Data visualisation for humanities researchers

Mia Ridge @mia_out
Elton Barker @eltonteb
Open University
CHASE Going Digital programme
Tips on getting started...

• Sign up to Many Eyes
  http://www-958.ibm.com/software/analytics/manyeyes/register (or check your login, if you’ve already signed up)

• Dig out your Google account details (if you have have data prepared suitable for spreadsheets)

• Have a look through these sites for inspiration:
  – http://www.visualcomplexity.com/vc/
  – http://selection.datavisualization.ch/
I’ll try to keep the bits of me talking at you to a minimum. The exercises start off quite simple and build from there; knowing your way around a browser will help but no hardcore technical skills are required. Making good visualisations takes time so it’s hard to tell how far you’ll get today. One of the most important things to take from today is a sense of how to figure out where to start then where to look for more information.

The most important skill in these digital workshop is actually your attitude - sometimes you might need to think your way through a frustrating bit of data cleaning or figuring out how tools work, but you’ll get there in the end if you keep at it. There are points where I’ll ask you to critique a tool, so be prepared for discussion. Keep your own work at the back of your mind, and ask questions when you want to check how the projects and ideas might relate to your own PhD. The more you put into this, the more you’ll get out of it.
Our ‘life’ visualisations are created in software called Neatline (http://neatline.org/), which is designed for hand-crafted visualisations with maps and timelines.

Currently PhD student in Digital Humanities in the Department of History, Open University
PhD and MSc (Human-Computer Interaction) research on crowdsourcing
‘Scholar in residence’ at Cooper Hewitt
Cultural heritage technologist (Science Museum, Museum of London, Melbourne Museum; programmer, analyst, user experience design and research)
Elton’s observations:

- Neatline is reasonably easy to use. On the other hand, Mia needed to explain a number of features for me, because the site itself doesn’t make them clear intuitively. This is something to bear in mind when producing your own visualisations.
- It’s quite good at bringing together time and place.
- As a snapshot, the timeline of this visualisation is more helpful than the map. In my case, since the data spreads across a geographically broad area, the map makes it difficult to picture the number of features in the UK. This is something else to think about: scale is always an issue in visualisation. The interface as a whole works much better if reading events alongside a detailed map. (Neatline uses an example from the US Civil war.)
Key starting questions

• What's your interest in data visualisation? What is it that you want to do? (e.g. explanation vs. exploration)
• Do you have any potential users in mind? What form will your visualisation take? (e.g. print vs. digital)
• What kind of data do you work with? Do those data need to be prepared in any way?
What is data visualisation?

- ‘...the graphical display of abstract information for two purposes: sense-making (also called data analysis) and communication’ (Stephen Few)
- ‘...fundamentally about showing quantitative and qualitative information so that a viewer can see patterns, trends, or anomalies, constancy or variation, in ways that other forms – text and tables – do not allow.’ (Michael Friendly)
- ‘...interactive, visual representations of abstract data to amplify cognition’

Data visualisation is about creating insight, or the formation of a mental model – a new way of thinking about data.


Michael Friendly quoted at http://www.visualcomplexity.com/vc/blog/?p=1076

If interested in the history of visualisation, find out more http://datavis.ca/milestones/ Milestones in the history of data visualisation or http://www.cabinetmagazine.org/issues/13/timelines.php CABINET // A Timeline of Timelines
In 1854, a major outbreak of cholera hit Soho. Thought to be caused by ‘miasma’ or bad air. Physician John Snow investigated the water supply. Used statistics and map to prove that outbreaks were related to water supply. Black lines in the detail view indicate number of associated deaths. ‘Cholera deaths are depicted along the road network in their correct locations by address, with quantities measured by parallel tick marks stacked next to the road.’ (Source: https://en.wikipedia.org/wiki/Dot_distribution_map)

The map was used to communicate and convince authorities to take action rather than to come up with his theory.

‘With regard to the deaths occurring in the locality belonging to the pump, there were 61 instances in which I was informed that the deceased persons used to drink the pump water from Broad Street, either constantly or occasionally... The result of the inquiry, then, is, that there has been no particular outbreak or prevalence of cholera in this part of London except among the persons who were in the habit of drinking the water of the above-mentioned pump well.’ John Snow, letter to the editor of the Medical Times and Gazette, quoted in Wikipedia

Source: https://upload.wikimedia.org/wikipedia/commons/thumb/2/27/Snow-cholera-map-1.jpg/1098px-Snow-cholera-map-1.jpg
Florence Nightingale needed to convince people that poor hygiene was causing preventable deaths in the Crimean War. Her petal diagram of *Mortality of the British Army* combined stacked bar and pie charts. Deaths that occurred from preventable diseases are in blue, those that were the results of wounds are in red, and those due to other causes are in black. The chart clearly shows that preventable deaths outnumbered deaths due to wounds, etc.

There’s a lot going on in Minard’s map of the successive losses in men of the French Army in the Russian campaign – movement and loss of men across geographical and temporal space, plus temperatures. Brown/red bar is the number of soldiers who went towards Russia - one millimeter for every ten-thousand men, started off with 422,000 men. The colour changes to black to represent their march back, and the temperature bar should be read from right to left as it marks the temperatures during their retreat. It takes time to read but he’s incorporated many variables. There are some heartbreaking details, like the big reduction in the size of the army after a river crossing – did more die in the river itself, or as a result of getting wet, or did it somehow prevent escape?

‘The numbers of men present are represented by the widths of the colored zones at a rate of; they are further written across the zones. The red [now brown] designates the men who enter into Russia, the black those who leave it. — The information which has served to draw up the map has been extracted from the works of M. M. Thiers, of Segur, of Fezensac, of Chambray, and the unpublished diary of Jacob, pharmacist of the army since October 28th. In order to better judge with the eye the diminution of the army, I have assumed that the troops of prince Jerome and of Marshal Davouss who had been detached at Minsk and Moghilev and have rejoined around Orcha and Vitebsk, had always marched with the army.

The scale is shown on the center-right, in "lieues communes de France" (common French league) which is 4,444m (2.75 miles).

The lower portion of the graph is to be read from right to left. It shows the temperature on the army’s return from Russia, in degrees below freezing on the Réaumur scale. (Multiply Réaumur temperatures by 1¼ to get Celsius, e.g. –30°R = –37.5 °C)’

Here's a version on a modern map that you can interact with...

In fact this map has enjoyed many different reincarnations, and each visualisation presents the data slightly differently (with different effects). See http://www.datavis.ca/gallery/re-minard.php 'Re-visions of Minard’. What visualisation might be most useful for your work? Why is that? And which one do you find the least helpful?
Is it easy to work out your route across London?

Beck realised that when you’re working out a route across London, the most important information is whether and where you need to change lines. Geographical relationships aren’t relevant (except of course that some tube stations are close geographically but on different lines, and some changes between lines take forever).
His topological map (diagram) make it easy to plan a route. Interchanges are visually distinct to normal stops, and even the choice of horizontal and vertical lines is easier to follow than the natural angles of the previous maps. Extraneous detail has been removed so the eye focuses on the important information.
It’s almost hard to remember a time before the Google Maps API made it so much easier to put geolocated information on a map, but it was only released in 2006.

This particular screen shows reported bomb sights during the Blitz and was hugely popular when it was released. http://www.bombsight.org/
Data:
128 paintings by Piet Mondrian (1905 - 1917).
151 paintings by Mark Rothko (1944 - 1957).

Mapping:
X-axis: brightness mean
Y-axis: saturation mean
The two image plots are placed side by side so they share the Y-axis.

‘This visualization demonstrates how image plots can be used to compare multiple data sets. In this case, the goal is to compare similar number of paintings by Piet Mondrian and Mark Rothko (produced over comparable time periods of 13 years) along particular visual dimensions. We have selected particular periods in the career of each artist which are structurally similar. In the beginning of a period each artist was imitating his predecessors and contemporaries. By the end period each developed his mature style for which he became famous. In between, each gradually moved moved from figurative representation to pure abstraction. The left image plot shows 128 paintings by Mondrian; the right shows 151 paintings by Rothko. The paintings are organized according to their brightness mean (X-axis) and saturation mean (Y-axis). These measurements were obtained with digital image processing software. Projecting sets of paintings of these two artists into the same coordinate space reveals their comparative “footprints” - the parts of the space of visual possibilities they explored. We can see the relative distributions of their works - the more dense and the more sparse areas, the presence or absence of clusters, the outliers, etc.’
Lots of different ways to think about types... Do you want to find new insights, or to communicate or convince? Can be exploratory (find stories)/explanatory (tell stories) in purpose, and range from analytic/pragmatic - abstract/emotive axis

Source: http://www.slideshare.net/visualisingdata/andy-kirks-facebook-talk

A Tale of Two Types of Visualization and Much Confusion

Robert Kosara: 'two major types of data-based visualization, and understanding the differences. ... Pragmatic Visualization ...even if understanding this requires some work and experience, the goal of this method is to communicate the data, as efficiently as possible. ... If a visualization is designed to visually represent data, and to do that in such a way as to gain new insights into that data, it shall be called a pragmatic visualization. The basic idea is that using the human visual system (instead of automatic means like data mining or statistics), we can gain insight into data, and develop an understanding of the data and the structures in it. To determine whether a visualization is pragmatic, we simply ask if it allows us to efficiently read the data (or at least the relationships between subsets) from the display.' Cf Artistic Visualization
http://www.emoto2012.org/ - timeline example that also shows sentiment analysis, represented in vertical spatial arrangement and colour. Showing both overview and detail by pulling out tweets that apparently represented particular moments.

For each: what types of data optimised is this optimised for? What are the strengths, weaknesses of this visualisation?

A timeline is a graphical or textual display of events in chronological order [Kumar et al., 1998] and is the most used technique for interacting with time-linear visual information. It also allows the user to explore relationships among historical events. [Silva and Catarci, 2000]

There are other Olympics twitter sentiment timelines, see also https://secure.flickr.com/photos/robhawkes/7661511816/
For each: what types of data optimised is this optimised for? What are the strengths, weaknesses of this visualisation?

This interactive map shows some of the places that printed publications that are listed in the British Library’s Bibliographic dataset - but if you look carefully you’ll probably notice it doesn’t include London – trying to work with messy data so reduced size of dataset - also to allow smaller numbers to be visible in comparison...
Visualisation types (based on Google Charts)
Scatterplots: good for relationships between variables
Matrix chart: good for multi-dimensional data
Bubble chart: good for data with big variations in numbers
Line, stack graphs: good for changes in numbers over time
Pie charts: good for showing proportions
Treemap: good for hierarchical structures
Word tree: good for unstructured text
Phrase Net: display common relationships between words in text
Maps: display data by location
Highly abstract visualisations of nodes and links between them. Placement and colour usually have no meaning. Size of node (blob) or edge (line) might represent the number of links or some other variable, but placement is usually pragmatically based around fitting it on the screen.

For each: what types of data optimised is this optimised for? What are the strengths, weaknesses of this visualisation?

Exercise: trying network visualisations

• In your browser, go to http://stanford.io/SiUC8N
• Follow the instructions to interact with it
• Questions to think about: does interacting with it confirm, enhance or confound your understanding of the characters? Does it open up new questions?

I’m interested in the history of science, so here I’ve plotted phlogiston against electricity to get a sense of when each was being discussed in books. ‘displays a graph showing how those phrases have occurred in a corpus of books…over the selected years.’

For each: what types of data optimised is this optimised for? What are the strengths, weaknesses of this visualisation? Plotting word frequencies. Better with longer phrases?


http://bookworm.culturomics.org/#%7B%22Ame_limits%22%3A%5B1700%2C1921%5D%2C%22smoothingSpan%22%3A%5B1%2C5%5D%22search_limits%22%3A%5B1%2C5%5D%22word%22%3A%5B1%2C5%5D%22language%22%3A%5B1%2C5%5D%22gender%22%3A%5B1%2C5%5D%22electricity%22%3A%5B1%2C5%5D%22time%22%3A%5B1%2C5%5D%22aLanguage%22%3A%5B1%2C5%5D%22language%22%3A%5B1%2C5%5D%22gender%22%3A%5B1%2C5%5D%22time%22%3A%5B1%2C5%5D%22aLanguage%22%3A%5B1%2C5%5D%22electricity%22%3A%5B1%2C5%5D%22time%22%3A%5B1%2C5%5D%22aLanguage%22%3A%5B1%2C5%5D%22gender%22%3A%5B1%2C5%5D%22time%22%3A%5B1%2C5%5D%22aLanguage%22%3A%5B1%2C5%5D

http://bookworm.culturomics.org/OL.html
Exercise: comparing N-gram tools

- Think of two words or phrases you’d like to compare over time
- Open two browser windows
- In one, go to http://books.google.com/ngrams
- In the other, go to http://bookworm.culturomics.org
- Enter your words or phrases in each and compare the results.

Can you tell which differences are due to different corpus, and which are due to different algorithms, or even due to OCR quality? Which do you think does full-text search and which is based only on metadata?

Has anything in the results piqued your curiosity and got you thinking about possible research questions?
Topics are words that frequently appear together in documents, found using statistics and probability. Able to tell you things about a text - very useful when metadata is poor or its content is unknown - ‘useful for analyzing large collections of unlabeled text.’ (http://mallet.cs.umass.edu/) Doesn't require a lot of setup or manual markup, and by focusing on the content and not how it's been recorded, it can challenge existing categories - emergent, not pre-determined (unlike n-gram searches).

http://dsl.richmond.edu/dispatch/pages/intro : ‘MALLET ...generates a specified number of topics from a group of documents. The specific topics are not predetermined by the researcher but instead emerge from the patterns uncovered by the statistical algorithm.'

http://www.cs.princeton.edu/~blei/topicmodeling.html ‘The structure uncovered by topic models can be used to explore the otherwise unorganized collection: dividing documents according to their topics and using the hidden structure to determine similarity between documents.’

For each: what types of data optimised is this optimised for? What are the strengths, weaknesses of this visualisation?
Text analysis sometimes a precursor to visualisation – timelines, maps, links to other datasets, etc, but also a useful result in their own right.

Looking for entities like known people, places, events, objects, concepts means documents can be placed in context, e.g. linked to visualisations of places or times - but without requiring manual markup - enabling new research questions. [e.g. Ian Gregory, Lake District]
Exercise: trying entity recognition

• In your browser, go to http://nlp.stanford.edu:8080/corenlp/process
• Find a short paragraph of text (e.g. from a news site or digitised text) to paste into the box
• How many of the things you recognise did it pick up? Is any of the other information presented useful?

Stanford CoreNLP An integrated suite of natural language processing tools for English in Java, including tokenization, part-of-speech tagging, named entity recognition, parsing, and coreference. Online CoreNLP demo

Stanford Named Entity Recognizer A Conditional Random Field sequence model, together with well-engineered features for Named Entity Recognition in English and German. Online NER demo

Could also try http://voyeurtools.org/
http://services.gate.ac.uk/annie/index.jsp
http://www.opencalais.com/
http://textalyser.net/

Text from news site, Internet Archive or Wikipedia

'ANNIE is one of many Information Extraction systems that have been developed using GATE. It uses finite state algorithms and the JAPE language. This demo shows ANNIE recognising entities in texts. Note: this demo uses a default set of components and IE resources; your mileage may vary! Also, complex HTML structures may prevent the system from being able to analyse the text they contain. The system does name recognition; see the IE User Guide for details of other forms of IE, and issues of domain-specificity and porting. Contact us about our cross-domain, multi-genre systems. To use ANNIE, enter a URL in the box below. Select the types of entities that you would like to mark. GATE will then retrieve the document and extract the required information. This process may take a few seconds.'
Infographics - can be used for good or evil! Special case – usually static, mixed formats.

‘What many of these infographics lack, unfortunately, is accuracy and depth. While the information graphics of the 1980s were generally useful for understanding the context of the data, many of today’s infographics just add eye candy that is of little practical use, while playing fast and loose with the data.’

http://eagereyes.org/criticism/fascinating-world-of-good-infographics on bad infographics: ‘These infographics and visualizations are easy to recognize, though: They throw together random facts without a story and without much of a connection between them. They use pie and bar charts to cheaply get the nice graphics real designers draw by hand. They leave you feeling empty and clueless about the purpose of the graphic.’
This example is from my residency at the Cooper Hewitt Design Museum in New York where I was trying to understand the shape of the collection by using different tools to examine it as an aggregate. The data was a bad combination of poor quality (inconsistent, missing, etc) and huge dataset, so difficult to manipulate. Working from the assumption that object accession numbers reflected the year of accession, I processed the data to extract just the year, then plotted it by department and total accessions by year. Like any museum collection, it’s hard to know what’s the result of different cataloguing standards between departments or object types, what’s not showing because it’s expressed too ambiguously for the software to interpret it, etc.

I don’t know the history of the Cooper Hewitt well enough to understand why certain years have huge peaks, but I can get a sense of the stories hidden behind the graph - changes of staff, the end of the war - and I now have some idea where to look for a story.


What have you learned from visualisation that you might not have learned from looking at the data?
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 - what do they leave out?

Hopefully have some ideas now for how visualisations can enable 'scholars to ask increasingly complex research questions by analysing large scale datasets with freely available tools.’ Thinking now about how visualisations can be used to understand, analyse and present large-scale datasets in the humanities and science, and the value of visualisation tools in understanding the shape of a data set.

In digital humanities, part of discourse around distant and close reading. Enables overview of many sources over long periods of time, highlighting changes in style, genre or content. Visualisation allows a view of large numbers of items and with tools like entity recognition, can help put them in spatial, historical or cultural context.

Ultimately about enabling spotting of patterns; patterns can lead to hypothesis.
This map offers a ‘mental model’ of the world based on the Histories of the ancient Greek writer Herodotus (http://en.wikipedia.org/wiki/File:Herodotus_world_map-en.svg). It’s useful for highlighting the division of the world into three separate units (Europe, Asia and Libya) and the importance of water bodies, in particular rivers, for organising that space. But there are at least two problems with it. First, this picture captures only a snapshot of places mentioned in Herodotus and gives no sense of the ‘narrative flow’ of space. Second, and a related point, this is a depiction of space without lines of interaction, where regions are separate from each other, and whose boundaries are rigidly policed by bodies of water.

In the Hestia project (http://www.open.ac.uk/Arts/hestia/), we investigated Herodotus’s representation of space, using web-mapping technologies to create a series of maps showing a world in flux, thereby allowing us to get a sense of space as something lived, not abstractly conceived. We were particularly interested in nuancing the view of the Histories as depicting a world divided into East and West.
To visualise all the places in Herodotus, we first took a digital text of the Histories from the online Perseus Classical Library (http://www.perseus.tufts.edu/hopper/collections), extracted all ‘placename’ data from it, and then stored those references in a PostgreSQL database, through which we were able to query and visualise the data in various ways.

We first used an open-source GIS (Geographical Information Systems). One important side-effect of being able to visualise all our place data on a map was the fact that it immediately drew to our attention problems with our data that needed addressing. In this scenario outlier places in Scotland and, even worse, the US, showed us that some places in the data that we had taken from Perseus had been misidentified: e.g. as Athens, Georgia and not (as it should have been) Athens, Greece. So, visualisation can be immensely helpful as a way of checking data quickly and flagging up potential discrepancies.

An early step in our analysis of the spatial data was to explore rapidly-generated network maps based on simple co-presence of terms within sections of the text. Such networks have no semantic content as such; they simply reflect the fact that Herodotus is mentioning places ‘in the same breath’. The purpose of producing such representations is to explore how strongly the narrative is bound to geographical territories and how clearly those territories are demarcated. It is also there to flag up potential ‘weak ties’ between them. An SQL query was written that generates a network map of lines, which extends between each pair of locations appearing in a single sentence (in the English translation) of Herodotus. The lines can be easily visualised in the GIS in conjunction with the town sites (as the visualisation shows).

As this description suggests, however, these queries and subsequent visualisations (both within GIS) require a high threshold for IT skills. In our case it was essential that we had a technically proficient researcher embedded on the project. You may want to think about whether it’s worth learning the skills for yourself or whether you can team up with someone who has the expertise already.
In addition to experimenting with visualising the placename data from the digital text in GIS, we also conducted a close textual study of all spatial relations in one part of Herodotus’s Histories, Book 5. That is, we were interested in identifying both spatial data broadly defined (not just placenames but also including peoples or individuals as proxies for places) and the relations between spatial data, based on co-occurrence in a single sentence (so with a semantic connection).

The result (visualised above) was a spaghetti monster! This is what can happen when you simply have too much data to depict...
Because of the illegibility of the previous visualisation, we had to simplify the network in order to be able to make sense of it. But this meant leaving data out. Behind every visualisation are choices made about what to show, how, and why.

With this being so, it would perhaps be attractive to allow each individual reader/user to manipulate the data, depict their own visualisations, and make their own inquiries of the data. If your data is in a digital format, that should be possible – and would have the added benefit of being able to show how a network changes dynamically over time. But in a print publication, only one image (perhaps) is allowed. The medium in which the final visualisation is going to be presented makes a difference too.
We were also concerned to make best use of the visualisation technologies that are already in use and, in particular, those that are easy to use. Google Earth is an obvious example, which we were able to use to represent our data by exposing them in KML (Keyhole Markup Language). This was achieved by installing GeoServer, an Open Source server that serves spatial data in a variety of web-friendly formats simultaneously, including KML, SVG, WMS, WFS and PDF. GeoServer has currently been configured to show all the references to spatial locations using a specially constructed view in the database likely to be of interest to the general public (column names have been altered to be more self-explanatory). Significantly, it should be automatically readable on any machine that has Google Earth installed. Furthermore, since the link is a network link, rather than a static KML file, any changes to the database result in automatic and more-or-less instantaneous updating of the data in the viewing application without any need on the user’s part to do anything.
In the Hestia project, we found it very difficult to visualise spatial changes through the narrative; QGIS does not have useful functionality in this regard beyond the ability to turn layers on and off which becomes impractical beyond book level. Google Earth provides limited timeline functionality, but the KML provided by GeoServer does not provide temporal coordinates and even if it did, we are more interested in change through ‘narrative time’ (i.e. book, and chapter) than real time. Instead, we discovered that the most likely candidate to provide this kind of functionality was a project called TimeMap.js: an Open Source JavaScript project that draws on several technologies in order to allow data plotted on Google Maps to appear and disappear as a timeline is moved. With its developer Nick Rabinowitz, we created a ‘Narrative map’ of Herodotus.

The figure above depicts the landing page for a particular text. A map depicts all the places mentioned in the text (the hotter the symbol, the more times it is mentioned). On the right hand side is a histogram cataloguing the places mentioned in the text, ranked according to most mentions, and visualised so that their occurrences through the narrative can be taken in at a glance.

This work, undertaken in the Google Ancient Places (GAP) project, builds on Hestia by using a fully automated means of discovering ancient places in texts (not just Herodotus) and visualising the results. See the blog http://googleancientplaces.wordpress.com/ for more details. For the visualisation interface, GapVis, see: http://gap.alexandriaarchive.org/gapvis/index.html
The second page of GapVis allows the reader/user to read through the narrative sequentially and see the places mentioned in that particular section of the text (depicted on the left hand side). The places, when first mentioned, appear flush on the right hand side of the Timeline (at the bottom) and bold in the Google Maps pane. These places ‘linger on’ in view until after 10 chapters, thereby attempting to give a more sensitive depiction of the reading process.
The third page of GapVis allows the reader/user to follow up on any particular place and view its summary network (i.e. those places mentioned most often in relation to it over the course of the narrative). Using the principles of linked open data, we also automatically pull in photos from Flickr, thereby providing richer content for the user. Being able to link data in this way automatically provides better content for your own data and allows users to link between different data sets and combine them. For the ancient world, we have been doing this in the Pelagios project: http://pelagios-project.blogspot.co.uk/
“Visualizing Emancipation organizes documentary evidence about when, where, and how slavery fell apart during the American Civil War. ...shows how emancipation occurred unevenly across the South, beginning before the first major battles and ending after the end of the Confederacy. It shows the complex interactions between federal policies, armies in the field, and the actions of enslaved men and women on countless farms and city blocks.” Source: http://dsl.richmond.edu/emancipation/about/

Give a few minutes to explore... What do you think is being presented here? What stories or trends can you start to see? Does it work better at one scale over another? Do you find it more effective at aggregate or detail level? Does it present an argument or provide a space to develop and explore one? What have you learned from visualisation that you might not have learned from looking at the data?

Things to point out: heatmap rather than markers - smooths out sometimes false precision, gives better sense of intensity – also used by Ian Gregory; effect...

Other... Animation of changes over time - what questions does it raise, what questions does it answer? Comments on search? (can you search over time?) Did you notice the drop-downs? Did you realise you could click on the red dots and get more information? [Introduce concept of affordances? How clear are they here?]

Some design points - labelling: 'Legality of Slavery Overlay' where's the key to say what each colour means? What is an 'Emancipation Events'? Does looking at the drop-down provide enough contextual clues? Is the 'i' a clear enough label for the key?

Production notes: 'personal sources' is in the database, but doesn't produce results? No docs for that source?
Correspondence networks...

'We told the students we wanted a web-based visualization of this correspondence collection on a map that could be filtered by author or recipient and by time.

...The tool enabled users to select from multiple views of graphs and animations to illustrate different attributes of the data. Geographic mapping of correspondence revealed the major highways of communication. The sliding timescale allowed for isolating periods of travel or exile. We could also compare correspondence collections by author, recipient, or both, revealing patterns of exchange that had not been revealed before.

By clicking on any vector on the map you see a list of letters identified by the name of the author and the recipient. Naturally, we asked the students to turn that string into a hyperlink to the full text of the letter in EE (available to anyone with a subscription).

With that, the RoFL viz correspondence visualization became a visual browser for a digital archive, giving scholars a familiar spatial and temporal orientation to a database that before could only be explored through a web form. The visualization, gives scholars new ways into an archive and also prompts new research questions. Source: http://republicofletters.stanford.edu/tools/

Quoted in http://danbri.org/words/2010/11/22/603 "Mapping the Republic of Letters has at its center a multidimensional data set which spans 300 years and nearly 100,000 letters. .... While we use software and computing techniques that were designed for scientific and statistical methods, we are seeking to develop computing tools to enhance humanistic methods, to help us to explore qualitative aspects of the Republic of Letters."

Give a few minutes to explore... What do you think is being presented here? What stories or trends can you start to see? Does it work better at one scale over another? Do you find it more effective at aggregate or detail level? Does it present an argument or provide a space to develop and explore one? What have you learned from visualisation that you might not have learned from looking at the data?

Looks wonderful, but left lots out so they have a newer solution: 'Corrispondenza is geographic correspondence viewer combined with a focusable timeline. Corrispondenza was our first "internally" developed visualization tool and was our attempt to show data that was hidden from view in RPLVIZ. Our first approach to this problem was to add, in the timeline, a graph that included two data measures by year: the letters plotted on the map and those not plotted. As you can see below, the results were dramatic. In the case of Voltaire, the vast majority of letters do not even appear on the map. We added to this a feature that shows on the map connections that do not have dates, so, letters that do not appear on the timeline. If there is no date for a letter, there is no place to put it on the timeline. As long as we have a source and a destination, we indicate that line as a gray line that is persistent, i.e. does not change with the change in time period.

In the process we began to uncover the limitations of existing visualization techniques and began to appreciate to potentially misleading nature of data visualization.

Re-thinking the map view to include letters not- mappable (not plotted). We discovered that many letters were missing from the first version of the correspondence map. The developers had simply eliminated any letters that did not include complete metadata with Author, Recipient, Source, Destination, Date. If any one of those elements was missing, the letter did not appear. This fact is not apparent in the visualization itself'
What types of data are suitable for visualisation? The issues researchers commonly encounter when applying tools designed for the commercial sector to typically fuzzy, incomplete and complex humanities data;

Data within one dataset might have been prepared by different departments, in different original systems or at different times, so when cleaning data, some content might be more likely to drop out than others.

BL data often lists place of publication as a city (as bibliographic data does) but many visualisations work at country level, so an additional step is needed to compare publications by country. Meanwhile, Excel doesn’t cope with dates before 1900.
Examples from the Cooper Hewitt collection. I spent 3/5 of my time at the Cooper Hewitt just trying to get the data clean enough to vaguely represent the collection. The problem is that computers think U.S., U. S., U.S.A., U. S. A., United States, United States of America are six different places.

Fields also contain things like internal notes about potential duplicates, unexpected extra information - notes on what type of location, etc. Lots of inconsistencies - uncertainty and date ranges expressed in different ways.

More common museum issues - What year is 'early 18th century'? What do you do with '1836 (probably)'?
Tools die when they encounter messy data. And not only do they die, they mysteriously take data down with them.
Cleaning data for visualisations

Humanities data often needs manual cleaning to:
- remove rows where vital information is missing
- tidying inconsistencies in term lists or spelling
- converting 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 one field)
- expanded coded values (e.g. countries, language)

Humanities data is only as good as the hundreds of people who’ve created, transcribed and recorded it over the centuries or decades...

Raw collections data often needs a bit of cleaning before it can be used. We can agree among ourselves that 'c1905' means 'within ten years of 1905' and 'c1900' means 'within fifty years of 1900' but how would a computer know that? And what does it do when it encounters ‘1900s’? (For that matter, would an ordinary person?) Specialised services that understand heritage data can help but there aren’t many of those. Data might also contain odd characters or unexpected empty fields that will break scripts.

Converting words to numbers - be careful when creating apparent precision from fuzziness – better to work to date ranges?

The necessity to leave out data that isn’t clean enough is one reason visualisations should be taken with a pinch of salt...

Tricky to load into tools that are limited by the number of rows (Excel), rows/columns (Google Docs) or size of file (Google Refine, ManyEyes). Be ruthless about splitting up files into manageable sizes; tidy up inconsistencies and uncertainties
Data Cleaning and Formatting Tools

- Excel
- Wrangler: [http://vis.stanford.edu/wrangler/]
- OpenRefine (was Google Refine)

Explain how cleaned data for use in exercises [compare and contrast exercises with cleaned and uncleaned data, or use Cooper Hewitt examples], have less tidy data available for Refine exercise or for more advanced people.

Data loss in cleaning, effect on accuracy, etc.

BL data (lodbnb book xml) contained non-UTF-8 characters that broke imports into Refine. Repeating attributes highlights the difference between row-based and object-based data - repeated attributes need to be fields in the same row in a table-style structure, but are fine in others. Crud in labels.

Tableau Public login is at [https://public.tableausoftware.com/auth/](https://public.tableausoftware.com/auth/) and the sheets are at [http://public.tableausoftware.com/workbooks](http://public.tableausoftware.com/workbooks) - not easily findable via the main site!
Google Refine is an amazing tool, and I wouldn’t have gotten anywhere without it. It will suggest ways to make the data more consistent. You can then export the data and keep working on it in other tools, or put it into Google Refine.

One issue is that museums tend to use question marks to record uncertainty in attribution, but Refine strips out all punctuation, so you have to be careful about preserving it (if that’s what you want).

Takes in TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents.

http://freeyourmetadata.org/cleanup/ useful advice
Don't go too far! It's been said that data cleaning is 80% of the work of preparing a visualisation, and that may be because sometimes you have to go back and look at individual rows and even perhaps research the values represented.
Data prep can also include making into a format like XML, JSON, CSV, TSV... Don't be put off by the acronyms, you can break it all down into smaller tasks and do them one after the other.

Getting hands on with reconciliation is a way of figuring out what decisions algorithms are making for you.

Linking to other datasets creates endless new possibilities for visualisations...
Using Google Refine, was able to automatically 'reconcile' 9000 names in the Cooper Hewitt Makers table. By linking them to a specific URL on a site like Freebase, those records can now be part of the web, not just on the web. However, as might be expected with a table that contains all kinds of famous and ordinary people, Refine couldn’t match everything. 66453 records are left as an exercise for the reader.
What other data can you join to yours?

- Information from general sites like Wikipedia, Freebase, VIAF
- Information from other humanities projects, GLAMs
- Other information about the same event, place, person, object, etc
- General contextualising information – science, history, reviews, citations?

Creates endless new possibilities for visualisations...
Best practice in data visualisations

• How effectively does the visualisation communicate and support cognitive tasks?
• Use spatial arrangement and visual variables
• Most important and frequent visual queries/pattern finding should be supported with the most visually distinct objects (e.g. colour (hue, lightness), elementary shape (orientation, size, elongation), motion, spatial grouping and stereoscopic depth)
Start with some poor practice. What’s wrong with this picture? Don’t use visualisations to misrepresent!
Correlation does not equal causation! Be careful about creating false patterns.
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

Could also be called 'should this really be a visualisation?' test.
Go through the relationship types - useful grounding. There’s an updated version of this, good basic building blocks Effective_Chart_Design.pdf

Including histograms, scatter plots, surface plots, tree maps, parallel coordinate plot. Don’t worry too much about the detail, just know that there’s advice on selecting graph types - you may want to refer back to this for the exercise later.

Article: **Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message**
Thinking about data for graphs

- Nominal data - can compare two data points, no inherent ordering. e.g. green, red.
- Ordinal data - can be compared and ordered, but not necessarily values associated with order e.g. S, M, L
- Continuous data
- Interval - equal distance between numbers or units on scale. No absolute 0 point (if it has one, it's arbitrary). e.g. 60 degrees F is not twice as hot as 30F.
- Ratio - has fixed zero point e.g. time. 10 hours is twice as long as 5 hours.

A bit detailed but you need to start thinking about the detail of your data structures...
Visually distinct objects

Source: John Krygier and Denis Wood via 'Information Display Tips', http://understandinggraphics.com/
There are also lots of software libraries for creating visualisations http://selection.datavisualization.ch/ lets you toggle between ones that require you to code and ones that don’t) but many require some programming knowledge so not discussing today.
Tools that don’t require programming

- Excel
- Google Fusion Tables, Google Drive
- IBM Many Eyes
- Tableau Public
Charts, graphs, maps... Make it easier to combine different datasets. Good way to geocode data. Copes better with CSV (save out from Excel).

Maps: 'The first column should contain location names or addresses. The second column should contain numeric values.'

Scatter chart: 'Two or more columns are required, all must be numeric. The values in the first column are used for the X-axis. The values in following columns are used for the Y-axis. Each column is displayed with a separate color. '

Bar chart: 'The first column in the table represents the label of a group of bars. Any number of columns can follow, all numeric, each representing the bars with the same color and relative position in each group. The value at a given row and column determines the height of the single bar represented by this row and column.

Timeline: 'The first column should contain dates. Subsequently, all columns should contain numbers or text. Each numeric column may be followed by one or two text columns.'

Sparkline: 'All columns must be numeric.'

Motion chart: 'The first column should contain entities (e.g. countries) the second is time (e.g. years) followed by 2-4 numeric or string columns.' (try with number of publications per country over time ###)

Candlestick/box chart: 'The first column should be the names of the stocks or categories. The second column represents the low or minimum value for the stock or category, the third column represents the opening or initial value for the stock or category, the fourth column represents the closing or final value for the stock or category, and the fifth column represents the high or maximum value for the stock or category. The optional sixth column contains tooltip text.'

Tree map: 'The first column should be the name of an entity in a hierarchy. Each entity is visualized by a box when the chart is rendered. The second column should be the name of the entity’s parent entity. (The value in the second column of each row should be found in the first column of some other row.) The optional third and fourth columns should be numerical values associated with the entity. The third column is visualized as the size of the box (must be a positive number), and the fourth column is visualized as the color of the box (may be a negative number).’ (try cities within countries ###)
Visualization Options Available in Many Eyes  

Data formats for uploading data
1] Prepare your data.
First, find the data set that you want to put into Many Eyes. The size limit is 5 megabytes.

Data tables
If your data is a list of values, first format it into a table with informative column headers. If your columns have different units of measure, be sure to include the units in the headers. Use a spreadsheet program such as Microsoft ExcelTM or a text file where columns are separated with tabs. If this is your first upload, read the format guidelines. If you have a specific visualization in mind, take a look at its explanation page for additional information.

Free Text
If your data is free text (such as an essay or a speech), open the data in a word processor or web browser, select the text, and copy it to the clipboard by typing control-C (Windows) or command-C (Macintosh).  
http://www-958.ibm.com/software/analytics/manyeyes/datasets/new
Downloadable tool, Windows only. Free with email address. 'Note, the visualisations cannot be saved locally, that is the ‘public’ essence of this free tool.'

'What each job gets paid: find yours and see how it compares' (2012 Annual Survey of Hours and Earnings data)

Copes better with Excel (save from CSV)
Planning a visualisation

- Who is the audience and what do you want them to know after seeing your visualisation?
- Are you telling a story or letting people explore?
- How interactive will it be?
- How do you need to organise your data for the tool?
- Do you need to bring in other data sources?
- Which are the most important variables, and which are supporting detail?

'Need to understand the message the sender wants to convey and the situation to design the visualisation. Also whether you know what you're looking for in the problem space'

URLs for Google Docs to use in other things:
PublicationsByPlace1534-1831_BNBsubset
https://www.google.com/fusiontables/DataSource?docid=14hSke3Xa_NhDDGxZ0Oyb2rBAa85rVBtvtxnMaU8

britishlibrarybnb_subset_listplacesyears_placescleanedinrefine
https://docs.google.com/spreadsheet/ccc?key=0AhekQVhIO1pLdEg3bWtDOFN6ekxYSWJ0anBnZnZQdEE
Exercise: visualising data

- Choose one of your datasets (or a subset)
  - (If you don't have a dataset, try http://www-958.ibm.com/software/data/cognos/manyeyes/datasets/cloud/tags or work with someone else)
- Decide: exploratory or explanatory? Static or dynamic? Small- or large-scale? Why?
- 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?
  - Optionally, sketch out your visualisation on paper to test it

Visualisation type - review previous slides, think about whether you're: Comparing categories; Assessing hierarchies & part-to-whole relationships; Showing changes over time; Charting connections and relationships; Mapping geo-spatial data

You might get further working in pairs...

[Exercises must include: creating a data visualisation (learn how to use online tools to create visualisations that explore British Library datasets such as the British National Bibliography or 19th Century books, designed to result in something to take home to mum); using Google Refine to clean and prepare data. Do, clean, re-do? How to design so that failure is a learning experience? Small, controlled 'compare and contrast' experiments with ManyEyes? Do exercise on discussing how visualisations are good or bad in terms of design?]
For each tool, walk-through of how to create a data visualisation - demo/group exercise. (prepared example) then short activity for people to try it out [? try it]

What were its strengths, weaknesses for your dataset and research goals? What kind of cleaning did you have to do? What scales did it work best at? What difference did size of data set make?

[Exercises must include: creating a data visualisation (learn how to use online tools to create visualisations that explore British Library datasets such as the British National Bibliography or 19th Century books, designed to result in something to take home to mum); using Google Refine to clean and prepare data. Do, clean, re-do? How to design so that failure is a learning experience? Small, controlled 'compare and contrast' experiments with ManyEyes? Do exercise on discussing how visualisations are good or bad in terms of design?]
Things you might have discovered:

Lots of cleaning is required, especially for dates. What decisions did you make about keeping or deleting data?
Some visualisations are better for numbers, others for text
Work out what the tool is good at to focus your cleaning efforts, or select which columns of data to use
You might need to try different things to work out what the visualisation tool does
Discussion: 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?
References and finding out more

Links are posted at http://wp.me/p34xJL-1D

Thank you!
Mia Ridge @mia_out
Elton Barker @eltonteb
Open University