SpaceMAP:
Visualizing High-dimensional Data by Space Expansion

Xinrui Zu Qian Tao

Department of Imaging Physics, Delft University of Technology, the Netherlands

TU Delft
ImPhys
Image to insight
Insight in imaging
Department of Imaging Physics
ICML
International Conference
On Machine Learning

SMACS 0723: Red arcs in the image trace light from galaxies in the very early Universe

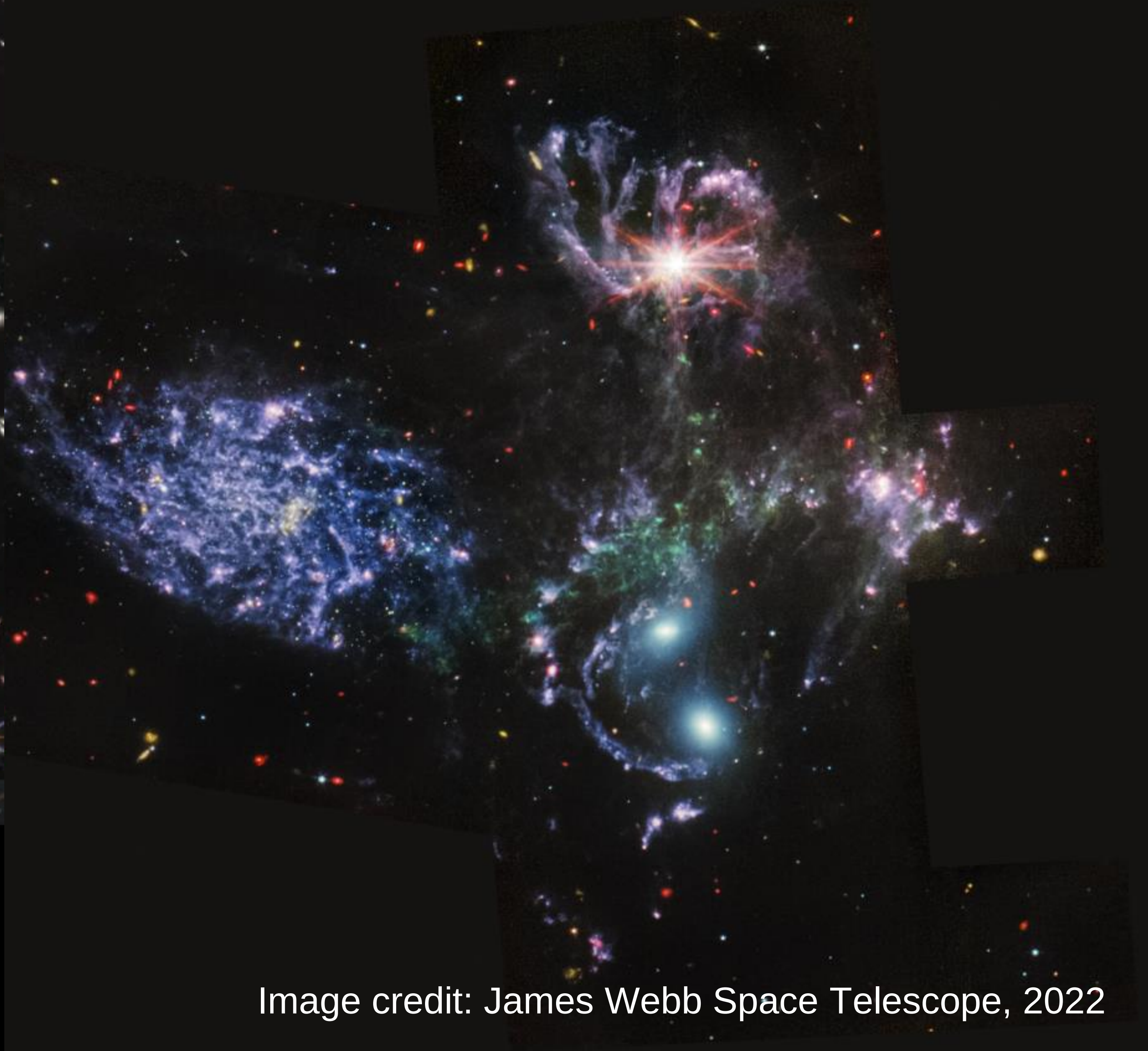


Image credit: James Webb Space Telescope, 2022

High-dimensional data visualization by SpaceMAP

What's SpaceMAP?

- SpaceMAP is a visualization / dimensionality reduction (DR) method that can see data of arbitrarily high dimension on a 2D map.

- It is based on understanding the capacity of SPACE.

- MAP refers to “manifold approximation and projection”.

Why is Visualization Interesting for ML?

- ML works based on the fundamental assumption that data lies on a low-dimensional manifold – otherwise “curse of dimensionality” holds

- Seeing is believing – as human we only can “see” well in 2 or 3 dimension, hence data visualization is super interesting for ML’ers!

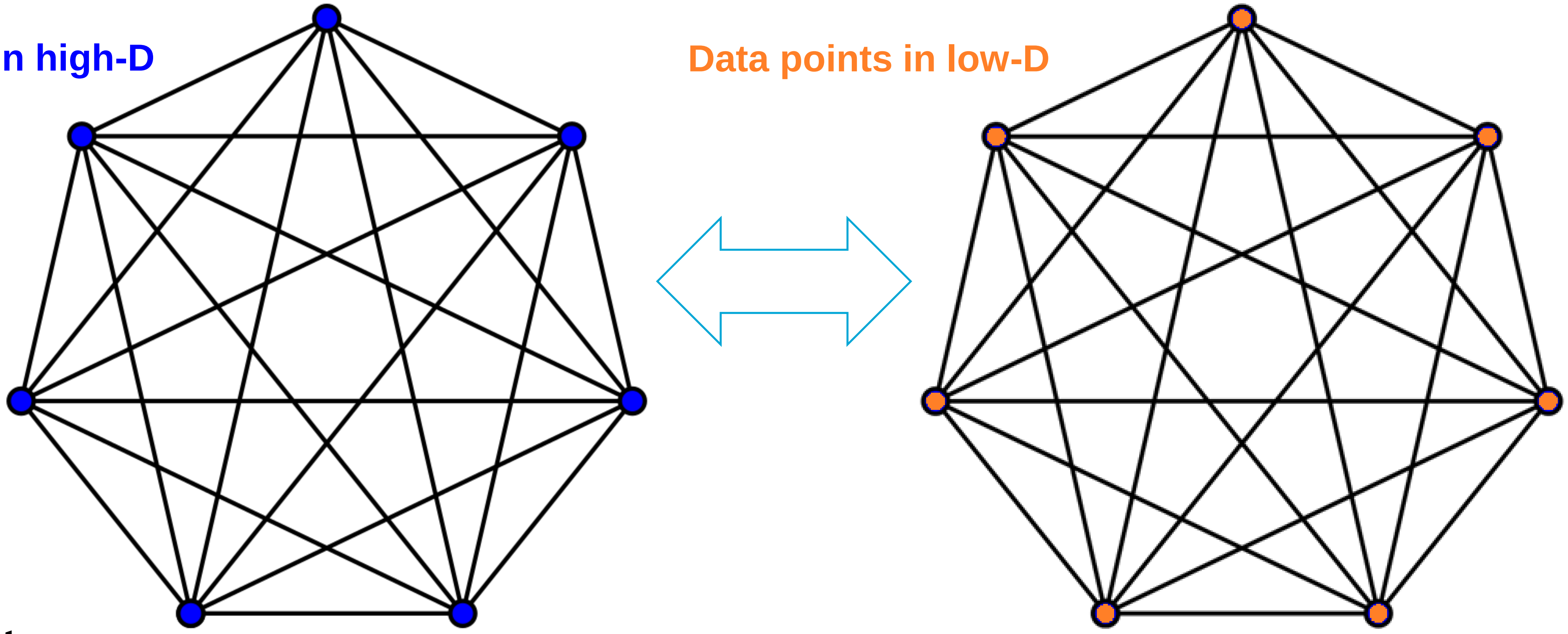
What's New in SpaceMAP?

- The discrepancy between high-D and low-D spaces is analytically studied, leading to transformation of similarity in a principled and explainable way.

- In contrast, previous methods such as t-SNE and UMAP transformed similarity implicitly.

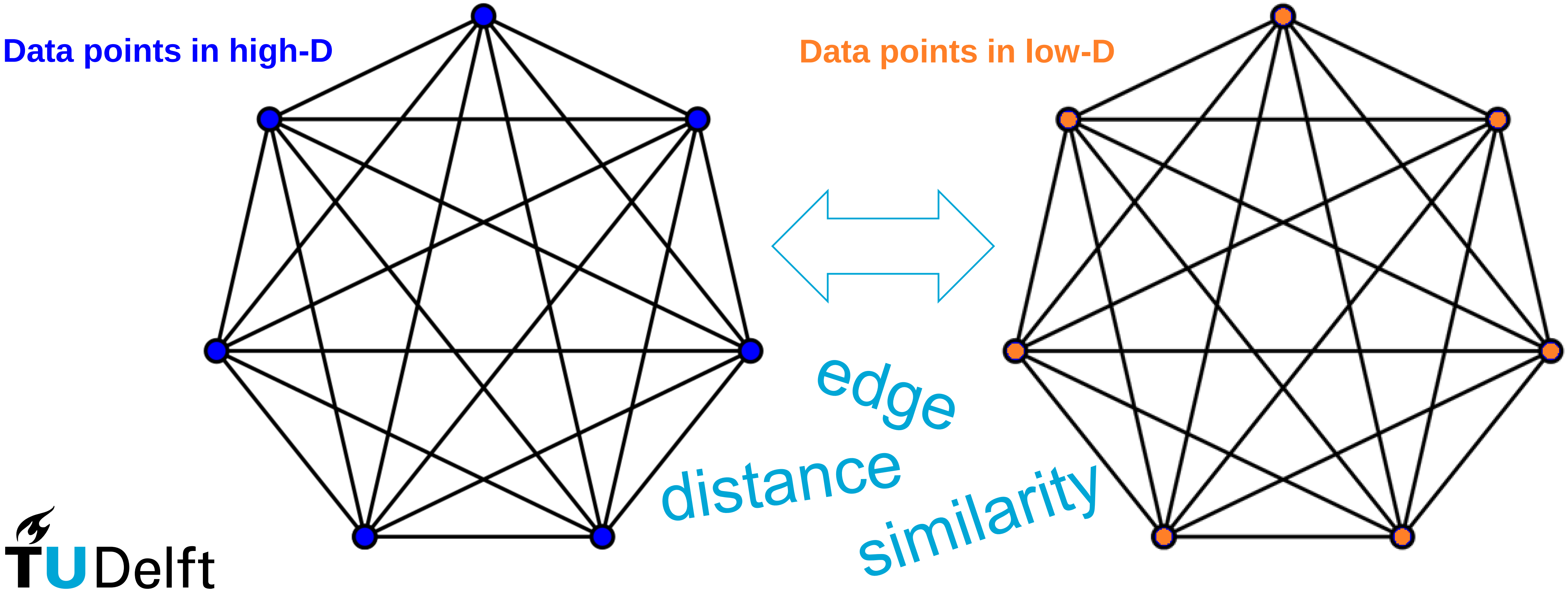
The Essence of Dimensionality Reduction

- Matching two graphs



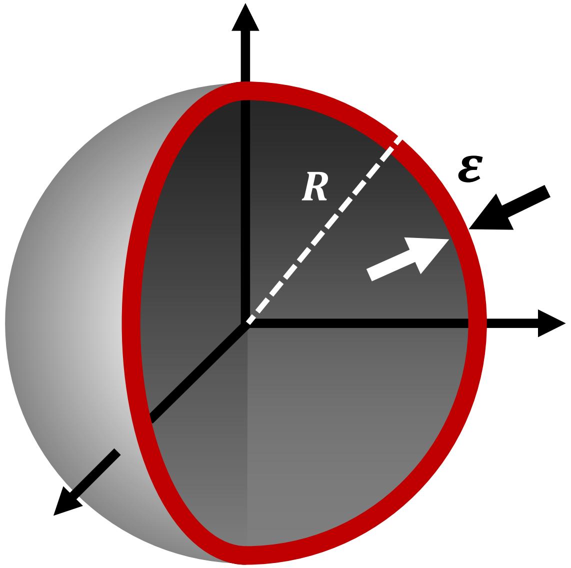
The Essence of Dimensionality Reduction

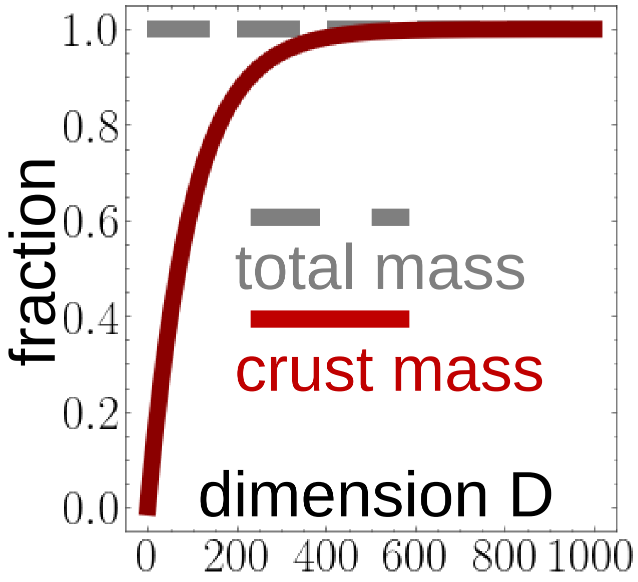
- Matching two graphs



The Crowding Problem of Dimensionality Reduction

- **high-dimensional** geometry: **“concentration on a crust”**



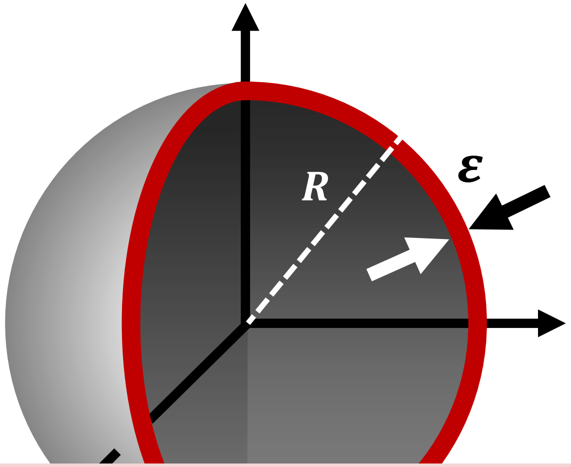


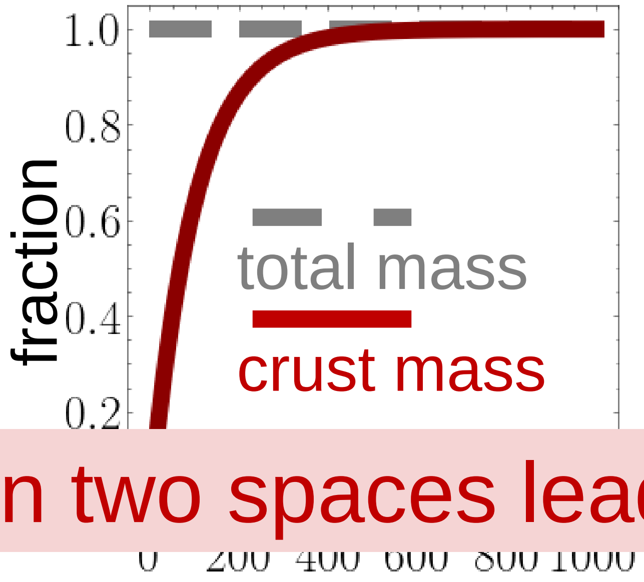
causes

- General difficulty in DR: **the “crowding problem”**

The Crowding Problem of Dimensionality Reduction

- **high-dimensional** geometry: **"concentration on a crust"**





Distance defined the same way in two spaces lead to problems!

causes

- General difficulty in DR: **the "crowding problem"**

TU Delft

SpaceMAP: the Theoretic Framework

- Space Capacity $\boldsymbol{\mathcal{V}_D(R_{ij})}$
  - A Hausdorff measure - volume of a D-dim ball
- Equivalent Extended Distance (EED) $\tilde{\mathcal{R}}_{ij,D\to d}$
  - Transform the distance in low-D space such that capacity matches to low-D space: $\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij})$

TU Delft

SpaceMAP: the Theoretic Framework

- Space Capacity $\mathcal{V}_D(R_{ij})$
  - A Hausdorff measure - volume of a D-dim ball

Definition 3.1 (Space Capacity). *Let $R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$ be the distance between data point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the D-dimensional space. The space capacity $\mathcal{V}_D(R_{ij})$ from point i to point j is defined as the volume of a D-dimensional ball with a radius of R_{ij}.*

- Equivalent Extended Distance
  - Transform the distance in low-D space such that capacity matches to low-D space: $\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij})$

$$\tilde{\mathcal{R}}_{D\to d} = \alpha R^{\frac{D}{d}}$$

Definition 3.2 (Equivalent Extended Distance: EED). *Let $R_{ij} = l(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathbb{R}$ be the distance between data point $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the D-dimensional space. The equivalent extended distance (EED) $\tilde{\mathcal{R}}_{ij,D\to d}$ is defined as the equivalent distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in d-dimensional space such that the Space Capacity matches: $\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij})$*

SpaceMAP: the Theoretic Framework

- Intrinsic Dimension (ID)
  - Inherent degrees of freedom of data

- EED provably transforms ID
  - By applying EED $\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij})$, the ID of data is transformed to be visualizable with mitigated “crowding problem”

SpaceMAP: the Theoretic Framework

- Intrinsic Dimension (ID) Estimation
  - Inherent degrees of freedom of data

$$\hat{d}_k(x_i; R) = \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{R_{ik}}{R_{ij}} \right)^{-1}$$

Levina & Bickel 2004

- EED provably transforms ID
  - By applying EED $\mathcal{V}_d(\tilde{\mathcal{R}}_{ij,D\to d}) = \mathcal{V}_D(R_{ij})$, the ID of data is transformed to be visualizable with mitigated “crowding problem”

Proposition 3.1 (EED transforms ID provably). *For any neighborhood size k, if the MLE of the intrinsic dimension around point x_i under the distance metric R is $\hat{d}_k(x_i; R) = D$, the MLE of the intrinsic dimension after applying EED to the distance metric is d: $\hat{d}_k(x_i; \tilde{\mathcal{R}}_{D\to d}) = d$.*

Proof. By replacing metric R with the EED-transformed metric $\tilde{\mathcal{R}}_{D\to d}$ (Equation 4) in Equation 5, we have:

$$\begin{aligned}\hat{d}_k(x_i; \tilde{\mathcal{R}}_{D\to d}) &= \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{\alpha R_{ik}^{D/d}}{\alpha R_{ij}^{D/d}} \right)^{-1} \\ &= \left(\frac{1}{k-1} \sum_{j=1}^{k-1} \frac{D}{d} \log \frac{R_{ik}}{R_{ij}} \right)^{-1} \\ &= \frac{d}{D} \hat{d}_k(x_i; R) = \frac{d}{D} D = d \qquad (7)\end{aligned}$$

□

TU Delft

SpaceMAP: the Method

■ Definitions in SpaceMAP:

Space Capacity

$\mathcal{V}_D(R_{ij})$ Volume of a D-dim ball

↓

Equivalent Extended Distance (EED)

$\tilde{\mathcal{R}}_{D\to d}$ s.t. $\mathcal{V}_D(R_{ij}) = \mathcal{V}_d(\mathcal{R}_{D\to d})$

Estimation of ID

$\hat{\mathbf{d}}_{\mathbf{k}}$ k: number of neighbors

↓

Local/Global ID

$d_{local}(x_i; k) = \hat{d}_k$

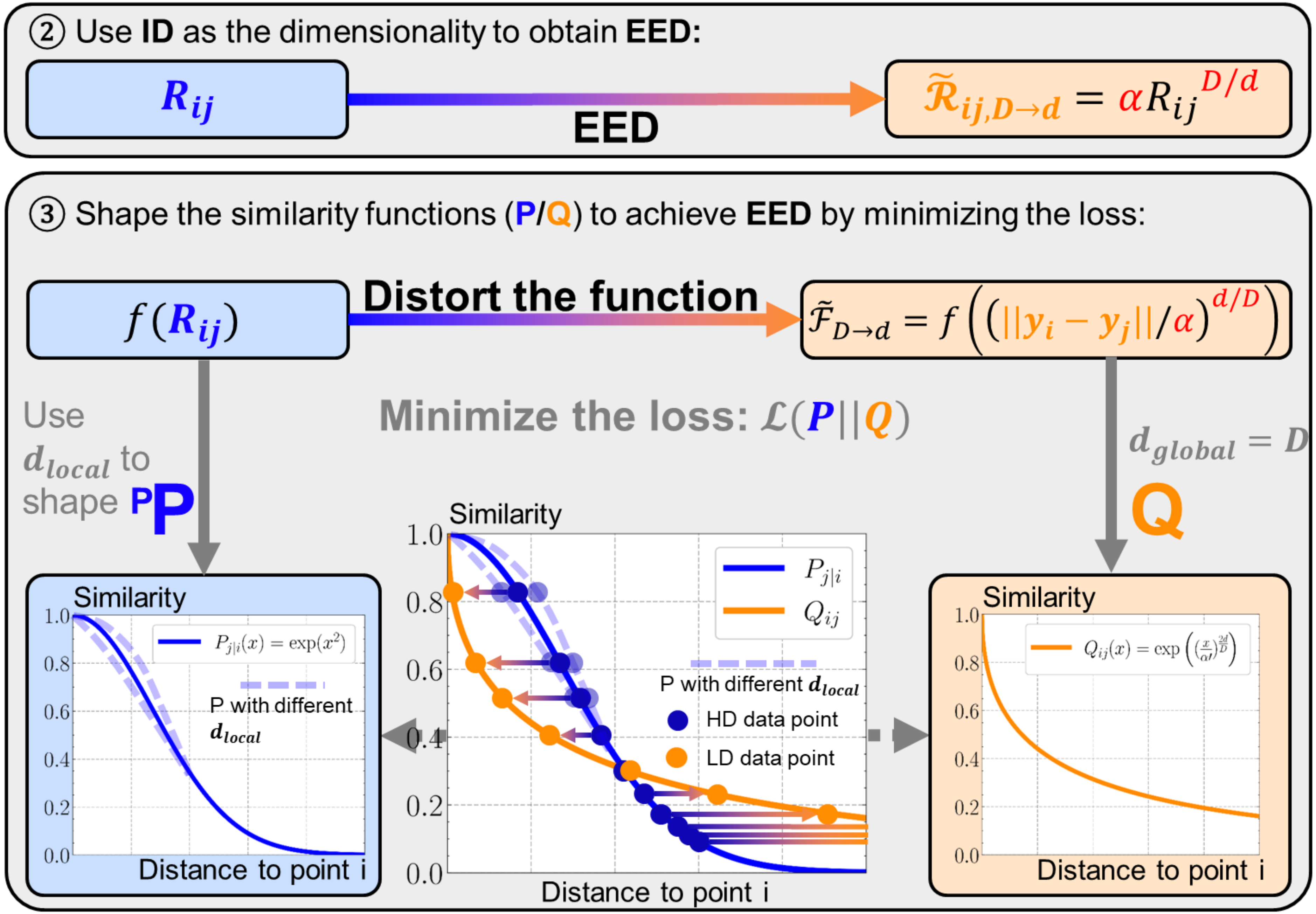
d_{global} Harmonic mean of d_{local}

■ Algorithm:

① Calculate local/global **ID** → ② Use **ID** to obtain **EED** among the data points → ③ Shape the similarity functions (**P**/**Q**) to achieve **EED** using **local/global ID**

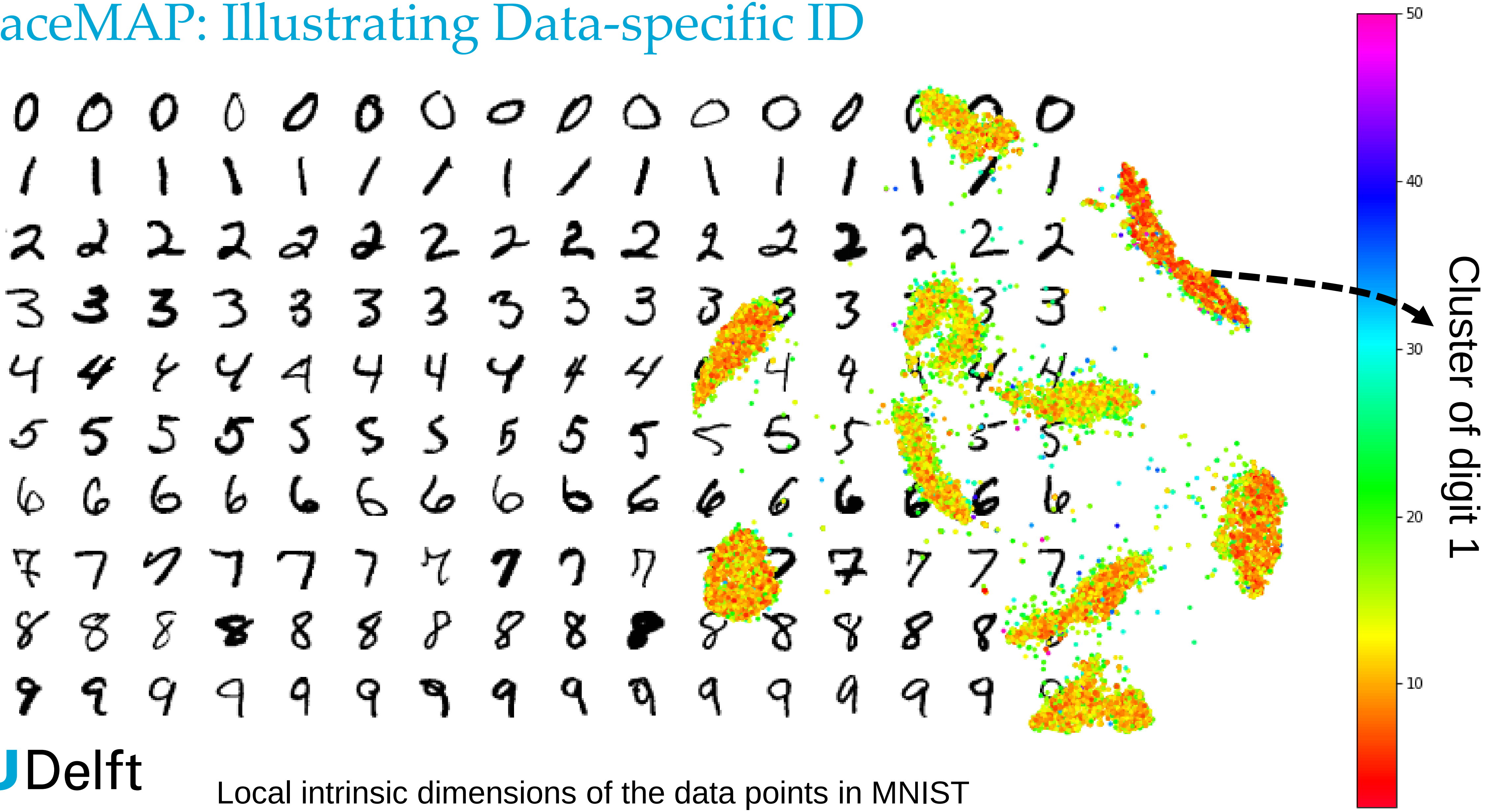
TU Delft

SpaceMAP: the Method



TUDelft

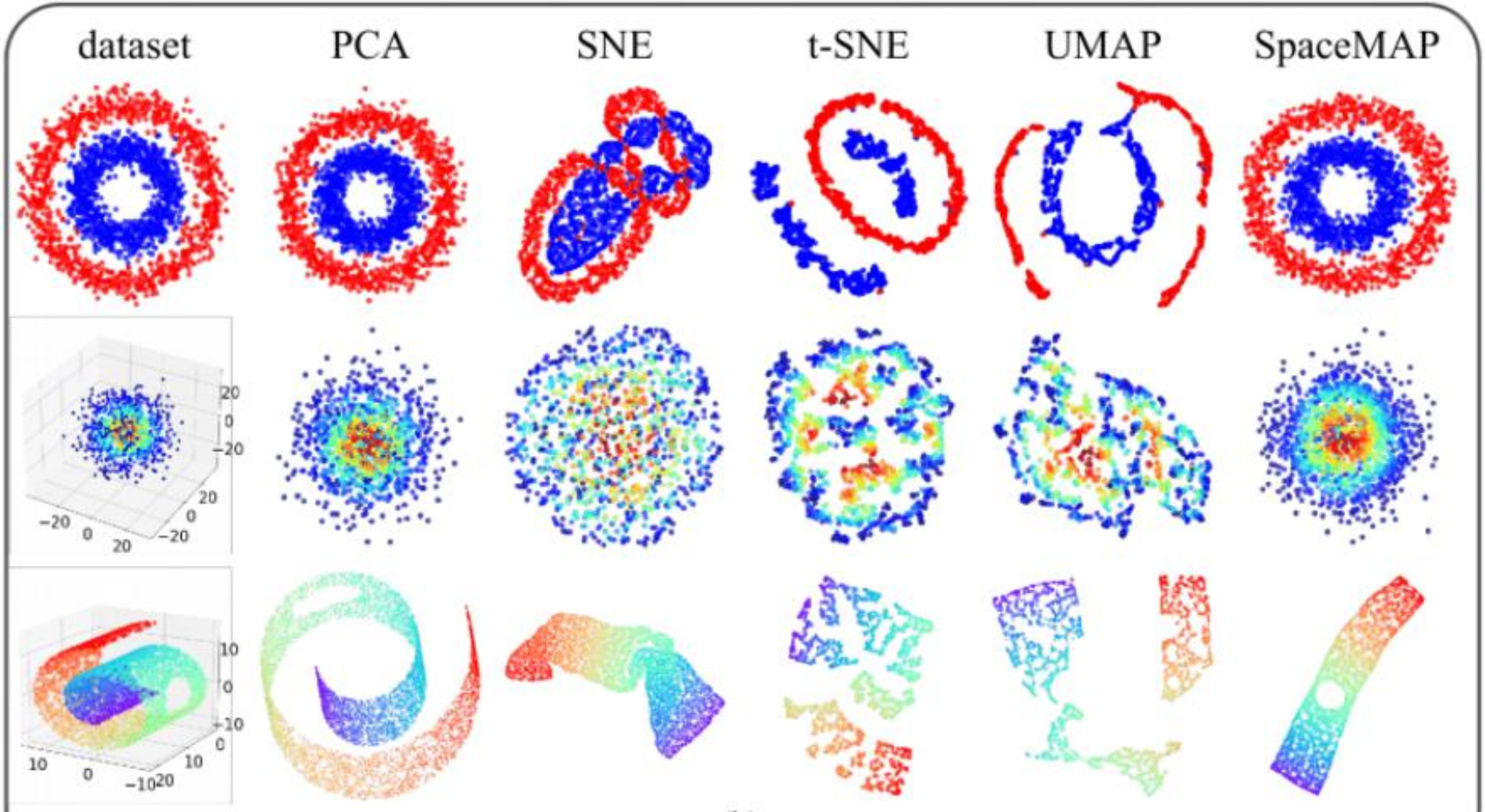
SpaceMAP: Illustrating Data-specific ID



TU Delft

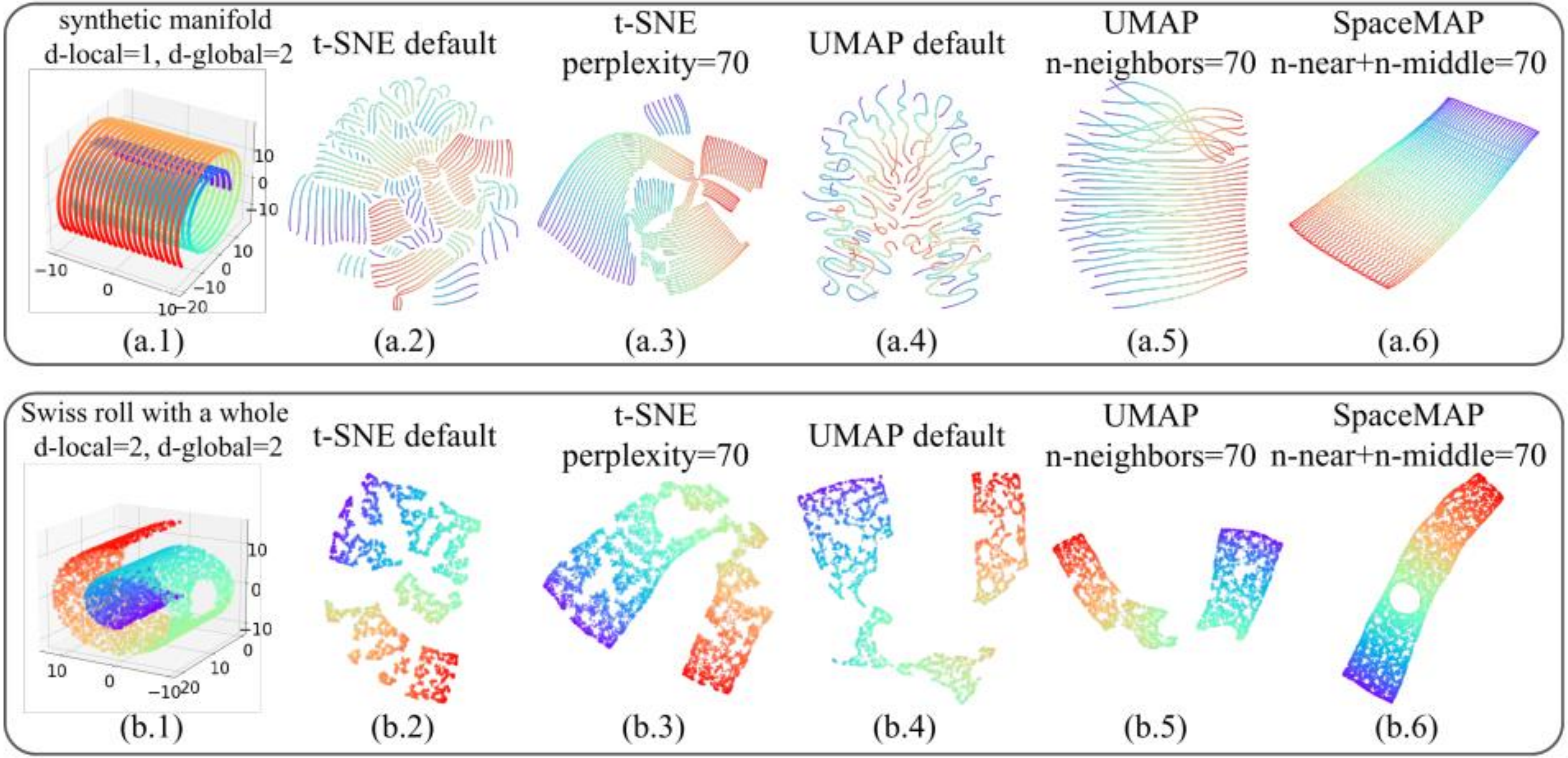
Local intrinsic dimensions of the data points in MNIST

SpaceMAP Results

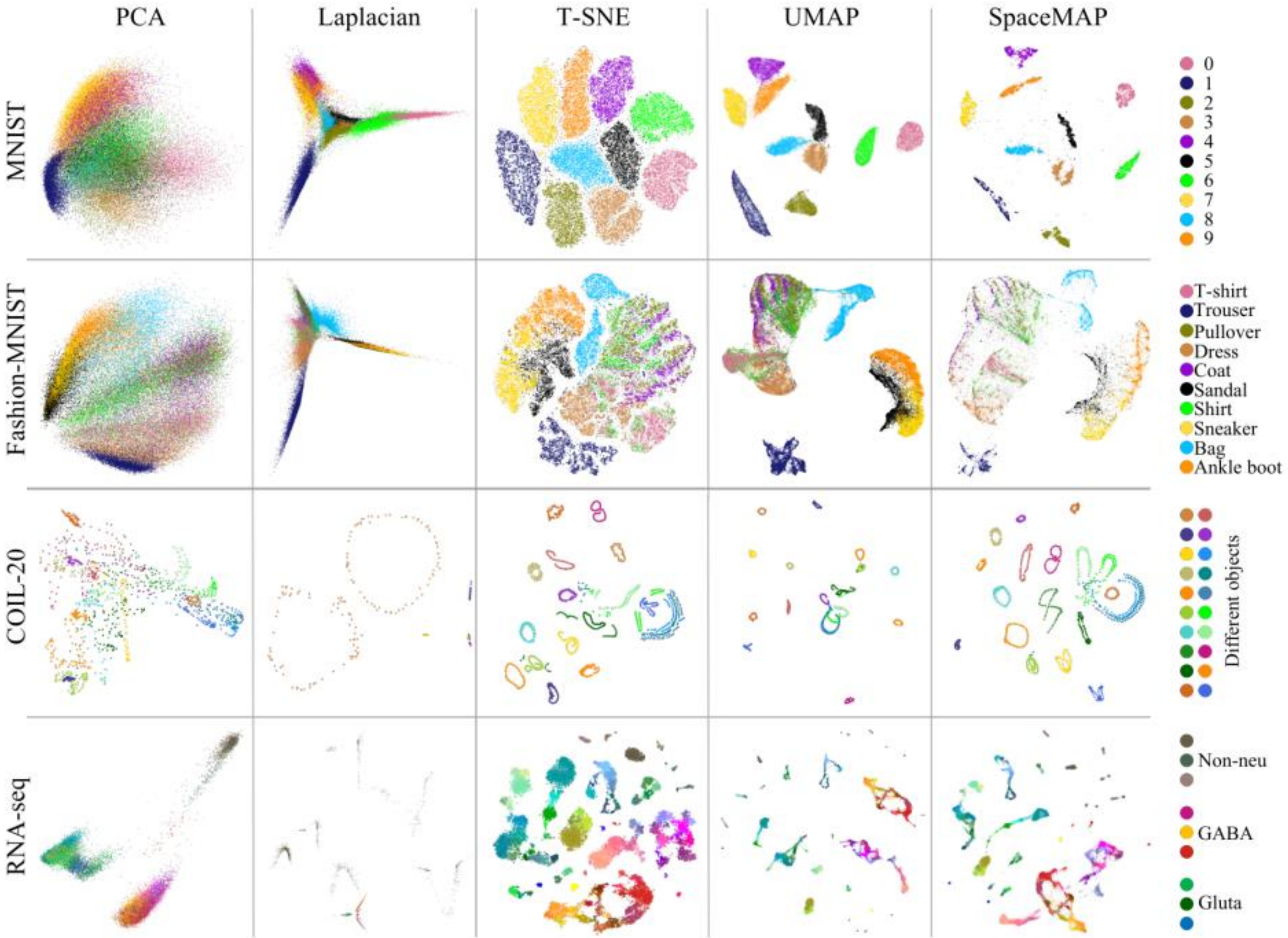


(b)

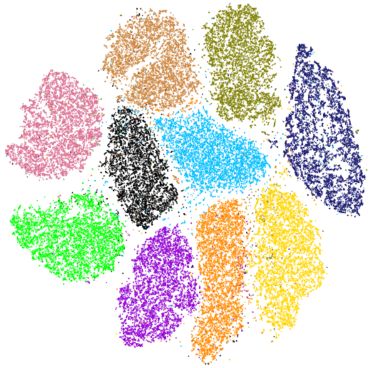
SpaceMAP Results



SpaceMAP


PCA
Laplacian
T-SNE
UMAP
SpaceMAP
MNIST
Fashion-MNIST
COIL-20
RNA-seq
0
1
2
3
4
5
6
7
8
9
T-shirt
Trouser
Pullover
Dress
Coat
Sandal
Shirt
Sneaker
Bag
Ankle boot
Different objects
Non-neu
GABA
Gluta

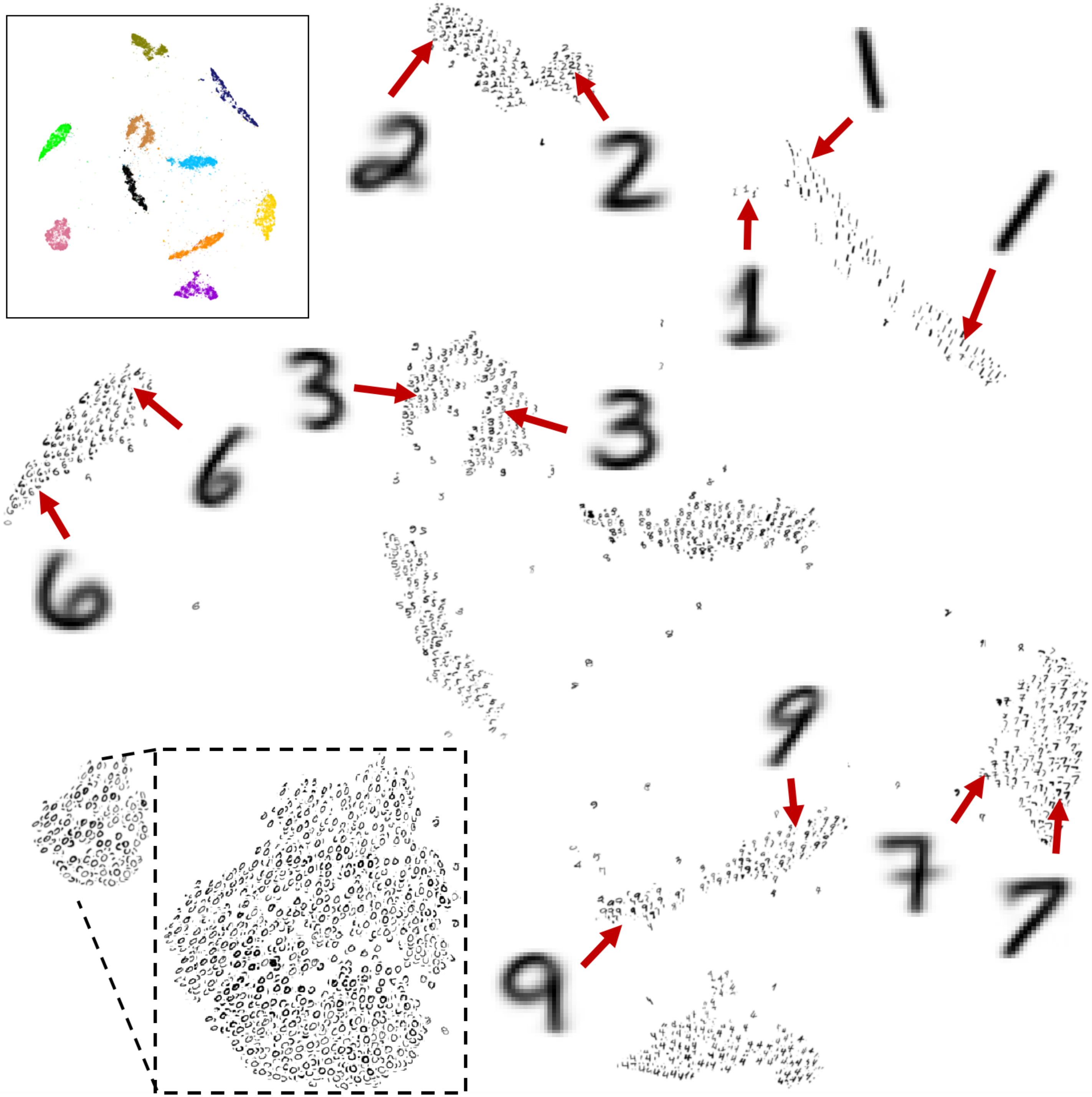
SpaceMAP Results



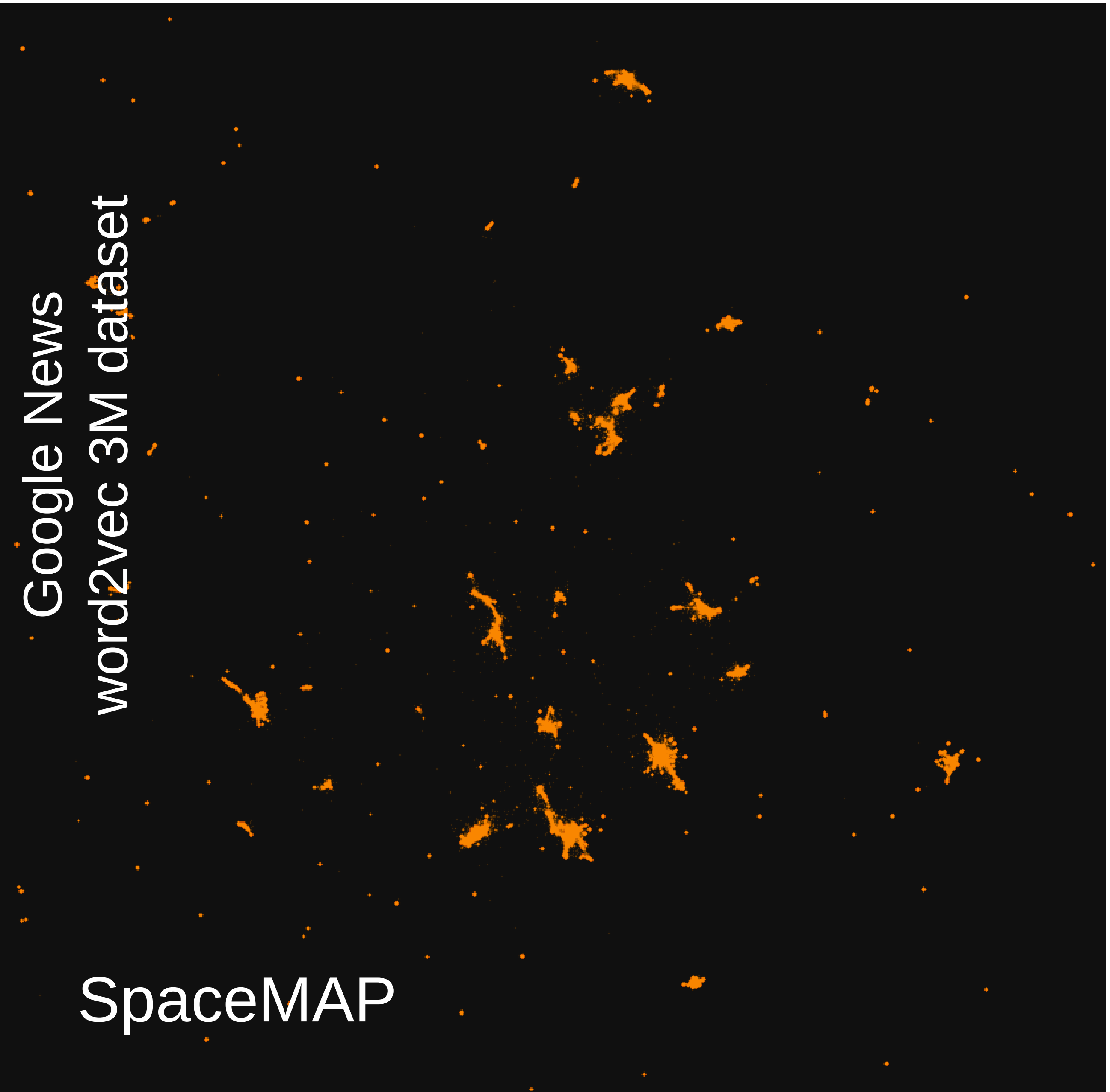
t-SNE result

UMAP result

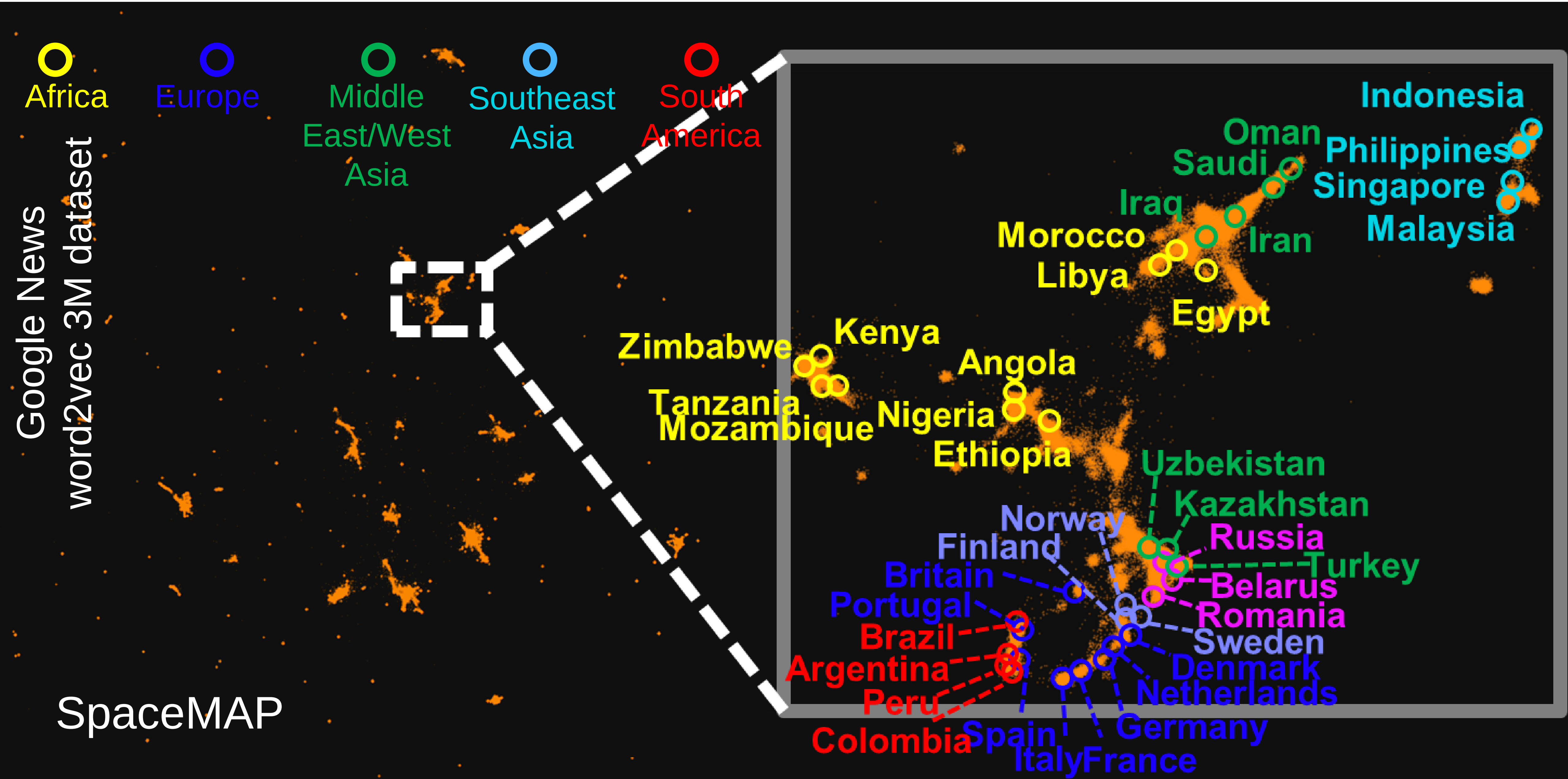
TU Delft



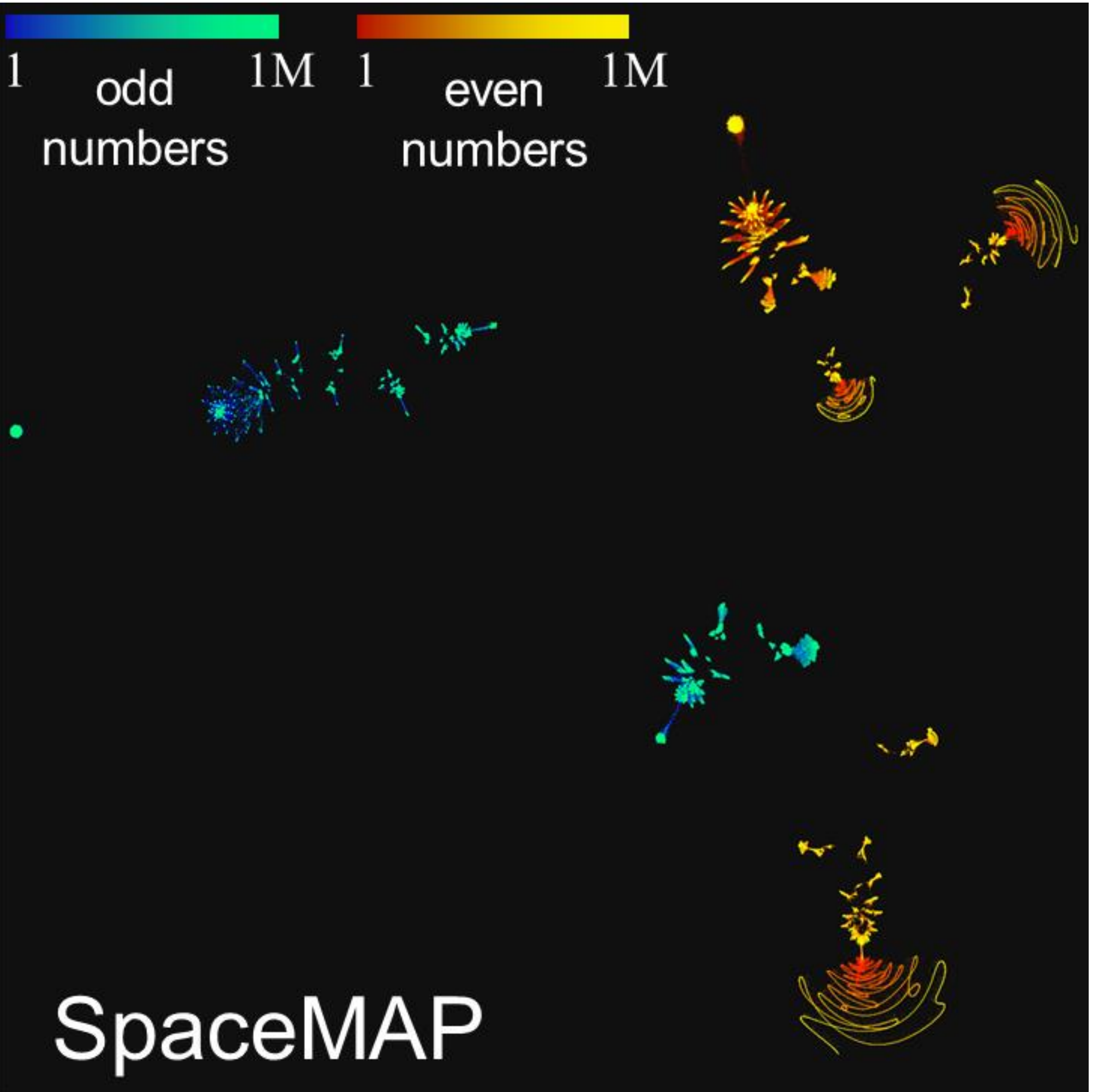
SpaceMAP Results: the Word Map



SpaceMAP Results: the Word Map – World Map



SpaceMAP Results



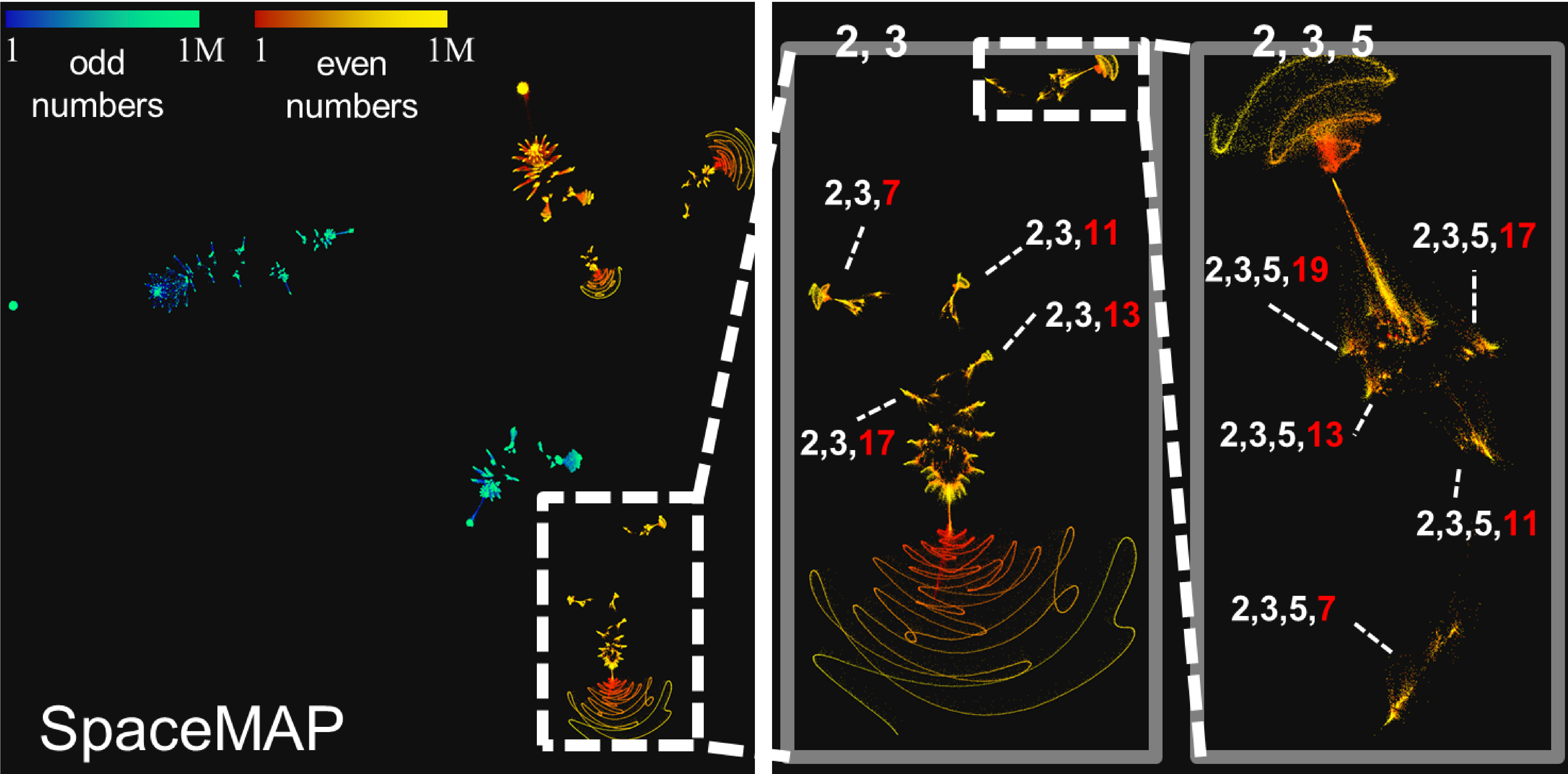
Divisibility by prime numbers:

- A binary vector showing the divisibility of positive integer from 1 to 1,000,000 by prime number 2, 3, 5,…, 999983

UMAP McInnes et al. 2018

- 78498-dimensional binary vector
- Visualized in 2D SpaceMAP

SpaceMAP Results



SpaceMAP: Visualizing High-dimensional Data by Space Expansion

Xinrui Zu Qian Tao
Department of Imaging Physics, Delft University of Technology

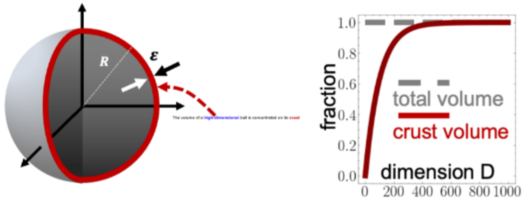
Poster #623

Dimensionality Reduction (DR) and Intrinsic Dimension (ID)

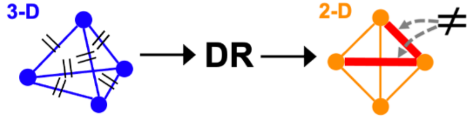
- Dimensionality Reduction (**DR**) translates **high-dimensional data** into **low-dimensional space** <2-D/3-D for visualization>.
- Intrinsic dimension (**ID**) is the internal degrees of freedom of data <usually larger than 3-D>.

'Concentration on a Crust' and the 'Crowding Problem'

- **high-dimensional** geometry: **'concentration on a crust'**:



- General difficulty in DR: **the 'crowding problem'**:



The distances between high-dimensional data points are **concentrated**, which are **difficult to preserve** in low-dimensional spaces.

Our Main Contribution

- **Analytically** alleviate the crowding problem in a data-specific manner.
- Hierarchical manifold approximation by estimating **local/global ID**.

Proposition (EED transforms ID provably)

- For any dataset with $ID = D$, if we apply EED $\tilde{\mathcal{R}}_{D\to d}$, then $ID = d$.

High-dimensional dataset with distances: R_{ij} — ID: D → EED → Dataset with EED distances: $\tilde{\mathcal{R}}_{ij,D\to d} = \alpha R_{ij}^{D/d}$ — ID: d

√ **The extended distances are easier to embed in the d-dimensional space!**

Methodology

- Definations in SpaceMAP:

Space Capacity $\mathcal{V}_D(R_{ij})$ Volume of a D-dim ball

Equivalent Extended Distance (EED) $\tilde{\mathcal{R}}_{D\to d}$ s.t. $\mathcal{V}_D(R_{ij}) = \mathcal{V}_d(\mathcal{R}_{D\to d})$

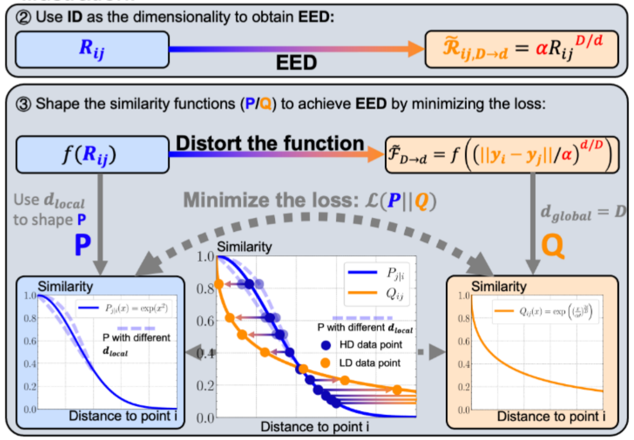
Estimation of ID $\hat{d}_k$ k: number of neighbors

Local/Global ID $d_{local}(x_i; k) = \hat{d}_k$; d_{global} Harmonic mean of d_{local}

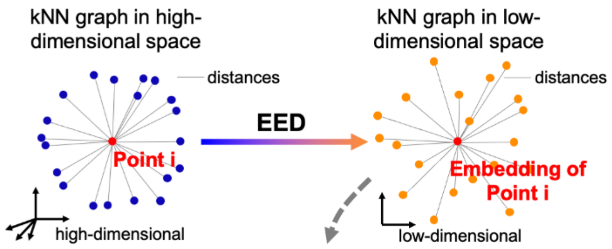
- Algorithm:

① Calculate local/global ID → ② Use ID to obtain EED among the data points → ③ Shape the similarity functions (P/Q) to achieve EED using local/global ID

- Illustration:

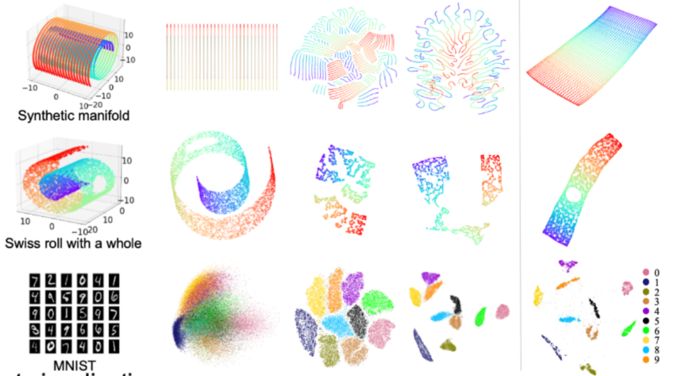


A Simple Example

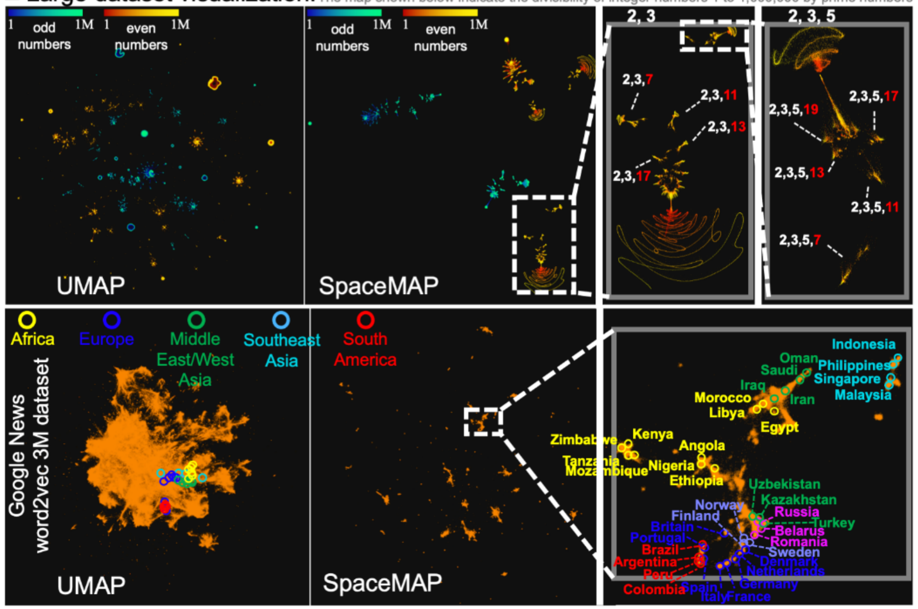


Results

- Visualization results:



- Large dataset visualization: The map shown below indicate the divisibility of integer numbers 1 to 1,000,000 by prime numbers



Conclusion

- We introduce the definitions of **space capacity**, **intrinsic dimension (ID)** and **equivelant extended distance (EED)** and utilize them to transform distances between high- and low-dimensional spaces and alleviate the 'crowding problem' analytically.
- We model the hierarchical structure in a dataset-specific manner based on the **local and global IDs** of data.

S P A C E M A P

- Space Capacity
- Intrinsic Dimension (ID)
- Equivalent Extended Distance (EED)
- Data-specific Manifold