

THE SEMANTIC WEB DEMANDS ‘NEW’ STATISTICS

Jim Ridgway, James Nicholson, and Sean McCusker
SMART Centre, Durham University, Leazes Road, Durham, DH1 1TA, UK
jim.ridgway@durham.ac.uk

Abstract

The concept of statistical literacy needs to be refreshed, regularly. Major changes in the ways that data can be accessed from government and non-government agencies (the ‘semantic web’) allow everyone to access huge databases, to create new variables, and to explore new relationships. New ways of visualising data provide further challenges and opportunities. The Open Data movement, and the rise of data driven journalism are increasing public access to large scale data via the media. Here, we map out the potential and pitfalls of the semantic web, and discuss the rebalancing of statistics curricula that is required. The most obvious challenge is the need to introduce students to the exploration and analysis of large scale multivariate data sets. We present examples of our visualizations of complex multivariate data, and describe some examples of use in classrooms. General issues of pedagogy and curriculum innovation are discussed.

ON STATISTICAL LITERACY

A number of authors offer generic descriptions of statistical literacy (SL) that are likely to be robust over time. Kathrine Wallman (1993) in her Presidential Address to ASA offers *the ability to understand and critically evaluate statistical results that permeate our daily lives... to appreciate the contribution that statistical thinking can make in public and private, professional and personal decisions...*

Other authors such as Gal (2002) and Schield (2010) have also made valuable contributions by mapping out some important components of SL that include:

- knowledge about data sources and the ability to evaluate the quality of evidence
- understanding concepts used to describe society (e.g. inflation, unemployment, GDP, poverty, GINI, ‘happiness’)
- ability to understand and draw conclusions from a variety of representations – tables, graphs and other visualisations
- knowledge about research methods (such as the relative merits of surveys, experiments, and meta-analyses).

However, of necessity, all of these papers offer just partial insights, as ‘number cultures’ evolve. The lacunae in this list are associated with recent developments in communication, driven by technology. In order to discuss changes to statistics curricula, we need to describe the ways that the ‘number culture’ has changed, and continues to evolve.

RESPONDING TO A CHANGING ‘NUMBER CULTURE’

Important developments in the number culture include: the open data movement; the graphical explosion; the rise of data driven journalism; and citizens’ increasing use of the web to improve their personal effectiveness. We consider these briefly, then focus attention on the implications of the semantic web for curriculum and pedagogy.

The Open Data movement

There is a philosophical tradition dating back as least as far as Condorcet (1792) which asserts that an important part of education is to inform citizens about governance, and evidence about the state of society, so that they will be aware of injustices and structural social inequalities. Condorcet believed in *savoir libérateur* – knowledge that would enable people to free themselves from social oppression. More recently, initiatives such as *data.gov* in the USA and *data.gov.uk* explicitly state political objectives associated with empowering democratic processes.

The graphical explosion

A great many ways have been developed to present data graphically since Playfair’s (1786) invention of statistical graphics, often for political purposes (e.g. the Neuraths’ work in the period

1930-45 designing graphical displays to demonstrate social inequalities (see Neurath, 2010)). Major data providers are providing powerful visualisations in the hope of making their data more accessible (e.g. OECD's data is accessible via *gapminder* and *eXplorer*). Key political targets such as the UN Millennium Development Goals are presented in the form of an interactive dashboard, to encourage public engagement.

Outlandish claims have been made about the power of pretty graphics to engage users with data, but almost no research (outside the SMART Centre) has been conducted into ways that multivariate data can be understood by people who are statistically naive.

Data driven journalism

A highly significant cultural trend is the emergence of 'data driven journalism' (e.g. Bradshaw, 2010; Brooke, 2010; Rogers, 2011; Gray, Bounegru, and Chambers, 2012). Print and TV media, increasingly, are providing interactive graphics on their websites to support their news reports. Media such as the BBC, Guardian, Washington Post, Stern, Economist and New York Times have excellent repositories of data embedded in interesting displays, and promote the use of high quality data in their reports. Good examples and resources include: Matthew Ericson's descriptions of the design processes at the New York Times <http://www.ericson.net/content/>; The Guardian's *datablog* <http://www.guardian.co.uk/news/datablog>; and Matt Stiles (NPR) *dailyviz*, that presents interesting data visualizations, and invites ideas for future visualisations <http://thedailyviz.com/>

Citizens' increasing use of the internet

Increased use for all manner of purposes is self evident.

THE SEMANTIC WEB: THE GOOD, THE NAÏVE, AND THE MENDACIOUS

The 'semantic web' (sometimes referred to as Web 3.0) is one of the most exciting and important developments in ICT in recent times. A major problem for anyone who wants to access data on social issues is that relevant data can be found in (literally) thousands of data bases, spread over (literally) hundreds of websites. The core ambition underpinning the semantic web is the provision of a portal which allows users to work actively on huge, disparate sources of information relevant to a common theme. Users can locate data that is available on distinct databases, via simple keyword searches. Portals often provide suggestions for related themes and links to other sources of data. Users can then conduct analyses, and can re-present data in new ways. They can create mash-ups (assemblies of data, graphics and text) on their own websites, where (for example) data displays are revised automatically, whenever the source data is updated. *ProgrammableWeb* provides links to over 6000 mashups.

An excellent example of a semantic web is provided by *data.gov* - a resource created at the behest of the US government, that is designed both to provide easy access to government data (from about 400,000 raw and geospatial data sets), and to make it easy to use and intelligible. Other examples include: *data.gov.uk*; *UNdata*; and OECD's *Statsportal*.

Many websites, however, present difficulties even for sophisticated users. For example, Simon Rogers, editor of the Guardian newspaper's *datablog* and *datastore*, offered these comments (Rogers, 2011) about the website of the UK's Office for National Statistics (ONS) 'It should be pointed out that the ONS has incredible info on this site – but it is also the world's worst website'. It is impossible on many websites to extract multivariate data that extends beyond three variables (e.g. ONS, Eurostat). So the Open Data movement has still got a great deal of work to do, simply to make data from single websites accessible to sophisticated (as opposed to unsophisticated) users.

An exciting feature of the semantic web is the provision of Standard Application Programming Interfaces (APIs). *ProgrammableWeb* gives access to over 3,500 APIs that users can use to display data – examples include 'heat maps' to display spatial data. Users can also perform calculations on data, and can present it in novel ways. APIs have been created that allow data to be displayed and distributed on a number of platforms – including mobile phones. It is easy to embed mashups in *Facebook*, *YouTube*, and other social networking sites.

The semantic web allows users to explore radical new conceptions about associations and causality between variables that were very hard to explore before. Data relevant to both practice and theory can appear on different websites, and can now be synthesised and explored. Social scientists from different disciplines and traditions can now explore phenomena in new ways, and can create multilevel theories at different grain sizes, and can incorporate variables in novel combinations.

The provision of powerful tools does not necessarily lead to empowered citizens. A range of conceptual and cognitive issues have to be overcome. The video used to introduce *data.gov* contains this assertion: *it is now as easy to analyse data as it is to buy a pair of shoes or a TV online* (<http://www.socrata.com/datagov/new-data-gov-platform-video-overview/> downloaded August 2011). It is certainly easy to create new variables, and to see how different values are distributed spatially. *Datamasher* is a site linked directly from *data.gov*, and provides several examples of ‘heat maps’ relevant to childhood obesity in the USA, state by state. A number of new variables have been created, using data from different government websites. These include plots of:

- Fast food restaurants PLUS adult obesity rate PLUS % overweight children
- % obese children DIVIDED by % obese adults
- Number of suicide deaths PLUS % obese children.

A fourth display purports to show “kids who are fat *because* their parents are fat” [italics added]. The variable displayed in the heat map is (State by State) the percentage of obese children multiplied by the adult obesity rate.

Such displays are evidence of the need for better SL in the community associated with open data.

The rise of data driven journalism will enhance the use of data in the media, and will exacerbate the problems of interpretation and misinterpretation. The extensive and increasing use of the internet to inform personal decisions will put similar demands on SL. We identify two sorts of challenge; first is to put more emphasis on components of SL related to data quality (conceived broadly); the second is to develop ways to help people with little formal training in statistics to understand multivariate data. Both of these need to be introduced in school and university curriculums.

DATA QUALITY AND THE SEMANTIC NET

A rose by any other name...

For the semantic web to work, there needs to be a way to describe variables that is common to all the databases that are accessed. As well as compatibility in terms of the ways that data are accessed, the deeper problem is the compatibility of metadata. Different agencies are likely to use identical terms (‘poverty’, ‘crime’, ‘educational attainment’ and the like) for key concepts, but to measure them in different ways. Ongoing work on the Resource Description Framework (RDF) addresses this problem directly – the RDF sets out to allow structured and semi-structured data from different sources to be combined (see WC3 Recommendations (2004)). From the viewpoint of SL, an awareness of the potential problems associated with the compatibility of metadata is probably sufficient. However, there is a pressing need for greater awareness of key issues around measurement.

Data accuracy

Challenging the source and reliability of data is important. Brennan, Watson, and Charles (2012) examined official UK Hospital Episode Statistics, and found, *inter alia*, in a one year period that there were over 17 000 male inpatient admissions to obstetric services, and nearly 8 000 male admissions to gynaecology. The authors remark ‘...although we applaud innovation, we suspect that the numbers may...reflect data errors.’

Astroturfing

The internet is not simply a repository of reliable (or unreliable) information. There are systematic attempts by some information providers to deceive and defraud others. Authors often pretend to be not what they are; examples include fraudsters claiming to be lotteries or

philanthropists, and companies who set out to create the impression of extensive web-based ‘grass-roots’ support for their products (so-called ‘astroturfing’). Inventing and selectively editing data has a long history, even in scientific circles.

The politics of measurement

Decisions about what to measure can be of great social and political importance (for example, the Istanbul Declaration (OECD, 2007), and the OECD Global Project are based on the idea that Gross Domestic Product (GDP), and other measures of economic performance, should not be the sole measure of progress within a country - see Stiglitz, Sen and Fitoussi (2009); the quality of life, sustainable development and the environment should also be considered. The International Monetary Fund appears to disagree: the Principal Global Indicators which they list at <http://www.principalglobalindicators.org/default.aspx> are exclusively economic and financial indicators.

Decisions about how to measure anything raise important scientific issues. Concepts and associated measures of variables such as speed, velocity, acceleration and temperature now seem self evident, and in no need of exploration or challenge. However, a great deal of theoretical and practical work was required to develop these ‘self-evident’ measures. In the social arena, there are a great many concepts with ‘self evident’ meanings, such as literacy, health, poverty, inequality, and crime. However, the definition of these concepts is problematic, and an awareness of these difficulties, and the importance of understanding something about the nature of measurement, and the importance of the choice of measures (at both a practical and political level) will be an increasingly important component of SL. A starting point can be found in the work of Swan and Ridgway (see Ridgway, Swan, and Burkhardt, 2001) who created resources for undergraduates entitled *Creating Measures*, where users invent, then critique, different measures of the same physical attribute (such as ‘squareness’ or ‘smoothness’). Critiques are designed to encourage the development of an aesthetic of measurement (what are good and poor properties of measures?).

Understanding representations

The large number of different sorts of visualizations that are being created means that an essential component of statistical literacy will be the ability to decode and work with novel representations. *Gapminder* is available as a free download, and a large number of data sets can be uploaded. *Open Statistics Explorer* is a generic platform that can be used (free of charge) by educational and research establishments. The SMART Centre provides a facility for users to use their own data sets to create interactive multidimensional displays. The Visualisation Unit within the UK Office for National Statistics makes its data displays available for use by non-profit organisations (see their dynamic population pyramid, for example <http://www.neighbourhood.statistics.gov.uk/HTMLDocs/dvc1/UKPyramid.html>). *Google* are increasing the number and quality of visual analytic tools they provide. The plethora of visualisations is likely to present considerable challenges for citizens (and educators), given the problems that users encounter when working with tables (e.g. Watson and Nathan, 2010), graphs (e.g. Swan and Philips, 1998) and box plots (e.g. delMas, Garfield, Ooms and Chance, 2007).

UNDERSTANDING MULTIVARIATE DATA

At school level (in the UK at least), the statistics curriculum can be seen to have its roots in the early 1900s. The pace of scientific discoveries has increased dramatically over the last 100 years – in the early 1900s, there was very little scientific knowledge compared to now. Key questions in agriculture posing binary oppositions (‘Is this fertiliser better than that one?’ ‘Which crops grow best here?’) provided vital information. The power of statistics was to make it possible to do small scale studies that could be generalised across a variety of settings. Consequently, the curriculum became dominated by important ideas such as sampling and hypothesis testing. The absence of computational power meant that heroic assumptions (such as linearity and Normality) had to be made, to make problems tractable.

In the UK, statistics teaching provides very little to support SL. The curriculum focuses on the mastery of statistical techniques suited to small sample, one and two variable problems, with very few illuminating examples where these techniques are or have been applied, and can be seen

to be useful. This comes at the expense of ignoring large scale multivariate data (See Ridgway, Nicholson and McCusker, 2007b). The result is that many of our most able and motivated statistics students are ill-prepared to understand data that are discussed increasingly in the media, and that can now be accessed via the semantic net. Given the difficulties that students experience with the current curriculum, it might seem eccentric to advocate the analysis of multivariate data. However, the need to build capacity here is urgent. We begin with a bald (but self evident) assertion.

Every interesting problem in health, crime, poverty, environment, education, and personal well being is multivariate, has non-linear relationships, and has confounding variables.

The logic of the analysis of large scale data sets is also rather different from the logic of drawing inferences from small samples that are applied to populations. In the case of census and some other data sets, data are available from the whole population of interest, and the inferences are often about the characteristics of subpopulations (is the association between poverty and attainment the same for boys and girls, across all ethnic groups?). Or, in the case of an individual looking for data on healthy lifestyles – *I'm a 40 year old-woman, with a blood pressure of 140/90 and high cholesterol. If I lose 20 pounds in weight, how much will this reduce my risk of a stroke/heart attack in the next 10 years?*

It follows that we might have some problems teaching about multivariate data for a number of reasons. Large scale complex data present different challenges. First, the data are primarily observational, and so conclusions about causality are tentative. Linear relationships are rare; data are multivariate, and interactions are common; assumptions about distributions can be wrong. At least as important, the sorts of question that can be asked are different to the ones asked before. A central issue often revolves around the wisdom of applying certain kinds of action to whole populations, where the impact on sub-populations is critically important. These considerations mean that the balance of knowledge and skills in the curriculum needs to be reweighted, away from an emphasis on linear models and significance testing, towards assessing effect sizes, establishing (non-linear) functional relationships and mapping interactions. Statistical ideas that require greater emphasis include: modeling functional relationships; confidence intervals; effect size; and Simpson's paradox.

Given the difficulties that students experience learning inferential statistics, these challenges might seem insurmountable. However, we have a variety of evidence to suggest that this is not the case. First, our studies (Ridgway, Nicholson, and McCusker, 2011) show that statistically naïve students (aged 14-16 years) working with large scale authentic data often have good insights into concepts such as effect size and interaction. In other studies (Ridgway, Nicholson, and McCusker, 2007a), we show that visualisations which pose questions based on multivariate data can actually be easier (in psychometric terms) than paper-based questions focused on one and two variables.

Our current work in classrooms (e.g. Nicholson, Ridgway and McCusker, 2011) with social science students engaged with authentic multivariate data on health and social inequalities also offers scope for optimism. This is described briefly, below, and some lessons for pedagogy are set out.

EXPLORING COMPLEX DATA IN SOCIAL SCIENCE CLASSROOMS

Educational Context

Statistics offers powerful tools to support the modeling and understanding of phenomena. Paradoxically, disciplines where there is often only modest teacher competence in statistics (such as the social sciences) require considerable statistical sophistication. We believe that the lack of engagement with quantitative aspects of evidence presents a barrier to understanding the phenomena under consideration, and, crucially, is a major contributing factor in inhibiting the development of skills in reasoning from evidence amongst those intending to work in social science areas (a further consideration is that many 'numerate' students do not maintain interest in the social science disciplines because they don't see any future in it for them. This represents a missed opportunity both for the students and for the social sciences community). Here, we describe our work in sociology and personal and social education.

Reasoning from evidence in sociology

The Nuffield Foundation has funded a project, *Reasoning from Evidence*, which is developing data visualisation tools to support the teaching of Sociology in courses for 16 – 18 year olds. In the UK, admission to university is determined by success on content-based examinations (GCSE) in a narrow range of subjects (for example, all university physics departments require students to have successfully completed courses in physics, and mathematics (at least) as a condition of entry). Students who want to take university courses in social sciences usually take preparatory courses in subjects such as sociology or psychology. A key theme in GCSE sociology is ‘inequalities in educational attainment’. We chose to use this as an example. The starting point for theory should be a clear understanding of phenomena. This poses a problem for teaching social sciences, particularly at the school level, for a number of reasons:

- every social phenomenon is influenced by a number of factors
- there is a wealth of data collected on a large scale by governmental and non-governmental agencies that could be used to describe phenomena and to provoke understandings around social phenomena, but these data have not been readily usable. Until recently, access to authentic data has been problematic; it can be hard to find; and it is not available in a form that is easy to use or interpret (e.g. presentations of large numbers of two-way tables are common). The semantic web has made access easier, but few teachers have the time or the knowledge to be able to access appropriate data sets, and to present them in class
- data in textbooks are (usually) out of date before they are used in class
- formal statistical analyses of multivariate situations are difficult to conduct and to understand. Even if the teacher understands the analysis (unlikely, given the background of most teachers), the students will not.

In class, the use of data has largely been restricted to headline statistics based on aggregated data, and explorations of the associations between single factors and some phenomenon of interest. We illustrate these problems in the context of educational attainment, and a solution, below.

Textbooks report well known phenomena:

- girls perform at least as well, and usually better, in every subject that is tested at the end of compulsory education
- pupils in lower socio-economic groups perform less well than students in higher-economic groups
- there are big differences in the performance of different ethnic groups.

Social phenomena often reveal interesting interactions, and so summary statements based on single factors, such as those above, can hide important aspects of the data. At the end of compulsory schooling (age 16 years) pupils take a number of subject-specific examinations (English, mathematics, geography, etc), that are set nationally. Performance is graded on a letter scale, and A* to C grades are viewed as ‘a good pass’. Schools are obliged to publish the examination results of their pupils, and league tables are created (some parents use these tables when deciding where to send their children). A commonly used measure of school success is the number of students who pass 5 or more subjects at grades A* to C, including English and mathematics. National data are available on student performance, disaggregated by ethnicity, sex, and eligibility for free school meals (FSM or NFSM) – a surrogate for relative poverty (see <http://www.education.gov.uk/rsgateway/DB/SFR/s000900/index.shtml>). The foreword to the 2010 Education White Paper by the Prime Minister and Deputy Prime Minister provide an example (see figure 1) which refers to these data; the authors do draw attention to interactions in the data.

- Children on free school meals do significantly worse than their peers at every stage of their education.
- They are just half as likely to get good GCSEs as the average.
- This vast gap between rich and poor is not pre-ordained. In Finland and Canada the gap is much smaller. Even in this country there are some groups – Chinese girls on free school meals for example – who significantly outperform the national average.

Figure 1: statements made in the foreword to the 2010 Education White Paper by the Prime Minister and Deputy Prime Minister.

The third bullet point either contradicts or qualifies the first, depending on your reading.

A report from the Equality and Human Rights Commission *How fair is Britain* (2010) offers the following description (see figure 2), and a graphical summary of data from high stakes examinations taken at the end of compulsory schooling (see figure 3).

Socio-economic groups

FSM eligibility reveals one of the starkest differences in achievement at age 16.

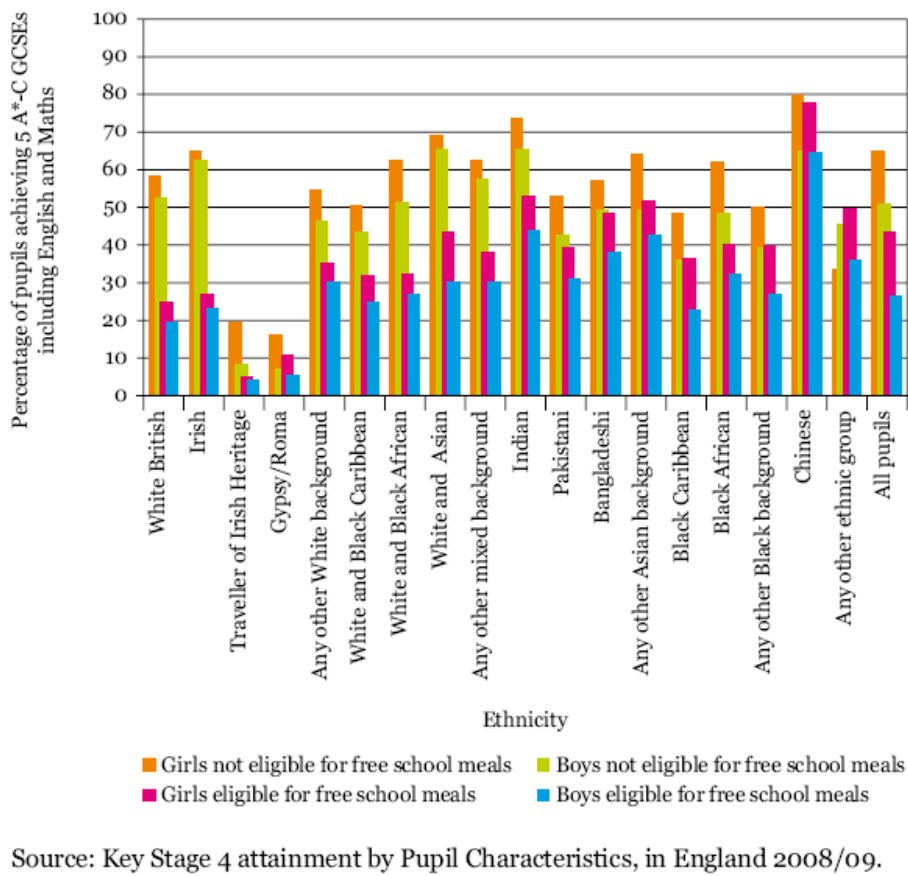
Being eligible for FSM is associated with a much lower probability of achieving 5+ good GCSEs including English and Maths. In 2009 only 27% of students eligible for FSM achieved this compared to 54% for those not eligible.

Figure 10.4.1 below shows that inequalities in England vary considerably within different ethnic groups when combined with eligibility for FSM. It reveals, for example, that while both Chinese and Indian students do very well overall, performance of the latter varies much more by socio-economic status. So whereas Indian boys from low income families perform significantly below the overall average, their Chinese counterparts are well above average.

When analysing how ethnicity, gender and FSM eligibility affect GCSE performance, the National Equality Panel found that the FSM effect was larger than any of the other associations shown, with the exception of the higher performance of Chinese pupils, and the lower performance of Gypsy and Traveller children.

Figure 2: text on education inequalities in How fair is Britain? (2010)

Figure 10.4.1 Percentage of pupils achieving 5 A*-C GCSEs including English and Maths 2009 by gender, FSM status and ethnic group in England, 2008-09⁷⁵



Source: Key Stage 4 attainment by Pupil Characteristics, in England 2008/09.

Figure 3: graph on education inequalities in *How Fair is Britain?* (2010)

To describe these data effectively, one needs to be sensitive to some big statistical ideas:

- There are huge differences in attainment between groups – pupils from traveller groups achieve about 5% passes; Chinese pupils achieve about 70% passes (*range, the foothills of effect size*)
- Within every group (ethnicity, school meals), girls outperform boys by about 5-10% (*simple main effects; the foothills of effect size*)
- Chinese pupils outperform all other ethnic groups – both when comparisons are made just between pupils who are eligible for free school meals or not, and just between pupils who are not eligible for free school meals, and when Chinese pupils on free school meals are compared with all other pupil groups (*interactions*)
- Eligibility for free school meals is associated with attainment in a dramatic way for the white population (FSM pupils achieve less than half as many passes as NFSM pupils; white pupils eligible for FSM have worse examination results than black children eligible for FSM), but hardly at all for the Chinese population (*interactions*).

Visualisation tools allow these issues to be explored by students, in class. An interactive display and guidance on how to use it, with classroom materials, can be found at www.dur.ac.uk/smart.centre/nuffield. A sample display is shown in figure 4.

The display shows the performance of girls and boys (and All pupils) for each of the main ethnic groups. Users can choose to compare all pupils in each ethnic group, or just those on FSM, or just NFSM pupils. Tabs at the top of the display provide more detailed data allowing the

exploration of differences between ethnic subgroups within each of the main ethnic groups. Variable names (*FSM?*; *ethnic group*; *sex*) can be dragged to different locations, to facilitate active data exploration.

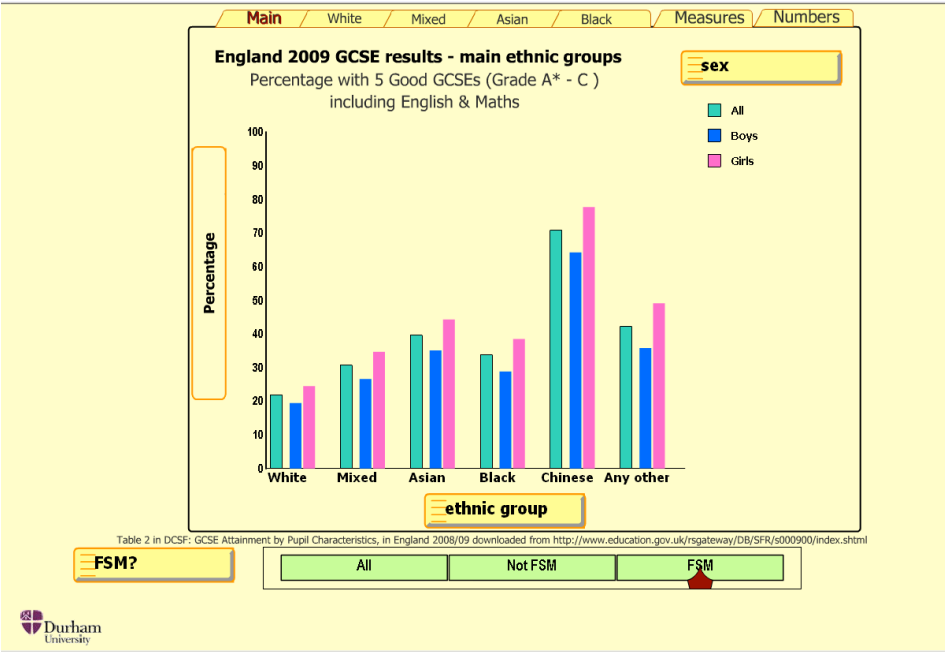
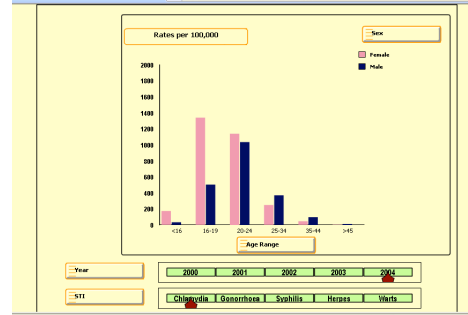
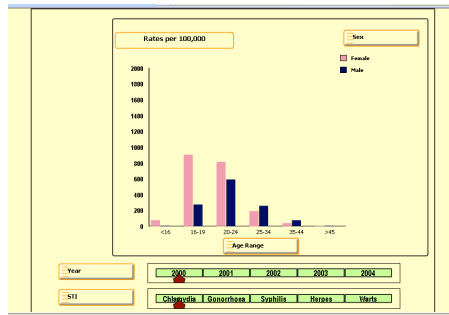


Figure 4: ethnic differences for pupils eligible for free school meals

The display offers the opportunity to let even young students get to grips (literally) with disaggregated data. By using drag and drop variable labels, they can manipulate the data to see which factors are associated with the biggest effects, and to describe patterns in the data. These patterns are often non-linear, and the patterns across values of one factor are sometimes different across values of another factor. Students can have the confidence to talk about the stories that the data tell, and can then judge the extent to which rival theories are consistent with the evidence. The relationships within the data in this interface are more complex than anything students have previously encountered, so building confidence in their ability to describe relationships is important.

Testing conjectures in personal and social education

If students generate conjectures about some aspect of the data, one test is whether the conjecture holds when the display is reordered. Classroom trials show that students find it easy to reorder displays and to explore conjectures. Figures 5 and 6 show data on the incidence of Sexually Transmitted Infections (STI) from all UK sexual health clinics (Health Protection Agency, 2009) for a number of STI. These materials were used in classrooms with students aged 14-16 years. Figures 5a and 5b show the incidence of one of the STI for males and females in different age groups. The STI is chlamydia (chosen on the bottom variable). Figure 5a shows data for the year 2000; Figure 5b shows the incidence for the year 2004 (chosen by sliding along the 'Year' variable).



Figures 5a, 5b: Rates of Chlamydia Infection in 2000 and 2004

Class discussions on changes in the incidence of chlamydia over time (dragging the top slider back and forth) drew out some interesting differences (in the incidence (especially the peak incidence) in different age groups of females and males) and some interesting similarities (the scale of the problem has increased dramatically for both groups over a short period of time).

Sliding along 'Year' creates a clear impression that the incidence is growing dramatically, and in all age groups. Students quickly get used to swapping the positions of the variable labels for sex, age range and time, and use these explorations to test conjectures, and to build up a stronger picture of the relationships in the data – see figures 6a and 6b for examples.

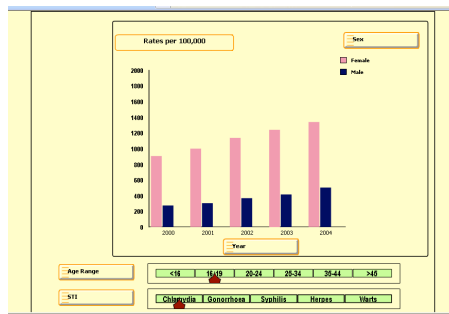


Figure 6a: All the Data for ages 16 – 19

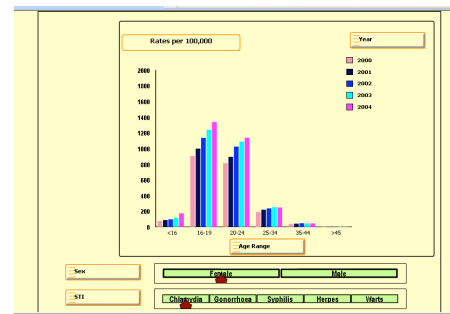


Figure 6b: All the Data for Females

Figure 6a shows that there is a steady linear increase in the recorded incidence of STI over time in the 16-19 year old group. Figure 6b shows roughly linear increases for all age groups, but with different slopes, the lower rates have a slower rate of increase – opening up possibilities for discussions about proportionate and absolute increases in measurements.

There are two striking features of the data on syphilis (not shown here). One is that this infection is extremely rare in comparison with the others. Second is that there are no obvious increases in the incidence over time. Students discussed these features sensibly. We conclude that 12-14 year olds can understand MV interactions in realistic contexts.

Encouraging a critical view of the nature of data in sociology

The sociology curriculum explicitly requires students to study research methods - including strengths and limitations of quantitative approaches and of the data they generate or use. Our visualisation tools describe the provenance of the data, and provide ways to access the metadata.

The tab labelled *Measures* in figure 4 shows the performance data when English and mathematics GCSE must be included in the 5 GCSE examinations to be passed. Unsurprisingly, it shows changes in attainment patterns in some ethnic groups more than others. The tab labelled *Numbers* shows the number of pupils in each of the classification groups, so that students can see that the 'White' sample comprises 480,000 students, and the Chinese sample comprises 2,275 students.

From the viewpoint of theory, we might want to use FSM as a surrogate for poverty, or social class, or both – students can be invited to discuss the plausibility and pitfalls of such assumptions.

The corruptibility of indicator systems provides an interesting topic for discussion. Goodhart's law (1975) refers to the idea that once a social or economic indicator is made a target for some aspect of policy, it then ceases to measure what it used to measure. This happens because stakeholders find ways to raise or lower scores on the indicator via superficial, rather than deep, changes in system activities. Data on educational attainment provides an example. When first introduced as a measure of school performance, the proportion of pupils gaining 5 good grades at GCSE was generally accepted as a reasonable measure of the raw academic performance of a school. However, very quickly schools explored ways of increasing their rating on this measure by a variety of means such as entering pupils for qualifications which counted for as many as 4 GCSEs (e.g. in Art) with associated curriculum distortions (e.g. some students spending as many as half of their lessons studying Art) and by assigning a disproportionate amount of the school's resources to the borderline group of pupils who were identified as being capable of achieving 5 good GCSEs but were at risk of not doing so (by teaching them in very small classes or by offering one-to-one tutor support in one or two subjects).

CAN THESE APPROACHES PROMOTE NEW STATISTICAL LITERACIES?

The classroom materials we have developed do begin to address some of the key elements of SL that were identified earlier. We believe (and have evidence to suggest) that they foster appropriate dispositions and habits of mind about data and evidence that should be components of SL. The materials focus directly on the use of multivariate data to understand complex phenomena. We believe that these materials provoke some important statistical 'habits of mind'.

The ability to evaluate the quality of evidence

Providing information about data sources, and embedding information about metadata into displays, and discussion points for teachers, draw attention to data provenance and quality.

Handling new representations

The explosion in techniques for data representation means that the ability to critique and interpret new displays has become important. Exploring our interfaces provides some experience of this.

A sophisticated approach to measurement

Again, teacher notes on topics such as Goodhart's law, and the appropriateness of using FSM as a surrogate for poverty and/or social class, provoke discussions around important themes.

Conceiving statistics as modeling

Because of the power of statistical models to describe a wide range of phenomena, it is easy to design a curriculum where students are taught statistical techniques and models for their own sake, devoid of context. Here, the focus is primarily on describing phenomena. Students are unlikely to believe that all phenomena can be modeled via linear algebra (although we have not tested this conjecture directly).

Action-oriented statistics

Statistics is seen in the context of an investigative cycle, where the end point is a theoretical account and some action designed to change the current situation. A central idea is to use population data to draw conclusions and plans for action that will be applied to all subpopulations (so interaction, effect size and Simpson's paradox are important). It also encompasses ideas such as risk, and utility.

STRATEGIC AND PEDAGOGIC PRINCIPLES FOR CURRICULUM REFORM

Any educational innovation depends on teachers changing some aspect of classroom activities. People are more likely to change practices if the new practices: make like more fun;

make life easier; solve a problem the person knows they have got; has a good deal of social approval. We set out some principles below that characterise our approach.

Address a major educational goal

The semantic net presents challenges to pedagogy. An important educational goal is to help students become sophisticated users of evidence. This, clearly, involves shaping dispositions – in particular, to increase engagement, and a willingness to explore new data, new representations, and new ideas.

Begin by showing students the potential payoff from learning about statistics

Vygotsky (1978) argues that people need to see the value of a tool before they invest effort in its acquisition. The use of contexts that are of direct interest to students (e.g. crime, alcohol use by young people, incidence of sexually transmitted infections) is a sensible starting point.

Solve a problems teachers know they have got

We offer access to current, authentic, data, and exploratory tools that are easy to use in a class setting (so we act as a bridge between the classroom and the semantic net). We have also chosen to present content that is central to the curriculum.

Use constructivist approaches to teaching

We encourage active exploration of data, hypothesis generation and testing, rather than memorisation of other people's summaries.

Make life easier for teachers

We give teachers resources that reflect well on them in class, and that lead to valuable classroom activities. We provide guidance in sufficient detail to allow them to succeed, but in a way that will be compatible with a range of teaching styles.

Make life easier for students

We present very extensive data sources relevant to the theme in question in a single location that are easy to navigate.

Make life more fun for teachers

Resources lead to enjoyable classroom activities.

Encourage responsible modeling

The nature of the data presented contains implicit messages that encourage users not to make the statistical assumptions that are commonly made, such as linearity, and a belief that complex phenomena can be understood via an assembly of one and two variable analyses. Data are presented in a responsible way: the source of the data is stated clearly (with links), and the metadata are displayed in a comprehensible way (with links to 'real' metadata descriptions for the bravest students).

The Trojan mouse

We are presenting information about phenomena where statistical thinking should play a big role, in order to provoke statistical thinking. A key idea is to bring out big statistical ideas early, often by eliciting them from the students. Big conceptual ideas are emphasised, not technical detail.

HIGHER AND DEEPER?

It is reasonable to ask about the components of the school statistics curriculum that should receive less attention, if multivariate data is to be explored. We offer some starting points; Schield (2012) offers more. Statistical inference is conceptually difficult. More use should be made of exact probability tests, and heuristic devices of the sort advocated by Wild, Pfannkuch, Regan and Horton, (2011). Sampling distributions should be introduced via computer simulations.

CONCLUSIONS

The semantic web offers unprecedented access to large scale, authentic data sets on a huge variety of topics, relevant to public policy and personal happiness. Successful use of such data requires a rather different skill set to skills encapsulated in many current views of 'statistical literacy'. Key skills involve a critical appreciation of data provenance and quality, and understanding of statistical ideas associated with multivariate analysis of large data sets.

The (UK) school curriculum is grounded in 1920s statistics – it is focused largely on one and two variable problems, and the emphasis is on testing hypotheses based on small samples. The semantic web offers multivariate population data, and can form the basis for theory, and for political decision making. Key 'new' SL skills are to describe the qualitative and quantitative features of complex phenomena. Ideas such as effect size and interaction are central to these descriptions. There are two distinct challenges to face; one is to address SL in the public domain; the other is to shape the school curriculum so that school leavers are equipped to benefit from (rather than to be victims of) the semantic web. We have developed some data rich resources that can be embedded in a number of curriculum areas which can be the basis for appropriate curriculum change.

Post Script

Other free teaching materials such as interactive displays on alcohol use by young people, poverty, heart disease and TB can be viewed at <http://www.dur.ac.uk/smart.centre/freeware/>.

ACKNOWLEDGEMENT

This research was funded by the Nuffield Foundation Grant EDU/33713. The views expressed here are entirely those of the authors.

REFERENCES

- Bradshaw, P. (2010). <http://www.guardian.co.uk/news/datablog/2010/oct/01/data-journalism-how-to-guide>
- Brennan, L., Watson, M. and Charles, T. (2012). The importance of knowing context of hospital episode statistics when configuring the NHS. *British Medical Journal*, 344 (7851). *BMJ* 2012;344:e2432
- Brooke, H. (2010). *The Revolution will be Digitised: Dispatches from the Information War*. London: Heinemann.
- Condorcet, J. (1994). *Foundations of social choice and political theory*. Aldershot and Brookfield, VT: Elgar (original work published in 1792).
- data.gov/ <http://www.data.gov/>
- Data.gov introductory video <http://www.socrata.com/datagov/new-data-gov-platform-video-overview/> (downloaded August 2011)
- data.gov.uk <http://data.gov.uk/>
- datamasher <http://www.datamasher.org/>
- delMas, R., Garfield, J., Ooms, A. and Chance, B. (2007). Assessing Students' Conceptual Understanding After a First Course in Statistics. *Statistics Education Research Journal*, 6(2), 28-58. <http://www.stat.auckland.ac.nz/serj>
- Department for Education (2010). *The Importance of Teaching: The Schools White paper 2010*. London: HMSO. <http://www.education.gov.uk/schools/toolsandinitiatives/schoolswhitepaper/b0068570/the-importance-of-teaching/>
- Equality and Human Rights Commission (2010). *How fair is Britain? Equality, Human Rights and Good Relations in 2010*. London: EHRC. <http://www.equalityhumanrights.com/key-projects/triennial-review/>
- Gal, I. (2002). Adult statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70(1), 1-25.
- Gapminder <http://www.gapminder.org/downloads/>

- Goodhart, C.A.E. (1975). "Monetary Relationships: A View from Threadneedle Street". *Papers in Monetary Economics* (Reserve Bank of Australia) I. cited in http://en.wikipedia.org/wiki/Goodhart%27s_law.
- Health Protection Agency www.hpa.org.uk
- Neurath, O. (2010) *From Hieroglyphics to Isotype: A Visual Autobiography*. London: Hyphen Press.
- Nicholson, J., Ridgway, J. and McCusker, S. (2011). Visualise then Conceptualise. *Social Science Teacher*, 40(3), 8-12.
- OECD (2007). *Istanbul Declaration*. Online: www.oecd.org/dataoecd/14/46/38883774.pdf
- OECD Statsportal <http://www.oecd.org/statsportal/>
- Open Data Handbook <http://opendatahandbook.org/en/>
- Open Knowledge Foundation <http://okfn.org/>
- Open Statistics eXplorer <http://ncva.itn.liu.se/explorer/openexp?l=en>
- Playfair, W. (1786, 2005) *The Commercial and Political Atlas and Statistical Breviary* Cambridge University Press
- ProgrammableWeb <http://www.programmableweb.com/>
- Ridgway, J., Nicholson, J. and McCusker, S. (2007a). Reasoning with Multivariate Evidence. *International Electronic Journal of Mathematics Education* 2(3), 245-269.
- Ridgway, J., Nicholson, J. and McCusker, S. (2007b) Teaching statistics - despite its applications. *Teaching Statistics*, 28(2), 44-48.
- Ridgway, J., Nicholson, J. and McCusker, S. (2011). Developing Statistical Literacy in Students and Teachers. In C. Batanero, G. Burrill and C. Reading (Eds.). *Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education*. Springer: Dordrecht. New ICMI Study Series, Vol. 15, 311-322.
- Ridgway, J., Swan, M. and Burkhardt, H. (2001). Assessing Mathematical Thinking Via FLAG. In: D. Holton and M. Niss (eds.): *Teaching and Learning Mathematics at University Level - An ICMI Study*. Dordrecht: Kluwer Academic Publishers. pp. 423-430.
- Rogers, S. (2011) *How to get to grips with data journalism*. <http://www.journalism.co.uk/skills/how-to-get-to-grips-with-data-journalism/s7/a542402/>
- Schild, M. (2012). Personal website. <http://web.augsburg.edu/~schild/>
- SMART Centre <http://www.dur.ac.uk/smart.centre/>
- Stiglitz, J., Sen, A. and Fitoussi, J-P. (2009). *Report by the Commission on the Measurement of Economic Performance and Social Progress*. OECD. http://www.stiglitz-sen-fitoussi.fr/documents/rapport_anglais.pdf
- Swan, M. and Phillips, R. (1998). Graph interpretation skills among lower-achieving school leavers. *Research in Education*, (60), 10-20
- UNdata <http://data.un.org/>
- Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wallman, K.K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association*, 88 (421), 1-8.
- Watson, J. and Nathan, E. (2010). *Assessing the Interpretation of two-way tables as part of statistical literacy*. Proceedings of ICOTS8 http://www.stat.auckland.ac.nz/~iase/publications/icots8/ICOTS8_5E1_WATSON.pdf
- WC3 Recommendations (2004). http://www.w3.org/standards/techs/rdf#w3c_all
- Wild, C.J., Pfannkuch, M., Regan, M., and Horton, N.J. (2011). Towards more accessible conceptions of statistical inference. *Journal of the Royal Statistical Society Series A* 174(2), pp 1-23.