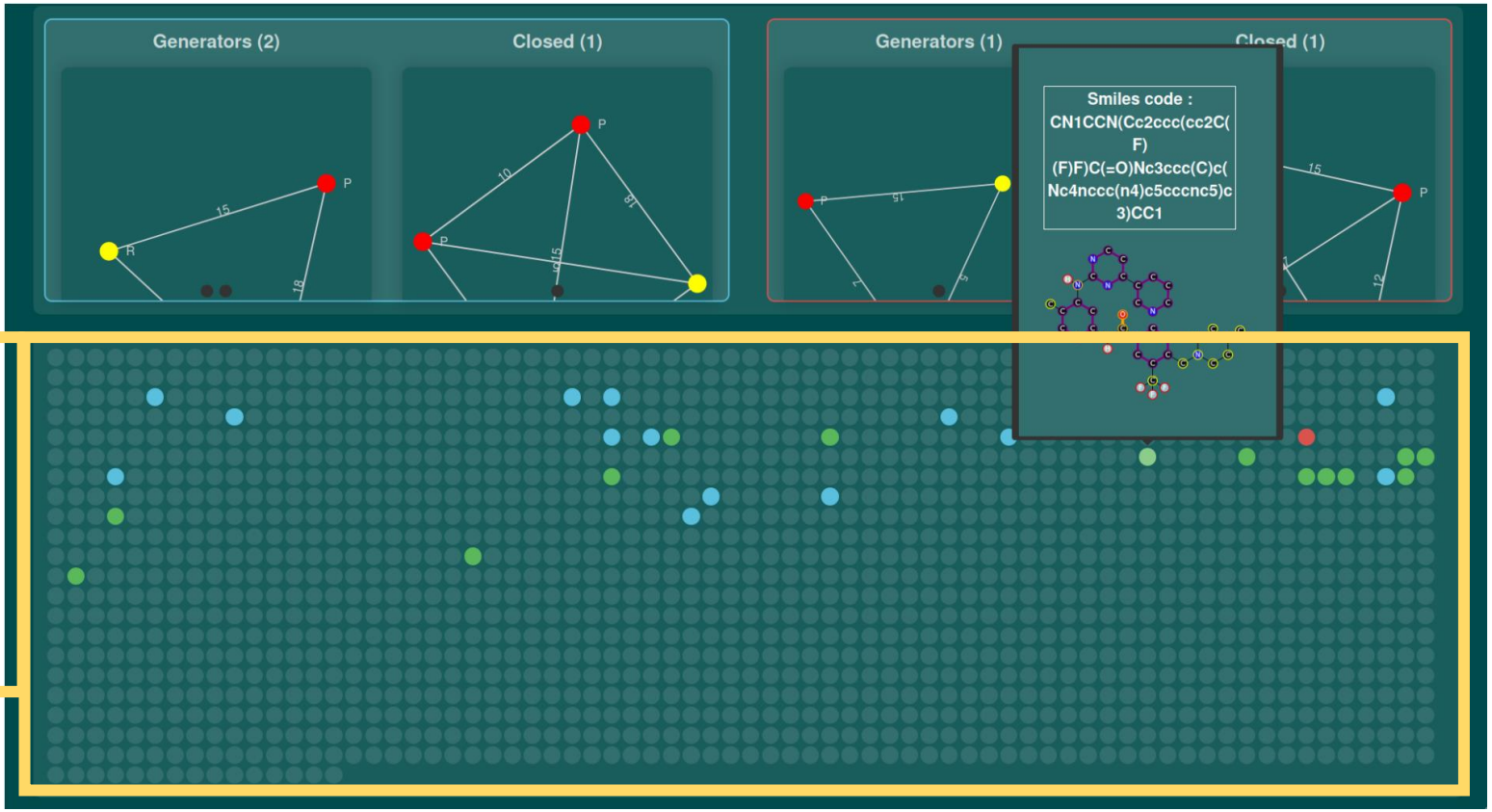
Visualisation compacte de données multidimensionnelles

David Auber

En collaboration avec:

Adrien Halnaut, Romain Giot et Romain Bourqui

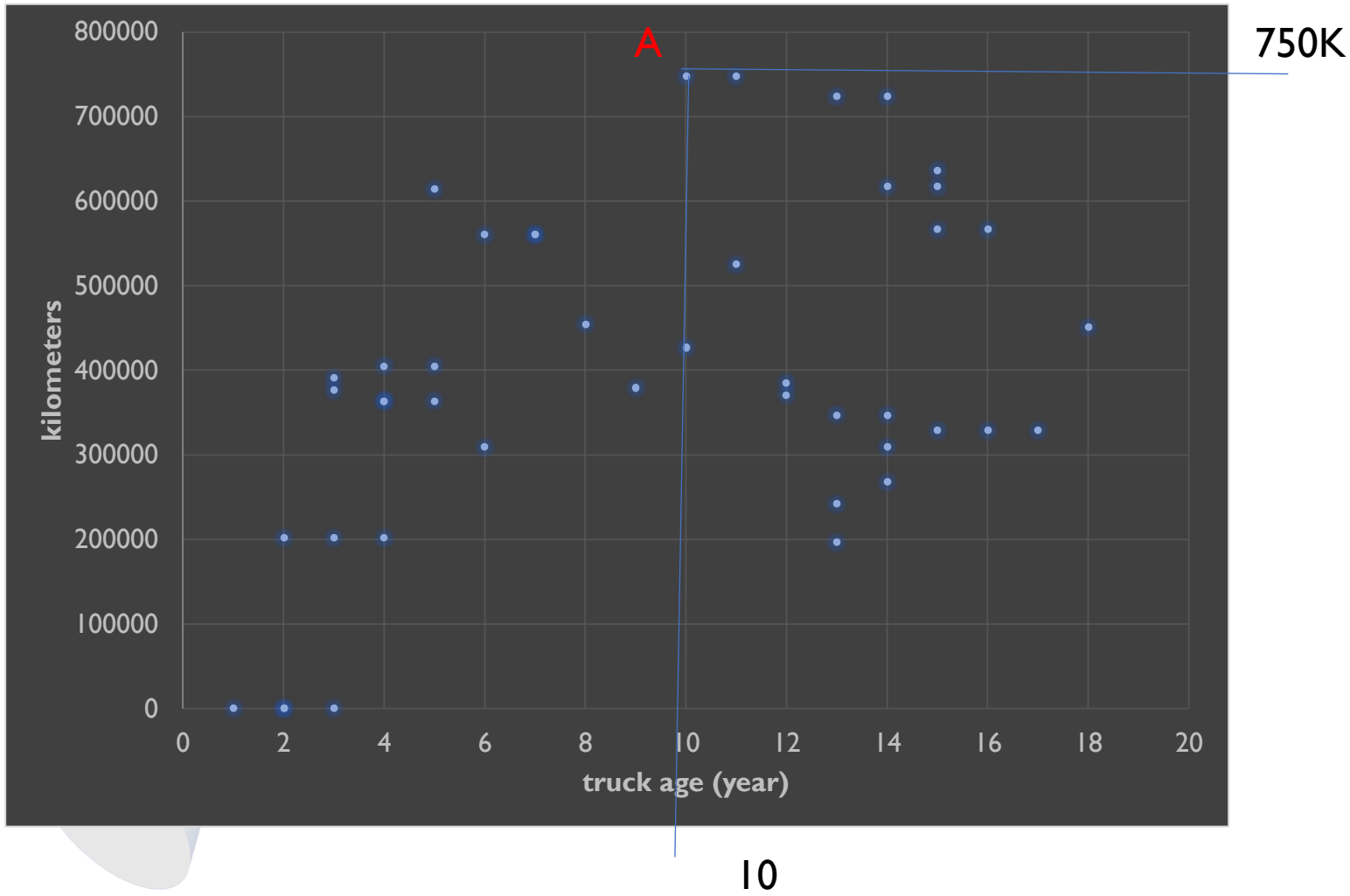
Données multidimensionnel dans InvolvD



Représentation 2D de donnée multi dimensionnelles des molécules (mindmap)

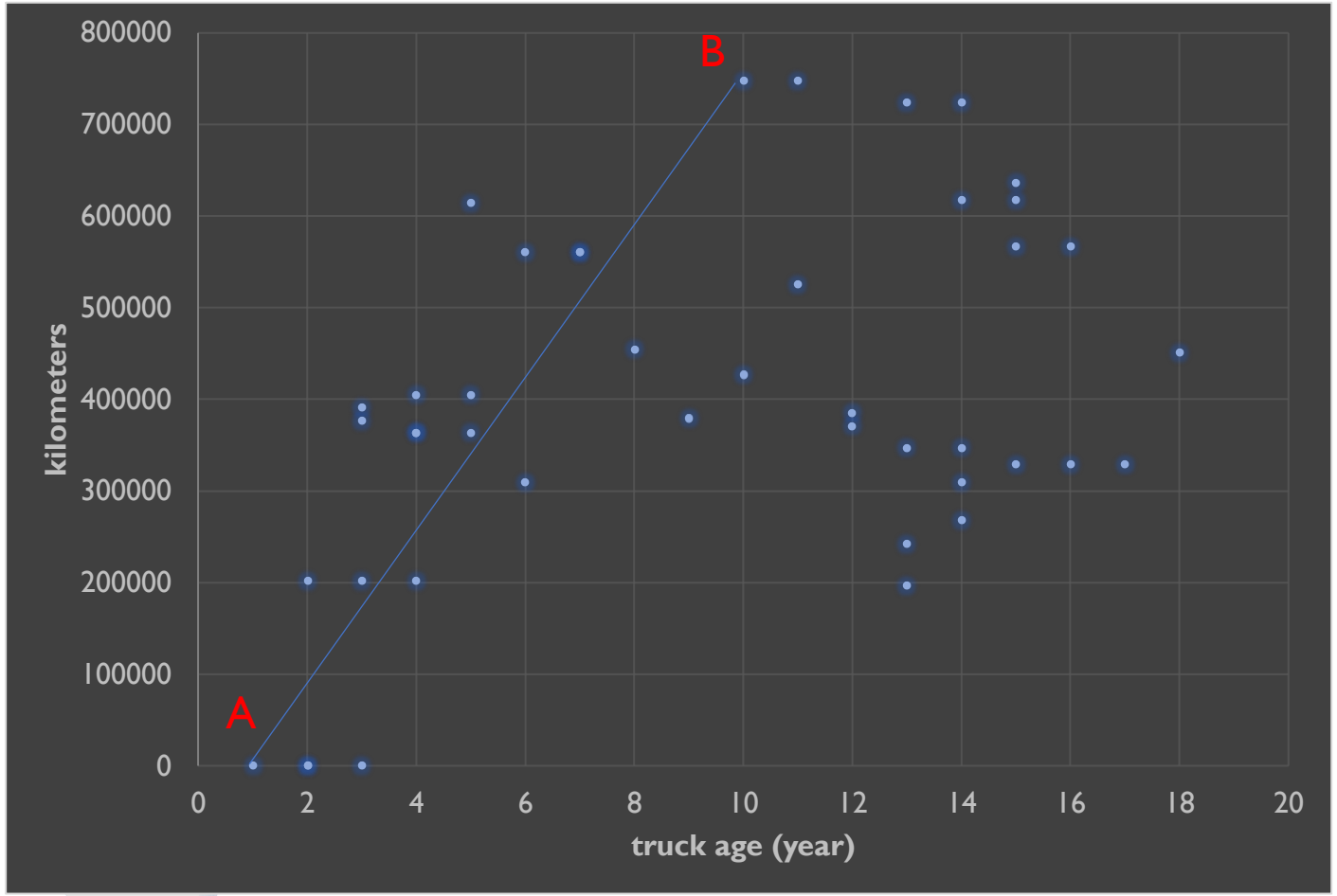
What are we looking for ?

- Read values of data points ?
- Compare data points to each other ?
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- Make assumptions ?
- Make decisions ?



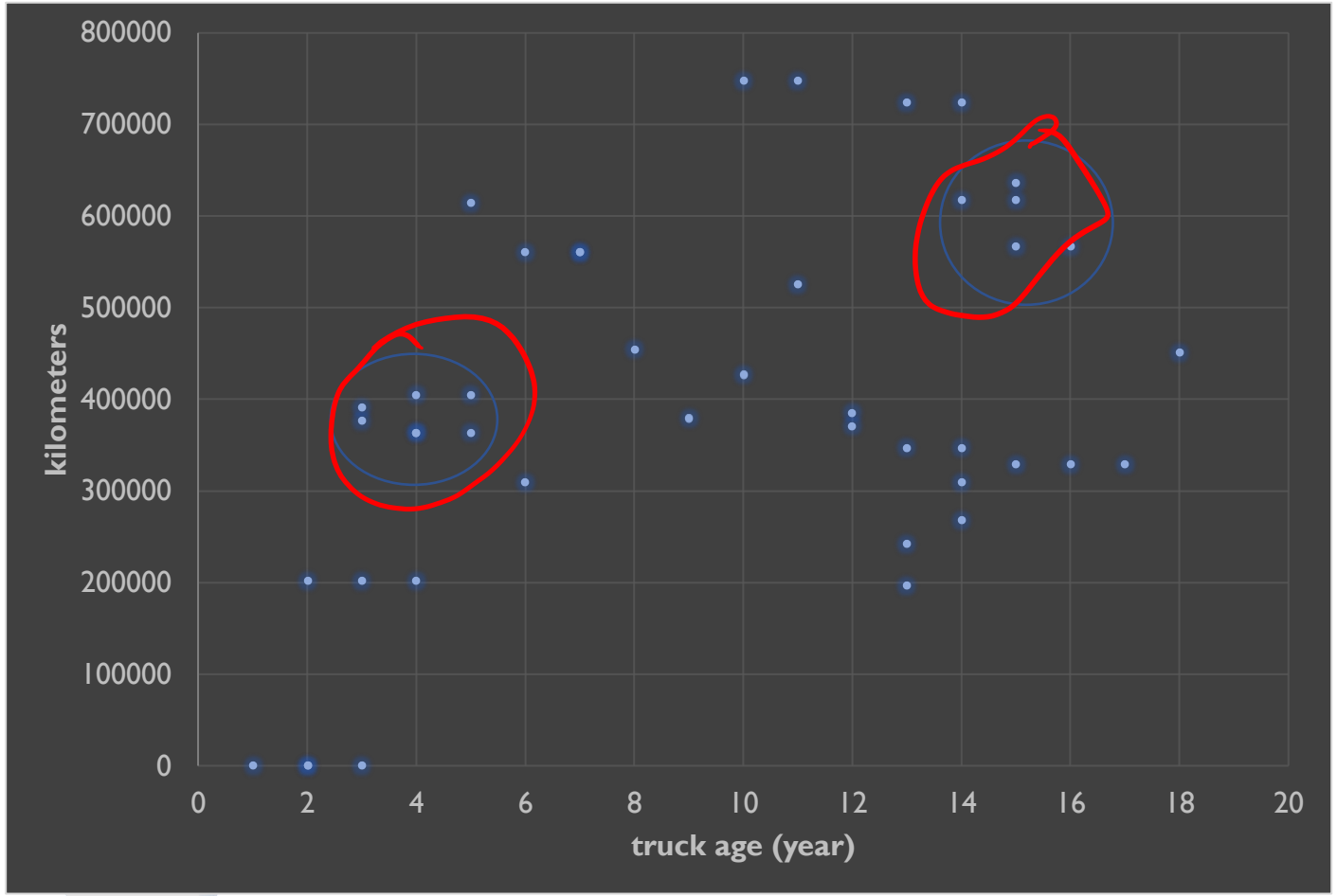
What are we looking for ?

- Read values of data points ?
- Compare data points to each other ?
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- Make assumptions ?
- Make decisions ?



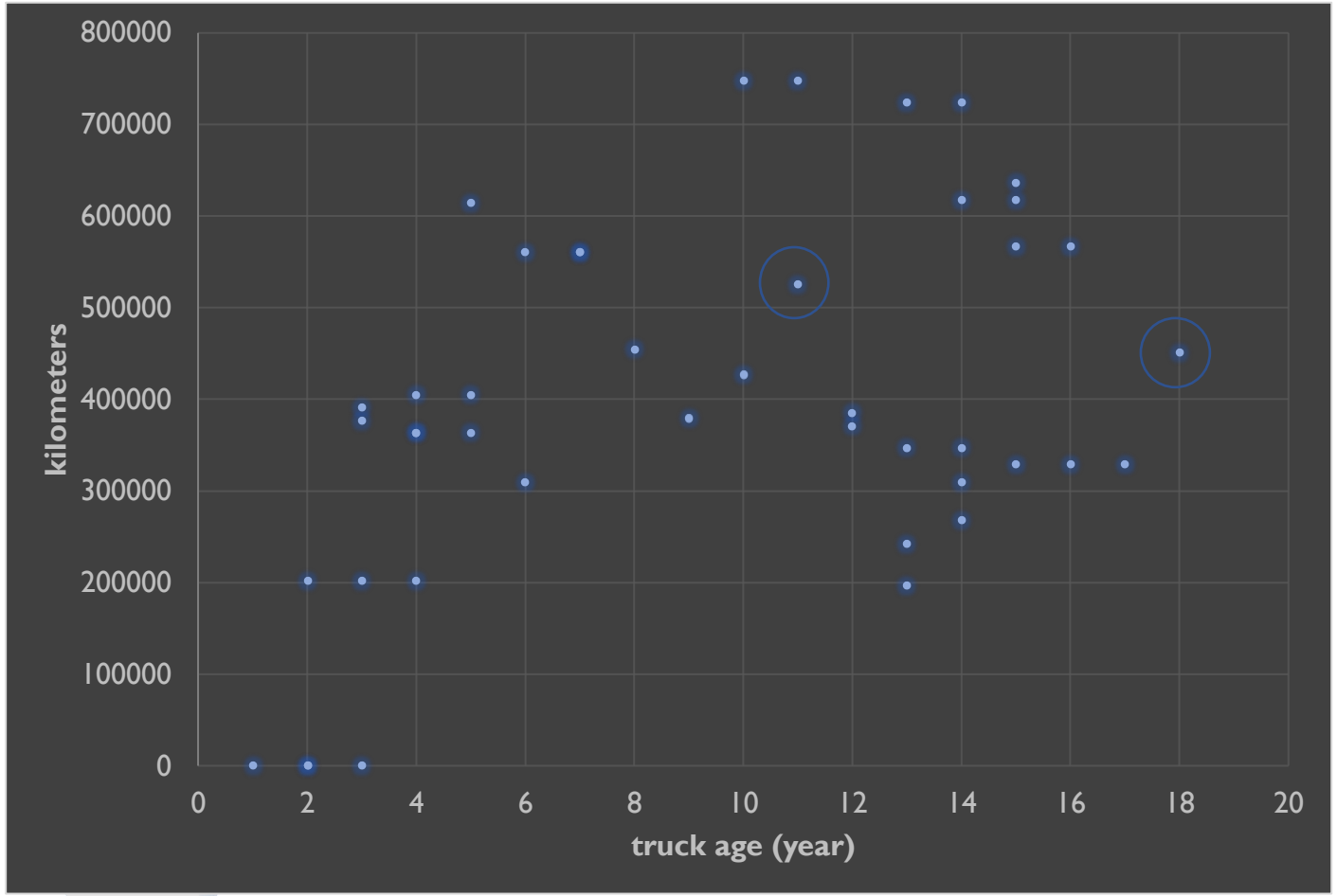
What are we looking for ?

- Read values of data points ?
- Compare data points to each other ?
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- Make assumptions ?
- Make decisions ?



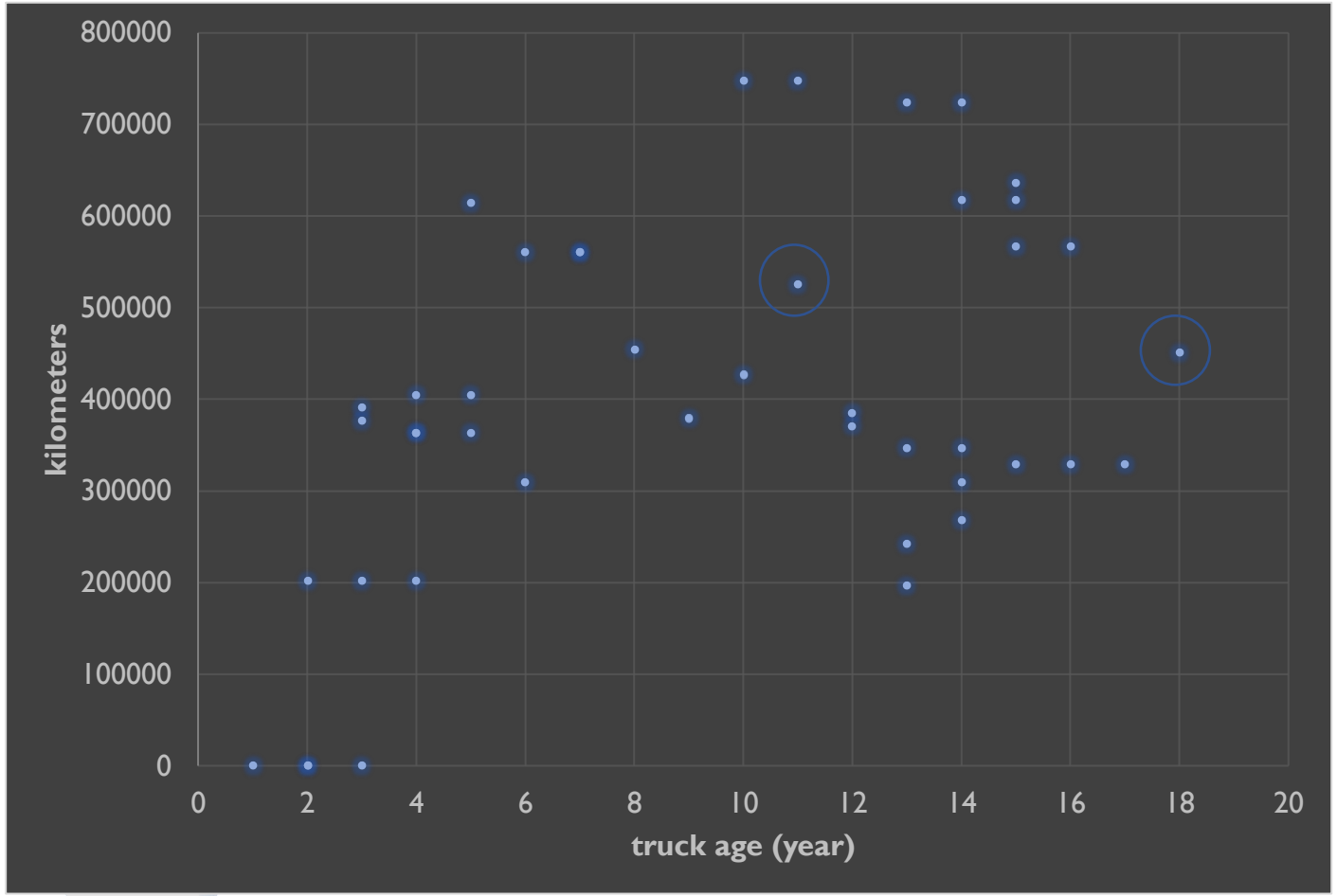
What are we looking for ?

- Read values of data points ?
- Compare data points to each other ?
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- Make assumptions ?
- Make decisions ?

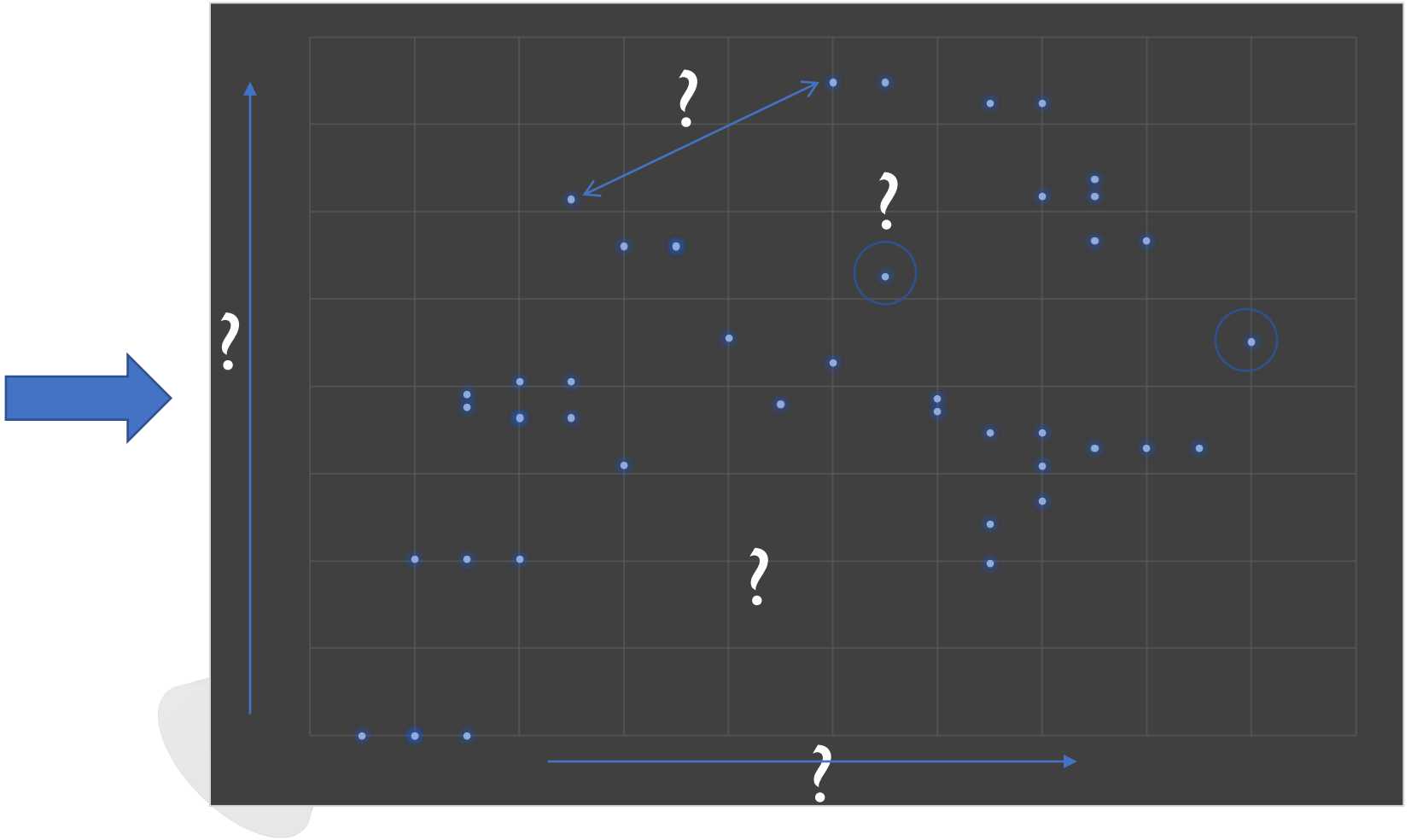


What are we looking for ?

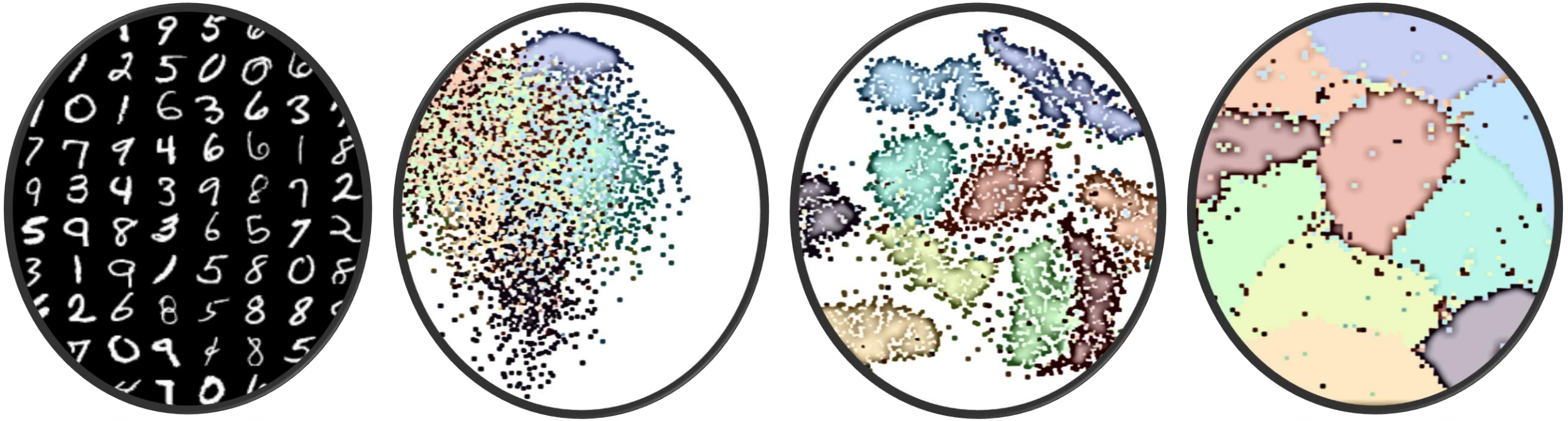
- Read values of data points ?
- Compare data points to each other ?
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- Make assumptions ?
- Make decisions ?



Curse of dimensionality



Curse of dimensionality (1)



Void-Vis Vs Compact-Vis

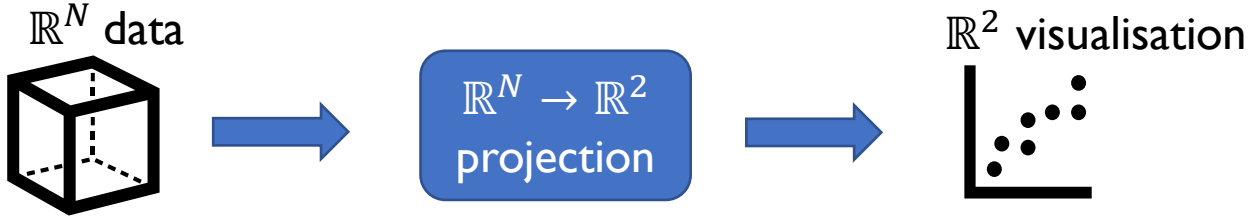
Dimensionality reduction R^n -> R^2

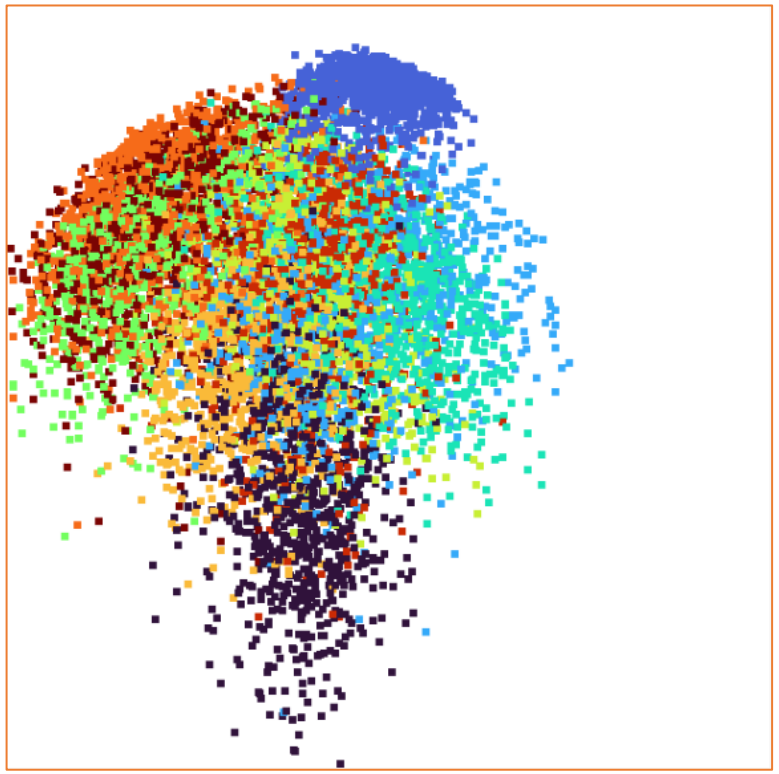
Linear :

- Orthogonal projection
- PCA
- …

Non linear

- Multi dimensinal scaling
- Graph-based kernel
- T-Sne
- Autoencoder
- …





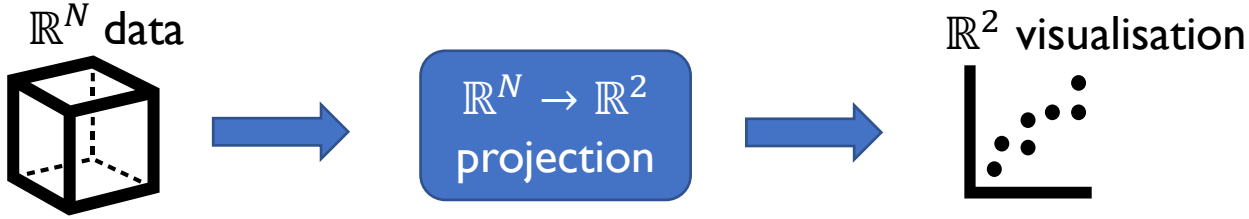
Dimensionality reduction R^n -> R^2

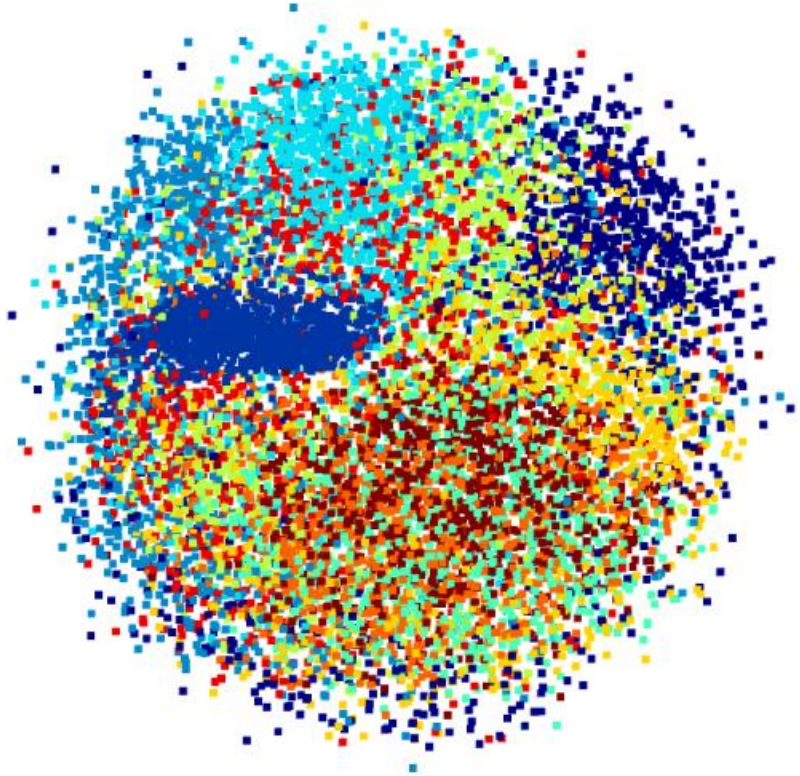
Linear :

- Orthogonal projection
- PCA
- …

Non linear

- Multi dimensinal scaling
- Graph-based kernel
- T-Sne
- Autoencoder
- …





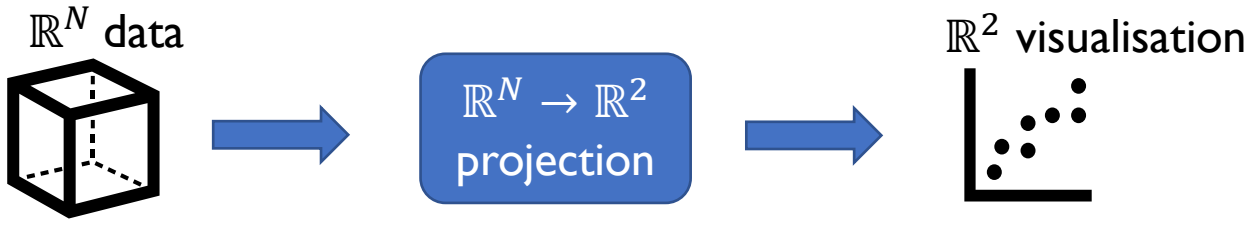
Dimensionality reduction R^n -> R^2

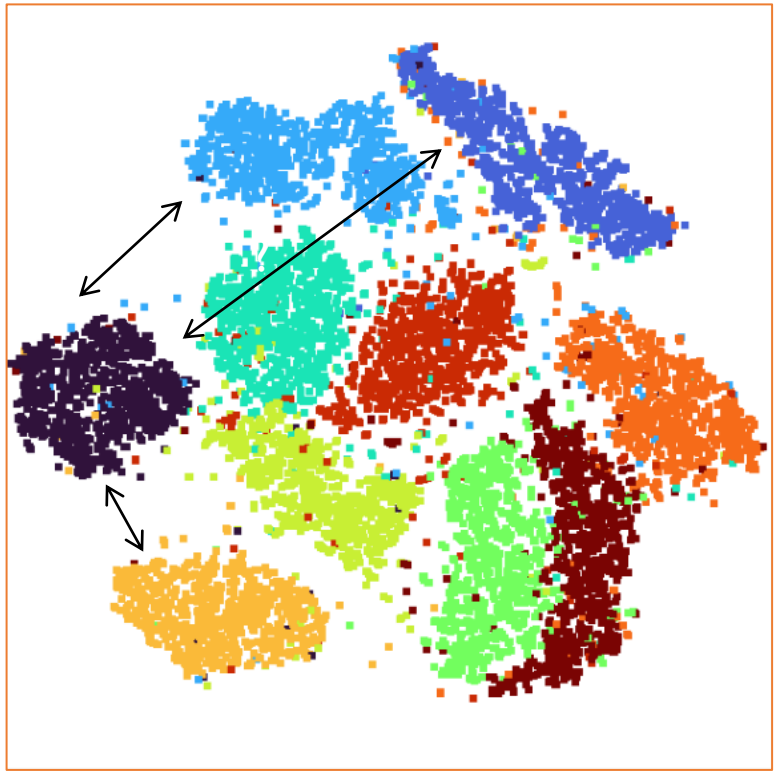
Linear :

- Orthogonal projection
- PCA
- …

Non linear

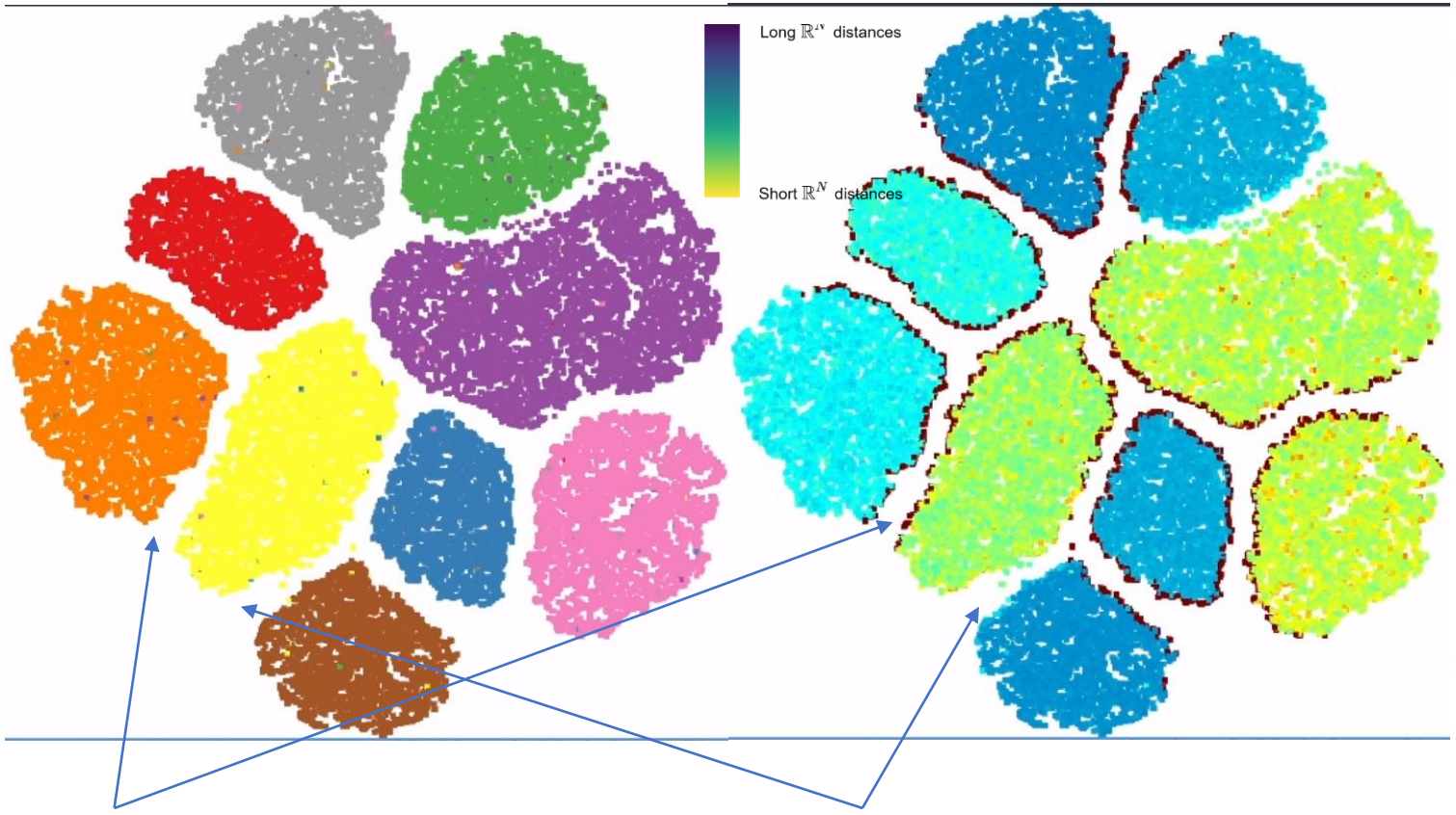
- Multi dimensinal scaling
- Graph-based kernel
- Tsne
- Autoencoder
- ….





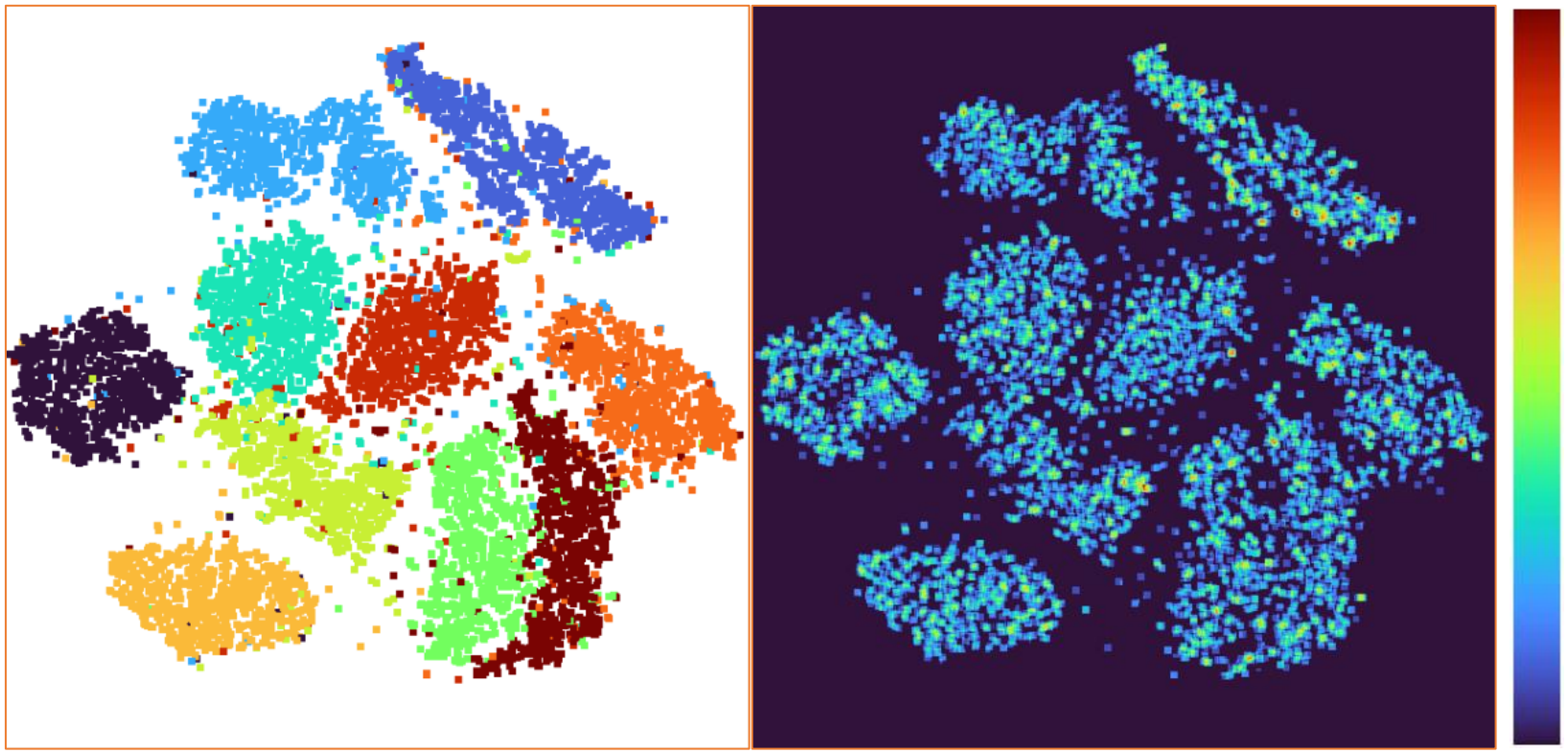
Void visualisation is misleading

- ~~Read values of data points ?~~
- Compare data points ~~to each other ?~~ neighboors
- ~~Estimate density or distribution ?~~
- ~~Find outliers ?~~
- Find clusters of data points ?
- ~~Compare clusters to each other ?~~
- ~~Make assumptions ?~~
- ~~Make decisions ?~~



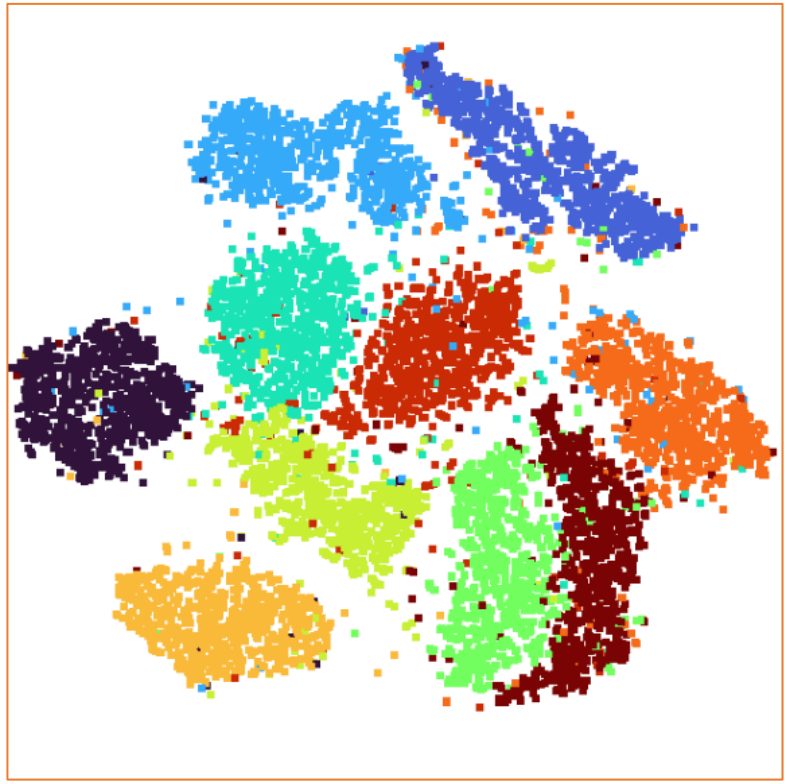
Void visualisation creates cluttering

- ~~Read values of data points ?~~
- Compare data points ~~to each other ?~~ neighboors
- ~~Estimate density or distribution ?~~
- ~~Find outliers ?~~
- Find clusters of data points ?
- ~~Compare clusters to each other ?~~
- ~~Make assumptions ?~~
- ~~Make decisions ?~~

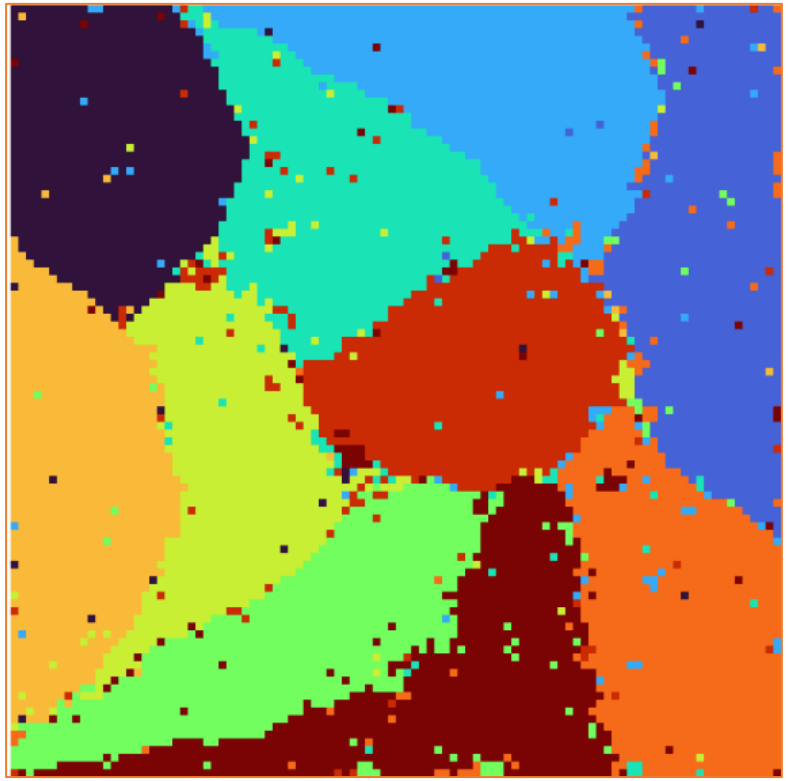


Void-Vis Vs Compact-Vis

- ~~Read values of data points ?~~
- Compare data points ~~to each other ?~~ neighboors
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- ~~Make assumptions ?~~
- ~~Make decisions ?~~



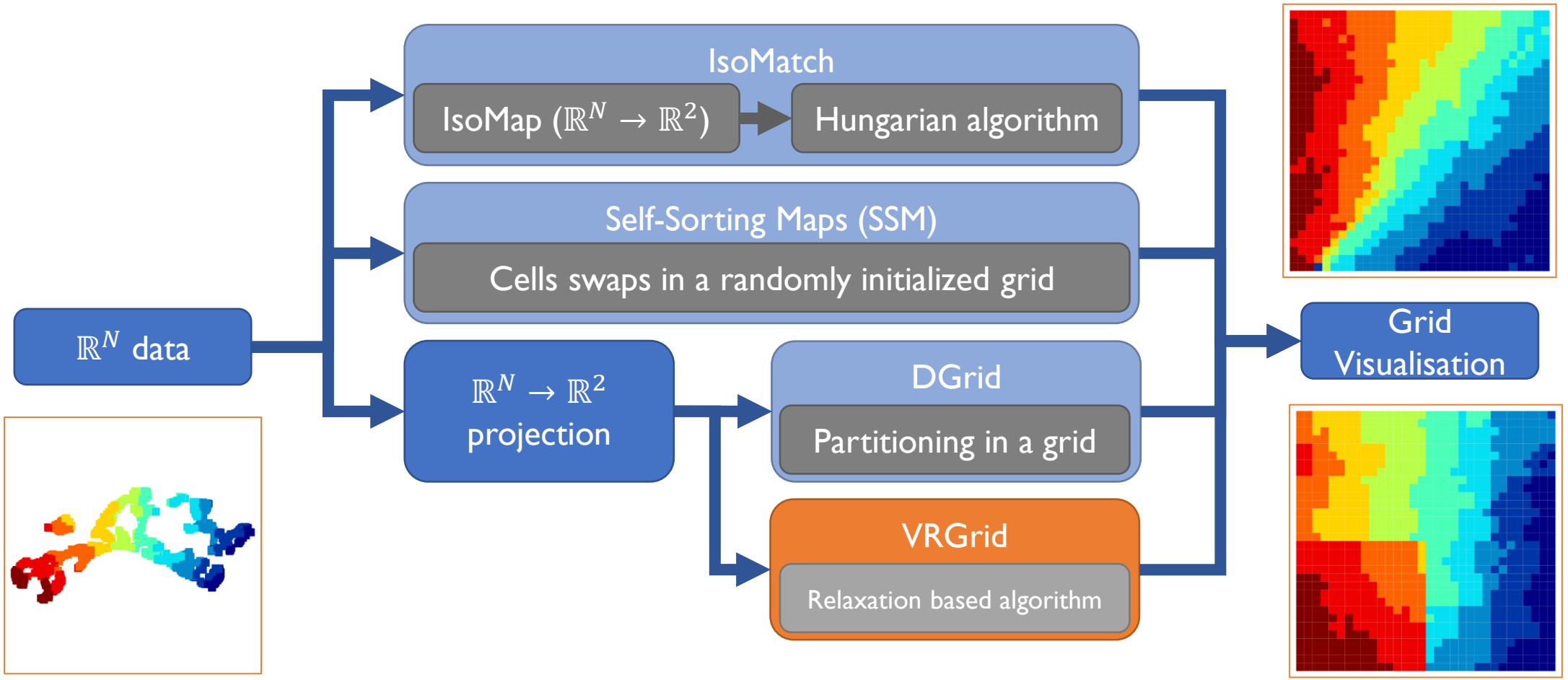
Data points & void visualisation



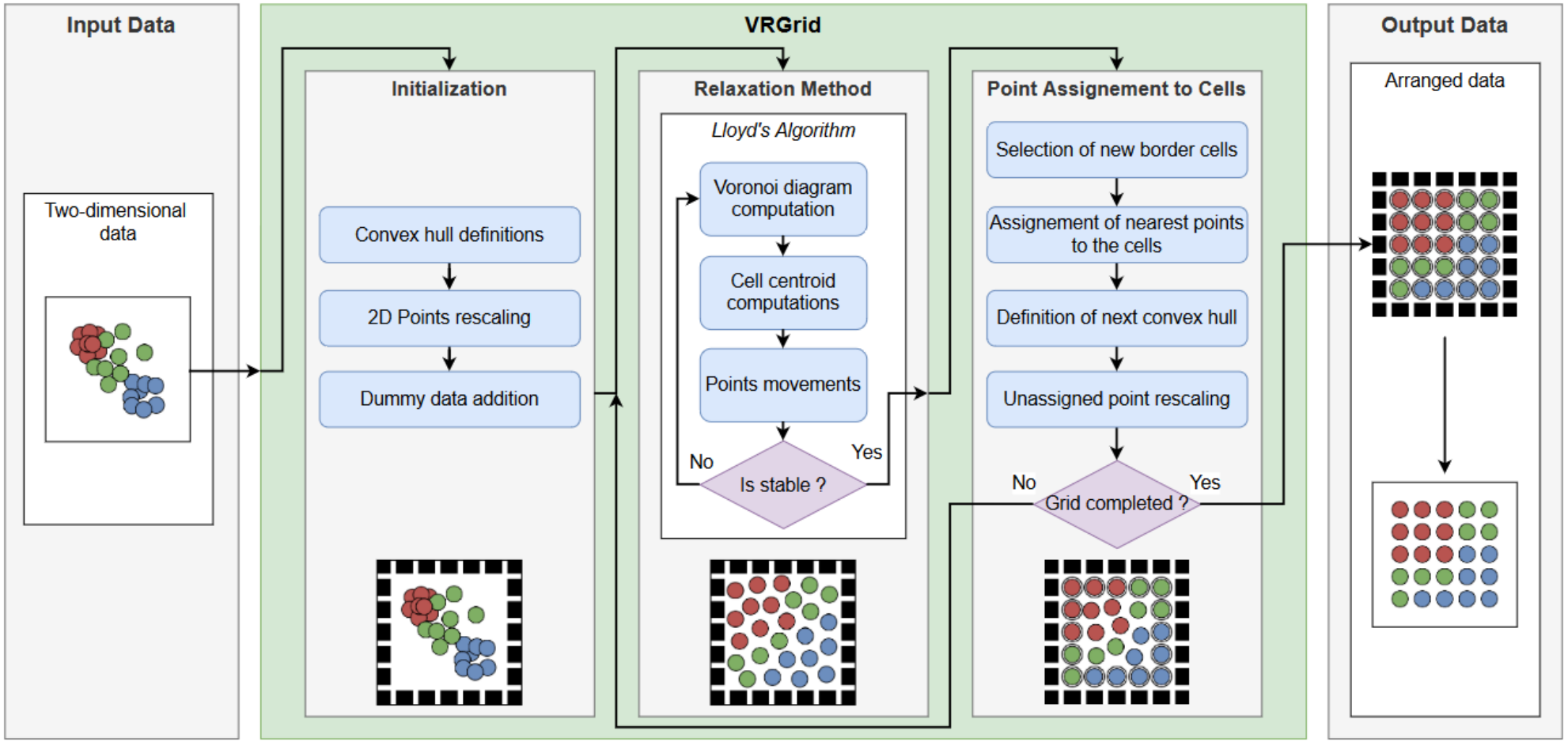
Data points visualisation

How to transform void to non void vis

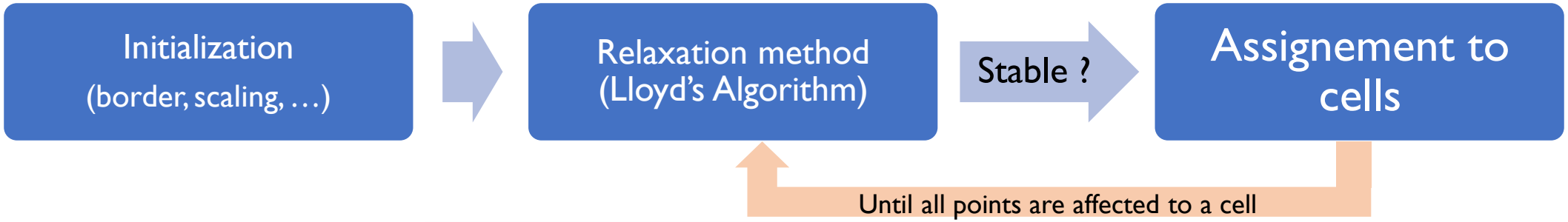
Grid arrangement methods

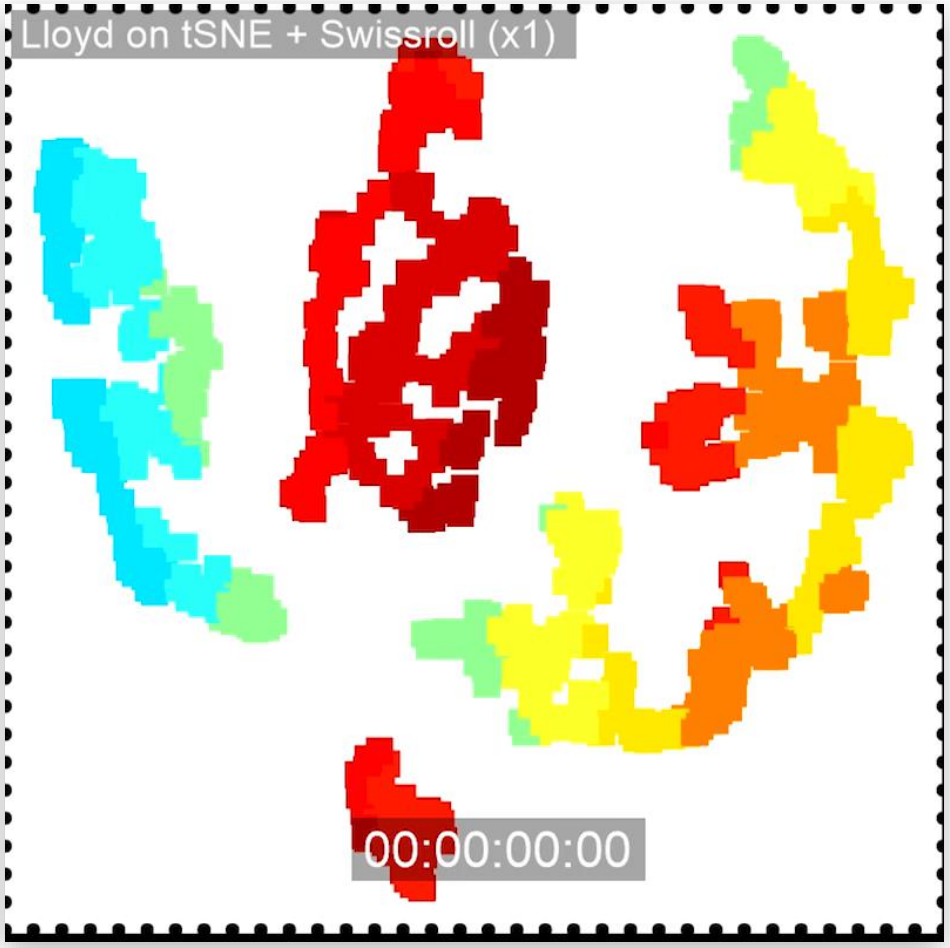


VRGrid presentation



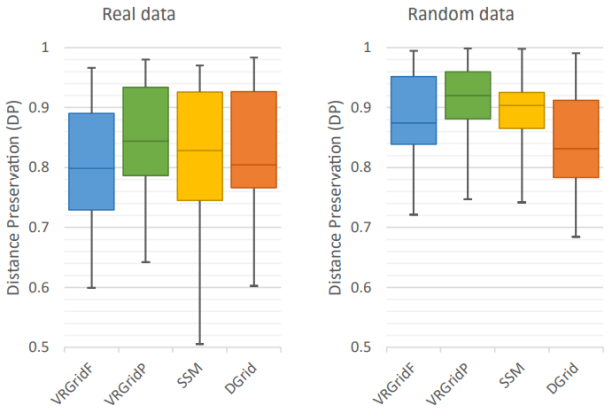
$$O(\sqrt{n} \,.\, i \,.\, n \,.\, \log(n)) = O\left(i \,.\, n^{\frac{3}{2}} \,.\, \log(n)\right)$$


Initialization
(border, scaling, …)
Relaxation method
(Lloyd's Algorithm)
Stable ?
Assignement to cells
Until all points are affected to a cell

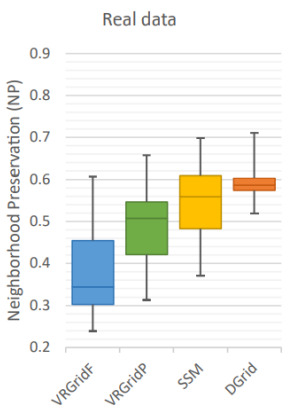

Lloyd on tSNE + Swissroll (x1)
00:00:00:00

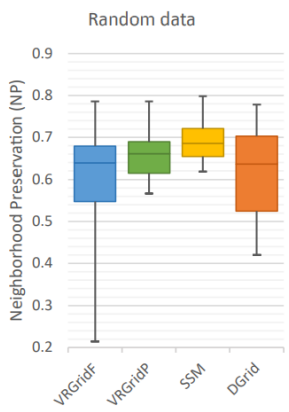
Results

Higher is better



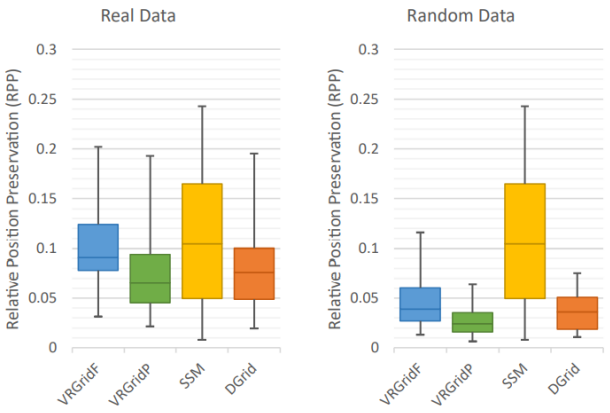
(a) Distance Preservation measure.



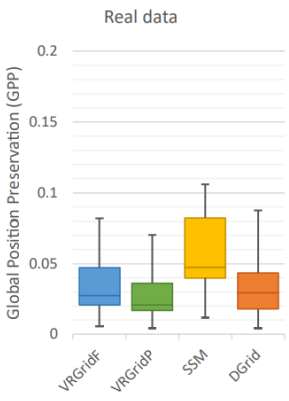


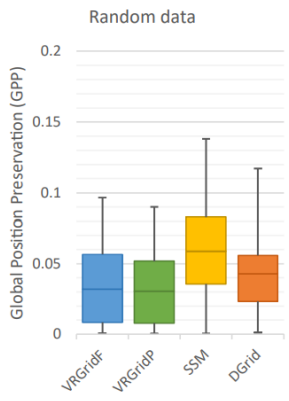
(b) Neighborhood Preservation measure.

Lower is better



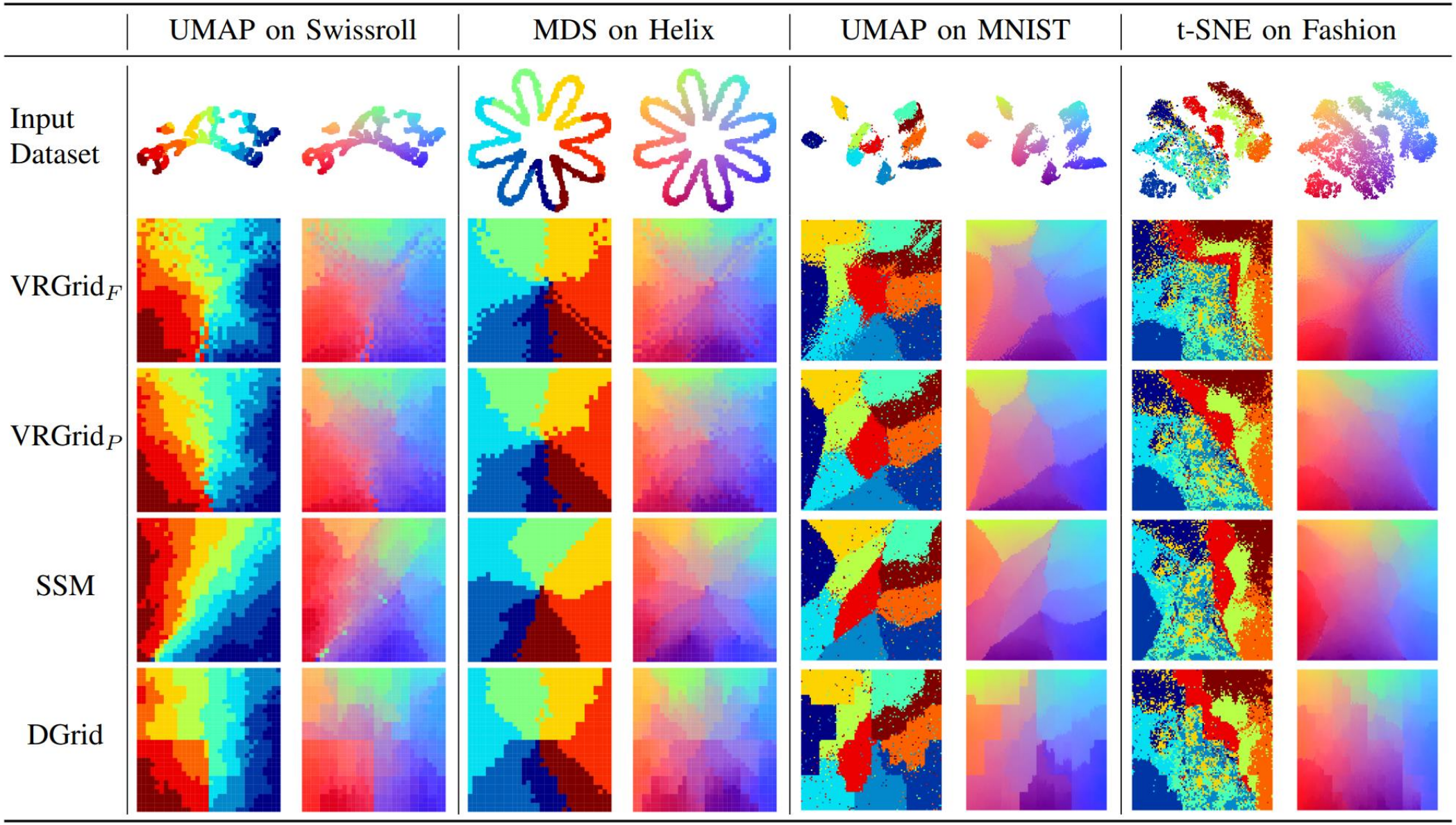
(c) Relative Position Preservation measures.





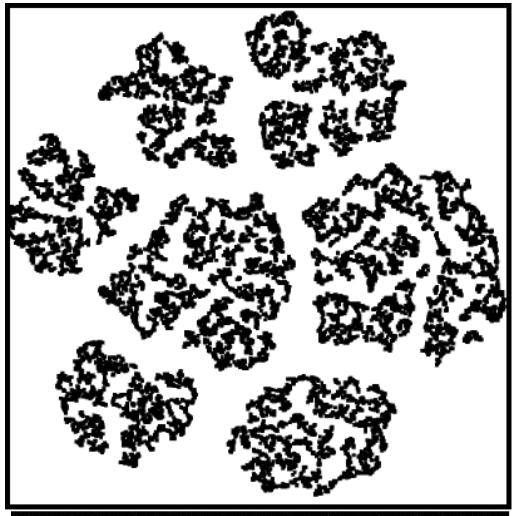
(d) Global Position Preservation measure.

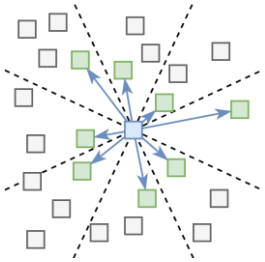
- VRGridP has the best results in all categories except Neighborhood Preservation
- VRGridF's fast approach has a high cost on quality
- DGrid has relatively good results despite the visually strong distortions (→ wrong metrics ?)
- SSM has bad* results on preservation of positions.
- However, VRGridP and VRGridF are **both significantly slower than SSM and DGrid on computation time**.


UMAP on Swissroll
MDS on Helix
UMAP on MNIST
t-SNE on Fashion
Input Dataset
VRGrid$_F$
VRGrid$_P$
SSM
DGrid

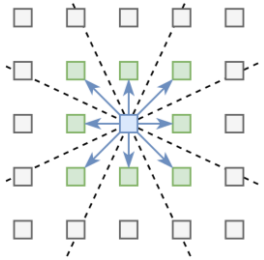
Void-Vis Vs Compact-Vis

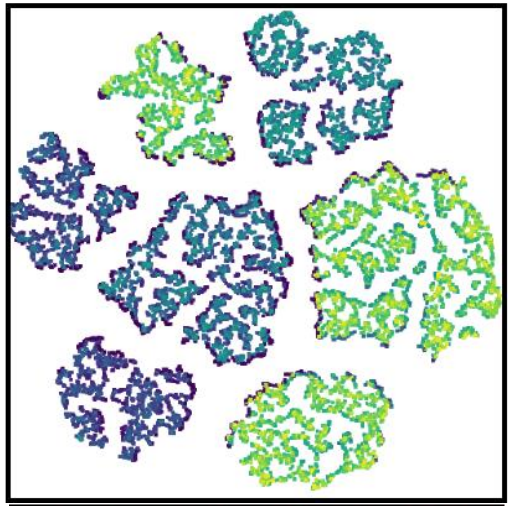
- ~~Read values of data points ?~~
- Compare data points ~~to each other ?~~ neighboors
- Estimate density or distribution ?
- Find outliers ?
- Find clusters of data points ?
- Compare clusters to each other ?
- ~~Make assumptions ?~~
- ~~Make decisions ?~~

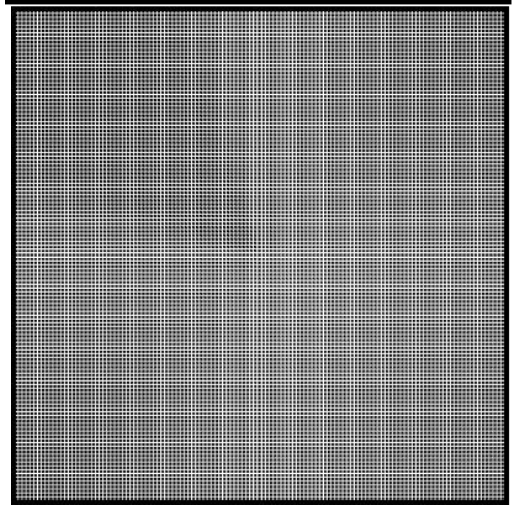


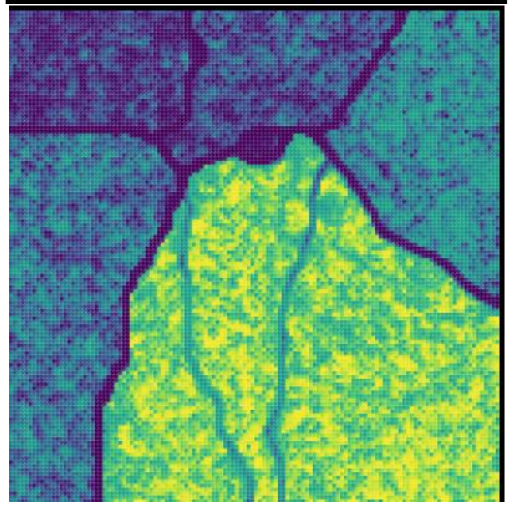


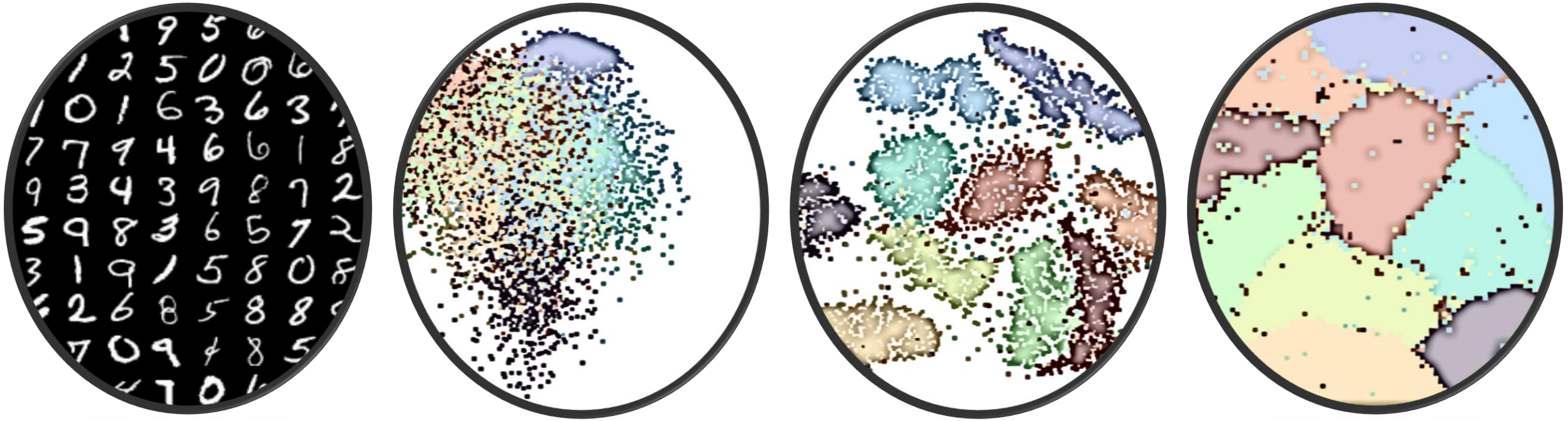
Mean distance to 2d neighboors in the high dimensional space







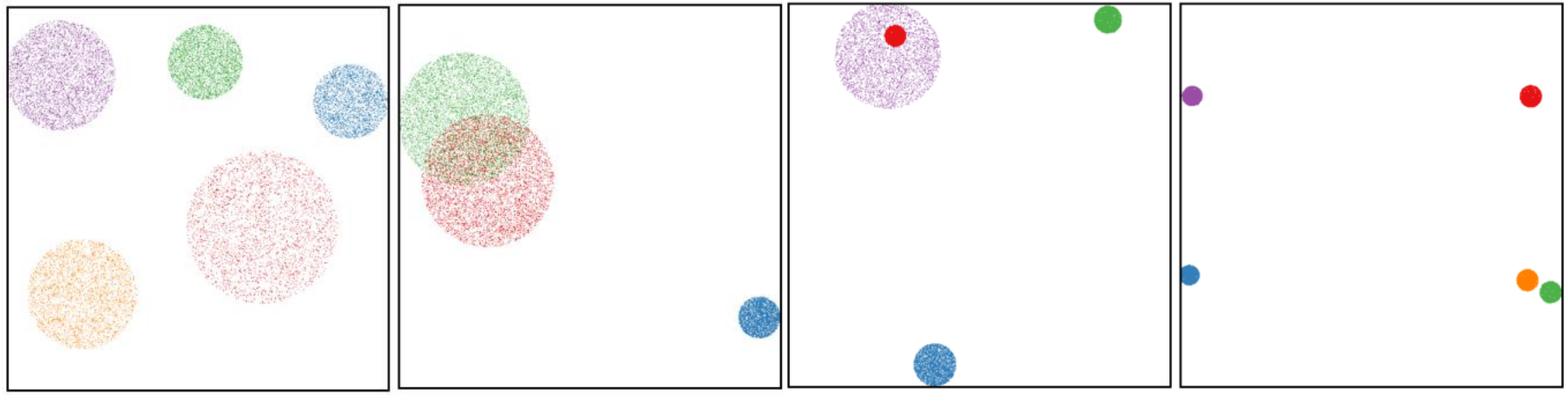




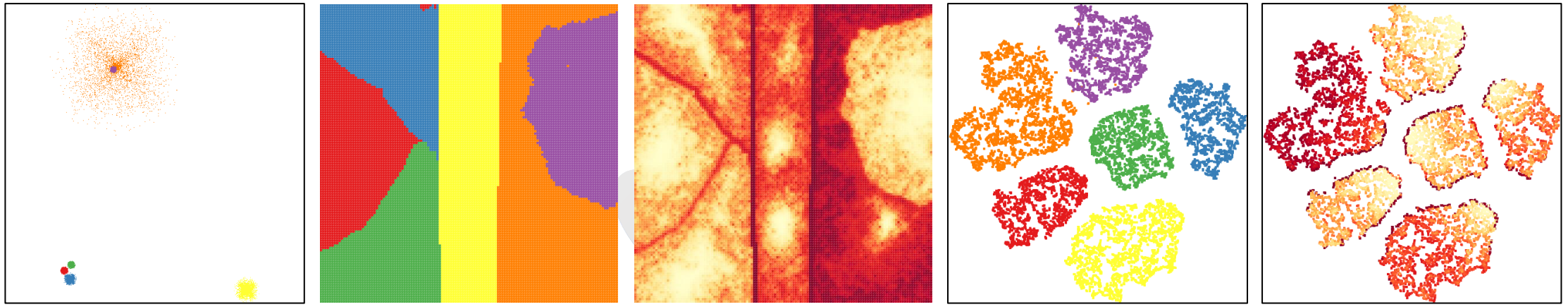
User evaluation

- https://pivert.labri.fr/eval-en/index.html

Random generation of dataset



Parameters	clusters sizes	Overlaping clusters	Nested clusters	Close clusters



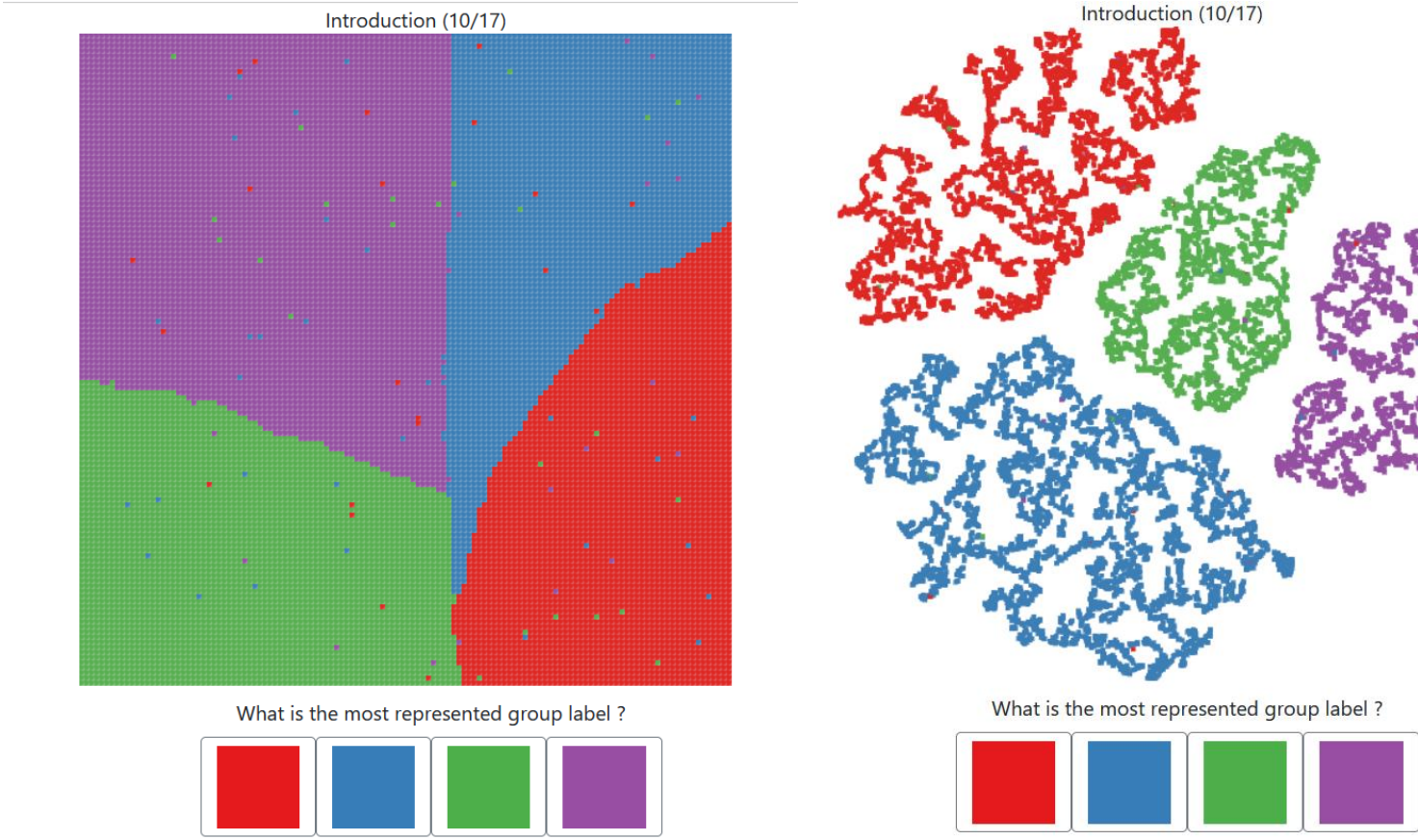
User tasks

Find largest group

Find closest group

Find number of outliers

Find the type of topogical structure



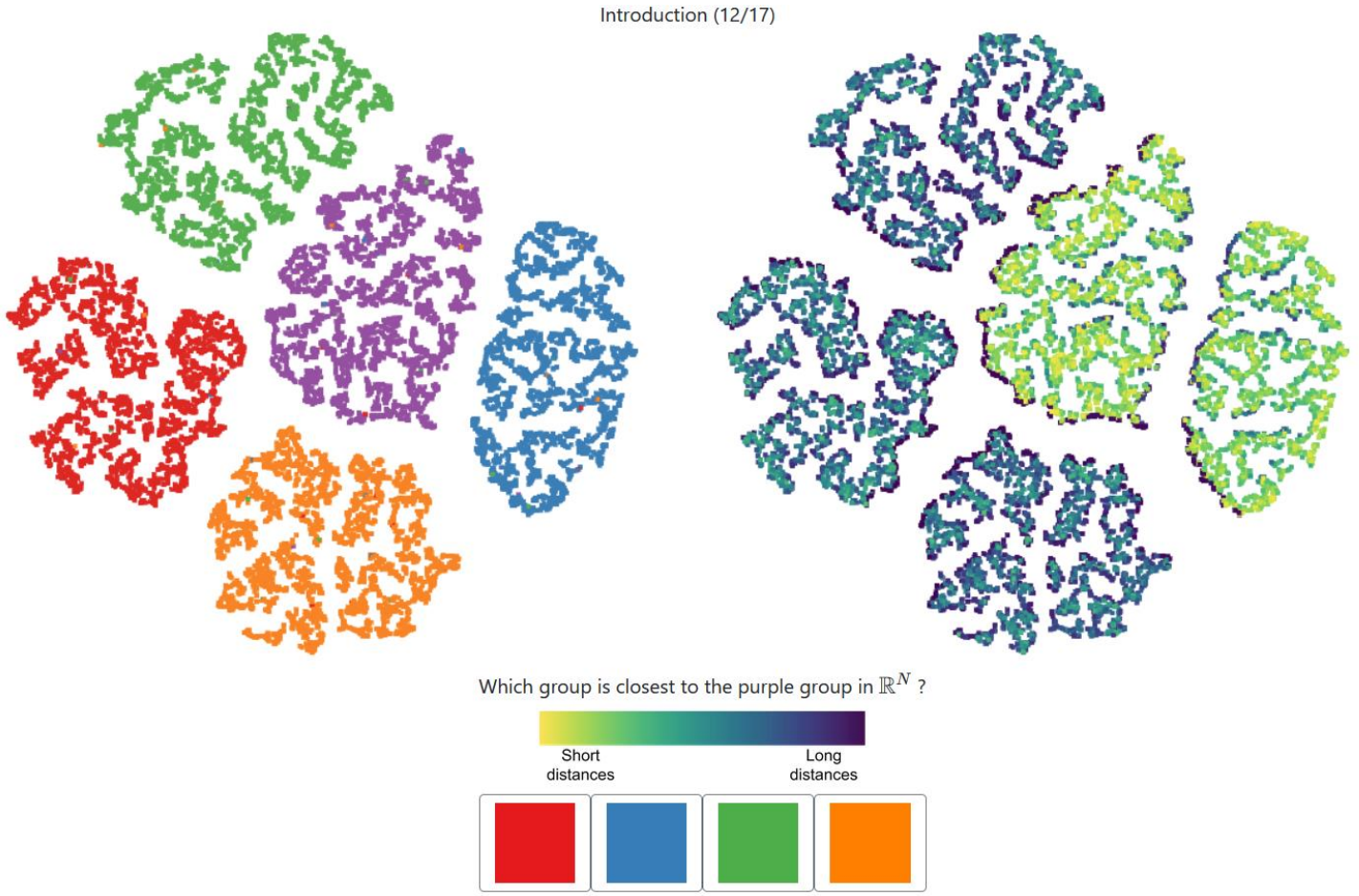
User tasks

Find largest group

Find closest group

Find number of outliers

Find the type of topogical structure



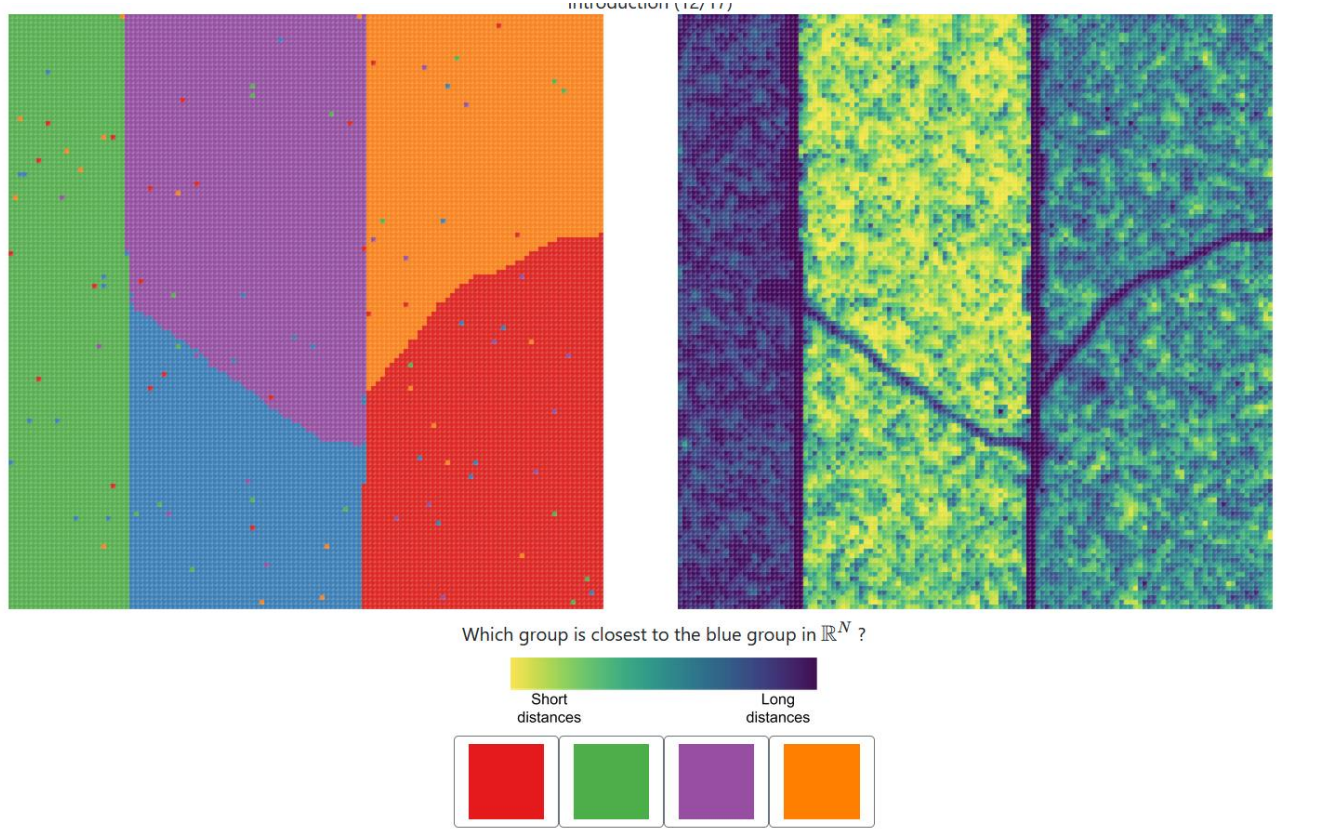
User tasks

Find largest group

Find closest group

Find number of outliers

Find the type of topogical structure



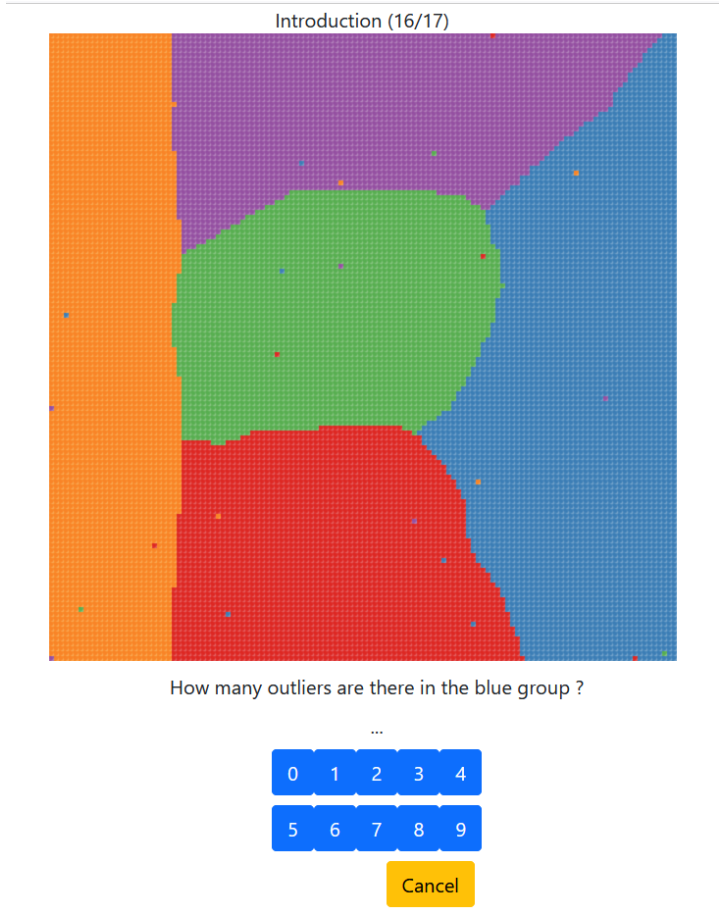
User tasks

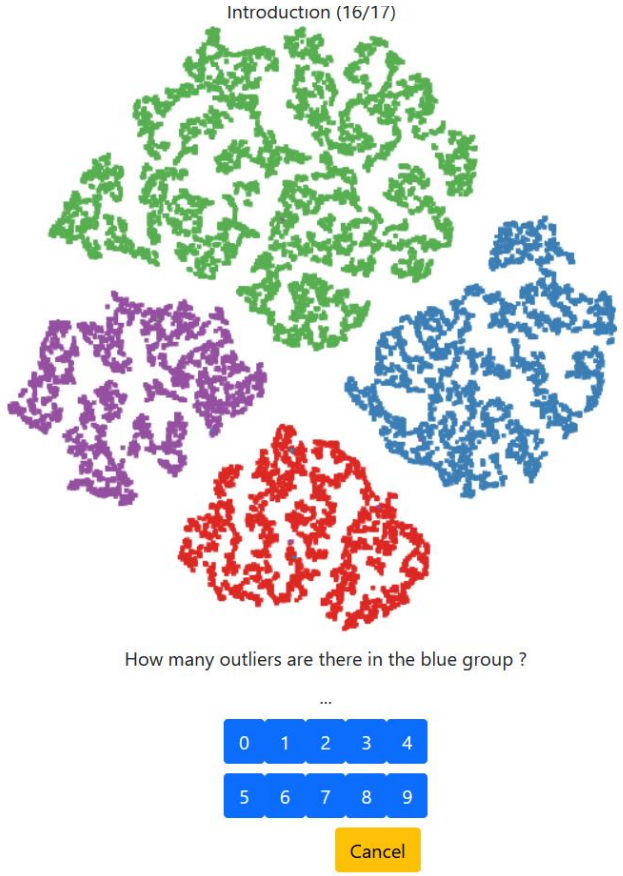
Find largest group

Find closest group

Find number of outliers

Find the type of topogical structure





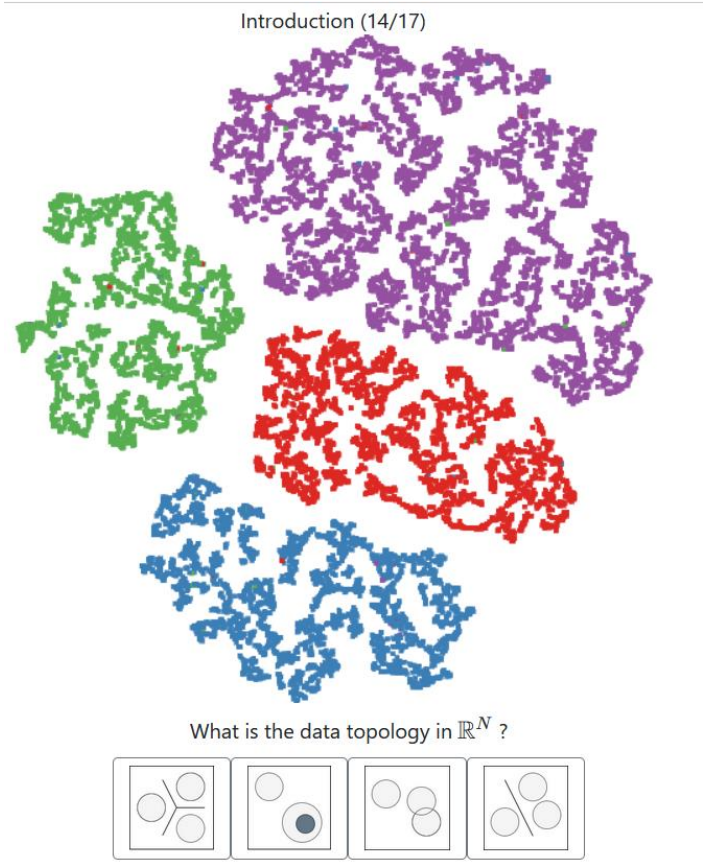
User tasks

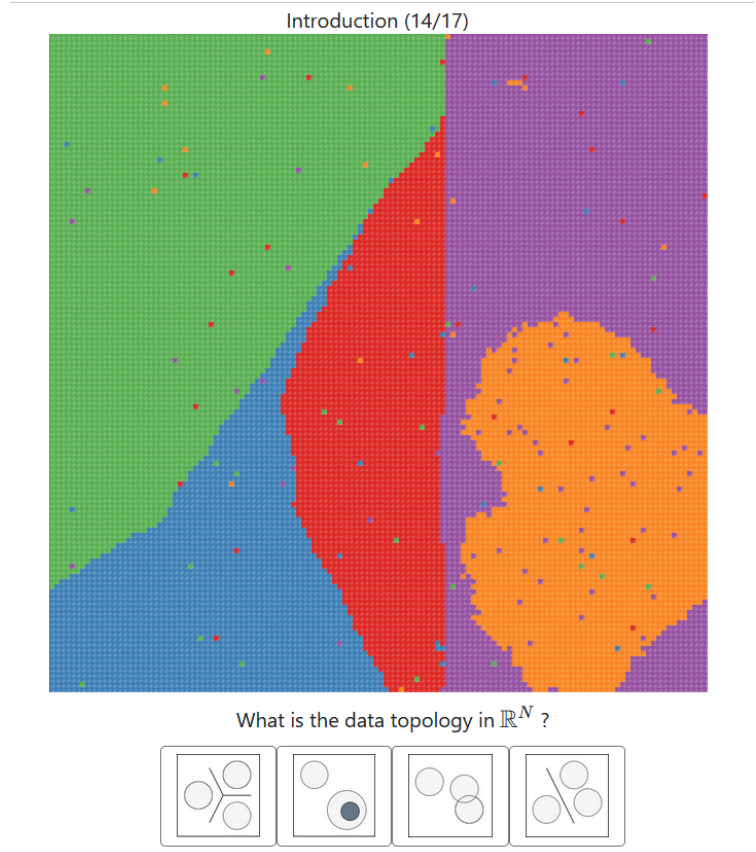
Find largest group

Find closest group

Find number of outliers

Find the type of topogical structure





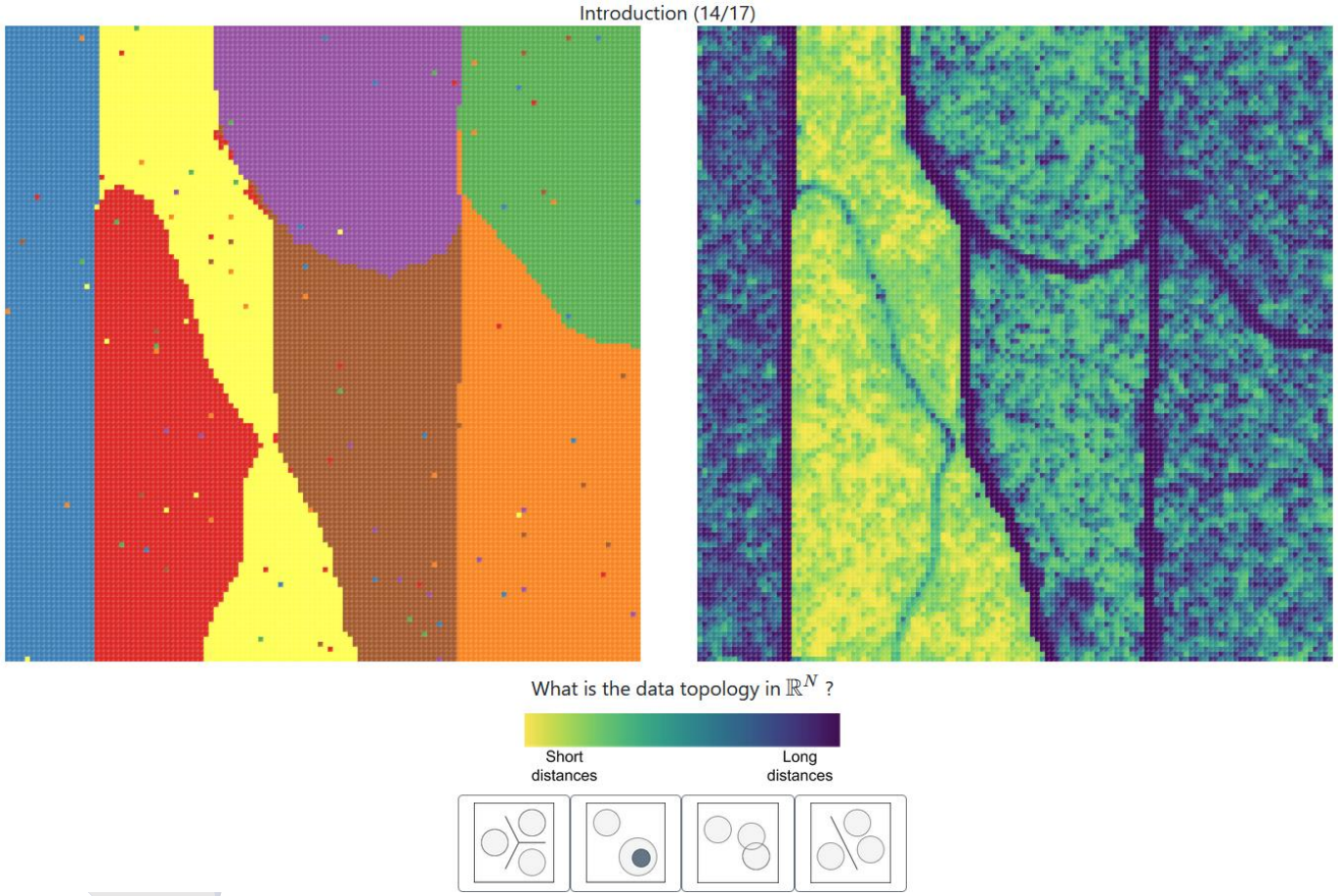
User tasks

Find largest group

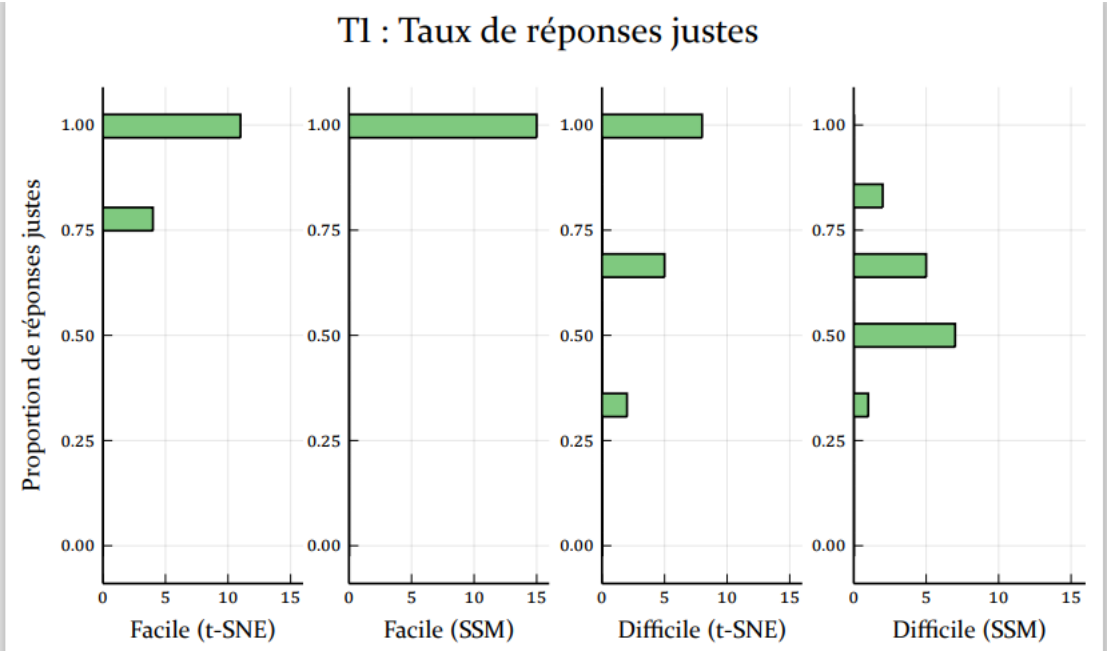
Find closest group

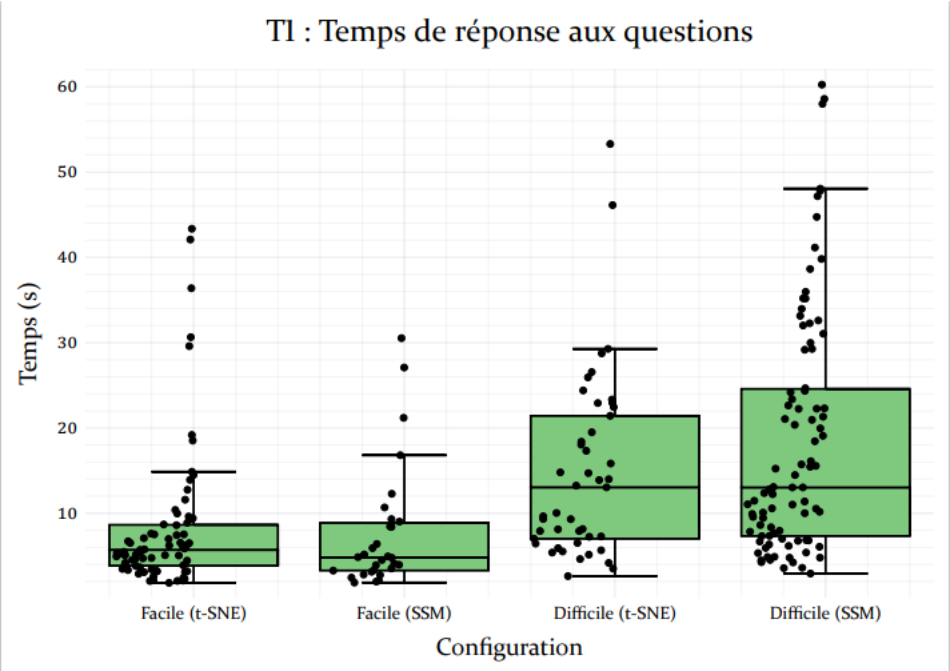
Find number of outliers

Find the type of topogical structure

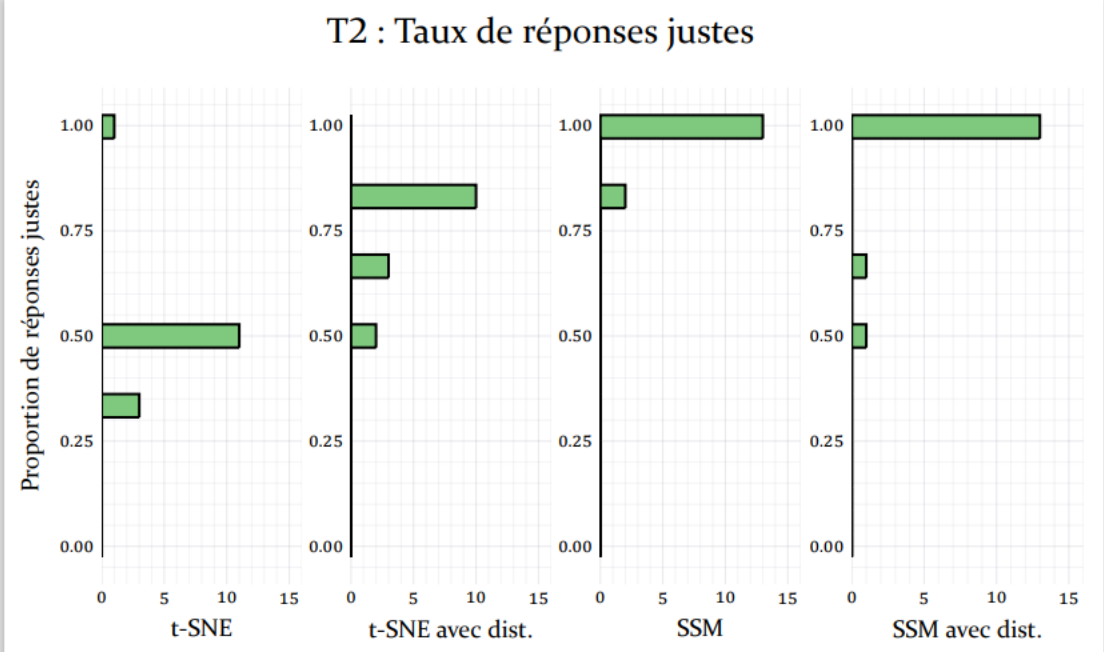


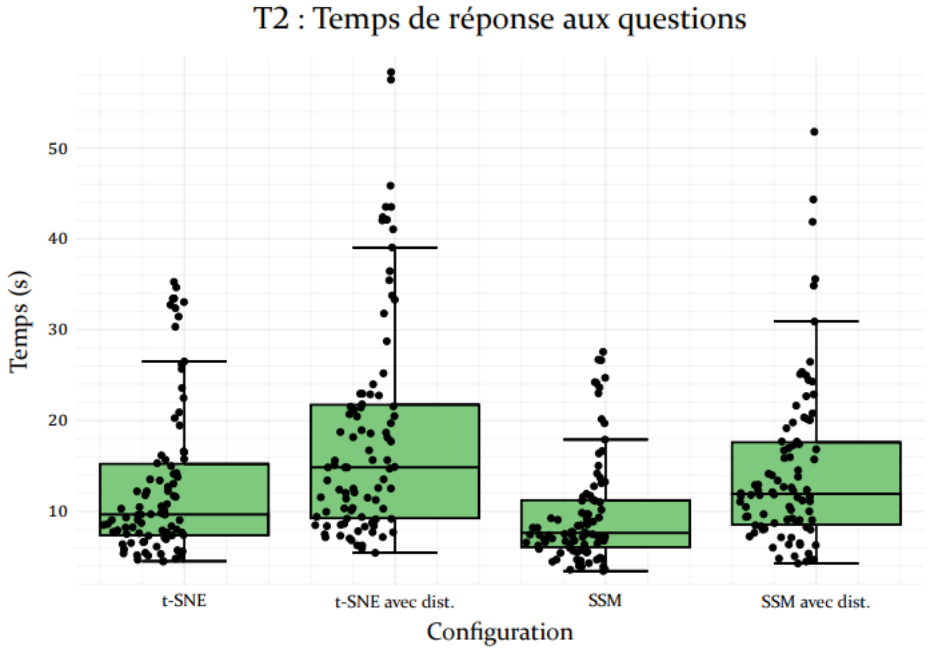
Find larger group



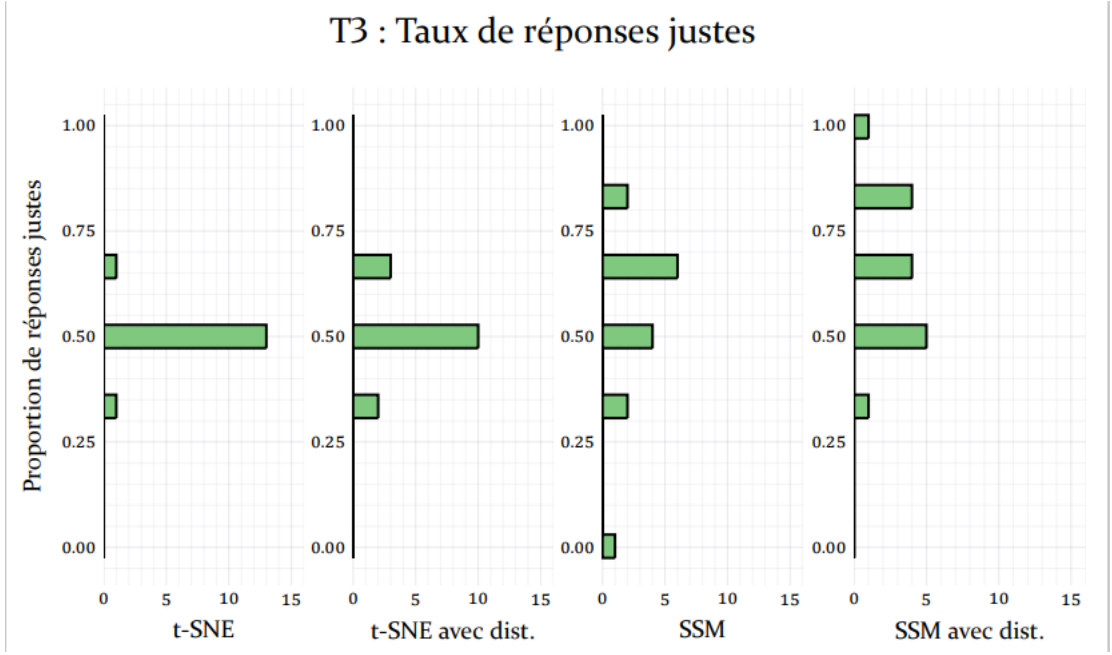


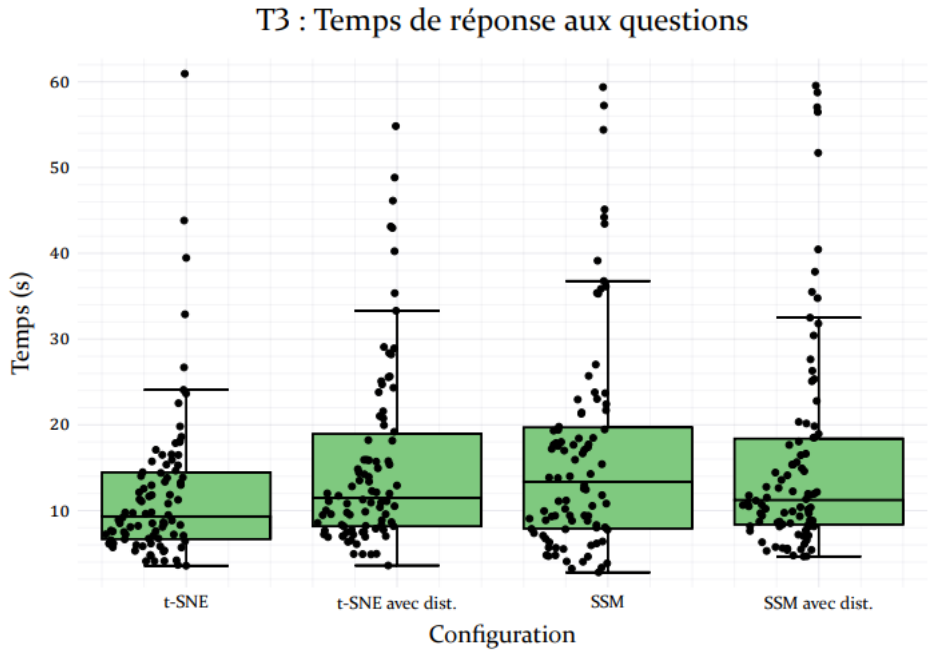
Find closest group



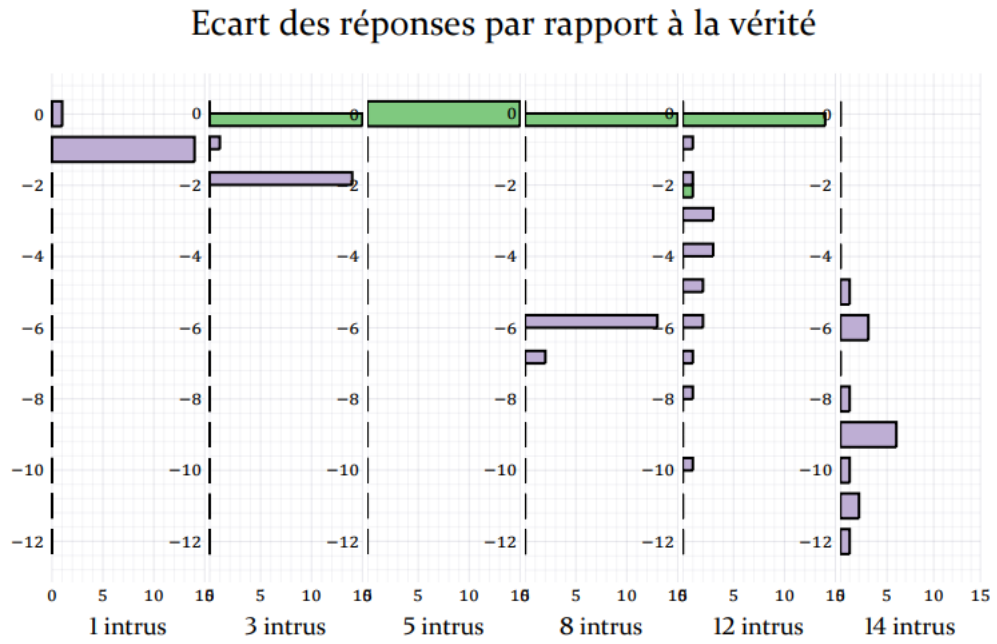


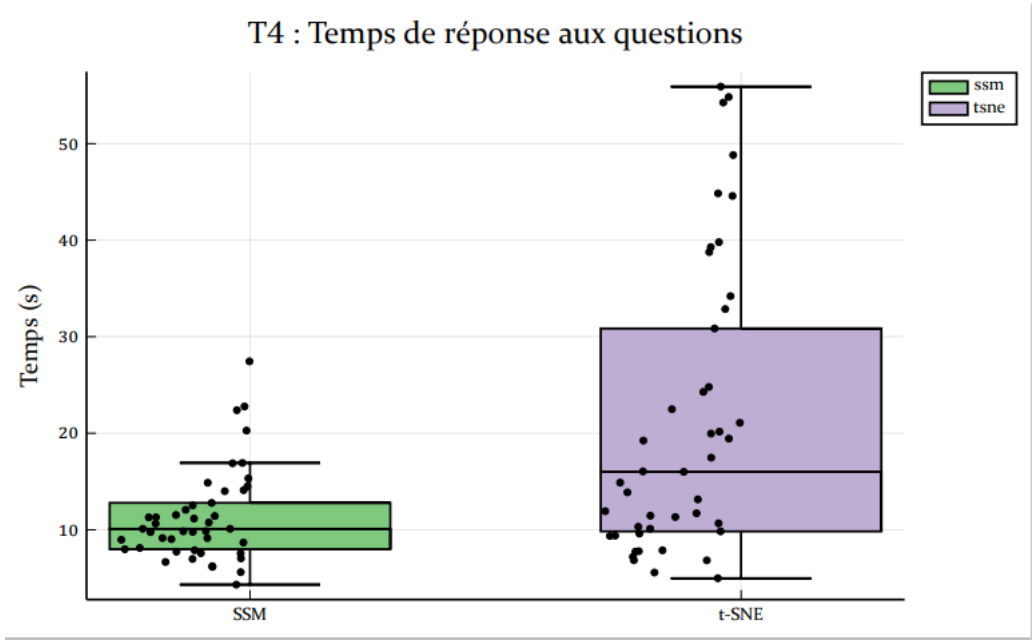
Find the type of topogical structure

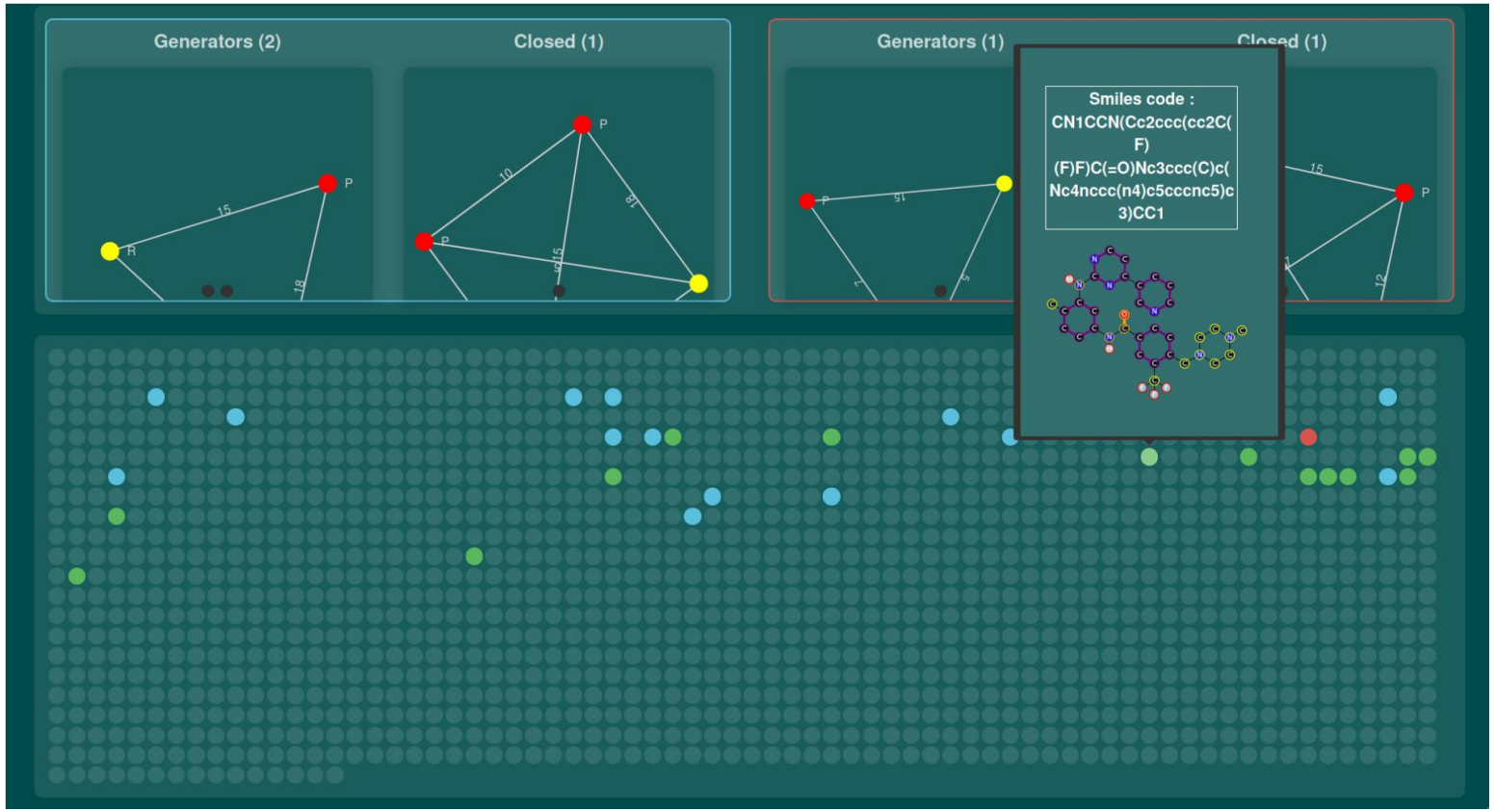




Find number of outliers





Future work InvolvD

- Intégrer la méthode dans notre prototype
- Evaluer l'impact