Queen Mary
University of London

ECS766 Data Mining

Week 4: Data exploration and data visualisation

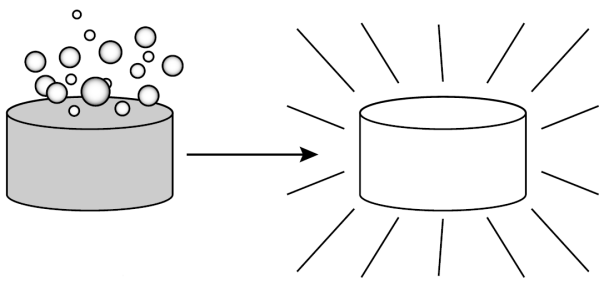
Emmanouil Benetos
emmanouil.benetos@qmul.ac.uk

October 2021

School of EECS, Queen Mary University of London

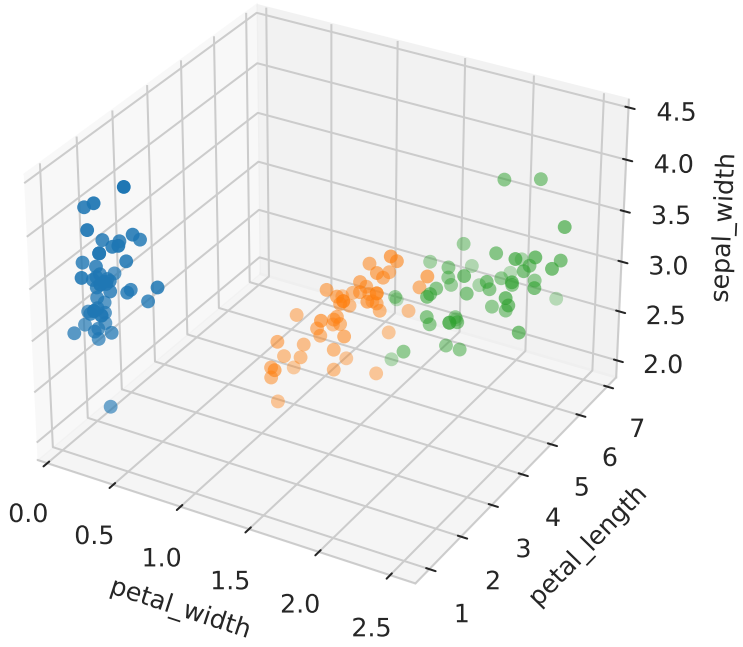
Last week: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretisation



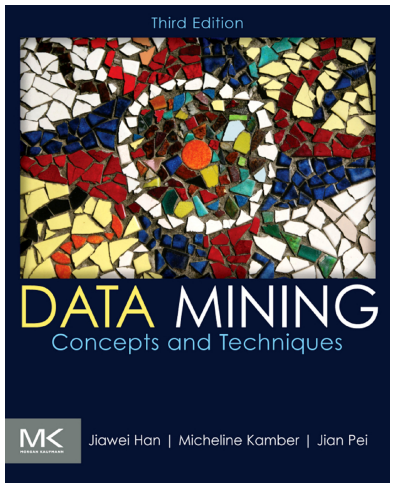
This week's contents

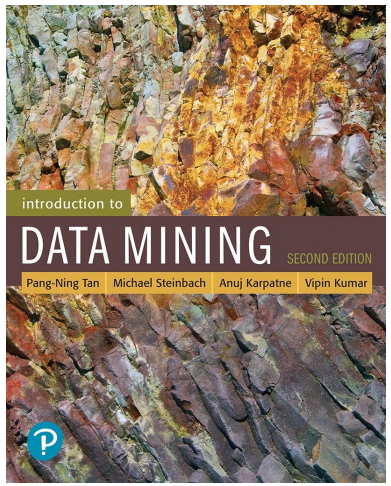
- Data Exploration
- Data Summarisation
- Data Visualisation



Reading

- Section 2.3 of J. Han, M. Kamber, J. Pei, “Data Mining: Concepts and Techniques”, 3rd edition, Elsevier/Morgan Kaufmann, 2012
- Chapter “Data exploration” of P.-N. Tan, M. Steinbach, A. Karpatne, V. Kumar, “Introduction to Data Mining”, 2nd edition, Pearson, 2019 [online]





Data exploration

Data exploration

- Data exploration refers to a preliminary investigation of the data
- This investigation typically has the following goals:
  - Revealing the need for **data pre-processing**
  - Answering simple questions about the data
  - Identifying applicable data analysis techniques
- There are two main methods for data exploration: **data summarisation** and **data visualisation**

Example: Iris flower dataset

The Iris flower dataset is used as a recurring example:

- 150 **observations** (flowers)
- 5 **features**: petal/sepal width and length, and *species*
- 3 species: I. setosa, I. versicolor, I. virginica



Data summarisation

Data summarisation

- Data summarisation represents characteristics of a potentially large number of values by a much smaller number of values
- A univariate summary characterises a single feature
- A multivariate summary characterises relationships between multiple features
- **Example:** the mean value of a feature is a common univariate summary, while the correlation between two features is a common multivariate summary

Notation: dataset

- Consider a dataset $\mathcal{D} = \mathbf{x}_1, \ldots, \mathbf{x}_N$
- The i-th **observation** $\mathbf{x}_i \in \mathbb{R}^d$ represents some **object**
- **Example**: each flower may be represented by a vector where the species is encoded by a number

1 (sepal_length)	2 (sepal_width)	3 (petal_length)	4 (petal_width)	5 (**species**)
5.1	3.5	1.4	0.2	0 (setosa)
4.9	3.0	1.4	0.2	0 (setosa)
4.7	3.2	1.3	0.2	0 (setosa)
4.6	3.1	1.5	0.2	0 (setosa)
5.0	3.6	1.4	0.2	0 (setosa)
...	...	...	...	...
6.7	3.0	5.2	2.3	2 (virginica)
6.3	2.5	5.0	1.9	2 (virginica)
6.5	3.0	5.2	2.0	2 (virginica)
6.2	3.4	5.4	2.3	2 (virginica)
5.9	3.0	5.1	1.8	2 (virginica)

Data summarisation

Summarisation techniques

Frequency

- The frequency $f_j(a)$ of value a for a **categorical feature** $j \in \{1, \ldots, d\}$ is given by

$$f_j(a) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[x_{i,j} = a],$$

  where $\mathbb{I}[e] = 1$ if e is true, and $\mathbb{I}[e] = 0$ otherwise

- **Example**: the frequency of value 0 (setosa) for feature 5 (species) is $f_5(0) = 50/150 = 1/3$

Mode

- The value a^* is a mode of the **categorical feature** j if it is among the most frequent values of that feature:

$$f_j(a^*) = \max_a f_j(a)$$

- A categorical feature may have **multiple modes**
- **Example**: feature 5 (species) has three modes, since there are exactly 50 observations of each species.

Mean

- The mean μ_j of feature j is given by

$$\mu_j = \frac{1}{N} \sum_{i=1}^{N} x_{i,j}$$

- The mean is a maximum likelihood estimate of parameters of many common models.
- However, the mean is highly influenced by **outliers**.

Median

- A feature is ordinal if its possible values can be **meaningfully** ordered
- Let j be an **ordinal feature**, and let S_j denote the sequence that results from ordering $x_{1,j}, \ldots, x_{N,j}$
- If N is odd, the median of feature j is the value in the middle of S_j
- If N is even, the median of feature j is the **mean** of the two values in the middle of S_j
- The median is robust to outliers, and often represents a *typical* value better than the mean

Quartile

- Let j be an **ordinal feature**, and let S_j denote the sequence that results from ordering $x_{1,j}, \ldots, x_{N,j}$
- There are three quartiles for feature j:
  - The median of feature j
  - If N is even, the lower/upper quartile is the median of the left/right half of the elements in S_j
  - If N is odd, the lower/upper quartile is the median of the elements to the left/right of the element in the middle of S_j

Percentiles

- The *p*-th percentile for an ordinal feature *j* is the smallest **occurring** value a_p such that at least *p*% of the values for feature *j* are less or equal to a_p
- **Example**: the 75-th percentile for petal length is 5.1, so **at least** 75% of the flowers have petals shorter or equal to 5.1

Range

- The **maximum** and **minimum** observed values of a feature are often worth inspecting
- The range of a feature j is the difference between these two values

$$\left(\max_i x_{i,j}\right) - \left(\min_i x_{i,j}\right)$$

- The range measures the *spread* of a feature
- However, it is highly influenced by **outliers**

Variance

- The (biased) variance σ_j^2 of feature j is given by

$$\sigma_j^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{i,j} - \mu_j)^2,$$

  where μ_j is the mean of feature j
- The standard deviation σ_j is the square root of the variance σ_j^2 of feature j
- The variance and the standard deviation also measure the *spread* of a feature
- They are both highly influenced by **outliers**

Covariance matrix

- The (biased) covariance $\sigma^2_{j,k}$ between features j and k is given by

$$\sigma^2_{j,k} = \frac{1}{N}\sum_{i=1}^{N}(x_{i,j} - \mu_j)(x_{i,k} - \mu_k)$$

  where μ_j is the mean of feature j, and μ_k is the mean of feature k
- The covariance of two attributes is a measure of the degree to which two attributes vary together.
- A $d \times d$ (symmetric) covariance matrix Σ contains the covariance of every pair of features

Correlation matrix

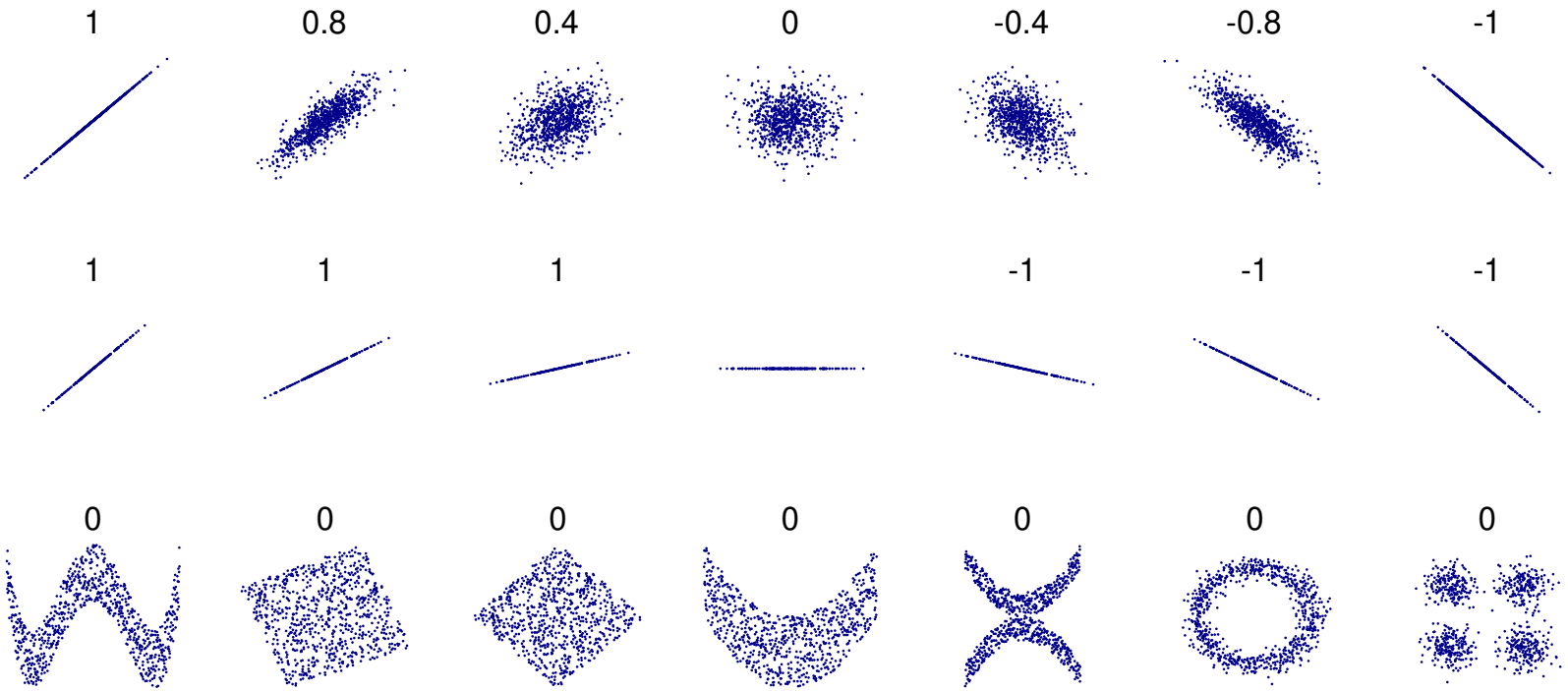
- The (biased) correlation $c_{j,k}$ between features j and k is given by

$$c_{j,k} = \frac{\sigma^2_{j,k}}{\sigma_j \sigma_k}$$

- The correlation is always between -1 and 1
- The correlation between two features attempts to measure the extent to which they are **linearly** related
- A $d \times d$ (symmetric) correlation matrix contains the correlation of every pair of features

Correlation: non-linear relationships

- **Example: scatter plots** for different two-dimensional datasets and the correlations between their features



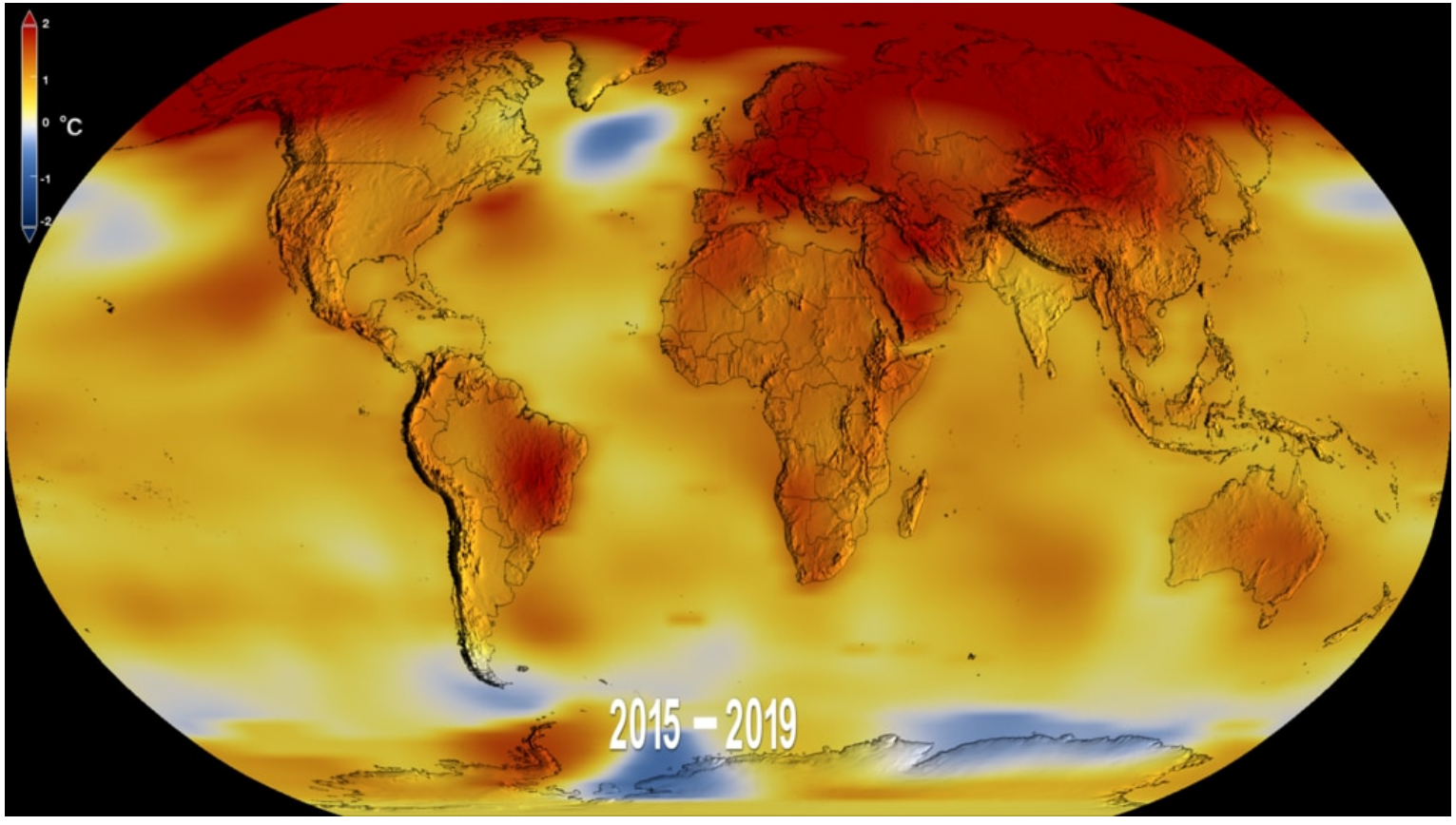
Data visualisation

Data visualisation

- Data visualisation refers both to the creation and the interpretation of visual data representations
- **Motivation:** humans are typically able to quickly acquire large amounts of visual data
- Data visualisation may aid both **hypothesis creation** and **hypothesis validation**

Data visualisation: quick data acquisition

- **Example:** imagine a table that records the value for each coordinate used to create the image below



[1]

[1] Image from NASA

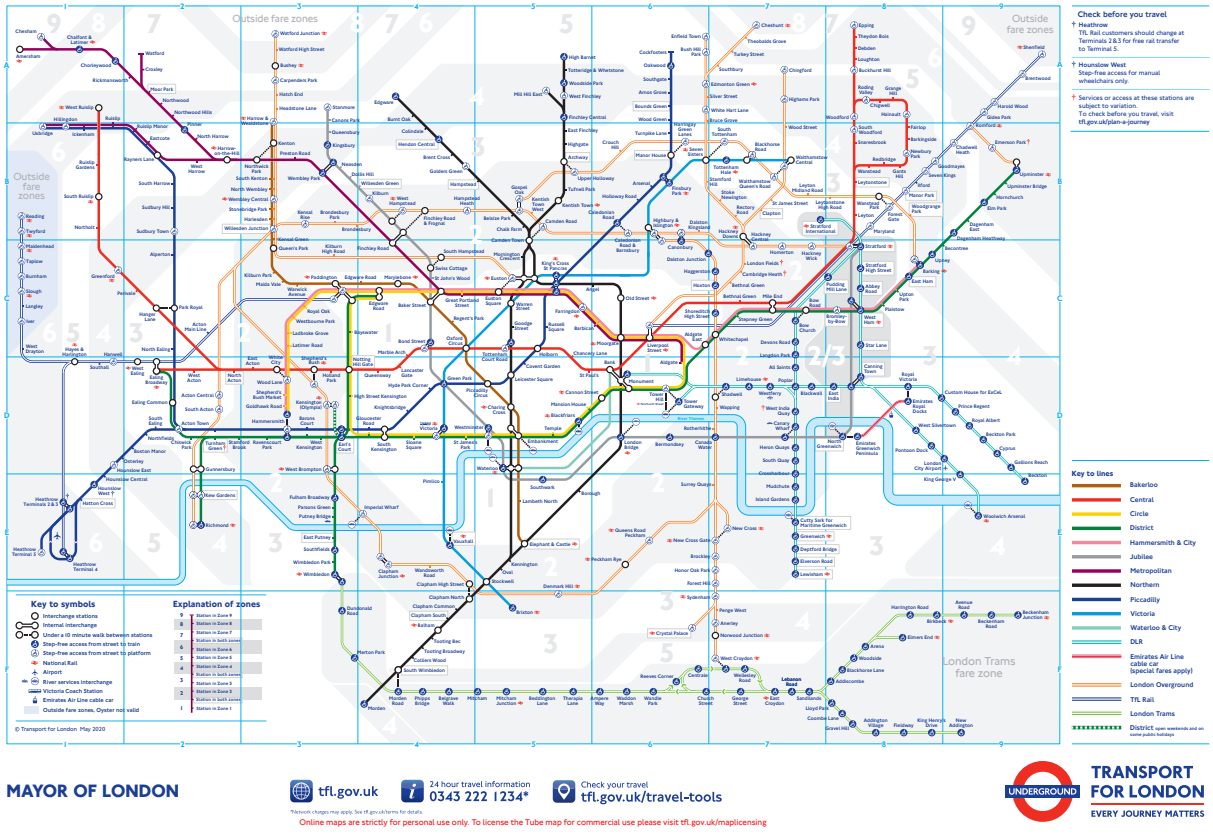
Data visualisation: specificity

- Visualisation techniques are often adapted or developed specifically for a dataset
- Different tasks may also require different visualisations of the same dataset
- **Domain expertise** is often required to guide the (iterative) development of visualisations

Data visualisation: specificity

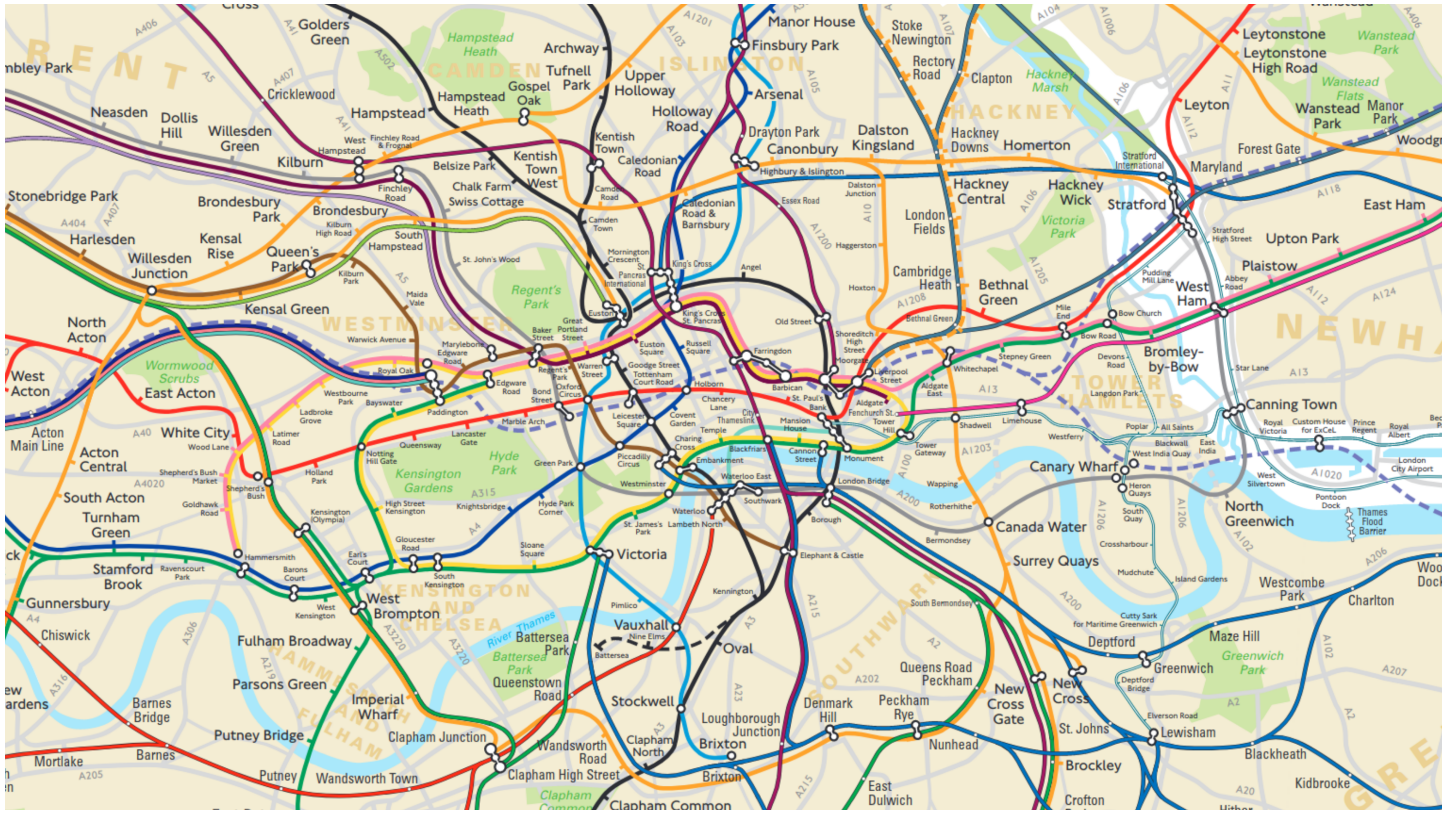
- **Example: connections** are more important than **distances** for a rapid transit system map

Tube map



Data visualisation: specificity

- London tube map with physical distances:



Data visualisation

Visualisation techniques

Table visualisation

- Most data can be represented by a **table**
- The traditional visual representation is not scalable, but some simple **operations** often make it useful

id	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
…	…	…	…	…	…
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

Table visualisation: sorting

- Sorting a table may suggest simple **correlations**

id	sepal_length	sepal_width	petal_length	**petal_width (↑)**	species
32	5.2	4.1	1.5	0.1	setosa
13	4.3	3.0	1.1	0.1	setosa
37	4.9	3.6	1.4	0.1	setosa
9	4.9	3.1	1.5	0.1	setosa
12	4.8	3.0	1.4	0.1	setosa
...	...	...	...	...	...
140	6.7	3.1	5.6	2.4	virginica
114	5.8	2.8	5.1	2.4	virginica
100	6.3	3.3	6.0	2.5	virginica
144	6.7	3.3	5.7	2.5	virginica
109	7.2	3.6	6.1	2.5	virginica

Table visualisation: slicing

- Slicing a range of rows is often combined with sorting
- **Example**: selecting the ten observations with the longest petals

id	sepal_length	sepal_width	**petal_length (↓)**	petal_width	species
118	7.7	2.6	6.9	2.3	virginica
122	7.7	2.8	6.7	2.0	virginica
117	7.7	3.8	6.7	2.2	virginica
105	7.6	3.0	6.6	2.1	virginica
131	7.9	3.8	6.4	2.0	virginica
107	7.3	2.9	6.3	1.8	virginica
130	7.4	2.8	6.1	1.9	virginica
109	7.2	3.6	6.1	2.5	virginica
135	7.7	3.0	6.1	2.3	virginica
100	6.3	3.3	6.0	2.5	virginica

Table visualisation: conditional indexing

- Conditional indexing is useful to test simple **hypotheses**
- **Example**: selecting I. versicolor observations where the sepal length is less than twice the sepal width

id	**sepal_length**	**sepal_width**	petal_length	petal_width	species
56	6.3	3.3	4.7	1.6	versicolor
59	5.2	2.7	3.9	1.4	versicolor
61	5.9	3.0	4.2	1.5	versicolor
64	5.6	2.9	3.6	1.3	versicolor
66	5.6	3.0	4.5	1.5	versicolor
70	5.9	3.2	4.8	1.8	versicolor
84	5.4	3.0	4.5	1.5	versicolor
85	6.0	3.4	4.5	1.6	versicolor
88	5.6	3.0	4.1	1.3	versicolor
95	5.7	3.0	4.2	1.2	versicolor
96	5.7	2.9	4.2	1.3	versicolor

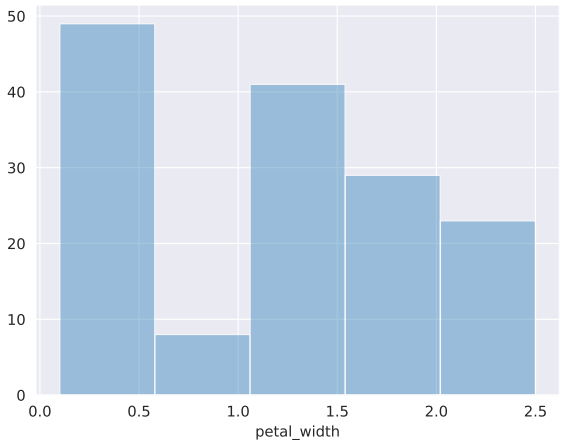
Table visualisation: grouping

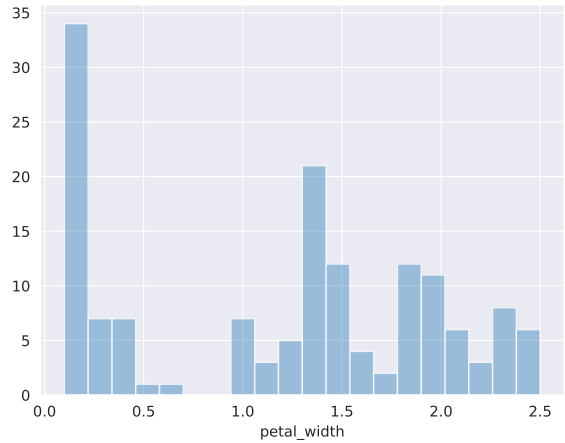
- Grouping involves partitioning the observations into groups based on the values of chosen **categorical features** and operating independently on each group
- **Example**: grouping observations based on species, and computing the mean of each other feature

species	sepal_length	sepal_width	petal_length	petal_width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

Histograms

- A histogram divides the range of a feature into a number of **bins** (disjoint intervals)
- The height of a rectangle represents the number of observations that fall into each bin
- **Purpose:** representing frequency of values
- **Note:** the number of bins may highly affect conclusions





Pie charts

- A pie chart divides a circle into sectors according to the number of **categories** of a feature
- The area of each sector is also proportional to the frequency of each category
- **Purpose:** representing frequency of values
- **Note:** it is difficult to quickly measure sector areas

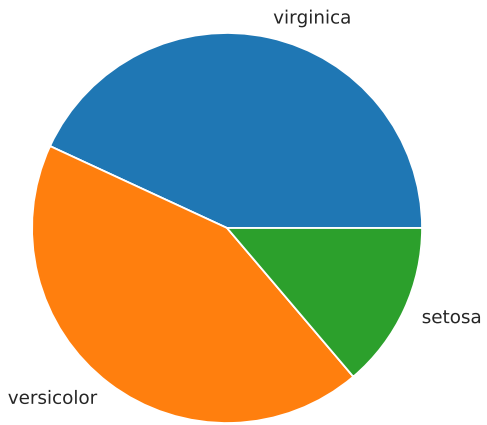
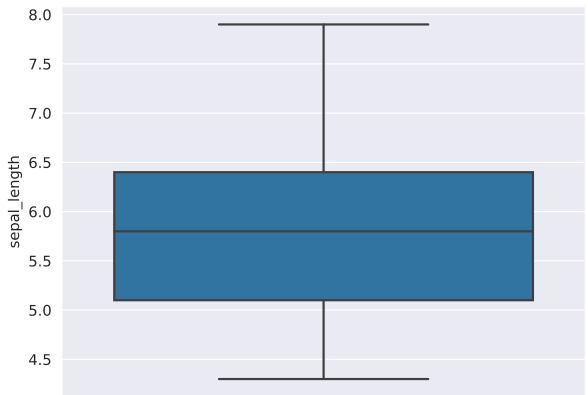
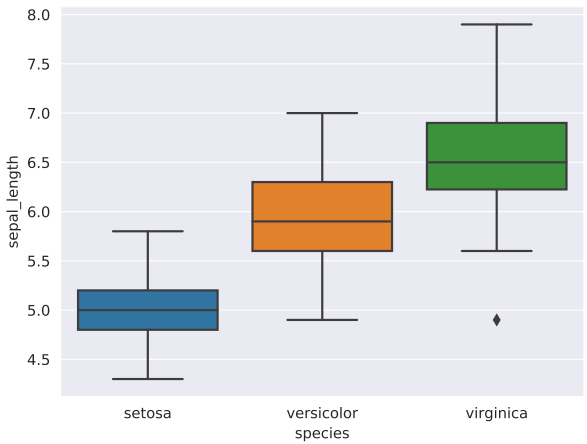


Figure 1: Species of flowers that exceed petal width 0.2

Box plots

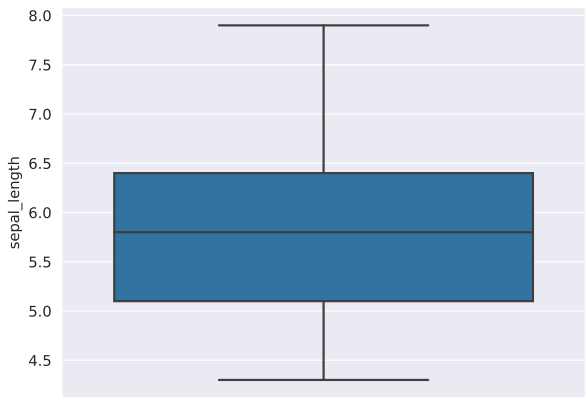
- A box plot represents the lower and upper quartiles of a feature by a rectangle divided across the median
- Outlier observations are represented by points
- The minimum and maximum values **excluding outliers** are represented by parallel line segments connected by a perpendicular line segment
- **Purpose:** displaying outlier-robust statistics

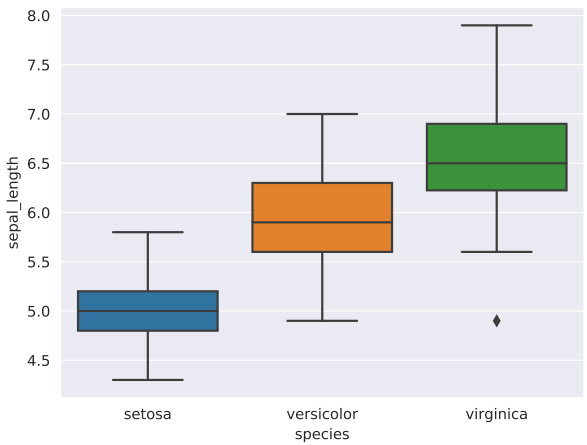




Box plots

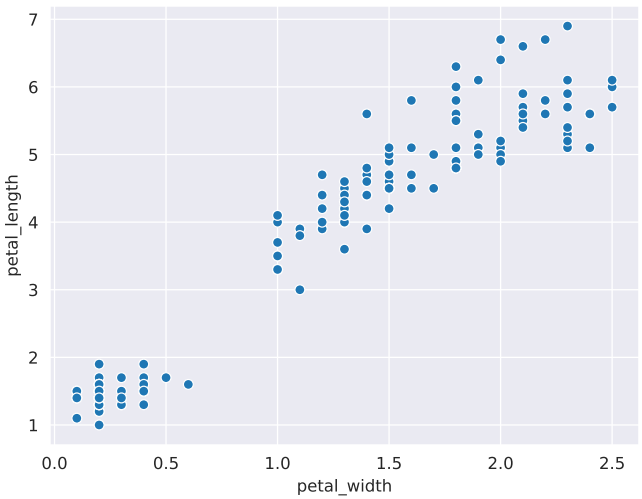
- The interquartile range is the difference between the upper quartile and the lower quartile
- An observation may be considered an **outlier** if it is above/below the upper/lower quartile plus/minus a constant (such as 1.5) times the interquartile range
- **Note: grouping** may reveal relationships between a categorical feature and other features

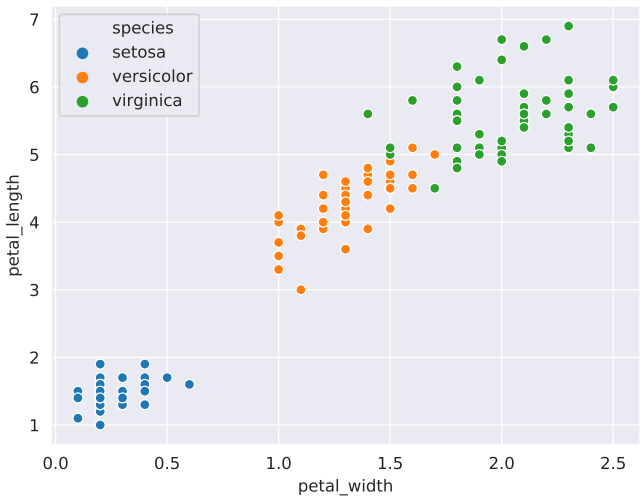




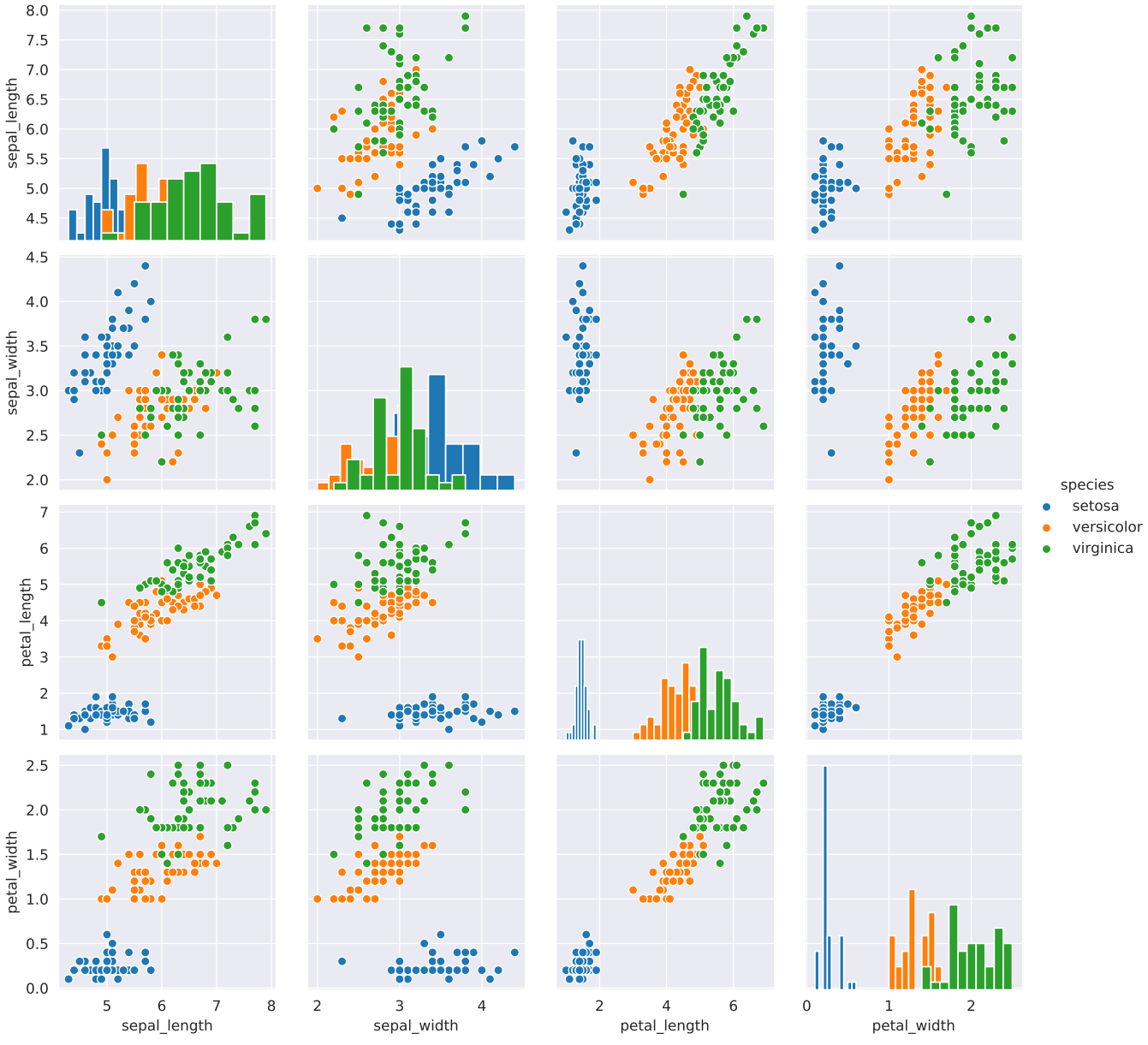
Scatter plots

- For each observation in a dataset, a scatter plot displays a point positioned according to the values of a fixed pair of features
- **Purpose:** revealing pairwise feature relationships
- **Note:** other features can be encoded into each point by using different shapes, colours, and sizes



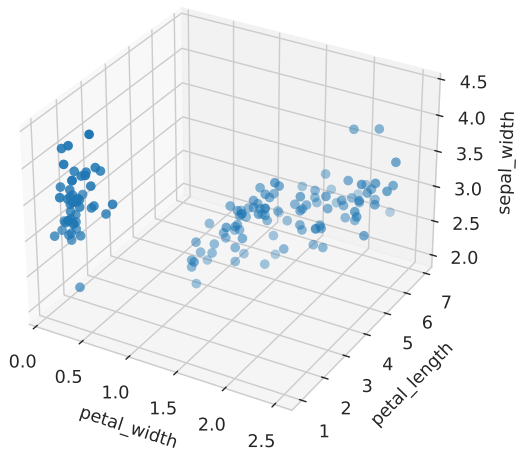


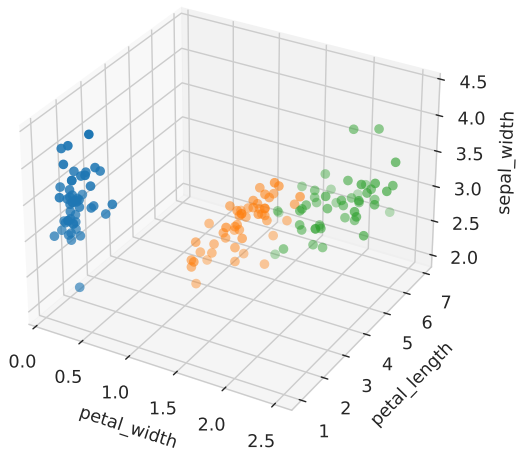
Scatter plot matrices



Three-dimensional scatter plots

- A three-dimensional scatter plot can be used to represent relationships between three features
- Three-dimensional plots are typically interactive
- **Note:** viewpoints and occlusions may highly affect conclusions





Parallel coordinates

- In parallel coordinates, each feature has a vertical line segment that represents its range
- Each observation $\mathbf{x}_i$ has a **polyline** that intersects the line segment for feature j where $x_{i,j}$ would lie
- **Purpose:** revealing groups of observations, outliers, and relationships between *adjacent* features
- **Note:** the order of the line segments (features) may highly affect conclusions

Distance matrices

- In an $N \times N$ distance matrix D, the distance between observation $\mathbf{x}_i$ and observation $\mathbf{x}_j$ is given by $D_{i,j}$
- A distance matrix D can be represented by an $N \times N$ table where cell (i, j) is coloured according to $D_{i,j}$
- This visualisation is often only useful if the observations are **sorted** by some criterion

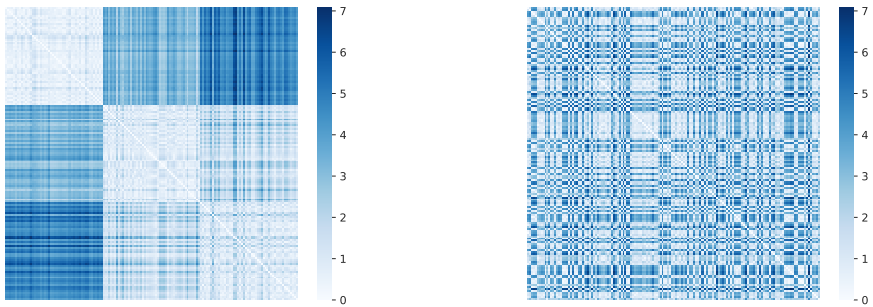


Figure 2: Distance matrices for the Iris flower dataset excluding species, sorted (left) and unsorted (right) by species

Dimensionality reduction

- For the purposes of visualisation, dimensionality reduction attempts to represent a dataset $\mathcal{D} = \mathbf{x}_1, \ldots, \mathbf{x}_N$, where $\mathbf{x}_i \in \mathbb{R}^d$, by a projection $\mathcal{P} = \mathbf{p}_1, \ldots, \mathbf{p}_N$, where $\mathbf{p}_i \in \mathbb{R}^c$, $c < d$, and each point $\mathbf{p}_i$ corresponds to an observation $\mathbf{x}_i$
- A projection attempts to preserve the **structure** of the data, which is defined by the relationships between observations, presence of clusters and outliers, or overall spatial distribution
- Typically, $c = 2$ or $c = 3$ such that the resulting projection may be represented by a scatter plot

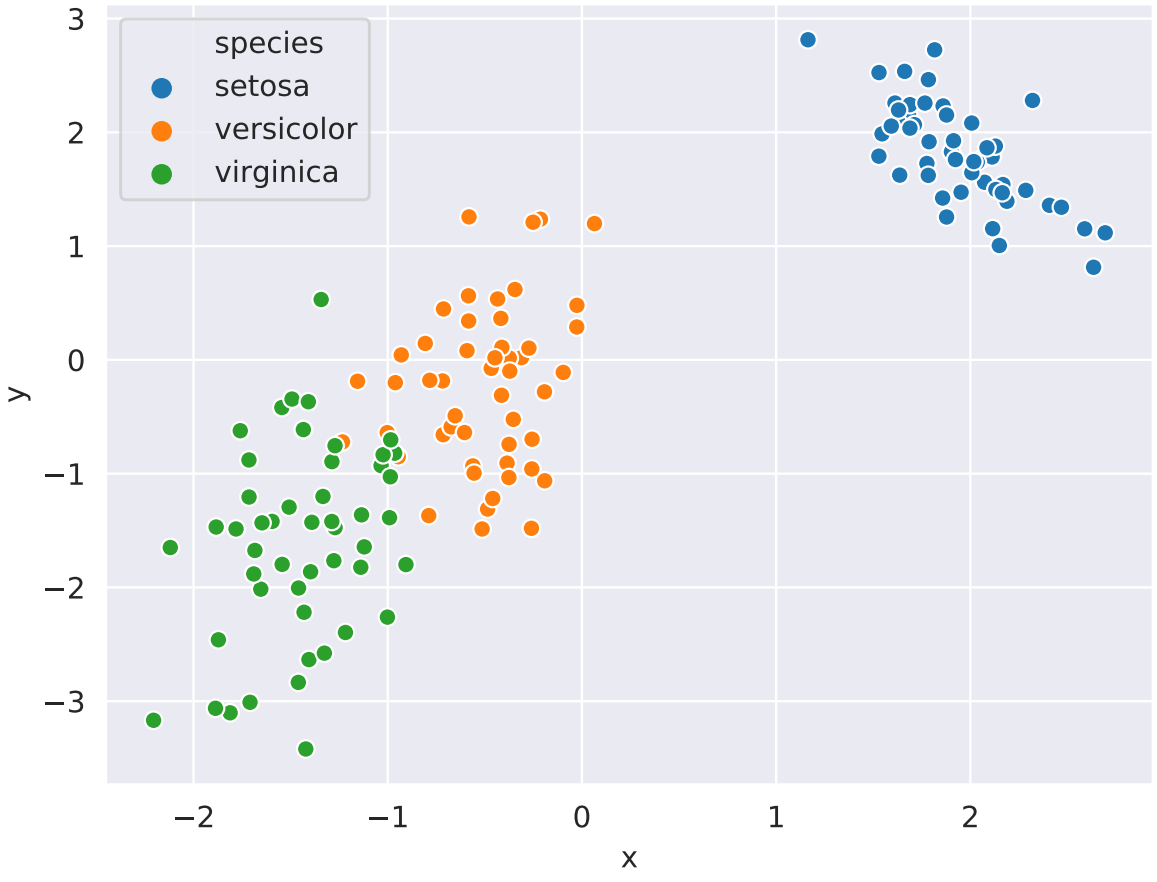
Multidimensional scaling

- The goal of absolute (metric) multidimensional scaling is to compute a projection $\mathcal{P} = \mathbf{p}_1, \ldots, \mathbf{p}_N$ where the **distances** in $\mathcal{D} = \mathbf{x}_1, \ldots, \mathbf{x}_N$ are preserved
- This goal may be achieved by minimizing the raw stress cost C with respect to the projection $\mathcal{P}$, where

$$C = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (d_{i,j} - r_{i,j})^2,$$

  $d_{i,j} = ||\mathbf{x}_i - \mathbf{x}_j||$, and $r_{i,j} = ||\mathbf{p}_i - \mathbf{p}_j||$
- There are efficient algorithms to solve this problem

Multidimensional scaling

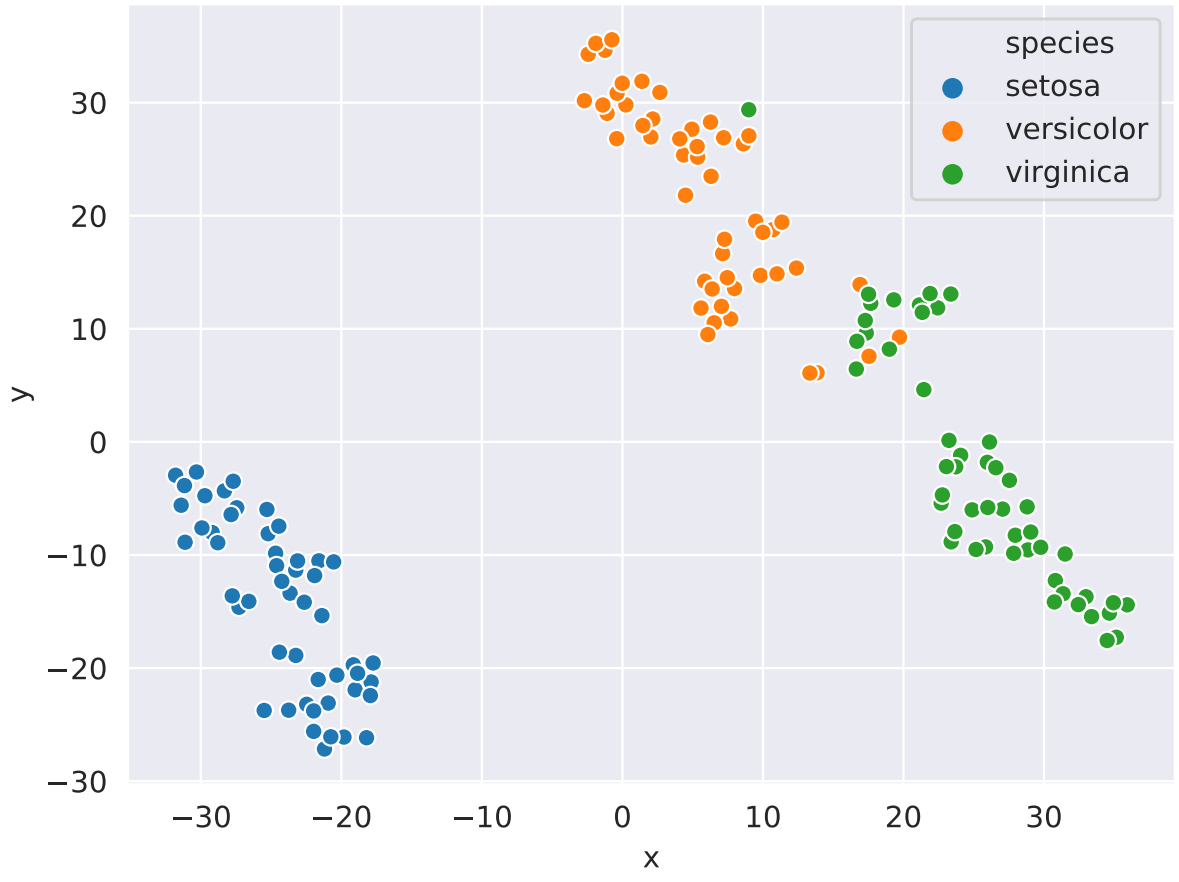


- **Note:** the axes are not easily interpretable, only placement relative to other points matters

t-distributed stochastic neighbor embedding (t-SNE)

- The goal of t-distributed stochastic neighbor embedding (t-SNE) is to compute a projection $\mathcal{P} = \mathbf{p}_1, \ldots, \mathbf{p}_N$ where the **neighbourhoods** in $\mathcal{D} = \mathbf{x}_1, \ldots, \mathbf{x}_N$ are preserved
- Uses the Kullback-Leibler divergence to compute the distribution between pairs of objects.
- Preserving neighbourhoods instead of distances may be preferable in high-dimensional spaces.

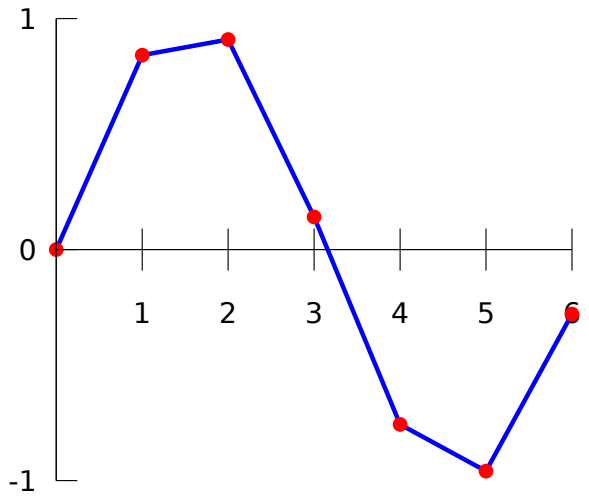
t-distributed stochastic neighbor embedding (t-SNE)



- **Note:** the axes are not easily interpretable, only placement relative to other points matters

Scalar fields

- A (real) scalar field is a function $f : \mathbb{R}^d \to \mathbb{R}$
- A dataset $\mathcal{D} = \mathbf{x}_1, \ldots, \mathbf{x}_N$ may contain **samples** from a scalar field such that $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d}, f(\mathbf{x}_{i,1:d}))$
- Interpolation may be used to approximate a scalar field f by a scalar field $\tilde{f}$ given a dataset $\mathcal{D}$

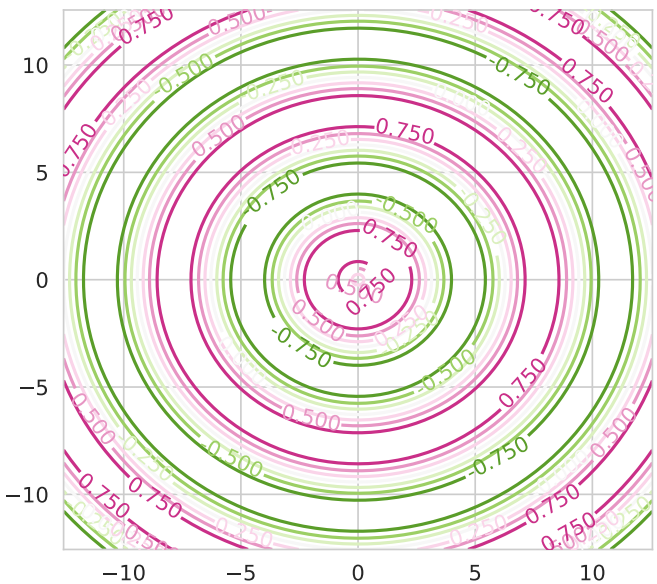


Heat maps

- A heat map uses a colour to represent the value $f(x, y)$ of each point (x, y) in a scalar field $f : \mathbb{R}^2 \to \mathbb{R}$

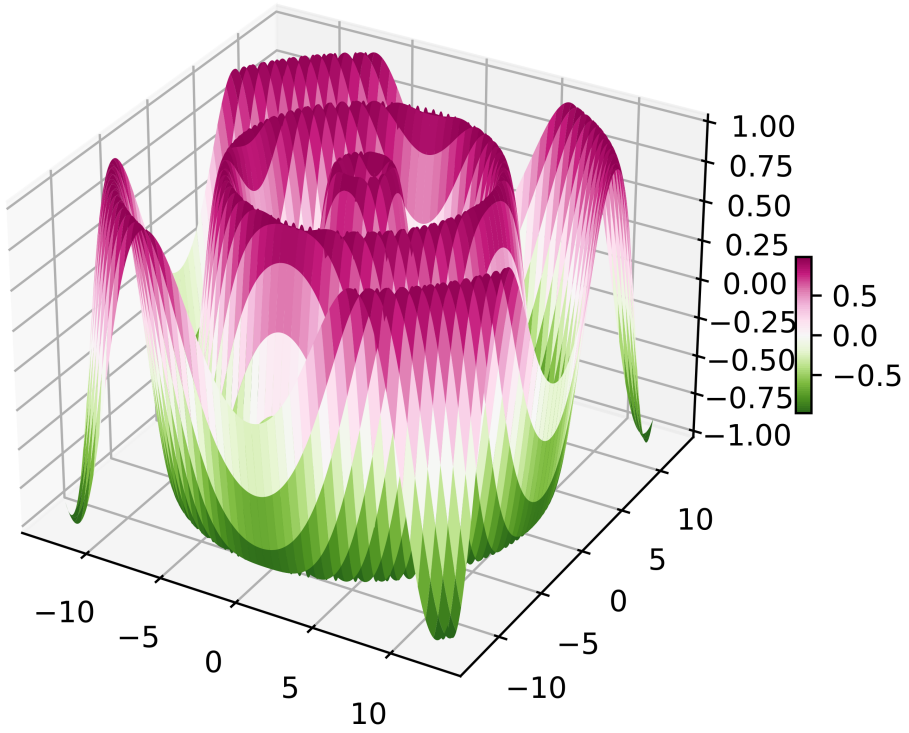
Contour plots

- A contour line $C(y)$ for the value y of a scalar field $f : \mathbb{R}^d \to \mathbb{R}$ is the set $C(y) = \{\mathbf{x} \in \mathbb{R}^d \mid f(\mathbf{x}) = y\}$
- A contour plot displays several contour lines of a scalar field $f : \mathbb{R}^2 \to \mathbb{R}$
- A contour plot contains less information than a heat map, which sometimes is an advantage



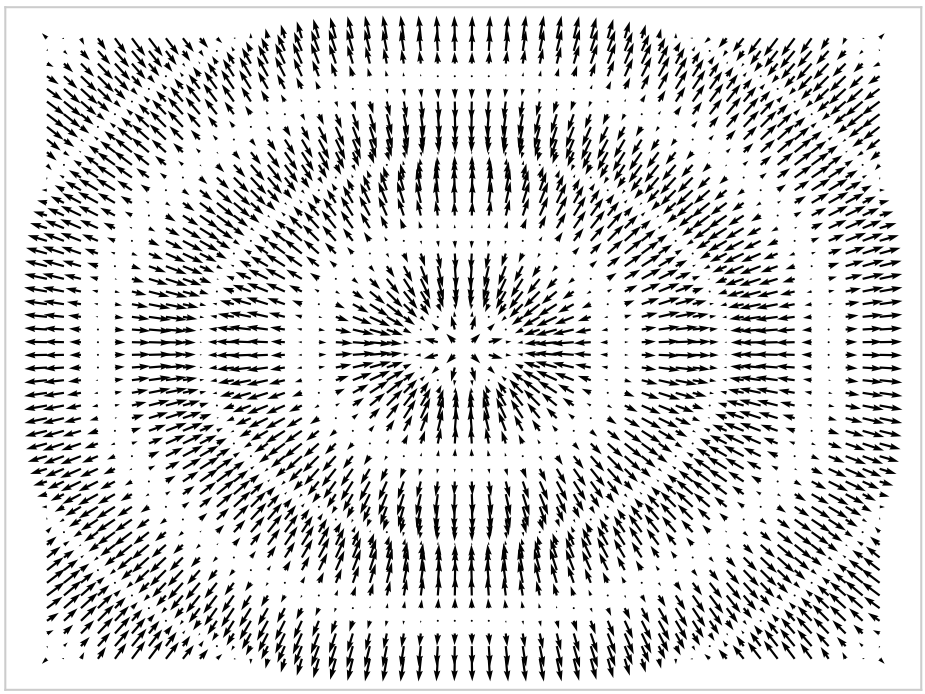
Surface plots

- A surface plot displays a **two-dimensional** scalar field $f : \mathbb{R}^2 \to \mathbb{R}$ by representing each point $(x, y) \in \mathbb{R}^2$ in the field by a point $(x, y, f(x, y)) \in \mathbb{R}^3$



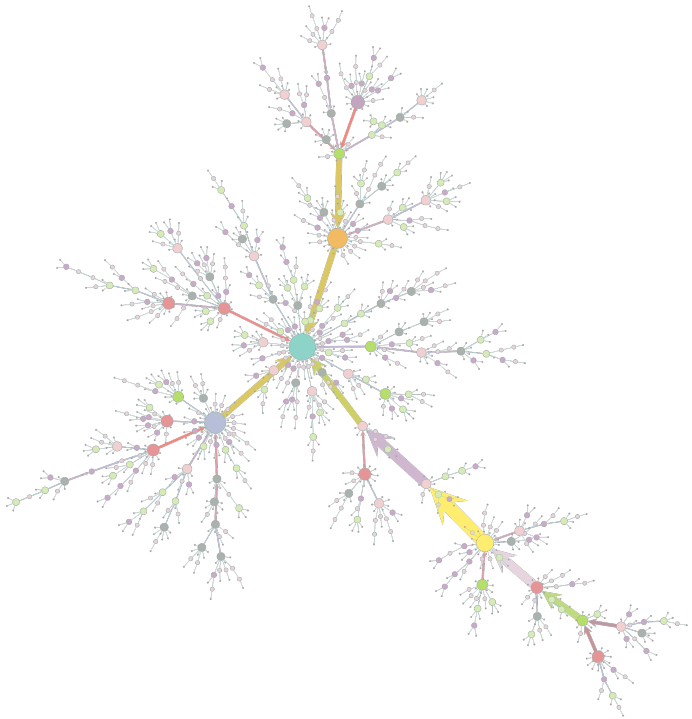
Vector field plots

- In some data, a characteristic may have both a magnitude and a direction associated with it (e.g. flow of a substance).
- Vector field plots (or quiver plots) display both direction and magnitude.
- **Example:** the gradient of the previous scalar fields.



Graph visualisations

- Graph drawing is concerned with obtaining visual representations of graphs based on nodes and links
- There are many different graph layout algorithms that attempt to maximise different **quality measures**



Data visualisation

Visualisation principles

Representation

- Representation is the process of encoding information into **visual objects** (such as points, lines, or polygons), **visual attributes** (such as size, orientation, colour, or texture), and **visual relationships** (such as relative positions).
- Representation often requires incorporating domain knowledge, adopting existing conventions, and minimizing the potential for misinterpretation.
- **Example:** many domains have conventions for colour usage, and the similarity between colours should match the similarity between what they represent.

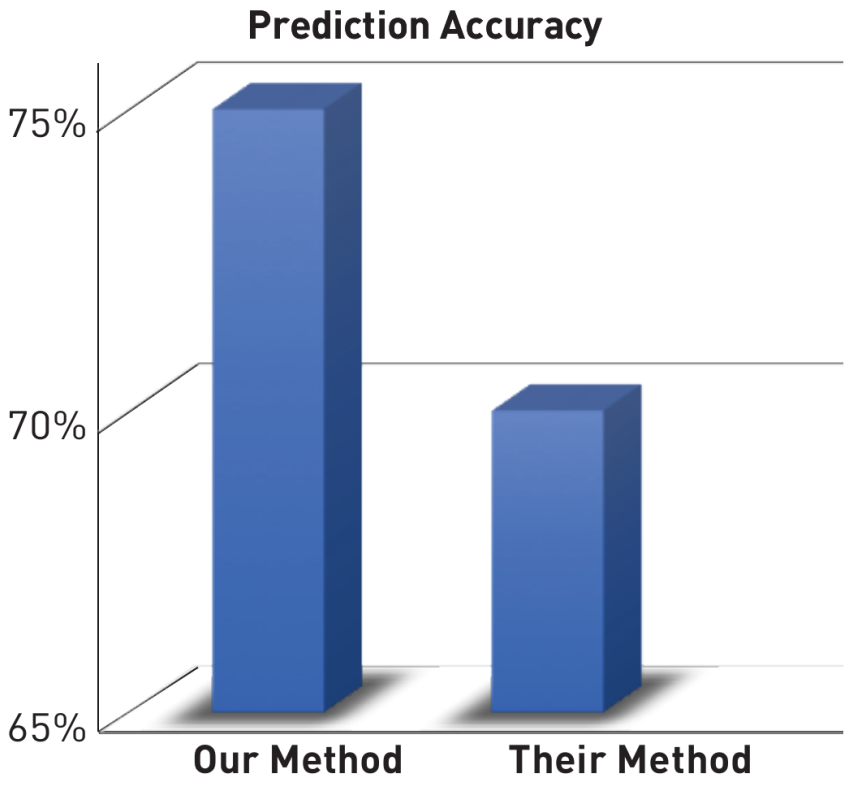
Arrangement

- Arrangement is the process of placing visual objects
- Many visualisation techniques are highly dependent on appropriate arrangement:
  - Table visualisations
  - Distance matrices
  - Parallel coordinates
  - Graph visualisations

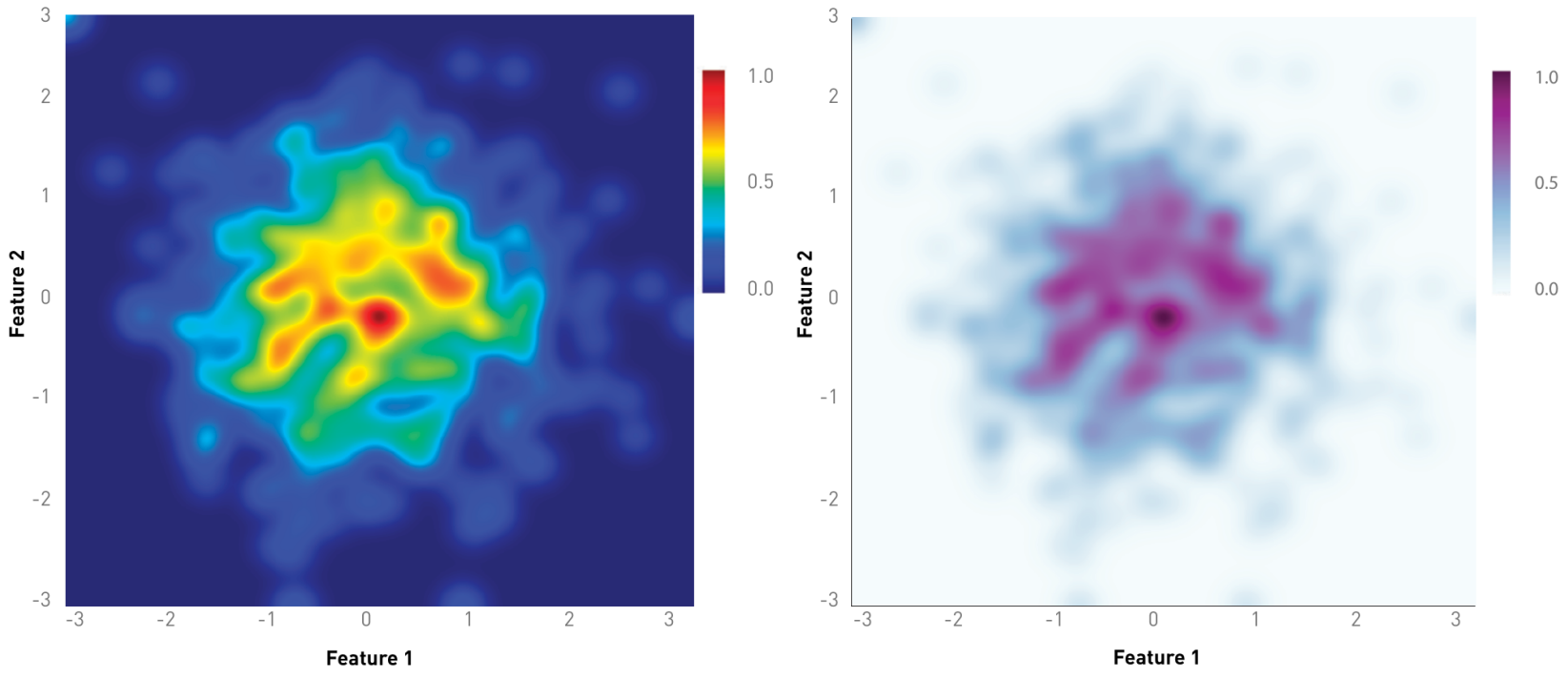
Selection

- Selection is the process of deciding which information should be encoded in a visualisation
- It may be impossible to convey all useful information in a single visualisation
- Visual objects, attributes, and relationships should emphasize the important aspects of the data
- **Animation** may be useful to allow the quick detection of visual changes
- **Interactivity** may be useful to enable control over the level of detail

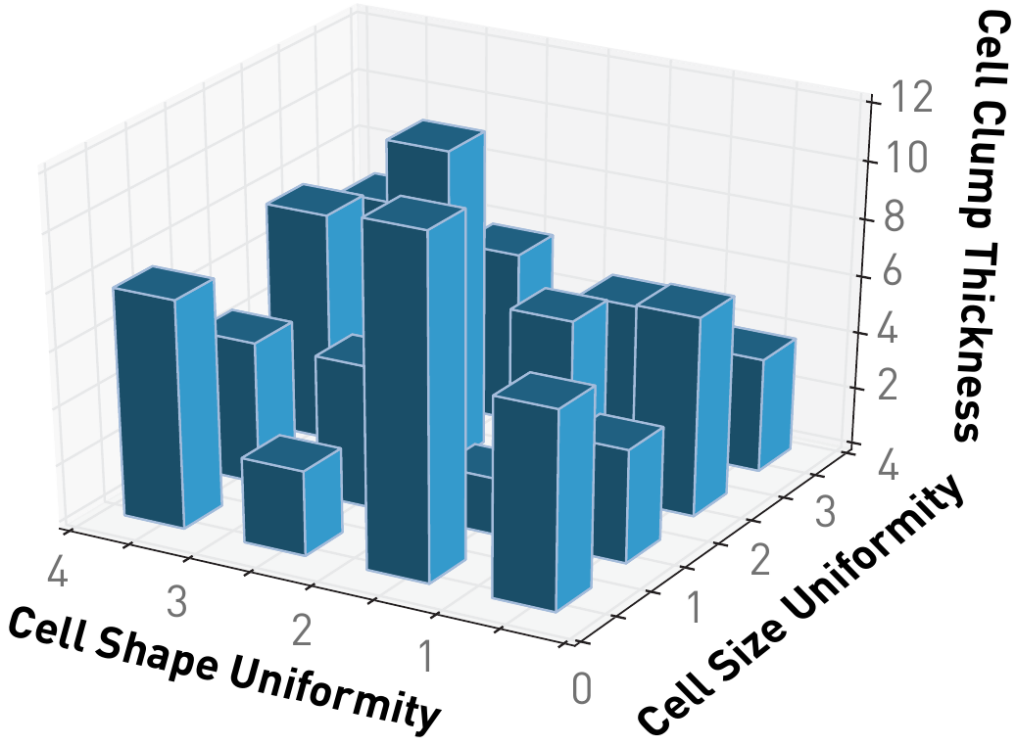
(Bad) Example: misleading axis



(Bad) Example: misleading colours (left)



(Bad) Example: occlusion



ACCENT checklist: evaluating a visualisation

- Apprehension: does the visualisation maximize the understanding of the relationships in the data?
- Clarity: are the most important relationships the most visually prominent?
- Consistency: are the visual elements consistent with their use in previous visualisations?
- Efficiency: are the visual elements economically used and easy to interpret?
- Necessity: is the visualisation more useful than alternative ways to represent the data?
- Truthfulness: are the visual elements accurate and unambiguous?

Questions?
also please use the forum on QM+