Edinburgh Napier UNIVERSITY

Data Analytics
SET10109

Multi Variable Visualisation

David Hunter
(d.hunter@napier.ac.uk)

Edinburgh Napier UNIVERSITY

Process Recap

- Project Understanding
  - What is the context of the data?
    - What should the data look like
    - What is the domain of the data?
  - What do you wish to learn from the data?

Edinburgh Napier UNIVERSITY

Spotting Patterns

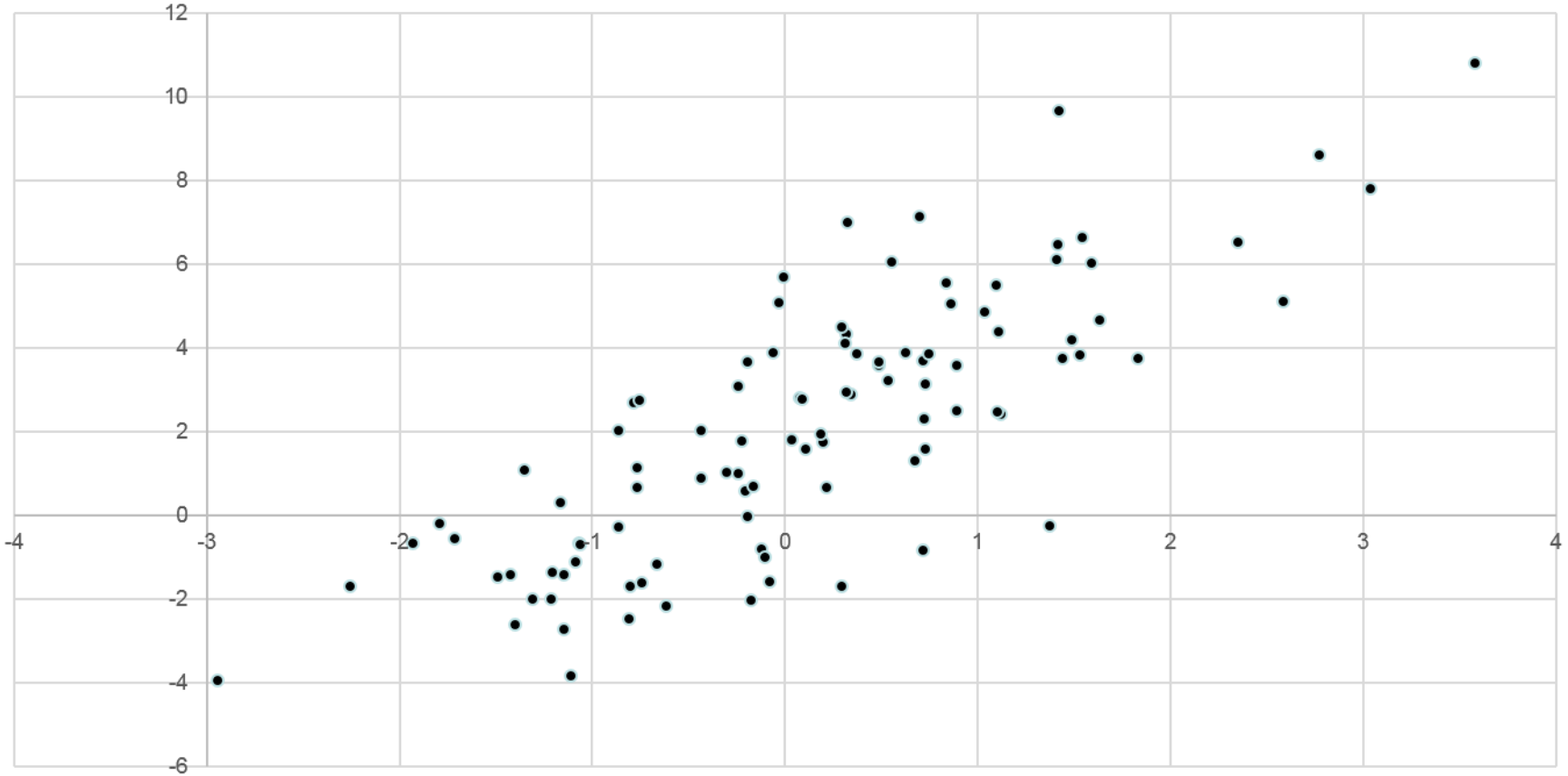
Shape of data is important to understanding it.

In scatterplots interactions can understood as the general shape of the dots

In a histogram patterns are visible in the heights of the bars.

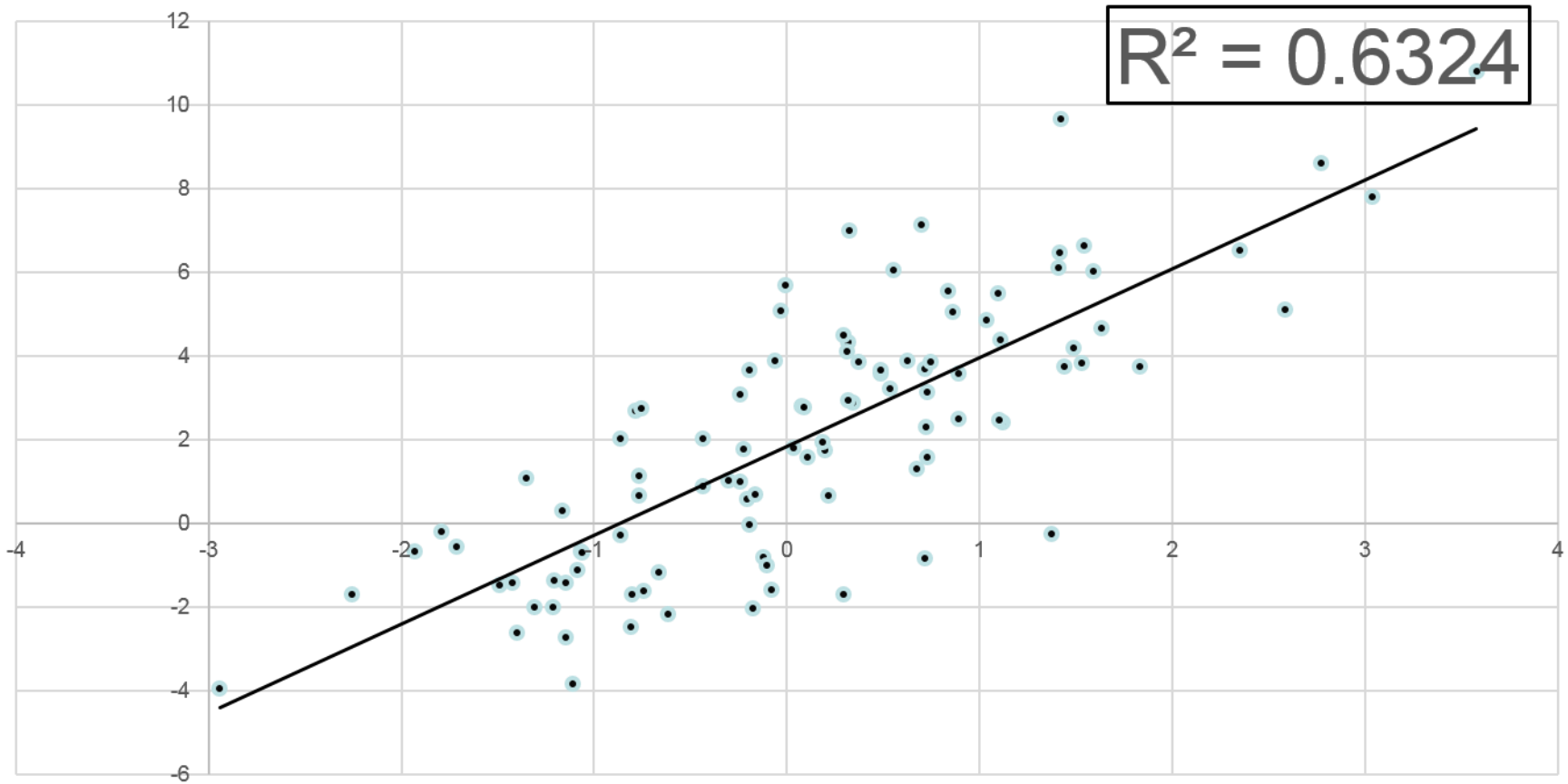
Edinburgh Napier UNIVERSITY

Regression

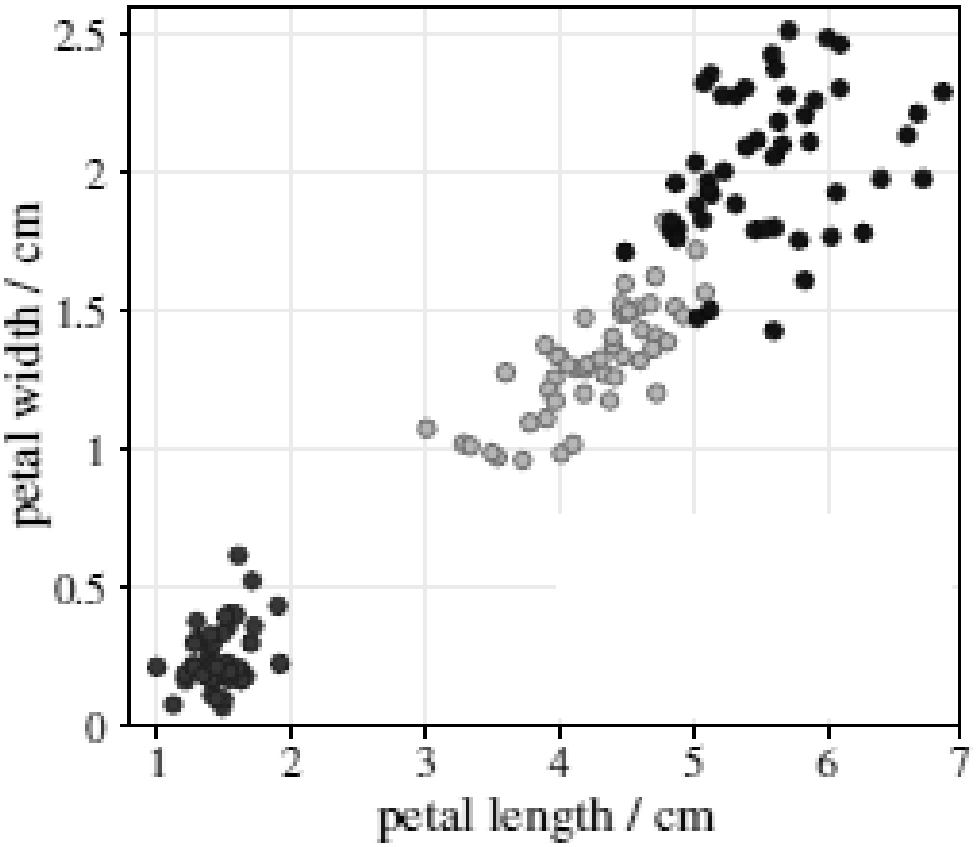


Edinburgh Napier UNIVERSITY

Regression



Scatterplots



Relationships between two variables can appear in a scatterplot

If two variables are correlated the point will approximate a line.

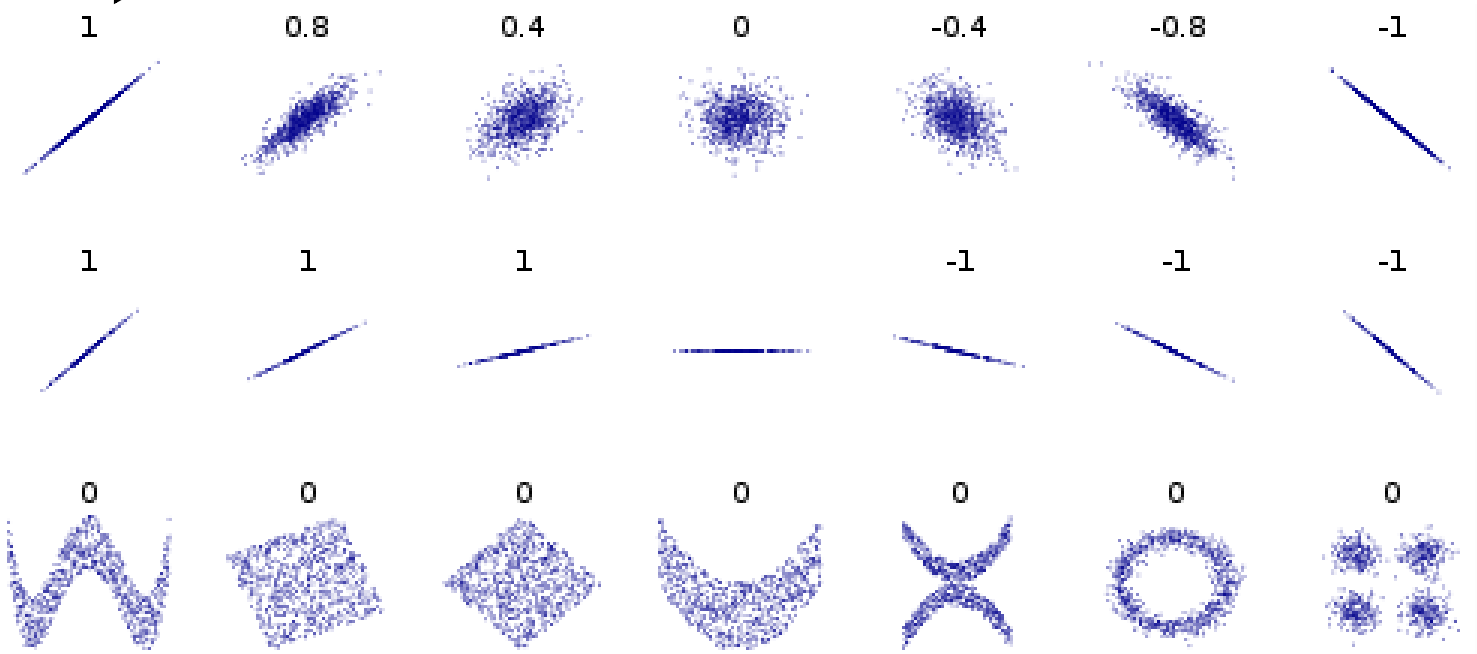
The narrower this line is the more correlated the data is.

Note if the line is horizontal or vertical the data is not correlated.

Edinburgh Napier UNIVERSITY

Correlation in scatterplots

R – coefficient of correlation
1 completely correlated
0 completely uncorrelated,
-1 completely correlated by in the opposite direction



Example data set: Iris data

iris setosa

iris versicolor

iris virginica

- collected by E. Anderson in 1935
- contains measurements of four real-valued variables:
- sepal length, sepal widths, petal lengths and petal width of 150 iris flowers of types Iris Setosa, Iris Versicolor, Iris Virginica (50 each)
- The fifth attribute is the name of the flower type.

What the heck is a sepal??

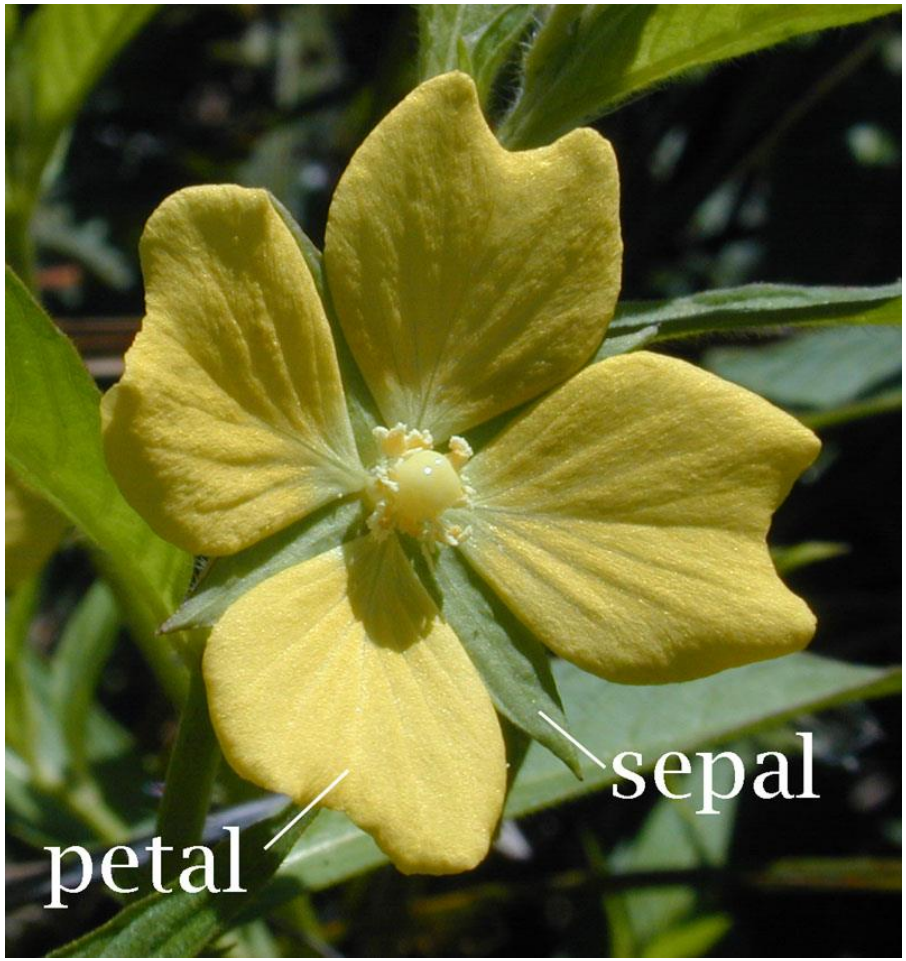


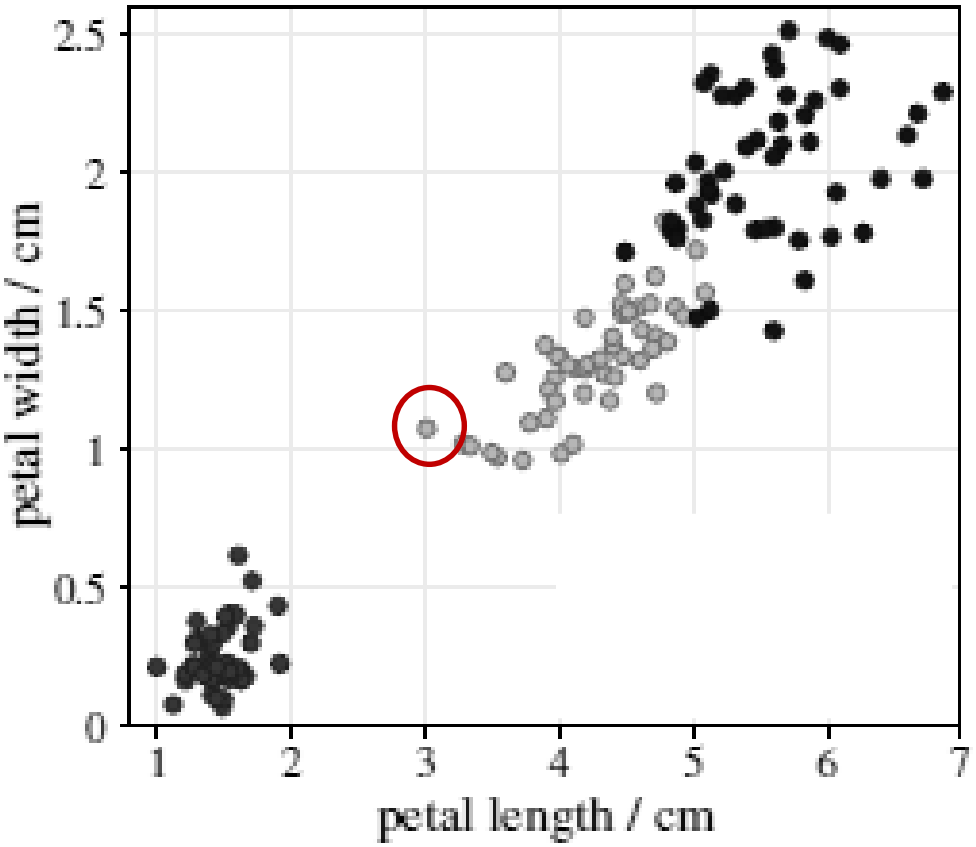
Image of a primrose willowherb *Ludwigia octovalvis* (family Onagraceae), flower showing petals and sepals. Photograph made in Hawai'i by Eric Guinther (Marshman at en.wikipedia) and released under the GNU Free Documentation License.

Example data set: Iris data

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
5.1	3.5	1.4	0.2	Iris-setosa
...				
...				
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
...				
...				
5.1	2.5	3.0	1.1	Iris-versicolor
5.7	2.8	4.1	1.3	Iris-versicolor
...				
...				
5.9	3.0	5.1	1.8	Iris-virginica

Scatterplots

2 dimensions are encoded using position (x,y)
Usually used for numeric data



Each axis corresponds to one variable, in this case:

x axis = petal length

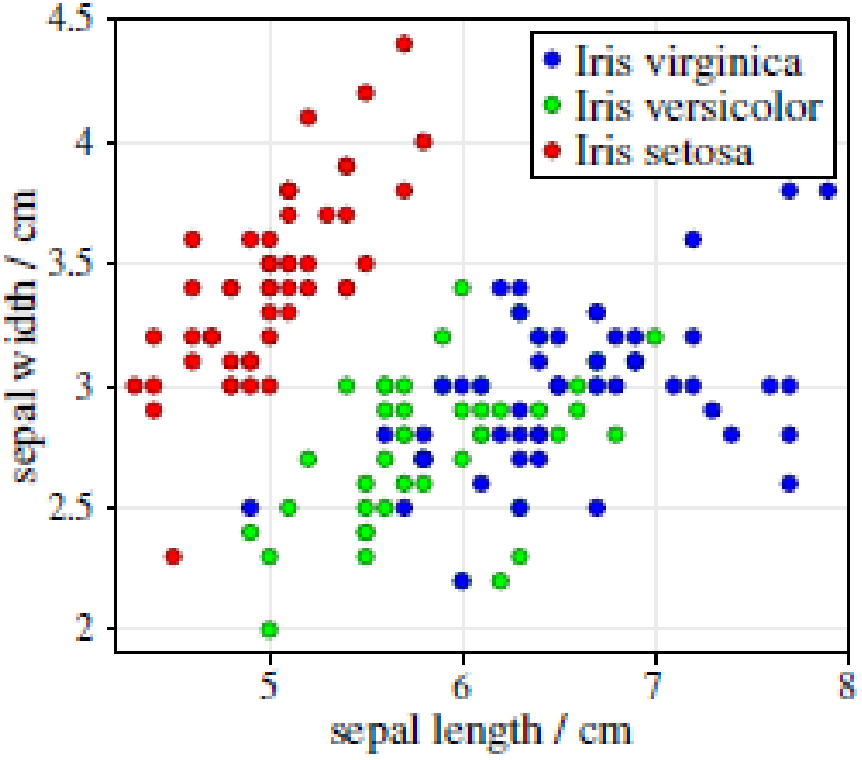
y axis = petal width

Each point represents the data values (petal length and width) for one data item e.g. for the circled value, petal length is 3cm, petal width is just over 1cm

(should have 150 points in total)

Scatter plots

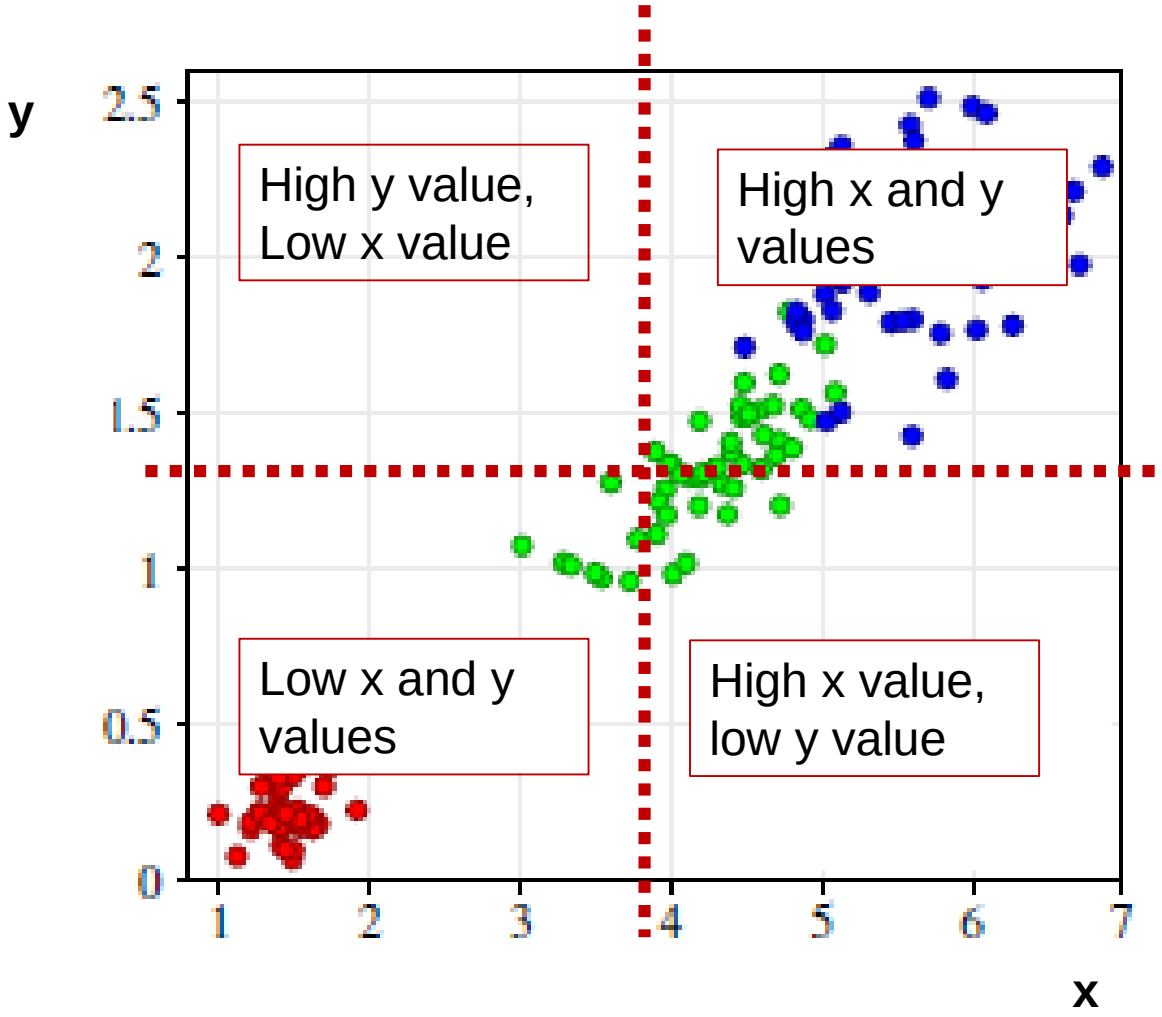
Encoding a 3rd dimension



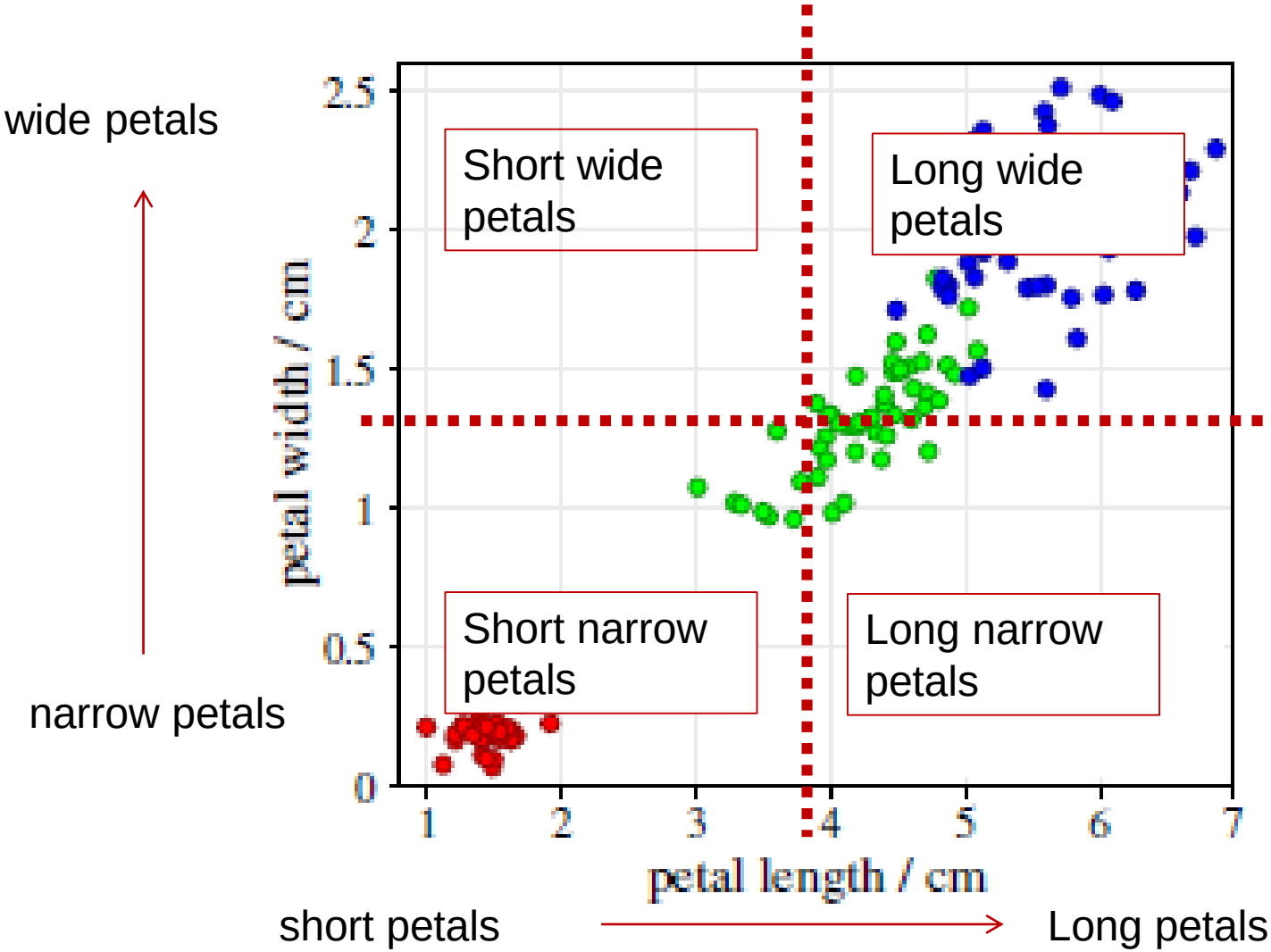
Scatter plots can be enriched with additional information: Colour or different symbols to incorporate a third attribute in the scatter plot.

Edinburgh Napier UNIVERSITY

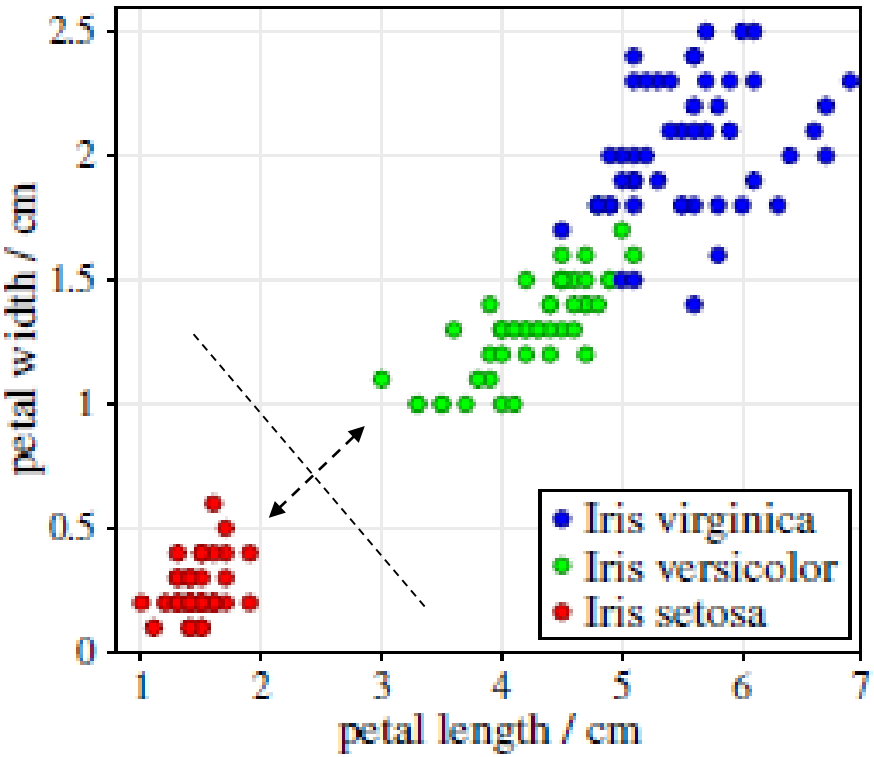
A note on reading scatterplots…



For this data set…



Scatter plots

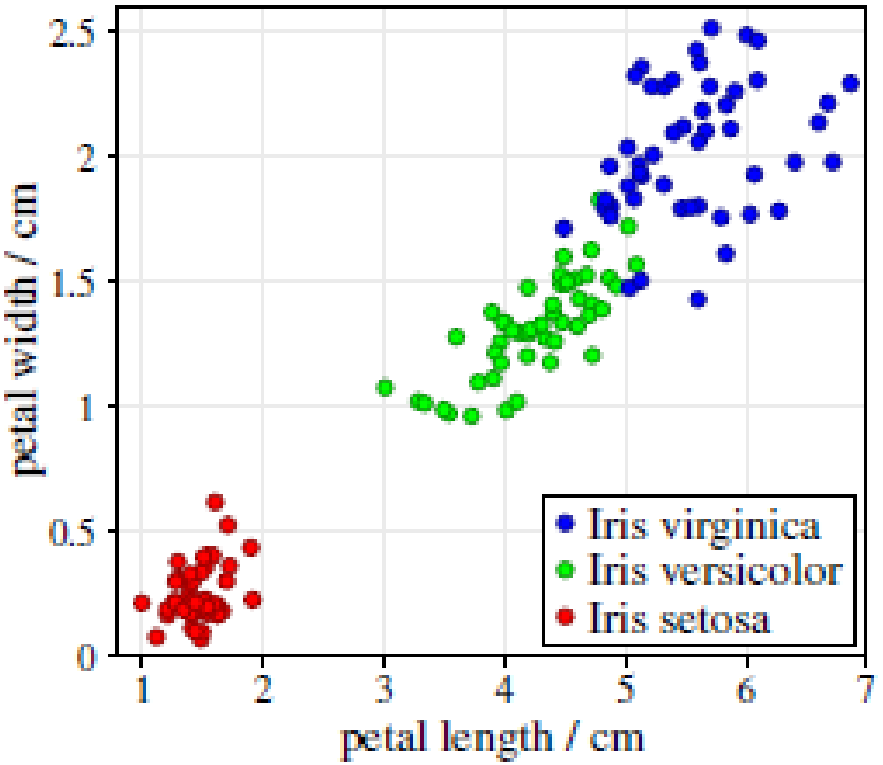


The two attributes petal length and width provide a better separation of the classes Iris versicolor and Iris virginica than the sepal length and width.

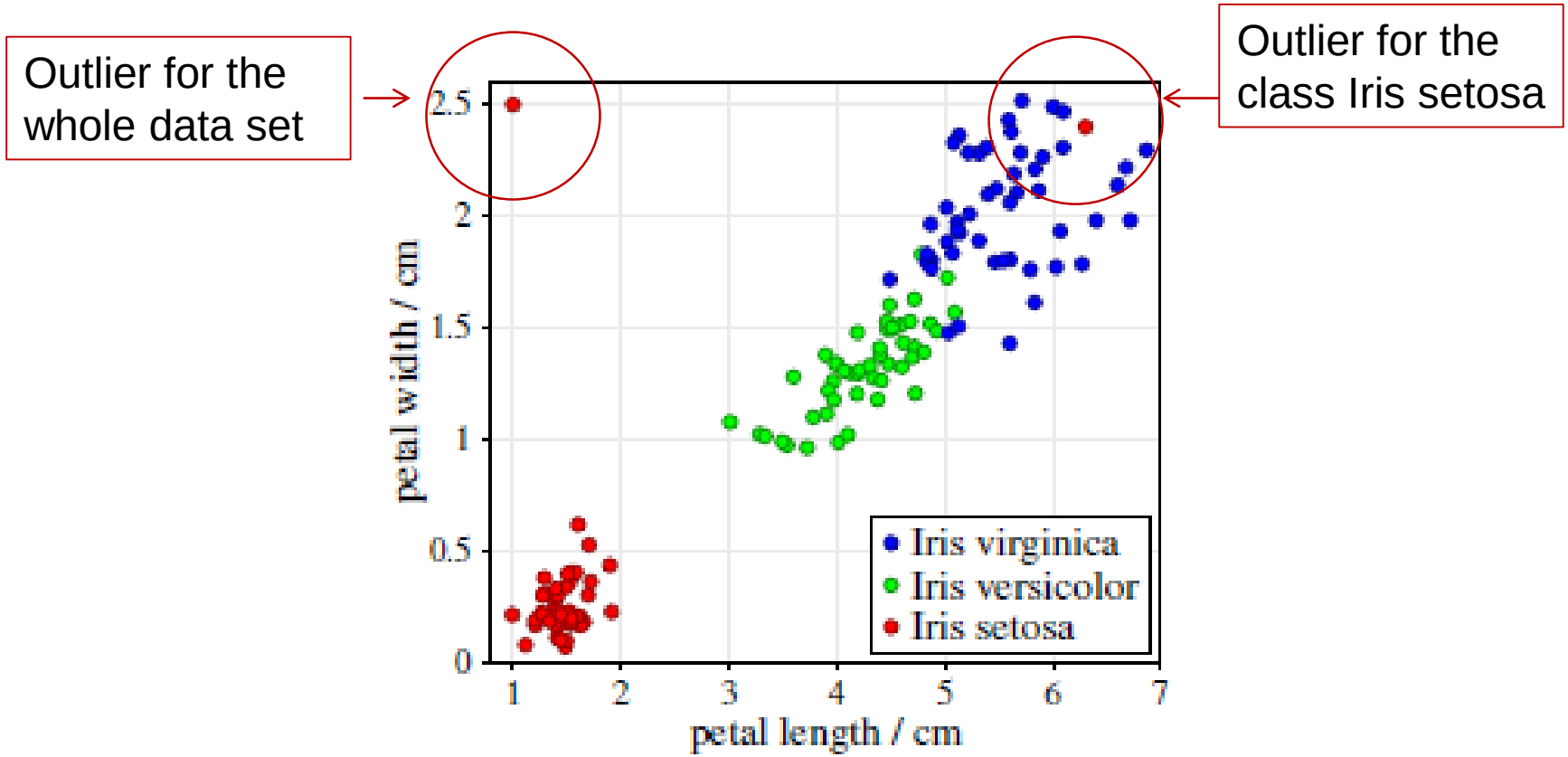
Scatter plots

Problem

Also, issue of occlusion when large numbers of points are plotted



Scatter plots

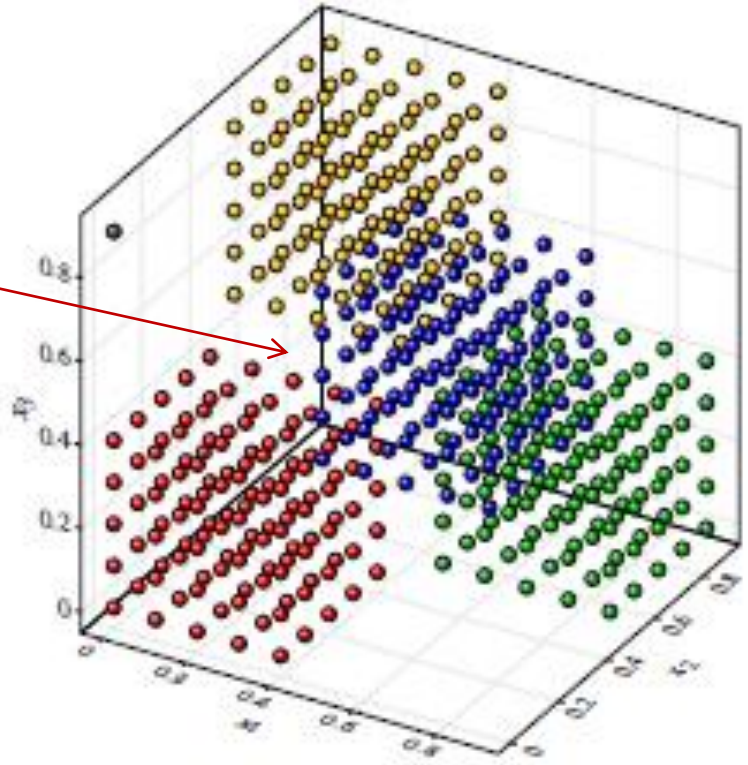


The Iris data set with two (additional artificial) outliers. One is an outlier for the whole data set, one for the class Iris setosa.

3D scatter plots

For data sets of moderate size, scatter plots can be extended to three dimensions.

Note problem of occlusion: what is going on in the middle here??!



Avoid !!

A 3D scatter plot of an artificial data set filling a cube in a chessboard-like manner with one outlier.

Parallel co-ordinates

Each axis is drawn parallel

-> no limit to the number of dimensions that can be shown

Each tuple (object in the data) becomes a line that intersects the axes according to its values for that dimension

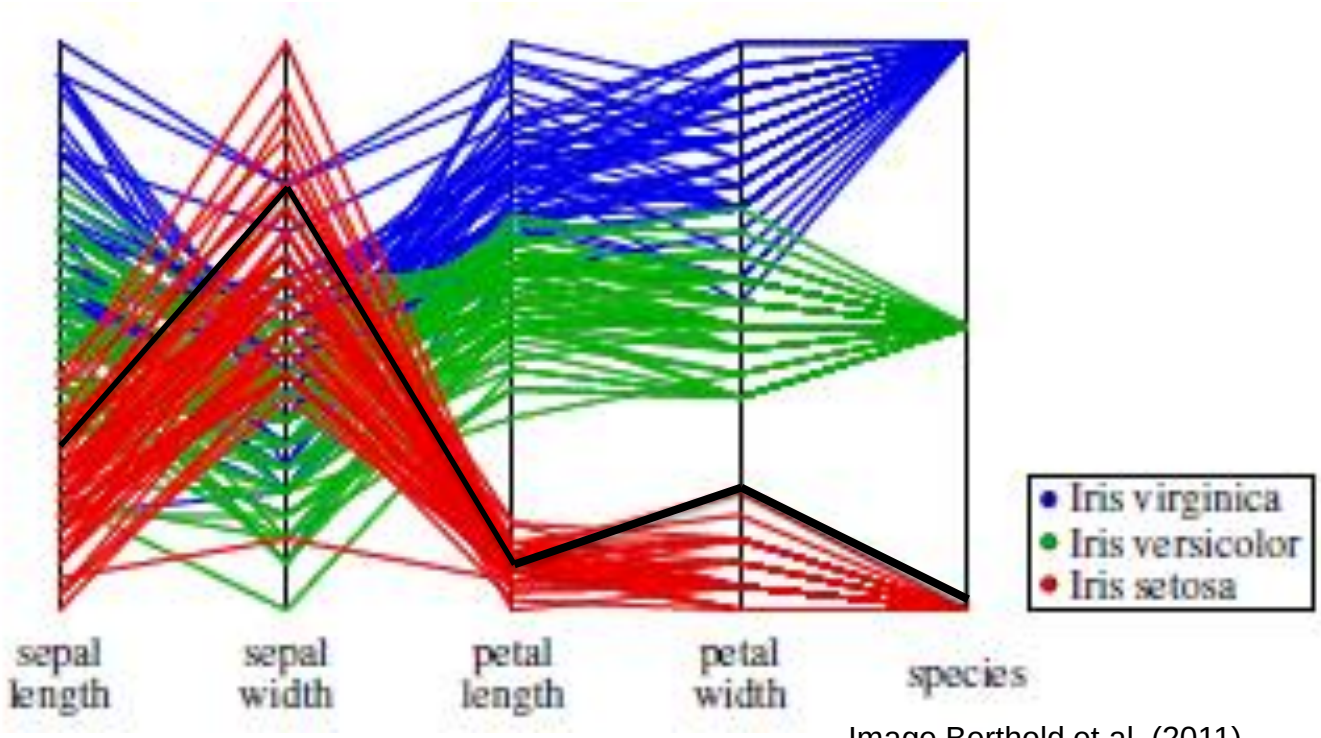


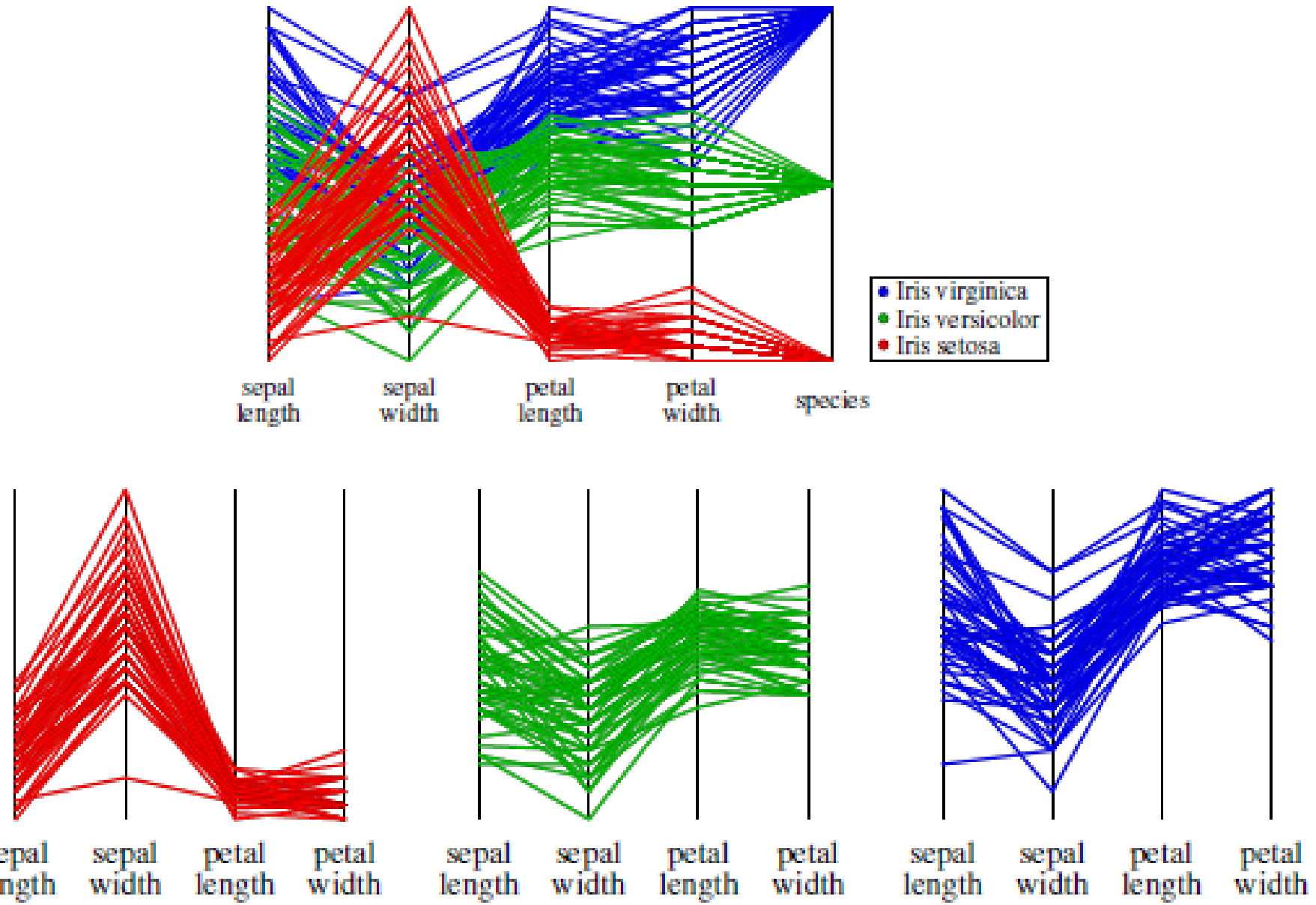
Image Berthold et al. (2011)

Similar items trace similar paths across the display, giving an overview of the dataset

Disadvantages:

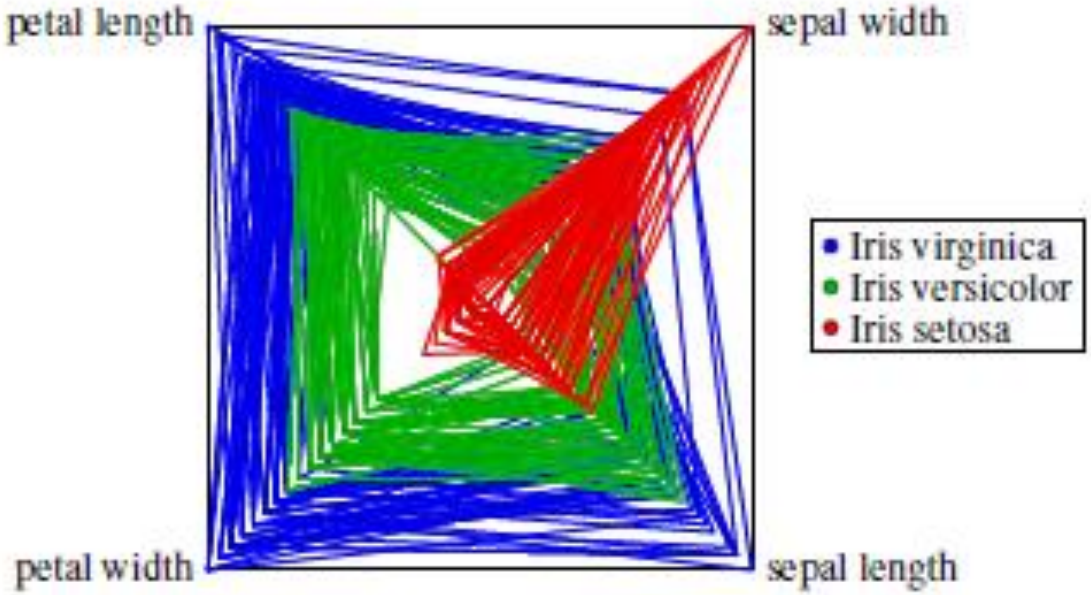
- Ordering effect. Can see correlations between neighbouring dimensions, but other combos?
- Over plotting/occlusion in large datasets (can use interaction to help with this)

Parallel coordinates: Iris data



Radar plots

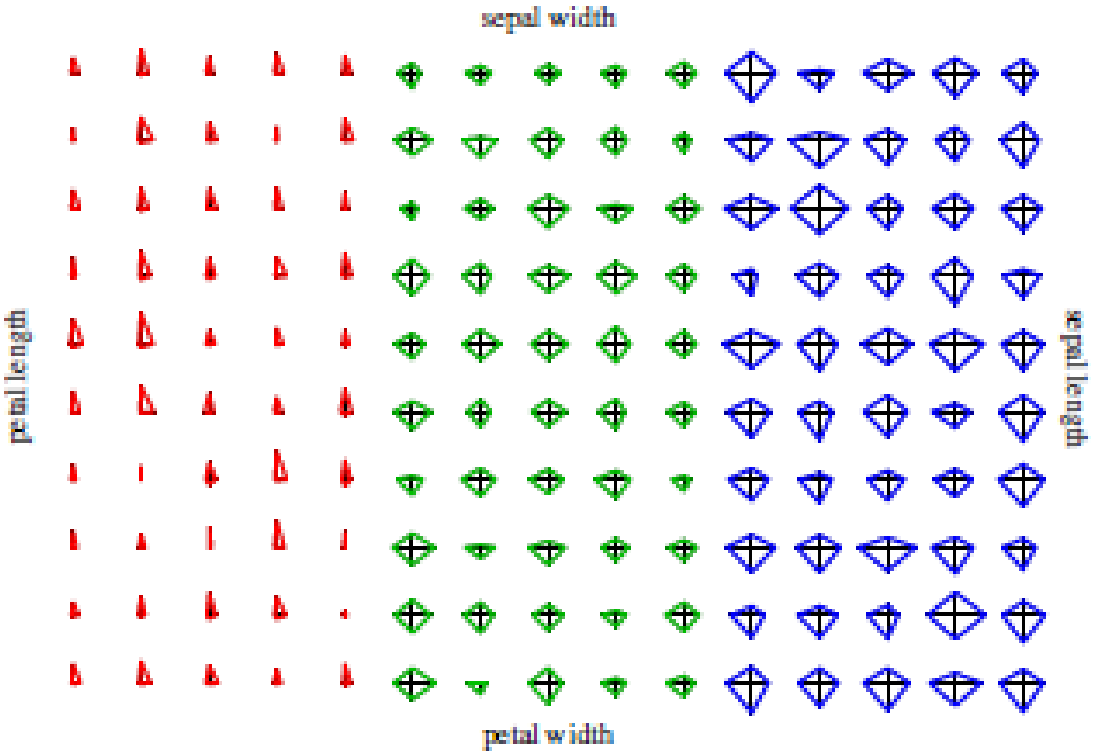
Radar plots are based on a similar idea as parallel coordinates with the difference that the coordinate axes are drawn as parallel lines, but in a star-like fashion intersecting in one point.



Radar plot for the Iris data set

Star plots

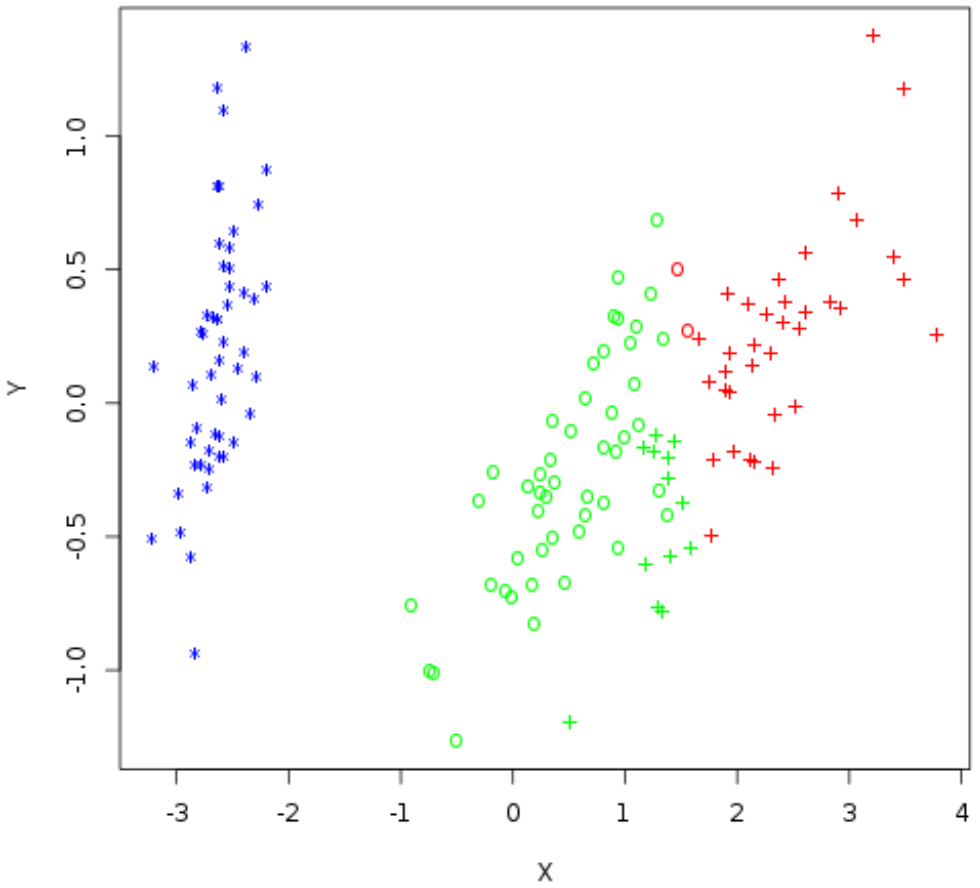
Star plots are the same as radar plots where each data object is drawn separately.



Star plot for the Iris data set

Multi-Dimensional Scaling (N dimensions)

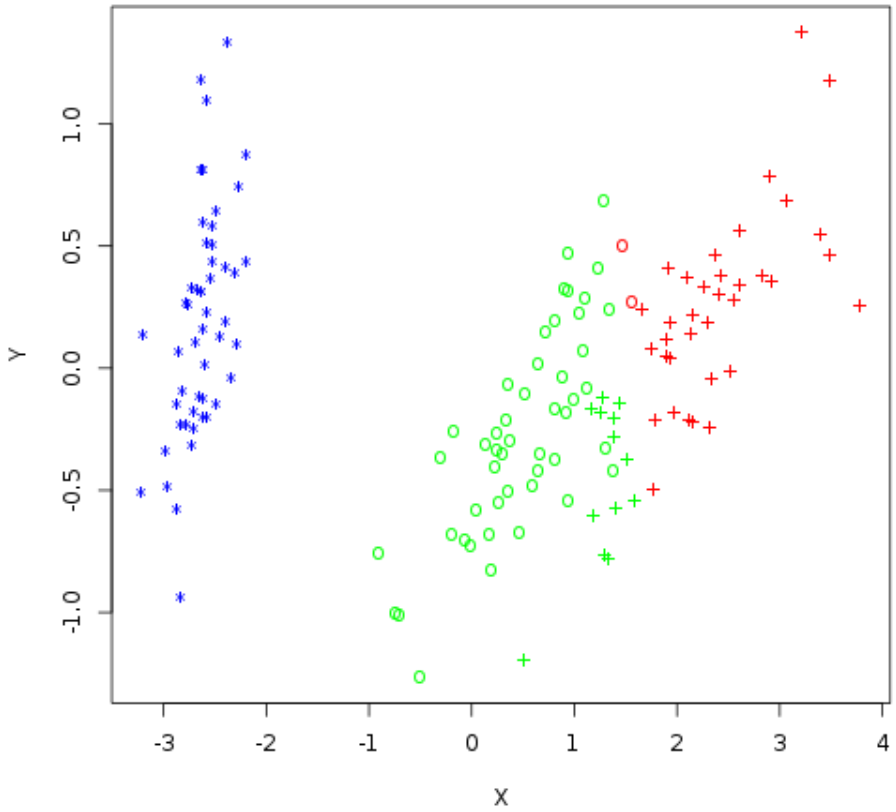
- Distances between items are calculated over **all** dimensions and then flattened to a 2D/3D view



Edinburgh Napier UNIVERSITY

Multi-Dimensional Scaling

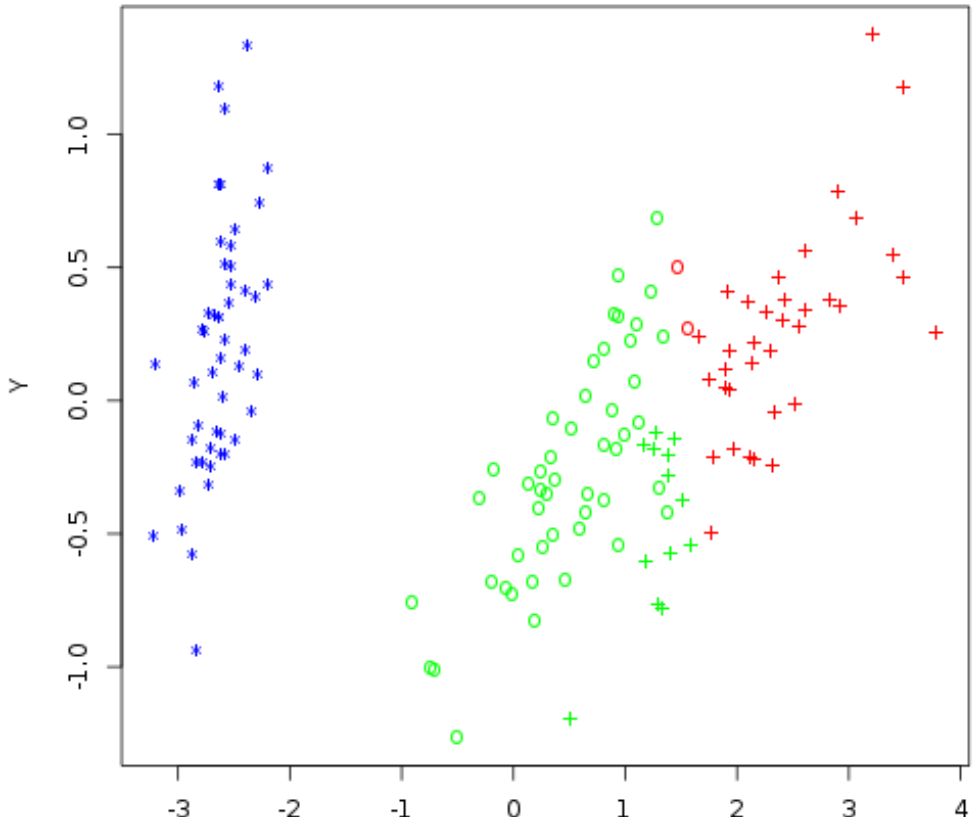
MDS algorithms try as much as possible to preserve the N-D distance between every pair of items in the 2D plot, but it's never perfect



Edinburgh Napier UNIVERSITY

Multi-Dimensional Scaling

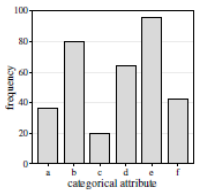
Effect is like that of a single scatterplot, though the axes themselves have no meaning, only the distances between items



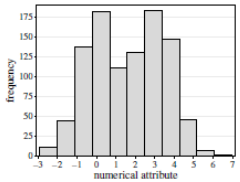
Some Common Layouts for Tabular Data

Recap

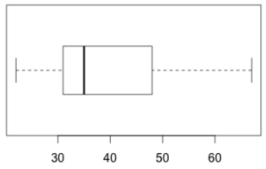
1D



Bar Chart

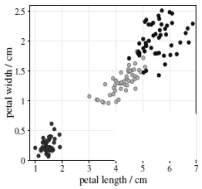


Histogram



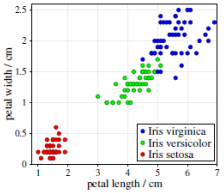
Boxplot

2D



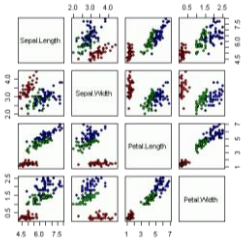
Scatter Plot

3D

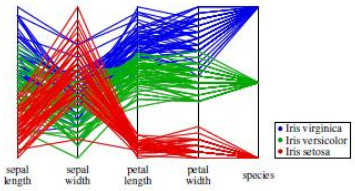


Scatter Plot +
1 non-spatial encoding

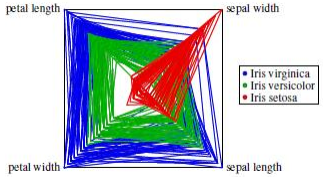
ND (3+)



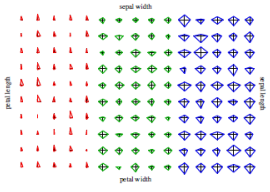
Scatter Plot Matrix



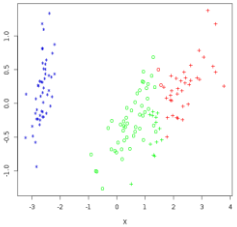
Parallel Co-ordinates



Radar Plot



Star Plot



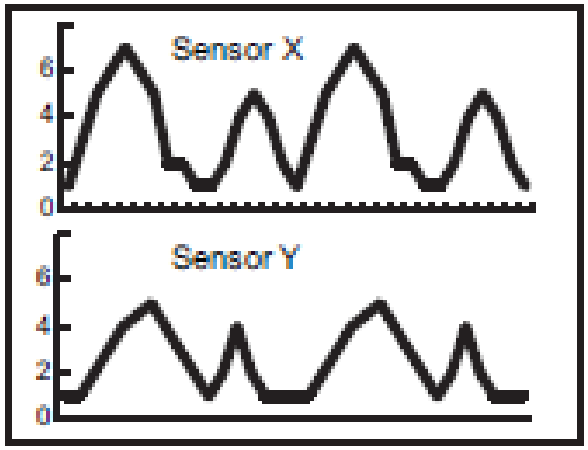
MDS

A more systematic approach

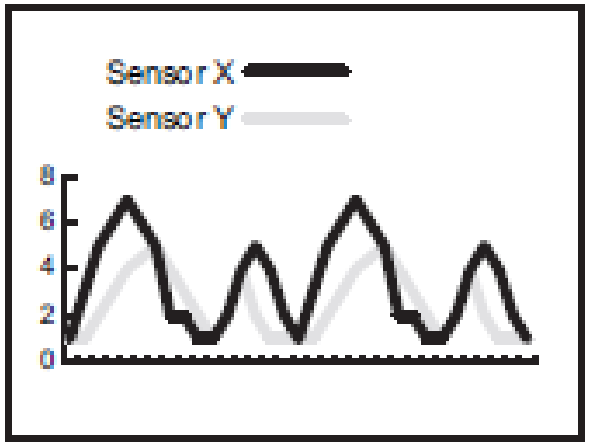
- What type of variable?
  - See previous slides
- What is the main variable?
  - Select a variable to base the visualisation on
- What are the important interactions?
  - Can’t visualise every interaction select most important.

Visual comparison

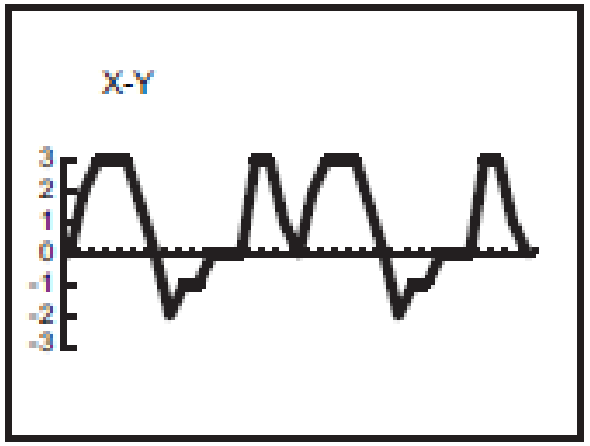
- Three basic categories of design for visual comparison (hybrids also possible):



a) Juxtaposition



b) Superposition



c) Explicit Encoding: Difference

Uses…

Peripheral Vision
viewer's memory

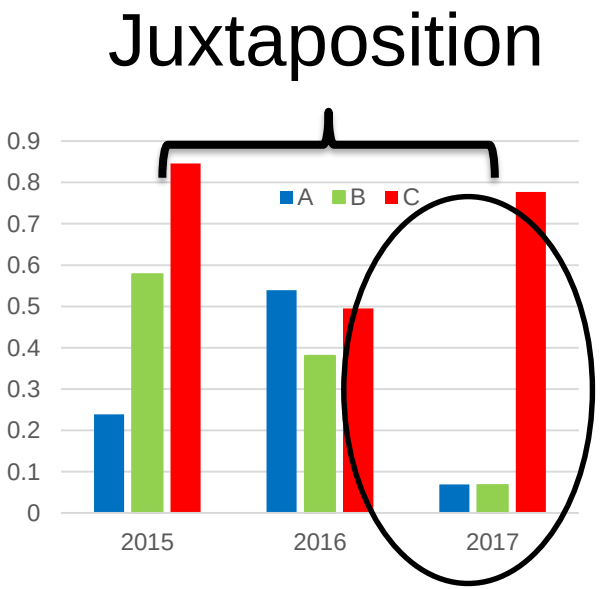
visual system

computation

(Gleicher et al., 2011)

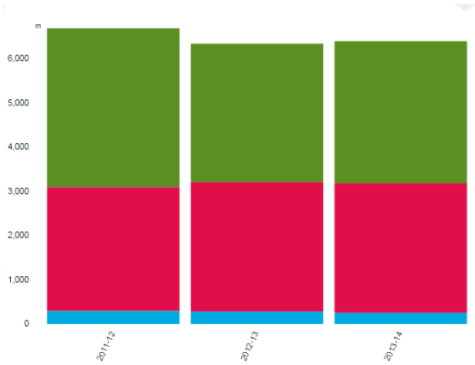
…to determine relationships

Example

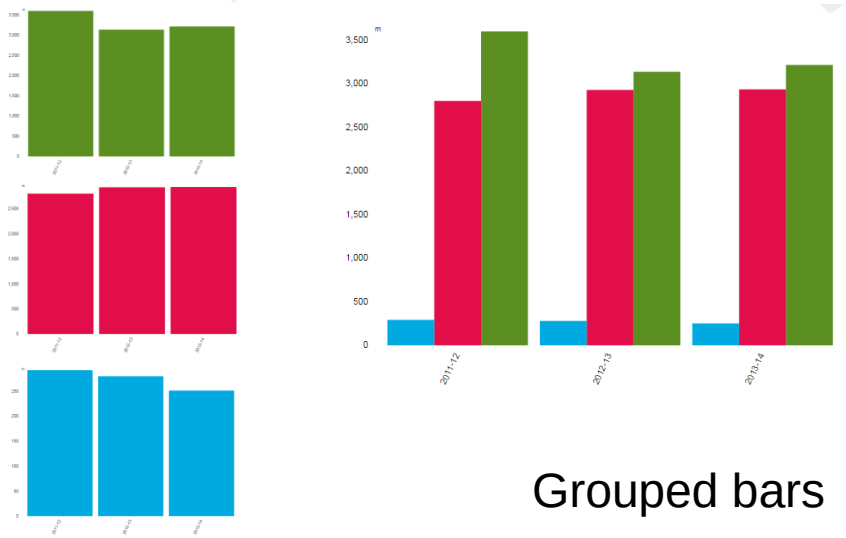

Juxtaposition
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A
B
C
2015
2016
2017
Superposition
(approx)
Grouped bars

Multiple categorical variables – some options

Edinburgh Napier UNIVERSITY

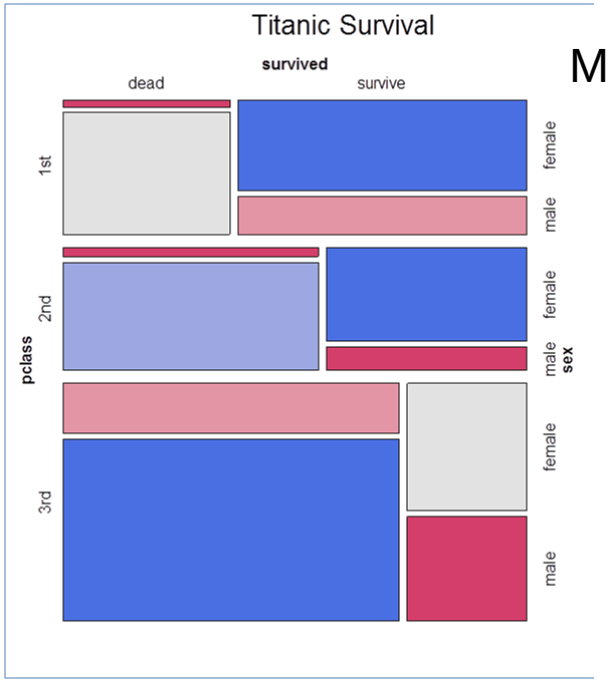


Stacked bars



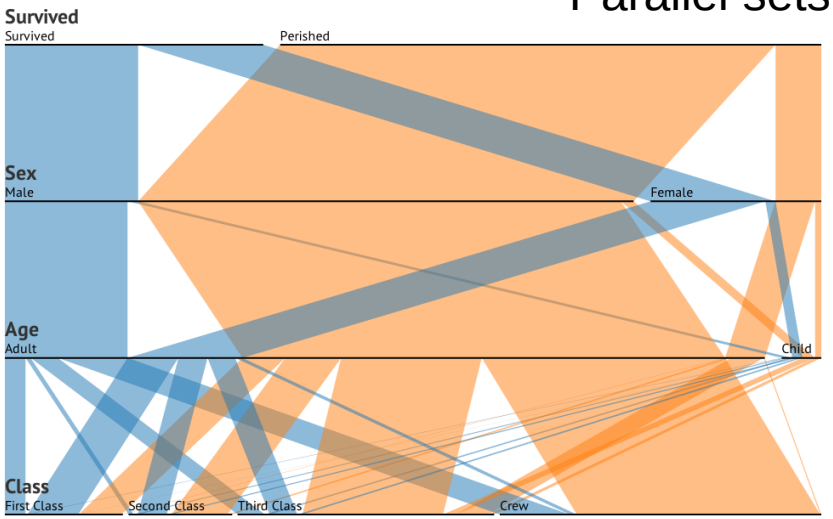
Grouped bars

Multiple/separate plots



Mosaic plot

Parallel sets



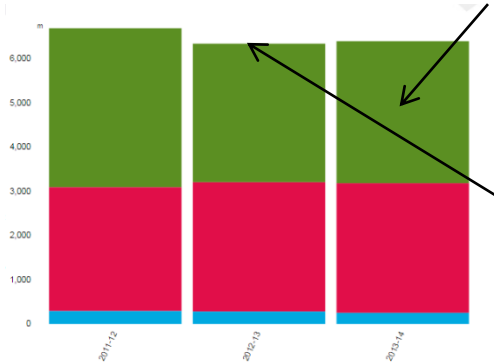
Parallel sets see: https://www.jasondavies.com/parallel-sets/

Stacked bars

- A multi-dimensional version of the histogram/bar chart. Shows counts in each **combination** of category.

Compares Categories by size

Category 2
- Colour channel

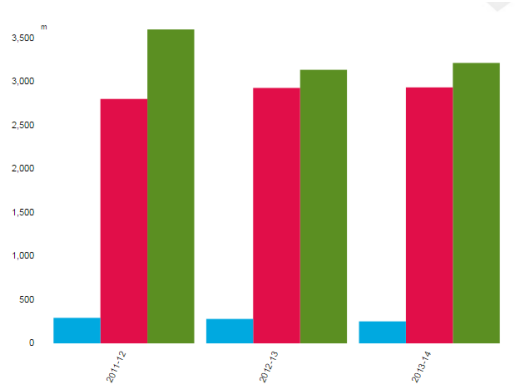


Compare 1 category by bar height

Category 1
Spatial Channel

Grouped bars

- Here multiple categories are combined into a single spatial channel (the horizontal axis)



Grouped bars
2 Categories Combined

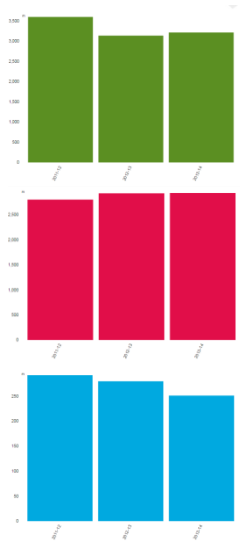
Comparisons within the coloured variable are easy.

Comparisons across categories are possible by drawing a mental line, (medium)

Ability to say anything about the total size of variables lost

Separate/Multiple plots

- Separate by a variable, one category per plot



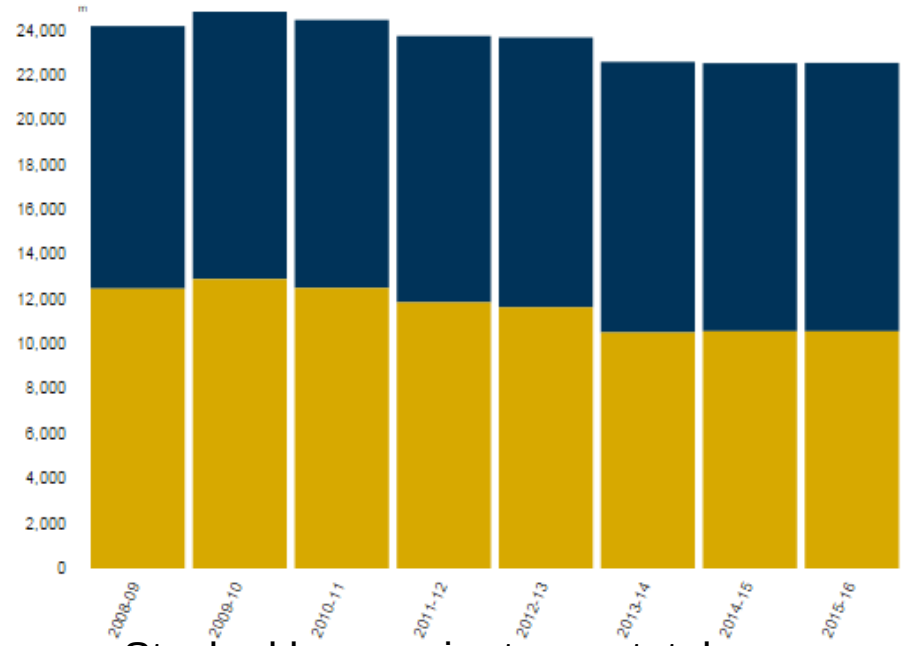
Multiple/separate plots

Comparisons within the coloured variable are easy.

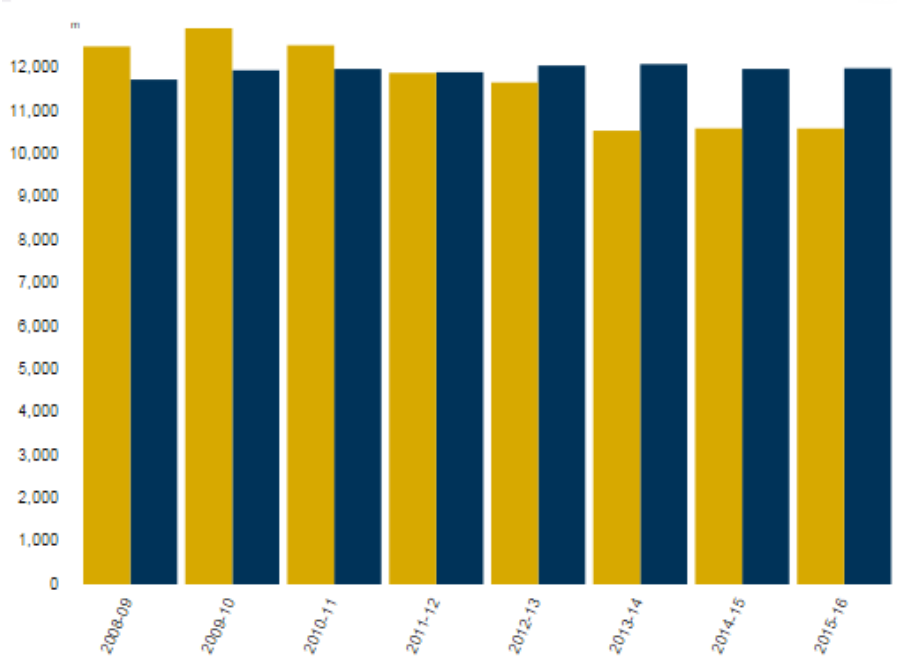
Comparisons across categories are hard

Some limited ability to comment on relative sizes

Example 1: more than one view for full understanding of the data



Stacked bar: easier to see total trend over time (and individual trend over time for bottom stack)



Grouped bar: easier to compare individual years

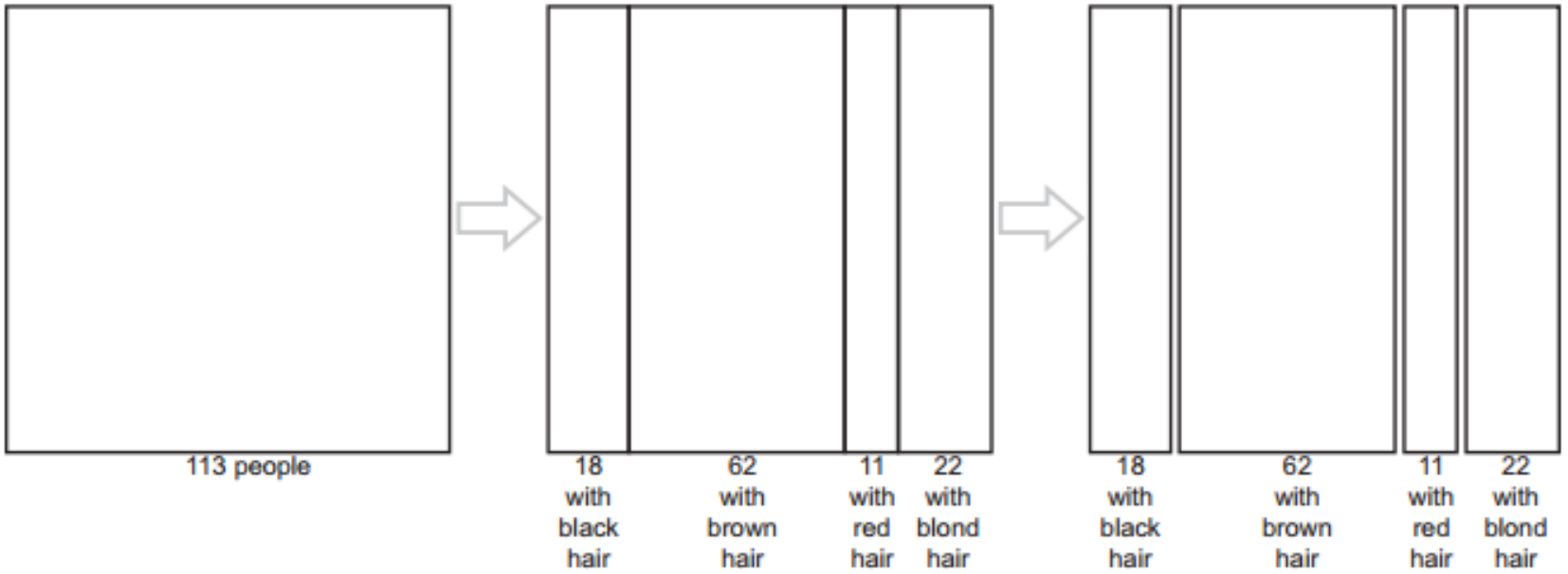
Different views help answer different questions

http://www.scottish.parliament.uk/parliamentarybusiness/82906.aspx

Mosaic plots - explanation

	Hair Color				
Eye Color	Black	Brown	Red	Blond	Total
Brown	10	25	4	2	41
Blue	4	18	3	15	40
Hazel	3	13	2	2	20
Green	1	6	2	3	12
Total	18	62	11	22	113

1. Divide the area proportionally according to one attribute (hair colour)



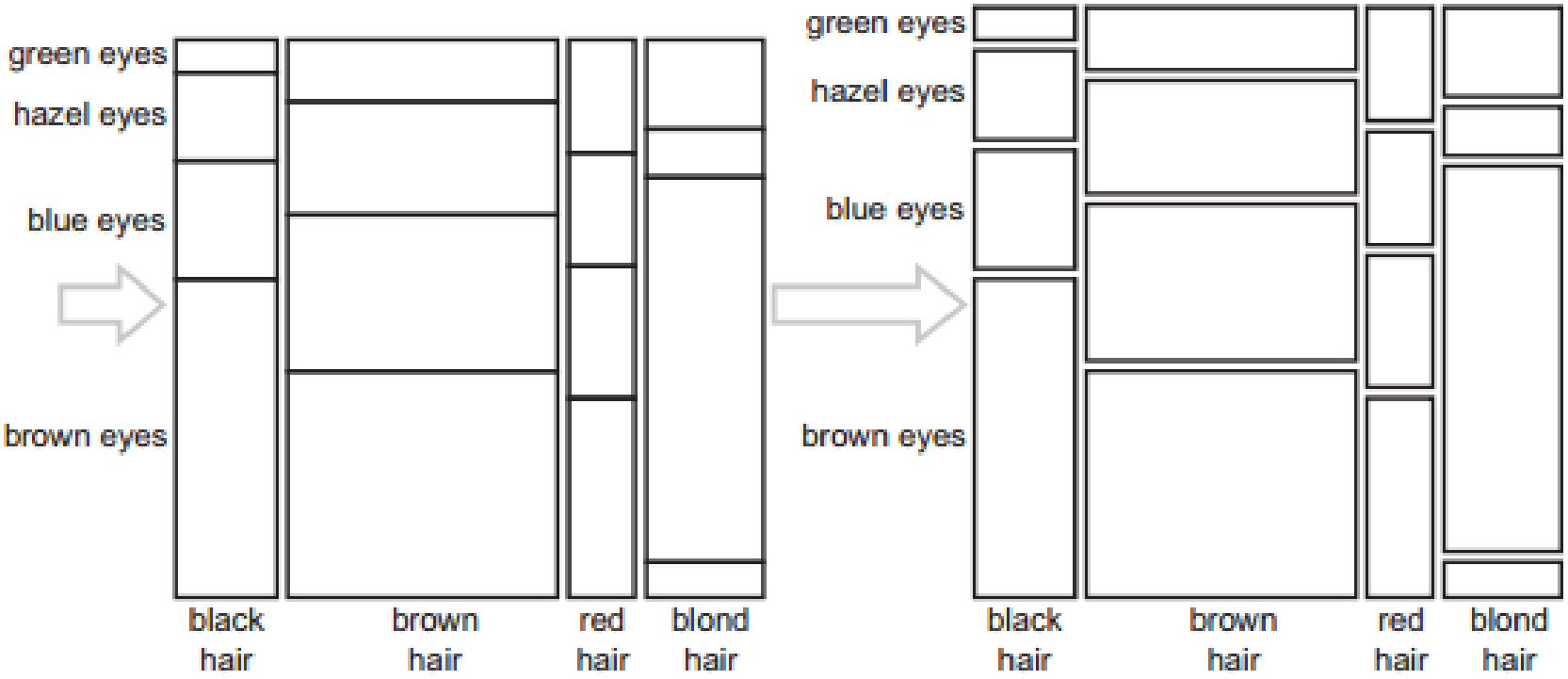
Source: Stephen Few (2014). Are Mosaic Plots Worthwhile?
https://www.perceptualedge.com/articles/visual_business_intelligence/are_mosaic_plots_worthwhile.pdf

See also: Michael Friendly (1999) Visualising Categorical Data http://www.datavis.ca/papers/casm/casm.pdf

Mosaic plots - explanation

	Hair Color				
Eye Color	Black	Brown	Red	Blond	Total
Brown	10	25	4	2	41
Blue	4	18	3	15	40
Hazel	3	13	2	2	20
Green	1	6	2	3	12
Total	18	62	11	22	113

2. Divide each area horizontally according to second attribute (eye colour)

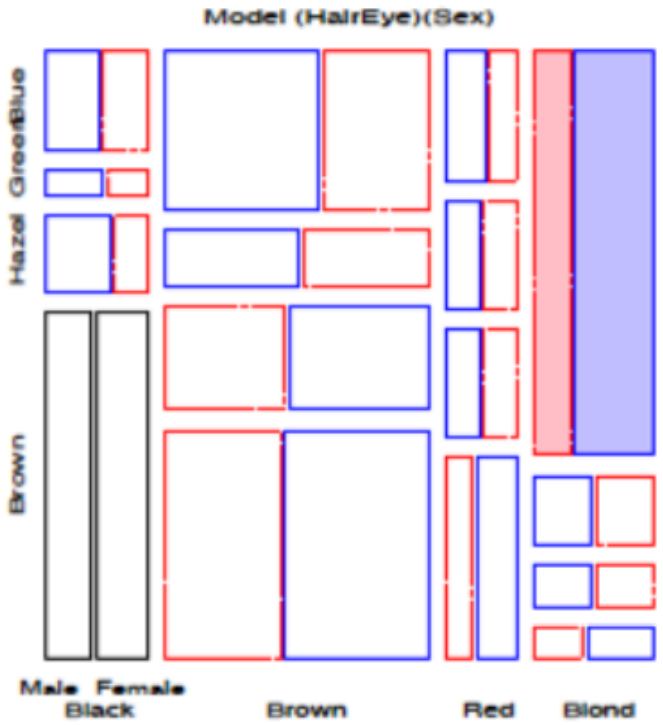


Source: Stephen Few (2014). Are Mosaic Plots Worthwhile?
https://www.perceptualedge.com/articles/visual_business_intelligence/are_mosaic_plots_worthwhile.pdf
See also: Michael Friendly (1999) Visualising Categorical Data http://www.datavis.ca/papers/casm/casm.pdf

Mosaic plots - explanation

	Hair Color				
Eye Color	Black	Brown	Red	Blond	Total
Brown	10	25	4	2	41
Blue	4	18	3	15	40
Hazel	3	13	2	2	20
Green	1	6	2	3	12
Total	18	62	11	22	113

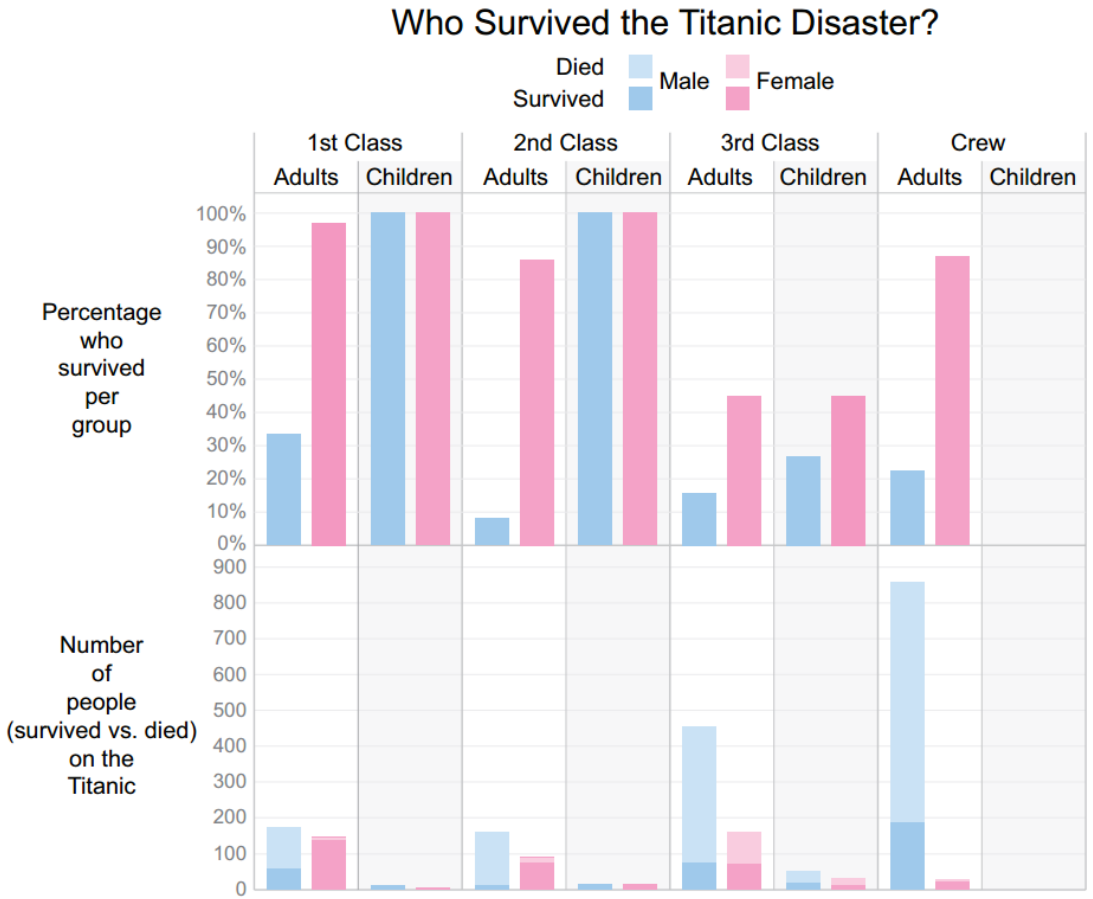
3. Could further subdivide the areas, e.g. vertically, by gender (NB ignore the colouring here)



Source: Stephen Few (2014). Are Mosaic Plots Worthwhile?
https://www.perceptualedge.com/articles/visual_business_intelligence/are_mosaic_plots_worthwhile.pdf

See also: Michael Friendly (1999) Visualising Categorical Data http://www.datavis.ca/papers/casm/casm.pdf

Alternative - multiple/grouped bar charts



Source: Stephen Few (2014). Are Mosaic Plots Worthwhile?
https://www.perceptualedge.com/articles/visual_business_intelligence/are_mosaic_plots_worthwhile.pdf

Multiple views

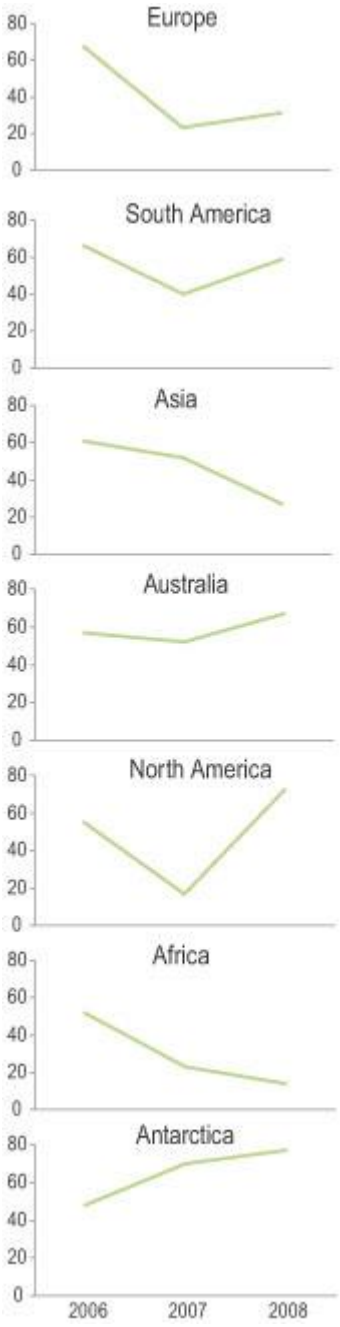
Small multiples example - timeslices

Moderate to extreme drought in the US from 1899 to 2012, The New York Times
http://www.nytimes.com/interactive/2012/07/20/us/drought-footprint.html

Multiple views

- Small multiples (Tufte)
- Enable comparison across variables
- Discussed in lectures 5 & 6

Heer and Shneiderman (2012) figure 11: employment figures by economic sector



Edinburgh Napier UNIVERSITY

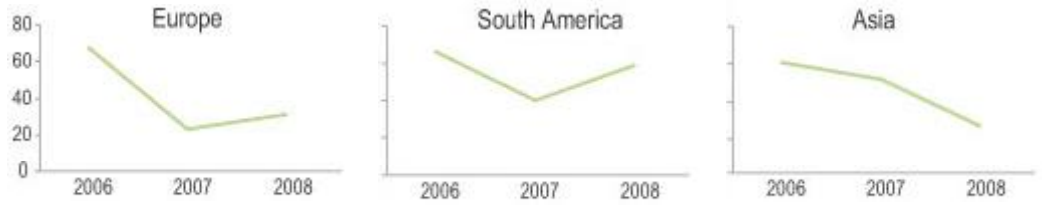
Juxtaposition and importance ordering

Example:

Small multiples arranged …

…vertically (aligned on x axis)
For comparison of patterns of change over time

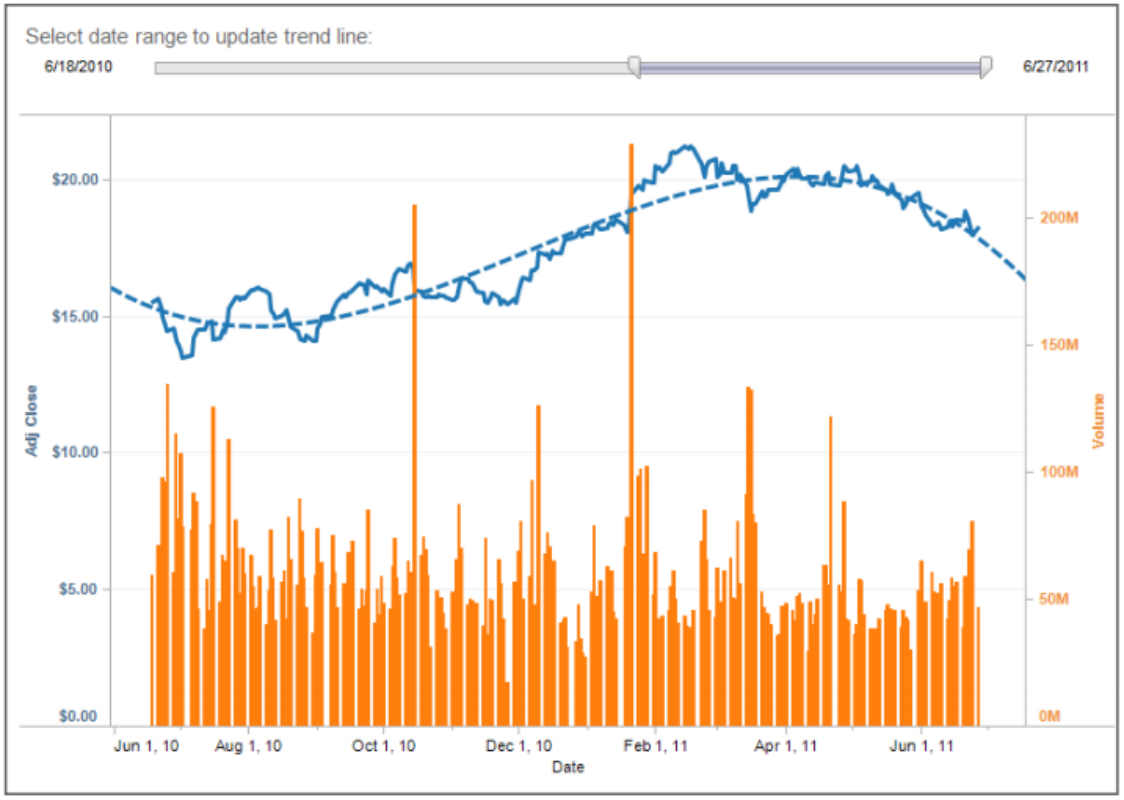
…horizontally (aligned on y axis)
For comparison of magnitudes



Source: http://www.perceptualedge.com/example11.php

Combining views using superposition

Edinburgh Napier UNIVERSITY



Also possible to combine different plot types in the same space.

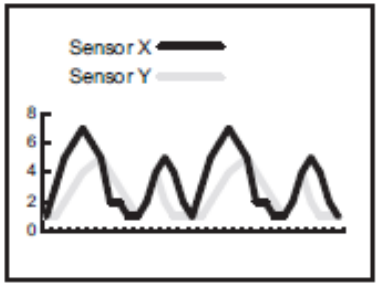
Line chart shows stock price over time

Bar chart shows volume sold per day

Allows us to see the relationship between two different attributes

Shows there were two significant events: one resulting in a sell-off and the other a gain for shareholders.

Source: Tableau (2012) Which chart or graph is right for you? Figure 5.
http://www.tableausoftware.com/learn/whitepapers/which-chart-or-graph-is-right-for-you



b) Superposition

Edinburgh Napier UNIVERSITY

IMPORTANT!!

WHAT YOU SHOULD HAVE LEARNED DURING THIS PART OF THE MODULE…

Edinburgh Napier UNIVERSITY

What you should take away from this section of the module…

- What visualisation is; why we would use it; when its use is appropriate and not appropriate.
- The differences between using visualisation for analysis and for presentation (in terms of audience, the purpose/goal of the visualisation, the techniques used).
- The properties of data that we need to consider when creating a visualisation.
- An awareness of the range of visual encodings (marks, channels, layouts) available when creating a visualisation.
- How to choose an appropriate visual encoding for the data: the considerations that we need to make when mapping data to a visual representation.
- Juxtaposition, Superposition, explicit coding.

Edinburgh Napier
UNIVERSITY

QUESTIONS?

Edinburgh Napier UNIVERSITY

Some useful resources

- Glasgow University's STEPS project – basic overview of data types and charts http://www.stats.gla.ac.uk/steps/glossary/presenting_data.html
- Tableau whitepaper: Which chart or graph is right for you? http://www.tableausoftware.com/learn/whitepapers/which-chart-or-graph-is-right-for-you explains which chart to use when, and useful tips on combining chart types.
- Tableau online help for building each chart type: http://onlinehelp.tableausoftware.com/v8.0/pro/online/en-us/help.htm#dataview_examples.html%3FTocPath%3DExamples%7C_____0
- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. ACM*, *53*(6), 59-67. Available at http://queue.acm.org/detail.cfm?id=1805128

Study guide for this lecture

Edinburgh Napier UNIVERSITY

Required reading:

- Berthold, Borgelt, Höppner, and Klawonn. Guide to intelligent data analysis. Vol. 42. Springer, 2010. Chapter 4.

Recommended:

- Heer, J., Bostock, M., & Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. ACM*, *53*(6), 59-67. Available at http://queue.acm.org/detail.cfm?id=1805128
- Munzner (2014). *Visualization Analysis and Design.* Chapter 7.

Reflective questions:

- Discuss layouts that can be used to show:
  - a single data dimension
  - two data dimensions
  - multiple dimensions
- For each layout, describe how the attributes are encoded, discuss any advantages or limitations of the layout, and give an example of a situation where it would be appropriate to use the layout.

Useful resource - chart types: http://www.datavizcatalogue.com/

References

Tukey, J.W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading

Compendium slides for Guide to Intelligent Data Analysis, Springer 2011. Michael R. Berthold, Christian Borgelt, Frank Höppner, Frank Klawonn and Iris Ad. http://www.informatik.uni-konstanz.de/gidabook/teaching-material/?print=1