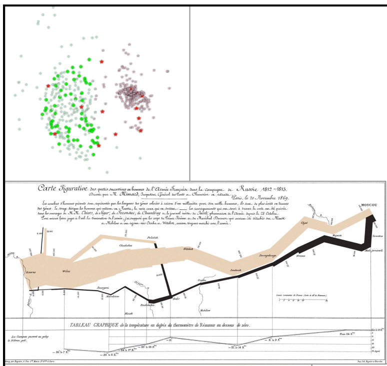
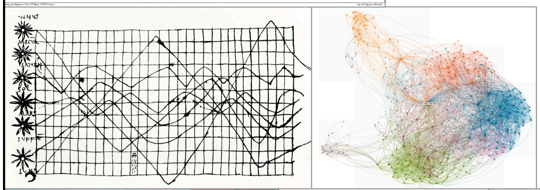
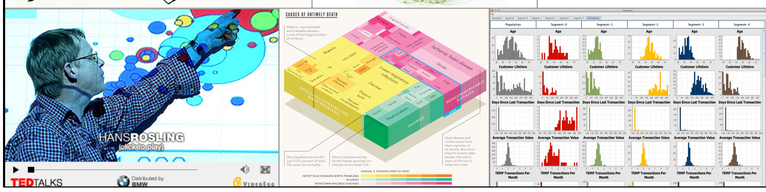






Visualising Multivariate Data

Information Visualisation COMP40610

Dr. Brian Mac Namee

Origins

This course curates material from multiple online and published sources

When this is the case full citations will be given

Agenda

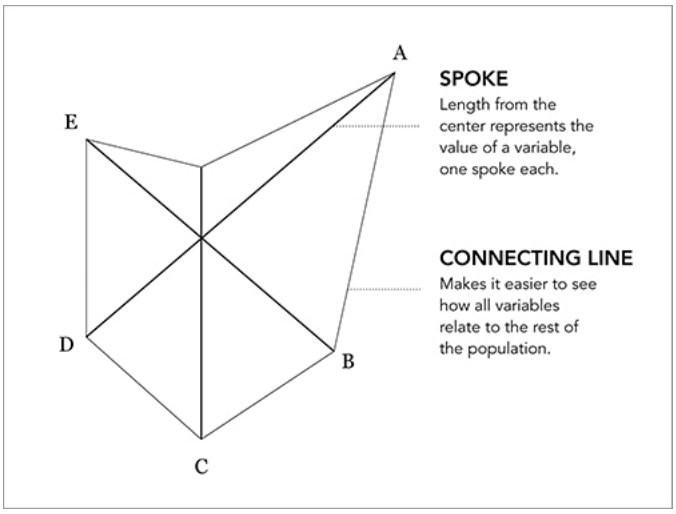
Visualising multi-variate data in two or three dimensions is difficult

We will look at a number of approaches:

- Point-Based Techniques
- Line-Based Techniques
- Dimensionality Reduction Techniques

POINT-BASED TECHNIQUES

Star Charts

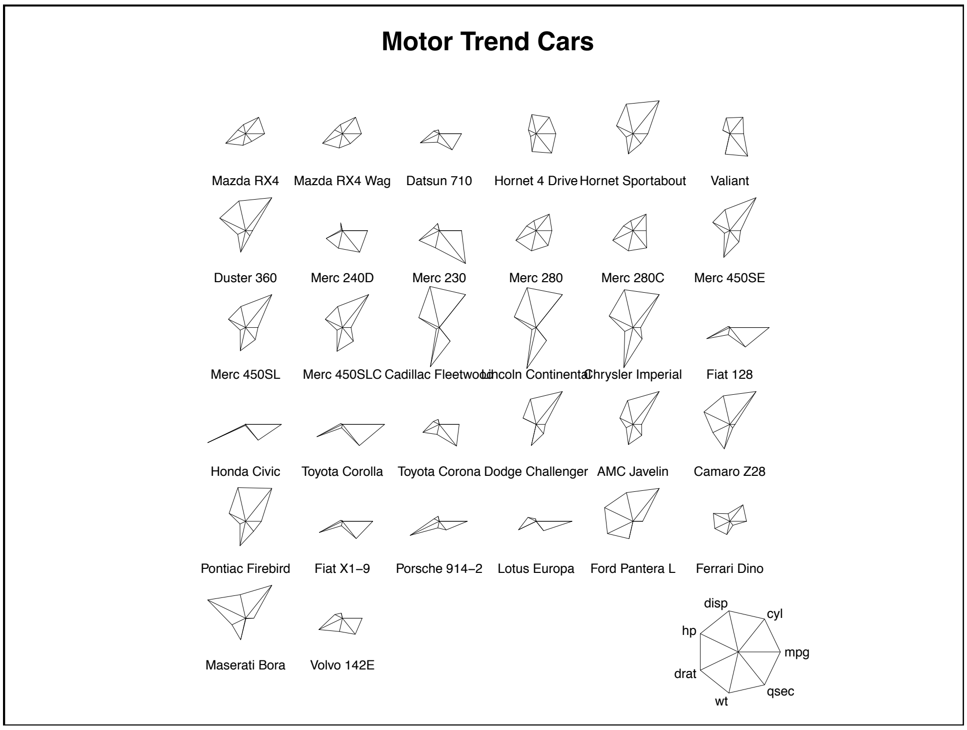


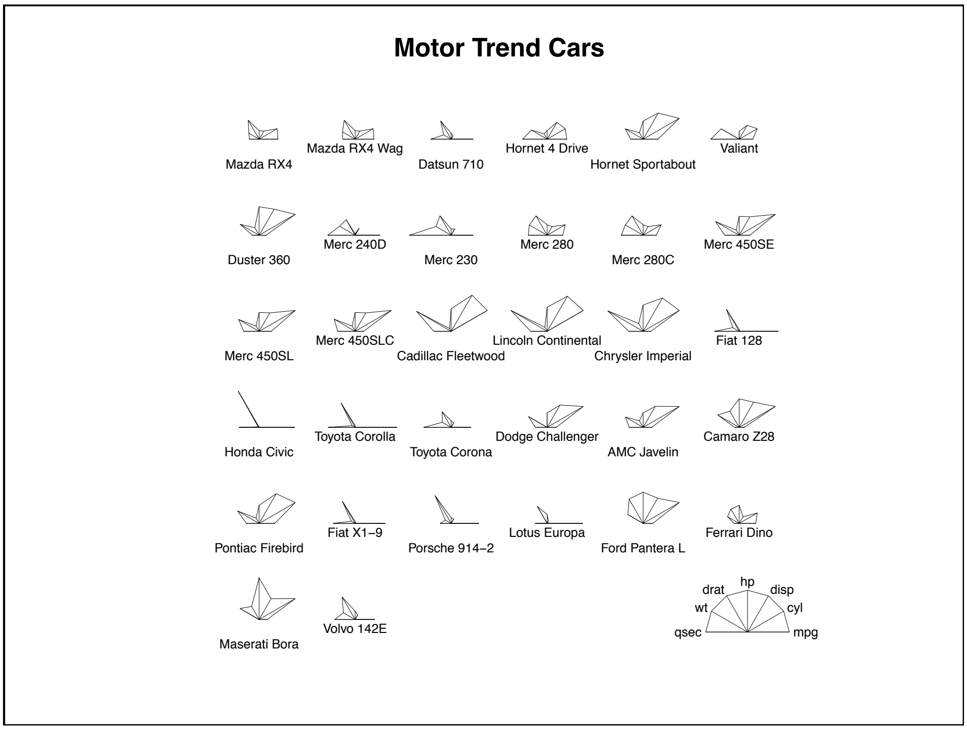
"Visualize This", N. Yau, Wiley, 2011

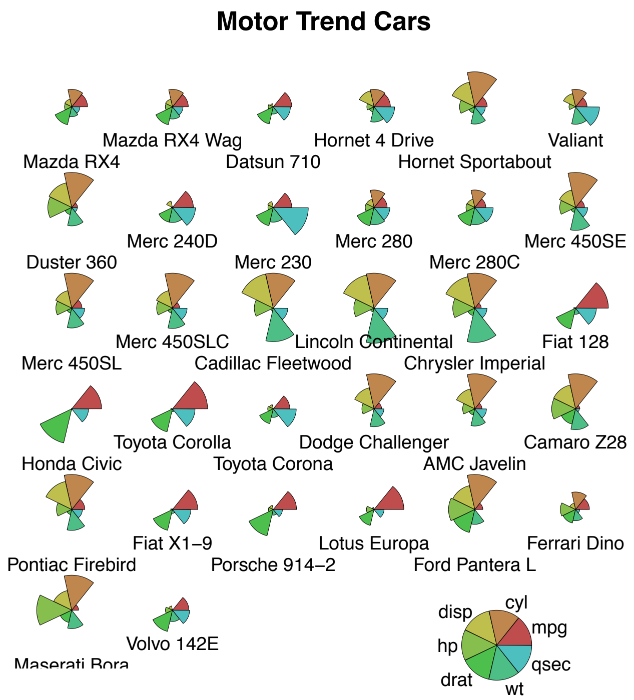
http://shop.oreilly.com/product/0636920022060.do

Car	mpg	cyl	disp	hp	drat	wt	qsec
Mazda RX4	21.0	6	160	110	3.9	2.62	16.46
Mazda RX4 Wag	21.0	6	160	110	3.9	2.875	17.02
Datsun 710	22.8	4	108	93	3.85	2.32	18.61
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02
Valiant	18.1	6	225	105	2.76	3.46	20.22
Duster 360	14.3	8	360	245	3.21	3.57	15.84
Merc 240D	24.4	4	146.7	62	3.69	3.19	20
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98

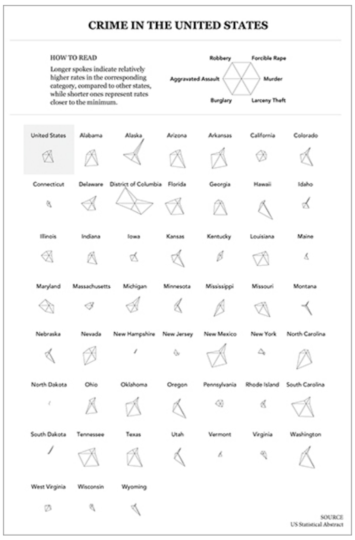
mtcars dataset (from R)


Motor Trend Cars
Mazda RX4
Mazda RX4 Wag
Datsun 710
Hornet 4 Drive
Hornet Sportabout
Valiant
Duster 360
Merc 240D
Merc 230
Merc 280
Merc 280C
Merc 450SE
Merc 450SL
Merc 450SLC
Cadillac Fleetwood
Lincoln Continental
Chrysler Imperial
Fiat 128
Honda Civic
Toyota Corolla
Toyota Corona
Dodge Challenger
AMC Javelin
Camaro Z28
Pontiac Firebird
Fiat X1−9
Porsche 914−2
Lotus Europa
Ford Pantera L
Ferrari Dino
Maserati Bora
Volvo 142E
disp
cyl
hp
mpg
drat
qsec
wt

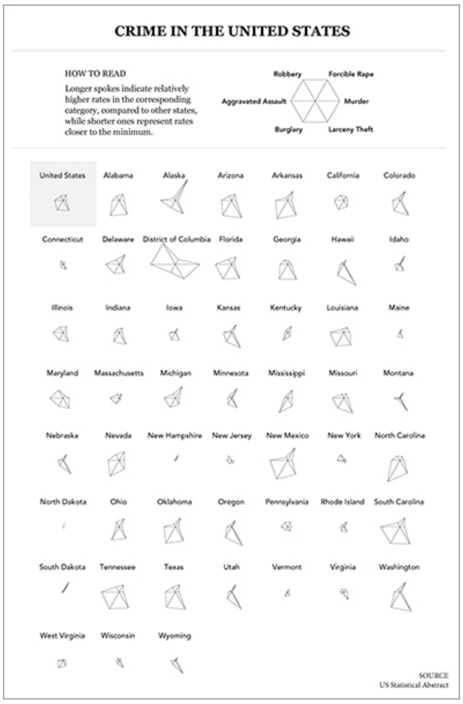

Motor Trend Cars
Mazda RX4
Mazda RX4 Wag
Datsun 710
Hornet 4 Drive
Hornet Sportabout
Valiant
Duster 360
Merc 240D
Merc 230
Merc 280
Merc 280C
Merc 450SE
Merc 450SL
Merc 450SLC
Cadillac Fleetwood
Lincoln Continental
Chrysler Imperial
Fiat 128
Honda Civic
Toyota Corolla
Toyota Corona
Dodge Challenger
AMC Javelin
Camaro Z28
Pontiac Firebird
Fiat X1−9
Porsche 914−2
Lotus Europa
Ford Pantera L
Ferrari Dino
Maserati Bora
Volvo 142E
hp
drat
disp
wt
cyl
qsec
mpg



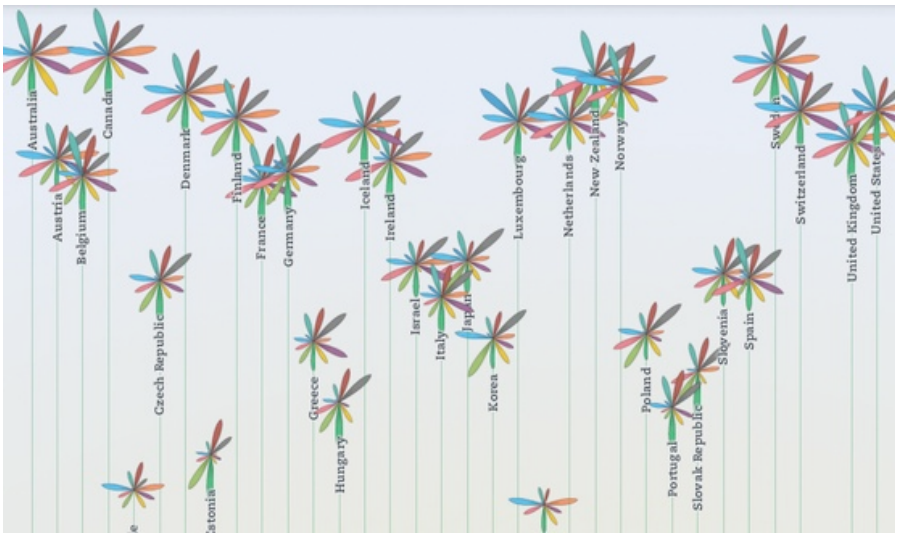
Star Charts



“Visualize This”, N. Yau, Wiley, 2011
http://shop.oreilly.com/product/0636920022060.do

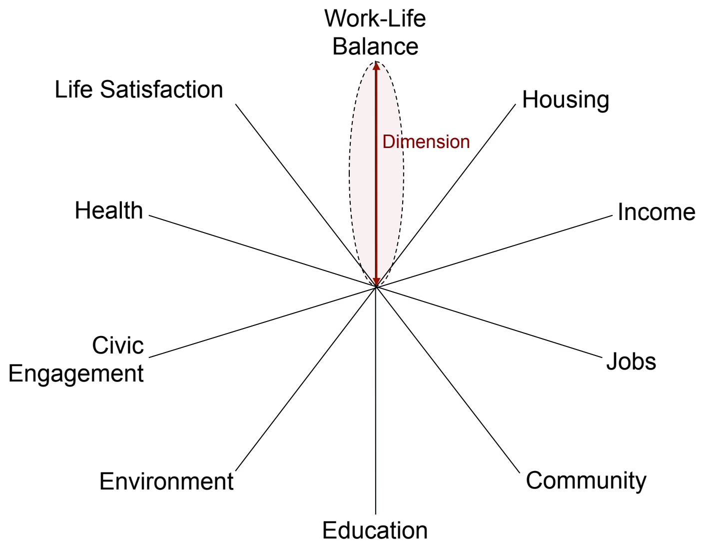


Star Charts



OECD Better Life Index: http://moritz.stefaner.eu/projects/oecd-better-life-index/

Star Charts

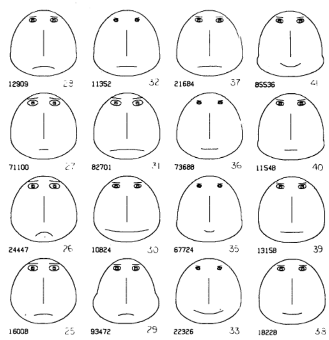


OECD Better Life Index: http://moritz.stefaner.eu/projects/oecd-better-life-index/

Chernoff Faces

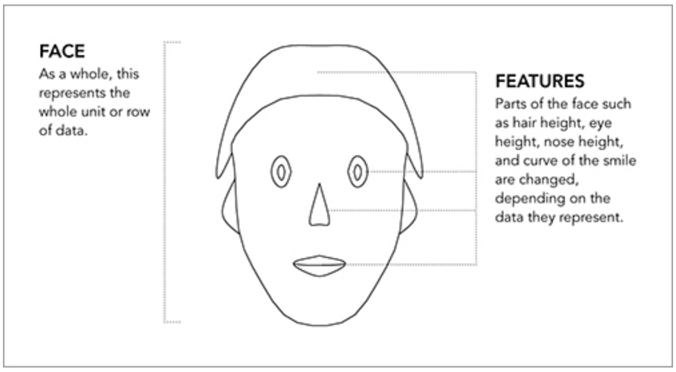
Chernoff faces are a novel approach to visualising multivariate data

Lots of interesting argument about how useful they are

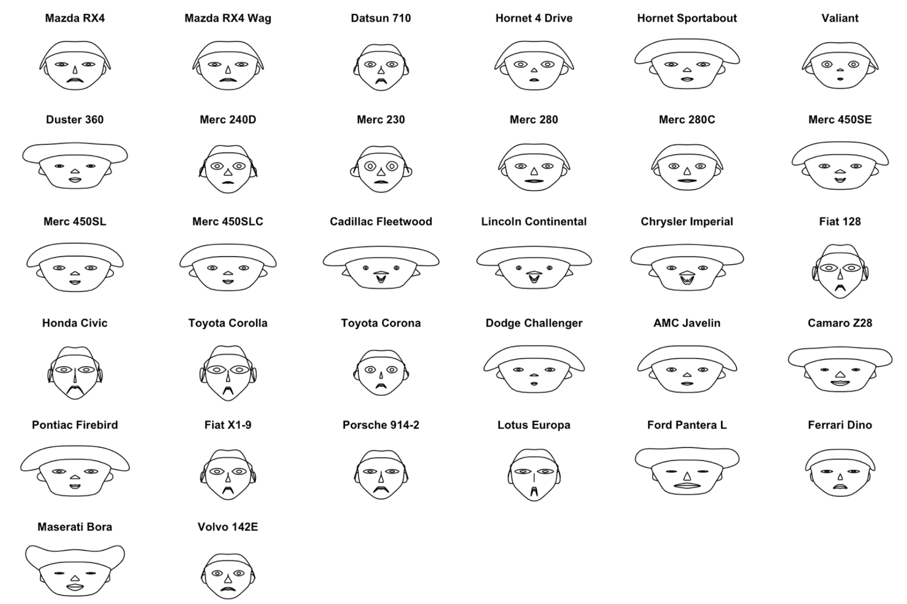


“The Use of Faces to Represent Points in K-Dimensional Space Graphically" Herman Chernoff, Journal of the American Statistical Association, vol 68 (342), pp 361–368, 1973.
http://wexler.free.fr/library/files/chernoff%20%281973%29%20the%20use%20of%20faces%20to%20represent%20points%20in%20k-dimensional%20space%20graphically.pdf

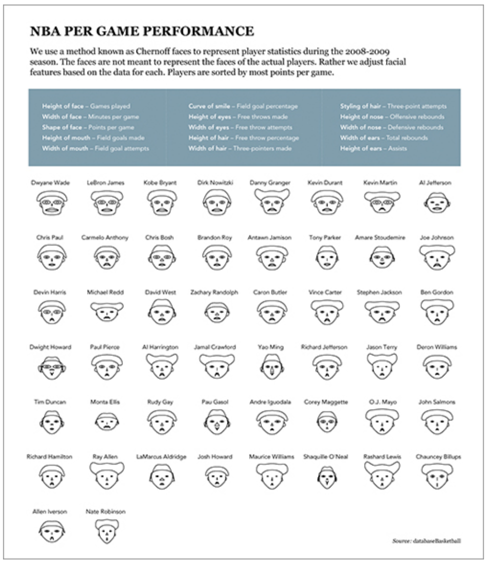
Chernoff Faces



"Visualize This", N. Yau, Wiley, 2011
http://shop.oreilly.com/product/0636920022060.do

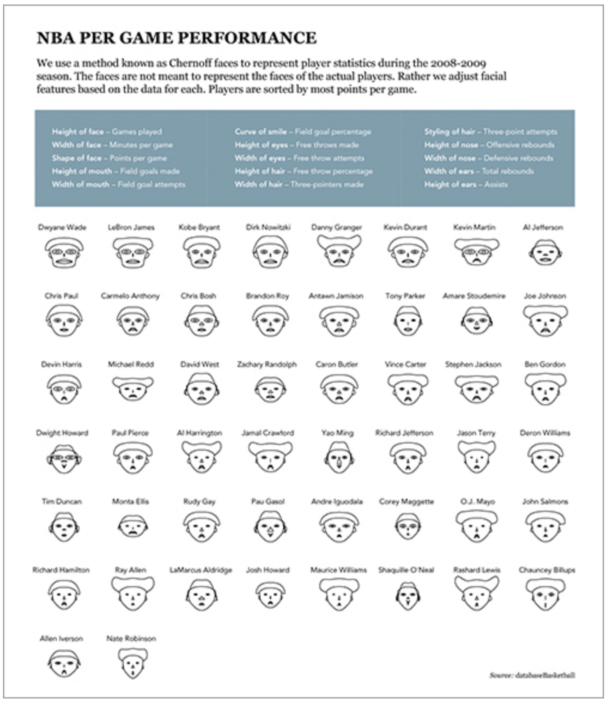


Example:



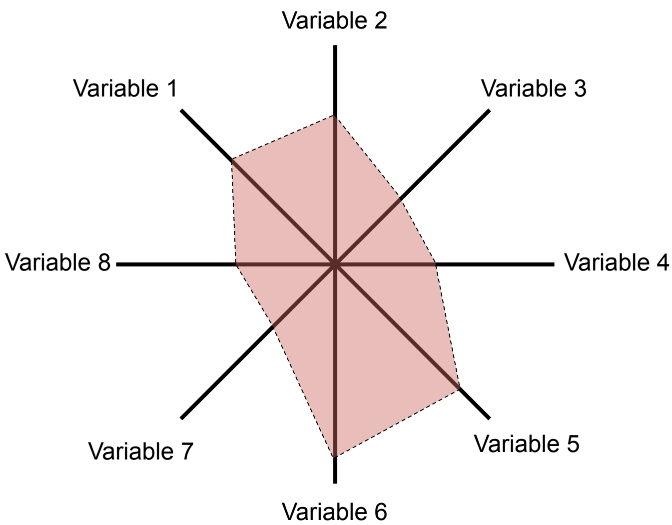
"Visualize This", N. Yau, Wiley, 2011

http://shop.oreilly.com/product/0636920022060.do

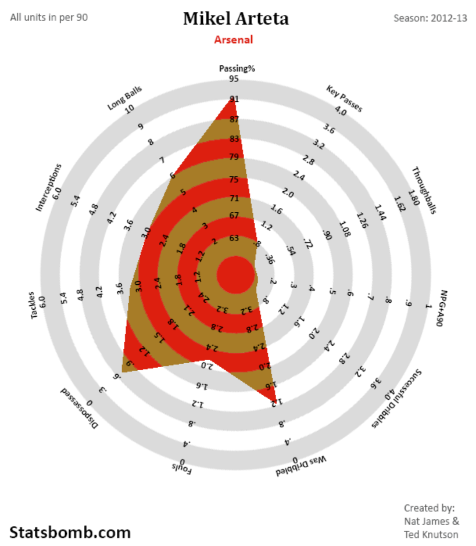


LINE-BASED TECHNIQUES

Radar Charts

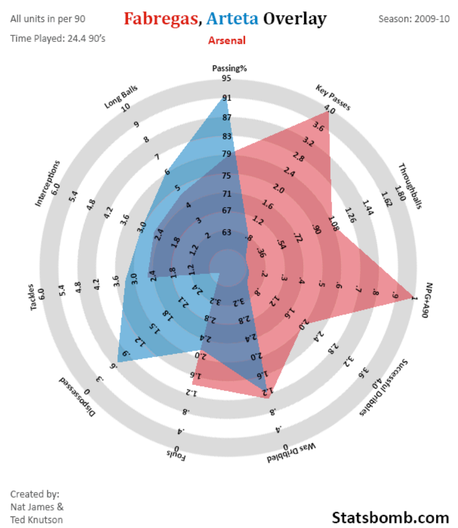


Radar Charts



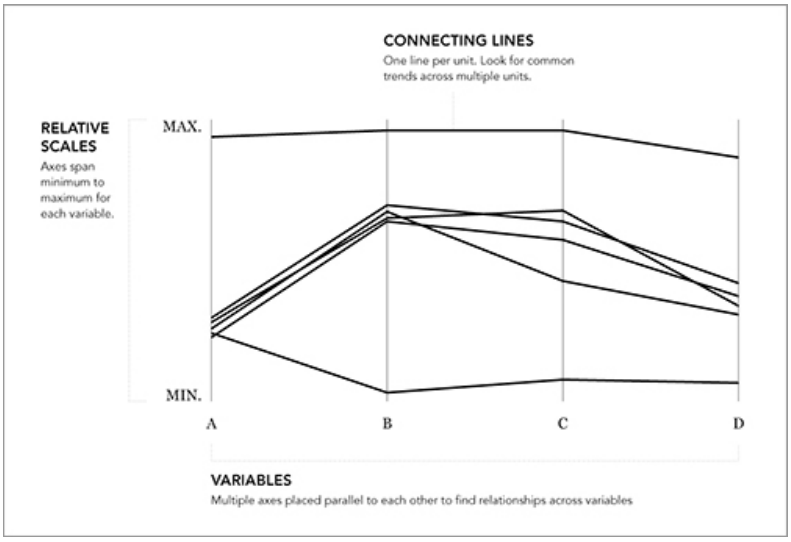
Comparing Arsenal Midfielders + Explaining CM Radar Charts
http://statsbomb.com/2014/01/comparing-arsenal-midfielders-explaining-cm-radar-charts/

Radar Charts



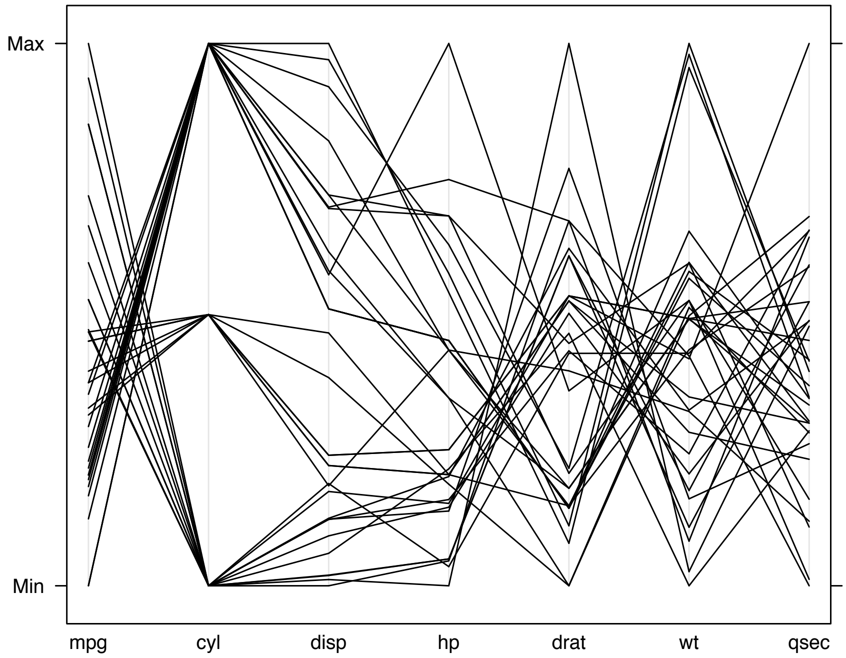
Comparing Arsenal Midfielders + Explaining CM Radar Charts
http://statsbomb.com/2014/01/comparing-arsenal-midfielders-explaining-cm-radar-charts/

Parallel Coordinates

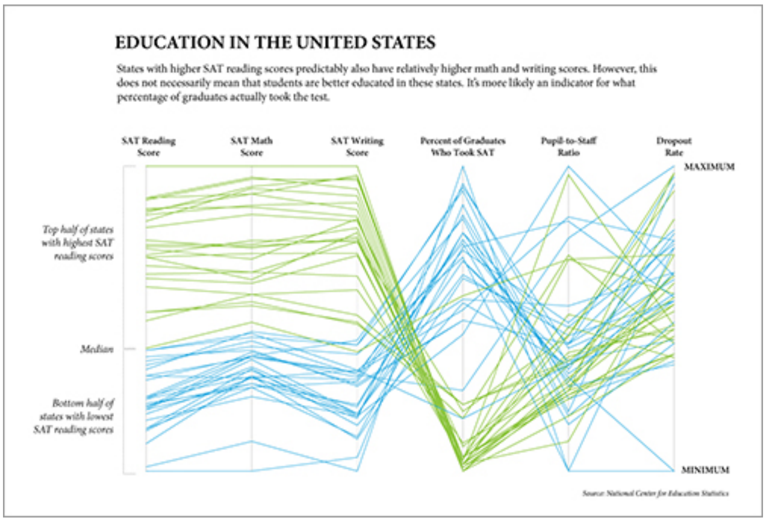


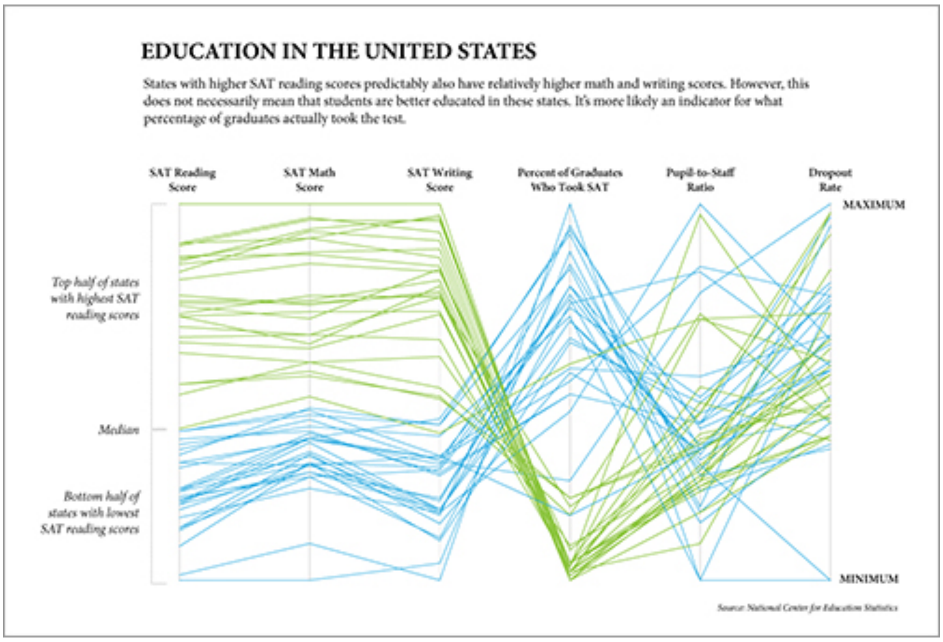
"Visualize This", N. Yau, Wiley, 2011

http://shop.oreilly.com/product/0636920022060.do

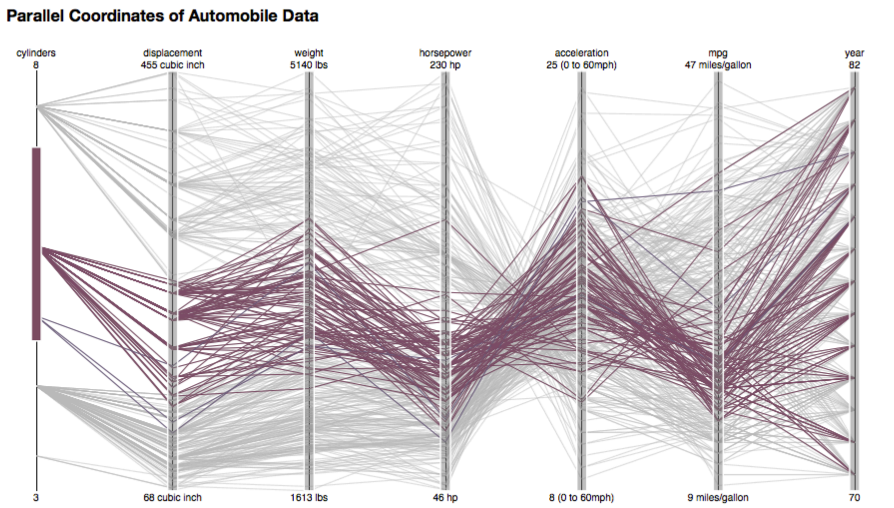


Example: Education Performance



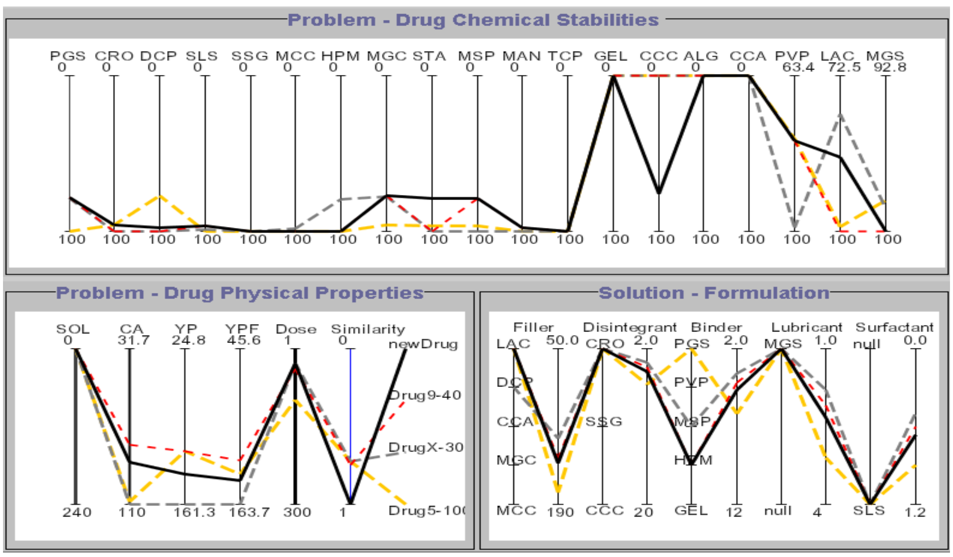


Parallel Coordinates: Best When Interactive



From GGobi: http://hci.stanford.edu/jheer/files/zoo/ex/stats/parallel.html

FormuCaseViz, Massie et al



Visualisation of Case-Based Reasoning for Explanation, S. Massie, S. Craw & N. Wiratunga, In Proceedings of the ECCBR 2004 Workshops, pp 135-144, 2004.

Nutrient Contents – Parallel Coordinates

An interactive visualization of the USDA Nutrient Database. For information on parallel coordinates, read this tutorial.

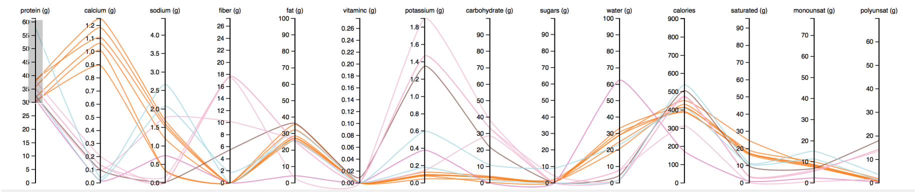
Hide Ticks | Dark | Shadows | Opacity: 75%

Per 100g of Food

Selected 14 rows | Keep | Remove | Export

Group Breakdown | Total Selected

Dairy and Egg Products, Fats and Oils, Poultry Products, Soups, Sauces, and Gravies, Vegetables and Vegetable Products, Sausages and Luncheon Meats, Breakfast Cereals, Fruits and Fruit Juices, Nut and Seed Products, Beverages, Finfish and Shellfish Products, Legumes and Legume Products, Baked Products, Sweets, Cereal Grains and Pasta, Fast Foods, Meals, Entrees, and Sidedishes, Snacks, Restaurant Foods



name	group	protein (g)	calcium ...	sodium ...	fiber (g)	vitaminc...	potassiu...	carbohy...	sugars (g)	fat (g)	water (g)	calories	saturate
:heese, goat, hard type	Dairy and Egg Products	30.52	0.895	0.346		0	0.048	2.17	2.17	35.59	29.01	452	24.60
:heese, gruyere	Dairy and Egg Products	29.81	1.011	0.336		0	0.081	0.36	0.36	32.34	33.19	413	18.91
:heese, parmesan, grated	Dairy and Egg Products	38.46	1.109	1.529		0	0.125	4.06	0.9	28.61	20.84	431	17.30
:heese, parmesan, hard	Dairy and Egg Products	35.75	1.184	1.602		0	0.092	3.22	0.8	25.83	29.16	392	16.41
:heese, parmesan, shredded	Dairy and Egg Products	37.86	1.253	1.696		0	0.097	3.41	0.9	27.34	25	415	17.37
:heese, romano	Dairy and Egg Products	31.8	1.064	1.2		0	0.086	3.63	0.73	26.94	30.91	387	17.11
eeds, cottonseed kernels, roasted (glandless)	Nut and Seed Products	32.59	0.1	0.025	5.5	0.009	1.35	21.9		36.29	4.65	506	9.699
ish, sturgeon, mixed species, smoked	Finfish and Shellfish Prod...	31.2	0.017	0.739		0	0.379			4.4	62.5	173	1.037
acon bits, meatless	Legumes and Legume Pro...	32	0.101	1.77	10.2	0.0019	0.145	28.6		25.9	8	476	4.055

Nutrient Contents - Parallel Coordinates http://bl.ocks.org/syntagmatic/raw/3150059/

Parallel Coordinates

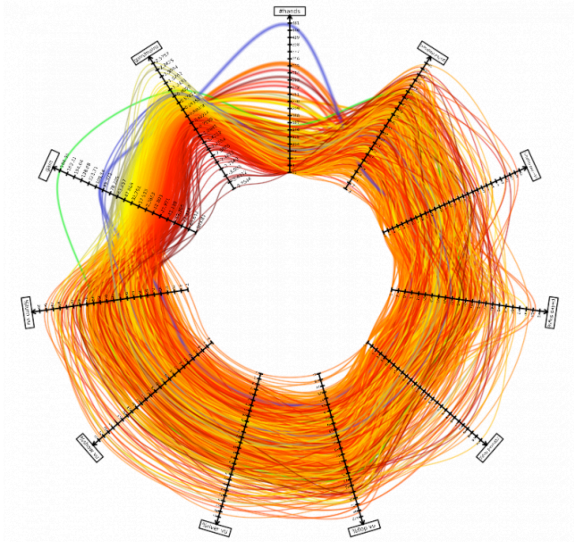
Parallel coordinates are a very good visualisation tool to help viewers explore multivariate relationships

Interaction (especially brushing) is required for parallel coordinate plots to be really useful

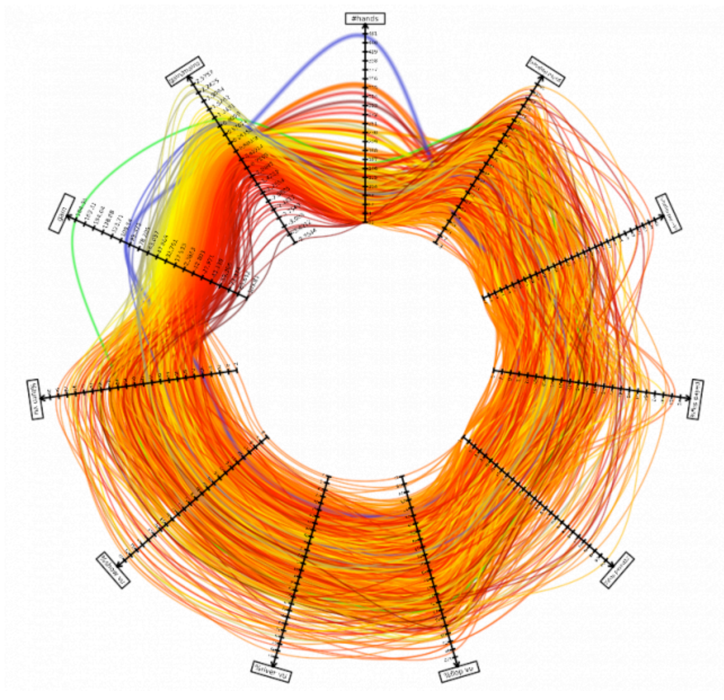
See paper below for a detailed exploration of the usefulness of parallel coordinates

"Perceiving Patterns in Parallel Coordinates: Determining Thresholds for Identification of Relationships", J Johansson, C. Forsell, M. Lind, M. Cooper, Information Visualization, vol. 7, no. 2, 152-162, 2008.
http://webstaff.itn.liu.se/~jimjo/papers/IVJ08/IVJ08.pdf

Circular Parallel Coordinates



Tulip visualisation Toolkit: http://tulip.labri.fr/TulipDrupal/?q=node/321

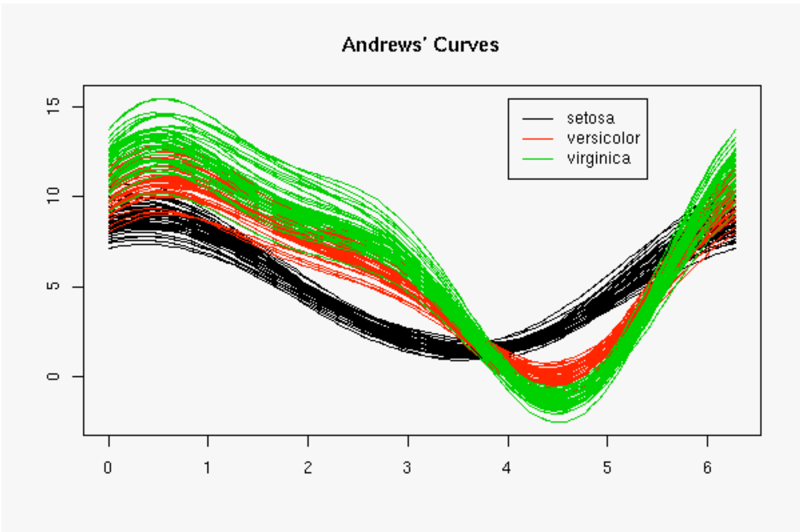


Andrews' Curves

Each multi-variate data point $D = (d_1, d_2, \ldots, d_N)$ is used to create a curve of the form:

$$f(t) = \frac{d_1}{\sqrt{2}} + d_2 \sin(t) + d_3 \cos(t) + d_4 \sin(2t) + d_4 \cos(2t) + \ldots$$

Andrews' Curves



DIMENSIONALITY REDUCTION APPROACHES

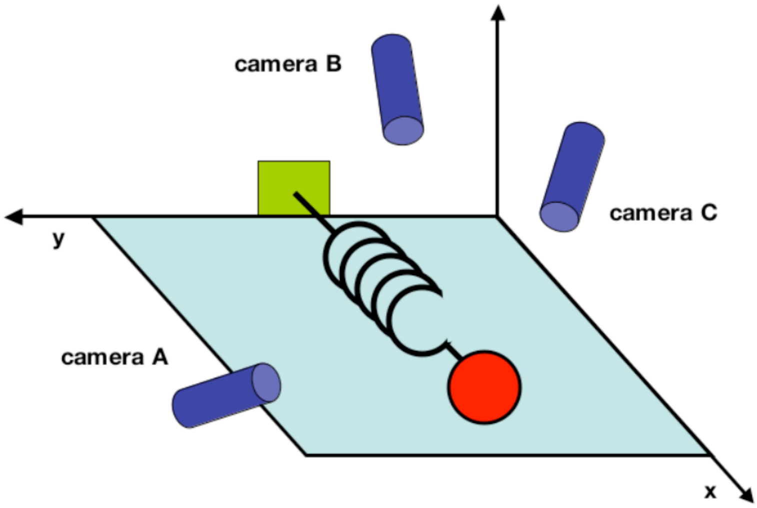
Principle Components Analysis

Principle Components Analysis (PCA) is a very common dimensionality reduction technique based on finding dimensions along which the maximum variance occurs in datasets

PCA is widely used for dimensionality reduction in statistics and machine learning

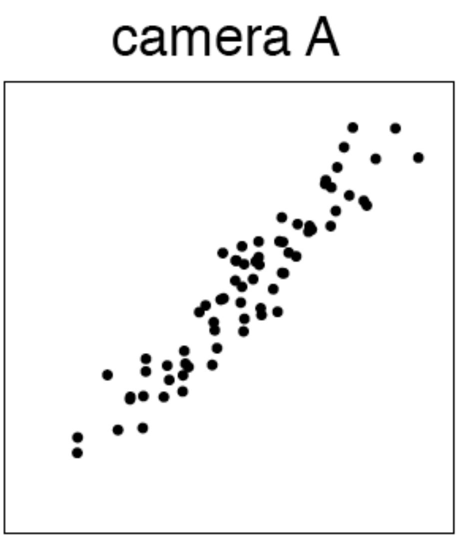
Also useful for visualizations

Explaining PCA



“A Tutorial on Principal Component Analysis”, Jonathon Shlens
https://arxiv.org/abs/1404.1100

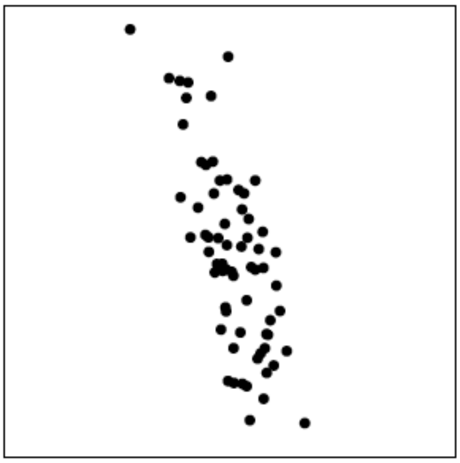
Explaining PCA



“A Tutorial on Principal Component Analysis”, Jonathon Shlens
https://arxiv.org/abs/1404.1100

Explaining PCA

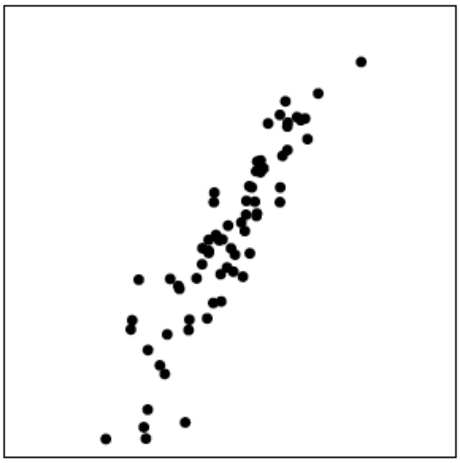
camera B



"A Tutorial on Principal Component Analysis", Jonathon Shlens
https://arxiv.org/abs/1404.1100

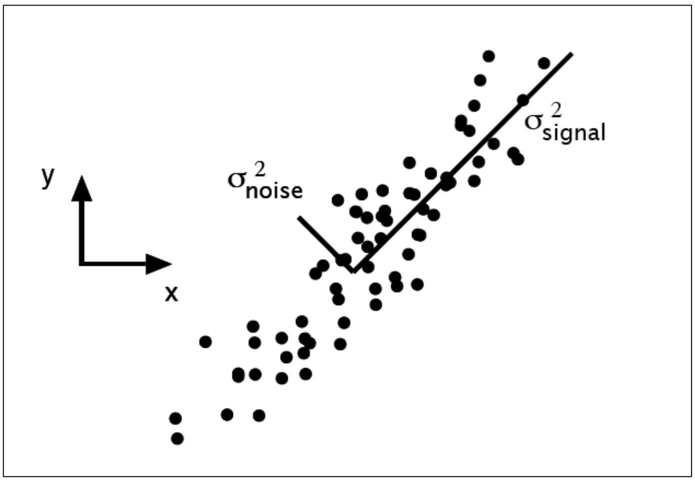
Explaining PCA

camera C



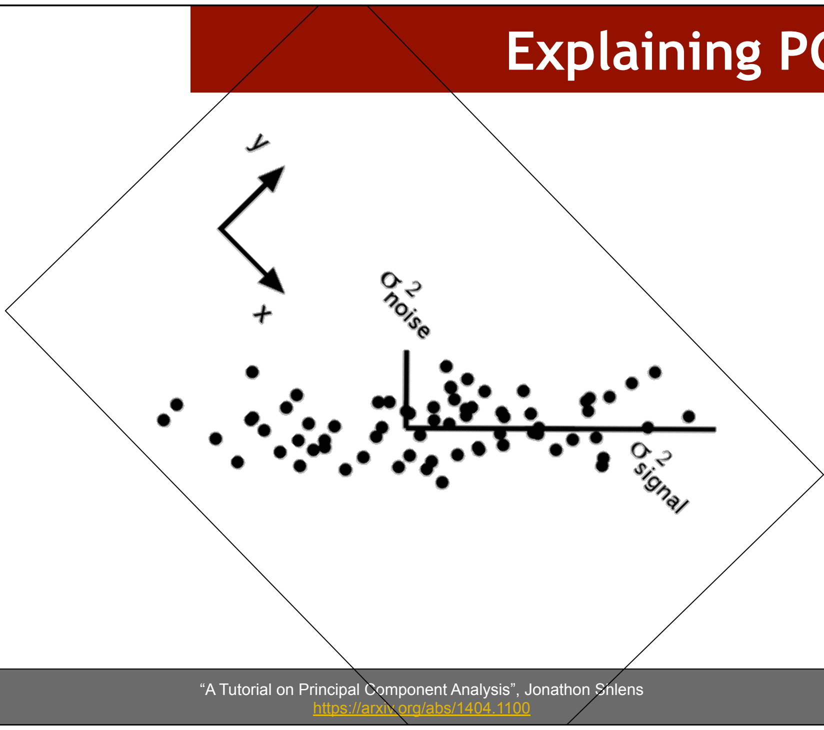
"A Tutorial on Principal Component Analysis", Jonathon Shlens
https://arxiv.org/abs/1404.1100

Explaining PCA



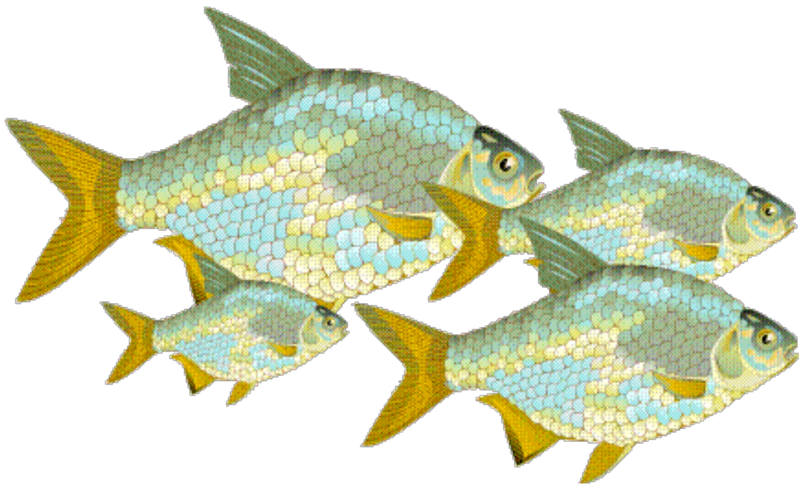
“A Tutorial on Principal Component Analysis”, Jonathon Shlens
https://arxiv.org/abs/1404.1100

Explaining PCA



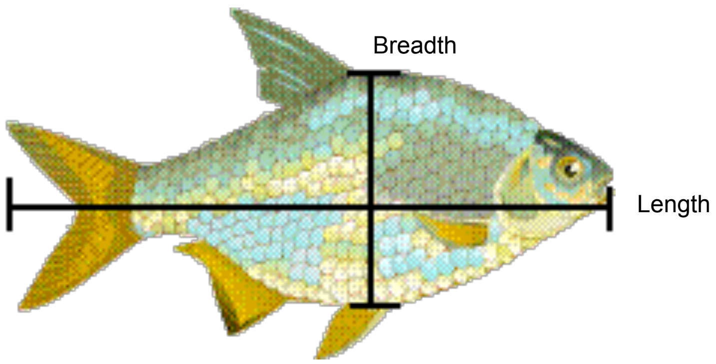
“A Tutorial on Principal Component Analysis”, Jonathon Shlens
https://arxiv.org/abs/1404.1100

Explaining PCA



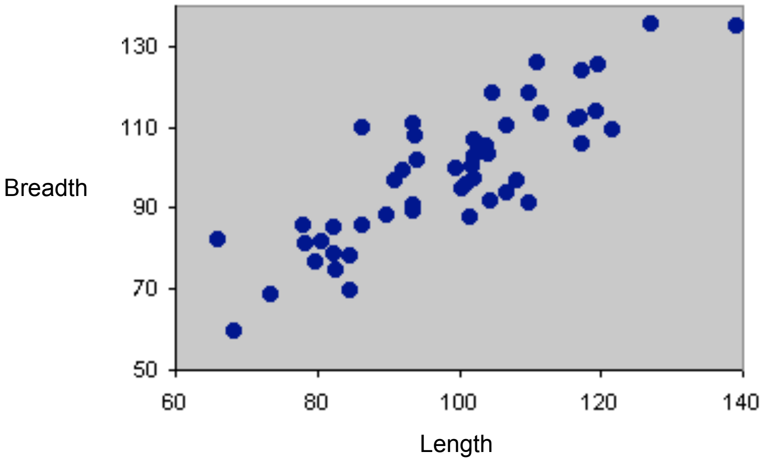
PCA can also be explained by a graphical method:
http://www.alanfielding.co.uk/multivar/pca_graf.htm

Explaining PCA



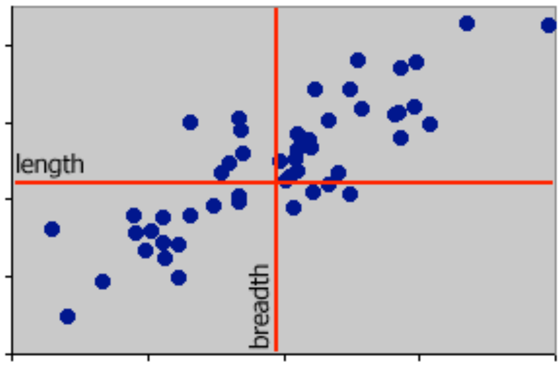
PCA can also be explained by a graphical method:
http://www.alanfielding.co.uk/multivar/pca_graf.htm

Explaining PCA



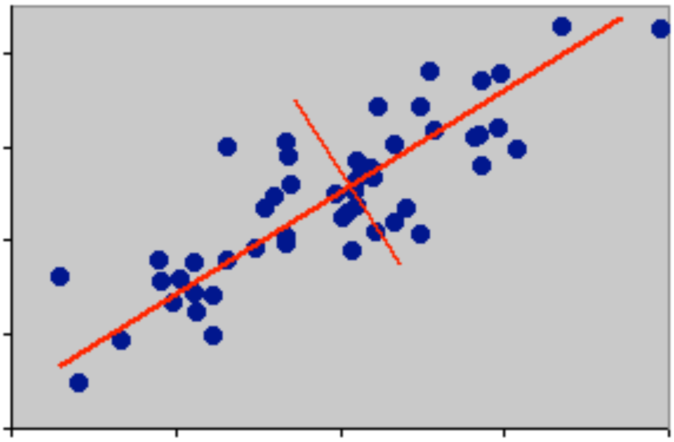
PCA can also be explained by a graphical method:
http://www.alanfielding.co.uk/multivar/pca_graf.htm

Explaining PCA



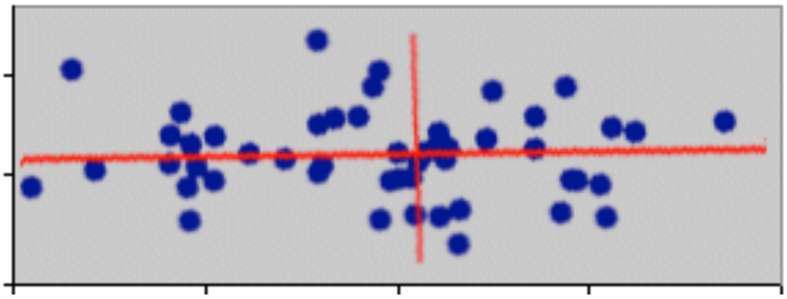
PCA can also be explained by a graphical method:
http://www.alanfielding.co.uk/multivar/pca_graf.htm

Explaining PCA



PCA can also be explained by a graphical method:
http://www.alanfielding.co.uk/multivar/pca_graf.htm

Explaining PCA



size = length + breadth

size = 0.75 x length + 0.25 x breadth

PCA can also be explained by a graphical method:
http://www.alanfielding.co.uk/multivar/pca_graf.htm

PCA Process

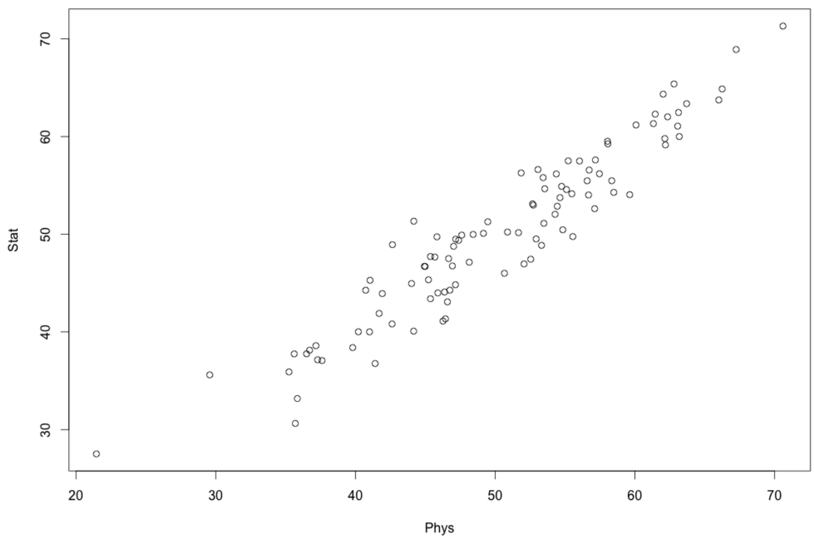
1. Center the data according to the mean.
2. Build the covariance matrix
3. Find the Eigenvalues and Eigenvectors for the above matrix
4. Plot the scaled data along with the Eigenvectors
5. Express the scaled data in terms of the Eigenvectors (principal components)
6. Plot the transformed data

PCA Process

Physics and statistics results for a group of students

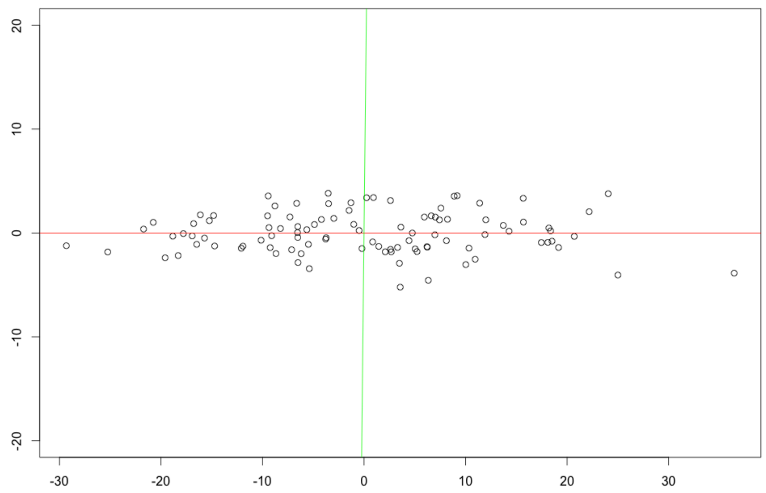
Phys	Stat
45.83286	49.734122
66.24545	64.873573
54.84116	50.454854
52.73309	52.980775
55.22804	57.510526
36.71771	38.129407
70.60068	71.311078
…	…

PCA Process

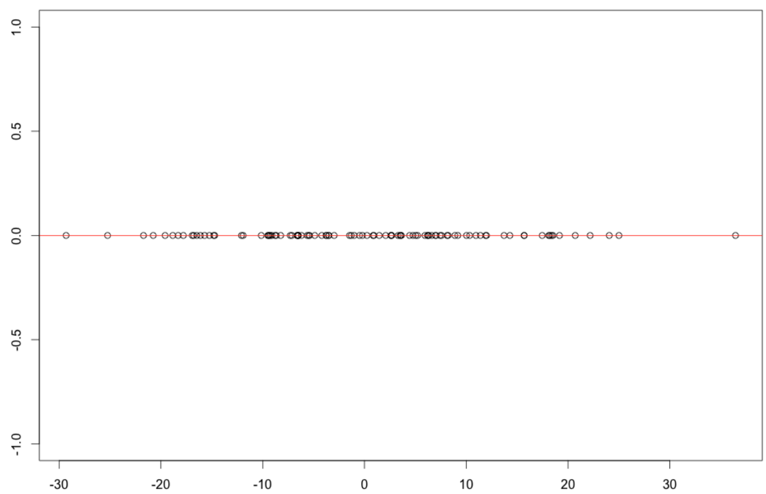


PCA Process

PCA Process



PCA Process



Iris PCA Example

Visualising Fisher's famous iris dataset

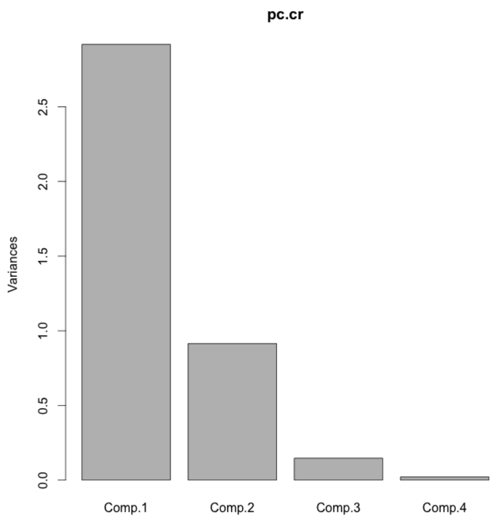
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	I. setosa
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
4.6	3.1	1.5	0.2	I. virginica
5.0	3.6	1.4	0.2	I. virginica
5.4	3.9	1.7	0.4	I. virginica
4.6	3.4	1.4	0.3	I. virginica
5.0	3.4	1.5	0.2	I. versicolor
4.4	2.9	1.4	0.2	I. versicolor
4.9	3.1	1.5	0.1	I. versicolor
5.4	3.7	1.5	0.2	I. versicolor

Iris PCA Example

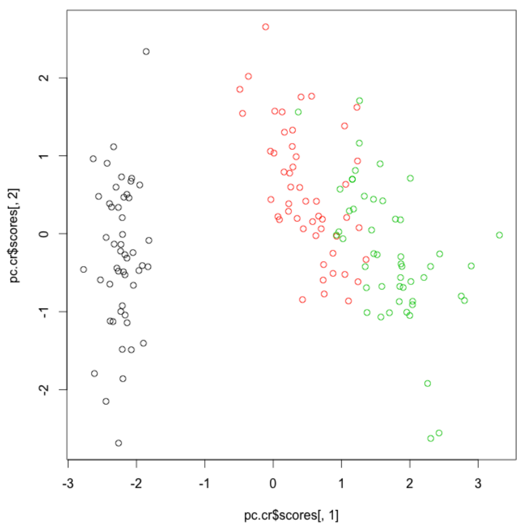
A simple R programme:

```
data("iris")
pc.cr <- princomp(iris[1:4], cor = TRUE)
summary(pc.cr)
plot(pc.cr)
biplot(pc.cr)
plot(pc.cr$scores[, 1], pc.cr$scores[, 2], col=iris[, 5])
```

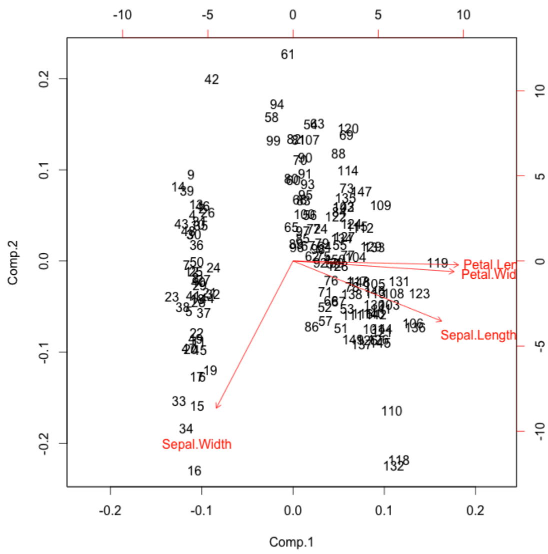
Iris PCA Example



Iris PCA Example



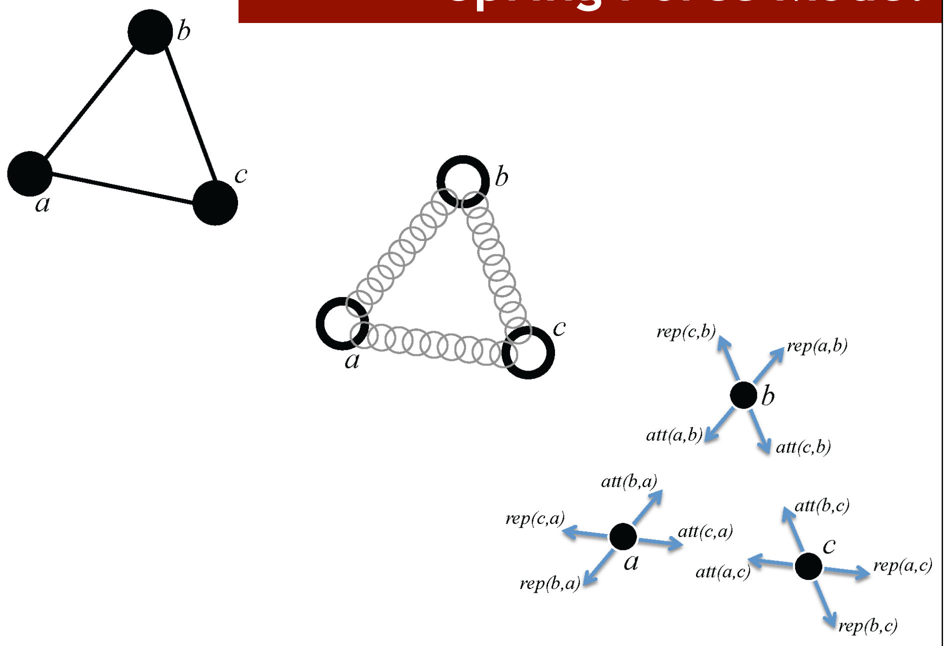
Iris PCA Example



Force Directed Graphs

Allow points arrange themselves within a two dimensional plane according to some set of simulated forces acting upon them

Spring Force Model



Spring Force Model

The total force exerted on any case *v*, *f(v)*, can be calculated as:

$$f(v) = \sum_{u \in N(v)} att(u,v) + \sum_{u \in N(v)} rep(u,v)$$

where *att(u, v)* is the attractive force exerted on *v* by case *u*; *rep(u, v)* is the repulsive force exerted on *v* by *u*; and *N(v)* is the set of vertices emanating from *v*

Spring Force Model

Attractive forces are dictated by the strength of the springs between cases and are calculated using Hooke's law

- *att(u, v)* is proportional to the distance between *u* and *v* and the zero energy length of the spring
- The zero energy length of a spring is the length at which the spring will exert no attractive forces and is directly proportional to the similarities between the cases it connects

The repulsive forces between cases are modelled as Newtonian gravitational forces, and so follow an inverse square law

Spring Force Model

So we can expand the previous equation as:

$$f_x(v) = \sum_{u \in N(v)} k_{att} * \frac{(dist(u,v) - zero(u,v)) * (v_x - u_x)}{dist(u,v)} + \sum_{u \in N(v)} k_{rep} * \frac{(u_x - v_x)}{dist(u,v)^3}$$

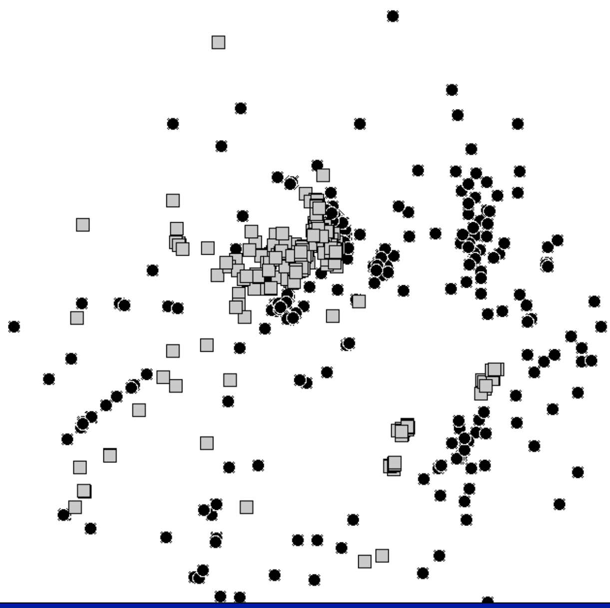
where *att(u, v)* is the attractive force where zero(u, v) is given as:

$$zero(u,v) = (1 - sim(u,v)) * maxZeroEnergyLength$$

Spring Force Model

The similarity between examples can be calculated using any suitable similarity metric – for example Euclidean distance

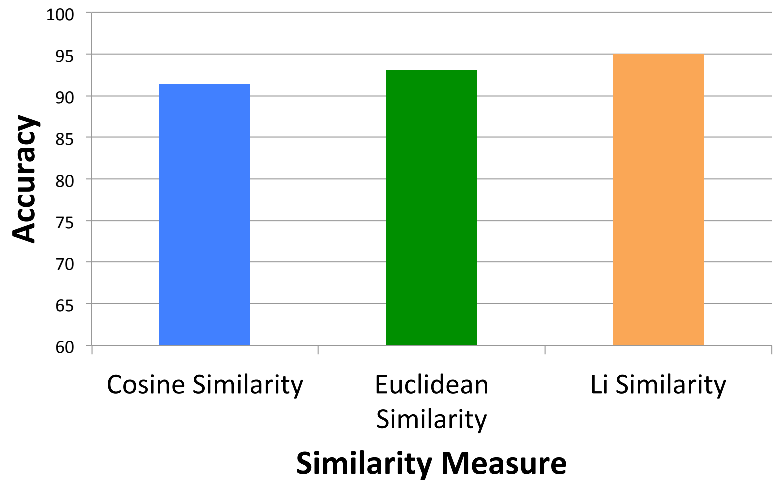
$$d = \sqrt{\sum_{i=1}^{k} (X_i - Y_i)^2}$$

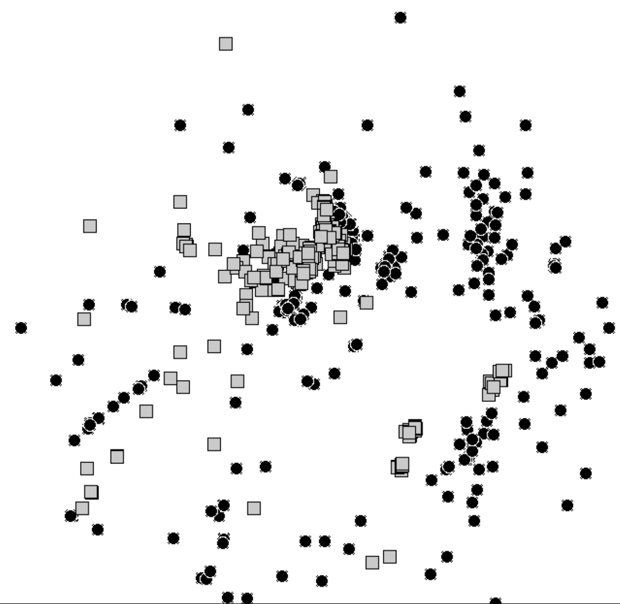


Visualisation of the Spam 2 Dataset

Instances: 500, # Features: 3,219, Cosine similarity

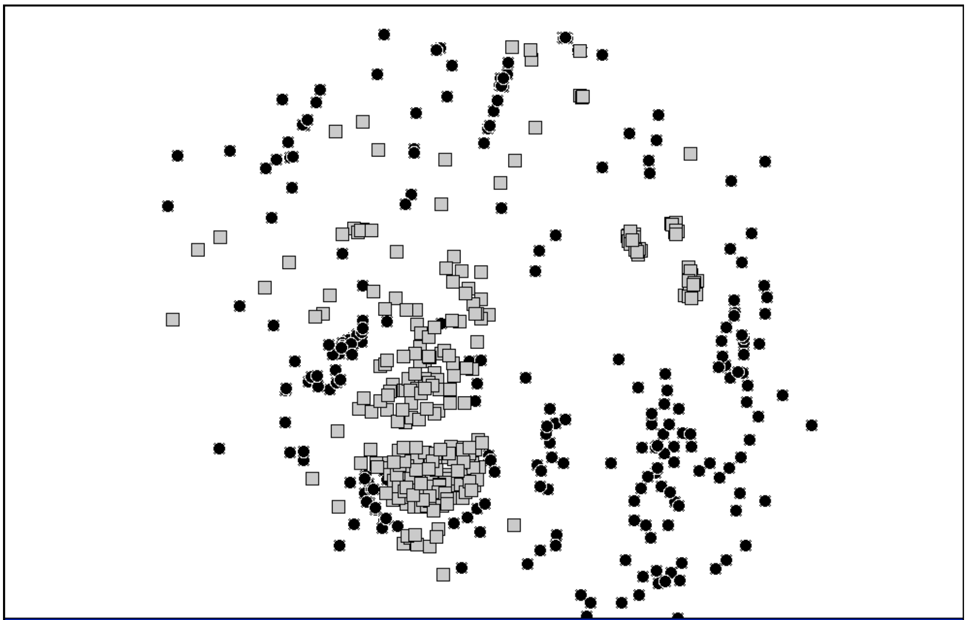
Similarity Measure Selection for Case-Based Reasoning





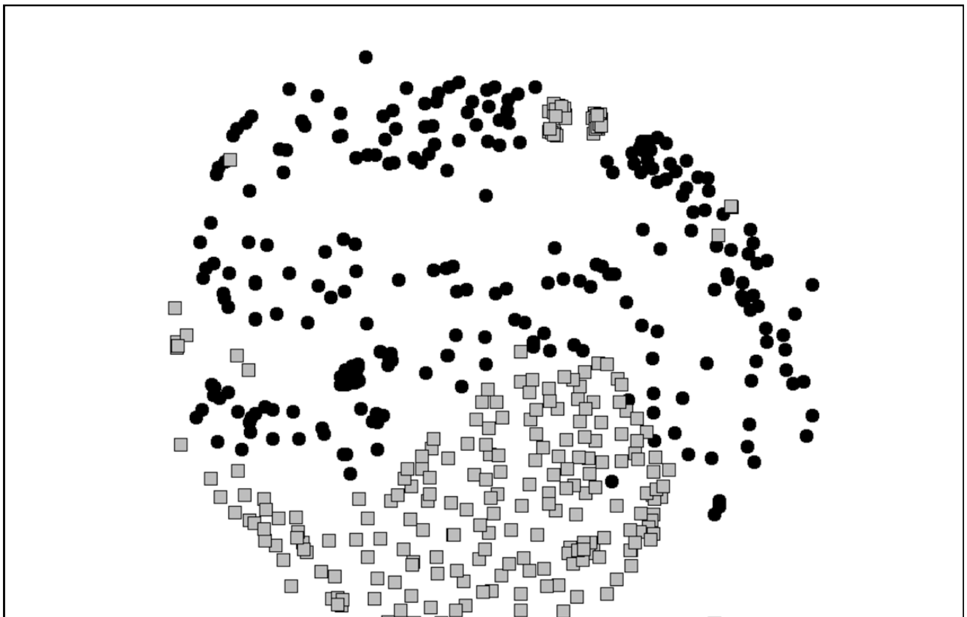
Visualisation of the Spam 2 Dataset

Instances: 500, # Features: 3,219, Cosine similarity



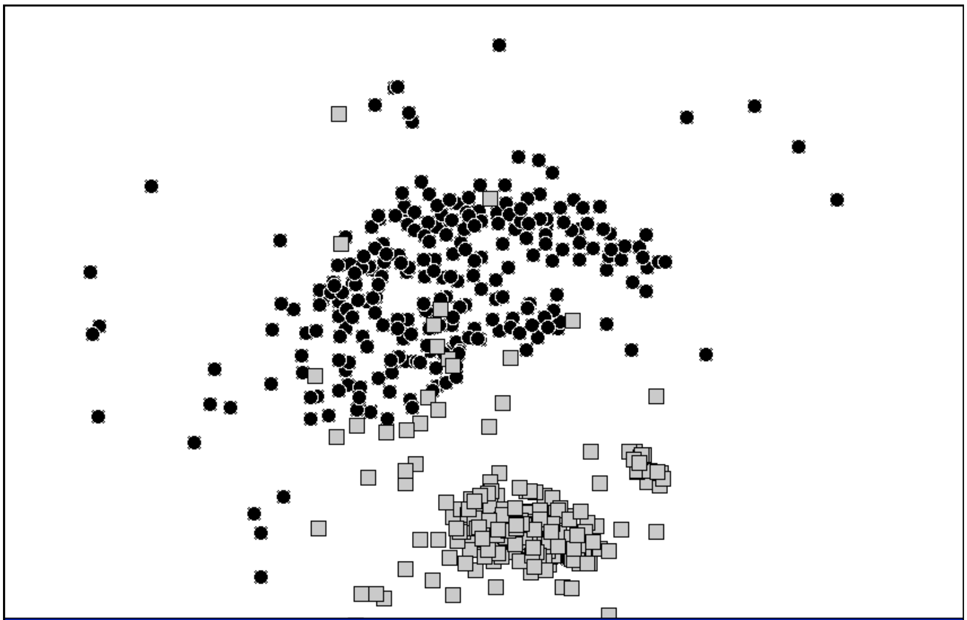
Visualisation of the Spam 2 Dataset

\# Instances: 500, # Features: 3,219, Euclidean similarity



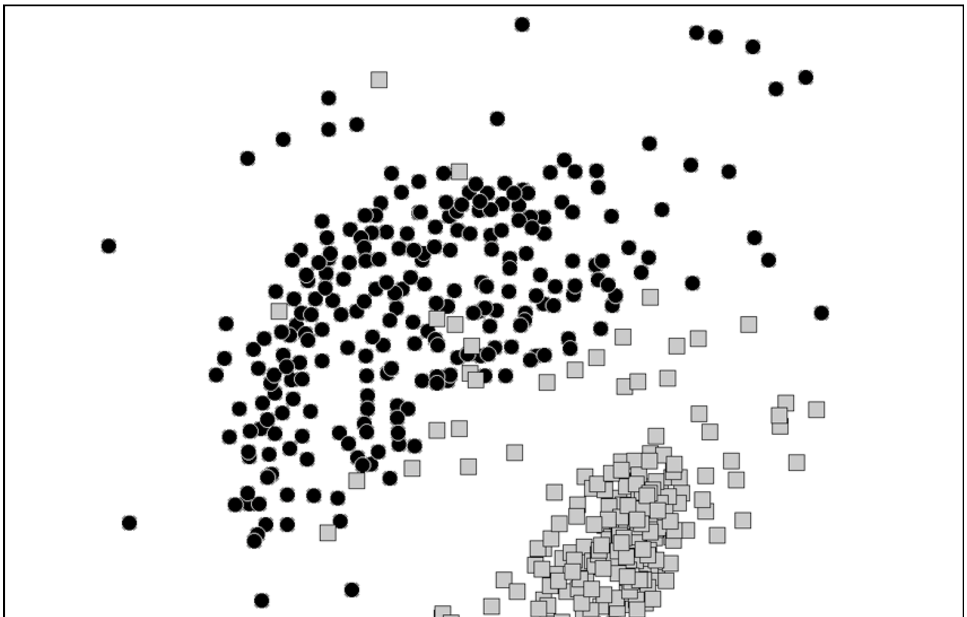
Visualisation of the Spam 2 Dataset

\# Instances: 500, # Features: 3,219, Li similarity



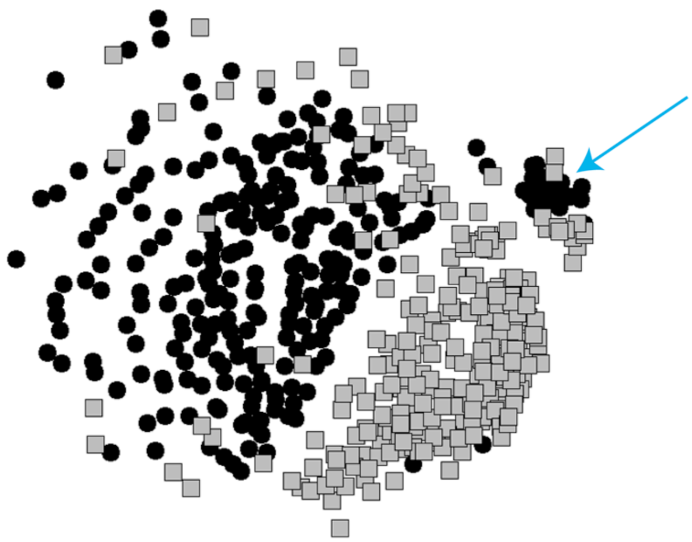
Visualisation of the Reuters Dataset

Instances: 500, # Features: 828, Cosine similarity



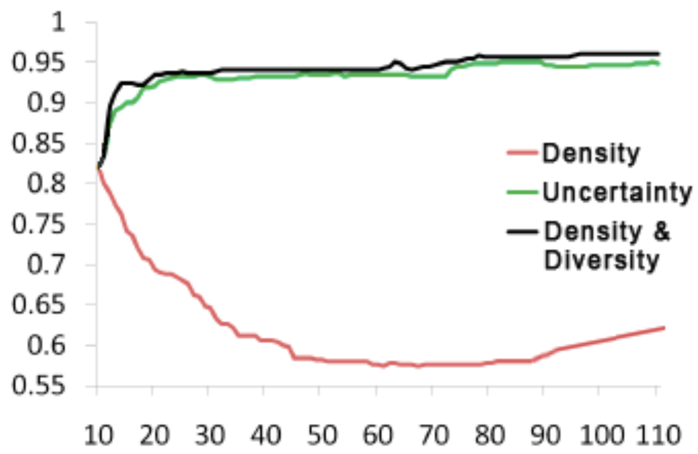
Visualisation of the Reuters Dataset

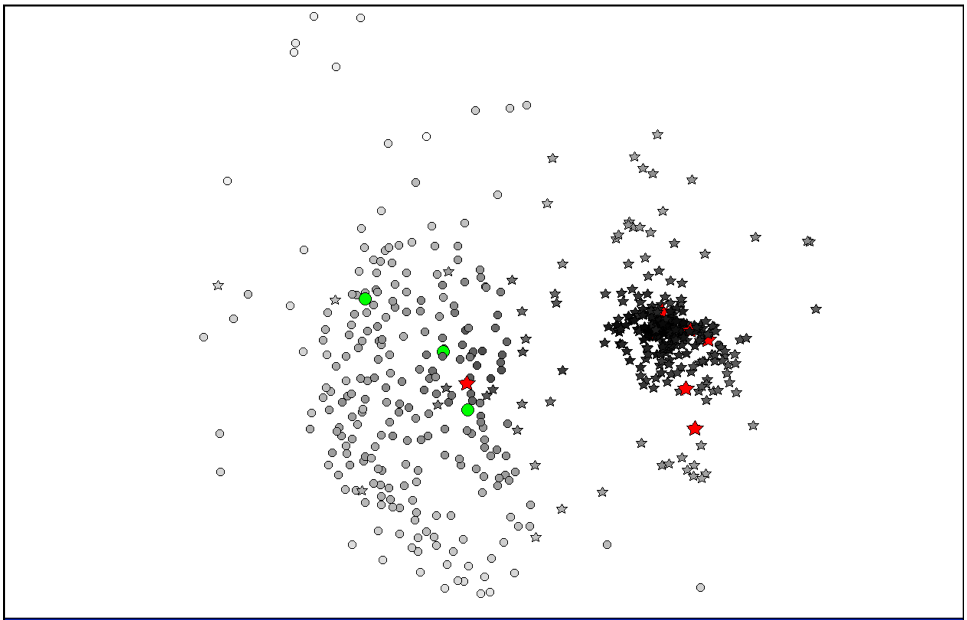
Instances: 500, # Features: 828, Euclidean similarity



Visualisation of the Reuters Dataset
Instances: 500, # Features: 828, Li similarity

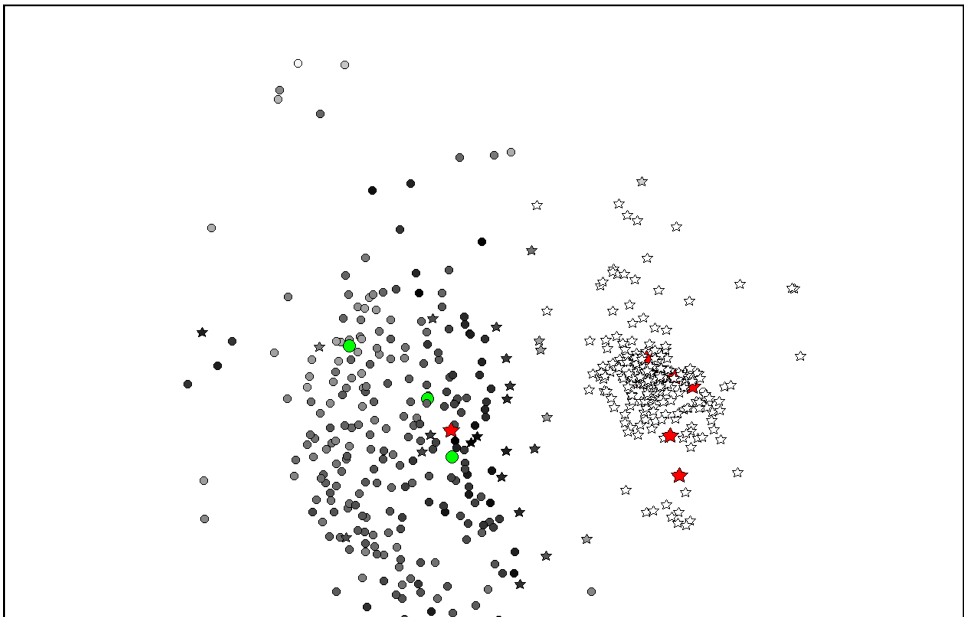
Choosing Selection Strategies for Active Learning





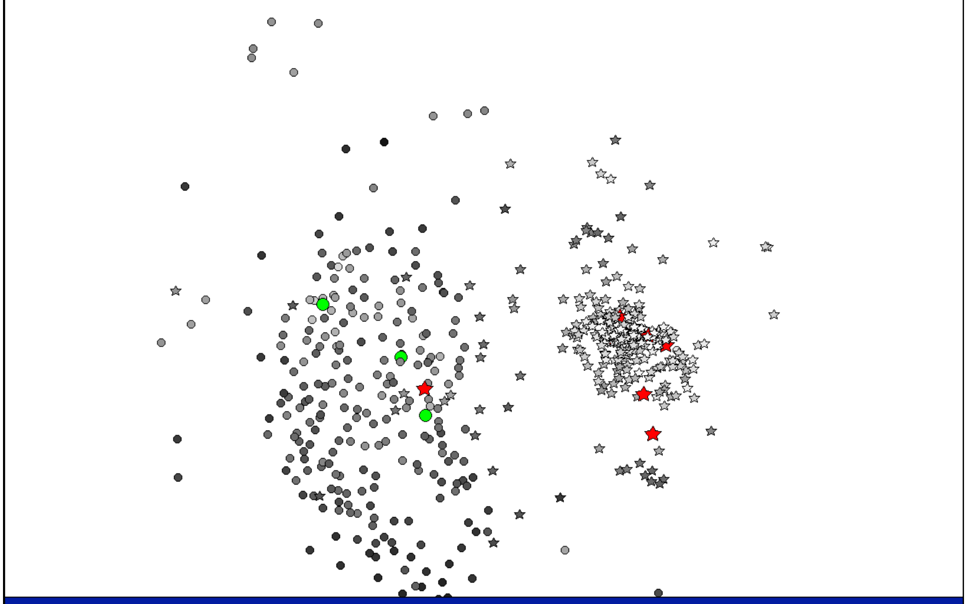
Visualisation of the Reuters Dataset

Instances: 500, # Features: 828, Density selection



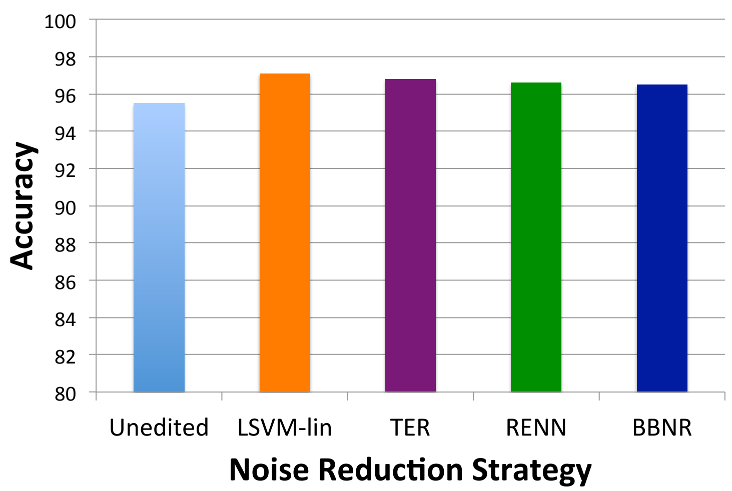
Visualisation of the Reuters Dataset

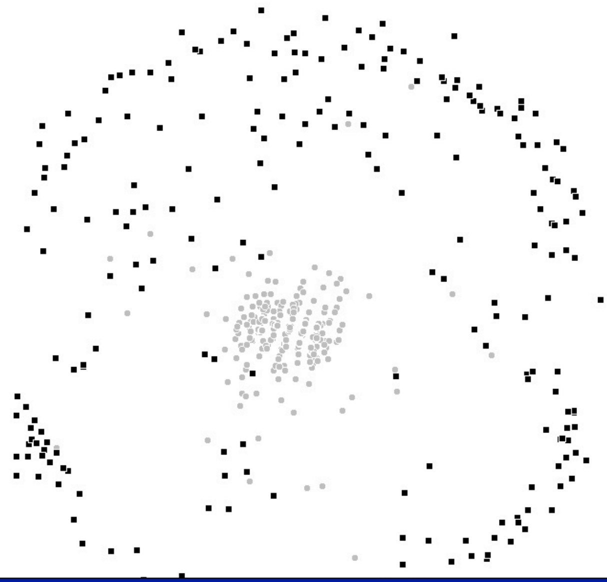
Instances: 500, # Features: 828, Uncertainty selection



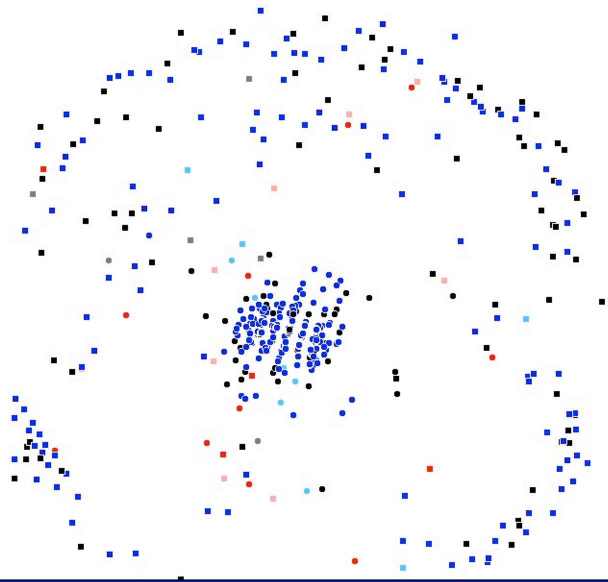
Visualisation of the Reuters Dataset
Instances: 500, # Features: 828, Cosine similarity

Choosing a Case Base Maintenance Strategy

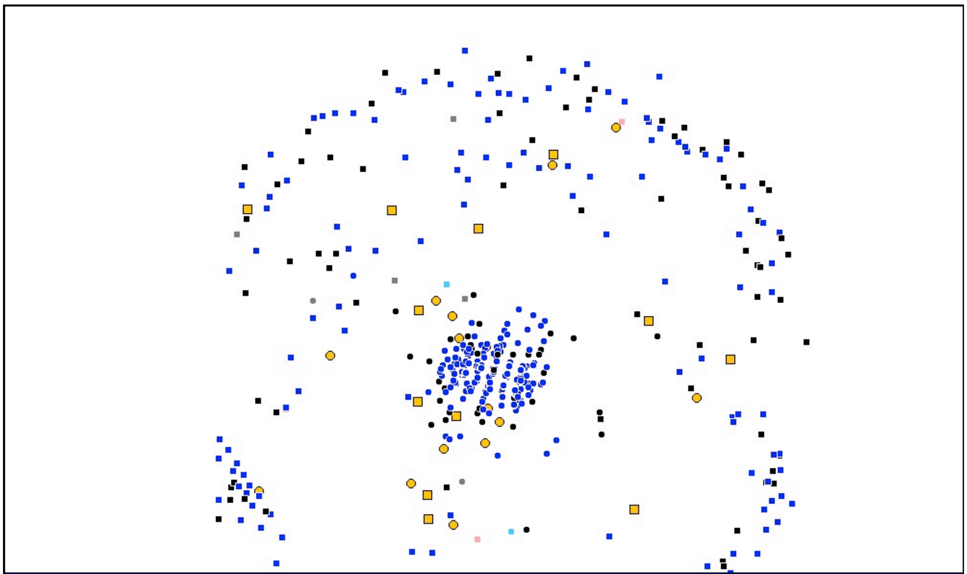




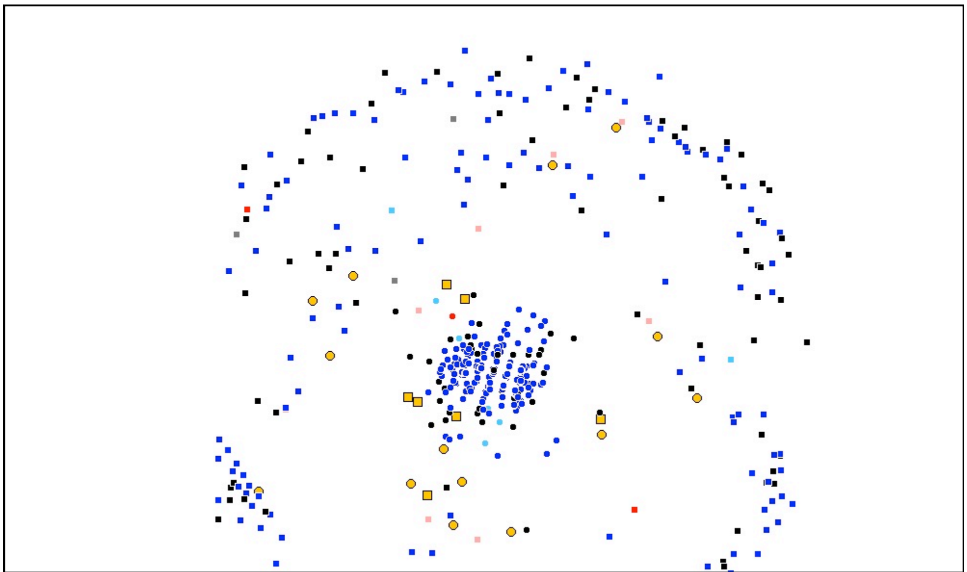
Visualisation of the Wisconsin Breast Cancer Dataset



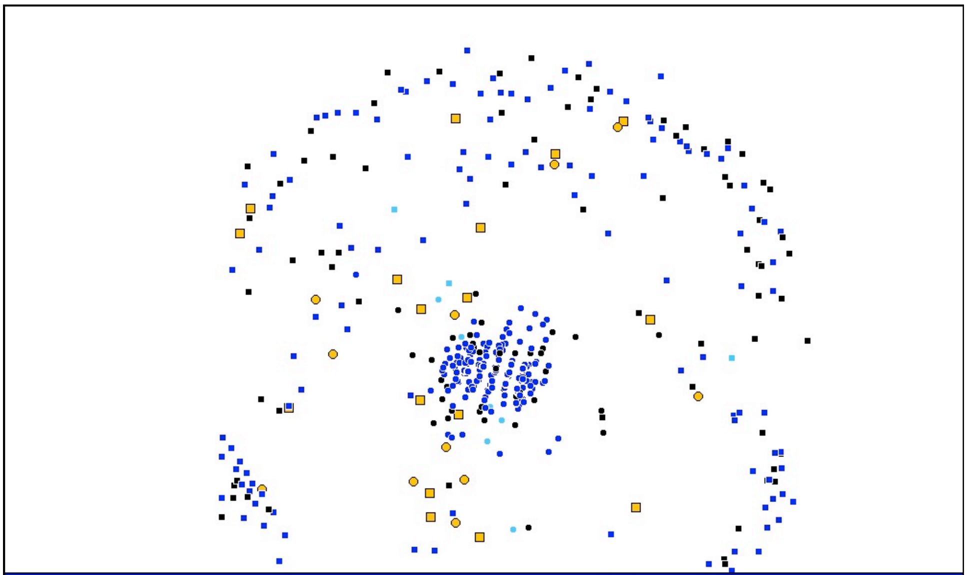
Visualisation of the Wisconsin Breast Cancer Dataset
Case profiles shown



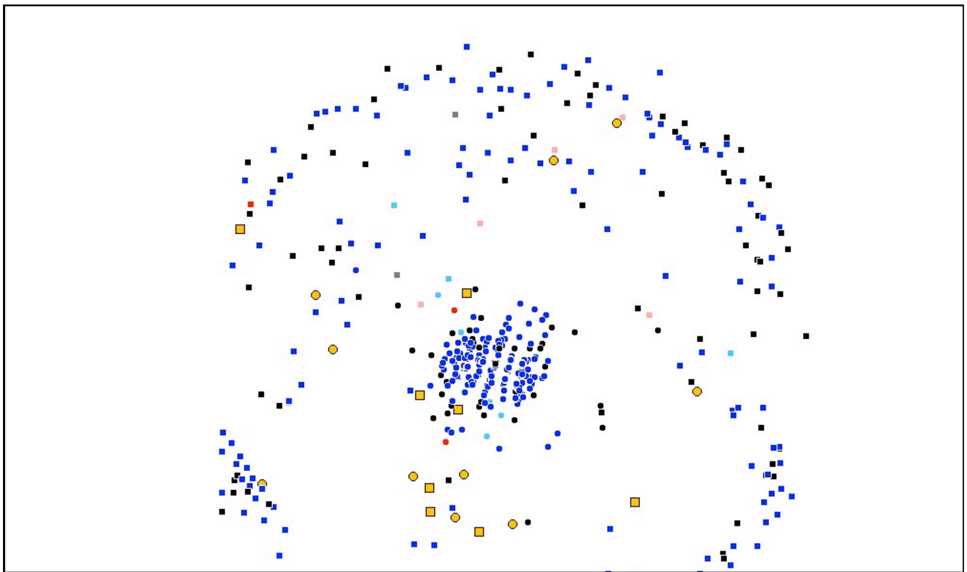
Visualisation of the Wisconsin Breast Cancer Dataset
Case profiles and Blame Based Noise Reduction noise reduction selected cases shown



Visualisation of the Wisconsin Breast Cancer Dataset
Case profiles and Local Support Vector Machine noise reduction selected cases shown



Visualisation of the Wisconsin Breast Cancer Dataset
Case profiles and Repeated Edited Nearest Neighbour noise reduction selected cases shown



Visualisation of the Wisconsin Breast Cancer Dataset
Case profiles and Threshold Error Reduction noise reduction selected cases shown

Conclusion: It does seem possible to create visualisations to help analysts better understand and use machine learning algorithms

Caveat: This relies on an accurate and representative visualisation

Future Work: Make examples more interactive and perform real evaluations with participant analysts

CONCLUSION

Conclusion

Coping with multi-dimensional data is one of the key challenges in data visualisation

We deal with a constant trade-off between losing information and gaining interpretability

We have looked at a number of approaches:

- Point-Based Techniques
- Line-Based Techniques
- Dimension Reduction Techniques

Many variants and combinations of these approaches exist