An Introduction to Data Visualisation: A Handbook of Notes and Ideas

Dr Jon Minton

Applied Quantitative Methods Network (AQMeN)

University of Glasgow

***[UNFINISHED DRAFT AS OF 6/7/2014]***

|  |  |
| --- | --- |
| http://the-sra.org.uk/img/sralogo120x83.jpg | https://www.aqmen.ac.uk/sites/default/files/aqmen_logo.jpg |

# Introduction

This document will present some ideas and further resources to help you on your journey to become more adept and efficient data visualisers. As you may gather, I’m not particularly linear-minded, and as a result the information and ideas are not presented in a clear and simple sequence throughout. However, I hope that, even though the notes may not always be in the right order, they are still the right notes, and will help you make further progress in applying and understanding data visualisation, and know where to turn for further advice and help.

Broadly, the structure of the document is as follows:

* In part one, I introduce some fundamental questions and ideas about what data visualisation is, and what it should aim to achieve, by discussing the distinction between data visualisation and information visualisation.
* In part two, I introduce the **Grammar of Graphics**, and the **Three Hats**. Two complementary sets of ideas for helping to understand what data visualisation is about in theory and in practice.
* In part three, I move onto talking more about some of the practical issues involved in data visualisation, with a focus on data management.

# Part One: Data Visualisation is not Information Visualisation

Misinformation can be beautiful. Tim Harford (pictured), author of The Undercover Economist book series and presenter of Radio 4’s statistics More or Less programme, says 200 years of statistical science should not be cast aside for the blandishments of big data.

Harford, a senior columnist at the Financial Times, cautioned delegates at Teradata’s recent Universe conference in Prague against falling too easily for the “dazzle” of data visualisation. While infographics enthusiasts, such as David McCandless, posit that information is beautiful, Harford encourages his readers and listeners to dig beneath the surface.

His Prague talk took its inspiration from the zebra-like dazzle camouflage applied to allied ships in the First World War to evade torpedo attack from German submarines. Data visualisation often functions like that, he contends.

**Source**: McKenna, B (12 April 2014) ‘Undercover economist Tim Harford decries data visualisation dazzle’, Available from: <http://www.bigdata-madesimple.com/undercover-economist-tim-harford-decries-data-visualisation-dazzle/>

I will start this document on data visualisation by saying that I largely agree with Tim Harford’s attack on data visualisation. Except for one point: Tim Harford used the wrong term, and was not describing or attacking data visualisation. Instead, he seemed to be attacking infographics, which are a form of *information* visualisation. Data visualisation is a specialist type of information visualisation, but most forms of information visualisation are not data visualisations. However, the aim of all data visualisations should be to inform.

A primary aim of this workshop is to help delegates understand this distinction between information visualisation and data visualisation better. Given that even an Oxford-educated economist can fail to understand the distinction, it sounds like there is a clear need to understand these differences.

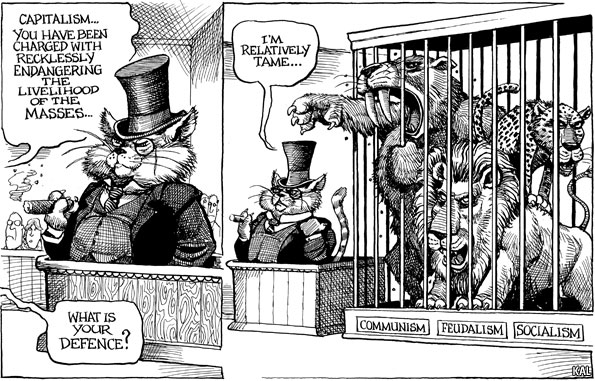
So, to reiterate, it is important to make clear at the outset that data visualisation and information visualisation are not the same. Put simply, information is not data, and data is not information. However, although data visualisation and information visualisation are not the same, they are closely related. Good data visualisation makes data more informative, just as bad data visualisation can be misinformative.

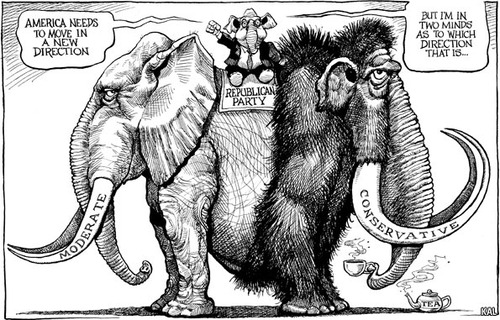
## Thought and imagination

It is because people can visualise information effectively that we can visualise data effectively. When we use data visualisations to turn numbers into images, we are building on solid cognitive foundations. We can usually understand more, process more and compare more if information is presented visually rather than in other forms.

There is a growing body of evidence that we think in images, and that we understand abstract and invisible notions and concepts by applying cognitive hooks to them, latching them onto objects that are concrete and visible. As an example of this, consider the political cartoons featured in the economist, in which entire countries and systems of government become substituted for animals – the bear, the panda – and cartoon characters – the Uncle Sam. Some examples are shown below:

[](https://www.google.co.uk/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&cad=rja&docid=l0F0r9CPt9SFtM&tbnid=Ff3wjWJkv6w5uM:&ved=0CAQQjB0&url=http://anticap.wordpress.com/2011/08/11/stocks-and-flows/&ei=RTaKUuqEGcSS1AWoxIHoDA&psig=AFQjCNG-sE90_pK)

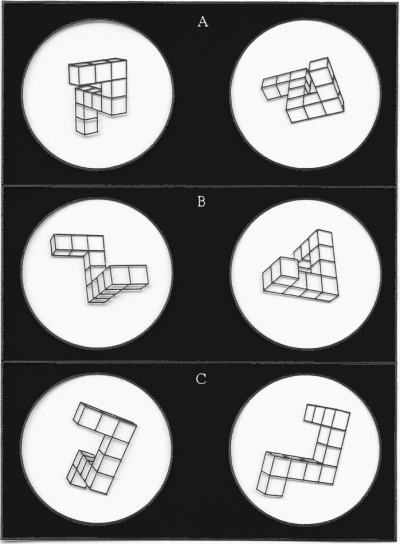
[](http://catallaxyfiles.com/2012/07/23/whats-wrong-with-the-ecommunist/)



[](http://balneus.wordpress.com/2011/04/23/old-cartoon-more-relevant-than-ever/)

Read the text surrounding the cartoons, and the need for visual imagery is still apparent. Economies ‘grow’ and ‘shrink’, speed up and slow down, move closer together and further apart. Without visual metaphors our brains we would be without any means to understand (or misunderstand) such things.

There is some good evidence from experiments in cognitive science that we often aim to understand things by constructing mental visual representations of them, and examining and visualizing these imaginary representations. For example, the cognitive scientist Stephen Kosslyn presented a large number of test subjects with pairs of images, such as those shown below:

[](http://plato.stanford.edu/entries/mental-imagery/mental-rotation.html)

He asked the subjects whether, for each of the pairs of images, whether the image on the left was of the same object as the image on the right, but shown from a different angle. Importantly, he found that the time taken to answer the question was proportional to the amount by which the structure represented by the image on the left had to be rotated to match up with the image on the right. This showed that people produce answers to these questions by producing three dimensional mental representations of what was presented, and turning, spinning, and no doubt growing, shrinking and stretching, such mental structures as an integral part of the process of understanding the world.

As we have seen with the economist cartoons, the implications of this process of mental visualisation being so integral to our capacity to make sense of the world are very broad. For example, the US political scientist George Lakoff suggested, in his book *Don’t Think of an Elephant,* that part of the reason for the continuing political strength of Republicans was that they were better able to evoke strong mental images in the electorate than Democrats, and that as a result Democrats needed to produce more effective and visual ways to communicate their messages. Robert McKee, author of *Story*, the ‘bible’ of screenwriting for film, often presents his arguments about how stories should be structured and sequenced in visual terms. For example, his visualisation of a narrative structure which he calls ‘The Quest’ is shown below:

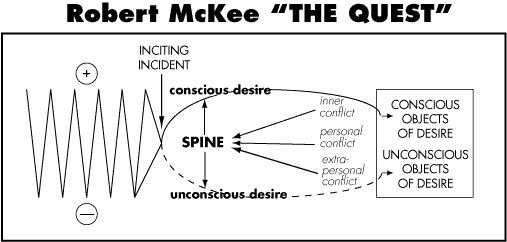
[](http://mckeestory.com/)

Figure 1 The Quest: Robert McKee (http://mckeestory.com/)

Two additional visualisations which spring to mind include a visual representation of the rules of (verbal) grammar, and a visual description of the process and context of gaining a PhD. Parts of these visualisations are shown below, with links to the websites shown in the captions:

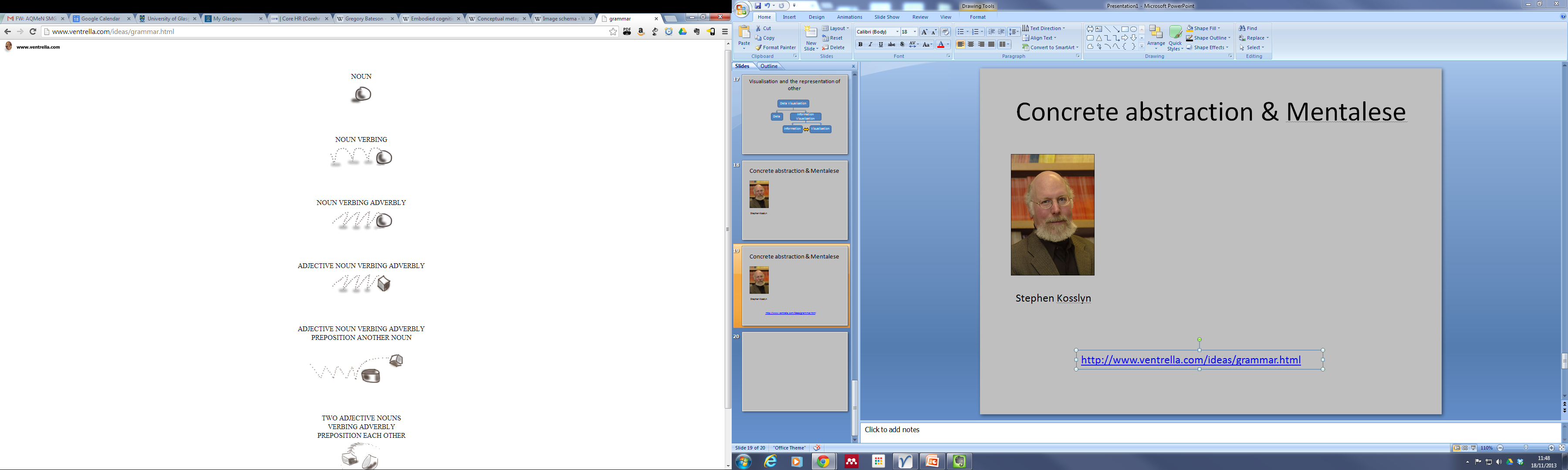
[](http://www.ventrella.com/ideas/grammar.html)

Figure 2 http://www.ventrella.com/ideas/grammar.html

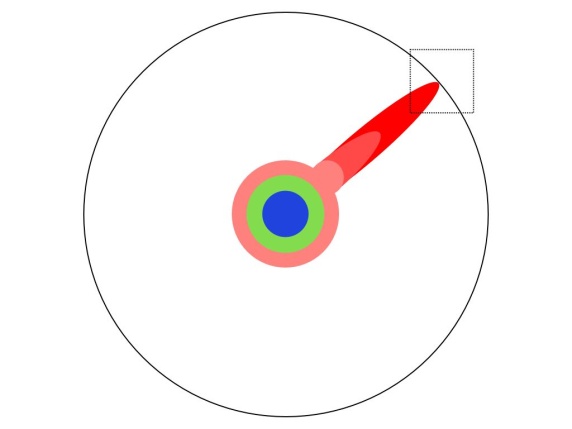
[](http://matt.might.net/articles/phd-school-in-pictures/)

Figure 3 The illustrated guide to a PhD (http://matt.might.net/articles/phd-school-in-pictures/)

## Data and Information: Making Data Informative

The following section will try to make the distinction and relationship between data and information clearer, and discuss the linked roles that statistics and data visualisation have in making data more informative.

### There cannot be ‘too much information’, but there can be too much data.

Imagine you asked someone if there had been any problems over the last month with a particular project at work. In response, this person sent you a sixty page document listing exactly how many minutes he spent working on the project, for each of the last thirty days, exactly which files were accessed on each day, and what text was changed within each file on each day.

The answer to the question would likely be somewhere in this report, but the report is not itself informative. The report is data, not information. Something has to be done to the data to make it informative.

Imagine if instead the response was “It’s all going according to plan”. This is more informative (although it might be misinformative), even though it contains much less data than the sixty page report.

Information is not data. Instead, information is what people want to get from data. Information is what allows people to make informed decisions about the world and the piece of it they can hope to influence and shape. Information is that which allows people to better understand and engage with bigger and broader things. We cannot be ‘too informed’, although it is easy to drown in facts and data.

Data visualisation, along with its cousin statistical inference, helps to make data more informative. Both statistical inference and data visualisation work on different sides of the equation, with statistical inference reducing the amount of data we have to fit inside our heads, and data visualisation increasing our capacity to handle larger amounts of data. Both of these tools, and their complementary roles, will be discussed below:

### Statistics is data reduction…

Returning to the example of the sixty page report earlier, imagine if, instead of being presented the entire report, we are presented with a half page summary. This summary shows, for each of the files involved in the project, the average amount of time per day spent on writing to the file, the number of lines added to that file, and a ratio of the two: average lines per day. Tens or hundreds of thousands of pieces of data have been reduced to just a handful of numbers. The tool that has managed to convert a lot of numbers into a handful is statistics. In this case, the calculation of arithmetic means. The few summary statistics are easier to think about, place, compare and make decisions about than the full data. Because of this, they are more informative.

Without access to statistical tools, people will also attempt to reduce the data to bite sized infomorsels. However, the filters people apply without help from quantitative methods are different: we might only tend to remember the first few observations, the last few observations, and one or two observations in the middle that hook inside our heads as they’re just plain bizarre. In short, we remember anecdotes, and we tend to remember anecdotes because they’re unrepresentative, not because they’re representative. If we rely on anecdotes alone to inform use, then we think we live in a world where men bite dogs, vaccines cause autism and diseases, and most unemployed people have eight children, 60” plasma screen TVs, and tattoos on their necks.

### … But data visualisation means we do not have to reduce the data as much

Some level of data reduction is almost always necessary, because of the limits on the number of pieces of information we can engage with and juggle at a time. However, so long as we do not become overwhelmed, more data is better than less data. One of the primary benefits of data visualisation is that it helps to raise the cognitive ceiling, allowing people to understand, engage with, and be informed by more pieces of data than if the same facts were presented numerically alone.

## What makes a good visualisation? … is a bad question.

Although there are various schools of thought about data visualisation, which will be discussed later, they do not tend to point towards a single solution to the problem of what makes a good visualisation. This is because the question is a bad question, and is bad because it assumes there should be a single answer.

The right kind of visualisation depends on the purpose it needs to fulfil. This is easiest to see when thinking about the appropriateness of different forms of text. If you were tasked with producing a pithy by-line for a product, a three or four word piece of copy positioned alongside an image of a new car, then a hundred thousand word thesis on the psychology of automobile colour preference will not be the right solution. Likewise, if you were asked to produce a comprehensive review of the effectiveness of statins in reducing the risk of heart disease in males aged between fifty and sixty years, then the sentence “It’s not as bad as butter, but it’s not as spreadable, either”, is unlikely to satisfy the funders.

In the context of text, the need to think of effectiveness in terms of fit quickly becomes obvious, but for some reason many people do not extend this lesson to data visualisation. A frequent and erroneous notion I need to disabuse people of is that a good data visualisation should only take four or five seconds, at most, to understand and absorb. If this criterion were the only one applied to text then the Nobel Prize for Literature would be awarded to the producers of adverts and road signs. An effective data visualisation may take not just seconds, but minutes, or hours, or even days or weeks of training to understand correctly, let alone to gleam all available information from. As a case in point, consider the form of data visualisation known as ‘musical notation’, where sounds are mapped onto spaces according to a series of precise rules that can take years of specialist training to easily encode and decode.

The data visualisation journalist Albert Cairo has some suggestions about how to match visualisations to viewers. He suggests we try to frame the visualisation by thinking about the following binary oppositions:

* Unidimensional vs multidimensional:
* Light vs dense
* Familiar vs original
* Figurative vs abstract
* Decorative vs functional
* Redundancy vs novelty

The more the visualisation tends towards the left side of the binary, the more likely it is to be easily understood by a wide audience. The more it tends to the right hand binaries, the more likely it is to require the additional knowledge and engagement of a specialist audience, and the less suitable it may be as a visualisation for broadcast rather than narrowcast.

# Part Two: The Grammar of Graphics, and the Three Hats: two tools for understanding data visualisation in theory and practice

## There is a grammar of graphics

Grammar is a set of tools for describing the hidden structure of sequences of words. Just as there is a grammar of language, so there is a grammar of graphics. The grammar of language distinguishes words and sequences as adjectives, nouns, pronouns, propositions, adverbs, and so on, based on the roles and functions words and sequences have in the context of sentences and paragraphs. Similarly, the grammar of graphics divides different markings and spaces within a data visualisation into distinct functional components. Although there is less standardization of how these different visual components are named and defined, there is agreement that data visualisations contain many distinct parts performing distinct roles.

A book, ‘The Grammar of Graphics’, was written by the statistician Leland Wilkinson in 2005. The book represents a highly formalized approach to understanding what goes into producing a data visualisation, providing terms and ideas for helping researchers distinguish between different component parts of a visualisation. The components of the grammar of graphics that Wilkinson identifies include:

* **Variables**: these are pieces of data, associated with observations, that contain values.
* **Scales**: these are choices of transformations that can be applied to the values associated with variables. For example, an identity scale produces as output whatever was given as its input, and a logarithmic transformation produces as output the logarithm of the input.
* **Geometries**: these are specific shapes and forms which you, as producer of the data visualiser, must choose between. Your choice will depend on your understanding of the types of variables being represented, and the types of relationship between them. For example, a line is a different geometry to a point, and a line which connects two points is different to a line which reaches upwards or sideways from an axis towards a point.
* **Coordinate systems**: these define how space is represented in an image. For example, a Cartesian coordinate system represents space in terms of latitude and longitude, whereas a Polar coordinate system represents space in terms of distance from origin and angle. Many visualisations that look very different are in fact the same, but for the coordinate system used.
* **Aesthetics**: these represent the rules by which different variables are mapped onto different geometric features placed in different positions within the space as defined by the coordinate system(!). This is an initially confusing idea, but in many ways the most important take-home message from the approach. The difference between data visualisation and other forms of information visualisation can mainly be understood in terms of this mapping process.
* **Facets**: these are tabular arrangements of multiple graphs. They are a vital tool for data visualisation meant for internal use (i.e. for exploratory data analysis), and are a simple but very important tool for high information density data visualisation more generally.
* **Guides**: these are additional pieces of information which help the viewer to understand the context of the data visualisation. Examples of guides include legends for distinguishing between different types of line and point, and scales for helping people work back from a position or other aesthetic feature of a graph and the underlying values.

The ‘Grammar of Graphics’ approach is perhaps now more famous and widely applied through the work of data scientist Hadley Wickham, who has implemented suite of tools for producing visualisations structured around Wilkinson’s terms in the statistical programming language R.

As with the grammar of language, people can be highly adept at applying a grammar without an advanced capacity to describe that grammar formally. In fact, too much focus on formalization and standardization can backfire, leading to a focus on form rather than content, ignoring the message to focus on the messenger. Pedantry is good, but only within limits.

Though this course will not aim to teach you to become graphical grammar pedants, it will introduce, define, and apply the terms above in order to help you become more analytical and able to think about and communicate data visualisation design decisions both with others and with oneself. Without a precise enough vocabulary for talking about data visualisation decisions, it can be difficult to be clear about what one wants to communicate with a data visualisation. As I will argue towards the end of the course, good data visualisation is often a team effort, and so the ability to clearly communicate graphical design preferences can be vital.

## The Three Hats of Data Visualisation

To be effective and efficient at data visualisation, we need to learn to swap between three very different frames of mind, three ‘hats’. We need to swap between the **Artist**’s hat, the **Scientist’**s hat, and the **Engineer**’s hat.

* As an **Artist**, you will be concerned with beauty, with proportion, with form, with shape, with engagement. You will approach the task of producing a data visualisation as an opportunity to engage and inspire audiences, to bring feeling to numbers, to express yourself, to make art.
* As a **Scientist**, you will be concerned with accuracy, with precision, with efficiency, with testing and exploring theories, with disentangling and teasing apart complex causal relationships between variables, with learning something new about the world, and explaining what you have found to others clearly and honestly.
* As an **Engineer**, you will have to implement the plans agreed upon by the **Artist** and the **Scientist**, turning sketches and aspirations into reality. You will be faced with overcoming constant, annoying, niggling technical challenges and obstacles to realizing the blueprints. You will spend your time answering the question how. How do I access that data? How to I derive this variable? How do I change the line width? How do I label that point? How do I change the tick-mark on the axis? How do I do this? How can I do that?

As an **Engineer**, you will have a raw deal. If you fail to figure out how to do something, you will be blamed for your lack of knowledge and relevant skills. If you succeed, however, you will not be rewarded. Instead, the **Artist** and the **Scientist** will be thanked for the inspiration, boldness and insight of their plan. You merely carried out their instructions correctly, as you are expected to.

However – and this point is critical for understanding what data visualisation is in practice – *you will spend much more time wearing an* ***Engineer****’s hat than an* ***Artist****’s hat or a* ***Scientist****’s hat*. As a rule of thumb: data visualisation is about 80% engineering, 10% art, and 10% science.

## Graphic designers like curves; statisticians like rectangles

There is a persistent faultline in the world of data visualisation; a wide chasm separating two tribes that both lay claim to the Best Visualisation title. On one side of this divide are people with training in graphic design, typography and art, independent café-dwelling iphiles, sipping filter coffee from freshly ground beans from Guatamala. On the other side of the chasm are people with training in programming and statistics, dwelling somewhere lacking natural light, staring intently at lines of code on their homemade Linux desktops, part buried by fast food containers and well-thumbed text books on Java and C#.

For the greater good of data visualisation, these two tribes need to start talking to each other. Along the way, they’ll run up against a wide range of challenges, differences both of style and of substance.

Though there are many differences between these groups, realizing that graphics people tend to prefer curves and statisticians prefer straight edges and rectangles can be a great starting point for understanding these differences and challenges.

Curves tend to be preferred because there’s an aesthetic charm to them: they’re soft, organic, not sharp and jagged. By contrast, straight edges offer precision, concision, simplicity and visual acuity. Statisticians charge that curved and circular forms of data visualisation are difficult to understand and usually inefficient uses of the space available, producing visualisations with low information density. Data visualisers more focused on graphic design, however, recognize that the loyalty and level of engagement of the audience can’t simply be assumed, but must be earned, especially when communicating statistics to a general audience.

Of course, both tribes are correct, which is why it is important to be able to cross the chasm and change perspectives many times throughout the data visualisation production process. To use the metaphor applied just before, to produce good visualisations we need to change hats often, swapping frequently between thinking like a **Scientist** and thinking like an **Artist**.

# Wearing an Engineer’s Hat: What is Tidy Data: An Introduction and Extended Example

More than 80% of the time will be spent wearing the engineer’s hat, thinking about issues such as formatting, tidying, merging, and reshaping data, converting file formats, saving, loading graphical templates, and so on. Because of this it is important to spend some time thinking carefully about data management.

The data scientist Hadley Wickham has recently suggested that much can be learned from the principles of database management. Most people who work in large organizations have some passing familiarity with these principles through Microsoft Access, which is often used for managing personnel records, amongst other things. Important tasks in database management include joining tables and aggregation.

As an example, we may need to join three tables: Table A, showing when people joined the organization; Table B, showing when people left the organization; and Table C, showing employees’ current salaries. In order to be able to join these pieces of information, each table must have something in common, which allows a particular row in one table to be linked with a row in another table. This piece of information is known as a ‘key’. In the case of employees, the key should be unique: otherwise we may end up paying someone who’s left or vice versa. In most cases, people don’t have the same names, so joining tables by name might work in most cases, most of the time. However, in and of itself, it’s not a very robust key, because of course multiple people can have the same name. More information is usually needed in order to make the key unique. In this example, we might want to add date of birth and national insurance number to the key as well: though we might not be very confident that no two people share the same name, we can be more confident that if two records refer to the same name, the same birth date, and the same national insurance number, they refer to the same person.

Data containing keys only are of no use. They play a supporting role, and what they support are **variables** and **values**. Crudely, if **keys** show *something that was* measured, **variables** record *what* was measured (of the ‘*something that was measured’)*, and **values** report *the result* of measuring (what was measured of the something that was measured).

## Some untidy data

Given the above, there is a very simple way of storing almost all kinds of data, commonly known as the ‘long’ format. Long format data usually contain far more rows than columns. The columns only contain keys, variables, and values. As a simple example, consider the data shown in the table below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Surname** | **First name** | **Month of trial** | **Height** | **Condition** | **Weight** |
| Smith | Mary | 1 | 78 | Healthy | 45 |
| Smith | John | 2 | 62 | Sick | 52 |
| Jones | William | 1 |  | Healthy |  |
| Jones | William | 2 | 86 |  | 84 |
| Smith | Mary | 3 | 78 | Sick | 43 |

Figure 4 Some 'untidy data'

Given this table, the first task in data management is to work out which pieces of the data are the key, which are the variable, and which are the values. An important thing to realise, in this case as well as many others, is that the key – i.e. the information needed to uniquely identify an observation – is spread across a number of columns. These are Surname, First Name, and Month of trial. This is because neither name nor day of trial, on their own, represents a unique observation: A person may be observed on more than one occasion, and more than one person may be observed on any one day. A key could be formed from these variables by combining joining the fields, for example using the concatenate command in Excel. Adding additional underscore characters to improve readability, this would produce the following:

|  |  |  |  |
| --- | --- | --- | --- |
| **Surname** | **First name** | **Month of trial** | **Key** |
| Smith | Mary | 1 | Smith\_Mary\_1 |
| Smith | John | 2 | Smith\_John\_2 |
| Jones | William | 1 | Jones\_William\_1 |
| Jones | William | 2 | Jones\_William\_2 |
| Smith | Mary | 3 | Smith\_Mary\_3 |

Figure 5 Producing unique keys

Now to the variables: these are typically, but not always, the columns in the table which are not part of the key. In our case the variables are height, condition, and weight.

Finally, the values: in this case, these are the cells ranging from the second row to the last row and, from the left, the fourth to the last column. Schematically, the way these three pieces of information are arranged is shown in the table below:

|  |  |
| --- | --- |
| **Keys** | **Variables** |
| *Values* |

Figure 6 A visual representation of the 'anatomy' of the untidy data

In long format, the variables and values are placed side by side rather than on top of each other. The long format equivalent of table X therefore becomes:

|  |  |  |
| --- | --- | --- |
| **Key** | **Variable** | **Value** |
| Smith\_Mary\_1 | Height | 78 |
| Smith\_Mary\_1 | Condition | Healthy |
| Smith\_Mary\_1 | Weight | 45 |
| Smith\_John\_2 | Height | 62 |
| Smith\_John\_2 | Condition | Sick |
| Smith\_John\_2 | Weight | 52 |
| Jones\_William\_1 | Condition | Healthy |
| Jones\_William\_2 | Height | 86 |
| Jones\_William\_2 | Weight | 84 |
| Smith\_Mary\_3 | Height | 78 |
| Smith\_Mary\_3 | Condition | Sick |
| Smith\_Mary\_3 | Weight | 43 |

Figure 7 Long format data

The data in this format is generally less human readable. However, long format data like this is usually easier and quicker to work with than data in other formats. As, in practice, the majority of the work involved in data visualisation is data management, learning more about these database management tools and principles can help you become much more agile as data visualisers.

Long format data can be thought of as the clay from which data visualisations can be formed. Data management tools, such as the filter and pivot table features of Excel, are therefore the tools with which this clay can be worked into the most appropriate shape. Given data in long format, a large number of possible alternative data structures can be formed. All of these can be thought of as one or other form of ‘wide’ format data, containing more columns but fewer rows.

Producing wide format tables is itself a rudimentary but important form of data visualisation. Certain forms of relationship are easier to identify if the associated values are arranged some ways rather than other ways. For example, if we have observations of the same individuals on many trial dates, we may find it easier to see trends over time if we split the keys into individuals and days, and arrange the days along the columns, and the individuals along the rows. In the example we are using, there seem to be two good candidates for which variable we wish to see in the cells intersected by these rows and columns: the condition, and the weight. Imagine we use weight as our variable. If we had a few more observations in the dataset, then the wide format dataset we produce might look as follows:

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | Month | | | | | | | | | |
|  | | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| Person | **A** | 78 | 79 | 75 | 69 | 64 | 65 | 62 | 59 | 59 | 62 |
| **B** | 56 | 59 | 62 | 61 | 55 | 58 | 63 | 67 | 69 | 62 |
| **C** | 92 | 99 | 105 | 101 | 104 | 111 | 114 | 105 | 97 | 99 |
| **D** | 48 | 49 | 49 | 56 | 59 | 55 | 63 | 66 | 61 | 52 |
| **E** | 66 | 70 | 64 | 62 | 68 | 63 | 64 | 64 | 62 | 67 |
| **F** | 99 | 98 | 87 | 88 | 86 | 85 | 86 | 82 | 79 | 85 |
| **G** | 154 | 160 | 161 | 154 | 148 | 149 | 142 | 139 | 137 | 133 |

Figure 8 An arrangement of data to facilitate understanding and thinking about changes over time.

Although we might be able to imagine the changes over time for these people, it will not be easy to pick out a general trend in these numbers. Imagine we have not just seven people, but seven hundred. If we have no missing observation, then this table would be a matrix of 7,000 values. This is too much data to easily cope with or make sense of. It is for this reason we need to produce some summary statistics, to cut down the number of values we have to think about and compare at one time. It is also for this reason we need data visualisation, to make it easier to cope with more data.

What would be an appropriate summary statistic to use? We might be tempted to look at mean weight observed in each month. If we did this, we would have turned either 70 (seven people) or 7,000 values into 10, which is definitely more manageable. However, would these means be meaningful? We know we are observing people with very different starting weights, and quite likely other characteristics too. It might instead be more appropriate to convert each person’s value into an index of its initial observation: 100 meaning the same weight, 110 meaning 10% greater than initial weight; 90 meaning 90% of the initial weight, and so on. Doing this in the above case produces the following:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Person | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| A | 100 | 101 | 96 | 88 | 82 | 83 | 79 | 76 | 76 | 79 |
| B | 100 | 105 | 111 | 109 | 98 | 104 | 113 | 120 | 123 | 111 |
| C | 100 | 108 | 114 | 110 | 113 | 121 | 124 | 114 | 105 | 108 |
| D | 100 | 102 | 102 | 117 | 123 | 115 | 131 | 138 | 127 | 108 |
| E | 100 | 106 | 97 | 94 | 103 | 95 | 97 | 97 | 94 | 102 |
| F | 100 | 99 | 88 | 89 | 87 | 86 | 87 | 83 | 80 | 86 |
| G | 100 | 104 | 105 | 100 | 96 | 97 | 92 | 90 | 89 | 86 |
| Mean | 100 | 104 | 102 | 101 | 100 | 100 | 103 | 102 | 99 | 97 |

Figure 9 Adding a summary margin to the table

By including the mean as an additional row, we have a clear summary of the trend over time. Because the values are next to each other, presented left-to-right, we are able to imagine the trend fairly easily without plotting it. However, to be able to compare the mean against each of the individuals’ trends would be a feat not just of imagination but of memory. Not only would we have to mentally construct each of the individuals’ lines, but imagine overlaying each on top of each other.

# Data Visualisation as ‘Box Wiring’

In data visualisation, ‘dimension’ is an overloaded term. Within the dataset, the term dimension refers a variable: something that varies, can be more than one thing. For example, sex is a variable, as it can be male or female (or ‘other’). Age is a variable. Car ownership is a variable. Being in receipt of benefits is a variable. And so on and so on.

Data visualisation is at its core about the data later. And the data layer is fundamentally the result of deciding on a set of rules by which dimensions of the data map onto dimensions of aesthetic and graphical form. In this usage, a dimension refers to more than simply the position of something along the x, y or z axis. Instead a dimension could be colour, it could be line width, it could be bubble size, or it could be a multitude of other graphical forms.

Given this definition, many apparently simple, two-dimensional graphs on closer inspection may contain five or six or more dimensions. For example, how many dimensions does the graph below have?

[An external file that holds a picture, illustration, etc.
Object name is bjpg54-899S1.jpg](http://www.ncbi.nlm.nih.gov/pubmed/15588533)

Figure 11 Shaw M & Dorling D (2004) "Who cares in England & Wales? The Positive care Law: cross sectional study" *British Journal of General Practice*

The answer is on the next page.

The answer is four.

How can we tell? Well, we know it must have at least two dimensions: both the vertical axis and the horizontal axis are obviously mapped to data variables: the vertical to “% of population providing free care”, and the horizontal to “% of population with health needs”. The other two dimensions are a little but harder to identify, but clear enough once we know what to look for.

Firstly, note that each of the bubbles is of a different size. Notice also that there seems to be a wide range of bubble sizes. Secondly, note that some of the bubbles are shaded with a dark grey, and others with a light grey.

Unless we assume the data visualiser has sized the bubbles and coloured them in arbitrarily, then we should conclude that the bubble size and bubble shade are both also the result of mapping from data to aesthetics. So, we can conclude that this bubble plot is in fact a four dimensional data visualisation.

There are a couple of other facts we can tease out from inspecting the image: firstly, there appear only to be two shades of grey, but more than two sizes of bubble. We might therefore feel safe to assume the bubble shade is mapped from a binary variable (or at least a variable where only two distinct values were observed); and that the bubble size is mapped from an ordinal or cardinal variable.

Now, to see if we were right, by looking at the associated description (which also applies to another graph, not reproduced here):

An external file that holds a picture, illustration, etc.
Object name is bjpg54-899S1.jpg

So, the binary variable is ‘North’ or ‘South’, and the ordinal or cardinal variable turns out to be population. (Technically an ordinal variable, as we can’t have fractions of a person, but in practice this is not usually an issue.)

To help think about the mapping processes involved in this visualisation, it can help to perform some simple information visualisation of the data visualisation. Imagine we have a box, with inputs on one side – the data variables – and outputs – the graph aesthetics – on the other side. Our job, as data visualisers, is to choose which inputs to map to which outputs. We choose how the box is ‘wired’. In this case, we can represent the ‘box’ as follows:

|  |  |  |
| --- | --- | --- |
| **Data Variable** |  | **Graphical Aesthetic** |
| % of population providing free care | Position along horizontal axis |
| % of population with health needs | Position along vertical axis |
| Areal unit population | Size of bubble |
| Geographical location (North or south) | Colour of bubble |

In the current example, the box is ‘wired’ as follows:

|  |  |  |
| --- | --- | --- |
| **Data Variable** |  | **Graphical Aesthetic** |
| % of population providing free care | Position along horizontal axis |
| % of population with health needs | Position along vertical axis |
| Areal unit population | Size of bubble |
| Geographical location (North or south) | Colour of bubble |

However, even with only the variables and aesthetics listed here, there are many other ways in which the visualisation could be wired. For example, we could map areal unit population to horizontal axis and % of population providing free care to size of bubble, or location to vertical axis and % of population with health needs, and so on and so on. Each of these alternate mappings would make different relationships within the data salient, and some would be more effective than others.

As data visualisers, we exercise our judgement at all stages: in choosing the Data Variables (the input side of the box), in selecting the geometric forms (lines, points, bars etc) and graphical aesthetics which are to be data controlled (the output side of the box), and in selecting the mapping rules which link data to aesthetics (the wiring inside the box). Fundamentally, this is what data visualisation is about, and what distinguishes data visualisation from other forms of information visualisation. Although at first glance this definition may seem fairly reductionistic, thinking more carefully we can see that it is not, as it still leaves us with uncountably many possibilities for visualisations, just as uncountably many melodies can be made with a finite number of notes, and uncounably many books can be made with a finite number of characters. Data visualisation as ‘box wiring’ is a definition, not a restriction.

# Data Visualisations for Internal Consumption: Exploratory Data Analysis

Data visualisations are not all meant for public or generalist consumption. Likewise, they are not all meant to communicate a series of messages or facts that the information ‘sender’ already knows. Instead, data visualisations are often used by data scientists and specialist in order to explore and understand subtle and complex patterns in the data. Exploratory data analysis has, over the last twenty five or so years, become a well respected means of scientific research, and data visualisation is a vital component of this means of enquiry. My own forms of visualisation – lexis plots of demographic data – probably fall much closer to this end of the data visualisation spectrum than any Guardian infographic.

Data visualisation for exploratory data analysis (EDA) tends to have some different aims than visualisations for external consumption. EDA graphics do not have to be pretty, but they do need to be accurate and quick: they should ‘tell the data’ simply and cleanly. They need to be quick and easy enough for an analyst to use that they encourage feedbackEDA visualisations need to be quick to make and adapt, and make it easy and even fun for researchers to want to and know how to ask questions of the data. They must make it easy for an iterative process of thinking of questions, thinking of ways of exploring these questions graphically, and using the results of the visualisations to generate more questions.

# Graphs within Tables: Lattice plots and small multiples

Hans Rosling’s data visualisations are well known. The type of visualisation he is most famous for using is known as a motion chart. In a motion chart, a long series of graphs are presented, one after another, which share the same axes, scales, formatting, and so on, but where one variable is varied incrementally. Most often, the variable that varies between images is time, and motion charts are best thought of as a form of animation.

Because of the way they work, Rosling’s motion charts effectively map time onto time, but typically using a very large scaling factor. For example, one year per second, so that changes that emerged over decades can be presented in under a minute. Motion charts can be an effective method of visualisation, but they are not necessarily the best way of showing changes over time.

Firstly, just as people tend to remember the beginning and end of films or stories better than the material in between, so it becomes easy for our memories to deceive us when thinking about have values have changed over time, if we do not have those intermediate graphs at hand. Secondly, there are usually computational and technical challenges to presenting motion charts: for example this piece of paper is not yet technologically advanced enough to show animations. Thirdly, without someone as charismatic, knowledgeable and engaging as Hans, it is easy to get lost in a motion chart, especially one where much is going on at once, to become overwhelmed by the amount and speed of the data presented, and as a result to become disengaged by them.

The simple but surprisingly effective solution to these problems is to use a method of visualisation which will be common to anyone who’s read a comic book. The method was illustrated by Eadweard Muybridge in 1886, as shown below:

# http://www.hrc.utexas.edu/exhibitions/permanent/windows/southeast/images/muybridge_large.jpg

Figure 12 Horse in motion, Eadwaerd Muybridge, ca. 1886 (<http://www.hrc.utexas.edu/exhibitions/permanent/windows/southeast/eadweard_muybridge.html>)

These photographs showed something about the way horses run that surprised people: at some stages within the gallop, shown in the first four panes, all four limbs are off the ground. Of course, horses are not invisible, and every frame in the horses’ motion had been observed directly by people for thousands of years, it took both photography, and this particular, tiling arrangement of photographs, in order to prove this aspect of the gallop to observers.

This form of tiling of images is a simple but powerful technique in data visualisation, especially visualisation used for exploratory data analysis. There are a number of different names for this approach, and different rules which can be used to determine the arrangement. Edward Tufte calls these tiling arrangements ‘small multiples’, other names for them include ‘lattice plots’ and ‘trellis plots’.

In the horses in motion graphic above, the arrangement of images is determined by time, starting at the top left, moving right, and then wrapping to the next row once the last column in the row has been reached. This is the same kind of arrangement that is applied in text, such as the text you are reading now. The ‘rule’ for mapping from data to visuals, therefore, is that the position along a long strip of tiles is determined by the variable ‘time’. The tiling adds one more dimension to the visualisation, even though the tiles extend in two dimensions, both rightwards and downwards.

A second variant of the lattice plot is also possible. In this, both column position *and* row position are determined separately by a different variable. In this arrangement, two additional dimensions have been added to the visualisation through tiling. For example, if the row position could be determined by gender, and the column position could be determined by age group. We have therefore added two more dimensions to the visualisation.

If we think about it, was can see that the tiling arrangement represents a series of graphs within a table. As we saw earlier, the arrangements of columns and rows and groupings within tables is itself a form of data visualisations.

# Graphs within Graphs: Representing more than one variation within space

In most areas of data visualisation, graphs within graphs are less common either than graphs within tables (small multiples) or graphs alone. The exception tends to be when looking at geographical variations in more than one variable at a time, and a little thought will show why.

In a visualisations of spatial relationships, the x and y axes are already ‘used up’ for plotting latitude and longitude. A third aesthetic element is needed in order to show how a single additional variable, such as mean rainfall, varies over space. Typically, and in some ways problematically, the third aesthetic element is colour, with darker shades perhaps indicating more of something, and lighter shades indicating less of something.

This visualisation is fine for a single value, but for multiple values something else is needed. One approach is to use the small multiple approach: plotting heatmaps on the same spatial scale, side by side. However, such arrangements may well be at the limits of human information processing density. Additionally, people are not good at distinguishing between different shades of colour in an objective and consistent way: the same shade of grey can look light or dark depending on context. An approach that is sometimes used in order to represent more than one variation over geographical space is to plot graphs on top of graphs.

In these graphs-within-graphs arrangements, the background graph is of the geographical region, for example, the United States of America. Other graphs would then be placed within the centre of each state, or some other subsection of land. For example, the

# Decomposing visualisations into layers

Let’s look at a few data visualisations. Let’s imagine they’re like jokes, or special effects, or magic tricks, and we want to appear both smug and pedantic by explaining how they work. We can do this by decomposing them, slicing them first into layers, and then looking at what each layer contributes to the overall effect.

* At its most minimal, a data visualisation needs just one layer: the **data layer**. The data layer is simply the geometric form created from applying the mapping rules to the data. Typically, it might just be a series of dots or a line formed by connecting these dots in a particular sequence. Tufte’s sparklines are about as close as data visualisations get to the ‘naked’ data layer. Even they contain some minimal annotation: a starting value, a closing value. Without these minimal additions, the data layer alone is without context, and without context it’s not particularly informative.
* ‘Beneath’ the data layer is the **support layer**. This contains things like gridlines, vertical and horizontal tickmarks and labels, and so on. Like bass and percussion in a song, the support layer is something that, if it’s doing its job properly, you won’t notice, but if it’s doing its job badly, you will. Just as bass and percussion and can be too loud, so the background layer can be too thick, too heavy, too distracting. A typical example of this would be if there are too many background gridlines, which are too thick, and the same colour as the geoms being represented in the data layer.
* The third type of layer is the **annotation layer**. For better or worse, this is where there is greatest scope for the designer to be an artist, a buddy, a guide, a helping hand, or an annoyance, a distraction, a biased commentator.

Visualisations can have more than three layers. They can have multiple data layers, multiple support scales – such as different axes for different series – and multiple forms of annotation. However, there can be some subjectivity in terms of whether two shapes are part of the same layer or different layer, as ultimately the distinction can be semantic. Just as thinking too much about verbal grammar can lead to spending more time thinking about the form than the content of sentence - wincing at the sight of errant apostrophes, for example- so it’s possible to go too far with visual grammatical formalism, and spend more time on decomposition and description than on construction and communication. Use the analytical approach – splitting visualisations into layers based on their visual roles and grammars – when it helps improve clarity of thought and communication. If and when it becomes a barrier, then switch to a more informal and intuitive mindset: More artist; less scientist.

# Heat Maps in Excel

Something that bridges the divide between tables and graphs is the heat map, in which different values are represented by different colours. Think of an image from a thermal camera: hotter areas are represented by reds, cool areas by blues or blacks. Similarly, we can use different colours to represent different values. The choice of colour matters a lot for interpretation, and as discussed elsewhere colour should not usually play a support rather than central role in communicating data visually, due to the prevalence of colour blindness in the general population. However, producing a rudimentary heatmap in Excel is relatively straightforward, and can be a useful way of exploring and presenting large amounts of data, despite the limitations.

To produce a heat map in Excel you need to do the following:

1. Select a series of cells you want to be coloured
2. Select conditional formatting
3. Select ‘colour scales’
4. Then select ‘more rules’
5. Then, in the dialogue box titled ‘New Formatting Rule’, select ‘format all cells based on their values’ in the ‘Select a Rule Type’ box; and choose the appropriate options in the ‘edit the rule description’ box in the bottom half of the dialogue box.
6. The options in the ‘edit the rule description’ box include:
   1. Format style: 2 colour scale – usually better when there are no negative values; 3 colour scale – usually better when there are some negative values
   2. Type: select ‘value’ for consistency across heatmaps
   3. ‘color’ (sic): the colours which represent the lowest, medium, and highest value
7. A preview colour bar appears at the bottom of the dialogue box showing the continuous colour scale produced.

I have found heatmaps particularly useful for highlighting correlations, which range from -1 to 1. I typically use a strong blue to represent -1, white to represent 0, and red to represent 1. Another, more tricky but ingenious, application of heatmaps is in producing fairly simple choropleth: maps of places which are coloured according to values associated with those areas. For example, if you want to produce a choropleth of crime rates across the UK, then each cell could represent a particular part of the country, and be coloured according to its crime rate.

# Templates

Something that can potentially save a lot of reinventing the wheel is the template feature. Once you have customized the graphical features – colours, axes, formatting and so on – of a single graph, you can save those customisations as a template, which you can then apply to later visualisations. When you need to produce a large number of graphs with exactly the same format, then using templates can save a lot of time. Note, however, that templates need to be saved to the relevant directory locations on any machine you wish to make use of them, and sadly they don’t store all of the customisations you might hope they would. Often, some level of further manual manipulation and customization will still be necessary.

# Excel: useful things to know about

Excel is a much used, if not much loved, application because of the wide range of features and options it includes, and because many of these options and features can be used in ways that the creators of the program likely did not have in mind when they included such options. Which features it’s useful to know about can depend on how you intend to use it. However, a short list of useful features includes:

* Knowing how to turn off gridlines: If you are using Excel to develop something like a small multiple or other arrangement of graphs, then being able to turn off the gridlines means you can use a worksheet as your canvas: placing text within cells inside the worksheet along with , including multiple smaller visualisations on a single
* Knowing how to add a second y axis
* Using templates
* Important and exporting text documents
* Filters and sorting
* Pivot-tables

# Lens Plots

Steven Few, whose books on Data Visualisation within business focus extensively on graph and table design in Excel, recommends producing what he calls ‘lens plots’. These are plots of two or more outcomes associated with individuals or groups (e.g. regions, business sectors etc) that are presented side-by-side, in which the groups are sorted according to one of the outcomes for all plots. An example of a lens plot is shown below: Centres have been shorted according to discharge rate, and this order is applied to the other outcome too – Mean LOS.

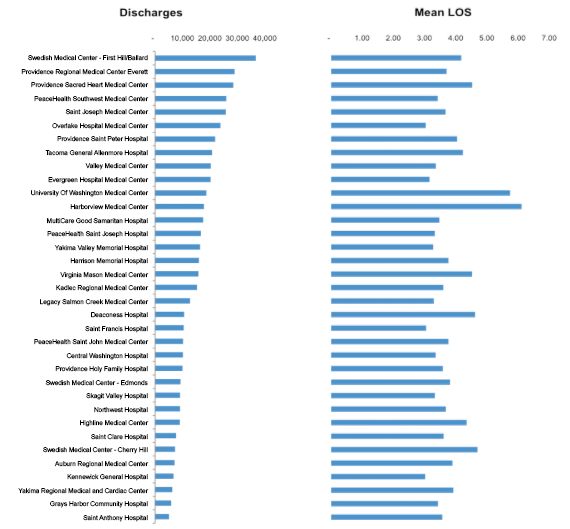


Figure 12 An example of a lens plot. (Source: http://ksrowell.com/blog-visualizing-data/page/4/)

Lens plans are a fairly simple and effective way of looking for correlations between variables. In this example, if there were positive correlation between discharges and mean LOS, then both the left and right image would have similar shapes: widest at the top, narrowest at the bottom. If there were a negative correlation, then the mean LOS bars would be like a mirror image of the discharges bars: widest at the bottom and narrowest at the top. Instead, in this case there appears to be no clear correlation between the variables.

# Useful Excel Add-ons

There are a number of macros and add in programs available for Excel. Some of these are free, and others cost considerable amounts of money.

One simple and free tool which I would recommend people use is the X-Y chart labeller. This is available below, and can be used with both older and newer versions of Excel.

<http://www.appspro.com/Utilities/ChartLabeler.htm>

To an extent, the need for a tool for making Excel do something should be taken as a clue that Excel might not be the most appropriate piece of software to use. Excel is flexible within limits, and for many data management and visualisations tasks can produce the results required given sufficient ingenuity.

Although it is simple to start with, and ‘hackable’ in many ways, before long trying to adapt Excel to the tasks can end up taking more effort than using something more specialist data visualisation software such as Tableau or R. As a rule of thumb, though Excel may be good for initial development of visualisation, it is generally better to bite the bullet and invest the financial and training resources necessary to learn more specialist software where the most appropriate visualisations cannot be produced easily in Excel.

# Take-Home Messages

The main take-home messages of the course are:

1. Data visualisation must involve the consist mapping of variables within a dataset to aesthetic elements on an image.
2. Data visualisations usually include a number of distinct layers, defined by the role they play in conveying information to the viewer. At its simplest, there are three layers: the background layer, the data layer, and the annotation layer. What makes a data visualisation a data visualisation, rather than some other kind of information visualisation, is the presence of the data layer.
3. You need three distinct mind-sets for data visualisation, three ‘hats’ that you need to swap between in order to complete different tasks involved in the process. These are the Scientist, the Artist, and the Engineer. On a day-to-day basis, you will spend most of your time wearing the Engineer’s hat. (Also, Scientists and Artists tend to squabble.)
4. There is a grammar of graphics, a way of describing the different elements that go into producing a data visualisation. Learning some of these terms, even if we don’t apply them consistently, can help us think and communicate more clearly about the elements, processes and aims of producing a visualisation.
5. What counts as a good data visualisation is heavily context dependent. A good data visualisation is one that is a good match for the audience. Audiences can be divided crudely into ‘internal’ and ‘external’ (we and they); and into ‘specialist’ and ‘generalist’. Often but not always, internal audiences are specialists, and external audiences are generalists. Data visualisations meant for internal audiences are visualisations we produce for ourselves. In these cases, visualisations are used as part of exploratory data analysis, to learn something new. By contrast, data visualisations meant for external audiences aim to communicate something we already know to other people.

# Tables are also Data Visualisation

This is an important point to stress: even something as apparently mundane as a table is a form of data visualisation. Tables contain a range of visual clues, suggesting to the viewer implicit relationships between variables. Although it’s not central to this course, effective and consistent table design is an important form of data communication. As we will see elsewhere, the same data can be presented in a large number of different ways. The ‘long’ data format may be one of the cleanest from the perspective of data management, but it tends to be a poor choice when it comes to communicating important relationships between and within variables. The arrangement of values within tables tells the viewer something about how we perceive the associated values to be related. The proximity of values within cells implies something about how they are grouped.

## Tables of tables

Groupings within categories can be presented using various visual and positional clues. For example, we might want to show mean consumption of a product varies for different age groups, different genders, and different regions. We are interested in presenting these variables over time, and importantly, over the same time periods. We might therefore choose to put time along the x axis, with different columns marked 2005, 2006 and so on; and category along the y axis, the different rows labeled with different age groups, genders, and regions.

There are two different ways in which this tabular arrangement could be presented. At one extreme, we could design three different tables: one for age, one for gender, and one for region. However, separating the tables this way has two disadvantages. Firstly, it uses more space than is necessary, because the same information has to be repeated in the column arrangement. Secondly, it makes it harder for the viewer to make assessments about the relative influence of gender, region, and age effects on the outcome of interest.

The solution is to present the information as a single table, containing three sub-tables. Each table shares the same column, but has different row labeling. The sub-tables are separated vertically either through lines, shading, or spaces. An additional column is added to indicate the super-groupings – gender, region, and age group – separating the three clusters of rows.

Although tabular arrangements like this are fairly common, it is important to think carefully about the way in which they present information, and how this facilitates or impedes the reader or viewer’s comprehension of the data. In a sense, grouped tables like these present data which varies within three dimensions: time, group, and variable within group. Tables like these are in fact tables nested within tables. In the case presented the nesting or supergrouping occurs at the level the row. A generalization of this is to also have some form of nesting by column. In this example, imagine that the column is now ‘super-grouped’ by region, and the rows ‘super-grouped’ only by gender and age group. Even with just a few variables, there are a large number of possible permutations of possible nested tabular arrangements.

As data visualisation tools, summary tables like the one described have strengths and weaknesses that balance well with graphs. It is, in general, always possible for someone to create a graph given the information in a table, but not as a easy to do the converse. The presentation of exact numeric quantities allows precise comparison between any two or three values within the table and, where the values are of the same variable as presented in other tables, with other tables. However, it tends to be much harder, given the information in a table, to get an intuitive sense of the relationship between variables, and to get a quick at-a-glance sense of general trends and patterns affecting many values and variables. Often, the process of decoding tables involves the reader ‘drawing an image’ of the associated visualisation in their heads – or for the particularly curious in Excel or a similar package – in order to fully absorb the information presented.

# Edward Tufte’s Mind tools

Edward Tufte is a statistician who has written some highly regarded and beautifully constructed books on the theory of data visualisation. He is also a sculptor, and a few years ago was profiled by the BBC, opening up a small arts studio in New York.

<http://www.bbc.co.uk/news/magazine-16976145>

Although a scientist, Tufte’s books on data visualisation are not particularly scientific. They are too personal for that. They are idiosyncratic visual essays on the ideas, aims and principles of data visualisation. The books describe a particular philosophy to data visualisation which alienate some practitioners but engage others.

If Tufte’s approach to data visualisation were to be summarized in a single word, it would be ‘minimalist’. Tufte railed in his books against what he considers ‘chart junk’: something included in a visualisation that either does not add to, or even subtracts from, the viewer’s understanding of the data itself. Instead, the visualisation should where possible present the data as clearly, simply and cleanly as itself. Tufte argues for a kind of Okham’s Razor position for data visualisation: where graphical elements are not needed, they should as a rule of thumb be removed.

Two concepts and one two types of visualisation follows from this commitment to minimalism. The two concepts are ‘data-ink ratios’ and ‘lie factors’; and the two types of visualisation are of ‘small multiples’ and ‘spark lines’. These will now be discussed in turn.

* The **data ink ratio** is an attempt to quantify the question: how much of this visualisation is telling me the data? For example, if a value could be encoded in a thick line, or a thin line, why not use the thin line? If the data contains only three points, but the graph contains twenty vertical and horizontal gridlines in addition to it, then are these gridlines needed? How much of what has been drawn plays a central, essential role; and how much plays a peripheral, non-essential, support role? If the data ink ratio is too low, then Tufte suggests ‘Chart Junk’ could be the culprit, something illustrated more through examples than a formal definition.
* A **lie factor** is an attempt to quantify how far wrong a viewer of a visualisation could go in trying to discern the values that were used as its input. For example, consider two bars, A and B, each representing a value. The value of A is 3 and the value of B is 6. The ratio of these two values is two, therefore, the viewer might expect the length of bar B to be twice that of bar A. If, instead, bar B is three times the length of bar A, or 1.5 times the length, then the visualisation is ‘lying’ to the viewer about the underlying data. The greater the discrepancy between the true values of the data and the values implied by the visualisation, the greater the lie factor.

Although the austerity and minimalism at the heart of Tufte’s suggestions may not seem particularly appealing, cutting the fat in this way means that very data rich visualisations become possible. Of particular interest are small multiples and sparklines.

Sparklines are graphs that contain almost nothing but the data layer. If these data plotted is the amount of something over time, then the sparkline might simply be the line produced by plotting the data on a grid, but without the grid itself. At the start of the line, the opening value in the series would be shown, and at the end of the line, the closing value of the series, but nothing else would be included. Because so many supporting graphical elements have been removed, sparklines take up very little physical space, meaning many of them can be placed in close proximity and fit on a single page.

The tiny footprint of sparklines means they can be a highly versatile and adaptable form of visualisation. Tufte, who printed his own books in order to be in full control of the typography and graphic design, shows how sparklines mean that text and data visualisation and be integrated seamlessly, as sparklines are small enough for graphics to be fitted not just alongside or close to paragraphs, but inside them. For example, rather than writing “stocks in A have been more volatile than stocks in B (See figures 13a and b)”, the associated sparklines for A and B could be embedded within the sentence itself.

Although this kind of integration of graphics within text has not so far taken off to the extent Tufte may have hoped for, minimal figures, by allowing high information density, make possible a much more common kind of visualisation. Tufte calls these visualisations ‘small multiples’, but they are also given names such as lattice plots and trellis plots. Rather than being figures within text, they are figures within tables, and discussed in more detail elsewhere in this manuscript. [LINK]

# Why is Facebook blue?

Facebook is blue because its founder, Mark Zuckerberg, is red-green colourblind. Although total colour blindness is rare a substantial proportion of the population is colour blindness in one form or another.

[Estimates of proportion colour blind; different ways of being colour blind; differences between sexes]

It is important to think carefully about the implications of colour blindness to effective data visualisations. Without awareness of this, your data visualisations will be accessible to fewer people. With proper consideration, it is possible to design colour visualisations that allow people without colour blindness to benefit from and engage with the visualisation, but that are not inaccessible to people whose eyes and brains do not comprehend the full palate.

Colour blindness accessibility is not the only reason to think carefully about how colour is best used within data visualisations. There are a number of optical and cognitive traps that the use of colour and shade, if applied without due consideration, can cause both the designer and viewer of data visualisations to fall into.

These traps are best avoided through a two fold strategy: Firstly, to be aware sufficiently of how people interpret shapes, forms, colours, hues and shades to be able to spot a visual trap at a distance. Secondly, we should think about how to use colour only as a supporting mapped graphical aesthetic. What I mean by this is that colour should, where possible, only reinforce information that is available elsewhere. For someone who is not colour blind, colour makes information even clearer, whereas someone who is colour blind will not be ‘blind’ to one of the dimensions of the data being mapped within the image. The use of colour as a secondary rather than primary visual cue is best understood by thinking about the box wiring diagrams presented elsewhere. Imagine one input, a single data dimension, mapping onto two aesthetics: colour, and something else. The double encoding of information, producing some level of redundancy and reinforcement, isn’t necessarily a bad thing. We’re frequently told, when presenting, that we should “Say what we’re going to say, say it, and say what we’ve said”, and to an extent double encoding is applying the same principle to the visual realm. People won’t necessarily pick up everything you’re saying to them first time around. However, we can also go over-board with visual repetition, just as we can with verbal repetition, saying the same things repeatedly, in different ways, reinforcing the same points again and again, treading the same ground, rehearsing the same old arguments, saying nothing new, just the same old, same old. (And so on, and so on!)

# Colour and cognitive science

An engaging and effective way of helping to recognize visual traps is to spend some time looking at optical illusions, thinking about how they deceive, and how their capacity to deceive sheds light on the kinds of post-processing and guesswork applied automatically by the mind in order to interpret a three dimensional world based only on information from two nearby patches of photosensitive cells lodged close to the front of our skulls.

# Escape the Tyranny of the Default Option

It’s easy to go with the flow, to order the dessert as everyone else is, to say you admire a musician or an author because someone else gushes with enthusiasm about them. Likewise, Excel likes to suggest visualisations for you: these are the options that are within easy reach, require little graphical rejigging and manipulation, and that Excel presents prominently to us as big icons near the top of the screen.

It is important, however, to remember that Excel has bad taste, and presents a range of in-built visualisations that satisfy neither the scientist nor the artist. It’s all too easy to simply accept these default options. On balance, this is a bad deal: you save a few seconds or minutes, but the dozens, hundreds or thousands of people viewing the visualisation can see this lack of effort, and tell the garish and dumb visual design choices as clearly as if the graphic was had ‘design by Microsoft’ written on it in 24 point comic sans wordart.

The best way to avoid falling for the mediocrity of the default option is to have designed the visualisation *before* you start producing it. Design the visualisation before you even switch the computer on, using pen and pencil. If it helps to, first draw out the aesthetic mappings you want to perform: which variable controls the x axis, which controls the y axis, which controls bubble size, point shape, and so on. Having done this, draw out what you think this visualisation should look like, given the mapping rules you have just specified.

Your paper sketch is your blueprint, and scientist and artist need to collaborate at this stage in producing it. Once produced, the scientist and artist need to make way for the engineer, who has the unenviable task of turning the blueprint into a reality, using the tool available: Microsoft Excel. This might involve a slow and frustrating learning process, searching help forums, testing out multiple options, hacking, tweaking and rearranging data and graphs for what may feel like far too long, eventually developing a series of templates and tricks and hints and tips for shaving off a bit more time what that particular visualisation is not needed.

Sometimes, we may find a particular data visualisation is just too impractical to produce in Excel, in which case we will be forced to either compromise by using the closest available match in Excel, or to make the painful and time-consuming transition to using a software program that is either more expensive (Tableau) or much less user friendly (such as R).

# The correct geom (shape) depends on the reality being represented

We will discuss the grammar of graphics, and the role of geoms within it, in more detail later. For now it is simply important to note that a line is more than just a line, a point is just a point, and a bar is more than just a bar. Each of geometric forms is an implicit declaration by the data visualiser about the state of the world. Or at least, about the little bit of the world that has been recorded, categorized and summarized by the data used in the production of the visualisation, about the relationship between variables, and the relationship between observations. For clear communication, both the data visualiser, encoding the data into geometric forms, and the viewer, decoding the visualisation, need to hold similar notions about what shapes imply what relationships.

# The Arrow of Time

People tend to imagine time running left to right, and so changes over time might best be laid out along rows rather than columns. The length of variable labels can also be a factor in thinking about whether a variable should be laid out by row or by column. Of course, people read left to right, and would prefer not to have to keep tilting their heads in order to make sense of a table. For this reason, nominal and ordinal descriptions, such as the type of a product or demographic category, are usually presented along rows rather than along columns.

# Different Geoms imply Different Relationships

It’s possible to make exactly the same decisions about how to map variables onto graphical elements, but suggest very different things about the relationships within and between the variables. This is because the type of the element matters as well as the position. The simplest example of this is the difference between two points, and one straight line. Both can be based on exactly the same data, but whereas the two points suggest separateness between the two observations, the lines suggests, quite literally, a connection between them.

# References

## Books

If you only want to read one book on data visualisation, and want to focus on using Excel, and using it to produce something professional and consistent now, then I would recommend the following book by Steven Few. One of the few ‘how-to’ book that would also look good on a coffee table:

* Few, Steven (2012) **Show Me the Numbers: Designing Tables and Graphs to Enlighten.** Second Edition, Analytics Press: Burlington

Other books by Steven Few include:

* Few, Steven (2006) **Information Dashboard Design: The Effective Visual Communication of Data.** O’Reilly: Sebastopol, California
* Few, Steven (2009) **Now you see it: Simple Visualisation Techniques for Quantitative Analysis.** Analytics Press: Burlington

If you would like to learn more about Edward Tufte’s data visualisation design principles and philosophy, then his books are a pleasure to read and view. The first and most famous of which is:

* Tufte, Edward (2001) **The Visual Display of Quantitative Information** Second Edition, Graphics Press: Cheshire, Conneticut

Two books from the ‘infographics’ end of the data visualisation spectrum, which have arguably done much to bring the idea of data visualisation to public consciousness, while also often misrepresenting the field in ways that cause more technically minded data visualisers to wince, are:

* McCandless, David (2012) **Information is Beautiful.** New Edition, HarperCollins: London (More so)
* Rogers, Simon (2013) **Facts are Sacred.** Faber & Faber: London (Less so)

(Declaration: I own both books.)

A more nuanced and balanced book from the world of data journalism is the following:

* Cairo, Alberto (2013) **The Functional Art: An introduction to information graphics and visualization**. New Riders: Berkeley, California

If you only want to buy two books, I would recommend buying Cairo’s after buying Few’s! Whereas Few’s book will help produce sturdy, clean, simple, honest graphics for helping organisations understand themselves better, Cairo’s book is an engaging and wide ranging tour popular data visualisation and infographics. A take-home message from the book is that the general public can probably cope with more information rich visualisations than they’re usually provided with.

If you are feeling brave, and want to learn about how to use statistical programming languages such as Javascript, Python and R to produce data visualisations, then I would recommend the following books by Nathan Yau, a US-based statistician who runs the website flowingdata.com

* Yau, N (2011) **Visualize This: The FlowingData Guide to Design, Visualization and Statistics.** Wiley: Indianapolis, Indiana
* Yau, N (2013) **Data Points: Visualization That Means Something.** John Wiley & Sons: Indianapolis, Indiana

Although Yau’s books are fairly technical, and producing visualisations as he does requires a lot of specialist knowledge in programming, his graphics also demonstrate a high level of care and attention to issues of aesthetics and graphic design. As an effective, all-round data visualiser, there are very few competitors to Yau.

The UK-based data visualisation expert Andy Kirk has written the following book:

* Kirk, Andy (2012) **Data Visualization: a successful design process**. Pakt Publishing: Birmingham

It is a book of two halves: the first (shorter) half discusses the process of arriving at the correct design choices for a graphical visualisation; the second half is a reference guide describing and providing a wide range of example of different types of visualisation, and suggestions about the applications for which they would be best suited. I see it as a ‘conceptual’ practical guide: something to help the Scientist and Artist work well together in drafting something suitable for the Engineer to produce, rather than something that directly makes the Engineering involved any easier.

If you would like an even more comprehensive reference of data visualisations, then I would recommend the following:

* Harris, R (1999) **Information Graphics: A Comprehensive Illustrated Reference.** Oxford University Press: New York

Much more a catalogue than a coffee table book. Nonetheless, a good book to skim through if you don’t want to rely on Excel’s default options to spark your imagination.

Finally, if you would like to learn to use a statistical package, ggplot2, which explicitly uses the grammar of graphics paradigm described by Leland Wilkinson, I would recommend:

* Chang, Winston (2013) **R Graphics Cookbook**. O’Reilly: Sebastopol
* Wickham, Hadley (2009) **ggplot2** Springer: London
* Wilkinson, Leland (2005) **The Grammar of Graphics.** Second Edition, Springer, Chicago

However, anyone taking this dangerous path would have to learn R, the statistical programming language which ggplot2 is built on, as well. Two good introductions are:

* Kabacoff, Robert (2011) **R in Action.** Manning: New York
* Field, A., Miles J & Field, Z. (2012) **Discovering Statistics using R.** Sage: London

## Software

It’s frustrating, never quite right, and has bad taste. It’s also ubiquitous, strangely compelling and often surprisingly versatile. Excel seems here to stay. It’s the technical (as opposed to conceptual) focus of this course, and even for someone like me who tends not to finish data visualisations using this package, it’s often where I start.

If you want to stray away from Excel, it’s going to cost you. The question is: are you more willing to pay in time, or in money? The answer to this question will lead your data visualisation journey down one of two very different paths:

* ***“I want to pay in time”***: Then firstly consider learning many of the more technical features of Excel, such as VBA – the underlying programming language – pivot tables, connecting with databases, using macros and so on. This will allow you to increase the flexibility and functionality of Excel for data management and visualisation a lot. Unfortunately, there can be diminishing returns as you push further past the limits of Excel. If you have the time and inclination, and are willing to suffer the constant frustrations involved, then the dedicated statistical programming language R may allow you to travel further in the longer term. R is free in terms of money, but definitely not in time and effort. A range of books have been listed above, but if you want to dip your toes in the language for a few minutes, then the Codeschool website offers a free, gently paced, and surprisingly addictive, nautically themed introduction to the language, available from the following link: <http://tryr.codeschool.com/>
* ***“I want to pay in money”***: Of course, time is money, so even what on the face of it looks like a lot of money for a dedicated and user friendly data visualisation program might turn out cheaper than spending weeks, months, or years learning something like R. If you want something that’s less limited than Excel, but less scary than R, then Tableau is the main option on the table. Tableau is a commercial data visualisation program which makes it relatively easy to produce attractive publication ready visualisations, and to explore data visually using a largely intuitive interface. Conceptually, try to imagine the pivot table and pivot chart options in Excel, but *much, much better*, with better handling both of data management and visual design. Tableau comes in both a public version, and a commercial version. The difference between the versions is not in the functionality, but in the storage of the data. In the public version, data has to uploaded to Tableau’s website in order to be used. In the commercial version, you can keep the data to yourself. The cost of secrecy is about £1,000. The public version of Tableau can help you decide whether, for your organisation, this is a price worth paying: <http://www.tableausoftware.com/>

# References

Codeschool website

Few books

Tufte Article

Tufte Books

Shaw & Dorling

Cairo

Tableau Website

Muybridge

Wickham

The Economist

McKee