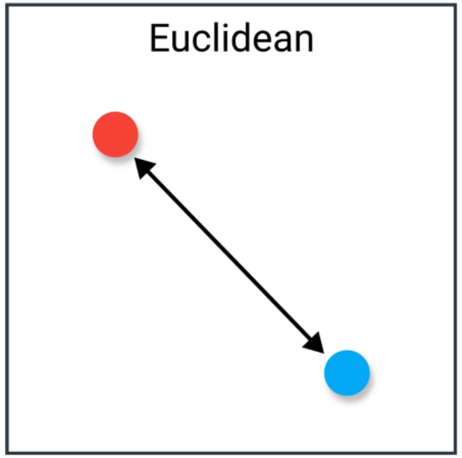
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

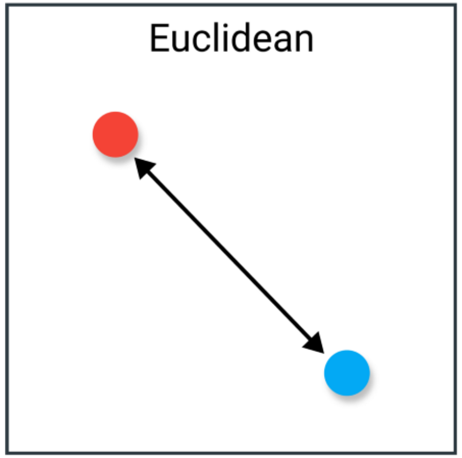
- Distance
- Other dimension reduction approaches (briefly)
    - Principal Coordinate Analysis (PCoA)
    - Non-Metric Multidimensional Scaling (NMDS)
    - UMAP: Uniform Manifold Approximation and Projection
    - t-SNE: t-distributed stochastic neighbor embedding

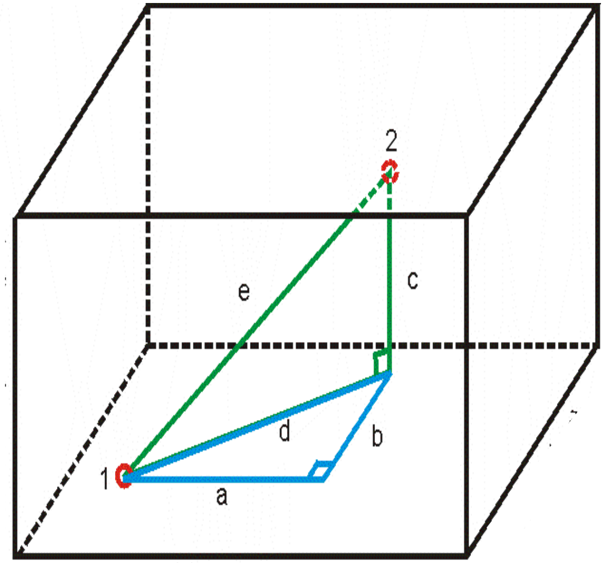
Measuring distance between observations



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

Measuring distance between observations





distance between points

$$2D: \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2}$$

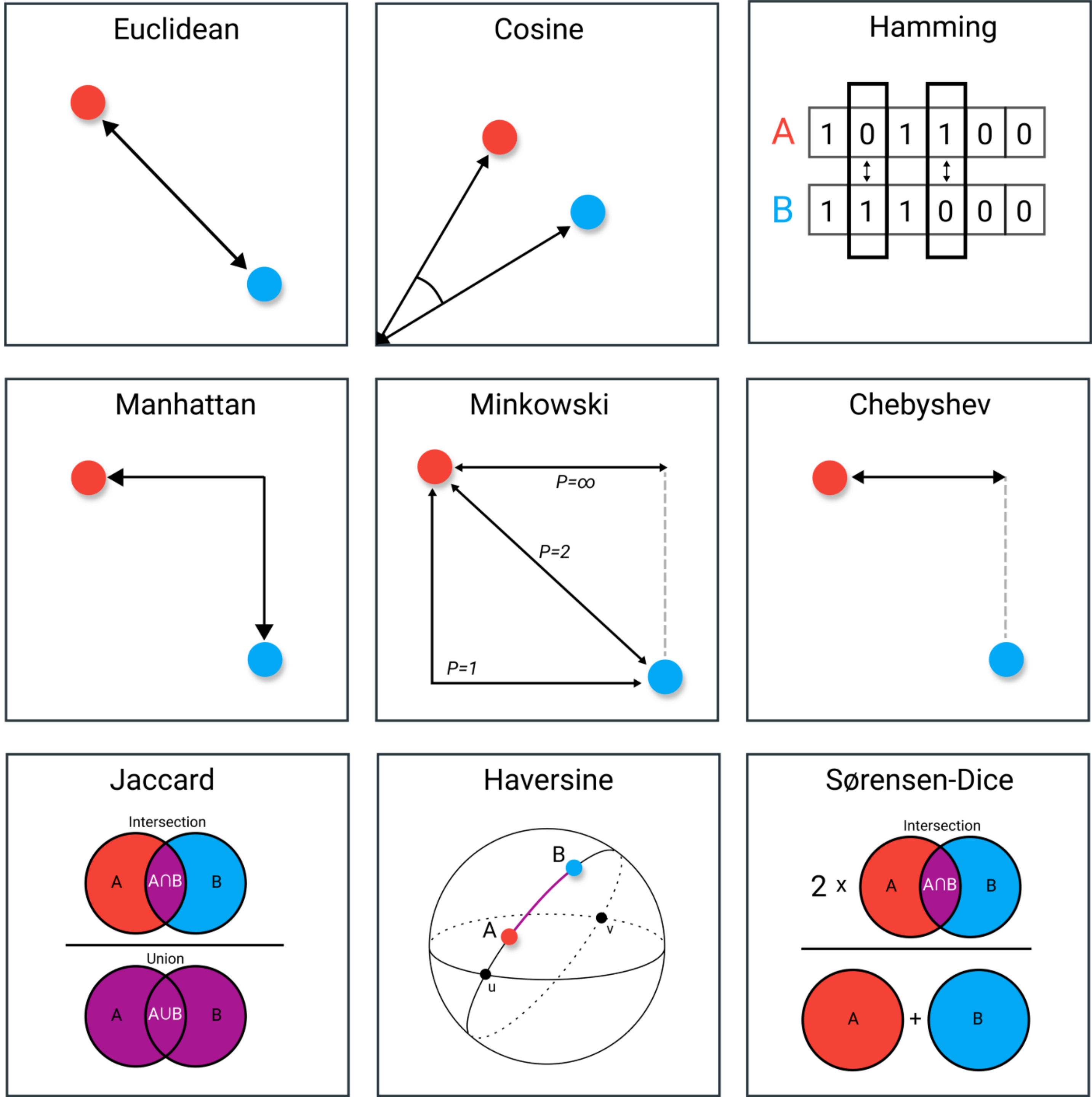
$$3D: \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2}$$

$$4D: \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2 + (a_1 - a_0)^2}$$

$$nD: \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2 + (a_1 - a_0)^2 + \cdots}$$

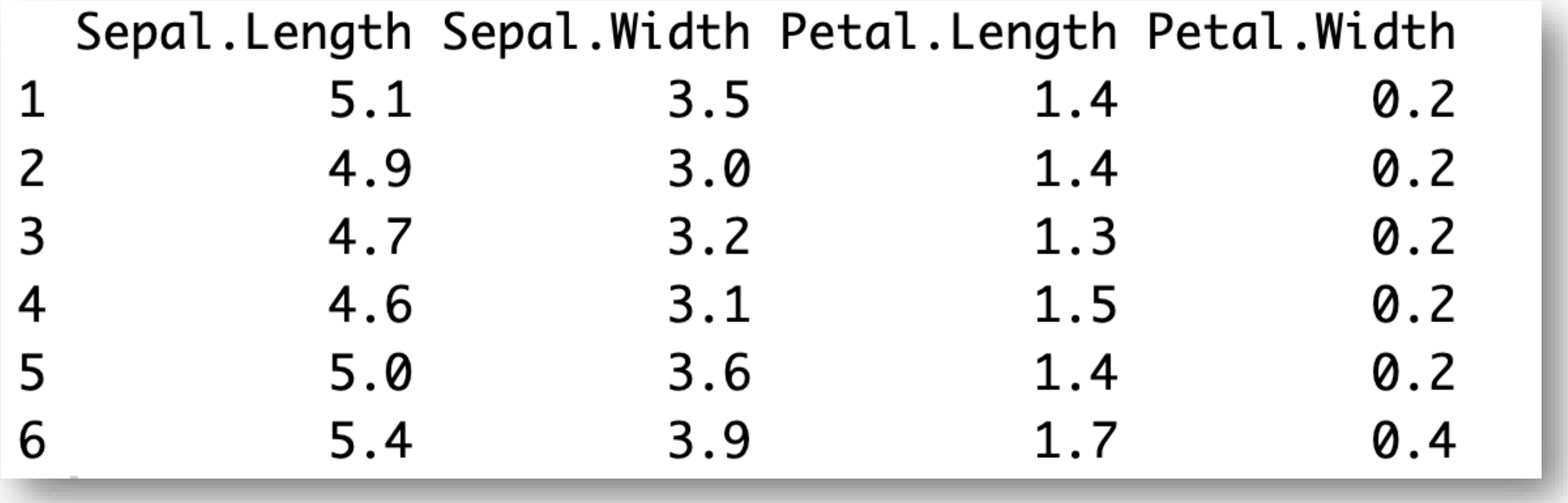
https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa
https://hlab.stanford.edu/brian/euclidean_distance_in.html
https://www.youtube.com/watch?v=K6Eu0kRolmA

Measuring distance between observations



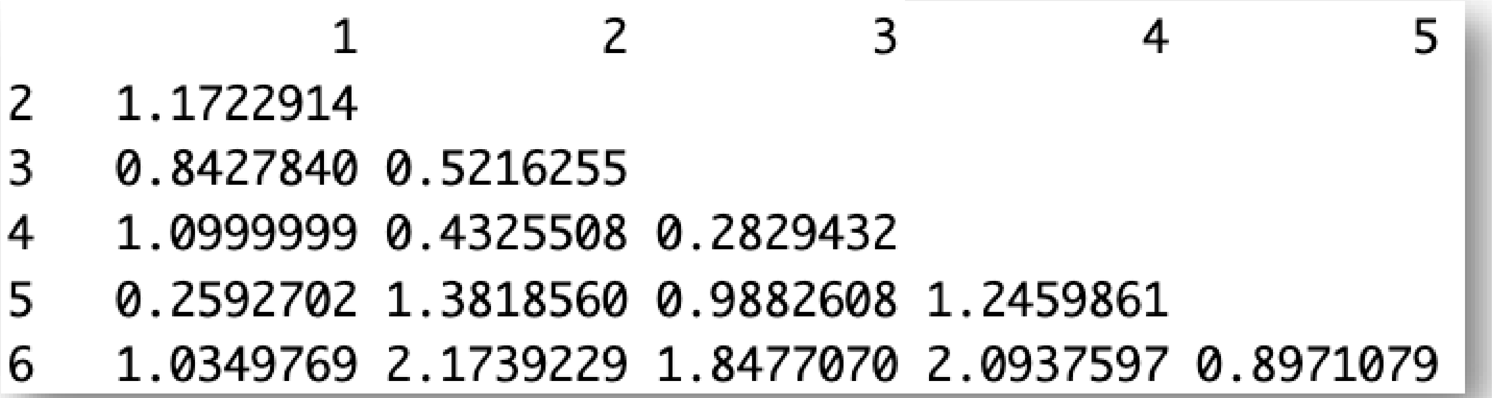
https://towardsdatascience.com/9-distance-measures-in-data-science-918109d069fa

```
head(iris[,1:4])
```



```
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4
```

```
dist.mat <- dist(scale(iris[,1:4]), method = "euclidean")
dist.mat
```

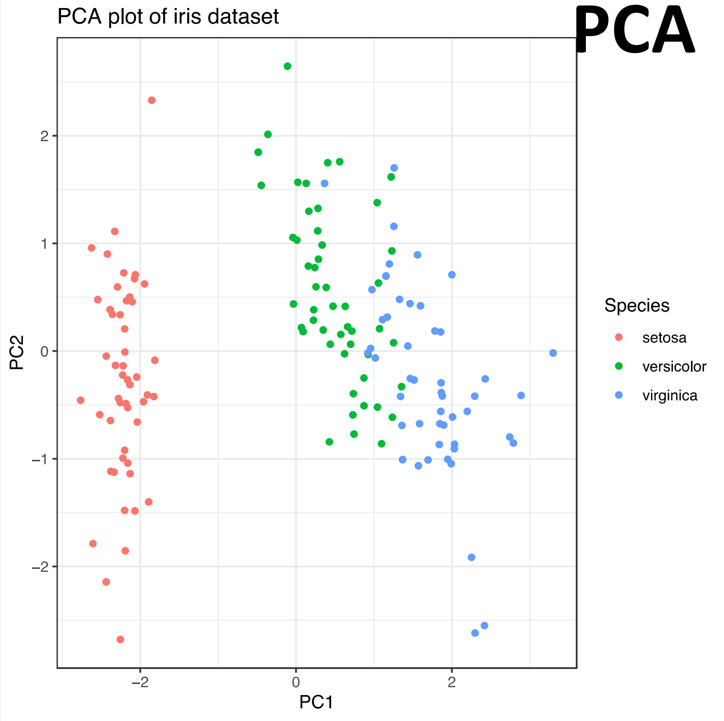


```
          1         2         3         4         5
2 1.1722914
3 0.8427840 0.5216255
4 1.0999999 0.4325508 0.2829432
5 0.2592702 1.3818560 0.9882608 1.2459861
6 1.0349769 2.1739229 1.8477070 2.0937597 0.8971079
```

Principal Coordinate Analyses (PCoA)

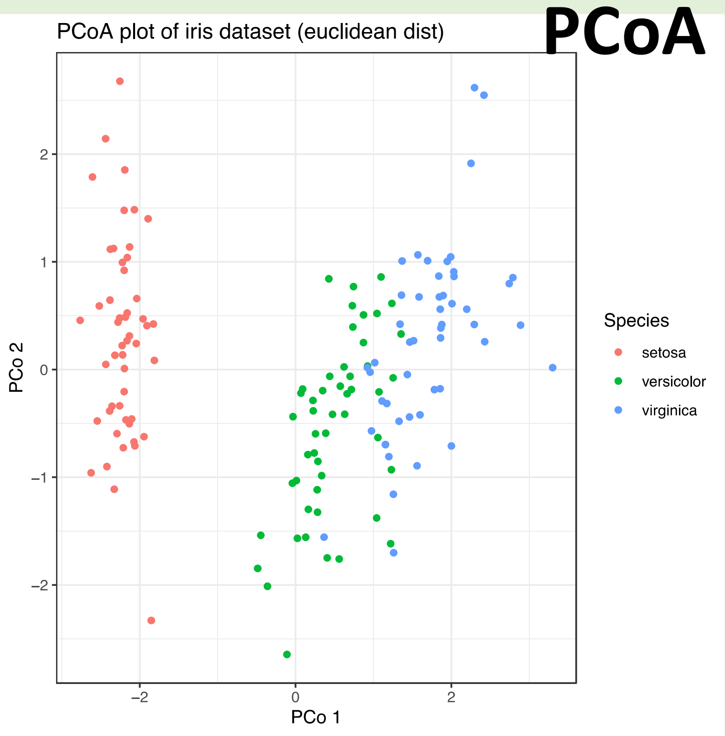
- Uses distance (dissimilarity) rather than correlation (similarity) to define new axes describing multivariate patterns of variation.

PCA and PCoA equal when using Euclidean distance (but note here that PC2 is reverse sign, a reminder that the sign of PCs are arbitrary)



PCA plot of iris dataset

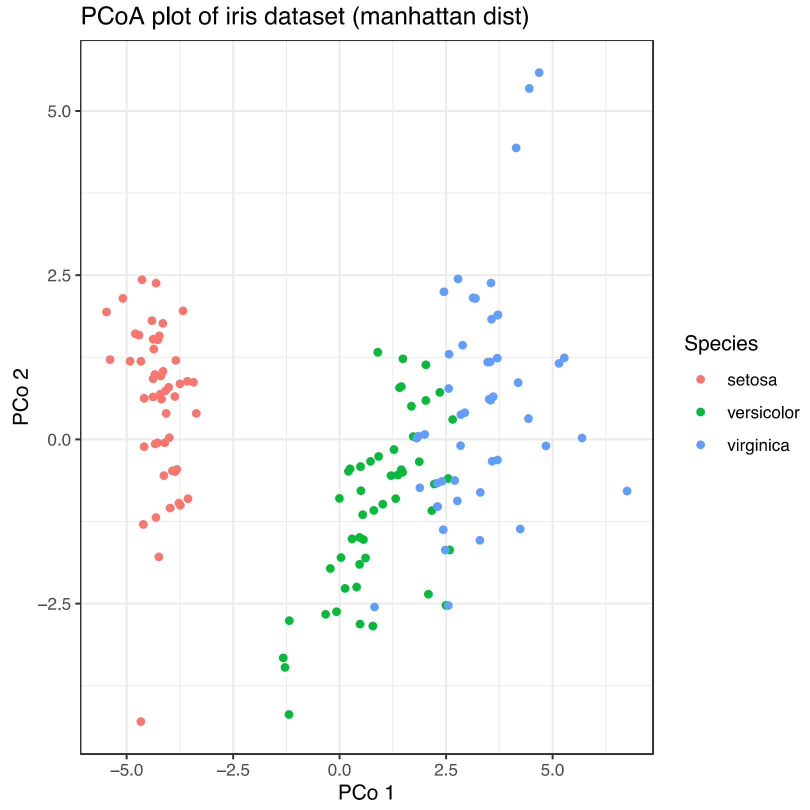
PCA



PCoA plot of iris dataset (euclidean dist)

PCoA

PCoA (Manhattan)

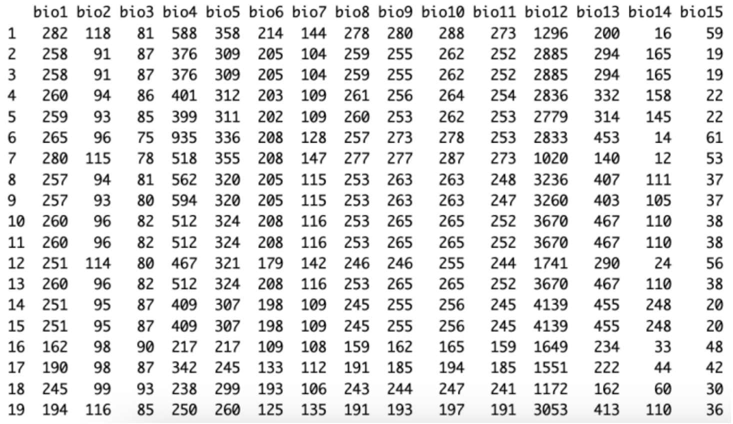


PCoA plot of iris dataset (manhattan dist)

Why use different measure of distance?

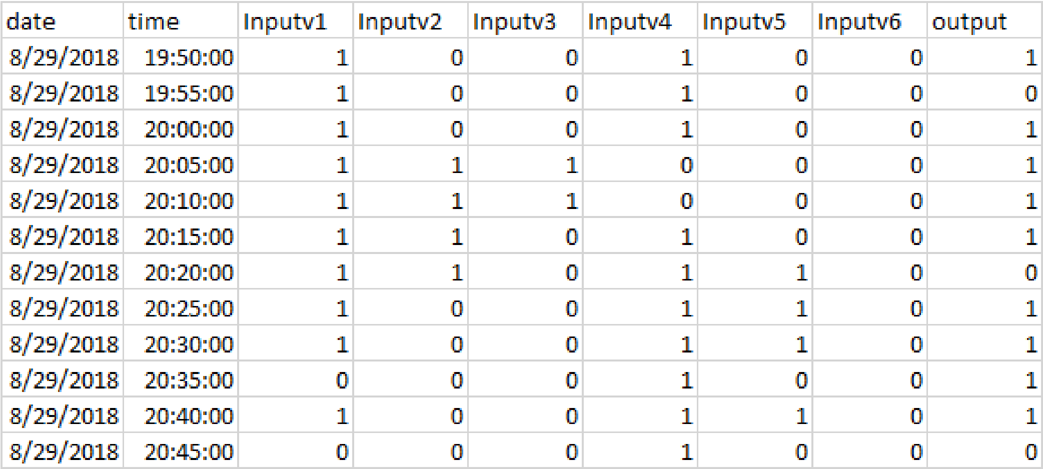
Different measures of distance may be more appropriate depending on the kind of data, your question, and your field!

Quantitative variables



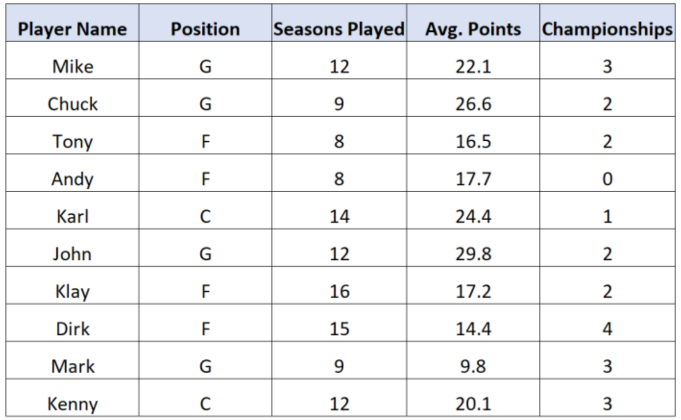
```
   bio1 bio2 bio3 bio4 bio5 bio6 bio7 bio8 bio9 bio10 bio11 bio12 bio13 bio14 bio15
1   282  118   81  588  358  214  144  278  280   288   273  1296   200    16    59
2   258   91   87  376  309  205  104  259  255   262   252  2885   294   165    19
3   258   91   87  376  309  205  104  259  255   262   252  2885   294   165    19
4   260   94   86  401  312  203  109  261  256   264   254  2836   332   158    22
5   259   93   85  399  311  202  109  260  253   262   253  2779   314   145    22
6   265   96   75  935  336  208  128  257  273   278   253  2833   453    14    61
7   280  115   78  518  355  208  147  277  277   287   273  1020   140    12    53
8   257   94   81  562  320  205  115  253  263   263   248  3236   407   111    37
9   257   93   80  594  320  205  115  253  263   263   247  3260   403   105    37
10  260   96   82  512  324  208  116  253  265   265   252  3670   467   110    38
11  260   96   82  512  324  208  116  253  265   265   252  3670   467   110    38
12  251  114   80  467  321  179  142  246  246   255   244  1741   290    24    56
13  260   96   82  512  324  208  116  253  265   265   252  3670   467   110    38
14  251   95   87  409  307  198  109  245  255   256   245  4139   455   248    20
15  251   95   87  409  307  198  109  245  255   256   245  4139   455   248    20
16  162   98   90  217  217  109  108  159  162   165   159  1649   234    33    48
17  190   98   87  342  245  133  112  191  185   194   185  1551   222    44    42
18  245   99   93  238  299  193  106  243  244   247   241  1172   162    60    30
19  194  116   85  250  260  125  135  191  193   197   191  3053   413   110    36
```

Binary data



date	time	Inputv1	Inputv2	Inputv3	Inputv4	Inputv5	Inputv6	output
8/29/2018	19:50:00	1	0	0	1	0	0	1
8/29/2018	19:55:00	1	0	0	1	0	0	0
8/29/2018	20:00:00	1	0	0	1	0	0	1
8/29/2018	20:05:00	1	1	1	0	0	0	1
8/29/2018	20:10:00	1	1	1	0	0	0	1
8/29/2018	20:15:00	1	1	0	1	0	0	1
8/29/2018	20:20:00	1	1	0	1	1	0	0
8/29/2018	20:25:00	1	0	0	1	1	0	1
8/29/2018	20:30:00	1	0	0	1	1	0	1
8/29/2018	20:35:00	0	0	0	1	0	0	1
8/29/2018	20:40:00	1	0	0	1	1	0	1
8/29/2018	20:45:00	0	0	0	1	0	0	0

Qualitative and Quantitative variables

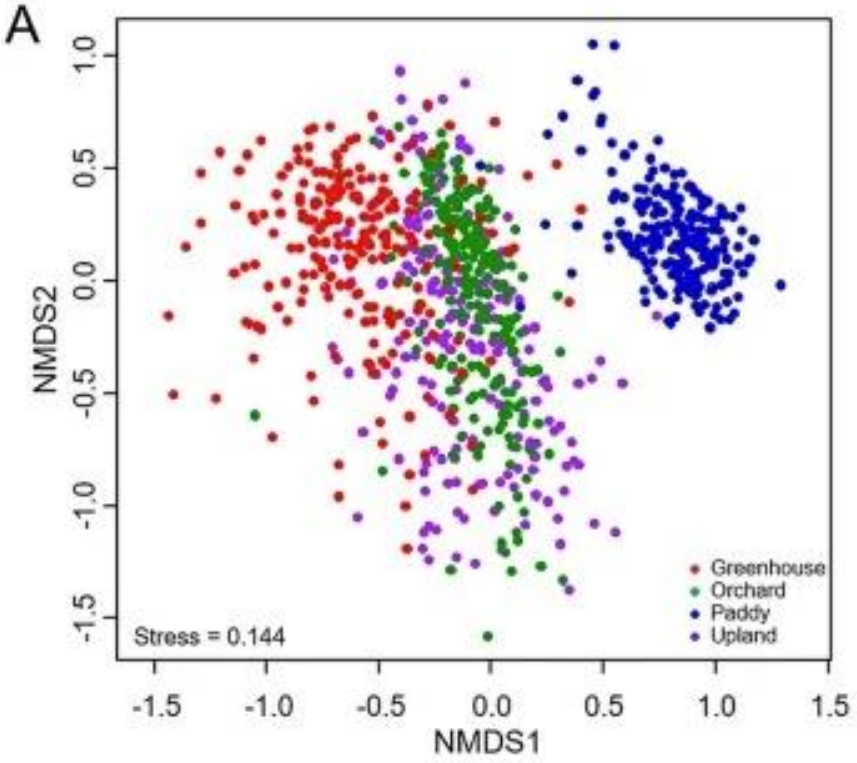


Player Name	Position	Seasons Played	Avg. Points	Championships
Mike	G	12	22.1	3
Chuck	G	9	26.6	2
Tony	F	8	16.5	2
Andy	F	8	17.7	0
Karl	C	14	24.4	1
John	G	12	29.8	2
Klay	F	16	17.2	2
Dirk	F	15	14.4	4
Mark	G	9	9.8	3
Kenny	C	12	20.1	3

Non-Metric Multidimensional Scaling (nMDS)

- nMDS uses a matrix of distances or dissimilarities among samples as its input
- It converts the raw dissimilarity values in the matrix into ranks
- It uses the rank order of the distances between samples to create the ordination plot
- It is useful if the variables in your data have non-linear relationships
- High flexibility for many kinds of data
- nMDS is for strictly for plotting to visualize relationships. It can’t be used for dimension reduction

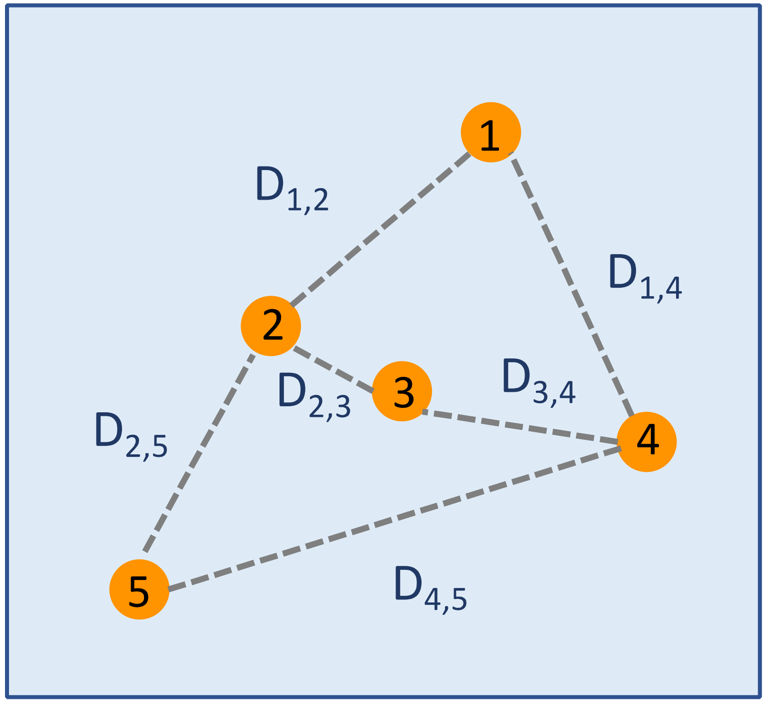
Lee et al. 2020. Sci Reports: 10, 17418



Beta diversity of soil bacterial communities

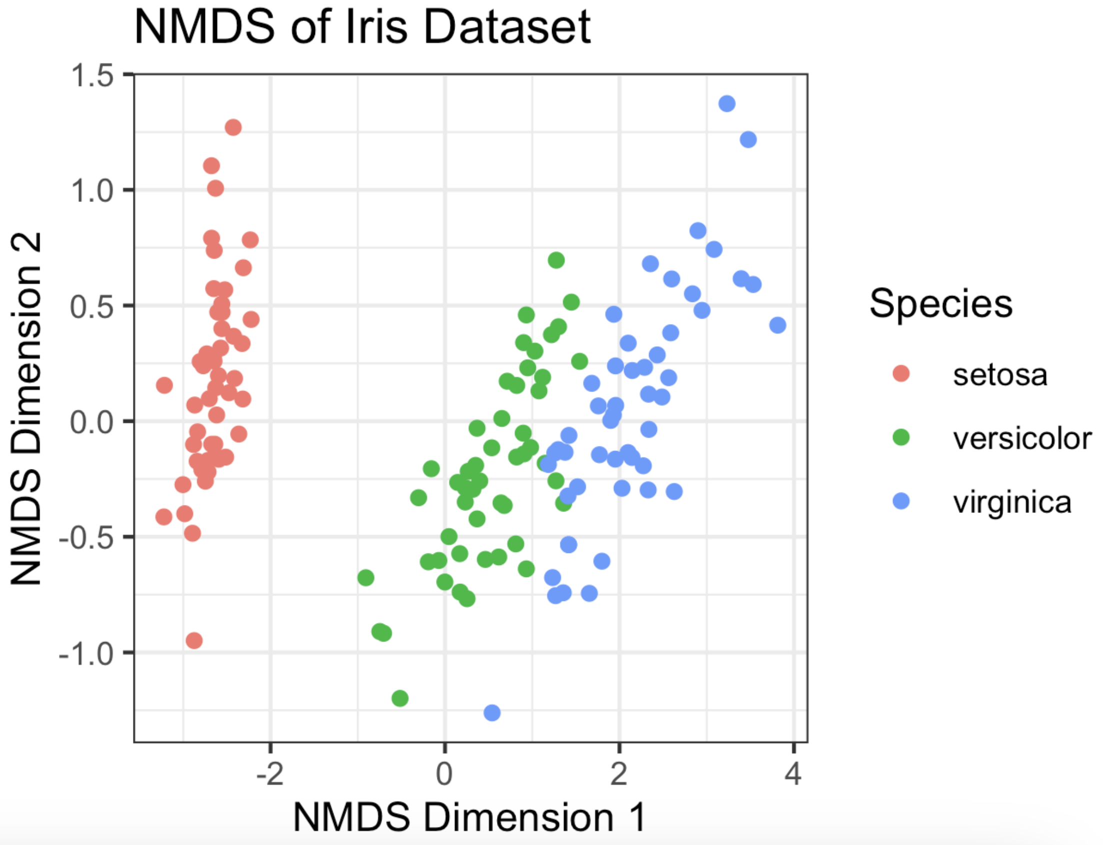
nMDS: How it Works

- Specify the number of dimensions
- nMDS uses an iterative process
- Places the samples in 2D (or 3D) space and calculates the distance between samples in the ordination plot
- Rearranges points to maximize the correlation between the original distance matrix (which is stored as ranks) and the distances in the plot
- Performs this process multiple times using a different starting point each time until it reaches the optimal solution
- Best solution: where correlation between points in 2D plot and original matrix is maximized
  - stress value is the lowest

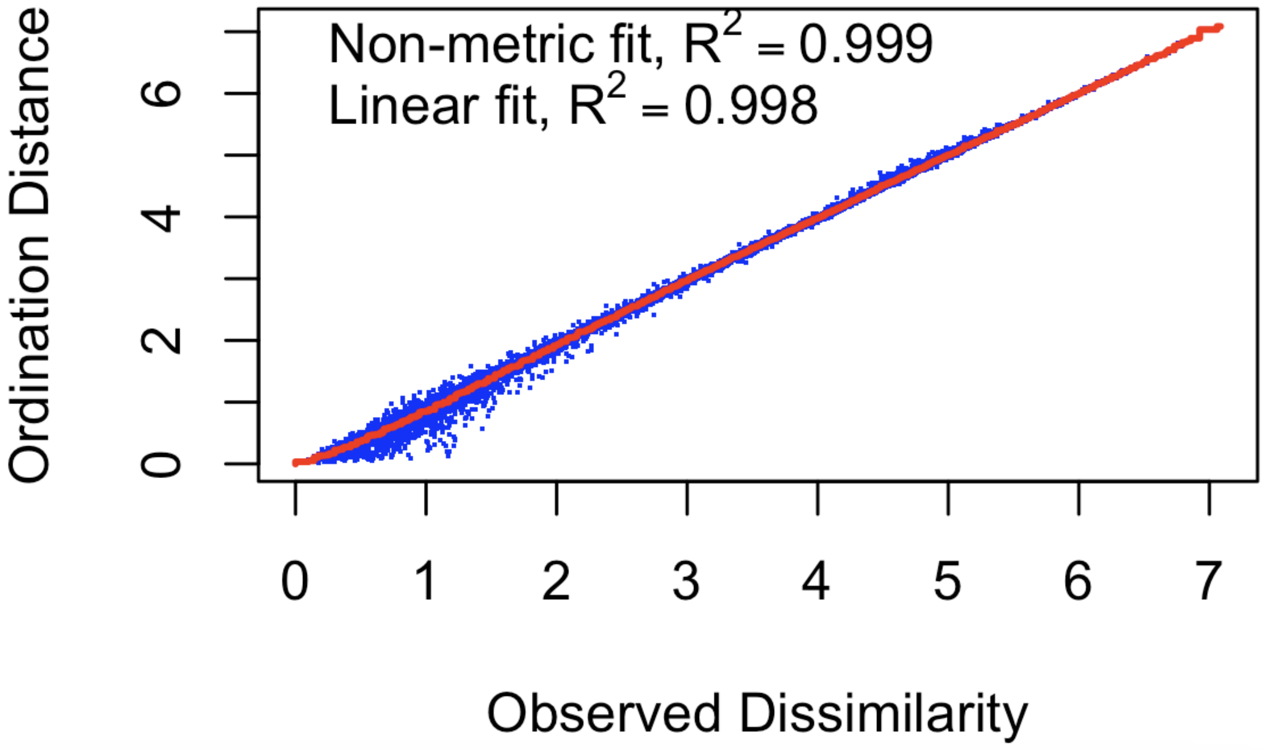


Missing some of the possible distances between points to keep plot from being too messy

nMDS: “stress” (goodness of fit)

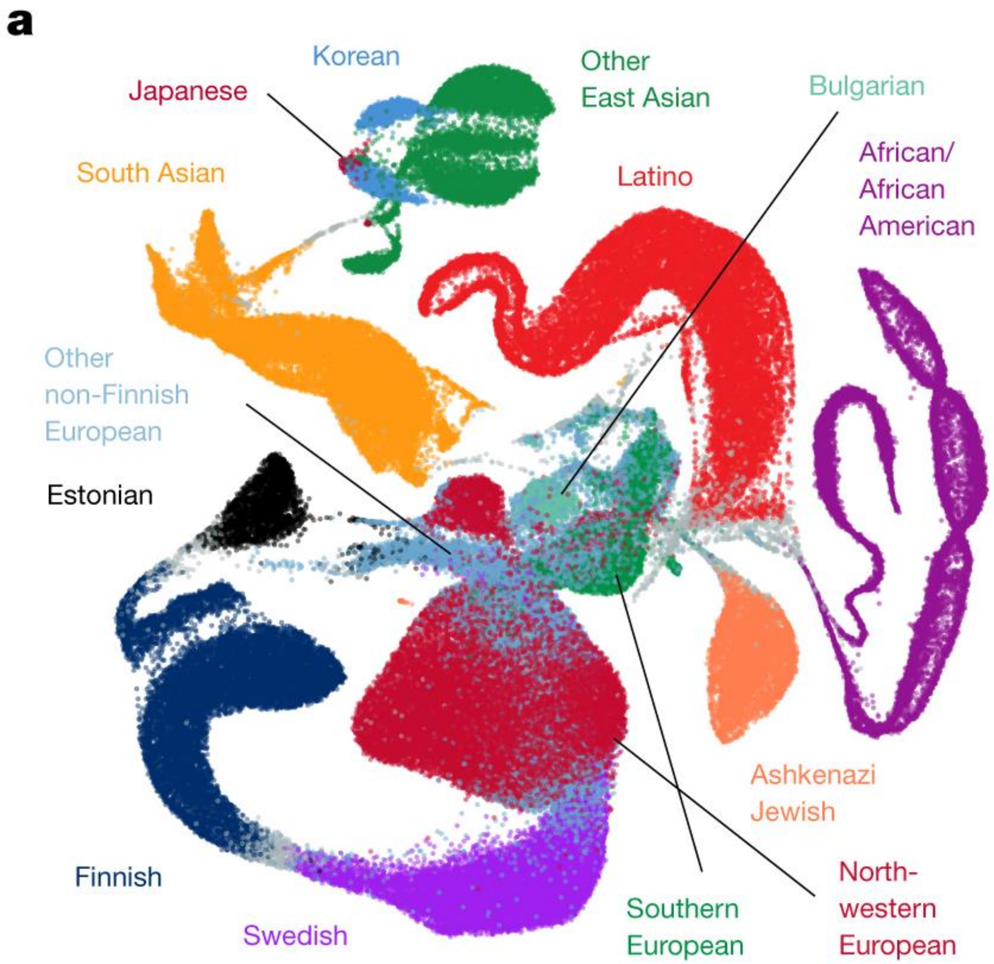


```
stressplot(iris_mds)
```



UMAP: Uniform Manifold Approximation and Projection

UMAP projection of human genetic relatedness



- UMAP tries to maintain clustering in multivariate data
- Deterministic (same result every time you run it)
- Sensitive to user set parameters (e.g. number of neighbors (n_neighbors))
- We can do in R with

  Library(umap)

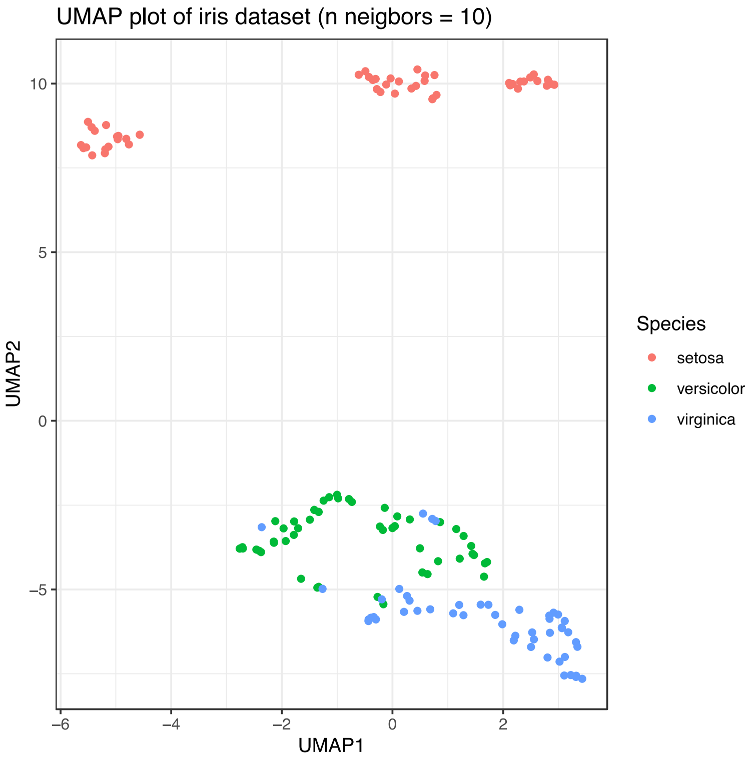
  iris_umap <- umap(iris)

https://www.nature.com/articles/s41586-020-2308-7/figures/1

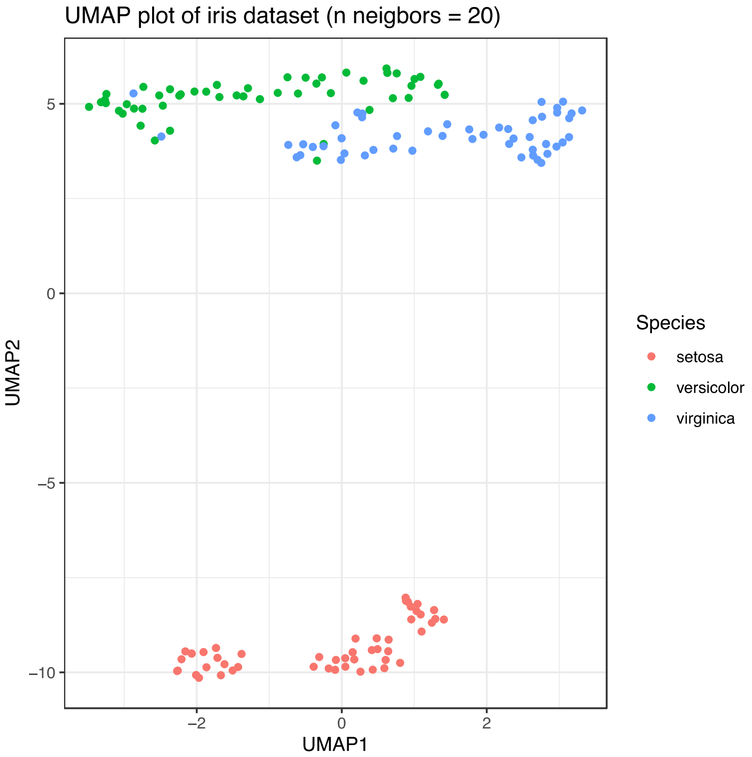
UMAP: Uniform Manifold Approximation and Projection

Results are different depending on the number of neighbors parameter

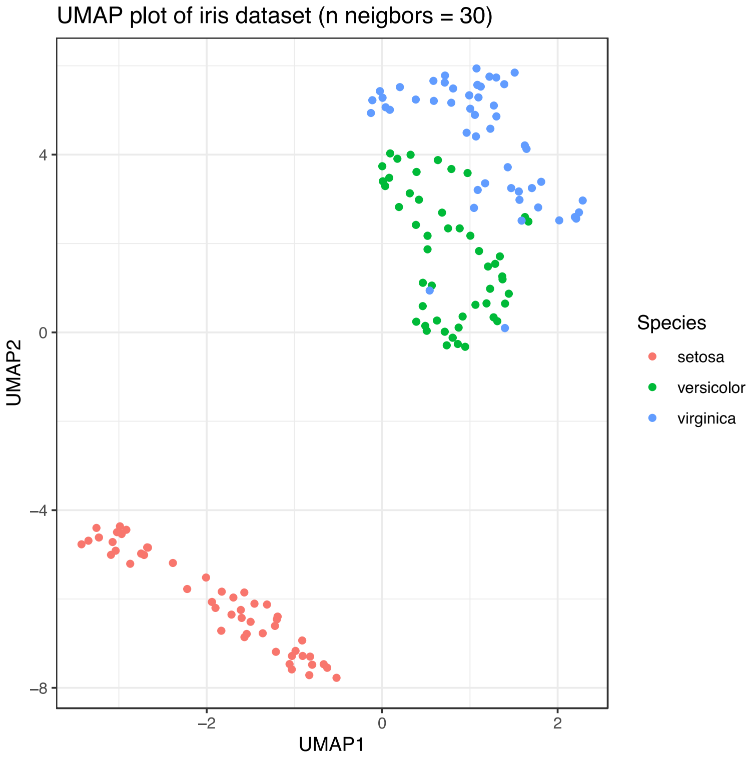
Neighbors = 10



Neighbors = 20

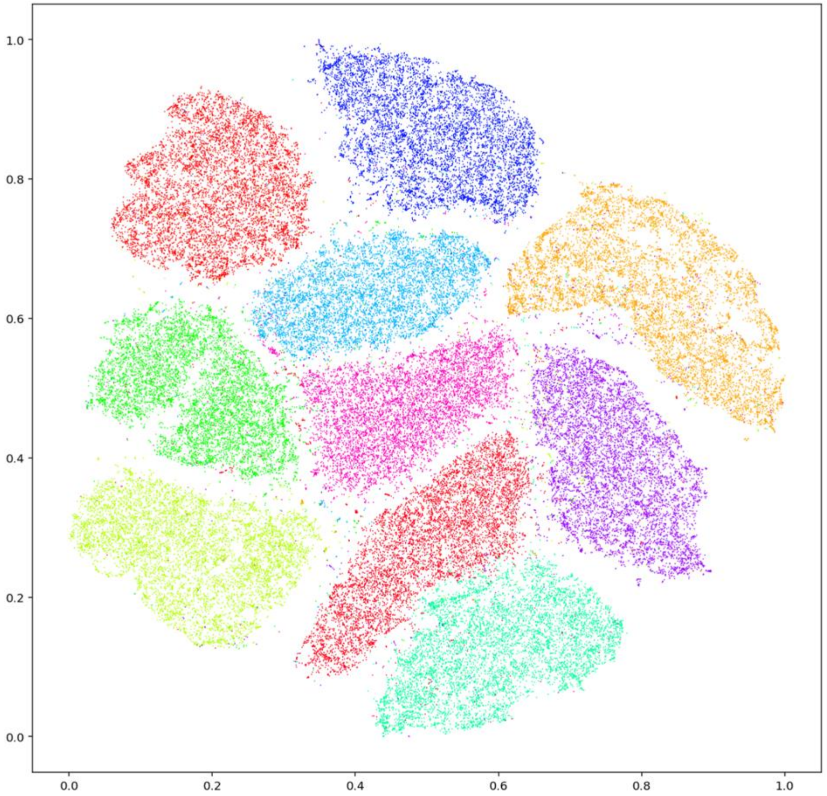


Neighbors = 30

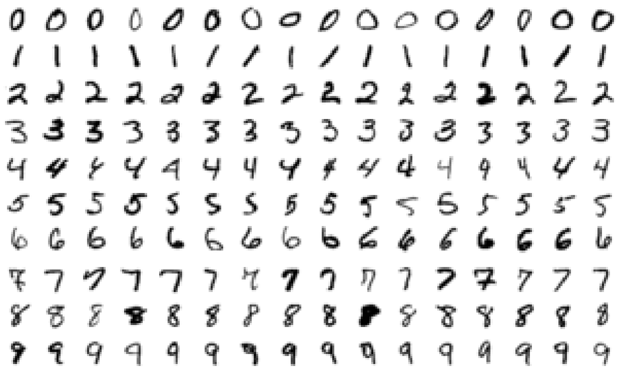


t-SNE (t-distributed stochastic neighbor embedding)

Colors represent different handwritten numbers



Eg. Modified National Institute of Standards and Technology database

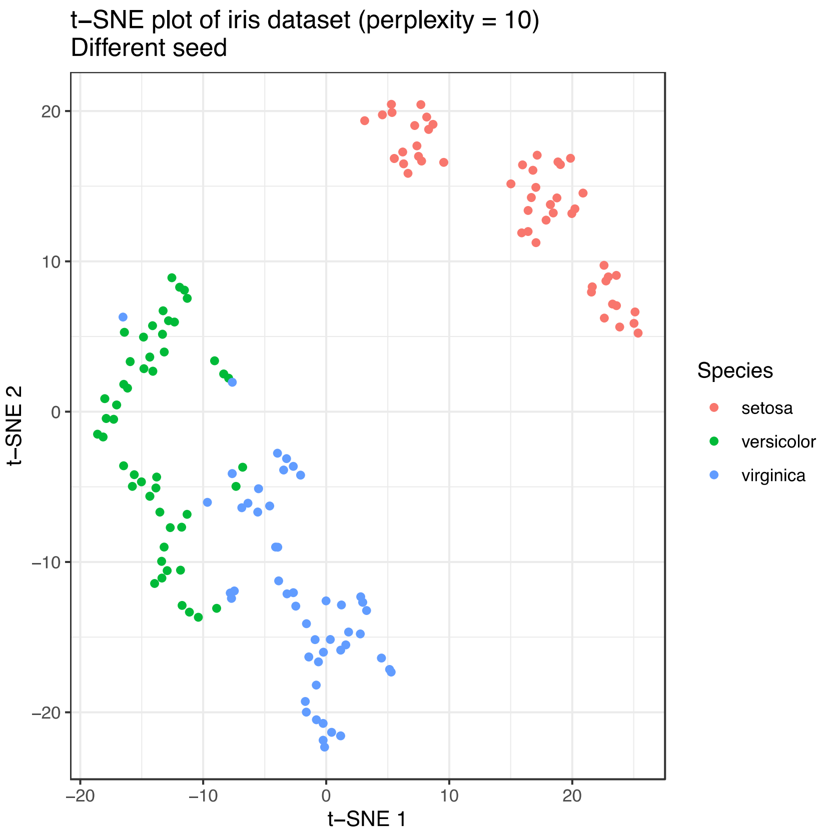


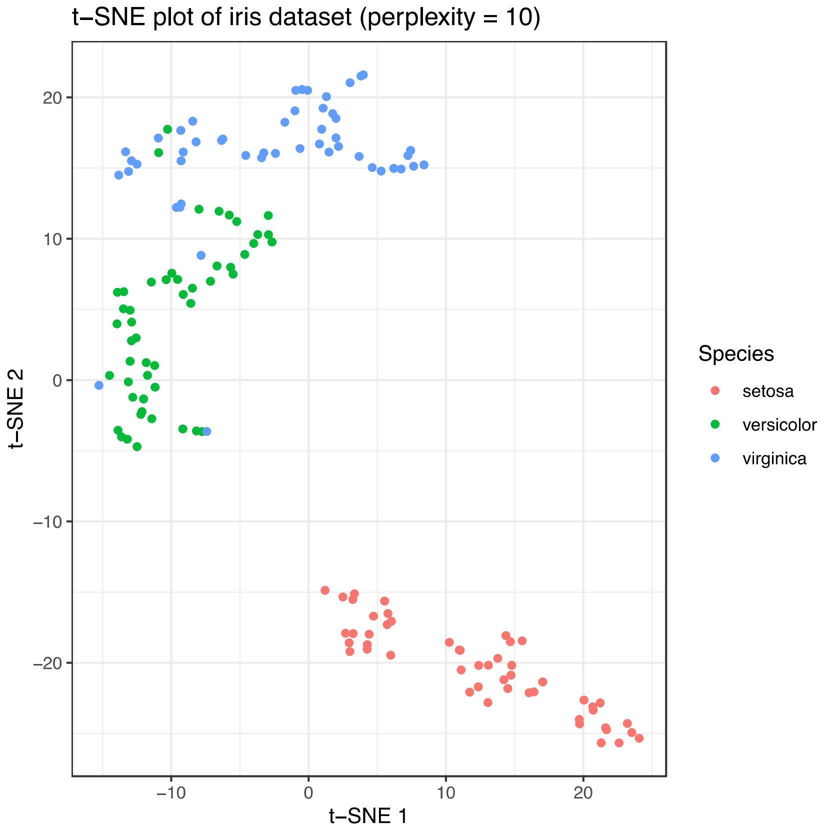
- Similar to UMAP
- Not deterministic (starts with random seed so results can change)
- Sensitive to user set parameters (e.g. perplexity)
- We can do in R with
  Library(Rtsne)
  iris_tsne <- Rtsne (iris)

https://en.wikipedia.org/wiki/MNIST_database
https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

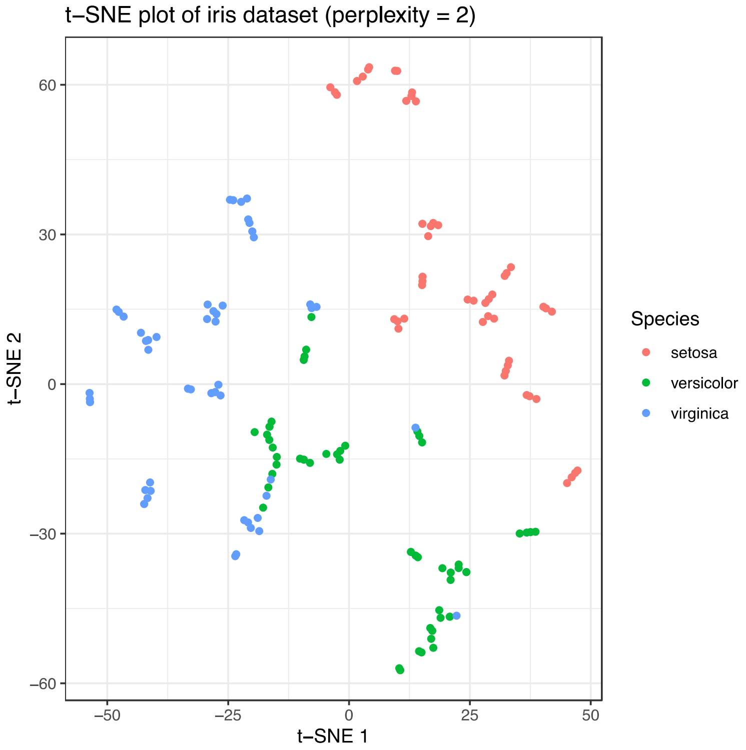
t-SNE (t-distributed stochastic neighbor embedding)

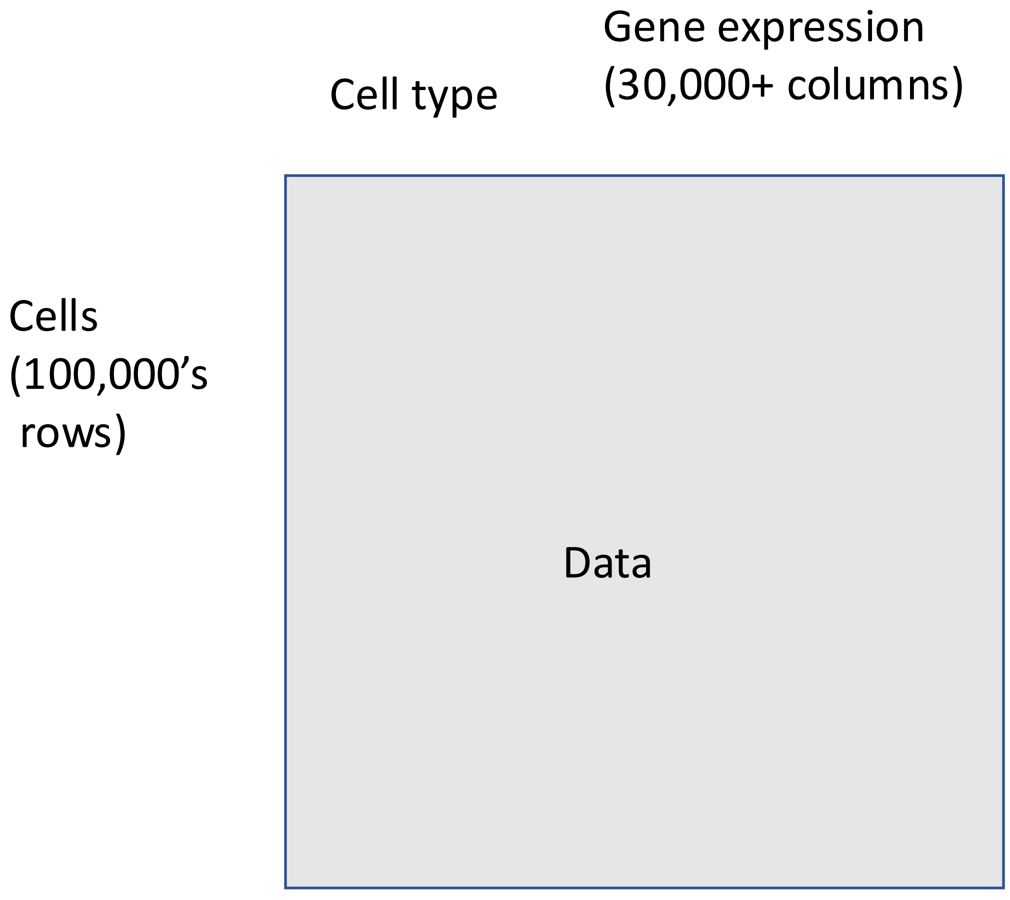
Same settings but different results due to non-determinacy of t-SNE





Different results with different perplexity



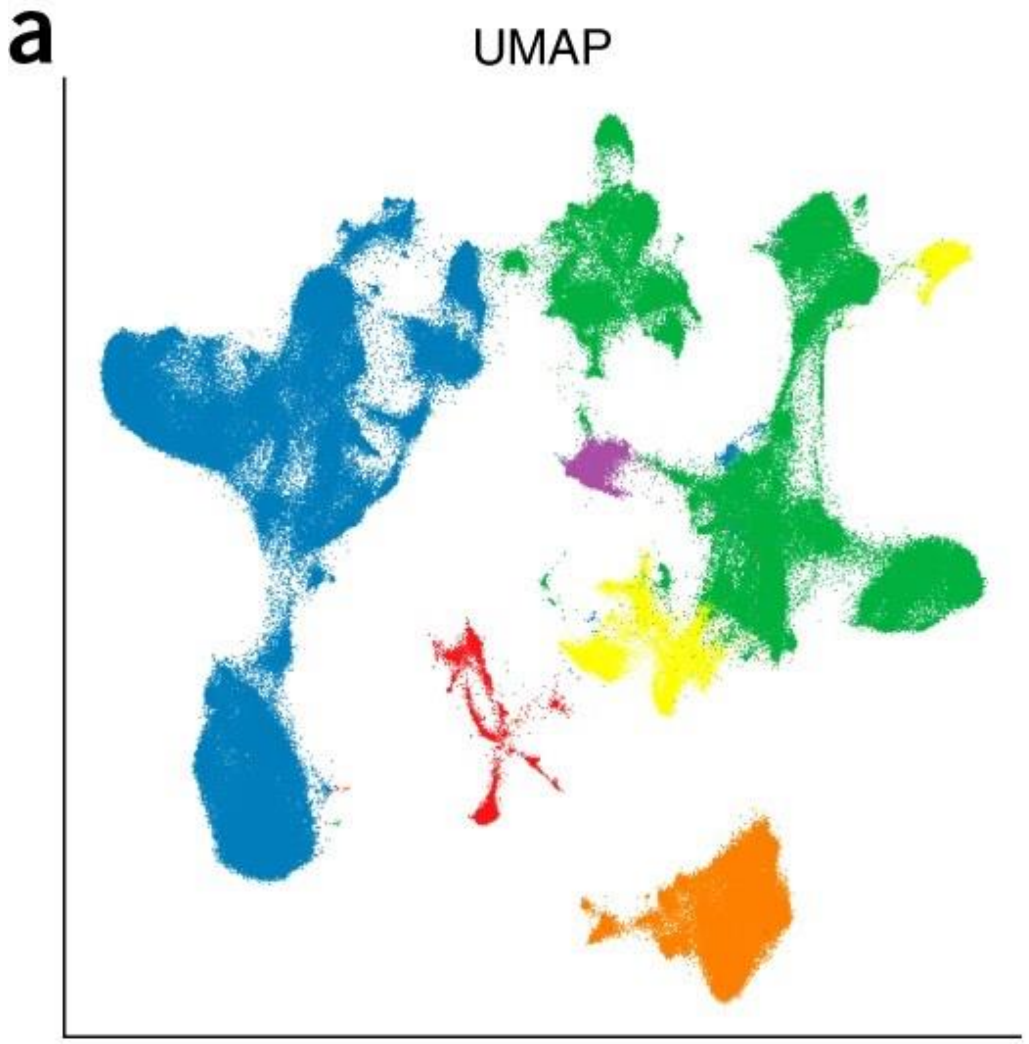


Cell type

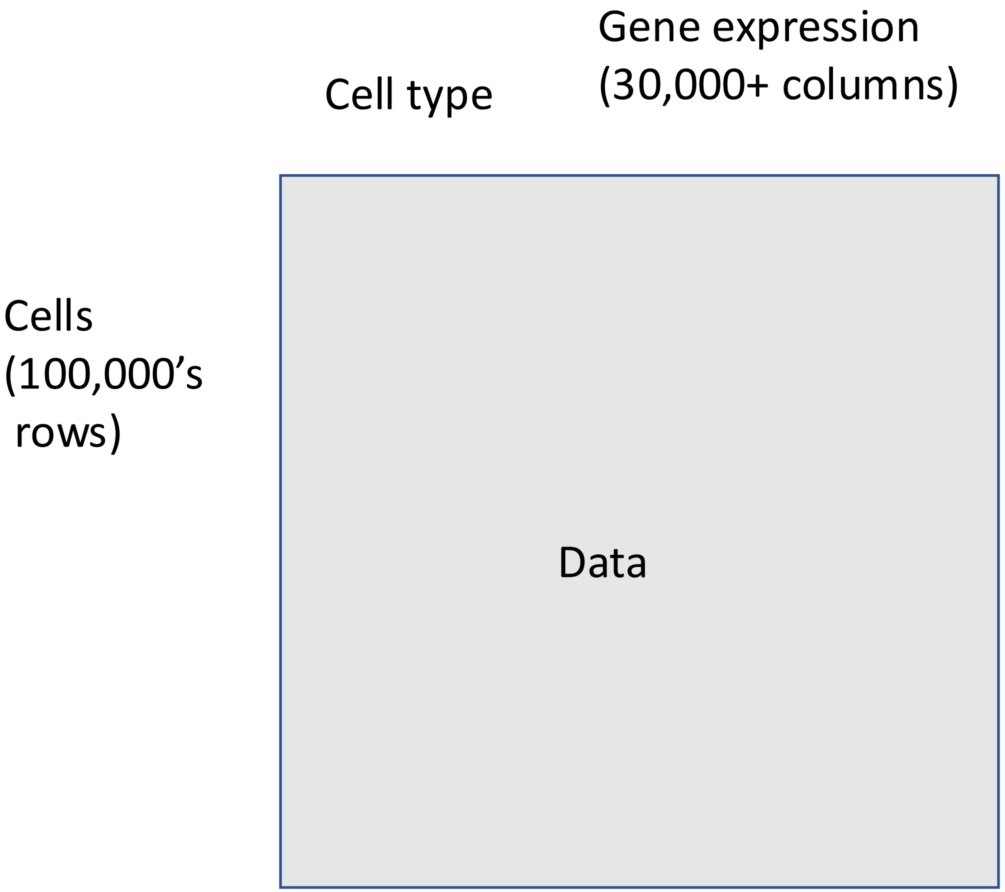
Gene expression
(30,000+ columns)

Cells
(100,000’s
rows)

Data



https://www.nature.com/articles/nbt.4314/figures/1

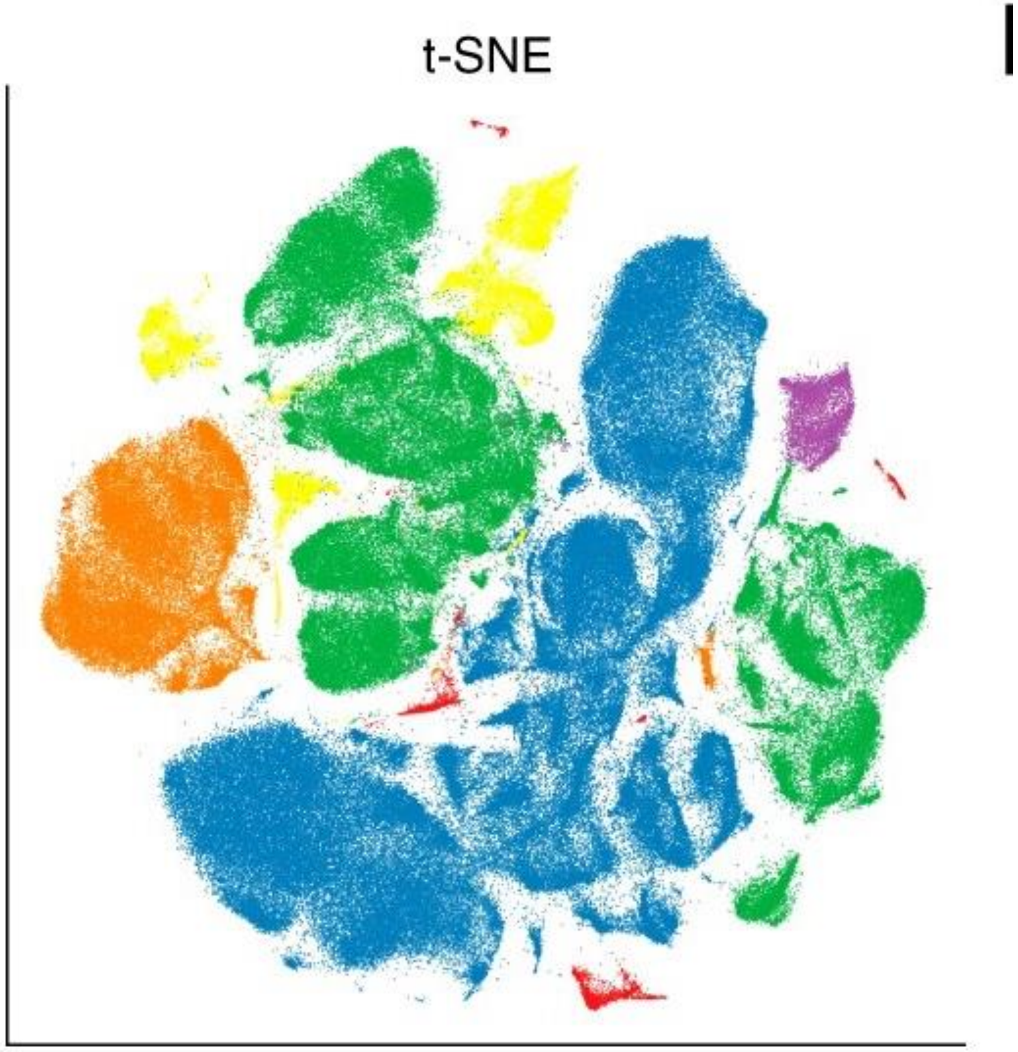


Cell type

Gene expression
(30,000+ columns)

Cells
(100,000's
rows)

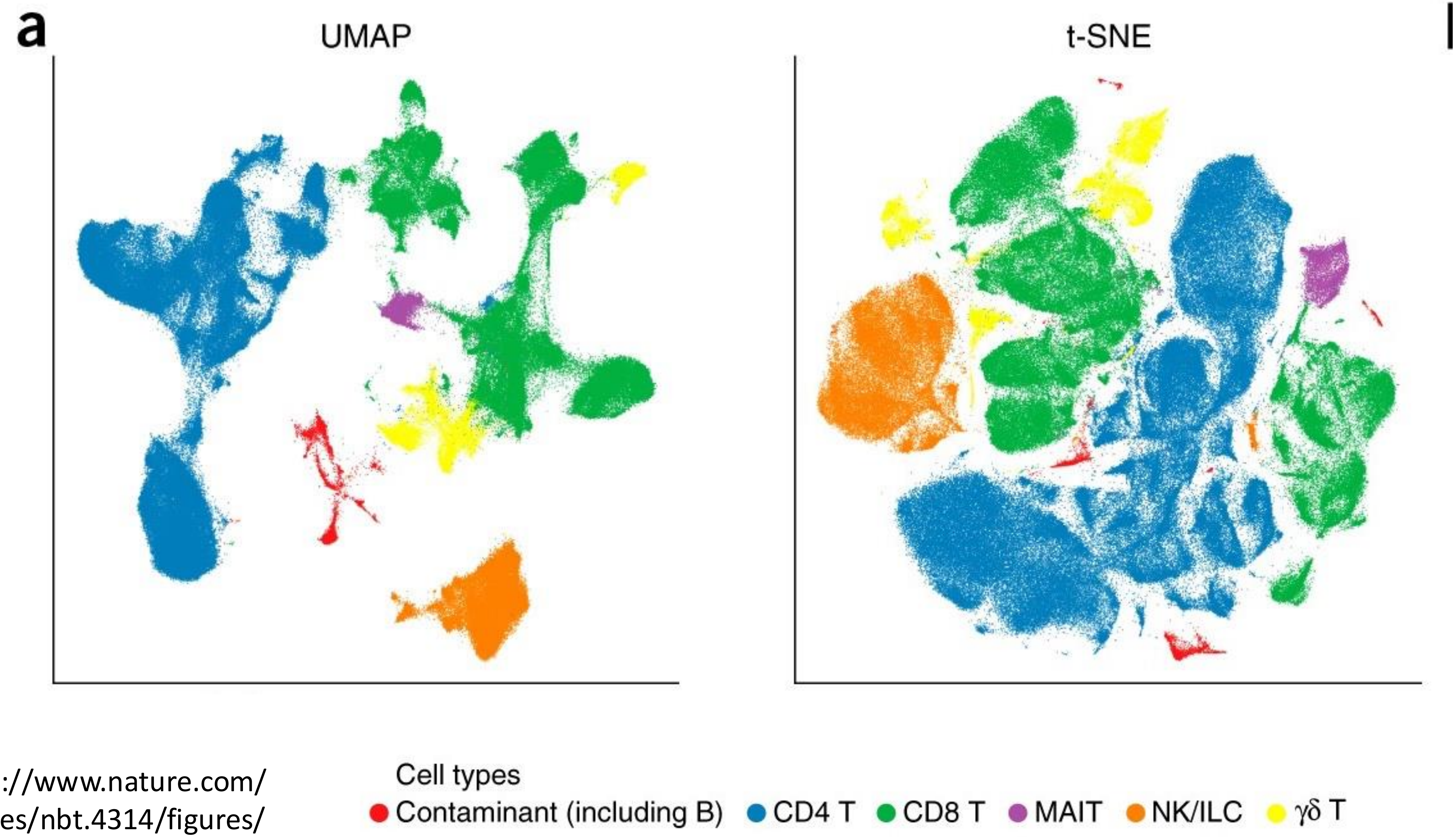
Data



Cell types

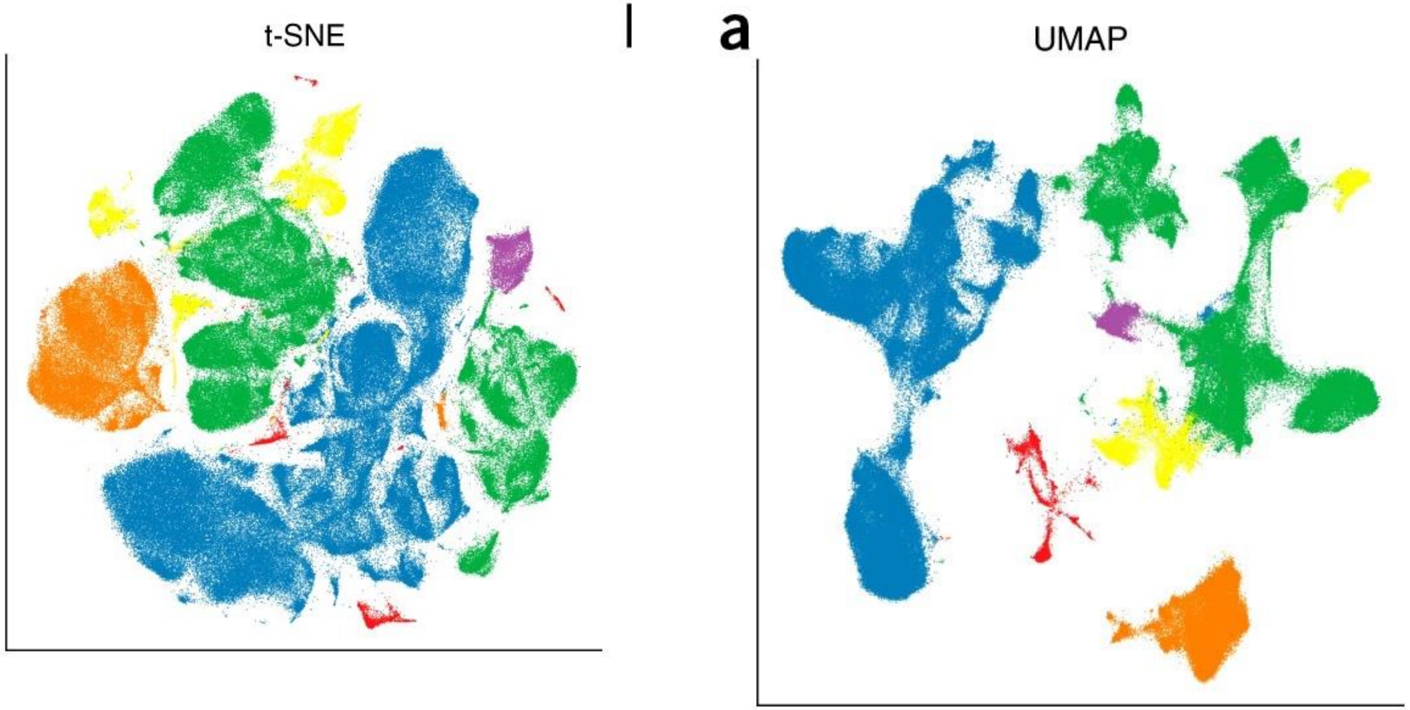
● Contaminant (including B) ● CD4 T ● CD8 T ● MAIT ● NK/ILC ● γδ T

https://www.nature.com/articles/nbt.4314/figures/1



https://www.nature.com/articles/nbt.4314/figures/1

UMAP and t-SNE



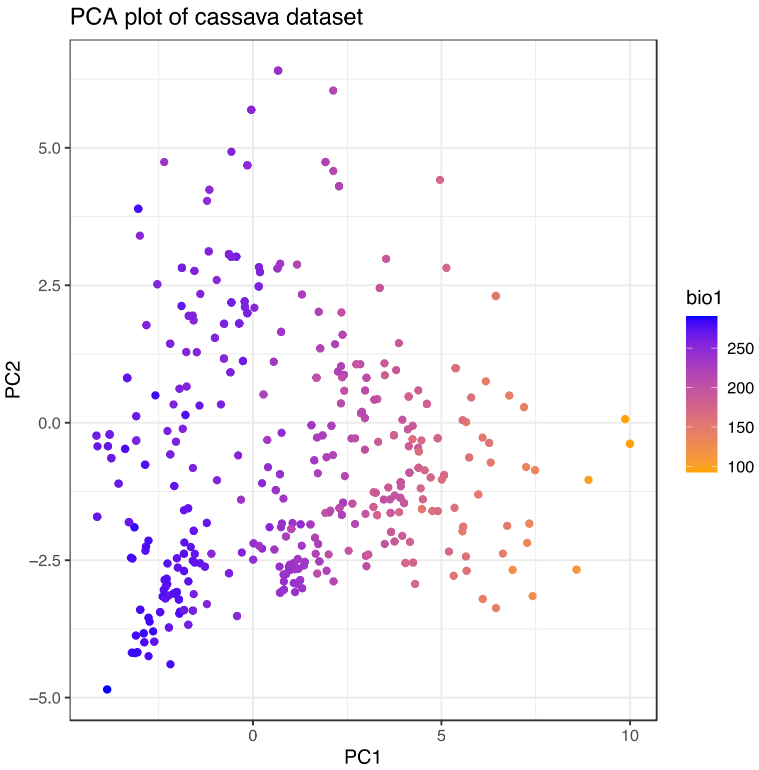
- Usually used for very high dimension data
- Meaning of axes are not inherently informative about effects of individual variables
- In contrast, PCA returns loadings which inform about the effect of variables on axes

PCA vs UMAP vs t-SNE example

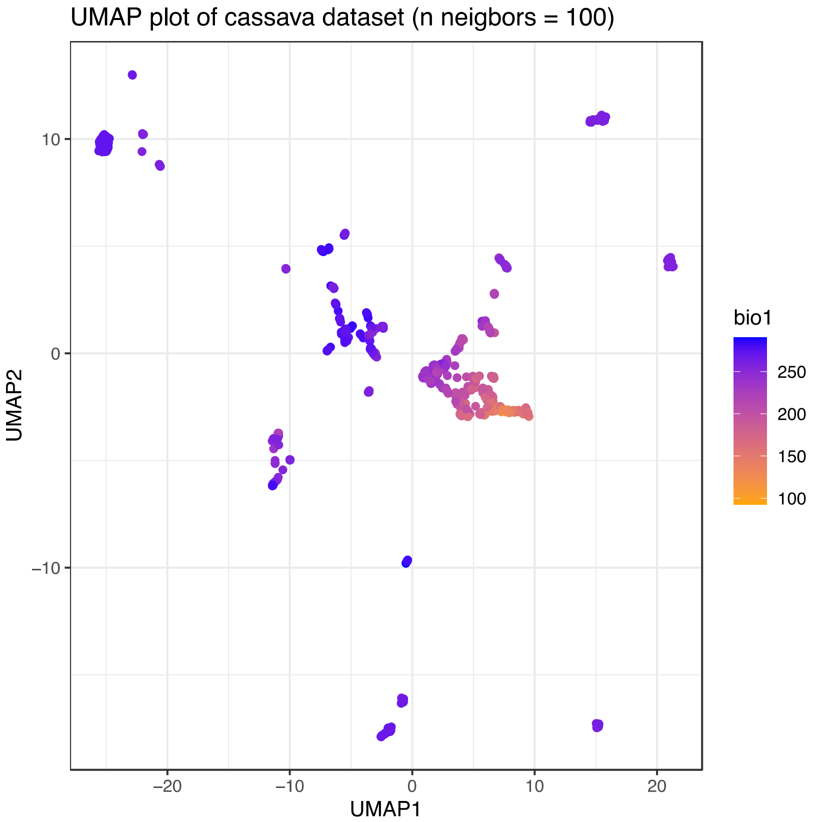
Cassava dataset: points colored by Mean Annual Temperature

You will notice that UMAP and t-SNE are biased toward identify clusters

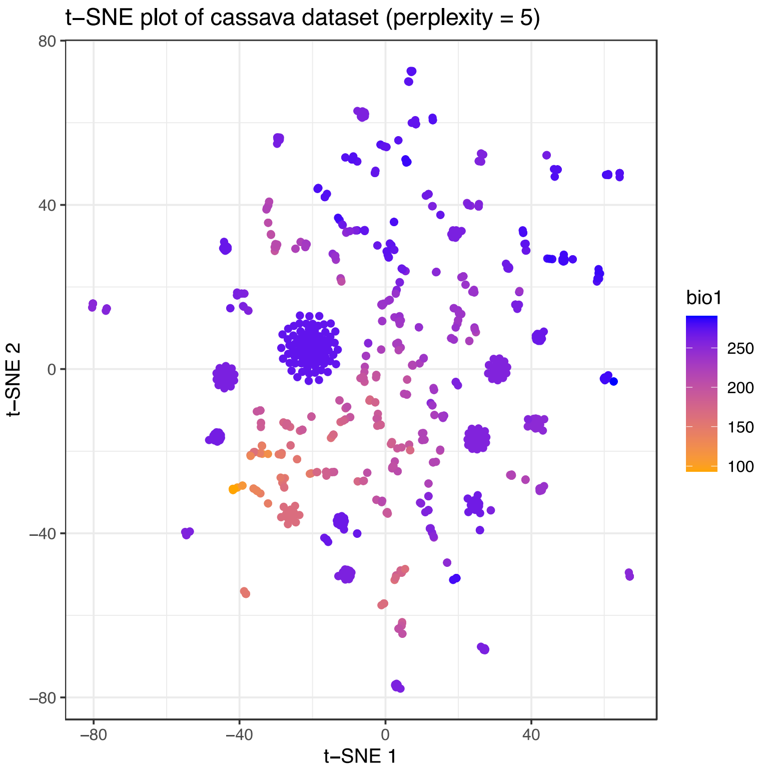
PCA



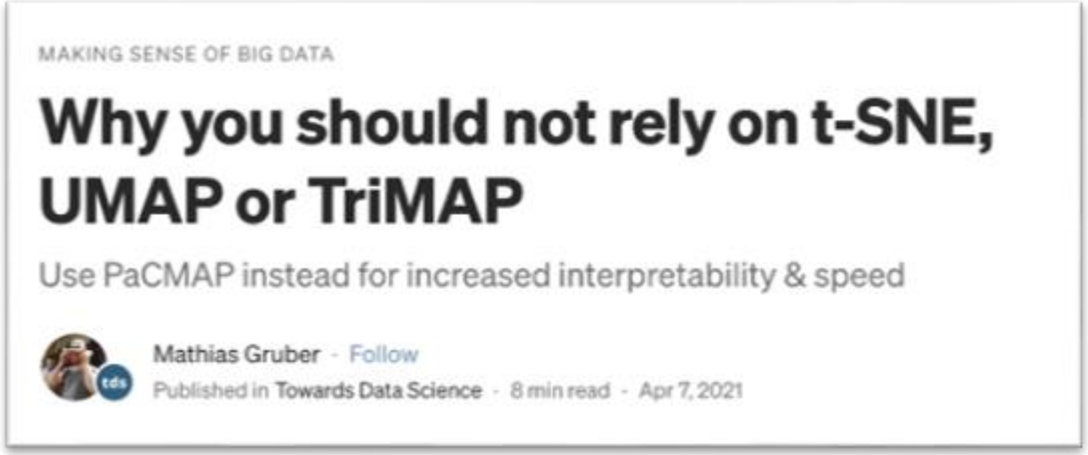
UMAP



t-SNE



Caution!



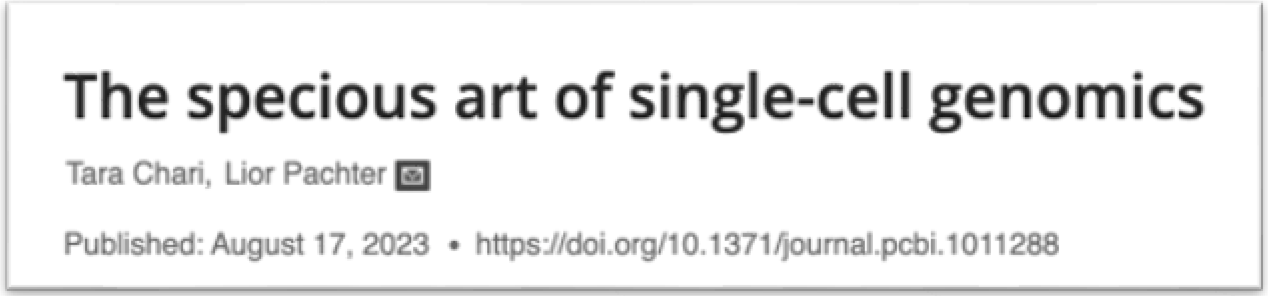
MAKING SENSE OF BIG DATA

Why you should not rely on t-SNE, UMAP or TriMAP

Use PaCMAP instead for increased interpretability & speed

Mathias Gruber · Follow

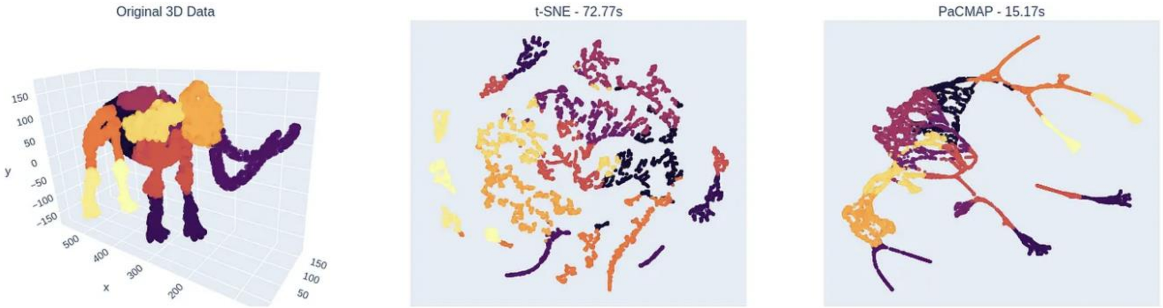
Published in Towards Data Science · 8 min read · Apr 7, 2021



The specious art of single-cell genomics

Tara Chari, Lior Pachter

Published: August 17, 2023 • https://doi.org/10.1371/journal.pcbi.1011288



Plot created by author