Visual Encoding, Data and Tasks

Jamie Holohan - 13325757

Part 1

Visual Encoding Channels

Position – Position as an encoding channel for data visualisation is probably the most useful. In almost any example of a visualised data set it is clear to see how the position of relevant data is selected and how other channels usually add further information on top of it. If we consider all the basic characteristics of data channels it can be shown that position has a very broad scope with regards to effective use of these characteristics.

For example, if there were no differences between each individual datum in a set being shown in some visualised model, other than their position in the rendering, we are clearly able to distinguish each point, we can easily associate points based on their position when compared to other points nearby and it is very simple to quantify the numbers of points which are likely to be categorised. Furthermore the range of the data being displayed can be quite large as one could essentially fill out an entire screen or display surface. This does not require a tremendous amount of creativity or artificial variety and as such it is very useful to those designing a visualised date set and those attempting to interpret that visualisation.

While position of data is a very useful tool in tandem with other encoding channel, by itself it has its limitations. As data sets grow larger it can often become difficult to fit each datum onto one surface. Furthermore, the position of each datum is not necessarily reflective of the intensity or significance of a certain piece of data. If each datum was only distinguishable by its location on a visualisation then users have no way of understanding which data maybe be of interest and which maybe be of less importance

Shape – Shape as an encoding channel can be tremendously useful in some circumstances, and useless or even detrimental in others. Unlike position which is applicable in almost any data visualisation, shape is most useful in displaying data where a user can easily associate that shape with an object or data set. If the data set in question related to different sports then simple icons that reflect the nature of that sport (i.e. a soccer ball for football, or a racket for tennis) would be quite natural for human interpretation. But if shape was used to indicate the size of a datum in comparison to others then it can become tricky to be used effectively. Shape does not allow for any simple ordering either and in this way, it is best suited to nominal datatypes as opposed to ordinal kinds. That being said, the range of shapes that could be used in a visualisation model is potentially infinite. Provided the data being displayed can be easily associated with a specific shape and is not an ordinal data set it is fair to say that shape can be an immensely useful encoding channel

Size – As a visual encoding channel, size is one that can be seen most commonly in visualised data sets, and for good reason. It is very simple to see if one mark or shape is larger than another and as such it drags a lot of attention to data that is generally the most significant. In terms of characteristics of an effective encoding channel for data visualisation, size can be extremely useful in making data selective and distinct from one another, it is extremely simple to group data based on its size and it is a visual way of quantifying a datum or data set. Ordering is also very simple. The only short coming of using size as a visual encoding channel is that it is limited by the screen space used to display the data. For data sets that vary greatly in size, large outliers can make smaller data too small to be effectively utilised. The range is limited by our display system

Brightness – The intensity of the shades being used to display data in a set is similar to size in many respects, but it is not limited by the real estate of space provided in the display area. Brightness can be useful in dragging attention to significant data and is very good at distinguishing different data types. It can make it very simple to categorise data which are of a similar intensity and in that same way it can be useful for quantifying similar data types. Just like size however, brightness has a very limited range and when we get into particularly large data sets with major outliers it can become difficult to provide unique values to different data types without the use of some other encoding channel. It in most useful in ordinal data sets as it can be very easily ordered very quickly by users.

Colour – As an encoding channel, colour is probably one of the most useful. After positioning, colour is probably the most widely used channel for distinguishing different data types. For particularly large data sets, individual colours cannot be assigned to each different value as the range of colours available to us is not that large. Most shades of colours are indistinguishable from other shades very similar to them in a visual sense and the human eye is not acutely able to separate them. For groups of data however colour is probably the most effective tool for someone attempting to visualise that data. Sharp changes in colour can be some of the most eye catching and attractive aspects of natural life and that translates very effectively when visualising data. Regions of the same colour can be very easily associated, while regions of differing colours are extremely easy to distinguish. Colour is not useful for ordering data when we don't consider the intensity of the colour, more so the hue and saturation. In that same way it is not useful for quantifying data. But if we were to only use position and colour as means of encoding a data set then there is a huge variety of ways to distinguish data and quantify and organise it in a very simple and easy to understand way.

Texture – When using texture as a visual encoding channel, it is very useful to ensure the texture(s) being used can be easily associated with the specific type of data it is trying to convey without any sort of key or indicator. If the user of a visualisation model can predict which texture applies to a particular type of data without any sort of specific key to make the associations then texture is highly useful. However if the textures being used are to abstracted from the data they are conveying then a variety in texture could very easily distract the user from the important information. Texture is probably best used in coordination with colour and intensity as it can provide variety within subsets of data without radically changing how they are perceived to be grouped by the user.

Orientation – Where data has a specific direction, its orientation in a visualised model can be the most effective way to display it. In any other circumstance the orientation of a data point can be hard to justify. The range of orientations available to a designer is small and finding a way of effectively mapping those orientations to the variety of data seems futile when we consider all the other channels at our disposal. This is not to say that orientation has no use as a visual encoding channel but unless there is a comprehensive way of mapping the direction of the orientation to some directional characteristic of the data being displayed, it is hard to justify using it as a way of distinguishing different types of data

Motion – This type of visual encoding is most effective for dynamic datasets, or data which changes or has changed by a significant amount over time. The human eye is very sensitive to movement and as such it it can be the most effective way to catch the attention of a user, but over use can be overwhelming or distracting. As such, movement must be used with care when attempting to visualise a data set, but when used correctly it can bring life to a visual model that cannot be achieved by any other encoding channel

Recurring Types of Datasets

Tables – These usually appear in two dimensional formats but extra dimensions can be used to provide depth to a table. This type of dataset is most useful when given a list of objects, each with the same attributes. It is not the most visually comprehensive method of storing data but it is possibly the easiest to index. It is a highly organised dataset and as such can be very quick for a human to search through but more importantly, computers will have no issue quickly sorting or searching through a data set if it as sorted this way and that is why it is the basis for almost every modern database. Tables are not efficiently at all for grouping data however. Other methods would need to be employed to group the data effectively but once this grouping has been achieved, tables are usually the most efficient way of storing that data.

Networks/Trees - This type of data set is an excellent compromise between human usability and machine navigation efficiency. It uses a hierarchal structuring type to group data. This ability to group datasets makes it useful for human consumption as it keeps relevant data close together but also provides a very clear structure for which data can be derived from others. It also for sub sets of sub sets of data sets which is how we as humans like to categorise information available to us.

In terms of machine navigation, hierarchy is a stable in how knowledge and data is stored. By being able to quickly navigate down through each level and narrow down the specific data to be used in certain tasks, machines can operate very efficiently with the data stored.

Fields – For human consumption and interpretation, using fields or spatially distinguished data types is paramount, for machine usability however it is almost pointless. A person can very quickly distinguish data on a canvass based on its position but trying to instruct a machine to measure the position of an object and then relate that to some attribute is computationally expensive and unnecessary. As such when one decides how one would wish to store or display our data it is important to consider how the data will be used and whom it will be used by.

Geometric – Geometrically distinguishable data sets like fields are intrinsically spatially defined and as such are highly effective for human consumption but are essentially useless for machine interpretation. Geometry is a diverse category and different geometric characteristics can be used in tandem to display a variety of attributes of a data set. Lines and curves can be used for connecting and correlating while shapes and sizes can be used for distinguishing and categorising.

For a machine, trying to measure and distinguish data based on its geometry is tedious and expensive. Then linking those measurements to data in some sort of organised and useful fashion is arduous.

Useful methods of storing data for computer consumption include sets, which are simple unordered groups of data, lists which are like sets except are stored in a specific order, and clusters which are just groups of sets or lists. This type of data can be static, i.e. non- changing and available at the one time, or dynamic, constantly changing and needing to be retrieved and displayed with regards to the time it is produced.

Recurring Types of Tasks

The types of tasks performed using visualised data can be divided into two categories, tasks where we try to consume the data and tasks where we intend to produce something from a data set

Consuming:

Discovery – This is about portraying data in a way that is easy to understand and interpret and as such, to deduce knowledge and prove or disprove our assumptions or hypotheses. This type of task would be performed if there is a large set of data available but the way in which the data is stored makes it almost impossible to gain any insight or meaning from it. Visualising it in a way that makes it easy for a person to distinguish, associate and correlate the data is most useful when performing these kinds of tasks

Presentation – This type of task is less about interpreting the data and more about effectively displaying what has already been interpreted. Once a hypothesis or assumption has been proven or atleast evidence has been found in favour of that assumption, it is useful to display the information in a way that clearly suggests that the assumption is true to someone who has not studied the data. As such this type of task is usually the result of other tasks being completed.

Entertainment – Similar to presentation, this type of task is not about return results that can be derived from data in existence but more so to give it purpose for human consumption. These types of tasks are about presenting information in a way that lure people into interacting or consuming it as is. Displaying data in a way that awakens the users curiosity is the objective here.

Producing:

Annotation – Essentially tagging data, by producing an annotated output it can be very useful as an input to another function or visualisation model. This type of task is about categorising existing data so that it can be easily sorted for use in further tasks

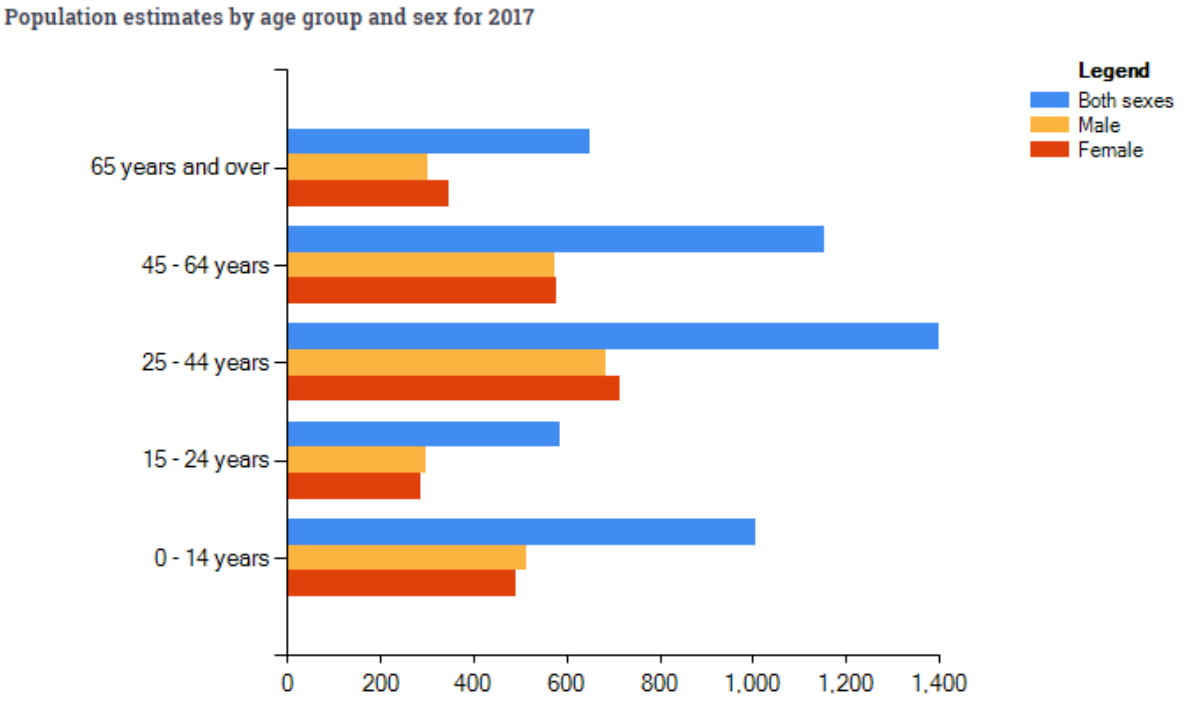
Recording – This type of task is about capturing data as persistent artefact for use in later tasks. How the data is recorded can effect how accurate it is or how easy it is to use. Often computers require data to be recorded in a very specific format. Badly recorded data can be tedious to sort through, well recorded data can reduce the stress and effort in subsequent tasks relying on the information provided.

Derivation – This is about deducing results from the information and data recorded. Data alone is not entirely useful, it is how one uses the data that gives it its value. These types of tasks usually provide an output to be used as an input for some other task.

Part 2

The first data set and corresponding visual model I would like to discuss is data gathered and presented by the Central Statistics Office (CSO) in Ireland (i). The data provided is of the size of the population citizens in Ireland within specified age ranges (i.e. 0-14 years old, 15-24, etc..). The data is visualised using a simple horizontal bar chart.

	Age Group	2016	2017
Both sexes			
	0 - 14 years	1,005.5	1,007.0
	15 - 24 years	574.7	584.8
	25 - 44 years	1,401.5	1,398.1
	45 - 64 years	1,128.0	1,152.7
	65 years and over	629.8	649.9
	All ages	4,739.6	4,792.5
Male			
	0 - 14 years	514.2	514.6
	15 - 24 years	292.8	297.9
	25 - 44 years	685.4	683.4
	45 - 64 years	561.0	572.7
	65 years and over	293.2	303.4
	All ages	2,346.5	2,372.1
Female			
	0 - 14 years	491.3	492.4
	15 - 24 years	281.9	286.8
	25 - 44 years	716.1	714.7
	45 - 64 years	567.0	580.0
	65 years and over	336.6	346.4
	All ages	2,393.1	2,420.4



The data is represented in thousands, meaning everyone is not being represented, instead the groups are rounded to the nearest 100, or 1 decimal place. The groups are not clustered based equal differences between the oldest and the youngest individual, but rather on what appears to be a social interpretation of which ages are most likely to be associated. The first cluster being 0-14-year olds (14 years), this incorporates preschool, primary school and more junior cert students in the Irish school system. The next cluster, 15-24-year olds (9 years) incorporates leaving cert students, secondary school graduates and college students. These clusters seem to indicate groups that society would consider as young children, and young adults. The third cluster which incorporates 25-44 years olds (19 years) is probably reflective of young to middle aged professionals, the people most likely to be building homes and families. The fourth cluster containing 45-64-year olds (19 years) are older professionals who have finished having children and are preparing for retirement. While the final cluster is the largest being all of those above 65 years old, that being the legal age of retirement in Ireland.

This simplified version of the population of Ireland by age is an indicator of the value and the intended use of the data. By understanding the sizes of the populations within these groups at this stage we can accurately predict the sizes of those groups in 10/20/50 years' time. This is useful in understanding the kind of services and infrastructure that the country will require to meet the demands of these groups. The types of tasks that this data would aid would be productive, this information is essentially an input for decision making with regards to government policy and spending. Large amounts of young people will require more jobs while large amounts of older people would require more healthcare. That's obviously a very simplified derivation but it is this kind of information that can be deduced and extracted from this data set.

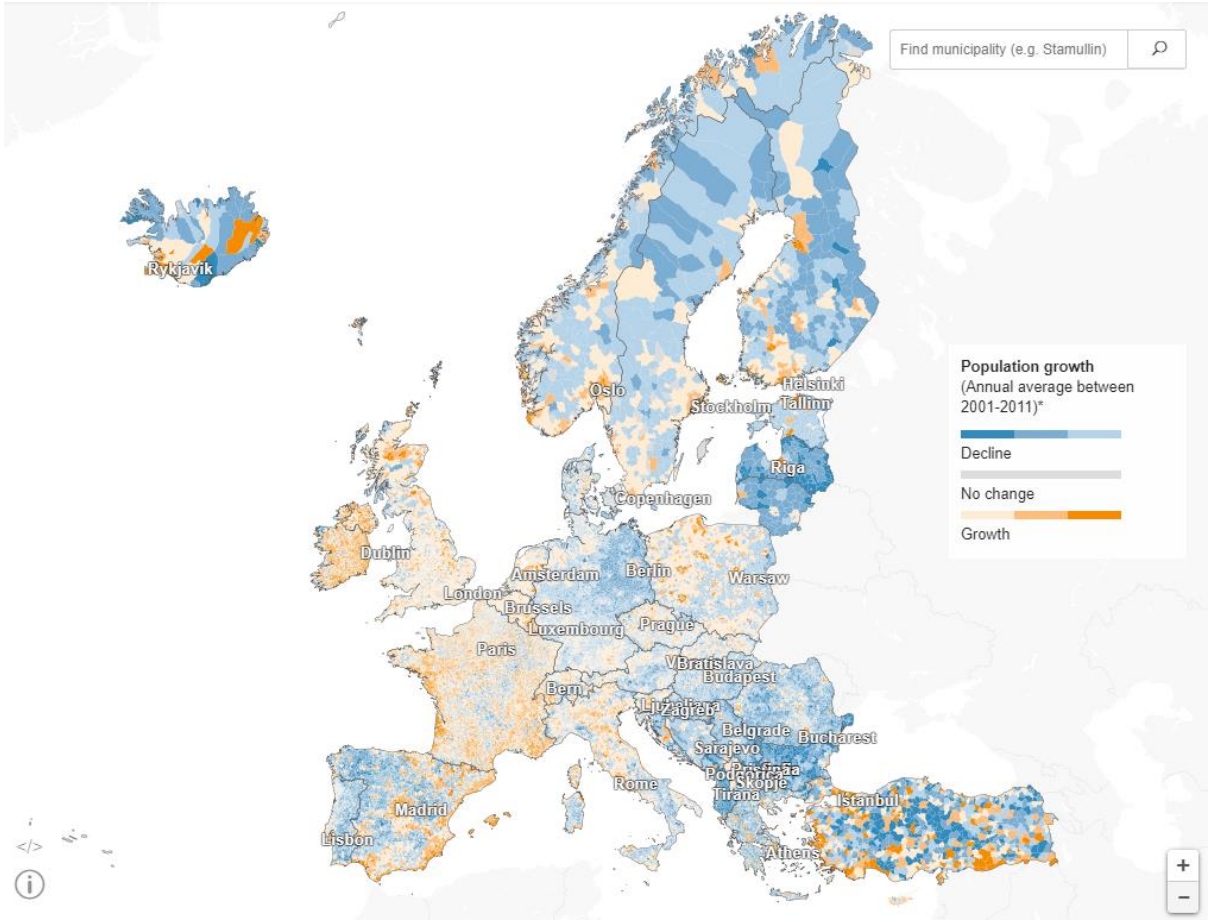
In terms of visual encoding channels, space and colour have been used very effectively to distinguish and cluster the data very clearly. Varying the sizes but maintaining the shapes works effectively for comparing cluster sizes. The use of the vertical scale to separate the age clusters in tandem with the 3 colours assigned to the sexes is clear and concise. Visually it is very easy to compare the populations by sex as they are adjacent to their counterparts of the same age. The use of bright primary colours is attractive and easy to distinguish. It is very easy to focus our eyes on the colour of interest.

The horizontal scale is assigned to the size of the populations. Once again, a very simple strategy but very effective, using essentially one plane of vision and by focusing our eyes on the colour of interest, usually blue in this case, we can rapidly compare the size of a population cluster.

While I believe this is an effective portrayal of the information made available I also believe this visualisation ha sits shortcomings. For a start, there seems to be no structure to the clusters other than social interpretations and as such I believe the data can be misleading. The group aged 15-24 contains only half the range of ages of most other clusters, as such it is no surprise it is only half the size. We must take this into account as we interpret the model and I don't believe that's an effective way to portray information. I would like to see the clusters made more fairly to give a more accurate representation of the breakdowns of the ages. I also think this diagram has a lot of wasted space. It effectively portrays the population of both sexes together and separately, but I feel this information could have been displayed on different charts. Or possibly could have been overlapped using different textures to fit more information onto the screen.

If I were to visualise the data I would break it up into groups of 10 years (0-9, 10-19, 20-29 etc..) and I would place the y-axis in the centre rather than the left. I would extend bars to the left and right, assigning one side for each sex (i.e. bars extending left are female and bars extending right are male). In this way the vertical axis would still cluster each group by age, but it would be easier to fit more age groups onto the canvas. I would still use colour to distinguish the bars by sex and the bottom axis would still be the scale for the size of the populations.

The second data set I am going to discuss is an interactive map of Europe compiled by the Berliner Morgenpost. The map is like a heat map of population growth and decline around the EU between 2001 – 2011. It uses colour and intensity to display regions of growth and decline. It is simple to zoom in and out of specific regions to gain a more accurate insight of the data in that zone. Hovering the cursor over an area provides details of that area.



Space and position are used to express geographic location while colour is used to represent the direction of change in population over time. Intensity is used to portray how significant the change in population has been in that area.

It is very easy to distinguish and categorise regions of population growth or decline both from a zoomed-out perspective and a very close up perspective of each individual country. Every country is broken down into 1000's of smaller areas, all of which are categorised by larger town or city names. The use of complementary colours blue and yellow allows for an attractive appearance while maintaining a simple visual feature for separating areas that have experienced different population shifts.

Space and position is based purely on geographic location. It is almost impossible to display the surface of the earth on a rectangular canvas due to the curvature of the earth but it seems that the position of the borders and regions is an accurate reflection of what they would look like in real life.

The types of tasks that can be performed using this information are similar to the previous example. Understanding where populations are growing and where populations are shrinking in the last 10 years can help us make assumptions about where similar behaviour will occur in the next 10 years. Knowing where we will need to build infrastructure and provide services is an important long-term investment. Similarly knowing which areas are seeing declines in population is valuable information. It can indicate areas that are not receiving the intention they require, or highlight changes in people's lifestyles, like moving from rural areas in urban areas. Understanding not only the size of a population but how dense that population is can be a major input for tasks and decision making that will affect the future of the European Union.

References:

(i)

Article title: Home - CSO - Central Statistics Office
Website title: Cso.ie
URL: http://www.cso.ie/en/index.html

(ii)

Article title: Where the population of Europe is growing – and where it's declining
Website title: interaktiv.morgenpost.de
URL: https://interaktiv.morgenpost.de/europakarte/#4/56.00/11.43/en