Data Driven Investigations

Notes, exercises, and ideas from the CIJ course

About this handbook

This book is intended as a handout for people who have attended the CIJ course "Data Driven Investigations". It contains reminders of some of the topics covered in the course, along with suggested exercises, and extra detail we don't have time to go into during the course. It is not a replacement for attending the course.

Contents

Why we need Data Driven Investigations

James Harkin
Director, Centre for Investigative Journalism

At the CIJ we interpret investigative journalism very broadly; sometimes it involves no more than the slow art of cultivating a source, or of picking up the phone.

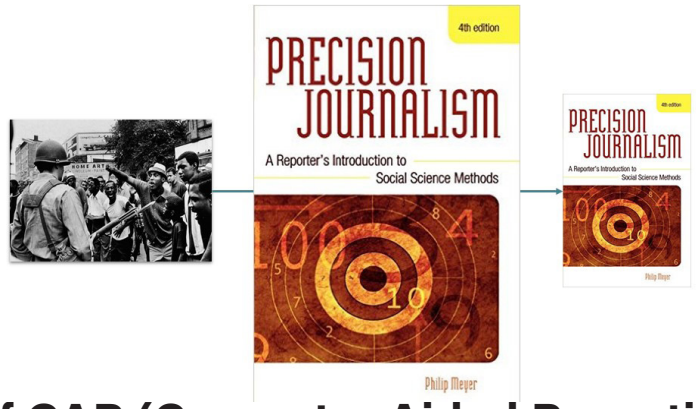
Our only criterion is the quantity and quality of investigative labour which goes into the product, and we want to use the foundation of our quality training to stand up for journalism as craft and properly resourced professional activity.

We're convinced that investigative research is the future of high-value quality journalism.

The best way to tackle fake news of any kind isn't only through fact-checking and schemes for "verification" of new media material after the event (though both can be useful) but by attacking the problem at source – by pouring more and better investigative journalism into the body politic.

One method of adding value through investigative research is via data journalism, at which the CIJ has been a beacon of best practice for over a decade.

In an age of ubiquitous data, it's becoming increasingly clear that the real value comes from connecting all those dots artfully into an impactful story



Brief History of CAR (Computer Aided Reporting)

Datajournalism, or data driven journalism has been a buzzword in news for several years: precisely how many probably depends on where in the world you practise journalism. Some people still use the original term – "Computer Aided Reporting" which is always shortened to CAR.

Computer Aided Reporting is a useful reminder of how long this sub-discipline of journalism has been in existence, and what its roots are. In the course we talk about the work done by Philip Meyer in investigating the Detroit Race Riots of 1967: as a social scientist by training, Meyer decided that rather than characterising the rioters' motives through assumption, he should ask them who they were and why they were rioting. His work won a Pulitzer Prize – and he later went on to write what many regard as the manifesto for data journalism, a book called "Precision Journalism".

Bear in mind, that when Meyer wrote that book in 1973 personal computers were a distant dream: Bill Gates and Steve Jobs were both 18 years old. Aiding one's journalism with a computer meant using punch cards to feed data into a mainframe computer which took up most of a room.

What happened next is well-known. You are probably reading this on a computer which is a distant relative of the one in the picture, and many times more powerful. More importantly, the range and power of the software available to help us has improved beyond all measure. Saying "Computer Aided Reporting" has become as odd as saying "Pen Aided Writing" would be. But it's an important reminder that the reporting is the end result, the computer is a means to an end. When we say datajournalism, the key part of the word is journalism.

If you want to learn more about the history of CAR – follow this link:

https://datajournalism.com/read/longreads/the-history-of-data-journalism

"A better solution is to push journalism toward science, incorporating both the powerful data gathering and analysis tools of science and ***its disciplined search for verifiable truth."***

\- Precision Journalism, 1973

Verifiable truth is a vital commodity these days. We are at a point in political history where "truth" is closer to "opinion" than it has ever been and it is more important than ever that journalists not only know how to check facts offered by others, as well as in their own work. Literacy in data - even just knowing how to examine a spreadsheet - is a key skill.

One other important thought worth adding at this point - data and statistics are often confused for one another. Raw data can be turned into statistics, but data can simply supply the journalist with ammunition for a question, or set of questions, to be posed to experts or politicians. Datasets often contain the seeds of stories which, once published, include no figures, no percentages and defnitely no statistics.

What can data do for Investigative Journalism

Why are we even doing this? If you look at Wikipedia's answer you will find this "Data-driven journalism, often shortened to "ddj", a term in use since 2009, is a journalistic process based on analyzing and filtering large data sets for the purpose of creating or elevating a news story. Many data-driven stories begin with newly available resources such as open source software, open access publishing and open data, while others are products of public records requests or leaked materials."

The Wikipedia article has recently been merged with the one on DataJournalism. The entry warns the reader that DDJ is "not to be confused with database journalism".

More frequently datajournalism is often confused with journalists "doing statistics" – some critics would say "journalists doing statistics badly". I think it's quite simple – we take datasets and use software to help us find facts or patterns which lead us to stories.

For example, we might take a large dataset and filter it to find companies with large government contracts which are owned by politicians, their families or friends. This isn't going to result in a story full of numbers and statistics, it's going to be a story with some names, and some eye-watering sums of money!

News is……

If you ask journalists why they took up the profession, many will make some reference to investigative journalism as a spur. (I would guess very few would claim to have been inspired by the idea of just writing down what people said and putting it in a paper or a news bulletin). I have always liked this definition – "News is something that someone, somewhere, doesn't want you to know. Everything else is just advertising".

In my mental notebook, I have attributed these words to Lord Northcliffe. But trying to confirm that he was the first person to say this takes one down an instructive rabbit hole – where similar ideas are attributed to a whole range of great figures in the field of journalism. Searching for similar definitions on Google one find things like "As Katherine Graham, the former publisher of the Washington Post, used to say: News is what someone wants suppressed. Everything else is advertising.

source:https://www.gov.uk/government/speeches/news-is-what-someone-wants-suppressed-everything-else-is-just-advertising

But there are yet more versions of the quotation out there. A little digression in search of the accurate quote is worthwhile:

News is something, someone, somewhere, doesn't want you to know. Everything else is [just advertising]

OR

News is something, someone, somewhere, doesn't want you to print. Everything else is [just advertising] Randolph Hearst

OR

Some people attribute something similar to George Orwell…. "Speaking the truth that somebody wants you not to publish is journalism. Everything else is marketing." though the questioner names Oscar Wilde as the possible source! Another user found (24 November 2012) a similar quote from Alfred Harmsworth, Lord Northcliffe: "News is what somebody somewhere wants to suppress; all the rest is advertising." And yet another user found another similar quote from Horacio Verbitsky (in Spanish):

"Journalism is to spread what someone does not want you to know; the rest is propaganda."

And finally - finally! - someone linked to the Quote Investigator page on this quote, from 20 January 2013.

(Quote Investigator, by the way, is an absolutely invaluable resource for answering quote-identification questions here: they do really in-depth research to puzzle out the true origins of a quote, and everything is properly sourced and verifiable).

So here's the "final" word on the matter:

> A version of this quote first appeared on 30 November 1918, in page 18, column 4 of The Fourth Estate: A Newspaper for the Makers of Newspapers, Ernest F, Birmingham, Fourth Estate Publishing Company, New York. The quote is as follows:
>
> "Whatever a patron desires to get published is advertising; whatever he wants to keep out of the paper is news," is the sentiment expressed in a little framed placard on the desk of L. E. Edwardson, day city editor of the Chicago Herald and Examiner.
>
> source: https://quoteinvestigator.com/2013/01/20/news-suppress/

Whoever defined news in this way is not really important - what matters is that our job should be to unearth the details which people, somewhere, don't want us to report.

With so much data now being made available online, many of the facts, data, and details which would previously have been hidden away, are now there for us all to use – hiding in plain sight.

All we have to do is to learn how to sift through them in the most efficient way, and join the dots to find the story which "somebody, somewhere, doesn't want us to know".

CONVENTIONAL JOURNALISM	INVESTIGATIVE JOURNALISM
Research	
Information is gathered and reported at a fixed rhythm (daily, weekly, monthly).	Information cannot be published until its coherence and completeness are assured.
Research is completed swiftly. No further research is done once a story is completed.	Research continues until the story is confirmed, and may continue after it is published.
The story is based on a necessary minimum of information and can be very short.	The story is based on the obtainable maximum of information, and can be very long.
The declarations of sources can substitute for documentation.	The reportage requires documentation to support or deny the declarations of sources.
Source relations	
The good faith of sources is presumed, often without verification.	The good faith of sources cannot be presumed; any source may provide false information; no information may be used without verification.
Official sources offer information to the reporter freely, to promote themselves and their goals.	Official information is hidden from the reporter, because its revelation may compromise the interests of authorities or institutions.
The reporter must accept the official version of a story, though he or she may contrast it to commentaries and statements from other sources.	The reporter may explicitly challenge or deny the official version of a story, based on information from independent sources.
The reporter disposes of less information than most or all of his sources.	The reporter disposes of more information than any one of his sources taken individually, and of more information than most of them taken together.
Sources are nearly always identified.	Sources often cannot be identified for the sake of their security.
Outcomes	
Reportage is seen as a reflection of the world, which is accepted as it is. The reporter does not hope for results beyond informing the public.	The reporter refuses to accept the world as it is. The story is aimed at penetrating or exposing a given situation, in order to reform it, denounce it or, in certain cases, promote an example of a better way.
The reportage does not require a personal engagement from the reporter.	Without a personal engagement from the reporter, the story will never be completed.
The reporter seeks to be objective, without bias or judgement toward any of the parties in the story.	The reporter seeks to be fair and scrupulous toward the facts of the story, and on that basis may designate its victims, heroes and wrongdoers. The reporter may also offer a judgment or verdict on the story.
The dramatic structure of the reportage is not of great importance. The story does not have an end, because the news is continuous.	The dramatic structure of the story is essential to its impact, and leads to a conclusion that is offered by the reporter or a source.
Errors may be committed by the reporter, but they are inevitable and usually without importance.	Errors expose the reporter to formal and informal sanctions, and can destroy the credibility of the reporter and the media.

We discover a subject.
We create a hypothesis to verify.
We seek open source data to verify the hypothesis.
We seek human sources.
As we collect the data, we organise it – so that it is easier to examine, compose into a story, and check.

Table from
Story Based Inquiry: a Manual for Investigative Journalists, Mark Lee Hunter & Luuk Sengers.
Source: https://bird.tools/wp-content/uploads/2020/01/SBI_Manual.pdf

Hypotheses, data and the story memo

Which comes first?

As journalists begin working with data and doing data-driven investigations they often ask: which comes first – the data or the story? There's no fixed answer: sometimes the story will come first – an idea, a hypothesis, a question to be answered, and the journalist will realise that data is probably available as one of the sources they can interrogate to get this story. So – story first.

But other times, the journalist happens upon data, or is perhaps given a dataset by a whistleblower, and the data leads to a story. So – data first.

The key point in either case is that the data is unlikely to be the only source of information, and the story is not necessarily going to end up being full of data, numbers or statistics. in some data-driven investigations the data may just provide the impetus for an investigation, the prompts for a set of questions which need asking. The questions might be posed of the data, or of people or the institutions they work for. Sometimes the data will give you the answers.

Shaping a data-driven investigation

When you begin an investigation, it's worth asking yourself three questions to decide whether it's worth the time and effort you are about to devote to a story:

Does it expose wrongdoing?

Does it inform public debate?

Is the story in the public interest?

Even if these are not the key questions you need to ask, you will need some guiding principles which will help you decide whether or not to embark on a potentially time-consuming journey.

The Story Memo

Once you (and your editor) have decided to commit, it's worth beginning a Story Memo – a living document which will help you remember what you are doing, and why. This is especially important when you are lost in thousands of details and questions and you are beginning to lose sight of what you set out to prove…

You need to keep in sight, your answers to these five questions. The answers may change as you gather the evidence and interrogate it:

Hypothesis	In a few words, what are you setting out to prove? What would be the headline or lede if you got the story you're aiming to get? The hypothesis may change over time as you make it more precise in the light of evidence available (but see also – Min/Max story)
Hook	Why this story, why now?
Questions to answer	What are the key questions you need answers to? This will help you identify people you need to interview, and data you need to interrogate. The list of questions will change a lot as the investigation progresses
Sources	The questions you need to answer will partly suggest the sources you need to consult. We will look at a method of generating a thorough list of sources later
Min/Max story	If all goes well, and you get everything you want, what's the best story you can hope to get? In theory this will be the answer to your hypothesis, the first question in the memo. But what if it doesn't pan out as you hope? What if not everyone will talk to you…? …the data doesn't answer the questions you hoped it would? What's the best story you can hope for if that happens, or your editor loses patience and says "just give me what you have now…"? Having a good minimum story to aim at will keep you going in the darker moments when you think you will never get it all done – "even if I can't prove xyz, I can still show that abc is true, and name the guilty parties"

Story Memo

Hypothesis

Hook

Questions to answer

Sources

Min/Max

What do you think is happening?

Write a one-line statement, or anticipated top-line, that you can work to prove or disprove.

Hypothesis

Hook

Questions to answer

Sources

Min/Max

Why this story and why now?

Does it meet the investigative criteria for relevance and impact?

Hypothesis

Hook

Questions to answer

Sources

Min/Max

What do you need to know in order to test your hypothesis?

What questions do you need to ask your sources?

Hypothesis

Hook

Questions to answer

Sources

Min/Max

If all else falls apart, what do you have left?

Is the minimum story still a solid publishable story that warrants the effort put in?

Exercises

It's worth practising the mental route from hypothesis to story using the story memo. If you're working alone it can be quite a struggle to come up with a workable hypothesis, ask the right questions and come up with maximum and minimum stories. Before you start trying to do this with a deadline, try some thought experiments -

Look at published stories - just the headline - use that as the hypothesis; imagine that you have to work on that story...fill in the rest of the story memo. Try to decide whether the published story is the maximum, the minimum, or somewhere in between.

If you need inspiration look at newspaper headlines, maybe the Pulitzer Prize winners' citations, and websites like the Global Investigative Journalism Network (GIJN.org)

"Interviewing data"

When talking to journalists about using data, many trainers draw the comparison with interviewing a spreadsheet as you would interview a person. It's an analogy which holds up well.

Firstly, data is just a source of information in the same way that a press release is a source, and so is an interview. What's more, an interview can be probing, or go into depth. An interview may be with a whistleblower who wants to tell you something but doesn't know how, or you may be trying to get answers out of someone who doesn't want to tell you what they know. Interviewing a dataset is very similar. Some data is so well organised that finding out what you need to know is simple, some is messy and disorganised, and it takes a lot of extra work to dig the information out.

Most importantly, the comparison with interviewing a person should serve as a reminder that we need to decide how much to trust them, we will almost always need to talk to other sources.

And, just as we do with interviews with people, we can go in unprepared, or well-prepared. We can interview someone just because they are there – perhaps a famous person who is offered to us by their PR team, or we can track someone down and ask them searching questions we have spent a long time preparing.

That said, when we ask questions of a dataset we will usually be doing the following:

- sorting
- filtering
- summarising

The key at this trick is to translate the questions we want to ask into the operations we need to perform to ask them. So:

- sorting – "what's the biggest, the smallest, the oldest, newest, most expensive…..?"
- filtering – "I only want to talk about what happened in 2020", or "I am only interested in events in Newcastle, between July and September last year" or "I need to know about purchases worth more than £50k, made by your company before 2015"
- summarising – "what was the total value of purchases worth more than £50k, made by your company before 2015", "how much did the Ministry spend on the top 5 items in total in 2020?"

In spreadsheets and analysis programs sorting and filtering are often called just that. Summarising is most easily and usually done by making a pivot table – a way of grouping together all records in your chosen category, adding up the numbers in them, and displaying them in a compact table.

Remember – the most important thing in all "computer aided reporting" is getting the computer to do as much of the work for you as you can – freeing your schedule and your brain to work on the story itself.

So – if you're doing something boring and repetitive – you're probably doing it wrong!

That is – for many processes there's a built-in tool or operator which will do it for you; if you don't know it yet, then use google search or youtube to find a more efficient way of doing what you need done. If you move on to script-based programs like Python or R you will start to learn what you yourself can automate to get the job done.

"Interviewing data"

When talking to journalists about using data, many trainers draw the comparison with interviewing a spreadsheet as you would interview a person. It's an analogy which holds up well.

Firstly, data is just a source of information in the same way that a press release is a source, and so is an interview. What's more, an interview can be probing, or go into depth. An interview may be with a whistleblower who wants to tell you something but doesn't know how, or you may be trying to get answers out of someone who doesn't want to tell you what they know. Interviewing a dataset is very similar. Some data is so well organised that finding out what you need to know is simple, some is messy and disorganised, and it takes a lot of extra work to dig the information out.

Most importantly, the comparison with interviewing a person should serve as a reminder that we need to decide how much to trust them, we will almost always need to talk to other sources.

And, just as we do with interviews with people, we can go in unprepared, or well-prepared. We can interview someone just because they are there – perhaps a famous person who is offered to us by their PR team, or we can track someone down and ask them searching questions we have spent a long time preparing.

That said, when we ask questions of a dataset we will usually be doing the following:

- sorting
- filtering
- summarising

The key at this trick is to translate the questions we want to ask into the operations we need to perform to ask them. So:

- sorting – "what's the biggest, the smallest, the oldest, newest, most expensive…..?"
- filtering – "I only want to talk about what happened in 2020", or "I am only interested in events in Newcastle, between July and September last year" or "I need to know about purchases worth more than £50k, made by your company before 2015"
- summarising – "what was the total value of purchases worth more than £50k, made by your company before 2015", "how much did the Ministry spend on the top 5 items in total in 2020?"

In spreadsheets and analysis programs sorting and filtering are often called just that. Summarising is most easily and usually done by making a pivot table – a way of grouping together all records in your chosen category, adding up the numbers in them, and displaying them in a compact table.

Remember – the most important thing in all "computer aided reporting" is getting the computer to do as much of the work for you as you can – freeing your schedule and your brain to work on the story itself.

So – if you're doing something boring and repetitive – you're probably doing it wrong!

That is – for many processes there's a built-in tool or operator which will do it for you; if you don't know it yet, then use google search or youtube to find a more efficient way of doing what you need done.

If you move on to script-based programs like Python or R you will start to learn what you yourself can automate to get the job done

PIVOT TABLES

What is a pivot table?

It's a feature embedded in most spreadsheet programs – notably MS Excel and Google sheets – which enables you to choose which variables (columns) to group together in order to summarise their contents. The name "pivot" refers to the notion of swiveling a table around your chosen point.

A pivot table takes the original data, but doesn't edit it. Instead it makes it possible for you to reshape it to meet your needs without having to do any copying or pasting of the data. Think of it as a kind of lens – you can focus on details, but you don't alter the data themselves.

Beware – when you first meet pivot tables and start to appreciate their power you may be tempted to pivot just about every datasheet you come across. But it's a bit like that saying "to someone with a hammer, everything looks like a nail". You can't, for example, pivot a pivot table – that is, you can't summarise something which is already a summary. (Well, you can – but you can't learn anything meaningful by doing this!)

How can you tell that something has already been summarised/pivoted? It's important to be able to see the difference between raw data and summary data. For example, if you can only see entries which are totals of something else then the chances are that you are looking at someone else's pivot table. Raw data will have the original entries, and, importantly, could usefully be pivoted.

For example – which of these tables show raw data, and which are the summaries/pivots?

Row Labels	F	M	Grand Total
blue	3	2	5
brown	2	2	4
green		3	3
Grand Total	**5**	**7**	**12**

	Column Labels			
	F		M	T
Row Labels	Average of age	Average of height	Average of age	Average of height
blue	59.33	1.48	45.50	1.58
brown	23.00	1.58	27.50	1.63
green			32.00	1.63
Grand Total	**44.80**	**1.52**	**34.57**	**1.62**

gender	age	height	eye colour
M	16	1.6	blue
M	23	1.7	green
F	34	1.65	brown
M	45	1.76	brown
F	86	1.6	blue
M	75	1.55	blue
M	60	1.7	green
F	85	1.65	blue
M	10	1.5	brown
F	7	1.2	blue
M	13	1.5	green
F	12	1.5	brown

Name	gender	age	height	eye colour
Brian	M	16	1.6	blue
Charles	M	23	1.7	green
Diana	F	34	1.65	brown
Eric	M	45	1.76	brown
Frances	F	86	1.6	blue
Francis	M	75	1.55	blue
Gerald	M	60	1.7	green
Gertrude	F	85	1.65	blue
Henry	M	10	1.5	brown
Henrietta	F	7	1.2	blue
James	M	13	1.5	green
Jemima	F	12	1.5	brown

That's right – the last two tables show raw data. Even without the names, the 3rd table still contains only raw details – whereas the first two contain summaries. So trying to pivot the first two tables is not going to get you very far – somebody else has already done that step.

Take another example - you open a table that you have downloaded and you see something like this –

Weekly pay - Gross (£) - For all employee jobsa: United Kingdom, 2019											
Description	**Code**	**thousand)**	**10**	**20**	**25**	**30**	**40**	**60**	**70**	**75**	**80**	**90**
All Employees	female part time	5,615	65.2	109.7	128.2	142.0	170.8	232.0	267.1	292.0	327.4	451.8
All Employees	male	13,331	234.0	357.3	392.7	427.7	498.1	661.5	766.6	829.7	914.3	1,193.4
All Employees	male part time	1,946	53.8	93.4	111.9	129.7	160.5	213.3	246.3	270.6	302.8	440.0
All Employees	female	13,373	124.2	193.1	229.0	262.9	331.0	457.3	542.9	598.1	668.3	854.0
All Employees	female part time	7,758	327.7	370.2	392.6	416.8	467.8	598.9	689.9	741.0	798.1	988.9
All Employees	all full time	19,144	345.0	400.1	428.6	458.4	517.5	669.1	763.5	820.4	896.3	1,151.5
All Employees	all part-time	7,561	61.8	104.9	124.1	138.0	166.1	228.5	261.8	287.2	321.1	449.5

Ask yourself – even if I could pivot it, what would the result look like? Put another way – what questions could I ask it? That's right – someone else, in this case the Office of National Statistics, has already summarised the original data for you. Just about the only thing your pivot of this table could tell you would be the grand totals for each row and column – which you can work out from the table itself.

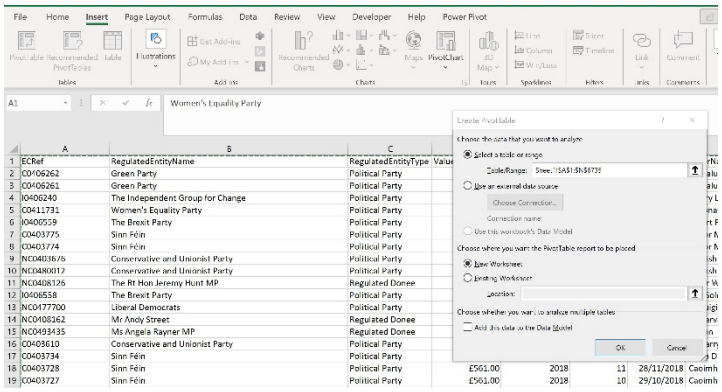
Making a pivot table

A quick reminder of how to make a pivot table. Slightly annoyingly, Excel and Googlesheets do the same thing in slightly different ways, so here are the two different methods –

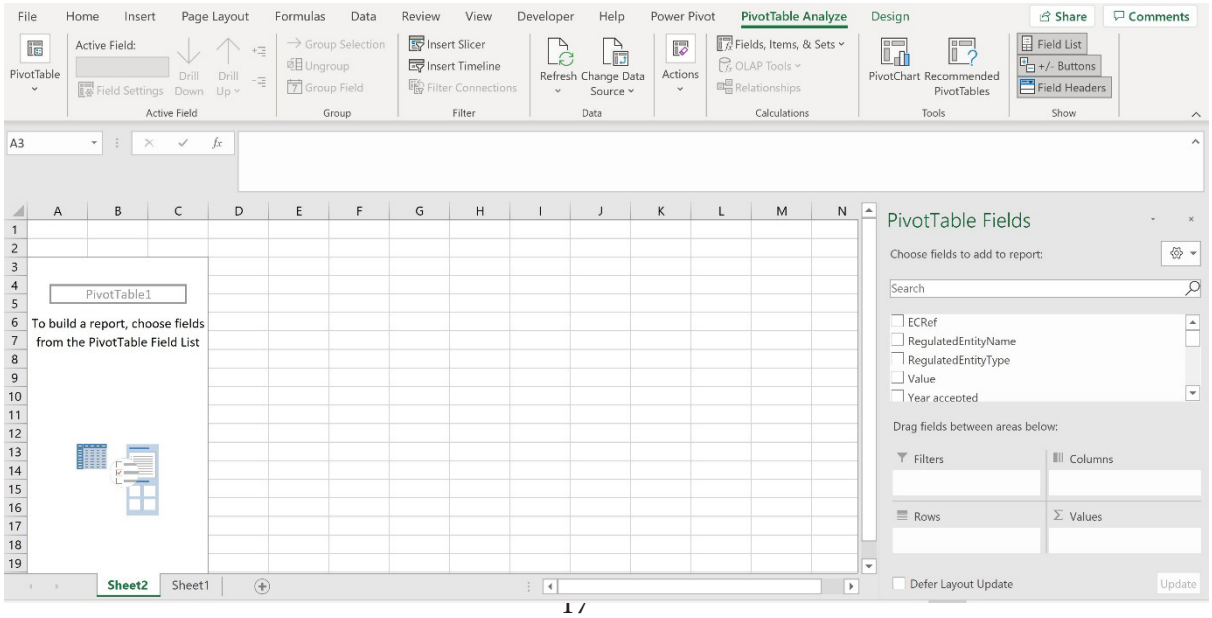
Excel

From your main data worksheet, click on Insert, and then choose Pivot Table – the icon on the far left near the top of the screen. (NB – don't be tempted by "Pivot Chart" in the middle of the same menu: it will take you on a rather confusing route to the same place!)

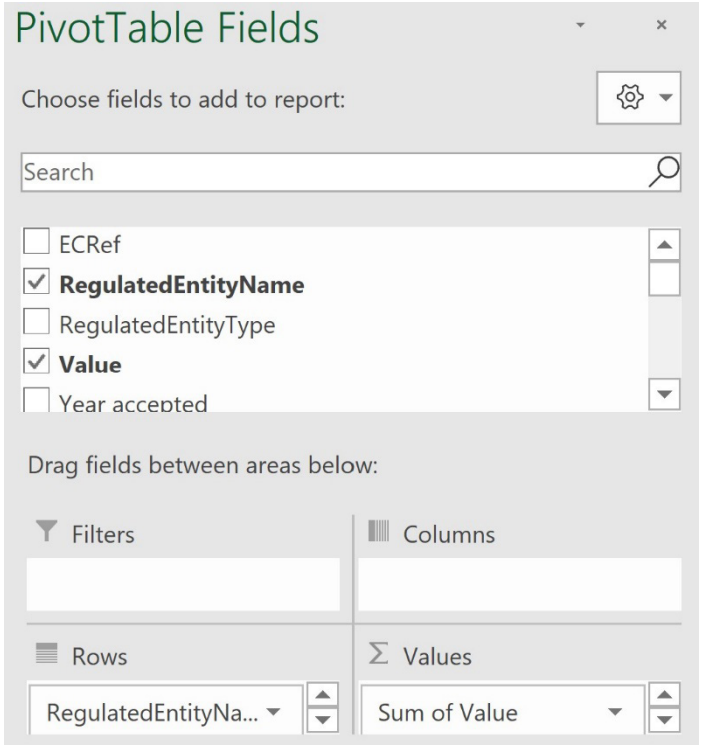
You should see a little dialogue box as below – make sure the cells detailed in "Table/ Range" are the whole table, not a part of it. You cannot edit this box by hand: if the entry is incorrect, you have to cancel and return to your datasheet, and start again. Make sure you only highlight one cell in the data you intend to pivot.



When you click OK you should see a new worksheet which looks like this –



The key to remaining sane, and knowing at all times what variables you have put where is the PIvot Table Fields selector:



To summarise all payments to political parties (“entities”) we drag the “Regulated Entity Name” entry in the PivotTable Fields chooser to Rows. And “Value” into the Values box:

Row Labels	Sum of Value
Advance Together	62500
Alan Mak MP	11683.81
Alec Shelbrooke	2000
Alex Sobel	1502.46
Alison McGovern MP	13750
Alliance - Alliance Party of Northern Ireland	288137.26
Alun Davies	2431.71
Alyn Smith	1852
Andrea Jenkyns MP	25628.31
Angela Eagle MP	5008.32
Anna McMorrin	9941.43
Ashfield Independents	8400
Barnet Labour Group	10000
Ben Bradley	19500
Blue Collar Conservatism Ltd	48000
Brendan O'Hara	2006.19

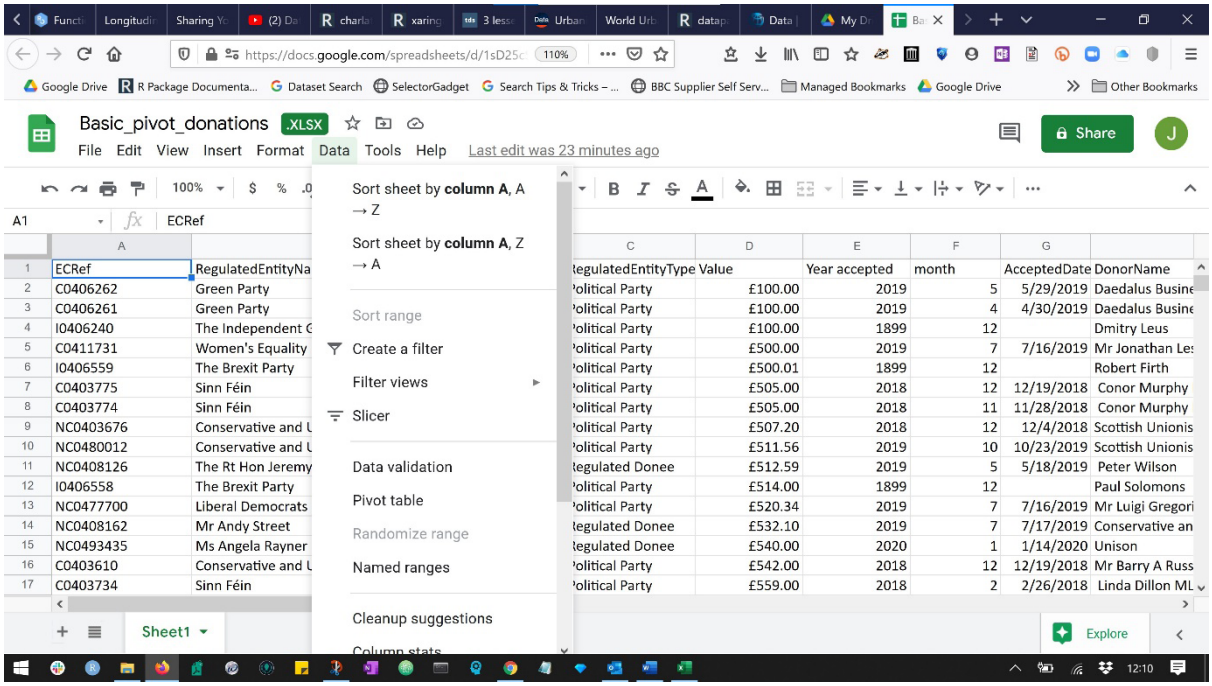
Then to sort the new Sum of Value column into descending order, right click on any one of the totals in that column to bring up a little menu containing the option to “sort highest to lowest”.

Remember – summary totals like this are clickable – double clicking on a total in a pivot table will ask the program to make a new sheet made up of all the numbers which make up that total – along with the other information associated with those payments.

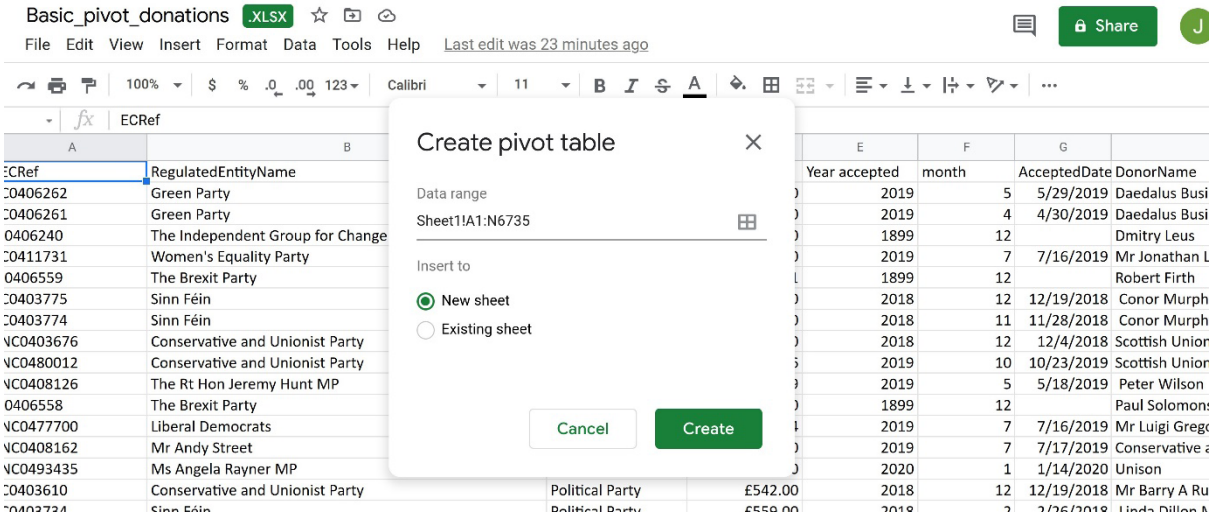
To go back over the basics of pivot tables with a demo, and an opportunity to watch short video demonstrations of the various techniques, click here

To make a Pivot Table in Googlesheets

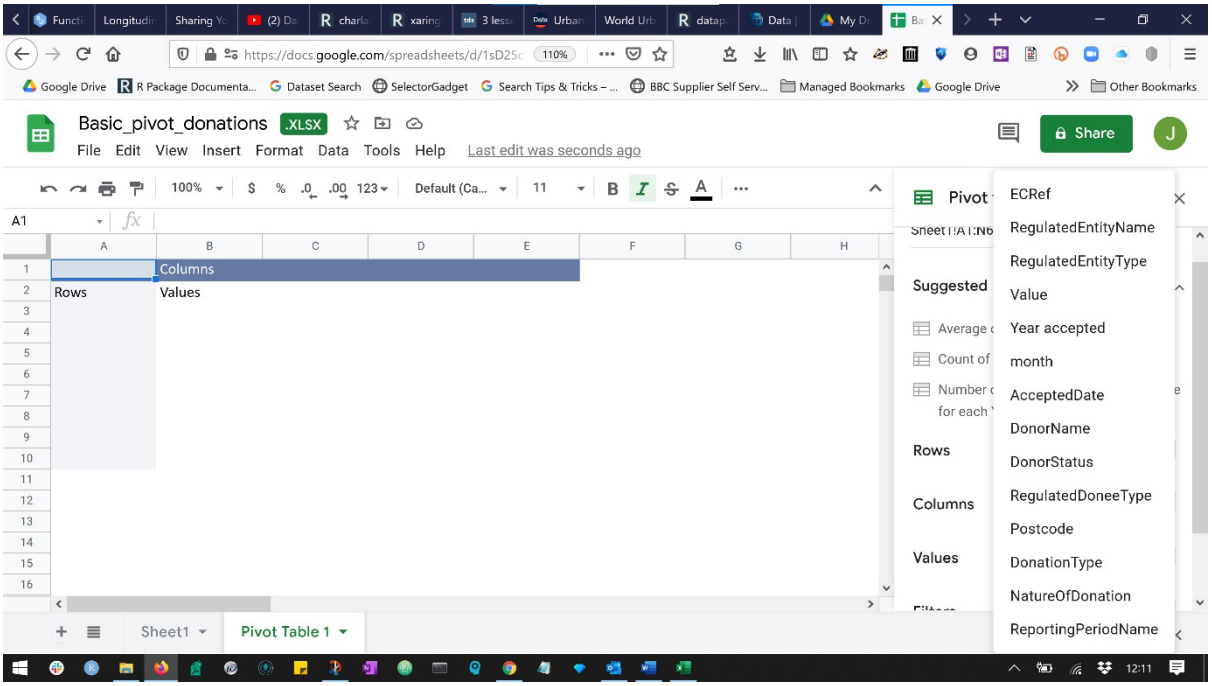
From your main data worksheet, click on Data, and then choose Pivot Table – about 2/3 of the way down the menu which appears.



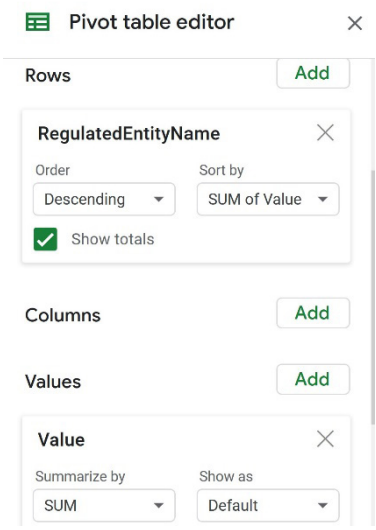
You should see a little dialogue box as below – make sure the cells detailed in "Table/ Range" are the whole table, not a part of it. You cannot edit this box by hand: if the entry is incorrect, you have to cancel and return to your datasheet, and start again. Make sure you only highlight one cell in the data you intend to pivot.



When you click OK you should see a new worksheet which looks like this –

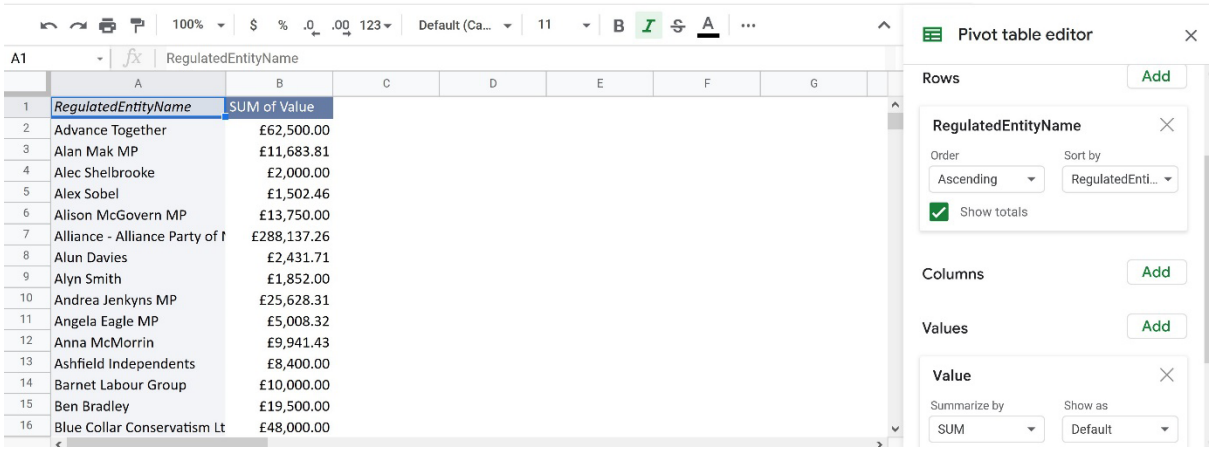


The key to remaining sane, and knowing at all times what variables you have put where is the PIvot Table Editor

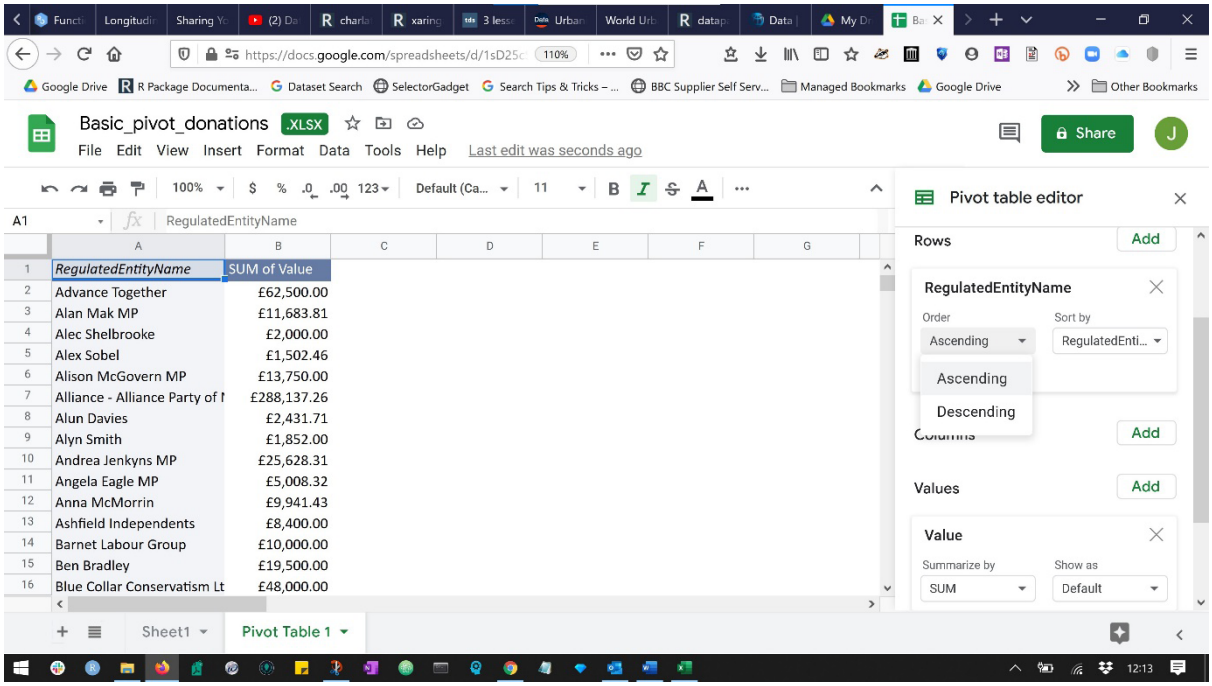


To summarise all payments to political parties (“entities”) click the Add button next to

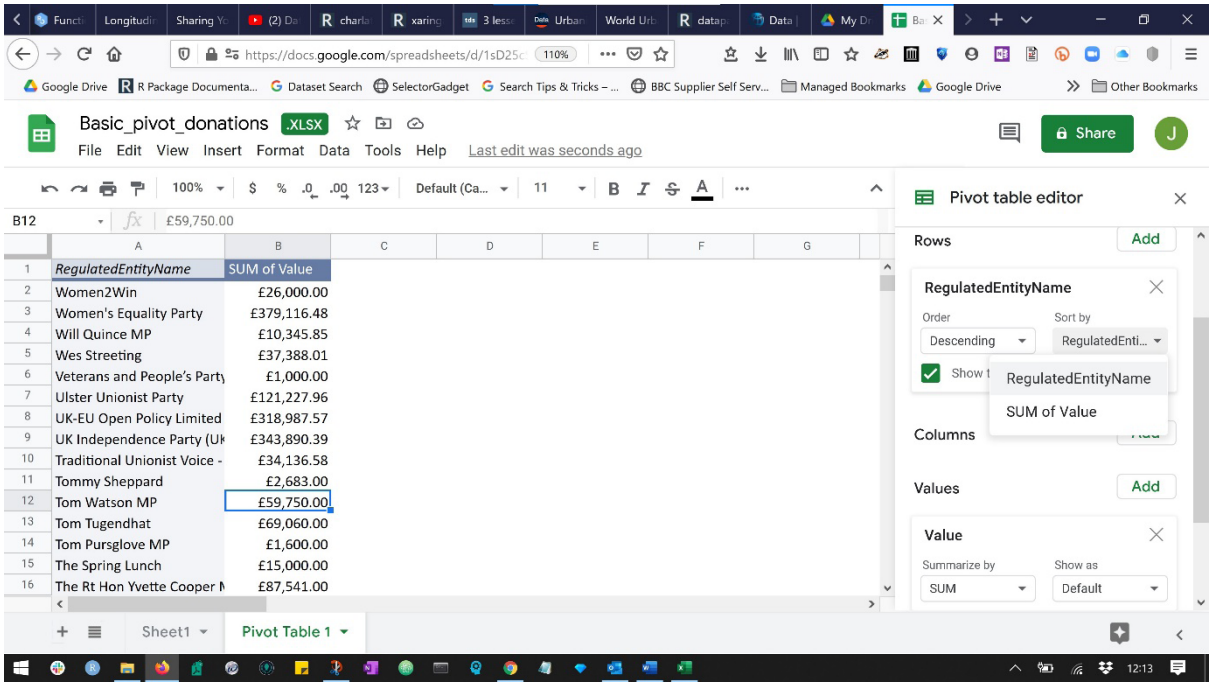
Rows in the Pivot Table Editor (right hand side of screen) and select “Regulated Entity Name”. Next to the Values box, we click on the Add button and select “Value”.



You should see a screen like this. To sort the results in descending order we need to go back to the Pivot Table Editor and change “Order”, in the box called RegulatedEntityName to Descending.



“Sort by” needs to be changed to “SUM of Value” :



Remember – summary totals like this are clickable – double clicking on a total in a pivot table will ask the program to make a new sheet made up of all the numbers which make up that total – along with the other information associated with those payments.

Further exercises

To go back over the basics of pivot tables with a demo, and an opportunity to watch short video demonstrations of the various techniques, click here

Incident Number
Crime code
Crime Description
Date Occurred
Time Occurred
Arrestee Name
Race
Gender
Arresting officer
Arrest Type
Arrest Location

For practice, do try to think of questions for your data – either the example in the exercise above, or your own. Don't just wade in and start clicking – make notes of questions as you would in preparing an interview with a person. If you don't, it's just too easy to get lost down rabbit holes and dead ends.

If you need more help on Pivot Tables don't forget that youtube is full of tutorials. The quality is mixed, of course, but you will almost always get visual help on the problem you need to solve.

The Atlantic

In Champaign-Urbana, Illinois, 89% of Those Arrested for Jaywalking Are Black

REBECCA J. ROSEN AUGUST 24, 2012

The power and promise of open data: A Freedom of Information Act request puts some numbers behind the charges of police discrimination.

RECOMMENDED READING

The Secret Internet of TERFs

KAITLYN TIFFANY

Dry Ice Is Hotter Than Ever

CHARLES FISHMAN

How Will the Future

It's always a good idea to start by looking at the variables (columns) you have in your data to help you think about the stories they could contain.

In the Champaign, Illinois arrest data the variables included - date, location, time of arrest, arrest description, name, age, and race of arrestee, badge numbers of arresting officers.

With that information you could work out the most common offence for that year, the busiest date, the location where most arrests are made just by analysing a single column for each story.

Then add a second and third variable and you can start answering much more complex questions - which arresting officer arrests most women of colour, and for what offences? Which officer makes most arrests by day, or by night, in a certain area of the city?

Interviewing data

https://features.propublica.org/walking-while-black/jacksonville-pedestrian-violations-racial-profiling/

The pictures on this page are examples of how the "Walking while Black" meme grew from data driven investigations like the one done in Champaign, Illinois, which we used in class: the "Improper Walking on Roadway" charge was not top of the league for arrests that year, but by some distance saw the most people of colour arrested as a proportion of all people charged with that offence.

Sources

Key steps in finding data

Think about it first

- What would the data contain? What variables might there be? What would they be called? What you call “spending” might be “expenditure” or “total amount”.
- Bear in mind that whenever you fill in a form online – such as registering your vehicle, or applying for a new passport, you are adding to somebody’s database. One day you may want to interrogate that database en masse – so make a note of who owns it, and what the input terms are – the language used on a web form is likely to become the names of the variables in the database itself.
- Who would gather and keep the kind of data you want? A government ministry? A national statistical office? A watchdog? An international organisation? The answer to this question will help you as you start to look for it.
- How can you get hold of the data? A lot of data is held as files on web portals for you to download, but an increasing number of datasets are held in databases which you need to interrogate via a webpage in order to generate a file you can download and work with.
- In the UK the Electoral Commission and the Land Registry are examples of sites where you “ask” for the files. Likewise, in the USA there are site like USASpending.gov which use the same interface system. In looking for them, this kind of site is not going to be findable using the “filetype:xls” google operator – you should try using a search string like “site:gov.uk intext:download” (along with a suitable search term, of course).
- For a complete guide to Google Operators – go to googleguide.com

Finding people

If you can't find the data anywhere, maybe you should try building your own database – you could create a google or Microsoft form, and invite your audience to supply details. When The Guardian wanted to track deaths at the hands of the police in the USA, as well as using open datasets, they asked their readers to send them details of killings which hadn't made it into the papers. The Counted was the resulting project.

Sources – mapping the territory

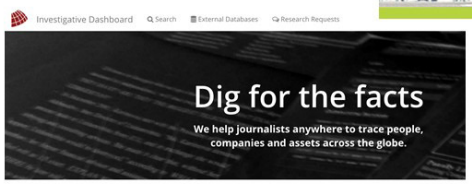
A really useful way of finding evidence, including data, is to "map the territory". This process will help you identify a whole range of possible sources you might not otherwise think of. It involves thinking of every step in a chain of events, and listing the sources, documents, people and data which could be involved at each stage.

For example – let's look at a controversial new government benefits scheme for people with disabilities. You want to prove that it's costing lives and wasting money. Consider all the links in the chain –

- Someone wrote the policy and it was approved by civil servants and/or parliamentarians.
- People apply for the benefit – there's probably an online form to complete. Anytime there's an online form, there's almost certainly a database of responses.
- Civil servants consider the application. Someone designed a questionnaire for applicants, maybe also an algorithm to help decide the level of payments.
- People are given the benefit, or are denied it, or given a partial award. Some react well, others badly. Some may even take their own life, or die as a result.
- Local media and social media may contain news of those deaths
- Coroners look into causes of death and may issue "prevention of future deaths" recommendations
- People who die as a result of the policy have relatives. Some post tributes and messages on social media
- Someone proposed the policy – either a government department, or sometimes a think tank or policy group
- Some civil servants may have misgivings about the policy and be trying to change it. If it's not being changed, someone is responsible for rejecting proposals for reform
- Have their been questions in parliament about the policy? What were the answers?
- Some people may be happy with the benefits they are receiving. You need to find them too!

See if you can think of more links in the chain. It helps to draw it as a chain, and try to make it as complete as possible.

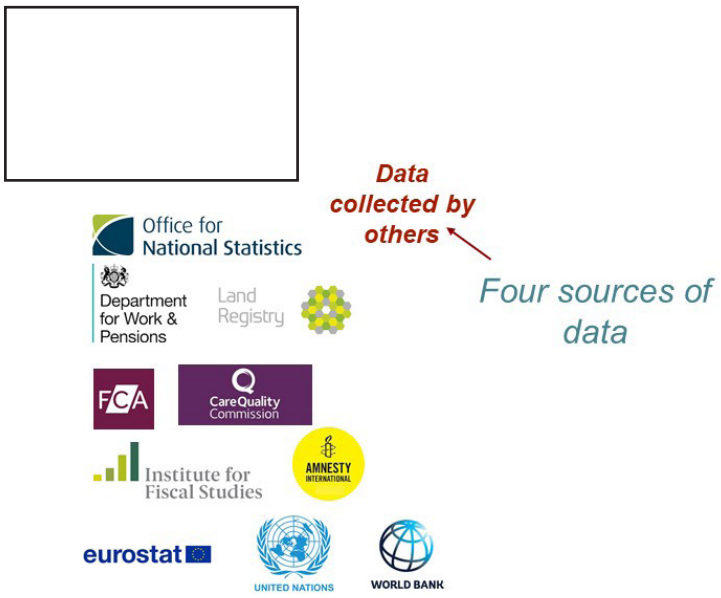
Company and charity information



Sources

Map your territory - take a look at the list on the previous page. Think about the territory for your own story idea.

What extra potential sources and questions to be answered does the exercise give you?



Sources

Data collected by others

Requested data

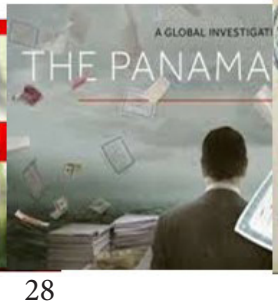
Four sources of data

- *Bespoke requests to official stats bodies*
- *Freedom of information requests*
- *Academics*
- *Third parties who collect data i.e. Local Data Company*

Data collected by others

Four sources of data

Leaks



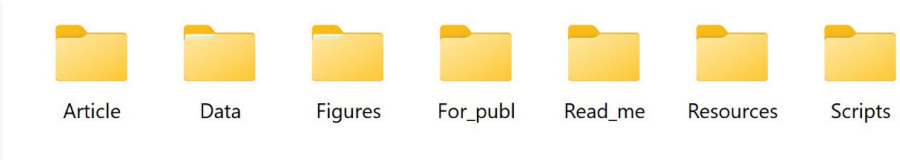
Workflow – keeping your investigation organised

As your investigation proceeds, you will build up more and more material, some of it written notes, some electronic, maybe audio and video recordings.

Keeping track of everything is vital. It will come as little surprise that I favour a spreadsheet – since you can interrogate a spreadsheet quite flexibly – looking at all conversations with one source one minute, then putting all your audio files in date order the next.

The most important thing is to stay organized. Make your story memo a key document and refer to it often.

Investigators at ***Der Spiegel*** in Hamburg talked about this folder template they use for every project:



The folders remain consistent for every project. Article contains all drafts of the article, Data contains all raw data for it, Figures contains the statistical analyses of the data, For_publ contains the finished article(s) (there may be online and print versions). Read_me is where the team puts notes for other team members and their future selves. Resources contains raw materials, contact lists, interview notes etc. And Scripts is for any computer scripts – R or Python, typically, which they use – scripts can, in turn, constistently refer to the Data folder.

Your Story Memo is a living document – you should update it as more facts come to light, and more questions suggest themselves. Even the "maximum and minimum stories" may change – but the minimum story should remain a beacon of hope to remind you, when everything is overwhelming and you being to wonder whether the investigation is worth all the time you are spending on it – "if it all falls apart, I still have the minimum story!"

For another, fuller, take on organising an investigation, read the CIJ Logan Handbook #1 ***The Hidden Scenario*** by Luuk Sengers and Mark Lee Hunter.

Dealing with Numbers

Let's start with a note I wrote about datajournalism and numbers back in 2016:

I got into data by accident. After 20 years in the BBC, I was working as a freelance trainer when I happened to talk to someone who was working on a complex data-driven project without knowing how to organise the data they were collecting.

After years of using data and spreadsheets as a management tool I retrained myself to use spreadsheets as a journalist – to ask questions of the data, and know what the answers meant. As I did so, I soon realised that there was a huge ecosystem data journalism I hadn't been aware of.

Up until 5 years ago or so, the situation in the UK reminded me of where we were when the web was just taking off: some journalists were rapidly learning to code, others preferred ready-made tools, and others were not quite sure what data could do for them, or what they could do with data.

That has changed since I first wrote this short intro in 2016 – but one remaining problem is a general confusion between data and statistics. Journalists tend to come from arts backgrounds and many are afraid of numbers and statistics. That still needs to change.

One thing which has not changed is that a huge amount of UK government data is being published with barely anyone noticing. This contains potential stories, leads to stories, evidence for investigations, evidence of what is really going on in society, rather than what we are told is happening. At the other end of the scale government departments are failing to open up useful data, or not publishing it fast enough. That failure is a big story too!

Meanwhile we are seeing the rise and rise of the infographic and data visualisation in general. On bad days I am reminded of the early days of powerpoint – people just pressing buttons to make use of the bells and whistles without paying attention to what it is they want to say.

There is still a challenge of how to raise data literacy, and statistical literacy, of journalists and audiences alike. When that happens data journalism will be truly mainstream – normal journalism driven by data telling vitally important stories based on facts and evidence. We all need to be up to data.

Big numbers

News > UK > Home News

NHS to ban the use of pagers within the next three years

Pagers are costing the NHS £6.6 million annually

The Independent

But how big is it really?

£6,600,000 / 150,000 doctors in UK hospitals

= £44 per doctor per year

Perhaps you consider yourself "number-phobic" or averse to maths. One colleague has wondered aloud whether journalists are allergic to numbers.

You cannot escape numbers, but there's no need to be afraid of them. We will, in the course of our data crunching come across numbers – raw figures, percentages, totals, averages and so on. And these numbers may well be essential to the story we are trying to tell. So it's worth thinking of a few simple rules to help you in assessing the importance of the figures you have, and therefore the place they will have in your story.

William Blundell gave similar advice on handling numbers in his book ***"The Art and Craft of Feature Writing":*** *We know that too many numbers are poison, so the writer's first impulse should be to omit unessential ones. But that's as painful as root-canal work to those of us whose days are filled with numbers-numbers vital to the breaking corporate and financial news stories inside The Wall Street Journal. In many such stories, numbers define the news or are the news.*

It's only natural, then, that some writers come to believe numbers per se possess a magical power of definition. They collect statistics by the bushel and, having gone to such pains, use them at the slightest excuse in every story they do. Then they wonder why editors find their features stupefyingly dull.

A wrenching change in attitude is required when doing features, where story values must be defined in other than just numerical ways. This change is easier to make if the writer remembers that fiction writers, who could bury us under invented figures to lend definition to their tales, never do so. They know better.

I don't imply that we should omit meaningful statistics to avoid boring readers. This would sacrifice substance for form, an error that a novelist may get away with but that we cannot. We need numbers in almost all our stories, and in some a number may be so important or startling that omitting or generalizing it would weaken the whole piece. I only argue that we be choosy in selecting figures and careful in their treatment.

In placing numbers in a story, the good writer tries not to stack too many in one paragraph; this builds a wall of abstraction difficult to breach. It becomes impossible to breach when two or more such paragraphs are butted together, a construction that may lead to more unread prose than any other writing fault. Don't do this. Don't ever do this.

The good writer also recasts as many numbers as he can in a simpler or more pictorial form that removes some of their abstraction. If a precise figure is not important, he rounds it off: $2.6 million is cleaner and easier than $2,611,423. If something increased by 36.7%, he may say it went up more than a third. If it increased 98%, he says it almost doubled. These expressions are pictorial in that they let the reader visualize a slice of a pie or two pies where there was one before.

Use ratios to simplify large numbers. Instead of saying that 14,654,231 American drivers out of a total of 58,013,261 own foreign cars, a writer may simply say that one in four owns a foreign auto. Smaller numbers can be grasped while large ones remain abstract.

"Spending on redundancy research by the Office of Unessential Affairs rose from $847M in 1983 to $1.26bn this year – a 49% increase"

"Over the past fiscal year the OUA increased spending on redundancy research by almost half, to $1.26bn"

So – when you're handling numbers or statistics, try to bear these rules in mind:

Who's counting? Who's asking?

Numbers may appear to be hard objective facts, but it matters who is doing the counting, what they are counting, and why. How many children live in the UK? Define a "child". Age may be the main factor: Are we counting under 18, the legal age of adulthood? under 16 – the point at which a young person can have sex, get married, join the milirary (not necessarily all at once) ? under 10 – an age at which behavioural scientists consider someone to become aware of rights and wrongs and so become capable of criminal responsibility? Maybe it's not a question of age, but of responsibility or independence?

Perhaps it will help to know why you're counting children – are you deciding how much financial support to award families depending on the number of children in a household? Perhaps you're trying to work out the number of children living in poverty – and if you want to make the number seem low, you might decide to define a child as someone still at primary school. If you want the number to seem higher, you count everyone under 18.

How big is the number? Compared to what?

The government announces a new initiative, to which it is allocating £50 million. Sounds a lot to me as an individual citizen. But let's look at the detail – the allocation is £50 M over 5 years, so £10 M a year. And the number of people affected by the initiative – let's say 25 million. So that's 50p a head – 10 p a year over the 5 years. Now how big does the sum seem?

Going up? How fast? And for how long?

Most figures don't live in isolation for a moment in time – they are part of a pattern: numbers go up, and they go down. Some go up rapidly, others fall gently. We need to compare figures to other figures – spending on health, say, is higher this year than it was last year. During the Covid pandemic, for example, it has been useful to get a handle on death tolls by comparing the number of people who died in one month with the same month a year ago. But is one year enough? Might it be even better to compare with the same month in five previous years? Or ten?

How far we look back to compare figures will often depend on what figures are available.

But it's important to avoid accusations of cherry-picking. It's often tempting to look at the performance of a government since it was elected – but many trends take a long time to change – unemployment, say, or inflation, won't go up or down the day after a new government is installed or a new policy is announced. When would be a good time to say the new policy has, or hasn't, worked?

Cause and effect

Talking of trying to work out when a new policy began to bite, the whole question of calculating cause and effect is worth a book of its own.

To put it at its bluntest – just because two variables appear to be linked doesn't mean that they are. It's easiest not to get involved at all – let your audience make the connection if they want, but don't try to work it out for them. In Michael Blastland's book, *The Tiger that Isn't*, the author uses the example of a housing estate in the English West Midlands where fourteen people died of a rare brain cancer, and the cause seemed obvious. As the Daily Mail reported in 2008 "14 die of cancer in seven years living next to phone mast with highest radiation levels in UK. Fourteen people living within a mile of a mobile phone mast that emits one of the highest levels of radiation in the country have died of cancer. Four of the deaths have been in a cul-de-sac yards from the site."

Tragic though the case is, and although the cause seems obvious, the radiation from mobile phone signals is not necessarily to blame. Fourteen deaths in seven years seems too many to be a coincidence. But in statistical terms it's just not "significant". The word "significant" is a statistical term which may sound callous when discussing dead people.

I am not going to discuss statistical significance in detail here, but the crucial paragraph in the relevant Wikipedia entry says –

"To determine whether a result is statistically significant, a researcher calculates a p-value, which is the probability of observing an effect of the same magnitude or more extreme given that the null hypothesis is true. The null hypothesis is rejected if the p-value is less than (or equal to) a predetermined level, α . Alpha is also called the significance level, and is the probability of rejecting the null hypothesis given that it is true. It is usually set at or below 5%."

In other words, if we would need to set out a hypothesis about the mobile phone mast's putative link with cancer. The hypothesis would need to specify proximity to the mast, and a period in which the people living there were to develop, or not develop, cancer. (We would probably also need to talk to oncologists to decide which cancers it might, reasonably, be linked to. And so on)

And then we would need to compare our chosen population with another population, of

about 700, with similar demographics, NOT living close to a phone mast (preferably several sample populations, some close to masts, others not.)

When all the necessary data has been collected and the analysis done, we would expect to see at least a 5% difference in the incidence of cancer between the populations living near masts, and those not living near masts.

Sound complicated? It's certainly a lot more rigorous than simply declaring the mast to be the cause of 14 cancers in 7 years.

As a datajournalist with no statistical qualificiations, when you see events which seem as clear cut as this one, your first thought should be to doubt the link, and to consider the possibility, the likelihood even, that for every housing estate with a horror story like this one, there will be others where there is no cancer at all.

-

Here's another contentious and complex example:

Do speed cameras prevent accidents, or cause them? To answer this question, what would you need to know? What data would you collect?

- Details of the roads where cameras were installed after a spate of accidents?
- Numbers of accidents afterwards?
- But how long do you want to go back before the cameras were put in?
- And how long after? Days? Months? Years? How many?
- And, anyway, how many accidents is a "spate"? 3? 4? 10?
- Within how many metres of each other?
- And within how many days or months of each other? Whatever you decide, you should then look at similar stretches of road where there were accidents, and no cameras were installed. Did numbers go down or up?

The most obvious problem in deciding whether a particular camera reduced accidents is the "regression to(wards) the mean" - a spate of accidents on a given stretch of road, can reasonably be expected to be followed by a safer period, of fewer to zero accidents even without installing a camera. Clearly there are some places which are dangerous, but the problem then is spotting the true danger spots within a mass of data. (Not to mention the ethical issue that arises if you find two danger areas and pick one to have a camera, and one to remain without a camera just in order to work out whether the presence of a camera makes the spot safer or more dangerous!)

Back in 2010 the RAC Foundation published a report by Professor Richard Allsop entitled The Effectiveness of Speed Cameras. Prof Allsop made decisions about the data to use along the lines described above.

It's worth reading the report in full. But here is an extract from the conclusions:

> But after allowing for all these factors, the judgement can be made that in the year ending March 2004, camera operations at more than 4,000 sites across Great Britain prevented some 3,600 personal injury collisions (PIC), saving around 1,000 people from being killed or seriously injured (KSI):

Type of site	Number prevented in year ending March 2004	
	PIC	KSI
Fixed urban Fixed rural	Between 1700 and 2200 Between 170 and 300	Between 500 and 560 Between 60 and 140
Mobile urban Mobile rural	Between 1000 and 1400 Between 180 and 300	Between 150 and 400 Between 90 and 200
All sites	Between 3050 and 4200	Between 800 and 1300

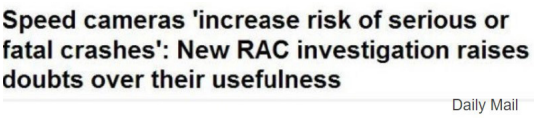
A lot depends on where you start and end the quotation from the report...

> Some of the reduction might also have been attributable to drivers diverting to avoid cameras, but the overall reduction might well have been greater had it not been for some collisions being caused by drivers suddenly braking and then accelerating in the vicinity of cameras.
>
> But after allowing for all these factors, the judgement can be made that in the year ending March 2004, camera operations at more than 4,000 sites across Great Britain prevented some 3,600 personal injury collisions (PIC), saving around 1,000 people from being killed or seriously injured (KSI):

Type of site	Number prevented in year ending March 2004	
	PIC	KSI
Fixed urban Fixed rural	Between 1700 and 2200 Between 170 and 300	Between 500 and 560 Between 60 and 140
Mobile urban Mobile rural	Between 1000 and 1400 Between 180 and 300	Between 150 and 400 Between 90 and 200
All sites	Between 3050 and 4200	Between 800 and 1300

The evidence and the judgements of the data's significance is weighed up and expressed with great care. So carefully, you might argue, that people on each side of the debate could use it as evidence for the side they took. Take your pick – The Guardian reading was that cameras save lives:

The Daily Mail seems to have read a different report...



Two diametrically opposed views from the same reporting of the same data!

(As you can see in the RAC Foundation post in the first link, Prof Allsop re-examined the question in 2013 and 2019. And you will see how he and others continue to consider the available data in different ways. We don't have space to consider all the details here, but you may want to practise with some of the datasets linked to from the RAC Foundation site.

If you remember nothing else, remember this – data can tell us what has happened. It cannot tell us why. And if you are tempted to link one event with another, get a statistician to advise you on what conclusions you can safely draw from the data you have.

Further reading

Michael Blastland – The Tiger that Isn't

Daniel Kahneman – Thinking Fast, Thinking Slow

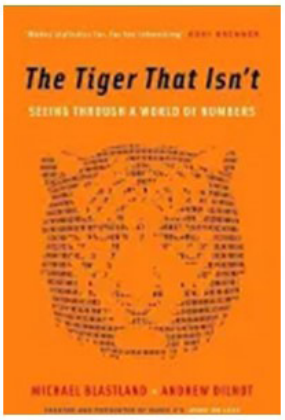
Daniel Levitin – A Field Guide to Lies and Statistics

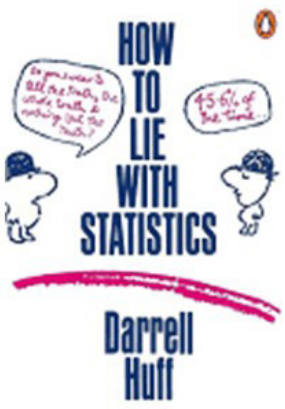
David Spiegelhalter – The Art of Statistics

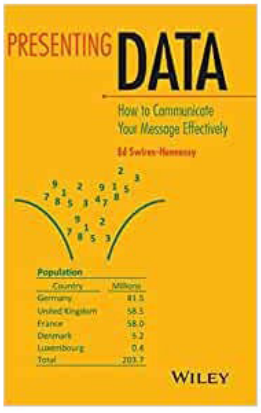
If you remember nothing else, remember this – data can tell us what has happened. It cannot tell us why.

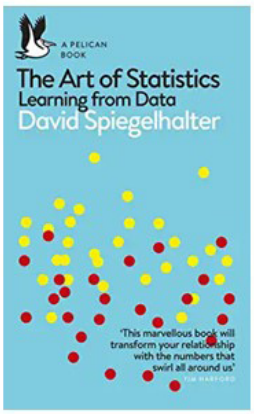
And if you are tempted to link one event with another, get a statistician to advise you on what conclusions you can safely draw from the data.

Books

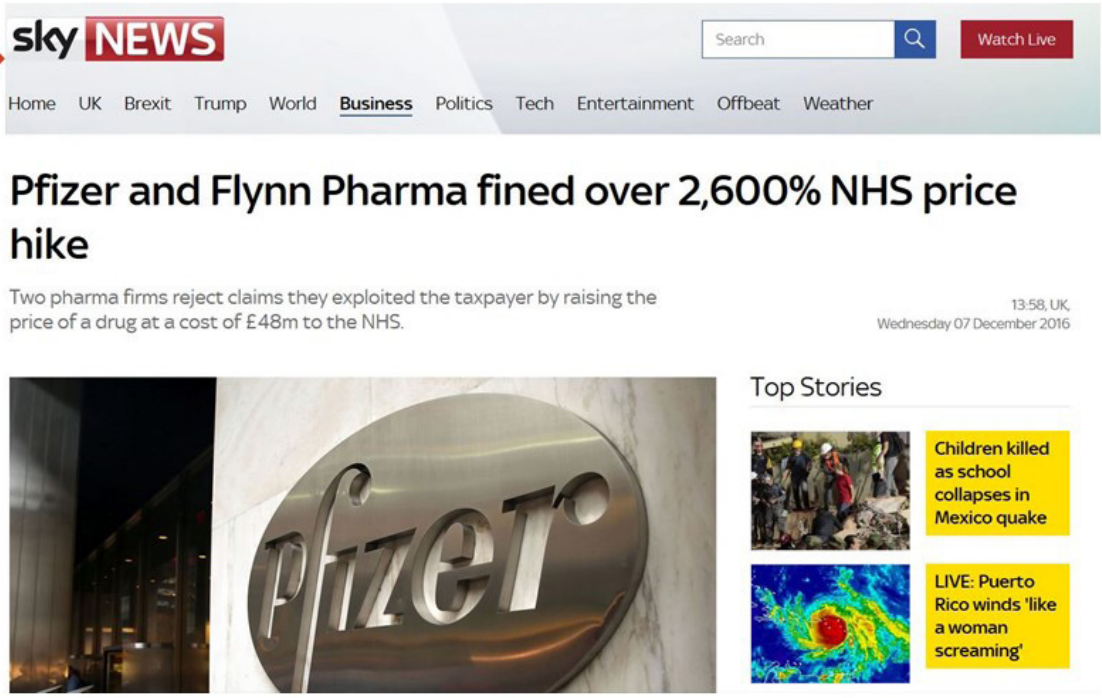








Percentages



Our advice is - don't be tempted to use percentages over 100%, especially when they get to this eye-watering level. You're making the reader do the maths, and you don't want to do that - keep it simple: prices have gone up from 50p to £25

Here the problem with the percentages is different - a 20% crime increase sounds worrying, but when you realise that the 20% increase is the same as 6 more crimes in a month that could be a blip, it could be the start of a seious increase, but sensationalising it by claiming it's a 20% increase is misleading - even in an effort to sell more papers! If you see a rate without the raw numbers, ask yourself why!

Percentages

*"Crimes reported in Honiton in December 2017 rose **by more than 20 per cent** compared to the same month in 2016."*

But...

Police statistics show **34 crimes** were recorded in December 2017 - including six reported incidents of criminal damage. Figures for December 2016 show **28 crimes** were recorded.

MIDWEEK Herald

Is a percentage really necessary?
Does it help the reader?

FOI

Who can you ask for information?

Who can you ask?

10 DOWNING STREET
LONDON SW1A 2AA

Department for Business, Energy & Industrial Strategy

Lewisham

CORNWALL COUNCIL

FCA

Ofcom

SPORT ENGLAND

Forestry Commission

Central Government departments

Local Government departments

Regulators

Quangos

LFB
LONDON FIRE BRIGADE

NHS

Public service bodies

TATE

HERITAGE FUND

Museums

The Royal Mint

hs2

Private firms owned by government

- *Correspondence (including emails and text messages)*
- *Internal reports, studies and audits*
- *Memos and diplomatic cables*
- *Images and plans*
- *Meeting schedules and minutes*
- *Audio and video recordings*
- *Data!*

Common Exemptions

Absolute	**Qualified**
Secret services/ Official Secrets	Formulation of policy
Most of the BBC	Information held in confidence
Royal family	Commercial sensitivity
Information covered by the DPA	International relations
These are rarely worth challenging!!	***These are always worth challenging!!***

Other grounds for refusal

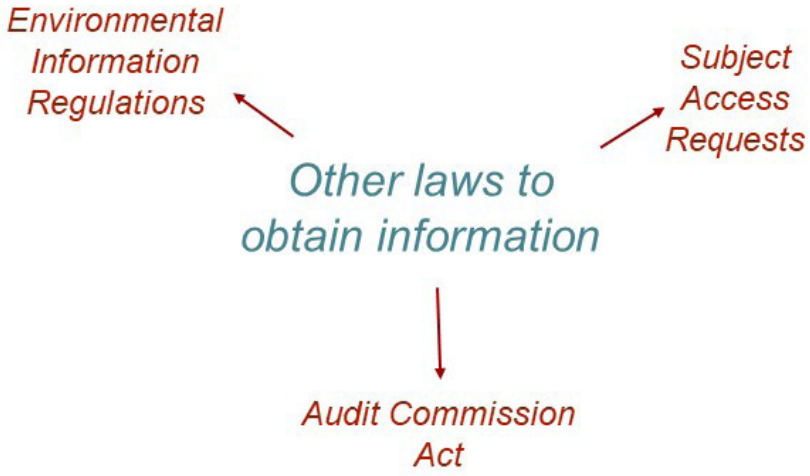
Limits:

- Central government: £600 (or 24 staff hours)
- All other bodies:£450 (18 staff hours)

Request considered vexatious - repeated, rude, causes undue stress

Information is already in the public domain

Other laws to obtain information



Extra considerations with data

- Ensure you ask for the relevant data dictionaries or accompanying documentation
- Be clear on what measurements you are asking for
- Be organised and keep a spreadsheet for round robin requests
- Specify the data format - if you ask for a spreadsheet, you should get a spreadsheet!

What do you want to do next?

The Data Driven investigations course is just a start. As you work on more and more investigations you will find obstacles to overcome. You may need to look for solutions online or take part in more training courses.

Here are a few thoughts as a guide to how to think about your next steps

HACKS HACKERS london

www.meetup.com/Journocoders

www.hhldn.co.uk

IRE INVESTIGATIVE REPORTERS & EDITORS

NICAR-LEARN

http://ire.org/
http://ire.org/nicar/

Knight Center for Journalism in the Americas
THE UNIVERSITY OF TEXAS AT AUSTIN
ENGLISH | ESPAÑOL | PORTUGUÊS

https://knightcenter.utexas.edu/

https://gijn.org/

Global Investigative Journalism Network

Newsletters

- Data is plural
- Data Elixir
- Quantum of Sollazzo
- Open Data Institute “Week in Data”