# Data Visualisation

Jingpeng Li



# Data Visualisation

- Our eyes are very good at data mining
- We can spot patterns, trends and clusters instantly in plotted data
- Problems begin when data covers more than a few dimensions
- Provides a good way to choose a more powerful data mining technique

# When to Use It

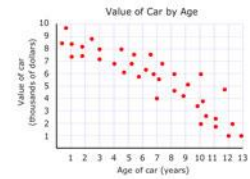
- Before starting a data mining project, to understand the problem
- To guide the data mining project and choice of technique
- To improve the use of data mining techniques, e.g. choosing a number of clusters
- To show the results of a data mining analysis



# Scatter Plots


- Perfect for seeing how one variable changes with another
- Can be used to see how well one variable predicts another
- Can be used to see how two variables combine to form clusters or a state space

# A Word on Graphs

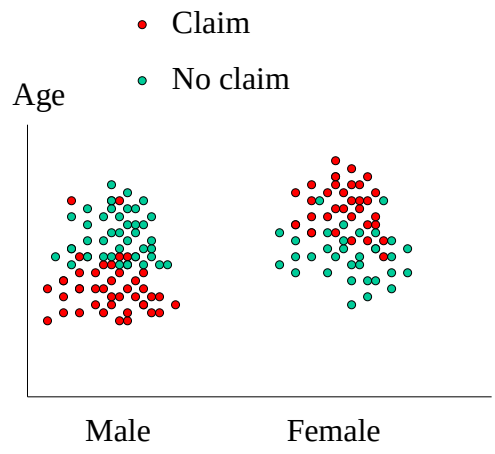



- Always give your graph a title
- Always label both axes with variable names and, if appropriate units (e.g. Spend in pounds or Number of products sold)
- Always show the scale of both axes
- Bar charts are for frequencies (counts of things)
- Line graphs are for continuous variables



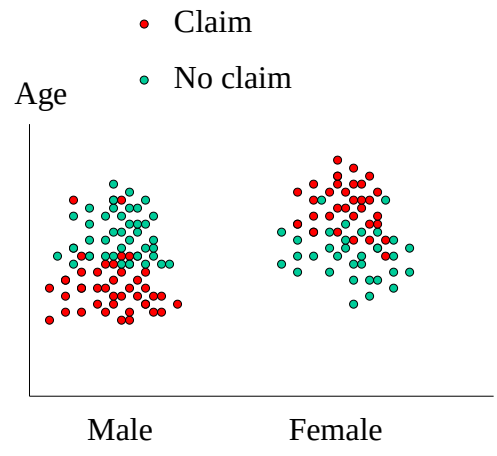
# Scatter Plots – Insurance Claims

- Here is an example from a previous lecture
- It is easy to see that younger males and older females make claims



# Class Labels

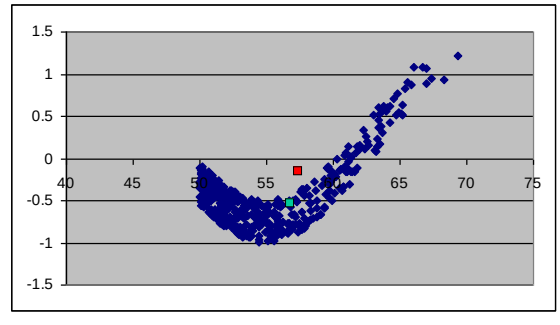
- Notice how the plot uses colour to represent the outcome class




# Scatter Plots – Machine Monitoring

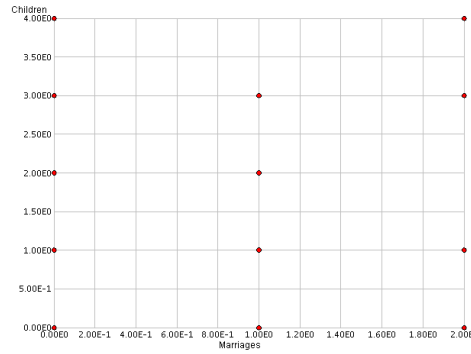
- Another previous example – machine health monitoring

This plot shows the operating relationship between temperature and pressure in a machine



# Overlap Problems

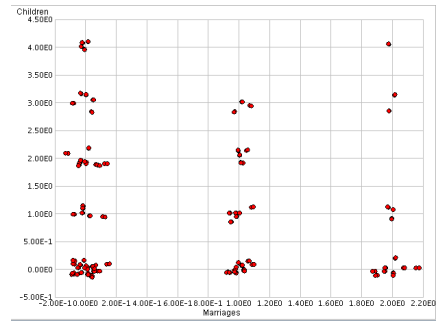
- Look at this plot, which plots the number of marriages a person has had against number of children they have
- We cannot tell if there are 1 or 100 examples at each point





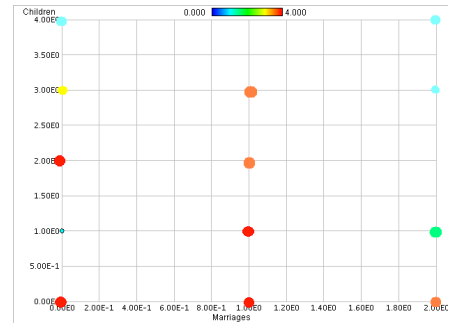

# Jitter

- This is the same data, but with small random amounts added to each value
- Notice how the distribution of points is revealed



# Colour as Frequency

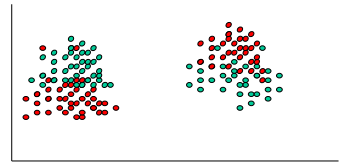
- By using a colour scale (red, orange, yellow, blue in this example), the number of times a data point is represented may be shown
- Size can also be used in place of colour





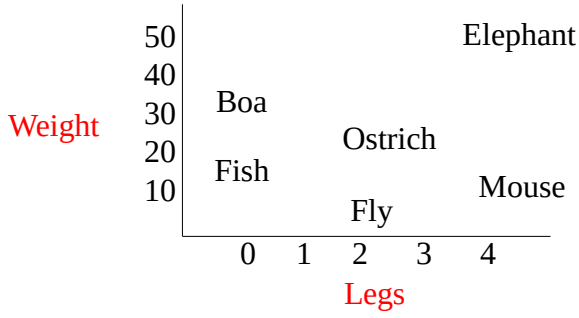

# Problems With Dimensions

- Plotting two things against each other is fine
- But what about looking at 3,4 5 or more variables?
- We have already seen one way of adding a third dimension – colour.



# Colour or Shape As a Dimension

- Category values can have their own colour or shape, or even a word or picture:




# Projection

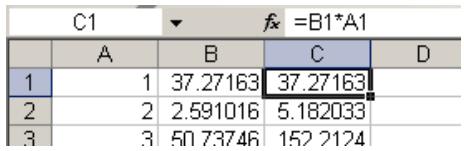
- If your data comes from a system that has more dimensions than you can plot, you will probably suffer problems with projection
- Imagine a cloud of moths flying in front of a projector. They occupy 3D space, but the shadow they project onto a wall is in 2D
- The third dimension (distance from the wall) is lost

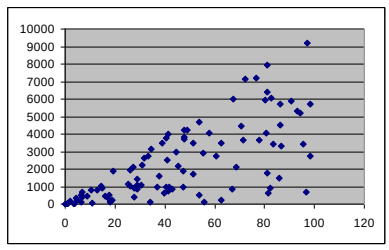
# Projection

- The same happens with plotting data
- Plotting data in fewer dimensions than it contains means that you see the ‘shadow’ of higher dimensions
- That spoils your plot



# Example

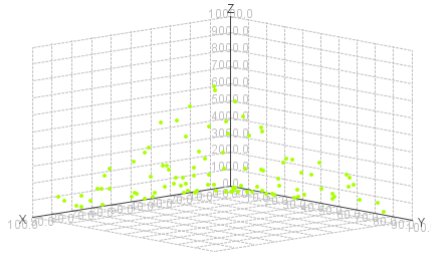


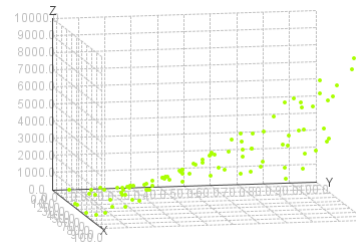

Column C is determined by A and B, but plotting B against C suggests only a weak relationship

If your plot could show A and B against C, the true shape of the relationship would appear.

# The Same Data in 3D




Software that can rotate 3D views helps you see that extra dimension



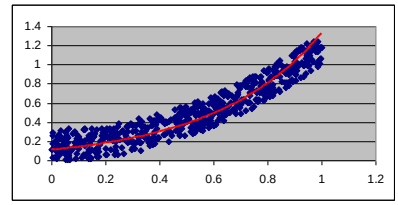

http://www.math.uri.edu/~bkaskosz/flashmo/graph3d/
where $x \in (0,1000)$, $y \in (0,10)$

# Solving Projection Problems

- Represent all the dimensions in some way
  - Colour, shape etc. as we have seen
  - Size to show the third dimension – larger things being closer
- Software that is able to rotate any 'fly' through data, switching dimensions to allow you to search for patterns
- Reduce the dimensionality

# Dimensionality Reduction

- If two or more dimensions are related, they can be reduced to a single, new dimension without loosing too much information
- This new single dimension can be plotted against others to allow deeper relationships to be found
- Always a loss of information

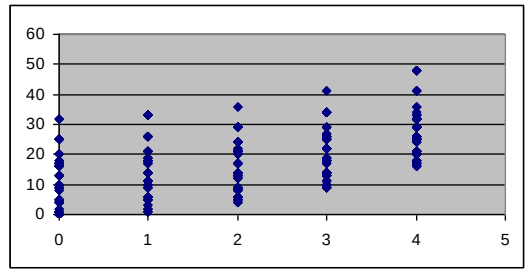



# Dimensionality Reduction

- Example techniques are:
  - Principal components reduction
  - Non-linear principal components reduction
  - Auto-associative neural networks
- Disadvantage is that the new dimensions are combinations of the original ones and might not make as much sense

# Keep Some Constant

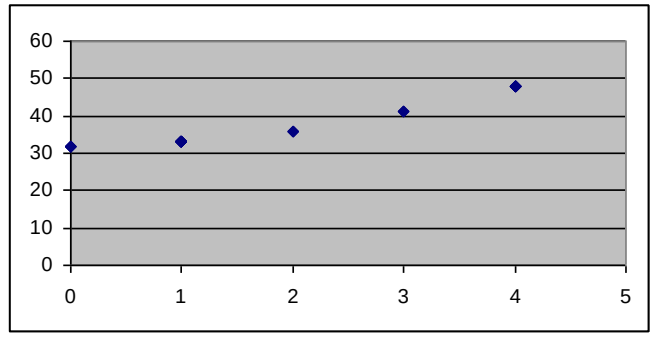
Here is an example with 3 inputs – *a*, *b*, *c*
and one output – *d*, which is affected by all 3 inputs




Here is a plot of input *c* against output *d*. The other variables are projected down onto the chart to show a mess of values



# Keep Some Constant




Now we keep *a* and *b* constant and just plot *c* against *d*.
In other words, we choose a combination of *a* and *b* that appear several times and plot *a* and *b* for just those points.

# Visualising Data for Users

- Scientific charts might not always be the best way to represent data to users or to the press
- Other visualisations can be more appropriate in the right setting



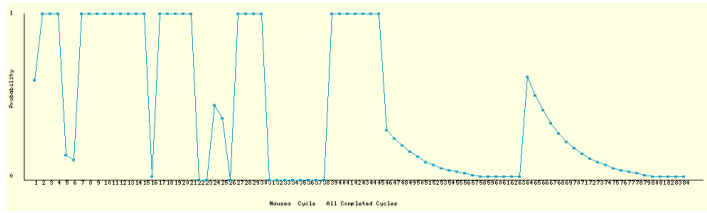
# Infographics

- Methods for displaying summaries of data in an attractive way
- Less of an analysis tool
- More of presentation tool
- Static or interactive

# Recent Example

- The project is to build a system that predicts the side effects that chemotherapy patients are likely to suffer on a daily basis throughout their treatment
- Here is a traditional time-series plot of a set of predictions:

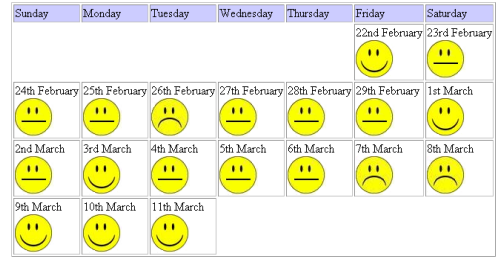
Probability of Nausea Over Time



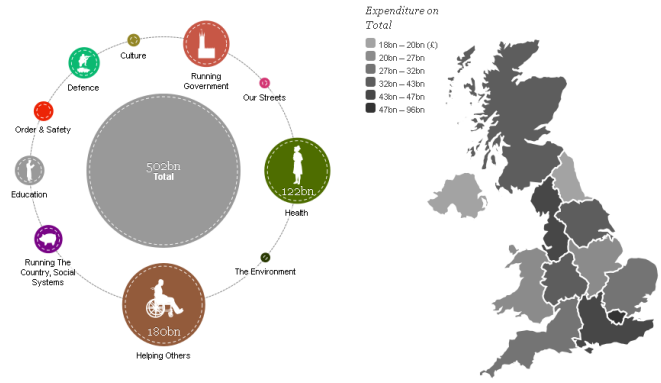
- Easy enough for us to understand – the higher the line, the larger the risk of suffering from the symptom, i.e. nausea.

# Recent Example

- For people who are not used to looking at charts, there might be a better way of presenting the same information
- In this example, we tried to use the familiar concept of a diary to present the same data
- Looks less like a scientific chart, but makes it much easier to see that planning a weekend away over the 7th and 8th of March might not be the best time to choose.



# WhereDoesMyMoneyGo




www.wheredoesmymoneygo.org/bubbletree-map.html#/~/total



# Hans Rosling's Famous Video

- It combine enormous quantities of public data to reveal the story of the world's past, present and future development.
  - https://www.youtube.com/watch?v=jbkSRLYSojo
- How many dimensions of the data are used?
  - Income (x) , life span (y)
  - population (circle size), country region (circle colour)
  - time, country name