Multivariate Data Visualization

Sungahn Ko
sako@unist.ac.kr

Disclaimer

- **The slides cannot be distributed, posted or used outside of this class**

- **Slides in this course courtesy of**
  - Dr. Abish Malik (Purdue)
  - Dr. Yun Jang (Sejong Univ.)
  - Dr. Ross Maciejewski (ASU)
  - Dr. Niklas Elmqvist (UMD)
  - Dr. David Ebert (Purdue)

Univariate Data Visualization

- **Univariate data**
  - Histograms
  - Box and Whisker Plots
  - Line Graphs
- **What if we have multivariate data?**
  - Today we want to discuss techniques for

Representation

- **What are the two main ways of presenting multivariate data sets?**
  - Directly (textually) – Tables
  - Symbolically (pictures) – Graphs

- **How do we decide which to use, and when?**

1 – Descriptions on this slide are borrowed from John Stasko's "Multivariate Data & Tables and Graphs" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/data.pdf

Tables?

- **Use tables when**
  - Precise values are required
  - We want to look at and compare individual values
  - 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
  - We want to reveal relationships among values

1 – Descriptions on this slide are borrowed from John Stasko's "Multivariate Data & Tables and Graphs" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/data.pdf
2 - S. Few, Show Me the Numbers: Designing Tables and Graphs to Enlighten,

Graphs?

- **Graph**
  - Visual display that illustrates one or more relationships among entities
  - Allows a trend, pattern or comparison to be easily comprehended
- **Critical to remain task-centric**
  - Why do you need a graph?
  - What questions are being answered?
  - What data is needed to answer those questions?
  - Who is the audience?

1 – Descriptions on this slide are borrowed from John Stasko's "Multivariate Data & Tables and Graphs" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/data.pdf

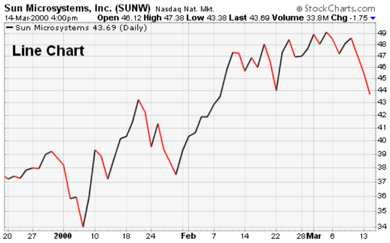
Name	Team	At Bats	Runs	RBI	Batting Ave
C. Gonzalez	COL	587	111	117	0.336
J. Votto	CIN	547	106	113	0.324
O. Infante	ATL	471	65	47	0.321
T. Tulowitzki	COL	470	89	95	0.315
M. Holiday	STl	596	95	103	0.312
A. Pujols	STL	587	115	118	0.312
M. Prado	ATL	599	100	66	0.307
R. Zimmerman	WSH	525	85	85	0.307
R. Braun	MIL	619	101	103	0.304
S. Castro	CHC	463	53	41	0.3
H. Ramirez	FLA	543	92	76	0.3
P. Polanco	PHI	554	76	52	0.298
A. Gonzalez	SD	591	87	101	0.298
J. Werth	PHI	554	106	85	0.296
M. Byrd	CHC	580	84	66	0.293
A. Ethier	LAD	517	71	82	0.292
A. Pagan	NYM	579	80	69	0.29
A. Huff	SF	569	100	86	0.29
J. Keppinger	HOU	514	62	59	0.288
D. Uggla	FLA	589	100	105	0.287

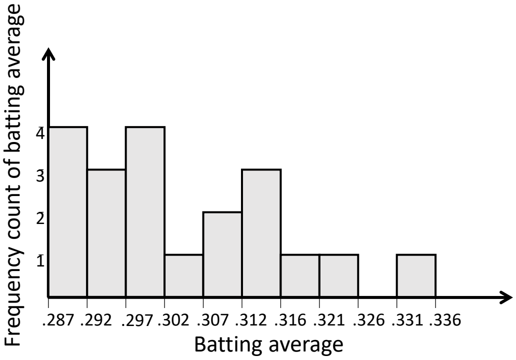
National League Batting Average Leaders 2010

6

Univariate Case

- **In univariate representations, we think of data case as being shown along one dimension and value in another**

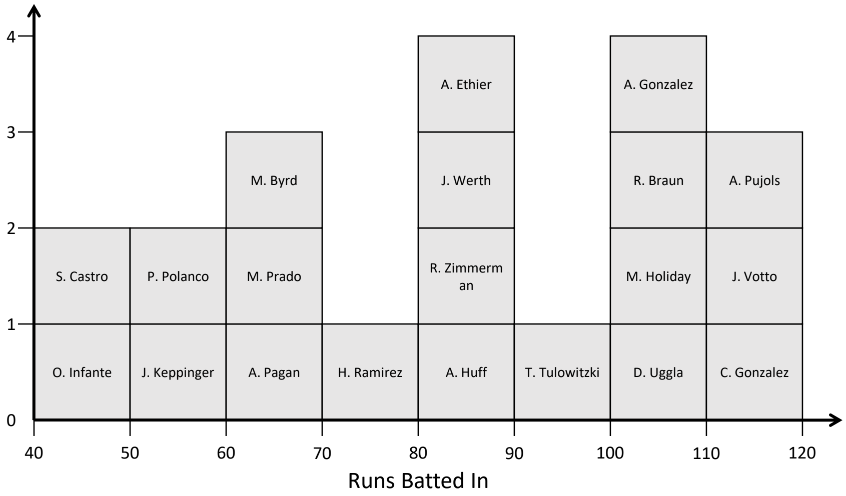




7

7

Bivariate Case – Stacked Bar Graph



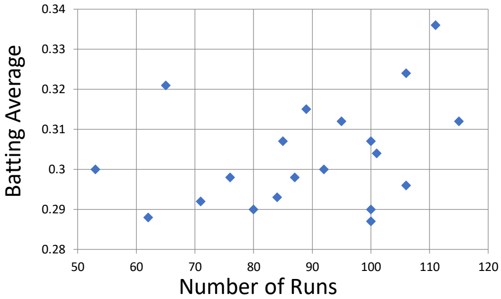
T. N. Dang, L. Wilkinson and A. Anand, "Stacking Graphic Elements to Avoid Over-Plotting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1044-1052, 2010.

Bivariate Case - Scatterplot

- Visualizes discrete data values along two axes
- Used as a means of analyzing bivariate relationships

Bivariate Case - Scatterplot

- **Quick means of assessing outliers, clusters and distributions**
- **Putting a line through the data can help assess trends, but can also mislead viewer**



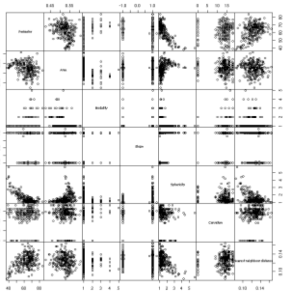
Scagnostics

- **Scatterplot diagnostics**
  - Graph-theoretic measures for detecting a variety of structural anomalies in a geometric graph representation of scatterplot data
  - Ratings can be used to pick views that show particular structures that are of interest to the user
  - Coined by Tukey, but never published, it is an exploratory graphical technique to help determine notable relationships between two variables

1 - J. Tukey and P. Tukey. Computing graphics and exploratory data analysis: An introduction. In Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics 85. In Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics, pages 773–785, 1985.
2 - L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics.In Proceedings Information Visualization, pages 157–164. IEEE CS Press, 2005.

Scagnostics

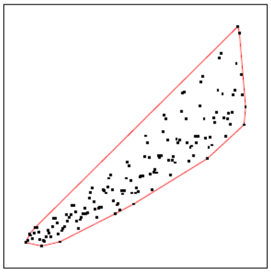
- **Measures have proven statistical properties that are computable for modern datasets (available as a free downloadable package in R)**



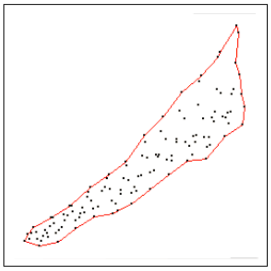
[1] - J. Tukey and P. Tukey. Computing graphics and exploratory data analysis: An introduction. In Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics 85. In Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics, pages 773–785, 1985.
[2] - L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics.In Proceedings Information Visualization, pages 157–164. IEEE CS Press, 2005.

Scagnostics - Methods

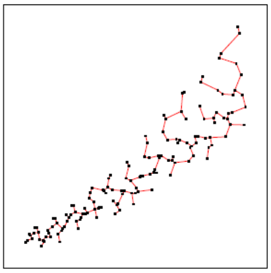
- **Graph Theoretic Scagnostics are based on Geometric Graphs**
  - Convex Hull (the outer edges of a Delaunay triangulation)



Convex Hull



Alpha Hull

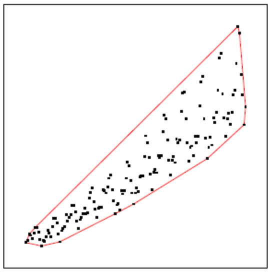


Minimum Spanning Tree

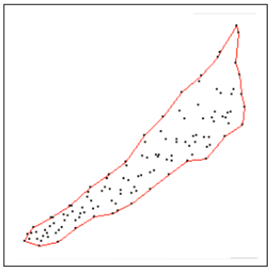
[1] - L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics.In Proceedings Information Visualization, pages 157–164. IEEE CS Press, 2005.

Scagnostics - Methods

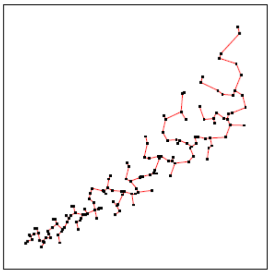
- **Graph Theoretic Scagnostics are based on Geometric Graphs**
  - Alpha Hull (a generalization of the convex hull and a subgraph of the Delaunay triangulation)



Convex Hull



Alpha Hull

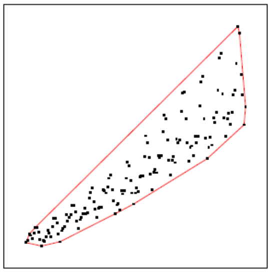


Minimum Spanning Tree

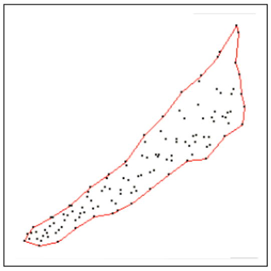
1 - L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics.In Proceedings Information Visualization, pages 157–164. IEEE CS Press, 2005.

Scagnostics - Methods

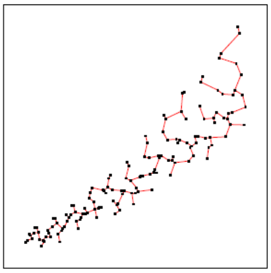
- **Graph Theoretic Scagnostics are based on Geometric Graphs**
  - Minimum Spanning Tree



Convex Hull



Alpha Hull

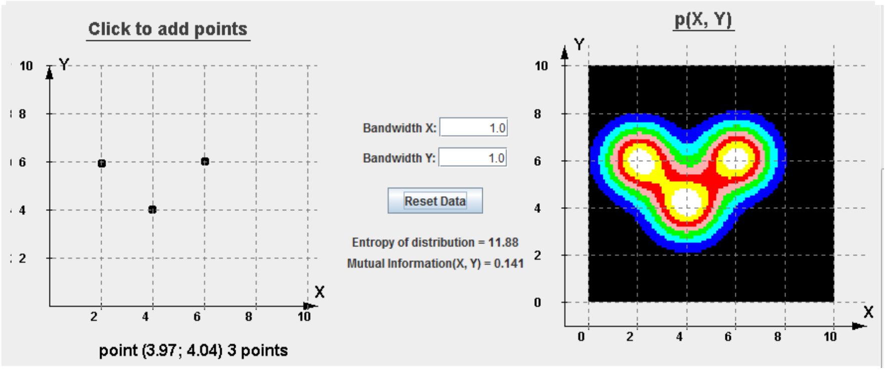


Minimum Spanning Tree

1 - L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics.In Proceedings Information Visualization, pages 157–164. IEEE CS Press, 2005.

Multivariate Density Estimation

- **Kernel Density Estimation**
- **Continuous Scatterplots**



Bandwidth X: 1.0

Bandwidth Y: 1.0

Entropy of distribution = 11.88

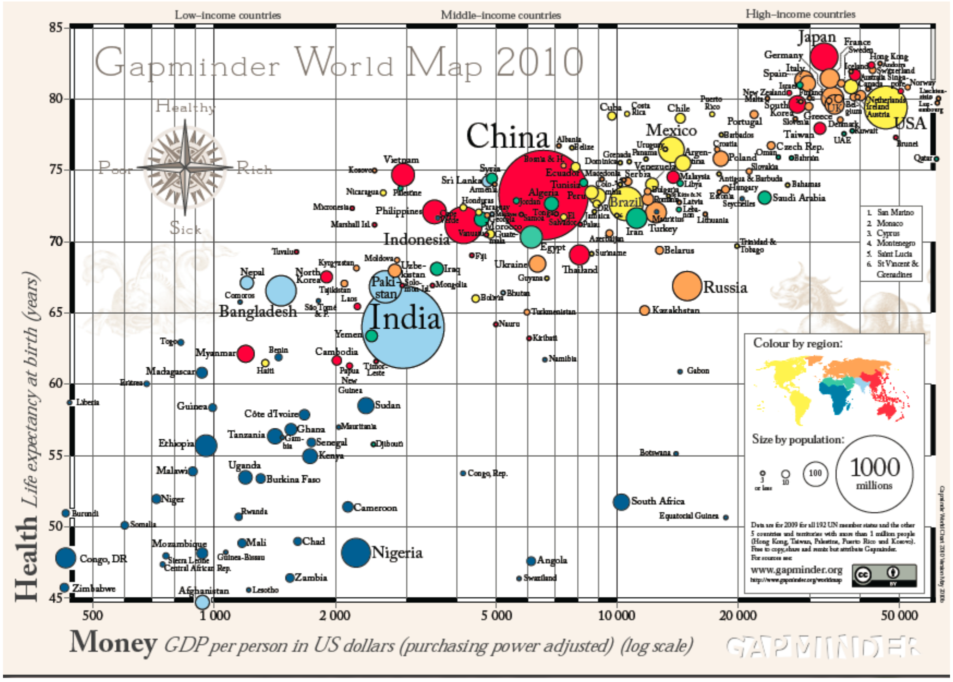
Mutual Information(X, Y) = 0.141

point (3.97; 4.04) 3 points

1 – B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

- - http://parallel.vub.ac.be/research/causalModels/tutorial/kde.html

2 - S. Bachthaler and D. Weiskopf, “Continuous Scatterplots," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1428-1 435, 2008.

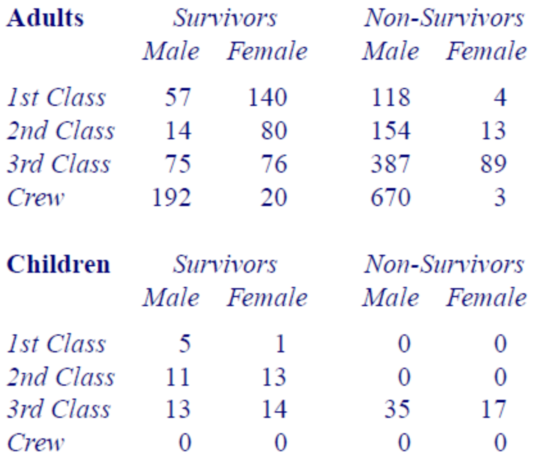


http://www.gapminder.org/GapminderMedia/wp-uploads/pdf_charts/GWM2010.pdf

Multivariate Case - Mosaic Plot

- **Graphical display that allows you to examine the relationship among two or more categorical variables**
- **Start as a square with length one**
  - Divide first into horizontal bars whose widths are proportional to the probabilities associated with the first categorical variable
  - Next each bar is split vertically by the conditional probability of the second categorical variable

The Titanic Example

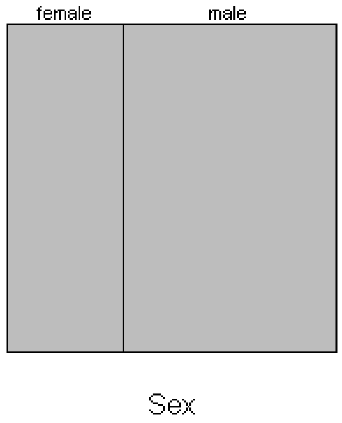


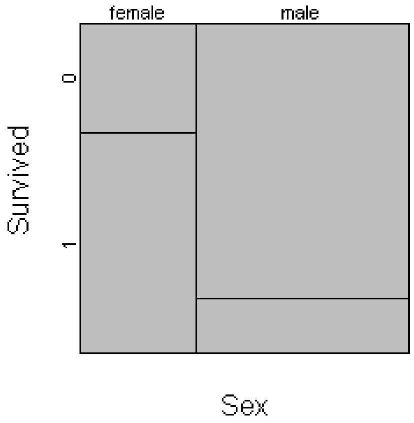
Adults	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	57	140	118	4
2nd Class	14	80	154	13
3rd Class	75	76	387	89
Crew	192	20	670	3

Children	*Survivors*		*Non-Survivors*	
	Male	*Female*	*Male*	*Female*
1st Class	5	1	0	0
2nd Class	11	13	0	0
3rd Class	13	14	35	17
Crew	0	0	0	0

The Titanic Example

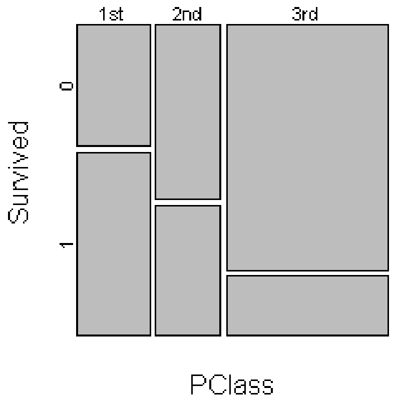
- **Mortality rates between men and women on the Titanic**

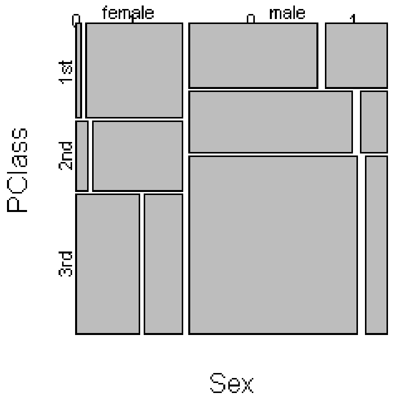




http://www.childrensmercy.org/stats/definitions/mosaic.htm.

The Titanic Example





Perceptual Basis for Mosaic Plot

- It is tempting to dismiss mosaic plots because they represent counts as rectangular areas and so provide a distorted perceptual encoding
- In fact, the important encoding is the length
- At each stage, the comparison of interest is of the length of the sides

Small Multiples

- Give each variable its own display (sometimes called Trellis Chart, Grid Chart)
- Use the same graphic to display different slices of a dataset
- We want to ask questions about our data, how does X compare to Y?

Tufte, Edward (1983). *Visual Display of Quantitative Information*.

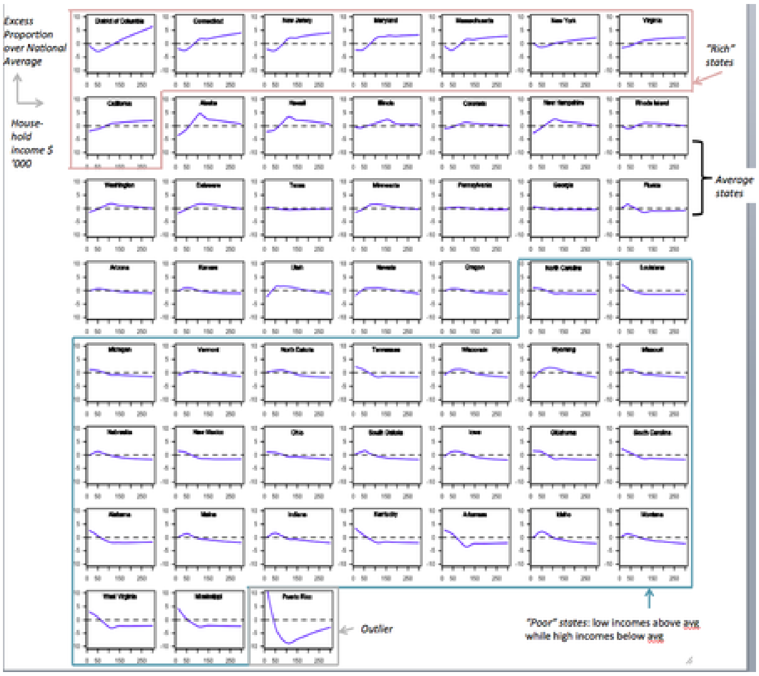
Small Multiples

- **Placement of the small multiples should reflect some logical order in order to guide user**
- **Should share the same measures, scales, size and shape**
- **Simplicity is key as users need to ingest large number of charts at once**

Tufte, Edward (1983). *Visual Display of Quantitative Information*.

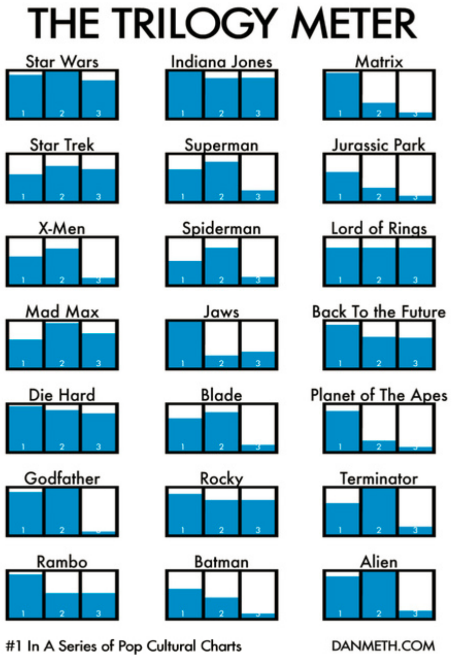
24

24



http://junkcharts.typepad.com/junk_charts/small_multiples/

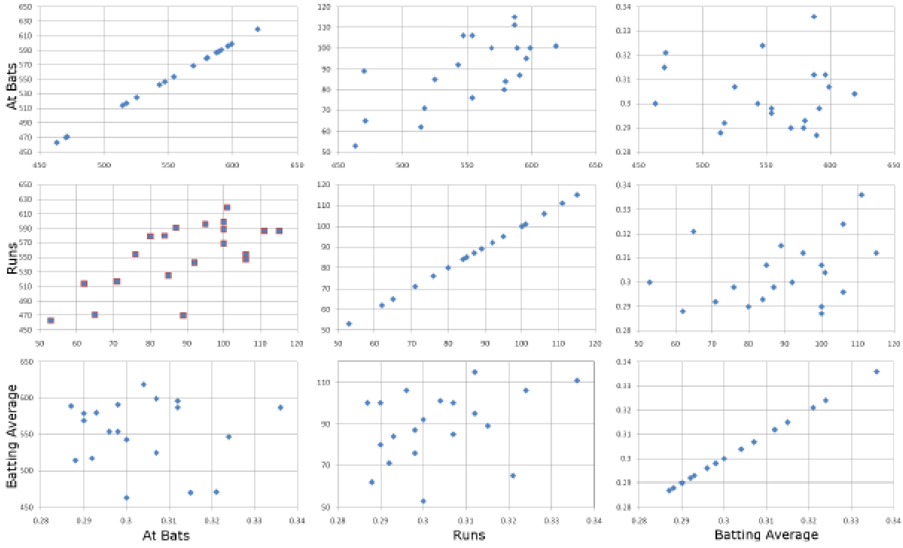
25



http://www.juiceanalytics.com/writing/better-know-visualization-small-multiples/

26

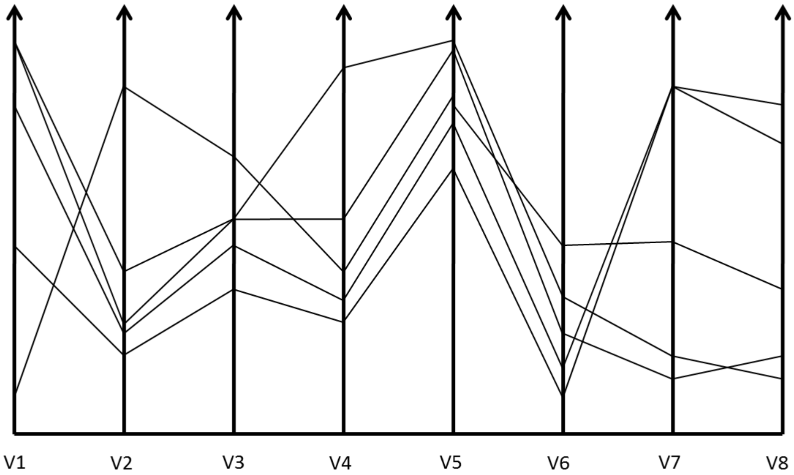
Multivariate Case - Scatterplot Matrix



27

27

Parallel Coordinate Plots



1 - Inselberg A (1985) The Plane with Parallel Coordinates. The Visual Computer1(4):69-91
2 - Ankerst M, Berchtold B, Keim DA (1998) Similarity clustering of dimensions for an enhanced visualization of multidimensional data. IEEE Symposium on Information Visualization pp 52-62

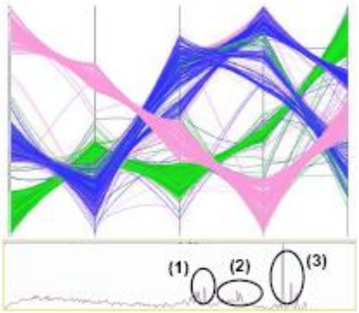
Issues With Parallel Coordinate Plots

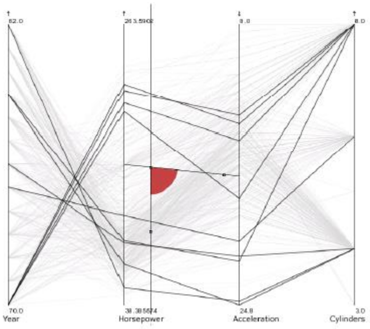
- **Different variables can take different values with very different ranges**
  - Need to normalize data ranges (maybe do a power transformation and then a normalization?)
- **Order of the parallel coordinate plots has a major impact on the resultant visualization**
- **The more variables we plot, the more lines we get and the more clutter that we get**

1 -Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. of IEEE Symposium on Information Visualization* (2003), pp. 105–112.
2 – Zhou, H., Yuan, X., Qu, Huamin, Cui, W., Chen, B., "Visual Clustering in Parallel Coordinates," Computer Graphics Forum 27(3) 1047-1054, 2008.

Attribute Ratios

- **Angular Brushing**
  - Angle between axes indicates level of correlation

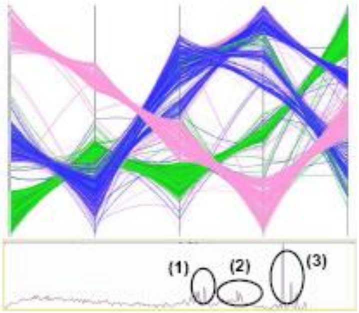


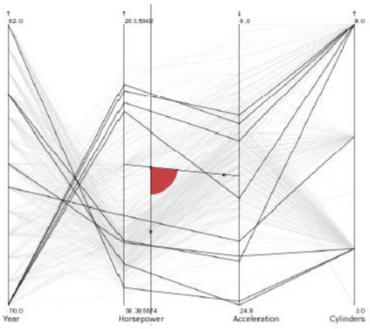


1 –Hauser, H., Ledermann, F., Doleisch, H., “Angular Brushing of Extended Parallel Coordinates,” *Proceedings of the IEEE Symposium on I nformation Visualization* (2002), pp. 127–130

Attribute Ratios

- **Angular Brushing**
  - Select subsets which exhibit a correlation along two axes by specifying angle of interest

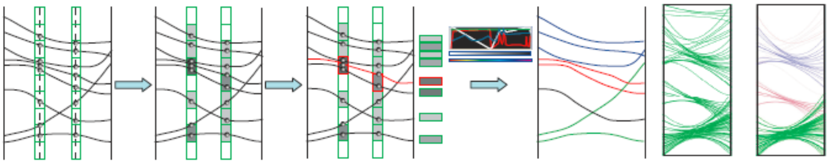




1 –Hauser, H., Ledermann, F., Doleisch, H., “Angular Brushing of Extended Parallel Coordinates,” *Proceedings of the IEEE Symposium on I nformation Visualization* (2002), pp. 127–130

Visual Clustering in Parallel Coordinate Plots

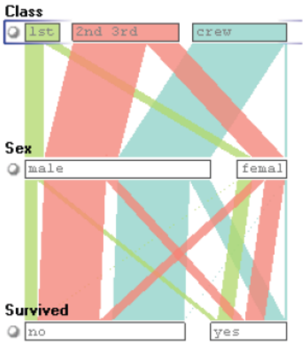
- Apply color and opacity based on line density
- Compute local density for each line by averaging the density values of all control points
- Apply color and opacity based on user specification



Visual Clustering in Parallel Coordinates, Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui, Baoquan Chen. Computer Graphics Forum (Proceedings of EuroVis'08), vol. 27, no. 3, 2008.

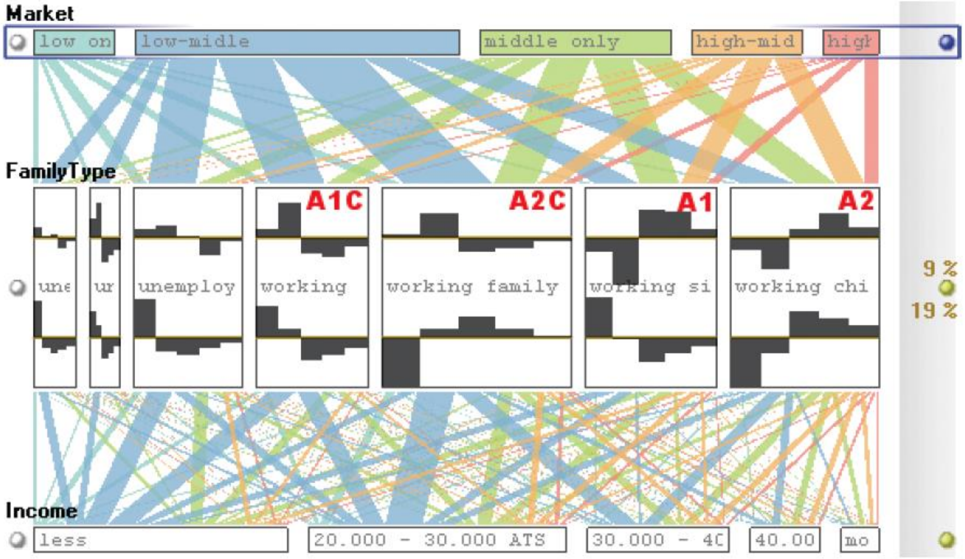
Parallel Sets

- Visualization method adopting parallel coordinate layout but uses frequency based representation
  - Layout similar to parallel coordinate plots
  - Continuous axes replaced with boxes
  - Used for categorical data



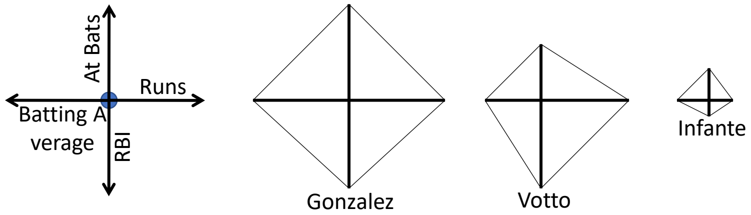
Robert Kosara, Fabian Bendix, and Helwig Hauser. 2006. Parallel Sets: Interactive Exploration and Visual Analysis of Categorical Data. *IEEE Transactions on Visualization and Computer Graphics* 12, 4 (July 2006), 558-568

Parallel Sets



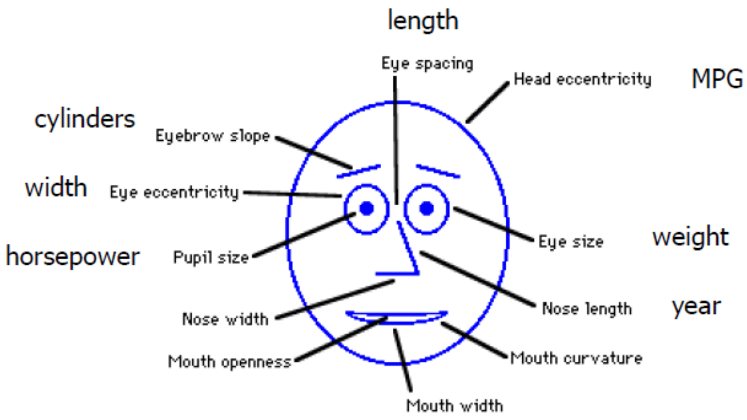
Star Plot

- Lay out axes in a radial layout, length of a ray emanates from a central point
- Rays are then joined together by a polyline drawn around the outside



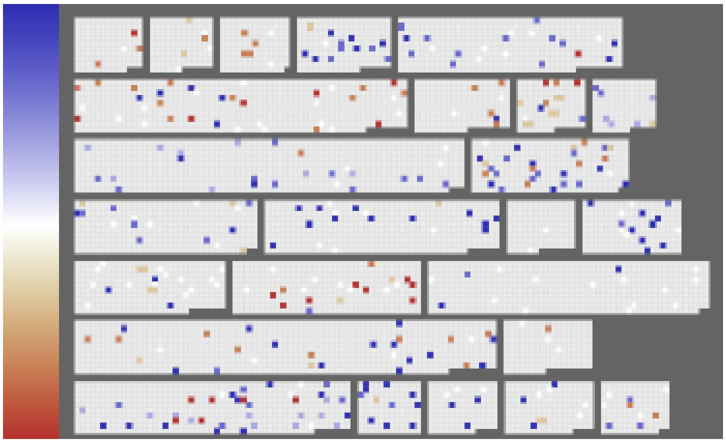
S. E. Fienberg,"Graphical Methods in Statistics," *The American Statistician*, vol. 33 pp. 156-178, 1979.

Multivariate Case – Chernoff Faces



1 - Herman Chernoff, The Use of Faces to Represent Points in K-Dimensional Space Graphically, *Journal of the American Statistical Association*, vol. 68, no. 342, pp. 361–368, 1973
2 - Christopher J. Morris, David S. Ebert, Penny Rheingans, *An Experimental Analysis of the Pre-Attentiveness of Features in Chernoff Faces*, Proceedings Applied Imagery Pattern Recognition, pp. 12–17, 1999.

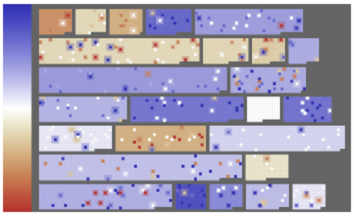
Pixel-Based Displays



1- Keim, DA.: Designing Pixel-oriented Visualization Techniques: Theory and Applications. IEEE Transactions on Visualization and Computer Graphics (TVCG) 6, 1 (2000), 59–78.
2 - Daniela Oelke, Halldor Janetzko, Svenja Simon, Klaus Neuhaus, Daniel A. Keim: Visual Boosting in Pixel-based Visualizations. Computer Graph. Forum 30(3): 871-880 (2011)

Visual Boosting of Pixel-Based Displays

- **Could modify the pixel based display to incorporate components that will draw attention to the salient aspects of the data**
  - Halo
  - Color
  - Distortion
  - Hatching



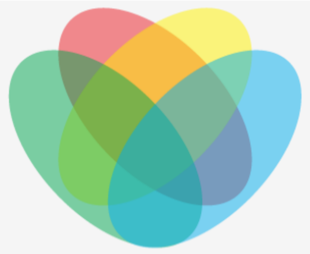
1- Keim, DA.: Designing Pixel-oriented Visualization Techniques: Theory and Applications. IEEE Transactions on Visualization and Computer Graphics (TVCG) 6, 1 (2000), 59–78.
2 - Daniela Oelke, Halldor Janetzko, Svenja Simon, Klaus Neuhaus, Daniel A. Keim: Visual Boosting in Pixel-based Visualizations. Computer Graph. Forum 30(3): 871-880 (2011)

Set Operations

- **Different type of problem**
  - Large set of items, each can be in one or more set
  - How do we visually represent the set membership?
    - Venn Diagrams
    - Euler Diagrams

Venn Diagram

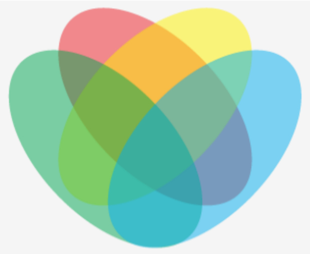
- Often used to illustrate a set
- By drawing partly overlapping shapes, different element combinations are assigned to areas



http://blog.visual.ly/euler-and-venn-diagrams/

Venn Diagram

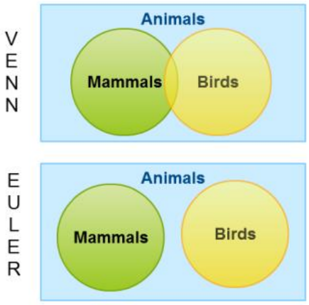
- As the number of set elements increases, this becomes impractical
- Hard to make all possible element combinations visible



http://blog.visual.ly/euler-and-venn-diagrams/

Euler Diagrams

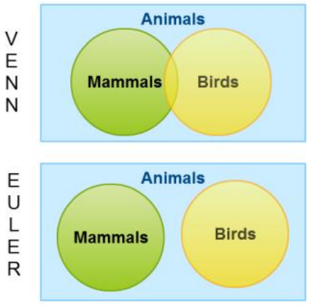
- All Venn diagrams are Euler diagrams, but not all Euler diagrams are Venn diagrams
- Venn diagrams have every hypothetically possible logical relationship between categories



http://blog.visual.ly/euler-and-venn-diagrams/

Euler Diagrams

- Euler Diagrams have only those combinations that exist in the real world

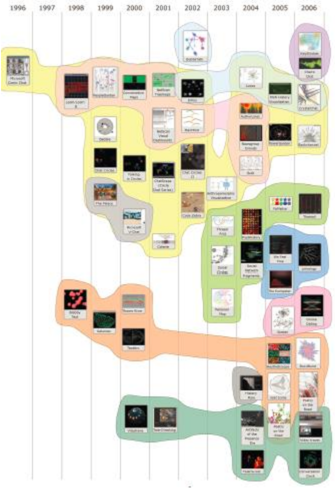


http://blog.visual.ly/euler-and-venn-diagrams/

Bubble Sets

Algorithm 1 Determining a Bubble Set Boundary

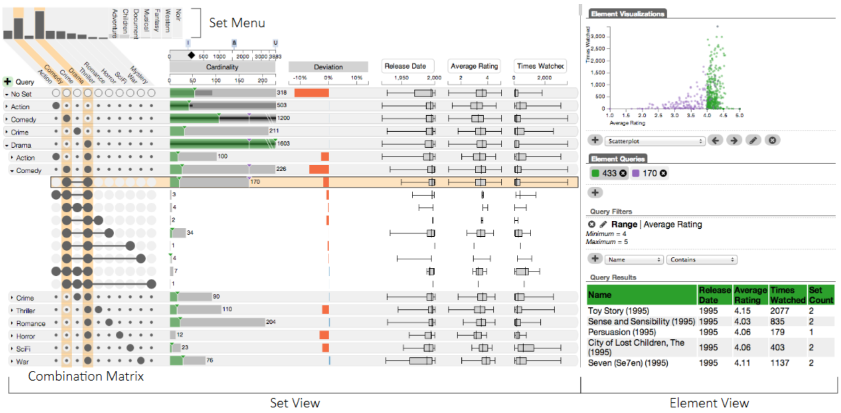
```
given items with positions
for all sets $s \in S$ do
    find centroid $c$ of $s$
    for all items $i \in s$, order ascending by distance to $c$ do
        find optimal neighbour $j \in s$
        find best route from $i$ to $j$
        for all cells (pixel or pixel group) within $R_1$ of $i$ do
            add potential due to $i$
            add potential due to nearest virtual edge $i \rightarrow j$
            subtract potential due to nearby non-set members $k \notin s$
        end for
    end for
    repeat
        perform marching squares to discover isopotential contour $\bar{s}$
        reduce threshold
    until $\forall i \in s$, isocontour $\bar{s}$ contains($i$)
    draw cardinal splines using every $N$th point on the contour
end for
```



Collins, Christopher; Penn, Gerald; Carpendale, Sheelagh. Bubble Sets: Revealing Set Relations over Existing Visualizations. IEEE Transactions on Visualization and Computer Graphics (Proceedings of the IEEE Conference on Information Visualization (InfoVis '09)), 15(6): November-December, 2009.
https://www.youtube.com/watch?v=P6CgBmIiXaE

44

Upset: Visualization of Intersecting Sets



Query Filters

Range | Average Rating
Minimum = 4
Maximum = 5

Query Results

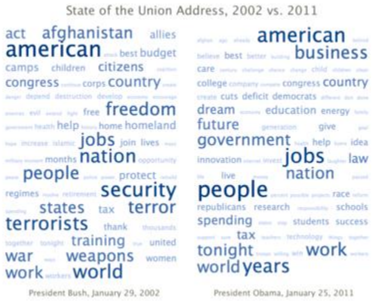
Name	Release Date	Average Rating	Times Watched	Set Count
Toy Story (1995)	1995	4.15	2077	2
Sense and Sensibility (1995)	1995	4.03	835	2
Persuasion (1995)	1995	4.06	179	1
City of Lost Children, The (1995)	1995	4.06	403	2
Seven (Se7en) (1995)	1995	4.11	1137	2

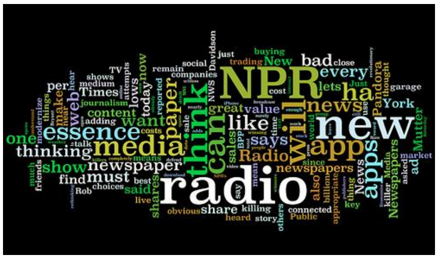
Let et al. UpSet: Visualization of Intersecting Sets, TVCG (InfoVis), 2014
https://www.youtube.com/watch?v=-IfF2wGw7Qk

45

Tag Clouds and Wordles

- Visual representation for text data where words are placed and scaled based on som e statistical measures
- Font size is typically determined by the number of instances a word is used





Text cloud comparing 2002 State of the Union Address by U.S. President Bush and 2011 State of the Union Address by President Obama. Created at TagCrowd.

Viegas, FB, Wattenberg, M, Feinberg, J, "Participatory Visualization with Wordle, IEEE Transactions on Visualization and Computer Graphics 15(6)1137-1144

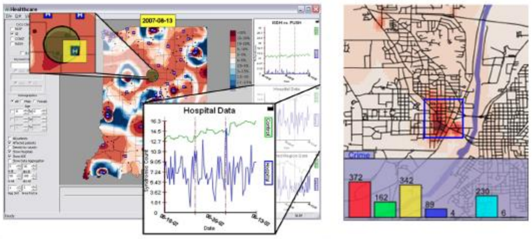
Coordinated Multiple Views

- Instead of trying to make the "best" visualization of all of our data, perhaps we can use multiple views
- Data can be expressed in a variety of ways
- Given the multivariate nature of data, a single statistical graphic may not be enough
- Interactive graphics systems can provide multiple representations of the data
- These representations can be *linked* or *coordinated*

North C, Shneiderman B (2000) Snap-together visualization Evaluating coordination usage and construction. International Journal of Human -Computer Studies 51:715-739

Brushing

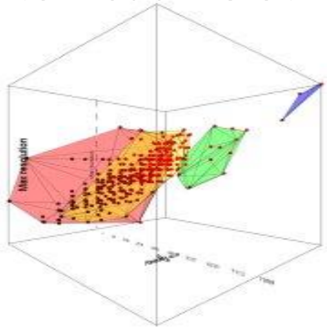
- Selecting data in one view highlights the same data points in other views
- Connecting multiple views through linked brushing provides more information than considering the component visualizations independently



Brushing Scatterplots. R. A. Becker and W. S. *Cleveland* (1987). Technometrics, 29:127-142

Interaction in Statistical Graphics

- Adding interaction can allow us to visualize other combinations of variables



1 - N. Elmqvist, P. Dragicevic, and J.-D. Fekete, “Rolling the Dice: MultidimensionalVisual Exploration Using Scatterplot Matrix Navigation,” *I EEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1141-1148, 2008..

2 - Jeffrey Heer and George Robertson. 2007. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Co mputer Graphics* 13, 6 (November 2007), 1240-1247.

Summary

- We've looked at methods that look at a fair number of methods for visualizing high-dimensional data
- Unfortunately, as the number of dimensions increase, we get more clutter and methods may fall apart
- Data that is similar in most dimensions ought to be drawn together
- Need to project the data down into the plane and give it some simplified representation
- Only look at certain aspects of the data at one time

Readings

- **Required Reading:**
  - L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics.In Proceedings Information Visualization, pages 157–164. IEEE CS Press, 2005.
  - Ankerst M, Berchtold B, Keim DA (1998) Similarity clustering of dimensions for an enhanced visualization of multidimensional data. IEEE Symposium on Information Visualization pp 52-62

Readings

- **Required Reading:**
  - Jeffrey Heer and George Robertson. 2007. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (November 2007), 1240-1247Homework