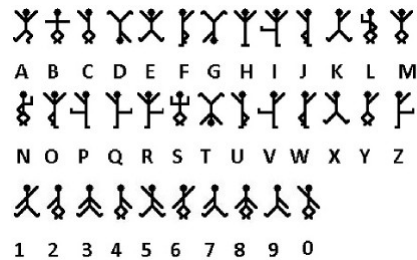


Encoding data

This document outlines how to encode numbers as visual elements and describes marks and attributes as used in graphs. An example is given showing how to identify the data encoded and the marks and attributes used in a graph.

"The idea is to go from numbers to information to understanding." (Hans Rosling)

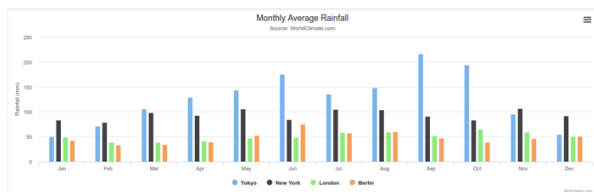


Did you ever play spies as a kid? Turning your secret messages into code? A graph *encodes* data. That is, it represents the information in a different way. While you might have used letters, numbers or stick figures to encode your secret messages, graphs use marks and attributes. In this article we will look at the different types of *marks* and their *attributes* and match them to the data that they can be used to encode.

2 100 39 91 93 98 94 89 30 82
name, age, id, colour, language



encoded



Some content in this article comes from section 6.1 of Andy Kirk, Visualising Data.

Marks

There are four variations of marks: point, line, area and form each of which can capture data by variations of different attributes.

Point



The point mark has no variation in the spatial dimension and is mostly used to represent quantitative data values either through position on a scale or the quantity of points used.

Example:

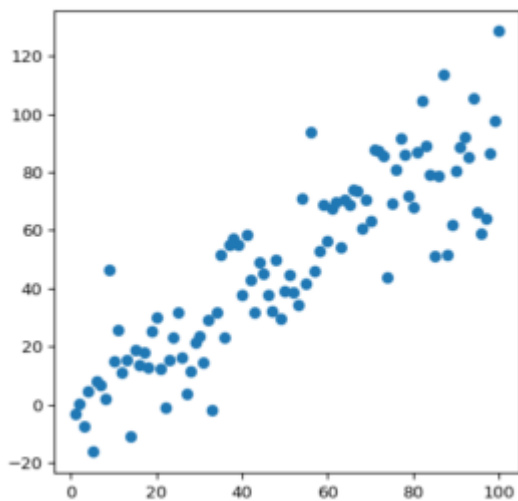


Fig: Basic matplotlib scatterplot (from: <https://python-graph-gallery.com/scatter-plot/>)

Note that while a point technically has no area (no spatial dimension) it still requires some to be visible. This area can be coloured to indicate categorical values but if the size is not connected to a quantitative value (ie, there's no variation in point size) then you can still consider the mark to be a point.

Line



The line mark has one spatial dimension and is commonly used to represent a quantitative value through variation in size or to show trends through variation in angle.

Example:

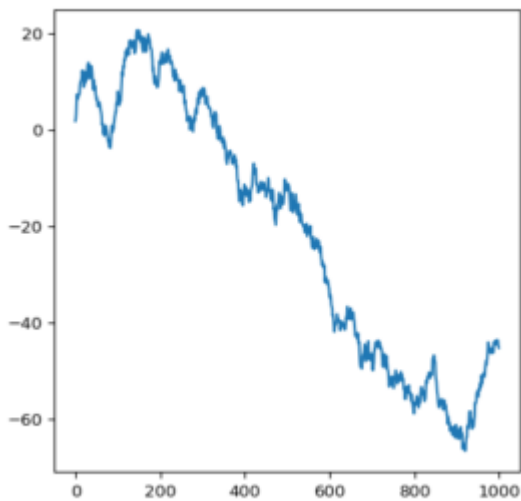


Fig: basic matplotlib line chart (from: <https://python-graph-gallery.com/line-chart/>)

Similarly to a point, a line technically has no width (a single spatial dimension - length) but some is required for it to be visible and this can also be coloured or dashed to indicate different categories. A line is also used to encode data in Bar or Column charts.

Area



The area mark has two spatial dimensions (typically width and height or radius) and can represent quantitative values by variations in both position and size. The texture, colour or shape can also be varied to represent categories.

Example:

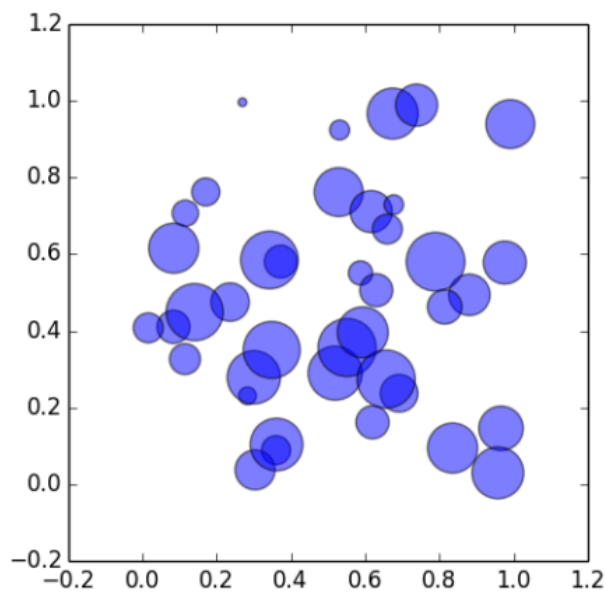


Fig: a simple bubble plot made using matplotlib (from: <https://python-graph-gallery.com/270-basic-bubble-plot/>)

Form



The form mark has three spatial dimensions and can represent quantitative values through variations in size (volume).

Example:

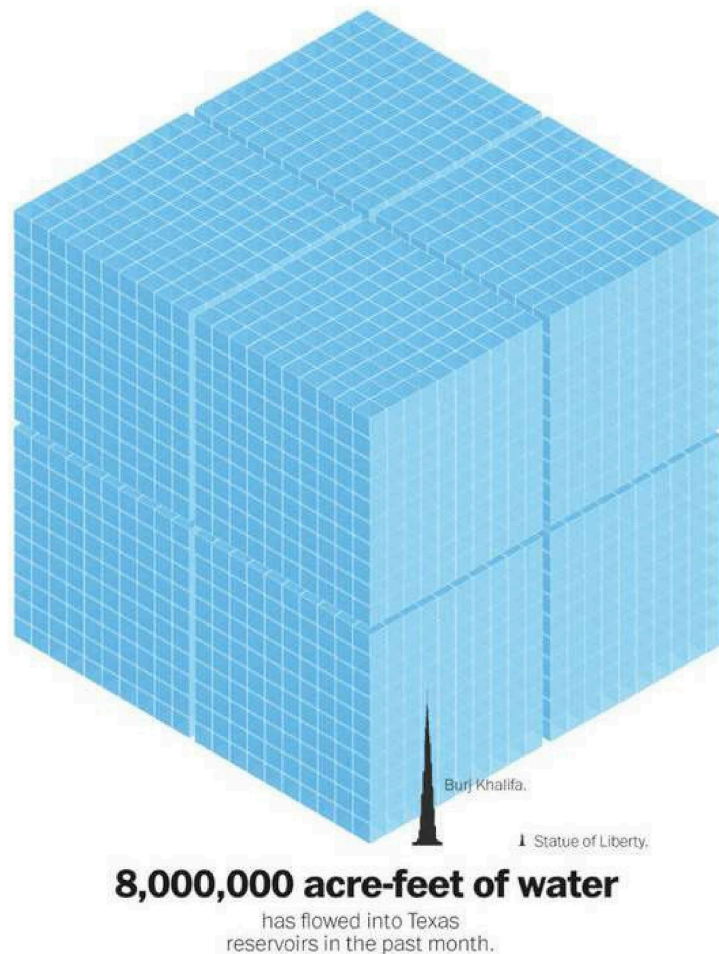


Fig: “How the Insane Amount of Rain in Texas Could Turn Rhode Island Into a Lake” from Chapter 6 of Kirk (2016) (Originally from the Washington Post)

Note: Form is rarely used and it was tough to find an example! This is from Kirk’s book. Remember to be wary of using 3D in a 2D visualisation. These work best in an interactive graph where the marks can be rotated and examined (especially if you have a VR interface).

Attributes

Now that we have the raw material to encode our data -- points, lines and area (and occasionally forms), let’s look at what attributes we can change to capture the data values. These can be divided into three classes: quantitative, categorical and relational. You should recognise these names from looking at data types. There is overlap between the attributes used in these different classes and it is possible to encode a categorical value using something like position or size.

The table below (from Kirk, 2016, p161) shows the attributes for each class with an example of how it might appear and a description.


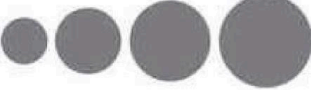
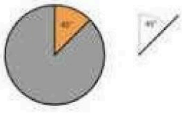



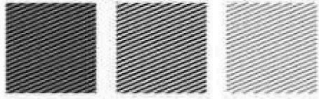

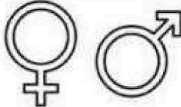


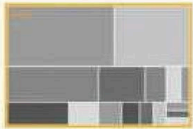
ATTRIBUTE	EXAMPLE	DESCRIPTION
QUANTITATIVE ATTRIBUTES		
Position		Position along a scale is used to indicate a quantitative value.
Size		Size (length, area, volume) is used to represent quantitative values based on proportional scales where the larger the size of the mark, the larger the quantity.
Angle/Slope		Variation in the size of angle forms the basis of pie chart sectors representing parts-of-a-whole quantitative values; the larger the angle, the larger the proportion. The slope of an incline formed by angle variation can also be used to encode values.
Quantity		The quantity of a repeated set of point marks can be used to represent a one-to-one or a one-to-many unit count.
Colour: Saturation		Colour saturation can be used (often in conjunction with other colour properties) to represent quantitative scales; typically, the greater the saturation, the higher the quantity.
Colour: Lightness		Colour lightness can be used (often in conjunction with other colour properties) to represent quantitative scales; typically, the darker the colour, the higher the quantity.
Pattern		Variation in pattern density or difference in pattern texture can be used to represent quantitative scales or distinguish between categorical ordinal states.
Motion		Motion is more rarely seen but it could be used as a binary indicator to draw focus (motion vs no motion) or by incorporating movement through speed and direction to represent a quantitative scale ramp.
CATEGORICAL ATTRIBUTES		
Symbol/shape		Symbols or shapes are generally used with point markers to indicate categorical association.
Colour: Hue		Colour hue is typically used for distinguishing different categorical data values but can also be used in conjunction with other colour properties to represent certain quantitative scales.
RELATIONAL ATTRIBUTES		
Connection/Edge		A connection or edge indicates a relationship between two nodes. Sometimes arrows may be added to indicate direction of relationship, but largely it is just about the presence or absence of a connection.
Containment		Containment is a way of indicating a grouping relationship between categories that belong to a related hierarchical 'parent' category.

Fig: Table showing the different attributes, an example and description. From p161 of Kirk, 2016.

Perceptual ranking

Some of these attributes are easier to identify and process than others. Later, we will look at attention and discuss how our brain prioritizes inputs. For now the following table ranks the attributes for each category by the ease with which the data values can be extracted or compared. Remember NOIR (Nominal, Ordinal, Interval, Ratio)?

<i>Qualitative</i>	<i>Qualitative</i>	<i>Quantitative</i>
Nominal	Ordinal	Interval, Ratio
Position	Position	Position
Colour (Hue)	Pattern (Density)	Size (Length)
Pattern (Texture)	Colour (Lightness)	Angle/Slope
Connection/Edge	Colour (Hue)	Size (Area)
Containment	Pattern (Texture)	Size (Volume)
Pattern (Density)	Connection/Edge	Pattern (Density)
Colour (Lightness)	Containment	Colour (Lightness)
Symbol/Shape	Size (Length)	Colour (Hue)
Size (Length)	Angle/Slope	Pattern (Texture)
Angle/Slope	Size (Area)	Connection/Edge
Size (Area)	Size (Volume)	Containment
Size (Volume)	Symbol/Shape	Symbol/Shape

As an illustration of the difference that choice of encoding attribute can make in the understandability of your graph, look at the figure below (based on Figure 6.58 in Kirk 2016). If A is 10 then can you tell me the value of B in the respective bar (line) and circle (area) displays?

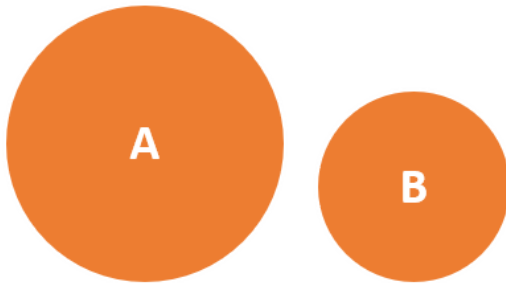






Fig: Comparison of judging line size vs area for data encoding

The answer is 5 in both cases. It's a lot easier to judge this in the line though!

The figures below summarise the main points about marks and attributes and might be useful as a reference or reminder.

Data representation: **Marks**

Point		No spatial variation	Eg. Quantity through position (scatter plot)
Line		1 spatial dimension	Eg. Quantity through variation in size (bar chart)
Area		2 spatial dimensions	Eg. Quantity through size and position (bubble chart)
Form		3 spatial dimensions	Eg. Quantity through variation in size/volume (proportional shape)

Suzanne Little, School of Computing, DCU

Data representation: **Attributes**

Quantitative

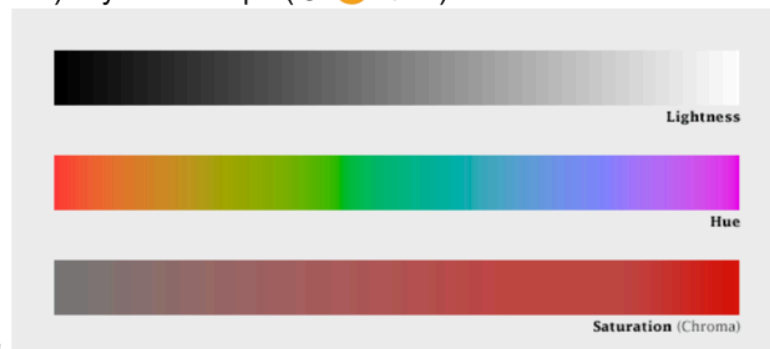
Position
Size (length, area, volume)
Angle/Slope
Quantity
Colour: Saturation
Colour: Lightness
Pattern
Motion

Categorical

Colour: Hue
Symbol/Shape (☺ ☹ ♻ Ω)

Relational

Connection/Edge
Containment



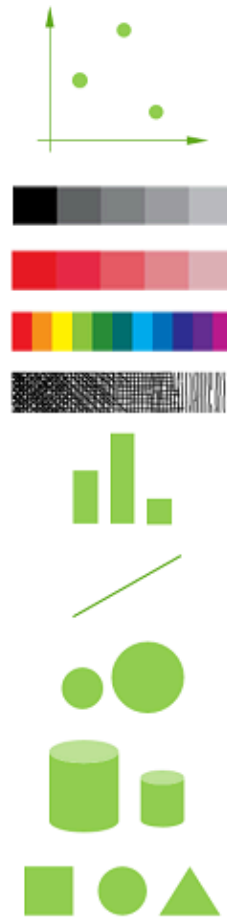
Suzanne Little, School of Computing, DCU

If you prefer a visual representation of the perceptual ranking, this graphic was created by Patrik Lundblad.

NOMINAL



ORDINAL



INTERVAL / RATIO



(from: <https://blog.qlik.com/visual-encoding>)

Example

Hopefully you recognise the graph below from earlier in the course. This is the chart created by Hans Rosling in gapminder. Let's identify the marks and attributes used in this graph and consider the density of data encoded.

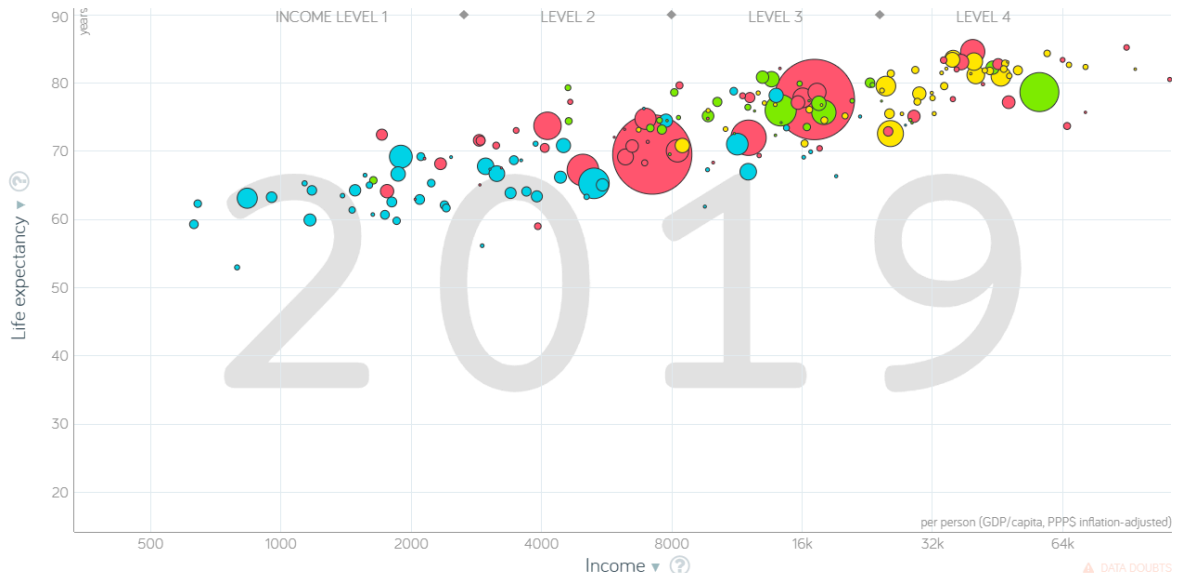


Fig: screenshot from gapminder showing Global Income vs Life expectancy for 2019 in a bubble chart.

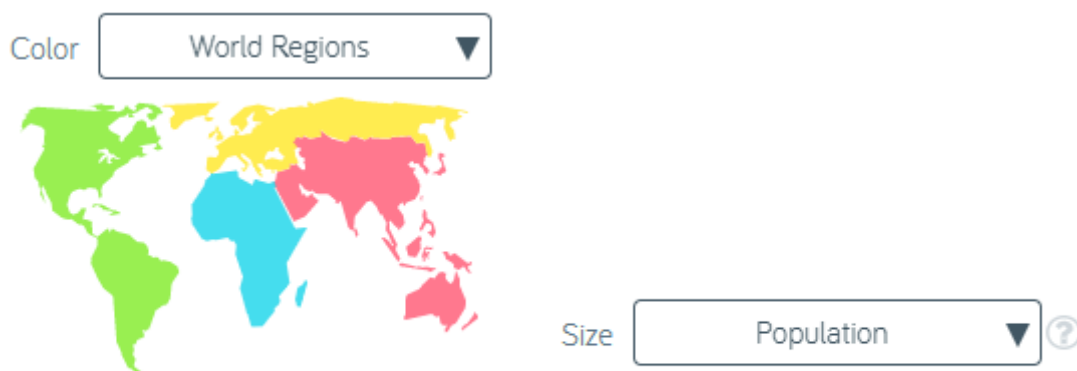


Fig: legends from the gapminder bubble chart

Let's start by considering which mark or marks are used to encode the data.



This could be a point but since the size is also used it is an *area* mark.

We know that area is a 2 dimensional encoding mark so what attributes is this mark using to encode data?

1. The position (X & Y)
2. The area (note: not the radius)

3. The colour? Specifically this is the Hue.
Look at that. There's actually a 3rd attribute used.

What data is being encoded?

1. Life expectancy (Y position)
2. Income (X position)
3. Population (Area)
4. World region (Categorical - Hue)

So we have 3 quantitative values and 1 qualitative value encoded in each mark.

You may remember that this is not a static graph but an interactive and dynamic one. It also uses motion to animate the change from year to year. If we include that element then it adds another piece of data to the marks and we have an associated year for each mark depending on the position of the slider at the bottom of the screen and the background label. The direction of the motion as you watch the animation also carries information. You can identify if a mark is rising or falling in either direction. This gives us

5. Year
6. Relation (rising/falling)

Bubble charts are particularly data dense so you can see why they are used in gapminder with the complexity of exploring this type of data.

Why is this useful? Similar to knowing the correct terms to label the graph components, knowing how data is encoded in a graph will help you understand the message you are sending (reminder: sender, medium, message, receiver). It's also useful to tune your eye for analysing and critiquing visualisations.

References

Andy Kirk (2016) "Data Visualisation", Sage

A good alternative explanation of encoding: <https://blog.qlik.com/visual-encoding>