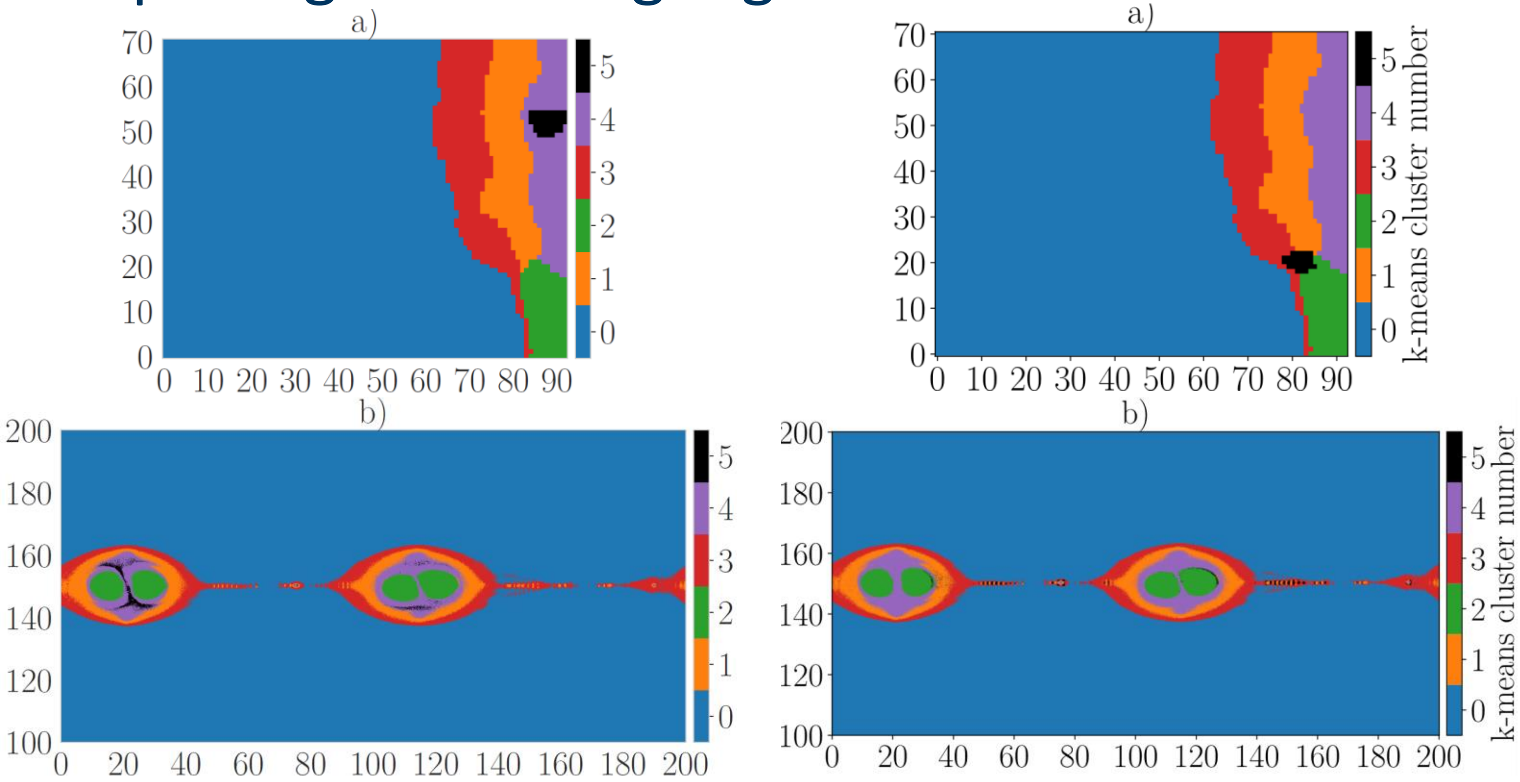
(Köhne et al., 2023, doi:10.1017/S0022377823000454)



(Köhne et al., 2023, doi:10.1017/S0022377823000454)

Results | Handpicking interesting regions

- extreme $J_{z,e}$ inside of purple cluster – **plasmoid merging** regions
- High-value $J_{z,e}$ at intersection of plasmoid clusters – **mini-plasmoids** at x-line

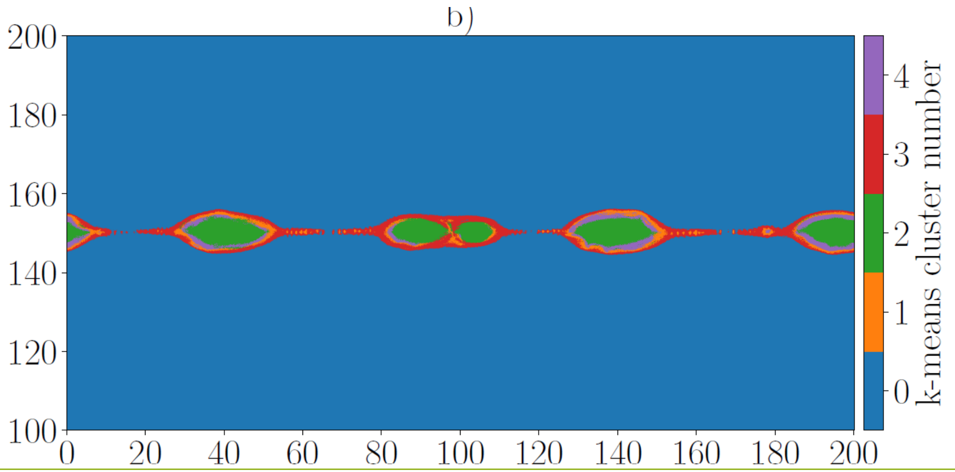


RUHR UNIVERSITÄT BOCHUM RUB

Results | Robustness of the results

- **Hyperparameter variation:** number of epochs, initial neighborhood radius, initial learning rate, initialization seed, number of nodes
  - Plasmoid classification stays very stable: run with largest deviation matches reference run to 86 %
  - most unstable clusters: intermediate plasmoid region (orange and purple)

- **Temporal variation:** classify data from earlier timestep of simulation using SOM trained on later timestep
  - General classification stays sensible
  - Plasmoid merging regions shrink



Summary and Outlook

Summary & Outlook

- To understand magnetic reconnection analysis of huge amounts of data is needed
- **Combination of unsupervised machine learning methods**: k-means and SOMs
  - Can identify **physically distinct regions** in fully kinetic simulations
  - SOMs offer many **intuitive investigation** possibilities

Further steps:

- Analysis and classification of **observational data** (Amaya et al., 2020)
- Usage in simulations to e.g. switch from one numerical method to another according to cluster

J. Plasma Phys. (2023), *vol.* 89, 895890301
Published by Cambridge University Press

doi:10.1017/S0022377823000454

Unsupervised classification of fully kinetic simulations of plasmoid instability using self-organizing maps (SOMs)

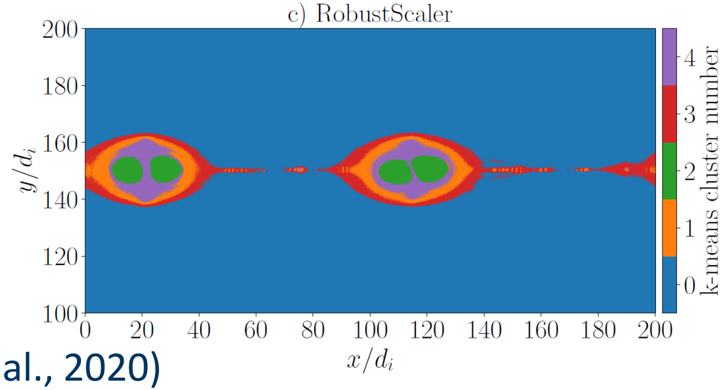
Sophia Köhne[1], **Elisabetta Boella**[2,3] **and Maria Elena Innocenti**[1,†]

[1]Institut für Theoretische Physik, Ruhr-Universität Bochum, Universitätstraße 150, 44801 Bochum, Germany

[2]Physics Department, Lancaster University, Bailrigg, Lancaster LA11NN, UK

[3]Cockcroft Institute, Sci-Tech Daresbury, Warrington WA44AD, UK

(Received 12 February 2023; revised 2 May 2023; accepted 3 May 2023)



RUHR UNIVERSITÄT BOCHUM RUB

Additional Slides

Clustering | K-means

Input:
number of clusters K
dataset x_i, $i \in (1,\ldots,n)$

Output:
Set of K clusters

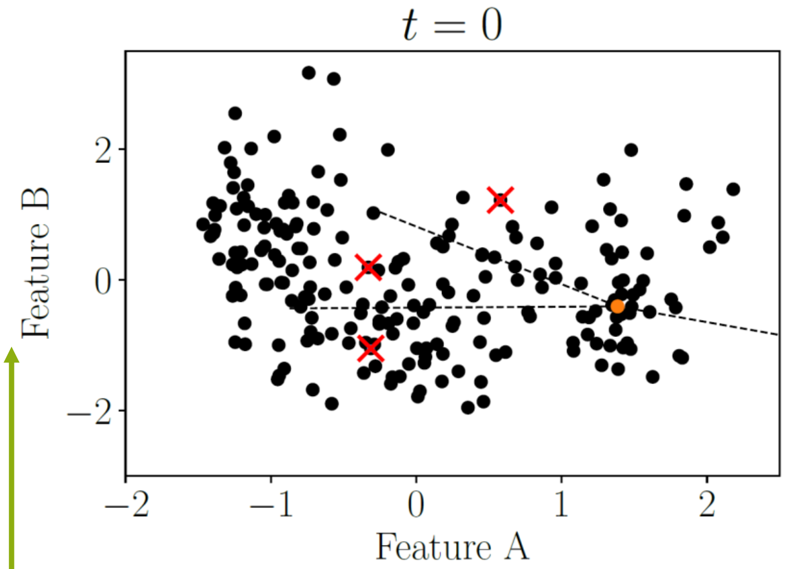
Algorithm:

1. Place k initial centroids randomly
2. Assign each data point to its closest centroid, minimizing:

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} ||x_i - c_j||^2$$

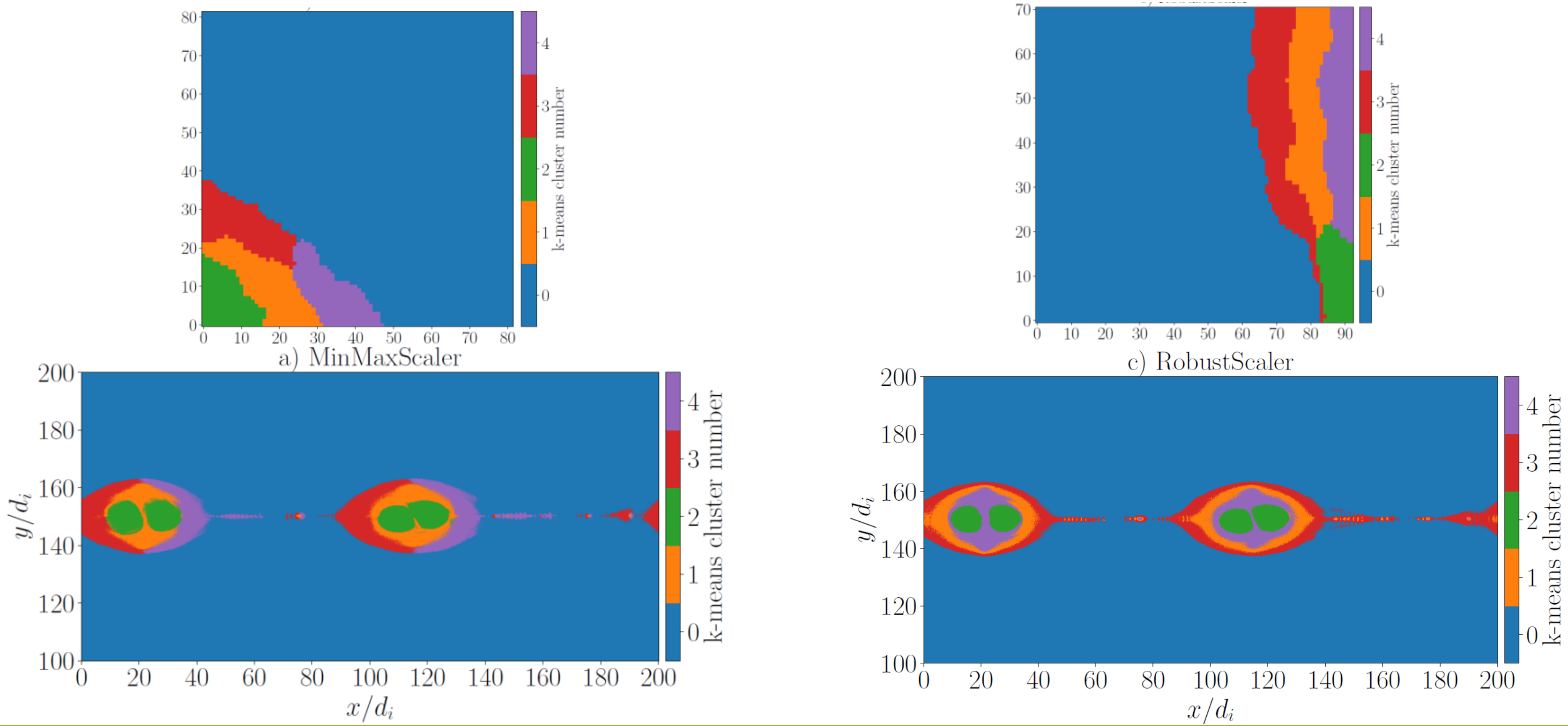
3. Update centroids - move them to mean of all xi assigned to them

Break if convergence criterion is matched



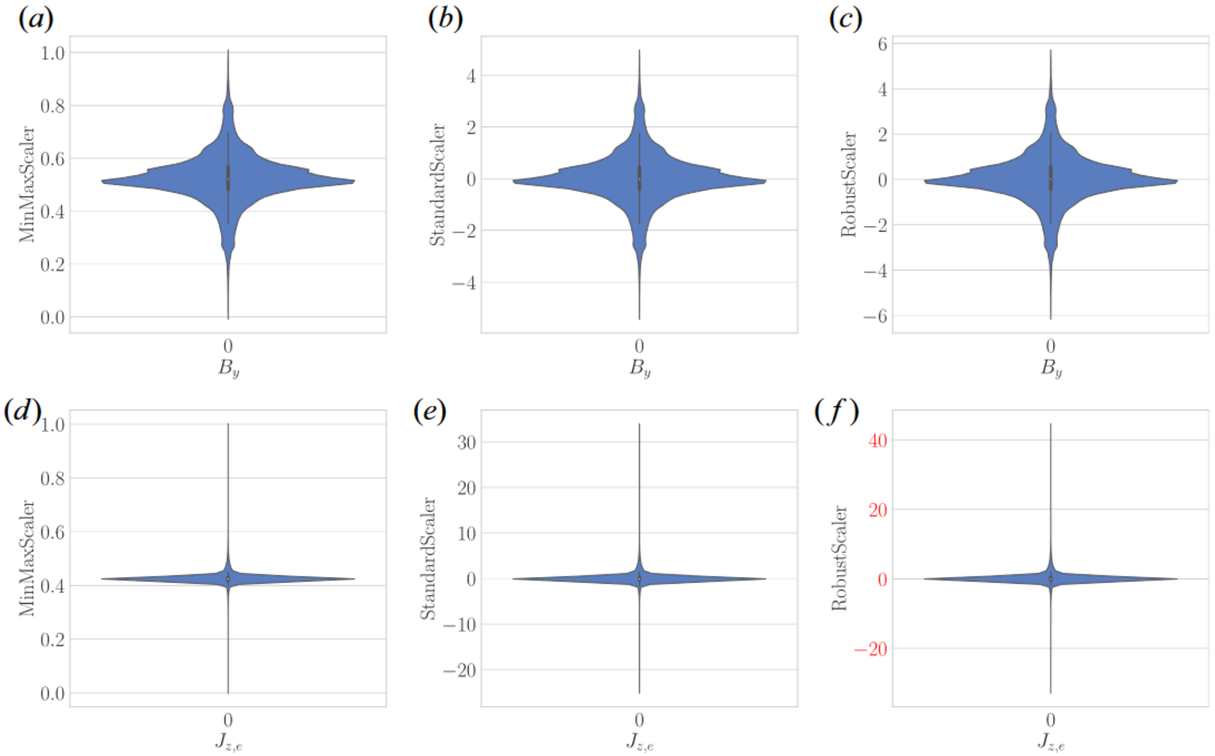
(Köhne et al., 2023,
doi:10.1017/S0022377823000454)

Results | Importance of Scalers



(Köhne et al., 2023,
doi:10.1017/S0022377823000454)

Results | Importance of Scalers



SOM | Convergence

Two convergence phases: Topographic ordering of weights, Convergence of those weights

- Both are strictly speaking not mathematically proven (Cotrell, Fort, et al., 1998)
- SOMs do not have an objective function *J* where $\Delta w_i \propto \frac{\partial J}{\partial w_i}$

Practical convergence criteria:

- Maximum number of iterations
- Stop when Δw_i is under threshold
- Quantization error: average distance between each input sample and its best matching unit → should be as low as possible
- Topographic error: percentage of samples whose second-best matching unit is not adjacent to their BMUs