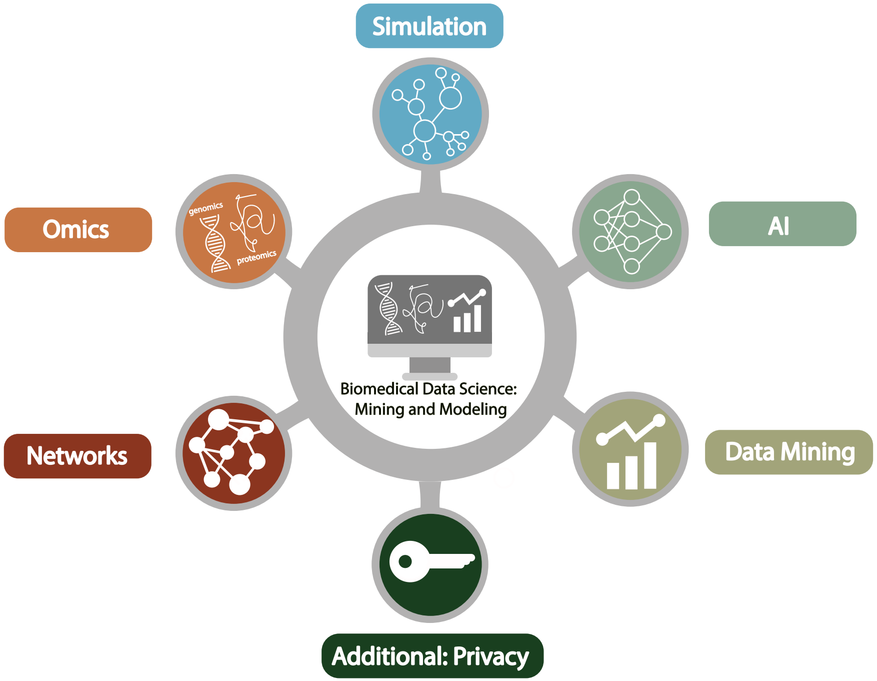
Biomedical Data Science (GersteinLab.org/courses/452)

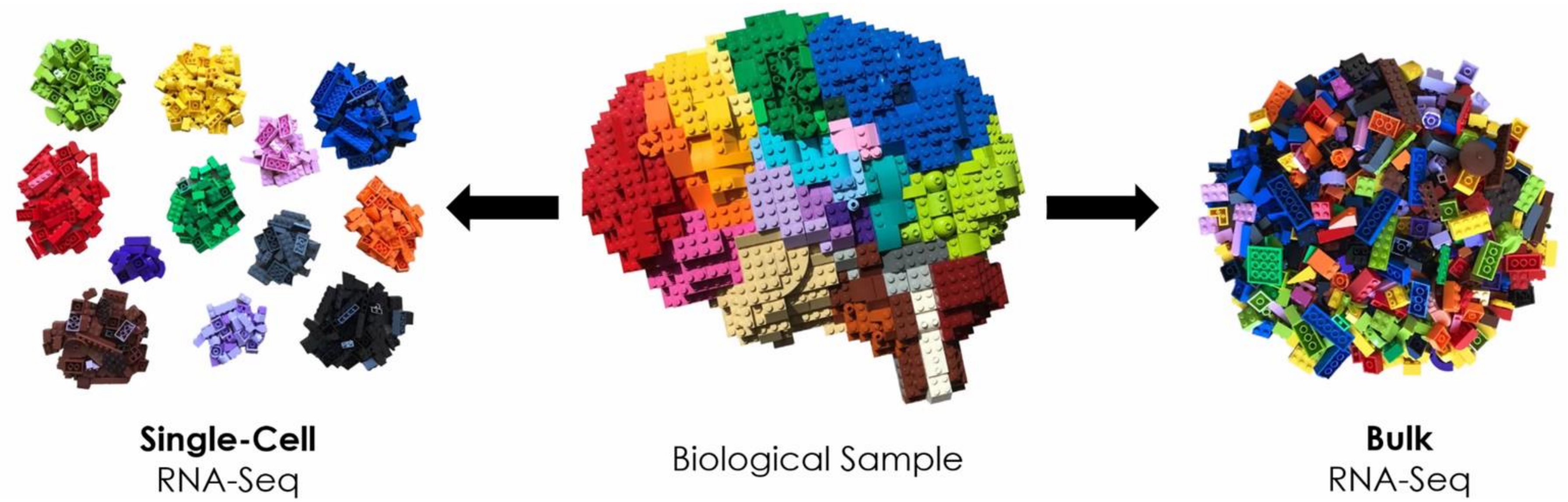
Single Cell Analysis (23m9e)



Mark Gerstein
Yale U.

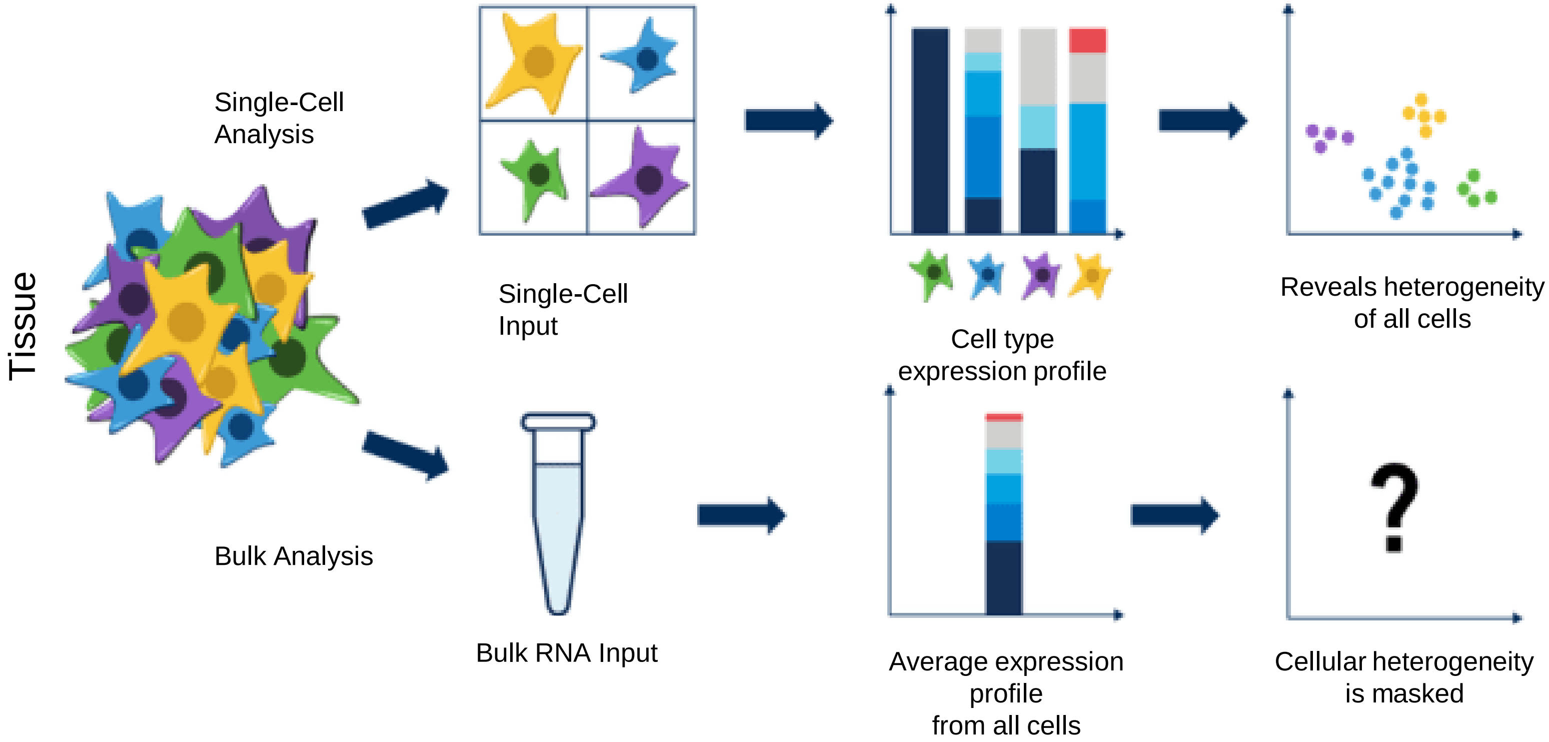
Last edit in spring ’23. All new pack which has a corresponding new video in 2023

Single-cell vs. bulk RNA



https://boxia2018.wixsite.com/boxia/momemts

Single-cell vs. bulk RNA



Tissue

Single-Cell Analysis

Single-Cell Input

Cell type expression profile

Reveals heterogeneity of all cells

Bulk Analysis

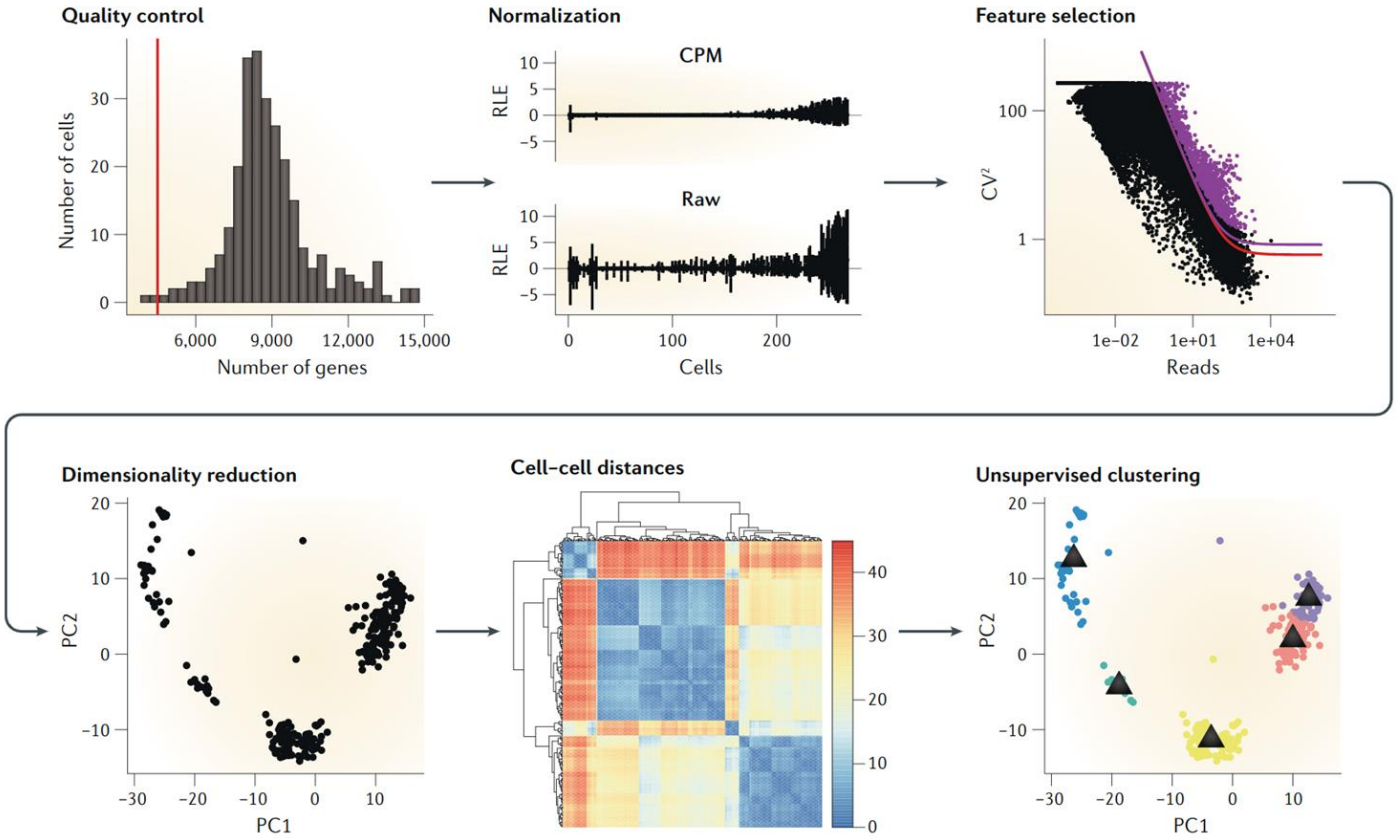
Bulk RNA Input

Average expression profile from all cells

Cellular heterogeneity is masked

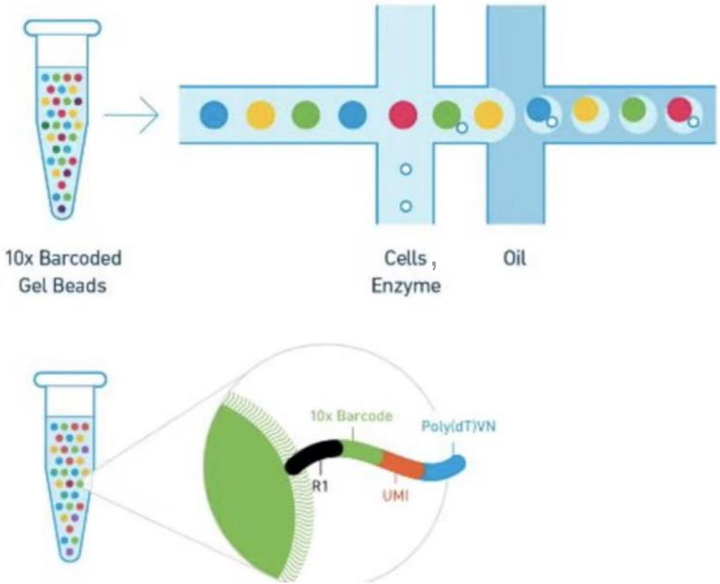
https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started

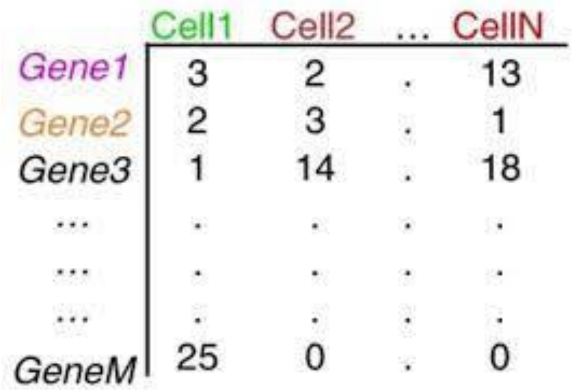
Overview of Single Cell Analysis Workflow



Kiselev et al. Nature Review Genetics 2019

Building the Expression (Count) Matrix





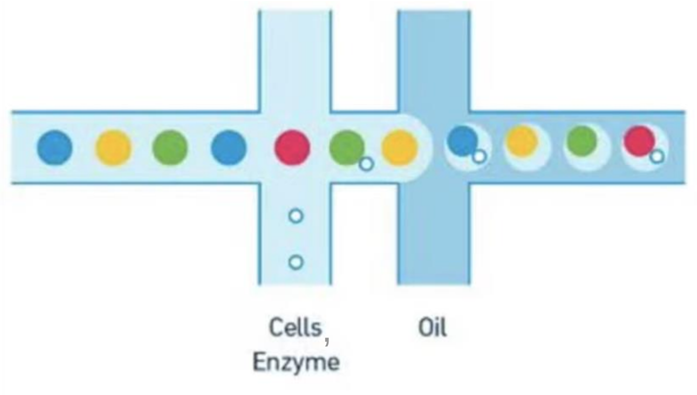
	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

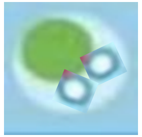
UMI (Unique Molecular Identifier)
counts the number of transcripts observed
for each gene and cell

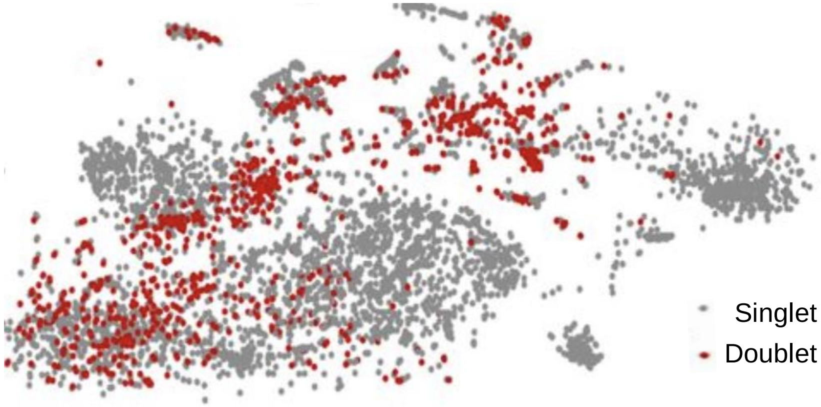
10x is specific to cell, UMI is specific to each RNA molecule.
UMI helps to differentiate between amplification copies and the original reads.

Gudodagi et al. *International Journal of Electrical and Computer Engineering* 2022

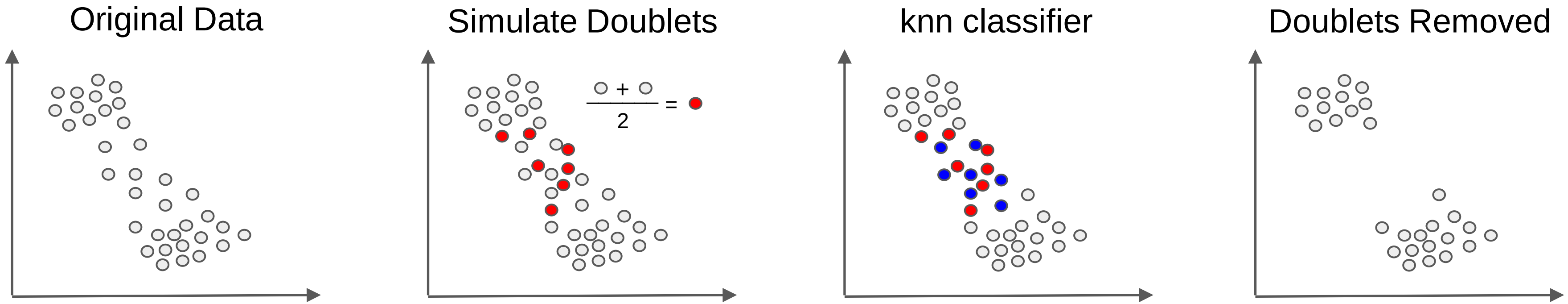
Doublet Detection & Removal







DePasquale et al. *BioRxiv* 2018



Original Data

Simulate Doublets

knn classifier

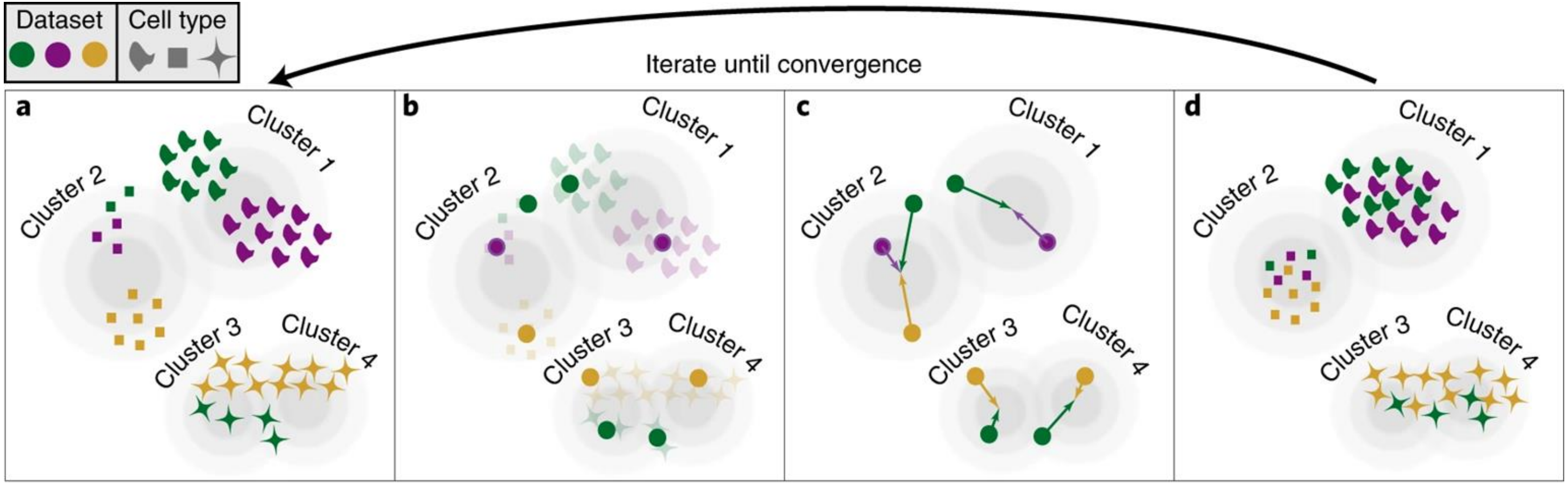
Doublets Removed

McGinnis et al. Cell Systems 2019

Batch Effect Correction

$log(\mu_{gcb}) \sim log(N_b) + \alpha_{gs} + \gamma_{sb}$ where g=gene, c=cell, b=batch, s=cluster

Mean genetic expression value is affected by total counts in each batch (N_b), natural expression (α_g), and cluster-dependent batch effect (γ_{sb})



Soft assign cells to clusters, favoring mixed dataset representation

Get cluster centroids for each dataset

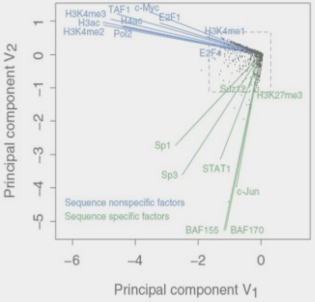
Get dataset correction factors for each cluster

Move cells based on soft cluster membership

Korsunski et al. Nature Methods 2019 (“Harmony”)

Dimensionality Reduction

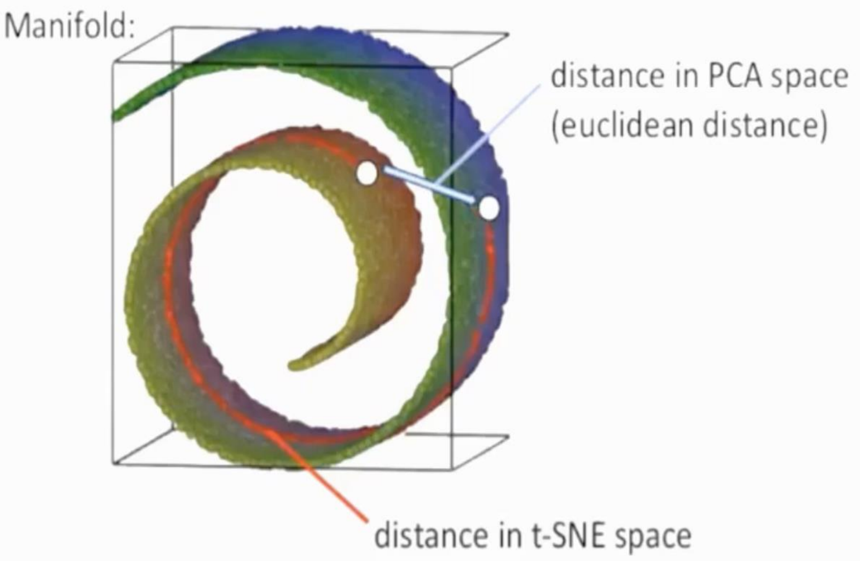
Previous
Discussion
of SVD/PCA



Each gene represents a dimension (~10k-D expression)

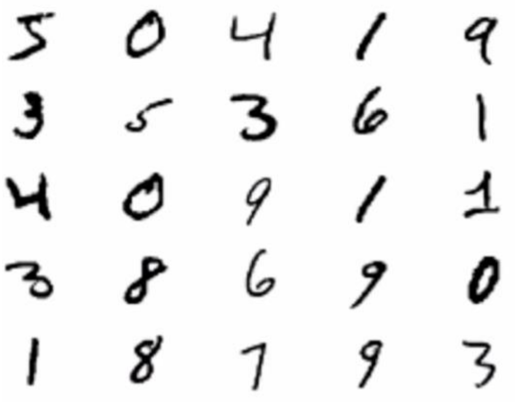
Dimensionality reduction is necessary for
visualization of high-dimensional datasets, and
distance estimates in high dimensions are unreliable

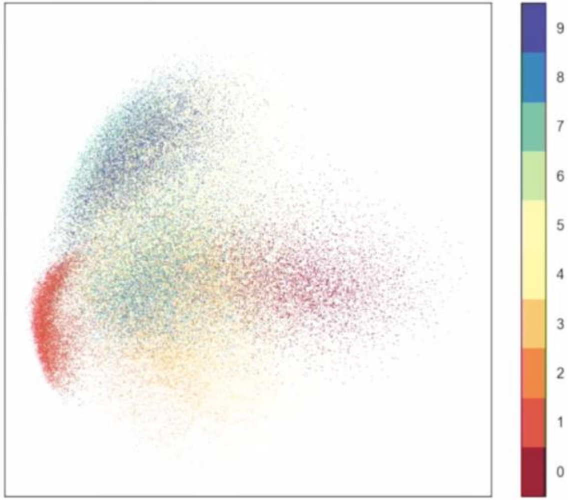
UMAP and t-SNE sacrifice global distance measurements
to better capture local distances,
so distances between clusters are not meaningful.

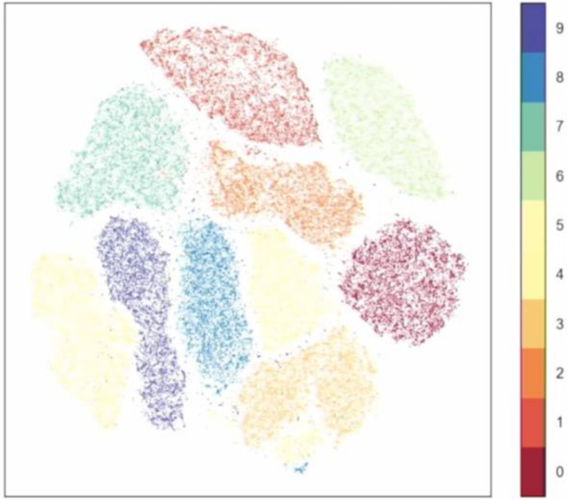


Takahashi et al. IEEE Transactions on Visualization and Computer Graphics 2009

PCA vs t-SNE



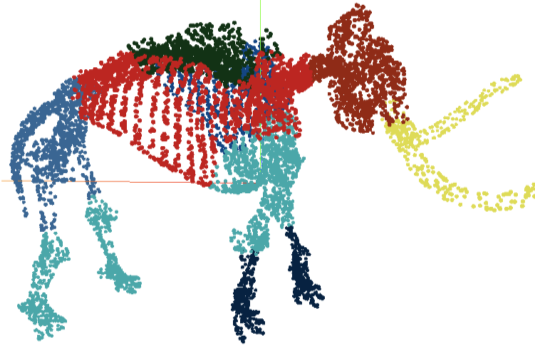


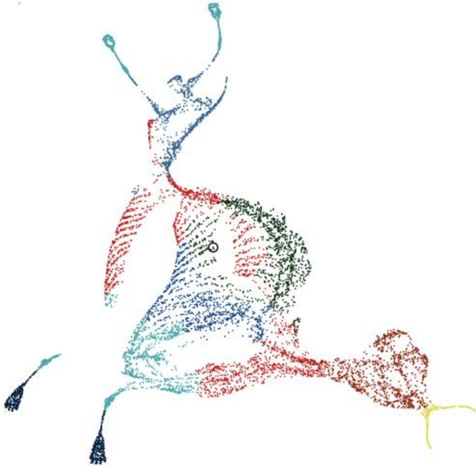


https://meta.caspershire.net/umap/

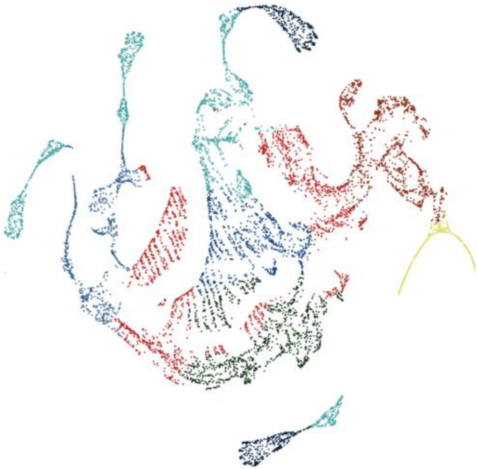
Loss of Global Distance in UMAP

‘perplexity’ represents significance of the global distance info.

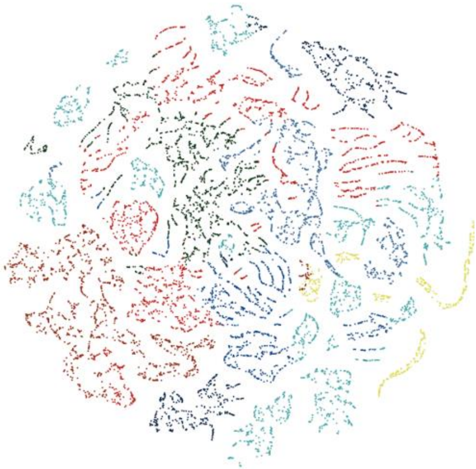




perplexity = 2000



perplexity = 500



perplexity = 50

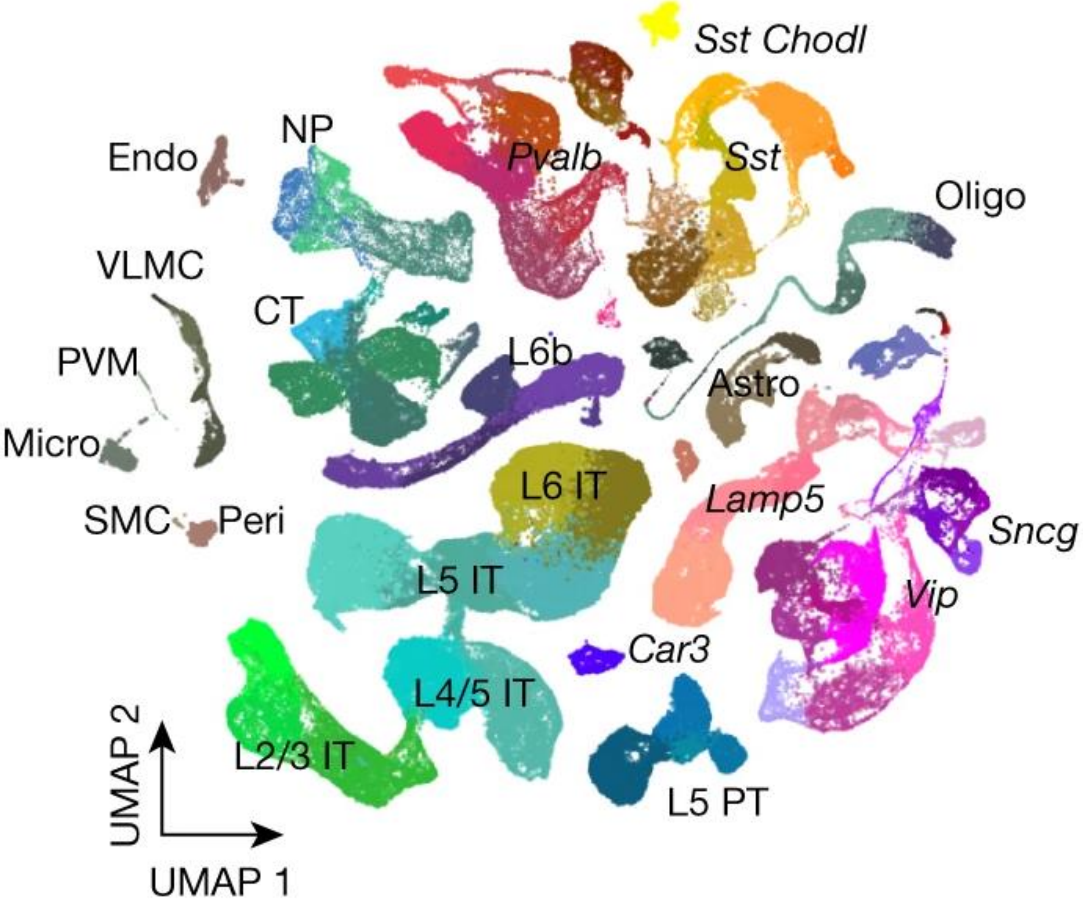
https://pair-code.github.io/understanding-umap/

Clustering to Determine Cell Types

Communities are dense groups of nodes

They could be related to cell types, cell states, or a disease. The goal is to identify communities of cells with similar expression profiles

Clustering for cell typing often uses **connectivity based approaches** (discussed earlier)…



[Yao et al. ('21) *Nature*]

Louvain Maximizes Modularity

(Calculation of Modularity Q)

$$Q = \frac{1}{2m}\sum_{i,j}\left(W_{ij} - \boxed{\frac{k_i k_j}{2m}}\right)\delta_{\sigma_i\sigma_j}$$

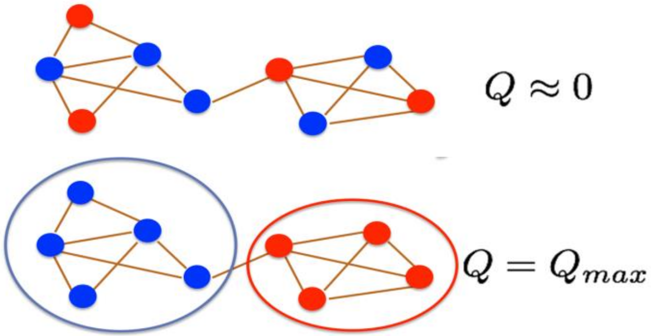
adjacency matrix

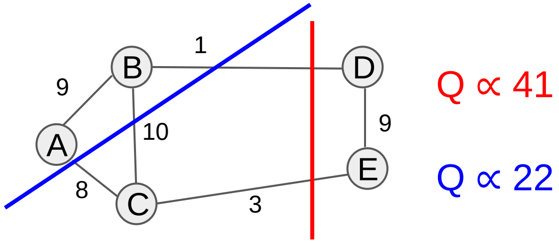
degree of node i

number of edges

expected number of edges between i and j

whether or not i, j are in the same module





W_ij	A	B	C	D	E
A	-	9	8	-	-
B	9	-	10	1	-
C	8	10	-	-	2
D	-	1	-	-	9
E	-	-	2	9	-

	A	B	C	D	E
k_i	17	20	21	10	12

Newman Phys. Rev. E 2004

Louvain maximizes modularity

(Overall Algorithm Flow)

#1 Start:

Each node (cell) having its own community.

#2 Moving nodes:

Repeat scanning all nodes until no change increases Q (from **a** to **b**)

```
{{
    Move each node to the one of its neighbor communities
        that maximizes ΔQ;
    Or start a new community
}}
```

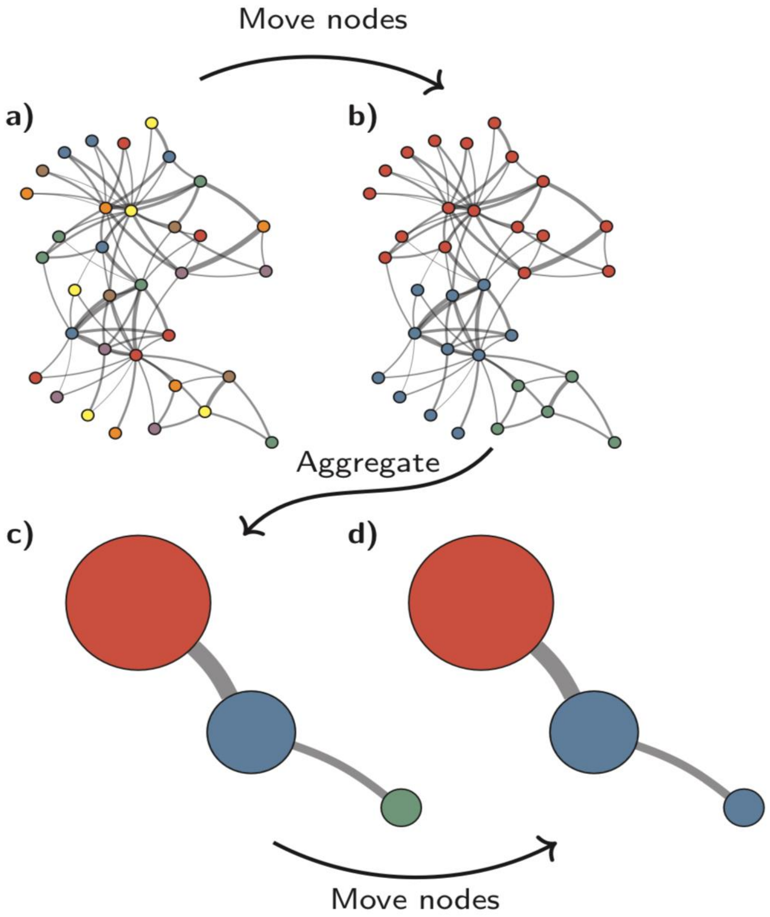
#3 Aggregate:

Turn each community into a node.

Edges between communities are added up as the weight.

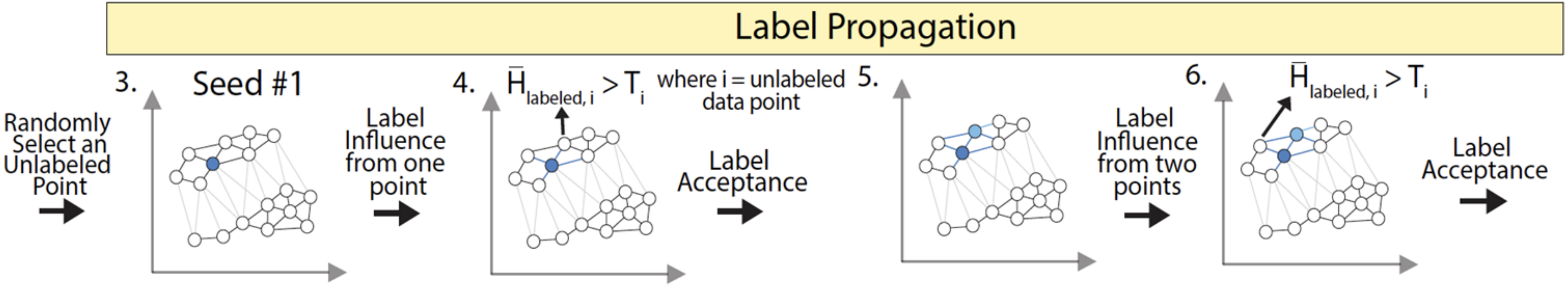
#4 Repeat from #2

Stop at desired resolution



Traag et al. Scientific Reports 2019

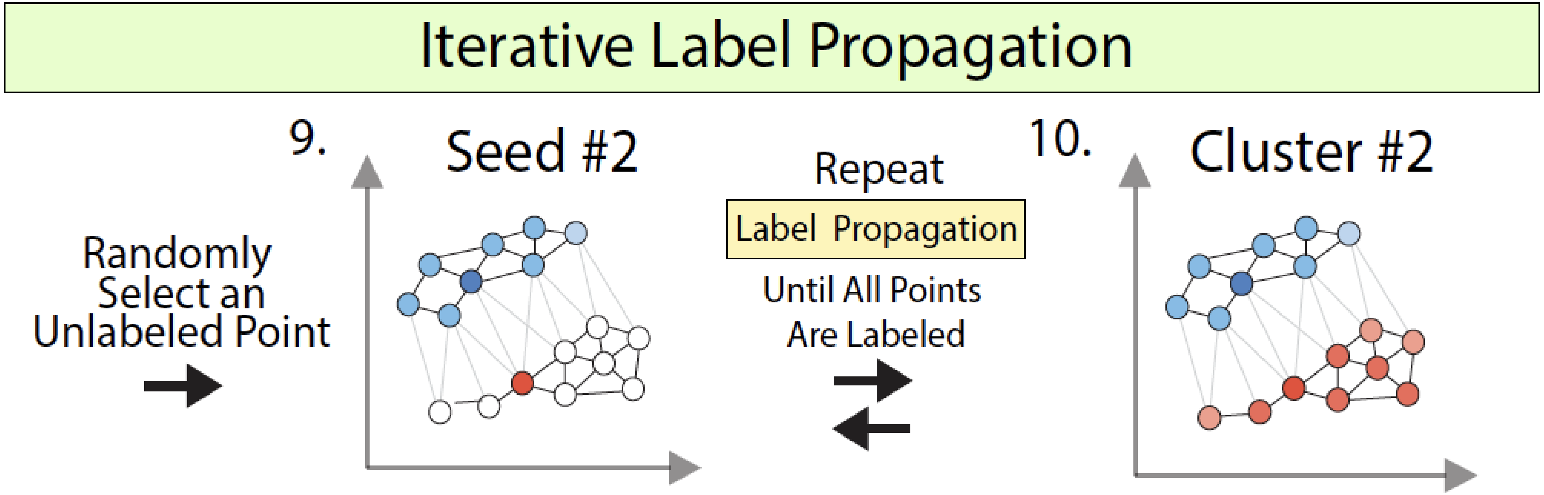
Forest Fire Clustering: To Find One Cluster


Label Propagation
3.
Seed #1
Randomly Select an Unlabeled Point
Label Influence from one point
4.
$\bar{H}_{labeled,\,i} > T_i$
where i = unlabeled data point
Label Acceptance
5.
Label Influence from two points
6.
$\bar{H}_{labeled,\,i} > T_i$
Label Acceptance

1. Randomly select a points (seed) to label
2. Label influence radiate from the labeled points
3. Check if other unlabeled points experience label influence higher than their threshold
   a. If so, the unlabeled points receives the same label as the seed.
   b. If not, check later when more points are labeled and see if the cumulative label influence is able to cross the threshold.
4. Repeat from step 2

Chen *et al.* Nat. Comm. '22

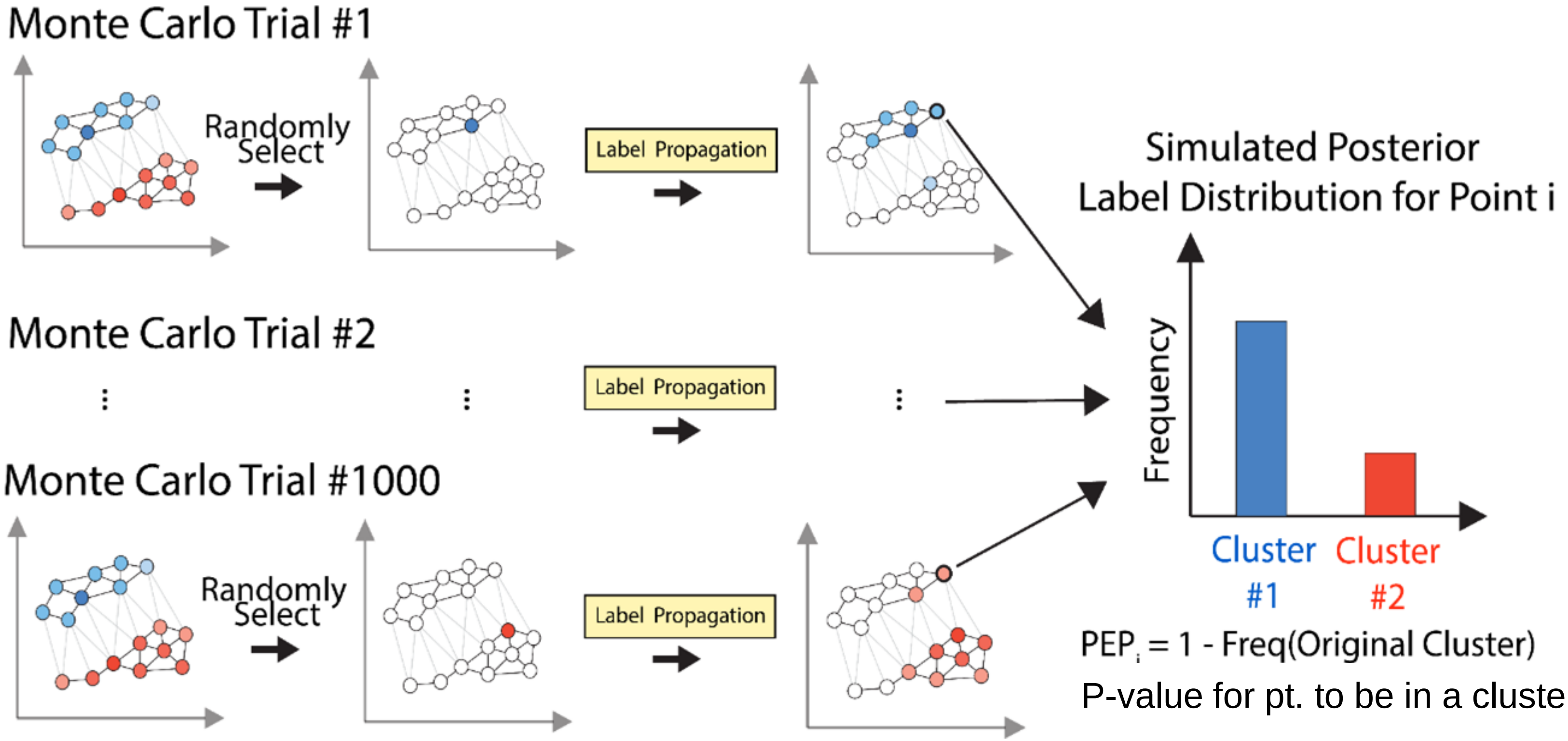
Forest Fire Clustering: To Find All Clusters


Iterative Label Propagation
9.
Seed #2
Randomly Select an Unlabeled Point
Repeat
Label Propagation
Until All Points Are Labeled
10.
Cluster #2

1. Select a new unlabeled points (seed) to label
2. Repeat Label Propagation for as many times as needed to label every data point.

Chen *et al.* Nat. Comm. '22

Forest Fire Clustering: Internal Validation via Monte Carlo Simulation

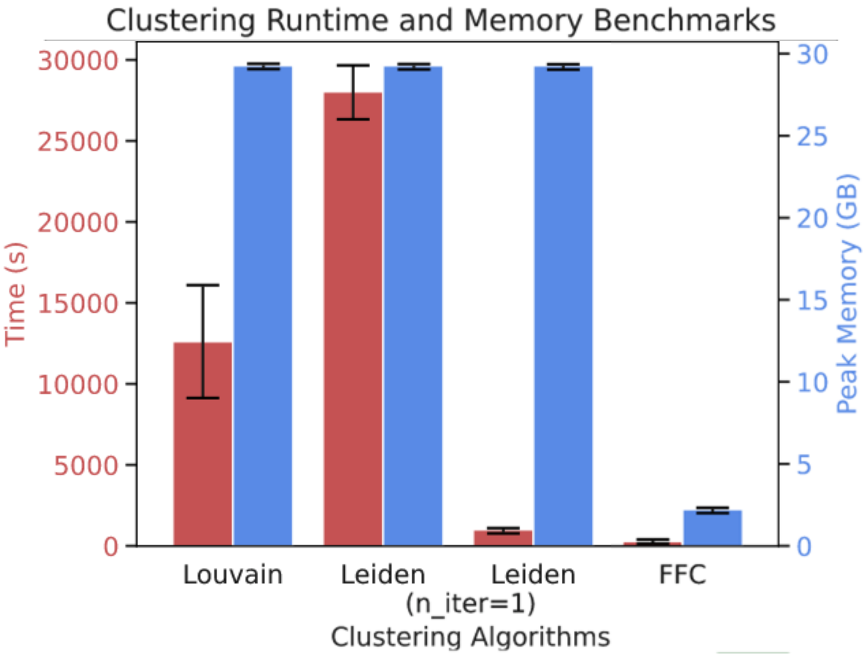


$PEP_i = 1 - Freq(Original Cluster)$

P-value for pt. to be in a cluster

Chen *et al.* Nat. Comm. '22

Forrest Fire is Substantially Faster than Louvain



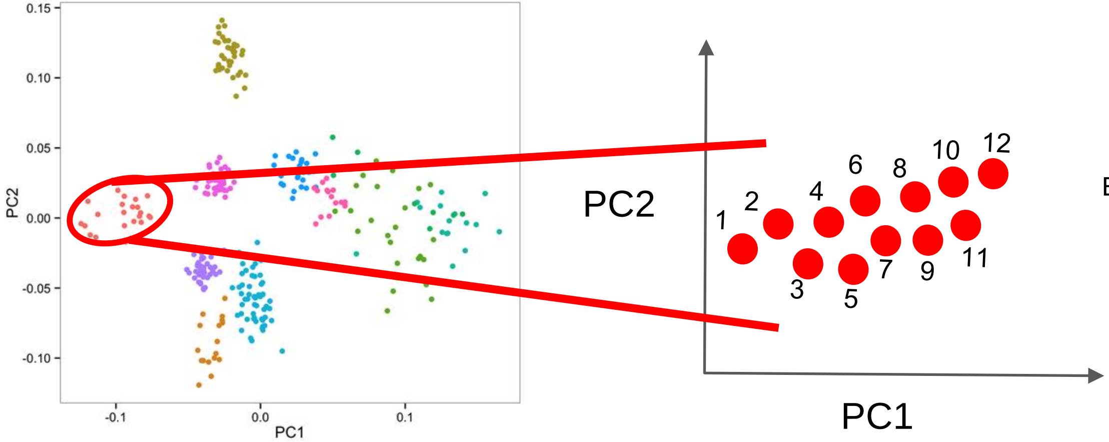
Chen *et al.* Nat. Comm. 2022

Imputation

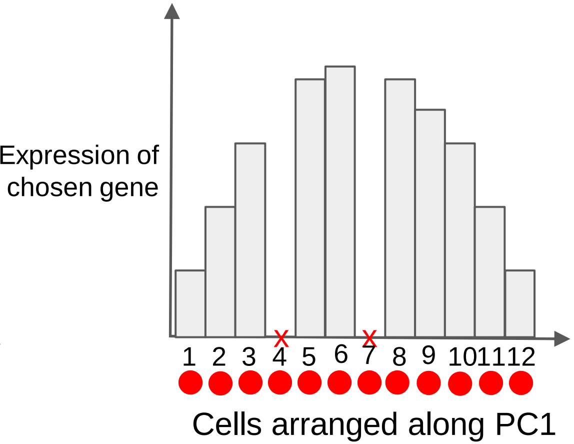
Some genes will fail to be detected, even if they are expressed.

Find a structure within the whole data
Fill in a derived mathematical estimate for undetected genes
Minimize ‘false’ effects of the underlying model

MAGIC, SAVER, DrImpute

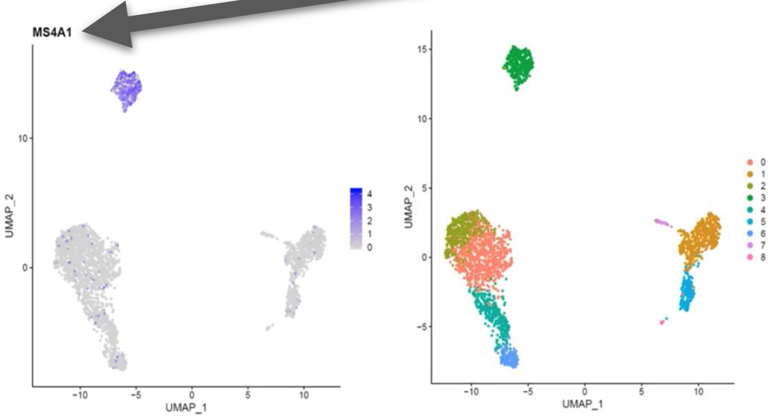


Cells 4 & 7 are dropouts

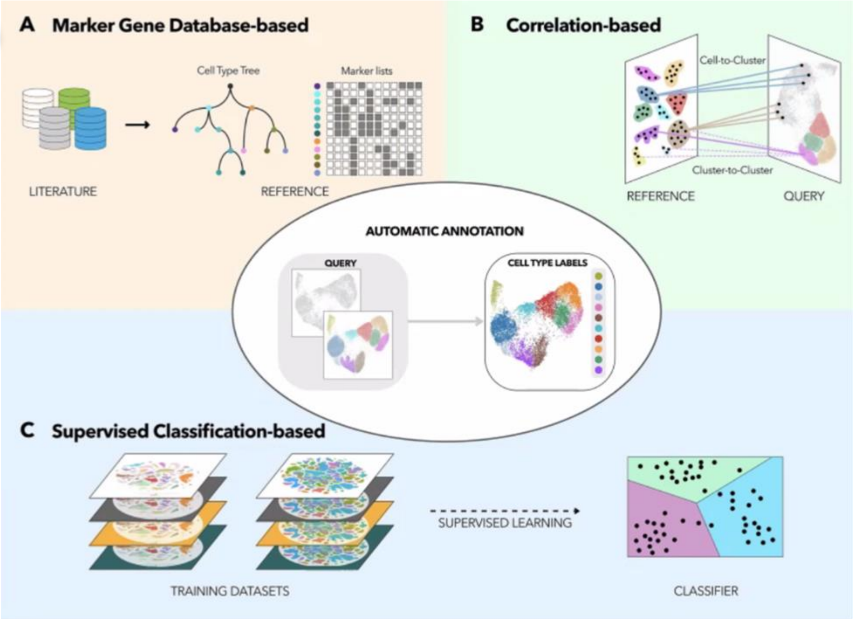


Van Dijk et al. Cell 2018 (“MAGIC”)

Transferring Cell-type Annotation



Marker Genes are active only in a specific cluster (i.e. cell type)



The alternative is to compare g.e. profiles of different experiments

Azimuth uses weighted-nearest neighbor method,
where the weights are determined for each cell and each modality
such that the marker genes are consistent (shared biological state)
across different data qualities and modalities.

https://chipster.csc.fi/material/single-cell/scrnaseq2020.pdf

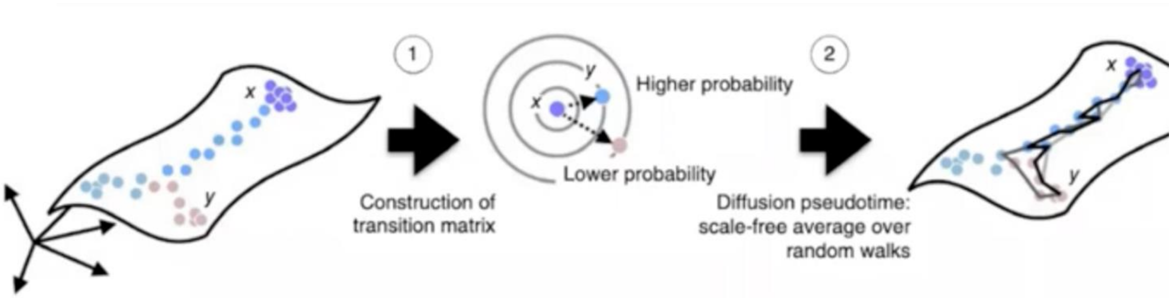
Pasquini et al. Comp. and Struct. Biotech. Journal 2021

Using Pseudotime –
going beyond discrete cell-type clusters

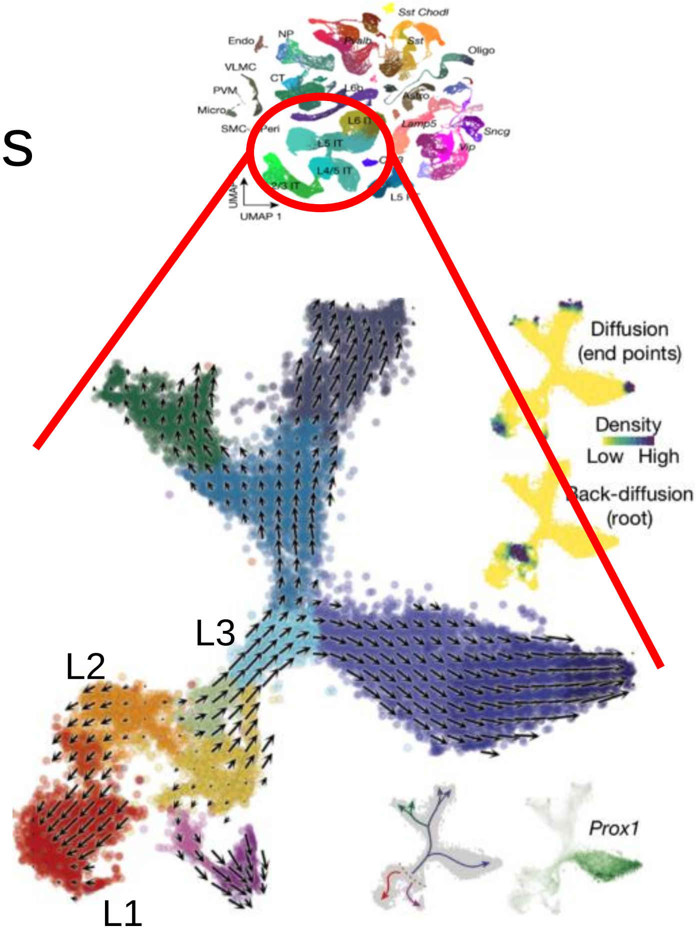
Where discrete categorization is not suitable;
pseudotime could represent time, chemical concentrations or spatial positions

Diffusion Pseudotime
Determine probability of transition between cell positions,
by constructing a weighted nearest-neighbor graph.
Find shortest random walk paths using transition matrix
Number of steps represents the amount the pseudotime

DeepVelo



Haghverdi et al. Nature Methods 2016



La Manno et al. Nature 2018

Key references

Single Cell overview [goes over every step]**:**

Andrews, Tallulah S., et al. "Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data." *Nat. protocols* (2021).

tSNE UMAP key concepts:

https://pair-code.github.io/understanding-umap/

Louvain clustering upgrade [method section]**:**

Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." *J. of Stat. Mech.: theory and experiment* (2008)

Pseudotime [first page summarizes the algorithm]**:**

Haghverdi, Laleh, et al. "Diffusion pseudotime robustly reconstructs lineage branching." *Nat. methods* (2016).

Annotation [Fig. 1, explained in the beginning of Results section]**:**

Stuart, Tim, et al. "Comprehensive integration of single-cell data." *Cell* (2019).