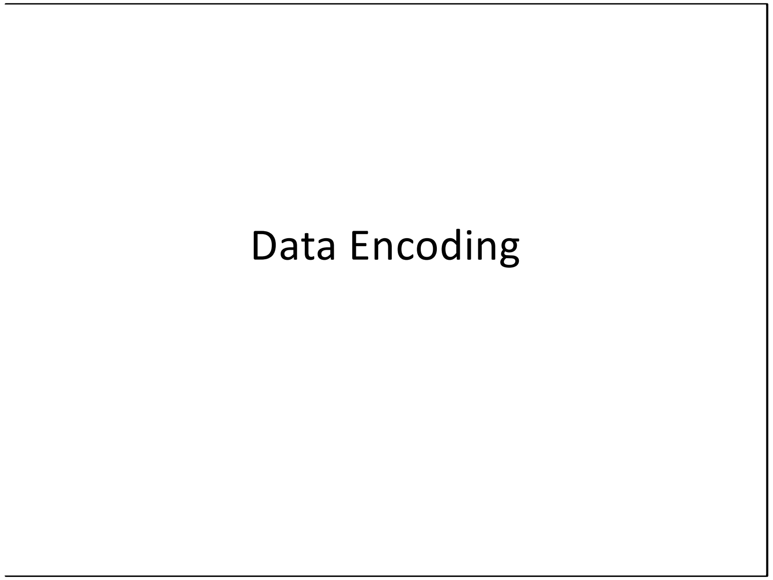
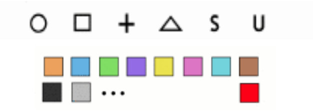


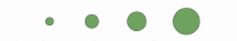
Data Encoding

Data attributes

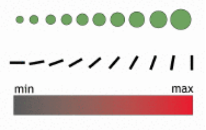
- categorical (or ‘nominal’)
  - apples, pears, bananas
  - sparrow, goldfinch, robin



- ordered (ordinal)
  - small, medium, large, XL



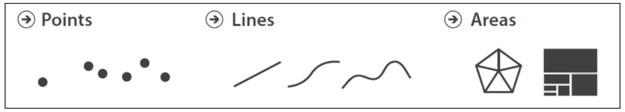
- ordered (quantitative)
  - 10cm, 17cm, 23cm



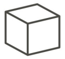
Stolte, C, and Hanrahan, P. Polaris. IEEE TVCG, 2002 (figures)

Data Marks

- The basic graphical element of an image
- Can be 0D, 1D, 2D, 3D

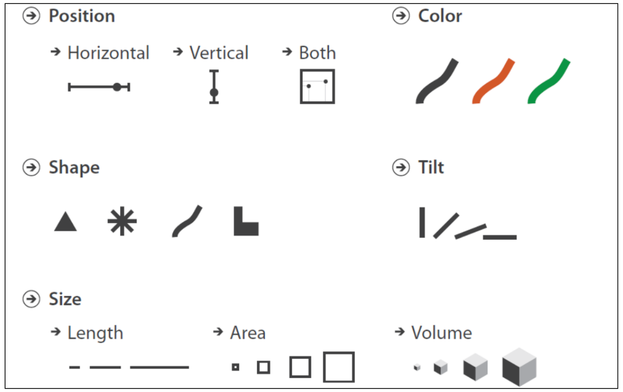


Volumes



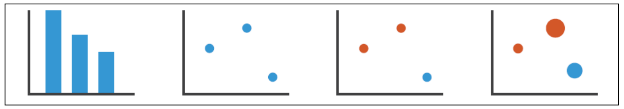
Visualisation Analysis & Design, T. Munzner, (Ch4)

Munzner calls Bertin's visual variables 'channels'



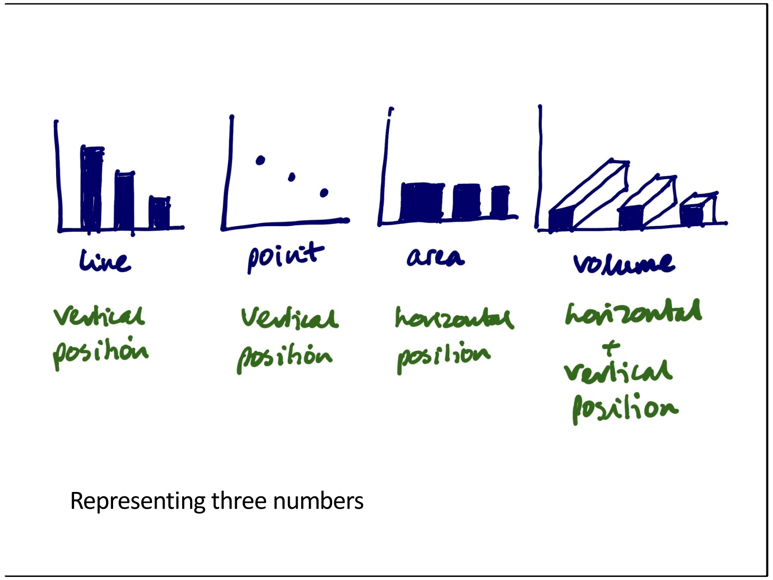
Visualisation Analysis & Design, T. Munzner, (Ch4)

Marks and Channels



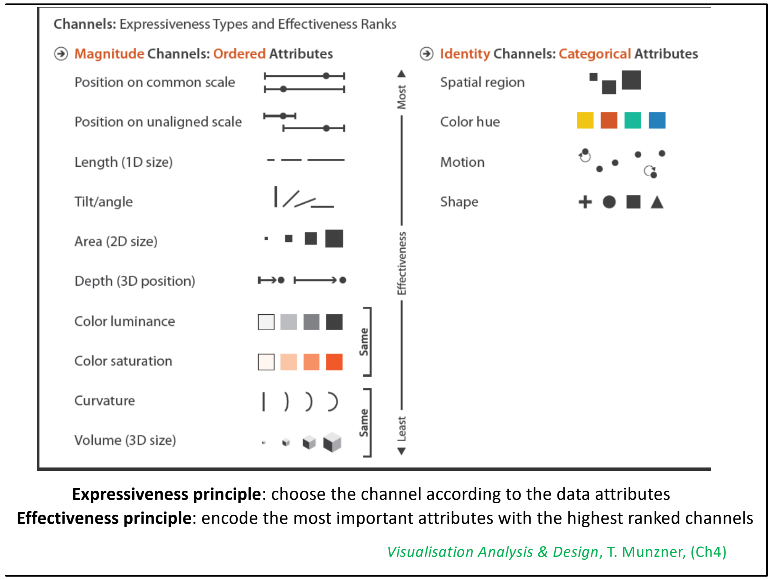
channel(s) /variables **used for showing the data**	vertical position	vertical position horizontal position	vertical position horizontal position color hue	vertical position horizontal position color hue size (area)
mark	line/ area	point	point	point

Visualisation Analysis & Design, T. Munzner, (Ch4)



Representing three numbers

There are clearly many choices with regard to how to encode data in a visualization. We need ways to prioritise the most important data dimensions, in encoding them with the most clearly perceived channels



Expressiveness principle: choose the channel according to the data attributes
Effectiveness principle: encode the most important attributes with the highest ranked channels

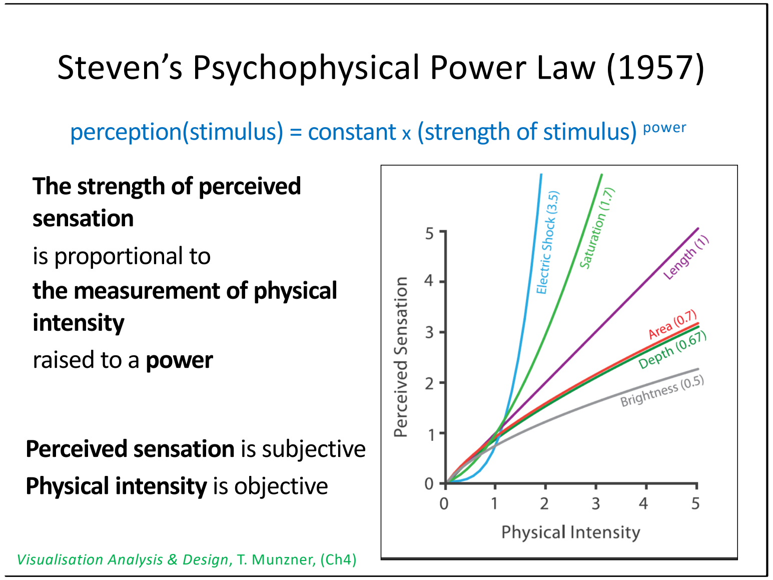
Visualisation Analysis & Design, T. Munzner, (Ch4)

There have been many studies, looking at how to prioritise such variables and encondings. A summary from the Munzner book is here.

The first thing to do is to consider the type of data – the attributes of the data. Is it categorical, or ordered in some way? Don't choose a channel that is misleading, or likely to create false interpretations. Choose a channel that can express the values in the data in ways that are accurate and clear. This is the 'expressiveness principle'.

The second thing is to encode the most important attributes (or data dimensions) with the channels that are highest ranked in terms of perception. The importance of attributes depends on the application context, for example the meaning of the data and the questions your visualisation will help users answer.

colour luminance = Bertin's value (shade and tint)
colour saturation was not represented/discussed by Bertin (but was added by Morrison)



Steven's Psychophysical Power Law (1957)

perception(stimulus) = constant x (strength of stimulus) $^{\text{power}}$

The strength of perceived sensation

is proportional to

the measurement of physical intensity

raised to a **power**

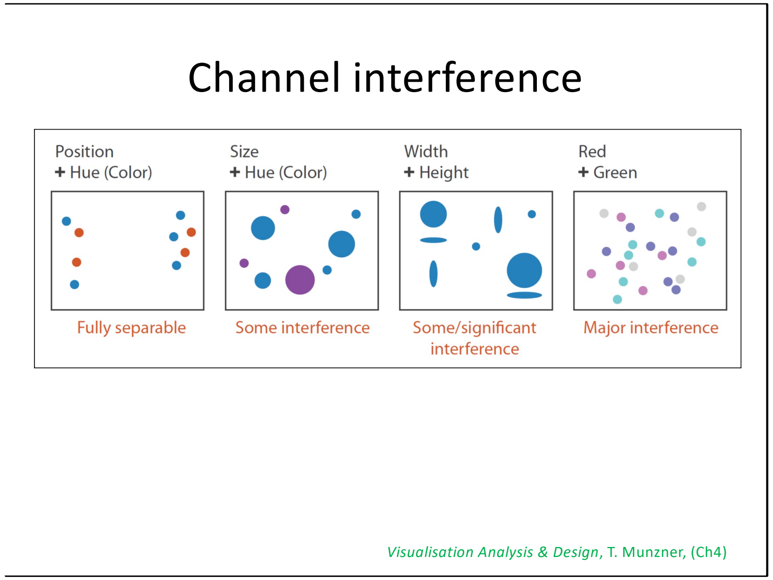
Perceived sensation is subjective

Physical intensity is objective

Visualisation Analysis & Design, T. Munzner, (Ch4)

The key thing to understand here is that there is a systematic bias in many forms of encoding of magnitudes… except length/distance. This is why bar charts (and similar charts) work so well, and bubble charts, pie charts and the like (that use 2D) are not so good… and anything 3D is even worse.

Some media tested are thankfully rare, like electric shock, but area of a shape (e.g. size of a bubble in a bubble chart), and saturation and brightness of colour, are often used… but have inherent problems. They are therefore ranked lower in the tables like the one in the previous slide.



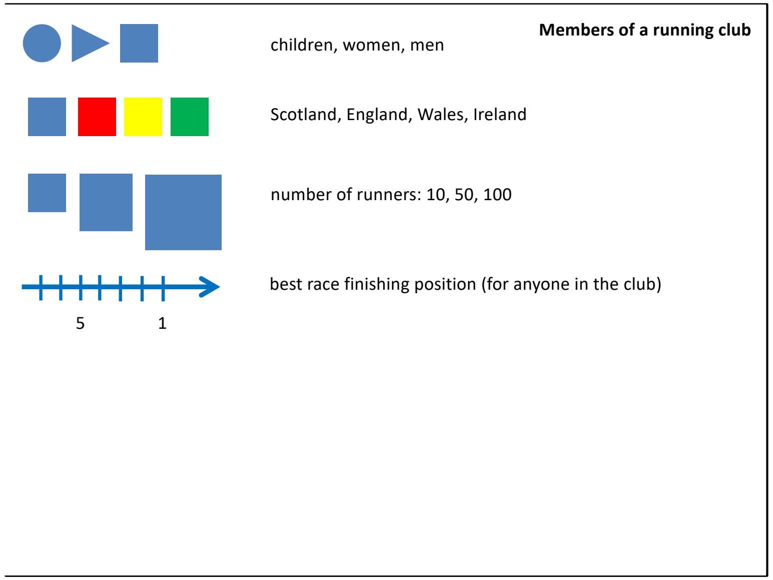
Channel interference

Visualisation Analysis & Design, T. Munzner, (Ch4)

Pairs of visual channels fall along a continuum from fully separable to intrinsically integral. Color and location are separable channels well suited to encode different data attributes for two different groupings that can be selectively attended to. However, size interacts with hue – hue is harder to perceive for small objects.

The horizontal size and and vertical size channels are automatically fused into an integrated perception of area, yielding three groups (when they do not really exist in the data here).

Attempts to code separate information along the red and green axes of the RGB color space fail, because we simply perceive four different hues… we tend to see categories instead of ordinal/quantitative data.



Members of a running club

children, women, men

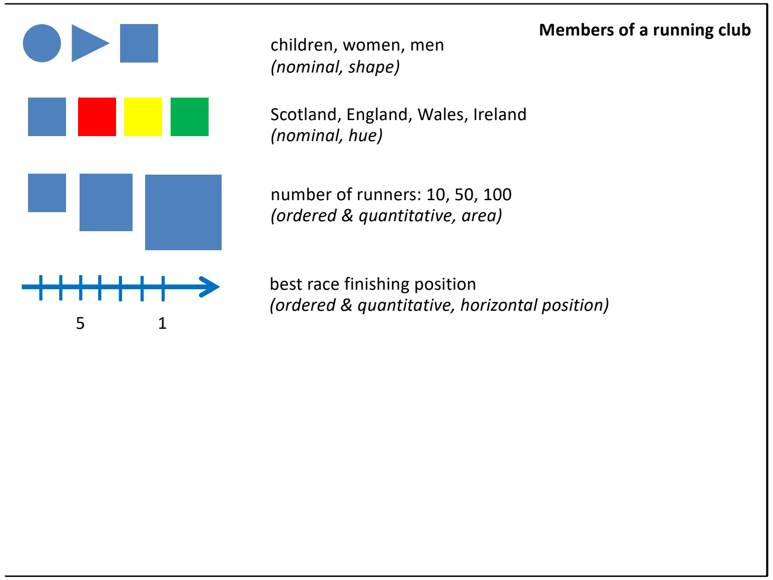
Scotland, England, Wales, Ireland

number of runners: 10, 50, 100

best race finishing position (for anyone in the club)

Would this be a good set of choices, for encoding data in a visualization?

Let’s assume that the best race finishing position is the most important attribute for the current task. Perhaps the number of runners is 2nd



Members of a running club

children, women, men
(nominal, shape)

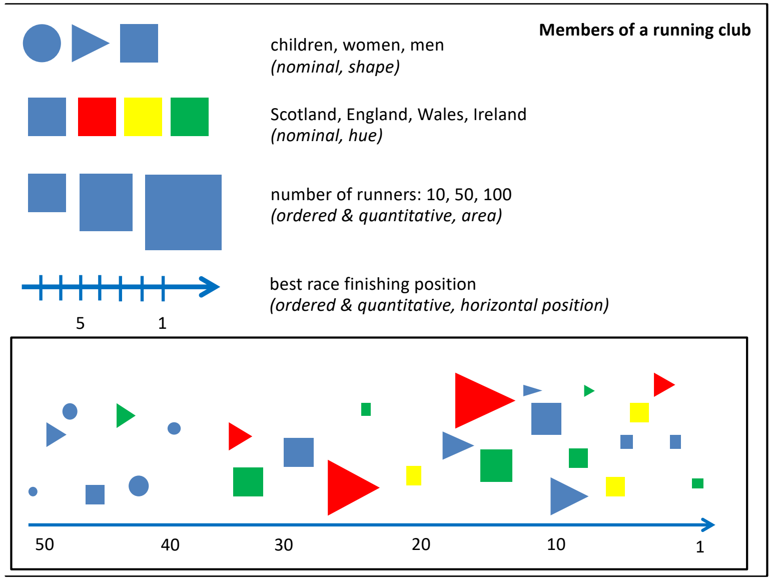
Scotland, England, Wales, Ireland
(nominal, hue)

number of runners: 10, 50, 100
(ordered & quantitative, area)

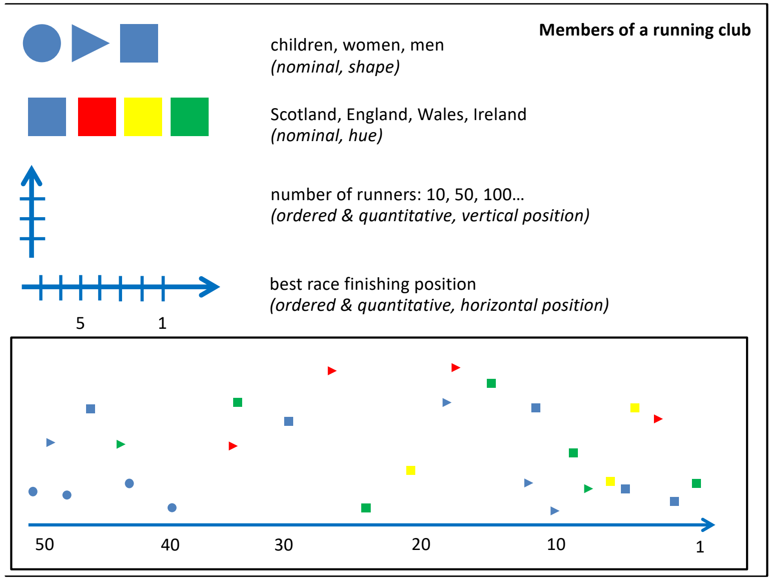
best race finishing position
(ordered & quantitative, horizontal position)

This would be a valid encoding… but not a very good one. There are some problems. Area is not a good visual variable (or channel) to use, because of bias. Also, here, it interferes with the shape variable quite a bit, I think… as the area of a square with side sized X is larger than the triangle of edge X or circle of diameter X.

Let’s look at a rough sketch of using this



One thing that stands out for me, is how vertical position in this chart *seems* to mean something… but *actually* doesn't mean anything, in this encoding. The eye is drawn most to the big shapes – the big clubs -- but we know that there is a bias there.



Members of a running club

children, women, men
(nominal, shape)

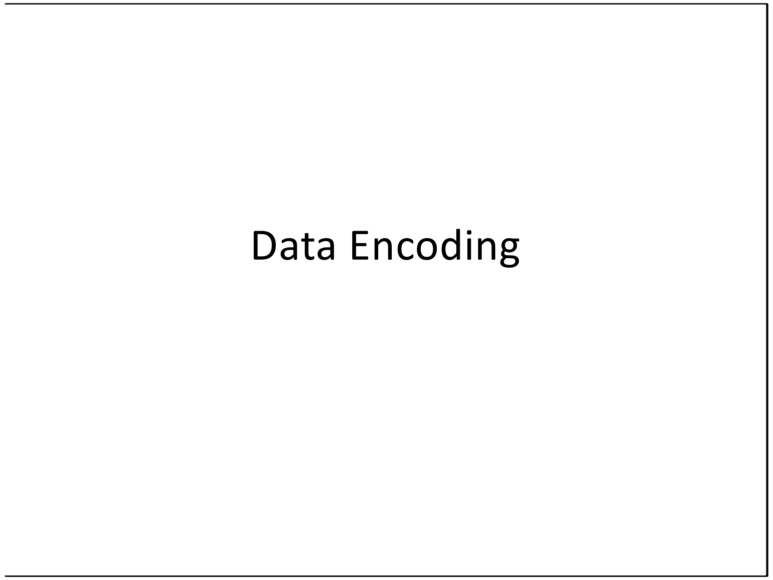
Scotland, England, Wales, Ireland
(nominal, hue)

number of runners: 10, 50, 100…
(ordered & quantitative, vertical position)

best race finishing position
(ordered & quantitative, horizontal position)

Perhaps this is a better encoding. I’ve tried to map the size in the last chart, into position here, and made all the symbols the same size.

Much like a simple scatterplot, we can use vertical position to encode an important attribute… so we can see clubs with many runners higher up the chart, and smaller clubs down low on the chart. If we are looking for patterns in this data now, I would expect this encoding to be more reliable.



Data Encoding