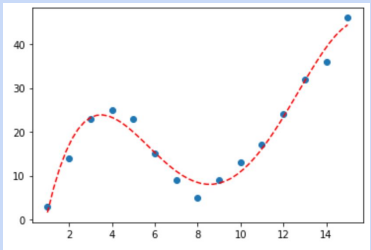
Unsupervised learning

Kasia Kedzierska https://kasia.codes/
Kaspar Märtens https://kaspar.website/

UNIVERSITY OF OXFORD

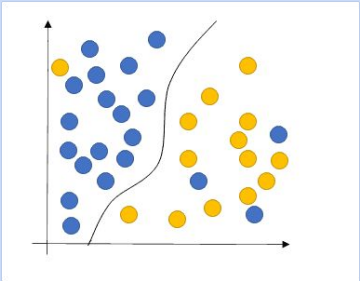
Supervised learning

Regression



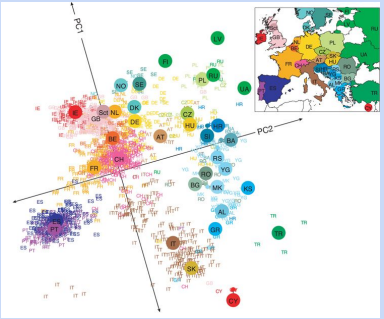
Classification

$$y = f(x)$$

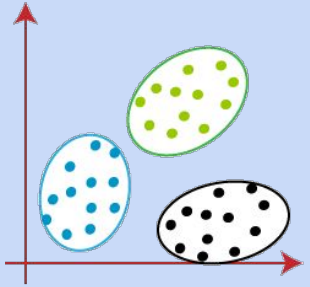


Unsupervised learning

Dimensionality reduction



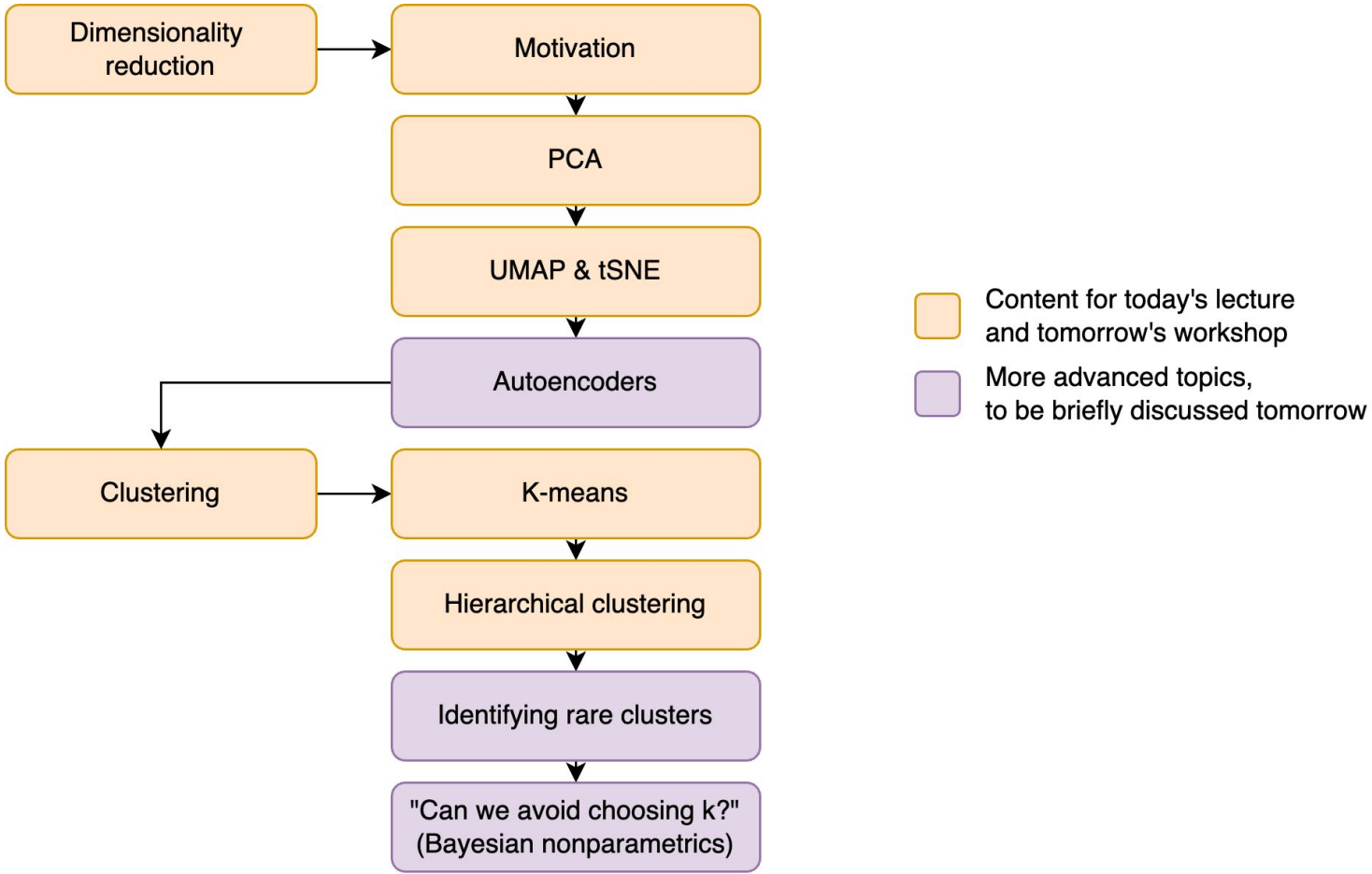
Clustering



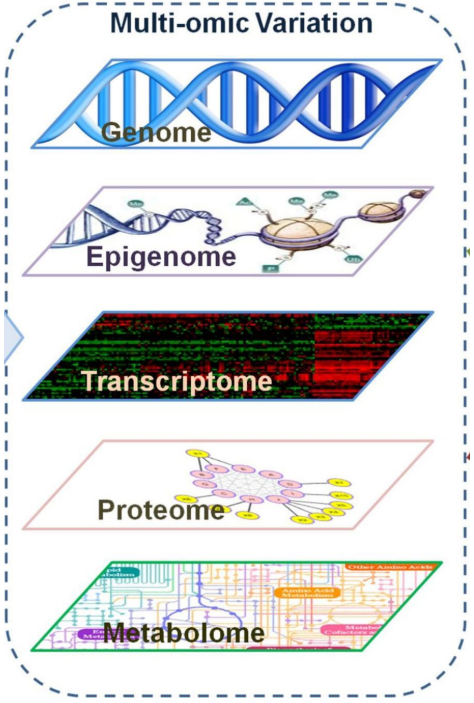
Sidenote: there also exist

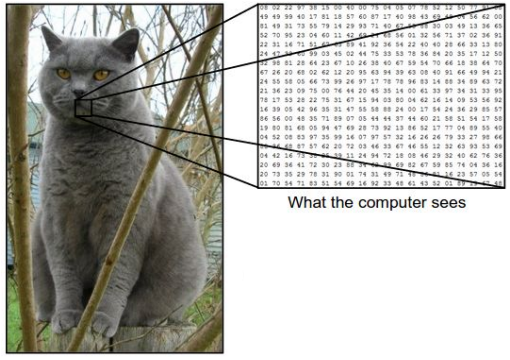
- Semi-supervised learning
- Self-supervised learning

Outline for this tutorial



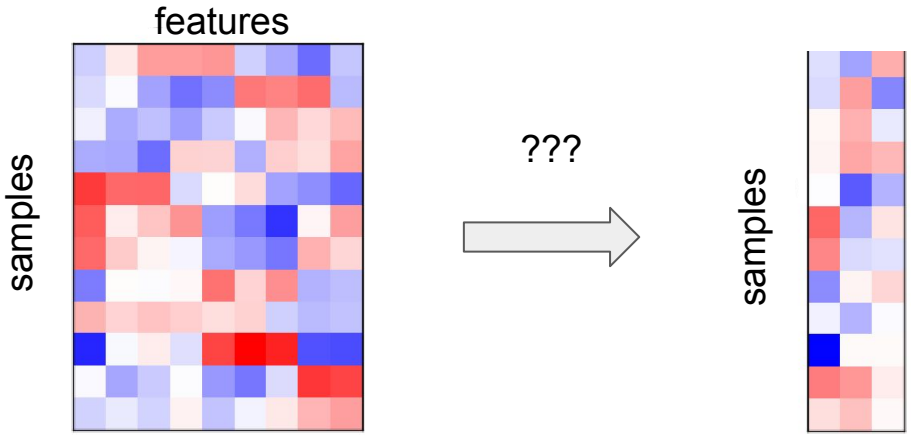
Examples of high-dimensional data sets





Dimensionality reduction

How to **summarise** a *large* number of *correlated* variables with a *small* number of features?



When is dimensionality reduction most meaningful?

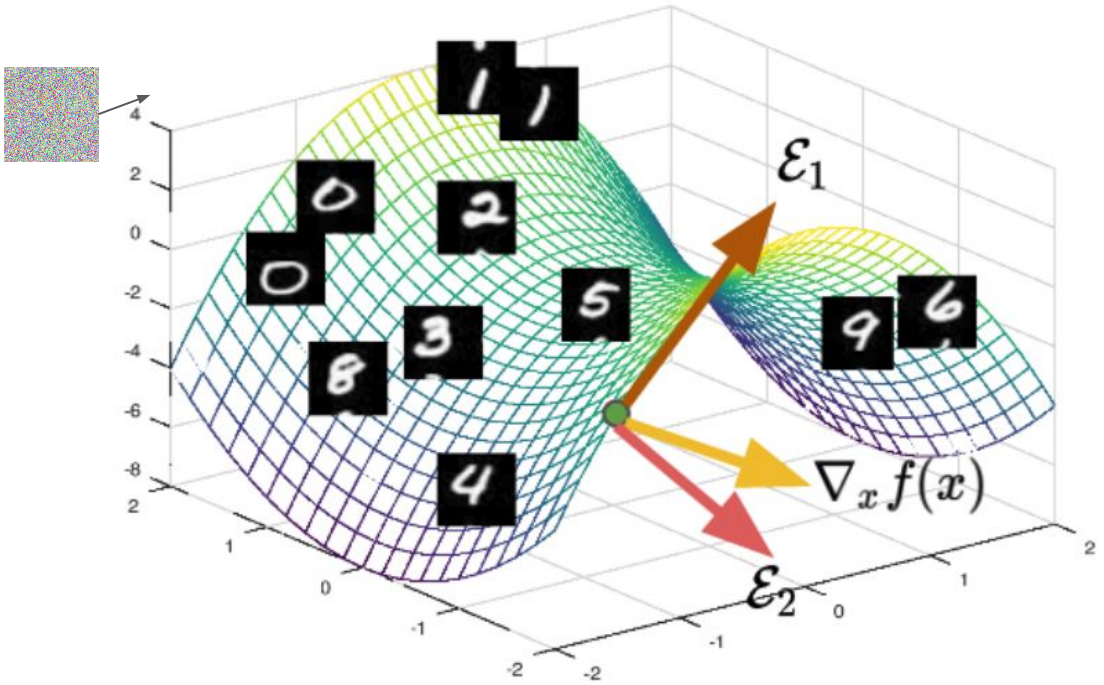


Fig from https://arxiv.org/abs/2206.07387

Example: Questionnaire data

National Cancer Patient Experience Survey

15. Before you started your treatment(s), were you also told about any side effects of the treatment that could affect you in the future rather than straight away?

1. ☐ Yes, definitely
2. ☐ Yes, to some extent
3. ☐ No, future side effects were not explained
4. ☐ I did not need an explanation
5. ☐ Don't know / can't remember

54. Did the different people treating and caring for you (such as GP, hospital doctors, hospital nurses, specialist nurses, community nurses) work well together to give you the best possible care?

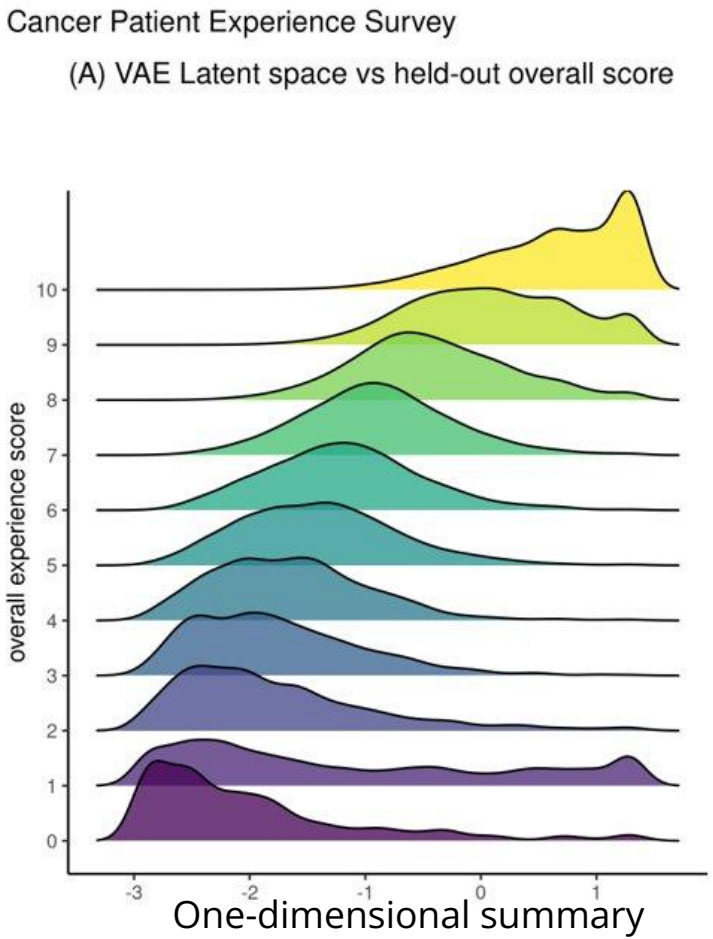
1. ☐ Yes, always
2. ☐ Yes, most of the time
3. ☐ Yes, some of the time
4. ☐ No, never
5. ☐ Don't know / can't remember

59. Overall, how would you rate your care?
(**Please circle a number**)

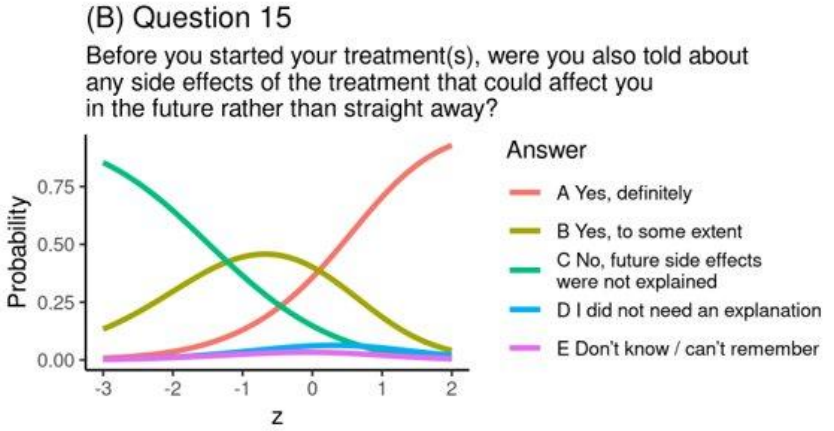
Very poor — Very good

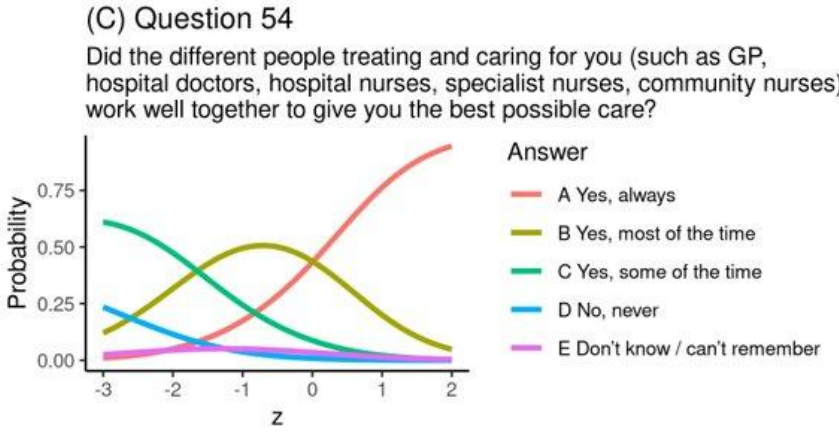
0 1 2 3 4 5 6 7 8 9 10

Example: Questionnaire data

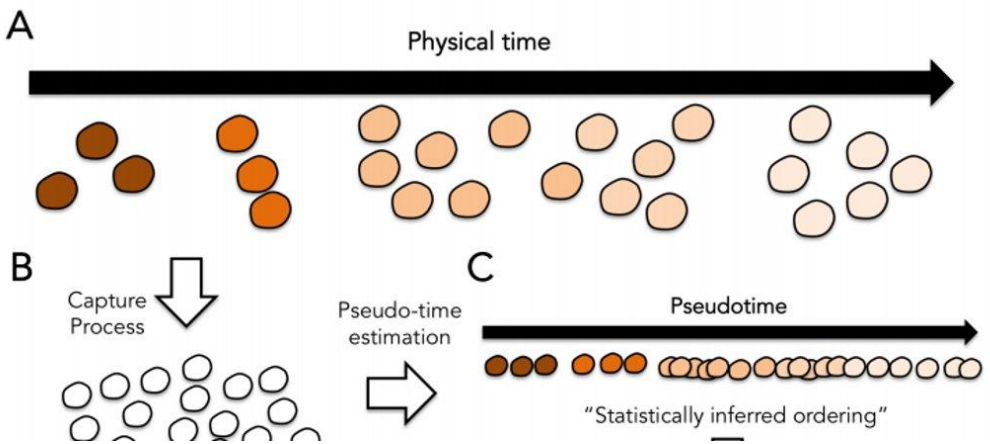


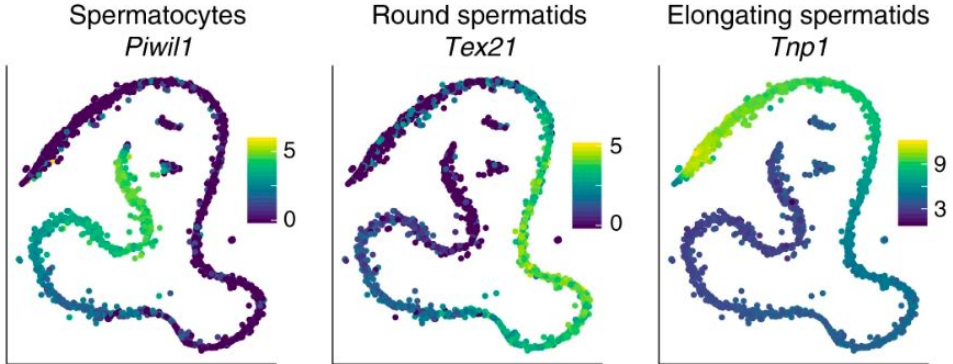
One-dimensional summary





Example: Pseudotime in single-cell data





Quiz time:

Please go to http://bit.ly/ngschool_quiz

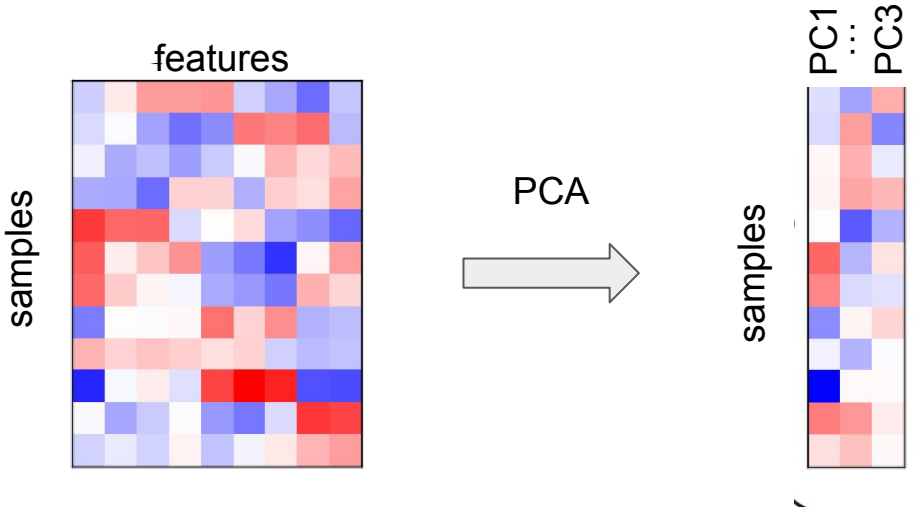
and enter room id “NGSCHOOL”

Principal Component Analysis (PCA) for dimensionality reduction

Principal Component Analysis (PCA) for dimensionality reduction

How to **summarise** a *large* number of *correlated* variables with a *small* number of features?

PCA is one solution: We can use the **top principal components** as a low-dimensional representation of the data



Principal Component Analysis (PCA)

How to **summarise** a *large* number of *correlated* variables with a *small* number of features?

PCA is a method to do this by

- Constructing **linear combinations** of original features
- Aiming to **maximise variance**

Principal Component Analysis (PCA)

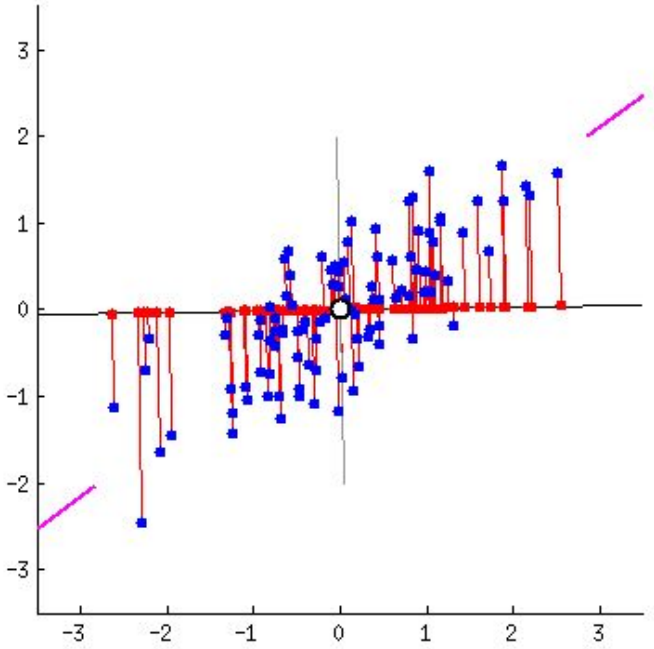
How to **summarise** a *large* number of *correlated* variables with a *small* number of features?

PCA is a method to do this by

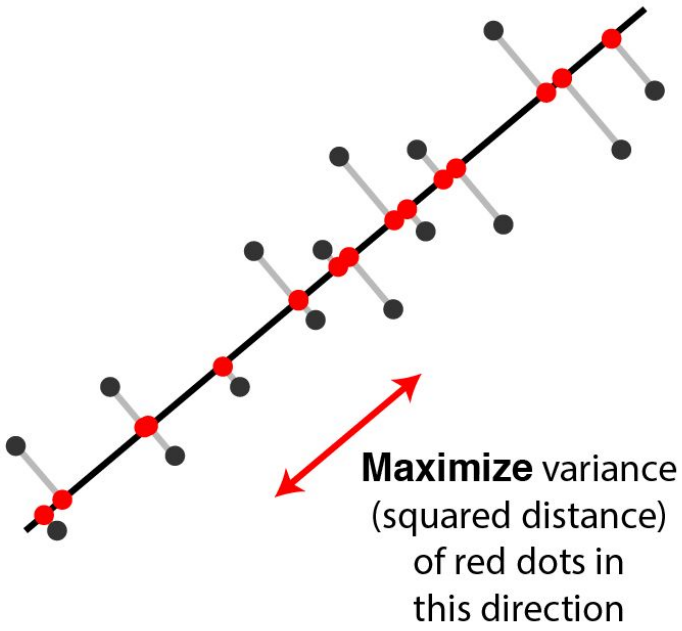
- Constructing **linear combinations** of original features
- Aiming to **maximise variance**

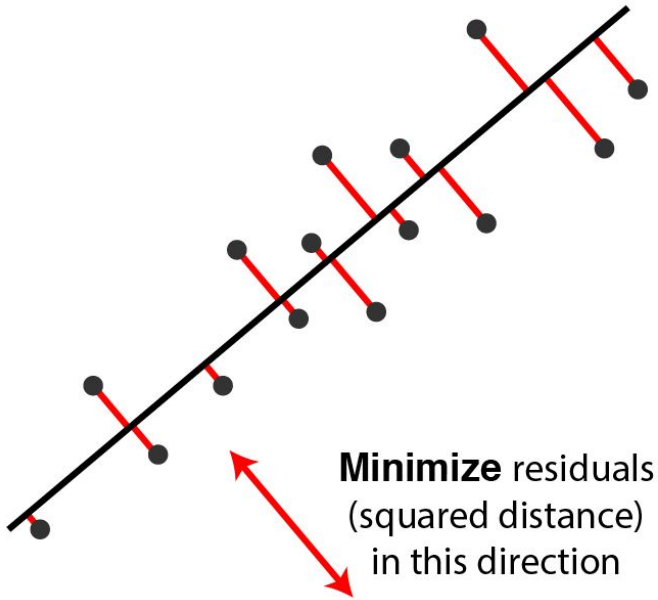
Turns out that **PCA** finds linear combinations that **best reconstruct original features**.

PCA: Which direction maximises variance?



Maximising variance is equivalent to minimising reconstruction error





Principal Component Analysis (PCA)

How to **summarise** a *large* number of *correlated* variables with a *small* number of features?

PCA is a method to do this by

- Constructing **linear combinations** of original features
- Aiming to **maximise variance**

Turns out that PCA finds linear combinations that best reconstruct original features.

Sidenote 1: PCA can be implemented e.g. via Singular Value Decomposition (SVD).

Principal Component Analysis (PCA)

How to **summarise** a *large* number of *correlated* variables with a *small* number of features?

PCA is a method to do this by

- Constructing **linear combinations** of original features
- Aiming to **maximise variance**

Turns out that PCA finds linear combinations that best reconstruct original features.

Sidenote 1: PCA can be implemented e.g. via Singular Value Decomposition (SVD).

Sidenote 2: This is a great explanation of PCA (ELI5 PCA): https://stats.stackexchange.com/a/140579

PCA

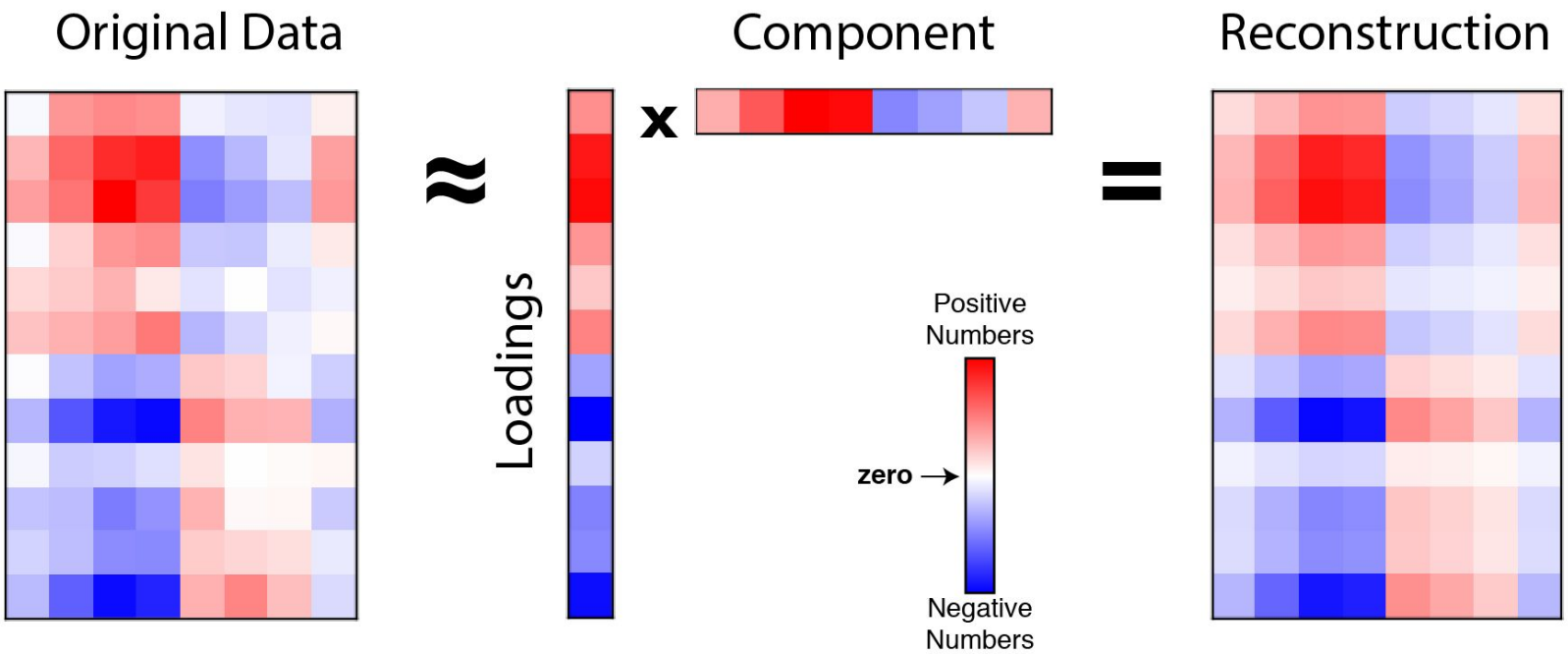


Figure from http://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/

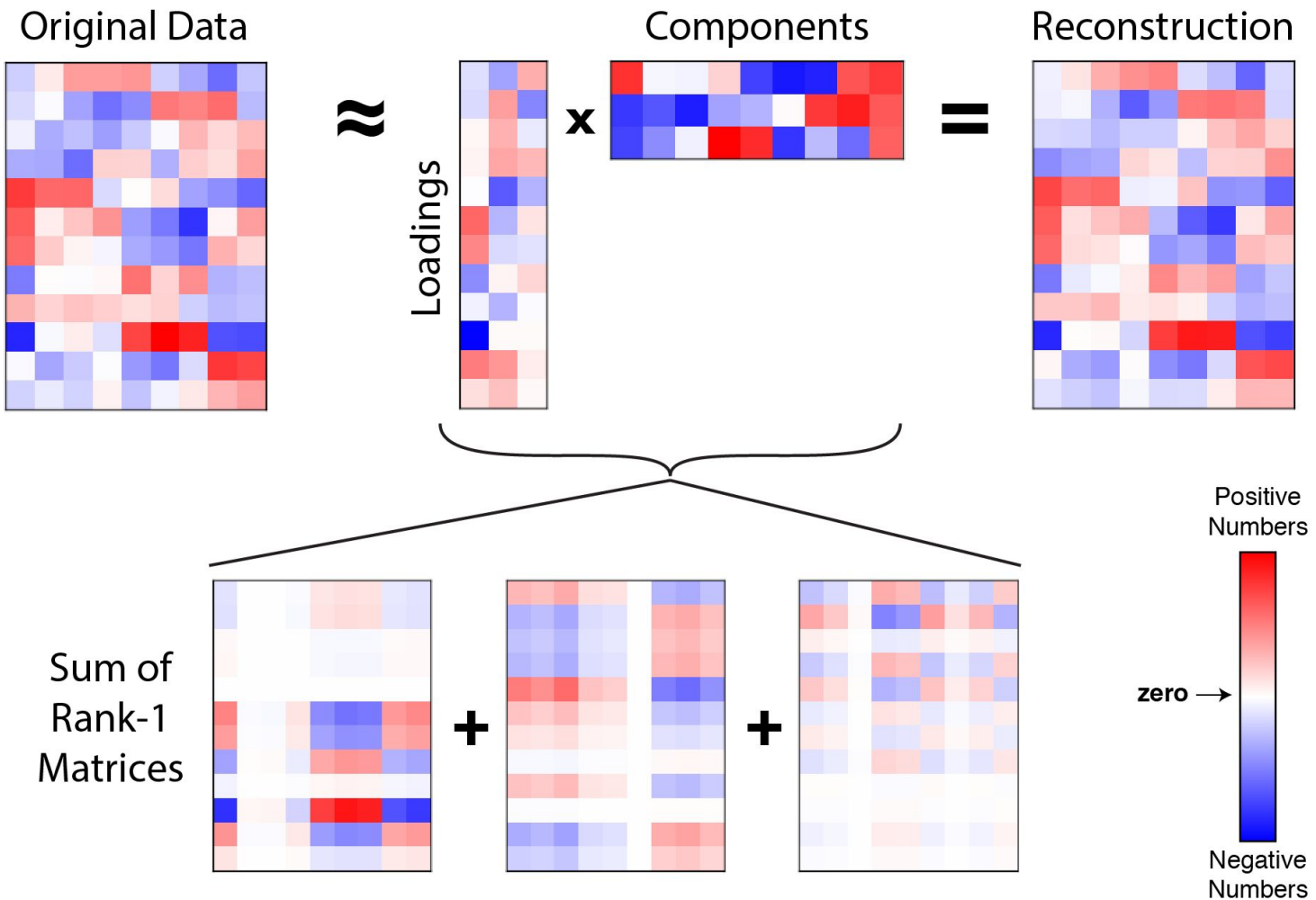


Figure from http://alexhwilliams.info/itsneuronalblog/2016/03/27/pca/

PCA reconstruction can be easily visualised on image data

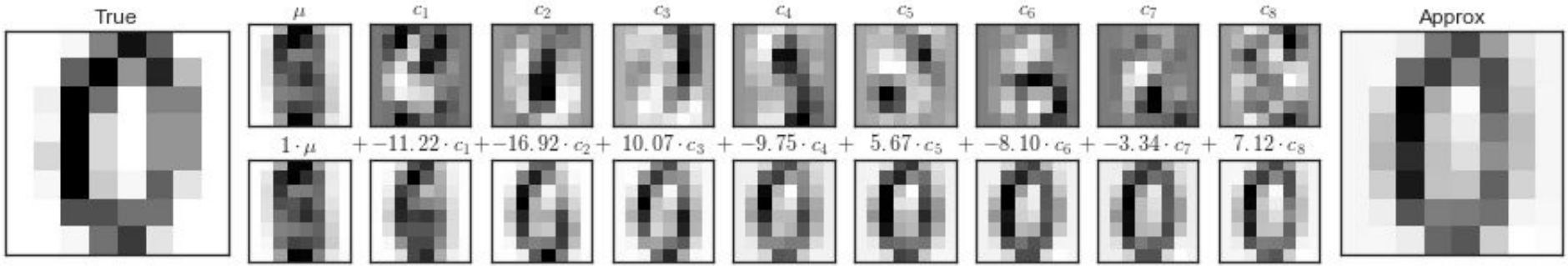
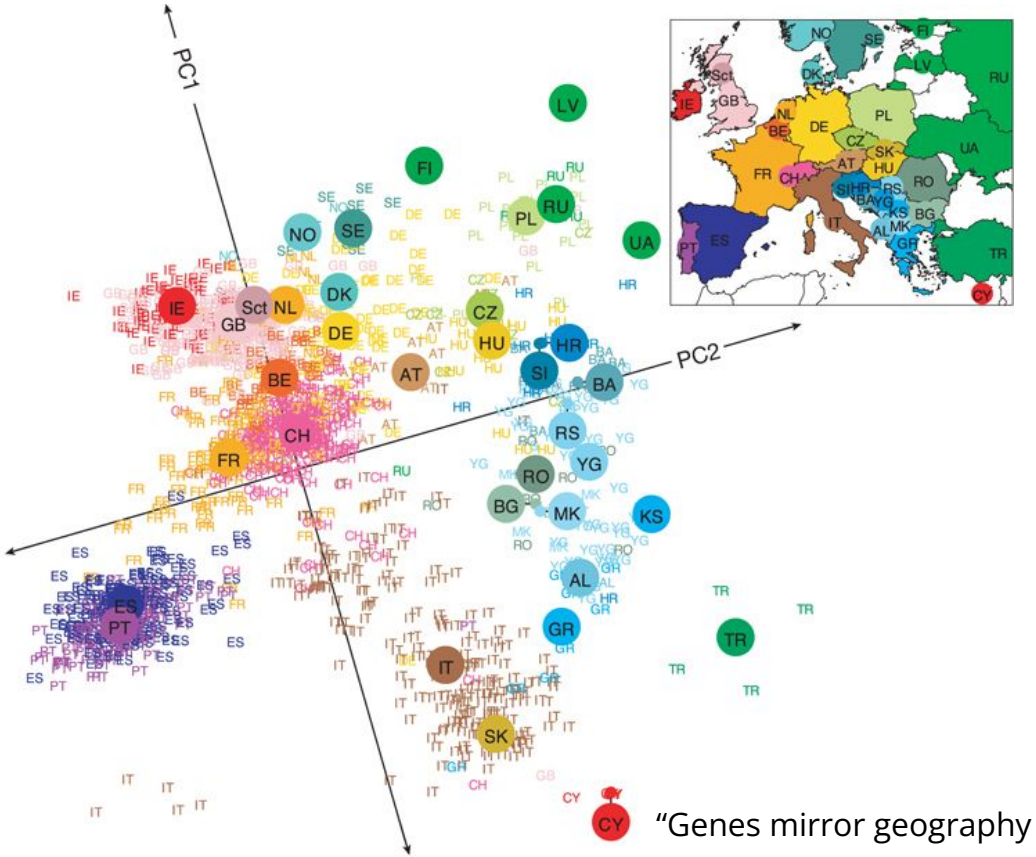


Figure from https://jakevdp.github.io/PythonDataScienceHandbook

PCA can discover meaningful structure in the data: Population genetics example

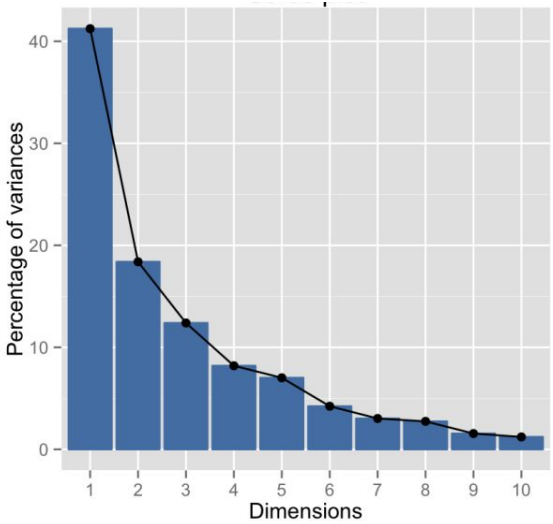


"Genes mirror geography within Europe" Novembre et al 2008

Any good ways to choose the number of PCs to keep?

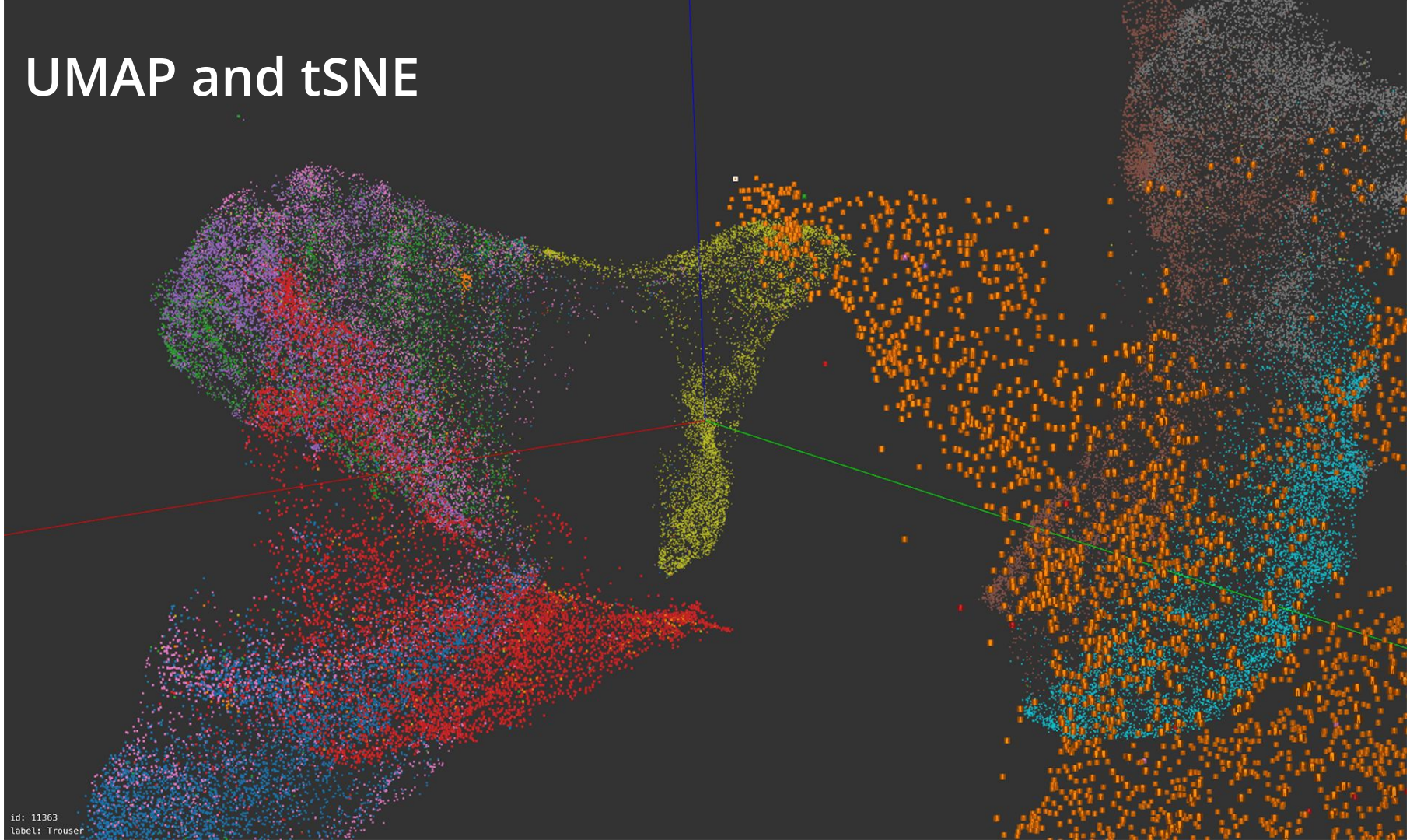
No 😔

- For visualisation purposes, we often use the top 2 (or 3) PCs
- We can set a desired threshold for "*total percentage-of-variance explained*"
- Choose an "elbow point" on the *percentage-of-variance-explained* plot

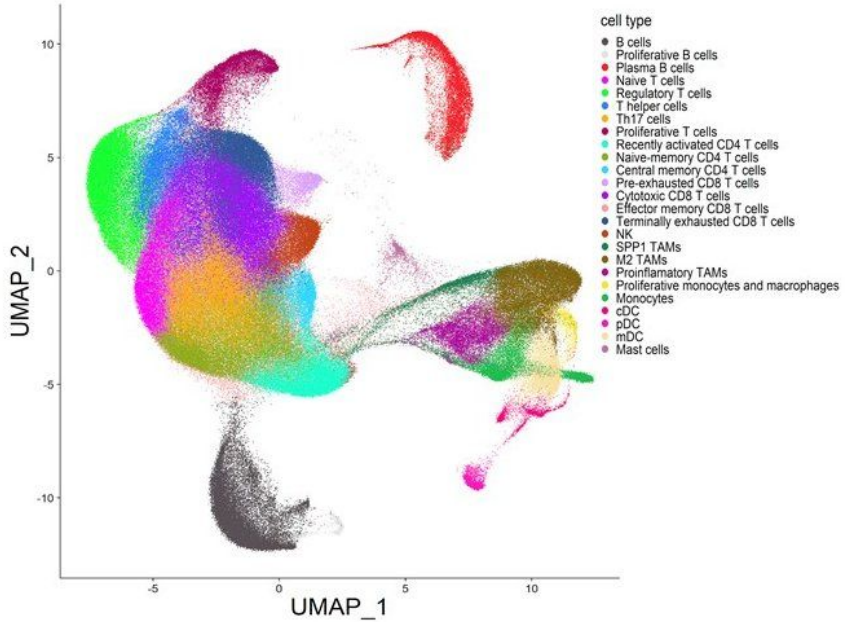


How to run PCA in R

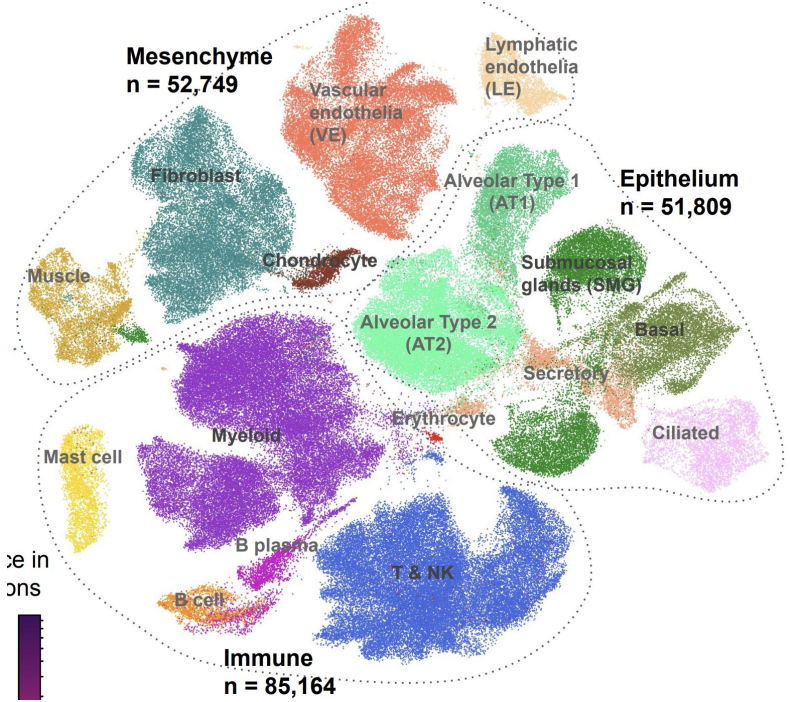
UMAP and tSNE



UMAP & tSNE:
especially common for visualising single-cell RNA-seq data

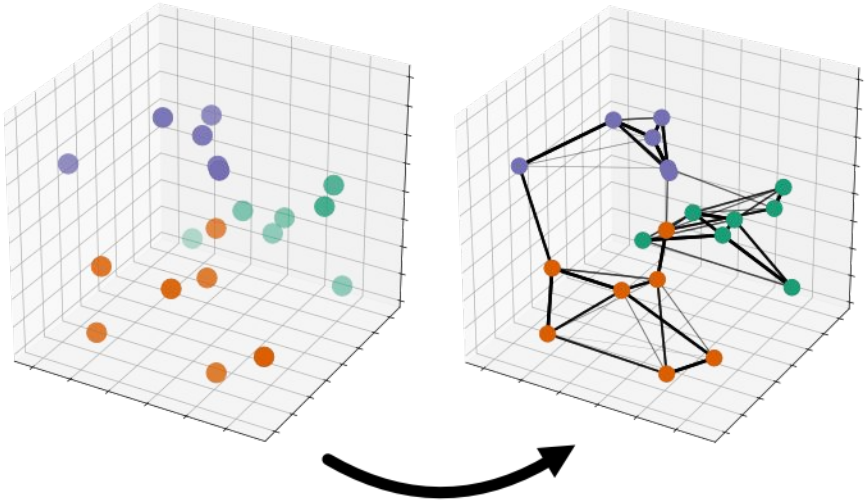


"Single-Cell Tumor Immune Atlas for Precision Oncology", Nieto et al 2021

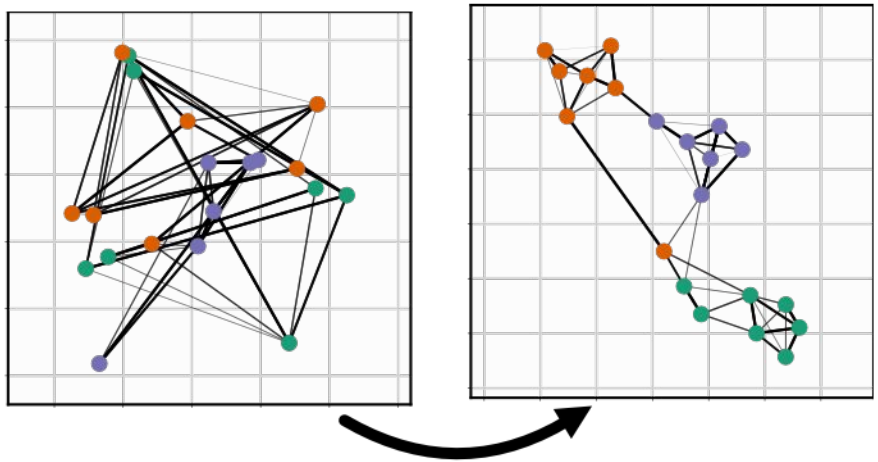


"An integrated cell atlas of the human lung in health and disease", Sikkema et al 2022

UMAP and tSNE



Compute a graphical representation of the dataset

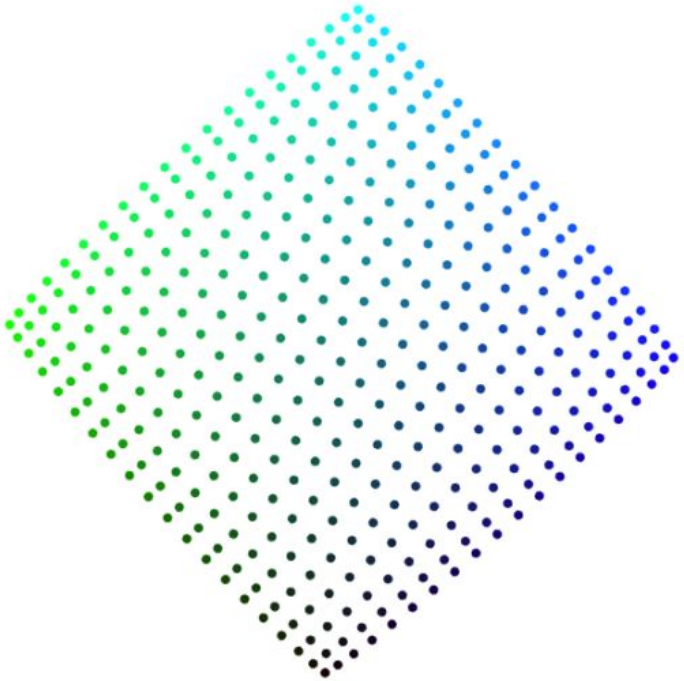


Learn an embedding that preserves the structure of the graph

How UMAP works: https://umap-learn.readthedocs.io/en/latest/how_umap_works.html

How to Use t-SNE Effectively

Although extremely useful for visualizing high-dimensional data, t-SNE plots can sometimes be mysterious or misleading. By exploring how it behaves in simple cases, we can learn to use it more effectively.



Step
1,940

A square grid with equal spacing between points. Try convergence at different sizes.

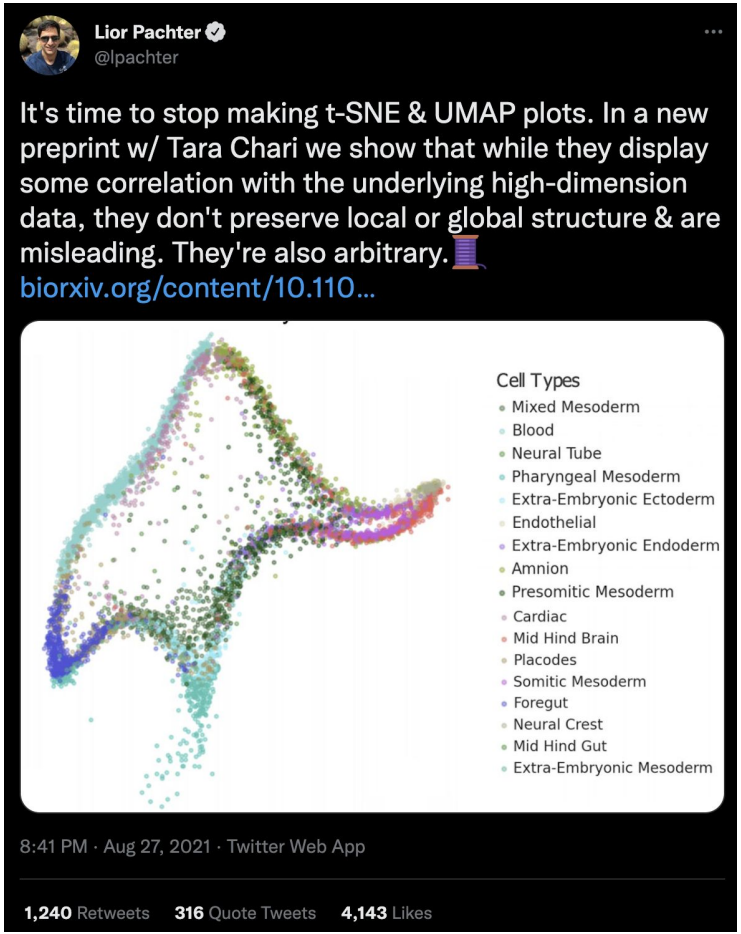
Points Per Side 20

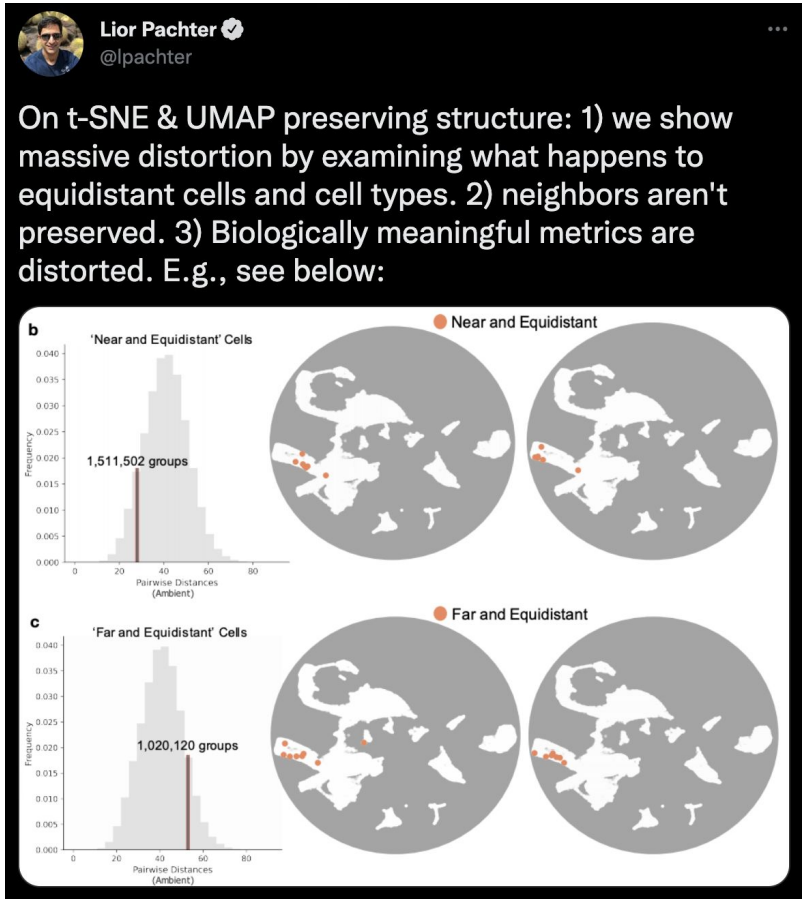
Perplexity 10

Epsilon 5

https://distill.pub/2016/misread-tsne/

There is ongoing debate on how well UMAP & tSNE preserve structure

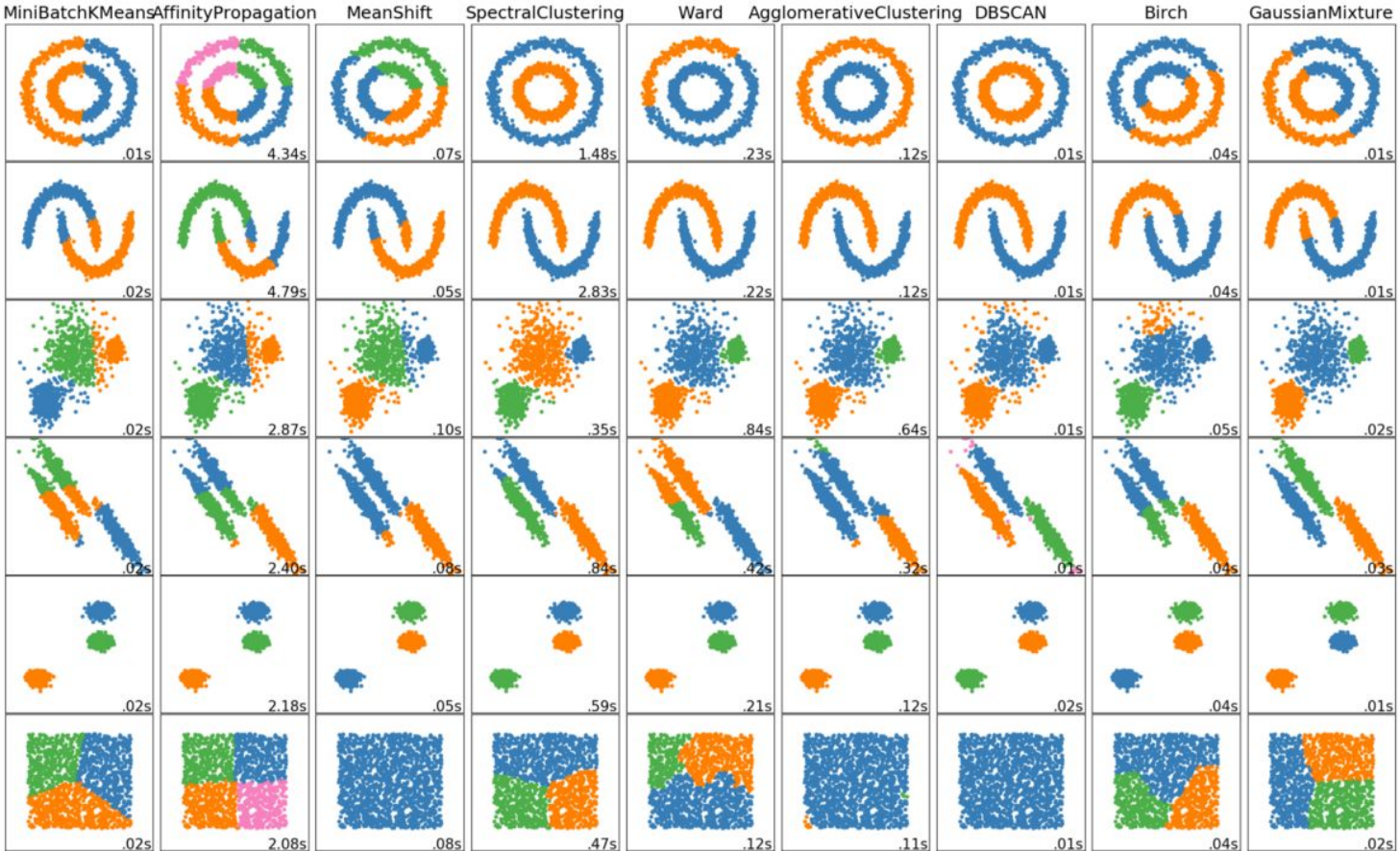




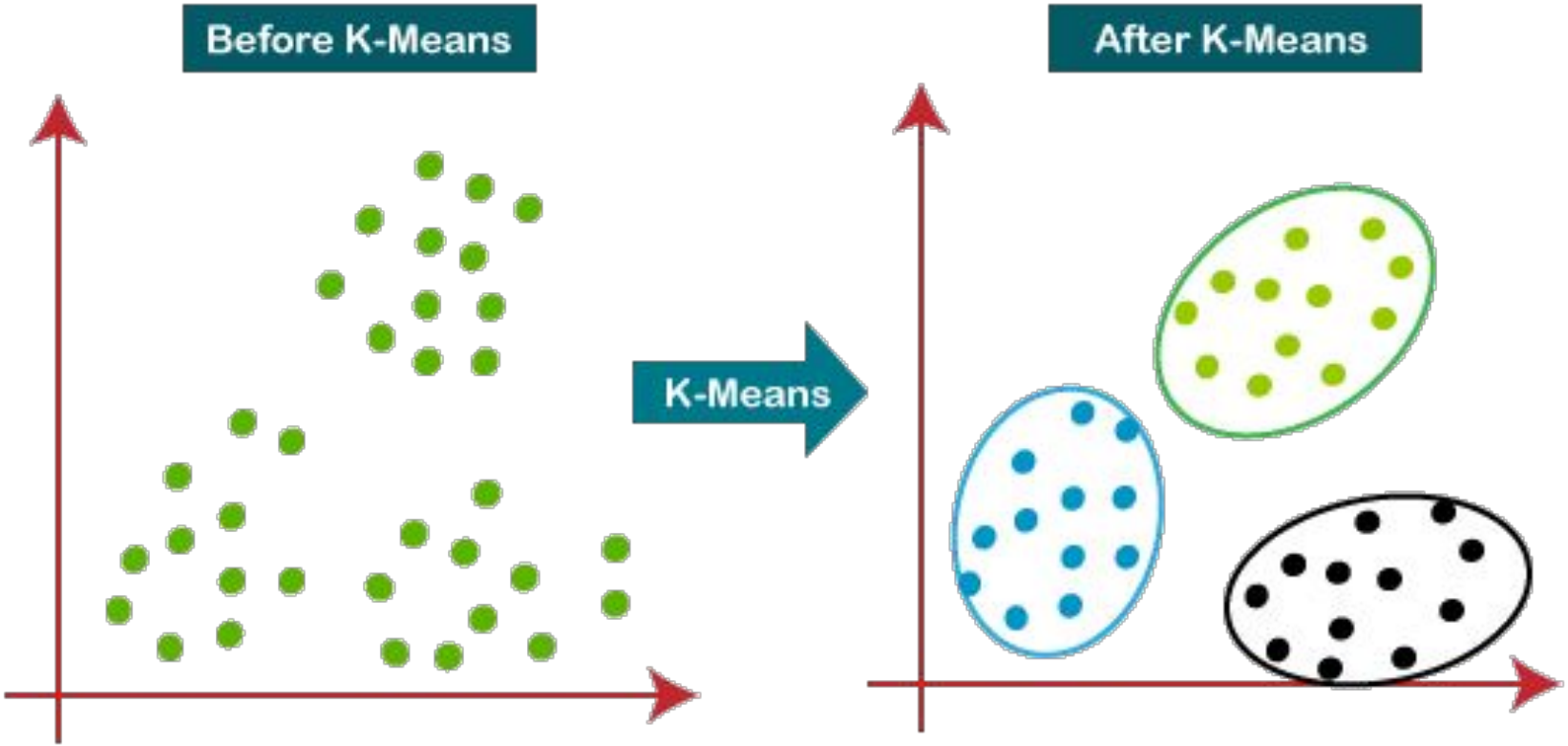
https://www.biorxiv.org/content/10.1101/2021.08.25.457696v1.full

How to run UMAP in R

Clustering


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
Birch
GaussianMixture
.01s
4.34s
.07s
1.48s
.23s
.12s
.01s
.04s
.01s
.02s
4.79s
.05s
2.83s
.22s
.12s
.01s
.04s
.01s
.02s
2.87s
.10s
.35s
.84s
.64s
.01s
.05s
.02s
.02s
2.40s
.08s
.84s
.42s
.32s
.01s
.04s
.03s
.02s
2.18s
.05s
.59s
.21s
.12s
.02s
.04s
.01s
.02s
2.08s
.08s
.47s
.12s
.11s
.01s
.04s
.02s

K-means

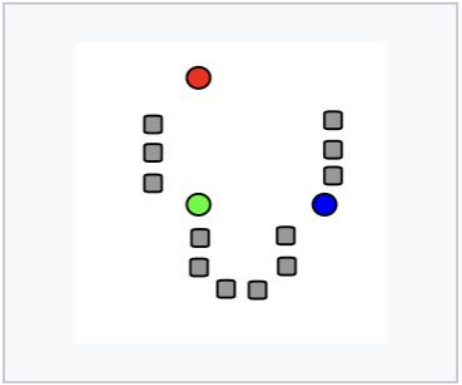


Algorithm for k-means

- Initialise cluster centroids
- Repeat until convergence:
  - Assign each point to its closest centroid
  - Update the cluster centroids

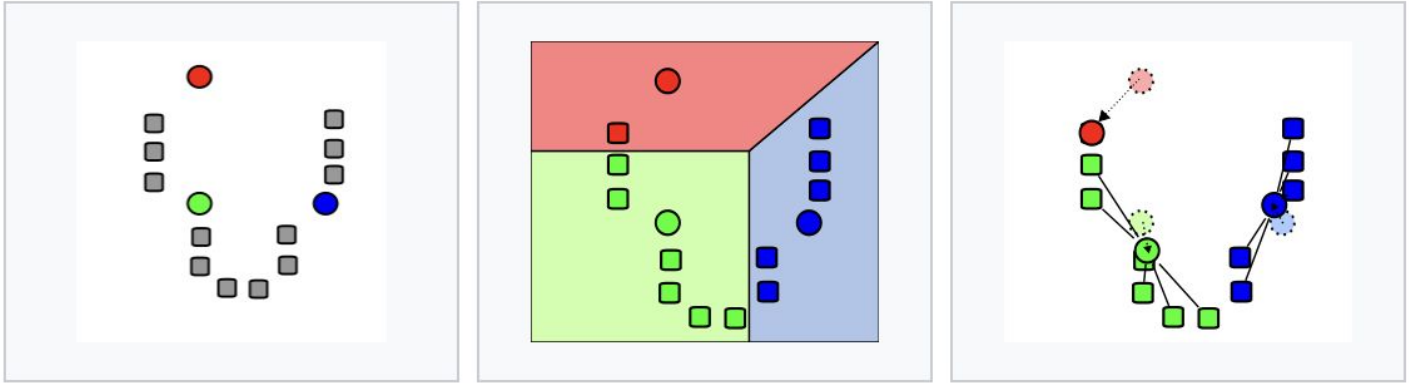
Algorithm for k-means

- Initialise cluster centroids
- Repeat until convergence:
  - Assign each point to its closest centroid
  - Update the cluster centroids



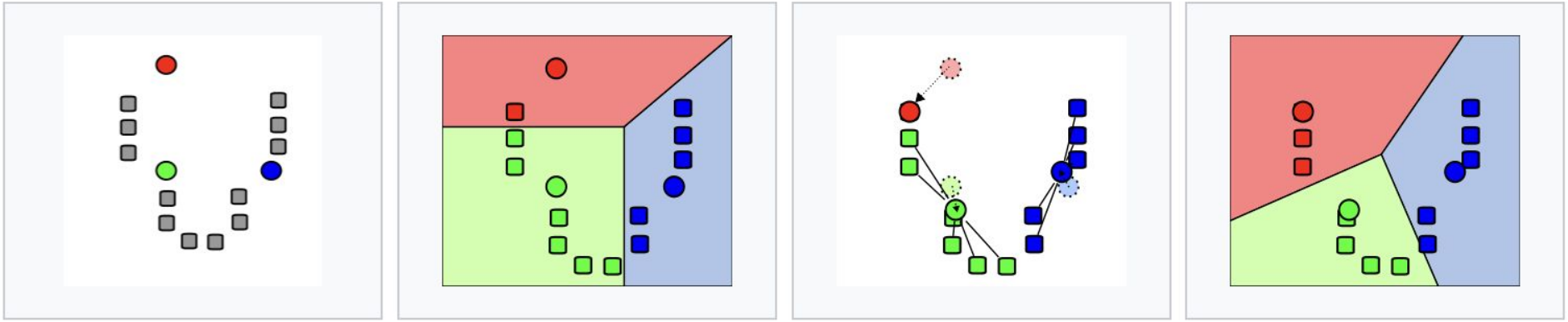
Algorithm for k-means

- Initialise cluster centroids
- Repeat until convergence:
  - Assign each point to its closest centroid
  - Update the cluster centroids



Algorithm for k-means

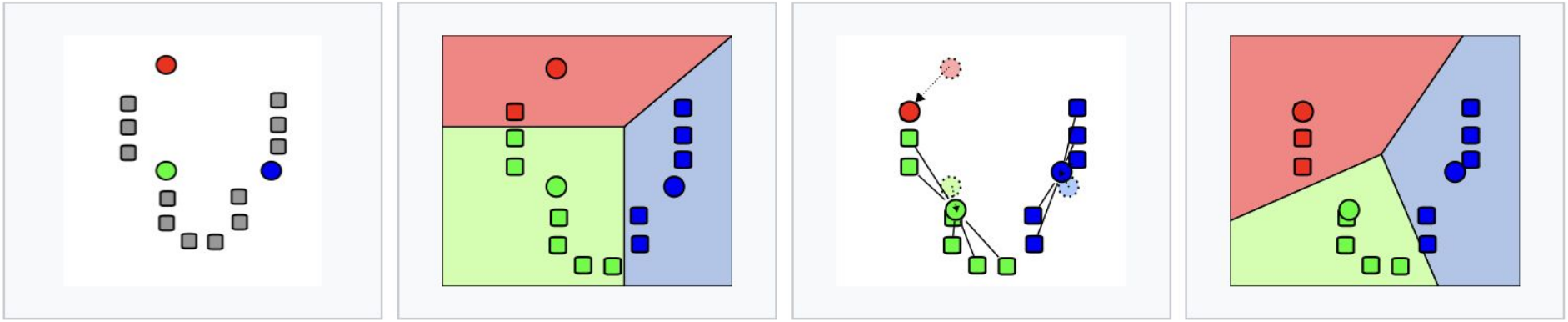
- Initialise cluster centroids
- Repeat until convergence:
  - Assign each point to its closest centroid
  - Update the cluster centroids



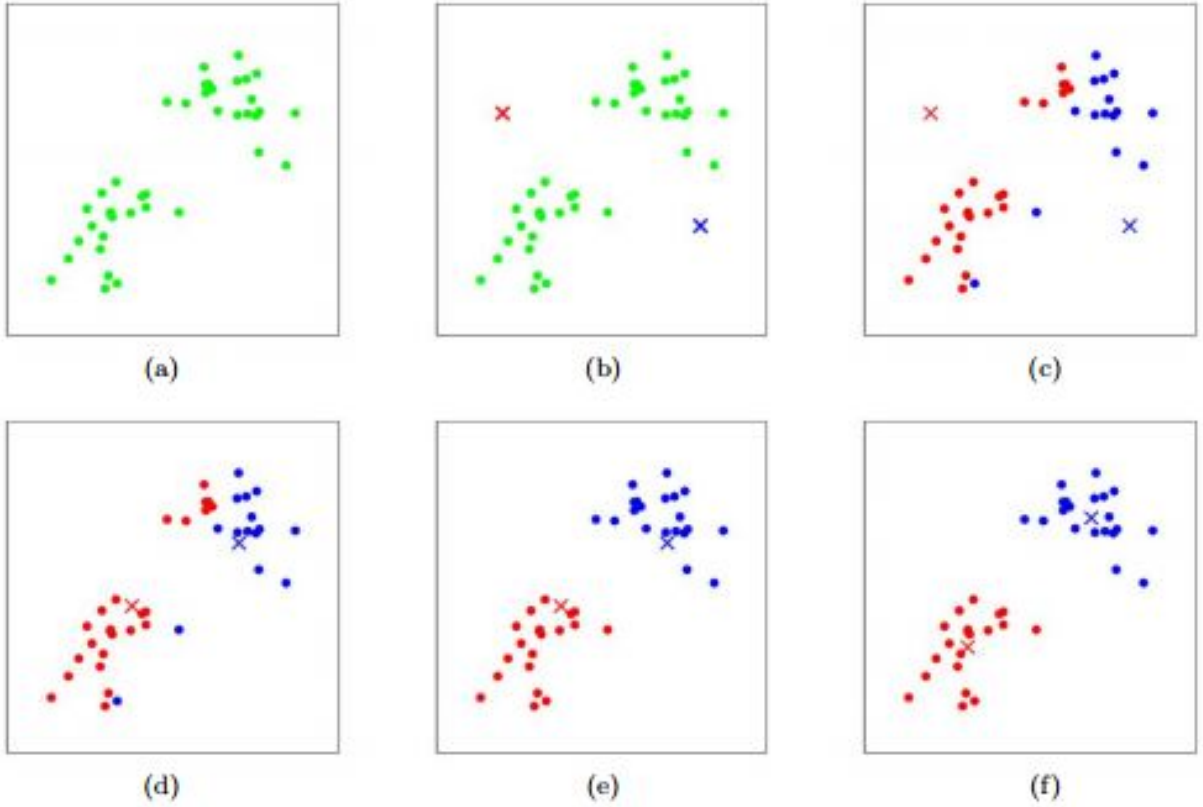
Algorithm for k-means

This solves an optimisation problem:
"**Minimise within-cluster sum-of-squares**"

- Initialise cluster centroids
- Repeat until convergence:
  - Assign each point to its closest centroid
  - Update the cluster centroids



K-means: another 2D example



🤔 Can you think of any failure modes of k-means?

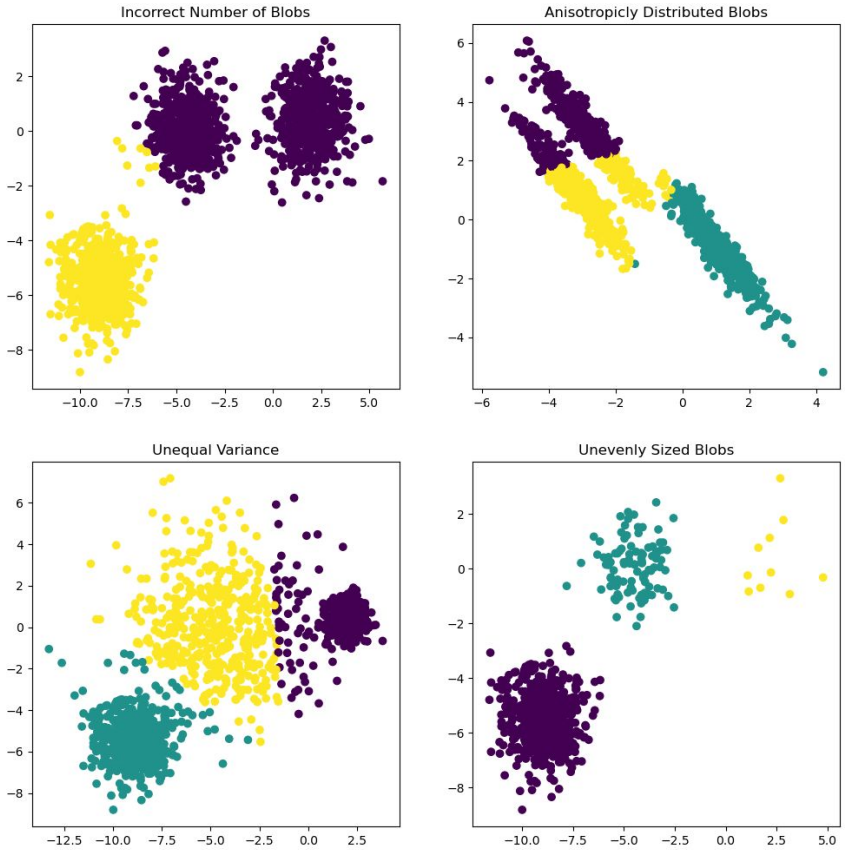


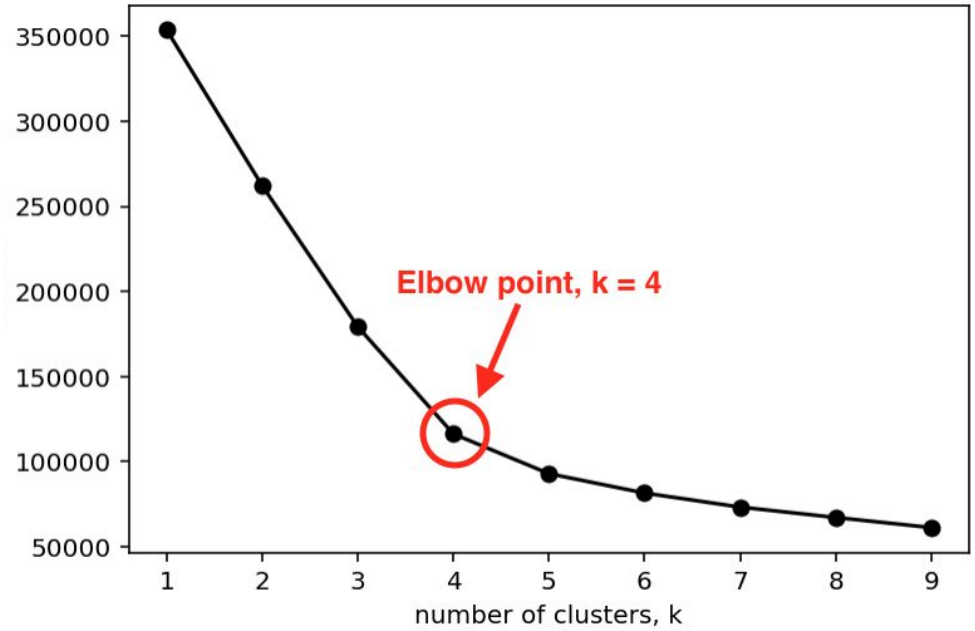
Figure from
http://scikit-learn.org/stable/modules/clustering.html

Any good way to choose the number of clusters *k*?

Any good way to choose the number of clusters *k*?

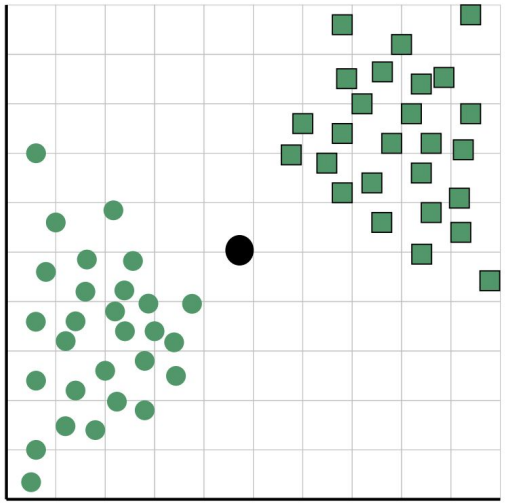
No 😔

A common heuristic is to plot the within-cluster sum-of-squares for varying k and find the “elbow point”:



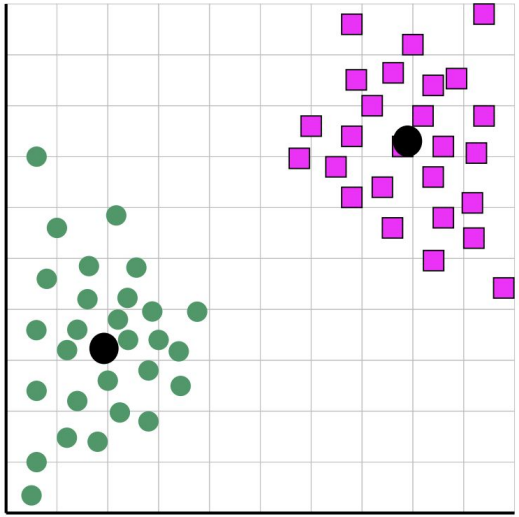
Intuition for the “elbow point”

k=1



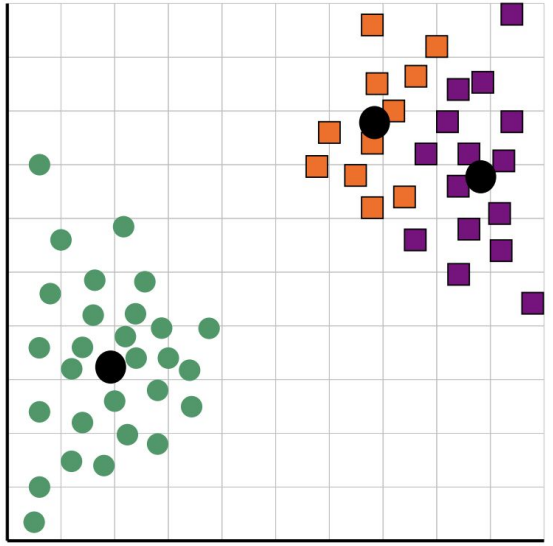
sum-of-squares=873

k=2



sum-of-squares=173

k=3



sum-of-squares=133

Figure from http://www.cs.cmu.edu/~tom/10601_fall2012/slides/clustering.pdf

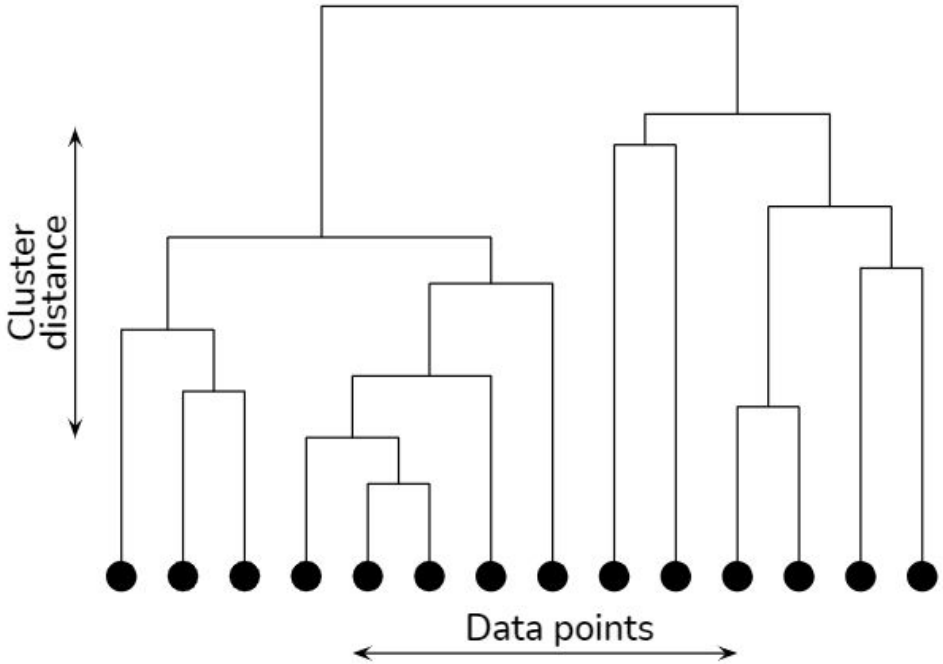
K-means clustering can be used for image segmentation

How to fit k-means in R

Hierarchical clustering

Involves construction of a dendrogram

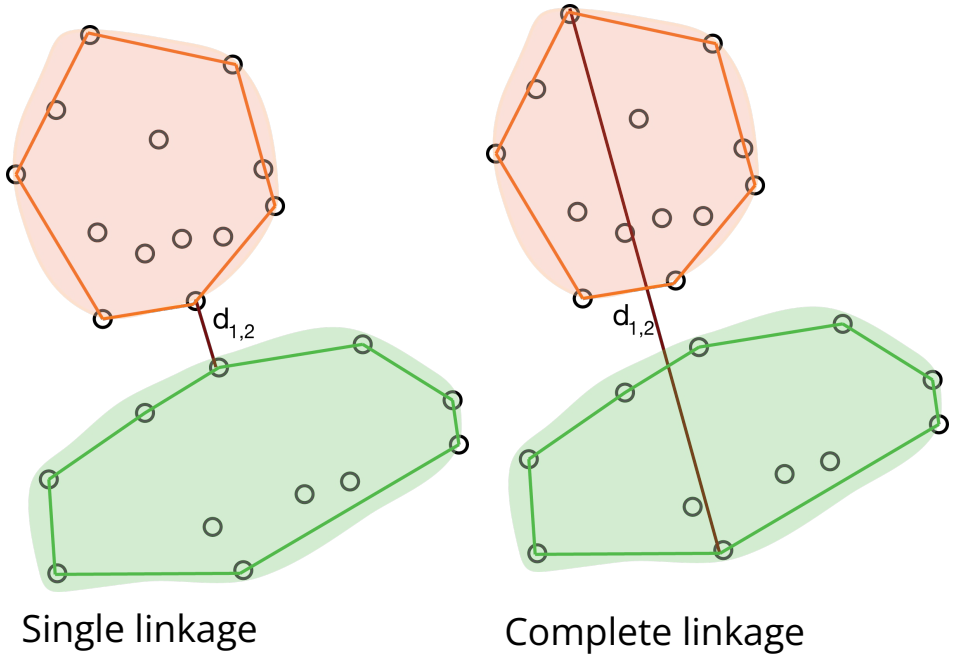
- Requires choice of:
  - Distance metric
  - A linkage criterion
    how to measure dissimilarity of points in a cluster



Hierarchical clustering

Involves construction of a dendrogram

- Requires choice of:
  - Distance metric
  - A linkage criterion
    how to measure dissimilarity of points in a cluster

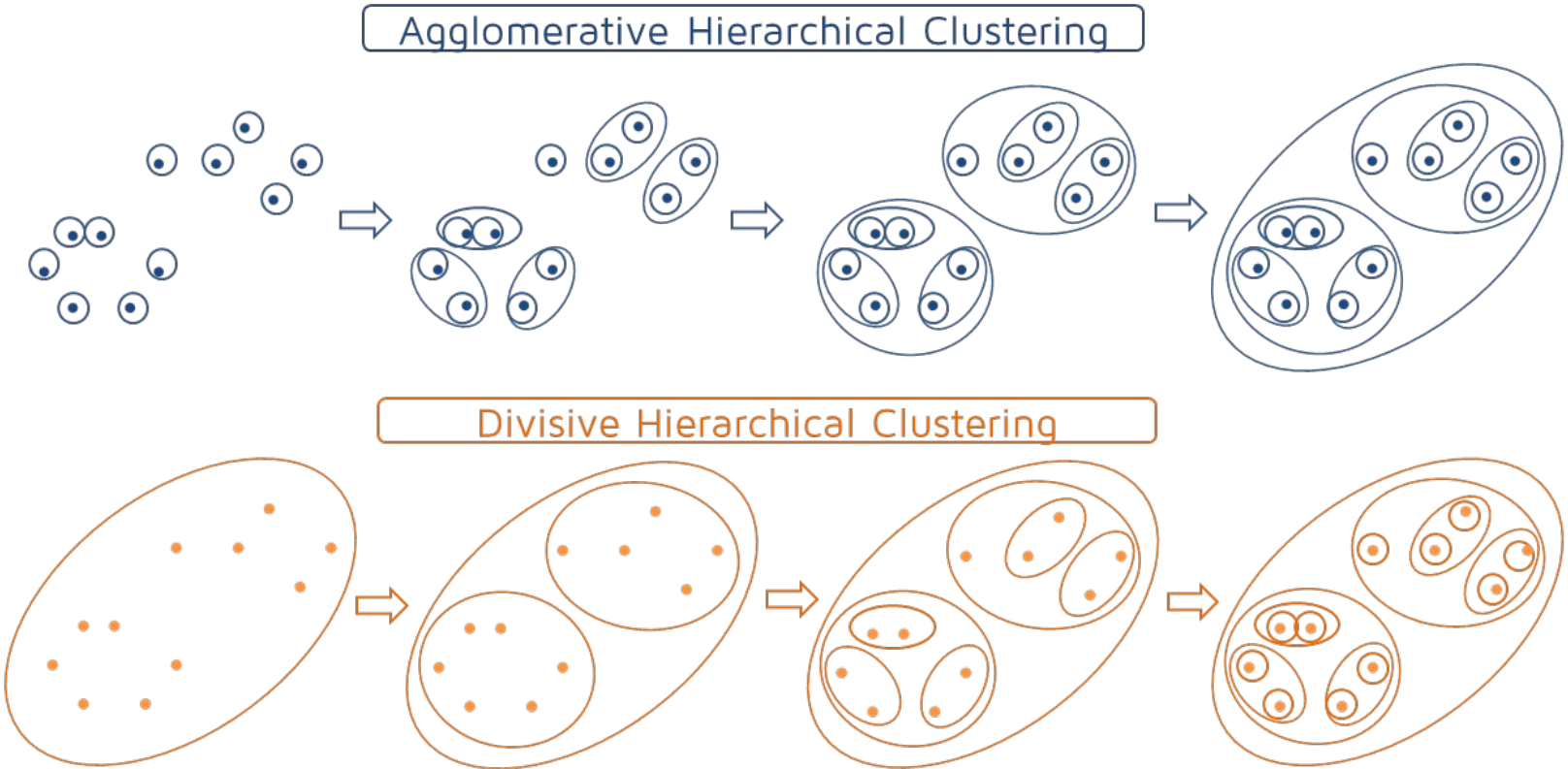


Single linkage

Complete linkage

Figure from https://web.stanford.edu/class/bios221/book/Chap-Clustering.html

Hierarchical clustering: agglomerative or divisive



Obtaining clusters from a dendrogram

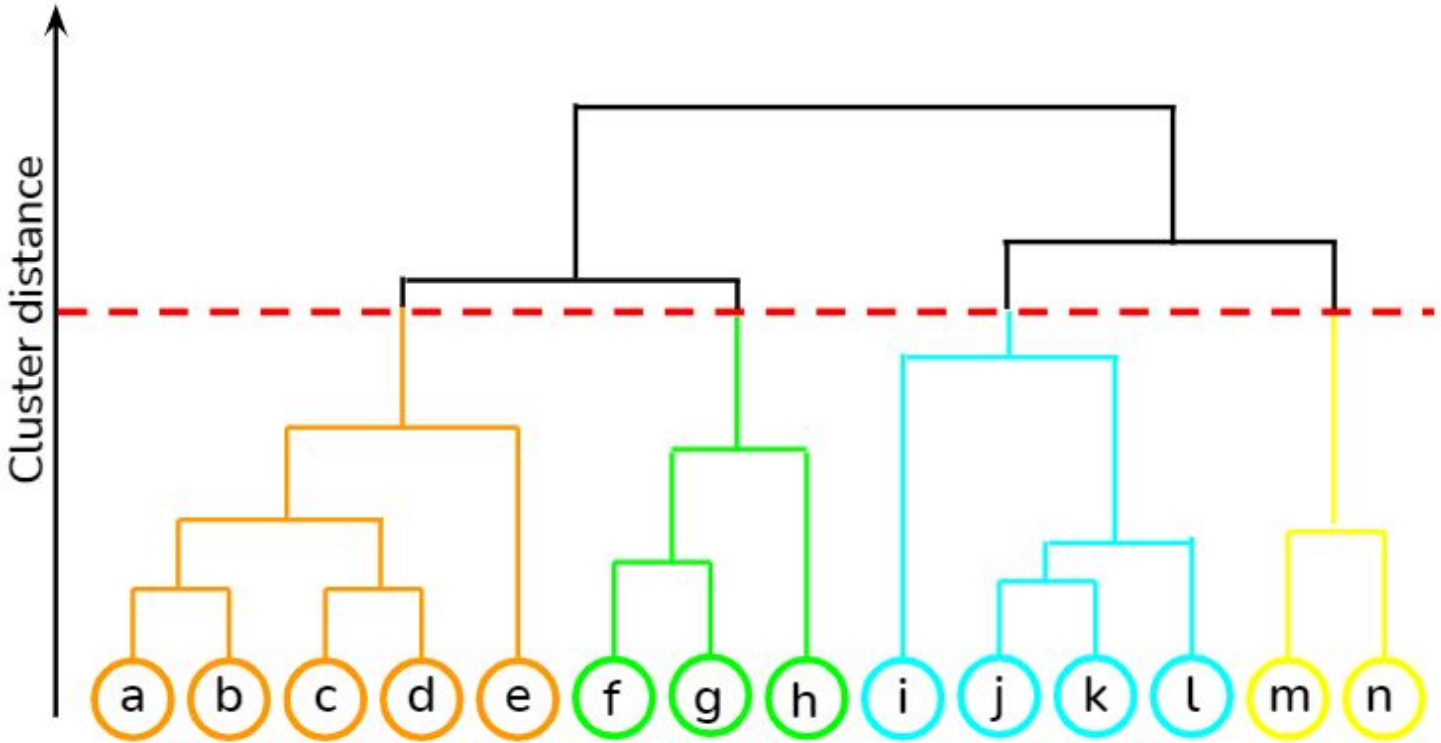


Figure from https://towardsdatascience.com/hierarchical-clustering-explained-e59b13846da8

How to perform hierarchical clustering in R

Summary of clustering algorithms

- Model-free (no distributional assumptions, require a distance metric):
  - K-means
    - Upsides: Runs quickly in practice
    - Downsides: Need to know k
  - Hierarchical clustering
    - Upsides: Don’t need to choose k to construct a dendrogram
    - Downsides: can be VERY slow (time-complexity is O(N^3) and memory O(N^2))
- Model-based approaches (not discussed today)
  - Gaussian mixture models
  - etc

Advanced topics in unsupervised learning (part II)

Kasia Kedzierska https://kasia.codes/
Kaspar Märtens https://kaspar.website/

UNIVERSITY OF OXFORD

Probabilistic and non-linear extensions of PCA

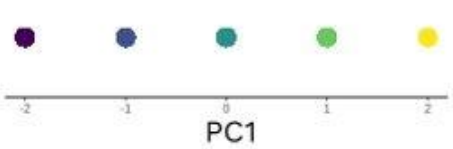
Probabilistic PCA: turning PCA into a generative model

PCA

Observed data

PCA

Low-dimensional latent representation

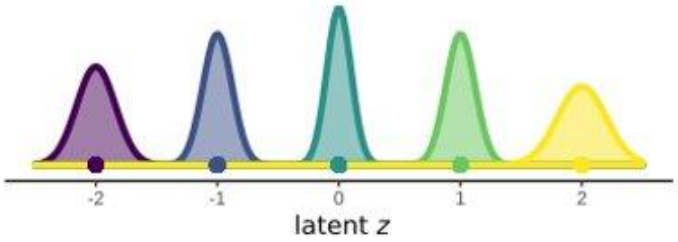


Probabilistic PCA

Observed data

Probabilistic PCA

Low-dimensional latent representation



Advantages of Probabilistic PCA

- Interpretation as a ***generative*** model
- **Uncertainty** in latent space
- Can handle **missing data**
- Can "automatically" choose the latent space dimension
- Naturally extendable to **binary/count/categorical data**

Citation: Tipping and Bishop, 1999

From PCA to Autoencoders

What is an autoencoder?

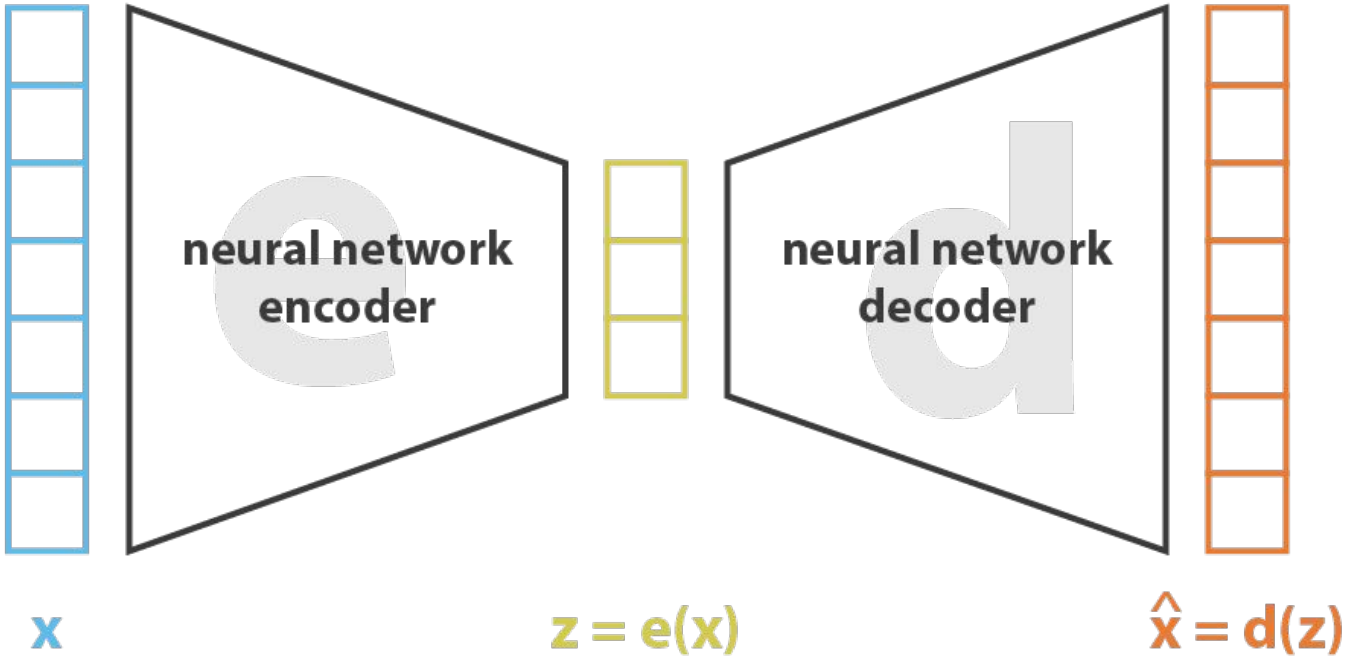


Figure from https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

From PCA to Autoencoders

Linear autoencoder implements PCA

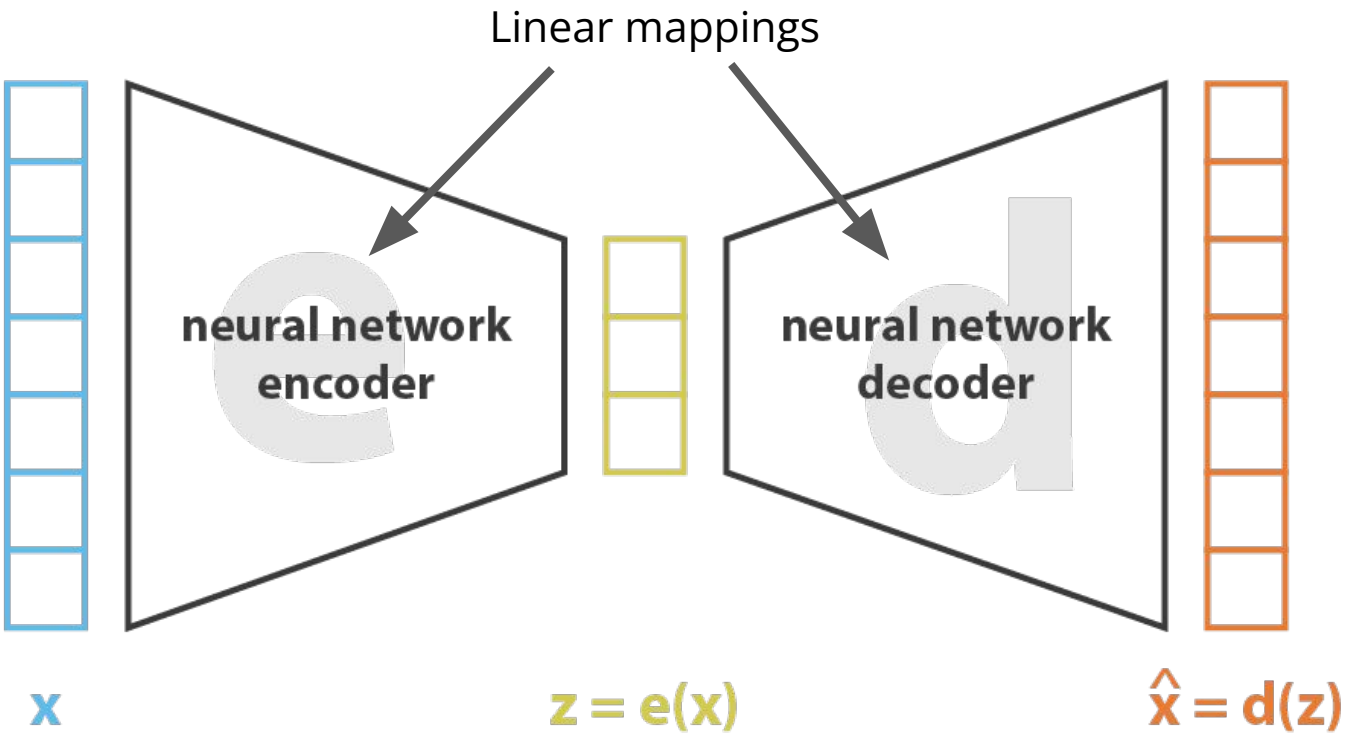
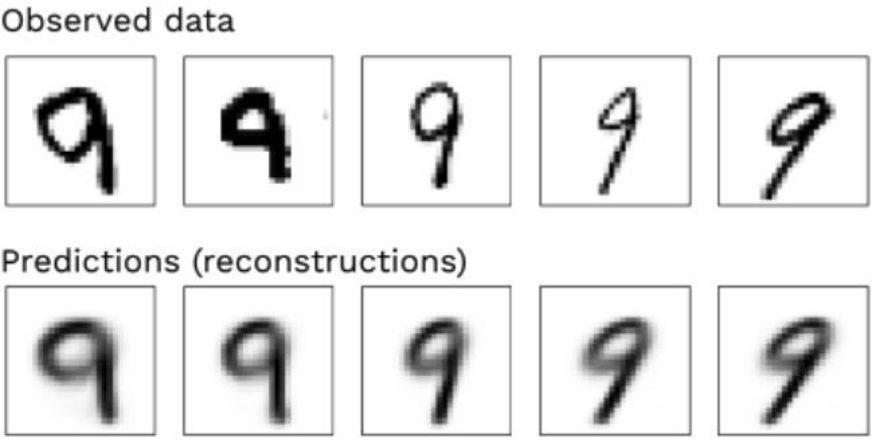
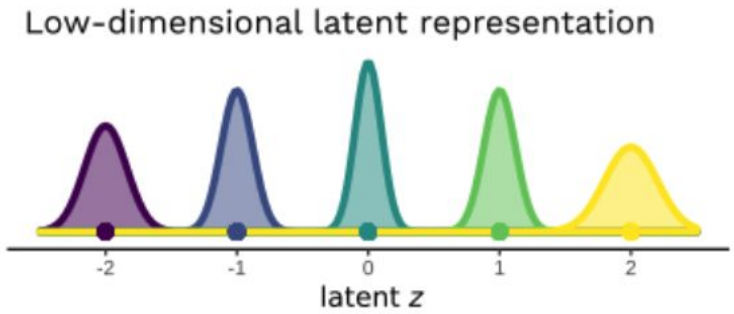


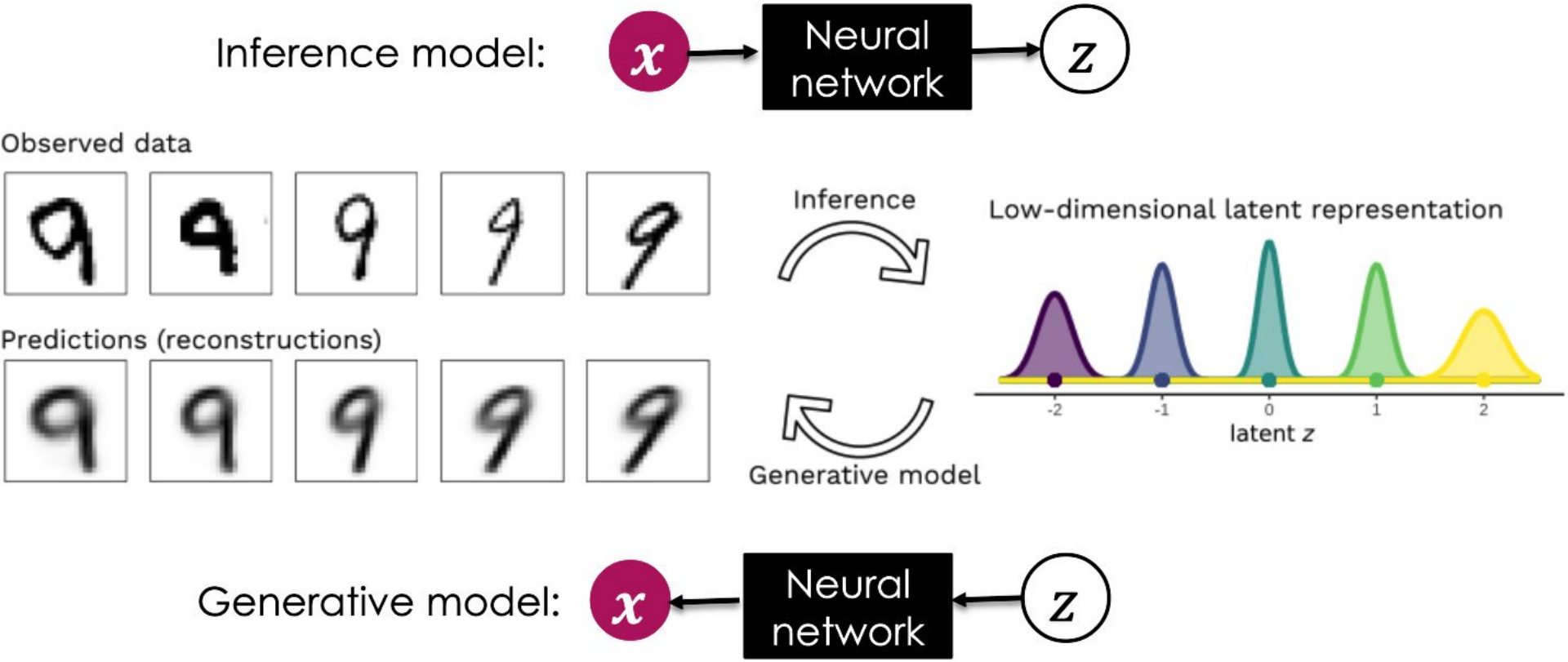
Figure from https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73

Variational Autoencoders are a non-linear extension of probabilistic PCA



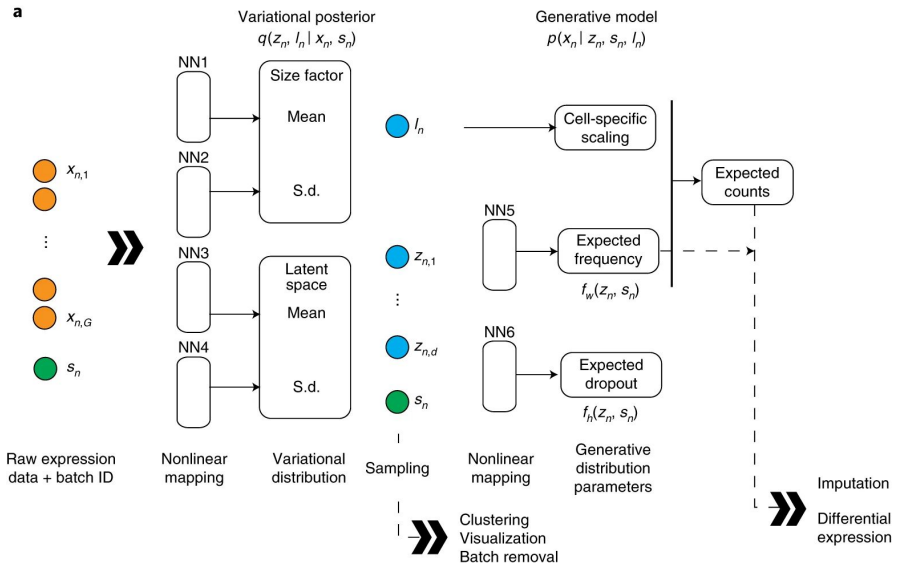


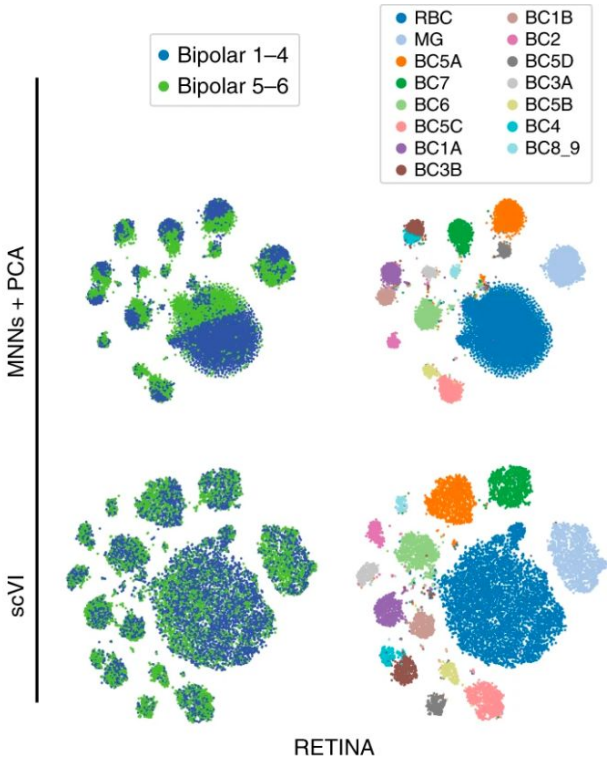
Variational Autoencoders are a non-linear extension of probabilistic PCA



Variational Autoencoders for single cell transcriptomics

Model setup/architecture:

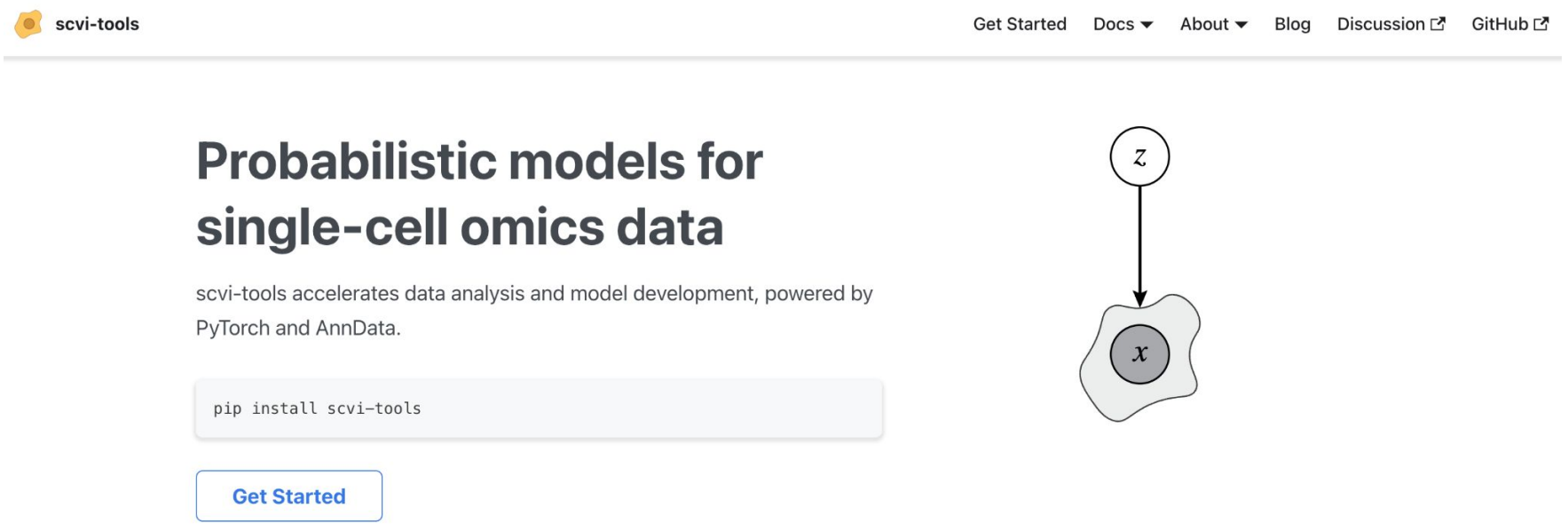




Deep generative modeling for single-cell transcriptomics https://www.nature.com/articles/s41592-018-0229-2

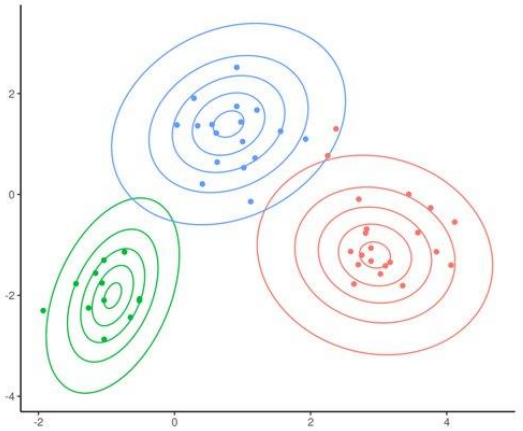
scVI-tools

VAE based methods can be useful for large single-cell omics data sets



https://docs.scvi-tools.org/en/stable/tutorials/index.html

The everlasting question:
How to choose “k” in clustering models



Can we circumvent choosing the number of clusters?

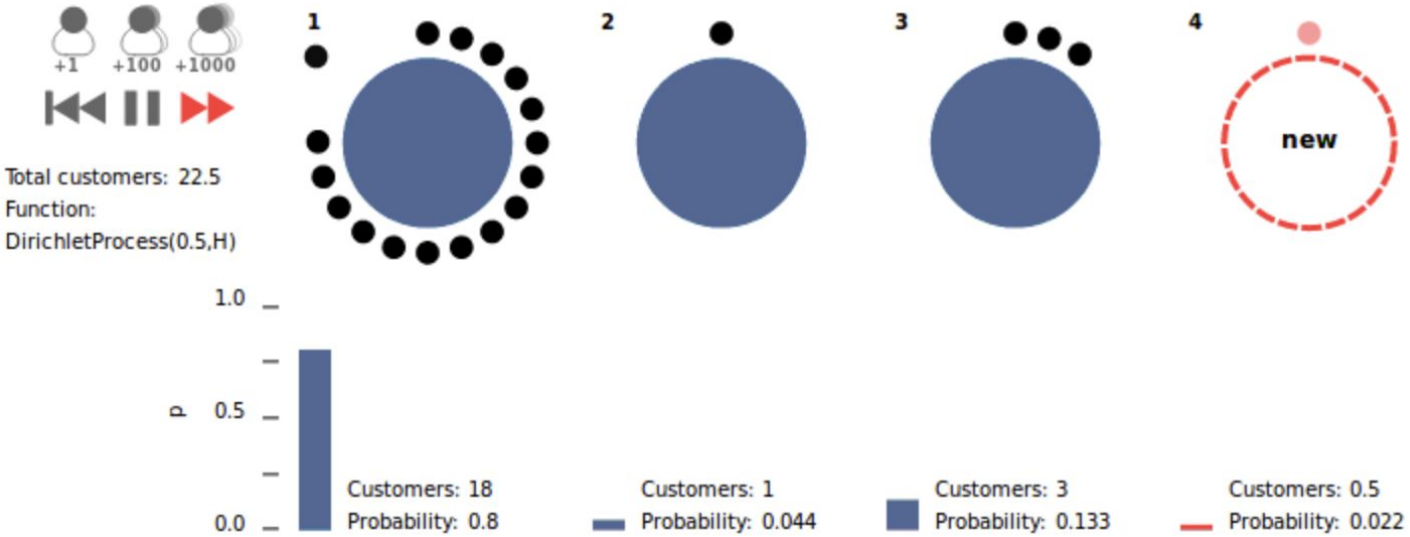
To some extent: there is an entire research field "**Bayesian nonparametrics**" trying to circumvent the issue by allowing a flexible (potentially infinite) number of clusters

Desirable properties of Bayesian nonparametric models:

- When new data comes in, ***new clusters*** should be ***allowed to appear*** (i.e. the number of clusters should not be fixed)
- We might want to capture ***uncertainty*** about the number of clusters

Flexible number of clusters with the Chinese Restaurant Process

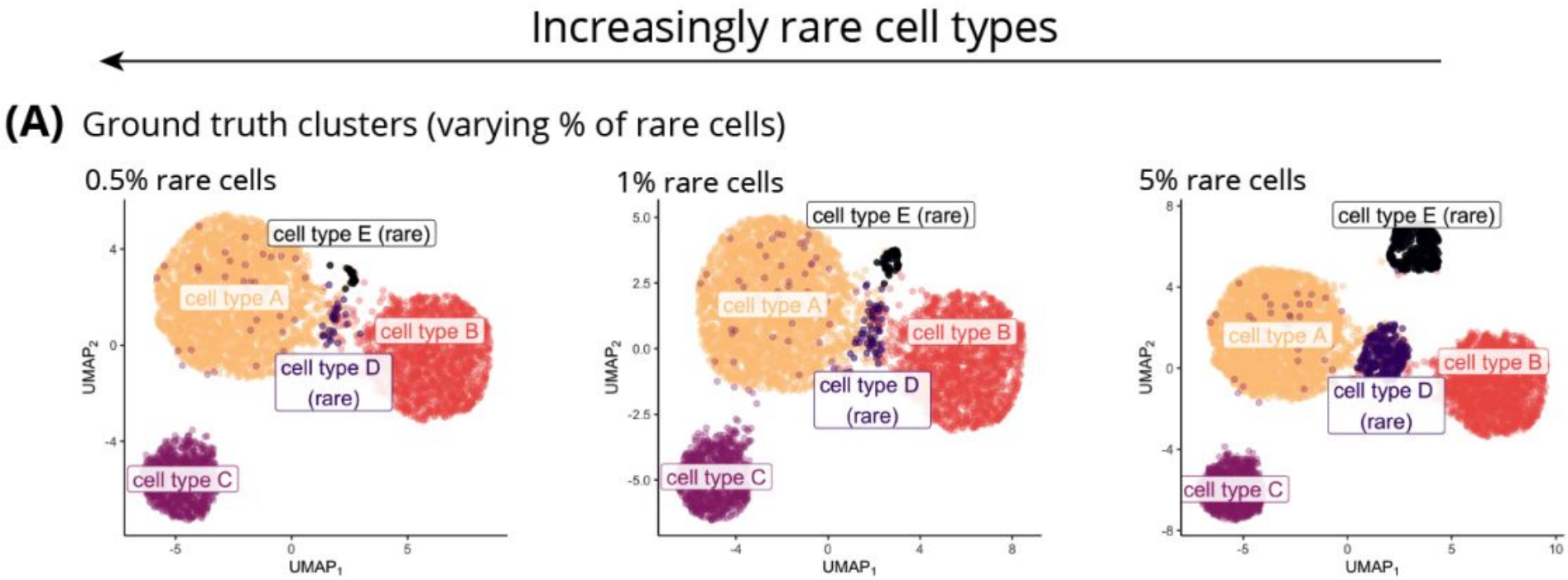
We could use a Gaussian Mixture Model, where the number of clusters is specified by the Chinese Restaurant Process (thus allowed to be flexible)



See gif in https://commons.wikimedia.org/wiki/File:Chinese_Restaurant_Process_for_DP(0.5,H).gif

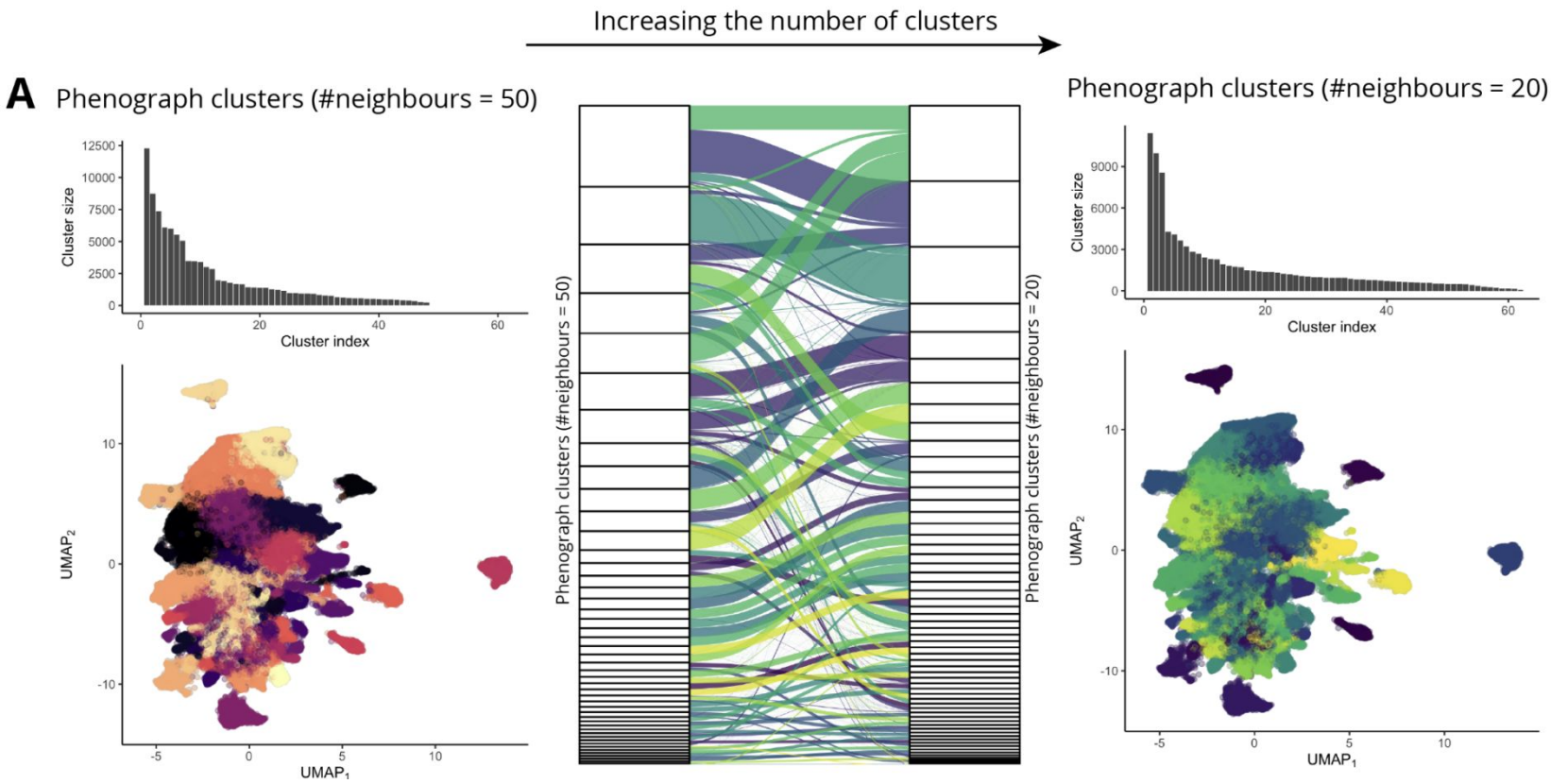
Identifying rare clusters

Detecting rare clusters is challenging



Rarity: Discovering rare cell populations from single-cell imaging data
https://www.biorxiv.org/content/10.1101/2022.07.15.500256v1

Can we simply increase *k* to increase granularity of clusters?



Example: discovering rare CD4- CD8- T cell clusters