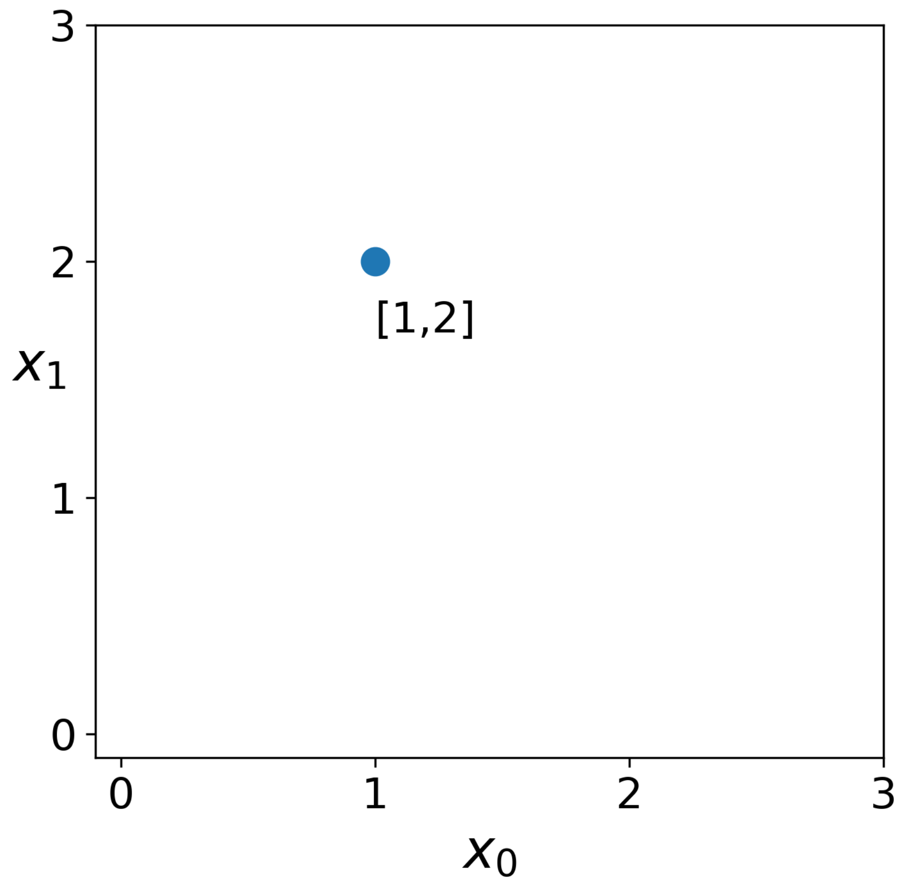
Thinking about high dimensional data

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

$\boldsymbol{a}_0 \in \mathbb{R}^2$

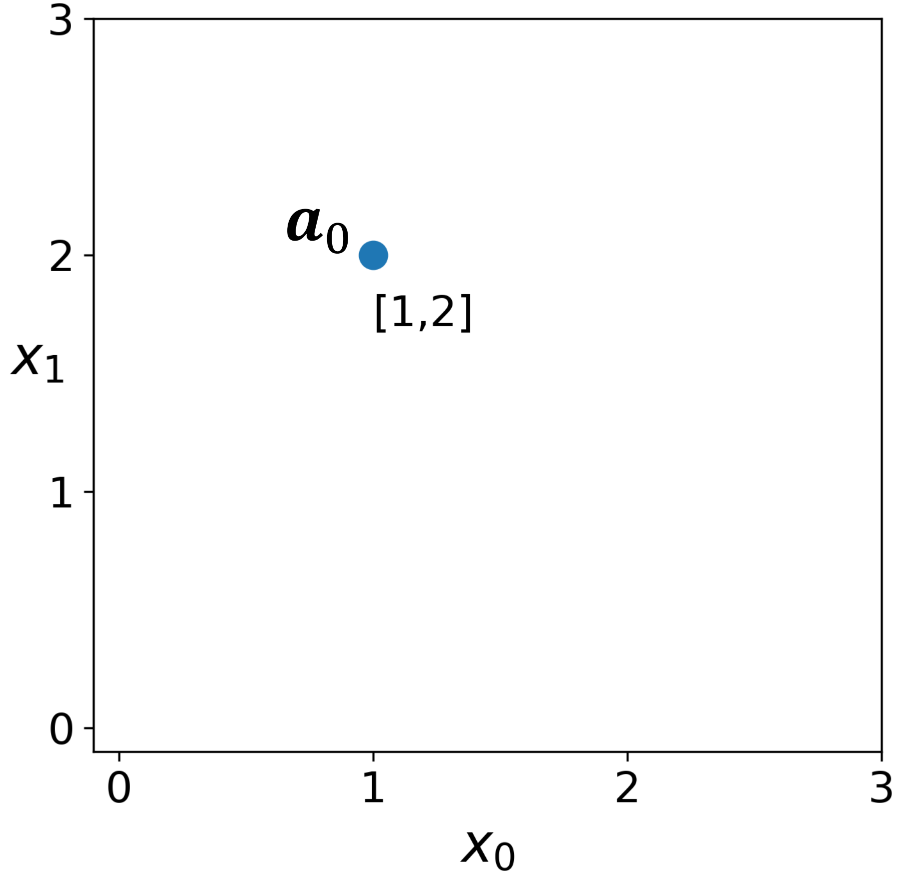
$\boldsymbol{a}_0 = [1,2]$

x_0　　x_1

features

$\boldsymbol{a}_{[:,0]}$

$\mathbb{R}^2$ two-dimensional space



What does "dimensionality" mean?

vector

coordinate space

$\boldsymbol{a}_0 \,\epsilon\, \mathbb{R}^2$

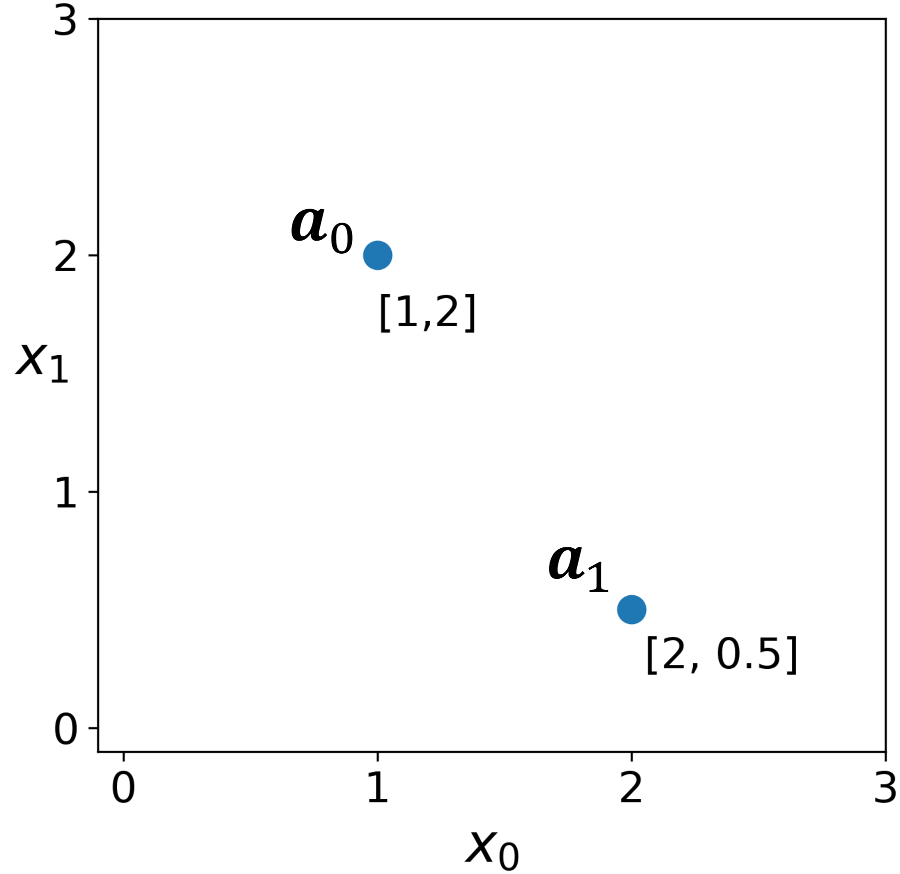
$\boldsymbol{a}_0 = [1,2]$

x_0 x_1

features

$\boldsymbol{a}_{[:,0]}$

$\mathbb{R}^2$ two-dimensional space



What does “dimensionality” mean?

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

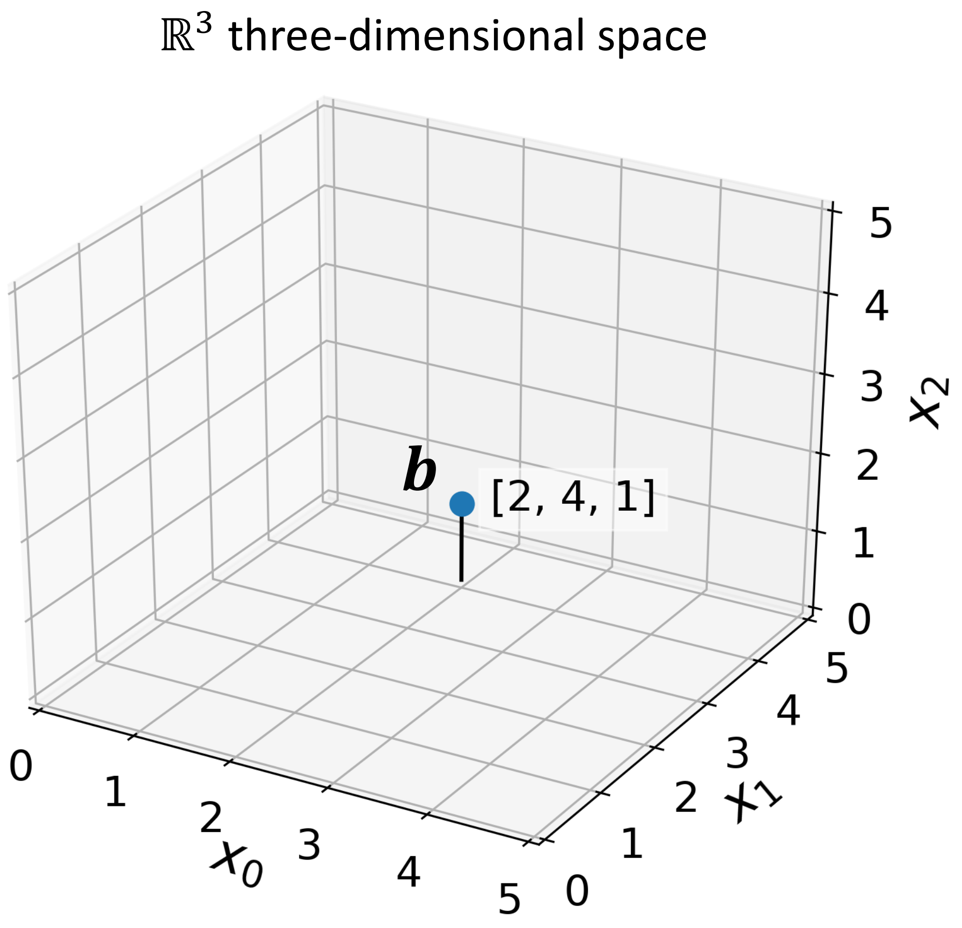
$\boldsymbol{a} = [1,2]$

x_0 x_1

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

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

x_0 x_1 x_2



What does "dimensionality" mean?

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

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

$x_0 \quad x_1$

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

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

$x_0 \quad x_1 \quad x_2$

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

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

$x_0 \quad x_1 \quad x_2 \quad x_3$

$\mathbb{R}^4$ four-dimensional space

???

What does “dimensionality” mean?

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

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

x_0 x_1

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

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

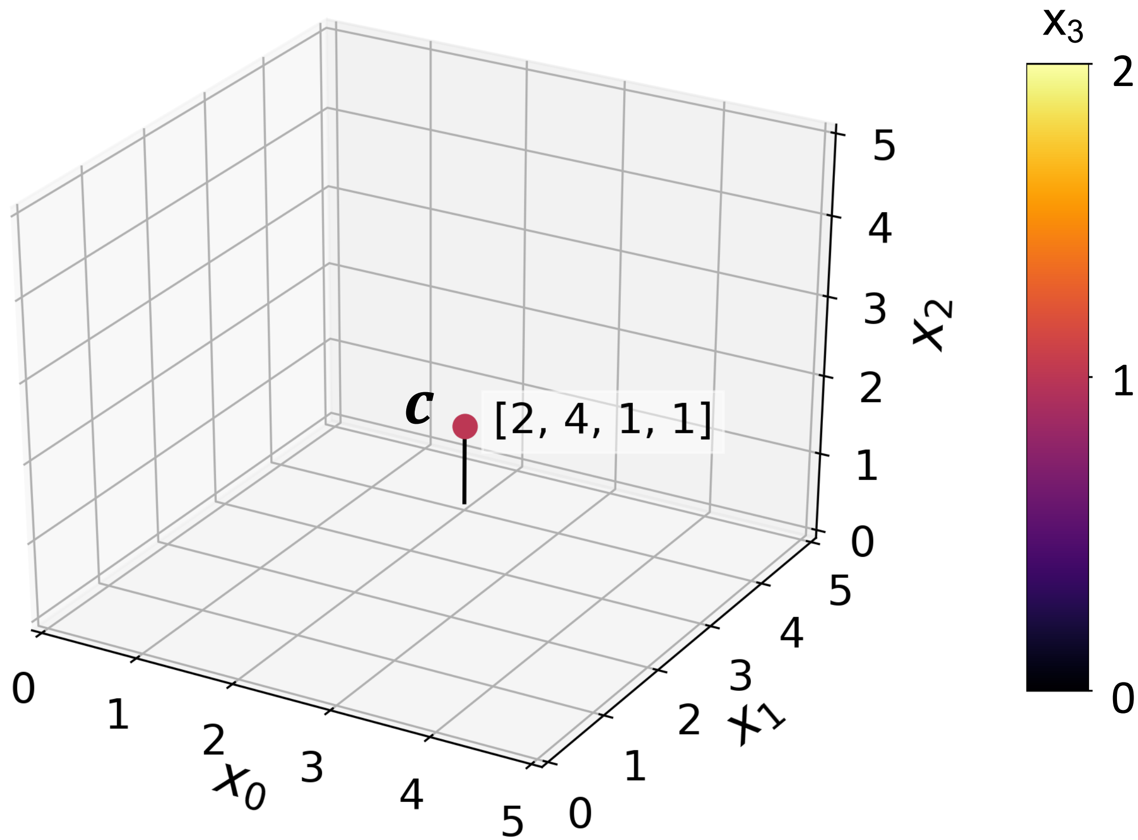
x_0 x_1 x_2

$\boldsymbol{c} \in \mathbb{R}^4$

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

x_0 x_1 x_2 x_3

$\mathbb{R}^4$ four-dimensional space



What does “dimensionality” mean?

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

$$\boldsymbol{a} = [\underset{x_1}{1}, \underset{x_2}{2}]$$

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

$$\boldsymbol{b} = [\underset{x_1}{2}, \underset{x_2}{4}, \underset{x_3}{1}]$$

$\boldsymbol{c} \in \mathbb{R}^4$

$$\boldsymbol{c} = [\underset{x_1}{2}, \underset{x_2}{4}, \underset{x_3}{1}, \underset{x_4}{1}]$$

$\mathbb{R}^5$ five-dimensional space

???

What do x_1 x_2, …, x_n represent?

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

$\boldsymbol{a} =$

	alcohol	hue
Wine0	14.23	1.04

$\boldsymbol{a}_0$ $\boldsymbol{a}_1$

$\boldsymbol{b} =$

	Age	Height (m)	Weight (kg)
Player0	23	2.31	91

$\boldsymbol{b}_0$ $\boldsymbol{b}_1$ $\boldsymbol{b}_2$

$\boldsymbol{c} =$

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
Iris0	5.1	3.5	1.4	0.2

UCI ML Wine Data Set

What if we have more than one observation?

$a =$

	alcohol	hue
Wine0	14.23	1.04

a_0 a_1

columns (features)

$A =$

	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

rows (observations)

matrix

UCI ML Wine Data Set

What if we have more than one observation?

$a =$

	alcohol	hue
Wine0	14.23	1.04

a_0 a_1

columns (features)

$A =$

	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

$\epsilon \ \mathbb{R}^2$

matrix

UCI ML Wine Data Set

How do we reference specific observations in a matrix?

x_0 x_1

$$A =$$

	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

$$\boldsymbol{A}_{[0,:]}$$

$$\boldsymbol{A}_{[\text{row, column}]}$$

What if we have more than one observation?

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

$$
\boldsymbol{A} = \begin{array}{c|cc}
 & \overset{x_0}{\textbf{alcohol}} & \overset{x_1}{\textbf{hue}} \\
\hline
\textbf{Wine0} & 14.23 & 1.04 \\
\textbf{Wine1} & 13.20 & 1.05 \\
\textbf{Wine2} & 13.16 & 1.03 \\
\textbf{Wine3} & 14.37 & 0.86 \\
\textbf{Wine4} & 13.24 & 1.04 \\
\textbf{Wine5} & 14.20 & 1.05 \\
\textbf{Wine6} & 14.39 & 1.02 \\
\textbf{Wine7} & 14.06 & 1.06 \\
\textbf{Wine8} & 14.83 & 1.08 \\
\textbf{Wine9} & 13.86 & 1.01
\end{array}
$$

$$
\boldsymbol{A}_{[0,:]} = a_0 = \begin{array}{c|cc}
 & \textbf{alcohol} & \textbf{hue} \\
\hline
\textbf{Wine0} & \underset{\boldsymbol{A}_{[0,0]}}{14.23} & \underset{\boldsymbol{A}_{[0,1]}}{1.04}
\end{array}
$$

What if we have more than one observation?

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

x_0 x_1

$\boldsymbol{A} =$

	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

$\boldsymbol{A}_{[0,:]} =$

	alcohol	hue
Wine0	14.23	1.04

$\boldsymbol{A}_{[0,0]}$ $\boldsymbol{A}_{[0,1]}$

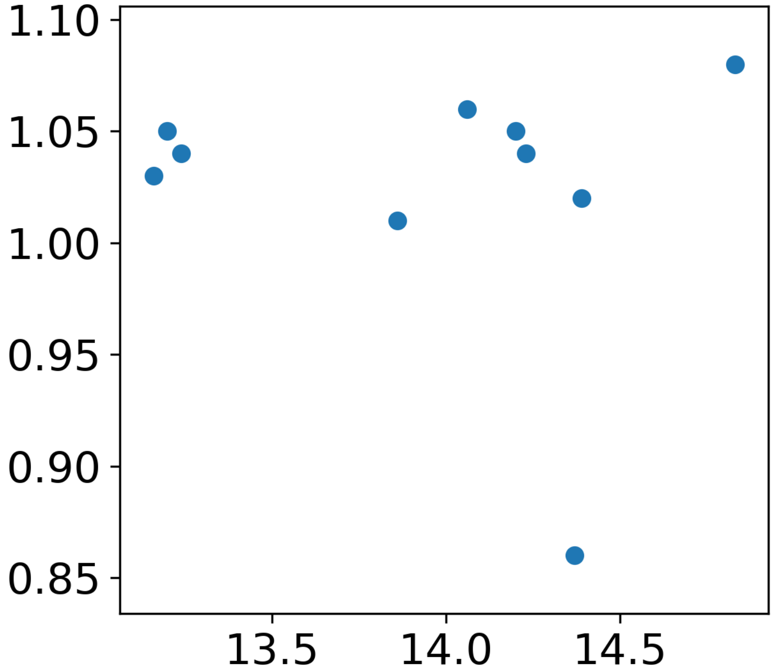
$\boldsymbol{A}_{[9,:]} =$

	alcohol	hue
Wine9	13.86	1.01

$\boldsymbol{A}_{[9,1]}$ $\boldsymbol{A}_{[9,1]}$

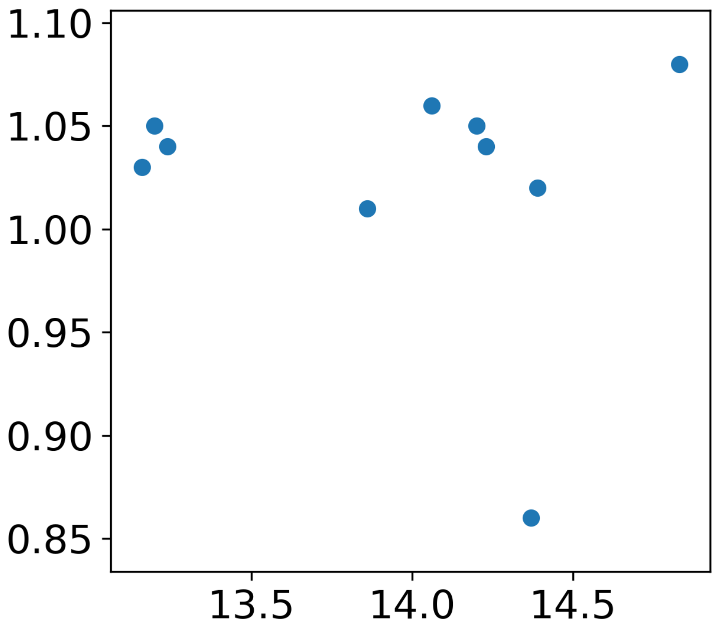
Exercise!

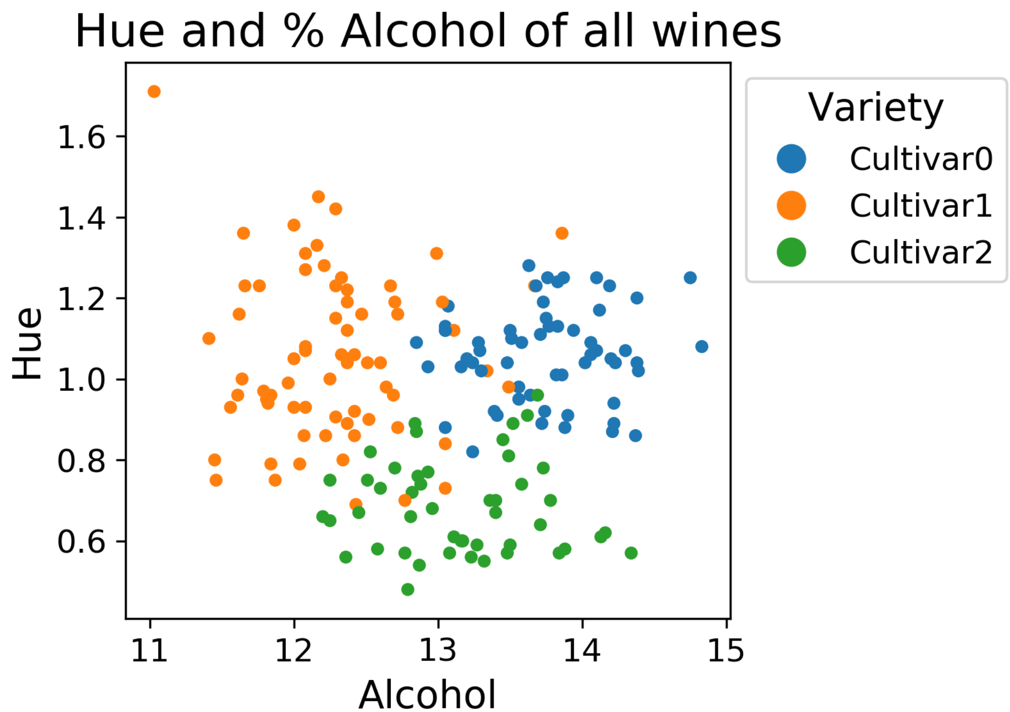
Loading and plotting the UCI wine dataset



Exercise!

Loading and plotting the UCI wine dataset





Introducing scprep

- Code: https://github.com/KrishnaswamyLab/scprep
- Docs: https://scprep.readthedocs.io

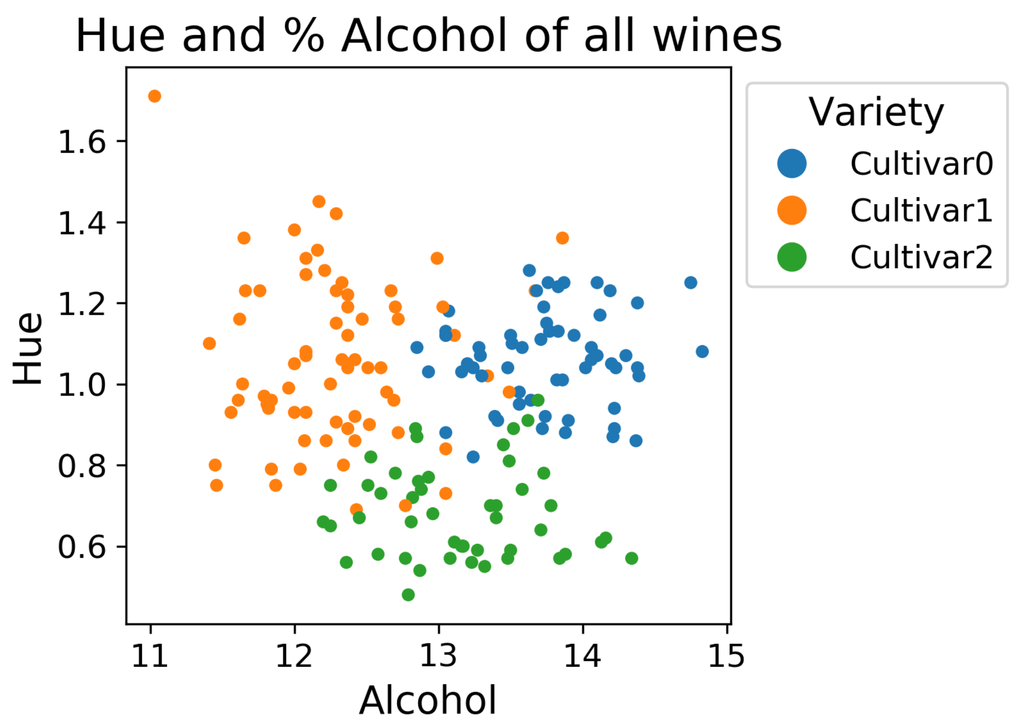
- Modules
  - filter
  - io
  - measure
  - normalize
  - **plot**
  - reduce
  - run
  - select
  - stats
  - transform
  - utils

Plotting functions:

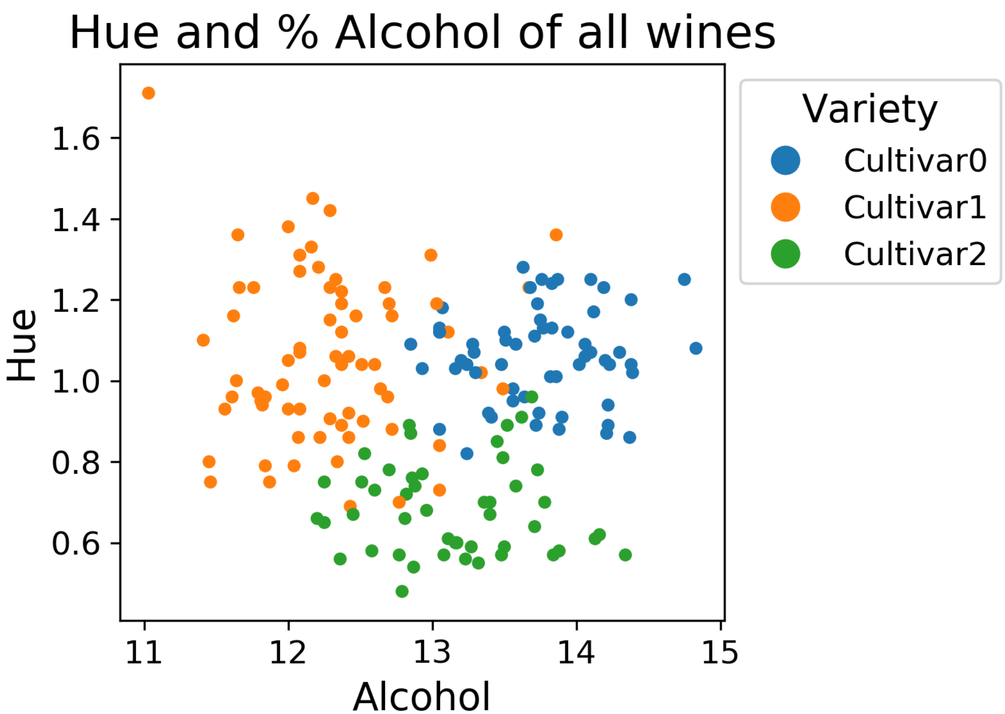
- histogram
- jitter
- marker
- marker_plot
- plot_gene_set_expression
- plot_library_size
- rotate_scatter3d
- scatter
- **scatter2d** = `scprep.plot.scatter2d`
- scatter3d
- scree
- scree_plot

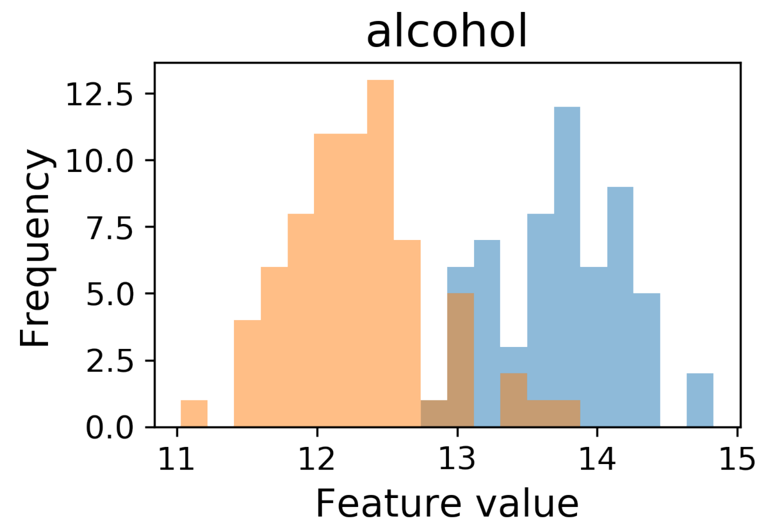
Exercise!

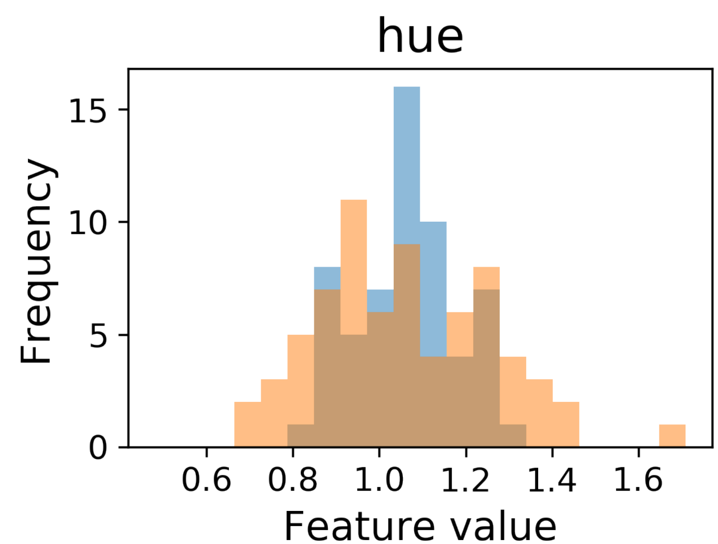
Loading and plotting the UCI wine dataset



How do features differentiate groups within a dataset?

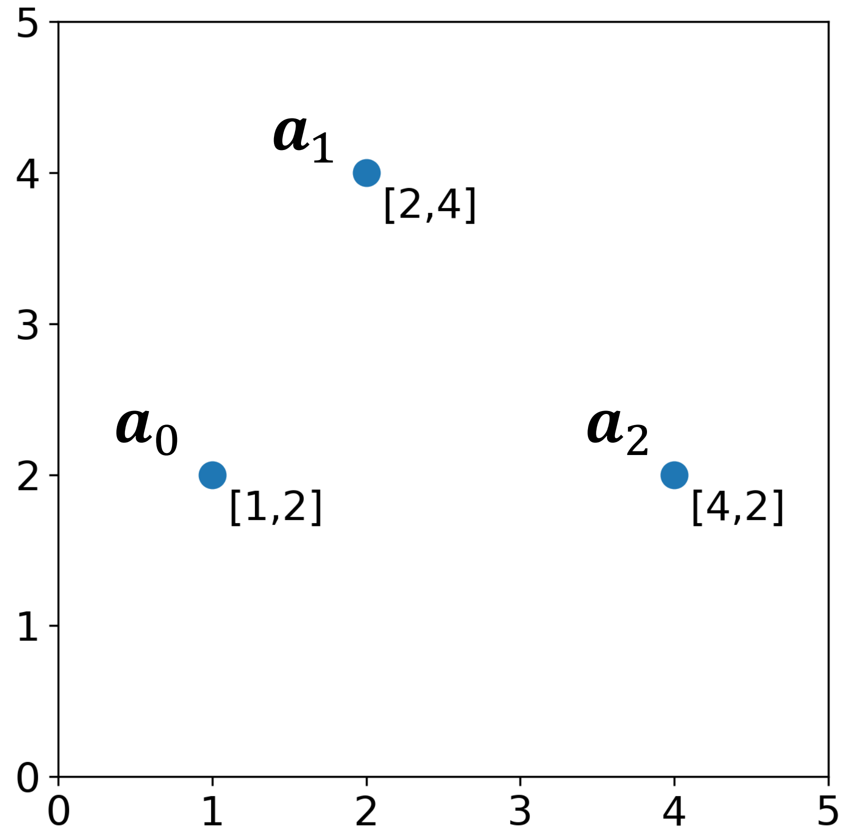




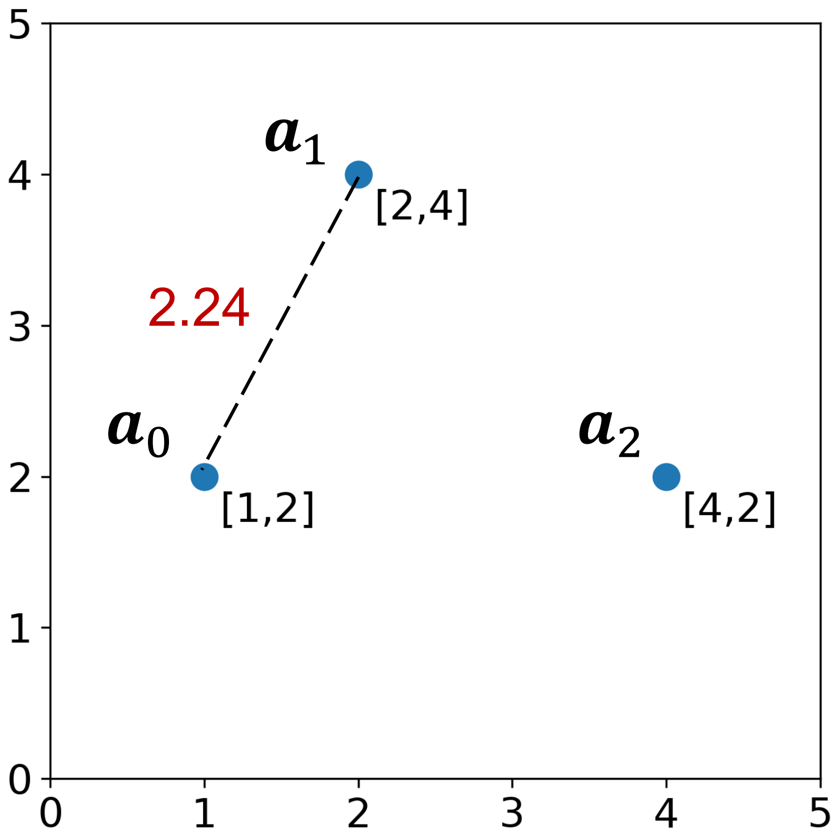


Introduction to Manifold Learning

How far away is each point from a_0?



How far away is each point from $\boldsymbol{a}_0$?



Euclidean distance

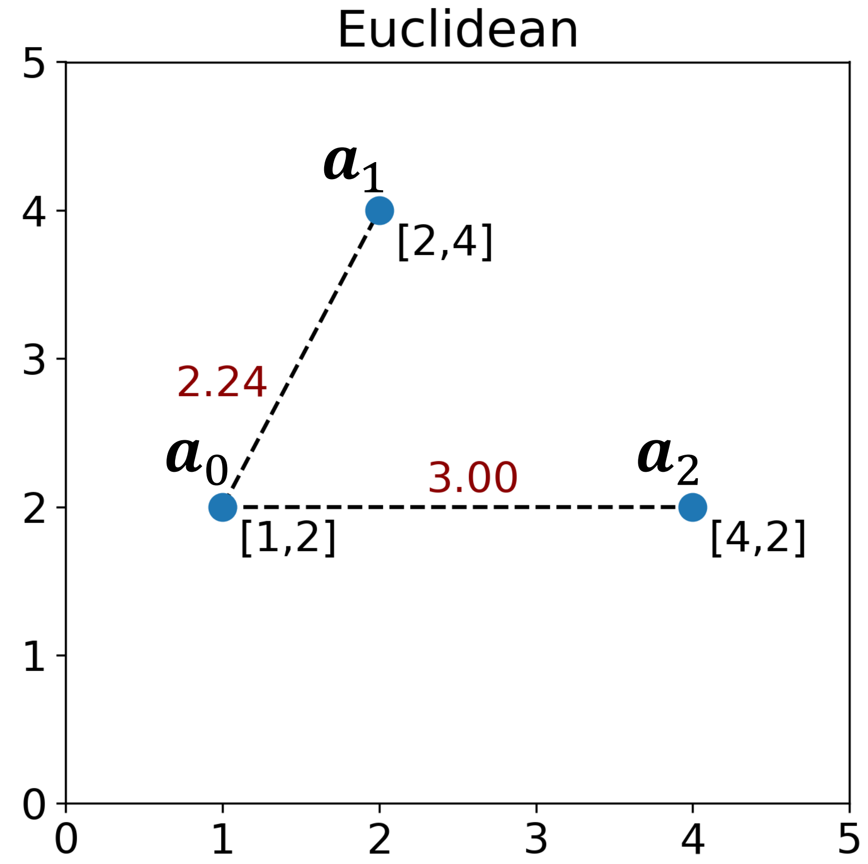
$$d_{euclidean}(a_0, a_1) = \ \|a_0 - a_1\|_2^2$$

$$= \sqrt{(a_{0,0} - a_{1,0})^2 + (a_{0,1} - a_{1,1})^2}$$

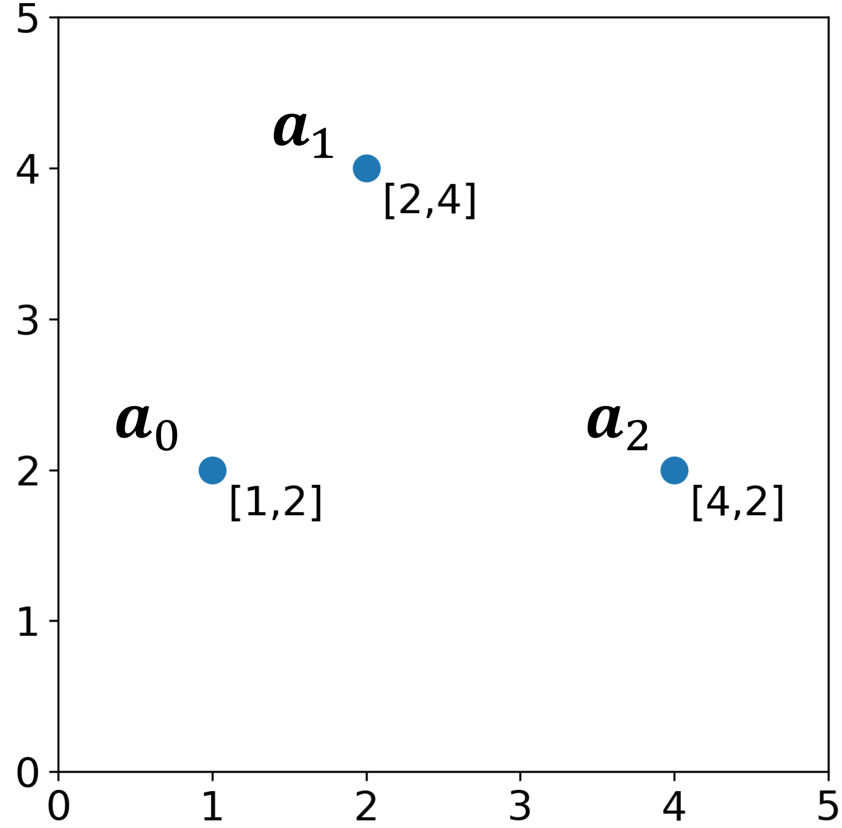
$$= \sqrt{(1 \ - 2)^2 + (2 \ - 4)^2}$$

$$\approx 2.24$$

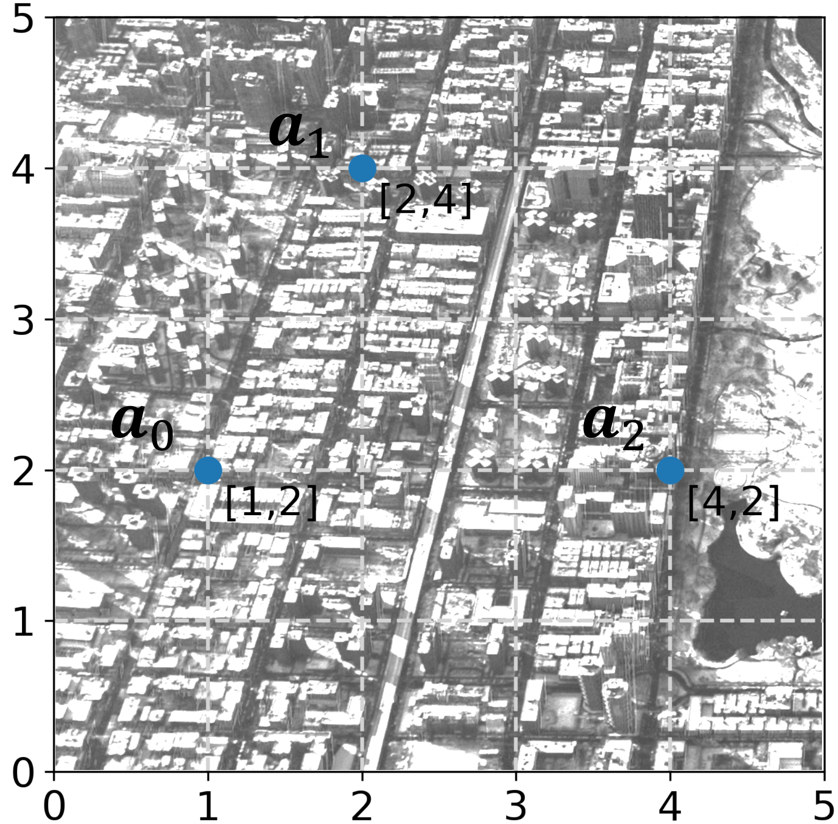
How far away is each point from a_0?



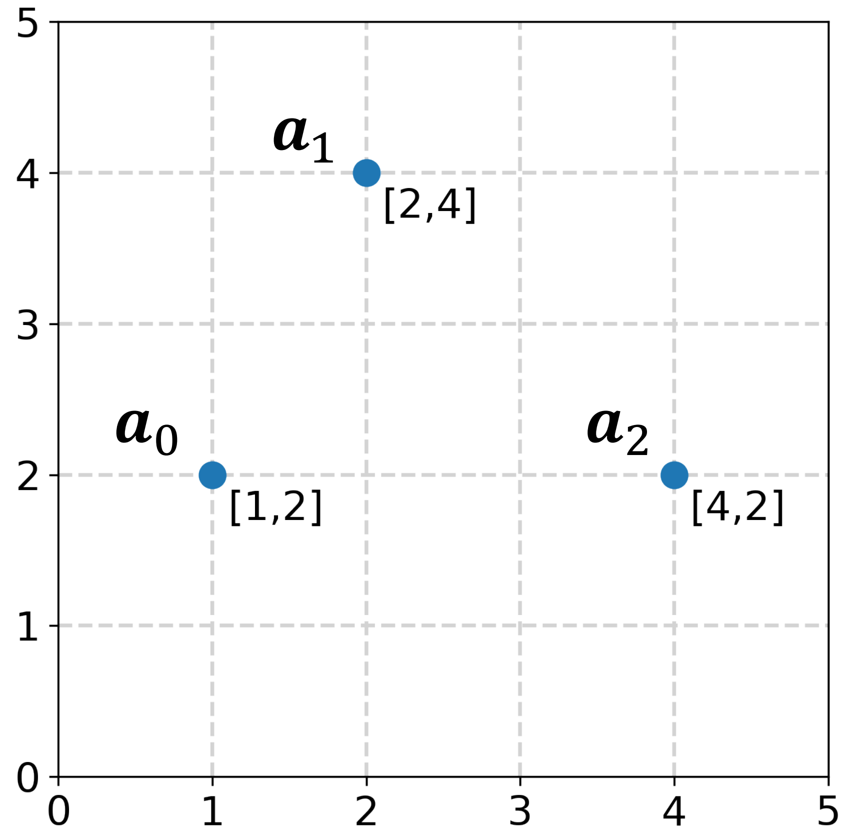
How far away is each point from a_0?



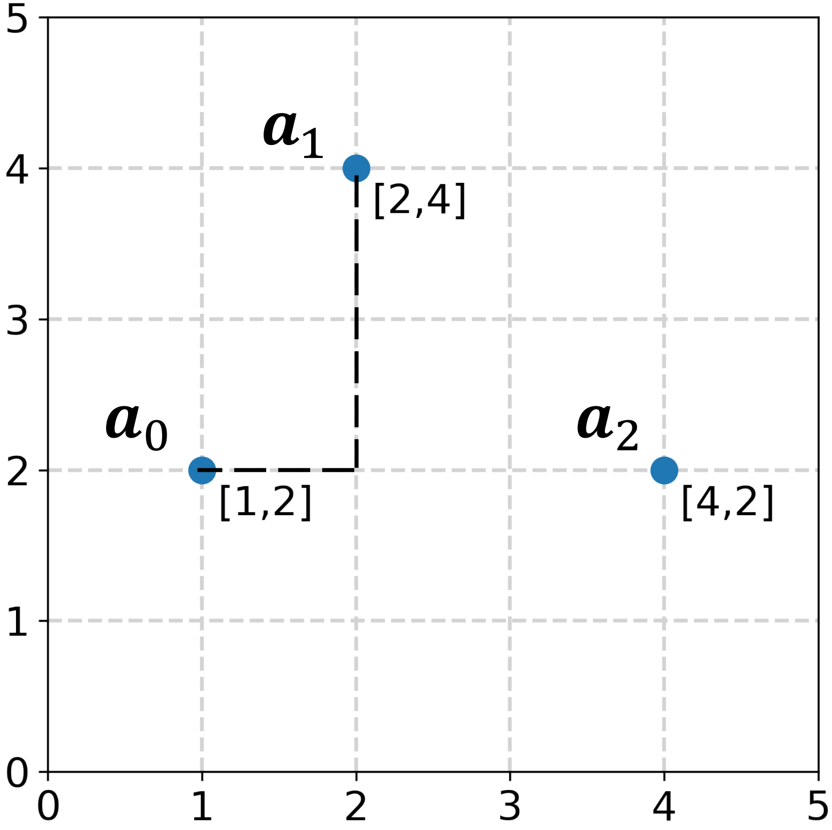
How far away is each point from $\boldsymbol{a}_0$?



How far away is each point from a_0?



How far away is each point from $\boldsymbol{a}_0$?



Manhattan distance

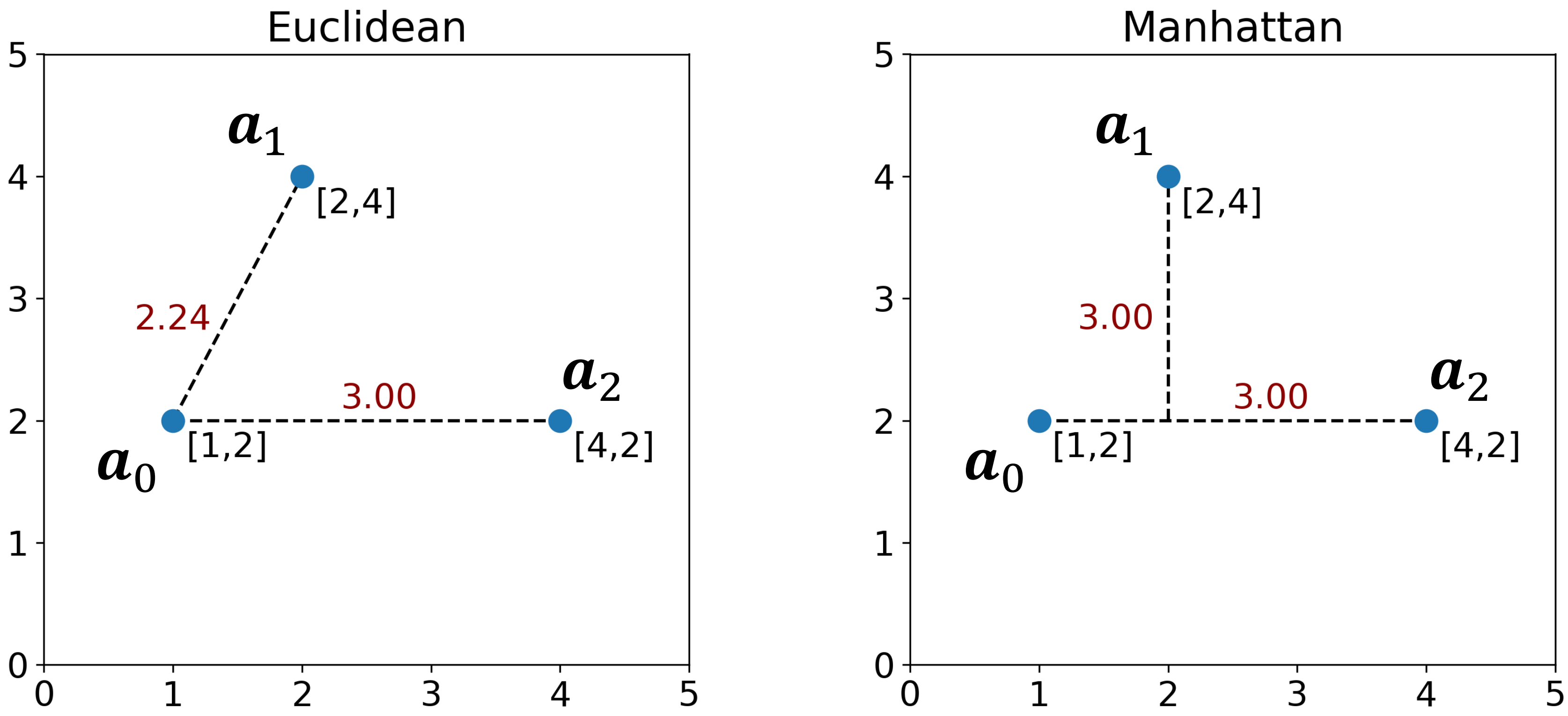
$$d_{manhattan}(a_0, a_1) = |a_{0,1} - a_{1,1}|$$

$$= |a_{0,0} - a_{1,0}| + |a_{0,1} - a_{1,1}|$$

$$= |1 - 2| + |2 - 4|$$

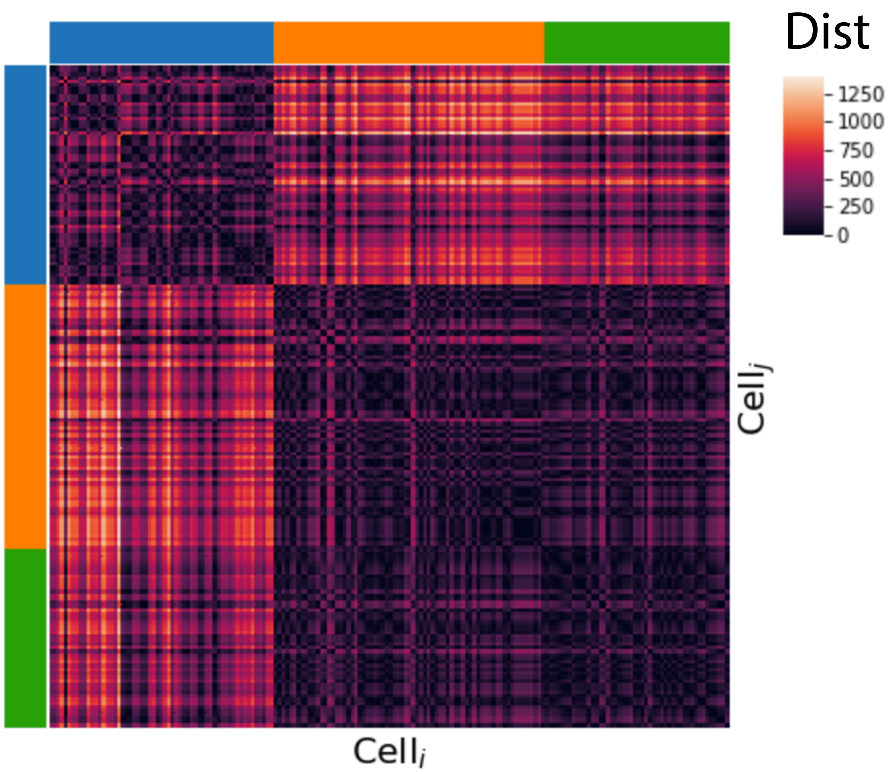
$$= 3$$

How far away is each point from a_0?



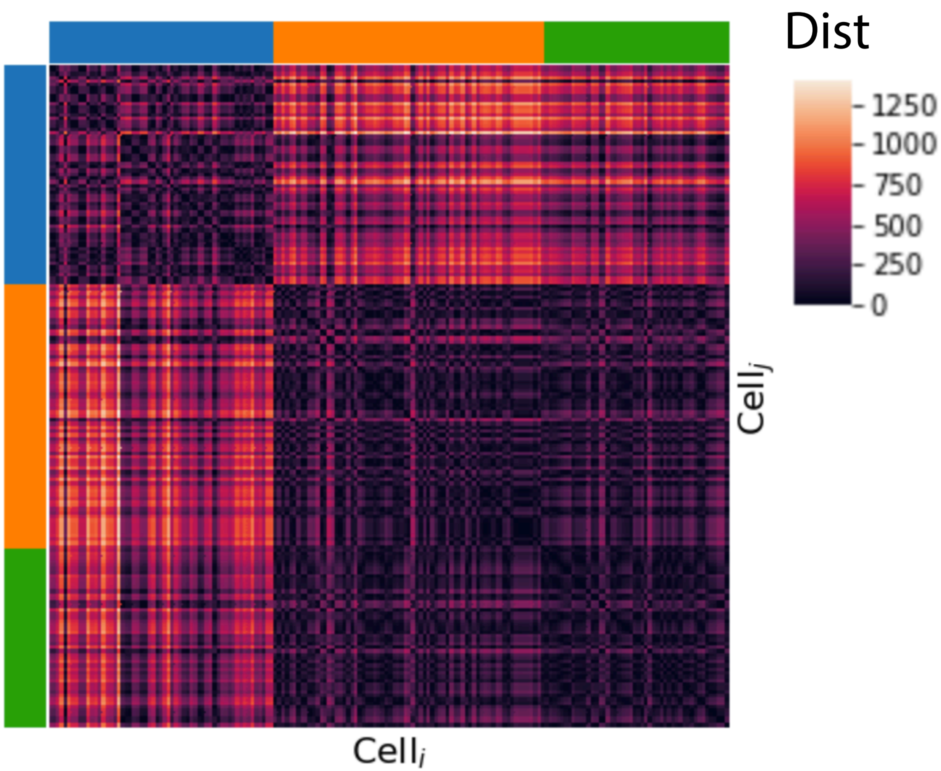
Exercise!

Calculating distances



Exercise!

Calculating distances

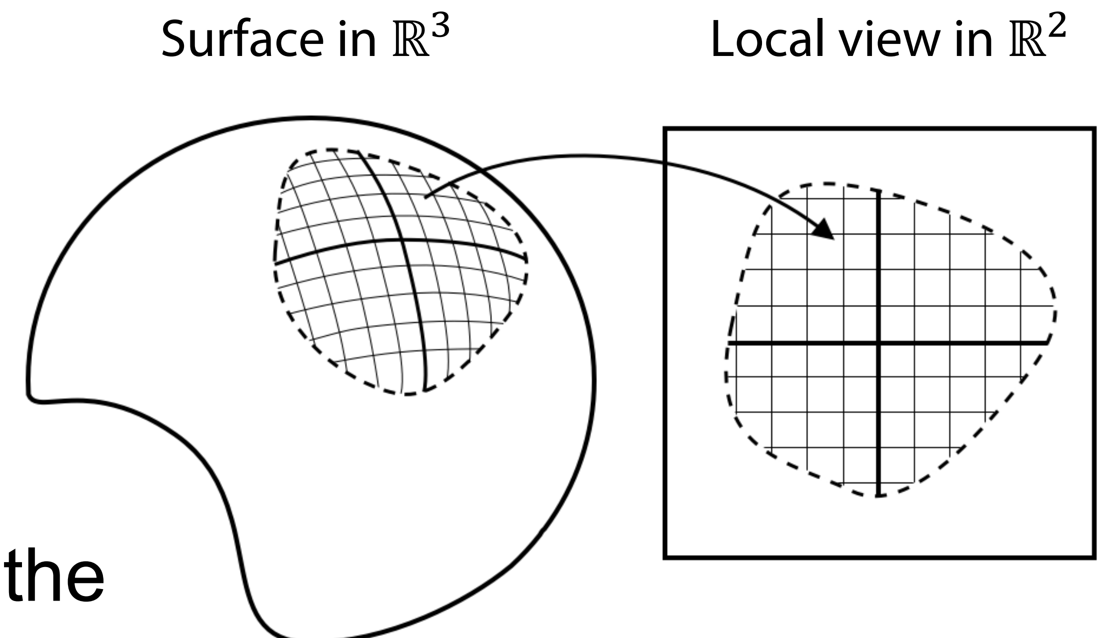


Distance defines **structure** in a dataset

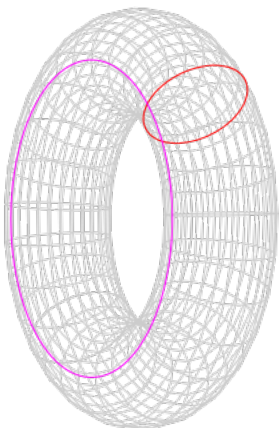
- How **different** are two data points?
- How many **types** exist in my dataset?
- Which types are **closest** to each other?
- Which **features** vary between dissimilar groups of data?

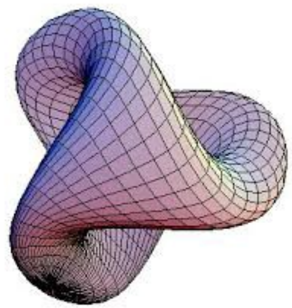
What is a manifold?

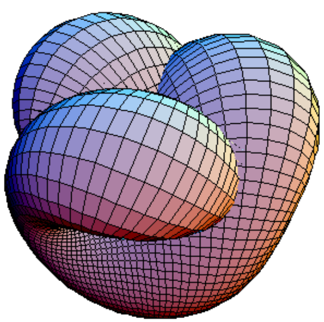
- Locally smooth
- Locally Euclidean
- Generally, lower dimensional than the ambient space (i.e. a subspace)

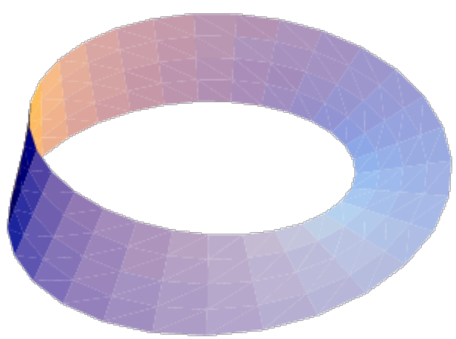


Examples of manifolds



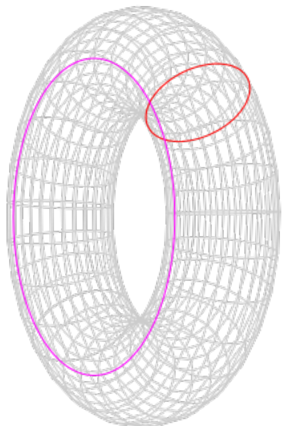


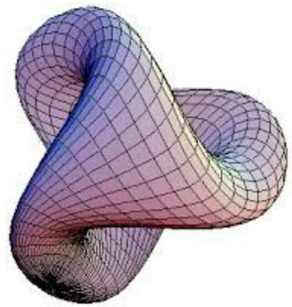


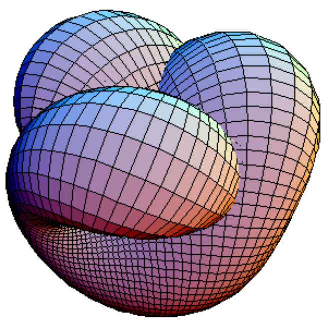


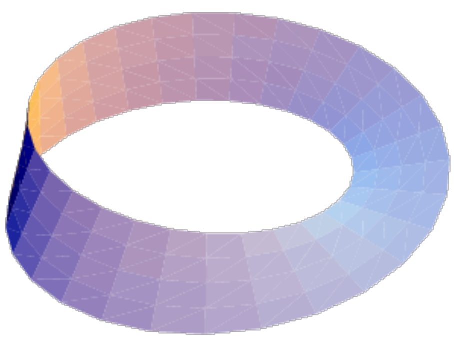
Wolfram

Examples of manifolds

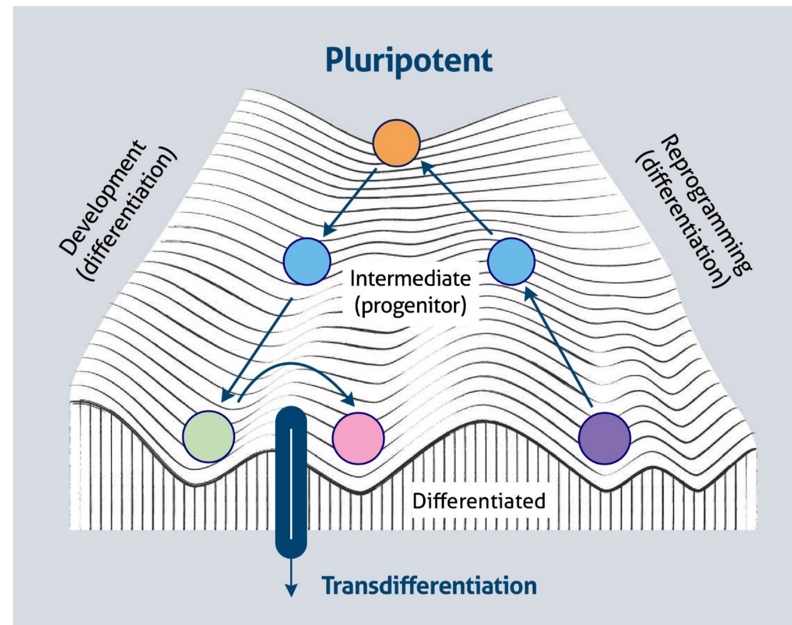




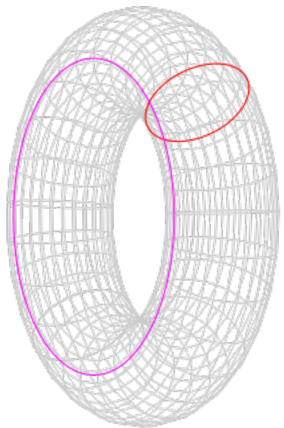


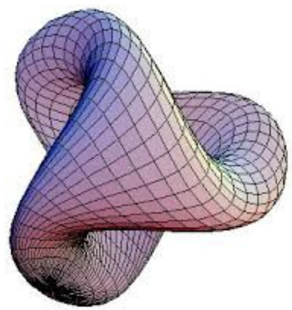


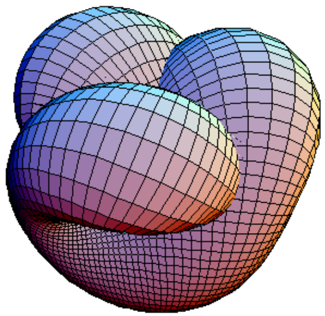
Waddington's landscape

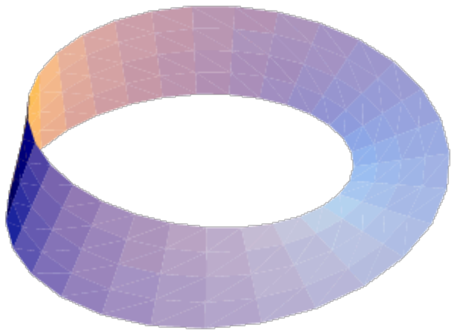


Examples of manifolds

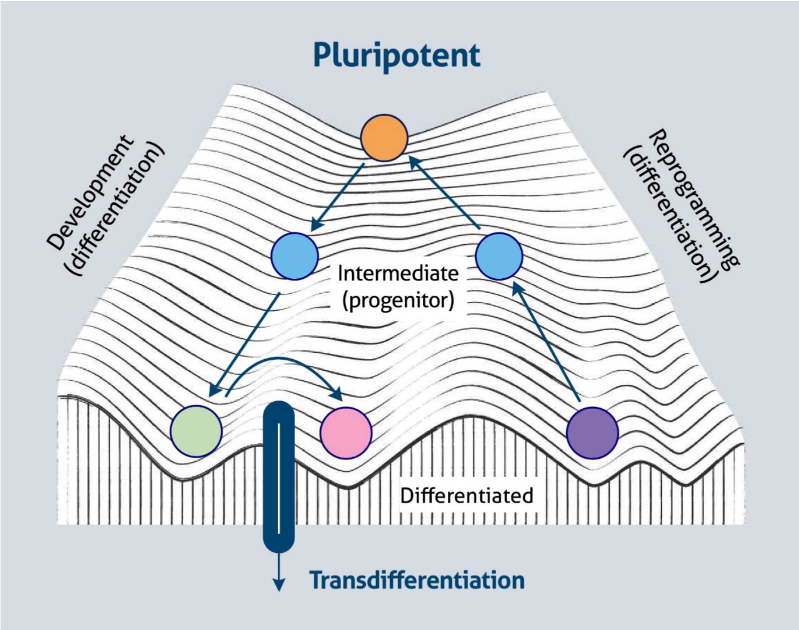








Waddington's landscape



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

Wolfram

ptglab.com

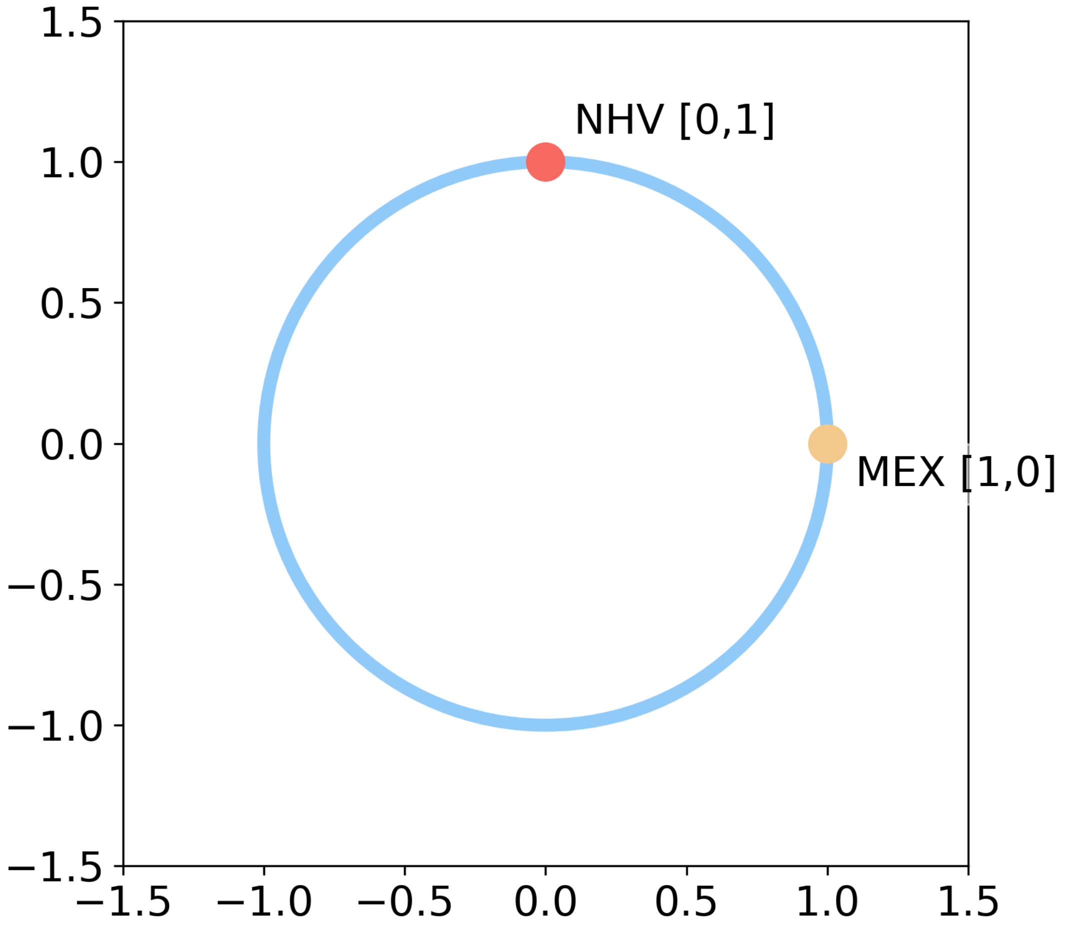
NHV

↓

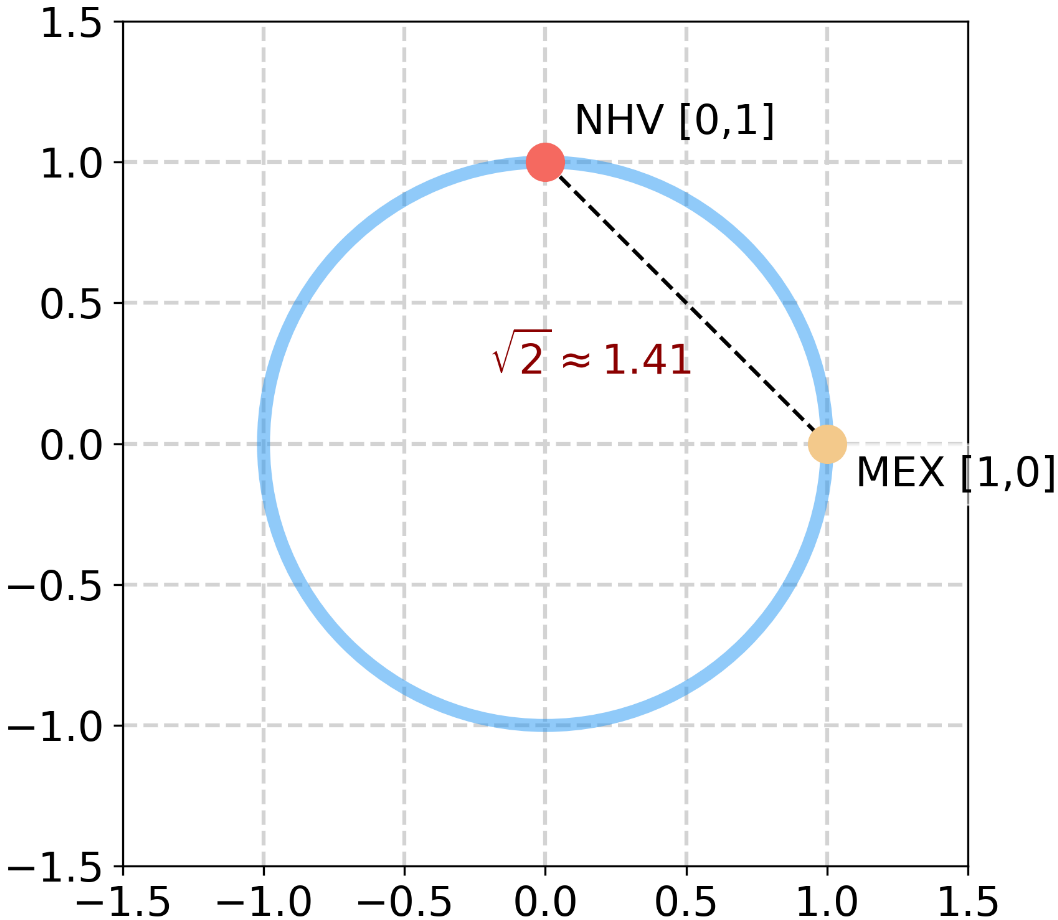
MEX

Arctic Ocean
Arctic Ocean
East Siberian Sea
Bering Sea
Greenland Sea
Baffin Bay
Northwestern Passages
Hudson Bay
Labrador Sea
Canada
Greenland
Iceland
Sweden
Norway
Belarus
Poland
United Kingdom
Ireland
Romania
Turkey
France
Italy
Spain
Portugal
Mediterranean Sea
Tunisia
Egypt
Libya
Algeria
Western Sahara
Mauritania
Mali
Chad
Nigeria
Guinea
Ghana
Gabon
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Nicaragua
Caribbean Sea
North Atlantic Ocean
Venezuela
Guyana
Suriname
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
South Atlantic Ocean

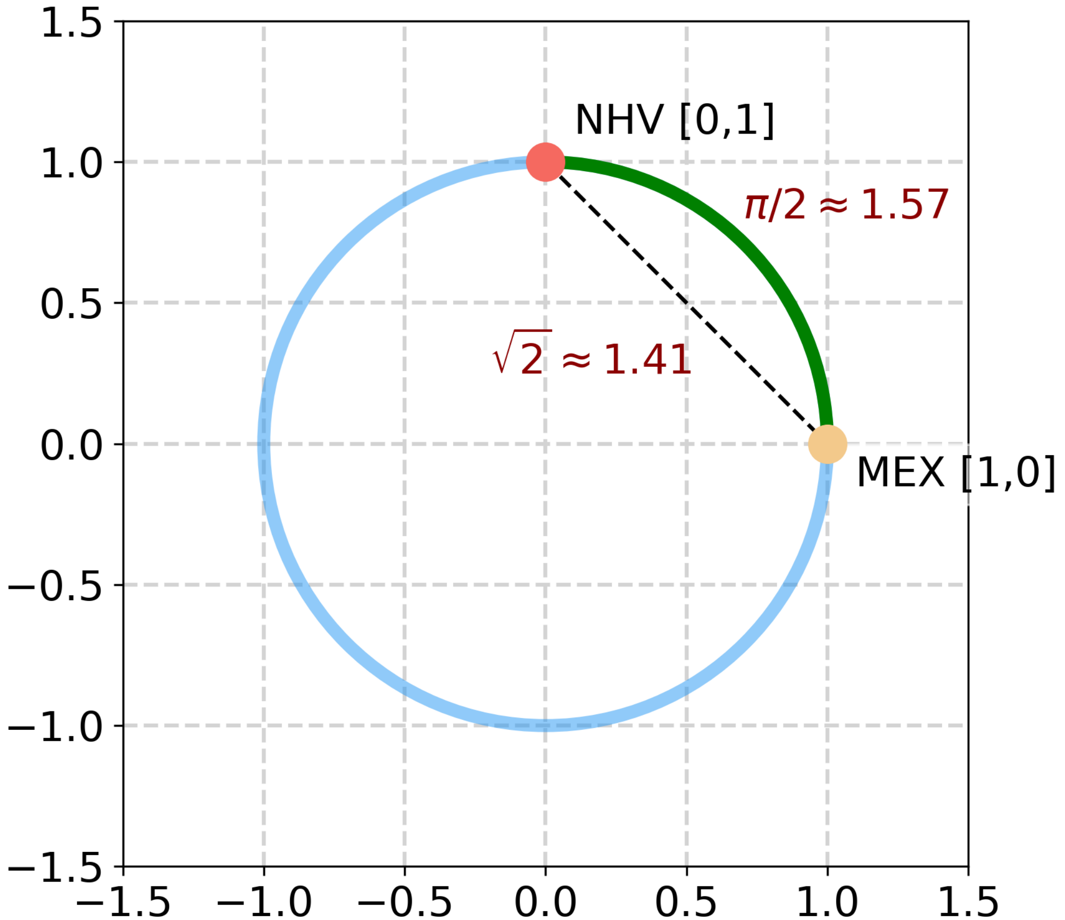
How far is NHV -> MEX?



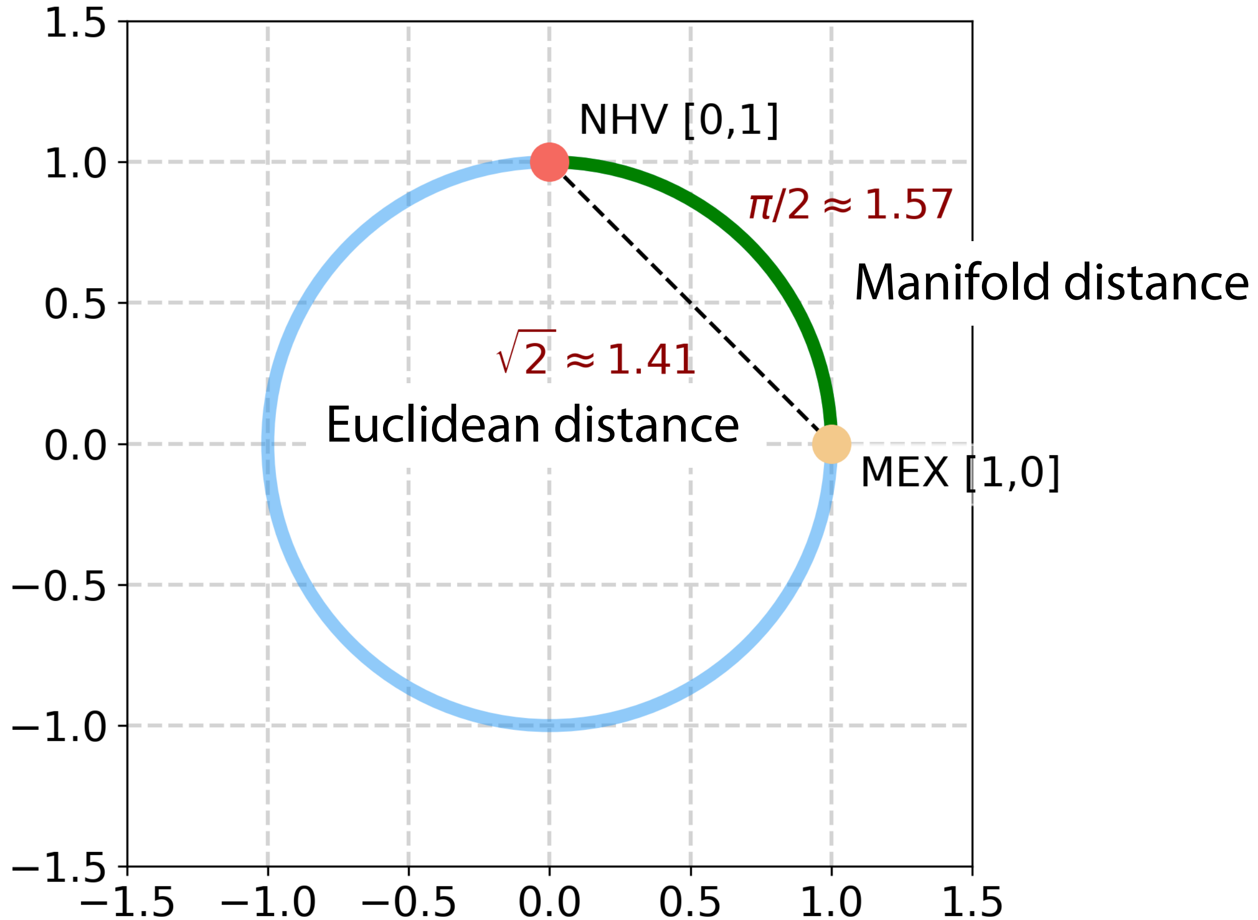
How far is NHV -> MEX?



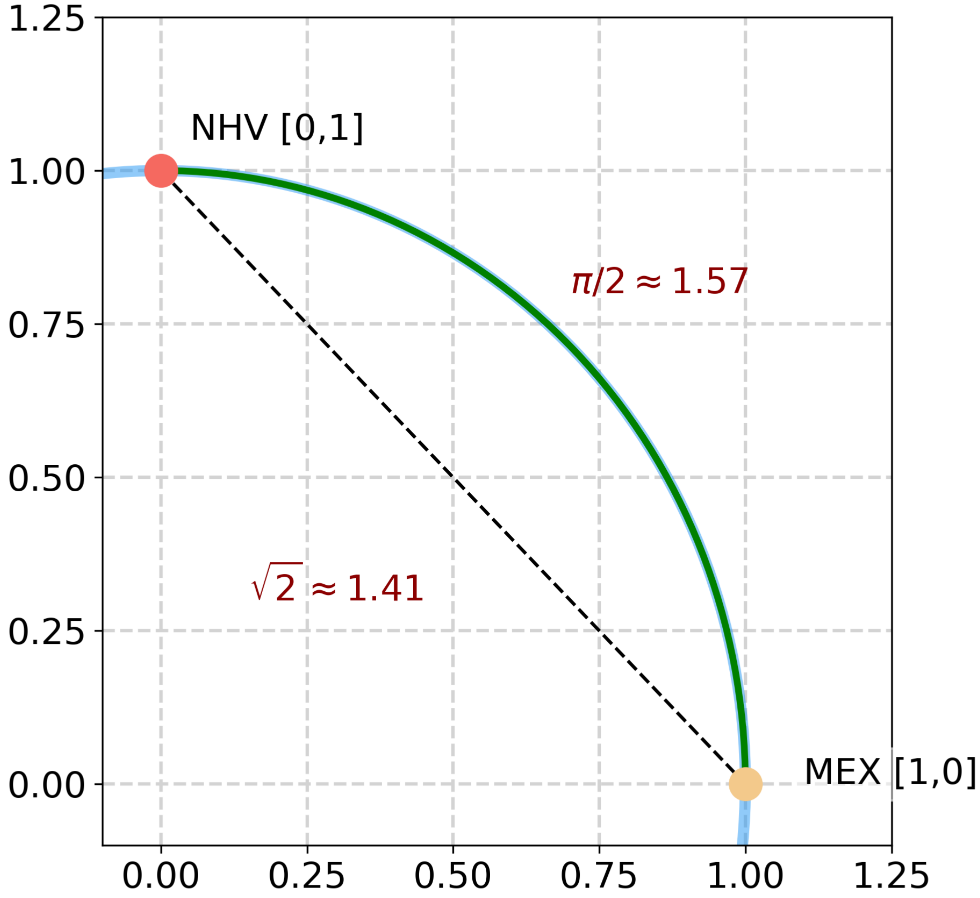
How far is NHV -> MEX?



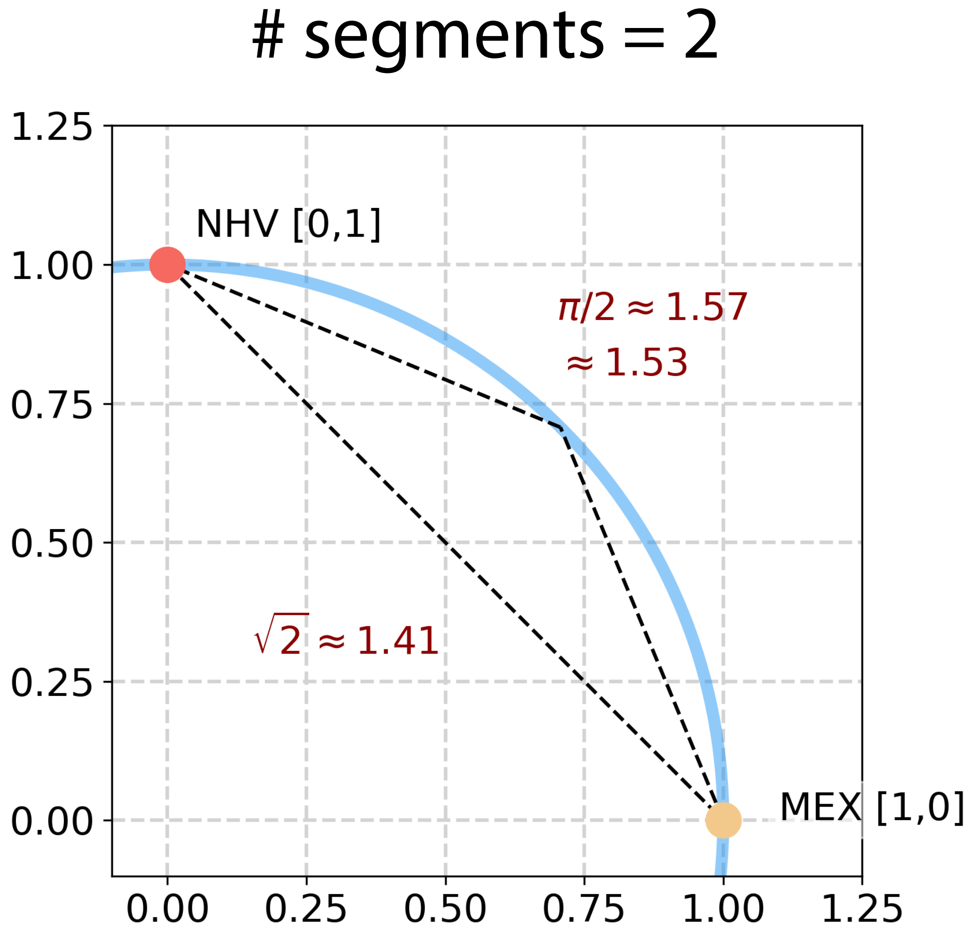
How far is NHV -> MEX?



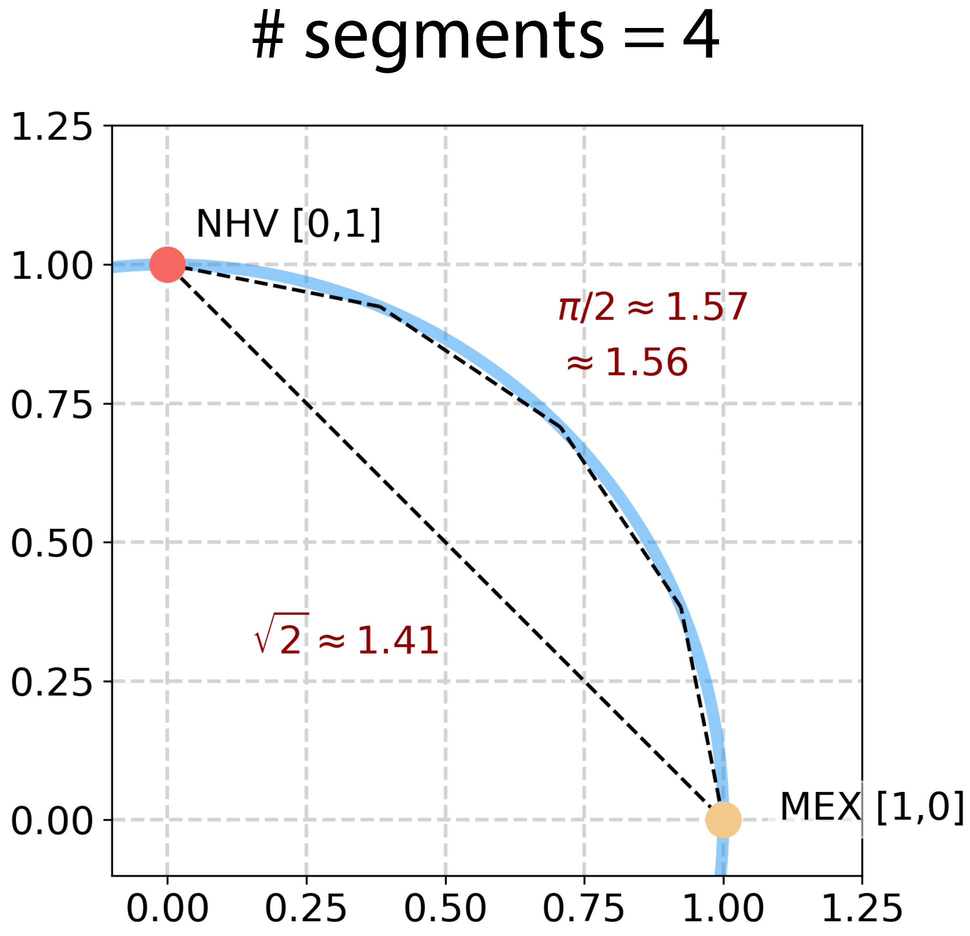
Is a circle locally Euclidean?



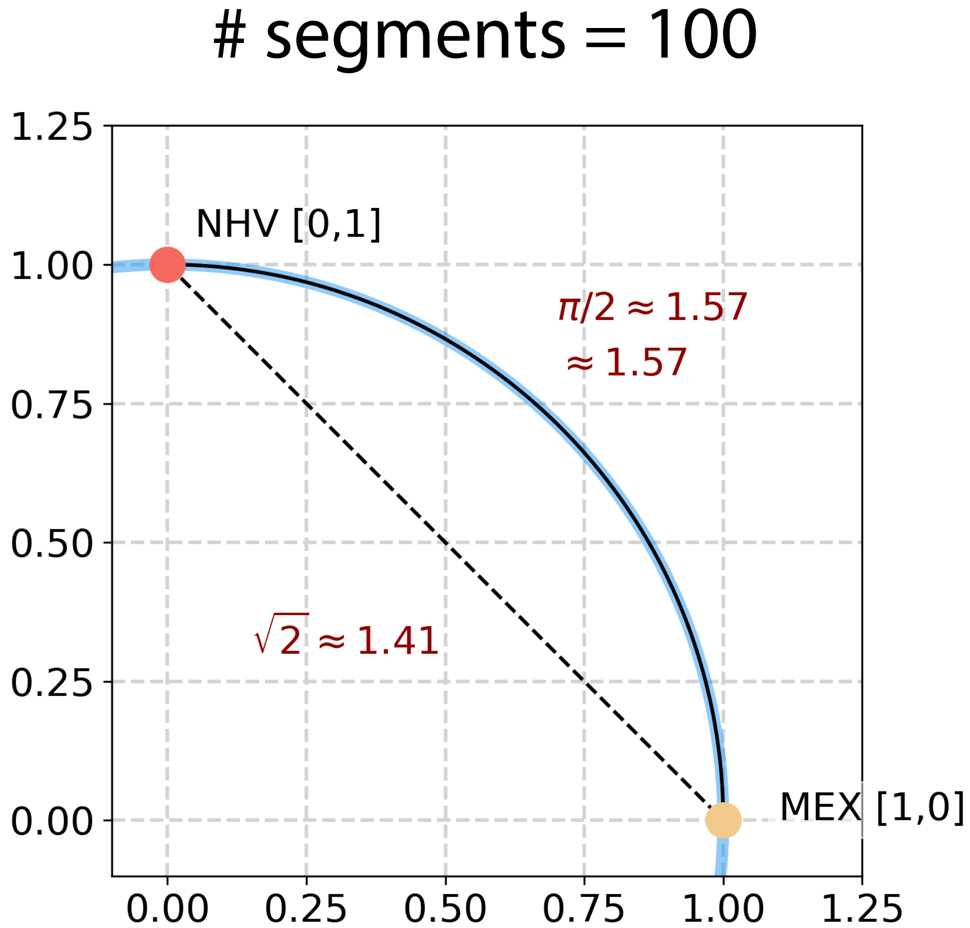
Is a circle locally Euclidean?



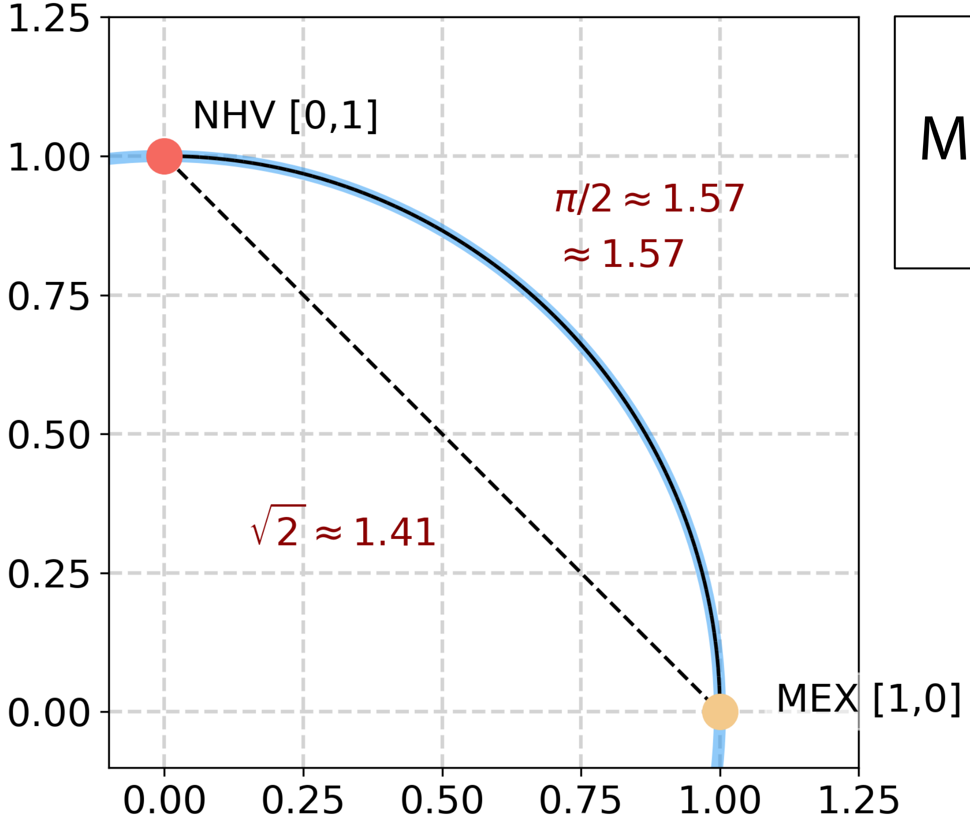
Is a circle locally Euclidean?



Is a circle locally Euclidean?

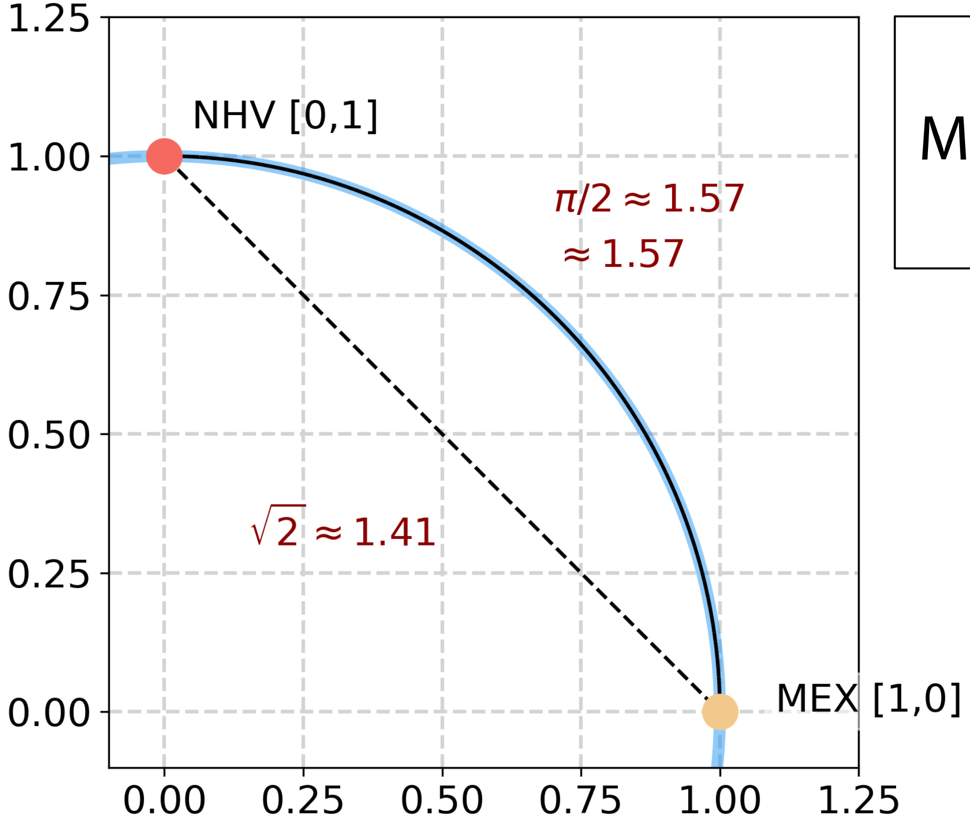


Is a circle locally Euclidean?



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

Is a circle locally Euclidean?

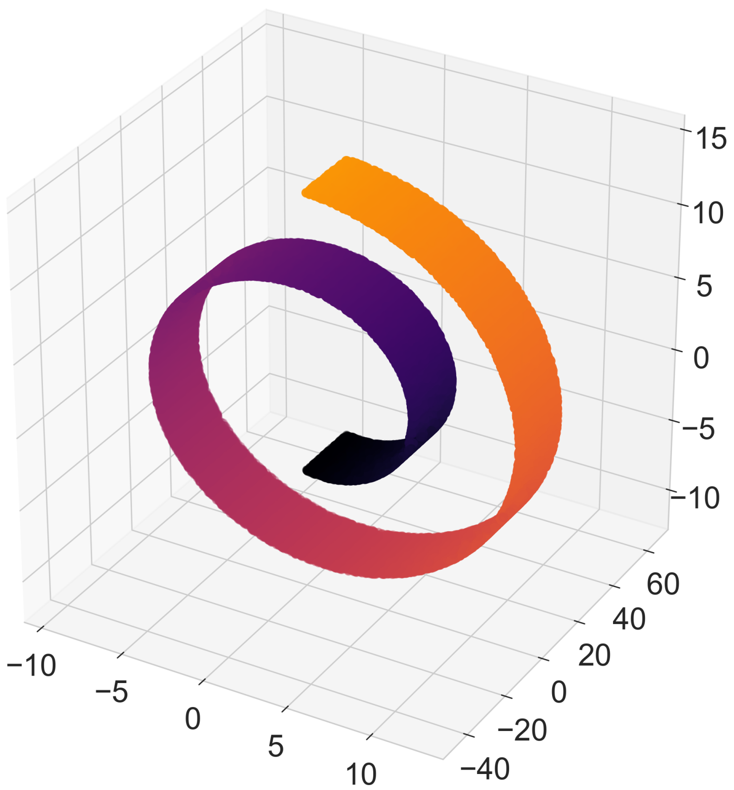


Key Insight:
Manifolds are **locally** Euclidean, but not **globally** Euclidean

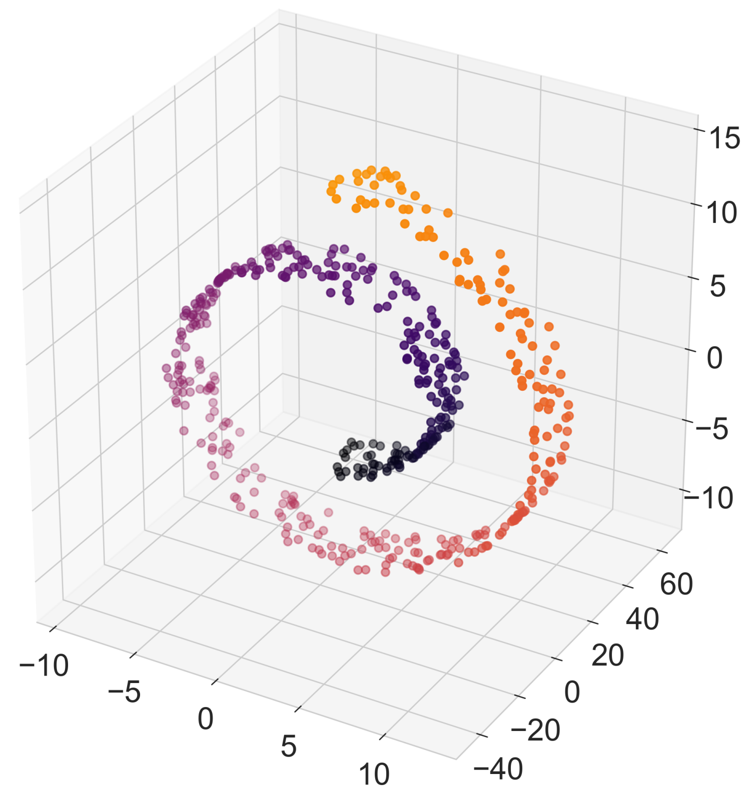
How do we define local versus global when we only have a collection of points?

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

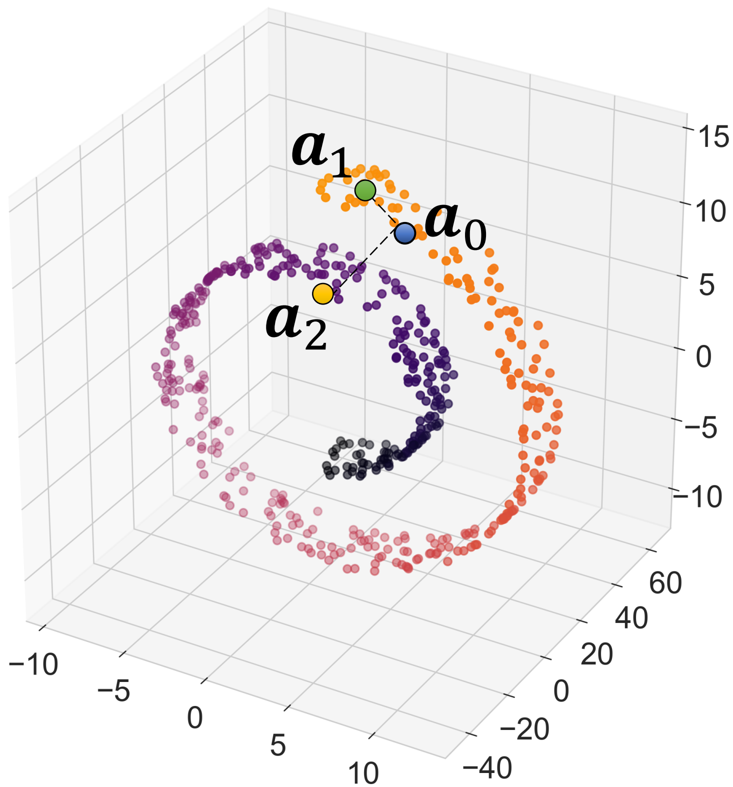
True manifold



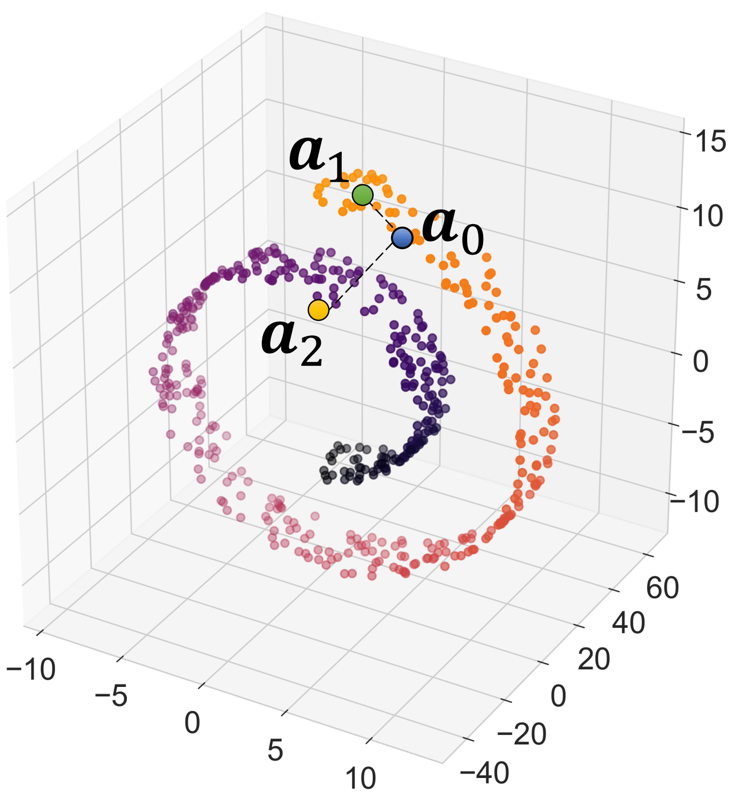
Data sampled from manifold

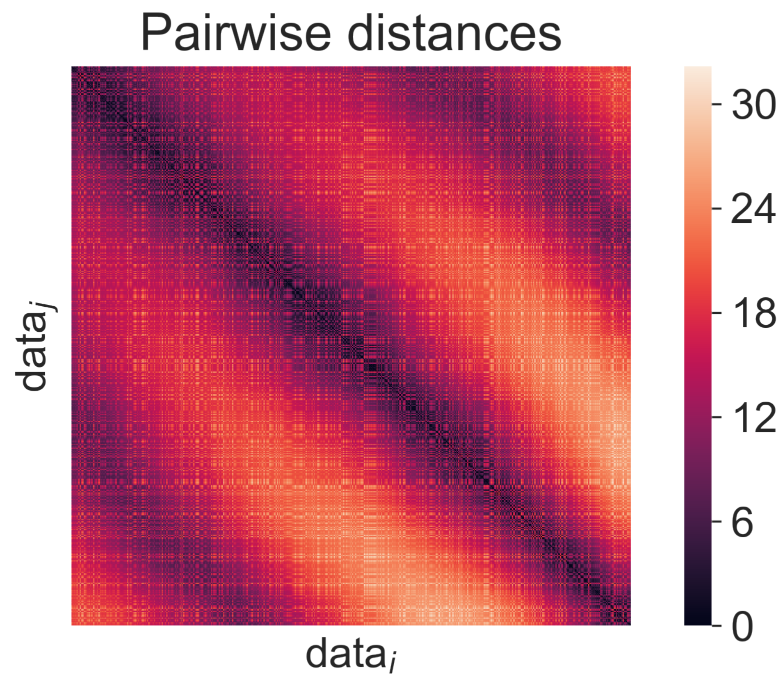


How do we quantify close (local) versus **far (global)** in a dataset?

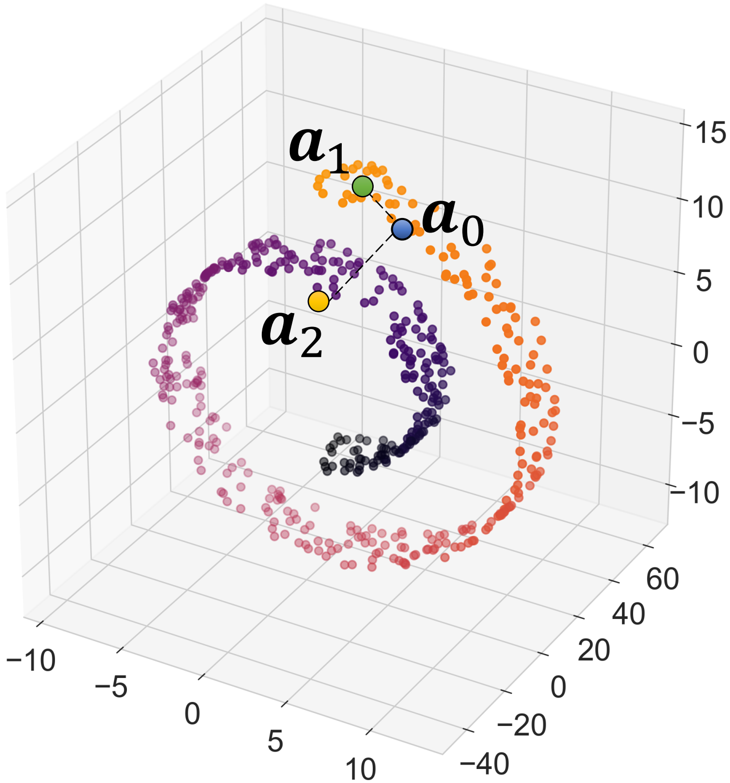


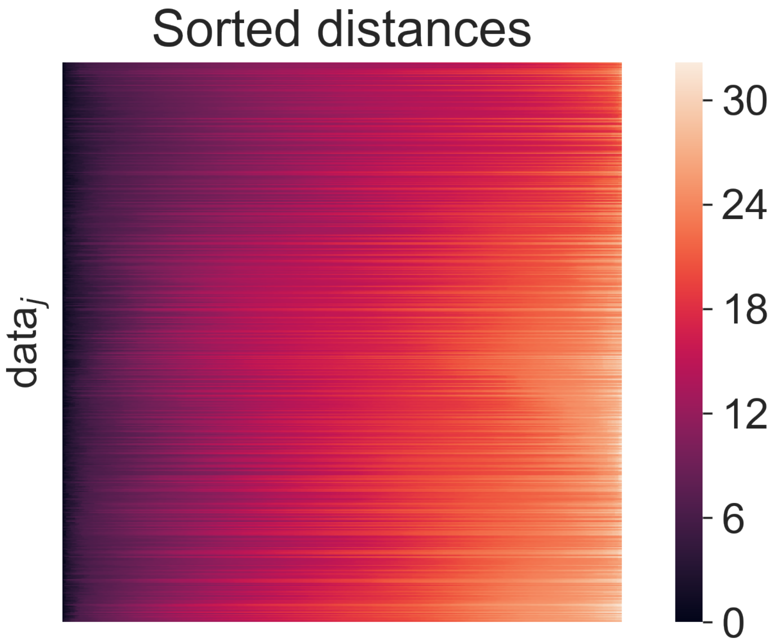
How do we quantify close (local) versus far (global) in a dataset?



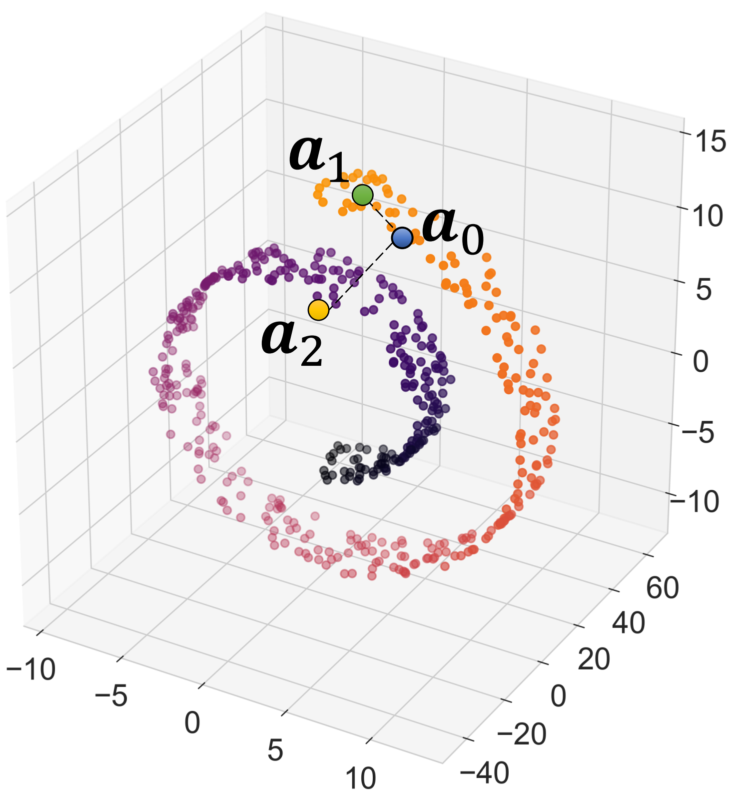


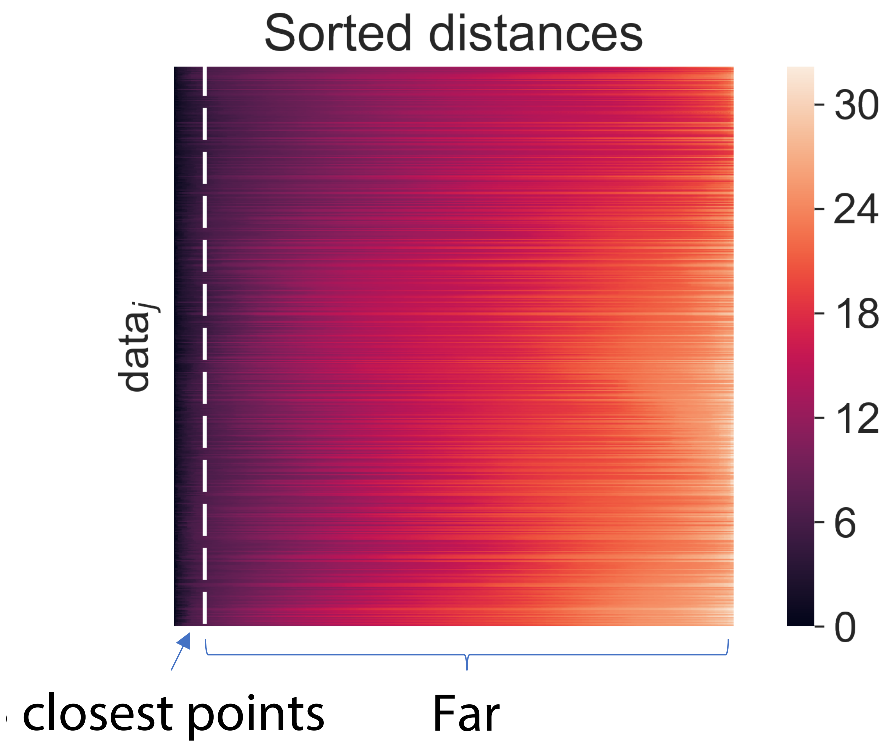
How do we quantify close (local) versus **far (global)** in a dataset?





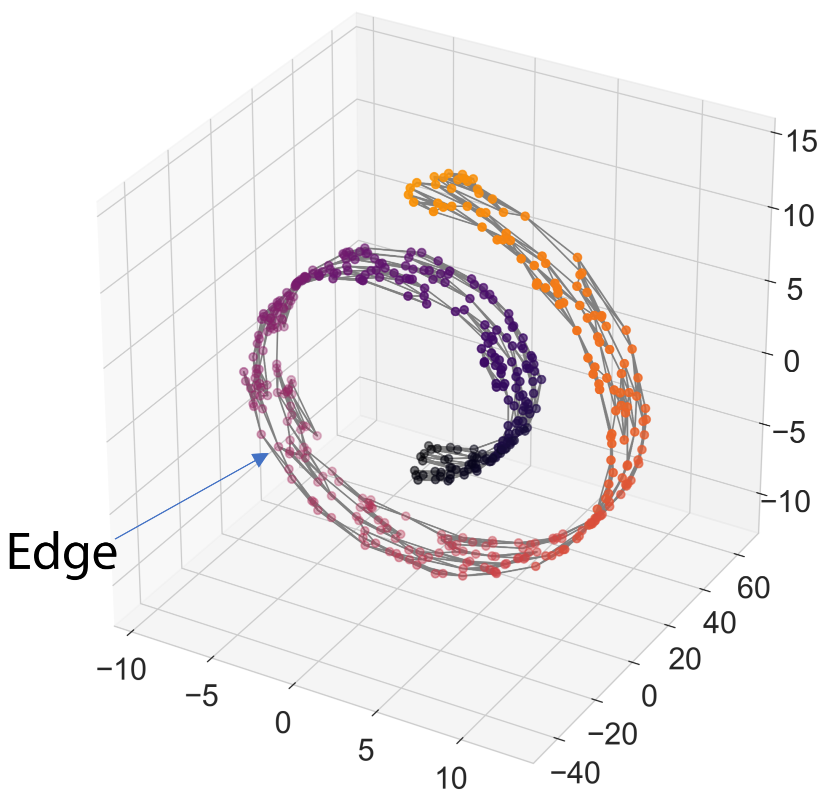
How do we quantify close (local) versus **far (global)** in a dataset?

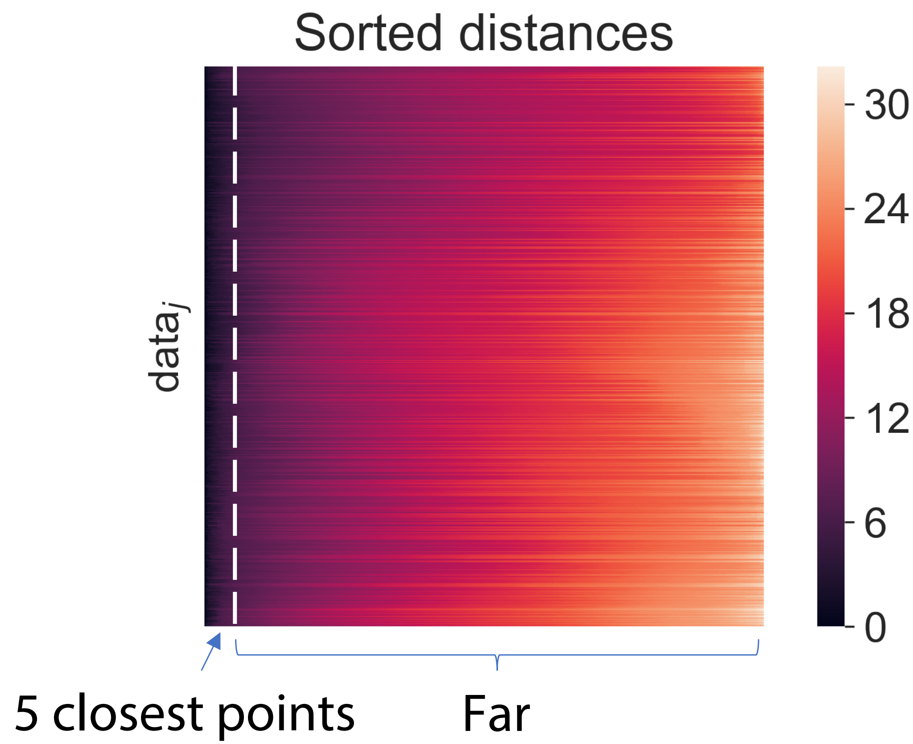




How do we quantify close (local) versus **far (global)** in a dataset?

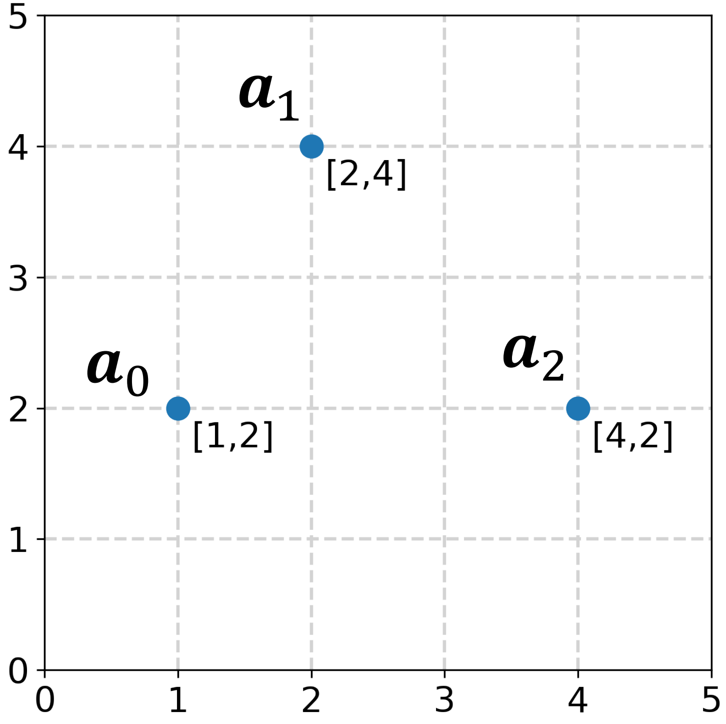
Point out that A2 would be “sort of close” in Euclidean space



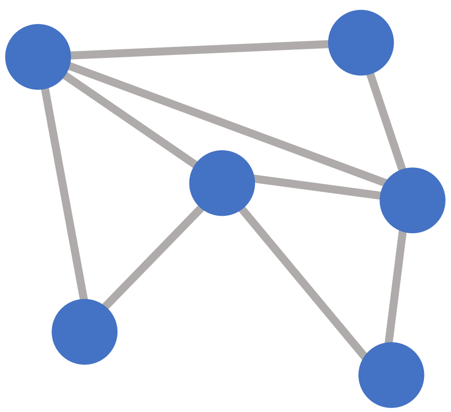


What is a “graph”?

Plot



Graph



$$G = (V, E)$$

● $= V =$ vertices / nodes

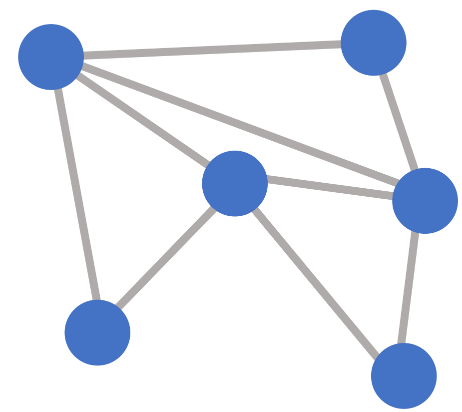
— $= E =$ edges

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

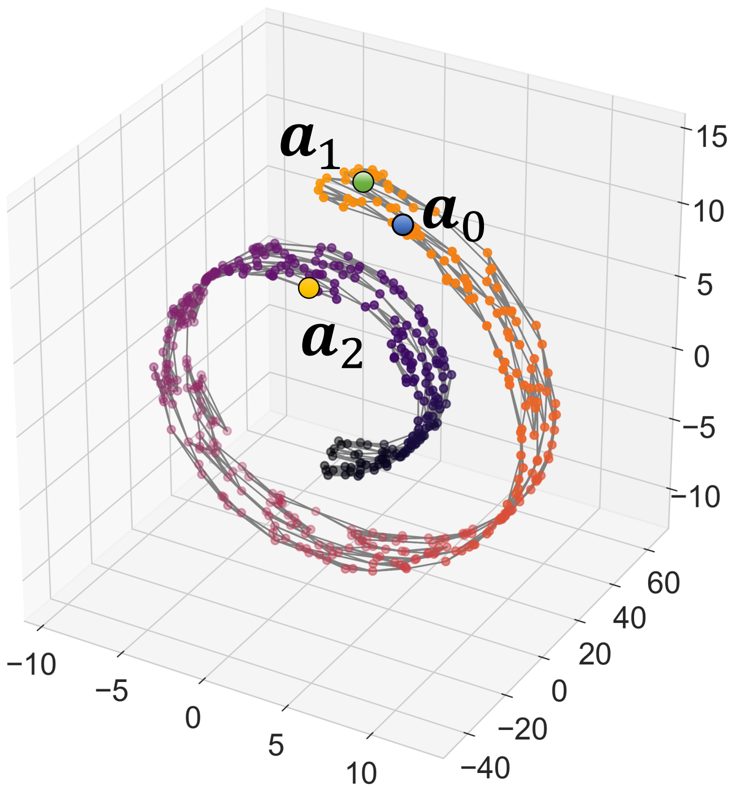


$G = (V, E)$

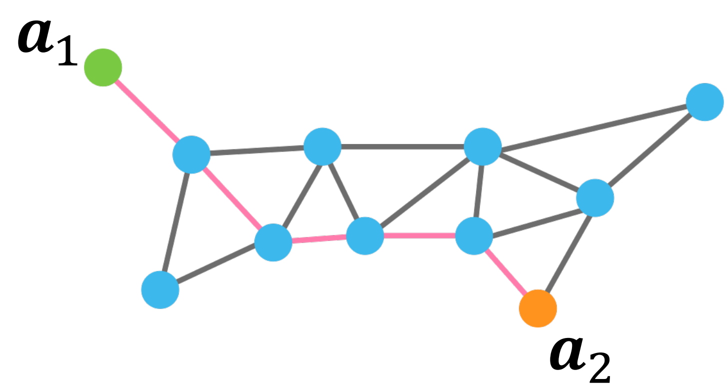
$= V =$ vertices

$= E =$ edges

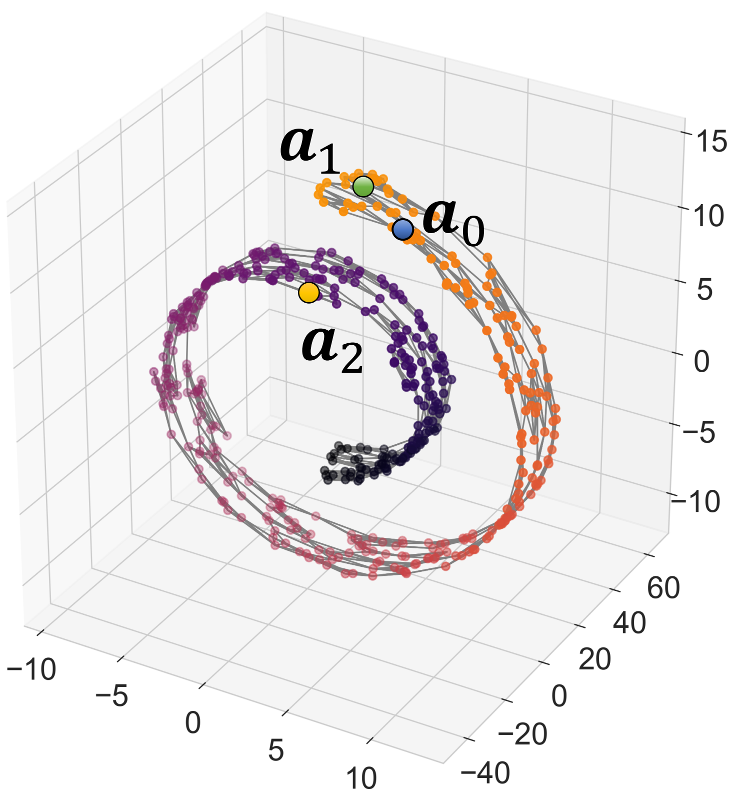
Graph walks approximate manifold distances



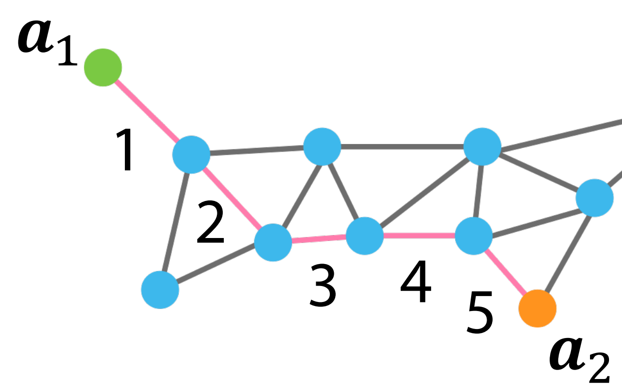
Shortest path between points



Graph walks approximate manifold distances

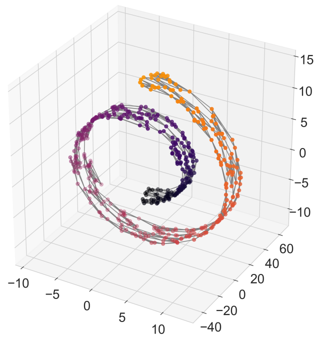


Shortest path between points

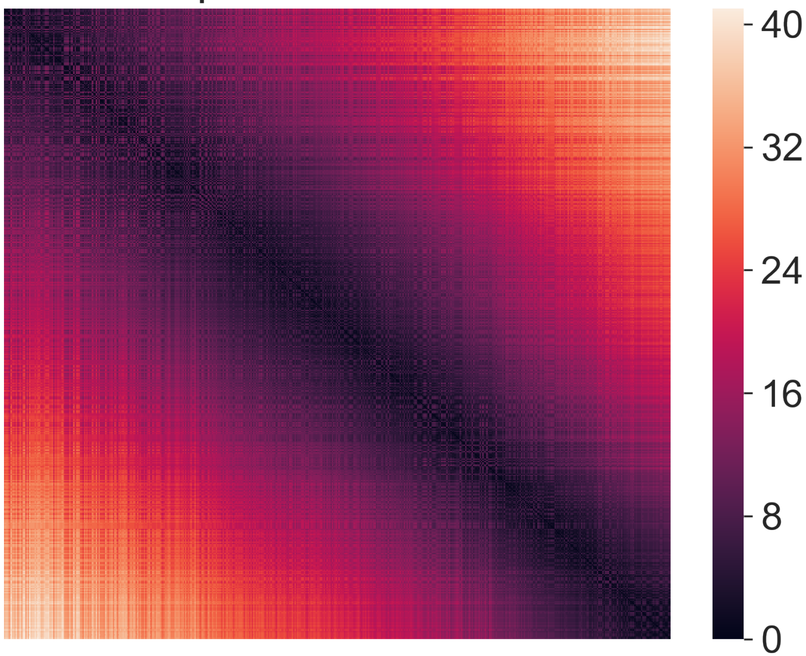


$$d_{geodesic}(a_1, a_2) = \text{shortestpath}(a_1, a_2)$$
$$= 5$$

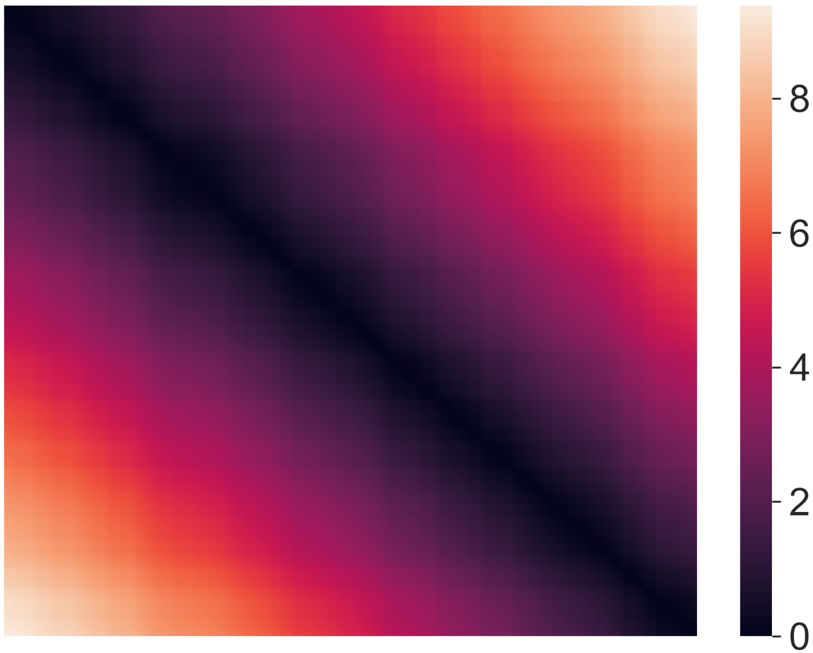
Graph walks approximate manifold distances



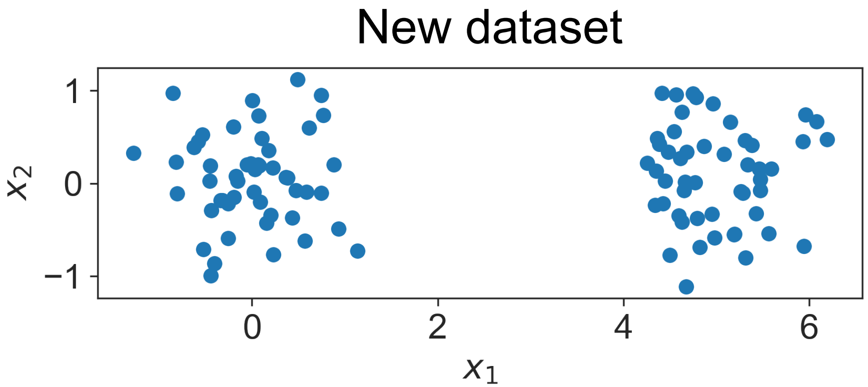
k-NN Graph Geodesic Distances



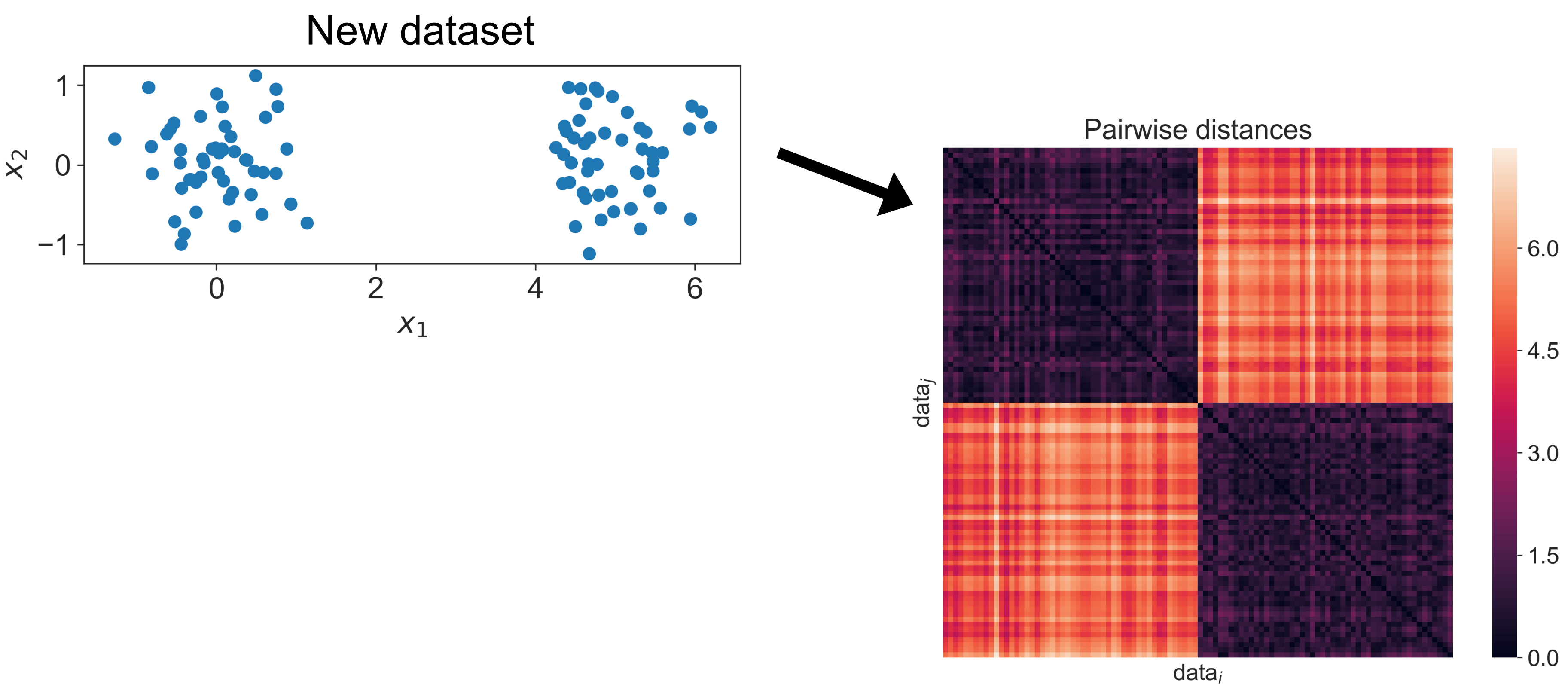
Manifold Distances



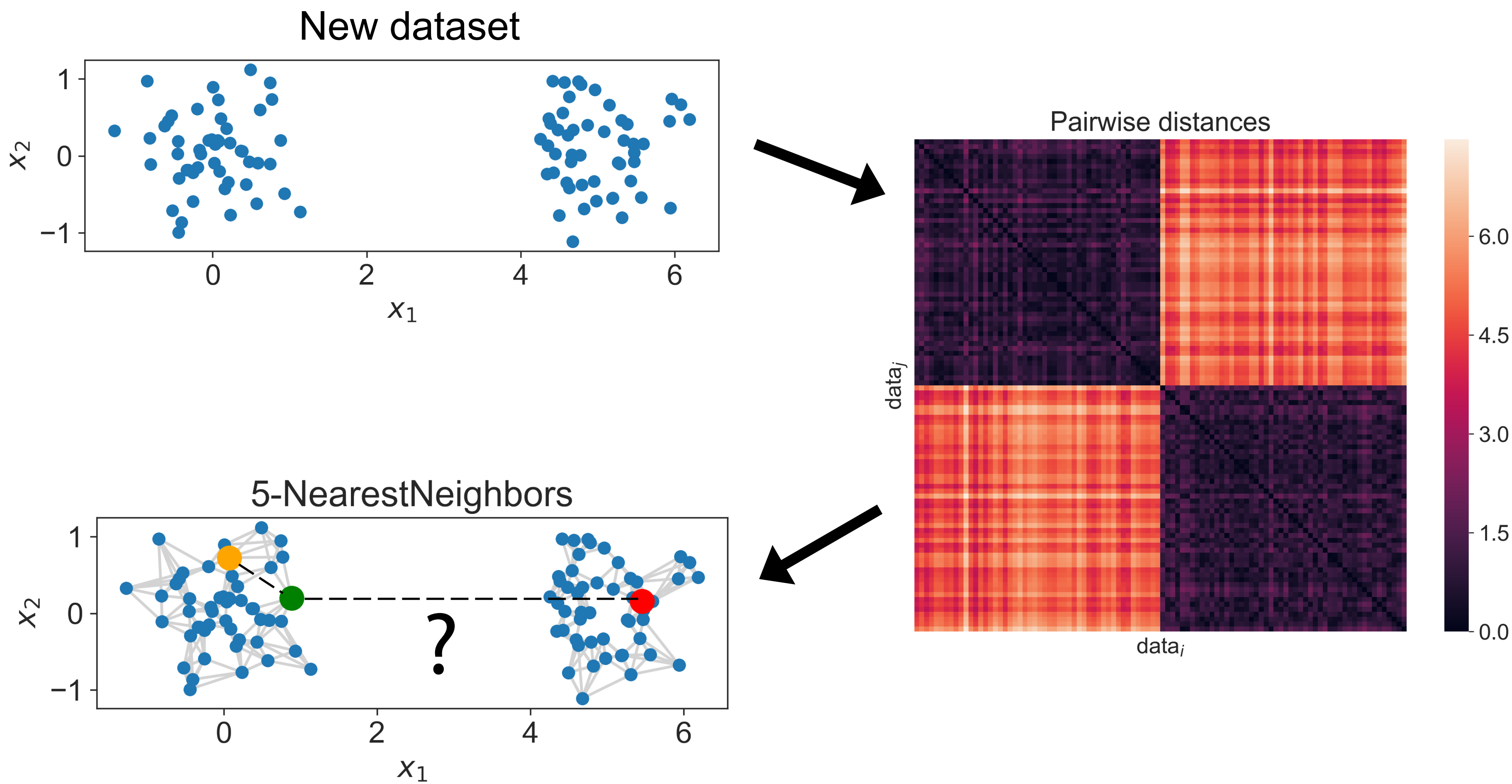
What if the data isn't entirely connected?



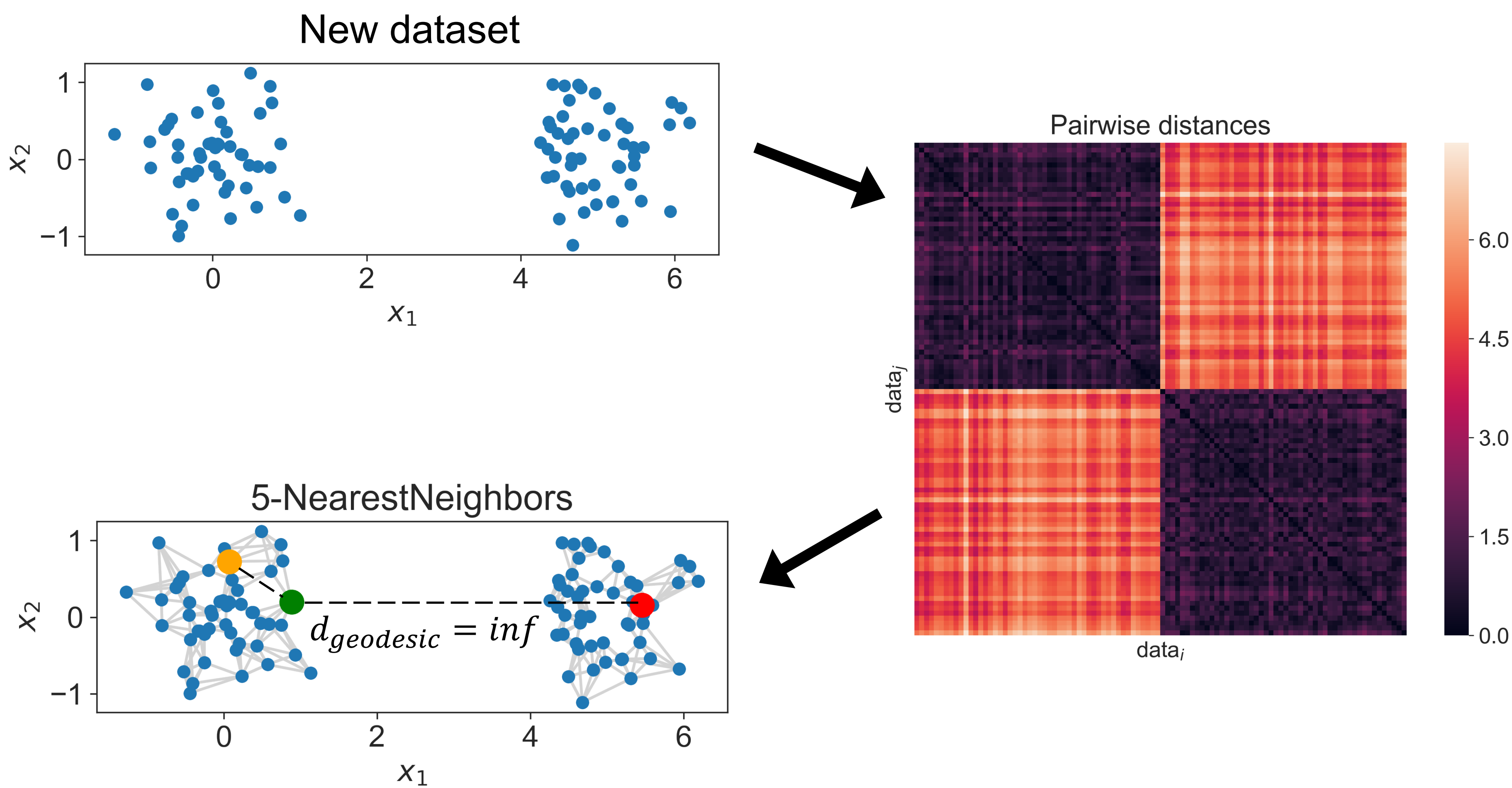
What if the data isn't entirely connected?



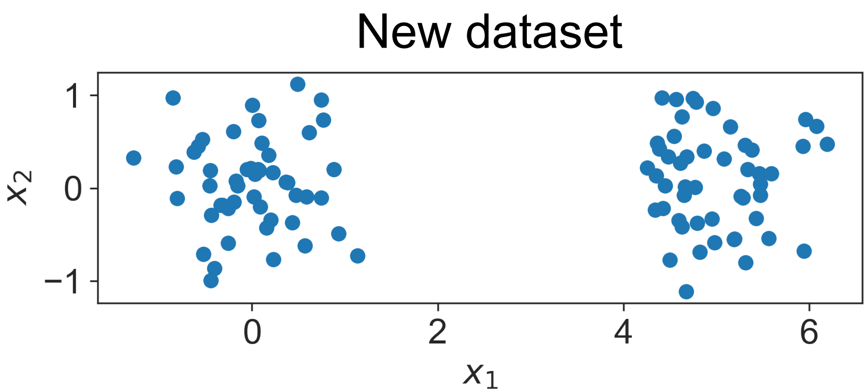
What if the data isn't entirely connected?

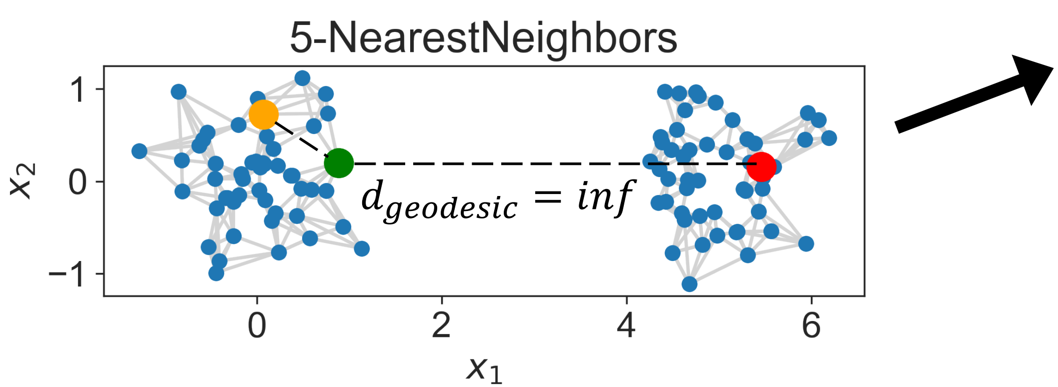


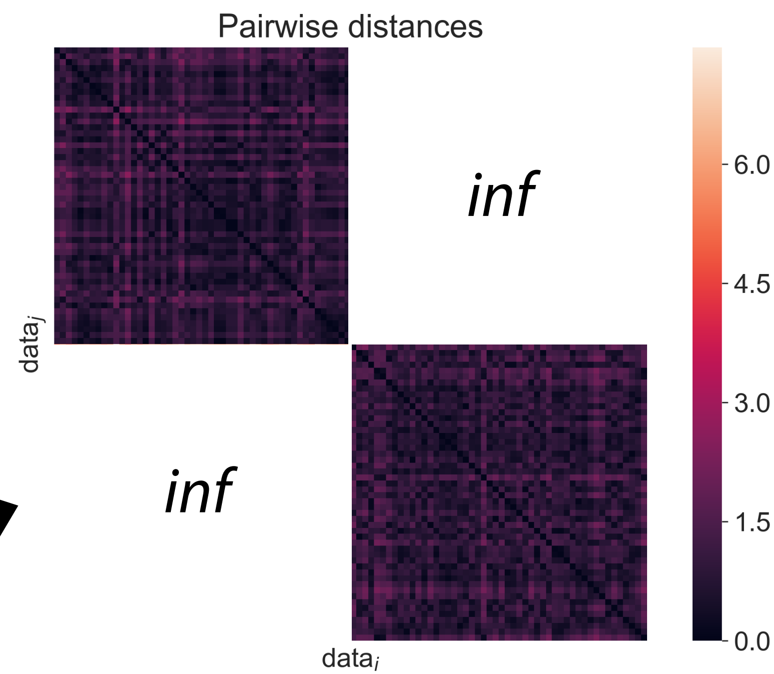
What if the data isn't entirely connected?



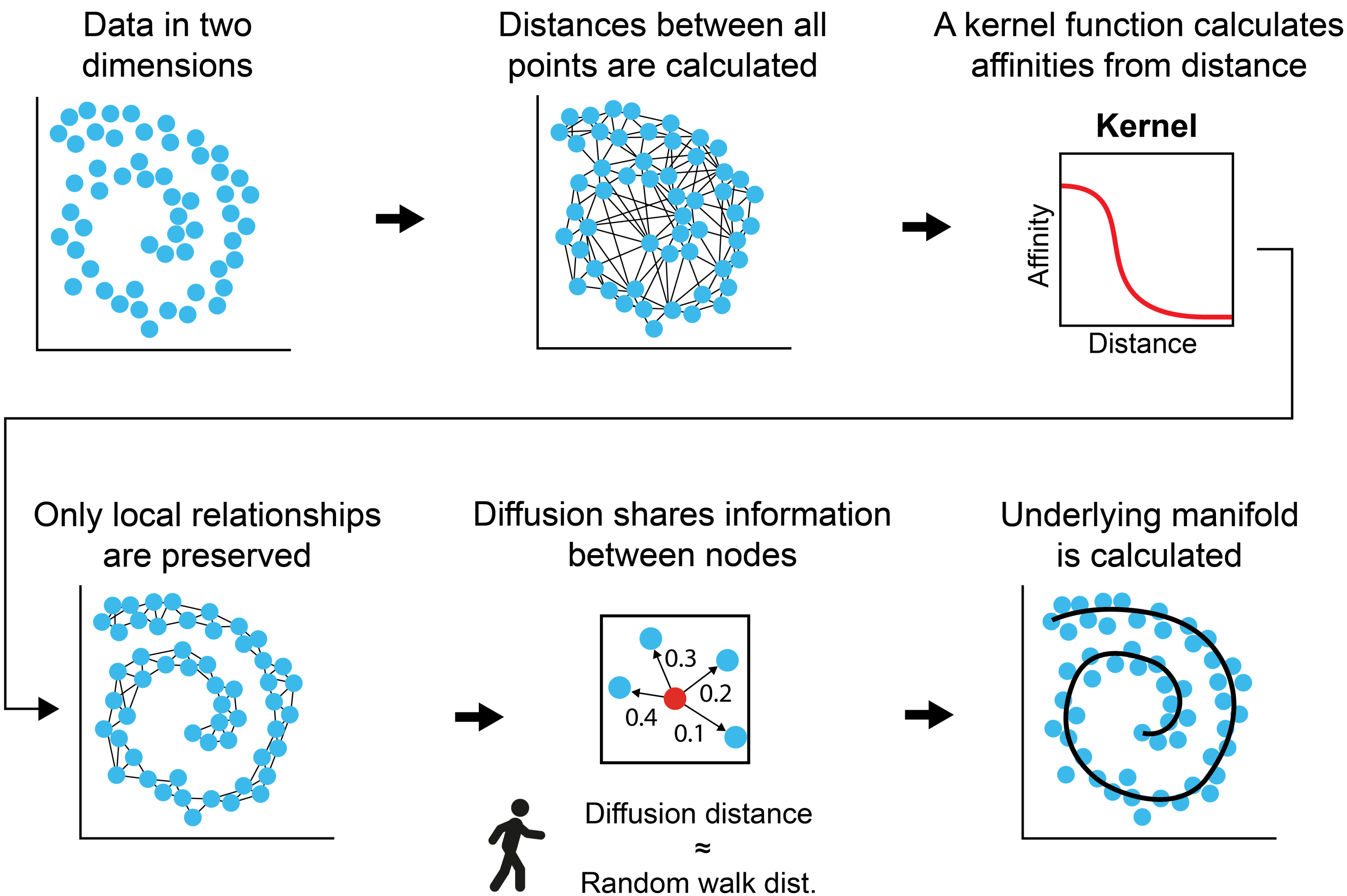
What if the data isn't entirely connected?



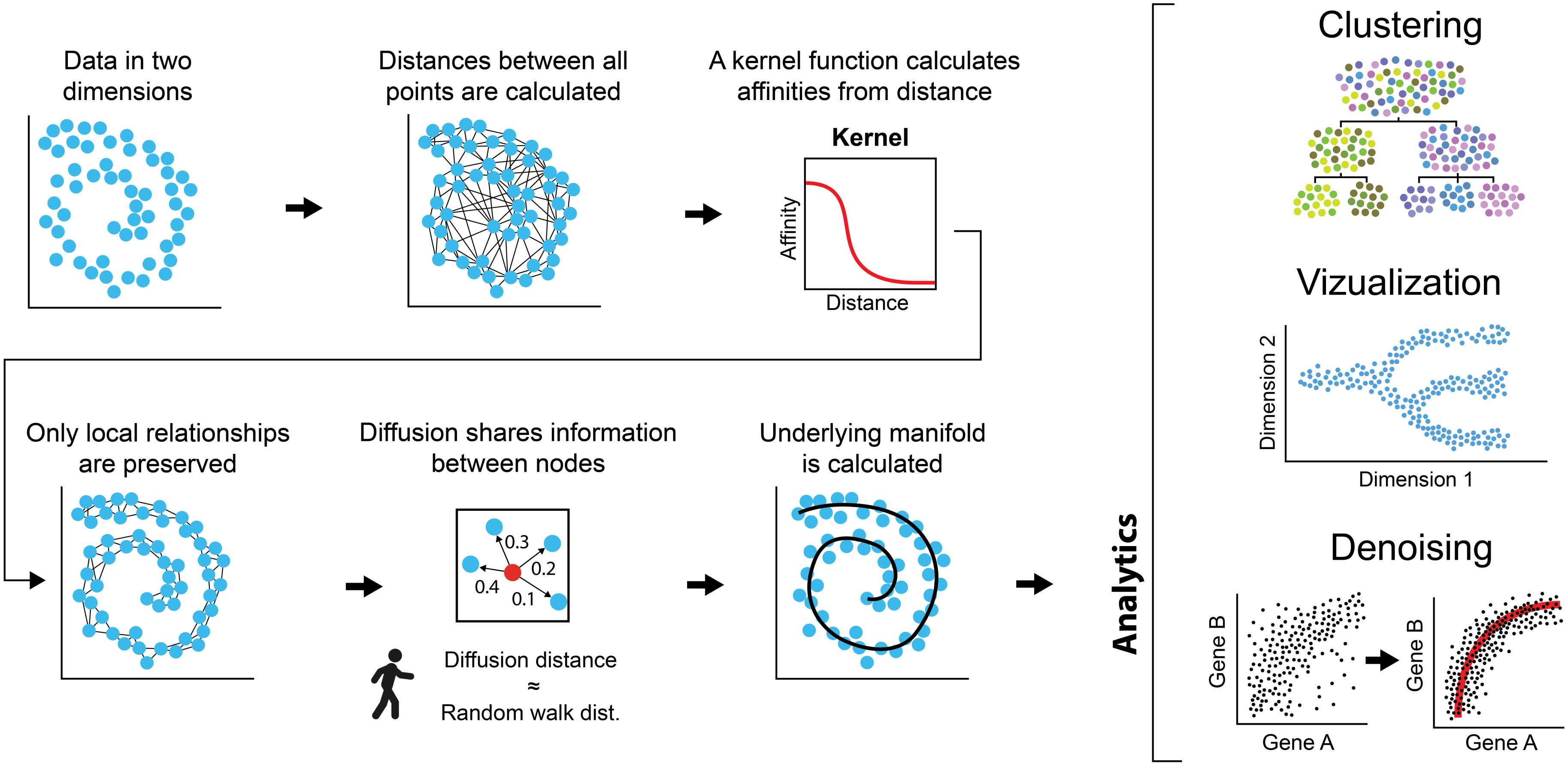




Summary: learning graphs from data

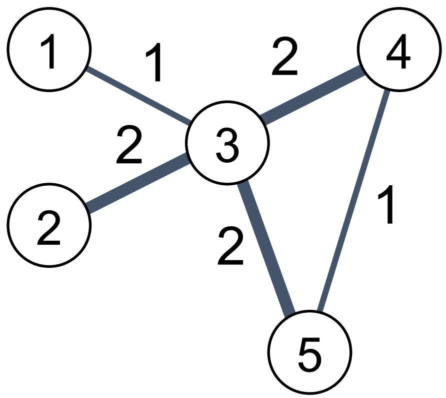


Summary: learning graphs from data



Matrix representations of graphs

G



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

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

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

Exercise – choosing k