



# Comparison of Multivariate Data Representations: Three Eyes are Better than One

Natsuhiko Kumasaka (Keio University)

Antony Unwin (Augsburg University)

# Content



- Visualisation of multivariate data
  - Parallel coordinate plots
  - Textile plots
  - Mosaic plots
- Visual data analysis
  - Some examples

# Parallel Coordinate Plots



- Each variable has its own vertical axis.
- Each case is represented by a set of line segments joining its points on the axes.
- The form, scaling and order of the axes influence the display a great deal.
- Interaction: querying, selecting and linking, rescaling, reordering



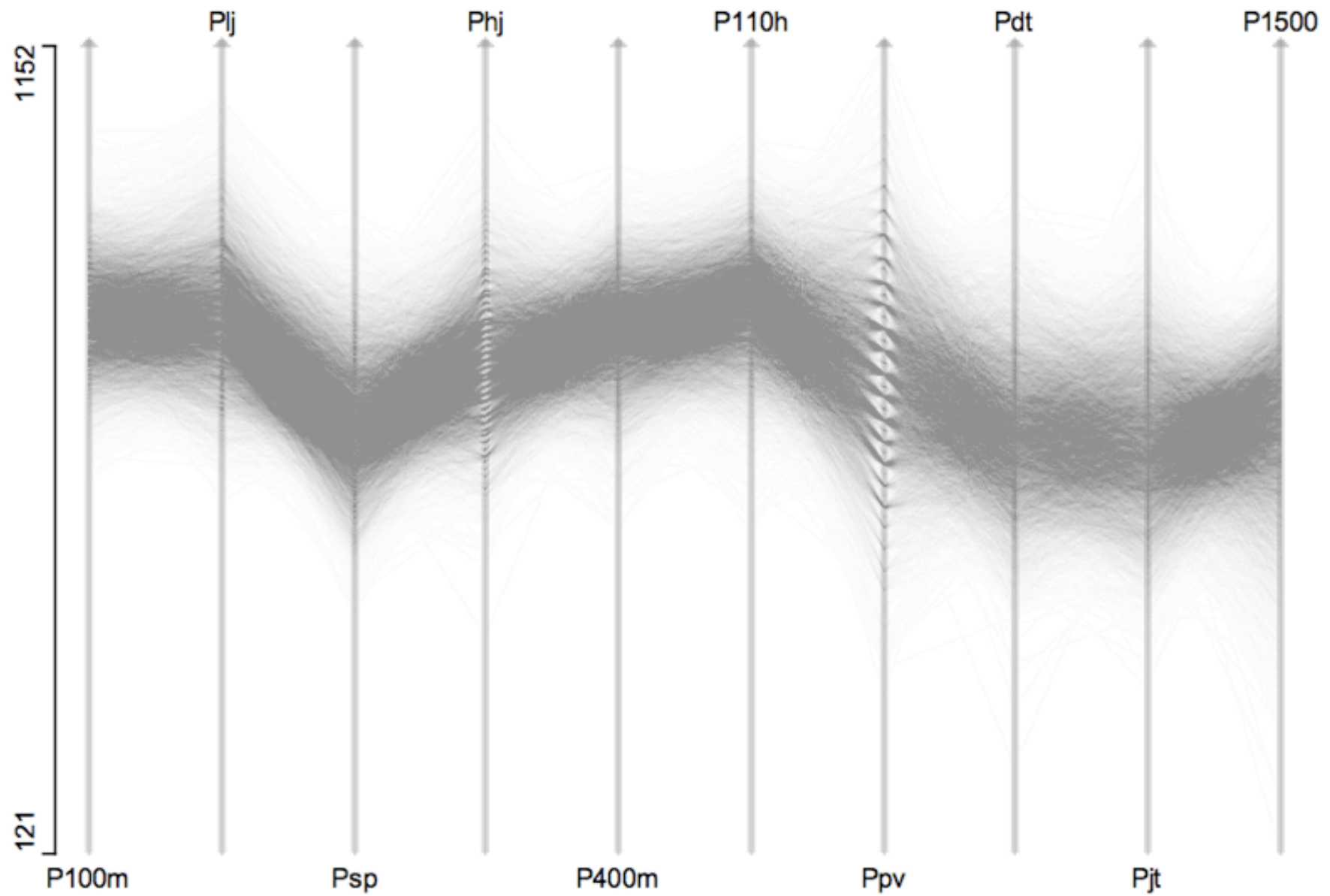
# Decathlon dataset

- Best performances each year, 1985 to 2006, by individual decathletes, 7968 cases
- Only complete, not hand-timed
- 10 events, results, points, competition dates
- Nationality, birthday
- Source: [www.decathlon2000.ee](http://www.decathlon2000.ee)

# Decathlon analysis goals



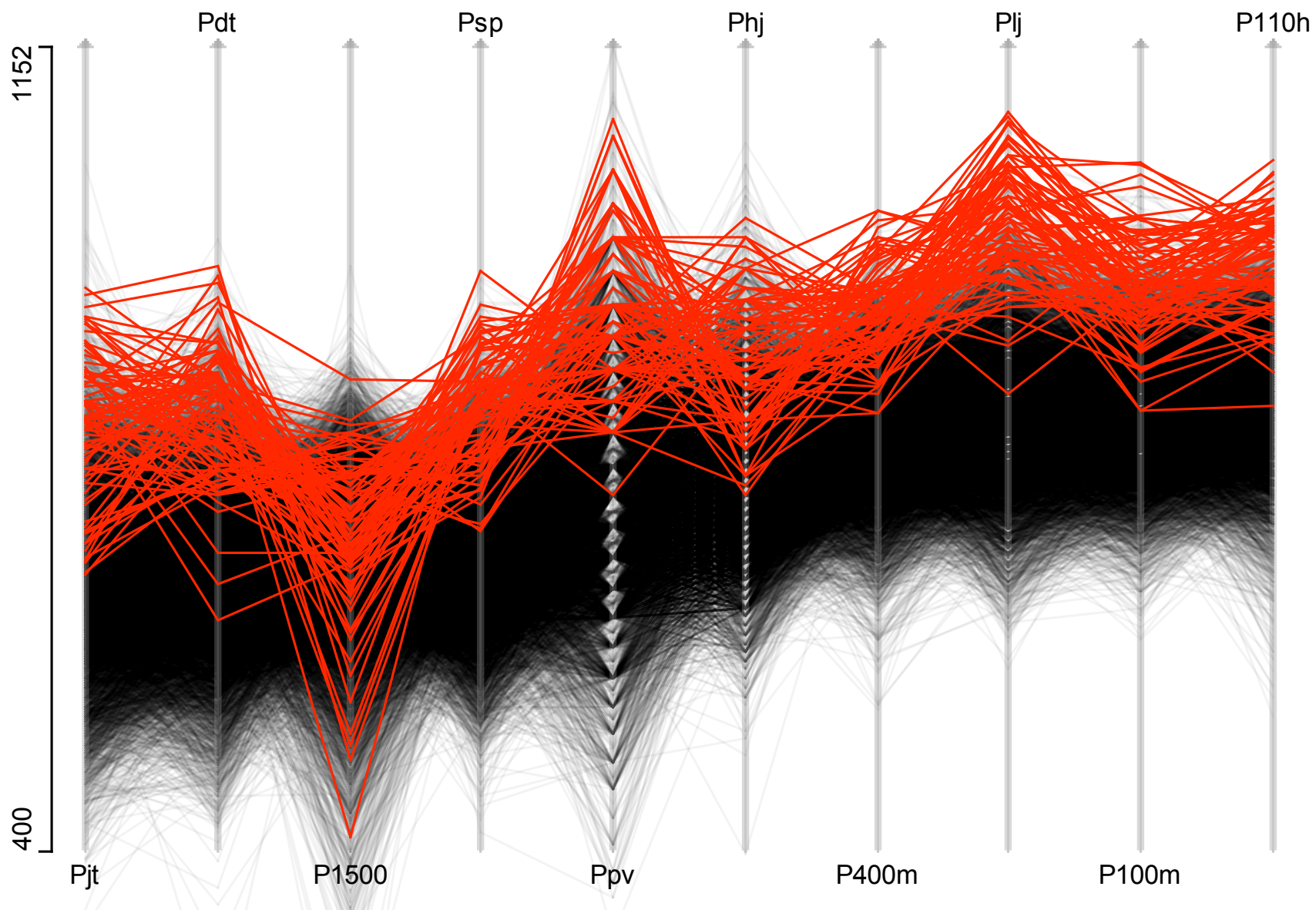
- Are the points' distributions the same for each event?
- Which events are most influential?
- Have the performances changed over the years?
- ...



Three eyes better than one

Kumasaka & Unwin

Cherry Bud 25th March, 2008



Three eyes better than one

Kumasaka & Unwin

Cherry Bud 25th March, 2008

# Wine dataset (1)



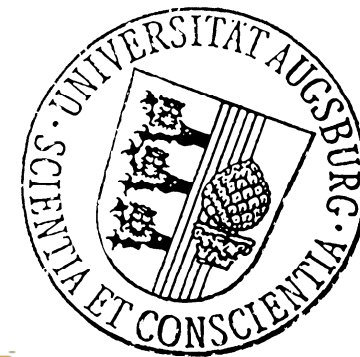
- Californian/French Tasting, July 1976
- 10 Cabernet Sauvignons (6 US, 4 French)
- 11 judges (9 French, 1 US, 1 English) scored the wines from 0 to 20
- The data may also be analysed as ranks
- Source: [www.liquidaset.com](http://www.liquidaset.com)

# Wine dataset (2)



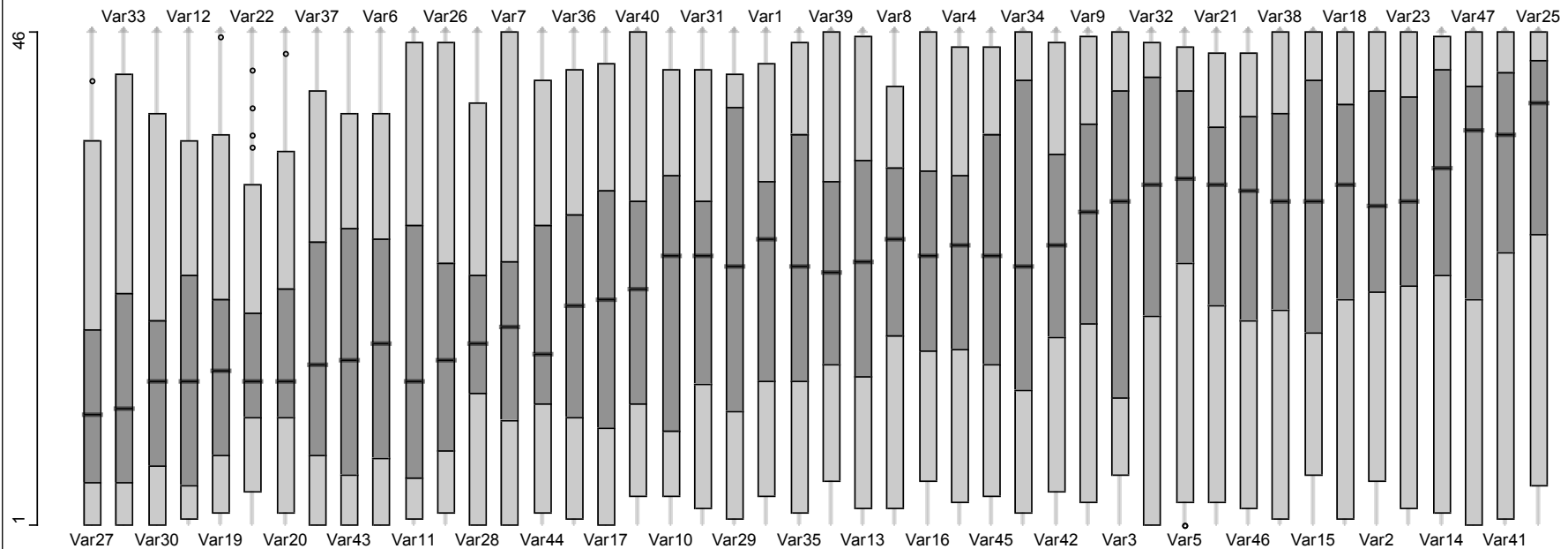
- Cabernet challenge 1999
- 47 (only 46 rated) Cabernet Sauvignons: 34 US, 9 French, 2 Italian, 2 others
- Vintages from 1994 to 1996
- 33 judges (Californian) ranked the wines
- Source: [www.liquidaset.com](http://www.liquidaset.com)

# Wine analysis goals



- Which wines were rated best?
- Is the ranking of wines clear-cut?
- Do the judges have similar opinions?
- Are there clusters of judges?
- ...

# Wine boxplots by mean



Three eyes better than one

Kumasaka & Unwin

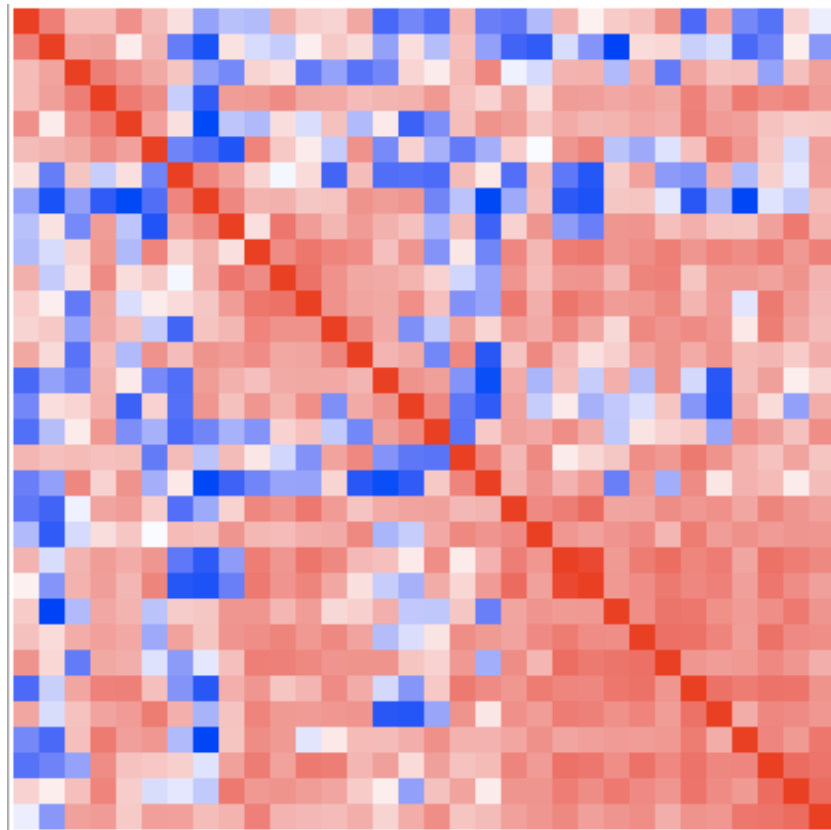
Cherry Bud 25th March, 2008

# Judge correlations



A heatmap of correlations between judges, after Ward clustering of the original data to order the judges.

(The display was drawn with Alex Gribov's SEURAT.)



# PCPs and Textile plots



- PCPs stick to the raw data
- Textile plots transform scales
- Textile plots offer informative defaults
- PCPs are flexible through interaction

# Mosaicplots (Classical)



- A rectangle is drawn for every combination of categories. Area is proportional to count.
- Divide the horizontal axis according to the category counts of the first variable.
- Divide each vertical column according to the relevant counts of the second variable.
- Continue dividing horiz / vert according to the relevant counts of the next variable.

# A zoo dataset



- 101 animal types
- 17 attributes (mostly binary)
- Created by Forsyth
- Source: [mllearn.ics.uci.edu/databases/](http://mllearn.ics.uci.edu/databases/)
- Analysis goals: What features best classify animals by type? How are the features related?



# Mosaicplot variants

- Classical for efficient use of space
- Fluctuation diagrams for cell sizes
- Same binsize to identify zeros and compare rates
- Multiple barcharts for comparisons
- Doubledecker plots for rates

# Mosaicplot interactions



- Querying
- Reordering variables
- Reordering categories
- Rotating variables
- Rotating plots
- Size and aspect ratio
- Censored (and quantum) zooming

# Titanic dataset



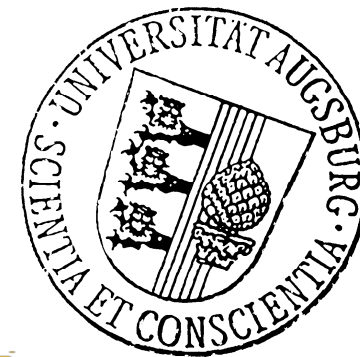
- 2201 passengers and crew
  - Class (First, Second, Third, Crew)
  - Age (Young, Old)
  - Gender (Male, Female)
  - Survived (Yes, No)
- Journal of Statistics Education (1995)

# Titanic analysis goals



- Which kinds of passenger survived?
- Did “Women and children first” apply?
- What was the effect of class?
- What was the combined effect of gender and class?

# Mosaicplots and Textiles



- Mosaicplots show variable combinations
- Mosaics are limited in number of variables
- Mosaics have many, many display options to reveal information
- Textile plots emphasise absolute numbers
- Textile plots can handle many variables

# Software



- Mondrian (Martin Theus)
  - interactive graphics
  - crossplatform, links to R via Rserve
- iPlots
  - available as R package

[www.stats.math.uni-augsburg.de](http://www.stats.math.uni-augsburg.de)



# Conclusions

- Multivariate displays have many options and making choices is difficult
- Textile plots can provide excellent defaults
- Interactive tools empower graphics, when they are fast, flexible and efficient
- No one display can show all information — “Three eyes are better than one”