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

**Antony Unwin**
Augsburg University

**Natsuhiko Kumasaka**
Keio University

# Multivariate Data Representation

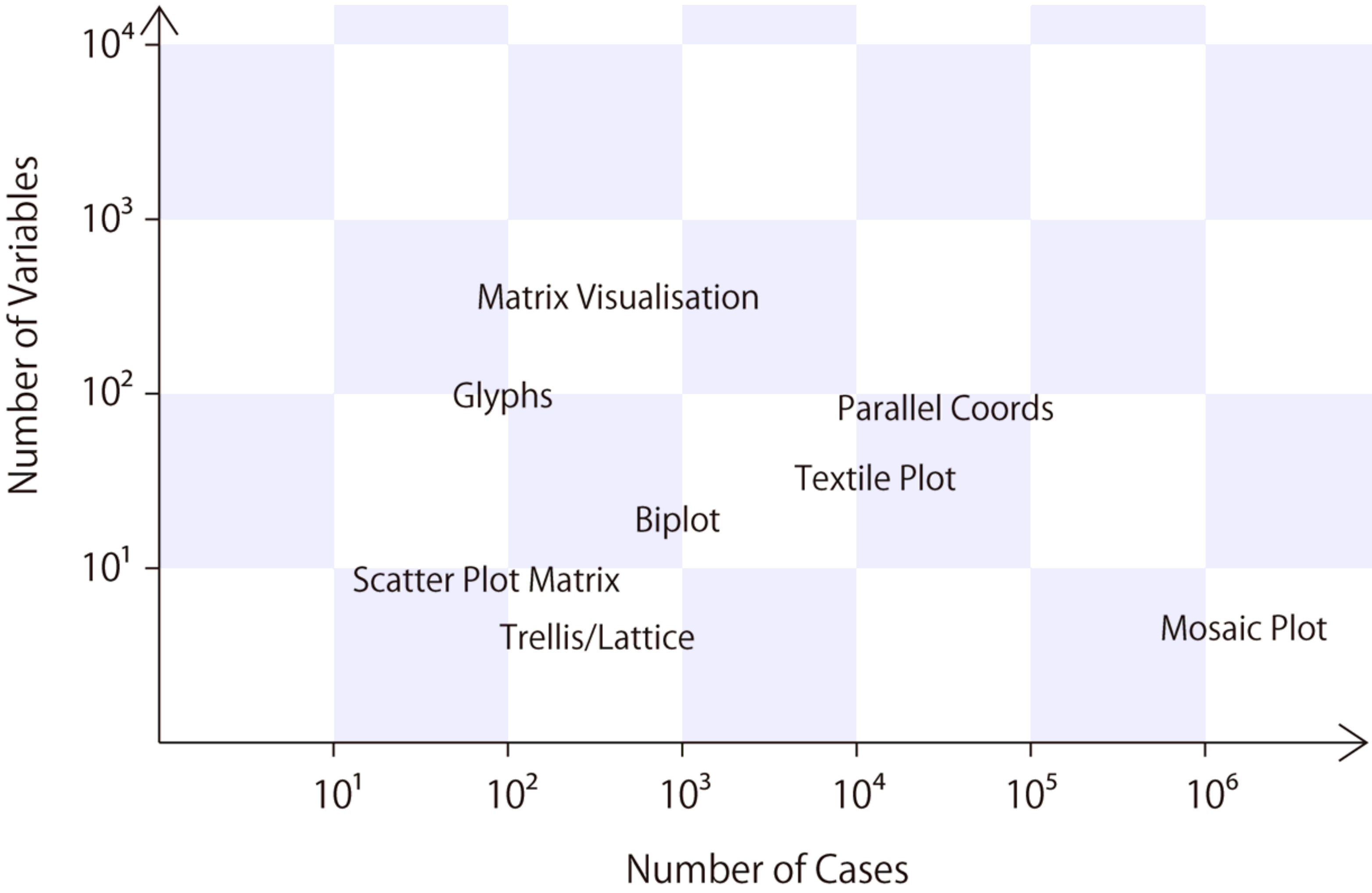
- **Numerical data**
  - Biplot (Gabriel 1971)
  - Scatter plot matrix (Cleveland 1984)
  - GGobi (Cook, D. et.al 2007)
  - Glyphs (Anderson 1957, Chernoff 1973, Fienberg 1979)
  - Parallel coordinate plot (Inselberg 1985, Wegman 1990)
  - Matrix visualization (Chen 2002)
- **Categorical data**
  - Mosaic plot (Hartigan 1981)
  - MANET (Unwin et.al. 1996)
- **Numerical/Categorical data**
  - Trellis/Lattice (Chambers 1992, Cleveland 1993)
  - Textile plot (Kumasaka and Shibata 2008)

- **General**
  - Graphics of Large Datasets (Unwin et.al. 2006)
  - Handbook of Data Visualization (Chen, et.al. 2008)

# CONTENT

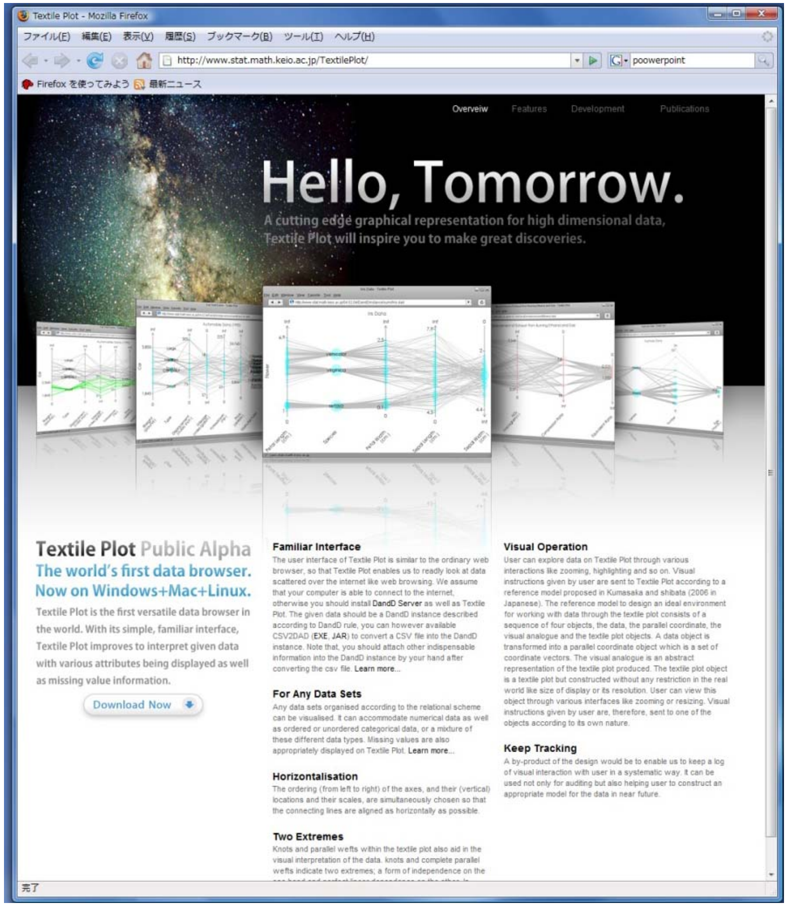
- 3 multivariate data representations
  - Parallel coordinate plot
  - Mosaic plot
  - Textile plot
- Visual data analysis
  - Decathlon data
  - Wine data
  - Animal data
  - Titanic data

# Capability of the Number of Cases and Variables on a Display



# TEXTILE PLOT (Kumaska and Shibata 2006, 2007, 2008)

- Parallel coordinate system
- Horizontalisation criterion
- Any type of data
- Order of Axes




http://stat.math.keio.ac.jp/TextilePlot/

# Transform Datavector into Coordinate Vector

$$\underset{n \times p}{\begin{pmatrix} \\ x_{ij} \\ \\ \end{pmatrix}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \quad \Rightarrow \quad \underset{n \times p}{\begin{pmatrix} \\ y_{ij} \\ \\ \end{pmatrix}} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)$$

Data Vector $\boldsymbol{x}_j$ ($n$- dimensional)

Numerical

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j$$

Non-numerical

$\mathbf{X}_j : n \times (q_j - 1)$ matrix

$q_j$: number of levels

contrasts

$$\mathbf{C} = \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \ddots & \vdots \\ \vdots & \ddots & 0 \\ 1 & \cdots & 1 \end{pmatrix}$$

Unordered

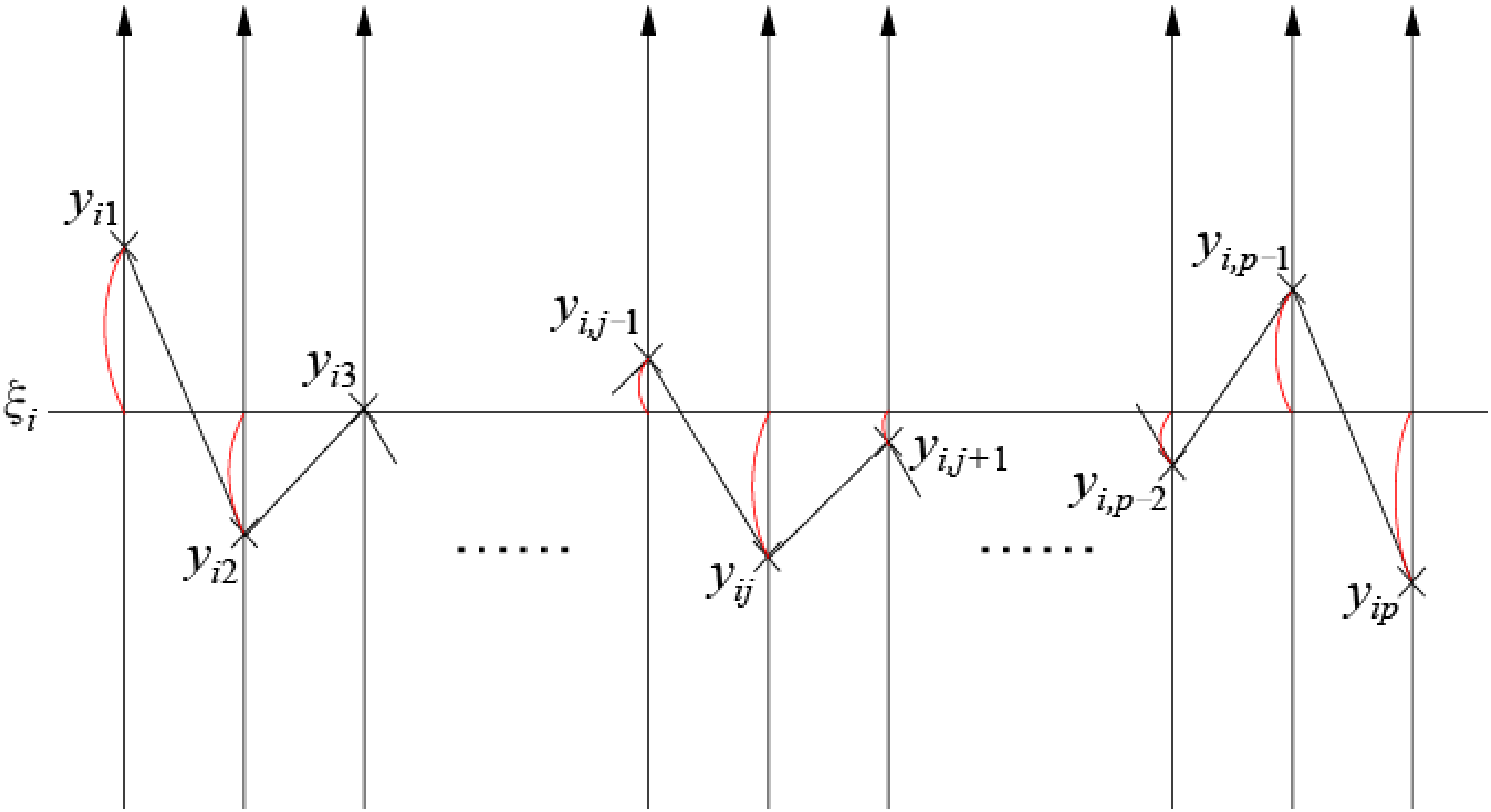
$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j$$

Ordered

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j$$

$$\boldsymbol{\beta}_j \leq \mathbf{0} \text{ or } \geq \mathbf{0}$$

# HORIZONTALISATION CRITERION

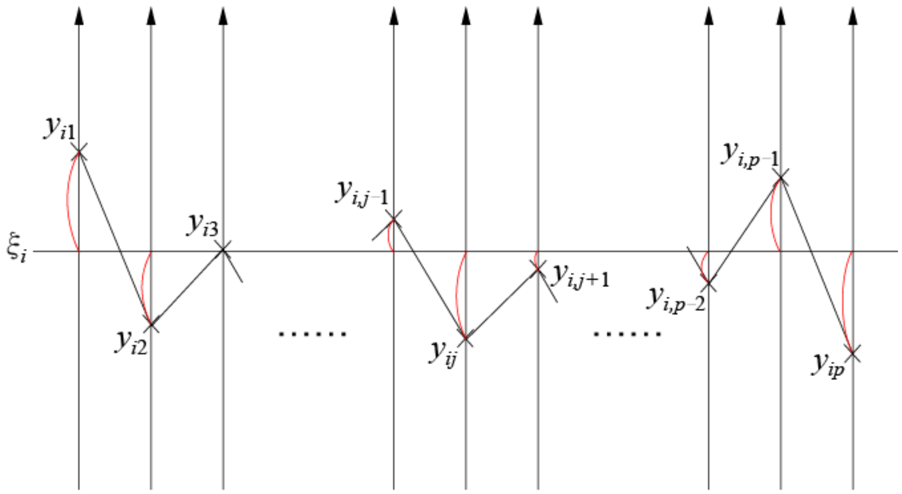



$$\sum_{j=1}^{p}(y_{ij}-\xi_i)^2$$

# Optimisation Problem

Minimise $$\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}-\xi_i)^2 = \sum_{j=1}^{p}\left\|\boldsymbol{y}_j-\boldsymbol{\xi}\right\|^2 \underset{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}{\rightarrow} \min$$

Subject to $$\sum_{j=1}^{p}\|\boldsymbol{y}_j-\bar{y}_{\cdot j}\mathbf{1}\|^2 = np$$




Location Parameter Vector

$$\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_p)^T$$

Scale Parameter Vector

$$\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^T$$

Ideal Coordinate Vector

$$\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)^T$$

# SOLUTION $(\mathbf{1}^T \boldsymbol{x}_j = 0 \text{ and } \|\boldsymbol{x}_j\| = 1;\ j = 1, \ldots, p)$

$$\sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{\xi}\|^2 = \sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2 + p\|\boldsymbol{m} - \boldsymbol{\xi}\|^2$$

$$\Rightarrow \quad \hat{\boldsymbol{\xi}} = \boldsymbol{m} = \frac{1}{p}\sum_{j=1}^{p} \boldsymbol{y}_j \qquad (\boldsymbol{m}:\ \text{mean vector})$$

$$\sum_{j=1}^{p} \|\boldsymbol{y}_j - \boldsymbol{m}\|^2 = \sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\bullet j}\mathbf{1}\|^2 - p\|\boldsymbol{m} - \bar{y}_{\bullet\bullet}\mathbf{1}\|^2 + \sum_{j=1}^{p} \|\bar{y}_{\bullet j}\mathbf{1} - \bar{y}_{\bullet\bullet}\mathbf{1}\|^2$$

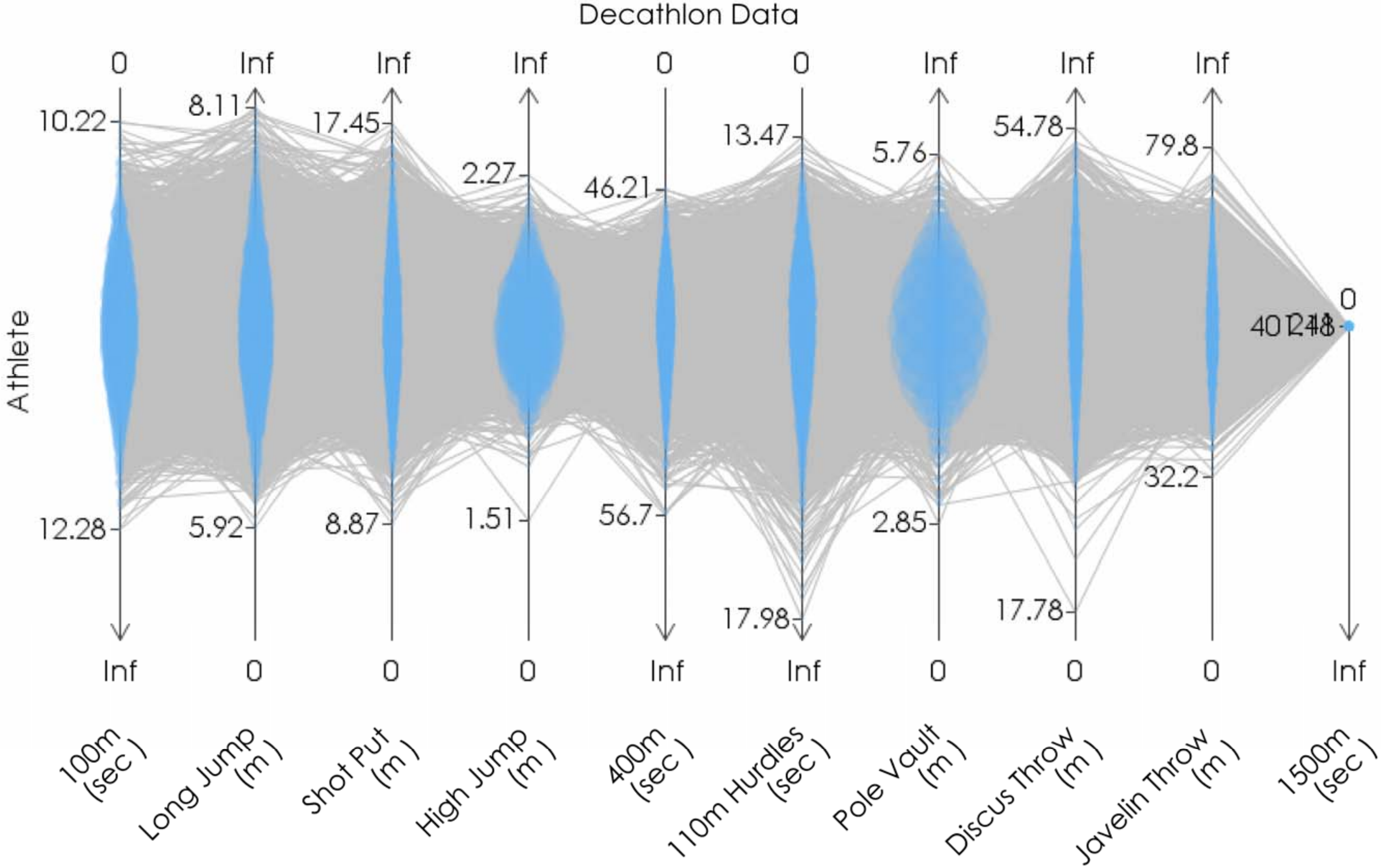
$$\Rightarrow \quad \hat{\alpha}_j = \alpha_0; \quad j = 1, \ldots, p \qquad (\bar{y}_{\cdot\cdot} = \textstyle\sum_{j=1}^{p} \bar{y}_{\cdot j}/p)$$

$$\sum_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\bullet j}\mathbf{1}\|^2 - p\|\boldsymbol{m} - \bar{y}_{\bullet\bullet}\mathbf{1}\|^2 = \|\boldsymbol{\beta}\|^2 - \frac{1}{p}\boldsymbol{\beta}^T\mathbf{R}\boldsymbol{\beta}$$

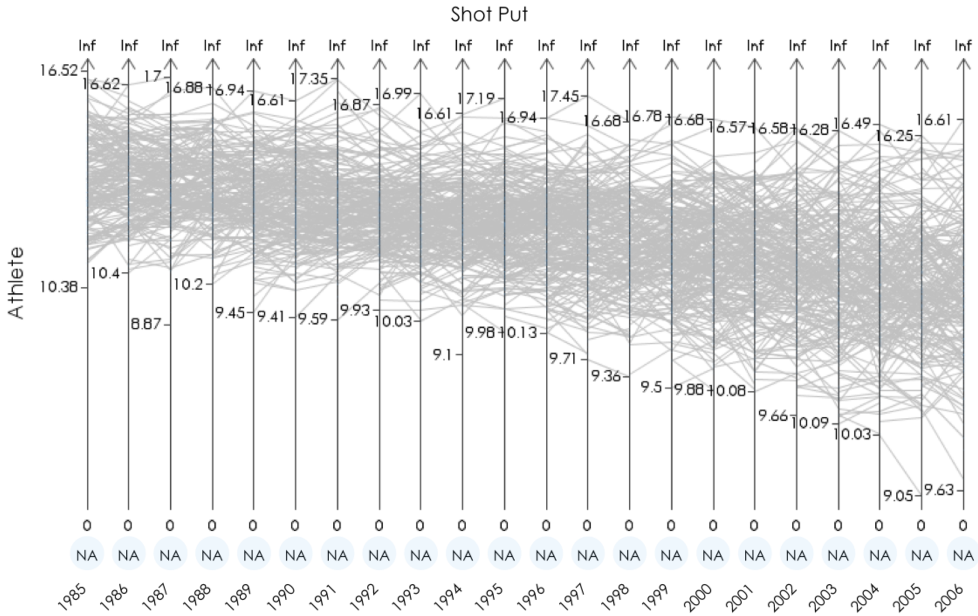
$\Rightarrow$ $\hat{\boldsymbol{\beta}}$: Eigenvector of sample correlation matrix $\mathbf{R}$ with the largest eigenvalue, such that $\|\hat{\boldsymbol{\beta}}\|^2 = np$.

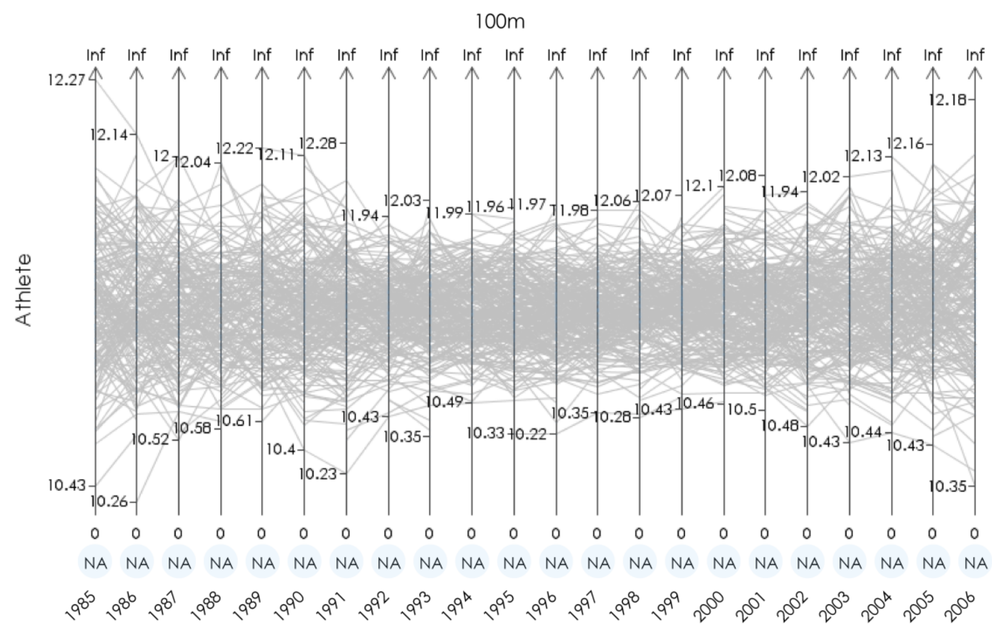
$(\Sigma_{j=1}^{p} \|\boldsymbol{y}_j - \bar{y}_{\bullet j}\mathbf{1}\|^2 = \|\boldsymbol{\beta}\|^2 = np)$

# TEXTILE PLOT OF DECATHLON DATASET (PERFORMANCES NOT POINTS)



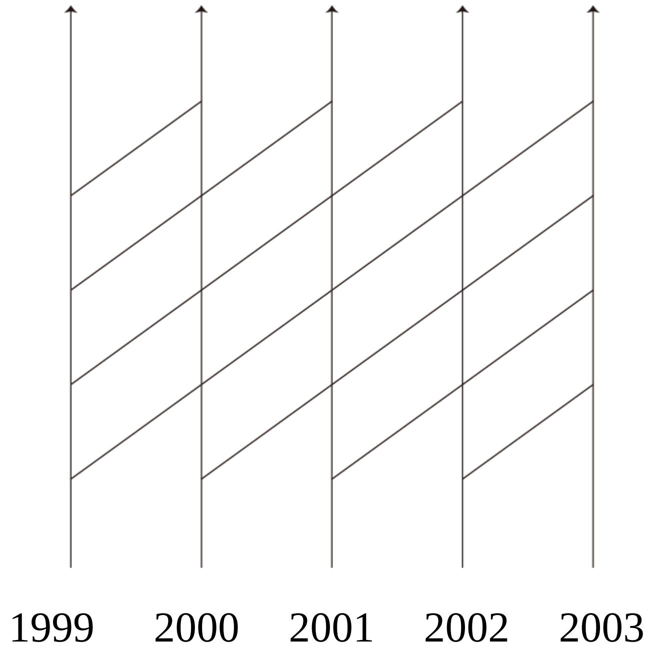
# Re-organised Decathlon Dataset



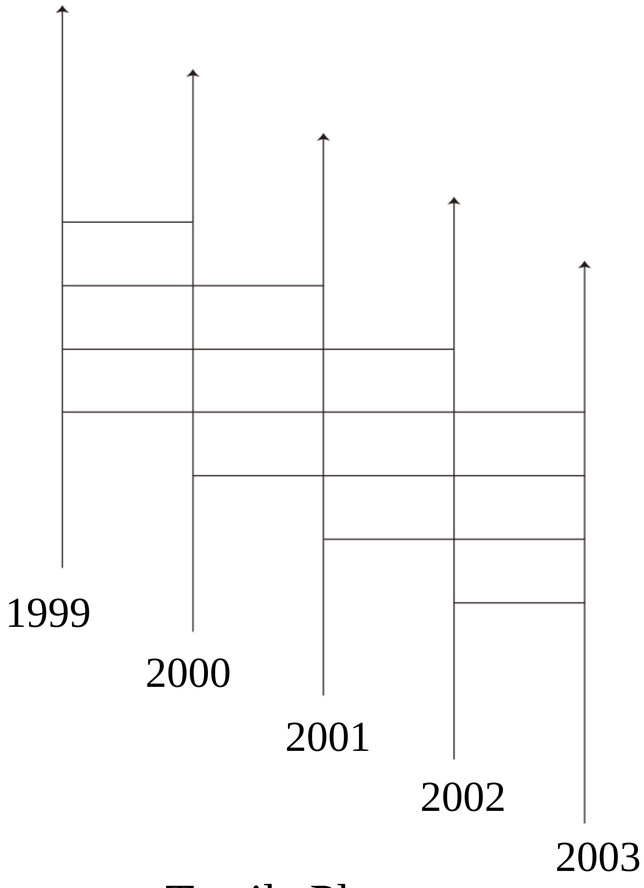
# Trend on Textile Plot

Assumption: All athletes improve their performances year by year, and stop their careers at their peak.




Parallel Coordinate Plot
with Common Scaling

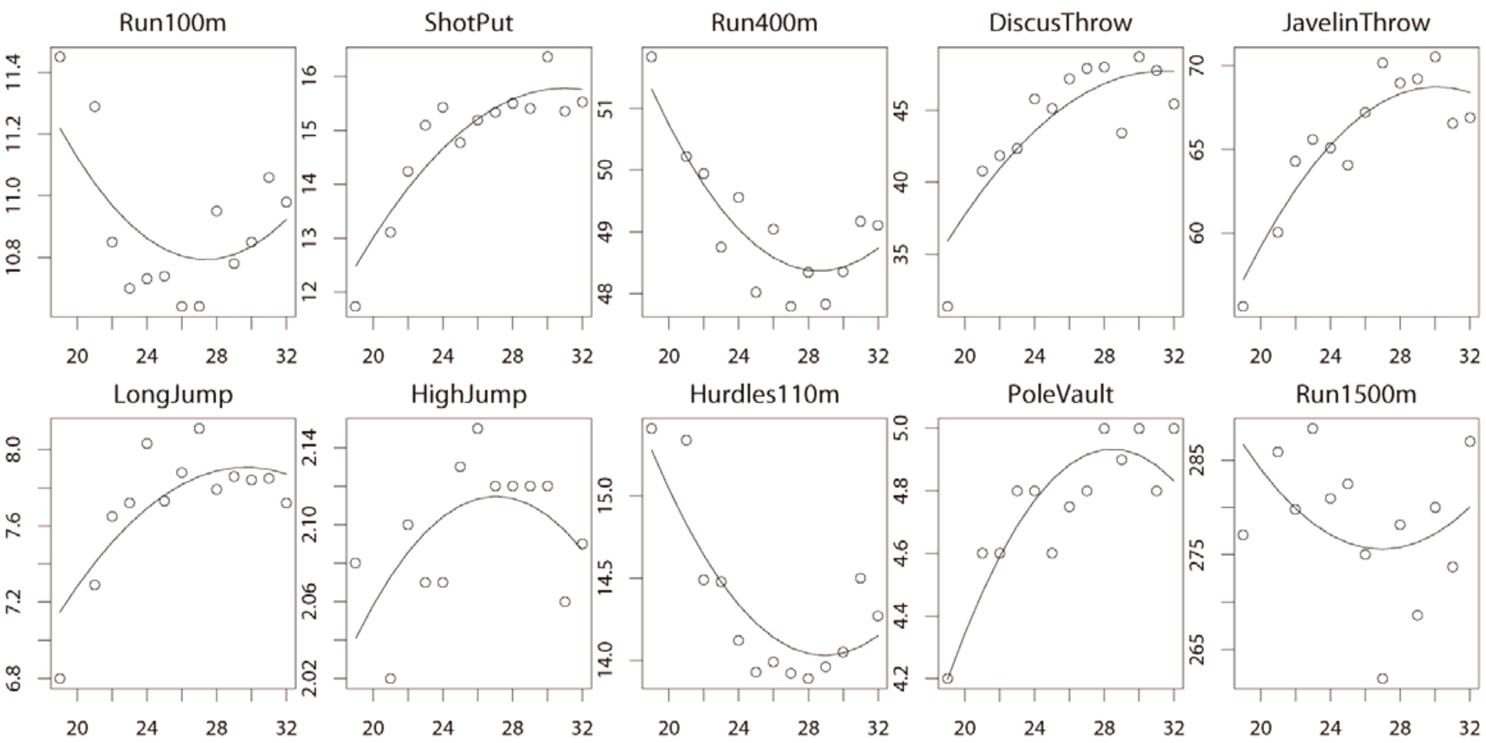
Horizontalisation




Textile Plot

# FURTHER CONFIRMATORY DATA ANALYSIS


- Performances of Mr. Roman Sebrle (best record holder)




$$Performance_{ik} = \alpha_{ik} + \beta_k (Age_i - \gamma_{ik})^2 + \varepsilon_{ik},$$

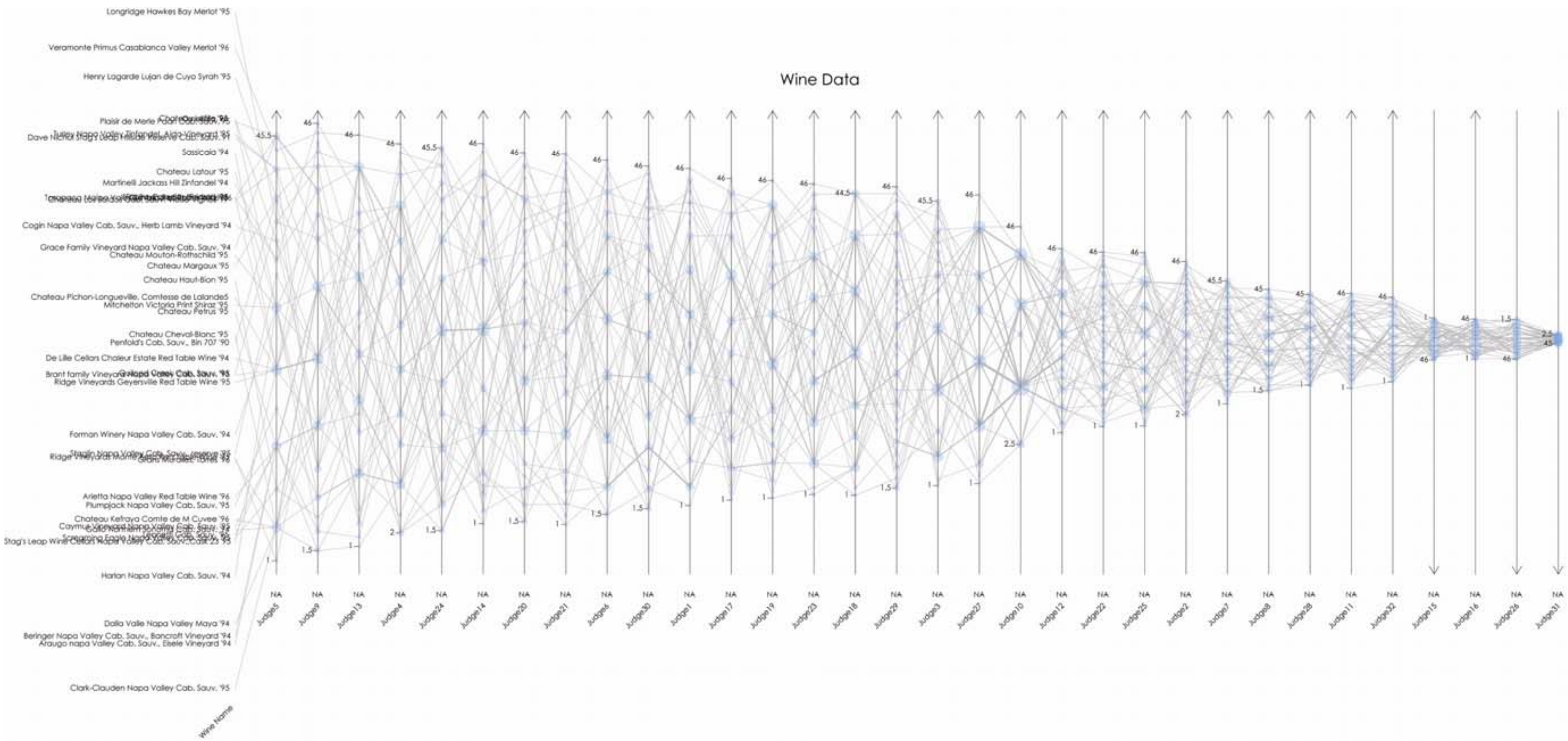
$i$: athlete, $k$: event

# Wine Dataset (Liquid Assets: www.liquidasset.com)

- Cabernet challenge 1999
- Case
  - 47 (only 46 rated) Cabernet Sauvignons: 34 US, 9 French, 2 Italian, 2 others
  - Vintages from 1994 to 1996
- Variable
  - 33 judges (Californian) ranked the wines

- 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?

# TEXTILE PLOT OF WINE DATASET



## Graphics for Multivariate Numerical Data

- Textile plot or parallel coordinate plot gives an overview of the data
- Re-organisation of data is always useful to know another aspects of the data
- Textile plots suggest potential avenues for subsequent further confirmatory data analysis

# TRANSFORM DATAVECTOR INTO COORDINATE VECTOR

$$\underset{n \times p}{\begin{pmatrix} \\ x_{ij} \\ \\ \end{pmatrix}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \quad \Rightarrow \quad \underset{n \times p}{\begin{pmatrix} \\ y_{ij} \\ \\ \end{pmatrix}} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_p)$$

Data Vector $\boldsymbol{x}_j$ ($n$- dimensional)

Numerical

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \beta_j \boldsymbol{x}_j$$

Non-numerical

$\mathbf{X}_j : n \times (q_j - 1)$ matrix

$q_j$: number of levels

contrasts

$$\mathbf{C} = \begin{pmatrix} 0 & \cdots & 0 \\ 1 & \ddots & \vdots \\ \vdots & \ddots & 0 \\ 1 & \cdots & 1 \end{pmatrix}$$

Unordered

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j$$

Ordered

$$\boldsymbol{y}_j = \alpha_j \mathbf{1} + \mathbf{X}_j \boldsymbol{\beta}_j$$

$\boldsymbol{\beta}_j \leq \mathbf{0}$ or $\geq \mathbf{0}$

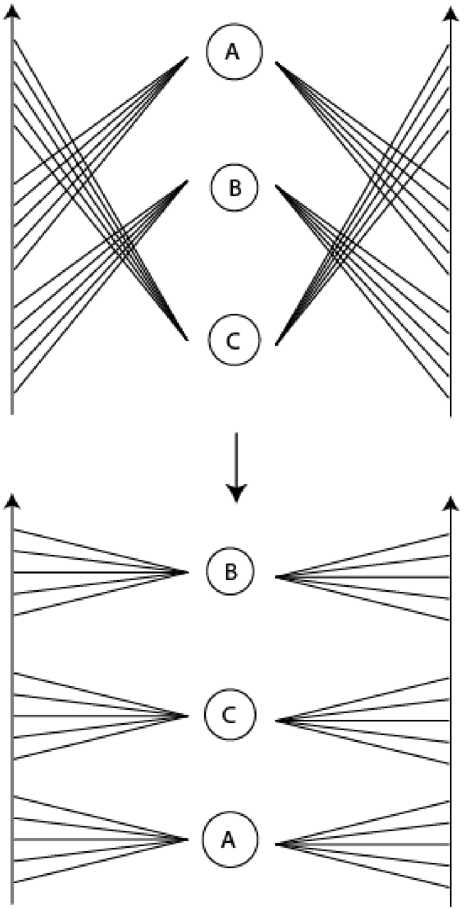
# Categorical Datavector

- Coordinate vector transformation
  - Determine optimised default position for each level
  - Introduction of a set of contrasts

Ex: Using a *treatment* contrast
$$\begin{array}{c} A \\ B \\ C \end{array} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\boldsymbol{x} = \begin{pmatrix} A \\ A \\ B \\ C \\ C \end{pmatrix} \Rightarrow \mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Coordinate vector:
$$\boldsymbol{y} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_2 \\ \alpha + \beta_2 \end{pmatrix}$$



# ORDERED CATEGORICAL DATAVECTOR

- Order of levels are retained on the axis
  - Introduction of the specific contrasts
  - Additional constraints

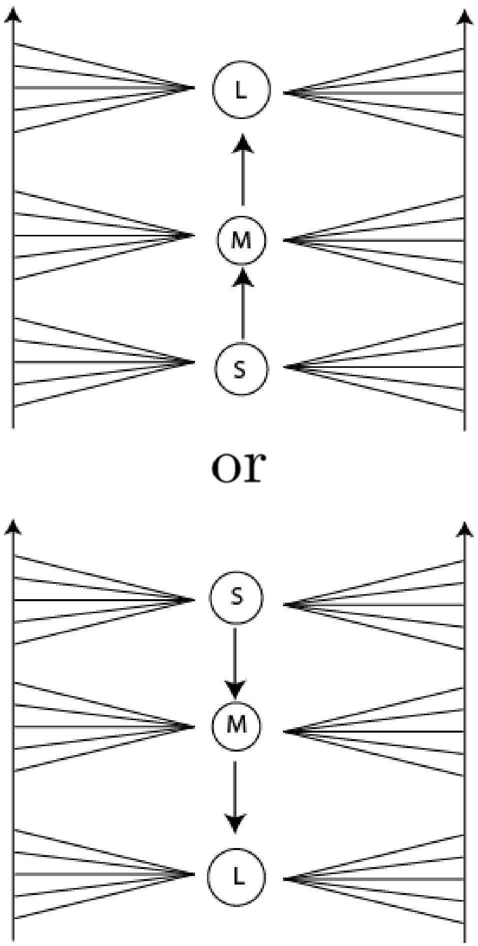
Ex:
$$\begin{array}{l} Small \\ Medium \\ Large \end{array} \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$\boldsymbol{x} = \begin{pmatrix} Small \\ Small \\ Medium \\ Large \\ Large \end{pmatrix} \Rightarrow \mathbf{X} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Coordinate vector:

$$\boldsymbol{y} = \alpha\mathbf{1} + \mathbf{X}\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \alpha \\ \alpha + \beta_1 \\ \alpha + \beta_1 + \beta_2 \\ \alpha + \beta_1 + \beta_2 \end{pmatrix}$$

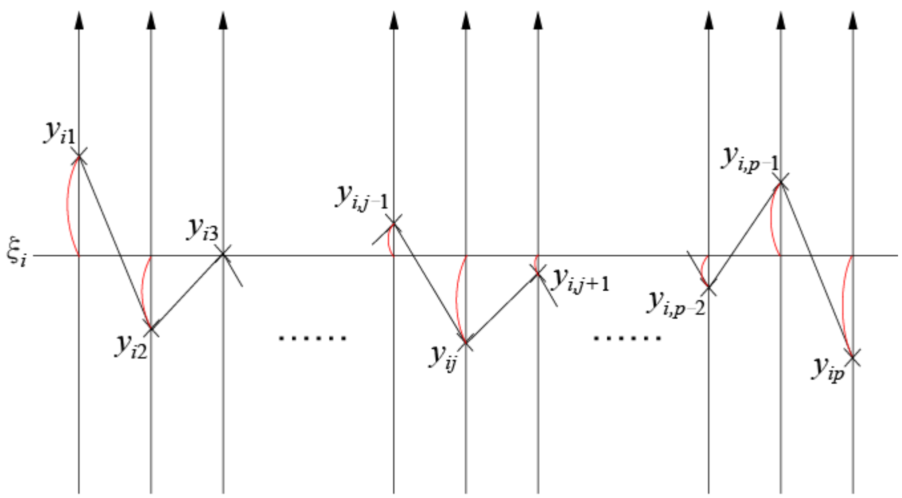
$$\beta_1, \beta_2 \geq 0 \quad \text{or} \quad \beta_1, \beta_2 \leq 0$$



# Optimisation Problem

Minimise $$\sum_{i=1}^{n}\sum_{j=1}^{p}(y_{ij}-\xi_i)^2 = \sum_{j=1}^{p}\left\|\boldsymbol{y}_j-\boldsymbol{\xi}\right\|^2 \underset{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\xi}}{\longrightarrow} \min$$

Subject to $$\sum_{j=1}^{p}\|\boldsymbol{y}_j-\bar{y}_{\cdot j}\mathbf{1}\|^2 = np \quad (\boldsymbol{\beta}_j \geq \mathbf{0} \text{ or } \boldsymbol{\beta}_j \leq \mathbf{0})$$




Location Parameter Vector
$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$$

Scale Parameter Vector
$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_p^T)^T$$

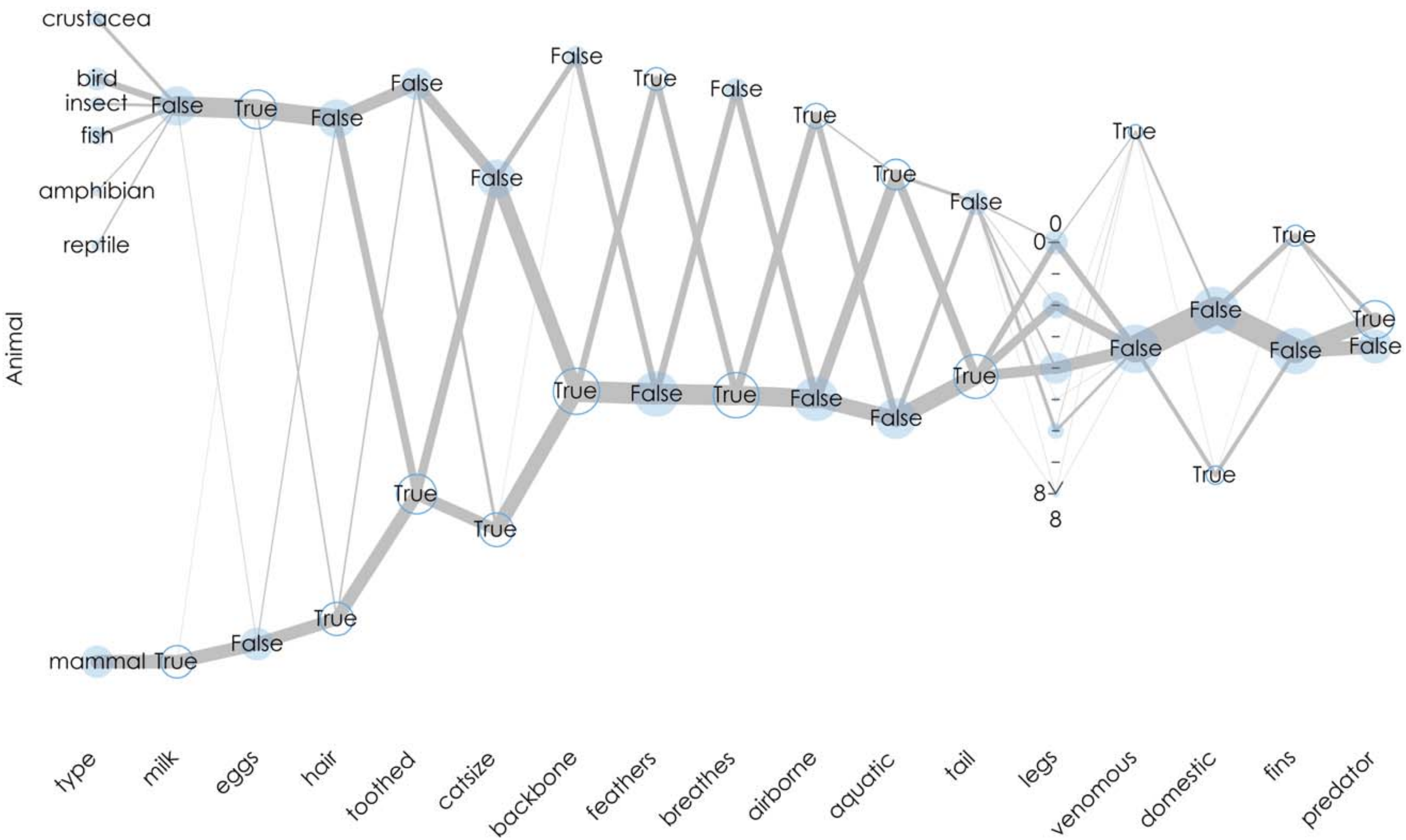
Ideal Coordinate Vector
$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$$

# Animal Data (UCI Machine Learning Group 2008)

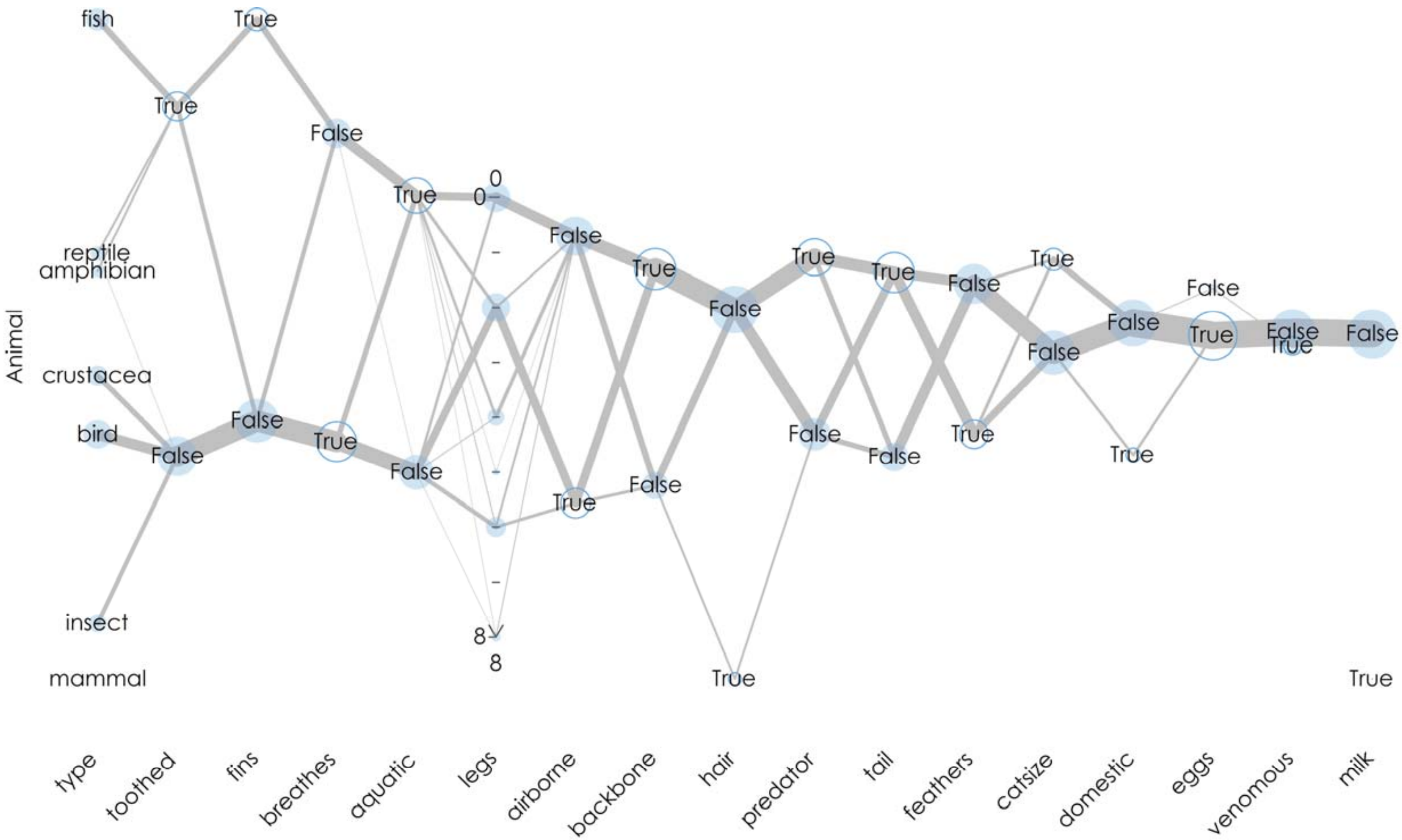
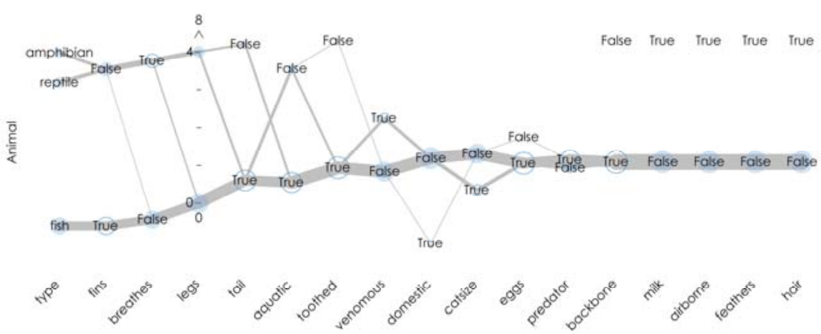
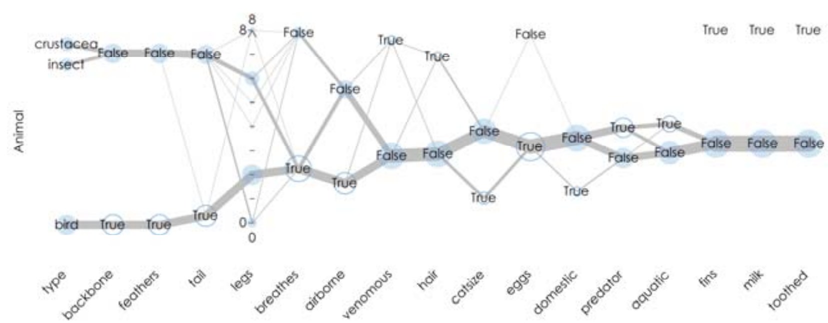
- Case
  - 101 animals
  - Invalid cases
    - Two frogs
    - Girl?
- Response
  - Animal type
    - Mammal
    - Reptile (爬虫類)
    - Amphibian (両生類)
    - Fish
    - Insect
    - Bird
    - Invertebrate (無脊椎動物)

- 16 Covariates (binary)
  - Hair/Feathers/Eggs/Milk/Airborne/Aquatic/Predator/Toothed/Backbone/Breathes/Venomous/Fins/Legs/Tail/Domestic/Cat-size

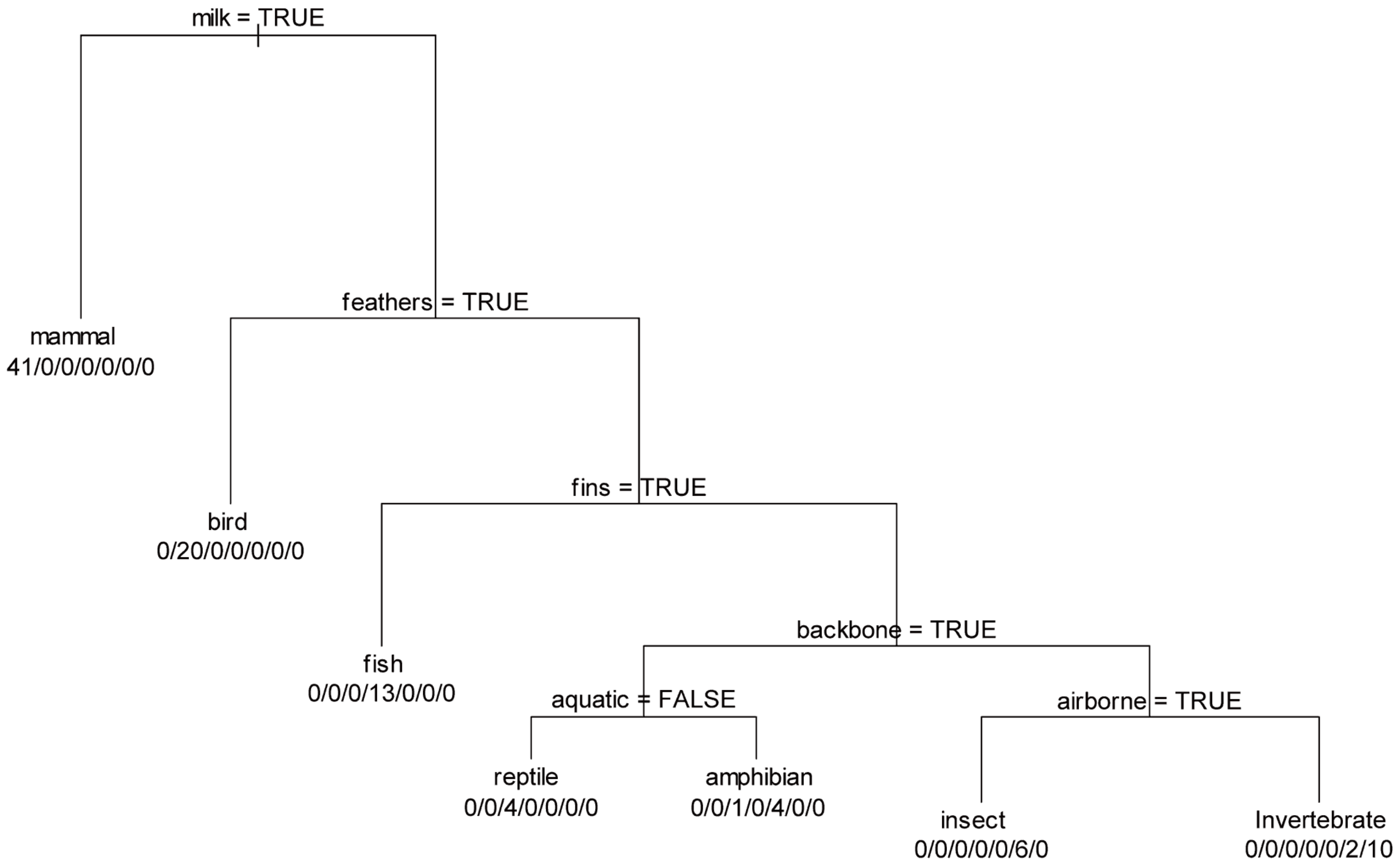
- Analysis goals:
  - What features best classify animals by type?
  - How are the features related?



Animal
crustacea
bird
insect
fish
amphibian
reptile
mammal
False
True
0
8
type
milk
eggs
hair
toothed
catsize
backbone
feathers
breathes
airborne
aquatic
tail
legs
venomous
domestic
fins
predator

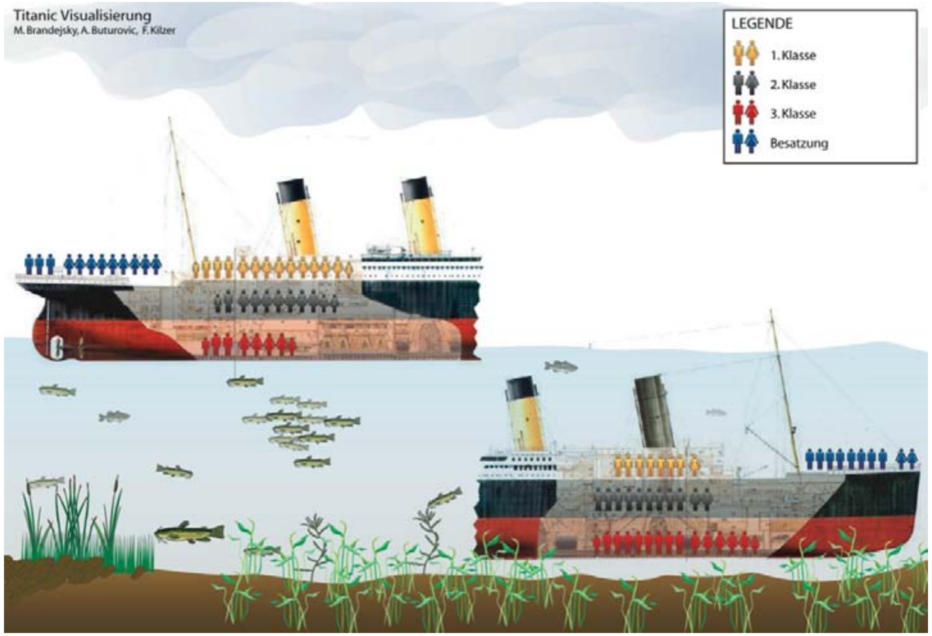

Animal
fish
reptile
amphibian
crustacea
bird
insect
mammal
True
False
0
8
type
toothed
fins
breathes
aquatic
legs
airborne
backbone
hair
predator
tail
feathers
catsize
domestic
eggs
venomous
milk

Animal
amphibian
reptile
fish
type
fins
breathes
legs
tail
aquatic
toothed
venomous
domestic
catsize
eggs
predator
backbone
milk
airborne
feathers
hair

Animal
crustacea
insect
bird
type
backbone
feathers
tail
legs
breathes
airborne
venomous
hair
catsize
eggs
domestic
predator
aquatic
fins
milk
toothed

# Relation to Partitioning Tree



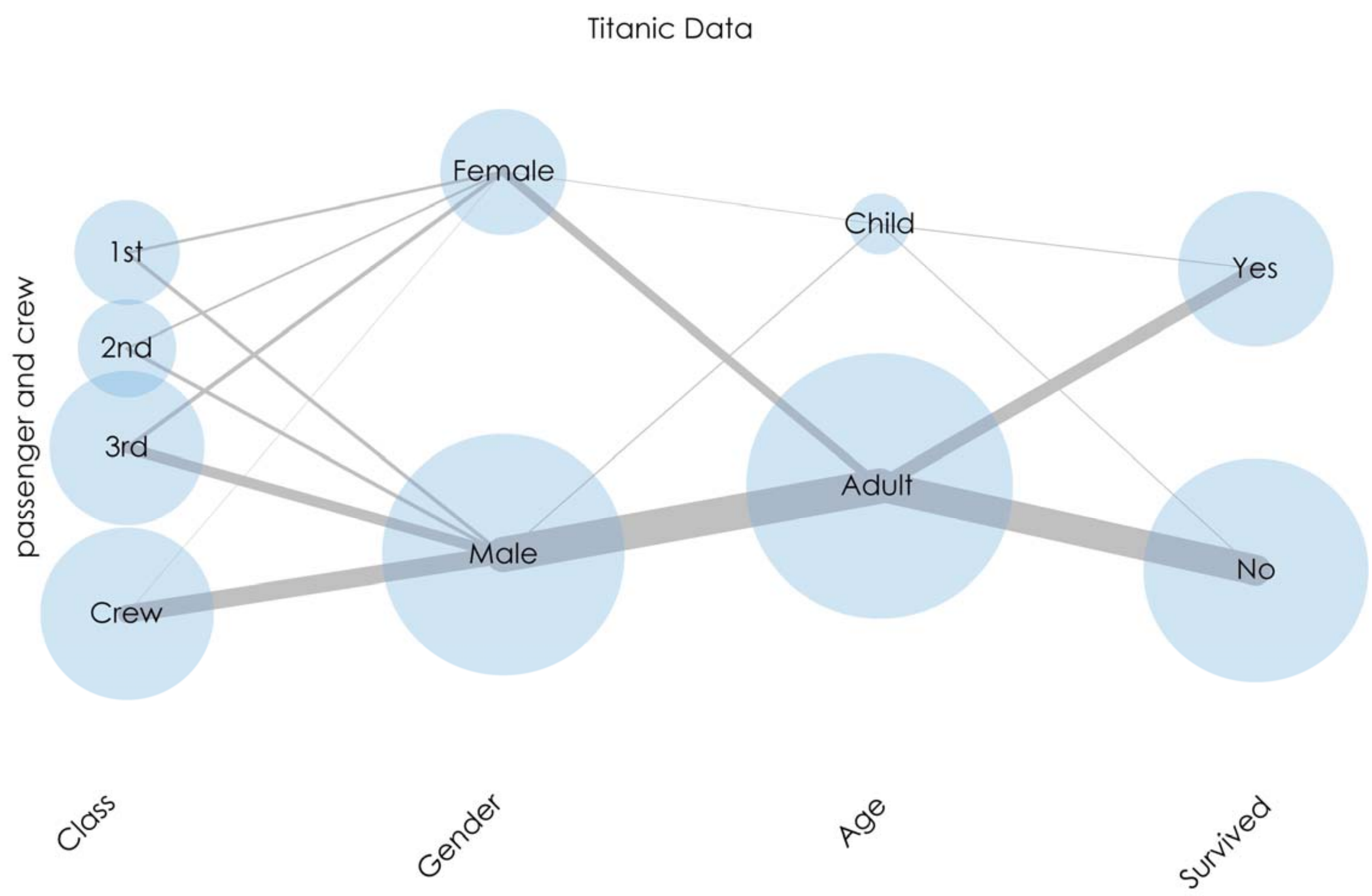
# Titanic Dataset (British Board of Trade 1990)

- Case
  - 2201 passengers and crew
- Variable
  - Class (First, Second, Third, Crew)
  - Age (Young, Old)
  - Gender (Male, Female)
  - Survived (Yes, No)




http://eagereyes.org/

# TEXTILE PLOT OF TITANIC DATASET

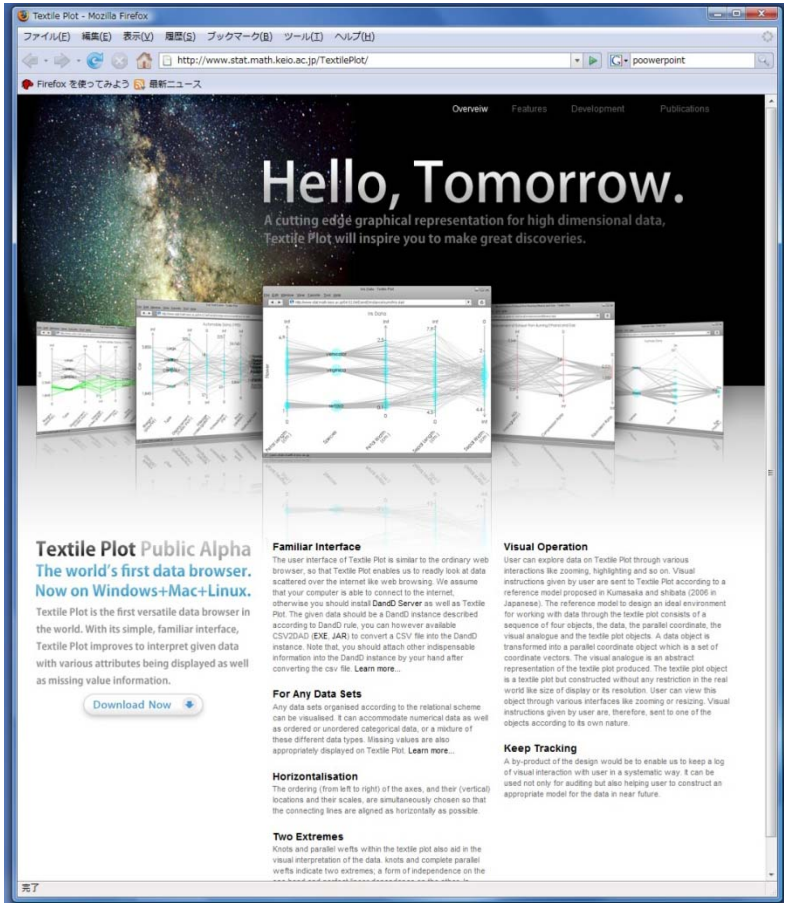


# Summary: Textile plot for Multivariate Categorical Datasets

- Textile plots provide rough idea of classifying cases
- Textile plots of multivariate categorical data emphasise absolute numbers
- Detailed conditional probability is difficult to interpret

# SOFTWARE: TEXTILE PLOT ENVIRONMENT

- Network ready
  - Based on DandD Client Server System (Yokouchi and Shibata 2004)
- Cross-platform
  - JAVA JRE 1.5
- Interactive user interfaces
  - Reference model (Kumasaka and Shibata, 2007)




http://stat.math.keio.ac.jp/TextilePlot/

# Conclusions

- Textile plot
  - Show an overview of the given data in an optimal way
  - Accommodate numerical and categorical data
  - Suggest several avenues for further exploratory or confirmatory data analysis
- Three eyes are better than one
  - Further investigations should be carried out with other graphical representations

# Bibliography


Anderson, E. (1957). A semigraphical method for the analysis of complex problems. *Proceedings of the National Academy of Sciences* **13** 923-927.

British Board of Trade (1990), *Report on the Loss of the `Titanic' (S.S.)*. British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing.

Chambers, J.M., and Hastie, T.J. (1992) *Statistical Models in S*, Wadsworth and Brooks/Cole, Pacific Grove CA.

Chen, C. H. (2002) Generalized Association Plots: Information Visualization via Iteratively Generated Correlation Matrices. *Statistica Sinica* **12** pp. 7--29.

Chen, C. H., haerdle, W. and Unwin, A.R. (2008) *Handbook of Data Visualization*, Springer, Berlin.

Chernoff, H. (1973) The use of faces to represent points in k-dimensional space graphically, *Journal of American Statistical Association* **68** 361-368.

Cleveland, W.S. and McGill, R. (1984). The Many Faces of a Scatterplot, *Journal of the American Statistical Association* **79** 807-822.

Cook, D. and Swayne, D. Interactive and dynamic graphics for Data Analysis, Springer New York.

DandD Project (2007) Home Page, http://www.stat.math.keio.ac.jp/DandD/.

Decathlon 2000 (2008) Decathlon 2000 Home Page, www.decathlon2000.ee.

Fienberg, S.E. (1979) Graphical Methods in Statis-tics, *The American Statistician* **33** 165-178.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal components analysis. *Biometrics* , **58** (3) pp. 453-467.

Hartigan, J. A. and Kleiner, B. (1981) Mosaics for Contingency Tables, *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, Springer-Verlag, pp.268-273.

# BIBLIOGRAPHY


Hurley, C. B. (2004) Clustering Visualizations of Multidimensional Data, *Journal of Computational and Graphical Statistics* **13** 788-806.

Inselberg, A. (1985) The plane with parallel coordinates, *The Visual Computer* **1** 69-91.

Kumasaka, N., Shibata, R. (2007) Textile Plot Environment, 統計数理特集号「統計データの可視化」, **55,** 47-68.

Kumasaka, N. and Shibata, R. (2008) High Dimesional Data Visualisation: the Textile Plot, *Computational Statistics & Data Analysis*, Submitted, **52**, 3616-3644.

Kumasaka, N. ANOVA on Textile Plot, *Proceedings in COMPSTAT 2008*, submitted.

Liquid Assets (2008) Liquid Assets Home Page, http://www.liquidasset.com/.

Unwin, A.R., Hawkins, G., Hofmann, H., and Siegl, B. (1996) *Interactive Graphics for Data Sets with Missing Values* – MANET, Journal of Comp and graph Stat, 5 113-122.

Wegman, E. (1990) Hyperdimensional data analysis using parallel coordinates. *Journal of The American Statistical Association* **85** 664-675.

Yokouchi, D. and Shibata, R. (2004), DandD: Client Server System, *Proceedings in COMPSTAT 2004*, Physica-Verlag.