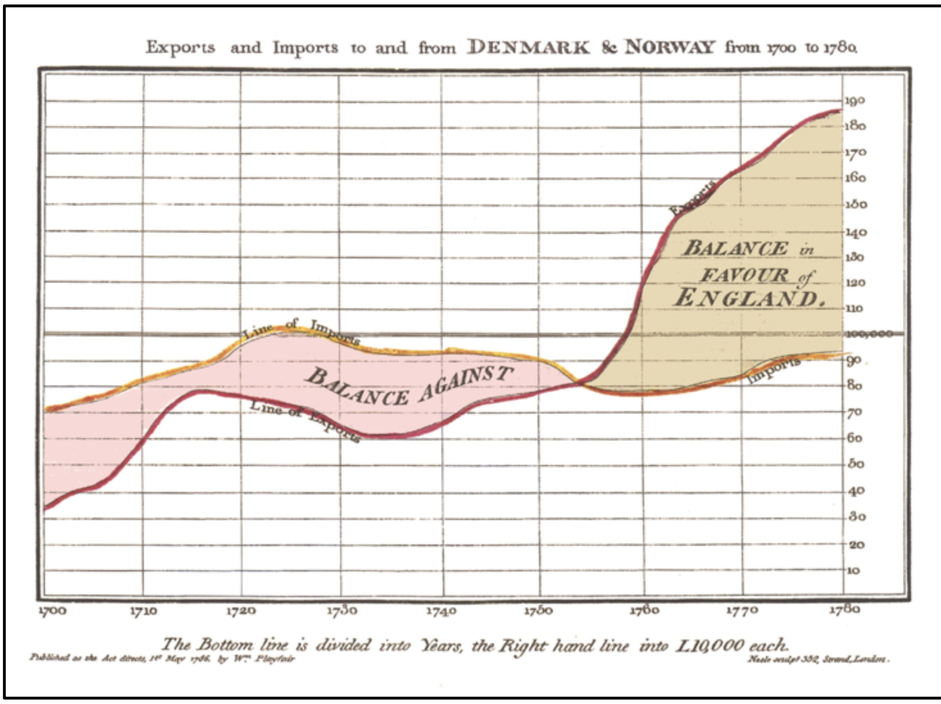
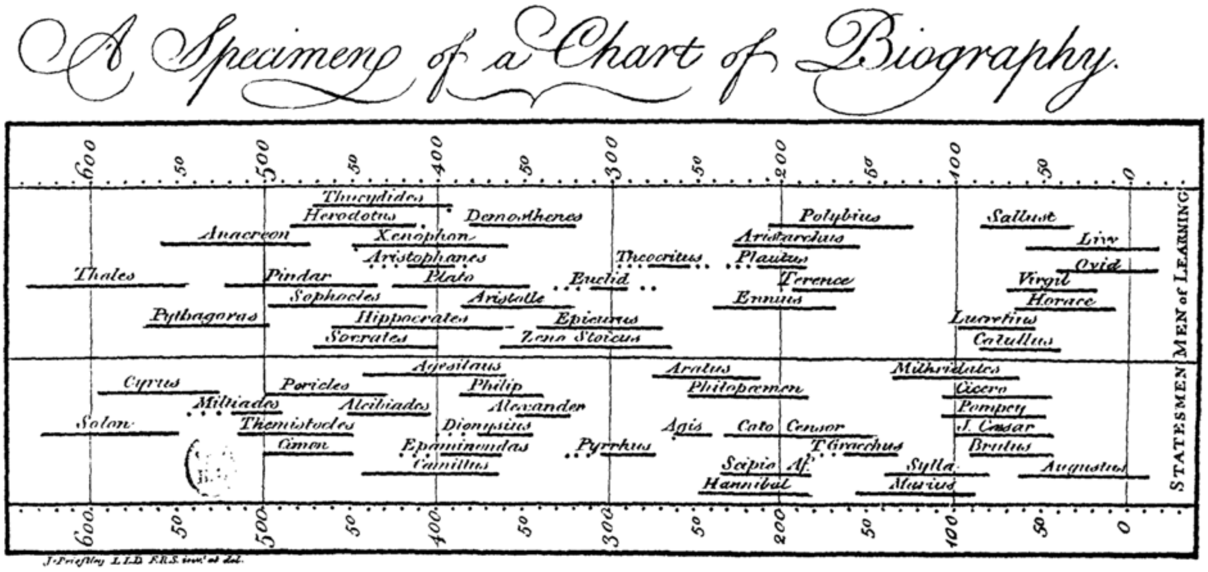
VISUALISATION

Interaction with Machine Learning
Cambridge MPhil ACS 2018-2019



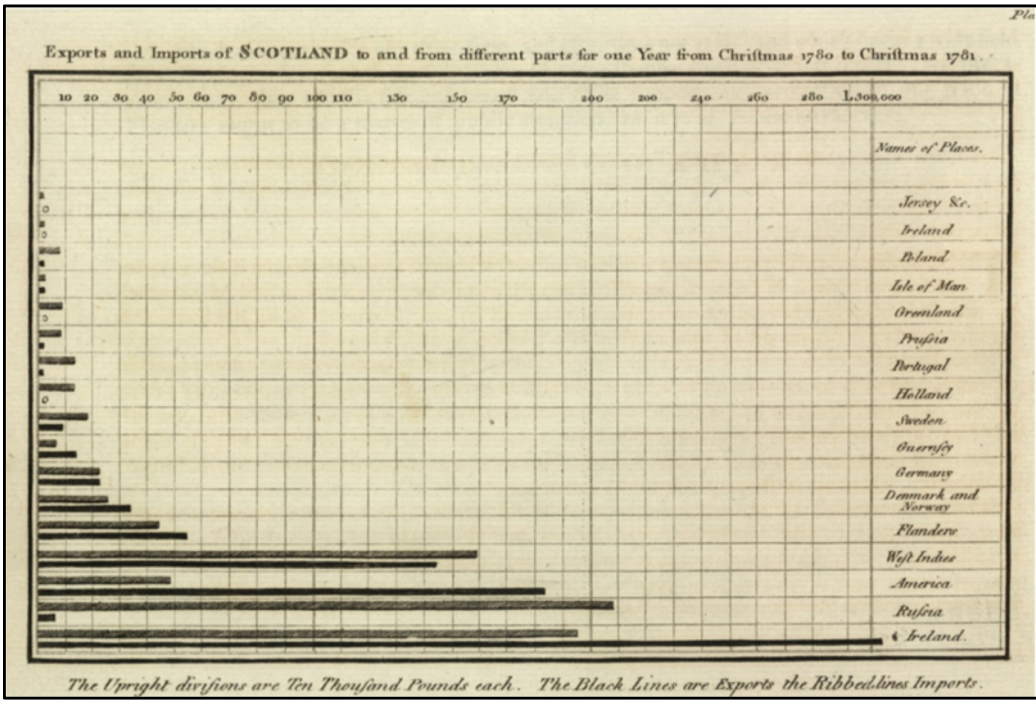
Time Series area chart
William Playfair, c1786

William Playfair (22 September 1759 – 11 February 1823) was a Scottish engineer and political economist, the founder of graphical methods of statistics.[1] He invented several types of diagrams: in 1786 the line, area and bar chart of economic data, and in 1801 the pie chart and circle graph, used to show part-whole relations.



Lifespan chart
Joseph Priestly, c1765

Two decades before Playfair's first achievements, in 1765 Joseph Priestley had created the innovation of the first timeline charts, in which individual bars were used to visualise the life span of a person, and the whole can be used to compare the life spans of multiple persons. According to James R. Beniger and Robyn (1978) "Priestley's timelines proved a commercial success and a popular sensation, and went through dozens of editions".

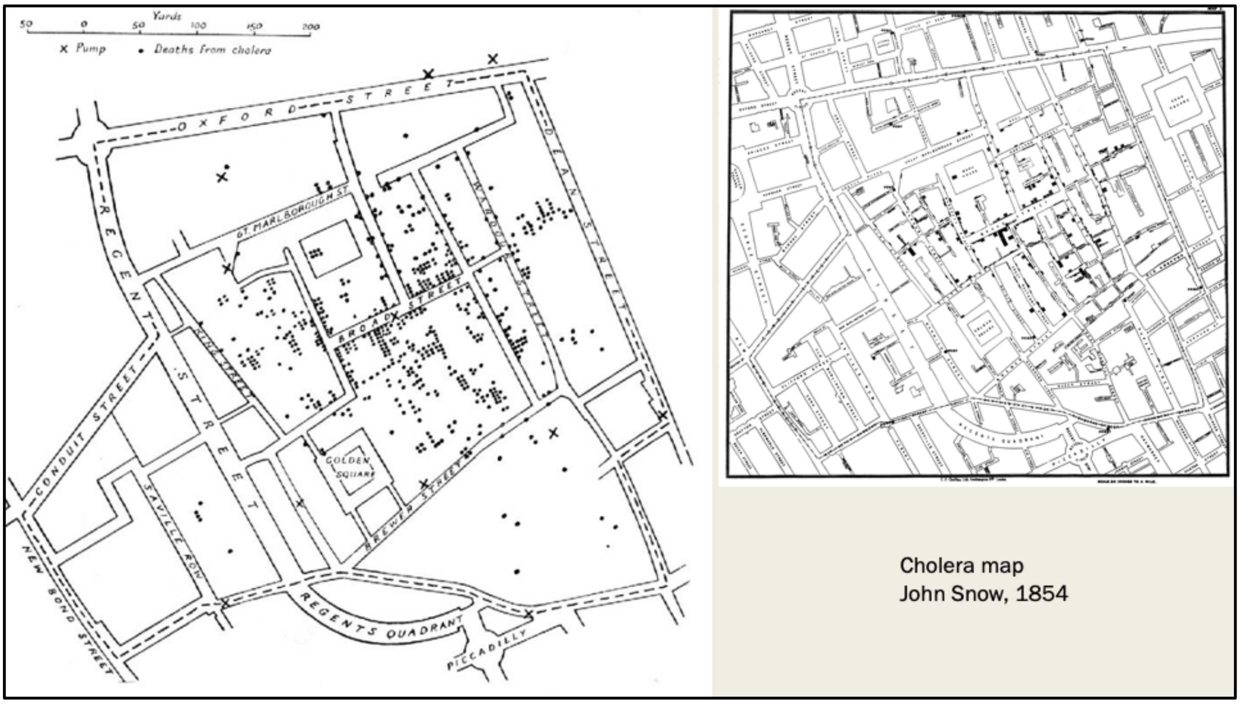


Bar chart
William Playfair, c1786

These timelines directly inspired Wiliam Playfair's invention of the bar chart, which first appeared in his *Commercial and Political Atlas*, published in 1786.

Playfair was driven to this invention by a lack of data. In his Atlas he had collected a series of 34 plates about the import and export from different countries over the years, which he presented as line graphs or surface charts: line graphs shaded or tinted to show the difference [skip back to slide].

Because Playfair lacked the necessary series data for Scotland, he graphed its trade data for a single year as a series of 34 bars, one for each of 17 trading partners, In this bar chart Scotland's imports and exports from and to 17 countries in 1781 are represented. "This bar chart was the first quantitative graphical form that did not locate data either in space, as had coordinates and tables, or time, as had Priestley's timelines. It constitutes a pure solution to the problem of discrete quantitative comparison".



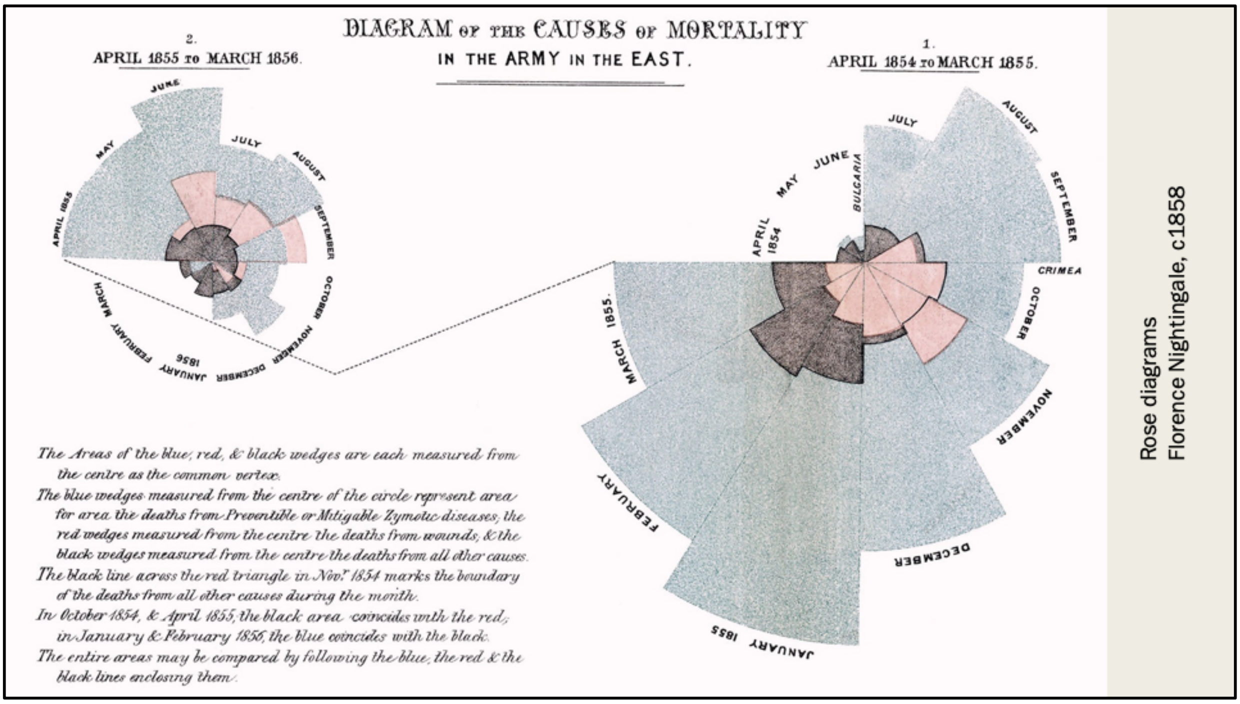
Cholera map
John Snow, 1854

John Snow (15 March 1813 – 16 June 1858) was an English physician and a leader in the adoption of anaesthesia and medical hygiene. He is considered one of the fathers of modern epidemiology, in part because of his work in tracing the source of a cholera outbreak in Soho, London, in 1854

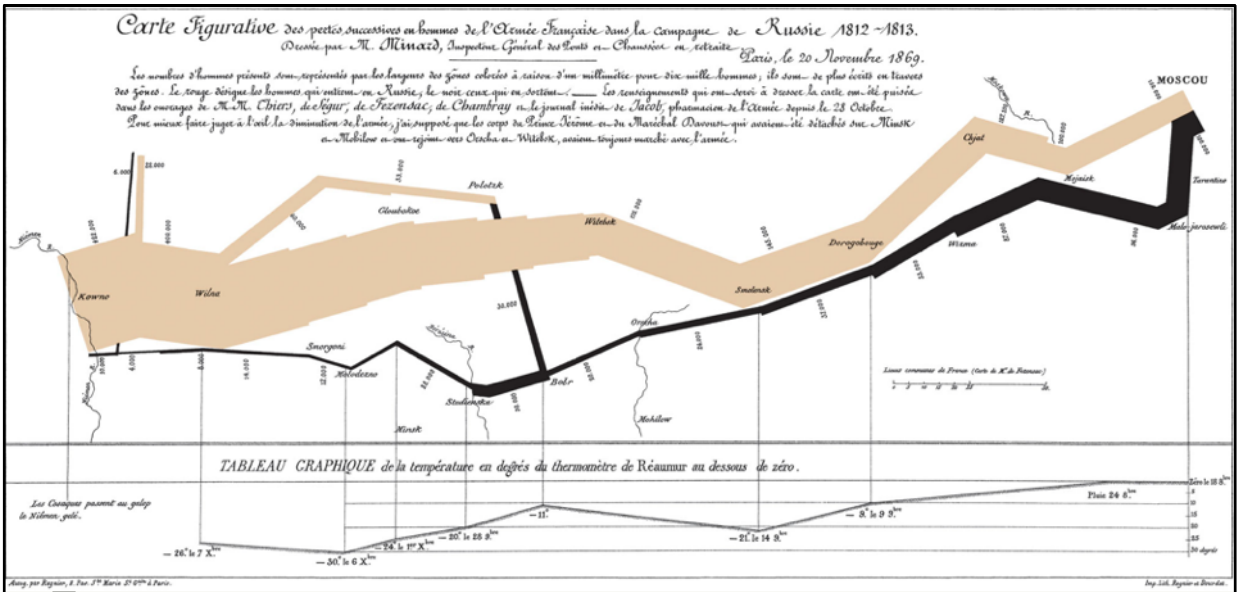
Snow was a skeptic of the then-dominant miasma theory that stated that diseases such as cholera and bubonic plague were caused by pollution or a noxious form of "bad air". The germ theory of disease had not yet been developed, so Snow did not understand the mechanism by which the disease was transmitted. His observation of the evidence led him to discount the theory of foul air. He first publicised his theory in an 1849 essay, *On the Mode of Communication of Cholera*,[14] followed by a more detailed treatise in 1855 incorporating the results of his investigation of the role of the water supply in the Soho epidemic of 1854.[15][16]

By talking to local residents (with the help of Reverend Henry Whitehead), he identified the source of the outbreak as the public water pump on Broad Street (now Broadwick Street). Although Snow's chemical and microscope examination of a water sample from the Broad Street pump did not conclusively prove its danger, his studies of the pattern of the disease were convincing enough to persuade the local council to disable the well pump by removing its handle.

Snow used a dot map to illustrate the cluster of cholera cases around the pump. He also used statistics to illustrate the connection between the quality of the water source and cholera cases. He showed that the Southwark and Vauxhall Waterworks Company was taking water from sewage-polluted sections of the Thames and delivering the water to homes, leading to an increased incidence of cholera. Snow's study was a major event in the history of public health and geography. It is regarded as the founding event of the science of epidemiology. Snow's map, demonstrating the spatial clustering of cholera deaths around the Broad Street well, provided strong evidence in support of his theory that cholera was a water-borne disease. Snow used some proto-GIS methods to buttress his argument: first he drew Thiessen polygons around the wells, defining straight-line least-distance service areas for each. A large majority of the cholera deaths fell within the Thiessen polygon surrounding the Broad Street pump, amd a large portion of the remaining deaths were on the Broad Street side of the polygon surrounding the bad-tasting Carnaby Street well. Next, using a pencil and string, Snow redrew the service area polygons to reflect shortest routes along streets to wells. An even larger proportion of the cholera deaths fell within the shortest-travel-distance area around the Broad Street pump.



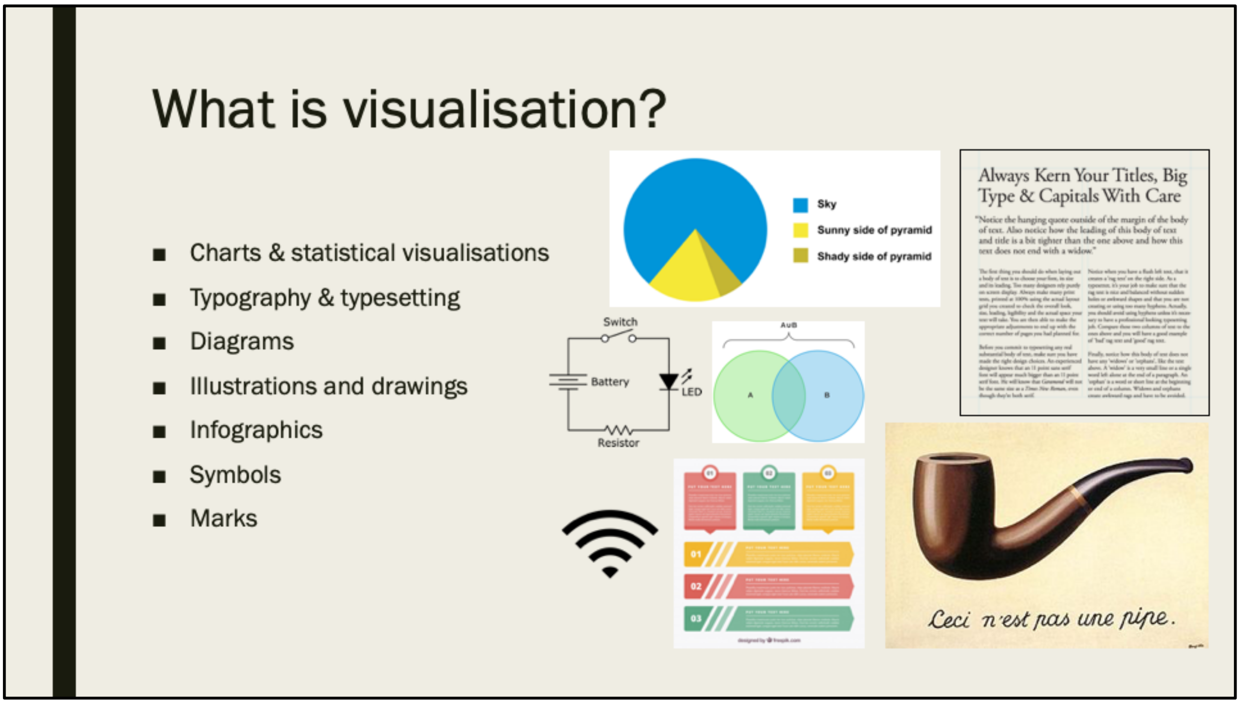
In 1858 nurse, statistician, and reformer Florence Nightingale published *Notes on Matters Affecting the Health, Efficiency, and Hospital Administration of the British Army. Founded Chiefly on the Experience of the Late War. Presented by Request to the Secretary of State for War*. This privately printed work contained a color statistical graphic entitled "Diagram of the Causes of Mortality in the Army of the East" which showed that epidemic disease, which was responsible for more British deaths in the course of the Crimean War than battlefield wounds, could be controlled by a variety of factors including nutrition, ventilation, and shelter. The graphic, which Nightingale used as a way to explain complex statistics simply, clearly, and persuasively, has become known as Nightingale's "Rose Diagram."



Figurative Map of the successive losses in men of the French Army in the Russian campaign 1812-1813.

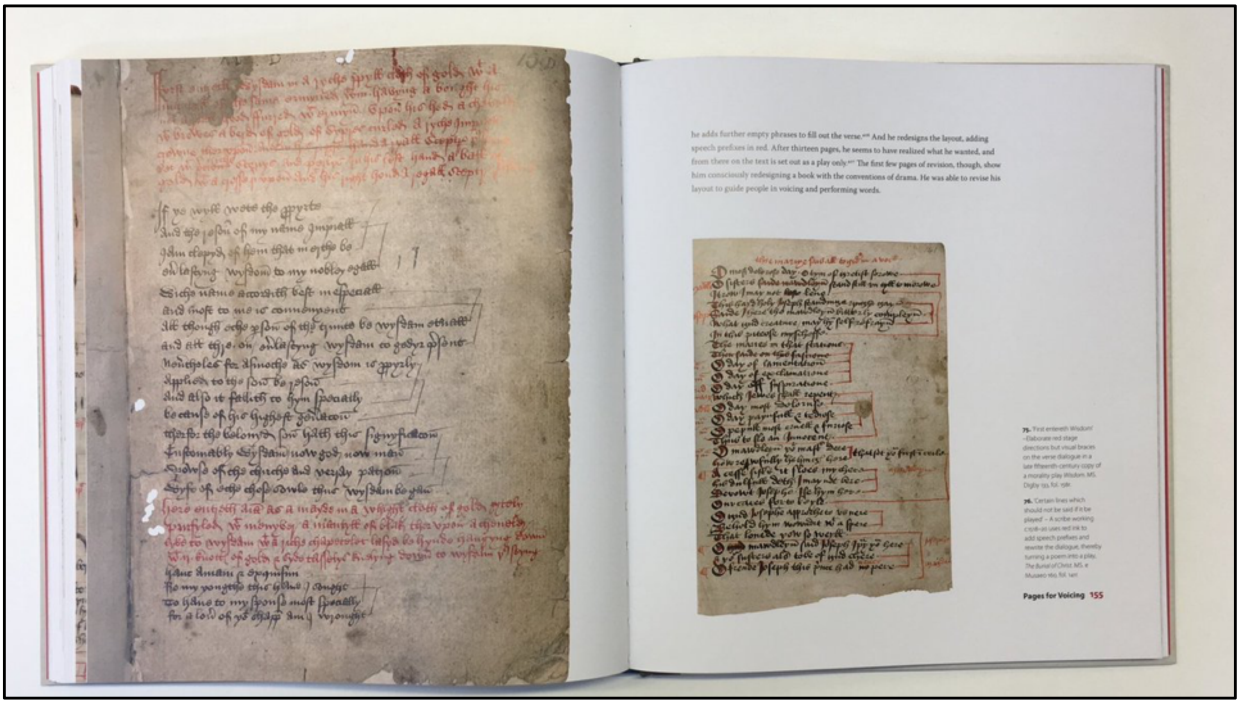
Drawn by Charles Minard, Inspector General of Bridges and Roads (retired). Paris, November 20, 1869.

Map of Napoleon's army by Charles Joseph Minard. Minard was a pioneer of the use of graphics in engineering and statistics. He is most well known for his cartographic depiction of numerical data on a map of Napoleon's disastrous losses suffered during the Russian campaign of 1812. The illustration depicts Napoleon's army departing the Polish-Russian border. A thick band illustrates the size of his army at specific geographic points during their advance and retreat. This graphic is notable for displaying six types of data in two dimensions: the number of Napoleon's troops; the distance traveled; temperature; latitude and longitude; direction of travel; and location relative to specific dates.[2] This type of band graph for illustration of flows was later called a Sankey diagram, although Matthew Sankey used this visualisation 30 years later and only for thematic energy flow).

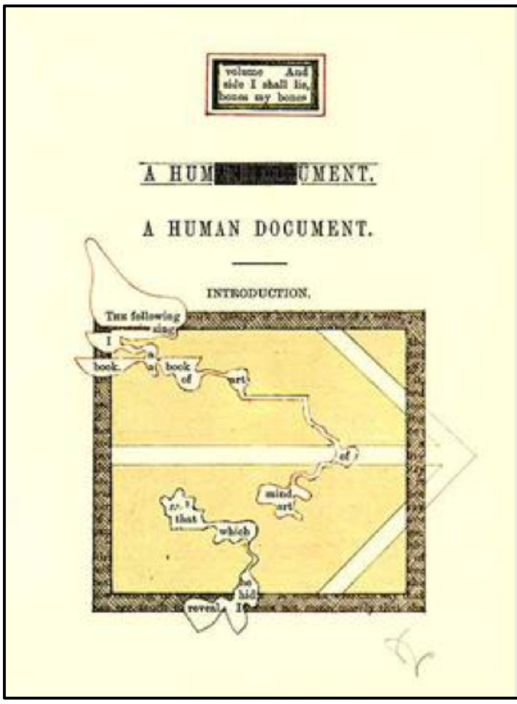


When you hear the word visualisation, you might think of a bar chart or a pie chart.

Daniel Wakelin



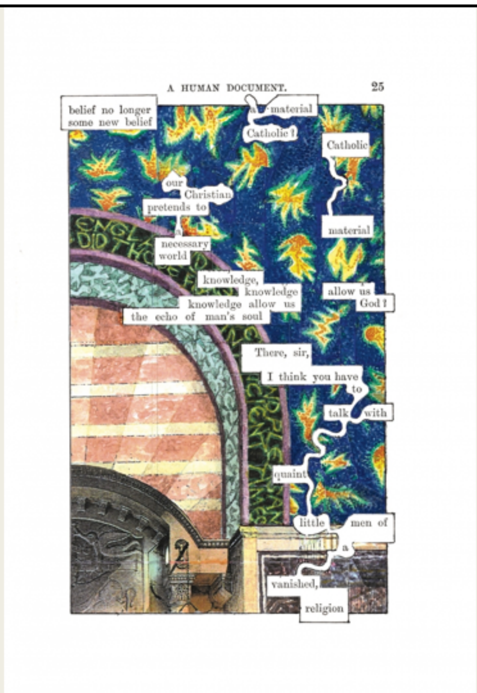
Late 15th century morality play.
A poem converted into a play.



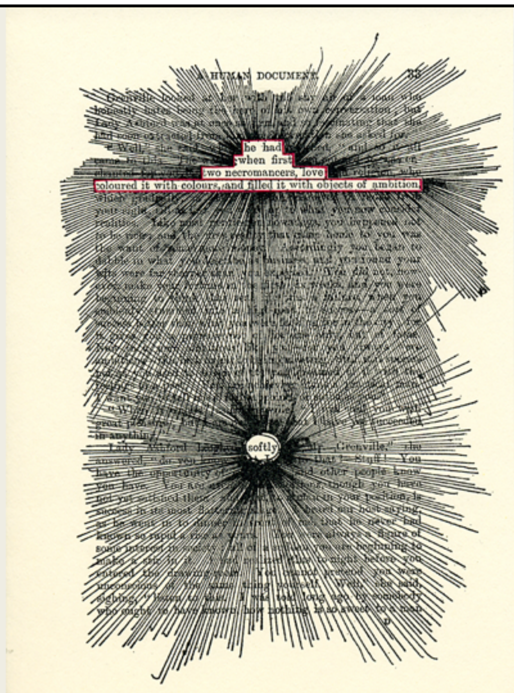
A Humument
Tom Phillips, 1960s

"It is a forgotten Victorian novel found by chance … plundered, mined, and undermined its text to make it yield the ghosts of other possible stories, scenes, poems and replaced the text [he'd] stripped away with visual images of all kinds."

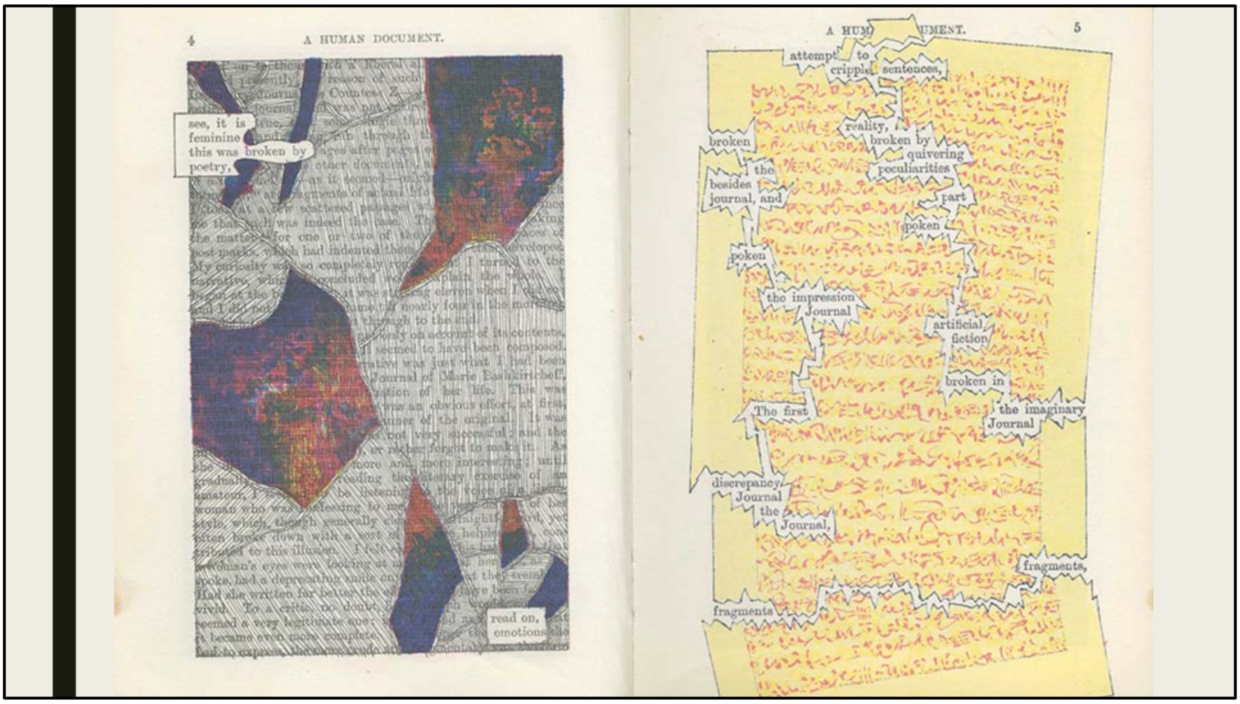
Tom Phillips, 1960s



A HUMAN DOCUMENT.
25
belief no longer
some new belief
material
Catholic!
Catholic
our
Christian
pretends to
a
necessary
world
material
knowledge,
knowledge
allow us
knowledge allow us
God!
the echo of man's soul
There, sir,
I think you have
to
talk
with
quaint
little
men of
a
vanished,
religion



A HUMAN DOCUMENT.
he had
when first
wo necromancers, love
coloured it with colours, and filled it with objects of ambition.
softly



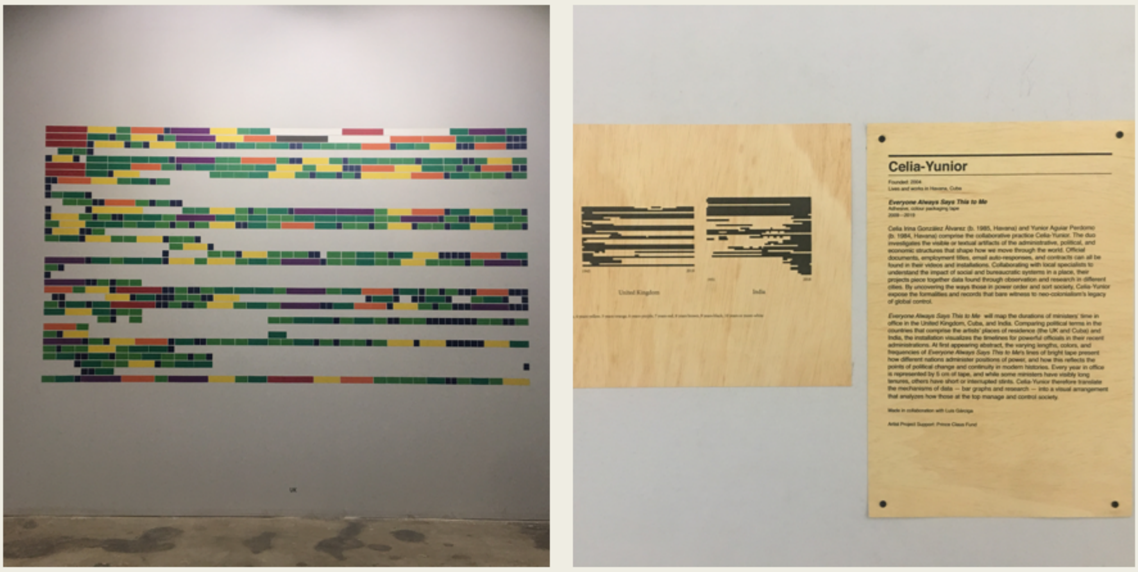
4 A HUMAN DOCUMENT.
see, it is
feminine
this was broken by
poetry,
read on,
emotions
A HUM UMENT. 5
attempt to
cripple sentences,
reality,
broken by
quivering
peculiarities
broken
the
besides
journal, and
part
poken
poken
the impression
Journal
artificial
fiction
broken in
The first
the imaginary
Journal
discrepancy
Journal
the
Journal,
fragments,
fragments


POSSIBILITIES
അന്യതയിൽ
FOR A NON-
നിന്നും അന്യോന്യത
ALIENATED LIFE
യിലേക്ക്
CURATORIAL NOTE

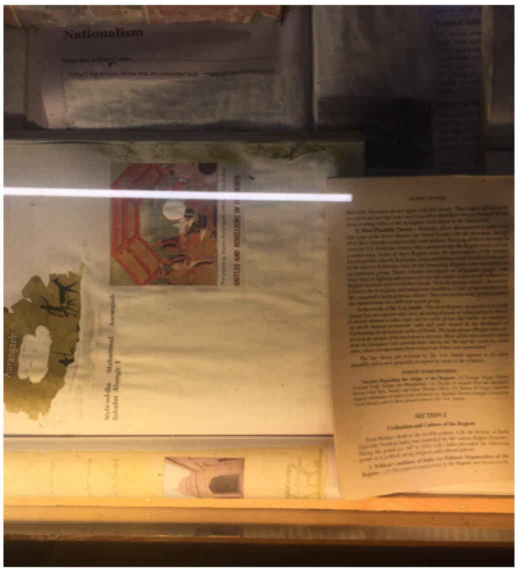
India infographic 1950s - Chittaprosad Bhattacharya - cartoonist

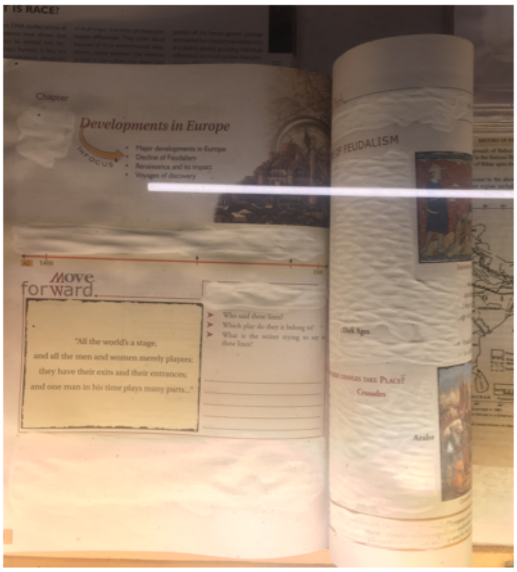
Celia Yunior – growth of IT industry in Kerala, income disparity

Celia Yunior – growth of IT industry in Kerala, income disparity



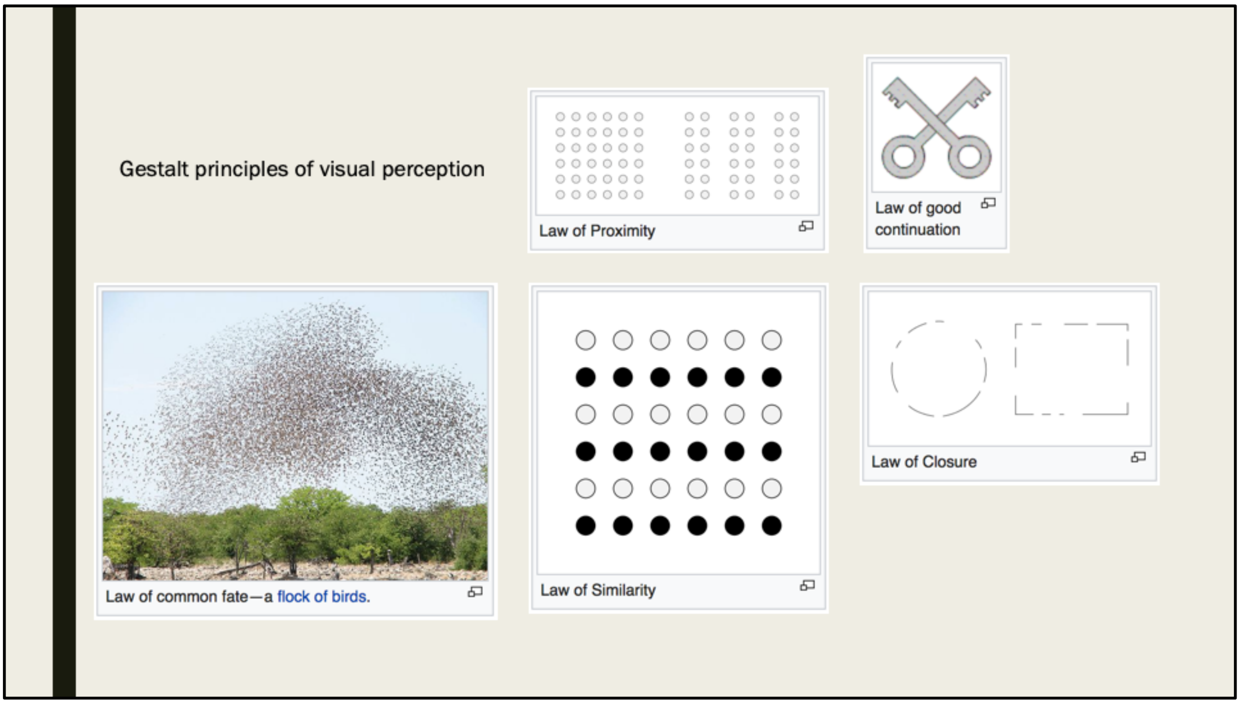
Celia Yunior – positions of power and control in administrations





History – a construct

THEORIES OF VISUALISATION



The **principles of grouping** (or **Gestalt laws of grouping**) are a set of principles in psychology, first proposed by Gestalt psychologists in the early 20th century to account for the observation that humans naturally perceive objects as organized patterns and objects, a principle known as Prägnanz. Gestalt psychologists argued that these principles exist because the mind has an innate disposition to perceive patterns in the stimulus based on certain rules.

For example, the law of common fate. Birds may be distinguished from their background as a single flock because they are moving in the same direction and at the same velocity, even when each bird is seen—from a distance—as little more than a dot. The moving 'dots' appear to be part of a unified whole. The law of common fate is used extensively in user-interface design, for example where the movement of a scrollbar is synchronised with the movement (i.e. cropping) of a window's content viewport; The movement of a physical mouse is synchronised with the movement of an on-screen arrow cursor, and so on.

The principle of similarity states that, all else being equal, perception lends itself to seeing stimuli that physically resemble each other as part of the same object, and stimuli that are different as part of a different object.

The Gestalt law of proximity states that "objects or shapes that are close to one another appear to form groups".

The principles of similarity and proximity often work together to form a Visual Hierarchy. Either principle can dominate the other, depending on the application and combination of the two. For example, in the grid to the left, the similarity principle dominates the proximity principle and you probably see rows before you see columns.

The principle of closure refers to the mind's tendency to see complete figures or forms even if a picture is incomplete

The law of good continuation. When there is an intersection between two or more objects, people tend to perceive each object as a single uninterrupted object.

	Graphic Resources	Correspondence	Design Uses
Marks	Shape Orientation Size Texture Saturation Colour Line	Literal (visual imitation of physical features) Mapping (quantity, relative scale) Conventional (arbitrary)	Mark position, identify category (shape, texture colour) Indicate direction (orientation, line) Express magnitude (saturation, size, length) Simple symbols and colour codes
Symbols	Geometric elements Letter forms Logos and icons Picture elements Connective elements	Topological (linking) Depictive (pictorial conventions) Figurative (metonym, visual puns) Connotative (professional and cultural association) Acquired (specialist literacies)	Texts and symbolic calculi Diagram elements Branding Visual rhetoric Definition of regions
Regions	Alignment grids Borders and frames Area fills White space Gestalt integration	Containment Separation Framing (composition, photography) Layering	Identifying shared membership Segregating or nesting multiple surface conventions in panels Accommodating labels, captions or legends
Surfaces	The plane Material object on which marks are imposed (paper, stone) Mounting, orientation and display context Display medium	Literal (map) Euclidean (scale and angle) Metrical (quantitative axes) Juxtaposed or ordered (regions, catalogues) Image-schematic Embodied/situated	Typographic layouts Graphs and charts Relational diagrams Visual interfaces Secondary notations Signs and displays

Bertin, J. (1967). Semiologie graphique. Paris: Editions Gauthier-Villars. English translation by WJ. Berg (1983)as Semiology of graphics, Madison, WI: University of Wisconsin Press

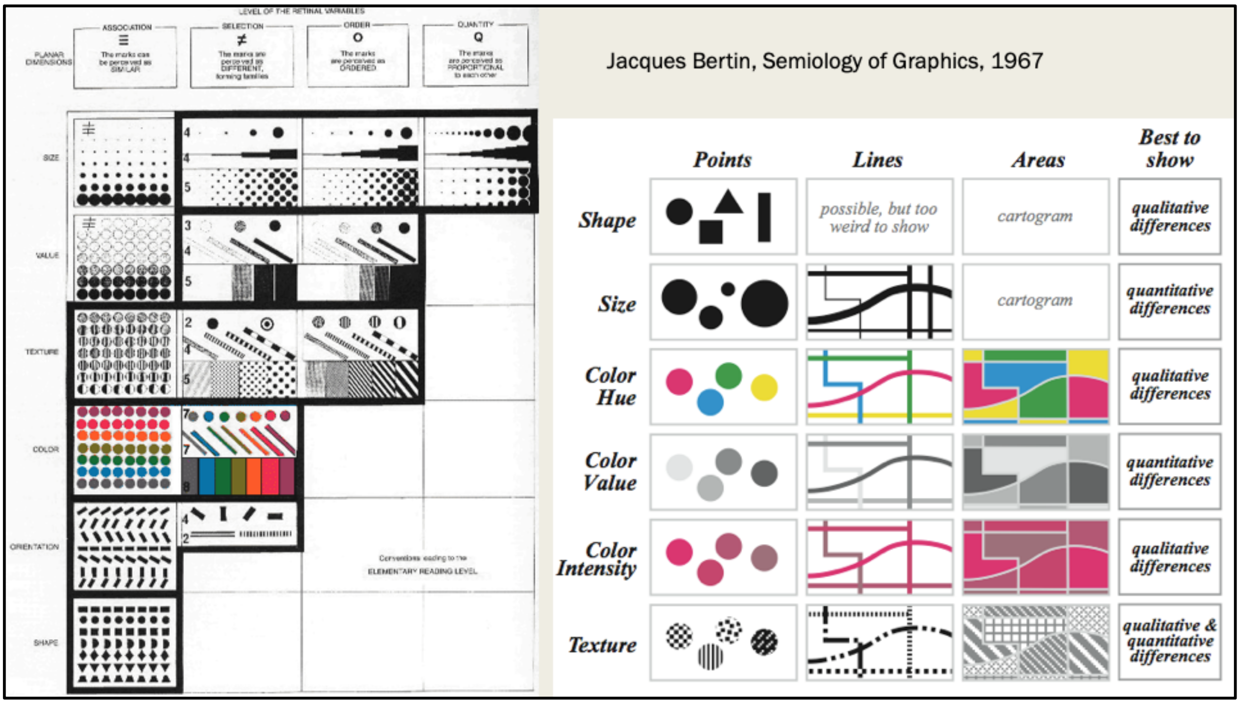
Blackwell, A.F. and Engelhardt, Y. (2002). A meta-taxonomy for diagram research. In M. Anderson&B. Meyer&P. Olivier (Eds.), Diagrammatic Representation and Reasoning, London: Springer-Verlag, pp. 47-64.

Engelhardt, Y. (2002). The Language of Graphics. A framework for the analysis of syntax and meaning in maps,charts and diagrams. PhD Thesis, University of Amsterdam.

MacEachren, A.M. (1995). How maps work: Representation, visualization, and design. Guilford.

Bertin, Richards, MacEachren, Blackwell&Engelhardt and Engelhardt.

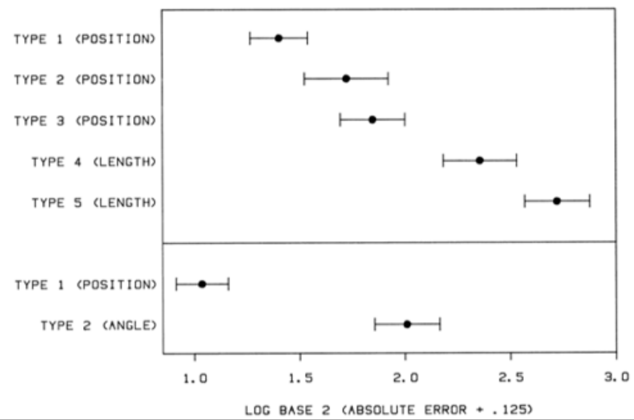
One approach is to take a holistic perspective on visual language, information design, notations, or diagrams. Specialist research communities in these fields address many relevant factors from low-level visual perception to critique of visual culture. Across all of them, it can be necessary to ignore (or not be distracted by) technical and marketing claims, and to remember that all visual representations simply comprise marks on a surface that are intended to correspond to things understood by the reader. The two dimensions of the surface can be made to correspond to physical space (in a map), to dimensions of an object, to a pictorial perspective, or to continuous abstract scales (time or quantity). The surface can also be partitioned into regions that should be interpreted differently. Within any region, elements can be aligned, grouped, connected or contained in order to express their relationships. In each case, the correspondence between that arrangement, and the intended interpretation, must be understood by convention or explained. Finally, any individual element might be assigned meaning according to many different semiotic principles of correspondence.



Graphic resources
“Planar dimensions”
Retinal variables

Cleveland, W. S.,&McGill, R. (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods.*Journal of the American Statistical Association*,*79*(387), 531–554. https://doi.org/10.2307/2288400

Heer, J.,&Bostock, M. (2010). Crowdsourcing graphical perception: using {Mechanical Turk} to assess visualisation design.*ACM Human Factors in Computing Systems (CHI)*, 203–212.



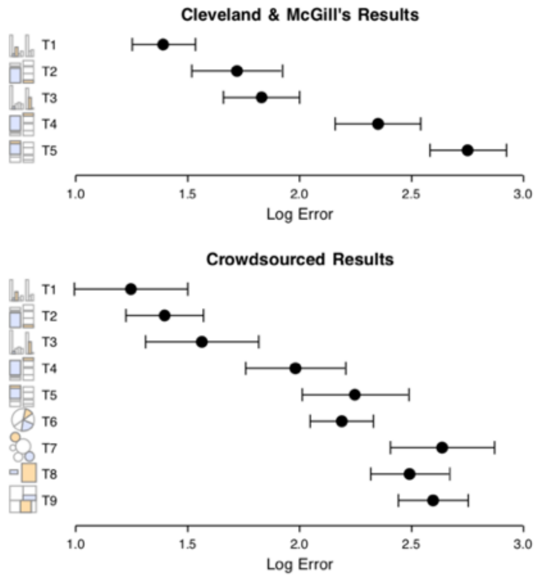


Figure 4: Proportional judgment results (Exp. 1A & B). Top: Cleveland & McGill's [7] lab study. Bottom: MTurk studies. Error bars indicate 95% confidence intervals.

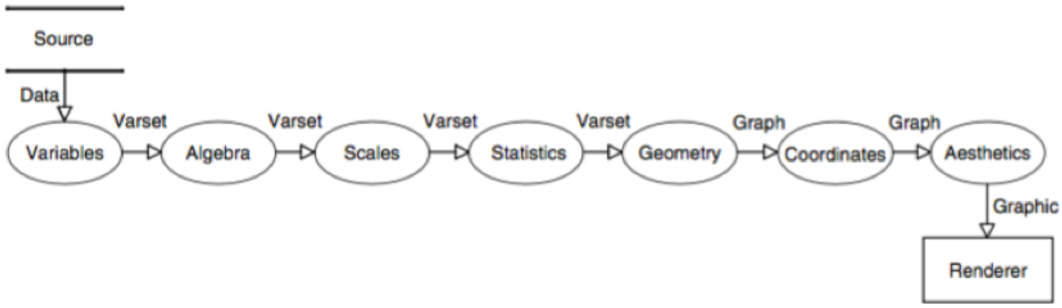
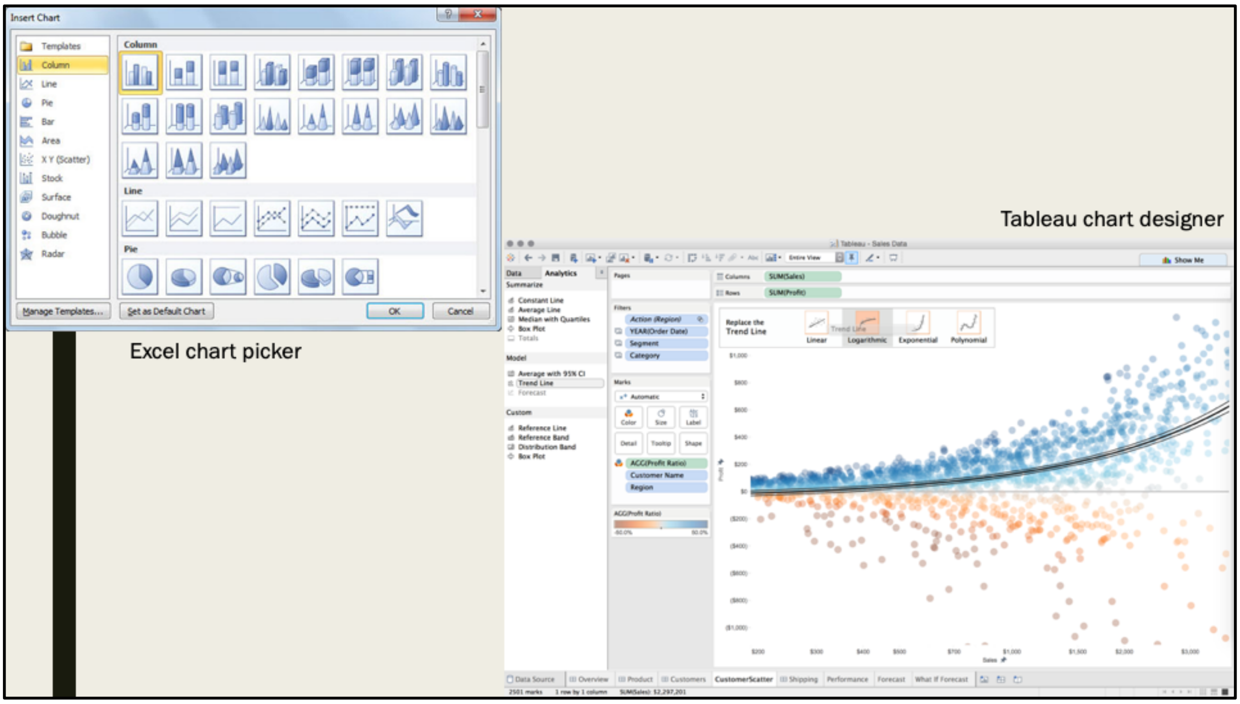


FIGURE 1 | The grammar of graphics data flow.

Leland Wilkinson, The Grammar of Graphics, 1999
Later extended by Hadley Wickham

Take a framework like this and formally encode it.

The grammar of graphics was the foundation for the R package ggplot2



Grammar of graphics is great for people who think about visualisation in such rareified planes of abstraction, but it is not really suited to the mental models and expertise of most end-users. So we have simplified alternatives such as the Excel chart picker, which reframe the pipeline in terms of concrete examples. This is perhaps limiting in terms of the types of visualisations you can achieve, but is vastly more usable. Another point in the spectrum is Tableau's chart designer. This came out of Christopher Stolte's PhD work at Stanford in the late 90s, early 2000s.

Mariana Mărășoiu & Alan F. Blackwell & Advait Sarkar & Martin Spott / Clarifying hypotheses by sketching data

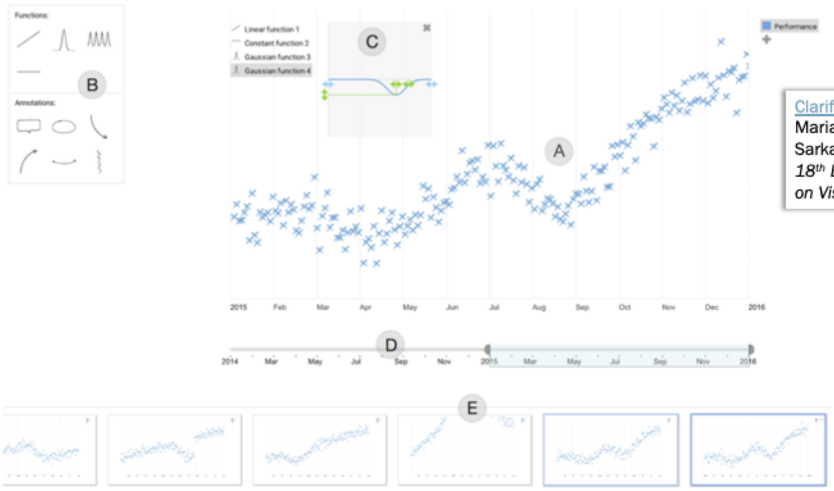


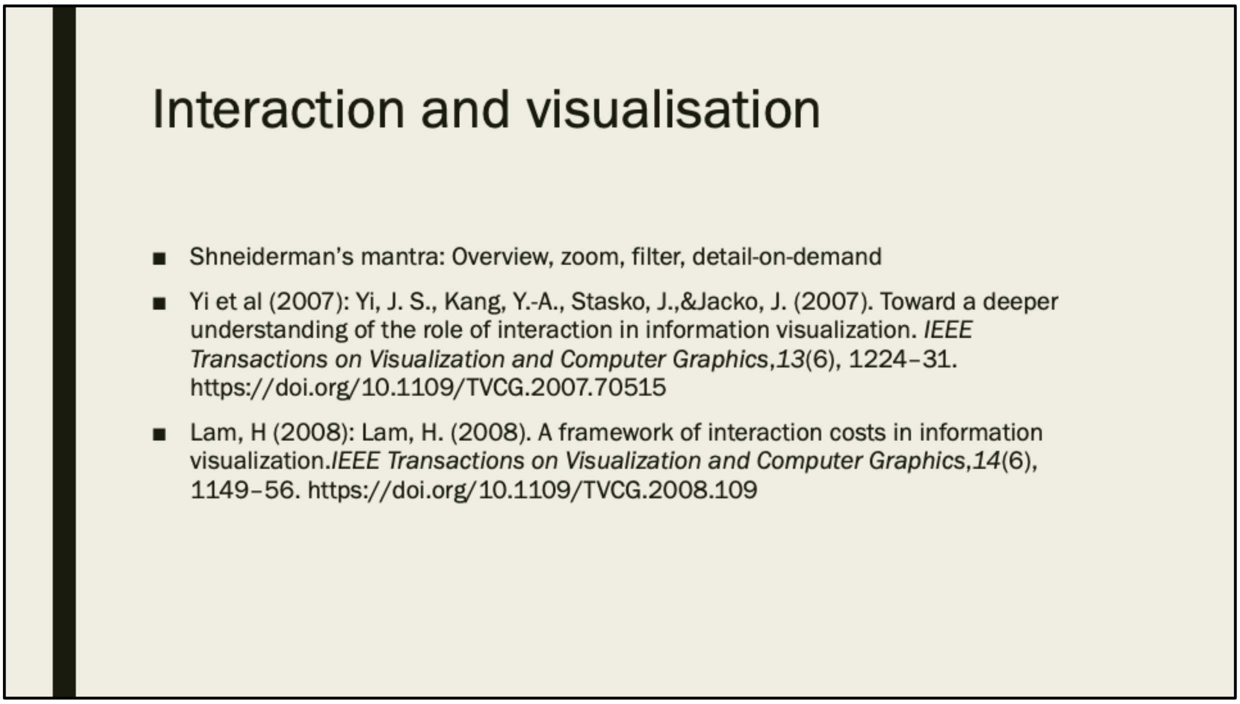
Figure 1: The SelfRaisingData prototype with the main components highlighted. (A) The time series chart with the fictional data points generated around the shape described through function composition, as presented in Section 4.1. (B) The tool panel containing functions and annotations (Section 4.2). (C) The function editor allows interactive modification of the mathematical parameters of the function and the time range for which it applies, as discussed in Section 4.3. (D) The time axis range selector (see Section 4.4). (E) Graphical history using a comic strip metaphor allows branching and visualising previous states (see Section 4.5).

Clarifying hypotheses by sketching data
Mariana Mărășoiu, Alan F. Blackwell, Advait Sarkar, Martin Spott
18th Eurographics/IEEE VGTC Conference on Visualization, 2016 (EuroVis 2016)

The directionality of data -> visualisation in the grammar of graphics can also be limiting. What about visualisation -> data?

Principles of visualisation

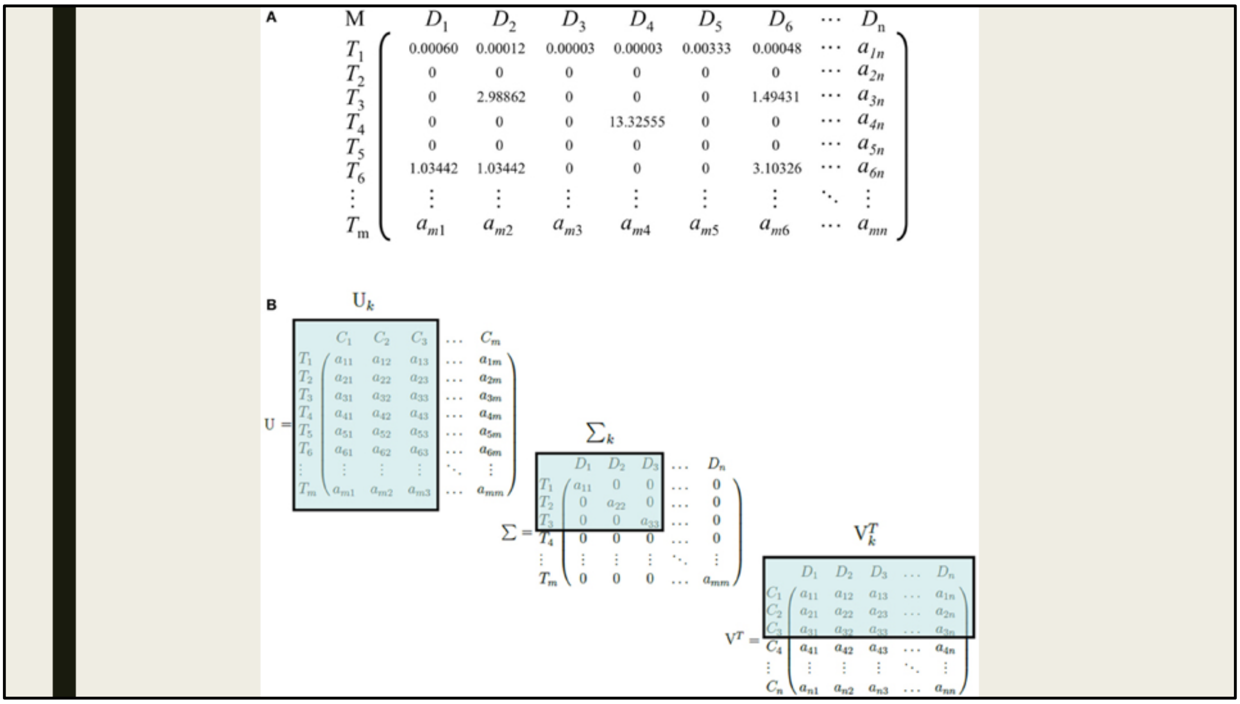
- Structural: e.g., Bertin, Wilkinson/Wickham
- Perceptual/cognitive: e.g., Bertin, Cleveland & McGill
- Aesthetic/designerly: e.g., Edward Tufte (Visual Display of Quantitative Information)



Yi, J. S., Kang, Y.-A., Stasko, J.,&Jacko, J. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*,*13*(6), 1224–31. https://doi.org/10.1109/TVCG.2007.70515

Lam, H. (2008). A framework of interaction costs in information visualization.*IEEE Transactions on Visualization and Computer Graphics*,*14*(6), 1149–56. https://doi.org/10.1109/TVCG.2008.109

LATENT SEMANTIC ANALYSIS



An example of a term-document matrix with a weighting function (tf-idf). M, D, and T refer to the term-document matrix, the set of all documents in the corpus, and the set of all terms in the corpus, respectively. T_1 is an example of a common word that occurs frequently in documents, whereas T_3, T_4, and T_6 are comparatively rarer words and receive a higher weight. **(B)** An illustration of the dimensionality-reduction step of LSI. U, Σ, and V^T are truncated and become Σ_k, U_k, and $V^T{}_k$, respectively. C, D, and T refer to the set of LSI topics, documents, and terms, respectively. Here, we illustrate a reduction to three dimensions.

These matrices can then be used as a distance metric for both terms and documents. Any two documents can be compared by computing the cosine distance between their corresponding column vectors in V^T. Likewise, any two terms can be compared by computing the cosine distance between their corresponding row rectors in U.

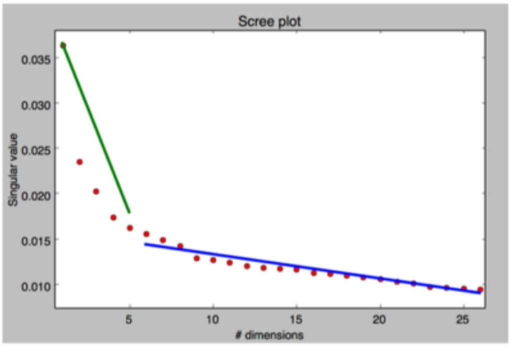
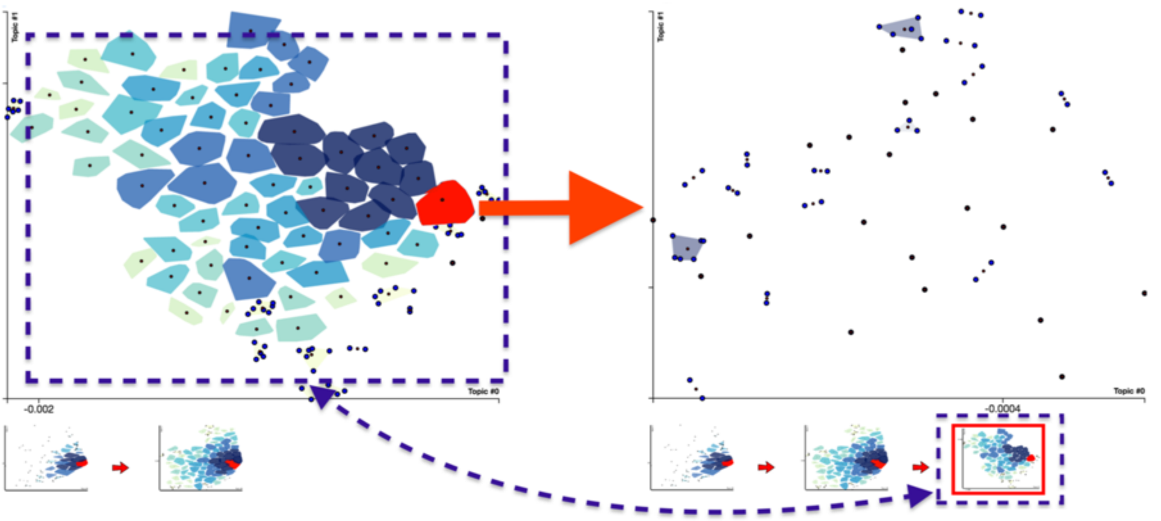


Figure 3.1: Singular value scree plot with a knee found by L-method at the 5^{th} singular value



Overview + detail, Semantic zooming, Graphical interaction histories

Figure 3.9: Random sampling performed in each scatte
very dense areas and a number of potentially interestin
and the shape is distorted. Sampling more values in w
performance.

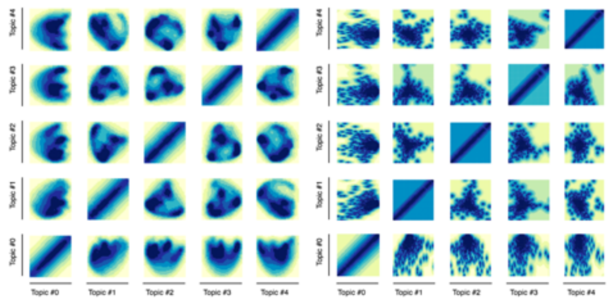


Figure 3.10: Two examples of heat map matrices. The colour scale ranging from light yellow to dark blue indicates the estimated probability density of the data distribution. Blue areas indicate higher probabilities of data points at that position.

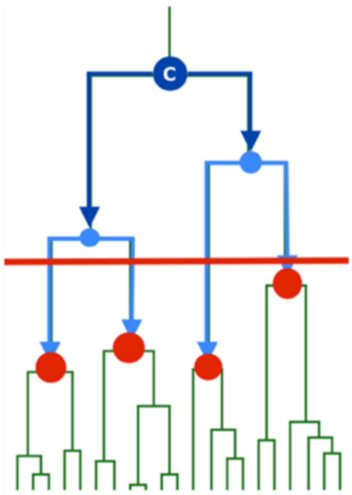


Figure 3.15: Obtaining the expansion of a cluster. To determine which clusters would become C's children in the expansion tree, a cut (in red) is made at the height corresponding to the minimum displayable distance between clusters. C's children are then expanded until the clusters immediately below the cut are reached; these are then chosen as C's expansion.

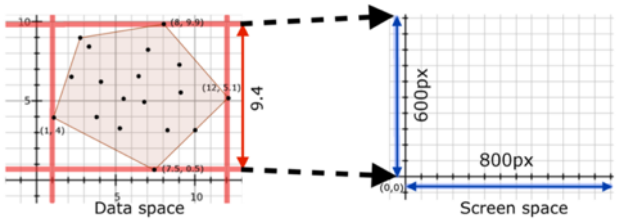


Figure 3.16: Mapping from data to screen space. The cluster shown is a cluster we want to expand and will be fragmented into its descendant clusters. By knowing the extent on one dimension in data space and the size of the y-axis in screen space, we can obtain a linear mapping between the two spaces. We can do the same for the other data dimension and x-axis.

BRAINCEL

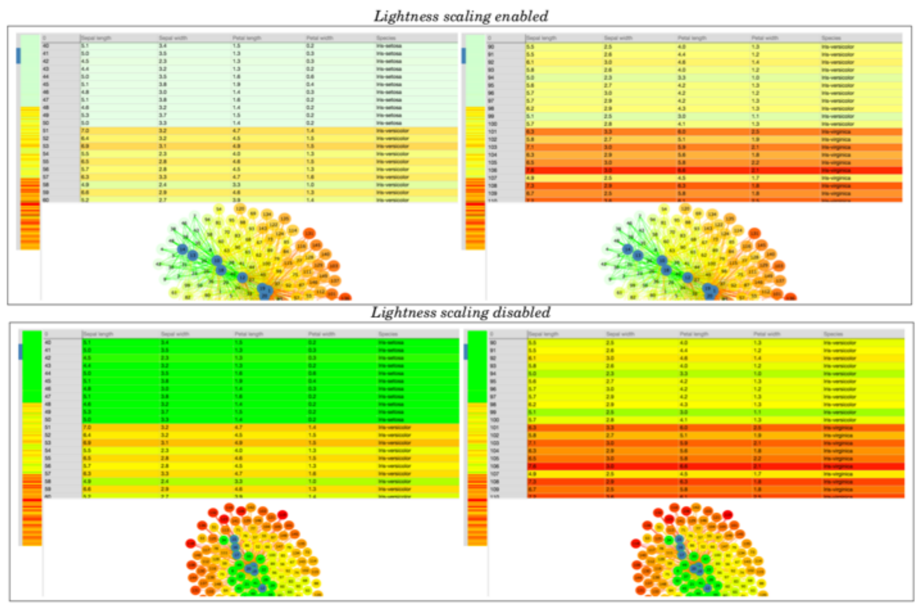
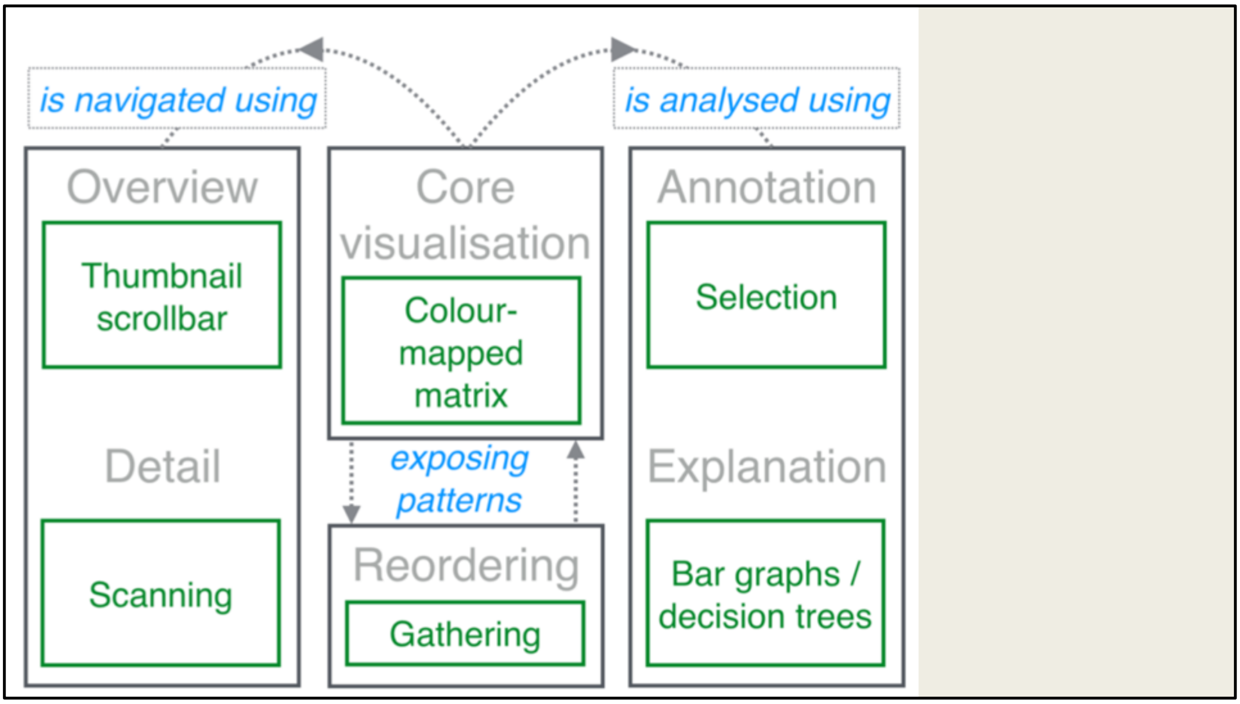
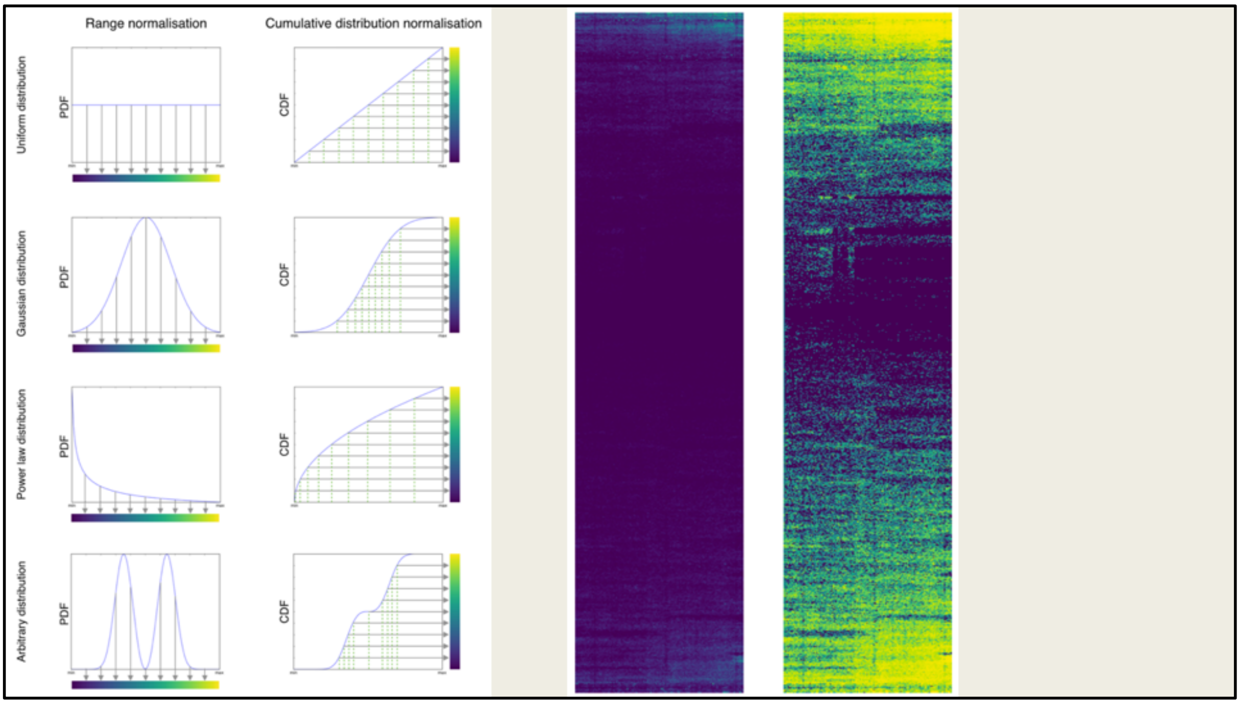
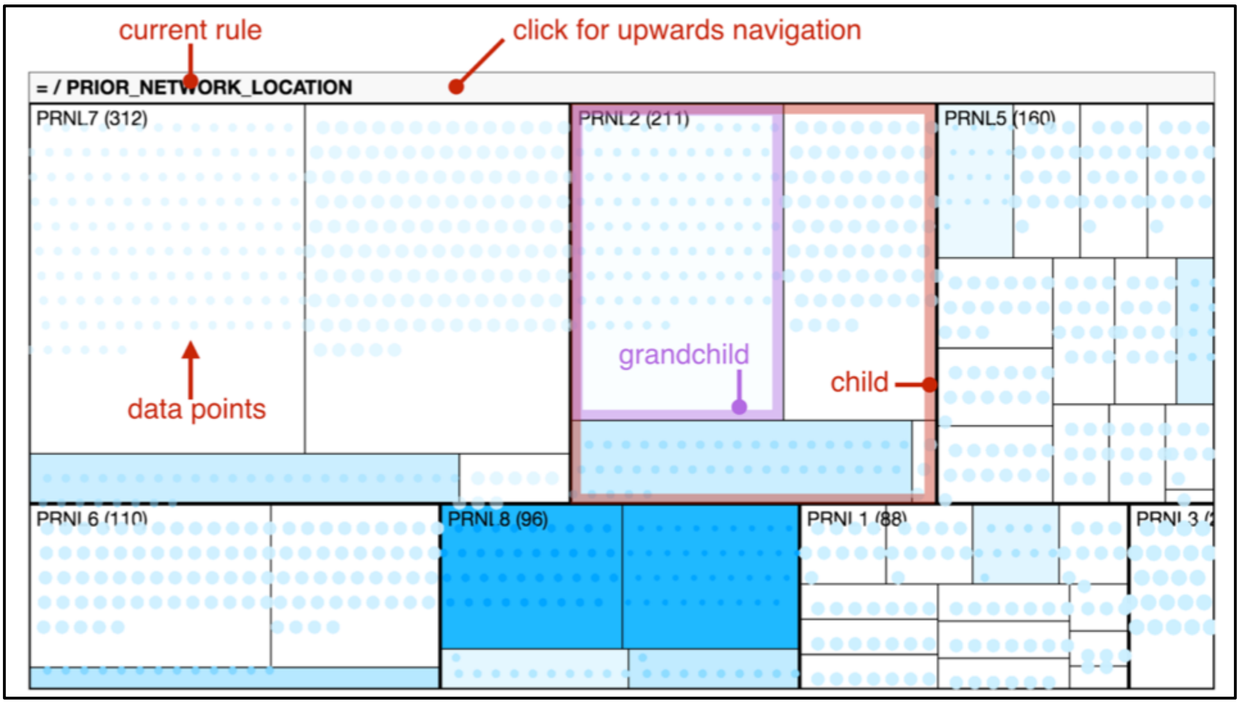


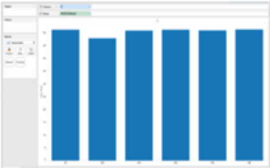
Figure 5.5: The effect of lightness scaling. Without lightness scaling, high-confidence (green) rows command disproportionately greater visual attention (the effect is most apparent onscreen).

GATHERMINER


is navigated using
is analysed using
Overview
Thumbnail scrollbar
Detail
Scanning
Core visualisation
Colour-mapped matrix
exposing patterns
Reordering
Gathering
Annotation
Selection
Explanation
Bar graphs / decision trees


Range normalisation
Cumulative distribution normalisation
Uniform distribution
Gaussian distribution
Power law distribution
Arbitrary distribution
PDF
CDF


current rule
click for upwards navigation
= / PRIOR_NETWORK_LOCATION
PRNL7 (312)
PRNL2 (211)
PRNL5 (160)
grandchild
child
data points
PRNL6 (110)
PRNL8 (96)
PRNL1 (88)

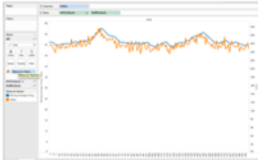


(a) This strategy involved comparing bar charts of each attribute-value pairing, aggregated over the entire span of time. Since the interesting features in our time series consisted of unusual spikes/troughs, this usually reflected in a higher/lower overall sum or average for those series – easily spotted in an unusually tall or short bar.

(b) This strategy involved comparing aggregate line charts of each attribute-value pairing. Here, any attribute-value that caused spikes or dips was clearly reflected.

(c) This interesting strategy also compared aggregate line charts of each attribute-value pairing. Here, by creating a 2D matrix of small multiples, the analyst was able to investigate the interaction of any two attributes.

Figure 4.15: Three successful strategies in Tableau.



(a) This strategy involved inspecting a completely aggregated line graph. In this dataset, we prepared a number of time series that had spikes at about 1/3 and 2/3 the duration of the series, which are clearly visible in the aggregate chart. However, there are also a number of series which have an upward spike in the halfway mark, and an equal number which have an equal and opposite downward spike at the same position. The two cancel each other out and become invisible in the aggregate line graph, and so the analyst never discovers them.

(b) This strategy, similar to the first successful strategy, uses summary bar graphs to represent the time series. However, since the series are completely disaggregated (one bar is generated per series), it is impossible to seek out global patterns.

(c) This strategy involved scanning through the entire list of time series, represented as line graphs, and manually noting down the attributes of any which appeared interesting. Needless to say, this is extremely ineffective and led to several false correlations being "discovered".

Figure 4.16: Three unsuccessful strategies using Tableau.