Data Mining and Visualization

Jeremy Walton

NAG Ltd, Oxford

Overview

- Data mining components
  - Functionality
  - Example application
    - Quality control
- Visualization
  - Use of 3D
  - Example application
    - Market research
- Statistics and visualization in Excel
  - What's the problem?

Overview

- Data mining components
  - Functionality
  - Example application
    - Quality control
- Visualization
  - Use of 3D
  - Example application
    - Market research
- Statistics and visualization in Excel
  - What’s the problem?

NAG Data Mining Components (DMC)

Data Cleaning

- *Data imputation* adding missing values
- *Outlier detection* finding suspect data records

Data Transformation

- *Scaling Data* before distance computation
- *Principal Component Analysis* reducing # of variables

Cluster Analysis

- *k-means* analyst decides # of clusters in data
- *Hierarchical* stepwise agglomeration of data

DMC: Classification techniques

- *Classification Trees* Two types of decision tree:
  - binary (Gini index)
  - n-ary (entropy-based)
- *Generalized Linear Models* Fitting of
  - Binomial distribution (for binary classification tasks)
  - Poisson distribution (for count data)
- *k-Nearest Neighbours*
  - Predict values using k most similar records in a training dataset
  - Set prior probabilities for data classes
  - Also used for regression (see below)

DMC: Regression techniques

- Regression Trees
  - Minimise sum of squares about mean
    - robust estimate of the mean, or sample average
- Linear Regression
  - Automatic selection of model variables
- Multi-Layer Perceptron Neural Networks
  - Flexible non-linear models
    - Free parameters in MLP optimised using conjugate gradients
- Nearest Neighbours (see above)
- Radial Basis Function Models
  - function of distance from centre location to data records

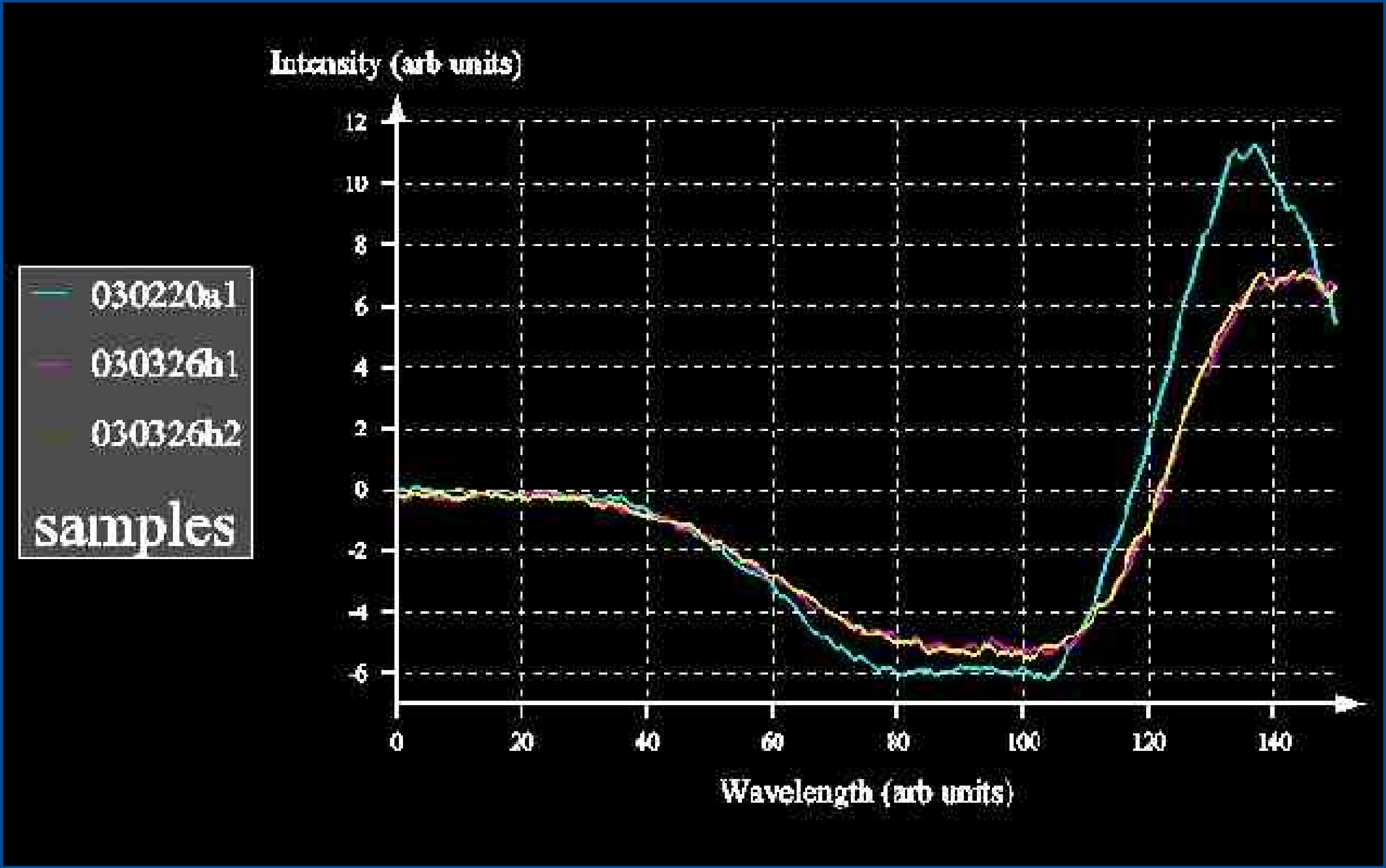
DMC: other techniques

- Association rules
  - Determine relationships between nominal data values
- Utility functions
  - Random number generators
  - Rank ordering
  - Sorting
  - Mean and sum of squares updates
  - Two-way classification comparison
  - Save and load models

Example application: Quality control

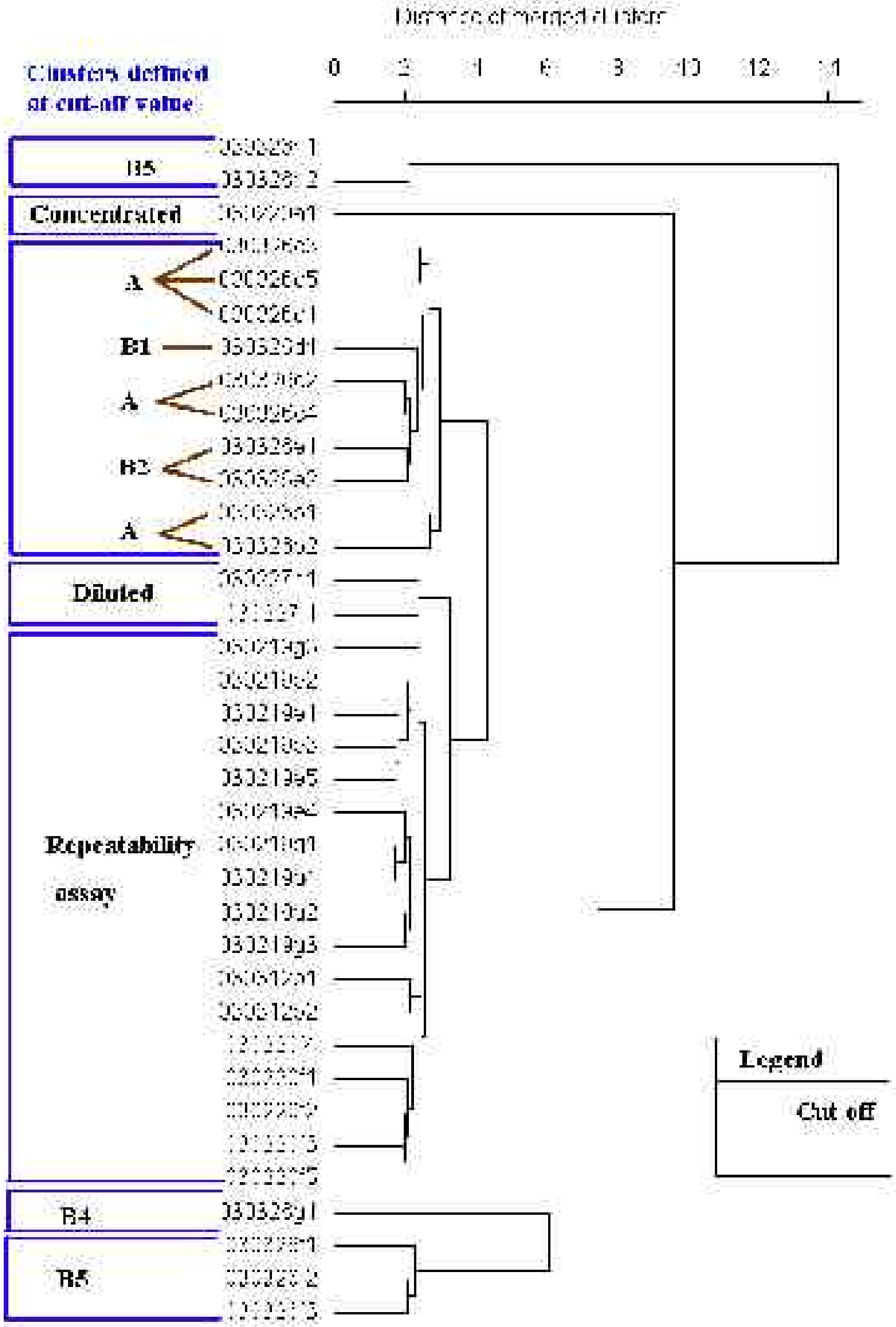
- Detection of changes in sample
  - due to e.g. heating
- Use circular dichroism spectroscopy
  - measures difference in absorbance by left & right polarized light
- Generates spectrum for each sample
  - Intensity vs wavelength
- Some spectra look similar, others don’t
- How to classify them?

Spectra display



Classification

- 36 spectra, 152 intensity values each
- Read into 36 x 152 matrix
- Passed to hierarchical cluster analysis routines
  - Euclidean distances between data points
  - Average link distances between clusters
- Output displayed as dendogram
  - tree plot showing merging of clusters with distance
- Introduce a cut-off to define “natural” clusters


Distance of merged clusters
0
2
4
6
8
10
12
14
Clusters defined at cut-off value
B5
Concentrated
A
B1
A
B2
A
Diluted
Repeatability assay
B4
B5
Legend
Cut off

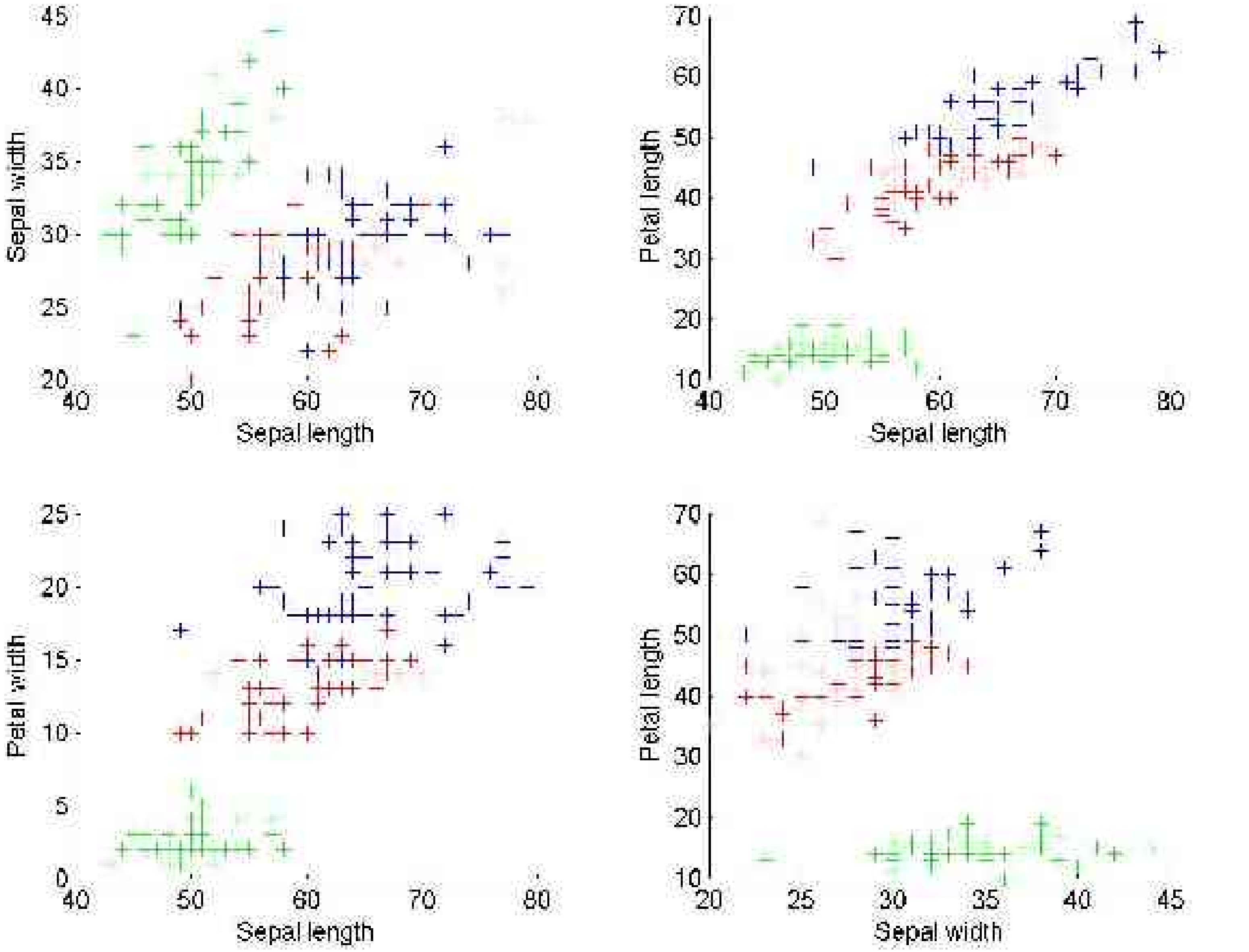
Analysis

- Cut off gives seven natural clusters
  - not v. sensitive to distance functions
- Some of the results can be understood w.r.t experimental conditions
  - e.g. 030220a1 - concentrated sample (evaporation)
  - e.g. 030319g5 to 030330f5 - repeatable experiment
- But there are some outliers
  - e.g. 030326e1, 030326e2 in normals
  - needs consultation with domain experts

Example application: Classification

- Fisher’s iris dataset
  - 4 measurements made on 50 iris specimens from each of three species
  - petal length, petal width, sepal length, sepal width
  - How to classify the species?
- 150 data points
- Each point
  - has 4 independent variables
  - belongs in one of three classes
    - red, green, blue
- How to display dataset?

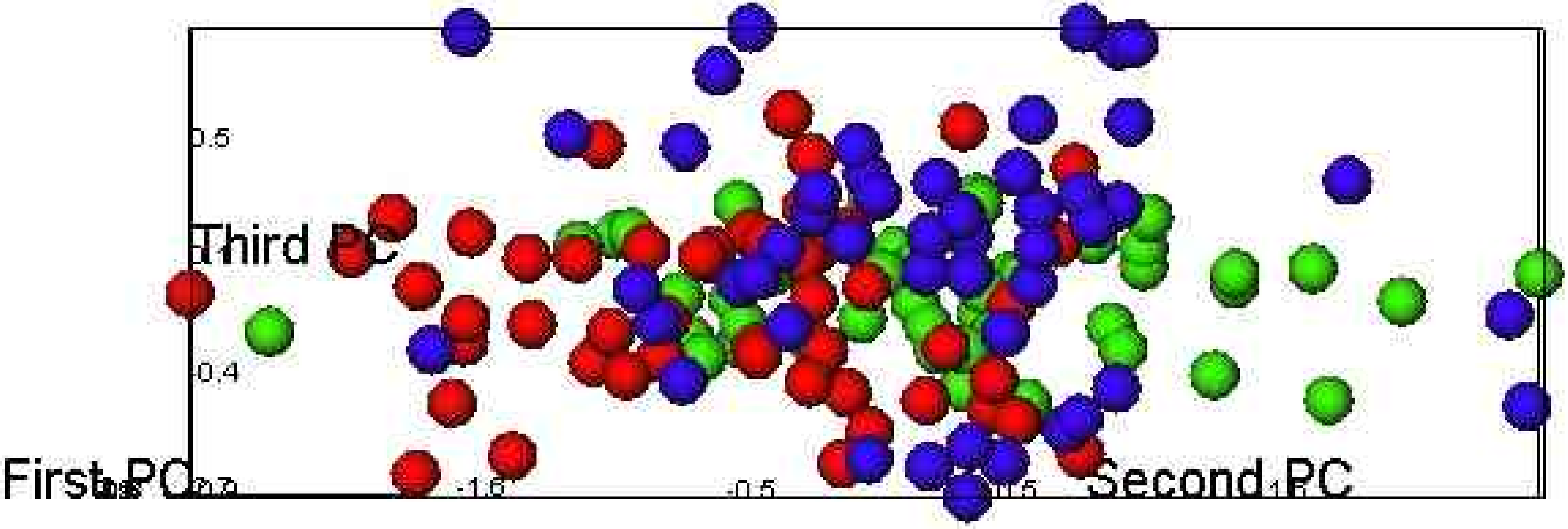
2D scatterplots



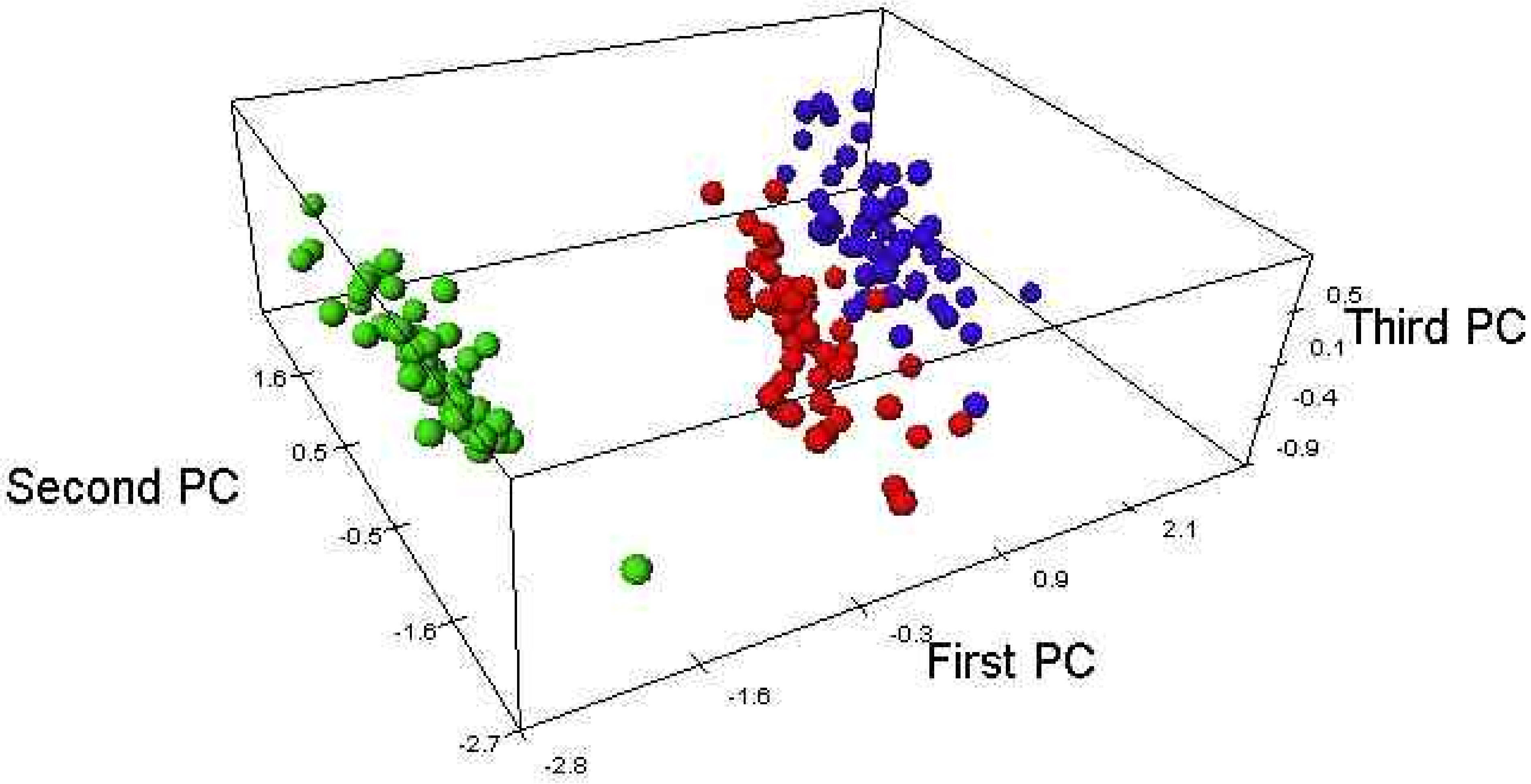
Scatterplots?

- In 2D, need 6 plots
  - to show each variable vs every other one
- Need to consider them all at the same time
- Can reduce the number of variables
  - using principal components analysis
  - first three explain ~95% of the variance
- Scatterplot in 3D

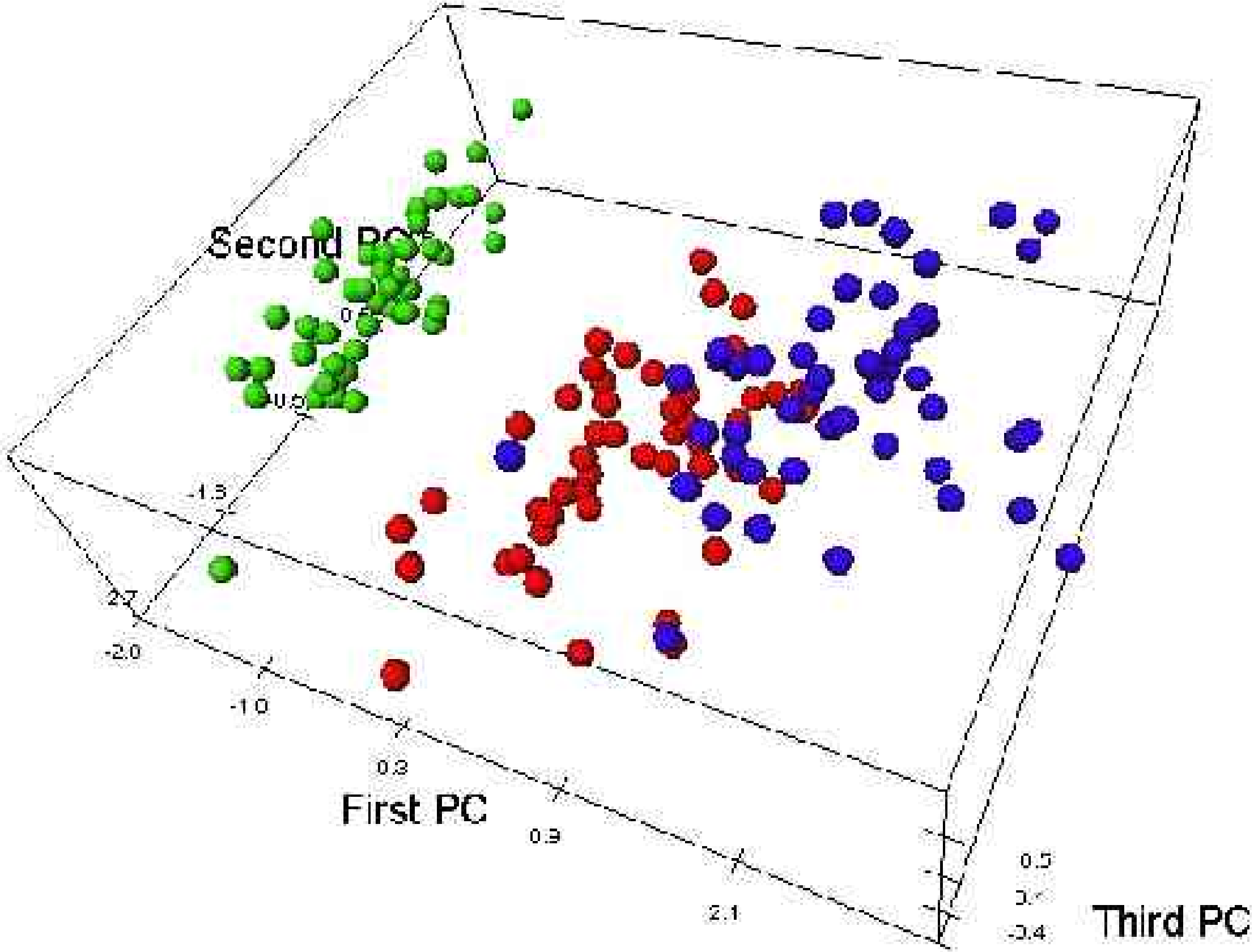
3D scatterplot



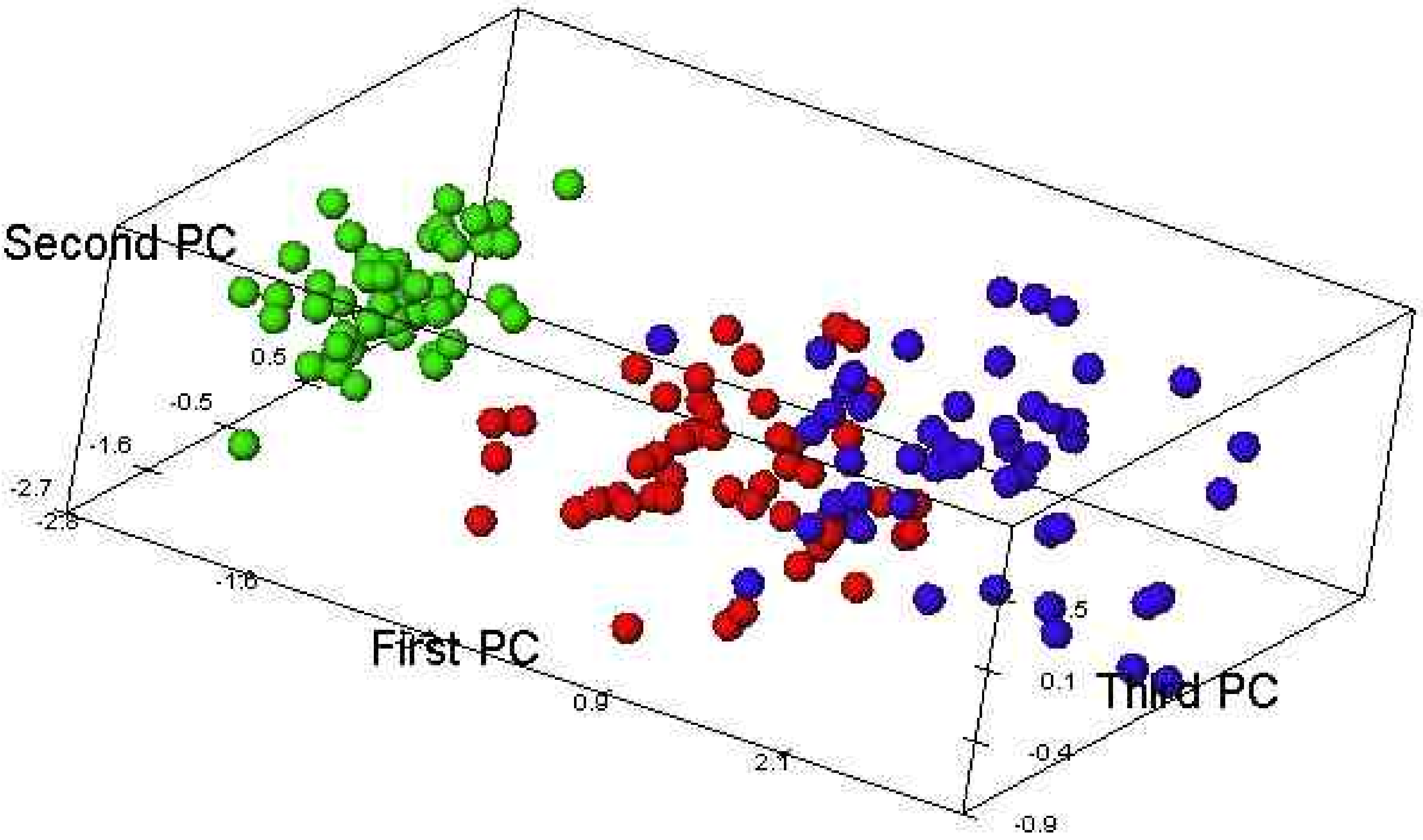
3D scatterplot



3D scatterplot



3D scatterplot

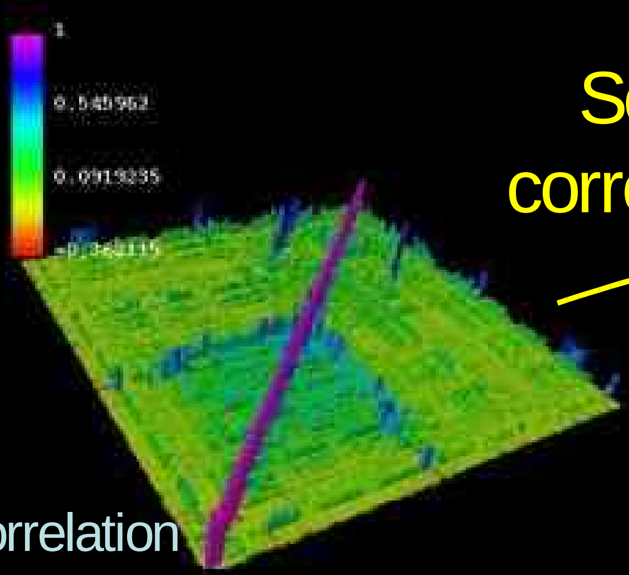


Overview

- Data mining components
  - Functionality
  - Example application
    - Quality control
- Visualization
  - Use of 3D
  - Example application
    - Market research
- Statistics and visualization in Excel
  - What’s the problem?

Example: Visual datamining

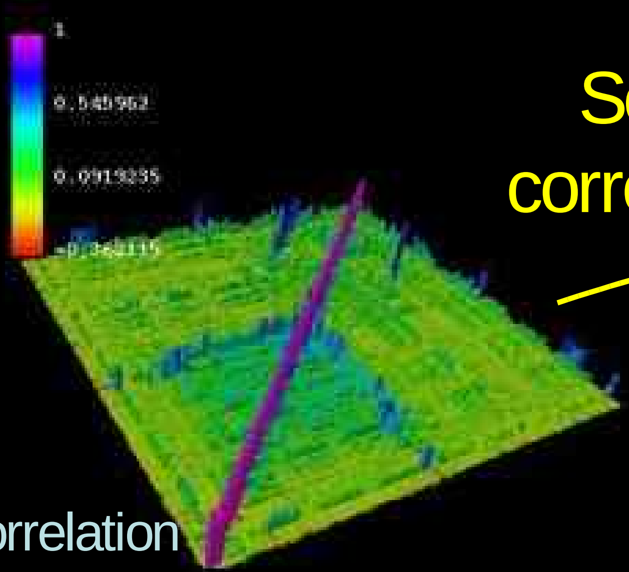
- Dutch financial asset management company
  - Keen to target products at entrepreneurs
  - How does entrepreneurship relate to other customer characteristics?
- Marketing dataset
  - 25,000 customers (sampled from full customer base)
  - Each characterised by values for 100 variables
    - Investment history = “entrepreneurship”
    - Income
    - Age
    - …



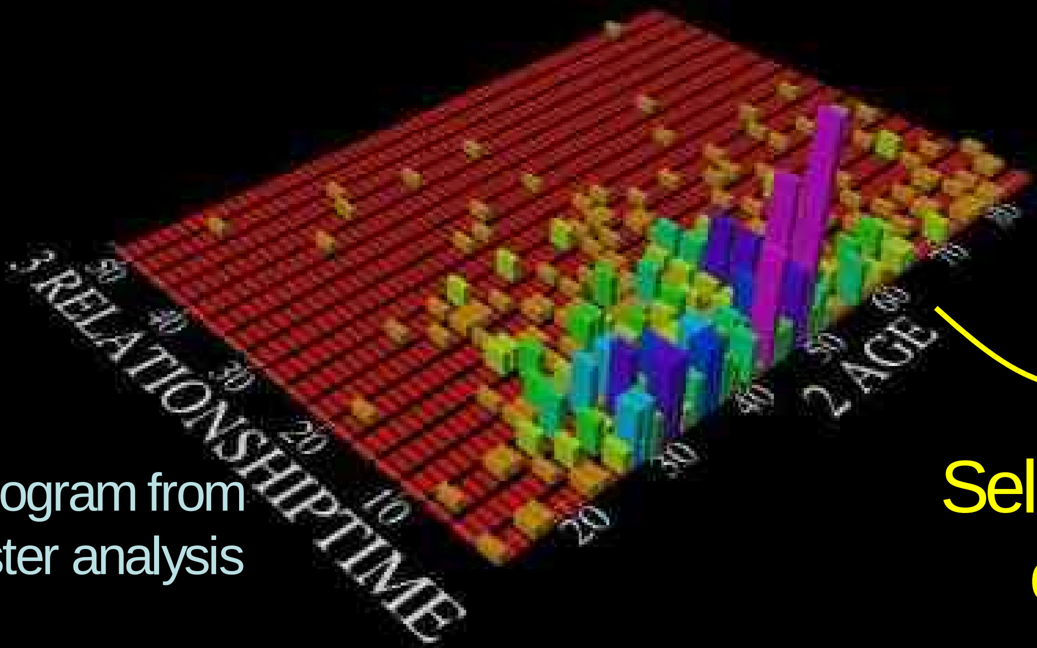
Correlation matrix

Select strongly correlated variables

Select strongly correlated variables

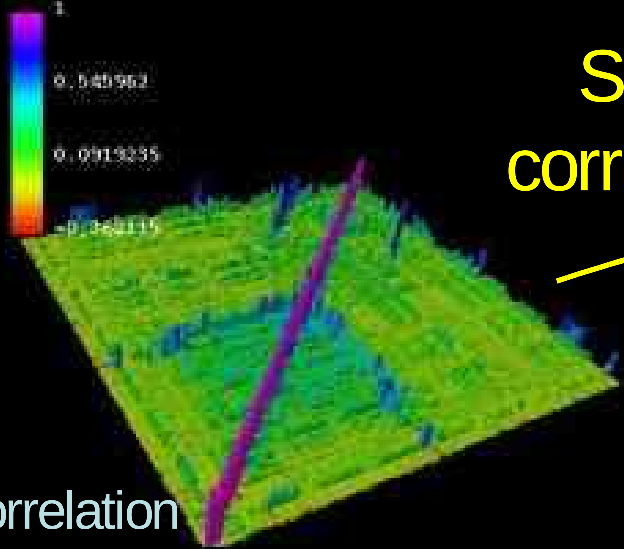


Correlation matrix



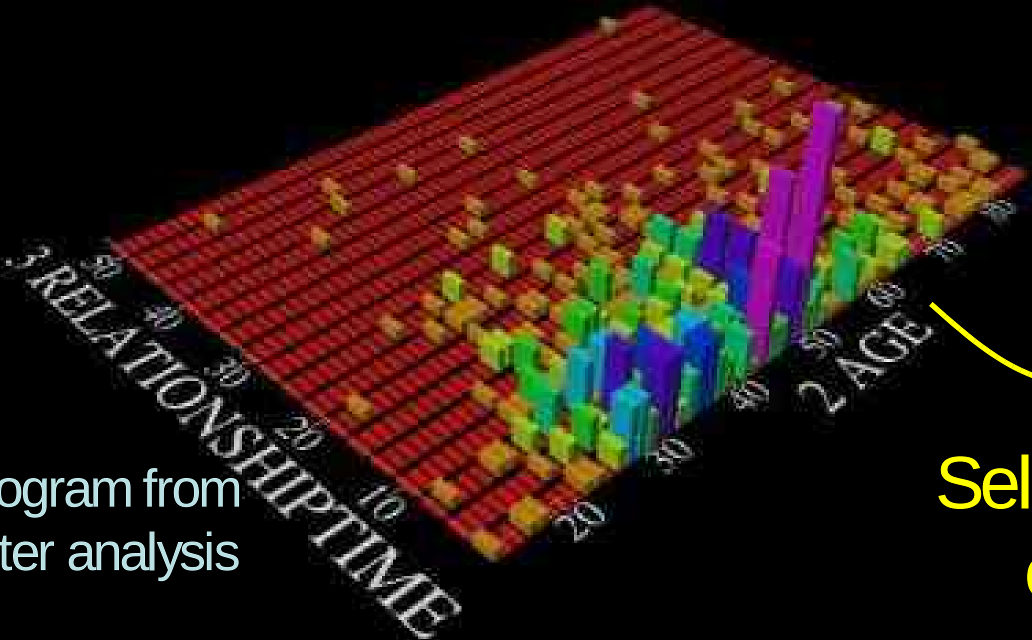
Histogram from cluster analysis

Select variables of interest



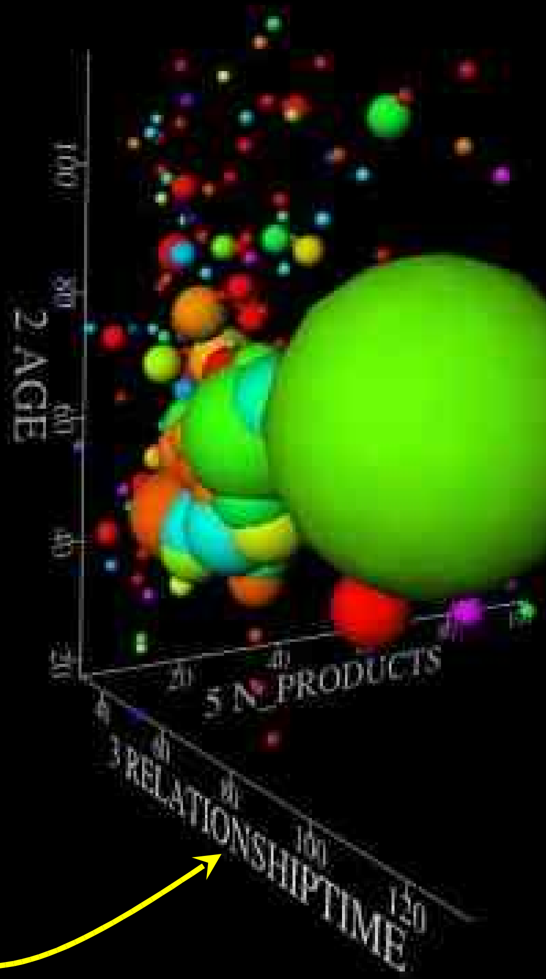
Correlation matrix

Select strongly correlated variables



Histogram from cluster analysis

Select variables of interest



Scatter plot

Lessons learnt

- 3D correlation landscape useful
  - Identifying significant variables
  - Focus on data distributions
  - Select appropriate ranges for cluster analysis
- Cluster visualization helpful
  - Non-linear relationships in data revealed
- 3D visualization combined with direct interaction
  - Selection of correlated variables
  - Binning and sorting
- Done with IRIS Explorer

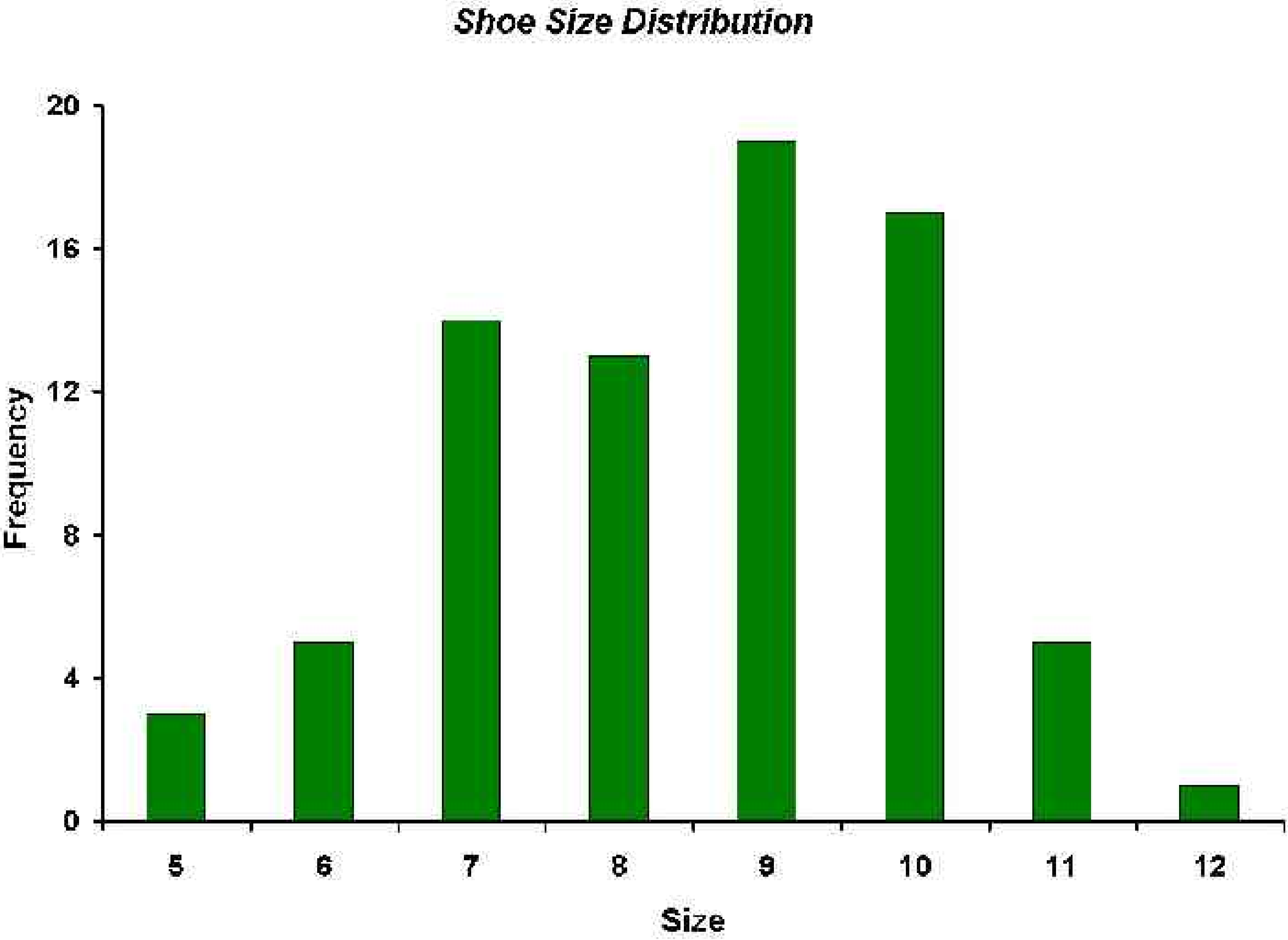
Overview

- Data mining components
  - Functionality
  - Example application
    - Quality control
- Visualization
  - Use of 3D
  - Example application
    - Market research
- Statistics and visualization in Excel
  - What’s the problem?

Simple visualization: shoe size survey

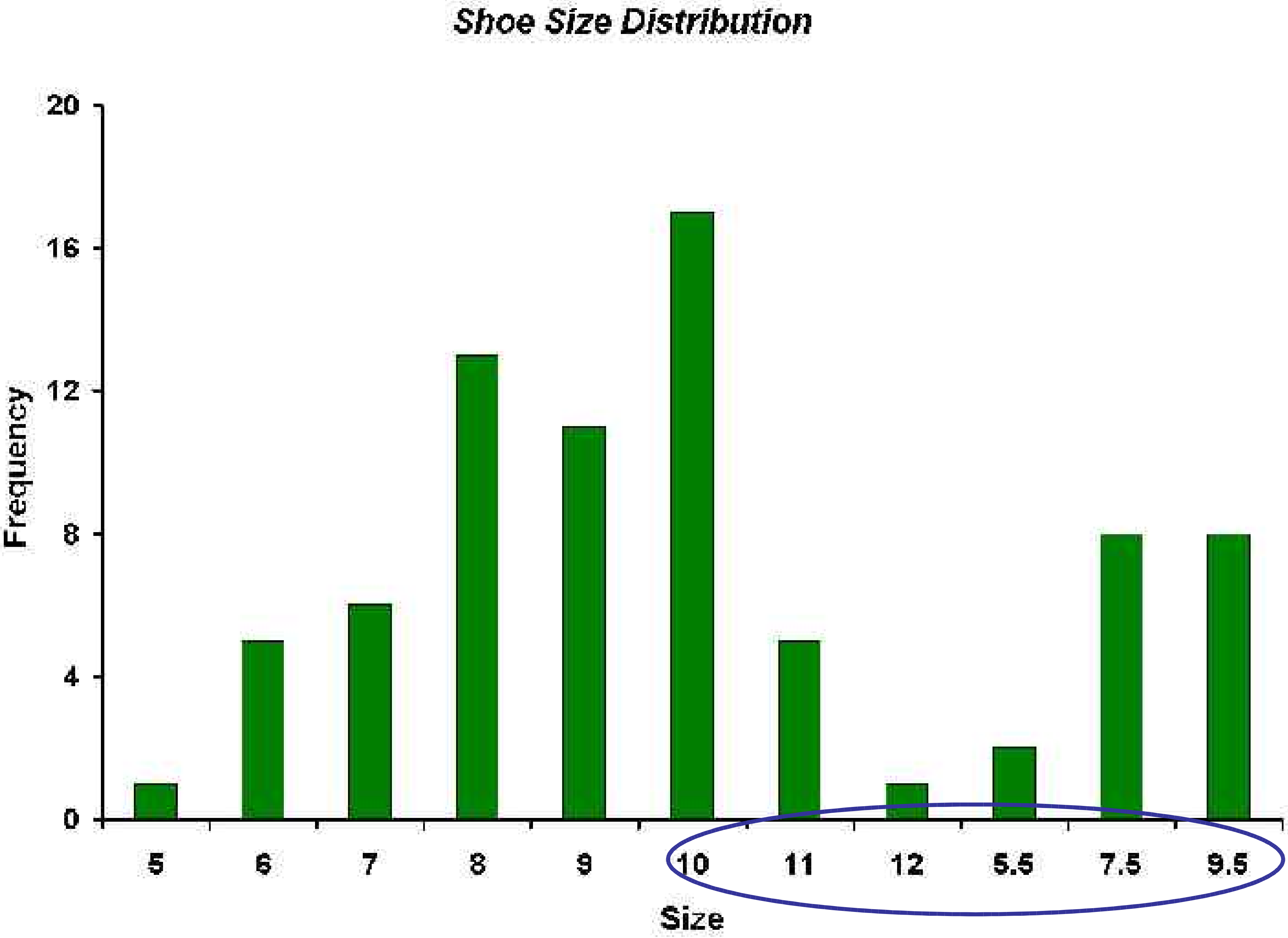
Shoe size	5	6	7	8	9	10	11	12
Frequency	3	5	14	13	19	17	5	1

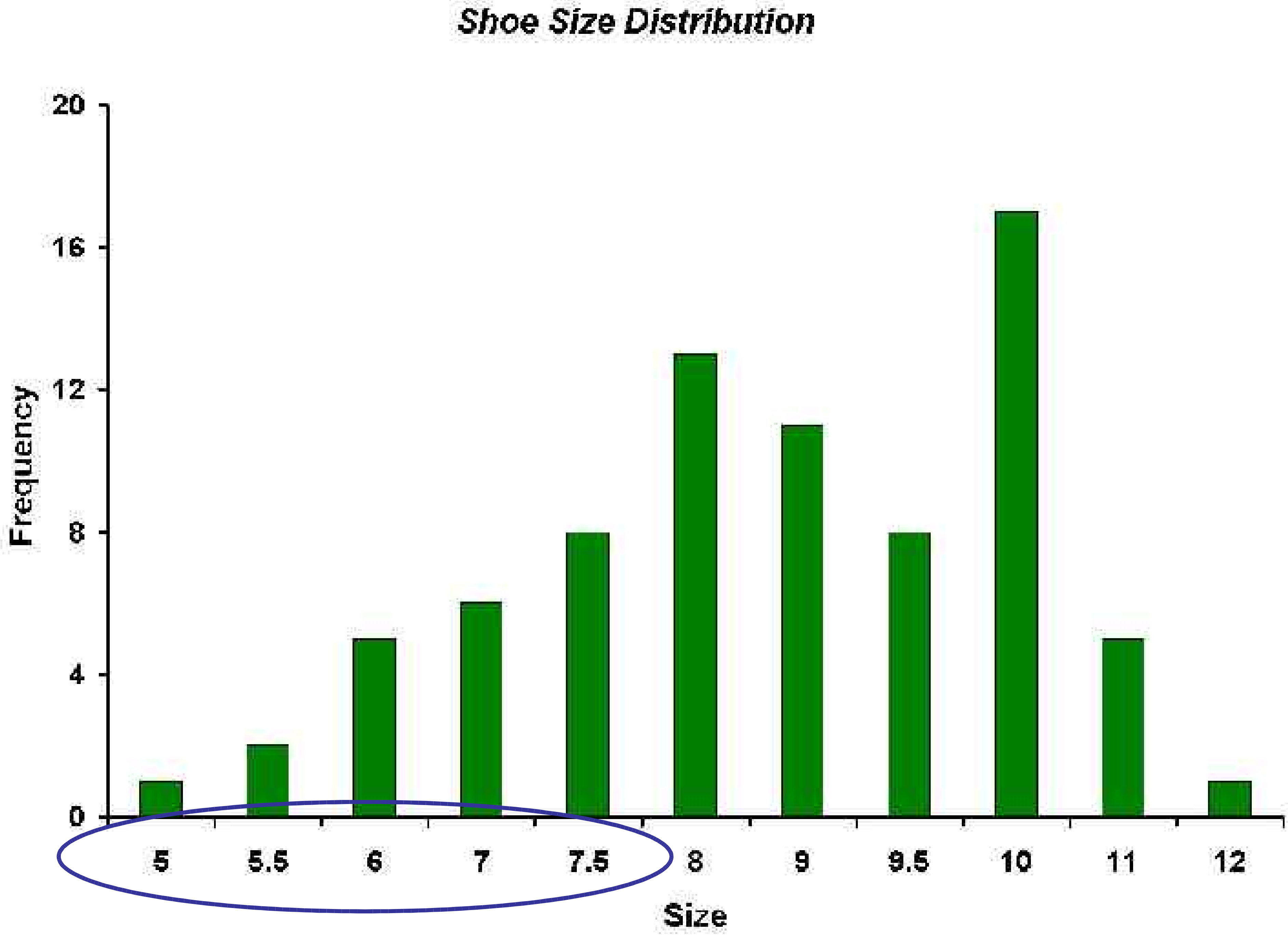
- Visualize using Excel
  - Chart wizard creates visualization easily
  - Interactive control over appearance
    - Colours, line width, text, fonts, placement
  - Data and visualization linked together


Shoe Size Distribution
Frequency
20
16
12
8
4
0
5
6
7
8
9
10
11
12
Size

Second survey (now with half-sizes)

Shoe size	5	6	7	8	9	10	11	12	5.5	7.5	9.5
Frequency	3	5	14	13	19	17	5	1	2	8	8


Shoe Size Distribution
Frequency
20
16
12
8
4
0
5
6
7
8
9
10
11
12
5.5
7.5
9.5
Size


Shoe Size Distribution
Frequency
0
4
8
12
16
20
5
5.5
6
7
7.5
8
9
9.5
10
11
12
Size

What’s wrong with this picture?

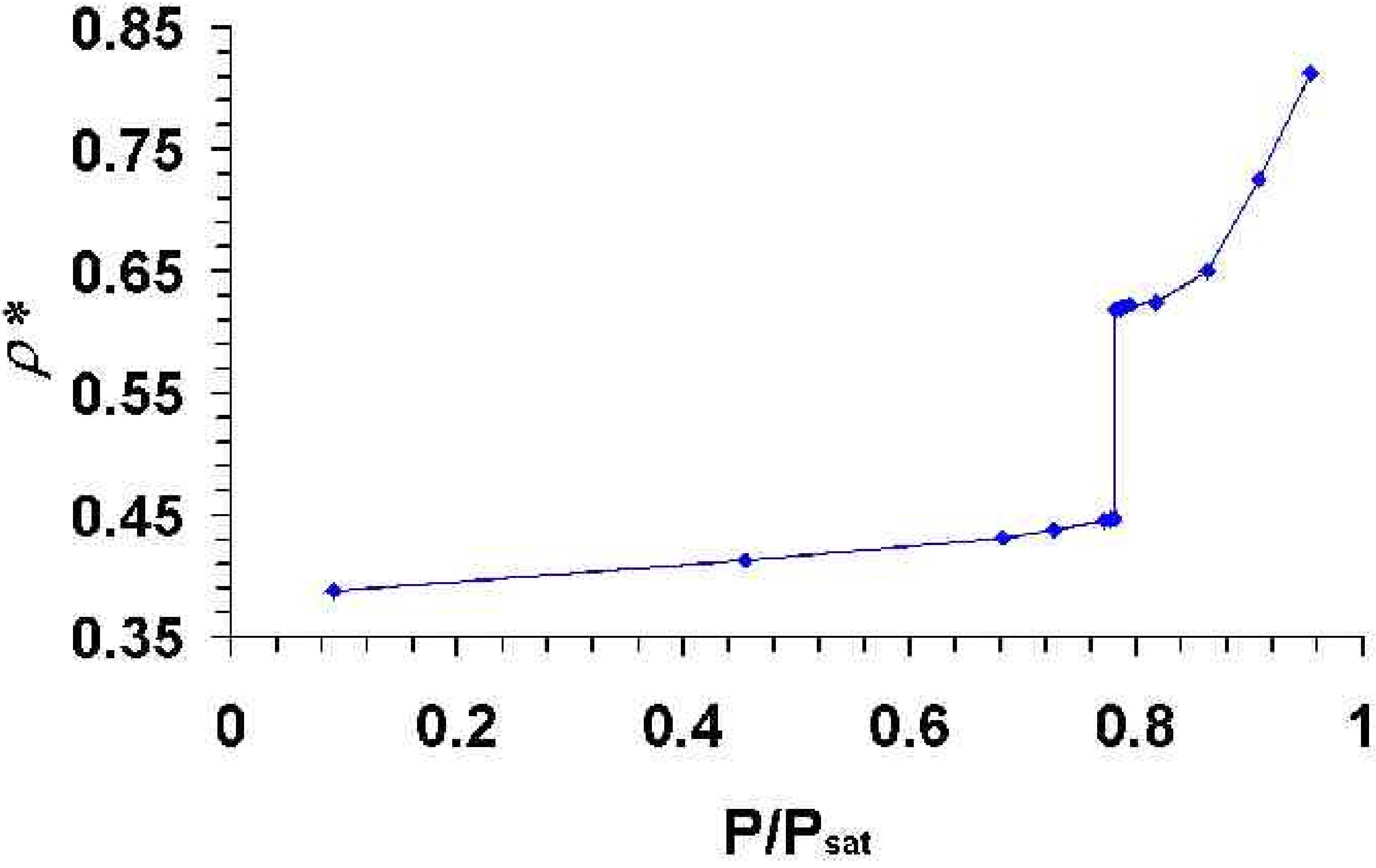
- Ordering of values on X axis reflects order in spreadsheet
  - not numerical order
- Spacing on X axis doesn’t reflect difference between values
  - spacing is everywhere the same
- Could pay close attention to labels
  - But might be harder for more complex data
  - Obscures missing data

Another example – adsorption isotherm

- Measurement of fluid density inside porous solid as a function of fluid pressure
- Confined fluid can condense before saturation
  - Capillary condensation
  - Vertical jump in isotherm
- Plot data using
  - Scatter plot
  - Line graph

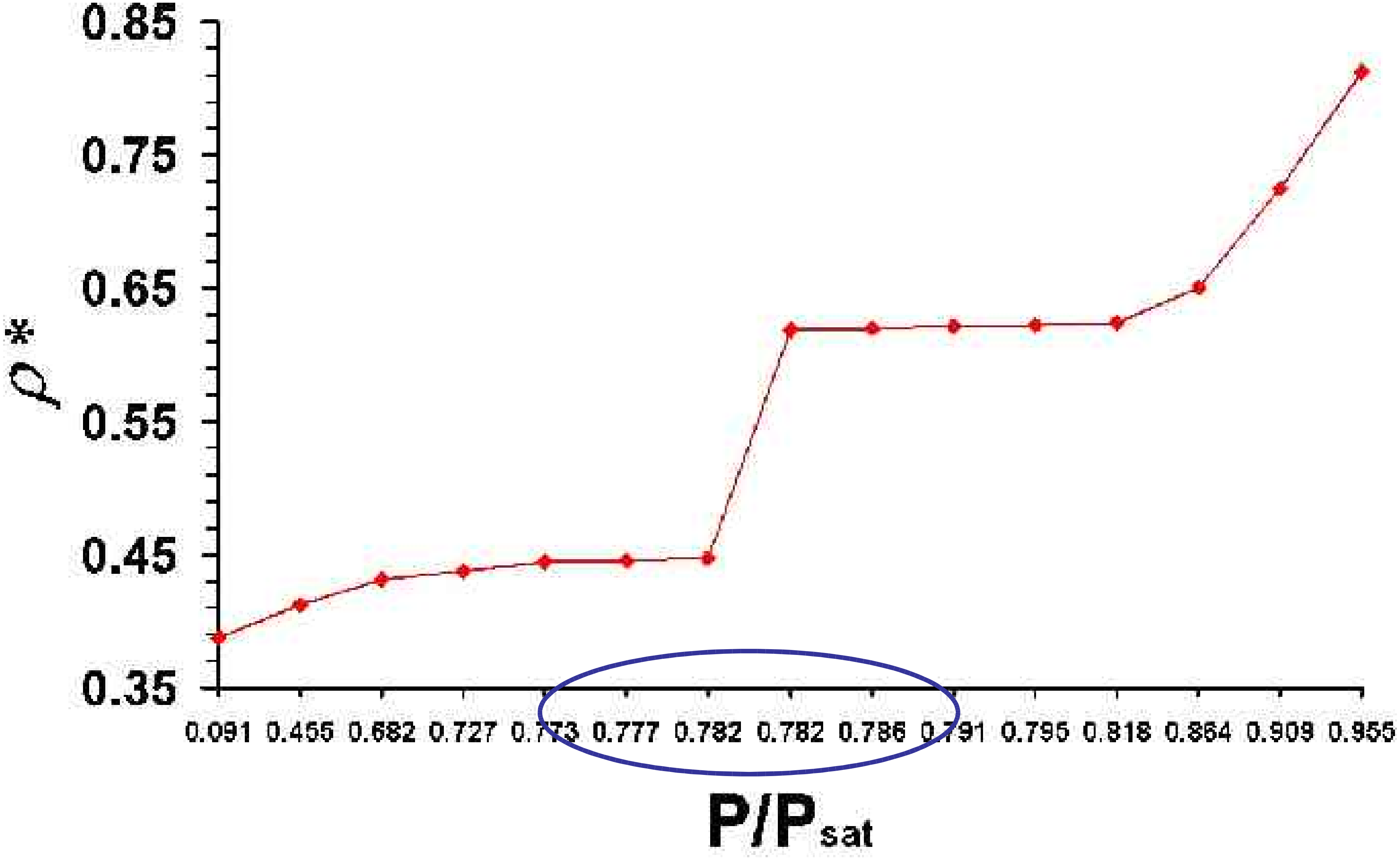
Scatter plot

Adsorption isotherm



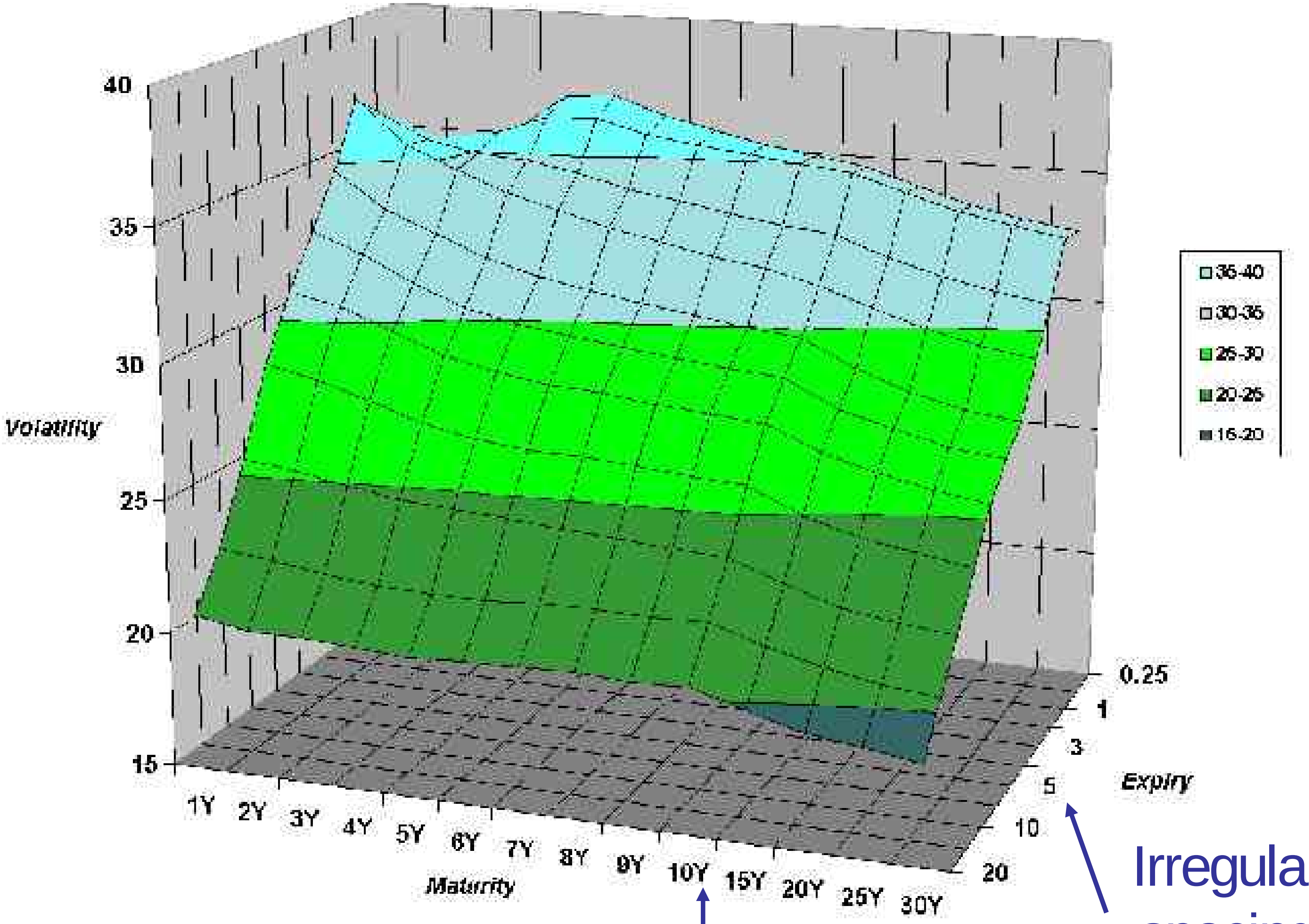
Line graph

Adsorption isotherm



Another example – derivative calculation

- Black-Scholes modelling of options on swap agreements (“swaptions”)
  - Option volatility as a function of
    - Swap maturity
    - Option expiry time
- Display in Excel as a surface plot

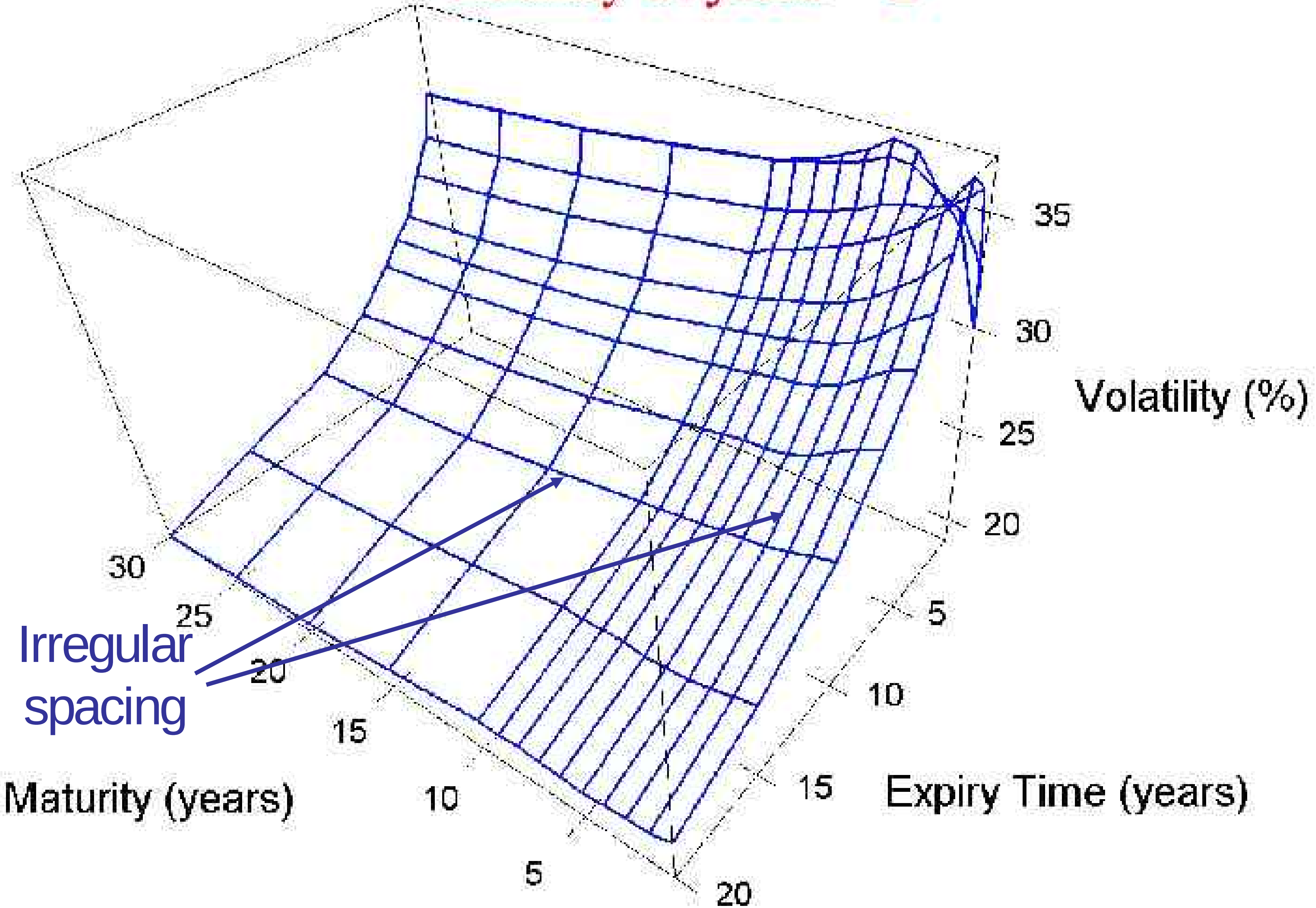


Irregular spacing

What’s wrong with this picture?

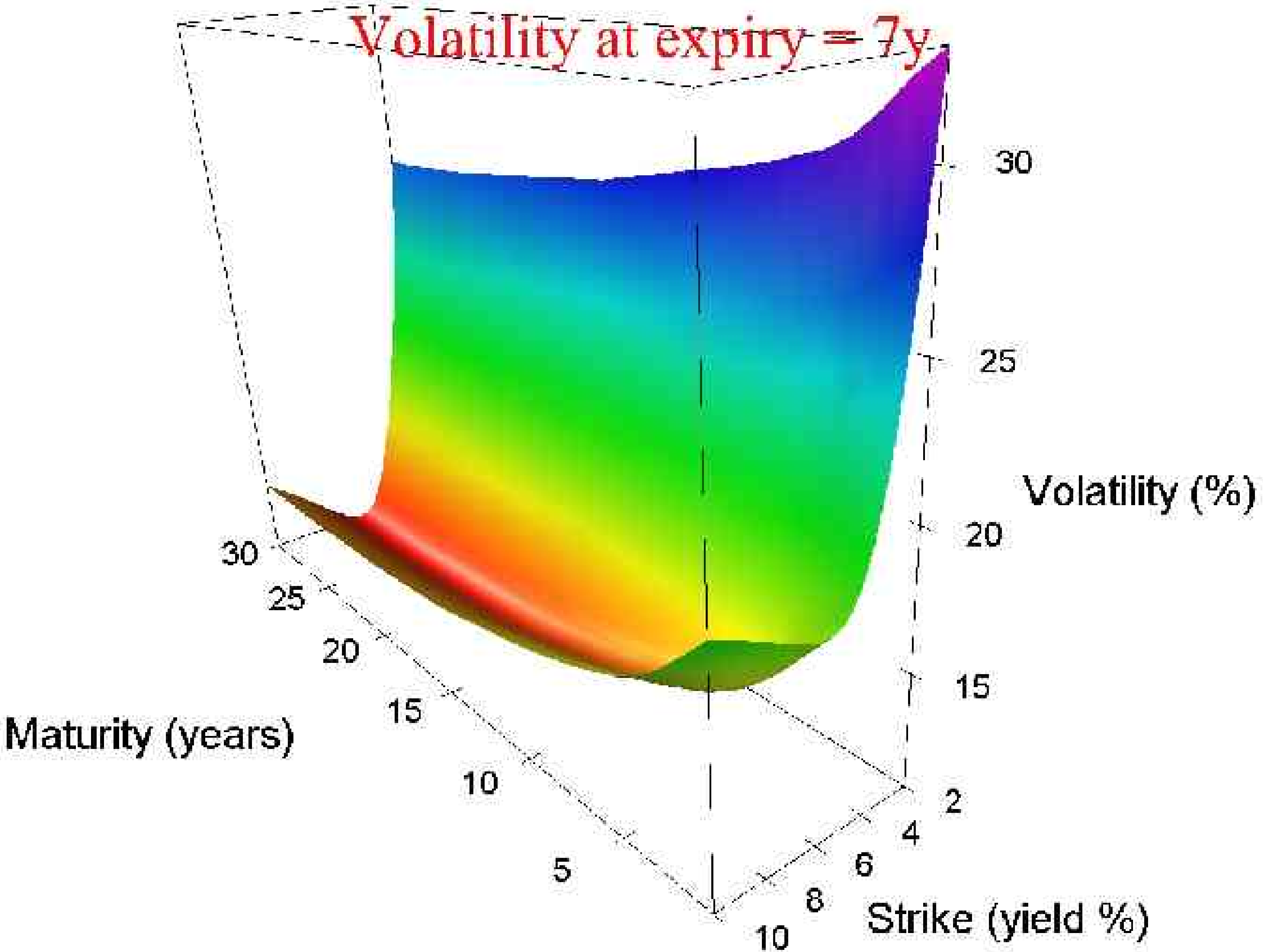
- See above
  - Irregular spacing, discontinuities
- This is only one part of the dataset
  - Volatility also depends on strike value
  - Want plots at other strike values
  - Want to see other relationships
    - e.g. volatility (strike) = “volatility smile”
- Use IRIS Explorer

Volatility at yield = 2

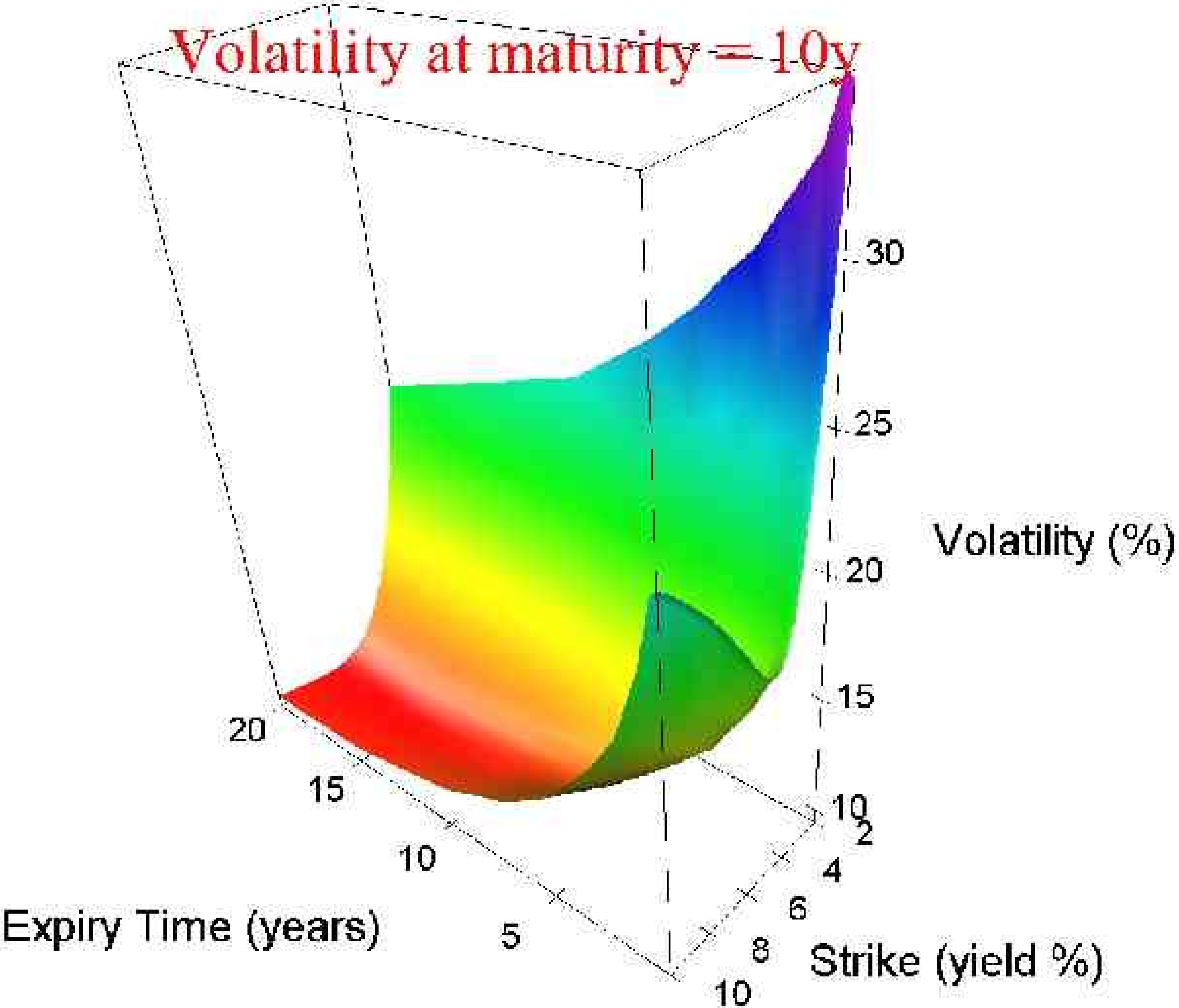


Irregular spacing

Volatility at expiry = 7y



Volatility at maturity = 10y



Two types of axis in Excel charts

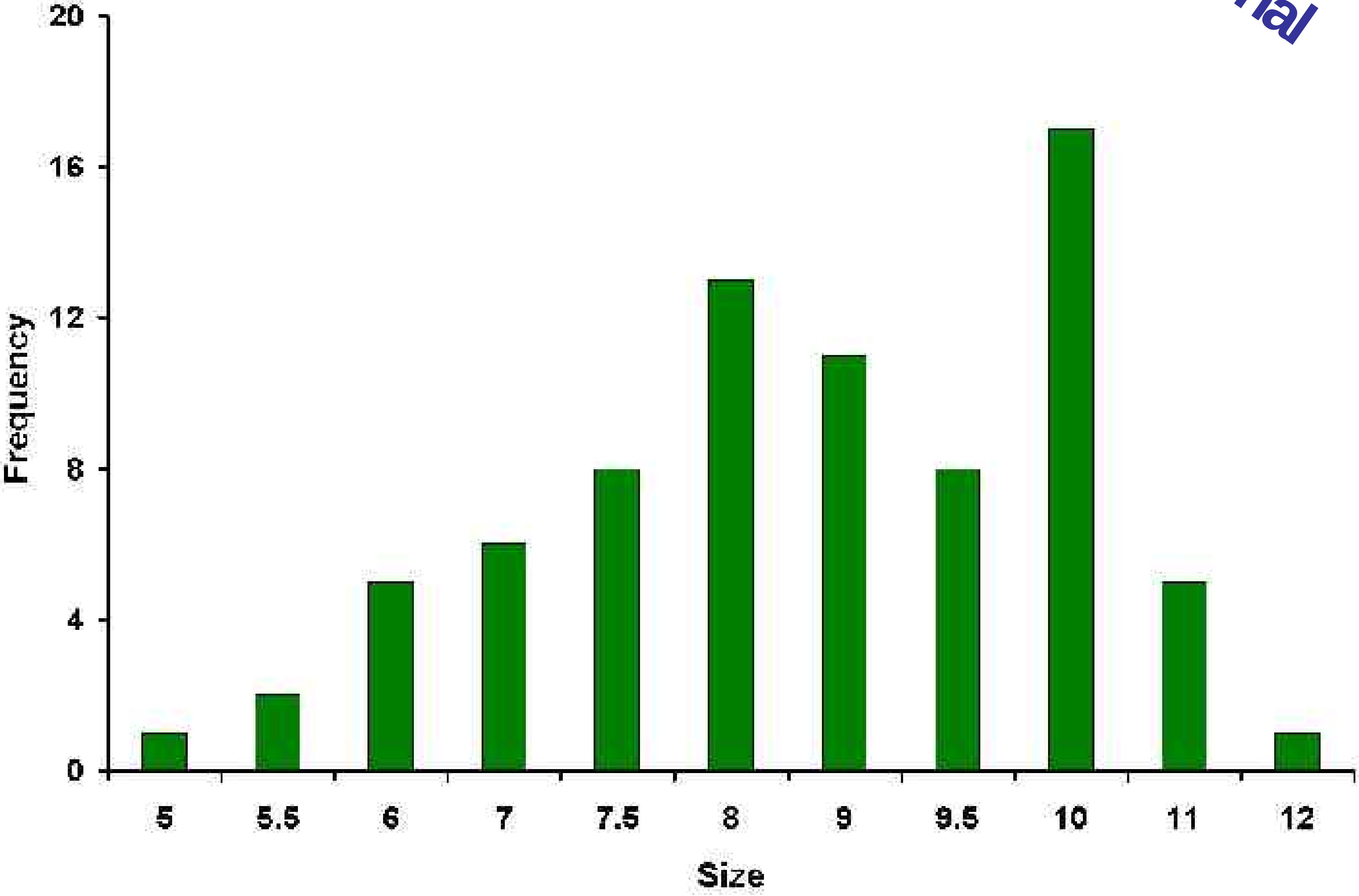
- Value
  - Data treated as continuously varying numerical values
  - Marker placed at location reflecting its value
  - Used in Excel Scatter plots
- Category
  - Data treated as sequence of non-numerical text labels
  - Marker location reflects position in sequence
  - Points distributed evenly along axis
  - Used in Excel Bar chart, Line chart, Surface plot…

NAG Schools Excel Add-in (N-SEA)

- Supports instruction in statistics
- Functionality for
  - Data sampling
  - Frequency plots
  - Box and whisker plots
  - Histograms
  - Continuous bar charts
- Allows (X/Y) ordering of data points in plotting
  - and inclusion of points with zero Y values

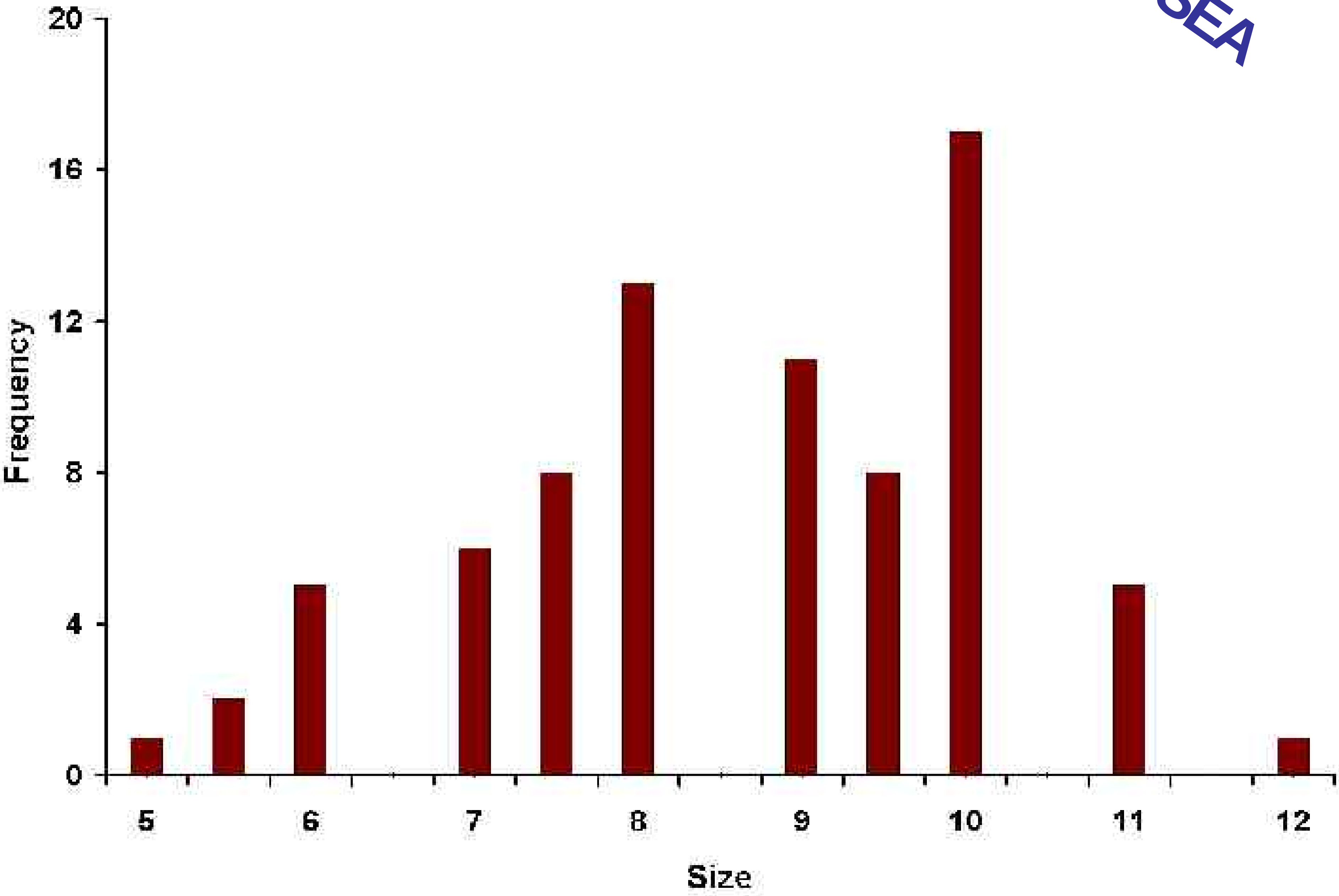
Original

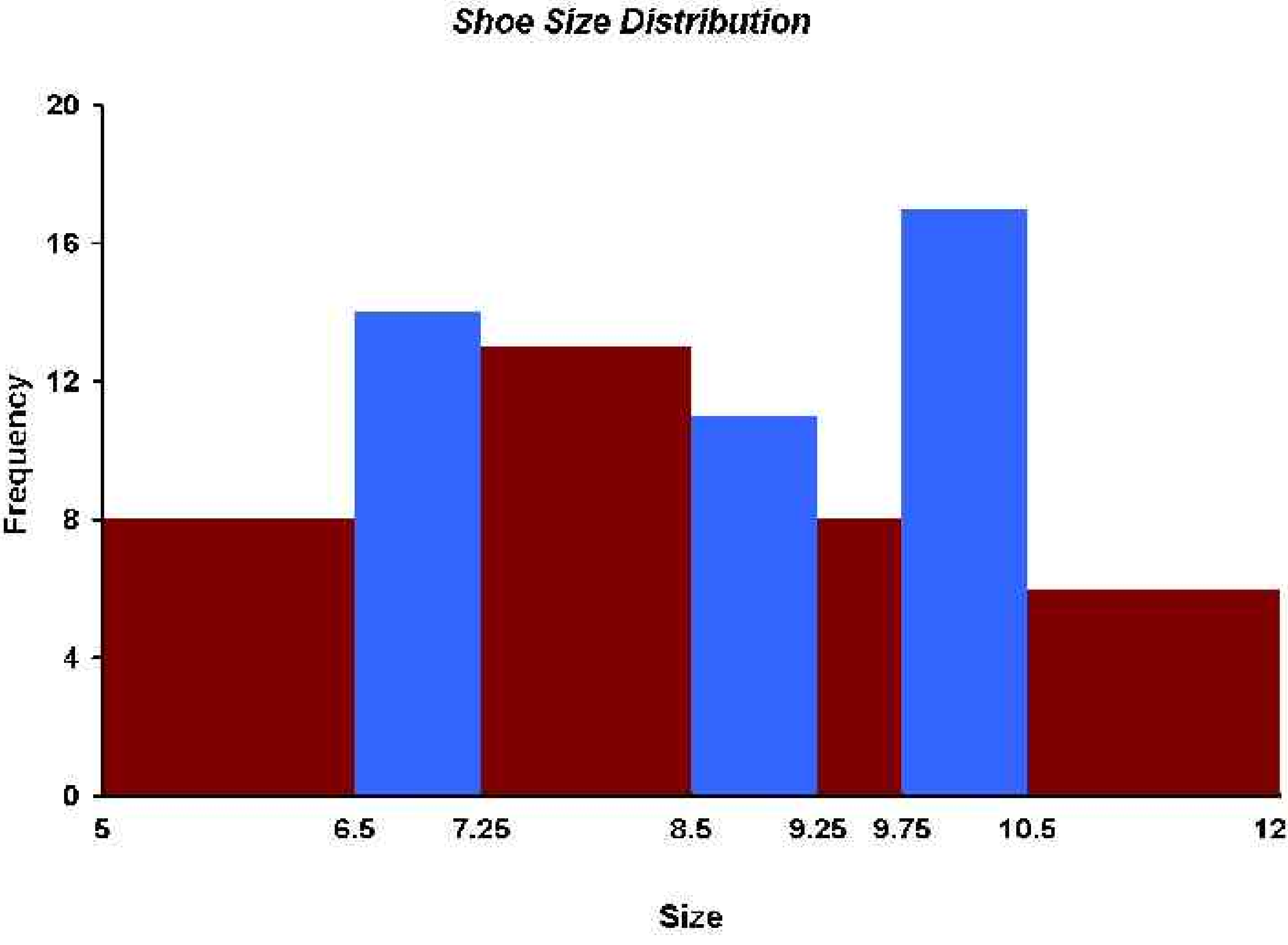
Shoe Size Distribution



Shoe Size Distribution

N-SEA




Shoe Size Distribution
Frequency
20
16
12
8
4
0
5
6.5
7.25
8.5
9.25
9.75
10.5
12
Size

Conclusions

- Data mining components offer basic routines
  - Developers can incorporate them into applications
  - No wheel-reinvention, stone canoes, chocolate teapots
    - cf NAG numerical library
- Visualization is crucial for analysis
  - Integration of data mining & visualization is application-dependent
  - Interactivity important
- Problems with (even) well-known tools
  - Be aware
  - Work around