Graphics with a cause

Robert Grant

Almost every day new exciting graphics appear online, conveying some message in a powerful and engaging way. Many have an interactive element, allowing the viewer to choose the details they want to see. This has become a boom industry in a short space of time and, for most statisticians and data analysts, it seems another world. Few of us know how to translate our work into animations and interactive graphics. We were simply never taught those skills, and perhaps we acquired in our education and work experience a certain disdain for style over substance. I contend we are right to be sceptical about exciting graphics, but wrong not to engage with them. Whenever we have a worthwhile subject of enquiry and a message that cries out to be communicated to a wider audience, is it not our duty to communicate it as effectively as possible?

I have been following this new breed of graphics, and the people who make them, for a few years now. Particularly in the case of the online interactive ones, they are perhaps better described as data visualisation. At first I sought to make some innovative images of my own using my familiar data analysis software, but gradually I have come to understand that there is also a lot to be gained from learning about the tools used to put visualisations online and make them interactive. In fact, I would go so far as to say that spending a train journey from London to Manchester and back with an introductory book on the web programming language JavaScript is probably the most profitable few hours I have ever invested in learning new skills. I would urge all Radstats readers to consider investing a few hours like this. More about the web later - what does this have to offer the radical statistician with a progressive cause?

Innovation in graphics has often come from those who had a progressive cause and were looking to make their message as clear as possible to politicians and the public. Florence Nightingale's rose diagrams showing the death toll from disease among soldiers in the Crimean war (figure 1) [1], Charles Booth's maps of poverty in London [2] and Friedrich Engels's maps of cramped and unsanitary living conditions in Manchester (figure 2) [3,4] are all fine examples from yesteryear of visualisations intended to convey a powerful message to readers without the confidence of numeracy or the patience to read a lot of text and numbers. It is rare to see as much passionate effort made in the name of profit or war [5], although some excellent work is

carried out in 'industry', where there is a need to communicate quickly and effectively to company directors [6].

Figure 2: Florence Nightingale on causes of death in the Crimean War

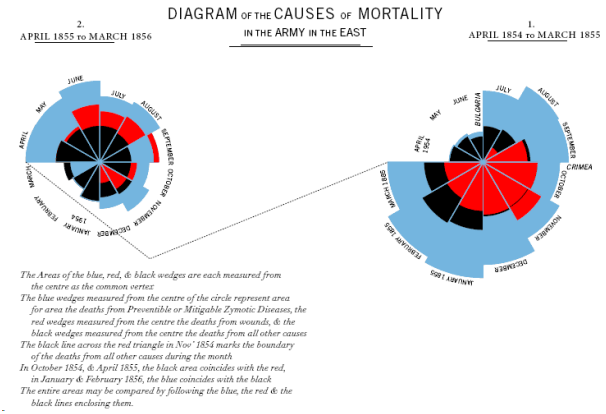
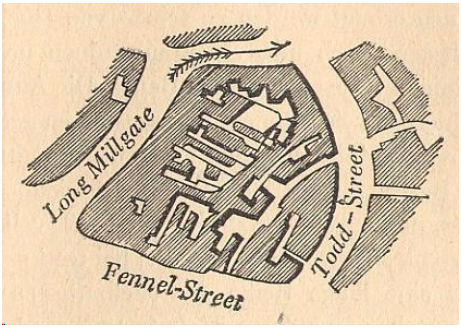


Figure 3: Friedrich Engels on poor housing in Manchester



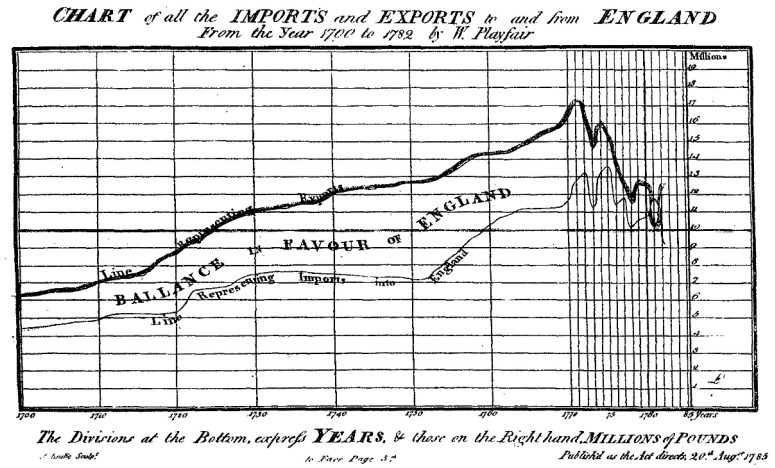
Lessons from history

The history of graphics holds some valuable lessons for contemporary data analysts. The early days, from the mid 18th century, were characterised by experiment and innovation in how results were communicated to a broader audience. At first, the role of the statistician (one who studies the state) was to compile extensive tables of facts and figures which would then be consulted by others with

more exciting roles: politicians, scientists, scholars and well-heeled dilettanti; in fact in post-revolution France, the Bureau of Statistics was split by heated debate over whether it was appropriate for them even to summarise data at all [7]. Graphics were absent, yet, as Leland Wilkinson among others has pointed out, a well-constructed table is a kind of graphic [8]. Proximities in row and column variables become Euclidean distances on the page (correctly or not). As numbers get magnitudes of scale larger, they look bigger and put more ink on the page. Although the stem-and-leaf plot is somewhat unfashionable now, it has an elegant simplicity that utilises this, and of course in the days before data analysis software, could be created by anyone with a typewriter, or indeed a pencil. The ease of production is a theme I will return to throughout this article.

Among the early experimenters, William Playfair's name looms large. By the early years of the 19th century, he was crunching economic data and producing line graphs, bar charts, pie charts and some other hybrids that have not stood the test of time quite so well. They were hand-drawn and intended to convey an overall pattern, not precise values. Axes were not always straight, and he was not inclined to let a lack of hard data get in the way of a good story (figure 3), for example, drawing amazingly detailed curves for fluctuations in the economic productivity of the Babylonian empire [9].

Figure 4: William Playfair on the faltering English economy

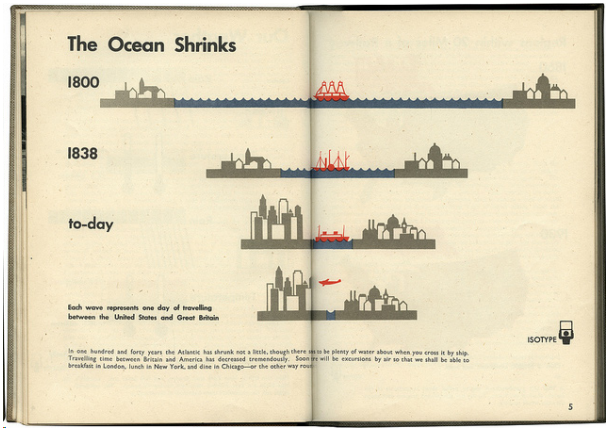


Maps were also fertile ground for experimentation; Charles Booth's poverty map of London and John Snow's cholera map are notable for

the impact they had on policy. The concepts they wanted to communicate were far clearer in a map than they could ever be in text, and clarity remains a good rule of thumb for choosing between verbal, tabular and graphical tools for communication (although it is still widely ignored). Small maps covering a small area, such as Engels's alleyways and courtyards crowded around Long Millgate, could be drawn by the author, but larger, more complex efforts required a skilled draftsman. Similarly, Playfair's hand-drawn maps gave way gradually to glossier, professionally produced graphics.

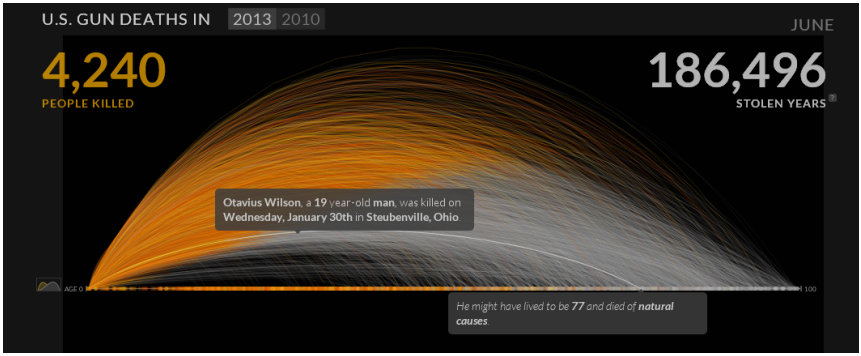
However, in the main, scientific publication was text-based. Perhaps the expense of employing a draftsman to make a diagram or map of sufficient quality that it could be used for printing was prohibitive, and perhaps a classical education inculcated dextrous wordplay. As a result of these tribal habits, it was not until 1841 that the first graph appeared in the *Statistical Journal* [10]. When pictograms and their many variants rose to popularity through the first half of the 20th century (figure 4), they were largely ignored by statisticians, who were on their way to establishing a toolbox of approved graphs: the line, the bar, the scatter, the box, all remain cornerstones of basic statistical education to this day and feature heavily in many subjects at school, despite Mr Gove's efforts to replace an understanding of data and evidence with an understanding of times tables (and quite possibly cold showers and the cane too).

Figure 5: Otto & Marie Neurath of the Isotype Institute on transport speed, image courtesy of Michael Stoll



Unfortunately, with an accepted canon of graphs, data analysts (a term I use to capture the many who analyse but are not qualified statisticians) lost the ability or confidence to innovate. It is only in recent years that a resurgence of interest in communicating statistics through the new medium of the internet has brought data analysts and graphic designers back into contact with each other. The results have at times valued style and novelty above substance and clarity, but there have been some genuine steps forward too. The latest trends, if they are at all discernible among the blooming of the proverbial thousand flowers, include real-time updating and 'scraping' of data from large sources such as Twitter (for example, *Flu Detector* [11]), adding individual stories to the data when you hover your mouse over a point in the graphic (for example, *US Gun Deaths* [12]), and all manner of network graphs (for example, *Health InfoScape* [13]).

Figure 5: US Gun Deaths



The overall trend, then, has been one of happy early experimentation, followed by a prolonged division between statistics and design. Now at the start of the 21st century, rapid changes are taking place and that division is being closed by a growing number of new experimenters. If I had to choose the lessons of history, they would be:

1. You have to communicate as well as calculate.
2. You can change the world with a pencil; don't be fooled into spending money.
3. Don't be content with the forms of communication you were taught - learn additional ones.

What is behind the rise of visualisation?

To my mind, the single most important factor driving this has been the democratizing effect of open-source software. It is now possible to download free software for almost anything, look at the details of the program to see how it works, amend it or write new programs that link to other software, and share the result instantly online. Of particular note in data analysis, R (www.r-project.org) has been built from a large number of packages written by enthusiasts, who can be faster to adopt new methods and to fix bugs than any commercial software house. The amount of e-mail, blog and online forum traffic about R vastly outstrips the commercial alternatives, reflecting the rapid growth of new packages and applications [14]. Python and Julia are other open-source programming languages amenable to demanding scientific calculations, which are being developed and extended all the time. For graphics, packages such as Gnuplot and Inkscape extend the capabilities of the number-crunching software, while there is as much appetite for maps as in Charles Booth's day through the collaborative OpenStreetMap and Geocommons.

The rise of open-source has brought the means of production (if I may appropriate a radical phrase of some pedigree) of engaging graphics back into the hands of those conducting research. While early graphics were made with pencils, the arrival of the professional draftsman and then the commercially produced software restricted graphics to those with a decent budget, and the scientific community aided this by requiring graphics of a certain 'standard' for publication. This is no longer the case, as anyone can make excellent high-quality graphics. Instead of investing money, you might have to invest some time in learning a new skill.

Often, the effort to develop new software is collaborative and hosted online, with the website GitHub providing the most lively community. Why would anyone work hard on something and give it away for free? I suspect readers of Radstats can answer this question quite easily, but there is another layer of incentive beyond altruistic dedication to the open source ethic, in that talented young developers can publish their work online from their homes anywhere in the world and catch the eye of a recruiter at a major software company.

Another way in which software and data are openly shared is through having web-based graphics and data analysis. Whenever you see an interactive chart or map online, there is a good chance it has been programmed using the web language JavaScript. This sends the data, and the instructions to turn them into the graphics, to your computer,

where it is put together by your web browser. This makes for far less strain on the web server but it also has the effect that someone who knows a little JavaScript can look at the code behind a visualisation that they like, adopt and amend it and use it again in a new way. At the time of writing, there has been a run on falling-object visualizations, with *US Gun Deaths* [12] inspiring *Out of Sight, Out of Mind: every drone attack in Pakistan since 2004* [15], which in turn inspired *Bolides: visualizing meteorites* [16]. An increasing number of web pages are constructed in this way; JavaScript is taking over in popularity from animations made in proprietary software such as Flash, where the data and the construction of the visualisation are hidden from the viewer.

Google is an interesting hybrid, producing a dizzying selection of maps and graphical products (for example, interactive graphs can very easily be produced from the cloud-based spreadsheet in Google Docs), all aimed at user-friendliness and simplicity, free but commercially funded through advertising. Those who would appropriate aspects of Google's outputs for their own open-source products often find they are stymied by constantly changing complex web addresses and JavaScript so tortuous it is hard to conclude anything other than deliberate obfuscation. Nevertheless, there are some very elegant solutions linking R to Google, such as the googleVis package, that handle the complexities.

There is perhaps another powerful motivation for people making web-based visualisations: the fun of experimentation and problem-solving. Not everyone making these visualisations is doing so for fame or fortune. In fact, much of the cutting-edge work is (if their blogs and social media posts are to be believed) done late at night, hunched over a computer, trying to get something to work that has been bugging them all day. To my mind, this spirit of experimentation is the real sea change taking place. Amanda Cox, one of the New York Times's leading data graphics contributors, said in a recent interview that:

> "I come from a statistics background, and I'm finding statistics students' portfolios are crazy weak compared to the computer science students, even though they're playing with the same problems. I think it's because comp sci students are encouraged to play, whereas stats majors it's, 'here's your rule book, now make things.' I don't think that's the good model for making better visualization" [17]

Although the best universities produce statisticians who are willing to experiment with new approaches and software, most do not. Cox's description rang true for me, but I experienced some culture shock in

conversation with Don Rubin at Harvard University recently, who was in no doubt that his students were not like that and experimented, taught themselves new skills, and made new solutions and tools if one did not exist. Perhaps all statistics educators should aim for that culture of experimentation, not rule book.

As a final discordant note in this theme of ever-increasing openness and accessibility, it is not a one-way direction of travel, as shown by Microsoft's adoption of a 'Windows Store' where users of Windows 8 and Windows RT can buy programs - as long as they are pre-approved by Microsoft. In a way, this is no more than coming into line with Apple and Google, but it is a long way from the heady days of MS-DOS powered shareware. For those who remember sending off a £1 postal order for a floppy disk full of programs of varying quality, written by enthusiasts, the call of a simpler era is not just nostalgia. A large proportion of web-based visualisers and programmers go fully open-source and run the free Linux operating system, dispensing with commercial software entirely.

What's next?

Not surprisingly, the best places to follow new data visualisations are online: blogs such as flowingdata.com, eagereyes.org, infosthetics.com and seeingcomplexity.wordpress.com specialise in visualisation, while andrewgelman.com and www.theatlanticcities.com sometimes feature insightful comments on the subject. Some newspapers have a regular data slot, principally the New York Times and the Guardian. The home pages of JavaScript libraries such as d3js.org, leafletjs.com and raphaeljs.com contain inspiring galleries of visualisations.

A recent paper by Gelman and Unwin in the Journal of Computational and Graphical Statistics, with responses from some prominent visualisers, and rejoinder, raised some interesting debates about the purpose of good visualisation [18]. The conclusion the authors settled on was that one fixed image is not enough; some people want little detail and others lots. An interactive display allows the viewer to choose for themselves, and means they can be guided into the layers of complexity by a storytelling approach. That certainly is the approach taken by one of my all-time favourites, a comparison of predicted and real balance of the USA budget, from Amanda Cox at the New York Times [19]. Story-telling was subsequently explored in more detail by Robert Kosara, one of the responders to Gelman and Unwin [20]. Taking viewers through the visualisation step-by-step with

narration / annotation before letting them experiment for themselves is notably the approach of Hans Rosling's famous Gapminder bubble plot talks, and it appears in some other very popular visualisations such as *US Gun Deaths*. I believe this will become a standard part of all good visualisations in the near future. Those with scant explanation are invariably confusing as a result; consider for example *MoneyBombs* [21].

The flurry of activity at the interface between web designers and data analysts suggests to me that we will also see an increasing uptake not only of art by data people, but also of data by art people. Data art has been a fringe activity but there are some notable examples, and interestingly it has often been linked to a socially progressive purpose. Explore the real-time social media text summary of *We Feel Fine* and you will soon find a heart of darkness [22]. In 2001, Damien Hirst announced plans to collaborate with the environmental charity Future Forests and create an installation of 441 carbon dioxide gas canisters, each 6 feet tall, representing the 15 tons of CO_2 he produced in a year [23]. Although the installation never came to fruition, it doesn't really matter; the concept is the art and lives on. This is clearly data art to me as its intention is to represent data (15 tons) in sculpture, it is essentially the wonderful video by CarbonVisuals writ large [24].

Figure 6: We Feel Fine

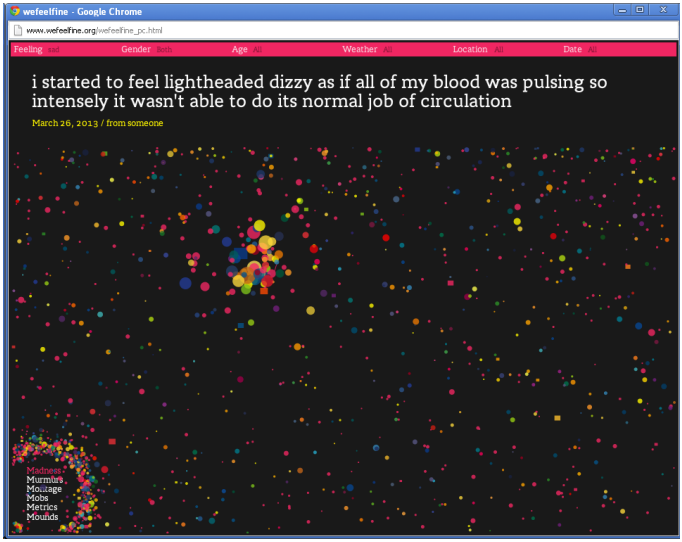
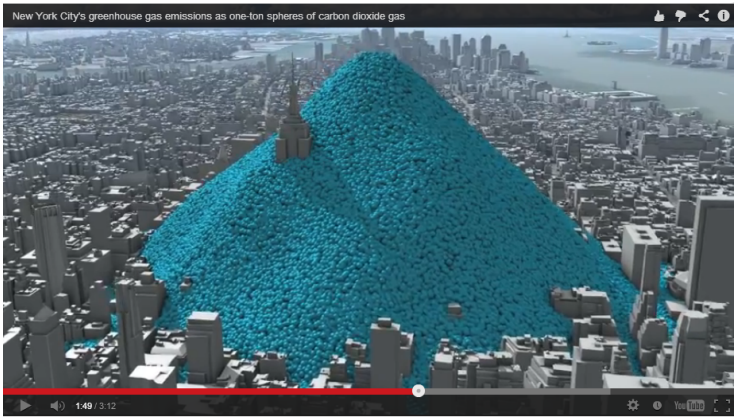


Figure 6: CarbonVisuals on New York's carbon dioxide emissions for one day



There have been a few tentative forays into representing data as sound. This is in its infancy, and whether it will prove useful will not be known for some time. There is not even agreement on a name, with 'sonification' and 'audiblization' both in use. I see this as being at the stage of growth that Playfair's graphs were at 230 years ago. It may yet provide another boom of activity to follow on from the interactive visualisations of today.

How can we spread these skills to get greater impact for our causes?

At present, we have a silent majority of data analysts who do not critique (constructively) data visualisations, perhaps because the purpose and techniques are opaque to them. To get wider use of these visualisations, and hence I believe, better understanding of our data among the public and policy-makers, we will all need to devote some time to experimenting and learning new skills in this area. In this way, we can all produce them without the need for a specific budget or a helpful IT / communications person down the corridor. In particular, planning from the outset to have an online presentation of research findings in parallel to scientific routes such as the peer-reviewed journal can only be a good thing. The end-users of data are far more likely to engage with these outputs but it is not something that can be squeezed into occasional spare moments, and that focus on

communication needs to be acknowledged as a valuable contribution by funding bodies and employers.

There is no doubt in my mind that the next five years will see the arrival of a really good user-friendly graphical user interface for R, which is the hub connecting all the software I have discussed. At present there are some packages that allow you to point-and-click through the menus, while they do the unpleasant programming in the background, but none has won over a broad enough section of data analysts to challenge the commercial software. Any programmer who achieves this will have made a name for themselves, and several are working towards this goal. When they get there, the barrier stopping most of us from taking up more open-source software - namely, that we find the programming impenetrable and unpleasant - will have been overcome.

In the meantime, there can be no doubt that the popularity of online visualisations will encourage more people and organisations to contribute their own efforts. Hans Rosling's *200 Countries, 200 Years* talk with the animated bubble plot has accumulated 5.8 million views on YouTube at the time of writing, after two years online. CarbonVisuals's video of New York's carbon dioxide emissions has had over 260,000 visitors in 8 months, and US Gun Deaths has had over 500,000 views, and has been online for just over 4 months. Those of us still producing black-and-white boxplots and bar charts can only dream of that sort of impact.

References

1. Understanding Uncertainty (2008) Nightingale's "Coxcombs". http://understandinguncertainty.org/coxcombs.

2. Booth C (1899) Booth Poverty Map & Modern map (Charles Booth Online Archive). http://booth.lse.ac.uk/cgi-bin/do.pl?sub=view_booth_and_barth&args=531000,180400,6,large,5.

3. Engels F (1891) The Condition of the Working-Class in England in 1844. http://www.gutenberg.org/ebooks/17306.

4. Pfalzgraf F (2002) Manchester: The Old Town. http://home.arcor.de/friedrichengels/old_town.htm. Accessed 30 May 2013.

5. Yau N (2012) Netanyahu knows his diagrams. http://flowingdata.com/2012/09/28/netanyahu-knows-his-diagrams/. Accessed 30 May 2013.

6. Jones W, Spence M (2011) GWSDAT Ground Water Spatio Temporal Analysis Tool. http://web.warwick.ac.uk/statsdept/user2011/TalkSlides/Contributed/17Aug_1705_FocusV_1-Hydrology_1-Jones.pdf.

7. Desrosières A (2002) The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press.

8. Wilkinson L (2005) The Grammar of Graphics. New York: Springer-Verlag.

9. Alonso J (2011) A short visual history of charts and graphs. http://seeingcomplexity.wordpress.com/2011/02/03/a-short-visual-history-of-charts-and-graphs/. Accessed 5 June 2013.

10. Aldrich J (2011) Graphs before graphics: visualisation and presentation in Victorian statistics. http://mcs.open.ac.uk/su379/VIPS/Aldrich.pdf.

11. Bristol University (2013) GeoPatterns - Flu Detector - Tracking Epidemics on Twitter. http://geopatterns.enm.bris.ac.uk/epidemics/.

12. Periscopic (2013) U.S. Gun Deaths in 2013. http://guns.periscopic.com/?year=2013.

13. GE Data Visualization (2011) Health InfoScape. http://visualization.geblogs.com/visualization/network/. Accessed 30 May 2013.

14. Muenchen RA (2013) The Popularity of Data Analysis Software. http://r4stats.com/articles/popularity/. Accessed 30 May 2013.

15. Pitch Interactive (2013) Out of Sight, Out of Mind: A visualization of drone strikes in Pakistan since 2004. http://drones.pitchinteractive.com/.

16. Zapponi C (2013) BOLIDES - Visualizing meteorites. http://www.bolid.es/.

17. Berinato S (2013) The Power of Visualization's "Aha!" Moments. Harvard Business Review. http://blogs.hbr.org/hbr/hbreditors/2013/03/power_of_visualizations_aha_moment.html. Accessed 4 June 2013.

18. Kosara R (2012) All Responses to Gelman and Unwin in One Convenient Posting. http://eagereyes.org/blog/2012/responses-gelman-unwin-convenient-posting.

19. Cox A (2010) Budget Forecasts, Compared With Reality. New York Times. http://www.nytimes.com/interactive/2010/02/02/us/politics/20100201-budget-porcupine-graphic.html. Accessed 11 December 2012.

20. Kosara R (2013) Paper: Storytelling, The Next Step for Visualization. http://eagereyes.org/papers/paper-storytelling-step-visualization. Accessed 4 June 2013.

21. VisPolitics (2012) MoneyBombs. http://www.vispolitics.com/project/moneybombs/. Accessed 5 June 2013.

22. Harris J, Kamvar S (2006) We Feel Fine. http://www.wefeelfine.org/.

23. BBC (2001) Cleaner dumps Hirst installation [sic]. ttp://news.bbc.co.uk/1/hi/entertainment/1608322.stm. Accessed 4 June 2013.

24. CarbonVisuals (2012) New York City's greenhouse gas emissions as one-ton spheres of carbon dioxide gas. http://www.youtube.com/watch?v=DtqSIplGXOA.

Robert Grant, Senior Research Fellow in Quantitative Methods, Center for Health & Social Care Research, St George's, University of London & Kingston University

Email: robert.grant@sgul.kingston.ac.uk