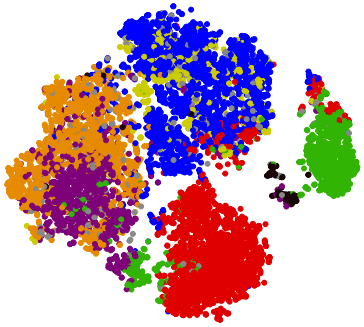
Understanding Algorithms

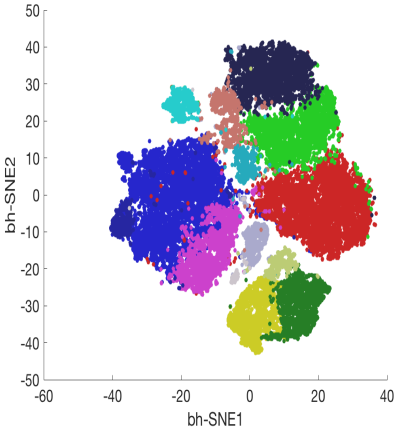
viSNE

Single cell resolution



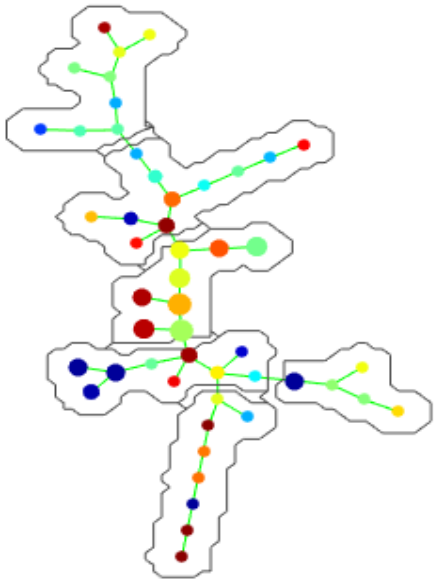
PhenoGraph

Automated population identification



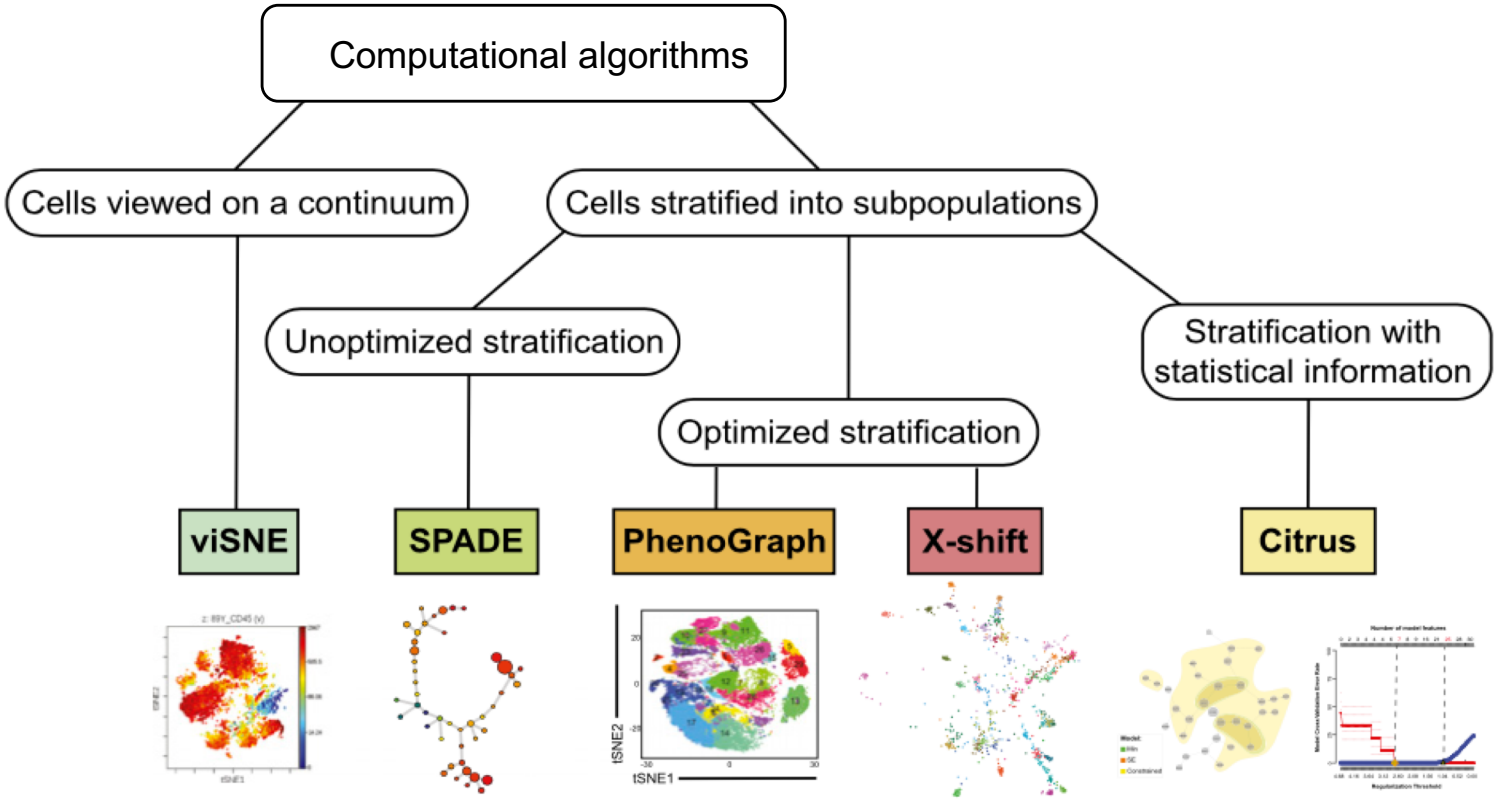
SPADE

Tree of relationships



Lisa Borghesi
Professor of Immunology
Director, Unified Flow Core

Choosing an Algorithm



Utility:

Rapid visual analysis, limited quantitative analysis

In depth visual analysis, rapid quantitative analysis

Rapid statistical analysis, limited visual analysis

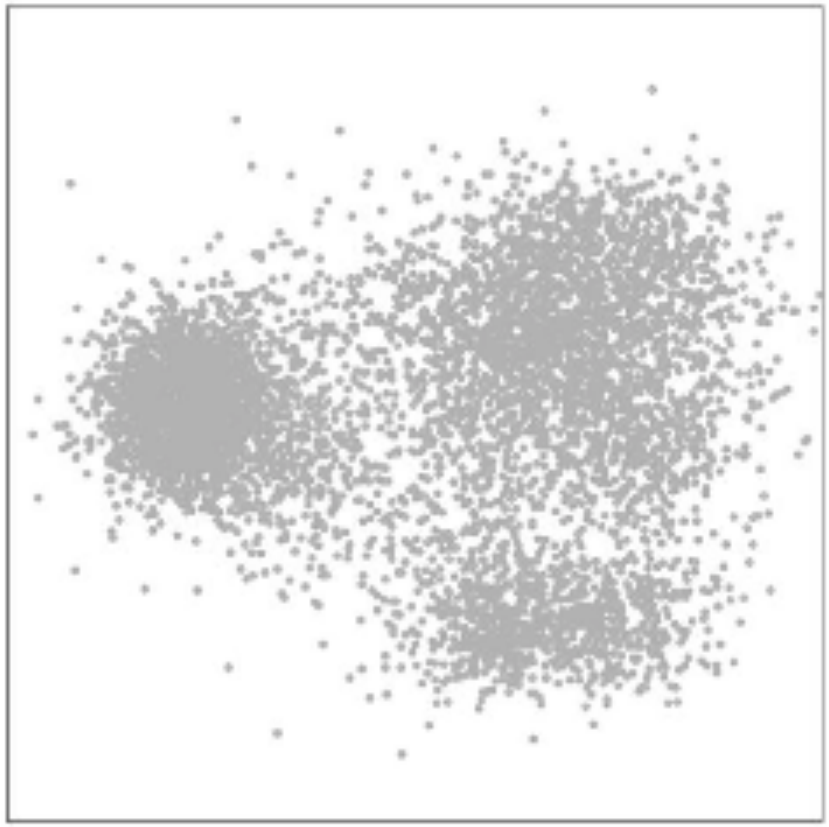
Data Output:

- Abundance
- Expression
- Cellular Identity
- Population Structure

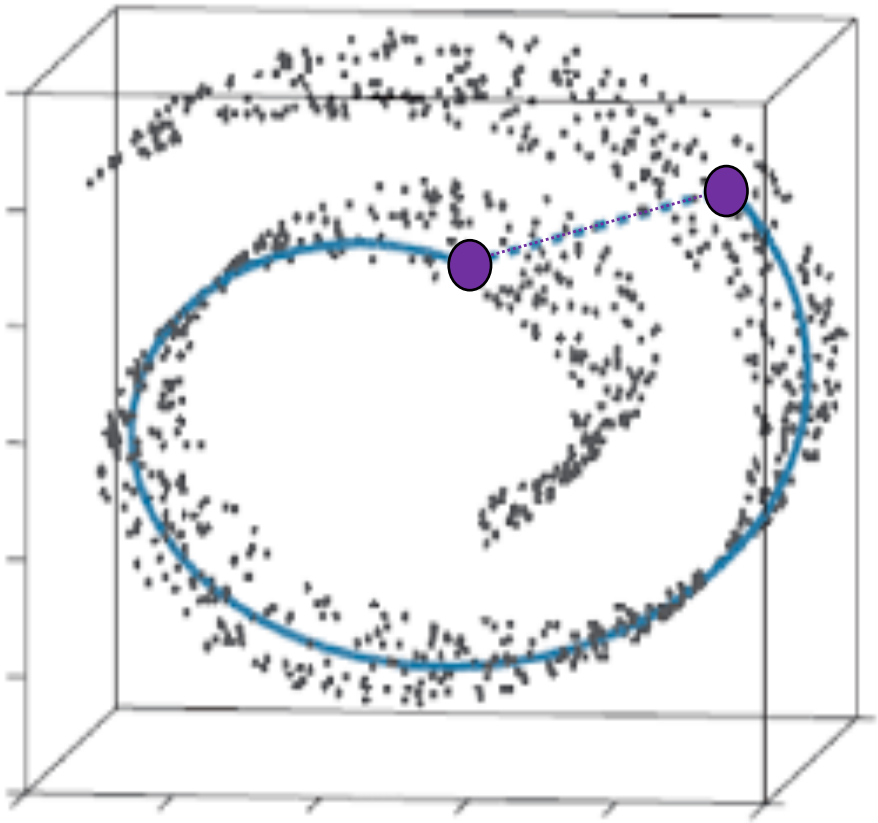
Kimball 2017 JI

Why not just use PCA?

PCA



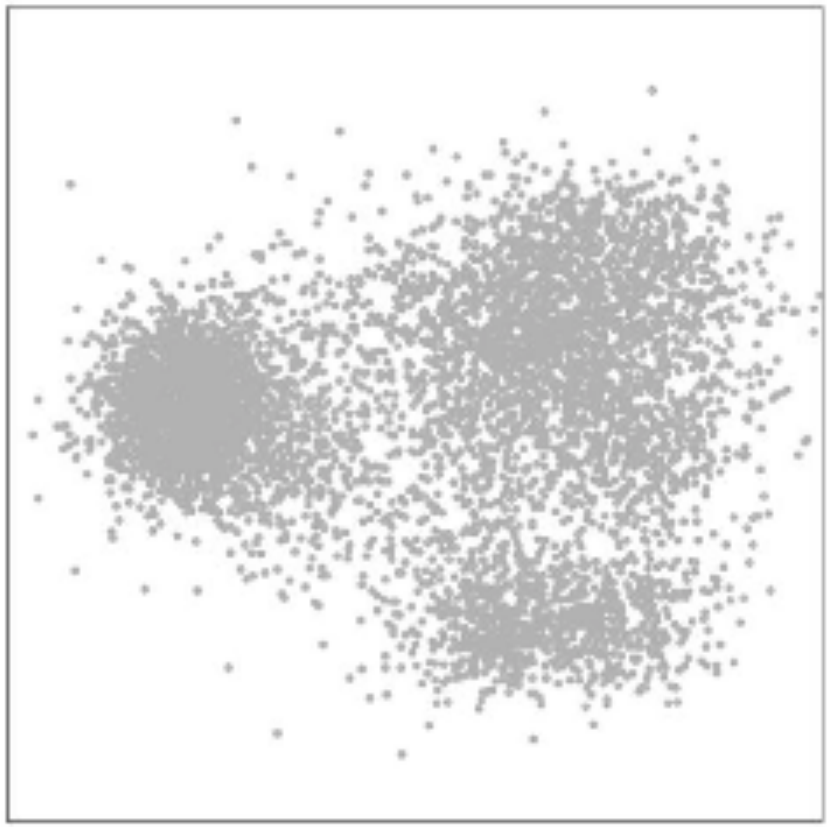
Swiss roll problem



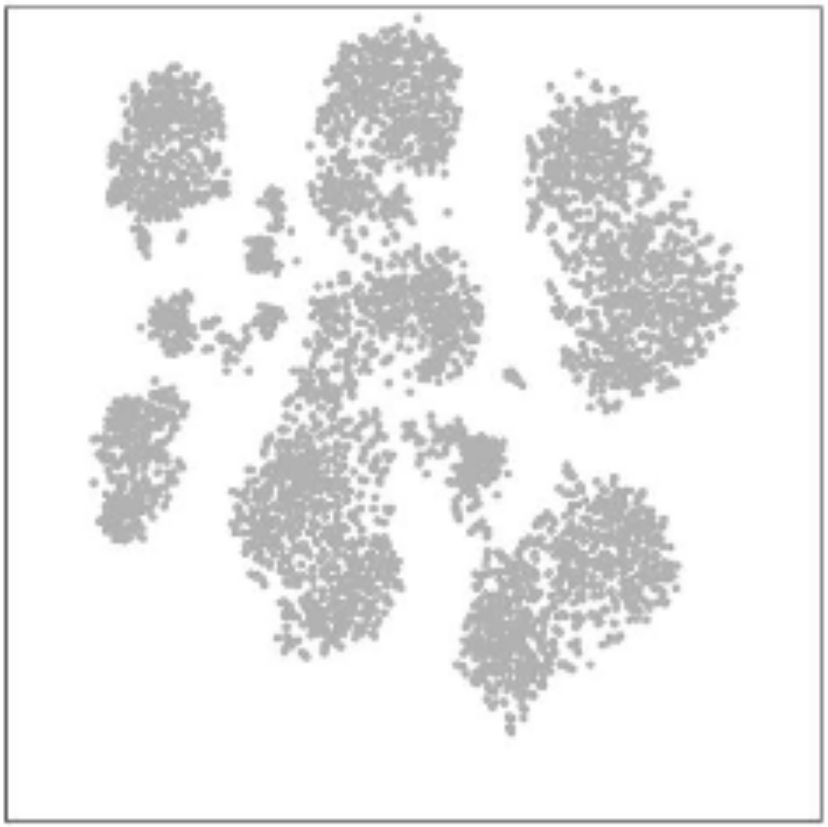
- Principal Components Analysis (PCA) preserves large pairwise distances
- Euclidean distance between two points on the Swiss roll does not accurately reflect local structure

t-SNE preserves local distances and global distances

PCA



tSNE

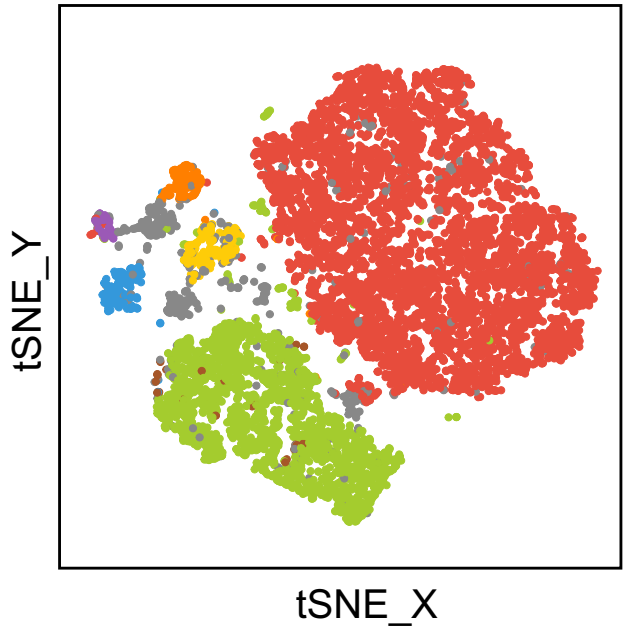


Amir 2013 Nature Biotech, Suppl.

t-SNE – dimensionality reduction algorithm

Goal: find a low dimensional visualization that best reflects population structure in high dimensional space

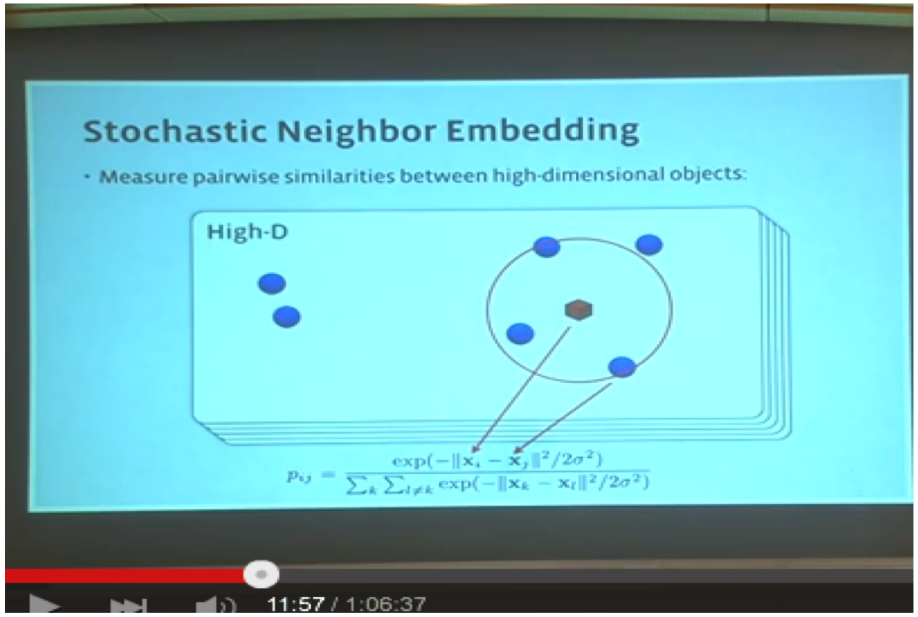
→ colloquially, get a feel for how objects are arranged in data space

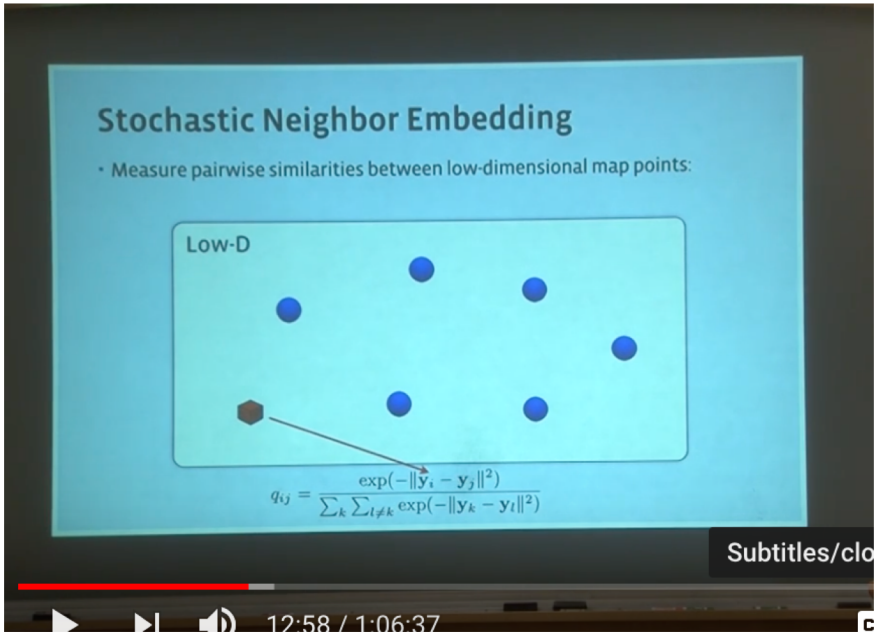


Laurens van der Maaten explains t-SNE (UCSD seminar) – fun and informative!!

https://www.youtube.com/watch?v=EMD106bB2vY

t-SNE operation



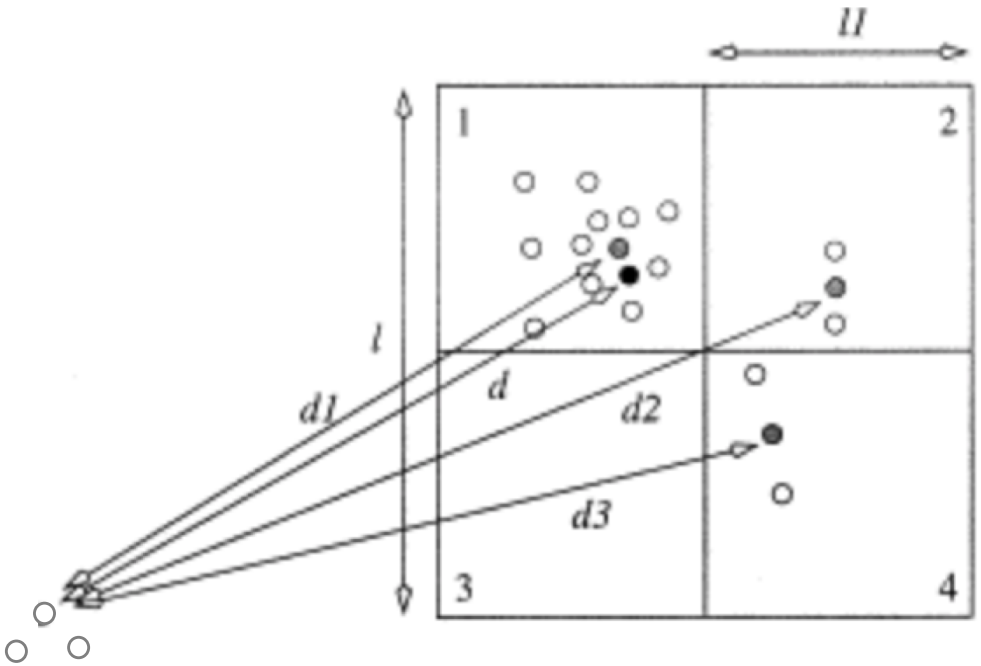


High-D data space. Draw Gaussian bell (circle) around data point. Measure density of all other points relative to that Gaussian bell, and establish probability distribution that represents their similarity. Computes local densities to get a distribution of pairs of points.
→ **Pij**

Low-D 2D map. Repeat above.
→ **Qij**

Mathematically minimize **P||Q** difference. Zero would be if two points were the same.

Barnes-Hut Modification of t-SNE (bh-SNE)



- Center of mass of domain
- Centers of mass of subdomains
- Source particle

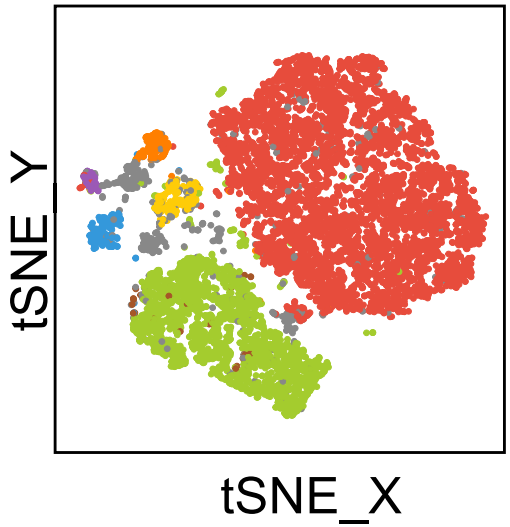
if (l/d < a)
compute direct force interaction
with the center of mass of domain.
else
if (ll/d1 < a)
compute direct force computation
with center of mass of subdomain 1
else
expand subdomain 1 further

Apply similar criteria to domains 2, 3, and 4

t-SNE

Advantages

- single cell information
- non-linear assumptions (as opposed to PCA)
- preserves local and global structure



Limitations

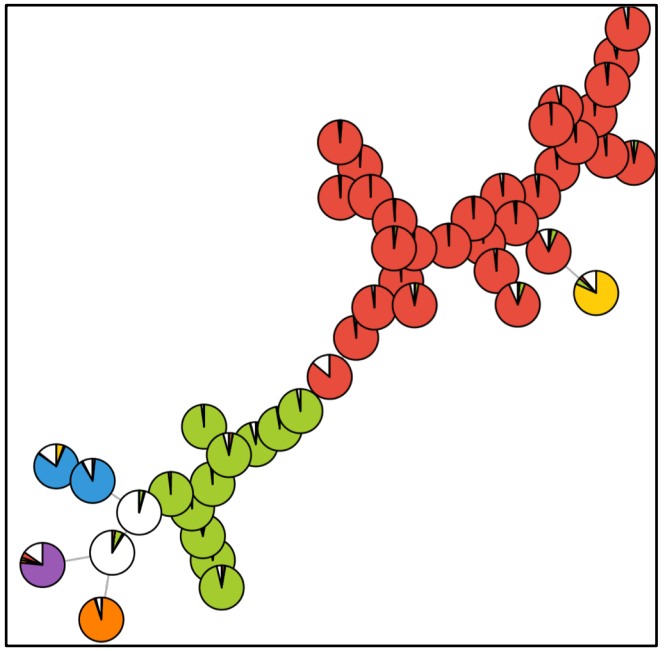
- computationally expensive; obligate downsampling means data are discarded
- plot axes are arbitrary and have no intrinsic meaning
- no population identification; follow up approaches required to assign identity to clusters and cells
- distance between clusters is not meaningful; no hierarchy

SPADE: Hierarchical clustering algorithm

spanning tree progression of density normalized events

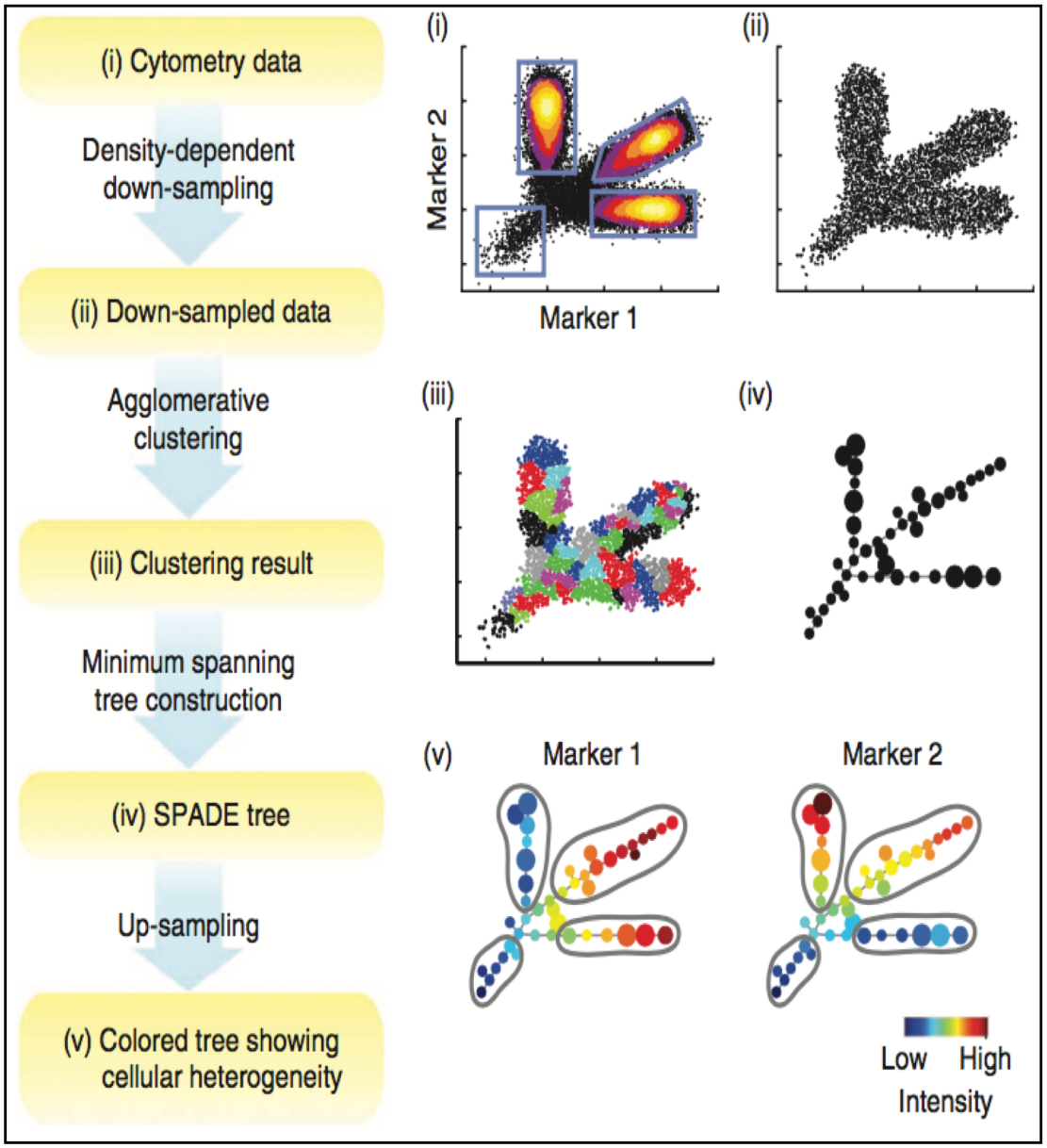
Goal: organize cells into a **hierarchy** using unsupervised approaches

→ colloquially, generate a tree of relationships



Output minimal spanning tree (MST) highlights the relationships between most closely related cell type clusters

SPADE



SPADE views **data as a cloud** of points (cells) where the dimensions = # markers

Density-dependent downsampling to equalize density in different parts of cloud, ensures rare cells not lost

Agglomerative clustering based on marker intensity

Connect clusters in minimal spanning tree that best reflects geometry of the original cloud

Upsampling, map each cell in the original data set to the clusters

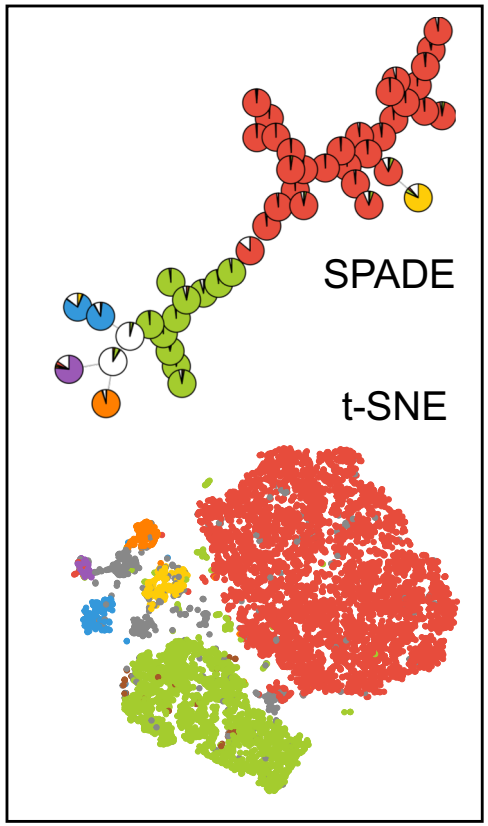
SPADE

Advantages

- rare pops preserved through density-dependent downsampling
- enables visualization of continuity of phenotypes
- can combine data sets that share common markers, and then co-map any markers unique to each data set (see orig. paper)

Limitations

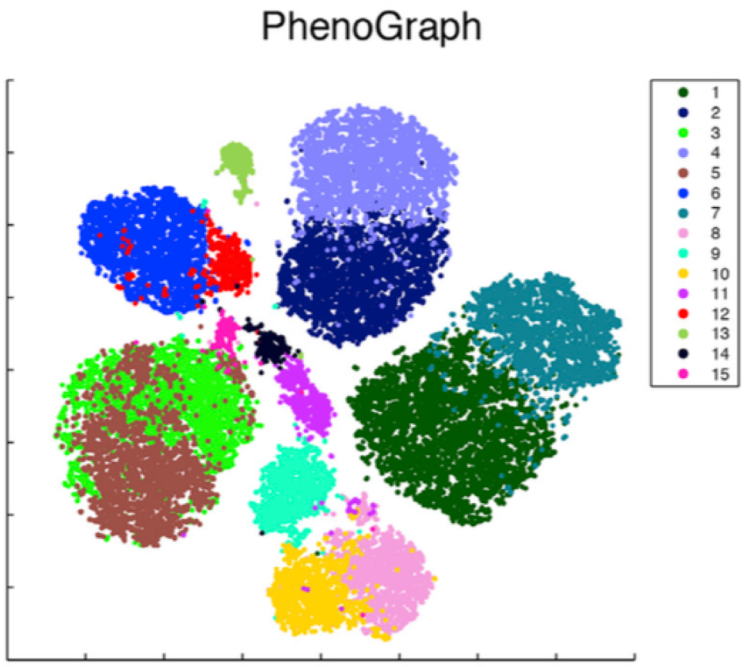
- loss of single cell information
- user chooses cluster number
- MSTs are non-cyclic and paths can be artificially split



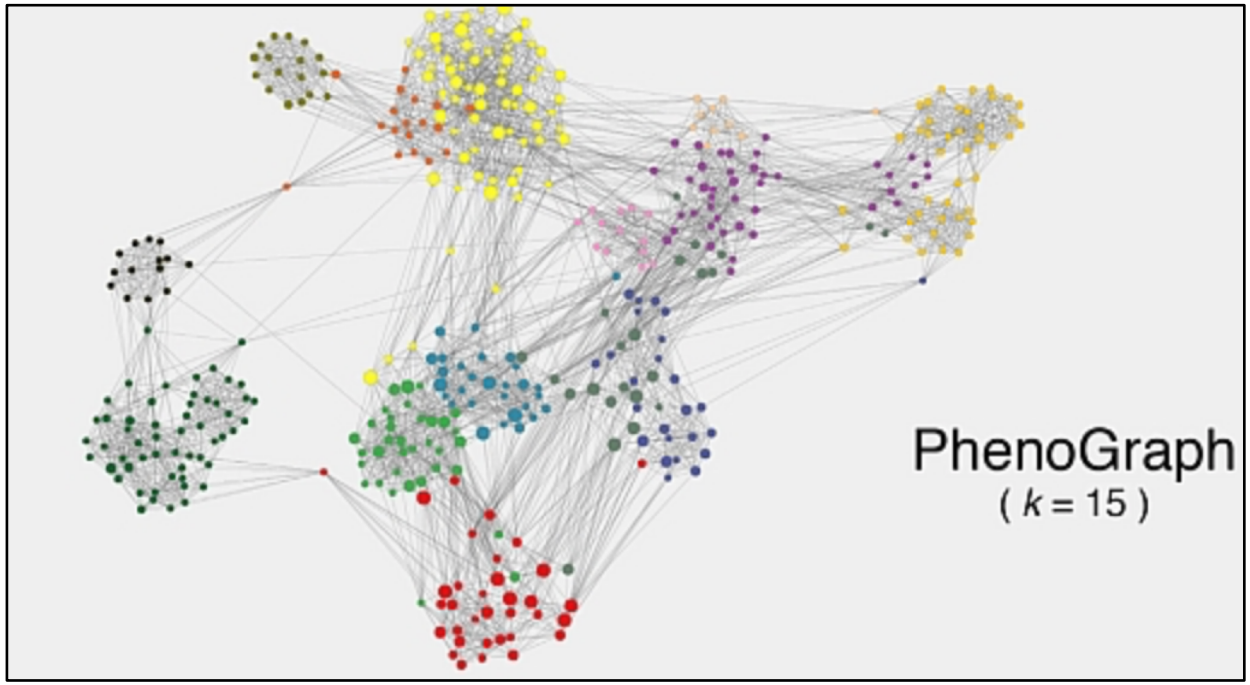
PhenoGraph

Goal: automated partitioning of high-dimensional single-cell data into subpopulations

→ colloquially, map nearest neighbors



PhenoGraph



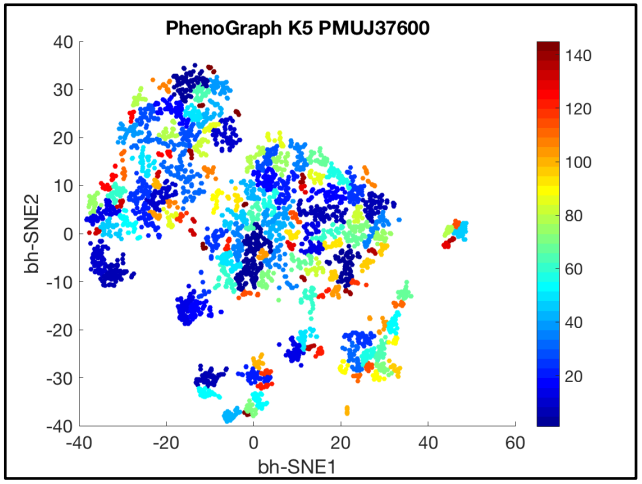
First order relationship – find the *k* nearest neighbors for each cell using Euclidean distance

Second order relationships – cells with shared neighbors should be placed near one another

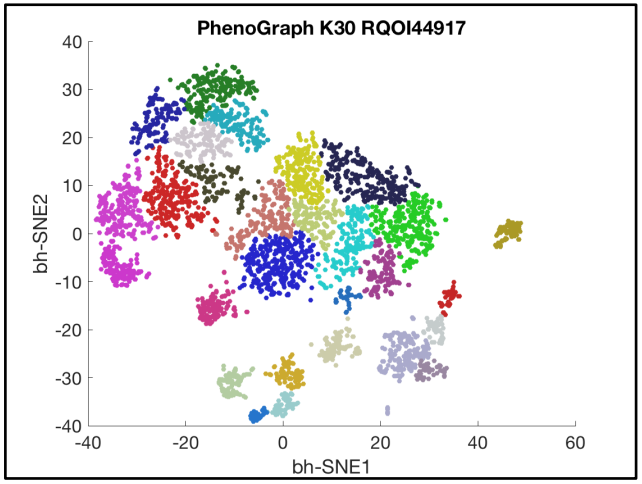
Third, identify communities – Louvain method that measures the density of edges inside communities to edges outside communities

PhenoGraph: Number of neighbors

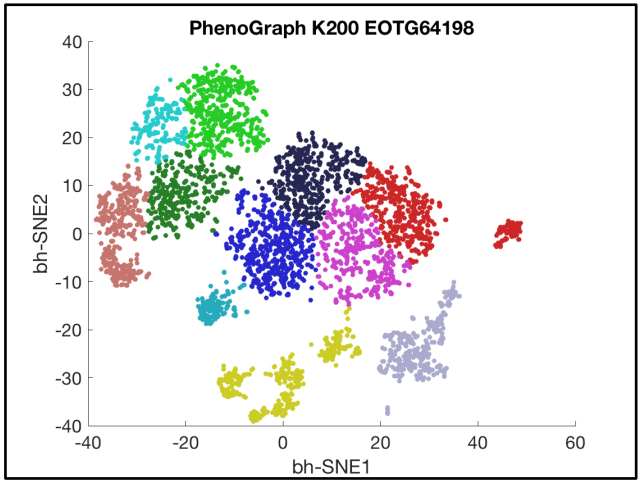
Neighbors = 5



Neighbors = 30

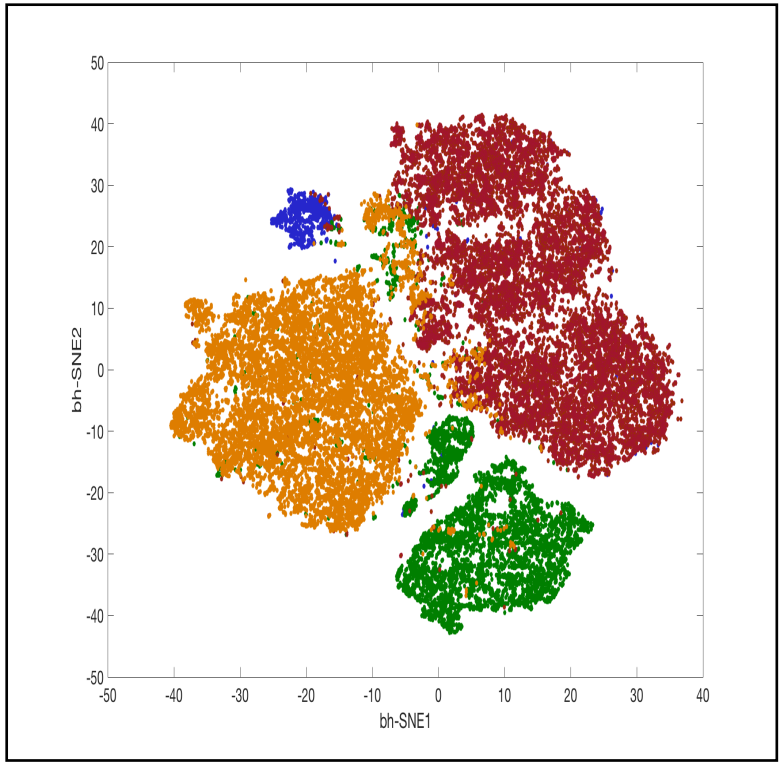


Neighbors = 200



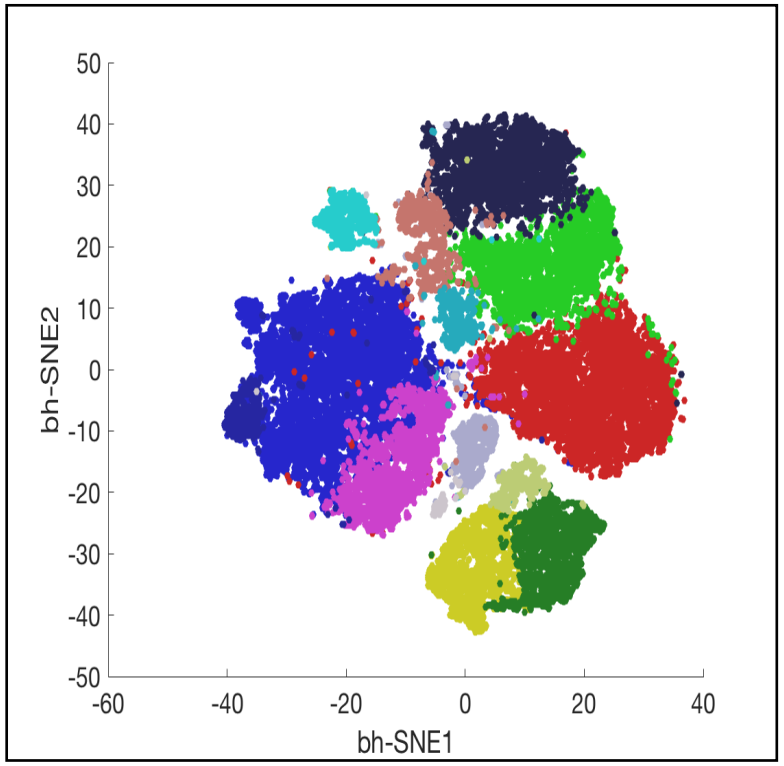
PhenoGraph – population discovery

Manual gate overlays



- Naive B
- ASC
- MBC (total)
- Ag-exper.

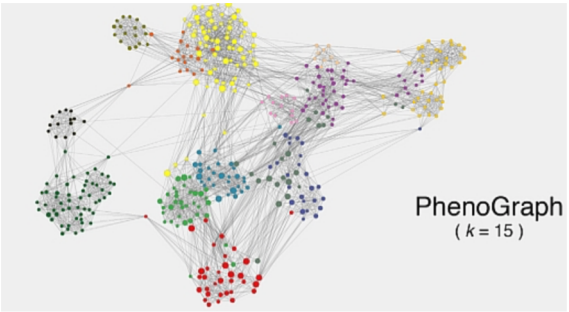
PhenoGraph



PhenoGraph

Advantages

- opportunity for population discovery
- can resolve subpopulations as rare as 1 in 2000 cells
- robust to cluster shape (e.g., need not be spherical)



Limitations

- user specifies number of neighbors
- ideal cluster number, or biologically relevant cluster number, is largely unconstrained

Next Presentation: Install the Algorithms on your Personal Computer

Resources

Useful starting places - reviews

1. Kimball AK, Oko LM, Bullock BL, Nemenoff RA, van Dyk LF, Clambey ET. A Beginner's Guide to Analyzing and Visualizing Mass Cytometry Data. J Immunol. 2018 Jan 1;200(1):3-22.
2. Saeys Y, Gassen SV, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016 Jul;16(7):449-62.
3. Mair F, Hartmann FJ, Mrdjen D, Tosevski V, Krieg C, Becher B. The end of gating? An introduction to automated analysis of high dimensional cytometry data. Eur J Immunol. 2016 Jan;46(1):34-43.
4. Chester C & Maecker HT. J Immunol. 2015 Aug 1;195(3):773-9. doi: 10.4049/jimmunol.1500633. Algorithmic Tools for Mining High-Dimensional Cytometry Data. J Immunol. 2015 Aug 1;195(3):773-9.

Original application of algorithms

1. Qiu P, Simonds EF, Bendall SC, Gibbs KD Jr, Bruggner RV, Linderman MD, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011 Oct 2;29(10):886-91. **SPADE**
2. Amir el-AD, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Amir el-AD, Davis KL, 3. Tadmor MD, Simonds EF, Levine JH, Bendall SC, Shenfeld DK, Krishnaswamy S, Nolan GP, Pe'er D. Nat Biotechnol. 2013 Jun;31(6):545-52. **viSNE**
3. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015 Jul 2;162(1):184-97. **PhenoGraph**
4. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci U S A. 2014 Jul 1;111(26):E2770-7. **CITRUS**
5. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. Wishbone identifies bifurcating developmental trajectories from single-cell data. Nat Biotechnol. 2016 Jun;34(6):637-45. **Wishbone**
6. Bendall SC, Davis KL, Amir el-AD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014 Apr 24;157(3):714-25. **Wanderlust**
7. McInnes L, Healy J, Melvile J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. https://arxiv.org/abs/1802.03426 **UMAP (for advanced Matlab Users, no GUI interface)**