Neighborhood-based analysis of self-organizing maps

Tyler J Burns, Ph.D.
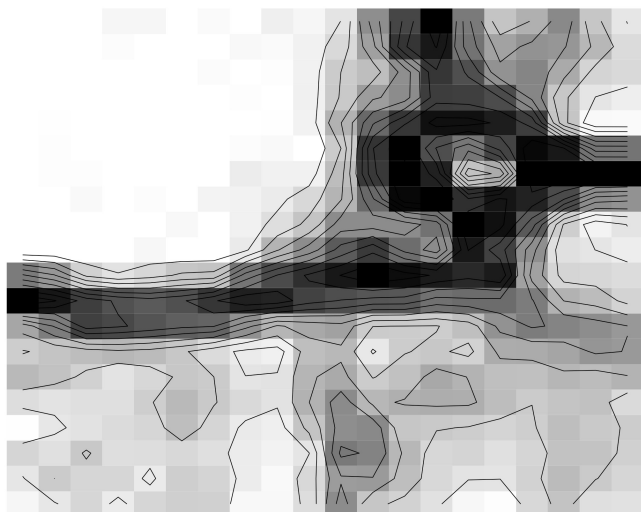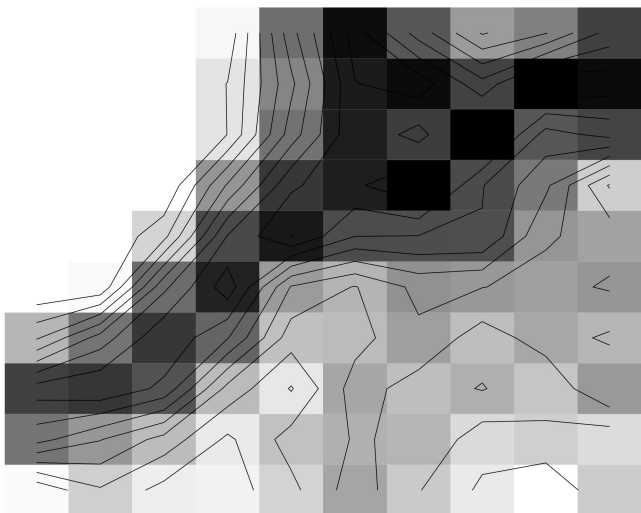
AG Mei at German Rheumatism Research Center

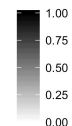# Exploring the visual capabilities of large self-organizing maps

U-Matrix visualizations reveal dataset complexity in a manner much different than t-SNE
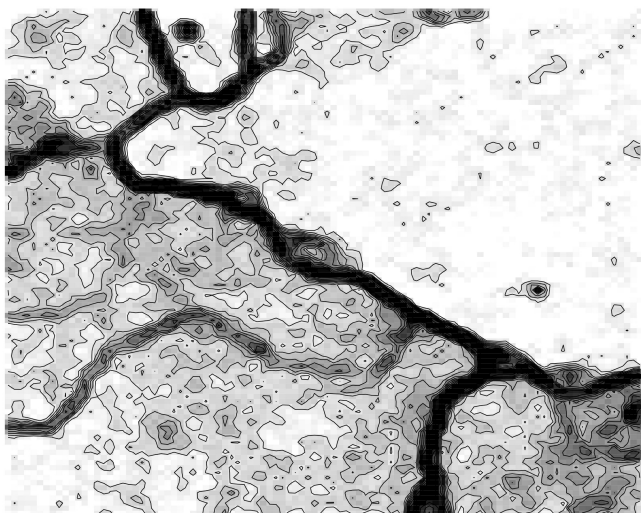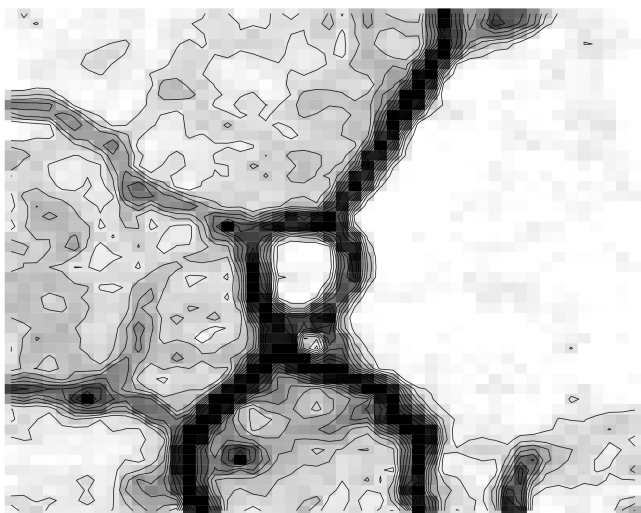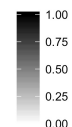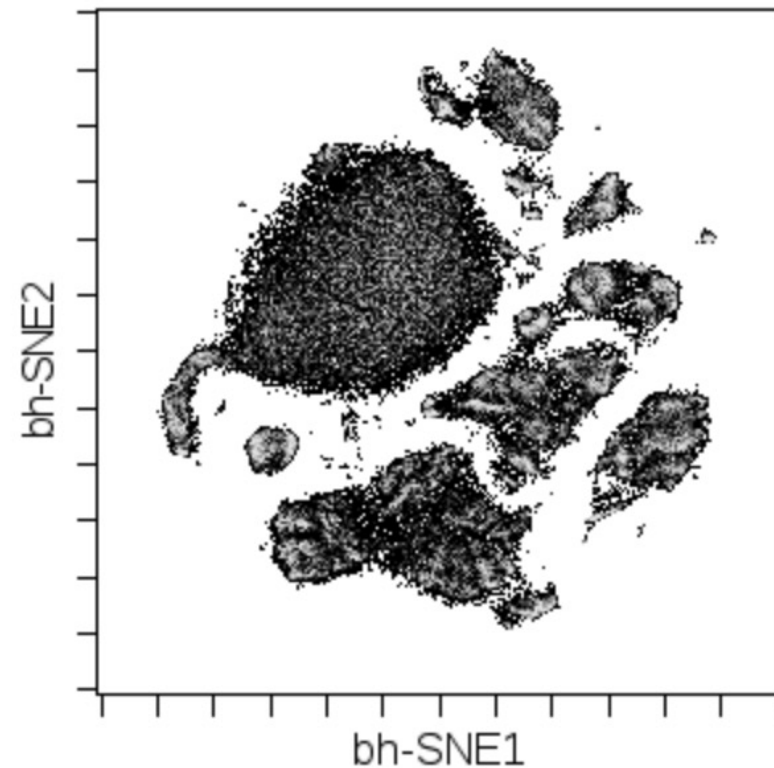

10 x 10 (Flow-SOM default)

20 x 20

50 x 50

100 x 100

t-SNE map

bh-SNE2

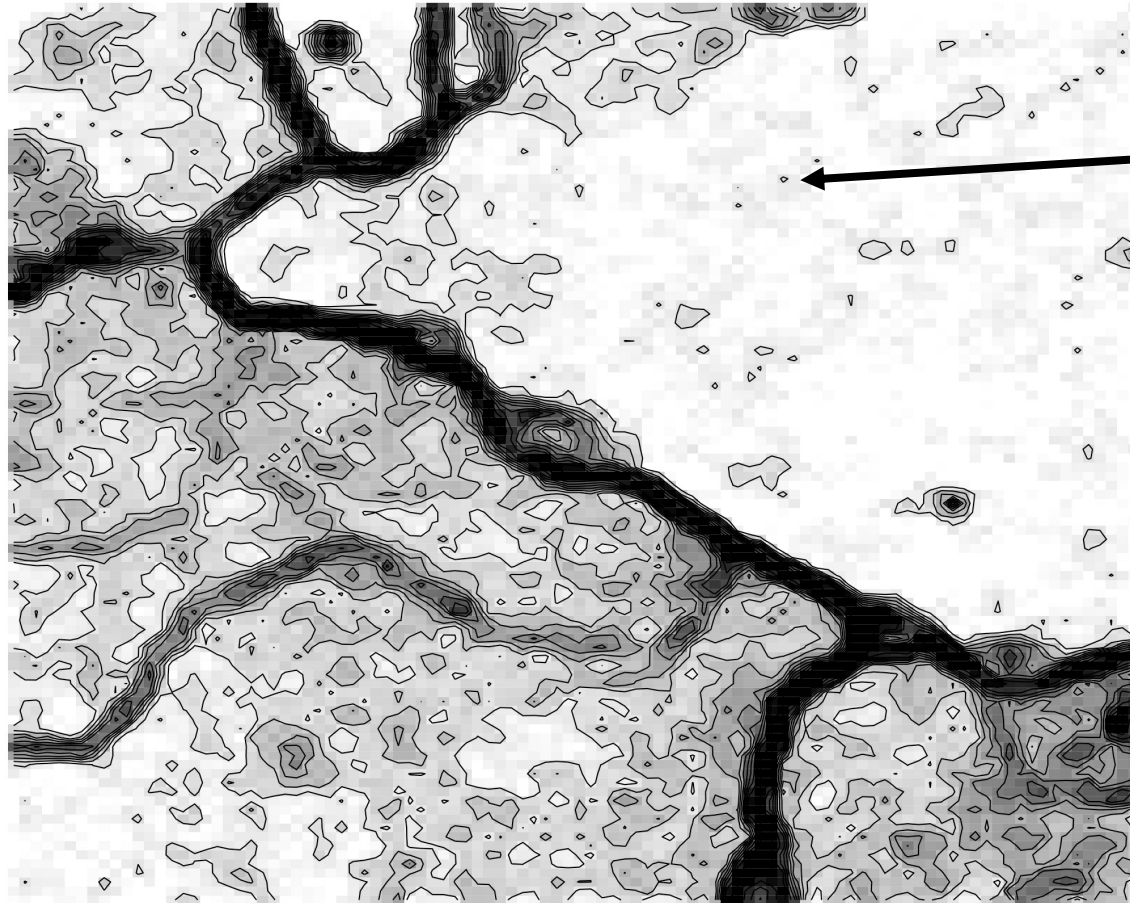bh-SNE1

Data type: Mass cytometry
Dims: 42, surface markers
Cell type: PBMCs
Cells: 100,000

Hypothesis: there are regions of the self-organizing map that are not populated
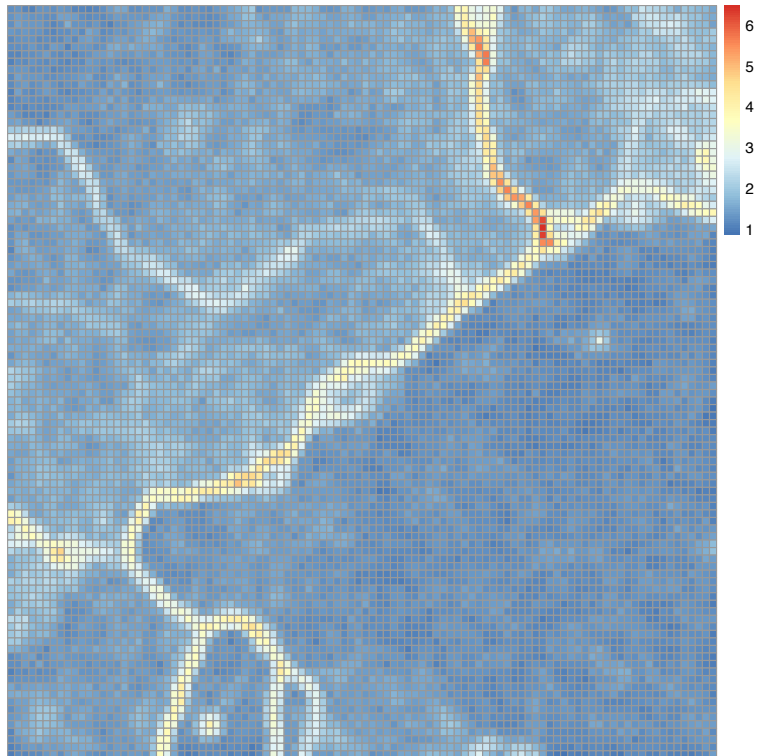

???????

Method:

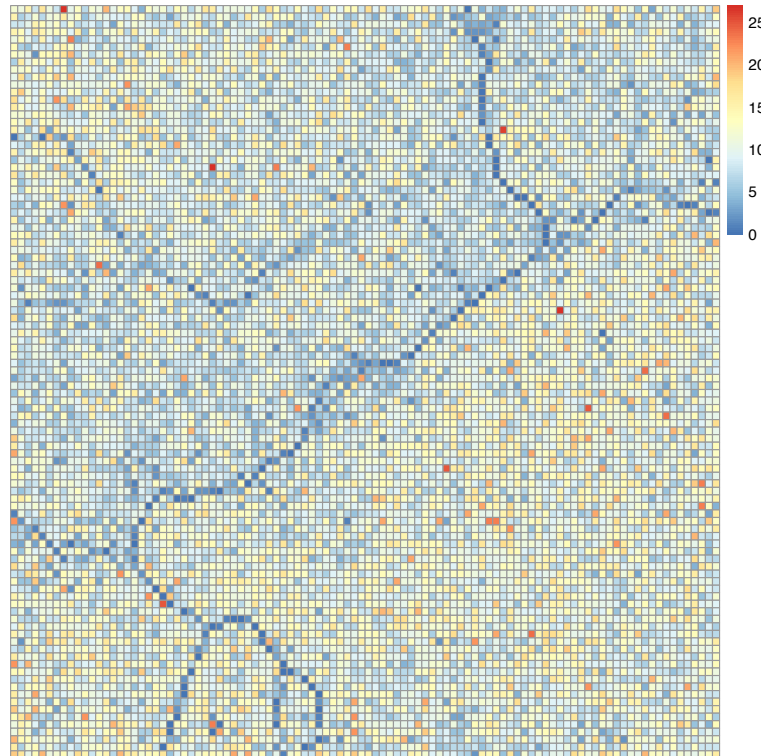Color the SOM by number of cells per node (average should be 10).

Threshold by 0 and greater than 0 cells

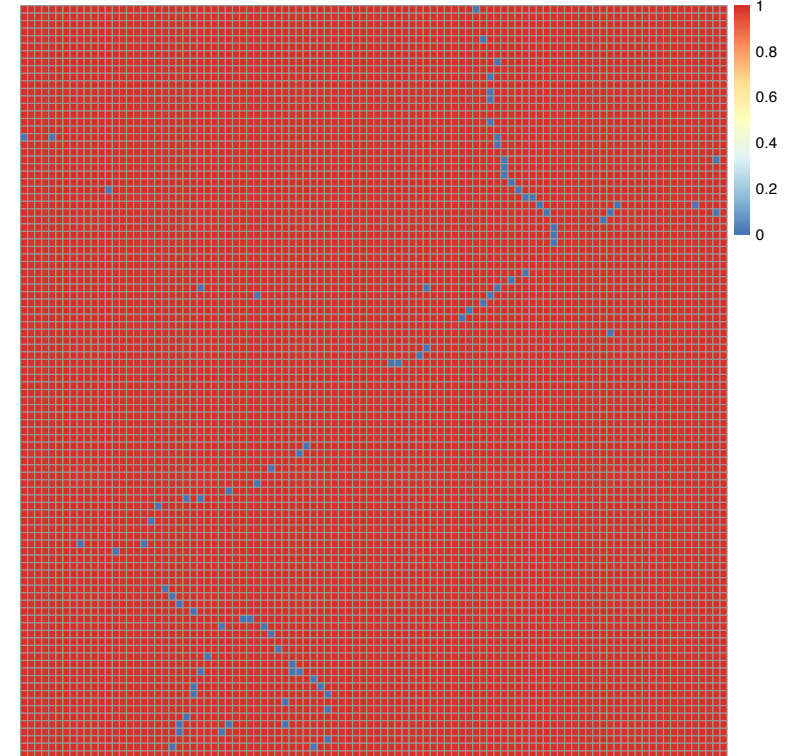# The majority of the SOM is populated with at least one cell
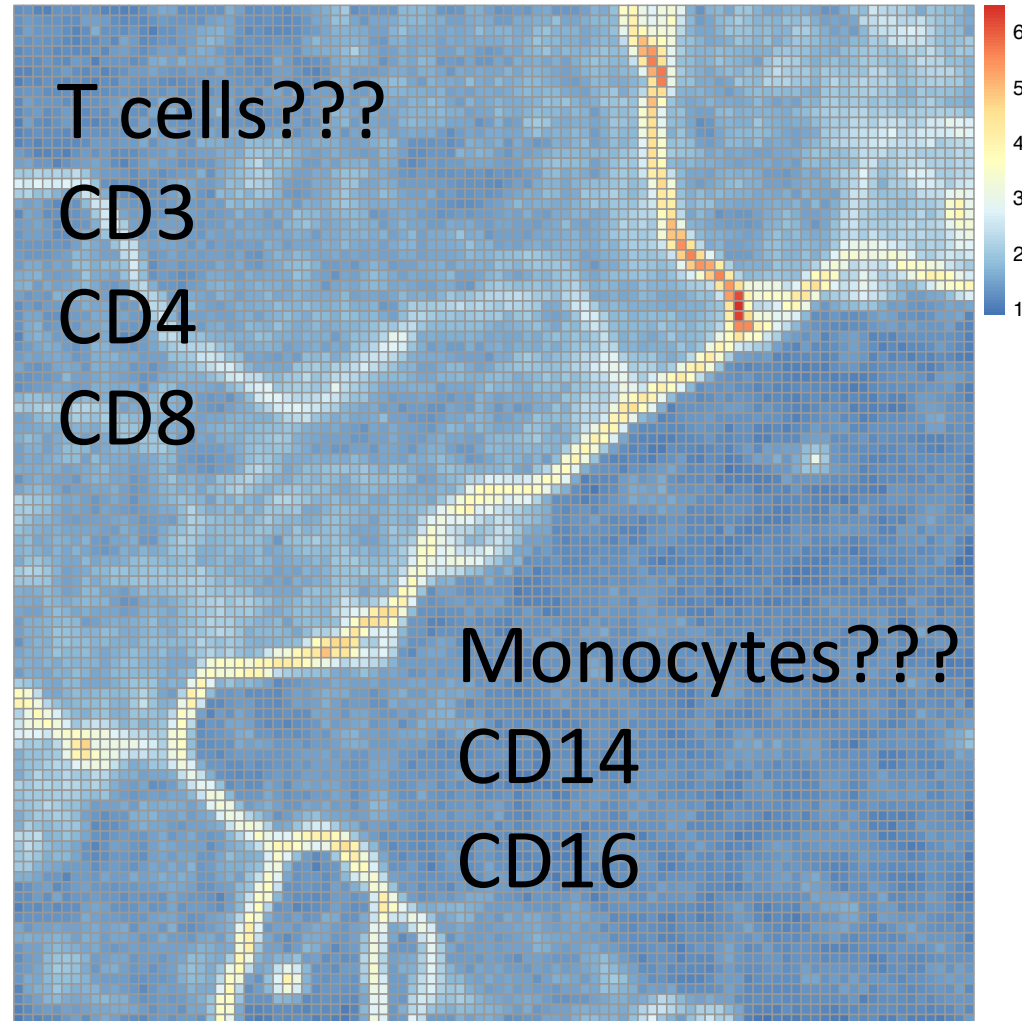
U-Matrix

Abundance

0-1 Thresholded Abundance

SOM size: 100 x 100

Data type: Mass cytometry
Dims: 42, surface markers
Cell type: PBMCs
Cells: 100,000

Hypothesis: distinct regions of the SOM are defined by distinct markers
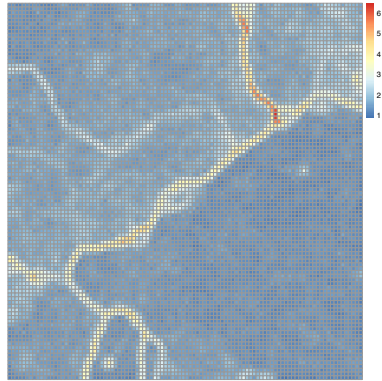

T cells???

CD3

CD4

CD8

Monocytes???

CD14

CD16

Method:

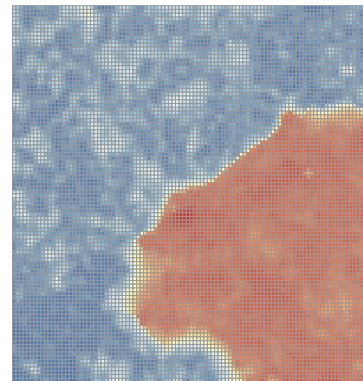1) Color the SOM by the marker expression levels at each node

2) Compare these types of patterns to that of a t-SNE map, the "best practices" visualization scheme for CyTOF data.

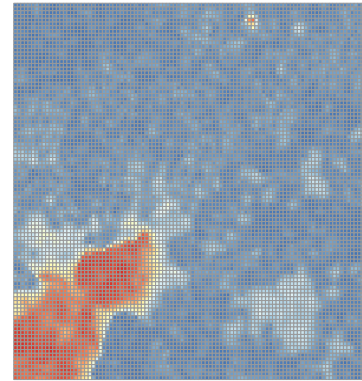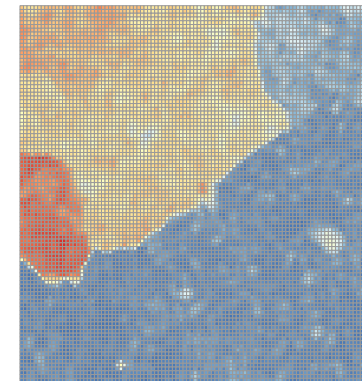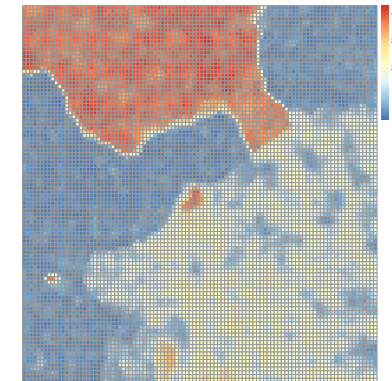# Marker expression patterns are concordant with the U-Matrix
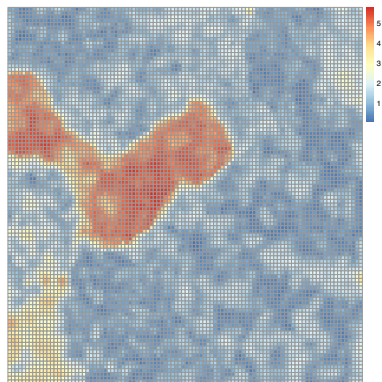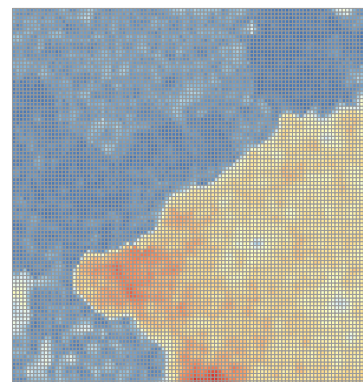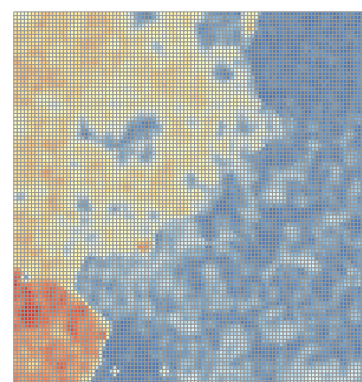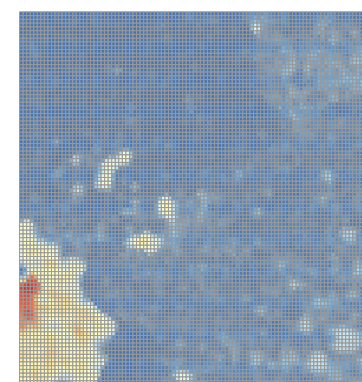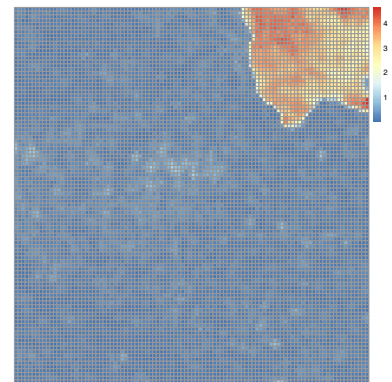
SOM vs t-SNE: colors of marker expression
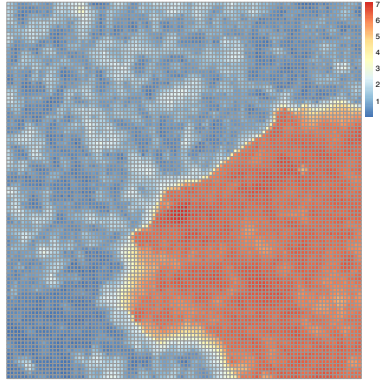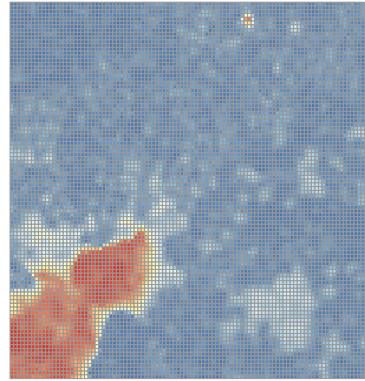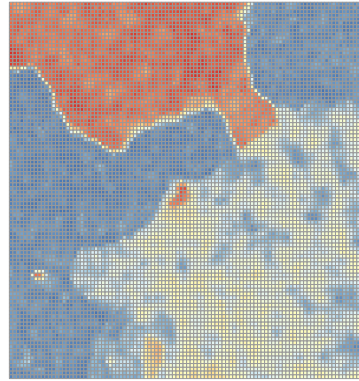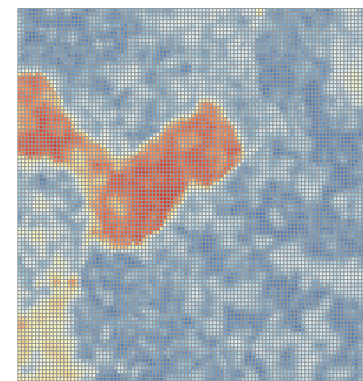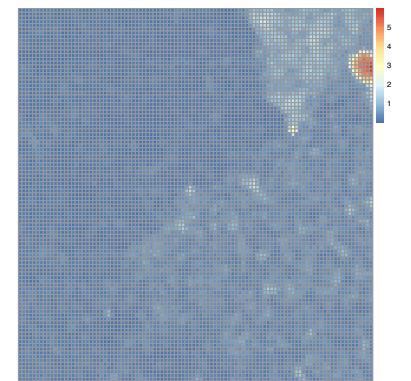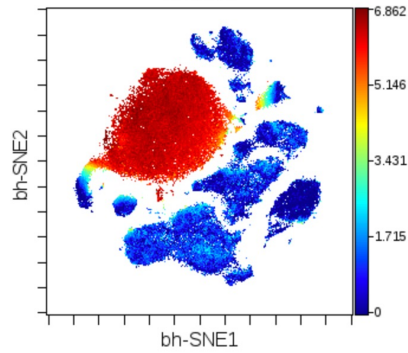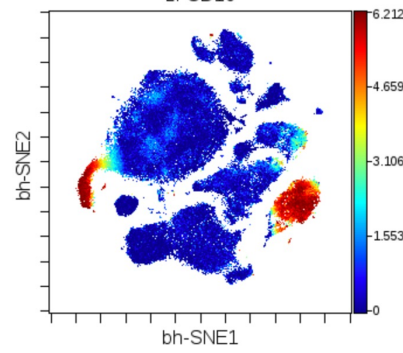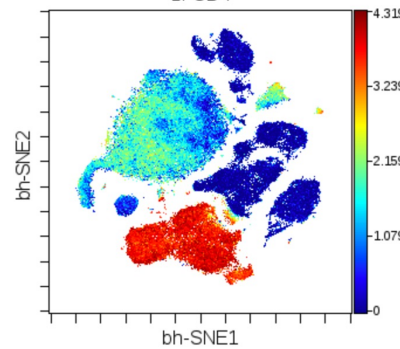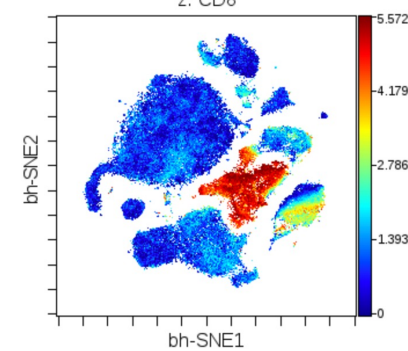
Hypothesis: U-matrix patterns could be brought out by analyzing Chebyshev distance between nodes
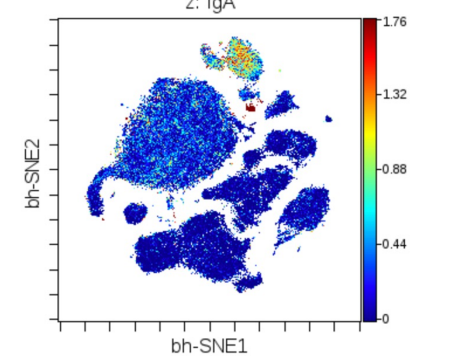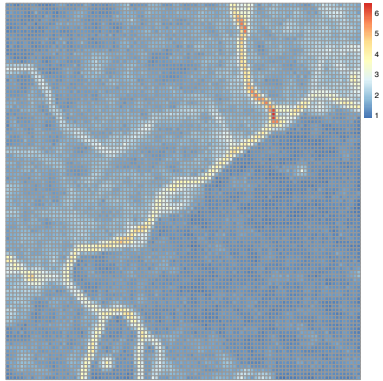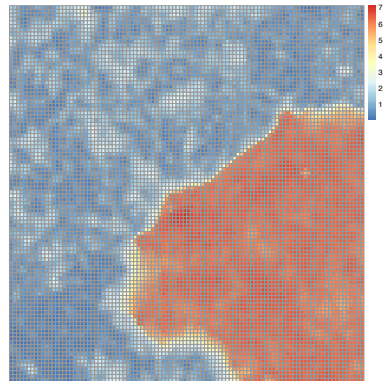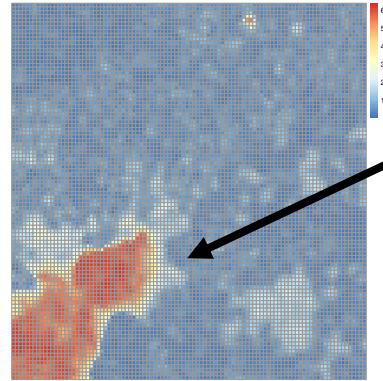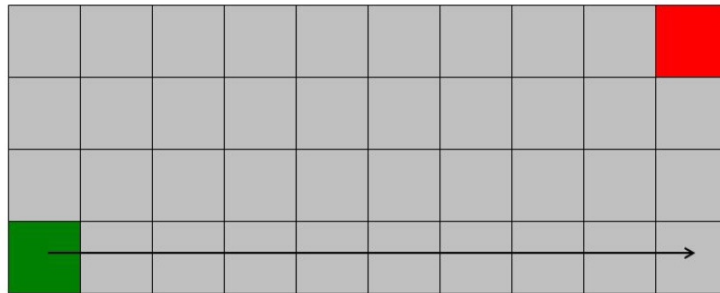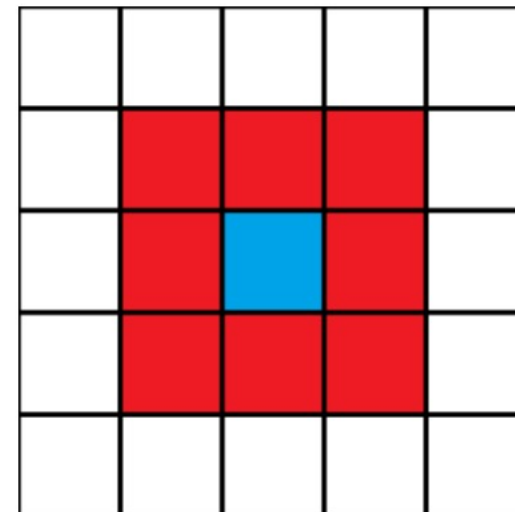
U-matrix

CD14

CD16

I don't see this boundary in the U-matrix!!

Method 1:
Calculate U-matrix with Chebyshev distance

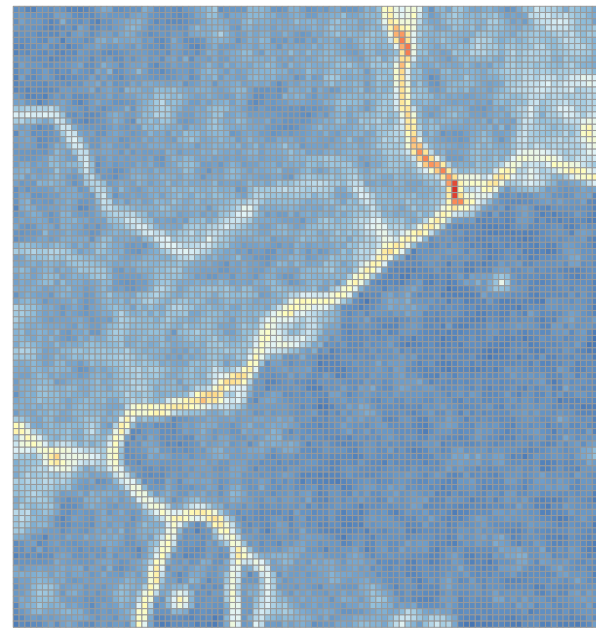- Chebyshev distance

max(xDiff, yDiff) = max(9,3) = 9

Method 2:
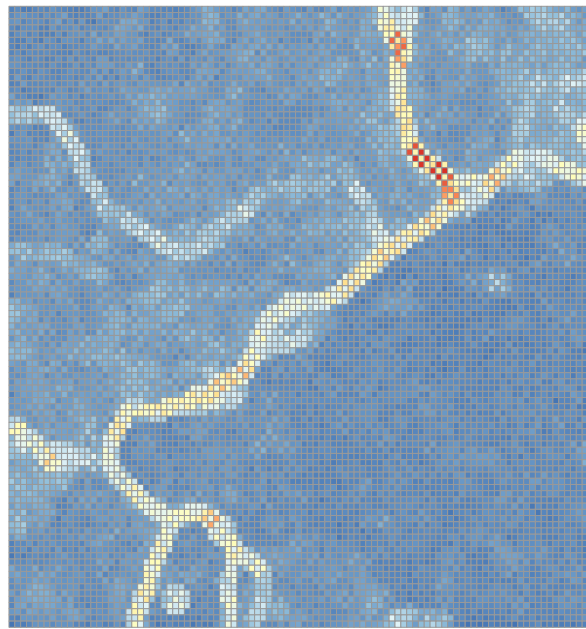Use the max not the mean of all distances in the Moore neighborhood

U-matrix visualization with alternate distance and neighbor aggregation functions
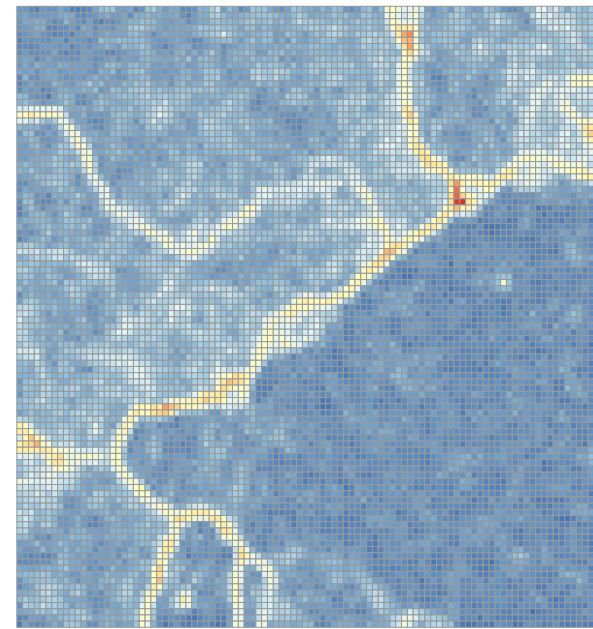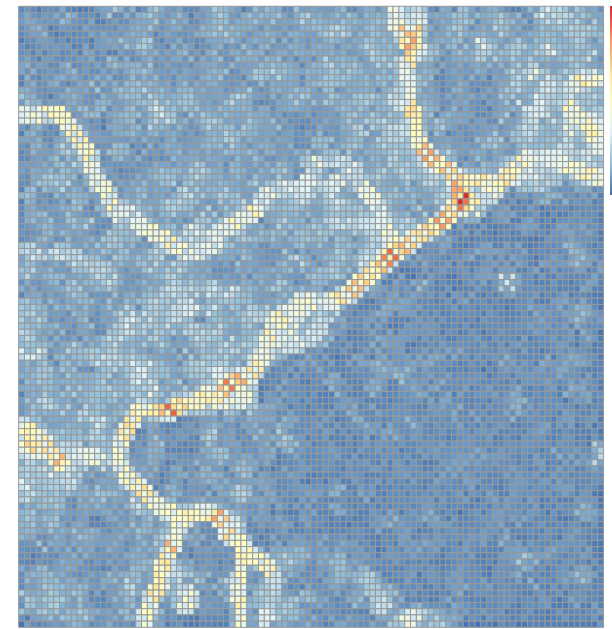

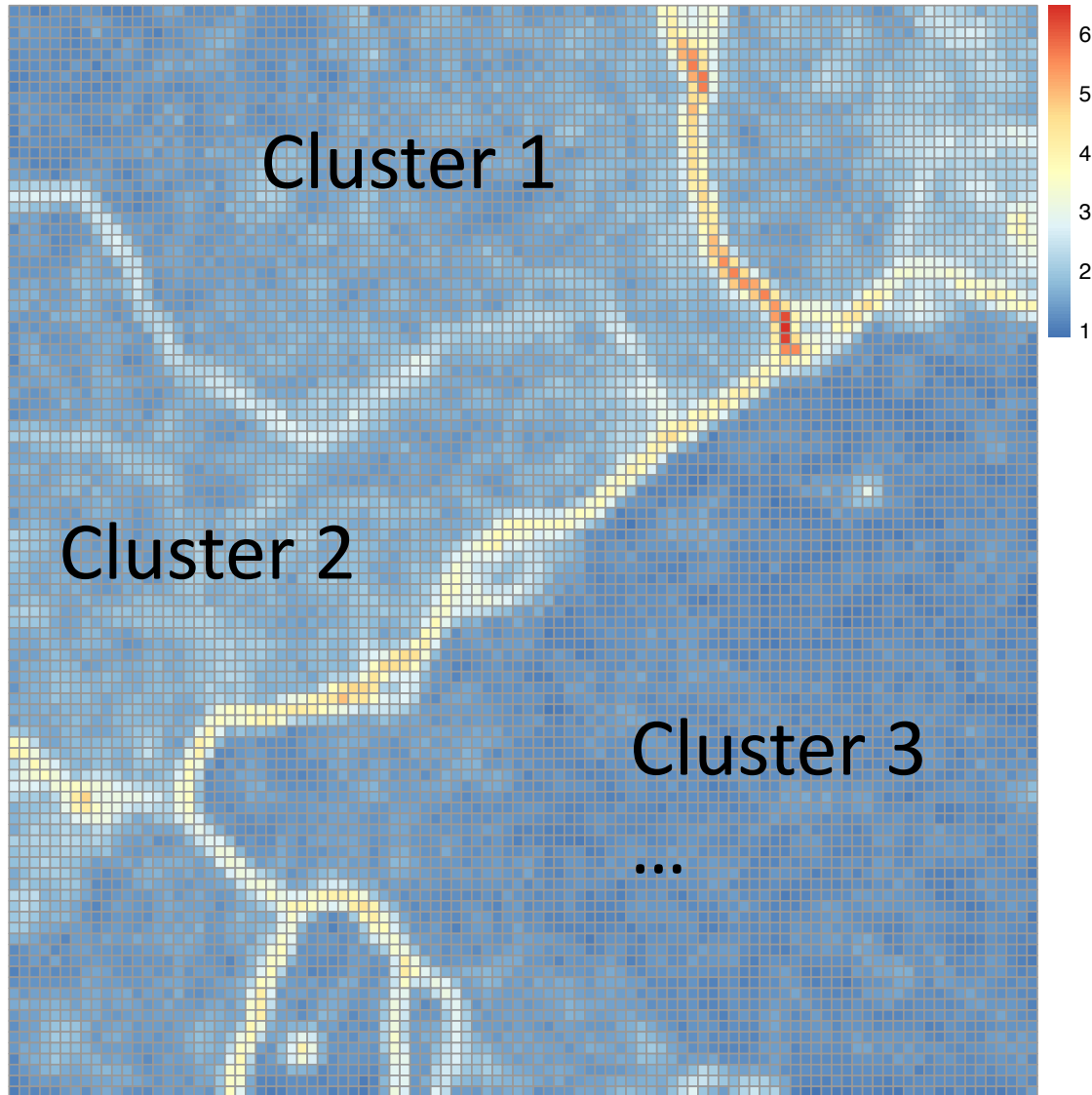Euclidean mean   Euclidean max   Chebyshev mean   Chebyshev max

Hypothesis: Meta-clustering will draw boundaries in concordance with the U-matrix
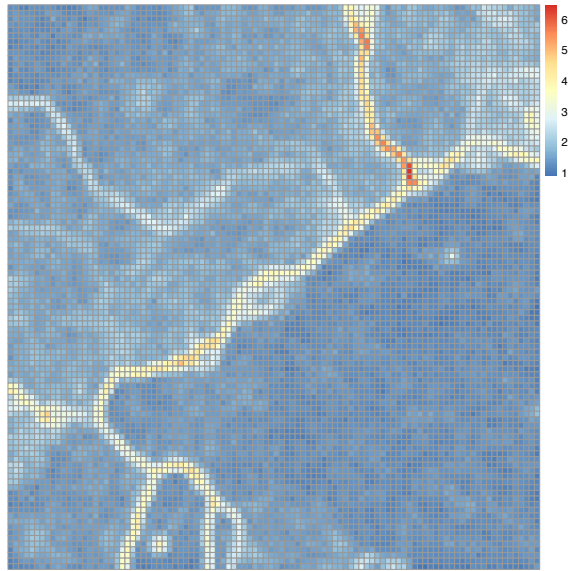

Cluster 1

Cluster 2

Cluster 3

...

Method:

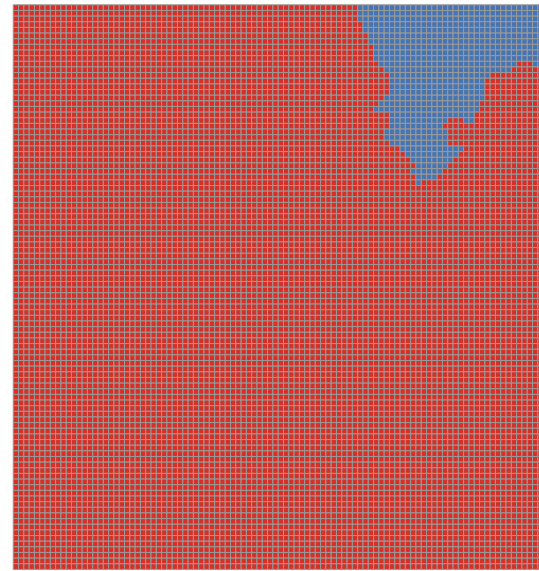1) Perform hierarchical clustering on the SOM nodes based on their marker expression

2) Visualize the meta-clusters on the map itself to compare with the U-matrix
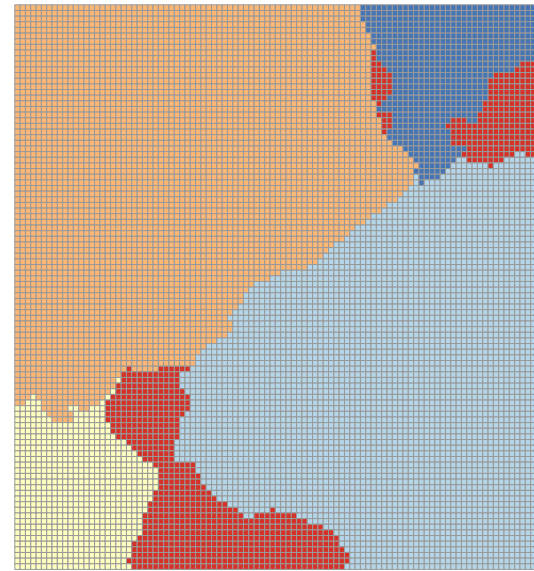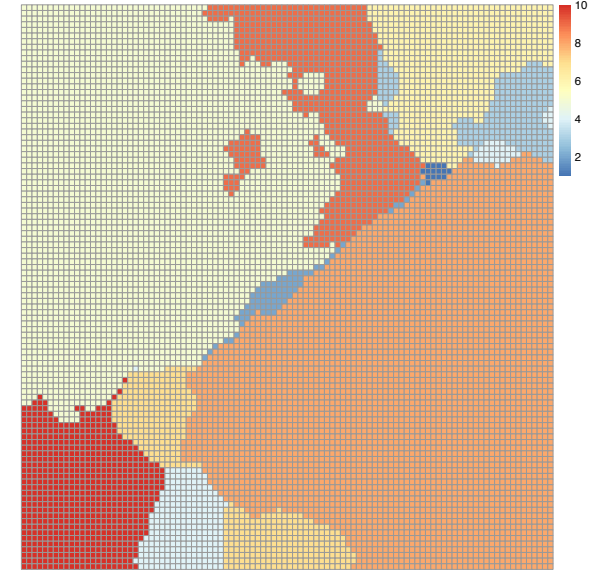
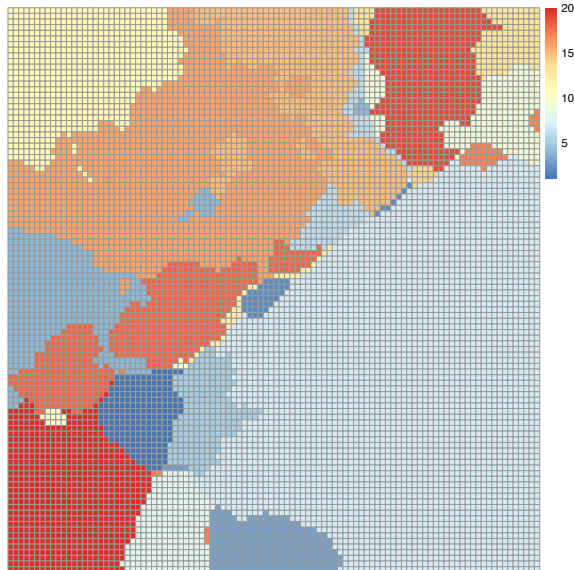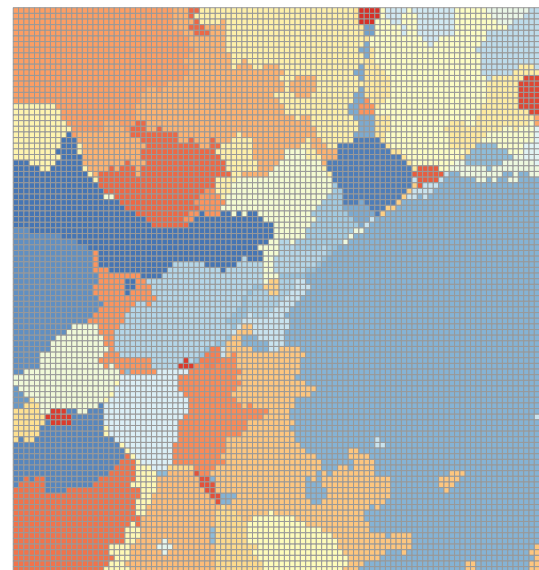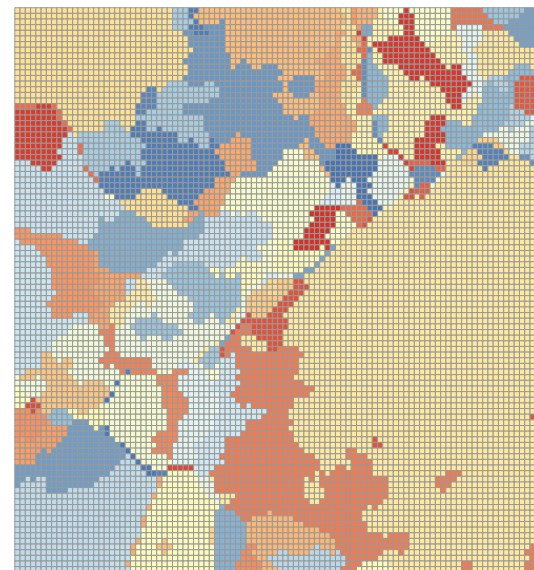# Color of U-Matrix by identity of the meta-clusters

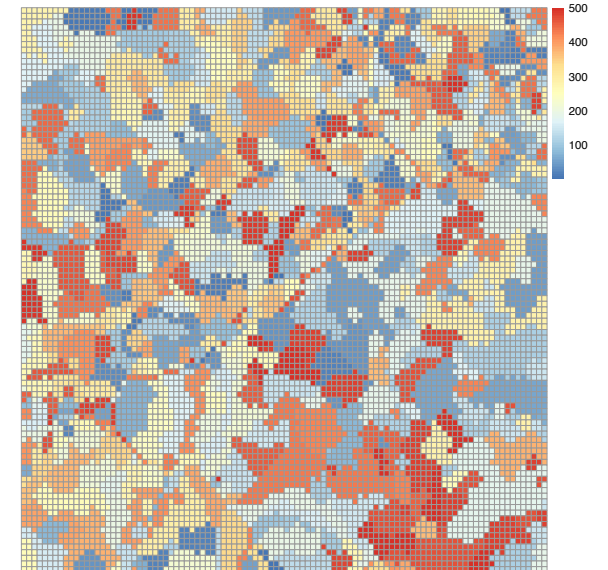U-matrix   2 clusters   5 clusters   10 clusters
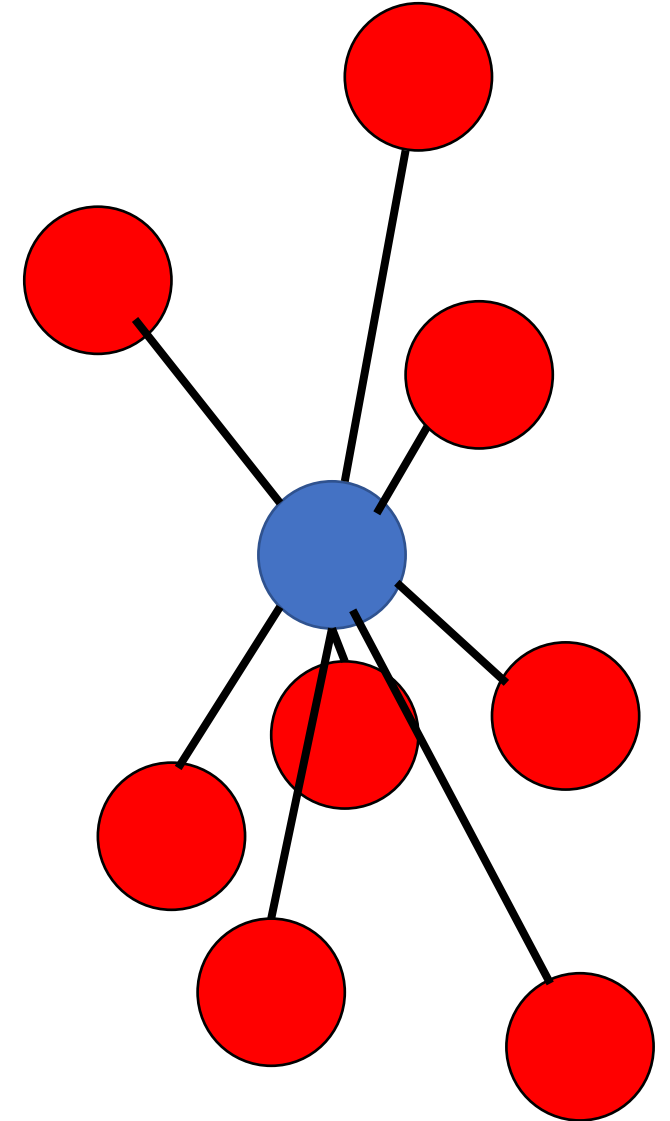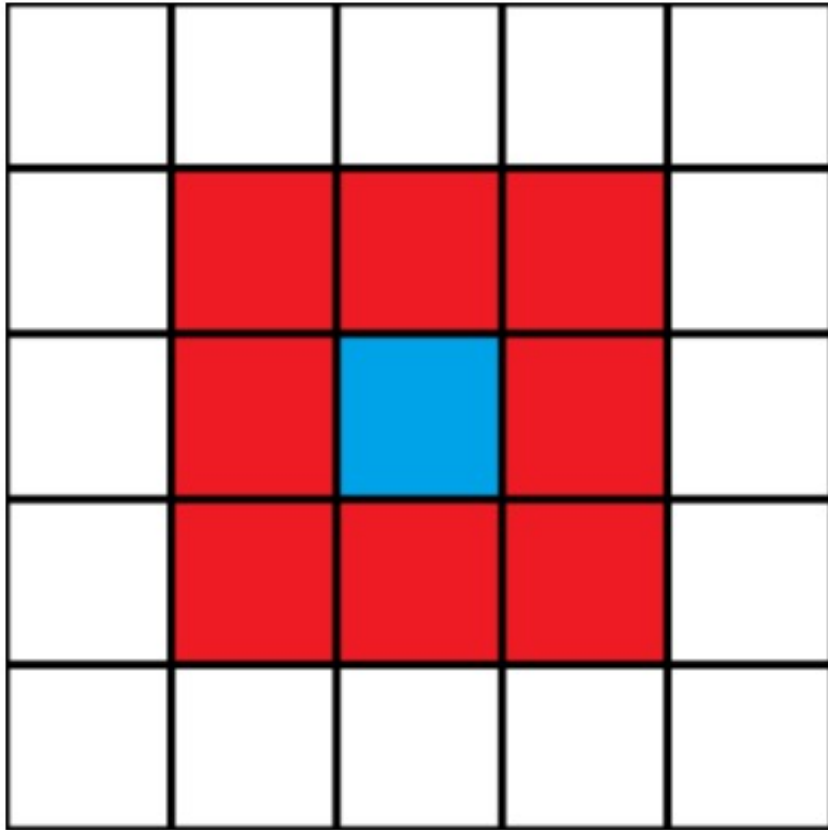
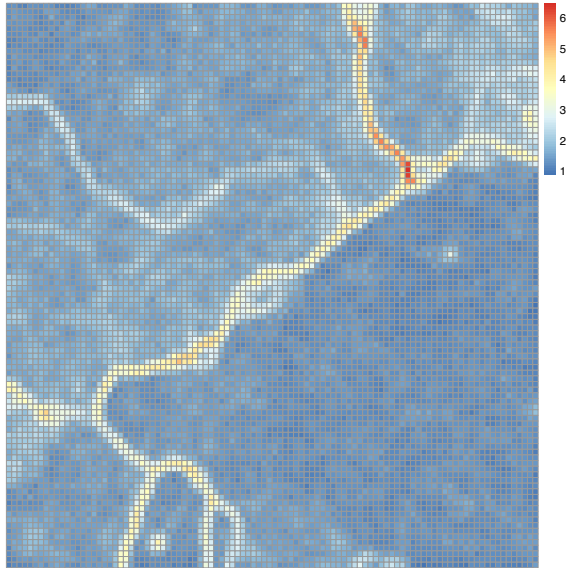20 clusters   50 clusters   100 clusters   500 clusters

Hypothesis: The Moore neighborhood distance averages provides similar information to the KNN distance averages
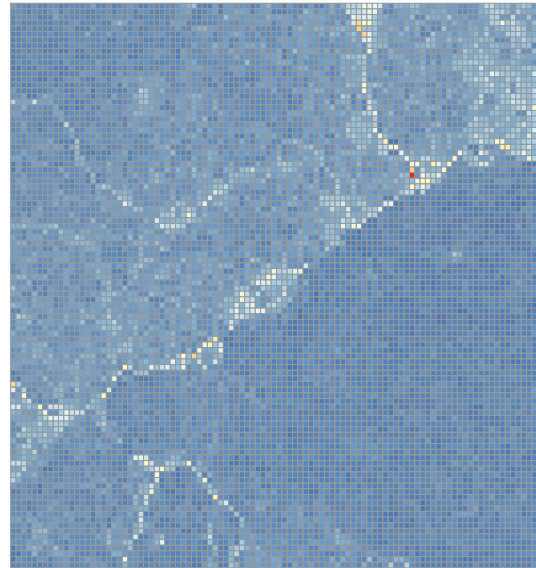
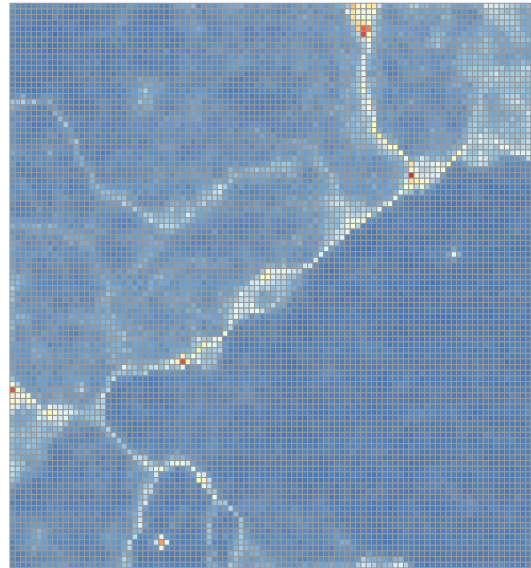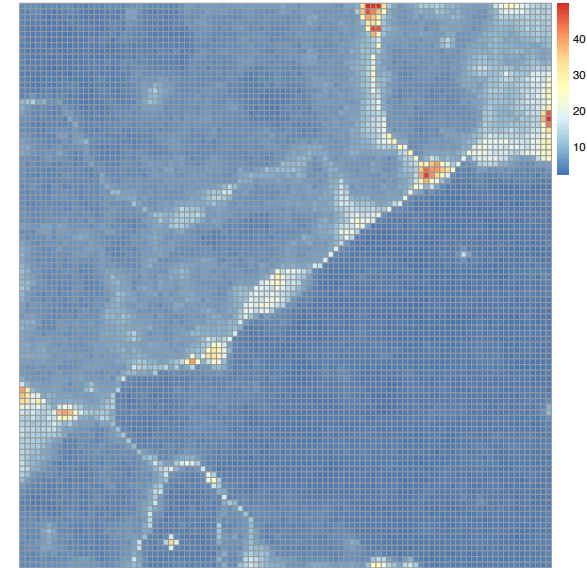# U-Matrix versus KNN-density estimation for SOM
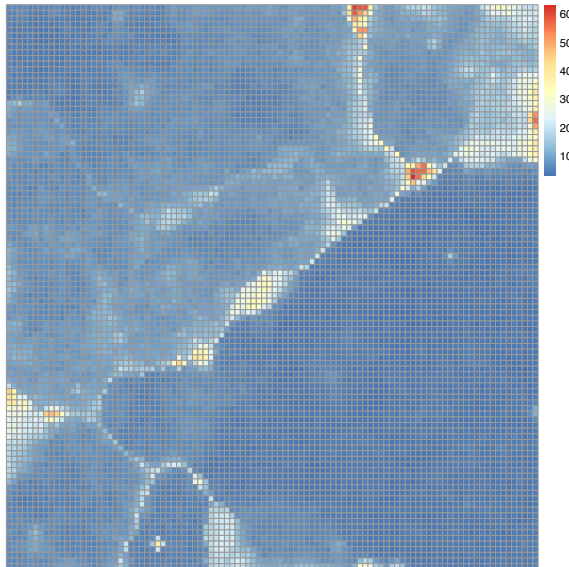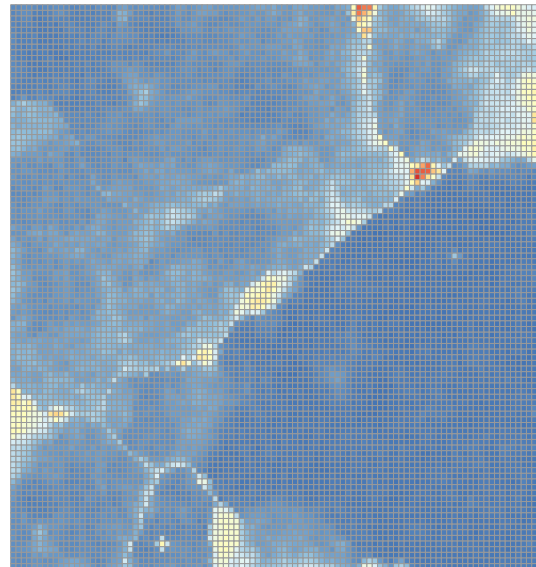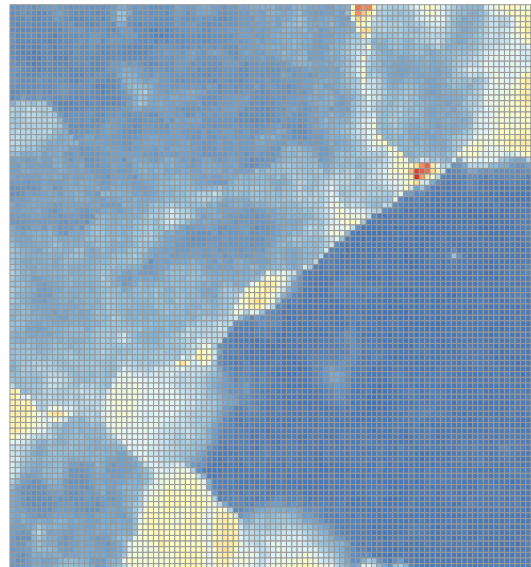
Hypothesis: U-matrix positioning fidelity more closely approximates the manifold locally than t-SNE positioning

Method:

1) Perform KNN with K = 8 on marker expression per-node
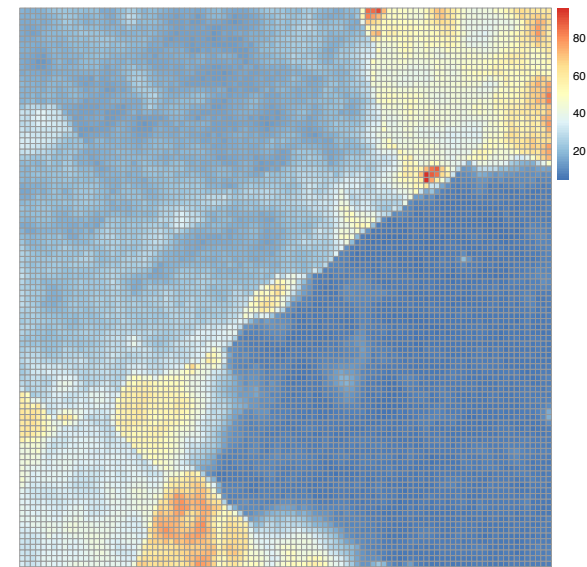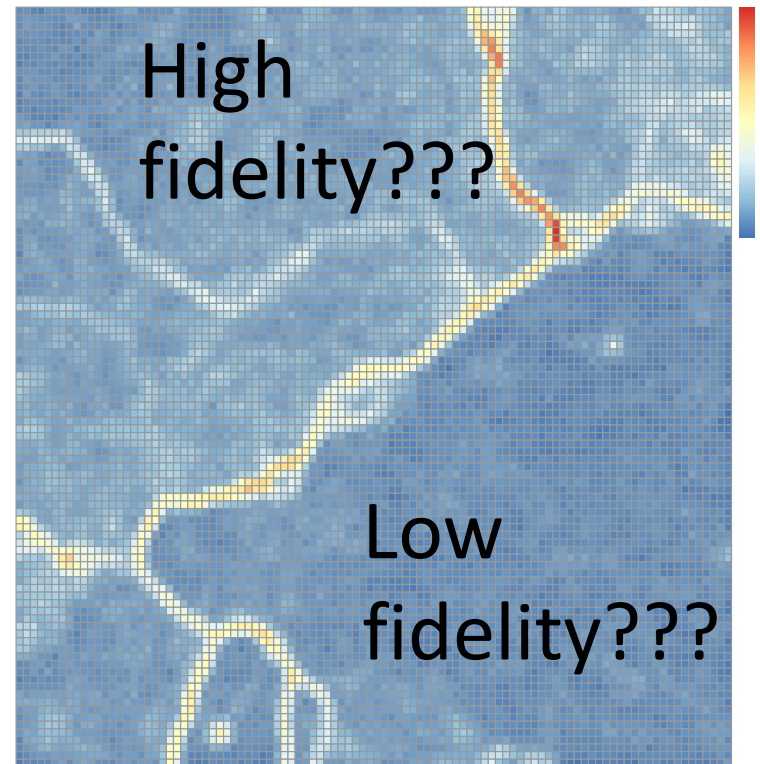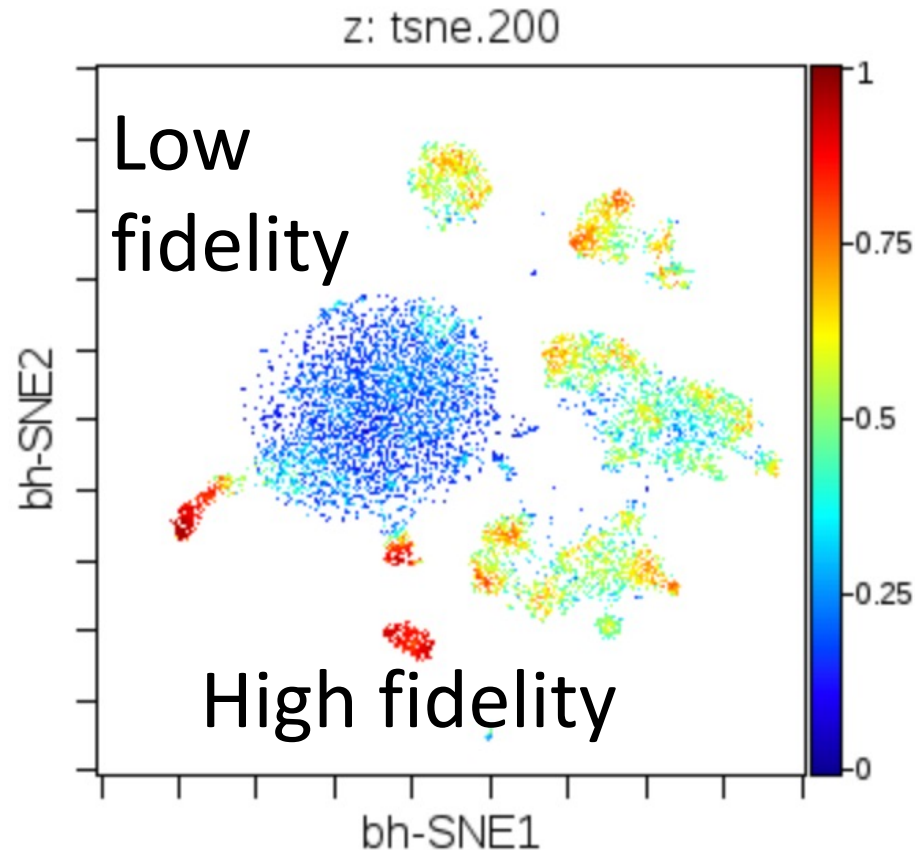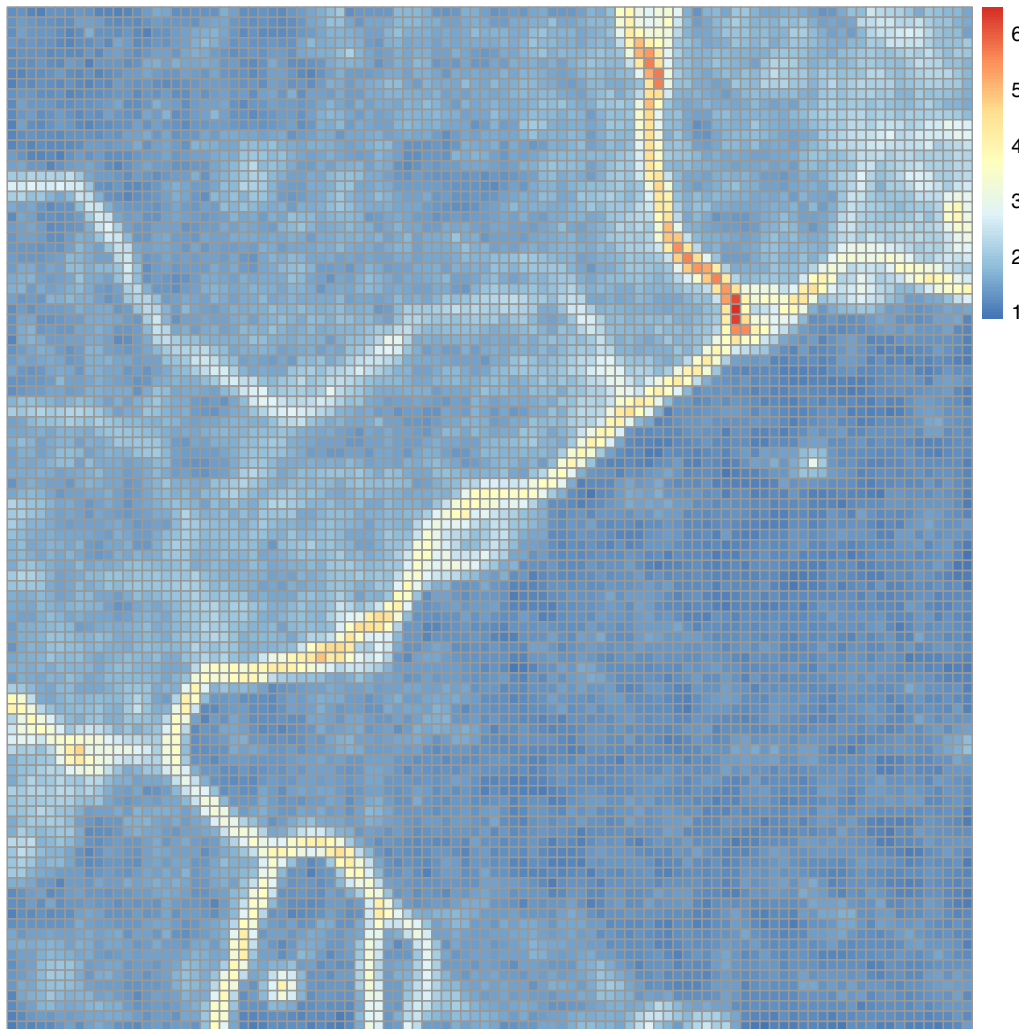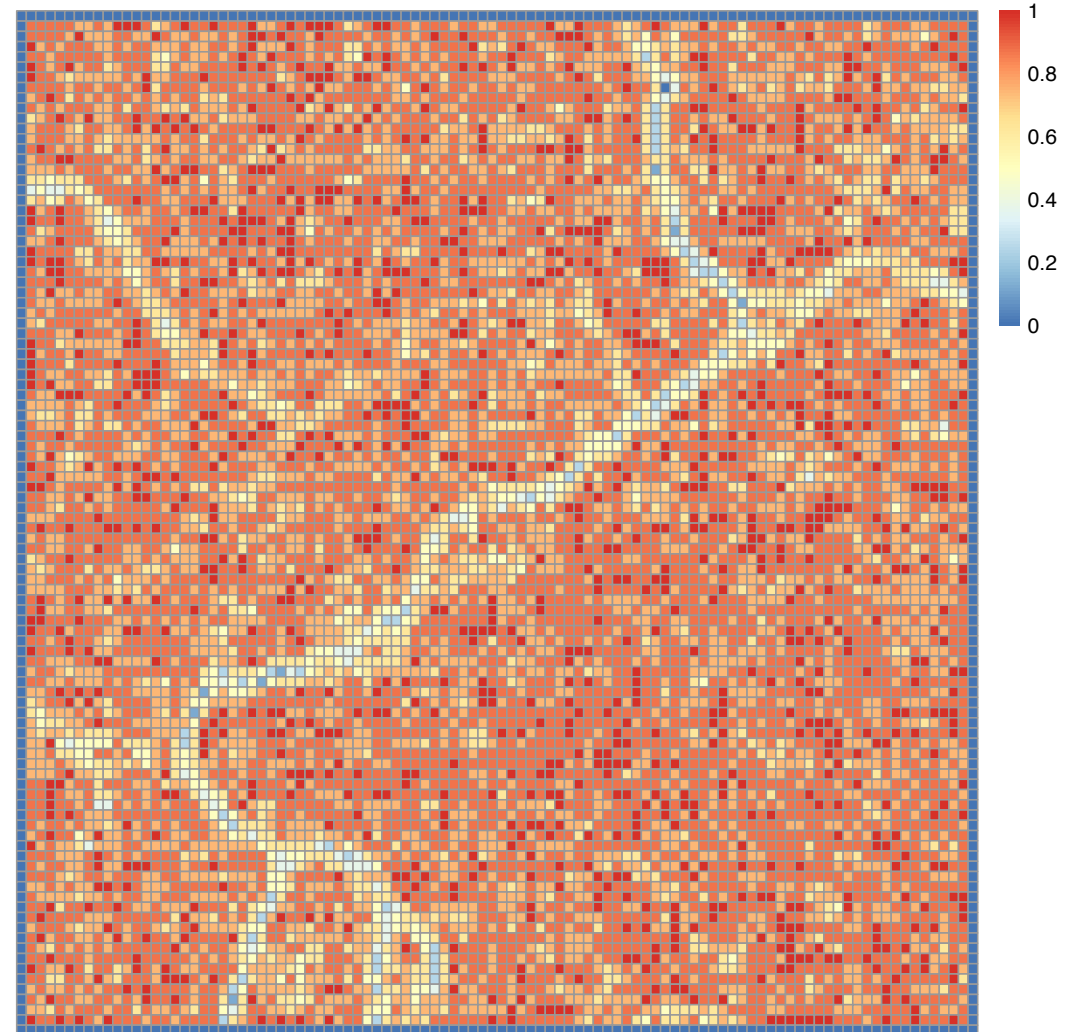
2) Compare KNN ID with the Moore Neighborhood ID

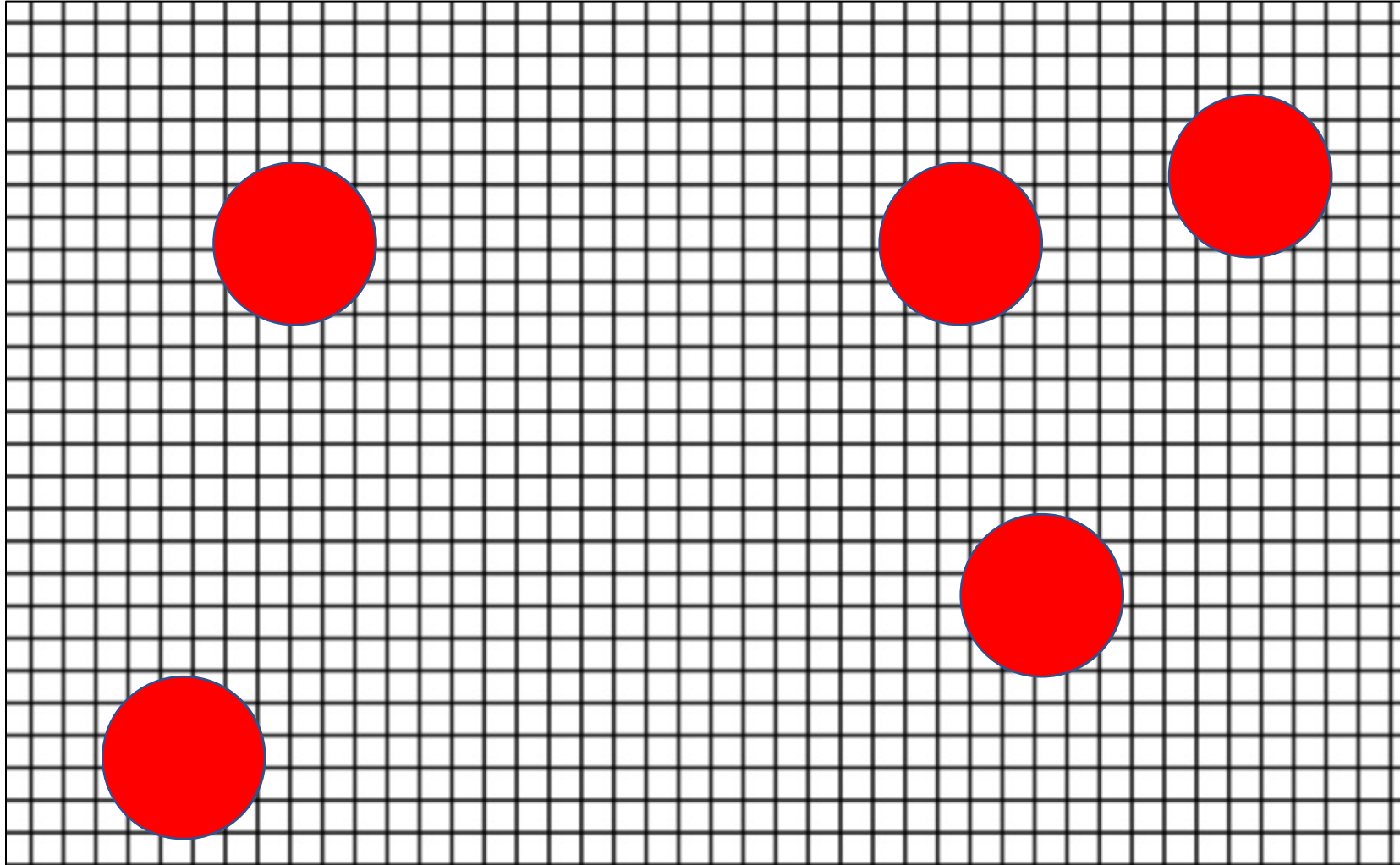# U-matrix fidelity: detecting topology of the SOM

U-matrix

Similarity of Outer Moore Neighborhood vs KNN with K of 8

Hypothesis: Emergent self-organizing map (more nodes than number of cells) will provide helpful visualization for users
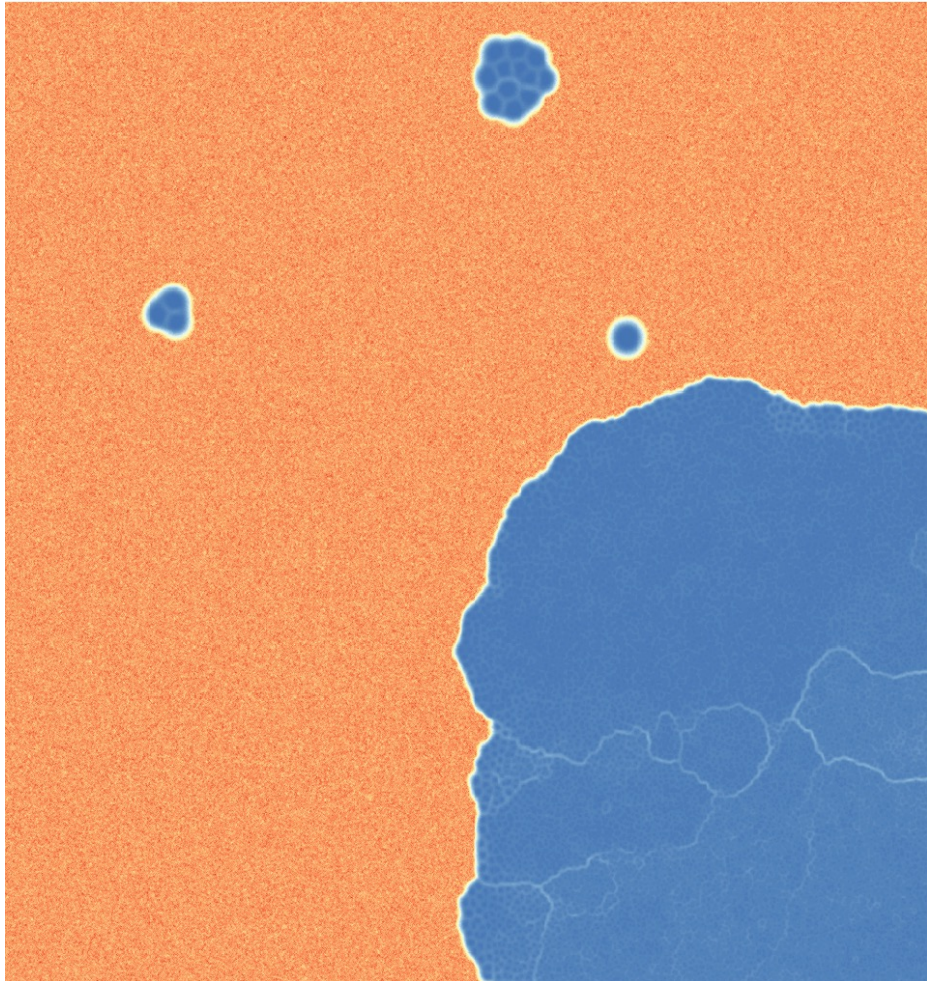
Method:

1) Train a 1000 x 1000 SOM on 10,000 cells. This makes 100 nodes for every cell.

(Runtime: 5 hours!)

2) Visualize the U-matrix, and cell occupancy

U-matrix and cell abundance profile of ESOM…still needs to be optimized
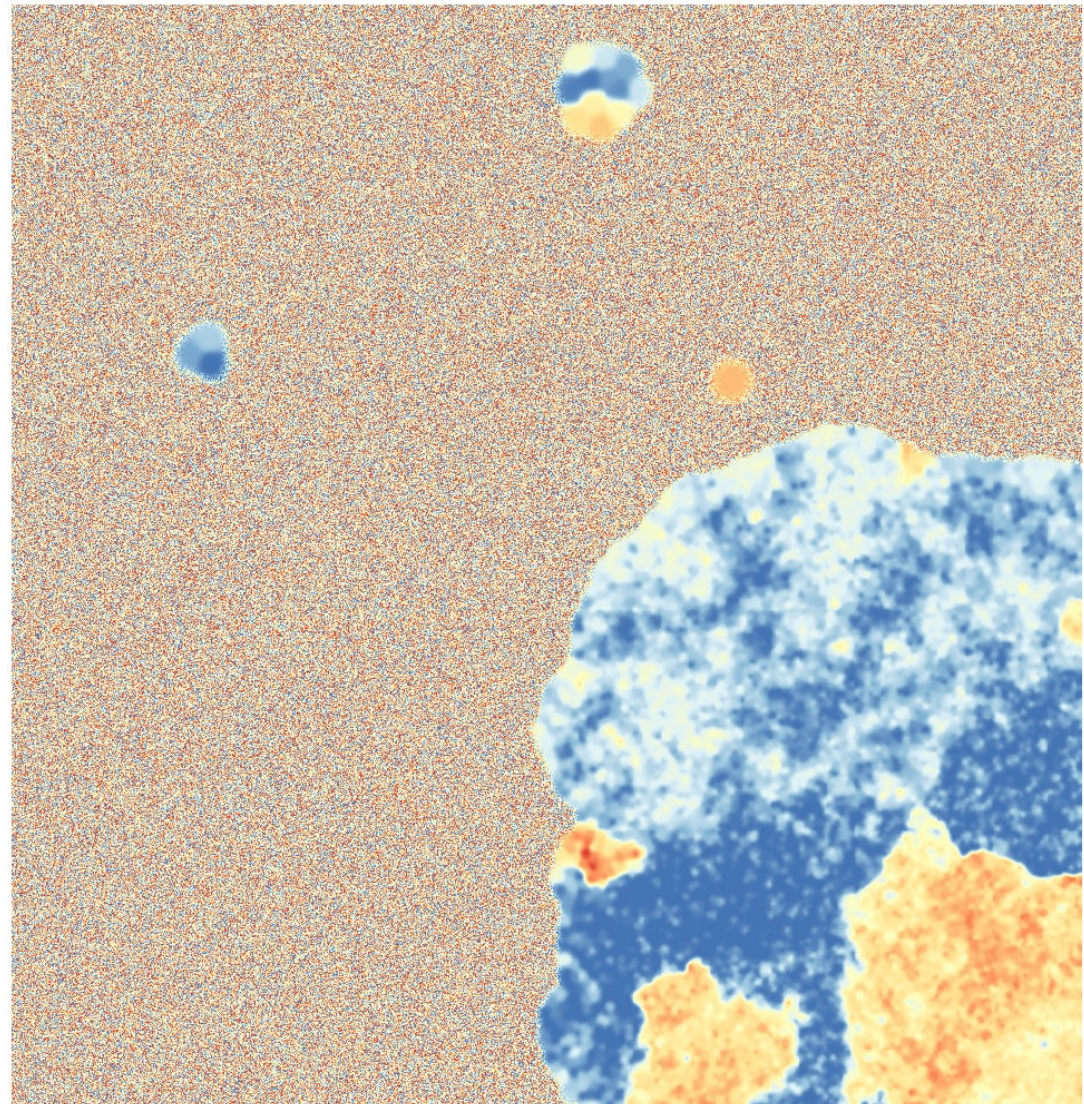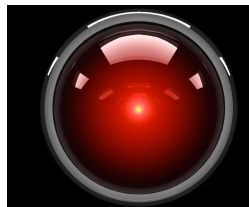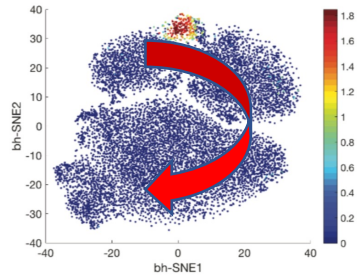

U-matrix

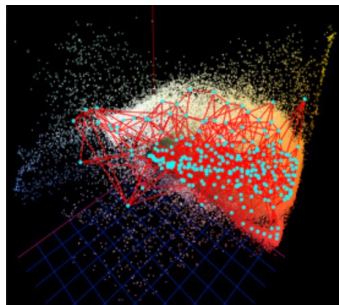Abundance

2001: A Marker Space Odyssey

Future directions

- Explore the use of U-Matrix visualizations on trajectory-based datasets

- Compare the results I have with that of a growing neural gas, or hierarchical SOM

Acknowledgement