

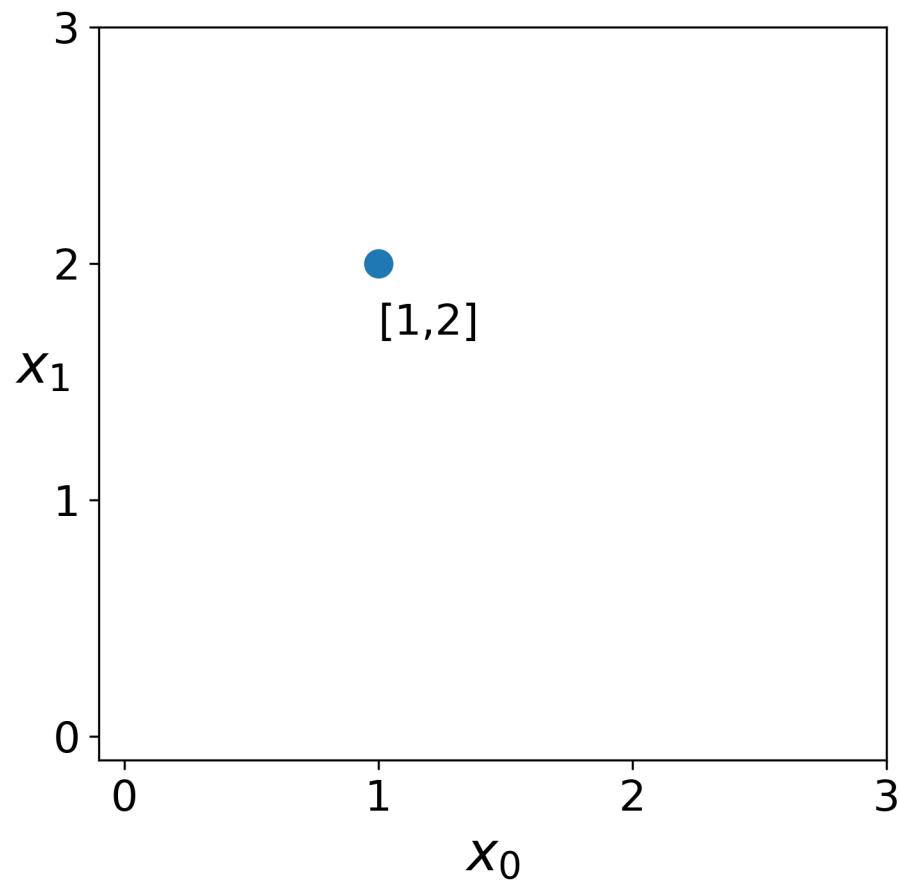
Thinking about high dimensional data

Machine Learning for Single Cell Analysis

Day 1 – Morning session

What does “dimensionality” mean?

\mathbb{R}^2 two-dimensional space



What does “dimensionality” mean?

vector coordinate space

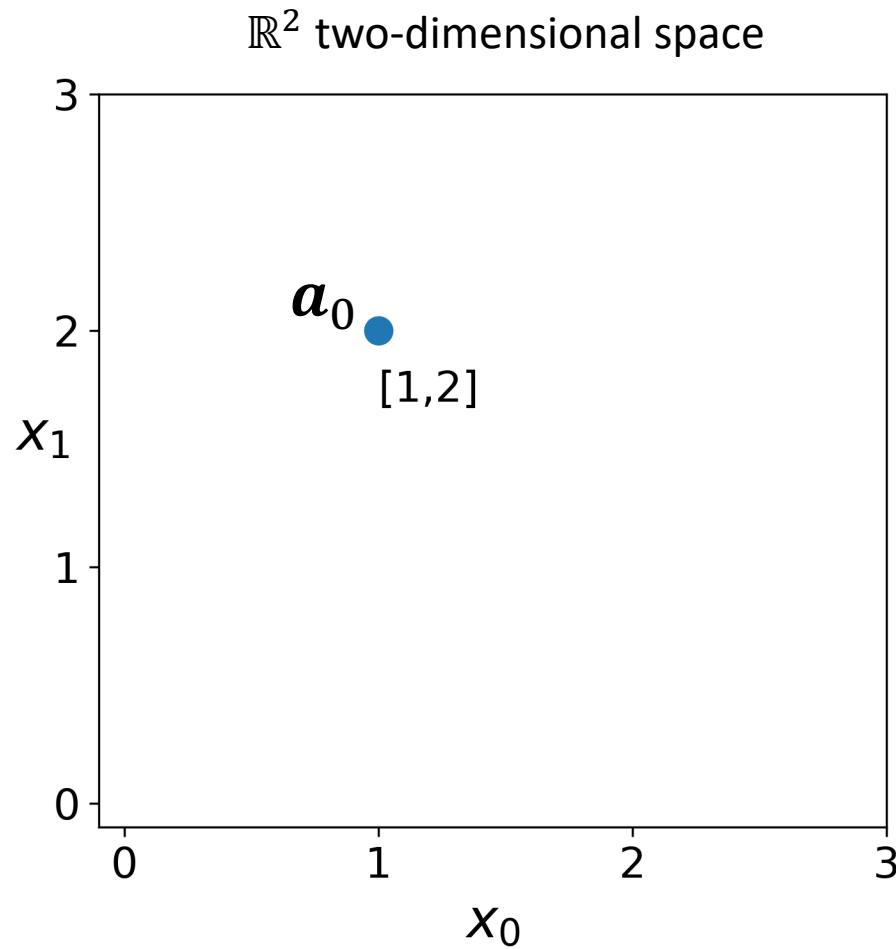
$a_0 \in \mathbb{R}^2$

$a_0 = [1, 2]$

$x_0 \quad x_1$

features

$a_{[:,0]}$



What does “dimensionality” mean?

vector coordinate space

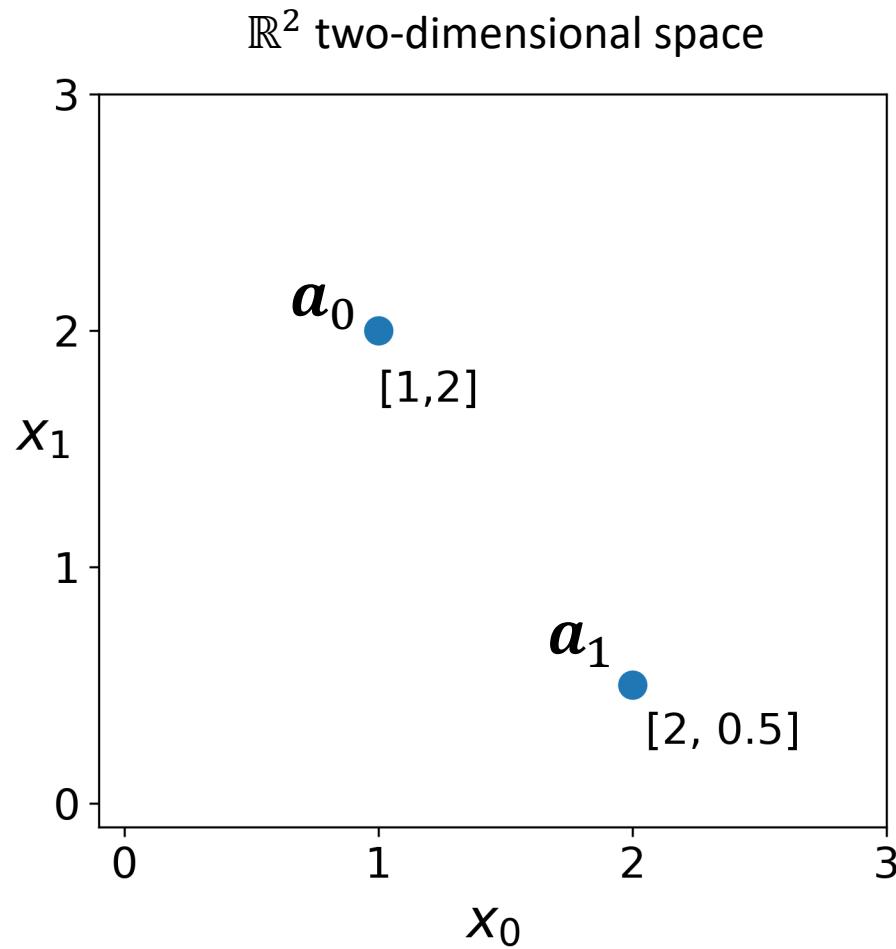
$a_0 \in \mathbb{R}^2$

$a_0 = [1, 2]$

$x_0 \quad x_1$

features

$a_{[:,0]}$



What does “dimensionality” mean?

$$a \in \mathbb{R}^2$$

$$a = [1, 2]$$

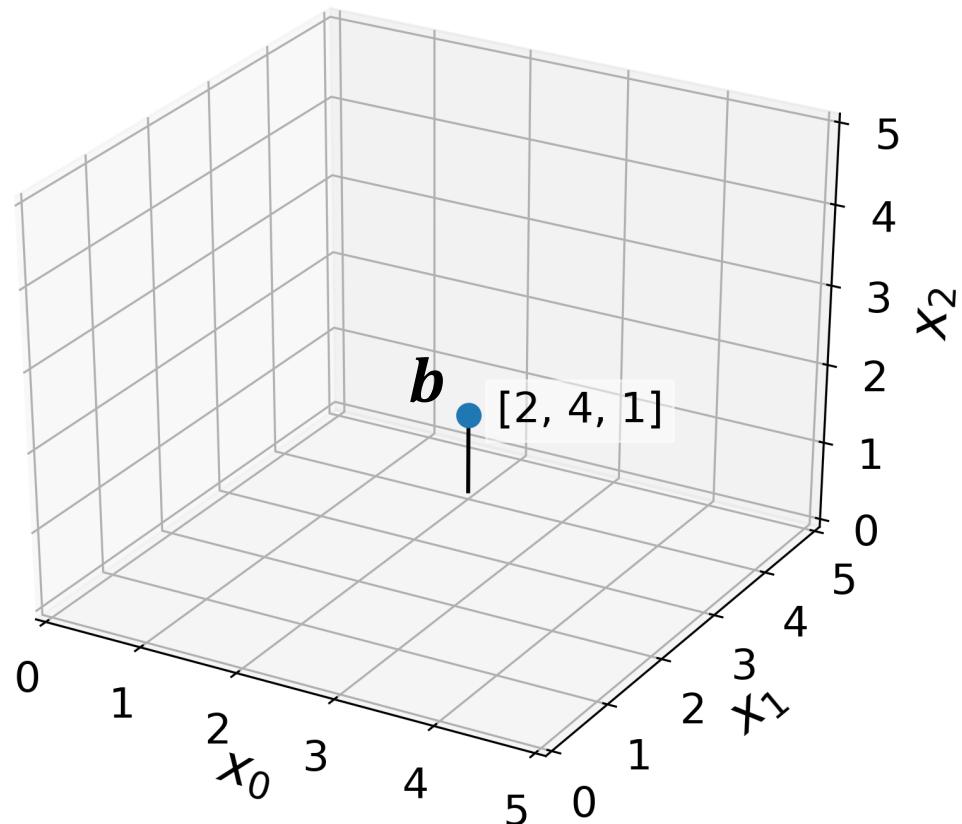
x_0 x_1

$$b \in \mathbb{R}^3$$

$$b = [2, 4, 1]$$

x_0 x_1 x_2

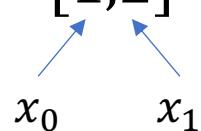
\mathbb{R}^3 three-dimensional space



What does “dimensionality” mean?

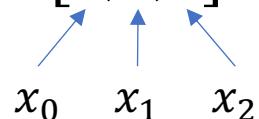
$$\mathbf{a} \in \mathbb{R}^2$$

$$\mathbf{a} = [1, 2]$$



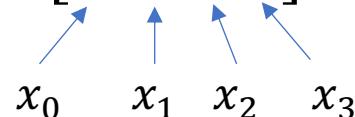
$$\mathbf{b} \in \mathbb{R}^3$$

$$\mathbf{b} = [2, 4, 1]$$

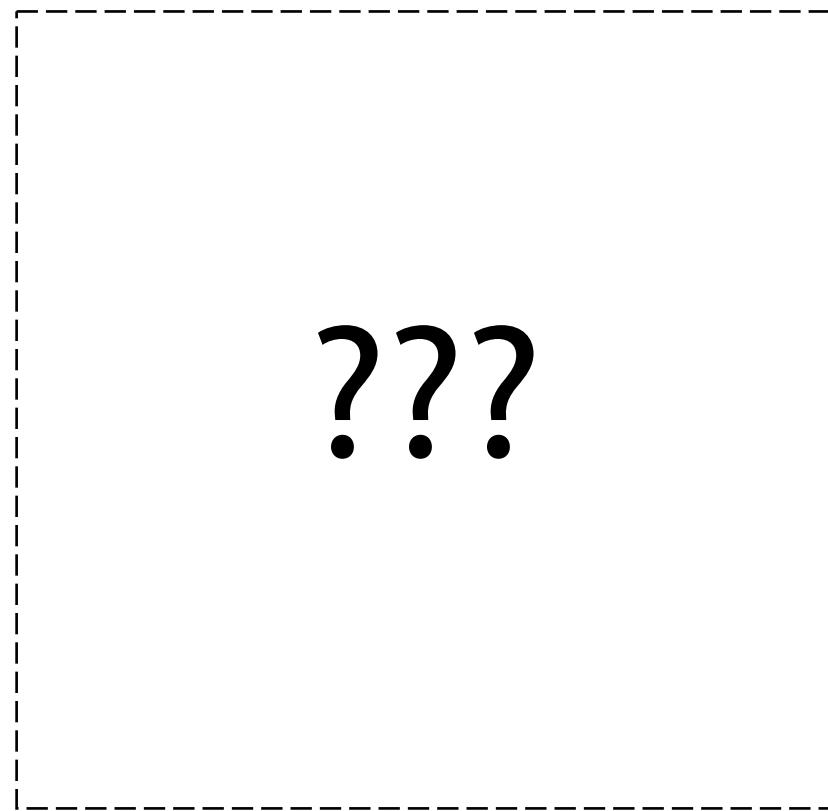


$$\mathbf{c} \in \mathbb{R}^3$$

$$\mathbf{c} = [2, 4, 1, 1]$$



\mathbb{R}^4 four-dimensional space



What does “dimensionality” mean?

$$a \in \mathbb{R}^2$$

$$a = [1, 2]$$

x_0 x_1

$$b \in \mathbb{R}^3$$

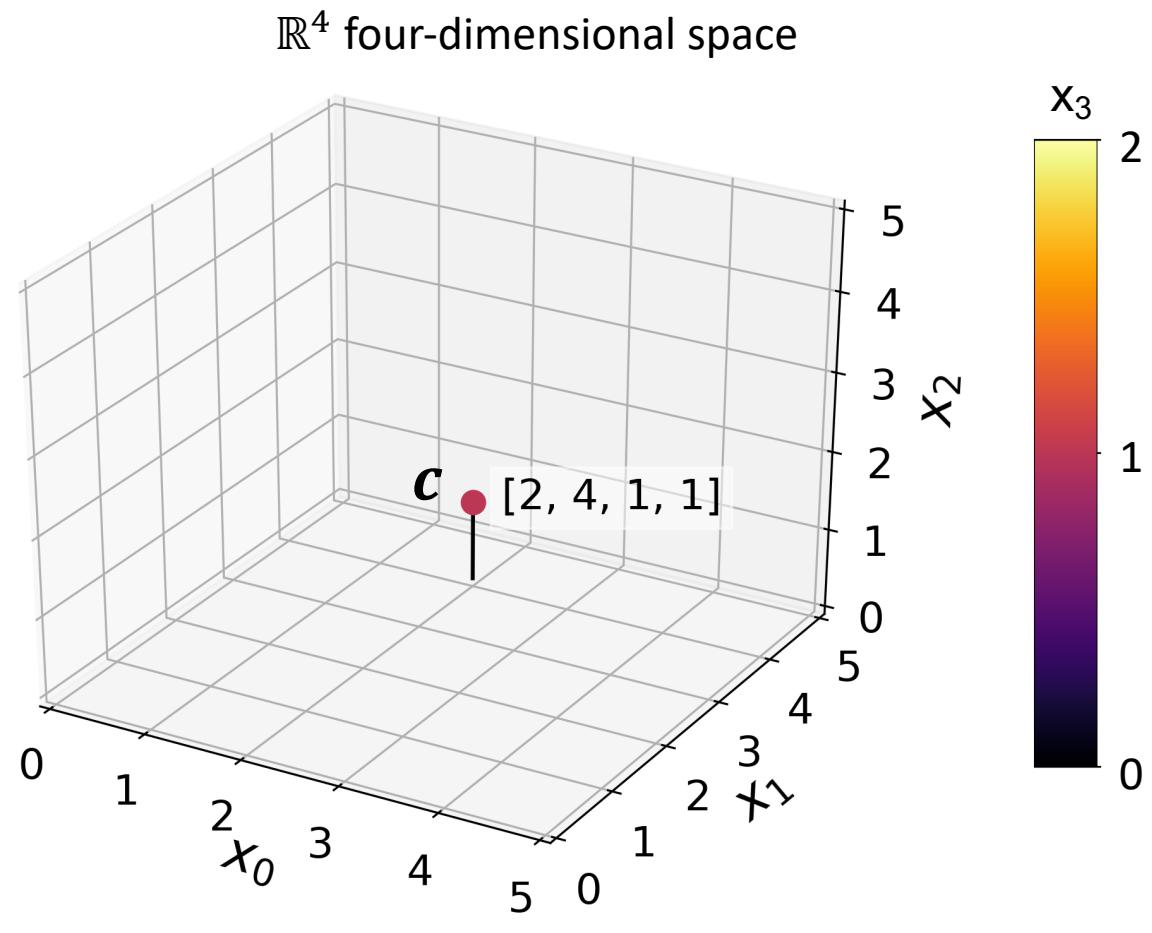
$$b = [2, 4, 1]$$

x_0 x_1 x_2

$$c \in \mathbb{R}^3$$

$$c = [2, 4, 1, 1]$$

x_0 x_1 x_2 x_3



What does “dimensionality” mean?

$$\mathbf{a} \in \mathbb{R}^2$$

$$\mathbf{a} = [1, 2]$$

$x_1 \quad x_2$

$$\mathbf{b} \in \mathbb{R}^3$$

$$\mathbf{b} = [2, 4, 1]$$

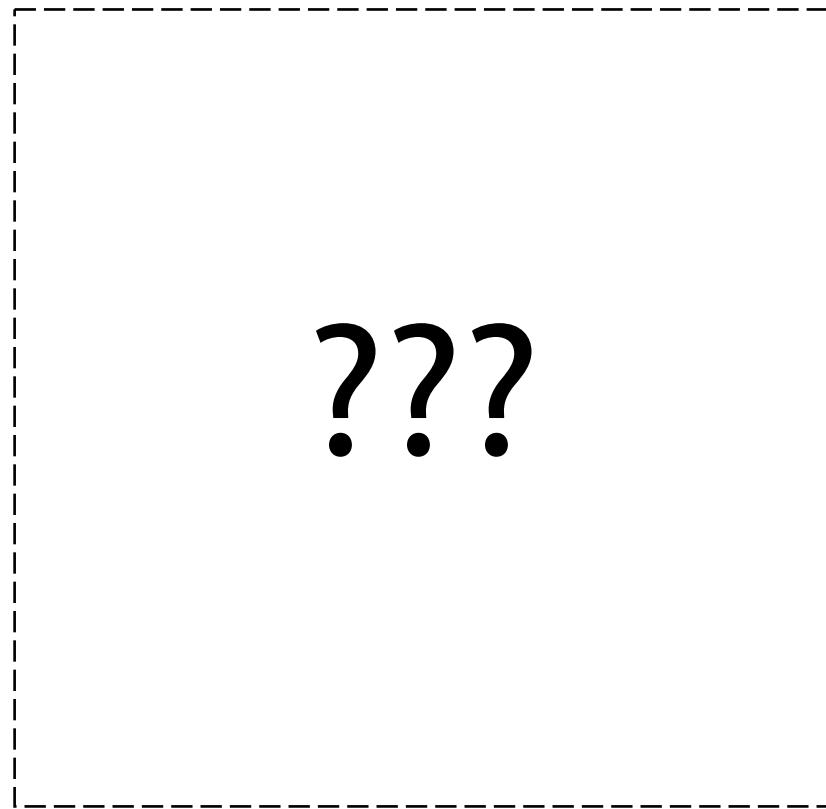
$x_1 \quad x_2 \quad x_3$

$$\mathbf{c} \in \mathbb{R}^3$$

$$\mathbf{c} = [2, 4, 1, 1]$$

$x_1 \quad x_2 \quad x_3 \quad x_4$

\mathbb{R}^5 five-dimensional space



What do x_1 x_2 , ..., x_n represent?

We describe coordinate spaces of n dimensions as \mathbb{R}^n

$$a = \begin{array}{c} \begin{matrix} & \text{alcohol} & \text{hue} \\ \text{Wine0} & 14.23 & 1.04 \end{matrix} \\ \begin{matrix} a_0 & & a_1 \end{matrix} \end{array}$$
$$b = \begin{array}{c} \begin{matrix} & \text{Age} & \text{Height (m)} & \text{Weight (kg)} \\ \text{Player0} & 23 & 2.31 & 91 \end{matrix} \\ \begin{matrix} b_0 & b_1 & b_2 \end{matrix} \end{array}$$
$$c = \begin{array}{c} \begin{matrix} & \text{sepal length (cm)} & \text{sepal width (cm)} & \text{petal length (cm)} & \text{petal width (cm)} \\ \text{Iris0} & 5.1 & 3.5 & 1.4 & 0.2 \end{matrix} \end{array}$$

What if we have more than one observation?

$$a = \begin{array}{c} \text{alcohol} & \text{hue} \\ \hline \text{Wine0} & 14.23 & 1.04 \\ \end{array}$$

a_0 a_1

$A =$
matrix

columns
(features)

rows
(observations)

	alcohol	hue
Wine0	14.23	1.04
Wine1	13.20	1.05
Wine2	13.16	1.03
Wine3	14.37	0.86
Wine4	13.24	1.04
Wine5	14.20	1.05
Wine6	14.39	1.02
Wine7	14.06	1.06
Wine8	14.83	1.08
Wine9	13.86	1.01

What if we have more than one observation?

$$a = \begin{array}{c} \text{alcohol} & \text{hue} \\ \hline \text{Wine0} & 14.23 & 1.04 \\ \end{array}$$

a_0 a_1

$A =$

matrix

$\in \mathbb{R}^2$

	columns (features)	
	alcohol	hue
Wine0	14.23	1.04
Wine1	13.20	1.05
Wine2	13.16	1.03
Wine3	14.37	0.86
Wine4	13.24	1.04
Wine5	14.20	1.05
Wine6	14.39	1.02
Wine7	14.06	1.06
Wine8	14.83	1.08
Wine9	13.86	1.01

How do we reference specific observations in a matrix?

$A =$

	x_0	x_1	
	alcohol	hue	
Wine0	14.23	1.04	$A[0,:]$
Wine1	13.20	1.05	
Wine2	13.16	1.03	
Wine3	14.37	0.86	$A[\text{row}, \text{column}]$
Wine4	13.24	1.04	
Wine5	14.20	1.05	
Wine6	14.39	1.02	
Wine7	14.06	1.06	
Wine8	14.83	1.08	
Wine9	13.86	1.01	

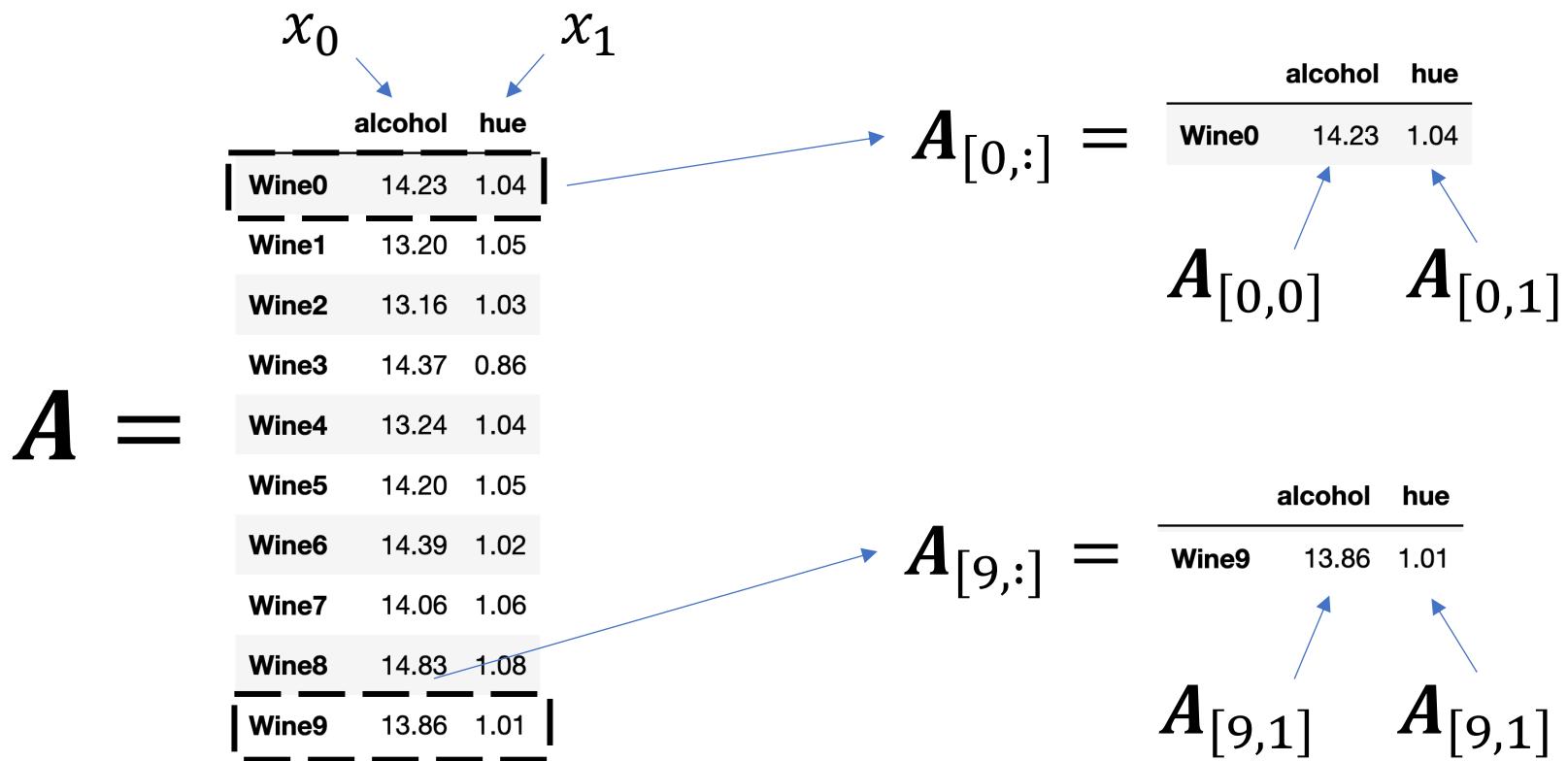
What if we have more than one observation?

We can describe coordinate spaces of n dimensions as \mathbb{R}^n

$$A = \begin{array}{c} \begin{matrix} & x_0 & & x_1 \\ & \downarrow & & \downarrow \\ \begin{matrix} & \text{alcohol} & \text{hue} \\ \hline \text{Wine0} & 14.23 & 1.04 \\ \text{Wine1} & 13.20 & 1.05 \\ \text{Wine2} & 13.16 & 1.03 \\ \text{Wine3} & 14.37 & 0.86 \\ \text{Wine4} & 13.24 & 1.04 \\ \text{Wine5} & 14.20 & 1.05 \\ \text{Wine6} & 14.39 & 1.02 \\ \text{Wine7} & 14.06 & 1.06 \\ \text{Wine8} & 14.83 & 1.08 \\ \text{Wine9} & 13.86 & 1.01 \end{matrix} & \xrightarrow{\quad} & A_{[0,:]} = a_0 = \begin{matrix} & \text{alcohol} & \text{hue} \\ \hline \text{Wine0} & 14.23 & 1.04 \end{matrix} \\ A_{[0,0]} & & A_{[0,1]} \end{array}$$

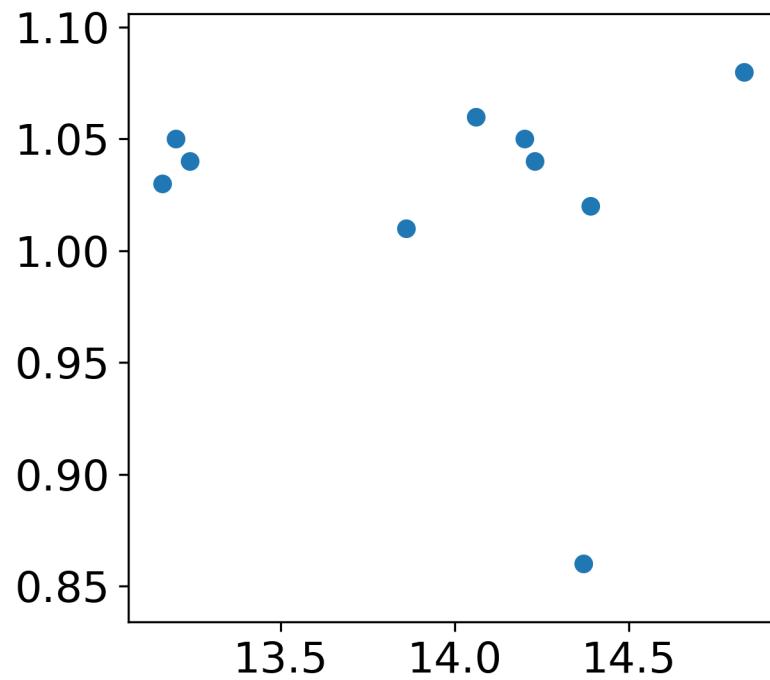
What if we have more than one observation?

We can describe coordinate spaces of n dimensions as \mathbb{R}^n



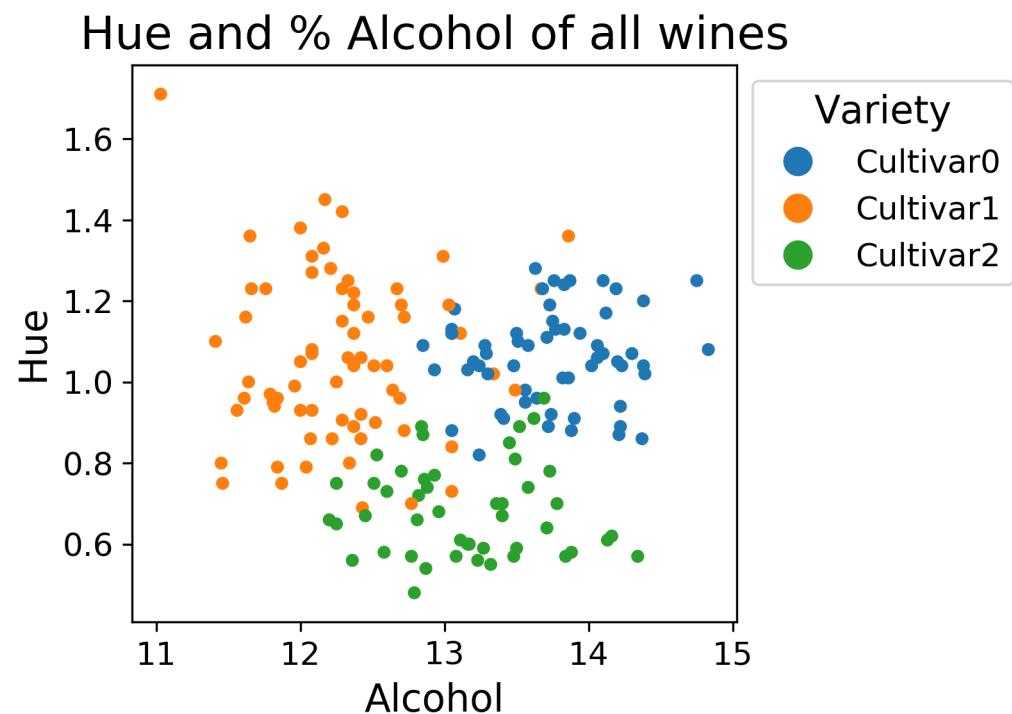
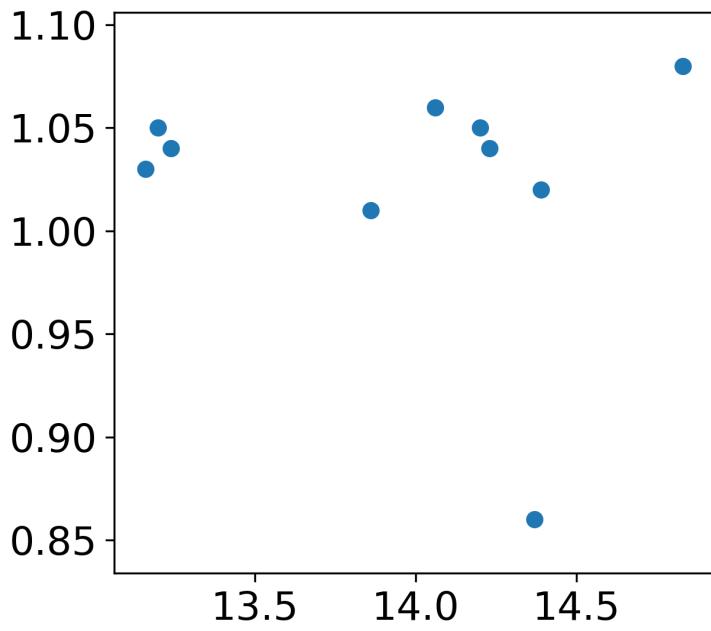
Exercise!

Loading and plotting the UCI wine dataset



Exercise!

Loading and plotting the UCI wine dataset

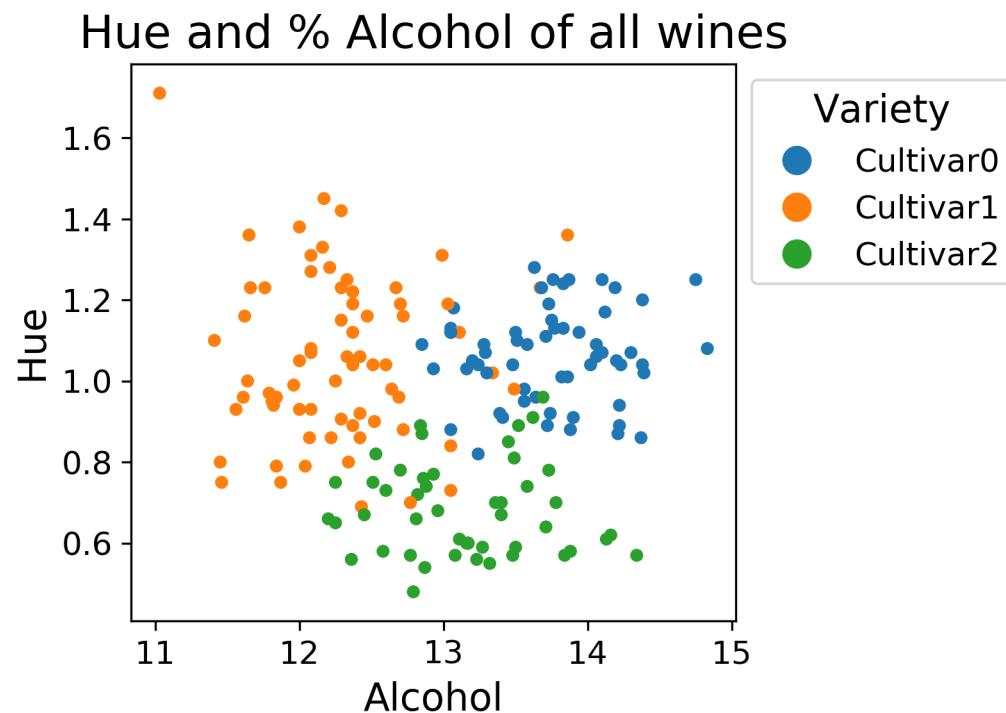


Introducing scprep

- Code: <https://github.com/KrishnaswamyLab/scprep>
 - Docs: <https://scprep.readthedocs.io>
 - Modules
 - filter
 - io
 - measure
 - normalize
 - **plot**
 - reduce
 - run
 - select
 - stats
 - transform
 - utils
- 
- Plotting functions:
 - histogram
 - jitter
 - marker
 - marker_plot
 - plot_gene_set_expression
 - plot_library_size
 - rotate_scatter3d
 - scatter
 - **scatter2d** = scprep.plot.scatter2d
 - scatter3d
 - scree
 - scree_plot

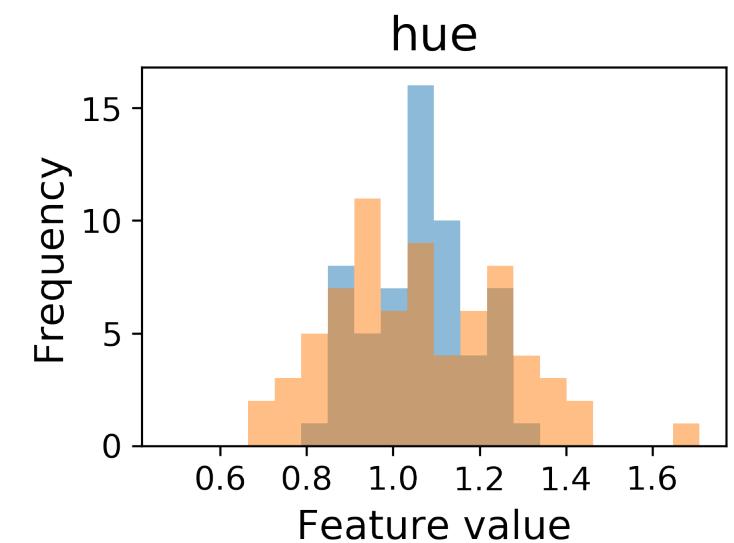
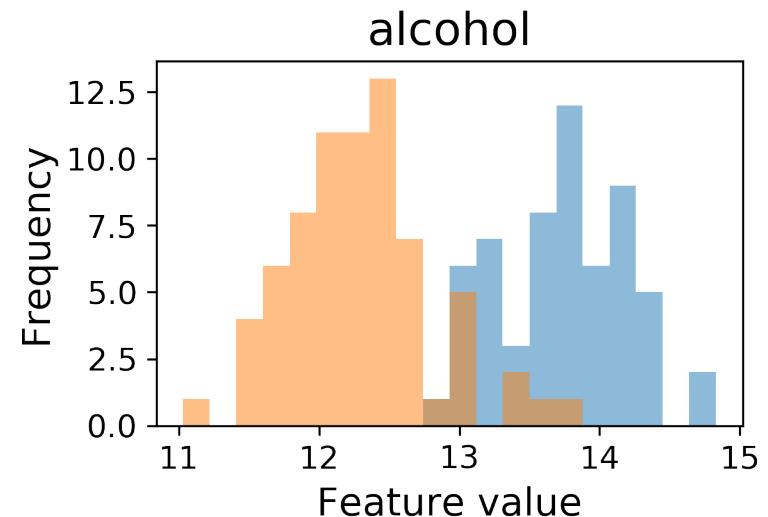
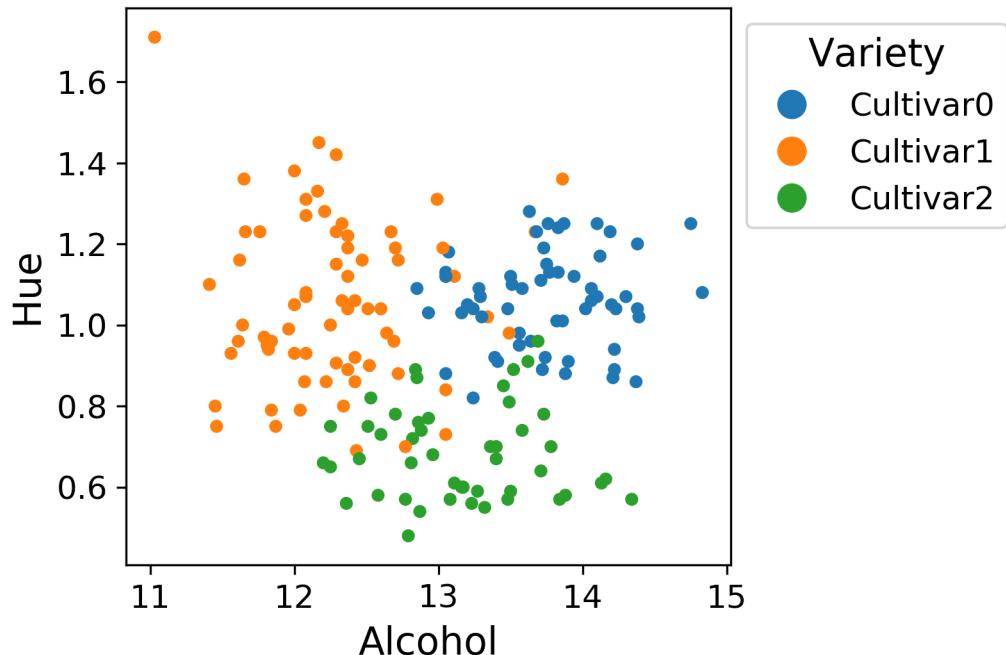
Exercise!

Loading and plotting the UCI wine dataset



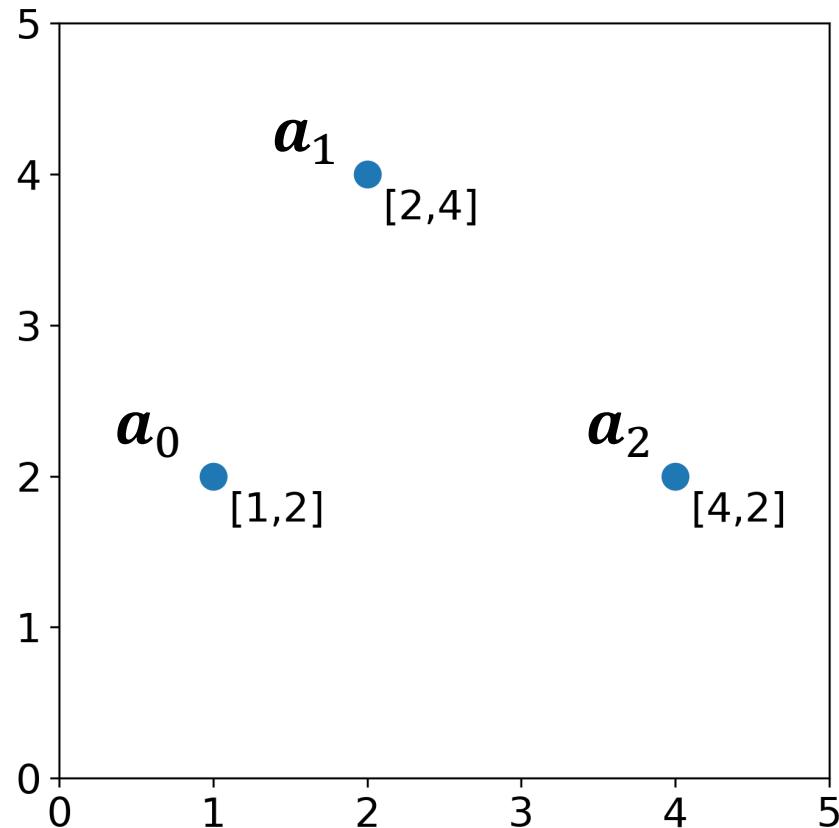
How do features differentiate groups within a dataset?

Hue and % Alcohol of all wines

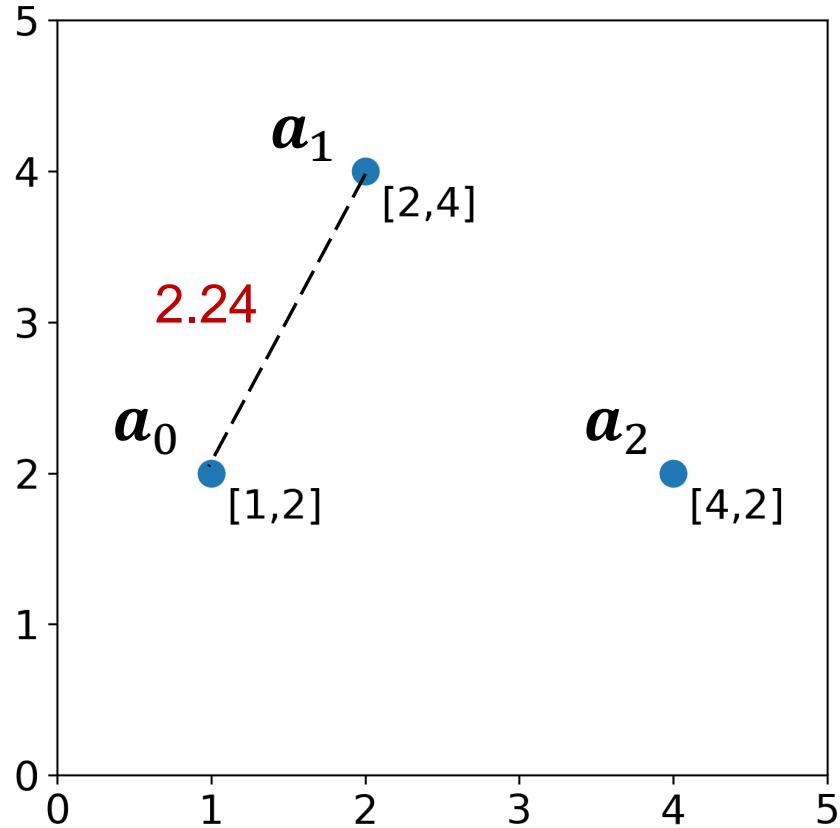


Introduction to Manifold Learning

How far away is each point from a_0 ?



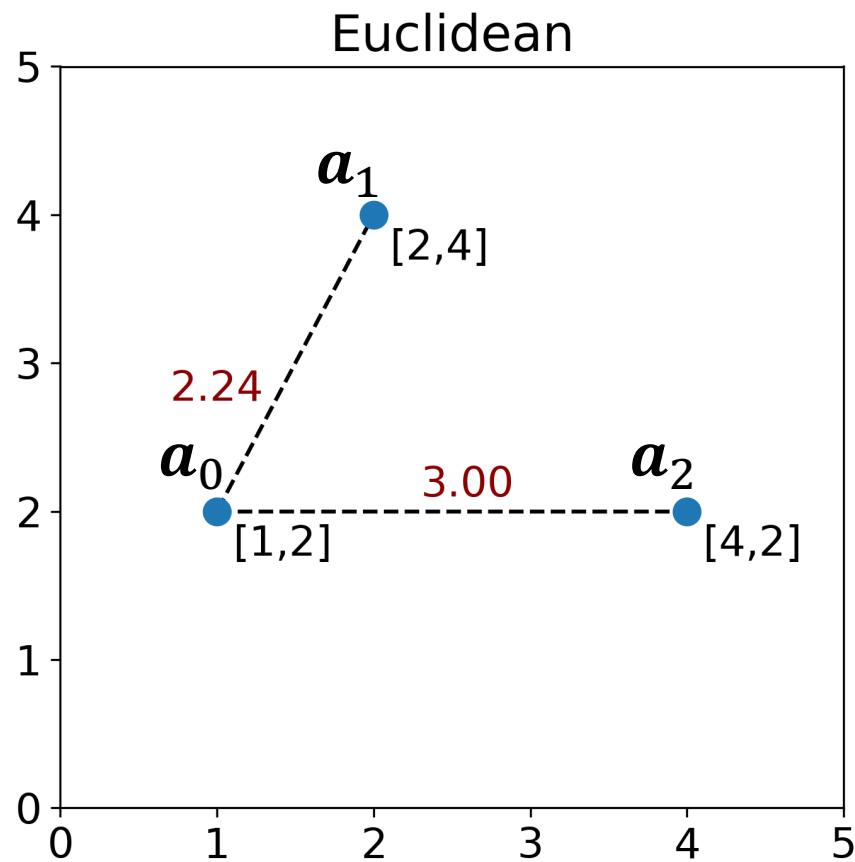
How far away is each point from a_0 ?



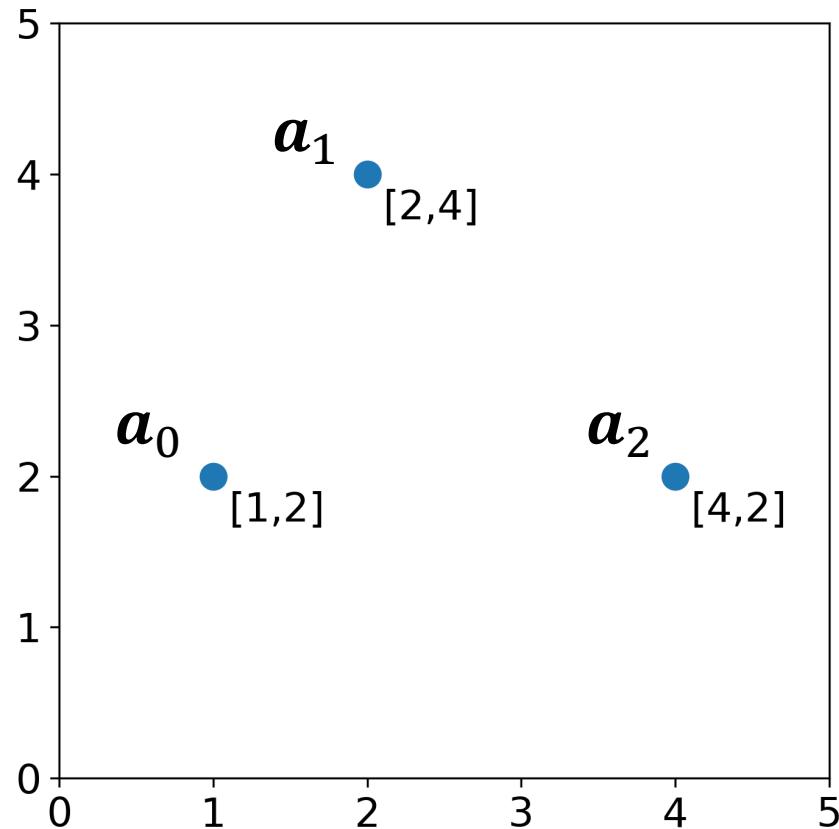
Euclidean distance

$$\begin{aligned}d_{euclidean}(a_0, a_1) &= \|a_0 - a_1\|_2^2 \\&= \sqrt{(a_{0,0} - a_{1,0})^2 + (a_{0,1} - a_{1,1})^2} \\&= \sqrt{(1 - 2)^2 + (2 - 4)^2} \\&\approx 2.24\end{aligned}$$

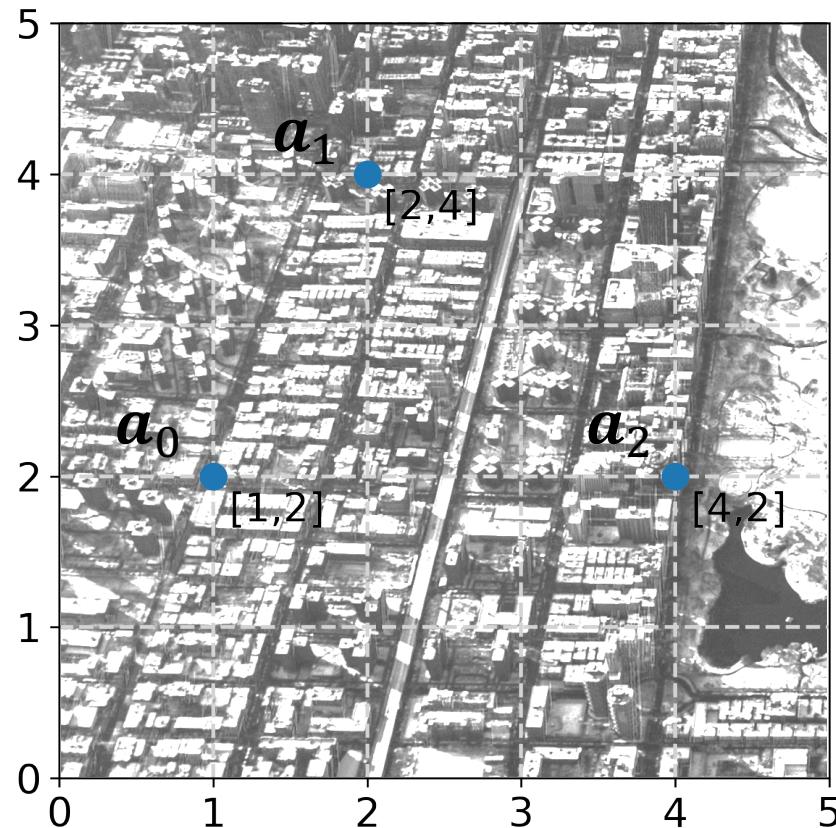
How far away is each point from a_0 ?



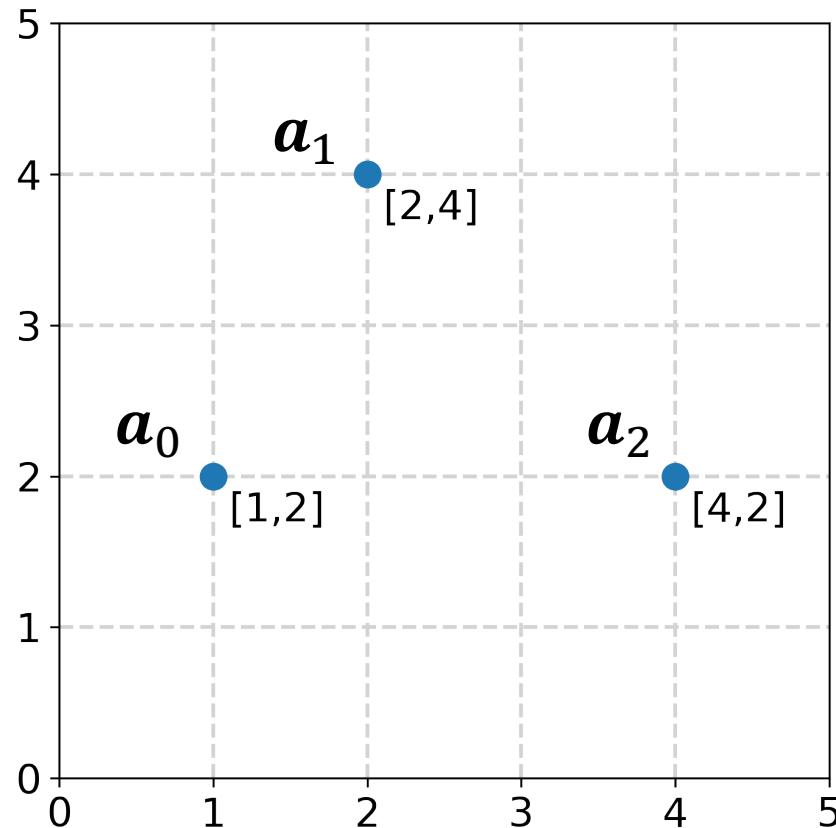
How far away is each point from a_0 ?



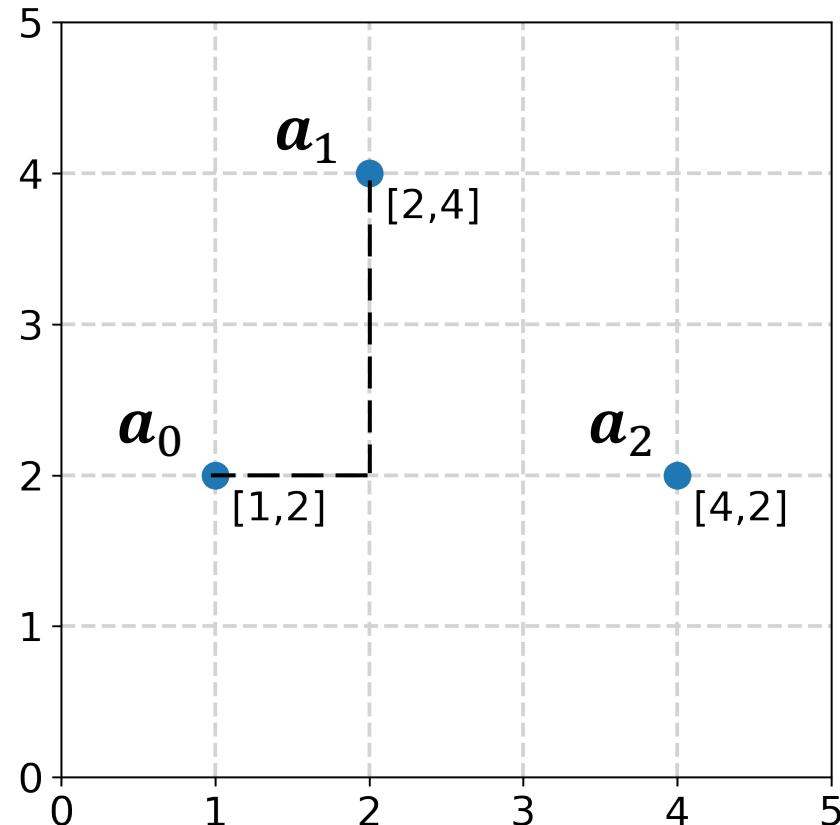
How far away is each point from a_0 ?



How far away is each point from a_0 ?



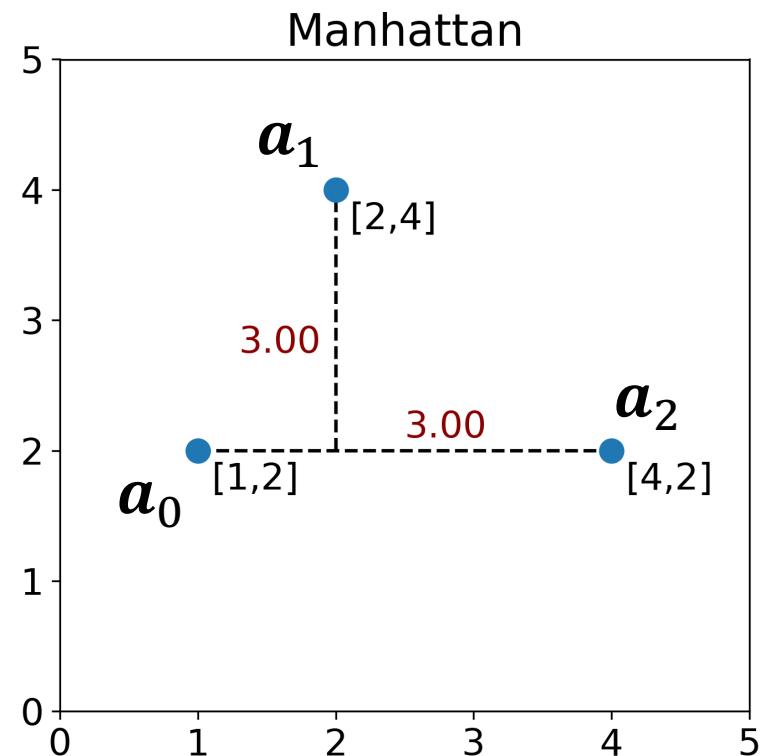
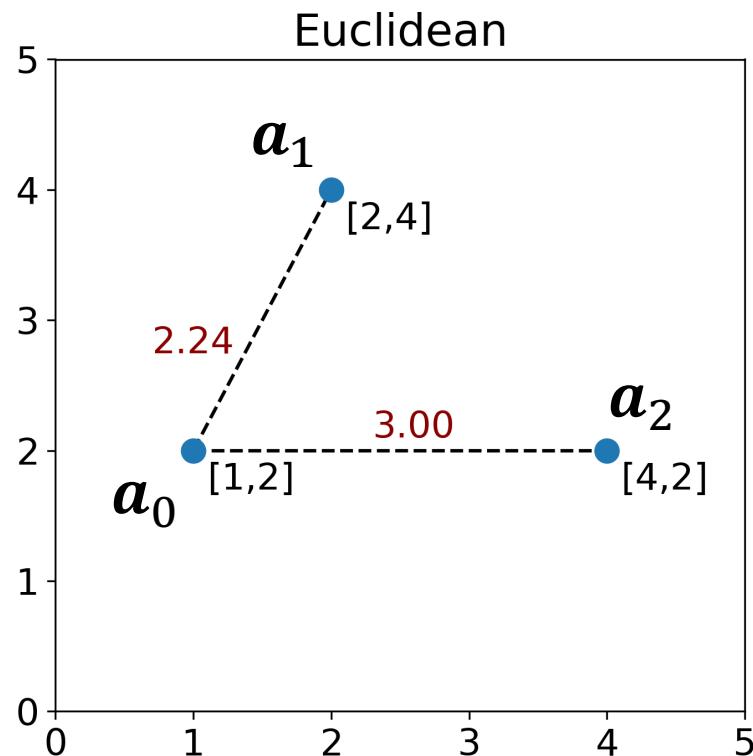
How far away is each point from a_0 ?



Manhattan distance

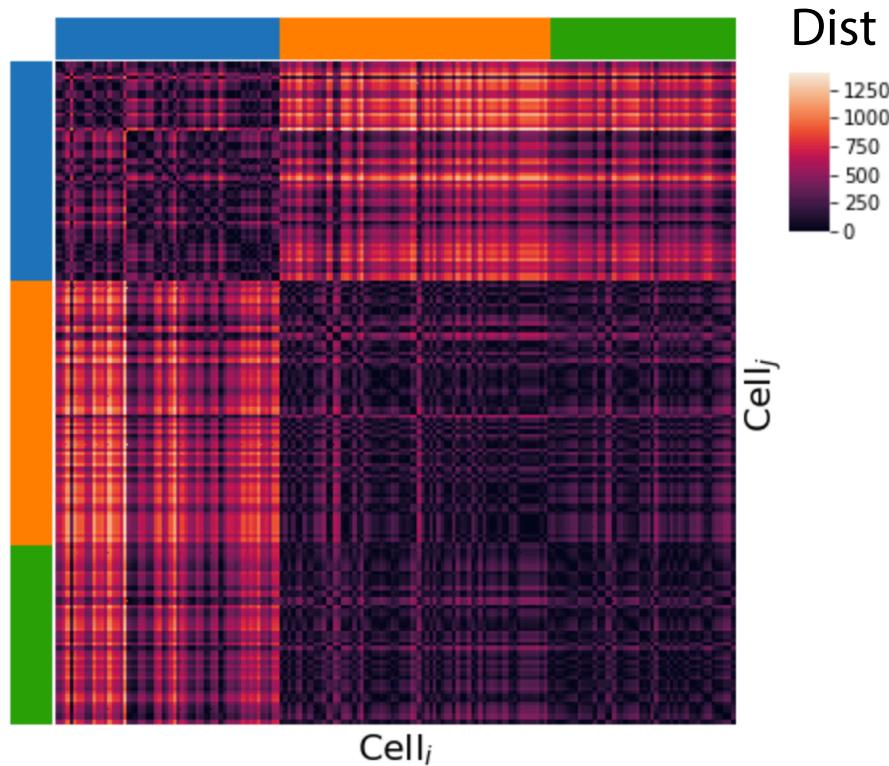
$$\begin{aligned}d_{manhattan}(a_0, a_1) &= |a_{0,1} - a_{1,1}| \\&= |a_{0,0} - a_{1,0}| + |a_{0,1} - a_{1,1}| \\&= |1 - 2| + |2 - 4| \\&= 3\end{aligned}$$

How far away is each point from a_0 ?



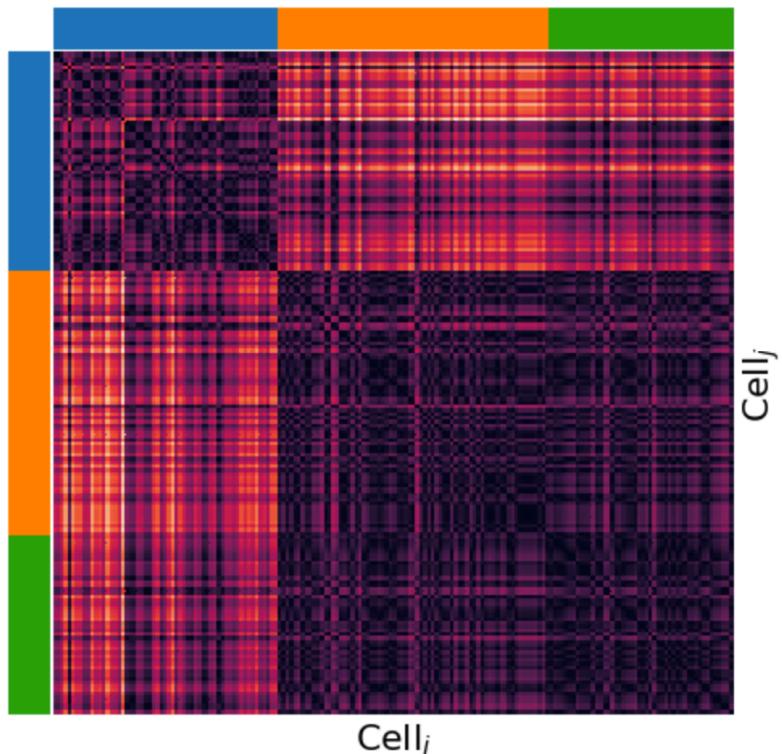
Exercise!

Calculating distances



Exercise!

Calculating distances



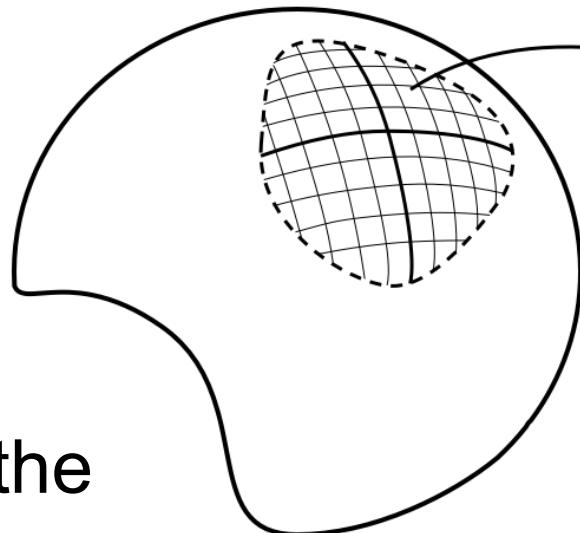
Distance defines
structure in a dataset

- How **different** are two data points?
- How many **types** exist in my dataset?
- Which types are **closest** to each other?
- Which **features** vary between dissimilar groups of data?

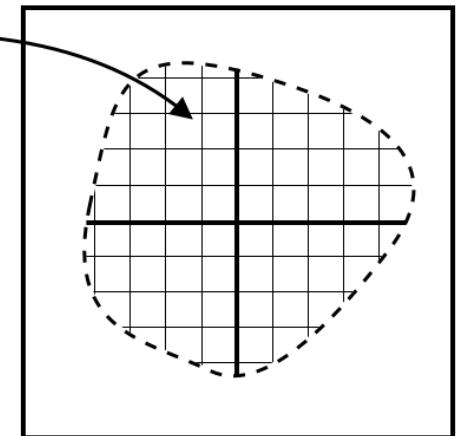
What is a manifold?

- Locally smooth
- Locally Euclidean
- Generally, lower dimensional than the ambient space (i.e. a subspace)

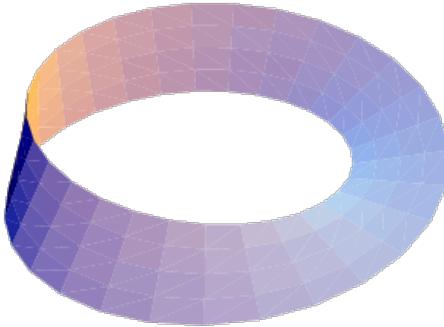
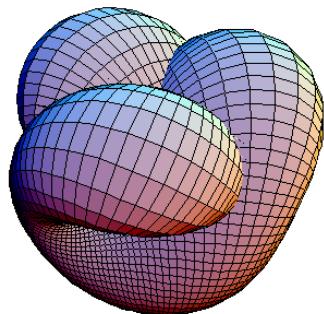
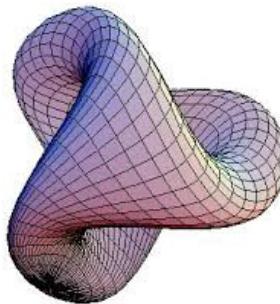
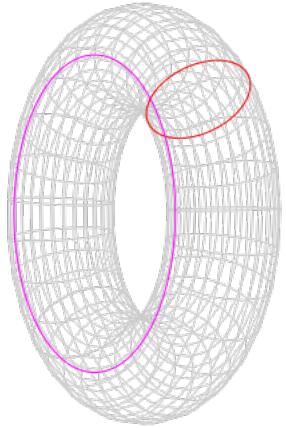
Surface in \mathbb{R}^3



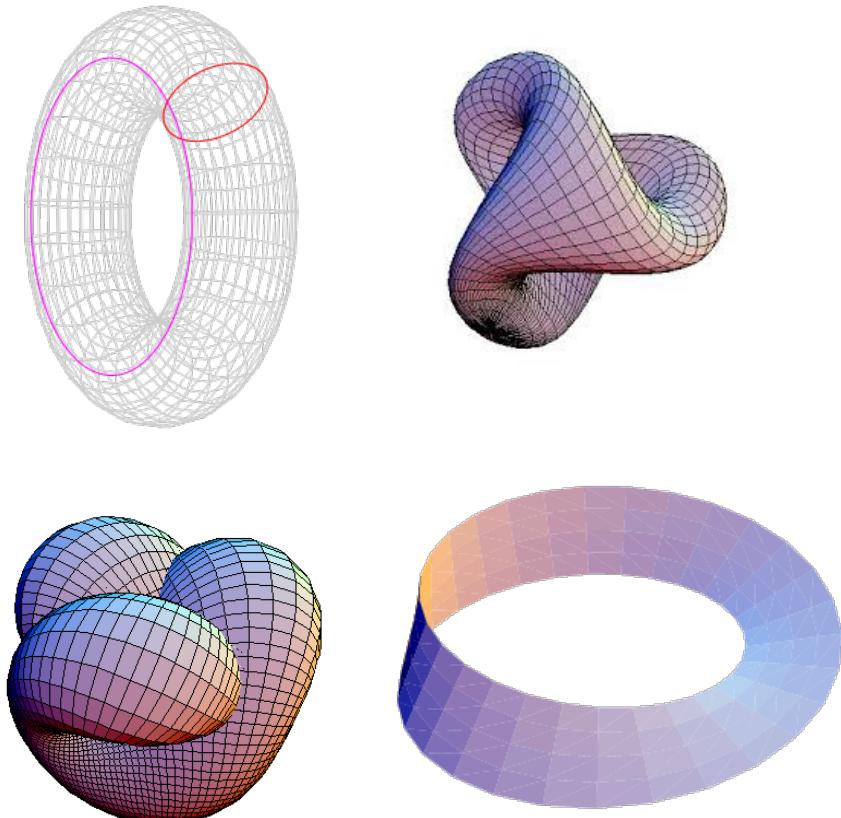
Local view in \mathbb{R}^2



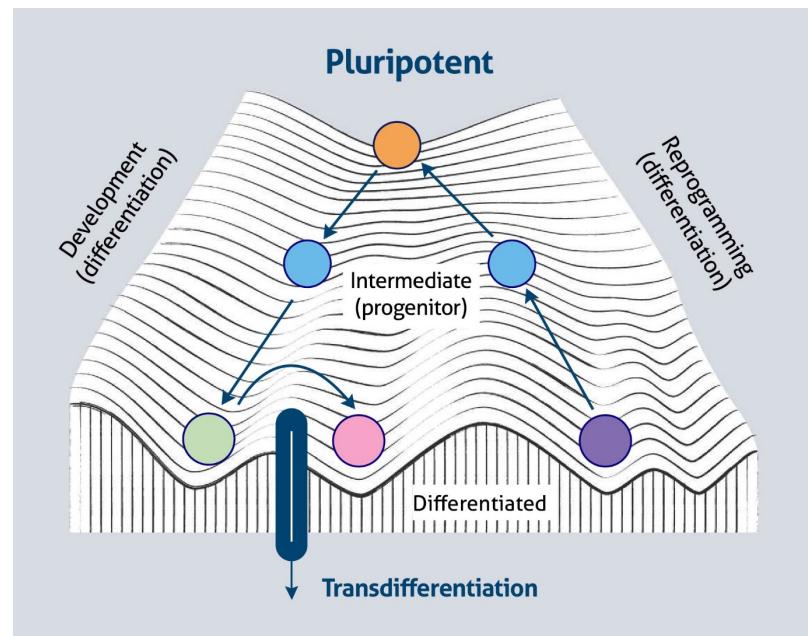
Examples of manifolds



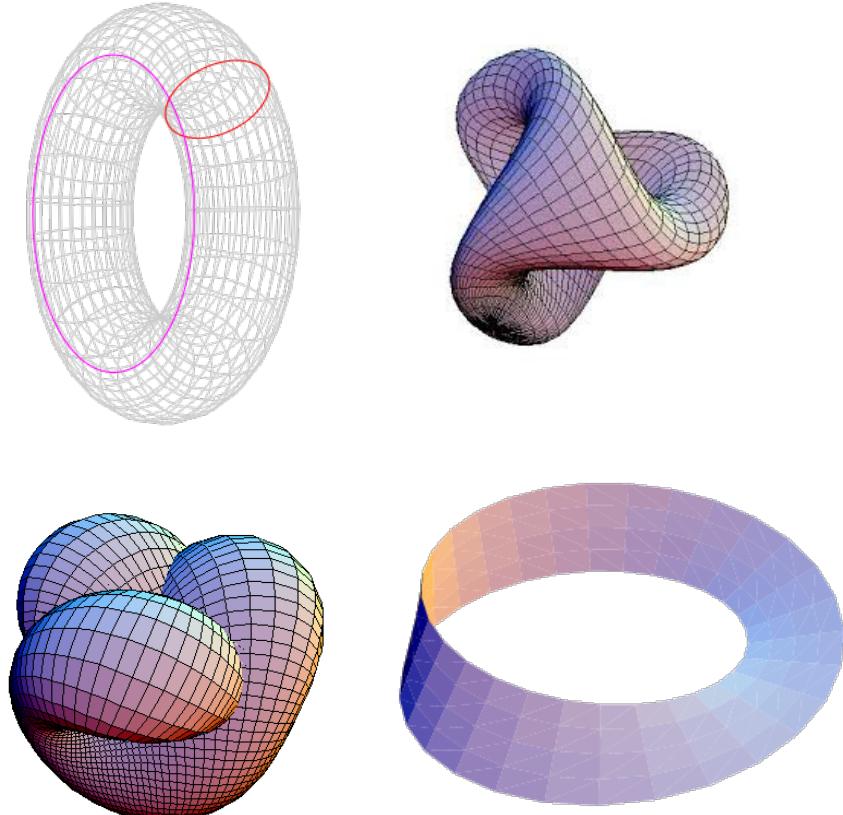
Examples of manifolds



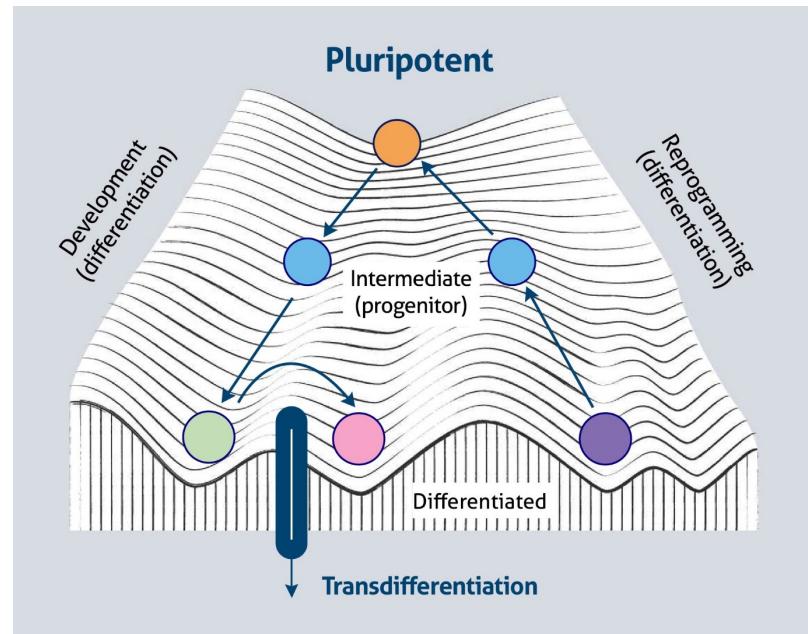
Waddington's landscape



Examples of manifolds



Waddington's landscape



- Smooth transitions between states
- Small changes in gene expression are linear

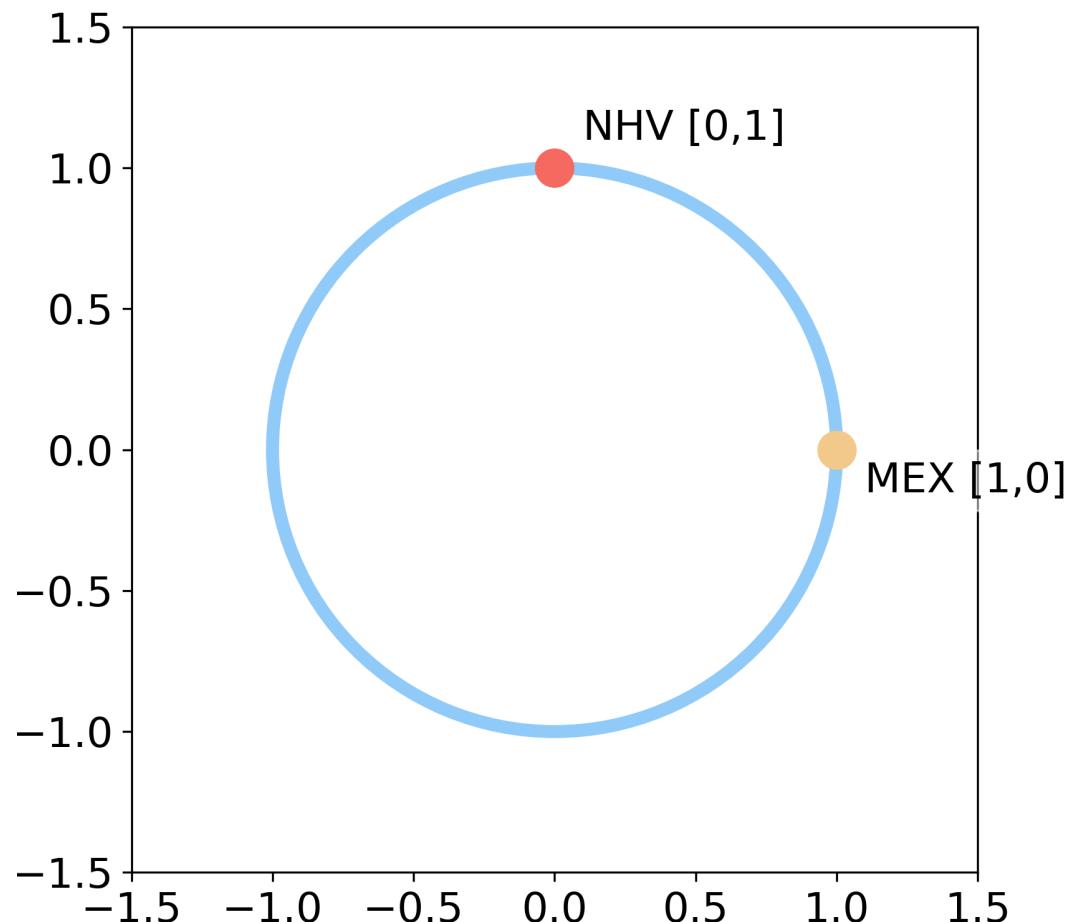
NHV



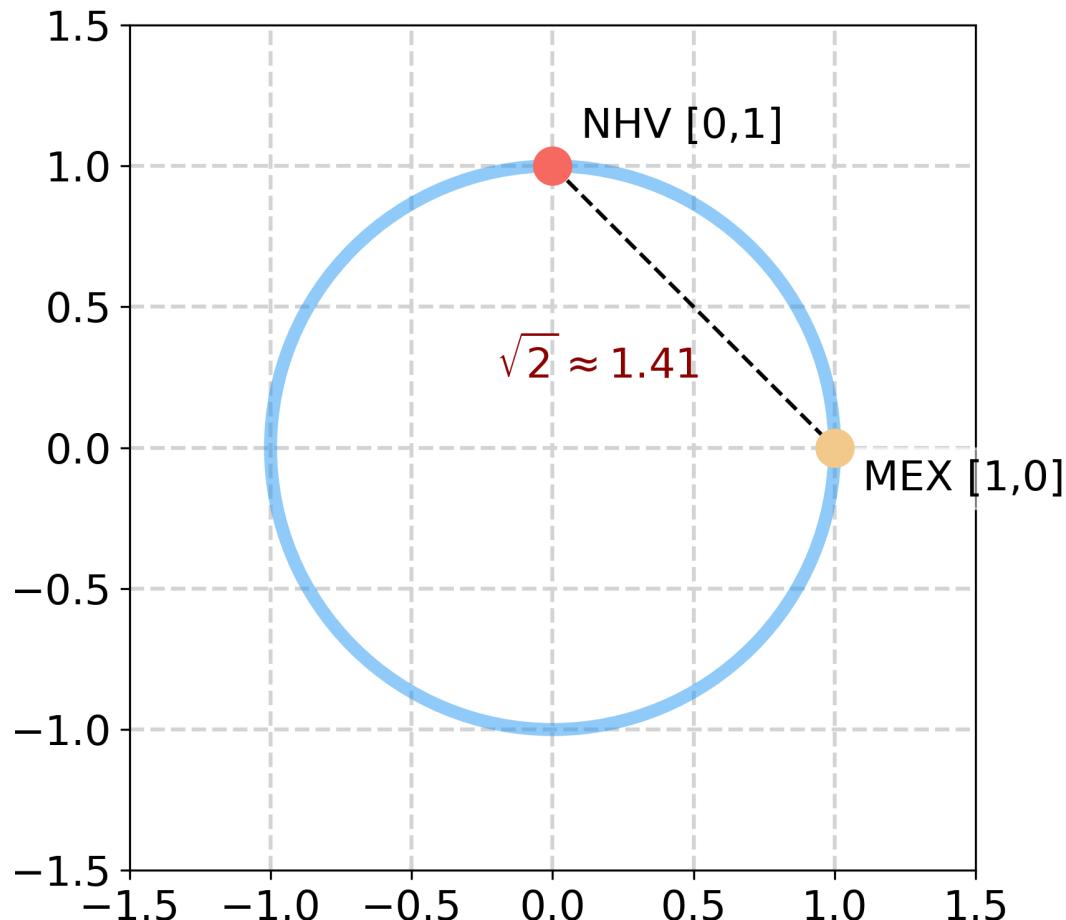
MEX



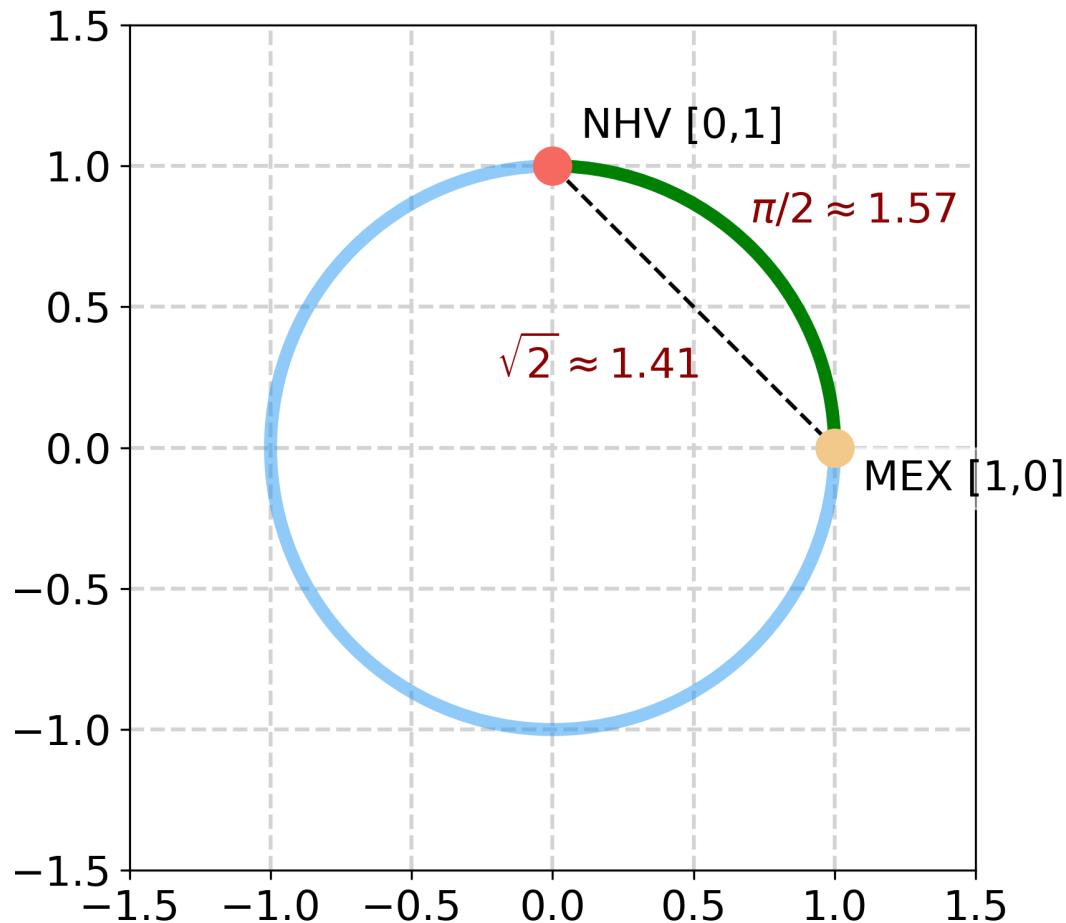
How far is NHV \rightarrow MEX?



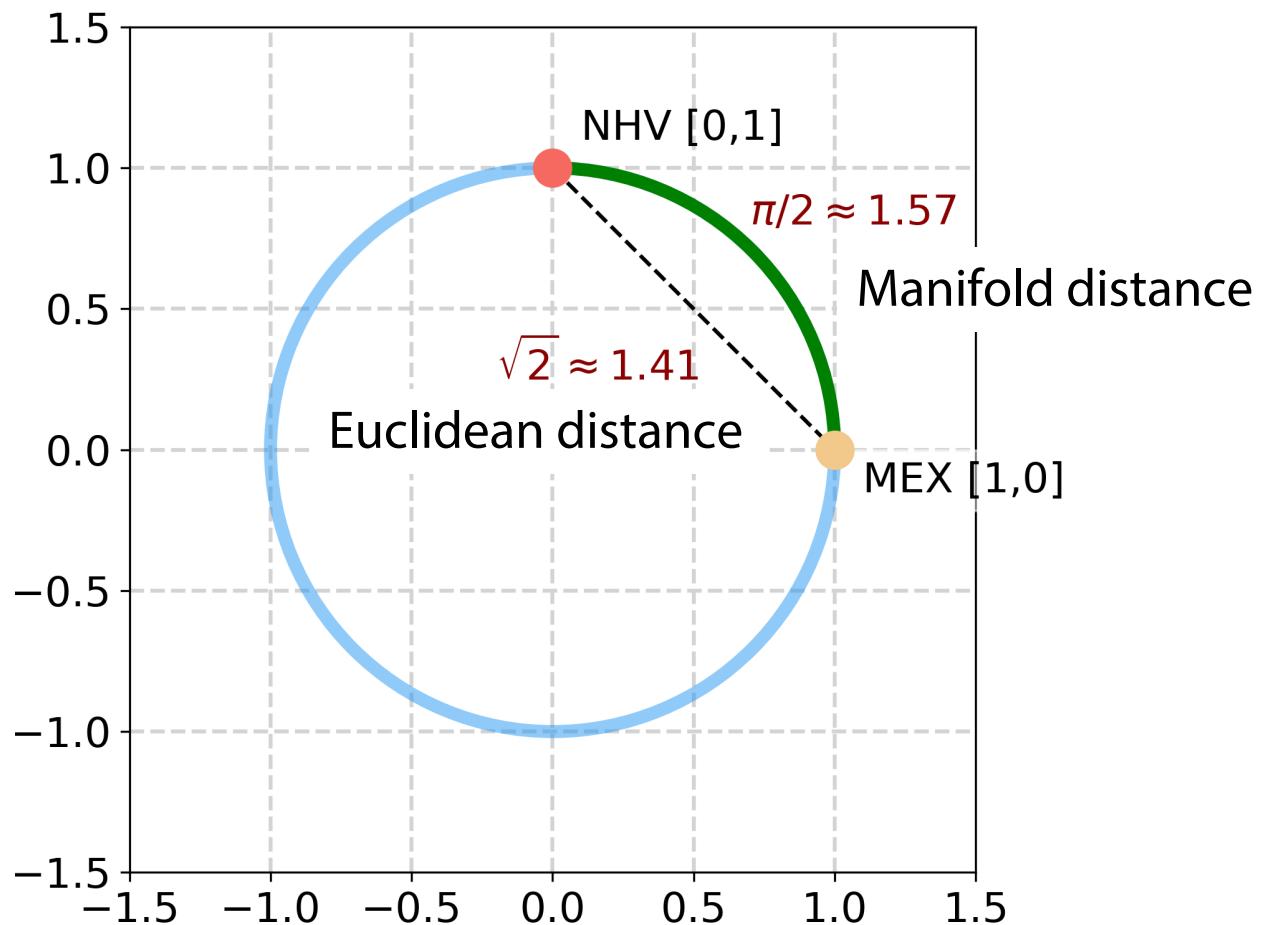
How far is NHV \rightarrow MEX?



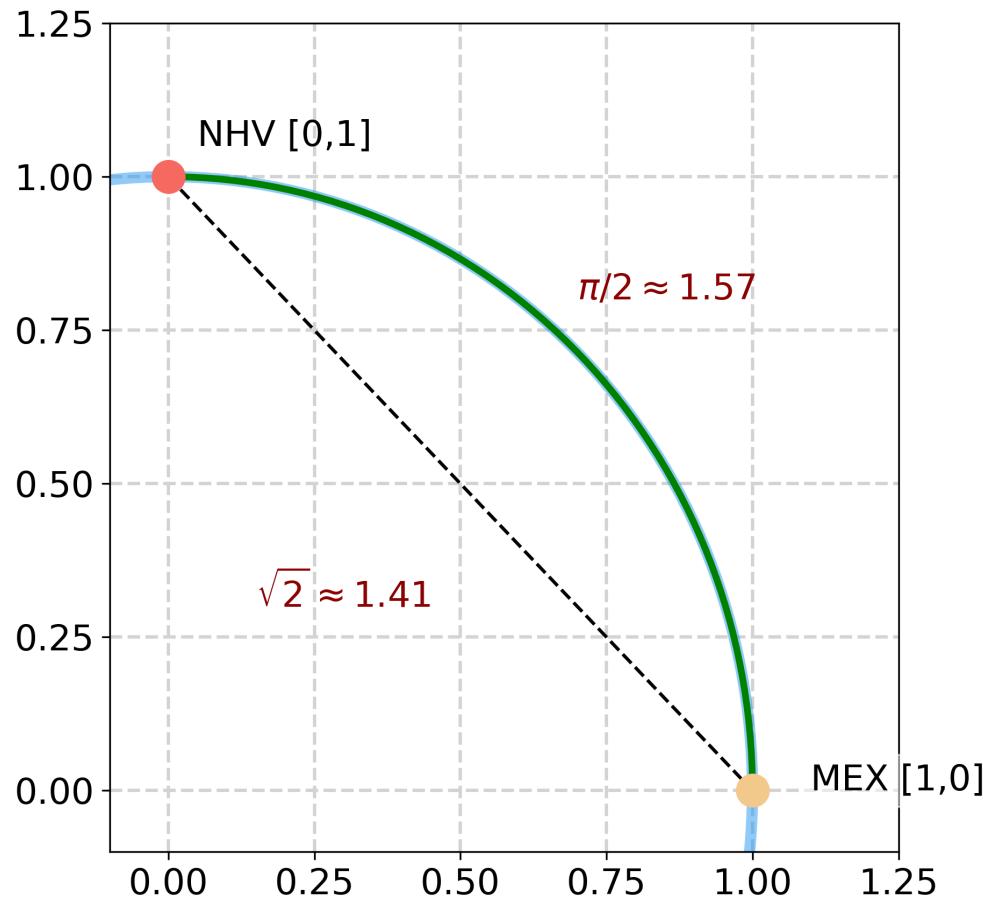
How far is NHV \rightarrow MEX?



How far is NHV \rightarrow MEX?

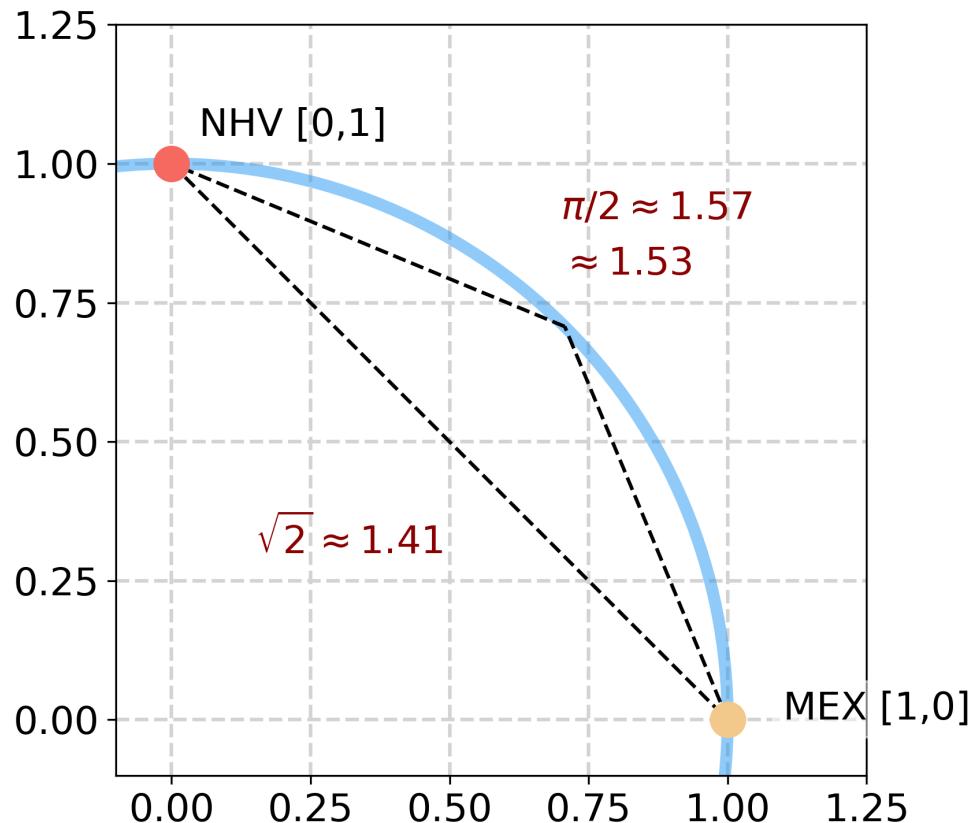


Is a circle locally Euclidean?



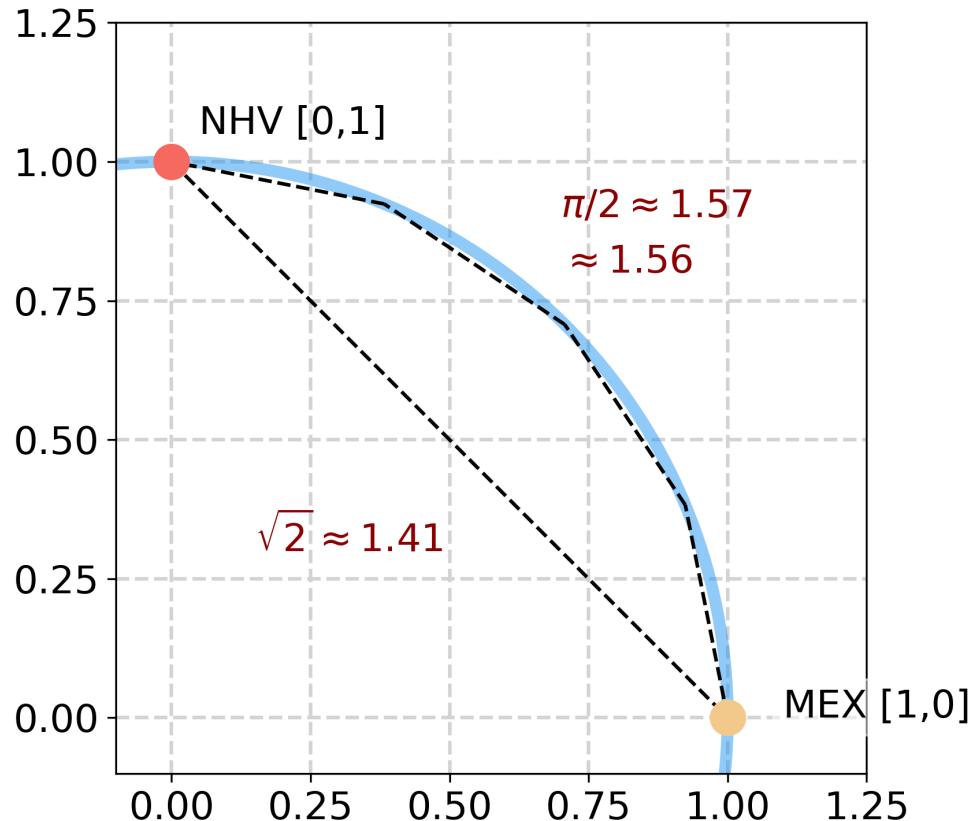
Is a circle locally Euclidean?

segments = 2



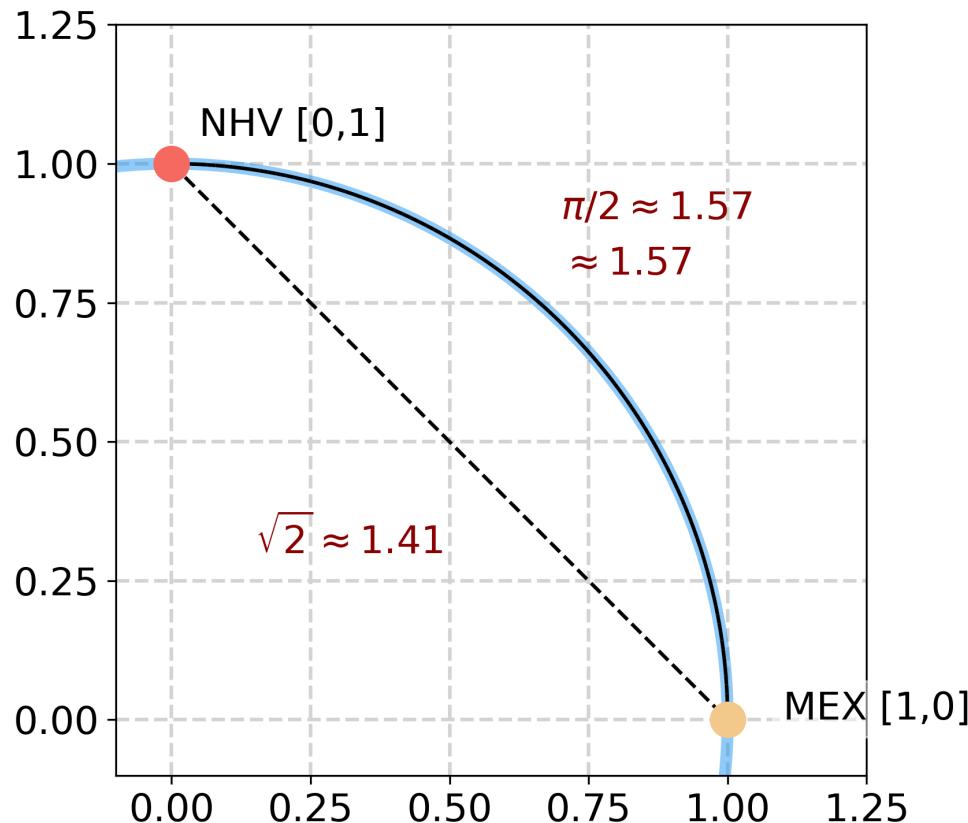
Is a circle locally Euclidean?

segments = 4

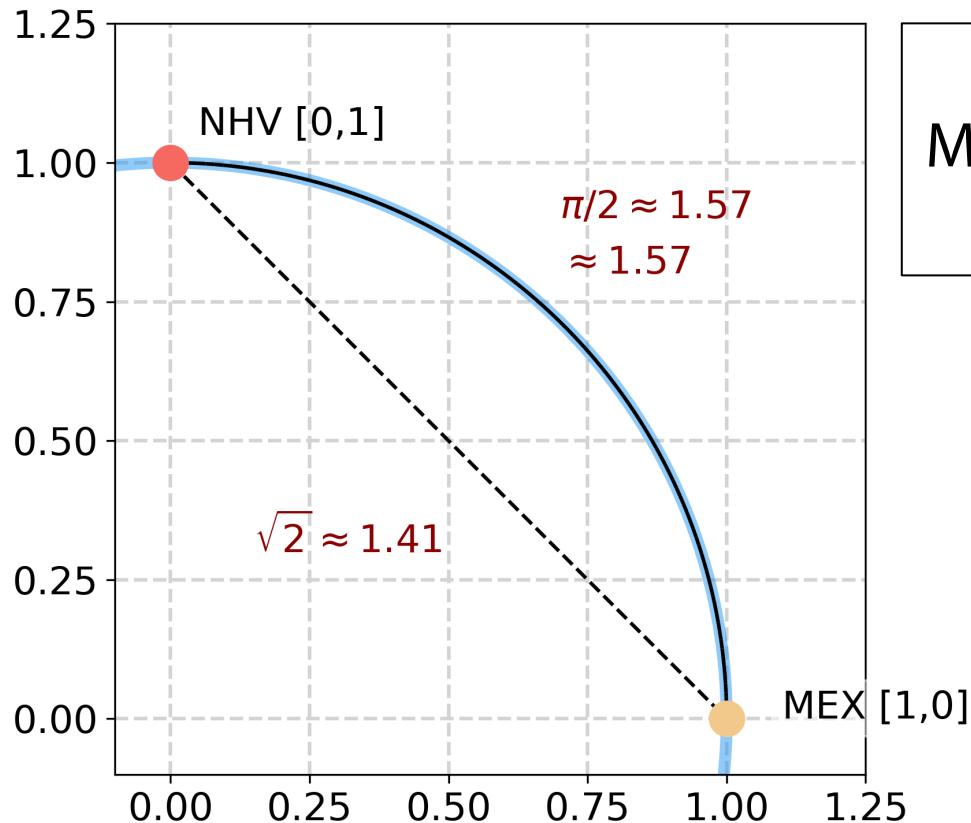


Is a circle locally Euclidean?

segments = 100

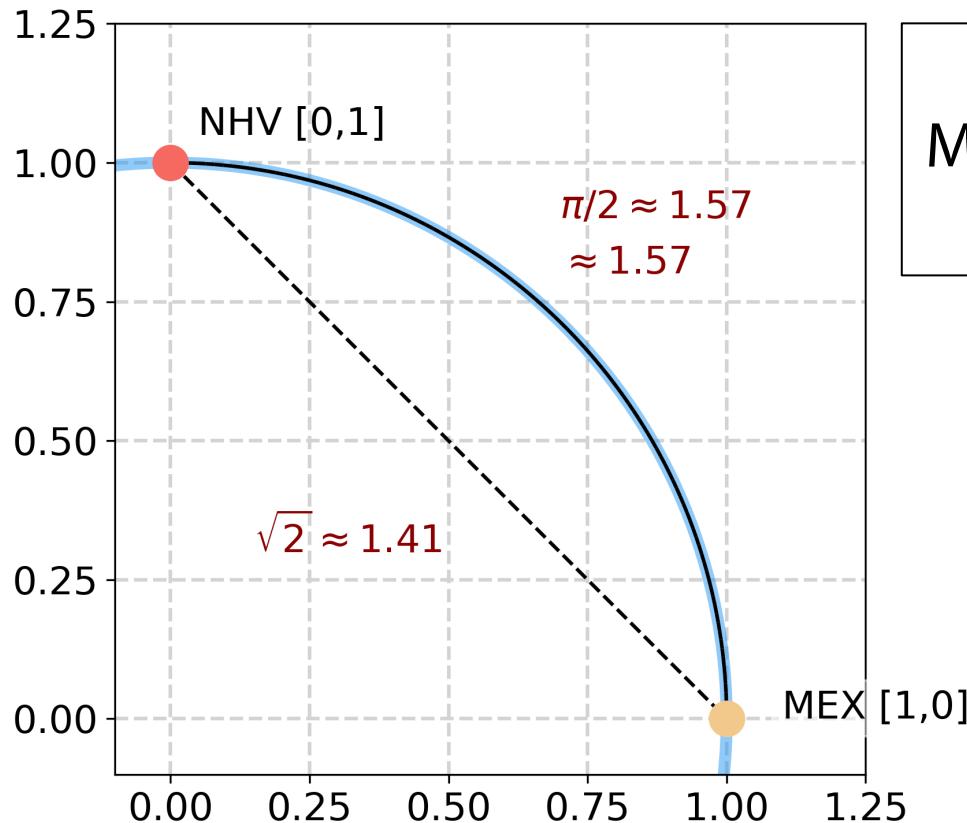


Is a circle locally Euclidean?



Key Insight:
Manifolds are **locally** Euclidean,
but not **globally** Euclidean

Is a circle locally Euclidean?

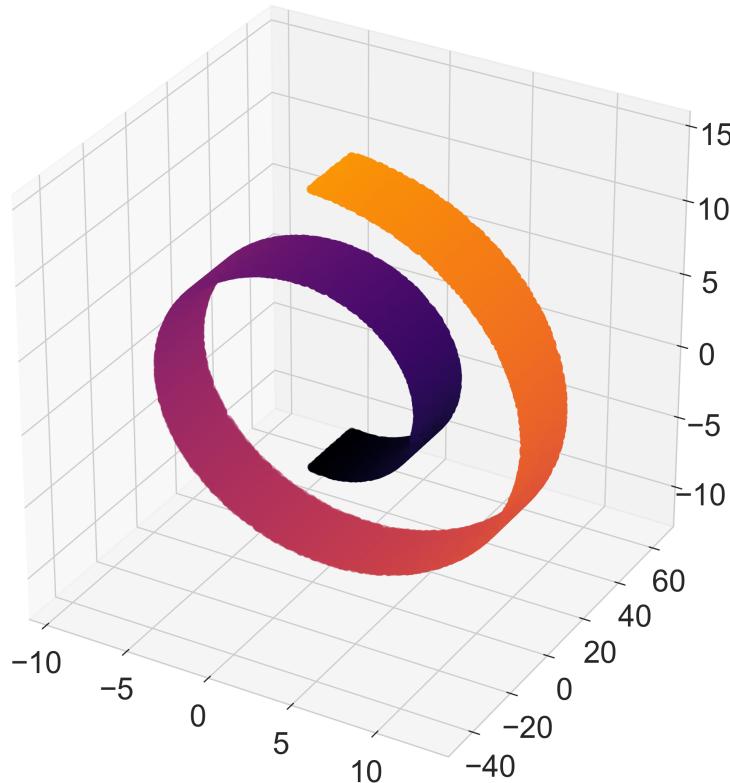


Key Insight:
Manifolds are **locally** Euclidean,
but not **globally** Euclidean

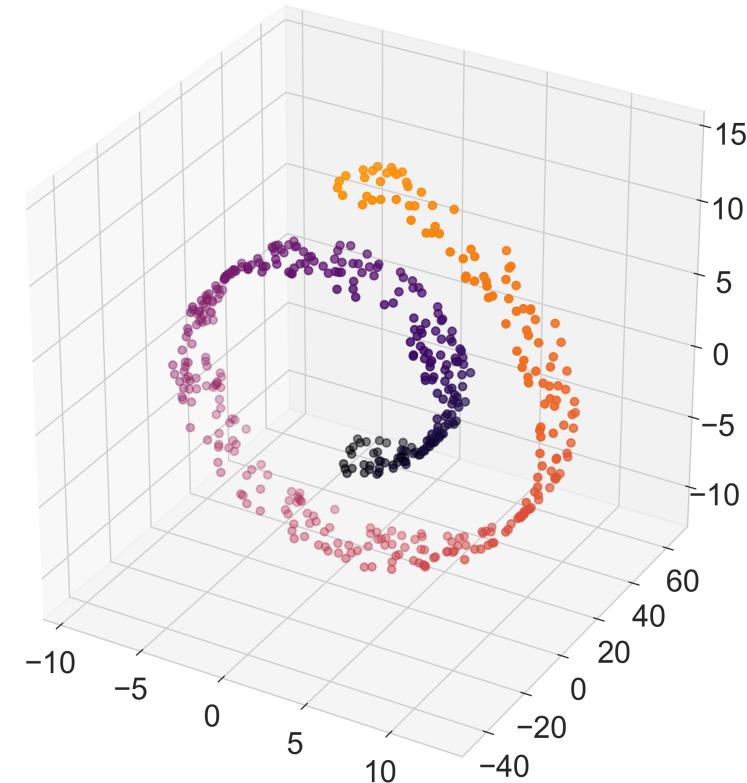
How do we define local
versus global when we only
have a collection of points?

The swiss roll is a plane (\mathbb{R}^2) embedded in \mathbb{R}^3

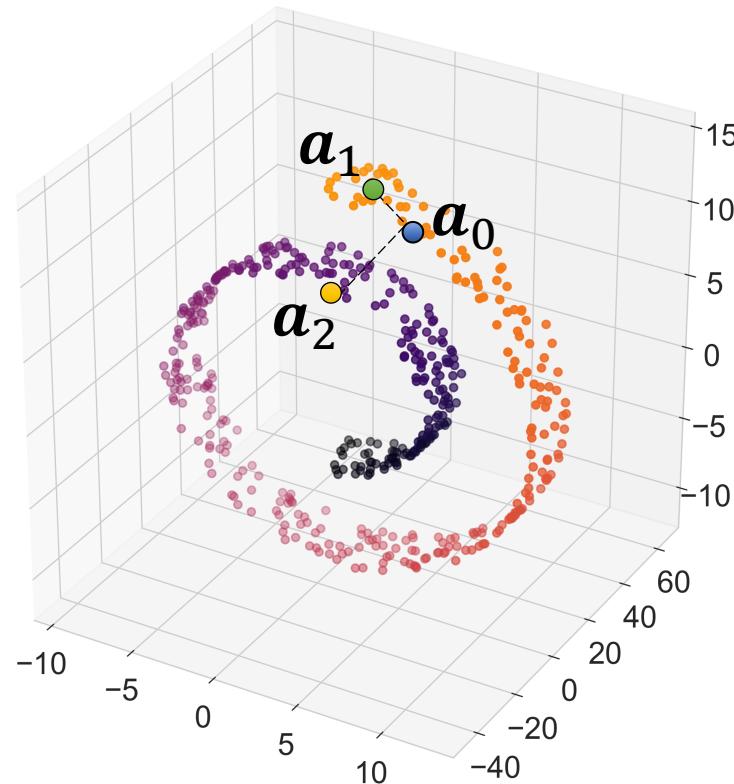
True manifold



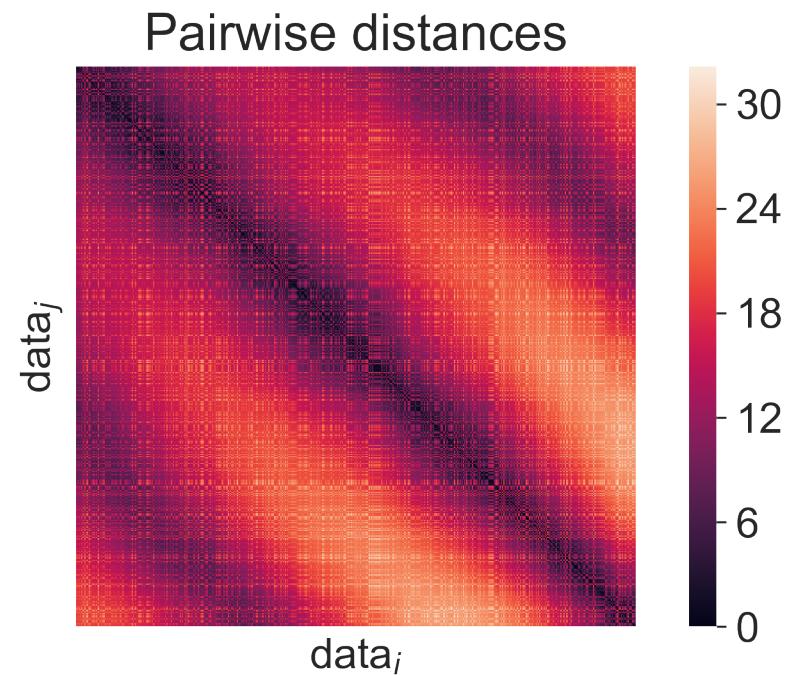
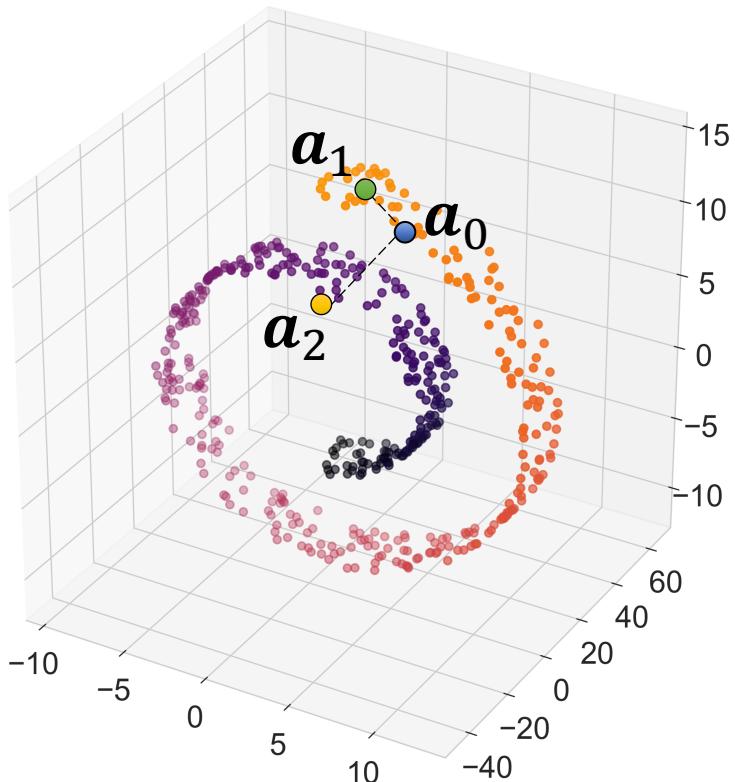
Data sampled from manifold



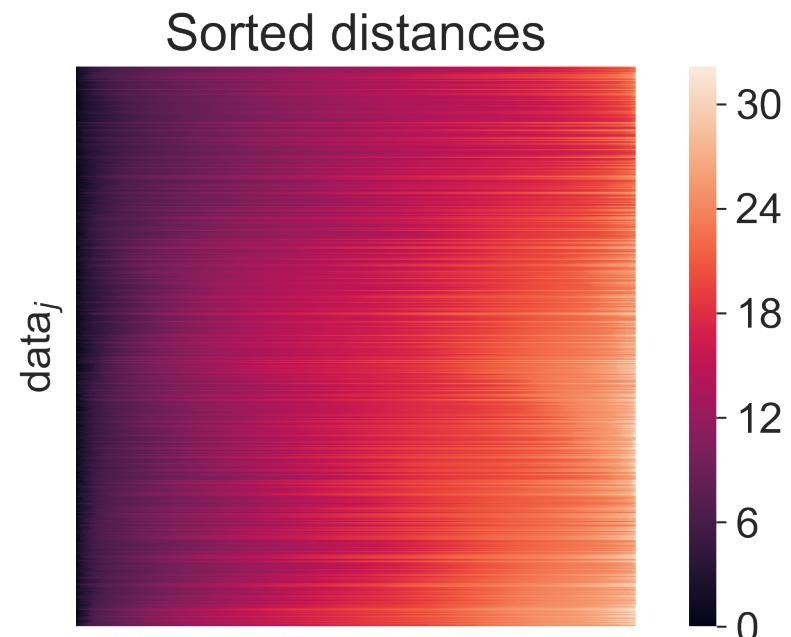
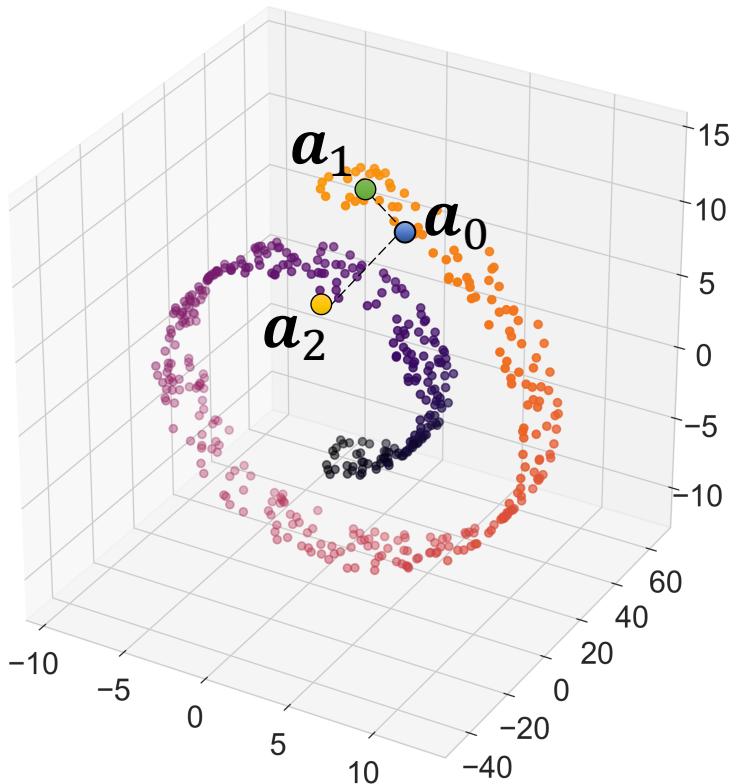
How do we quantify close (local) versus far (global) in a dataset?



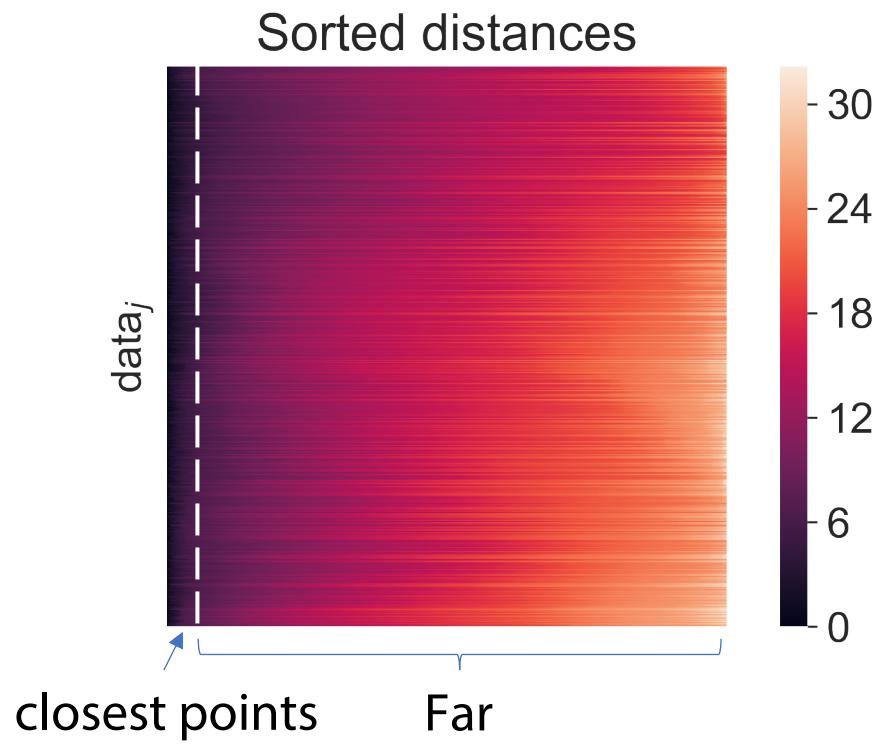
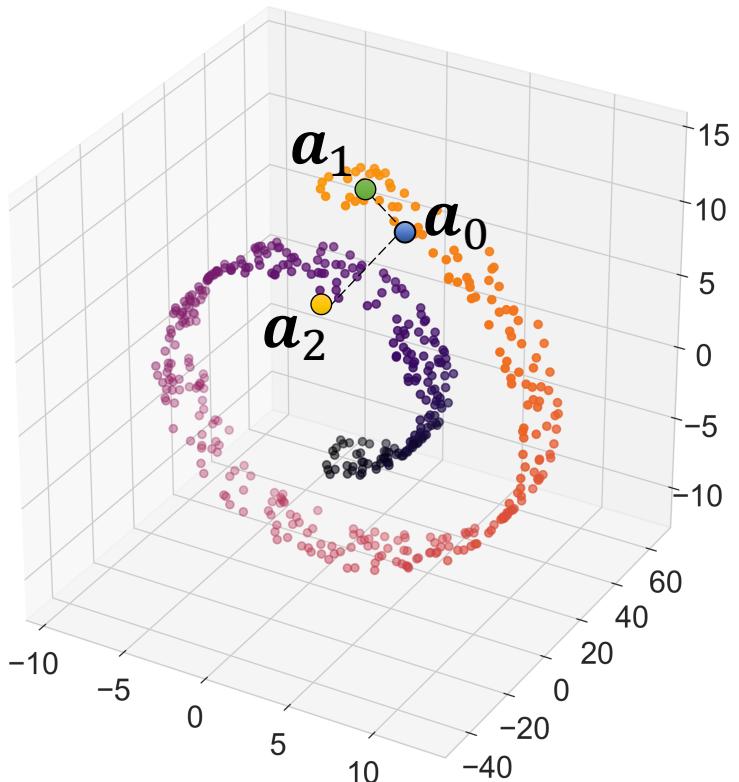
How do we quantify close (local) versus far (global) in a dataset?



How do we quantify close (local) versus far (global) in a dataset?

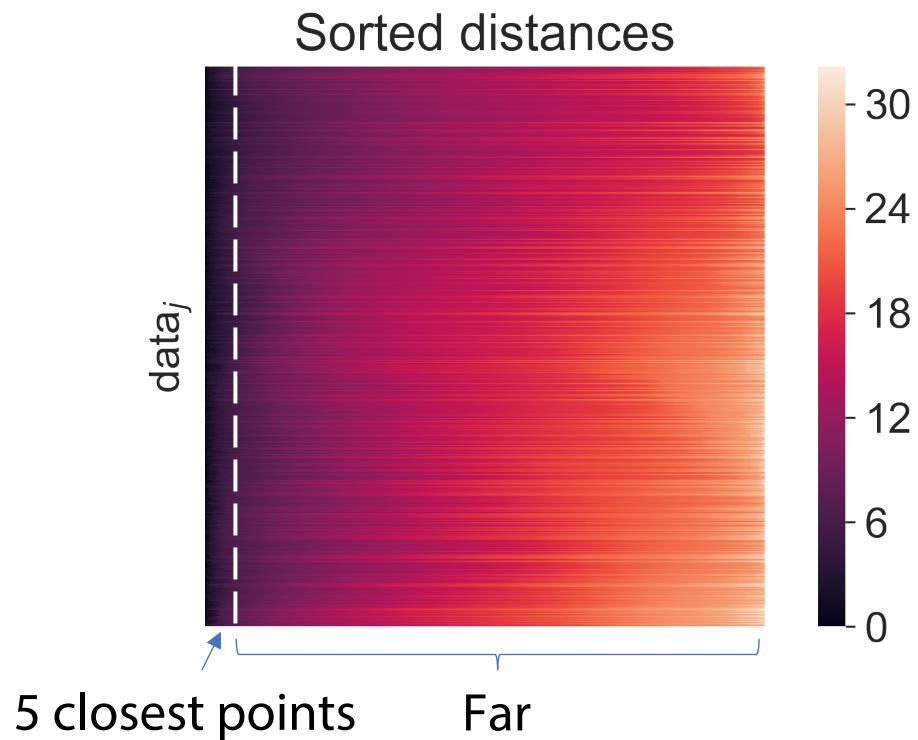
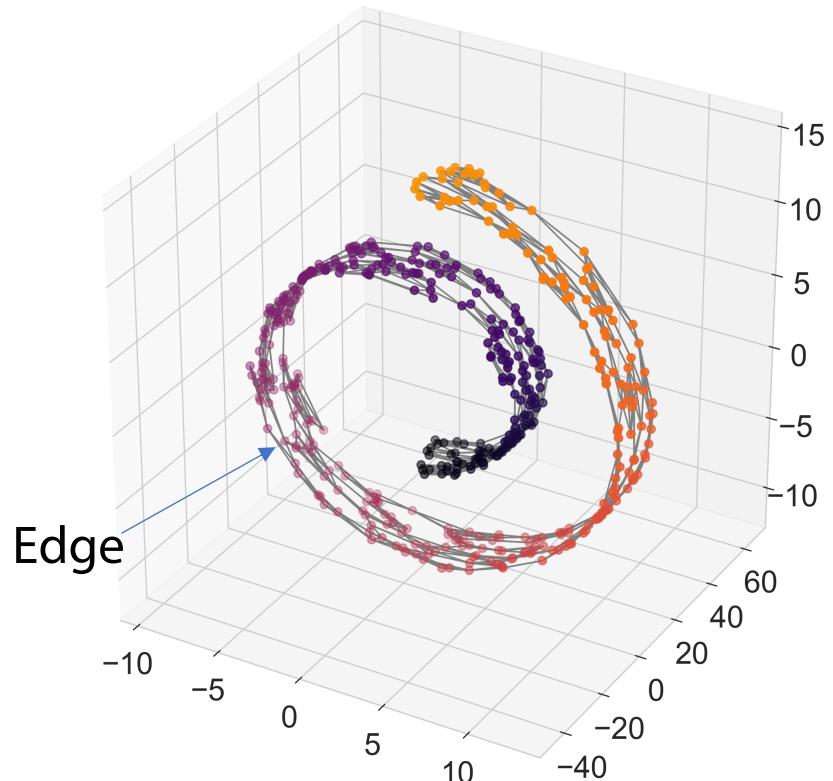


How do we quantify close (local) versus far (global) in a dataset?

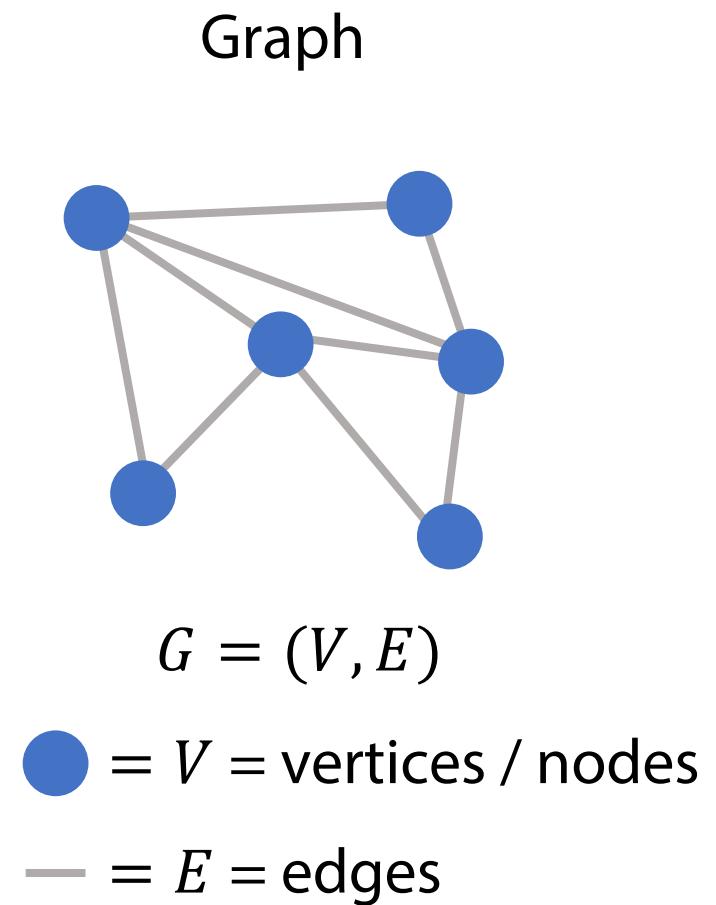
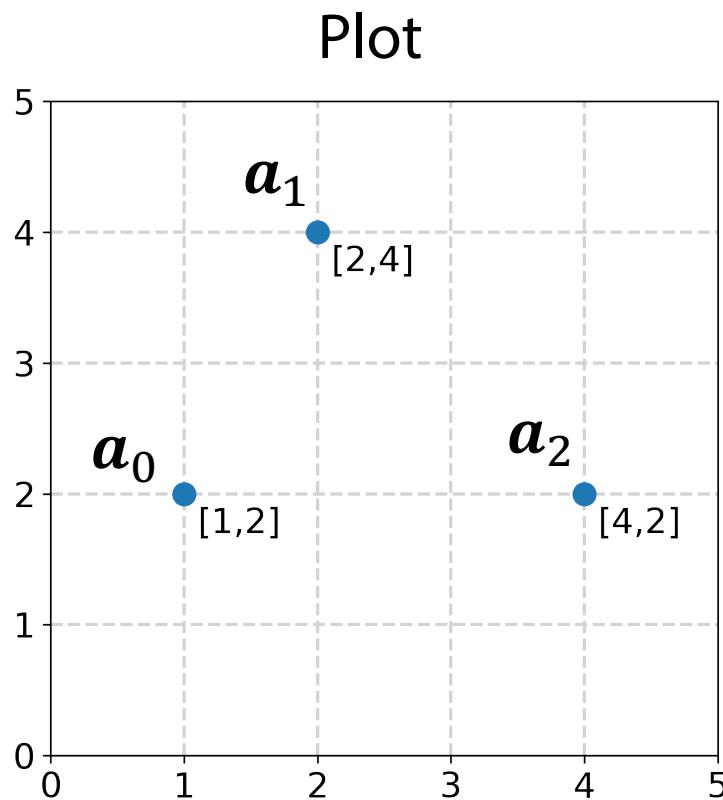


How do we quantify close (local) versus far (global) in a dataset?

Point out that A2 would be “sort of close” in Euclidean space



What is a “graph”?

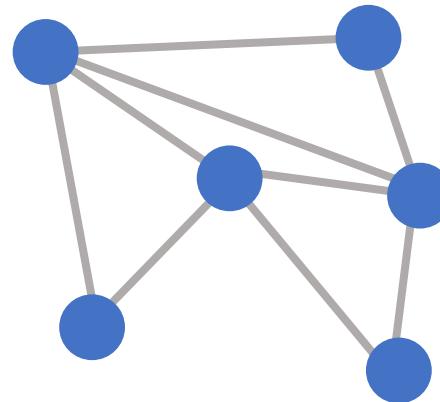


What is a “graph”?

Things that can be modelled as a graph

Thing	Vertices	Edges
Internet	Computers	Network connections
Traffic	Intersections	Roads
Social network	People	Friendships
Cell similarities	Cells	Similarity relationships

Graph

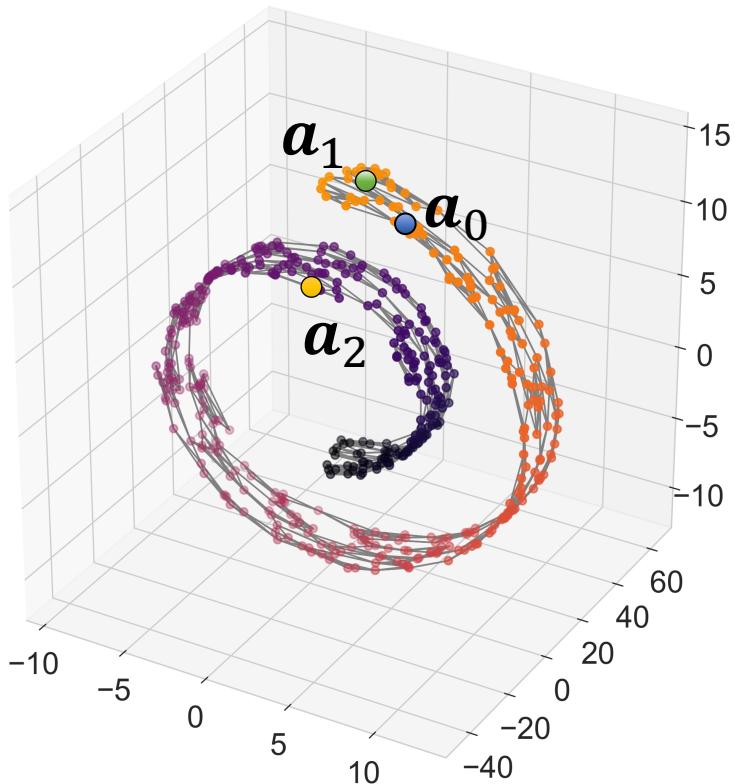


$$G = (V, E)$$

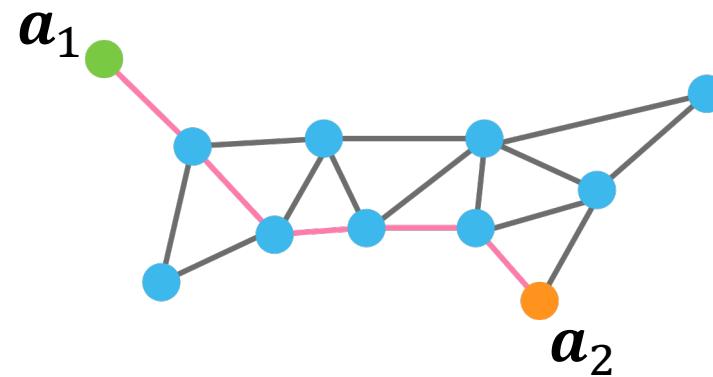
● = V = vertices

— = E = edges

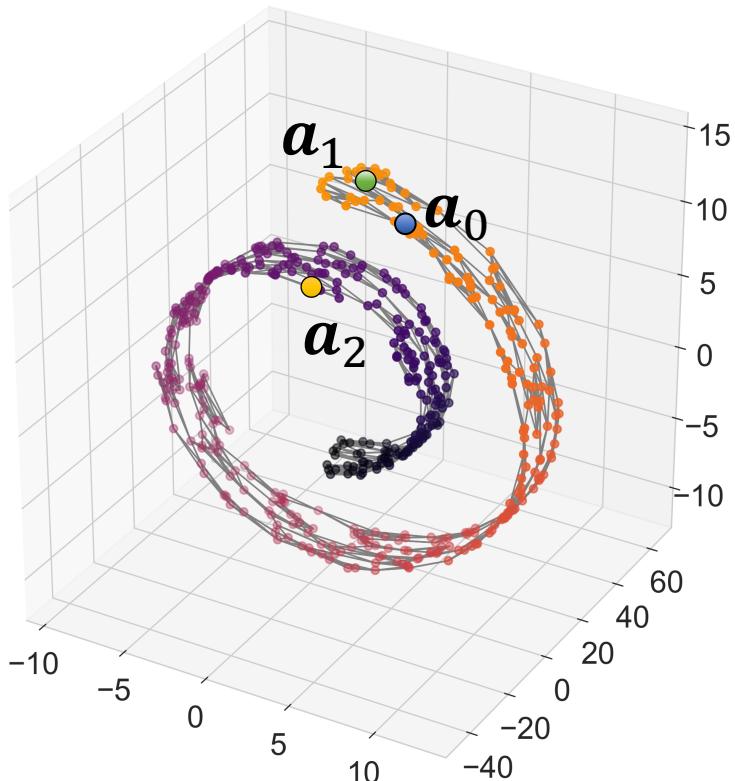
Graph walks approximate manifold distances



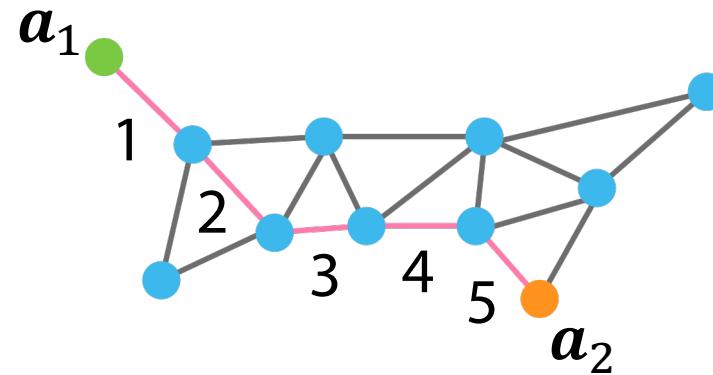
Shortest path between points



Graph walks approximate manifold distances

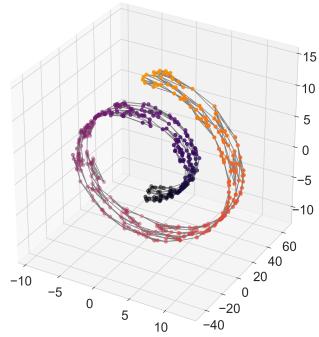


Shortest path between points

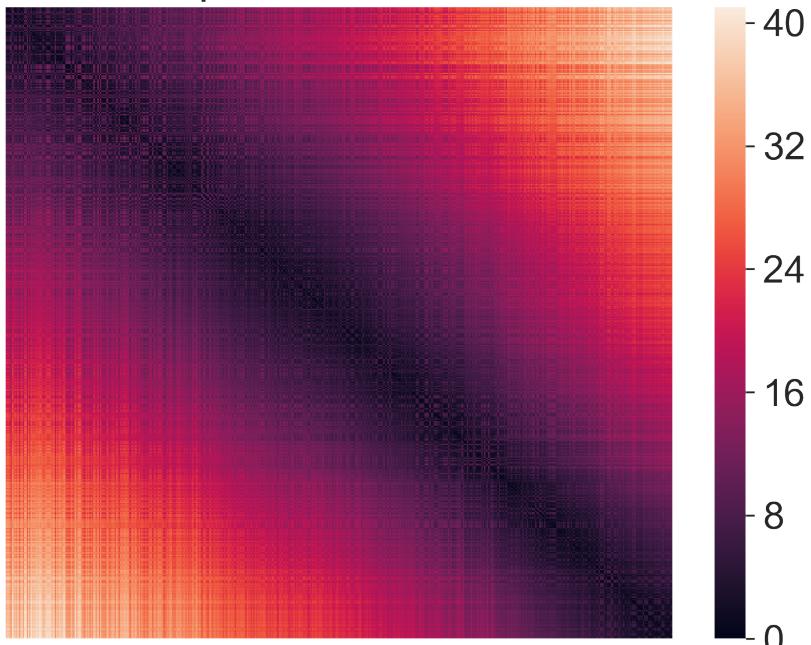


$$\begin{aligned} d_{geodesic}(a_1, a_2) &= \text{shortestpath}(a_1, a_2) \\ &= 5 \end{aligned}$$

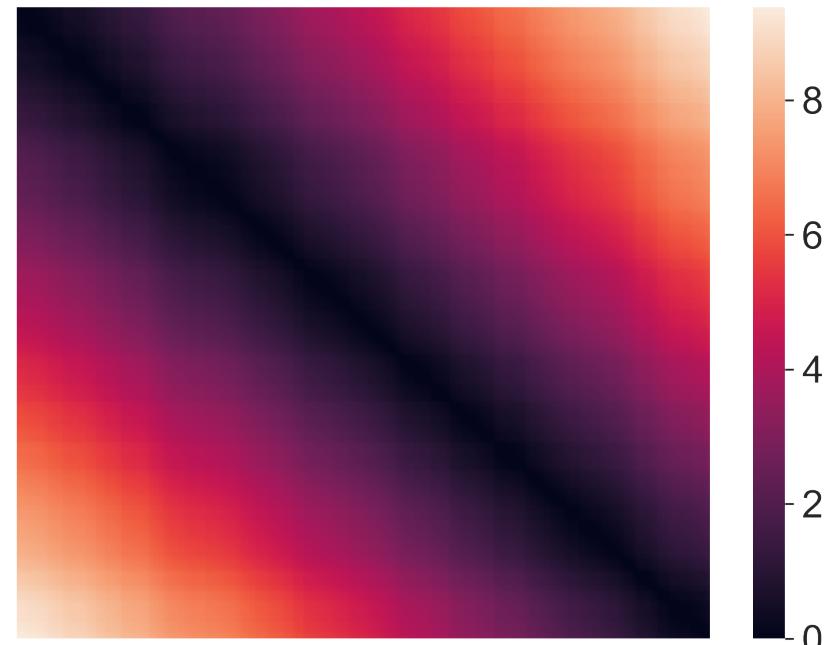
Graph walks approximate manifold distances



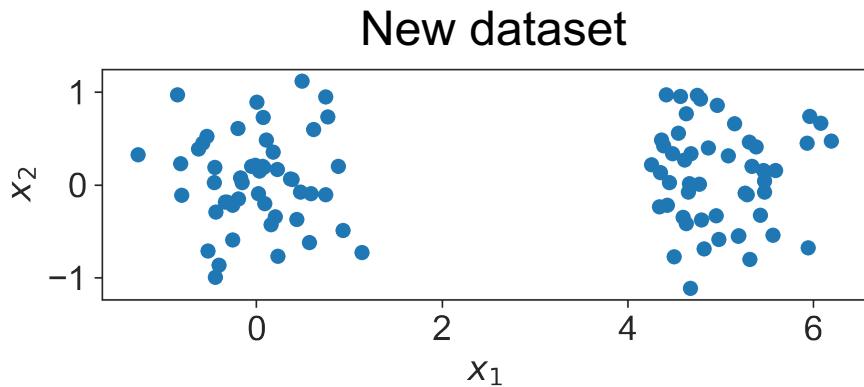
k-NN Graph Geodesic Distances



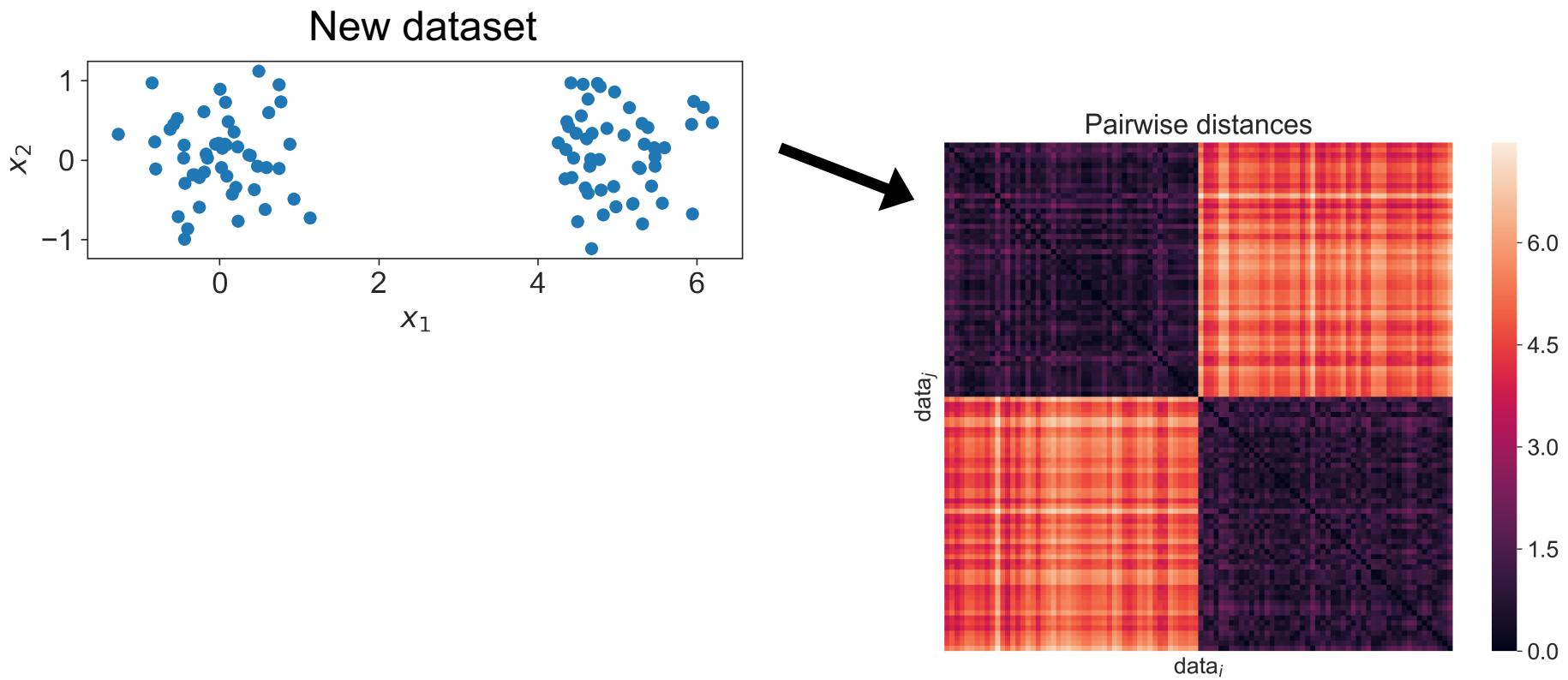
Manifold Distances



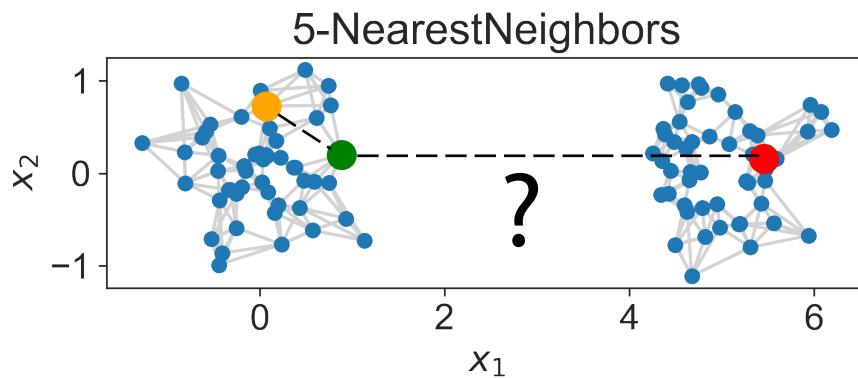
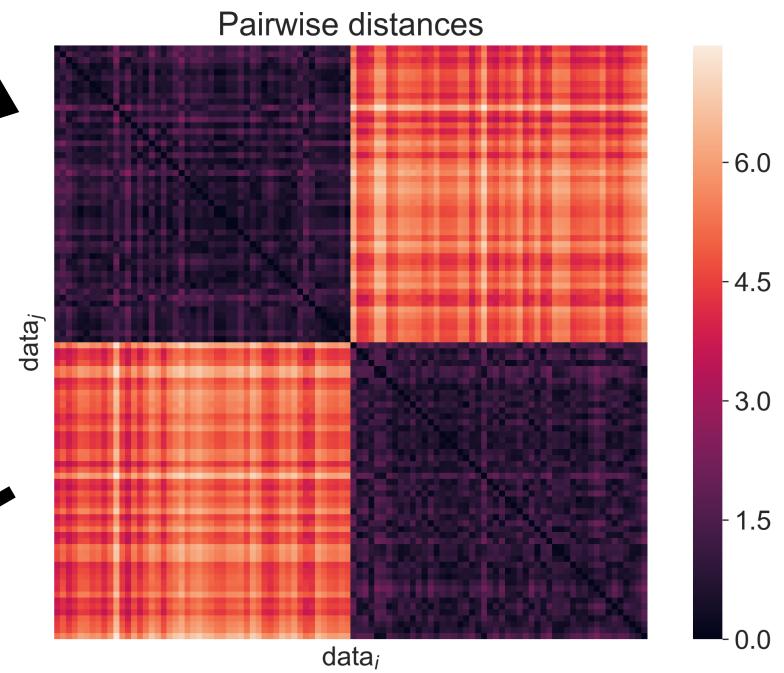
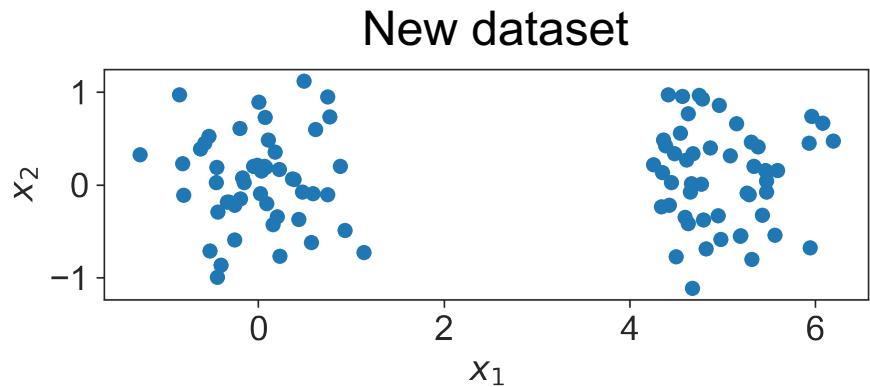
What if the data isn't entirely connected?



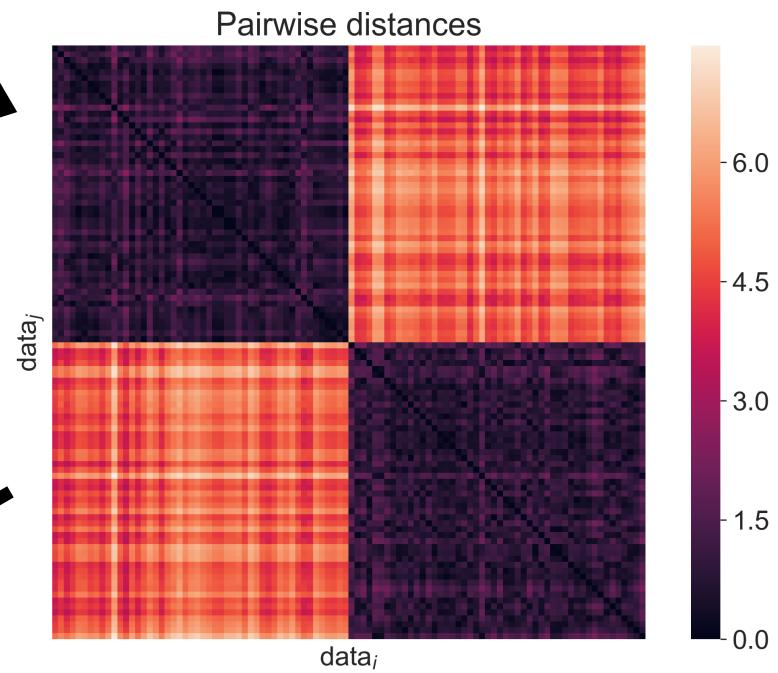
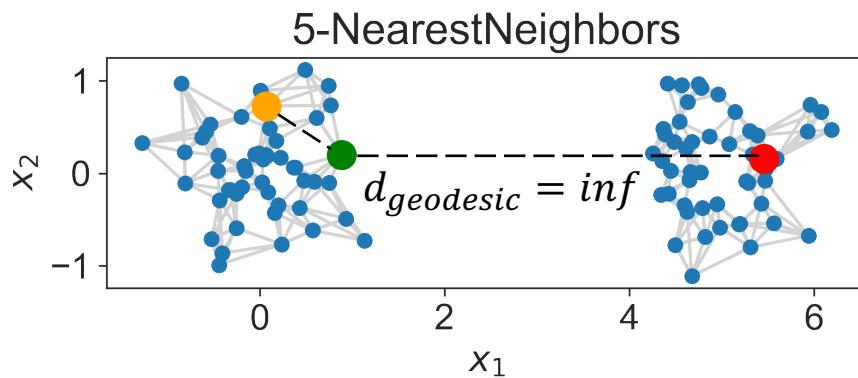
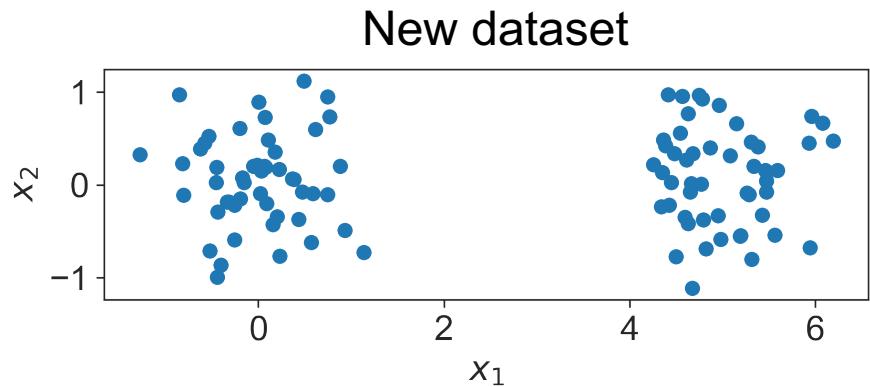
What if the data isn't entirely connected?



What if the data isn't entirely connected?

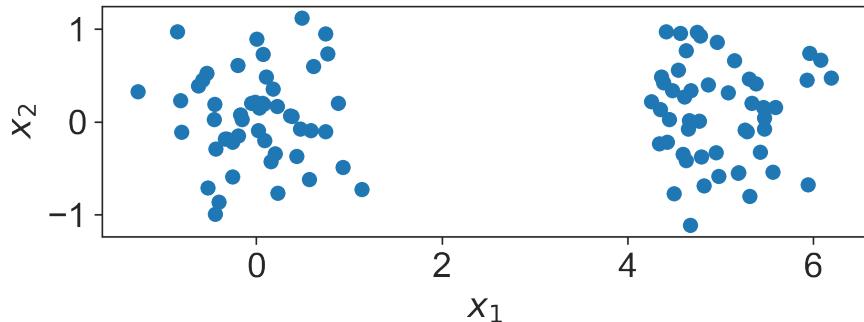


What if the data isn't entirely connected?

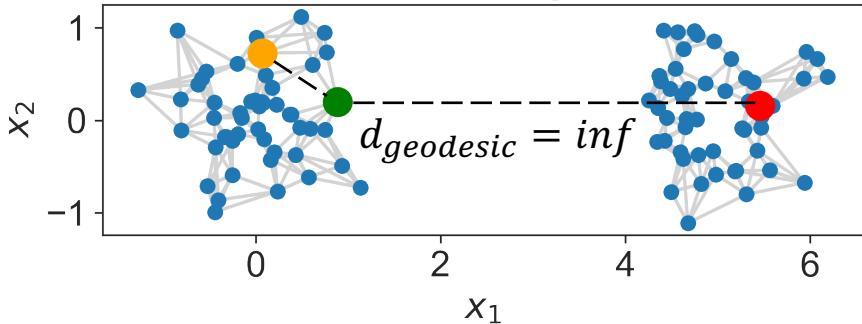


What if the data isn't entirely connected?

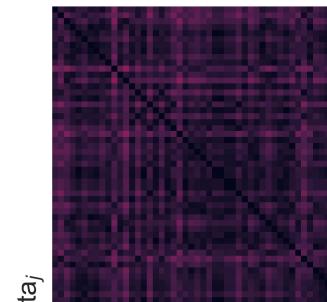
New dataset



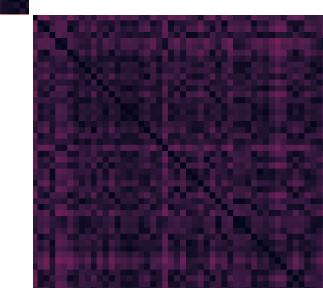
5-NearestNeighbors



Pairwise distances



inf



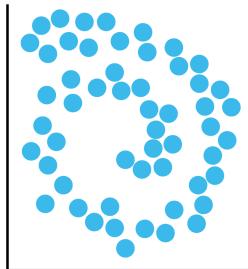
inf

data_j

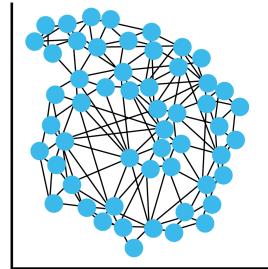
data_i

Summary: learning graphs from data

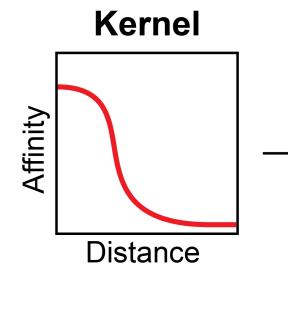
Data in two dimensions



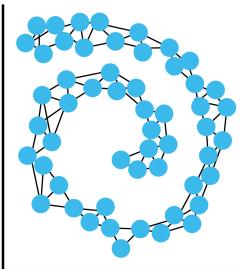
Distances between all points are calculated



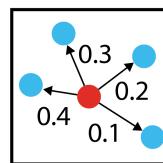
A kernel function calculates affinities from distance



Only local relationships are preserved

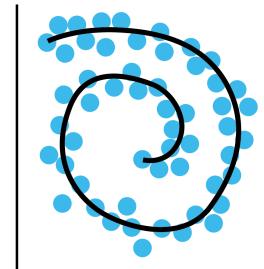


Diffusion shares information between nodes

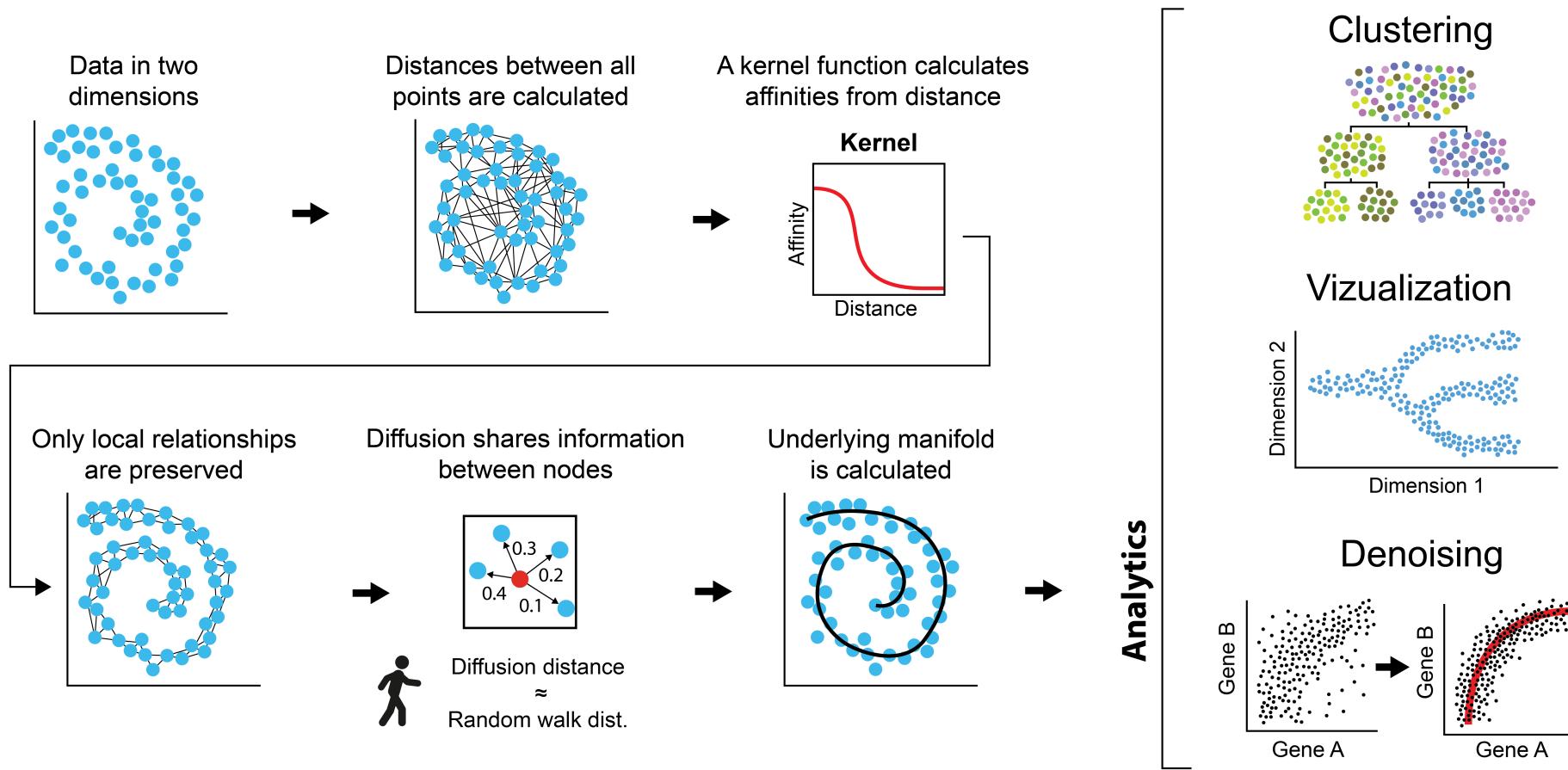


Diffusion distance
≈
Random walk dist.

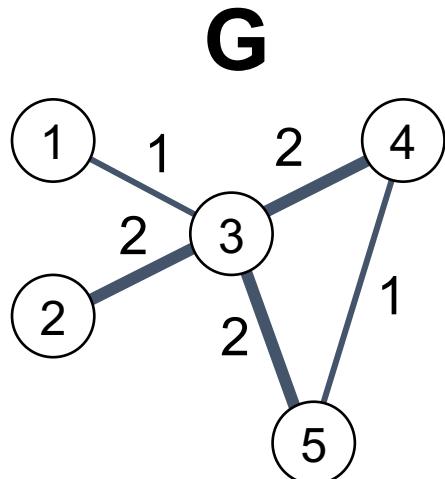
Underlying manifold is calculated



Summary: learning graphs from data



Matrix representations of graphs



A adjacency matrix

$$\mathbf{A} = \begin{vmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 1 & 2 & 0 & 2 & 2 \\ 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 & 0 \end{vmatrix}$$

D degree matrix

$$\mathbf{D} = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 7 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{vmatrix}$$

Laplacian matrix

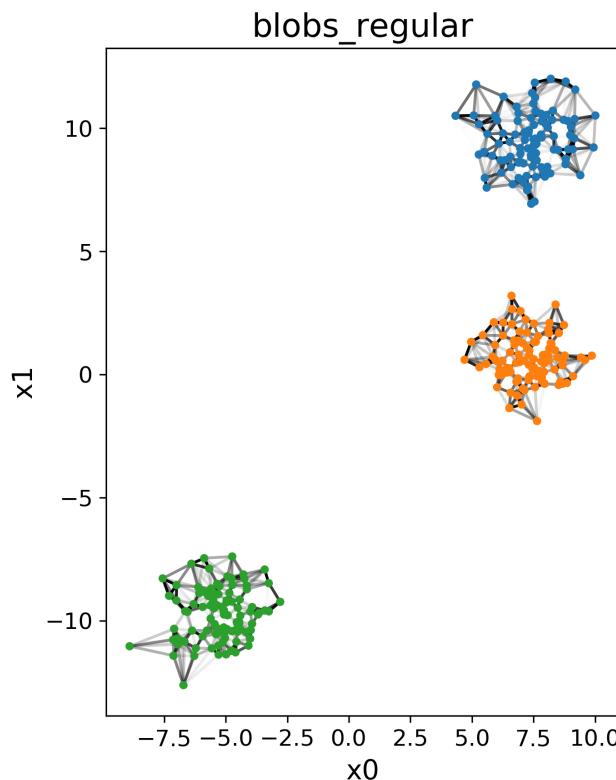
$$\mathcal{L} = \mathbf{D} - \mathbf{A} = \begin{vmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 2 & -2 & 0 & 0 \\ -1 & -2 & 7 & -2 & -2 \\ 0 & 0 & -2 & 3 & -1 \\ 0 & 0 & -2 & -1 & 3 \end{vmatrix}$$



- Creating visualizations
- Clustering
- Denoising data associated with the nodes

Exercise – choosing k

$k = 5$



$k = 10$

