3000788 Intro to Comp Molec Biol

Lecture 14: Single-cell data analysis

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today’s agenda

- Recap: QC and batch effect removal
- Visualization of single-cell data
- Cell type inference
- Trajectory (pseudotime) reconstruction

Recap: Key steps in single-cell data processing

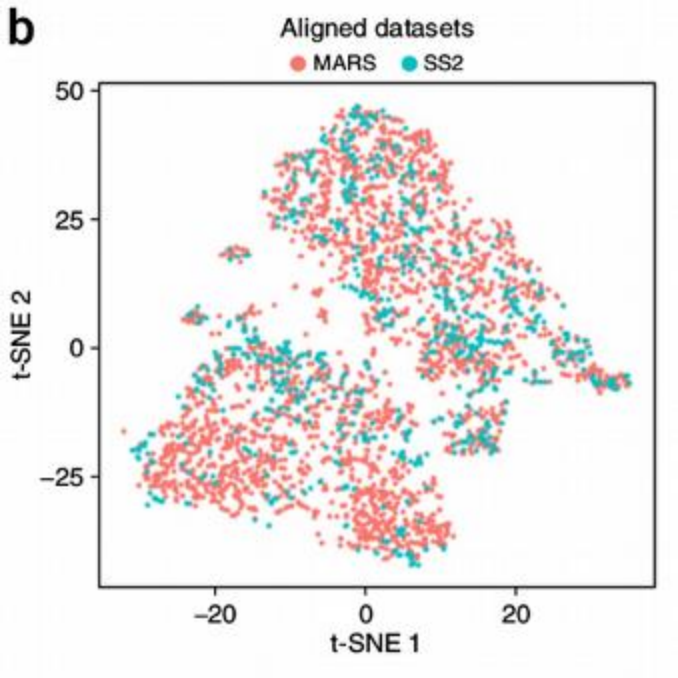
- Quality filter
  - Low read count & gene count = non-cells
  - Very high read count & gene count = multi-cells (doublets)
  - High mitochondrial expression ~ dead cells (or special cell types)

- Normalization and impute missing values

- Multi-sample integration
  - Some cell types/genes are detected in some batches
  - Some genes are affected by conditions → affect visualization and clustering

Visualization of single-cell data

How were these plots generated?



- Each cell is described by >5,000 genes
- But visualized on the screen, each cell is represented by just (*x*, *y*)-coordinates
- How to reduce >5,000 dimension to 2 while still retaining key information about the data?

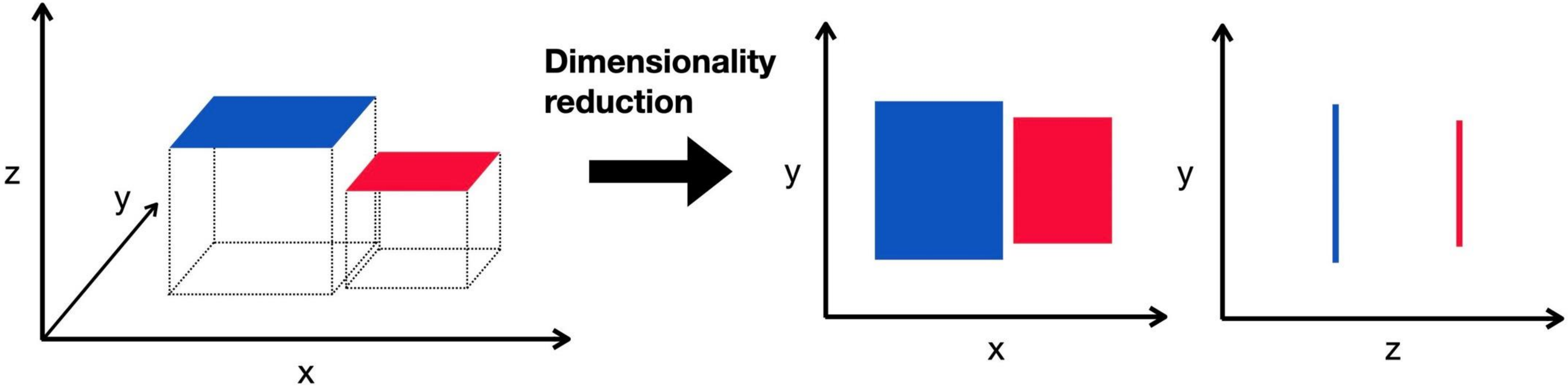
https://towardsdatascience.com/how-to-batch-correct-single-cell-7bad210c7ae1

What is the dimension of a transcriptomics dataset?

	Gene 1	Gene 2	Gene 3	Gene 4	…
Sample 1					
…					

- Number of genes?
- Number of non-redundant genes (aggregated into pathways)?
- The minimal number of values from which the entire transcriptomics profile can be accurately recreated?

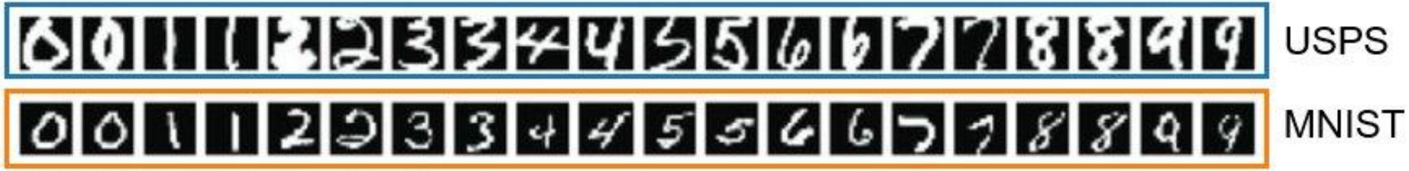
Dimensionality reduction



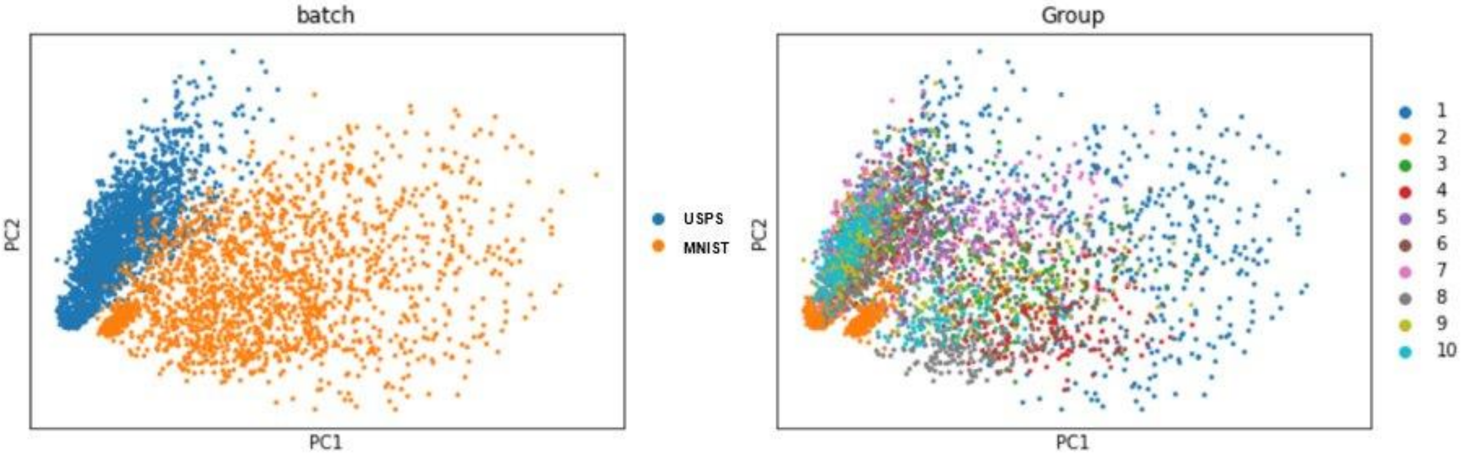
https://www.sc-best-practices.org/preprocessing_visualization/dimensionality_reduction.html

- Reduce dimension (number of features) while maintaining information
  - We measured more genes than needed
  - To distinguish cell types, a few gene combination may be enough

A digit dataset example

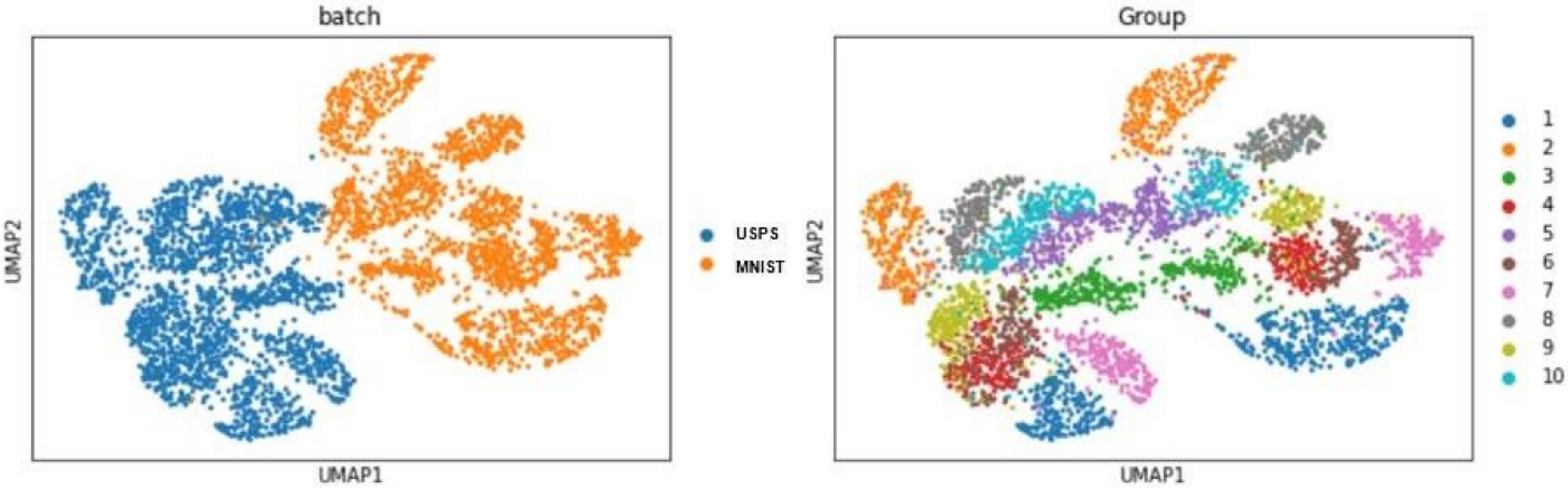


*adapted from doi:10.1109/TKDE.2017.2669193



https://twitter.com/lkmklsmn/status/1436357177887895555

A digit dataset example

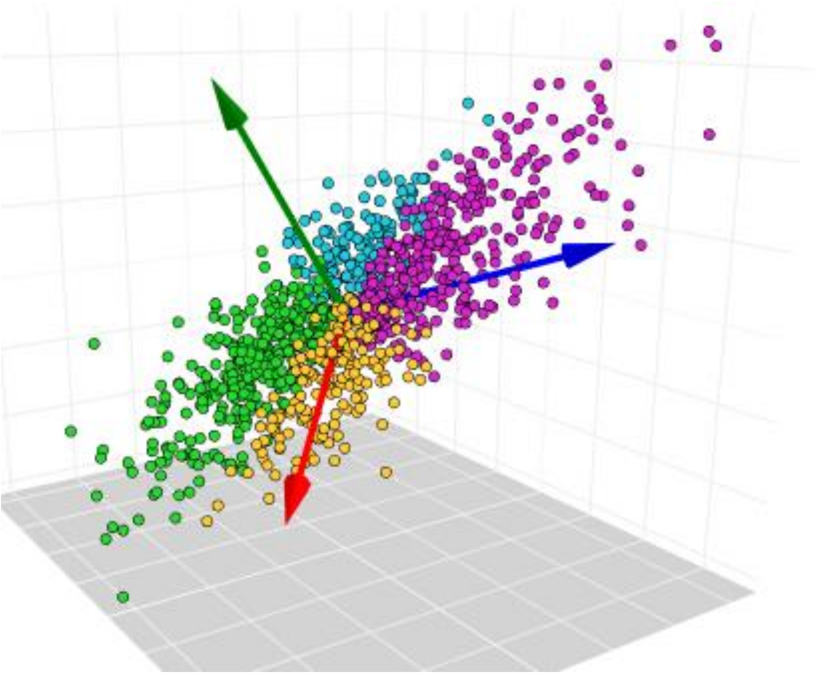


https://twitter.com/lkmklsmn/status/1436357177887895555

- Both data source and digit identity can be distinguished

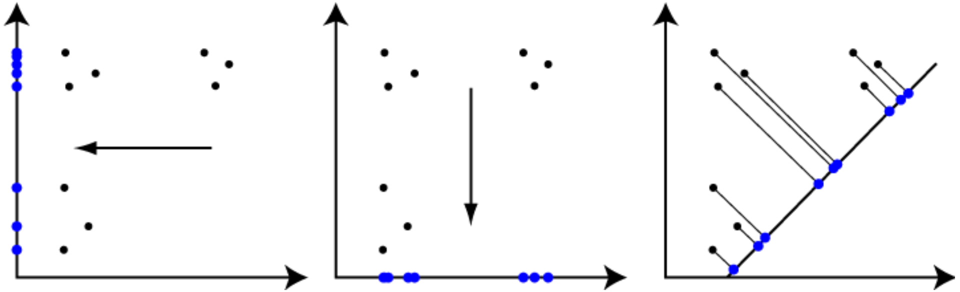
Principal component analysis (PCA)

Variance is information



https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d

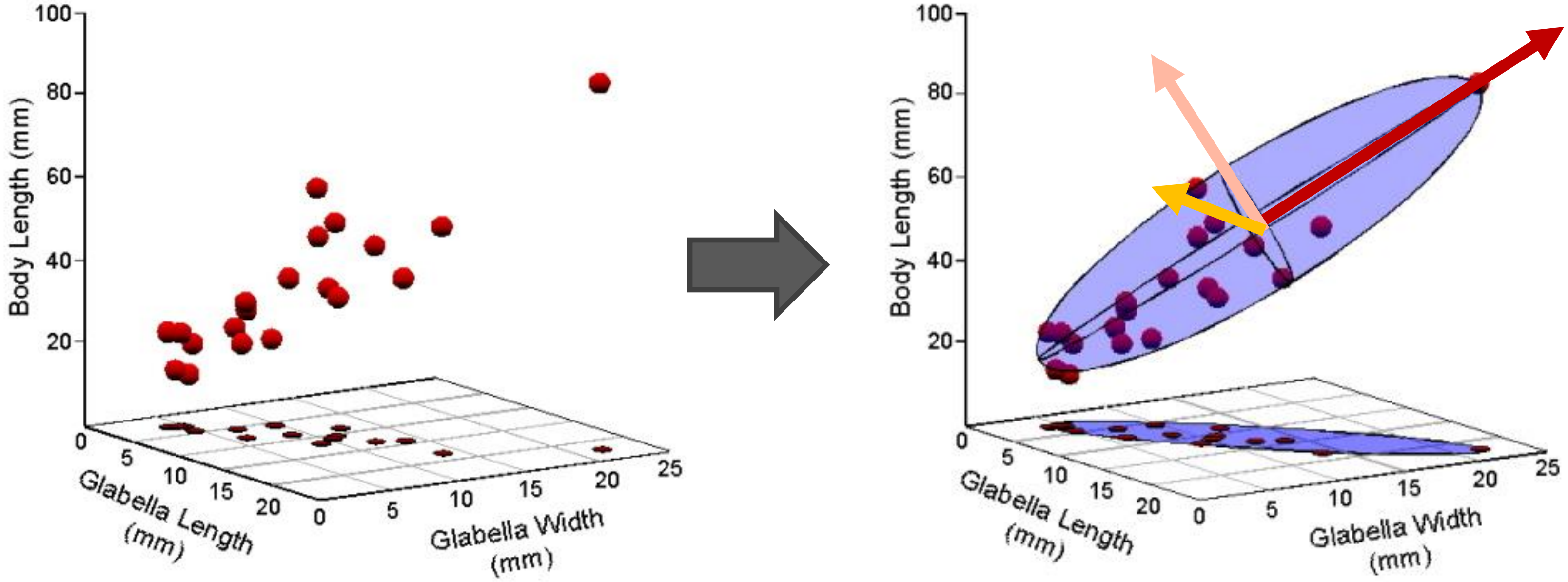
Projection



https://shapeofdata.wordpress.com/2013/04/16/visualization-and-projection/

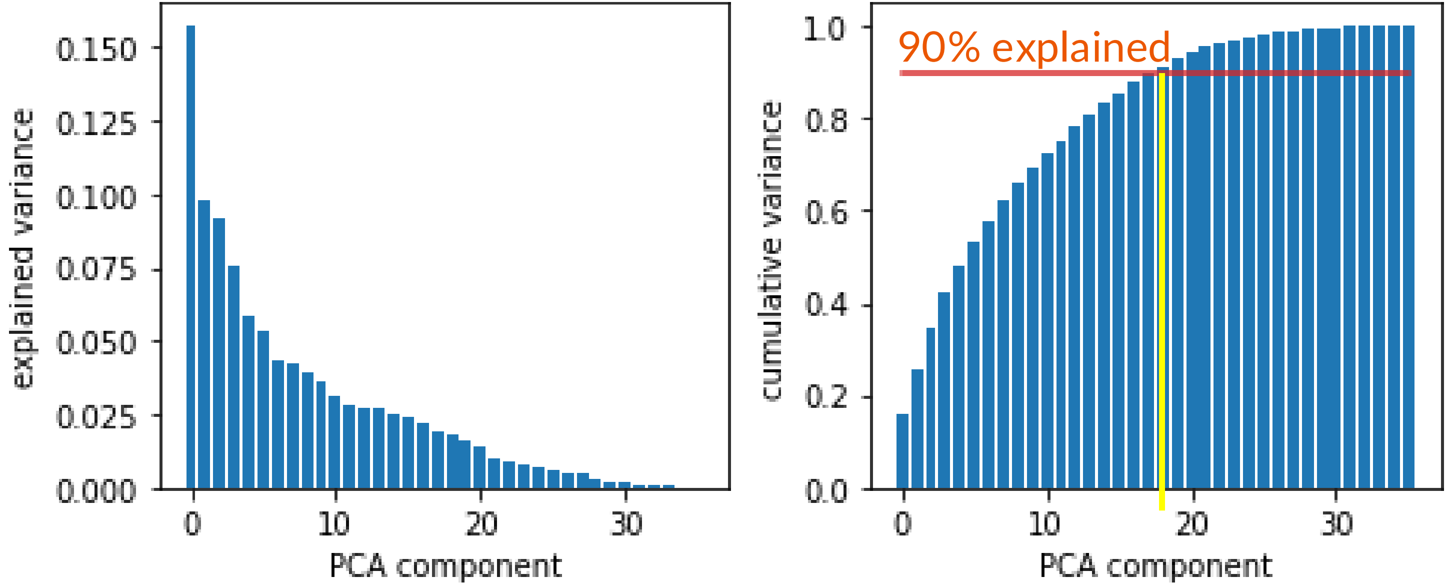
- High variances = more power to distinguish groups of data points

PCA follows the directions of maximal variance



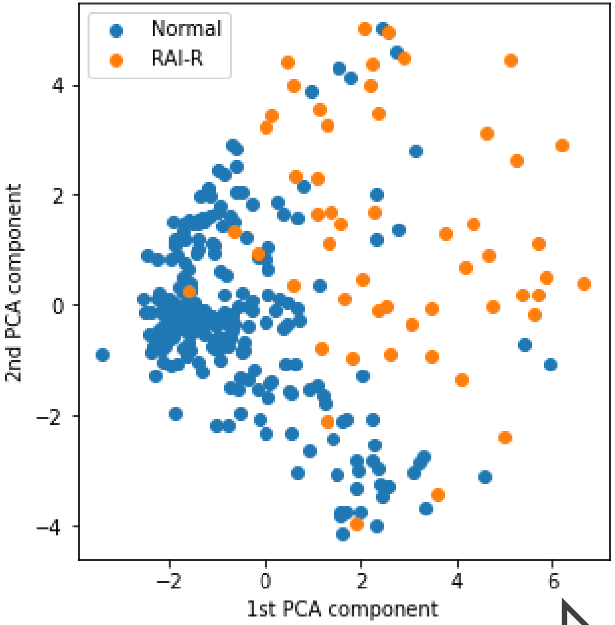
Source: the paleontological association

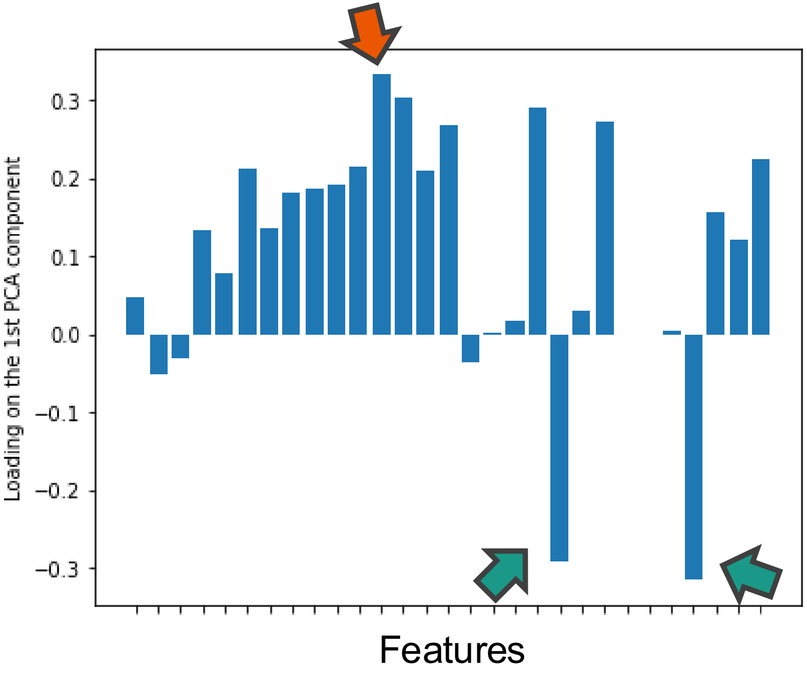
PCA for dimensionality reduction



- By default, PCA does not reduce the number of dimensions
- We can select only the first *k* PC for downstream analyses

Interpretation of PCA result



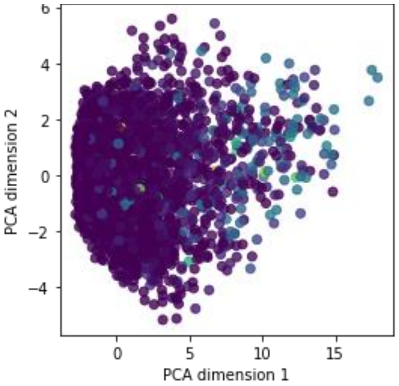


Resistance to treatment

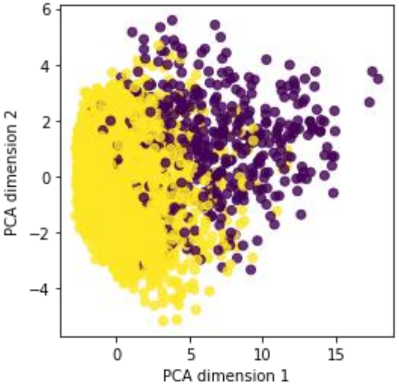
Each principal component is associated with a linear combination of input variables

Exploring PCA results

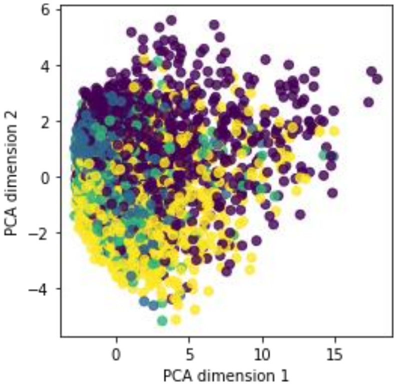
Feature 1



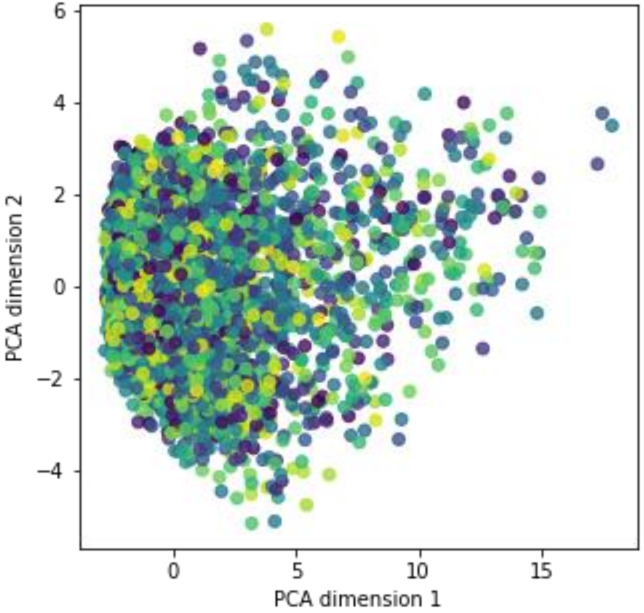
Feature 2



Feature 3



Batch ID

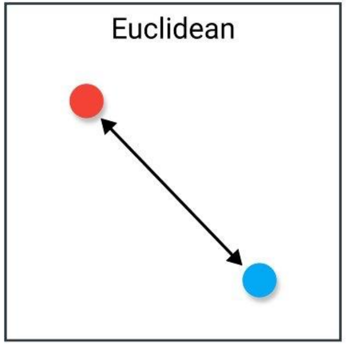


- Color by feature values to understand how PCA group data points
- Color by potential confounding factors

Pros and cons of PCA

- Each PC can be interpreted from the loadings
- Highly correlated features tend to be grouped into the same PC
- PCA is a good initial dimensionality reduction step

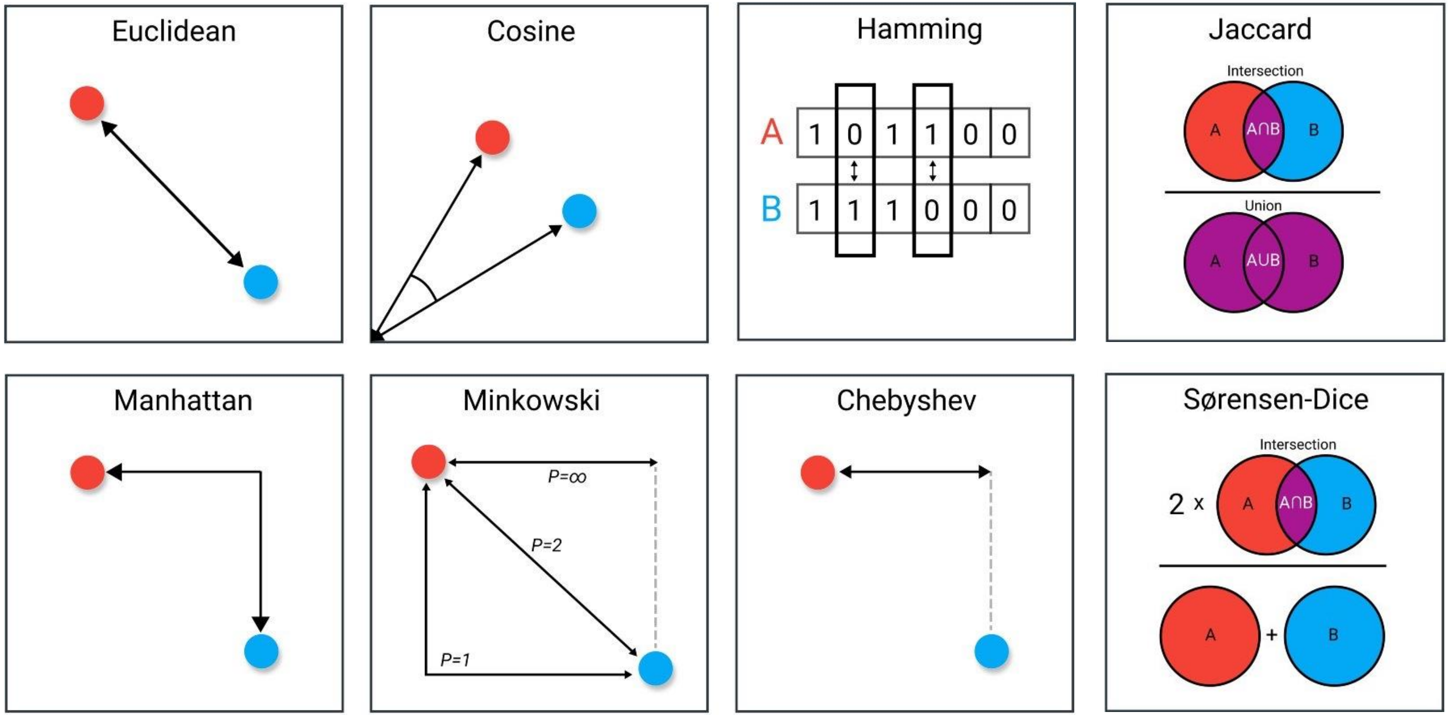
- PCA strictly preserves Euclidean distance
  - But some datasets require different distance metric!



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

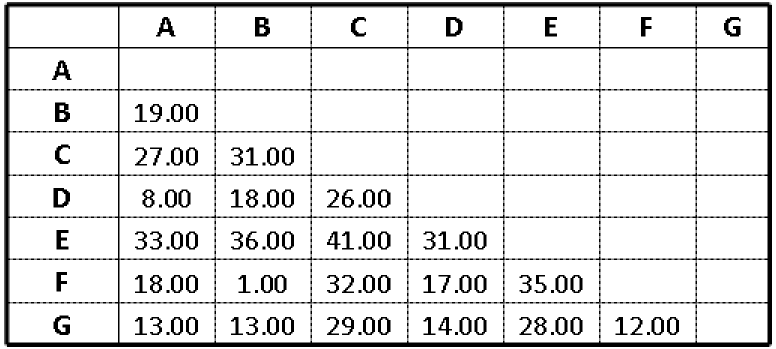
Multidimensional Scaling (MDS)
Principal Coordinate Analysis (PCoA)

Different distances for different data types



https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

Pairwise distance matrix



	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	1.00	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

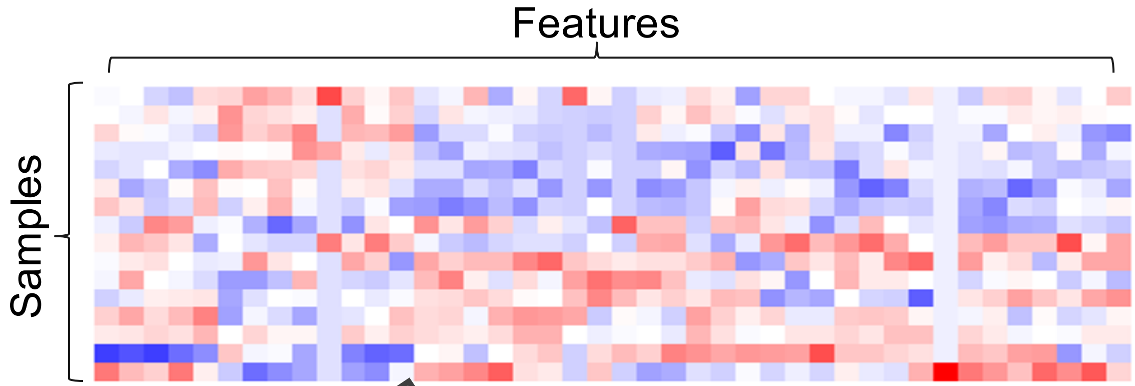
http://www.slimsuite.unsw.edu.au/teaching/upgma/

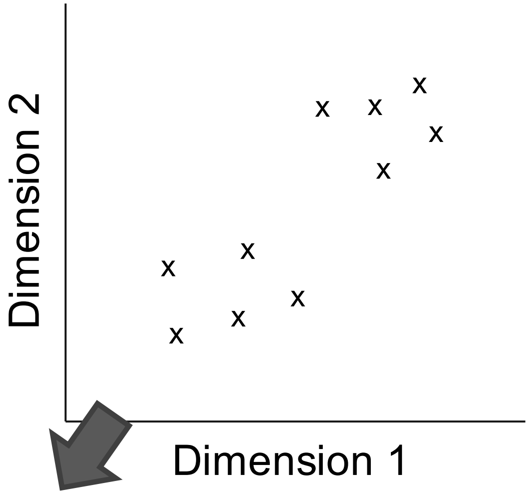
- For each pair of samples, calculate the distance between them

- Some samples are similar to each other, some are not

- When we reduce the dimension or visualize the data on the plot, we want **similar samples to remain closer to each other than to dissimilar samples**

MDS / PCoA general principles





User-defined distance

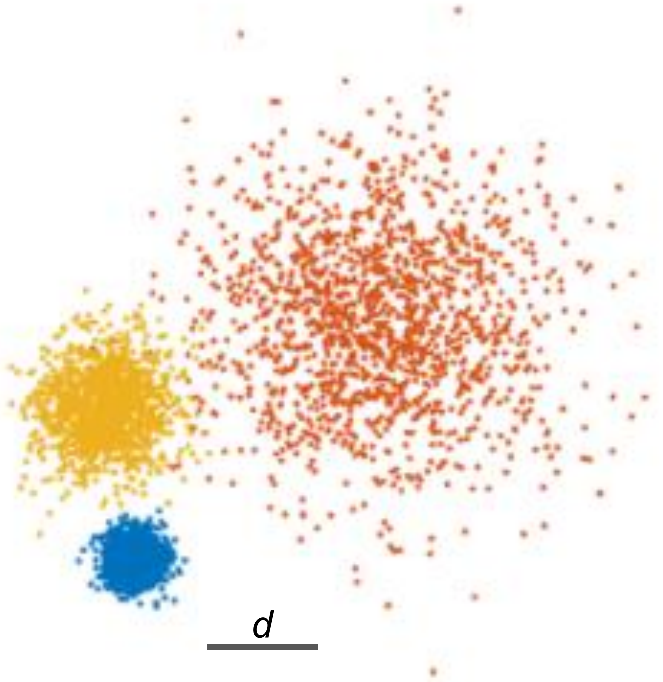
	x_1	x_2	x_3	x_4
x_1	0	3	3	3
x_2	3	0	1	1
x_3	3	1	0	1
x_4	3	1	1	0

Correlation

	x_1	x_2	x_3	x_4
x_1	0	9	9	9
x_2	9	0	1	1
x_3	9	1	0	1
x_4	9	1	1	0

Euclidean distance

Limitation of MDS / PCoA



- A single definition of distance is used throughout the dataset

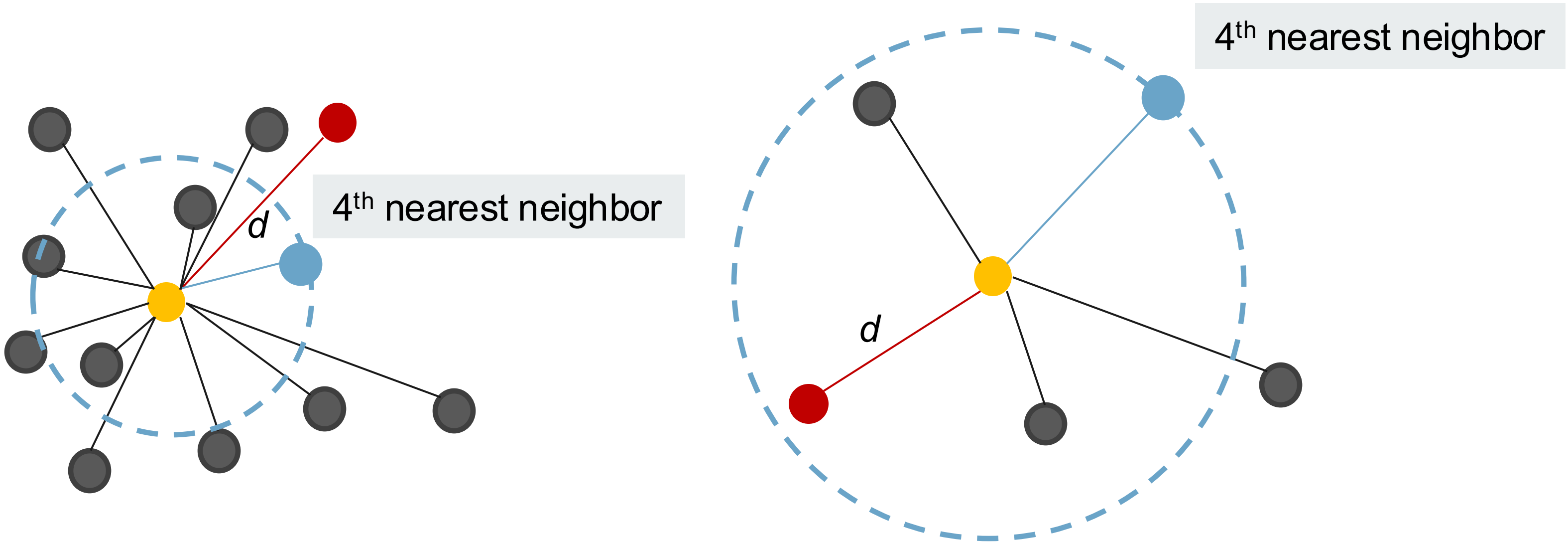
- What if some data groups are noisier or have higher variances than the others?

- Distance *d* can mean either similar or dissimilar depending on cell types

- **Can we account for the density of the data?**

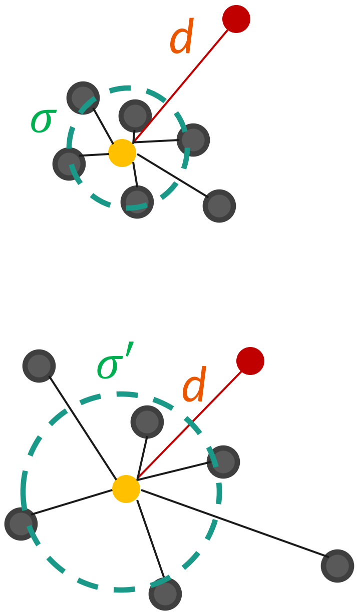
t-distributed Stochastic Neighbor Embedding (*t*-SNE)
Uniform Manifold Approximation and Projection (UMAP)

How to measure data density?



- Naïve approach: Count number of data points within a distance
- Measure distance to the *k*-th nearest neighbor
- Different interpretations of distance *d*

Probability of being a neighbor



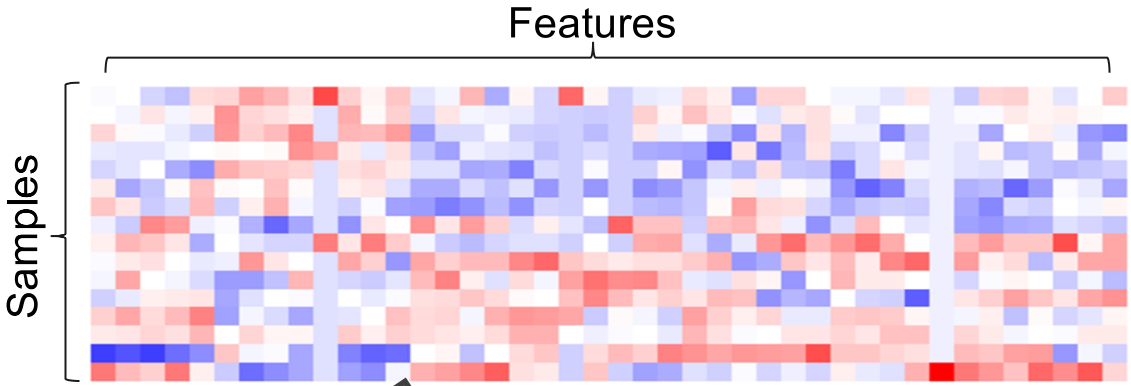
score(o | o) = probability that o would pick o as neighbor under a **normal distribution** center at o

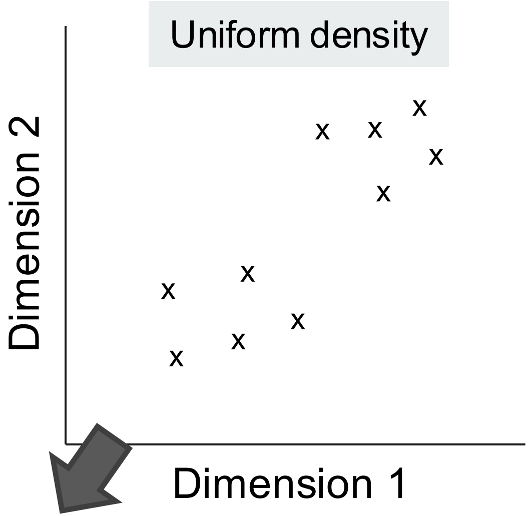
$$= \frac{e^{-\frac{d^2}{2\sigma^2}}/\sigma}{\sum e^{-\frac{(\text{dist}(o,o))^2}{2\sigma^2}}/\sigma}$$

o = other data points

- Distance *d* is normalized against density σ and distances from o to other nearby data points

t-SNE's general framework





Probability of being neighbor (various σ)

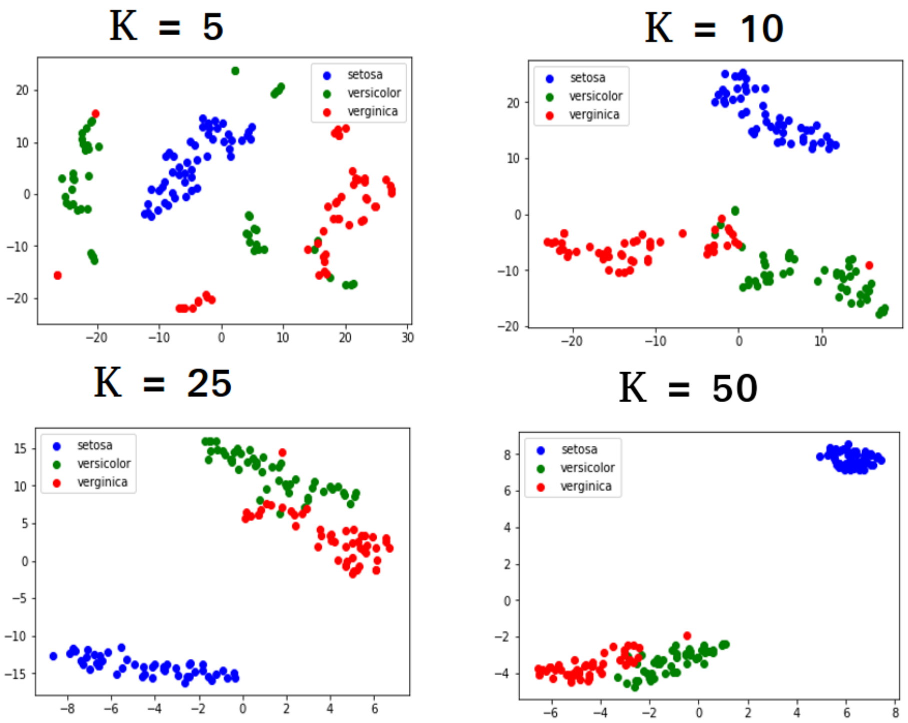
$$\begin{array}{c|cccc} & x_1 & x_2 & x_3 & x_4 \\ \hline x_1 & 0 & 3 & 3 & 3 \\ x_2 & 3 & 0 & 1 & 1 \\ x_3 & 3 & 1 & 0 & 1 \\ x_4 & 3 & 1 & 1 & 0 \end{array}$$

Divergence

$$\begin{array}{c|cccc} & x_1 & x_2 & x_3 & x_4 \\ \hline x_1 & 0 & 9 & 9 & 9 \\ x_2 & 9 & 0 & 1 & 1 \\ x_3 & 9 & 1 & 0 & 1 \\ x_4 & 9 & 1 & 1 & 0 \end{array}$$

Probability of being neighbor (σ = 1)

Perplexity: Which *k*th nearest neighbor to consider?



K = 5

K = 10

K = 25

K = 50

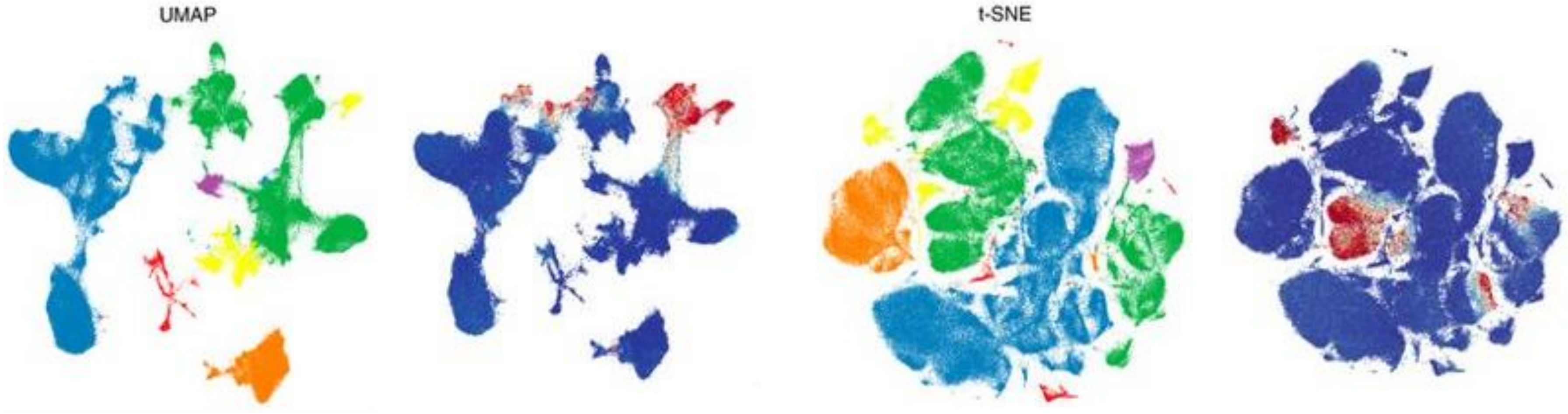
Source: blog.paperspace.com/dimension-reduction-with-t-sne/

- Too small perplexity = poor estimate of density, resulting in a lot of scatted data clusters

- Optimal perplexity varies across datasets

- **Advice**: Vary the perplexity and identify **patterns that appear consistently**

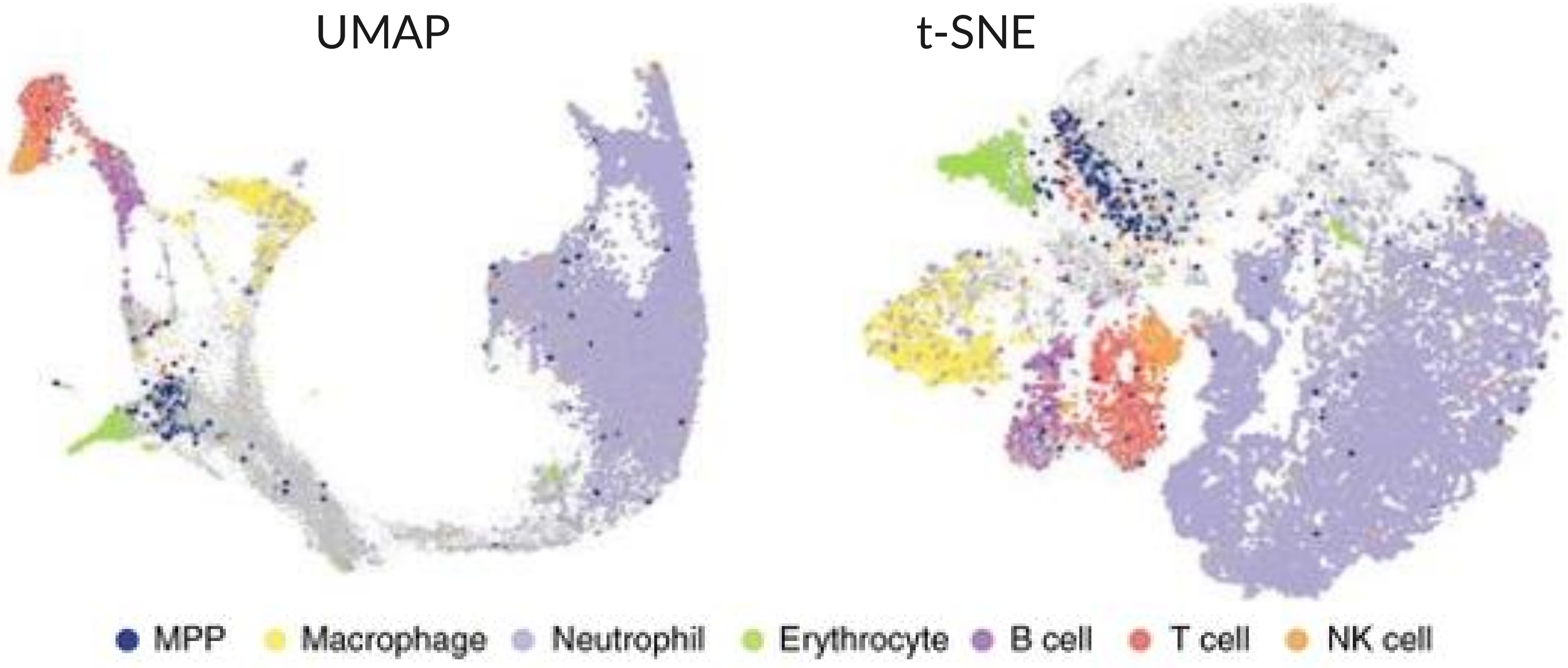
t-SNE vs UMAP on single-cell data



Becht, E. et al. Nature Biotechnology 37:38-44 (2019)

- Both are good for visualization
- *t*-SNE focuses more on clustering the cells
- UMAP also displays transitions across cell clusters

t-SNE vs UMAP on single-cell data



UMAP

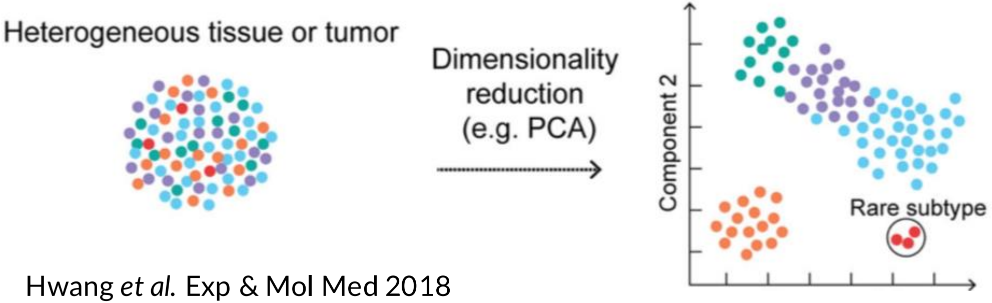
t-SNE

Cell-cell transition is easier to infer from UMAP

Becht, E. et al. Nature Biotechnology 37:38-44 (2019)

Analysis of single-cell data

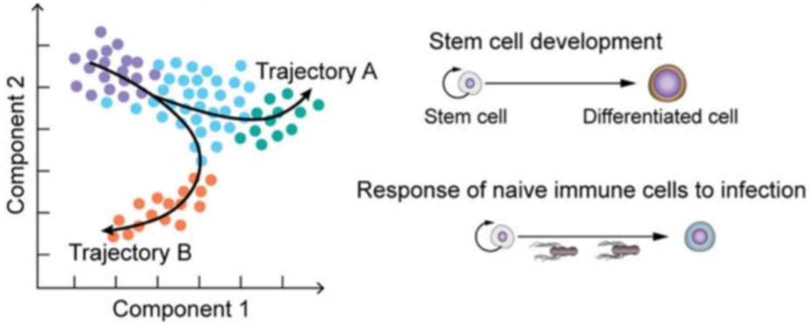
Cell type clustering and trajectory reconstruction


Heterogeneous tissue or tumor
Dimensionality reduction (e.g. PCA)
Component 2
Rare subtype

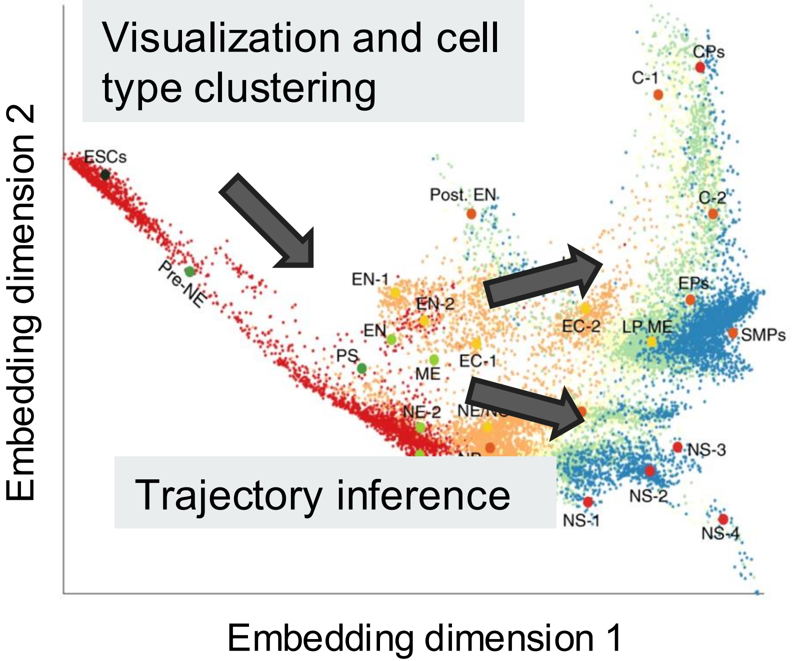
Hwang *et al.* Exp & Mol Med 2018

Clustering of cells with similar omics signatures reveals distinct cell types

Trajectory reconstruction (pseudotime) reveals the developmental pathways across cell types

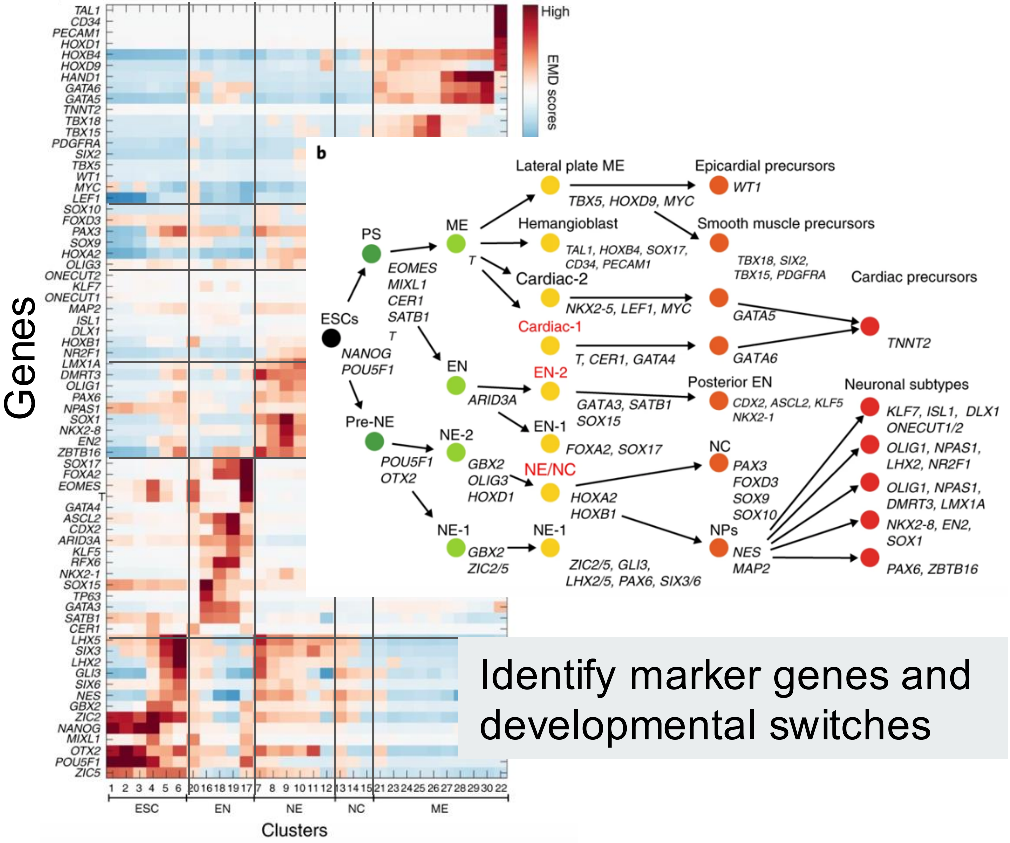


An end-to-end example



Visualization and cell type clustering

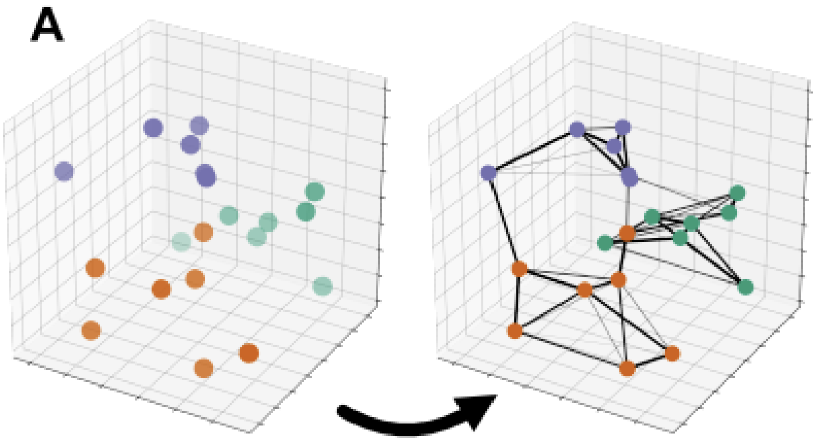
Trajectory inference



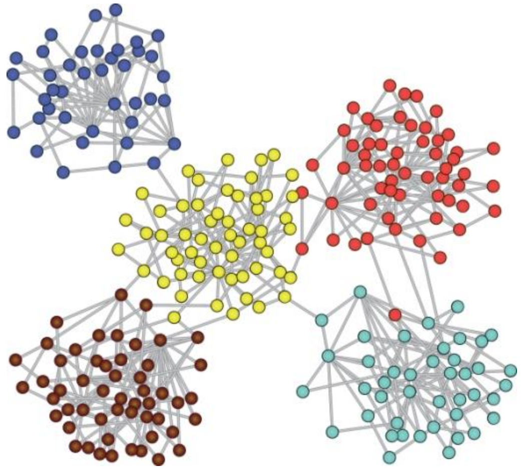
Identify marker genes and developmental switches

Source: Moon et al. Nature Biotechnology 37:1482-92 (2019)

Network clustering



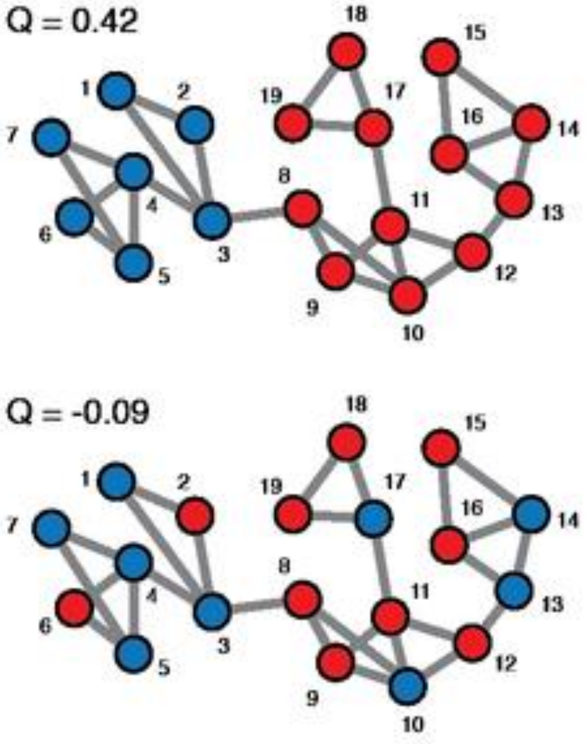
Sainburg, T. *et al.*, Neural Comput 33(11):2881-2907 (2021)



https://github.com/topics/graph-clustering

- **View cell-cell distance as a network**
- Apply some threshold on the distance to create sparse network
- Split network into modules with dense edges

Network clustering with modularity score



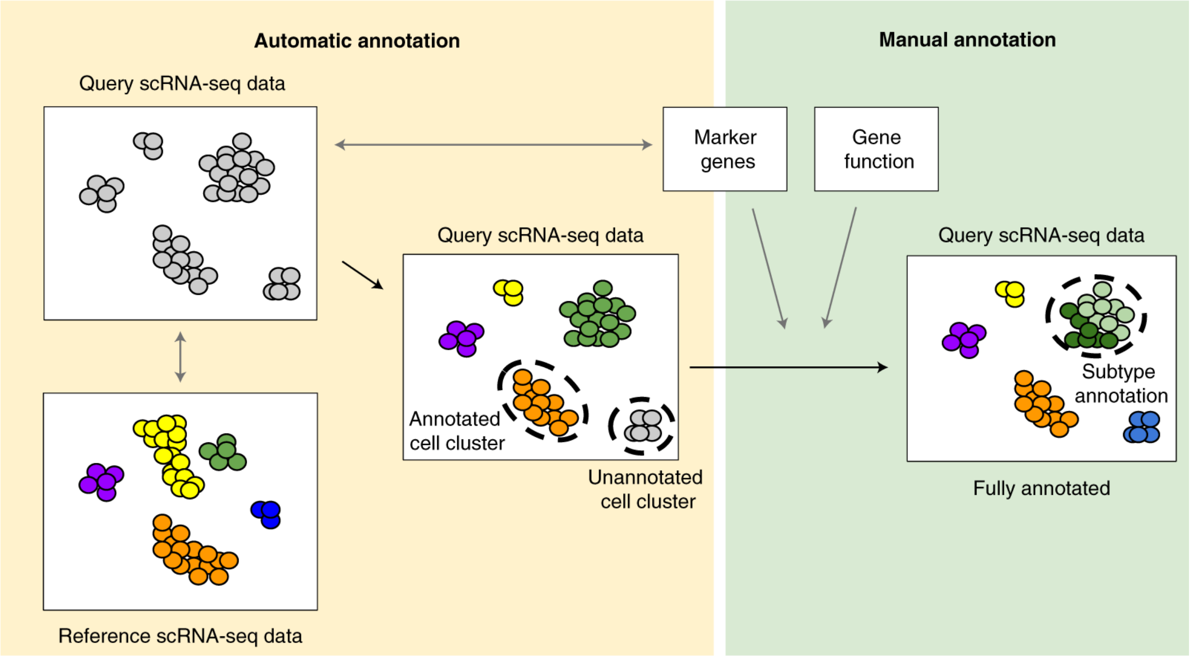
Betzel *et al*. (2015)

- Good clustering characteristics
  - Cells within the same cluster are highly connected
  - Cells across cluster are not connected

- **Modularity score** = normalized ratio of within-cluster edges versus between-cluster edges

- Louvain / Leiden algorithms

Cell type annotation

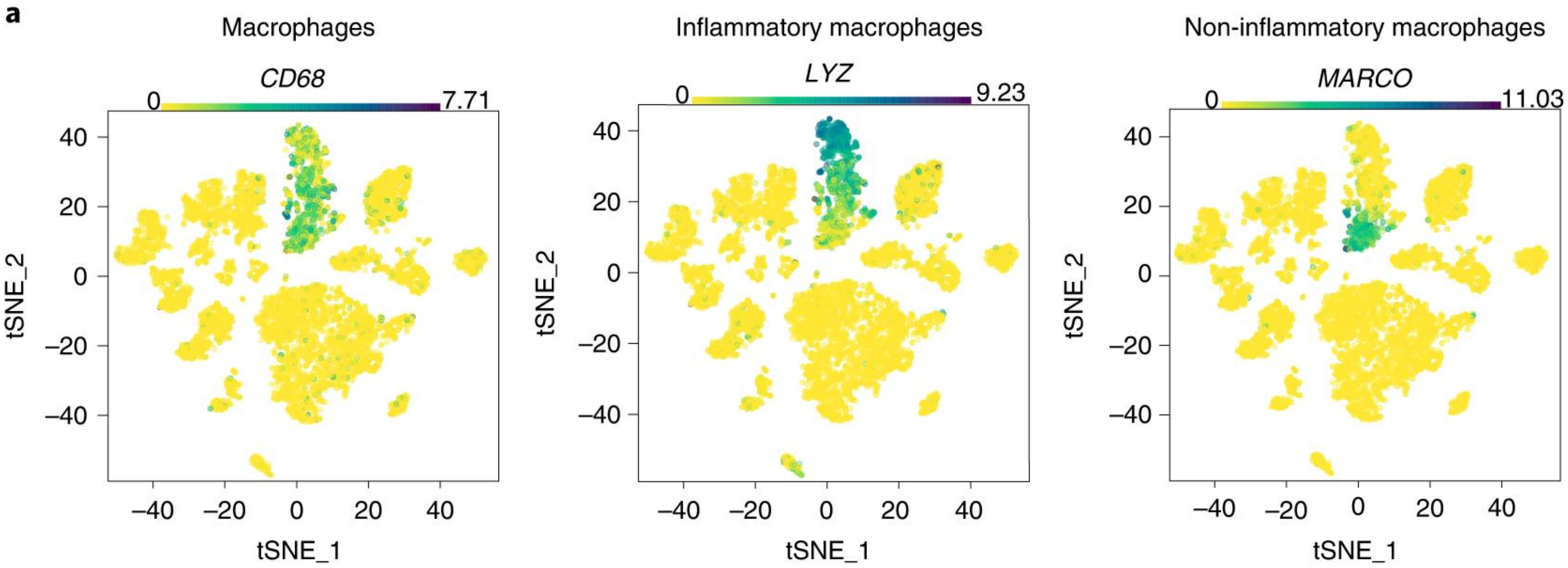


Clarke, Z.A. et al. Nature Protocols 16:2749-2764 (2021)

- Compare to previous scRNA-seq
- Use known marker genes and gene functions
- **Advice**: Always visualize the expression of marker genes on your data

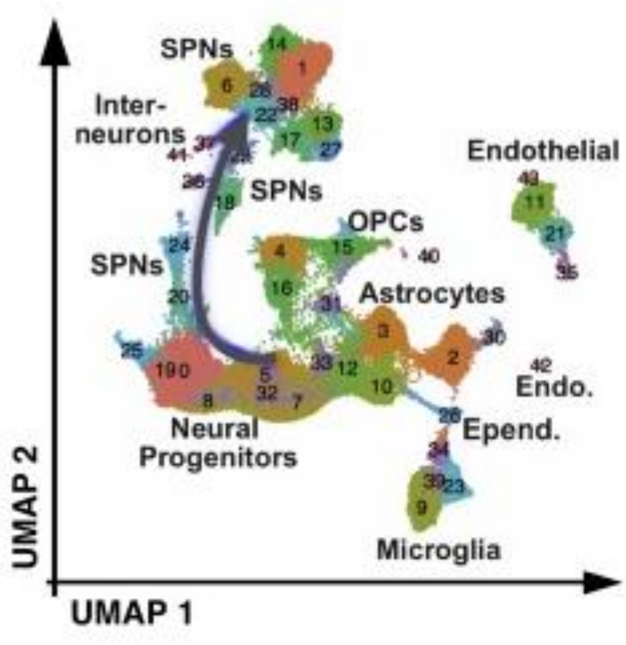
Multiple cell sub-clusters

Clarke, Z.A. et al. Nature Protocols 16:2749-2764 (2021)



- Expression pattern of marker genes across cells can guide the discovery

Trajectory (pseudotime) analysis

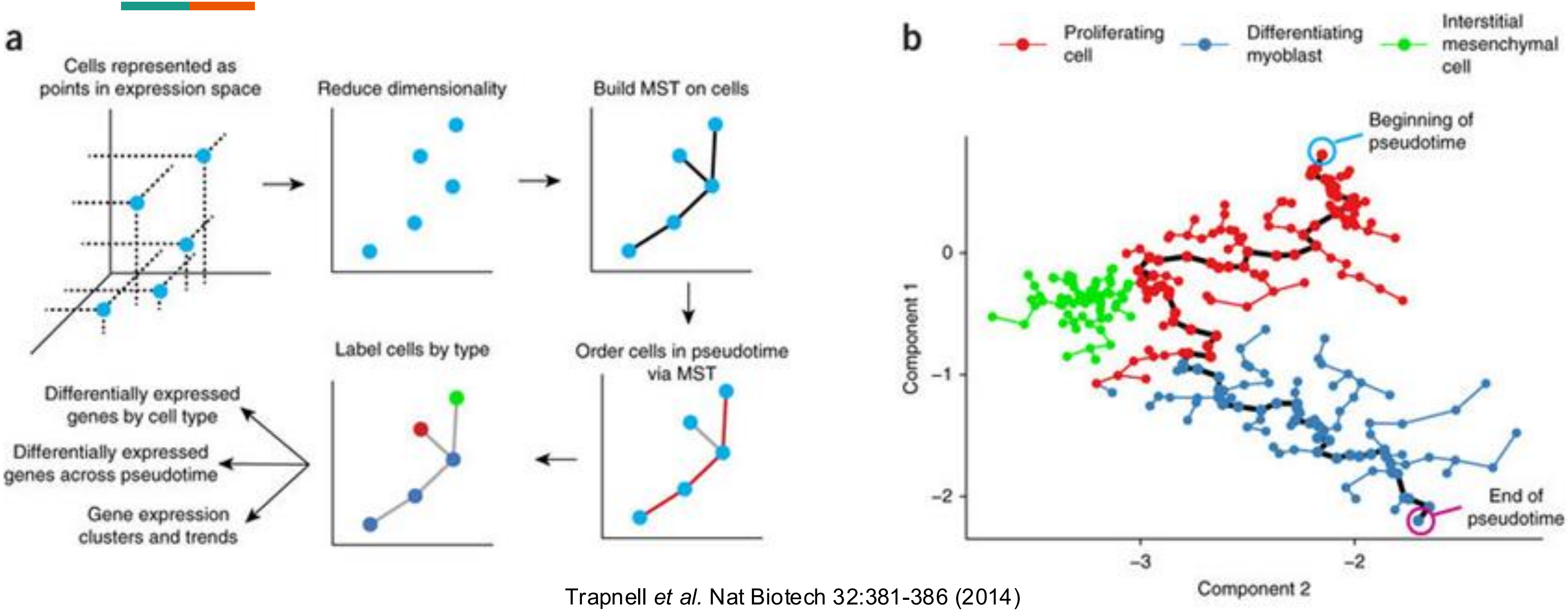


- Adjacent cells are likely in adjacent developmental stages

- Reconstruct path through cells, from one type to another

- Called **pseudotime** because all cells are collected at the same time (not an actual time-series experiment)

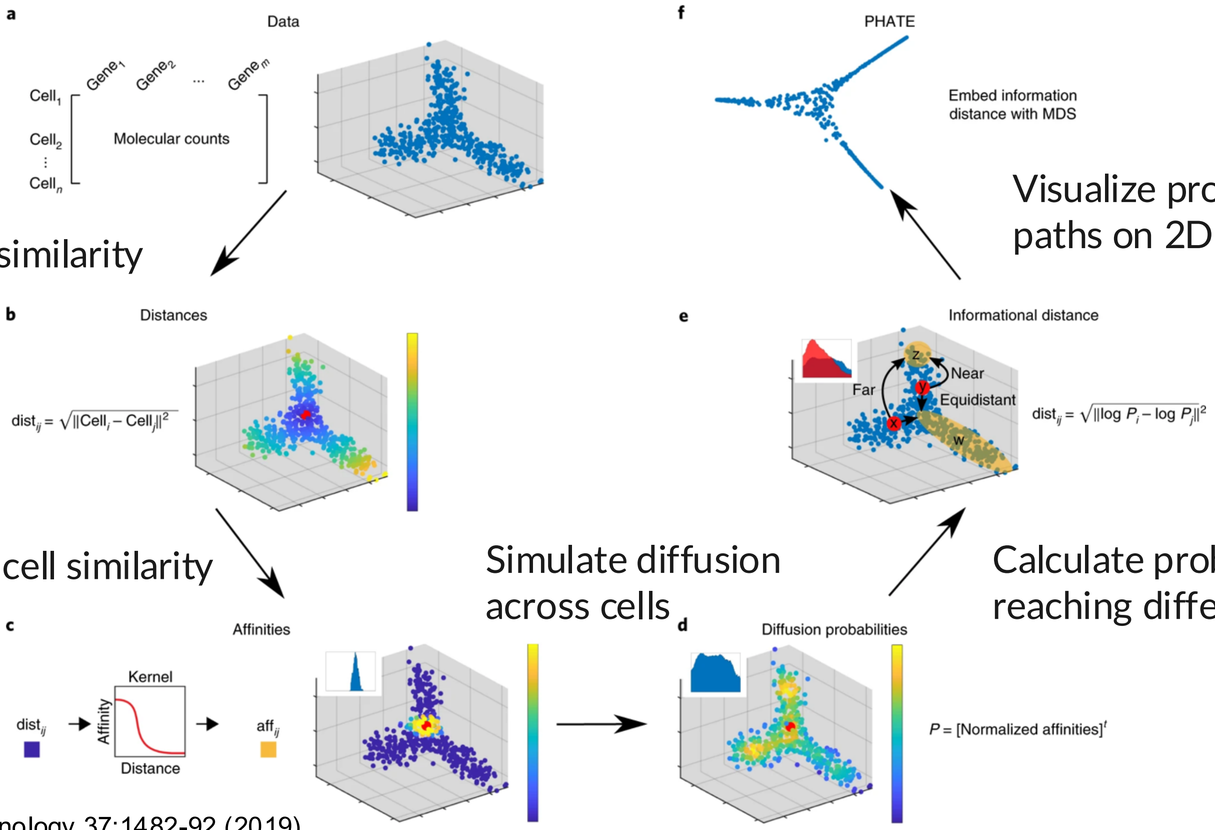
Minimum spanning tree


a
Cells represented as points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime via MST
Label cells by type
Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends
b
Proliferating cell
Differentiating myoblast
Interstitial mesenchymal cell
Beginning of pseudotime
End of pseudotime
Component 1
Component 2
0
−1
−2
−3
−2

Trapnell *et al.* Nat Biotech 32:381-386 (2014)

- **Assumption**: Cell development follows the simplest, shortest routes

Diffusion modeling for cell-cell transitions



Calculate cell-cell similarity

Normalize cell-cell similarity

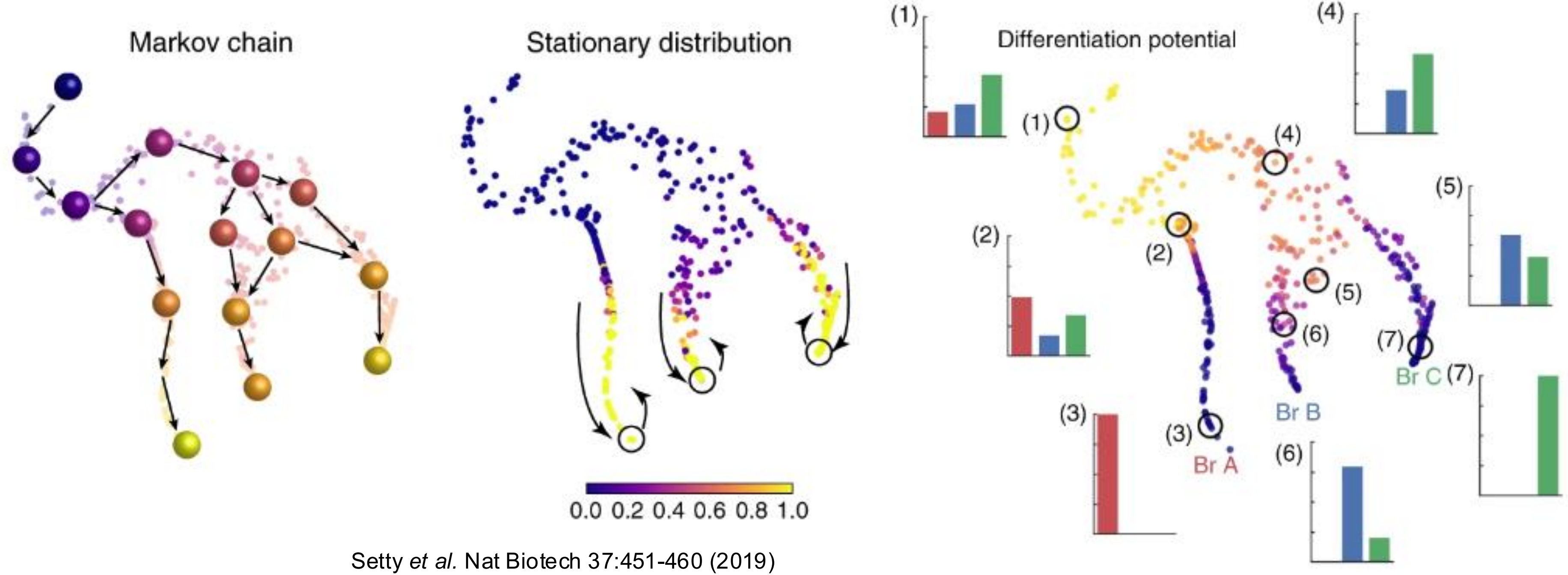
Simulate diffusion across cells

Calculate probability of reaching different cells

Visualize probability paths on 2D/3D

Moon et al. Nature Biotechnology 37:1482-92 (2019)

Estimating differentiation potential



Setty *et al.* Nat Biotech 37:451-460 (2019)

- Capability to differentiate into multiple cell types

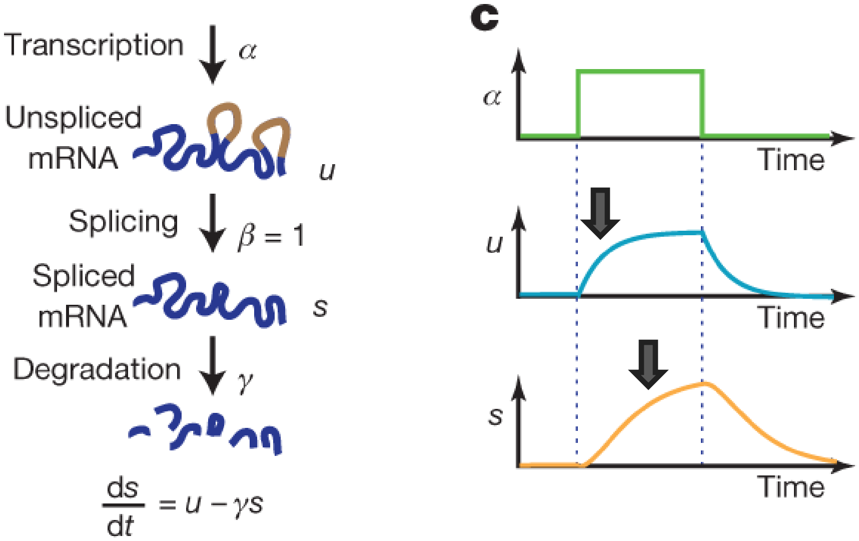
Limitation of trajectory reconstruction

- The model does not know the direction of development
  - Require user to specify

- Assume that similarity in overall transcriptomics profiles define the direction of development
  - Development is driven by a few genes and pathways

RNA velocity model

Interpreting unspliced transcripts



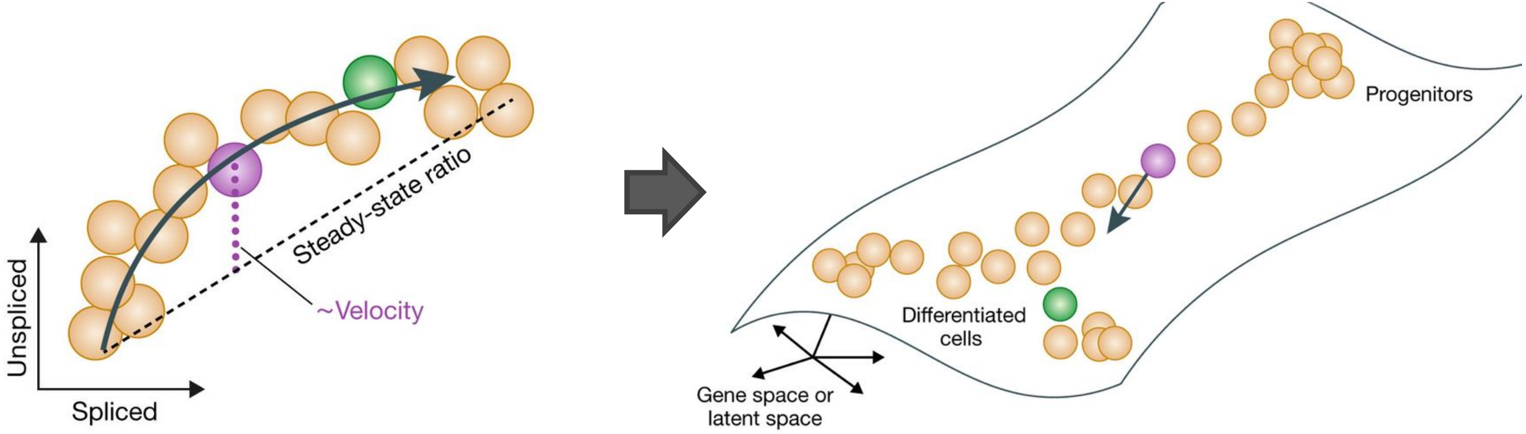
La Manno *et al.* Nature 2018

- When a gene is activated, the level of unspliced transcripts will increase first, followed by the spliced form

- **Assumption**: Cell development is driven mostly by activation (rather than repression)

- Increase in unspliced transcripts ~ direction of development!

RNA velocity model

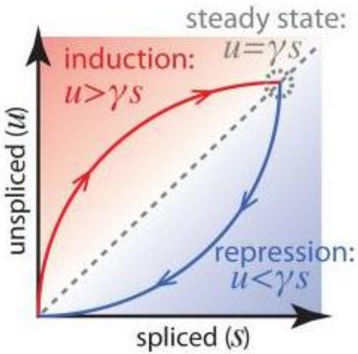


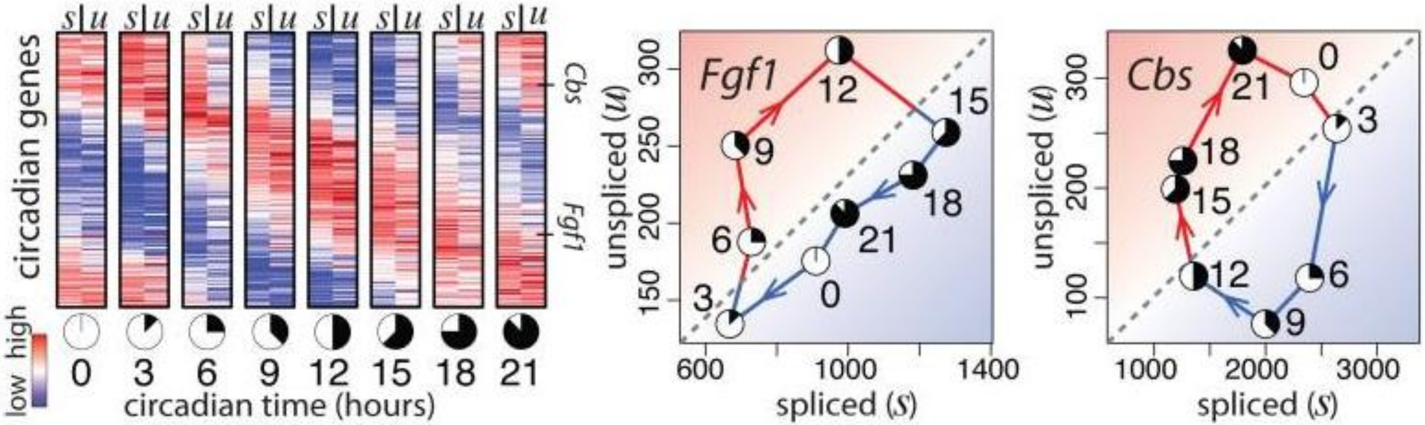
Bergen, V. *et al.* Mol Sys Biol 17:e10282 (2021)

- Use the ratio of spliced and unspliced isoforms to estimate whether a gene is being activated or repressed
- Compare ratios between nearby cells to identify direction of changes

Support of RNA velocity model

Expected pattern

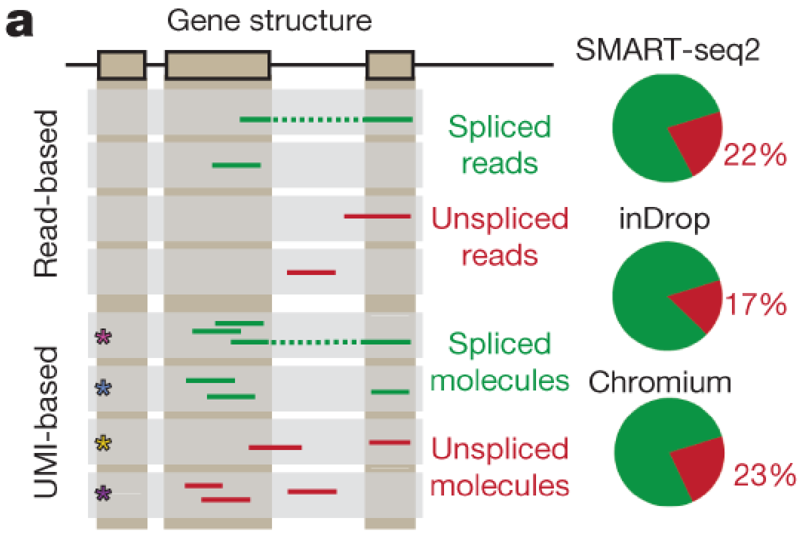




La Manno *et al.* Nature 2018

- Analyze the expression of circadian genes, whose expression cycle with the time of day

scRNA-seq already capture unspliced transcripts



La Manno *et al.* Nature 2018

- Require full-length scRNA-seq protocols

- Specialized tools identify reads that map to introns versus reads that map to splice junctions

The use of highly variable genes

Gene sets affect analysis result

- A key component of single-cell analysis is the similarity between cells
  - Similarity metric
  - **Gene set**

- Some gene are differentially expressed across cell types, some are differentially expressed across treatment conditions, some are differentially expressed across donors
  - Different gene sets reveal different biological information
  - **Advice**: Use **highly-variable genes (within sample)** that are consistently observed in multiple samples → reveal cell types

Any question?

- See you next time