University of BRISTOL

Visual Analytics: View Reduction

Ian Nabney

ian.nabney@bristol.ac.uk

bristol.ac.uk

University of BRISTOL

Overview

- Reading: Chapter 13 and 14 of Munzner
- Understand the design choices to reduce (or increase) what is shown in a view
- Able to apply appropriate choices to design challenges

Big Picture

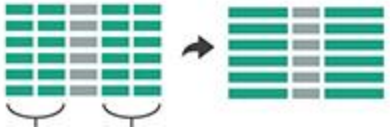
Reducing Items and Attributes

- Filter
  - Items

  - Attributes

- Aggregate
  - Items

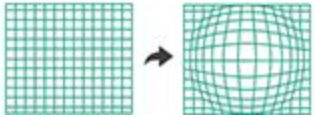
  - Attributes



- Embed
  - Elide Data

  - Superimpose Layer

  - Distort Geometry



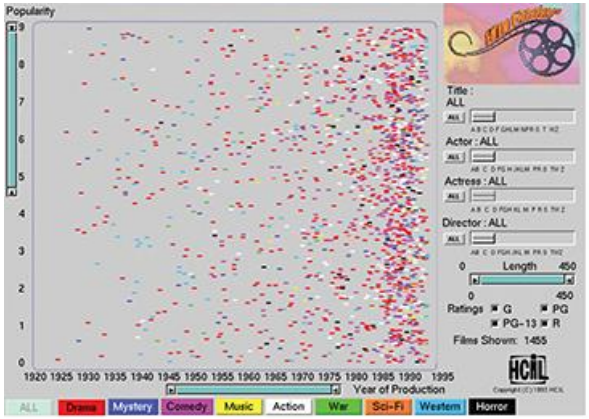
Why reduce?

- Reduction is one of five major strategies for managing complexity in visualizations (others are deriving new data, changing a view over time, faceting data, and embedding focus)
- **Static** data reduction idioms only reduce what is shown
- In the **dynamic** case, the outcome of changing a parameter or a choice may be an increase in the number of visible elements
- Reduction can be applied to both items and attributes
- **Filtering** simply eliminates elements, whereas **aggregation** creates a single new element that stands in for multiple others that it replaces
- Filtering is very straightforward for users to understand, and typically also to compute. However, people tend to forget to take into account elements that have been filtered out.
- Aggregation can be somewhat safer from a cognitive point of view because the stand-in element is designed to convey information about the entire set of elements that it replaces. However, by definition, it cannot convey all omitted information; the challenge with aggregation is how and what to summarize in a way that matches well with the dataset and task.

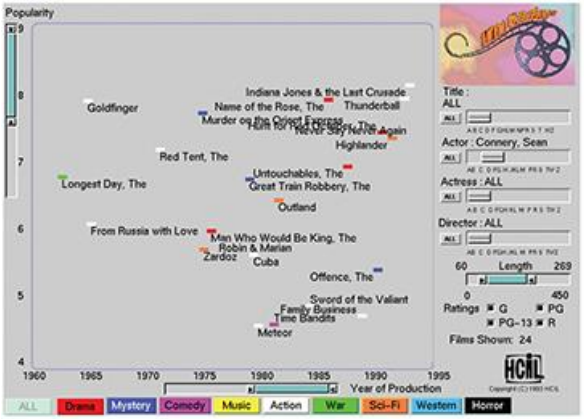
Filtering

- Allow the user to select one or more ranges of interest in one or more of the elements. The range might mean what to show or what to leave out.
- Consider the simple case of filtering the set of items according to their values for a single quantitative attribute.
  - The goal is to select a range within in terms of minimum and maximum numeric values.
  - From the **programmer's** point of view, a very simple way to support this functionality would be to simply have the user enter two numbers, a minimum and maximum value.
  - From the **user's** point of view, this approach is very difficult to use: how do they know what numbers to type? After they type, how do they know whether that choice is correct?
- In an **interactive** vis, filtering is often accomplished through dynamic queries, where there is a tightly coupled loop between visual encoding and interaction, so that the user can immediately see the results of the intervention.

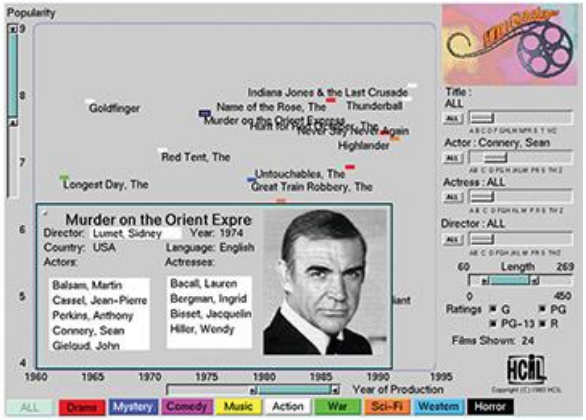
Item filtering



(a)



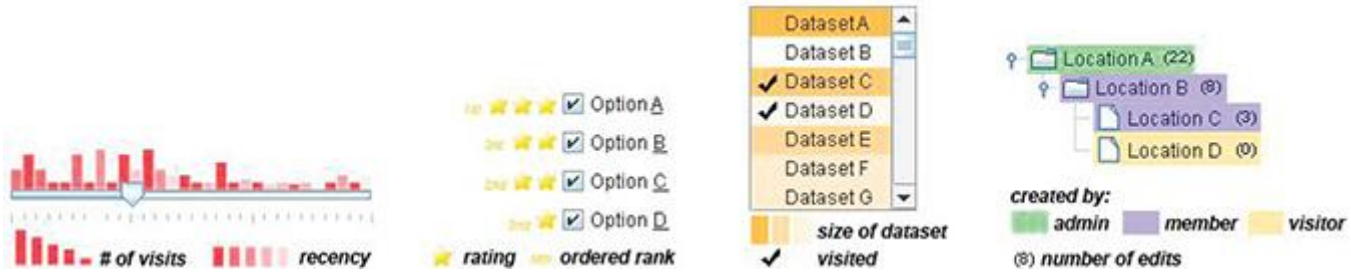
(b)



(c)

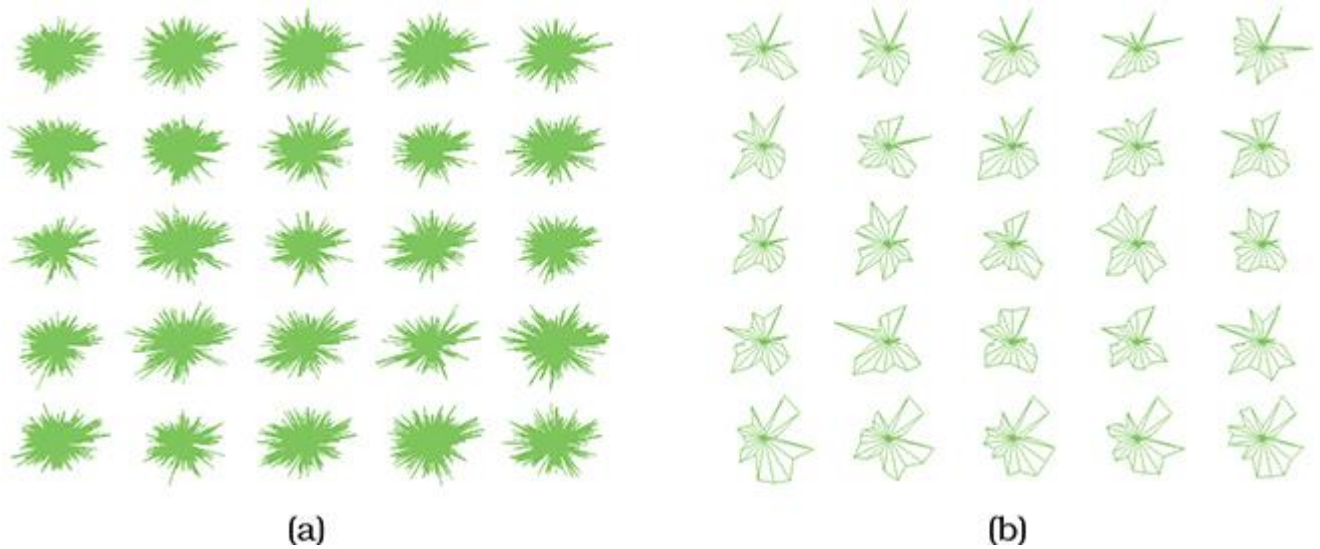
- FilmFinder system
- Dataset has nine attributes: genre, year made, title, actors, actresses, directors, rating, popularity, and length.
- Vis contains an interactive scatterplot where the items are movies colour coded by genre, with scatterplot axes of year made versus movie popularity.

Filter controls



- Standard widgets for filtering controls can be augmented by concisely visually encoding information about the dataset
- The idea is to create displays that have high information density. These augmented widgets are called **scented widgets**
- Cues that help a searcher decide whether there is value in drilling down further into a particular information source, versus looking elsewhere

Filtering attributes



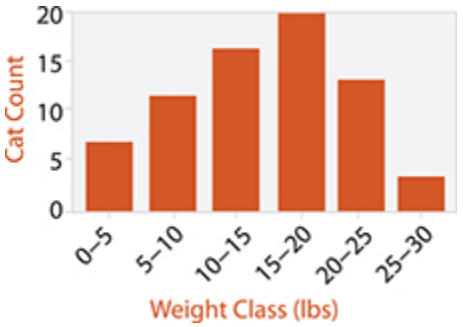
- Dimensional Ordering, Spacing, and Filtering Approach (DOSFA) idiom
- In Figure (a), plot axes are so densely packed that little structure can be seen.
- Figure (b) shows the plots after the dimensions are ordered by similarity and filtered by both similarity and importance thresholds. The filtered display does show clear visual patterns.
- Attribute similarity can be measured using correlation.

Aggregation

- A group of elements is represented by a new derived element that represents the entire group.
- Elements are merged together with aggregation, as opposed to eliminated completely with filtering.
- Aggregation typically involves the use of a derived attribute. A very simple example is computing an average; the four other basic aggregation operators are minimum, maximum, count, and sum.
- Several standard plots can be viewed as aggregation.

Histograms

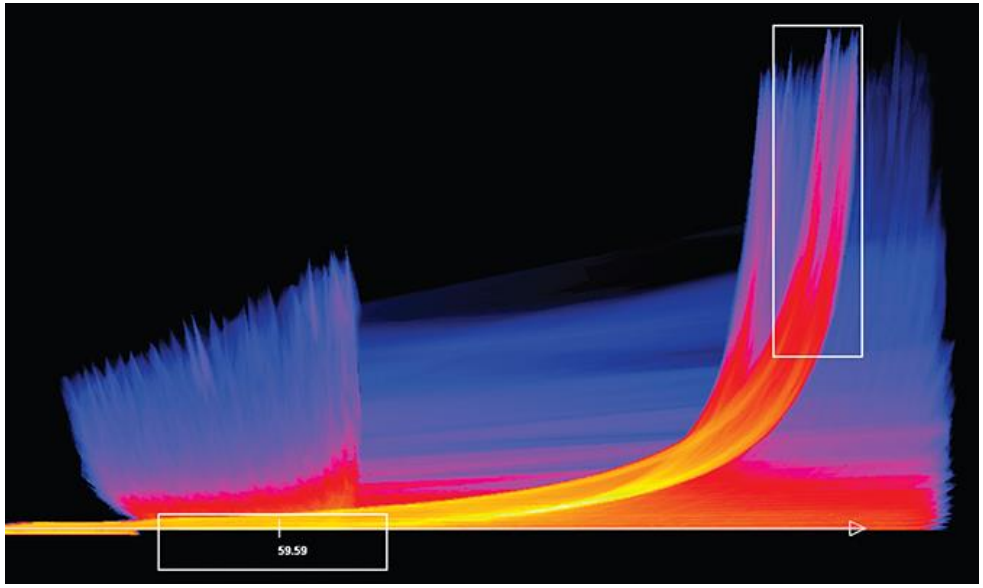
- Range of the original attribute is partitioned into bins, and the number of items that fall into each bin is counted
- Visual encoding of a histogram is very similar to bar charts
- One difference is that histograms are sometimes shown without space between the bars to visually imply continuity, whereas bar charts have spaces between the bars to imply discretization
- Choice of bin size is crucial and tricky: a histogram can look quite different depending on the discretization chosen.
  - One possible solution is to compute the number of bins based on dataset characteristics;
  - another is to provide the user with controls to easily change the number of bins interactively, to see how the histogram changes
- Variable width bins can be used to avoid noisy histograms (e.g. by making the count in each bin roughly equal
- An interesting rule of thumb for the number of bins is $\sqrt[3]{n}$ where *n* is the number of data points



Histogram of the distribution of weights for all of the cats in a neighbourhood, binned into 5-pound blocks

Continuous scatterplots

- Scatterplots for large numbers of data points because the marks occlude each other and because interest shifts to the density of points (hard to estimate by eye).
- Solved by plotting an aggregate value at each pixel rather than drawing every single item as an individual point
- Continuous scatterplots use colour coding at each pixel to indicate the density of overplotting, often in conjunction with transparency
- Continuous scatterplots use a dense, space-filling 2D matrix alignment, where each pixel is given a different colour

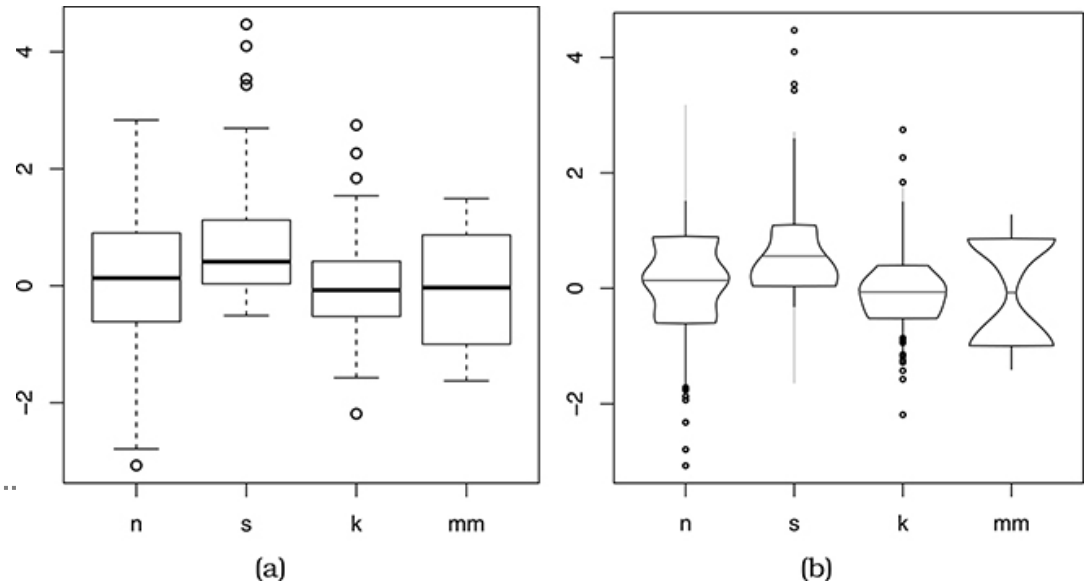


- Tornado air-flow dataset, with the magnitude of the velocity on the horizontal and the z-direction velocity on the vertical
- Derived table: two ordered key attributes (x, y pixel locations), one quantitative attribute (overplot density).
- Dense space-filling 2D matrix alignment, sequential categorical hue + ordered luminance colormap.

Boxplots

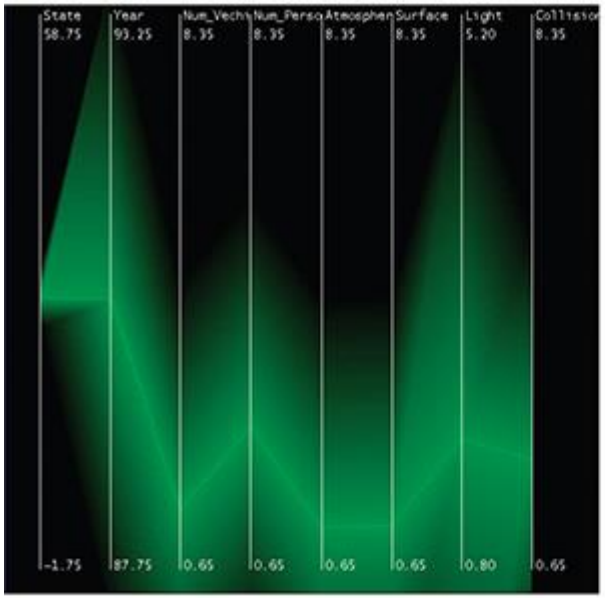
- **Boxplots** shows an aggregate statistical summary of all the values that occur within the distribution of a single quantitative attribute.
- Five derived variables carefully chosen to provide information about the attribute's distribution: the median (50% point), the lower and upper quartiles (25% and 75% points), and the upper and lower fences (chosen values near the extremes, beyond which points should be counted as outliers).

- These plots illustrate four kinds of distributions: normal (n), skewed (s), peaked (k), and multimodal (mm).
- (a) Standard box plots.
- (b) Vase plots, which use horizontal spatial position to show density directly.

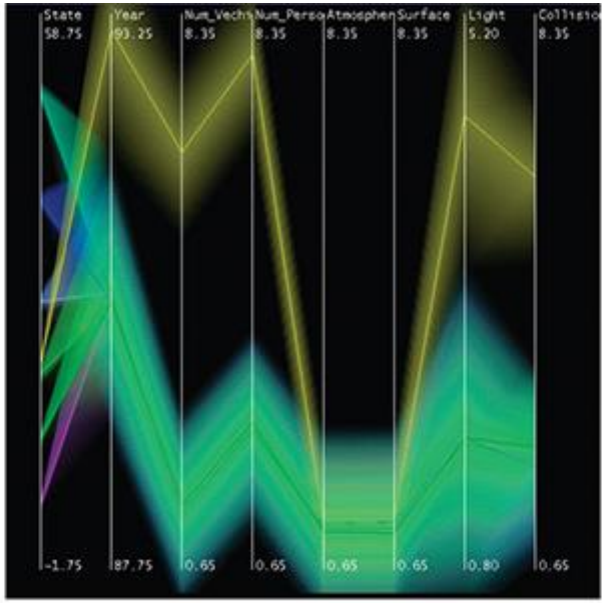


From [Wickham and Stryjewski 12, Figure 5].

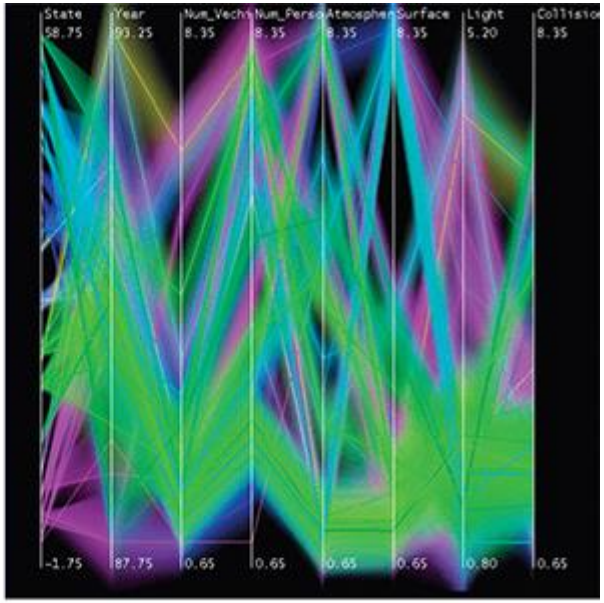
Hierarchical parallel coordinates



(a)



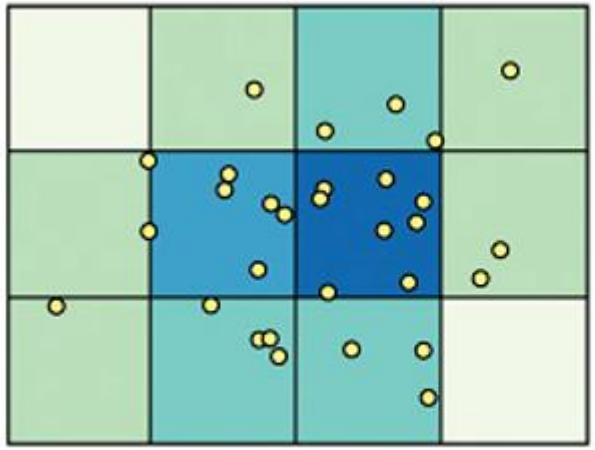
(b)



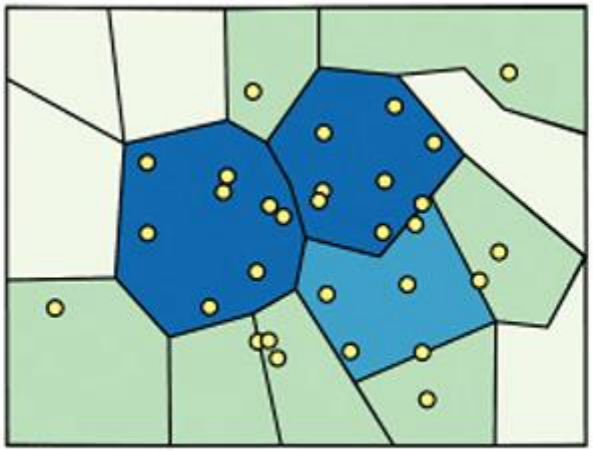
(c)

- The dataset is transformed by computing derived data: a hierarchical clustering of the items.
- A cluster is represented by a band of varying width and opacity, where the mean is in the middle and width at each axis depends on the minimum and maximum item values for that attribute within the cluster.
- The cluster bands are colored according to their proximity in the cluster hierarchy, so that clusters far away from each other have very different colors.

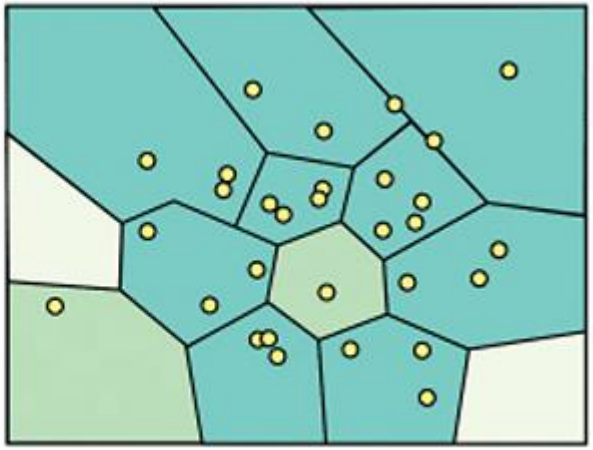
Spatial aggregation



(a)



(b)



(c)

- **Modifiable areal unit problem (MAUP)**: changing the boundaries of the regions used to analyse data can yield dramatically different results
- The same location near the middle of the map has a different density level depending on the region boundaries: high in Figure (a), medium in Figure (b), and low in Figure (c).
- A well-known example of this is **gerrymandering** with the boundaries of voting districts.

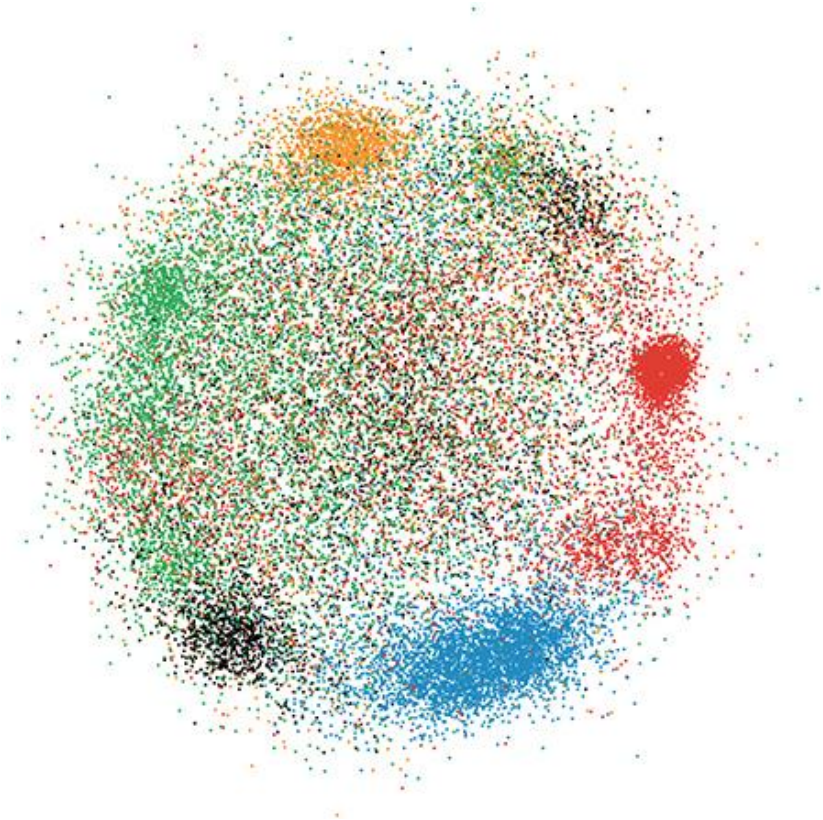
Illinois congressional district 4

Attribute aggregation

- A new attribute is synthesized to take the place of multiple original attributes.
- A very simple approach to aggregating attributes is to group them by some kind of similarity measure, and then calculate an average across that similar set.
- A more complex approach to aggregation is **dimensionality reduction**, where the goal is to preserve the meaningful structure of a dataset while using fewer attributes to represent the items.
- Nonlinear methods for dimensionality reduction are used when the new dimensions cannot be expressed in terms of a straightforward combination of the original ones. The **multidimensional scaling** (MDS) family of algorithms includes both linear and nonlinear variants, where the goal is to minimize the differences in distances between points in the high-dimensional space versus the new lower-dimensional space.
- Much more on this in second half of course

Example: document collections

- Use **bag of words** feature space: huge number of quantitative attributes
- The user can interactively navigate the scatterplot
- The user's goal is cluster discovery: either to verify an existing conjecture about cluster structure or to find previously unknown cluster structure
- Selecting a point shows document keywords in a popup display and the full text of the document in another view
- In the third step, the user's goal is to produce annotations by adding text labels to the verified clusters



Dimensionality reduction of a large document database using Glimmer for multidimensional scaling

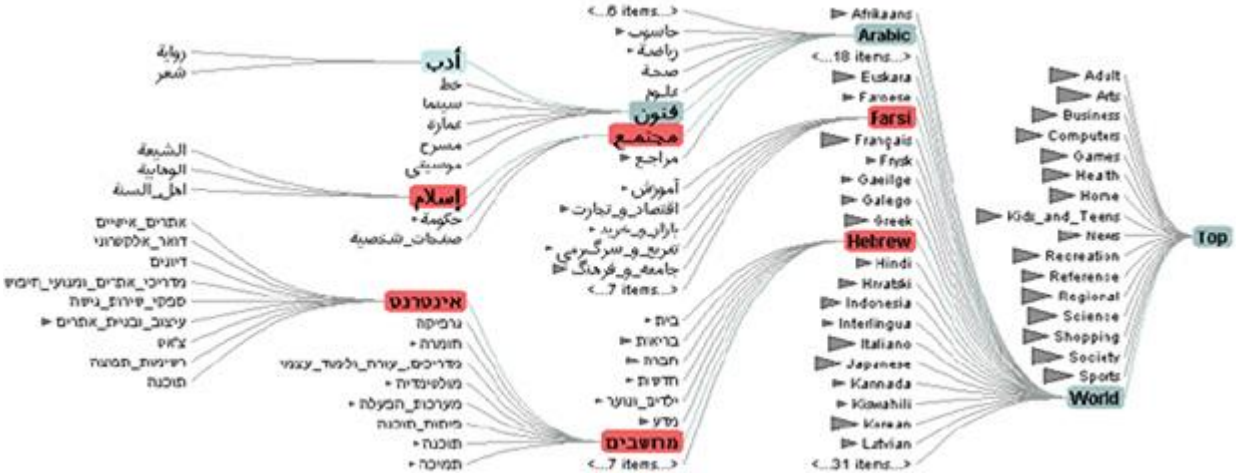
Embed: focus and context

- The family of idioms known as **focus+context** are based on the design choice to embed detailed information about a selected set—the focus—within a single view that also contains overview information about more of the data—the context.
- These idioms reduce the amount of data to show in the view through sophisticated combinations of filtering and aggregation.
  - One design choice for embedding is to **elide** items, where some items are filtered out completely while others are summarized using dynamic aggregation for context; only the focus items are shown in detail.
  - Another choice is to **superimpose** layers, where a local region of focus information can be moved against the background layer of contextual information.
  - A third choice is to **distort** the geometry, where context regions are compressed to make room for magnified focus regions.

Why embed?

- The goal of embedding focus and context together is to mitigate the potential for disorientation that comes with standard navigation techniques such as geometric zooming.
- With realistic camera motion, only a small part of world space is visible in the image when the camera is zoomed in. With geometric navigation and a single view that changes over time, the only way to maintain orientation is to internally remember one's own navigation history.
- Focus+context idioms attempt to support orientation by providing contextual information as recognizable landmarks, using external memory to reduce internal cognitive load.
- Embedding idioms are fundamentally a synthesis of visual encoding and interaction. The key idea of focus+context is that the focus set changes dynamically as the user interacts with the system, and thus the visual representation also changes dynamically.
- Many of the idioms involve indirect control, where the focus set is inferred via the combination of the user's navigation choices and the inherent structure of the dataset.

Elision

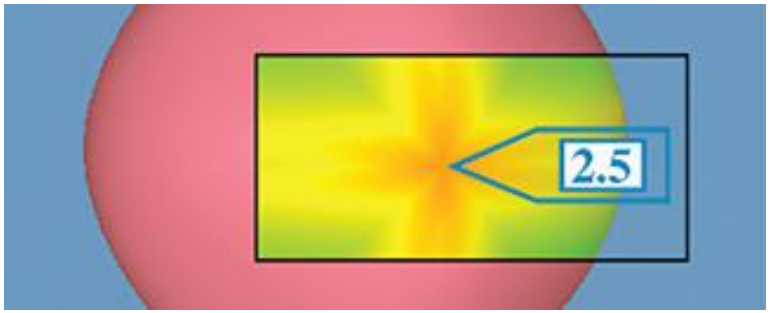


Degree of Interest Trees: http://krisrs1128.github.io/treelapse_expers/doi_supp/doi_exper.html
Abundance between taxonomic groups at varying levels of locality

- Some items are omitted from the view completely, in a form of dynamic filtering. Other items are summarized using dynamic aggregation for context, and only the focus items are shown in detail.
- **Degree of interest (DOI) function**: DOI = I(x) D(x, y). I is an interest function; D is the distance, either semantic or spatial; x is the location of an item; y is the current focus point
- The example uses multiple foci to show an elided version of a 600,000 node tree.

Superimpose

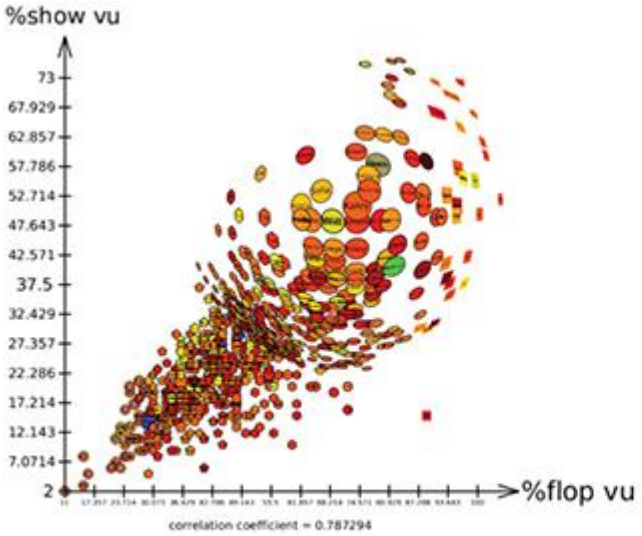
- The Toolglass and Magic Lenses system uses a see-through lens to show color-coded Gaussian curvature in a foreground layer, atop the background layer consisting of the rest of the 3D scene.
- Within the lens, details are shown, and the unchanged remainder of the other view provides context. The lens layer occludes the region beneath it.
- The system handled many different kinds of data with different visual encodings of it; this example shows 3D spatial data.



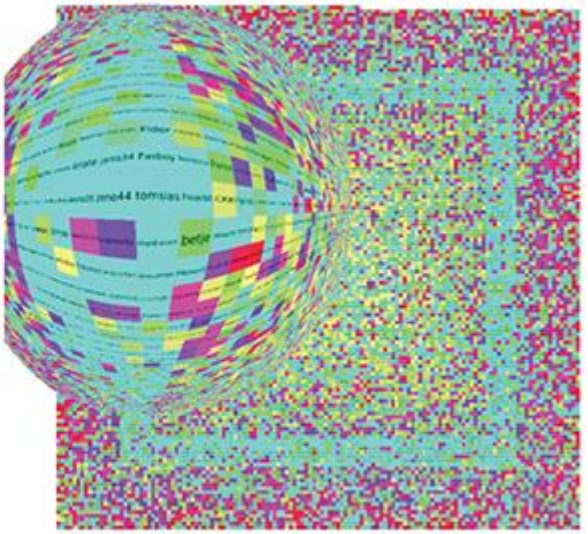
Distortion

- Integrating focus and context into a single view using **geometric distortion** of the contextual regions to make room for the details in the focus regions.
- Choice of the number of focus regions: is there only a single region of focus, or does the idiom allow multiple foci?
- Shape of the focus: is it a radial, rectangular, or a completely arbitrary shape?
- Extent of the focus: is it global across the entire image, or constrained to just a local region?
- Interaction metaphor:
  - constrained geometric navigation
  - moveable lenses, evocative of the real-world use of a magnifying glass lens
  - stretching and squishing a rubber sheet
  - vector fields.

Example: fisheye lens



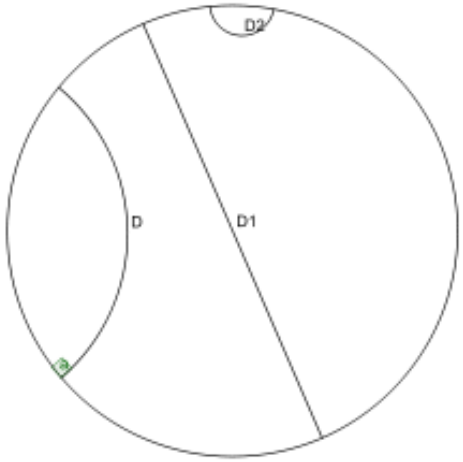
(a)



(b)

- Single focus with local extent and radial shape and the interaction metaphor of a draggable lens on top of the main view
- Foreground layer that completely replaces what is beneath it, like the magic lens
- Under the fisheye lens, the labels are large enough to read; that focus region remains embedded within the surrounding context, showing the global pattern within the rest of the dataset

Hyperbolic geometry



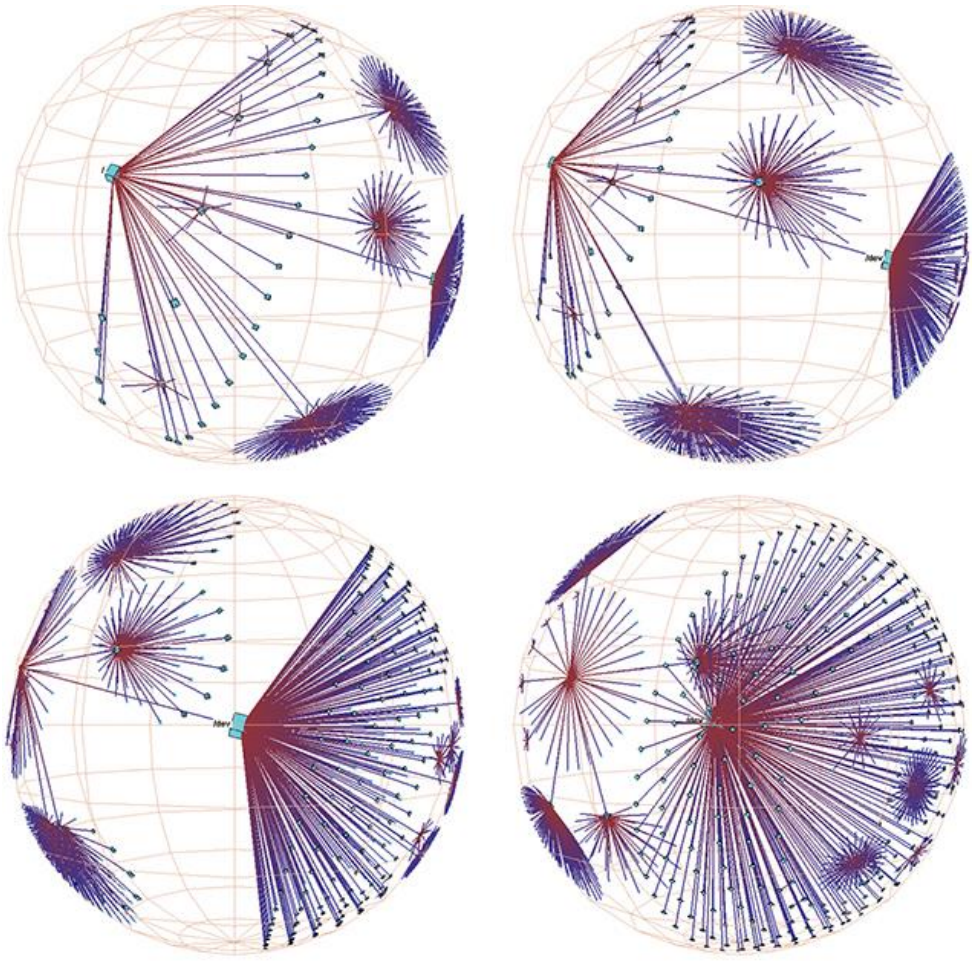
Poincaré disc with 3 hyperbolic straight lines

Circle Limit III

Circle Limit IV

- Hyperbolic geometry is a non-Euclidean geometry where for any given line L and point P not on L, in the plane containing both line L and point P there are at least two distinct lines through P that do not intersect L.
- In the Poincaré model, the plane is represented by a disc, and straight lines by diagonals or arcs of (Euclidean) circles that intersect the boundary at right angles.
- There are interesting infinite tessellations (tilings) of the plane. The two pictures were created by M.C. Escher with advise from the mathematician Coxeter.

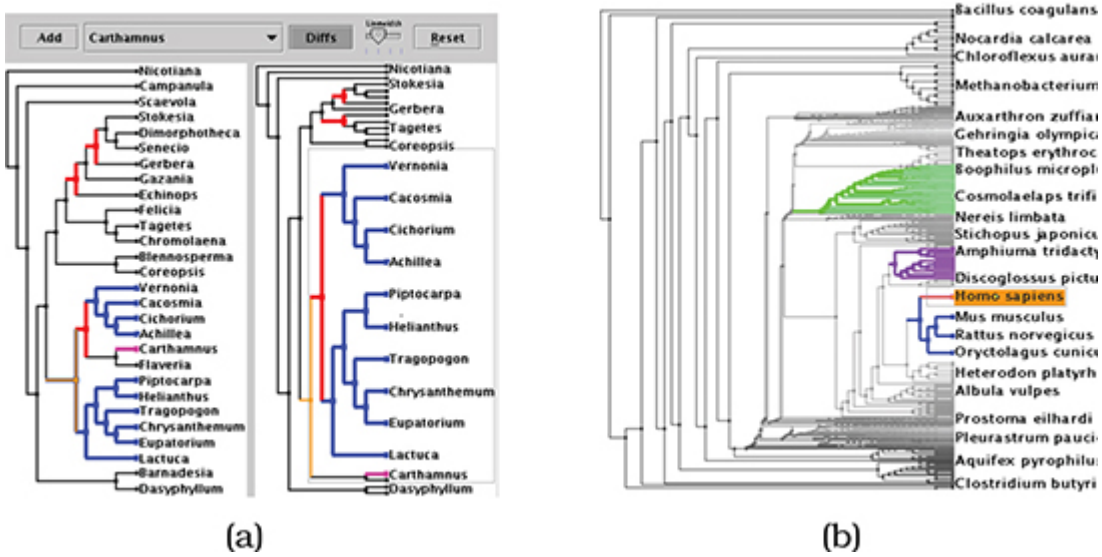
Hyperbolic geometry for vis



- 3D hyperbolic node–link tree representing the structure of a file system laid out with the H3 idiom, through a sequence of frames from an animated transition as the view changes over time.
- The first three frames show hyperbolic translation to change what part of the tree is magnified, where the subtree on the right side gets larger as it moves toward the centre of projection.

Stretch and squish

- **Stretch and squish** navigation uses multiple rectangular foci of global extent for distortion.
- Enlarging some regions causes others to shrink.
- The borders of the sheet stay fixed so that all items stay visible within the viewport, although they may be projected to arbitrarily small regions of the image.
- The user can choose to separately stretch the rectangular focal regions in the horizontal and vertical directions.



(a) (b)

TreeJuxtaposer uses stretch and squish navigation with multiple rectangular foci for exploring phylogenetic trees.
(a) Stretching a single region
(b) Stretching multiple regions

Evaluation of distortion

- Distortion-based focus+context in particular has measurable costs.
- Distance or length judgements are severely impaired, so distortion is a poor match with any tasks that require such comparisons. Thus, one of the most successful use cases for geometric distortion is with exploration of node–link layouts for networks and trees. The task of understanding the topological structure of the network is robust to distortion because precise angle and length judgements are not necessary.
- Users may not be aware of the distortion, and thus misunderstand the underlying object structure. This risk is highest when the user is exploring an unfamiliar or sparse structure, and many idioms incorporate explicit indications of distortion to lessen this risk. Hyperbolic views typically show the enclosing circle or sphere, magnification fields often show a superimposed grid or shading to imply the height of the stretched surface.
- Internal overhead of maintaining object constancy, which is the understanding that an item seen in two different frames represents the same object. Understanding the underlying shape of a complex structure could require mentally subtracting the effect of the transformation in order to recognize the relationship between the components of an image before and after the transformation. Although in most cases we do this calculation almost effortlessly for standard 3D perspective distortion, the cost of mentally tracking general distortions increases as the amount of distortion increases.

Conclusions

- Reduction is one of the five major strategies for managing complexity in visualisations
- Filtering and aggregating are commonly used methods for simplifying what the user sees, but removing information must be done with care
- Dimensionality reduction is a powerful method for maintaining the structural distance-based relationships in high-dimensional data while visualising it in 2D
- Embedding supports a user-defined simplification or overlay of information
- Distortion should be used with care: it is best when **what** is shown is more important than **how** it is shown (so the idiom matters less)