

Data Visualisation for Analysis in Scholarly Research

University of St Andrews Library, May 2015

Mia Ridge, Open University
@mia_out <http://miaridge.com>

While we're getting started...

- Check that the mouse on your laptop works and that you can get online with the browsers Firefox or Chrome
- Unzip ('extract') the file containing the slides and exercise handouts and copy the folder to your desktop
- Dig out your GMail/Google login details (if you have an account)

Timetable

- 10am Start
 - 11am Coffee
 - 1pm Lunch
 - 3pm Afternoon tea
 - 4pm Conclude
- Sources and further reading
<http://bit.ly/UJwgEz>

What is data visualisation?

'**sense-making** (also called data analysis) and **communication**' Stephen Few

'...showing quantitative and qualitative information so that a viewer can see **patterns, trends, or anomalies, constancy or variation**' Michael Friendly

'...interactive, visual representations of abstract data to **amplify cognition**' Card et al

Visualisations as intersection of format and purpose

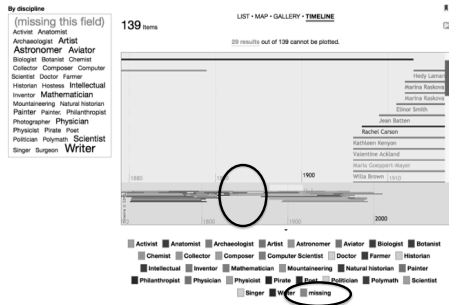
- Exploratory or explanatory: find new insights, or tell a story?
- Product or process?
- Pragmatic, emotive?
- Static or interactive? Print or digital?
- 'Distant reading' - focus on the shape rather than detail of a collection Franco Moretti

Data visualisation can help you...

Explore your data

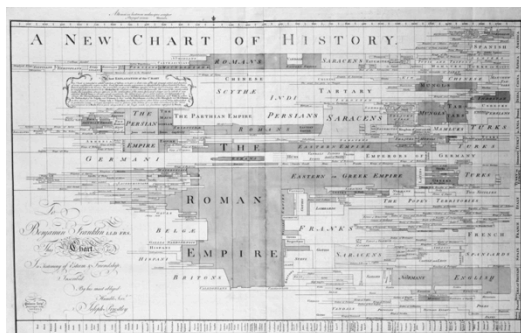
Explain your results

Exploring data



HISTORY AND TYPES OF VISUALISATIONS

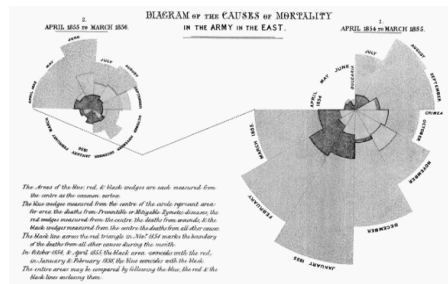
Joseph Priestley, 1769



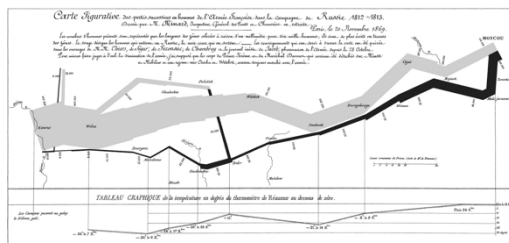
John Snow's cholera map, 1854



Florence Nightingale's petal charts, 1857



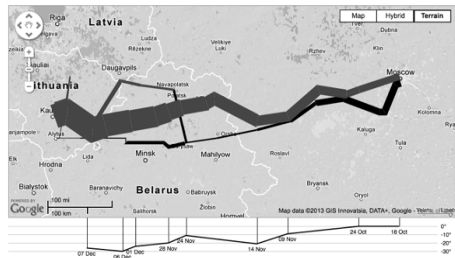
Charles Minard's figurative map, 1869



'Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812-1813'. Drawn up by M. Minard, Inspector General of Bridges and Roads in retirement. Paris, November 20, 1869.

...translated

Flow Map of Napoleon's March on Moscow



<http://hci.stanford.edu/jheer/files/zoo/ex/maps/napoleon.html>

The old tube map



Harry Beck, 1931



Visualising images and video, 2012



Data types

- quantitative / qualitative
- geographic, time series
- media
- entities (people, places, events, concepts, things)

CRITIQUING VISUALISATIONS

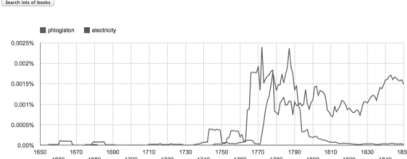
Exercise 1: network visualisations

Instructions on the hand-out.

N-grams

Google books Ngram Viewer

Graph these two sensitive comma-separated phrases: *prisoners electricity*
between 1650 and 1850 from the corpus (English) with smoothing of 3



Search in Google Books:
1650...1702 1708...1778 1778...1790 1791...1842 1843...1850 Electricity English
1650...1778 1778...1790 1791...1800 1801...1850 Electricity English

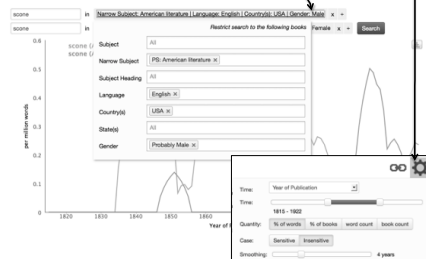
<http://books.google.com/ngrams/>

Exercise 2: comparing N-gram tools

Bookworm tip: click here to change options

bookworm Open Library

Search for trends in hundreds of thousands of books on Open Library



Topic modelling

Topics

future man people men world great life day time present years country things make work good past human true
australia year wool industry future trade production years world prices australian present war price cent time market increase great
it li the lie ti ta tile are good be oil li al mid game mi ni li ha
air future motor power our speed miles made engine day flying years light cars high great ship machine oil
land wheat good farmers sheep sugar years stock time country cattle mr queensland soil farm acres fruit farmers agricultural
city railway water future land town north miles port south building river years area present street sydney line great
war british empire germany britain great german peace states united nations france europe india world russia government military china
mine company mines mining gold and ore is work we they mr futuro them been there which were ft mr government state australia future party made minister members public present air general council federal meeting labour work commonwealth
school women future home children young work years day woman good house boys hers time man year made miss

<http://discontents.com.au/mining-for-meanings/>
<http://wraggelabs.com/shed/presentations/nla/#slide-24>

Other forms of text analysis

JB/002/004/001
1818 April 19 4
Annuitant Notes. Advertisement
4

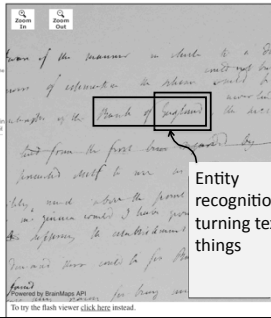
Assure of the manner in which to a degree beyond my power of estimation the scheme would not but be prejudicial to the interests of the Bank of England, the success, however, of the adoption of the scheme, I had been led to me as standing in the scale of probability much above the point of hopelessness. Not so much as one guinea would I have given to receive a hundred in time of success. But in comparison of the advantage, which presented itself to my eyes, the offering (case) was as little as the Treasury was led to mind.

As Supposing the establishment of this currency, what use or demand there could be for Bank paper I neither saw nor saw I had any reason for being anxious to enquire.

But the particular interest of that corporation is in a very new understanding which is not now a secret to any body when even truth is always understood with the interest of subservient to the most available.

of all but the last of ambition, and in that way inappropriately linked and interwoven with that not only with the interest of any administration, but with every the minutest flow of the sinister interest of the ruling class.

To have supposed the Bank would have any impracticable



Exercise 3: trying entity recognition

Instructions on the hand-out.

Entity recognition examples

Named Entity Recognition:

1	Speaking at a UN conference in Sendai, Japan, on Monday, said 90% of buildings in Port Vila had been damaged or destroyed by the category five storm, which saw winds of up to 250km/h (150mph).
2	'It's a setback for the government and for the people of Vanuatu,' he said.
3	'After all the development that has taken place, all this development has been wiped out.'
4	Communications in the Port Vila province of the island have now been 'almost fully restored', according to telecommunications provider Digicel, allowing information to flow more freely to and from disaster areas.
5	Several countries have also pledged additional aid and funding for the stricken island nation.
6	The Australian foreign affairs minister, Julie Bishop, pledged \$5m in support, and New Zealand has offered \$2.5m.

VISUALISATIONS FOR SCHOLARLY ANALYSIS

Scholarly data visualisations

- Visualisations as 'distant reading' where distance is 'a specific form of knowledge: fewer elements, hence a sharper sense of their overall interconnection' Moretti, 2005
- Inspiring curiosity and research questions
- But - which questions do they privilege and what do they leave out?

Exercise 4: explore scholarly visualisations

Pair up and discuss together before reporting back.

Instructions on the hand-out.

University of Richmond, "Visualizing Emancipation"



<http://www.americanpast.org/emancipation/>

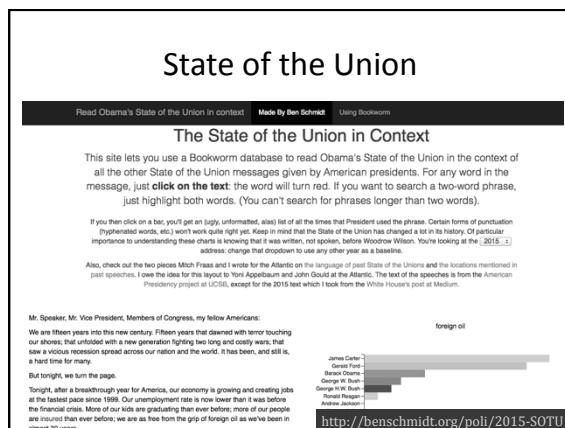
Stanford "Mapping the Republic of Letters"



<http://www.stanford.edu/group/toolingup/rplviz/rplviz.swf>







Comments or questions?

**ISSUES WITH HISTORICAL,
CULTURAL DATA**

Considerations for GLAM data

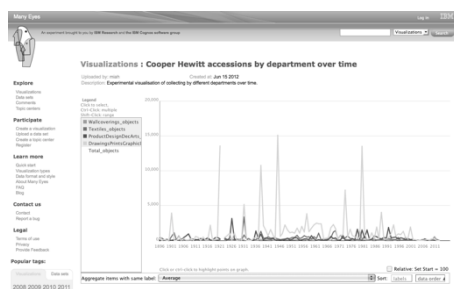
(GLAM: galleries, libraries, museums, archives)

- Commercial tools often assume complete, born-digital datasets – no missing fields or changes in data entry over time
- GLAM records often contain uncertainty and fuzziness (e.g. date ranges, multiple values, uncertain or unavailable information)
- Includes metadata, data, digital surrogates

Messiness in GLAM data

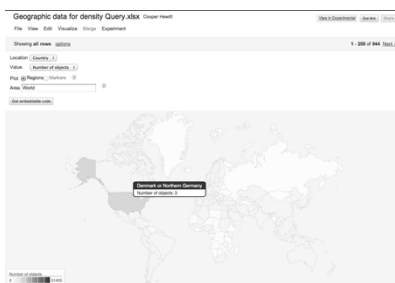
- 'Begun in Kiryu, Japan, finished in France'
- 'Bali? Java? Mexico?'
- Variations on USA:
 - U.S.
 - U.S.A
 - U.S.A.
 - USA
 - United States of America
 - USA ?
 - United States (case)
- Inconsistency in uncertainty
 - U.S.A. or England
 - U.S.A./England ?
 - England & U.S.A.

When were objects collected?



<http://ibm.co/OS3HBa>

Computers don't cope



Preparing data for visualisations

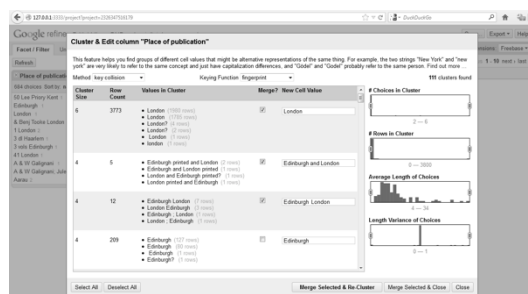
GLAM data often needs manual cleaning to:

- remove rows where vital information is missing
- tidy inconsistencies in term lists or spelling
- convert words to numbers (e.g. dates)
- remove hard returns and non-ASCII characters (or change data format)
- split multiple values in one field into other columns (e.g. author name, date in single field)
- expand coded values (e.g. countries, language)

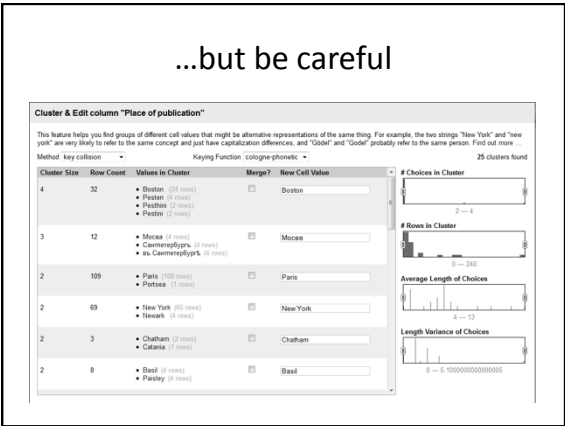
Data Preparation

- Generally needs to be in tables, one row per item, one column per value
- Might need to calculate values in advance
- Data should be made as consistent as possible with tools like
 - Excel
 - OpenRefine <http://openrefine.org>

Open Refine

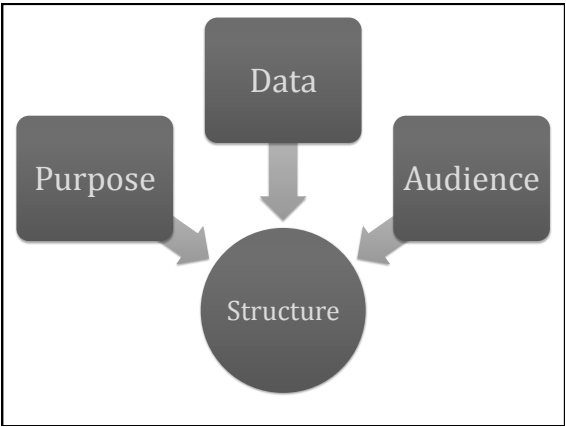


...but be careful



PLANNING VISUALISATIONS

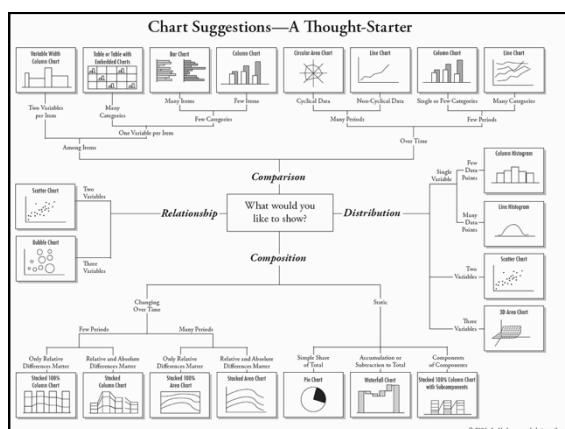




Purpose, data, audiences (revision)

- Intersections of format and purpose
- Data types: quantitative, qualitative, geographic, time series, media, entities (people, places, events, concepts, things)
- Static, interactive; print, digital; product, process
- Exploratory, explanatory: find new insights, or tell a story? Pragmatic, emotive?

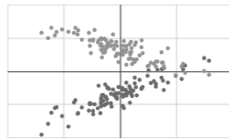
Choosing a structure



See relationships among data points

- Scatterplot
- Matrix
- Network diagram

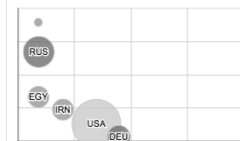
Scatter Chart



Compare a set of values

- Bar chart
- Bubble chart
- Histogram

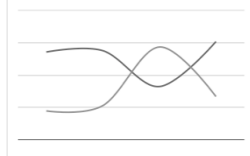
Bubble Chart



Track change over time

- Line graph
- Stack graph

Line Chart



See the parts of a whole

- Pie chart
- Treemap

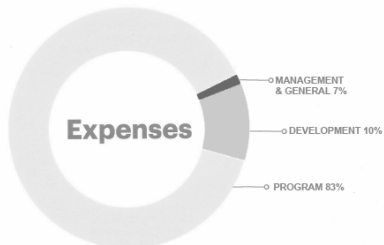


Exercise 5: create a chart using Google Fusion Tables

- Instructions on the hand-out
- If you would rather try an exercise in Excel, see instructions for [creating simple graphs with Excel's Pivot Tables and Tate's artist data](#)

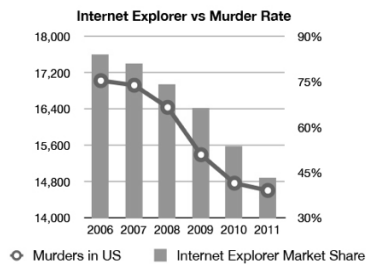
DESIGNING VISUALISATIONS

Worst practice in data visualisations



Source: <http://www.forbes.com/sites/naomirobbins/2013/01/03/deceptive-donut-chart/>

Worst practice in data visualisations



Source: <https://twitter.com/altonnncf/status/293392615225823232>

Best practice for design

- How effectively does the visualisation support cognitive tasks?
- The most important and frequent visual queries/pattern finding should be supported with the most visually distinct objects

Visually distinct objects

- Colour (hue, lightness)
- Elementary shape (orientation, size, elongation)
- Motion
- Spatial grouping

	Points	Lines	Areas	Best to show
Shape		<i>possible, but too weird to show</i>	<i>cartogram</i>	<i>qualitative differences</i>
Size			<i>cartogram</i>	<i>quantitative differences</i>
Color Hue				<i>qualitative differences</i>
Color Value				<i>quantitative differences</i>
Color Intensity				<i>quantitative differences</i>
Texture				<i>qualitative & quantitative differences</i>

Bertin's retinal variables via Making Maps: A Visual Guide to Map Design for GIS by John Krygier and Denis Wood

Properties and Best Uses of Visual Encodings							
Example	Encoding	Ordered	Useful values	Quantitative	Ordinal	Categorical	Relational
	position, placement	yes	infinite	Good	Good	Good	Good
1, 2, 3, A, B, C	text labels	optional (alphabetical or numbered)	infinite	Good	Good	Good	Good
	length	yes	many	Good	Good		
	size, area	yes	many	Good	Good		
	angle	yes	medium/few	Good	Good		
	pattern density	yes	few	Good	Good		
	weight, boldness	yes	few		Good		
	saturation, brightness	yes	few		Good		
	color	no	few (< 20)			Good	
	shape, icon	no	medium			Good	
	pattern texture	no	medium			Good	
	enclosure, connection	no	infinite			Good	Good
	line pattern	no	few				Good
	line endings	no	few				Good
	line weight	yes	few		Good		

Noah Iliinsky • ComplexDiagrams.com/properties • 2012-06

Dealing with complex data

- Find a visualisation type that can harbour the data in a meaningful way or reduce the data in a meaningful way.
 - e.g. go from individual values to distribution of values
 - e.g. introduce interaction: overview, zoom and filter, details on demand Ben Shneiderman

Do you really need a visualisation?

- Use tables when:
 - doc will be used to look up individual values
 - to compare individual values
 - precise values are required
 - the quantitative info to be communicated involves more than one unit of measure
- Use graphs when:
 - the message is contained in the shape of the values
 - the document will be used to reveal relationships among values

Publishing visualisations

- How can you contextualise, explain any limitations of your visualisations? e.g.
 - provenance and qualities of original dataset;
 - what you needed to do to get it into software (how transformed, how cleaned);
 - what's left out of the visualisation, and why?

Tools that don't require programming

- Excel
- Viewshare
- Google Fusion Tables, Google Drive
- IBM Many Eyes
- Tableau Public

Exercise 6: geocoding data and creating a map

- Instructions on the hand-out

Review: planning a visualisation

- With a dataset in mind, consider...
- Exploratory or explanatory? Static or dynamic? Small- or large-scale?
- Choose a type of visualisation (map, timeline, chart, etc)
 - Is your dataset in a suitable format for your visualisation type? How can you clean it?
 - Is more cleaning or transformation needed? You may need to iterate with different versions of your data

If all else fails...

- Sketch out your visualisation on paper to test it
- Iteration is key, and...
- Stubbornness is a virtue!

Exercise 7: taking things further

- Instructions on the hand-out

Review: visualisation tools

- Any data cleaning tips?
- What did you learn about the data?
- What did the tool do well? Poorly?
- Were the tool and the data a good match for each other?
- What other data could you link to?

Thank you!

Mia Ridge, Open University
<http://miaridge.com>
@mia_out
