Single-Cell RNA-seq Analysis
Methods Overview

Horacio Gómez-Acevedo
Department of Biomedical Informatics
University of Arkansas for Medical Sciences

August 16, 2021

UAMS

Paper?

Today's session is lossely based on this paper

Figure: Paper

sc-RNA seq

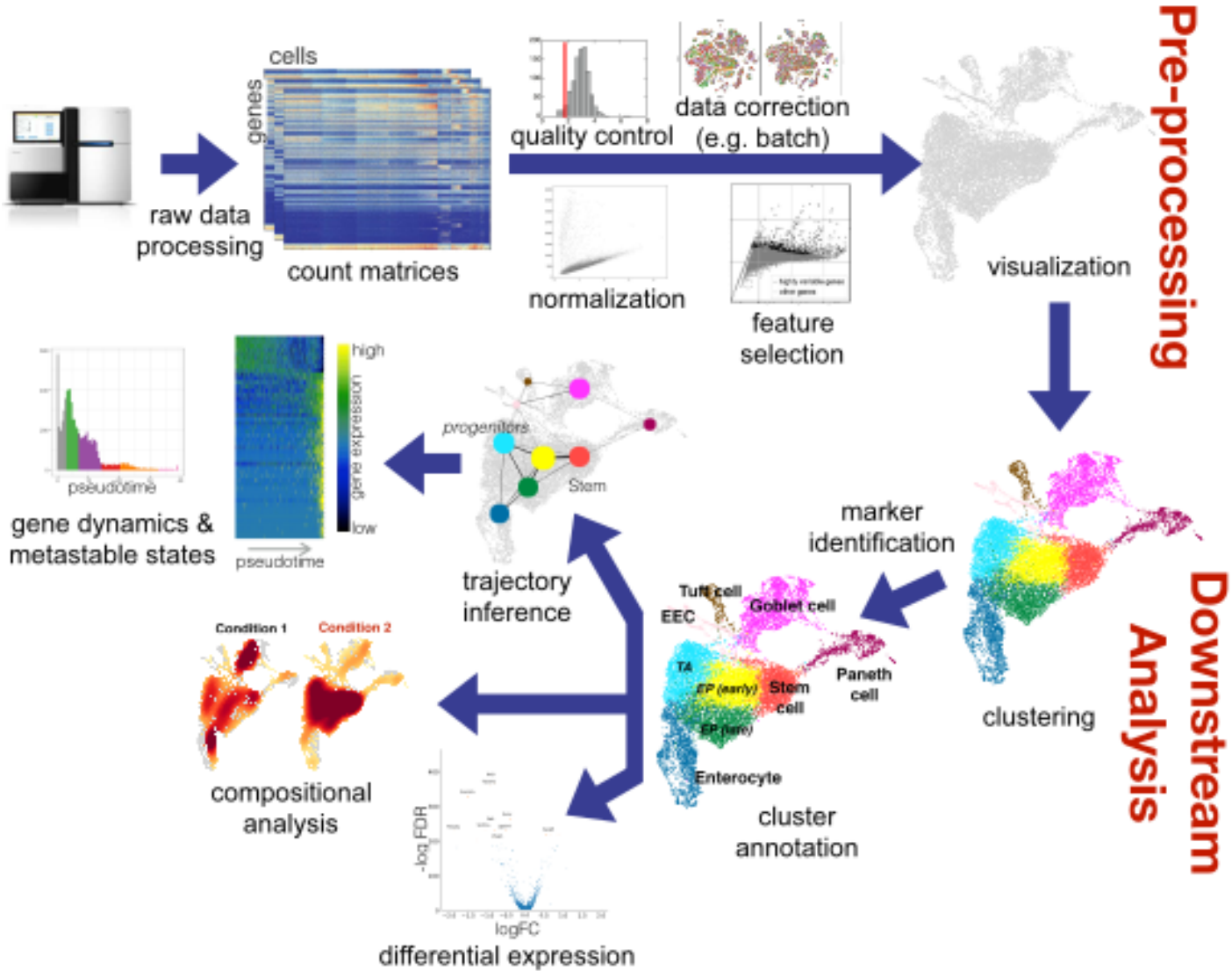
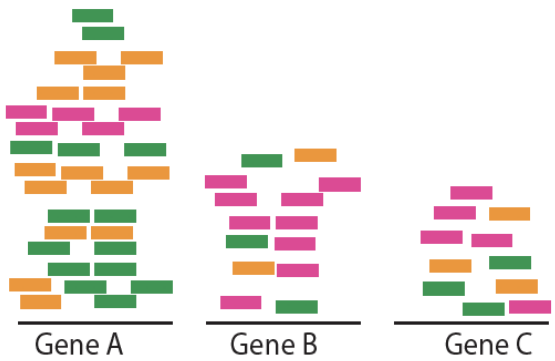
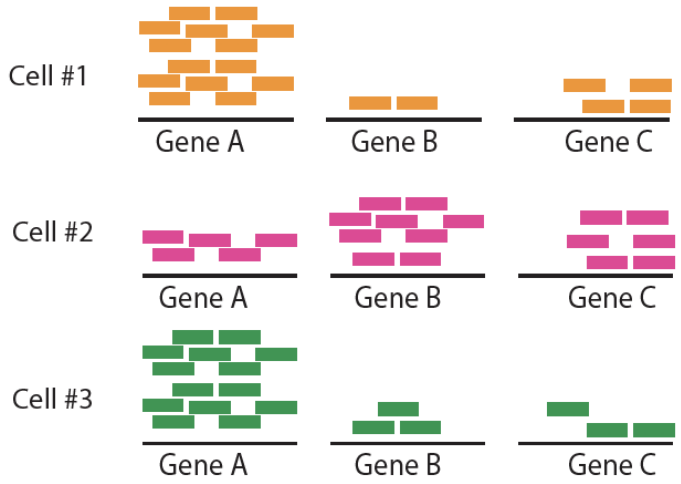


Figure: Overview

Bulk vs Single-cell RNAseq



(a) Bulk RNAseq



(b) Sc-RNAseq

Space: the final frontier

The mathematical representation of the data is given by placing the readings (RPKM) of each individual gene in an "axis" of an n dimensional space. In this case n represents a large number of genes.

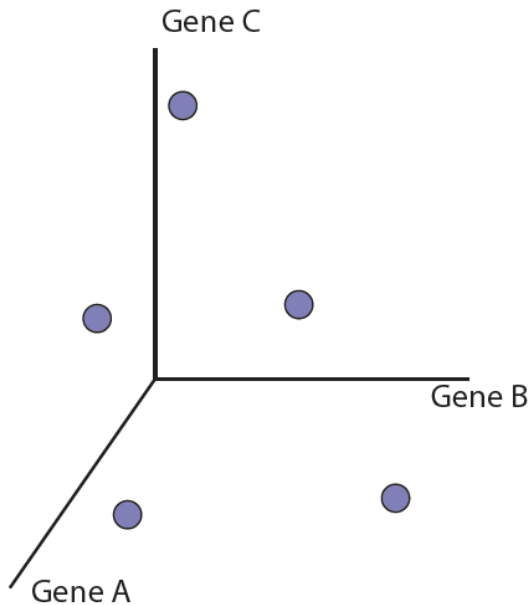


Figure: Spatial representation of scRNAseq data

PCA interlude

Principal Component Analysis is one of the most commonly used methodologies to "inspect" high dimensional data.

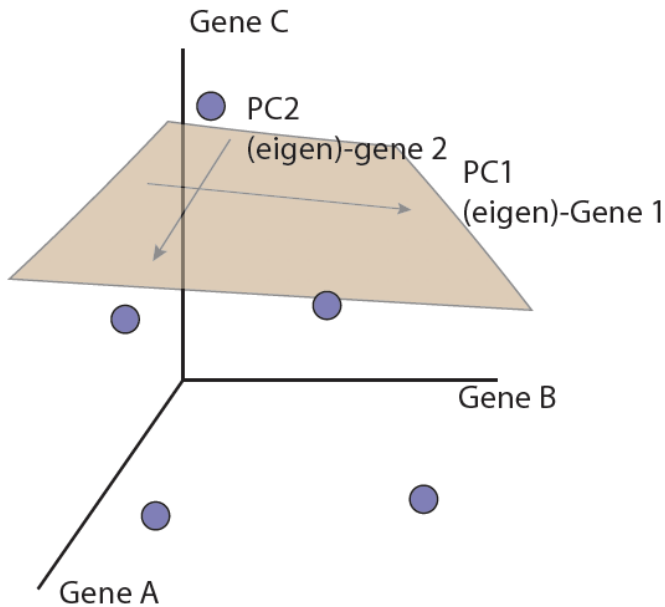


Figure: Plane projection of data

PCA projection

The main goal of PCA is to find the representation of our data in a lower dimensional space (mostly a plane). But the selection of such plane should preserve original data variability.
Formally, the two directions represent the directions of the maximal variability of our data.

RGB example

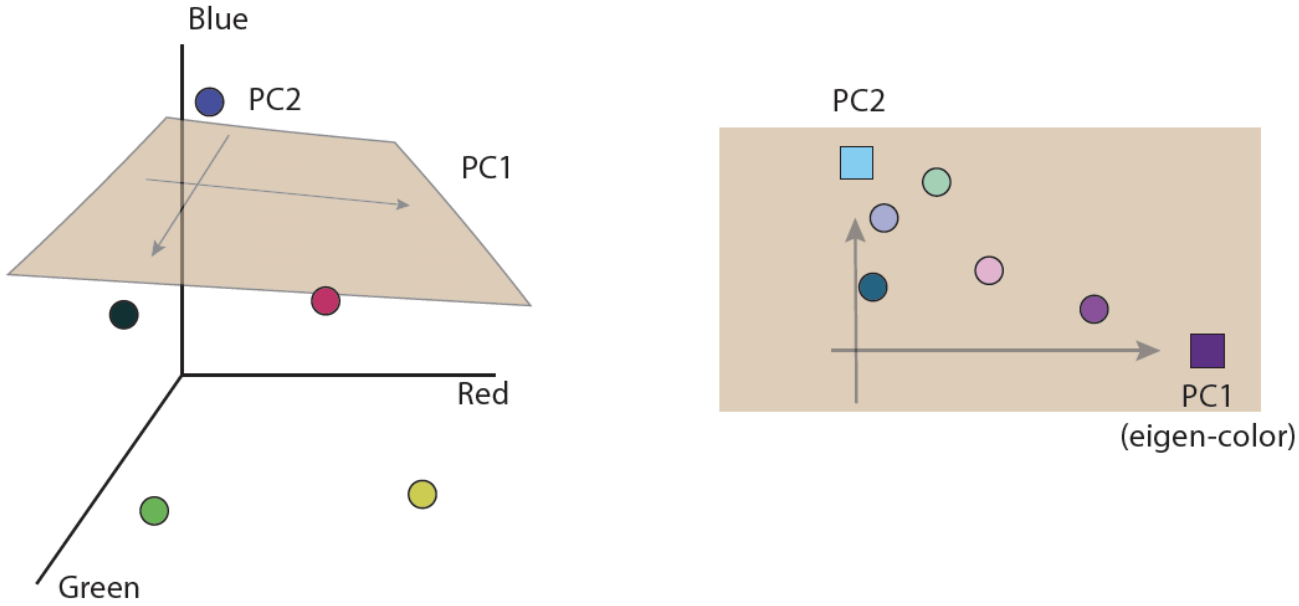


Figure: Plane projection of RGB data

Clustering

Clustering refers to a set of computational techniques for finding subsets or *clusters* in a data set. Clustering is among the so-called **unsupervised learning methodologies** .

$$\text{Unsupervised} \neq \text{Automatic}$$

Thus, the main goal of clustering is to find homogeneous subgroups among the data.

Clustering terminology

Let's define a distance between two points (in a plane), say d.
Also, the *centroid* of a cluster is the mean of their observations.

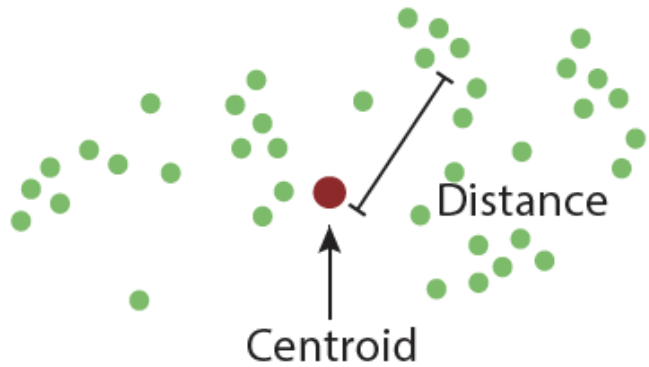


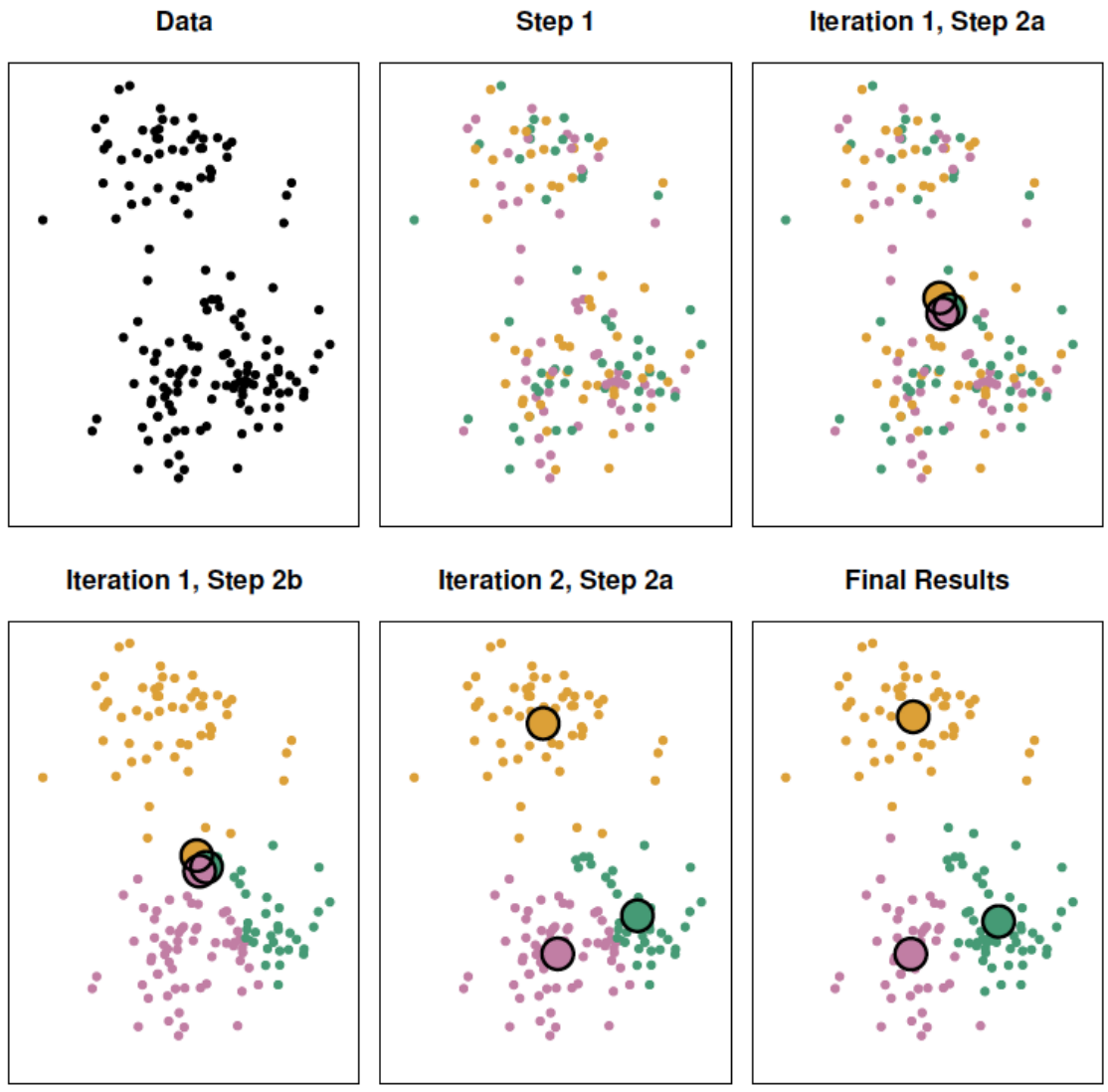
Figure: Distance and Centroid of a cluster

K-means clustering

One of the simplest methods for clustering is K-means clustering. We begin with a data set and a value K fixed by a human (say $K = 3$).

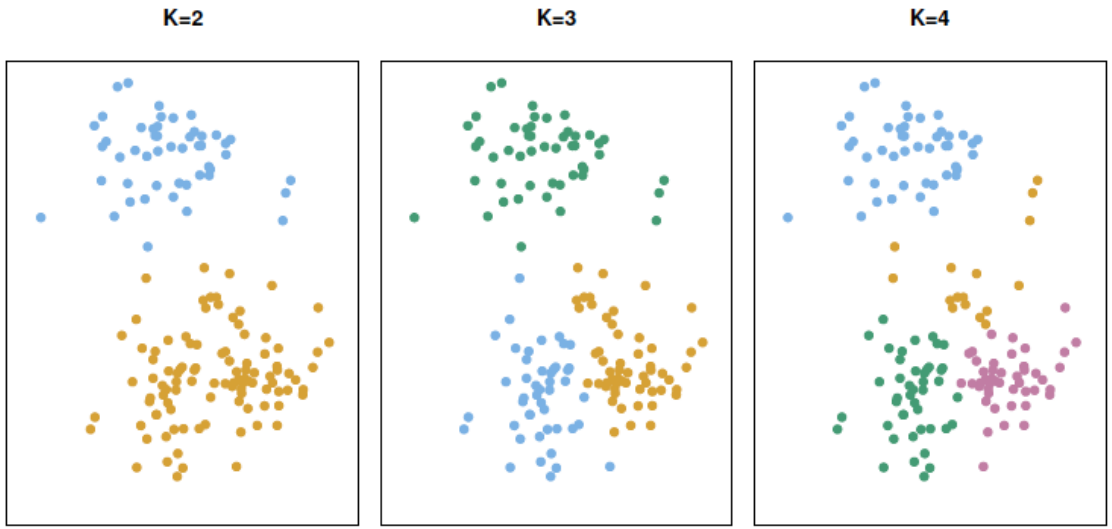
1. Assign randomly a value between 1 and K to each data point.
2. Iterate the following procedure until the clusters assignments stop changing
   - 2.1 Find the centroid for each of the K clusters.
   - 2.2 Each point will be assigned to the cluster K whose distance is the smallest. If two or more are equidistant, select randomly the cluster among the equidistant clusters.

K-means clustering picture



Problems with K-means clustering

- Selection of the distance function d (Euclidean, Pearson correlation, arccos, etc.)
- Selection of K.



An Introduction to Statistical Learning (Chapter 10)

t Stochastic Neighbor Embedding

t-SNE maps a set of high-dimensional points to a plane, such that ideally, close neighbors remain close and distant points remain distant.
Informally, the algorithm places all points on the 2D plane, initially at random positions, and lets them interact as if they were physical particles.
The interaction is governed by two laws:

- all points are repelled from each other
- each point is attracted to its nearest neighbors

This methodology is governed by a parameter called **perplexity**.

tSNE vs PCA

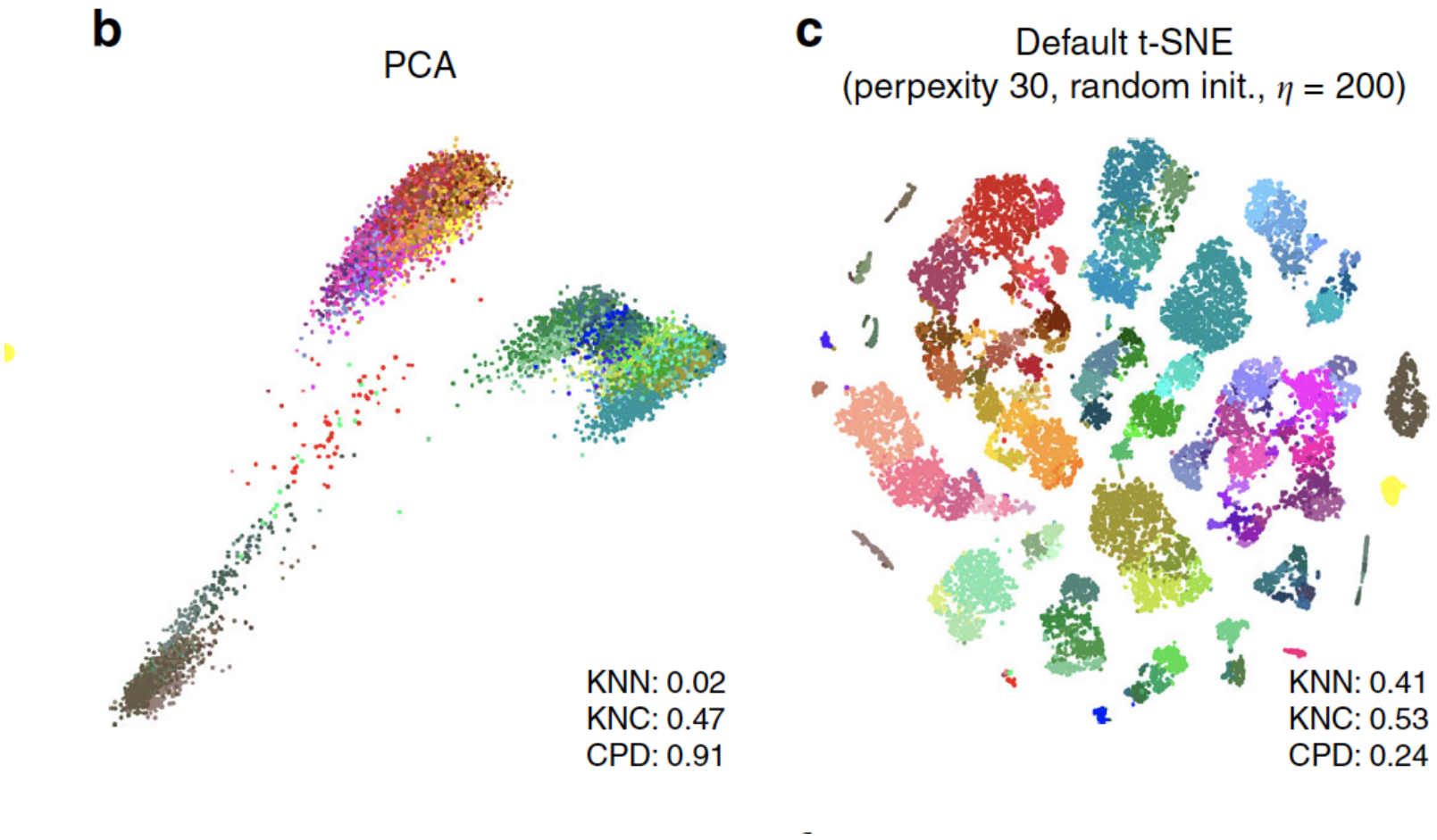
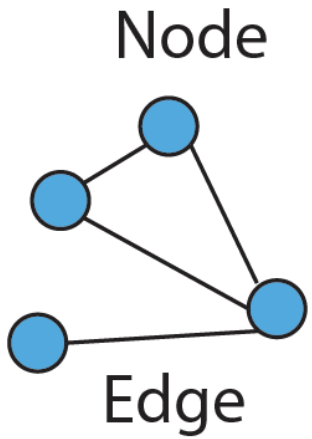
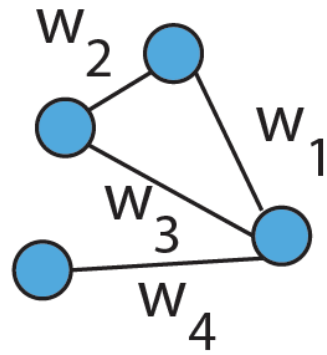


Figure: A Visual Comparison of PCA and tSNE

Community Detection

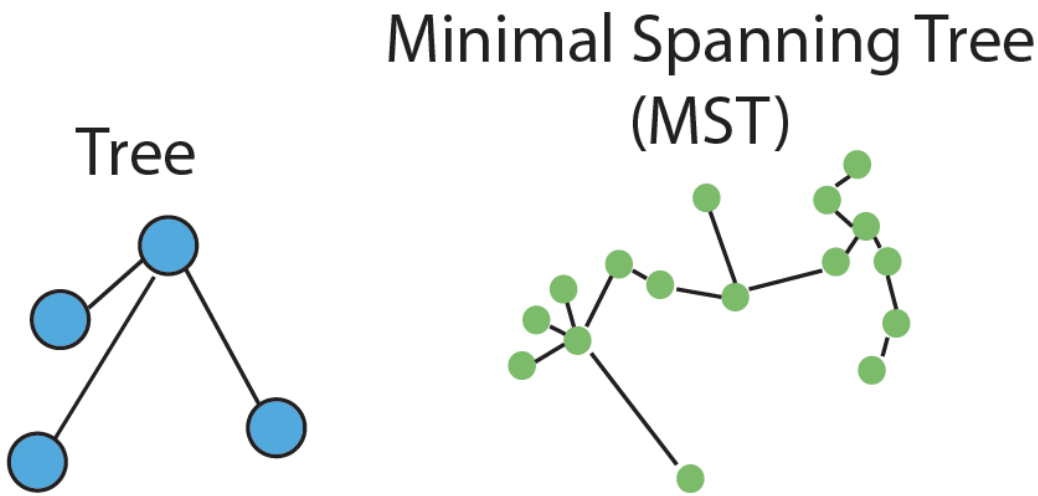
We can also add another layer of complexity to our clustered 2d data by using **graphs**.





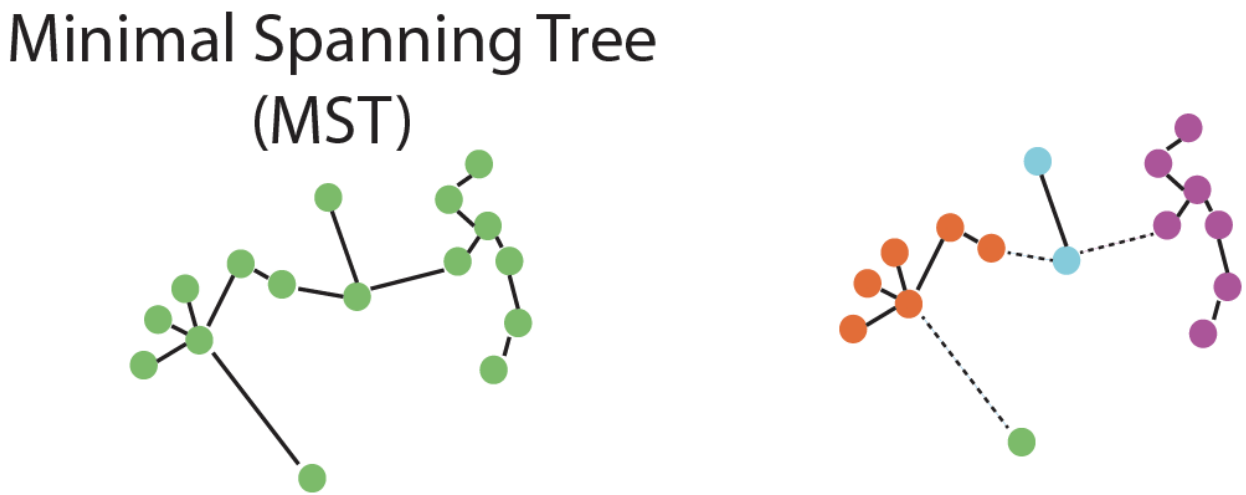
MST clustering

A tree is a graph that when adding an edge will form a cycle. A minimum spanning tree (MST) of a weighted graph is the minimum weight set or edges that connects all the vertices.



Clustering via MST

The advantage to use graph algorithms is speed which is more evident as the number of interactions increases.
Also, graphs are very intuitive as it is shown with the clustering using MST. Namely, deleting the longest K edges, we can cluster the data ($K = 3$ for the graph on the right)

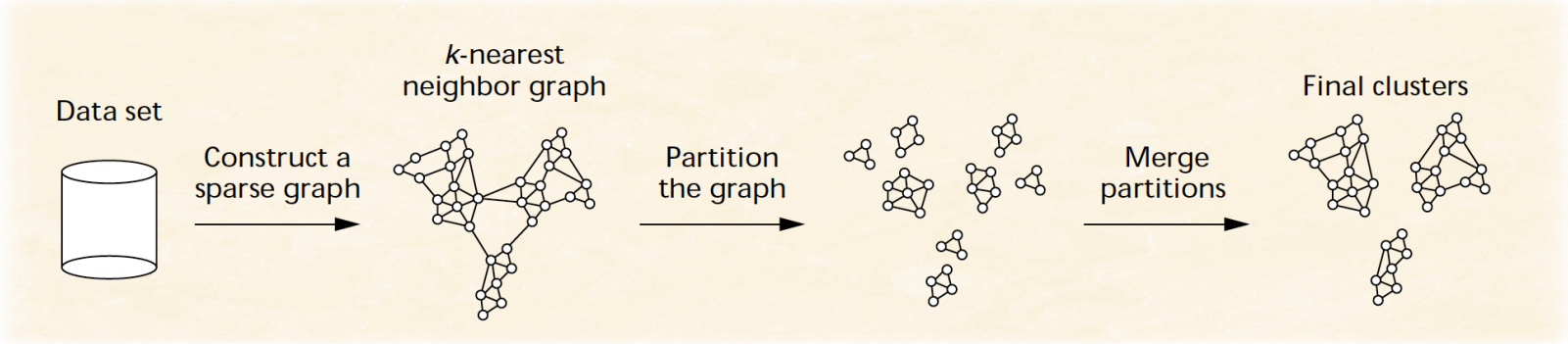


KNN graph

The K-nearest neighbors (KNN) graph is a graph which two vertices p and q are connected by an edge if the distance between p and q are among the Kth smallest distances.
A word of caution. KNN is a method commonly for classification. This means that we already know something about our data and we will classify an unknown sample. It is not the same as KNN graph.

Chameleon

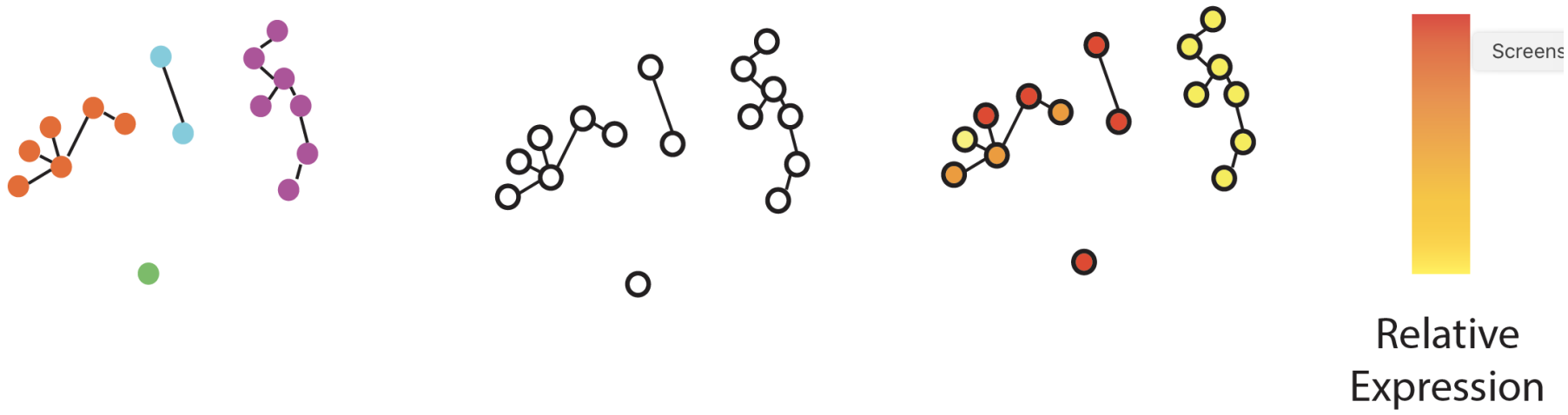
An important graphic method for clustering is called Chameleon



Chameleon uses a dynamic modeling framework to determine the similarity between pairs of clusters by looking at their relative interconnectivity (RI) and relative closeness (RC).

Cluster Annotation

Once we have done the clustering, we are ready to give **annotation** .



Cluster Annotation(cont)

"Clustered data are analysed by finding the gene signatures of each cluster. These so-called *marker genes* characterize the cluster and are used to annotate it with a meaningful biological label. "
One of the bigest warnings for using this approach is that the cluster do not represent cell types. That is why they are referred to as **cell identities** and not **cell types**.
External data should be used try to identify the expected expression profiles of individual cell identities.
From the statistical standpoint, one should use some permutation test (bootstrap) to determine the fidelity of the clustering step.

References

Materials and some of the pictures are from (1).

1. Gareth James et al. *An Introduction to Statistical Learning with applications in R*. Springer (2015)
2. Robert Sedgewick. *Algorithms in C* , 3rd edition. Addison-Wesley (2002) .
3. Ke-Lin Du et al. *Neural networks and Statistical Learning*, 2nd edition. Springer (2019)

I have used some of the graphs by hacking TiKz code from StakExchange, Inkscape for more aesthetic plots and other old tricks of TEX