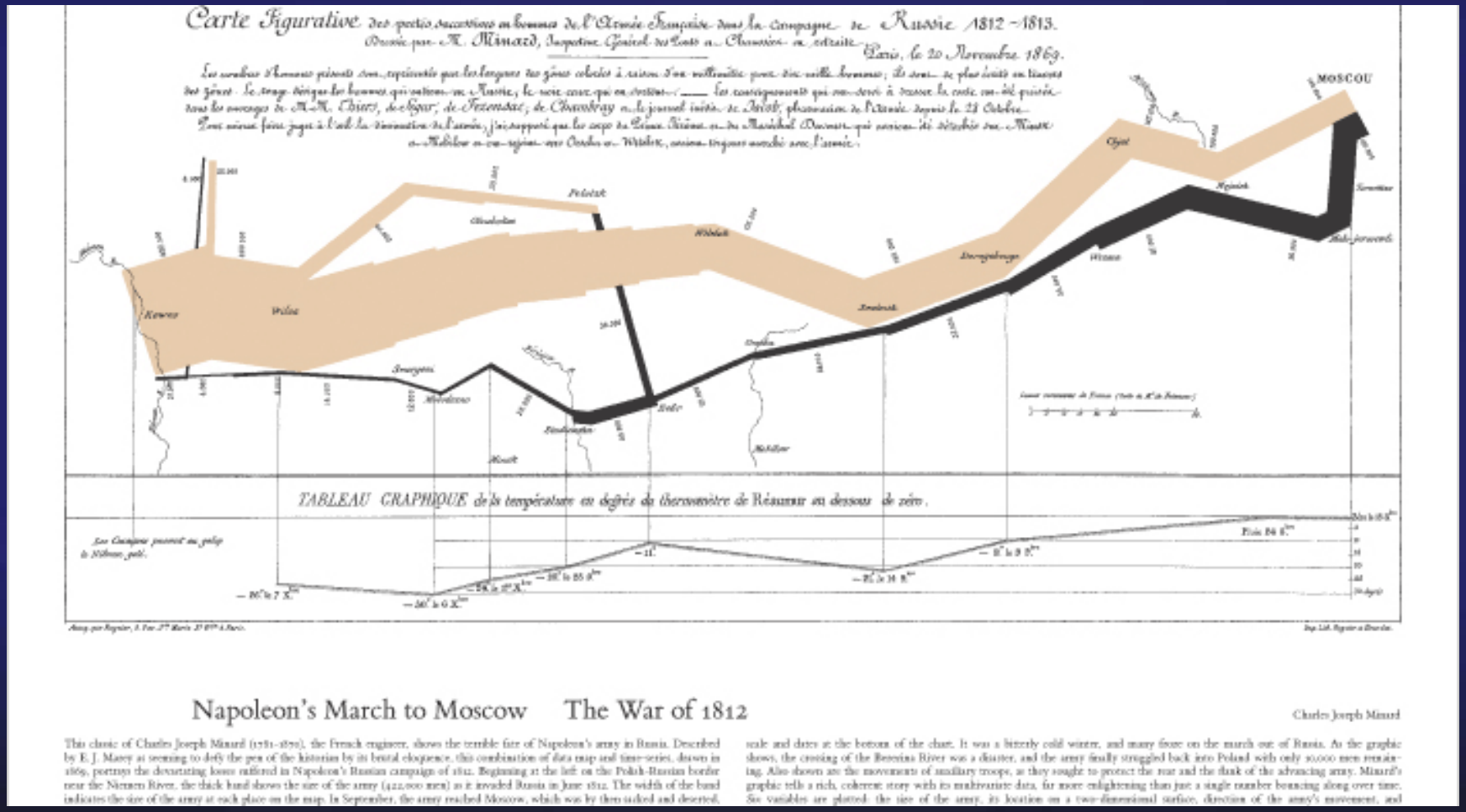
Data Analysis and Information Visualization

Texas Advanced Computing Center

Napoleon Vs. Russia, 1812-1813

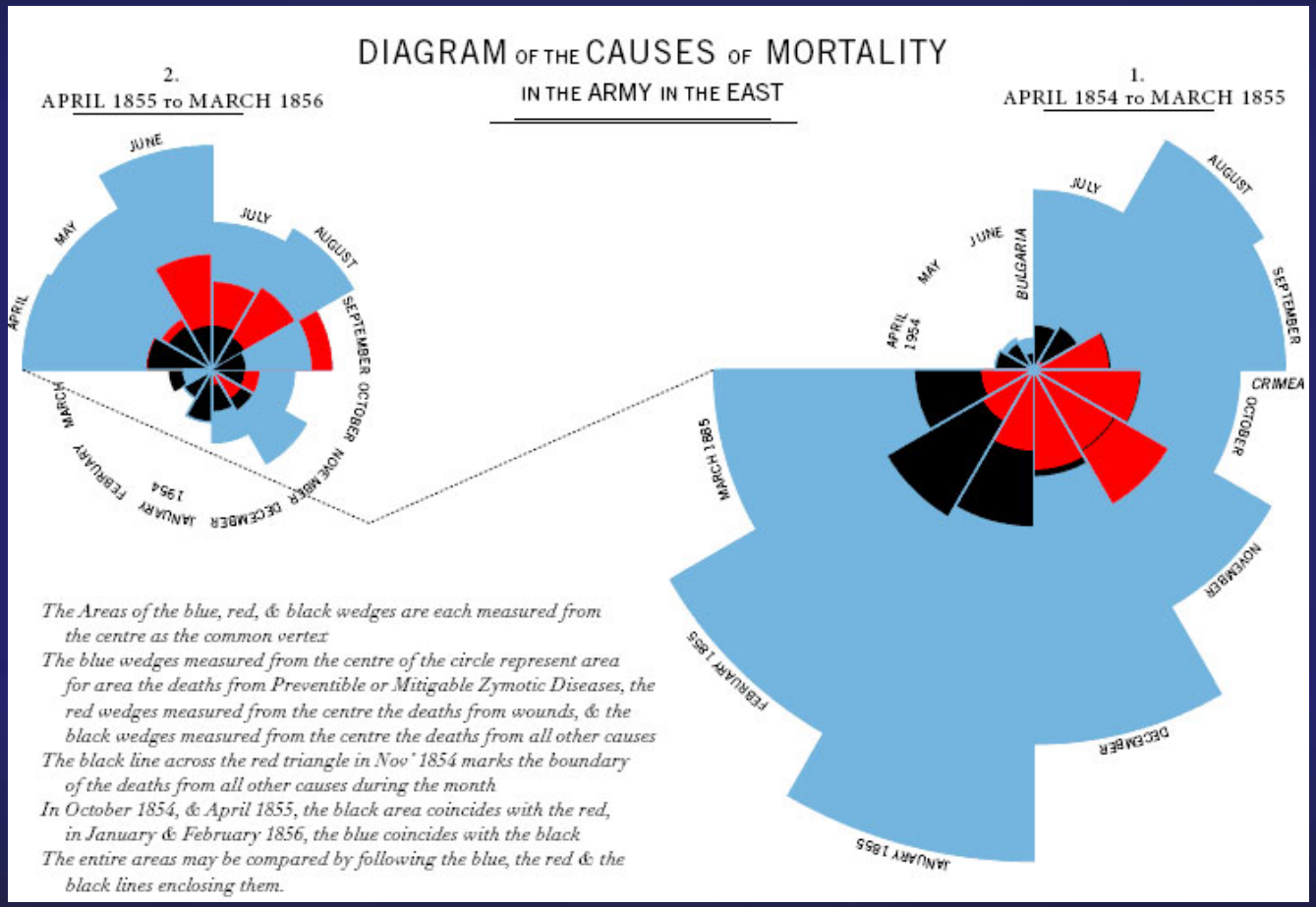


Napoleon's March to Moscow The War of 1812

Charles Joseph Minard

This classic of Charles Joseph Minard (1781–1870), the French engineer, shows the terrible fate of Napoleon's army in Russia. Described by E. J. Marey as seeming to defy the pen of the historian by its brutal eloquence, this combination of data map and time-series, drawn in 1869, portrays the devastating losses suffered in Napoleon's Russian campaign of 1812. Beginning at the left on the Polish-Russian border near the Niemen River, the thick band shows the size of the army (422,000 men) as it invaded Russia in June 1812. The width of the band indicates the size of the army at each place on the map. In September, the army reached Moscow, which was by then tinder and deserted, scale and dates at the bottom of the chart. It was a bitterly cold winter, and many froze on the march out of Russia. As the graphic shows, the crossing of the Berezina River was a disaster, and the army finally struggled back into Poland with only 10,000 men remaining. Also shown are the movements of auxiliary troops, as they sought to protect the rear and the flank of the advancing army. Minard's graphic tells a rich, coherent story with its multivariate data, far more enlightening than just a single number bouncing along over time. Six variables are plotted: the size of the army, its location on a two-dimensional surface, direction of the army's movement, and

Florence Nightingale Cox Comb



Data Analysis and Information Visualization

- Data Sources
  - CSV
  - Excel
  - Databases – SQL, no-SQL, Map-Reduce…
- Data Analysis – Extract Information from Data
  - Statistics: PCA, regression…
  - Machine Learning: clustering, classification…
  - Data Mining
- Visualization – Information Representation
  - histograms, dendograms, tree maps..

Analysis vs. Visualization

- Anscombe's Quartet

Data Set I		Data Set II		Data Set III		Data Set IV	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (Feb 1973), pp17-21

Why Visualize?

- Simple statistical analysis

Data Set I		Data Set II		Data Set III		Data Set IV	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
correlation	0.816		0.816		0.816		0.816	
regression	Y=3+0.5x		Y=3+0.5x		Y=3+0.5x		Y=3+0.5x	

Why Visualize?

Data Set I		Data Set II		Data Set III		Data Set IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (Feb 1973), pp17-21

Data Set I		Data Set II		Data Set III		Data Set IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (Feb 1973), pp17-21

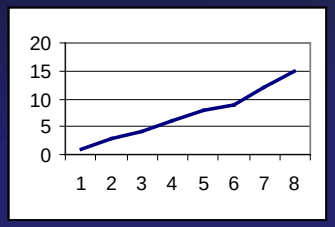
Data Set I		Data Set II		Data Set III		Data Set IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

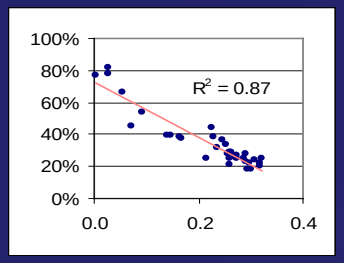
F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (Feb 1973), pp17-21

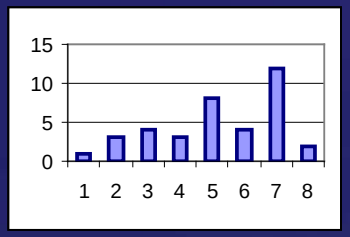
Data Set I		Data Set II		Data Set III		Data Set IV	
X	Y	X	Y	X	Y	X	Y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

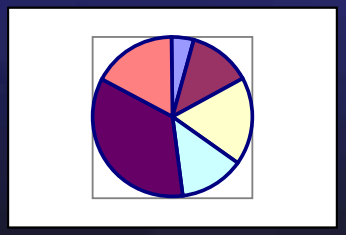
F.J. Anscombe, "Graphs in Statistical Analysis," *American Statistician*, 27 (Feb 1973), pp17-21

In the simple case









Line Graph
- x-axis requires quantitative variable
- Variables have contiguous values
- Familiar/conventional ordering among ordinals

Scatter Plot
- Convey overall impression of relationship between two variables

Bar Graph
- Comparison of relative point values

Pie Chart
- Emphasizing differences in proportion among a few numbers
- Histogram vs. Pie

From Data to Graph

- Information Type:
  - Easy case: 1D, 2D, 3D spatial
  - What about more dimensions?
  - *Structured* data
    - Tree
    - Network
    - Graph
  - Text and document collections

Example: A movie database

Attributes

Types:
- Quantitative
- Ordinal
- Nominal/Categorical

Items

(aka: cases, tuples, data points, …)

Note: No spatial info!

Microsoft Excel - film.txt

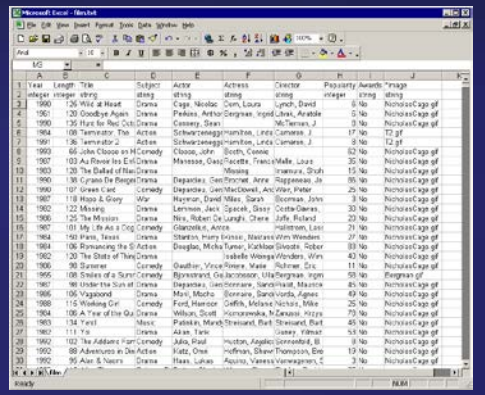
	A	B	C	D	E	F	G	H	I	J
1	Year	Length	Title	Subject	Actor	Actress	Director	Popularity	Awards	*Image
2	integer	integer	string	string	string	string	string	integer	string	string
3	1990	125	Wild at Heart	Drama	Cage, Nicolas	Dern, Laura	Lynch, David	6	No	NicholasCage.gif
4	1961	120	Goodbye Again	Drama	Perkins, Anthor	Bergman, Ingrid	Litvak, Anatole	6	No	NicholasCage.gif
5	1990	135	Hunt for Red Octo	Drama	Connery, Sean		McTiernan, J.	8	No	NicholasCage.gif
6	1984	108	Terminator, The	Action	Schwarzenegge	Hamilton, Linda	Cameron, J.	17	No	T2.gif
7	1991	136	Terminator 2	Action	Schwarzenegge	Hamilton, Linda	Cameron, J.	8	No	T2.gif
8	1993	65	John Cleese on H	Comedy	Cleese, John	Booth, Connie		62	No	NicholasCage.gif
9	1987	103	Au Revoir les Enf	Drama	Manesse, Gasp	Racette, Francii	Malle, Louis	35	No	NicholasCage.gif
10	1983	128	The Ballad of Nar	Drama		Missing	Imamura, Shoh	15	No	NicholasCage.gif
11	1990	138	Cyrano De Berger	Drama	Depardieu, Ger	Brochet, Anne	Rappeneau, Je	86	No	NicholasCage.gif
12	1990	107	Green Card	Comedy	Depardieu, Ger	MacDowell, And	Weir, Peter	25	No	NicholasCage.gif
13	1987	118	Hope & Glory	War	Hayman, David	Miles, Sarah	Boorman, John	3	No	NicholasCage.gif
14	1982	122	Missing	Drama	Lemmon, Jack	Spacek, Sissy	Costa-Gavras,	30	No	NicholasCage.gif
15	1986	125	The Mission	Drama	Niro, Robert De	Lunghi, Cherie	Joffe, Roland	20	No	NicholasCage.gif
16	1987	101	My Life As a Dog	Comedy	Glanzelius, Anton		Hallstrom, Lass	21	No	NicholasCage.gif
17	1984	150	Paris, Texas	Drama	Stanton, Harry I	Kinski, Nastass	Wim Wenders	27	No	NicholasCage.gif
18	1984	106	Romancing the S	Action	Douglas, Micha	Turner, Kathleer	Silvestri, Rober	83	No	NicholasCage.gif
19	1982	120	The State of Thin	Drama		Isabelle Weinga	Wenders, Wim	40	No	NicholasCage.gif
20	1986	98	Summer	Comedy	Gauthier, Vince	Riviere, Marie	Rohmer, Eric	11	No	NicholasCage.gif
21	1955	108	Smiles of a Sumr	Comedy	Bjornstrand, Gu	Jacobsson, Ulla	Bergman, Ingm	58	No	Bergman.gif
22	1987	98	Under the Sun of	Drama	Depardieu, Ger	Bonnaire, Sandi	Pialat, Maurice	45	No	NicholasCage.gif
23	1985	105	Vagabond	Drama	Meril, Macha	Bonnaire, Sandi	Varda, Agnes	49	No	NicholasCage.gif
24	1988	115	Working Girl	Comedy	Ford, Harrison	Griffith, Melanie	Nichols, Mike	25	No	NicholasCage.gif
25	1984	106	A Year of the Qui	Drama	Wilson, Scott	Komorowska, M	Zanussi, Krzys	78	No	NicholasCage.gif
26	1983	134	Yentl	Music	Patinkin, Mandy	Streisand, Barb	Streisand, Barb	46	No	NicholasCage.gif
27	1982	111	Yol	Drama	Akan, Tarik		Guney, Yilmaz	53	No	NicholasCage.gif
28	1992	102	The Addams Fam	Comedy	Julia, Raul	Huston, Anjelic	Sonnenfeld, B.	8	No	NicholasCage.gif
29	1992	88	Adventures in Din	Action	Katz, Omri	Hoffman, Shawr	Thompson, Bre	19	No	NicholasCage.gif
30	1992	95	Alan & Naomi	Drama	Haas, Lukas	Aquino, Vaness	Vanwagenen, S	3	No	NicholasCage.gif

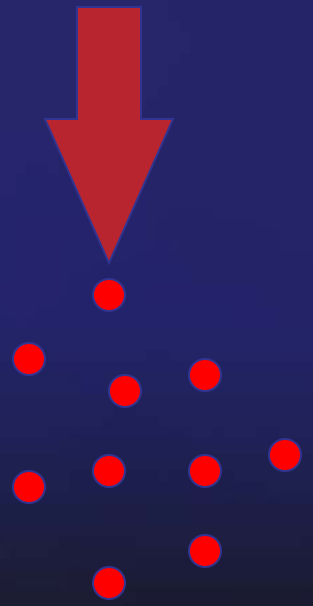
Visual Mapping

1. Map: data items → visual marks

   Visual marks:

   - Points
   - Lines
   - Areas
   - Volumes
   - Glyphs





Visual Mapping

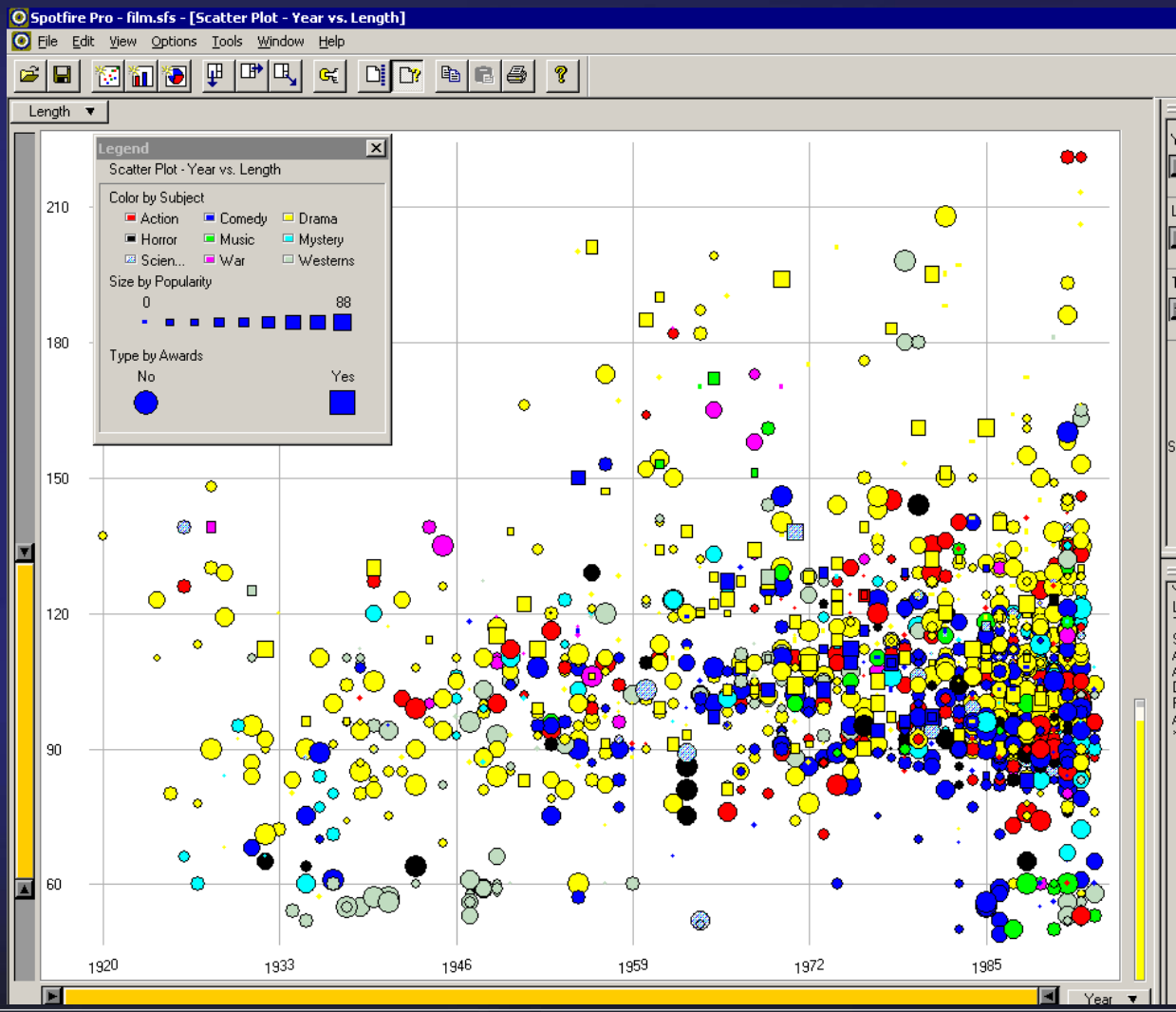
1. Map: data items → visual marks
2. Map: data attributes → visual properties of marks

Visual properties of marks:

- Position, x, y, z
- Size, length, area, volume
- Orientation, angle, slope
- Color, gray scale, texture
- Shape
- Animation, blink, motion
-

Example: A Movie database

- Year → X
- Length → Y
- Popularity → size
- Subject → color
- Award? → shape



Accuracy of Visual Attributes

- Position
- Length
- Angle, Slope
- Size
- Color
- Shape

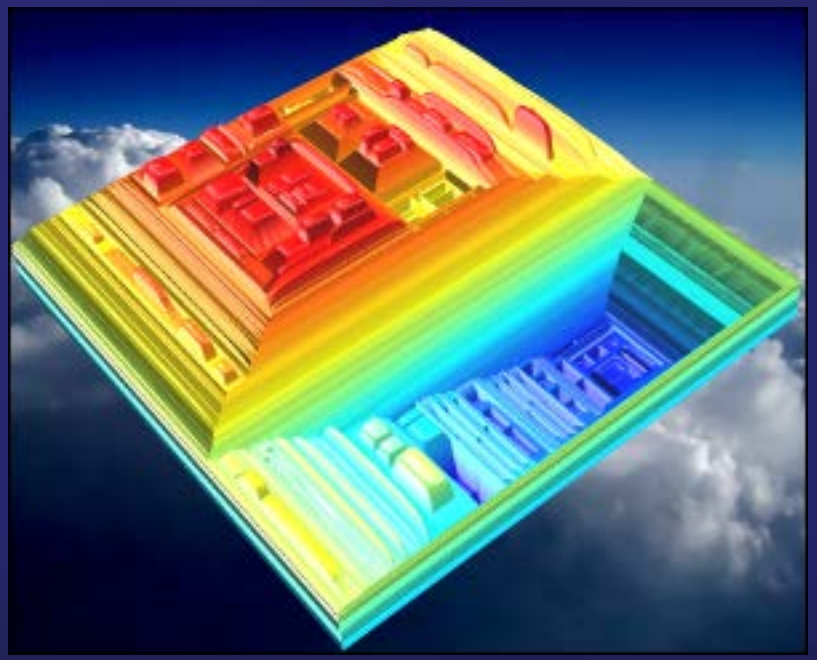
Increased accuracy for quantitative data

Map n-D space onto 2-D screen

- Visual representations:
  - Continuous
    - Heatmap, heightfield, volume
  - Multiple views
    - E.g. plot matrices, brushing histograms, …
  - Complex glyphs
    - E.g. star glyphs, faces …
  - More axes
    - E.g. Parallel coords, star coords, …

Continuous approximations

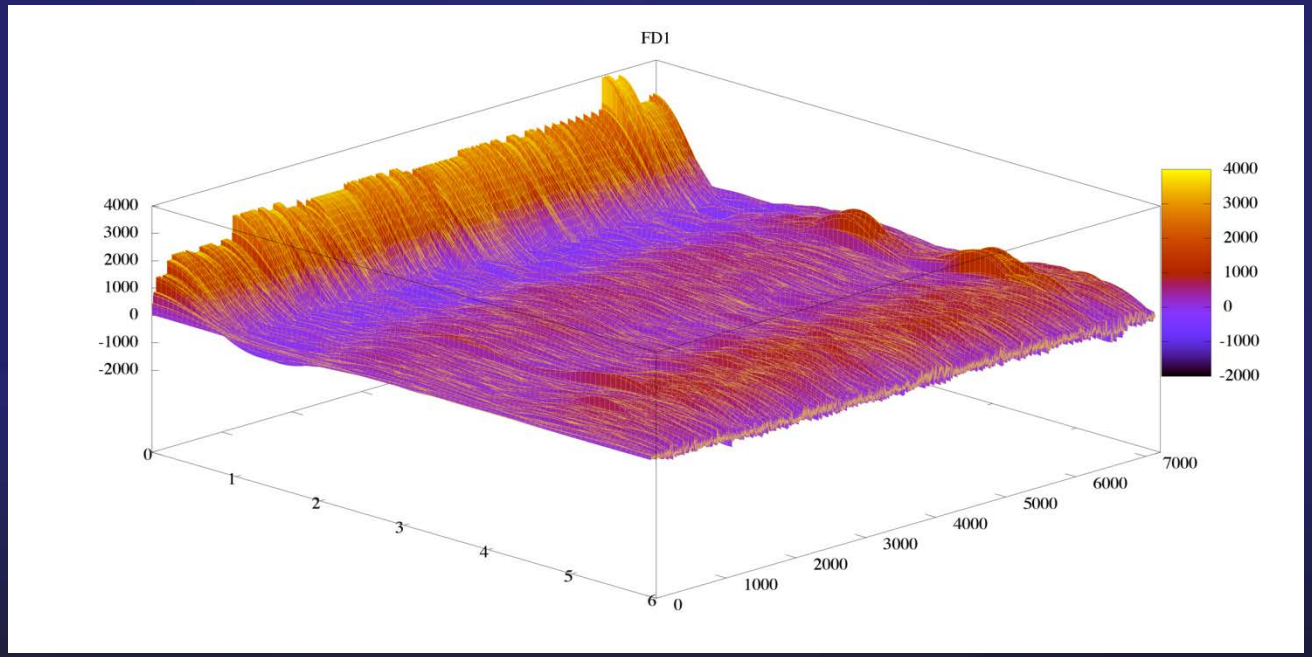
- Reduce a high-dimensional data set to 2D or 3D
- Principal component analysis (PCA):
  - determine 2-3 significant vectors
  - Represent data as linear combinations of those vectors
- Topological Landscapes (Weber et al. 07, Harvey et al. 10)



- Are PCA axes relevant?

Continuous Descriptors

- Transform spatial data into another doman
  - Histogram
  - Fourier transform, other spectra
- Fourier spectra of 7000 carbon molecules with 6 atoms or less



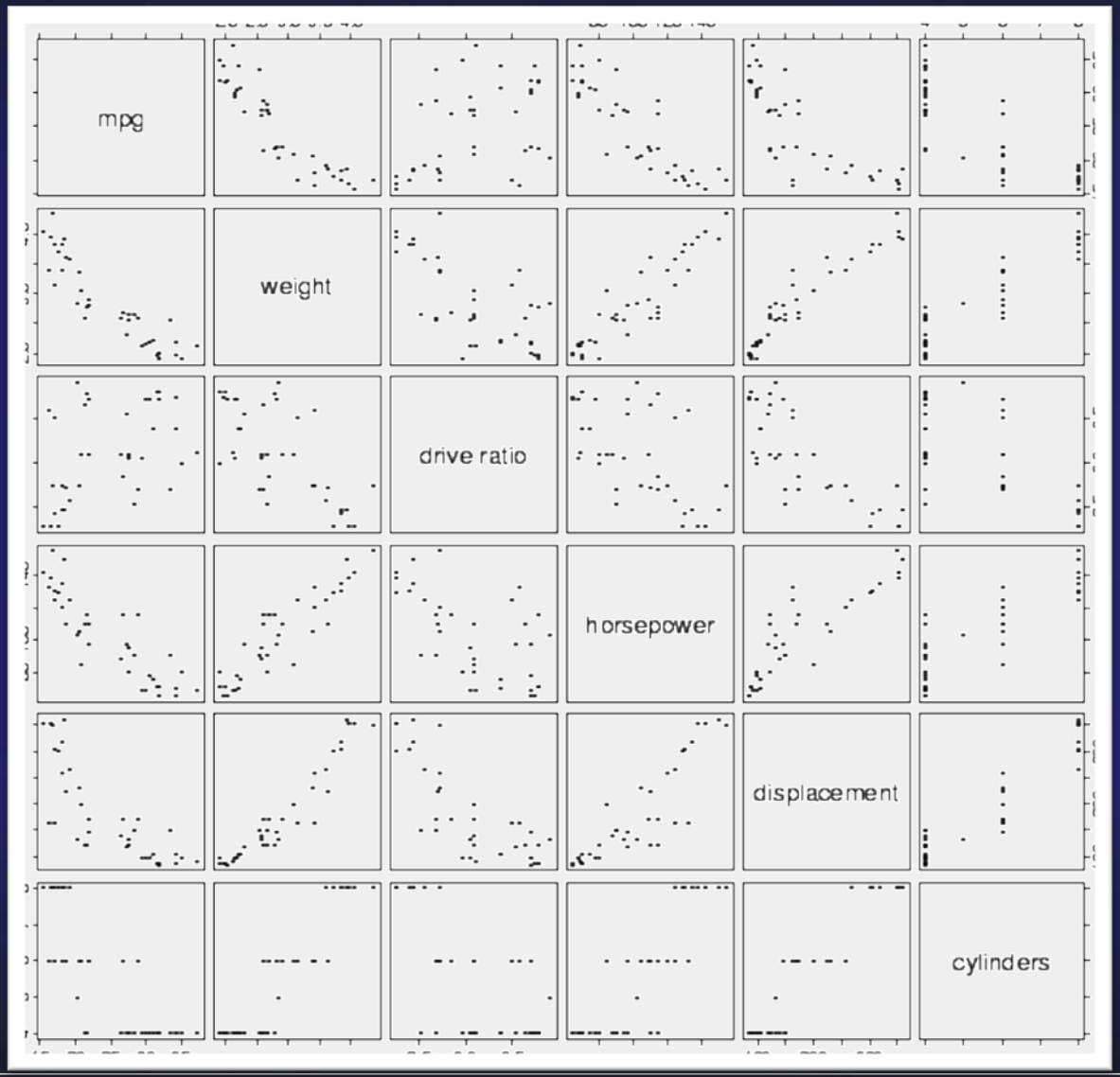
Multiple Views

- Basic idea:
  - Showing multiple views of same data set at the same time.
  - Each individual visualizations might be of same or different types.
  - Brushing and linking
    - With interactive visualizations, All views might be linked so that action, such as selection, on one view might be reflected in all other views.
- Example: Scatter plot matrix
  - Create a 2d views for all attributes pairs

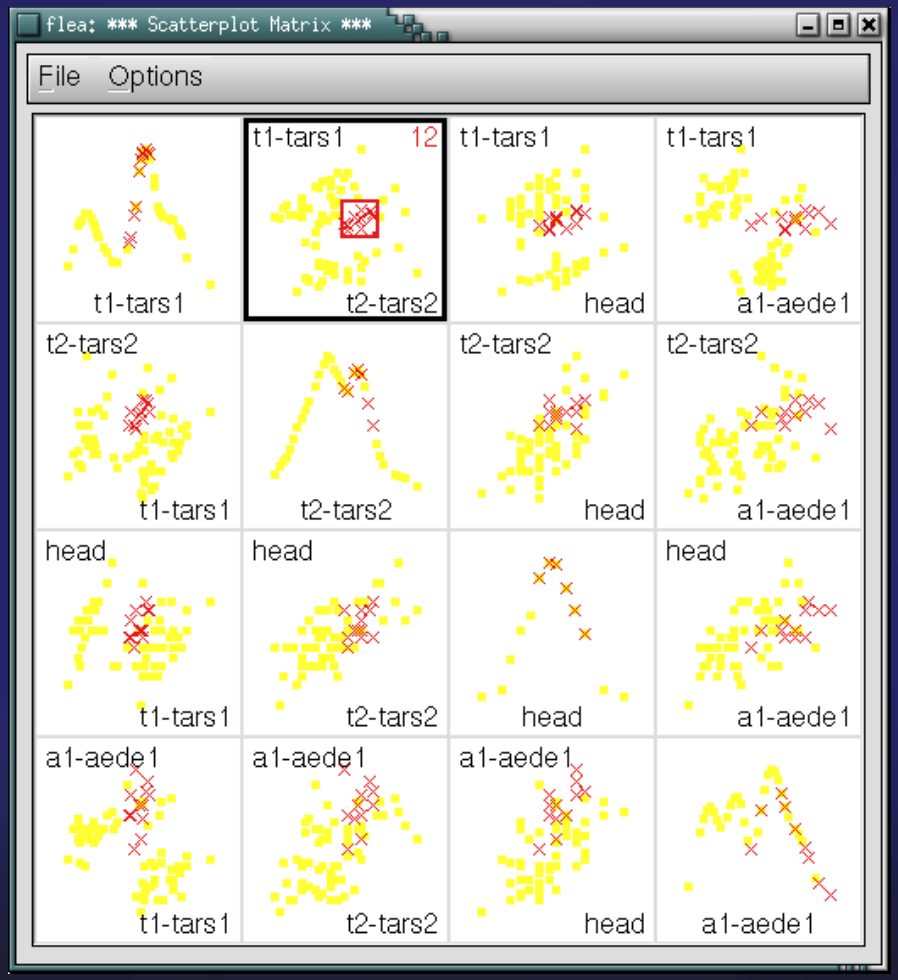
Example Data

country	model name	mpg	weight	ratio	hp	disp.	cyl.
USA	Buick Estate Wagon	16.9	4.360	2.73	155	350	8
USA	Ford Country Squire Wagon	15.5	4.054	2.26	142	351	8
USA	Chevy Malibu Wagon	19.2	3.605	2.56	125	267	8
USA	Chrysler LeBaron Wagon	18.5	3.940	2.45	150	360	8
USA	Chevette	30.0	2.155	3.70	68	98	4
Japan	Toyota Corona	27.5	2.560	3.05	95	134	4
Japan	Datsun 510	27.2	2.300	3.54	97	119	4
USA	Dodge Omni	30.9	2.230	3.37	75	105	4
Germany	Audi 5000	20.3	2.830	3.90	103	131	5
Sweden	Volvo 240 GL	17.0	3.140	3.50	125	163	6
Sweden	Saab 99 GLE	21.6	2.795	3.77	115	121	4
France	Peugeot 694 SL	16.2	3.410	3.58	133	163	6
USA	Buick Century Special	20.6	3.380	2.73	105	231	6
USA	Mercury Zephyr	20.8	3.070	3.08	85	200	6
USA	Dodge Aspen	18.6	3.620	2.71	110	225	6

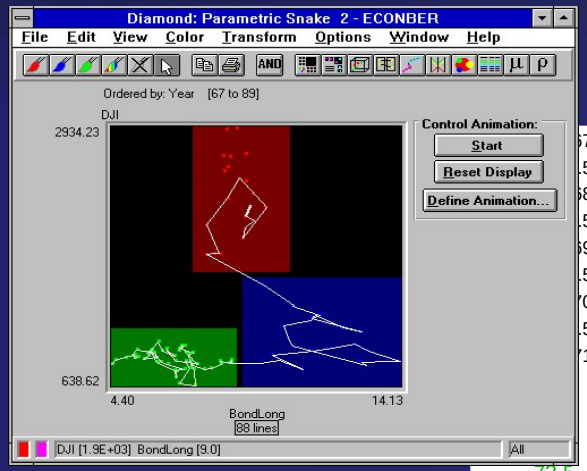
Scatter plot Matrix Example

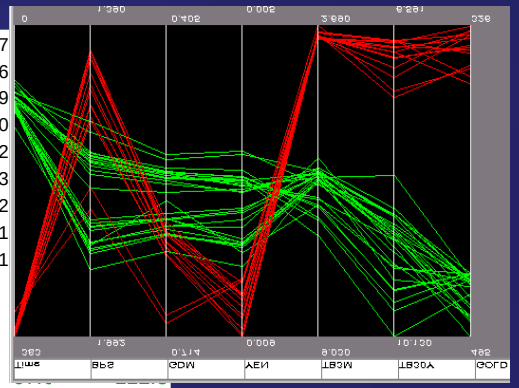


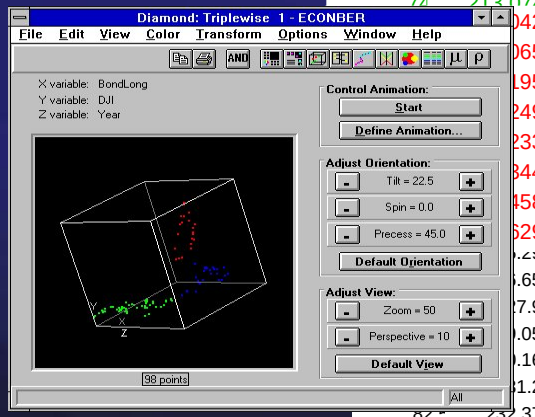
Brushing

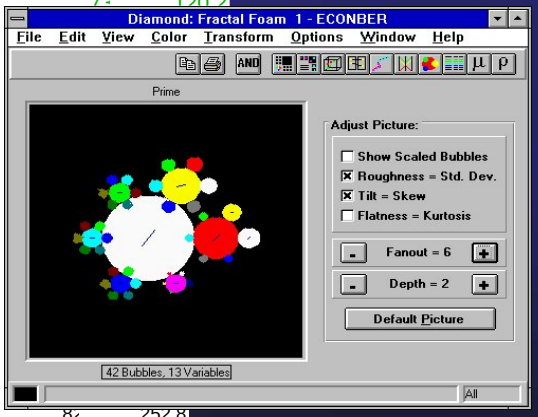

flea: *** Scatterplot Matrix ***
File
Options
t1-tars1
12
t2-tars2
head
a1-aede1

Brushing Across Different Projection Types









73	211.12	0.427	41.542	2734	71.8	126.3
73.5	212.092	0.442	42.542	2738.3	74.4	125
83	233.462	0.979	91.607	3186.6	80.8	266.7
83.5	234.49	0.998	98.885	3306.4	85.3	295.2

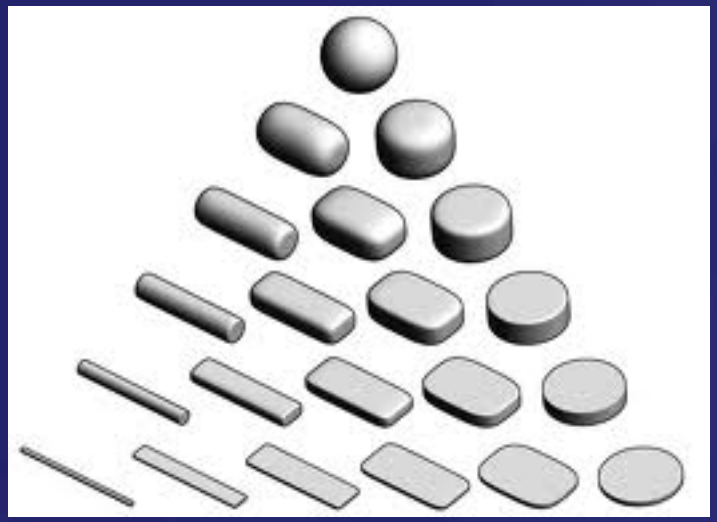
Glyphs

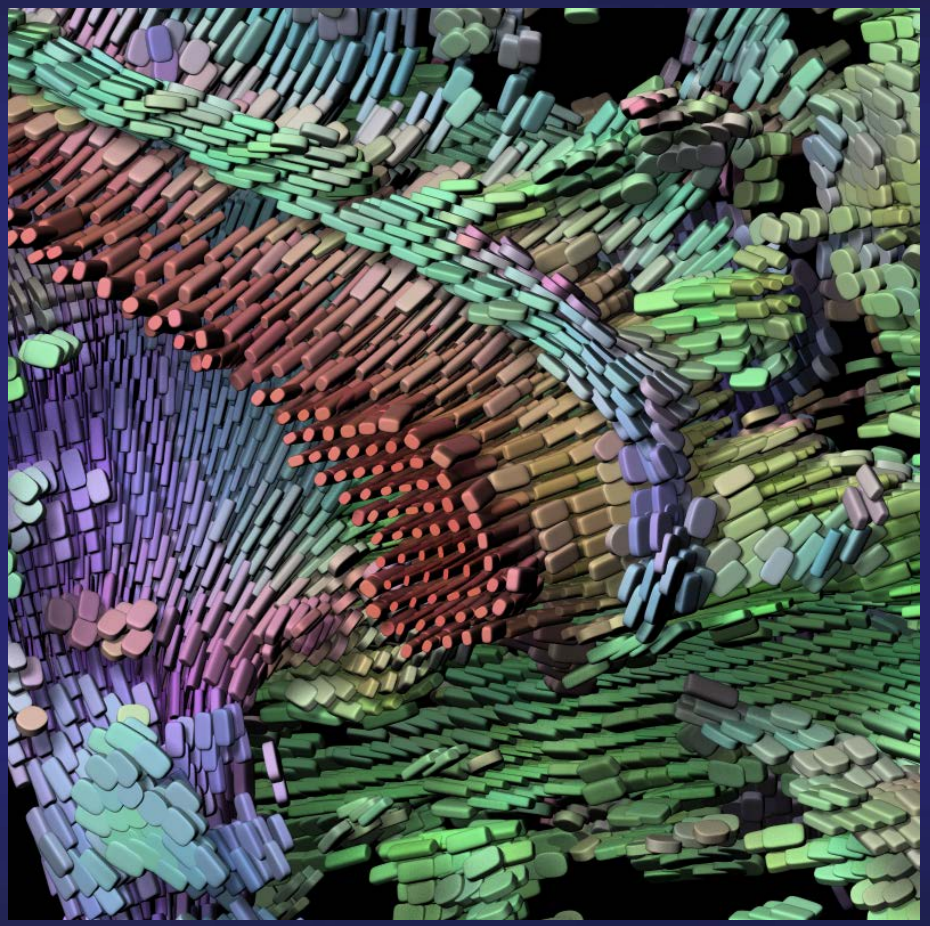
- Glyph
  - composite graphical objects where different geometric and visual attributes are used to encode multidimensional data structures in combination.

- Examples:
  - Superquadrics for DTI
  - Chernoff Face*
    - mapping k-dimensions to facial features

*Herman Chernoff, "The use of faces to represent points in k-dimensional space graphically," *J. Am. Stat. Assoc.*, v68, 361-368 (1973).

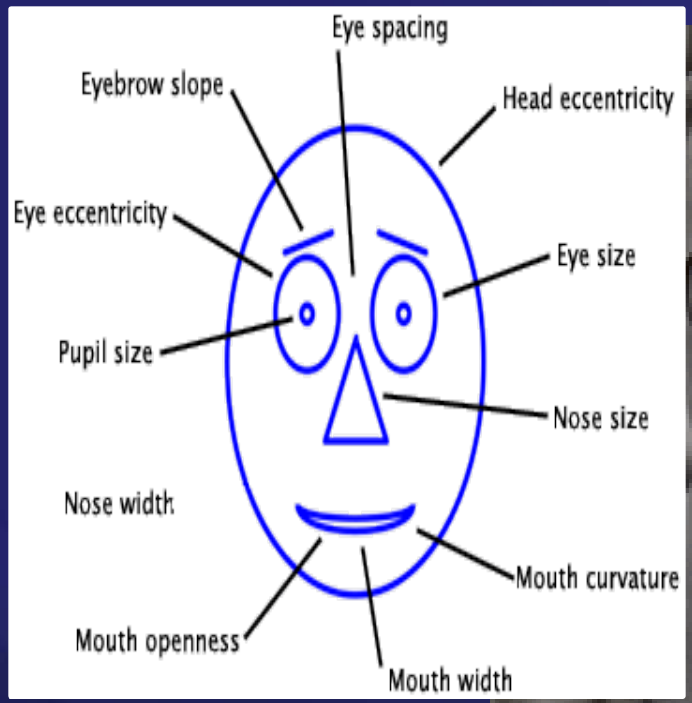
Superquadric glyphs for DT-MRI

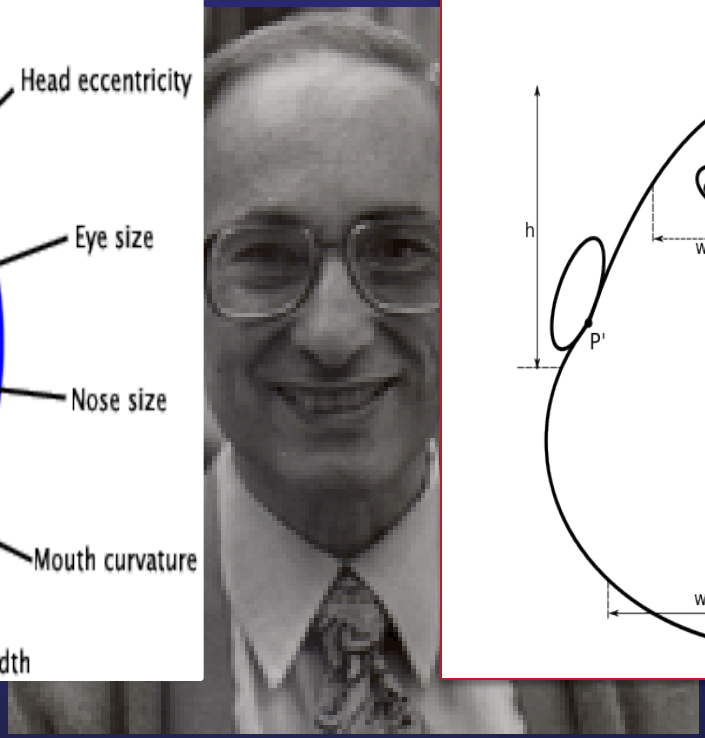


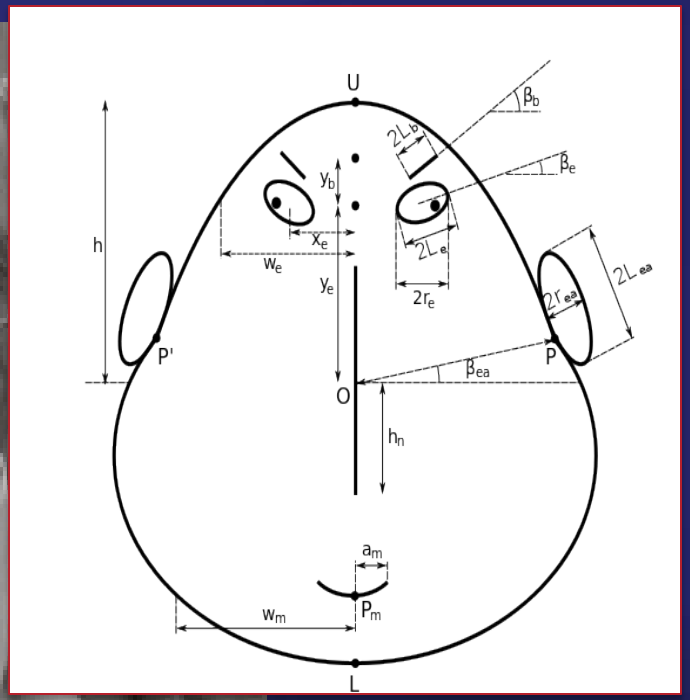


- Determine structure of brain tissue, examining movement along N different axes
- G. Kindlmann, University of Utah / University of Chicago

Glyphs: Chernoff Faces







- http://www.stat.harvard.edu/People/Faculty/Herman_Chernoff/

- http://hesketh.com/schampeo/projects/Faces/chernoff.html

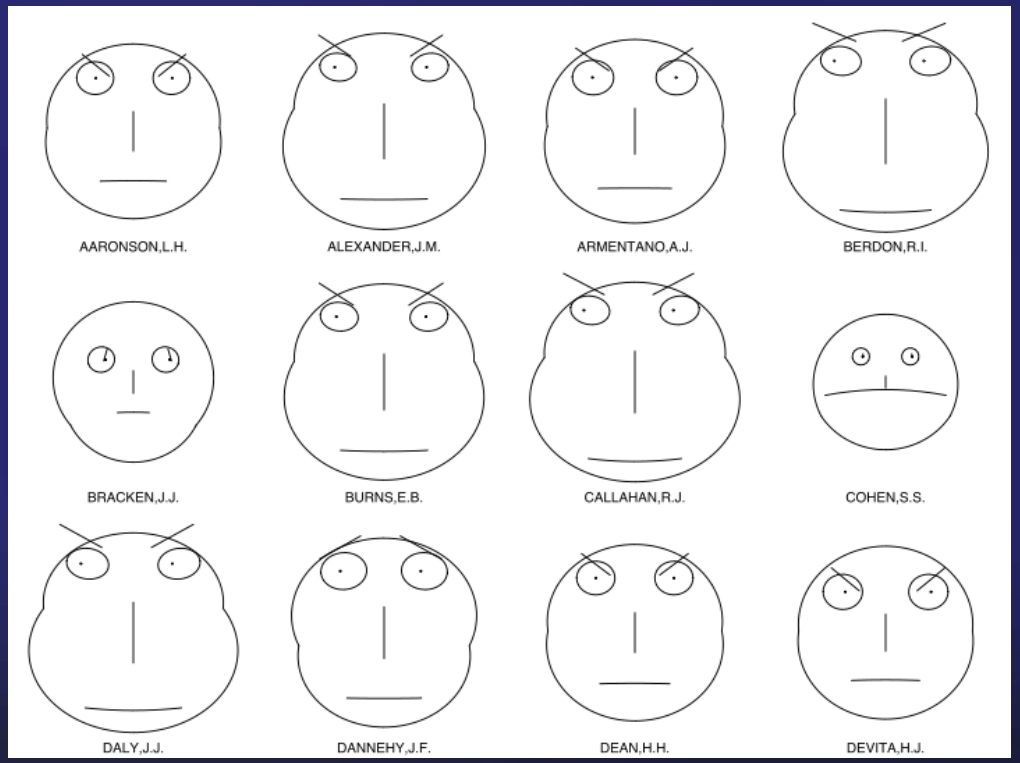
Chernoff Face Example

- Map to 10 dimension binary vector

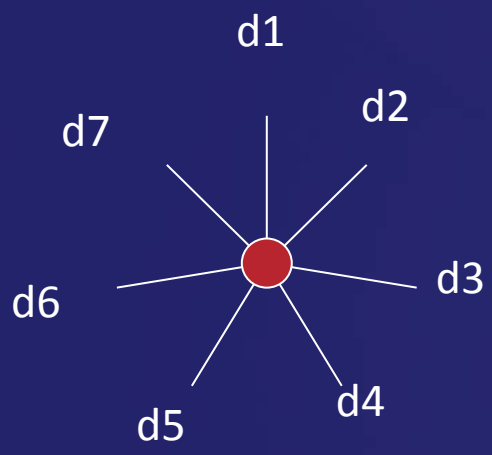
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

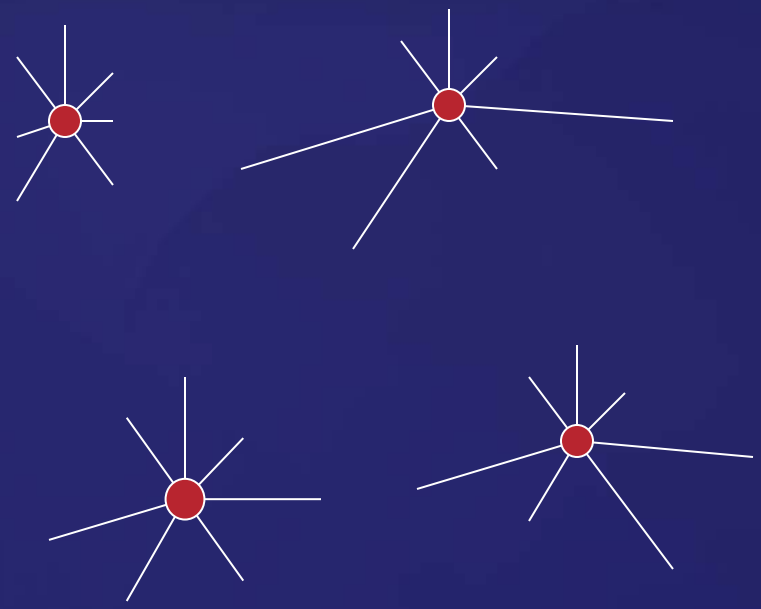
[1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

- Evaluation Of Judges



Star Glyph

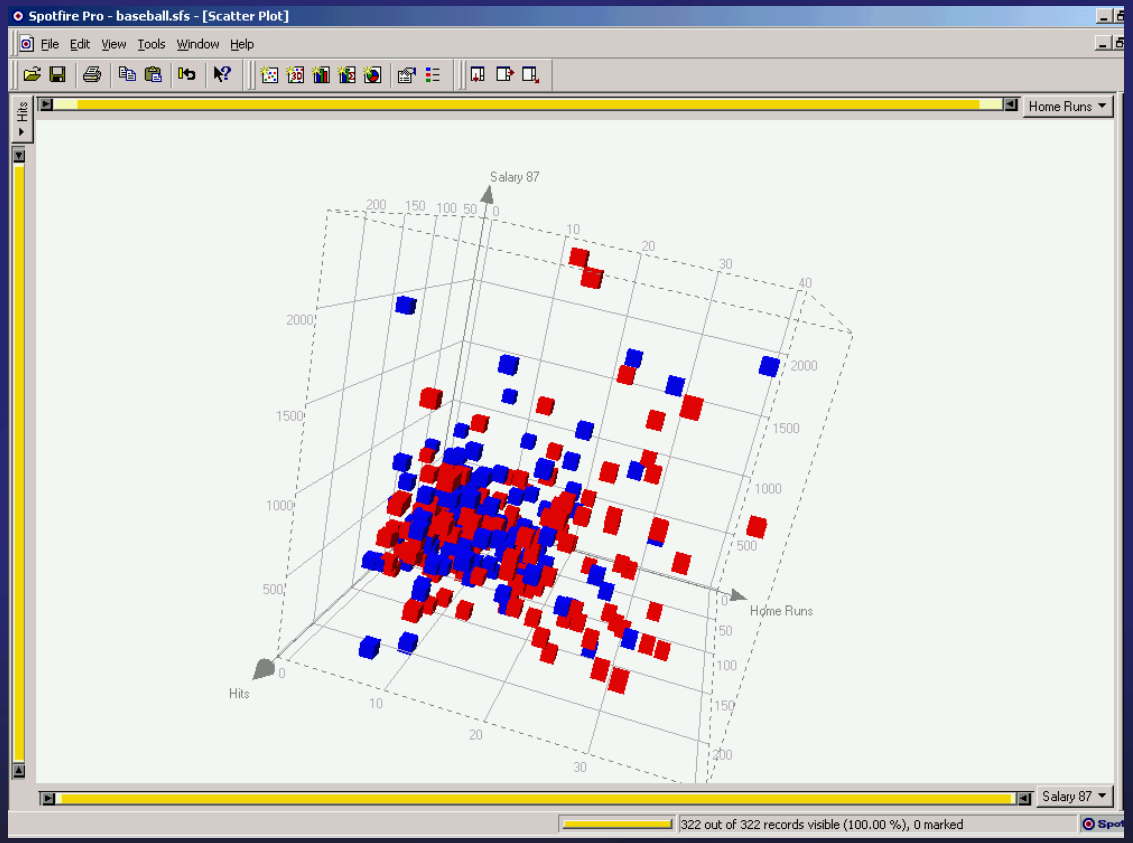




- What's a problem with using star glyphs?

Using additional axes

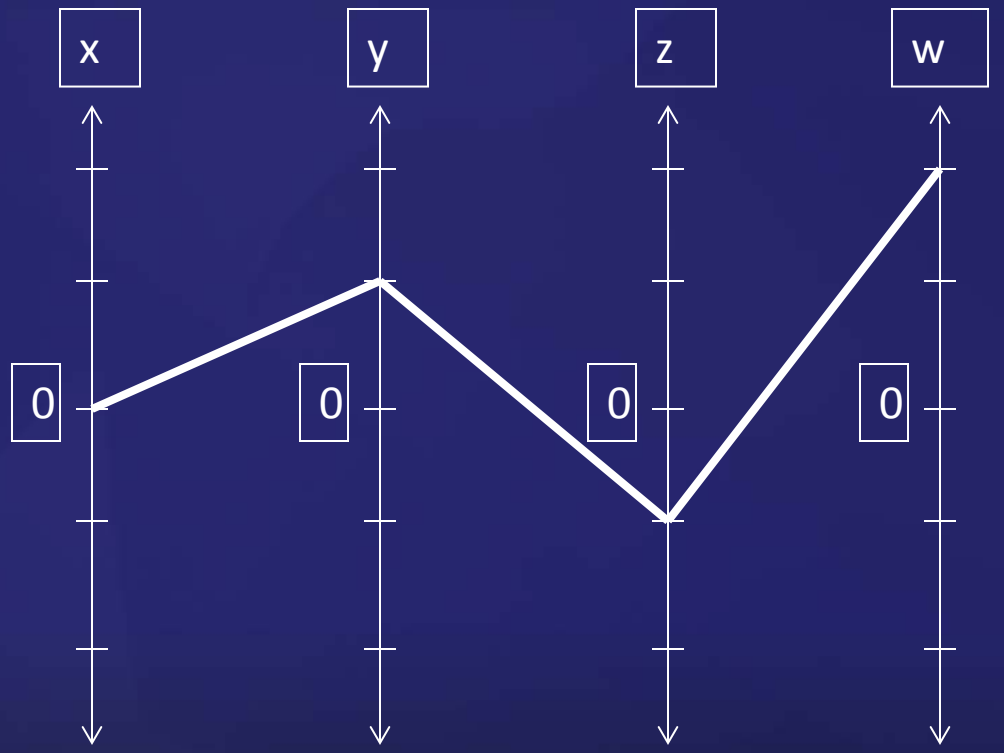
- Easy example:
  - 2D scatter plot → 3D scatter plot



- Space > 3D ?

Parallel Coordinates

- Instead of orthogonal axes, let's go parallel
- (0,1,-1,2)=

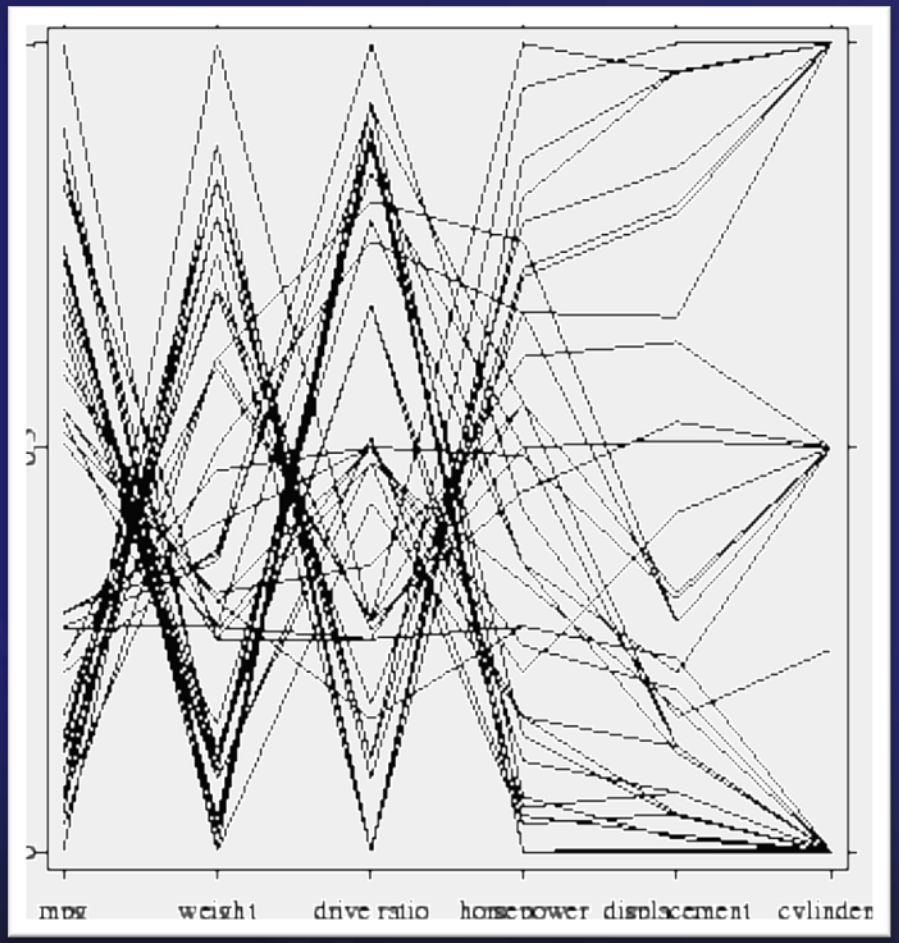


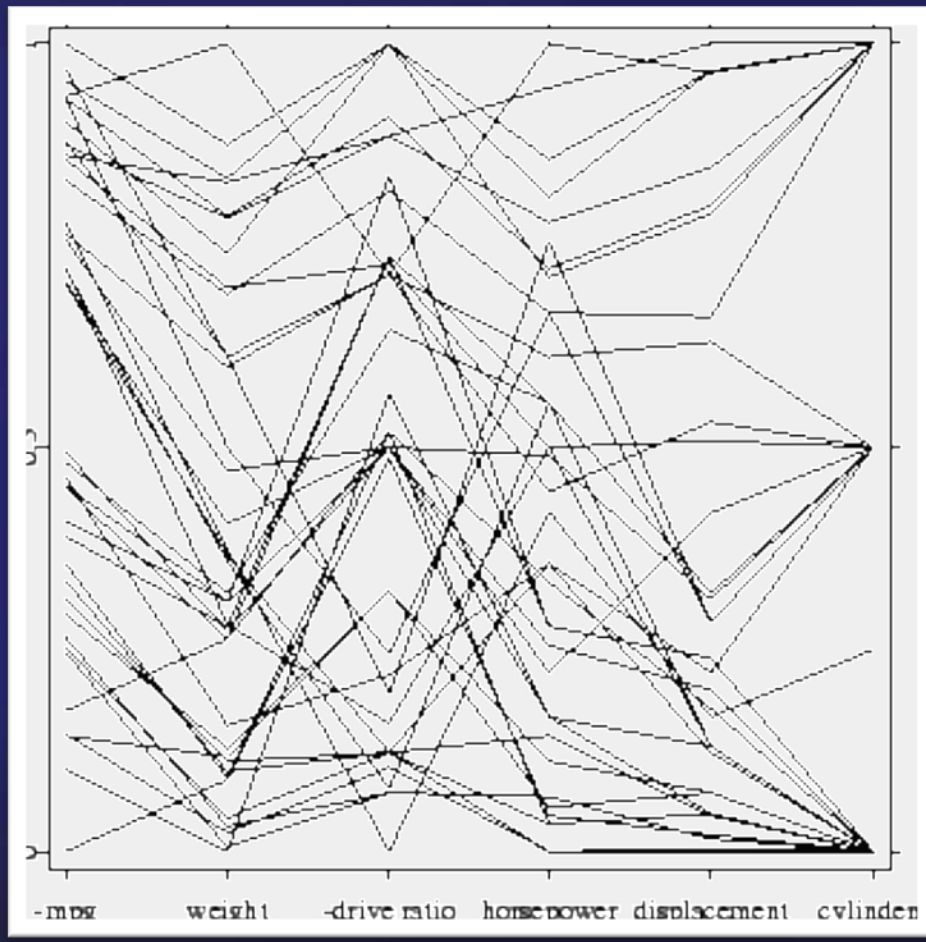
Inselberg, "Multidimensional detective" (parallel coordinates)

Parallel Coordinates

- Important factors:
  - the scaling of the axes.
  - the order of the axes
  - the rotation of the axes

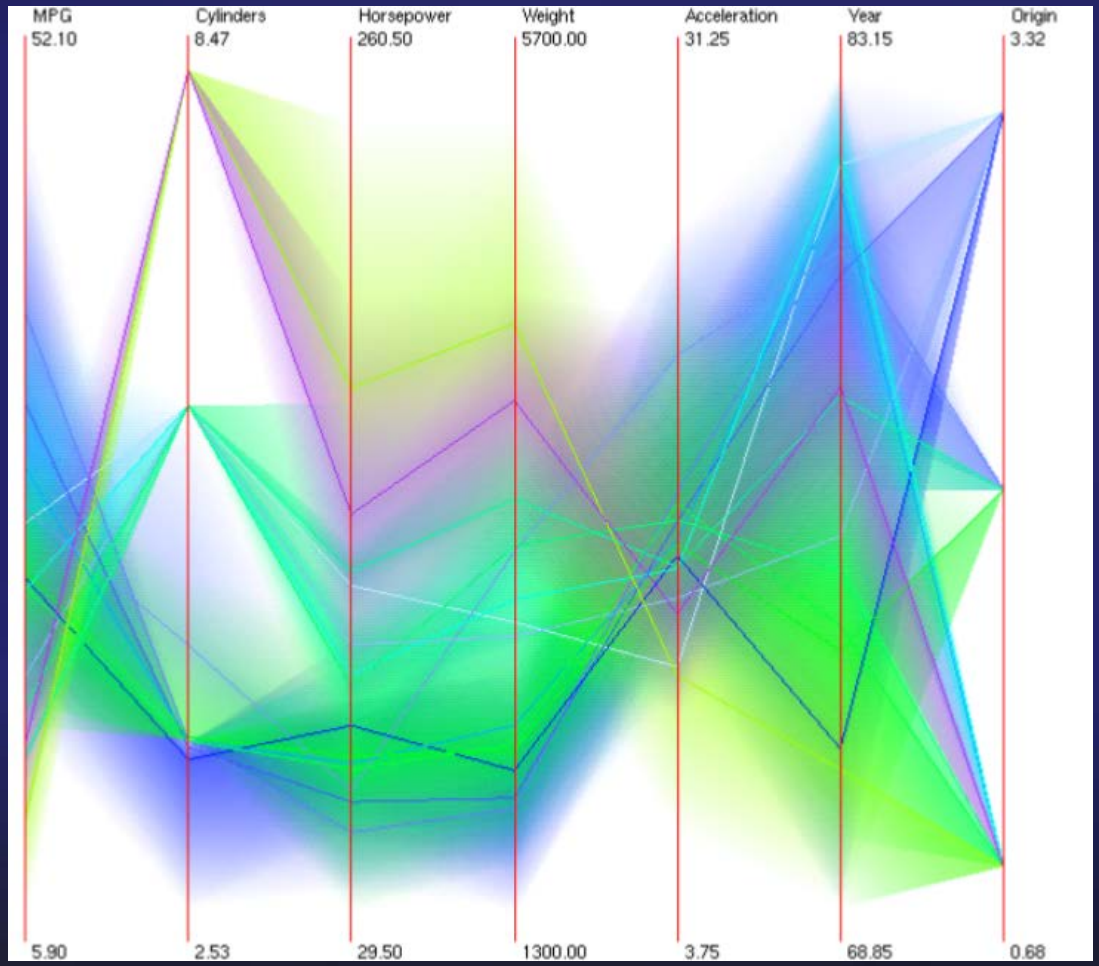
Parallel Coordinates





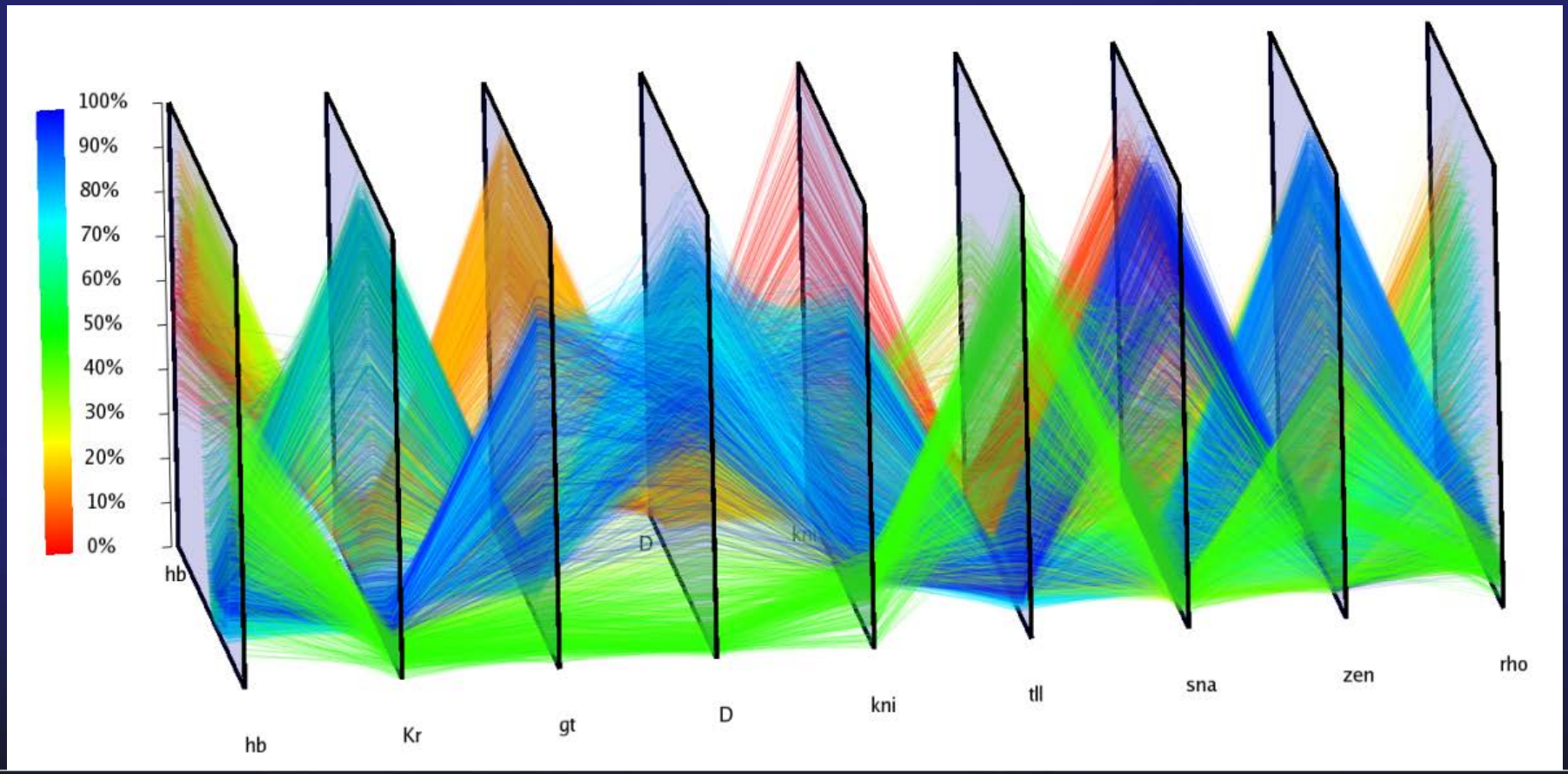
Parallel Coordinates

- Better visualizations



Parallel Coordinates

- 3D parallel coordinates
  - http://www-vis.lbl.gov/Events/SC07/Drosophila/3DParallelCoordinates.png

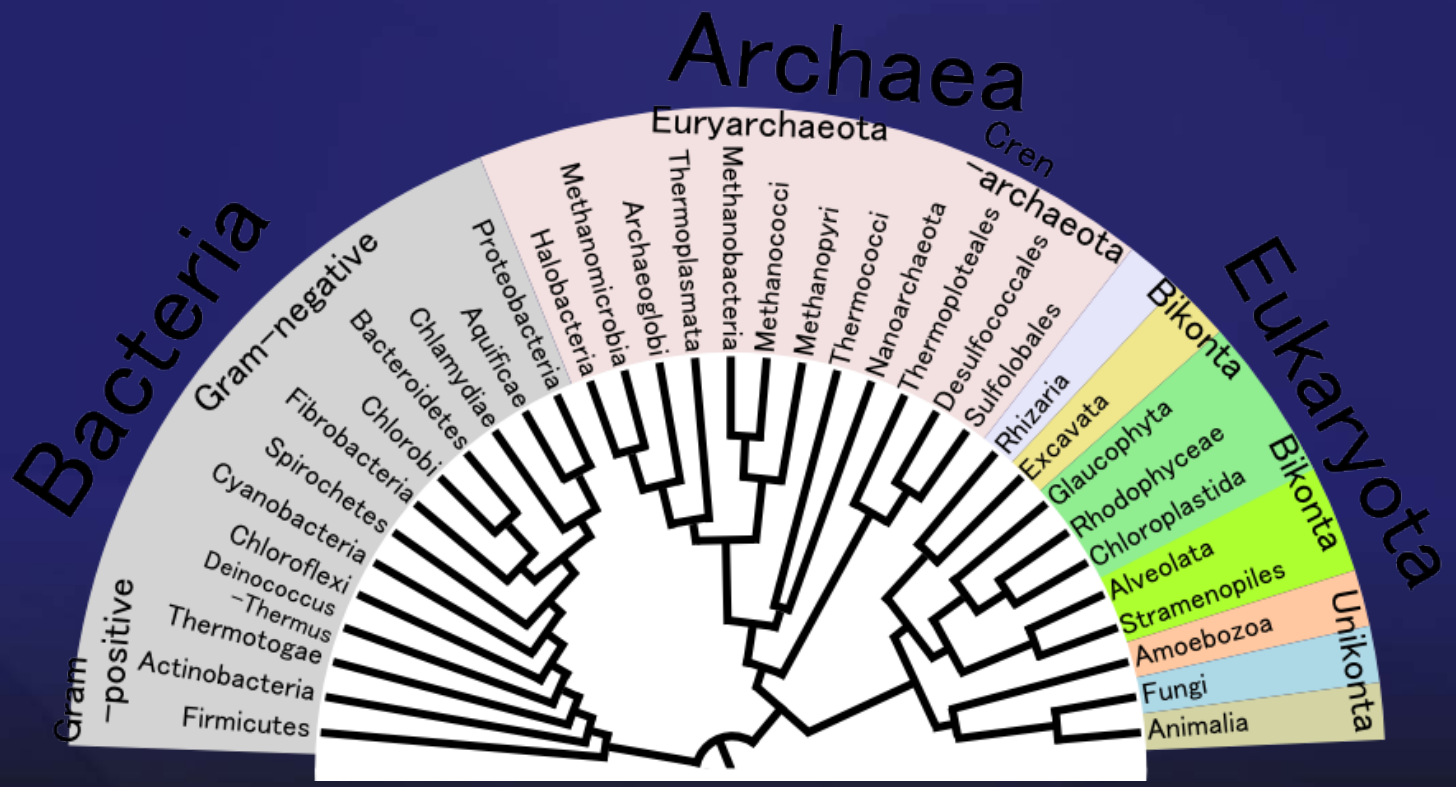


Visualizing Structured Data

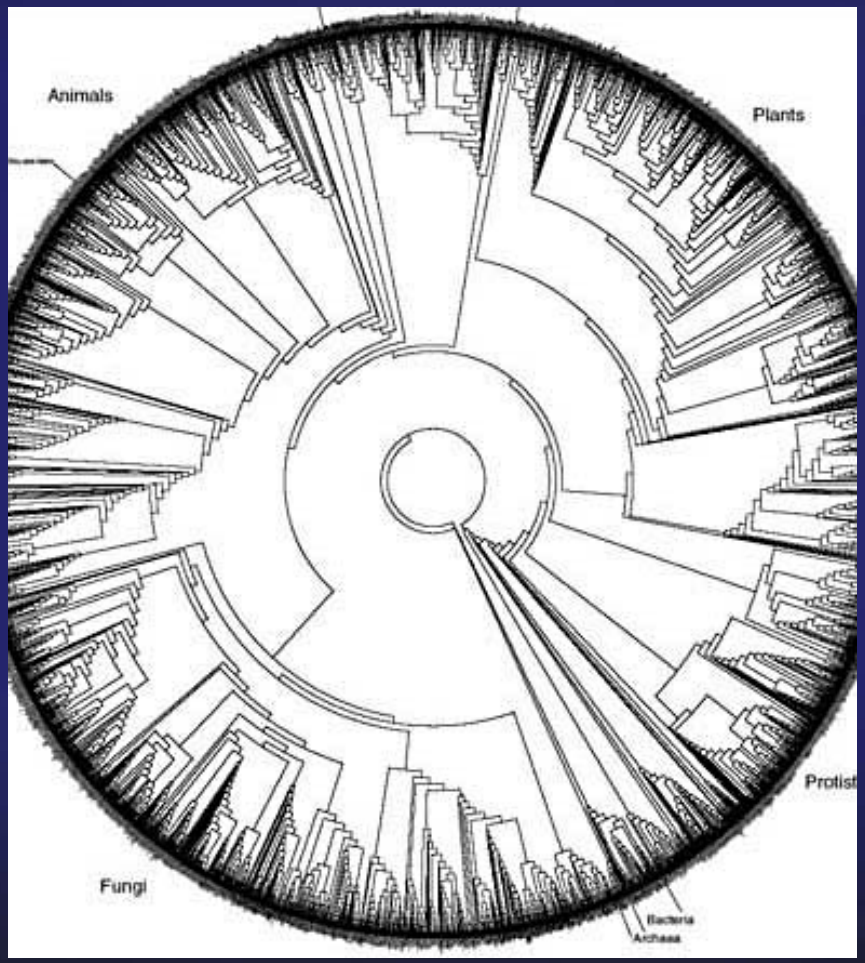
- Some data contains relationships between entities
  - Tree Structures
    - Phylogenetic Trees
    - Presidential voting by state, county and precinct
  - Generalized relations
    - Who knows whom in a college dorm
    - Who follows whom on Twitter

Tree-Structured Data

- Phylogenetic Trees

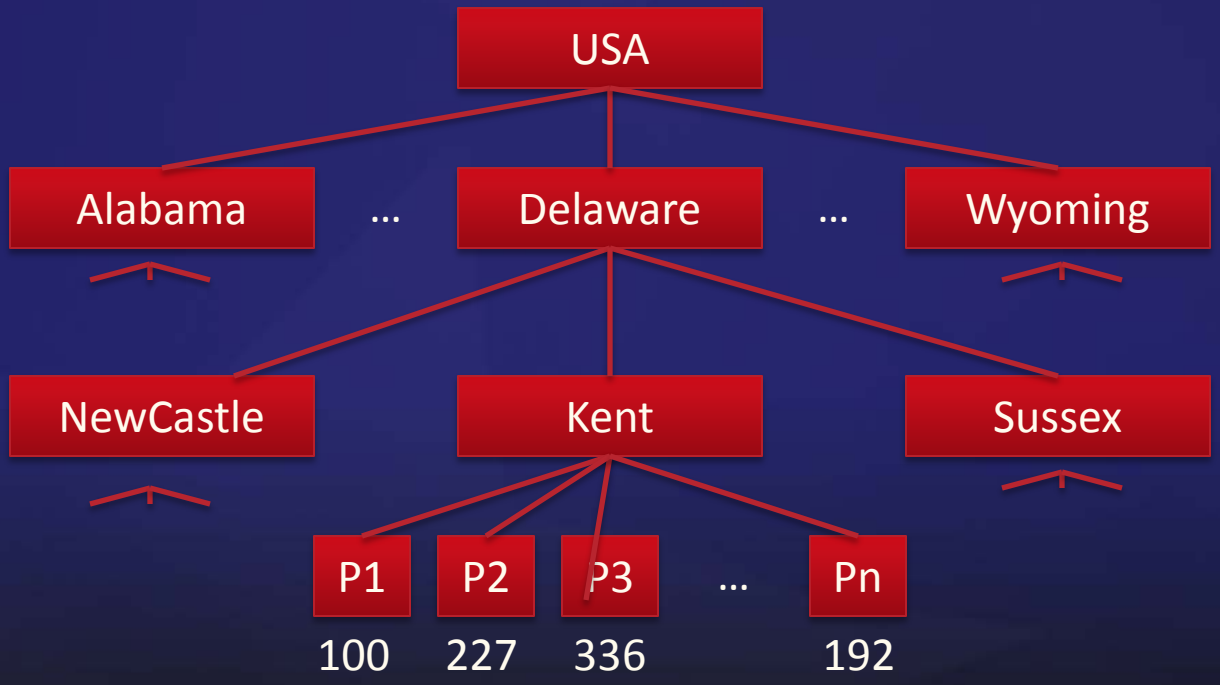


Can Get Busy

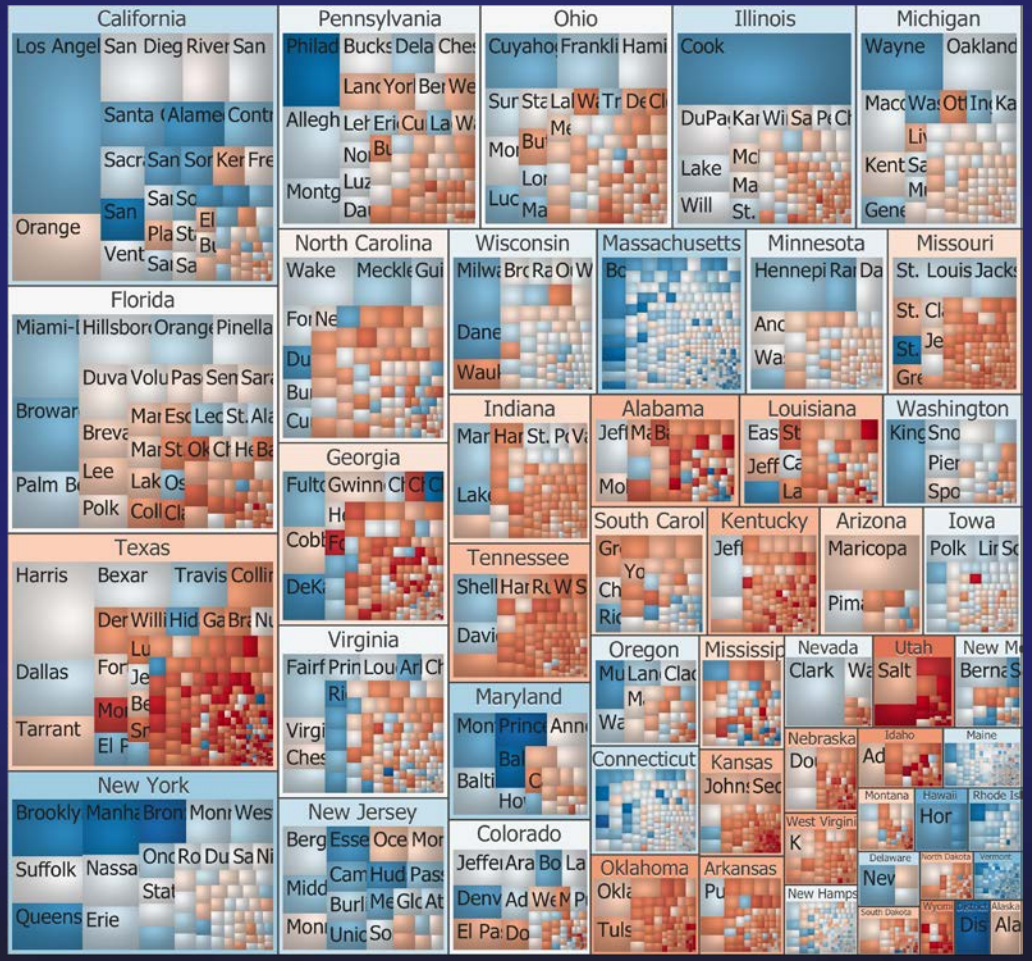


Tree-Structure With Values

- Suppose we know the breakdown of votes for each precinct in the country….



Treemaps


California
Los Angel
San Dieg
River
San
Santa
Alame
Contr
Sacr
San
Sor
Ker
Fre
San
Orange
Vent
Pennsylvania
Philad
Bucks
Dela
Ches
Lanc
York
Ber
We
Allegh
Montg
Ohio
Cuyaho
Frankli
Hami
Illinois
Cook
DuPa
Kar
Wi
Sa
Lake
Will
St.
Michigan
Wayne
Oakland
Kent
Gene
Florida
Miami-
Hillsbor
Orange
Pinella
Duva
Volu
Pas
Sen
Sara
Browar
Breva
Lee
Polk
Palm B
North Carolina
Wake
Meckle
Gui
Wisconsin
Milwa
Bro
Ra
Dane
Waul
Massachusetts
Minnesota
Hennepi
Ran
Da
Anc
Missouri
St. Louis
Jacks
St. Cl
St.
Gre
Texas
Harris
Bexar
Travis
Collin
Dallas
Tarrant
Georgia
Fulto
Gwinn
Cobb
DeKa
Indiana
Mar
Har
St. P
Lake
Alabama
Jeff
Mo
Louisiana
Eas
St
Jeff
La
Washington
King
Sno
Pier
Spo
Tennessee
Shell
Har
Davi
South Carol
Kentucky
Jeff
Arizona
Maricopa
Pima
Iowa
Polk
Virginia
Fairf
Prin
Lou
Virgi
Ches
Maryland
Mon
Princ
Balti
Oregon
Mississip
Nevada
Clark
Utah
Salt
New M
Berna
Connecticut
Kansas
John
Nebraska
Idaho
Maine
Montana
Hawaii
Rhode Is
Hor
New York
Brookly
Manha
Bron
Mon
Wes
Suffolk
Nassa
Queens
Erie
New Jersey
Berg
Esse
Oce
Mor
Cam
Hud
Pass
Midd
Burl
Mon
Unic
Colorado
Jeffe
Ara
Bo
Denv
Ad
El Pa
Do
Oklahoma
Okla
Tuls
Arkansas
West Virgini
New Hamps
Delaware
North Dakota
Vermont
South Dakota
Wyomi
Alaska
Dis
Ala

Generalized Relation Data

Bob knows Bill

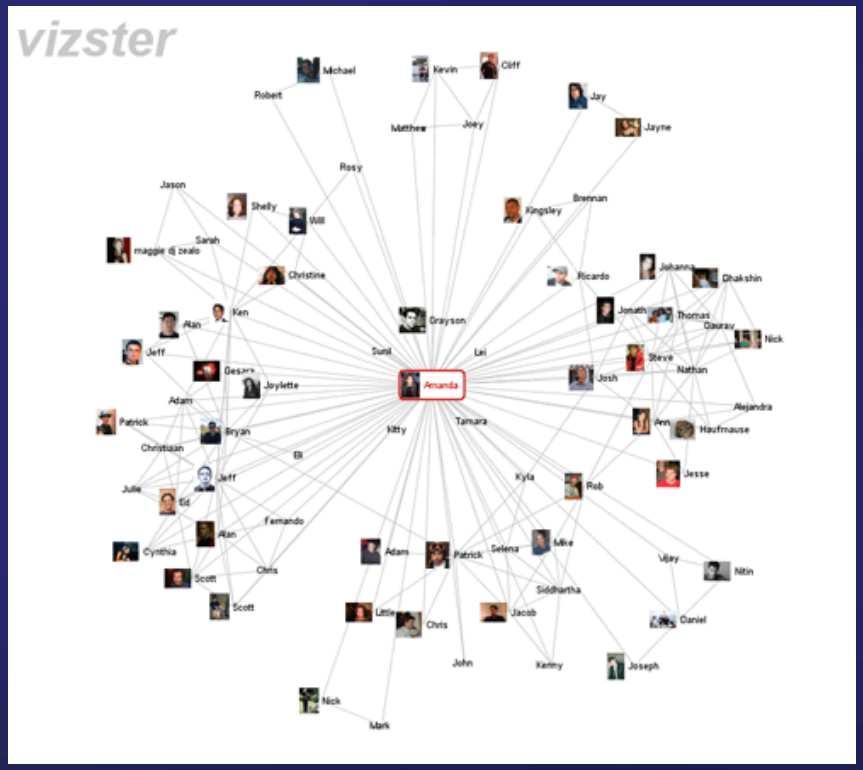
Bill knows Ted

Ted knows Ann

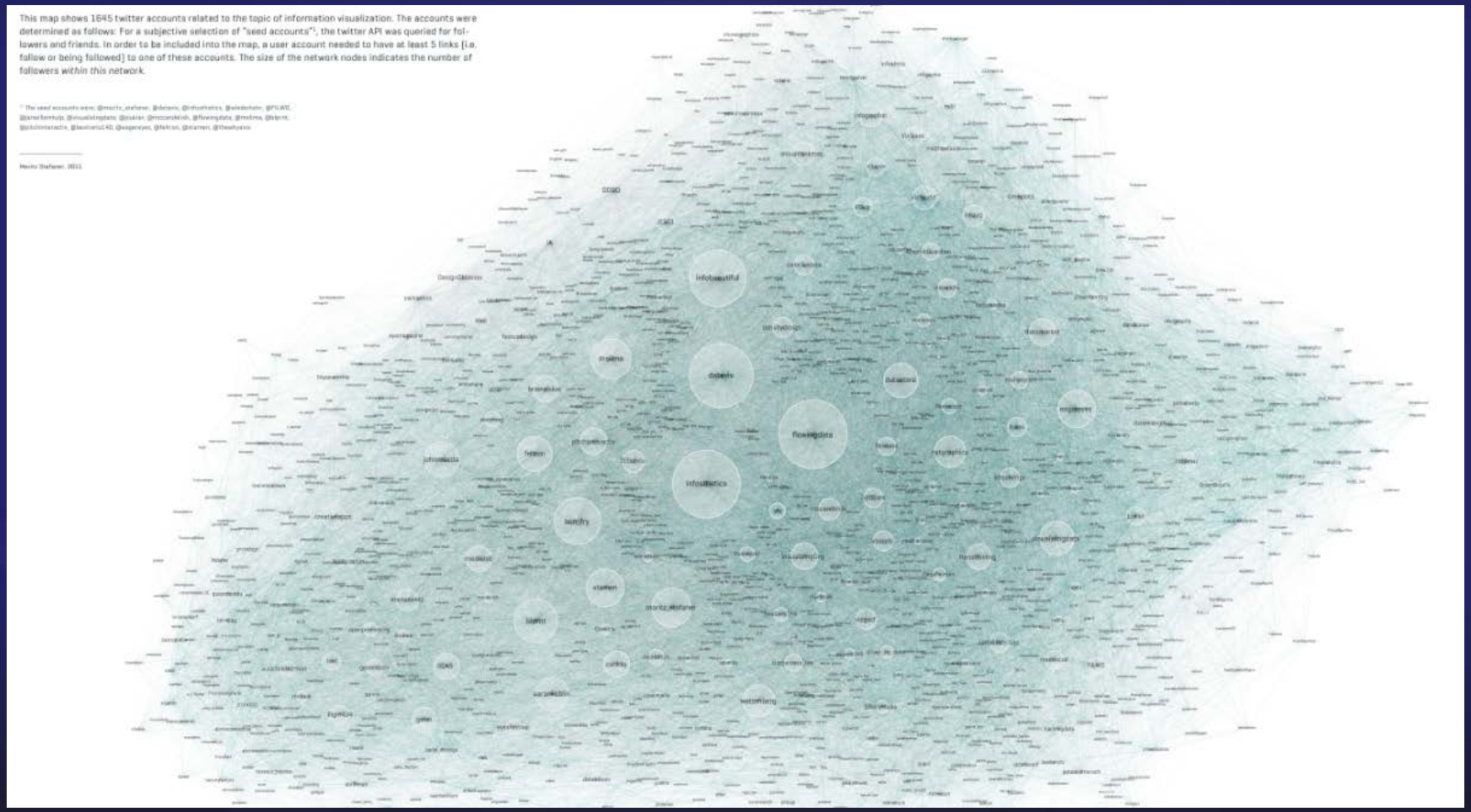
Bill knows Ann

Bob knows Ken

...



Gets Busy Fast…



Relation Data With Weights

Bob likes Bill a little

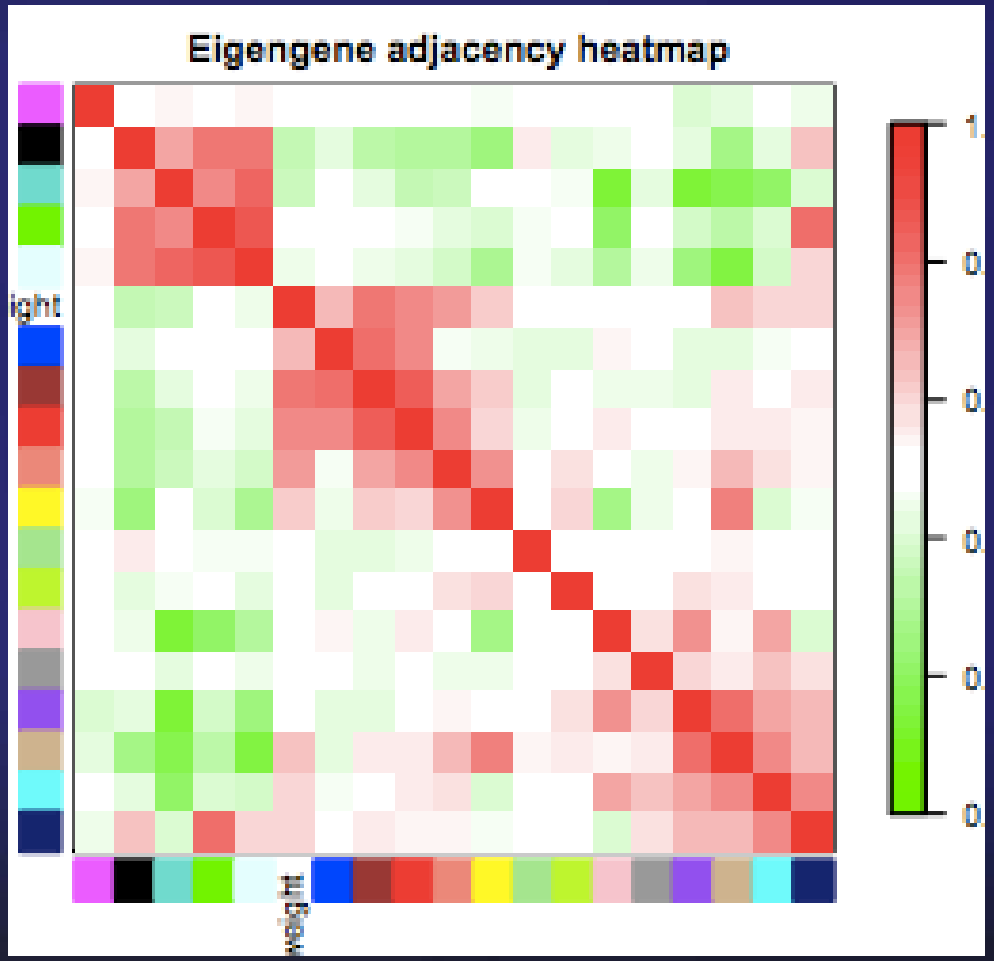
Bill likes Ted a lot

Ted dislikes Ann

Bill *really* likes

Bob despises Ken

…



Information visualization …

- General Aims
  - Use human perceptual capabilities
  - To gain insights into large and abstract data sets that are difficult to extract using standard query languages
- Exploratory Visualization
  - Look for structure, patterns, trends, anomalies, relationships
  - Provide a qualitative overview of large, complex data sets
  - Assist in identifying region(s) of interest and appropriate parameters for more focused quantitative analysis

Analysis and Visualization Tools

- Data Management and Applications
  - Database systems
  - Statistical packages: R
- Visualization Tools
  - Geographical Information Systems (GIS) - ESRI
  - Tableau, Spotfire…
- Toolkits
  - Web-native Javascript: D3, Protovis, OpenLayers
  - Python interfaces to DBs, R, Matplotlib…

Techniques

- Lots and Lots
  - And as many permutations as there are graduate students
- Few really general applications
  - Excel
- Lots of Toolkits
  - In particular, in Javascript for web applications

Good visualization

- Use of computer-supported, interactive, visual representations of abstract data to amplify cognition
  - Visual representation can enhance recognition
    - Recognition of patterns
    - Abstraction and aggregation
    - Perceptual interference
  - Facilitate data exploration
    - Interactive medium
    - High data density
    - Greater access speed
  - Increased analytic resources
    - Parallel perceptual processing
    - Offload work from cognitive to perceptual system

Fun Websites

- Atlas of Science (Katy Borner, IU)
  - http://scimaps.org/atlas/maps
- Many Eyes: a project to encourage sharing and conversation around visualizations (need java)
  - http://manyeyes.alphaworks.ibm.com/
- New York Times Infographics
  - http://www.smallmeans.com/new-york-times-infographics/
- Gap Minder http://www.gapminder.org
- How to visualize data with Chernoff face using R
  - http://flowingdata.com/2010/08/31/how-to-visualize-data-with-cartoonish-faces/

Reference materials

- References
  - E.R. Tufte, *The Visual Display of Quantitative Information,* Graphics Press, 1983.
  - S.K. Card, J.D. Mackinlay, and B. Shneiderman, *Information Visualization: Using Vision to Think, Morgan Kaufmann* Publishers, 1999.
- Software
  - Matplotlib/Python
  - Google charts/Javascript
  - InfoVis ToolKit
  - Prefuse
  - Titan Libraries/VTK InfoVis Libraries