Unsupervised Data Visualization for Big Data Exploratory Analysis

Kevin R. Moon, PhD

Utah State University

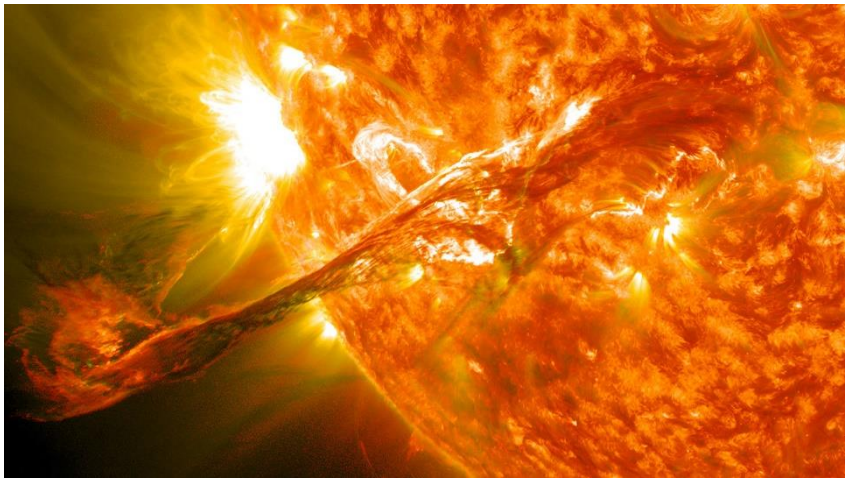
Yale SCHOOL OF MEDICINE

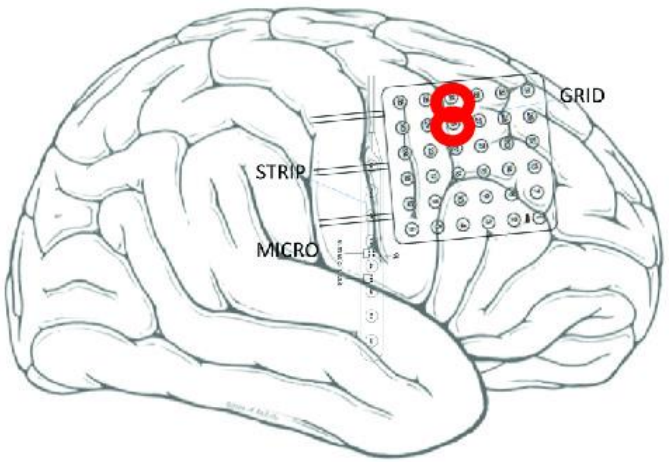
MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

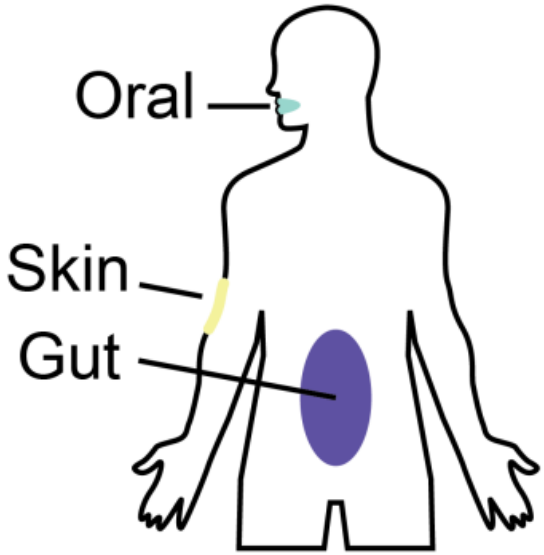
PHATE Github Repo

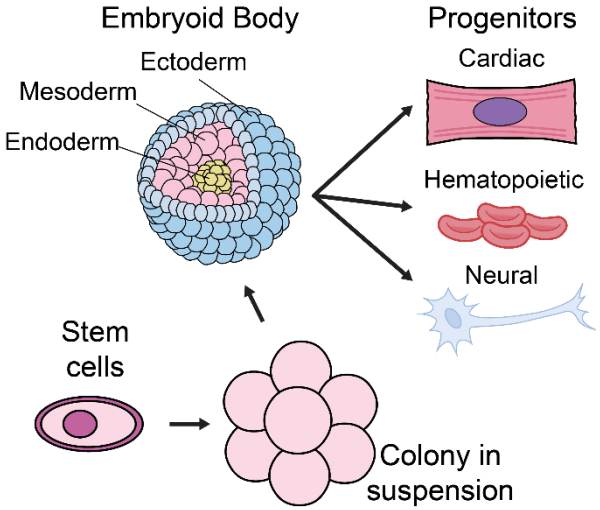
- github.com/KrishnaswamyLab/PHATE

The Era of Big Data



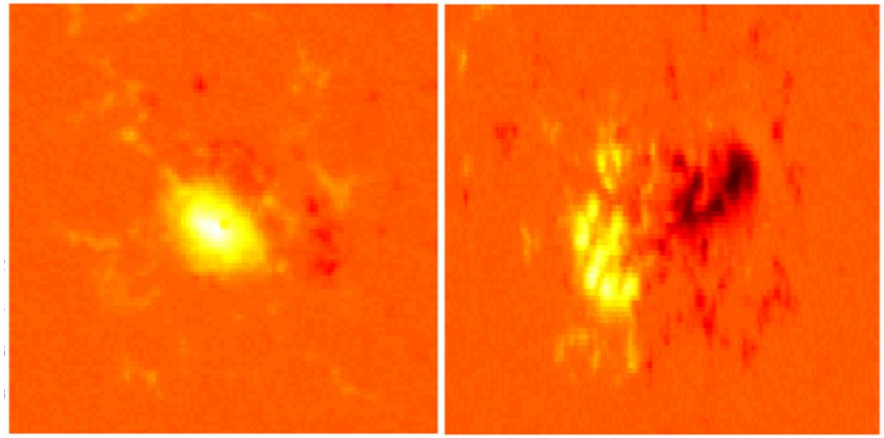






Solar flare prediction

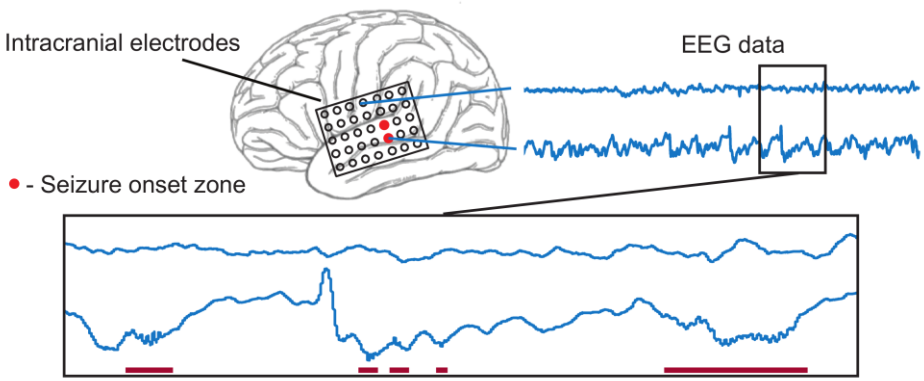
- Solar flares – potentially dangerous to electronics
- Active region – locally increased magnetic flux
  - Associated with sunspots and solar flares
- Can we predict solar flare strength and occurrence?

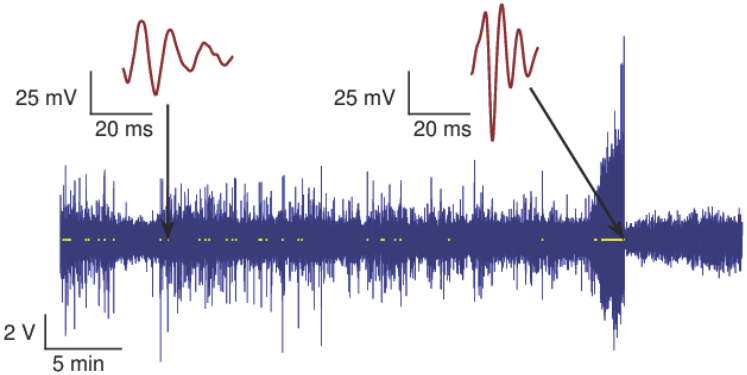


Magnetogram images of two active regions

Moon et al. (SWSC, 2016ab)

Seizure onset zone (SOZ) detection



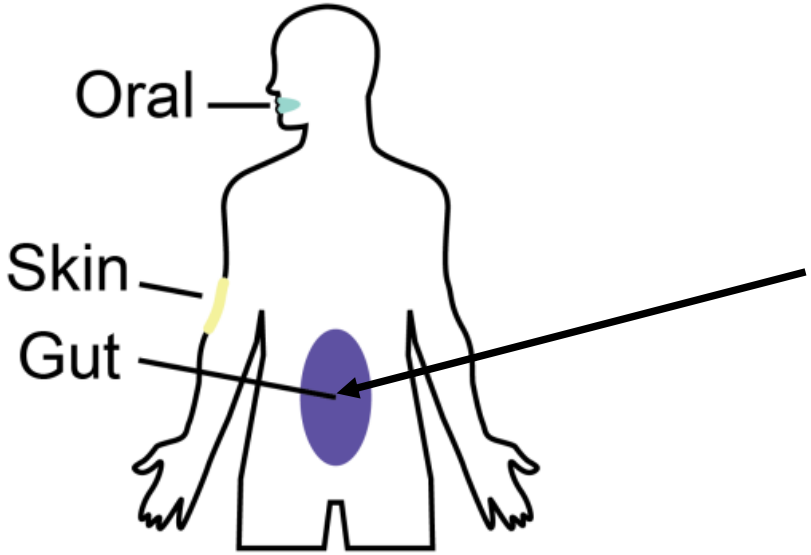


- Strong correlation between high frequency oscillation (HFO) rate and the SOZ (Gliske et al., 2015)
- Can we accurately localize the SOZ from HFO measurements?

Gliske & Moon et al. (*ICASSP*, 2016)

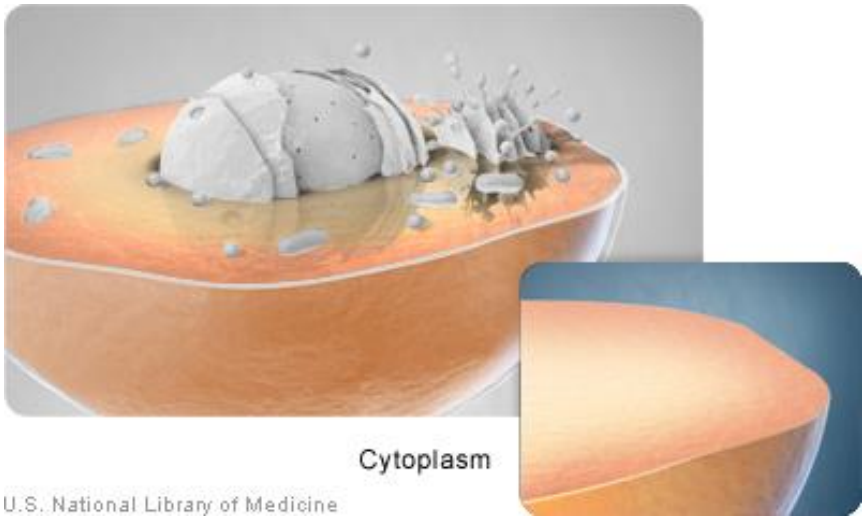
Human microbiome

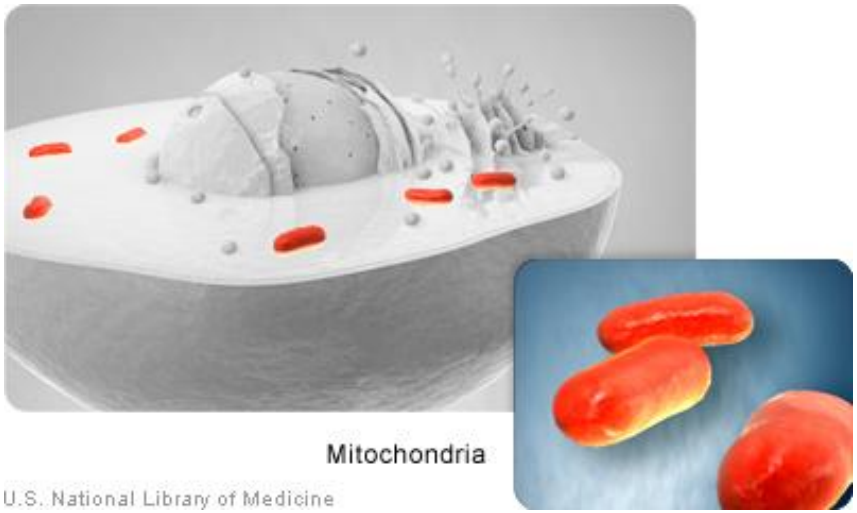
- Microbiome linked to various diseases
  - e.g. autism (Li et al., 2017), arthritis (Scher et al., 2016)

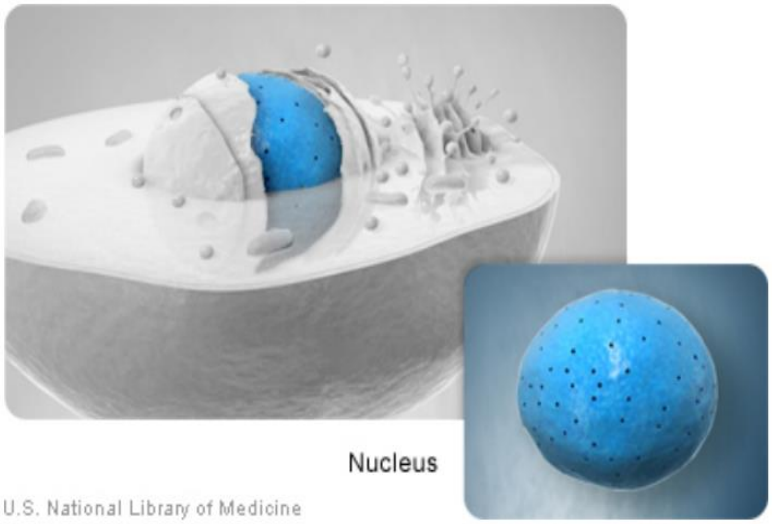


(Pennisi, 2016)

Parts of Cell



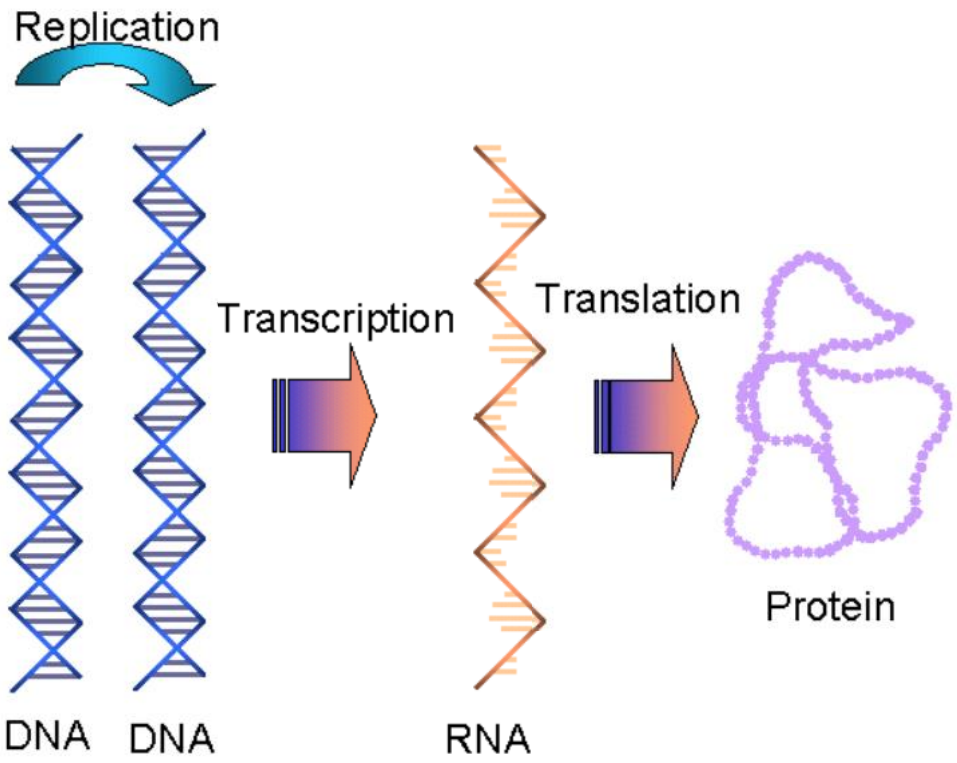




https://ghr.nlm.nih.gov/primer/basics.pdf

Central Dogma of Molecular Biology

- DNA
  - Information across generations
- RNA
  - Information around the cell
- Protein
  - Functional gene product

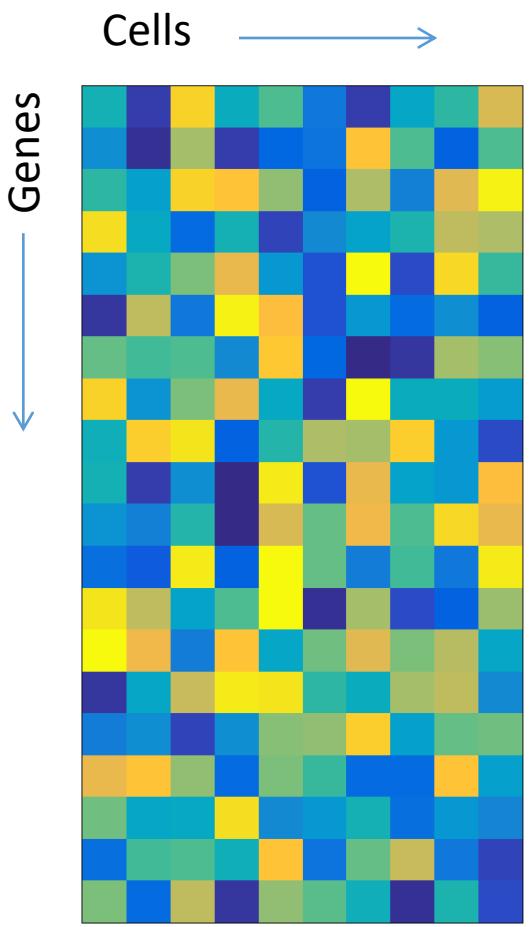


So what?

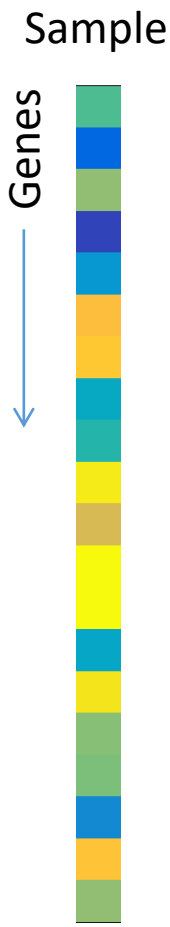
- Cancer metastasis
- Autoimmune diseases
- Immunological responses to diseases and treatment
- Effects of aging
- Effects of genetic mutations
- Etc.

Bulk vs. Single-Cell Data

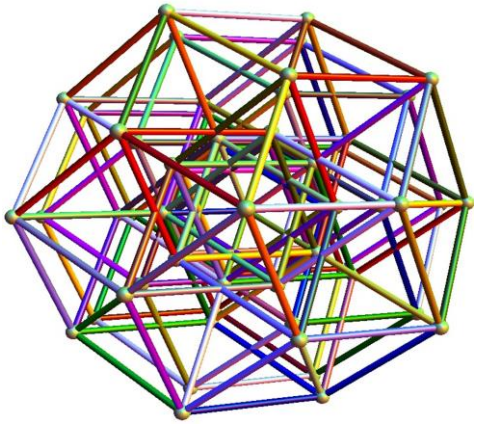
Single-cell Measurement



Bulk Measurement



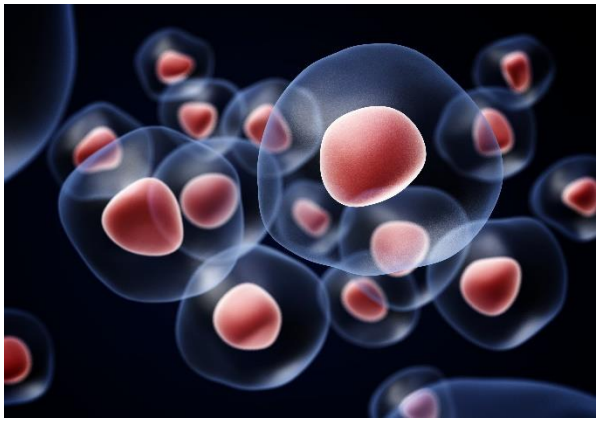
Single-Cell Data



High Dimensional

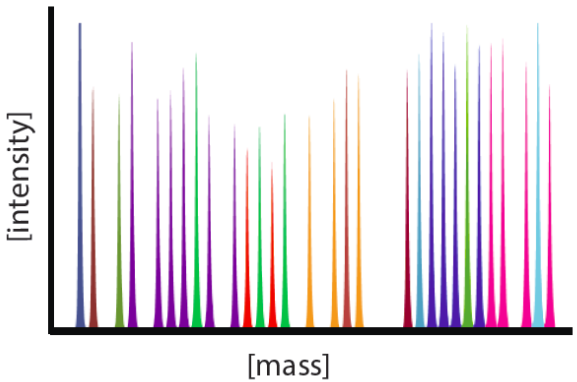
High Throughput

1000s to 100,000s of cells



Heterogeneous

Current Single-Cell Measurement Technologies



- Mass Cytometry (CyTOF)
  - 30 to 40 targets
  - 10,000 to 100,000s of cells
- Single-cell RNA-sequencing (scRNA-seq)
  - 2000+ genes
  - 100 to 10,000s of cells

Challenges in Big Data Analysis

- Interpretability
- Labels or samples may be expensive
- Often, little is known about the data
  - Data exploration is necessary

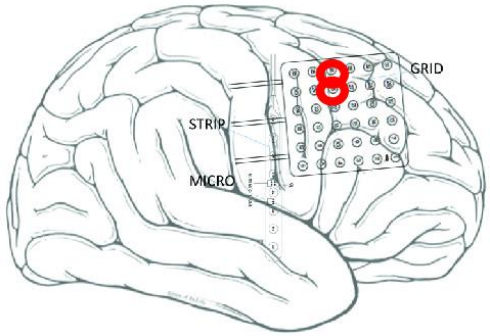
Supervised methods are insufficient in many cases

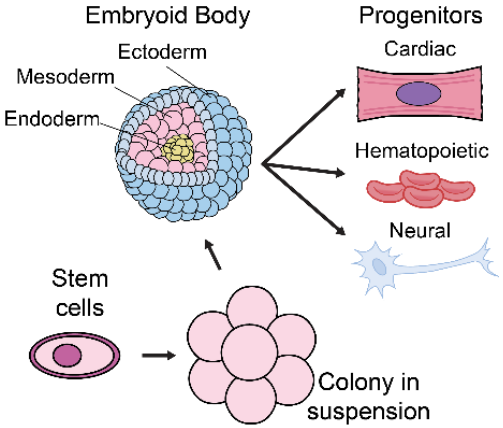
xkcd.com/1838

My Research

Develop data analysis methods for analyzing and extracting information from large complex datasets

- Techniques
  - Information Theory
  - Density estimation
  - Manifold learning
  - Deep learning
  - Nonparametric estimation theory
  - Signal processing filtering theory



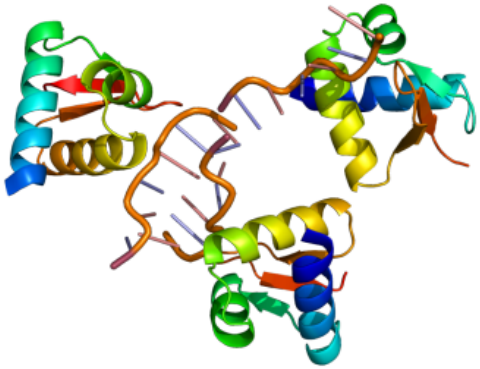


Exploratory Data Analysis

Descriptive Exploration

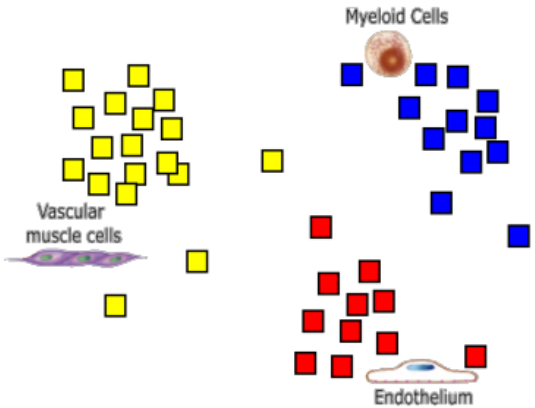
Single-cell data:

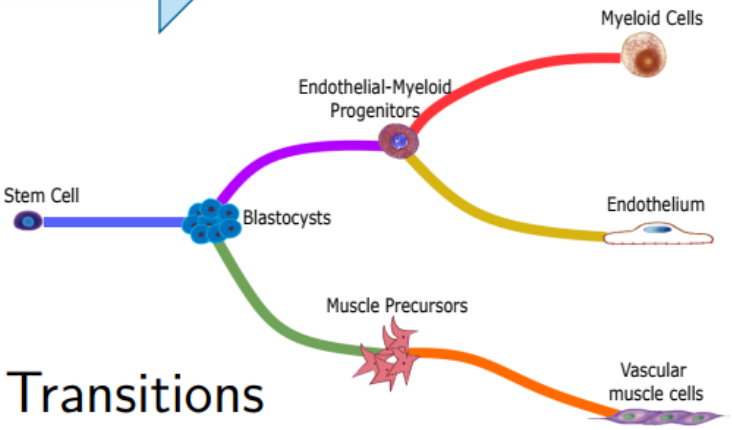
Gene counts



Protein counts

Clusters





Transitions

Methods for Exploratory Data Analysis

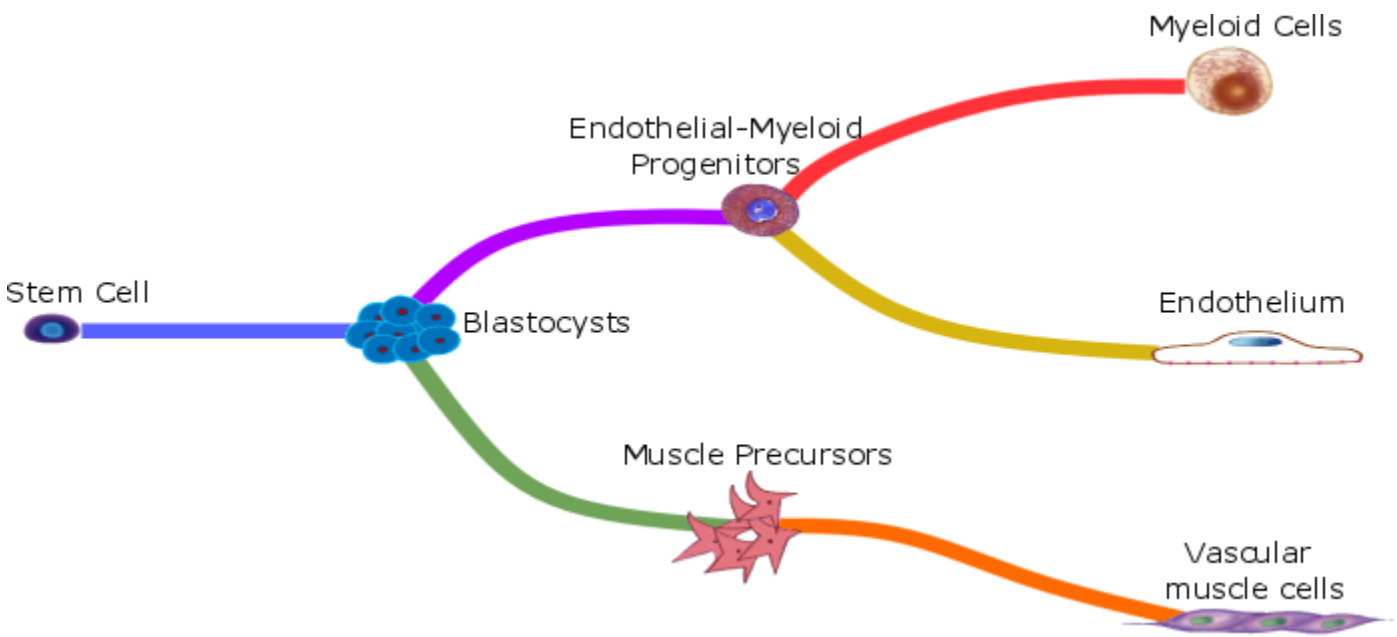
- Unsupervised
  - Labels unnecessary
  - Minimal assumptions
- Robust to noise
- Scalable
  - Dimension
  - Data points
- Usable by domain experts

Data Visualization

- Humans are very visual
- Data visualization is a necessary tool for exploration
  - Develop intuitive understanding of the structure
  - Generate hypotheses

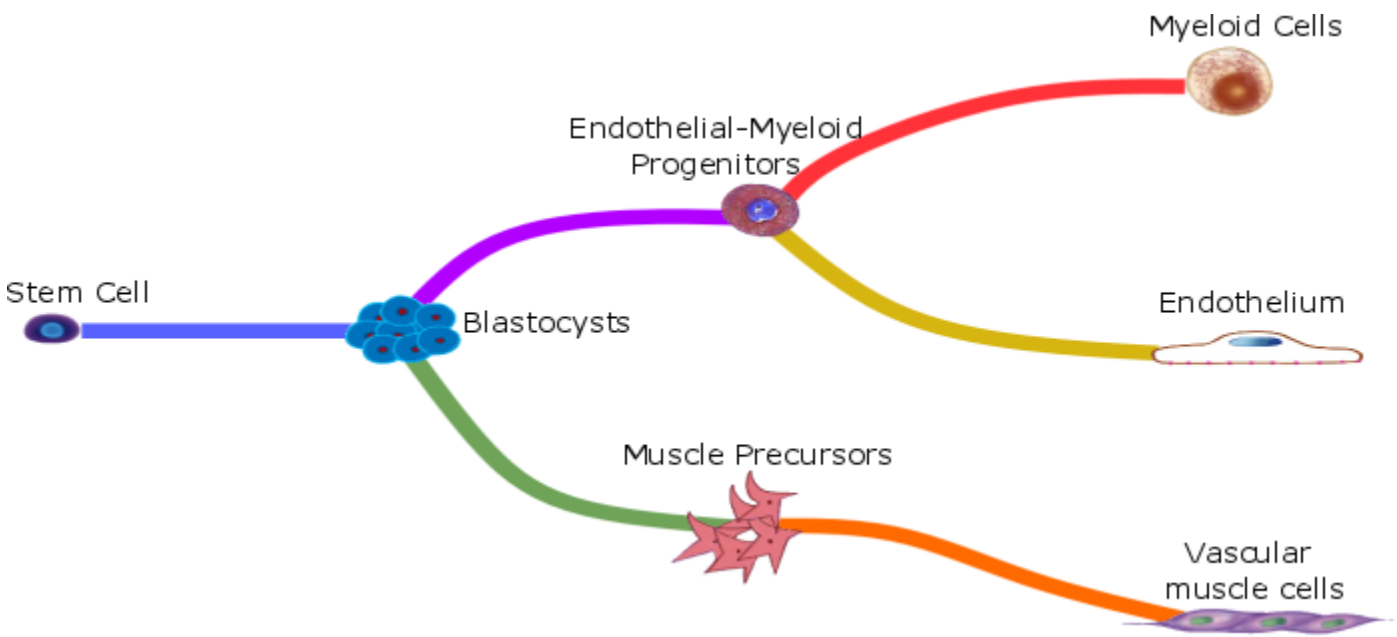
Challenges in Visualizing (Biomedical) Data

- High dimensions
- Noise/artifacts
- Nonlinearities
- Transitional structure

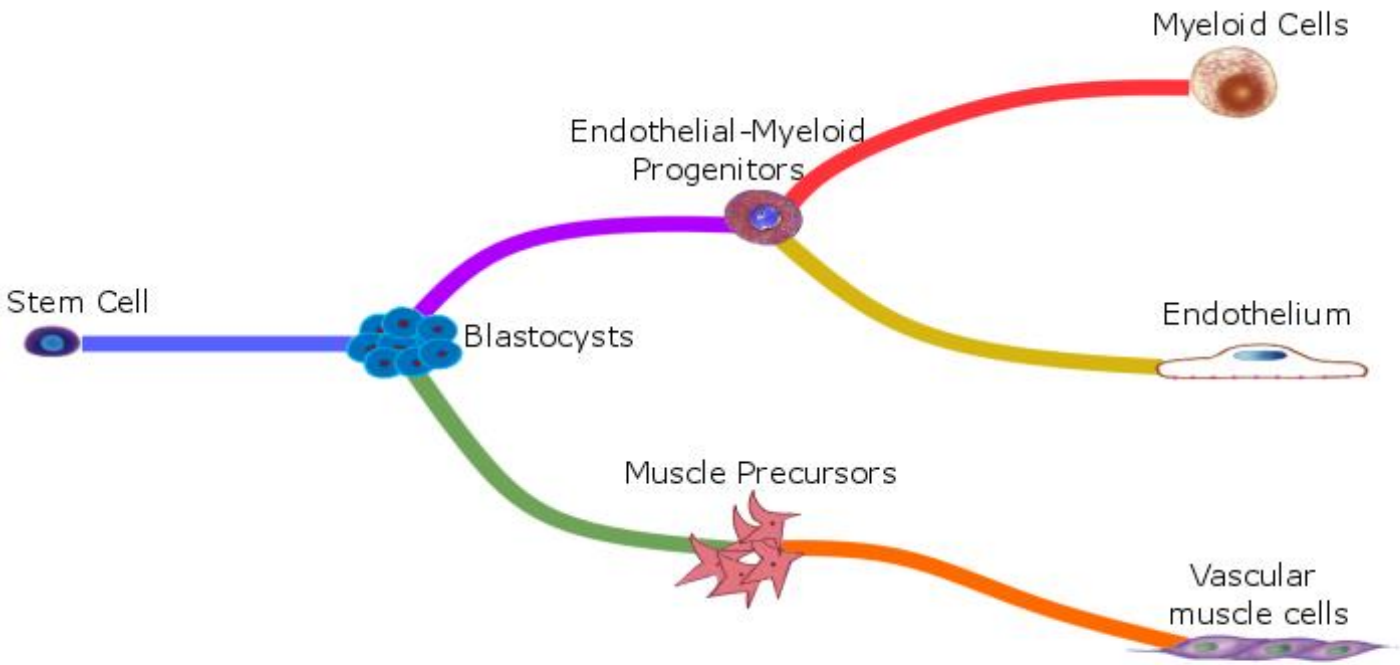


Transitional Structure

- Many biological systems have a dominant transitional structure
  - Developmental branches
  - Cancer
  - Sleep stage



Visualizing Structure



Existing Methods: PCA

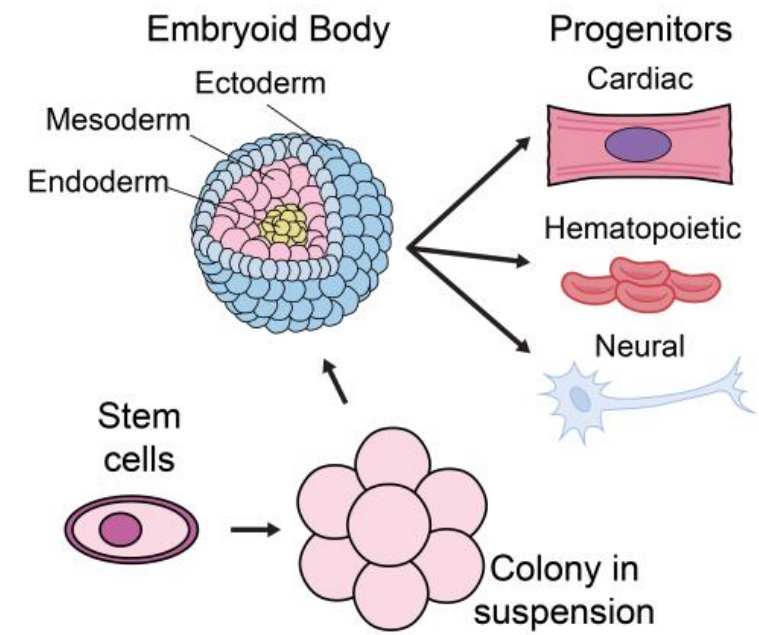
Artificial Tree Data

PCA

1400 points, 60 dimensions

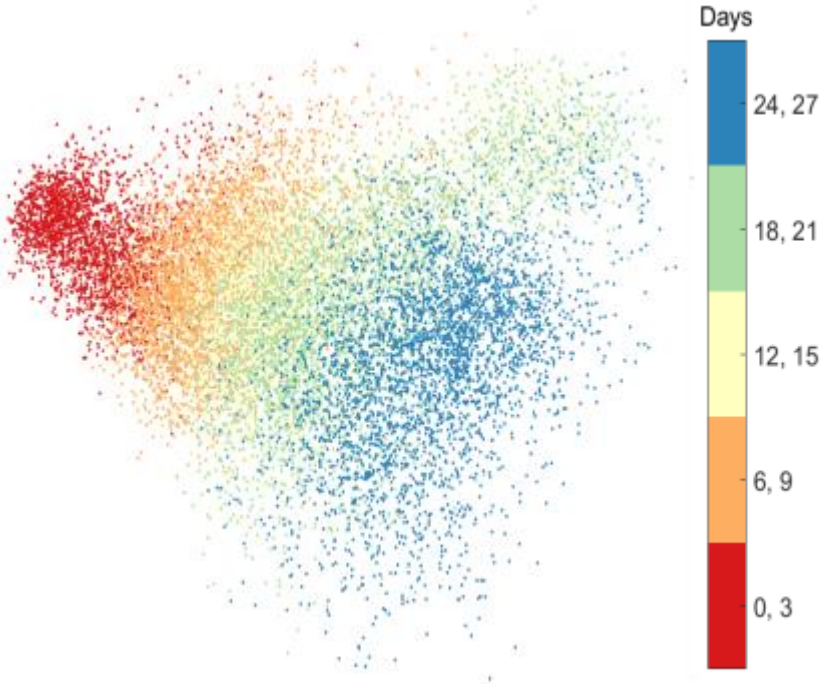
Existing Methods: PCA

Newly generated scRNA-seq data (27 days)



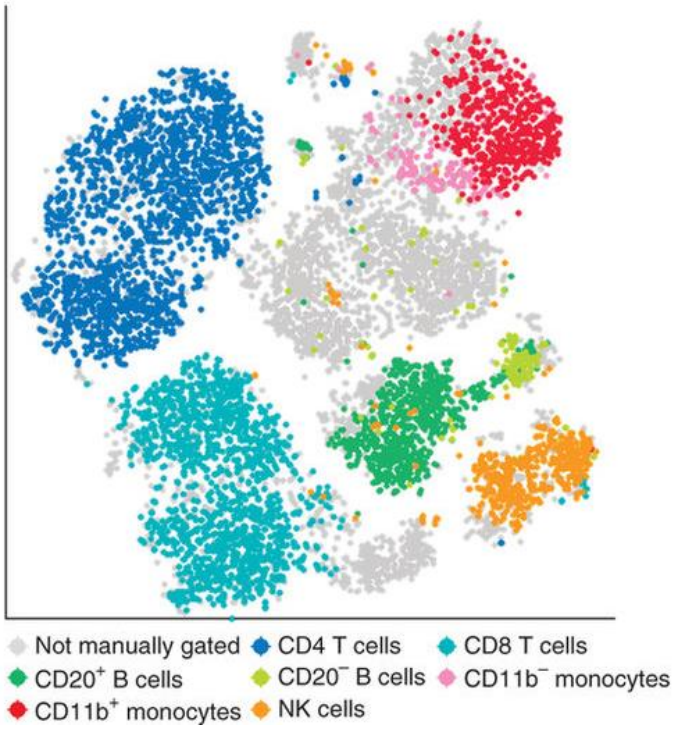
≈31k cells, more than 17k genes

PCA



Existing Methods: t-SNE (van der Maaten & Hinton, *JMLR*, 2008)

- Widely used on single cell data
  - Biology version (Amir et al, *Nature Biotech*, 2013) has 500+ citations
- Attempts to preserve local relationships in both the high and low-dimensional spaces
  - Designed for separating clusters



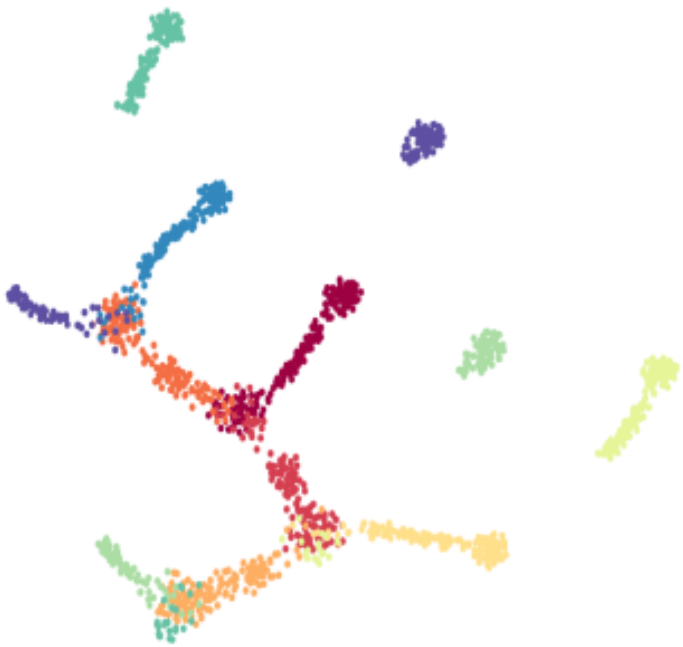
(Amir et al, 2013)

Existing Methods: t-SNE (van der Maaten & Hinton, *JMLR,* 2008)

Artificial Tree Data

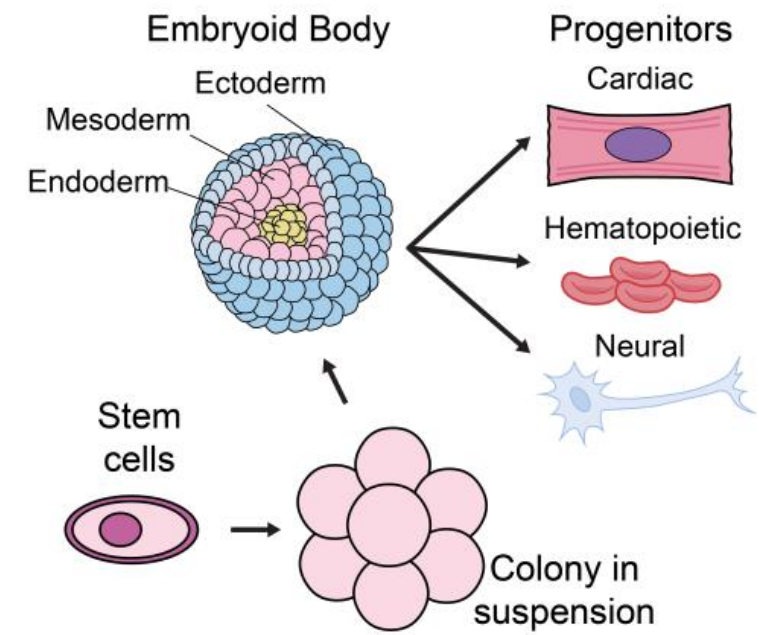
1400 points, 60 dimensions

t-SNE



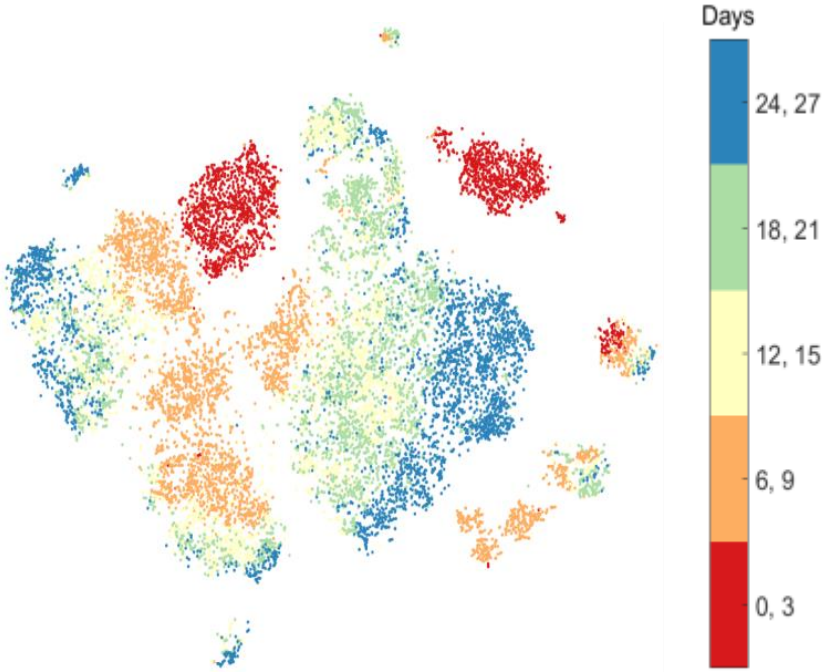
Existing Methods: t-SNE (van der Maaten & Hinton, *JMLR*, 2008)

Newly generated scRNA-seq data (27 days)



≈31k cells, more than 17k genes

t-SNE



Need a New Visualization

- Shows ground truth:
  - Clusters
  - Transitions
  - Local relationships AND global structure
- Denoises

PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding)

Visualizing Transitions and Structure for High Dimensional Data Exploration

Preprint (bioRxiv): `https://doi.org/10.1101/120378`

Code (GitHub): `http://phate.net`

Yale Applied Math & Genetics

Joint work with:

Guy Wolf, David van Dijk, Smita Krishnaswamy,

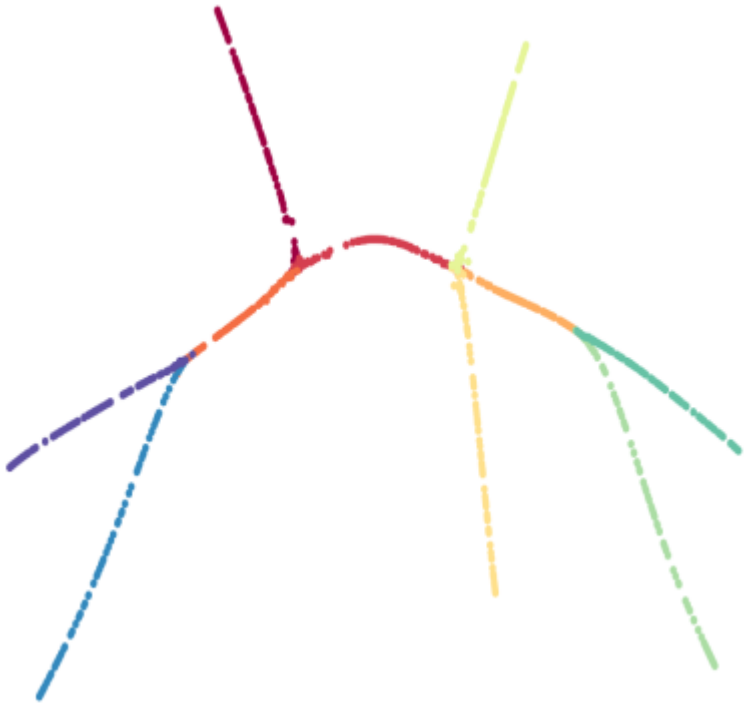
Z. Wang, W. Chen, M.J. Hirn, R.R. Coifman, N.B. Ivanova et al.

Under review at *Nature Biotechnology*

PHATE on Artificial Tree Data

Artificial Tree Data

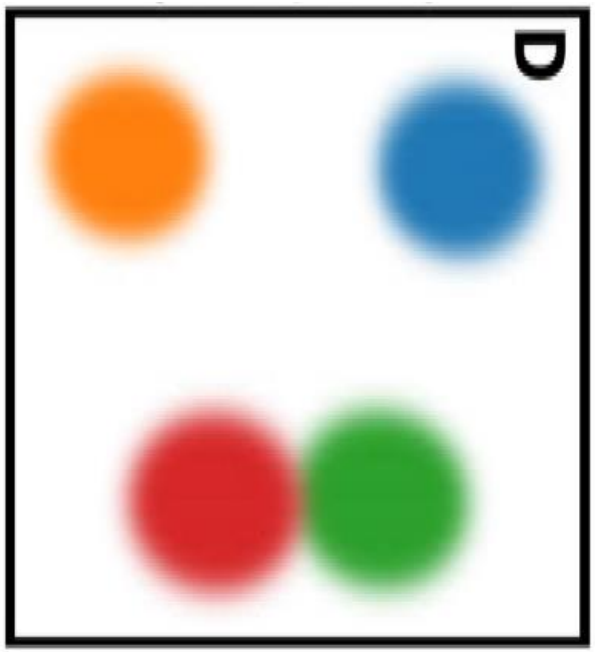
PHATE



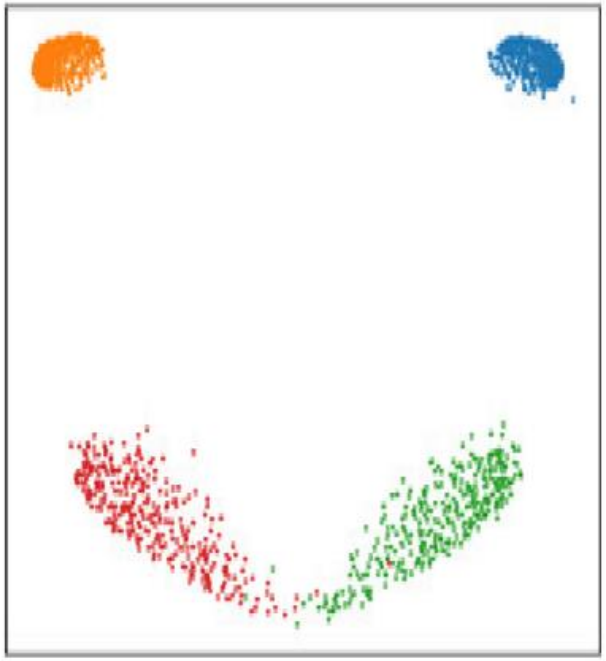
1400 points, 60 dimensions

PHATE on GMM

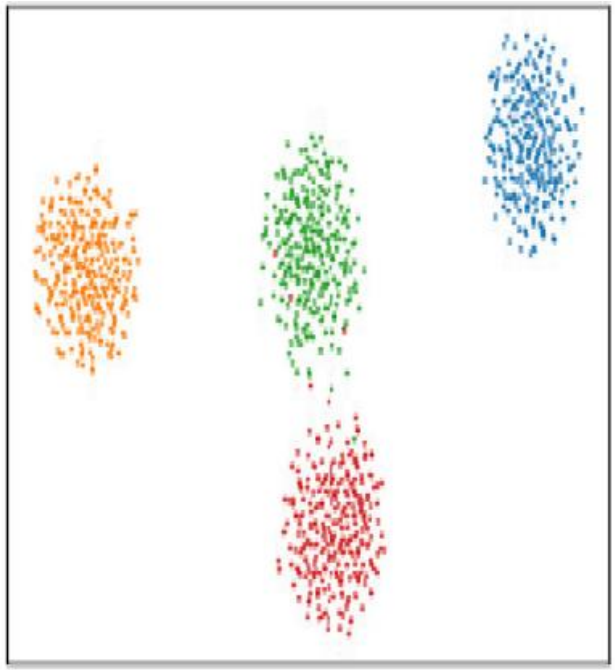
Truth



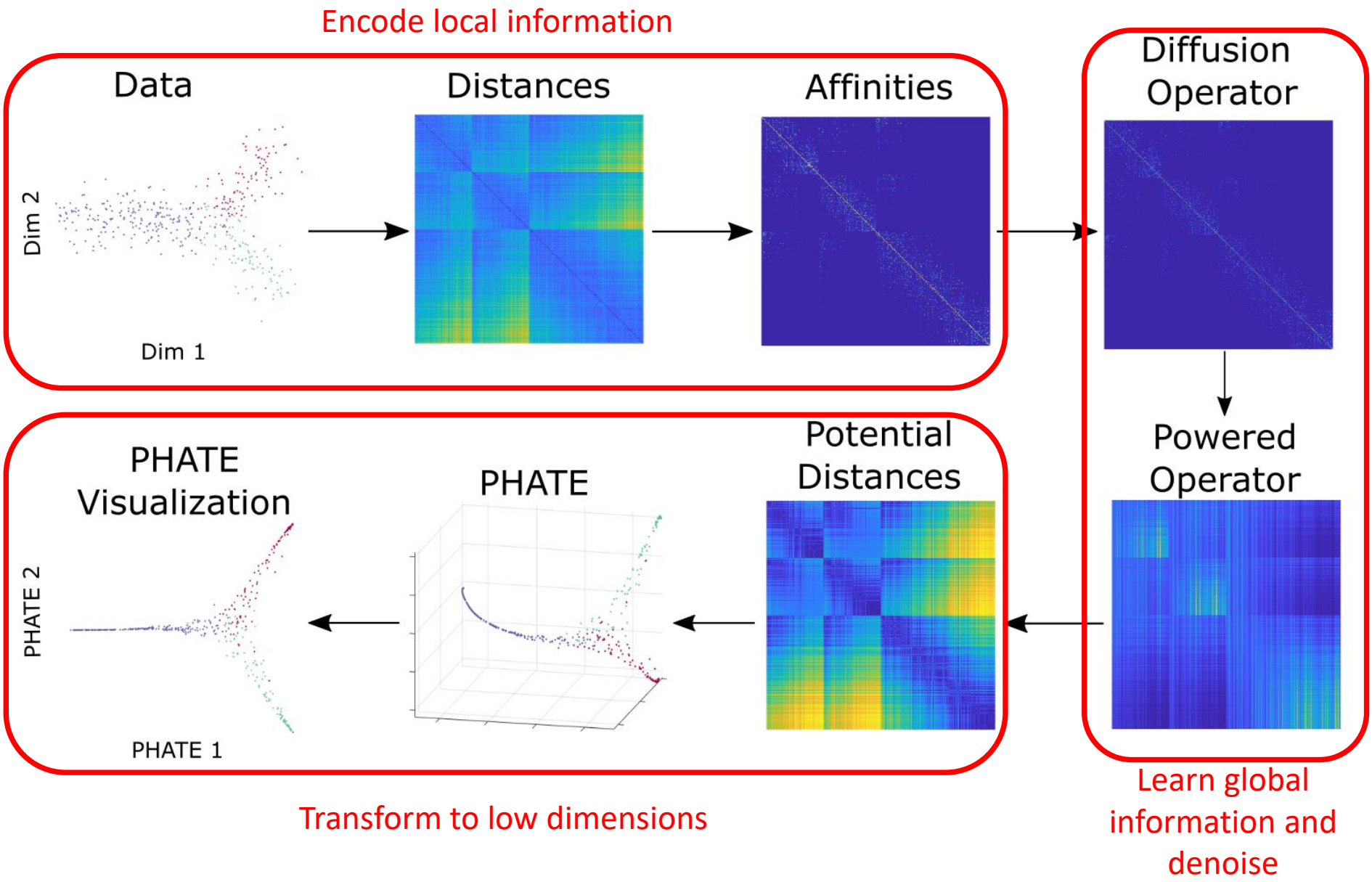
PHATE



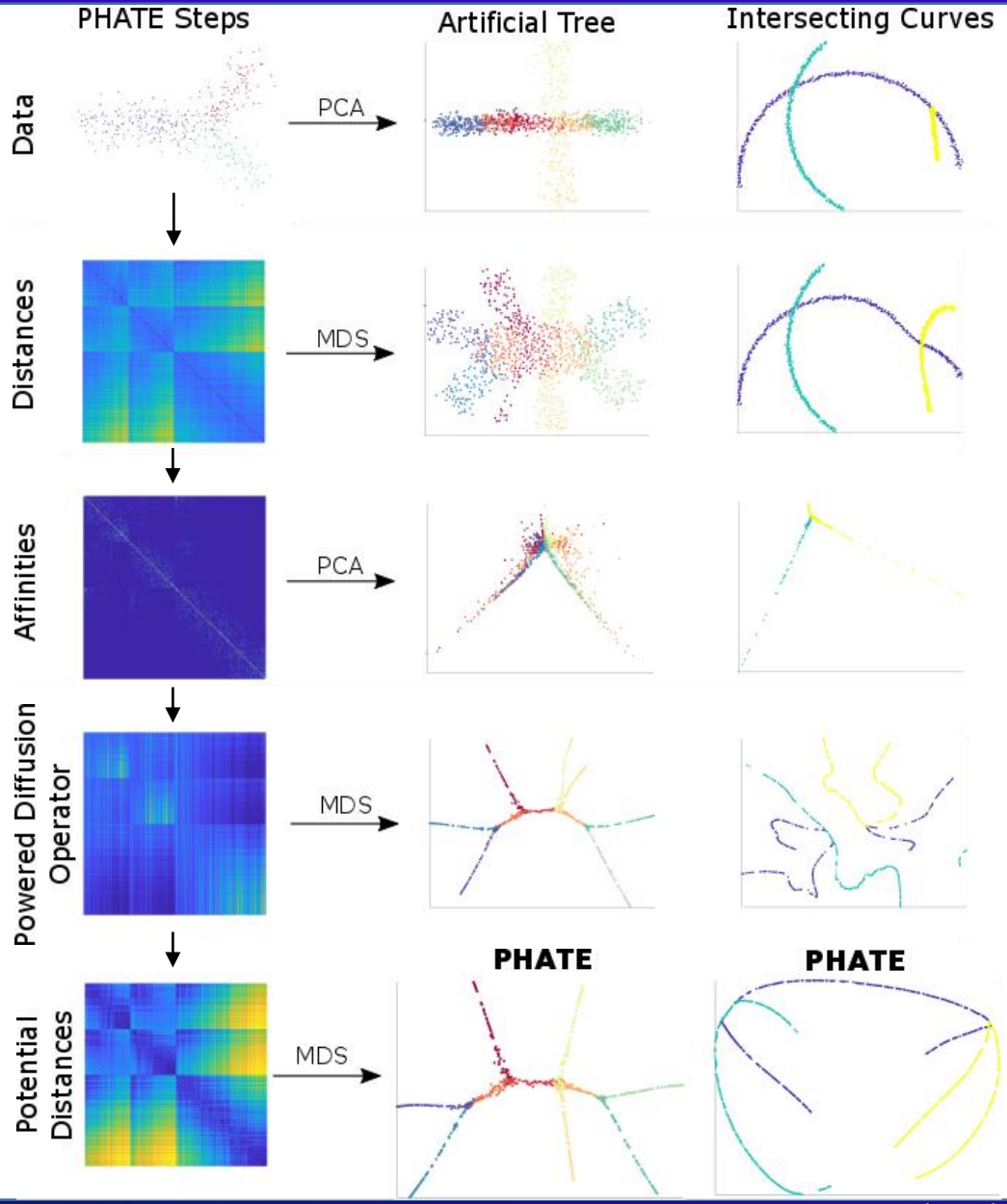
t-SNE



The PHATE Algorithm



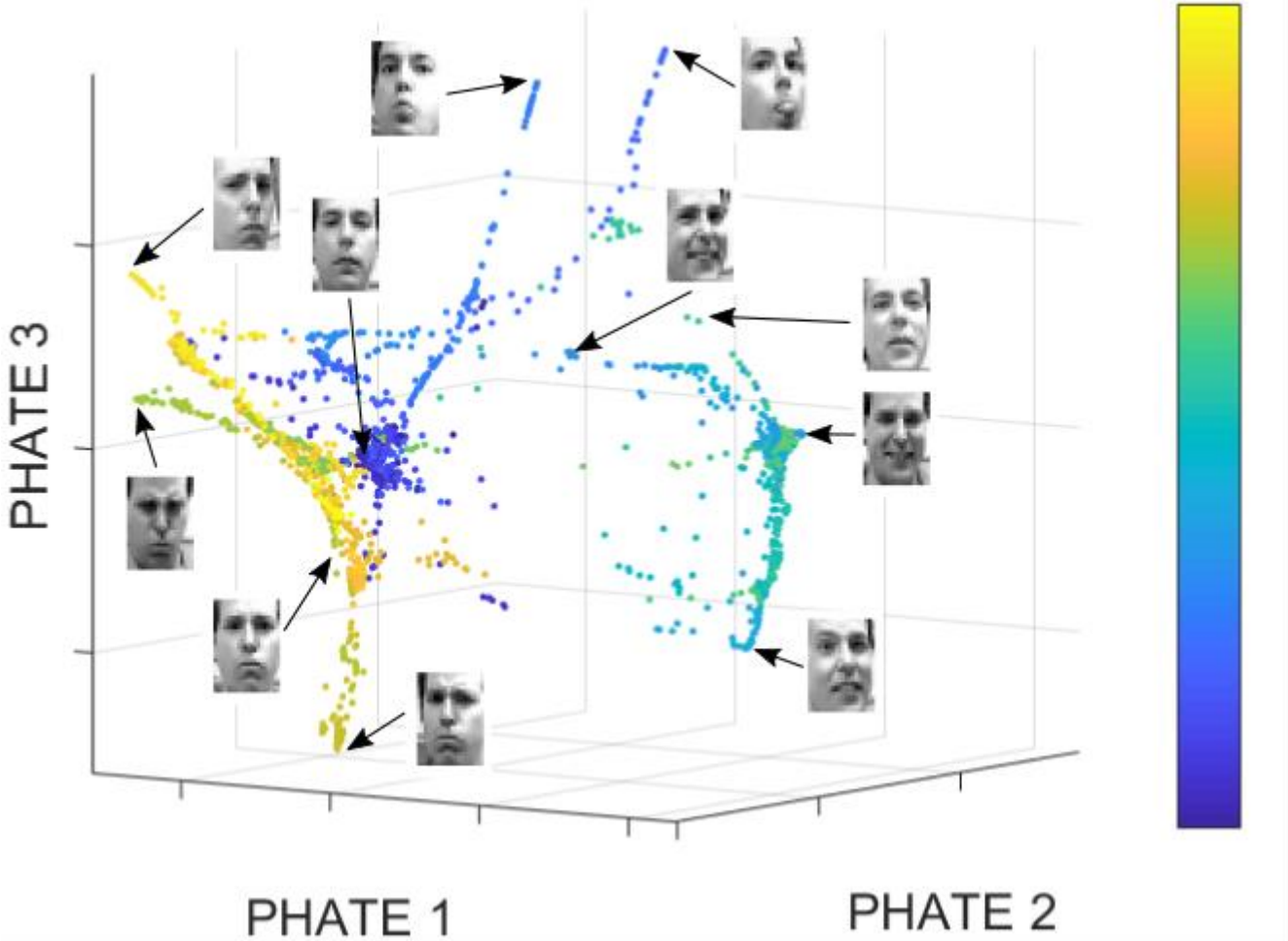
The PHATE Algorithm



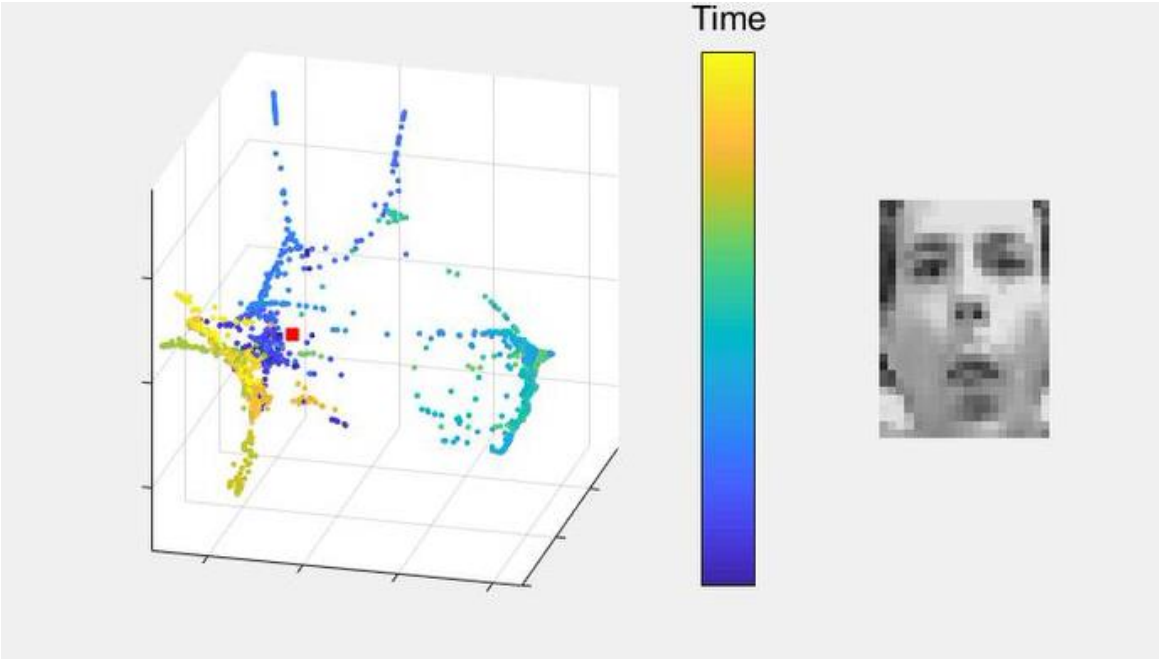
PHATE on Frey Faces

- Frey Faces dataset
  - 1965 frames taken from a video of Brendan Frey making various faces in front of a camera
  - Resolution: 20x28
- Frames are given out of order to PHATE w/o information about sequential ordering

PHATE on Frey Faces

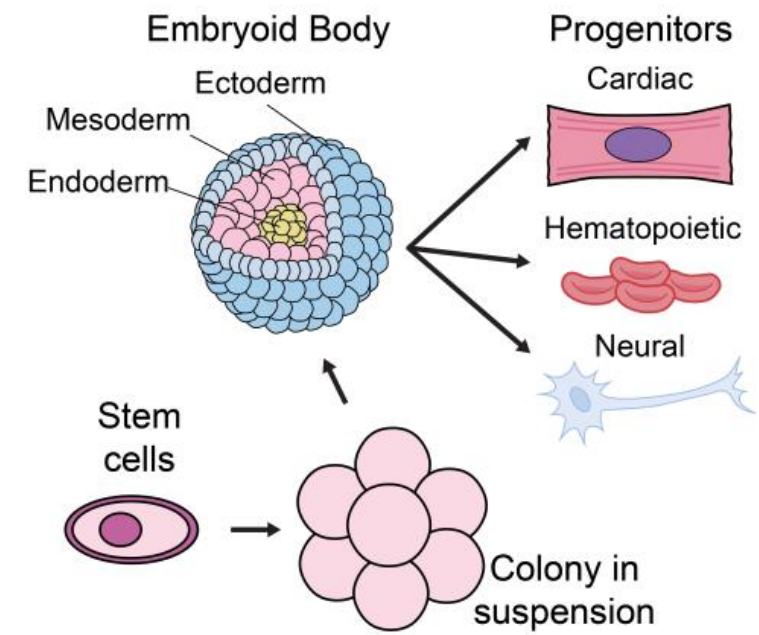


PHATE on Frey Faces

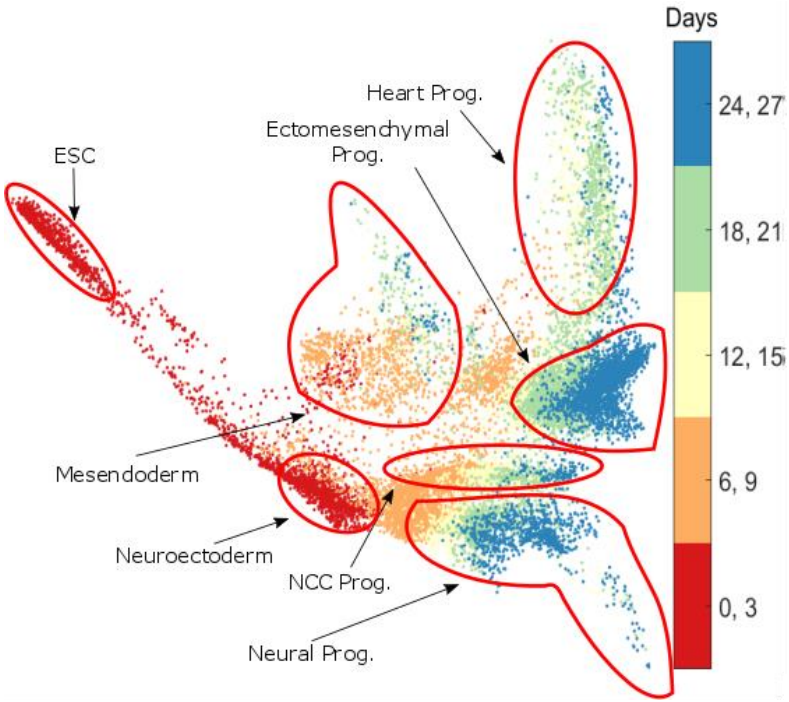


PHATE on EB scRNA-seq Data

Newly generated scRNA-seq data (27 days)



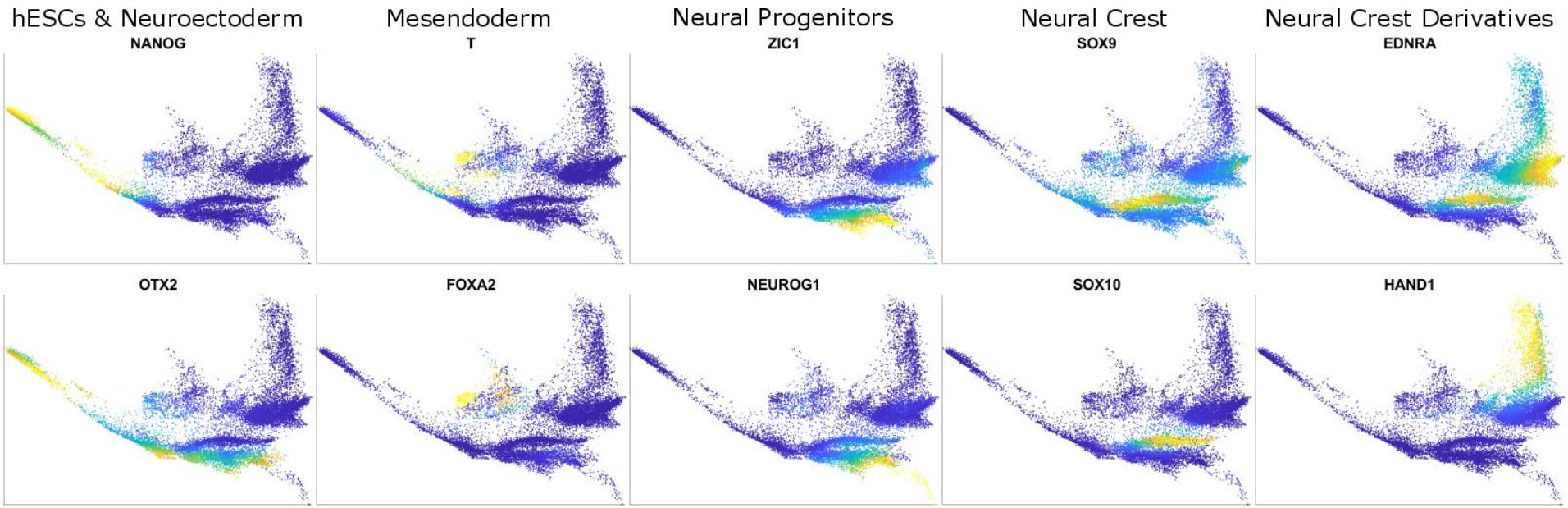
PHATE



≈31k cells, more than 17k genes

Exploratory Data Analysis with PHATE

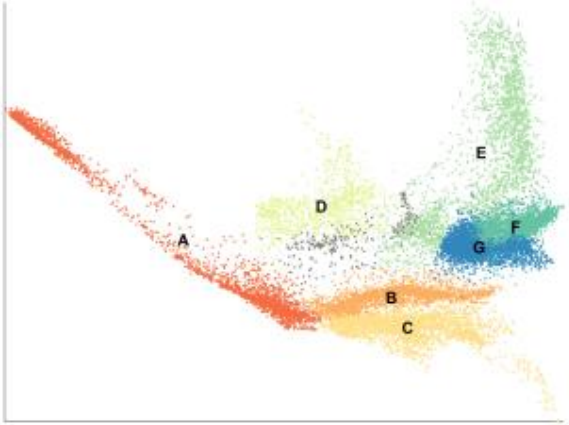
- Coloring the embedding by gene expression after MAGIC* (van Dijk,…Moon et al., 2018) reveals lineages



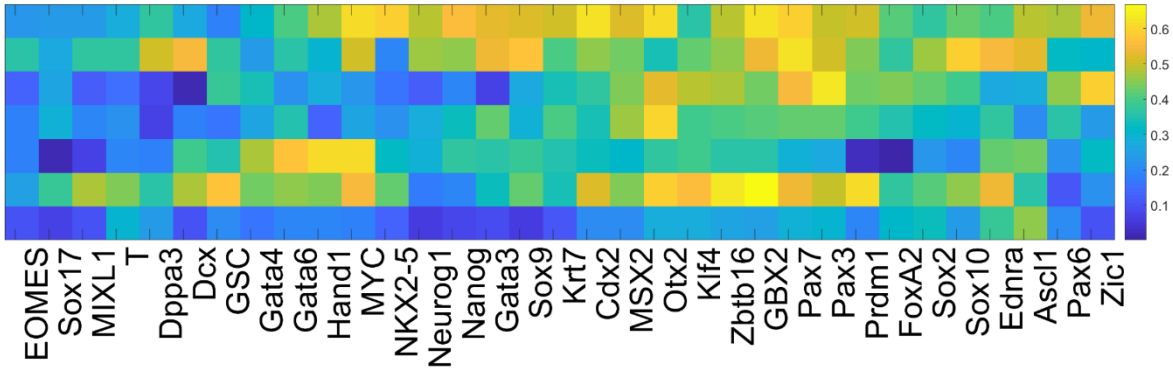
- Interactive web tool: krishnaswamylab.org/phatewebtool

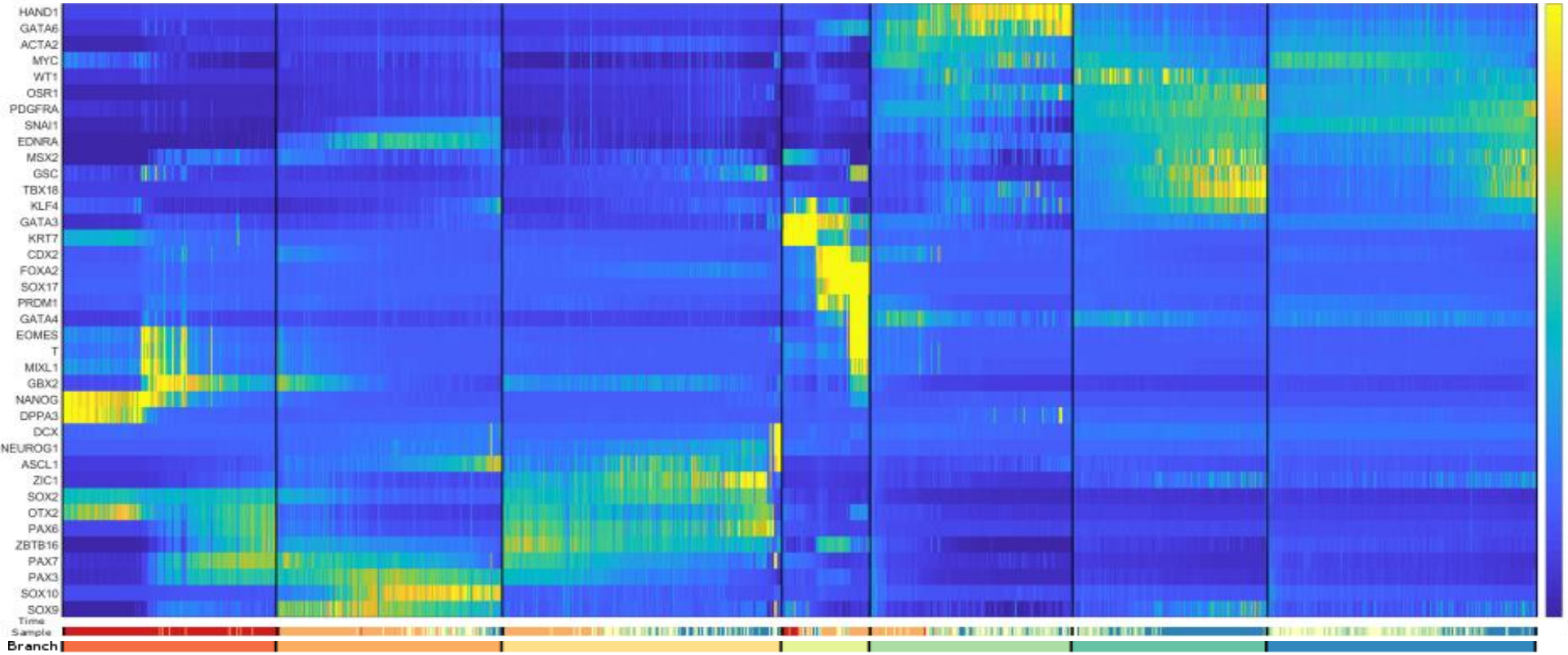
Published in *Cell

Exploratory Data Analysis with PHATE



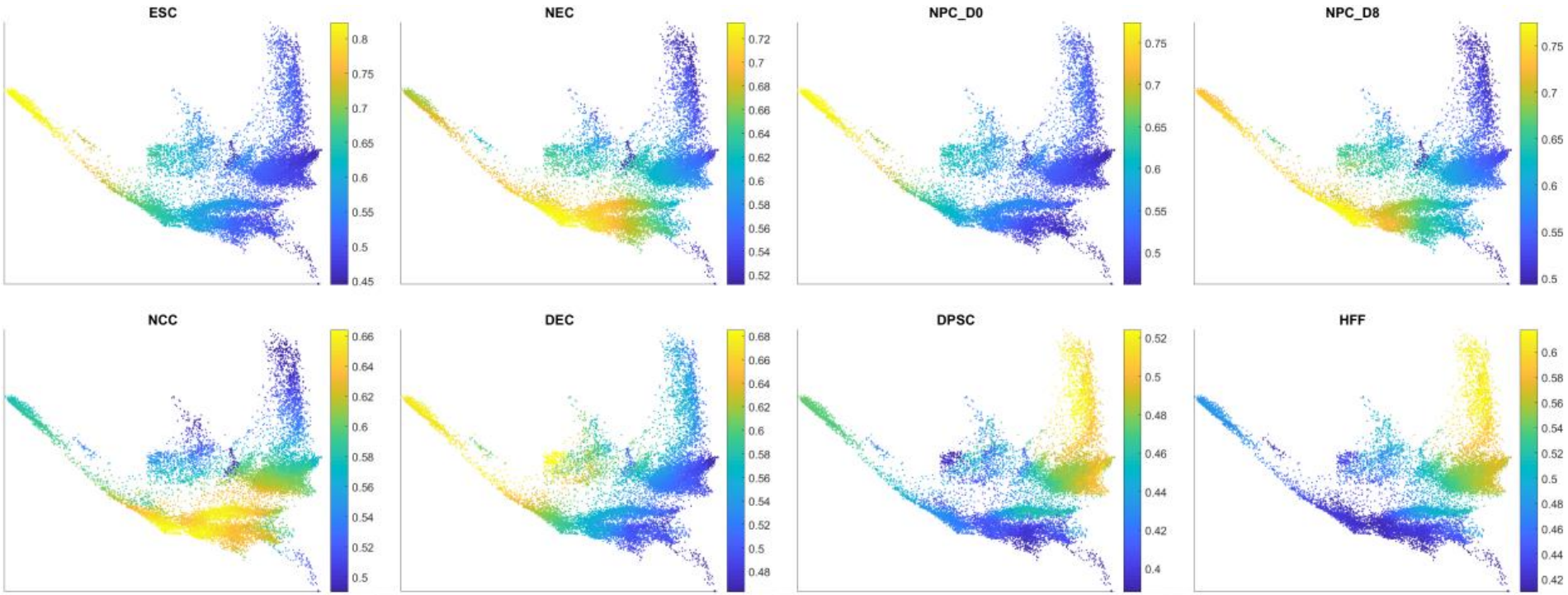
DREMI Score (Krishnaswamy et al., 2014)



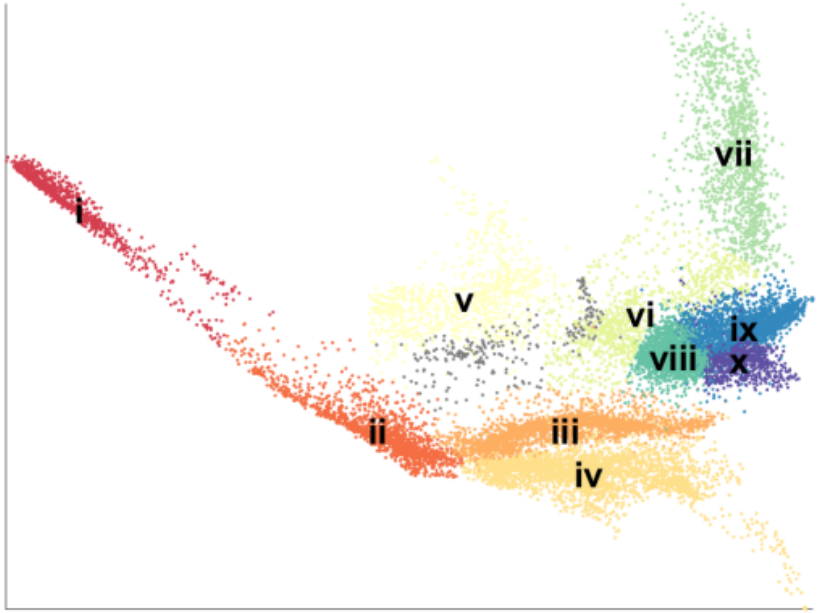


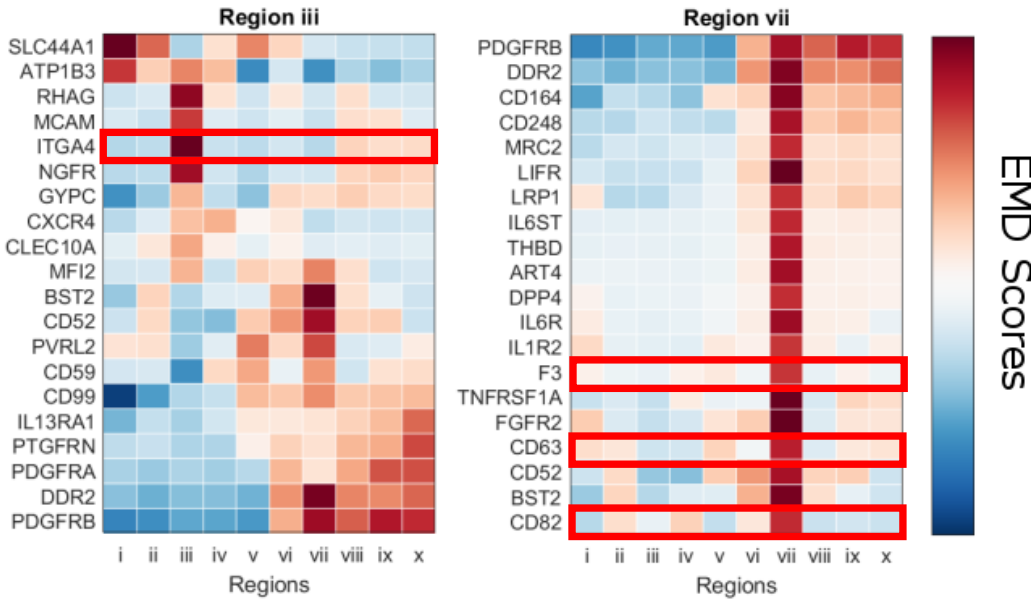
Validation with Bulk RNA-seq

- Spearman correlation with bulk RNA-seq samples

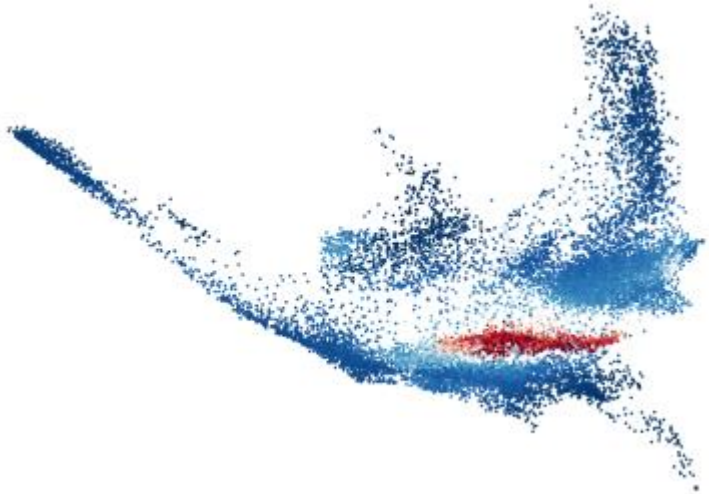


Discovering New Surface Markers for Sorting Populations

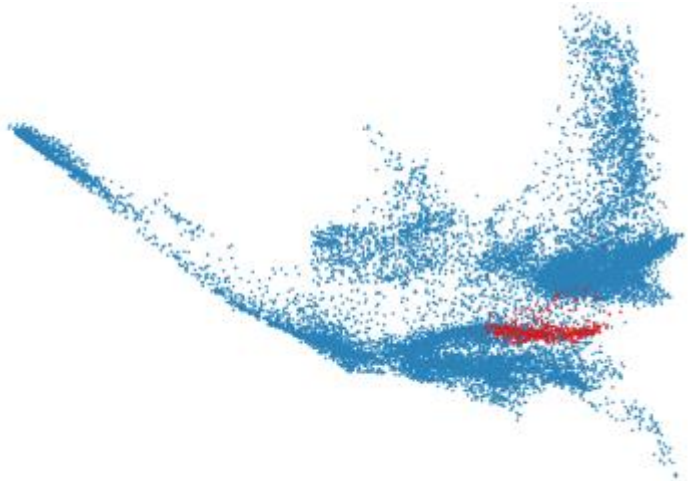




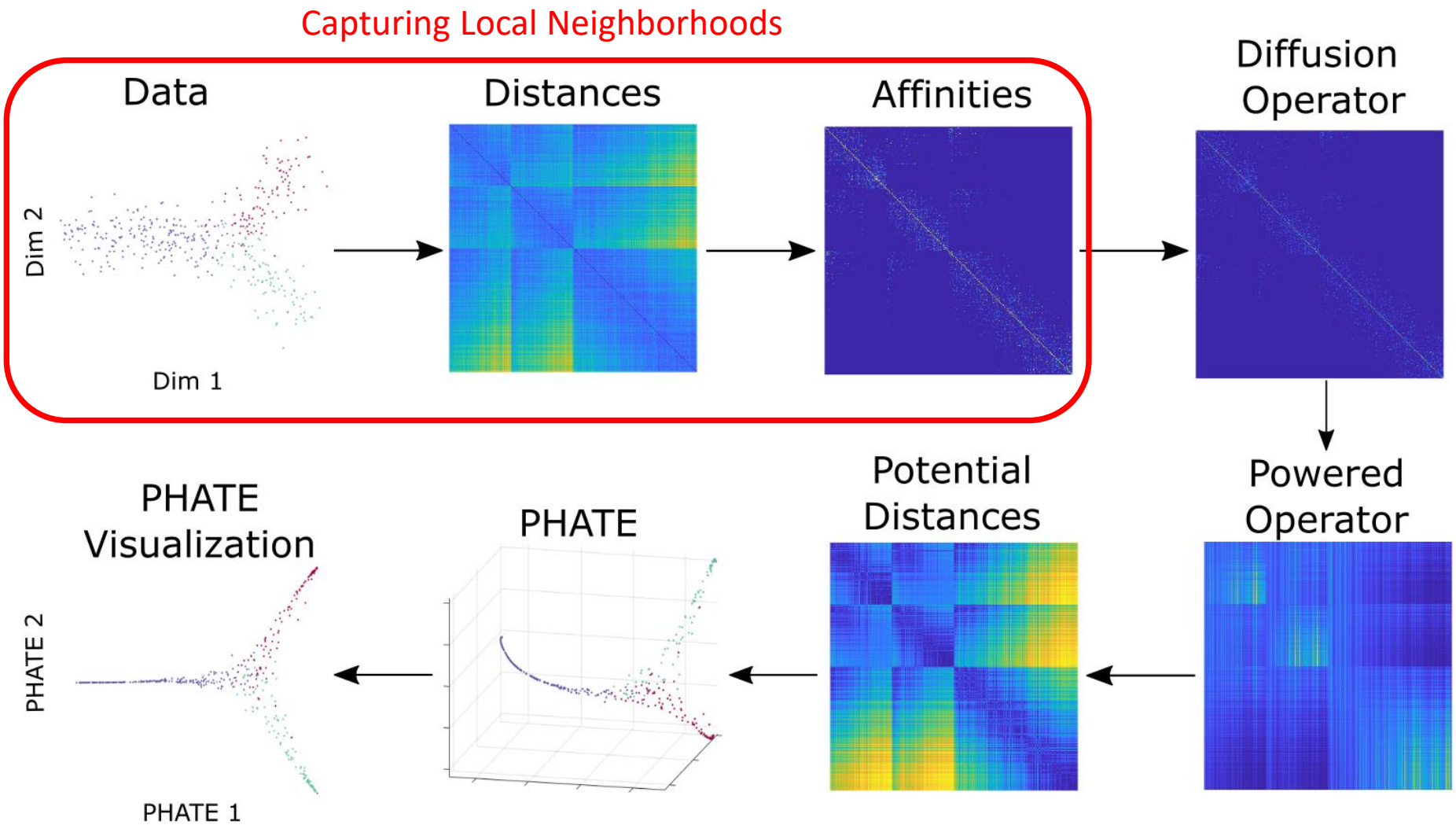
ITGA4



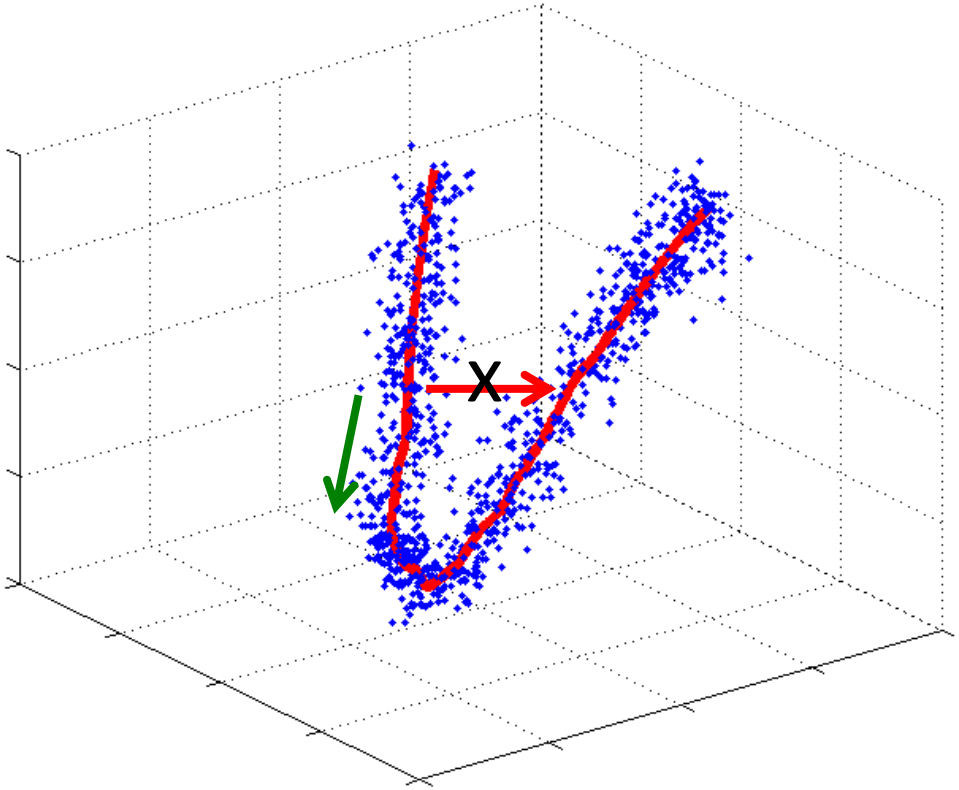
ITGA4 sorted bulk RNA-seq correlation



The PHATE Algorithm



Capturing local neighborhoods



- Large Euclidean distance gives wrong global structure for visualizing
- Small Euclidean distances ok
  - Encodes local structure

From distances to affinities

- Step 1: Calculate all pairwise Euclidean distances
- Step 2: Convert the distances to pairwise affinities using a kernel function
  - E.g. the Gaussian kernel
  - Kernel function must be near zero for large distances and nonzero for small distances
  - Affinity: a measure of similarity between points
- Problem: many data have both dense and sparse regions
- A kernel bandwidth fixed for dense regions doesn't work well for sparse regions and vice versa
- Idea: use locally adaptive bandwidth

Capturing Local Neighborhoods

- Nonuniform densities
- **Adaptive α-decaying kernel**

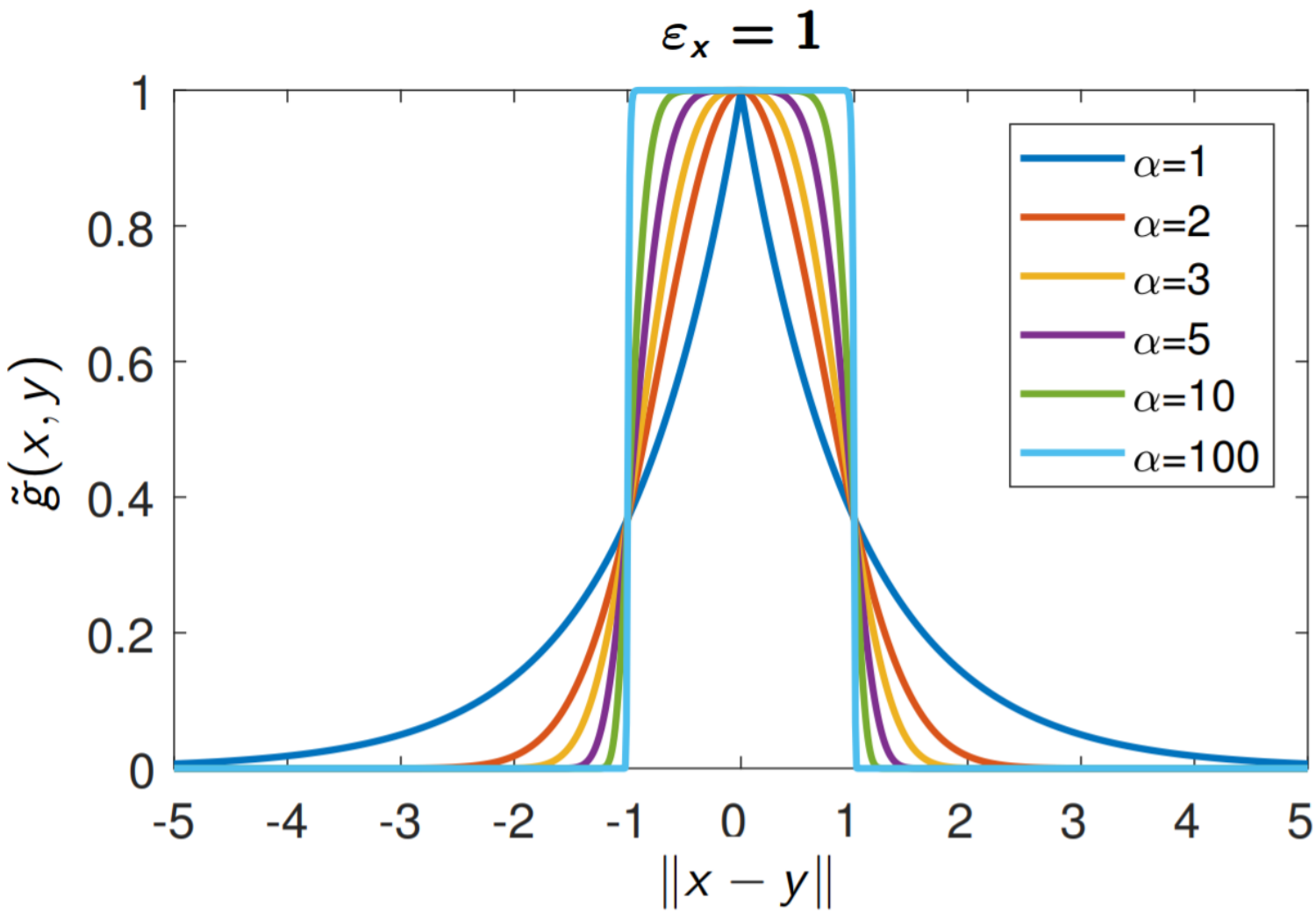
$$\tilde{g}(x,y) = \exp\left(-\left(\frac{||x-y||}{\epsilon_x}\right)^{\alpha}\right), \rightarrow g(x,y) = \frac{\tilde{g}(x,y)+\tilde{g}(y,x)}{2}$$

Where

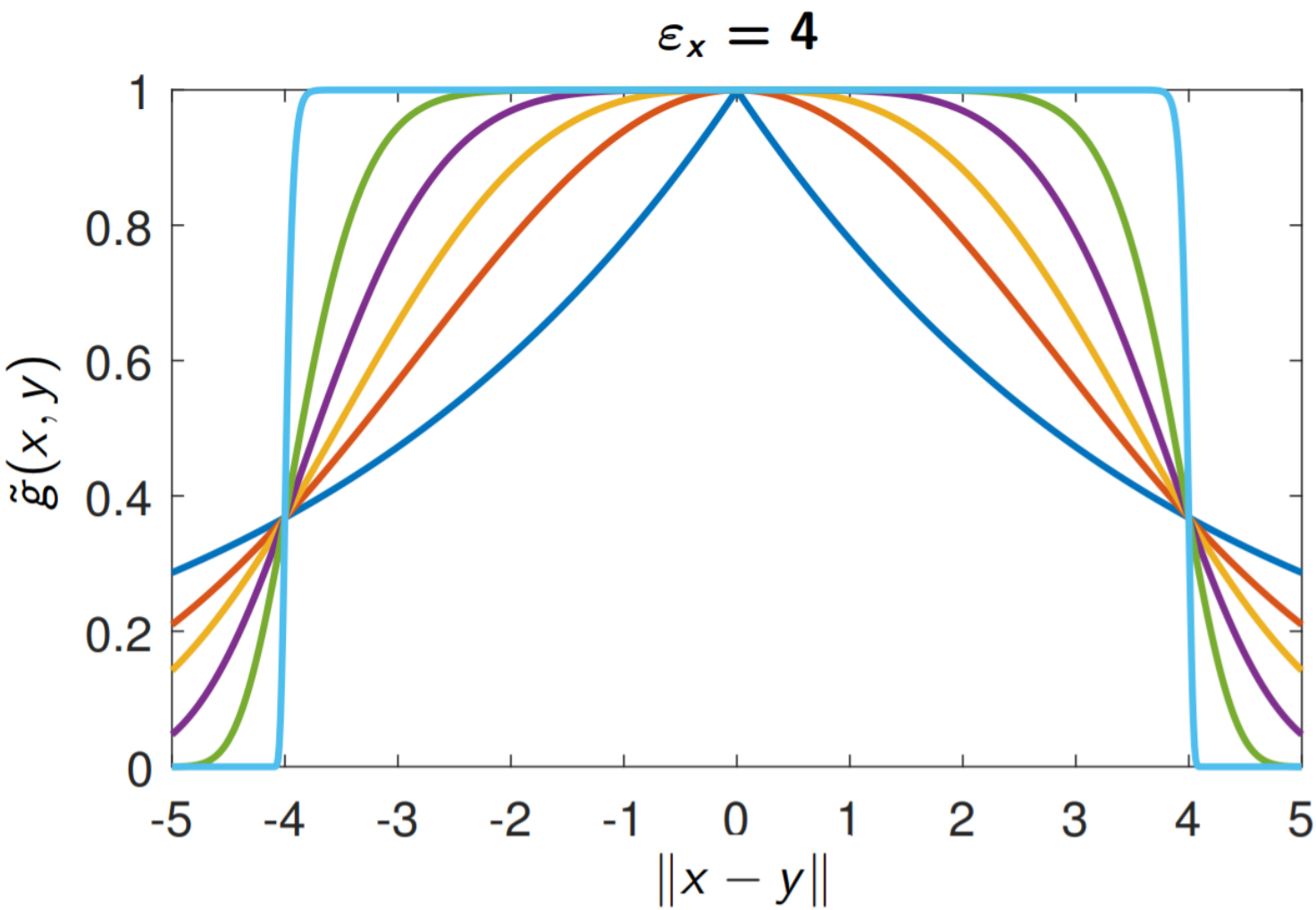
- ϵ_x = distance from x to its kth nearest neighbor
- α controls the decay rate of $\tilde{g}$

- Provides a robust notion of adaptive locality

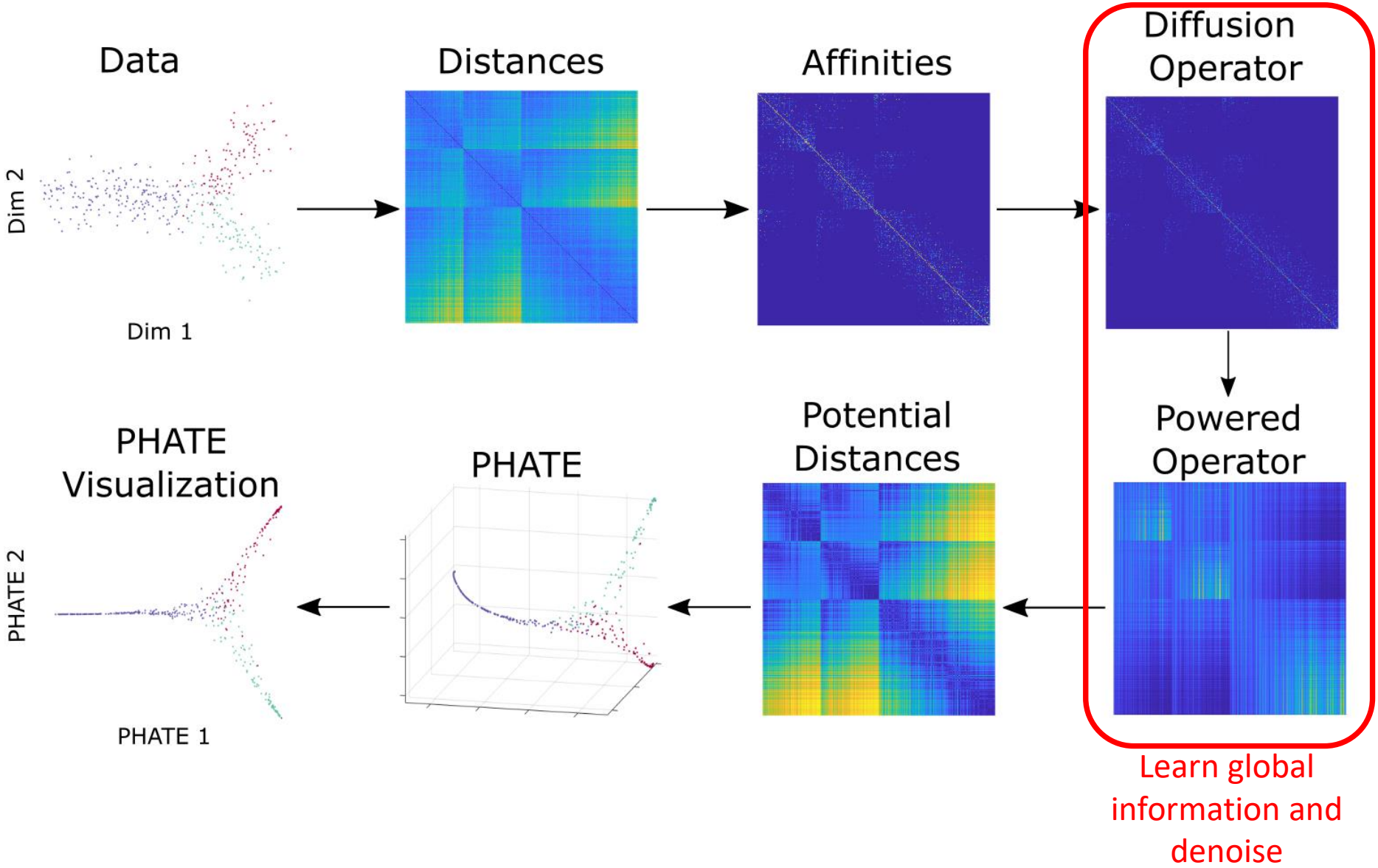
Capturing Local Neighborhoods



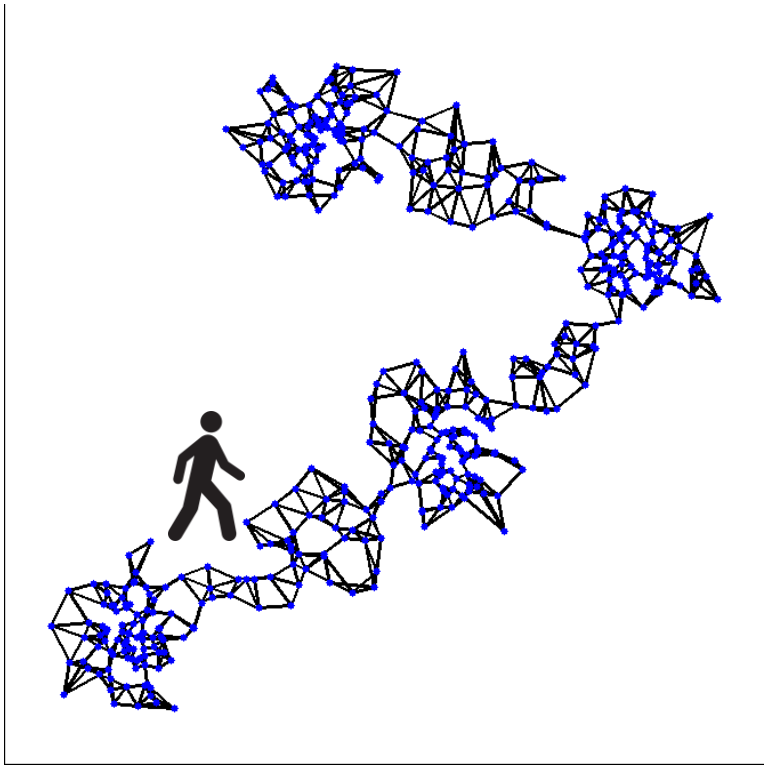
Capturing Local Neighborhoods



The PHATE Algorithm



Diffusion Denoises and Recovers Global Structure



- Big Euclidean steps bad, likely to exit structure
- Small steps good, likely to stay within structure
- To learn global structure via small local steps we use random walk (diffusion)
  - Normalize affinity matrix to create Markov transition matrix (the diffusion operator) P (Coifman & Lafon, *ACHA*, 2006)
  - Power P by time step t

How much diffusion?

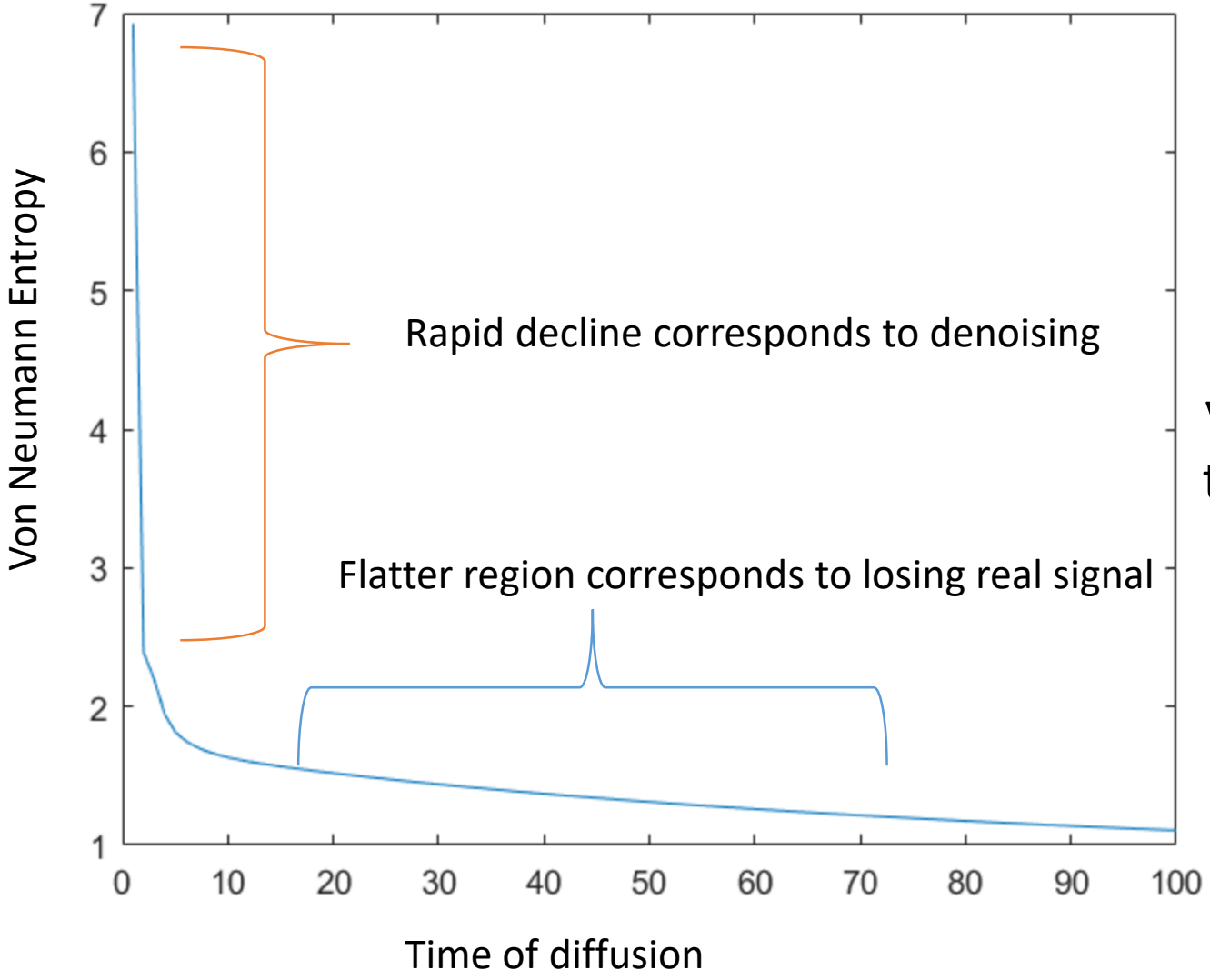
- Von Neumann entropy (von Neumann, 1932) of diffused operator P^t

$$VNE(P^t) = -\sum_j \eta_j \log \eta_j \,, \eta_j = \lambda_j^t / \left|\left|\boldsymbol{\lambda}^t\right|\right|_1$$

Where $\boldsymbol{\lambda}^t = \{\lambda_0^t, \lambda_1^t, \dots\}$ are the eigenvalues of P^t

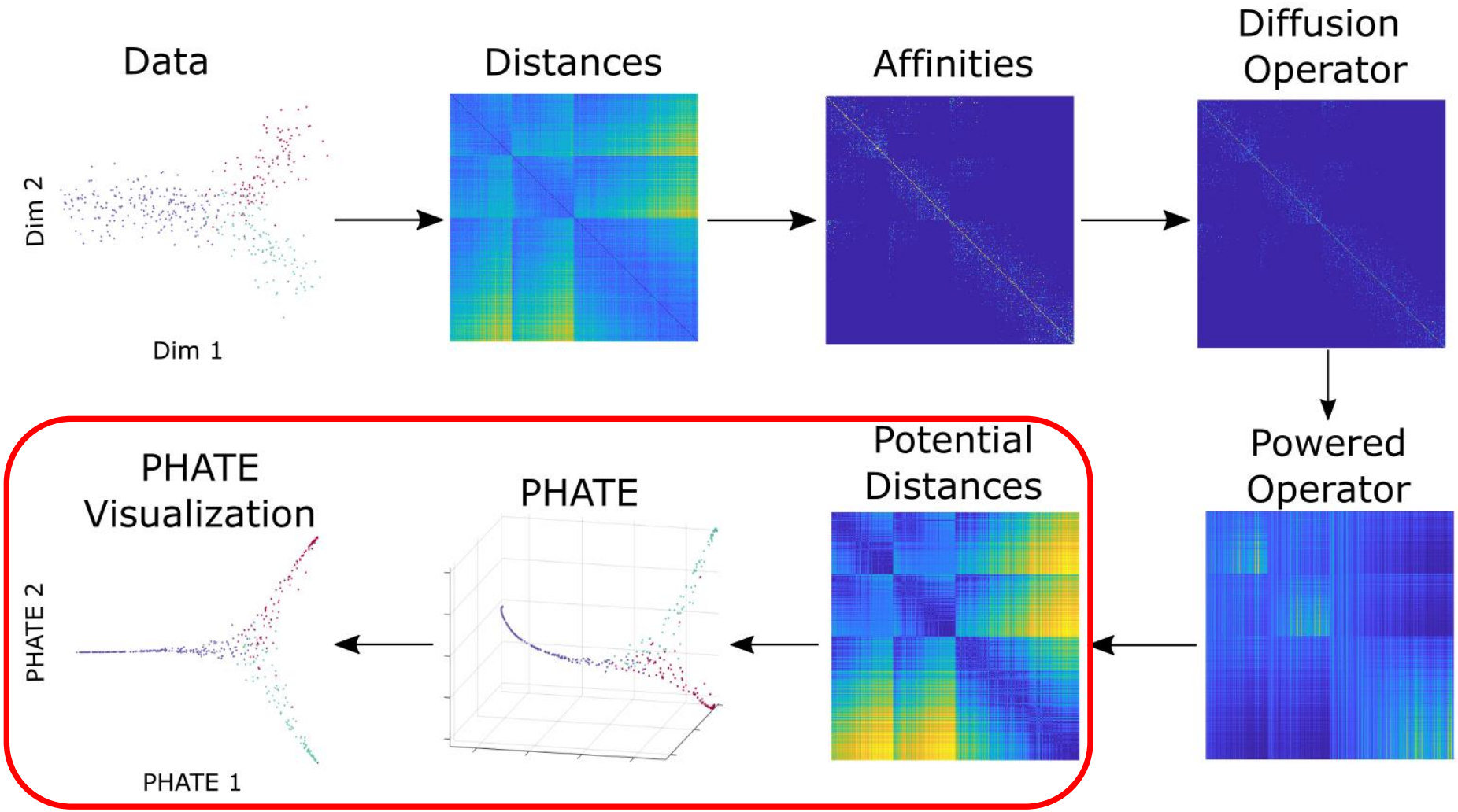
- VNE is a soft proxy of numerical rank
- Decays as diffusion time increases, $\lim_{t \to \infty} VNE(P^t) = 0$
- Use rate of decay to choose time scale

How much diffusion?



Embedding is visually robust in the flatter region

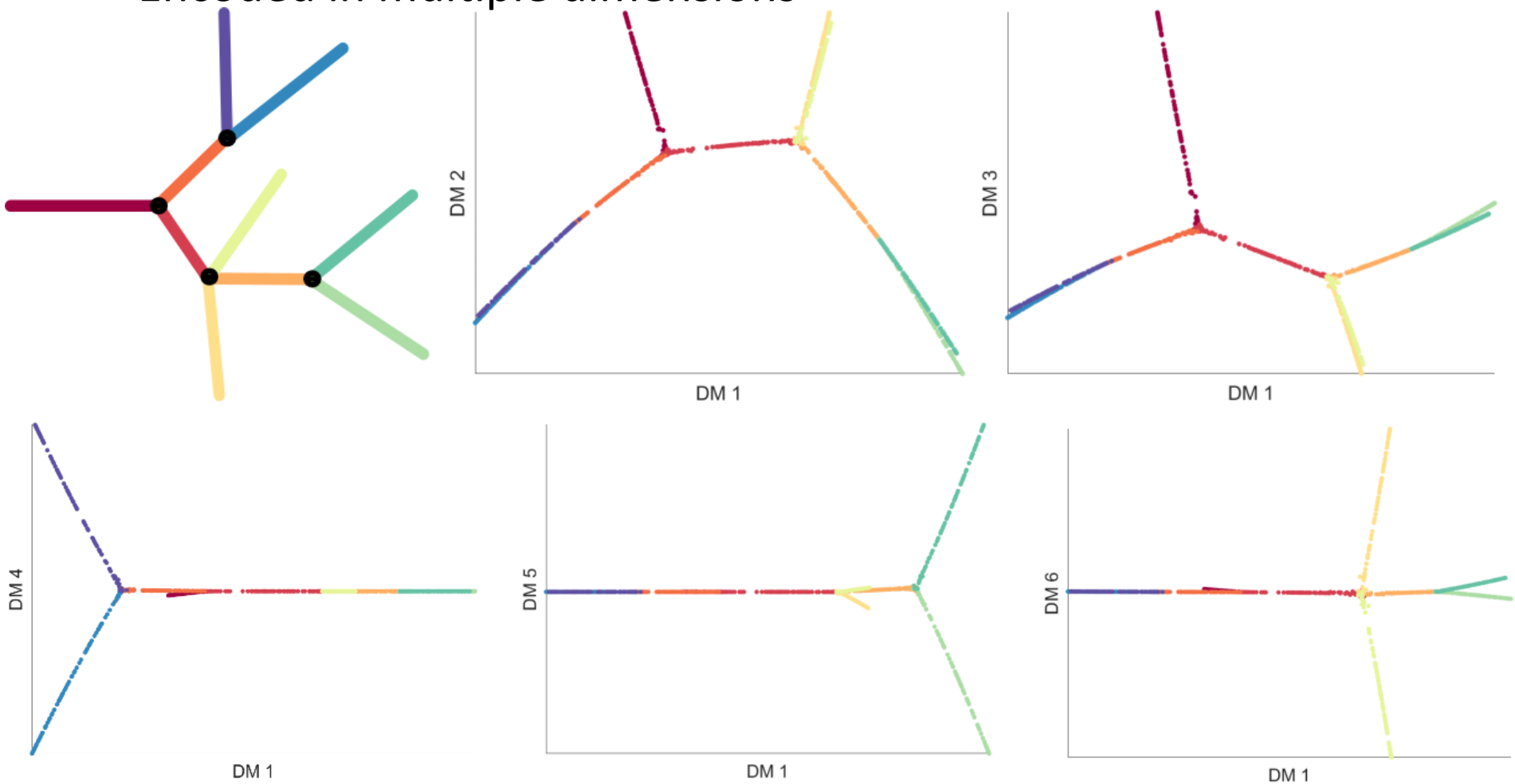
The PHATE Algorithm



Transform to low dimensions

Potential Distances

- Diffusion operator contains global and local structure
  - Encoded in multiple dimensions



Potential Distances

- Diffusion operator contains global and local structure
  - Encoded in multiple dimensions
- To extract this information, we transform diffused probabilities using a ***potential transformation***
  - Forms an information distance between diffused probabilities

Proposition 1 (Moon, van Dijk, Wang et al., 2018)

The squared potential distance is an M-divergence on the diffusion probabilities.

Potential Distances

- Diffusion operator contains global and local structure
  - Encoded in multiple dimensions
- To extract this information, we transform diffused probabilities using a ***potential transformation***
  - Forms an information distance between diffused probabilities
  - Connected to heat potential

Proposition 2 (Moon, van Dijk, Wang et al., 2018)

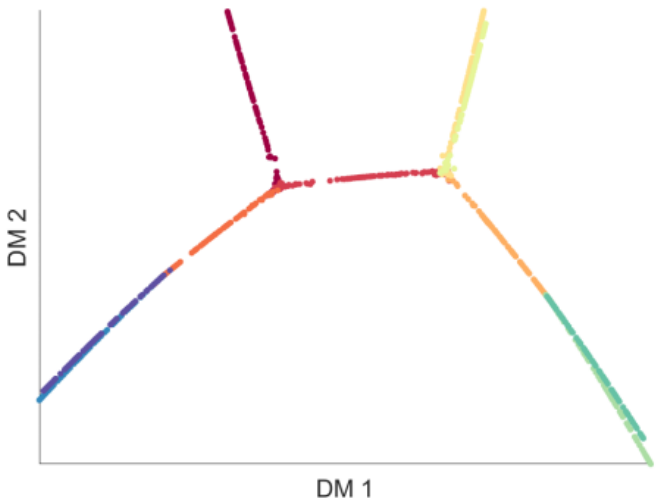
Given a diffusion process defined by a fixed-bandwidth Gaussian kernel, the potential distance at time-scale t can be written as

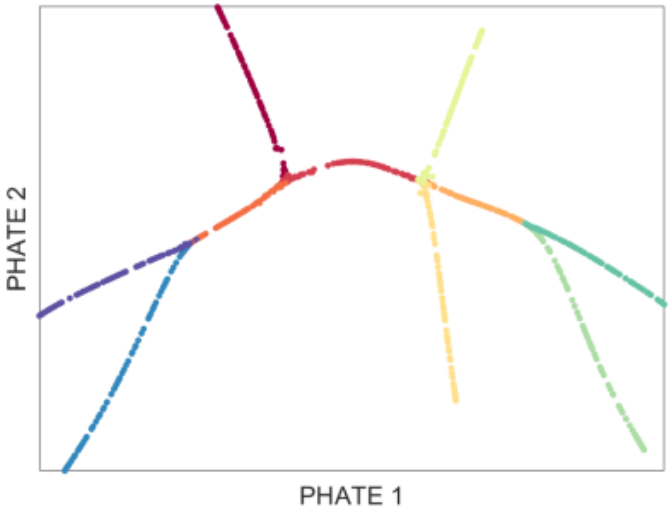
$$[\mathcal{D}_t(x,y)]^2 = \sum_z \log^2\left(\frac{1+\langle\phi(x),\phi(z)\rangle}{1+\langle\phi(y),\phi(z)\rangle}\right),$$

where $\phi(\cdot)$ is the Diffusion Map embedding at time-scale $t/2$.

Potential Distances

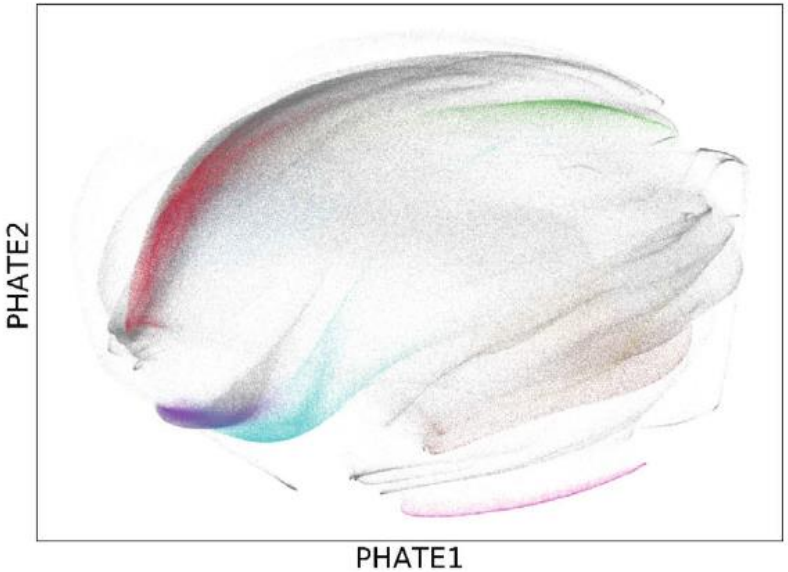
- Diffusion operator contains global and local structure
  - Encoded in multiple dimensions
- To extract this information, we transform diffused probabilities using a ***potential transformation***
  - Forms an information distance between diffused probabilities
  - Connected to heat potential
- Embed for visualization using multidimensional scaling (MDS)





Scalability of PHATE

- Storing and performing operations on the diffusion matrix can be difficult for large samples
- Can reduce computation by diffusing through “landmarks”
  - Landmarks are chosen by clustering
  - Obtain an embedding of all points by projection



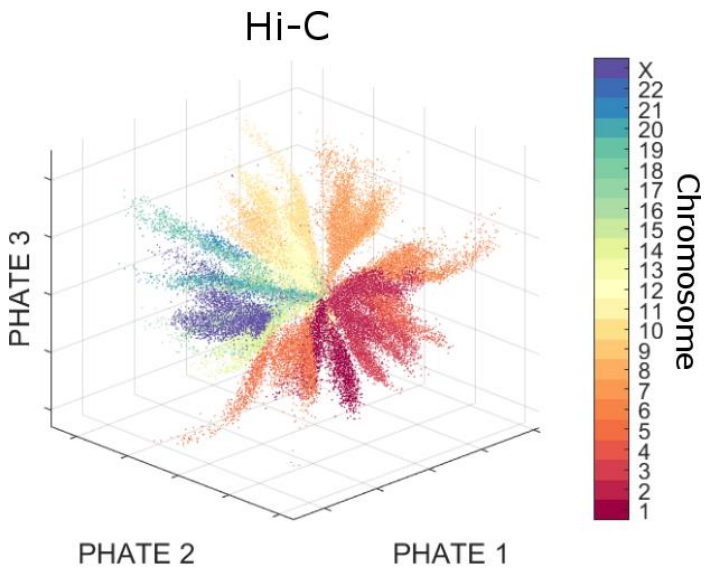
PHATE applied to the 10x megacell mouse brain data (>1 million cells)

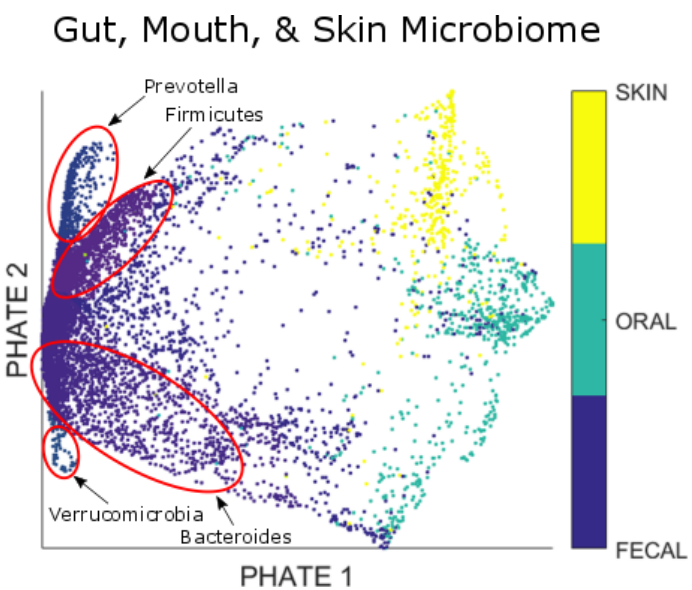
Impact of PHATE

- Visualizing data from microbiome, Hi-C chromatin conformation maps, Facebook networks, brain organoid scRNA-seq, Zika/Dengue Fever CyTOF, flu CyTOF, mouse iPSC CyTOF, mouse bone marrow scRNA-seq, human SNP

Other researchers are using PHATE

- Studying autism in neural cortex of developing mouse embryo at Yale
  - PHATE highlights change in key gene
- EEG data in sleep study at Michigan
  - PHATE highlights "spectrum of how the EEG changes vs. sleep stage in neonates"





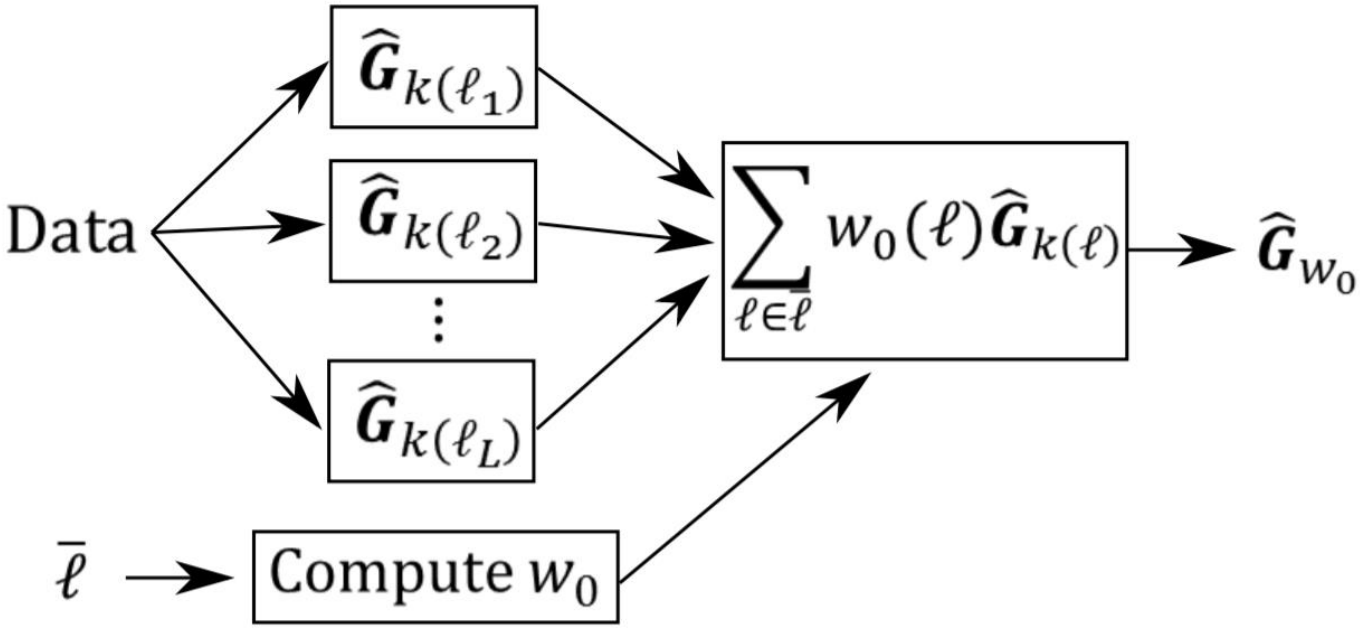
Future Work: Further Analysis of PHATE

- Is the potential distance the optimal distance for visualization?
- Can we use the VNE in a more general setting for choosing scale?
- Can we better relate the choice of k, α, and t to other problems such as density estimation?
- Can we choose something better than Euclidean distance to better denoise the data?
- Can we extend the visualization to a systems view?
  - Compare systems in different states (e.g. diseases) to determine similarities

Future Work: Higher-Order Gene Interactions

- Deep learning-based transcoder
  - Learn a mapping between distributions of cells in different states
  - Learn the genes driving transitions
- Information measures
  - Use **ensemble estimation** techniques to overcome the curse of dimensionality (Moon, Sricharan et al, 2017; Moon, Noshad et al., 2017)

Ensemble Estimation



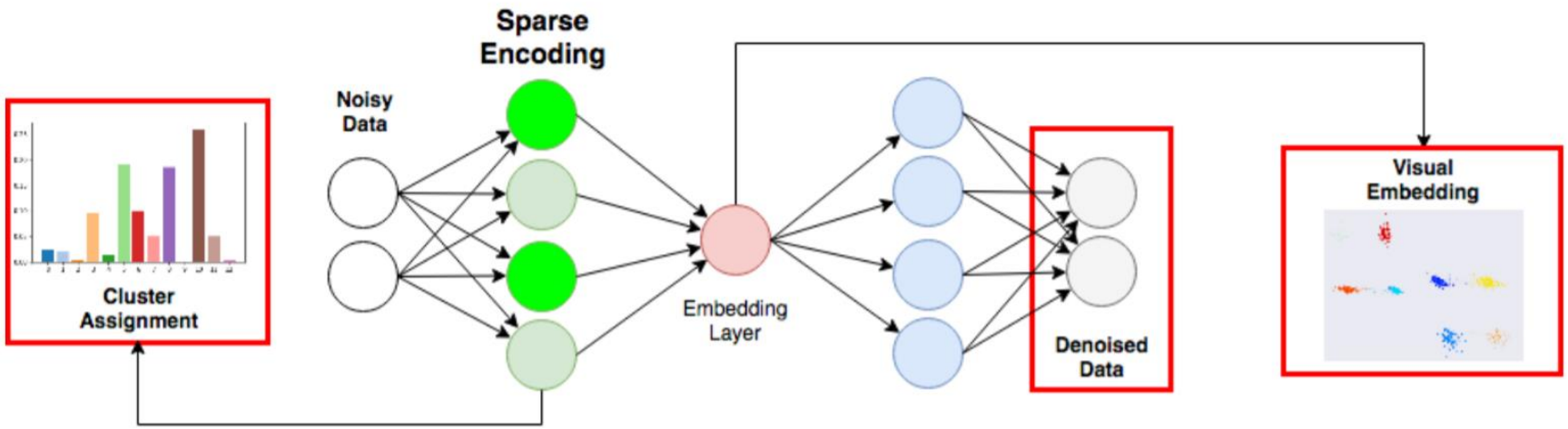
- A weighted average of weak learners/estimators to achieve superior performance
  - E.g., k-nearest neighbor plug-in estimators of entropy with varying k
  - Weights computed based on the performance measure (e.g. MSE for estimation) of the weak learner/estimator
  - Achieve the parametric MSE rate for sufficiently smooth densities
  - Simpler than competing estimators

Ensemble Estimation

- KDE plug-in divergence estimation (ISIT 2016)
  - Ensemble of bandwidth parameters
  - Central limit theorem
  - Uniform convergence over class of bounded Holder smooth densities
- KDE plug-in mutual information estimation (ICASSP 2017; ISIT 2017)
  - Continuous case
  - Mixture of discrete and continuous random variables (first nonparametric estimators that achieve the parametric rate)
  - Central limit theorem
- k-nn approaches (ISIT 2014, NIPS 2014, ISIT 2017, ICASSP 2018)
  - Ensemble of k values
  - Central limit theorem

SAUCIE

- Sparse Autoencoder of Unsupervised Clustering, Imputation, and Embedding (on bioArxiv)



- Novel batch-level regularizations
  - Information Dimension: adapt # of clusters by penalizing entropy of neuron activations over each mini-batch
  - Max-Mean Discrepancy: resolve batch effects by encouraging cross-batch distribution consistency

Joint work w/ M. Amodio, K. Srinivasan, D. van Dijk, S. Krishnaswamy, G. Wolf, et al.

Future Work: Comparing Systems in Different States

- Perturbations (e.g. Crispr/CAS9)
- Different time points
- Diseases/treatment

- Need to compare continuous systems
  - Dimensionality reduction and system matching
  - Information divergence (new theory required)
- Visualize the perturbations/states to obtain larger understanding

- Batch effects/machine effects can be big
  - Normalization via spike-in samples
  - Mutual nearest neighbors

Future Work: Extension to Image Analysis

- Single-cell imaging technologies are becoming popular
  - Provides spatial information about cells
  - CyTOF imaging (Giesen et al., *Nature Methods*, 2014)
  - Multiplexed ion beam imaging (Angelo et al., *Nature Methods*, 2014)

- Challenges
  - Symmetries
  - Deformations

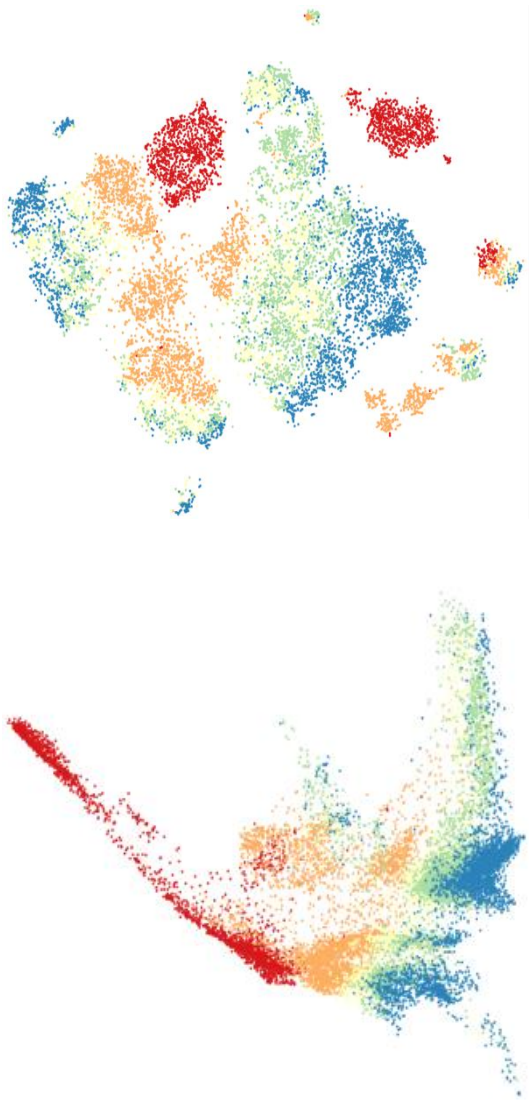
- Potential approaches
  - Image patches
  - Information divergence between images

Other projects I am working on

- Semi-supervised ensemble estimation of information measures
- Deep learning applied to financial data
- Neural networks for analyzing single-cell data
- Model selection in ecological networks for disease prediction
- Machine learning applied to control problems

Conclusion

- Data have structure at different scales
  - Local branching structure
  - Global relationship between branches
- Existing visualization methods fail to account for all scales
- PHATE captures both global and local structure
  - Innovative diffusion process
  - Selection of t
  - Potential transformation
  - Branch detection
- PHATE reveals new biology
- Lots of exciting research directions remain



Questions?

References

- Gliske et al. (2015)
- Li et al, “The Gut Microbiota and Autism Spectrum Disorders,” *Front Cell Neurosci,* 2017.
- Scher et al, “Microbiome in Inflammatory Arthritis and Human Rheumatic Diseases,” *Arthritis Rheumatol*, 2016.
- Pennisi, “Your gut bacteria are more than what you eat,” *Science Mag*, 2016.
- Van der Maaten & Hinton, “Visualizing Data using t-SNE,” *J. Machine Learning Research*, 2008.