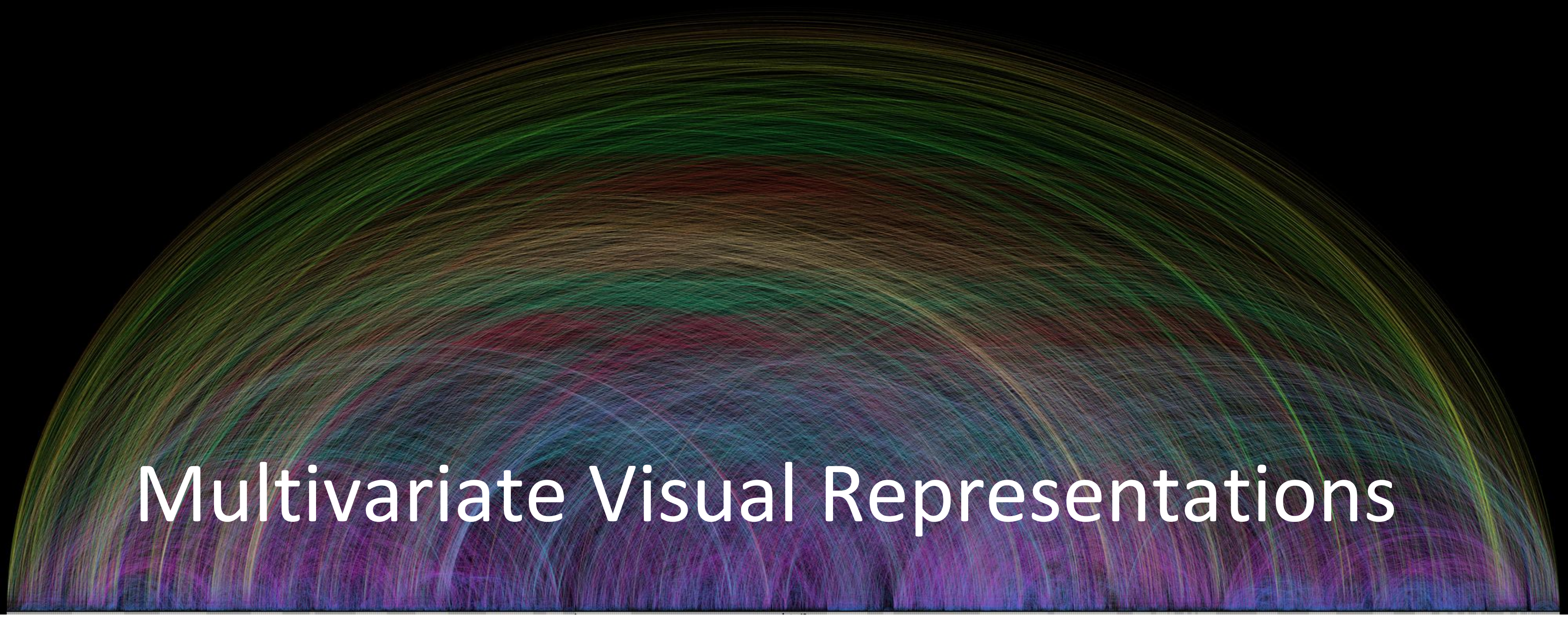


Multivariate Visual Representations

BAIS 6140 – Information Visualization

L. Miguel Encarnação

Agenda

- General representation techniques for multivariate (>3) variables per data case
  - But not lots of variables yet…

Multivariate data

Revisit

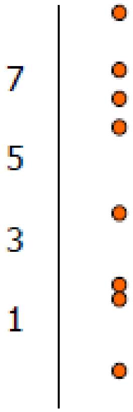
- Data sets of dimensions 1, 2, 3 are common
- Number of variables per class
  - 1 -Univariate data
  - 2 -Bivariate data
  - 3 -Trivariate data
  - >3 -Hypervariate data Focus Today

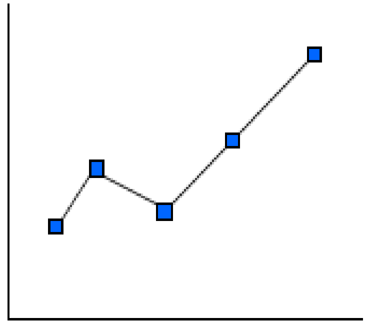
Previously …

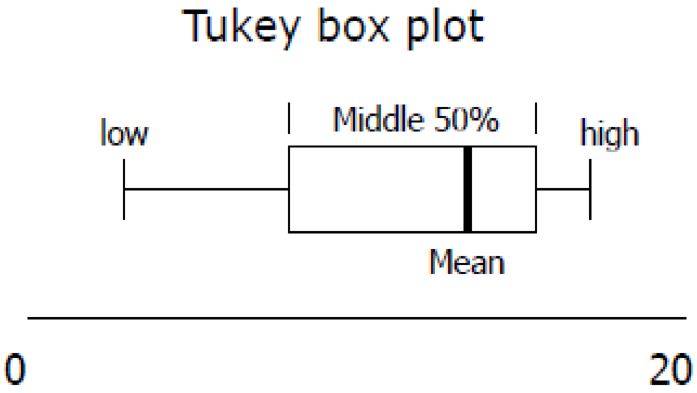
- We examined a number of tried-and-true techniques/visualizations for presenting multivariate (typically <=3) data sets
  - Hinted at how to go above 3 dimensions

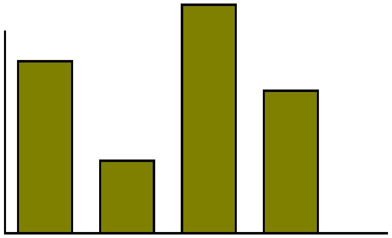
Representations

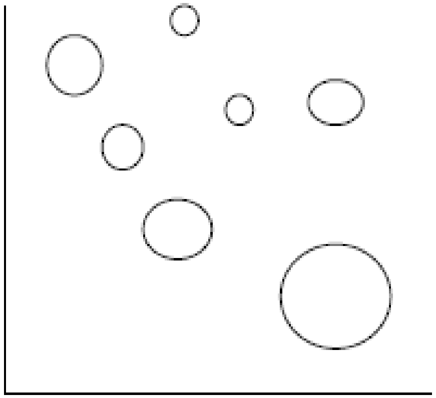
- Some standard ways for low-d data



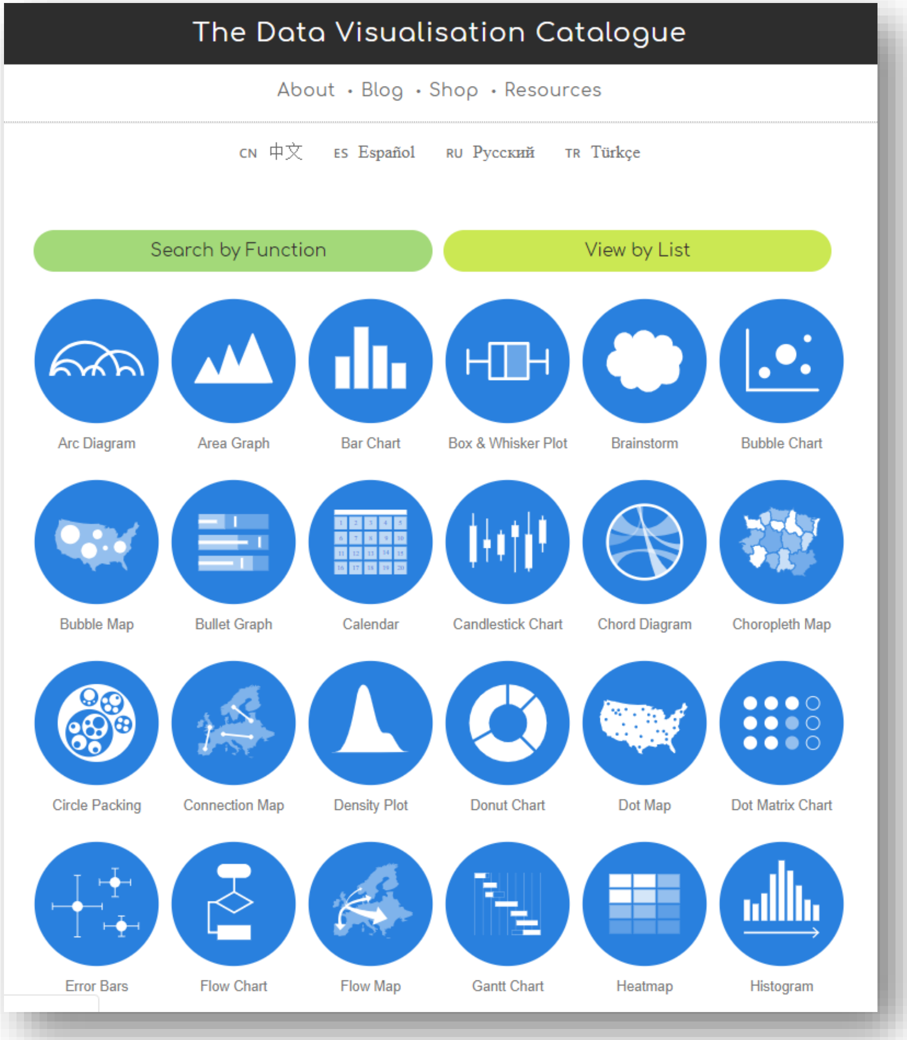








Data Visualization Catalogue



The Data Visualisation Catalogue

About · Blog · Shop · Resources

CN 中文 ES Español RU Русский TR Türkçe

https://datavizcatalogue.com/

- Most visualization techniques - discussed and compared!

Hypervariate Data

- What about 4 to 20 or so variables (for instance)?
  - Lower-dimensional hypervariate data
  - (Much higher dimensions next week)
  - Many data sets fall into this category

Too many dimensions to display …

- Fundamentally, we have 2 geometric (position) display dimensions
- For data sets with >2 variables, we must project data down to 2D
- Come up with visual mapping that locates each dimension into 2D plane
- Computer graphics: 3D->2D projections

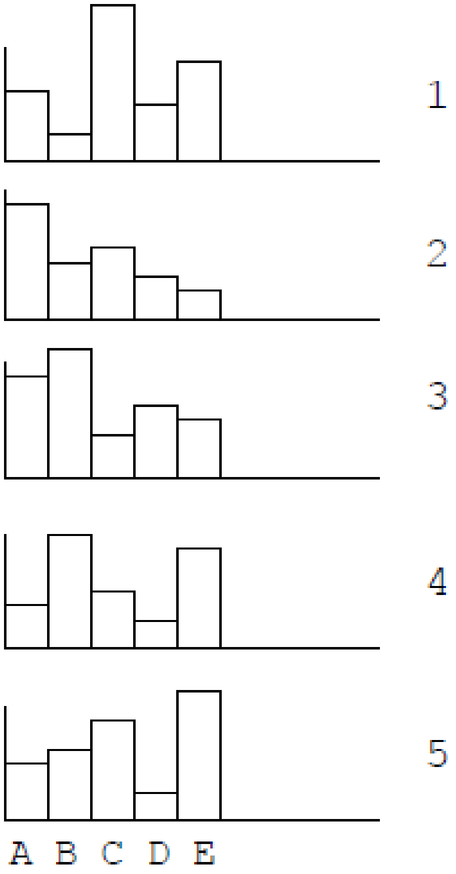
Actually …

- A spreadsheet already does that
  - Each variable is positioned into a column
  - Data cases in rows
  - This is a projection (mapping)
- What about some other techniques?
  - Already seen a couple

Revisit

- Give each variable its own display

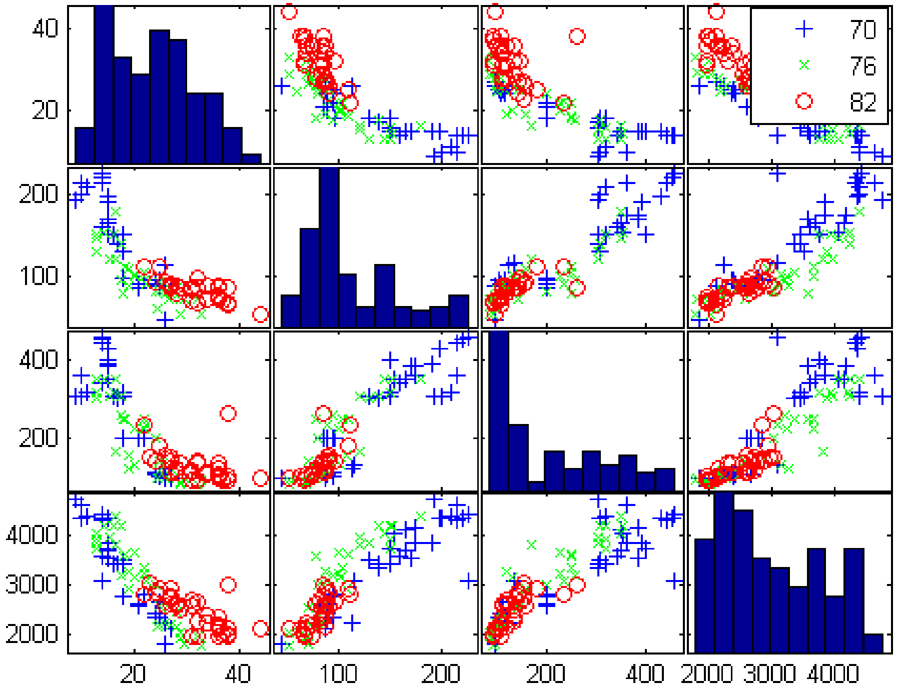
	A	B	C	D	E
1	4	1	8	3	5
2	6	3	4	2	1
3	5	7	2	4	3
4	2	6	3	1	5
5	3	4	5	1	7



Scatterplot Matrix

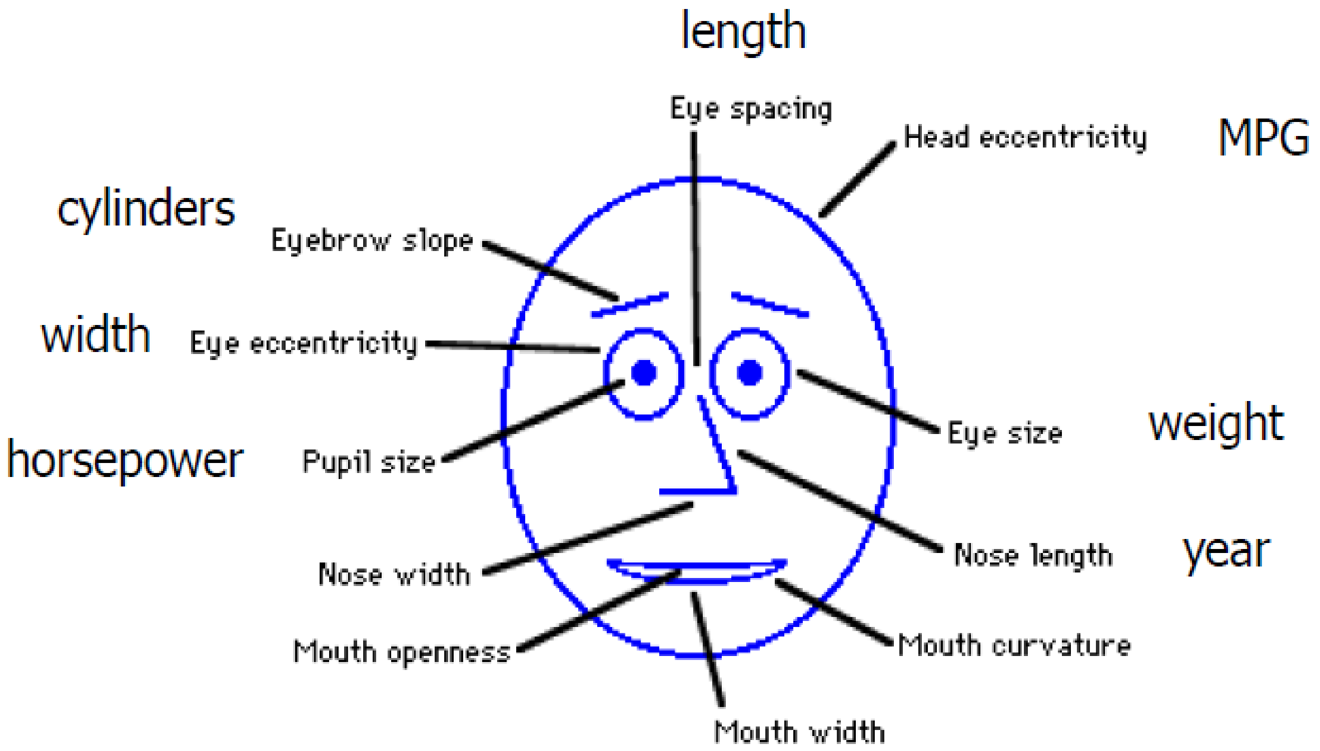
Revisit

- Represent each possible pair of variables in their own 2-D scatterplot

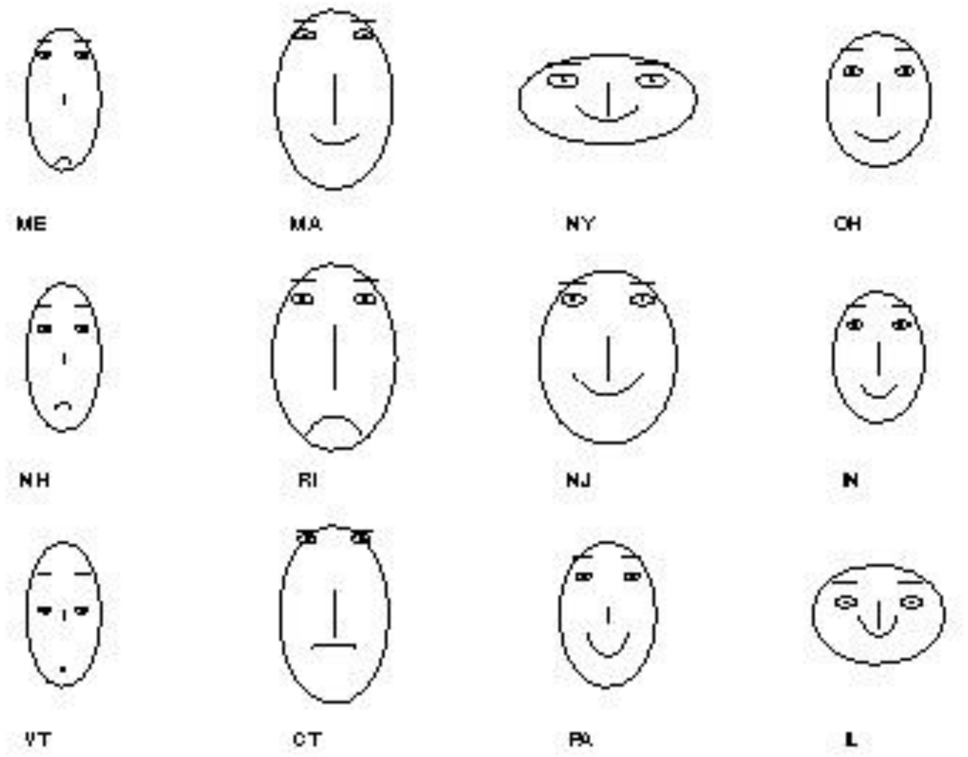


Chernoff Faces

- Encode different variables' values in characteristics of human face



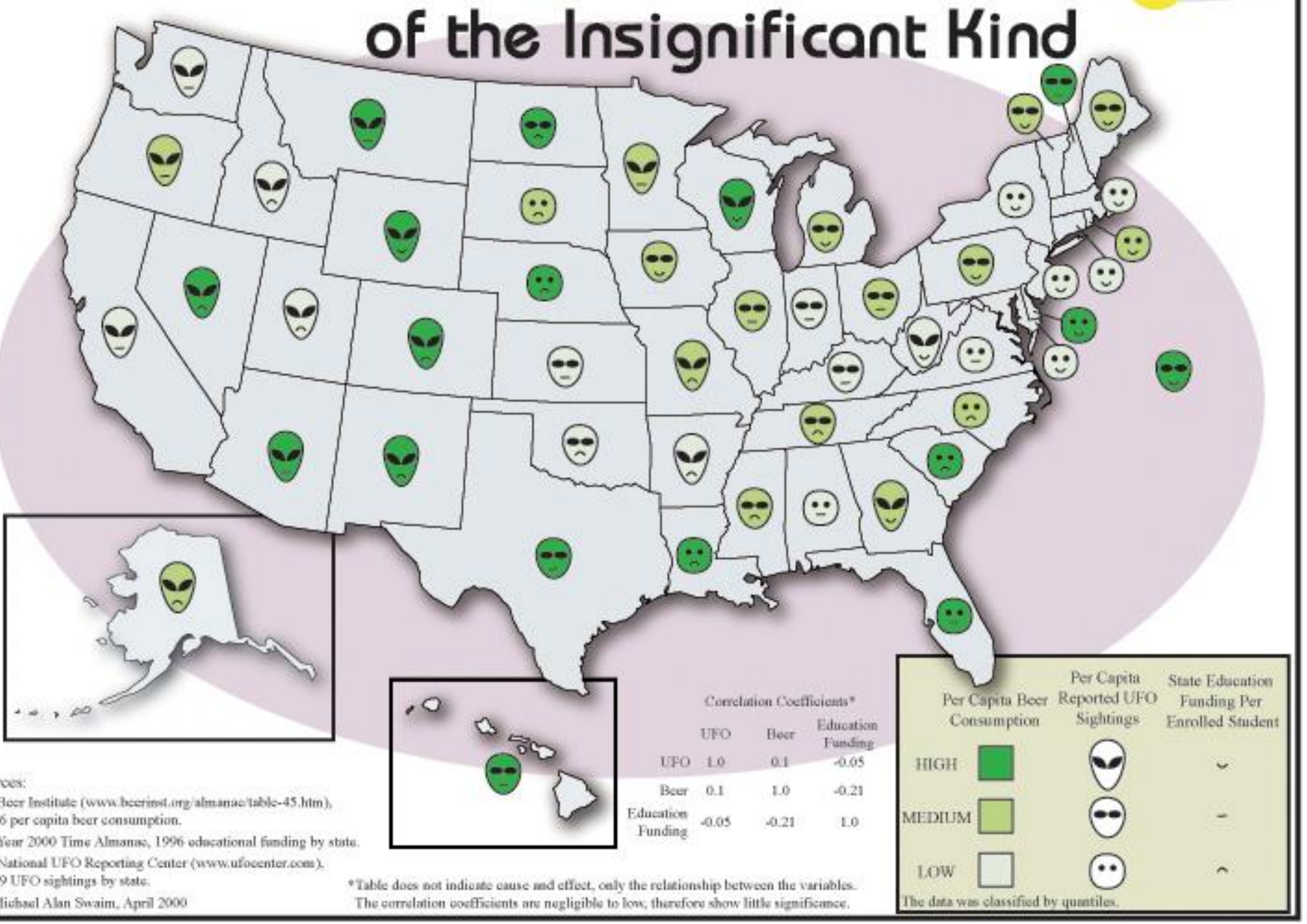
Examples



Critique:

https://eagereyes.org/criticism/chernoff-faces

Close Encounters of the Insignificant Kind


Per Capita Beer Consumption
Per Capita Reported UFO Sightings
State Education Funding Per Enrolled Student
HIGH
MEDIUM
LOW

The data was classified by quantiles.

Correlation Coefficients*

	UFO	Beer	Education Funding
UFO	1.0	0.1	-0.05
Beer	0.1	1.0	-0.21
Education Funding	-0.05	-0.21	1.0

*Table does not indicate cause and effect, only the relationship between the variables. The correlation coefficients are negligible to low, therefore show little significance.

Sources:

The Beer Institute (www.beerinst.org/almanac/table-45.htm), 1996 per capita beer consumption.

The Year 2000 Time Almanac, 1996 educational funding by state.

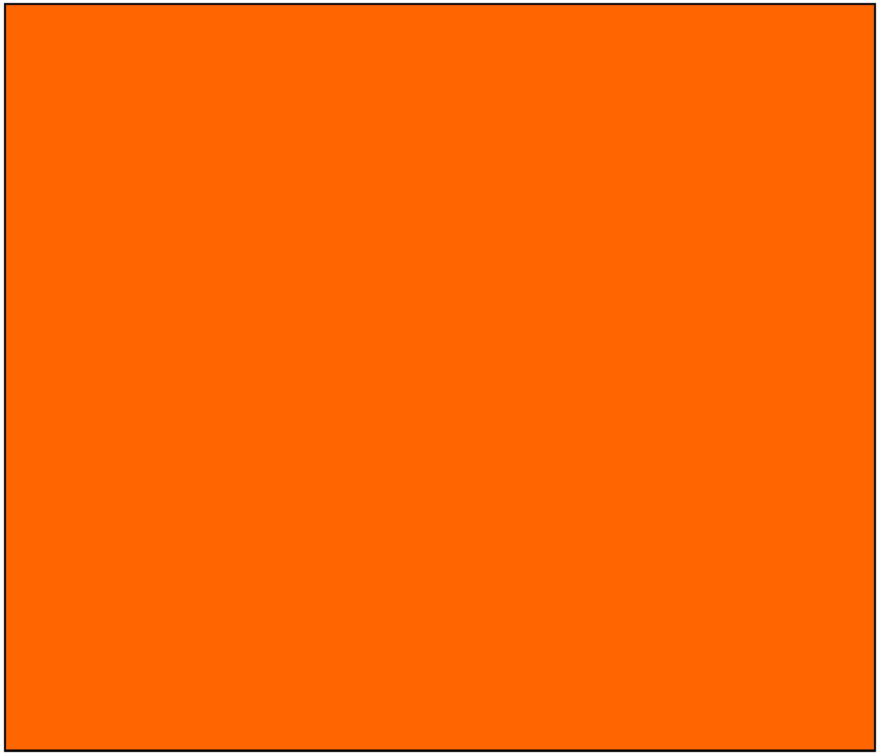
The National UFO Reporting Center (www.ufocenter.com), 1999 UFO sightings by state.

By Michael Alan Swaim, April 2000

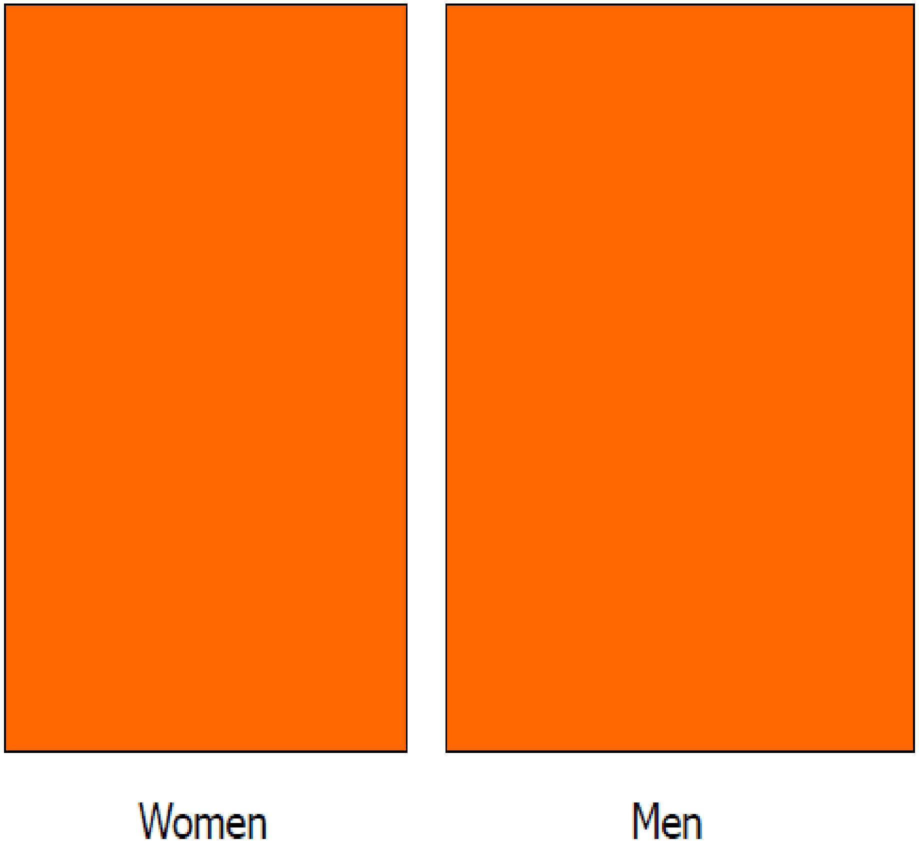
Categorical data

- How about multivariate categorical data?
- Students
  - Gender: Female, male
  - Eye color: Brown, blue, green, hazel
  - Hair color: Black, red, brown, blonde, gray
  - Home country: USA, China, Italy, India, …

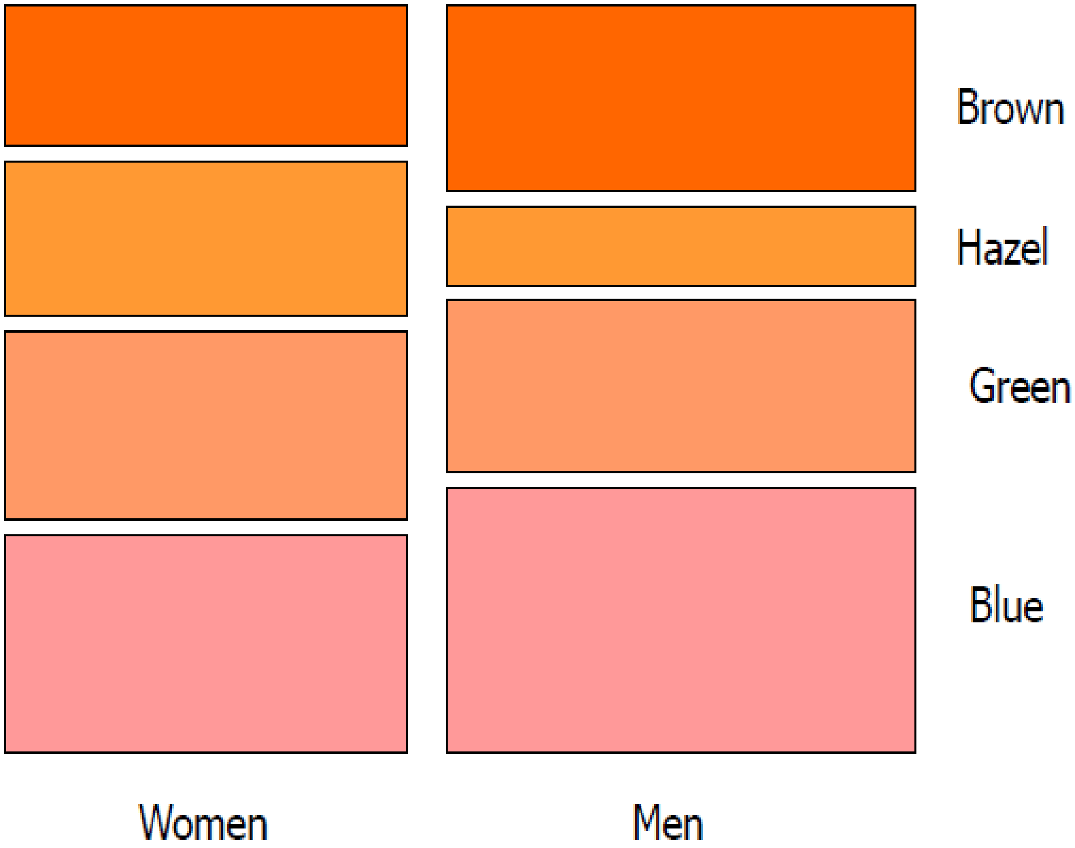
Mosaic Plot / Marimekko Diagram



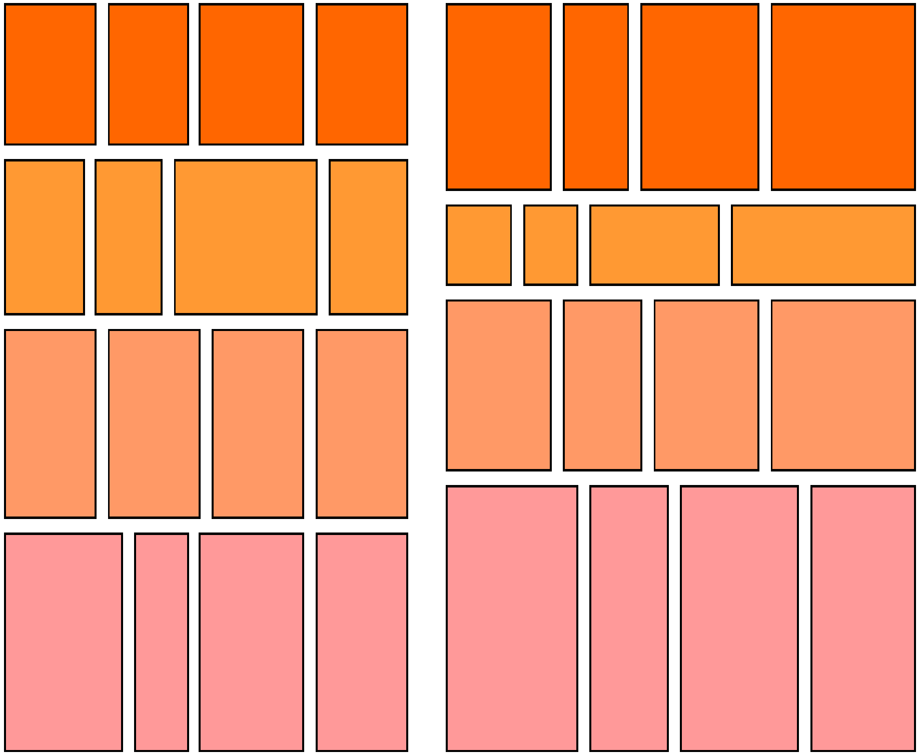
Mosaic Plot / Marimekko Diagram

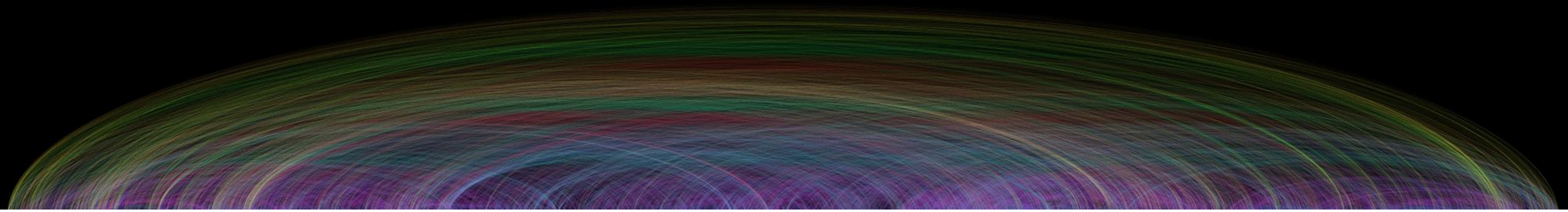


Mosaic Plot / Marimekko Diagram



Mosaic Plot / Marimekko Diagram



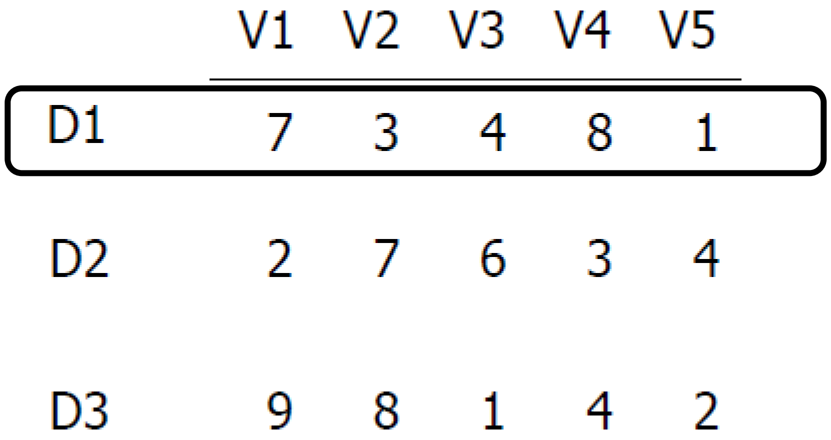


PARALLEL COORDINATES (AND MORE)

Previously

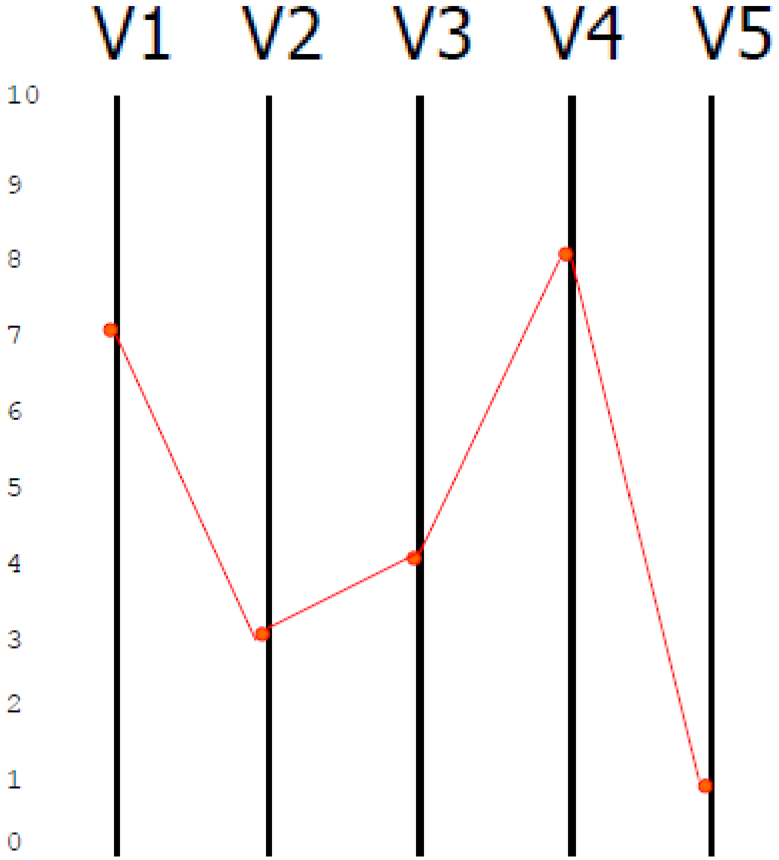
- Viewed a number of techniques for portraying low-dimensional data (about $3<x<20$)
  - scatterplot matrix
  - Table Lens
  - sliding rods
  - Attribute Explorer
  - Dust & Magnet
  - etc.

Parallel Coordinates



	V1	V2	V3	V4	V5
D1	7	3	4	8	1
D2	2	7	6	3	4
D3	9	8	1	4	2

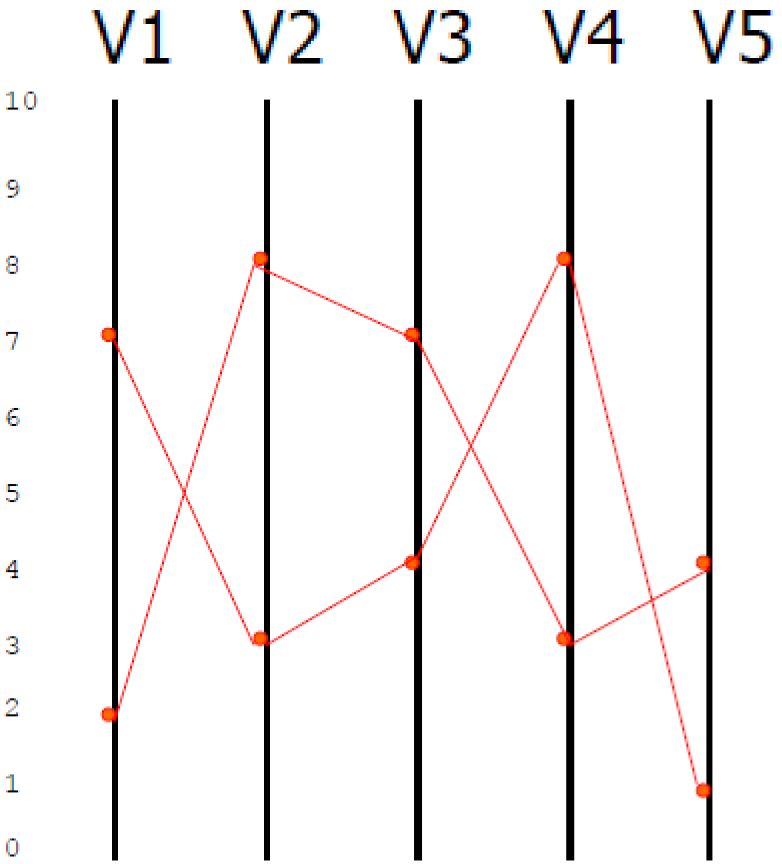
Parallel Coordinates



D1

7 3 4 8 1

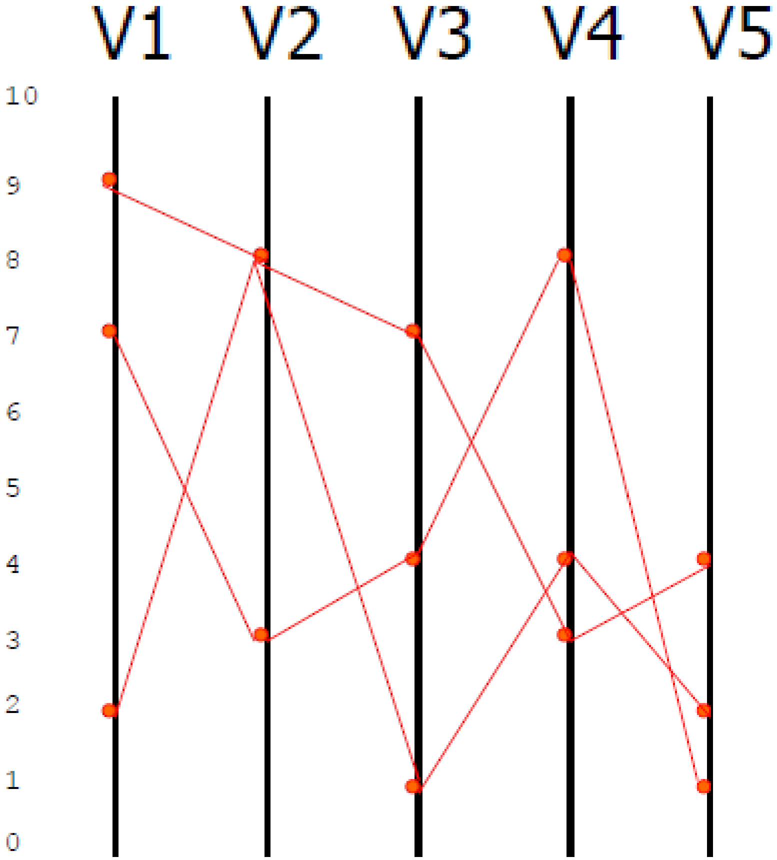
Parallel Coordinates



D2

2 7 6 3 4

Parallel Coordinates

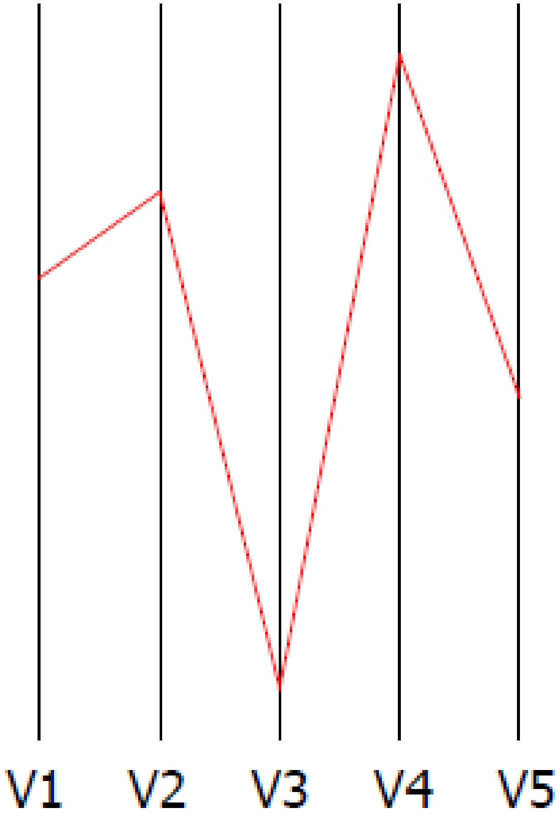


D3

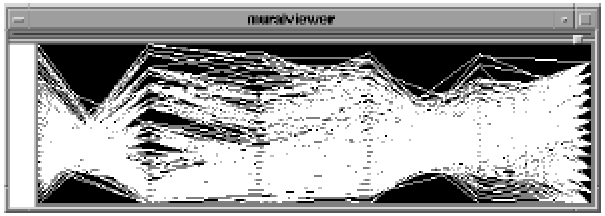
9 8 1 4 2

Parallel Coordinates

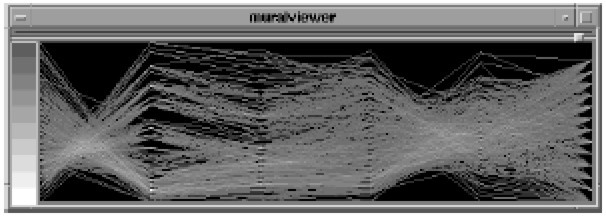
- Encodes variables along a horizontal row
- Vertical line specifies different values that variable can take
- Data point represented as a polyline



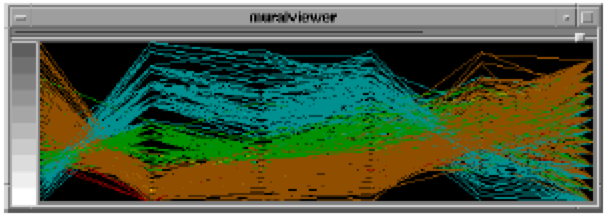
Parallel Coords Examples



Basic



Grayscale



Color

Issue

- Different variables can have values taking on quite different ranges
- Must normalize all down (e.g., 0..1)
- It’s about pattern recognition not precision

*

Example VLSI chip manufacture

- Application: System that uses parallel coordinates for information analysis and discovery
- Interactive tool
  - Can focus on certain data items
  - Color

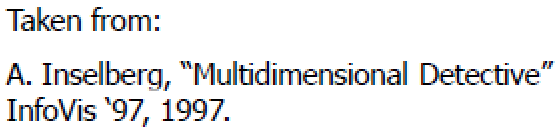
Taken from:

A. Inselberg, "Multidimensional Detective" InfoVis '97, 1997.

*

Example VLSI chip manufacture

- Want high quality chips (high speed) and a high yield batch (% of useful chips)
- Able to track defects
- Hypothesis: No defects gives desired chip types
- 473 batches of data


Taken from:

A. Inselberg, "Multidimensional Detective" InfoVis '97, 1997.

Example VLSI chip manufacture

- The data
  - 16 variables
    - X1 -yield
    - X2 -quality
    - X3-X12 -# defects (inverted)
    - X13-X16 -physical parameters

Example VLSI chip manufacture

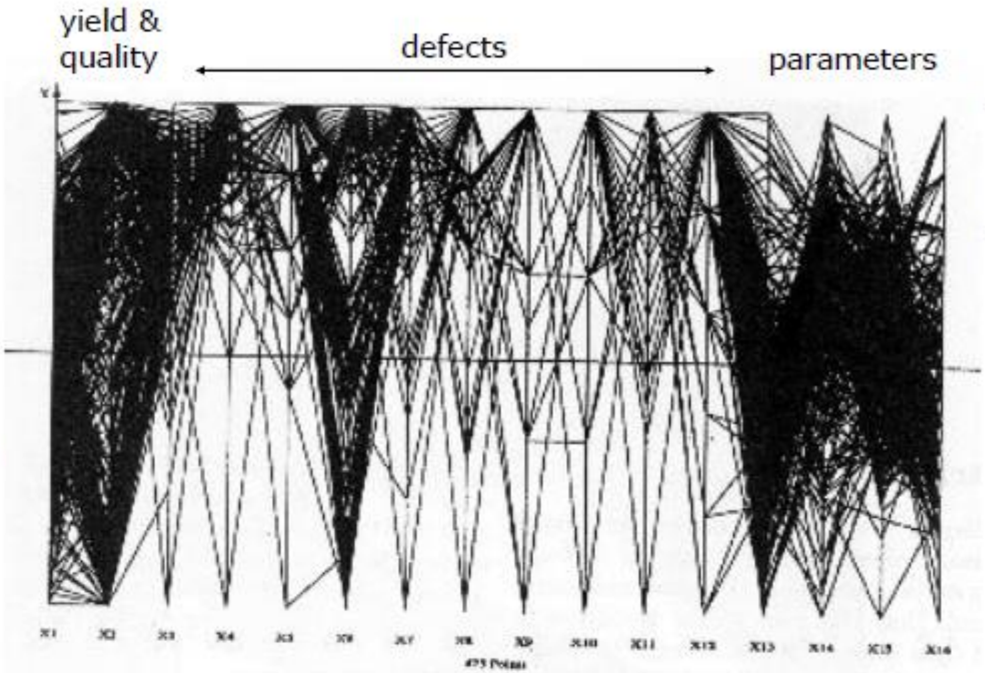


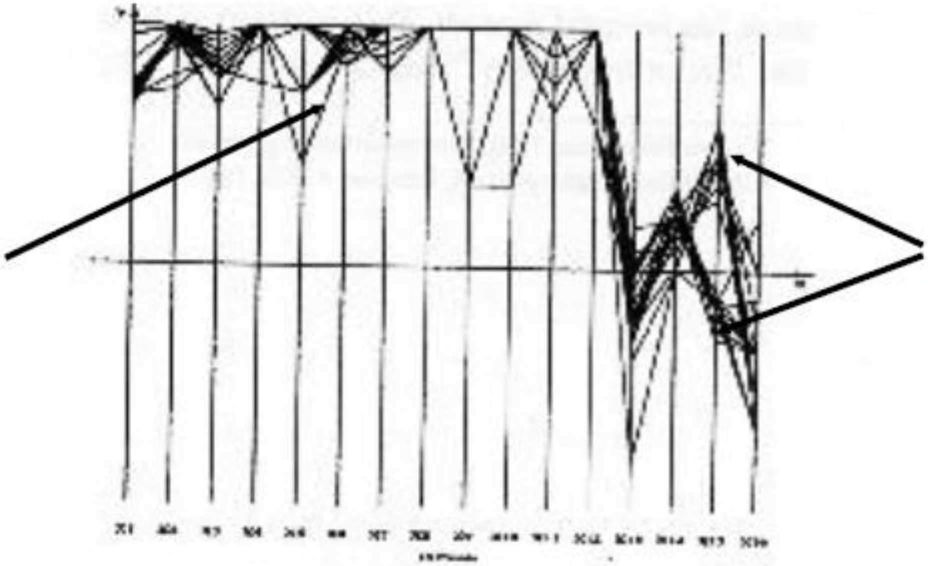
Figure 1: The full dataset consisting of 473 batches

Yikes!
But not
that bad

Distributions
x1 - normal
x2 - bipolar

Example VLSI chip manufacture

- Top Yield and Quality



defects

split

Figure 2: The batches high in Yield, $X1$, and Quality, $X2$.

Have some defects

*

Example VLSI chip manufacture

- Minimal defects

Not the highest yields and quality

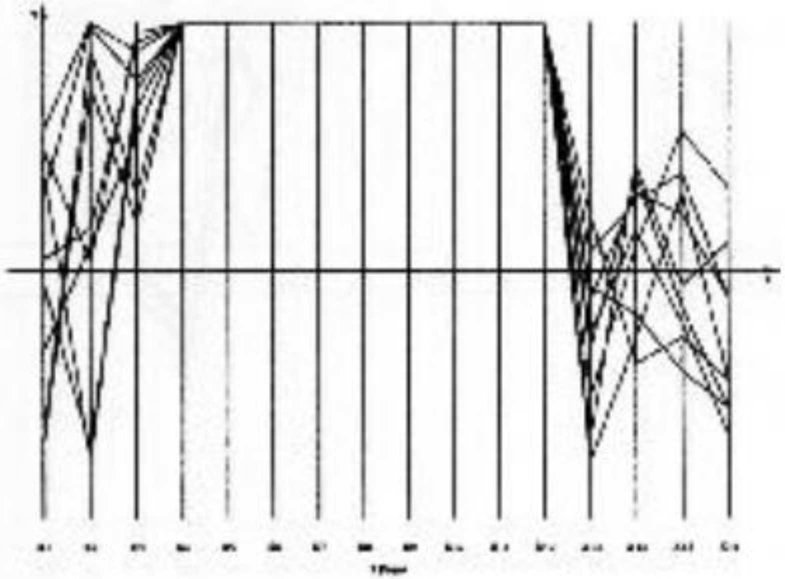


Figure 3: The batches with zero in 9 out of the ten defect types.

Example VLSI chip manufacture

- Best yields

Appears that some defects are necessary to produce the best chips

Non-intuitive!

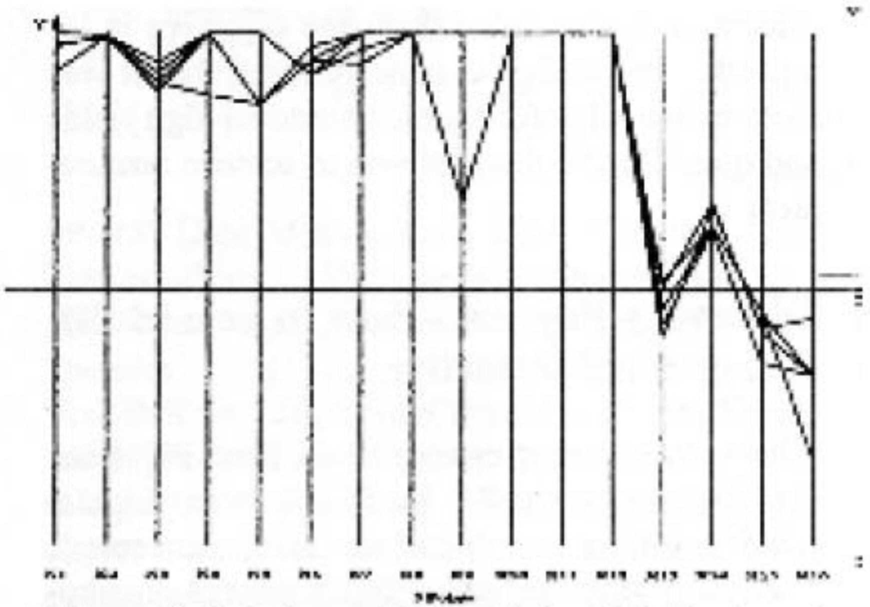
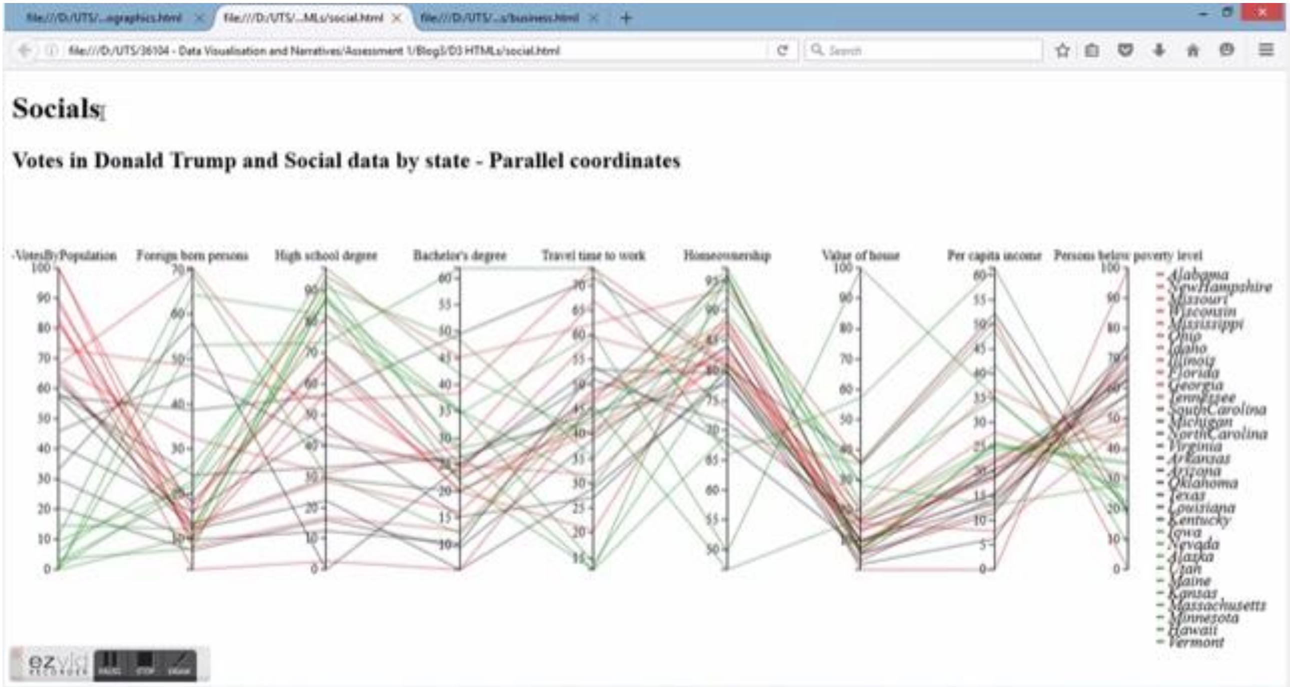


Figure 6: Batches with the highest Yields do not have the lowest defects in $X3$ and $X6$.

D3 (Video)



file:///D:/UTS/...ographics.html
file:///D:/UTS/...MLs/social.html
file:///D:/UTS/...s/business.html
file:///D:/UTS/36104 - Data Visualisation and Narratives/Assessment 1/Blog3/D3 HTMLs/social.html
Search

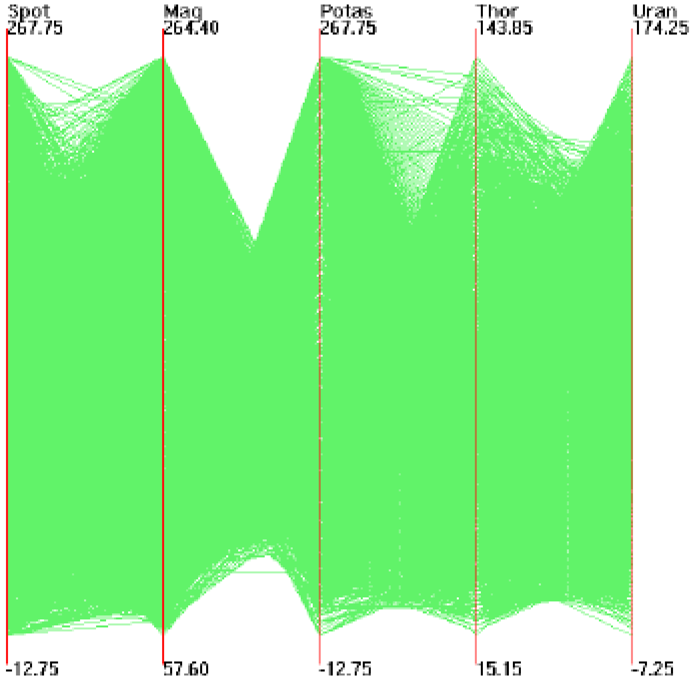
Socials

Votes in Donald Trump and Social data by state - Parallel coordinates

https://www.youtube.com/watch?v=qrij1-d4RYk

Challenges

Too much data



Out5d dataset (5 dimensions, 16384 data items)

Dimensional Reordering

Can you reduce clutter and highlight other interesting features in data by changing order of dimensions?

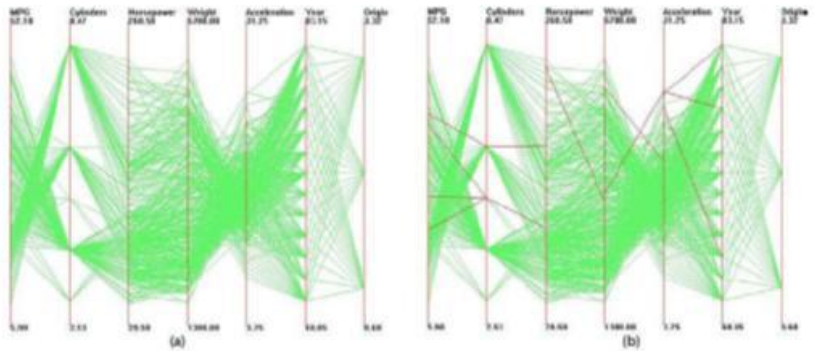


Figure 1: Parallel coordinates visualization of Cars dataset. Outliers are highlighted with red in (b).

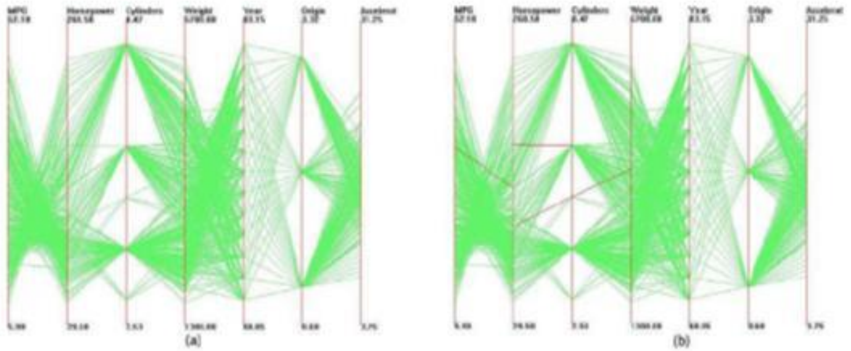


Figure 2: Parallel coordinates visualization of Cars dataset after clutter-based dimension reordering. Outliers are highlighted with red in (b).

Peng et al
InfoVis '04

Reducing Density

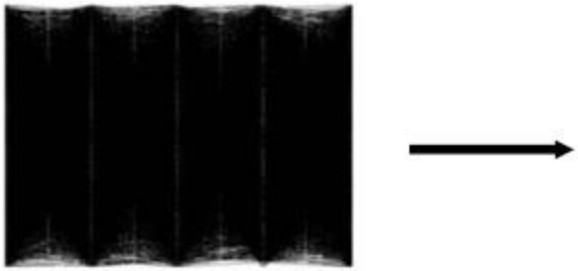
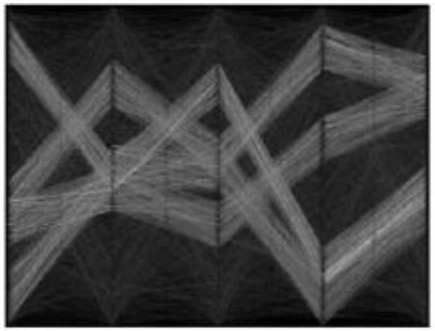
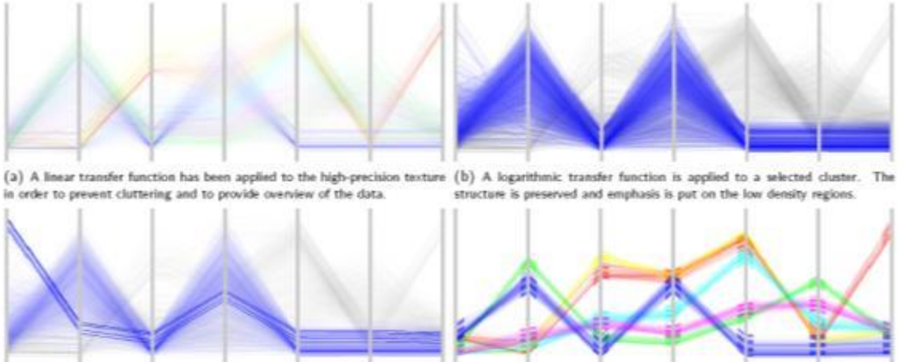


Figure 1 – Parallel Coordinates visualization of the *Sint1* data set (7,500 five-attribute records).



(a)

Artero et al, 04



(a) A linear transfer function has been applied to the high-precision texture in order to prevent cluttering and to provide overview of the data.

(b) A logarithmic transfer function is applied to a selected cluster. The structure is preserved and emphasis is put on the low density regions.

(c) Local cluster outliers are enhanced. A square root transfer function is used and the outliers are visible even through high-density regions.

(d) A complementary view of the clusters with uniform bands. 'Feature animation' presents statistics about the clusters and acts as a guidance.

Johansson et al, '05

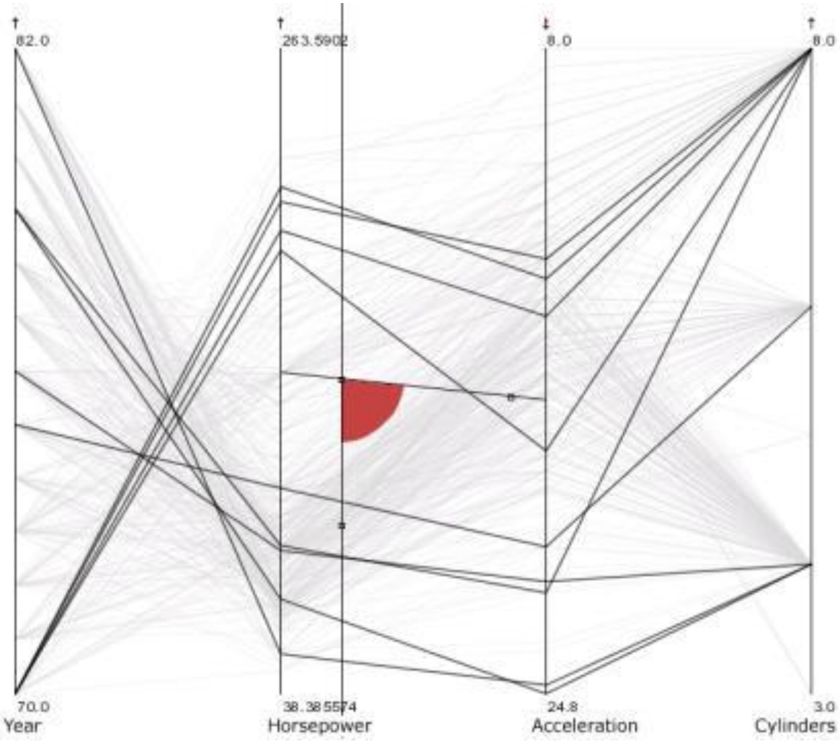
Jerding and Stasko, '95, '98
Wegman & Luo, '96

Improved Interaction

- How do we let the user select items of interest?
- Obvious notion of clicking on one of the polylines, but how about something more than that?

Attribute Ratios

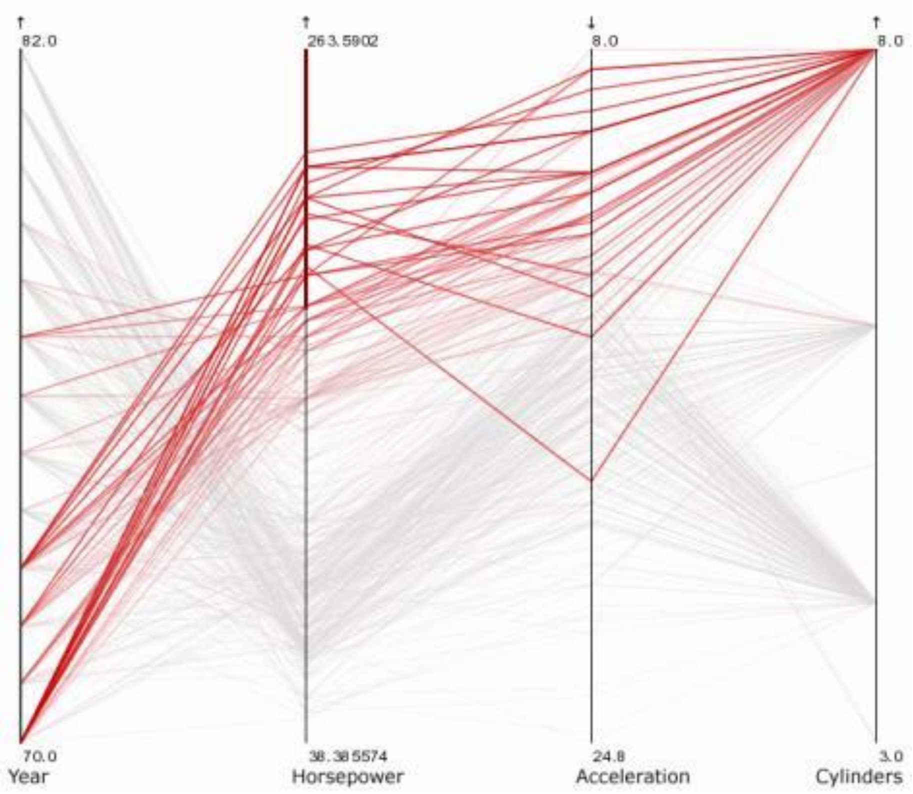
- Angular Brushing
  - Select subsets which exhibit a correlation along 2 axes by specifying angle of interest



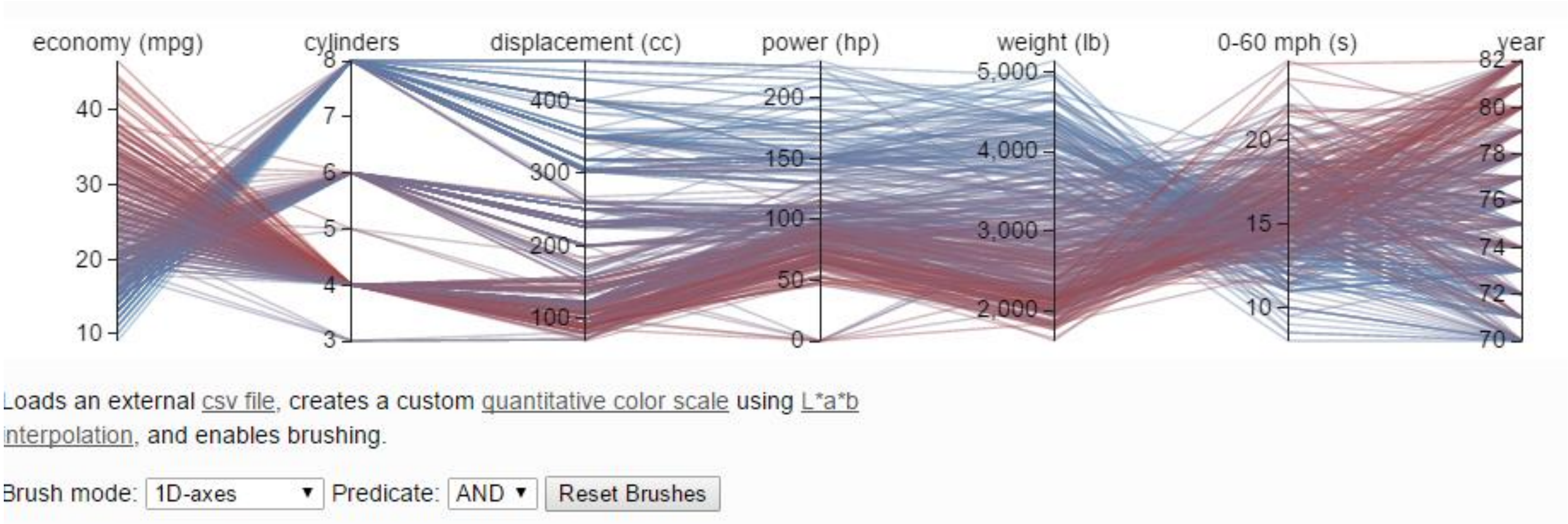
Hauser, Ledermann & Doleisch
InfoVis '02

Range Focus

- Smooth Brushing
  - Specify a region of interest along one axis



“Multidimensional Detective” on GitHub



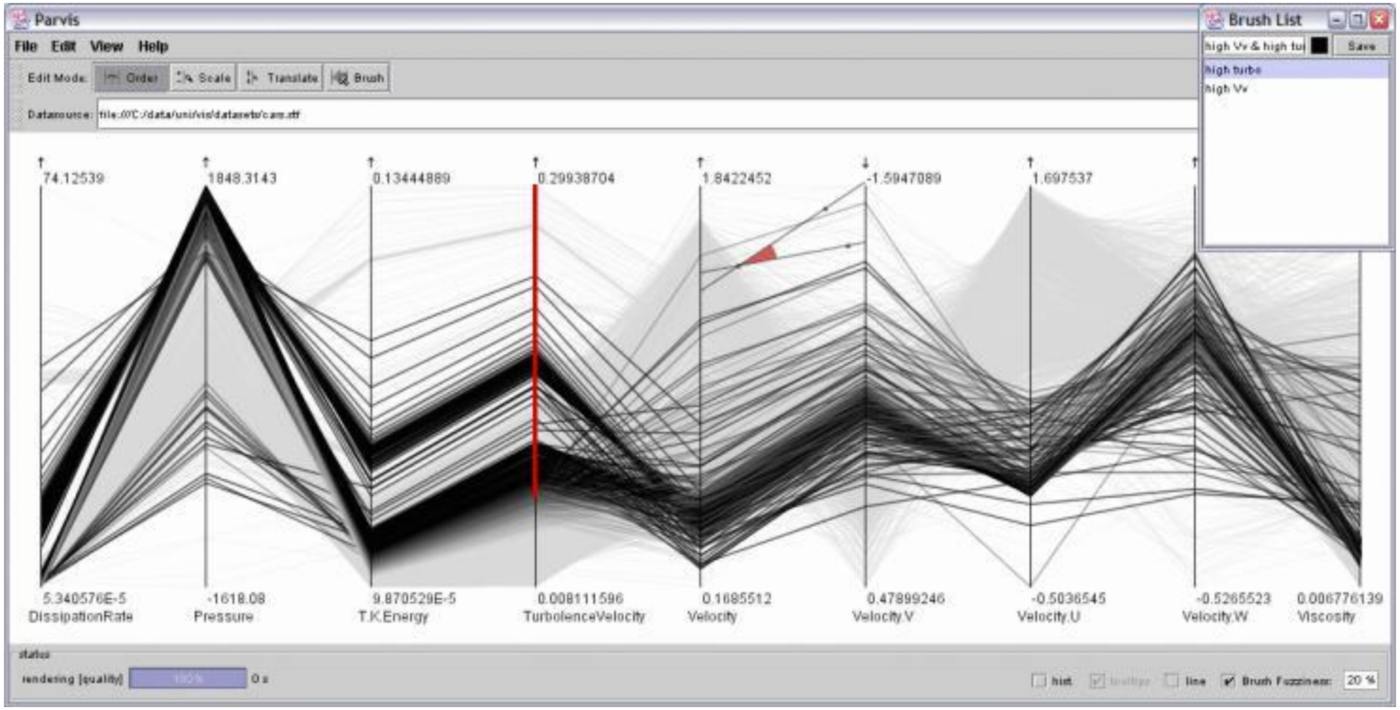
Loads an external csv file, creates a custom quantitative color scale using L*a*b interpolation, and enables brushing.

Brush mode: 1D-axes ▾ Predicate: AND ▾

https://syntagmatic.github.io/parallelcoordinates/examples/brushing.html

Composite Brushing

- Combine brushes and DOI functions using logical operators



Datasource: file:///C:/data/uni/vis/datasets/cars.stf

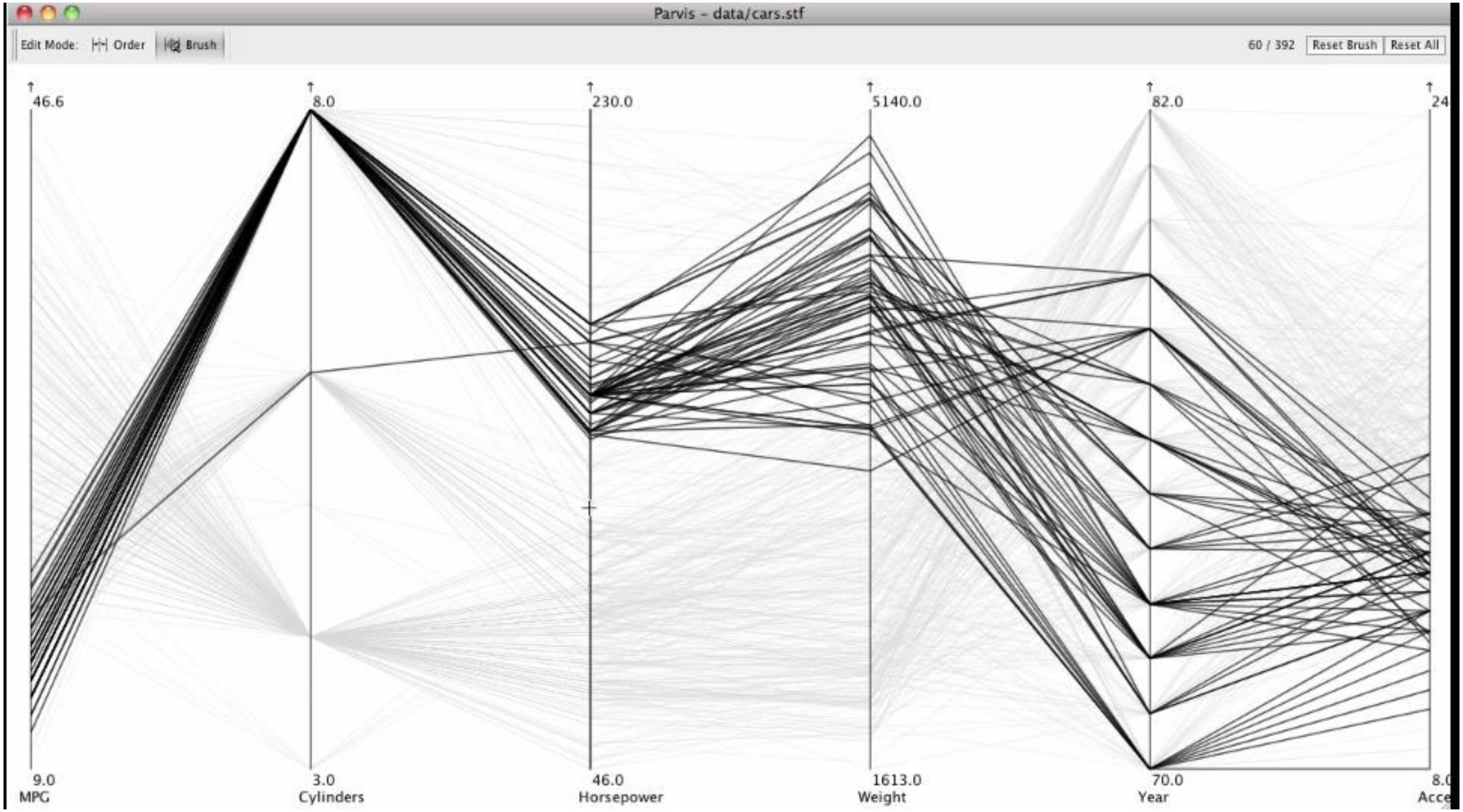
high Vv & high tu

high turbo

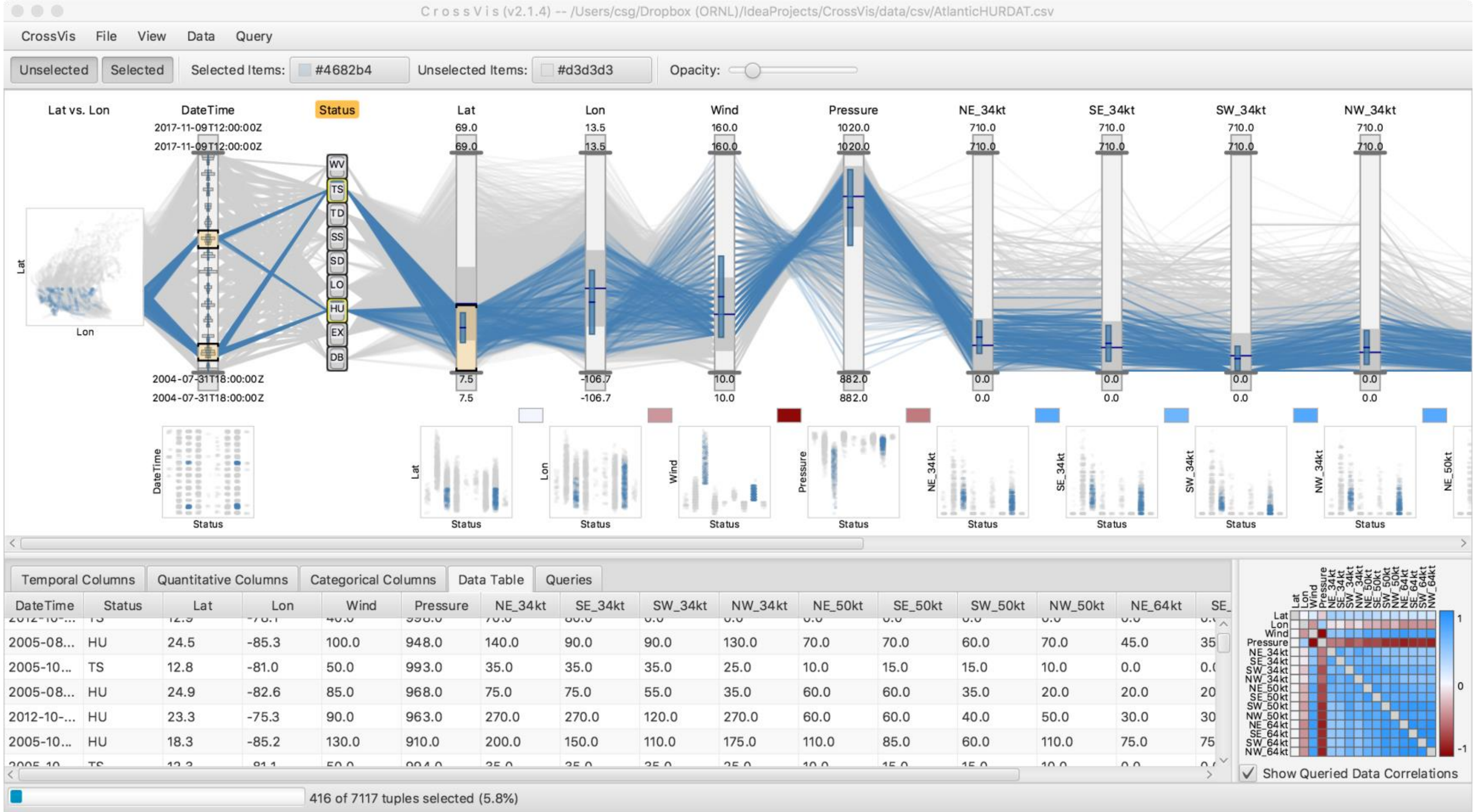
high Vv

☐ hist ☑ tooltips ☐ line ☑ Brush Fuzziness: 20 %

Video


Parvis – data/cars.stf
Edit Mode:
Order
Brush
60 / 392
Reset Brush
Reset All
46.6
8.0
230.0
5140.0
82.0
24
9.0
MPG
3.0
Cylinders
46.0
Horsepower
1613.0
Weight
70.0
Year
8.0
Acce

Lots of advanced parallel coordinate applications

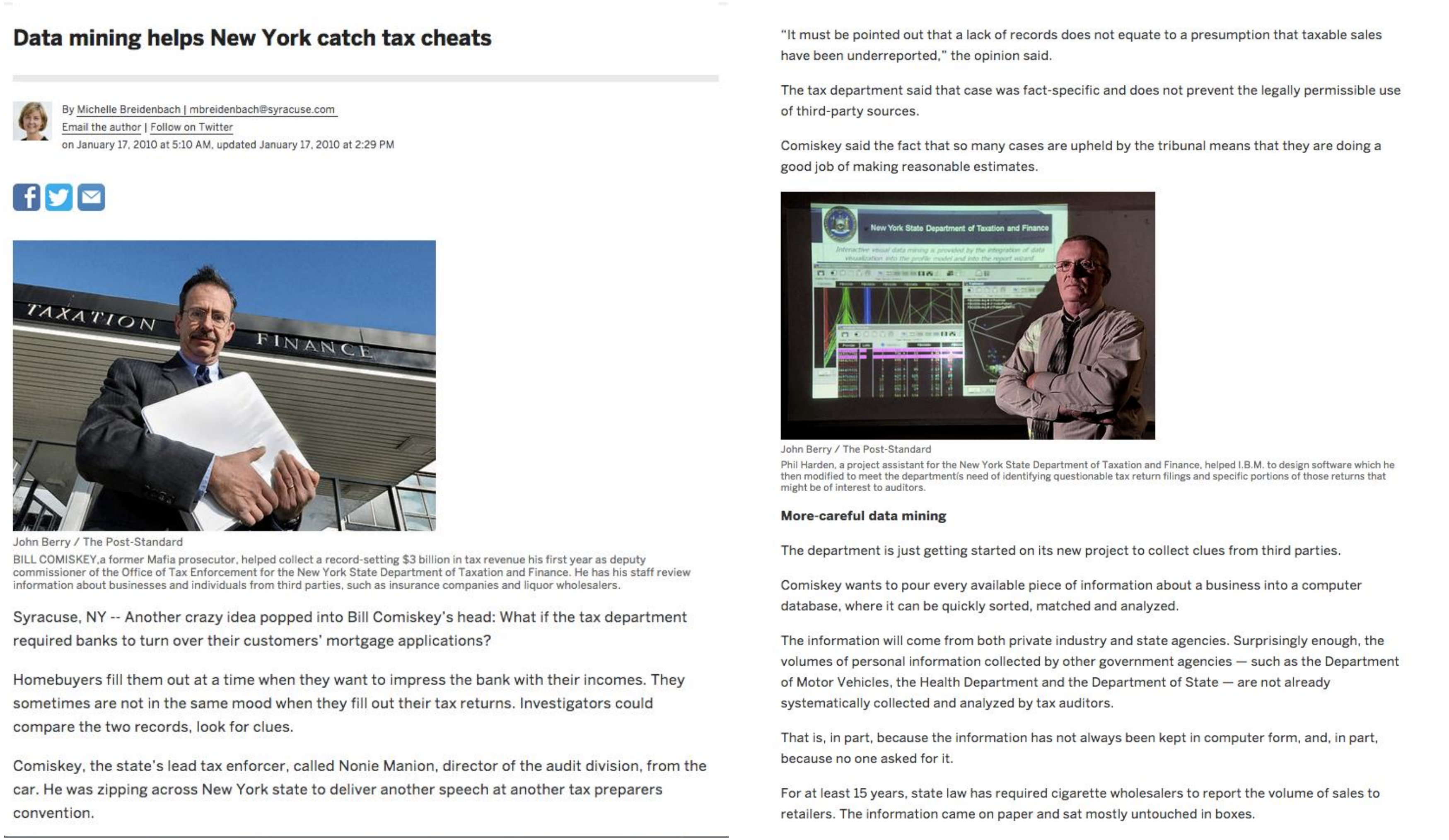


DateTime	Status	Lat	Lon	Wind	Pressure	NE_34kt	SE_34kt	SW_34kt	NW_34kt	NE_50kt	SE_50kt	SW_50kt	NW_50kt	NE_64kt	SE_
2005-08...	HU	24.5	-85.3	100.0	948.0	140.0	90.0	90.0	130.0	70.0	70.0	60.0	70.0	45.0	35
2005-10...	TS	12.8	-81.0	50.0	993.0	35.0	35.0	35.0	25.0	10.0	15.0	15.0	10.0	0.0	0.(
2005-08...	HU	24.9	-82.6	85.0	968.0	75.0	75.0	55.0	35.0	60.0	60.0	35.0	20.0	20.0	20
2012-10-...	HU	23.3	-75.3	90.0	963.0	270.0	270.0	120.0	270.0	60.0	60.0	40.0	50.0	30.0	30
2005-10...	HU	18.3	-85.2	130.0	910.0	200.0	150.0	110.0	175.0	110.0	85.0	60.0	110.0	75.0	75

416 of 7117 tuples selected (5.8%)

http://csteed.com/projects.html

What about categorical data?



Data mining helps New York catch tax cheats

By Michelle Breidenbach | mbreidenbach@syracuse.com
Email the author | Follow on Twitter
on January 17, 2010 at 5:10 AM, updated January 17, 2010 at 2:29 PM

John Berry / The Post-Standard
BILL COMISKEY,a former Mafia prosecutor, helped collect a record-setting $3 billion in tax revenue his first year as deputy commissioner of the Office of Tax Enforcement for the New York State Department of Taxation and Finance. He has his staff review information about businesses and individuals from third parties, such as insurance companies and liquor wholesalers.

Syracuse, NY -- Another crazy idea popped into Bill Comiskey's head: What if the tax department required banks to turn over their customers' mortgage applications?

Homebuyers fill them out at a time when they want to impress the bank with their incomes. They sometimes are not in the same mood when they fill out their tax returns. Investigators could compare the two records, look for clues.

Comiskey, the state's lead tax enforcer, called Nonie Manion, director of the audit division, from the car. He was zipping across New York state to deliver another speech at another tax preparers convention.

"It must be pointed out that a lack of records does not equate to a presumption that taxable sales have been underreported," the opinion said.

The tax department said that case was fact-specific and does not prevent the legally permissible use of third-party sources.

Comiskey said the fact that so many cases are upheld by the tribunal means that they are doing a good job of making reasonable estimates.

John Berry / The Post-Standard
Phil Harden, a project assistant for the New York State Department of Taxation and Finance, helped I.B.M. to design software which he then modified to meet the departmentís need of identifying questionable tax return filings and specific portions of those returns that might be of interest to auditors.

More-careful data mining

The department is just getting started on its new project to collect clues from third parties.

Comiskey wants to pour every available piece of information about a business into a computer database, where it can be quickly sorted, matched and analyzed.

The information will come from both private industry and state agencies. Surprisingly enough, the volumes of personal information collected by other government agencies — such as the Department of Motor Vehicles, the Health Department and the Department of State — are not already systematically collected and analyzed by tax auditors.

That is, in part, because the information has not always been kept in computer form, and, in part, because no one asked for it.

For at least 15 years, state law has required cigarette wholesalers to report the volume of sales to retailers. The information came on paper and sat mostly untouched in boxes.

http://www.syracuse.com/news/index.ssf/2010/01/data_mining_helps_new_york_cat.html

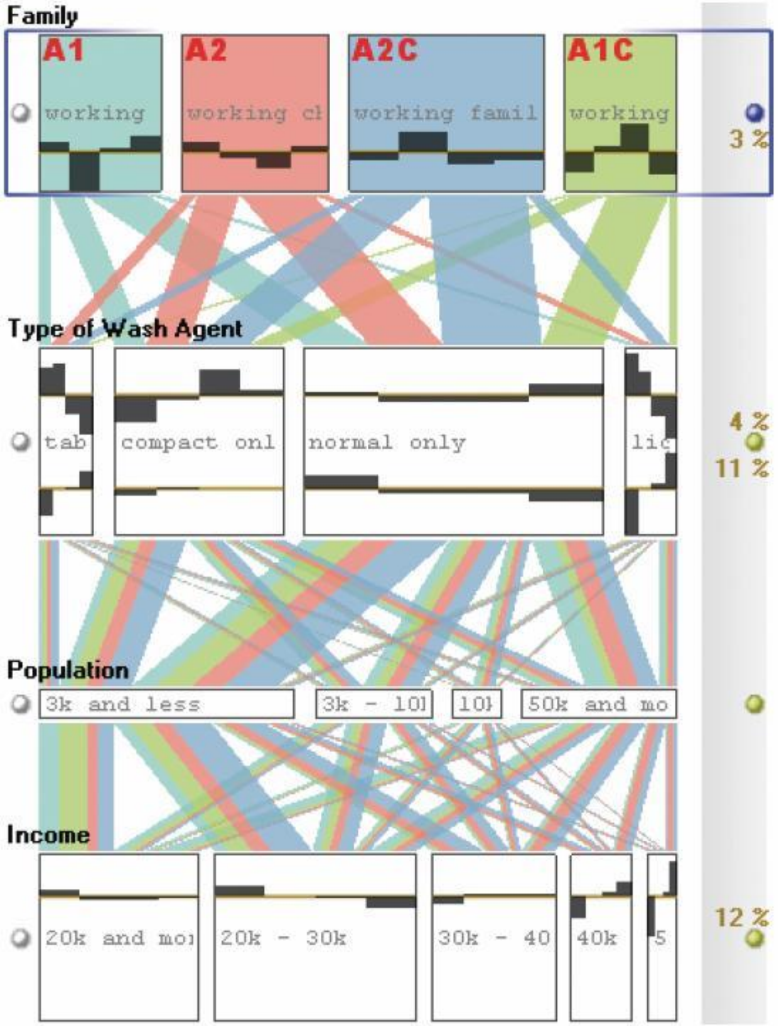
Parallel Sets

- Visualization method adopting parallel coordinates layout but uses frequency-based representation
- Visual metaphor
  - Layout similar to parallel coordinates
  - Continuous axes replaced with boxes
- Interaction
  - User-driven: User can create new classifications

Kosara, Bendix, & Hauser
TVCG ‘05

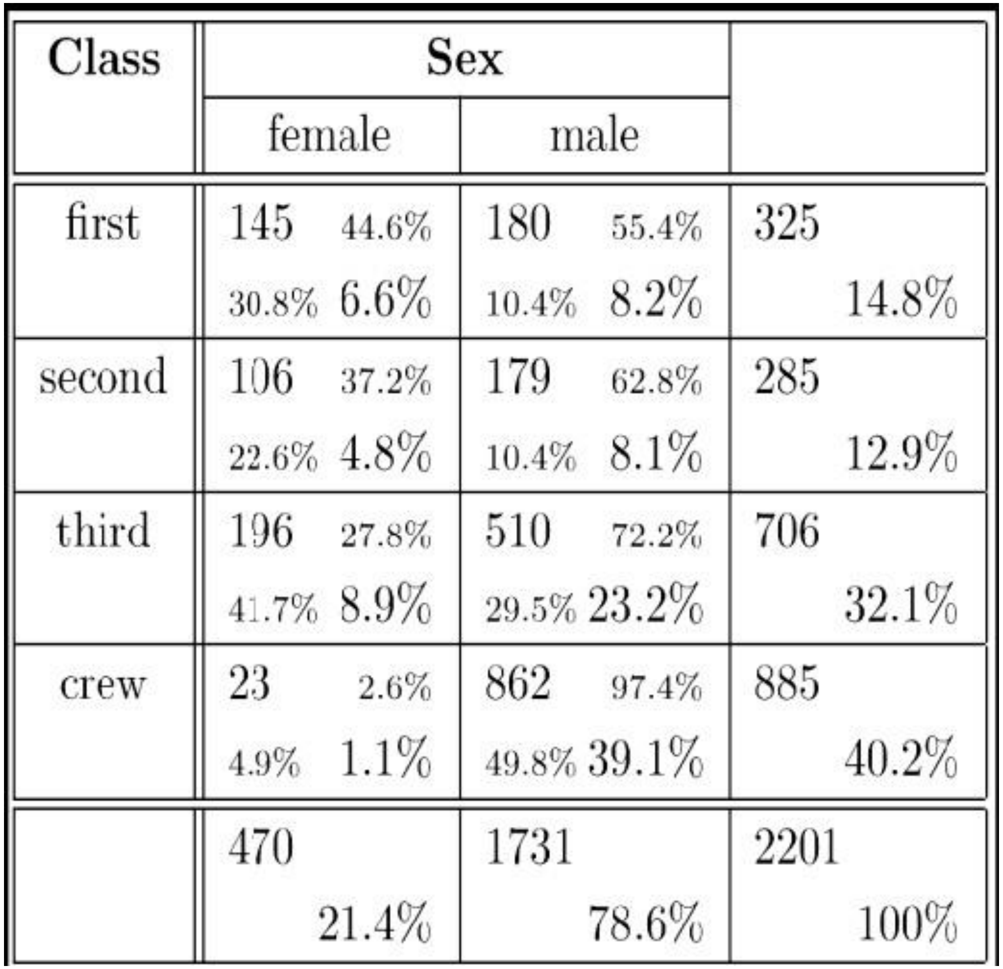
Representations

- Color used for different categories
- Those values flow into other variables



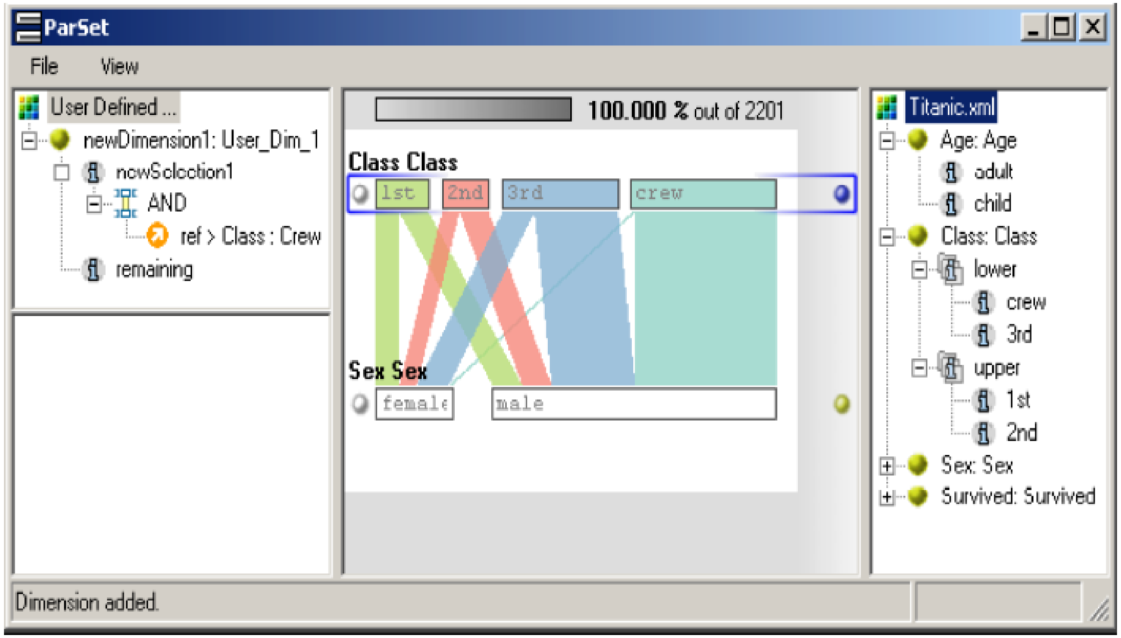
Example

- Titanic passengers set



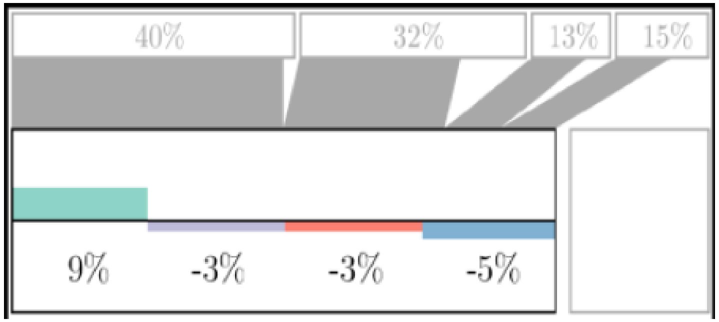
Class	Sex		
	female	male	
first	145 44.6% 30.8% 6.6%	180 55.4% 10.4% 8.2%	325 14.8%
second	106 37.2% 22.6% 4.8%	179 62.8% 10.4% 8.1%	285 12.9%
third	196 27.8% 41.7% 8.9%	510 72.2% 29.5% 23.2%	706 32.1%
crew	23 2.6% 4.9% 1.1%	862 97.4% 49.8% 39.1%	885 40.2%
	470 21.4%	1731 78.6%	2201 100%

Titanic Data Set



100.000 % out of 2201

Dimension added.



Interactions

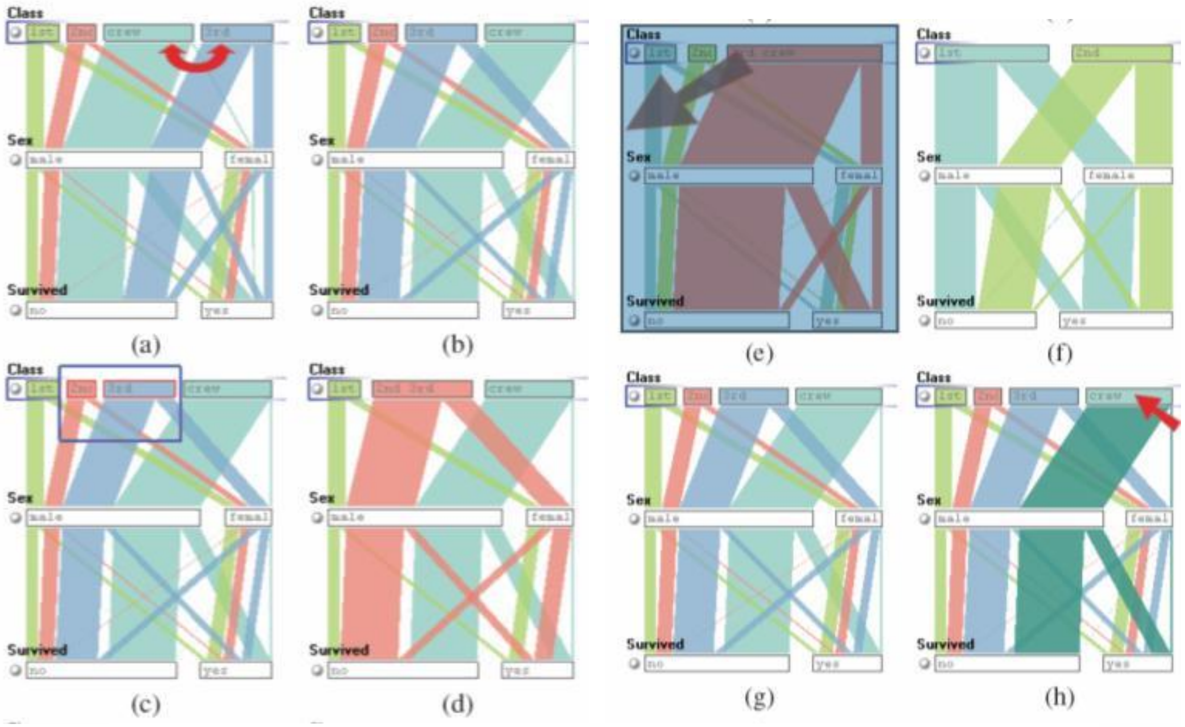
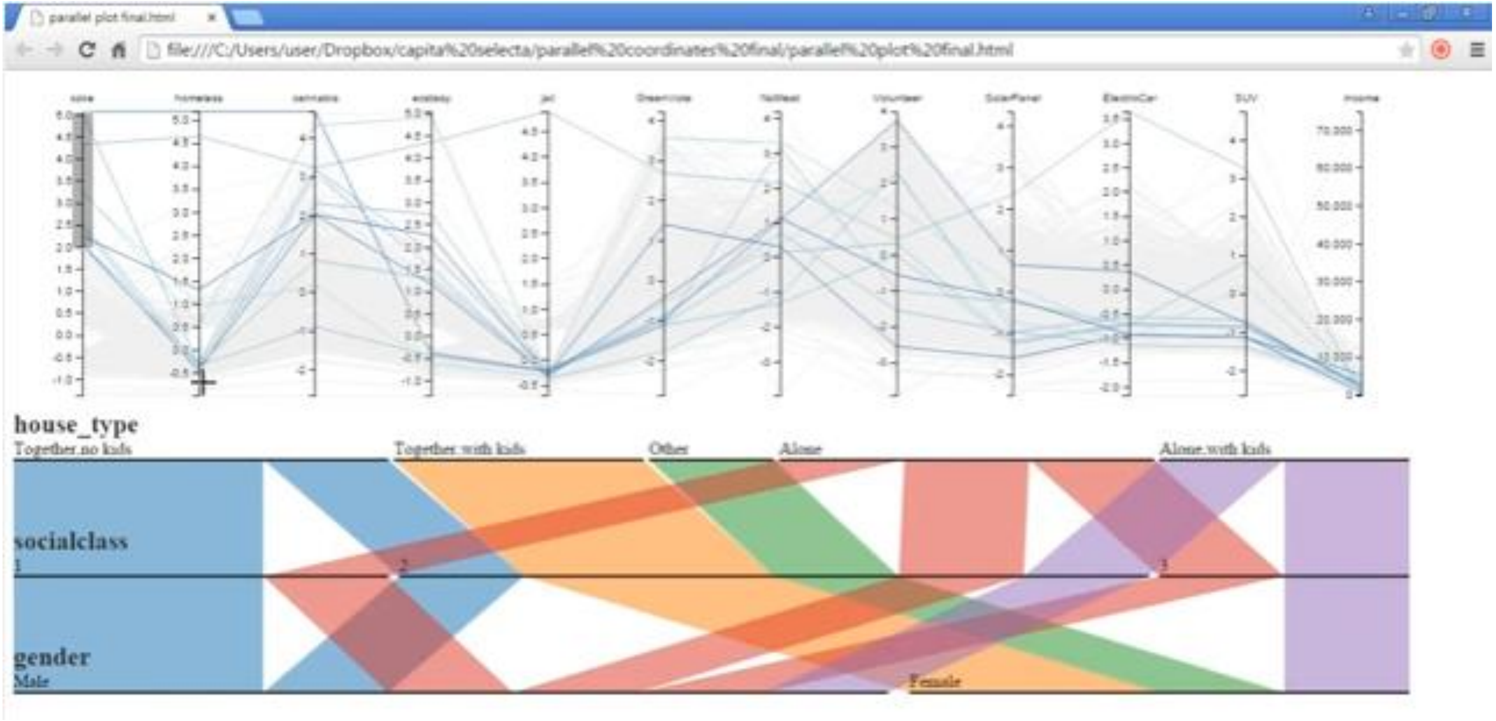


Fig. 7. Basic interaction elements in Parallel Sets: reordering categories (a, b) helps to generate a more meaningful layout; grouping categories (c, d) enables a hierarchical analysis/exploration; excluding categories from the visualization (e, f) allows for interactive filtering; and category highlighting (g, h) enables the selective investigation of high-dimensional relations.

http://code.google.com/p/parsets/

Video

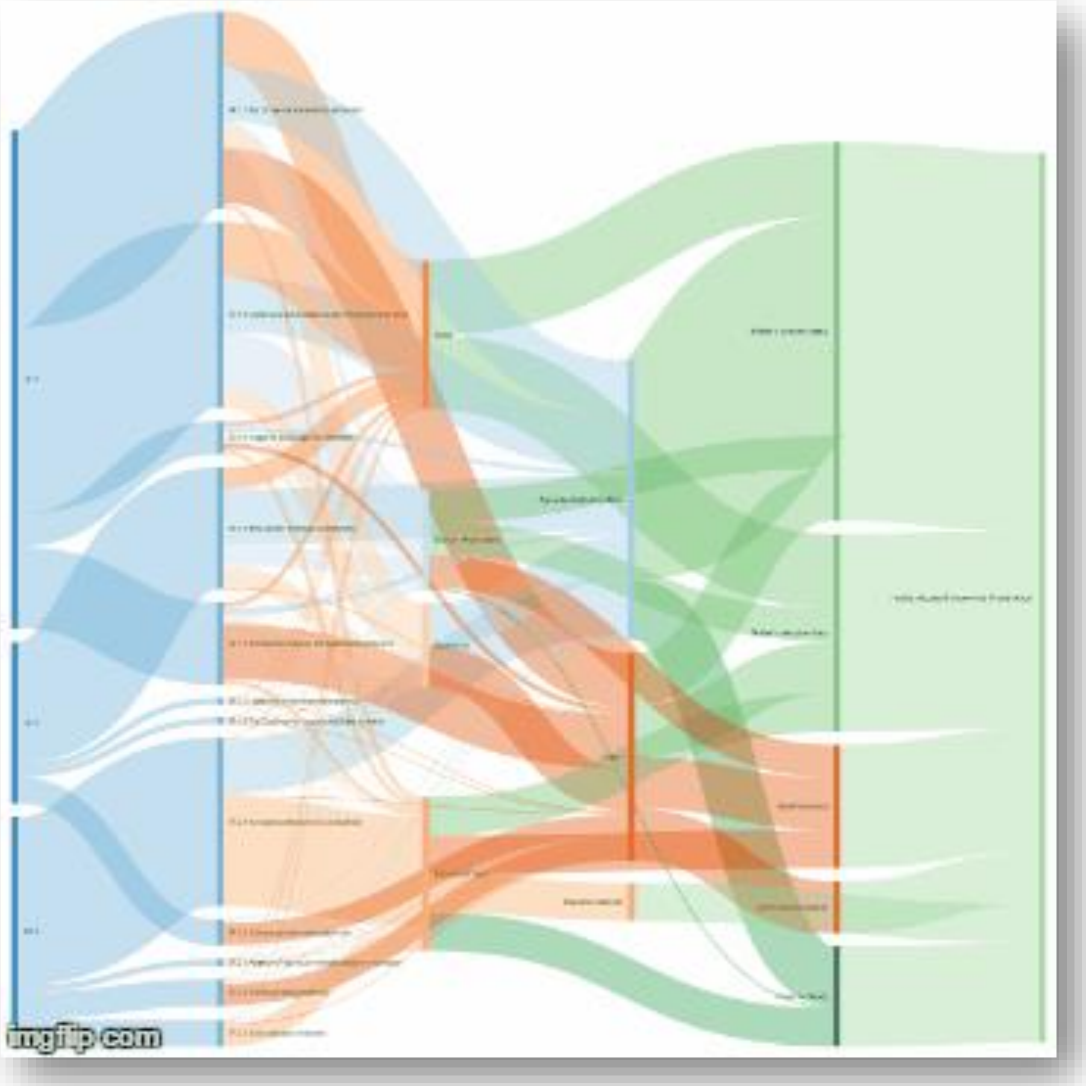


parallel plot final.html

file:///C:/Users/user/Dropbox/capita%20selecta/parallel%20coordinates%20final/parallel%20plot%20final.html

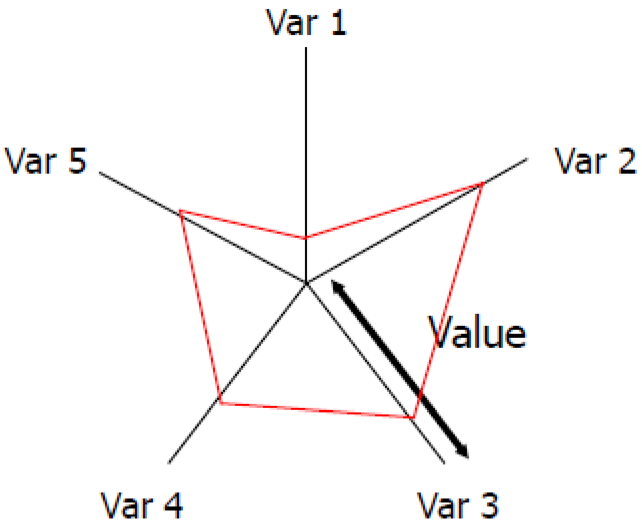
https://www.youtube.com/watch?v=SphrIOU76o0

Similar: Sankey Diagrams (flow)

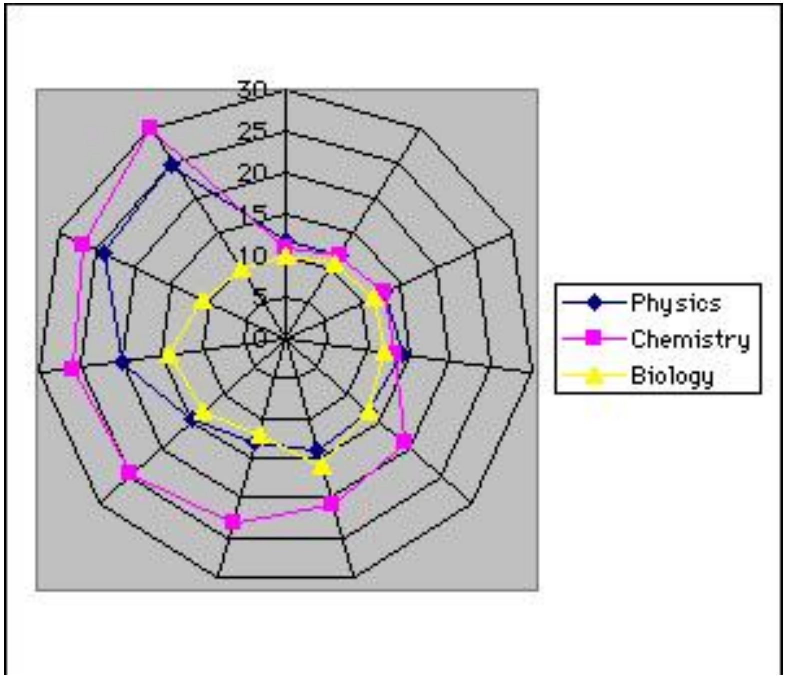


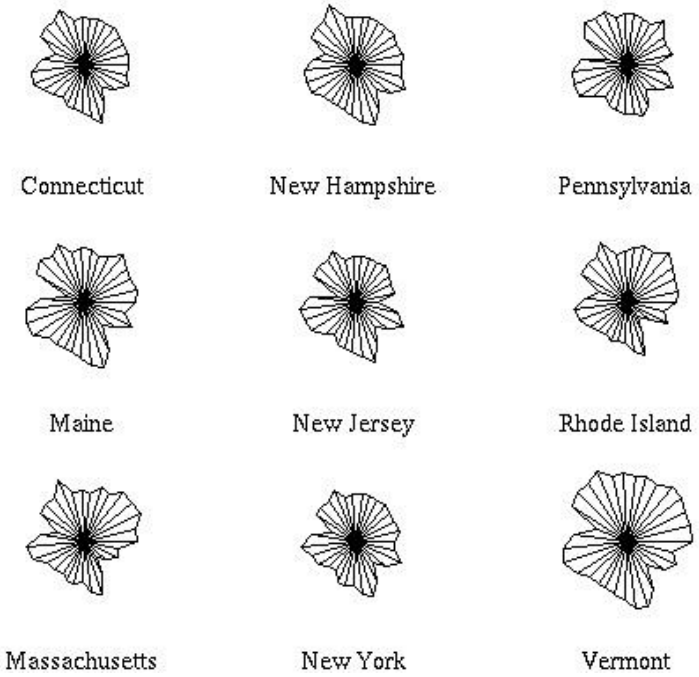
Star Plots

- Alternative representation
  - Space out the n variables at equal angles around a circle
  - Each “spoke“ encodes a variable’s value
  - Data point is now a “shape“!

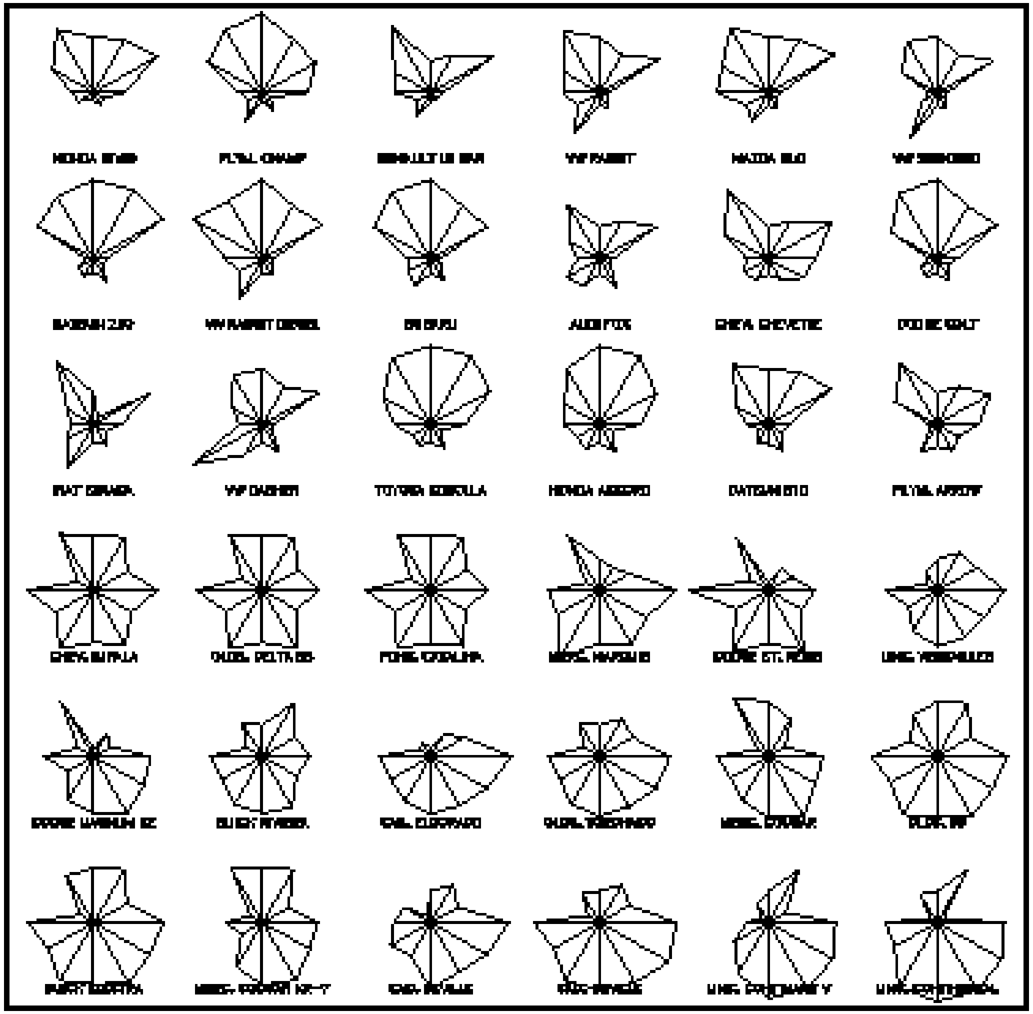


Star Plot examples



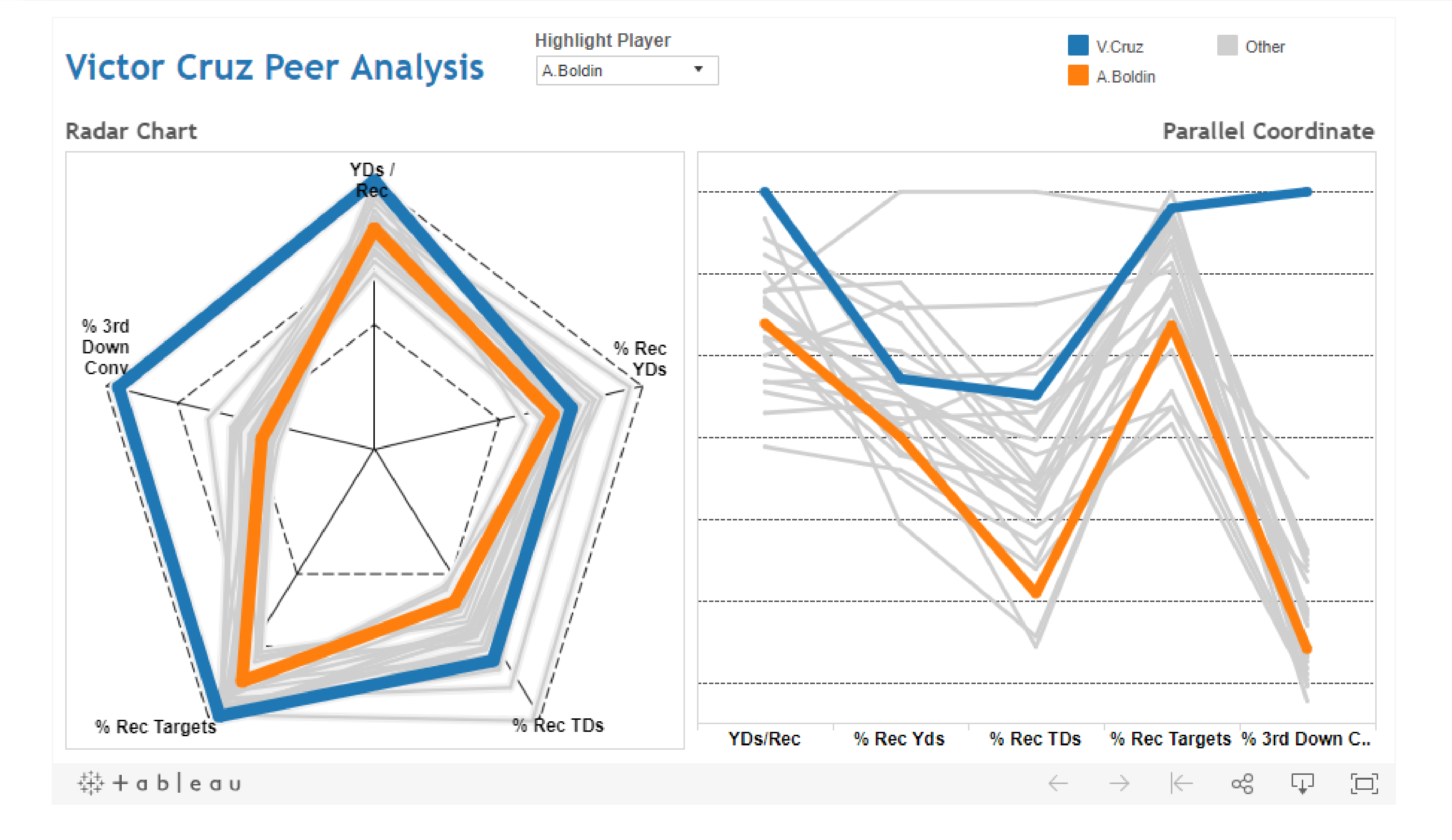


Star Plots



https://developer.ibm.com/predictiveanalytics/2012/03/14/reference-lines-for-star-plots-aid-interpretation/

Star/Radar Plot vs.Parallel Coordinates

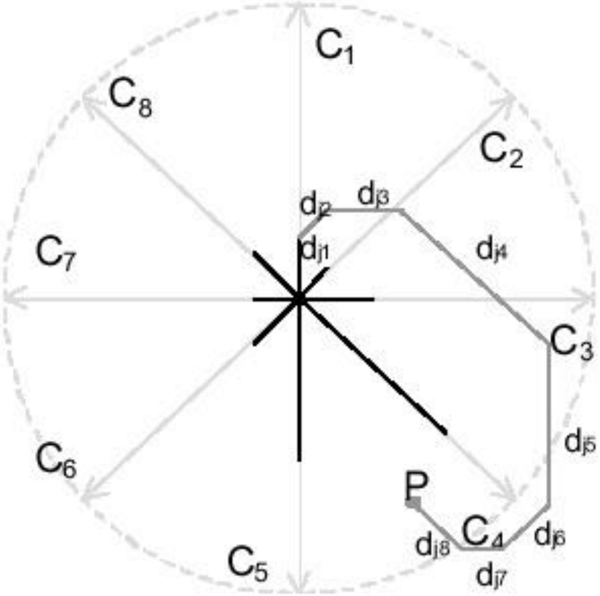


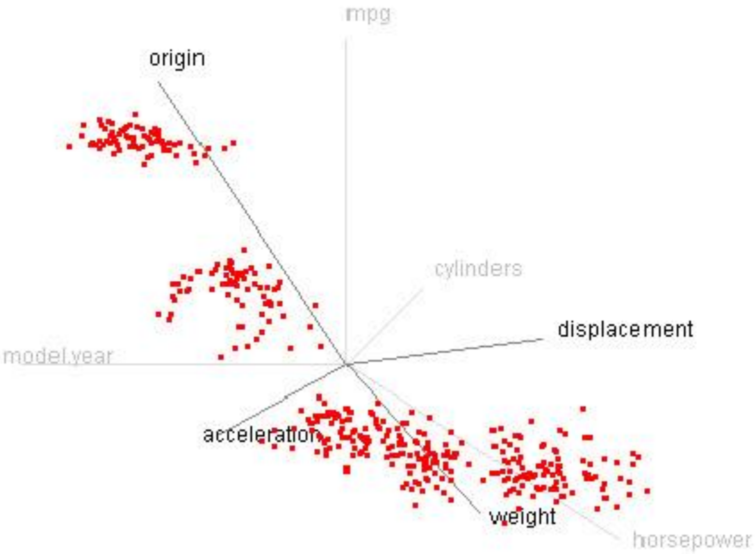
https://public.tableau.com/profile/adam.e.mccann#!/vizhome/RadarvsParallelCoordinate/RadarChartvsParallelCoordinate

Star Coordinates

- Same ideas as star plot
- Rather than represent case as polyline, just accumulate values along a vector parallel to particular axis
- Data case then becomes a point

Star Coordinates





E. Kandogan, “Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions”, *InfoVis 2000* Late-Breaking Hot Topics, Oct. 2000

Star Coordinates

- Data cases with similar values will lead to clusters of points
- Problem: Multi-dimensional scaling or projection down to 2D