Chapter Seven:

Data-driven journalism

Maureen Henninger

Simply put, data journalism is the use of data in journalistic analysis and reportage. This is not a new concept as journalists have used data such as statistics and incorporated graphs to support reportage for many years. The use of data for support and clarification is still an important part of data journalism therefore data visualisation design and its appropriate use is covered in this chapter. However the analysis of data (and text) using data mining techniques to create new knowledge, while not new concept, is relatively new in journalism and enables journalists to find or discover a story. And finally, data journalism is digital story-telling that is very rich in content and, when delivered in the online environment enables readers to explore the story interactively.

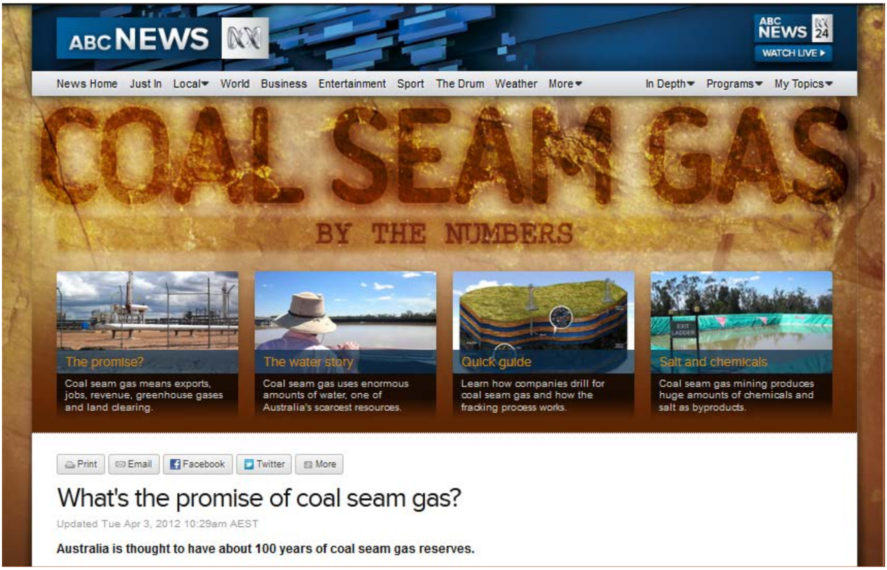
The momentum for data journalism is being driven by two major factors: the exponential growth of online access to datasets, particularly through open government data initiatives, and the development of powerful tools and technologies that enable the mining of data and information, and its subsequent visualisation.

Data journalism requires large amounts of time and many skills; in fact it is often the product of an investigative team, including journalists, researchers, statisticians, information designers, data and text analysts, information visualisation specialists, and web developers. This chapter addresses the principles of many of these skills but does not aim to make journalists experts in any; it is assumed that they will follow up any area that is of particular interest—the Web is awash with such expertise.

By way example I would point to a project done by an investigative team is the data journalism project done by the Australian Broadcasting Corporation *Coal Seam Gas by the Numbers* (Figure 1). As Wendy Carlisle stated (2012) “it wasn’t exclusively data journalism — but a hybrid of journalisms that was born of the mix of people on the team”.[1]

Figure 1 *Coal Seam Gas by the Numbers*, ABC News Online (data journalism project, 2011)

[1] The author of this chapter was the “academic consultant with expertise in data mining, graphic visualization, and advanced research skills” (Carlise, 2012, p. 27).



COAL SEAM GAS

BY THE NUMBERS

The promise?

Coal seam gas means exports, jobs, revenue, greenhouse gases and land clearing.

The water story

Coal seam gas uses enormous amounts of water, one of Australia's scarcest resources.

Quick guide

Learn how companies drill for coal seam gas and how the fracking process works.

Salt and chemicals

Coal seam gas mining produces huge amounts of chemicals and salt as byproducts.

What's the promise of coal seam gas?

Updated Tue Apr 3, 2012 10:29am AEST

Australia is thought to have about 100 years of coal seam gas reserves.

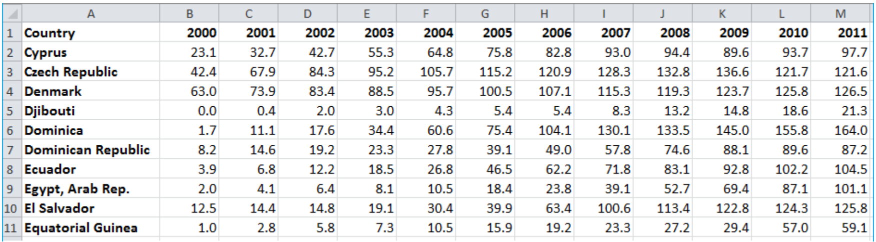
So let’s look at the various resources, processes, skills and tools that are required for a data journalism project.

Public data sets

In chapter 5 we looked a range of tools for discovering information, including finding statistics and datasets. To briefly reprise, statistics are generated by national and international government and non-government organisations and many of these raw datasets are made available online for re-use under license. They can be found at government and non-government websites or by using general and specialised subject directories such as Infomine and OFFSTATS (see Tables 3, 4 and 5 in Chapter 5). In general open government data (OGD) is available under a Creative Commons or similar licensing agreement; however you must always check on the organisations reuse policy, for example the United Nations COMTRADE (Commodities Trade Statistics) datasets reuse agreement states “UN COMTRADE data are provided for internal use only and may not be re-disseminated in any form without the written permission of the United Nations Statistics Division (UNSD)”.[2]

Many of these datasets are time series, a sequence of well-defined data items collected over a period of time at uniform intervals (daily, weekly, quarterly, etcetera). For example the World Bank on worldwide mobile/cellular subscriptions as shown in Figure 2.

Figure 2 Mobile cellular subscriptions (per 100 people)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	**Country**	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
2	**Cyprus**	23.1	32.7	42.7	55.3	64.8	75.8	82.8	93.0	94.4	89.6	93.7	97.7
3	**Czech Republic**	42.4	67.9	84.3	95.2	105.7	115.2	120.9	128.3	132.8	136.6	121.7	121.6
4	**Denmark**	63.0	73.9	83.4	88.5	95.7	100.5	107.1	115.3	119.3	123.7	125.8	126.5
5	**Djibouti**	0.0	0.4	2.0	3.0	4.3	5.4	5.4	8.3	13.2	14.8	18.6	21.3
6	**Dominica**	1.7	11.1	17.6	34.4	60.6	75.4	104.1	130.1	133.5	145.0	155.8	164.0
7	**Dominican Republic**	8.2	14.6	19.2	23.3	27.8	39.1	49.0	57.8	74.6	88.1	89.6	87.2
8	**Ecuador**	3.9	6.8	12.2	18.5	26.8	46.5	62.2	71.8	83.1	92.8	102.2	104.5
9	**Egypt, Arab Rep.**	2.0	4.1	6.4	8.1	10.5	18.4	23.8	39.1	52.7	69.4	87.1	101.1
10	**El Salvador**	12.5	14.4	14.8	19.1	30.4	39.9	63.4	100.6	113.4	122.8	124.3	125.8
11	**Equatorial Guinea**	1.0	2.8	5.8	7.3	10.5	15.9	19.2	23.3	27.2	29.4	57.0	59.1

Source: The World Bank[3]

This dataset is well-defined and is a complete time series for twelve years. However sometimes you may need to combine several statistical sources to obtain a complete time series with no null data points (empty cells). As we shall see in the section on cleaning data there are several tools and techniques for doing this. Before beginning any data journalism project you need to search for and retrieve a copy of the data; if you have not been able to download a public dataset, sometimes it can be ‘scraped’ from other online sources, for example an HTML web page, or a pdf file and put into a spread sheet format such as Excel.

Screen scraping Web data

Datasets are often embedded in published documents on the Web, either in an HTML page or in a pdf (portable document format) format. If it is in an HTML

[2] For further details see http://comtrade.un.org

[3] The World Bank terms of use: “You may extract, download, and make copies of the information contained in the Datasets, and you may share that information with third parties”.

Web page it is relatively easy to copy and paste from your browser into a spread sheet. A pdf file however can be more problematic, since this is a display format which is created in two ways—either it is text-based, that is converted from a text file and therefore has the text embedded in it, or is a scanned image file. The data in a scanned image file cannot be copied and pasted and techniques for this doing are complicated and not covered in this overview.[4]

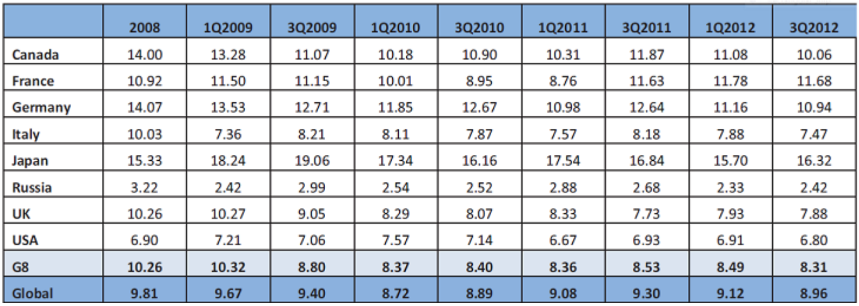
The following is a simple example of extracting the data in a text-based pdf file (Figure 3); it demonstrates how to screen scrape and import into Excel to create a well-defined, usable dataset (if you have Adobe Professional the table should copy instantly into Excel). The process has several steps shown in Figures 3–5:

copying and pasting into a Word file;
converting the data into a table;
copying the table into Excel;
cleaning it if required.

If the table can be copied and pasted it generally appears in ASCII (American Standard Code for Information Interchange) text format, that is, all display formatting is removed, and instead will have a space between each data point (Figure 4). This can easily be converted into a Word table by specifying that the separator is a space—this is similar to the CSV (comma-separated variable) format. Once the data is in a table it can be pasted into Excel. Alternatively the data pasted into Word can be saved as a text (.txt) file and imported into Excel (Figure 5).

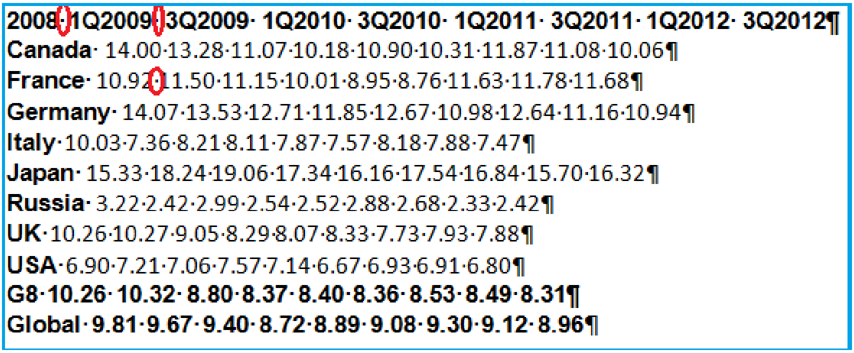
[4] See the tutorial by ProPublica's Dan Nguyen, *Chapter 5: Getting Text Out of an Image-Only PDF* (2010) viewed 10 February 2013, http://www.propublica.org/nerds/item/image-to-text-ocr-and-imagemagick

Figure 3 Table embedded in a pdf file)[5]



	2008	1Q2009	3Q2009	1Q2010	3Q2010	1Q2011	3Q2011	1Q2012	3Q2012
Canada	14.00	13.28	11.07	10.18	10.90	10.31	11.87	11.08	10.06
France	10.92	11.50	11.15	10.01	8.95	8.76	11.63	11.78	11.68
Germany	14.07	13.53	12.71	11.85	12.67	10.98	12.64	11.16	10.94
Italy	10.03	7.36	8.21	8.11	7.87	7.57	8.18	7.88	7.47
Japan	15.33	18.24	19.06	17.34	16.16	17.54	16.84	15.70	16.32
Russia	3.22	2.42	2.99	2.54	2.52	2.88	2.68	2.33	2.42
UK	10.26	10.27	9.05	8.29	8.07	8.33	7.73	7.93	7.88
USA	6.90	7.21	7.06	7.57	7.14	6.67	6.93	6.91	6.80
G8	10.26	10.32	8.80	8.37	8.40	8.36	8.53	8.49	8.31
Global	9.81	9.67	9.40	8.72	8.89	9.08	9.30	9.12	8.96

Figure 4 PDF table screen scrape, pasted into Word (Source: Costs of Migrant Remittances Services)

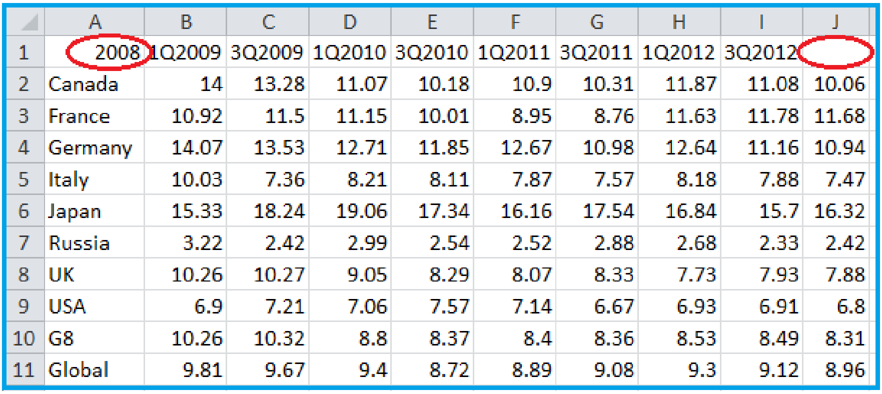


```
2008·1Q2009·3Q2009· 1Q2010· 3Q2010· 1Q2011· 3Q2011· 1Q2012· 3Q2012¶
Canada· 14.00·13.28·11.07·10.18·10.90·10.31·11.87·11.08·10.06¶
France· 10.92·11.50·11.15·10.01·8.95·8.76·11.63·11.78·11.68¶
Germany· 14.07·13.53·12.71·11.85·12.67·10.98·12.64·11.16·10.94¶
Italy·10.03·7.36·8.21·8.11·7.87·7.57·8.18·7.88·7.47¶
Japan· 15.33·18.24·19.06·17.34·16.16·17.54·16.84·15.70·16.32¶
Russia· 3.22·2.42·2.99·2.54·2.52·2.88·2.68·2.33·2.42¶
UK·10.26·10.27·9.05·8.29·8.07·8.33·7.73·7.93·7.88¶
USA·6.90·7.21·7.06·7.57·7.14·6.67·6.93·6.91·6.80¶
G8·10.26·10.32·8.80·8.37·8.40·8.36·8.53·8.49·8.31¶
Global·9.81·9.67·9.40·8.72·8.89·9.08·9.30·9.12·8.96¶
```

Note that in Figure 4 the right column (3Q2012) appears to have no data, and once imported into Excel the headings are not aligned; this can be corrected by simply inserting another cell at A1 and adding the variable label Country.

Figure 5 Screen scraped table imported from Word into Excel

[5] While this data is available from the World Bank Remittances database, the report, a pdf file, is used as an example of screen scraping. Source: Table I *An analysis of Trends in the average Cost of Migrant Remittance Services*, *2012*, viewed 24 January 2013, <http://remittanceprices.worldbank.org/~/media/FPDKM/Remittances/Documents/RemittancePriceWorldwide-Analysis-Nov2012.pdf>



	A	B	C	D	E	F	G	H	I	J
1	2008	1Q2009	3Q2009	1Q2010	3Q2010	1Q2011	3Q2011	1Q2012	3Q2012	
2	Canada	14	13.28	11.07	10.18	10.9	10.31	11.87	11.08	10.06
3	France	10.92	11.5	11.15	10.01	8.95	8.76	11.63	11.78	11.68
4	Germany	14.07	13.53	12.71	11.85	12.67	10.98	12.64	11.16	10.94
5	Italy	10.03	7.36	8.21	8.11	7.87	7.57	8.18	7.88	7.47
6	Japan	15.33	18.24	19.06	17.34	16.16	17.54	16.84	15.7	16.32
7	Russia	3.22	2.42	2.99	2.54	2.52	2.88	2.68	2.33	2.42
8	UK	10.26	10.27	9.05	8.29	8.07	8.33	7.73	7.93	7.88
9	USA	6.9	7.21	7.06	7.57	7.14	6.67	6.93	6.91	6.8
10	G8	10.26	10.32	8.8	8.37	8.4	8.36	8.53	8.49	8.31
11	Global	9.81	9.67	9.4	8.72	8.89	9.08	9.3	9.12	8.96

In the above example the data required very little tidying up but often this is not the case, and needs extensive cleaning.

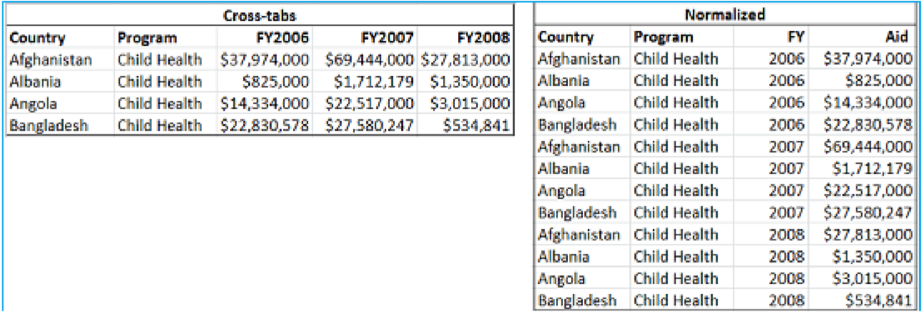
Cleaning data

Cleaning data can be very tedious in order to create a well-defined dataset with consistently named variables, data points in constant values with no null instances, and uniform intervals. In many cases the data may have missing data, misspellings, typos and spelling and naming variations, and other inconsistencies, and its cleaning is time-consuming and painful. While data may originate in a messy state, more often this occurs when a number of different datasets are merged in order to manipulate them (often called data wrangling) using semi-automated tools, or to work with them in pivot tables (see further down).
There are a number of tools and programs available for data cleaning, many of which require sophisticated scripting and programing which is beyond this introductory chapter. Instead I will concentrate on examining Excel, an invaluable tool and one which all data journalists should be very familiar. Two other excellent tools are Data Wranger and Google Refine.

Tip: Excel has a pivot table wizard that normalises cross-tabs data; using the multiple consolidation ranges. See Debra Dalgleish, 2011 *Normalize Data for Excel Pivot Table* (video), www.youtube.com/watch?v=xmqTN0X-AgY

Data Wrangler: This is an online utility developed by the Stanford University Visualization Group in February 2011. At the time of writing it can ‘wrangle’ datasets of only 40 columns and 1000 lines. One of its features is the ability to convert a dataset from a cross-tab format to a normalised format where there is only one column for the FY (Financial Year). If you are using pivot tables in Excel you will have trouble if the data is not normalised (Figure 6). It is available at http://vis.stanford.edu/wrangler/.

Figure 6 Cross-tabs vs. normalised formats



Cross-tabs				
Country	Program	FY2006	FY2007	FY2008
Afghanistan	Child Health	$37,974,000	$69,444,000	$27,813,000
Albania	Child Health	$825,000	$1,712,179	$1,350,000
Angola	Child Health	$14,334,000	$22,517,000	$3,015,000
Bangladesh	Child Health	$22,830,578	$27,580,247	$534,841

Normalized			
Country	Program	FY	Aid
Afghanistan	Child Health	2006	$37,974,000
Albania	Child Health	2006	$825,000
Angola	Child Health	2006	$14,334,000
Bangladesh	Child Health	2006	$22,830,578
Afghanistan	Child Health	2007	$69,444,000
Albania	Child Health	2007	$1,712,179
Angola	Child Health	2007	$22,517,000
Bangladesh	Child Health	2007	$27,580,247
Afghanistan	Child Health	2008	$27,813,000
Albania	Child Health	2008	$1,350,000
Angola	Child Health	2008	$3,015,000
Bangladesh	Child Health	2008	$534,841

Google/Open Refine: This is an open source data cleaning and wrangling tool, originally Freebase Gridworks that developed from an MIT pilot project into Metaweb, it purchased by Google in 2010. In early February 2013 Google announced that it would discontinue its active support and release its code onto GitHub under the new name Open Refine.[6] It is an excellent tool and is easy to use; there are many tutorials online and the best to get started with are those developed by Google at http://code.google.com/p/google-refine/.

Excel

The following examples demonstrates how to merge two quite different datasets about earthquakes in Japan downloaded from the NOAA Significant Earthquake Database which has 26 variables and ANSS (Advanced National Seismic System), with 12 variables. Table 1 shows a selection of fields (variables) from both datasets will be cleaned and fused into one complete dataset using Excel.

Table 1 Selected fields from two earthquakes datasets

NOAA sample variables		**ANSS sample variables**	
Year	2011	DateTime	11/03/2011

[6] At the time of writing you can download the software from http://openrefine.org/

Mo	3		5:46:24 AM
Dy	11		
Hr	5		
46	32		
Sec	24		
Latitude	38.297	38.297	
Longitude	142.373	142.373	
Focal depth	29	Depth	29
Mag	91	Magnitude	9.1
Num (deaths)	2000		
		Nst [No. of seismic stations reporting]	541
Damage $Mill	210000.000		

Sources: NOAA Significant Earthquake Database and ANSS Worldwide Earthquake Catalog

As you can see the data is messy; the problems are:

variables with different names;

in the NOAA dataset the date and time is more granular, that is each component is separate;

in the ANSS dataset the format of date and time is dd/mm/yyyy and a 12 hour clock, which cannot easily be sorted chronologically;

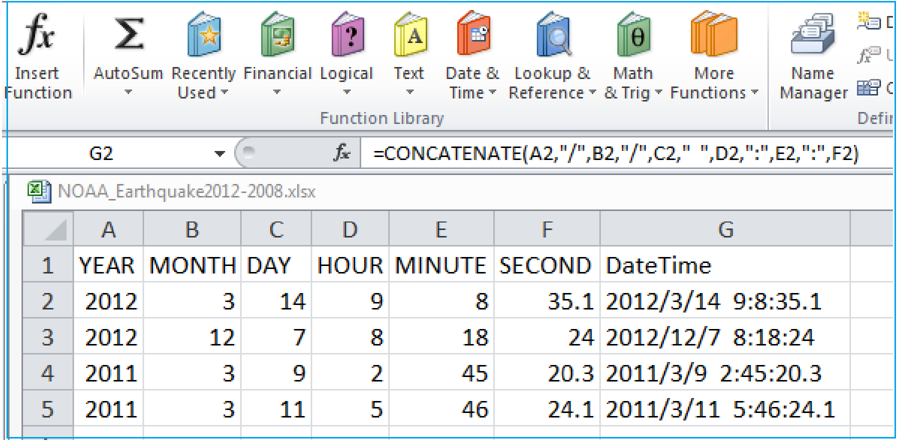
there are empty cells in both datasets implying there is no data available (no of deaths, no. of seismic stations reporting, and damage in $m).

These problems can easily be corrected using Excel and while this chapter is not an extensive guide to all Excel's features, the following steps are indicative of its powerful functionality. However a caveat: some of the choices I have made might not match data requirements in some contexts, rather they have been made to illustrate the use of Excel for data cleaning.

To harmonise the variable names use Excel's search and replace facility to change the names to the full names (Mag to Magnitude; Depth to Focal depth, No. of deaths, etcetera).

To clean the date and time is more problematic; the simplest solution is to have only one variable, DateTime. Use the Excel function CONCATENATE in the dataset with the more granular data. Insert a new column, DateTime and input the function—many of the functions in Excel have wizards which are a set of dialog boxes to lead you through the steps. Copy the function through the entire column; insert another new column and copy into it the created *values only*. Finally delete the column with the concatenation function (see Figure 7 for the concatenation function).

Figure 7 Example of an Excel function (Concatenate, to combine several fields into one field)



G2 =CONCATENATE(A2,"/",B2,"/",C2," ",D2,":",E2,":",F2)

NOAA_Earthquake2012-2008.xlsx

	A	B	C	D	E	F	G
1	YEAR	MONTH	DAY	HOUR	MINUTE	SECOND	DateTime
2	2012	3	14	9	8	35.1	2012/3/14 9:8:35.1
3	2012	12	7	8	18	24	2012/12/7 8:18:24
4	2011	3	9	2	45	20.3	2011/3/9 2:45:20.3
5	2011	3	11	5	46	24.1	2011/3/11 5:46:24.1

Convert the date and time in both datasets to the international standard format (ISO 8601),[7] 2011/11/3 5:46:24. Select the entire DateTime column and using the number formatting, select the *custom* Date facility, type in the format you want—yyyy/mm/dd h:mm:ss.

In both datasets add extra columns where required and in the same order to match the total number of variables, naming them correctly.

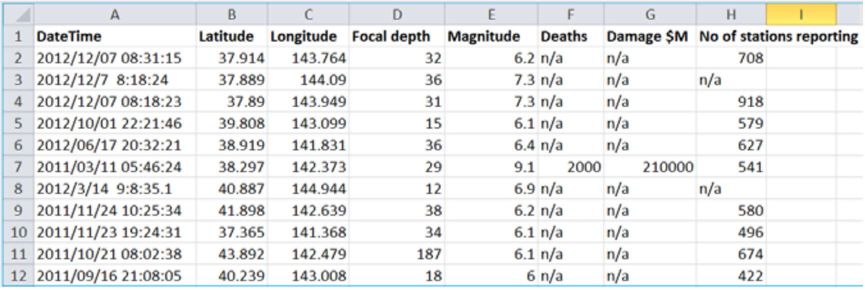
When both datasets are complete and matching, using copy and paste, append one dataset to the bottom of the other.

[7] The International Standards Organisation enables harmonisation of many types of standards to facilitate globalisation and inter-changeability.

Replace the missing data points (empty cells) with the value n/a (not available), using Excel's special search and replace facility as follows:
Home > Find & select > Go to specials > Blanks > OK > type in the character(s) you want > Ctrl + Enter.

Finally to remove the duplicates, select the Latitude column (here I have assumed that earthquakes rarely hit at the exact latitude) and from the Home menu use the Conditional Formatting > Highlight Cells Rules > Duplicate Values. Once these cells are highlighted you can easily scan the entire dataset to see which are describing the same event. Select one of the duplicate rows and add the correct data, then delete the other. See Figure 8 for the final cleaned dataset.

Figure 8 Partial section of the final dataset cleaned with Excel



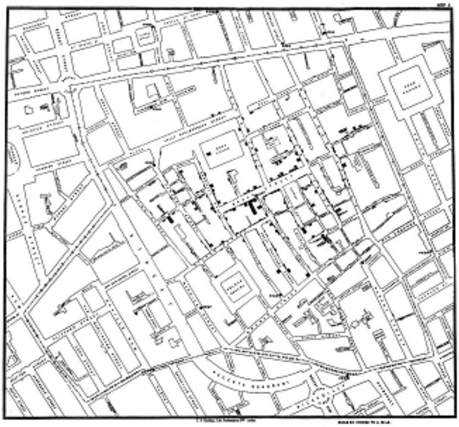
	A	B	C	D	E	F	G	H	I
1	DateTime	Latitude	Longitude	Focal depth	Magnitude	Deaths	Damage $M	No of stations reporting	
2	2012/12/07 08:31:15	37.914	143.764	32	6.2	n/a	n/a	708	
3	2012/12/7 8:18:24	37.889	144.09	36	7.3	n/a	n/a	n/a	
4	2012/12/07 08:18:23	37.89	143.949	31	7.3	n/a	n/a	918	
5	2012/10/01 22:21:46	39.808	143.099	15	6.1	n/a	n/a	579	
6	2012/06/17 20:32:21	38.919	141.831	36	6.4	n/a	n/a	627	
7	2011/03/11 05:46:24	38.297	142.373	29	9.1	2000	210000	541	
8	2012/3/14 9:8:35.1	40.887	144.944	12	6.9	n/a	n/a	n/a	
9	2011/11/24 10:25:34	41.898	142.639	38	6.2	n/a	n/a	580	
10	2011/11/23 19:24:31	37.365	141.368	34	6.1	n/a	n/a	496	
11	2011/10/21 08:02:38	43.892	142.479	187	6.1	n/a	n/a	674	
12	2011/09/16 21:08:05	40.239	143.008	18	6	n/a	n/a	422	

Once you have a clean dataset you can begin the process of analysing it.

Data analysis

The notion of finding evidence or new knowledge in a body of information or data has been an essential part of research for centuries. By the mid-18th century the development of probability and statistical theories by the mathematicians Bayes and LaPlace enabled analytical problem-solving based on research data. The concept of visualising data provided a way to

Figure 9 John Snow's map showing the clusters of cholera cases in the London

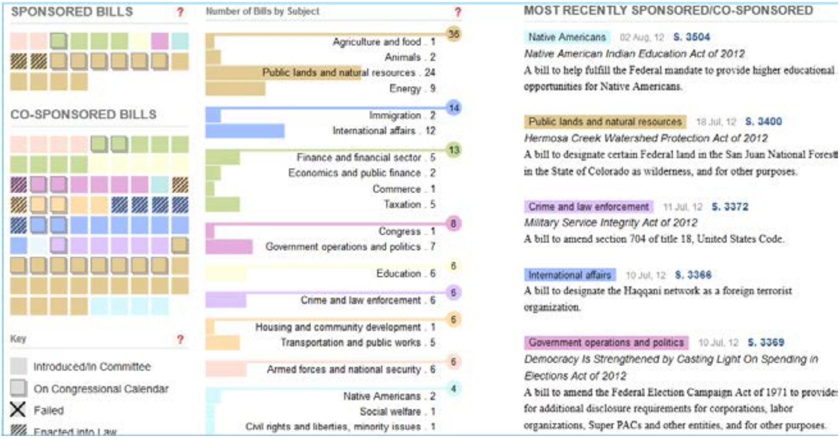


discover patterns, connections and relationships in the data. For example in 1854 the epidemiologist John Snow meticulously collected data on deaths from a cholera outbreak in London. By mapping the data on his now famous map (Figure 9) he made the connection between cholera deaths and a water pump; this connection ultimately led to the knowledge that cholera is a water-borne disease. More recently the increased power of computing has enabled large-scale reuse of data by data mining, and, along with new software tools has made data analysis much easier. However, before looking at methods for analysing datasets, we should remember that information is also in contained in text, and that often this has to be analysed, either manually (what investigative journalists all do by reading appropriate reports and documents), or by using text mining software.

Text analysis

Text or content analysis relies on natural language processing (NLP), a science combining computer science, linguistics and artificial intelligence. Simply put, a computer program parses a sentence (separates into its component parts) and describes their syntactic roles—verb, noun, adjective, etcetera, and then by applying complex rules ascertains its meaning (semantics). More complex applications of NLP facilitate the mining of a large corpus of text which may then be visualised, for example IBM’s Many Bills visualisation of US Congressional legislation (Figure 10).

Figure 10 Text mining and visualisation (IBM's ManyBills)


SPONSORED BILLS
?
CO-SPONSORED BILLS
Key
?
Introduced/In Committee
On Congressional Calendar
Failed
Enacted into Law
Number of Bills by Subject
?
36
Agriculture and food . 1
Animals . 2
Public lands and natural resources . 24
Energy . 9
14
Immigration . 2
International affairs . 12
13
Finance and financial sector . 5
Economics and public finance . 2
Commerce . 1
Taxation . 5
8
Congress . 1
Government operations and politics . 7
6
Education . 6
6
Crime and law enforcement . 6
6
Housing and community development . 1
Transportation and public works . 5
6
Armed forces and national security . 6
4
Native Americans . 2
Social welfare . 1
Civil rights and liberties, minority issues . 1

MOST RECENTLY SPONSORED/CO-SPONSORED

Native Americans 02 Aug, 12 S. 3504
Native American Indian Education Act of 2012
A bill to help fulfill the Federal mandate to provide higher educational opportunities for Native Americans.

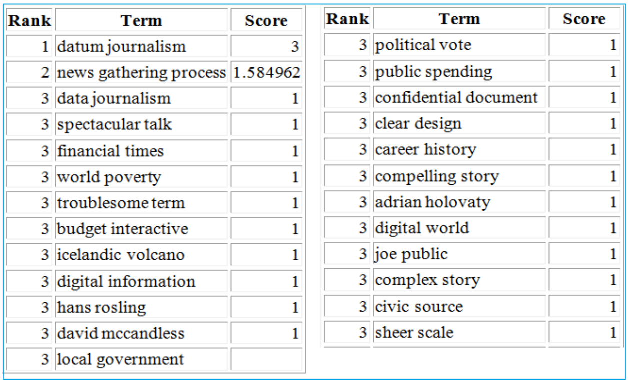
Public lands and natural resources 18 Jul, 12 S. 3400
Hermosa Creek Watershed Protection Act of 2012
A bill to designate certain Federal land in the San Juan National Forest in the State of Colorado as wilderness, and for other purposes.

Crime and law enforcement 11 Jul, 12 S. 3372
Military Service Integrity Act of 2012
A bill to amend section 704 of title 18, United States Code.

International affairs 10 Jul, 12 S. 3366
A bill to designate the Haqqani network as a foreign terrorist organization.

Government operations and politics 10 Jul, 12 S. 3369
Democracy Is Strengthened by Casting Light On Spending in Elections Act of 2012
A bill to amend the Federal Election Campaign Act of 1971 to provide for additional disclosure requirements for corporations, labor organizations, Super PACs and other entities, and for other purposes.

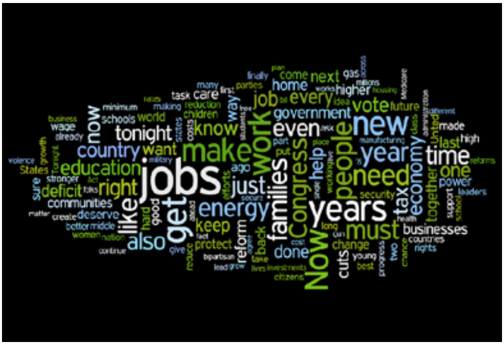
Figure 11 List of terms generated by TerMine from the Introduction to the Data Journalism Handbook



Rank	Term	Score	Rank	Term	Score
1	datum journalism	3	3	political vote	1
2	news gathering process	1.584962	3	public spending	1
3	data journalism	1	3	confidential document	1
3	spectacular talk	1	3	clear design	1
3	financial times	1	3	career history	1
3	world poverty	1	3	compelling story	1
3	troublesome term	1	3	adrian holovaty	1
3	budget interactive	1	3	digital world	1
3	icelandic volcano	1	3	joe public	1
3	digital information	1	3	complex story	1
3	hans rosling	1	3	civic source	1
3	david mccandless	1	3	sheer scale	1
3	local government				

Unfortunately most text mining programs are proprietary and expensive, however the National Centre for Text Mining (NaCTeM) at the University of Manchester has developed TerMine a Web demonstration utility that identifies key phrases in a text and weights (ranks) them according to the number of occurrences in the text—see Figure 11, (Frantzi et al., 2000). TerMine is now available to industry under licence.

Figure 12 Themes in President Obama's 2013 State of the Union speech (created by

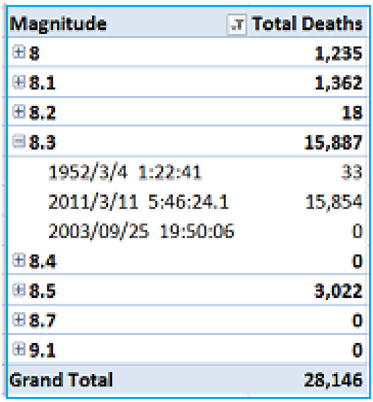


However, very basic visualisations of words in texts can be done using Web-based programs such as Wordle. For example supposing you wanted to visualise, or simply know, what were the major themes of a speech; you paste the text into the template and based on the frequency of the words (unfortunately it cannot construct phrases) Wordle produces a word cloud—Figure 12 shows President Obama's 2013 State of the Union speech; I edited the cloud to remove high frequency non-thematic words *applause* and *America*. This simple tool could be used to compare texts.

Large data sets

Managing large data sets with a large number of variables can be difficult. For example the Japanese earthquakes dataset (Figure 8) has over 2000 rows. Obviously it is impossible to see the entire dataset at once and therefor would be very difficult to do any data analysis without an enormous amount of mind-numbing fiddling. Pivot tables[8] are very handy solutions that enable you to select, drag and drop different variables into a template in order to summarise or filter values. In the case of the Japanese earthquakes dataset creating a pivot table makes it very easy to select three variables—DateTime, Magnitude and No of deaths, and to then filter or summarise in a new table the number of deaths caused by earthquakes with magnitudes equal to and greater than 8 (Figure 13).

Figure 13 No. of deaths in Japanese earthquakes with



Magnitude	Total Deaths
⊞ 8	1,235
⊞ 8.1	1,362
⊞ 8.2	18
⊟ 8.3	15,887
1952/3/4 1:22:41	33
2011/3/11 5:46:24.1	15,854
2003/09/25 19:50:06	0
⊞ 8.4	0
⊞ 8.5	3,022
⊞ 8.7	0
⊞ 9.1	0
Grand Total	28,146

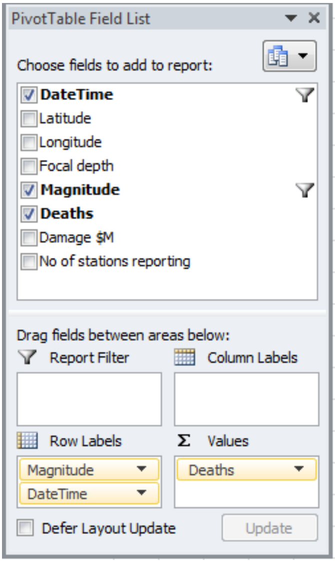
There are many books and online tutorials for creating pivot tables in Excel but the main steps are:

make sure you have a clean dataset with no empty cells;
format the dataset as a table;
insert the table as a pivot table;
using the pivot table template drag and drop the required variables in one of three positions—rows, columns and values;
select the calculation to be performed on the value;
apply appropriate filters to further summarise the data.

As can be seen in Figure 15 the row labels (variables) are Magnitude with DateTime in the secondary position. There is no column variable but the No of Deaths is the value required; the calculation is to Sum, that is, to give the total number of earthquakes, rather than Count, the number of earthquakes in any year (other calculations include Average, Minimum, Maximum, etcetera are available). Filters have been applied to the variables Deaths and Magnitude—to show only those earthquakes with a magnitude of 8.0 or greater that caused at least one death. Finally the table can be formatted in many ways; in this case the subtotals and grant totals have been included and labels have been changed. Lastly you can create a variety of graphs such as that shown in Figure 14.

[8] Excel has a patent on the term Pivot Table, however it is a function of many software vendors.

Figure 14 Excel's pivot table template



PivotTable Field List

Choose fields to add to report:

☑ **DateTime**
□ Latitude
□ Longitude
□ Focal depth
☑ **Magnitude**
☑ **Deaths**
□ Damage $M
□ No of stations reporting

Drag fields between areas below:

Report Filter	Column Labels

Row Labels	Values
Magnitude	Deaths
DateTime	

□ Defer Layout Update

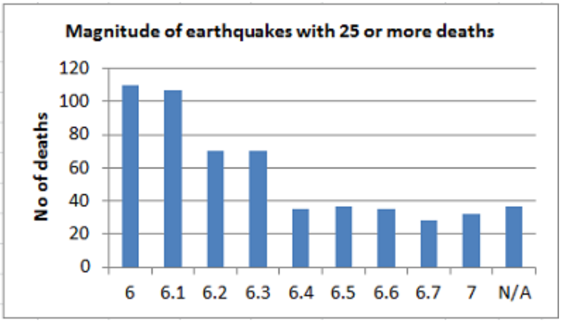
Update

Pivot tables are brilliant for working with large datasets; one or several variables can be selected, moved around and grouped, and values can be calculated in different ways in order to analyse the data. And they can produce one-dimensional or two-dimensional comparisons. The trick is having a firm idea of what you want to analyse in order to select and place the variables in the appropriate rows, columns and values. Once data is analysed visualisations can be created to support the story, or to let the data tell you there is a story to be written.

Data visualisation

Modern data graphics can do much more than simply substitute for small statistical tables. At their best, graphics are instruments for reasoning about quantitative information. Often the most effective way to describe, explore, and summarize a set of numbers –even a very large set –is to look at pictures of those numbers. Furthermore, of all methods for analyzing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful.

Figure 15 Graph created from an Excel pivot table



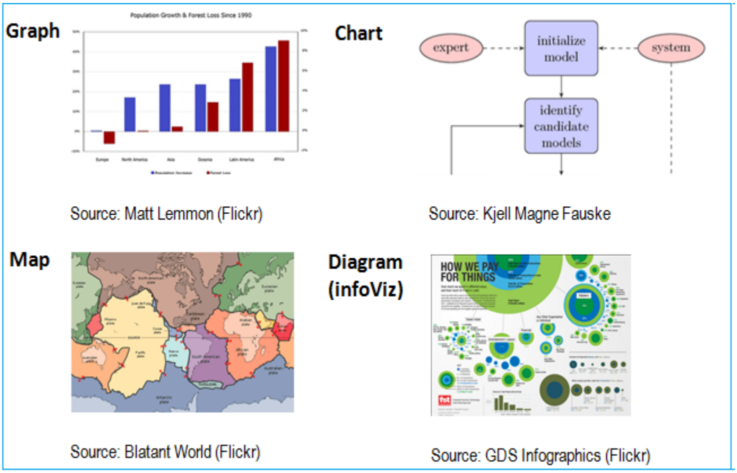
Edward Tufte, 2010, Introduction to The Visual Display of Quantitative Information

Creating a visualization requires a number of nuanced judgments. One must determine which questions to ask, identify the appropriate data, and select effective visual encodings to map data values to graphical features such as position, size, shape, and color

Heer, Bostock, & Ogievetsky, 2010, p. 59

Visualisations are symbolic displays that "reveal the data at several levels of detail from a broad overview to the fine structure" (Tufte, 2001), and there is a wide variety—graphs, charts, maps and diagrams, each of which has properties that make it appropriate to a particular type of data. In this section we examine each of these data visualisations and their usefulness.[9]

Figure 16 Examples of the four types of visual displays (examples are available under Creative Commons licence)



Source: Matt Lemmon (Flickr)

Source: Kjell Magne Fauske

Source: Blatant World (Flickr)

Source: GDS Infographics (Flickr)

The terms graphs and charts are often used synonymously—indeed Excel does—they are in fact different. Graphs such as scatter plots and bar 'charts', require at least two scales, whereas as charts have an internal structure, for example a flowchart. However in the following discussion I also will use the terms synonymously. Maps, on the other hand, generally are determined by spatial relations, and diagrams (often called infoVizs) are schematic pictures of objects and often include some or all of the other types of visual displays.[10]

[9] NOTE: all the illustrations in the following sections are either my own or are licensed under Creative Commons.

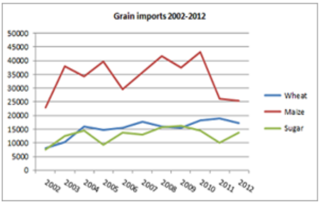
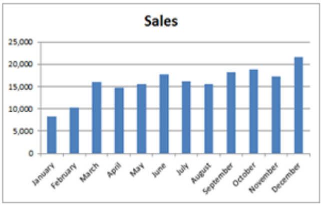
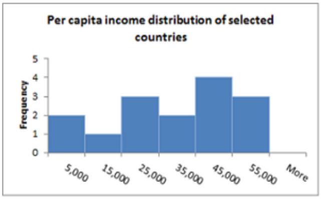
[10] For further discussion on graphs and charts, see Kosslyn, S.M. 1989, 'Understanding

Ben Shneiderman (1996) divided data into a taxonomy of seven data types, each data type required a different type of visualisation. Three of these data types are relevant to this chapter:

2-dimensional—planar or map data;
temporal—time series;
multi-dimensional—relational and statistical databases in which the data can be manipulated (for example in pivot tables).
Data-driven journalism may have occasion to use any of these types of data and therefore need to understand their basic properties in order to select the most appropriate type of visualisation.

Graphs and charts
There are many types of graphs and when to use which depends on the type of data. Here are some of the main ones and when they are used (Figure 17):
line and index charts,
bar charts,
histograms, and
bubble charts.
Line charts and index charts are used for time series data and plot data points at regular interval over time. Index charts are helpful when it is important to display value *changes*

Figure 17 Data types and appropriate graphs & charts

Data type	Appropriate graph/chart	
Time series data	Line and index charts	
Categorical variables	Bar chart	
Quantitative variables	Histograms	

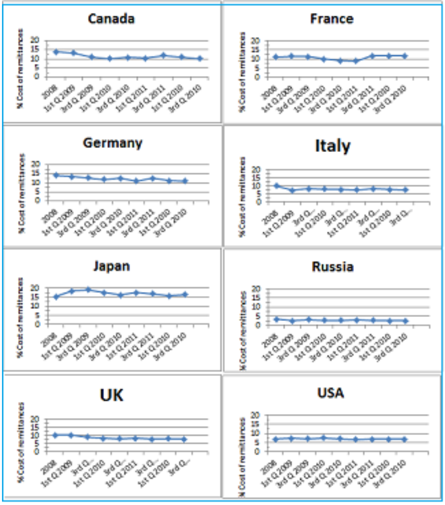
charts and graphs', *Applied cognitive psychology*, vol. 3, no. 3, pp. 185-225.

rather than individual data points. They are interactive, in that in the Web environment you are able to drag a slider over the time frequencies.
Bar charts are on the most frequently used graphs and are typically used for variables that are categories, that is, they have no numerical meaning such as months or countries.

Histograms are special bar charts show qualitative categories such as statistical frequencies and distributions, for example the per capital income distribution of selected countries, in which the income data is 'classed' or put into categories.
Bubble charts are used to display 3 dimensional data, that is, three data series, two of which are located on the X and Y axis and the third appears as filled-in circles in proportional sizes.

Often when you have several series in the one dataset, or if you have very dense data, placing them in a single visualization may result in an unreadable graph. One way of solving this problem is to use a series of graphics, what Tufte (2001) calls small multiples which he describes as "resemble[ing] the frames of a movie: a series of graphics, showing the combination of variables, indexed by changes in another variable . . . the design remains constant through all the frames, so that attention is devoted entirely to shifts in the data." This type of visualisation is also call a trellis or panel chart and the series can consist of any type of visualisation (Figure 18).

Figure 18 Small multiples (from Figure 2)



Excel is an excellent tool to create many different charts and graphs and you should be familiar with its charting features; to create a histogram you need to make sure that the Analysis ToolPak *add*-in is available and creating small multiples or trellis charts in Excel is complicated. But no matter what type of graph you use you must make sure that it doesn't lie.

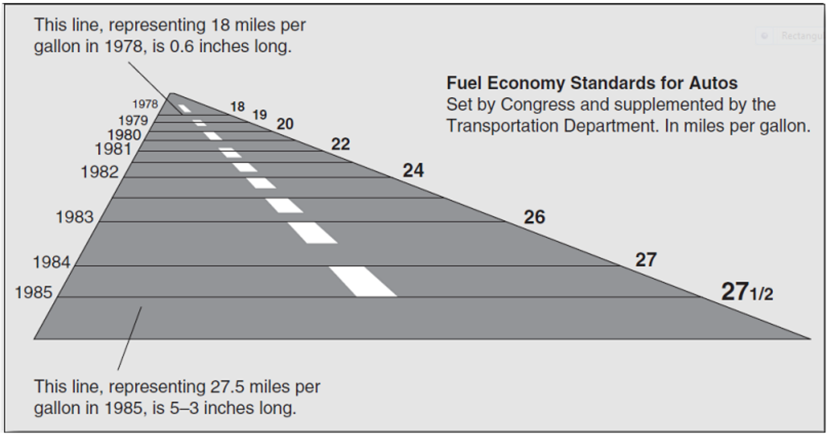
Graphical integrity

Edward Tufte in his discussion of graphical integrity notes that “graphical excellence begins with telling the truth about data” (2001); that the “representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented" (2001) He suggested that such violations can be measured by the “lie factor”, that is, the as the ratio of the size of an effect shown in the *graphic* to the size of the effect in the data and should always be less than one. He devised the following formula

$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

Thus in Figure 19 the data shows the changes fuel economy standards for automobiles in miles per gallon went from 18 in 1978 to 27.5 in 1985, an increase of 53%, whereas the length graph lines representing this data (0.6 inches and 5.3 inches) is an increase of 783%. Tufte calculated that the lie factor of the graphic is 783%/53% = 14.8!

Figure 19 Extreme example of graphic misrepresentation



Source: Tufte 2001, p. 57. Used by permission by Graphics Press.

We have now examined data visualisations in the form of graphs and charts, and by implication how these can be done using Excel. We will now examine another of the more valuable types of methods of visualising information and data—maps.

Maps

For most of human history maps have been used to communicate complex ideas and knowledge. Much of this is spatial and/or temporal information; one of the earliest is the 5th century BC Babylonian map of the world, now in the British Museum. Early methods for finding one's position or location (latitude and longitude) were generally based the on altitude of the sun, on star position and dead reckoning, but it was not until the invention of an accurate method of calculating longitude—Harrison's marine chronometer— that accurate positions could be calculated.

In the early 1970s, with the availability of satellites the US Department of Defense[11] developed a 24 satellite system, the Global Positioning System (GPS) that broadcasts two radio signals that give the satellite's precise position and time. The second signal is the data that has been made available for civilian use and is accessible to anyone with a GPS receiver; it is the basis of applications which require geographic coordinates (latitude and longitude) for location on Earth. The representation of geographic coordinates has been standardised[12], for example the location from where I am writing this is 33°51′35.9″S 151°12′40″E (Sydney, Australia—southern latitude, eastern longitude).

For the journalist (or any other information professional) the ability to create maps on which information can be superimposed automatically at a precise spatial point is an enormous development. A variety of software tools have been developed that can process other geographic data such as postcodes or street addresses to their equivalent GPS coordinates, a process called geocoding. For example the addresses of several journalism schools (Table 2) can be automatically geocoded by the 'multiple address locators' function at GPSVisualizer (see Table 3 for results).

[11] Other countries such as Russia, China and India are also developing GPS systems and the European Union's Galileo system is being built by the European Space Agency to be independent from these and that of the US.

[12] ISO 6709 *Standard Representation of Geographic Point Location by Coordinates*

Table 2 Addresses of selected schools of journalism

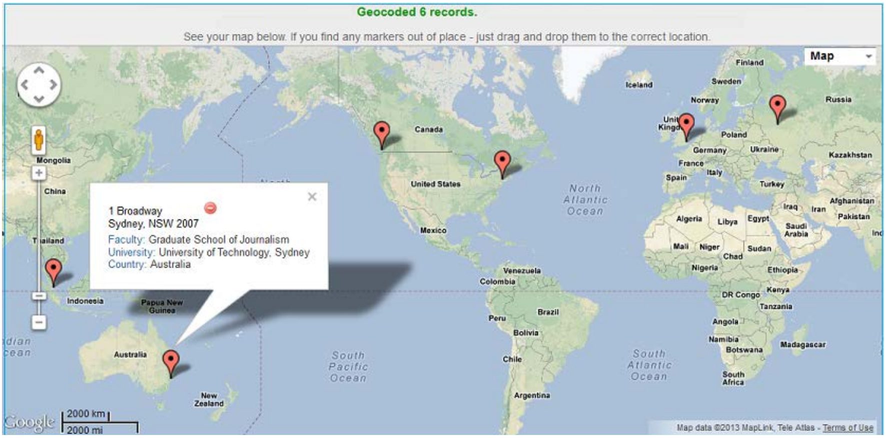
University	**Address**	**City**	**State**	**Postcode**	**Country**
University of Technology, Sydney	1 Broadway	Sydney	NSW	2007	Australia
Columbia University	2950 Broadway	New York	NY	10027	USA
Graduate School of Journalism	6388 Crescent Road	Vancouver	BC	V6T 1Z2	Canada
Moscow State University	9 Mokhovaya Str.	Moscow	n/a	125009	Russia
University of Kent	n/a	Medway	Kent	ME4 4AG	UK
Nanyang Technological University	50 Nanyang Ave	Singapore	n/a	639798	Singapore

Table 3 Addresses of selected journalism schools automatically geocoded by GPSVisualizer

Latitude	Longitude
-33.884331	151.19725
40.807648	-73.964027
49.264763	-123.244623
55.753956	37.612015
51.39731	0.54133
1.331355	103.663651

Generally speaking Web tools such as GPSVisualizer and Batchgeo, once they have processed the addresses into geographic coordinates, include all the other information in the dataset in a Google or Yahoo map (see Figure 20). Maps created by these tools are saved online, on their servers and you embed the links in your Web pages; you can also save the map as an image file using a snipping or screen shot tool for paper publication.

Figure 20 A very basic map of selected journalism schools created by Batchgeo



Geocoded 6 records.

See your map below. If you find any markers out of place - just drag and drop them to the correct location.

1 Broadway
Sydney, NSW 2007
Faculty: Graduate School of Journalism
University: University of Technology, Sydney
Country: Australia

The map above is a dot map, in which the geographic coordinates are represented by a shape, in this case the standard Google ‘paddle’ shape; there are many other shapes and logically any icon can be used as a map shape—dots, pins, flags— as long as the map generator is customisable. Dot maps are very useful for representing quantity and density, for example dots of graduated sizes can represent large and small quantities. As in all information design, when designing maps and graphics with shapes, you need to be sure the symbology being conveyed is appropriate (see the section Colour and Information Design below). More complex types of maps are cartograms, choropleths and heat maps (see Figure 21). Let’s look briefly at these maps and when they are used.

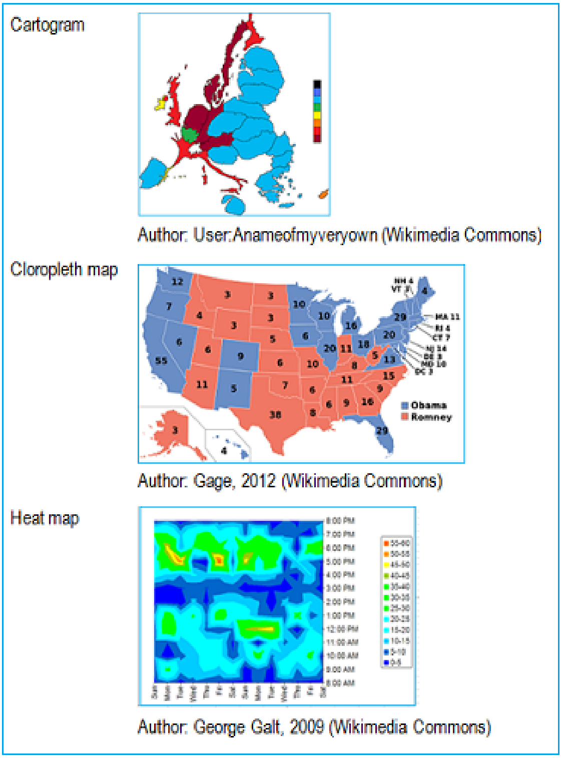
Cartograms and cloropleth maps are thematic maps, that is, they indicate a theme concerning a particular geographic area. The theme can be any data such as energy consumption, world education levels, or language distribution.

A cartogram is a map on which the data, represented by symbols, is superimposed on geographic areas that are scaled to show the relative sizes of the data.

The choropleth map, using colour coding, aggregates data by geographic areas that have been previously defined, for example countries or postcode areas. In order to create the filled geographic area you need a polygon file defining the boundaries of the area.

A heat map is data visualisation where the individual data values are contained in a matrix (2 dimensional data) and are represented as colour gradients. Although not necessarily as effective as a map, a chart with graduated colour coding can be done in Excel using conditional formatting.

Figure 21 Examples of complex colour maps



Cartogram

Author: User:Anameofmyveryown (Wikimedia Commons)

Cloropleth map

Author: Gage, 2012 (Wikimedia Commons)

Heat map

Author: George Galt, 2009 (Wikimedia Commons)

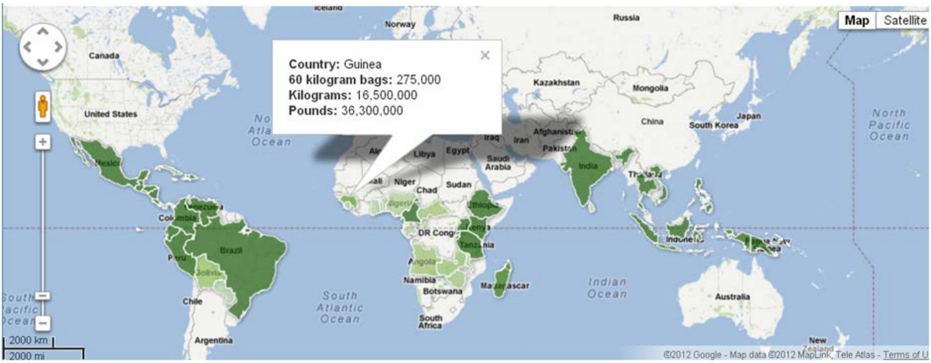
To create these complex maps specialised software is required, the most sophisticated are referred to as Geographic Information Systems (GIS). However in recent years other software programs have been developed and two such programs widely used for data journalism are Google's Fusion Tables and Tableau Public.

Google fusion tables

This Google product, which is still experimental, is used in conjunction with your Google Docs (now Google Drive) account. At its most simple it is a spread sheet application that can create charts and graphs. However it additional functionality to enable you to fuse (merge) different datasets from disparate sources, including polygon files—in KML (Keynote Markup Language) format— which are needed to create a cloropleth map. The map below, showing world coffee production, is the one that I created by doing the Google tutorials.[13]

Figure 22 Cloropleth map of world coffee production created with Google Fusion Tables

[13] Another very good tutorial is Sreeram Balakrishnan's *Fusion Tables Workshop*, 2012, available at https://sites.google.com/site/fusiontablestalks/talks/fusion-tables-where-2-0-workshop, viewed 21 January 2013.

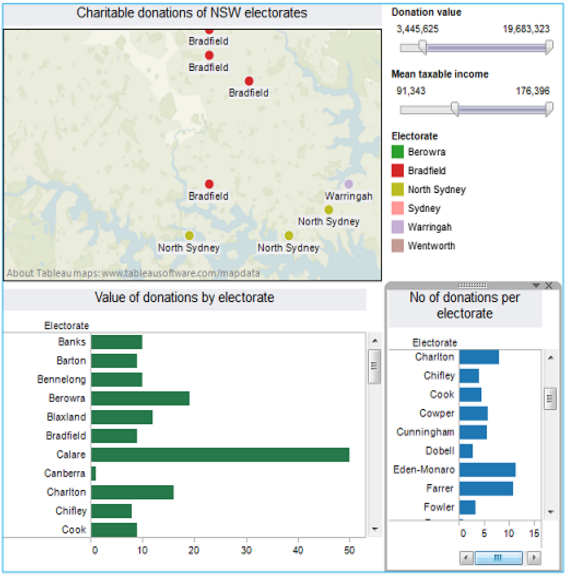


Country: Guinea
60 kilogram bags: 275,000
Kilograms: 16,500,000
Pounds: 36,300,000

Tableau Public

Tableau Software is a data visualisation program that is widely used by data journalists from newspapers like The Guardian and the New York Times. It combines all the tools that this chapter has discussed, and more—spread sheets, pivot tables, automated geocoding, relational databases, the ability to merge tables, and of course, to create very good visualisations. In 2010 Tableau released a free product, Tableau Public that has most of the functionality of the professional package, but your visualisations cannot be saved on your computer. They must be saved on the Tableau servers (you can of course save any

Figure 23 Tableau Public interactive dashboard



visualisation as an image file as I have done for this chapter). Unfortunately there is not yet a MAC version, although apparently one is being developed.

Tableau Public's basic functions are very easy to use and its 'getting started' online training modules are very helpful. One of the great features of Tableau Public is the ability to create an interactive dashboard for online stories. Figure 23 above is a visualisation I created to show data on charitable donations based on electorates in the state of New South Wales, Australia.[14] The dashboard includes a map and two bar charts, which the reader can interactively explore using the sliders and scroll bars to filter the data and change the displays.

While Tableau includes a polygon files of geographic areas of the United States down to the zip code level for creating cloropleth maps of United States, currently these are not built-in for other geographic regions. There is however a great Tableau community which are very active and generous in sharing workarounds and tutorials for Tableau use; one of these is the Clearly and Simply blog.[15]
Finally no matter what type of visualisation you create—map, graph, infoviz—it is important to understand that for information and data to be communicated effectively is must be well-designed.

Colour and information design

Information design is not graphic design; graphic design is only part of the effective communication of information. It enables the reader to make sense of what is being communicated. Information design is "the *integrator* that brings other disciplines together to create excellent information solutions . . . to provide the most possible clarity, understanding and effectiveness."(Knemeyer, 2003 para 5-6). Once you have your data, your analysis is done and you have your story here are some points to consider when you create your visualisations.
Consider if the data is suited to a visualisation—if comparing specific values is required perhaps a well-designed table would be better.
Use colour sparingly—one spot of colour draws the eye to what is important, a plethora of colour causes confuses.

[14] Data sources: Australian Taxation Office, the Australian Electoral Commission and the Australian Bureau of Statistics Census data. I wish to acknowledge much of the data cleaning done by one of my Masters student Katrina Stolk.
[15] Clearly and Simply [blog] http://www.clearlyandsimply.com/clearly_and_simply/

Colour can encode the data—consider monotone gradations for comparative data. Avoid ‘chart junk’ and reduce non-data ink—remember it is the data that you need to see, not the design.

If you have high-information graphics consider reducing the size of the graph. You can reduce data graphics by 50% with no loss of legibility—what Tufte refers to as the shrink principle.

Using the shrink principle consider creating a small multiples visualisation to show data density effectively.

Conclusion

This chapter has provided an introductory overview of data journalism, with the sections generally following the logical progression of the processes involved, from the finding of data through its cleaning to its visualisation. And as befits a chapter on data visualisation it relies on many visualisations to illustrate these processes.

Emphasis is placed on the finding of data sources, particularly public sector datasets now being released through the open government data initiatives and the importance of checking any restrictions on the reuse of these data. There is a section on data that is not in a useable format and the steps that are required to wrangle and clean it; while there are several tools to do this, I have concentrated on the use of Excel, an invaluable and indispensable tool in the tool kit of data journalists.

A key section is a discussion of analysis of the information and data, and while there is a brief introduction to the concept and usefulness of text analysis, the emphasis is on data analysis. The notion of using data visualisations for discovering connections, relationships and new knowledge is explored, thus demonstrating the power of data visualisation to uncover a new story.
The majority of the chapter is an exploration of the various kinds of graphical representations of data and the appropriateness matching of each to the data available. Of serious importance is the discussion of the need for graphical integrity, that is, that the visualisation does not distort the data.

Finally I have briefly considered the importance of information design in the effective communication of information and have given some points of consideration for good, effective design.

In conclusion this chapter provides an introduction for journalism students as well as more seasoned journalists for understanding data journalism and has provided a basis on which to hone their skills in the use of tools for creating meaningful graphics which tell an interesting and factual story.

Further readings

FEW, S. 2009. *Now you see it: simple visualization techniques for quantitative analysis*, Oakland, Calif, Analytics Press.

GRAY, J., BOURNEGRU, L. & CHAMBERS, L. (eds.) 2012. *The data journalism handbook : how journalism can use data to improve the news,* Sebastopol, CA: O'Reilly & Associates Inc.

TUFTE, E. R. 2001. *The visual display of quantitative information,* Cheshire, Conn., Graphics Press.

TUFTE, E. R. 1990. *Envisioning information,* Cheshire, Conn., Graphics Press.
STILES, M. *The Daily Viz,* [blog], National Public Radio (NPR), <http://thedailyviz.com/>

References

CARLISLE, W. 2012. The ABC’s Data Journalism Play. *In:* GRAY, J., BOURNEGRU, L. & CHAMBERS, L. (eds.) *The data journalism handbook : how journalism can use data to improve the news.* Sebastopol, CA: O'Reilly & Associates Inc.

FRANTZI, K., ANANIADOU, S. & MIMA, H. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries,* 3**,** 115-130.

KNEMEYER, D. 15 July 2003 2003. Information Design: The Understanding Discipline. *Boxes and Arrows* [Online]. Available from: http://boxesandarrows.com/information-design-the-understanding-discipline/].

SHNEIDERMAN, B. Year. The eyes have it: A task by data type taxonomy for information visualizations. *In:* Visual Languages, 1996. Proceedings., IEEE Symposium on, 1996. IEEE, 336-343.

TUFTE, E. R. 2001. *The visual display of quantitative information,* Cheshire, Conn., Graphics Press.