



# Data Visualizations - Introduction

Pattabiraman V

SCOPE

# Topics

- Definition
- Importance & Benefits
- Limitation
- Data Types
- Data Visualisation
- Dashboard
- Tools Available

# Objectives

- Core principles of visualization
- Visual representation of data
- Visualization tools

# Data Visualisation

- Data visualization is the use of abstract, non-representational pictures to show numbers
- It can include points, lines, symbols, words, shading and colour
- Is the process of converting raw data into easily understood pictures of information that enable fast and effective decisions
- Visual Representation of Data
- For exploration, discovery, insight, ...
- Interactive component provides more insight as compared to a static image

# Data Visualisation

“Transformation of the symbolic into the geometric”

(McCormick et al., 1987)

“... finding the artificial memory that best supports our natural means of perception.”

(Bertin, 1983)

The depiction of information using spatial or graphical representations, to facilitate comparison, pattern recognition, change detection, and other cognitive skills by making use of the visual system

# Objectives of Information Visualization

- Make large datasets coherent  
(Present huge amounts of information compactly)
- Present information from various viewpoints
- Present information at several levels of detail  
(from overviews to fine structure)
- Support visual comparisons
- Tell stories about the data

# Visualisation – Types

- Scientific Visualization
  - Structural Data – RDBMS, DW, Seismic, Medical, ..
- Information Visualization
  - No inherent structure – Streaming, RNews, stock market, top grossing movies, facebook connections
- Visual Analytics
  - Use visualization to understand and synthesize large amounts of multimodal data – audio, video, text, images, networks of people

# Importance / Benefits

- Data visualization allows users see several different perspectives of the data.
- Data visualization makes it possible to interpret vast amounts of data
- Data visualization offers the ability to note exceptions in the data
- Data visualization allows the user to analyze visual patterns in the data
- Exploring trends within a database through visualization by letting analysts navigate through data and visually orient themselves to the patterns in the data
- Huge impact on policy, planning and disaster avoidance



# Importance / Benefits

- Data visualization can help translate data patterns into insights, making it a highly effective decision-making tool.
- Data visualization equips users with the ability to see influences that would otherwise be difficult to find.
- With all the data available, it is difficult to find the nuances that can make a difference.
- By simplifying the presentation, Data Visualization can reduce the time and difficulty it takes to move from data to decision making

# Impact of Visualisation – Classic Examples

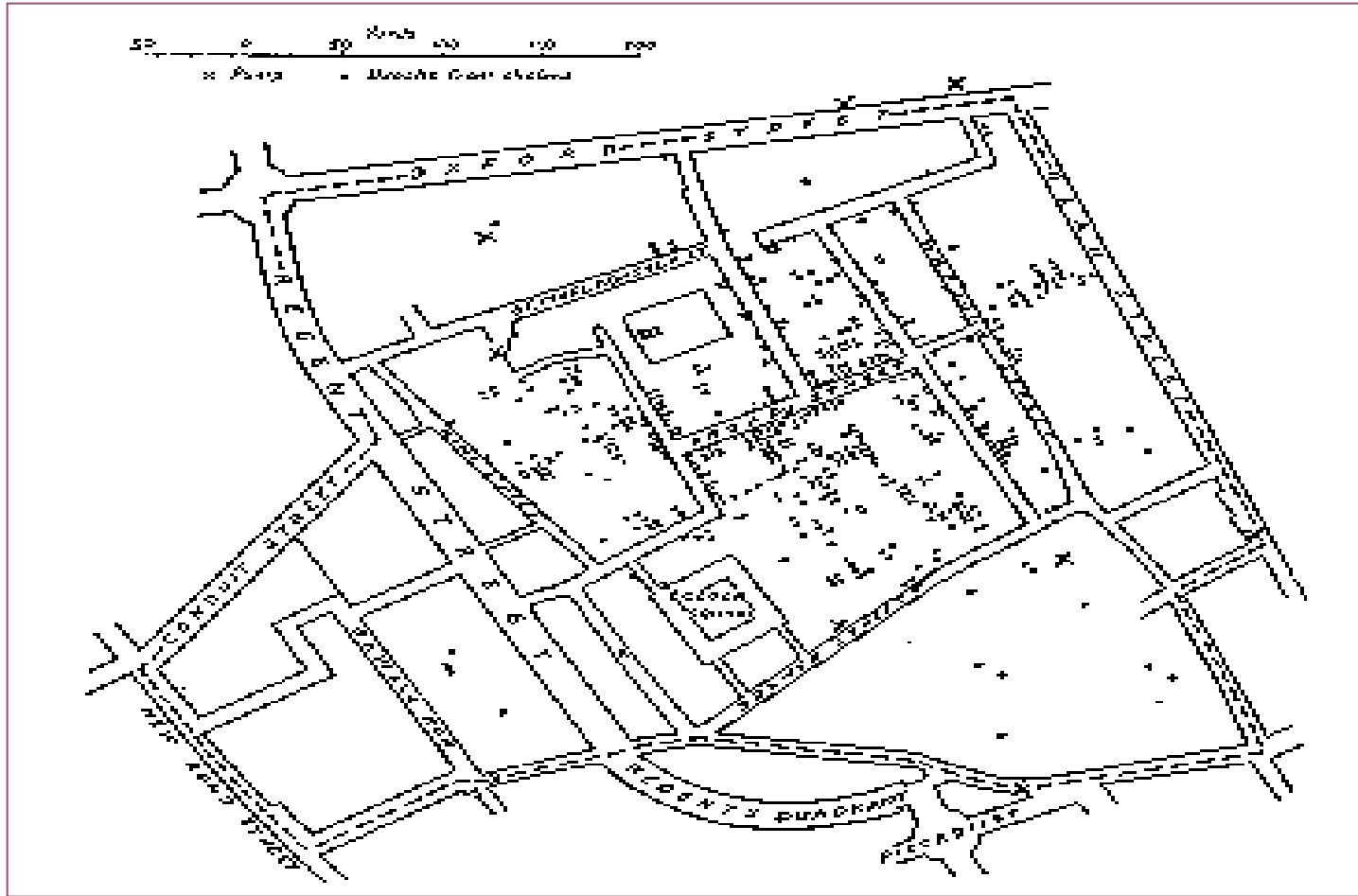


Illustration of John Snow's deduction that a cholera epidemic was caused by a bad water pump, circa 1854.

Horizontal lines indicate locations of deaths.

[illegible]

# Why Visualisation

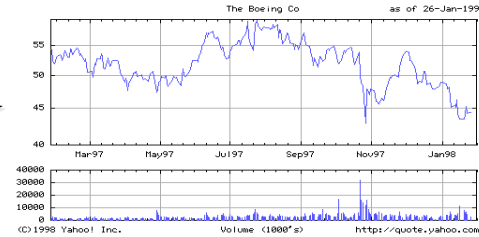
- Use the eye for pattern recognition; people are good at
  - Scanning, recognizing & remembering images
- Animation shows changes across time
- Color helps make distinctions
- Aesthetics help maintain interest
- Graphical elements facilitate comparisons via
  - Length, shape, orientation & texture

# Visual Principles

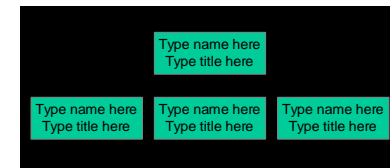
- Types of Graphs
- Pre-attentive Properties
- Relative Expressiveness of Visual Cues
- Visual Illusions
- Tufte's notions
  - Graphical Excellence
  - How to Lie with Visualization
  - Data-Ink Ratio Maximization

# Types of Symbolic Displays

- Graphs



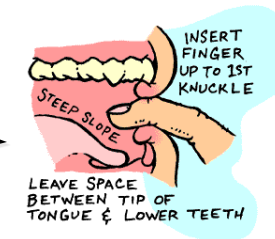
- Charts



- Maps



- Diagrams



# Types of Symbolic Displays

## Graphs

- at least two scales required
- values associated by symmetric “paired with” relation
- Examples: scatter-plot, bar-chart, layer-graph



# Elements of a Graph

- Framework
  - sets the stage
  - kinds of measurements, scale, ...
- Content
  - marks
  - point symbols, lines, areas, bars, ...
- Labels
  - title, axes, tic marks, ...

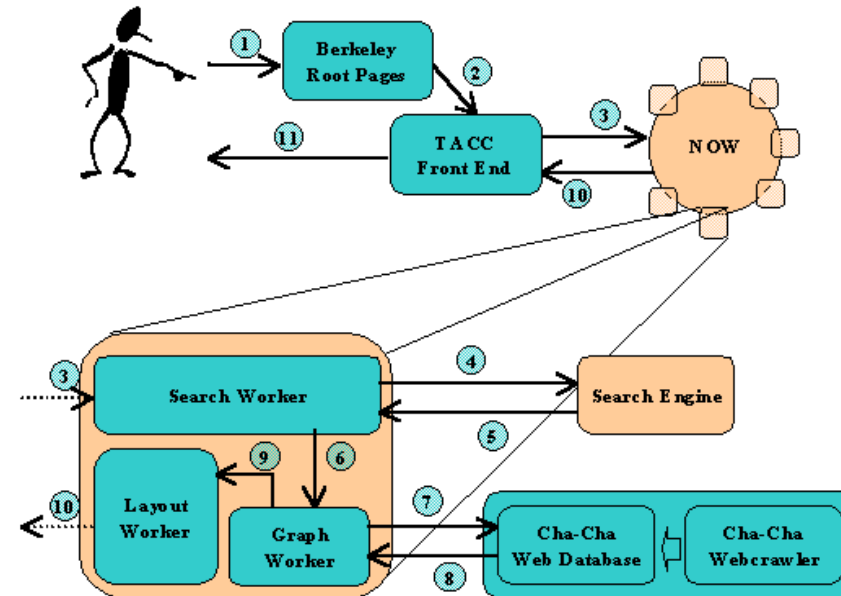


# Types of Symbolic Displays

## Charts

- discrete relations among discrete entities
- structure relates entities to one another
- lines and relative position serve as links
- Examples:

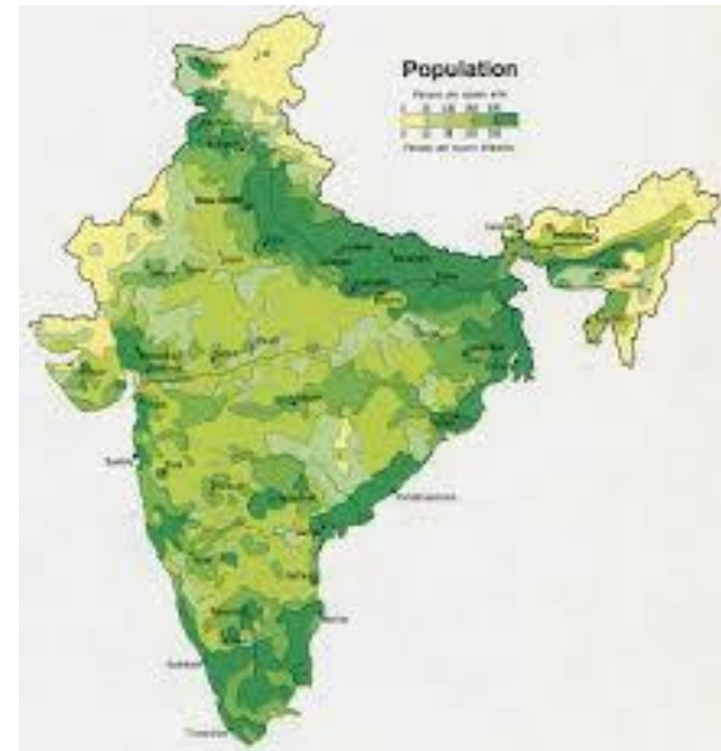
family tree,  
flow chart,  
network diagram



# Types of Symbolic Displays

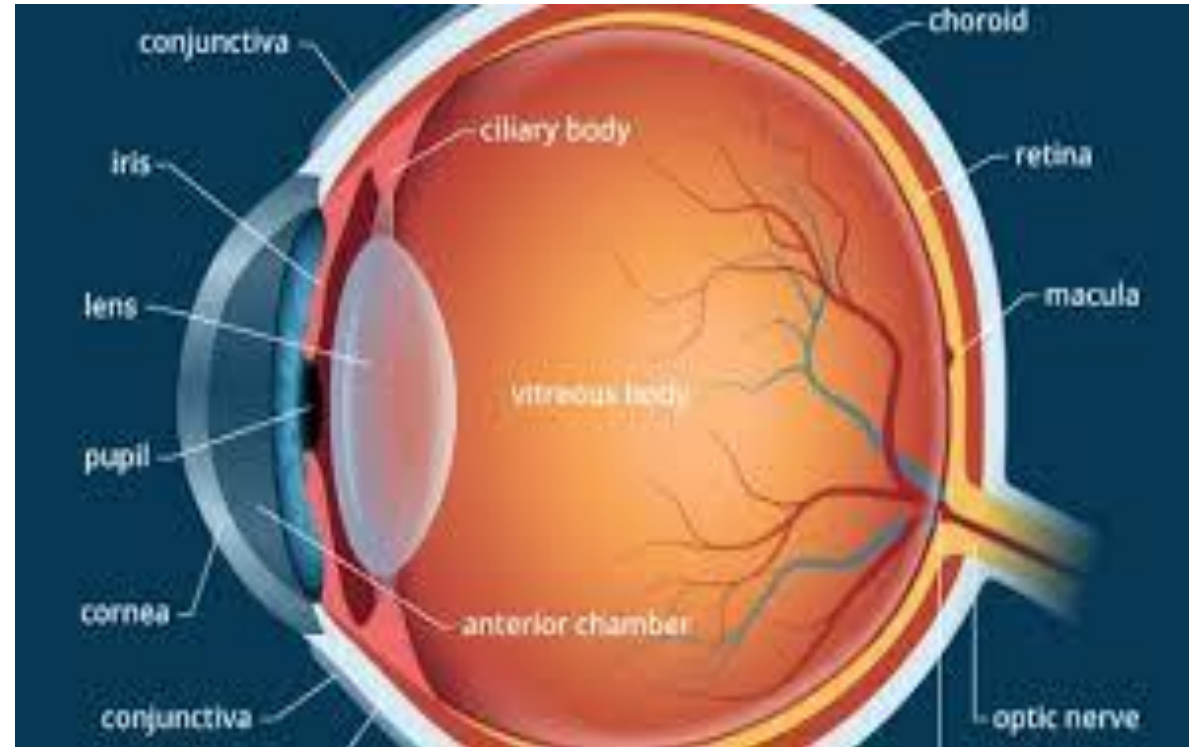
## Maps

- internal relations determined (in part) by the spatial relations of what is pictured
- labels paired with locations
- Examples:
  - physical maps,
  - topographic maps,
  - political maps,
  - maps of census data



# Types of Symbolic Displays

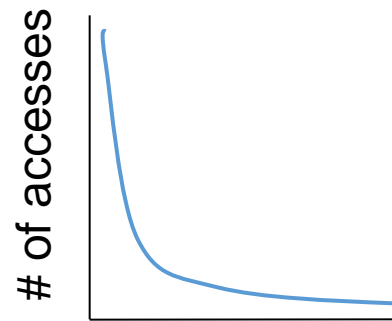
- Diagrams
  - schematic pictures of objects or entities
  - parts are symbolic (unlike photographs)
  - Examples:
    - how-to illustrations,
    - figures in a manual



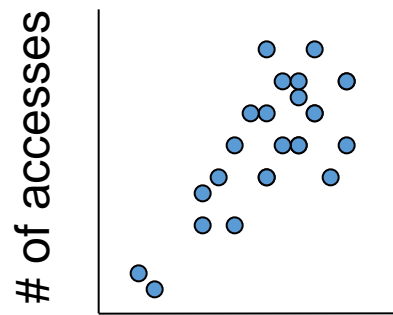
# Basic Types of Data

- Nominal (qualitative)
  - (no inherent order)
  - city names, types of diseases, ...
- Ordinal (qualitative)
  - (ordered, but not at measurable intervals)
  - first, second, third, ...
  - cold, warm, hot
- Interval (quantitative)
  - list of integers or reals

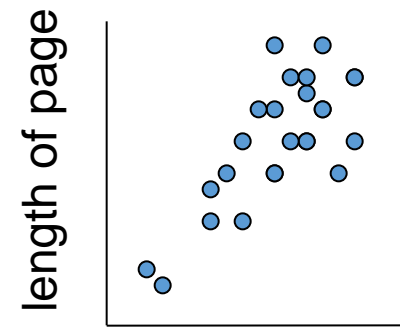
# Common Graph Types



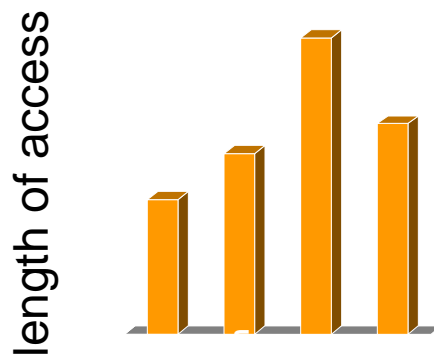
URL



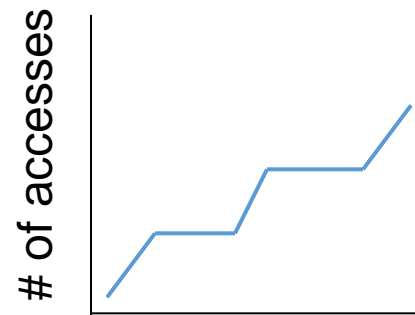
length of access



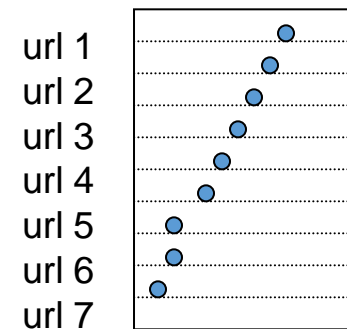
length of access



length of page



days



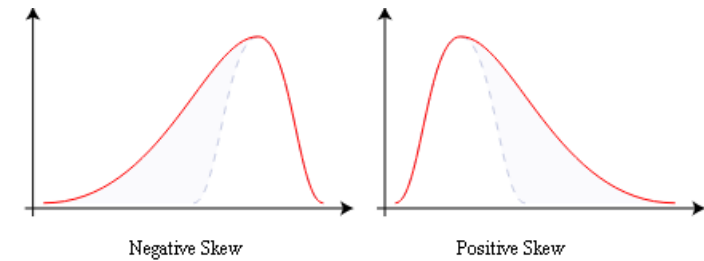
# of accesses

# Exploratory Data Analysis (EDA)

- Gives a general sense of the data
  - means, medians, quantiles, histograms, boxplots
- Especially useful in early stages of data mining
  - detect outliers (e.g. assess data quality)
  - test assumptions (e.g. normal distributions or skewed?)
  - identify useful raw data & transforms (e.g.  $\log(x)$ )

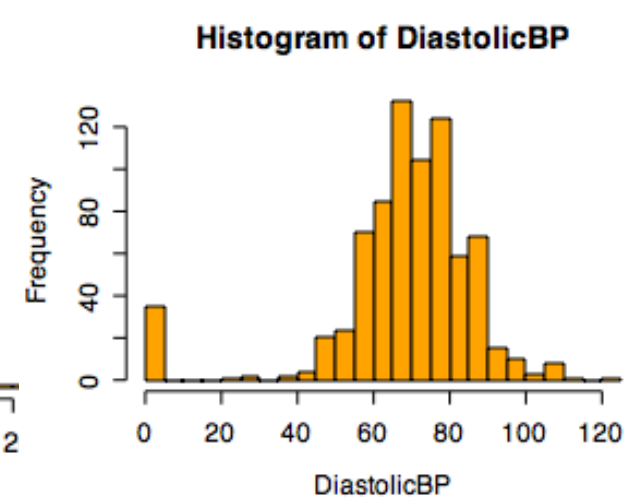
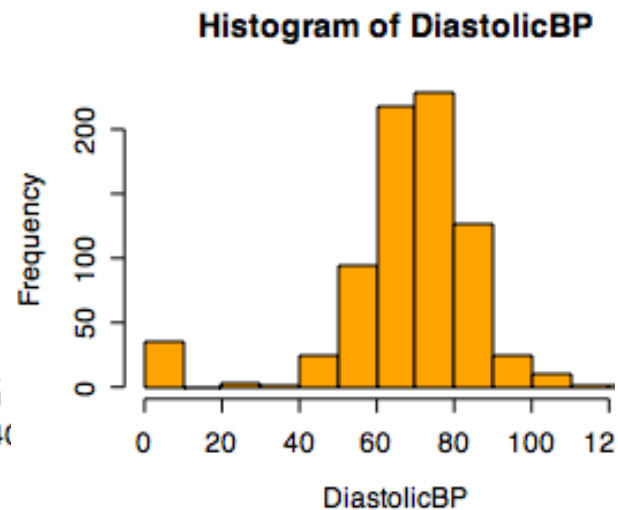
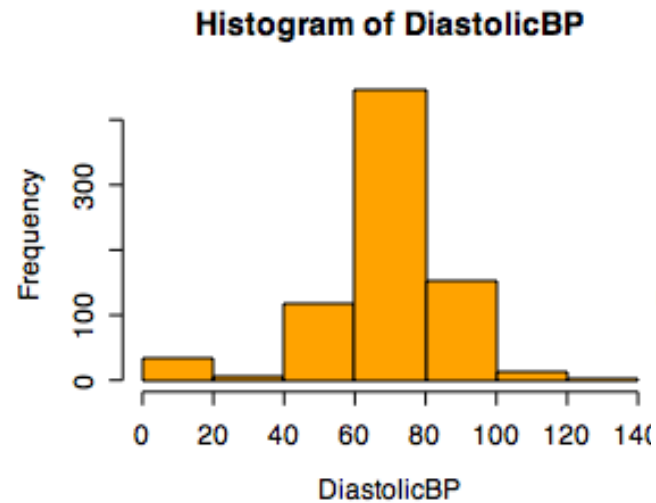
# Summary Statistics

- *not* visual
- sample statistics of data  $X$ 
  - mean:  $\mu = \sum_i X_i / n$
  - mode: most common value in  $X$
  - median:  $\mathbf{X} = \text{sort}(X)$ , median =  $\mathbf{X}_{n/2}$  (half below, half above)
  - quartiles of sorted  $\mathbf{X}$ : Q1 value =  $\mathbf{X}_{0.25n}$ , Q3 value =  $\mathbf{X}_{0.75n}$ 
    - interquartile range: value(Q3) - value(Q1)
    - range:  $\max(X) - \min(X) = \mathbf{X}_n - \mathbf{X}_1$
  - variance:  $\sigma^2 = \sum_i (X_i - \mu)^2 / n$
  - skewness:  $\sum_i (X_i - \mu)^3 / [(\sum_i (X_i - \mu)^2)^{3/2}]$ 
    - zero if symmetric; right-skewed more common (what kind of data is right skewed?)
  - number of distinct values for a variable (see `unique()` in R)



# Single Variable Visualization

- Histogram:
  - Shows center, variability, skewness, modality,
  - outliers, or strange patterns.
  - Bin width and position matter
  - Beware of real zeros

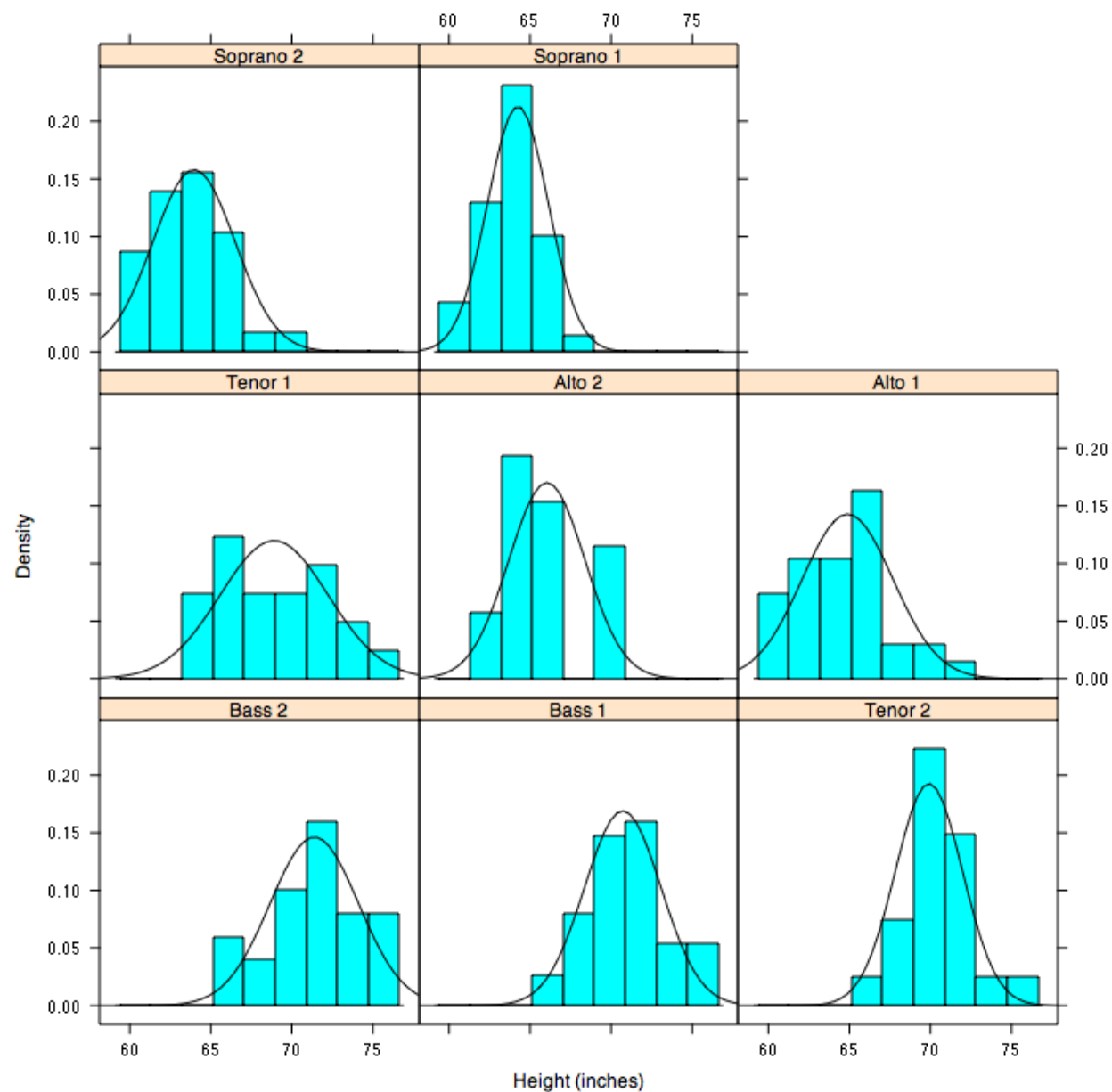




# Issues with Histograms

- For small data sets, histograms can be misleading
  - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution
- Histograms effectively only work with 1 variable at a time
  - But 'small multiples' can be effective

But be careful with  
axes and scales!



# Smoothed Histograms - Density Estimates

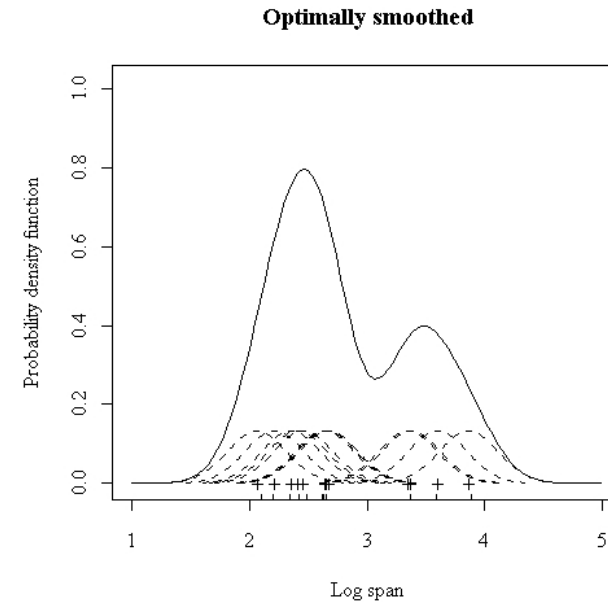
- Kernel estimates smooth out the contribution of each datapoint over a local neighborhood of that point

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

$h$  is the kernel width

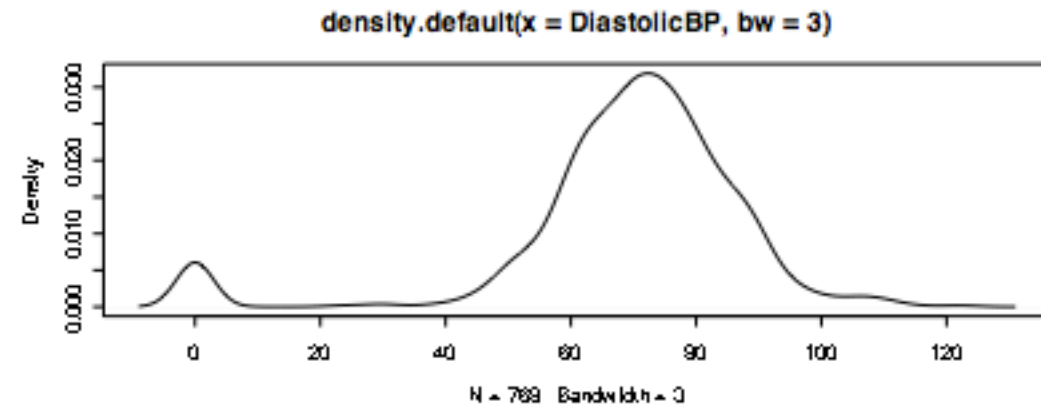
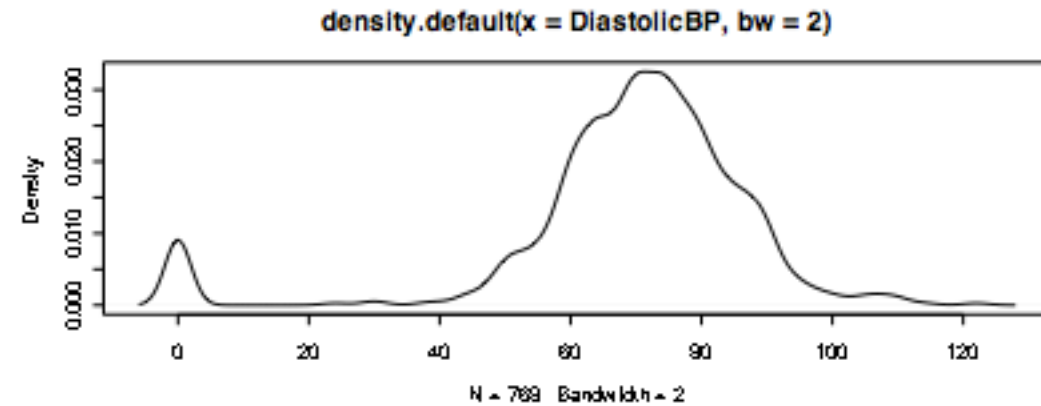
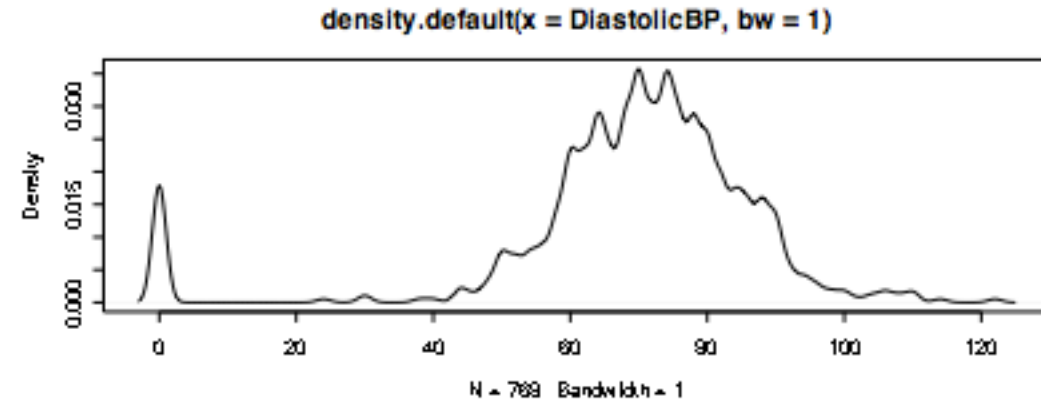
- Gaussian kernel is common:

$$Ce^{-\frac{1}{2}\left(\frac{x-x(i)}{h}\right)^2}$$



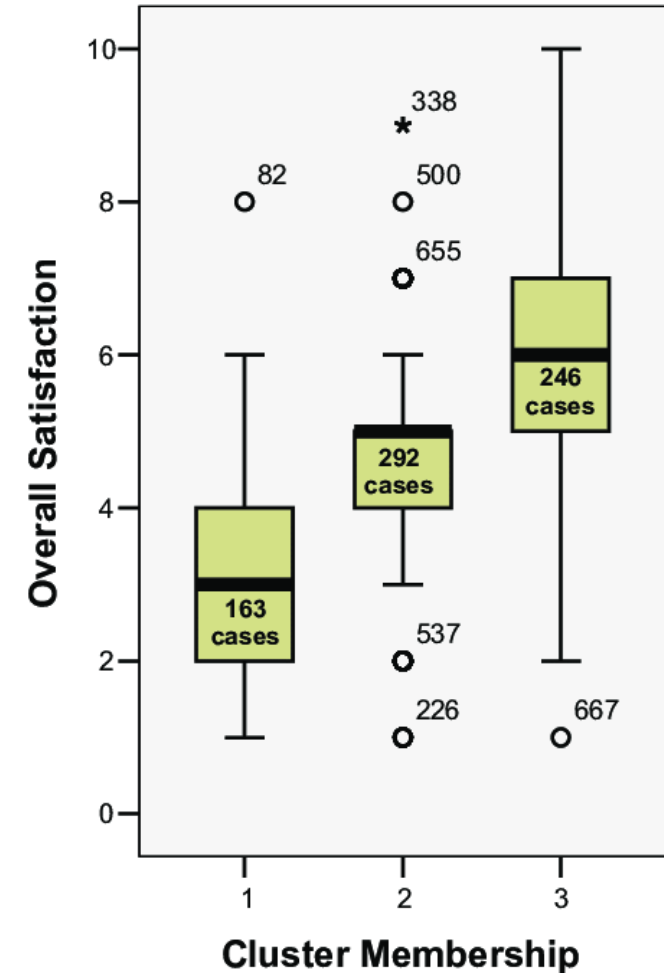
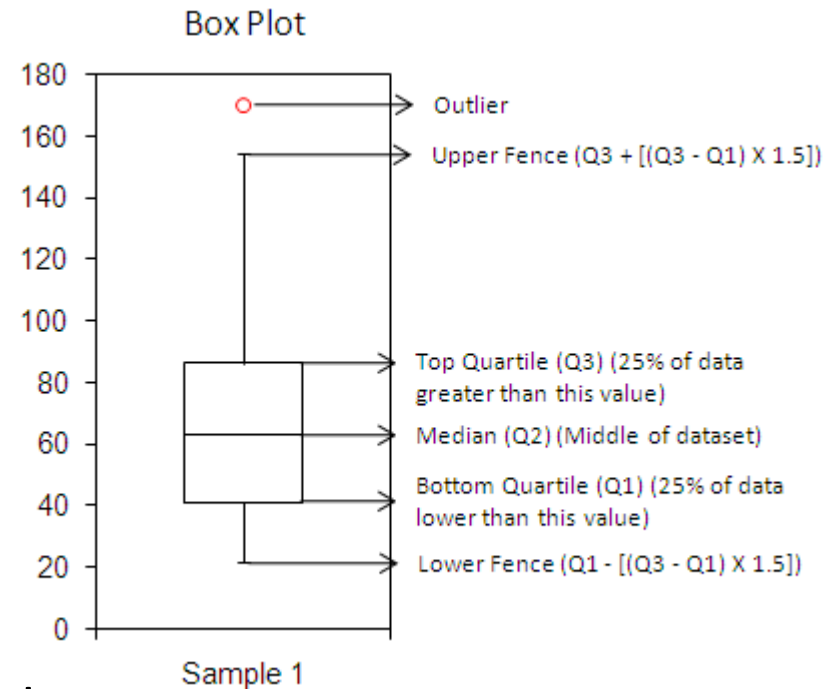
Bandwidth  
choice is an art

Usually want to  
try several



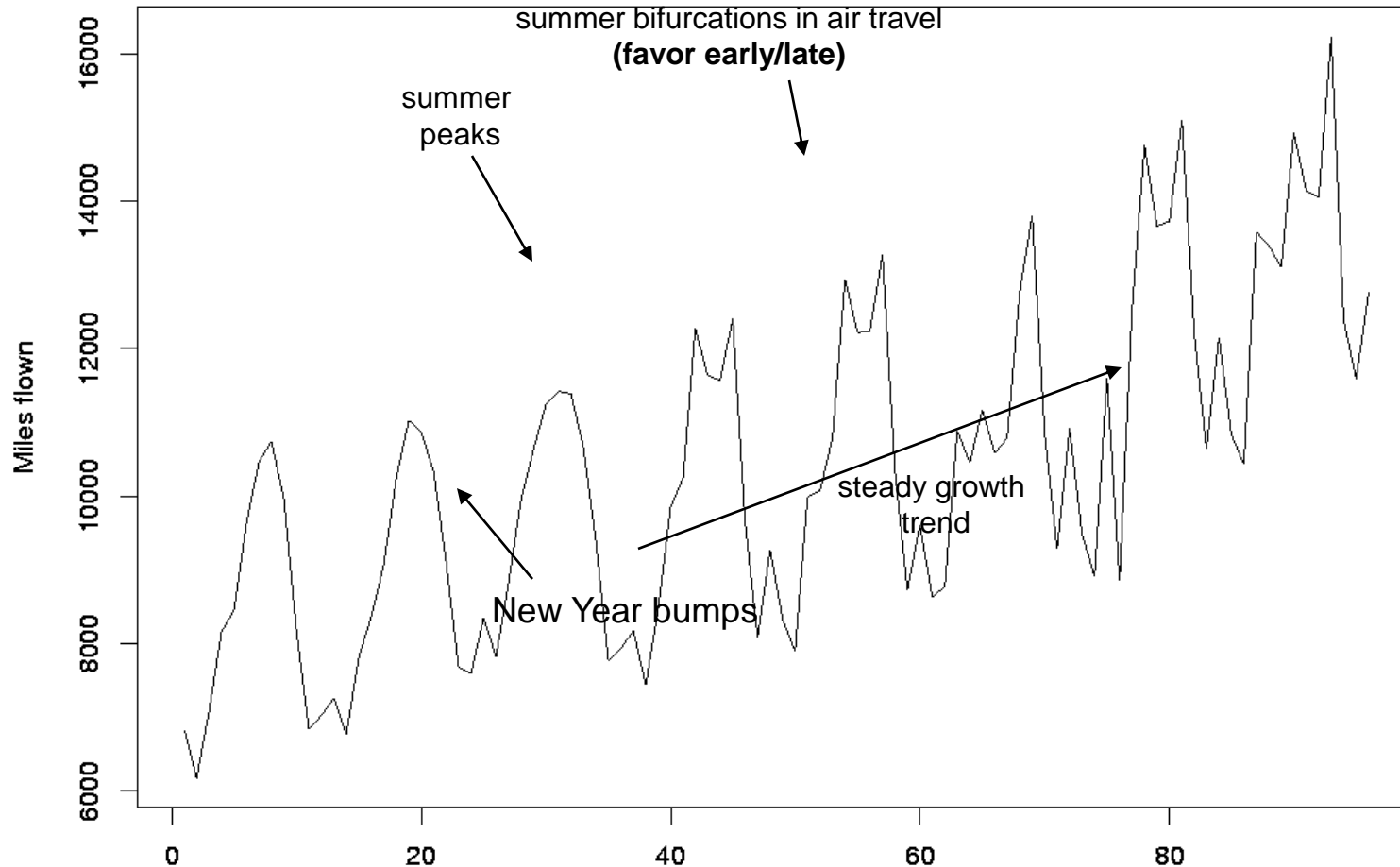
# Boxplots

- Shows a lot of information about a variable in one plot
  - Median
  - Inter Quartile Range (IQR)
  - Outliers
  - Range
  - Skewness
- Negatives
  - Overplotting
  - Hard to tell distributional shape
  - no standard implementation in software (many options for whiskers, outliers)



# Time Series

If your data has a temporal component, be sure to exploit it



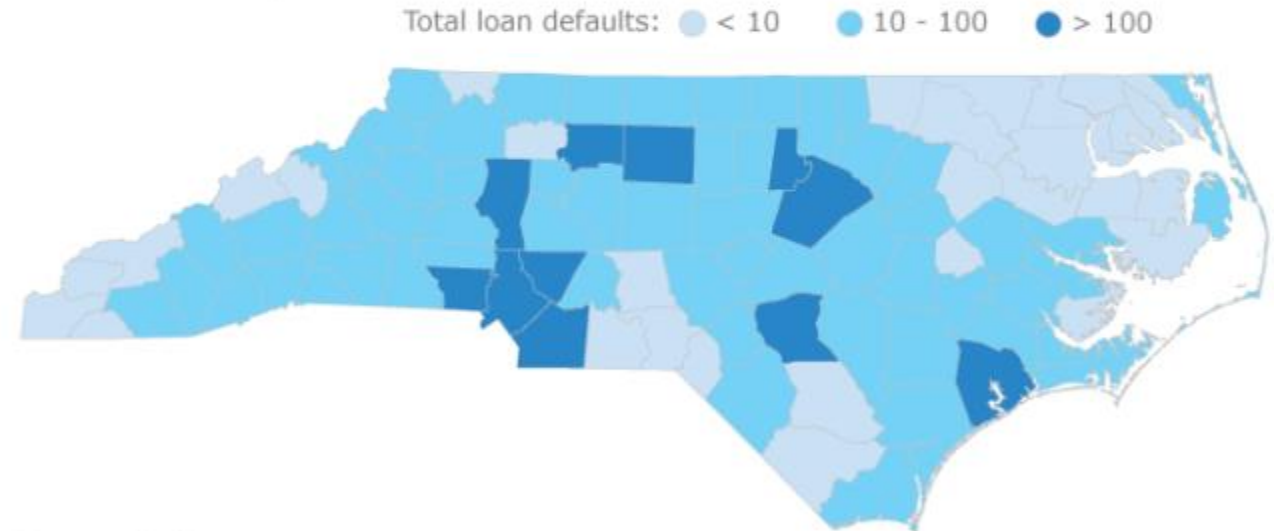
**Visually reveals  
pattern not  
apparent from  
raw data table**

# Spatial Data

- If your data has a geographic component, be sure to exploit it
- Data from cities/states/zip cods – easy to get lat/long
- Can plot as scatterplot

## Visual aid

Banks are using mapping software to help with such decisions as where to open and close branches and how to serve low-income communities. One unnamed bank used this map to track loan defaults in each North Carolina county

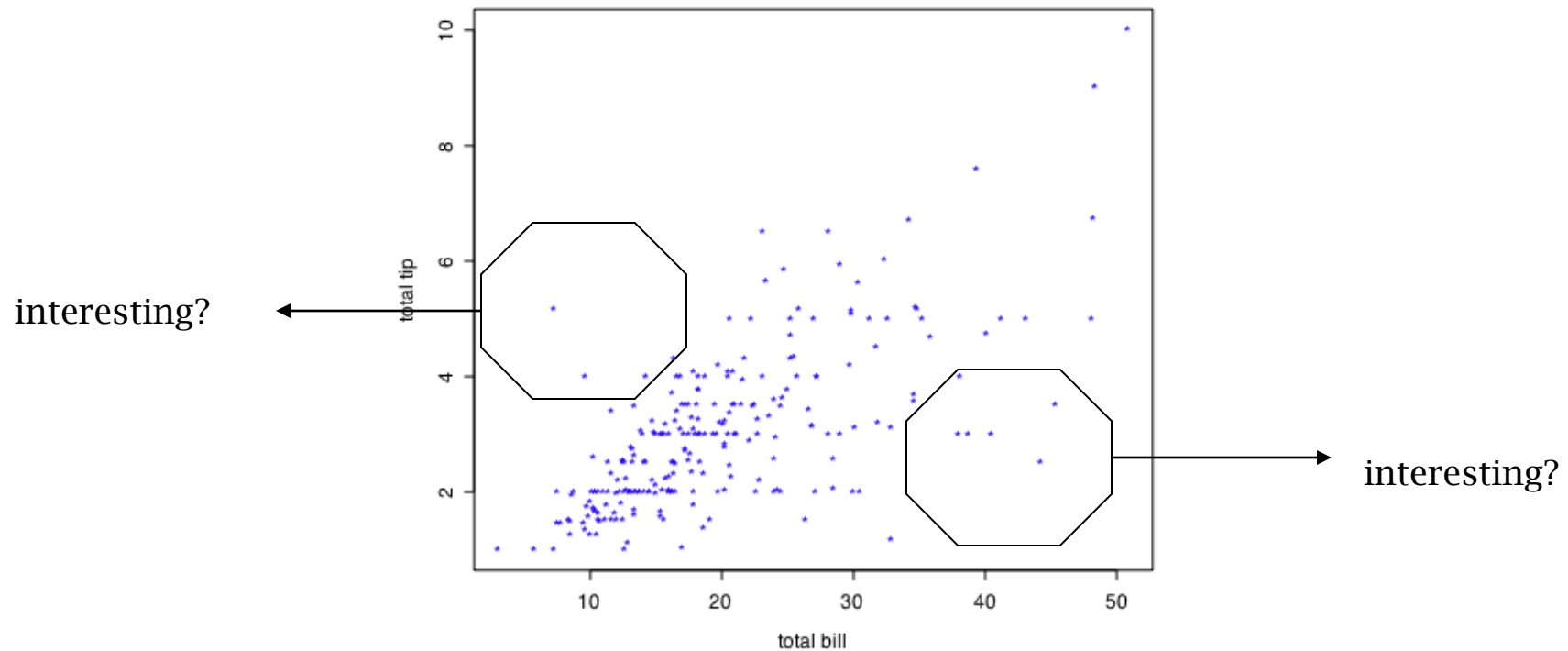


Source: Esri Inc.

- When Bank of America needed to identify current and future locations of its banking centres in low-income neighbourhoods, it turned to mapping and geographic analysis application
- Central Bank of Nigeria relied heavily on GIS mapping to mark the financial access points of the entire country

# Two Continuous Variables

- For two numeric variables, the scatterplot is the obvious choice

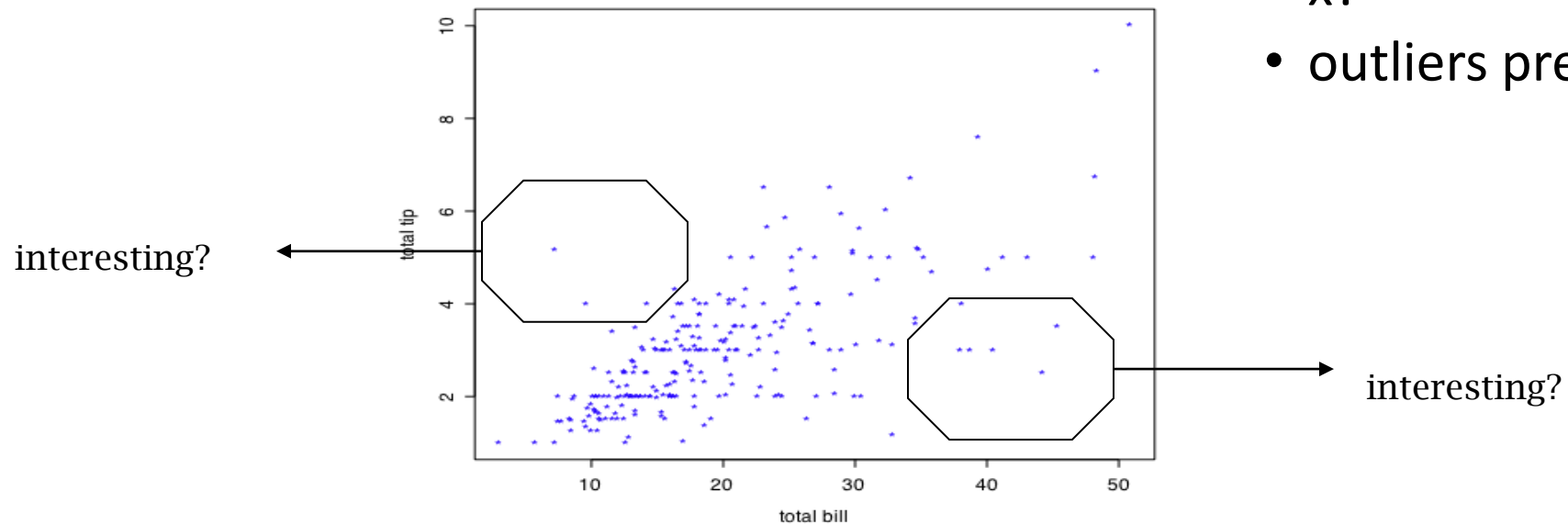




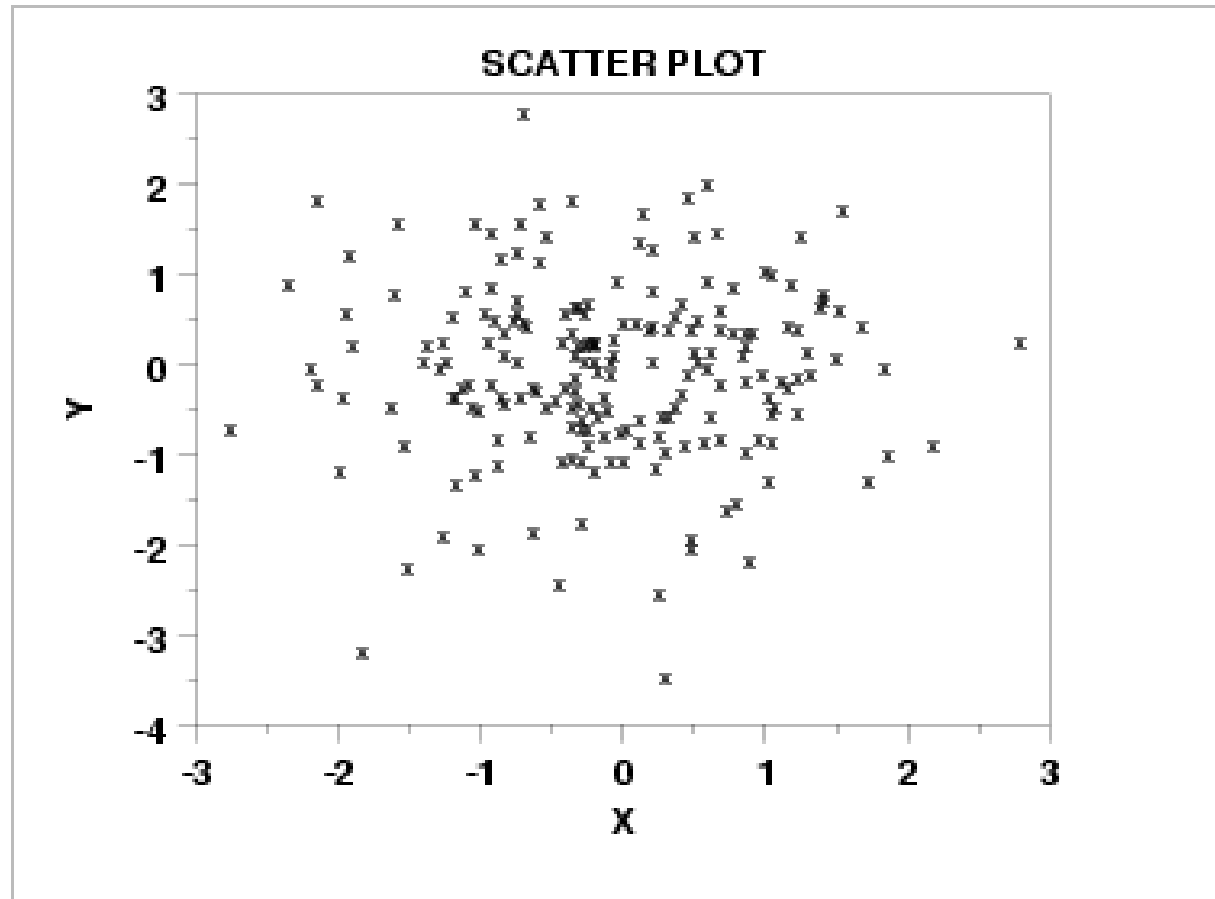
# 2D Scatterplots

- standard tool to display relation between 2 variables
  - e.g. y-axis = response, x-axis = suspected indicator

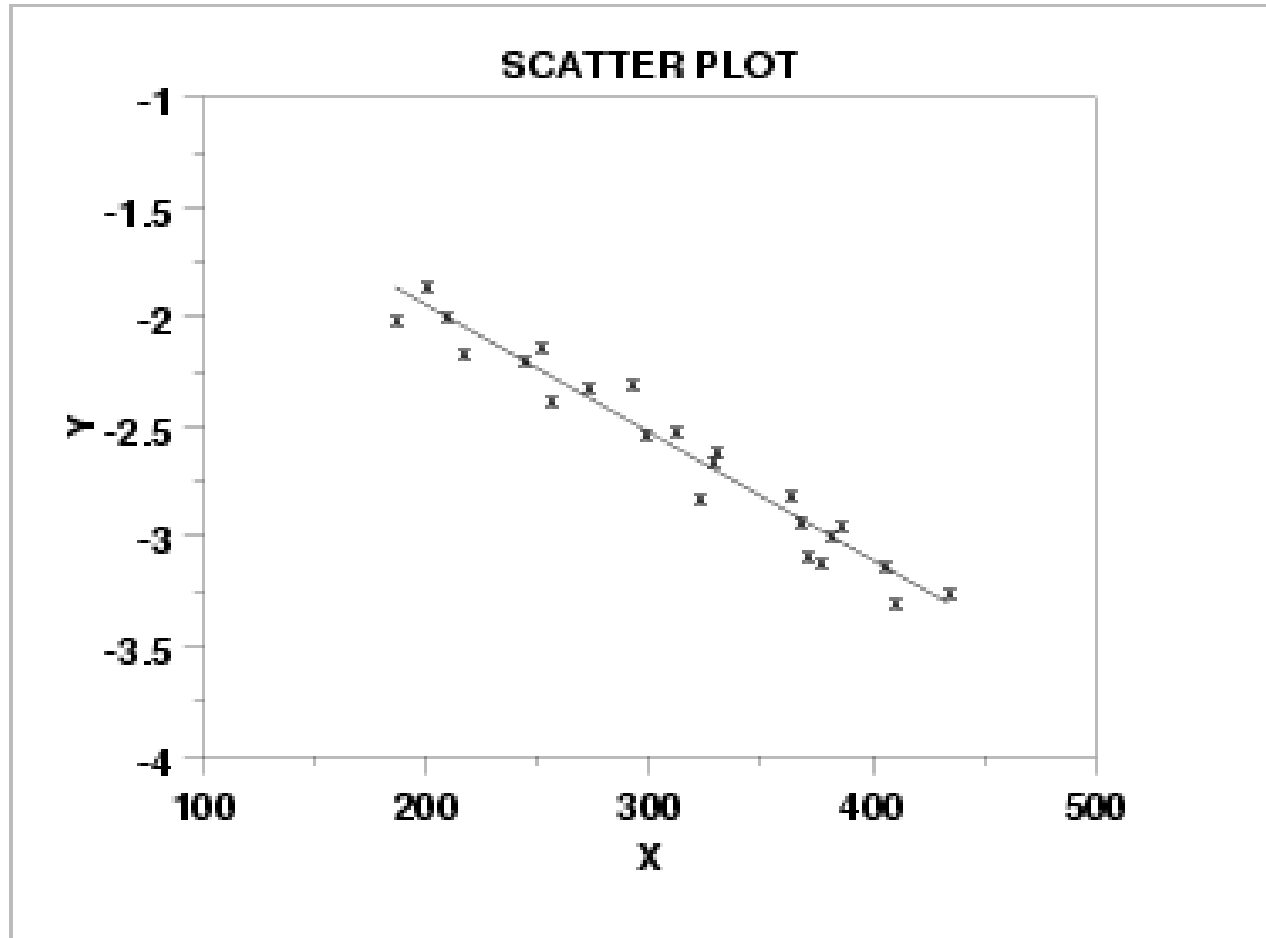
- useful to answer:
  - x,y related?
    - linear
    - quadratic
    - other
  - variance(y) depend on x?
  - outliers present?



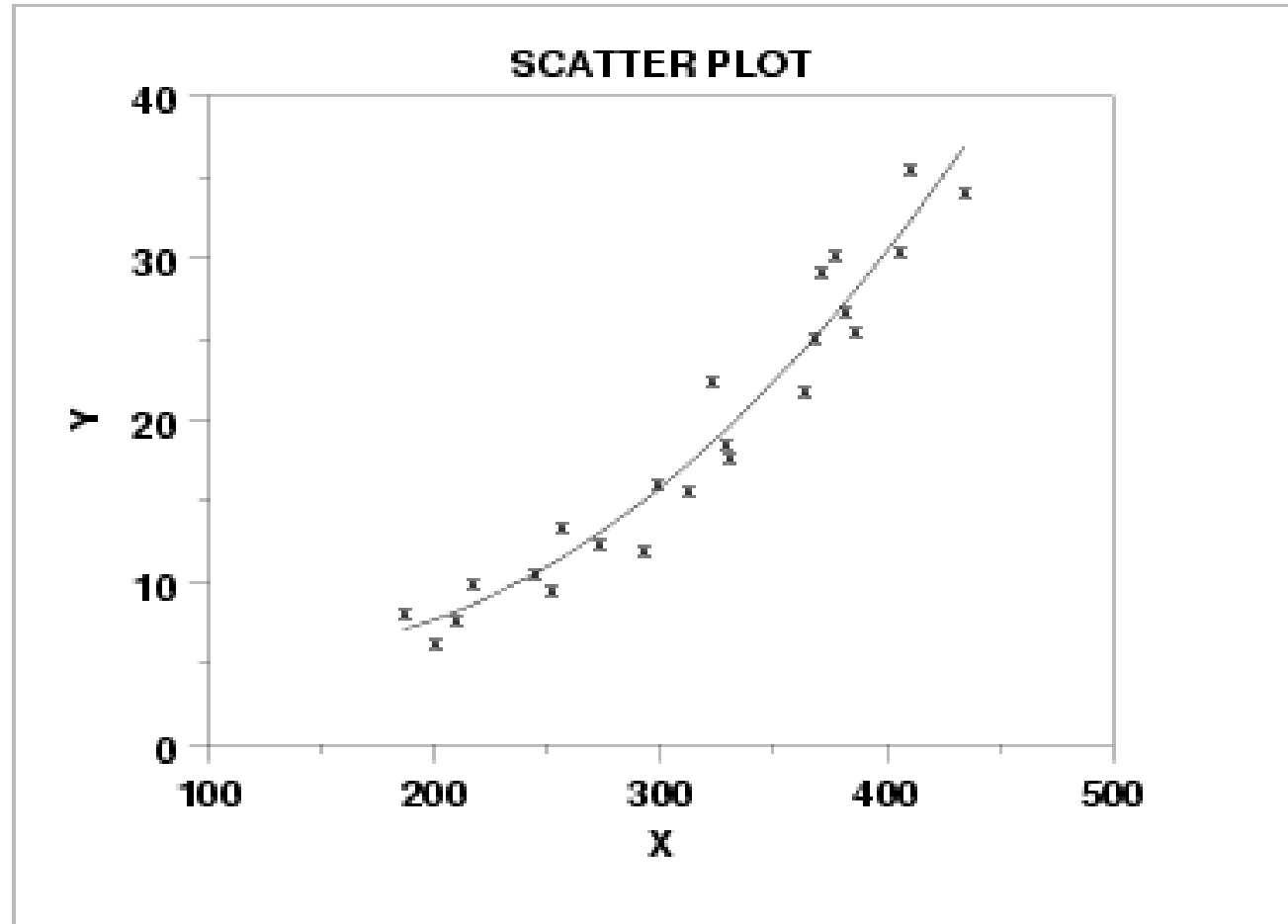
# Scatter Plot: No apparent relationship



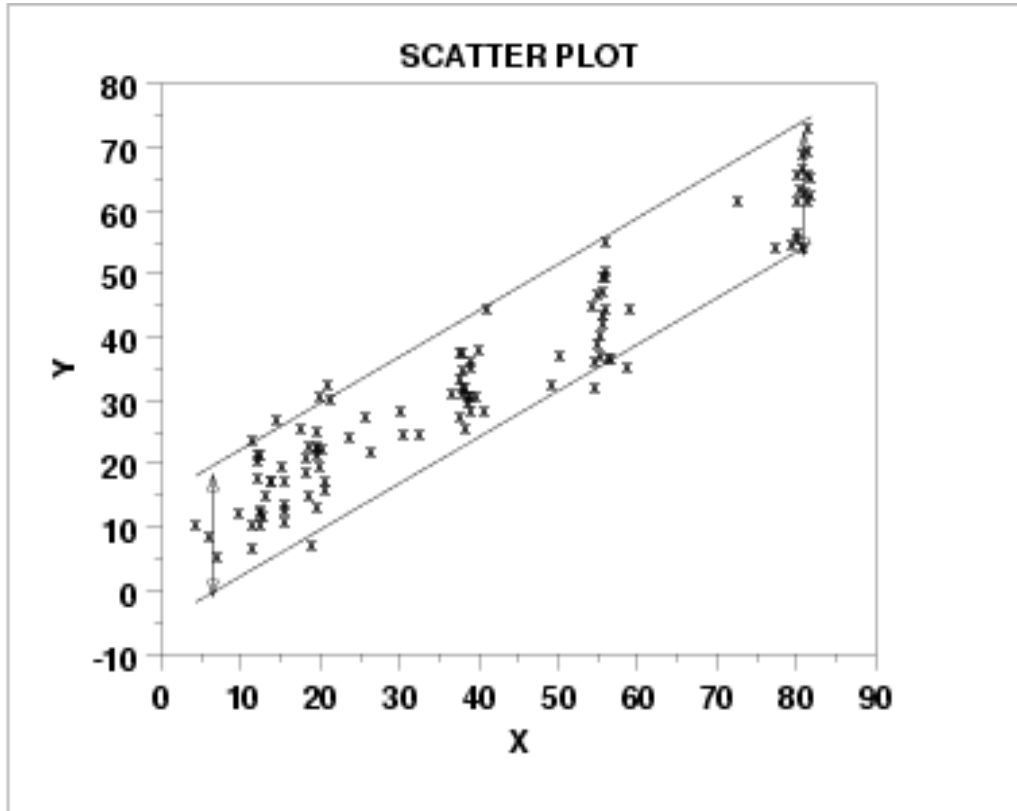
# Scatter Plot: Linear relationship



# Scatter Plot: Quadratic relationship



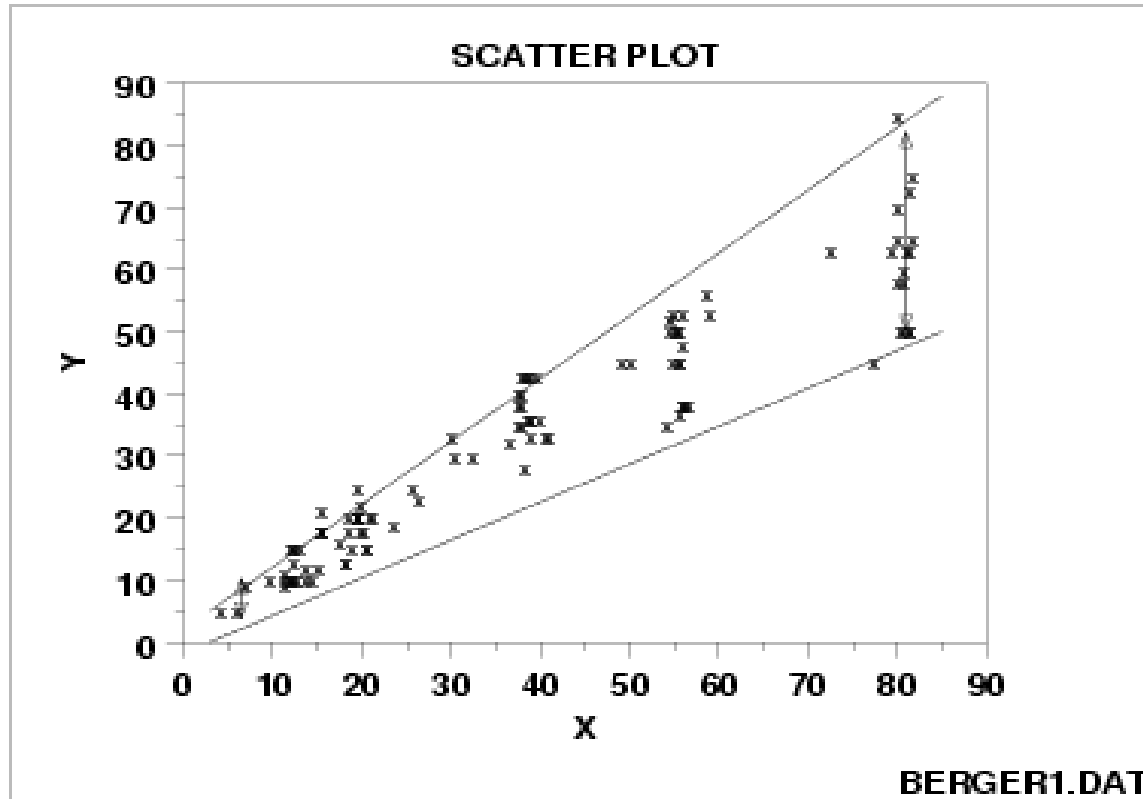
# Scatter plot: Homoscedastic



Why is this important in classical statistical modelling?

- a situation in which the error term (that is, the “noise” or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables

# Scatter plot: Heteroscedastic



**variation in  $Y$  differs depending on the value of  $X$**   
***e.g.,  $Y = \text{annual tax paid}$ ,  $X = \text{income}$***

# Two variables - continuous

- Scatterplots
  - But can be bad with lots of data

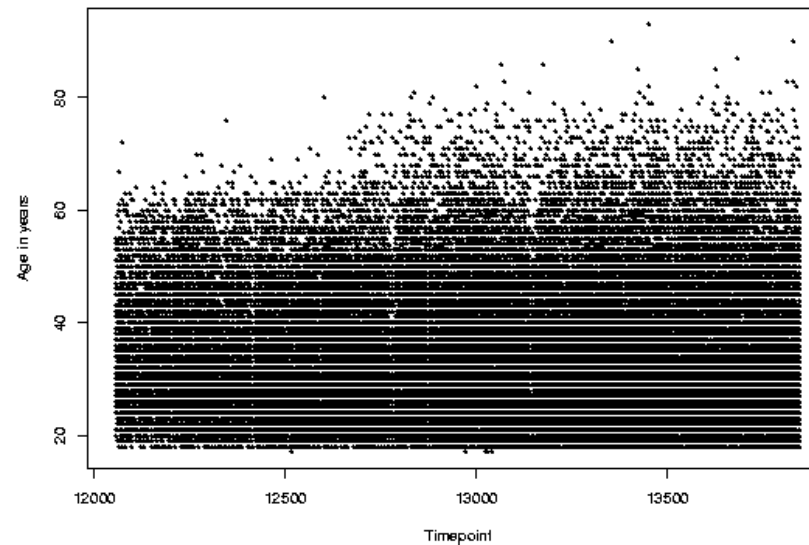
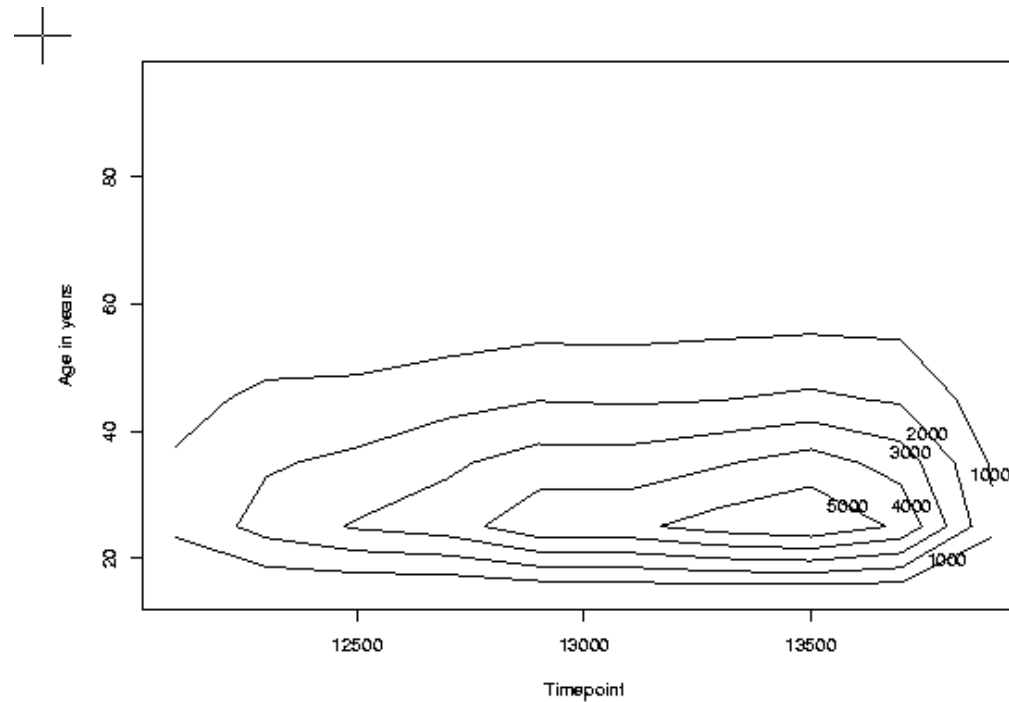


Figure 3.7: A scatterplot of 96,000 cases, with much overprinting. Each data point represents an individual applicant for a loan. The vertical axis shows the age of the applicant, and the horizontal axis indicates the day on which the application was made.

# Two variables - continuous

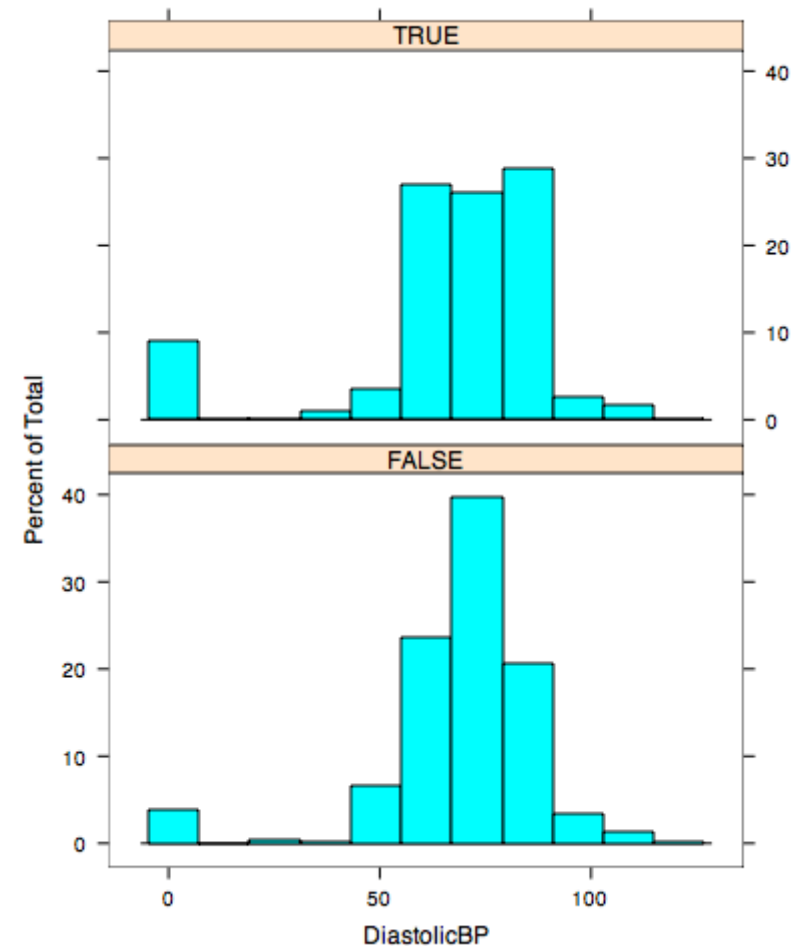
- What to do for large data sets
  - Contour plots





# Displaying Two Variables

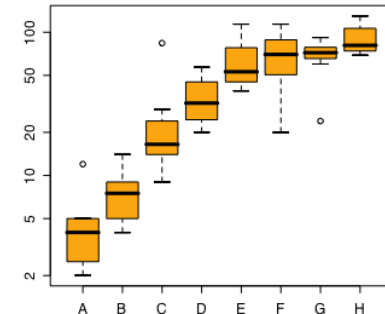
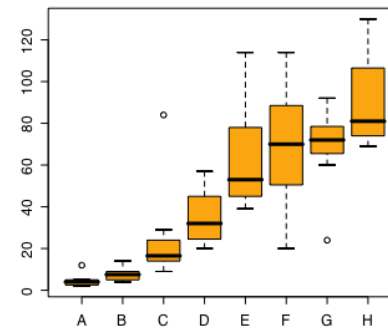
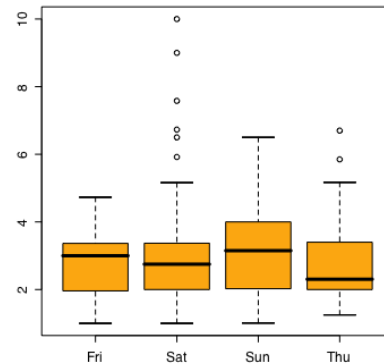
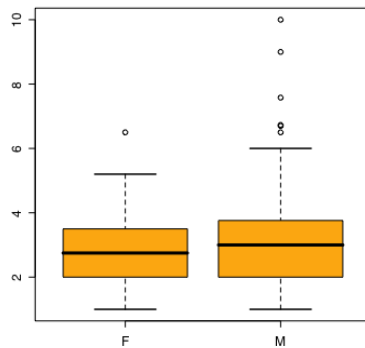
- If one variable is categorical, use small multiples
- Software packages have this implemented as 'lattice' or 'trellis' packages



```
library('lattice')  
histogram(~DiastolicBP | TimesPregnant==0)
```

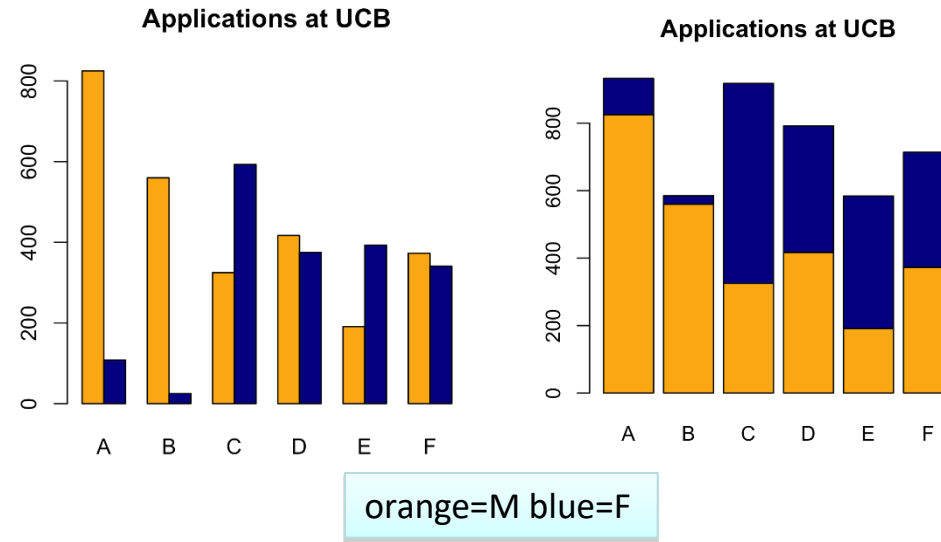
# Two Variables - one categorical

- Side by side boxplots are very effective in showing differences in a quantitative variable across factor levels
  - tips data
    - do men or women tip better
  - orchard sprays
    - measuring potency of various orchard sprays in repelling honeybees

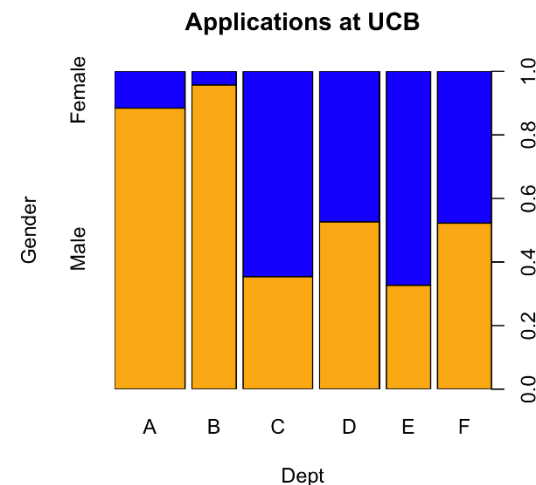


# Barcharts and Spineplots

stacked barcharts  
can be used to  
compare continuous  
values across two or  
more categorical  
ones.



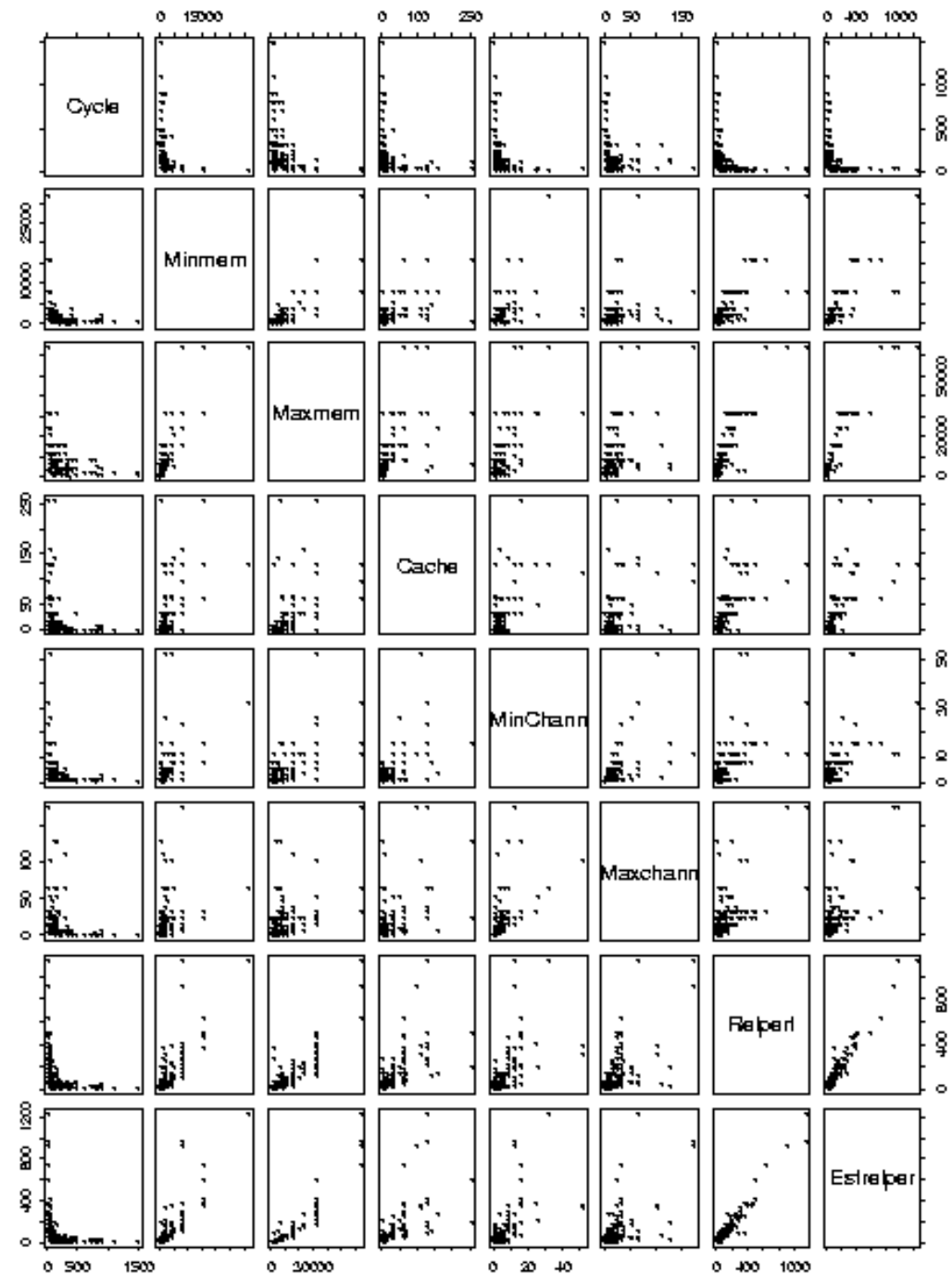
spineplots show  
proportions well, but  
can be hard to  
interpret

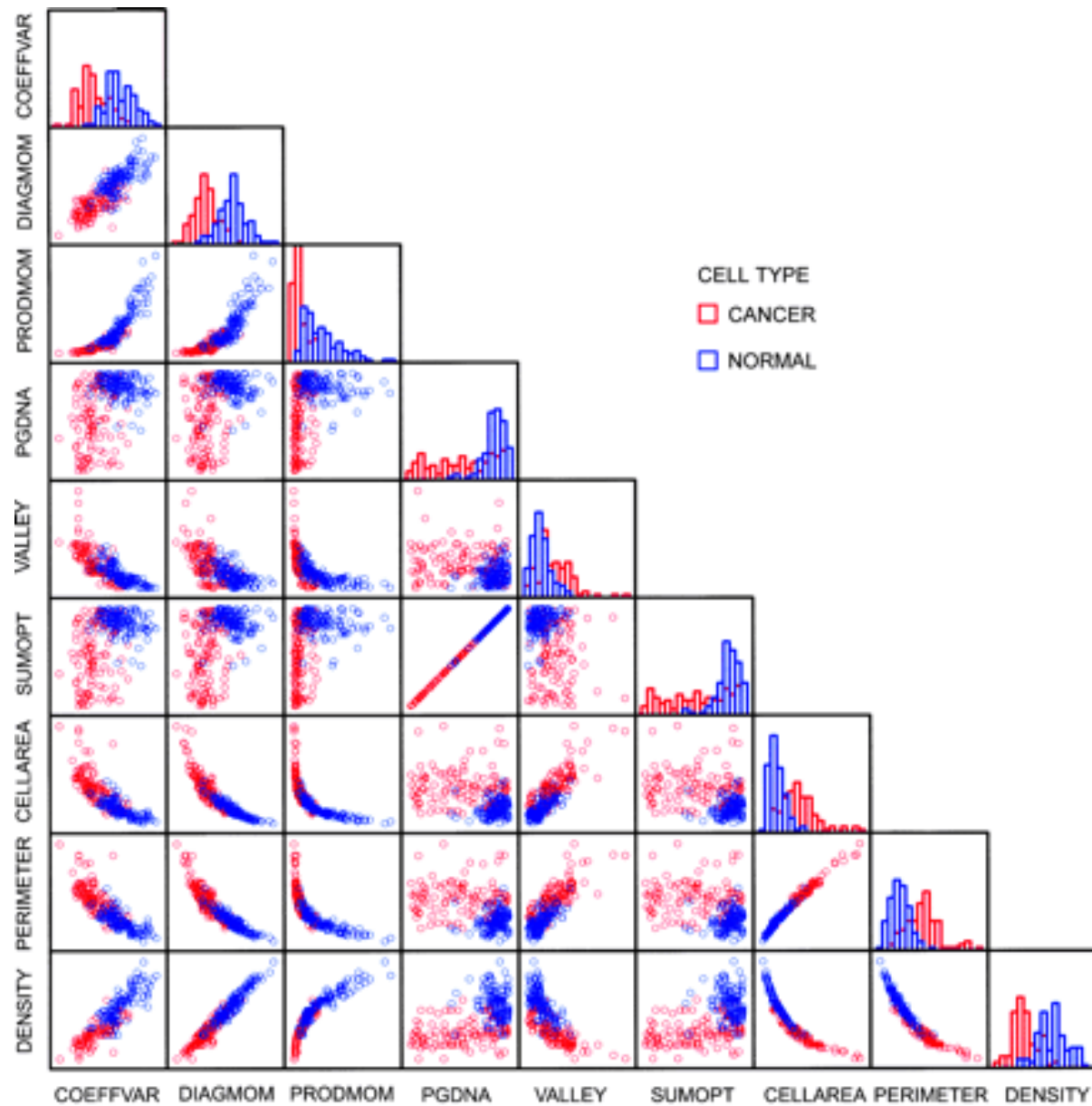


# More than two variables

Pairwise scatterplots

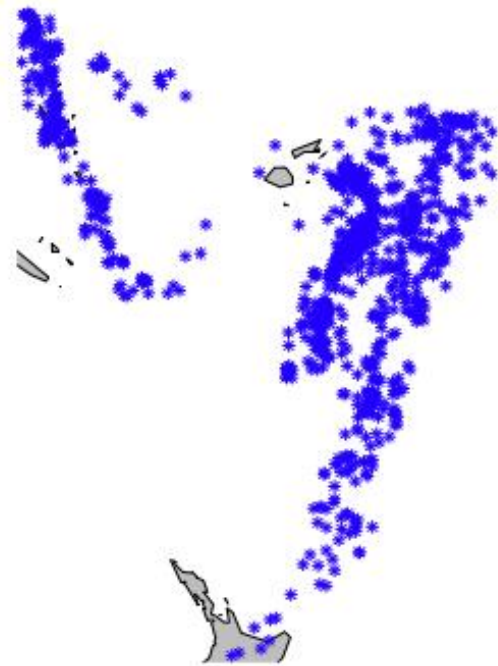
Can be somewhat ineffective for categorical data



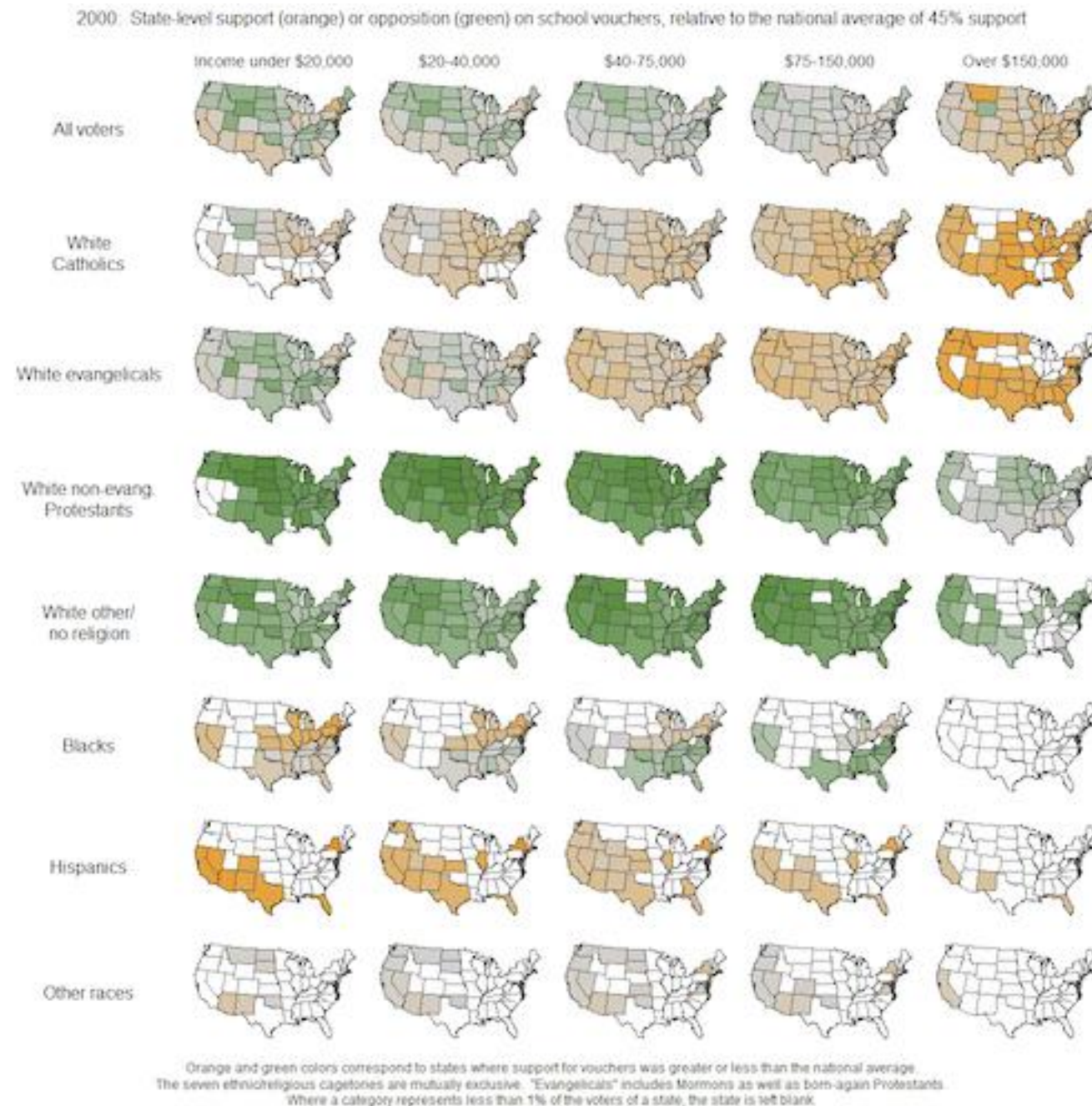


# Multivariate: More than two variables

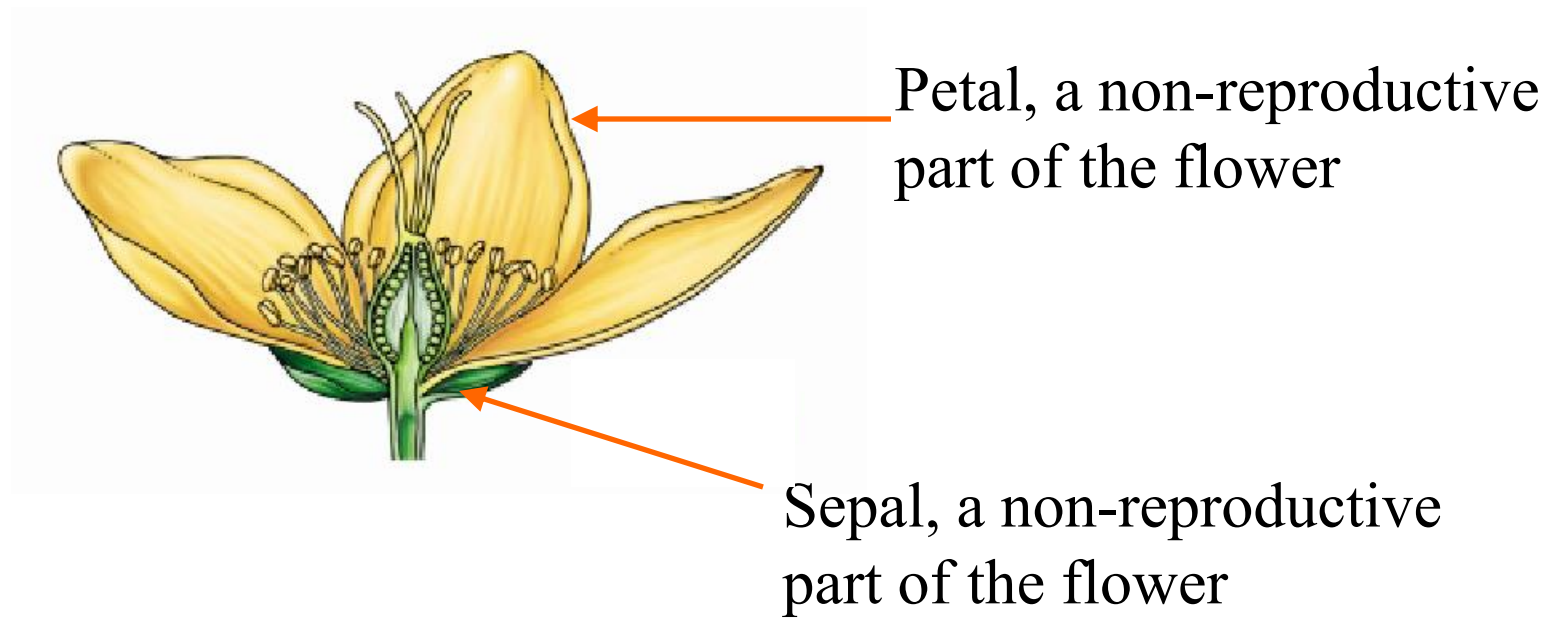
- Get creative!
- Conditioning on variables
  - trellis or lattice plots
  - Cleveland models on human perception, all based on conditioning
  - Infinite possibilities
- Earthquake data:
  - locations of 1000 seismic events of MB > 4.0. The events occurred in a cube near Fiji since 1964
  - Data collected on the severity of the earthquake



How many  
dimensions  
are  
represented  
here?



# Multivariate: Parallel Coordinates






# Parallel Coordinates

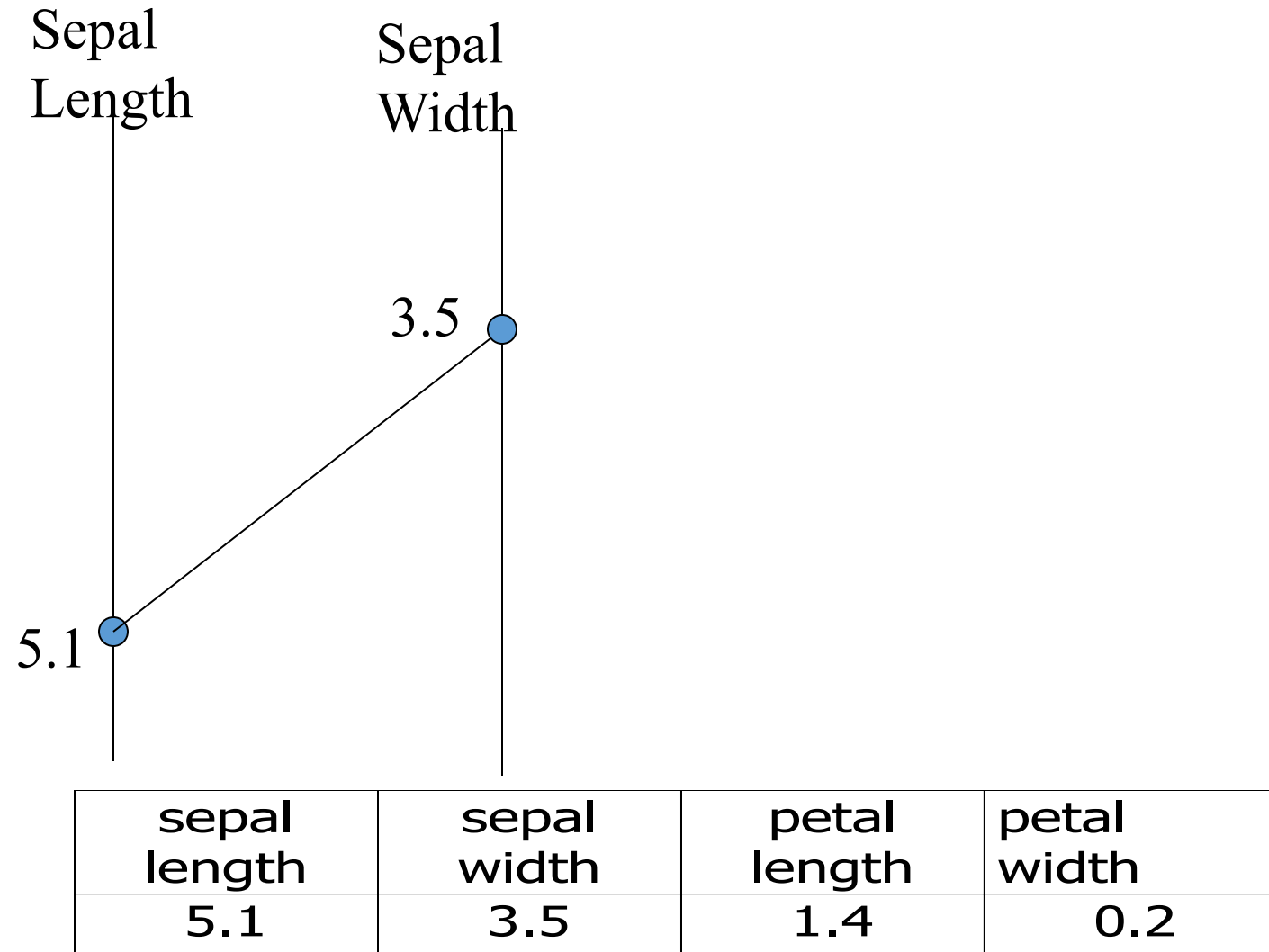
Sepal  
Length

5.1

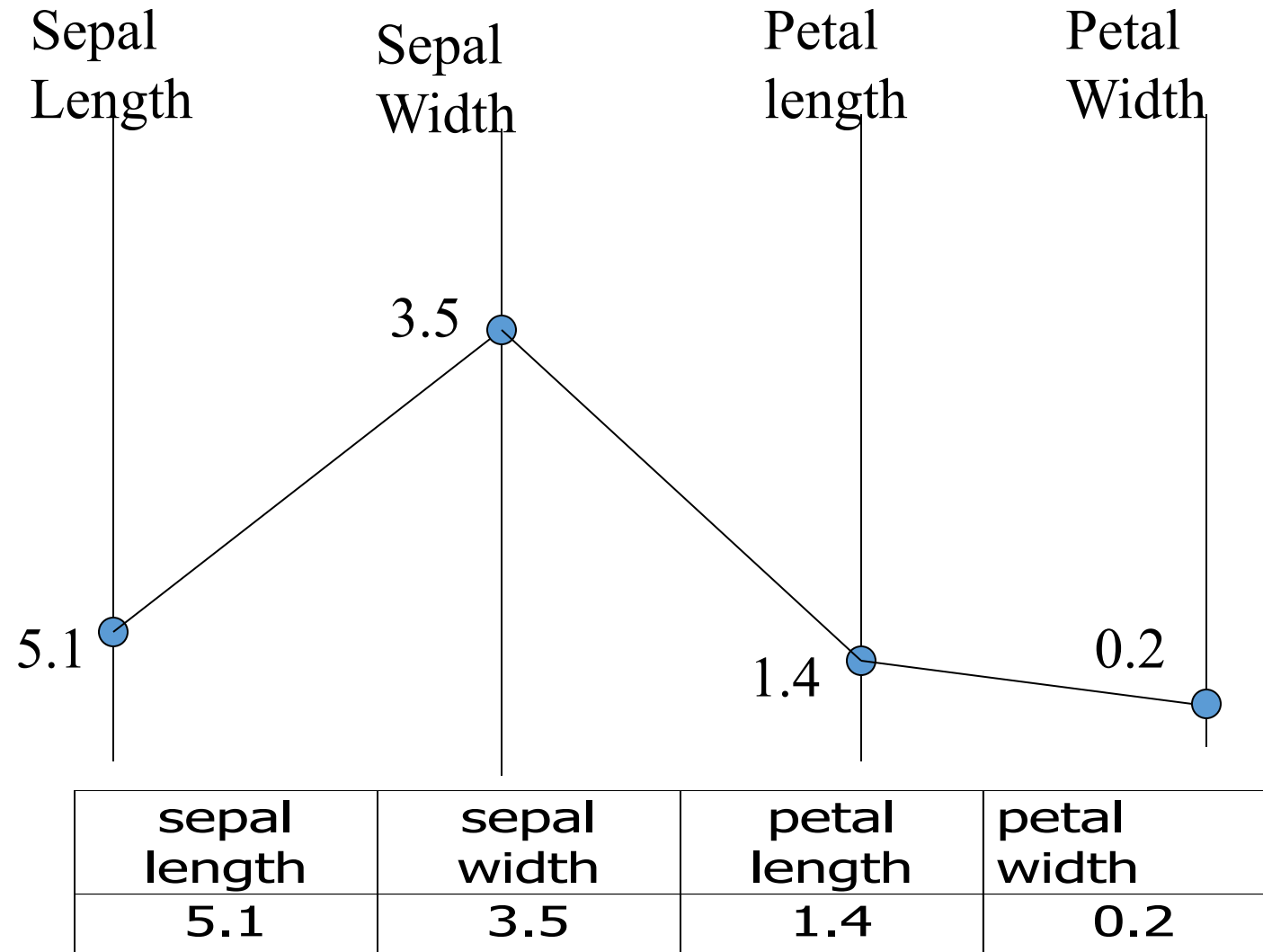


sepal length	sepal width	petal length	petal width
5.1	3.5	1.4	0.2

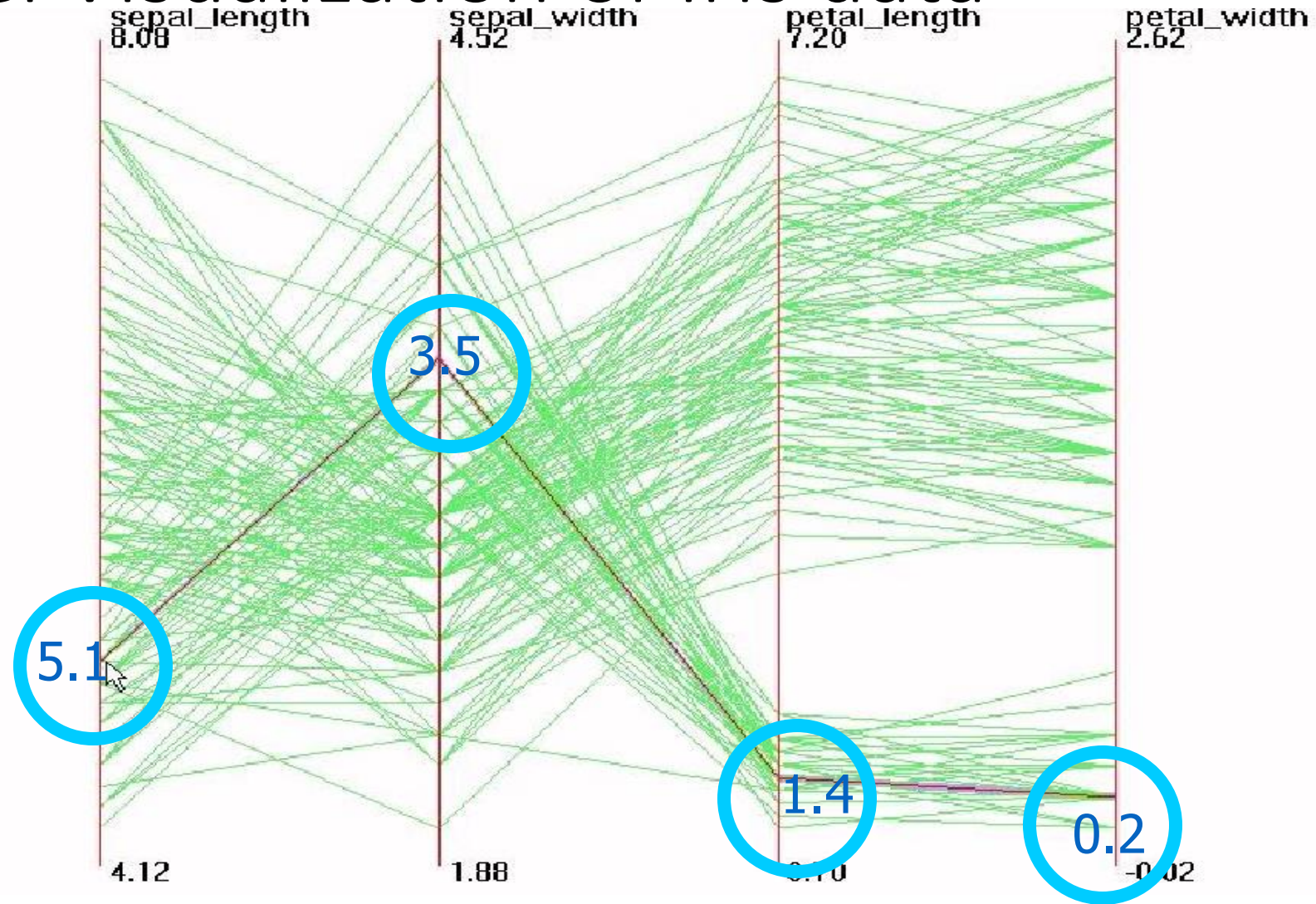
# Parallel Coordinates: 2 D



# Parallel Coordinates: 4 D

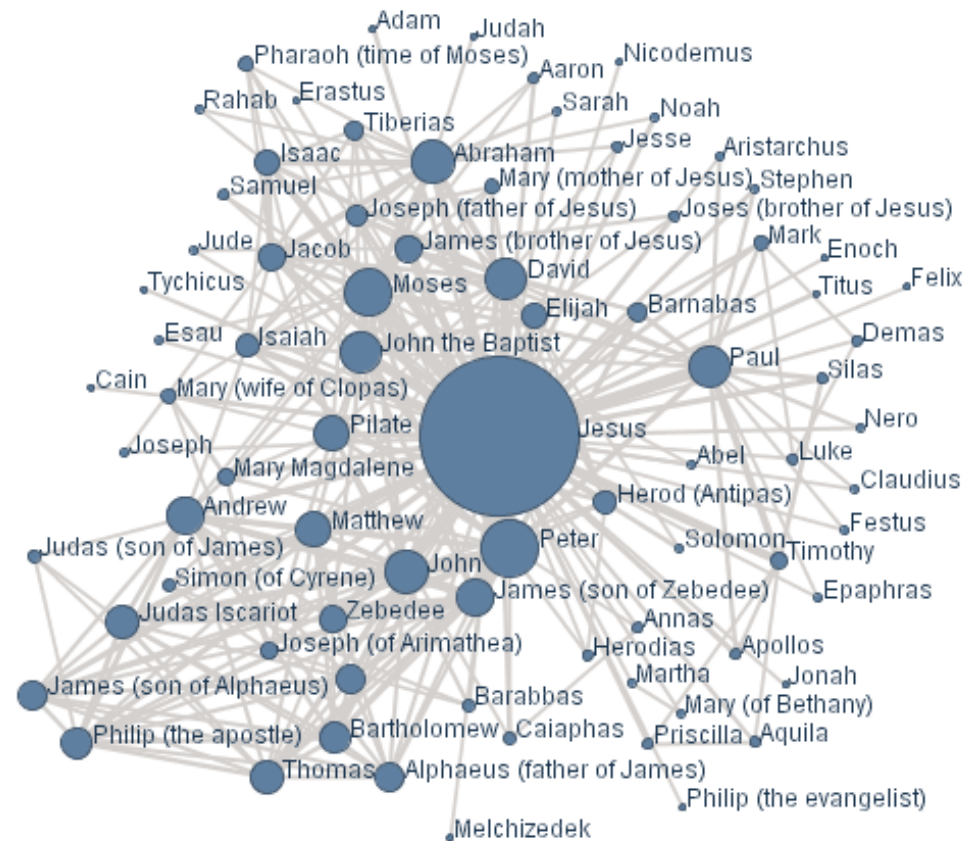


# Parallel Visualization of Iris data



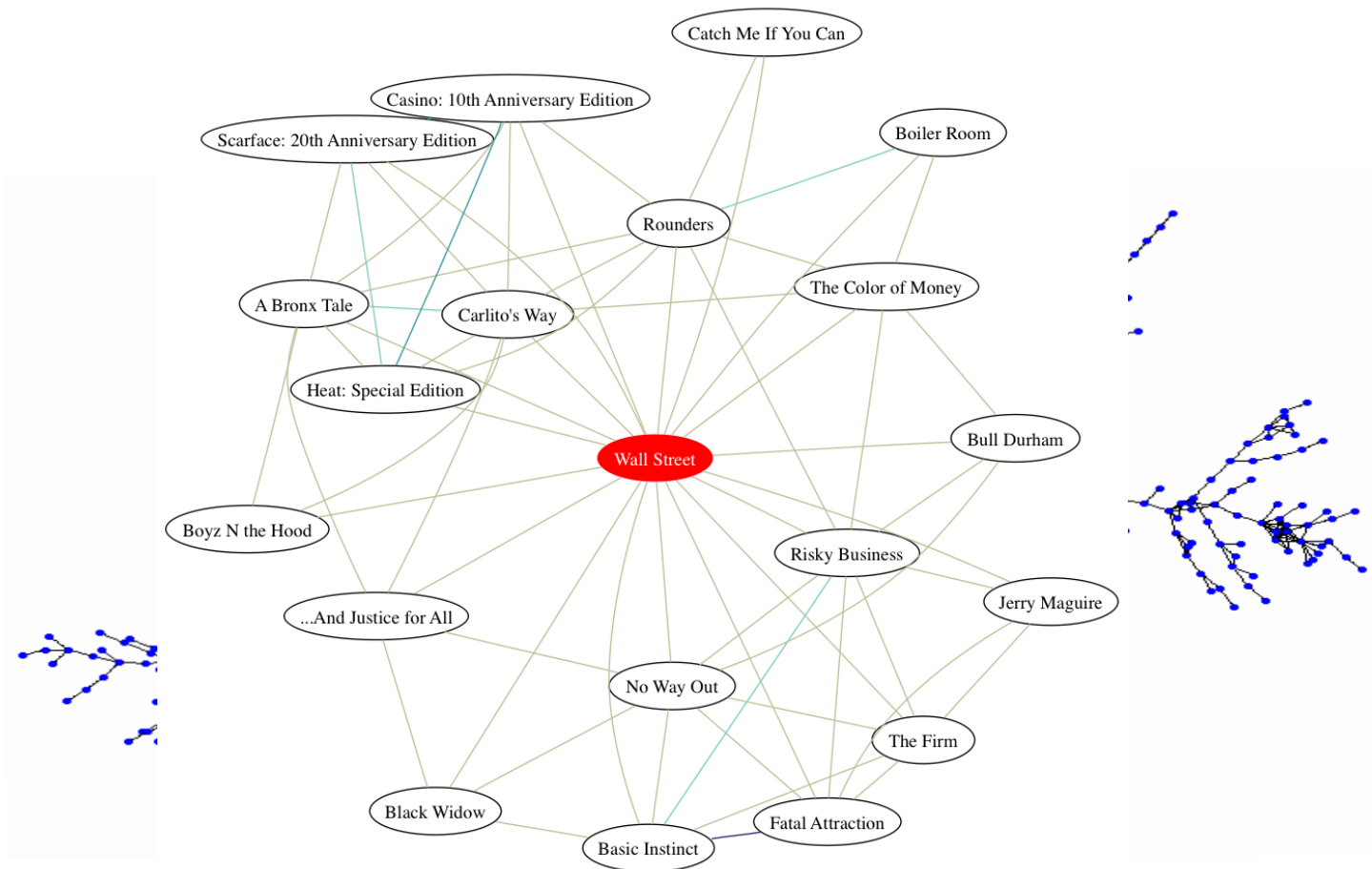
# Networks and Graphs

- Visualizing networks is helpful, even if it is not obvious that a network exists



# Network Visualization

- Graphviz (open source software) is a nice layout tool for big and small graphs



# Other Charts

- Pie charts
  - very popular
  - good for showing simple relations of proportions
  - human perception not good at comparing arcs
  - barplots, histograms usually better (but less pretty)
- 3D
  - nice to be able to show three dimensions
  - hard to do well
  - often done poorly

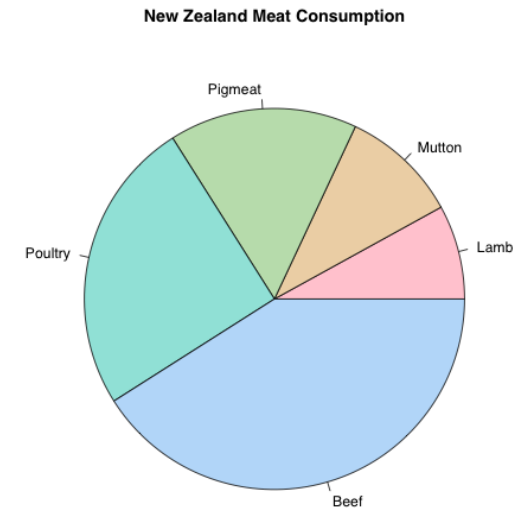
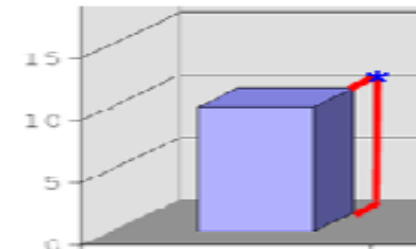
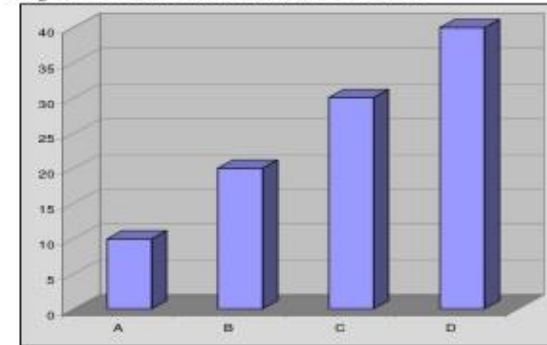


Figure 1. Three-dimensional bar chart.



# Dimension Reduction

- One way to visualize high dimensional data is to reduce it to 2 or 3 dimensions
  - Variable selection
    - e.g. stepwise
  - Principle Components
    - find linear projection onto p-space with maximal variance
  - Multi-dimensional scaling
    - takes a matrix of (dis)similarities and embeds the points in p-dimensional space to retain those similarities

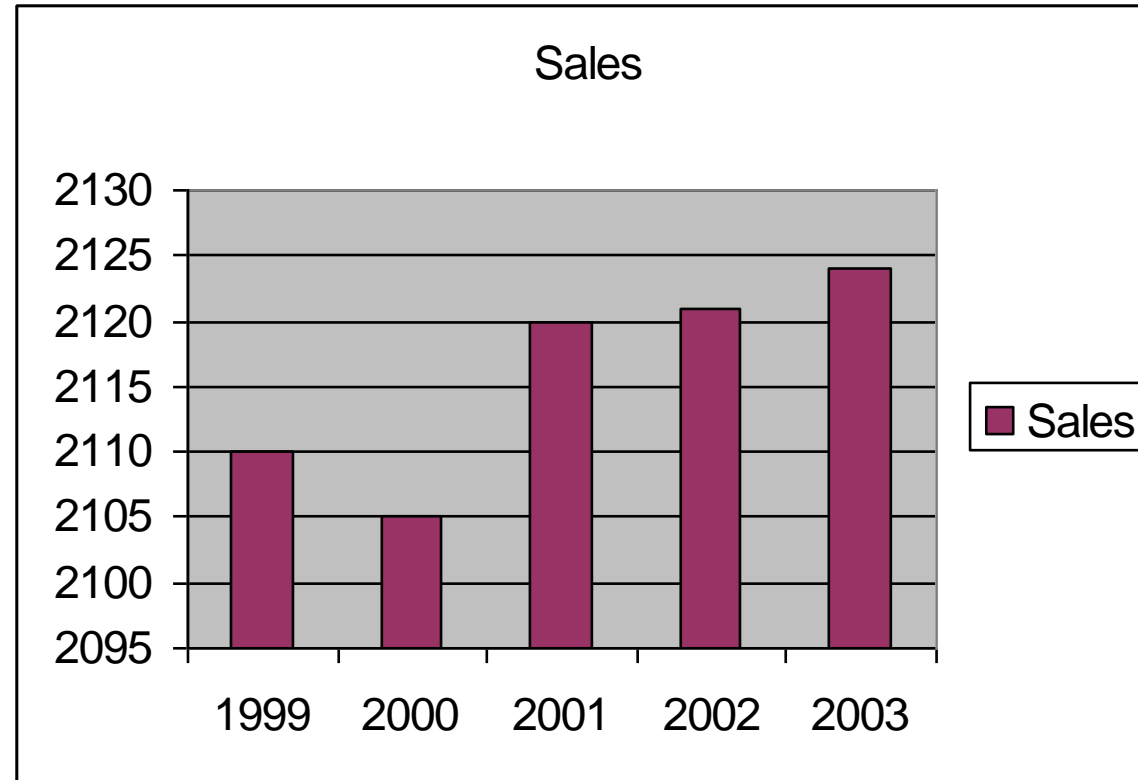


# When to use which type?

- Line graph
  - x-axis requires quantitative variable
  - Variables have contiguous values
  - familiar/conventional ordering among ordinals
- Bar graph
  - comparison of relative point values
- Scatter plot
  - convey overall impression of relationship between two variables
- Pie Chart?
  - Emphasizing differences in proportion among a few numbers

# Bad Visualization: Spreadsheet

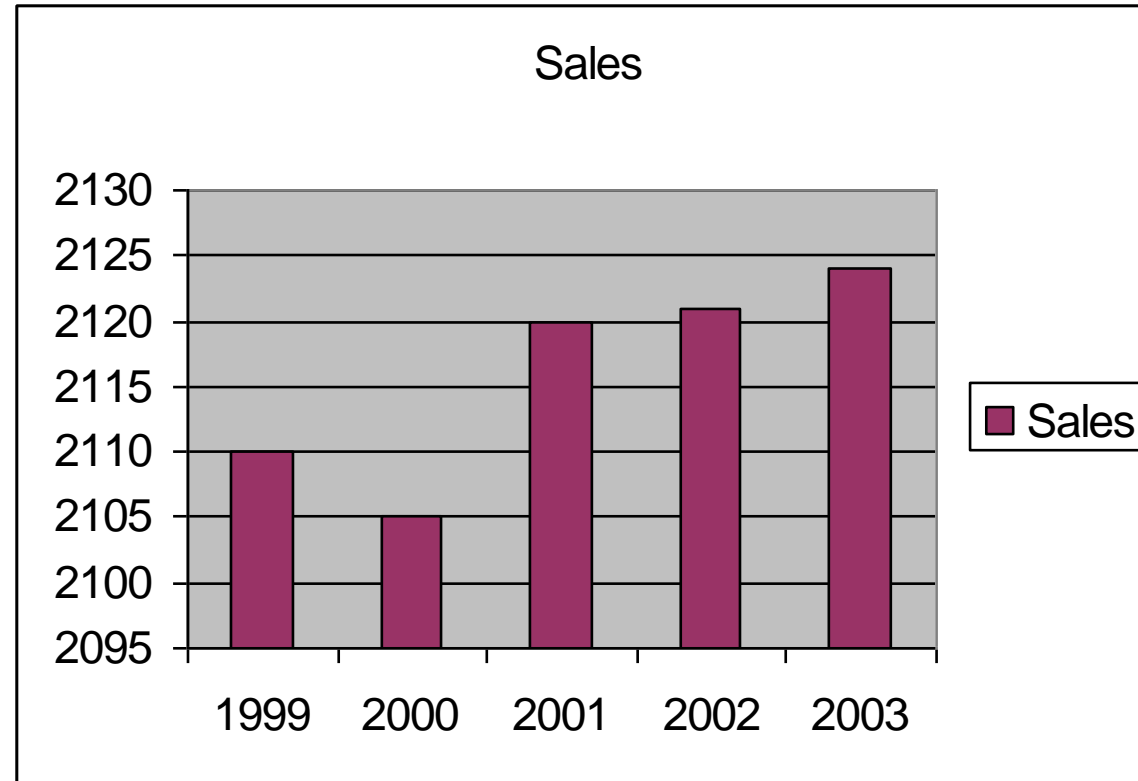
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



What is wrong with this graph?

# Bad Visualization: Spreadsheet with misleading Y –axis

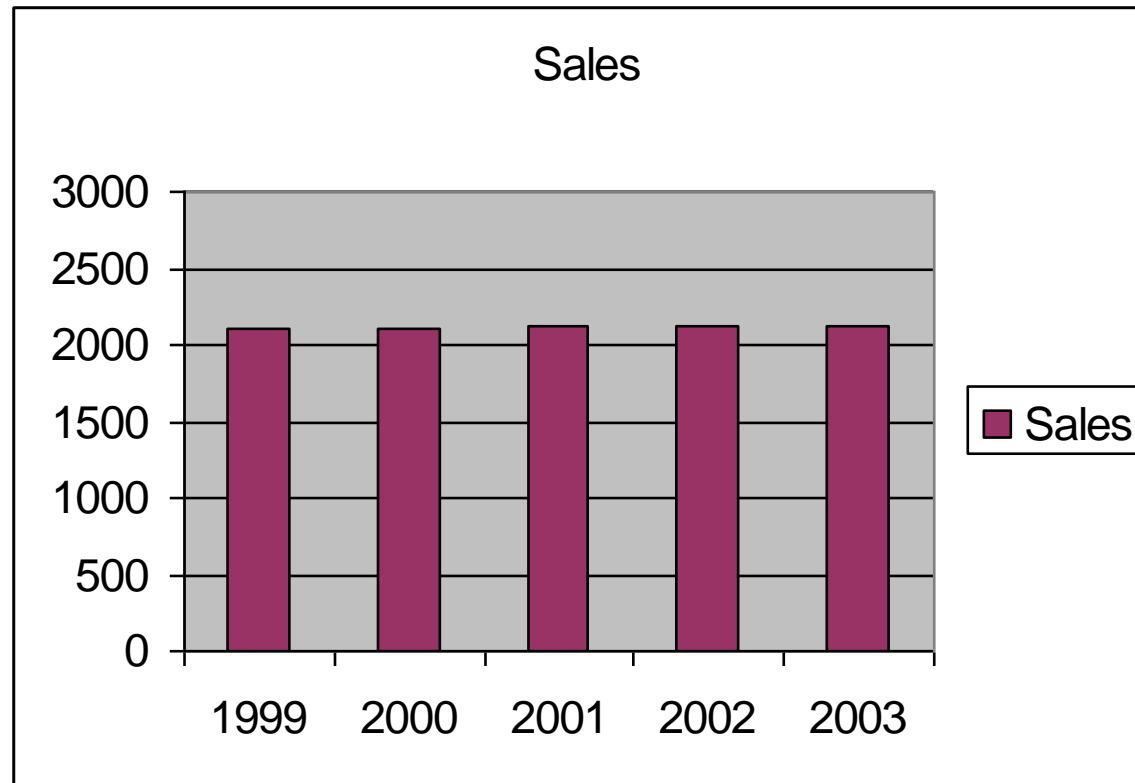
Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



Y-Axis scale gives **WRONG**  
impression of big change

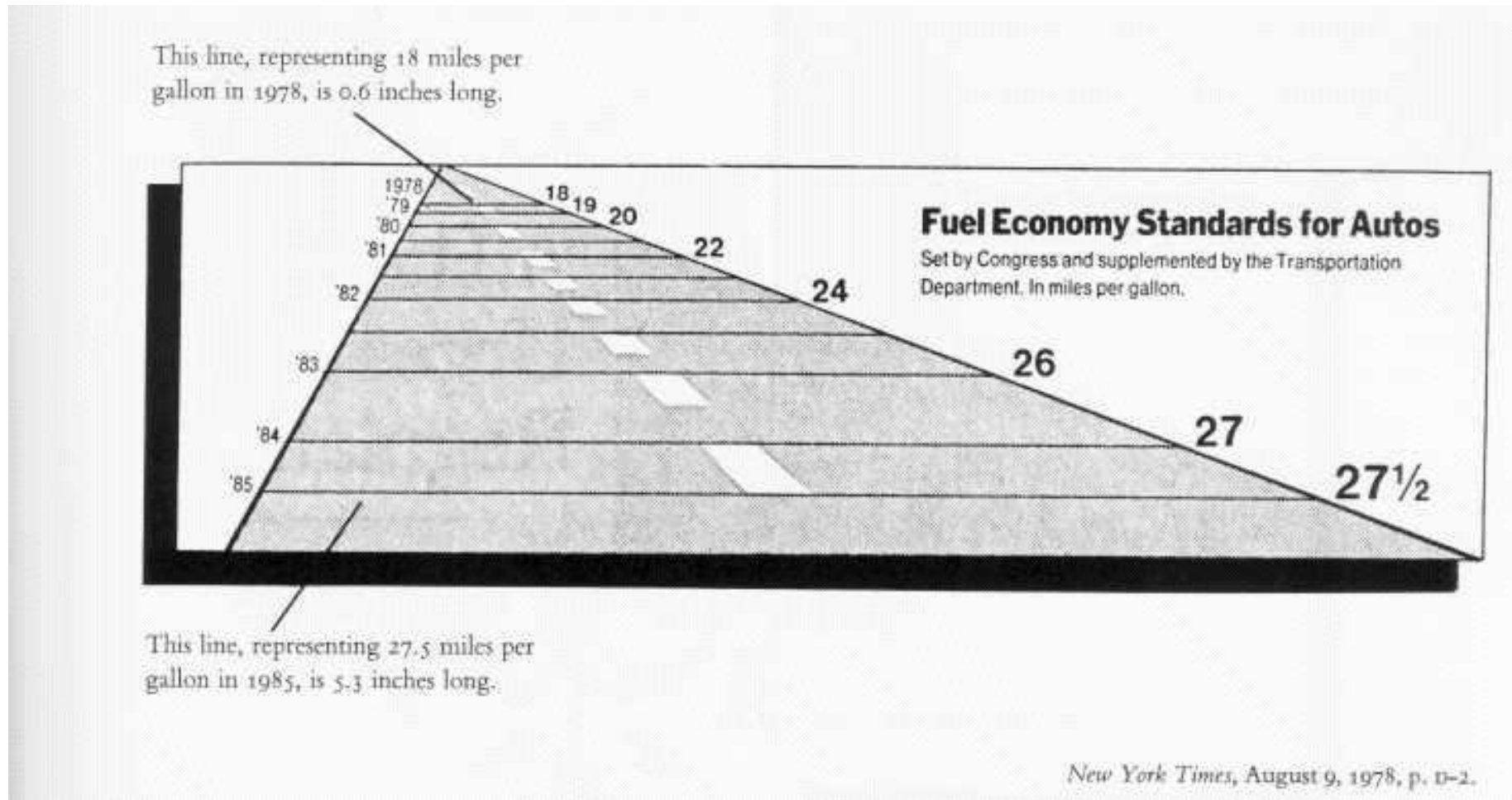
# Better Visualization

Year	Sales
1999	2,110
2000	2,105
2001	2,120
2002	2,121
2003	2,124



Axis from 0 to 2000 scale gives  
correct impression of small change

# Lie Factor – Classic Example



# Lie Factor

$$\text{Lie Factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}} =$$

$$= \frac{\frac{(5.3 - 0.6)}{0.6}}{\frac{(27.5 - 18.0)}{18}} = \frac{7.833}{0.528} = 14.8$$

Tufte requirement:  $0.95 < \text{Lie Factor} < 1.05$

# What is a Dashboard?

- A dashboard is a visual display of the most important information [...] which fits entirely on a single computer screen [...]
- **Who is the audience of this dashboard?**
  - Top management or project sponsors or team members or other departments?
- **What are they interested to know?**
  - Day to day issues or High level stuff or Plans or Budgets?
- **What is the frequency for updating the dashboard?**
  - Weekly, Bi-weekly or Monthly

# Dashboard Design – Best Practices

- Primary Focus of Dashboard: Top Left
- Dashboard Title: Top Right
- Dashboard Description: Provide short description regarding contents
- Dashboard Size: Printable to one page or sized to fit within constraints of web page
- Number of Visualizations per Dashboard: 3-6 graphs/charts to ensure content is interesting, but not cluttered
- Filters: Placed across the top or down the left-hand side. Should be no more than two rows of filters. Apply to every visualization on the dashboard if possible
- Provide tool tips where needed: Make sure they are easy to read and understand.
- Online Dashboard: Minimize scrolling, both horizontal and vertical
- Information Icon: Include instructions or explanations about the dashboard. Note the source and date of the data displayed
- Limit unnecessary color and decoration



# Colour Tips

- Use color key on dashboards
- Use soft, natural colors for normal use
- Use colors equal in tone to not emphasize one area over another
- Use fully saturated, dark colors for emphasis only
- Use lightest colors for large bars, dark for smaller

# Data Visualisation- Tools

- QlikView
- Klipfolio
- Tableau
- Geckoboard
- Power BI
- Google Data Studio

# Data Visualisation- Tools

Tools	Strength	Weakness
<b>QlikView</b>	Attractive User Interface	Difficult to self service for non technical users
	Easy to setup filtering for any kind of visuals	Syntax are not clear and obvious
	Fast rendering of both graphs & tables	Features are complicated to program
	Supports data imports from a comprehensive variety of sources	Reporting (email) functionality - needs to be improved
	Undisputed speed of loading and processing of data, even with millions of instances	
<b>Tableau</b>	Intuitive and attractive user interface	Initial data prep is required
	Seamless integration with big data platforms	Features may seem too specialised and restricting
	Supports mobile platform	Implementation of "row level" security is complicated
	Reliable customer support & community collaboration	Requires IT consultancy
	Vast library of video materials about the tool, online courses, learning blogs etc	

Thanks!!!