Visual Analytics

Ryan G

March 11, 2020

Contents

(Wk 1) Introduction to Data Visualisation

Tutorial

- 01. Tutorial Sheet

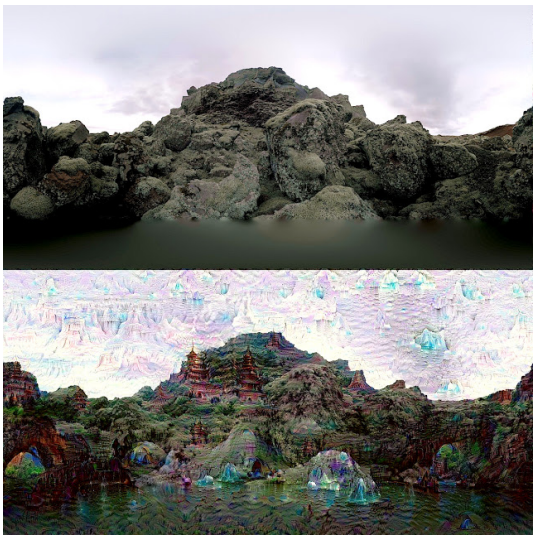
Using *Zathura*

When using *Zathura* the copied text will be placed in the primary * register not the + register, so paste it in with `" * p`.

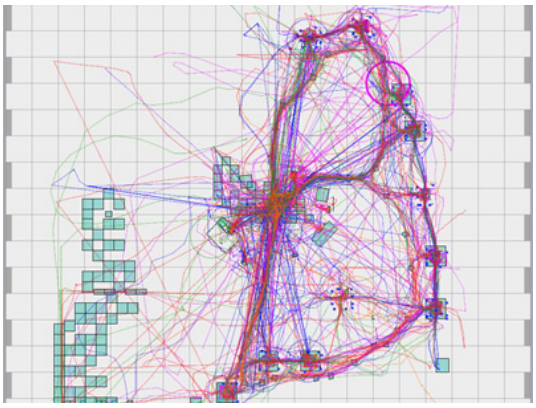
Question 1

1. Problem Visit the website: http://www.visualcomplexity.com/vc/. Have a close look at the available visualisation techniques. In your opinion, which techniques are among the most useful? Or which one is the most pretty visual display? Explain and justify your preference. Note: there is no right or wrong answer in this question.

2. Working This Visual from the *Google AI* blog highlighting the important features detected by a neural network is a unique insight into the *predictive modelling* technique. It is not easy to understand the behaviour of a Neural Network and this visual offers a uniqe insight that could not easily be understood:



This Visual however is arguably one of the most visually striking because of the vibrant colour choice:



Question 2

1. Problem Explore the online demo: http://graphs.gapminder.org/world/. What have you discovered or found from the visualization of the "Wealth and Health of Nations data set"? E.g. is there any correlation between GDP and Life Expectancy? etc.

2. Working There is a clear positive correlation between income and life expectancy, other features that are readily observable are:

   (a) Regions It can be observed that Europe is a wealthy region (Yellow) while Africa and Asia are poor regions (cyan, red))

   This could be further observed to be a function of distance from the equator, generally regions lose to the equator like Europe, UK, USA are quite wealthy where as regions nearer the equator are more likely to be impovrished.

   Exceptions to this are:

   - Australia due to the rich resources such as coal/uranium and the head start in education owing to the European Descent.

- Russia is impovrished because the October Revolutions that followed the first

World War greatly modified there economy relative to the rest of Europe, this decayed into authoritarianism and eventually extreme wealth inequality (with the rise of homelessness and oligarchs) following the collapse of the Soviet Union at the end of the cold-war.

This could be a functoin of temperature, a higher temperature may lead to :

- increased spread of bacterial disease
- increased perishability of food

(b) Population It isn't possible to say much about the influence population has because the population is measured by region without taking account of size (e.g. China is quite large while Vietnam is relatively small) or habitability (e.g. Japan is very mountainous and hard to farm or ranch on).

This is made more difficult because the income is measured per person but the population is not the relative space that person might enjoy (i.e. the population density), the GDP per country rather than capita might be more instructive.

(c) Time Trends

It appears that Europe (and Japan) have always enjoyed a higher GDP per capita while asia and Africa appear to lag in this respect.

Other than the "reshuffling" caused by the Great Depression, WW I, WW II and the collapse of the soviet union Europe has always enjoyed a higher GDP hence a higher life expectancy.

During the first two world wars (with the exception of Russia and Poland respectively, for obvious reasons) it appears that the influence of currency on life expectancy became even stronger, somewhat unfortunate for other regions given that these were European wars. The great depression also showed a similar effect on this correlation, going from a linear trend to an exponential.

(d) China as an Exception

China appears to not follow a correlation between GDP and Life Expectancy until the new milennia, life expectancy rises following the second world war despite no change in GDP per Capita.

In the 70's China begins to follow this trend moving in a diagonal fashion indicative of a correlation, this is presumably due to manufacturing exported to China perhaps due to the development of the micro-processor in the US.

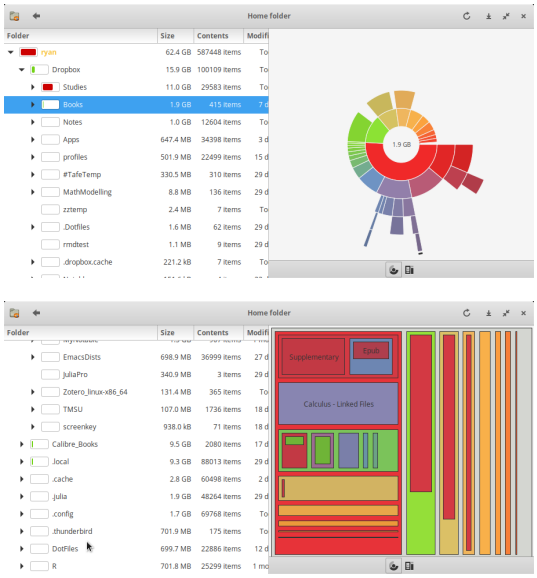
In the 2000's China began to improve GDP per capita more significantly, perhaps due to free market policies, and followed a trend where GDP per capita would strongly correlated with life expectancy.

Question 3

1. Problem Using a search engine, explore 3 applications/projects/tools that use visualisations. You are required to write a short summary about the applications/projects/tools. Why do you think they are significant? What are good and bad about these applications/projects/tools? Etc.

2. Working

   (a) Gnome's Baobab / WinDirStat / KDirStat Gnome' *Baobab*, *KDirStat* and *WinDirStat* are disk usage analysers for Gnome, KDE and *Windows* respectively.



i. Significance All three tools use a descending list of directories with a bar chart indicating the proportion of disk space consumed by that directory, to the right is a graphic shosing this distribution.

ii. Good Design All three tools have correctly implement an ordered list and bar chart, making it easy to understand and manage disk space on a system.
Gnome's Baobab uses a ring chart with a popup overlay to describe the corresponding directory.
This choice of graph makes it easy to understand which directories are consuming the most space while still having an overview of the structure of the directories, the popup prevents the graph from becoming too busy and showing unnecessary information.
The ring chart will also re-centre following a selection of a directory allowing deeply nested structures to be understood easily.

iii. Bad Design Unfourtunately KDirstat and Windirstat only offer treemap visualisations, this choice of graph is vastly inferior to a ring chart because it can only clearly show a certain amount of information at an overview, it is difficult do understand deeply nested folder structures they won't re-generate the plot without rescanning the drive.

(b) GitHub Visualizer http://ghv.artzub.com/#repo=ranger-assets&climit=100&user=ranger
The Github Visualiser is a way of visualising the proportionate activity of various repositiories of various projects.

i. Significance This provides a novel way to understand the:
   - relative popularity of repos
   - The language most composing a repo / project
   - How Active a project is generally and over time.

ii. Good Design Using Bubbles to illustrate the is the overall share of a repo in a project provides a quick understanding at a glance.
Using a Time Series Chart over time is an easy way to show trends.

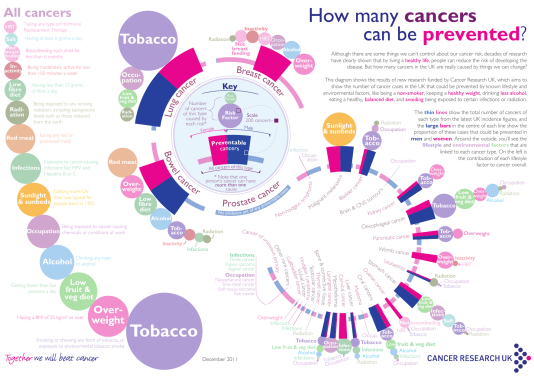
iii. Bad Design The visualisation is far too busy to understand what is going on, for instance the size of the bubbles are not made clear, are they popularity, size, frequency of commits or frequency of clones / pulls of the repo?

For instance the visualisatoin of one of my favourite pieces of software Ranger is such that I can infer nothing from it other than the fact that the web page is written in HTML and the project is written in python, which is totally obvious and not particularly helpful to get a deeper understanding of the *Ranger* project as opposed to say the Midnight Commander Project, although the visualisation for Midnight Commander is significantly better and so this may be a scaling issue.

(c) Cancer Graphic http://web.archive.org/web/20140526084103/http://www.cancerresearchuk.org:80/cancer-info/cancerstats/causes/attributable-risk/visualisation/

This Plot attached here shows the attributes most likely to cause cancer:

```
cd /tmp
#wget "http://web.archive.org/web/20140801035734/http://publications.cancerresearchuk.org/downloads/product/CS_POSTER_ATTRIB.pdf"
command -v pdftoppm >/dev/null 2>&1 || { echo >&2 "command -v foo >/dev/null 2>&1 || { echo >&2 "I require pdftoppm but its not installed.  Aborting."; exit 1; }I require foo but it's not installed.  Aborting."; exit 1; }
pdftoppm CS_POSTER_ATTRIB.pdf  CS_POSTER_ATTRIB -png
mv CS_POSTER*png ~/Notes/Org/Attachments/Statistics/
ls ~/Notes/Org/Attachments/Statistics/CS_POSTER_ATT*png
```



i. Significance This visualisation is significant because it effectively describes both the magnitude and interaction of various risk factors and behaviours on the probability of developing cancer.
Statistically this can be a very difficult thing to describe and explain but this graphic very clearly shows what to look out or

ii. Good Design Having a key in the centre of the graph makes it very easy to determine what the individual elements of the visualisation mean.
The relative size of individual risk factors means that at a galnce it is very easy to determine risk factors for cancer.

iii. Bad Design The graph unfourtunately is a little busy but this doesn't appear to be in a way that is disproportionate to the amount of information conveyed.

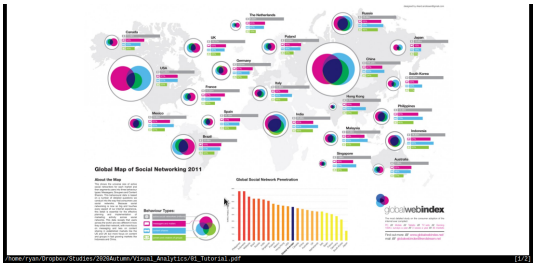
iv. Inferences An interesting component of this visualisation is that it clearly shows interactions of various items, for example the following are large to moderate risk factors for cancer generaly:

- Overweight (BMI $>$ 25 kg / m^2, distninct from obesity which is \$$\approx >$ 30)
- Inactivity

While HRT is a very small risk factor for all types of cancer.
This clearly illustrates that HRT may be medically appropriate in men despite the historical belive that such treamtent was correlated with prostate cancer in men [**bell2018**] . Recent studies suggest that this might not be the case [**loeb2017**] and HRT is (somewhat obviously) correlated with increased physical activity, weight loss [**traish2014**] and general health metrics [**saad2020**] .
This visualisation clearly shows that this treatment might generally lower the risk of cancer in men and provides a simple way to convey complex interactions between attributes and the complex statistics involved in this type of research in a way that clearly shows the important facts of the matter..
Moreover this plot also only shows HRT as a risk factor in cancers that tend to overwhelmingly affect women, a distinction that is very important in that area of medical science.

Question 4

1. Problem For the following visualisations, in your opinion, are they good or bad; Justify the answer:
2. Working
   (a) Plot A



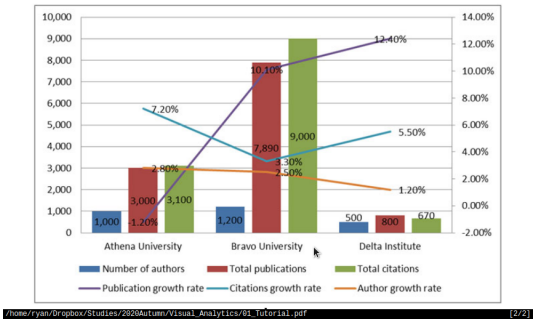
This plot is overly busy but also not very descriptive, meaning at a glance it is not possible to determine what the differences between different regions actually are.
Upon closer inspection it is not clear what the bar charts are trying to illustrate, clearly the venn diagram is trying to illustrate the proportion of communication and overlap, relative to thhe number of users. This is an interesting way to try and show the interactions between the different communication strategies but the choice of a Venn diagram was misguided. A venn diagram can only represent 3 sets with circles [1] and this has seemingly artificially restricted the number of potential behavriou types that the plot has tried to convey to such an extent that they may as well be disregarded entirely because they are just not descriptive enough to understand.

[1] combinatorics - Why can a Venn diagram for $4+$ sets not be constructed using circles? - Mathematics Stack Exchange

Interestingly Eastern countries use a combination of content/messges while western countries tend to use or the other, this could be related to the pictographical written laguage implemented by Eastern / Asian Countries and could lead to insights on human behaviour, unfourtunately the plot does not clearly describe the meaning of the colours and so no inferences can be easily made.

Moreover the size of the circles are relative to the total number of users, this is somewhat misleading because the plot shows that 33% of the USA uses Social Media while only 15% of China uses social media, countries/regions are not uniformally distributed or constructed so choosing to use regions as a delimiter when using absolute size of users is potentially misleading. This is partially addressed however by the *Global Social Network penetration* bar chart below the plot.

(b) Plot B



- The amount of text overlayed on this graph makes it difficult to read.
- The Growth rate of the publication should not be represented as a line because

it indicates some type of continuous connection between the universities, it should be represented as either another column or ideally a seperate time series plot should be produced:

- Such a plot should show the market share and allow the growth rate to be

interpreted from the slope of the line in an organic fashion, that way the different universities could have both market share and growth rate compared between each other without trying to mentally visualise a cumullative summation.

Question 5

Visualisation of the CoronaVirus:

- http://rocs.hu-berlin.de/viz/sgb/

1. Problem Optional: If you still have time, play around this website to see how visualization help to find patterns: http://rocs.hu-berlin.de/viz/sgb/ (Coronavirus Geographic and Network visualisations)
2. Working

   :HideRef:

References

../../Dropbox/Studies/Papers/references