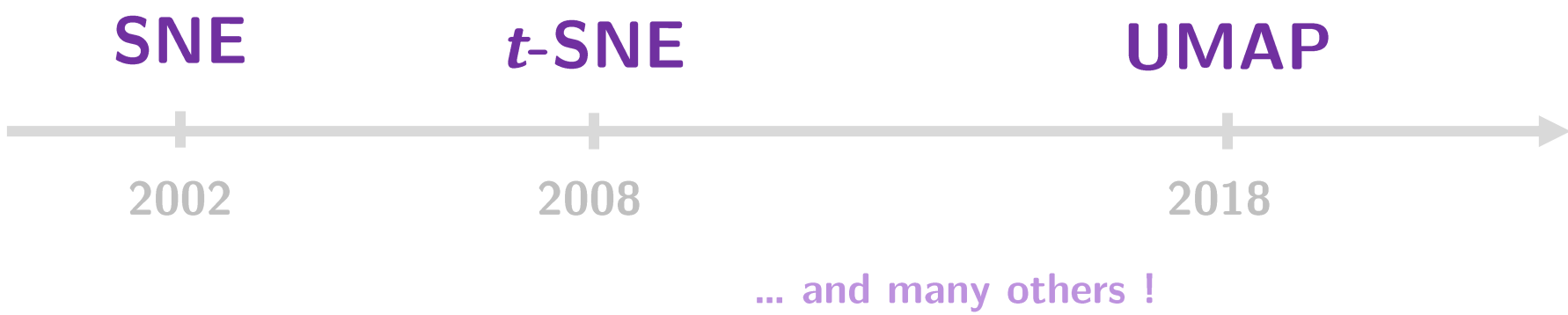
Natively Interpretable t-SNE

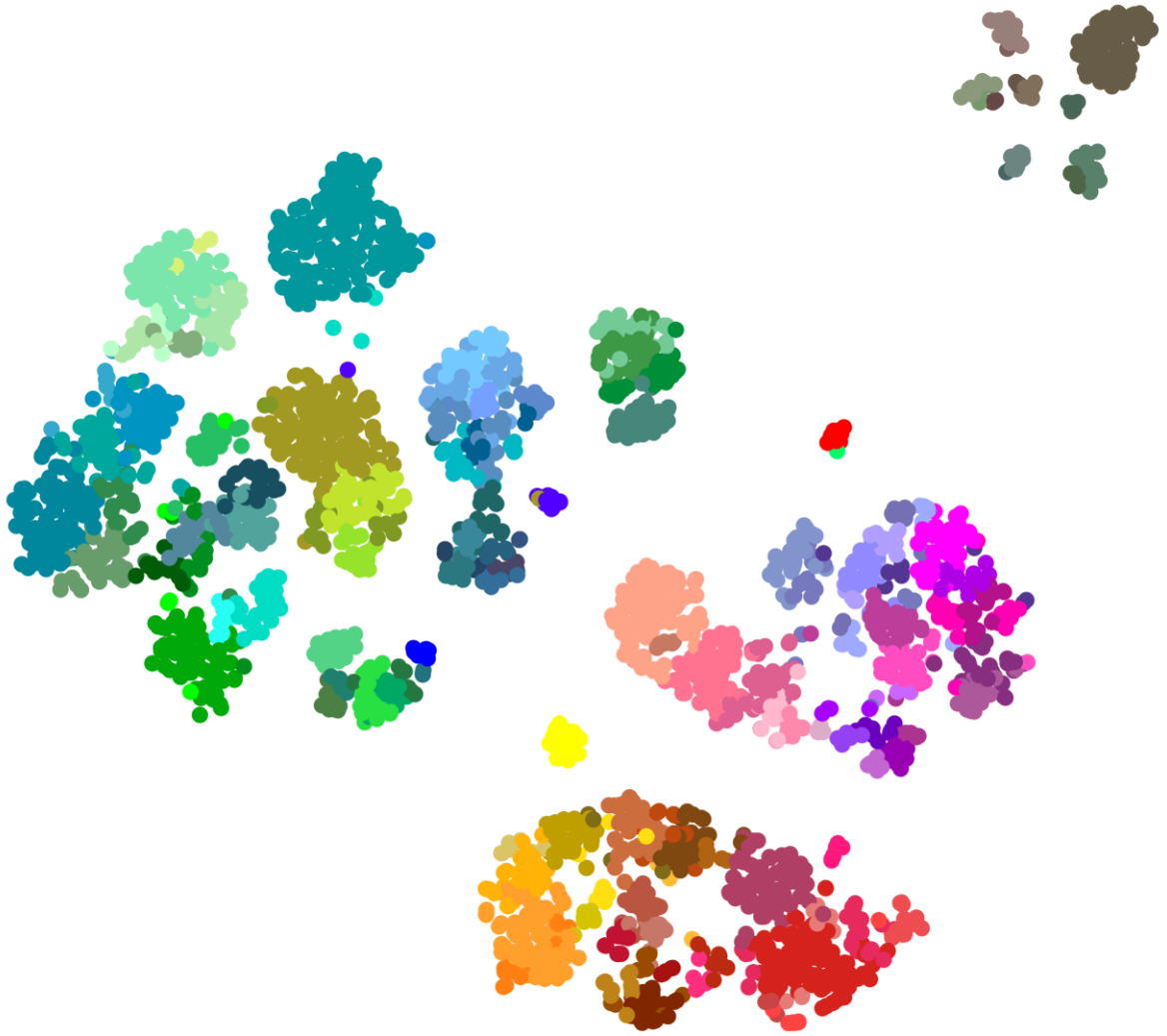
Neighbor Embedding



Mouse cortex dataset [1]

Single cell RNA sequencing

Dim = ~ 9500



[1] Tasic, Bosiljka, et al. "Shared and distinct transcriptomic cell types across neocortical areas." *Nature* 563.7729 (2018)

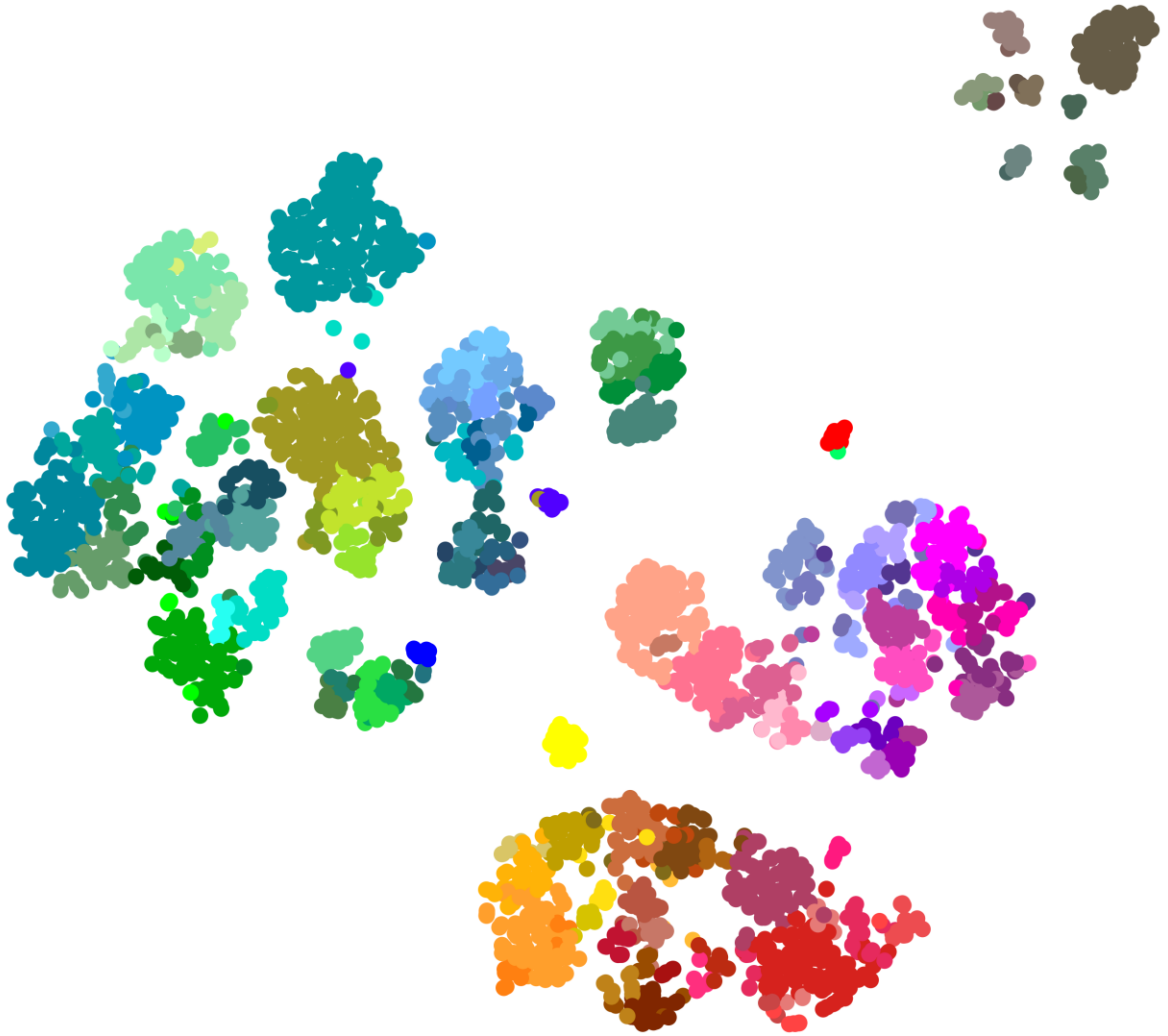
Mouse cortex dataset [1]

Single cell RNA sequencing

Dim = ~ 9500

Preprocessing [2]

Dim = 50



[2] Kobak, Dmitry, and Philipp Berens. "The art of using t-SNE for single-cell transcriptomics." *Nature communications* 10.1 (2019)

Mouse cortex dataset [1]

Single cell RNA sequencing

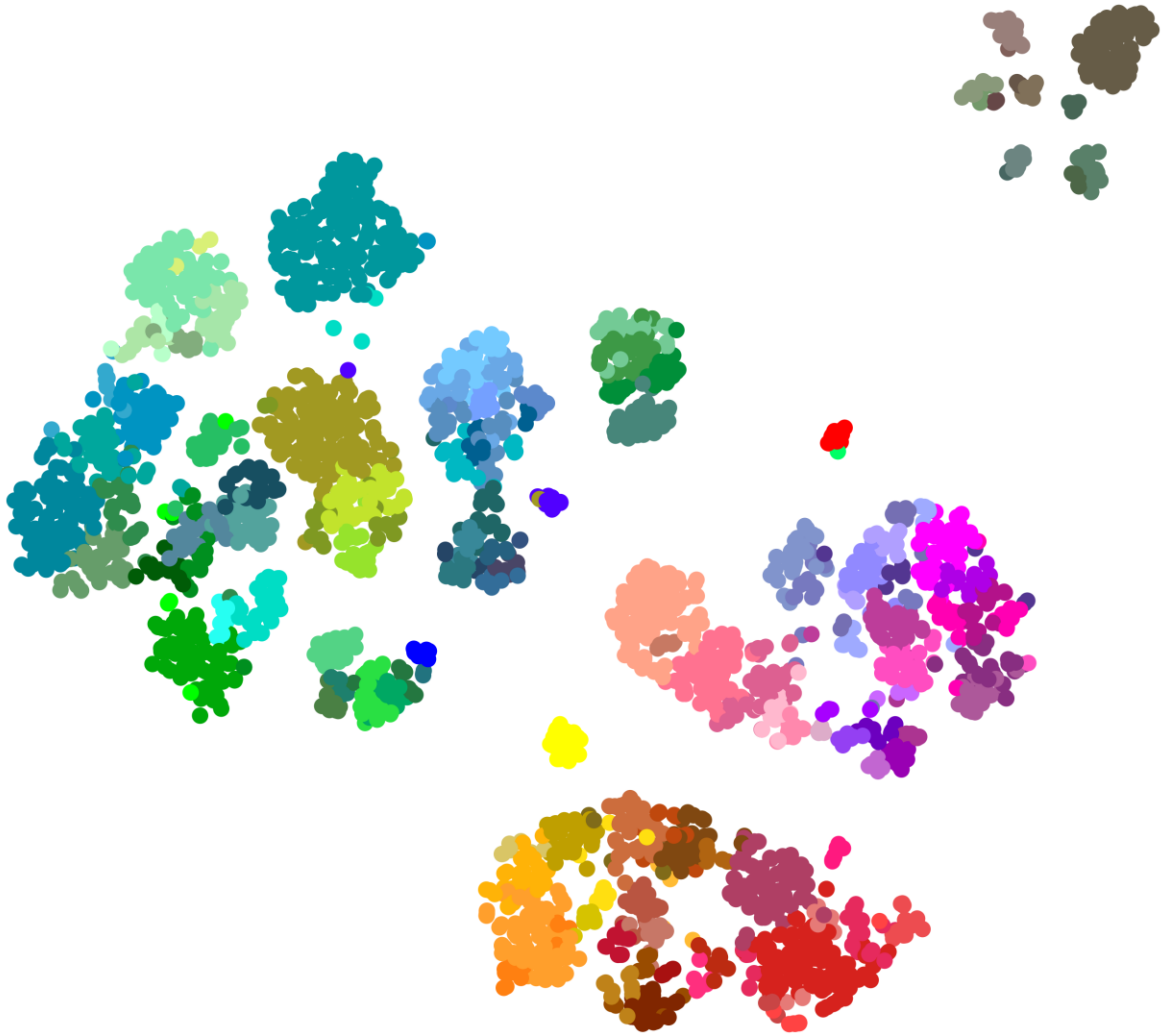
Dim = ~ 9500

Preprocessing [2]

Dim = 50 (HD)

Ms.*t*-SNE

Dim = 2 (LD)



Mouse cortex dataset [1]

Single cell RNA sequencing

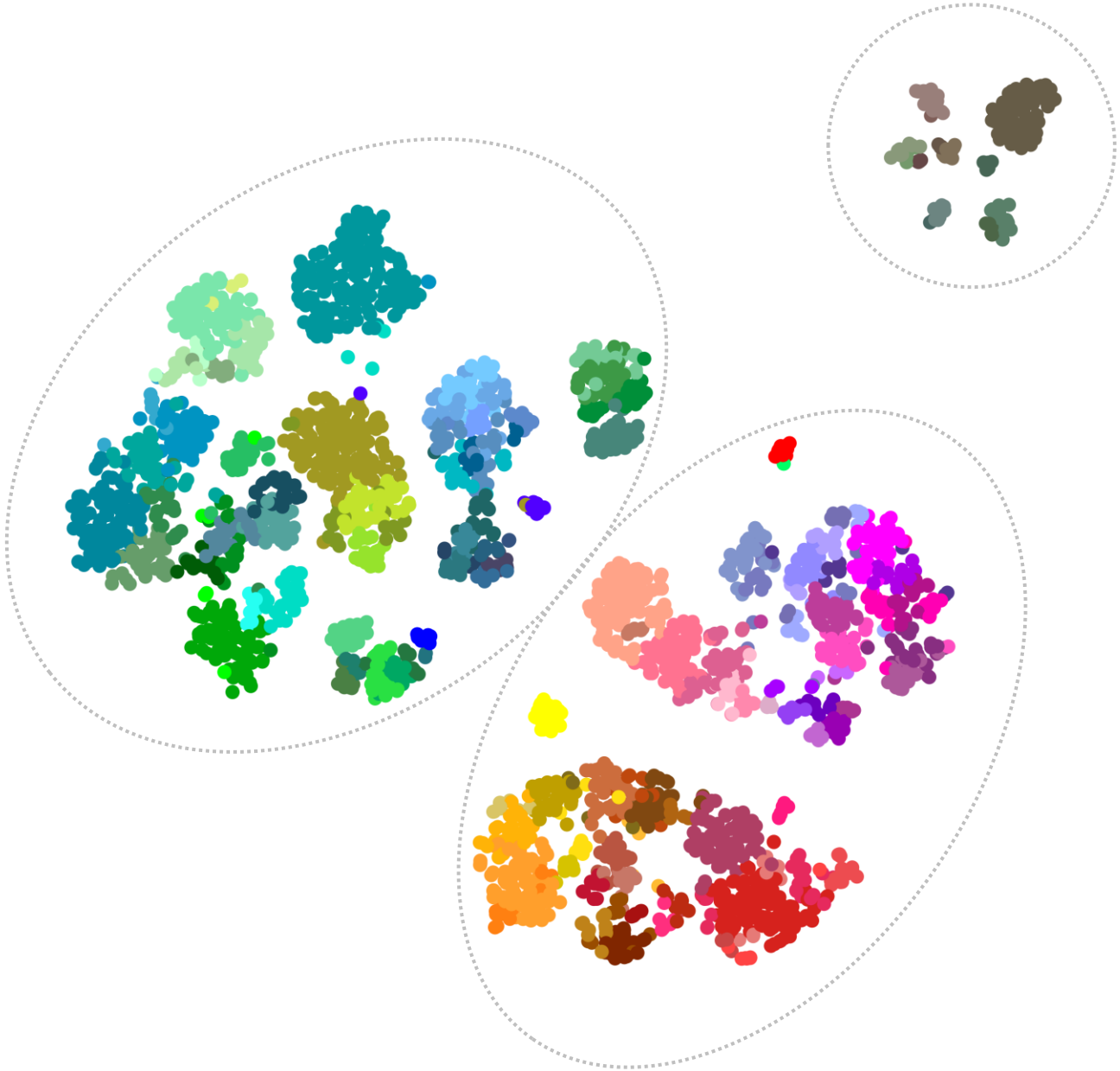
Dim = ~ 9500

Preprocessing [2]

Dim = 50 (HD)

Ms.*t*-SNE

Dim = 2 (LD)



Excitatory neurons

Inhibitory neurons

Non neural cells

Mouse cortex dataset [1]

Single cell RNA sequencing

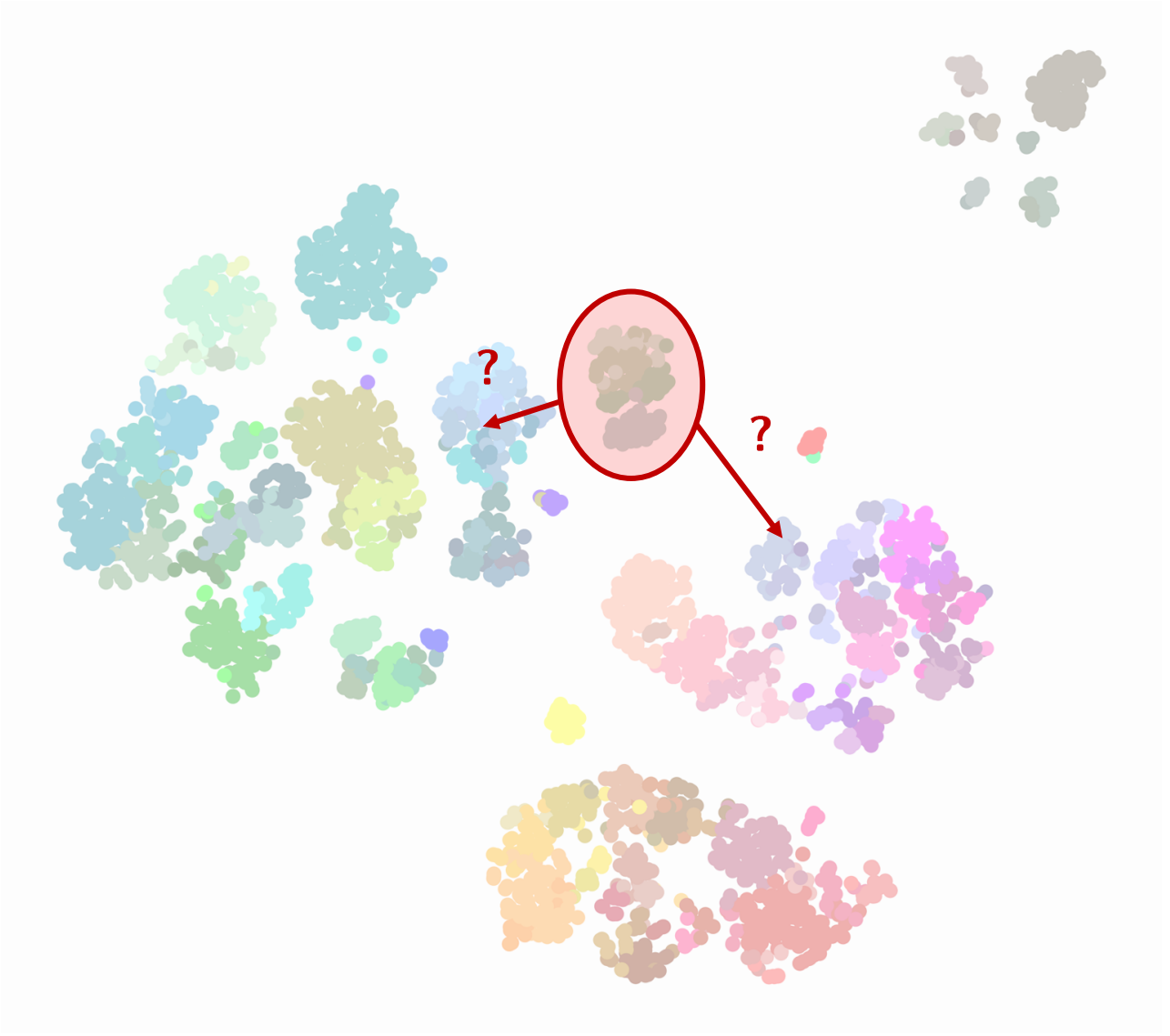
Dim = ~ 9500

Preprocessing [2]

Dim = 50 (HD)

Ms.*t*-SNE

Dim = 2 (LD)



Excitatory neurons

Inhibitory neurons

Non neural cells

Mouse cortex dataset [1]

Single cell RNA sequencing

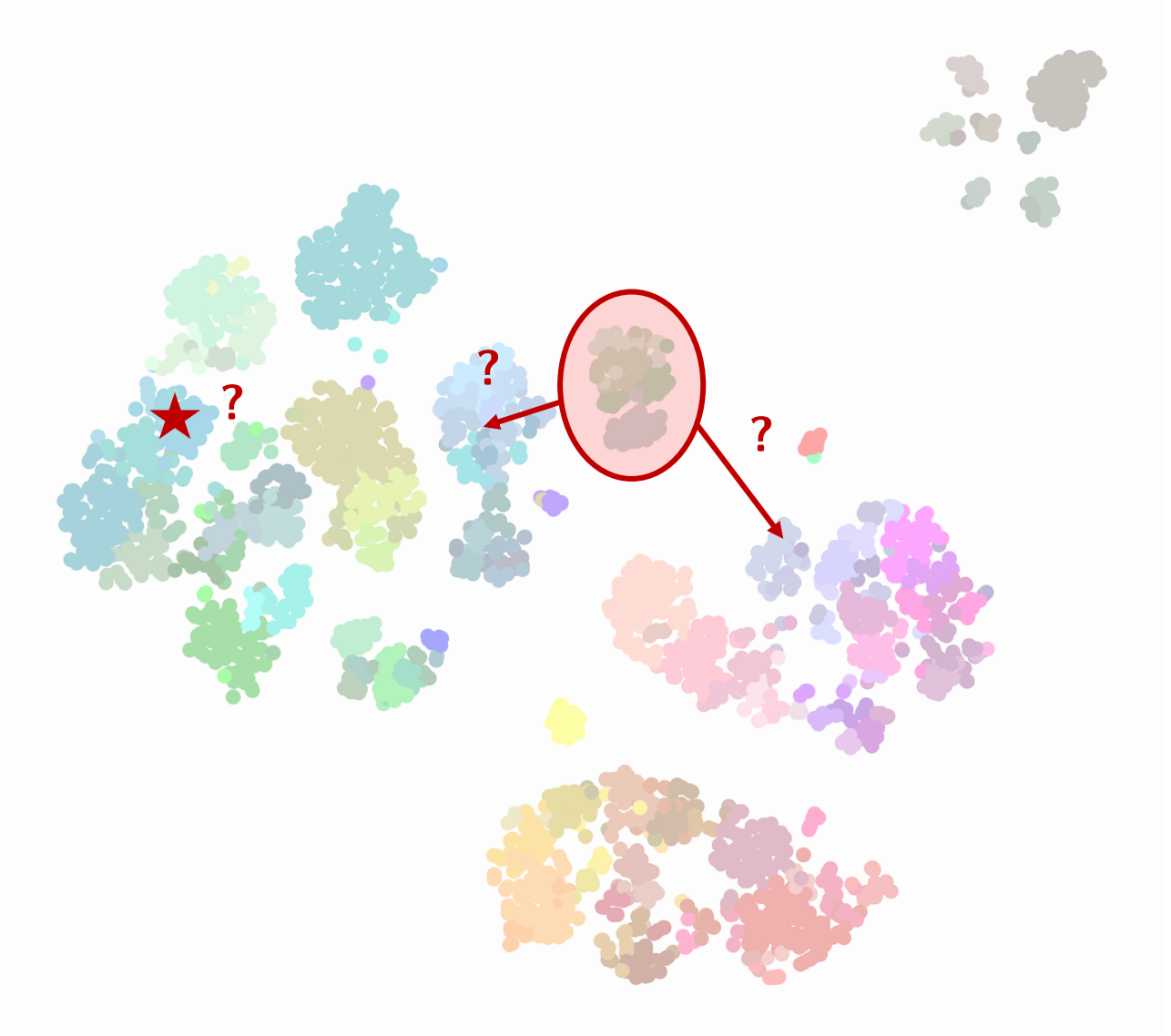
Dim = ~ 9500

Preprocessing [2]

Dim = 50 (HD)

Ms.*t*-SNE

Dim = 2 (LD)



Excitatory neurons

Inhibitory neurons

Non neural cells

Mouse cortex dataset [1]

Single cell RNA sequencing

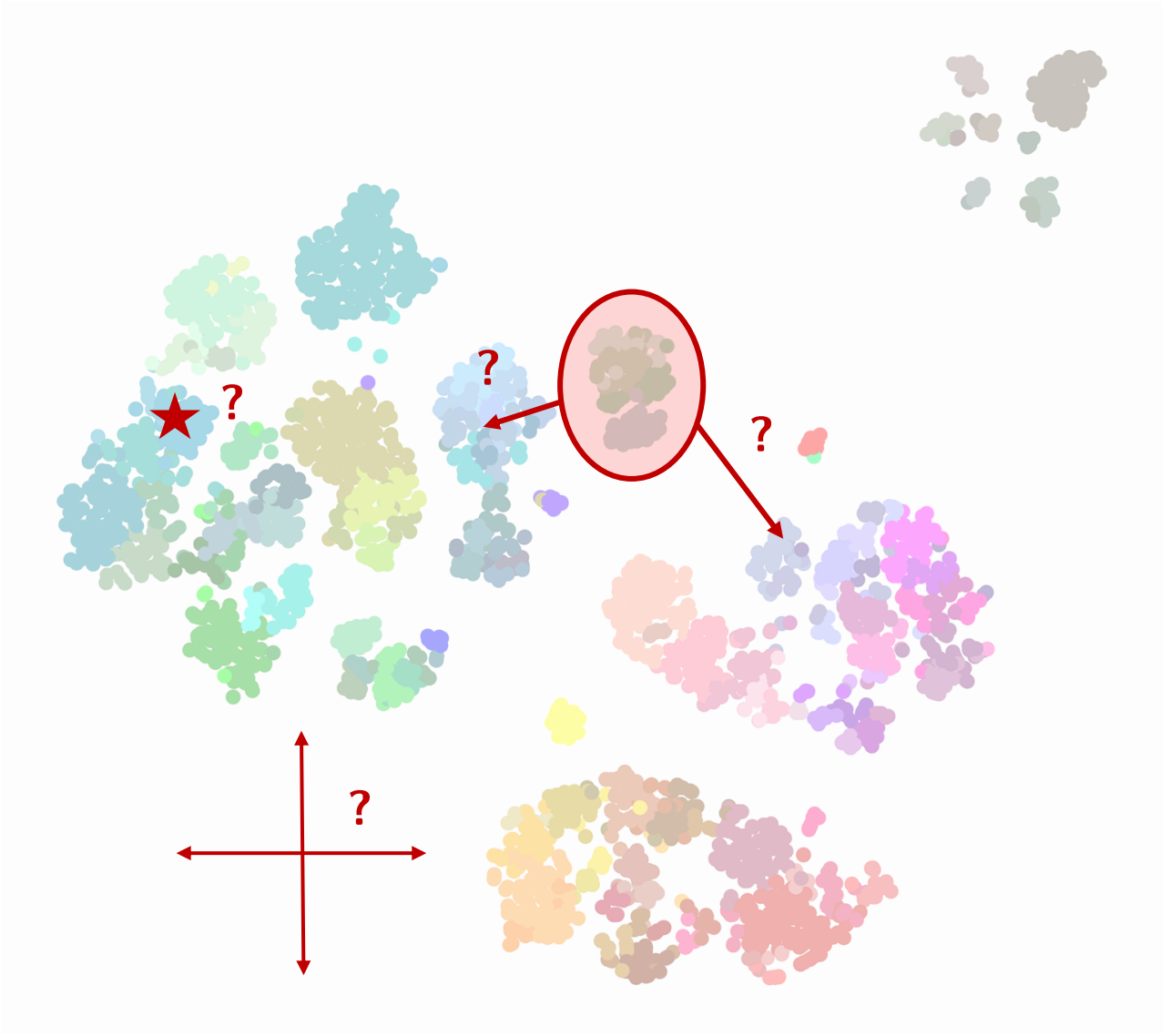
Dim = ~ 9500

Preprocessing [2]

Dim = 50 (HD)

Ms.*t*-SNE

Dim = 2 (LD)



Excitatory neurons

Inhibitory neurons

Non neural cells

M-dim space

Data

$$\{\xi_i\}_{i=1}^{N}$$

M-dim space

Data

$$\{\xi_i\}_{i=1}^N$$

2-dim space

Initialization

$$\{x_i\}_{i=1}^N$$

M-dim space

Data

$$\{\xi_i\}_{i=1}^{N}$$

Similarities

$$\tau_{ij}$$

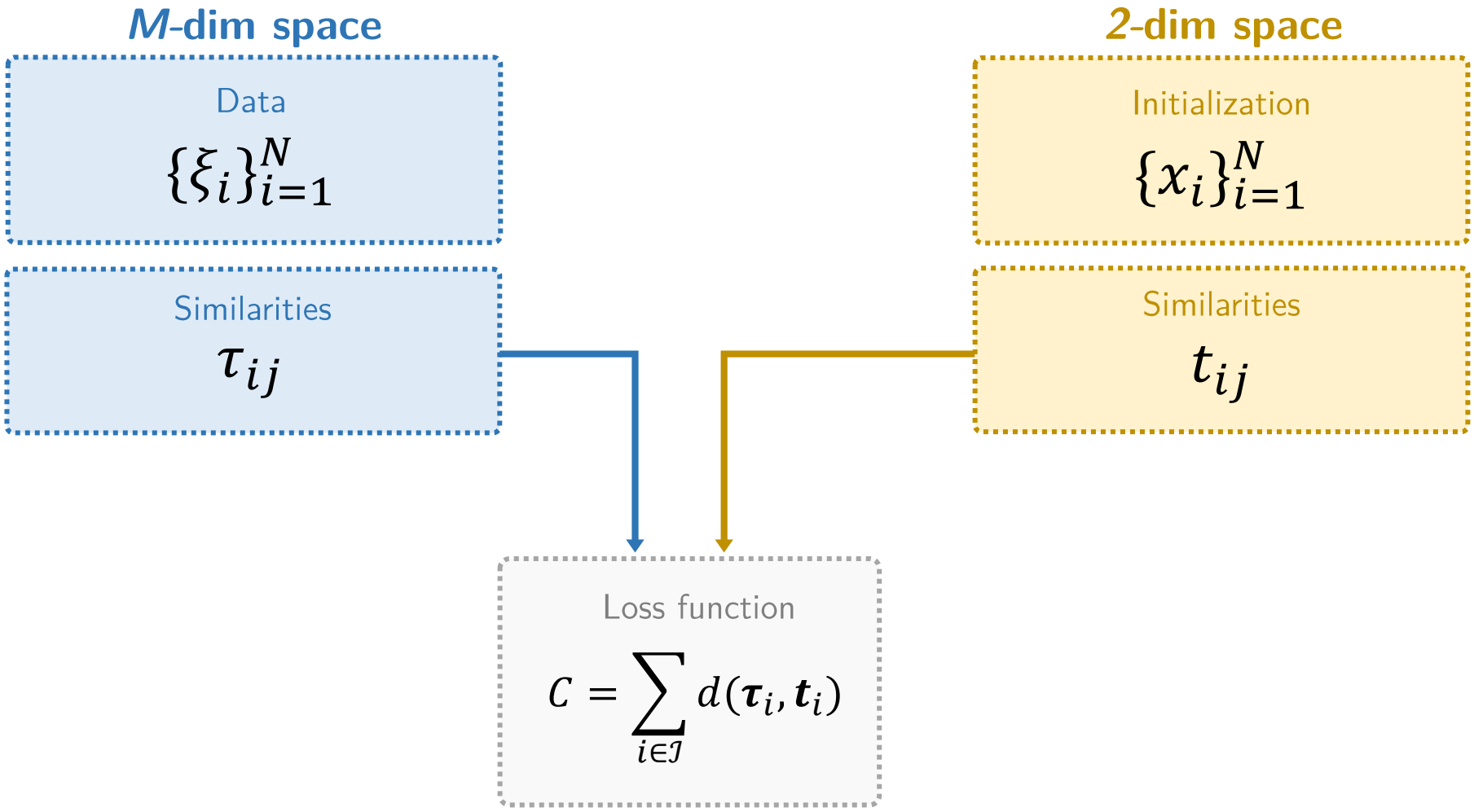
2-dim space

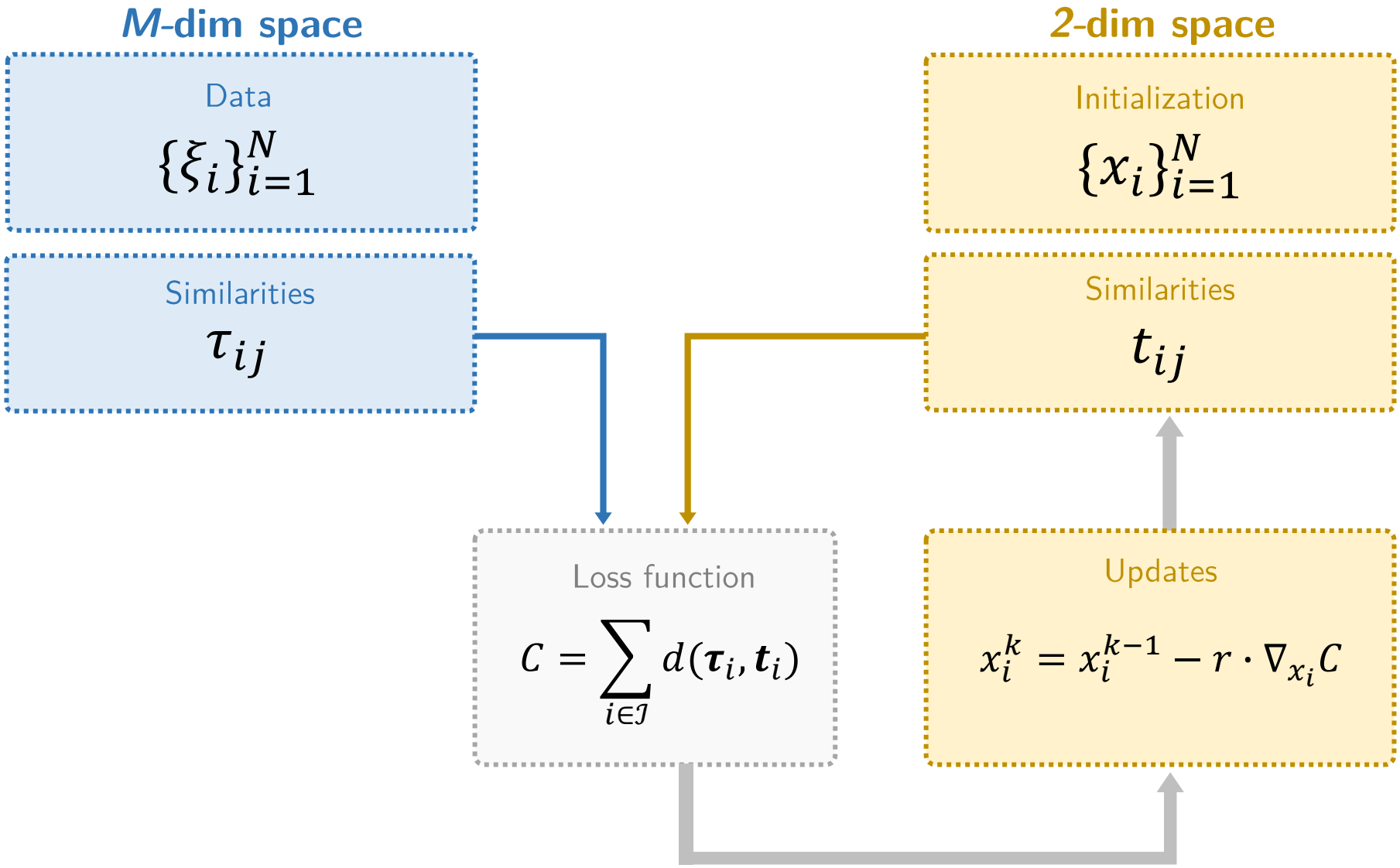
Initialization

$$\{x_i\}_{i=1}^{N}$$

Similarities

$$t_{ij}$$

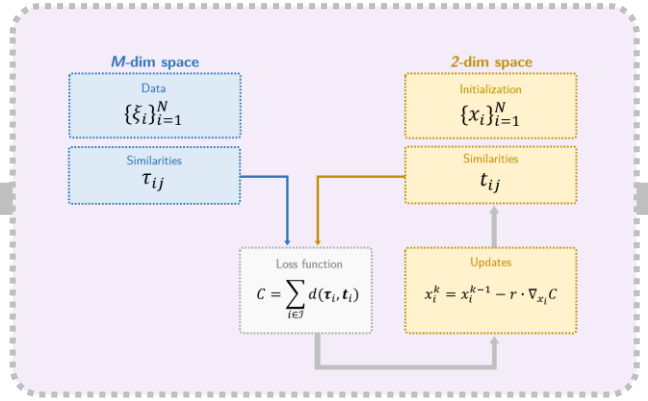

M-dim space
Data
$\{\xi_i\}_{i=1}^N$
Similarities
τ_{ij}
2-dim space
Initialization
$\{x_i\}_{i=1}^N$
Similarities
t_{ij}
Loss function
$C = \sum_{i \in \mathcal{I}} d(\boldsymbol{\tau}_i, \boldsymbol{t}_i)$


M-dim space
Data
$\{\xi_i\}_{i=1}^N$
Similarities
τ_{ij}
2-dim space
Initialization
$\{x_i\}_{i=1}^N$
Similarities
t_{ij}
Loss function
$C = \sum_{i \in \mathcal{J}} d(\boldsymbol{\tau}_i, \boldsymbol{t}_i)$
Updates
$x_i^k = x_i^{k-1} - r \cdot \nabla_{x_i} C$

HD Features

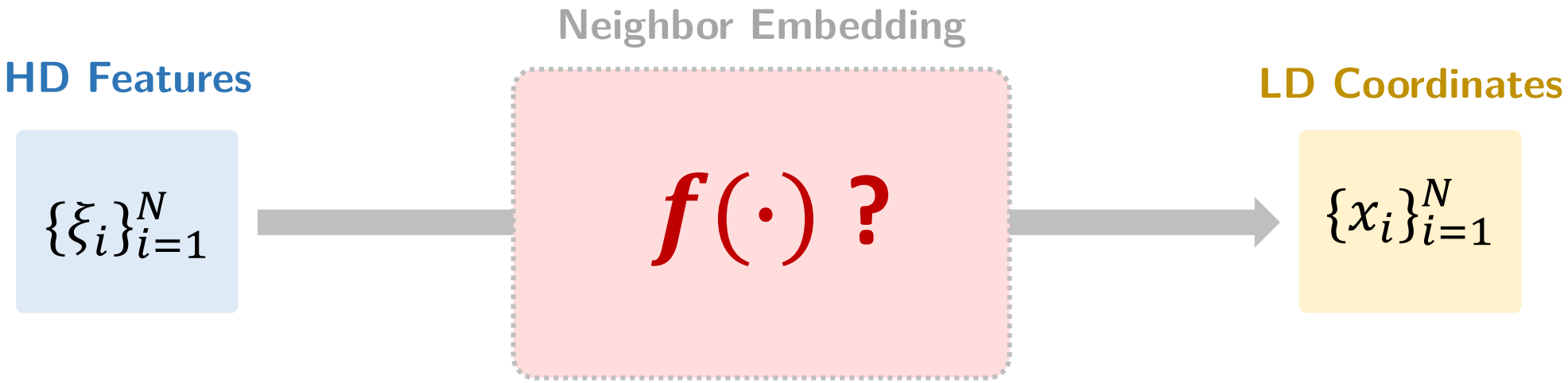
$$\{\xi_i\}_{i=1}^N$$

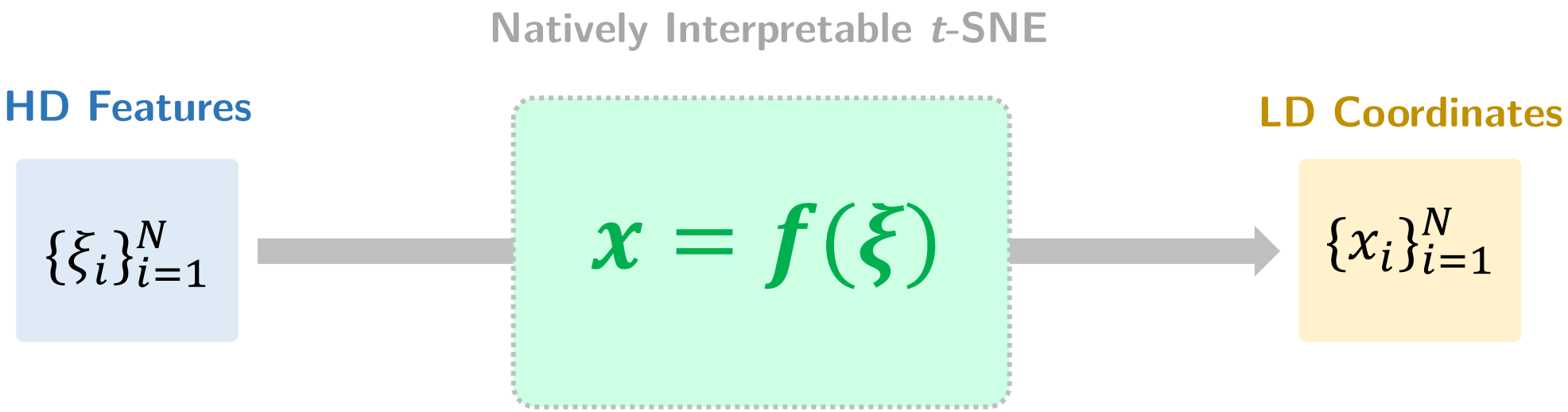
Neighbor Embedding



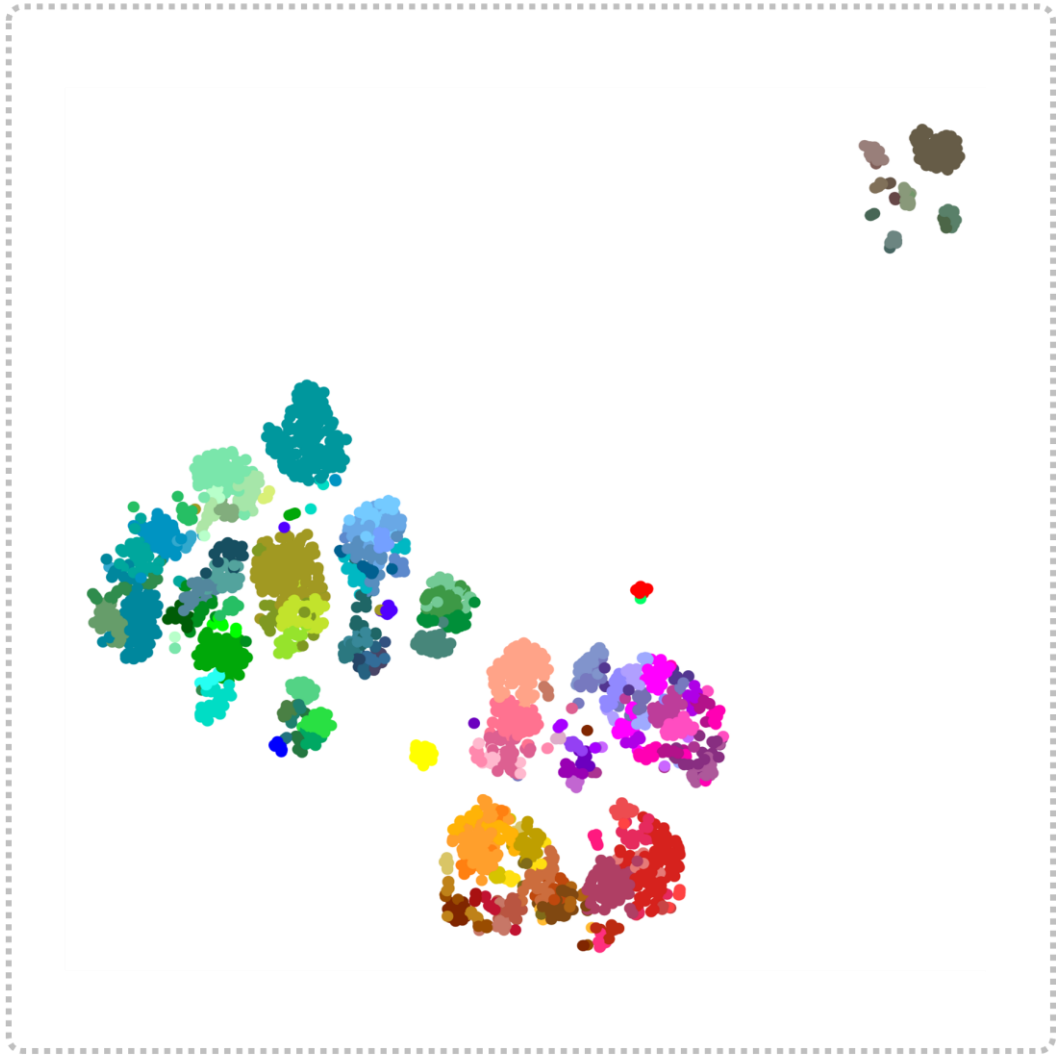
LD Coordinates

$$\{x_i\}_{i=1}^N$$


Neighbor Embedding
HD Features
LD Coordinates
$\{\xi_i\}_{i=1}^N$
$f(\cdot)$?
$\{x_i\}_{i=1}^N$

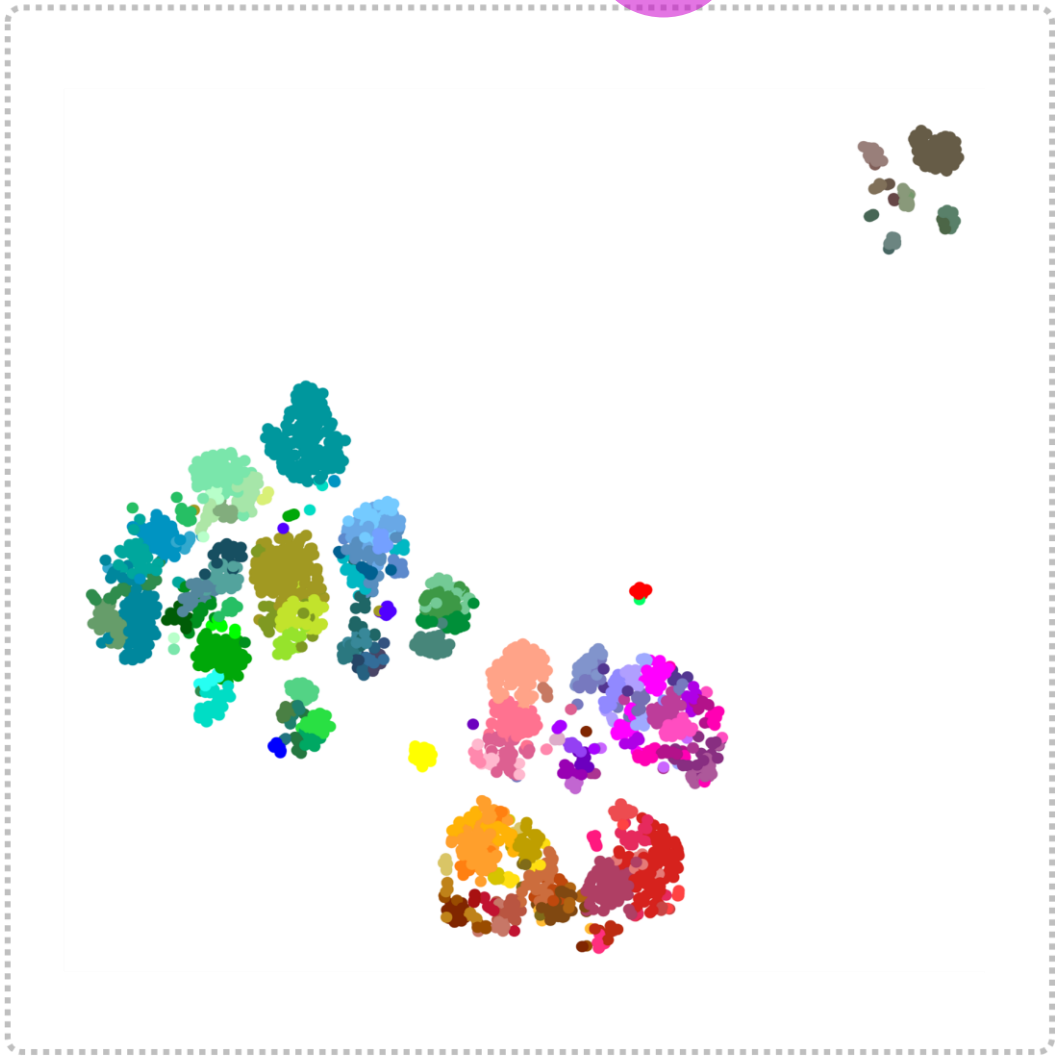

Natively Interpretable t-SNE
HD Features
$\{\xi_i\}_{i=1}^N$
$x = f(\xi)$
LD Coordinates
$\{x_i\}_{i=1}^N$

Ms. *t*-SNE [-]

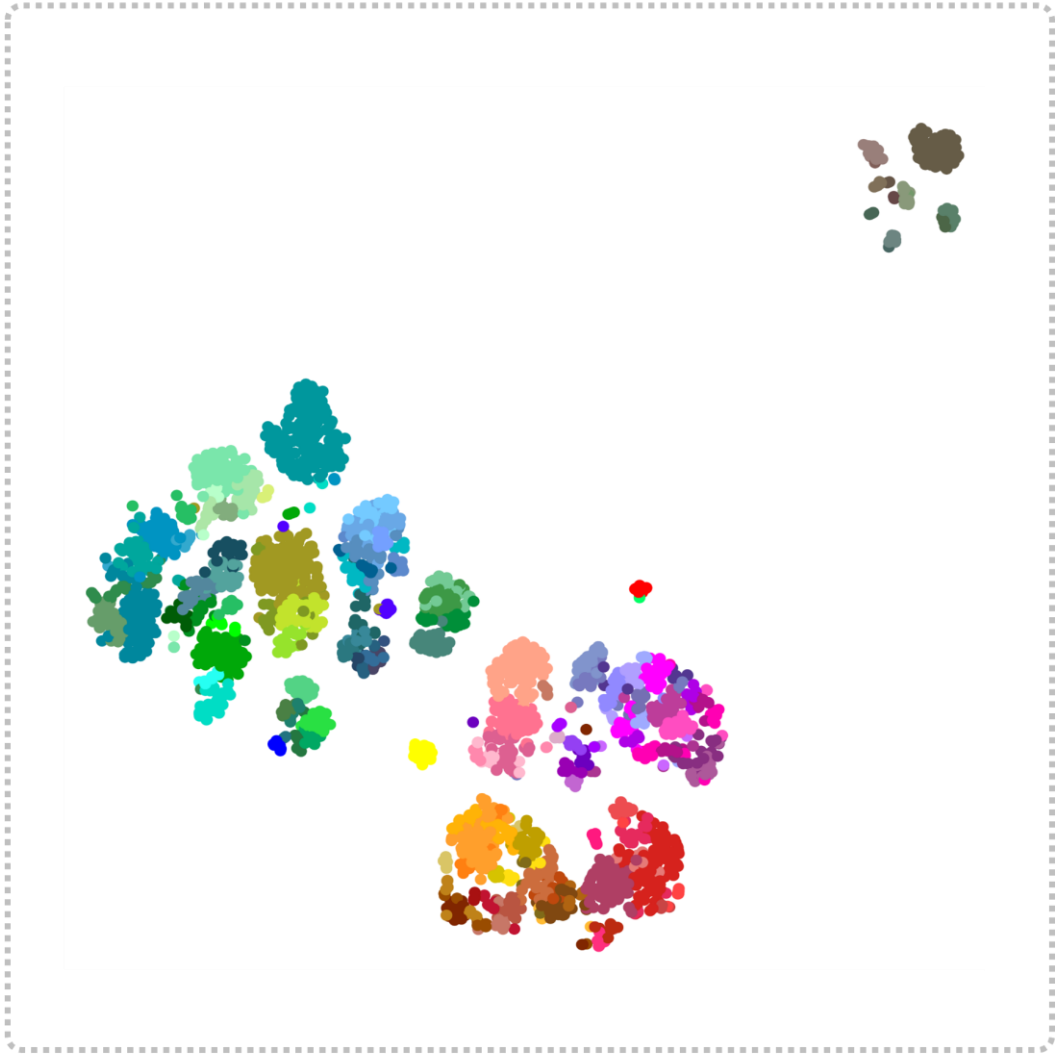


Ms. *t*-SNE [-]

Nonparametric

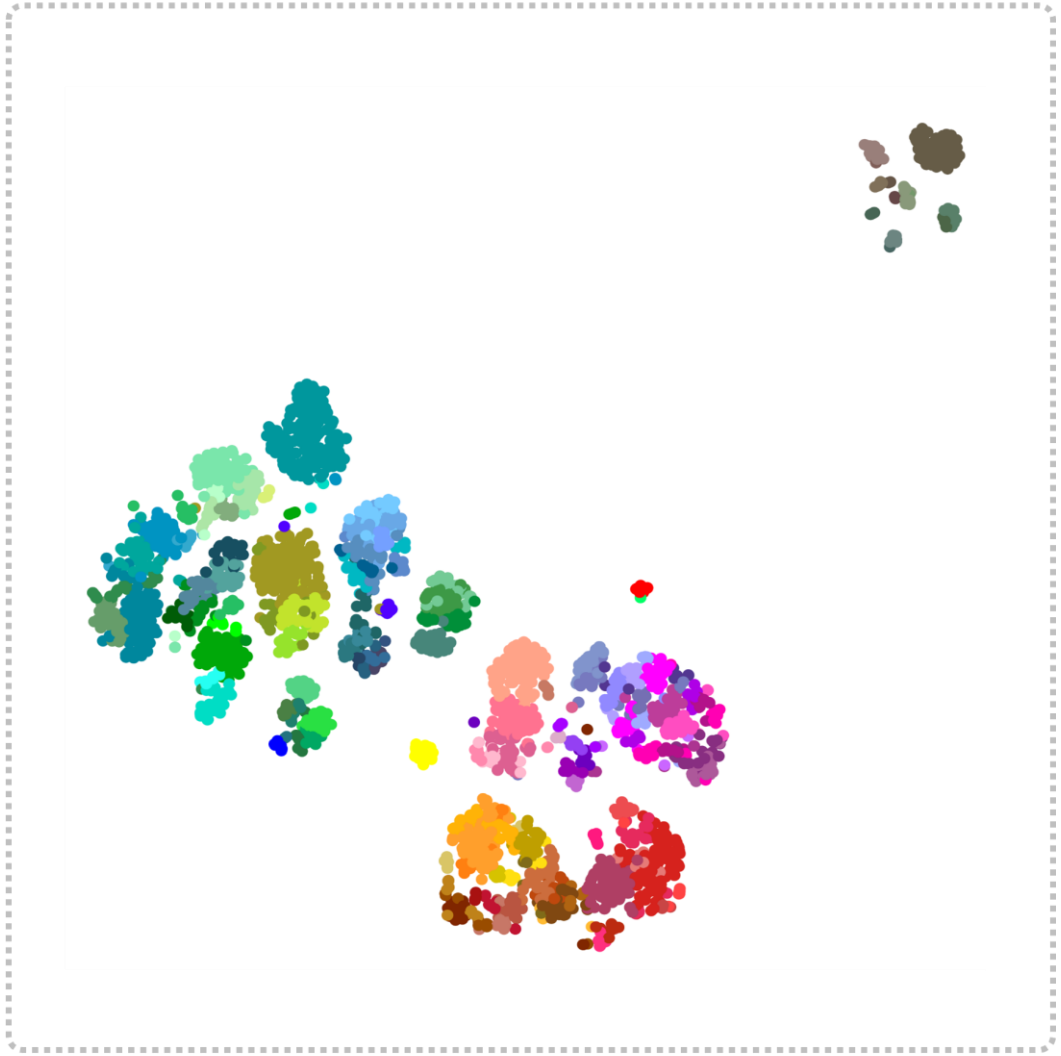


Ms. *t*-SNE [-]



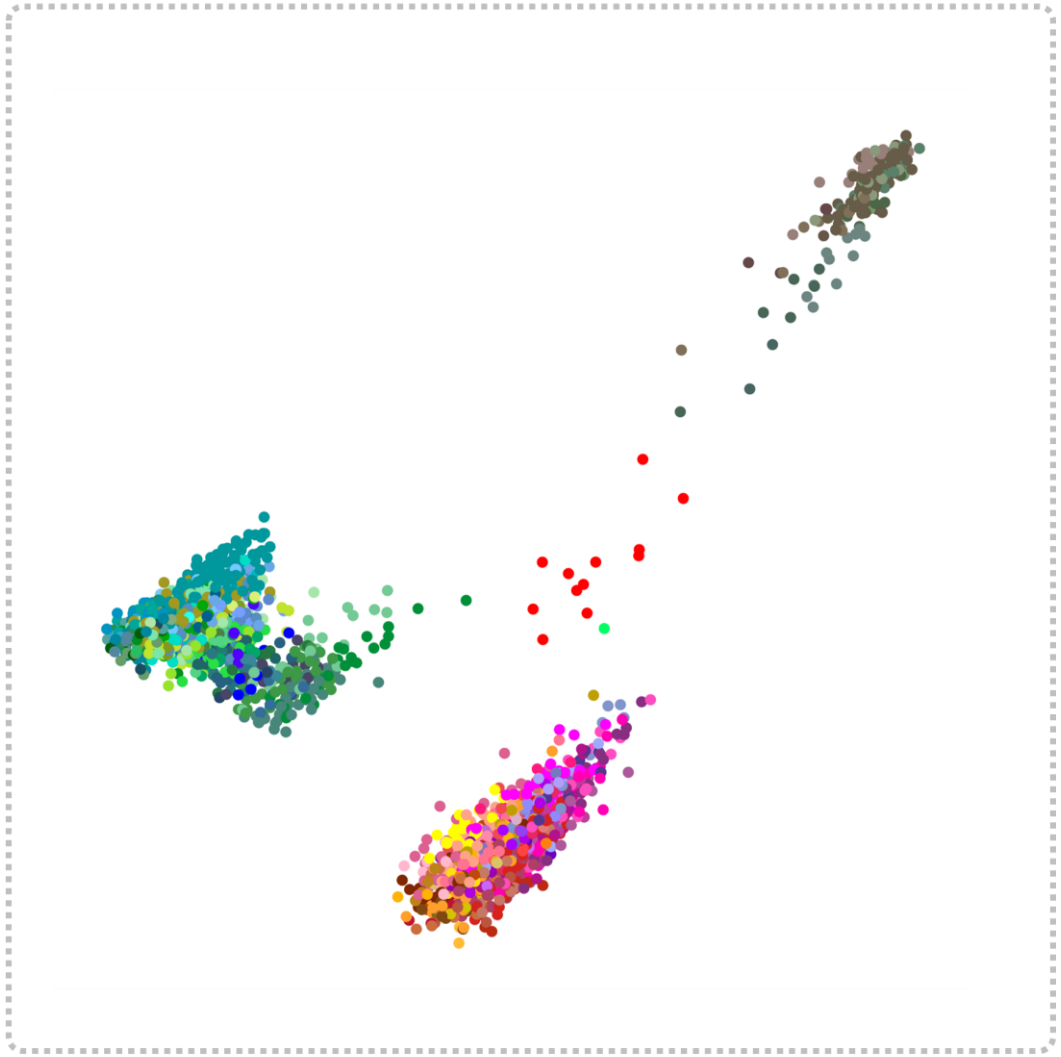
Interpretability

Ms. *t*-SNE [-]



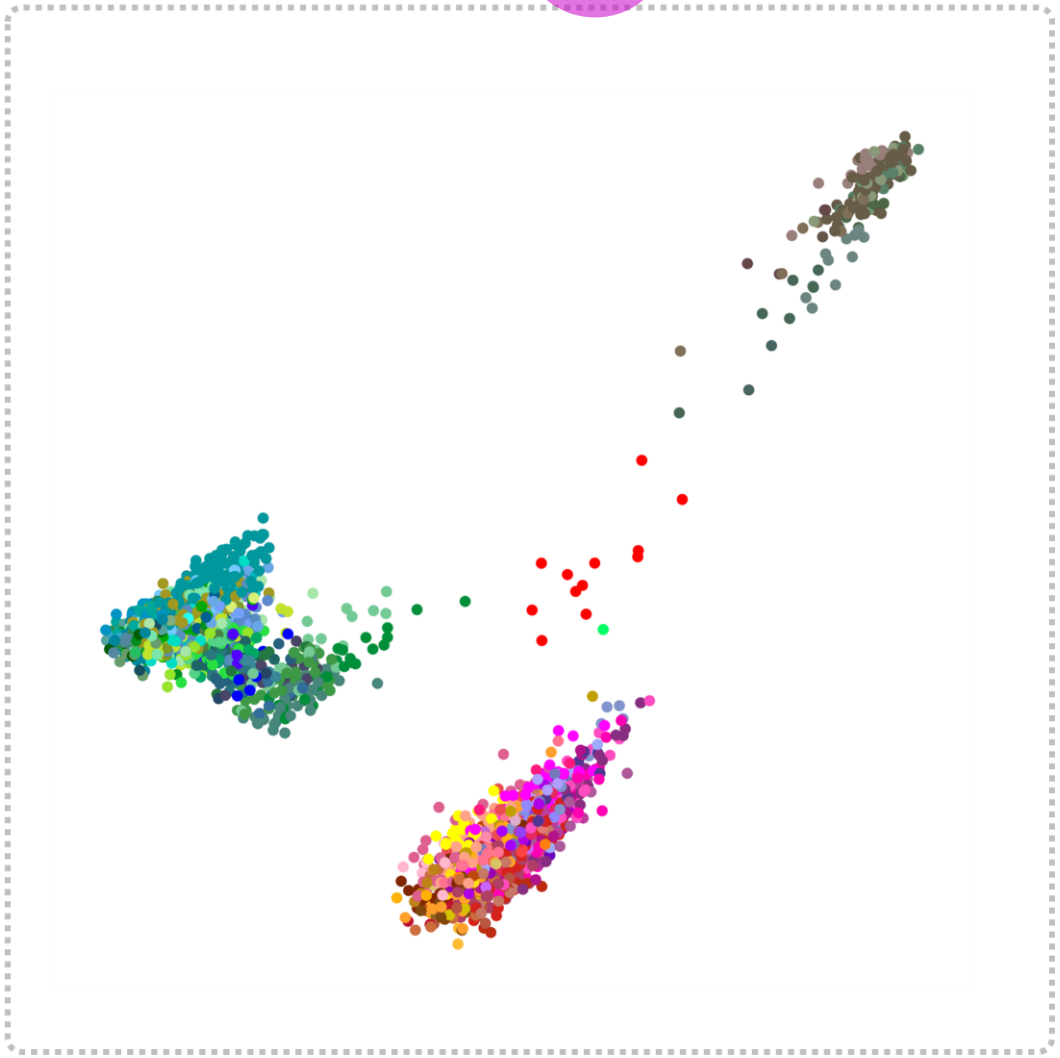
Interpretability − − Quality + +

PCA [W]

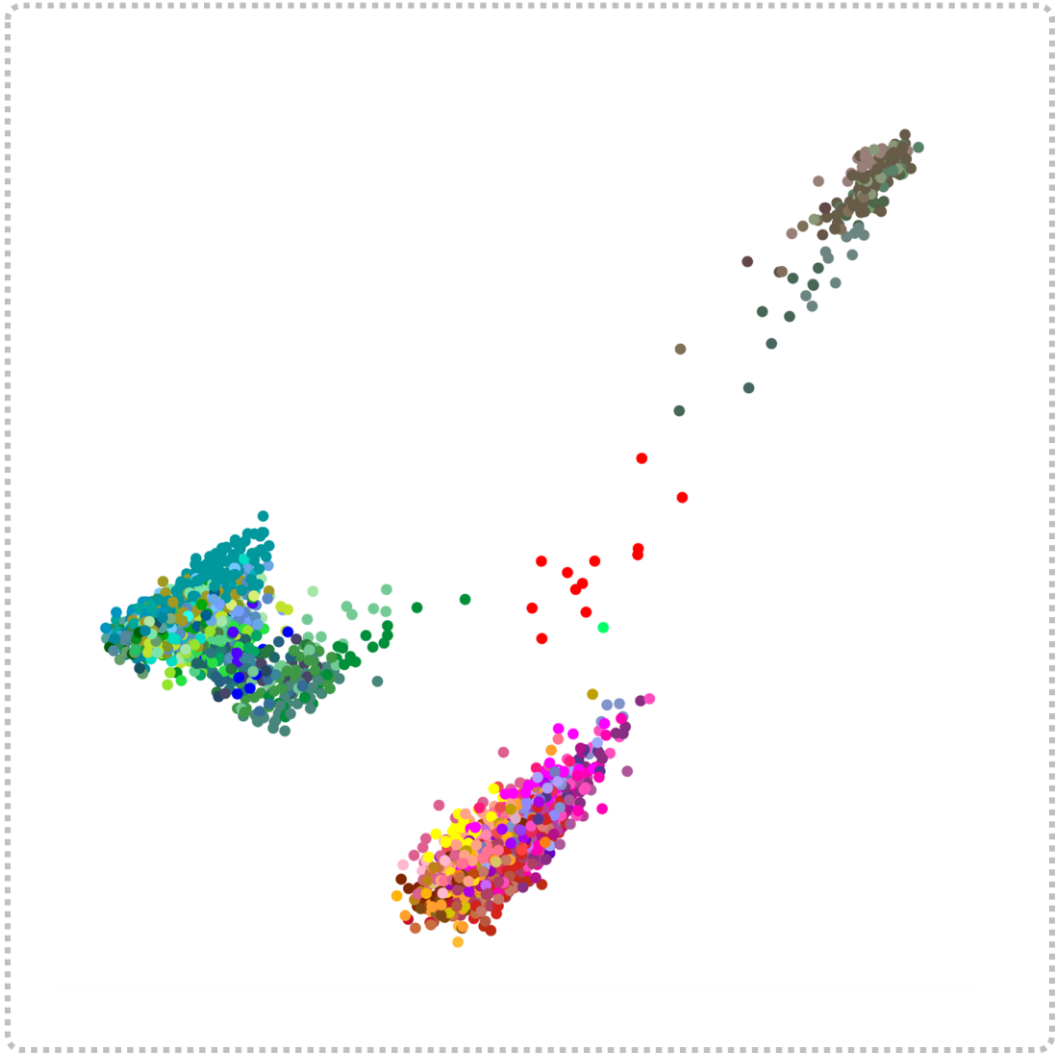


PCA [W]

Parametric

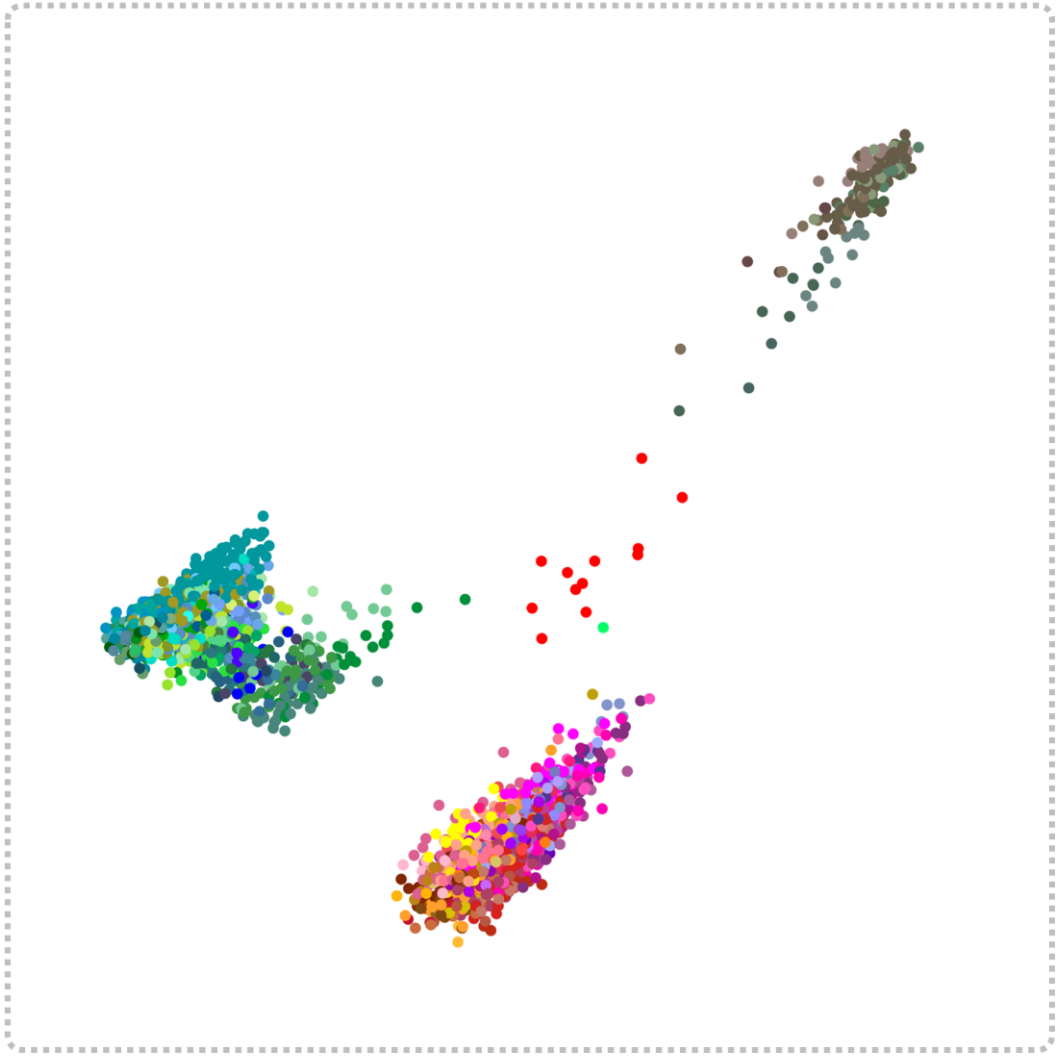


PCA [W]



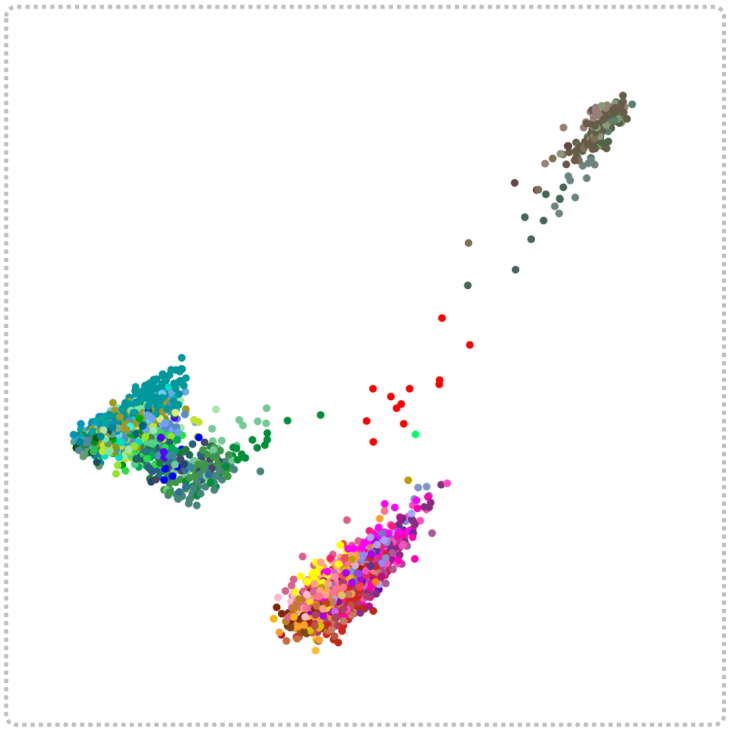
Interpretability + +

PCA [W]



Interpretability + + Quality − −

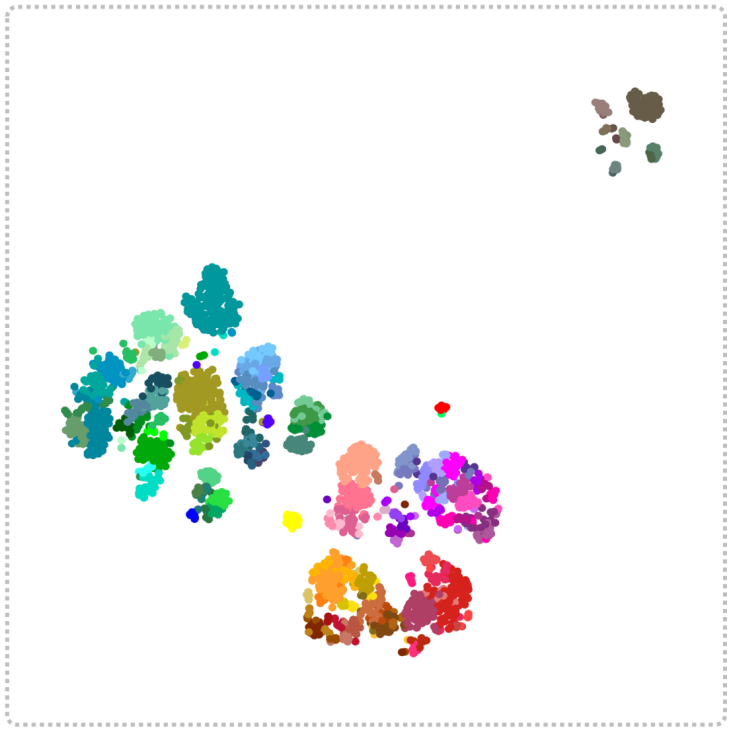
PCA [W]



Interpretability + + Quality − −

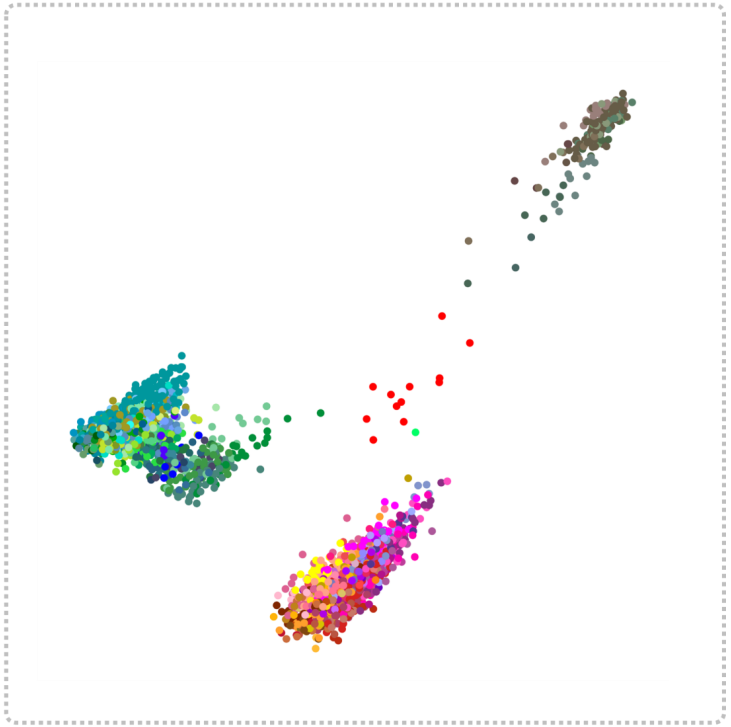
?

Ms. *t*-SNE [-]



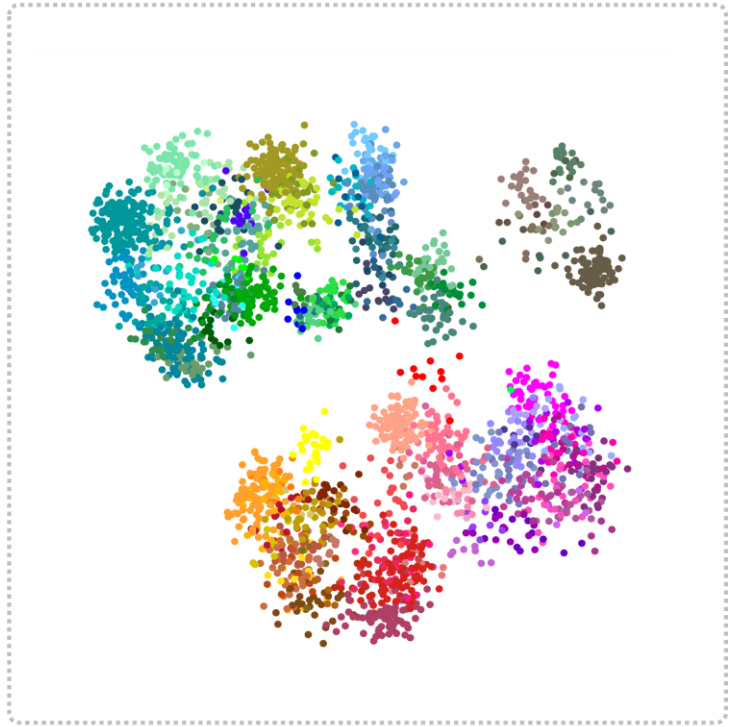
Interpretability − − Quality + +

PCA [W]

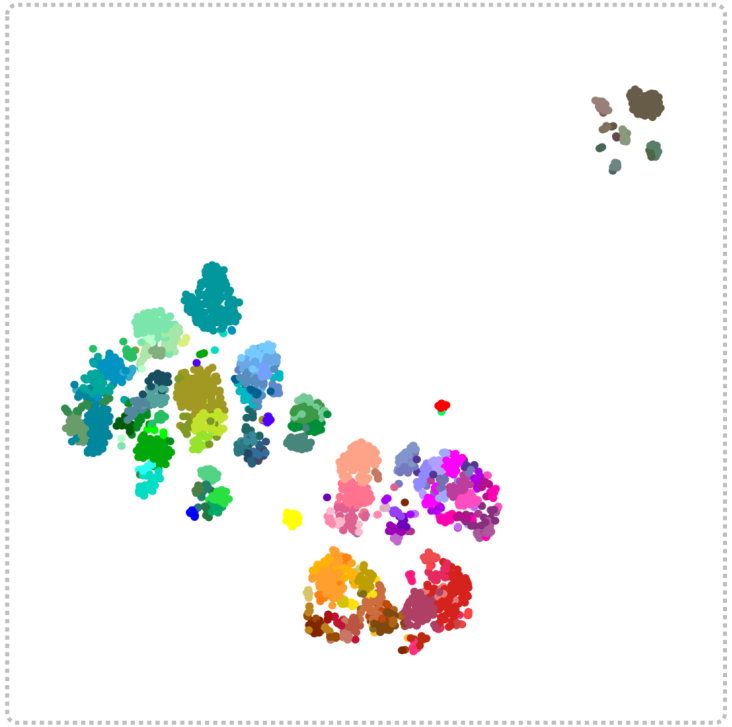


Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]

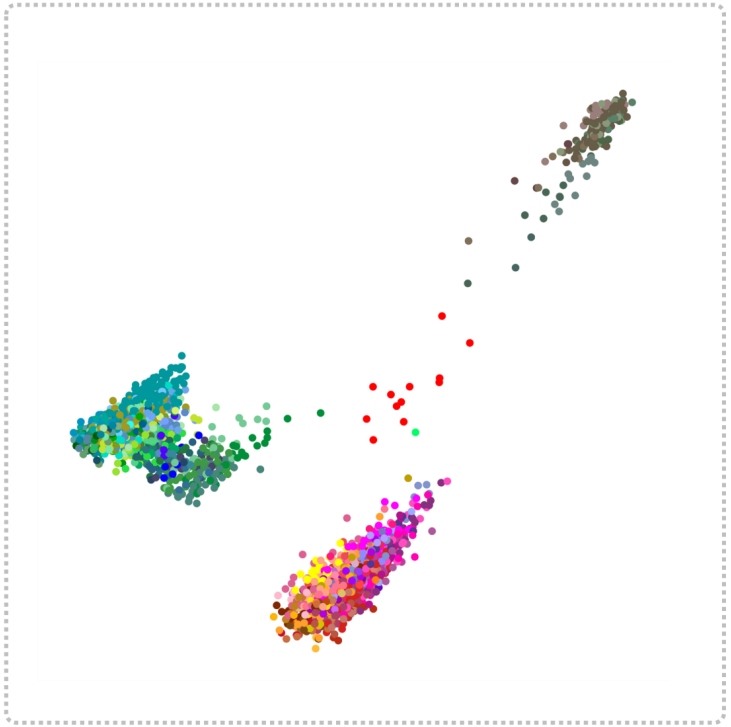


Ms. *t*-SNE [-]



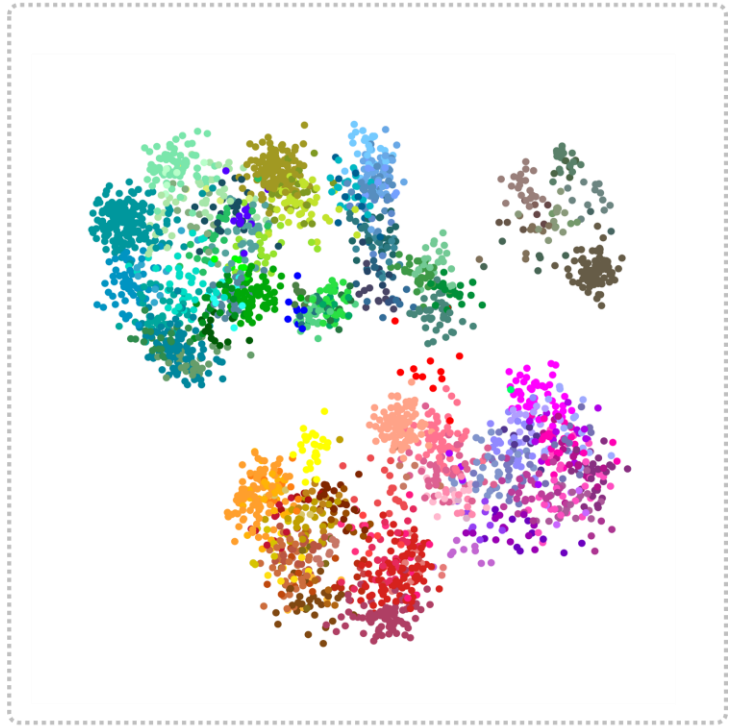
Interpretability ⊖ ⊖ Quality ⊕ ⊕

PCA [W]



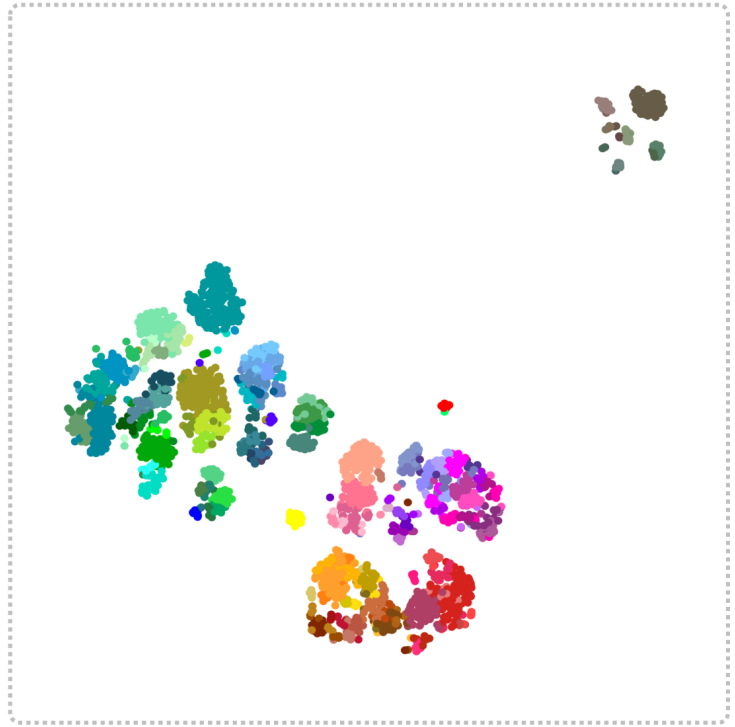
Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]



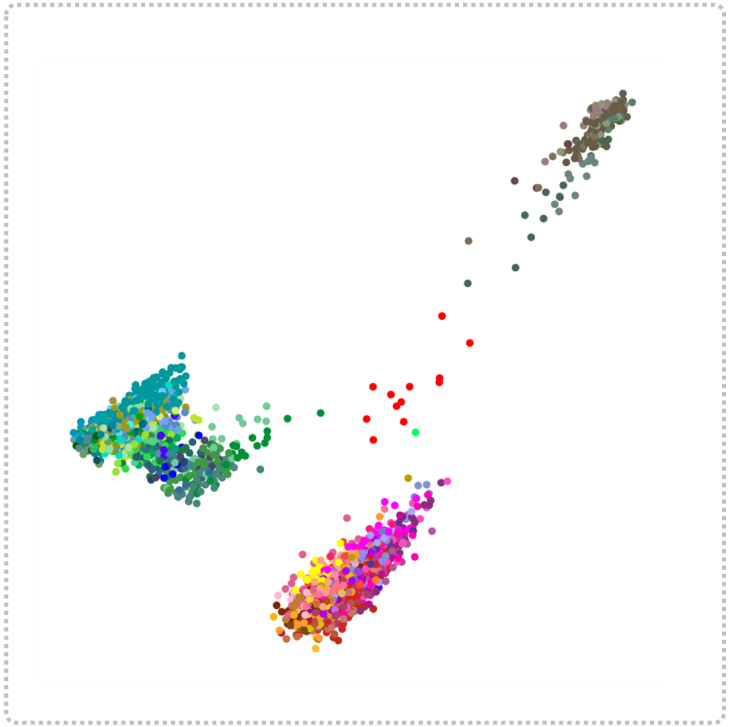
Interpretability ⊕ ⊕

Ms. *t*-SNE [-]



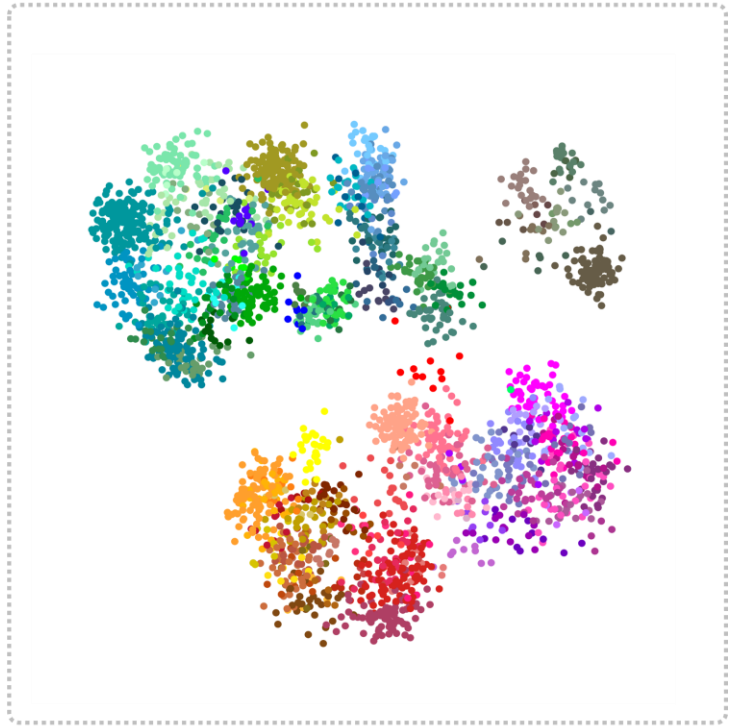
Interpretability ⊖ ⊖ Quality ⊕ ⊕

PCA [W]



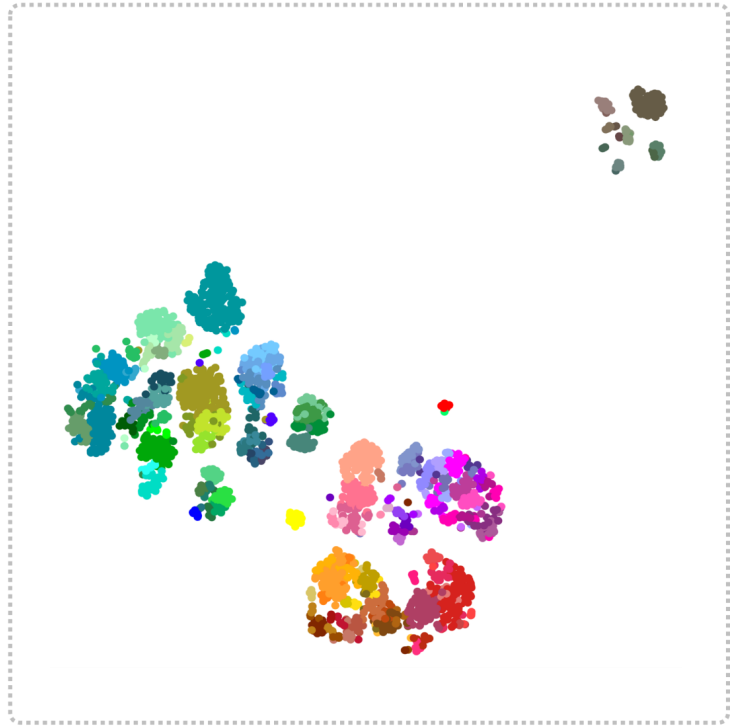
Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]



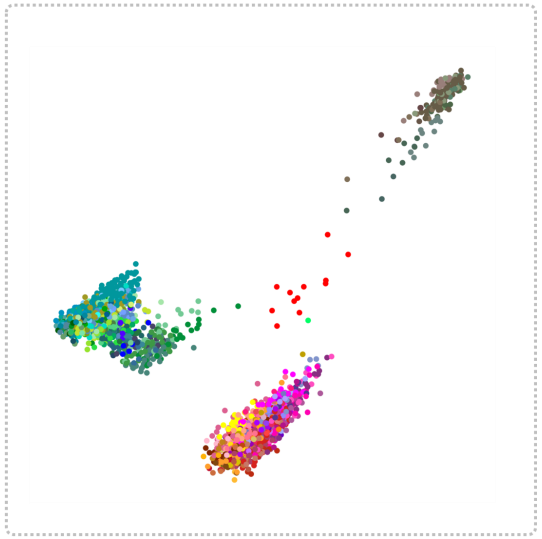
Interpretability ⊕ ⊕ Quality ⊖

Ms. *t*-SNE [-]



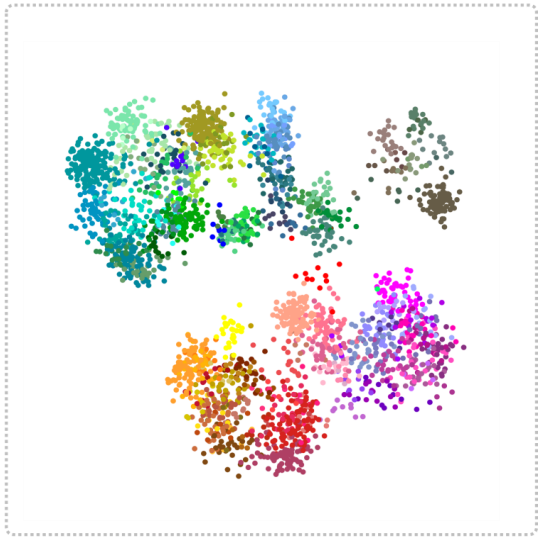
Interpretability ⊖ ⊖ Quality ⊕ ⊕

PCA [W]



Interpretability + + Quality − −

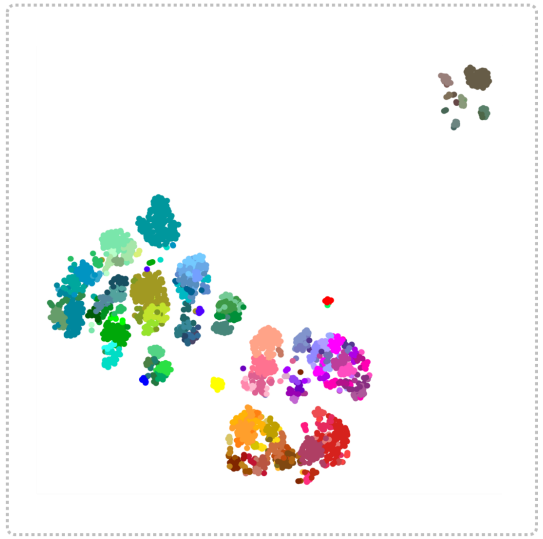
Ms. *t*-SNE [W]



Interpretability + + Quality −

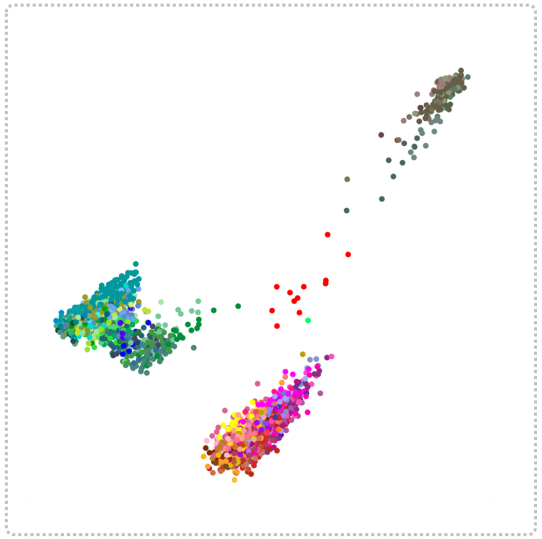
?

Ms. *t*-SNE [-]



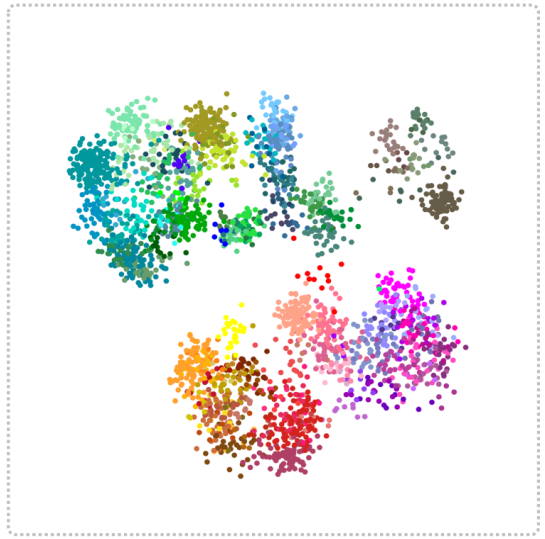
Interpretability − − Quality + +

PCA [W]



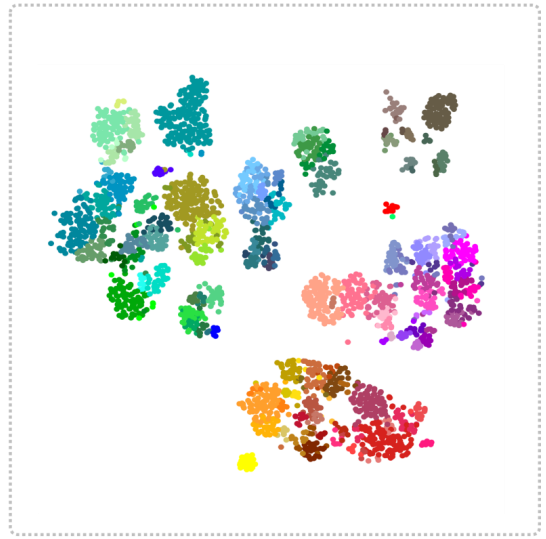
Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]

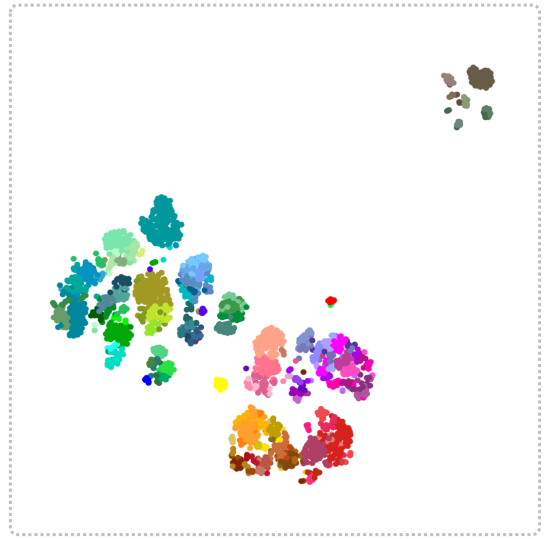


Interpretability ⊕ ⊕ Quality ⊖

Ms. *t*-SNE [W_i]

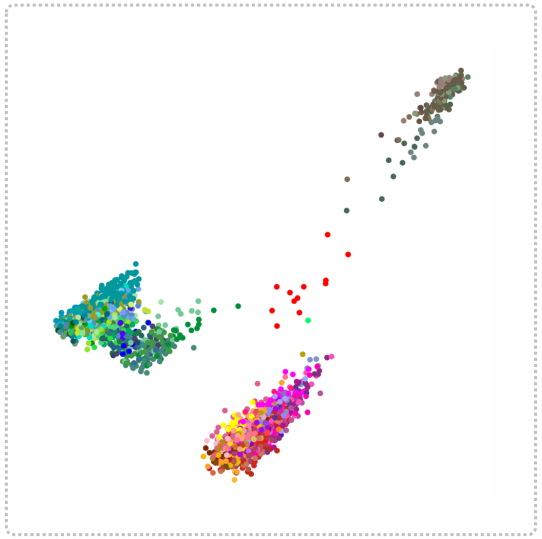


Ms. *t*-SNE [-]



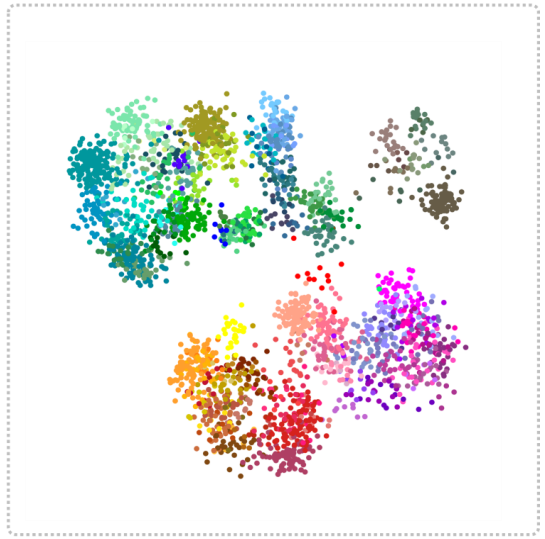
Interpretability ⊖ ⊖ Quality ⊕ ⊕

PCA [W]



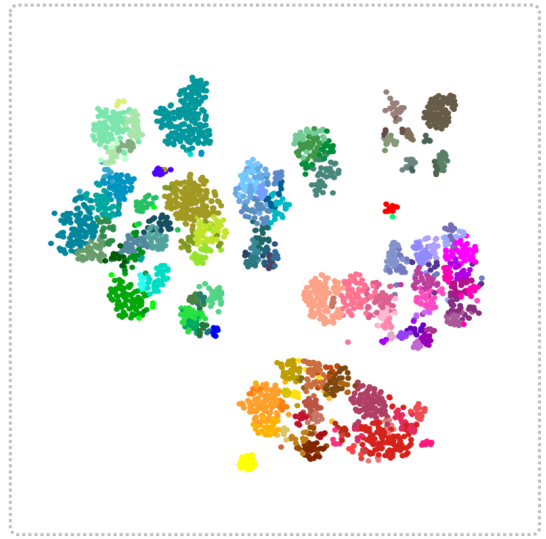
Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]



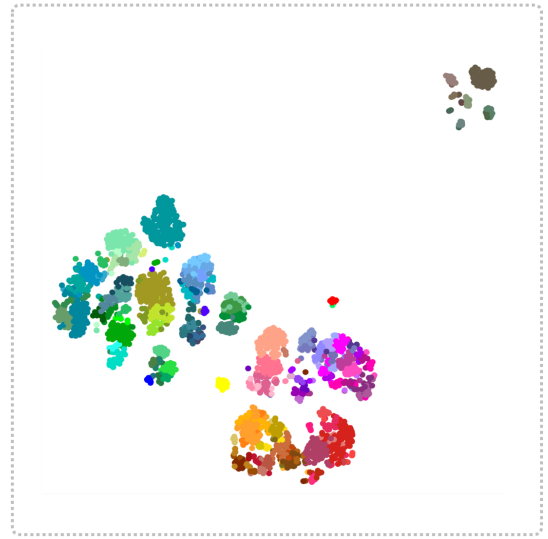
Interpretability ⊕ ⊕ Quality ⊖

Ms. *t*-SNE [W_i]



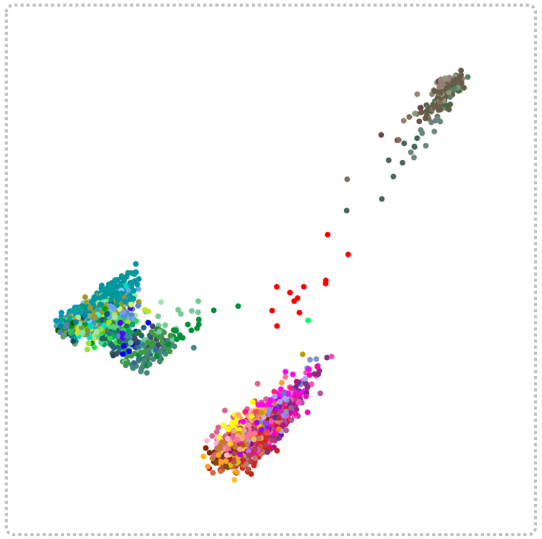
Interpretability ⊕

Ms. *t*-SNE [-]



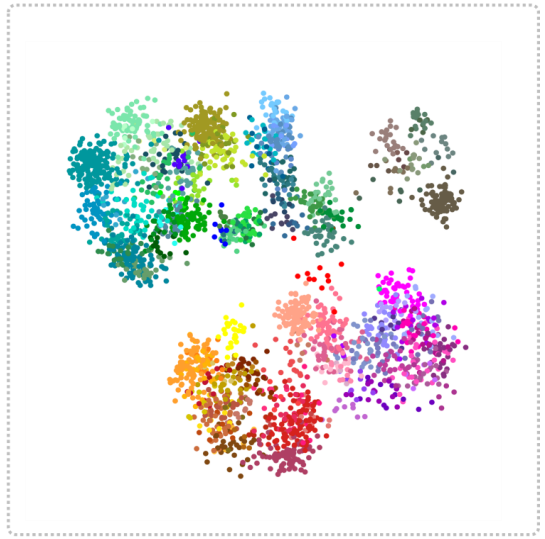
Interpretability ⊖ ⊖ Quality ⊕ ⊕

PCA [W]



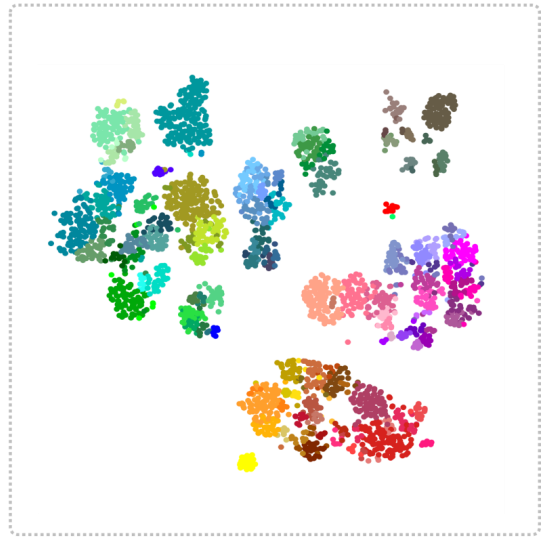
Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]



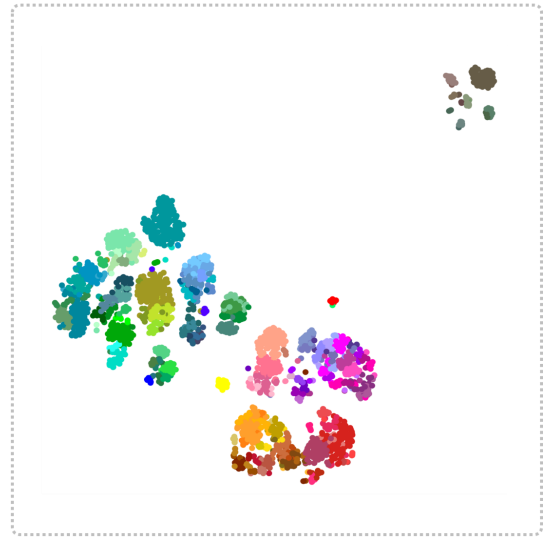
Interpretability ⊕ ⊕ Quality ⊖

Ms. *t*-SNE [W_i]

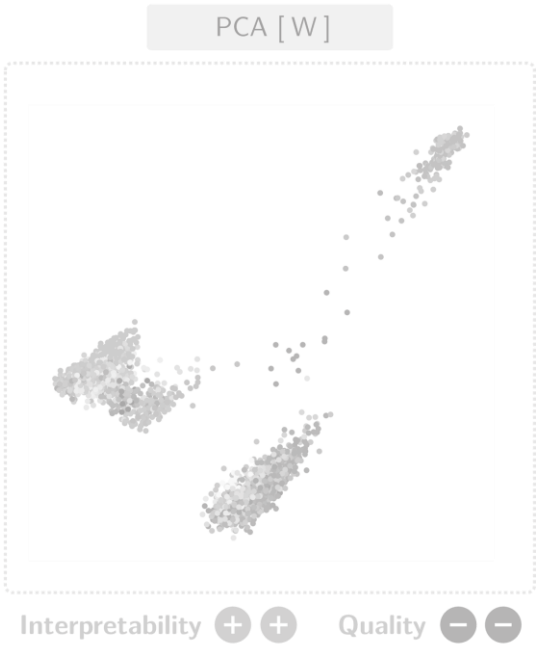


Interpretability ⊕ Quality ⊕

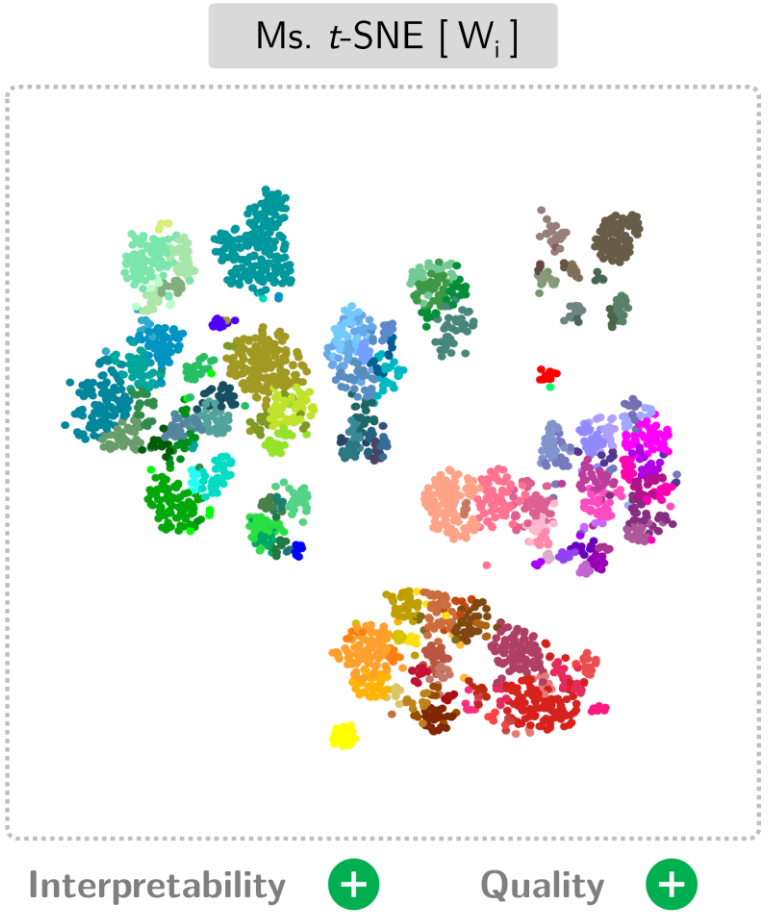
Ms. *t*-SNE [-]

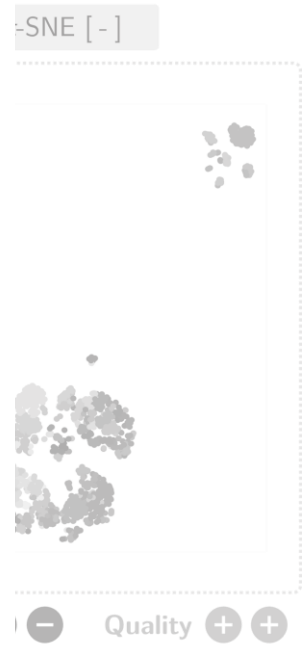


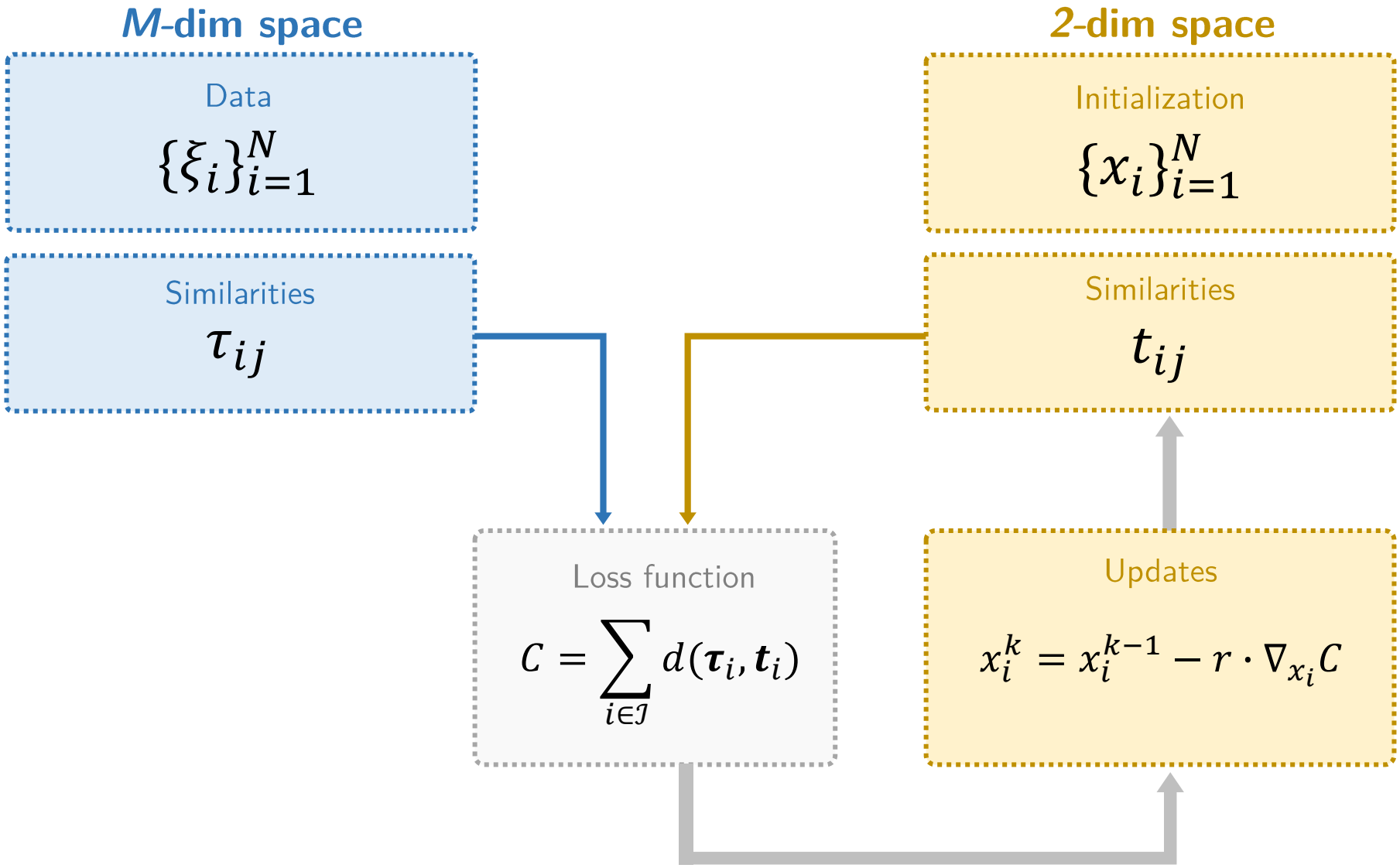
Interpretability ⊖ ⊖ Quality ⊕ ⊕

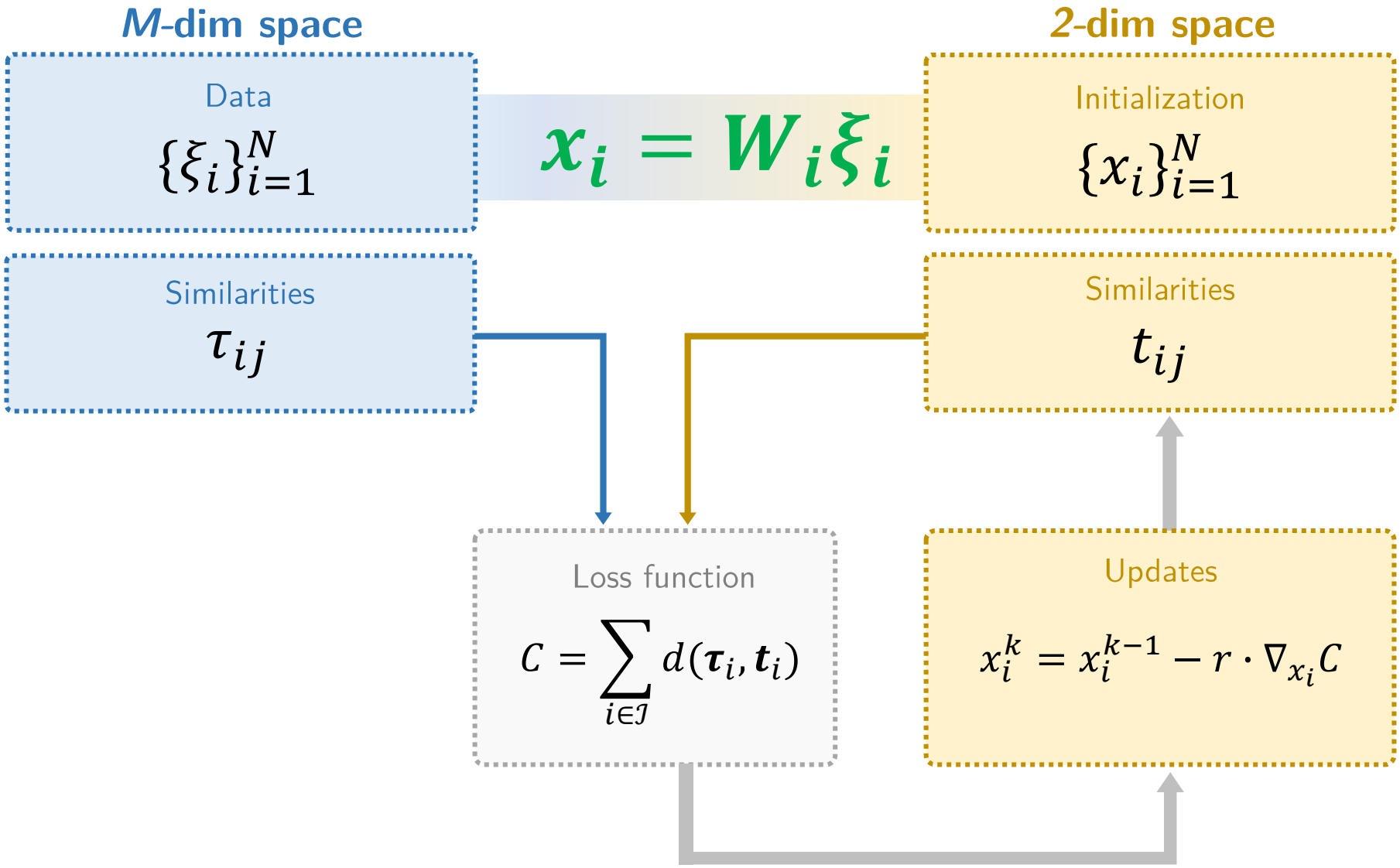

PCA [W]
Interpretability
Quality

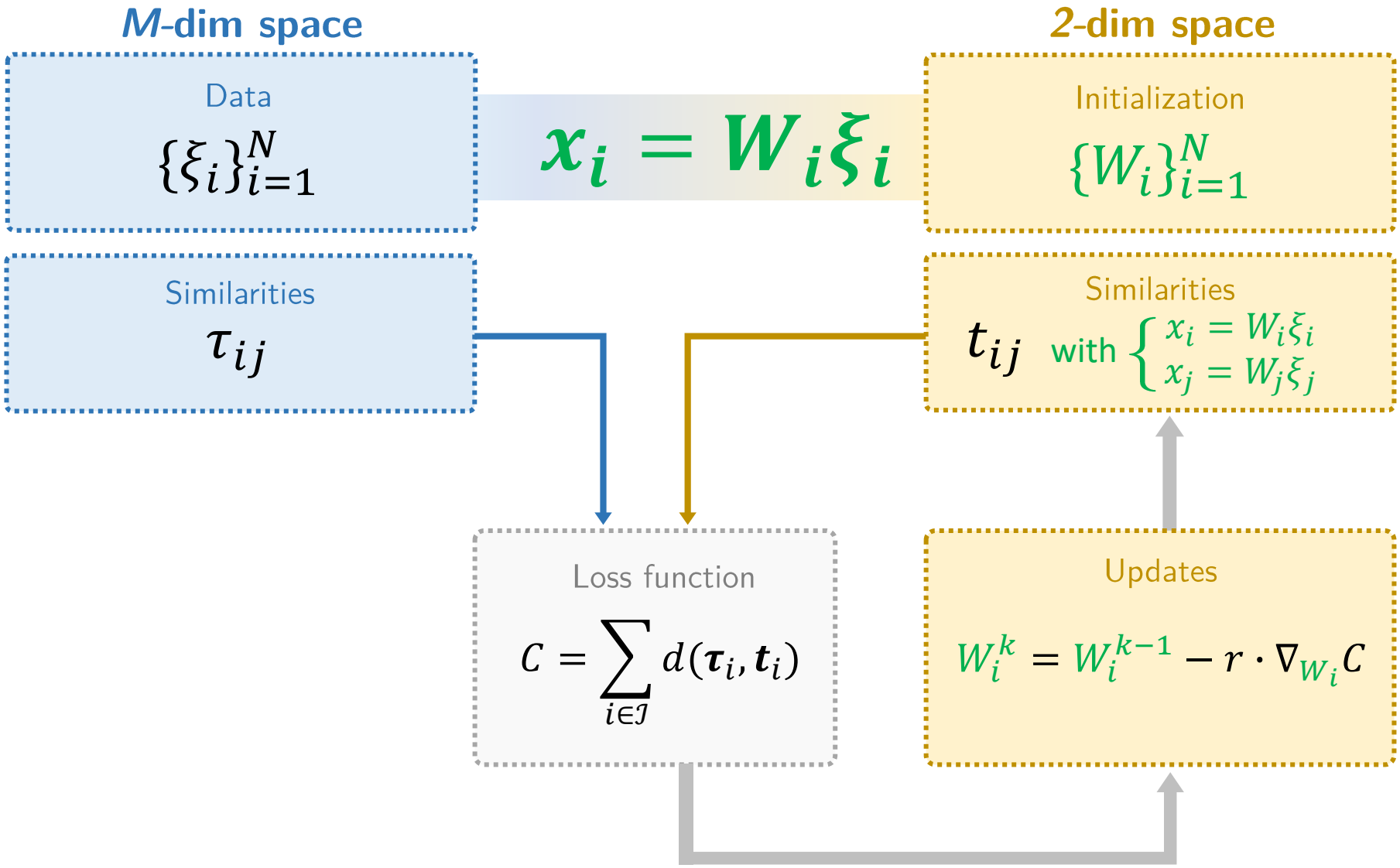

Ms. t-SNE [W]
Interpretability
Qu

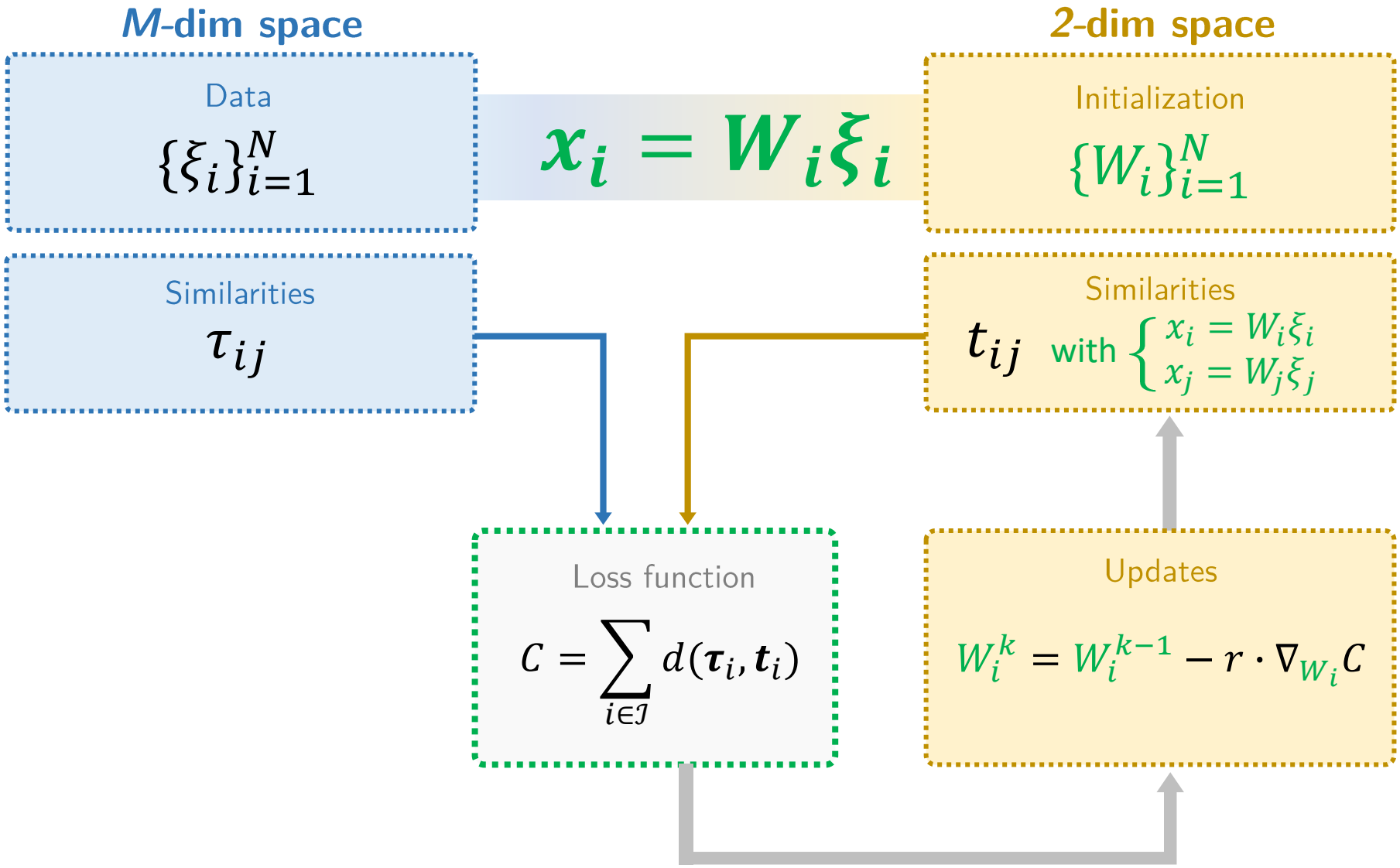

Ms. t-SNE [W_i]
Interpretability
Quality


-SNE [-]
Quality


M-dim space
Data
$\{\xi_i\}_{i=1}^N$
Similarities
τ_{ij}
2-dim space
Initialization
$\{x_i\}_{i=1}^N$
Similarities
t_{ij}
Loss function
$C = \sum_{i \in \mathcal{J}} d(\boldsymbol{\tau}_i, \boldsymbol{t}_i)$
Updates
$x_i^k = x_i^{k-1} - r \cdot \nabla_{x_i} C$


M-dim space
2-dim space
Data
$\{\xi_i\}_{i=1}^N$
$x_i = W_i \xi_i$
Initialization
$\{x_i\}_{i=1}^N$
Similarities
τ_{ij}
Similarities
t_{ij}
Loss function
$C = \sum_{i \in \mathcal{J}} d(\tau_i, t_i)$
Updates
$x_i^k = x_i^{k-1} - r \cdot \nabla_{x_i} C$


M-dim space
2-dim space
Data
$\{\xi_i\}_{i=1}^N$
$x_i = W_i \xi_i$
Initialization
$\{W_i\}_{i=1}^N$
Similarities
τ_{ij}
Similarities
t_{ij} with $\begin{cases} x_i = W_i \xi_i \\ x_j = W_j \xi_j \end{cases}$
Loss function
$C = \sum_{i \in \mathcal{J}} d(\boldsymbol{\tau}_i, \boldsymbol{t}_i)$
Updates
$W_i^k = W_i^{k-1} - r \cdot \nabla_{W_i} C$


M-dim space
2-dim space
Data
$\{\xi_i\}_{i=1}^N$
$x_i = W_i \xi_i$
Initialization
$\{W_i\}_{i=1}^N$
Similarities
τ_{ij}
Similarities
t_{ij} with $\begin{cases} x_i = W_i \xi_i \\ x_j = W_j \xi_j \end{cases}$
Loss function
$C = \sum_{i \in \mathcal{J}} d(\boldsymbol{\tau}_i, \boldsymbol{t}_i)$
Updates
$W_i^k = W_i^{k-1} - r \cdot \nabla_{W_i} C$

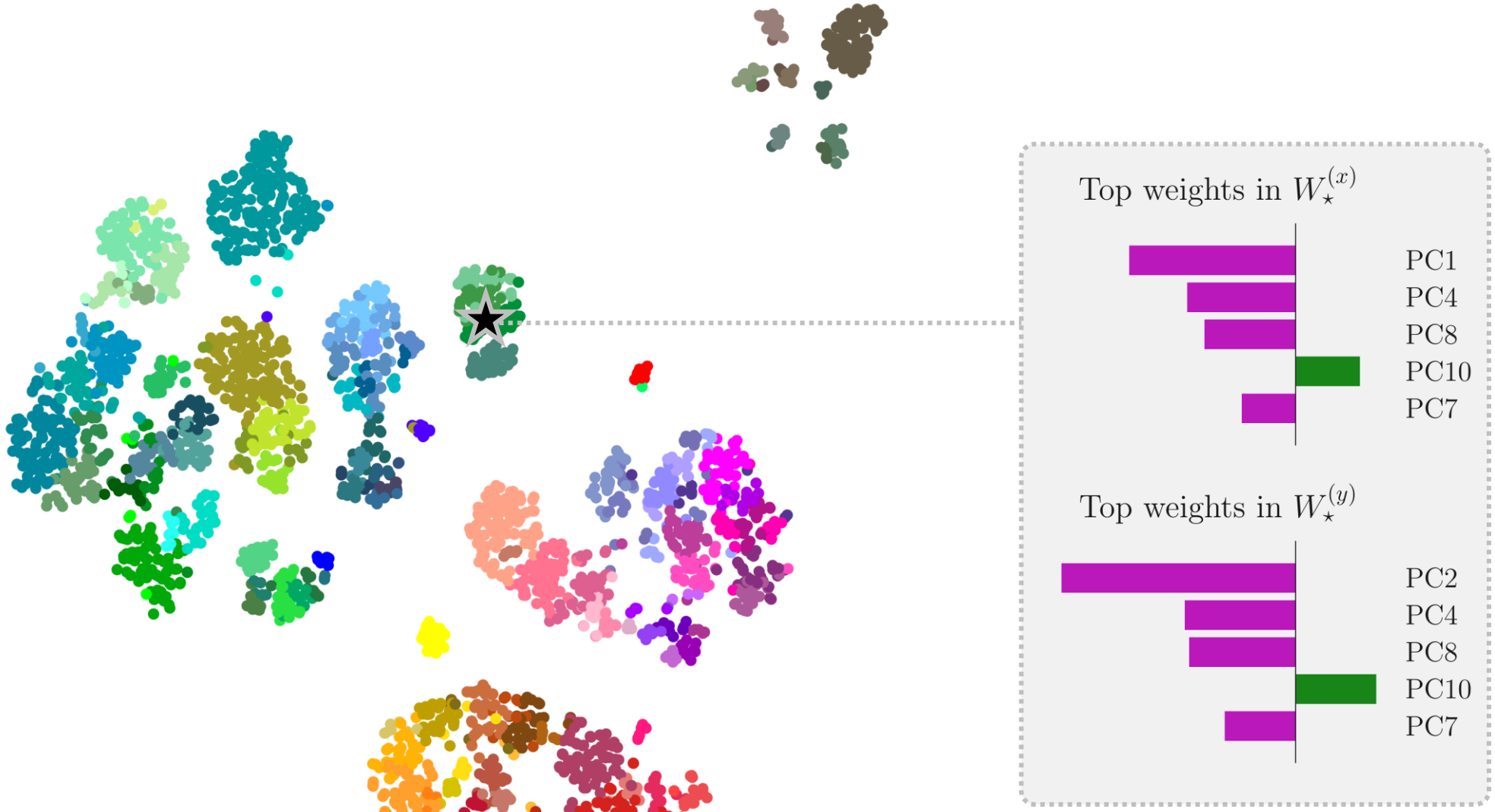
$$C = \sum_{i \in \mathcal{I}} \left(d(\boldsymbol{\tau}_i, \boldsymbol{t}_i) \right)$$

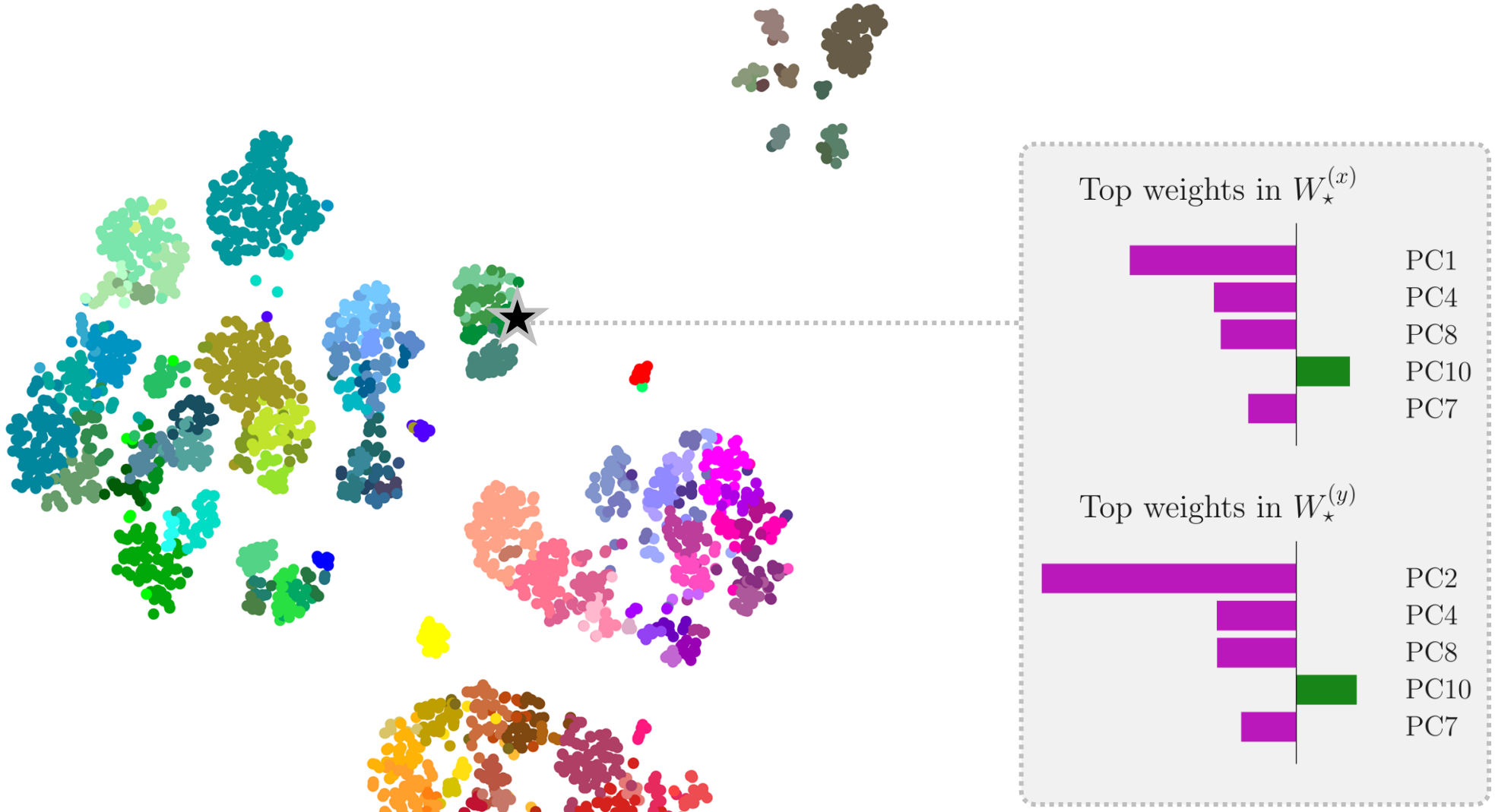
$$C = \sum_{i \in \mathcal{I}} \left(d(\boldsymbol{\tau}_i, \boldsymbol{t}_i) \; + \; \alpha \cdot \sum_{j \in \mathcal{J}} t_{ij} \left\| W_i - W_j \right\|_F \right)$$

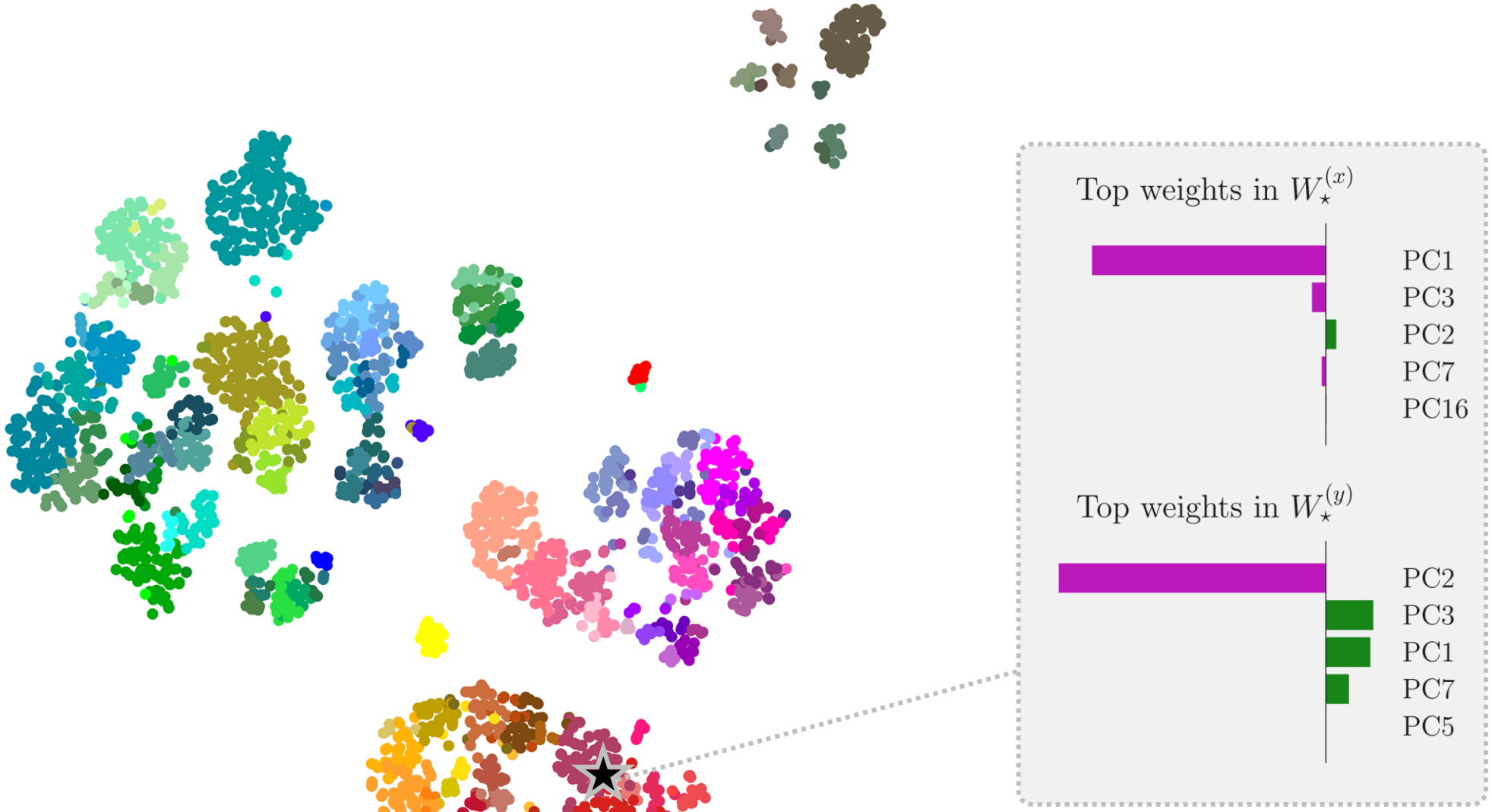
Coherence

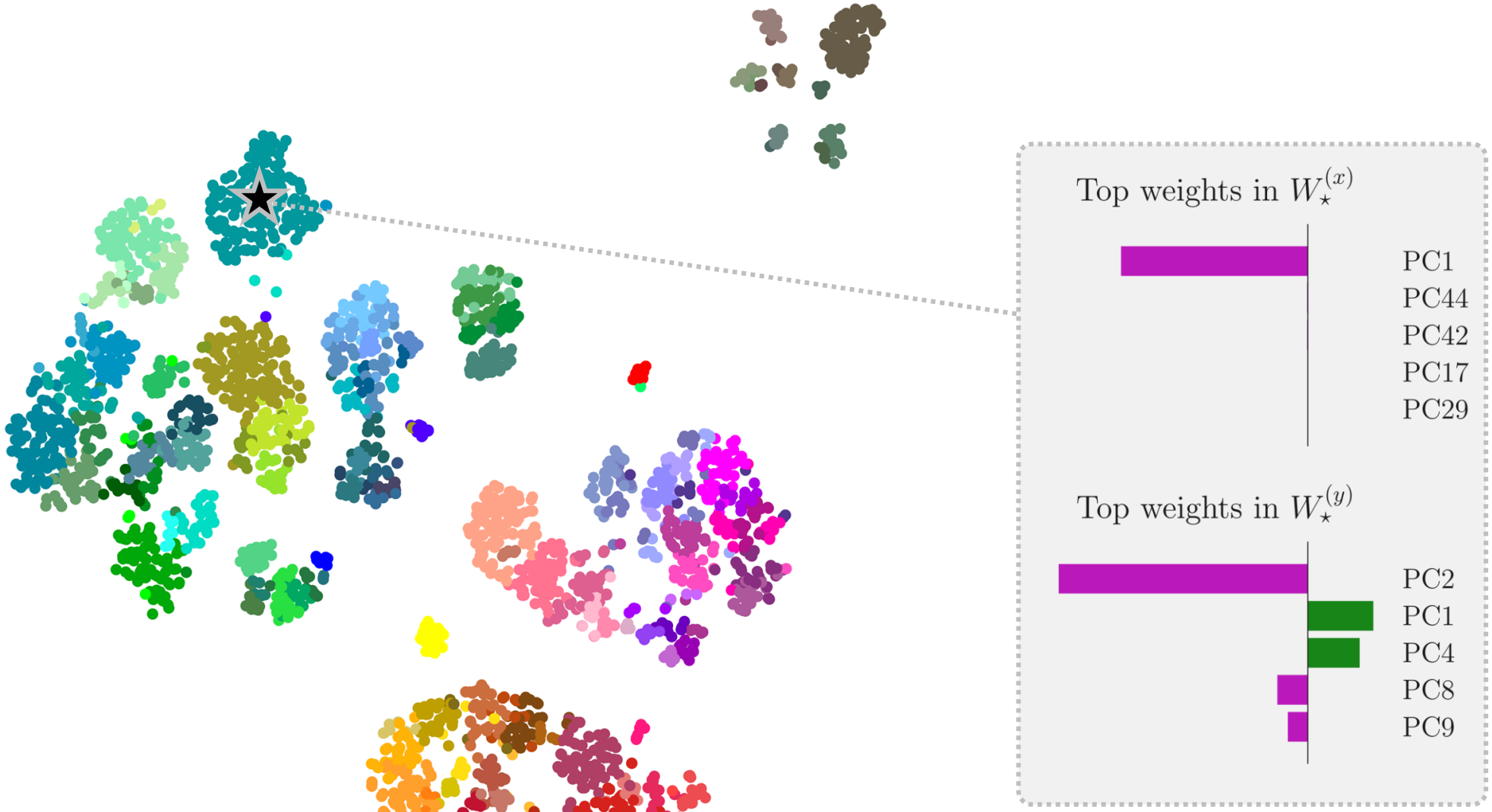
$$C = \sum_{i \in \mathcal{I}} \left(d(\boldsymbol{\tau}_i, \boldsymbol{t}_i) \; + \; \alpha \cdot \sum_{j \in \mathcal{I}} t_{ij} \left\| W_i - W_j \right\|_F + \beta \cdot \| W_i \|_1 \right)$$

Coherence **Sparsity**


Top weights in $W_\star^{(x)}$
PC1
PC4
PC8
PC10
PC7
Top weights in $W_\star^{(y)}$
PC2
PC4
PC8
PC10
PC7


Top weights in $W_\star^{(x)}$
PC1
PC4
PC8
PC10
PC7
Top weights in $W_\star^{(y)}$
PC2
PC4
PC8
PC10
PC7


Top weights in $W_\star^{(x)}$
PC1
PC3
PC2
PC7
PC16
Top weights in $W_\star^{(y)}$
PC2
PC3
PC1
PC7
PC5


Top weights in $W_\star^{(x)}$
PC1
PC44
PC42
PC17
PC29
Top weights in $W_\star^{(y)}$
PC2
PC1
PC4
PC8
PC9

Main takeaways

Main takeaways

- A natively **interpretable** version of t-SNE

Main takeaways

- A natively **interpretable** version of *t*-SNE
- Explicit locally linear mapping $\boldsymbol{x_i = W_i \xi_i}$

Main takeaways

- A natively **interpretable** version of t-SNE
- Explicit locally linear mapping $\boldsymbol{x}_i = \boldsymbol{W}_i \boldsymbol{\xi}_i$
- Navigable trade-off: **interpretability/quality**

UCLouvain

Natively Interpretable *t*-SNE

icteam

Don't hesitate to reach out if you'd like to learn more!

edouard.couplet@uclouvain.be

Edouard Couplet

Pierre Lambert

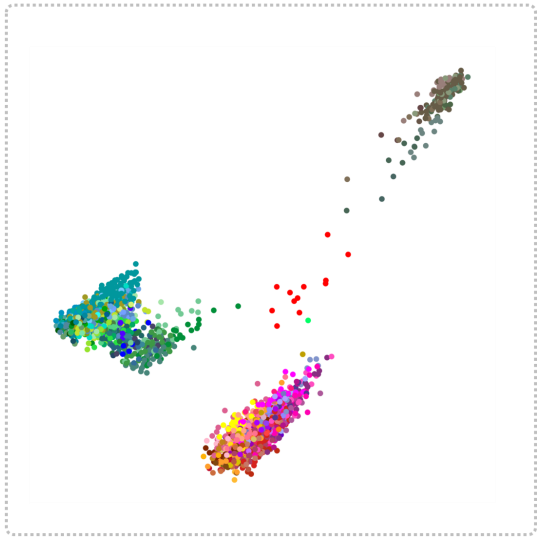
Michel Verleysen

Dounia Mulders

John Lee

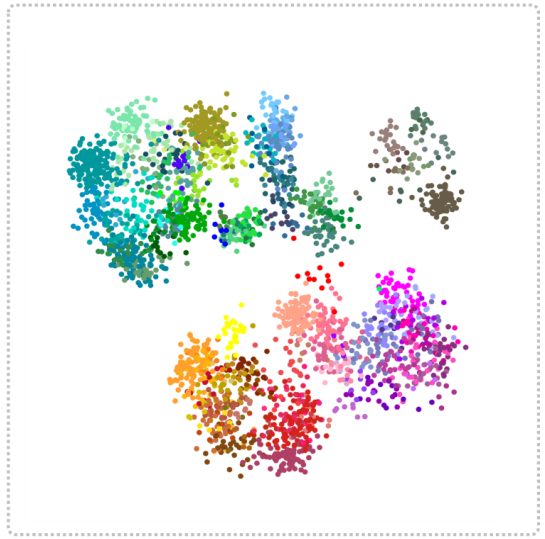
Cyril de Bodt

PCA [W]



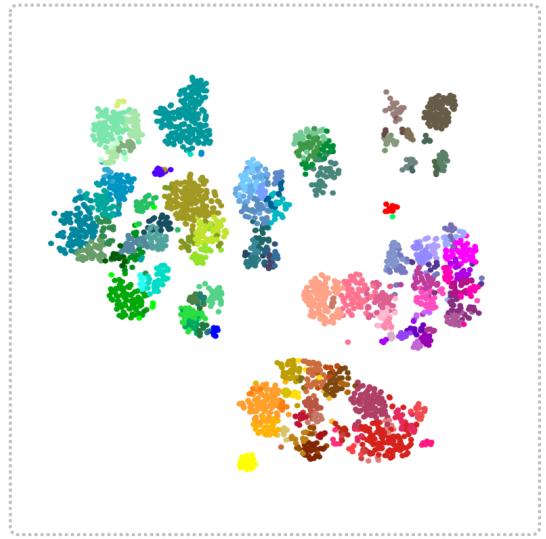
Interpretability ⊕ ⊕ Quality ⊖ ⊖

Ms. *t*-SNE [W]



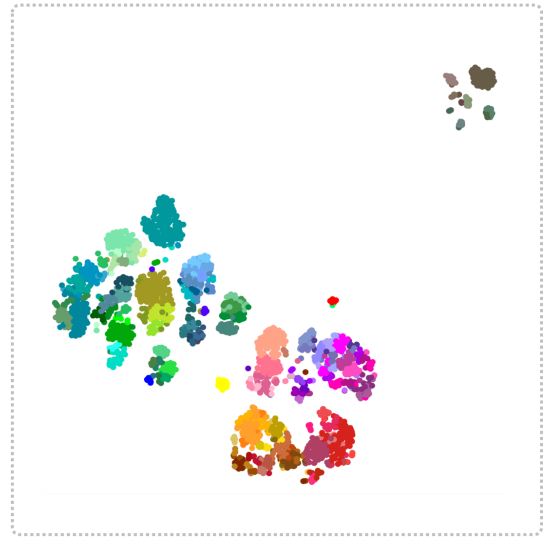
Interpretability ⊕ ⊕ Quality ⊖

Ms. *t*-SNE [W_i]

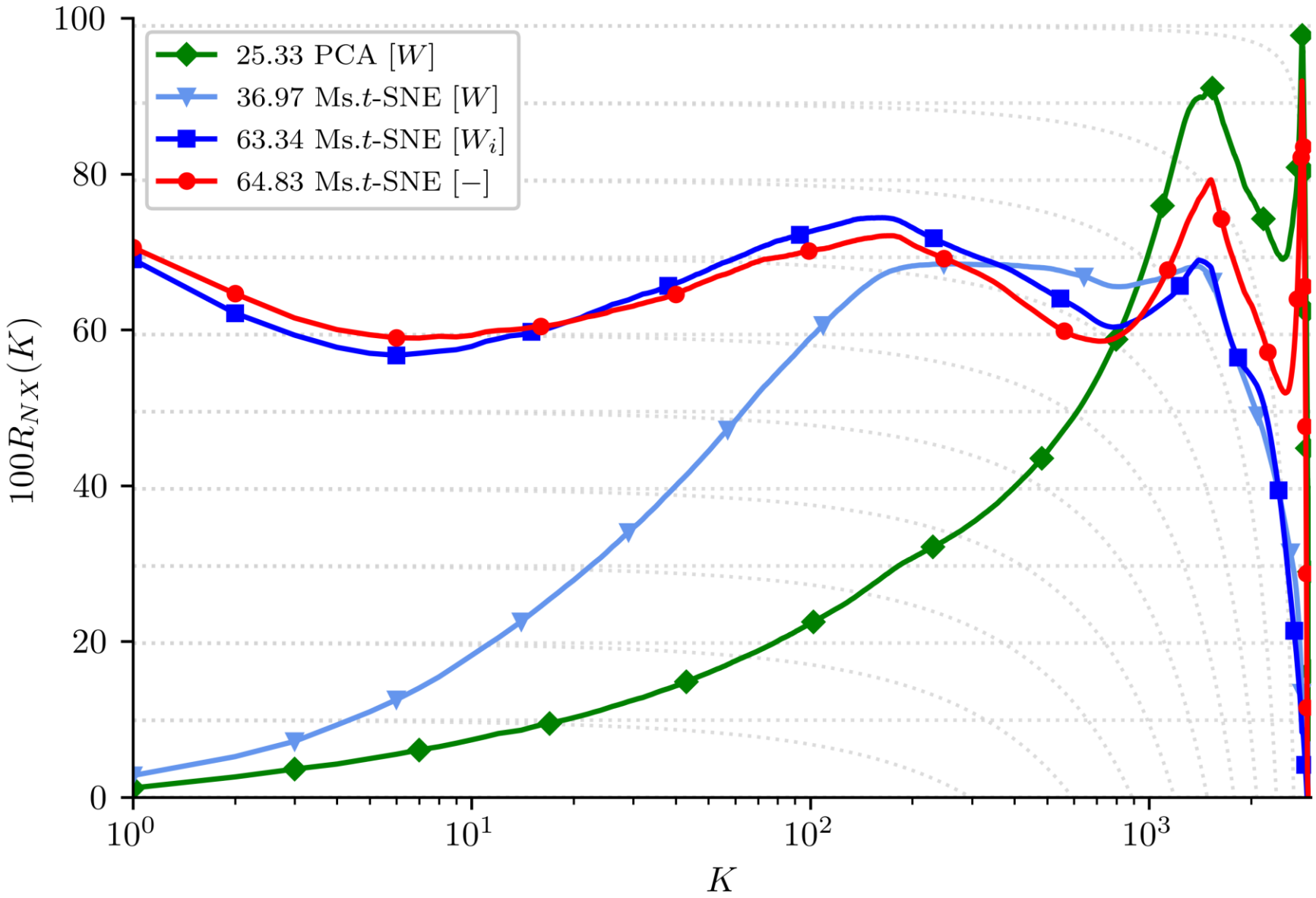


Interpretability ⊕ Quality ⊕

Ms. *t*-SNE [-]

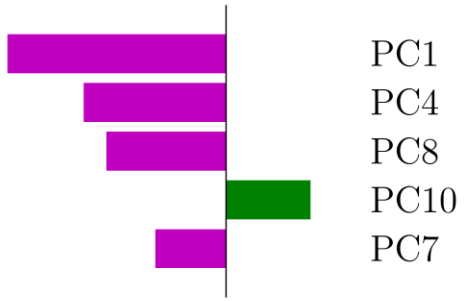


Interpretability ⊖ ⊖ Quality ⊕ ⊕

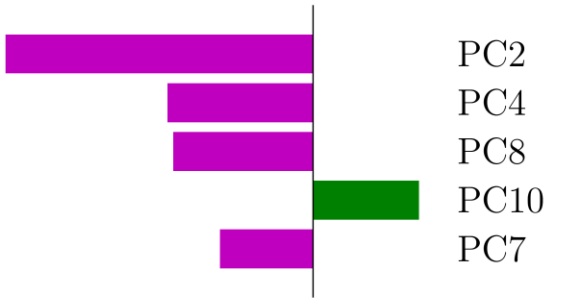

25.33 PCA [W]
36.97 Ms.t-SNE [W]
63.34 Ms.t-SNE [W_i]
64.83 Ms.t-SNE [$-$]
$100R_{NX}(K)$
K
0
20
40
60
80
100
10^0
10^1
10^2
10^3

$\alpha = 10^6$ $\beta = 10^2$

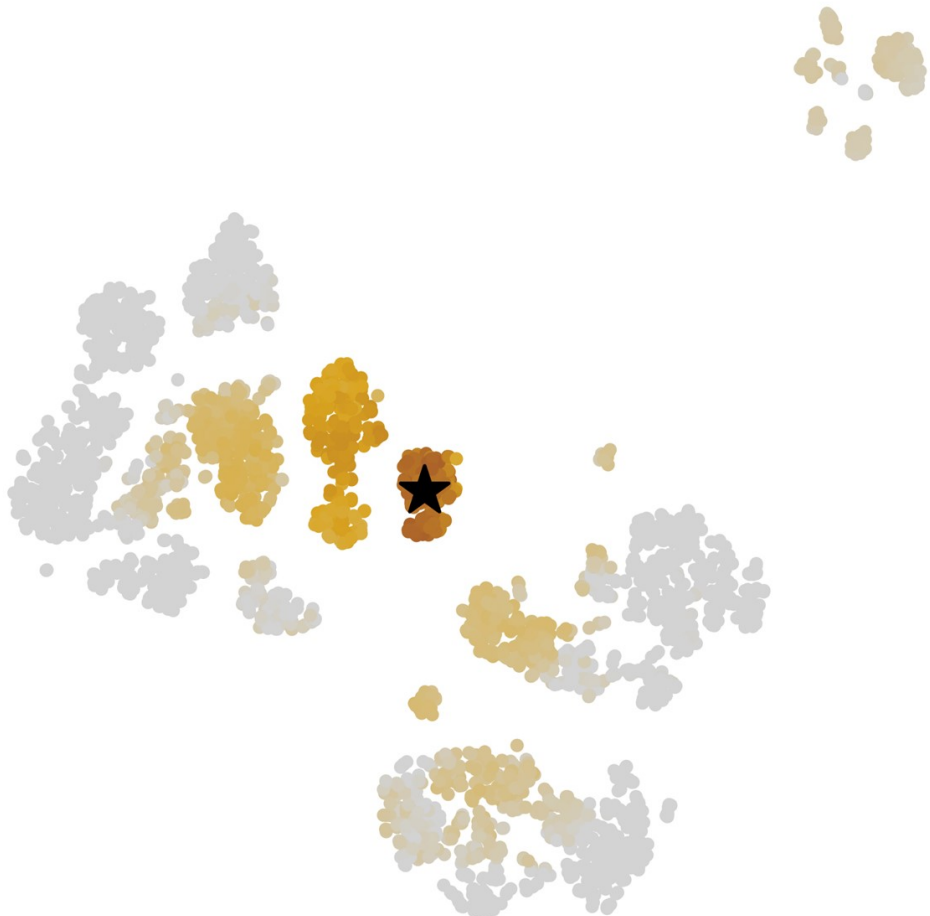
Top weights in $W_\star^{(x)}$

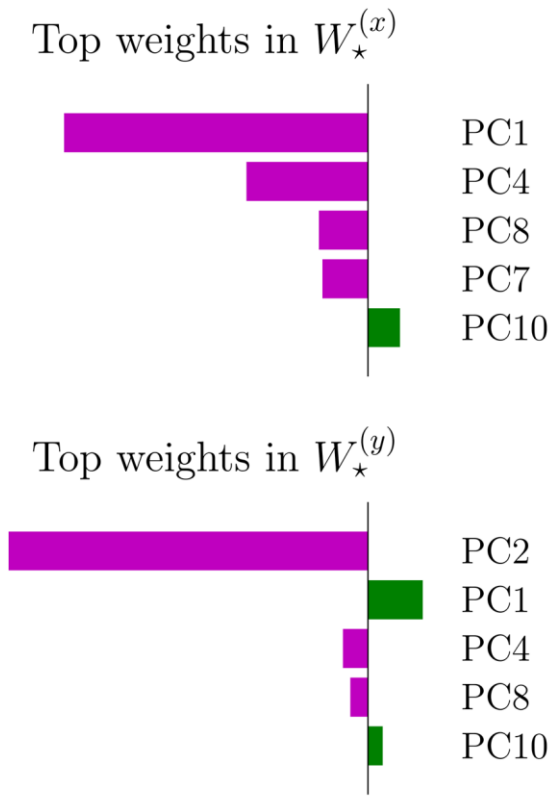


Top weights in $W_\star^{(y)}$

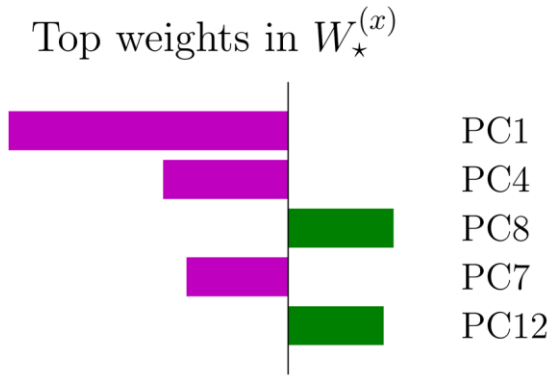


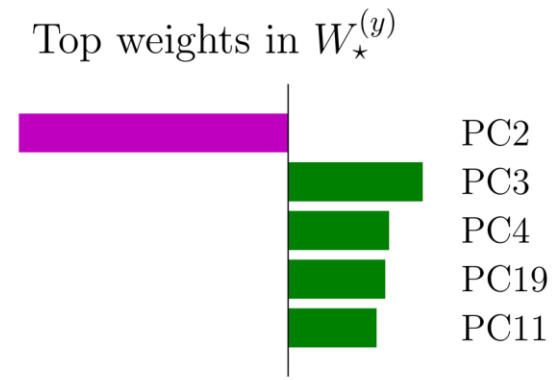
$\alpha = 10^7$ $\beta = 10^2$



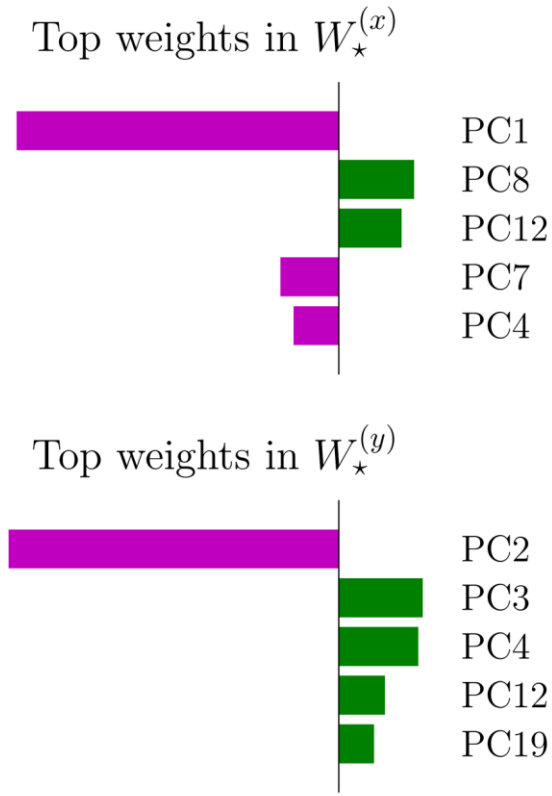


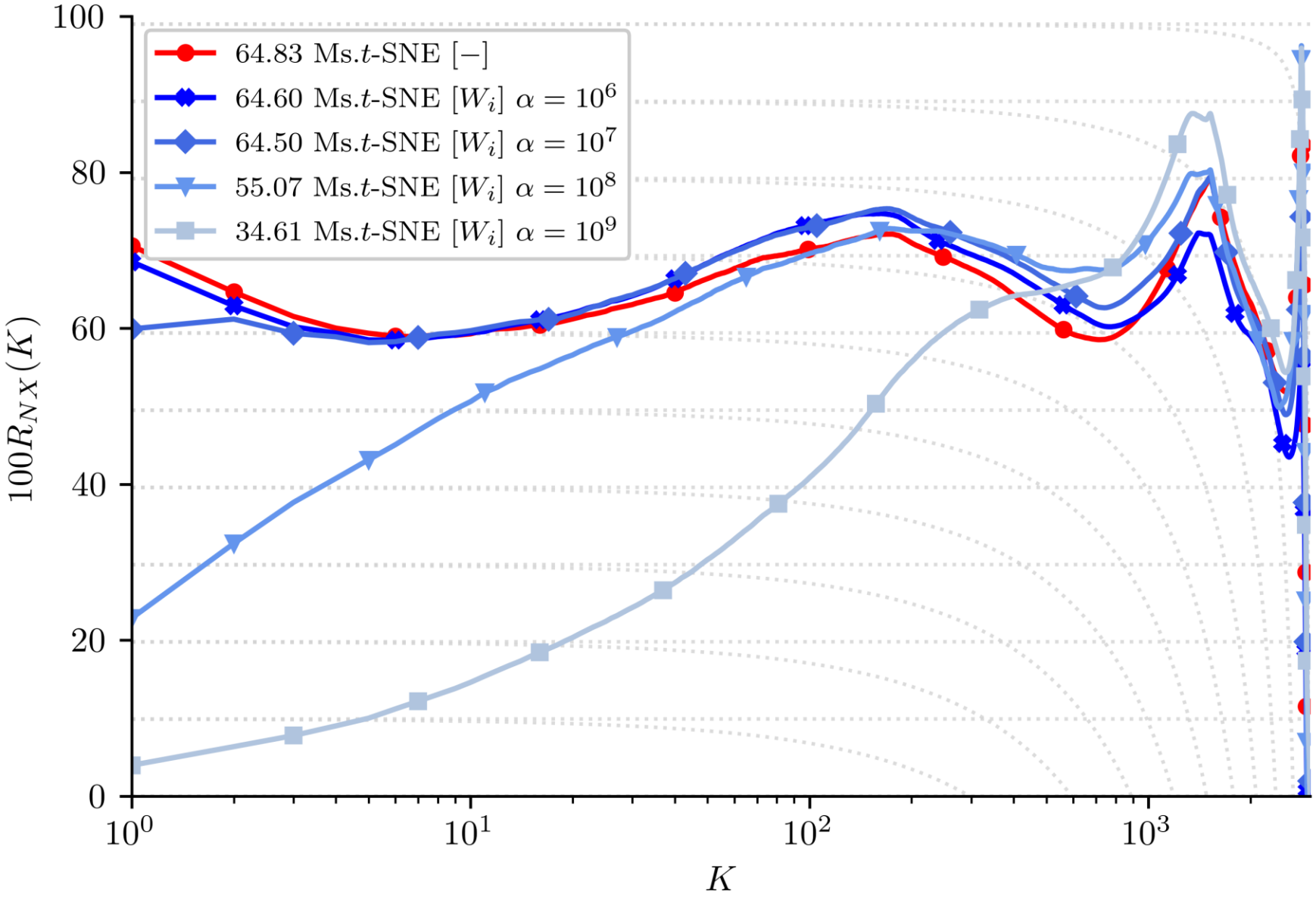
$\alpha = 10^8$ $\beta = 10^2$



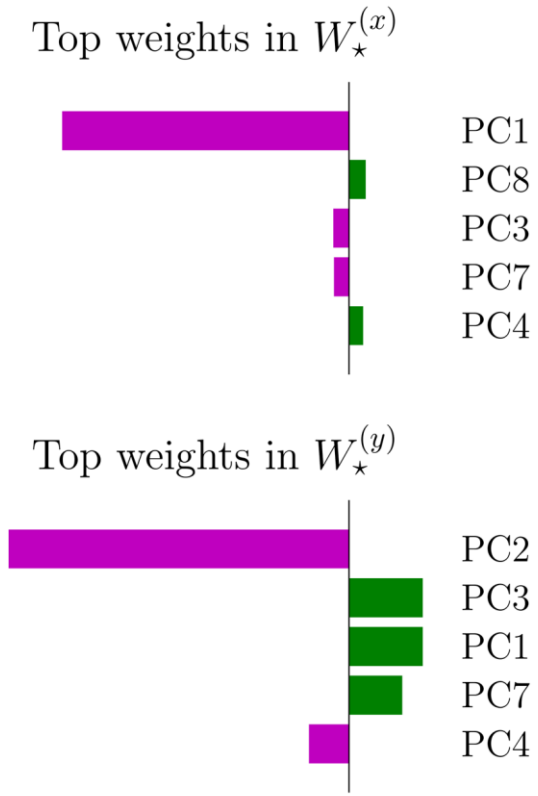


$\alpha = 10^9$ $\beta = 10^2$

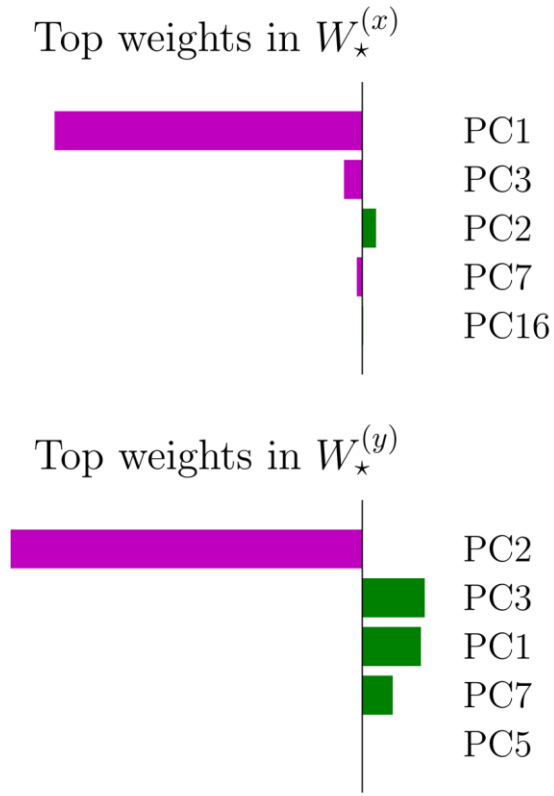



64.83 Ms.t-SNE [−]
64.60 Ms.t-SNE [W_i] $\alpha = 10^6$
64.50 Ms.t-SNE [W_i] $\alpha = 10^7$
55.07 Ms.t-SNE [W_i] $\alpha = 10^8$
34.61 Ms.t-SNE [W_i] $\alpha = 10^9$
$100R_{NX}(K)$
K
0
20
40
60
80
100
10^0
10^1
10^2
10^3

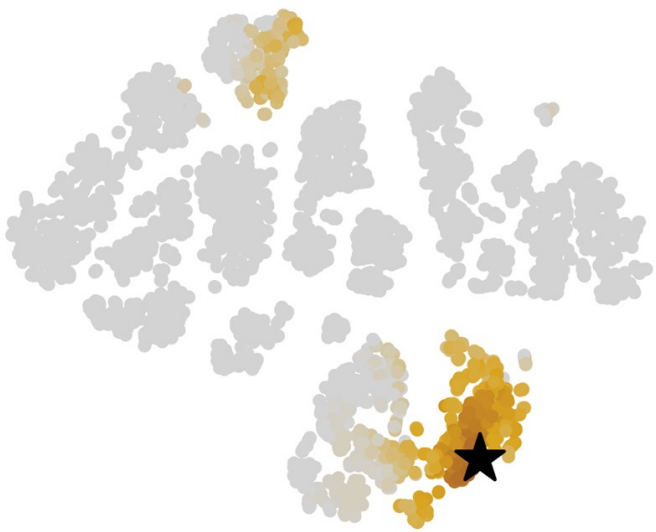
$\alpha = 10^6$ $\beta = 10^1$



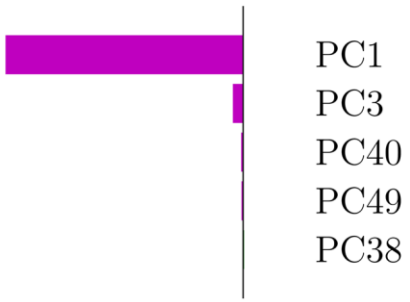
$\alpha = 10^6$ $\beta = 10^2$



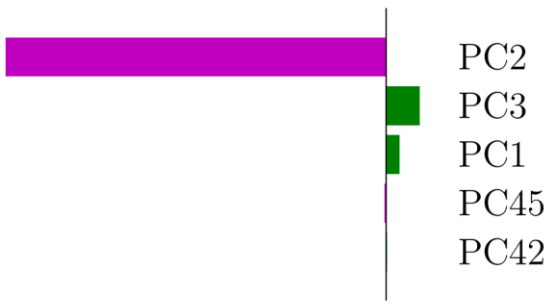
$\alpha = 10^6$ $\beta = 10^3$



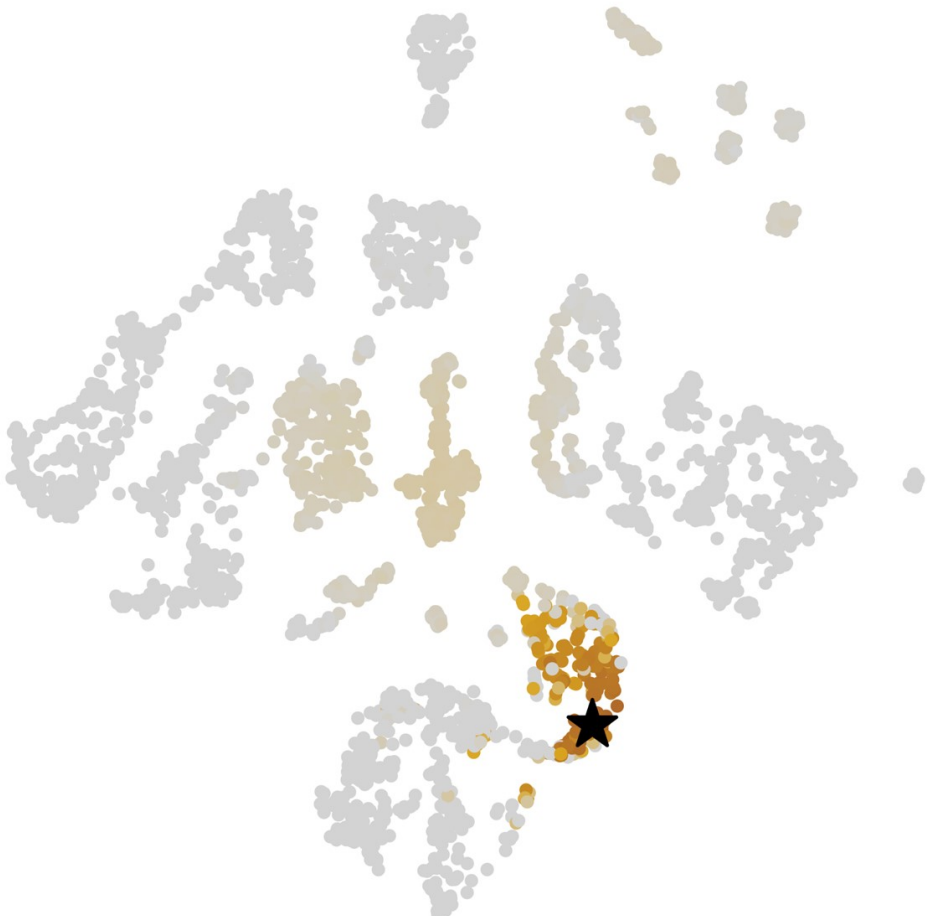
Top weights in $W_\star^{(x)}$



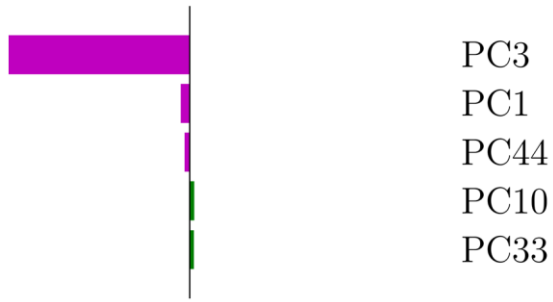
Top weights in $W_\star^{(y)}$



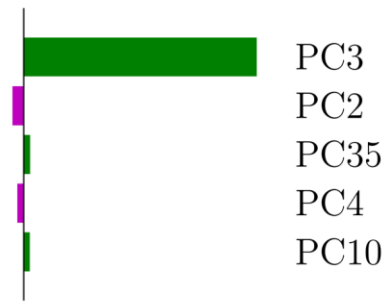
$\alpha = 10^6$ $\beta = 10^4$

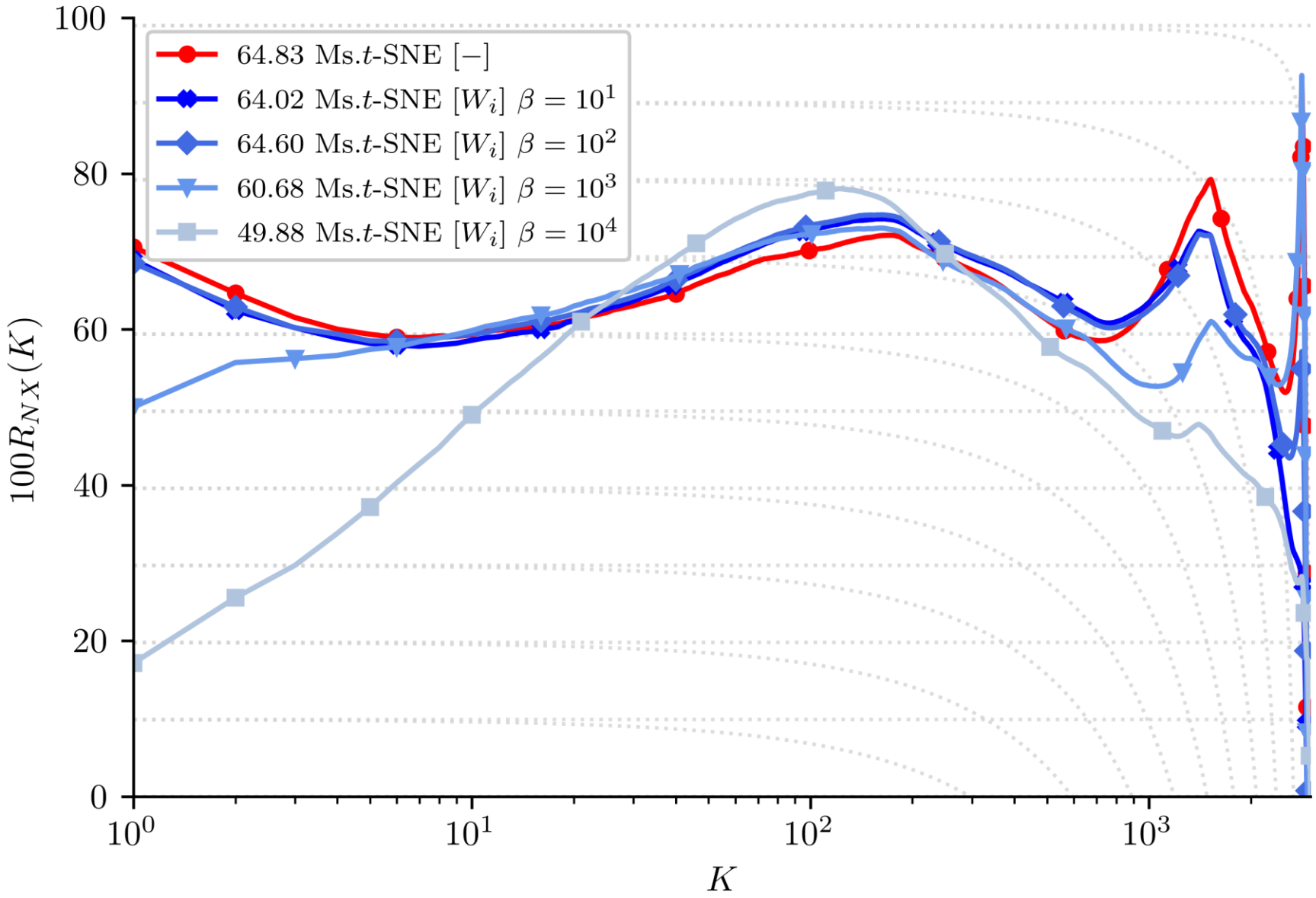


Top weights in $W_\star^{(x)}$



Top weights in $W_\star^{(y)}$




64.83 Ms.t-SNE [−]
64.02 Ms.t-SNE [W_i] $\beta = 10^1$
64.60 Ms.t-SNE [W_i] $\beta = 10^2$
60.68 Ms.t-SNE [W_i] $\beta = 10^3$
49.88 Ms.t-SNE [W_i] $\beta = 10^4$
100
80
60
40
20
0
$100R_{NX}(K)$
10^0
10^1
10^2
10^3
K