University of BRISTOL

Visual Analytics: Arranging Tables

Ian Nabney

ian.nabney@bristol.ac.uk

bristol.ac.uk

University of BRISTOL

Overview

- Reading: Chapter 7 of Munzner
- Understand the four ways of arranging table data
- Understand the three ways axes can be oriented
- Able to select appropriate methods for displaying table data for a particular task

Why spatial arrangement first?

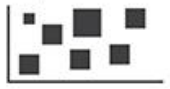
- The **arrange** design choice covers all aspects of the use of spatial channels for visual encoding
- It is the most crucial visual encoding choice because the use of space dominates the user's mental model of the dataset
- The three highest ranked effectiveness channels for quantitative and ordered attributes are all related to spatial position: planar position against a common scale, planar position along an unaligned scale, and length
- The highest ranked effectiveness channel for categorical attributes, grouping items within the same region, is also about the use of space
- There are no nonspatial channels that are highly effective for all attribute types: they are either suitable for ordered or categorical attributes, but not both, because of the principle of expressiveness.

The big picture

Arrange Tables

- **Express Values**

- **Separate, Order, Align Regions**
  - Separate
  - Order
  - Align

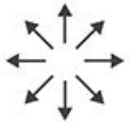


  - 1 Key — List
  - 2 Keys — Matrix
  - 3 Keys — Volume
  - Many Keys — Recursive Subdivision

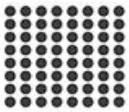



- **Axis Orientation**
  - Rectilinear
  - Parallel
  - Radial





- **Layout Density**
  - Dense
  - Space-Filling




Keys and values

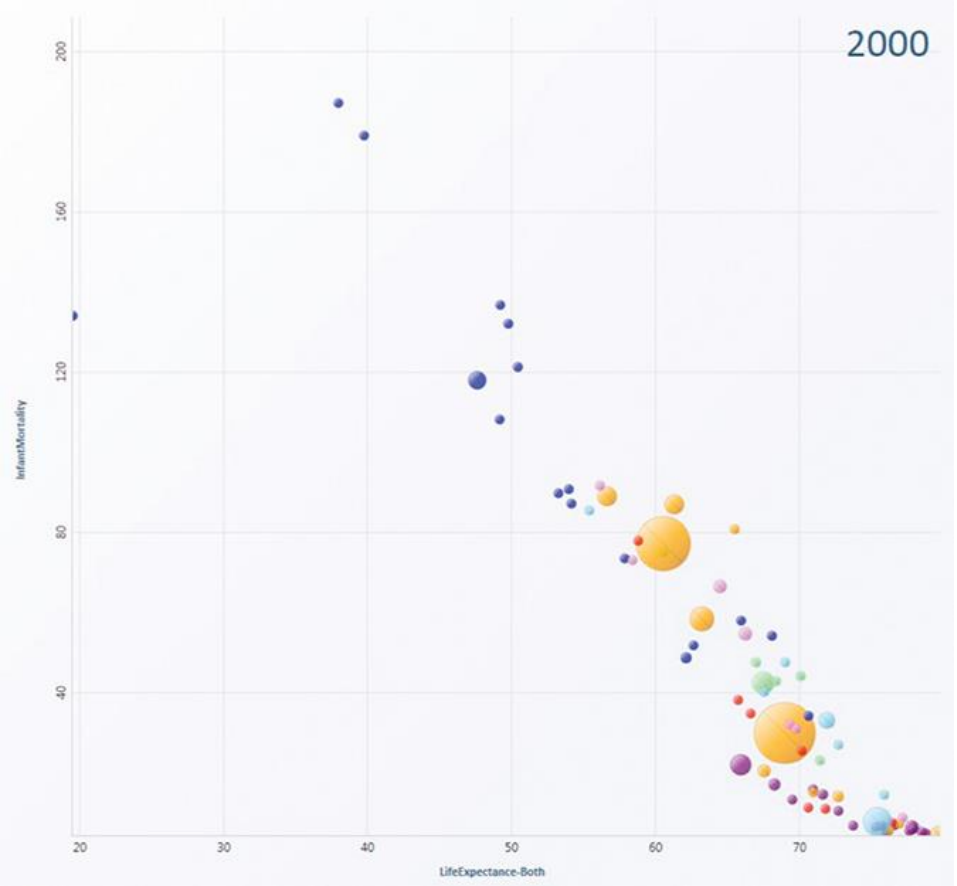
- A **key** is an independent attribute that can be used as a unique index to look up items in a table, while a **value** is a dependent attribute: the value of a cell in a table.
  - Key attributes can be categorical or ordinal
  - Values can be all three types: categorical, ordinal, or quantitative.
- The unique values for a categorical or ordered attribute may be called **levels**, to avoid the confusion of overloading the term value.

Keys and values 2

- The core design choices for visually encoding tables directly relate to the semantics of the table's attributes: how many keys and how many values does it have?
  - An idiom could only show values, with no keys; **scatterplots** are the canonical example of showing two value attributes.
  - An idiom could show one key and one value attribute; **bar charts** are the best-known example.
  - An idiom could show two keys and one value; for example, **heatmaps**.
  - Idioms that show many keys and many values often recursively subdivide space into many regions, as with **scatterplot matrices**.
- Keys are typically used to define a region of space for each item in which one or more value attributes are shown.

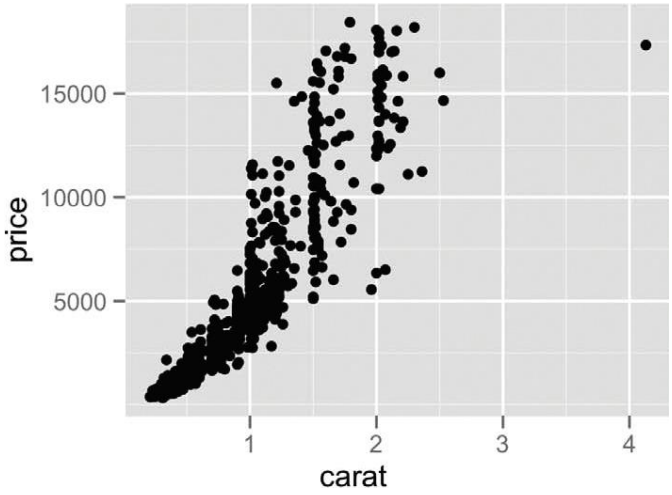
Expressing quantitative values

- The attribute is mapped to spatial position along an axis
- A single value is encoded with a mark at some position along the axis
- Additional attributes can be encoded on the same mark with other nonspatial channels such as colour and size
- Hans Rosling was a great exponent of (dynamic) scatterplots with additional attributes
- https://www.gapminder.org/downloads/updated-gapminder-world-poster-2019/

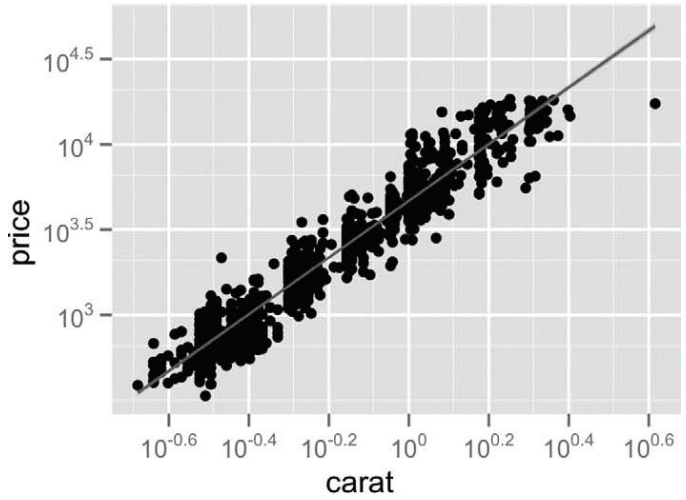


Scatterplot tasks

- What are scatterplots good for?
  - providing overviews and characterizing distributions
  - for finding outliers and extreme values
  - judging the correlation between two attributes
  - locate clusters
  - additional transformations can be used to shed more light on the data



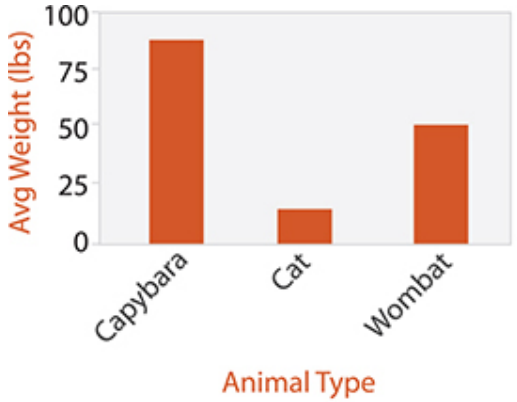
(a) Original diamond price/carat data.



(b) Log-scale data is more linearly correlated.

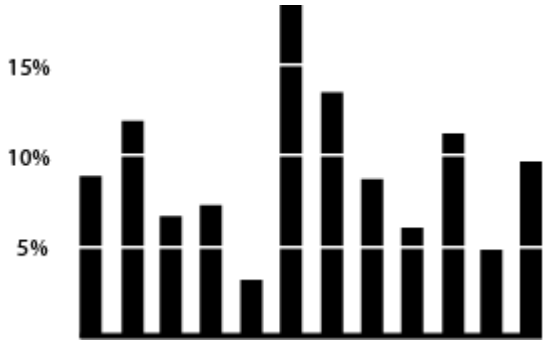
Categorical regions: bar charts

- Using space to encode categorical variables is harder than for quantitative variables because 1-d position is ordered, but categorical variables are not
- The semantics of categorical attributes does match up well with the idea of **spatial regions**: distinct contiguous bounded areas
- Drawing them can be broken down into three operations:
  - separating into regions,
  - aligning the regions (optional),
  - ordering the regions
- Separation should be done on a categorical attribute
- Alignment and ordering should be done on a different attribute that is ordered
- Bar charts are simple example: region is the bar



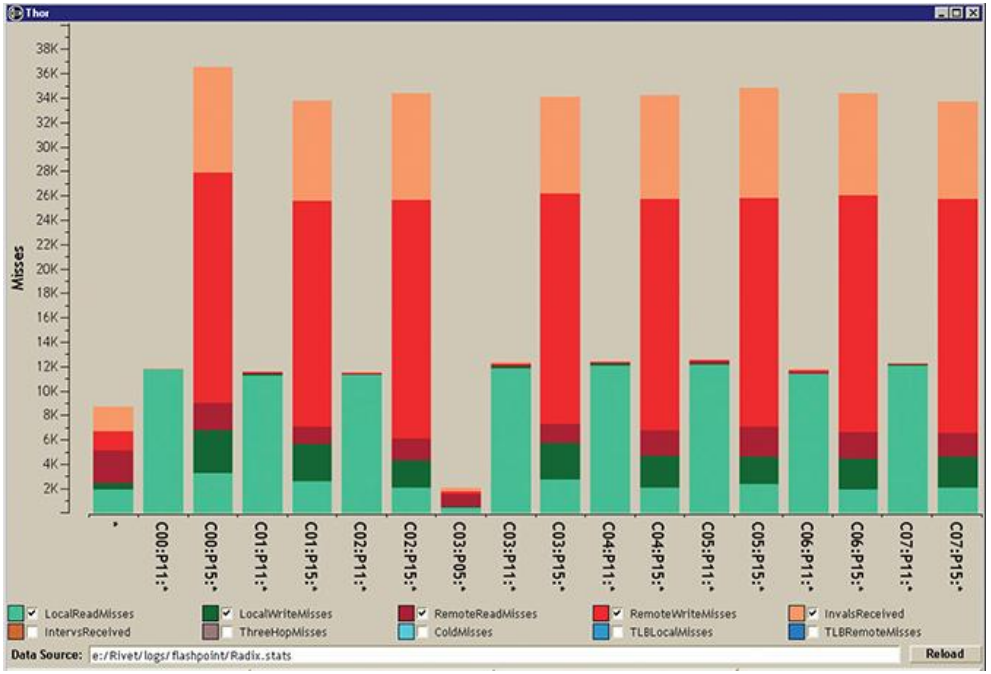
Tufte's redesign of the bar chart

- Tufte has a theory (almost an obsession) of data-ink (i.e. ration of data to ink) maximisation
- He used this to redesign the standard bar chart, removing the frame and the axes, and creating a white grid which shows the coordinate lines more precisely than ticks on the axis
- A thin baseline shows the alignment clearly



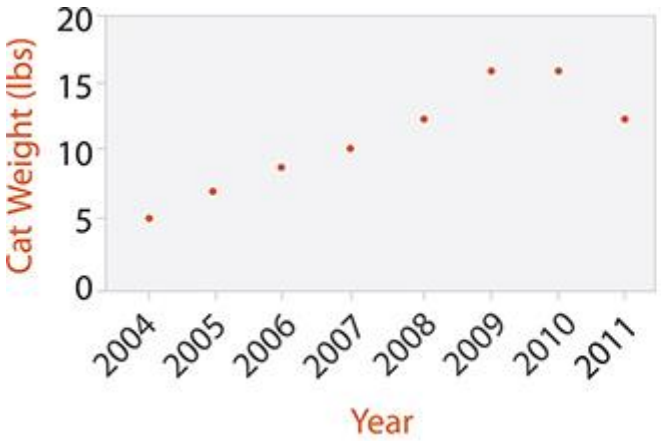
Stacked bar charts

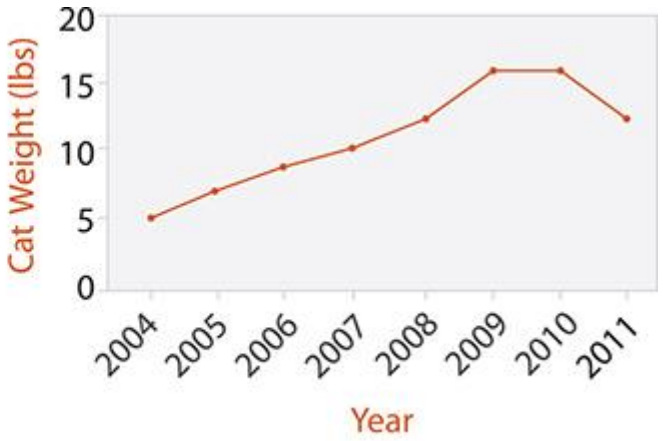
- Multiple sub-bars are stacked vertically
- Two-dimensional table with two keys
- Secondary key is used in constructing the vertical structure of the glyph
- Each subcomponent is coloured according to the same key that is used to determine the vertical ordering; since the subcomponents are all abutted end to end without a break and are the same width, they would not be distinguishable without different colouring.
- Texture can be used if colour is not available



Dot and line charts

- The **dot chart** idiom is a visual encoding of one quantitative attribute using spatial position against one categorical attribute using point marks, rather than the line marks of a bar chart
- The idiom of **line** charts augments dot charts with line connection marks running between the points.
- Line charts should be used for ordered keys but not categorical keys



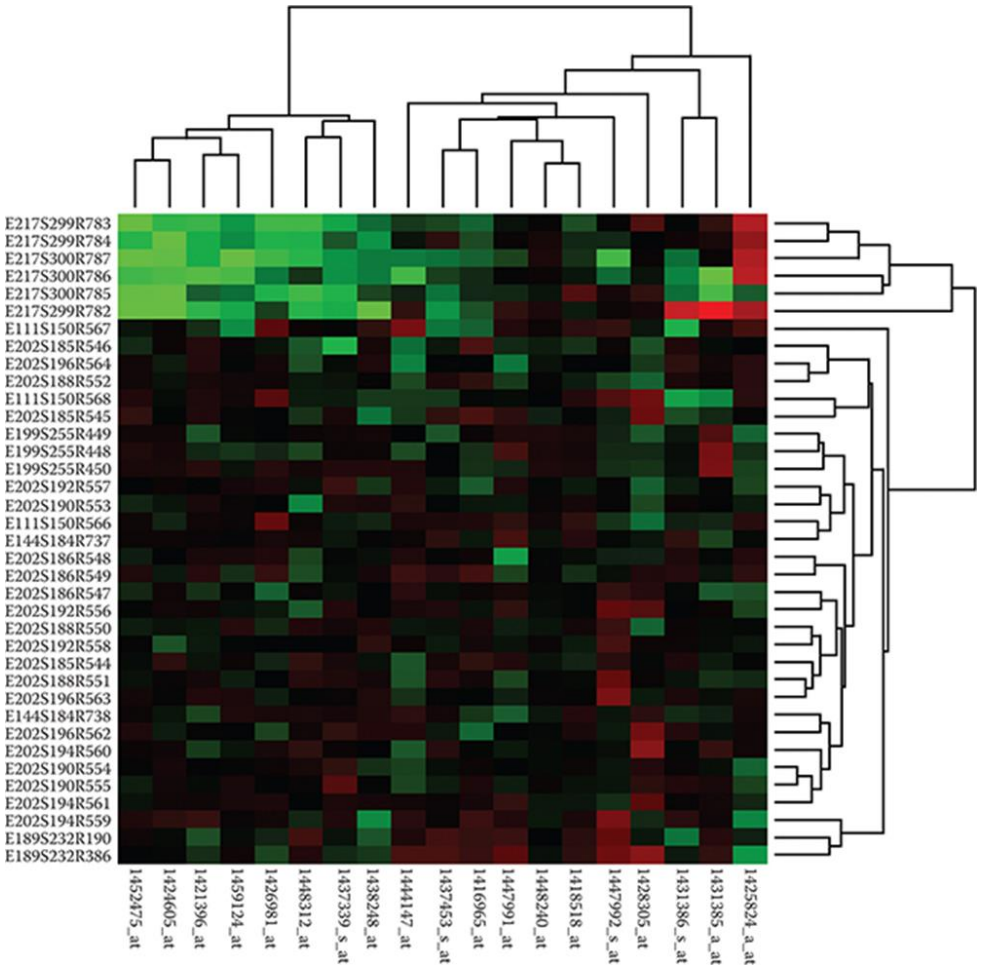


Matrix alignment: heatmap

- Datasets with two keys are often arranged in a two-dimensional matrix alignment where one key is distributed along the rows and the other along the columns
- In **heatmaps** each cell is fully occupied by an area mark encoding a single quantitative value attribute with colour
- A **scatterplot matrix** (SPLOM) is a matrix where each cell contains an entire scatterplot chart. A SPLOM shows all possible pairwise combinations of attributes, with the original attributes as the rows and columns

Cluster heatmaps

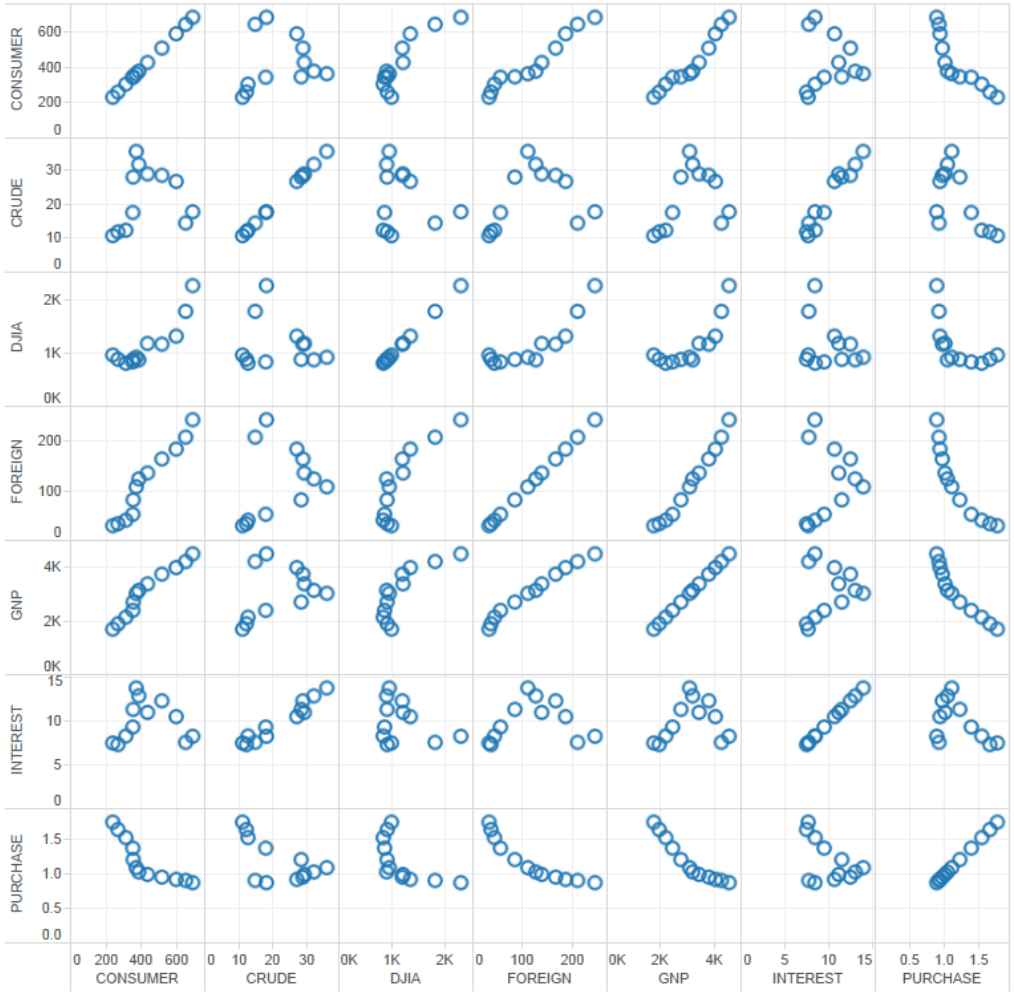
- Often used in bioinformatics
- Visually encoding quantitative data with colour using small area marks is very compact, so they provide high information density
- The area marks in a heatmap are often several pixels on a side for easy distinguishability, so a matrix of 200 x 200 with 40,000 items is easily handled
- Only a small number of different levels of the quantitative attribute can be distinguished, because of the limits on colour perception in small non-contiguous regions: between 3 and 11 bins



Scatterplot matrix

How to interpret the graphs on the diagonal from top-left to bottom-right?

Think about what attributes are in the pairs down that diagonal.

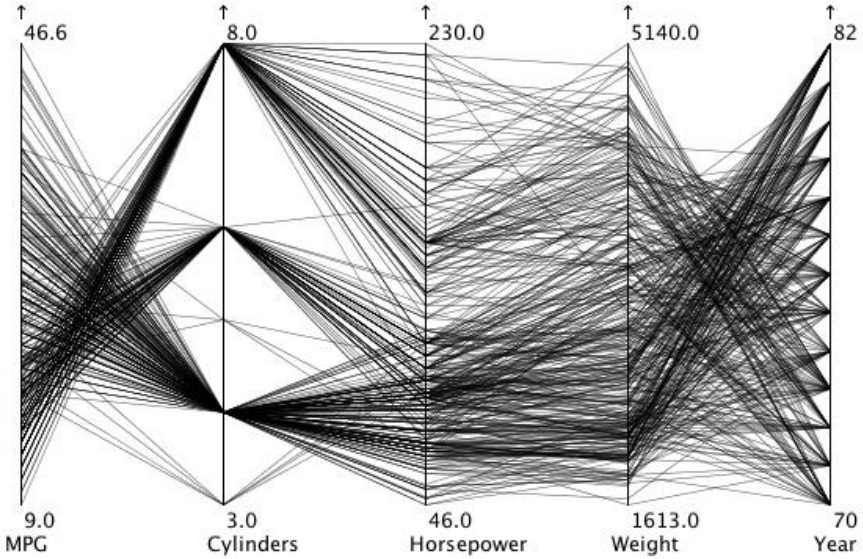


Spatial axis orientation

- Rectilinear, parallel or radial layout
- In a **rectilinear** layout, regions or items are distributed along two perpendicular axes, horizontal and vertical, that range from minimum value at one end of the axis to a maximum value at the other
- Rectilinear layouts are used in many common statistical charts
- The scatterplot is only usable for two data attributes. Even if the low-precision visual channel of a third spatial dimension is used, then only three data attributes can be shown using spatial position channels
- Although additional nonspatial channels can be used for visual encoding, the problem of channel inseparability limits the number of channels that can be combined effectively in a single view
- Of course, many tables contain far more than three quantitative attributes

Parallel axis orientation

- The idiom of **parallel coordinates** is an approach for visualizing many quantitative attributes at once using spatial position
- The axes are placed parallel to each other, rather than perpendicularly at right angles
- A single item is represented by a jagged line that zigzags through the parallel axes, crossing each axis exactly once at the location of the item's value for the associated attribute
- Interactive example https://public.tableau.com/views/ParallelCoordinates_15626823354580/NBASTATS?:embed=y&:display_count=yes&publish=yes&:origin=viz_share_link&:showVizHome=no



https://eagereyes.org/techniques/parallel-coordinates

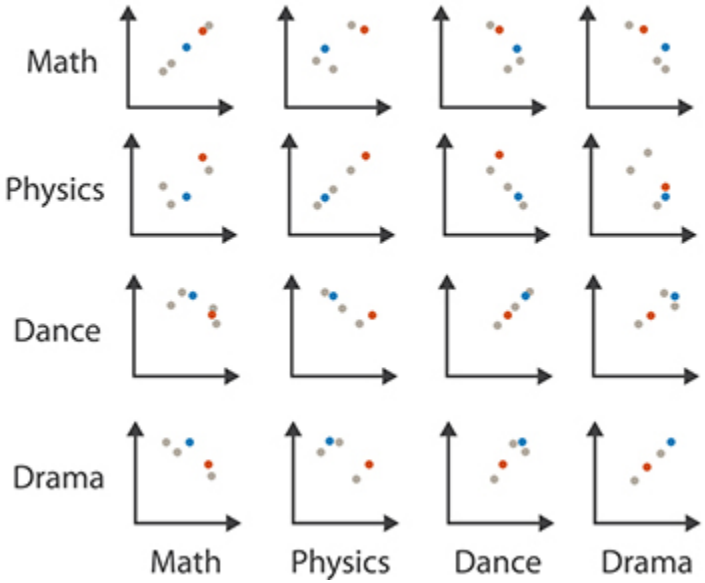
Comparison with SPLOM

- If two **neighbouring** axes have high positive correlation, the line segments are mostly parallel. If two axes have high negative correlation, the line segments mostly cross over each other at a single spot between the axes. The pattern in between uncorrelated axes is a mix of crossing angles.
- SPLOMs are better for the task of finding correlation. Parallel coordinates are more often used for other tasks, including overview over all attributes, finding the range of individual attributes, selecting a range of items, and outlier detection

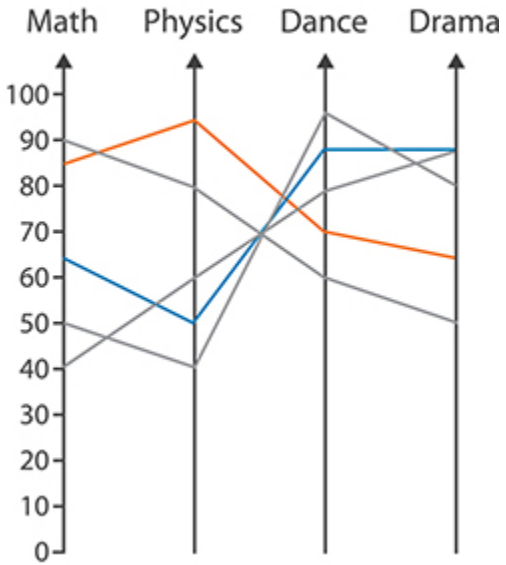
Table

Math	Physics	Dance	Drama
85	95	70	65
90	80	60	50
65	50	90	90
50	40	95	80
40	60	80	90

Scatterplot Matrix

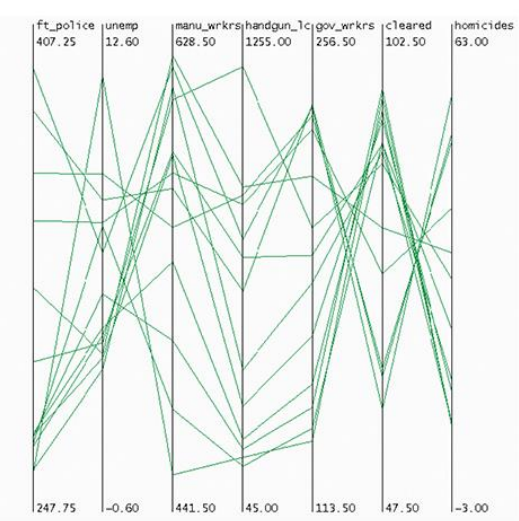


Parallel Coordinates

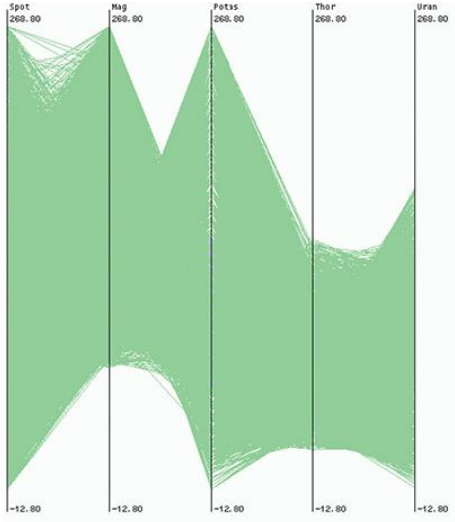


Features of parallel coordinates

- The idiom scales to showing hundreds of items, but not thousands.
- The patterns made easily visible by parallel coordinates have to do with the pairwise relationships between neighbouring axes. The crucial limitation of parallel coordinates is how to determine the order of the axes.
- Most implementations allow the user to interactively reorder the axes.



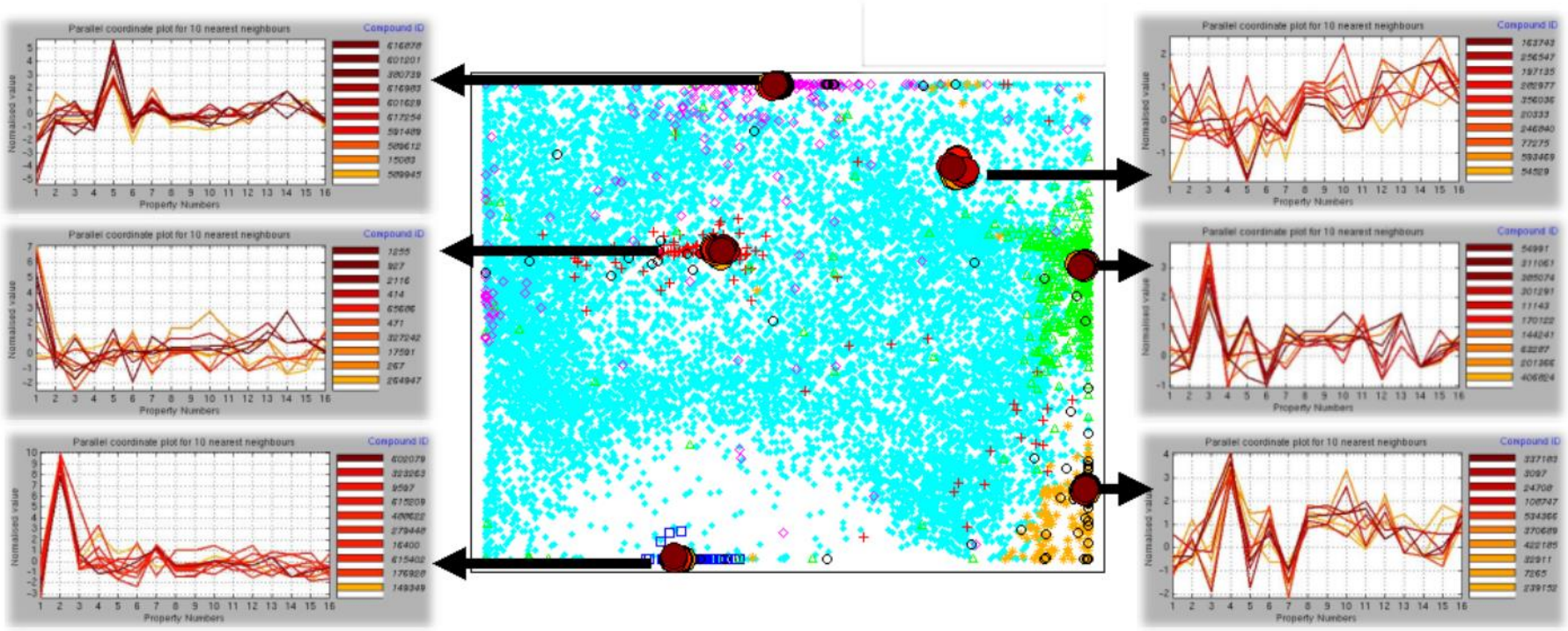
(a)



(b)

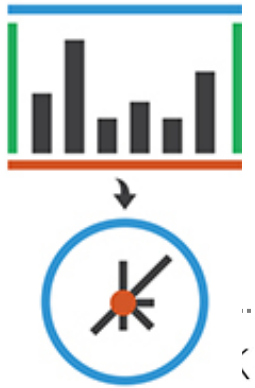
However, exploring all possible configurations of axes through systematic manual interaction would be prohibitively time consuming as the number of axes grows, because of the exploding number of possible combinations.

Local parallel coordinates



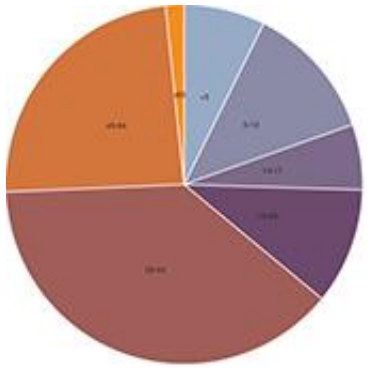
Radial layout

- In a radial spatial layout, items are distributed around a circle using the angle channel in addition to linear spatial channels
- The natural coordinate system in radial layouts is polar coordinates, where one dimension is measured as an angle from a starting line and the other is measured as a distance from the centre
- From a perceptual point of view, two major drawbacks
  - the angle channel is less accurately perceived than a rectilinear spatial position channel
  - the angle channel is inherently cyclic, because the start and end point are the same, as opposed to the inherently linear nature of a position channel
- Best reserved for genuinely cyclic data (expressiveness rule)

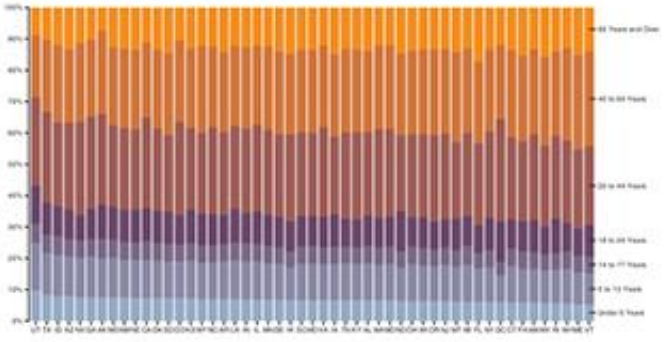


Pie charts

- Don’t use them!
- Accuracy of angle judgement is low – often augmented by writing numbers on the slices
- They do show relative contribution of parts to a whole, but this can be achieved by normalised stacked bar charts showing percentages rather than counts
- One pie chart shows the aggregate results of 50 stacked bar charts in a similar sized space



(a)



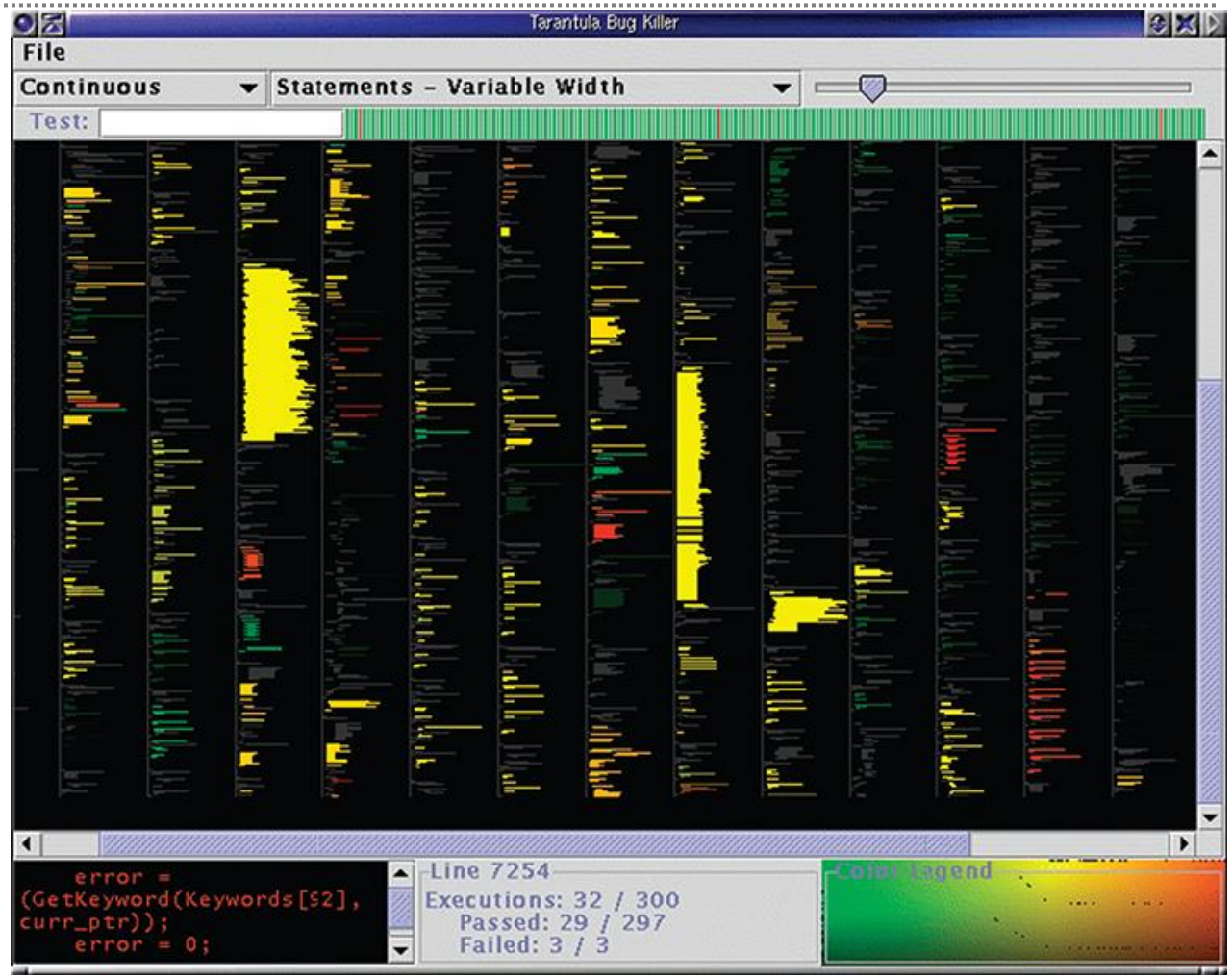
(b)

Spatial layout density

- A **dense** layout uses small and densely packed marks to provide an overview of as many items as possible with very high information density
- The small size of the marks implies that only the planar position and colour channels can be used in visual encoding; size and shape are not available
- A **space-filling** layout has the property that it fills all available space
  - Maximize the amount of room available for colour coding, increasing the chance that the coloured region will be large enough to be perceptually salient to the viewer
  - designer cannot make use of white space in the layout (valuable for readability, emphasis, relative importance, and visual balance)

Case study: Tarantula

- Visualising test coverage
- Overview of source code using one-pixel tall lines, colour coded to show whether it passed, failed, or had mixed test results
- Brightness encodes the percentage of coverage by the test cases, where dark lines represent low coverage and bright ones are high coverage
- Hue encodes the relative percentage of passed versus failed tests



Summary

- Respect expressiveness principle
- More complex chart types can represent more than two attributes at a time
- Layout of axes can help with different tasks
- Don’t use pie charts!