http://www.aida-space.eu

A VERSATILE TECHNIQUE FOR UNSUPERVISED CLASSIFICATION AND PRELIMINARY ANALYSIS OF SIMULATION RESULTS

MARIA ELENA INNOCENTI[1]

Sophia Köhne[1], Simon Hornisch[1], Rainer Grauer[1], Jorge Amaya[2], Joachim Raeder[3], Romain Dupuis[2], Banafsheh Ferdousi[2], Giovanni Lapenta[2]

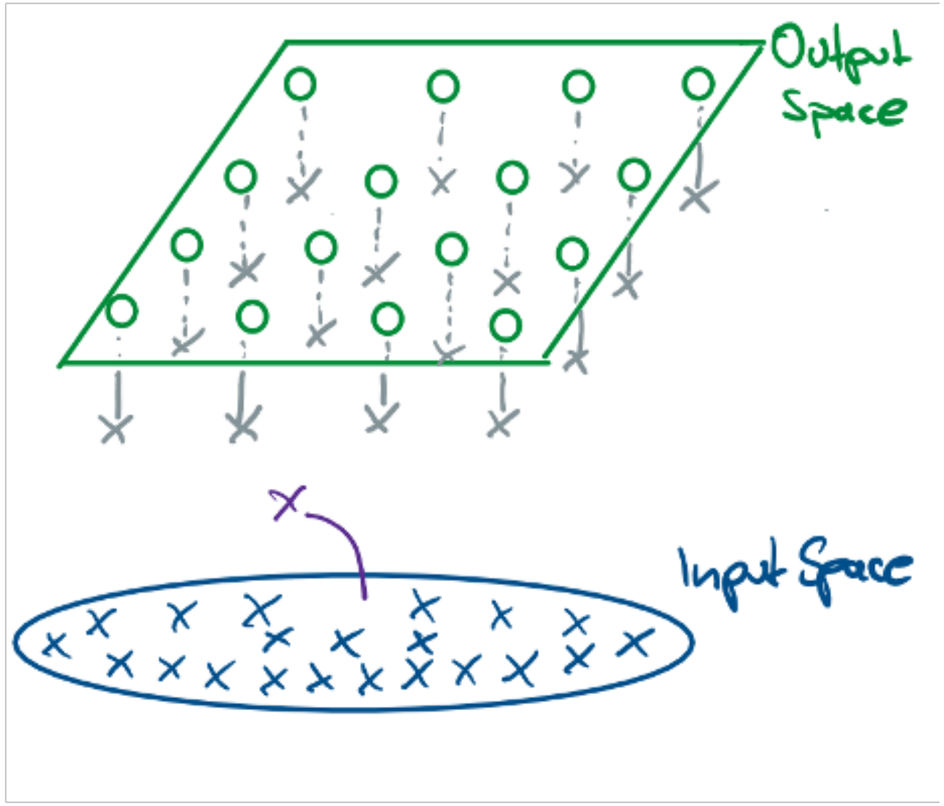
[1] **Ruhr-Universität Bochum, Germany**
[2] KULeuven, Belgium
[3] University of New Hampshire, USA

OUTLINE

- **Self-Organizing Maps:** a recap
- **To begin**: classification of global MagnetoHydroDynamics simulations of the terrestrial magnetosphere
- **Let's add kinetic scale physics**: classification of Particle-In-Cell simulations of plasmoid instability
- **Let's move across models**: training the map on a Vlasov-Maxwell simulation, classifying a multi-fluid one
- Discussion and conclusions

1. SELF-ORGANISING (KOHONEN) MAPS

The idea: represent a large set of high-dimensional data as a 2D *ordered* lattice



Drawings: S. Köhne

2. **SOM**: a 2D map of nodes/ neurons/ units, each characterized by a weight vector $\mathbf{w} \in R^n$; each node is a *model,* a local average of the data

The map is **ordered**: nearby models are “similar”, far away models are “different”

$1 \rightarrow 2$ **Training**: from (randomly initialised) weights, to models that represent the input data well; the training is **unsupervised**

1. **Input data**: large set of high- dimensional data
n: dimensionality of the data/ number of features associated with each input point

Kohonen, 1982

A VERY-SHORT INTRODUCTION TO SOM TRAINING

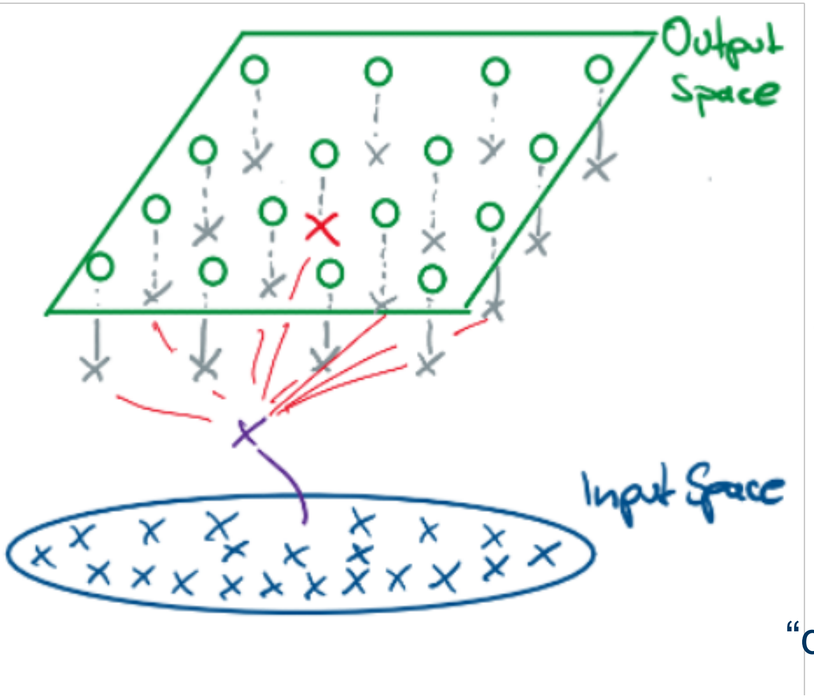
The goal: modify the node weights, so that the input data are well represented by the map, and adjacent units have similar values/ map to similar data points

Training process:

1. **Input data are presented to the map (multiple times): the "Best Matching Unit" of each input is identified**
2. The neighbors of the BMU are identified
3. The weights of the BMU and of the nearest neighbors are updated, so that they are more similar to the input data

Drawings: S. Köhne

Competition



Best Matching Unit:

$$\mathbf{w}_s = \underset{\mathbf{w}_i \in \mathbf{W}}{\arg\min} \left(\|\mathbf{x}_\tau - \mathbf{w}_i\| \right)$$

"distance" has to be defined

A VERY-SHORT INTRODUCTION TO SOM TRAINING

The goal: modify the node weights, so that the input data are well represented by the map, and adjacent units have similar values/ map to similar models

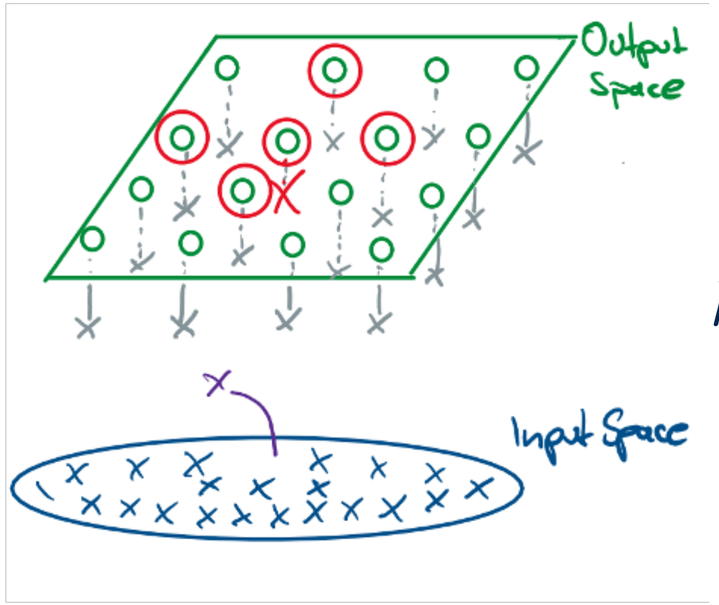
Collaboration

Training process:

1. Input data are presented to the map (multiple times): the “Best Matching Unit” of each input is identified

2. The neighbors of the BMU are identified

3. The weights of the BMU and of the nearest neighbors are updated, so that they are more similar to the input data



Neighborhood function

$$h_{ij\tau} = e^{-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{2\sigma(\tau)^2}}$$

Time-dependent lattice neighborhood width

Drawings: S. Köhne

A VERY-SHORT INTRODUCTION TO SOM TRAINING

The goal: modify the node weights, so that the input data are well represented by the map, and adjacent units have similar values/ map to similar models

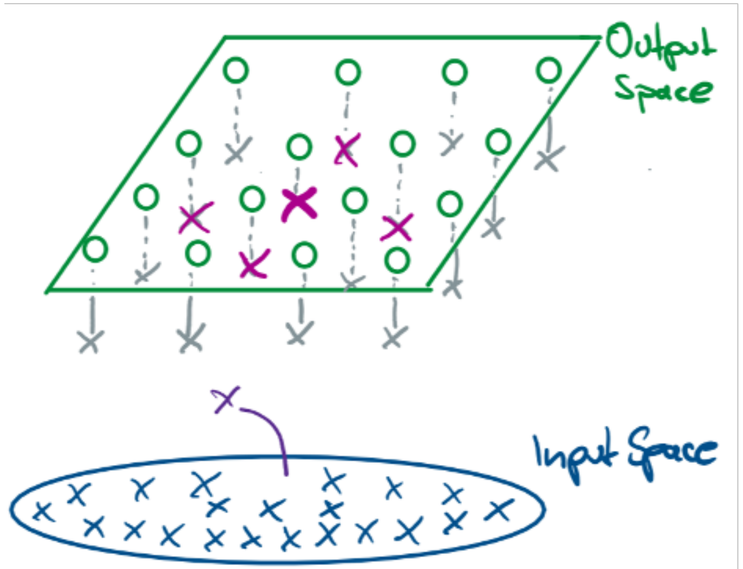
Collaboration

Training process:

1. Input data are presented to the map (multiple times): the “Best Matching Unit” of each input is identified

2. The neighbors of the BMU are identified

3. **The weights of the BMU and of the nearest neighbors are updated, so that they are more similar to the input data**



$$\mathbf{w}_j(\tau) = \mathbf{w}_j(\tau - 1) + \Delta\mathbf{w}_j$$

$$\Delta\mathbf{w}_j(\tau) = \eta(\tau) h_{sj\tau}(\mathbf{x}_\tau - \mathbf{w}_j)$$

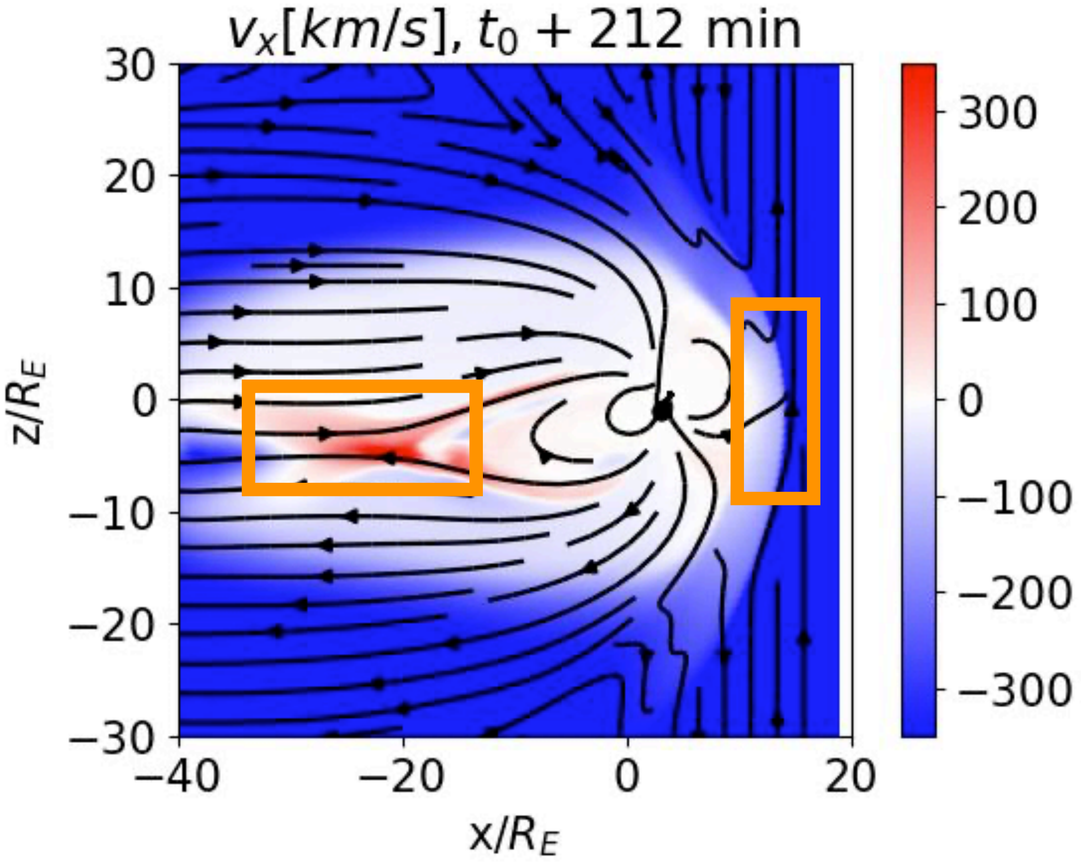
Time-dependent learning rate

The closer to the input, the largest the update

Drawings: S. Köhne

2. CLASSIFICATION OF GLOBAL MAGNETOSPHERIC SIMULATIONS

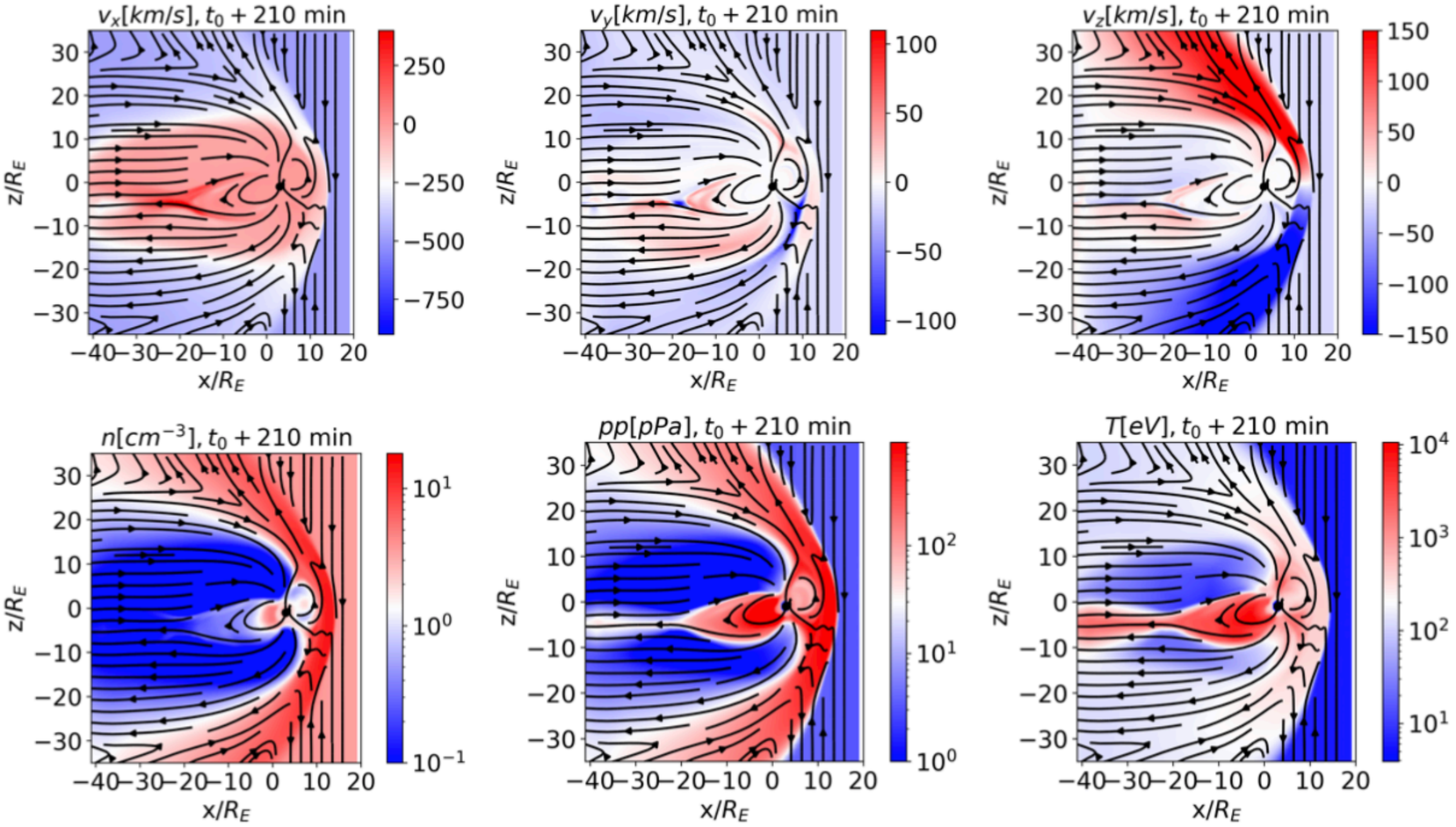
- OpenGGCM-CTIM-RCM (magnetosphere + ionosphere + thermosphere; Reader et al 2003) global magnetospheric simulation
- **MagnetoHydroDynamic** simulation
- Stretched Cartersian grid, with 325 x 150 x150 cells
- [- 3000 R_E - 18 R_E] x [-36 R_E - 36 R_E] x [-36 R_E - 36 R_E]
- Sunwards boundary conditions from ACE
- t_0= 8 May 2004, 9:00 UTC



GLOBAL MAGNETOSPHERIC SIMULATIONS

From simulation results, different magnetospheric regions can be easily identifie

- Bow shock
- Magnetosheath
- Magnetopause
- Magnetotail
- Inner magnetosphere
- Plasma sheet



CLASSIFICATION PROCEDURE

- Data preprocessing:
  - Feature scaling (minmax scaler)
  - Dimensionality reduction via Principal Component Analysis, PCA: from 9 to 3 features* (>93% of the original variance)
  - Feature scaling
- SOM training
  - Selection of SOM hyper-parameters (number of nodes, initial lattice neighbor width σ_0, initial learning rate η_0)
  - **Analysis of SOM feature map**
- **K-means classification of SOM nodes**
  - Identification of large scale magnetospheric regions

SOM hyperparameters:

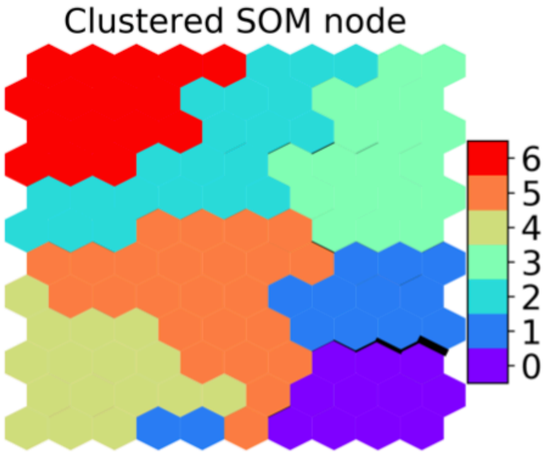
10x12 nodes, η_0= 0.25, σ_0=1, 3+ epoch

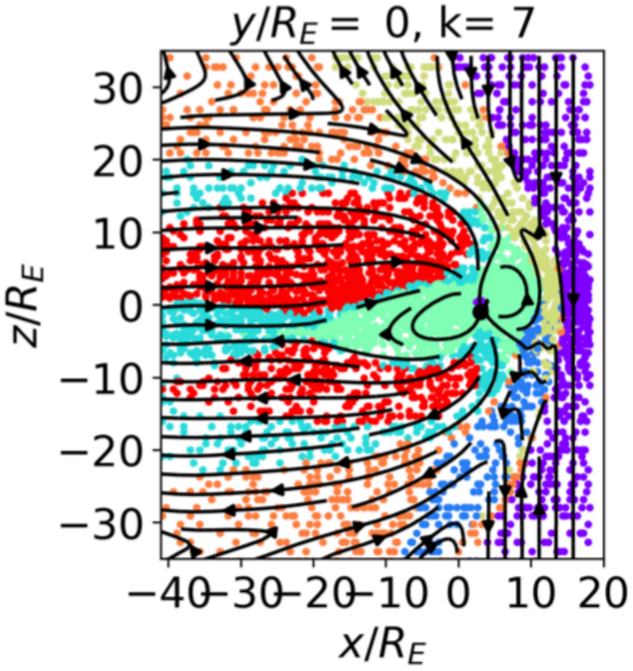
N= 1% of 5.5 10^6 data points

Features: B_x, B_y, B_z, v_x, v_y, v_z, log(n), log(pp), log(T)

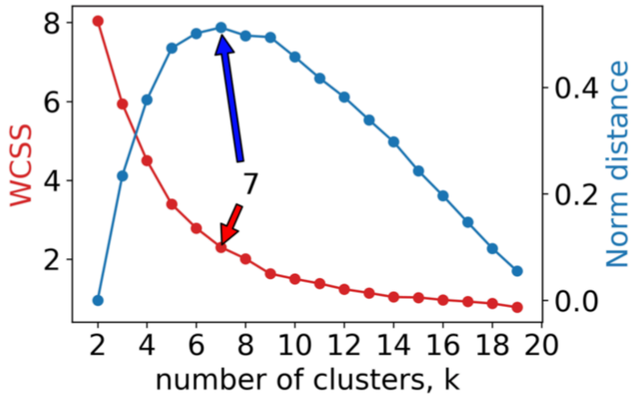
K-MEANS CLASSIFICATION OF SOM NODES

- Each SOM node is a “model”, representing a certain number of input points
- We use K-means classification to classify the trained SOM node
- We select the optimal cluster number with the Kneedle method
- A posteriori, we identify the regions associated with each cluster





Kneedle determination of the optimal number of K-means clusters (Satopaa et al, 2011)

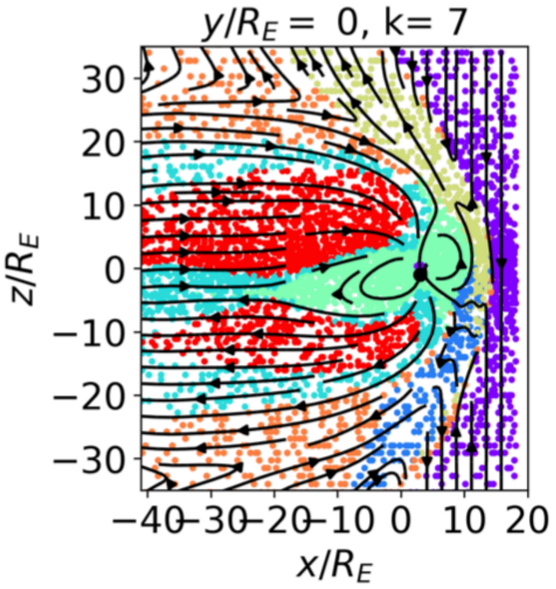


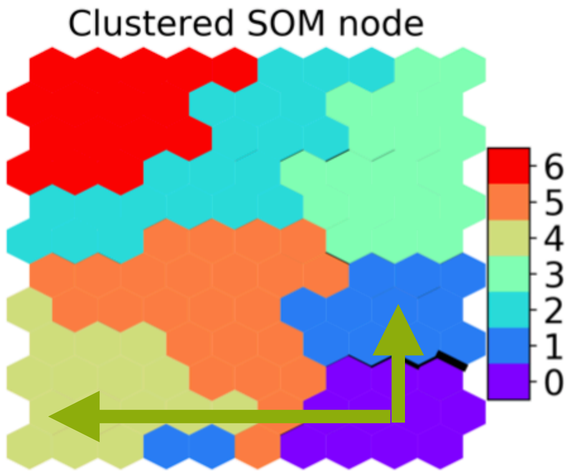
Class	MS region
0	Pristine solar wind
1	Magnetosheath (downstream bowshock)
2	Boundary layer
3	Inner magnetosphere
4	Magnetosheath (downstream bowshock)
5	Magnetosheath
6	Lobes

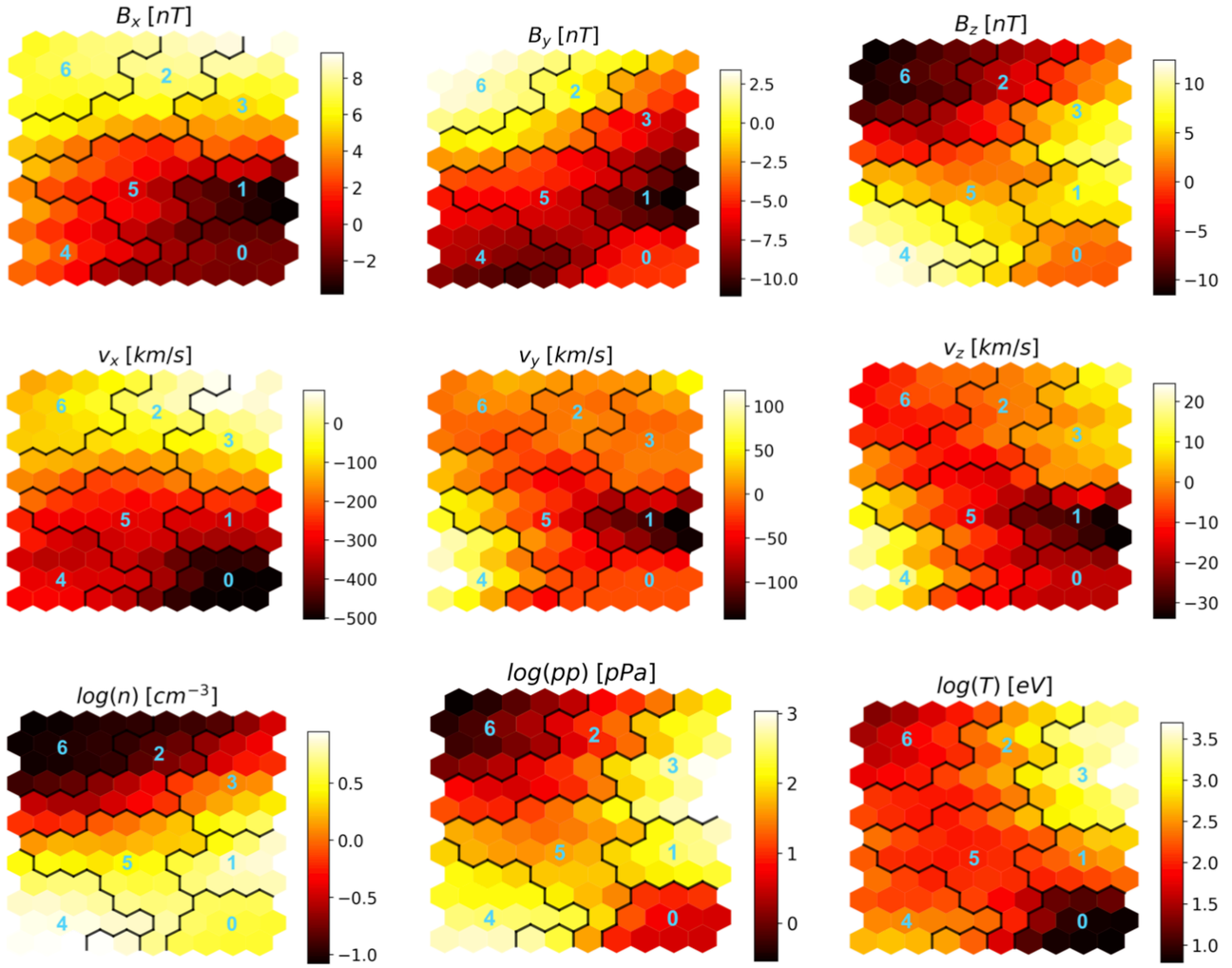
FEATURE MAP ANALYSIS

The map is ordered: we can “walk” through the feature map and observe how the features change across the clusters

Example: crossing the bow shock





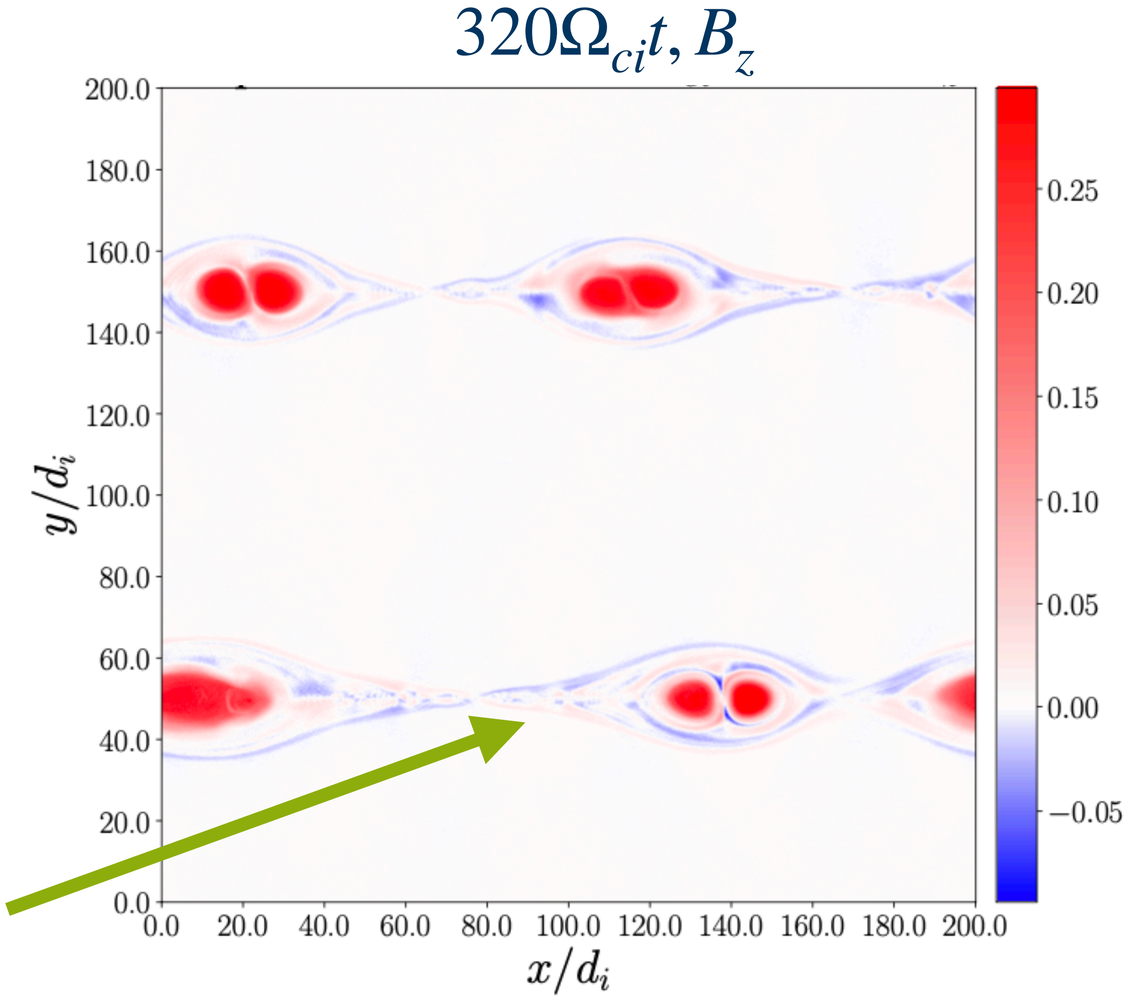


Amaya, J., Dupuis, R., Innocenti, M. E., and Lapenta, G.: Visualizing and Interpreting Unsupervised Solar Wind Classifications, Front. Astron. Space Sci., 7, 66, https://doi.org/10.3389/fspas.2020.553207, 2020.

3. PLASMOID INSTABILITY

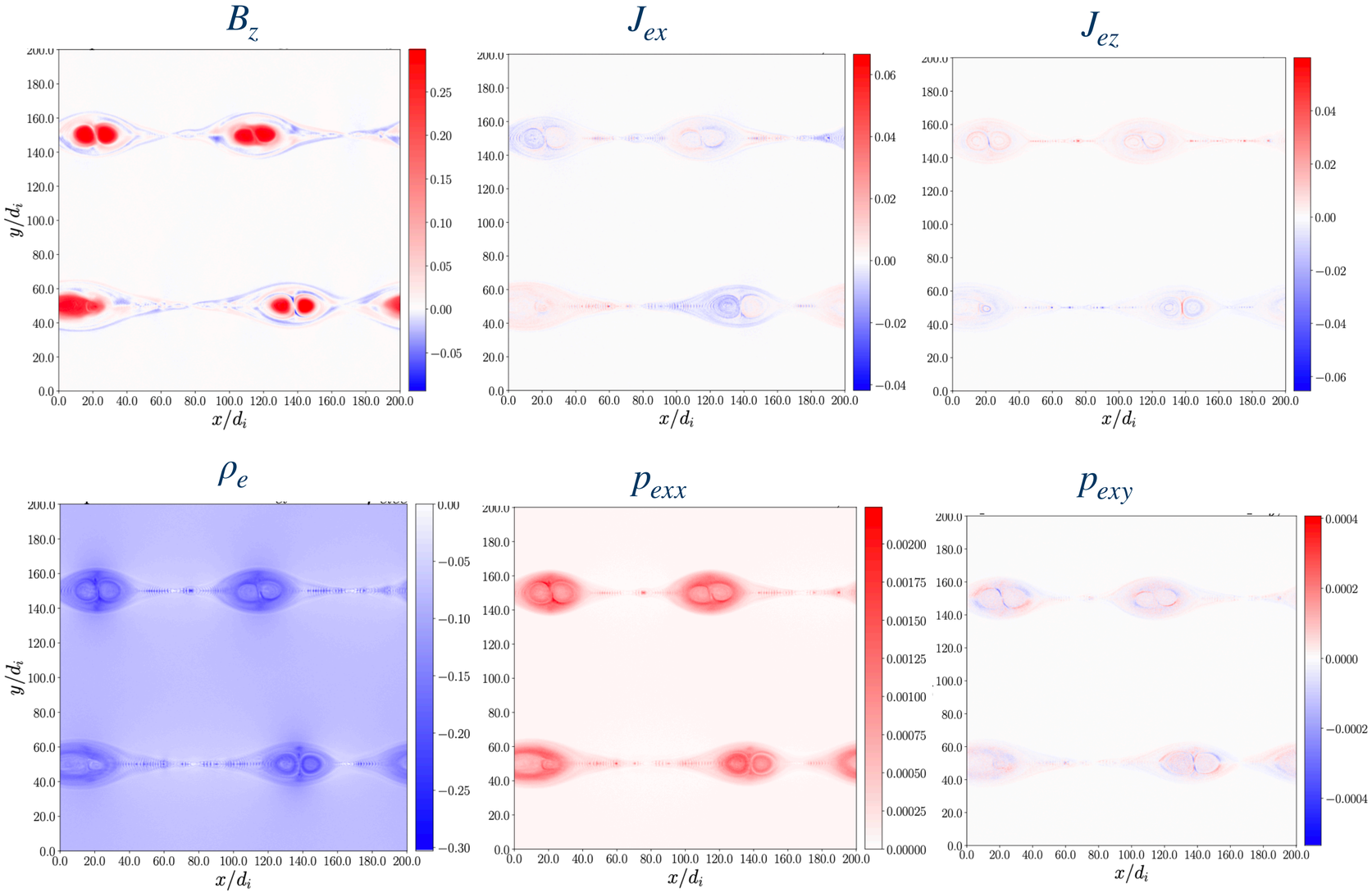
- "Related" to tearing instability
- Formidable link between large and small scales in plasmas
- Converts magnetic energy into kinetic energy and heat, particle heating and acceleration

quadrupole magnetic field signature, signature of collisionless magnetic reconnection



FULLY KINETIC SIMULATIONS OF PLASMOID INSTABILITY

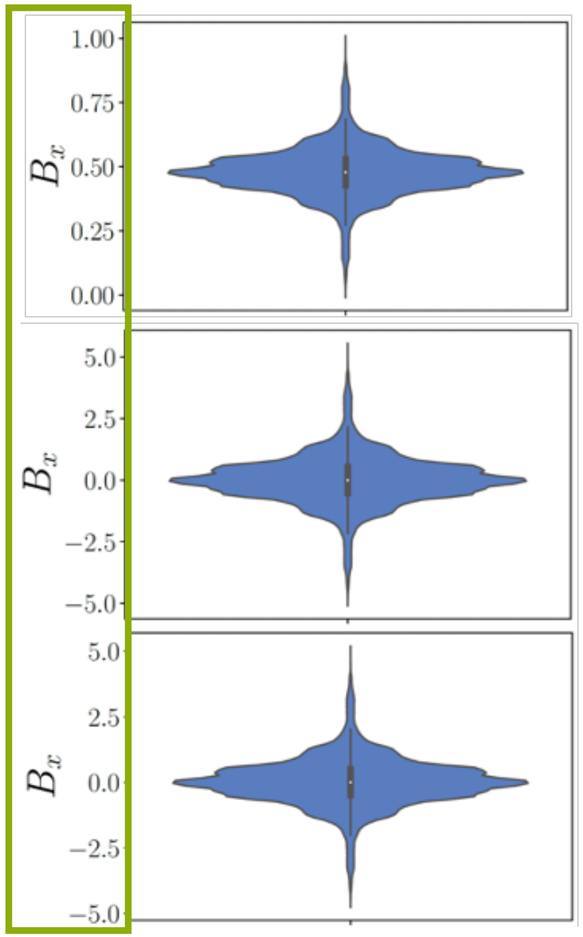
- **Fully kinetic simulations** of plasmoid instability done with the Particle-In-Cell, fully kinetic, semi-implicit, energy conserving code ECsim (Lapenta, 2017)
- 200 x 200 ion skin depths d_i, hundreds of inverse ion gyro frequencies Ω_{ci}, reduced mass ratio mr=25
- Force-free initialization, $dx/d_i = 0.1$, $\Omega_{ci}t = 0.16$
- Periodic boundary conditions, lower current sheet is perturbed for faster instability onset
- **Features**: all info available from a PIC simulations: field + electron and ion moments, up to pressure tensor included



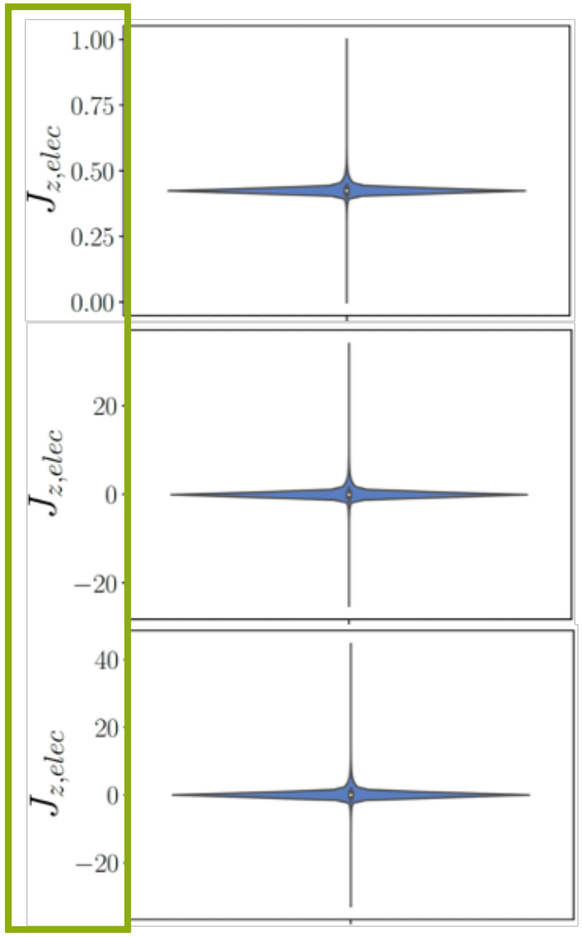
CLASSIFICATION PROCEDURE & THE ROLE OF SCALERS

Same classification procedure as before, but we test the **role of different scalers**

Feature w/o many outliers



Feature with many outliers



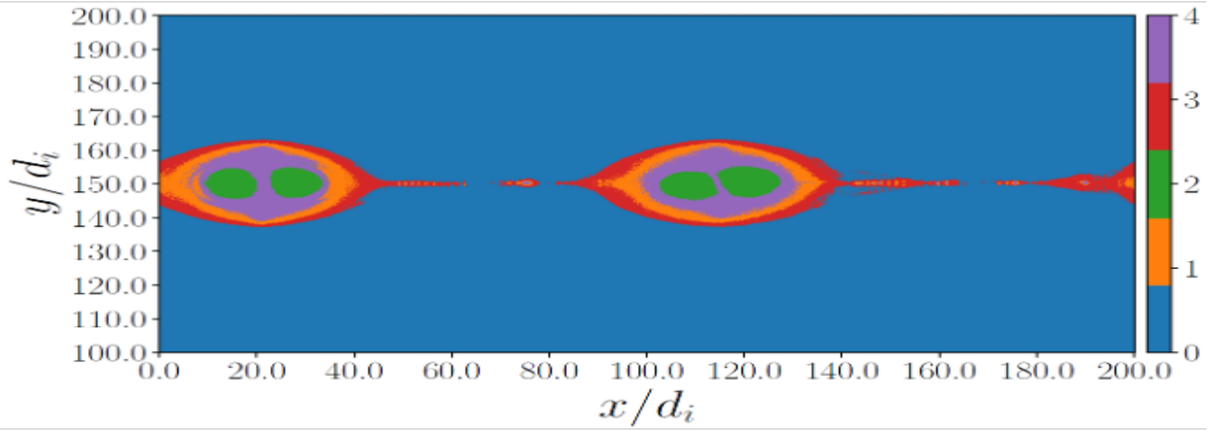
MinMax: rescales so all data are between [0,1]

Standard: removes the mean and scales to unit variance

Robust: removes median and rescales according to percentiles, here 1st to 3rd

→ robust to outliers

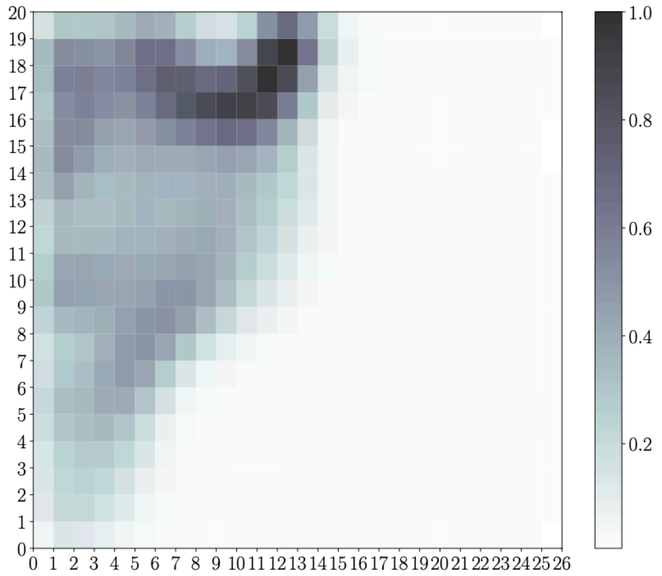
CLASSIFICATION & UNIFIED DISTANCE MATRIX ANALYSIS

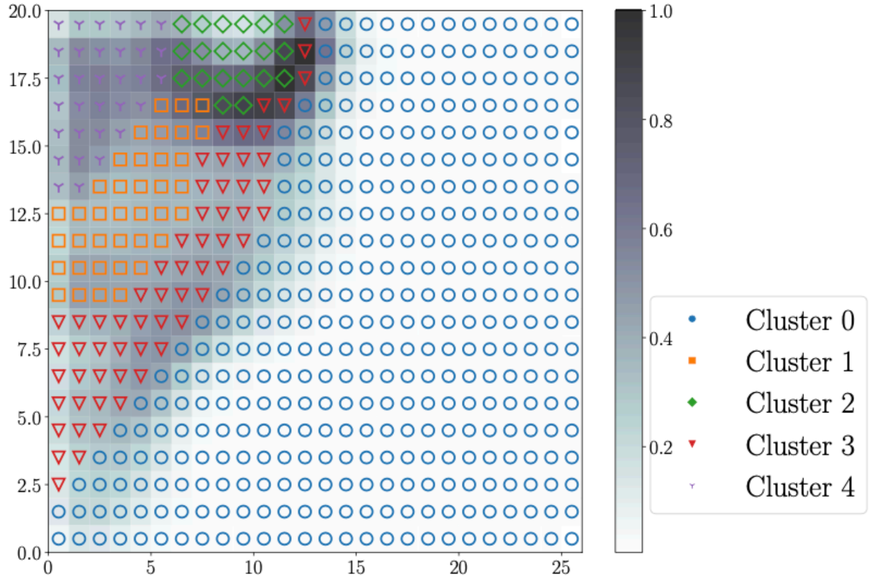


SOM parameters:

number of nodes $\sim 5\sqrt{N}$, with N number of points [Tian et al, 2014], aspect ratio same as the ratio of the two largest eigenvalues of the covariance matrix of the data [Tian et al, 2014], $\sigma_0 \sim$ 20% of bigger matrix side, η_0=0.5, ~2.5 epochs, Gaussian neighbourhood function, Euclidean distance, PCA weight init

Unified distance matrix: node map showing the distance between the weights of nearby nodes

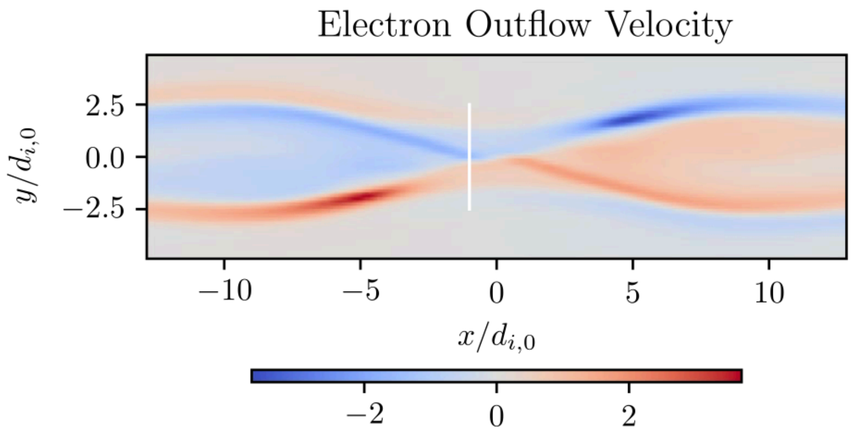


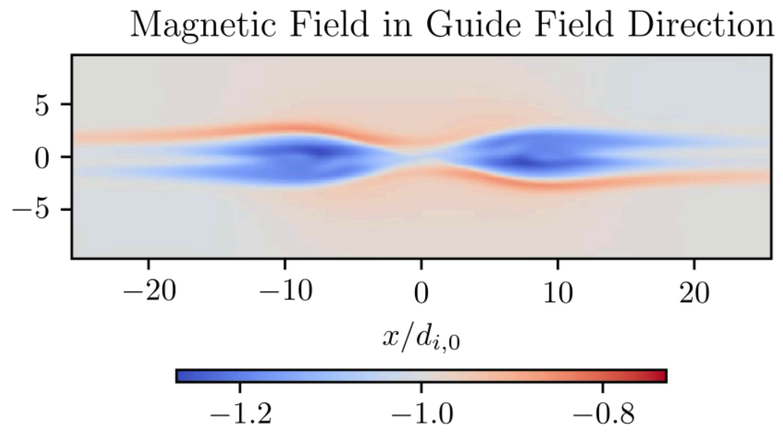


- The region of very similar nodes maps to the inflow, where “nothing much” occurs in the data
- The “walled-in“ nodes map to the plasmoid interior, arguably the simulation region most different from the others
- Nearby points in space (arguably, similar points) map to nearby SOM nodes

4. TRAINING A SOM ON A VLASOV SIMULATION OF MAGNETIC RECONNECTION

- The SOM is trained on a **Vlasov-Maxwell simulation** of magnetic reconnection done with the Muphy II code (Allmann-Rahn et al, 2022)
- 8π x 4π ion skin depths d_i reduced mass ratio mr=25, c=20 $v_{A,0}$
- Harris equilibrium with background species + initial perturbation
- $dx/d_i = 0.05$,
  -12.5 $v_{A,0}$<v_e<12.5$v_{A,0}$, 32^3 points
  -5.5 $v_{A,0}$<v_i<5.5 $v_{A,0}$, 18^3 points





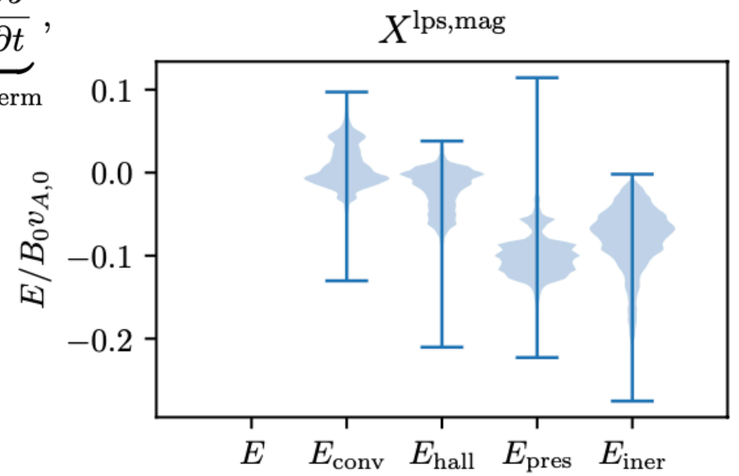
Feature used for training: components of the generalized Ohm's law

$$\vec{E} + \underbrace{\vec{v}\times\vec{B}}_{\text{Convective term}} = -\underbrace{\frac{\vec{J}\times\vec{B}}{n_e q_e}}_{\text{Hall term}} + \underbrace{\frac{\vec{\nabla}\cdot\vec{P}}{n_e q_e}}_{\text{Pressure term}} + \underbrace{\hat{\eta}\vec{J}}_{\text{Joule heating}} + \underbrace{\frac{m_e}{n_e q_e^2}\frac{\partial \vec{J}}{\partial t}}_{\text{Inertia term}},$$

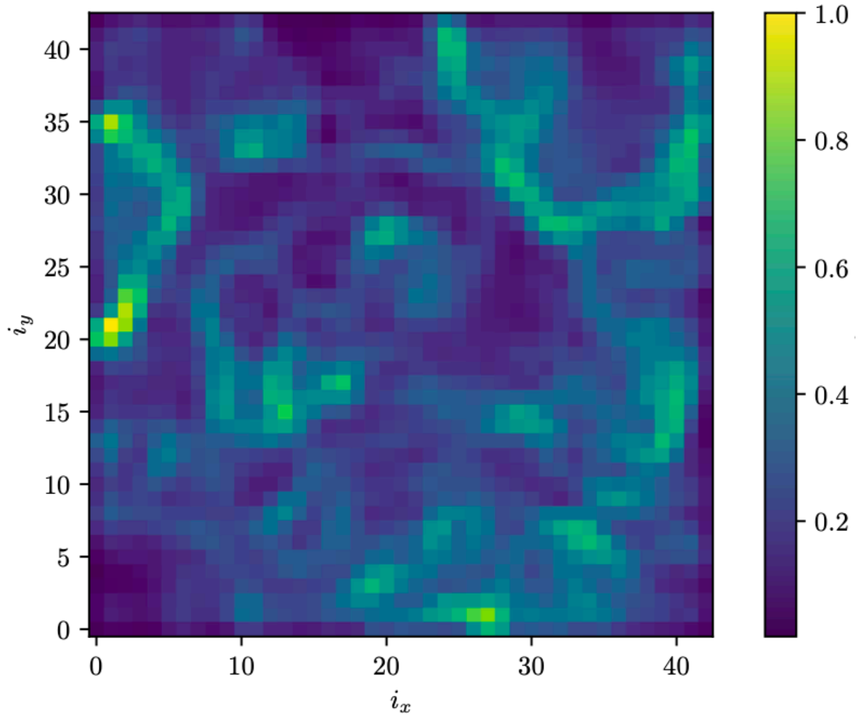
$\hat{\eta}\vec{J} \rightarrow 0$

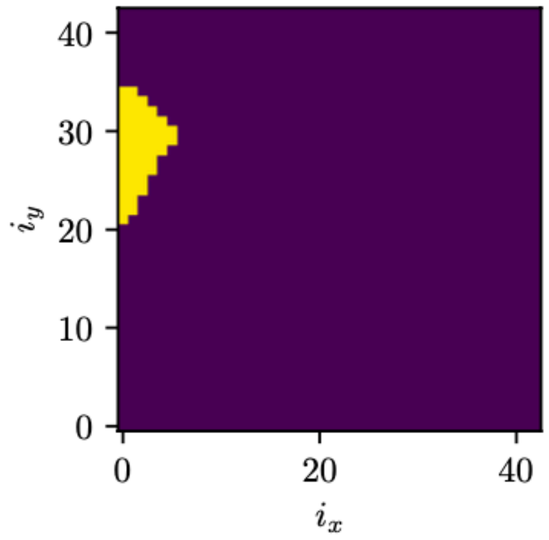
scaled to preserve their relative magnitude

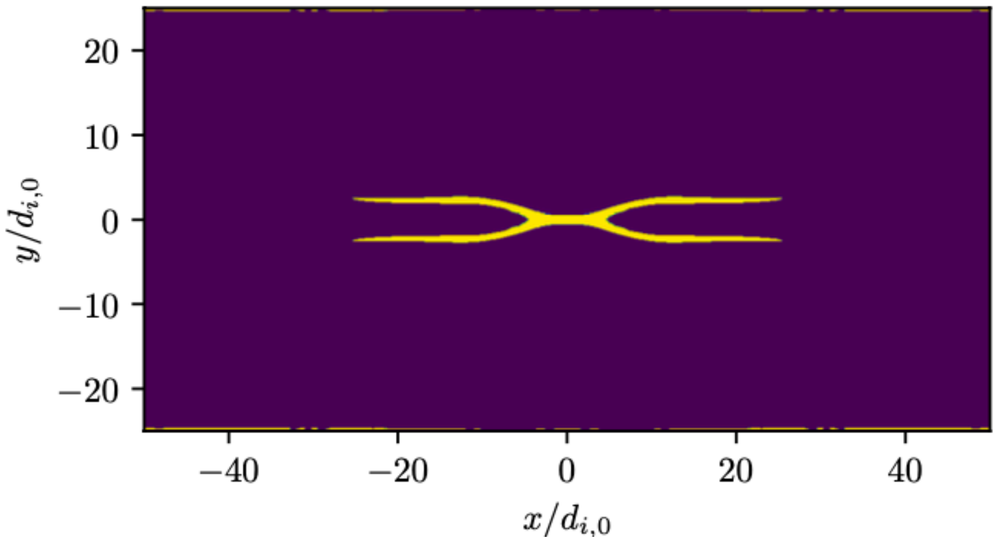
$$x_n^{\text{lps}} = \frac{1}{s}\ln\left(\frac{x_n}{\sum_{n\neq i} x_i}\right)$$



UNIFIED DISTANCE MATRIX ANALYSIS







SOM hyperparameters:

42x42 nodes [Tian et al, 2014]

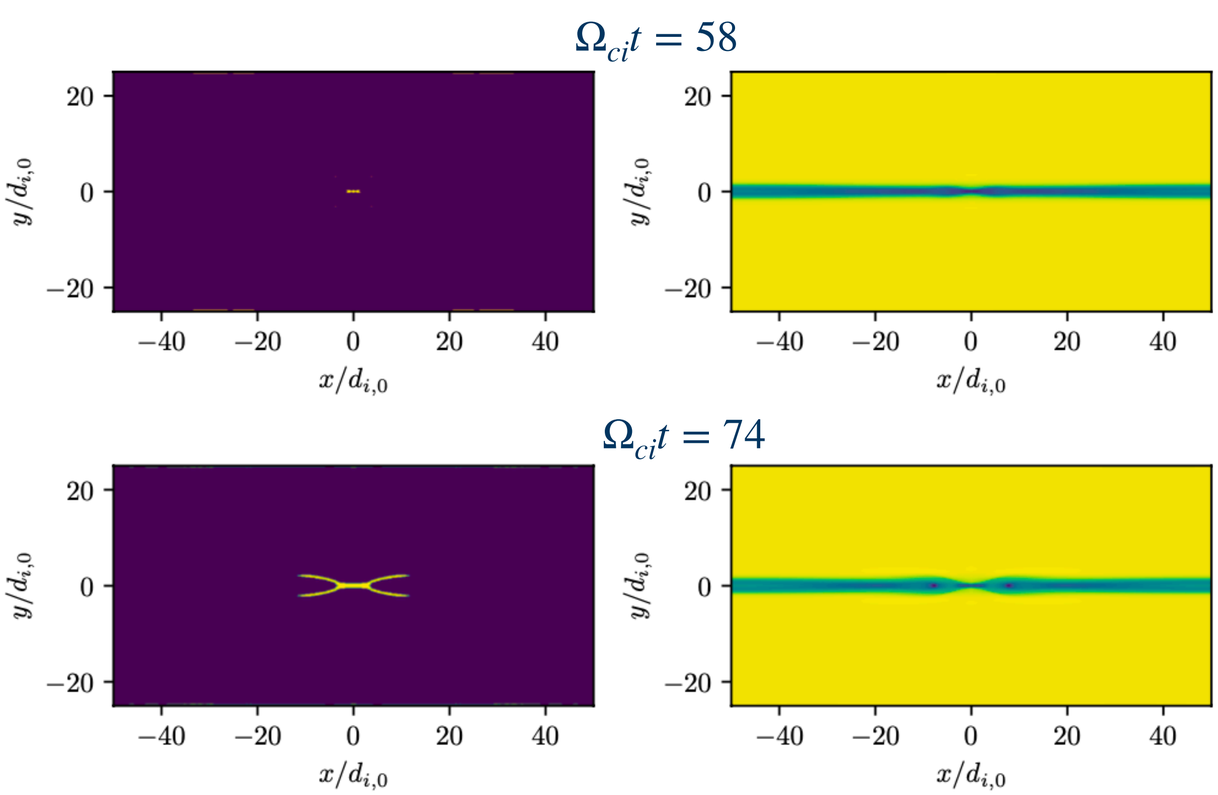
η_0= 0.1, σ_0=2.5,

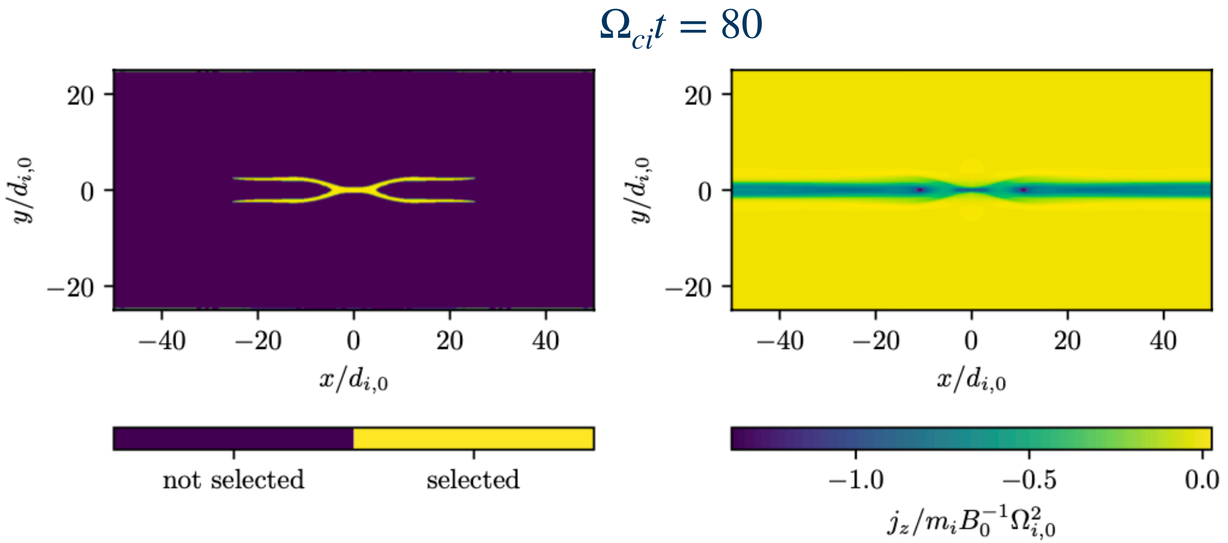
10 epochs

The “walled-in” region in the unified distance map maps to diffusion region + separatrices

A ROBUST WAY TO IDENTIFY DIFFUSION REGION + SEPARATRICES

Robustness to **temporal evolution**

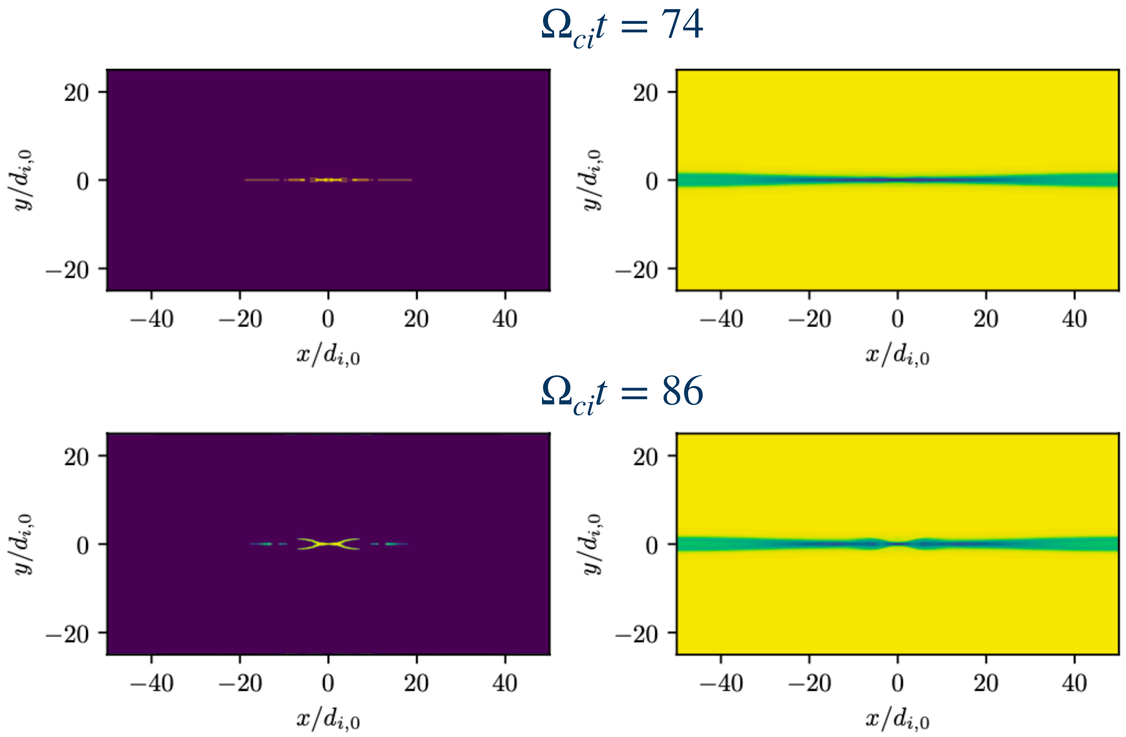


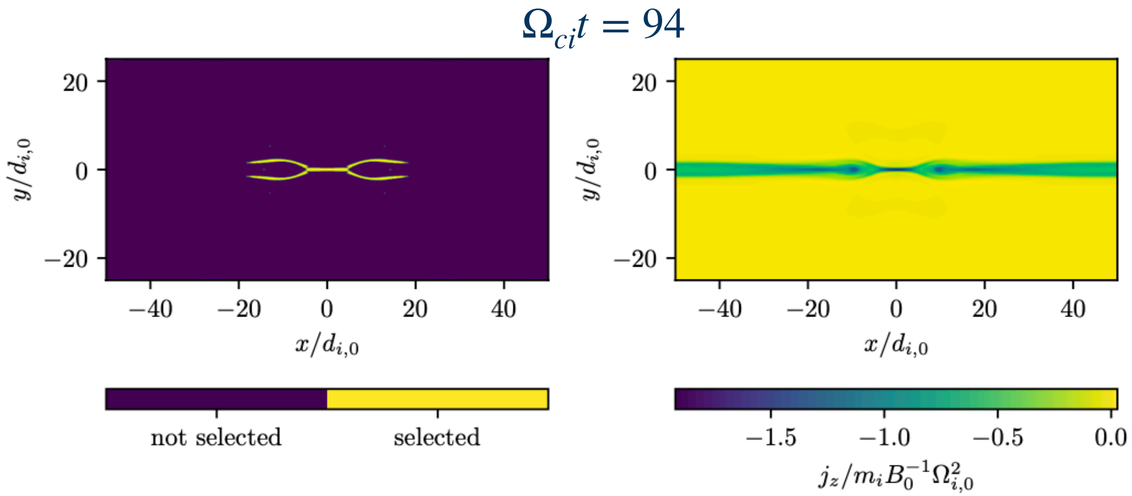


Also at **different simulated times**, diffusion region + separatrices map to the same SOM nodes

A ROBUST WAY TO IDENTIFY DIFFUSION REGION + SEPARATRICES

Robustness to **different simulation methods**





Diffusion region + separatrices are identified also in **multifluid simulations**, using the map trained on Vlasov-Maxwell simulations

SUMMARY

- We have applied an unsupervised classification technique based on Self Organizing Maps to simulated data: MHD simulations of the terrestrial magnetosphere, Particle In Cell simulations of plasmoid instability, Maxwell-Vlasov and multifluid simulations of magnetic reconnection
- We identify macro-scale regions that map well to our a priori knowledge of the processes
- Analysing feature maps and unified distance matrix we unlock information on the data
- The classification method appears robust to temporal variation and even to the numerical method used for the simulations
- This technique then emerges as an **effective and versatile method to classify simulated data**

Amaya, J., Dupuis, R., Innocenti, M. E., and Lapenta, G.: Visualizing and Interpreting Unsupervised Solar Wind Classifications, Front. Astron. Space Sci., 7, 66, https://doi.org/10.3389/fspas.2020.553207, 2020.

Innocenti, M. E., Amaya, J., Raeder, J., Dupuis, R., Ferdousi, B., & Lapenta, G. (2021). Unsupervised classification of simulated magnetospheric regions. Annales Geophysicae Discussions, 1-28.
https://angeo.copernicus.org/articles/39/861/2021/angeo-39-861-2021.pdf