CS 277, Data Mining

Exploratory Data Analysis

Padhraic Smyth
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine

Outline

Assignment 1: Questions?

Today's Lecture: Exploratory Data Analysis

- Analyzing single variables
- Analyzing pairs of variables
- Higher-dimensional visualization techniques

Next Lecture: Clustering and Dimension Reduction

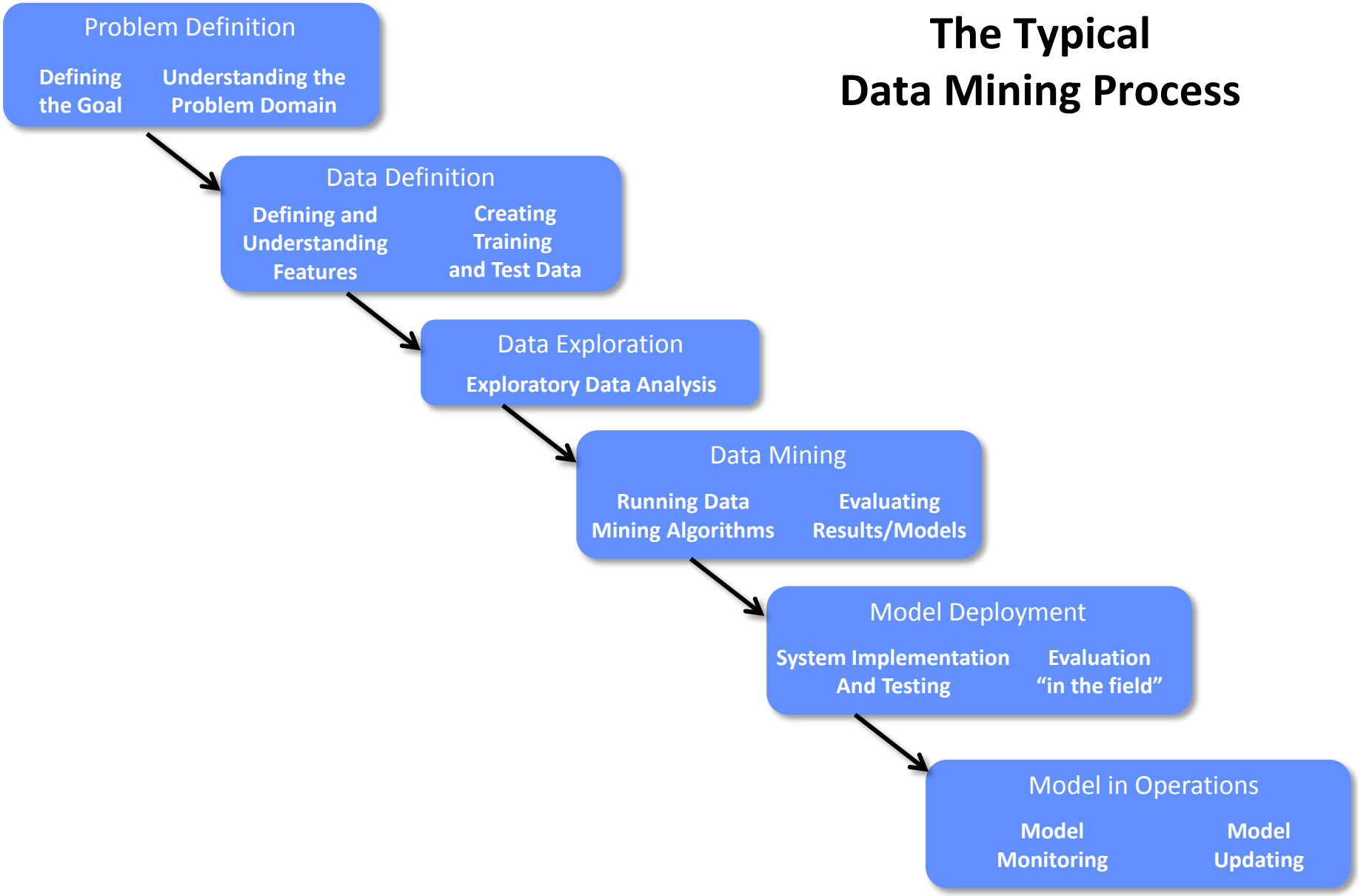
- Dimension reduction methods
- Clustering methods

What are the main Data Mining Techniques?

- Descriptive Methods
  - Exploratory Data Analysis, Visualization
  - Dimension reduction (principal components, factor models, topic models)
  - Clustering
  - Pattern and Anomaly Detection
  - ….and more

- Predictive Modeling
  - Classification
  - Ranking
  - Regression
  - Matrix completion (recommender systems)
  - …and more

The Typical Data Mining Process



Exploratory Data Analysis: Single Variables

Summary Statistics

Mean: “center of data”

Mode: location of highest data density

Variance: “spread of data”

Skew: indication of non-symmetry

Range: max - min

Median: 50% of values below, 50% above

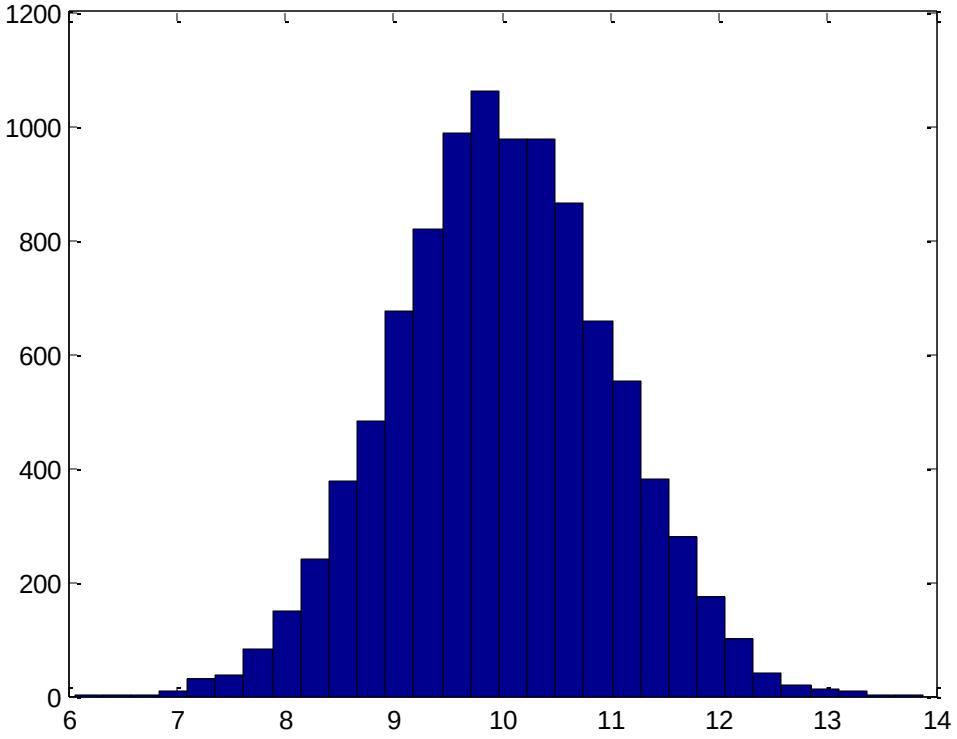
Quantiles: e.g., values such that 25%, 50%, 75% are smaller

Note that some of these statistics can be misleading

E.g., mean for data with 2 clusters may be in a region with zero data

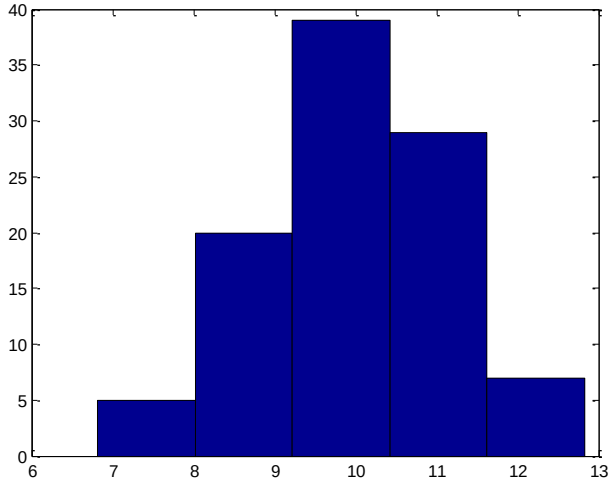
Histogram of Unimodal Data

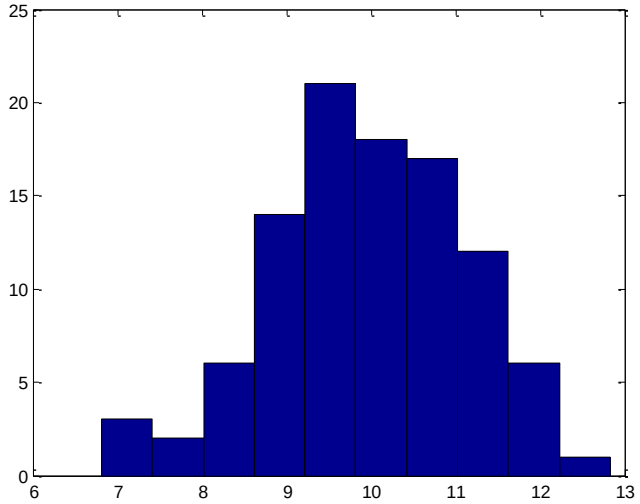
1000 data points simulated from a Normal distribution, mean 10, variance 1, 30 bins

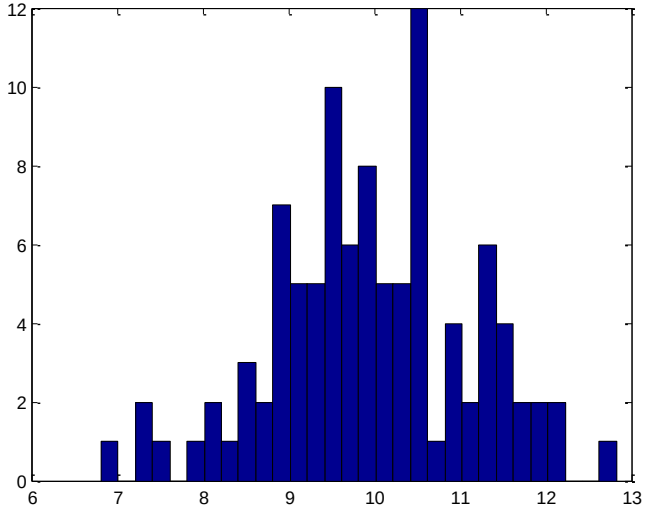


Histograms: Unimodal Data

100 data points from a Normal, mean 10, variance 1, with 5, 10, 30 bins

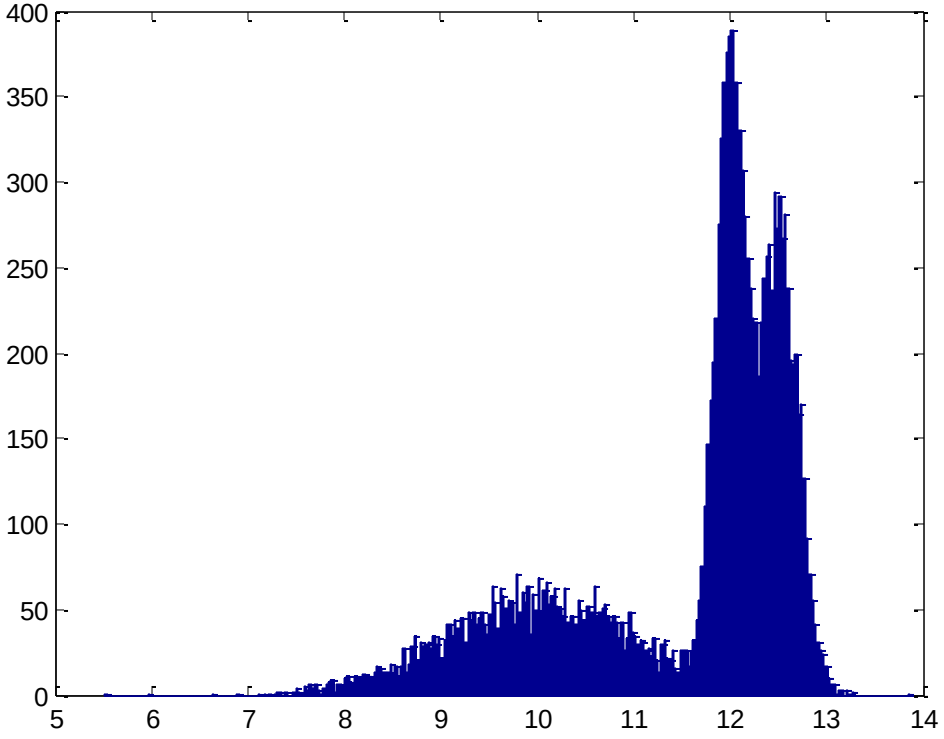






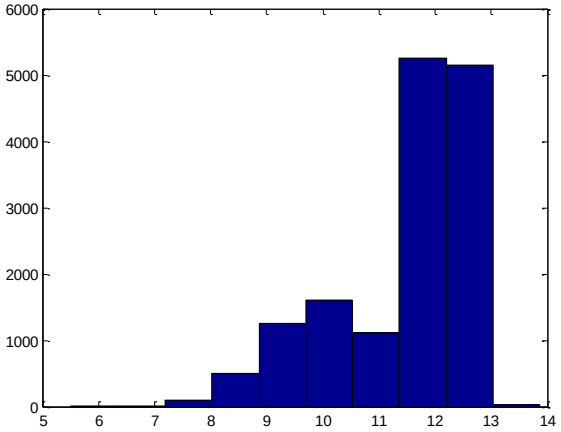
Histogram of Multimodal Data

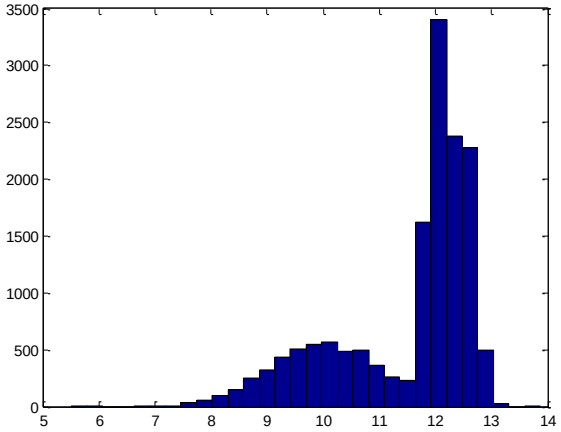
15000 data points simulated from a mixture of 3 Normal distributions, 300 bins

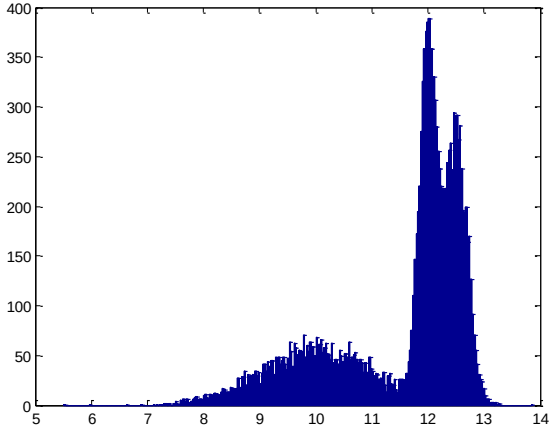


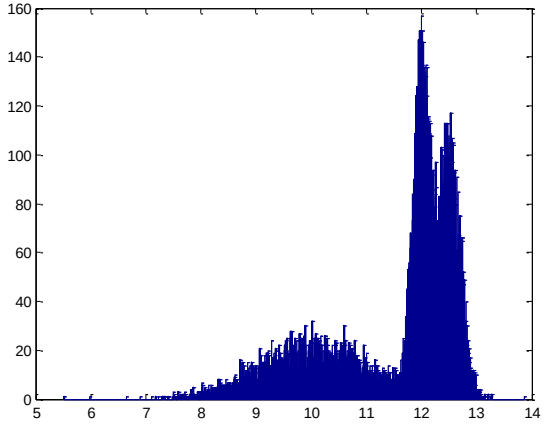
Histogram of Multimodal Data

15000 data points simulated from a mixture of 3 Normal distributions, 300 bins



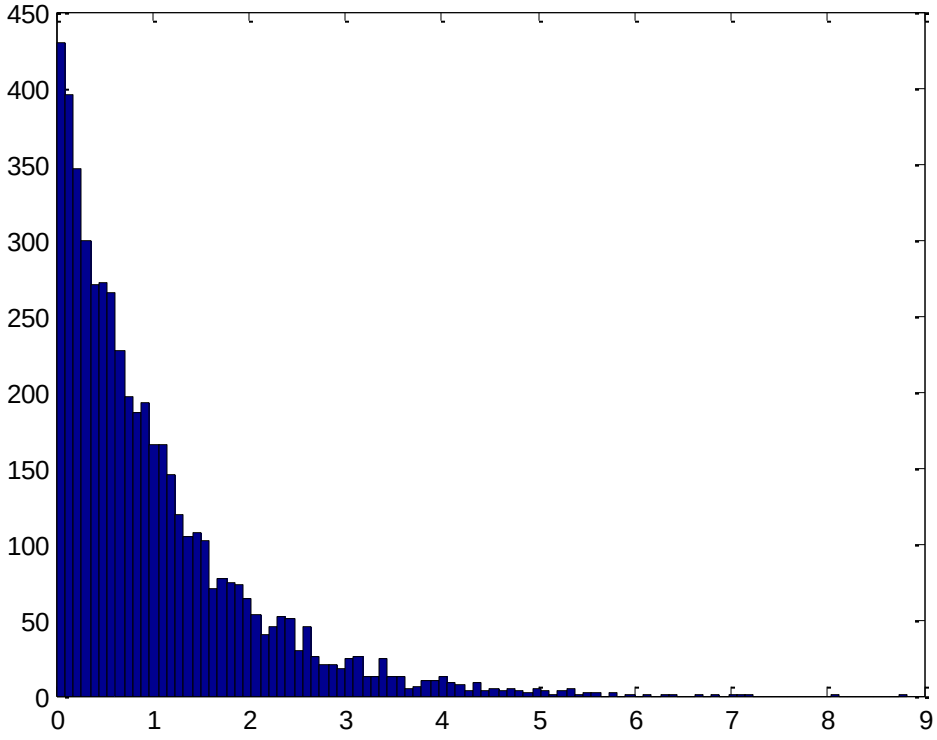






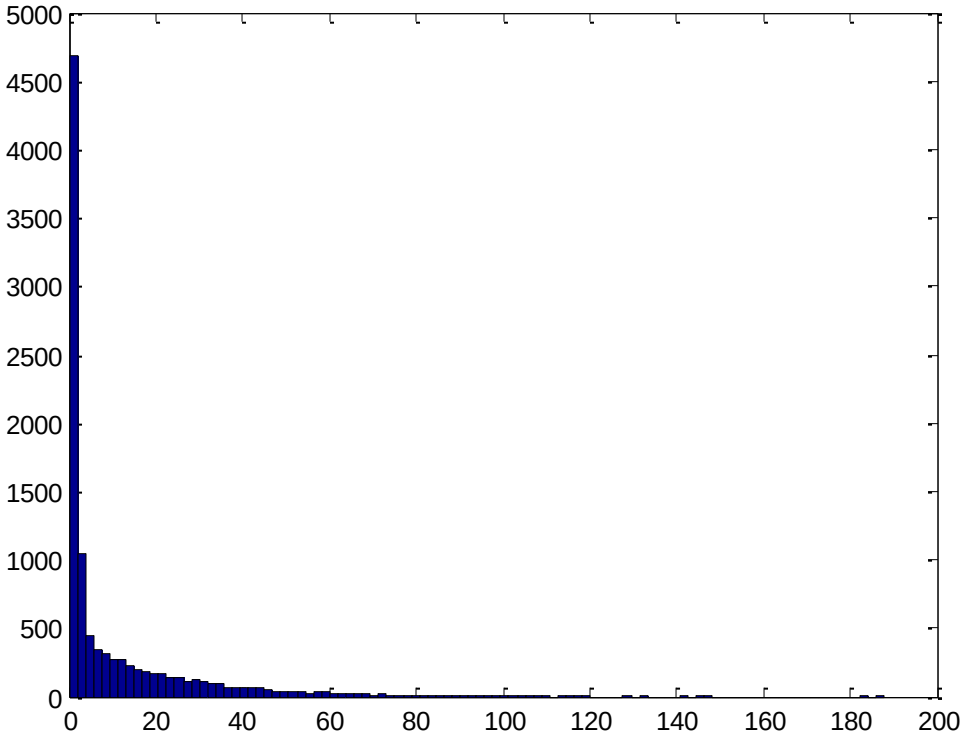
Skewed Data

5000 data points simulated from an exponential distribution, 100 bins



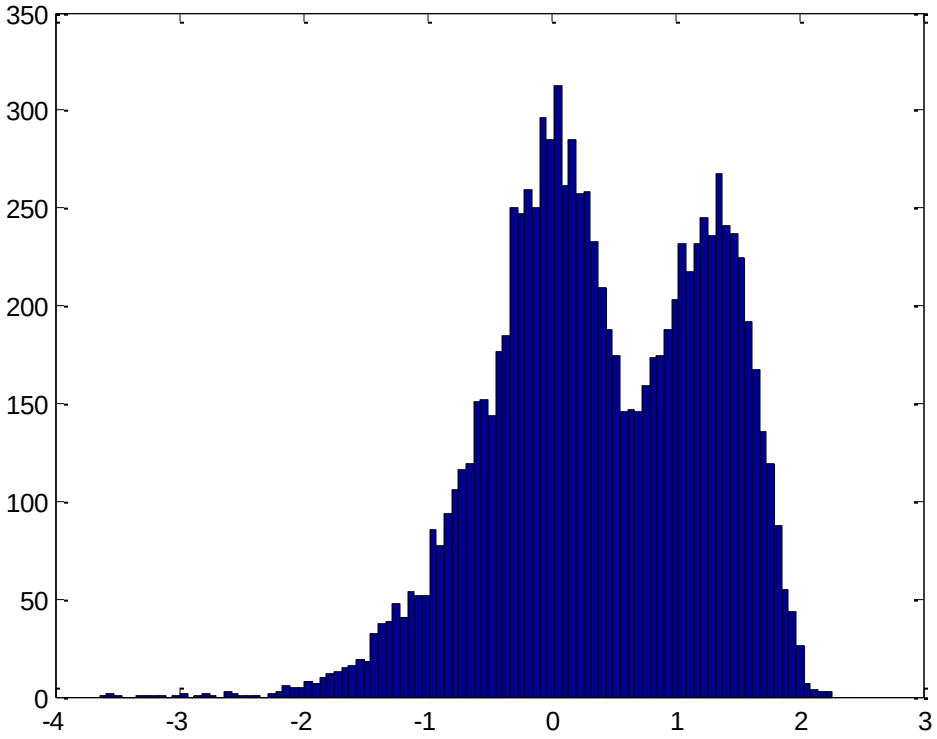
Another Skewed Data Set

10000 data points simulated from a mixture of 2 exponentials, 100 bins

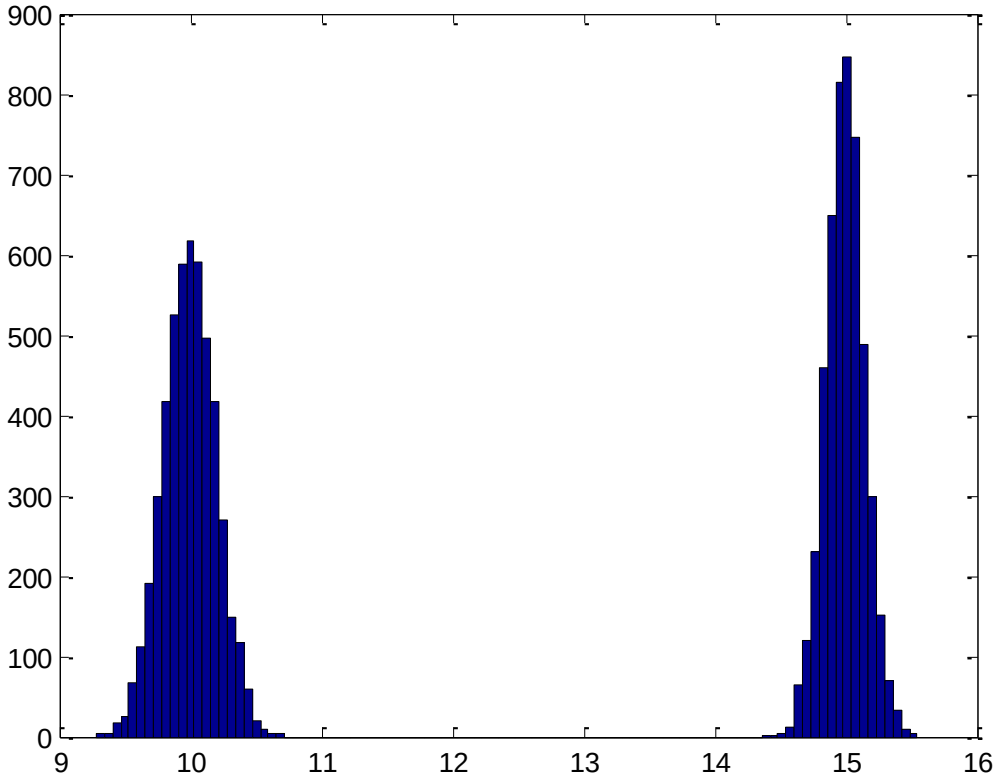


Same Skewed Data after taking Logs (base 10)

10000 data points simulated from a mixture of 2 exponentials, 100 bins



What will the mean or median tell us about this data?



Issues with Histograms

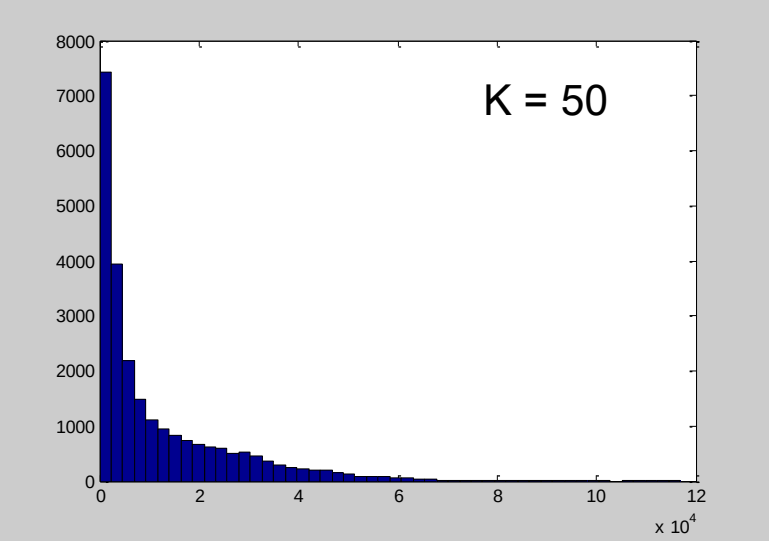
- For small data sets, histograms can be misleading. Small changes in the data or to the bucket boundaries can result in very different histograms.

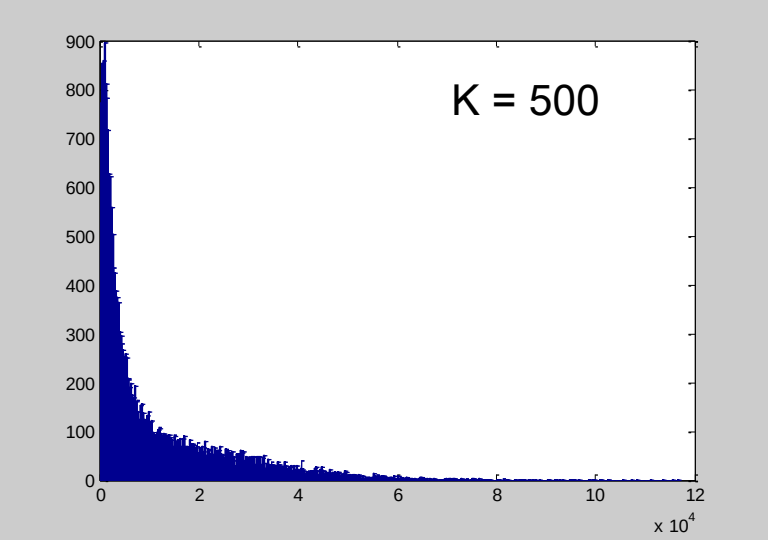
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.

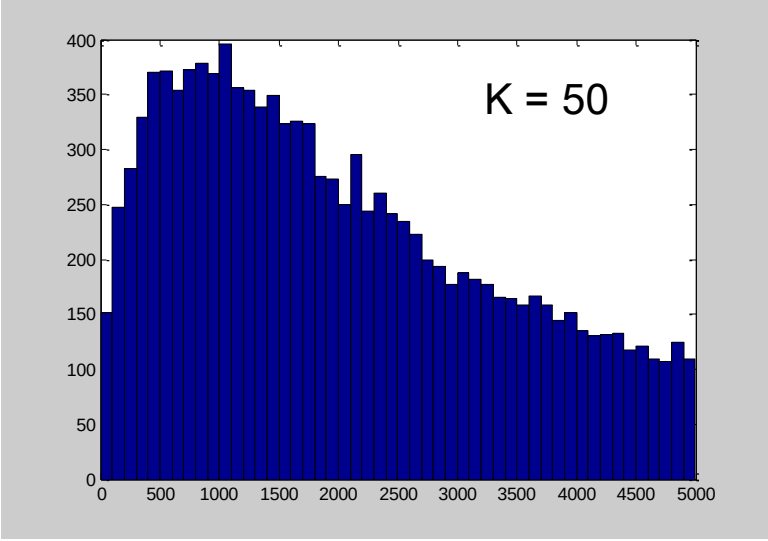
- Can smooth histogram using a variety of techniques
  - E.g., kernel density estimation, which avoids bins – but requires some notion of "scale"

- Histograms effectively only work with 1 variable at a time
  - Difficult to extend to 2 dimensions, not possible for >2
  - So histograms tell us nothing about the relationships among variables

US Zipcode Data: Population by Zipcode

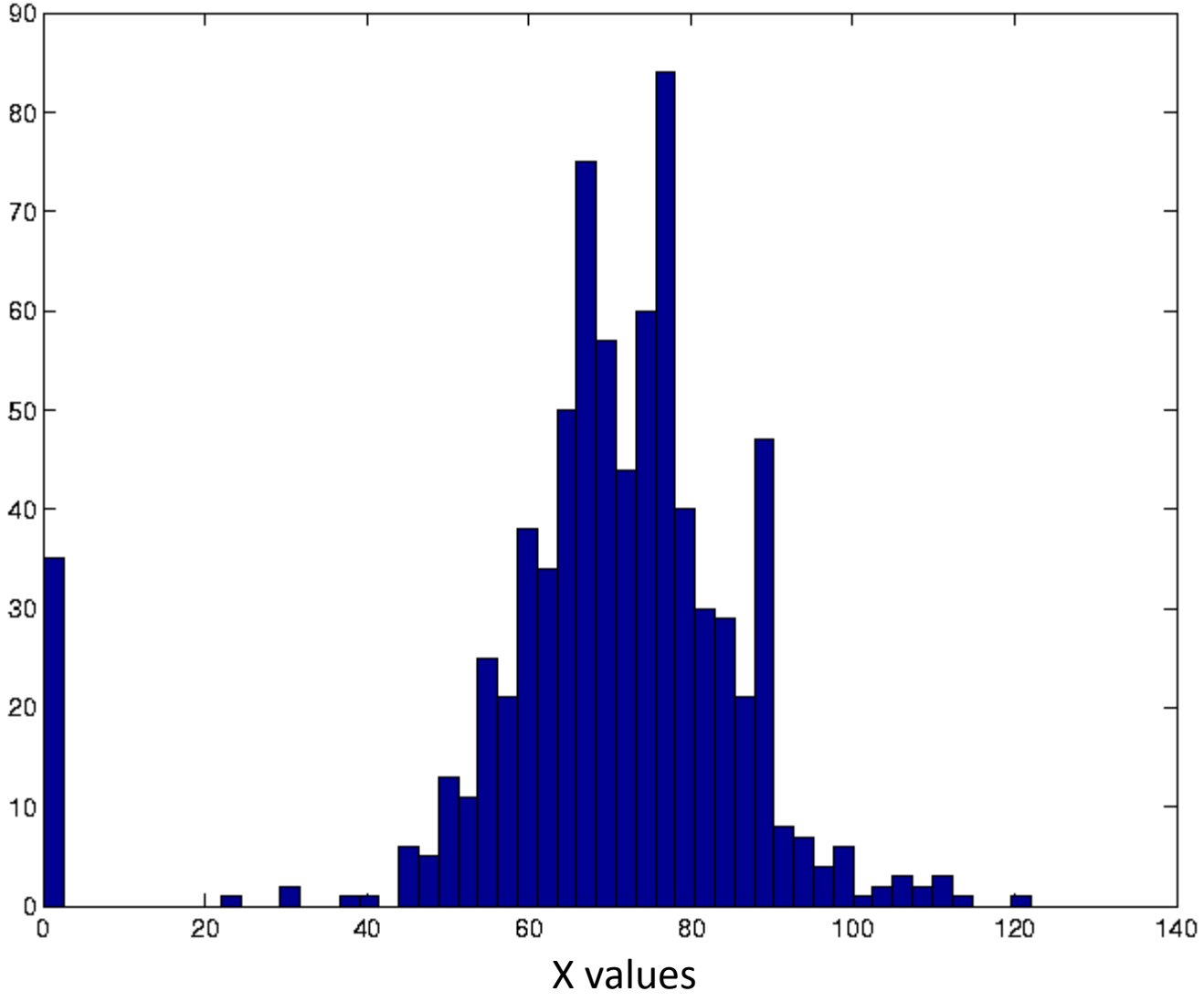






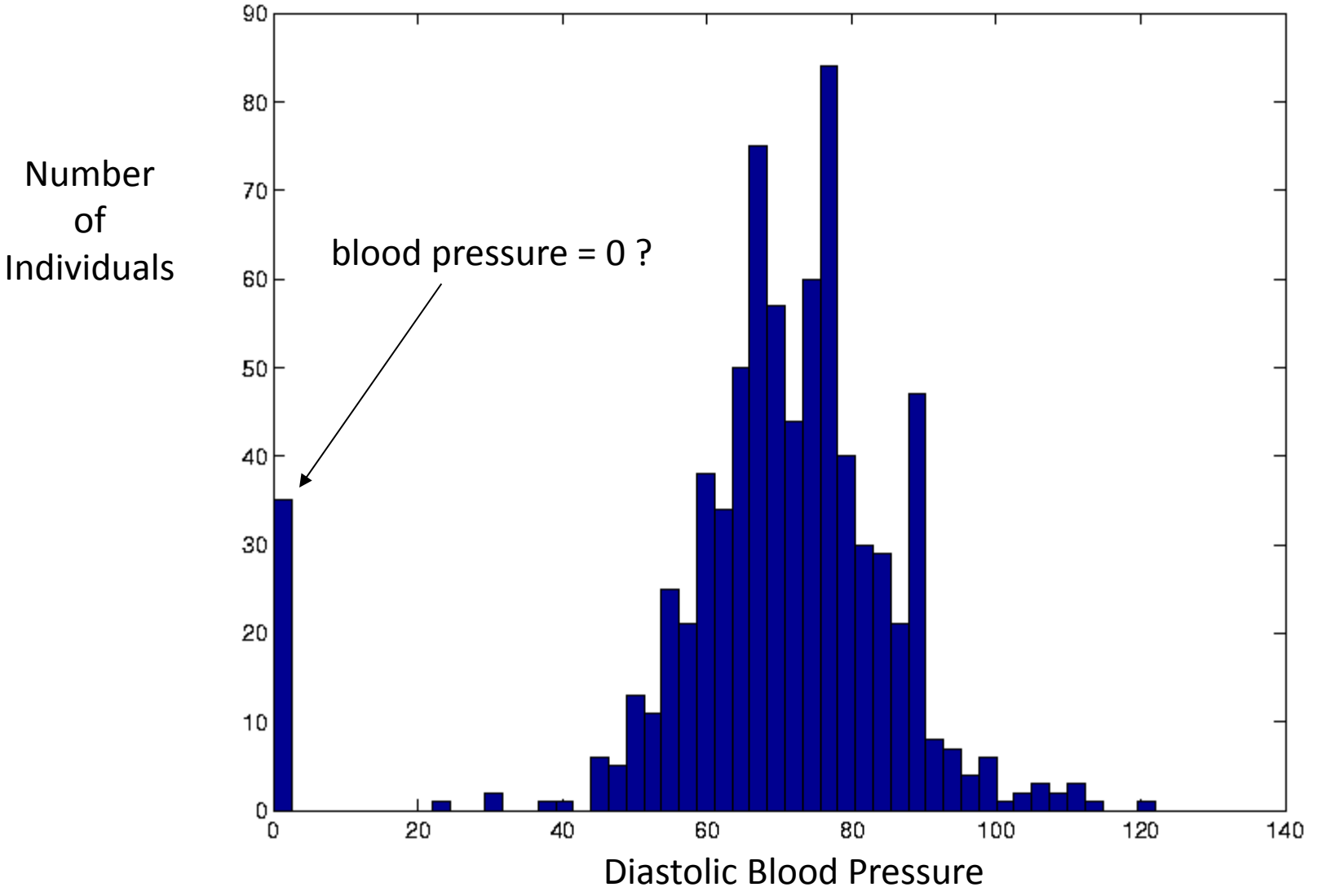
Histogram with Outliers

Pima Indians Diabetes Data,
From UC Irvine Machine Learning Repository



Histogram with Outliers

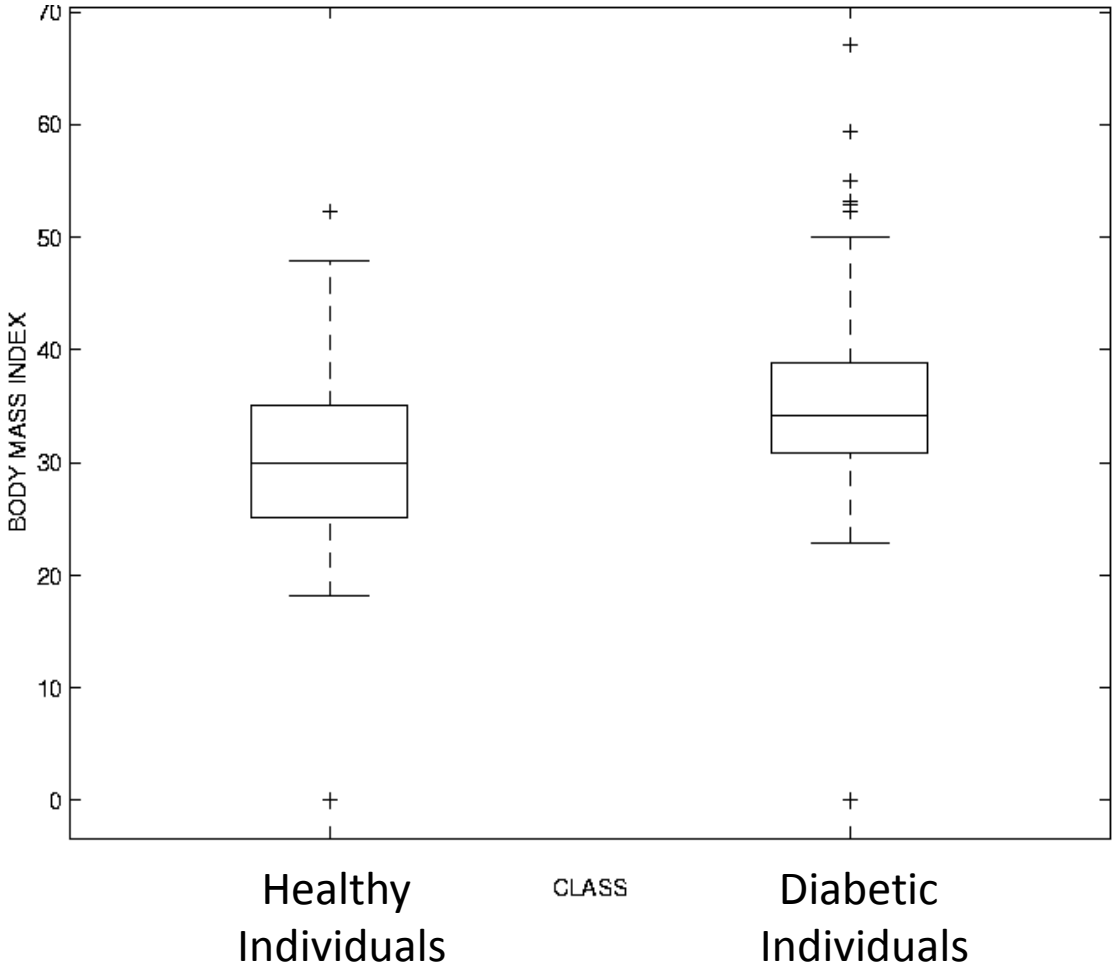
Pima Indians Diabetes Data,
From UC Irvine Machine Learning Repository



Box Plots: Pima Indians Diabetes Data

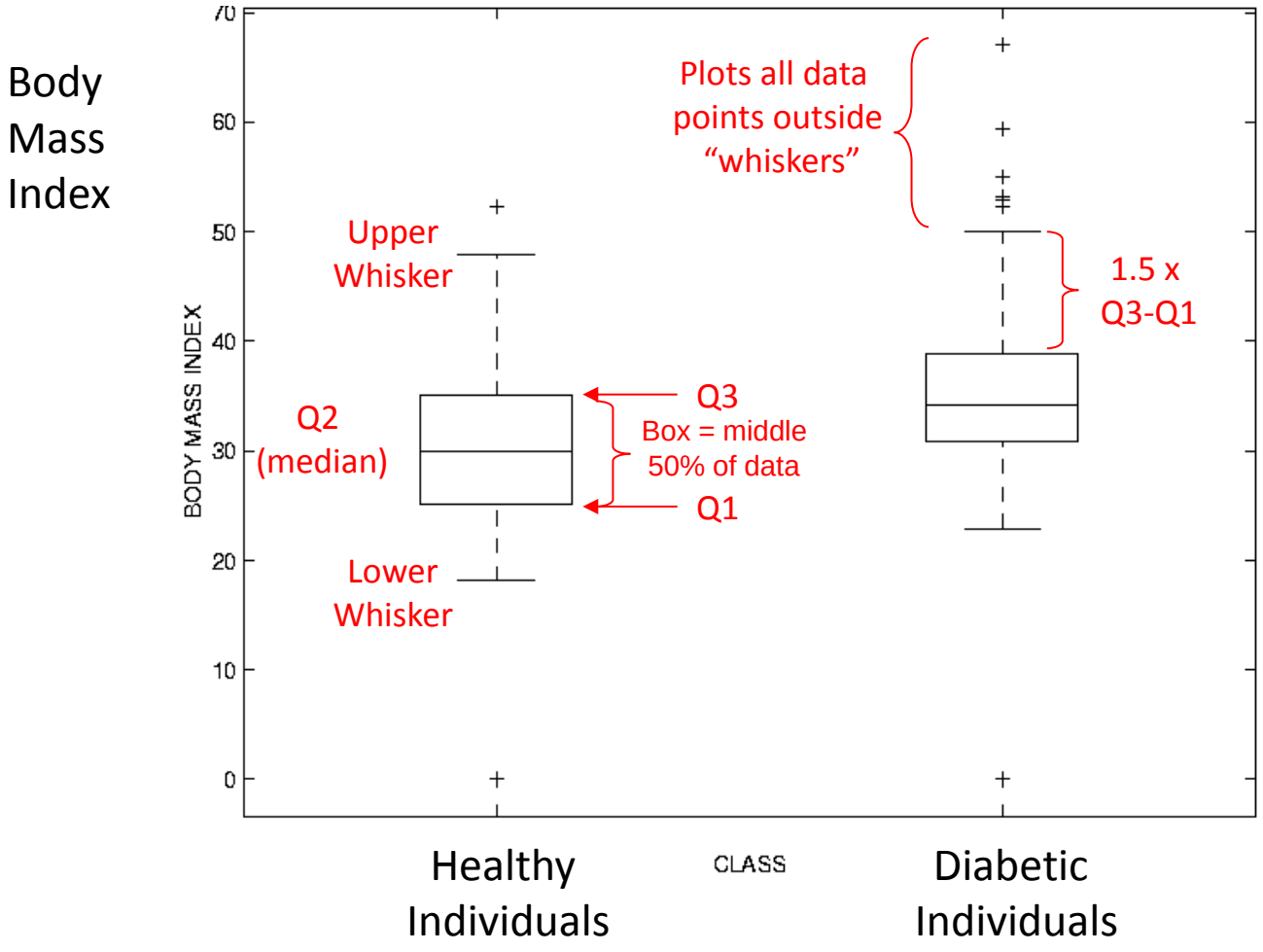
Two side-by-side box-plots of individuals from the Pima Indians Diabetes Data Set

Body Mass Index

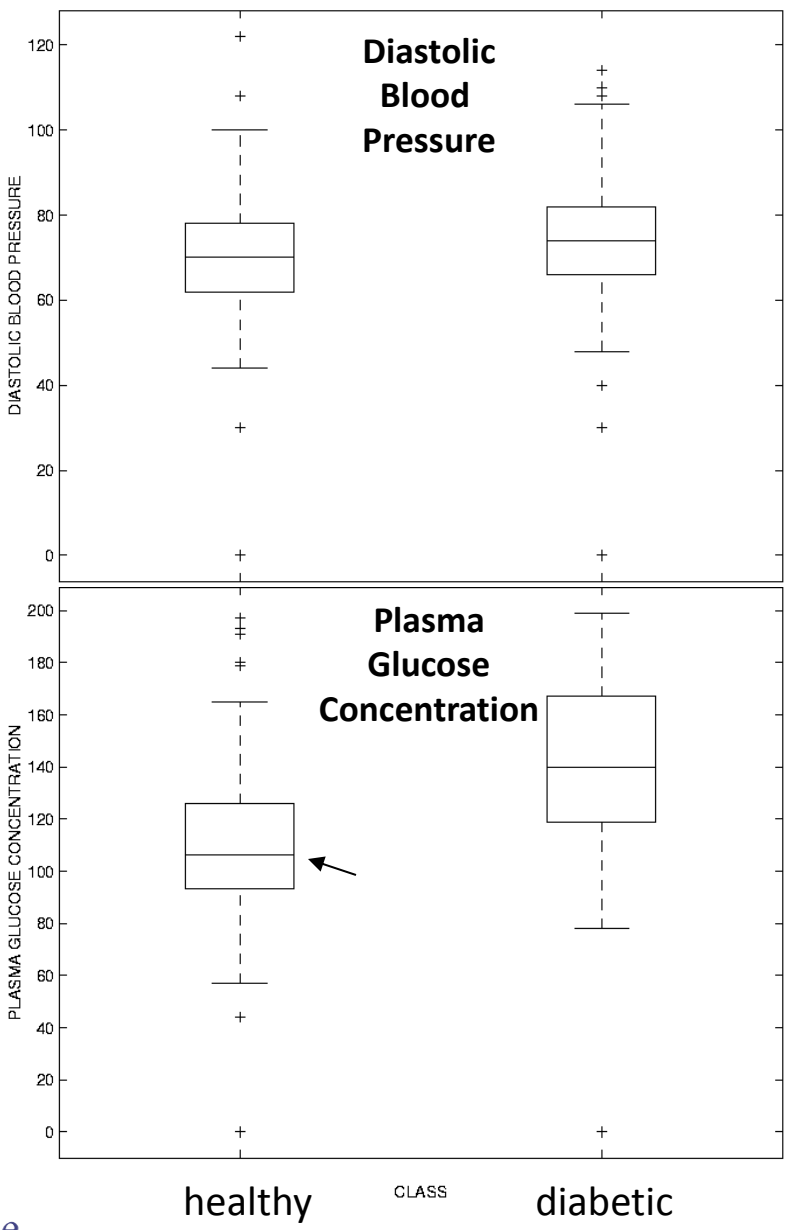


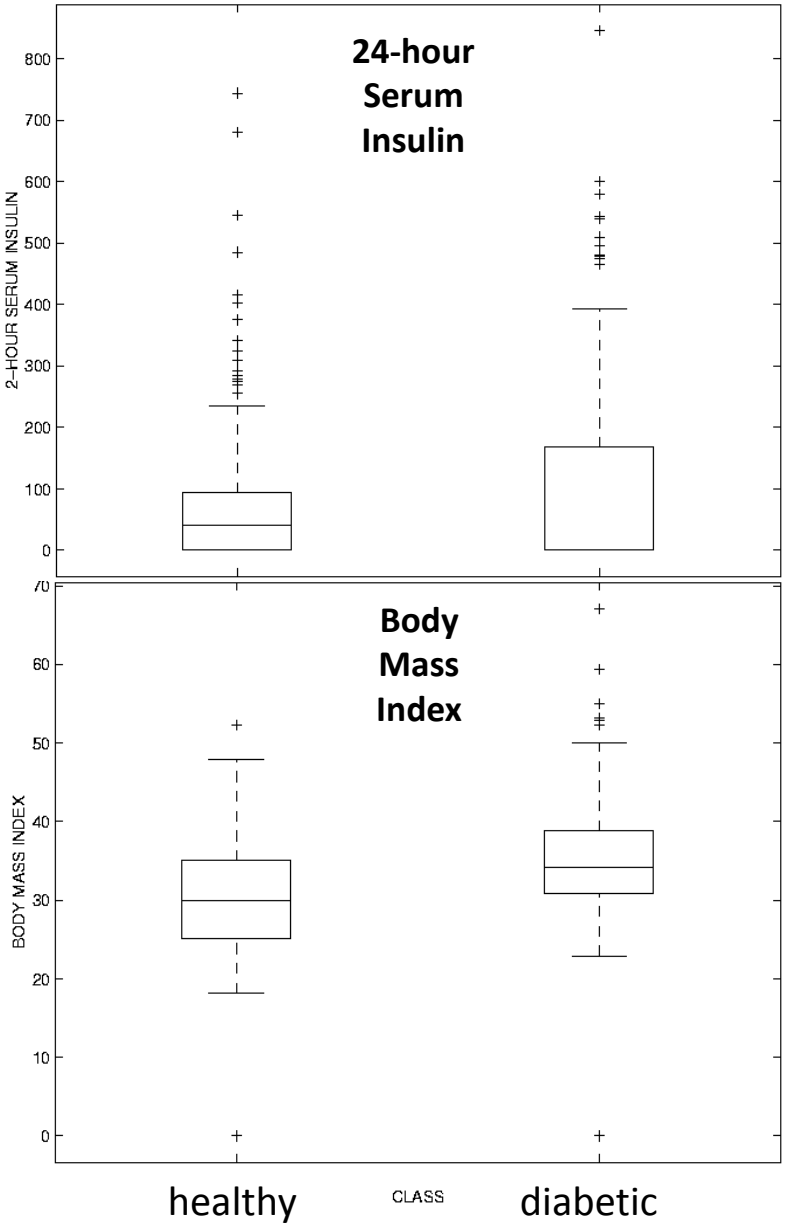
Box Plots: Pima Indians Diabetes Data

Two side-by-side box-plots of individuals from the Pima Indians Diabetes Data Set



Box Plots: Pima Indians Diabetes Data





Exploratory Data Analysis

Analyzing more than 1 variable at a time…

Relationships between Pairs of Variables

- Say we have a variable Y we want to predict and many variables X that we could use to predict Y

- In exploratory data analysis we may be interested in quickly finding out if a particular X variable is potentially useful at predicting Y

- Options?
  - Linear correlation

  - Scatter plot: plot Y values versus X values

Linear Dependence between Pairs of Variables

- Covariance and correlation measure linear dependence

- Assume we have two variables or attributes X and Y and n objects taking on values x(1), …, x(n) and y(1), …, y(n). The sample covariance of X and Y is:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})$$

- The covariance is a measure of how X and Y vary together.
  - it will be large and positive if large values of X are associated with large values of Y and small X $\Rightarrow$ small Y

- (Linear) Correlation = scaled covariance, varies between -1 and 1

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\left(\sum_{i=1}^{n}(x(i)-\bar{x})^2\sum_{i=1}^{n}(y(i)-\bar{y})^2\right)^{\frac{1}{2}}}$$

Correlation coefficient

- Covariance depends on ranges of X and Y
- Standardize by dividing by standard deviation
- Linear correlation coefficient is defined as:

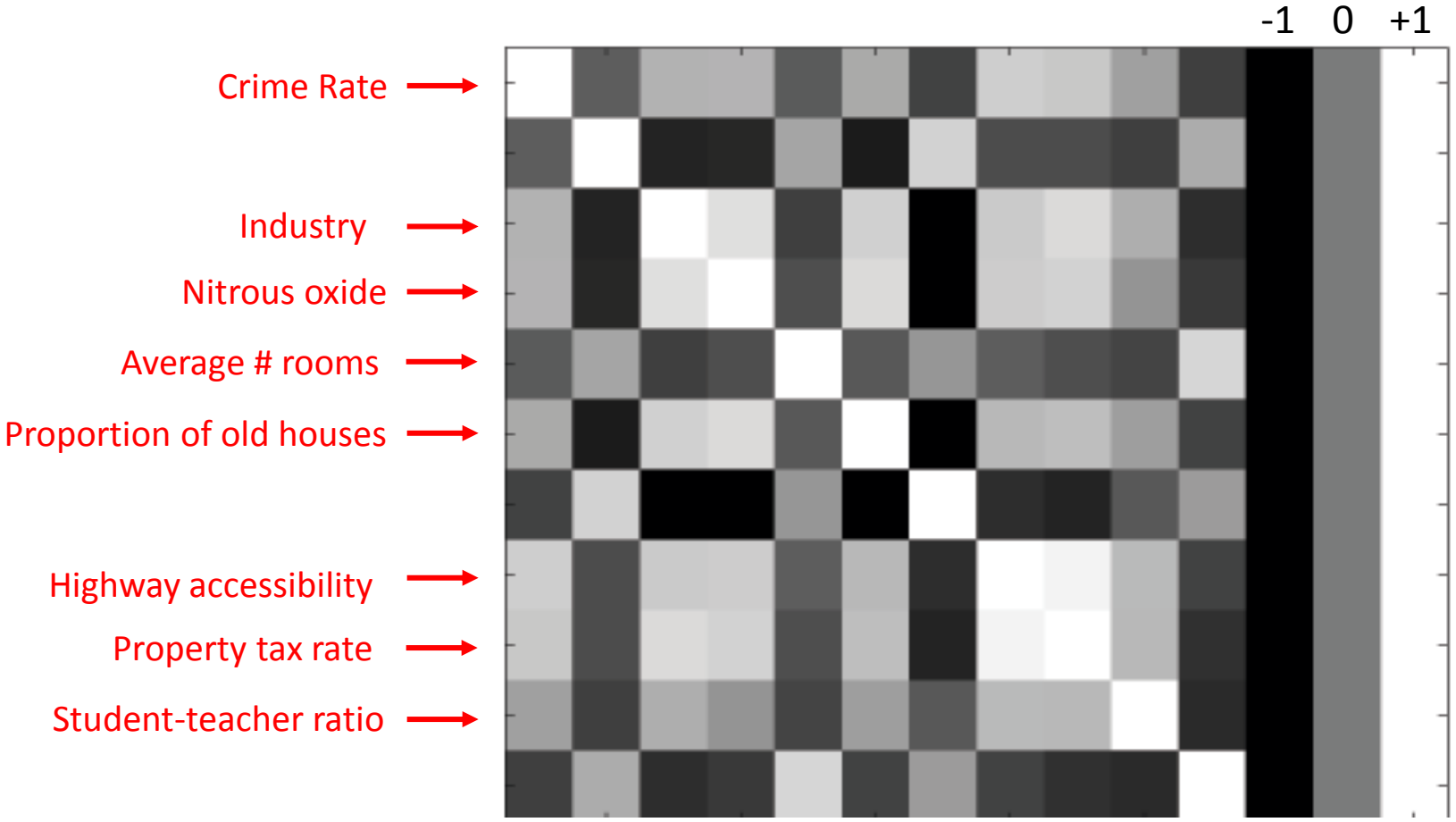
$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x(i)-\bar{x})(y(i)-\bar{y})}{\left(\sum_{i=1}^{n}(x(i)-\bar{x})^2 \sum_{i=1}^{n}(y(i)-\bar{y})^2\right)^{\frac{1}{2}}}$$

Data Set on Housing Prices in Boston

(widely used data set in research on regression models)

1	CRIM	per capita crime rate by town
2	ZN	proportion of residential land zoned for lots over 25,000 ft^2
3	INDUS	proportion of non-retail business acres per town
4	NOX	Nitrogen oxide concentration (parts per 10 million)
5	RM	average number of rooms per dwelling
6	AGE	proportion of owner-occupied units built prior to 1940
7	DIS	weighted distances to five Boston employment centres
8	RAD	index of accessibility to radial highways
9	TAX	full-value property-tax rate per $10,000
10	PTRATIO	pupil-teacher ratio by town
11	MEDV	Median value of owner-occupied homes in $1000's

Matrix of Pairwise Linear Correlations



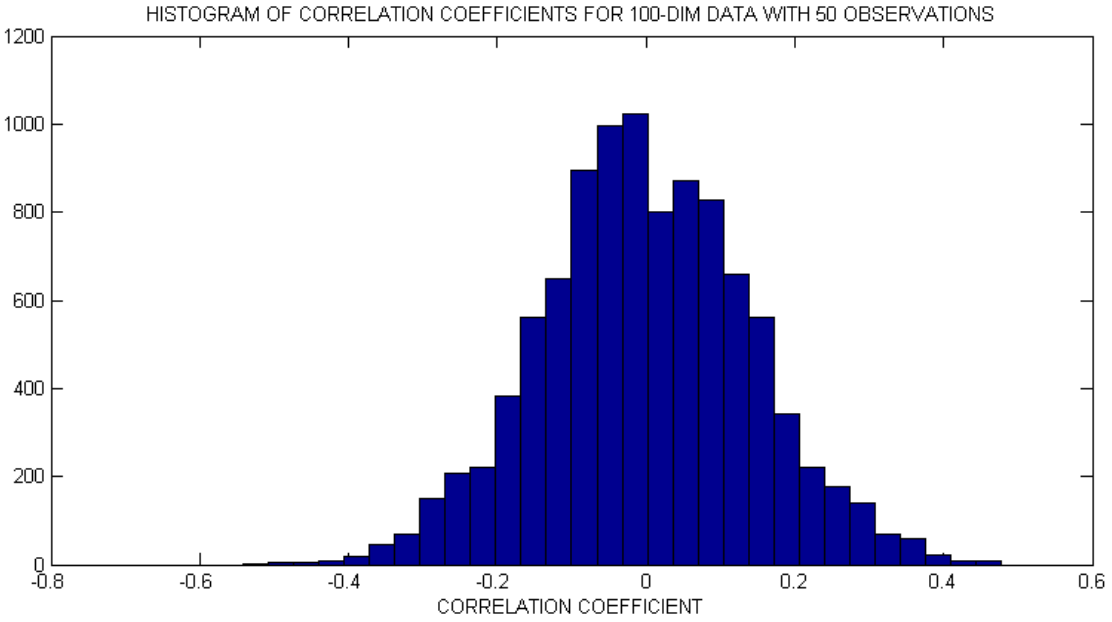
Data on characteristics of Boston housing

Dangers of searching for correlations in high-dimensional data

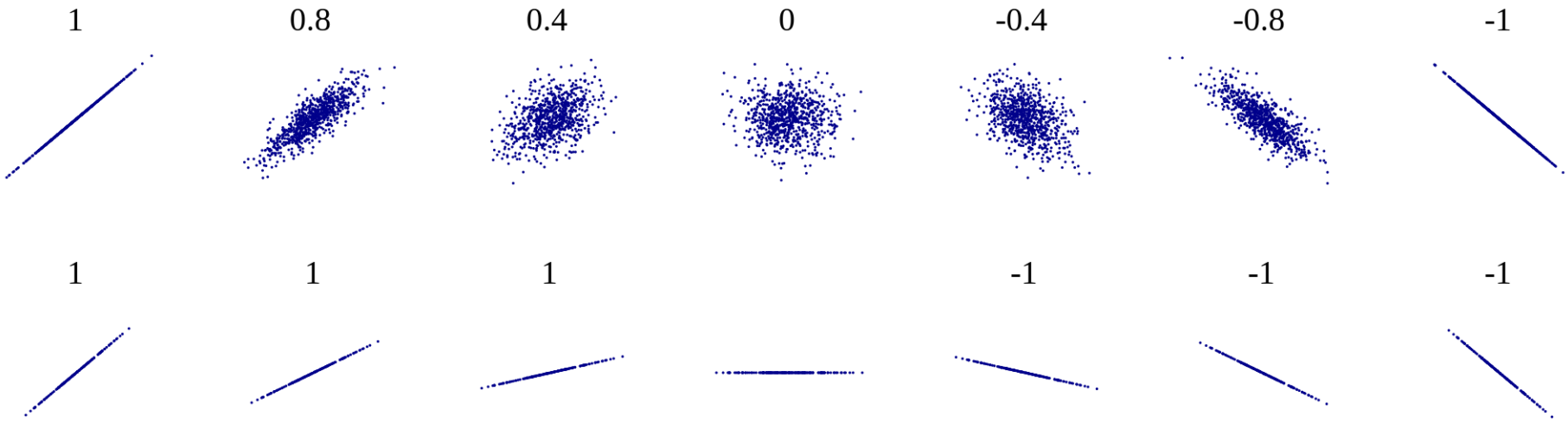
Simulated 50 random Gaussian/normal data vectors, each with 100 variables
Results in a 50 x 100 data matrix

Below is a histogram of the 100 choose 2 pairs of correlation coefficients

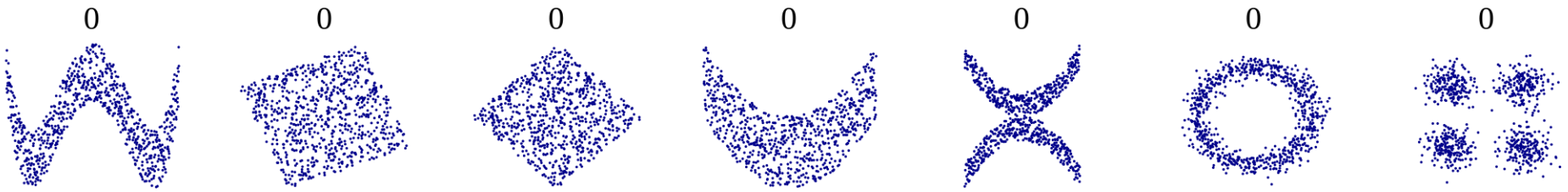
Even if data are entirely random (no dependence) there is a very high probability some variables will appear dependent just by chance.



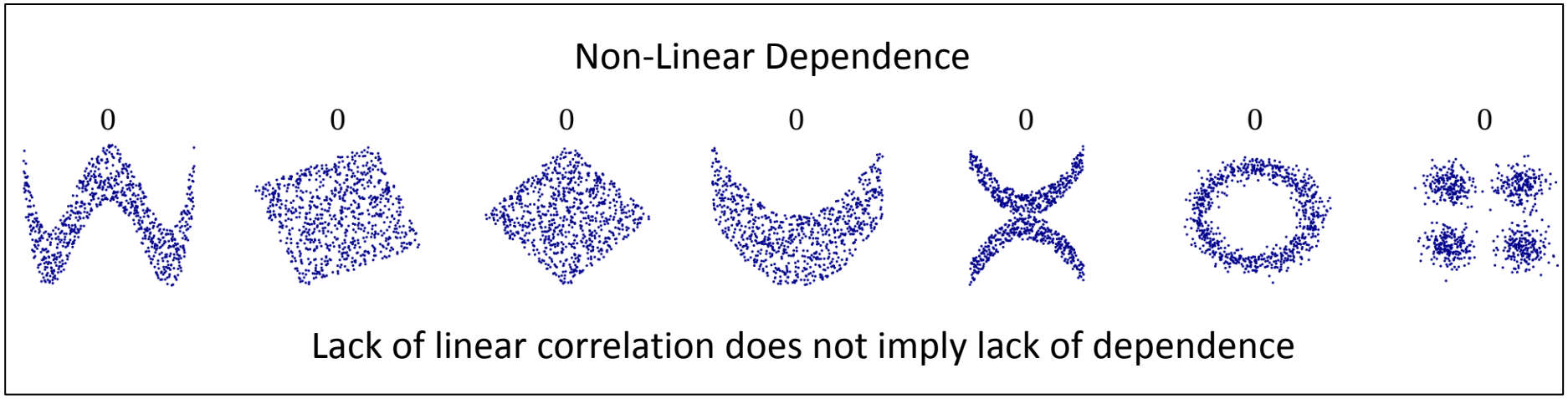
Examples of X-Y plots and linear correlation values

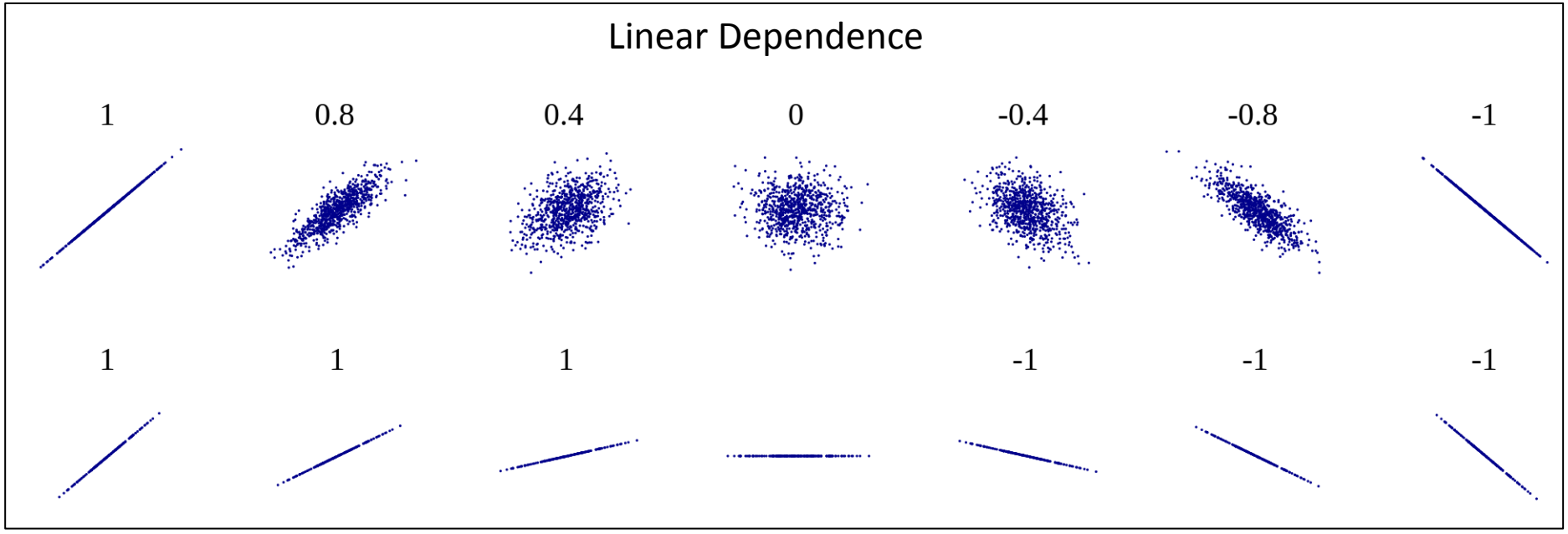


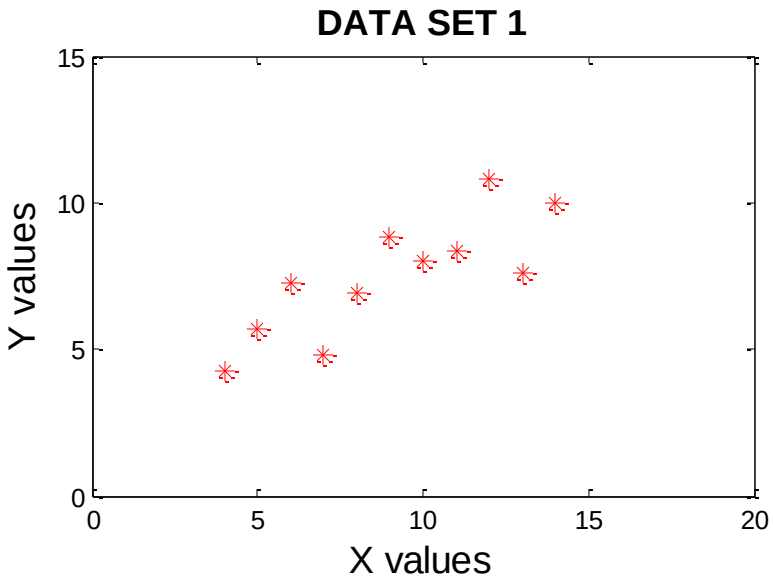
Examples of X-Y plots and linear correlation values

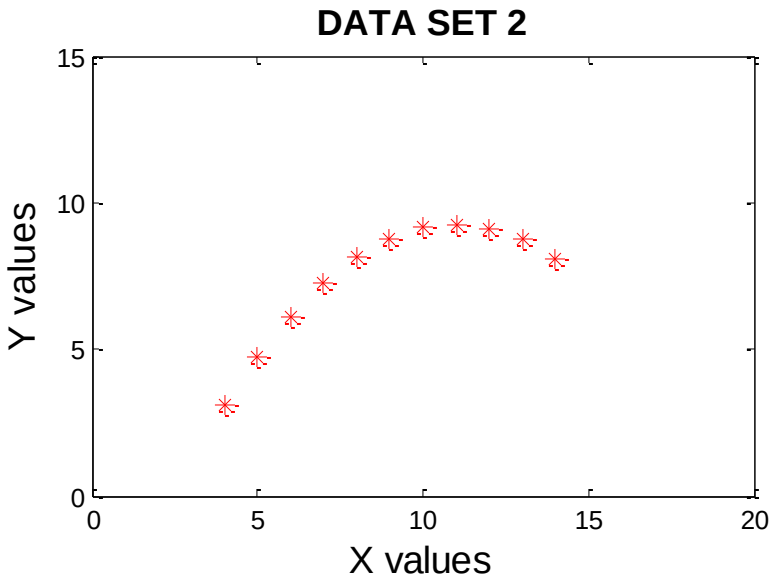


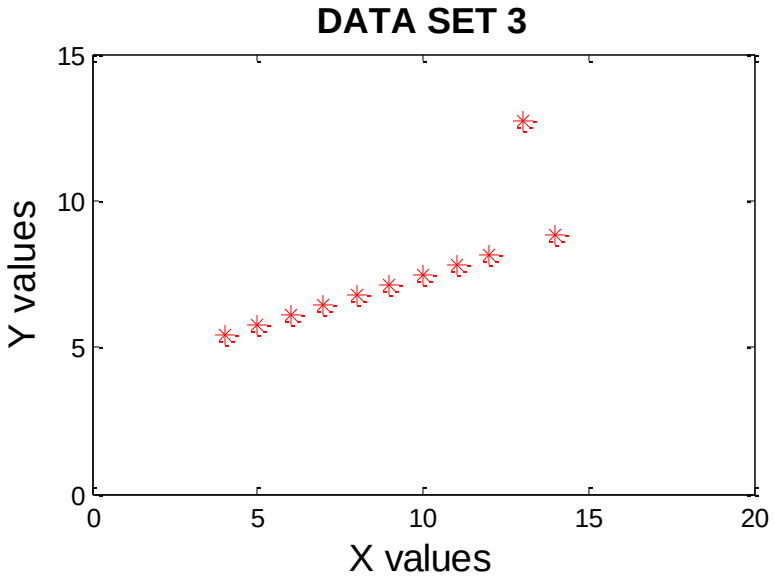
Examples of X-Y plots and linear correlation values

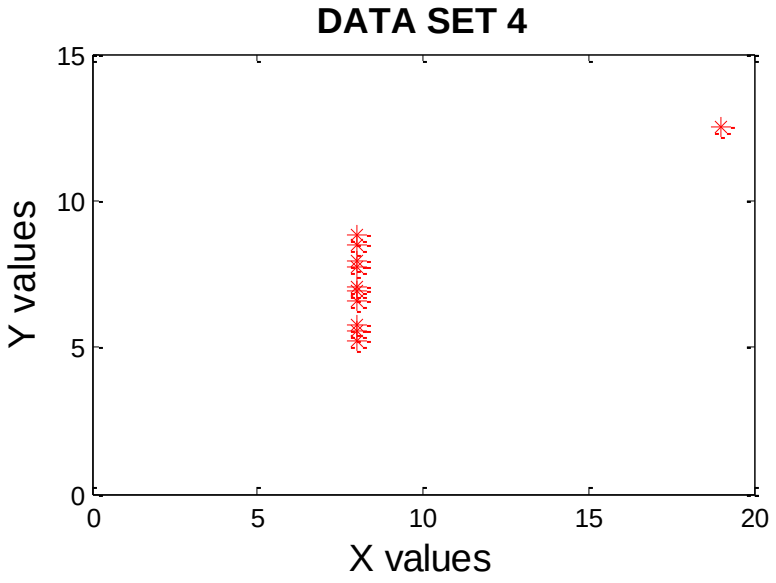






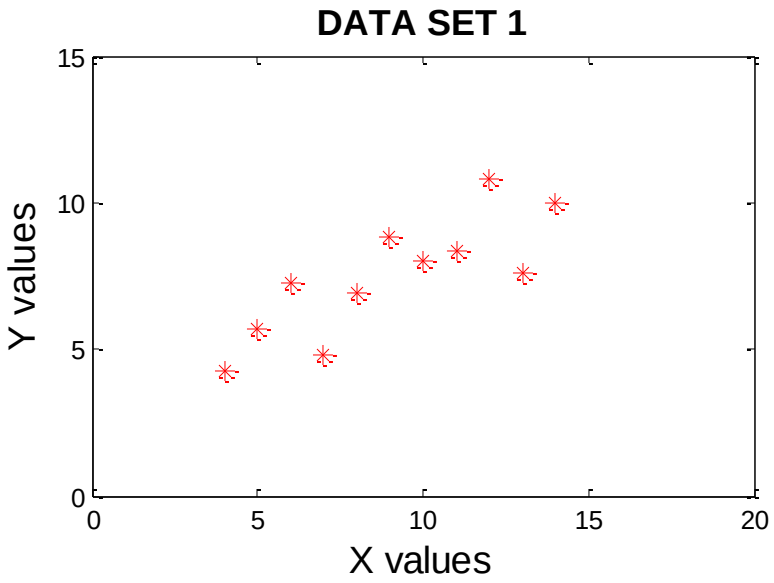


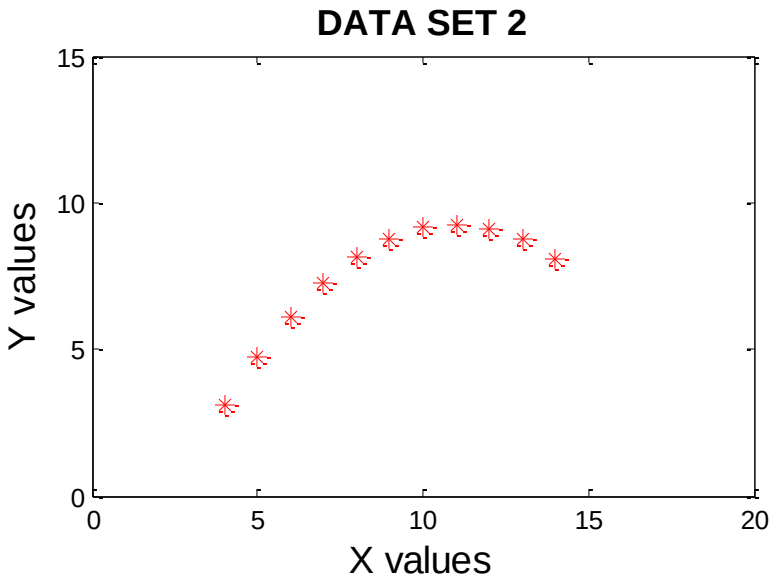


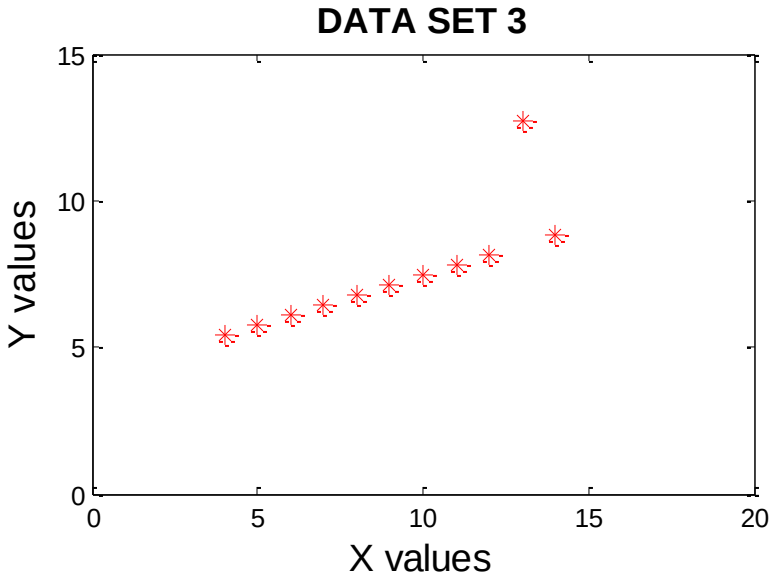


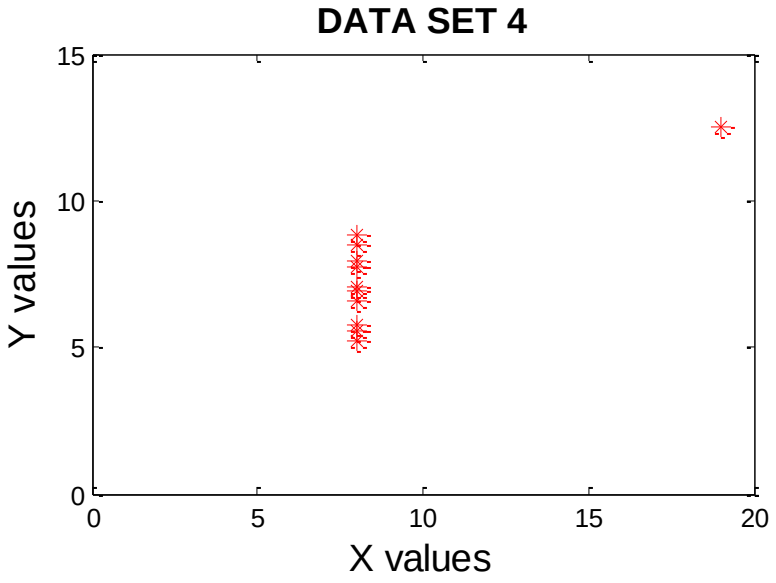
Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Guess the Linear Correlation Values for each Data Set



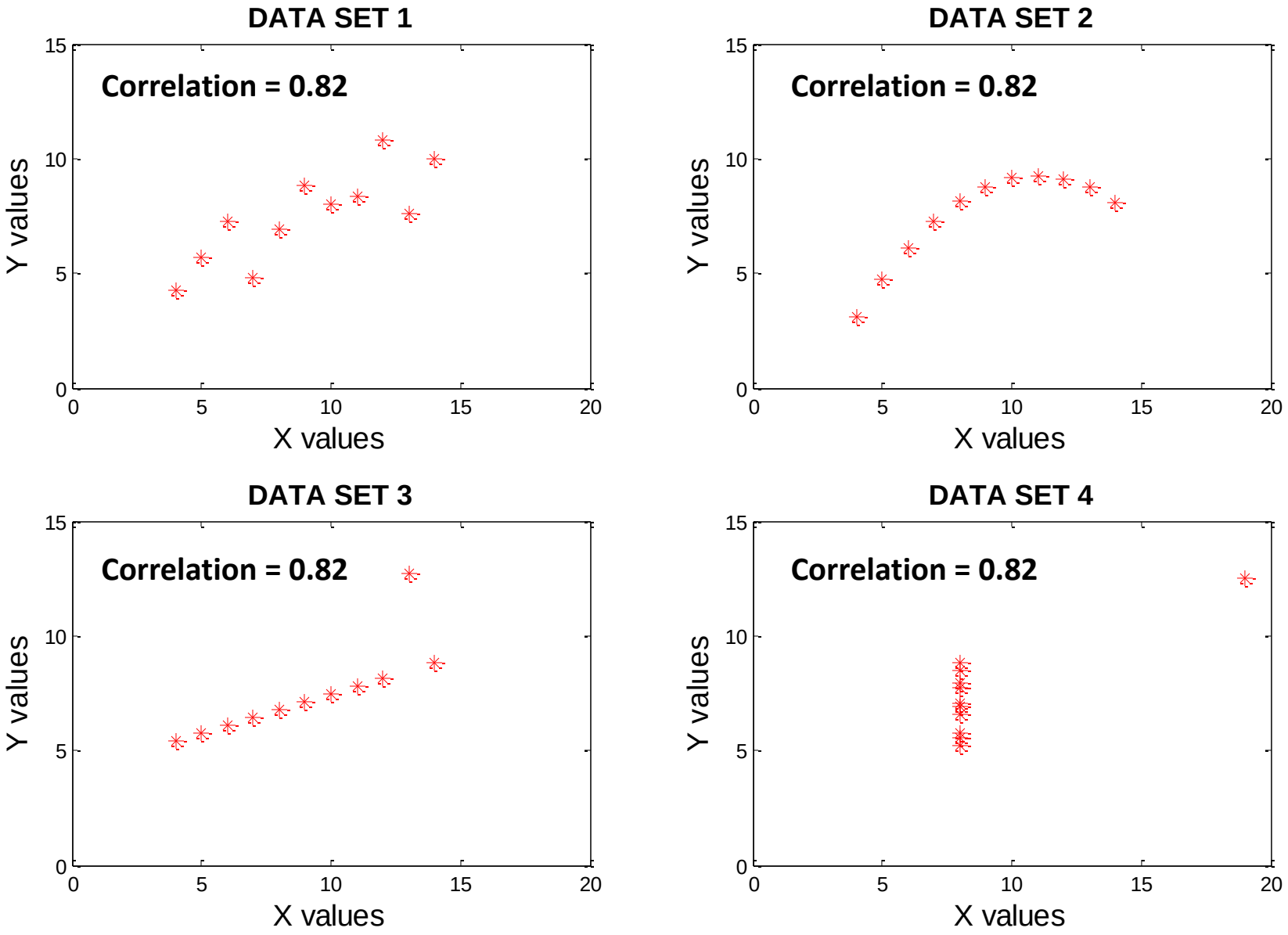






Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Actual Correlation Values



Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Summary Statistics for each Data Set

Summary Statistics of Data Set 1
N = 11
Mean of *X* = 9.0
Mean of *Y* = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

Summary Statistics of Data Set 2
N = 11
Mean of *X* = 9.0
Mean of *Y* = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

Summary Statistics of Data Set 3
N = 11
Mean of *X* = 9.0
Mean of *Y* = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

Summary Statistics of Data Set 4
N = 11
Mean of *X* = 9.0
Mean of *Y* = 7.5
Intercept = 3
Slope = 0.5
Correlation = 0.82

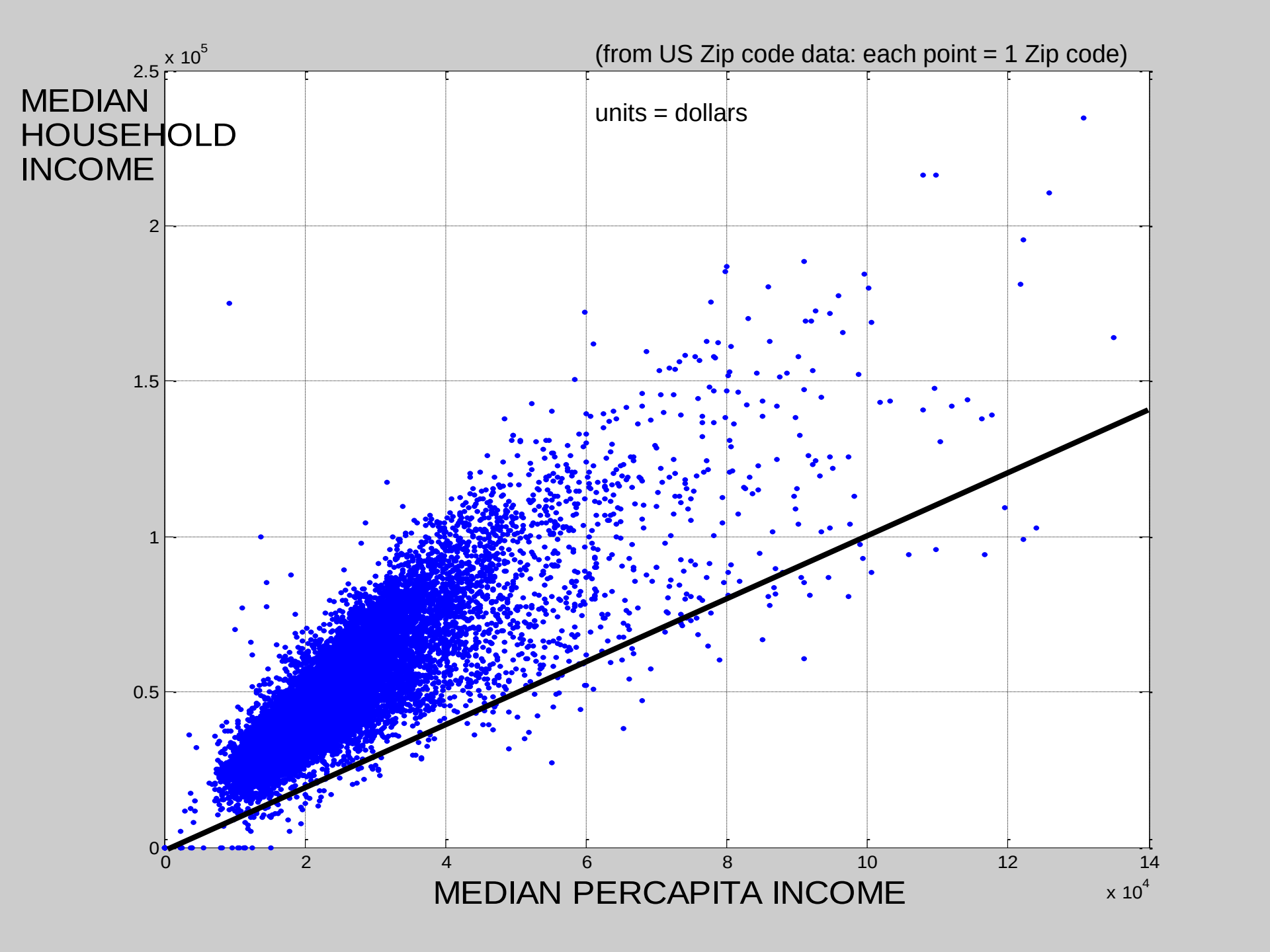
Anscombe, Francis (1973), *Graphs in Statistical Analysis*, The American Statistician, pp. 195-199.

Conclusions so far?

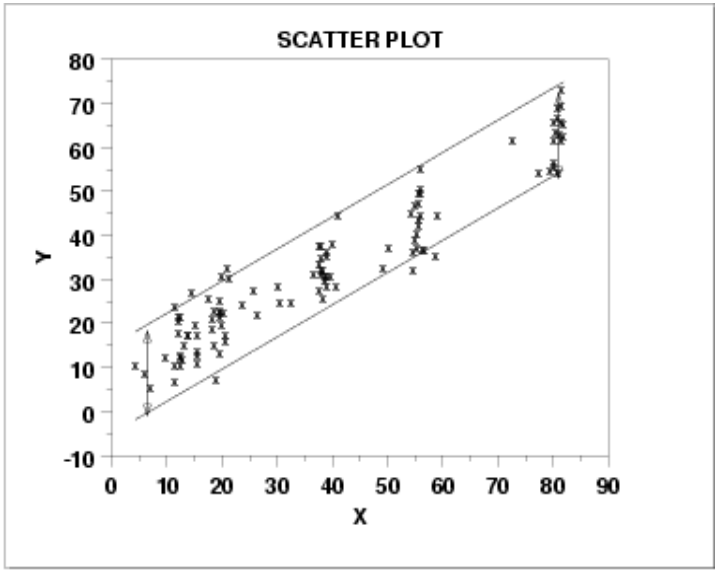
- Summary statistics are useful…..up to a point
- Linear correlation measures can be misleading
- There really is no substitute for plotting/visualizing the data

Scatter Plots

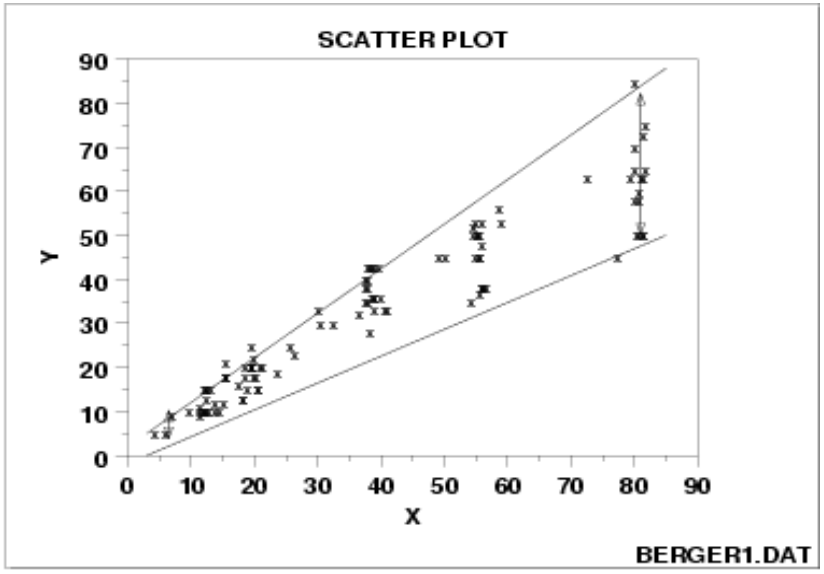
- Plot the value of one variable against the other
- Simple…but can be very informative, can reveal more than summary statistics
- For example, we can…
  - See if variables are dependent on each other (beyond linear dependence)
  - Detect if outliers are present
  - Can color-code to overlay group information (e.g., color points by class label for classification problems)


(from US Zip code data: each point = 1 Zip code)
MEDIAN HOUSEHOLD INCOME
units = dollars
x 10^5
2.5
2
1.5
1
0.5
0
0
2
4
6
8
10
12
14
MEDIAN PERCAPITA INCOME
x 10^4

Constant Variance versus Changing Variance



variation in Y does not depend on X

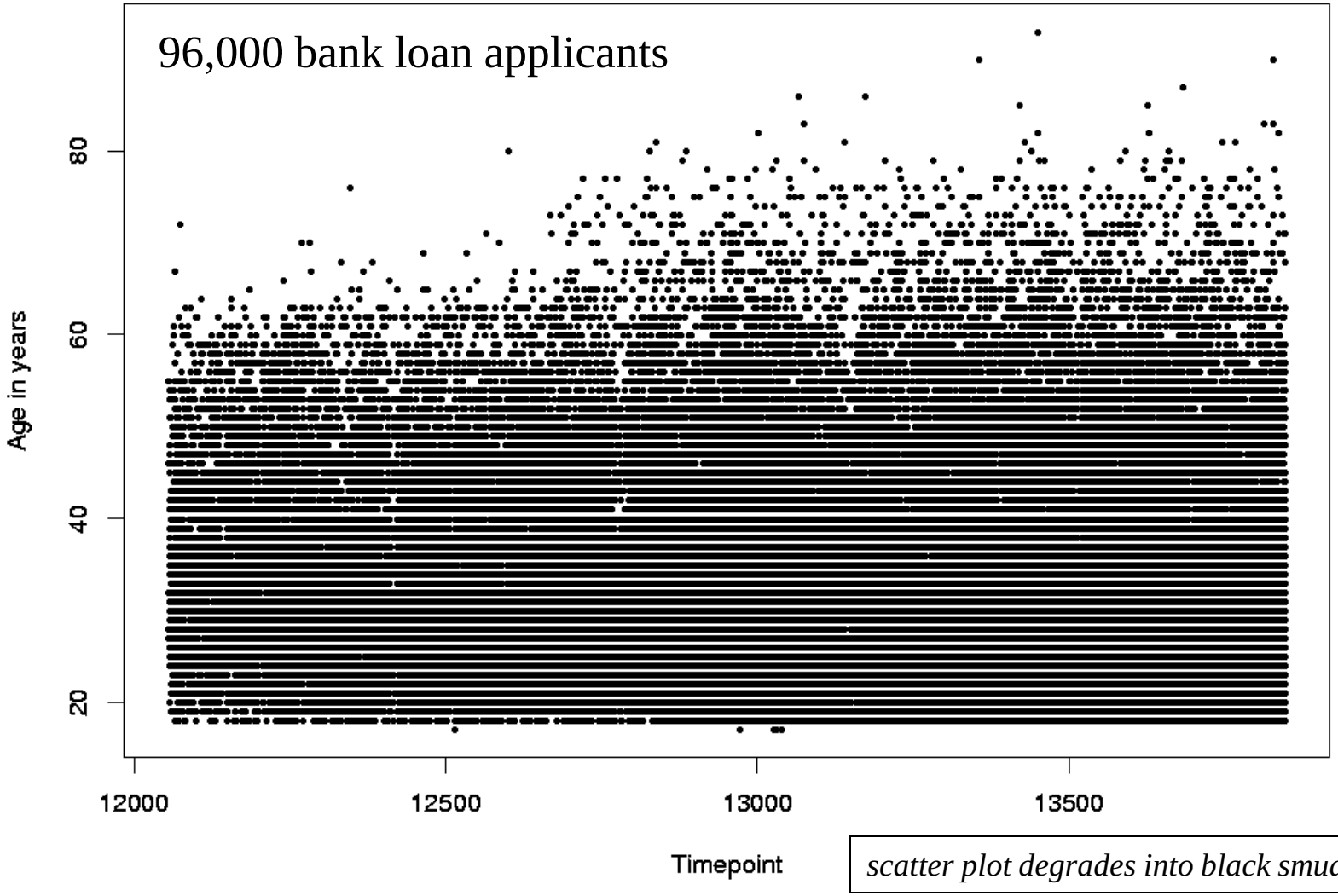


variation in *Y* changes with the value of *X*

e.g., Y = annual tax paid, X = income

Problems with Scatter Plots of Large Data

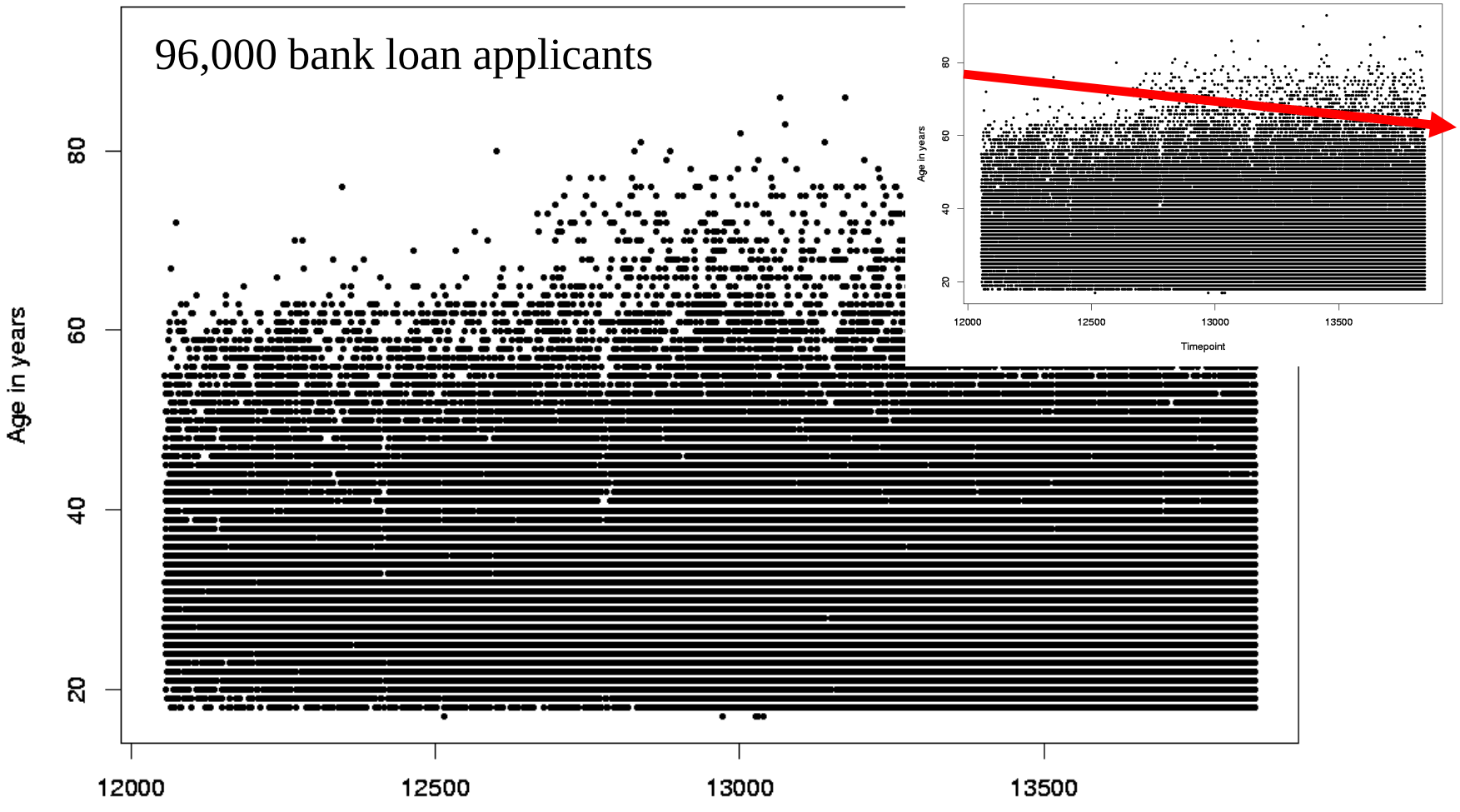
appears: later apps older; reality: downward slope (more apps, more variance)



scatter plot degrades into black smudge ...

Problems with Scatter Plots of Large Data

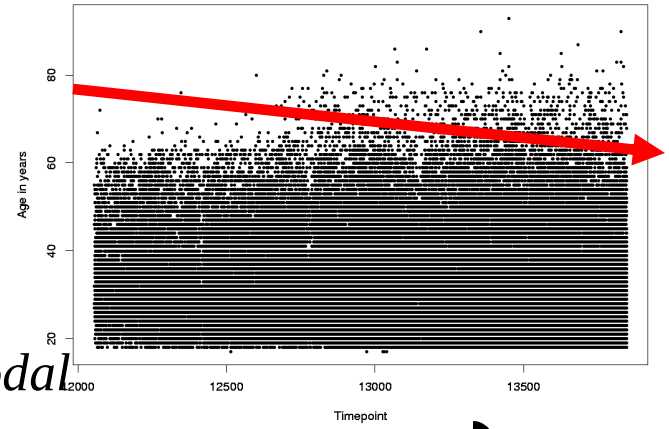
appears: later apps older; reality: downward slope (more apps, more variance)



scatter plot degrades into black smudge ...

Contour Plots (based on local density) can help

recall:



(same 96,000 bank loan apps as before)

shows variance(y) ↑ with x ↑ is indeed due to horizontal skew in density

unimodal

Age in years

Timepoint

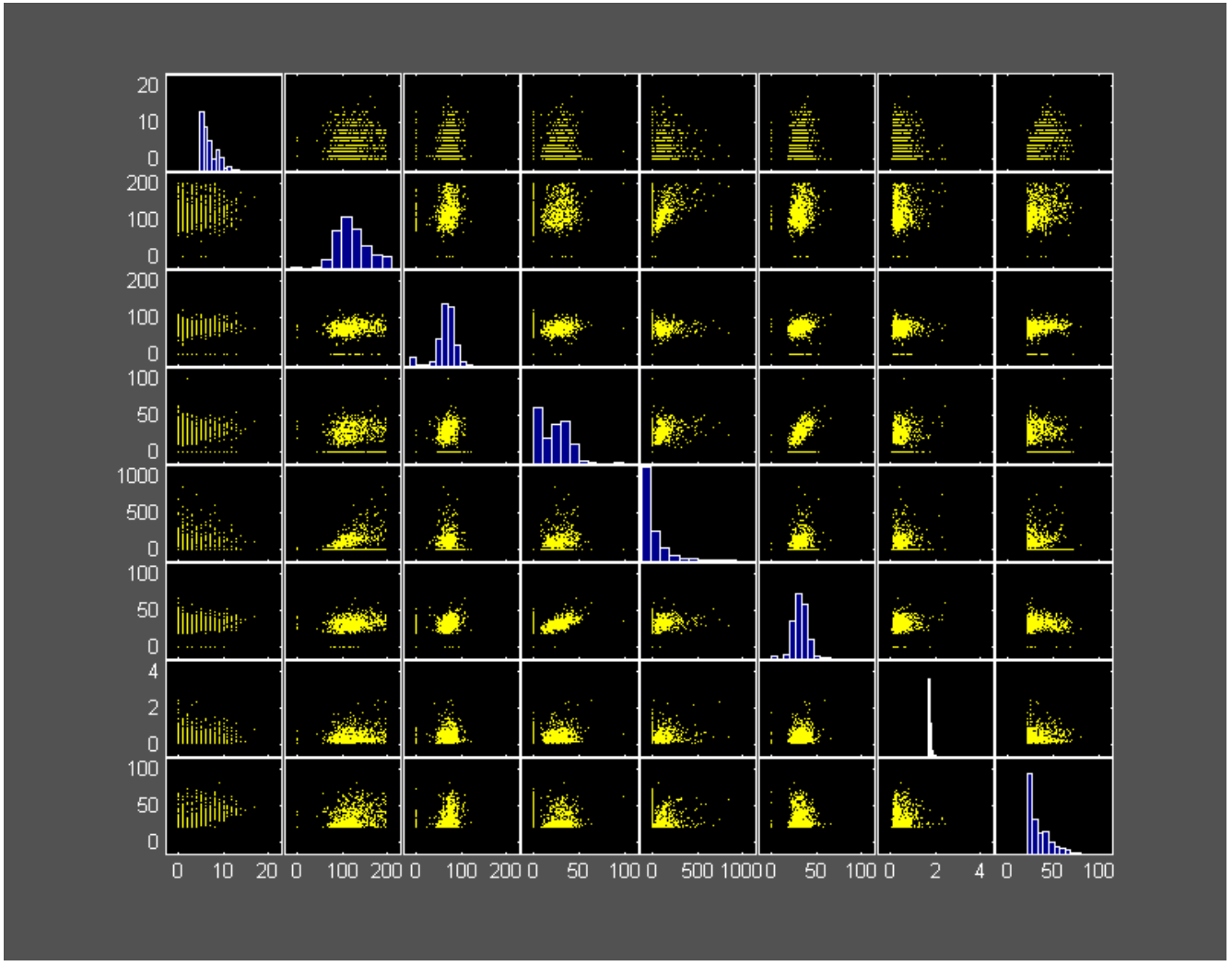
2000
3000
1000
5000
4000
1000

skewed ↓

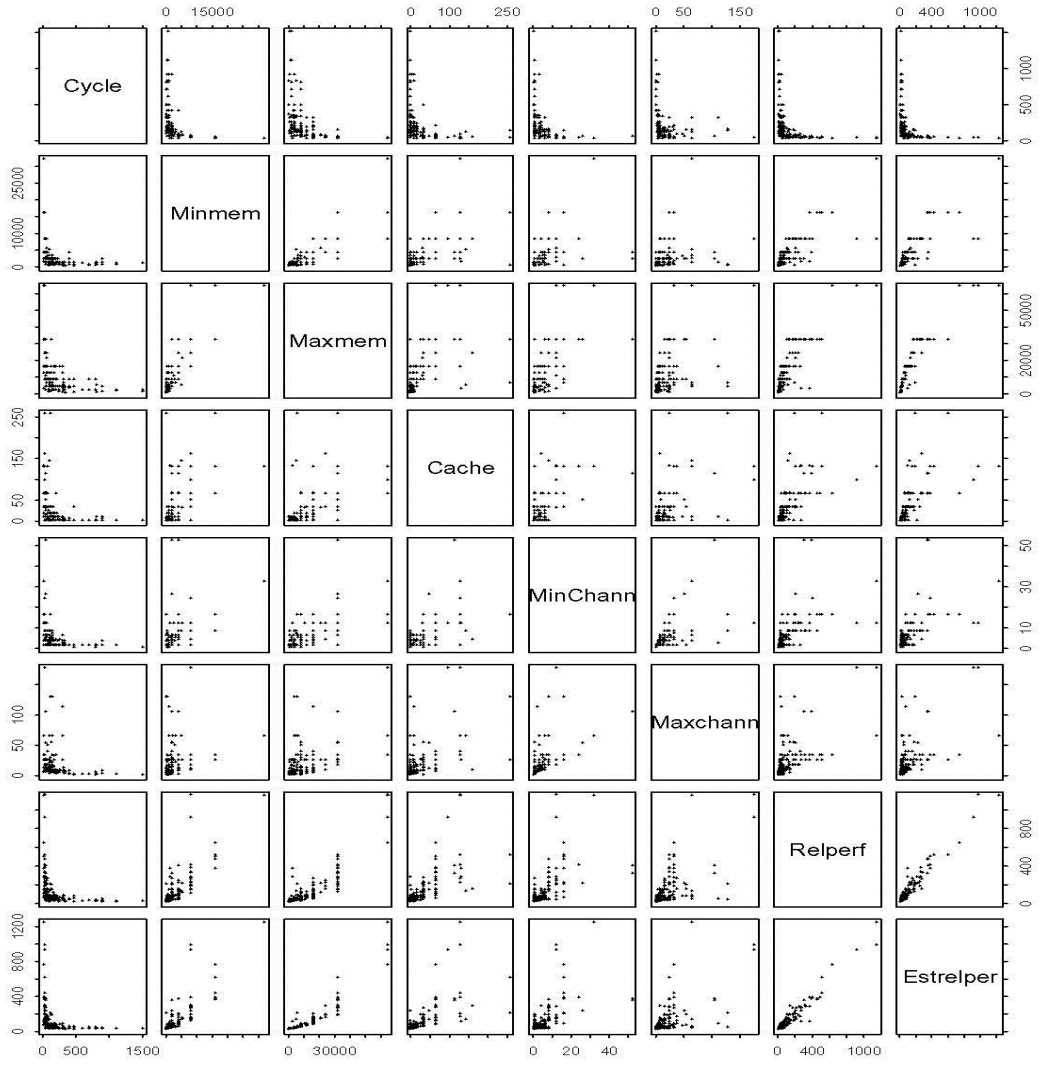
skewed →

Scatter-Plot Matrices

Pima Indians Diabetes data



Another Scatter-Plot Matrix



For interactive visualization the concept of "linked plots" is generally useful, i.e., clicking on 1 or more points in 1 window and having these same points highlighted in other windows

Using Color to Show Group Information in Scatter Plots

Iris classification data set, 3 classes

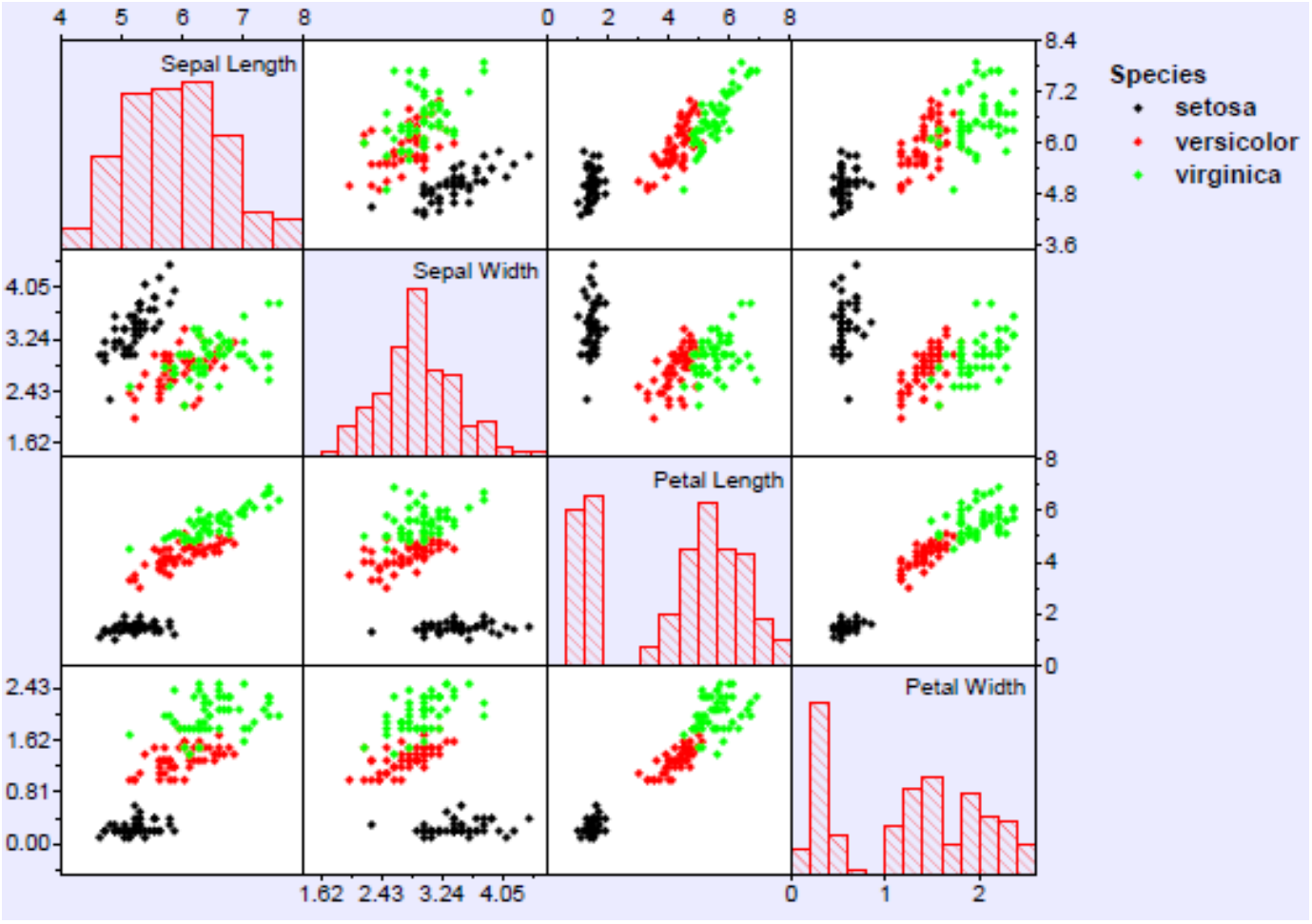


Figure from www.originlab.com

Another Example with Grouping by Color

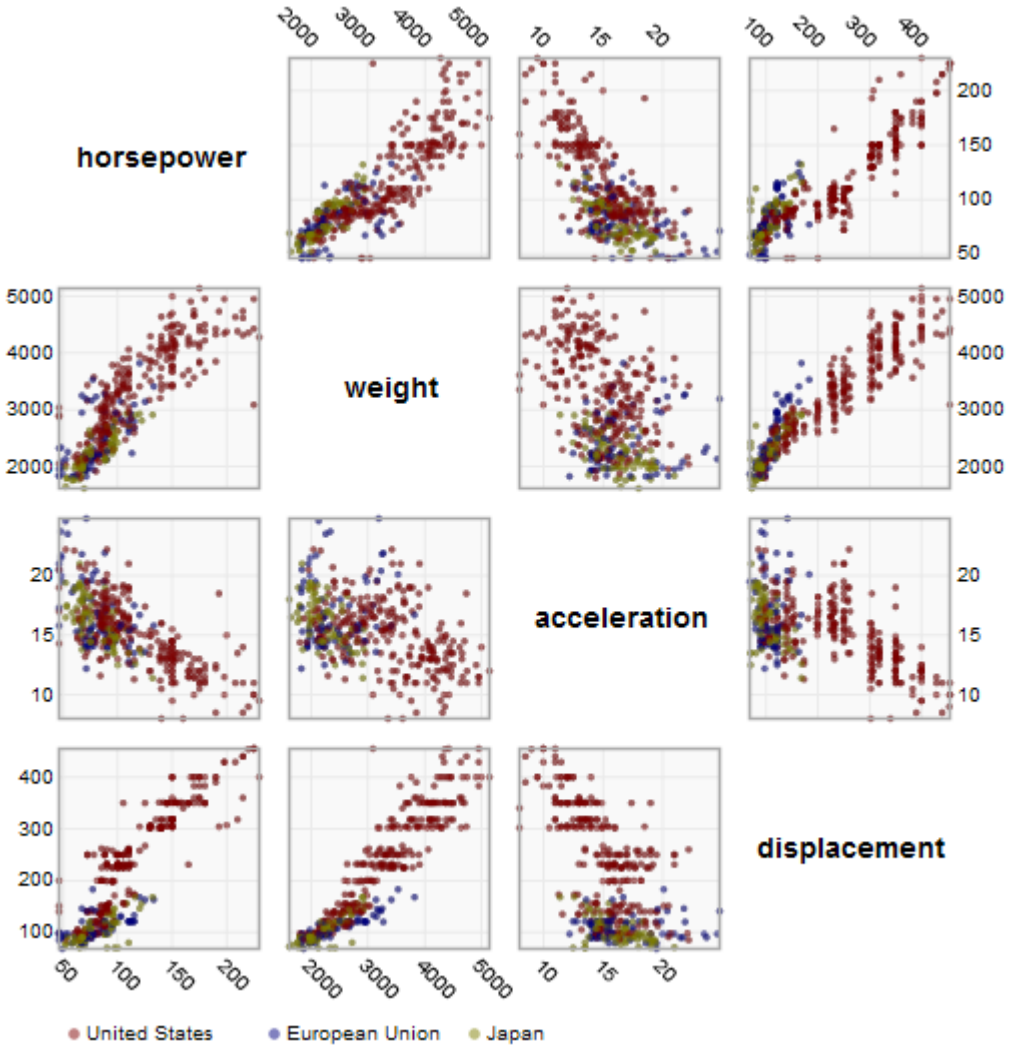
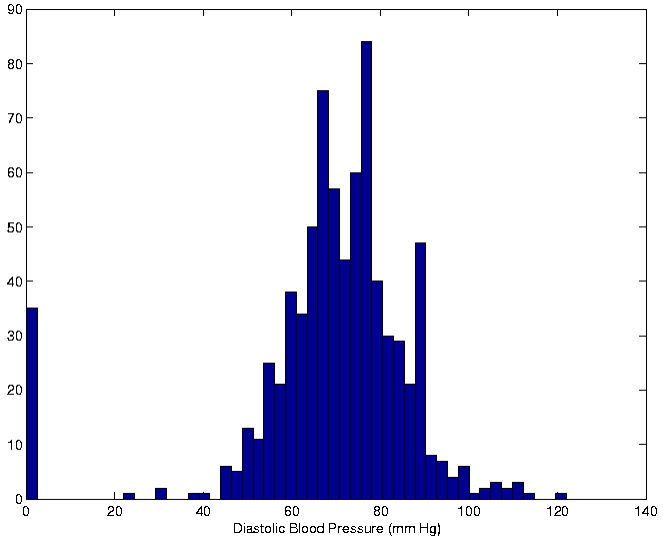


Figure from hci.stanford.edu

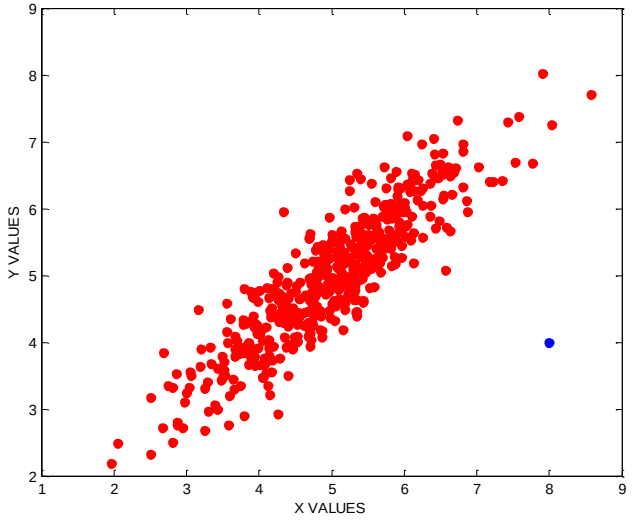
Outlier Detection

- Definition of an outlier?
  - No precise definition
  - Generally…."A data point that is significantly different to the rest of the data"
  - But how do we define "significantly different"? (many answers to this…..)
  - Typically assumed to mean that the point was measured in error, or is not a true measurement in some sense

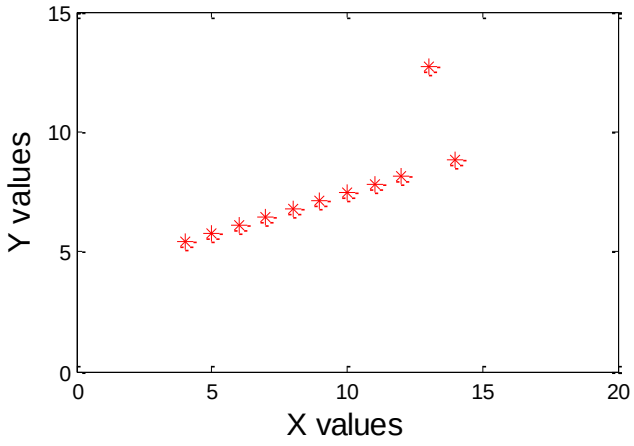
Outliers in 1 dimension



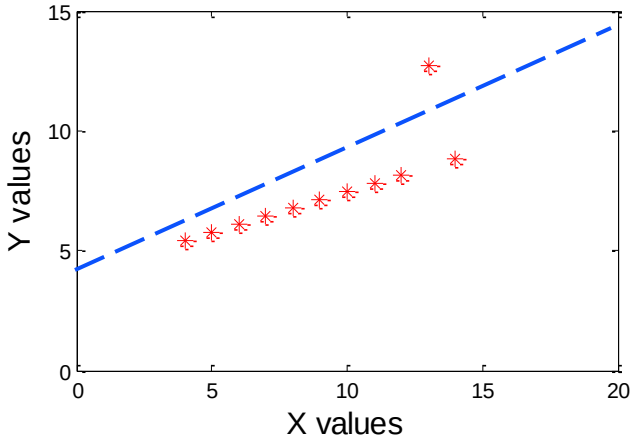
Outlier in 2 dimensions



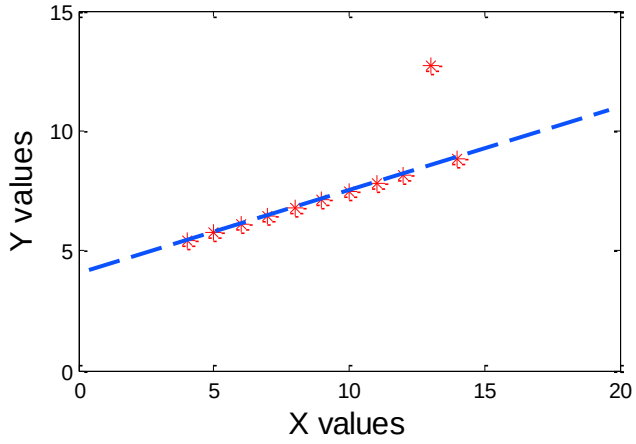
Example 1: The Effect of Outliers on Regression



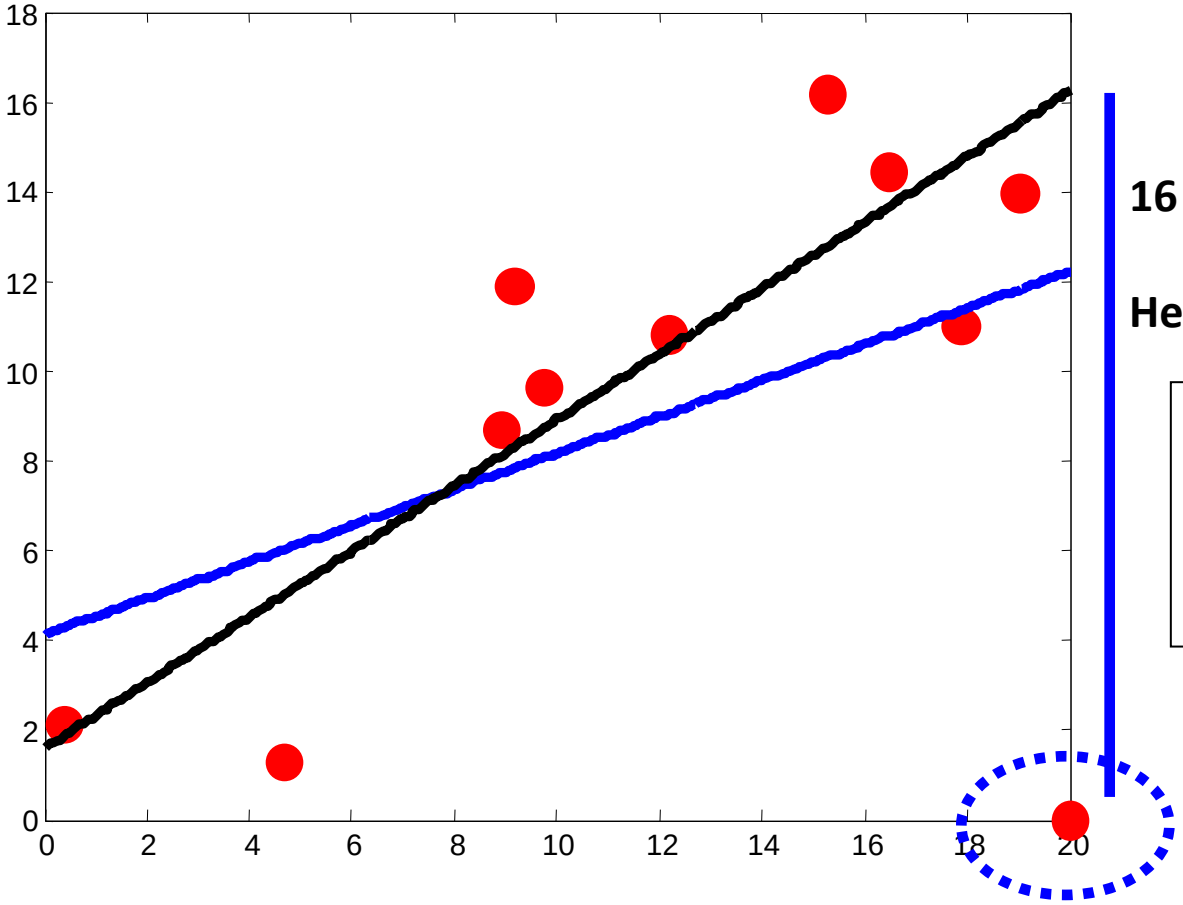
Least Squares Fit with the Outlier



Least Squares Fit without the Outlier

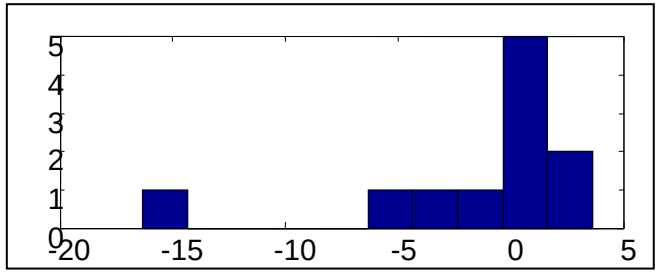


Example 2: Least Square Fitting is Sensitive to Outliers



16^2 cost for this one datum

Heavy penalty for large errors



Slide courtesy of Alex Ihler

More Robust Cost Functions for Training Regression Models

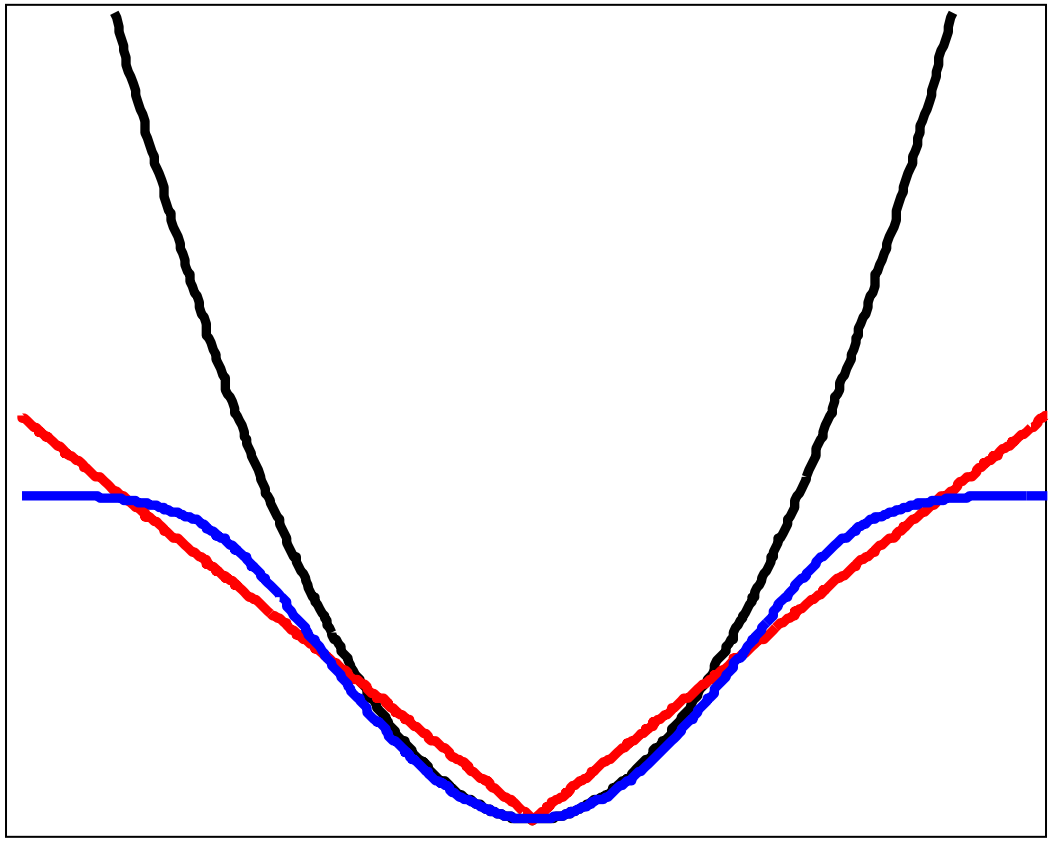
$$\ell_2 \ : \ (y - \hat{y})^2 \qquad \textbf{(MSE)}$$

$$\ell_1 \ : \ |y - \hat{y}| \qquad \textbf{(MAE)}$$

Something else entirely, e.g.,

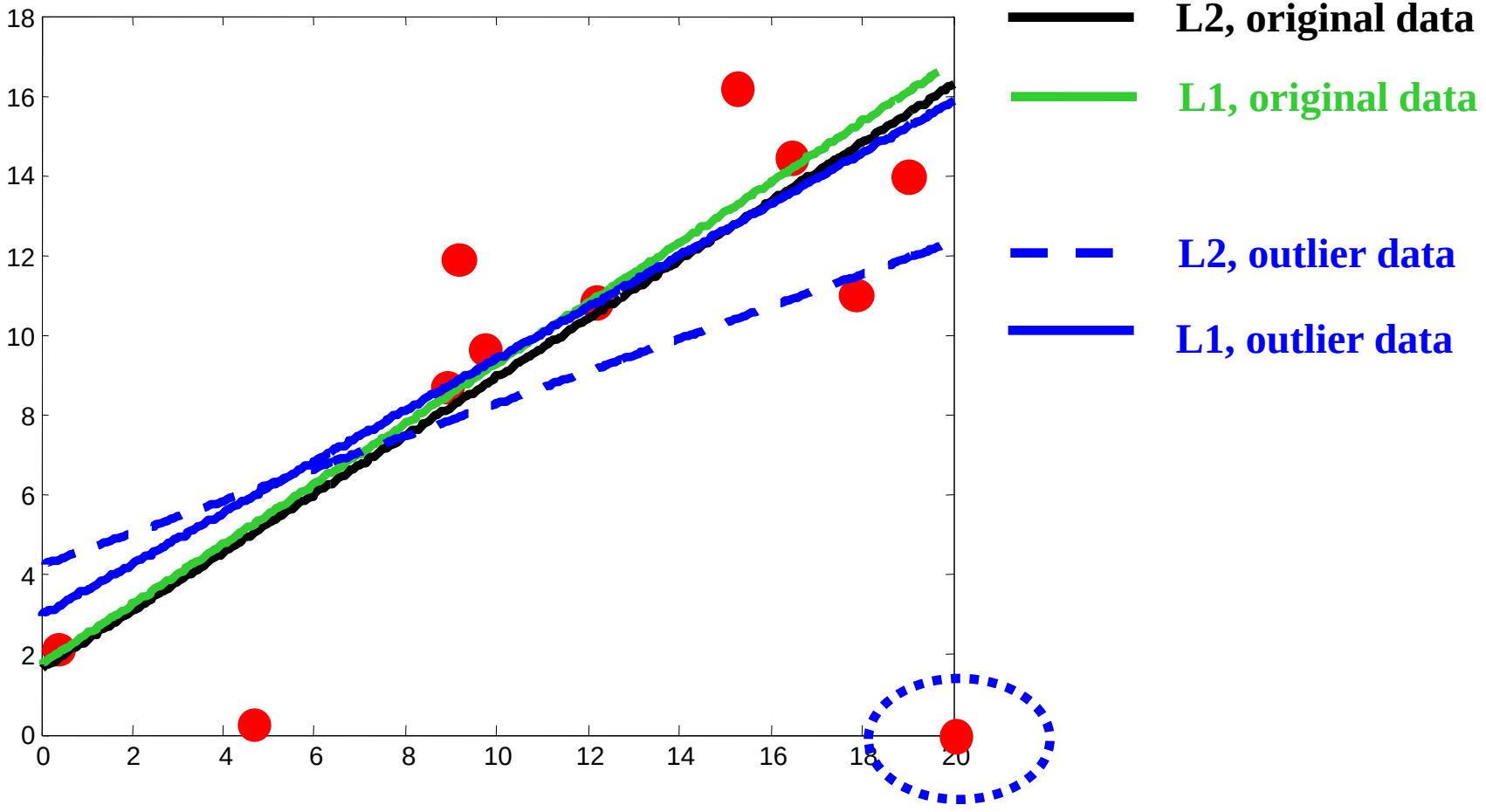
$$c - \log(\exp(-(y - \hat{y})^2) + c)$$

(Blue Line)



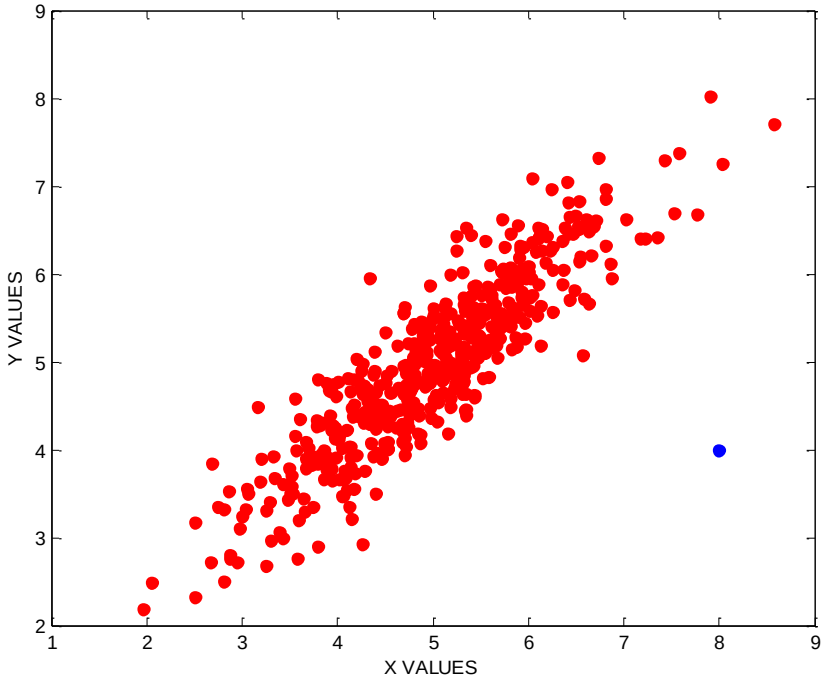
Slide courtesy of Alex Ihler

L1 is more Robust to Outliers than L2



Slide courtesy of Alex Ihler

Detection of Outliers in Multiple Dimensions



- Detecting "multi-dimensional outliers" is generally difficult
- In the example above, the blue point will not look like an outlier if we were to plot 1-dimensional histograms of Y or X – it only stands out in the 2d plot
- Now consider the same situation but in 3 or more dimensions

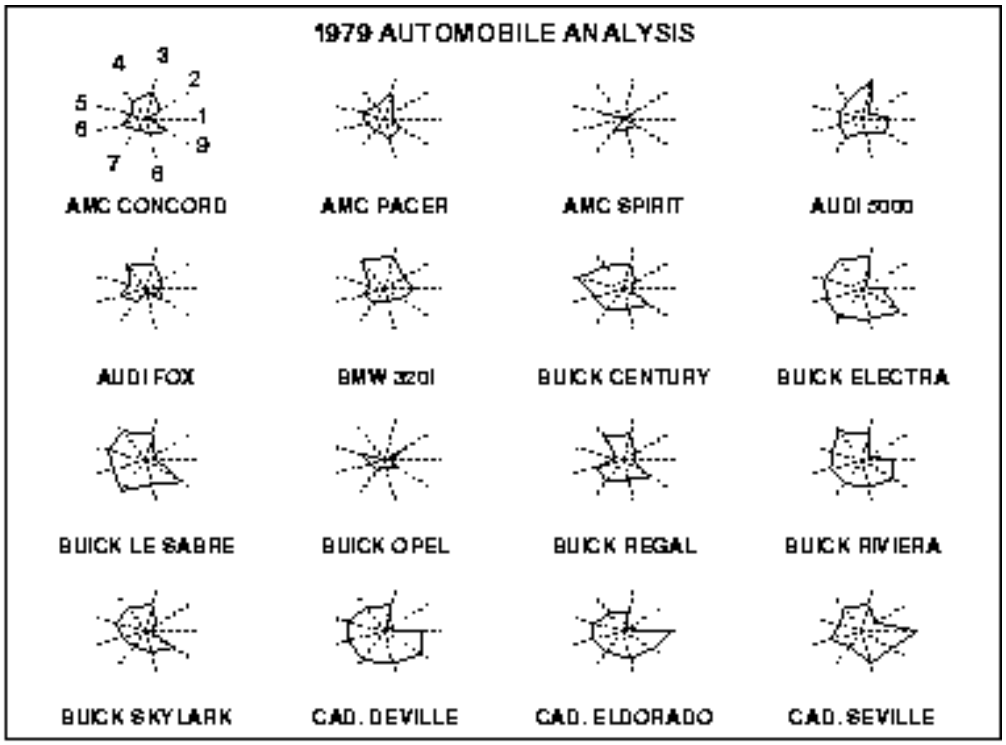
Some Advice on Outliers

- Use visualization (e.g., in 1d, 2d) to spot obvious outliers
- Use domain knowledge and known constraints
  - E.g., Age in years should be between 0 and 120
- Use the model itself to help detect outliers
  - E.g., in regression, data points with errors much larger than the others may be outliers
- Use robust techniques that are not overly sensitive to outliers
  - E.g., median is more robust than the mean, L1 is more robust than L2, etc
- Automated outlier detection algorithms? …not always useful
  - E.g., fit probability density model to N-1 points and determine how likely the Nth point is
  - May not work well in high dimensions and/or if there are multiple outliers
- In general: for large data sets outliers you can probably assume that outliers are present and proceed with caution….

Multivariate Visualization

- Multivariate -> multiple variables
- 2 variables: scatter plots, etc
- 3 variables:
  - 3-dimensional plots
  - Look impressive, but often not that useful
  - Can be cognitively challenging to interpret
  - Alternatives: overlay color-coding (e.g., categorical data) on 2d scatter plot
- 4 variables:
  - 3d with color or time
  - Can be effective in certain situations, but tricky
- Higher dimensions
  - Generally difficult
  - Scatter plots, icon plots, parallel coordinates: all have weaknesses
  - Alternative: “map” data to lower dimensions, e.g., PCA or multidimensional scaling
  - Main problem: high-dimensional structure may not be apparent in low-dimensional views

Using Icons to Encode Information, e.g., Star Plots



1 Price
2 Mileage (MPG)
3 1978 Repair Record (1 = Worst, 5 = Best)
4 1977 Repair Record (1 = Worst, 5 = Best)
5 Headroom
6 Rear Seat Room
7 Trunk Space
8 Weight
9 Length

Each star represents a single observation. Star plots are used to examine the relative values for a single data point

The star plot consists of a sequence of equi-angular spokes, called radii, with each spoke representing one of the variables.

Useful for small data sets with up to 10 or so variables

Limitations?

- Small data sets, small dimensions
- Ordering of variables may affect perception

Another Example of Icon Plots

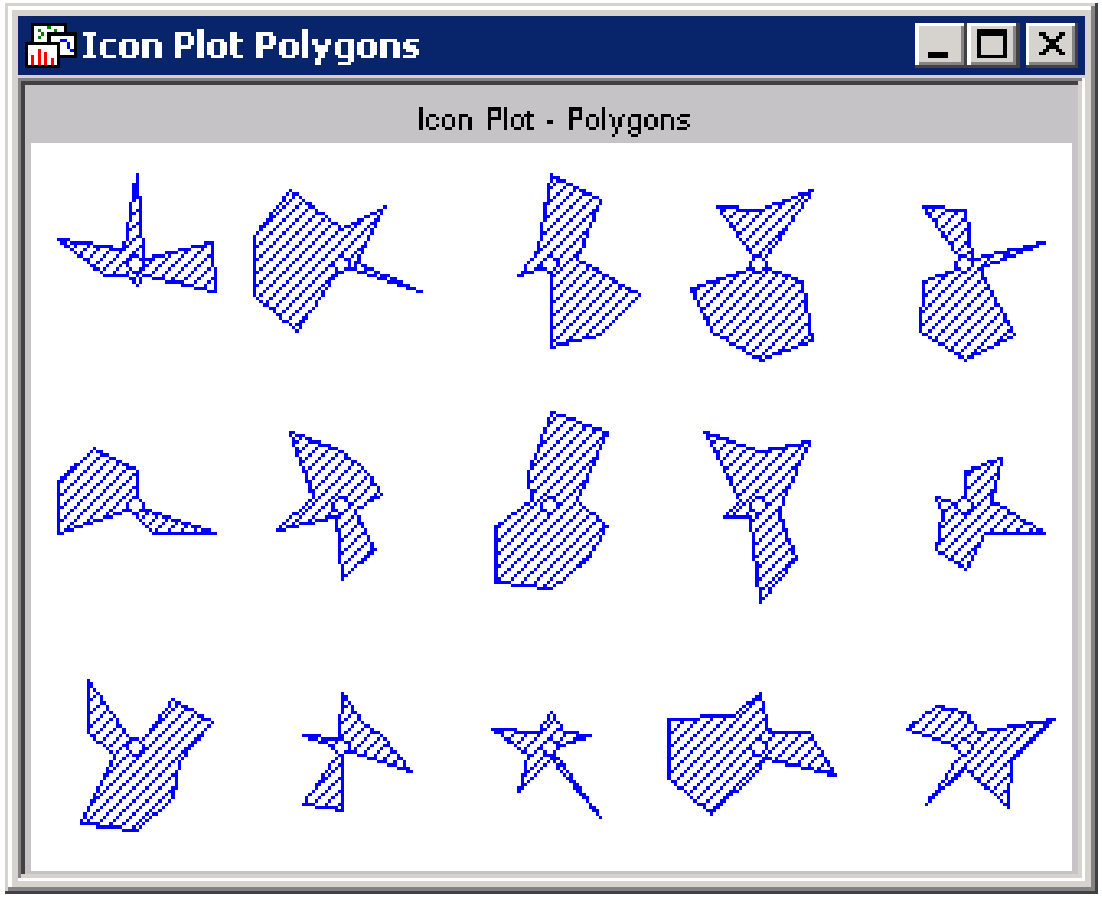


Figure from statsoft.com

Combining Scatter Plots and Icon Plots

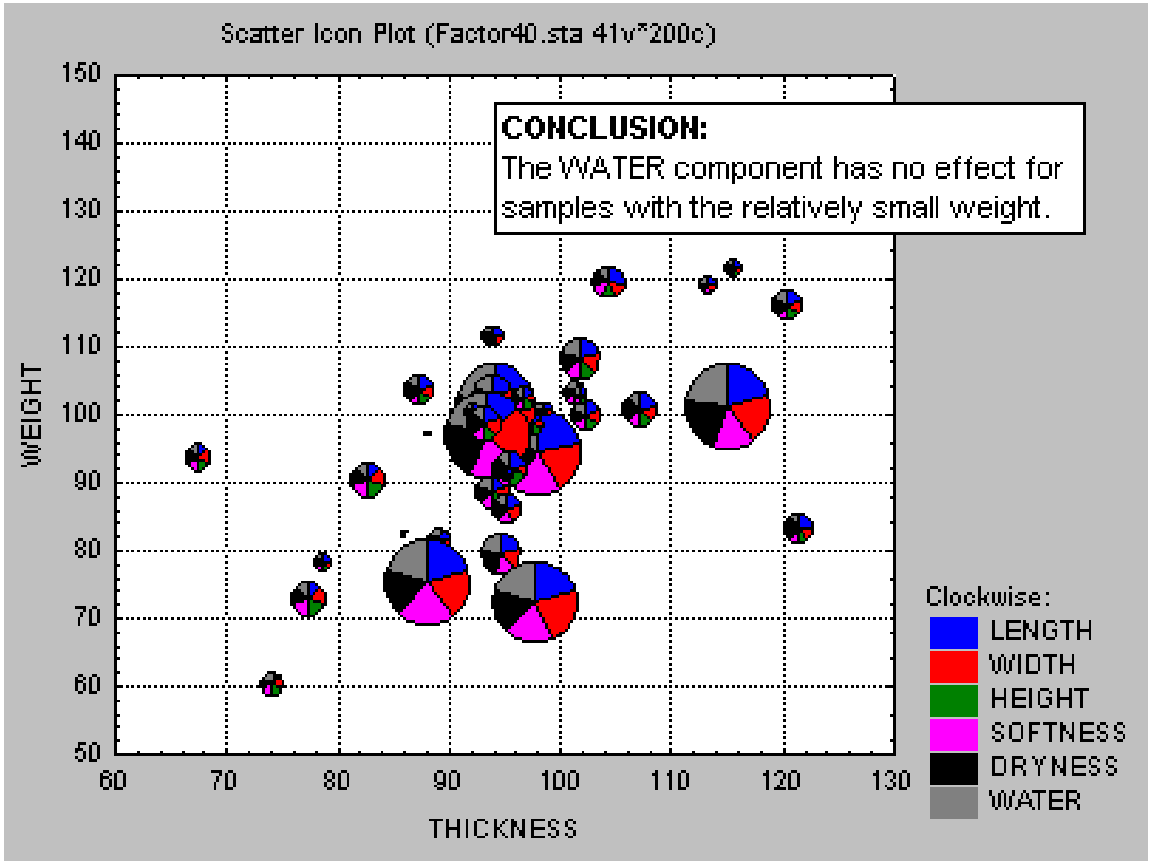
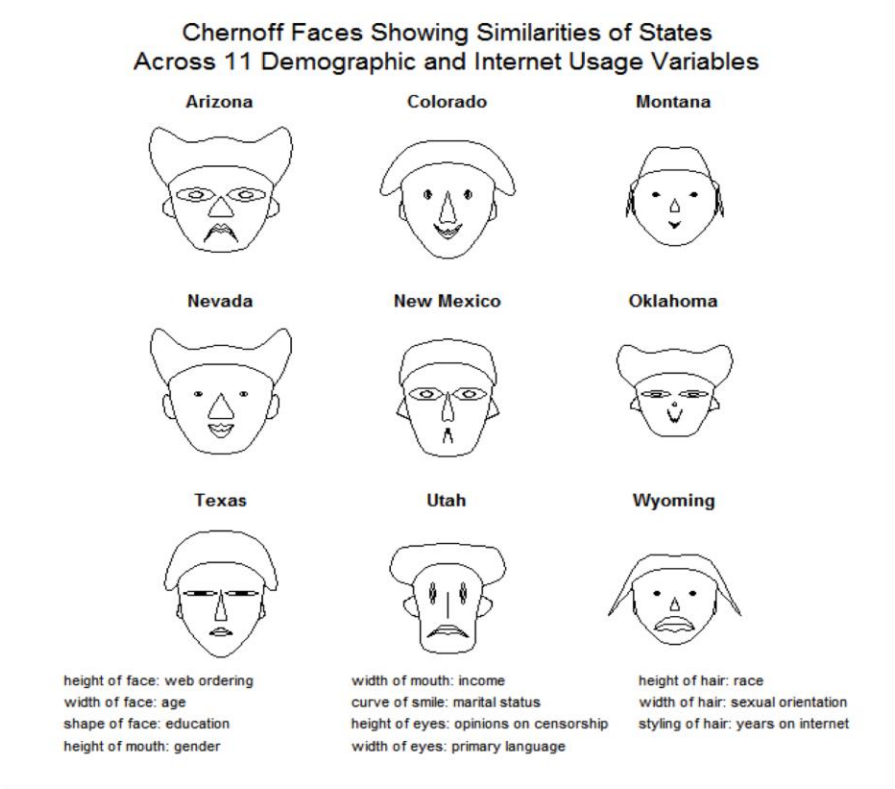


Figure from statsoft.com

Chernoff Faces

- Variable values associated with facial characteristic parameters, e.g., head eccentricity, eye eccentricity, pupil size, eyebrow slant, nose size, mouth shape, eye spacing, eye size, mouth length and degree of mouth opening

- Limitations?
  - Only up to 10 or so dimensions
  - Overemphasizes certain variables because of our perceptual biases



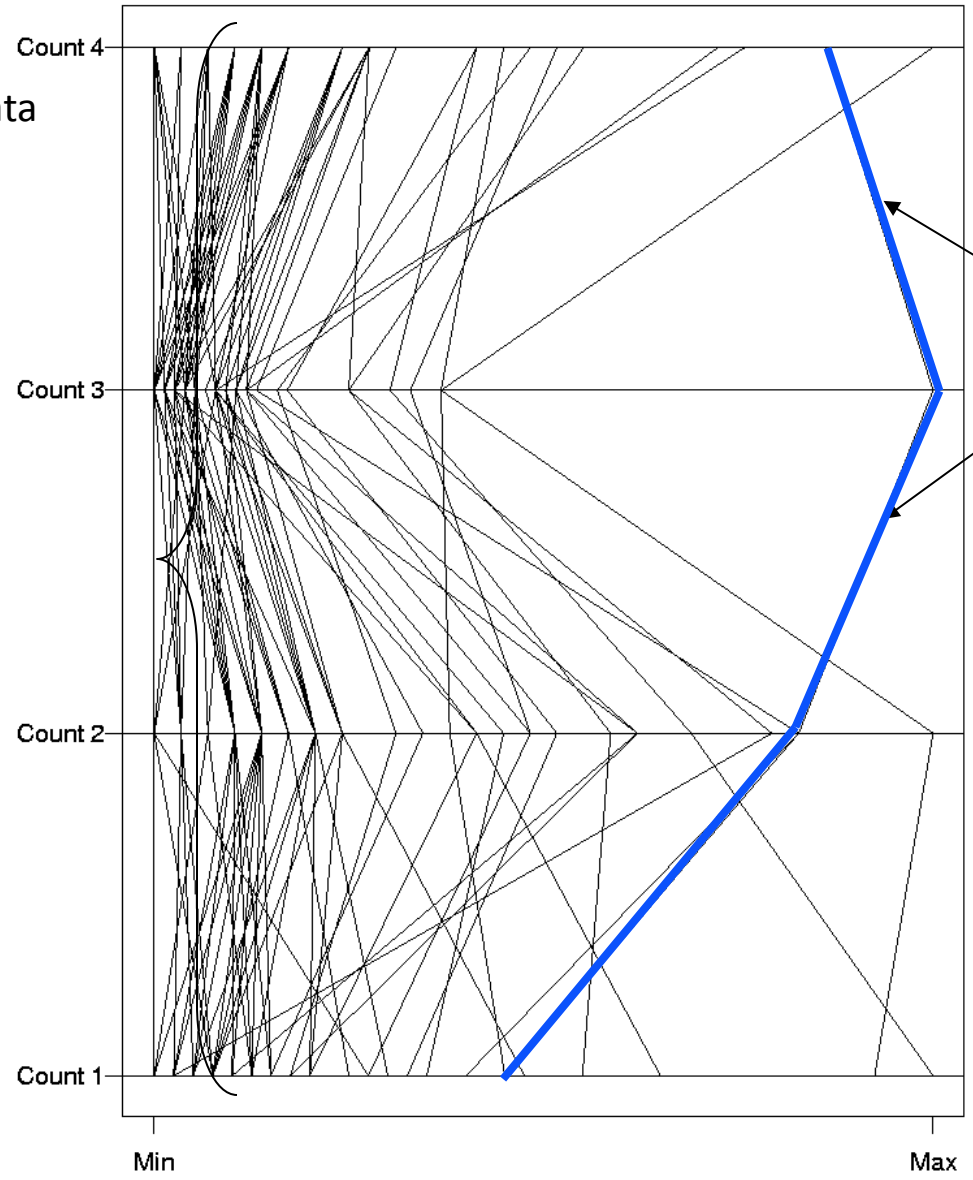
Chernoff Faces
2005 National League

alexreisner.com/baseball/stats/chernoff

	PCT	H	HR	BB	SB
ARI	0.475	1419	191	606	67
ATL	0.556	1453	184	534	92
CHI	0.488	1506	194	419	65
CIN	0.451	1453	222	611	72
COL	0.414	1477	150	509	65
FLO	0.512	1499	128	512	96
HOU	0.549	1400	161	481	115
LAD	0.438	1374	149	541	58
MIL	0.500	1413	175	531	79
NYM	0.512	1421	175	486	153
PHI	0.543	1494	167	639	116
PIT	0.414	1445	139	471	73
SDP	0.506	1416	130	600	99
SFG	0.463	1427	128	431	71
STL	0.617	1494	170	534	83
WAS	0.500	1367	117	491	45

Parallel Coordinates Method

Epileptic Seizure Data

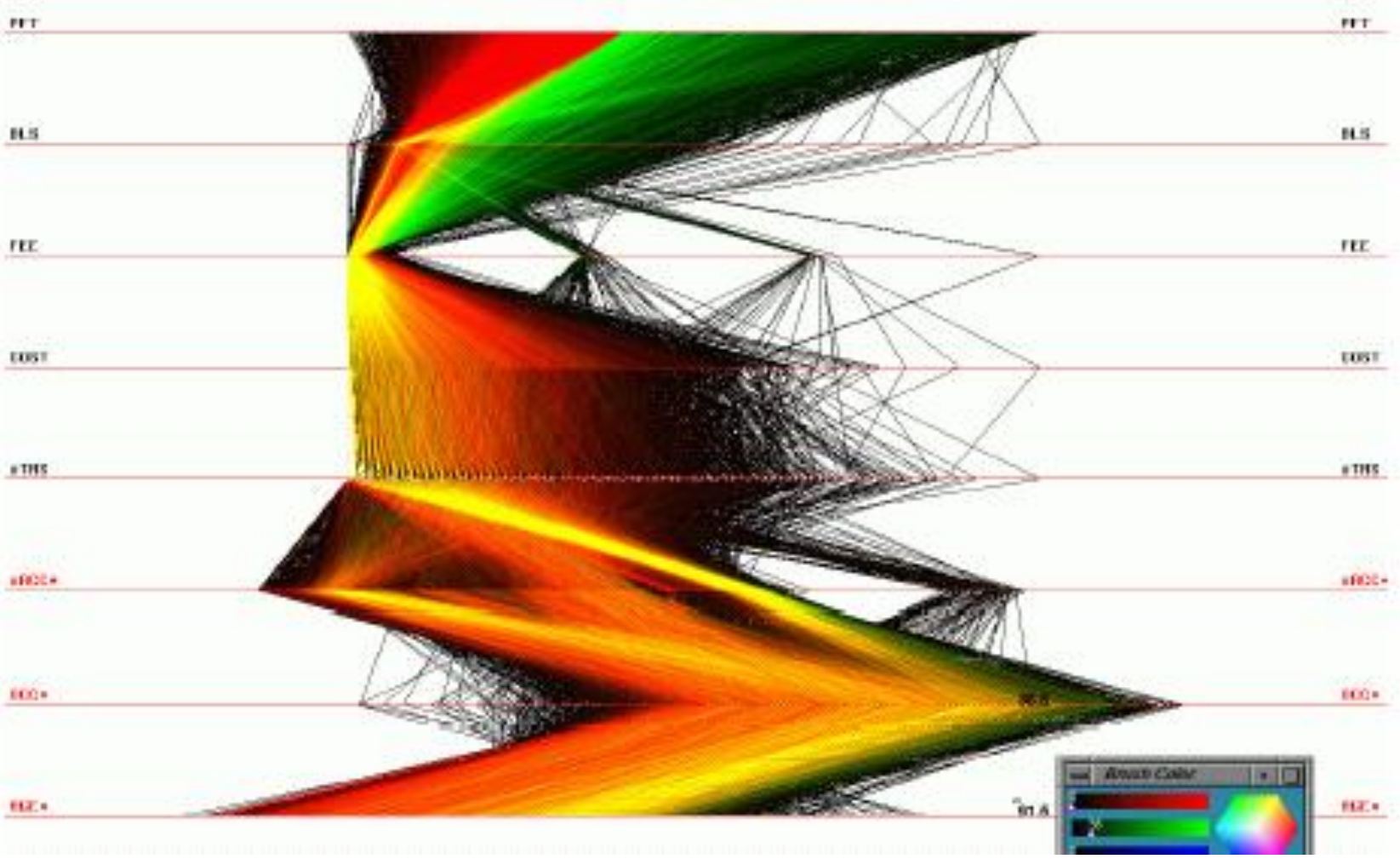


Interactive "brushing" is useful for seeing distinctions

1 (of n) cases

(this case is a "brushed" one, with a darker line, to standout from the n-1 other cases)

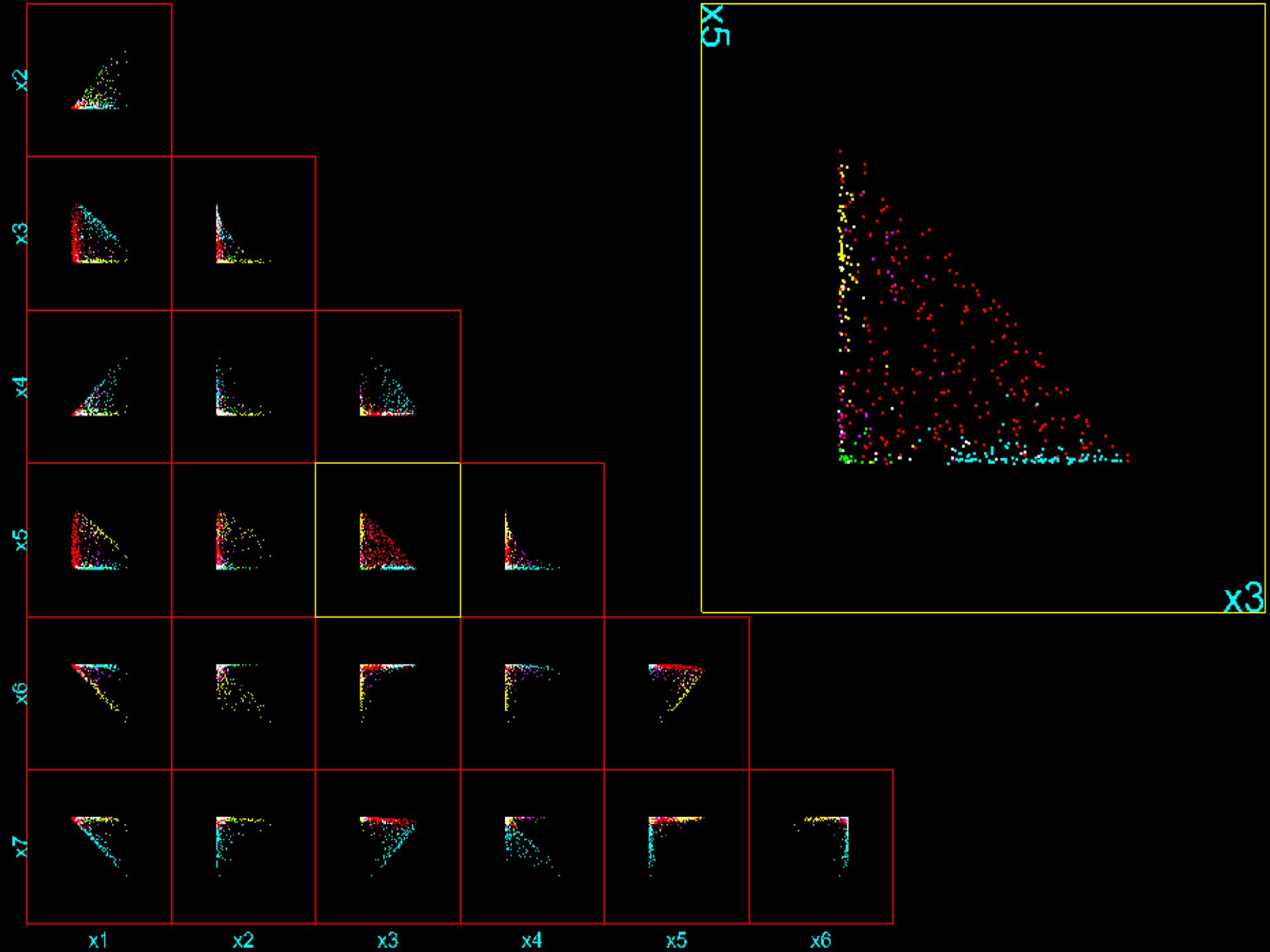
More elaborate parallel coordinates example (from E. Wegman, 1999).
12,000 bank customers with 8 variables
Additional "dependent" variable is profit (green for positive, red for negative)

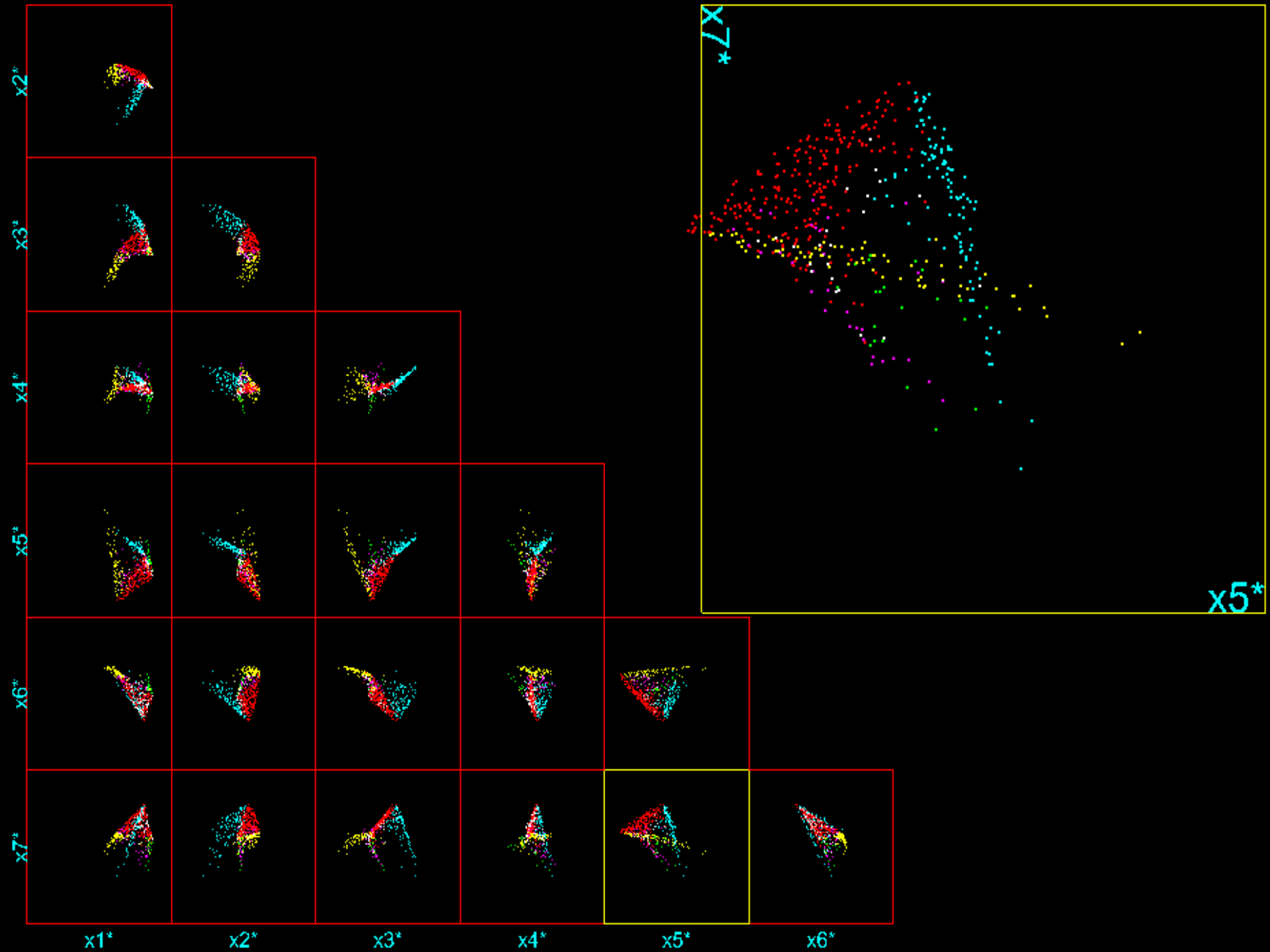


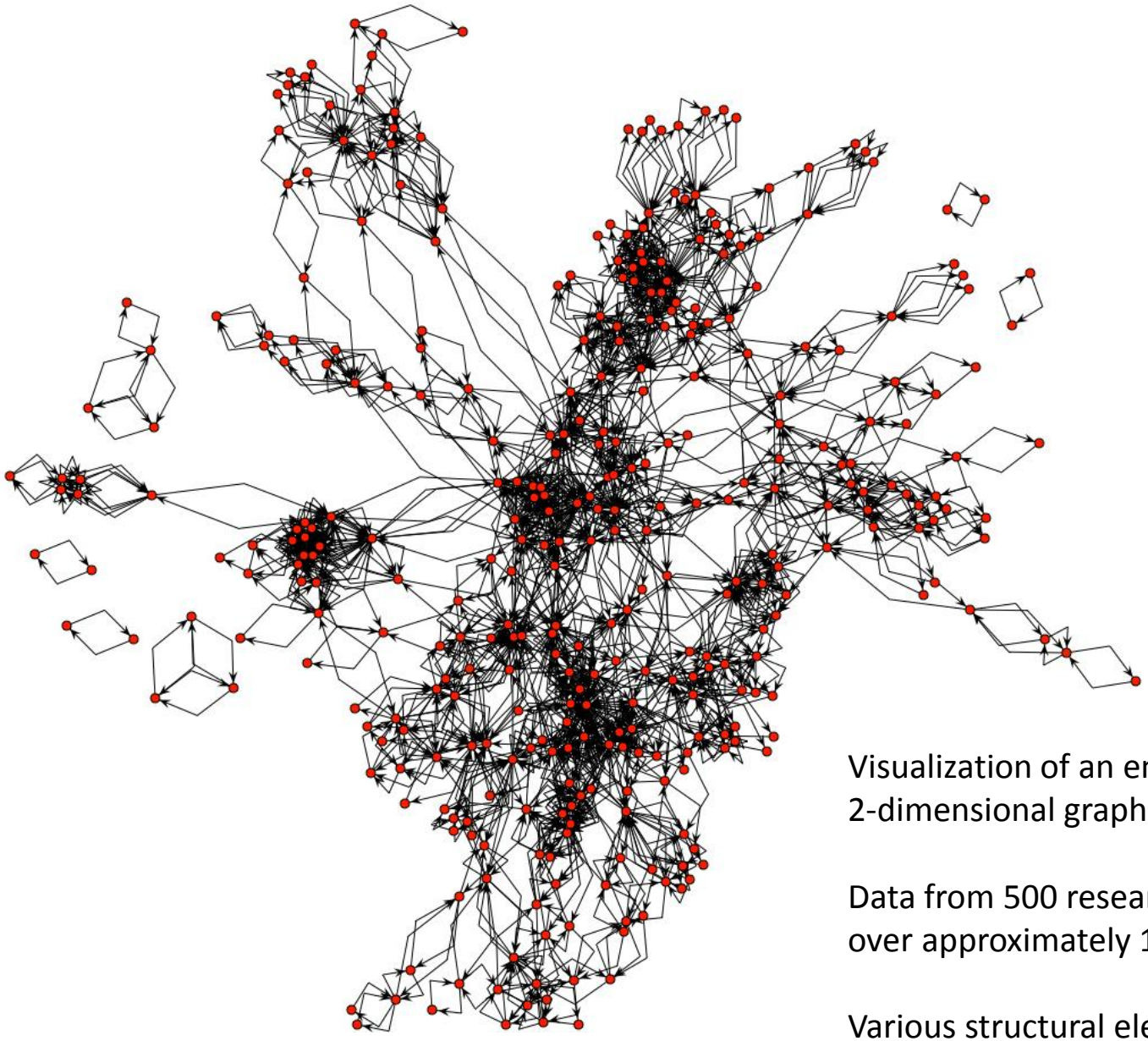
Interactive "Grand Tour" Techniques

- "Grand Tour" idea
  - Cycle continuously through multiple projections of the data
  - Cycles through all possible projections (depending on time constraints)
  - Projects can be 1, 2, or 3d typically (often 2d)
  - Can link with scatter plot matrices (see following example)
  - Asimov (1985)

- Example on following 2 slides
  - 7 dimensional physics data, color-coded by group, shown with
    (a) Standard scatter matrix
    (b) 2 static snapshots of grand tour


x2
x3
x4
x5
x6
x7
x1
x2
x3
x4
x5
x6
x5
x3


x2*
x3*
x4*
x5*
x6*
x7*
x1*
x2*
x3*
x4*
x5*
x6*
x7*
x5*



Visualization of an email network using 2-dimensional graph drawing or “embedding”

Data from 500 researchers at Hewlett-Packard over approximately 1 year.

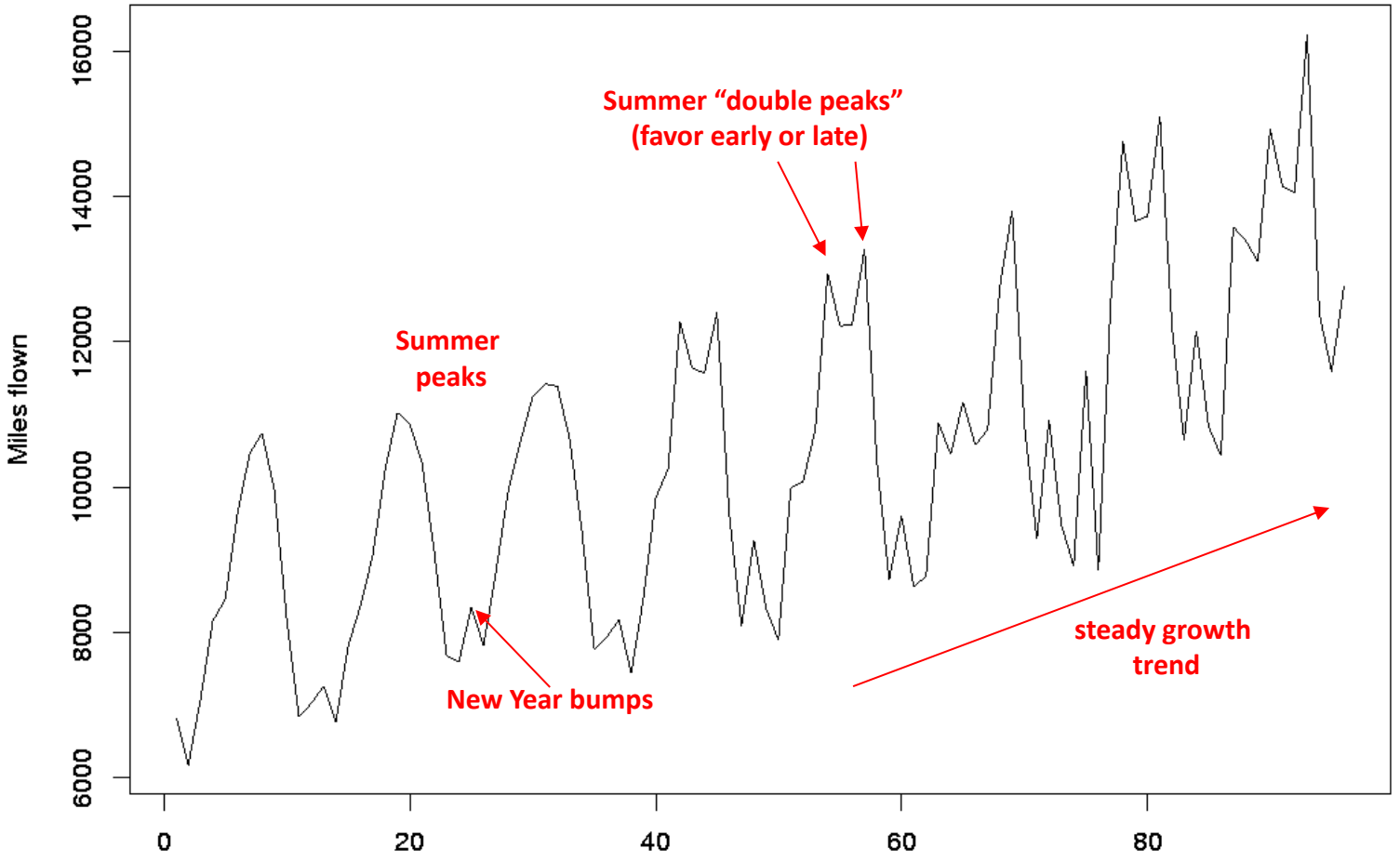
Various structural elements of the network are apparent

Exploratory Data Analysis

Visualizing Time-Series Data

Time-Series Data: Example 1

Historical data on millions of miles flown by UK airline passengers
…..note a number of different systematic effects

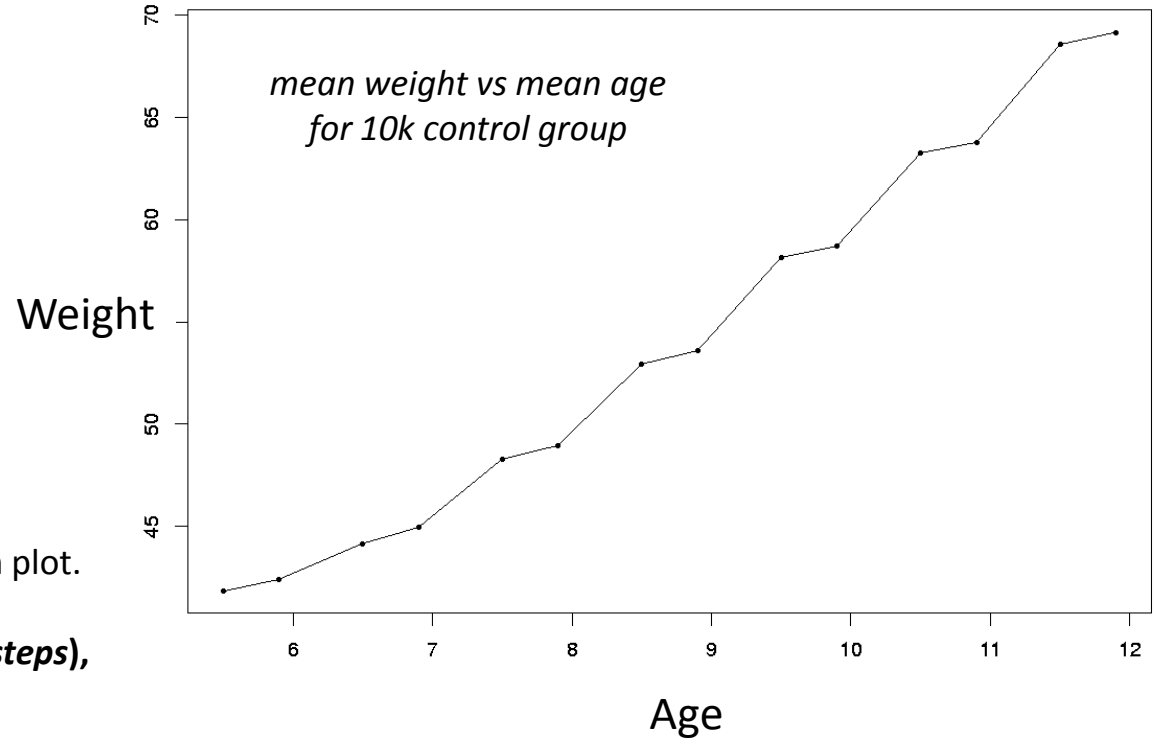


Time-Series Data: Example 2

Data from study on weight measurements over time of children in Scotland

Experimental Study:
More milk -> better health?

20,000 children:
5k raw, 5k pasteurize,
10k control (no supplement)



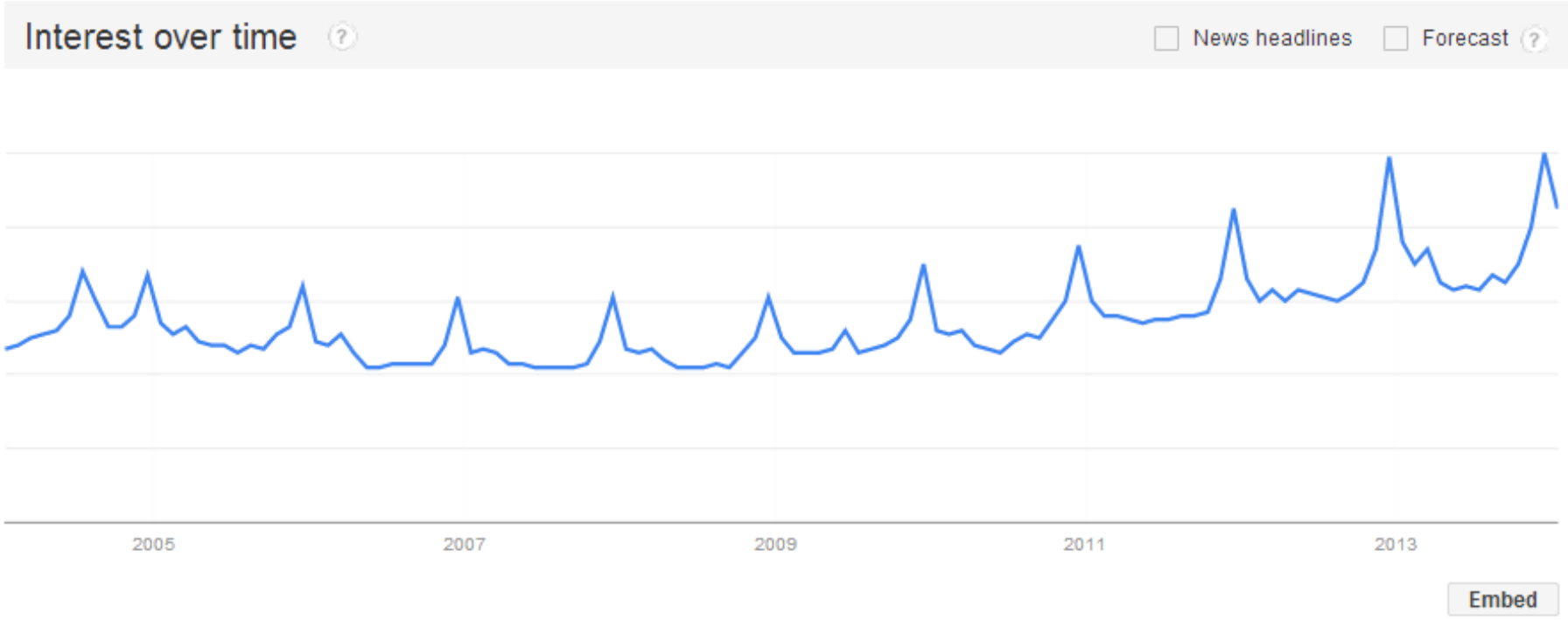
Would expect smooth weight growth plot.

Plot shows an unexpected pattern (*steps*), not apparent from raw data table.

Why do the children appear to grow in spurts?

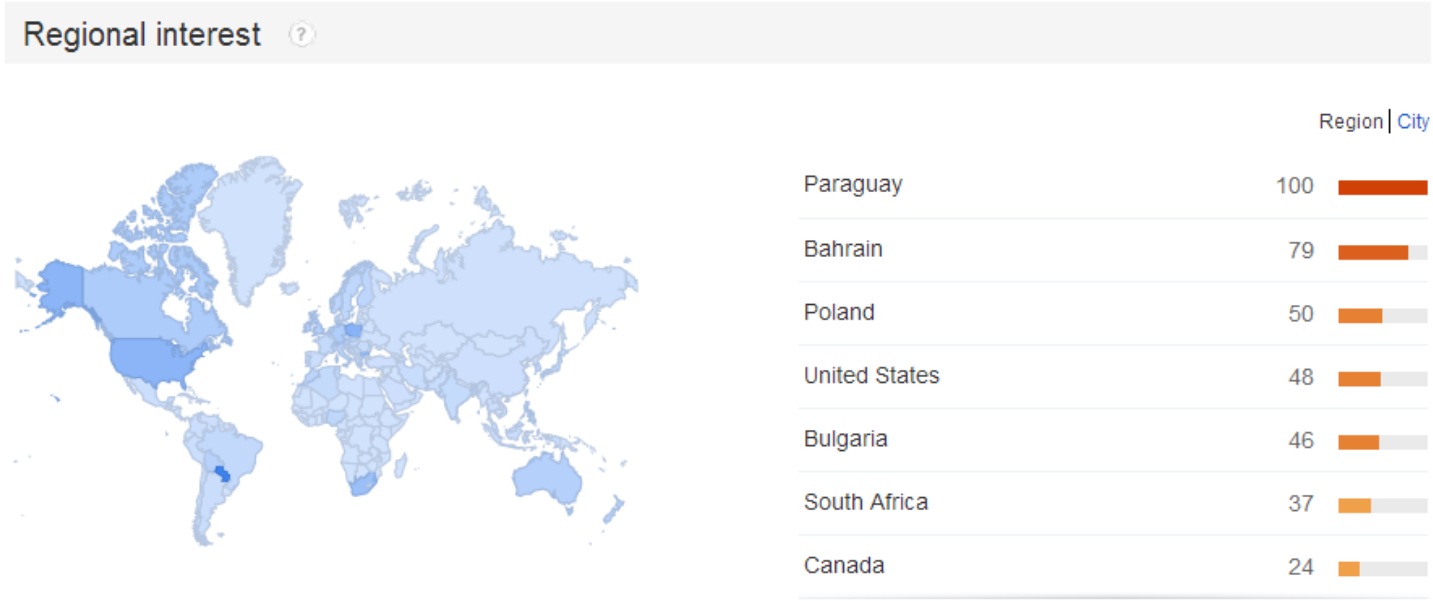
Time-Series Data: Example 3 (Google Trends)

Search Query = whiskey



Time-Series Data: Example 4 (Google Trends)

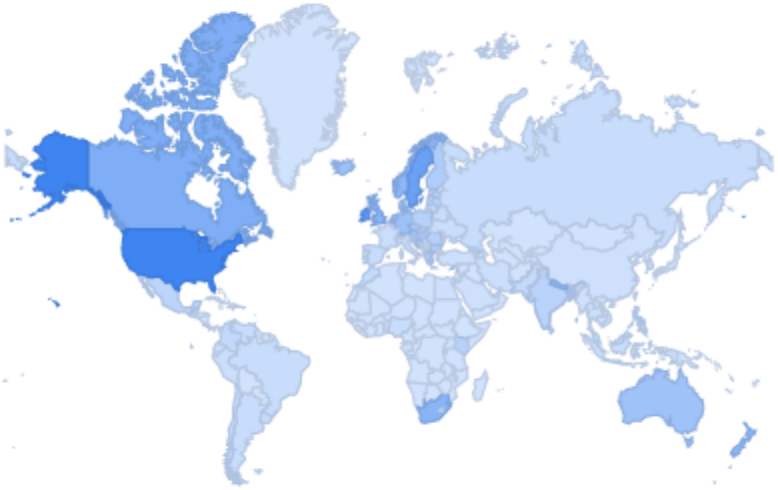
Search Query = NSA



Spatial Distribution of the Same Data (Google Trends)

Search Query = whiskey

Regional interest



Region | City

United States	100
Ireland	79
Sweden	70
Canada	56
South Africa	49
United Kingdom	49
Norway	46

View change over time

Embed

Embed

Summary on Exploration/Visualization

- Always useful and worthwhile to visualize data
  - human visual system is excellent at pattern recognition
  - gives us a general idea of how data is distributed, e.g., extreme skew
  - detect “obvious outliers” and errors in the data
  - gain a general understanding of low-dimensional properties
- Many different visualization techniques
- Limitations
  - generally only useful up to 3 or 4 dimensions
  - massive data: only so many pixels on a screen - but subsampling is useful