IRDS: Visualization

Charles Sutton
University of Edinburgh

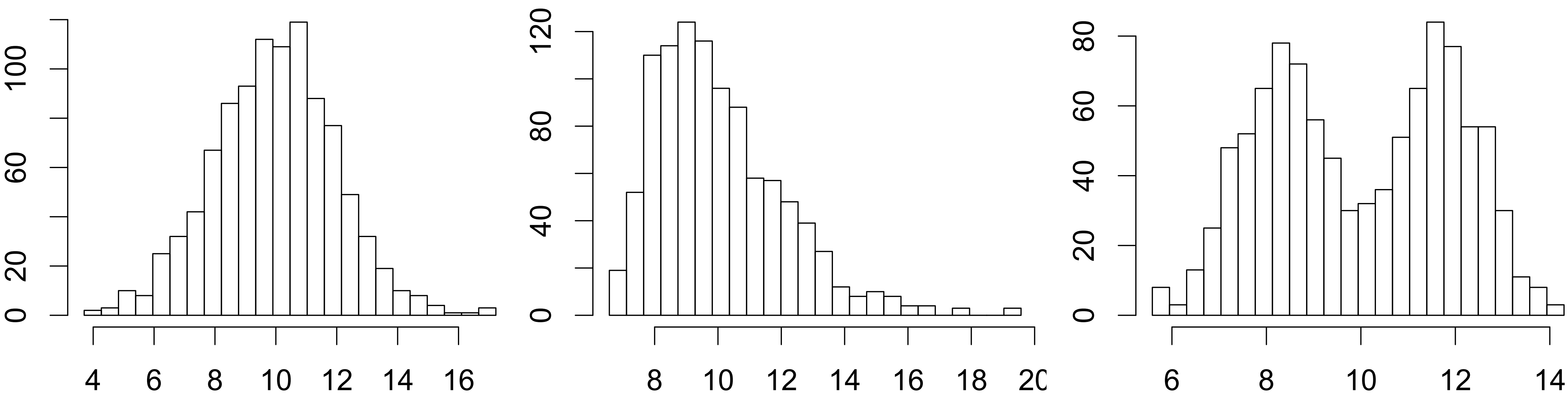
Why visualisation?

- Exploratory
  - What’s in the data? What’s wrong with it?
  - Today’s lecture.
- Presentation
  - Display results of algorithms for publication
- Engagement
  - Infographics from web sites, word clouds, etc

Summary Statistics

- Univariate
  - Mean
  - Median
  - Variance
  - Quantiles
- Multivariate
  - Correlation
  - Covariance

Histograms

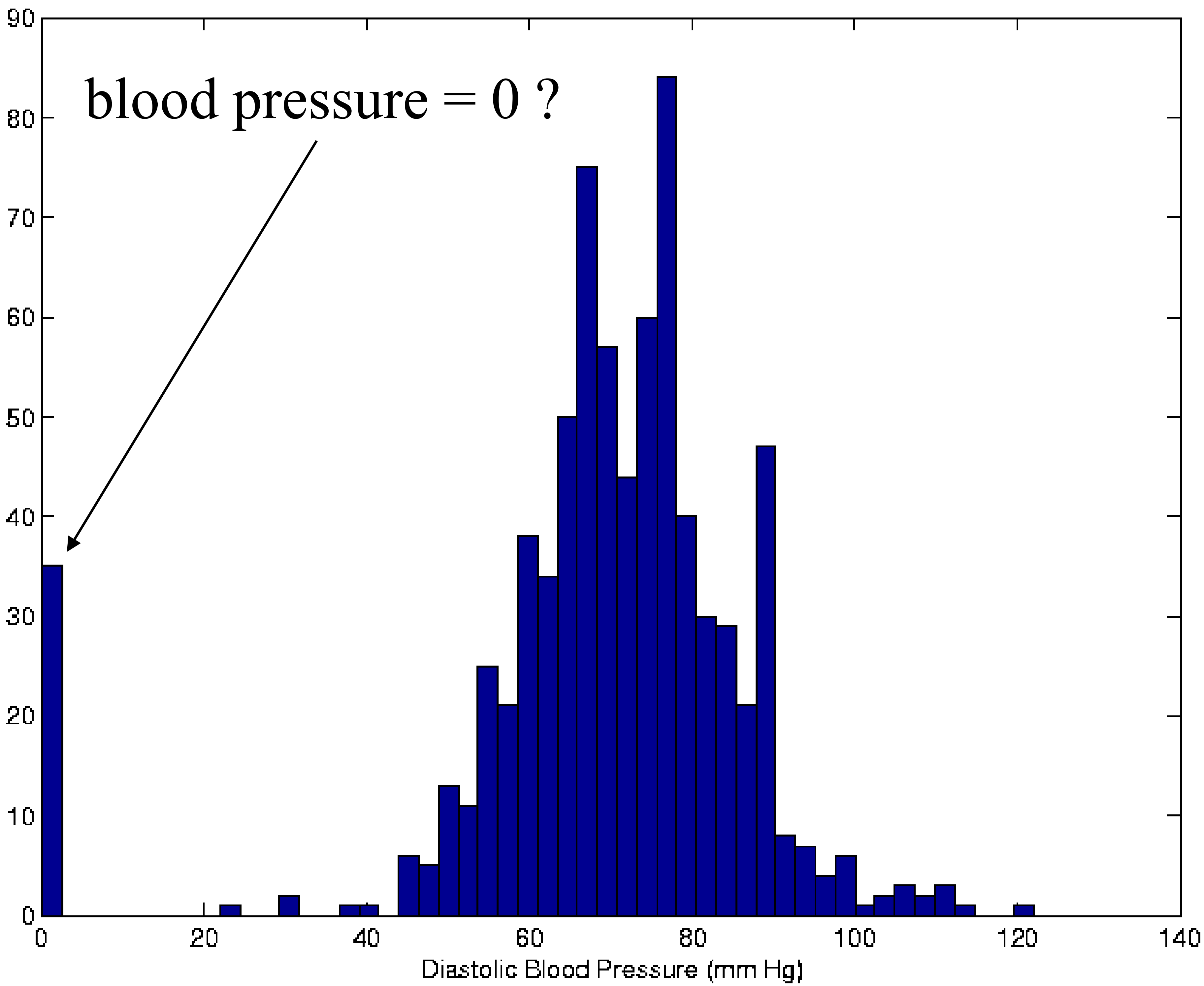


skew

multimodality

these three have same summary statistics!

Outliers in histograms

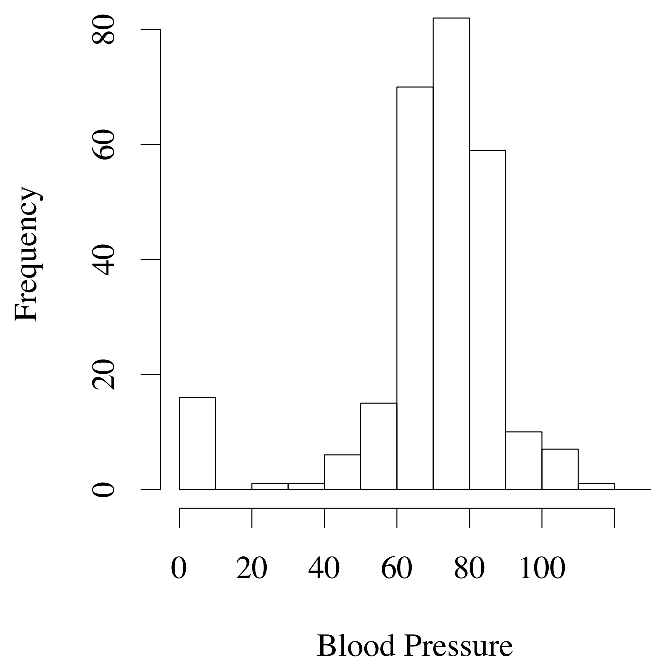


Blood pressure data set

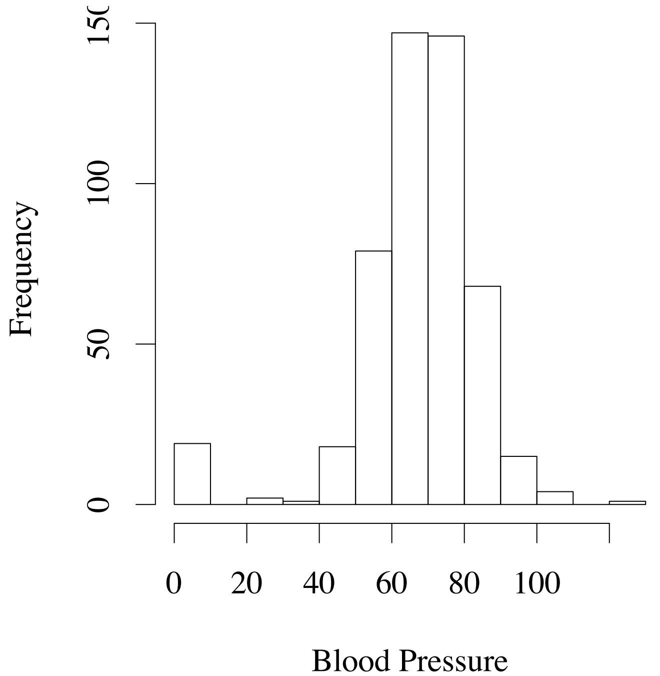
UCI ML repository says no missing data
(well, for 20 years it did)

[Source: Padhraic Smyth]

Class-Conditional Histograms

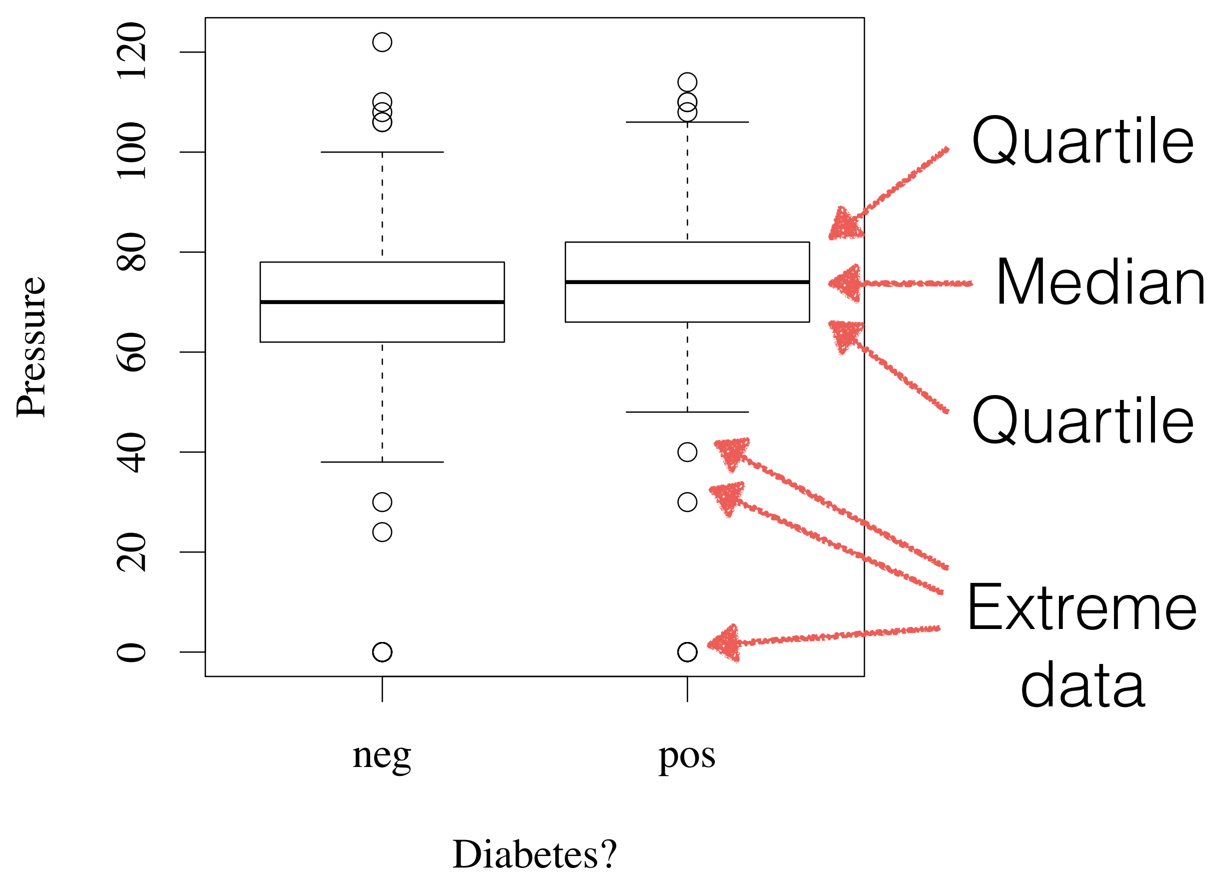


Positive
(diabetes)



Negative

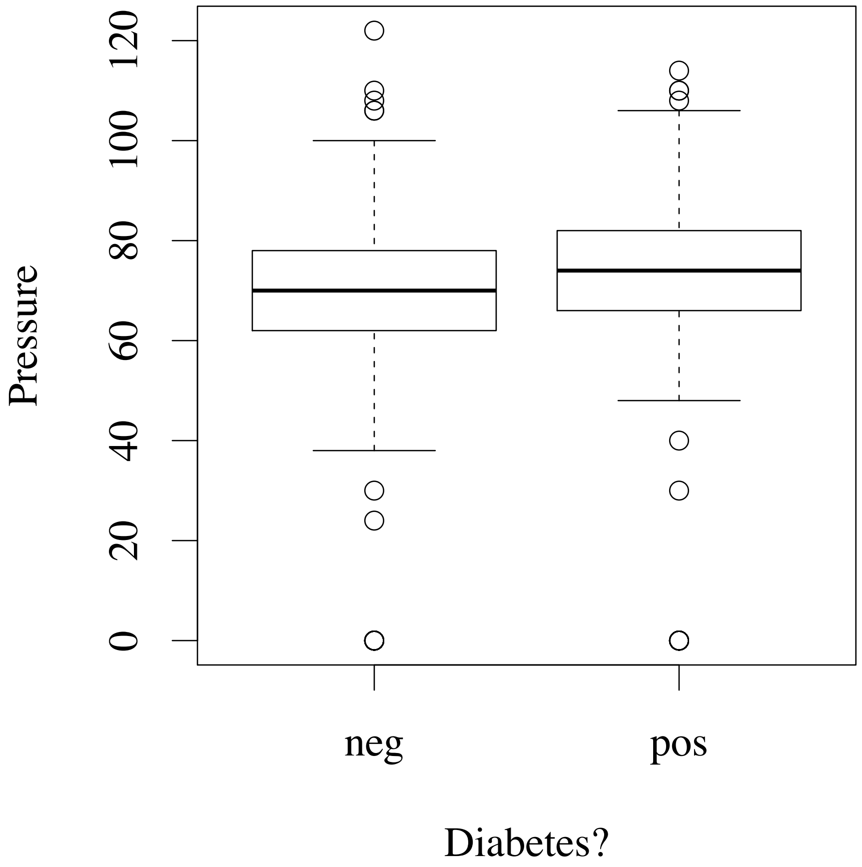
Alternative: Box plot



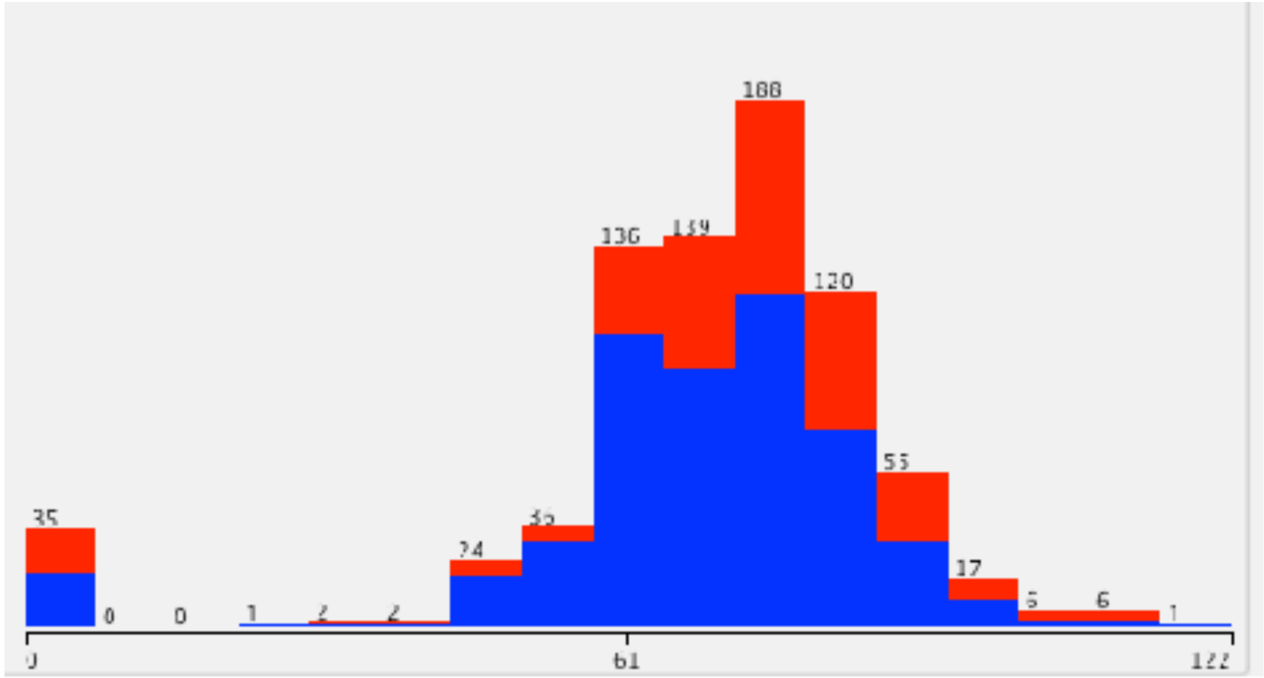
Maybe for only 2 groups, graphs not necessary.
For more visual comparisons, can be helpful.

Slight rant about bar charts

Here's my boxplot

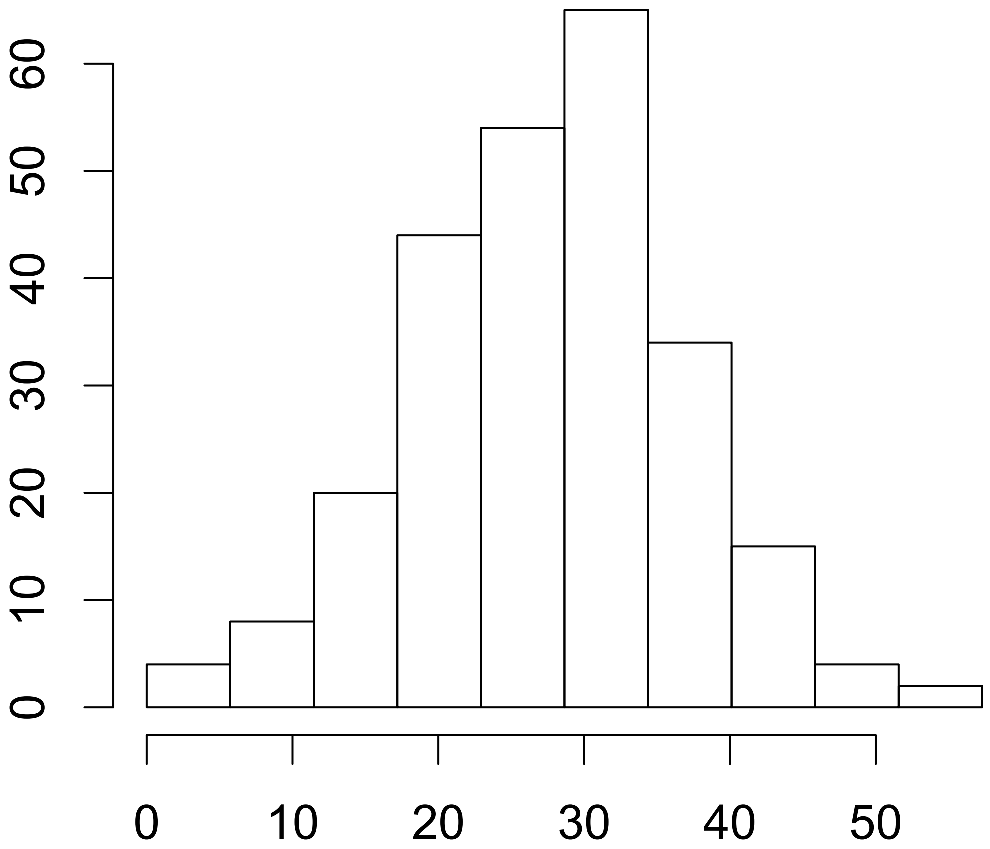


Weka's automatic visualisation

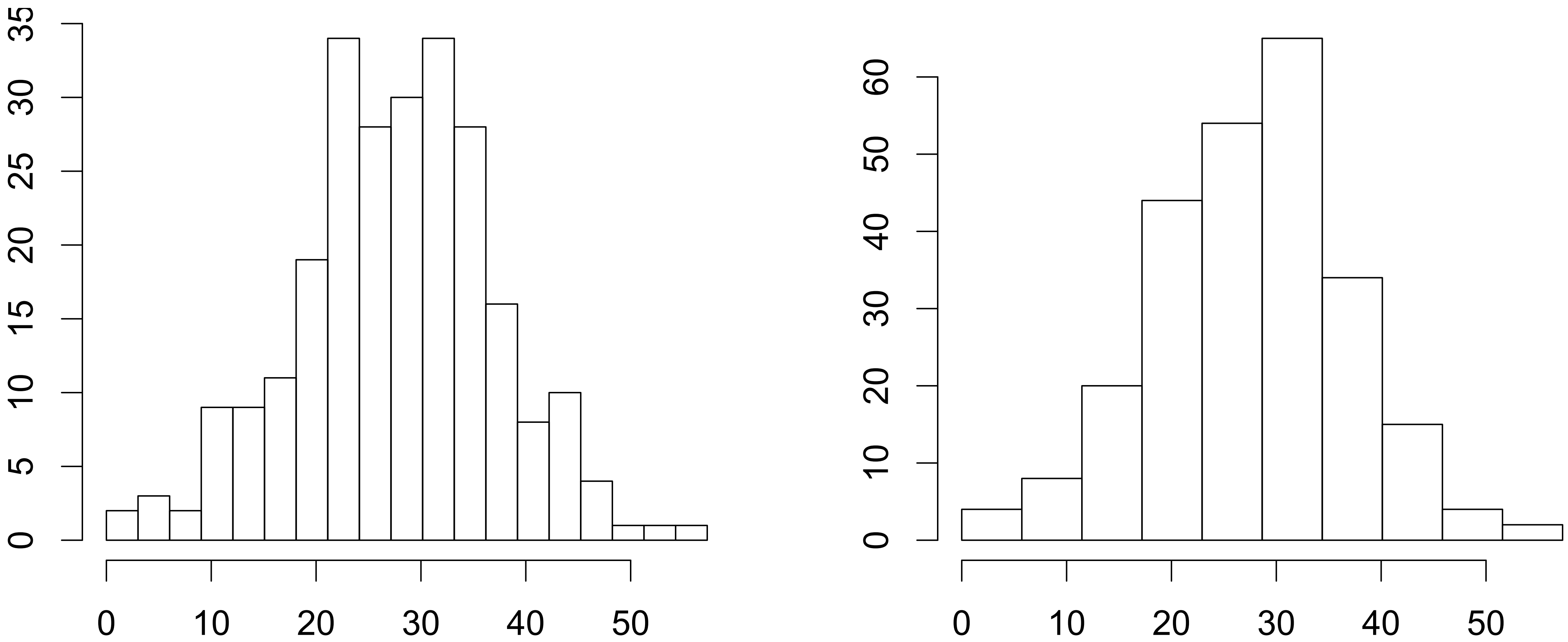


Bar charts often seem like a better idea than they are

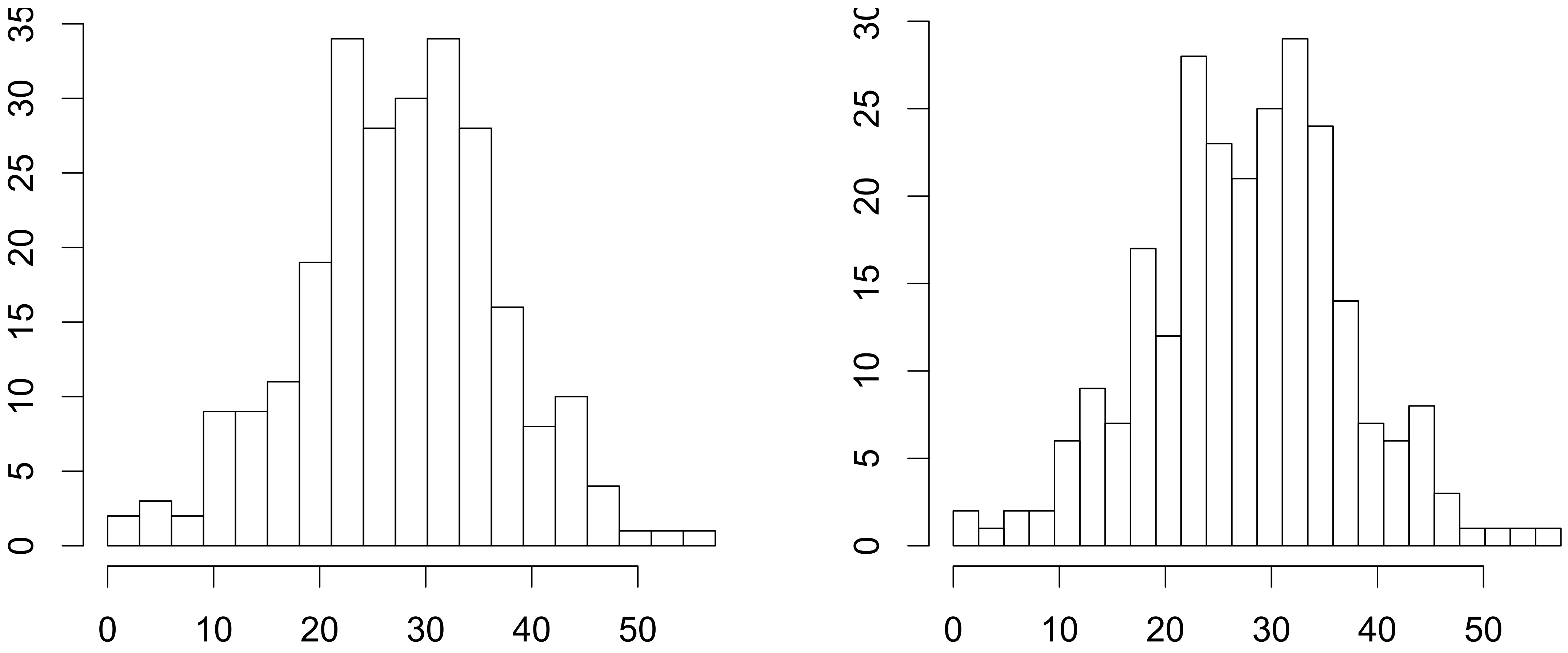
Effect of bin size



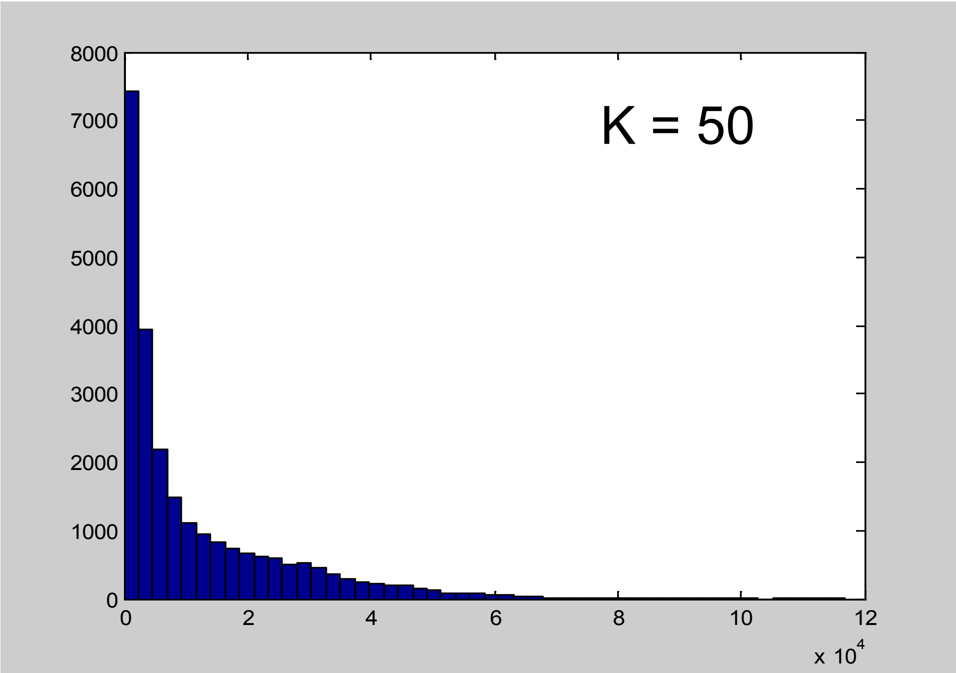
Effect of bin size

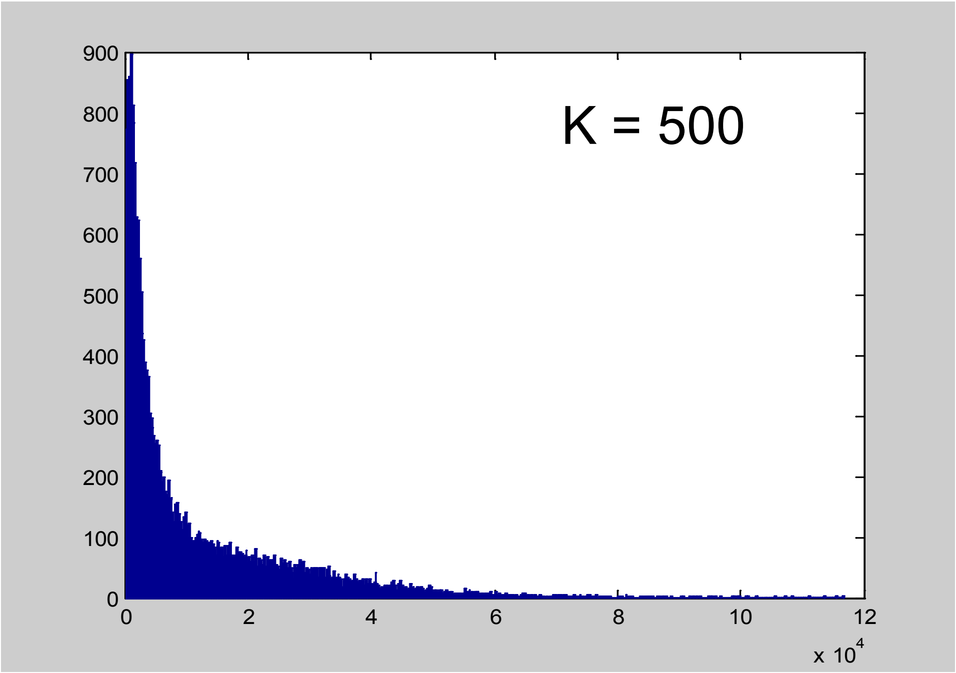


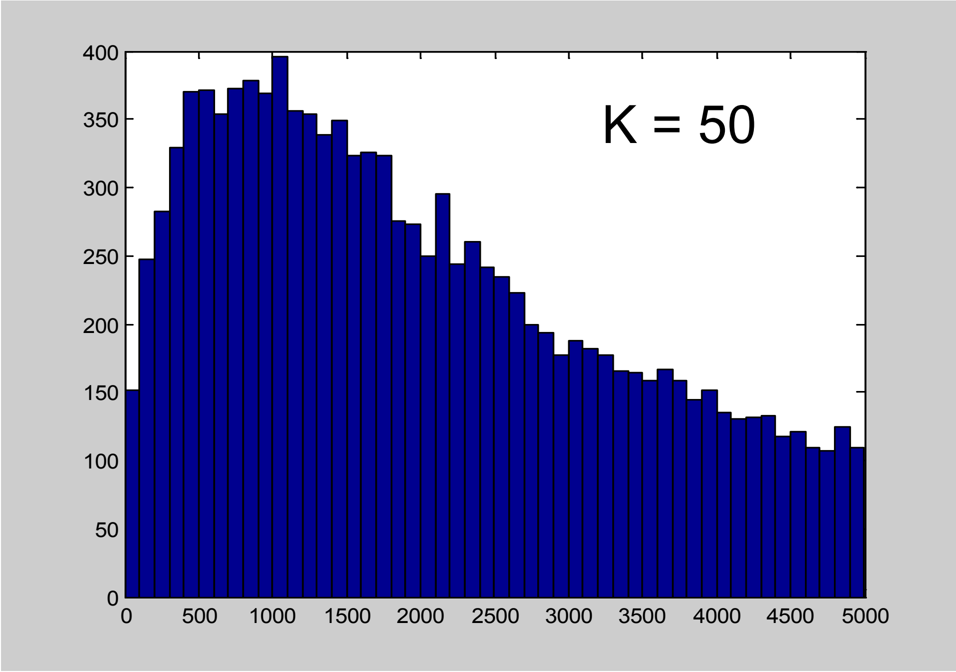
Effect of bin size



More misleading histograms







Data: US Post Codes

[Source: Padhraic Smyth]

Bivariate data

- Numerical summaries about linear dependence
- Histograms sort of scale to 2-D but not really higher
- More common to use scatter plots

Numerical bivariate summaries

Data are $(x_1, y_1), (x_2, y_2), \ldots (x_N, y_N)$

Sample covariance:

$$s_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})$$

Sample correlation:

$$\rho_{xy} = \frac{s_{xy}}{s_x s_y}$$

where as before

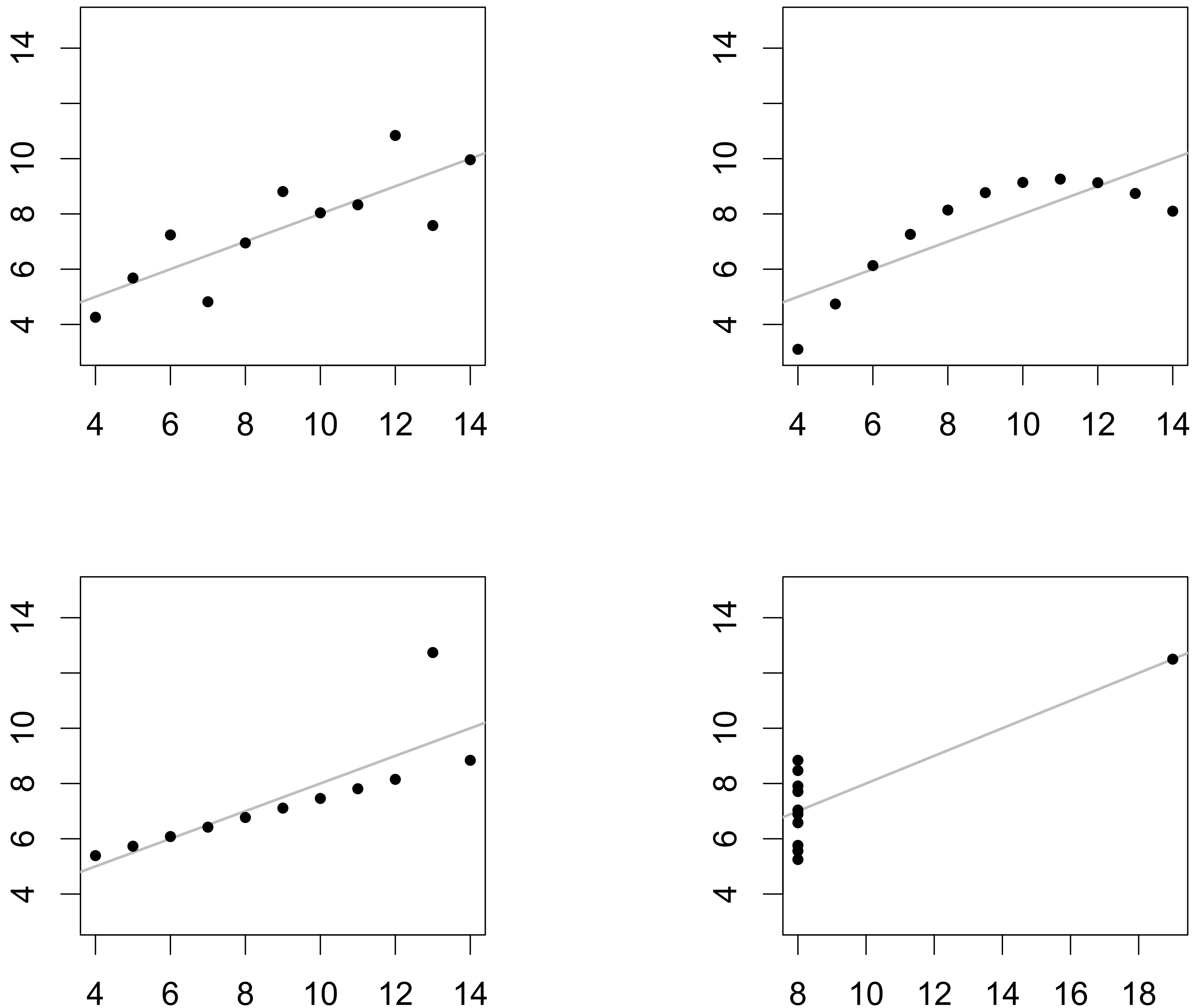
$$\bar{x} = \frac{1}{N} \sum_i x_i$$

$$\bar{y} = \frac{1}{N} \sum_i y_i$$

$$s_x = \sqrt{\frac{1}{N-1} \sum_i (x_i - \bar{x})}$$

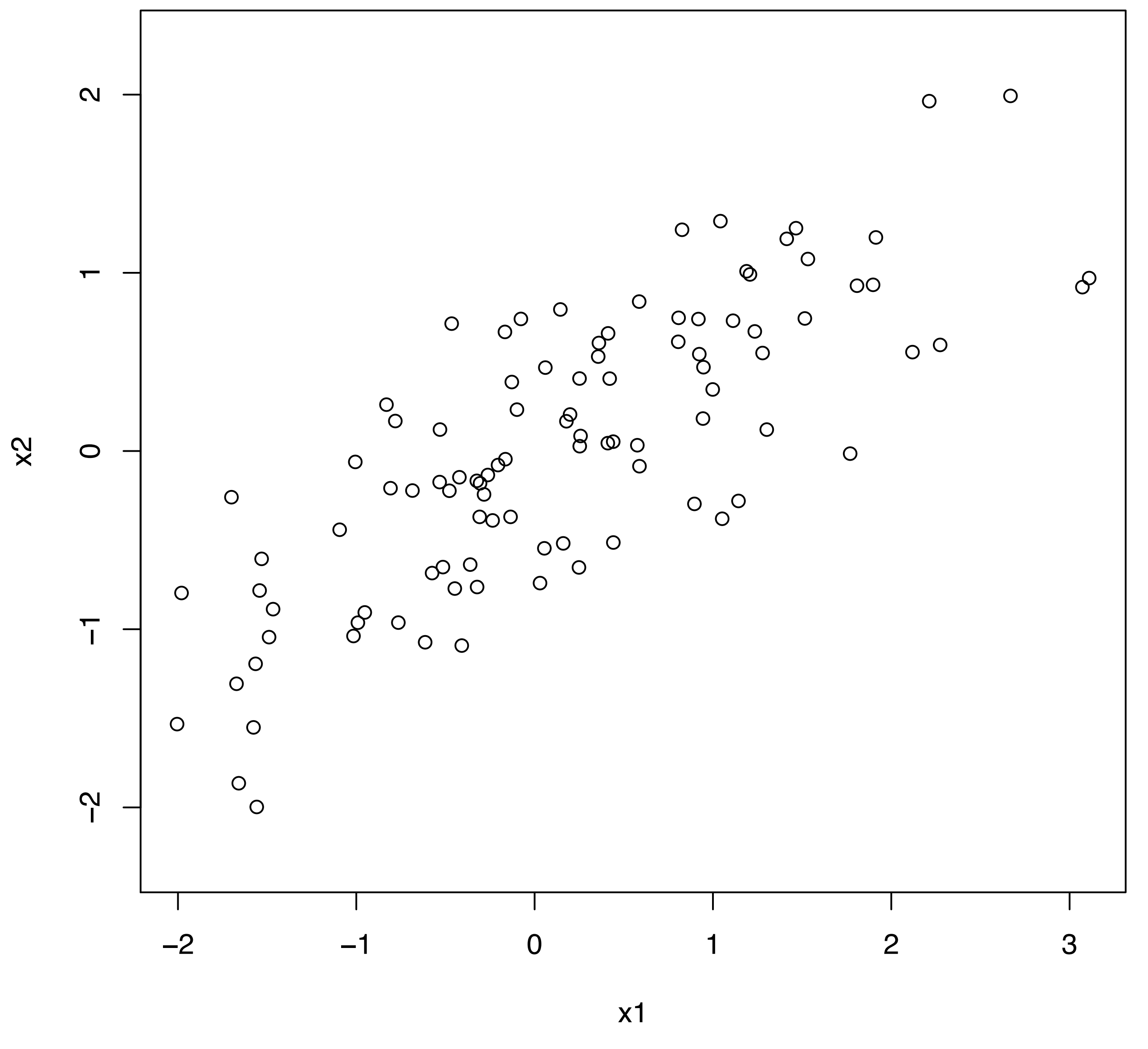
$$s_y = \sqrt{\frac{1}{N-1} \sum_i (y_i - \bar{y})}$$

Dangers of correlation



[Anscombe, 1973]

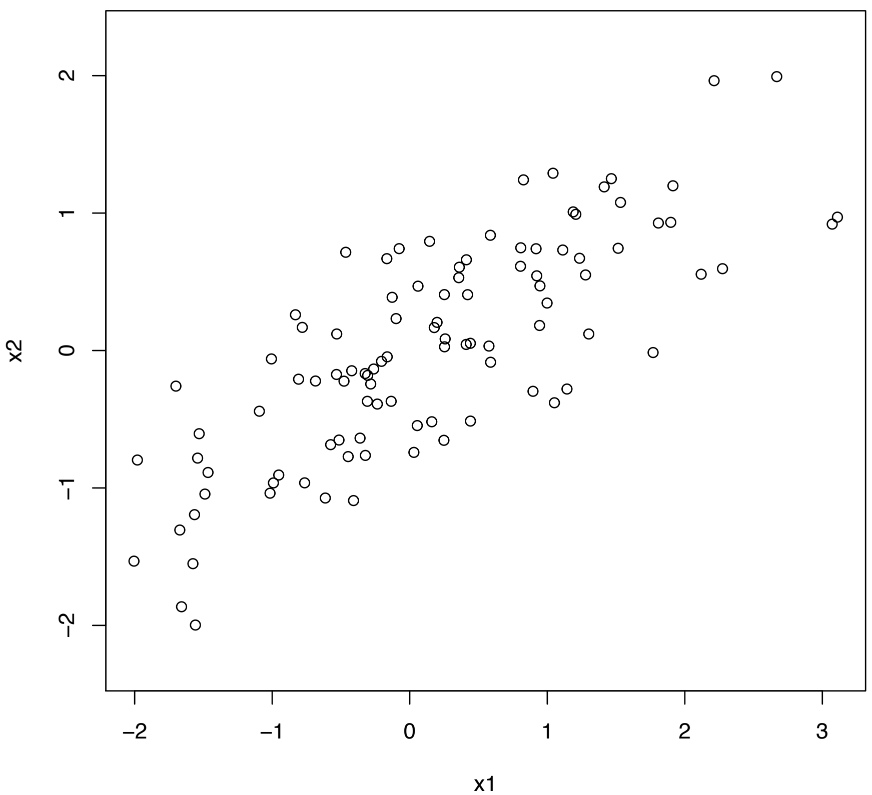
Scatterplots



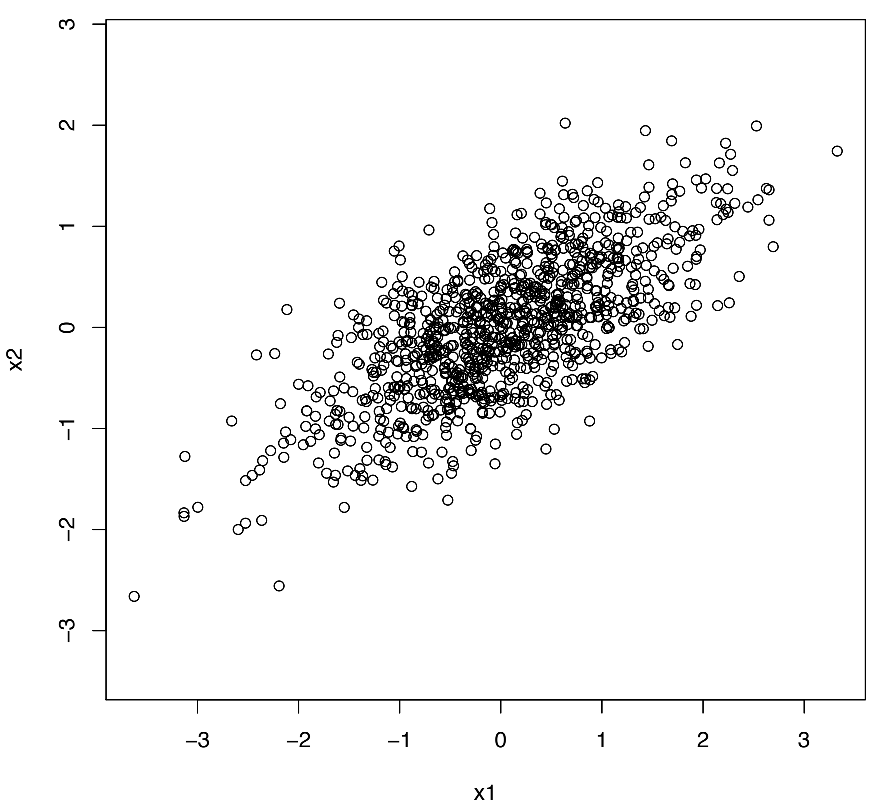
Overplotting

samples from bivariate normal

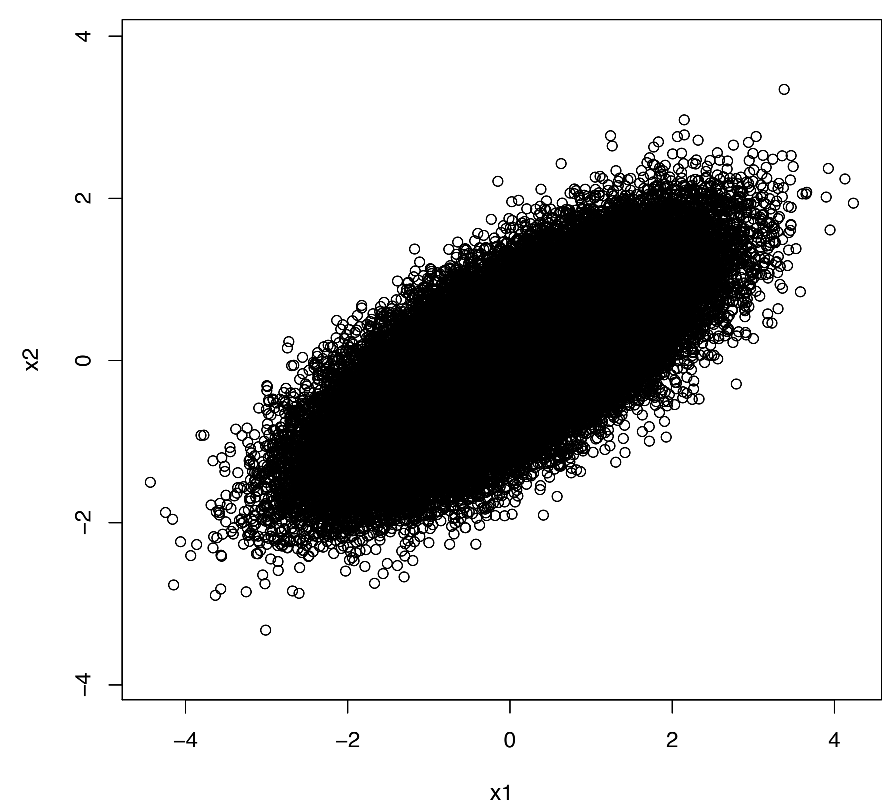
also: notice the axes!



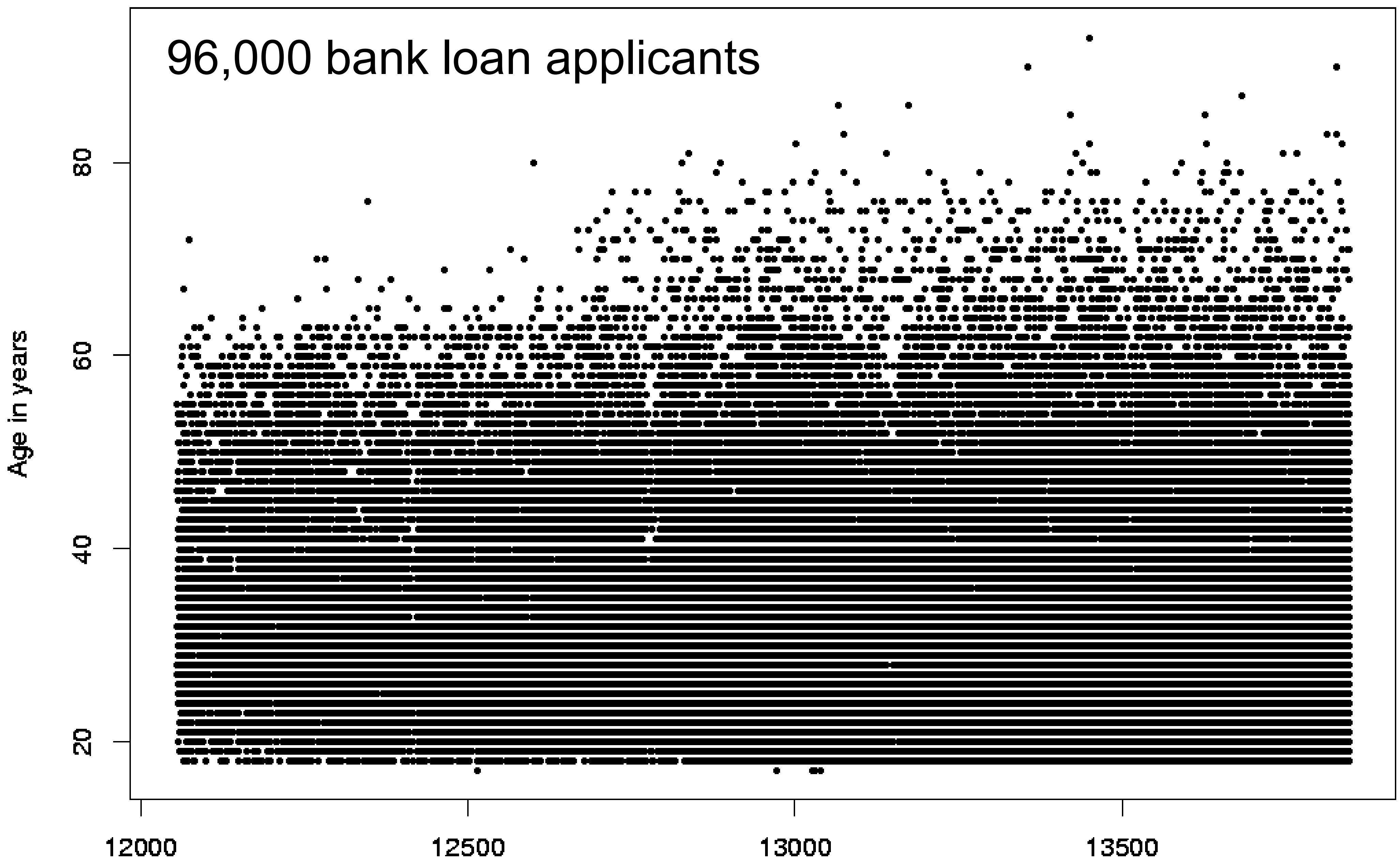
100 data points



1000 data points



100,000 data points



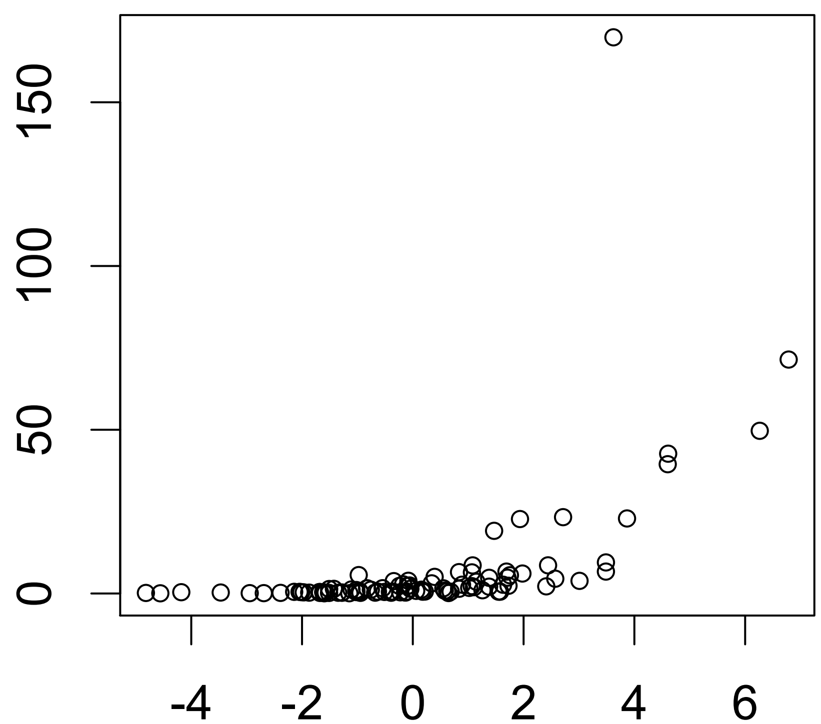
[Source: Hand, Manila, and Smyth]

To fix overplotting, could consider:

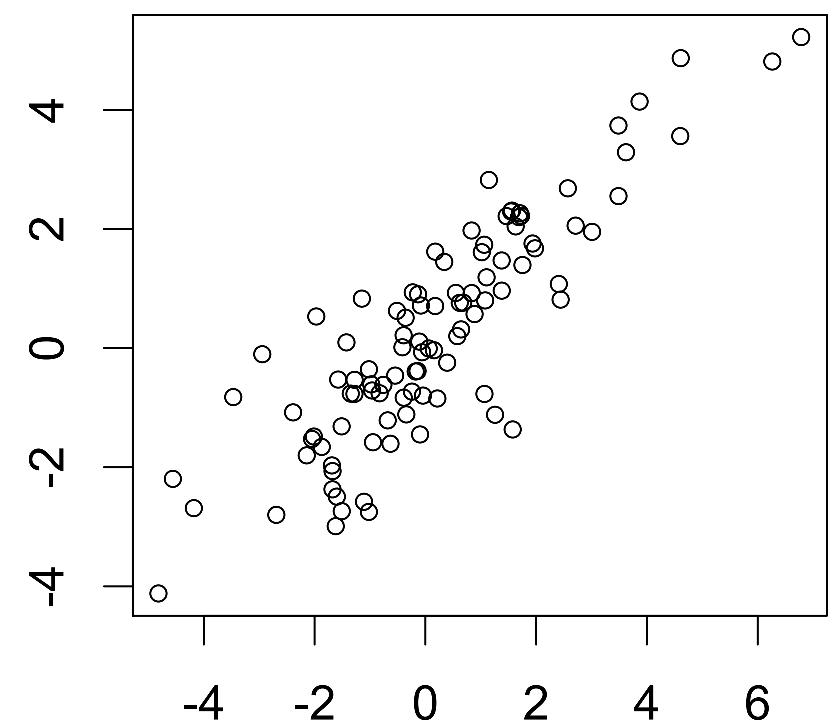
- Jittering points
- Subsampling points (i.e., plot only 10%)
- Averaging (if this makes sense)
- Add trend lines (e.g., quantile lines)

Transformations

Consider powers, logs.
Occasionally reciprocals (e.g., rates).
Also square root

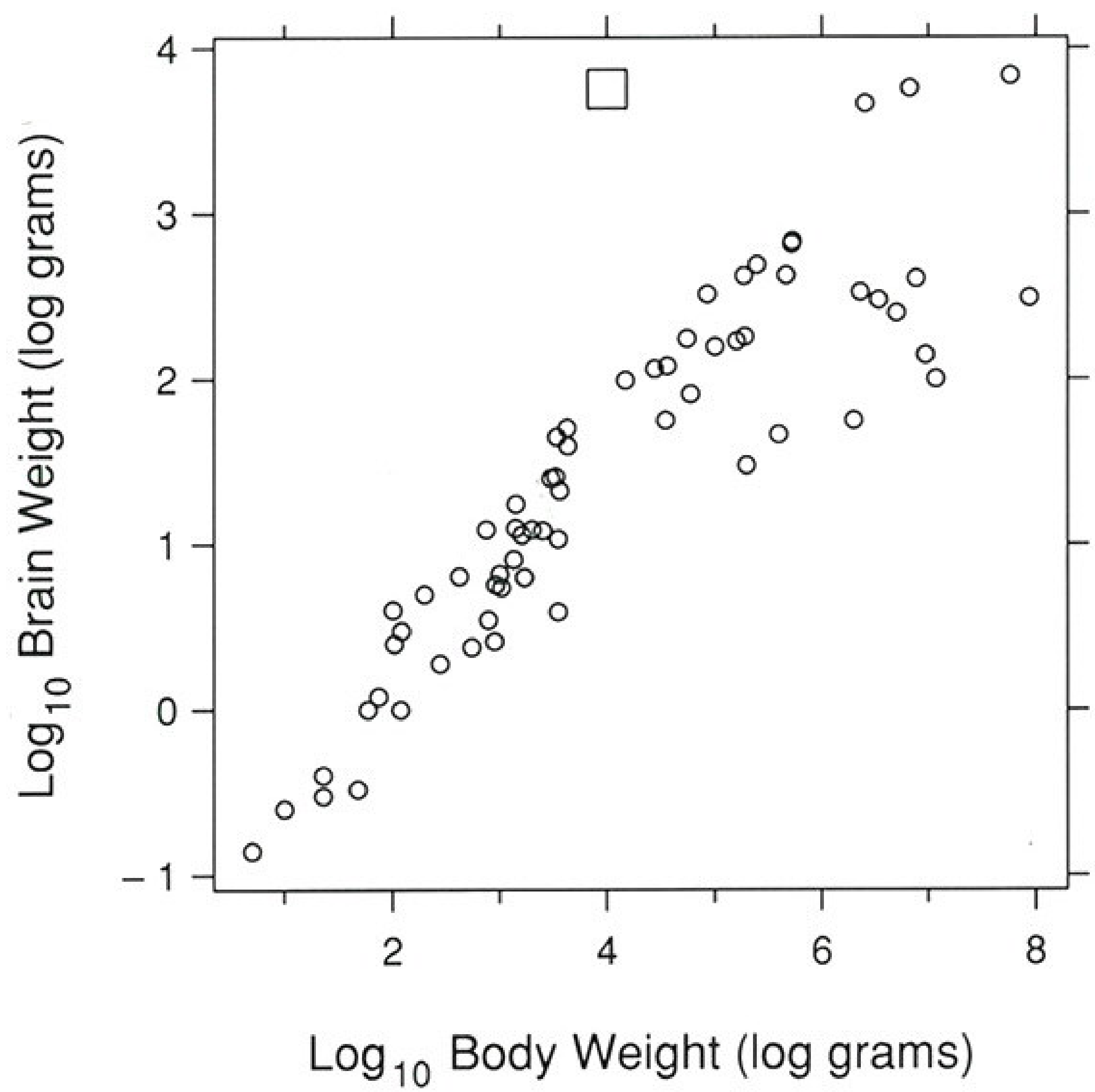


Before



After

Example Transformation



Why log log here? Hint: Imagine a spherical cow

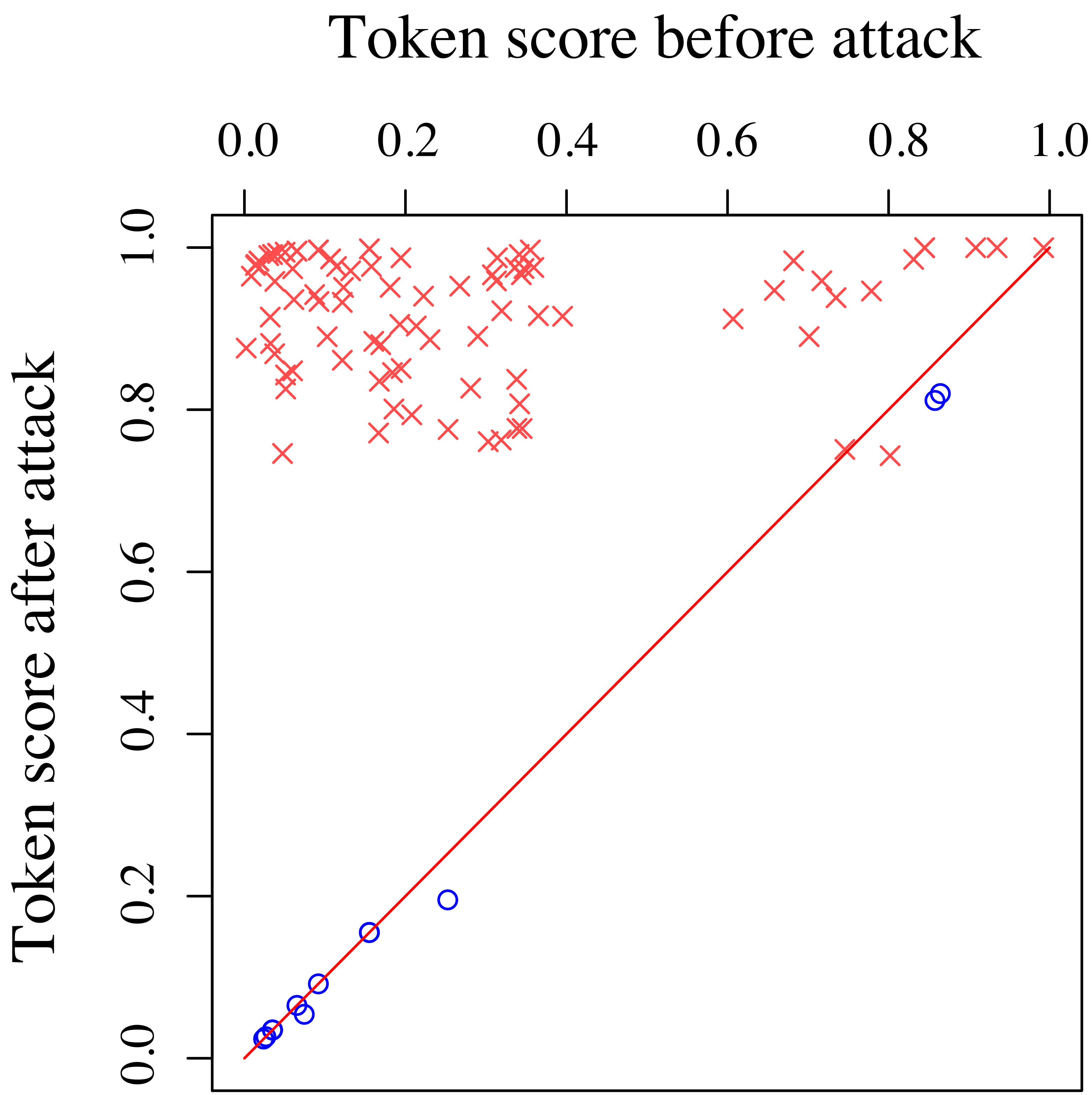
[Source: William Cleveland, Visualizing Data]

Wait, what if you have categorical data?

Tools here include:

- Colour
- Contingency tables
- Multiple plots (e.g., class-conditional histograms)

Colour in Scatterplots

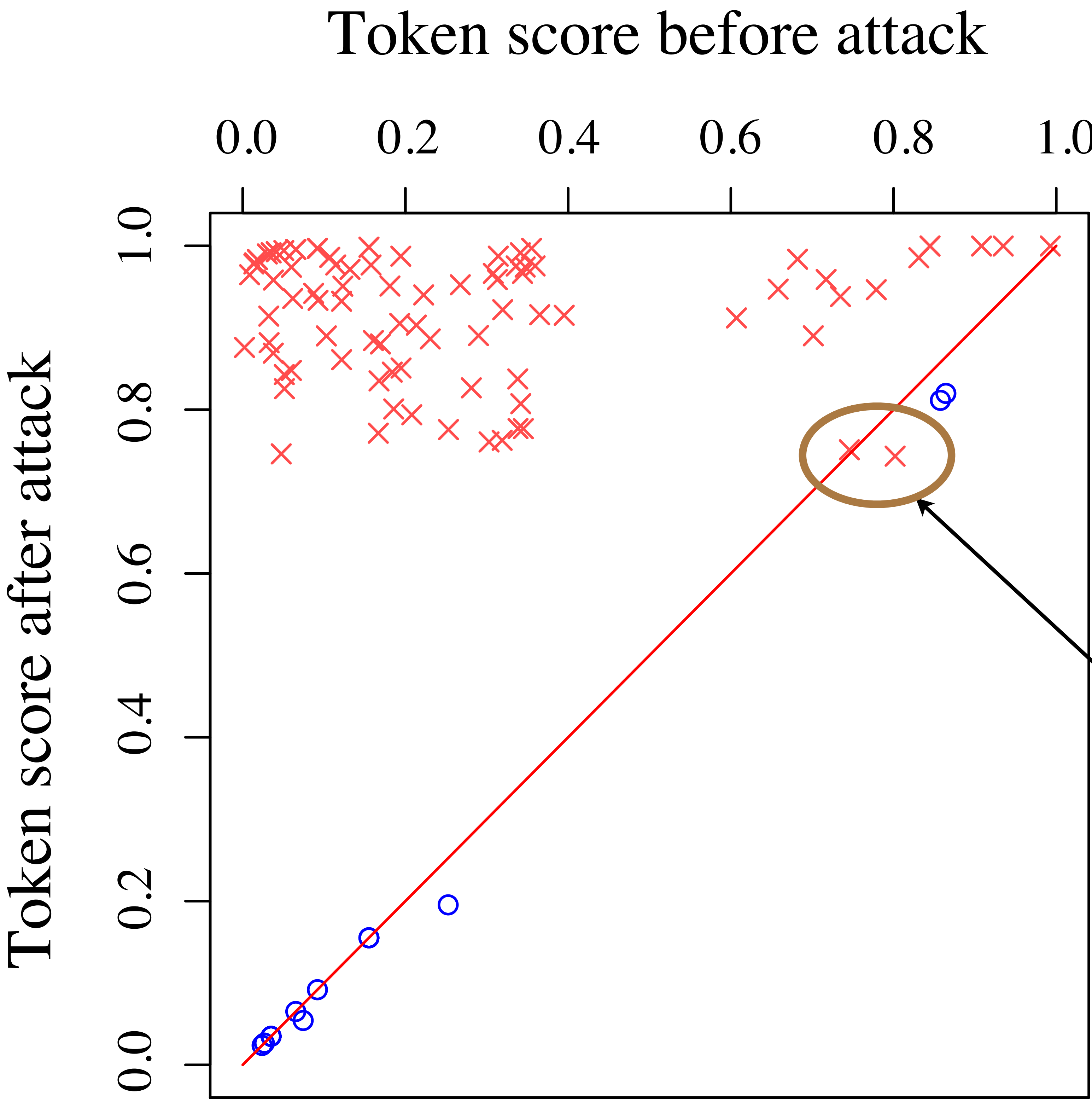


Each point is a word
Entire plot: one email
Axes: “Spam score”

Colour: Whether token was part of an attack on the spam filter

[Nelson et al, 2008]

Colour in Scatterplots



For our purposes, note:

- Use of colour to add a categorical variable
- Without this colour would not have seen these two outliers
- Use of y=x line to add the eye

[Nelson et al, 2008]

Thoughts about using color

- Think about data: sequential, diverging, qualitative
- Intensity of color conveys information
- Colorblindeness
- Cartographers care about this
  - colorbrewer2.org

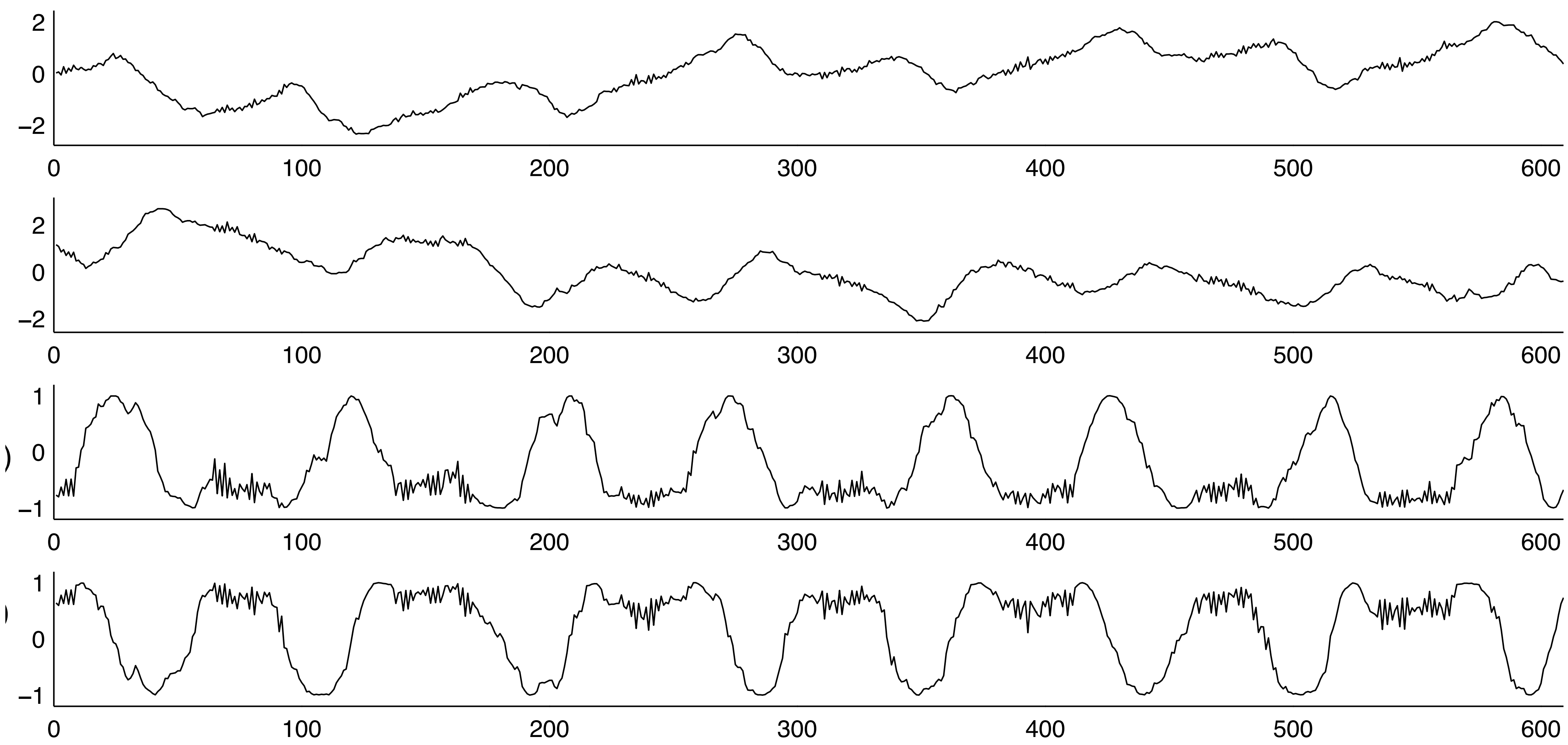
Time Series

Examples

- Financial data
- Network traffic
- Energy usage
- Human traffic
- Building occupancy

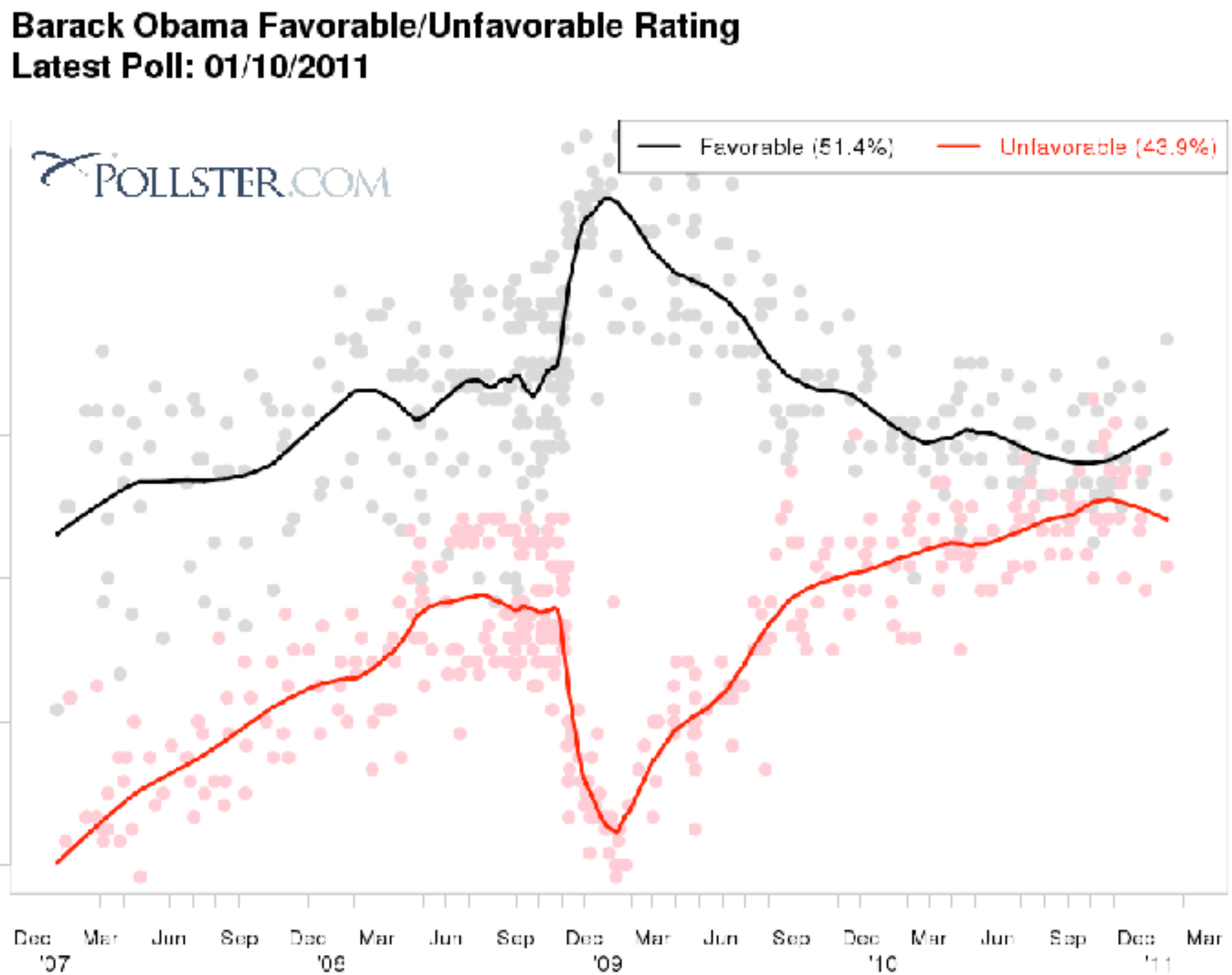
Visualization tricks include:

- Smoothing
  - (running mean, median)
- Repeated multiples



[Oh et al, 2006], figure from [Xuan and Murphy, 2007]

Fitted line



This fit is from loess (local linear regression).

Power Law

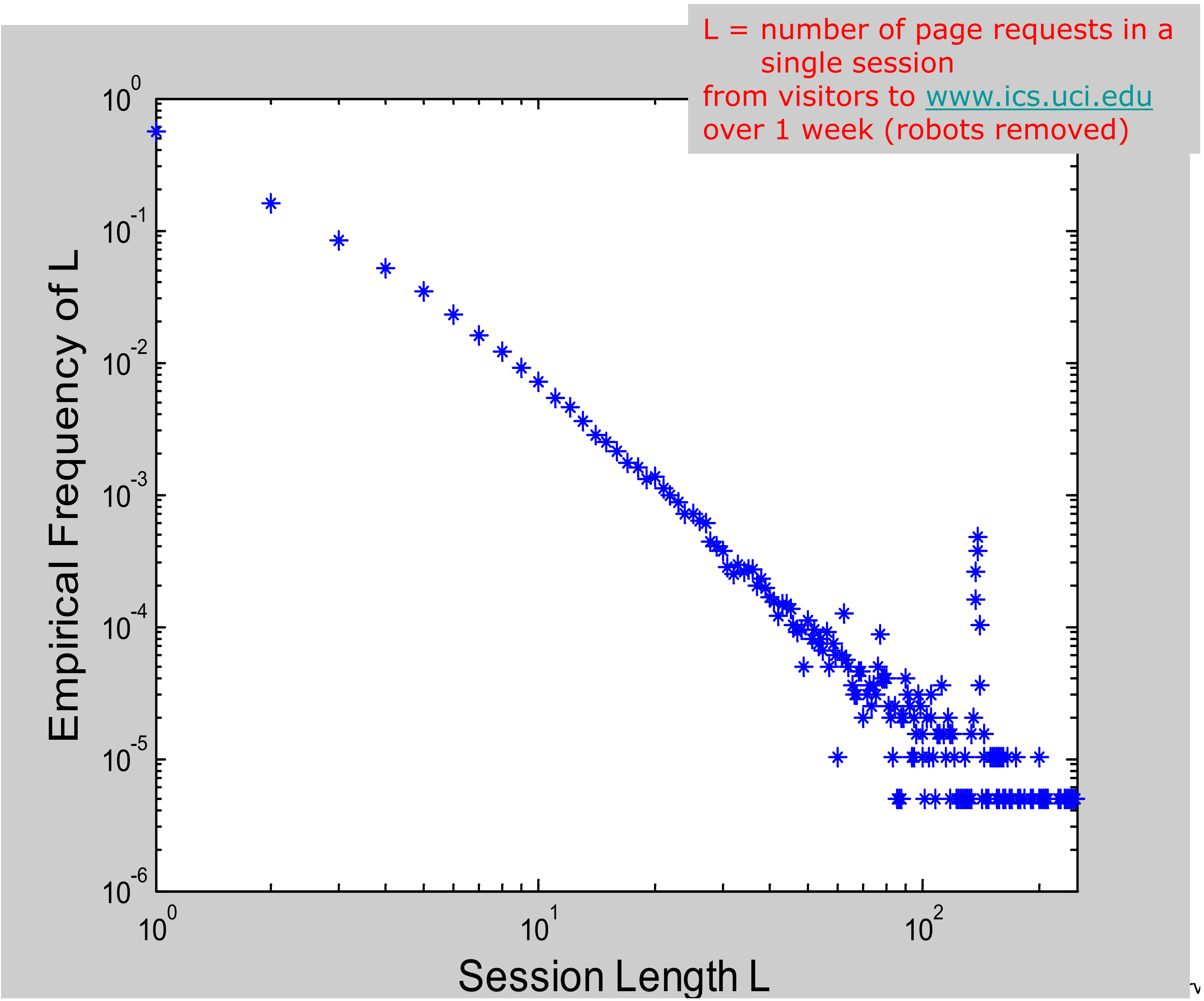
- Often data have skewed distribution, e.g., accesses to Web site, number of friends a person has, sizes of files on your computer.

- One way to model these is a power law

$$p(x) = C^{-1} x^{-\alpha} \qquad \text{where} \quad \alpha > 1$$

- Key point, this is linear when you take logs

Example of Power Law



[Source: Padhraic Smyth]

Three-Dimensional Data

- Generally hard
- 3-D plots are not usually useful
- Usually better to use colour on a 2-D plot
- Or show multiple 2D plots for each value of third variable

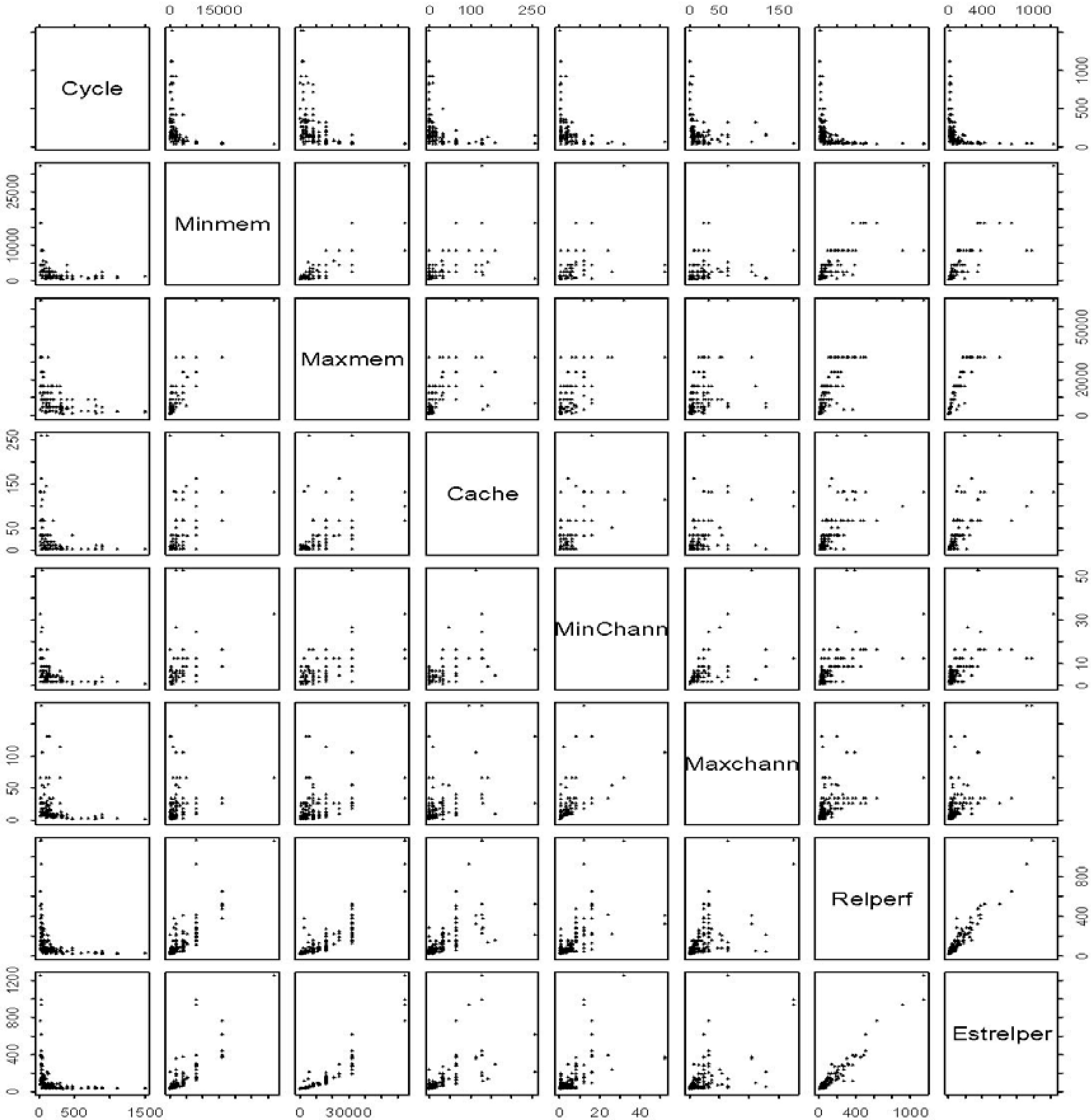
High-Dimensional Data

Always going to be hard.

Reason: Visualisation does not scale. (in number of pixels)

- Project the data down to 2-D
  - Many techniques
    - Principal Components Analysis (IAML, MLPR)
    - Multidimensional scaling
    - Modern nonlinear methods: t-SNE, LLE, Isomap, Eigenmaps
  - Problem: Sometimes this will obscure high-D structure and nonlinear structure
- Another option: Scatterplot matrix (see next)

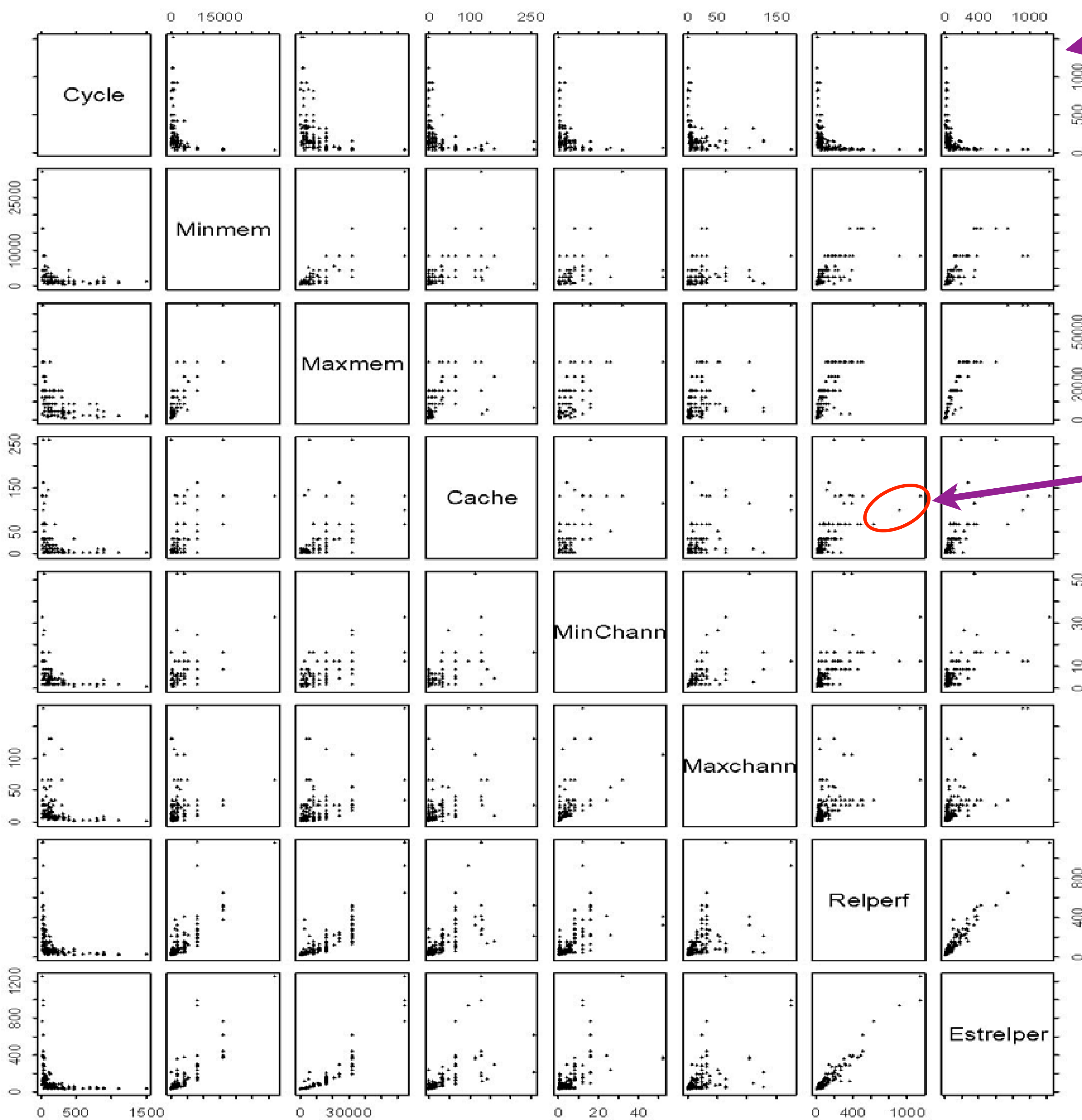
Scatterplot matrix



This is performance data for (very old) CPUs

Important:
Scales must be matched

Scatterplot matrix



Maybe want to use transformed variables up here

Might be worth understanding points like these

This row is the variable we want to predict

This is the prediction according to somebody's model (explains strong relationship)

Another complicated case

What if individual features don't mean as much?

Text, images, audio

- Text: Can summarise individual documents (tf-idf, topics)
- For images / audio, probably reduced to navigation via clustering
- Or can project the data down to 2-D again

Principles

- Visualisation is essential but not scalable (in dimension or data size)
- For exploration, simple is good
  - Histograms and scatterplots rule. Fancy 3-D graphs, meh.
- Principle of small multiples
- Color is the fourth dimension.
  - Time not so useful.
- Understand the axes. Scaling and transforming.
  - (always check the axes on small multiples)
- If something looks weird, figure out why.
  - Find anomalies. Data are always suspect.
- Relationships are about managing expectations
  - What relationships do you *expect* to exist? Can you *see* them?
- Use visualization to inform models and vice versa
  - Feature construction, debugging

“Dual-purpose” principles

(both exploratory and presentation analysis)

- Make sure the axis labels aren’t too small
- What the reader wants to compare, make it easy to compare
- Revise figures often, same as text

not

Making it easy to compare

Something I often see in MSc theses…

	Size of training set	100	10,000	100,000
SVM	Features A	0.3	0.15	0.12
	Features B	0.32	0.17	0.13
	Features C	0.35	0.19	0.14
k-NN	Features A	0.4	0.25	0.20
	Features B	0.42	0.27	0.18
	Features C	0.38	0.3	0.25
Logistic regression	Features A	0.35	0.2	0.14
	Features B	0.36	0.22	0.16
	Features C	0.33	0.19	0.15

If you really like this stuff

- Tukey, Exploratory Data Analysis
- Bill Cleveland, Visualizing Data
- Edward Tufte, all books