6 Data Representation

In this chapter you will explore in detail the first, and arguably the most significant, layer of the visualisation design anatomy: data representation. This is concerned with deciding in what visual form you wish to show your data.

To really get under the skin of data representation, we are going to look at it from both theoretical and pragmatic perspectives. You will start by learning about the building blocks of visual encoding, the real essence of this discipline and something that underpins all data representation thinking. Whereas visual encoding is perhaps seen as the purist 'bottom-up' viewpoint, the 'top-down' perspective possibly offers more pragmatic value by framing your data representation thinking around the notion of chart types. For most people facing up to this stage of data representation, this is conceptually the more practical entry point from which to shape their decisions.

To substantiate your understanding of this design layer you will take a tour through a gallery of 49 different chart type options, reflecting the many common and useful techniques being used to portray data visually in the field today. This gallery will then by supplemented by an overview of the key influencing factors that will inform and determine the choices you make.

6.1 Introducing Visual Encoding

As introduced in the opening chapter, data representation is the act of giving visual form to your data. As viewers, when we are perceiving a visual display of data we are *decoding* the various shapes, sizes, positions and colours to form an understanding of the quantitative and categorical values represented. As visualisers, we are doing the reverse through visual *encoding*, assigning visual properties to data values. Visual encoding forms the basis of any chart or map-based data representation, along with the components of chart apparatus that help complete the chart display.

There are many different ways of encoding data but these always comprise combinations of two different properties, namely *marks* and *attributes*. Marks are visible features like dots, lines and areas. An individual mark can represent a record or instance of data (e.g. your phone bill for a given month). A mark can also represent an aggregation of records or instances (e.g. a summation of individual phone charges to produce the bill for a given month). A set of marks would therefore represent a set of records or instances (e.g. the 12 monthly phone bills for 2015).

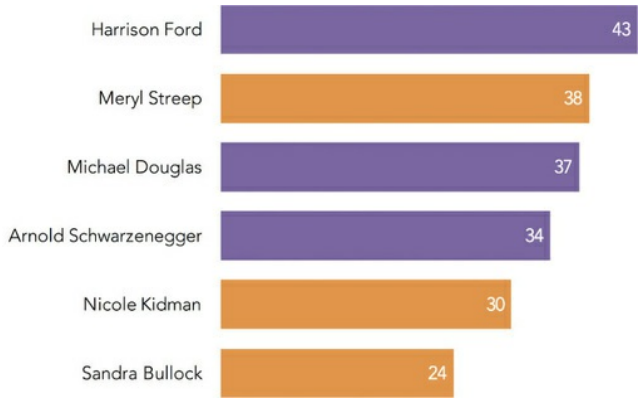
Attributes are variations applied to the appearance of marks, such as the size, position, or colour. They are used to represent the values held by different quantitative or categorical variables against each record or instance (or, indeed, each aggregation). If you had 12 marks, one for each phone bill during 2015, you could use the size attribute of each mark to represent the various phone bill totals.

Figure 6.1 offers a more visual illustration. In the dataset there are six records, one for each record listed. 'Gender' is a categorical variable and 'Years Since First Movie' is a quantitative variable. 'Male' and '43' are the specific values of these variables associated with Harrison Ford. In the associated chart, each actor from the table is represented by the mark of a line (or bar). This represents their record or instance in the table. Harrison Ford's bar is proportionally sized in scale to represent the 43 years since his first movie and is coloured purple

to distinguish his gender as 'Male'. Each of the five other actors similarly has a bar sized according to the years since their first movie and coloured according to their gender.

Figure 6.1 Illustration of Visual Encoding

ACTOR	GENDER	YEARS SINCE FIRST MOVIE
Harrison Ford	Male	43
Meryl Streep	Female	38
Michael Douglas	Male	37
Arnold Schwarzenegger	Male	34
Nicole Kidman	Female	30
Sandra Bullock	Female	24



The objective of visual encoding is to find the right blend of marks and attributes that most effectively will portray the angle of analysis you wish to show your viewers. The factors that shape your choice and define the notion of what is considered 'effective' are multiple and varied in their influence. Before getting on to there, let's take a closer look at the range of different marks and attributes that are commonly found in the data representation toolkit.

It is worth noting upfront that while the organisation of the 'attributes', in particular, suggests a primary role, several can be deployed to encode both categorical (nominal, ordinal) variables and quantitative variables. Furthermore, as you see in the bar chart in Figure 6.1, combinations of several attributes are often applied to marks (such as colour and size) to encode multiple values.

> Although beyond the scope of this book, there are techniques being developed in the field exploring the use of non-visual senses to portray data, using variations in properties for auditory (sound), haptic (touch), gustatory (taste) and olfactory (smell) senses.

Figure 6.2 List of *Mark* Encodings

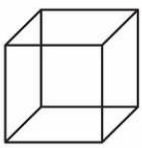
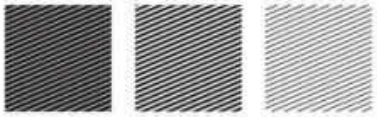
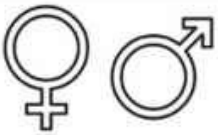
MARK	EXAMPLE	DESCRIPTION
Point		The *point* mark has no variation ('constant') in the spatial dimension. It is largely a placeholder commonly used to represent a quantity through position on a scale, forming the basis of, for example, scatter plots.
Line		The *line* mark has one ('linear') spatial dimension. It is commonly used to represent quantitative value through variation in size, forming the basis of, for example, the bar chart.
Area		The *area* mark has two ('quadratic') spatial dimensions. It is commonly used to represent quantitative values through variation in size and position, forming the basis of, for example, bubble plots.
Form		The *form* mark has three ('cubic') spatial dimensions. It might be used to represent quantitative values through variation in size (specifically, through volume), forming the basis of, for example, a 3D proportional shape chart.

Figure 6.3 List of *Attribute* Encodings

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.

Grasping the basics of visual encoding and its role in data visualisation is one of the fundamental pillars of understanding this discipline. However, when it comes to the reality of considering your data representation options you do not necessarily need to always approach things from this somewhat bottom-up perspective. For most people's needs when creating a data visualisation it is more pragmatic (and perhaps more comprehensible) to think about data representation from a top-down perspective in the shape of chart types.

If marks and attributes are the ingredients, a chart 'type' is the recipe offering a predefined template for displaying data. Different chart types offer different ways of representing data, each one comprising unique combinations of marks and attributes onto which specific types of data can be mapped.

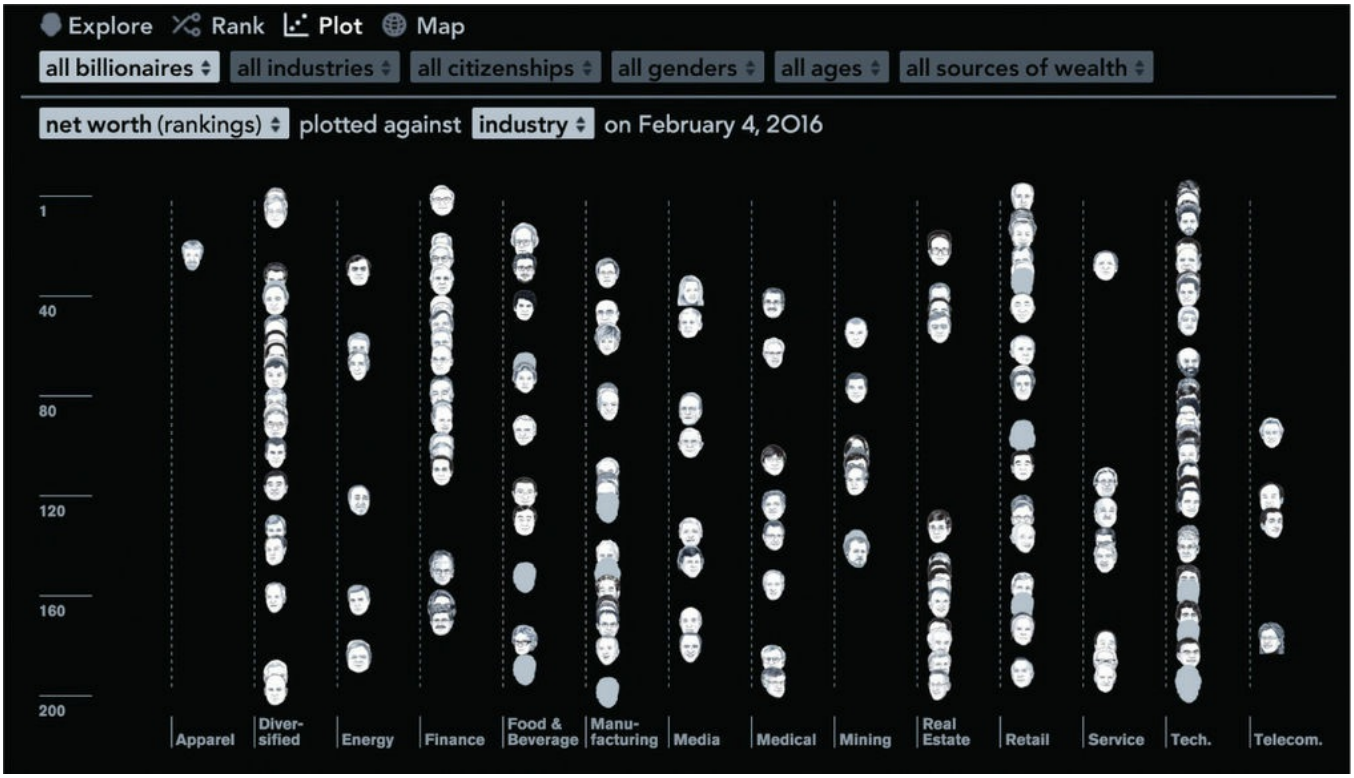
> Recall that I am using chart type as the all-encompassing term, though this is merely a convenient singular label to cover any variation of map, graph, plot and diagram based around the representation of data.

Let's work through a few examples to illustrate the relationship between some selected chart types demonstrating different combinations of *marks* and *attributes*.

To begin with Figure 6.4, visualises the recent fortunes of the world's billionaires. The display shows the relative ranking of each profiled billionaire in the rich list, grouping them by the different sectors of industry in which they have developed their wealth. This data is encoded using the *point* mark and two attributes of *position*. The point in this deployment is depicted using small caricature face drawings representative of each individual – effectively unique symbols to represent the distinct 'category' of each different billionaire. Note that these are points, as distinct from area marks, because their size is constant and insignificant in terms of any quantitative implication. The position in the allocated column signifies the industry the individuals are associated with, while the vertical position signifies the rank (higher position = higher rank towards number 1).

> For reference, this is considered a derivative of the univariate scatter plot, which usually shows the dispersal of a range of absolute values rather than rank.

Figure 6.4 Bloomberg Billionaires



As seen in Chapter 1, the clustered bar chart in Figure 6.5 displays a series of *line* marks (normally described as bars). There are 11 pairs of bars, one for each of the football seasons included in the aggregated analysis. The attribute of *colour* is used to distinguish the bars between the two quantitative measures displayed: blue is for 'games', purple is for 'goals'. The *size* dimension of 'height' (the widths are constant) along the y-axis scale then represents the quantitative values associated with each season and each measure.

Figure 6.6 is called a bubble chart and displays a series of geometric *area* marks to represent the top 100 blog posts on my website based on their popularity over the previous 100 days. Each circle represents an individual post and is *sized* to show the quantitative value of 'total visits' and then *coloured* according to the seven different post categories I use to organise my content.

Figure 6.5 Lionel Messi: Games and Goals for FC Barcelona

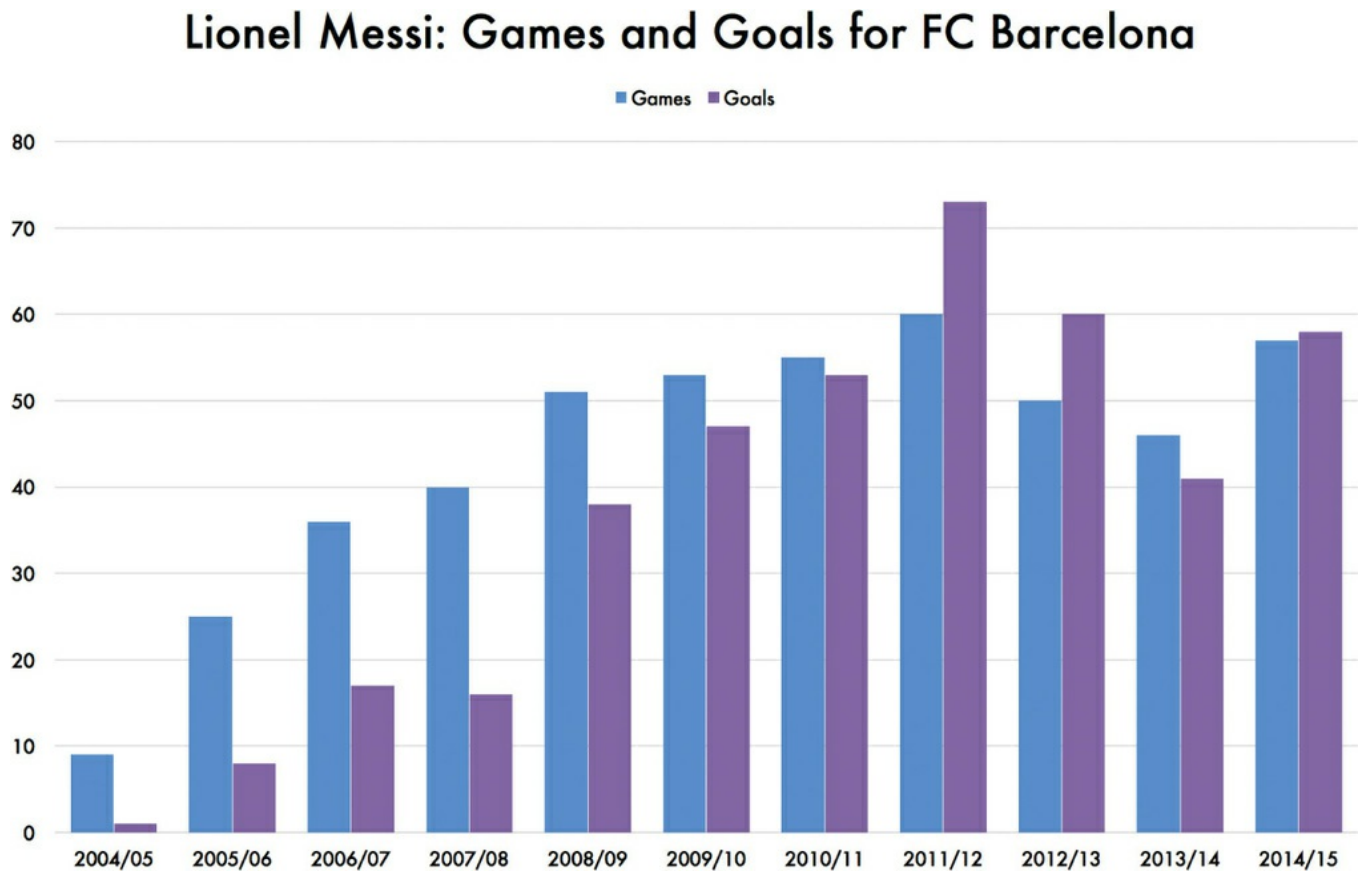


Figure 6.6 Image from the home page of visualisingdata.com

Figure 6.7 How the Insane Amount of Rain in Texas Could Turn Rhode Island Into a Lake

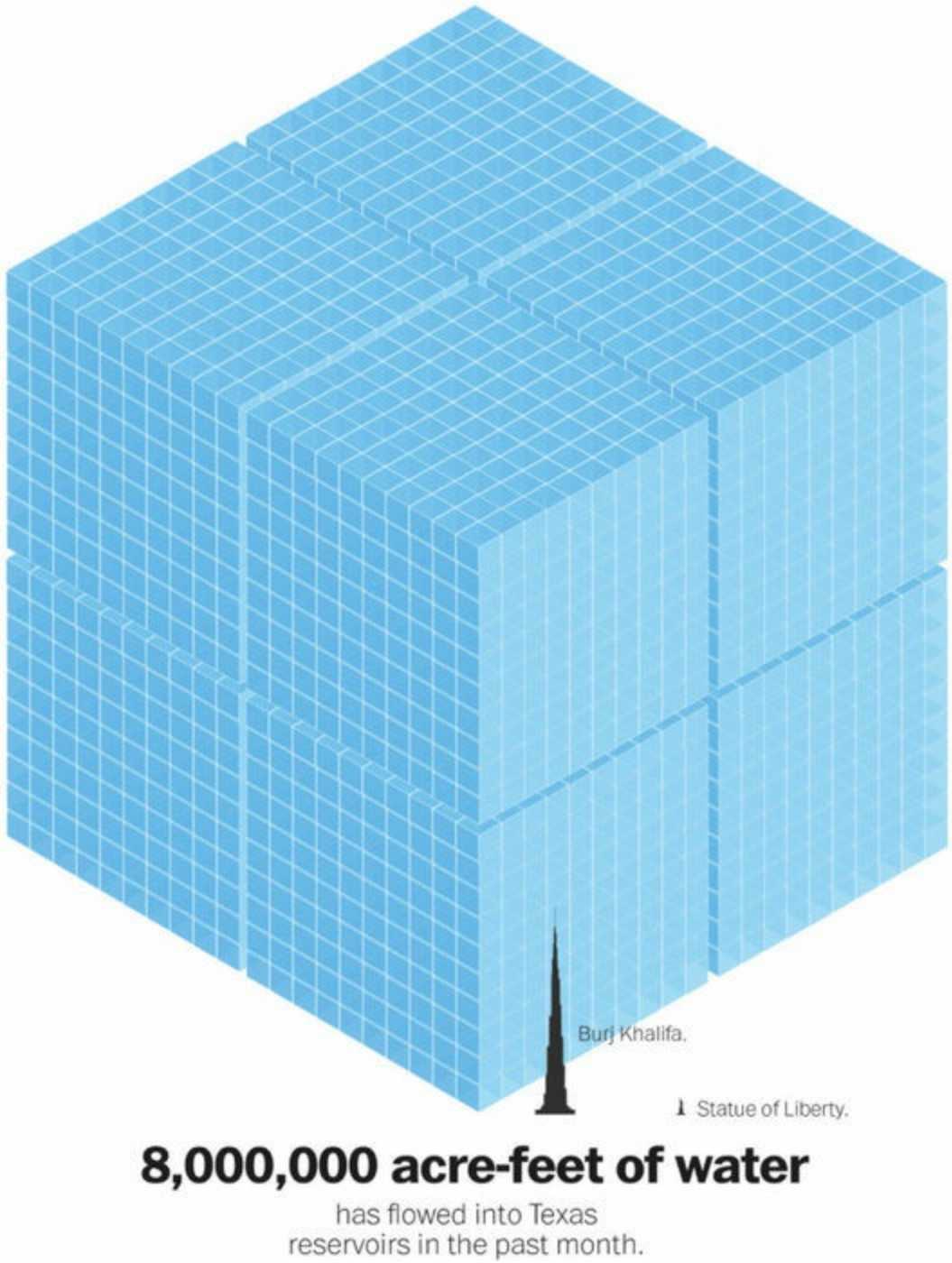


Figure 6.7 demonstrates the use of the *form*, which is more rarely used. My advice is that it should remain that way as it is hard for us to judge scales of volume in 2D displays. However, it can be of merit when values are extremely diverse in size as in this good example. The chart displayed contextualises the amount of water that had flowed into Texas reservoirs in the 30 days up to 27 May 2015. The *size* (volume) of a cube is used to display the amount of rain, with 8000 small cubes representing 1000 acre-feet of water (43,560,000 cubic feet or 1233.5 mega litres) to create the whole (8 million acre-feet), which is then compared against the heights of the Statue of Liberty and what was then the world's tallest building, the Burj Khalifa, to orient in height terms at least.

6.2 Chart Types

For many people, creating a visualisation involves using tools that offer chart menus: you might select a chart type and then 'map' the records and variables of data against the marks and attributes offered by that particular

chart type. Different tools will offer the opportunity to work with a different range of chart types, some with more than others.

As you develop your capabilities in data visualisation and become more 'expressive' – trying out unique combinations of marks and attributes – your approach might lean more towards thinking about representation from a bottom-up perspective, considering the visual encodings you wish to deploy and arriving at a particular chart type as the destination rather than an origin. This will be especially likely if you develop or possess a talent for creating visualisations through programming languages.

As the field has matured over the years, and a greater number of practitioners have been experimenting with different recipes of marks and attributes, there is now a broad range of established chart types. Once again I hesitate to use the universal label of chart type (some mapping techniques are not chart types per se) but it will suffice. While all of us are likely to be familiar with the 'classic three' – namely, the bar, pie and line chart – there are many other chart type options to consider.

To acquaint you with a broader repertoire of charting options, over the coming pages I present you with a gallery. This offers a curated collection of some of the common and useful chart types being used across the field today. This gallery aims to provide you with a valuable reference that will directly assist your judgements, helping you to pick (conceptually, at least) from a menu of options.

I have attempted to assign each chart to one of five main families based on their primary analytical purpose. What type of angle of analysis does each one principally show? Using the five-letter mnemonic CHRTS this should provide a useful taxonomy for organising your thinking about which chart or charts to use for your data representation needs.

> I know what you're thinking: 'well that's a suspiciously convenient acronym'! Honestly, if it was as intentional as that I would have tried harder to somehow crowbar in an 'A' family. OK, I did spend a lot of time, but I couldn't find it and it's now my life's ambition to do so. Only then will my time on this planet have been truly worthwhile. In the meantime, CHRTS is close enough. Besides, vowels are hugely overrated.

CATEGORICAL	Comparing categories and distributions of quantitative values
HIERARCHICAL	Charting part-to-whole relationships and hierarchies
RELATIONAL	Graphing relationships to explore correlations and connections
TEMPORAL	Showing trends and activities over time
SPATIAL	Mapping spatial patterns through overlays and distortions

Each chart type presented is accompanied by an array of supporting details that will help you fully acquaint yourself with the role and characteristics of each option.

A few further comments about what this gallery provides:

- The primary name used to label each chart type as well as some further alternative names that are often used
- An indication of which CHRTS family each chart belongs to, based on their specific primary role, as well as a sub-family definition for further classification

- An indicator for each chart type to show which ones I consider to be most useful for undertaking Exploratory Data Analysis (the black magnifying glass symbol)
- An indicator for whether I believe a chart would typically require interactive features to offer optimum usability (the black cursor symbol)
- A description of the chart's representation: what it shows and what encodings (marks, attributes) it is comprised of
- A working example of the chart type in use with a description of what it specifically shows
- A 'how to read' guide, advising on the most effective and efficient approach to making sense of each chart type and what features to look out for
- Presentation tips offering guidance on some of the specific choices to be considered around interactivity, annotation, colour or composition design
- 'Variations and alternatives' offer further derivatives and chart 'siblings' to consider for different purposes

Exclusions: It is by no means an exhaustive list: the vast permutations of different marks and attributes prevents any finite limit to how one might portray data visually. I have, however, consciously excluded some chart types from the gallery mainly because they were not different enough from other charts that have been profiled in detail. I have mentioned charts that represent legitimate derivatives of other charts where necessary but simply did not deem it worthy to assign a whole page to profile them separately. The *Voronoi treemap*, for example, is really just a circular treemap that uses different algorithms to arrange its constituent pieces. While the construction task is different, its usage is not. The *waterfall chart* is a single stacked bar chart broken down into sequenced stages.

Inclusions: I have wrestled with the rights and wrongs of including some chart types, unquestionably. The radar chart, for example, has many limitations and flaws but is not entirely without merit if deployed in a very specific way and only for certain contexts. By including profiles of partially flawed charts like these I am using the gallery as much to signpost their shortcomings so that you know to use them sparingly. There will be some purists gathering in angry mobs and foaming at the mouth in reaction to the audacity of my including the pie chart and word cloud. These have limited roles, absolutely, but a role nonetheless. Put down your pitchforks, return to your homes and have a good read of my caveats. Rather than being the poacher of all bad stuff, I think a gamekeeper role is equally important.

Although I have excluded several charts on grounds of demonstrating only a slight variation on profiled charts, there are some types included that do exhibit only small derivations from other charts (such as the bar chart and the clustered bar, or the scatter plot and the bubble plot). In these cases I felt there was sufficient difference in their practical application, and they were in common usage, to merit their separate inclusion, despite sharing many similarities with other profiled siblings.

'Interestingly, visualisations of textual data are not as developed as one would expect. There is a great need for such visualisations given the amount of textual information we generate daily, from social media to news media and so on, not to mention all the materials generated in the past and that are now digitally available. There are opportunities to contribute to the research efforts of humanists as well as social scientists by devising ways to represent not only frequencies of words and topics, but also semantic content. However, this is not at all trivial.' Isabel Meirelles, Professor, OCAD University (Toronto), discussing one of the many remaining unknowns in visualisation

Categorical comparisons: All chart types can feasibly facilitate comparisons between categories, so

why have a separate C family? Well, the distinction is that those charts belonging to the H, R, T and S families offer an additional dimension of analysis *as well* as providing comparison between categories.

Dual families: Some charts do not fit just into a single family. Showing connected relationships (e.g. routes or flows) on a map is ticking the requirements across at least two or family groups (Relational, Spatial). In each case I have tried to best-fit the family classifications around the primary angle of analysis portrayed by each chart – what is the most prominent aspect that characterises each representation technique.

Text visualisation: As I noted in the discussion about data types, when it comes to working with textual-based data you are almost always going to need to perform some transformation, maybe through value extraction or by applying a statistical technique. The text itself can otherwise largely function only as an annotated device. Chart types used to visualise text actually visualise the properties of text. For example, the word cloud visualises the quantitative frequency of the use of words: text might be the subject, but categories (words) and their quantities (counts) are the data mappings. Varieties of network diagrams might show the relationship between word usage, such as the sequence of words used in sentences (word trees), but these are still only made possible through some quantitative, categorical or semantic property being drawn from the original text.

Dashboard: These methods are popular in corporate settings or any context where you wish to create instrumentation that offers both at-a-glance and detailed views of many different analytical and information monitoring dimensions. Dashboards are not a unique chart type themselves but rather should be considered projects that comprise multiple chart types from across the repertoire of options presented in the gallery. Some of the primary demands of designing dashboards concern editorial thinking (what angles to show and why) and composition choices (how to get it all presented in a unified page layout).

Small multiples: This is an invaluable technique for visualising data but not necessarily a chart type per se and, once again, more a concern for about editorial thinking and composition design. Small multiples involve repeated display of the same chart type but with adjustments to the framing of the data in each panel. For example, each panel may show the same angle of analysis but for different categories or different points in time. Small multiples are highly valued because they exploit the capabilities of our visual perception system when it comes to comparing charts in a simultaneous view, overcoming our weakness at remembering and recalling chart views when consumed through animated sequences or across different pages.

A note about 'storytelling': Storytelling is an increasingly popular term used around data visualisation but I feel it is often misused and misunderstood, which is quite understandable as we all have different perspectives. I also feel it is worth clarifying my take on what I believe storytelling means practically in data visualisation and especially in this discussion about data representation, which is where it perhaps most logically resides in terms of how it is used.

Stories are constructs based on the essence of movement, change or narrative. A line chart shows how a series of values have changed over a temporal plane. A flow map can reveal what relationships exist across a spatial plane between two points separated by distance – they may be evident of a journey. However, aside from the temporal and spatial families of charts, I would argue that no other chart family realistically offers this type of construct in and of itself.

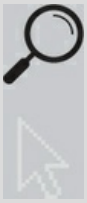
The only way to create a story from other types of charts is to incorporate a temporal dimension (video/slideshow) or provide a verbal/written narrative that itself involves a dimension of time through the sequence of its delivery.

For example, a bar chart alone does not represent a story, but if you show a 'before' and 'after' pair of bar charts side by side or between slides, you have essentially created 'change' through sequence. If you

show a bar chart with a stack on top of it to indicate growth between two points in time, well, you have added a time dimension. A network diagram shows relationships, but stood alone this is not a story – its underlying structure and arrangement are in abstract space. Just as you do when showing friends a photograph from your holiday, you might use this chart as a prop to explain how relationships between some of the different entities presented are significant. Making the chart a prop allows *you* to provide a narrative. In this case it is the setting and delivery that are consistent with the notion of storytelling, not the chart itself. I made a similar observation about the role of exhibitory visualisations used as props within explanatory settings.

A further distinction to make is between stories as being presented and stories as being interpreted. The famous six-word story 'for sale: baby shoes, never worn' by Ernest Hemingway is not presented as a story, the story is triggered in our mind when we dissect this passage and start to infer meaning, implication and context. The imagined bar chart I mentioned earlier in the book that could show the 43 white presidents and 1 black president is only presenting a story if it is accompanied by an explanatory narrative (in which case the chart was again really just a prop) or if you understand the meaning of the significance of this statistic without this description and are able to form the story in your own mind.

Charts Comparisons



Bar chart

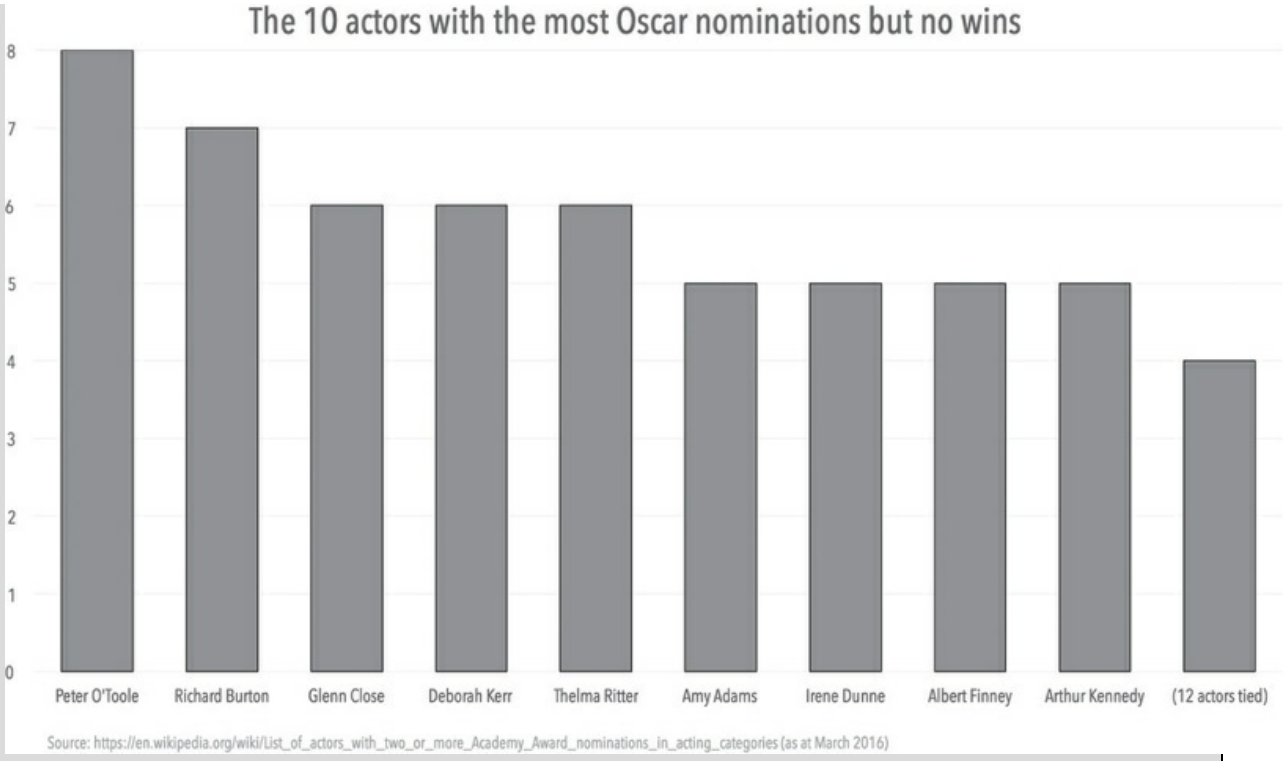
ALSO KNOWN AS Column chart, histogram (wrongly)

REPRESENTATION DESCRIPTION

A bar chart displays quantitative values for different categories. The chart comprises line marks (bars) – not rectangular areas – with the size attribute (length or height) used to represent the quantitative value for each category.

EXAMPLE Comparing the number of Oscar nominations for the 10 actors who have received the most nominations without actually winning an award.

Figure 6.8 The 10 Actors with the Most Oscar Nominations but No Wins



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with which categorical value each bar is associated and what the range of the quantitative values is (min to max). Think about what high and low values mean: is it 'good' to be large or small? Glance across the entire chart to locate the big, small and medium bars and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring bars, to identify larger than and smaller than relationships and estimate the relative proportions. Estimate (or read, if labels are present) the absolute values of specific bars of interest. Where available, compare the quantities against annotated references such as targets, forecast, last year, average, etc.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like tick marks and gridlines, in particular, can be helpful to increase the accuracy of the reading of the quantitative values. If you have axis labels you should not need direct labels on each bar – this will lead to label overload, so generally decide between one or the other.

COMPOSITION: The quantitative value axis should always start from the origin value of zero: a bar should be representative of the true, full quantitative value, nothing more, nothing less, otherwise the perception of bar sizes will be distorted when comparing relative sizes. There is no significant difference in perception between vertical or horizontal bars though horizontal layouts tend to make it easier to accommodate and read the category labels for each bar. Unlike the histogram, there should be a gap, even if very small, between bars to keep each category's value distinct. Where possible, try to make the categorical sorting meaningful.

VARIATIONS & ALTERNATIVES

A variation in the use of bar charts is to show changes over time. You would use a bar chart when the focus is on individual quantitative values over time rather than (necessarily) the trend/change between points, for which a line-chart would be best. 'Spark bars' are mini bar charts that aim to occupy only a word's length amount of space. They are often seen in dashboards where space is at a premium and there is a desire to optimise the density of the display. To show further categorical subdivisions, you might consider the 'clustered bar chart' or a 'stacked bar chart' if there is a part-to-whole angle. 'Dot plots' offer a particularly useful alternative to the bar chart for situations where you have to show large quantitative values with a

narrow range of differences.

Charts Comparisons

Clustered bar chart

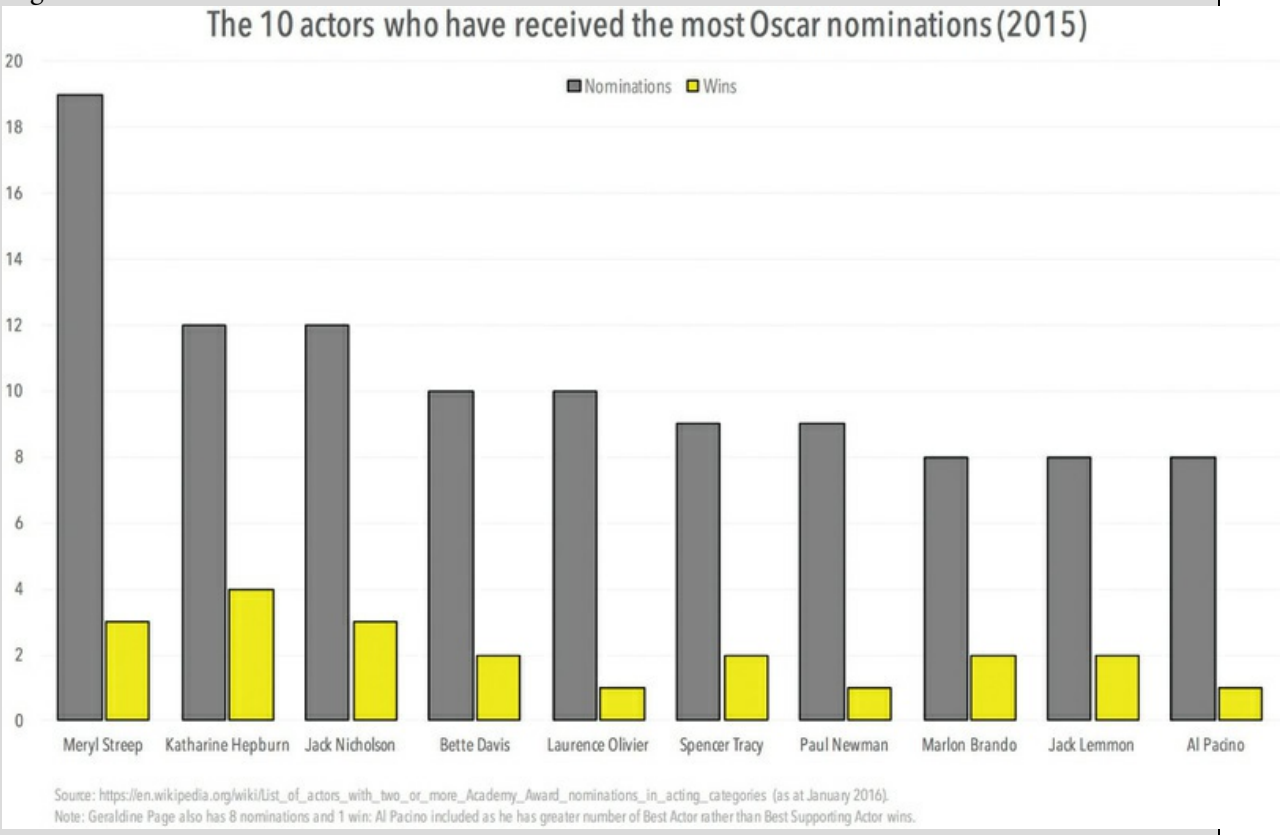
ALSO KNOWN AS Clustered column chart, paired bar chart

REPRESENTATION DESCRIPTION

A clustered bar chart displays quantitative values for different major categories with additional categorical dimensions included for further breakdown. The chart comprises line marks (bars) – not rectangular areas – with the size attribute (length or height) used to represent the quantitative value for each category and colours used to distinguish further categorical dimensions.

EXAMPLE Comparing the number of Oscar nominations with the number of Oscar awards for the 10 actors who have received the most nominations.

Figure 6.9 The 10 Actors who have Received the Most Oscar Nominations



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with which categorical value each bar is associated and what the range of the

quantitative values is (min to max). Learn about the colour associations to understand what sub-categories the bars within each cluster represent. Glance across the entire chart to locate the big, small and medium bars and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Perform local comparisons within clusters to identify the size relationship (which is larger and by how much?) and estimate (or read, if labels are present) the absolute values of specific bars of interest.

PRESENTATION TIPS

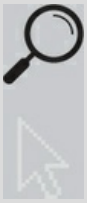
ANNOTATION: Chart apparatus devices like tick marks and gridlines, in particular, can be helpful to increase the accuracy of the reading of the quantitative values. If you have axis labels you should not need direct labels on each bar – this will lead to label overload, so generally decide between one or the other.

COMPOSITION: The quantitative value axis should always start from the origin value of zero: a bar should be representative of the true, full quantitative value, nothing more, nothing less, otherwise the perception of bar sizes will be distorted when comparing relative sizes. If your categorical clusters involve a breakdown of more than three bars, it becomes a little too busy, so you might therefore consider giving each cluster its own separate bar chart and using small multiples to show a chart for each major category. Sometimes one bar might be slightly hidden behind the other, implying a before and after relationship, often when space is at a premium – just do not hide too much of the back bar. There is no significant difference in perception between vertical or horizontal bars though horizontal layouts tend to make it easier to accommodate and read the category labels for each bar. The individual bars should be positioned adjacent to each other with a noticeable gap and then between each cluster to help direct the eye towards the clustering patterns first and foremost. Where possible try to make the categorical sorting meaningful.

VARIATIONS & ALTERNATIVES

Clustered bar charts are also sometimes used to show how two associated sub-categories have changed over time (like the Lionel Messi bar chart discussed in Chapter 1). Alternatives would include the 'dot plot' or, if you have just two categories forming the clusters and these categories have a binary state (male, female or yes %, no %), the 'back-to-back bar chart' would be effective.

Charts Comparisons



Dot plot

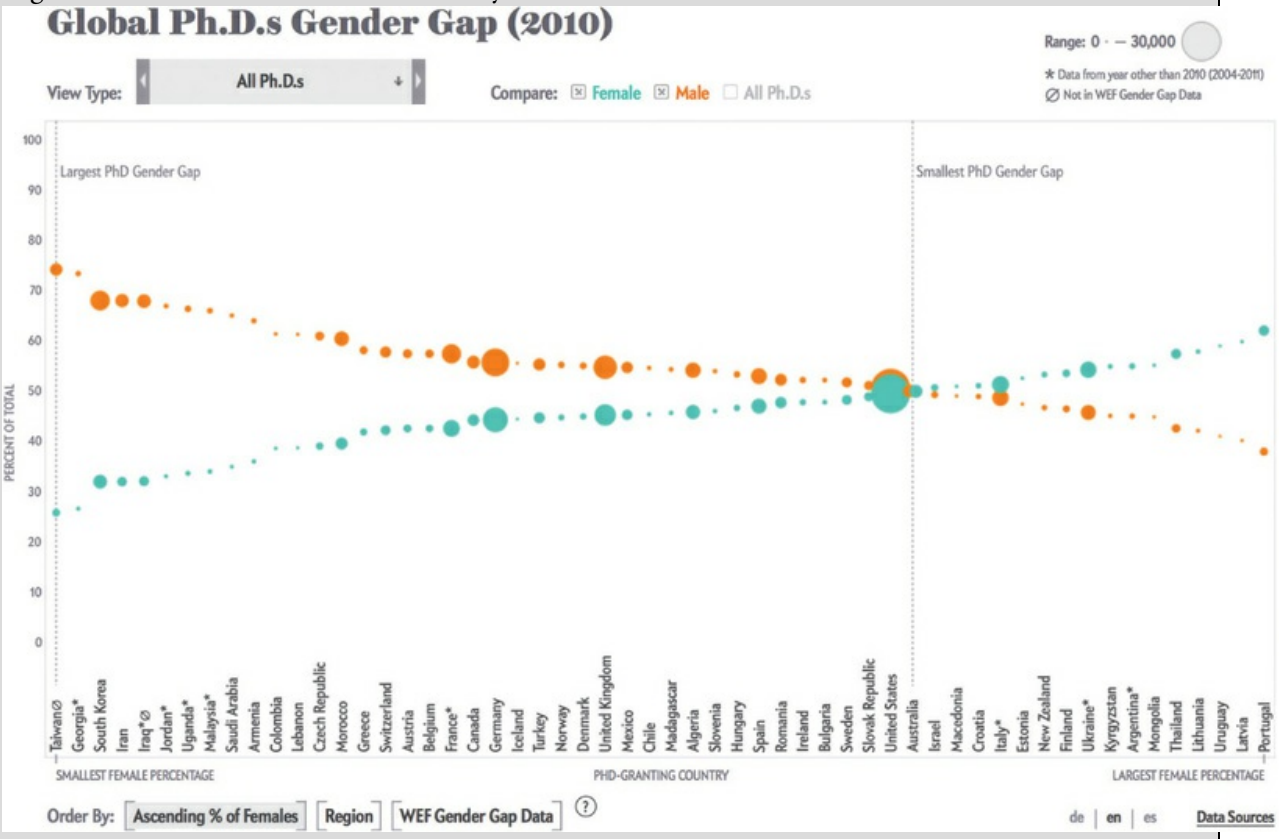
ALSO KNOWN AS Dot chart

REPRESENTATION DESCRIPTION

A dot plot displays quantitative values for different categories. In contrast to the bar chart, rather than using the size of a bar, point marks (typically circles but any 'symbol' is legitimate) are used with the position along a scale indicating the quantitative value for each category. Sometimes an area mark is used to indicate one value through position and another value through size. Additional categorical dimensions can be accommodated in the same chart by including additional marks differentiated by colour or symbol.

EXAMPLE Comparing the number and percentage of PhDs awarded by gender across different academic subjects.

Figure 6.10 How Nations Fare in PhDs by Sex



HOW TO READ IT & WHAT TO LOOK FOR

For single-series dot plots (i.e. just one dot per row), look at the axes so you know with which categorical value each row is associated and what the range of the quantitative values is (min to max). Where you have multiple series dot plots (i.e. more than one dot), establish what the different colours/symbols represent in terms of categorical breakdown. Glance across the entire chart to locate the big, small and medium values and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Where you have multiple series look across each series of dot values separately and then perform local comparisons within rows to identify the relative position of each dot, observing the gaps, big and small. Estimate the absolute values of specific dots of interest. Where available, compare the quantities against annotated references such as targets, forecast, last year, average, etc.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like tick marks and gridlines, in particular, can be helpful to increase the accuracy of the reading of the quantitative values.

COMPOSITION: Given that the quantitative value axis does not need to commence from a zero origin it is important to label clearly the axis values when the baseline is *not* commencing from a minimum of zero. There is no significant difference in perception between vertical or horizontal arrangement though horizontal layouts tend to make it easier to accommodate and read the category labels for each row. Where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order.

VARIATIONS & ALTERNATIVES

Alternatives would include the 'bar chart', to show the size of quantitative values for different categories. The 'connected dot plot' would be used to focus on the difference between two measures. The 'univariate scatter

plot' would be used to show the range of multiple values across categories, to display the diversity and distribution of values.

Charts Comparisons

Connected Dot Plot

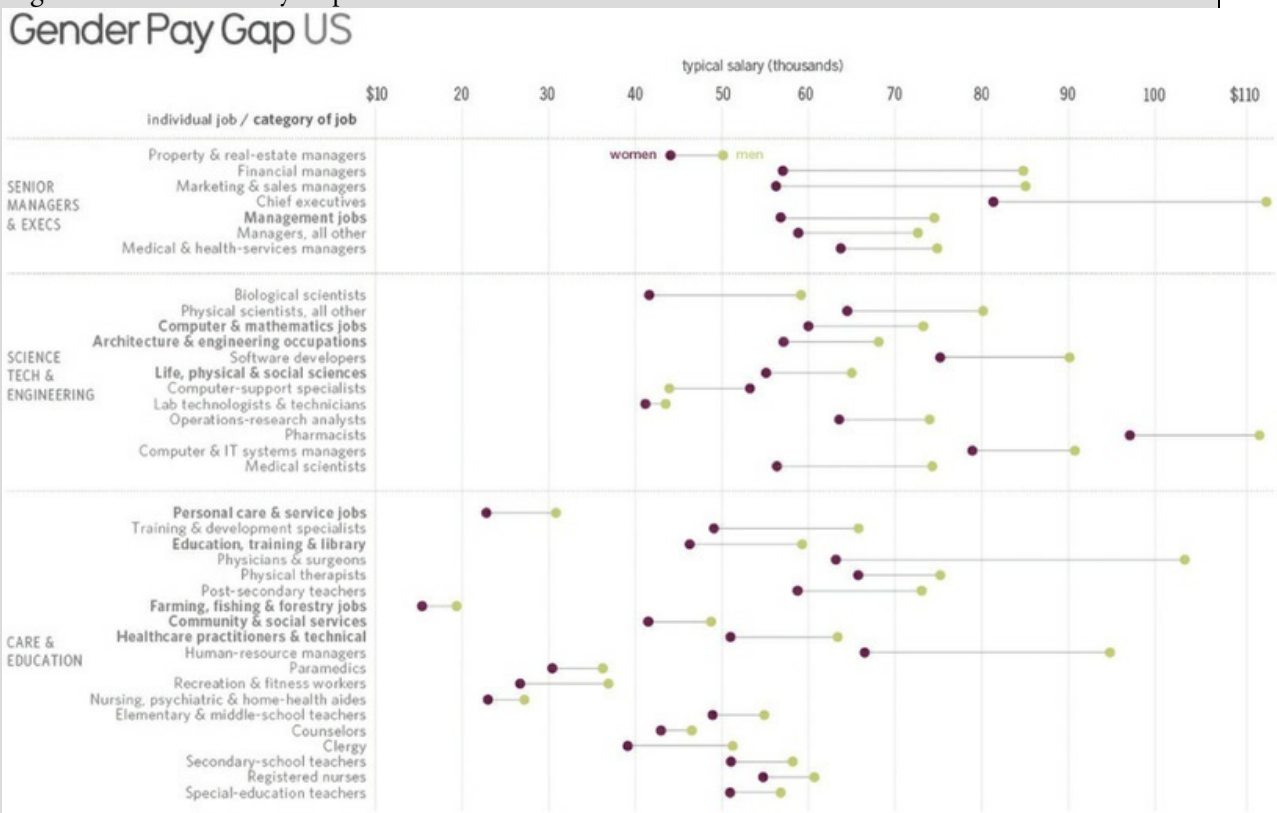
ALSO KNOWN AS Barbell chart, dumb-bell chart

REPRESENTATION DESCRIPTION

A connected dot plot displays absolute quantities and quantitative differences between two categorical dimensions for different major categories. The display is formed by two points (normally circles but any 'symbol' is legitimate) to mark the quantitative value positions for two comparable categorical dimensions. There is a row of connected dots for each major category. Colour or difference in symbol is generally used to distinguish these points. Joining the two points together is a connecting line which effectively represents the 'delta' (difference) between the two values.

EXAMPLE Comparing the typical salaries for women and men across a range of different job categories in the US.

Figure 6.11 Gender Pay Gap US



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with which major categorical values each row is associated and what the range of the quantitative values is (min to max). Determine which dots resemble which categorical dimension (could be colour, symbol or a combination) and see if there is any meaning behind the colouring of the connecting bars. Think about what the quantitative values mean to determine whether it is a good thing to be higher or lower. Glance across the entire chart to locate the big, small and medium connecting bars in each direction. Perform global comparisons to establish the high-level ranking of biggest > smallest differences as well as the highest and lowest values. There may be deliberate sorting of the display based on one of the quantitative measures. Identify any noticeable exceptions and/or outliers. Estimate (or read, if labels are present) the absolute values, direction and size of differences for specific categories of interest.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like tick marks and gridlines, in particular, can be helpful to increase the accuracy of the reading of the quantitative values. Consider labelling categories adjacent to the plotted points rather than next to the axis line (and possibly far away from the values) to make it easier for the reader to understand the category–row association.

COLOUR TIPS: Colour may be used to indicate and emphasise the directional basis of the connecting line differences.

COMPOSITION: If the two plotted measures are very similar, and the point markers effectively overlap, you will need to decide which should be positioned on top. As the representation of the quantitative values is through position along a scale and not size (it is the difference that is sized, not the absolutes) the quantitative axis does not need to have a zero origin. However, a zero origin can be helpful to establish the scale of the differences. Where possible try to make the sorting meaningful using any one of the three quantitative measures to optimise the layout.

VARIATIONS & ALTERNATIVES

Variations in the use of the 'connected dot plot' would show before and after analysis between two points in time, possibly using the 'arrow chart' to indicate the direction of change explicitly. Similarly, the 'carrot chart' uses line width tapering to indicate direction, the fatter end the more recent values. The 'univariate scatter plot' would be used to show the range of multiple values across categories, to display the diversity and distribution of values rather than comparing differences between values.

Charts Comparisons

Pictogram

ALSO KNOWN AS Isotype chart, pictorial bar chart, stacked shape chart, tally chart

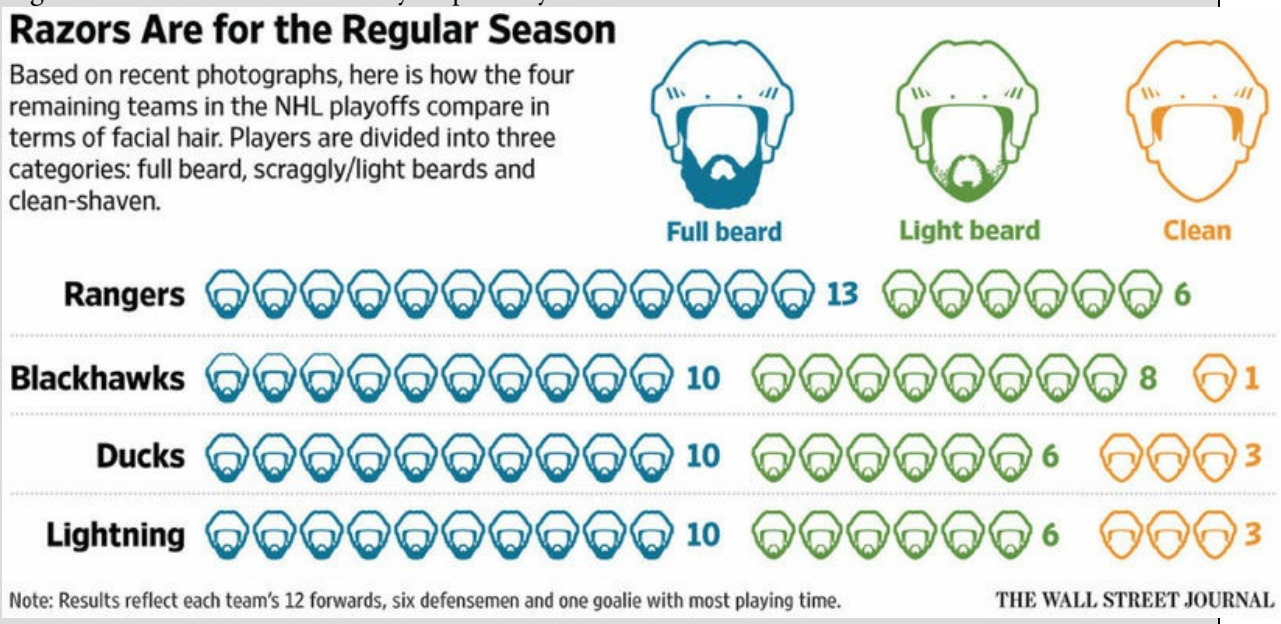
REPRESENTATION DESCRIPTION

A pictogram displays quantitative values for different major categories with additional categorical

dimensions included for further breakdown. In contrast with the bar chart, rather than using the size of a bar, quantities of point marks, in the form of symbols or pictures, are stacked to represent the quantitative value for each category. Each point may be representative of one or many quantitative units (e.g. a single shape may represent 1000 people) but note that, unless you use symbol portions, you will not be able to represent decimals. Pictograms may be used to offer a more emotive (humanising or more light-hearted) display than a bar can offer. Additional categorical dimensions can be accommodated in the same chart by using marks differentiated by variations in colour, symbol or picture. Always ensure the markers used are as intuitively recognisable as possible and consider minimising the variety as this makes it cognitively harder for the viewer to identify associations easily and make sense of the quantities.

EXAMPLE Comparing the number of players with different facial hair types across the four teams in the NHL playoffs in 2015.

Figure 6.12 Who Wins the Stanley Cup of Playoff Beards?



HOW TO READ IT & WHAT TO LOOK FOR

Look at the major categorical axis to establish with which category each row is associated. Establish the mark associations to understand what categorical dimensions each colour/shape variation represents. Glance across the entire chart to locate the big, small and medium stacks of shapes and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring categories, to identify larger than and smaller than relationships and estimate the relative proportions. Estimate (or read, if labels are present) the absolute values of specific groups of markers of interest.

PRESENTATION TIPS

ANNOTATION: The choice of symbol/ picture should be as recognisably intuitive as possible and locate any legends as close as possible to the display.

COLOUR TIPS: Maximise the variation in marker by using different combinations in both colour and shape, rather than just variation of one attribute.

COMPOSITION: If the quantities of markers exceed a single row, try to make the number of units per row logically 'countable', such as displaying in groups of 5, 10 or 100. To aid readability, make sure there is a sufficiently noticeable gap between rows, otherwise sometimes the eye struggles to form the distinct clusters of shapes for each category displayed. Where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order.

VARIATIONS & ALTERNATIVES

Extending the idea of using repeated quantities of representative symbols, some applications take this further by using large quantities of individual symbols to get across the feeling of magnitude and scale. When showing a part-to-whole relationship, the 'waffle chart' can use simple symbol devices to differentiate the constituent parts of a whole.

Charts Comparisons

Proportional shape chart

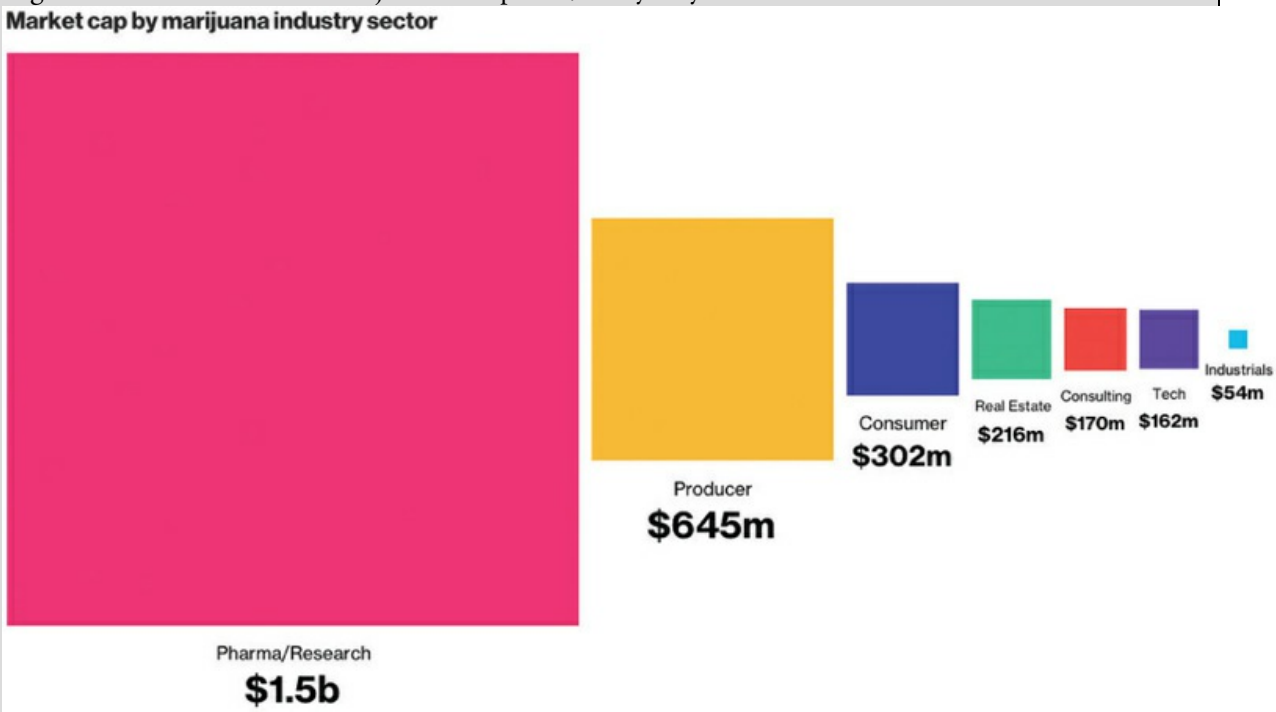
ALSO KNOWN AS Area chart (wrongly)

REPRESENTATION DESCRIPTION

A proportional shape chart displays quantitative values for different categories. The chart is based on the use of different area marks, one for each category, sized in proportion to the quantities they represent. By using the quadratic dimension of area size rather than the linear dimension of bar length or dot position, the shape chart offers scope for displaying a diverse range of quantitative values within the same chart. Typically the layout is quite free-form with no baseline or central gravity binding the display together.

EXAMPLE Comparing the market capitalisation ($) of companies involved in the legal sale of marijuana across different industry sectors.

Figure 6.13 For These 55 Marijuana Companies, Every Day is 4/20



HOW TO READ IT & WHAT TO LOOK FOR

Look at the shapes and their associated labels so you know with what major categorical values each is associated. If there are only direct labels, find the largest shape to establish its quantitative value as the maximum and do likewise for the smallest – this will help calibrate the size judgements. Otherwise, if it exists, acquaint yourself with the size key. Glance across the entire chart to locate the big, small and medium shapes and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring shapes to identify larger than and smaller than relationships and estimate the relative proportions. Estimate (or read, if labels are present) the absolute values of specific shapes of interest.

PRESENTATION TIPS

ANNOTATION: Sometimes a quantitative size key will be included rather than direct labelling (usually when there are many shapes and limited empty space) though direct labels will help overcome some of the limitations of judging area size. You will have to decide how to handle label positioning for those shapes with exceptionally small sizes.

COLOUR TIPS: Colours are not fundamentally necessary to encode category (the position/separation of different shapes achieves that already) but they can be useful as redundant encodings to make the category even more immediately distinguishable.

COMPOSITION: Estimating and comparing the size of areas with accuracy is not as easy as it is for judging bar length or dot position, so only use this chart type if you have a diverse range of quantitative values. The geometric accuracy of the size calculations is paramount. Mistakes are often made, in particular, with circle size calculations: it is the area you are modifying, not the diameter/radius. Arrangement approaches vary: sometimes you see the shapes anchored to a common baseline (bottom or central alignment) while on other occasions they might just 'float'. If you use an organic shape, like a human figure, to represent different quantities you need to adjust the entire shape area, not just the height. Often the approach for this type of display is to treat the figure as a rudimentary rectangular shape. Sometimes the volume of a shape is used rather than area to represent quantitative values (especially if there are almost exponentially different values to show) but this increases the perceptual difficulty in estimating and comparing values. Where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order.

VARIATIONS & ALTERNATIVES

The 'bubble chart' uses clusters of sized bubbles to compare categorical values and, sometimes, to represent part-to-whole analysis. The 'nested shape chart' might include secondary, smaller area sizes nested within each shape to display local part-to-whole relationships.

Charts Comparisons

Bubble chart

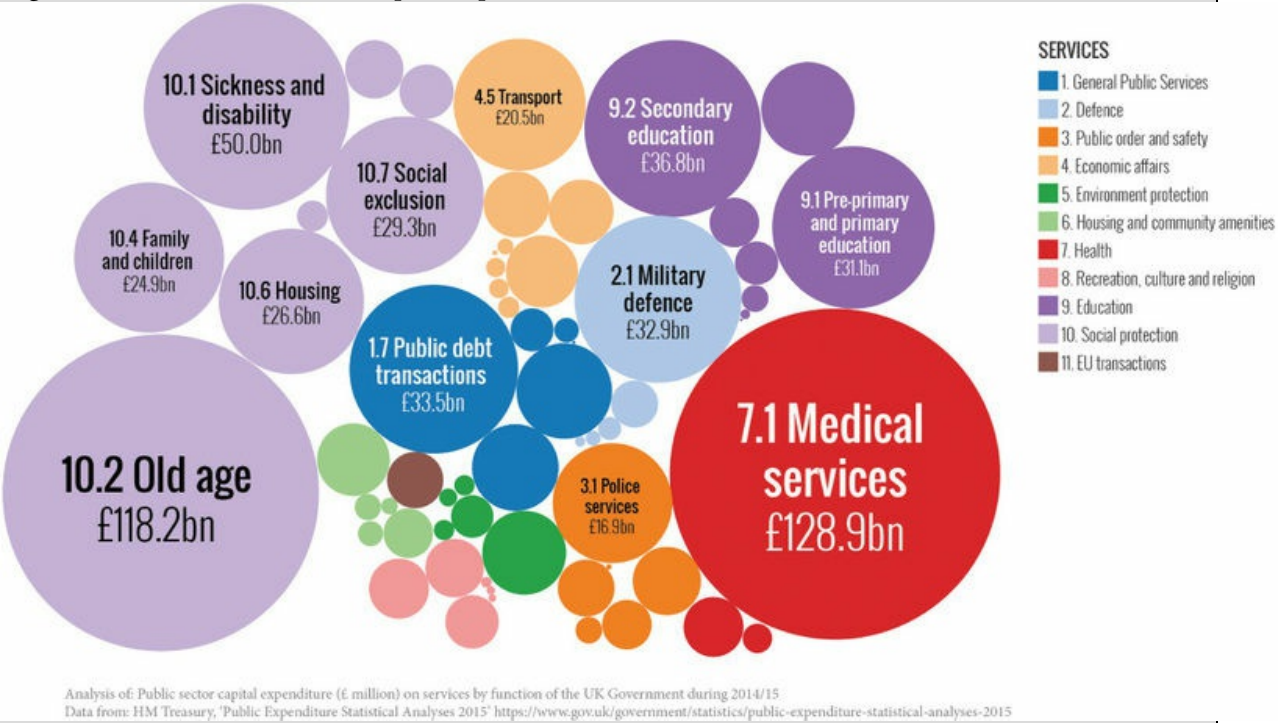
ALSO KNOWN AS Circle packing diagram

EXAMPLE Comparing the Public sector capital expenditure (£ million) on services by function of the UK Government during 2014/15.

REPRESENTATION DESCRIPTION

A bubble chart displays quantitative values for different major categories with additional categorical dimensions included for further breakdown. It is based on the use of circles, one for each category, sized in proportion to the quantities they represent. Sometimes several separate clusters may be used to display further categorical dimensions, otherwise the colouring of each circle can achieve this. It is similar in concept to the proportional shape chart but differs through the typical layout being based on clustering, which therefore also enables it as a device for showing part-to-whole relationships as well.

Figure 6.14 UK Public Sector Capital Expenditure, 2014/15



HOW TO READ IT & WHAT TO LOOK FOR

Look at the shapes and their associated labels so you know with what major categorical values each is associated, noting any size and colour legends to assist in forming associations. If there are multiple clusters, learn about the significance of the grouping/separation in each case. If there are direct labels, find the largest shape to establish its quantitative value as the maximum and do likewise for the smallest – this will help calibrate other size judgements. Glance across the entire chart to locate the big, small and medium shapes and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring shapes to identify larger than and smaller than relationships and estimate the relative proportions. Estimate (or read, if labels are present) the absolute values of specific shapes of interest. If there are multiple clusters, note the general relative size and number of members in each case.

PRESENTATION TIPS

INTERACTIVITY: Bubble charts may often be accompanied by interactive features that let users select or mouseover individual circles to reveal annotated values for the quantity and category.

ANNOTATION: If interactivity is not achievable, a quantitative size key should be included or direct labelling; the latter may make the display busy (and be hard to fit into smaller circles) but will help overcome

some of the limitations of judging area size.

COLOUR TIPS: Colours are sometimes used as redundant encodings to make the quantitative sizes even more immediately distinguishable.

COMPOSITION: Estimating and comparing the size of areas with accuracy is not as easy as it is for judging bar length or dot position, so only use this chart type if you have a diverse range of quantitative values. The use of this chart will primarily be about facilitating a gist, a general sense of the largest and smallest values. The geometric accuracy of the circle size calculations is paramount. Mistakes are often made with circle size calculations: it is the area you are modifying, not the diameter/radius. If you wish to make your bubbles appear as 3D spheres you are essentially no longer representing quantitative values through the size of a geometric area mark; rather the mark will be a 'form' and so the size calculation will be based on volume, not area. There is no categorical or quantitative sorting applied to the layout of the bubble chart, instead the tools that offer these charts will generally use a layout algorithm that applies a best-fit clustering to arrange the circles radially about a central 'gravity' force.

VARIATIONS & ALTERNATIVES

When the collection of quantities represents a whole, this evolves into a chart known as a 'circle packing diagram' and usually involves many parts that pack neatly into a circular layout representing the whole. Another variation of the packing diagram is when the adjacency between circle 'nodes' indicates a connected relation, offering a variation of the node–link diagram for showing networks of relationships. The bubble plot also uses differently sized circles but the position in each case is overlaid onto a scatter plot structure, based on two dimensions of further quantitative variables. Removing the size attribute (and effectively replacing area with point mark) you could simply use the quantity of points clustered together for different categories to create a 'tally chart'.

Charts Comparisons

Radar chart

ALSO KNOWN AS Filled radar chart, star chart, spider diagram, web chart

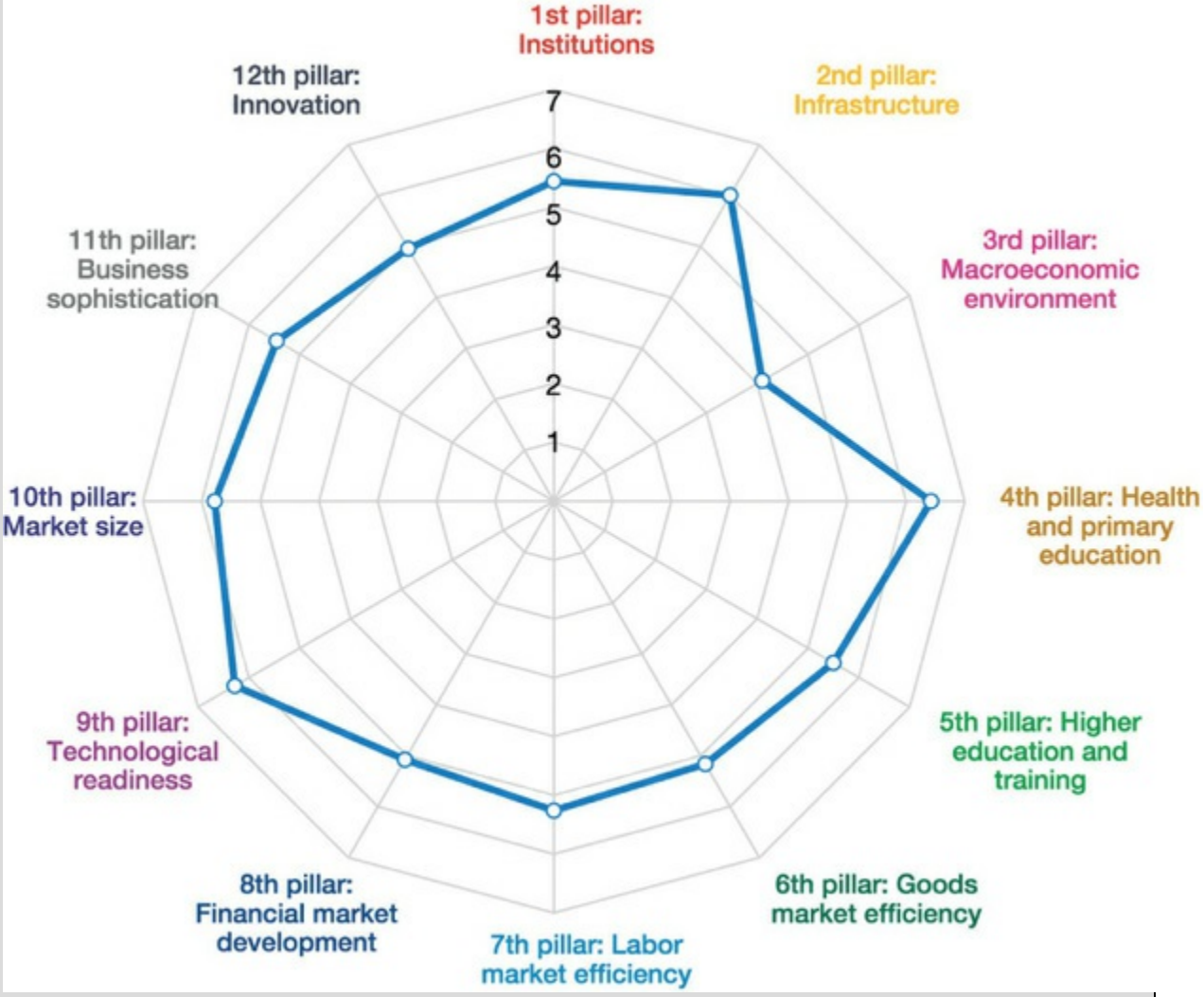
EXAMPLE Comparing the global competitive scores (out of 7) across 12 'pillars' of performance for the United Kingdom.

REPRESENTATION DESCRIPTION

A radar chart shows values for three or more different quantitative measures in the same display for, typically, a single category. It uses a radial (circular) layout comprising several axes emerging from the centre-like spokes on a wheel, one for each measure. The quantitative values for each measure are plotted through position along each scale and then joined by connecting lines to form a unique geometric shape. Sometimes this shape is then filled with colour. A radar chart should only be considered in situations where the cyclical ordering (and neighbourly pairings) has some significance (such as data that might be plotted around the face of a clock or compass) and when the quantitative scales are the same (or similar) for each

axis. Do not plot values for multiple categories on the same radar chart, but use small multiples formed of several radar charts instead.

Figure 6.15 Global Competitiveness Report 2014—2015



HOW TO READ IT & WHAT TO LOOK FOR

Look around the chart and acquaint yourself with the quantitative measure represented by each axis and note the sequencing of the measures around the display. Is there any significance in this arrangement that can assist in interpreting the overall shape? Note the range of values along each independent axis so you understand what positions along the scales mean in a value sense for each measure. Scan the shape to locate the outliers both towards the outside (larger values) and inside (smaller values) of the scales. It is more important to pay attention to the position of values along an axis than the nature of the connecting lines between axes, unless the axis scales are consistent or at least if the relative position along the scale has the same implied meaning. If the variable sequencing has cyclical relevance, the spiking, bulging or contracting shape formed will give you some sense of the balance of values. Perform local comparisons between neighbouring axes to identify larger than and smaller than relationships. Estimate (or read, if labels are present) the absolute values of specific shapes of interest.

PRESENTATION TIPS

ANNOTATION: The inclusion of visible annotated features like axis lines, tick marks, gridlines and value labels can naturally aid the readability of the radar chart. Gridlines are only relevant if there are common scales across each quantitative variable. If so, the gridlines must be presented as straight lines, not concentric arcs, because the connecting lines joining up the values are themselves straight lines.

COLOUR TIPS: Often the radar shapes are filled with a colour, sometimes with a degree of

transparency to allow the background apparatus to be partially visible.

COMPOSITION: The cyclical ordering of the quantitative variables has to be of optimum significance as the connectors and shape change for every different ordering permutation. This will have a major impact on the readability and meaning of the resulting chart shape. As the axes will be angled all around the radial display, you will need to make sure all the associated labels are readable (i.e. not upside down or at difficult angles).

VARIATIONS & ALTERNATIVES

A 'polar chart' is an alternative to the radar chart that removes some of the main shortcomings caused by connecting lines in the radar chart. If you have consistent value scales across the different quantitative measures, a 'bar chart' or 'dot plot' would be a better alternative. While not strictly a variation, 'parallel coordinates' display a similar technique for plotting several independent quantitative measures in the same chart. The main difference is that parallel coordinates use a linear layout and can accommodate many categories in one display.

Charts Comparisons

Polar chart

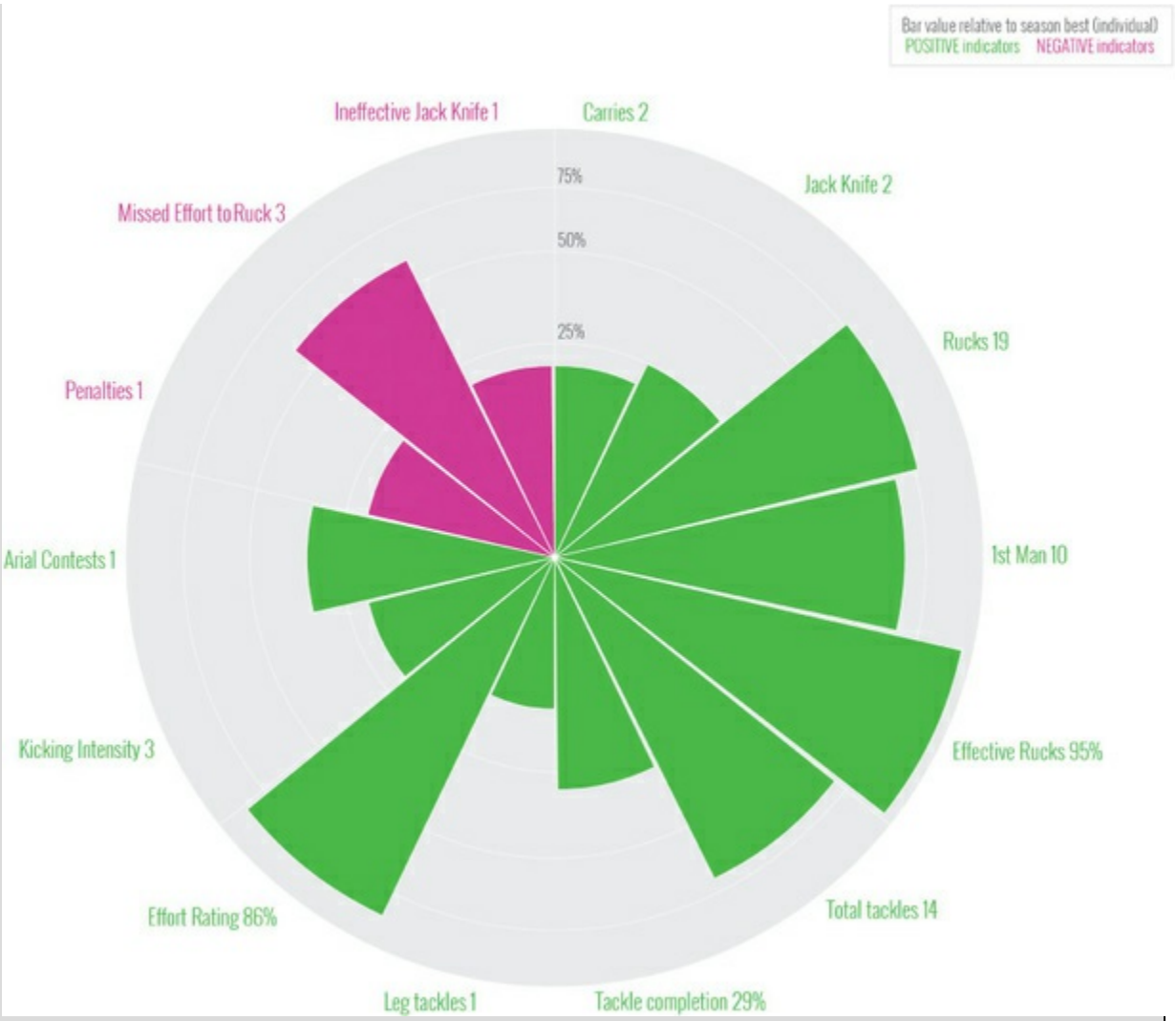
ALSO KNOWN AS Coxcomb plot, polar area plot

REPRESENTATION DESCRIPTION

A polar chart shows values for three or more different quantitative measures in the same display. It uses a radial (circular) layout comprising several equal-angled circular sectors like slices of a pizza, one for each measure. In contrast to the radar chart (which uses position along a scale), the polar chart uses variation in the size of the sector areas to represent the quantitative values. It is, in essence, a radially plotted bar chart. Colour is an optional attribute, sometimes used visually to indicate further categorical dimensions. A polar chart should only be considered in situations where the cyclical ordering (and neighbourly pairings) has some significance (such as data that might be plotted around the face of a clock or compass) and when the quantitative scales are the same (or similar) for each axis.

EXAMPLE Comparing the quantitative match statistics across 14 different performance measures for a rugby union player.

Figure 6.16 Excerpt from a Rugby Union Player Dashboard



HOW TO READ IT & WHAT TO LOOK FOR

Look around the chart and acquaint yourself with the quantitative measures each sector represents and note the sequencing of the measures around the display. Is there any significance in this arrangement that can assist in interpreting the overall shape? Note the range of values included on the quantitative scale and acquaint yourself with any colour associations. Glance across the entire chart to locate the big, small and medium sectors and perform global comparisons to establish the high-level ranking of biggest > smallest. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring variables to identify the order of magnitude and estimate the relative sizes. Estimate (or read, if labels are present) the absolute values of specific sectors of interest. Where available, compare the quantities against annotated references such as targets, forecast, last year, average, etc. If there is significance behind the sequencing of the variables, look out for any patterns that emerge through spiking, bulging or contracting shapes.

PRESENTATION TIPS

ANNOTATION: The inclusion of visible annotated features like tick marks and value labels can naturally aid the readability of the polar chart. Gridlines are only relevant if there are common scales across each quantitative variable. If so, the gridlines must be presented as arcs reflecting the outer shape of each sector. Connecting lines joining up the values are themselves straight lines. Each sector typically uses the same quantitative scale for each quantitative measure but, on the occasions when this is not the case, each axis will require its own, clear value scale.

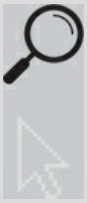
COLOUR TIPS: Often polar chart sectors are filled with a meaningful colour, sometimes with a degree of transparency to allow the background apparatus to be partially visible.

COMPOSITION: The cyclical ordering of the quantitative variables has to be of some significance to legitimise the value of the polar chart over the bar chart. As the sectors will be angled all around the radial display, you will need to make sure all the associated labels are readable (i.e. not upside down or at difficult angles). The quantitative values represented by the size of the sectors need to be carefully calculated. It is the area of the sector, not the radius length, that will be modified to portray the values accurately. If you make maximum quantitative value equivalent to the largest sector area, all other sector sizes can be calculated accordingly. Knowing how many different quantitative variables you are showing means you can easily calculate the angle of any given sector. The quantitative measure axes should always start from the origin value of zero: a sector should be representative of the true, full quantitative value, nothing more, nothing less, otherwise the perception of size will be distorted when comparing relative sizes.

VARIATIONS & ALTERNATIVES

Unless the radial layout provides meaning through the notion of a 'whole' or through the cyclical arrangement of measures, you might be best using a 'bar chart'. Variations in approach tend to see modifications in the sector shape with measure values represented by individual bars lengths or, in the example of the Better Life Index project, through variations in 'petal' sizes.

Charts Distributions



Range Chart

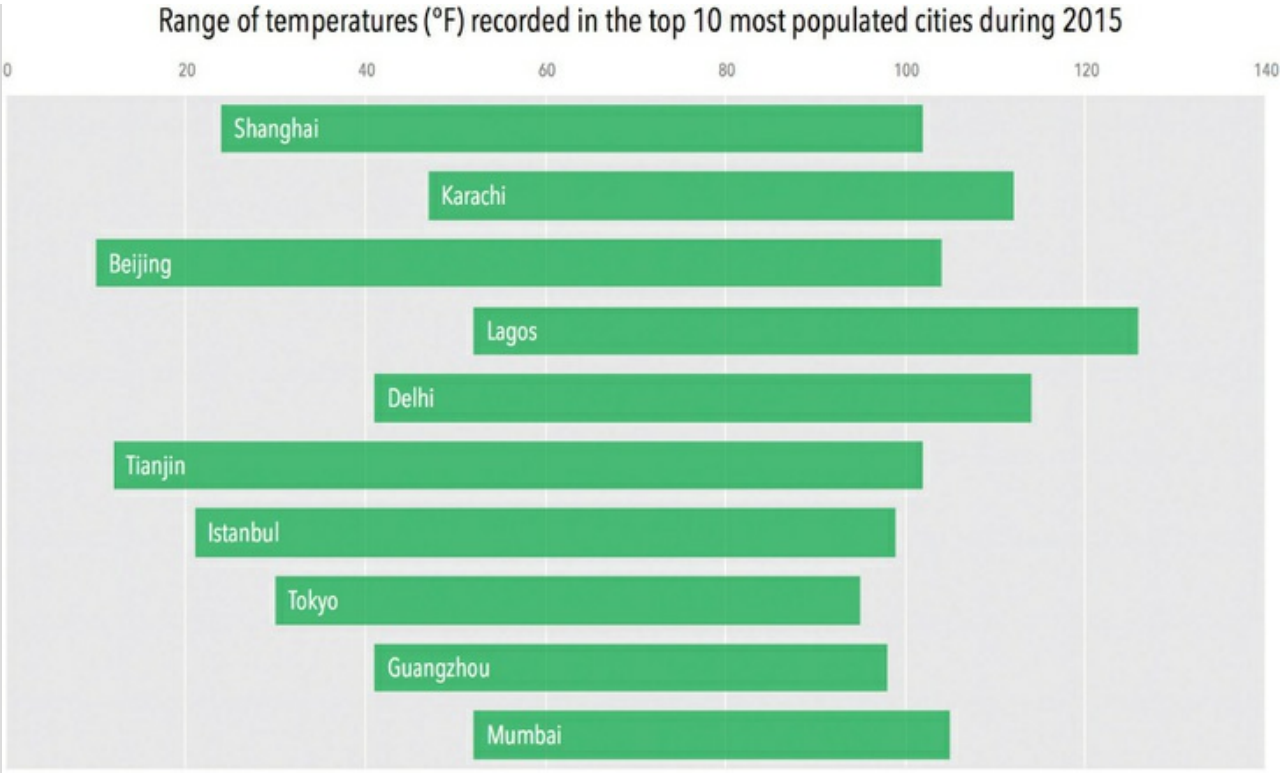
ALSO KNOWN AS Span chart, floating bar chart, barometer chart

REPRESENTATION DESCRIPTION

A range chart displays the minimum to maximum distribution of a series of quantitative values for different categories. The display is formed by a bar, one for each category, with the lower and upper position of the bars shaped by the minimum and maximum quantitative values in each case. The resulting bar lengths thus represent the range of values between the two limits.

EXAMPLE Comparing the highest and lowest temperatures (°F) recorded across the top 10 most populated cities during 2015.

Figure 6.17 Range of Temperatures Recorded in Top 10 Most Populated Cities (2015)



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with what major categorical values each range bar is associated and what the range of the quantitative values is (min to max). Glance across the entire chart to locate the big, small and medium bars and perform global comparisons to establish the high-level ranking of biggest > smallest differences as well as the highest and lowest values. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring bars, to identify larger than and smaller than relationships and estimate the relative proportions. There may be deliberate sorting of the display based on one of the quantitative measures. Estimate (or read, if labels are present) the absolute values of specific bars of interest. Where available, compare the quantities against annotated references such as targets, forecast, last year, average, etc.

PRESENTATION TIPS

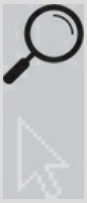
ANNOTATION: Chart apparatus devices like tick marks and gridlines, in particular, can be helpful to increase the accuracy of the reading of the quantitative values. If you have axis labels you may not need direct labels on each bar – this will be lead to label overload, so generally decide between one or the other.

COMPOSITION: The quantitative value axis does not need to commence from zero, unless it means something significant to the interpretation, as the range of values themselves does not necessarily start from zero and the focus is more on the range and difference between the outer values. There is no significant difference in perception between vertical or horizontal layouts, though the latter tend to make it easier to accommodate and read the category labels. Where possible, try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order.

VARIATIONS & ALTERNATIVES

'Connected dot plots' will also emphasise the difference between two selected measure values (as opposed to min/max) or where the underlying data is a change over time between two observations. 'Band charts' will often be used to show how the range of data values has changed over time, displaying the minimum and maximum bands at each time unit. These are often used in displays like weather forecasts.

Charts Distributions



Box-and-whisker plot

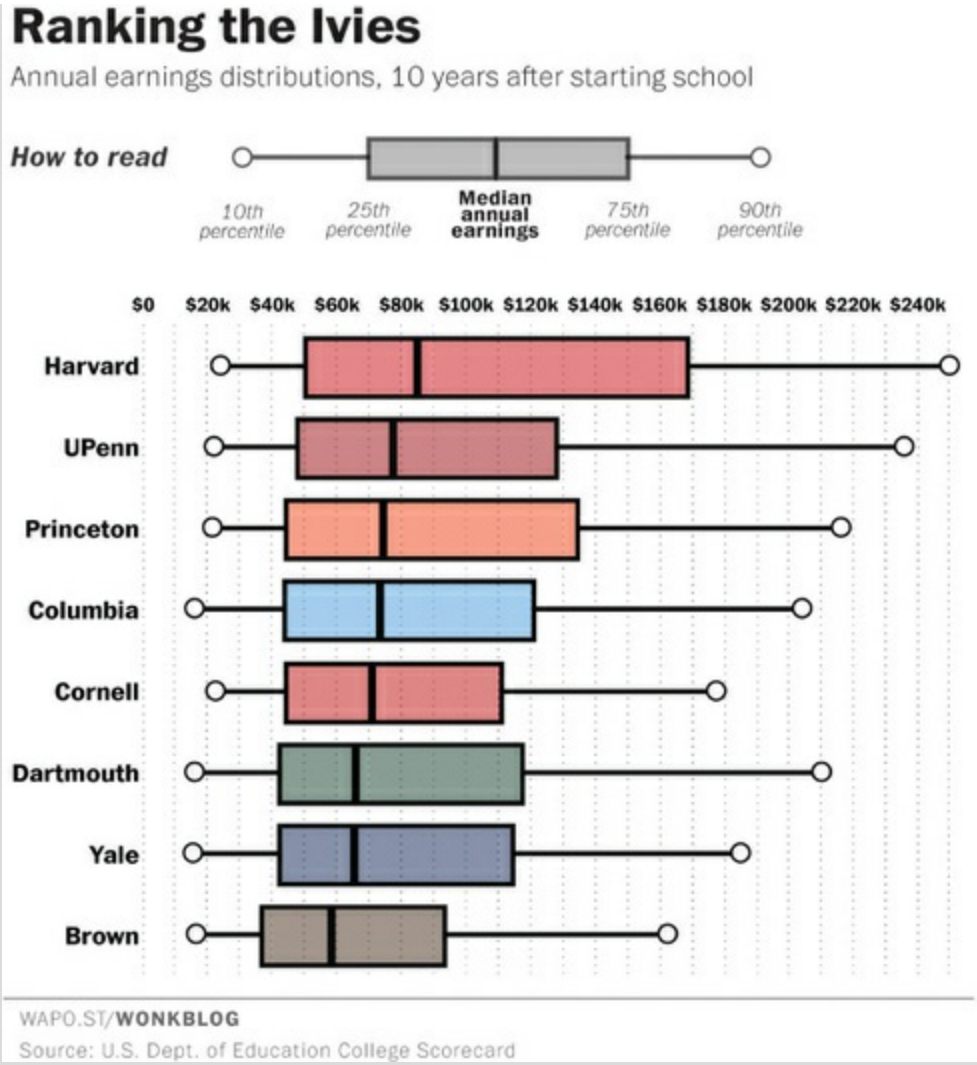
ALSO KNOWN AS Box plot

REPRESENTATION DESCRIPTION

A box-and-whisker plot displays the distribution and shape of a series of quantitative values for different categories. The display is formed by a combination of lines and point markers to indicate (through position and length), typically, five different statistical measures. Three of the statistical values are common to all plots: the first quartile (25th percentile), the second quartile (or median) and the third quartile (75th percentile) values. These are displayed with a box (effectively a wide bar) positioned and sized according to the first and third quartile values with a marker indicating the median. The remaining two statistical values vary in definition: usually either the minimum and maximum values or the 10th and 90th percentiles. These statistical values are represented by extending a line beyond the bottom and top of the main box to join with a point marker indicating the appropriate position. These are the whiskers. A plot will be produced for each major category.

EXAMPLE Comparing the distribution of annual earnings 10 years after starting school for graduates across the eight Ivy League schools.

Figure 6.18 Ranking the Ivies



HOW TO READ IT & WHAT TO LOOK FOR

Begin by looking at the axes so you know with which category each plot is associated and what the range of quantitative values is (min to max). Establish the specific statistics being displayed, by consulting any legends or descriptions, especially in order to identify what the 'whiskers' are representing. Glance across the entire chart to locate the main patterns of spread, identifying any common or noticeably different patterns across categories. Look across the shapes formed for each category to learn about the dispersal of values: starting with the median, then observing the extent and balance of the 'box' (the interquartile range between the 25th and 75th percentiles) and then check the 'whisker' extremes. Is the shape balanced or skewed around the median? Is the interquartile range wide or narrow? Are the whisker extremes far away from the edges of the box? Then return to comparing shapes across all categories to identify more precisely any interesting differences or commonalities for each of the five statistical measures.

PRESENTATION TIPS

ANNOTATION: If you have axis labels you may not need direct labels on each bar – this will lead to label overload, so generally decide between one or the other.

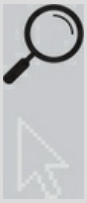
COMPOSITION: The quantitative value axis does not need to commence from zero, unless it means something significant to the interpretation, as the range of values themselves do not necessarily start from zero and the focus is on the statistical properties between the outer values. There is no significant difference in perception between vertical or horizontal box-and-whisker plots, though horizontal layouts tend to make it easier to accommodate and read the category labels. Try to keep a noticeable gap between plots to enable greater clarity in reading. When you have several or many plots in the same chart, where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending order based on

the median value.

VARIATIONS & ALTERNATIVES

Variations involve reducing the number of statistical measures included in the display by removing the whiskers to just show the 25th and 75th percentiles through the lower and upper parts of the box. The 'candlestick chart' (or OHLC chart) involves a similar approach and is often used in finance to show the distribution and milestone values of stock performances during a certain time frame (usually daily), plotting the opening, highest, lowest and closing prices, using colour to indicate an up or down trend.

Charts Distributions



Univariate scatter plot

ALSO KNOWN AS 1D scatter plot, jitter plot

REPRESENTATION DESCRIPTION

A univariate scatter plot displays the distribution of a series of quantitative values for different categories. In contrast to the box-and-whisker plot, which shows selected statistical values, a univariate scatter plot shows all values across a series. For each category, a range of points (typically circles but any 'symbol' is legitimate) are used to mark the position along the scale of the quantitative values. From this you can see the range, the outliers and the clusters and form an understanding about the general shape of the data.

EXAMPLE Comparing the distribution of average critics score (%) from the Rotten Tomatoes website for each movie released across a range of different franchises and movie theme collections.

Figure 6.19 Comparing Critics Scores for Major Movie Franchises

HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know what each scatter row/column relates to in terms of which category it is associated with and what the range of the quantitative values is (min to max). If colour has been used to emphasise or separate different marks, establish what the associations are. Also, learn about how the design depicts multiple marks on the same value – these may appear darker or indeed larger. Glance across the entire chart to observe the main patterns of clustering and identify any noticeable exceptions and/or outliers across all categories. Then look more closely at the patterns within each scatter to learn about each category's specific dispersal of values. Look for empty regions where no quantitative values exist. Estimate the absolute values of specific dots of interest. Where available, compare the quantities against annotated references such as the average or median.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like gridlines can be helpful to increase the accuracy of the reading of the quantitative values. Direct labelling is normally restricted to including values for specifically noteworthy points only.

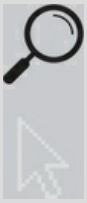
COLOUR: Colour may be used to establish focus of certain points and/or distinction between different sub-category groups to assist with interpretation. When several points have the exact same value you might need to use unfilled or semi-transparent filled circles to facilitate a sense of value density.

COMPOSITION: The representation of the quantitative values is based on position and not size, therefore the quantitative axis does not need to have a zero origin. There is no significant difference in perception between vertical or horizontal arrangement, though horizontal layouts tend to make it easier to accommodate and read the category labels. Where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order.

VARIATIONS & ALTERNATIVES

To overcome occlusion caused by plotting several marks at the same value, a variation of the univariate scatter plot may see the points replaced by geometric areas (like circles), where the position attribute is used to represent a quantitative value along a scale and the size attribute is used to indicate the frequency of observations of similar value. Adding a second quantitative variable axis would lead to the use of a 'scatter plot'.

Charts Distributions



Histogram

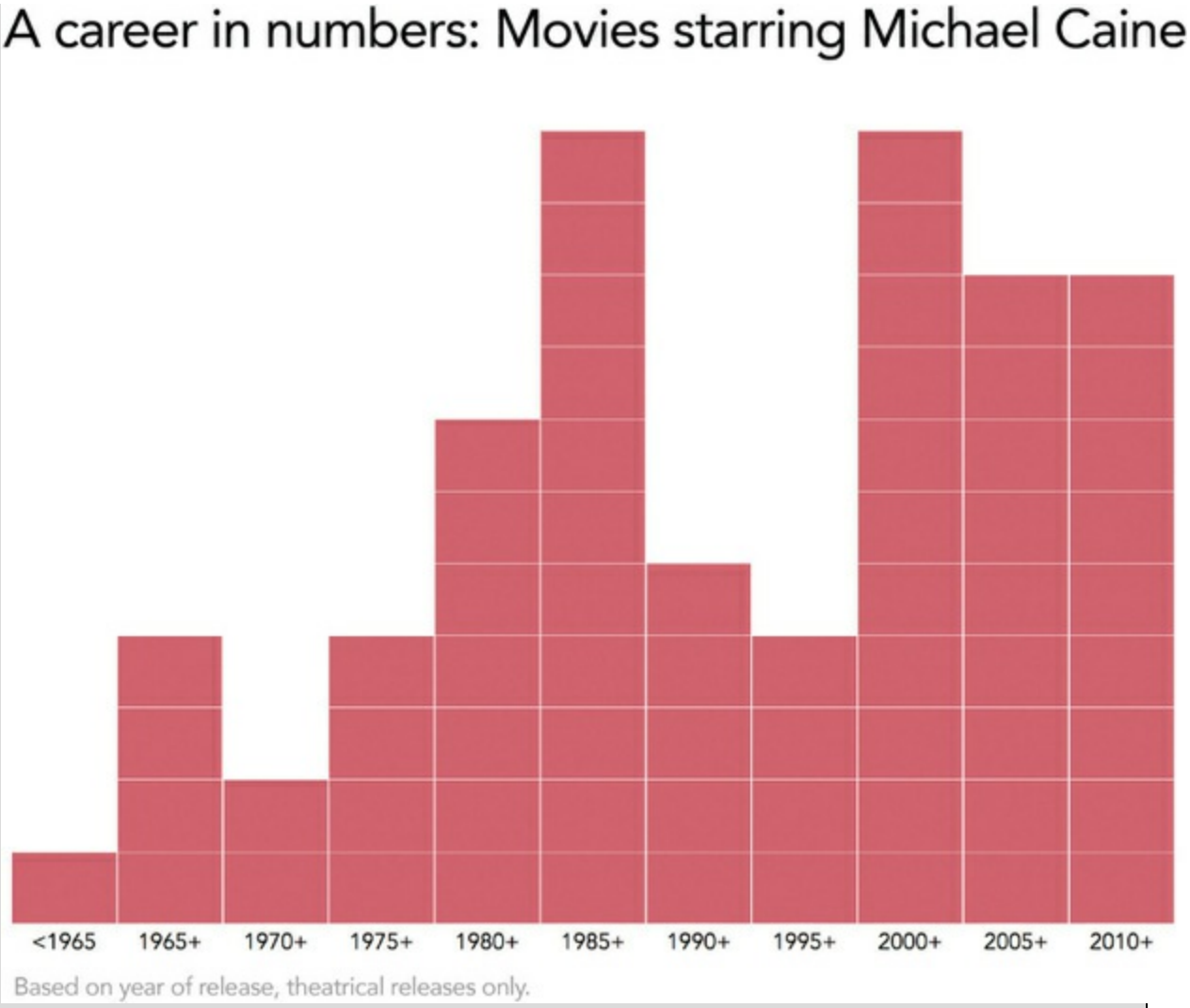
ALSO KNOWN AS Bar chart (wrongly)

REPRESENTATION DESCRIPTION

A histogram displays the frequency and distribution for a range of quantitative groups. Whereas bar charts compare quantities for different categories, a histogram technically compares the number of observations across a range of value 'bins' using the size of lines/bars (if the bins relate to values with equal intervals) or the area of rectangles (if the bins have unequal value ranges) to represent the quantitative counts. With the bins arranged in meaningful order (that effectively form ordinal groupings) the resulting shape formed reveals the overall pattern of the distribution of observations.

EXAMPLE Comparing the distribution of movies released over time starring Michael Caine across five-year periods based on the date of release in the US.

Figure 6.20 A Career in Numbers: Movies Starring Michael Caine



HOW TO READ IT & WHAT TO LOOK FOR

Begin by looking at the axes so you know what the chart depicts in terms of the categorical bins and the range of the quantitative values (zero to max). Glance across the entire chart to establish the main pattern. Is it symmetrically shaped, like a bell or pyramid (around a median or average value)? Is it skewed to the left or right? Does it dip in the middle and peak at the edges (known as bimodal)? Does it have several peaks and troughs? Maybe it is entirely random in its pattern? All these characteristics of 'shape' will inform you about the underlying distribution of the data.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like tick marks and gridlines in particular can be helpful to increase the accuracy of the reading of the quantitative values. Axis labels more than direct value labels tend to be used so as not to crowd the shape of the histogram.

COMPOSITION: Unlike the bar chart there should be no (or at most a very thin) gap between bars to help the collective shape of the frequencies emerge. The sorting of the quantitative bins must be in ascending order so that the reading of the overall shape preserves its meaning. The number of value bins and the range of values covered by each have a prominent influence over the appearance of the histogram and the usefulness of what it might reveal: too few bins may disguise interesting nuances, patterns and outliers; too many bins and the most interesting shapes may be abstracted by noise above signal. There is no singular best approach, the right choice simply arrives through experimentation and iteration.

VARIATIONS & ALTERNATIVES

For analysis that looks at the distribution of values across two dimensions, such as the size of populations

for age across genders, a 'back-to-back histogram' (with male on one side, female on the other), also commonly known as a 'violin plot' or 'population pyramid', is a useful approach to see and compare the respective shapes. A 'box-and-whisker plot' reduces the distribution of values to five key statistical measures to describe key dimensions of the spread of values.

Charts Distributions

Word cloud

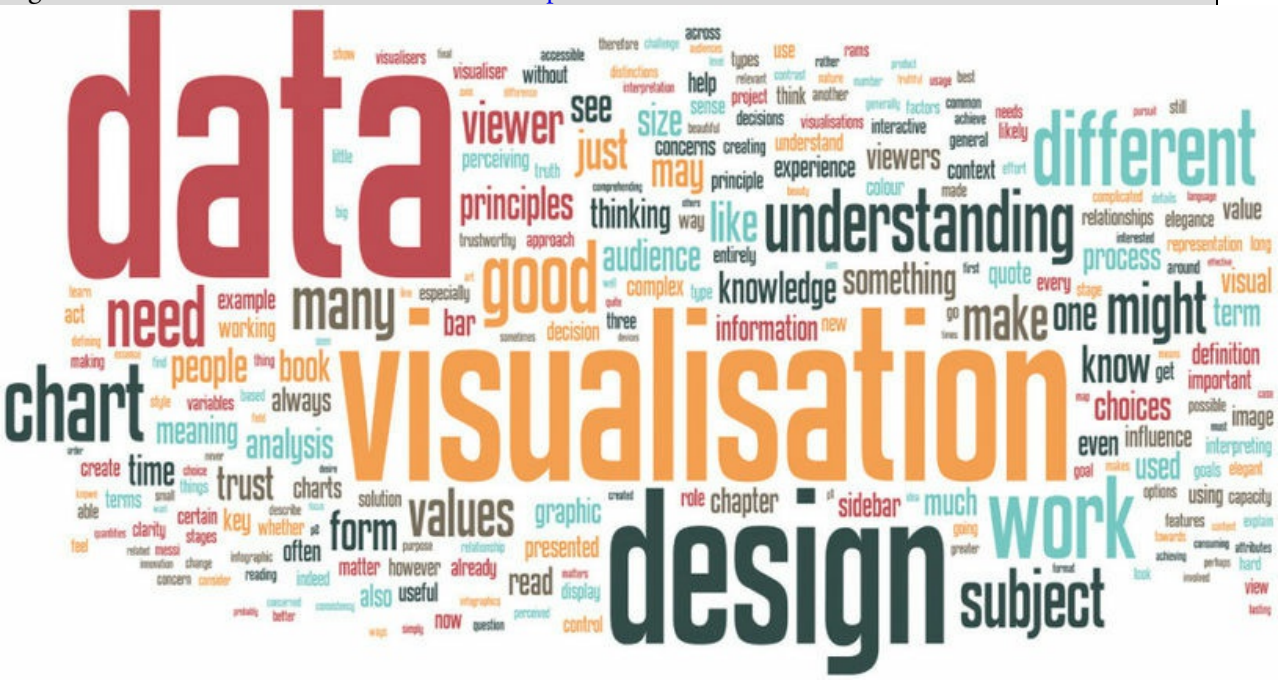
ALSO KNOWN AS Tag cloud

REPRESENTATION DESCRIPTION

A word cloud shows the frequency of individual word items used in textual data (such as tweets, comments) or documents (passages, articles). The display is based around an enclosed cluster of words with the font (not the word length) sized according to the frequency of usage. In modifying the size of font this is effectively increasing the area size of the whole word. All words have a different shape and size so this can make it quite difficult to avoid the prominence of long words, irrespective of their font size. Word clouds are therefore only useful when you are trying to get a quick and rough sense of some of the dominant keywords used in the text. They can be an option for working with qualitative data during the data exploration stage, more so as a means for reporting analysis to others.

EXAMPLE Comparing the frequency of words used in Chapter 1 of this book.

Figure 6.21 Word Cloud of the Text from Chapter 1



HOW TO READ IT & WHAT TO LOOK FOR

The challenge with reading word clouds is to avoid being drawn to the length and/or area of a word – they are simply attributes of the word, not a meaningful representation of frequency. It is the size of the font that you need to focus on. Scan the display to spot the larger text showing the more frequently used words. Consider any words of specific interest to see if you can find them; if they are not significantly visible, that in itself could be revealing. While most word cloud generators will dismiss many irrelevant words, you might still need to filter out perceptually the significance of certain dominantly sized text.

PRESENTATION TIPS

INTERACTIVITY: Interactive features that let users interrogate, filter and scrutinise the words in more depth, perhaps presenting examples of their usage in a passage, can be quite useful to enhance the value of a word cloud.

ANNOTATION: While the absolutes are generally of less interest than relative comparisons, to help viewers get as much out of the display as possible a simple legend explaining how the font size equates to frequency number can be useful.

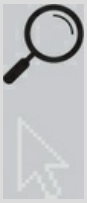
COLOUR: Colours may be used as redundant encoding to accentuate further the larger frequencies or categorically to create useful visual separation.

COMPOSITION: The arrangement of the words within a word cloud is typically based on a layout process. Although not random, this will generally prioritise the placement of words to occupy optimum collective space that preserves an overall shape (with essentially a central gravity) over and above any arrangement that might better enable direct comparison.

VARIATIONS & ALTERNATIVES

The alternative approach would be to use any other method in this categorical family of charts that would more usefully display the counts of text, such as a bar chart.

Charts Part-to-whole



Pie chart

ALSO KNOWN AS Pizza chart

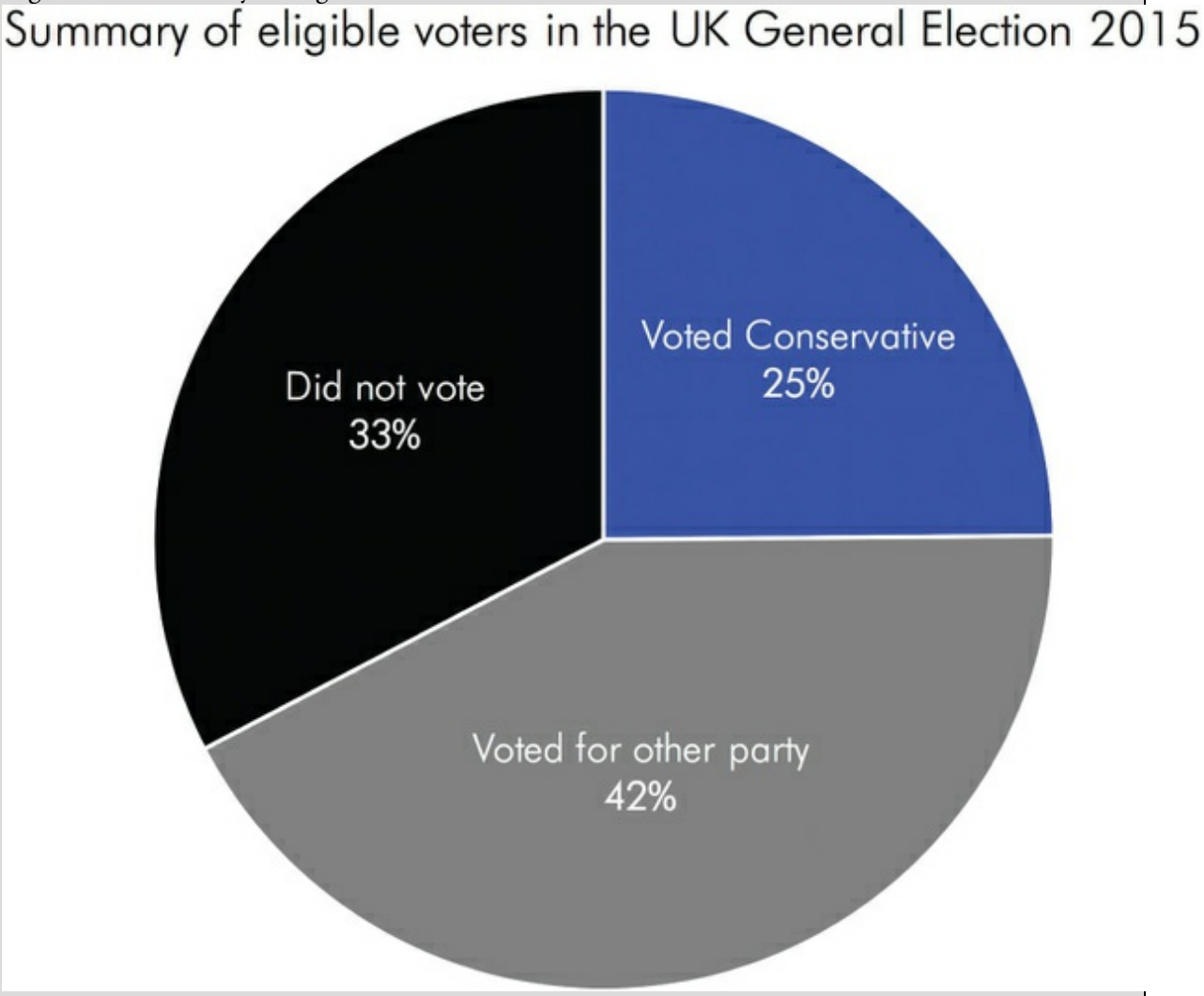
REPRESENTATION DESCRIPTION

A pie chart shows how the quantities of different constituent categories make up a whole. It uses a circular display divided into sectors for each category, with the angle representing each of the percentage proportions. The resulting size of the sector (in area terms) is a spatial by-product of the angle applied to each part and so offers an additional means for judging the respective values. The role of a pie chart is primarily about being able to compare a part to a whole than being able to compare one part to another part. They therefore work best when there are only two or three parts included. There are a few important rules for pie charts. Firstly, the total percentage values of all sector values must be 100%; if the aggregate is greater than or less than 100% the chart will be corrupted. Secondly, the whole has to be meaningful – often

people just add up independent percentages but that is not what a pie chart is about. Finally, the category values must represent exclusive quantities; nothing should be counted twice or overlap across different categories. Despite all these warnings, do not be afraid of the pie chart – just use it with discretion.

EXAMPLE Comparing the proportion of eligible voters in the 2015 UK election who voted for the Conservative Party, for other parties and who did not vote.

Figure 6.22 Summary of Eligible Votes in the UK General Election 2015



HOW TO READ IT & WHAT TO LOOK FOR

Begin by establishing which sectors relate to what categories. This may involve referring to a colour key legend or through labels directly adjacent to the pie. Quickly scan the pie to identify the big, medium and small sectors. Notice if there is any significance behind the ordering of the parts. Unless there are value labels, you next will attempt to judge the individual sector angles. This usually involves mentally breaking the pie into 50% halves (180°) or 25% quarters (90°) and using those guides to perceptually measure the category values. Comparing parts against other parts with any degree of accuracy will only be possible once you have formed estimates of the individual sector sizes. If you are faced with the task of judging the size of many parts it is quite understandable if you decide to give up quite soon.

PRESENTATION TIPS

ANNOTATION: The use of local labelling for category values can be useful but too many labels can become cluttered, especially when attempting to label very small angled sectors.

COLOUR: Colour is generally vital to create categorical separation and association of the different sectors so aim to use the difference in colour hue and not colour saturation to maximise the visible difference.

COMPOSITION: Positioning the first slice at the vertical 12 o'clock position gives a useful baseline to help judge the first sector angle value. The ordering of sectors using descending values or ordinal characteristics helps with the overall readability and allocation of effort. Do not consider using gratuitous decoration (like 3D, gradient colours, or exploding slices).

VARIATIONS & ALTERNATIVES

Sometimes a pie chart has a hole in the centre and is known as a 'doughnut chart', continuing the food-related theme. The function is exactly the same as a pie but the removal of the centre, often to accommodate a labelling property, removes the possibility of the reader judging the angles at the origin. One therefore has to derive the angles from the resulting arc lengths. If you want to display multiple parts (more than three) the bar chart will be a better option and, for many parts, the 'treemap' is best. Depending on the allocated space, a 'stacked bar chart' may provide an alternative to the pie. Unlike most chart types, the pie chart does not work well in the form of small multiples (unless there is only a single part being displayed). A 'nested shape chart', typically based on embedded square or circle areas, enables comparison across a series of one-part-to-whole relationships based on absolute numbers, rather than percentages, where the wholes may vary in size.

Charts Part-to-whole

Waffle chart

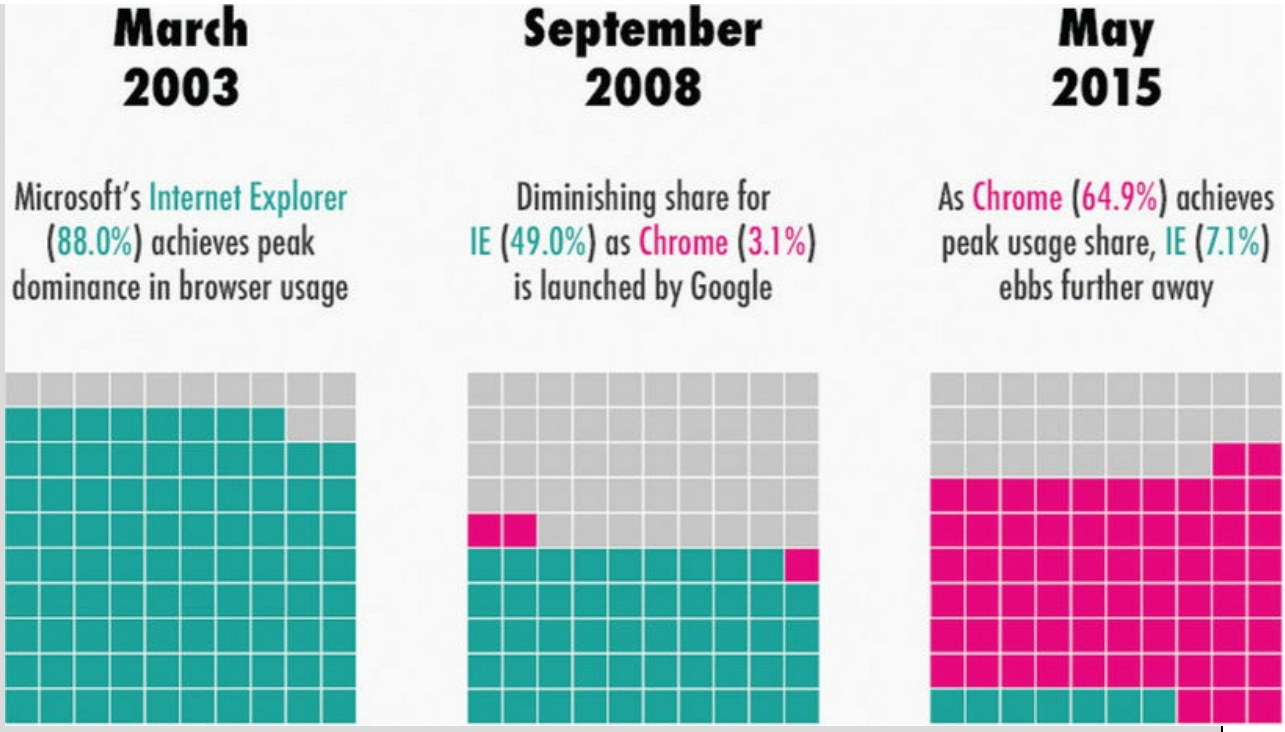
ALSO KNOWN AS Square pie, unit chart, 100% stacked shape chart

REPRESENTATION DESCRIPTION

A waffle chart shows how the quantities of different constituent categories make up a whole. It uses a square display usually representing 100 point 'cells' through a 10 × 10 grid layout. Each constituent category proportion is displayed through colour-coding a proportional number of cells. Difference in symbol can also be used. The role of the waffle chart is to simplify the counting of proportions in contrast to the angle judgements of the pie chart, though the display is limited to rounded integer values. This is easier when the grid layout facilitates quick recognition of units of 10. As with the pie chart, the waffle chart works best when you are showing how a single part compares to the whole and perhaps offers greater visual impact when there are especially small percentages of a whole. Rather than just colouring in the grid cells, sometimes different symbols will be used to associate with different categories. For example, you might see figures or gender icons used to show the makeup of a given sample population.

EXAMPLE Comparing the proportion of total browser usage for Internet Explorer and Chrome across key milestone moments.

Figure 6.23 The Changing Fortunes of Internet Explorer and Google Chrome



HOW TO READ IT & WHAT TO LOOK FOR

Begin by establishing how the different shapes or colours are associated with different categories. Assess the grid layout to understand the dimension of the chart and the quantity of cell 'units' forming the display (e.g. is it a 10 x 10 grid?). Quickly scan the chart to identify the big, medium and small sectors. Notice if there is any significance behind the ordering of the parts. Unless there are value labels, you will need to count/estimate the number of units representing each category value. Comparing parts against other parts will only be possible once you have established the individual part sizes. If several related waffle charts are shown, possibly for different categories or points in time, identify the related colours/shapes in each chart and establish the patterns of size between and across the various charts, looking for trends, declines and general differences.

PRESENTATION TIPS

ANNOTATION: Direct labelling can become very cluttered and hard to incorporate elegantly without the need for long arrows.

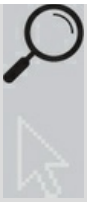
COLOUR: Borders around each square cell are useful to help establish the individual units, but do not make the borders too thick to the point where they dominate attention.

COMPOSITION: Always start each row of values from the same side, for consistency and to make it easier for people to estimate the values. When you have several parts in the same waffle chart, where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order or based on a logical categorical order.

VARIATIONS & ALTERNATIVES

Sometimes the waffle chart approach is used to show stacks of absolute unit values and indeed there are overlaps in concept between this variation in the waffle chart and potential applications of the pictogram. Aside from the pie chart, a 'nested shape chart' will provide an alternative way of showing a part-to-whole relationship while also occupying a squarified layout.

Charts Part-to-whole



Stacked bar chart

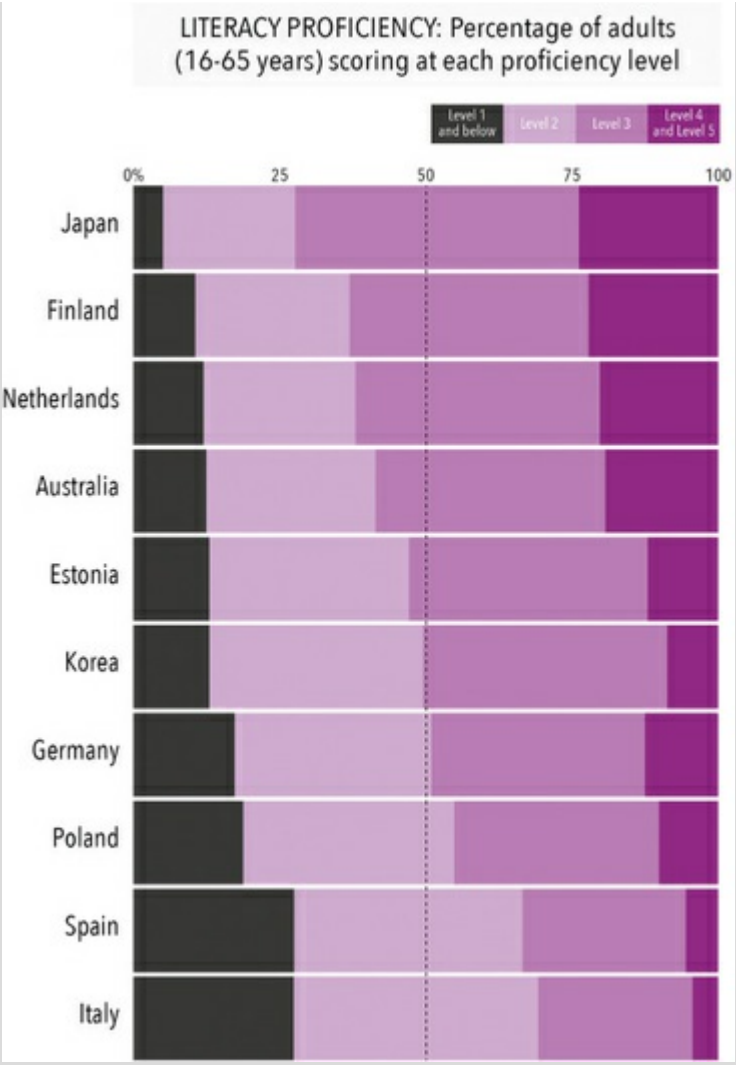
ALSO KNOWN AS

REPRESENTATION DESCRIPTION

A stacked bar chart displays a part-to-whole breakdown of quantitative values for different major categories. The percentage proportion of each categorical dimension or 'part' is represented by separate bars, distinguished by colour, that are sized according to their proportion and then stacked to create the whole. Sometimes the whole is standardised to represent 100%, at other times the whole will be representative of absolute values. Stacked bar charts work best when the parts are based on ordinal dimensions, which enables ordering of the parts within the stack to help establish the overall shape of the data. If the parts are representative of nominal data, it is best to keep the number of constituent categories quite low, as estimating the size of individual stacked parts when there are many becomes quite hard.

EXAMPLE Comparing the percentage of adults (16–65 year olds) achieving different proficiency levels in literacy across different countries.

Figure 6.24 Literarcy Proficiency: Adult Levels by Country



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with what major categorical values each bar is associated and what the quantitative values are, determining if it is a 100% stacked bar or an absolute stacked bar (in which case identify the min and the max). Establish the colour association to understand what categories the bars within each stack represent. Glance across the entire chart. If the categorical data is ordinal, and the sorting/colour of the stacks is intuitive, you should be able to derive meaning from the overall balance of colour patterns, especially where any annotated gridlines help to guide your value estimation. If the categorical data is nominal, seek to locate the dominant colours and the least noticeable ones. Comparing across different stacked bars is made harder by the lack of a common baseline for anything other than the bottom stack on the zero baseline (and for 100% stacked bars, those final ones at the top) and so a general sense of magnitude will be your focus. Study closer the constituent parts within each stack to establish the high-level ranking of biggest > smallest. Estimate (or read, if labels are present) the absolute values of specific stacked parts of interest.

PRESENTATION TIPS

ANNOTATION: Direct value labelling can become very cluttered when there are many parts or stacks and you are comparing several different major categories. You might be better with a table if that is your aim. Definitely include value axis labels with logical intervals and it is very helpful to annotate, through gridlines, key units such as the 25%, 50% and 75% positions when based on a 100% stacked bar chart.

COLOUR: If you are representing categorical ordinal data, colour can be astutely deployed to give a sense of the general balance of values within the whole, but this will only work if their sorting arrangement within the stack is logically applied. For categorical nominal data, ensure the stacked parts have sufficiently

different colours so that their distinct bar lengths can be efficiently observed.

COMPOSITION: Across the main categories, once again consider the optimum sorting option, maybe organising values in ascending/descending size order or based on a logical categorical order. Judging the size of the stacks with accuracy is harder for those that are not on the zero baseline, so maybe consider which ones are of most importance to be more easily read and place those on the baseline.

VARIATIONS & ALTERNATIVES

The main alternative would be to use 'multi-panel bar charts', where separate bar charts each include just one 'stack'/part and they are then repeated for each subsequent constituent category. In the world of finance the 'waterfall chart' is a common approach based on a single stacked bar broken up into individual elements, almost like a step-by-step narrative of how the components of income look on one side and then how the components of expenditure look on the other, with the remaining space representing the surplus or deficit. Like their unstacked siblings, stacked bar charts can also be used to show how categorical composition has changed over time.

Charts Part-to-whole

Back-to-back bar chart

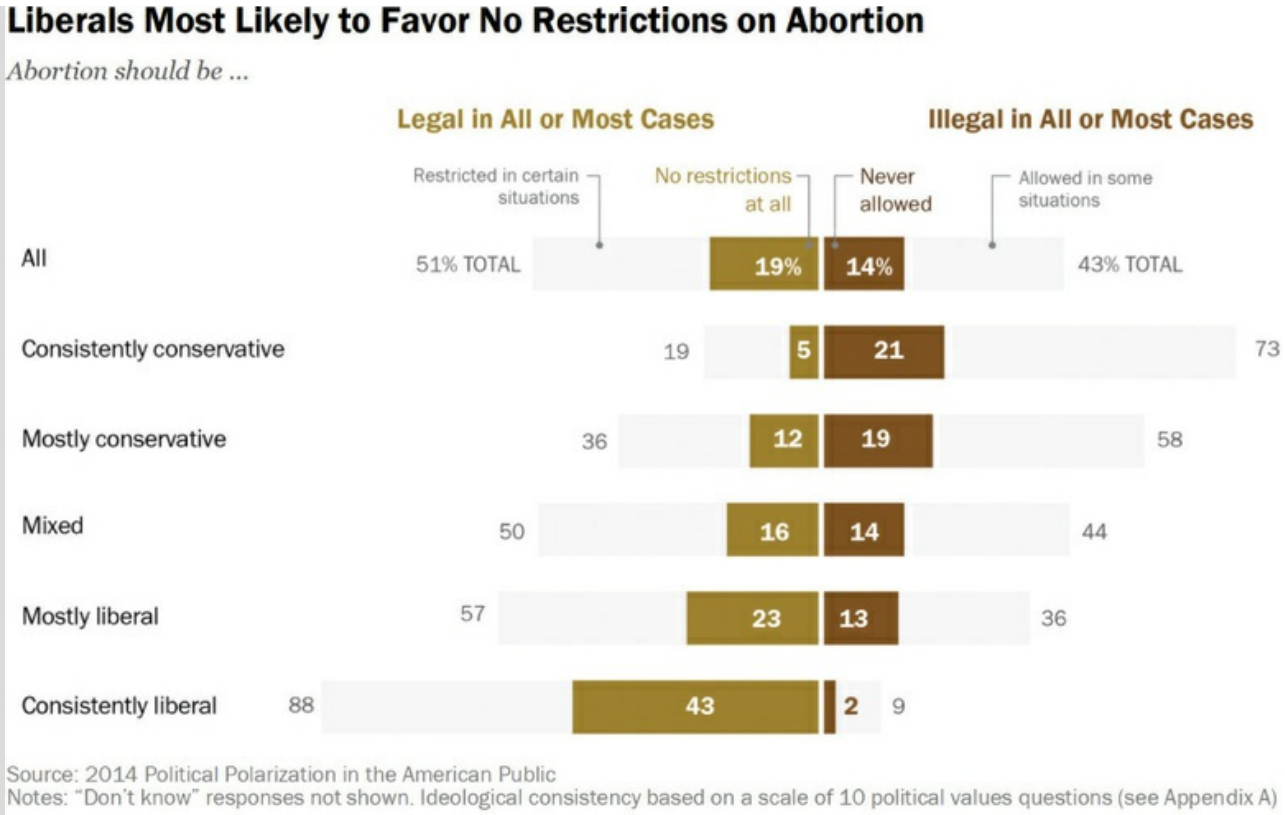
ALSO KNOWN AS Paired bar chart

REPRESENTATION DESCRIPTION

A back-to-back bar chart displays a part-to-whole breakdown of quantitative values for different major categories. As with any bar chart, the length of a bar represents a quantitative proportion or absolute value for each part and across all major categories. In contrast to the stacked bar chart, where the constituent bars are simply stacked to form a whole, in a back-to-back bar chart the constituent parts are based on diverging categorical dimensions with a 'directional' essence such as yes/no, male/female, agree/disagree. The values for each dimension are therefore presented on opposite sides of a shared zero baseline to help reveal the shape and contrast differences across all major categories.

EXAMPLE Comparing the responses to a survey question asking for opinions about 'the government collection of telephone and Internet data as part of anti-terrorism efforts' across different demographic categories.

Figure 6.25 Political Polarization in the American Public



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with which major categorical values each bar is associated and what the range of the quantitative values is (min to max). Establish what categorical dimensions are represented by the respective sides of the display and any colour associations. Glance across the entire chart to locate the big, small and medium bars and perform global comparisons to establish the high-level ranking of biggest > smallest. Repeat this for each side of the display, noticing any patterns of dominance of larger values on either side. Identify any noticeable exceptions and/or outliers. Perform local comparisons for each category value to estimate the relative sizes (or read, if labels are present) of each bar.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like tick marks and gridlines in particular can be helpful to increase the accuracy of the reading of the quantitative values.

COLOUR: The bars either side of the axis do not need to be coloured but often are to create further visual association.

COMPOSITION: The quantitative value axis should always start from the origin value of zero: a bar should be representative of the true, full quantitative value, nothing more, nothing less, otherwise the perception of bar sizes will be distorted when comparing relative sizes. There is no significant difference in perception between vertical or horizontal bars, though horizontal layouts tend to make it easier to accommodate and read the category labels. Where possible try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order or based on a logical categorical order.

VARIATIONS & ALTERNATIVES

Back-to-back bar charts facilitate a general sense of the shape of diverging categorical dimensions. However, if you want to facilitate direct comparison, a 'clustered bar chart' showing adjacent bars helps to compare respective heights more precisely. For analysis that looks at the distribution values across two dimensions, such as the size of populations for age across genders, a 'back-to-back histogram' (with male on one side, female on the other), also commonly known as a 'violin plot' or 'population pyramid', is a useful approach to

see and compare the respective shapes. Some back-to-back applications do not show a part-to-whole relationship but simply compare quantities for two categorical values. Further variations may appear as 'back-to-back area charts' showing mutual change over time for two contrasting states.

Charts Part-to-whole

Treemap

ALSO KNOWN AS Heat map (wrongly)

REPRESENTATION DESCRIPTION

A treemap is an enclosure diagram providing a hierarchical display to show how the quantities of different constituent parts make up a whole. It uses a contained rectangular layout (often termed 'squarified') representing the 100% total divided into proportionally sized rectangular tiles for each categorical part. Colour can be used to represent an additional quantitative measure, such as an indication of amount of change over a time period. The absolute positioning and dimension of each rectangle is organised by an underlying tiling algorithm to optimise the overall space usage and to cluster related categories into larger rectangle-grouped containers. Treemaps are most commonly used, and of most value, when there are many parts to the whole but they are only valid if the constituent units are legitimately part of the same 'whole'.

EXAMPLE Comparing the relative value of and the daily performance of stocks across the S&P 500 index grouped by sectors and industries.

Figure 6.26 FinViz: Standard and Poor's 500 Index

HOW TO READ IT & WHAT TO LOOK FOR

Look at the high-level groupings to understand the different containing arrangements and establish what the colour association is. Glance across the entire chart to seek out the big, small and medium individual rectangular sizes and perform global comparisons to establish a general ranking of biggest > smallest values. Also identify the largest through to smallest container group of rectangles. If the colour coding is based on quantitative variables, look out for the most eye-catching patterns at the extreme end of the scale(s). If labels are provided (or offered through interactivity), browse around the display looking for categories and values of specific interest. As with any display based on the size of the area of a shape, precise reading of values is hard to achieve and so it is important to understand that treemaps can only aim to provide a single-view gist of the properties of the many components of the whole.

PRESENTATION TIPS

INTERACTIVITY: Typically, a treemap will be presented with interactive features to enable selection/mouseover events to reveal further annotated details and/or drill-down navigation.

ANNOTATION: Group/container labels are often allocated a cell of space but these are not to be read as proportional values. Effective direct value labelling becomes difficult as the rectangles get smaller, so often only the most prominent values might be annotated. Interactive features will generally offer visibility of the relevant labels where possible.

COLOUR: Colour can also be used to provide further categorical grouping distinction if not already assigned to represent a quantitative measure of change.

COMPOSITION: As the tiling algorithm is focused on optimising the dimensions and arrangement of the rectangular shapes, treemaps may not always be able to facilitate much internal sorting of high to low values. However, generally you will find the larger shapes appear in the top left of each container and work outwards towards the smaller constituent parts.

VARIATIONS & ALTERNATIVES

A variation of the treemap sees the rectangular layout replaced by a circular one and the rectangular tiles replaced by organic shapes. These are known as 'Voronoi treemaps' as the tiling algorithm is informed by a

Voronoi tessellation. The 'circle packing diagram', a variation of the 'bubble chart', similarly shows many parts to a whole but uses a non-tessellating circular shape/layout. The 'mosaic plot' or 'Marimekko chart' is similar in appearance to a treemap but, in contrast to the treemap's hierarchical display, presents a detailed breakdown of quantitative value distributions across several categorical dimensions, essentially formed by varied width stacked bars.

Charts Part-to-whole

Venn diagram

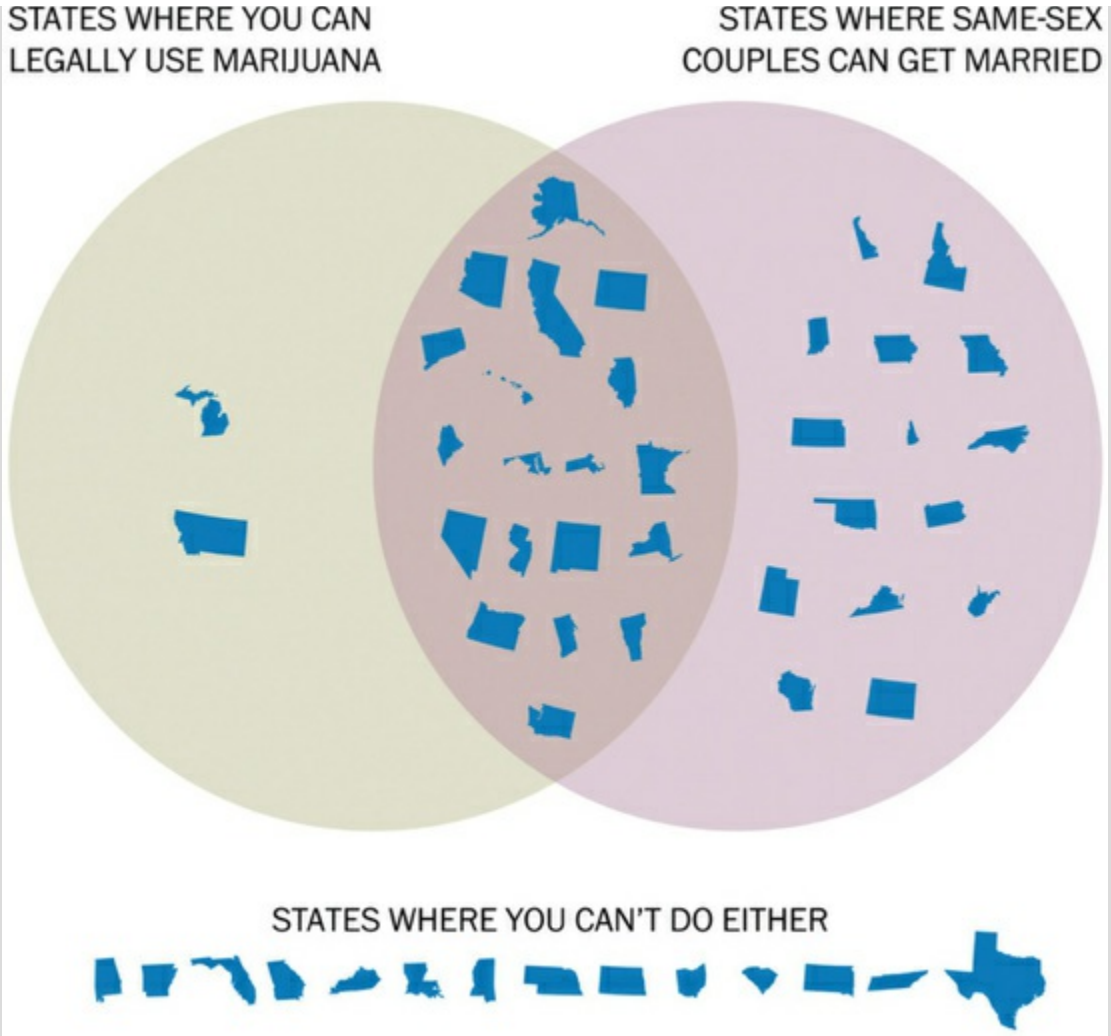
ALSO KNOWN AS Set diagram, Euler diagram (wrongly)

REPRESENTATION DESCRIPTION

A Venn diagram shows collections of and relationships between multiple sets. They typically use round or elliptical containers to represent all different 'membership' permutations to include all independent and intersecting containers. The size of the contained area is (typically) not important: what is important is in which containing region a value resides, which may be represented through the mark of a text label or 'point'.

EXAMPLE Comparing sets of permutations for legalities around marijuana usage and same-sex marriage across states of the USA.

Figure 6.27 This Venn Diagram Shows Where You Can Both Smoke Weed and Get a Same-Sex Marriage



HOW TO READ IT & WHAT TO LOOK FOR

To read a Venn diagram firstly establish what the different containers are representative of in terms of their membership. Assess the membership of the intersections (firstly 'all', then 'partial' intersections when involving more than two sets) then work outwards towards the independent container regions where values are part of one set but not part of others. Occasionally there will be a further grouping state outside of the containers that represents values that have no membership with any set at all.

PRESENTATION TIPS

ANNOTATION: Unless you are using point markers to represent membership values, clear labels are vital to indicate how many or which elements hold membership with each possible set combination.

COLOUR: Colour is often used to create more immediate distinction between the intersections and independent parts or members of each container.

COMPOSITION: As the attributes of size and shape of the containers are of no significance there is more flexibility to manipulate the display to fit the number of sets around the constraint of real estate you are facing and to get across the set memberships you are attempting to show. The complexity of creating containers to accommodate all combinations of intersection and independence states increases as the number of sets increases, especially to preserve all possible combinations of intersections between and independencies from all sets. As the number of sets increases, the symmetry of shape reduces and the circular containers are generally replaced with ellipses. While it is theoretically possible to exceed four and five set diagrams, the ability of readers to make sense of the displays diminishes and so they commonly involve only two or three different sets.

VARIATIONS & ALTERNATIVES

A common variation or alternative to the Venn (but often mistakenly called a Venn) is the 'Euler diagram'. The difference is that an Euler diagram does not need to present all possible intersections with and independencies from all sets. A different approach to visualising sets (especially larger numbers) can be achieved using the 'UpSet' technique.

Charts Hierarchies

Dendrogram

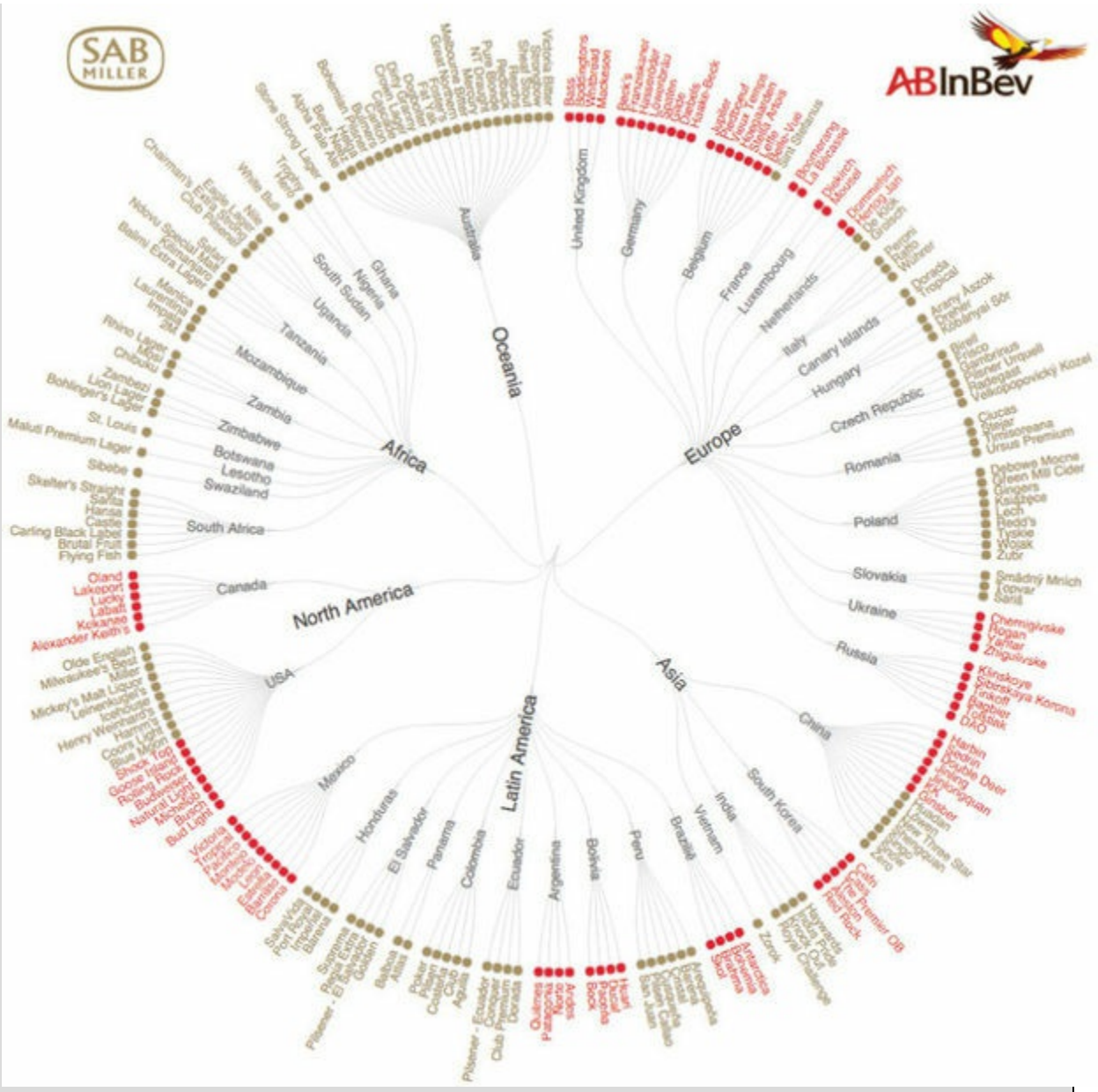
ALSO KNOWN AS Node–link diagram, layout tree, cluster tree, tree hierarchy

REPRESENTATION DESCRIPTION

A dendrogram is a node–link diagram that displays the hierarchical relationship across multiple tiers of categorical dimensions. It displays a hierarchy based on multi-generational 'parent-and-child' relationships. Starting from a singular origin root node (or 'parent') each subsequent set of constituent 'child' nodes, a tier below and represented by points, is connected by lines (curved or straight) to indicate the existence of a relationship. Each constituent node may have further sub-constituencies represented in the same way, continuing down through to the lowest tier of detail. Each 'generational' tier is presented at the same relative distance from the origin. The layout can be based on either a linear tree structure (typically left to right) or radial tree (outwards from the centre).

EXAMPLE Showing a breakdown of the 200+ beer brands belonging to SAB InBev across different countries grouped by continent.

Figure 6.28 The 200+ Beer Brands of SAB InBev



HOW TO READ IT & WHAT TO LOOK FOR

Reading a dendrogram will generally be a highly individual experience based on your familiarity with the subject and your interest in exploring certain hierarchical pathways. The main focus of attention will likely be to find the main clusters from where most constituent parts branch out and to contrast these with the thinner, lighter paths comprising fewer parts. Work left to right (linear) or in to out (radial) through the different routes that stoke your curiosity.

PRESENTATION TIPS

ANNOTATION: With labelling required for each node, depending on the number of tiers and the amount of nodes, the size of the text will need to be carefully considered to ensure readability and minimise the effect of clutter.

COLOUR: Colour would be an optional choice for accentuating certain nodes or applying some further visual categorisation.

COMPOSITION: There are several different layout options to display tree hierarchies like the dendrogram. The common choice is a cluster layout based on the 'Reingold–Tilford' tree algorithms that offers a tidying and optimisation treatment for the efficiency of the arrangement of the nodes and connections. The sequencing of sub-constituencies under each node could be logically arranged in some

more meaningful way than just alphabetical, though the cataloguing nature of A–Z may suit your purpose. The choice of a linear or radial tree structure will be informed largely by the space you have to work in as well as by the cyclical or otherwise nature of the content in your data. The main issue is likely to be one of legibility if and when you have numerous layers of divisions and many constituent parts to show in a single view.

VARIATIONS & ALTERNATIVES

More advanced applications of dendrograms are used to present hierarchical clustering (in fields such as computational biology) and apply more quantitative meaning to the length of the links and the positioning of the nodes. The 'tree hierarchy diagram' offers a similar tree structure but introduces quantitative attributes to the nodes using area marks, such as circles, sized according to a quantitative value. An alternative approach to the dendrogram could involve a 'linear bracket'. This might show hierarchical structures for data-related sporting competitions with knock-out format. The outer nodes would be the starting point representing all the participating competitors/teams. Each subsequent tier would represent those participants who progressed to the next round, continuing through to the finalists and eventual victors.

Charts Hierarchies

Sunburst

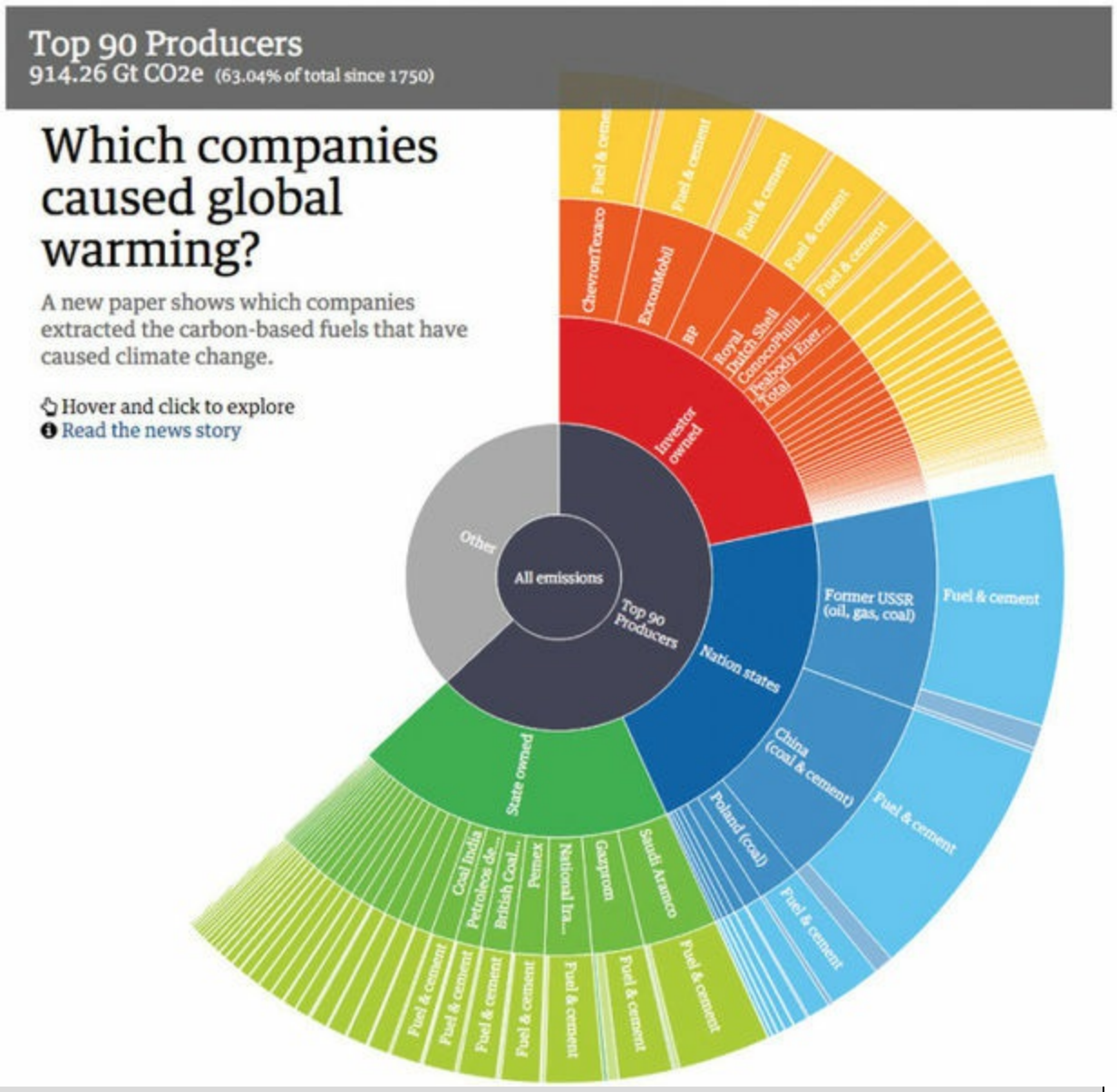
ALSO KNOWN AS Adjacency diagram, icicle chart, multi-level pie chart

EXAMPLE Showing a breakdown of the types of companies responsible for extracting different volumes of carbon-based fuels through various activities.

REPRESENTATION DESCRIPTION

A sunburst chart is an adjacency diagram that displays the hierarchical and part-to-whole relationships across multiple tiers of categorical dimensions. In contrast to the dendrogram, the sunburst uses layers of concentric rings, one layer for each generational tier. Each ring layer is divided into parts based on the constituent categorical dimensions at that tier. Each part is represented by a different circular arc section that is sized (in length; width is constant) according to the relative proportion. Starting from the centre 'parent' tier, the outward adjacency of the constituent parts of each tier represents the 'parent-and-child' hierarchical composition.

Figure 6.29 Which Fossil Fuel Companies are Most Responsible for Climate Change?



HOW TO READ IT & WHAT TO LOOK FOR

Reading a sunburst chart will be a highly individual experience based on your familiarity with the subject and your interest in exploring certain hierarchical pathways. The main focus of attention will likely be to find the largest arc lengths, representing the largest single constituent parts, and those layers or tiers with the most constituent parts. Work from the centre outwards through the different routes that stoke your curiosity. Depending on the deployment of colour, this may help you identify certain additional categorical patterns.

PRESENTATION TIPS

INTERACTIVITY: Often interactive mouseover/selection events are the only way to reveal the annotations here.

ANNOTATION: Labelling can be quite difficult to fit into the narrow spaces afforded by small proportion 'parts'. If interactivity is not an option you may decide to label only those parts that can accommodate the text space.

COLOUR: Colours are often used to achieve further categorical distinction.

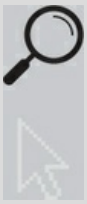
COMPOSITION: Sometimes the parent–child (and other generational) relationships could be legitimately reversed, so decisions need to be made about the best hierarchy sequencing to suit the

curiosities of the audience. The sequencing of sub-constituencies under each node could also be logically arranged in a meaningful way, more so than just alphabetical, unless the cataloguing nature of A–Z ordering suits your purpose.

VARIATIONS & ALTERNATIVES

Where the sunburst chart uses a radial layout, the 'icicle chart' uses a vertical, linear layout starting from the top and moving downwards. The choice of a linear or radial tree structure will be informed largely by the space you have to work in as well as by the legitimacy of the cyclical nature of the content in your data. A variation on the sunburst chart would be the 'ring bracket'. This might show a reverse journey for hierarchical data based on something like sporting competitions with knock-out formats. The outer concentric partitions would represent the participant competitors/teams at the start of the process. The length of these arc line parts would be equally distributed across all constituent parts with each subsequent tier representing 'participants' who progress forward to the next 'round', continuing through to the finalists and eventual victors in the centre.

Charts Correlations



Scatter plot chart

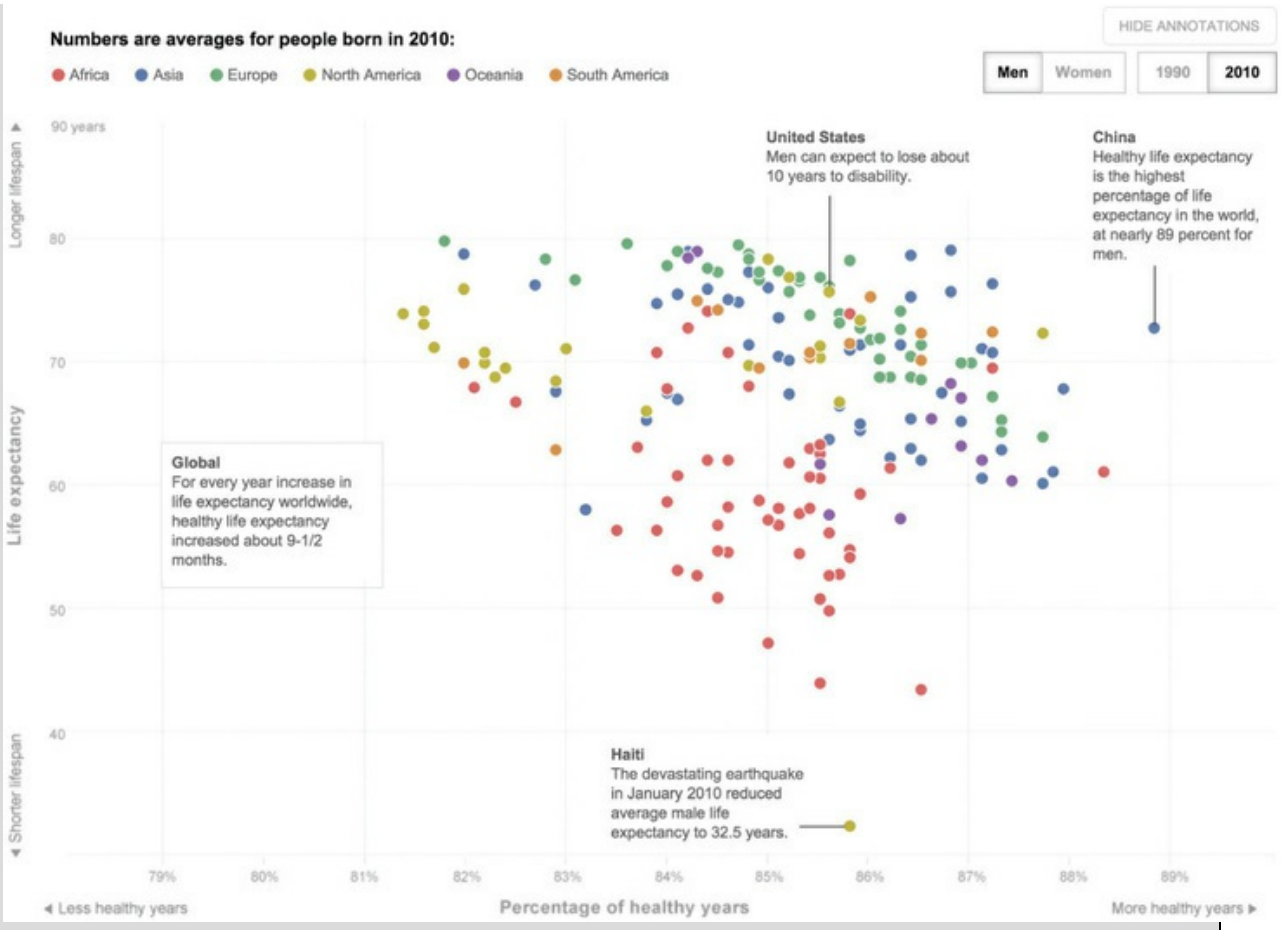
ALSO KNOWN AS Scatter graph

REPRESENTATION DESCRIPTION

A scatter plot displays the relationship between two quantitative measures for different categories. Scatter plots are used to explore visually the potential existence, extent or absence of a significant relationship between the plotted variables. The display is formed by points (usually a dot or circle), representing each category and plotted positionally along quantitative x- and y-axes. Sometimes colour is used to distinguish categorical dimensions across all the points. Scatter plots do not work too well if one or both of the quantitative measures has limited variation in value as this especially causes problems of 'occlusion', whereby multiple instances of the similar values are plotted on top of each other and essentially hidden from the reader.

EXAMPLE Exploring the relationship between life expectancy and the percentage of healthy years across all countries.

Figure 6.30 How Long Will We Live — And How Well?



HOW TO READ IT & WHAT TO LOOK FOR

Learn what each quantitative axis relates to and make a note of the range of values in each case (min to max). Look at what category or observation each plotted value on the chart refers to and look up any colour associations being used for categorical distinction. Scan the chart looking for the existence of any diagonal trends that might suggest a linear correlation between the variables, or note the complete absence of any pattern, to mean no correlation. Annotations will often assist in determining the significance of any patterns like this. Identify any clusters of points and also look at the gaps, which can be just as revealing. Some of the most interesting observations come from individual outliers standing out separately from others. Look out for any patterns formed by points with similar categorical colour. One approach to reading the 'meaning' of the plotted positions involves trying to break down the chart area into a 2 × 2 grid translating what marks positioned in those general areas might mean – which corner is 'good' or 'bad' to be located in? Remember that ruling out significant relationships can be just as useful as ruling them in.

PRESENTATION TIPS

ANNOTATION: Gridlines can be useful to help make the value estimates clearer and reference lines (such as a trend line of best fit) might aid interpretation. It is usually hard to make direct labelling of all values work well. Firstly, it can be tricky making it clear which value relates to which point, especially when several points may be clustered together. Secondly, it creates a lot of visual clutter. Labelling choices should be based on values that are of most interest to include editorially unless interactive features enable annotations to be revealed through selection or mouseover events. If possible, you might consider putting a number inside the marker to indicate a count of the number of points at the same position if this occurs.

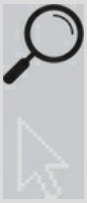
COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. On the occasion where multiple values may be plotted close to or on top of each other, you might need to use semi-transparency to enable overlapping of points to build up a recognisably darker colour compared to other points, indicating an underlying stack of values at the same location on the chart.

COMPOSITION: As the encoding of the plotted point values is based on position along an axis, it is not necessary to start the axes from a zero baseline, so just make the scale ranges as representative as possible of the range of values being plotted. Ideally a scatter plot will have a 1:1 aspect ratio (equally as tall as it is wide), creating a squared area to help patterns surface more evidently. If one quantitative variable (e.g. weight) is likely to be affected by the other variable (e.g. height), it is general practice to place the former on the y-axis and the latter on the x-axis. If you have to use a logarithmic quantitative scale on either or both axes, you need to make this clear to readers so they avoid making incorrect conclusions from the resulting patterns (that might imply correlation if the values were linear, for example).

VARIATIONS & ALTERNATIVES

A 'ternary plot' is a variation of the scatter plot through the inclusion of a third quantitative variable axis. The 'bubble plot' also incorporates a third quantitative variable, this time through encoding the size of a geometric shape (replacing the point marker). A 'scatter plot matrix' involves a single view of multiple scatter plots presenting different combinations of plotted quantitative variables, used to explore possible relationships among larger multivariate datasets. A 'connected scatter plot' compares the shifting state of two quantitative measures over time.

Charts Correlations



Bubble plot

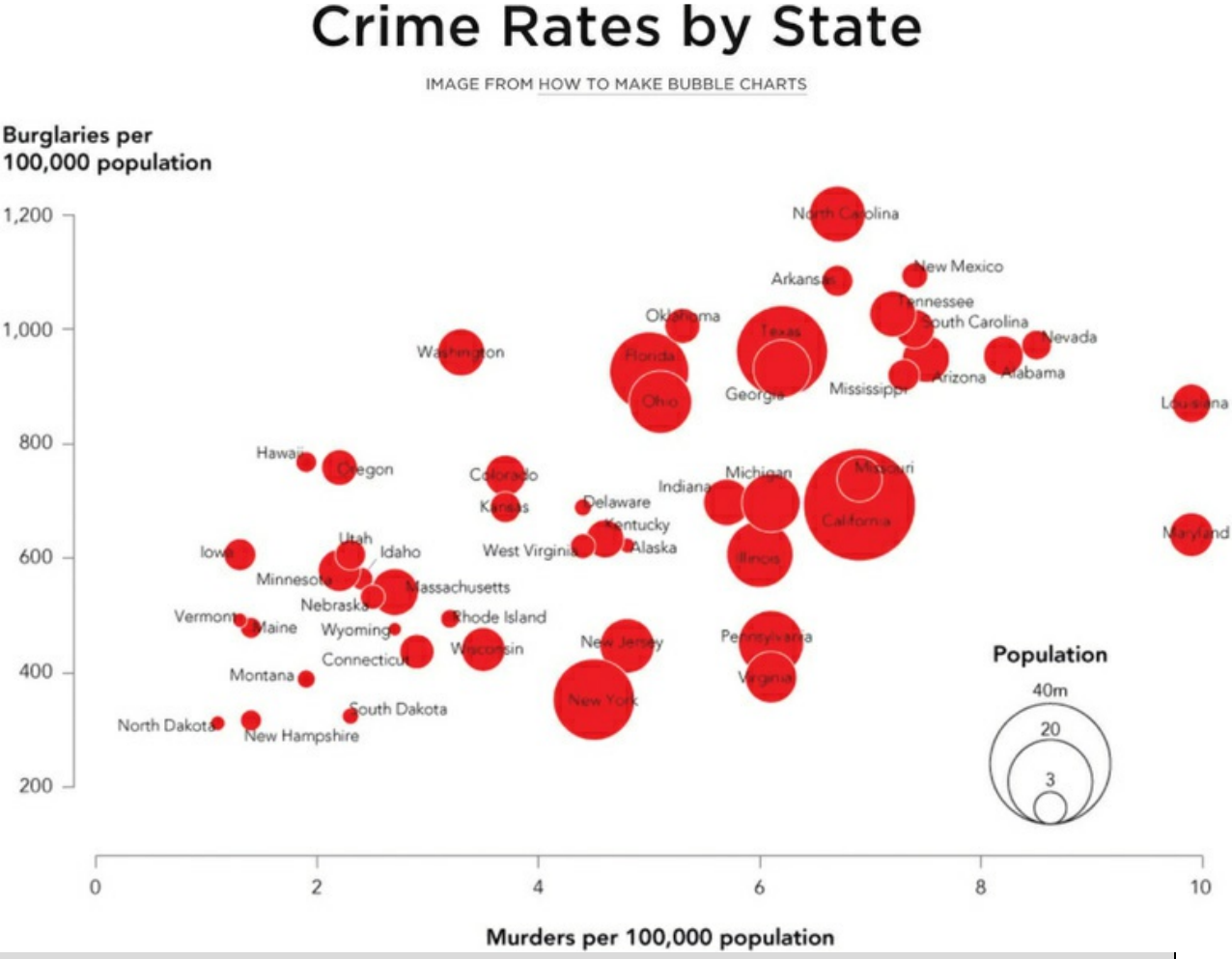
ALSO KNOWN AS Bubble chart

REPRESENTATION DESCRIPTION

A bubble plot displays the relationship between three quantitative measures for different categories. Bubble plots are used visually to explore the potential existence, extent or absence of a significant relationship between the plotted variables. In contrast to the scatter plot, the bubble plot plots proportionally sized circular areas, for each category, across two quantitative axes with the size representing a third quantitative measure. Sometimes colour is used to distinguish categorical dimensions across all the shapes.

EXAMPLE Exploring the relationship between rates of murders, burglaries (per 100,000 population) and population across states of the USA.

Figure 6.31 Crime Rates by State



HOW TO READ IT & WHAT TO LOOK FOR

Learn what each quantitative axis relates to and make a note of the range of values in each case (min to max). Look at what category or observation each plotted value on the chart refers to. Establish the quantitative size associations for the bubble areas and look up any colour associations being used for categorical distinction. Scan the chart looking for the existence of any diagonal trends that might suggest a linear correlation between the variables, or note the complete absence of any pattern, to mean no correlation. Annotations will often assist in determining the significance of any patterns like this. Identify any clusters of points and also look at the gaps, which can be just as revealing. Some of the most interesting observations come from individual outliers standing out separately from others. Look out for any patterns formed by points with similar categorical colour. What can you learn about the distribution of small, medium or large circles: are they clustered together in similar regions of the chart or quite randomly scattered? One approach to reading the 'meaning' of the plotted positions involves trying to break down the chart area into a 2 × 2 grid translating what marks positioned in those general areas might mean – which corner is 'good' or 'bad' to be located in? Remember that ruling out significant relationships can be just as useful as ruling them in. Estimating and comparing the size of areas is not as easy as it is for judging bar length or dot position. This means that the use of this chart type will primarily be about facilitating a gist – a general sense of the hierarchy of the largest and smallest values.

PRESENTATION TIPS

ANNOTATION: Gridlines can be useful to help make the value estimates clearer and reference lines (such as a trend line of best fit) might aid interpretation. It is usually hard to make direct labelling of all values work well. Firstly, it can be tricky making it clear which value relates to which point, especially when several points may be clustered together. Secondly, it creates a lot of visual clutter. Labelling choices should be based on values that are of most interest to include editorially unless interactive features enable annotations to be revealed through selection or mouseover events.

COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. When a circle has a large value its size will often overlap in spatial terms with other values. The use of outline borders and semi-transparent colours helps with the task of avoiding occlusion (visually hiding values behind others).

COMPOSITION: As the encoding of the plotted area marker values is based on position along an axis, it is not necessary to start the axes from a zero baseline – just make the scale ranges as representative as possible of the range of values being plotted. Make sensible decisions about how large to make the maximum bubble size; this will usually require trial and error experimentation to find the right balance. Ideally a bubble plot will have a 1:1 aspect ratio (equally as tall as it is wide), creating a squared area to help patterns surface more evidently. If one quantitative variable (e.g. weight) is likely to be affected by the other variable (e.g. height), it is general practice to place the former on the y-axis and the latter on the x-axis. Geometric accuracy of the circle size calculations is paramount, since mistakes are often made with circle size calculations: it is the area you are modifying, not the diameter/radius. If you wish to make your bubbles appear as 3D spheres you are essentially no longer representing quantitative values through the size of a geometric area mark, rather the mark will be a 'form' and so the size calculation will be based on volume, not area.

VARIATIONS & ALTERNATIVES

If the third quantitative variable is removed, the display would just become a 'scatter plot'. Variations on the bubble plot might see the use of different geometric areas as the markers, maybe introducing extra meaning from the underlying data through the shape, size and dimensions used.

Charts Correlations

Parallel coordinates

ALSO KNOWN AS Parallel sets

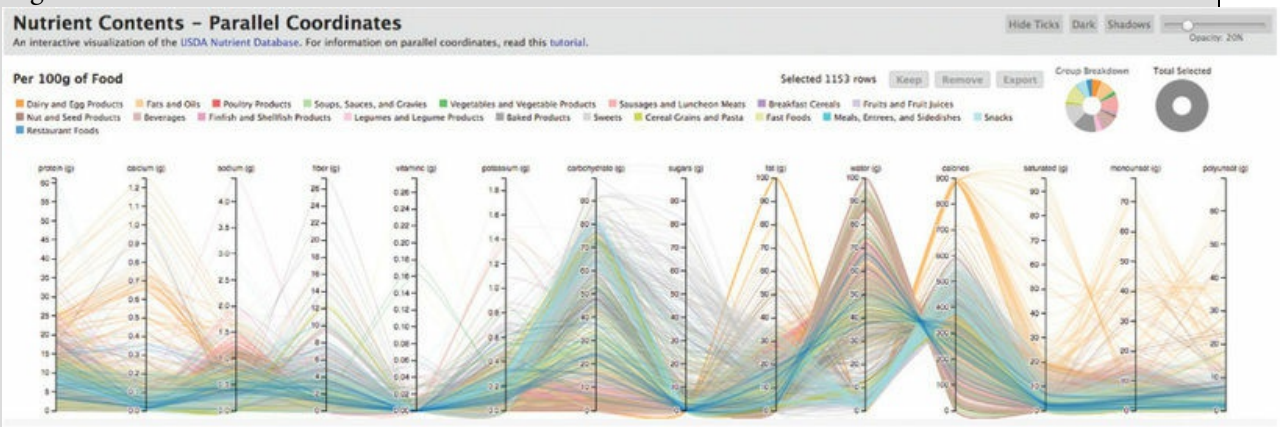
REPRESENTATION DESCRIPTION

Parallel coordinates display multiple quantitative measures for different categories in a single display. They are used visually to explore the relationships and characteristics of multi-dimensional, multivariate data. Parallel coordinates are based on a series of parallel axes representing different quantitative measures with independent axis scales. The quantitative values for each measure are plotted and then connected to form a single line. Each connected line represents a different category record. Colour may be used to differentiate further categorical dimensions. As more data is added the collective 'shape' of the data emerges and helps to inform the possibility of relationships existing among the different measures. Parallel coordinates look quite overwhelming but remember that they are almost always only used to assist in exploratory work of large and varied datasets, more so than being used for explanatory presentations of data. Generally the greater the number of measures, the more difficult the task of making sense of the underlying patterns will be, so be discerning in your choice of which variables to include. This method does not work for showing categorical (nominal) measures nor does it really offer value with the inclusion of low-range, discrete quantitative variables used (e.g. number of legs per human). Patterns will mean very little when intersecting with such

axes (they may be better deployed as a filtering parameter or a coloured categorical separator).

EXAMPLE Exploring the relationship between nutrient contents for 14 different attributes across 1,153 different items of food.

Figure 6.32 Nutrient Contents — Parallel Coordinates



name	group	protein (g)	calcium ...	sodium ...	fiber (g)	vitaminc ...	potassiu ...	carbohy ...	sugars (g)	fat (g)	water (g)	calories	saturate ...	monoun ...	polyuns ...
Butter oil, anhydrous	Dairy and Egg Products	0.28	0.004	0.002		0	0.005			99.48	0.24	876	61.924	28.732	3.694
Butter, salted	Dairy and Egg Products	0.85	0.024	0.714		0	0.024	0.06	0.06	81.11	15.87	717	51.368	21.021	3.043
Cheese fondue	Dairy and Egg Products	14.23	0.476	0.132		0	0.105	3.77		13.47	61.61	229	8.721	3.563	0.484
Cheese food, cold pack, american	Dairy and Egg Products	19.66	0.497	0.966		0	0.363	8.32		24.46	43.12	331	15.355	7.165	0.719
Cheese food, pasteurized process, swiss	Dairy and Egg Products	21.92	0.723	1.552		0	0.284	4.5		24.14	43.67	323	15.487	6.801	0.6
Cheese spread, cream cheese base	Dairy and Egg Products	7.1	0.071	0.673		0	0.112	3.5	3.5	28.6	58.5	295	18.02	8.071	1.033
Cheese, blue	Dairy and Egg Products	21.4	0.528	1.395		0	0.256	2.34	0.5	28.74	42.41	353	18.669	7.778	0.8
Cheese, brick	Dairy and Egg Products	23.24	0.674	0.56		0	0.136	2.79	0.51	29.68	41.11	371	18.764	8.598	0.784
Cheese, brie	Dairy and Egg Products	20.75	0.184	0.629		0	0.152	0.45	0.45	27.68	48.42	334	17.41	8.013	0.826

HOW TO READ IT & WHAT TO LOOK FOR

Look around the chart and acquaint yourself with what each quantitative measure axis represents. Also note what kind of sequencing of measure has been used: are neighbouring measures significantly paired? Note the range of values along each independent axis so you understand what positions along the scales represent and can determine what higher and lower positions mean. If colour has been used to group related records then identify what these represent. Scan the overall mass of lines to identify any major patterns. Study the patterns in the space between each pair of adjacent axes. This is where you will really see the potential presence or absence of, and nature of, relationships between measures. The main patterns to identify involve the presence of parallel lines (showing consistent relationships), lines converging in similar directions (some correlation) and then complete criss-crossing (negative relationship). Look out for any associations in the patterns across colour groupings. Remember that ruling out significant relationships can be just as useful as ruling them in.

PRESENTATION TIPS

INTERACTIVITY: Parallel coordinates are particularly useful when offered with interactive features, such as filtering techniques, enabling the user to interrogate and manipulate the display to facilitate visual exploration. Additionally, the option to rearrange the sequence of the measures can be especially useful.

ANNOTATION: The inclusion of visible annotated features like axis lines, tick marks, gridlines and value labels can naturally aid the readability of the data but be aware of the impact of clutter.

COLOUR: When you are plotting large quantities of records, inevitably there will be over-plotting and this might disguise the real weight of values, so the variation in the darkness of colour can be used to establish density of observations.

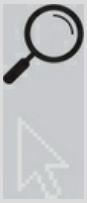
COMPOSITION: The ordering of the quantitative variables has to be of optimum significance as the connections between adjacent axes will offer the main way of seeing the local relationships: the patterns will change for every different ordering permutation. Remember that the line directions connecting records are often inconsequential in their meaning unless neighbouring measures have a common scale and similar

meaning: the connections are more about establishing commonality of pattern across records, rather than there being anything too significant behind the absolute slope direction/length.

VARIATIONS & ALTERNATIVES

The 'radar chart' has similarities with parallel coordinates in that they include several independent quantitative measures in the same chart but on a radial layout and usually only showing data for one record in the same display. A variation on the parallel coordinate would be the 'Sankey diagram', which displays categorical composition and quantitative flows between different categorical dimensions or 'stages'.

Charts Correlations



Heat map

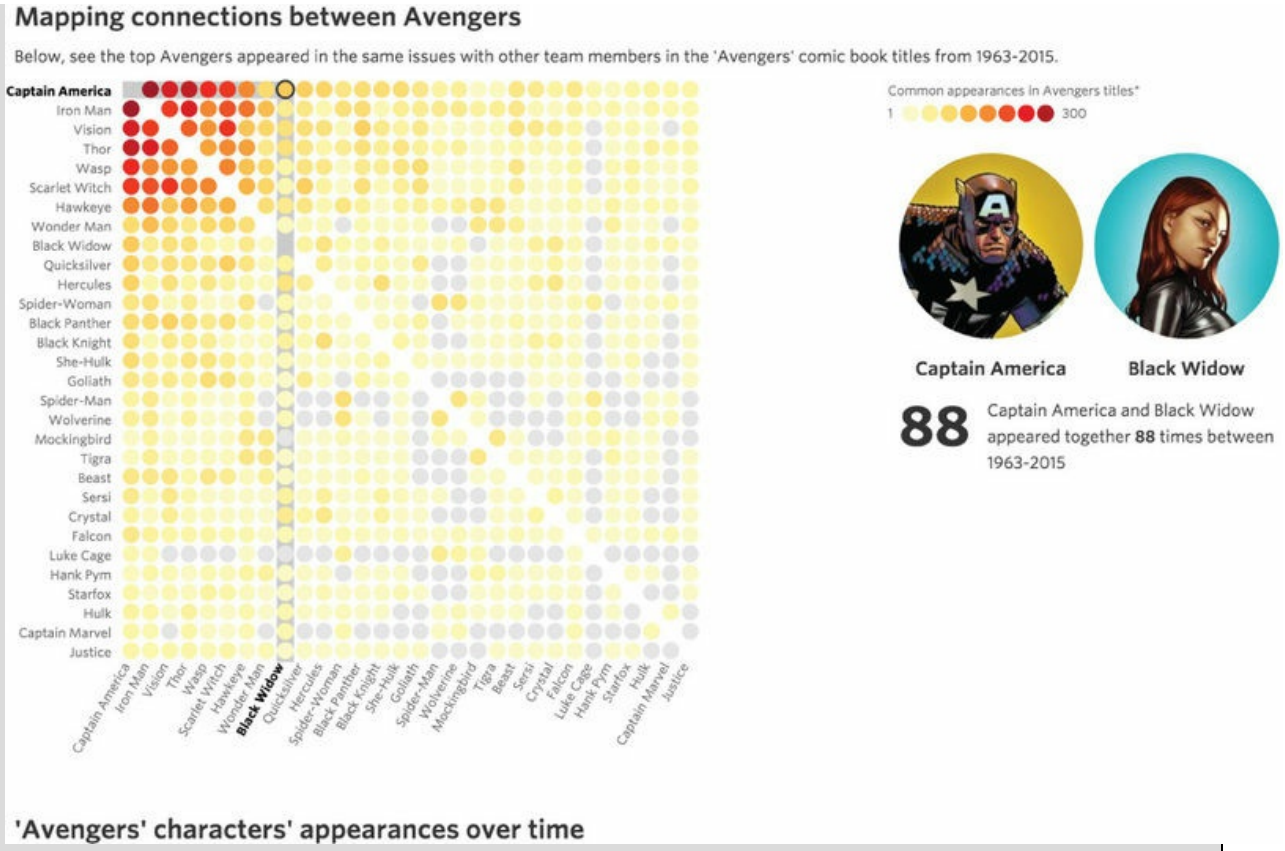
ALSO KNOWN AS Matrix chart, mosaic plot

REPRESENTATION DESCRIPTION

A heat map displays quantitative values at the intersection between two categorical dimensions. The chart comprises two categorical axes with each possible value presented across the row and column headers of a table layout. Each corresponding cell is then colour-coded to represent a quantitative value for each combination of category pairing. It is not easy for the eye to determine the exact quantitative values represented by the colours, even if there is a colour scale provided, so heat maps mainly facilitate a gist of the order of magnitude.

EXAMPLE Exploring the connections between different Avengers characters appearing in the same Marvel comic book titles between 1963 and 2015.

Figure 6.33 How the 'Avengers' Line-up Has Changed Over the Years



HOW TO READ IT & WHAT TO LOOK FOR

Learn what each categorical dimension relates to and make a note of the range of values in each case, paying attention to the significance of any ordering. Establish the quantitative value associations for the colour scales, usually found via a legend. Glance across the entire chart to locate the big, small and medium shades (generally darker = larger) and perform global comparisons to establish the high-level ranking of biggest > smallest. Scan across each row and/or column to see if there are specific patterns associated with either set of categories. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring cell's areas, to identify larger than and smaller than relationships and estimate the relative proportions. Estimate (or read, if labels are present) the absolute values of specific colour scales of interest.

PRESENTATION TIPS

ANNOTATION: Direct value labelling is possible, otherwise a clear legend to indicate colour associations will suffice.

COLOUR: Sometimes multiple different colour hues may be used to subdivide the quantitative values into further distinct categorical groups. Decisions about how many colour-scale levels and what intervals each relates to in value ranges will affect the patterns that emerge. There is no single right answer – you will arrive at it largely through trial and error/experimentation – but it is important to consider, especially when you have a diverse distribution of values.

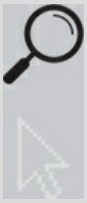
COMPOSITION: Logical sorting (and maybe even sub-grouping) of the categorical values along each axis will aid readability and may help surface key relationships.

VARIATIONS & ALTERNATIVES

A 'radial heat map' offers a structure variation whereby the table may be portrayed using a circular layout. As with any radial display this is only really of value if the cyclical ordering means something for the subject matter. A variation would see colour shading replaced by a measure of pattern density, using a scale of 'packedness' to indicate increasing quantitative values. An alternative approach would be the 'matrix chart' using size of a shape to indicate the quantitative or a range of point marker to display categorical

characteristics.

Charts Connections



Matrix chart

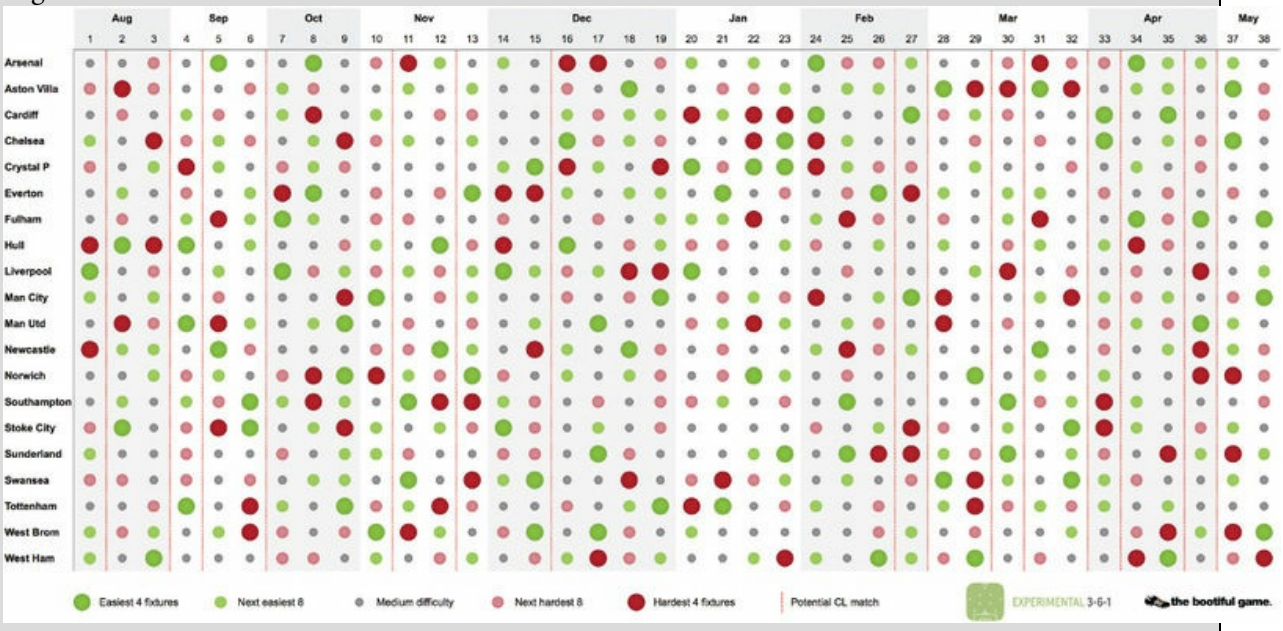
ALSO KNOWN AS Table chart

REPRESENTATION DESCRIPTION

A matrix chart displays quantitative values at the intersection between two categorical dimensions. The chart comprises two categorical axes with each possible value presented across the row and column headers of a table layout. Each corresponding cell is then marked by a geometric shape with its area sized to represent a quantitative value and colour often used visually to distinguish a further categorical dimension. While they are most commonly seen using circles, you can use other proportionally sized shapes.

EXAMPLE Exploring the perceived difficulty of fixtures across the season for teams in the premier league 2013–14.

Figure 6.34 Interactive Fixture Molecules



HOW TO READ IT & WHAT TO LOOK FOR

Learn what each categorical dimension relates to and make a note of the range of values in each case, paying attention to the significance of any ordering. Establish the quantitative size associations for the area marks and look up any colour associations being used, both usually found via a legend. Glance across the entire chart to locate the big, small and medium areas and perform global comparisons to establish the high-level ranking of biggest > smallest. Scan across each row and/or column to see if there are specific patterns associated with either set of categories. Identify any noticeable exceptions and/or outliers. Perform local

comparisons between neighbouring circular areas, to identify larger than and smaller than relationships and estimate the relative proportions. Estimate (or read, if labels are present) the absolute values of specific geometric areas of interest.

PRESENTATION TIPS

ANNOTATION: Direct value labelling is possible, otherwise be sure to include a clear size legend. Normally this will be more than sufficient as the reader may simply be looking to get a gist of the order of magnitude.

COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible.

COMPOSITION: If there are large outlier values there may be occasions when the size of a few circles outgrows the cell it occupies. You might editorially decide to allow this, as the striking shape may create a certain impact, otherwise you will need to limit the largest quantitative value to be represented by the maximum space available within the table's cell layout. Logical sorting (and maybe even sub-grouping) of the categorical values along each axis will aid readability and may help surface key relationships. The geometric accuracy of the circle size calculations is paramount. Mistakes are often made with circle size calculations: it is the area you are modifying, not the diameter/radius.

VARIATIONS & ALTERNATIVES

A variation may be to remove the quantitative attribute of the area marker, replacing it with a point marker to represent a categorical status to indicate simply a yes/no observation through the presence/absence of a point or through the quantity of points to represent a total. An application of this might be in calendar form whereby a marker in a date cell indicates an instance of something. It could also employ a broader range of different categorical options; in practice any kind of marker (symbol, colour, photograph) could be used to show a characteristic of the relationship at each coordinate cell. An alternative might be the 'heat map' which colour-codes the respective cells to indicate a relationship based on a quantitative measure.

Charts Connections

Node–link diagram

ALSO KNOWN AS Network diagram, graph, hairballs

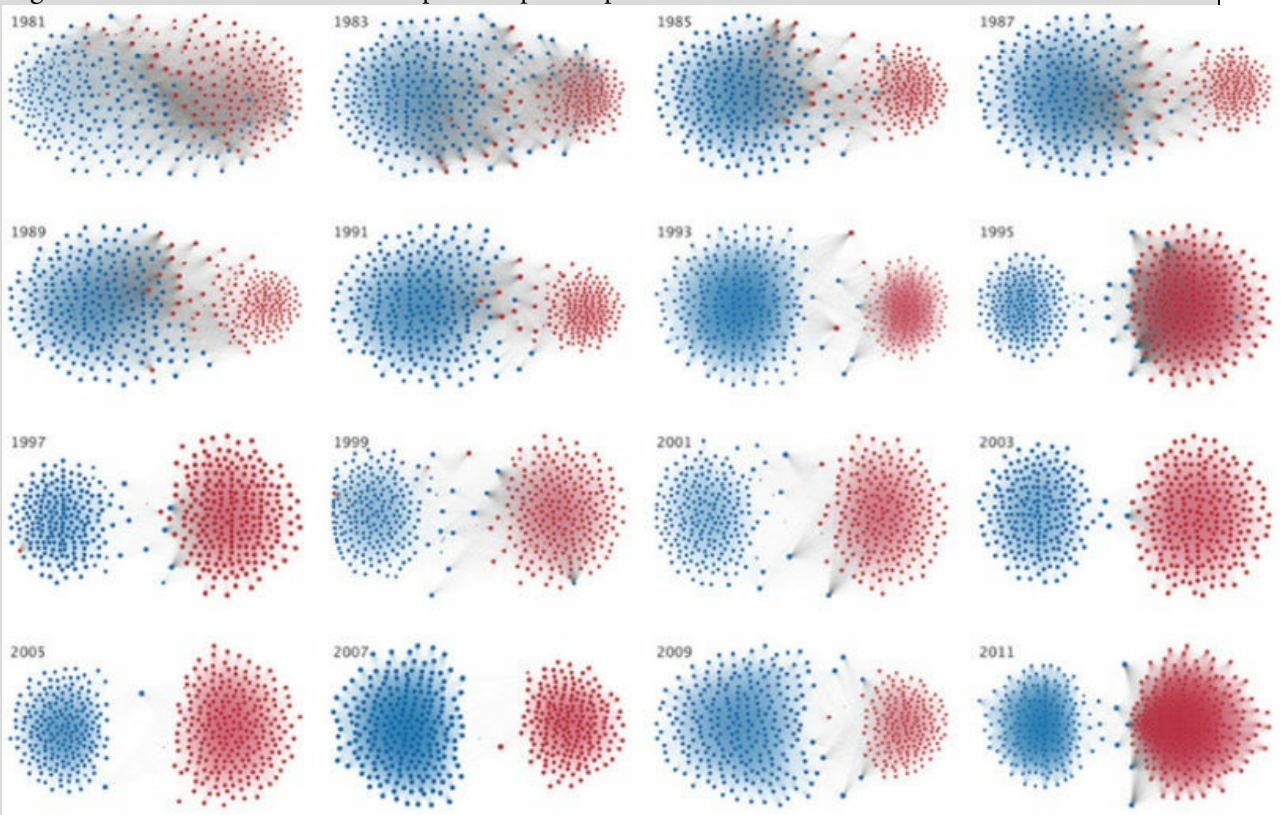
REPRESENTATION DESCRIPTION

Node–link diagrams display relationships through the connections between categorical 'entities'. The entry-level version of this type of diagram displays entities as nodes (represented by point marks and usually including a label) with links or edges (represented by lines) depicting the existence of connections. The connecting lines will often display an attribute of direction to indicate the influencer relationship. In some versions a quantitative weighting is applied to the show relationship strength, maybe through increased line width. Replacing point marks with a geometric shape and using attributes of size and colour is a further

variation. Often the complexity seen in these displays is merely a reflection of the underlying complexity of the subject and/or system upon which the data is based, so oversimplifying can compromise the essence of such content.

EXAMPLE Exploring the connections of voting patterns for Democrats and Republicans across all members of the US House of Representatives from 1949 to 2012.

Figure 6.35 The Rise of Partisanship and Super-cooperators in the U.S.



HOW TO READ IT & WHAT TO LOOK FOR

The first thing to consider is what entity each node (point or circular area) represents and what the links mean in relationship terms. There may be several other properties to acquaint yourself with, including attributes like the size of the node areas, the categorical nature of colouring, and the width and direction of the connections. Across the graph you will mainly be seeking out the clusters that show the nodes with the most relationships (representative of influencers or hubs) and those without (including outliers). Small networks will generally enable you to look closely at specific nodes and connections and easily see the emerging relationships. When datasets are especially large, consisting of thousands of nodes and greater numbers of mutual connections, the displays can seem overwhelmingly cluttered and will be too dense to make many detailed observations at node–link level. Instead, just relax and know that your readability will be about a higher level sense-making of the clusters/hubs and main outliers.

PRESENTATION TIPS

INTERACTIVITY: Node–link diagrams are particularly useful when offered with interactive features, enabling the user to interrogate and manipulate the display to facilitate visual exploration. The option to apply filters to reduce the busy-ness of the visual and enable isolation of individual node connections helps users to focus on specific parts of the network of interest.

ANNOTATION: The extent of annotated features tends to be through the inclusion of value labels for each entity. Accommodating the relative word sizes on each node can be difficult to achieve with real elegance (once again that is where interactivity adds value, through the select/mouseover event to reveal the label).

COLOUR: Aside from the possible categorical colouring of each node, decisions need to be made about the colour of the connecting lines, especially on deciding how prominent these links will be in contrast to the nodes.

COMPOSITION: Composition decisions are where most of the presentation customisation exists. There are several common algorithmic treatments used to compute custom arrangements to optimise network displays, such as force-directed layouts (using the physics of repulsion and springs to amplify relationships) and simplifying techniques (such as edge bundling to aggregate/summarise multiple similar links).

VARIATIONS & ALTERNATIVES

There are many derivatives of the node–link diagram, as explained, based on variations in the use of different attributes. 'Hive plots' and 'BioFabric' offer alternative approaches based on replacing nodes with vertices.

Charts Connections

Chord diagram

ALSO KNOWN AS Radial network diagram, arc diagram (wrongly)

REPRESENTATION DESCRIPTION

A chord diagram displays relationships through the connections between and within categories. They are formed around a radial display with different categories located around the edge: either as individual nodes or proportionally sized segments (arcs) of the circumference according to a part-to-whole breakdown. Emerging inwards from each origin position are curved lines that join with other related categorical locations around the edge. The connecting lines are normally proportionally sized according to a quantitative measure and a directional or influencing relationship is often indicated. The perceived readability of the chord diagram will always be influenced by the quantity and range of values being plotted. Small networks will enable a reader to look closely at specific categories and their connections to see the emerging relationships easily; larger systems will look busy through the network of lines but they can still provide windows into complex networks of influence. Often the complexity seen in these displays is merely a reflection of the underlying complexity of the subject and/or system upon which the data is based, so oversimplifying can compromise the essence of such content.

EXAMPLE Exploring the connections of migration between and within 10 world regions based on estimates across five-year intervals between 1990 and 2010.

Figure 6.36 The Global Flow of People

HOW TO READ IT & WHAT TO LOOK FOR

First determine how categories are displayed around the circumference, either as nodes or part-to-whole arcs, and identify each one individually. Consider the implication of the radial sorting of these categorical values and, if based on part-to-whole sizes, establish a sense of the largest > smallest arc lengths. Colour-coding may be applied to the categories so note any associations. Look inside the display to determine what relationships the connecting lines represent and check for any directional significance. Look closer at the tangled collection of lines criss-crossing this space, noting the big values (usually through line weight or width) and the small ones. Avoid being distracted by the distance a line travels, which is just a by-product of the outer categorical arrangement: a long connecting line is just as significant a relationship as a short one. For this reason, pay close attention to any connecting lines that have very short looping distances to adjacent categories. Are there any patterns of lines heading towards or leaving certain categories?

PRESENTATION TIPS

INTERACTIVITY: Chord diagrams are particularly useful when offered with interactive features, enabling the user to interrogate and manipulate the display to facilitate visual exploration. The option to apply filters to reduce the busy-ness of the visual and enable isolation of individual node connections helps users to focus on specific parts of the network of interest.

ANNOTATION: Annotated features tend to be limited to value labelling of the categories around the circumference and, occasionally, directly onto the base or ends of the connecting lines (usually just those that are large enough to accommodate them).

COLOUR: Aside from the categorical colouring of each node, decisions need to be made about the colour of the connecting lines, especially on deciding how prominent these links will be in contrast to the

nodes. Sometimes the connections will match the origin or destination colours, or they will combine the two (with a start and end colour to match the relationship).

COMPOSITION: The main arrangement decisions come through sorting, firstly by generating as much logical meaning from the categorical values around the edge of the circle and secondly by deciding on the sorting of the connecting lines in the z-dimension – if many lines are crossing, there is a need to think about which will be on top and which will be below. Showing the direction of connections can be difficult as there is so little room for manoeuvring many more visual attributes, such as arrows or colour changes. One common, subtle solution is to pull the destination join back a bit, leaving a small gap between the connecting line and the destination arc. This then contrasts with connecting lines that emerge directly from the categorical arcs, showing it is their origin.

VARIATIONS & ALTERNATIVES

The main alternatives would be to consider variations of the 'node–link diagram' or, specifically, the 'arc diagram', which offers a further variation on the theme of networked displays, placing all the nodes along a baseline and forming connections using semi-circular arcs, rather than using a graph or radial layout.

Charts Connections

Sankey diagram

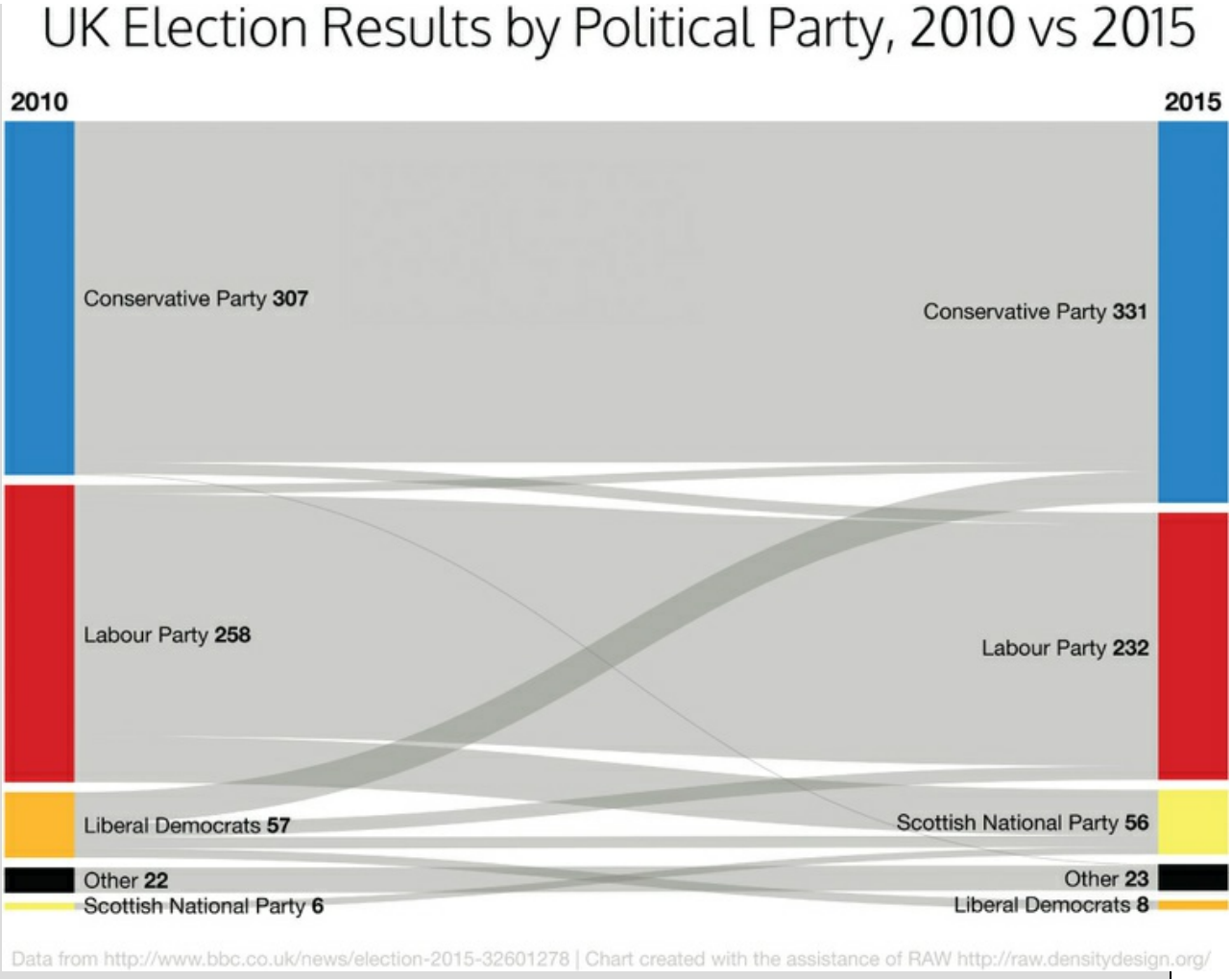
ALSO KNOWN AS Alluvial diagram

REPRESENTATION DESCRIPTION

Sankey diagrams display categorical composition and quantitative flows between different categorical dimensions or 'stages'. The most common contemporary form involves a two-sided display, with each side representing different (but related) categorical dimensions or different states of the same dimension (such as 'before and after'). On each side there is effectively a stacked bar chart displaying proportionally sized and differently coloured (or spaced apart) constituent parts of a whole. Curved bands link each side of the display to represent connecting categories (origin and destination) with the proportionally sized band (its thickness) indicating the quantitative nature of this relationship. Some variations involve multiple stages and might present attrition through the diminution size of subsequent stacks. Traditionally the Sankey has been used as a flow diagram to visualise energy or material usage across engineering processes. It is closely related to the 'alluvial diagram', which tends to show changes in composition and flow over time, but the Sankey label is often applied to these displays also.

EXAMPLE Exploring the seat changes among political parties between the 2010 and 2015 UK General Elections.

Figure 6.37 UK Election Results by Political Party, 2010 vs 2015



HOW TO READ IT & WHAT TO LOOK FOR

Based on the basic two-sided version of the Sankey diagram, look down both sides of the chart to learn what states are represented and what the constituent categories are. Pay close attention to the categorical sorting and pick out the large and small values on each side. Then look at the connecting lines, making observations about the largest and narrowest bands and noting any that seem to be mostly redistributed into a different category compared to those that just join with the same. Notice any small break-off bands that seem to cross the height of the whole chart, perhaps representing a more dramatic change or diversion between states. As with most network-type visualisations, the perceived readability of the Sankey diagram will always be influenced by the quantity and range of values being plotted, as well as the number of different states presented.

PRESENTATION TIPS

INTERACTIVITY: Sankey diagrams are particularly useful when offered with interactive features, enabling the user to interrogate and manipulate the display to facilitate visual exploration. The option to apply filters to reduce the busy-ness of the visual and enable isolation of individual node connections helps users to focus on specific parts of the network of interest.

ANNOTATION: Annotated features tend to be limited to value labelling of the categories that make up each 'state' stack.

COLOUR: Colouring is often used visually to indicate the categories of the connecting bands, though it can get a little complicated when trying to combine a sense of change through an origin category colour blending with a destination category colour when there has been a switch.

COMPOSITION: The main arrangement decisions come through sorting, firstly by generating as much

logical meaning from the categorical values within the stacks and, secondly, by deciding on the sorting of the connecting lines in the z-dimension – if many lines are crossing, there is a need to think about which will be on top and which will be below. There is no significant difference between a landscape or portrait layout, which will depend on the subject matter 'fit' and the space within which you have to work. Try to ensure that the sorting of the categorical dimensions is as logical and meaningful as possible.

VARIATIONS & ALTERNATIVES

The concept of a Sankey diagram showing composition and flow can also be mapped onto a geographical projection as one of the variations of the 'flow map'. You could use a 'chord diagram' as an alternative to show how larger networks are composed proportionally and in their connections. Showing how component parts have changed over time could just be displayed using a 'stacked area chart'. A 'funnel chart' is a much simplified display to show how a single value changes (usually diminishing) across states, for topics like sales conversion. This often is based on a funnel-like shape formed by a wide bar at the top (those entering the system) and then gradually narrower bars, stage by stage towards the end state.

Charts Trends

Line chart

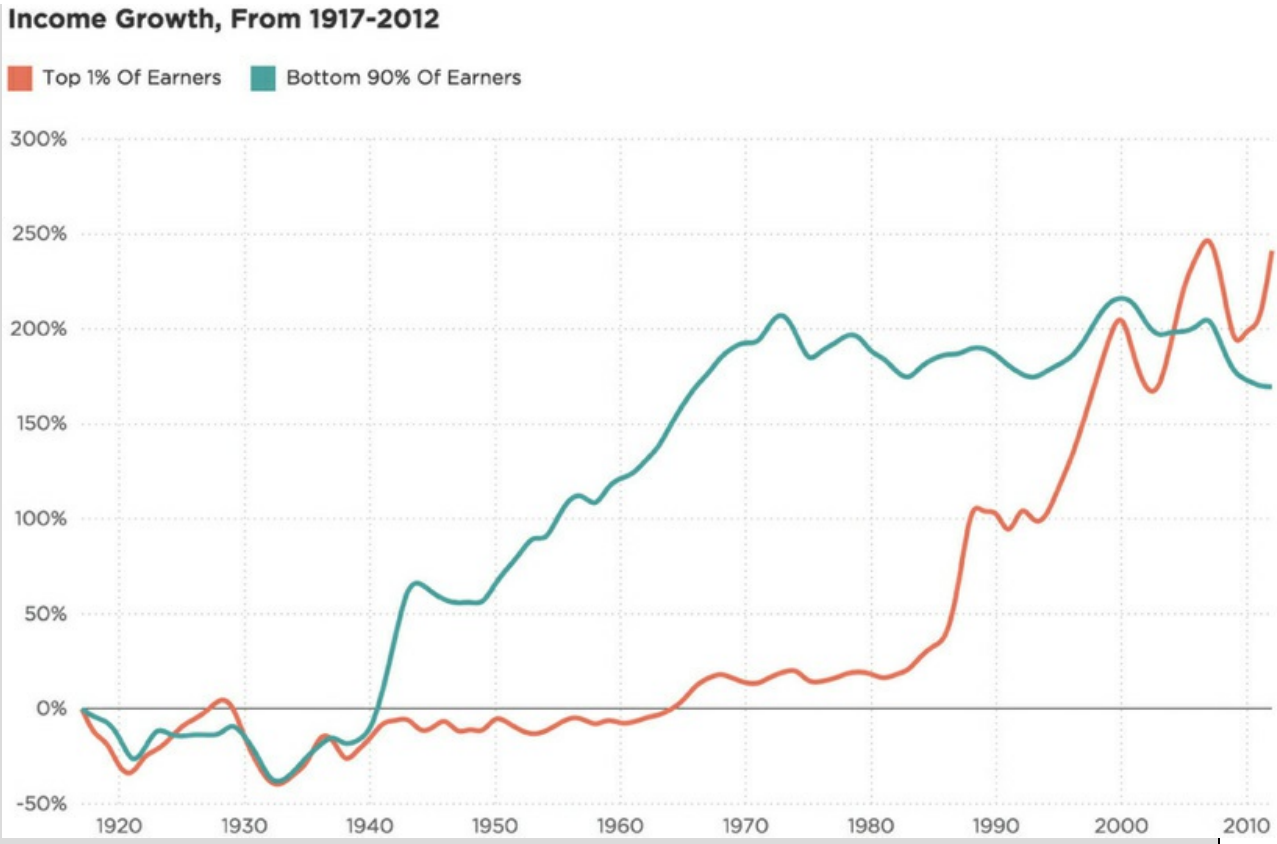
ALSO KNOWN AS Fever chart, stock chart

REPRESENTATION DESCRIPTION

A line chart shows how quantitative values for different categories have changed over time. They are typically structured around a temporal x-axis with equal intervals from the earliest to latest point in time. Quantitative values are plotted using joined-up lines that effectively connect consecutive points positioned along a y-axis. The resulting slopes formed between the two ends of each line provide an indication of the local trends between points in time. As this sequence is extended to plot all values across the time frame it forms an overall line representative of the quantitative change over time story for a single categorical value. Multiple categories can be displayed in the same view, each represented by a unique line. Sometimes a point (circle/dot) is also used to substantiate the visibility of individual values. The lines used in a line chart will generally be straight. However, sometimes curved line interpolation may be used as a method of estimating values between known data points. This approach can be useful to help emphasise a general trend. While this might slightly compromise the visual accuracy of discrete values if you already have approximations, this will have less impact.

EXAMPLE Showing changes in percentage income growth for the Top 1% and Bottom 90% of earners in the USA between 1917 and 2012.

Figure 6.38 The Fall and Rise of U.S. Inequality, in 2 Graphs



HOW TO READ IT & WHAT TO LOOK FOR

Firstly, learn about the axes: what is the time period range presented on the x-axis (and in what order) and what is the range of quantitative values shown on the y-axis, paying particular attention to the origin value (which may not be zero)? Inside the chart, determine what categories each line represents: for single lines this will usually be clear from the chart title, for multiple lines you might have direct labelling or a legend to learn colour associations. Think about what high and low values mean: is it 'good' to be large/small, increasing or decreasing? Glance at the general patterns (especially if there are many) looking for observations such as any trends (short or long term), any sudden moments of a rise or fall (V- or W -shapes, or inverted), any sense of seasonal or cyclical patterns, any points of interest where lines cross each other or key thresholds that are reached/exceeded. Can you mentally extrapolate from the values shown any sense of a forecasted trend? Avoid jumping to spurious interpretations if you see two line series following a similar pattern; this does not necessarily mean that one thing has caused the other, it might just be coincidence. Then look more closely at categories of interest and at patterns around specific moments in time, and pick out the peak, low, earliest and latest values for each line. Where available, compare the changing quantities against annotated references such as targets, forecast, previous time periods, range bands, etc.

PRESENTATION TIPS

INTERACTIVITY: Interactivity may be especially helpful if you have many categories and wish to enable the user to isolate (in focus terms) a certain line category of interest.

ANNOTATION: Chart apparatus devices like tick marks and gridlines in particular can be helpful to increase the accuracy of the reading of the quantitative values. If you have axis labels you should not need direct labels on each value point – this will be label overload. You might choose to annotate specific values of interest (highest, lowest, specific milestones). Think carefully about what is the most useful and meaningful interval for your time axis labelling. When several categories are being shown, if possible, try directly to label the categories shown by each line, maybe at the start or end position.

COLOUR: When many categories are shown it may be that only certain emphasised lines of interest possess a colour and a label – the rest are left in greyscale for context.

COMPOSITION: Composition choices are mostly concerned with the chart's dimensions: its aspect ratio, how high and wide to make it. The sequencing of values tends to be left to right for the sequence of the time-based x-axis and low rising to high values on the y-axis; you will need a good (and clearly annotated) reason to break this convention. Line charts do not always need the y-axis to start at zero, as we are not judging the size of a bar, rather the position along an axis. You should expect to see a zero baseline if zero has some critical significance in the interpretation of the trends. If your y-axis origin is not going to be zero, you might include a small gap between the x-axis and the minimum so that it is not implied. Be aware that the upward and downward trends on a line chart can seem more significant if the chart width is narrow and less significant if it is more stretched out. There is no single rule to follow here but a useful notion involves 'banking to 45°' whereby the average slope angle across your chart heads towards 45°. While it is impractical to actually measure this, judging by eye tends to be more than sufficient.

VARIATIONS & ALTERNATIVES

Variations of the line chart may include the 'cumulative line chart' or 'step chart'. 'Spark lines' are mini line charts that aim to occupy almost only a word's length amount of space. Often seen in dashboards where space is at a premium and there is a desire to optimise the density of the display. 'Bar charts' can also be used to show how values look over time when there is perhaps greater volatility in the quantitative values across the time period and when the focus is on the absolute values at each point in time, more so than trends. Sometimes a line chart can show quantitative trends over continuous space rather than time. For showing ranking over time, consider the 'bump chart', and for before and after comparisons, the 'slope graph'.

Charts Trends

Bump chart

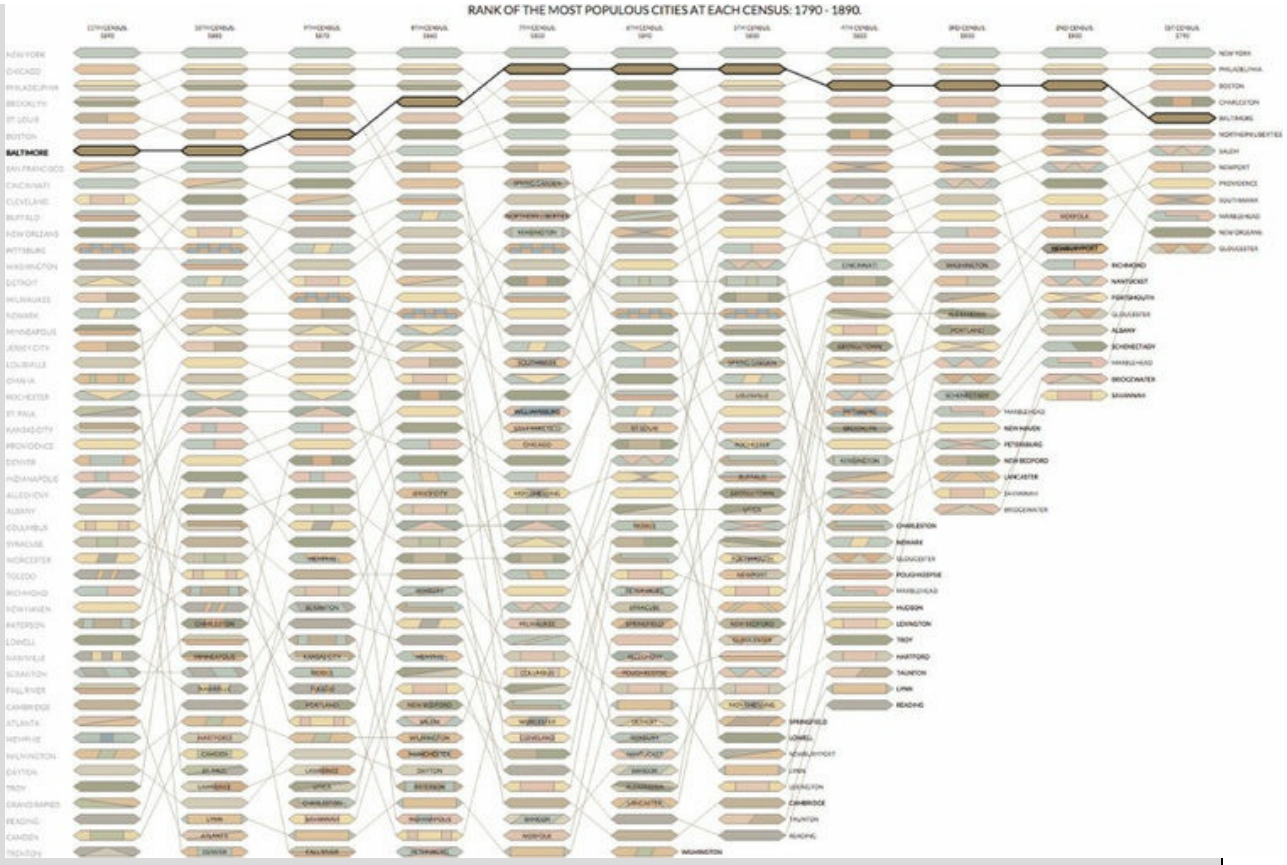
ALSO KNOWN AS

REPRESENTATION DESCRIPTION

A bump chart shows how quantitative rankings for categories have changed over time. They are typically structured around a temporal x-axis with equal intervals from the earliest to latest point in time. Quantitative rankings are plotted using joined-up lines that effectively connect consecutive points positioned along a y-axis (typically top = first). The resulting slopes formed between the two ends of each line provide an indication of the local ranking trends between points in time. As this sequence is extended to plot all values across the time frame it forms an overall line representative of the ranking story for a single categorical value. Multiple categories are often displayed in the same view, showing how rankings have collectively changed over time. Sometimes a point (circle/dot) mark is also used to substantiate the connected visibility of category lines, as is colour (for the lines and/or the points).

EXAMPLE Showing changes in rank of the most populated US cities at each census between 1790 and 1890.

Figure 6.39 Census Bump: Rank of the Most Populous Cities at Each Census, 1790—1890



HOW TO READ IT & WHAT TO LOOK FOR

Firstly, you need to learn about the axes. What is the time period range presented on the x-axis (and in what order)? What are the range of quantitative rankings shown on the y-axis (check that the ranks start at 1 from the top downwards)? Inside the chart, determine what categories each line represents: this might be explained through direct labelling, a colour legend, interactivity or through differentiating point marker attributes of colour/shape/pattern. Think about what high and low ranks mean: is it 'good' to be high up the rankings and is it better to be moving up or down? Consider the general patterns to look for observations such as consistent trends (largely parallel lines) or completely non-relational patterns (lines moving in all directions). Are there any prominent stories of categories that have had a sudden rise or fall (V- or W-shapes, or inverted)? Is there any evidence of seasonal or cyclical patterns, any key points of interest where lines cross each other or key thresholds that are reached/exceeded? Next, look more closely at categories of interest and at patterns around specific moments in time, and pick out the peak, low, earliest and latest values for each line.

PRESENTATION TIPS

INTERACTIVITY: Interactivity is usually necessary with bump charts, especially if you have many categories and wish to enable the user to isolate (in focus terms) a certain line category of interest.

ANNOTATION: The ranking labels can be derived from the vertical position along the scale so direct labelling is usually unnecessary. You might choose to annotate specific values of interest (highest, lowest, specific milestones). Think carefully about what is the most useful and meaningful interval for your time axis labelling.

COLOUR: Often, with many categories to show in the same chart, the big challenge is to distinguish each line, especially as they likely criss-cross often with others. Using colour association can be useful for less than 10 categories, but for more than that you really need to offer the interactivity or maybe decide that only certain emphasised lines of interest will possess a colour and the rest are left in greyscale for context.

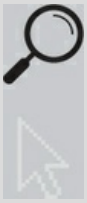
COMPOSITION: The sequencing of values tends to be left to right for the sequence of the time-based

x-axis with high rankings (low number) on the y-axis moving downwards. You will therefore need a good (and clearly annotated) reason to break this convention.

VARIATIONS & ALTERNATIVES

Alluvial diagrams (similar to Sankey diagrams) can show how rankings have changed over time while also incorporating a component of quantitative magnitude. This approach is effectively merging the 'bump chart' with the 'stacked area chart'. Consider 'line charts' and 'area charts' if the ranking is of secondary interest to the absolute values.

Charts Trends



Slope graph chart

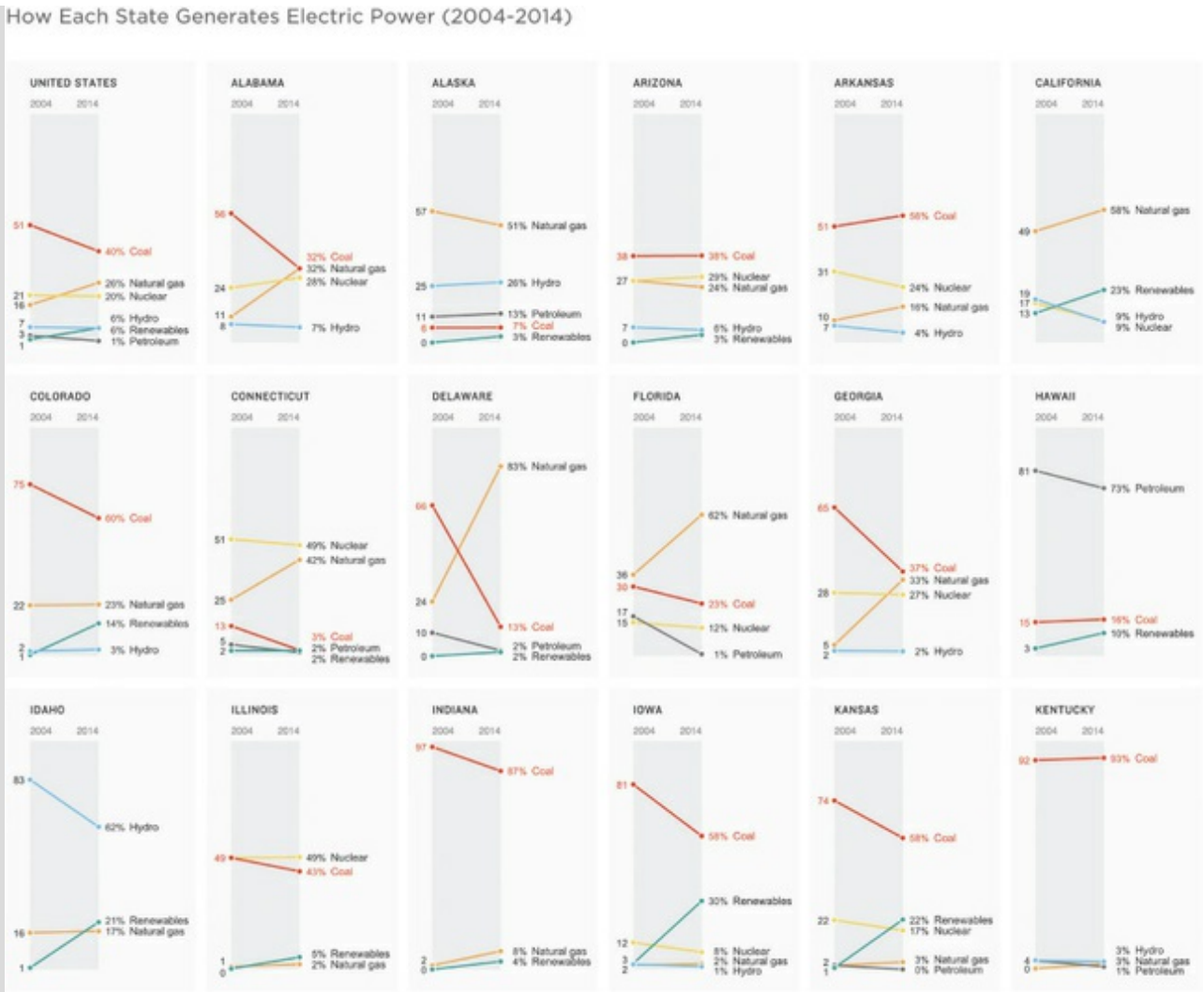
ALSO KNOWN AS Slope chart

REPRESENTATION DESCRIPTION

A slope graph shows a 'before and after' display of changes in quantities for different categories. The display is based on (typically) two parallel quantitative axes with a consistent scale range to cover all possible quantitative values. A line is plotted for each category connecting the two axes together with the vertical position on each axis representing the respective quantitative values. Sometime a dot is also used to further substantiate the visibility of the value positions. These connecting lines form slopes that indicate the upward, downward or stable trend between points in time. The resulting display incorporates absolute values, reveals rank and, of course, shows change between time. Colours are often used visually to distinguish different categorical lines, otherwise this can be used to surface visibly the major trend states (up, down, no change). A slope graph works less well when all values (or the majority) are going in the same direction; consider alternatives if this is the case.

EXAMPLE Showing changes in the share of power sources across all US states between 2004 and 2014.

Figure 6.40 Coal, Gas, Nuclear, Hydro? How Your State Generates Power



HOW TO READ IT & WHAT TO LOOK FOR

Firstly, learn about the axes: what are the two points in time being presented and what is the possible range of quantitative values shown on the y-axis, checking that the ranks start from the top down? Inside the chart, learn what each category line relates to and determine what categories each line represents: this might be explained through direct labelling, a colour legend, or through interactivity. Think about what upward, downward and stable trends mean: is it 'good' to be moving up or down? Is it more interesting to show no change? Look at the general patterns to observe such things as consistent trends (largely parallel lines in either direction) or completely non-relational patterns (lines moving in all directions). Colour may be used to accentuate the distinction between upward and downward trends. Are there any prominent stories of categories that have had a dramatic rise or fall? Even if no values have dramatically altered, that in itself can be an important finding, especially if change was expected. Next, look more closely at categories of interest and pick out the highest and lowest values on each side to learn about those stories. Look for the gaps where there are no values, and at outlier values too, to see if some sit outside the normal value clusters.

PRESENTATION TIPS

INTERACTIVITY: Depending on the number of category values being presented, slope graphs can become quite busy, especially if there are bunches of similar values and slope transitions. This also causes a problem with accommodating multiple labels on the same value. On these occasions you might find interactive slope graphs to help filter/exclude certain values.

ANNOTATION: Labelling of each category will get busy, especially when there are shared values, so you might choose to annotate specific values of interest (highest, lowest, of editorial interest).

COLOUR: Often when you have many categories to show in the same chart the big challenge is to

distinguish each line, especially as they likely criss-cross often with others. Using colour association can be useful for less than 10 categories usually with direct labelling on the left and/or right of the chart.

COMPOSITION: The aspect ratio of the slope graph (height and width) will often be determined by the space you have to work with.

VARIATIONS & ALTERNATIVES

Rather than showing a before and after story, some slope graphs are used to show the relationship between different quantitative measures for linked categories. In this case the connecting line is not indicative of a directional relationship, just the relationship itself. An alternative option would be the 'connected dot plot' which can also show before and after stories and is a better option when all values are moving in the same direction.

Charts Trends

Connected scatter plot

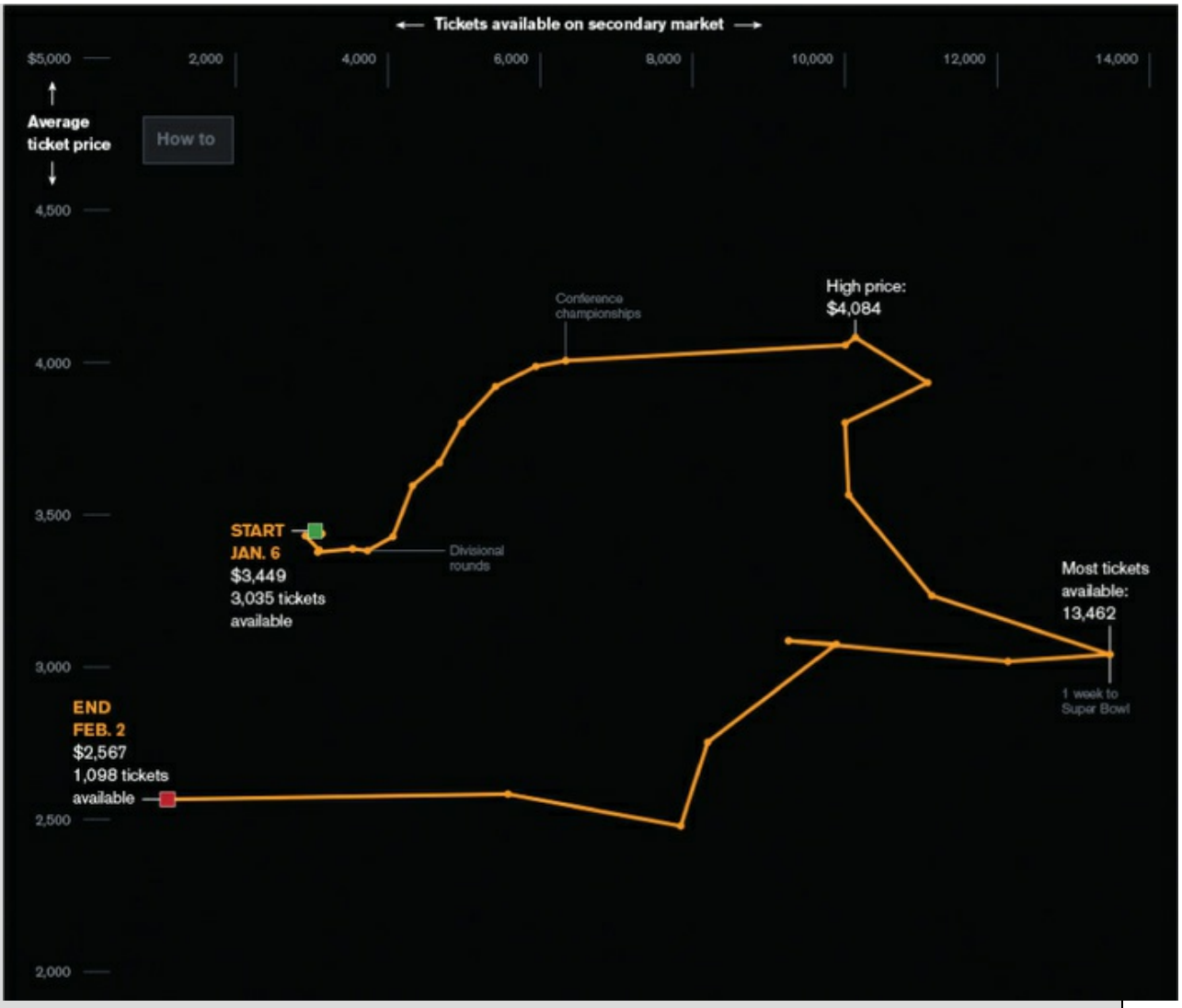
ALSO KNOWN AS Trail chart

REPRESENTATION DESCRIPTION

A connected scatter plot displays the relationship between two quantitative measures over time. The display is formed by plotting marks like a dot or circle for each point in time at the respective coordinates along two quantitative x- and y-axes. The collection of individual points is then connected (think of a dot-to-dot drawing puzzle) using lines joining each consecutive point in time to form a sequence of change. Generally there would only be a single connected line plotted on a chart to avoid the great visual complexity of overlaying several in one display. However, if multiple categories are to be included, colour is typically used to distinguish each series.

EXAMPLE Showing changes in the daily price and availability of Super Bowl tickets on the secondary market four weeks prior to the event across five Super Bowl finals.

Figure 6.41 Holdouts Find Cheapest Super Bowl Tickets Late in the Game



HOW TO READ IT & WHAT TO LOOK FOR

Learn what each quantitative axis relates to and make a note of the range of values in each case (min to max). Look at what each plotted value on the chart refers to in terms of its date label and determine the meaning of line direction. It usually helps to parse your thinking by considering what higher/lower values mean for each quantitative axis individually and then combining the joint meaning thereafter. Try to follow the chart from the start to the end, mapping out in your mind the sequence of a narrative as the values change in all directions and noting the extreme values in the outer edges of your line's reach. Look at the overall pattern of the connected line: is it consistently moving in one direction? Does it ebb and flow in all directions? Does it create a spiral shape? Compare consecutive points for a more focused view of change between two points.

PRESENTATION TIPS

INTERACTIVITY: The biggest challenge is making the connections and the sequence as visible as possible. This becomes much harder when values change very little and/or they loop back almost in spiral fashion, crossing back over themselves. It is especially hard to label the sequential time values elegantly. One option to overcome this is through interactivity and particularly through animated sequences which build up the display, connecting one line at a time and unveiling the date labels as time progresses. It is often the case that only one series will be plotted. However, interactive options may allow the user to overlay one or more for comparison, switching them on and off as required.

ANNOTATION: Connected scatter plots are generally seen as one of the most complex chart types for the unfamiliar reader to work out how to read, given the amount of different attributes working together in the display. It is therefore vital that as much help is given to the reader as possible with 'how to read' guides

and illustrations of what the different directions of change mean.

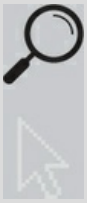
COLOUR: Colour is only generally used to accentuate certain sections of a sequence that might represent a particularly noteworthy stage of narrative.

COMPOSITION: As the encoding of the plotted point values is based on position along an axis, it is not necessary to start the axes from a zero baseline – just make the scale ranges as representative as possible of the range of values being plotted. Ideally a connected scatter plot will have a 1:1 aspect ratio (equally as tall as it is wide), creating a squared area to help patterns surface more evidently. If one quantitative variable (e.g. weight) is likely to be affected by the other variable (e.g. height), it is general practice to place the former on the y-axis and the latter on the x-axis.

VARIATIONS & ALTERNATIVES

The 'comet chart' is to the connected scatter plot what the 'slope graph' is to the 'line chart' – a summarised view of the changing relationships across two quantitative values between just two points in time. Naturally a reduced variation of the connected scatter plot is simply the 'scatter plot' where there is no time dimension or elements of connectedness.

Charts Trends



Area chart

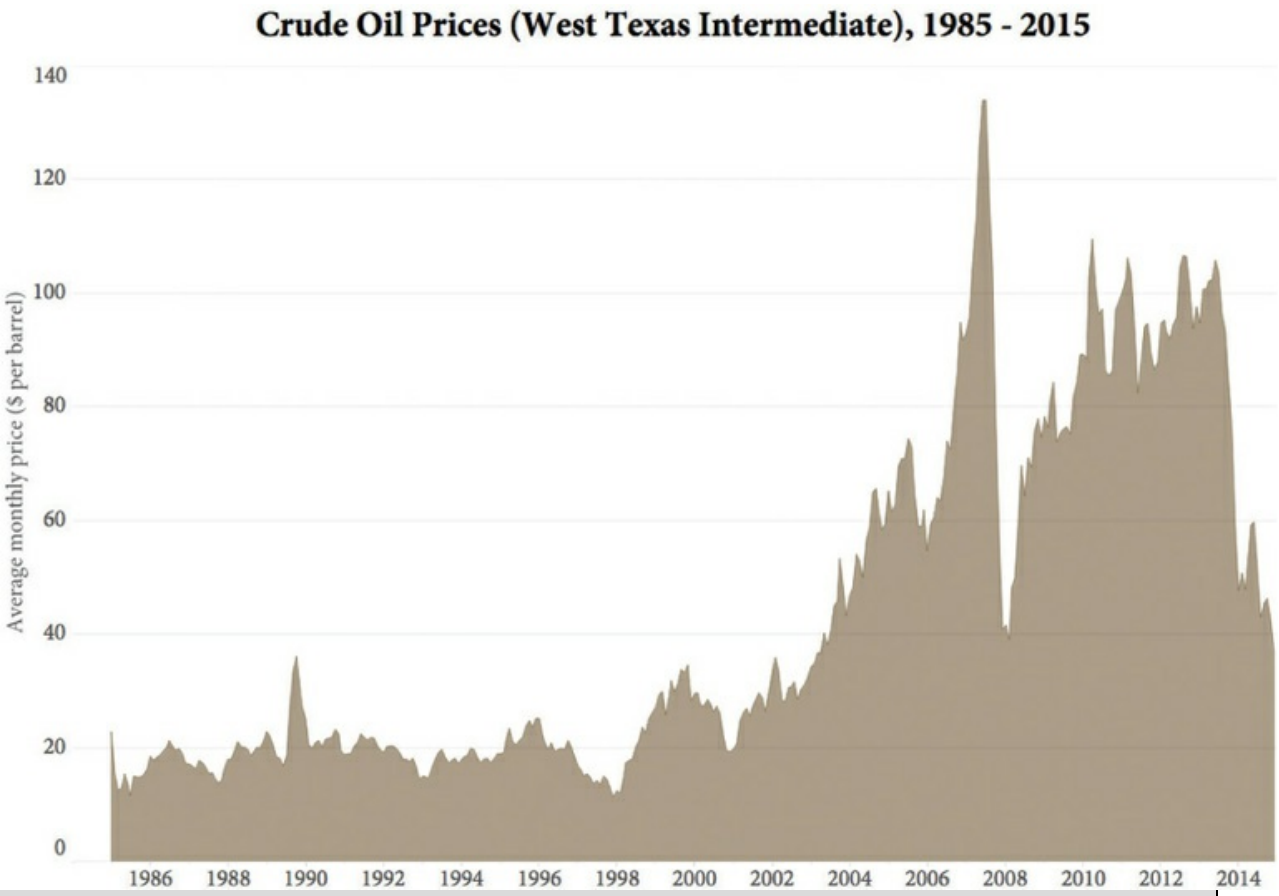
ALSO KNOWN AS

REPRESENTATION DESCRIPTION

A line chart shows how quantitative values for different categories have changed over time. They are typically structured around a temporal x-axis with equal intervals from the earliest to latest point in time. Quantitative values are plotted using joined-up lines that effectively connect consecutive points positioned along a y-axis. The resulting slopes formed between the two ends of each line provide an indication of the local trends between points in time. As this sequence is extended to plot all values across the time frame it forms an overall line representative of the quantitative change over time story for a single categorical value. To accentuate the magnitude of the quantitative values and the change through time the area beneath the line is filled with colour. The height of each coloured layer at each point in time reveals its quantity. Area charts can display values for several categories, using stacks, to show also the changing part-to-whole relationship.

EXAMPLE Showing changes in the average monthly price ($ per barrel) of crude oil between 1985 and 2015.

Figure 6.42 Crude Oil Prices (West Texas Intermediate), 1985—2015



HOW TO READ IT & WHAT TO LOOK FOR

Firstly, learn about the axes: what is the time period range presented on the x-axis (and in what order) and what is the range of quantitative values shown on the y-axis, paying particular attention to whether it is a percentage or absolute based scale? Inside the chart, determine what categories each area layer represents: for single areas this will usually be clear from the chart title, for multiple areas you might have direct labelling or a nearby legend to learn colour associations. Think about what high and low values mean: is it 'good' to be large/small, increasing or decreasing? Glance at the general patterns (especially if there are many layers), looking at the visible 'thickness' of the coloured layers. At what points are the values highest or lowest? When are they growing or shrinking as the time axis moves along? If there are multiple categories, which ones take up the largest and smallest slices of the overall total? Are there any trends (short or long term), any sudden moments of a rise or fall, any sense of seasonal or cyclical patterns? If there are multiple categories, look more closely at individual layers of interest.

PRESENTATION TIPS

ANNOTATION: Direct labelling of quantitative values will get far too busy so you might choose to annotate specific values of interest (highest, lowest, specific milestones). Think about the most useful interval for your axis labelling. As ever there is no single rule, so adopt the Goldilocks principle of not too many, not too few. If you have a stacked area chart, try directly to label the category layers shown as closely as possible (if the heights allow it) or at least ensure any colour associations are easily identifiable through a nearby legend. Think carefully about what is the most useful and meaningful interval for your time axis labelling.

COLOUR: If you are using a stacked area chart, ensure the categorical layers have sufficiently different colours so that their distinct reading can be efficiently performed.

COMPOSITION: Similar to the line chart, the area chart's dimensions should ideally utilise an aspect ratio that optimises the readability through 45° banking (roughly judging the average slope angle). The sequencing of values tends to be left to right for the sequence of the time-based x-axis and low rising to high

values on the y-axis; you will need a good (and clearly annotated) reason to break this convention. Unlike the line chart, the quantitative axis for area charts must start at zero as it is the height of the coloured areas under each line that helps readers to perceive the quantitative values. Do not have overlapping categories on the same chart because it makes it very difficult to see (imagine hills behind hills, peaking out and then hiding behind each other). Rather than stacking categories you might consider using small multiples, especially as this will present the respective displays from a common baseline (and make reading sizes a little easier).

VARIATIONS & ALTERNATIVES

Like area charts, 'alluvial diagrams' display proportional stacked layers for multiple categories showing the absolute value change over time. However, they also show the evolving ranks, switching the relative ordering of each layer of values based on the current magnitude. Some deployments of the area chart are not plotted over time but over continuous dimensions of space, perhaps showing the changing nature of a given quantitative measure along a given route. When you have many concurrent layers to show and these layers start and stop at different times, a 'slope graph' is worth considering.

Charts Trends

Horizon chart

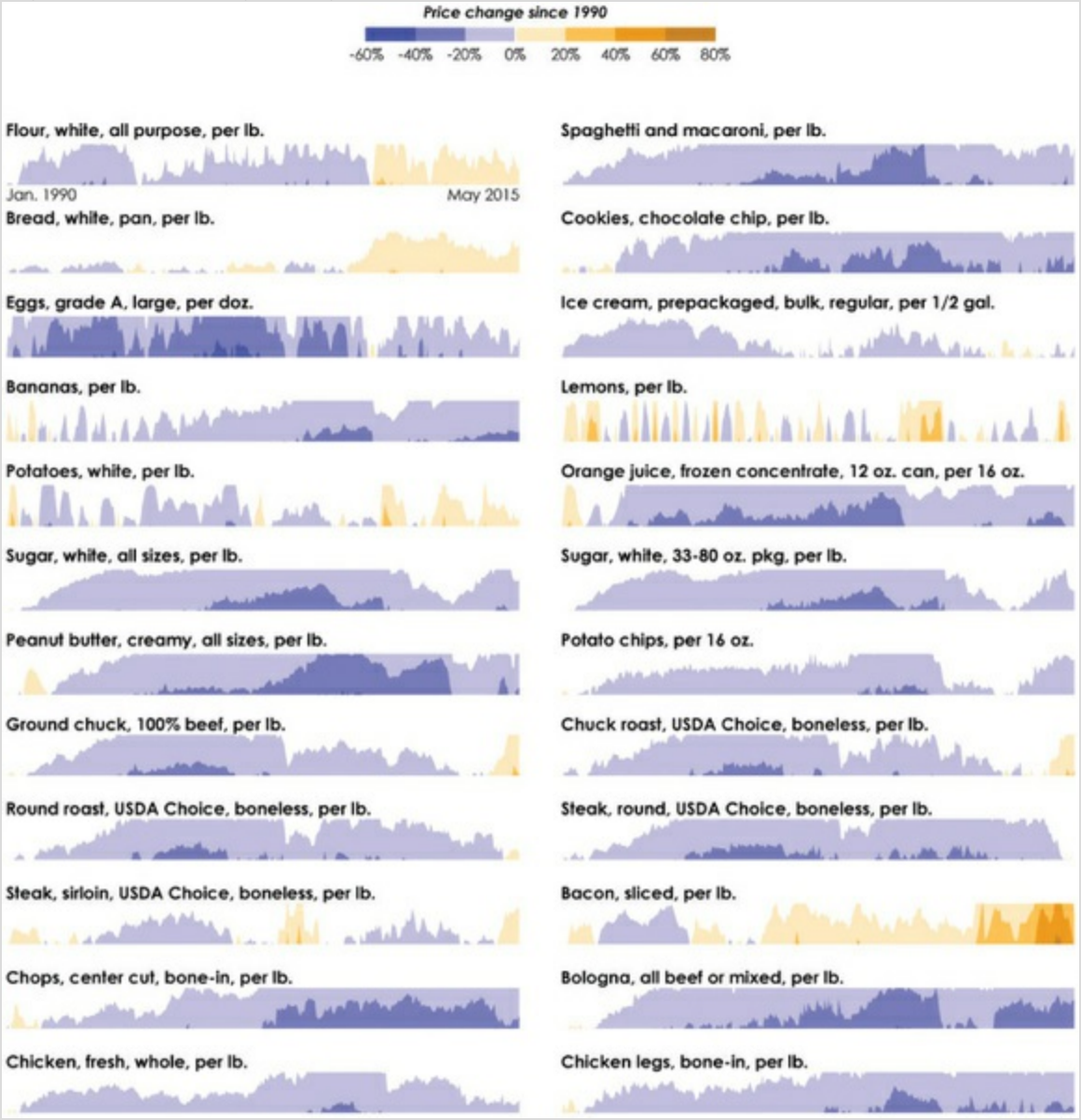
ALSO KNOWN AS

EXAMPLE Showing percentage changes in price for selected food items in the USA between 1990 and 2015.

REPRESENTATION DESCRIPTION

Horizon charts show how quantitative values for different categories have changed over time. They are valuable for showing changes over time for multiple categories within space-constrained formats (such as dashboards). They are structured around a series of rows each showing changes in quantitative values for a single category. The temporal x-axis has equal intervals from the earliest to latest point in time. Quantitative values are plotted using joined-up lines that connect consecutive points positioned along a value y-axis. The resulting slopes formed between the ends of each line provide an indication of the local trends between two points in time. As this sequence is extended to plot all values across the time frame it forms an overall line representative of the quantitative changes. To accentuate the magnitude of the quantitative values the area beneath the line is filled with colour. Negative values are highlighted in one colour, positive values in another colour. Variations in colour lightness are used to indicate different degrees or bands of magnitudes, with the extremes getting darker. Negative value areas are then flipped from underneath the baseline to above it, joining the positive values but differentiated in their polarity by colour. Finally, like slicing off layers of a mountain, each distinct threshold band that sits above the imposed maximum y-axis scale is chopped off and dropped down to the baseline, in front of its foundation base. The final effect shows overlapping layers of increasingly darker colour-shaded areas all occupying the same vertical space with combinations of height, colour and shade representing the values.

Figure 6.43 Percentage Change in Price for Select Food Items, Since 1990



HOW TO READ IT & WHAT TO LOOK FOR

Firstly, learn about the category rows: what do they represent and in what order are they presented? Next, the chart axes: what is the time period range presented on the x-axis (and in what order) and what is the range of quantitative values shown on the y-axis, paying attention to whether it is a percentage or absolute value scale? Next, what are the colour associations (for positive and negative values) and the different shaded banding thresholds? Think about what high and low values mean: is it 'good' to be large/small, increasing or decreasing? Glance at the general patterns over time, looking at the most visible dark areas of each colour polarity: where have values reached a peak in either direction? Maybe then separate your reading between looking at the positive value insights and then the negative ones: which chunks of colour are increasing in value (darker) or shrinking (getting lighter) as the time axis moves along? Where can you see most empty space, indicating low values? Are there any trends (short or long term), any sudden moments of a rise or fall, any sense of seasonal or cyclical patterns, any points of interest where lines cross each other or key thresholds that are reached/exceeded? Then look more closely at categories of interest, assessing their own patterns around specific moments in time and picking out the peak, low, earliest and latest values for each row.

PRESENTATION TIPS

ANNOTATION: The decisions around annotations are largely reduced to labelling the category rows.

Such is the busy-ness of the chart areas that any direct labelling is going to clutter the display too much: horizon charts are less about precise value reading and more about getting a sense of the main patterns, so avoid the temptation to over-label. Think carefully about what is the most useful and meaningful interval for your time axis labelling.

COLOUR: Colour decisions mainly concern the choices of quantitative scale bandings to show the positive and negative value ranges.

COMPOSITION: The height of the chart area in which you can accommodate a single row of data will have an influence on the entire construction of the horizon chart. It will often involve an iterative/trial and error process, looking at the range of quantitative values across each category, establishing the most sensible and meaningful thresholds within these range and then fixing the y-axis scales accordingly. Try to ensure the sorting of the main categorical rows is as logical and meaningful as possible.

VARIATIONS & ALTERNATIVES

An alternative to the horizon chart is the entry-level single category 'area chart', which does not suffer the same constraints of restrictions to the vertical scale. For space-constrained displays, 'spark lines' would offer an option suitable to such situations and easily accommodate multiple category displays.

Charts Trends

Stream graph

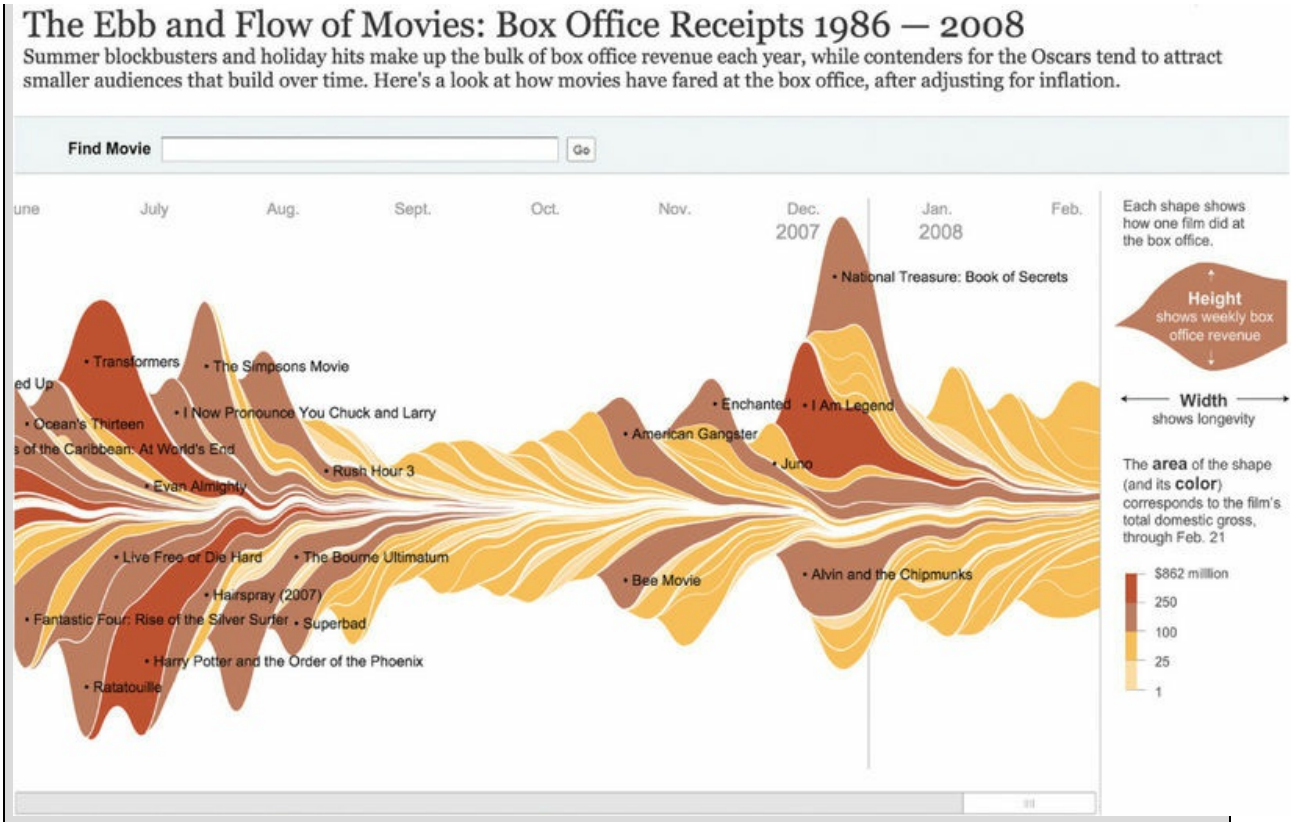
ALSO KNOWN AS Theme river

REPRESENTATION DESCRIPTION

A stream graph shows how quantitative values for different categories have changed over time. They are generally used when you have many constituent categories at any given point in time and these categories may start and stop at different points in time (rather than continue throughout the presented time frame). As befitting the name, their appearance is characterised by a flowing, organic display of meandering layers. They are typically structured around a temporal x-axis with equal intervals from the earliest to latest point in time. Quantitative values are plotted using joined-up lines that effectively connect consecutive points to quantify the height above a local baseline, which is not a stable zero baseline but rather a shifting shape formed out of other category layers. To accentuate the size of the category's height at any given point the area beneath the line is filled with colour. The height of each coloured layer at each point in time reveals its quantity. This colour is often used to further represent a quantitative value scale or to associate with categorical colours. The stacking arrangement of the different categorical streams goes above and below the central axis line to optimise the layout but not with any implication of polarity.

EXAMPLE Showing changes in the total domestic gross takings ($US) and the longevity of all movies released between 1986 and 2008.

Figure 6.44 The Ebb and Flow of Movies: Box Office Receipts 1986—2008



HOW TO READ IT & WHAT TO LOOK FOR

Firstly, determine what is the time period presented on the x-axis (and in what order). In most stream graphs you do not see the quantitative y-axis scale because the level of reading is more about getting a gist for the main patterns in a relative sense rather than an absolute one. You might find that the colouring of layers has a quantitative scale or categorical association so look for any keys. Also, you will often find guides to help estimate the quantitative heights of each layer. Think about what high and low values mean: is it 'good' to be large/small, increasing or decreasing? Glance at the general patterns over time. Remember that above or below means nothing in the sense of polarity of values, so your focus is on the entirety of the collective shape. Look for the largest peaks and the shallowest troughs, possible seasonal patterns or the significant moments of change. Note where these patterns occur in relation to the timescale. Can you see any prominently tall (big values) or wide (long-duration) layers? Notice when layers start and end, noting times when there are many concurrent categories and when there are few. Pick out the layers of personal interest and assess their patterns over time. Do not spend too much effort trying to estimate precise values of height, but keep your focus on the bigger picture level. It is often useful to rotate the display so the streams are travelling vertically, offering a different perspective and removing the instinct to see positive values above and negative values below the central axis.

PRESENTATION TIPS

INTERACTIVITY: If interactivity is a possibility, this could enable selection or mouseover events to reveal annotated values at any given point in time or to filter the view.

ANNOTATION: Chart apparatus devices are generally of limited use in a stream graph with the priority on a general sense of pattern more than precision value reading. Direct labelling of categories is likely to be quite busy but may be required, at least to annotate the most interesting patterns (highest, lowest, specific milestones). Think carefully about what is the most useful and meaningful interval for your time axis labelling.

COLOUR: Ensure any colour associations or size guides are easily identifiable through a nearby legend.

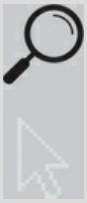
COMPOSITION: Composition choices are firstly concerned with the landscape or portrait layout. This

will largely be informed by the format and space of your outputs and the meaning of the data. The stream layers are often smoothed, giving them an aesthetically organic appearance, both individually and collectively. This is achieved via curved line interpolation.

VARIATIONS & ALTERNATIVES

The fewer categorical series you have in your data, the more likely a stacked 'area chart' is going to best-fit your needs. You could consider a stacked 'bar chart' over time also, but there is less chance of maintaining the connected visibility of continuous categorical series via a singular shape.

Charts Activities



Connected timeline

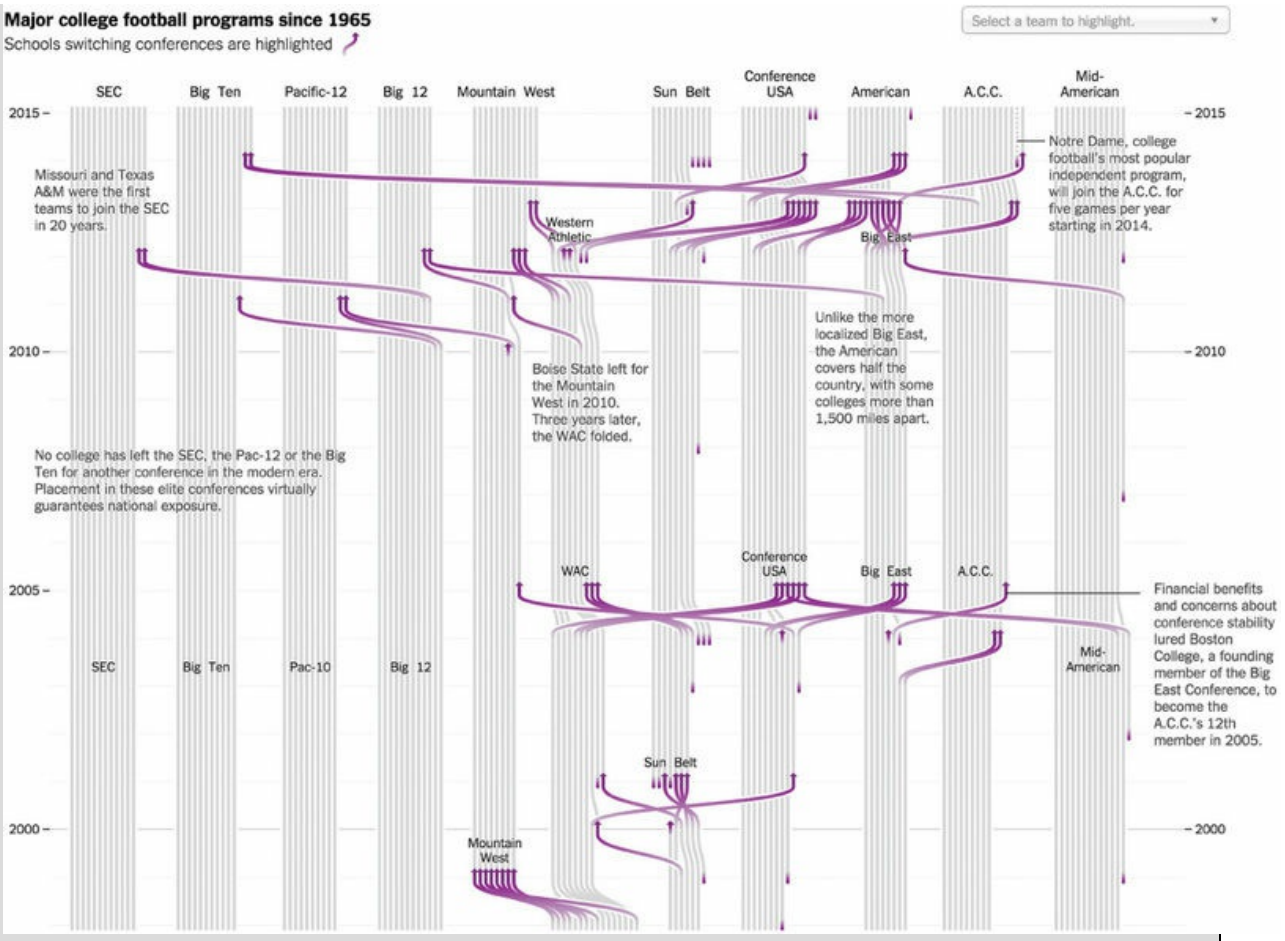
ALSO KNOWN AS Relationship timeline, storyline visualisations, swim-lane chart

REPRESENTATION DESCRIPTION

A connected timeline displays the duration, milestones and categorical relationships across a range of categorical 'activities'. It represents a particularly diverse and creative way of showing changes over time and so involves many variations in approach. The structure is generally formed of time-based quantitative x-axis and categorical y-axis lanes. Each categorical activity will commence at a point in time and from within a vertical category 'family'. Over time, the line will progress, possibly switching to a different categorical lane position as the nature of the activity alters. The lines may be of fixed width or proportionally weighted to represent a quantitative measure. Some activity lines may cease, restart or merge with others to build a multi-faceted narrative. Colour can also be used to present further relevant detail. The main issue with any connected timeline approach is simply the complexity of the content and the number of moving parts crossing over the display. As there are many entry points into reading such a timeline there can be inefficiency in the reading process, but this is usually proportional simply to the subject at hand and you may not wish to see these nuances being removed.

EXAMPLE Showing changes in US major college football programme allegiance to different conferences between 1965 and 2015.

Figure 6.45 Tracing the History of N.C.A.A. Conferences



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know what the major categorical 'lanes' represent and what the range of date values is (min to max). Then try to determine what each categorical activity line represents. As there are so many derivatives there is no single reading strategy, but generally glance across the entire chart noting the sequence of the activities; there is usually a sequential logic attached to their sorting based on the start date milestone in particular. Follow the narrative from left to right, noting observations about any big, small and medium weighted lines and spotting any moment when they connect with, overlap or detach from other activities. Are there any major convergences or divergences in pattern? Any hubs of dense activity and other sparse moments? Look for the length of lines to determine the long, medium and short durations of activity. Where available, compare the activities against annotated references about other key milestone dates that might hold some significance or influence.

PRESENTATION TIPS

ANNOTATION: Chart apparatus devices like tick marks and gridlines in particular can be helpful to increase the accuracy of the reading of both the quantitative values and the activity 'lanes', which may be coloured to help recognise divisions between categories. Direct labelling is usually seen in these timelines to help maintain associations across the display with the categories of characters or activities, perhaps annotating the consequence or cause of lines merging, etc. Think carefully about what is the most useful and meaningful interval for your time axis labelling.

COLOUR: Even if colour does not have a direct association with given activities, it can be a useful property to highlight certain features of the narrative, sometimes acting as a container device to group activities together, even if just for a momentary time period.

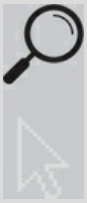
COMPOSITION: Where possible, try to make the categorical sorting meaningful, maybe organising values in ascending/descending size order. The vertical (y) or horizontal (x) sequencing of time will depend on the amount of data to show and the space you have to work with. Also, depending on the narrative, the

past > present ordering may be reversed.

VARIATIONS & ALTERNATIVES

There are similarities with the organic nature of the 'alluvial diagram', which shows ranking and quantitative change over time for a number of concurrent categories. When there are fewer inter-activity relationships and more discrete categories are involved, then the 'Gantt chart' offers an alternative way of showing this analysis.

Charts Activities



Gantt chart

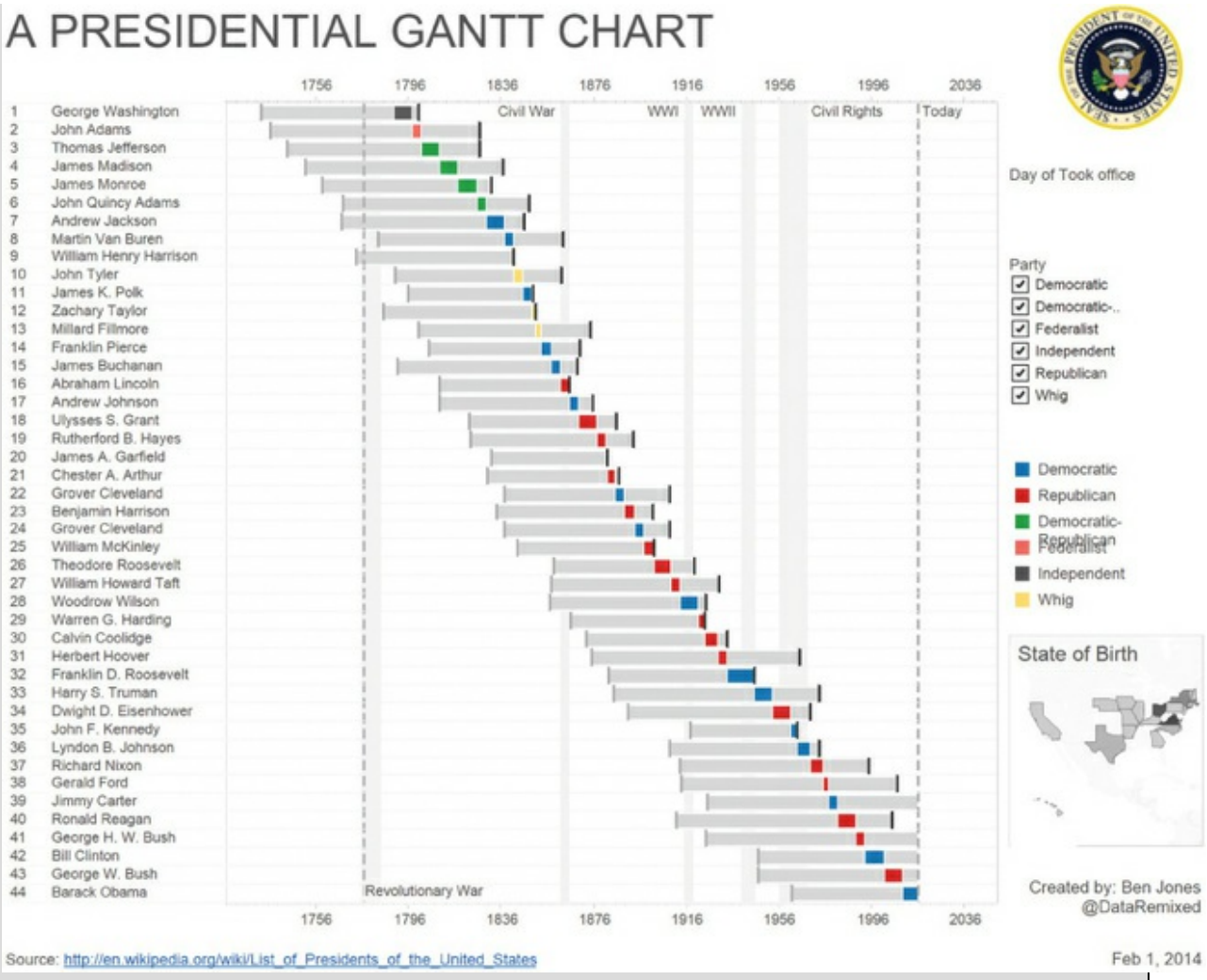
ALSO KNOWN AS Range chart, floating bar chart

REPRESENTATION DESCRIPTION

A Gantt chart displays the start and finish points and durations for different categorical 'activities'. The display is commonly used in project management to illustrate the breakdown of a schedule of tasks but can be a useful device to show any data based on milestone dates and durations. The chart is structured around a time-based quantitative x-axis and a categorical y-axis. Each categorical activity is represented by lines positioned according to the start moment and then stretched out to the finish point. There may be several start/finish durations within the same activity row. Sometimes points are used to accentuate the start/finish positions and the line may be coloured to indicate a relevant categorical value (e.g. separating completed vs ongoing).

EXAMPLE Showing the events of birth, death and period serving in office for the first 44 US Presidents.

Figure 6.46 A Presidential Gantt Chart



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with what major categorical values each Gantt bar is associated and what the range of the date values is (min to max). Follow the narrative, noting the sequence of the categories – there is usually a sequential sorting based on the start date milestone. Glance across the entire chart and perform global comparisons to establish the high-level ranking of biggest > smallest durations (based on the length of the line) as well as early and late milestones. Identify any noticeable exceptions and/or outliers. Perform local comparisons between neighbouring bars to identify proportional differences and any connected dependencies. Estimate (or read, if labels are present) the absolute values for specific categories of interest. Where available, compare the activities against annotated references about other key milestone dates that might hold some significance or influence.

PRESENTATION TIPS

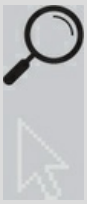
ANNOTATION: Chart apparatus devices like tick marks and gridlines (or row band-shading) in particular can be helpful to increase the accuracy of the reading of the start point and duration of activities along the timeline. If you have axis labels you may not need direct labels for the values shown with each duration bar – this will be label overload, so generally decide between one or the other. Think carefully about what is the most useful and meaningful interval for your time axis labelling.

COMPOSITION: There is no significant difference in perception between vertical or horizontal Gantt charts, though horizontal layouts are more metaphorically consistent with the concept of reading time. Additionally, these layouts tend to make it easier to accommodate and read the category labels. Where possible, try to sequence the categorical 'activities' in a way that makes for the most logical reading, either organised by the start/finish dates or maybe the durations (depending on which has most relevance).

VARIATIONS & ALTERNATIVES

Variations might involve the further addition of different point markers (represented by combinations of symbols and/or colours) along each activity row to indicate additional milestone details, using the 'instance chart'. An emerging trend in technique terms involves preserving the position of activity lines adjacent to other concurrent activities, rather than fixing them to stay within discrete rows. Sometimes there is much more fluidity and less 'discreteness' in the relationships between activity, so approaches like the 'connected timeline' may be more fitting.

Charts Activities



Instance chart

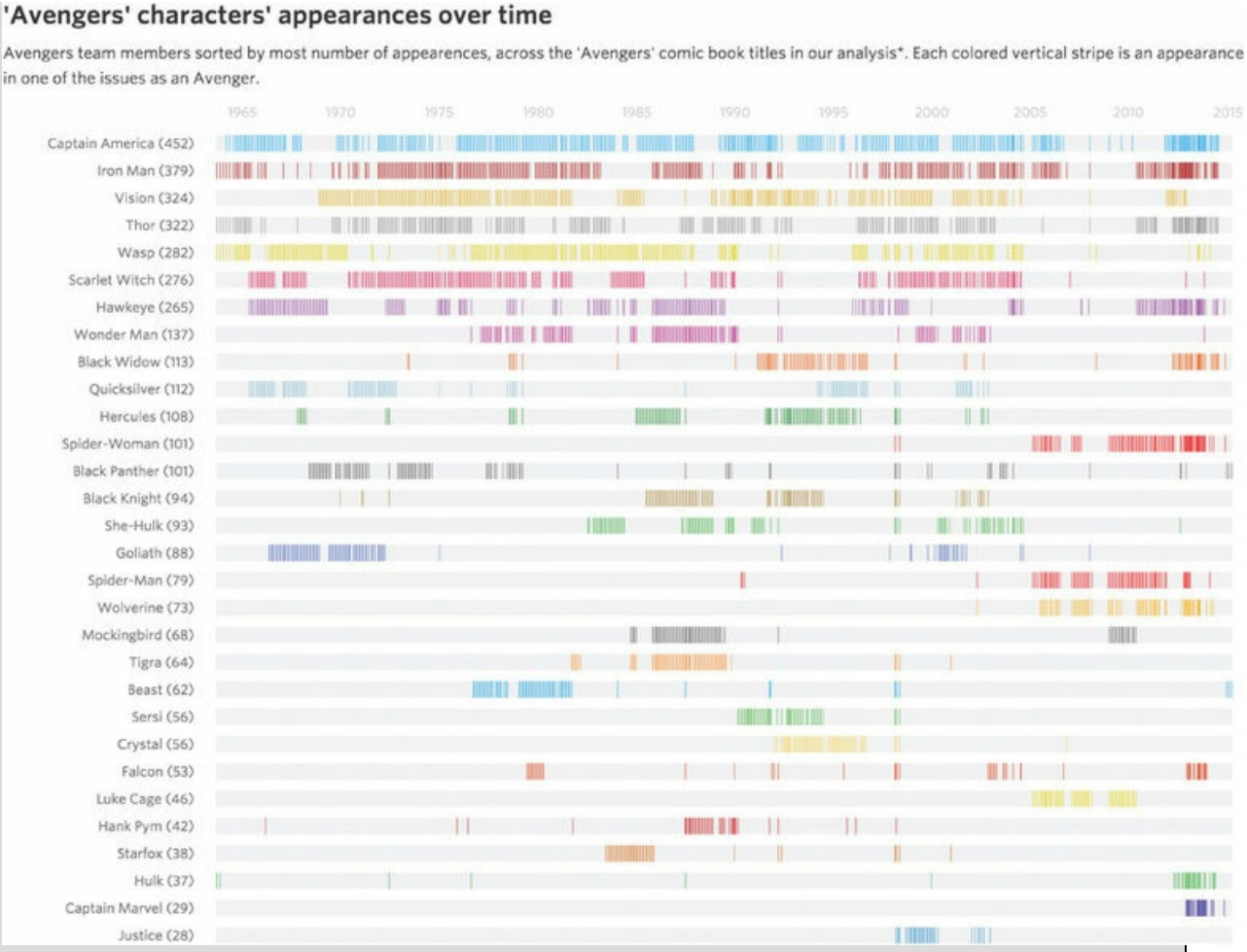
ALSO KNOWN AS Milestone map, barcode chart, strip plot

REPRESENTATION DESCRIPTION

An instance chart displays individual moments or instances of categorical 'activities'. There are many variations in approach for this kind of display but generally you will find a structure based on a time-based quantitative x-axis and a categorical y-axis. For each categorical activity, instances of note are represented by different point markers that indicate along the timeline when something has happened. The point markers may have different combinations of symbols and colours to represent different types of occurrences, but avoid having too many different combinations so that viewers do not have to learn an entirely new alphabet of meaning.

EXAMPLE Showing the instances of different Avengers characters appearing in Marvel's comic book titles between 1963 and 2015.

Figure 6.47 How the 'Avengers' Line-up Has Changed Over the Years



HOW TO READ IT & WHAT TO LOOK FOR

Look at the axes so you know with what major categorical values each row of instances is associated and what the range of the date values is (min to max). Look up any legend that will explain what (if any) associations exist between the instance markers and their colour/symbol. Glance down the y-axis noting the sequence of the categories; there is usually a sequential logic attached to their sorting based on the start date milestone in particular. Follow the narrative, noting observations about the type and frequency of instances being plotted. Look across the entire chart to locate the headline patterns of clustering and identify any noticeable exceptions and/or outliers. Look across the patterns within each row individually to learn about each category's dispersal of instances. Look for empty regions where no marks appear. How do all these patterns relate to the time frame displayed? Where available, compare the activities against annotated references about other key milestone dates that might hold some significance or influence.

PRESENTATION TIPS

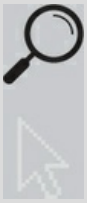
ANNOTATION: The main annotation properties will be used to serve the role of explaining the associations between marks and attributes through clear legends/keys.

COMPOSITION: Where possible, try to sequence the categorical 'activities' in a way that makes for the most logical reading, either organised by the start/finish dates or maybe the durations (depending on which has most relevance).

VARIATIONS & ALTERNATIVES

Some variations may see the size of a geometric shape used instead of just a point to indicate also a quantitative measure to go with the instance. The marking of an instance through a 'when' moment could also be based on data that talks about positional moments within a sequence. If the basic activity is reduced to a start/finish moment then the 'Gantt chart' will be the best-fit option.

Charts Overlays



Choropleth map

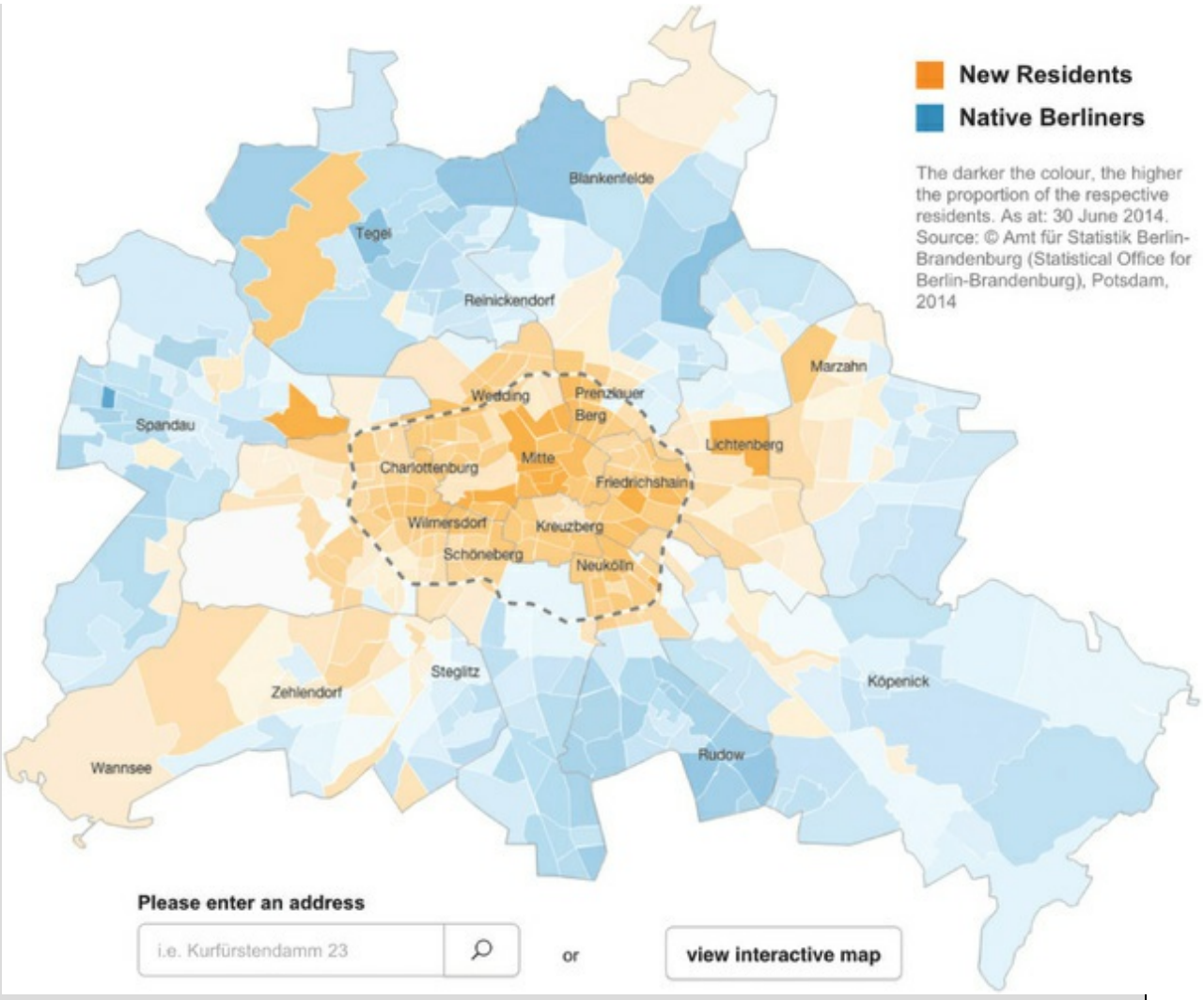
ALSO KNOWN AS Heat map

REPRESENTATION DESCRIPTION

A choropleth map displays quantitative values for distinct, definable spatial regions on a map. Each geographic region is represented by a polygonal area based on its outline shape, with each distinct shape then collectively arranged to form the entire landscape. (Note that most tools for mapping have a predetermined reference between a region name and the dimensions of the regional polygon.) Each area is colour-coded to represent a quantitative value based on a scale with colour variation intervals that (typically) go from a light tint for smaller values to a dark shade for larger values. Choropleth maps should only be used when the quantitative measure is directly associated with and continuously relevant across the spatial region on which it will be displayed. Similarly, if your quantitative measure is about or related to the consequence of more people living in an area, interpretations may be distorting, so consider transforming your data to per capita or per acre (or other spatial denominator) to standardise the analysis accordingly.

EXAMPLE Mapping the percentage change in the populations of Berlin's districts across new and native Berliners since the fall of the Berlin Wall.

Figure 6.48 Native and New Berliners — How the S-Bahn Ring Divides the City



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Establish the colour-scale value associations, usually found via a legend. Glance across the entire chart to locate the dark, light and medium shades (generally darker = larger) and perform global comparisons to establish the high-level ranking of biggest values > smallest. Identify any noticeable exceptions and/or outliers. Beware making judgements about the significance of prominent large geographical areas: size is an attribute of the underlying region, not the significance of the measure displayed. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate (or read, if labels are present) the absolute values of specific regions of interest.

PRESENTATION TIPS

ANNOTATION: Directly labelling the regional areas with geographical details and the value they hold is likely to lead to too much clutter. You might include only a limited number of regional labels to provide spatial context and orientation.

COLOUR: Legends explaining the colour scales should ideally be placed as close to the map display as possible. The border colour and stroke width for each spatial area should be distinguishable to define the shape but not so prominent as to dominate attention – usually a subtle grey- or white-coloured thin stroke will be fine. As well as variation in colour scales, sometimes pattern or textures may add an extra layer of detail to the value status of each region. When including a projected mapping layer image in the background, ensure it is not overly competing for visual prominence by making it light in colour and possibly semi-transparent. Do not include any unnecessary geographical details that add no value to the

spatial orientation or interpretation and clutter the display (e.g. roads, building structures).

COMPOSITION: With Earth being a sphere, there are many different mapping projections for representing the regions of the world on a plane surface. Be aware that the transformation adjustments made by some map projections can distort the size of regions of the world, inflating their size relative to other regions.

VARIATIONS & ALTERNATIVES

Some choropleth maps may be used to indicate categorical association rather than quantitative measurements. Alternative thematic mapping approaches to representing quantitative values might include the 'proportional symbol map' and the 'dot density map'. This is a variation that involves plotting a representative quantity of dots equally (but randomly) across and within a defined spatial region. The position of individual dots is therefore not to be read as indicative of precise locations but used to form a measure of quantitative density. This offers a useful alternative to the choropleth map, especially when categorical separation of the dots through colour is of value. 'Dasymetric mapping' is similar in approach to choropleth mapping but breaks the constituent regional areas into much more specific, almost custom-drawn, sub-regions to better represent the realities of the distribution of human and physical phenomena within a given spatial boundary.

Charts Overlays

Isarithmic map

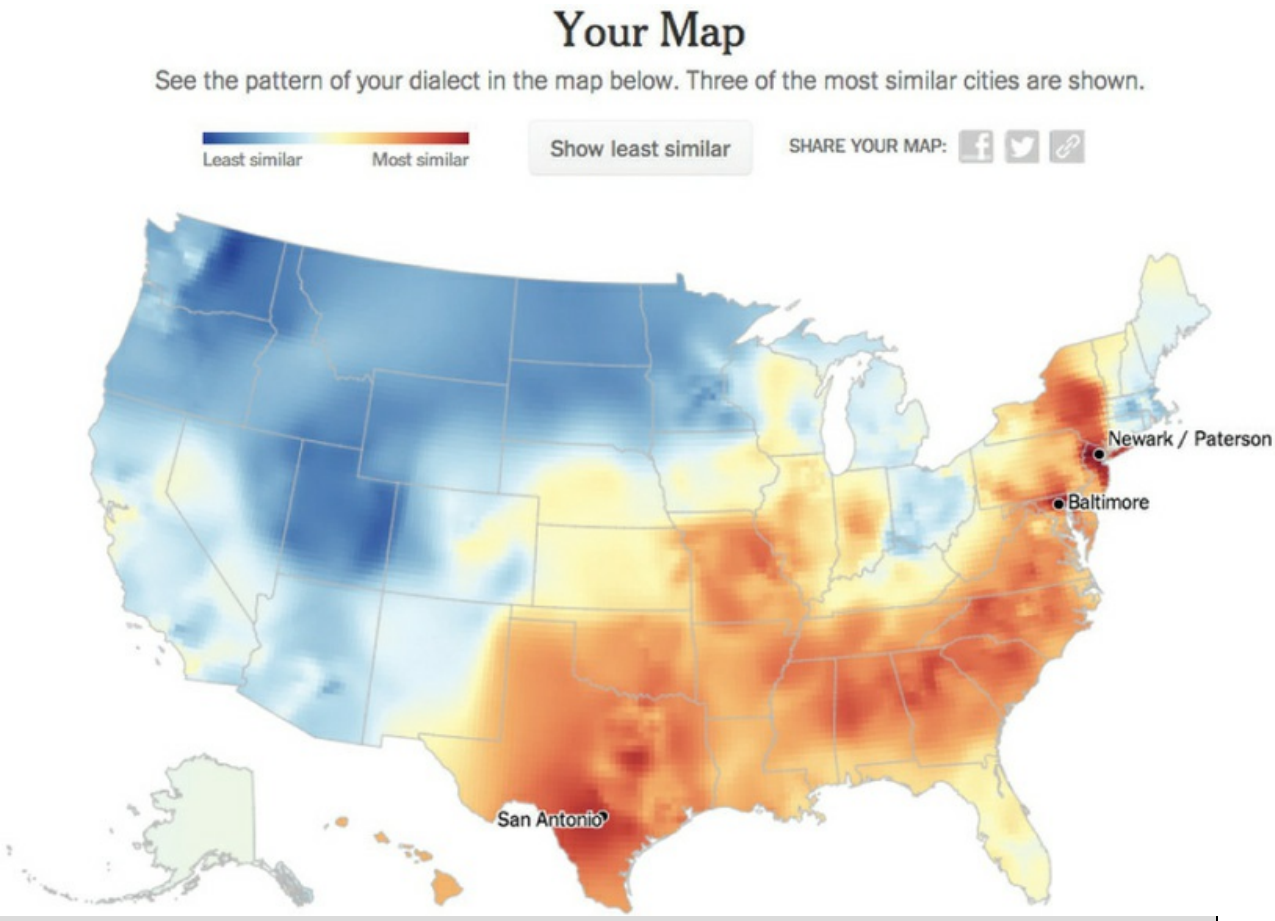
ALSO KNOWN AS Contour map, isopleth map, isochrone map

REPRESENTATION DESCRIPTION

An isarithmic map displays distinct spatial surfaces on a map that share the same quantitative classification. All spatial regions (transcending geo-political boundaries) that share a certain quantitative value or interval are formed by interpolated 'isolines' connecting points of similar measurement to form distinct surface areas. Each area is then colour-coded to represent the relevant quantitative value. The scale of colour variation intervals differs between deployments but will typically range from a light tint for smaller values to a dark shade for larger values. An isarithmic map would be used in preference to a choropleth map when the patterns of data being displayed transcend the distinct regional polygons. They could be used to show temperature bandings or smoothed regions of political attitudes.

EXAMPLE Mapping the degree of dialect similarity across the USA.

Figure 6.49 How Y'all, Youse and You Guys Talk



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Establish the colour scale value associations, usually found via a legend. Glance across the entire chart to locate the dark, light and medium shades (generally darker = larger) and perform global comparisons to establish the high-level ranking of biggest values > smallest. Identify any noticeable exceptions and/or outliers, including regions that appear in isolation from their otherwise related values and notable for their position adjacent to very different shaded regions. Note that any interpolation used to smooth the joins between data points to form organic surfaces will inevitably reduce the precision of the surfaces in their relationship to land position. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate the absolute values of specific regions of interest.

PRESENTATION TIPS

ANNOTATION: Directly labelling the surface areas to show the quantitative value or range they represent will be too cluttered. You might include only a limited number of regional labels to provide spatial context and orientation.

COLOUR: Legends explaining the colour scales should ideally be placed as close to the map display as possible. If using visible contour or boundary lines there is a clear implication of a location being inside or outside the line, so make these lines as prominent in colour as possible according to the precision of their representation. If the smoothing of the surface locations has been applied the representation of these areas should similarly avoid looking definitive. You therefore might consider subtle colour gradation/overlapping between different regions to capture appropriately the underlying 'fuzziness' of the data. As well as colour scales, sometimes pattern or textures may add an extra layer of detail to the value status of each surface region. When including a projected mapping layer image in the background, ensure it is not overly competing for visual prominence by making it light in colour and possibly semi-transparent. Do not include any unnecessary geographical details that add no value to the spatial orientation or interpretation and clutter

the display (e.g. roads, building structures).

COMPOSITION: Be aware that the transformation adjustments made by some map projections can distort the size of regions of the world, inflating their size relative to other regions.

VARIATIONS & ALTERNATIVES

There are specific applications of isarithmic maps used for showing elevation ('contour maps'), atmospheric pressure ('isopleth maps') or travel–time distances ('isochrone maps'). Sometimes you might use isarithmic maps to show a categorical status (perhaps even a binary state) rather than a quantitative scale.

Charts Overlays

Proportional symbol map

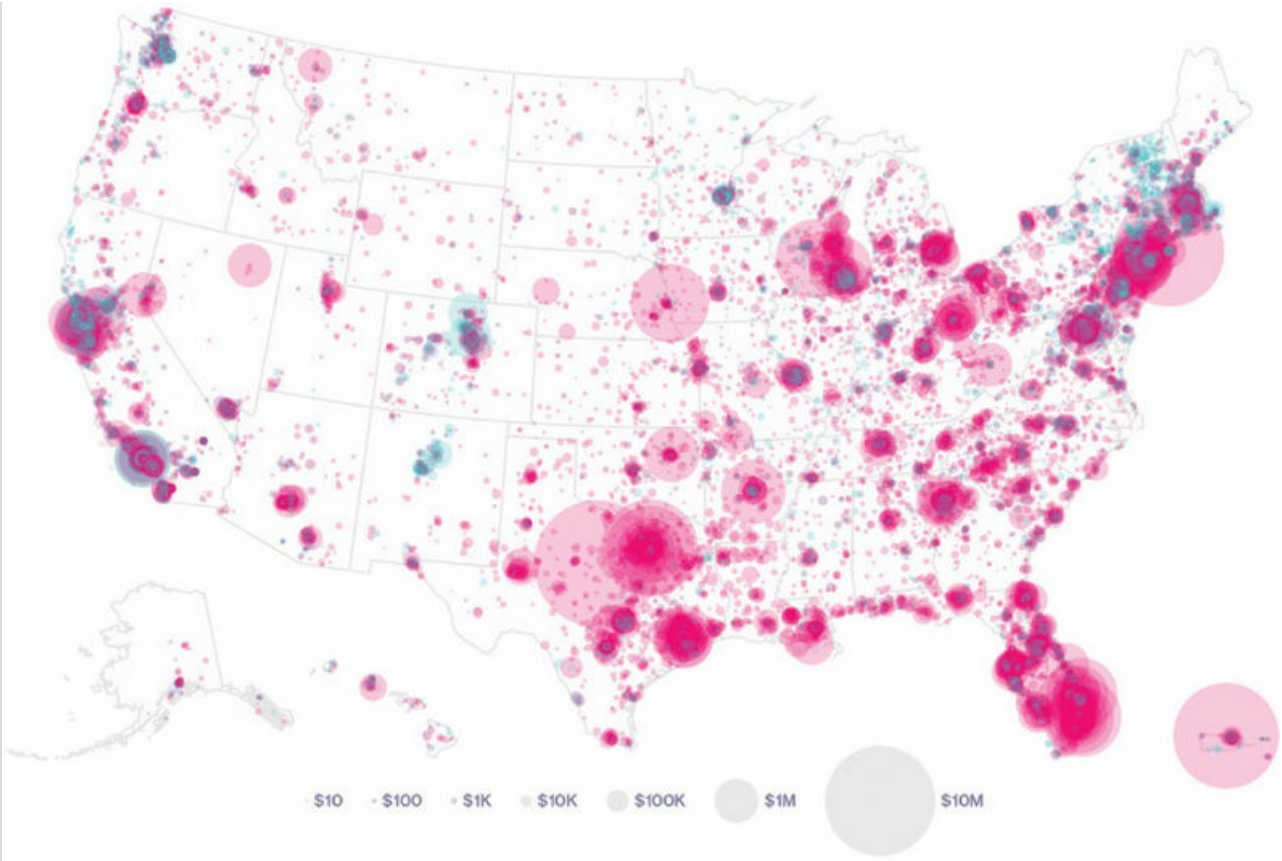
ALSO KNOWN AS Graduated symbol map

REPRESENTATION DESCRIPTION

A proportional symbol map displays quantitative values for locations on a map. The values are represented via proportionally sized areas (usually circles), which are positioned with the centre mid-point over a given location coordinate. Colour is sometimes used to introduce further categorical distinction.

EXAMPLE Mapping the origin and size of funds raised across the 22 major candidates running for US President during the first half of 2015.

Figure 6.50 Here's Exactly Where the Candidates' Cash Came From



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Establish the area size value associations, usually found via a legend. Glance across the entire chart to locate the large, medium and small shapes and perform global comparisons to establish the high-level ranking of biggest values > smallest. Identify any noticeable exceptions and/or outliers. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate (or read, if labels are present) the absolute values of specific regions of interest. Also note where there are no markers. If colour is being used to further break down the categories of the values shown, identify any grouped patterns that emerge.

PRESENTATION TIPS

INTERACTIVITY: Interaction may be helpful to reveal location and value labels through selection or mouseover events.

ANNOTATION: Directly labelling the shapes with geographical details and the value they hold is likely to lead to too much clutter. You might therefore include only a limited number of regional labels to provide spatial context and orientation. Legends explaining the size scales – and any colour associations – should ideally be placed as close to the map display as possible. Avoid including unnecessary geographical details that add no value to the spatial orientation or interpretation and clutter the display (e.g. roads, building structures).

COLOUR: Sometimes the circular shapes are filled, at other times they remain unfilled. If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. When a circle has a large value its shape will transgress well beyond the origin of its geographical location, intruding on and overlapping with other neighbouring values. The use of outline borders and semi-transparent colours helps with the task of avoiding occlusion (visually hiding values behind others). When including a projected mapping layer image in the background, ensure it is not overly competing for visual prominence by making it light in colour and possibly semi-transparent.

VARIATIONS & ALTERNATIVES

Variations may see the typical circle replaced by squares and geographical space replaced by anatomical regions. Alternatives to the proportional symbol map include the 'choropleth map', which colour-codes regions, or the 'dot map', which uses a dot to represent an instance of something. Avoid the temptation to turn the circle symbols into pie charts; it is not a good look. If you absolutely positively have to show a part-to-whole relationship, only show two categories, as per the recommended practice for pies.

Charts Overlays

Prism map

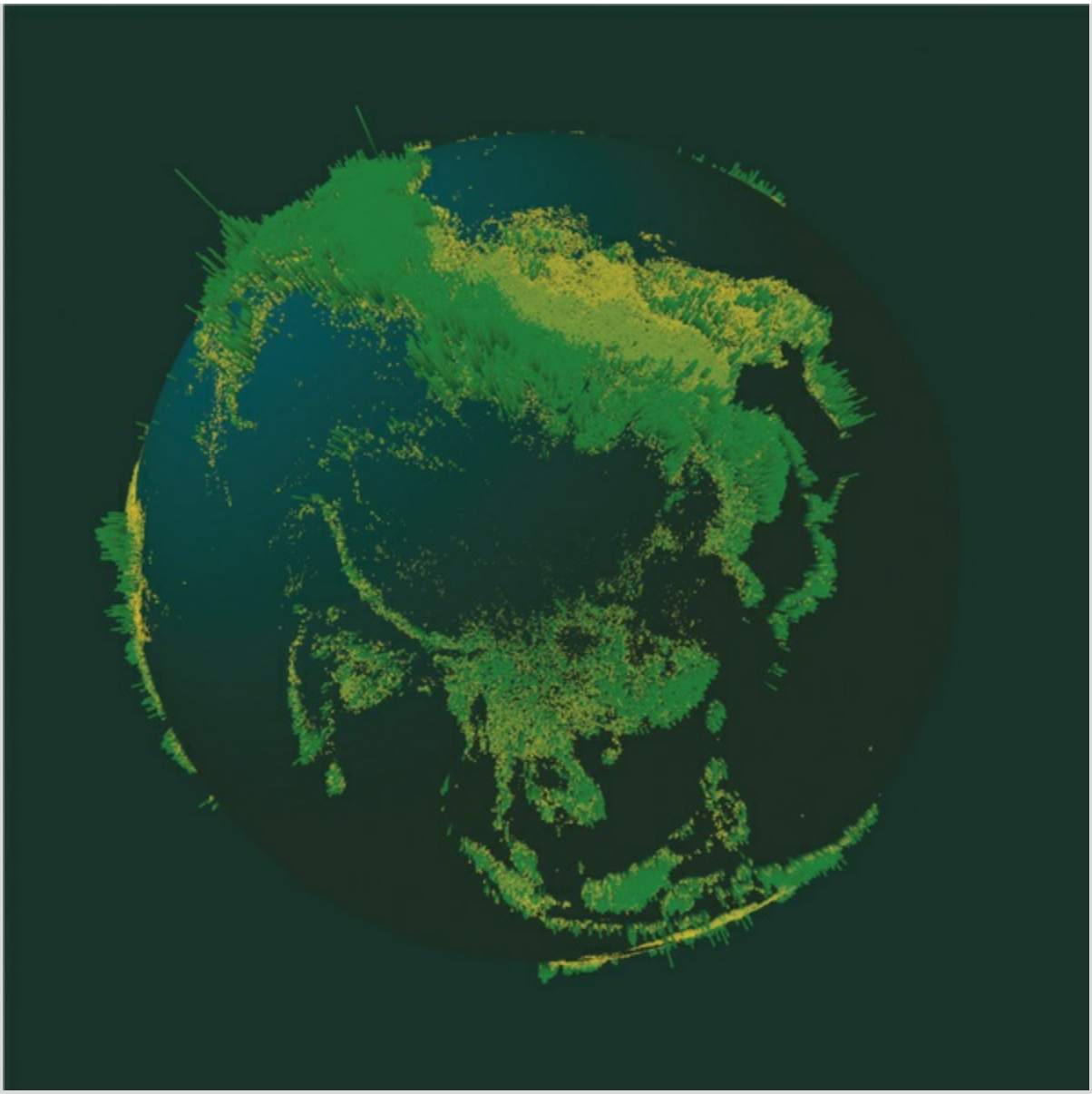
ALSO KNOWN AS Isometric map, spike map, datascape

REPRESENTATION DESCRIPTION

A prism map displays quantitative values for locations on a map. The values are represented via proportionally sized lines, appearing as 3D bars, that typically cover a fixed surface area of space and are just extended in height proportionally to represent the quantitative value for that location. Being able to judge the dimensions of 3D forms in a 2D view is very difficult, so they are only ever really used to create a gist of the profile of values, enabling recognition of the main peaks in particular.

EXAMPLE Mapping the population of trees for each 180 square km of land across the globe.

Figure 6.51 Trillions of trees



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Establish the area size value associations, usually found via a legend. Glance across the entire chart to locate the large, medium and small shapes and perform global comparisons to establish the high-level ranking of biggest values > smallest. Identify any noticeable exceptions and/or outliers. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate (or read, if labels are present) the absolute values of specific regions of interest. Also note where there are no bars emerging from the surface.

PRESENTATION TIPS

INTERACTIVITY: Ideally prism maps would be provided with interactive features that allow panning around the map region to offer different viewing angles to overcome the perceptual difficulties of judging the dimensions of 3D forms in a 2D view. Without this, smaller values will be hidden behind the larger forms, just as smaller buildings are hidden by skyscrapers in a city.

ANNOTATION: Directly labelling the prism shapes is infeasible – at most you might include only a limited number of labels to provide spatial context and orientation against the largest forms. Legends explaining the size scales should ideally be placed as close to the map display as possible.

COLOUR: Most tools that enable this type of mapping will likely have visual property settings for a faux light effect, helping the physical shapes to emerge more prominently through light and shadow. Ensure colour assist in helping the shape of the forms to be as visible as possible, maybe with opacity to enable smaller values to be not entirely hidden behind any larger ones. When including a mapping layer image on the surface, ensure it is not overly competing for visual prominence by making it light in colour and possibly semi-transparent. Do not include any unnecessary geographical details that add no value to the spatial orientation or interpretation and clutter the display (e.g. roads, building structures).

COMPOSITION: Be aware that the transformation adjustments made by some map projections can distort the size of regions of the world, inflating their size relative to other regions.

VARIATIONS & ALTERNATIVES

Alternatives to the prism map, especially to avoid 3D form, include the 'proportional symbol map', which uses proportionally sized geometric shapes, and the 'choropleth map', which colour-codes regional shapes.

Charts Overlays

Dot map

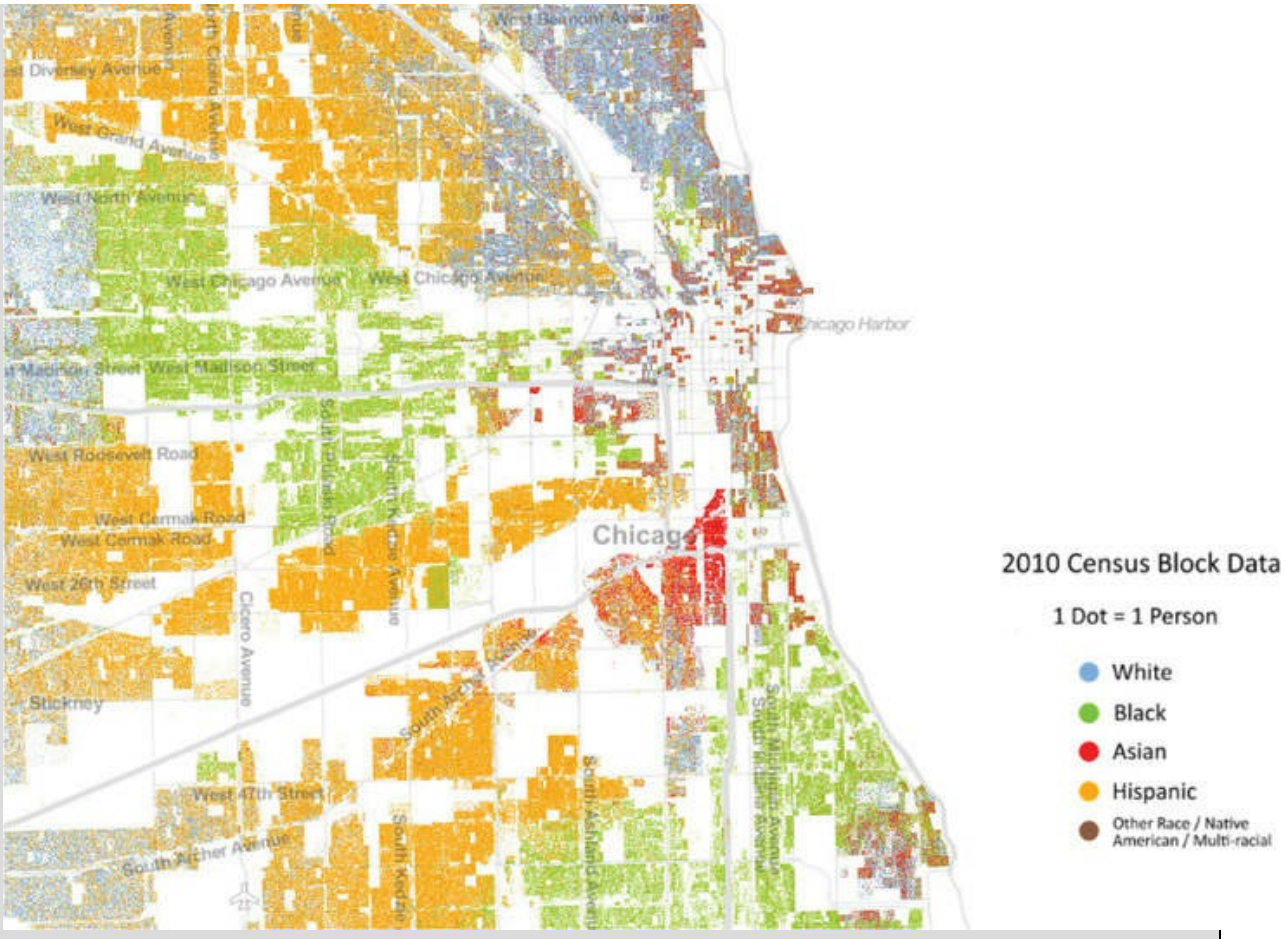
ALSO KNOWN AS Dot distribution map, pointillist map, location map, dot density map

REPRESENTATION DESCRIPTION

A dot map displays the geographic density and distribution of phenomena on a map. It uses a point marker to indicate a categorical 'observation' at a geographical coordinate, which might be plotting instances of people, notable sites or incidences. The point marker is usually a filled, small dot. Colour can be used to distinguish categorical classifications. Sometimes a dot represents a one-to-one phenomenon (i.e. a single record at that location) and sometimes a dot will represent one-to-many phenomena (i.e. for an aggregated statistic whereby the location represents a logical mid-point). As the proliferation of GPS recording devices increases, the accuracy and prevalence of detailed location marked incidences are leading to increased potential for this type of approach. However, think carefully about the potential sensitivity of directly plotting a phenomenon or data incidence at a given location.

EXAMPLE Mapping each resident of the USA based on the location at which they were counted during the 2010 Census across different ethnicities.

Figure 6.52 The Racial Dot Map



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the phenomenon that is being represented. Establish the unit of this measure (is it a one-to-one relationship or one-to-many?) by referring to a legend. If categorical colours have been deployed, establish the different classifications and associations. Scan the chart looking for the existence of noticeable clusters as well as the widely dispersed (and maybe empty) regions. Some of the most interesting observations come from individual outliers that stand out separately from others. Are there any patterns between the presence of dots and their geographical location? Are there any patterns across the points with similar categorical colour? Gradually zoom in your focus to perform increasingly local assessments between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their patterns.

PRESENTATION TIPS

INTERACTIVITY: One method for dealing with plotting high quantities of observations is to provide interactive semantic zoom features, whereby each time a user zooms in by one level of focus, the unit quantity represented by each dot decreases, from a one-to-many towards a one-to-one relationship.

ANNOTATION: Direct labelling is not necessary, just provide a limited number of regional labels to offer spatial context and orientation. Legends explaining the dot unit scale and any colour associations should ideally be placed as close to the map display as possible.

COLOUR: If colours are being used to distinguish the different categories, ensure these are as visibly different as possible. When including a mapping layer image in the background, ensure it is not overly competing for visual prominence by making it light in colour and possibly semi-transparent. Do not include any unnecessary geographical details that add no value to the spatial orientation or interpretation and clutter the display (e.g. roads, building structures).

COMPOSITION: Dot maps must always be displayed on a map that demonstrates an equal-area

projection as the precision of the plotted locations is paramount. From a readability perspective, try to find a balance between making the size of the dots small enough to preserve their individuality but not too tiny to be indecipherable.

VARIATIONS & ALTERNATIVES

A 'dot density map' is a variation that involves plotting a representative quantity of dots equally (but randomly) across and within a defined spatial region. The position of individual dots is therefore not to be read as indicative of precise locations but used to form a measure of quantitative density. This offers a useful alternative to the choropleth map, especially when categorical separation of the dots through colour is of value. Plotting the location of an incidence of a phenomenon can transcend geographical mapping to any spatial display, such as the seat layout and availability at a theatre or on a flight, or showing the key patterns of play across a sports pitch.

Charts Overlays

Flow map

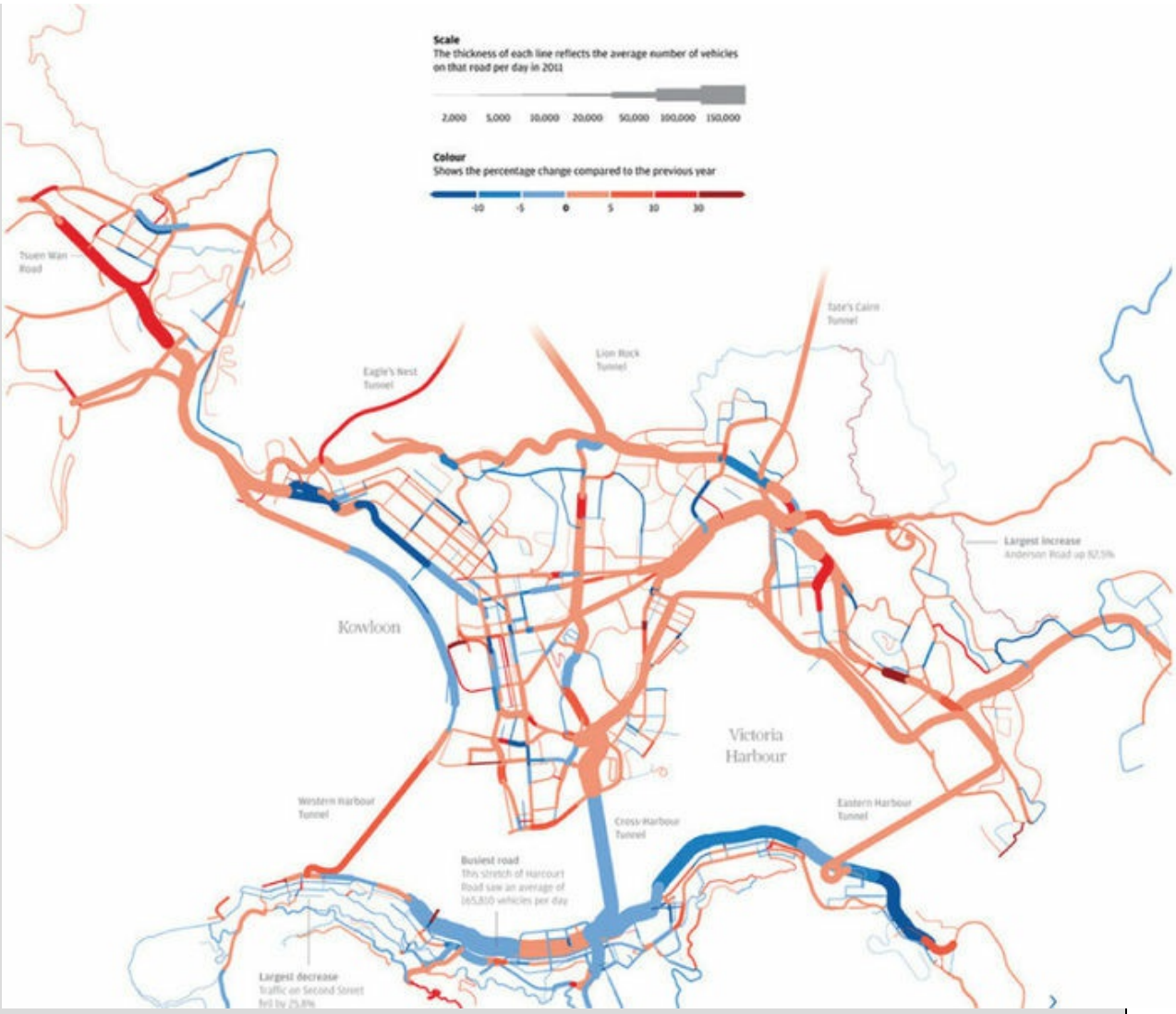
ALSO KNOWN AS Connection map, route map, stream map, particle flow map

REPRESENTATION DESCRIPTION

A flow map shows the characteristics of the movement or flow of a phenomenon across spatial regions. It is often formed using line marks to map flow and combinations of attributes to display the characteristics of this flow. Examples might include the patterns of traffic and travel across or between given routes, the dynamics of the patterns of weather, or the movement patterns of people or animals. There is no fixed template for a flow map but it generally displays characteristics of origin and destination (positions on a map), route (using organic or vector paths), direction (arrow or tapered line width), categorical classification (colour) and some quantitative measure (line weight or motion speed).

EXAMPLE Mapping the average number of vehicles using Hong Kong's main network of roads during 2011.

Figure 6.53 Arteries of the City



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the phenomenon that is being displayed. Establish the association of all visible attributes to understand fully their classification and representation, such as the use of quantitative scales (colour, line size or width) or categorical associations (colour). Scan the chart looking for the existence of patterns of movement, maybe through clustering or common direction, and identify any main hubs and densities within the network. Find the large and the small, the dense and the sparse, and draw out any patterns formed by colour classifications. Gradually zoom in your focus to perform increasingly local assessments between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their patterns.

PRESENTATION TIPS

INTERACTIVITY: Animated sequences will be invaluable to convey motion if the nature of the flow being presented has the relevant physics of movement.

ANNOTATION: Annotation needs will be unique to each approach and the inherent complexity or otherwise of the display. Often the general patterns may offer the sufficient level of readability without the need for imposing amounts of value labels.

COLOUR: The colour relationship needs careful consideration to get the right balance between the intricacies of the foreground data layer and the background mapping layer image. Ensure the background is not overly competing for visual prominence by making it light in colour and possibly semi-transparent. Do not include any unnecessary geographical details that add no value to the spatial orientation or interpretation, but do include those features that have a direct association with the subject matter (such as

roads, routes, etc.).

COMPOSITION: Some degree of geographic distortion of routes or connecting lines may be required practically to display flow data. Choices like interpolation of lines to smooth an activities route or the merging of relatively similar pathways may be entirely legitimate but ensure that this is made clear to the reader.

VARIATIONS & ALTERNATIVES

There are naturally many variations in how you might show flow. It generally differs between whether you are showing point A to point B 'connection maps', more nuanced 'route maps' or surface phenomena such as 'particle flow maps'.

Charts Distortions

Area cartogram

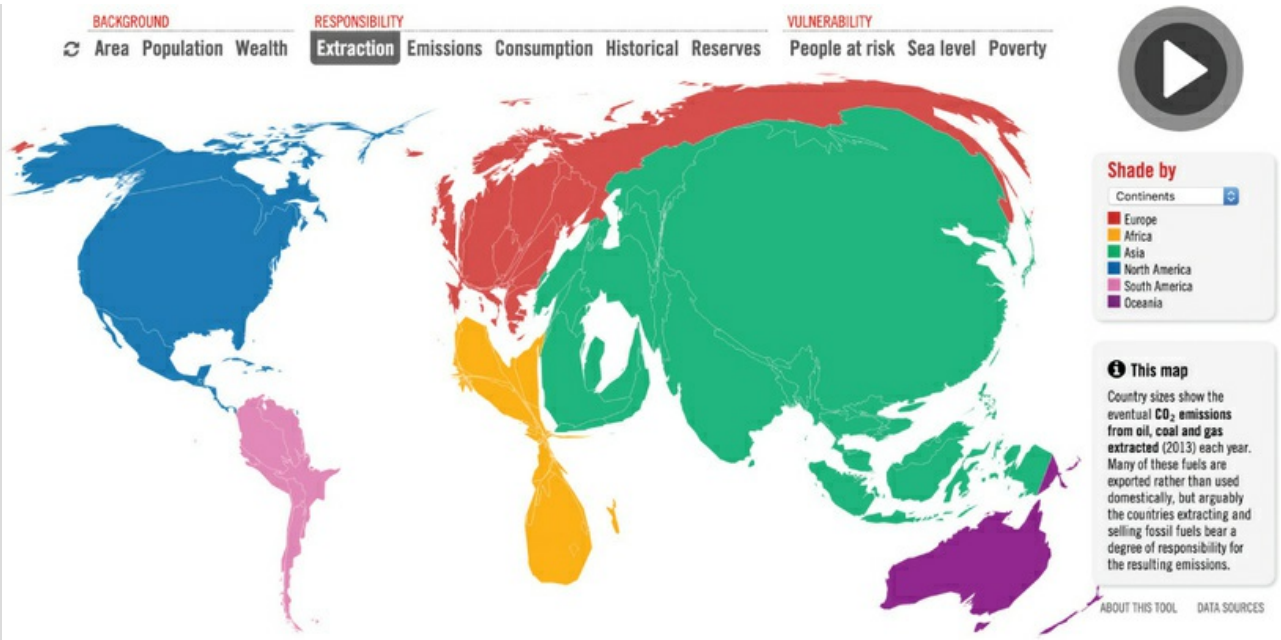
ALSO KNOWN AS Contiguous cartogram, density-equalizing map

EXAMPLE Mapping the measures of climate change responsibility compared to vulnerability across all countries.

REPRESENTATION DESCRIPTION

An area cartogram displays the quantitative values associated with distinct definable spatial regions on a map. Each geographic region is represented by a polygonal area based on its outline shape with the collective regional shapes forming the entire landscape. (Note that most tools for mapping have a predetermined reference between a region name and the dimensions of the regional polygon.) Quantitative values are represented by proportionately distorting (inflating or deflating) the relative size of and, to some degree, shape of the respective regional areas. Traditionally, area cartograms strictly aim to preserve the neighbourhood relationships between different regions. Colour is sometimes used to further represent the same quantitative value or to associate the region with a categorical classification. Area cartograms require the reader to be relatively familiar with the original size and shape of regions in order to be able to establish the degree of relative change in their proportions. Without this it is almost impossible to assess the degree of distortion and indeed to identify the regions themselves.

Figure 6.54 The Carbon Map



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Establish the quantitative value scales or categorical classifications associated with the colour scale, usually found via a legend. Glance across the entire chart to locate the big-, small- and medium-sized shapes according to their apparent distortion. Identify any noticeable exceptions and/or outliers. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate (or read, if labels are present) the absolute values of specific regions of interest.

PRESENTATION TIPS

INTERACTIVITY: Animated sequences enabled through interactive controls can help to better identify instances and degrees of change but usually only over a small set of regions and only if the change is relatively smooth and sustained. Manual animation will help provide more control over the experience.

ANNOTATION: Directly labelling the regional areas with geographical details and the value they hold is likely to lead to too much clutter. You might include only a limited number of regional labels to provide spatial context and orientation.

COLOUR: Legends explaining any colour scales should ideally be placed as close to the map display as possible. The border colour and stroke width for each spatial area should be distinguishable to define the shape but not so prominent as to dominate attention, usually a subtle grey- or white-coloured thin stroke will be fine.

COMPOSITION: To aid the readability of the size of the distortions, it can be useful to present a thumbnail view of the undistorted original geographical layout to help the readers orient themselves with the changes.

VARIATIONS & ALTERNATIVES

Unlike contiguous cartograms, non-contiguous cartograms tend to preserve the shape of the individual polygons but modify the size and the neighbouring connectivity to other adjacent regional polygon areas. The best alternative ways of showing similar data would be to consider using the ‘choropleth map’ or ‘Dorling cartogram’.

Charts Distortions

Dorling cartogram

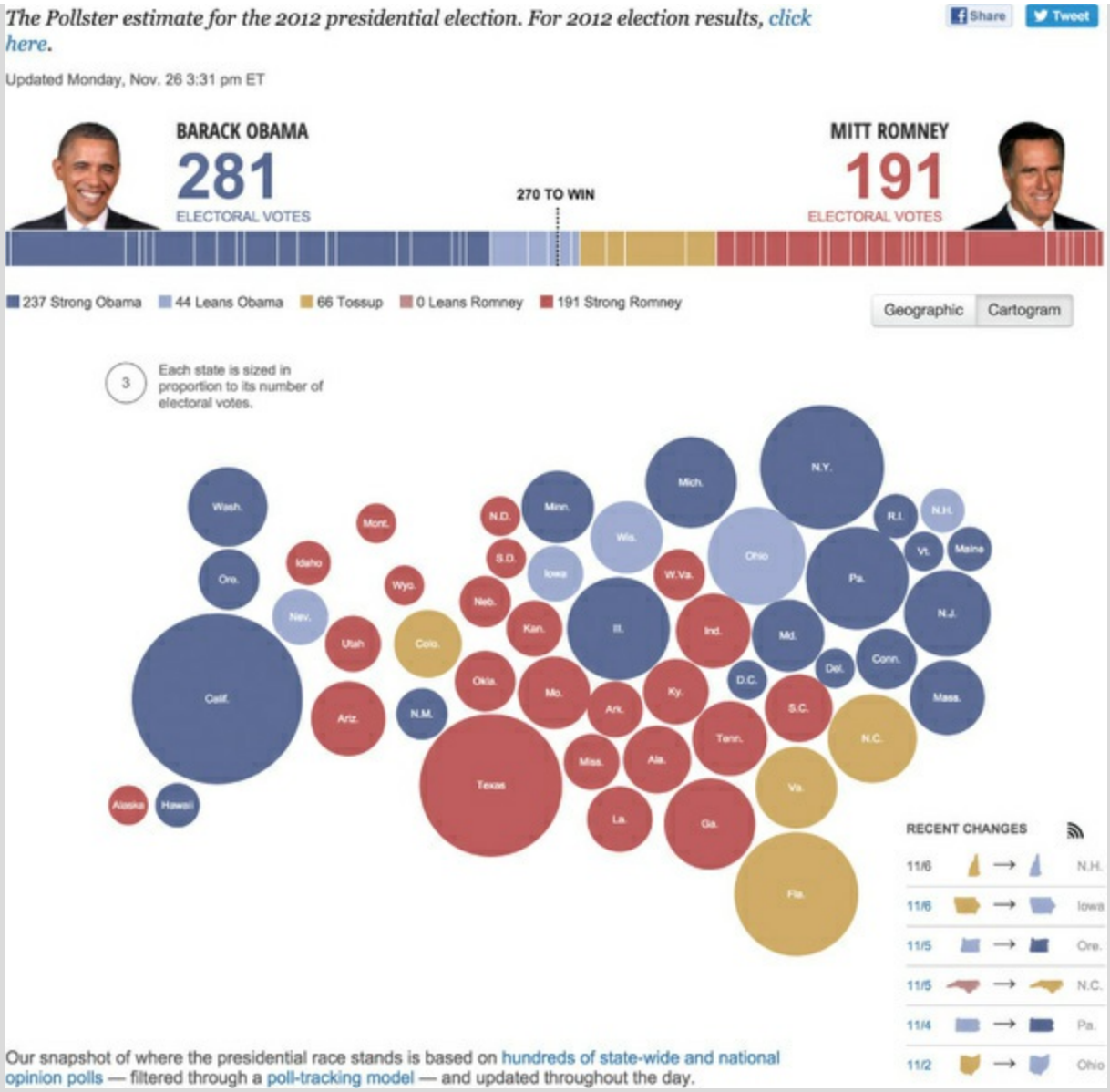
ALSO KNOWN AS Demers cartogram

REPRESENTATION DESCRIPTION

A Dorling cartogram displays the quantitative values associated with distinct, definable spatial regions on a map. Each geographic region is represented by a circle which is proportionally sized to represent a quantitative value. The placement of each circle generally resembles the region's geographic location with general preservation of neighbourhood relationships between adjacent shapes. Colour is used to associate the region with a categorical classification.

EXAMPLE Mapping the predicted electoral voting results for each state in the 2012 Presidential Election.

Figure 6.55 Election Dashboard



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Establish the quantitative value scales or categorical classifications associated with the colour scale, usually found via a legend. Glance across the entire chart to locate the big-, small- and medium-sized shapes. Identify any noticeable exceptions and/or outliers. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate (or read, if labels are present) the absolute values of specific regions of interest.

PRESENTATION TIPS

INTERACTIVITY: Interactive features that enable annotation for category and value labelling can be useful to overcome the difficulties associated with the geographic distortion.

ANNOTATION: Directly labelling the shapes with geographical details and the value they hold is common, though you might restrict this to the circles that have sufficient size to hold such annotation. Otherwise you will need to decide how to handle the labelling of small values.

COLOUR: Legends explaining the size scales and colour associations should ideally be placed as close to the map display as possible. If colours are being used to distinguish the different categories, ensure these are as visibly different as possible.

COMPOSITION: Remember that preserving the adjacency with neighbouring regions is important. Dorling cartograms tend not to allow circles to overlap or occlude, so some accommodation of large values might result in location distortion.

VARIATIONS & ALTERNATIVES

A variation on the approach, called the 'Demers cartogram', involves the use of squares or rectangles instead of circles, which offers an alternative way of connecting adjacent shapes. Other approaches would be through the 'area cartogram' and the 'choropleth map'.

Charts Distortions

Grid map

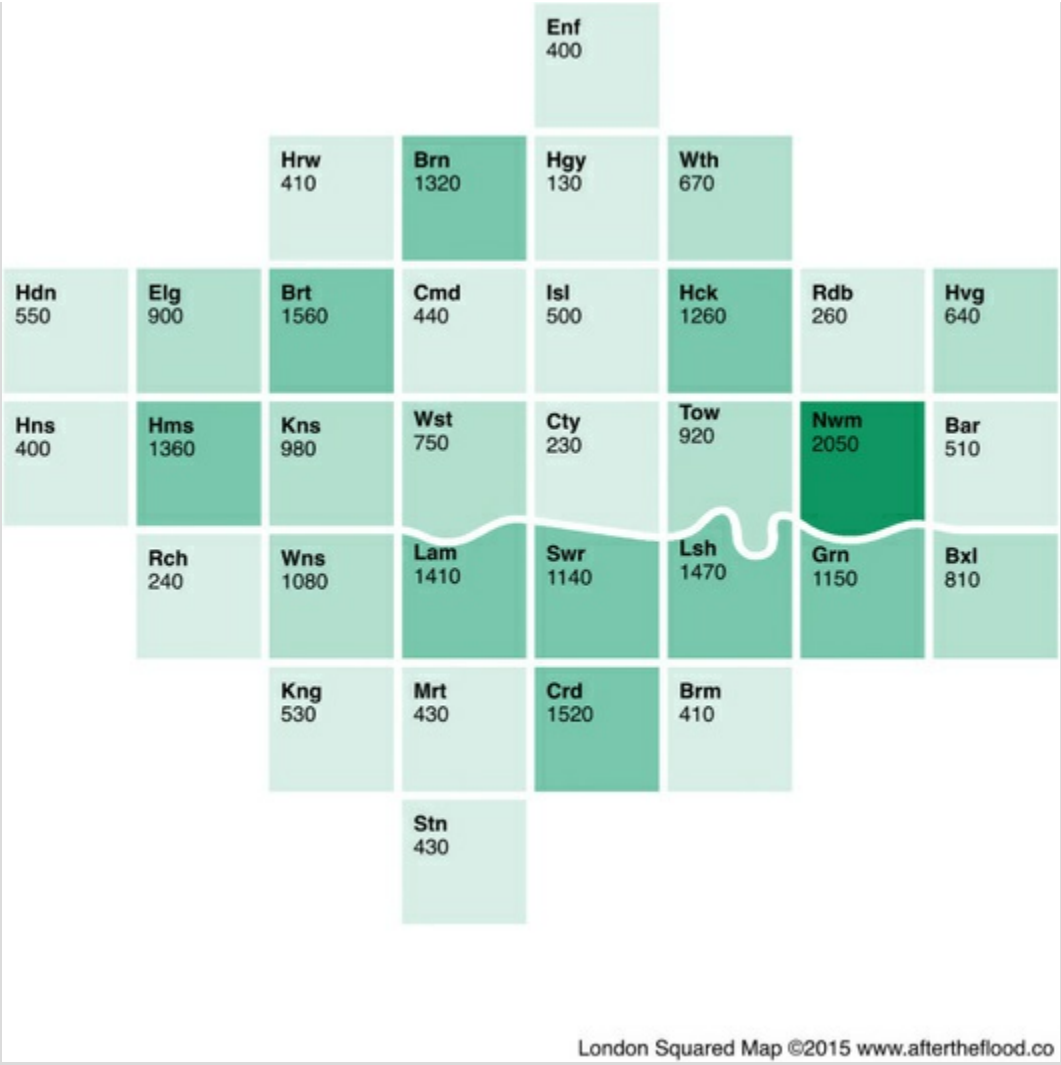
ALSO KNOWN AS Cartogram, bin map, equal-area cartogram, hexagon bin map

REPRESENTATION DESCRIPTION

A grid map displays the quantitative values associated with distinct definable spatial regions on a map. Each geographic region (or a statistically consistent interval of space, known as a 'bin') is represented by a fixed-size uniform shape, sometimes termed a 'tile'. The shapes used tend to be squares or hexagons, though any tessellating shape would work in theory in order to help arrange all the regional tiles into a collective shape that roughly fits the real-world geographical adjacency. Colours are applied to each regional tile either to represent a quantitative value or to associate the region with a categorical classification. Note that the mark used for this chart type is a point rather than an area mark as its size attributes are constant.

EXAMPLE Showing the percentage of household waste recycled in each council region across London between April 2013 to March 2014.

Figure 6.56 London is Rubbish at Recycling and Many Boroughs are Getting Worse



HOW TO READ IT & WHAT TO LOOK FOR

Acquaint yourself with the geographic region you are presented with and carefully consider the quantitative measure that is being represented. Identify the general layout of the constituent tiles to determine how good a fit they are with their adjacent regions in absolute and relative geographical terms. Establish the categorical or quantitative classifications associated with the colour scale, usually found via a legend. Glance across the entire chart to locate the big, small and medium shaded tiles (if quantitative) or the main patterns formed by the categorical colouring. Identify any noticeable exceptions and/or outliers. Gradually zoom in your focus to perform increasingly local comparisons between neighbouring regional areas to identify any noticeable consistencies or inconsistencies between their values. Estimate (or read, if labels are present) the absolute values of specific regions of interest.

PRESENTATION TIPS

INTERACTIVITY: Interactive features that enable annotation for category and value labelling can be useful to overcome the difficulties associated with the geographic distortion.

ANNOTATION: Directly labelling the shapes with geographical details is usually too hard. Some versions of the 'grid map' will include abbreviated labels, maybe two digits, to indicate the region they represent and to aid orientation. Otherwise it may require interactivity to facilitate such annotations. Legends explaining the colour associations should ideally be placed as close to the map display as possible.

COLOUR: If colour is being used to distinguish the different categories, ensure they are as visibly different as possible.

COMPOSITION: The main challenge is to find the most appropriate and representative tile–region

relationship (what is the right amount and geographical level for each constituent tile?) and to optimise the best-fit collective layout that preserves as many of the legitimate neighbouring regions as possible.

VARIATIONS & ALTERNATIVES

'Hexagon bin maps' are specific deployments of the grid map that offer a layout formed by a high resolution of smaller hexagons to preserve localised details. Beyond geographical space, the grid map approach is applicable to any spatial analysis such as in sports.

6.3 Influencing Factors and Considerations

Having covered the fundamentals of visual encoding and profiled many chart type options that deploy different encoding combinations you now need to consider the general factors that will influence your specific choices for which chart or charts to use for your data representation.

Choosing which chart type(s) to use is, inevitably, not a single-factor decision. Rather, as ever with data visualisation, it is an imperfect recipe made up of many ingredients. A pragmatic balance has to be found somewhere between taking on board the range of influencing factors that shape selections and not becoming frozen with indecision caused by the burden of having to consider so many different issues.

Firstly, you need to reflect on the relevant factors that emerge from the first three 'preparatory' stages of the design process and then supplement this by addressing the guidance offered by the three visualisation design principles introduced in Chapter 1. It must be emphasised that there are no direct answers provided for you here, simply guidance. How you might resolve the unique challenges posed by your project has to be something you arrive at yourself.

Formulating Your Brief

Skills and resources, frequency: What charts can you actually make and how efficiently can you create them? This is the big question. Having the ability to create a broad repertoire of different chart types is the vocabulary of this discipline, judging when to use them is the literacy. What will have a great influence on the ambitions of the type of charts you might employ is the 'expressiveness' of your abilities and that of the technology (applications, programs, tools) you have access to. Expressiveness is a term I first heard used in this context by Arvind Satyanarayan, a Computer Science PhD candidate at Stanford University. It describes the amount of variety and extent of control you are provided with by a given technology in the construction of your visualisation solution, so long as you also possess the necessary skills to exploit such features, of course:

- In a data representation context, maximum expressiveness means you can create any combination of mark and attribute encoding to display your data – that is, you can create many different charts. Programming libraries like D3.js and open source tools like R offer broad libraries of different chart options and customisations. The drawing-by-hand nature of Adobe Illustrator would similarly enable you to create a wide range of solutions (though unquestionably more manual in effort and less replicable).
- Restricted expressiveness means you have much more limited scope to adapt different mark and attribute encodings. Indeed you might be faced with assigning data to the fixed encoding options afforded by a modest menu of chart types. A tool like Excel has a relatively limited range of (useful) chart types in its menu. While there are ways of enhancing the options through plugins

and different 'workaround' techniques that broaden its scope, it is a relatively limited tool. It may, however, suffice for most people's visualisation ambitions. Elsewhere, there are many web-based visualisation creation tools which are of value for those who want quick and simple charting, though they certainly reduce the range of options and the capability to customise their appearance.

'The capability to cope with the technological dimension is a key attribute of successful students: coding - more as a logic and a mindset than a technical task - is becoming a very important asset for designers who want to work in Data Visualization. It doesn't necessarily mean that you need to be able to code to find a job, but it helps a lot in the design process. The profile in the (near) future will be a hybrid one, mixing competences, skills and approaches currently separated into disciplinary silos.' Paolo Ciuccarelli, discussing students on his Communication Design Master Programme at Politecnico di Milano

As you reflect on the gallery of charts, my advice would be to perform an assessment of the charts you can make using a scoring system as follows:

3 points	Charts you can personally create relatively easily
2	Charts you can make but involve a greater amount of time and effort, perhaps through your lack of confidence with a certain tool, and possibly involving some innovative workaround solution
1	Charts you can get collaborators or colleagues to create for you, but put you at the mercy of their capacity and availability to do so
0	Charts, currently at least, you might not be able to create at all

For any of the charts that fail to score 3 points, here are some strategies to dealing with this:

- Tools are continually being enhanced. The applications you use now that cannot create, for example, a Sankey diagram, may well offer that in the next release. So wait it out!
- For those charts that currently score 1 or 0 points, look around the web for examples of workaround approaches that will help you achieve them. For example, you might use conditional formatting in an Excel worksheet to create a rudimentary heat map. This is not a chart type offered as standard within the tool but represents an innovative solution through appropriating existing features intended to serve other purposes. Any such solutions, though, have to be framed by the frequency of your work – will this work realistically need to be replicable and repeatable (for example, every month) and does my solution make that achievable?
- Invest time in developing skills in the other tools to broaden your repertoire. Tools like R have a large community of users sharing code, tutorials and examples, resources that would greatly help to facilitate your learning.
- Lower your ambitions. Sometimes the most significant discipline to demonstrate is acknowledging what you cannot do and accepting that (at least, for now) you might need to sacrifice the ideal choices you would make for more pragmatic ones.

Purpose: Should you even seek to represent you data in chart form? Will it add any value, enabling new insights or greater perceptual efficiency compared with its non-visualised form? Will portraying your data via an elegantly presented table, offering the viewer the ability to look up and reference values, actually offer a more suitable solution? Do not rule out the value of a table. Additionally, perhaps you are trying to represent

something in chart form that would actually be better displayed through information-based (rather than data-based) explanations using imagery, textual anecdotes, video and photos? Most of the time the charting of data will be fit for purpose, but just keep reminding yourself that you do not have to chart everything – just make sure you are doing it to add value.

> 'I was in the middle of this huge project, juggling as fast and as focused as I could, and I had this idea of a set of charts stuck in my head that kept resurfacing. And then, as we were heading close to deadline, I realized I couldn't do it. I failed. I couldn't make it work. Because we had pictures of the children, and that was enough … I had to let it go.' Sarah Slobin, Visual Journalist, discussing a project profiling a group of families with children who have a fatal disease

Purpose map: In defining the 'tone' of the project, your were determining what the optimum perceptibility of your data would be for your audience. Your definitions were based on whether you were aiming to facilitate the *reading* of the data or more a general *feeling* of the data? Were you concerned with enabling precise and accurate perceptions of values or is it more about the sense-making of the big, medium and small judgments – getting the 'gist' of values more than reading back the values? Were there emotional qualities that you wanted to emphasise perhaps at the compromise of perceptual efficiency? Maybe there was a balance between the two?

How these tonal definitions apply specifically to data representation requires our appreciation of some fundamental theory about data visualisation. In his book *Semiology Graphique*, published in 1967, Jacques Bertin was the first, most notable author to propose the idea that different ways of encoding data might offer varying degrees of effectiveness in perception. In 1984 William Cleveland and Robert McGill published a seminal paper, 'Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods', that offered more empirical evidence of Bertin's thoughts. They produced a general ranking that explained which attributes used to encode quantitative values would facilitate the highest degree of perceptual accuracy. In 1986, Jock Mackinlay's paper, 'Automating the Design of Graphical Presentations of Relational Information', further extended this to include proposed rankings for encoding categorical nominal and categorical ordinal data types as well as quantitative ones. The table shown in Figure 6.57, adapted from Mackinlay's paper, presents the 'Ranking of Perceptual Tasks'.

In a nutshell, this ancestry of studies reveals that certain attributes used to encode data may make it easier, and others may make it harder, to judge accurately the values being portrayed. Let's illustrate this with a couple of examples. Looking at Figure 6.58, ask yourself: if A is 10, how big is B in the respective bar and circular displays?

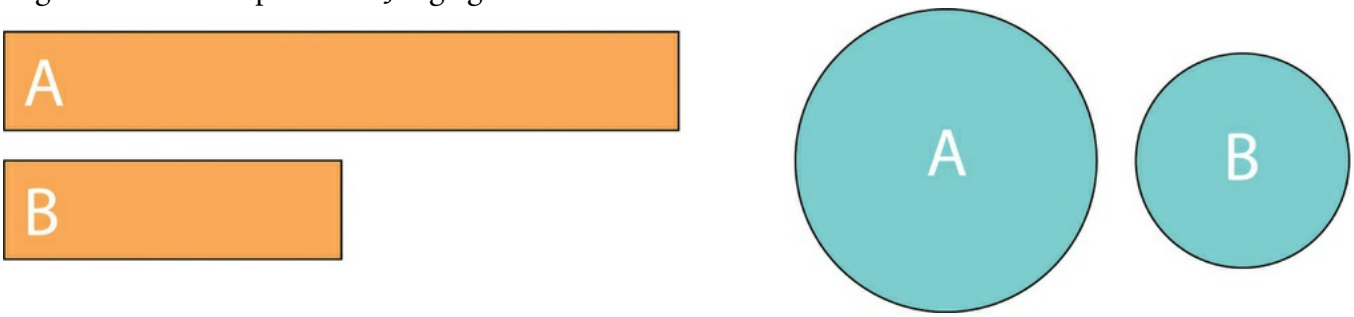
In both cases the answer is B = 5, but while the B 'bar' being 5 feels about right, the idea that the B 'circle' is 5 does not feel quite right. That is because our ability to perform relative judgements for the length of bars is far more precise and accurate than the relative judgements for the area of circles. This is explained by the fact that when judging the variation in size of a line (bar) you are detecting change in a linear dimension, whereas the variation in size of a geometric area (circle) occurs across a quadratic dimension. If you look at the rankings in Figure 6.57 in the 'Quantitative' column, you will see the encoding attribute of *Length* is ranked higher than the attribute of *Area*.

Figure 6.57 The Ranking of Perceptual Tasks

Qualitative **Nominal**	Qualitative **Ordinal**	Quantitative **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

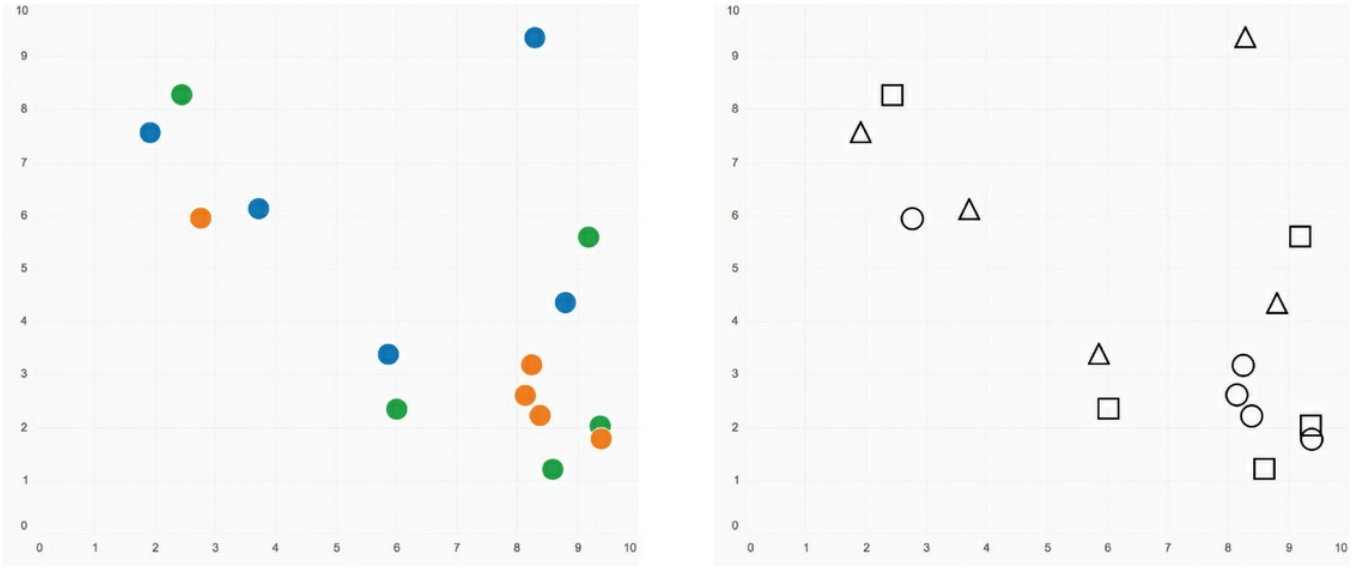
Note that the attribute of 'Motion' was not included in this study. For the purposes of this display, 'Angle' and 'Slope' are combined whereas they were distinguished as separate in the study.

Figure 6.58 Comparison of Judging Line Size vs Area Size



Now let's consider an example (Figure 6.59) that shows the relative accuracy of using different dimensions of colour variation to represent categorical nominal values. In the next pair of charts you can see different attributes being used to represent the categorical groupings of the points in the respective scatter plots. On the left you can see variation in the attribute of colour hue (blue, orange and green) to separate the categories visually; on the right you will see the attribute of shape (diamond, circle and square) applied to the same category groupings. What you should be experiencing is a far more immediate, effortless and accurate sense of the groupings of the coloured category markers compared with the shaped category markers. It is simply easier to spot the associations through variation in colour than variation in shape. This explains why *colour hue* is much higher in the proposed rankings for nominal data than *shape*.

Figure 6.59 Comparison of judging related items using variation in colour (hue) vs variation in shape



So you can see from these simple demonstrations that there are clearly ways of encoding data that will make it easier to read values accurately and efficiently. However, as Cleveland and McGill stress in their paper, this ranking should be taken as only one ingredient of guidance: 'The ordering does not result in a precise prescription for displaying data but rather is a framework within which to work'.

This is important to note because you have to take into account other factors. You have to decide whether precise perceiving is actually what you need to facilitate for your readers. If you do, then the likes of the bar chart – through the variation in length of a bar – will evidently offer a very precise approach. As stated in Chapter 3, that is why they are such an important part of your visual artillery.

However, sometimes getting a 'gist' of the data is sufficient. A few pages ago I presented an image of a bubble chart on my website's home page, showing the popularity of my blog posts over the previous 100-day period. The purpose of this display was purely to give visitors a sense of the general order of magnitude from the most popular to the relative least popular posts. I do not need visitors to form a precise understanding of absolute values or exact rankings. I just want them to get a sense of the ranking hierarchy. I can therefore justify moving down the quantitative attribute rankings proposed and deploy a series of circles that encode the visitor totals through the size of their area (colour is used to represent different article categories). The level of perceptibility (accuracy and efficiency) that I need to facilitate is adequately achieved by the resulting 'frogspawn'-like display. Furthermore, it offers an appealing and varied display that suits the purpose of this front-page navigation device.

In practice, what all this shows is that chart types vary in the relative efficiency and accuracy of perception offered to a viewer. Moreover, many of the charts shown in the gallery can therefore only ever facilitate a gist of the values of data due to the complexity of their mark and attribute combinations and the amount of data values they might typically contain (e.g. the treemap often has many parts of a whole in a single display). It is up to you to judge what the right threshold is for your purpose.

Working With Data

> Data examination: Inevitably, the physical characteristics of your data are especially influential. *What* types of data you are trying to display will have a significant impact on *how* you are able to show them. Only certain types of data can fit into certain chart types; only certain chart types can accommodate certain types of data. That is why it is often most useful practically to think of this task in

terms of chart types and particularly in terms of these as templates, able to accommodate specific types of data.

For example, representing data through a bar chart requires one categorical variable (e.g. department) and one quantitative variable (e.g. maximum age). If you want to show a further categorical variable (let's say, to break down departments by gender) you are going to need to switch 'template' and use something like a clustered bar chart which can accommodate this extra dimension.

I explained earlier how the shape of data influenced the viability of the flower metaphor used in the 'Better Life Index'. The range of categorical and quantitative values will certainly influence the most appropriate chart type choice. For example, suppose you want to show some part-to-whole analysis and you have only three parts (three sub-categories belonging to the major category or whole) then a treemap really does not make a great deal of sense – they are better at representing many parts to a whole. The unloved pie chart would probably suffice if the percentage values were quite diverse otherwise the bar chart would be best.

Beyond the size and shape of your data you also might be influenced by its inherent meaning. Sometimes, you will have scope in your encoding choices to incorporate a certain amount of visual immediacy in accordance with your topic. The flowers of the Better Life Index feel consistent in metaphor with the idea of better life: the more in bloom the flowers, the more colourful and proud each petal appears and the better the quality of life in that country. There is a congruence between subject matter and visual form. Think about the billionaires' project from earlier in the chapter, with rankings displayed by industry. Each point marking each billionaire was a small caricature face. This is not necessary – a small circular mark for each person would have been fine – but by using a face for the mark it creates a more immediate recognition that the subject matter is about people.

Data exploration: One consistently useful pointer towards how you might visually communicate your data to others is to consider which techniques helped *you* to unearth key insights when you were visually exploring the data. What chart types have you already tried out and maybe found to reveal interesting patterns? Exploratory data analysis is, in many ways, a bridge to visual communication: the charts you use to inform yourself often represent prototype thinking on how you might communicate with others. The design execution may end up being different once you introduce the influence of audience characteristics into your thinking, naturally, but if a method is already working, why not utilise the same approach again?

'Effective graphics conform to the Congruence Principle according to which the content and format of the graphic should correspond to the content and format of the concepts to be conveyed.' Barbara Tversky and Julie Bauer Morrison, taken from *Animation: Can it Facilitate?*

Establishing Your Editorial Thinking

Angle: When articulating the angles of analysis you intend to portray to your viewers, you are effectively dictating which chart types might be most relevant. If you intend to show how quantities have changed over time, for example, there will be certain charts best placed to portray that and many others that will not. By expressing your desired editorial angles of analysis in language terms, this will be extremely helpful in identifying the primary families of charts across the CHRTS taxonomy that will provide the best option.

It is vital to treat every representation challenge on its own merits – do not fall into the trap of going through the motions. Just because you have spatial data does not mean that the most useful portrayal of

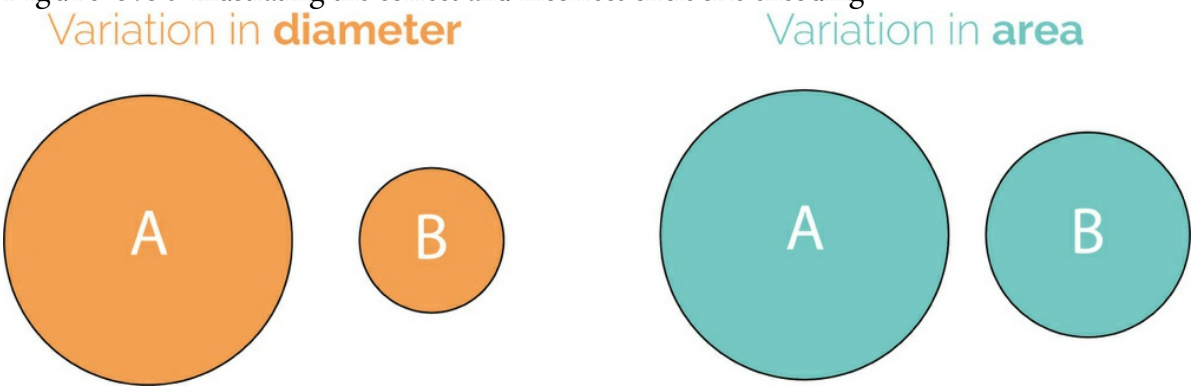
that data will be via a map. If the interesting insights are not regionally and spatially significant, then the map may not provide the most relevant window on that data. The composition of a map – the shape, size and positioning of the world's regions – is so diverse, inconsistent and truly non-uniform that it may hinder your analysis rather than illuminate it. So always make sure you have carefully considered the relevance of your chosen angle through your editorial thinking.

Trustworthy Design

Avoiding deception: In the discussion about tone I explained how variations in the potential precision of perception may be appropriate for the purpose and context of your work. Precision in perception is one thing, but precision in design is another. Being truthful and avoiding deception in how you portray data visually are fundamental obligations.
There are many ways in which viewers can be deceived through incorrect and inappropriate encoding choices. The main issues around deception tend to concern encoding the size of quantities. For beginners, these mistakes can be entirely innocent and unintended but need to be eradicated immediately.

- *Geometric calculations* – When using the area of shapes to represent different quantitative values, the underlying geometry needs to be calculated accurately. One of the common mistakes when using circles, for example, is simply to modify the diameters: if a quantitative value increases from 10 to 20, just double the diameter, right? Wrong. That geometric approach would be a mistake because, as viewers, when perceiving the size of a circle, it is the area, not the width, of the circle upon which we base our estimates of the quantitative value being represented.

Figure 6.60 Illustrating the correct and incorrect circle size encoding



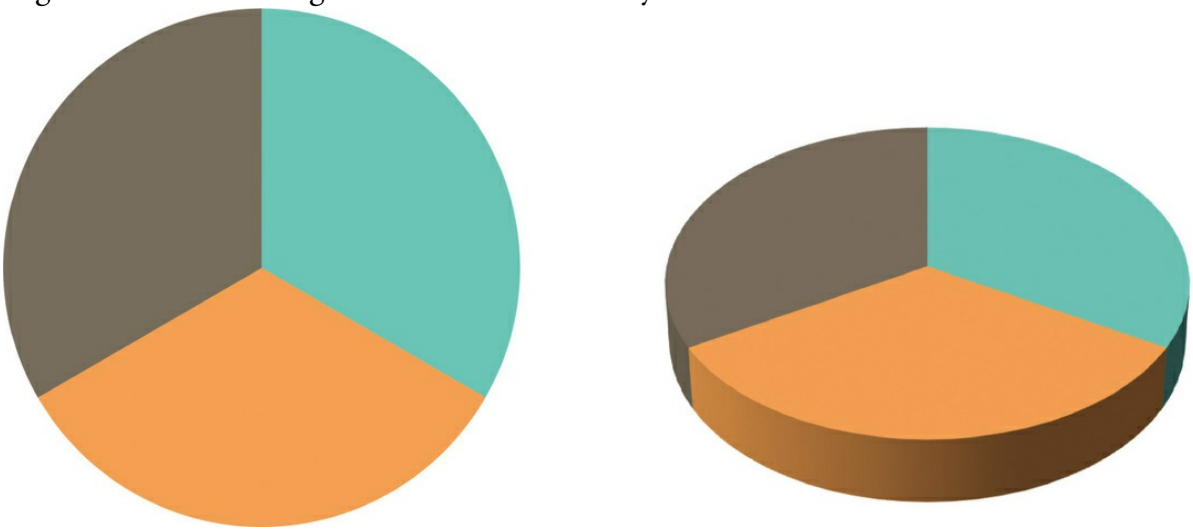
The illustration in Figure 6.60 shows the incorrect and correct ways of encoding two quantitative values through circle size, where the value of A is twice the size of B. The orange circle for B has half the *diameter* of A, the green circle for B has half the *area* of A. The green circle area calculations are the correct way to encode these two values, whereas the orange circle calculations disproportionately shrink circle B by halving the diameter rather than halving the area. This makes it appear much smaller than its true value.

- *3D decoration* – In the vast majority of circumstances the use of 3D charts is at best unnecessary and at worst hugely distorting in the display of data. I have some empathy for those who might volunteer that they have made and/or like the look of 3D charts. In the past I did too. Sometimes we don't know not to do something until we are told. So this is me, here and now, telling you.

The presence of 3D in visualisation tends to be motivated by a desire to demonstrate technical competence with the features of a tool in terms of 'look how many things I know how to do with this tool!' (users of Excel, I am pointing an accusatory finger at you right now). It is also driven by the appetite of rather unsophisticated viewers who are still attracted by the apparent novelty of 3D skeuomorphic form. (Middle and senior management of the corporate world, with your 'make me a fancy chart' commands, my finger of doom is now pointing in your direction.)

Using psuedo-3D effects in your charts when you have only two dimensions of data means you are simply decorating data. And when I say 'decorating', I mean this with the same sneer that would greet memories of avocado green bathrooms in 1970s Britain. A 3D visualisation of 2D data is gratuitous and distorts the viewer's ability to read values within any degree of acceptable accuracy. As illustrated in Figure 6.61, in perceiving the value estimates of the angles and segments in the respective pie charts, the 3D version makes it much harder to form accurate judgements. The tilting of the isometric plane amplifies the front part of the chart and diminishes the back. It also introduces a raised 'step' which is purely decorative, thus embellishing the judgement of the segment sizes.

Figure 6.61 Illustrating the Distortions Created by 3D Decoration



- Furthermore, for charts based on three dimensions of data, 3D effects should only be considered if – and only if – the viewer is provided with means to move around the chart object to establish different 2D viewing angles *and* the collective representation of all the 3D of data makes sense in showing a whole 'system'.
- *Truncated axis scales* – When quantitative values are encoded through the height or length components of size (e.g. for bar charts and area charts), truncating the value axis (not starting the range of quantitative values from the true origin of zero) distorts the size judgements. I will look at this in more detail in the chapter on composition because it is ultimately more about the size considerations of scales and deployment of chart apparatus than necessarily just the representation choices.

Accessible Design

The bullet chart is a derivative of the bar chart – the older, more sophisticated brother of the idiot gauge chart – but I didn't think it was necessary to profile as a separate chart type.

Encoded overlays: Beyond the immediate combinations of marks and attributes that comprise a given chart type, you may find value in incorporating additional detail to help viewers with the perceiving and interpretation task. *Encoded overlays* are useful to help explain further the context of values and amplify the interpretation of the good and the bad, the normal and the exceptional. In some ways these features might be considered forms of annotation, but as they represent data values (and therefore require encoding choices) it makes sense to locate these options within this chapter. There are many different types of visual overlays that may be useful to include:

Figure 6.62 Example of a Bullet Chart Using Banding Overlays

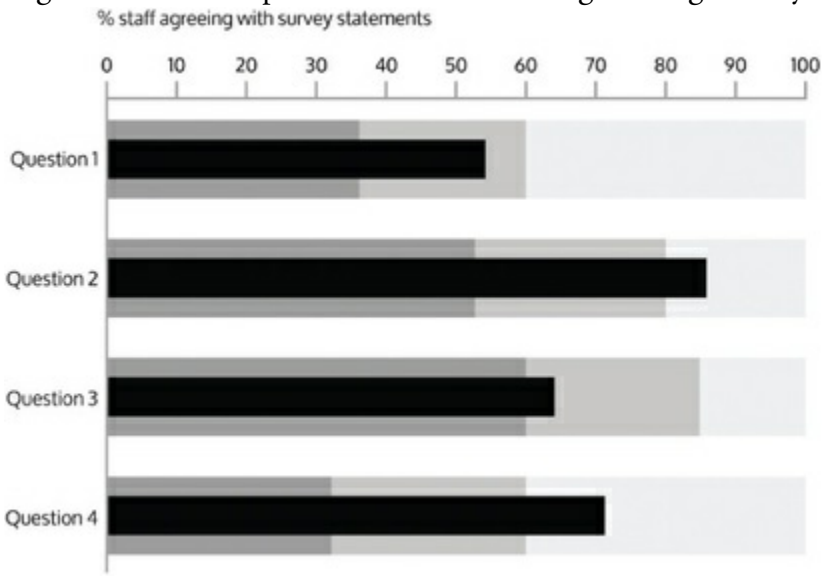
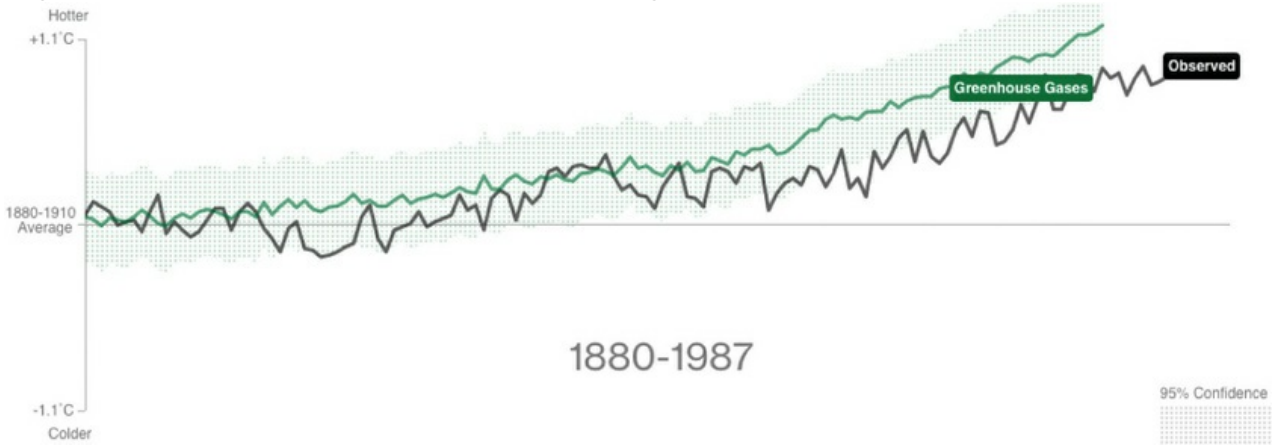


Figure 6.63 Excerpt from 'What's Really Warming the World?'



- *Bandings* – These are typically shaded areas that provide some sense of contrast between the main data value marks and contextual judgements of historic or expected values. In a bullet chart (Figure 6.62) there are various shaded bands that might help to indicate whether the bar's value should be considered bad, average or good. In the line chart (Figure 6.63) here you can see the observed rise in global temperatures. To facilitate comparison with potentially influencing factors, in the background there is a contextual overlay showing the change in greenhouse gases with banding to indicate the 95% confidence interval.
- *Markers* – Adding points to a display might be useful to show comparison against a target, forecast, a previous value, or to highlight actual vs budget. Figure 6.64 shows a chart that facilitates comparisons against a maximum value marker.

Figure 6.64 Example of Using Markers Overlays

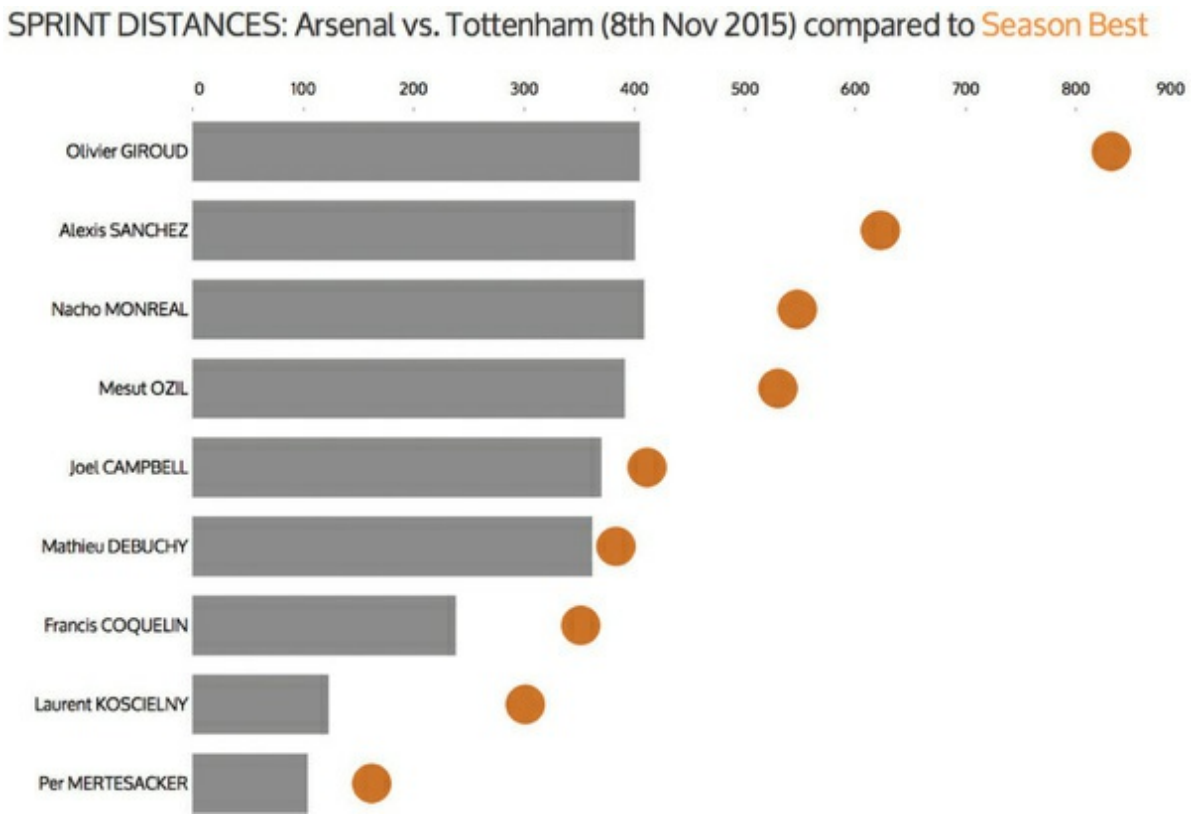
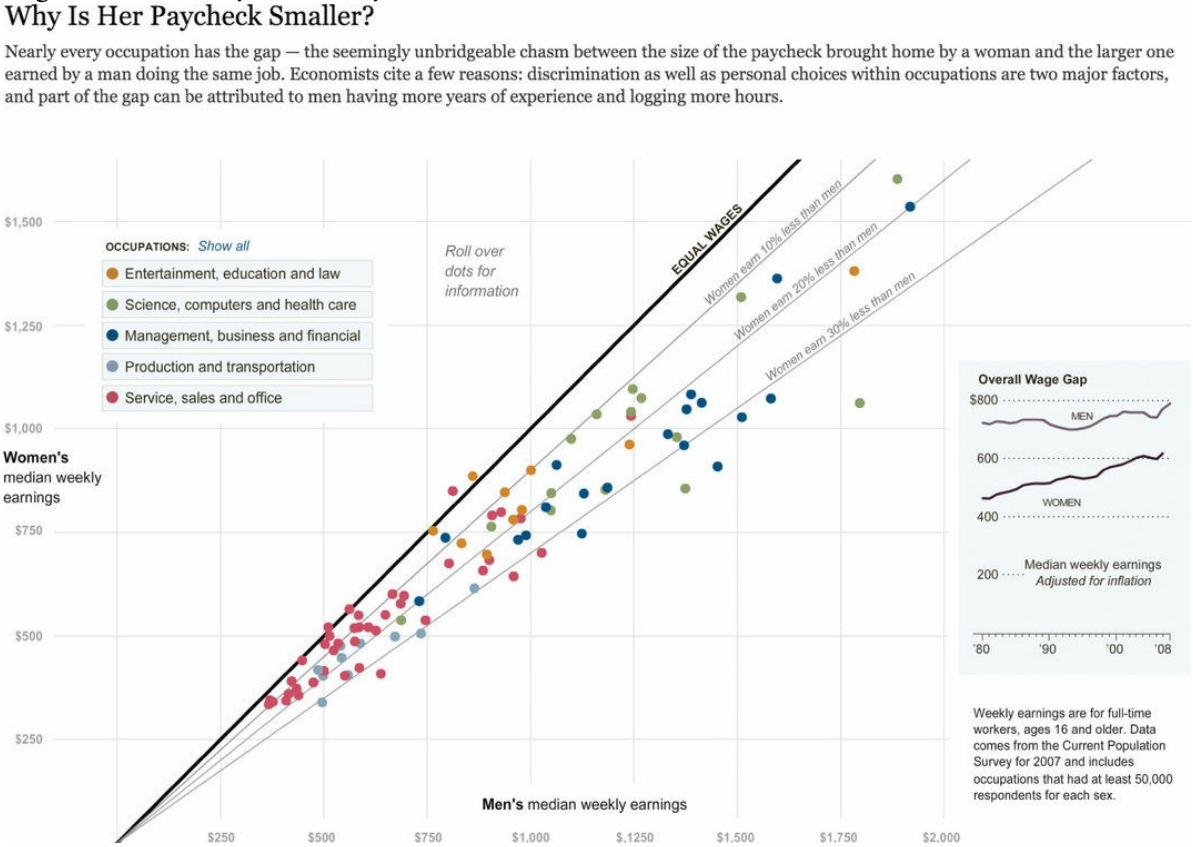


Figure 6.65 Why Is Her Paycheck Smaller?



- *Reference lines* – These are useful in any display that uses position or size along an axis as an attribute for a quantitative value. Line charts or scatter plots (Figure 6.65) are particularly enhanced by the inclusion of reference lines, helping to direct the eye towards calculated trends, constants or averages and, with scatter plots specifically, the lines of best fit or correlation.

Elegant Design

Visual appeal: This fits again with the thinking about 'tone' and may also be informed by some of the mental visualisations that might have formed in the initial stages of the process. Although you should not allow yourself to be consumed by ideas over the influence of the data, sometimes there is scope to squeeze out an extra sense of stylistic association between the visual and the content. For example, the 'pizza' pie chart in Figure 6.66 presents analysis about the political contributions made by companies in the pizza industry. The decision to use pizza slices as the basis of a pie chart makes a lot of sense. The graphic in Figure 6.67 displays the growth in online sales of razors. Like the pizzas, the notion of creating bar charts by scraping away lengths of shaving foam offers a clever, congruent and charming solution.

Figure 6.66 Inside the Powerful Lobby Fighting for Your Right to Eat Pizza

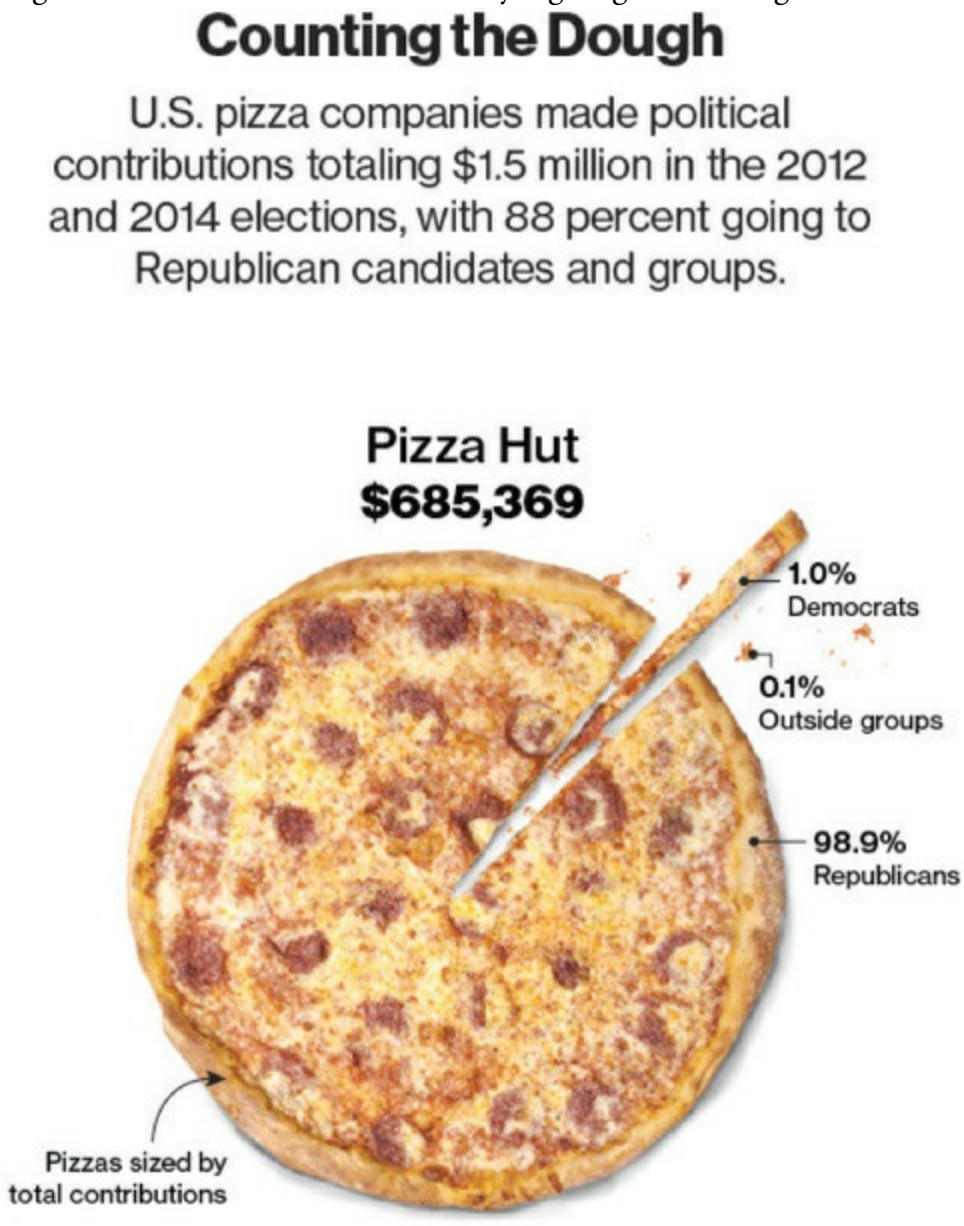


Figure 6.67 Excerpt from 'Razor Sales Move Online, Away From Gillette'

Summary: Data Representation

Visual Encoding All charts are based on combinations of marks and attributes:

- Marks: represent records (or aggregation of records) and can be points, lines, areas or forms.
- Attributes: represent variable values held for each record and can include visual properties like position, size, colour, connection.

Chart Types If visual encoding is the fundamental theoretical understanding of data representation, chart types are the practical application. There are five families of chart types (CHRTS mnemonic):

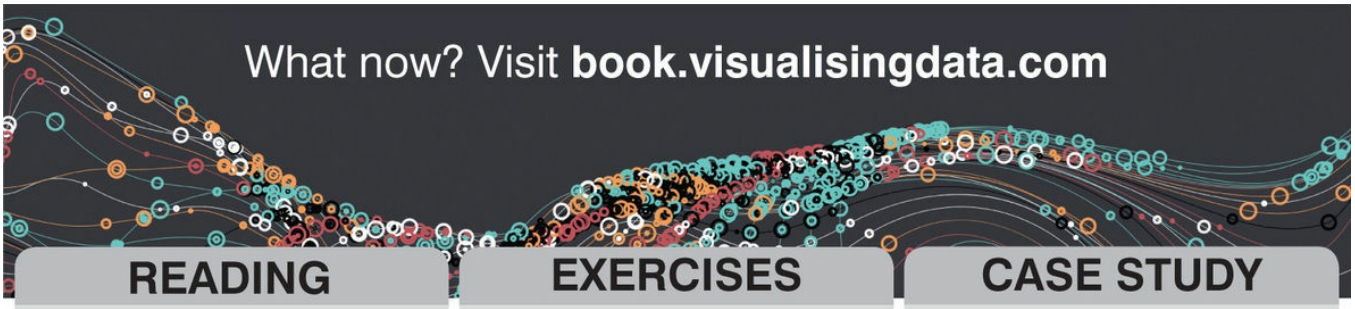
CATEGORICAL	Comparing categories and distributions of quantitative values
HIERARCHICAL	Charting part-to-whole relationships and hierarchies
RELATIONAL	Graphing relationships to explore correlations and connections
TEMPORAL	Showing trends and activities over time
SPATIAL	Mapping spatial patterns through overlays and distortions

Influencing Factors and Considerations

- Formulating the brief: skills and resources – what charts can you make and how efficiently? From the definitions across the 'purpose map' what 'tone' did you determine this project might demonstrate?
- Working with data: what is the shape of the data and how might that impact on your chart design? Have you already used a chart type to explore your data that might prove to be the best way to communicate it to others?
- Establishing your editorial thinking: what is the specific angle of the enquiry that you want to portray visually? Is it relevant and representative of the most interesting analysis of your data?
- Trustworthy design: avoid deception through mistaken geometric calculations, 3D decoration, truncated axis scales, corrupt charts.
- Accessible design: the use of encoded overlays, such as bandings, markers, reference lines, can aid readability and interpretation.
- Elegant design: consider the scope of certain design flourishes that might enhance the visual appeal through the form of your charts whilst also preserving their function.

Tips and Tactics

- Data is your raw material, not your ideas, so do not arrive at this stage desperate and precious about wanting to use a certain data representation approach.
- Be led by the preparatory work (stages 1 to 3) but do use the chart type gallery for inspiration if you need to unblock!
- Be especially careful in how you think about representing instances of zero, null (no available data) and nothing (no observation).
- Do not be too proud to acknowledge when you have made a bad call or gone down a dead end.



READING

Visit the chapter's library of further reading and references to continue your learning about data representation

EXERCISES

Undertake these practical exercises to help refine your skill and understanding about making effective data representation choices

CASE STUDY

Work through the next instalment of the Filmographics case-study narrative, discussing the data representation choices that were made