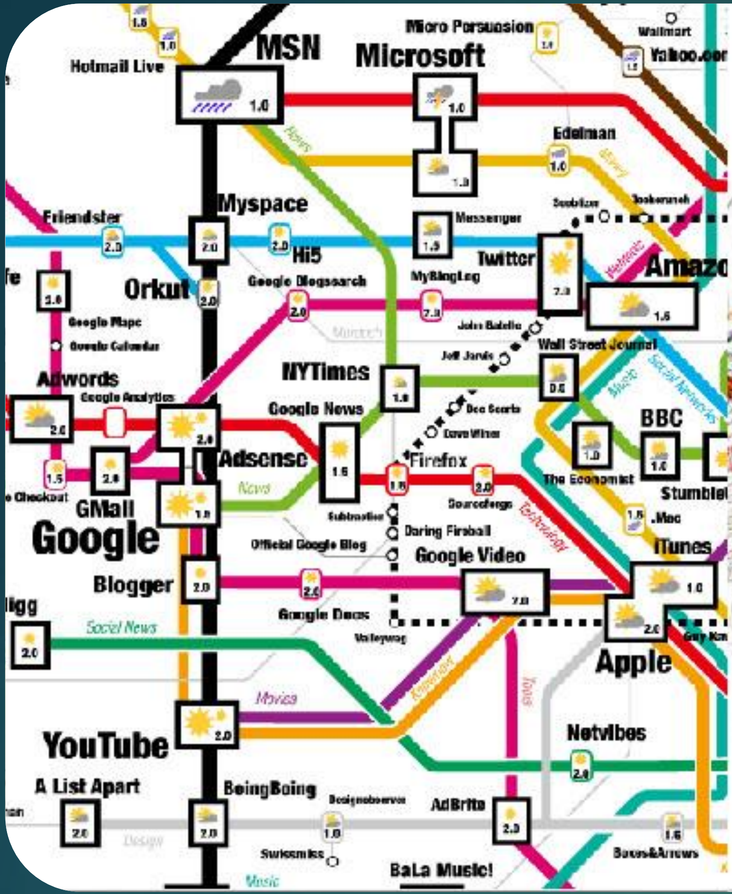
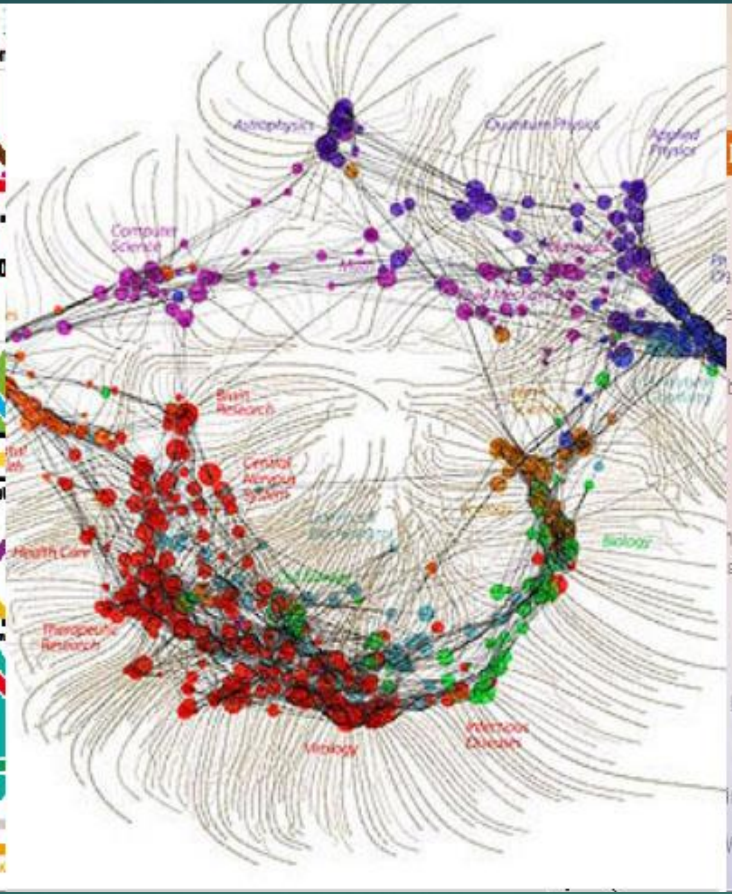
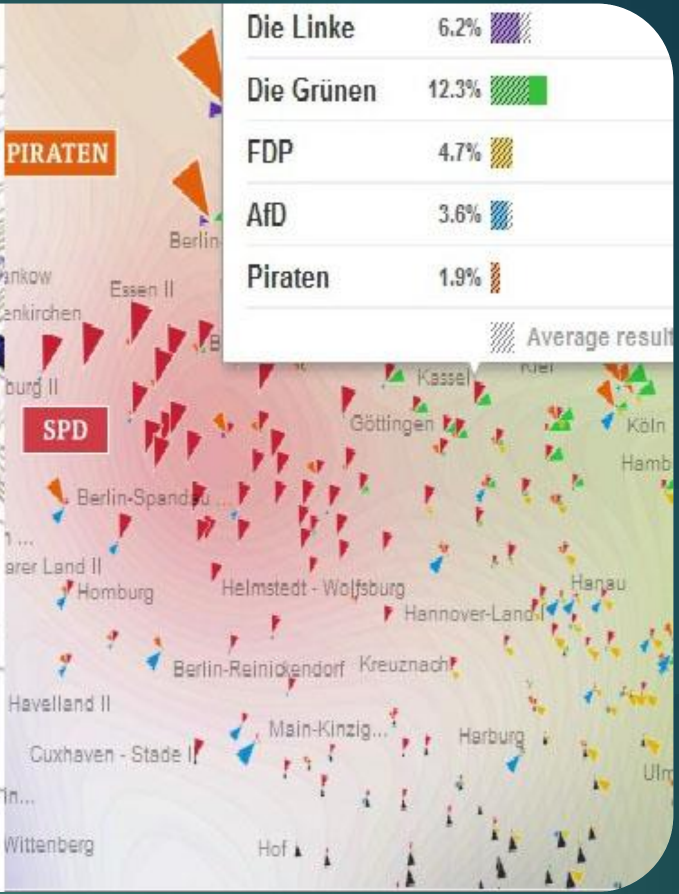






Data Visualization

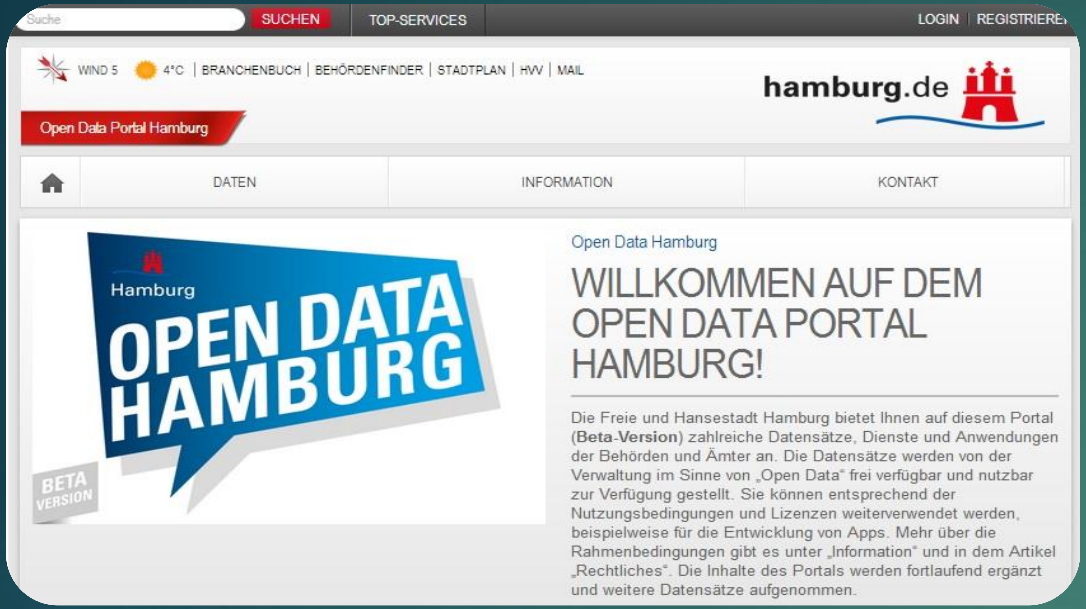
VINH PHAN

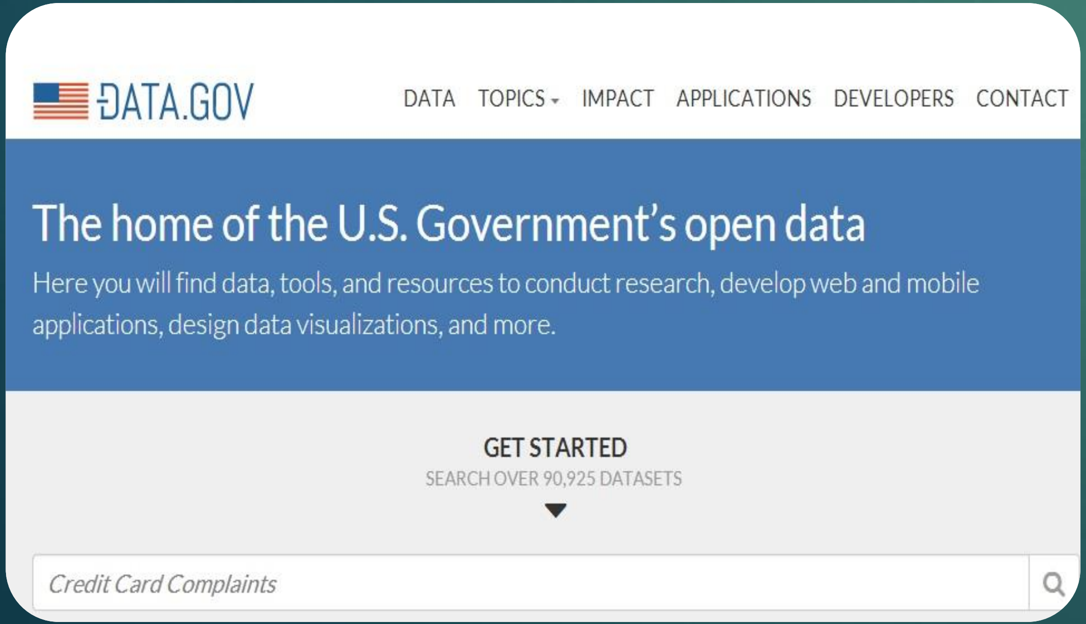
AW2 – 17/04/2014

HAW HAMBURG

Image Sources:
1. http://ia.net/blog/webtrends2007/
2. http://infosthetics.com/
3. http://www.zeit.de/politik/deutschland/wahlland

Introduction





- Trends: Open data/transparency
  - http://daten.hamburg.de
  - http://data.gov
- Huge datasets
- Multivariate data
- Exploratory graphics for data analysis
  - Used in bulk
  - Fast, informative
  - No detailed legends/captions, etc.
  - User interactions

Source: [Open Data Hamburg/Data.gov]

Agenda

- I. Tree Visualization with Tree-Maps: A 2-D Space-Filling Approach
- II. Trellis Displays for High-Dimensional Data Visualization
- III. Linked Views for Visual Exploration
- IV. Conclusion

Motivation

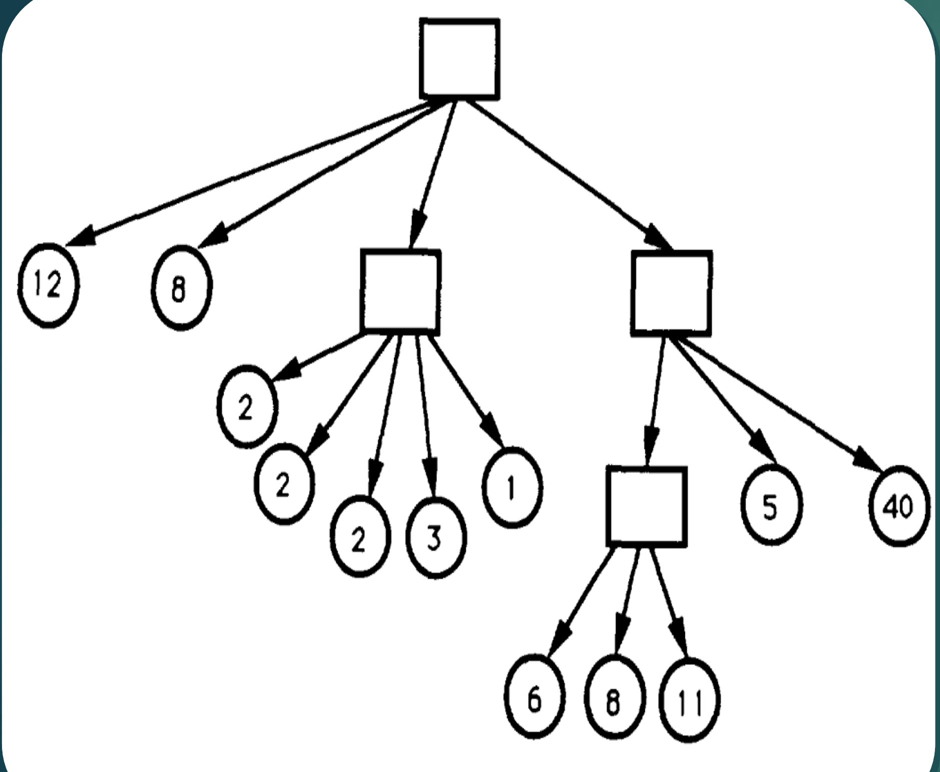


Fig. 1. Typical 3-level tree structure with numbers indicating size of each leaf node.

- Tree structures: traditionally represented as rooted, directed graph.
- Several alternatives were discussed, including a briefly-mentioned space-filling approach by [Knuth68], the rest mainly focused on node and edge representations.
- Related research projects include quad-trees, XY-trees, and k-d trees. But none focus on visualization aids for large tree structures.
- Motivation: to gain a better representation of complex traditional tree structures

Fig. 1. Typical 3-level tree structure with numbers indicating size of each leaf node.
Source: [shneiderman92]

Introduction

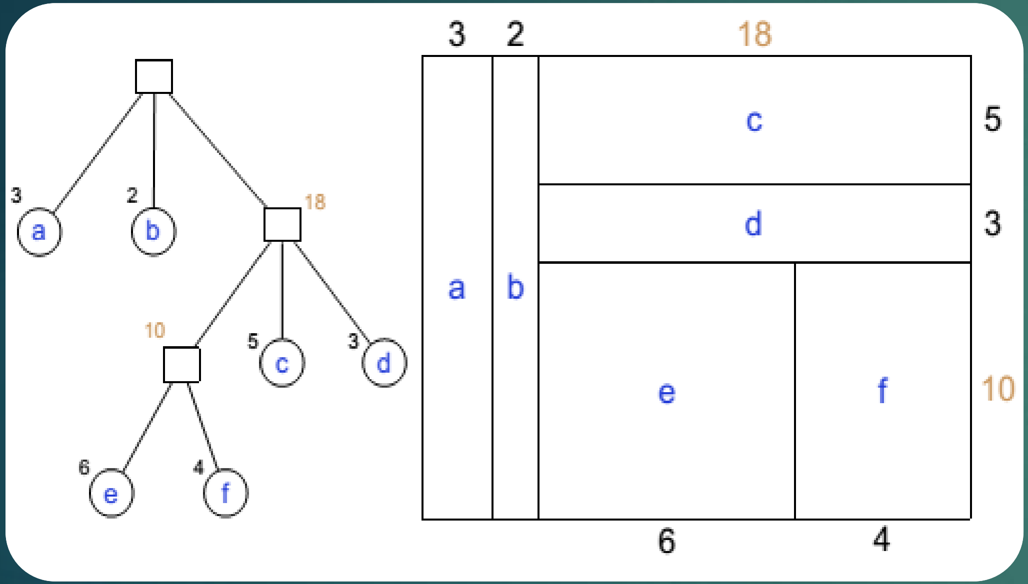


Fig. 2. Tree-map representation (right) of the tree on left with numerical values assigned to each leaf node and parent nodes report the sum of values of their child nodes.
Source: [Introduction to Information Visualization]

- Problem: limited screen estate for large tree structure, node info (size, importance) cannot be shown.
- Most OS show one node at a time with node info/tree structure using indentation.
- Goal: visualize entire set of files & show context info. Files can be of any size and at any level of tree structure.
- Tree-maps algorithm: 2-D space-filling approach, each node is a rectangle whose area is proportional to some variable.
  - Advantages: drawable, comprehensible and not computationally intensive.

I. Tree Visualization with Tree-Maps: A 2-D Space-Filling Approach

Algorithm

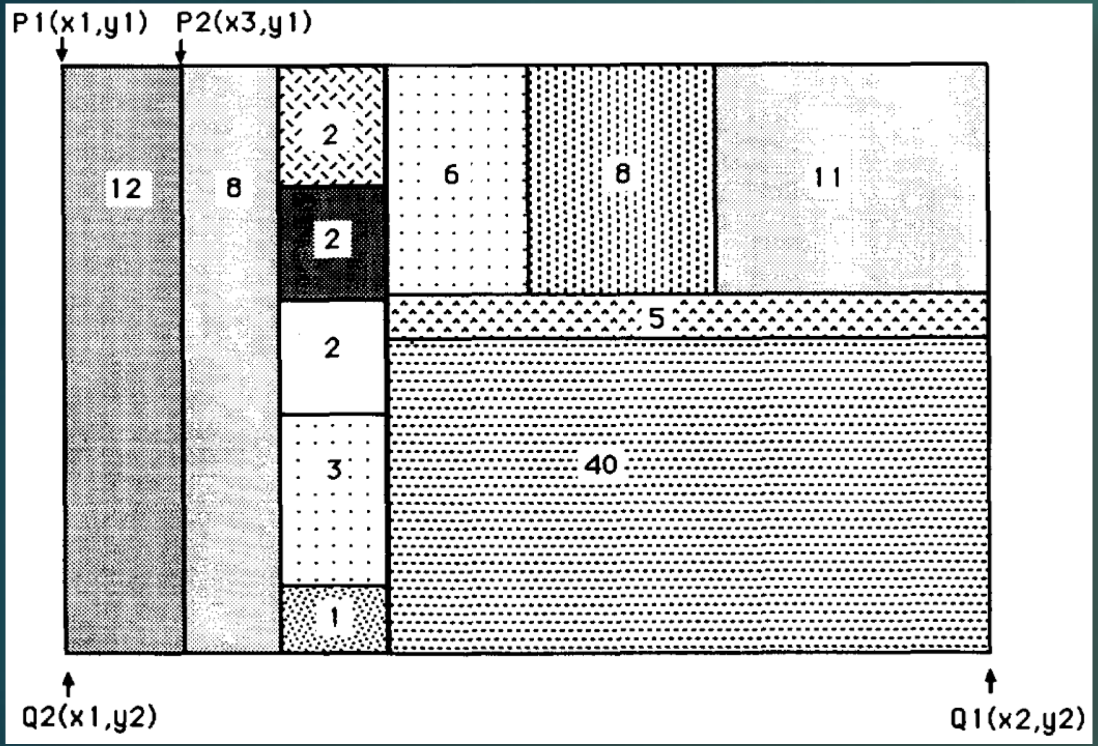


Fig. 3. Tree-map representation of Fig. 1
Source: [Shneiderman92]

$x3 = x1 + (Size(child[1] / Size(root)) * (x2 – x1)$

```
Treemap(root, P[0..1], Q[0..1], axis, color)
  Paint_rectangle(P, Q, color)        —paint full area
  width := Q[axis] - P[axis]          —compute location of next slice
  for i := to num_children do
    Q[axis] := P[axis] + (Size(child[i])/Size(root))*width
    Treemap(child[i],                                  recur on each slice,
    P, Q, 1 - axis, color)                             flipping axes
    P[axis] := Q[axis]
  endfor
```

- root: pointer to the root of the tree/subtree

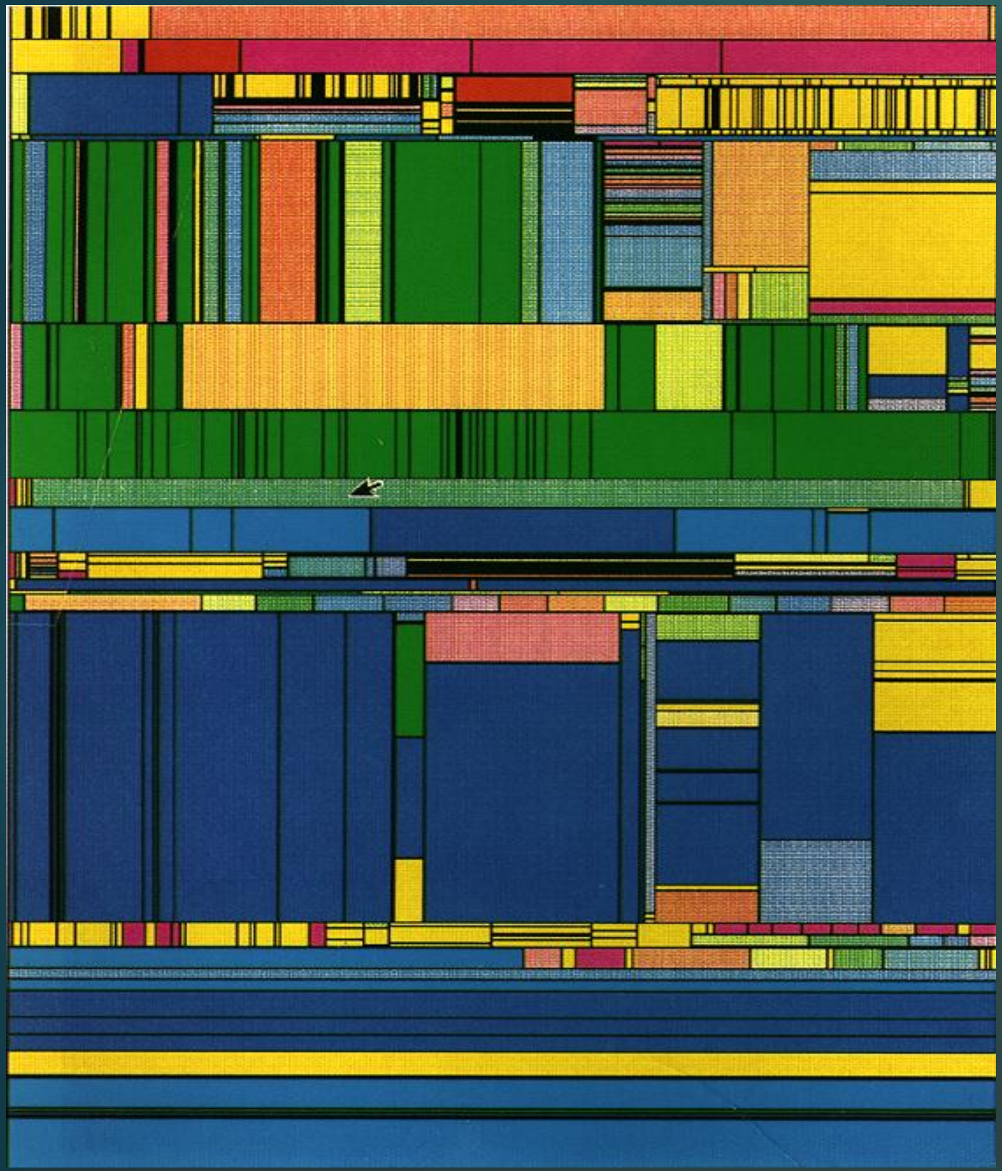
- P, Q: arrays of length 2 containing coordinate pairs (x,y) of opposite corners of the current rectangle

- axis: either 0 or 1, indicates cuts to be made vertically/horizontally

- color: color to be used for the current rectangle

Each node contains record with its fir/file name (name), number of children (num_children), array of pointers to next level (child[1..num_children]

Algorithm Assessment & Applications



- Performance: O(n)
- Depth-first traversal, rectangles are painted left to right, top to bottom (breadth-first also possible)
- Context info can be shown via mouse-over/mouse click, other operations via pop-up menus
- Display: VGA by 640 x 480 enough for one – two thousand files (small/zero byte files ignored)
- Applications: organization chart (employee/budget), library's holdings chart, stock portfolio, etc.

Fig. 4. 850 files at 4 levels with color coding by file types and context info on mouse-over. Source: [Shneiderman92]

I. Tree Visualization with Tree-Maps: A 2-D Space-Filling Approach
Newsmap

- Color coding indicates type of news
- Size indicates number of articles related to the news (from all the sources of Google News)
- Color intensity indicates freshness of news

Fig. 5. Newsmap using a treemap algorithm. Source: [newsmap.jp] taken on 16/04/2014

Definition

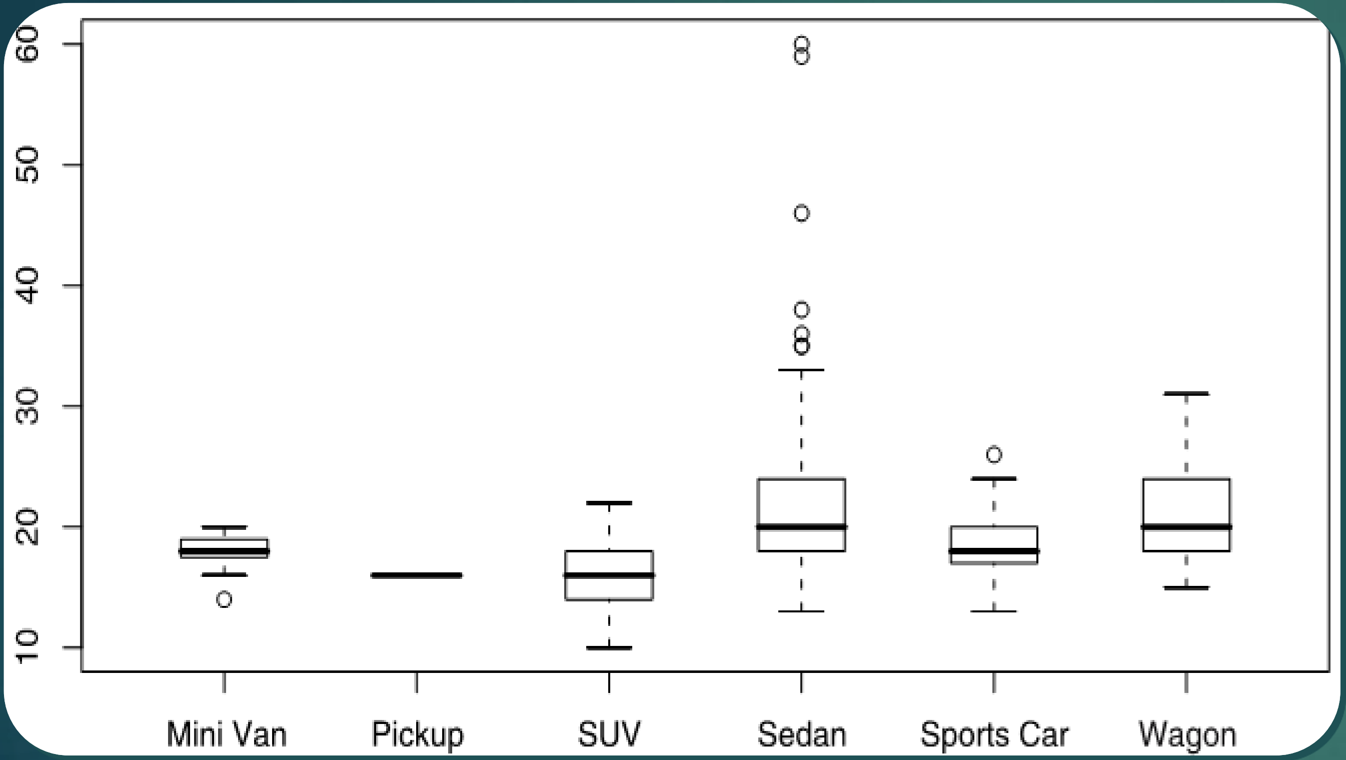


Fig. 6. Simplest form of a trellis display: a box plot y by x .
Souce: [Handbook of Data Visualization]

- Introduced by Becker et al. (1996) to visualize multivariate data.
- Use a lattice-like arrangement to place plots onto panels, called *panel plots,* which share the same scale.
- A single display can hold up to 7 variables. 5 of them must be categorical, other 2 called *axis variables* can be continuous.
- Up to 3 categorical variables can be used as *conditioning variables* to form row, columns and pages. The other 2, called *adjunct variables*, can be coded base on type of panel plot.

II. Trellis Displays for High-Dimensional Data Visualization
Definition

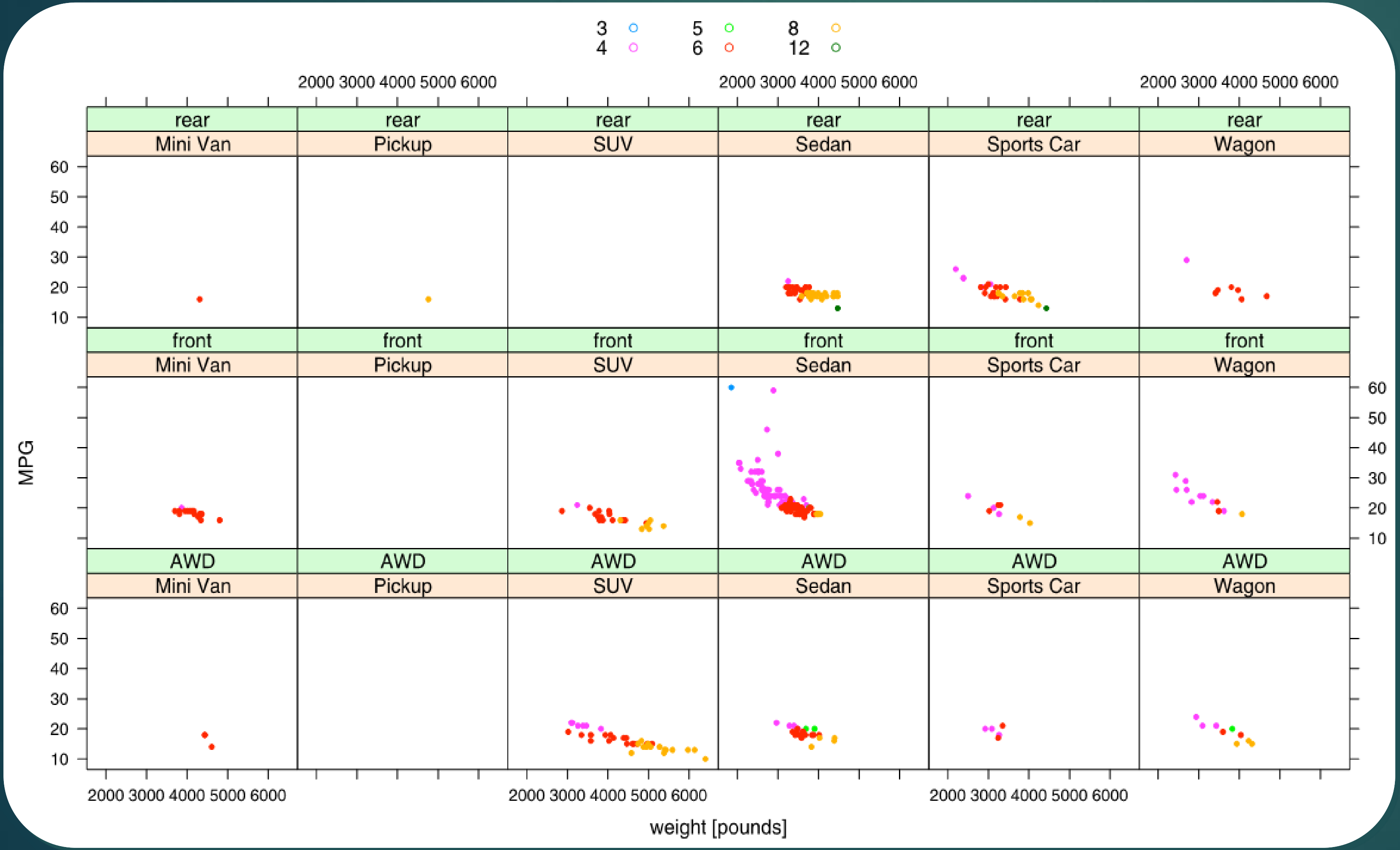


Fig. 7. Trellis display incorporating 5 variables of a car data-set.
Source: [Handbook of Data Visualization]

- Trellis displays introduce the concept of shingles.
- Shingling: dividing a continuous variable into (overlapping) intervals to convert it into a discrete variable.
- Overlapping shingles/intervals leads to multiple representations of data within a trellis display.
- Show the interval of a shingle using an interval of the strip label.
- Motivation for shingling: Brushing with linked highlighting

Trellis Displays and Interactivity

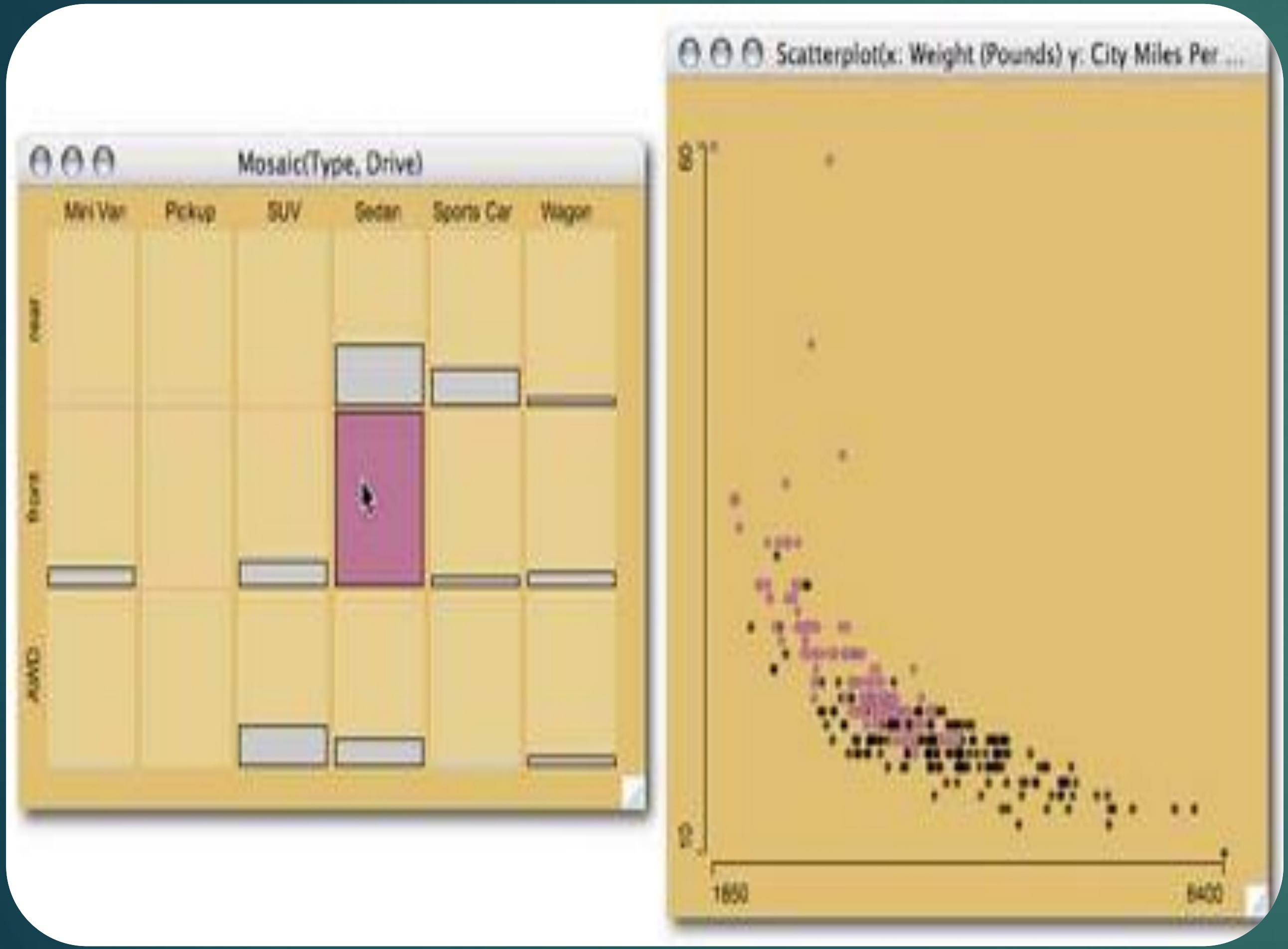
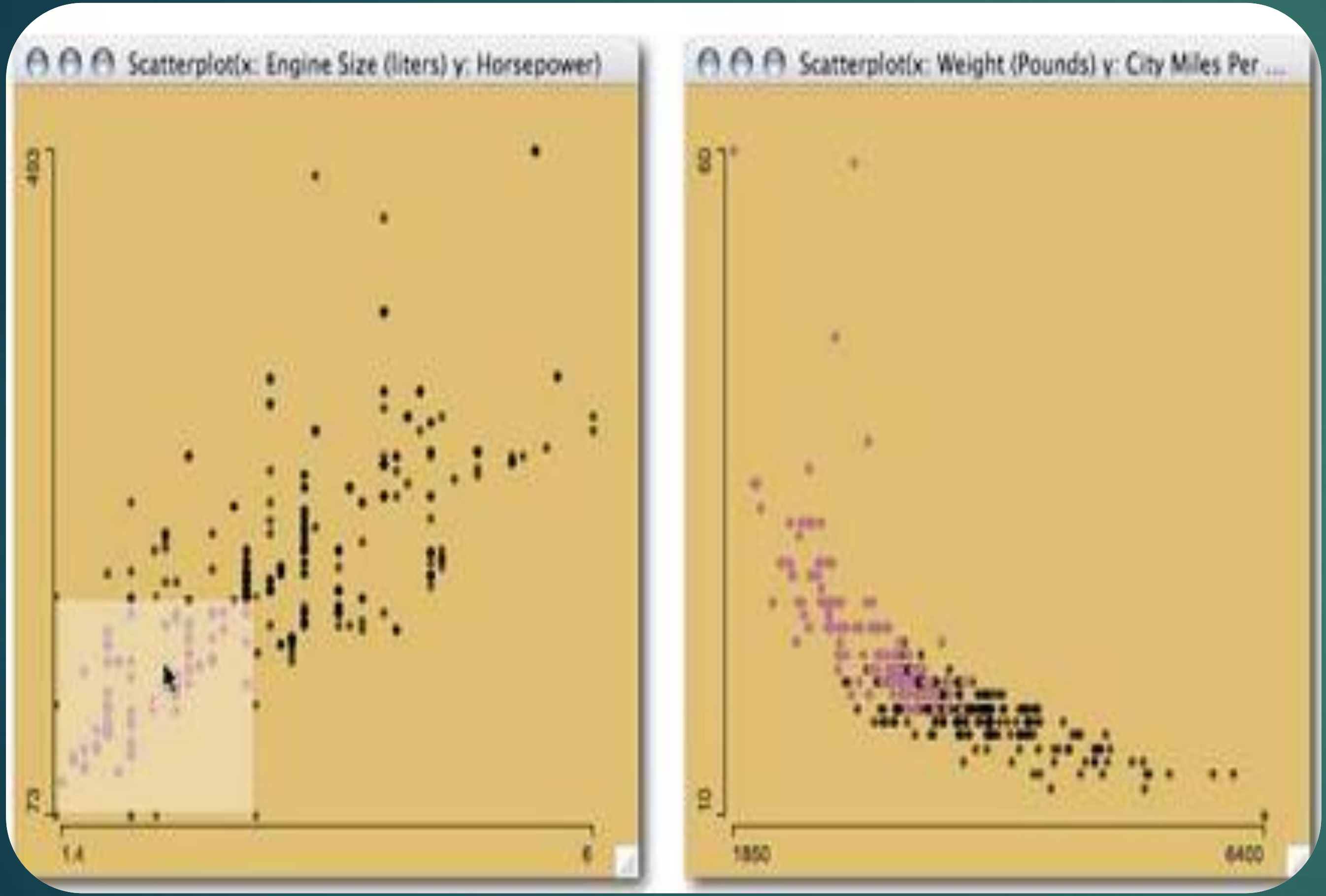


Fig. 8. Selecting the group of FWD sedans in the mosaic plot in multiple-barchart view (left), one gets the corresponding panel plot (scatterplot on right) in the highlighted subgroup
Source: [Handbook of Data Visualization]

- Single view in a panel := highlighted part of the graphics of panel plot for conditioned subgroup
- Possible interaction: brushing
- Brushing: steadily move an indicator of selection region along one/two axes of a plot
- Selected interval from brushing := interval of a shingle variable
- Subdivided intervals of continuous variable := snapshots of continuous brushing process from min - max

Trellis Displays and Interactivity



- Motivation for shingle variables: brushing with linked highlighting
- Much more flexible than static view in a trellis display.
- In contrast, trellis display can be easily reproduced in printed form.

Fig. 9. Brushing in the conditioning scatterplot (left), one gets the panel plot (scatterplot on right) in the highlighted subgroup
Source: [Handbook of Data Visualization]

Visualization of Models

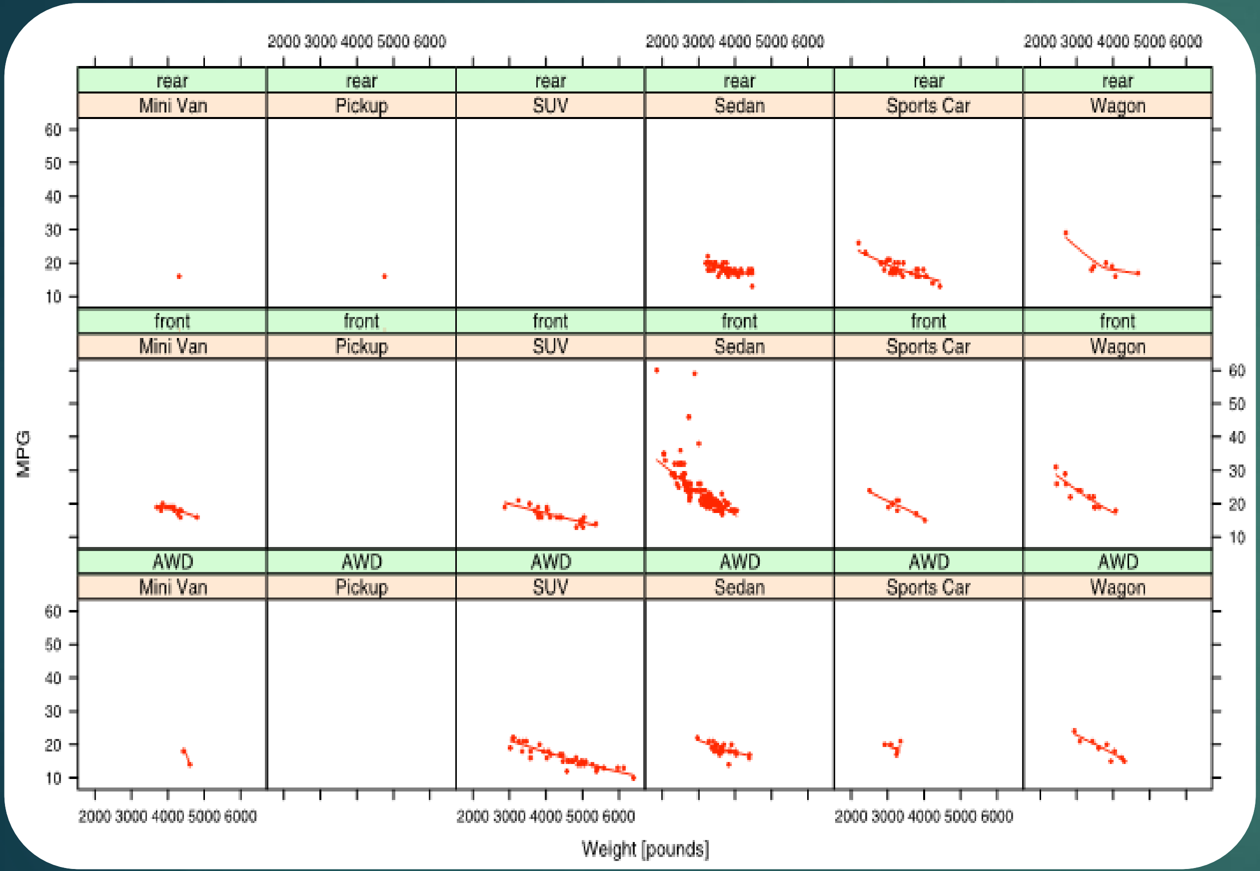


Fig. 10. Trellis display in Fig. 7 with lowess smoother overlaid
Source: [Handbook of Data Visualization]

- Main advantage of trellis displays: common scale among all plot panels → effective comparison of panel plots between rows/columns/pages
- Best used for model diagnostics
- Major advantage over other multivariate techniques: flat learning curve and ability of static reproduction
- Drawback: hard to judge the number of cases in a panel plot

III. Linked Views for Visual Exploration

Introduction

- Problem: limitation of 2-D space (paper, computer screen)
- Approaches: VR/pseudo 3-D environment, data projection, non-orthogonal coordinate system, linked displays
- Concept of linked graphic first implemented in software in [McDonald82] to connect 2 scatterplots.
- Scatterplots linking (scatterplot brushing) by now the most prominent case of linked views [Becker87]
- Pros:
  - easiness of underlying display, speed, flexibility → essential for exploratory data analysis
  - applicable for complex data structures
- Visual exploration requires flexible, adaptive framework with flexibility + stability

Linking Schemes & Structures

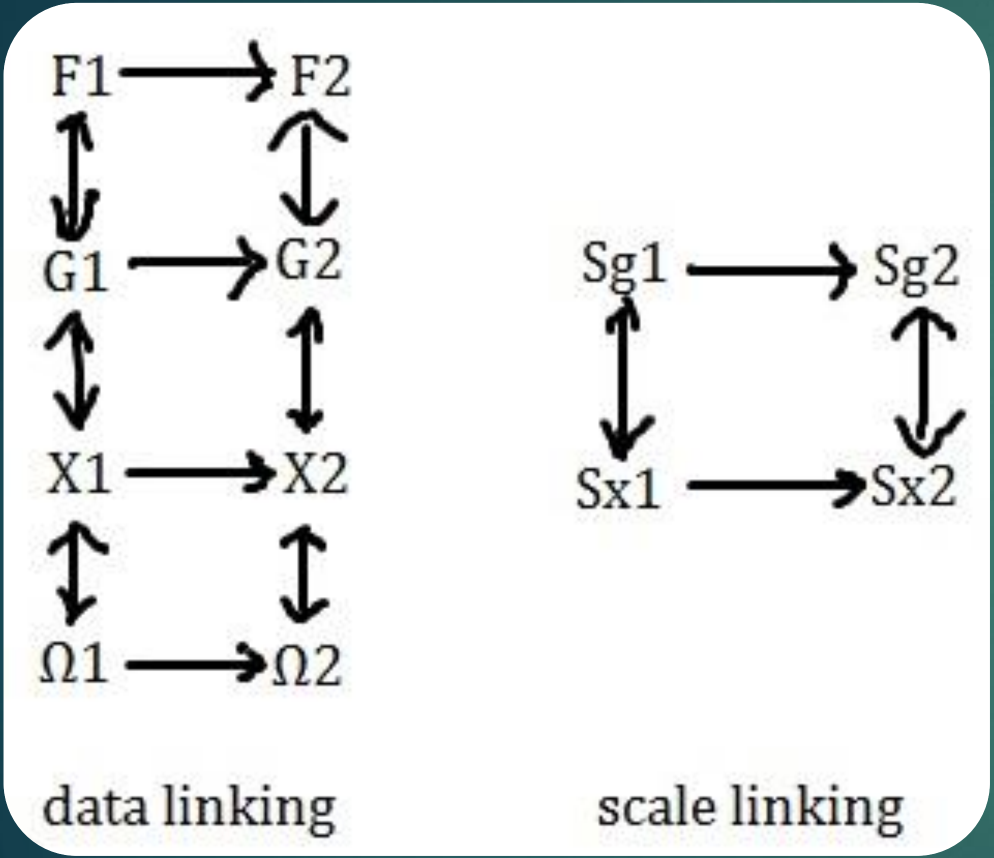


Fig. 11. Possible linking structures between active plot D1 & passive plot D2

- Linked views: 2 or more plots share & exchange information
- Question: which information is shared + how to realize it?
- [Wilhelm05] proposed a separation of data displays:

$$D = (F, (G, s_g), (\chi, s_\chi), \Omega)$$

D: data analysis display; F: frame; (G, s_g): set of graphical elements & set of scales; (χ, s_χ): model & its set of scale; Ω: sample population

- One "active" plot & the others "passive"
- Relevant: information sharing between identical layers → linking frames | types | models | sample populations
- Direct linking scheme & combined linking scheme

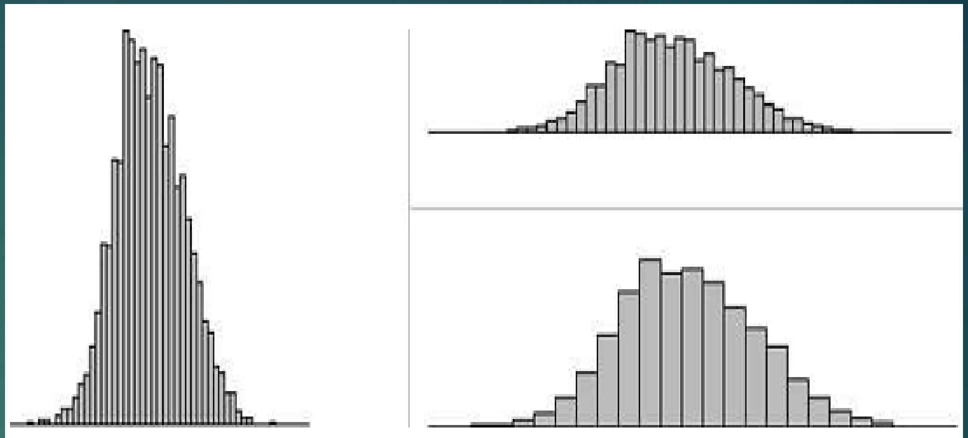
Linking Schemes & Structures

- **Linking Sample Populations**
- **Linking Models**
- **Linking Types**
- **Linking Frames**

Linking Schemes & Structures

- **Linking Sample Populations (SPL)**
  - Most widely used, most important
  - Defined as mapping $m\colon \Omega_1 \to \Omega_2$
  - Identity linking (empirical linking): most common case of SPL
    - Identity mapping: id: $\Omega \to \Omega$
    - Visualize connection between observations of the same individual/case
    - Use natural connection between features of the same set of cases
  - Hierarchical Linking
    - Data to be analyzed is heterogeneous, but typically has some hierarchy resulted from different aggregation levels (e.g. spatial data)
    - Visualize the hierarchical connection by establish relation: $m\colon \Omega_1 \to \Omega_2$ so that some filtration is generated
  - Neighborhood/Distance Linking
    - Used on geographic data to investigate local effects
    - Different neighborhood definitions lead to (somewhat) different linking relations. In general: $m\colon \Omega_1 \to \Omega_2, m(\omega^*) = \{\omega \in \Omega_2 \colon dist(\omega^*, \omega) \le d\}$

Linking Schemes & Structures



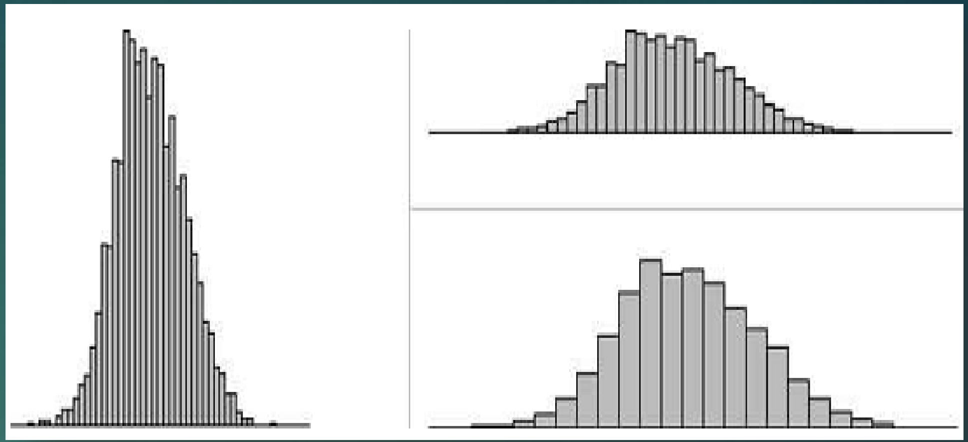
Source: Handbook of Data Visualization

- Linking Sample Populations
- **Linking Models**
  - Model: central part of data display definition, define amount of information to be visualized [Wilhelm05]
  - Example: categorization model (histogram of quantitative variable) $A \boxplus C = ([C_0, C_1], (C_1, C_2], \ldots, (C_{c-1}, C_c; count(AC) := \left(\sum_{i:C_0 \leq A_i \leq C_1} count(A_i), \ldots, \sum_{i:C_{c-1} \leq A_i \leq C_c} count(A_i)\right);\ s_\chi = (C, \pi_c, \max\left(count(AC)\right))$
  - Model link either via set of observations $A \boxplus C$ or scale s_χ (3 cases)
  - Order of categorization values less important for continuous data (histogram)
  - Scale component important for nominal categories
  - Linking scale parameters common for bar-charts & mosaic-plots
  - Categorization vector C belongs to both observation & scale component (histogram) → linking observations is restricted to variables used in the model
  - Young et al. (1993) created a system that combined various views of same data set in one window.

Linking Schemes & Structures

- Linking Sample Populations
- Linking Models
- **Linking Types**
  - Type layer: visible components, representing the model
  - Linked models → congruency at type level
  - Direct link between type levels (without model links) uncommon, except for color & size
  - Example: colors in pie charts to enhance differentiation between categories → link by assigning identical color to each slice
    - Type level link only if slice ordering does not reflect information of model level
    - Otherwise color linking could be a realization of a model link (if slices are ordered)
  - Linking type information necessary for comparing various plots

Linking Schemes & Structures



Source: Handbook of Data Visualization

- Linking Sample Populations
- Linking Models
- Linking Types
- **Linking Frames**
  - Frame: coarsest level of a data display
  - Determine size & shape of plot window
  - Frame linking: essential for screen-space-saving layout of displays & correct comparison of graphical displays
  - Linking other attributes (e.g. background color) not important, but ability to set and change those in a framework is valuable.

Visualization Techniques for Linked Views [Roberts et al. (2000)]

- **Replacement**
- **Overlaying**
- **Repetition**
- **Special Forms**

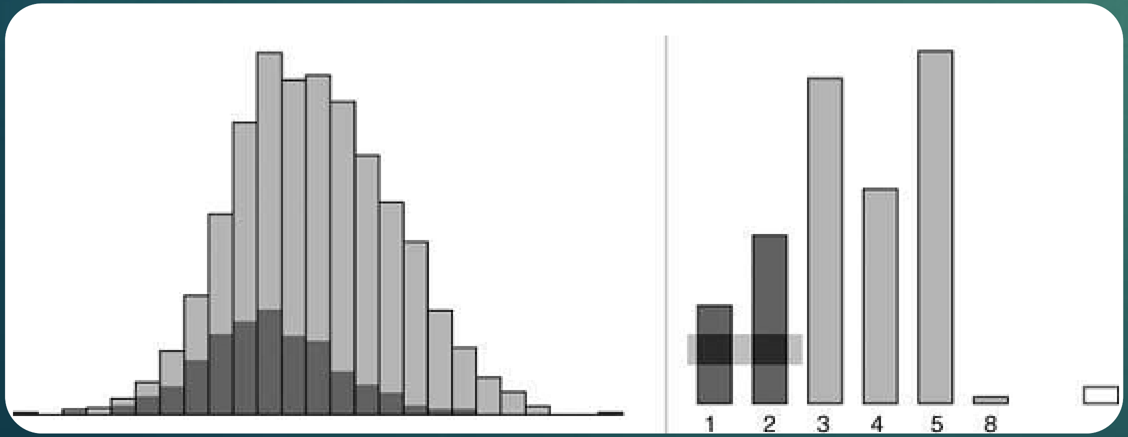
III. Linked Views for Visual Exploration

Visualization Techniques for Linked Views [Roberts et al. (2000)]

- **Replacement**
  - Old information is lost and replaced by new information
  - Useless for sub-setting/conditioning approach → marginal distribution information is lost
  - Unable to compare different plot versions
  - Exploratory data analysis: important to keep track of changing scenarios → history system in geo-visualization systems [Roberts04]

Visualization Techniques for Linked Views [Roberts et al. (2000)]

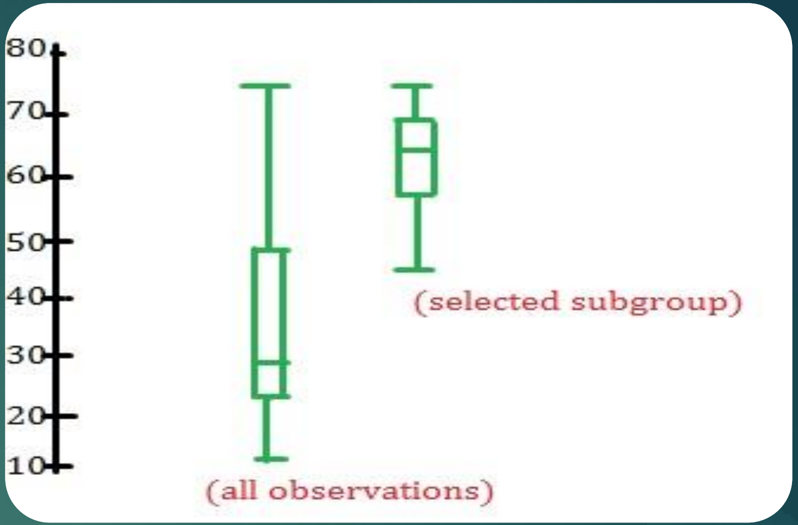
- **Replacement**
- **Overlaying**
  - Common strategy used to look at conditional distribution in area plots
  - Provide framework for comparison between conditional/marginal distributions
  - Problems
    - No freedom in parameter choice for selected subset (inherited from original plot)
    - Occlusion/Overplotting: original display is hidden by new overlaid plot → irrelevant for area-based displays/scatterplots but important for more complex plot, e.g. box-plots



Source: Introduction to Information Visualization

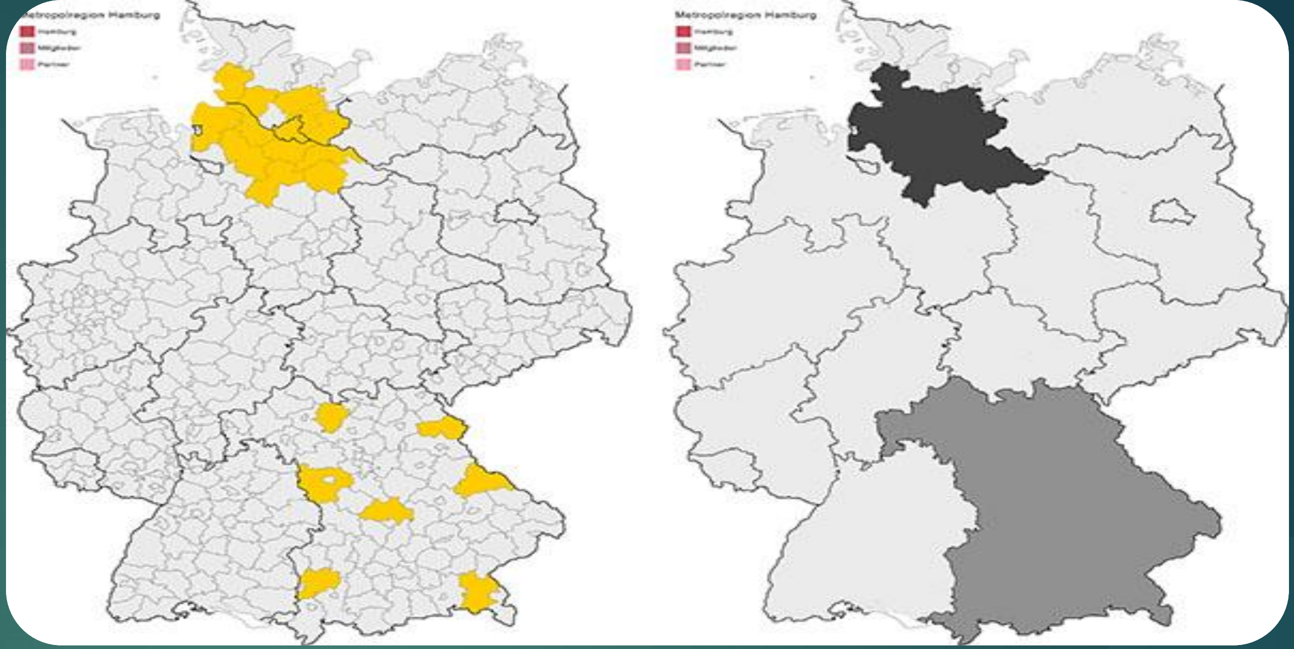
III. Linked Views for Visual Exploration

Visualization Techniques for Linked Views [Roberts et al. (2000)]



Two plots in juxtaposition

- **Replacement**
- **Overlaying**
- **Repetition**
  - Displays are cloned, showing different views of same data at the same time
  - Pros: comprehensive picture/complete overview of data, impact of parameter changes/user interactions are observable.
  - Cons: user confusion from multiple changing displays/views
  - Requirements:
    - Keep track of user changes/interactions
    - Easy, powerful method to rearrange displays on computer screen
  - Example: juxtaposition → common repetition strategy for sub-setting, well-known for static plots but not yet widely used in interactive graphical systems



Hierarchical linking scheme for 2 maps

- Replacement
- Overlaying
- Repetition
- **Special Forms**
  - Different problems with non 1-to-1 linking, e.g. m-to-1 linking (hierarchy)
    - Example: 2-level hierarchy (an aggregated macro level + a micro level)
  - Partial selection → partial highlighting by subdividing shape into components (for regular shapes)
  - General approach: different color intensities
  - Bi-directional links for plot parameters governing layout & size of display (e.g. by choosing joint axis parameters)

IV. Conclusion

- ✓ Exploratory data analysis often deals with large, multivariate datasets
- ✓ Multiple techniques/approaches exist for representing high-dimensional data, but not all are suitable for general audience
- ✓ Some techniques produce highly complex graphics, intended for scientific researchers/data analysts
- ✓ Linked views paradigm is well-known and a powerful tool to explore broad range of data, effective even for complex/high-dimensional datasets
  - ✓ Simplicity of individual plots + intuitive, easy to use & flexible UI → suitable for a broad audience with regard to data-driven journalism, which is defined as "journalistic process based on analyzing and filtering large data sets for the purpose of creating a news story" [Wikipedia]

IV. References

- Johnson, B. and Shneiderman, B. (1991). *Tree-Maps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures*
- Becker, R., Cleveland, W. and Shyu, M. (1996). *The Visual Design and Control of Trellis Displays*
- Chen, C.-H., Härdle, W. and Unwin, A. (eds) (2008). *Handbook of Data Visualization*
- Roberts, J.C. (2004). *Exploratory Visualization with Multiple Linked Views*, in J. Dykes, A.M. MacEachren and M.-J. Kraak (eds), *Exploring Geovisualization*
- Shneiderman, B. (1994). *Dynamic Queries for Visual Information Seeking*
- Tufte, E.R. (1983). *The Visual Display of Quantitative Information*
- Wilhelm, A.F.X (2005). *Interactive Statistical Graphics: The Paradigm of Linked Views,* in C. Rao, E. Wegman and J. Solka, *Handbook of Statistics*, Vol. 24
- Young, F.W., Faldowski, R.A. and McFarlane, M.M. (1993). *Multivariate Statistical Visualization,* in C. Rao, *Handbook of Statistics*, Vol. 9
- McDonald, J.A. (1982). *Interactive Graphics for Data Analysis*
- Becker, R.A., Cleveland, W.S. and Wilks, A.R. (1987). *Dynamic Graphics for Data Analysis*
- Hofmann, H. (2001). *Graphical Tools for the Exploration of Multivariate Categorical Data*