

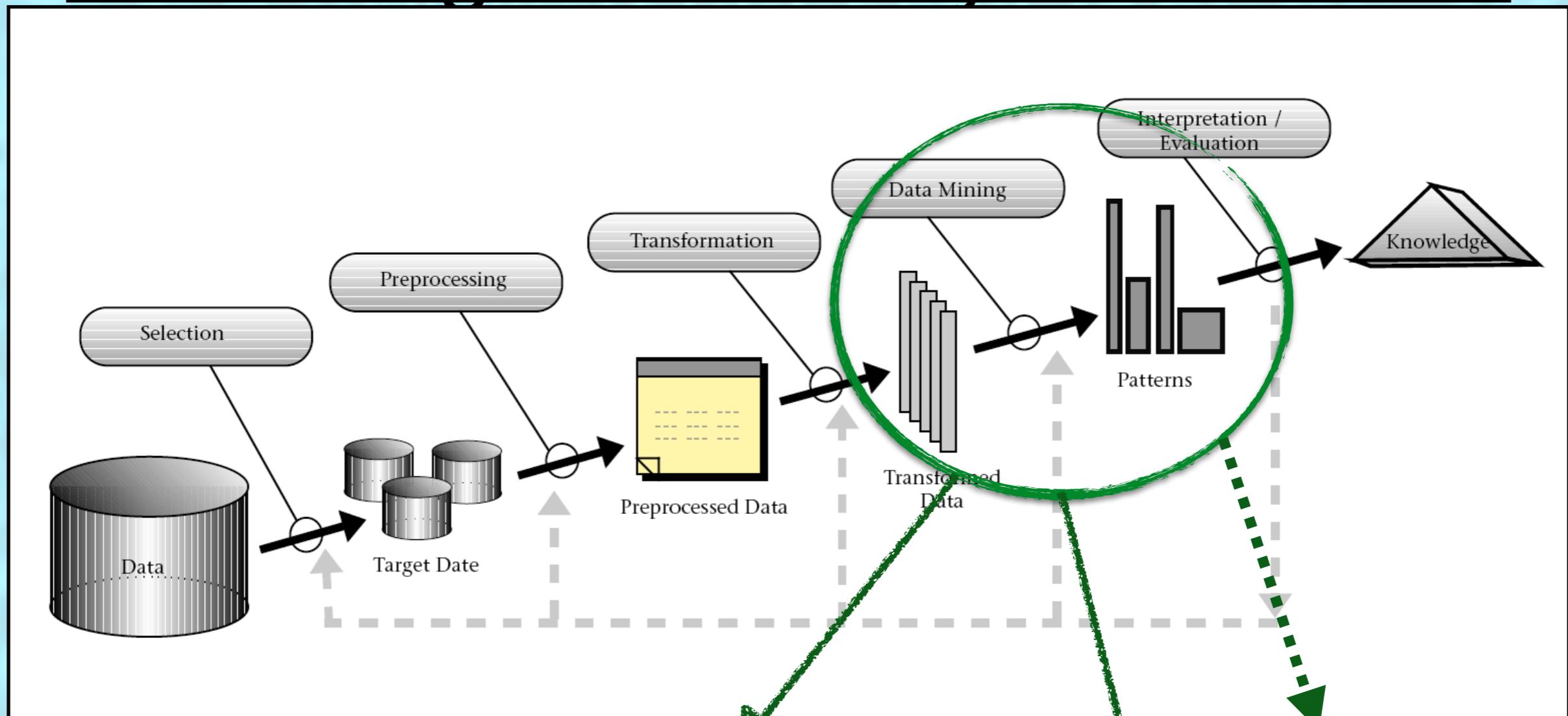
DATA Analysis con PYTHON



Exploratory Data Analysis

Exploratory Data Analysis

Knowledge Discovery in Datasets



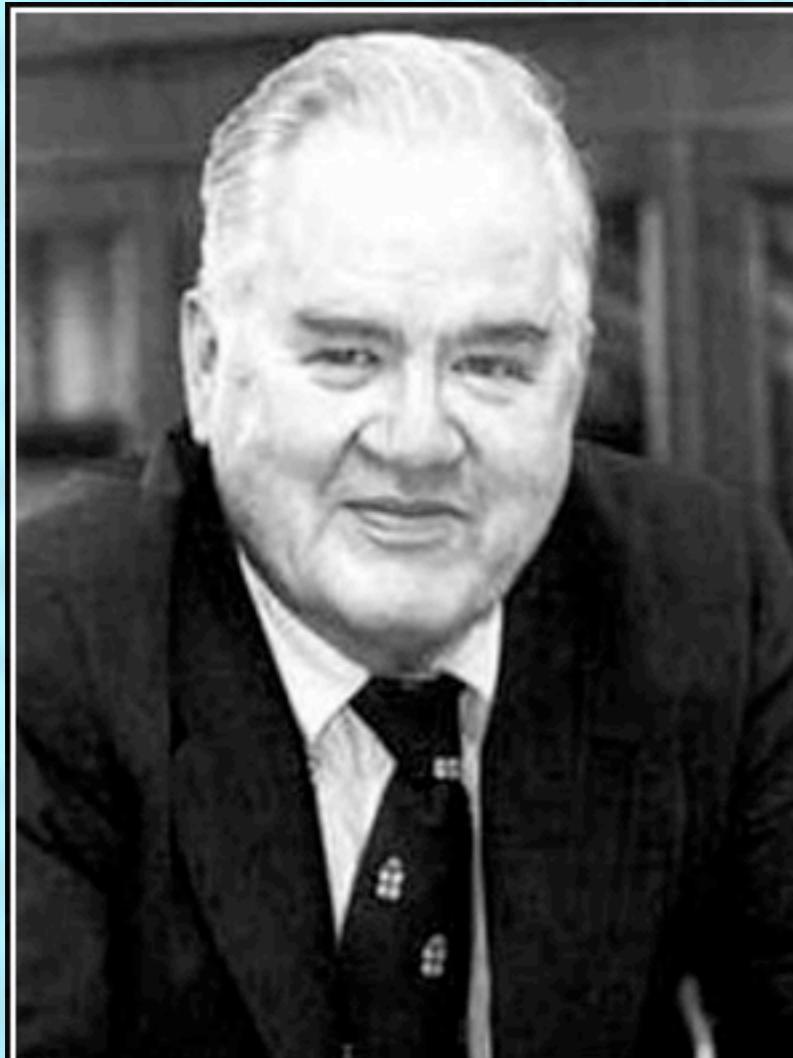
Exploratory Data Analysis

visualizati

Descriptive

...

Exploratory Data Analysis (EDA)



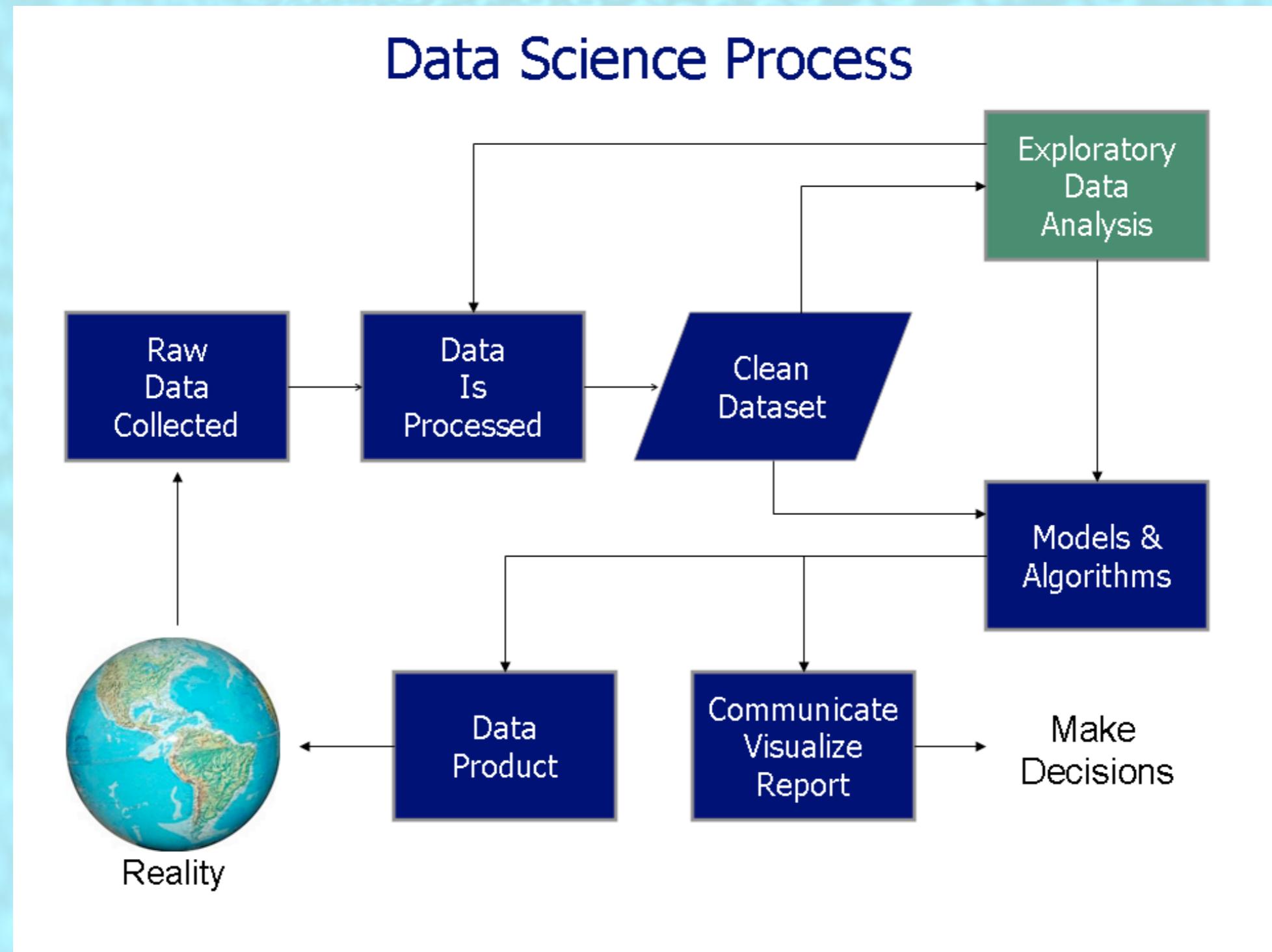
The greatest value of a picture is
when it forces us to notice what we
never expected to see.

— *John Tukey* —
(1915-2000)

AZ QUOTES

Tukey, John (1977), Exploratory Data Analysis, Addison-Wesley.

EDA Definition: an approach of analysing data that summarise their main characteristics without using a statistical model or having formulated a prior hypothesis [wikipedia]



EDA Definition: an approach of analyzing data that summarize their main characteristics without using a statistical model or having formulated a prior hypothesis [wikipedia]

Main motivations and tasks

maximise insight into a data set

extract important variables

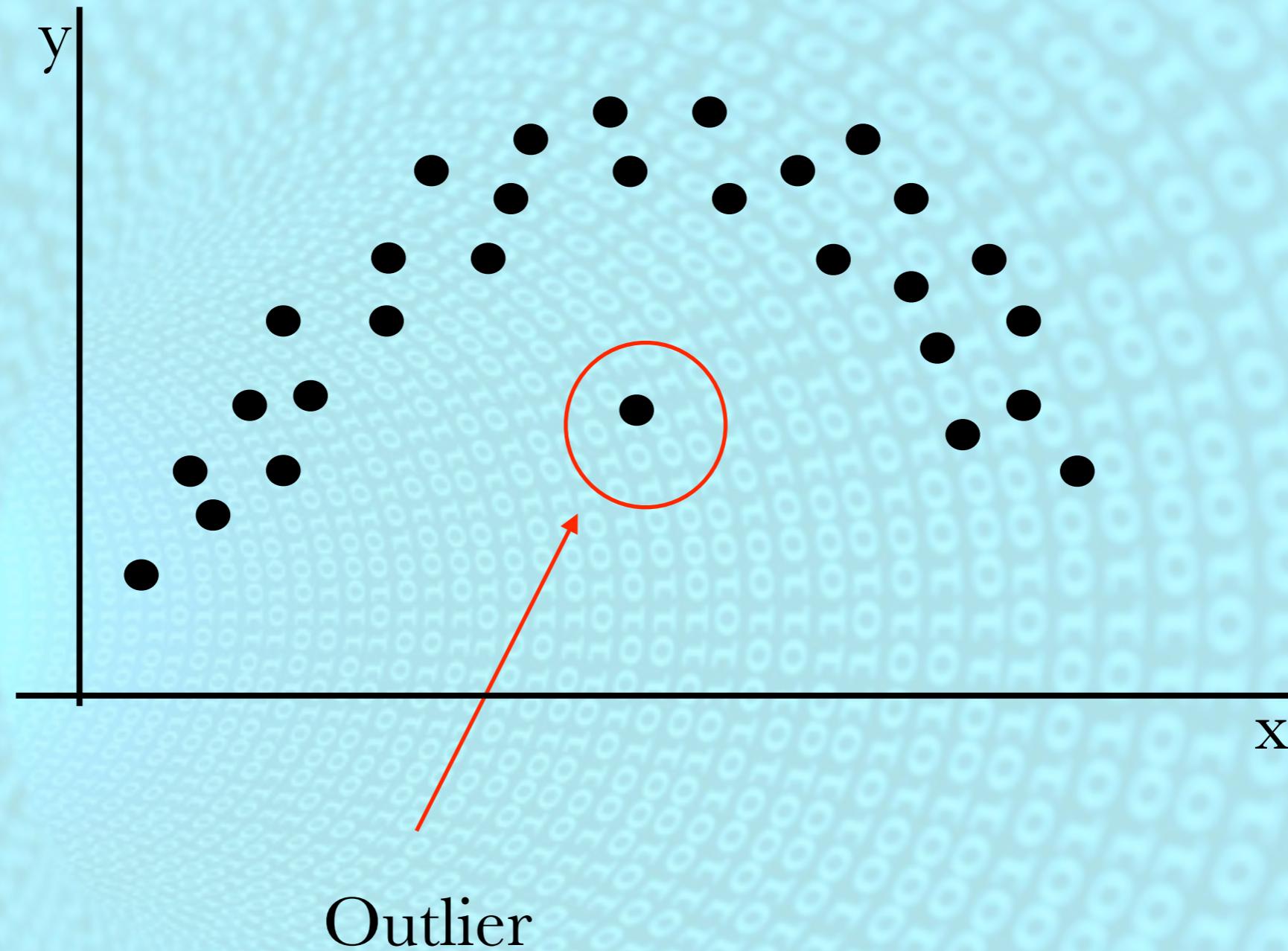
uncover underlying structure

detect outliers and anomalies

test underlying assumptions

address appropriate models of analysis

Example: detect outliers and anomalies

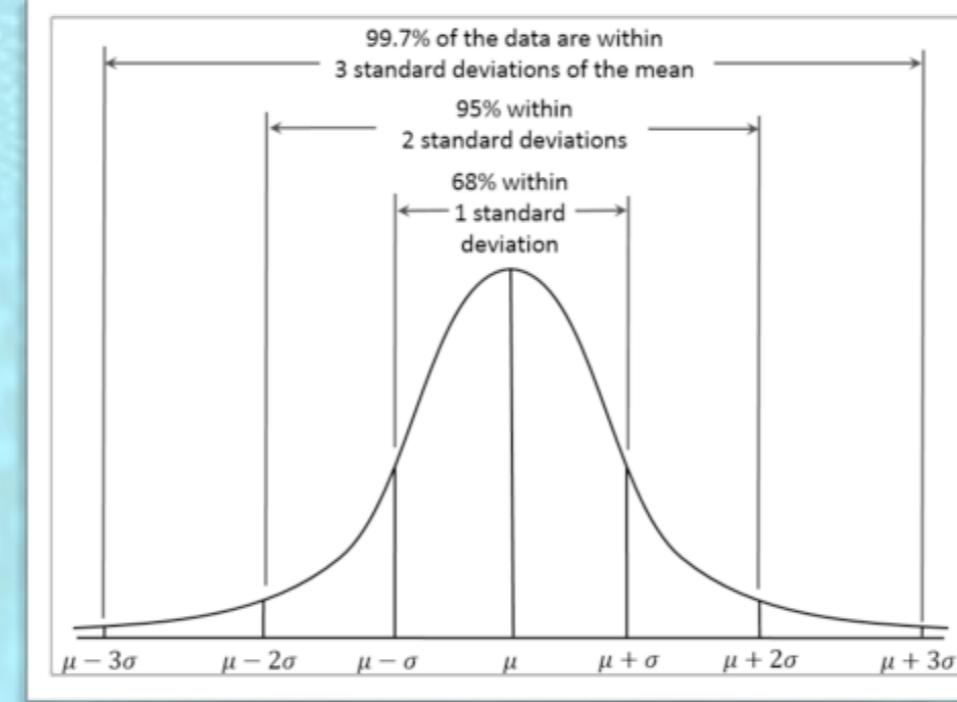
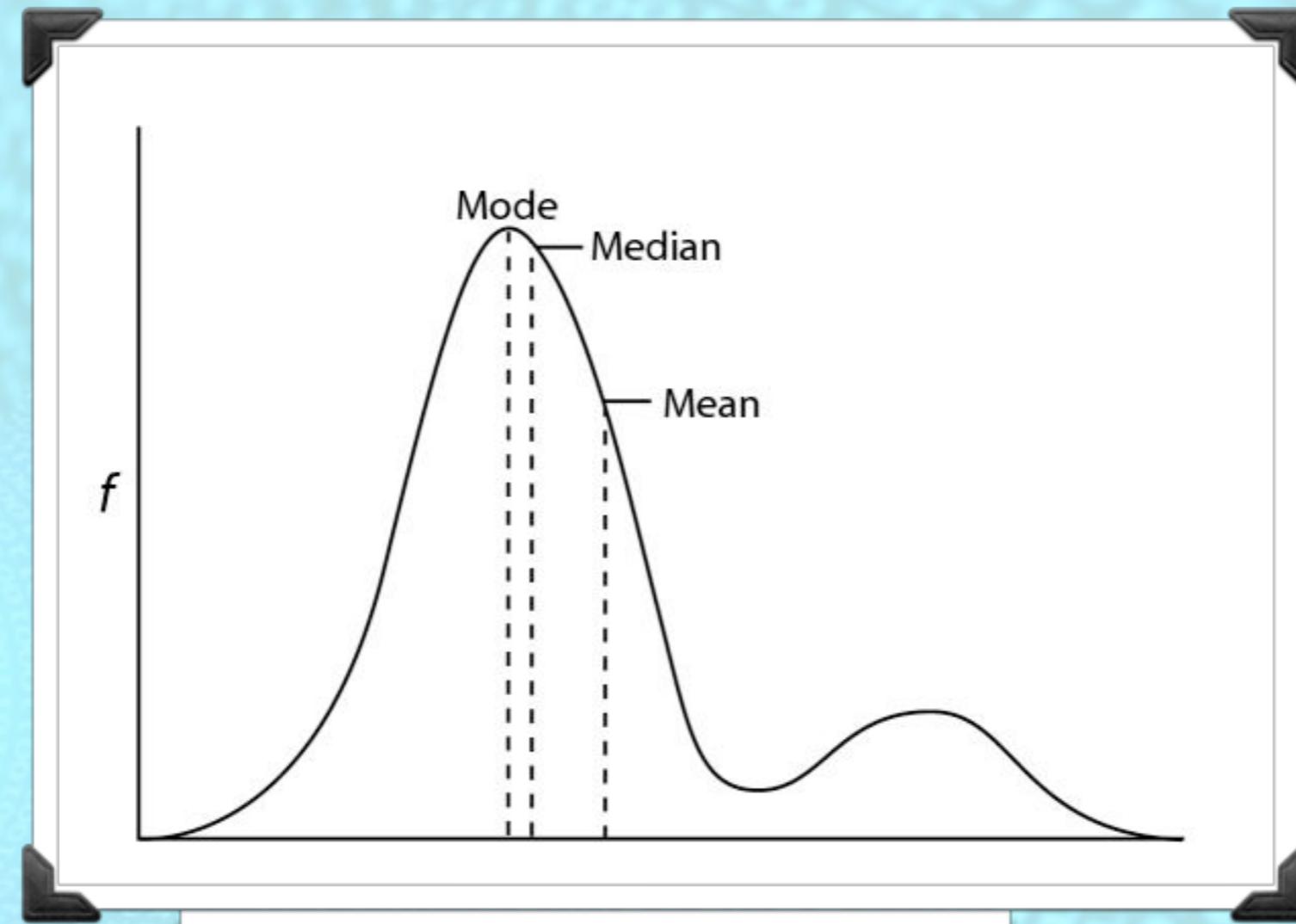


What are anomalies/outliers?

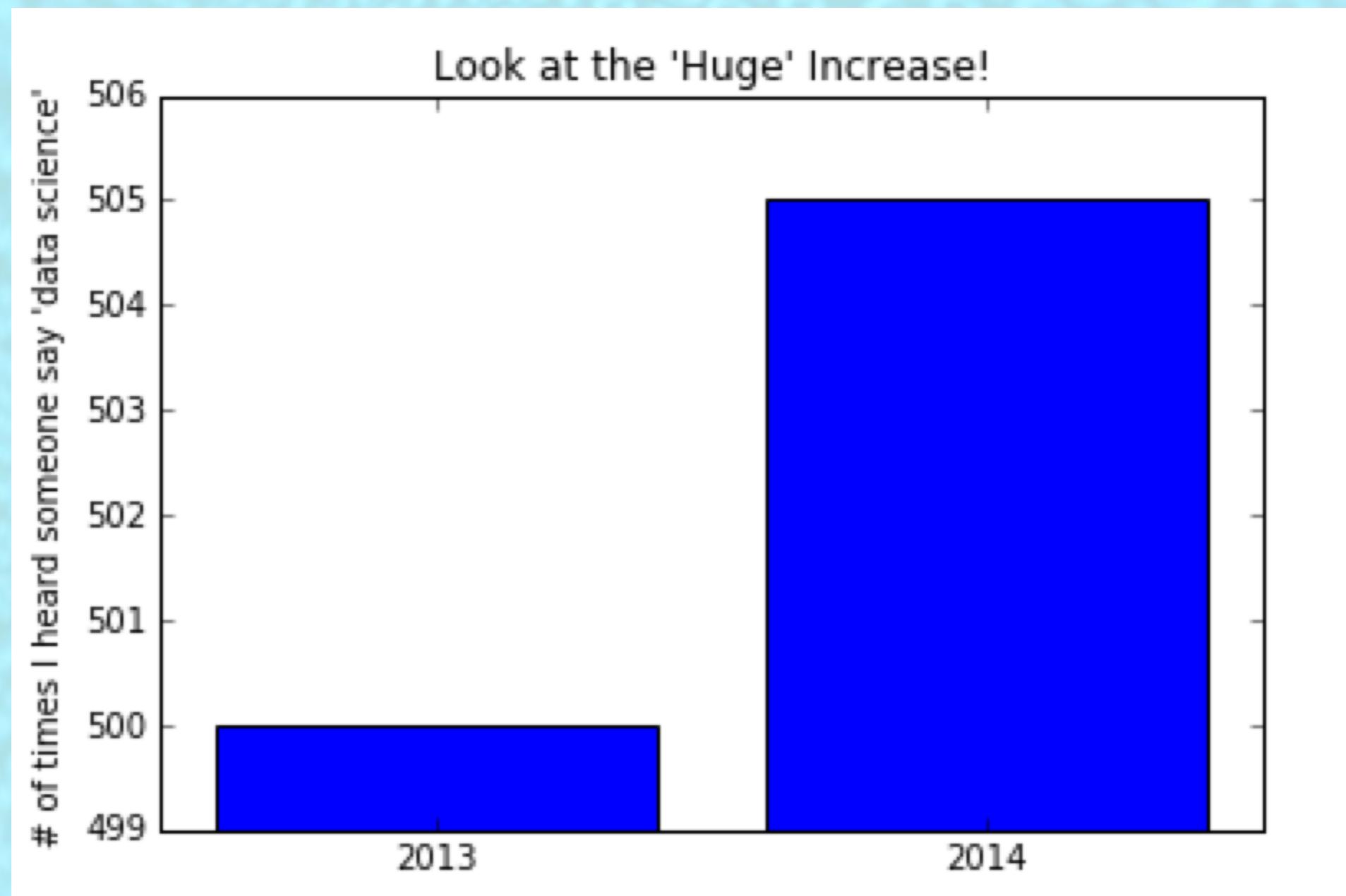
The set of data points that are considerably different than the remainder of the data

Applications: credit card fraud detection, telecommunication fraud detection, network intrusion detection, fault detection

Example: test underlying assumptions

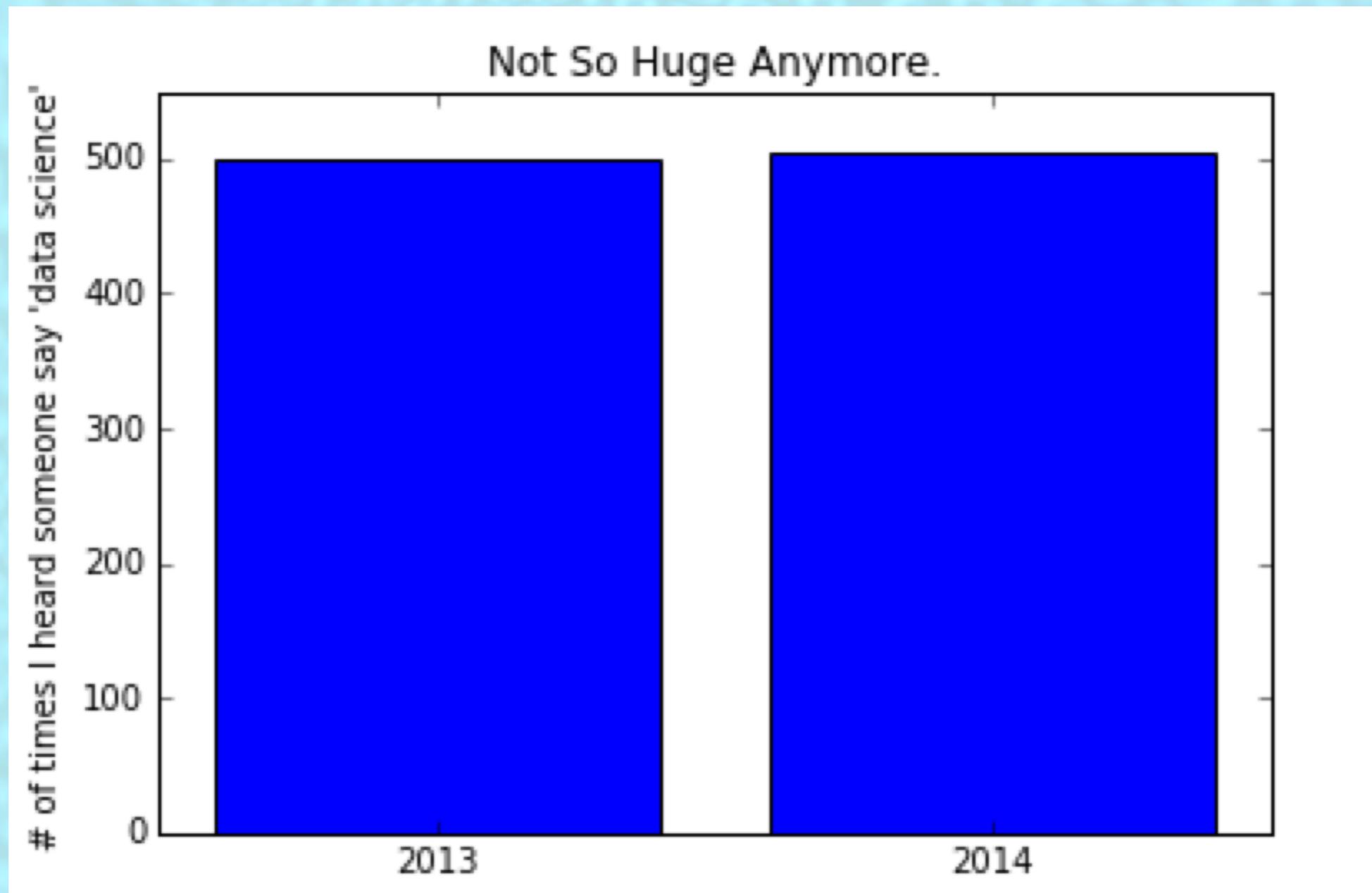


Few recommendations: 1) be fair



misleading bar-charts

Few recommendations: 1) be fair



misleading bar-charts

Facts are stubborn, but statistics are more pliable (Mark Twain)

Linear scale not always good

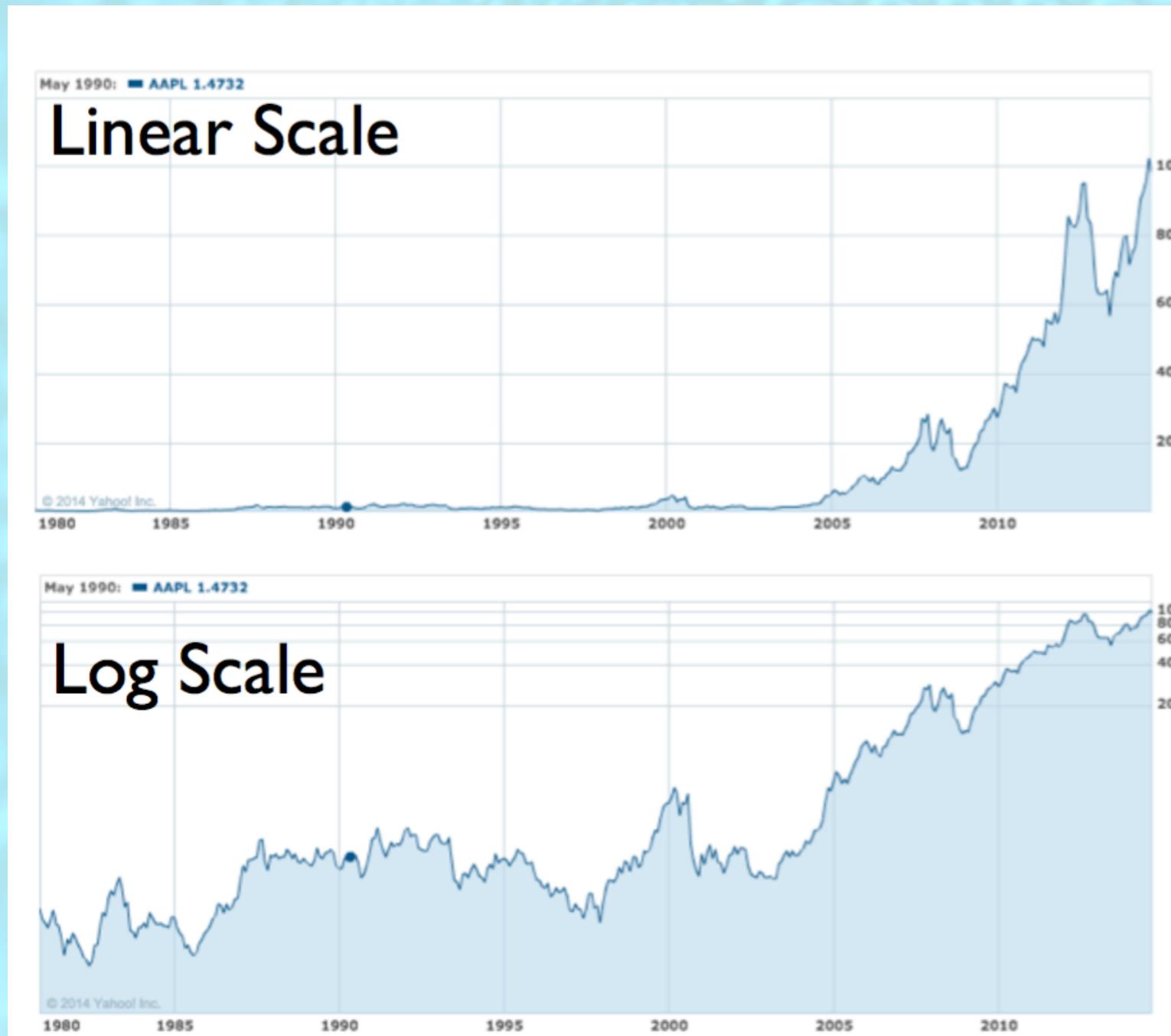
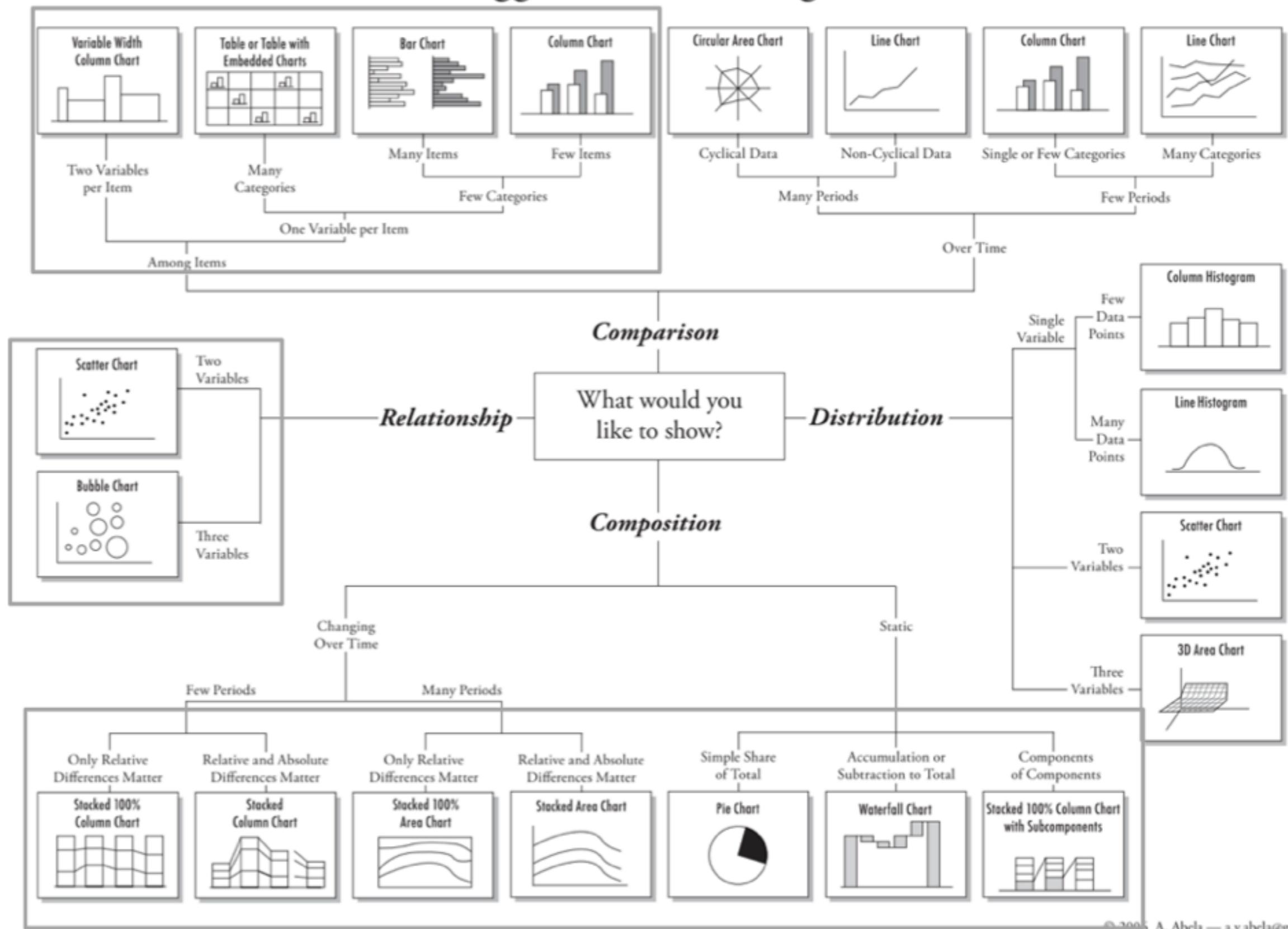


Chart Suggestions—A Thought-Starter



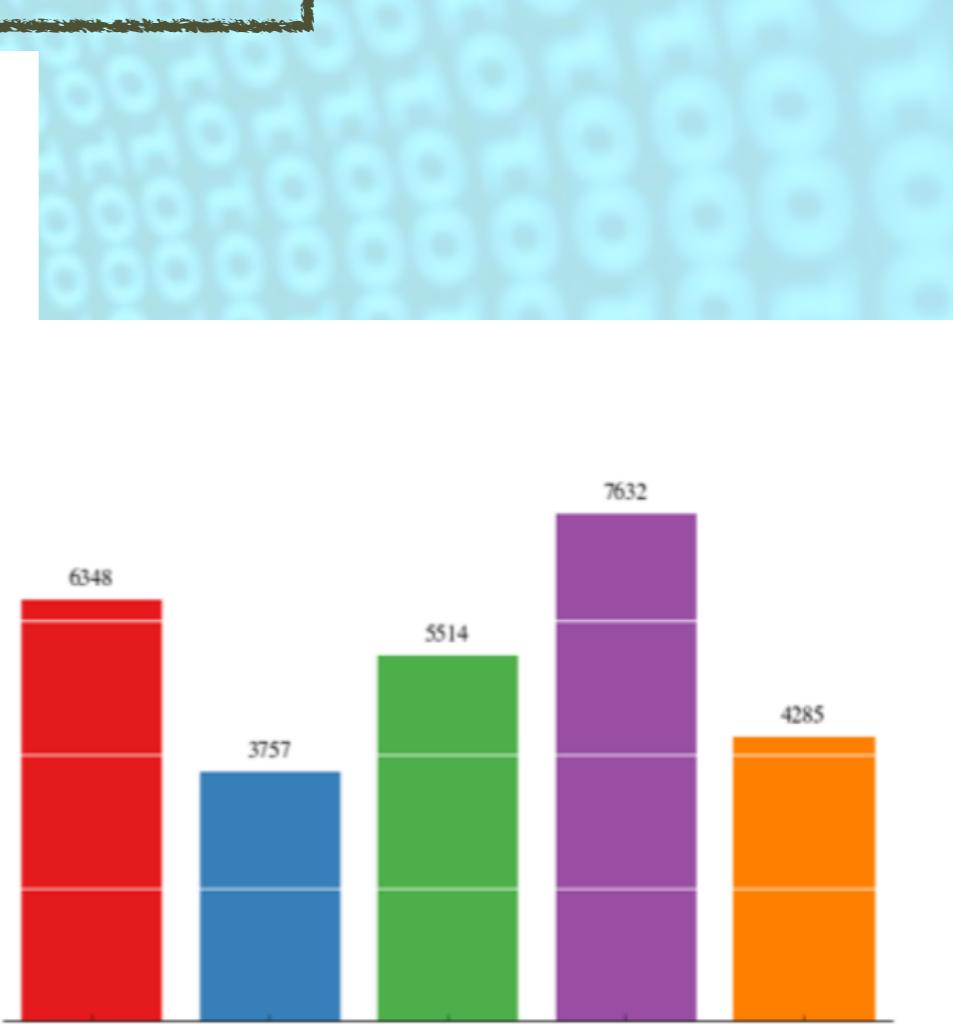
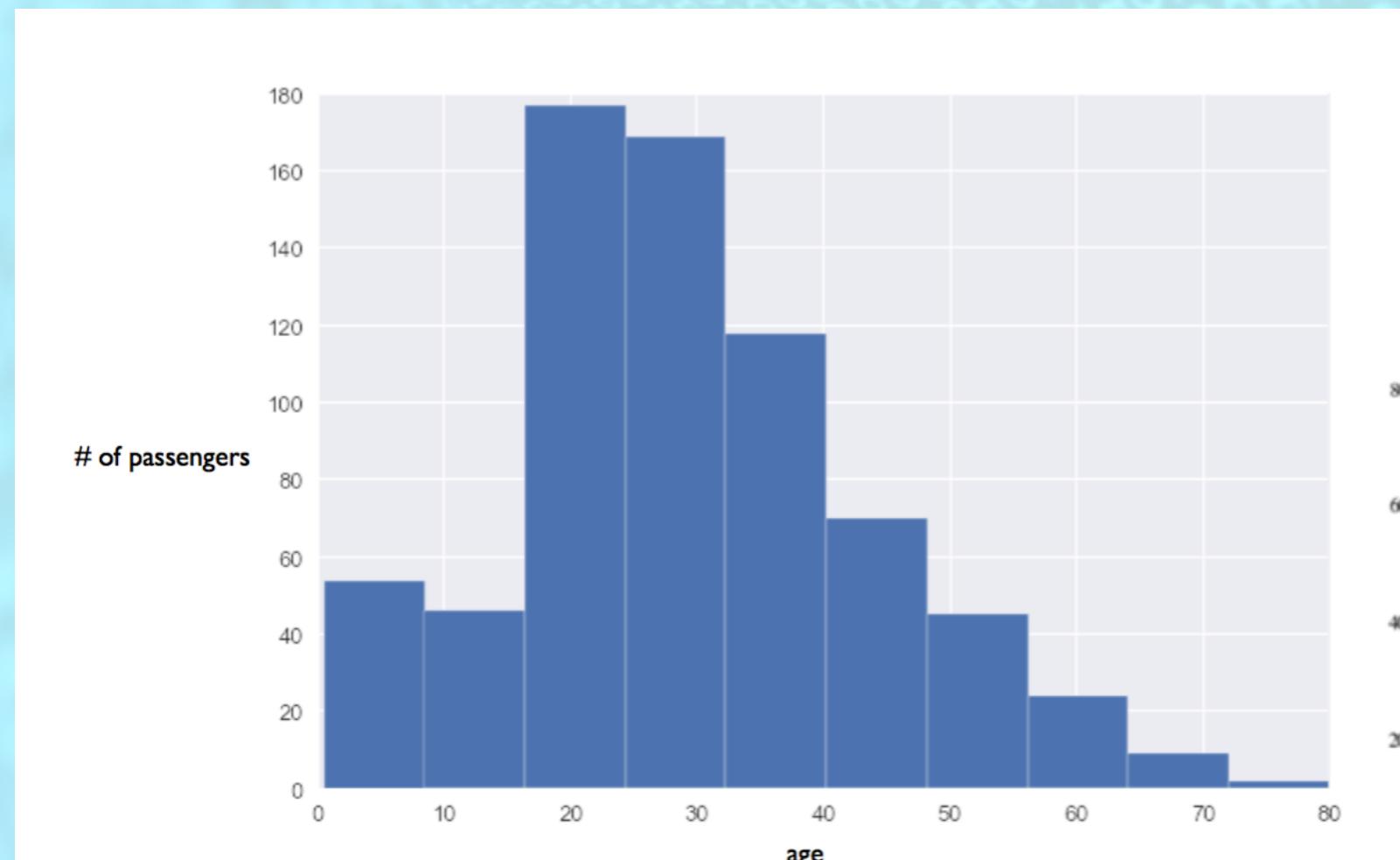
© 2006 A. Abela — a.v.abela@gmail.com

http://extremepresentation.typepad.com/blog/files/choosing_a_good_chart.pdf

Kind of plots

Bar - charts

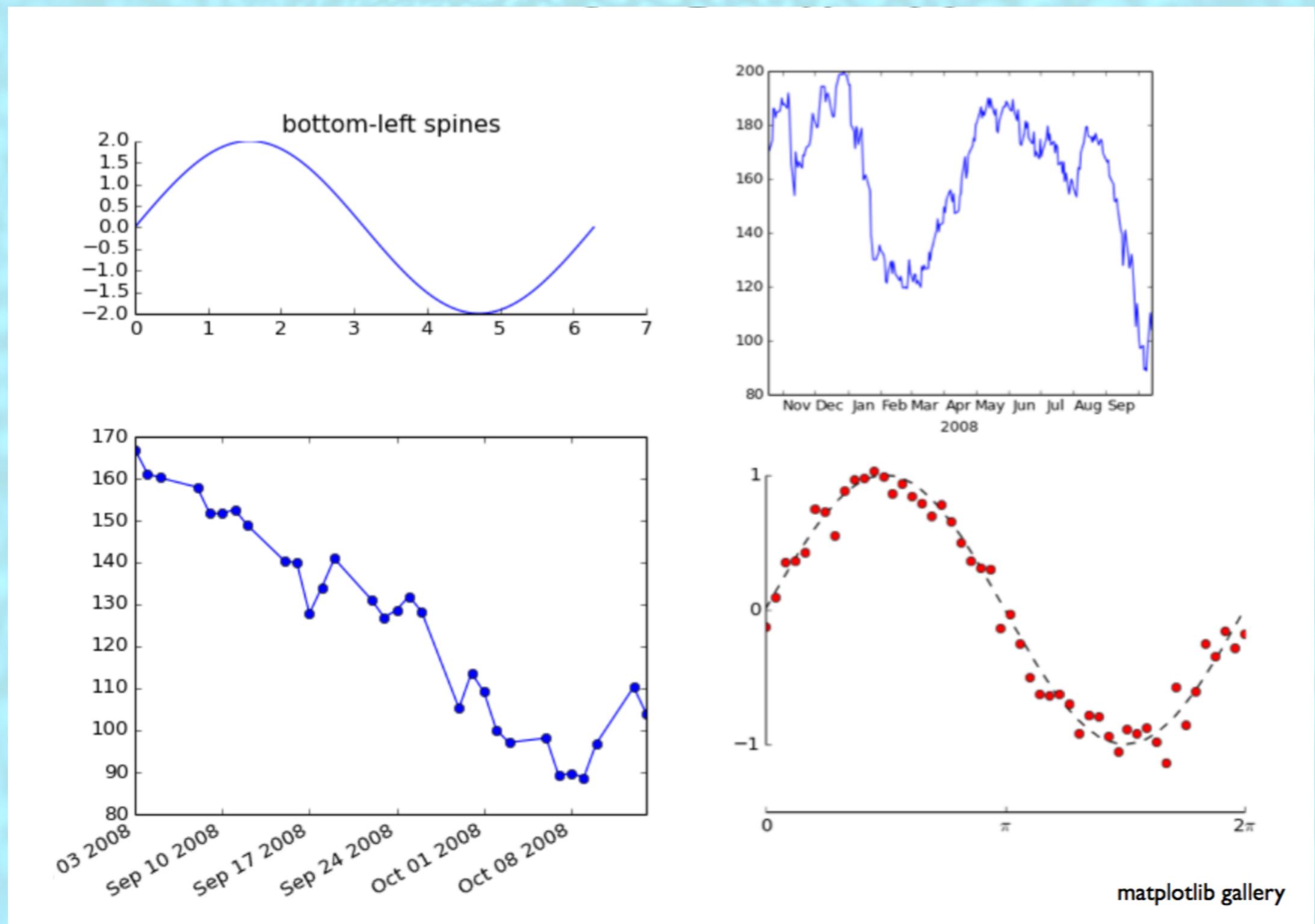
- 1) when you want to show how some quantities varies among a set of items
- 2) also good to explore how the values are distributed



<http://nbviewer.ipython.org/gist/olgabot/5357268>

Line-charts

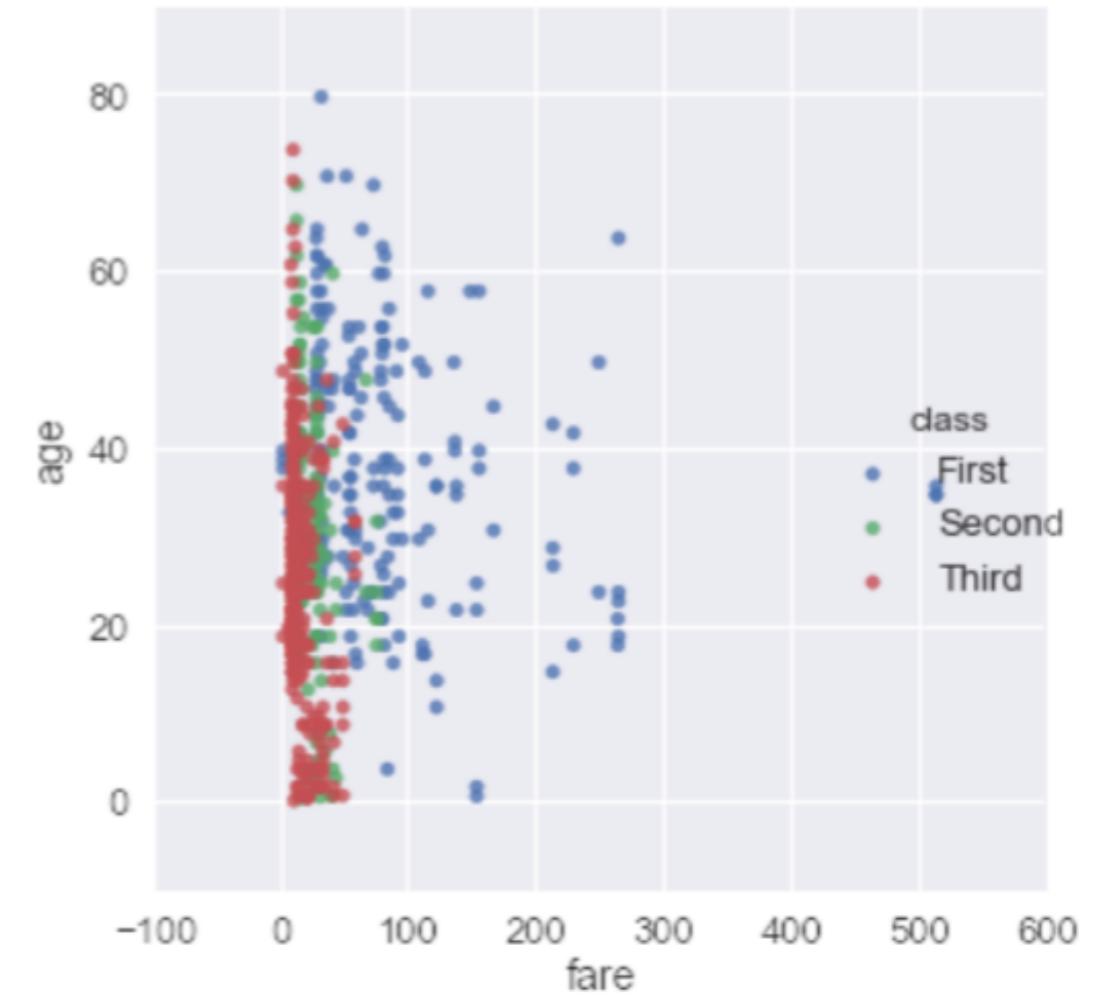
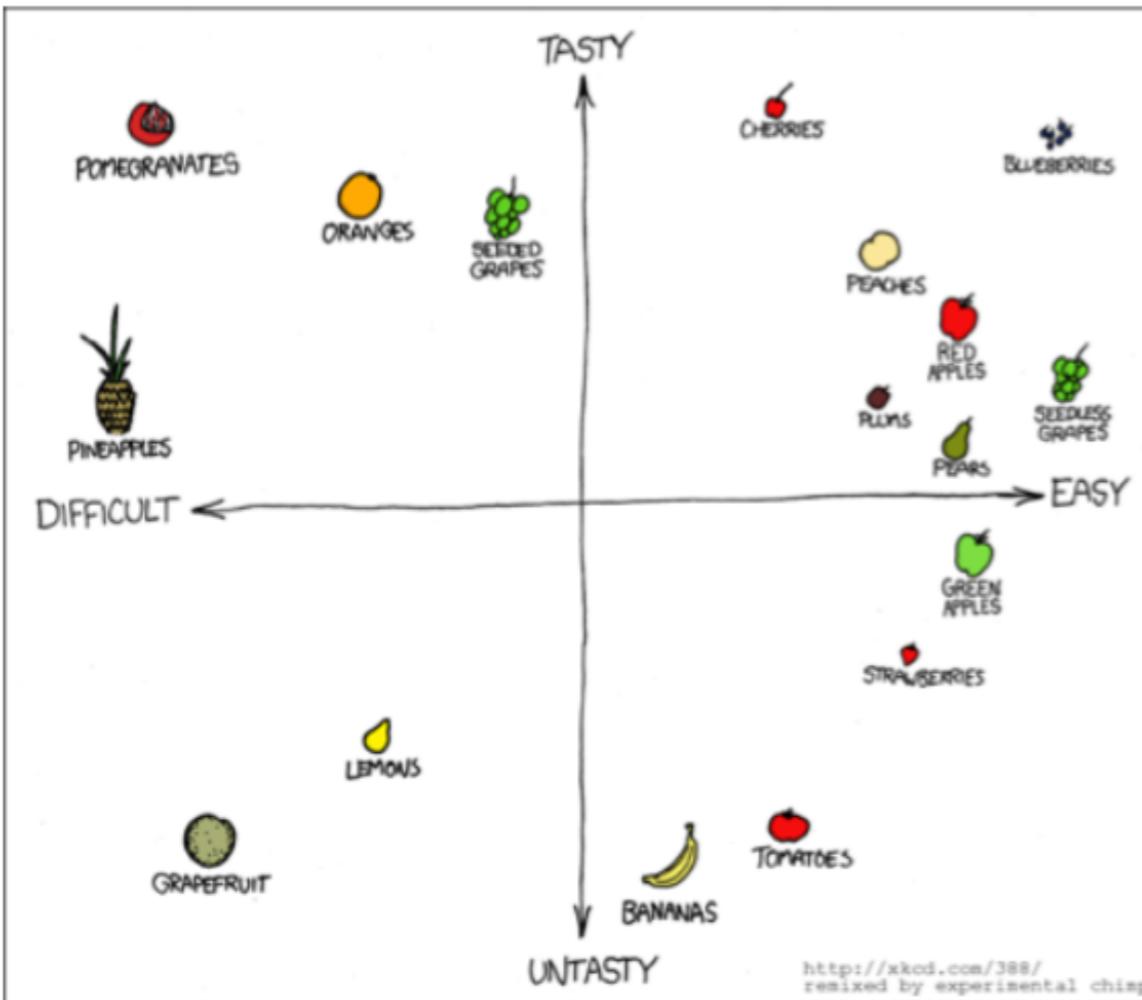
line chart: good choice for showing *trends*



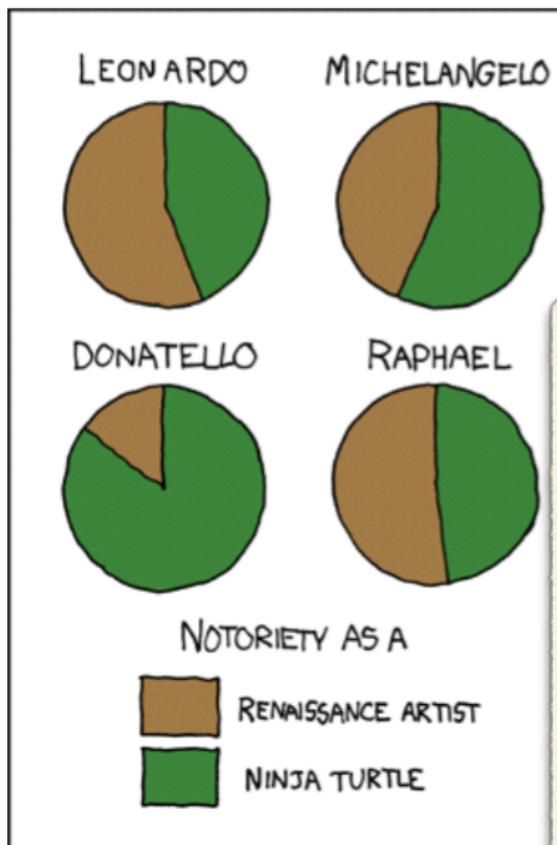
matplotlib gallery

Scatterplots

scatter plot: good choice for showing *relationship* between two paired sets of data



Pie Charts

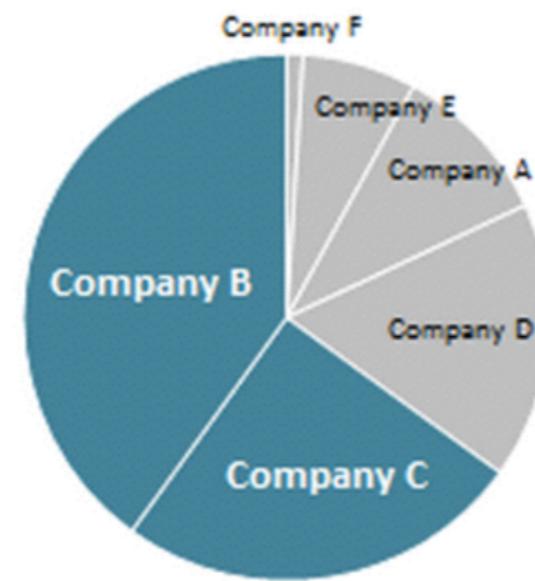


Passenger Class on the Titanic

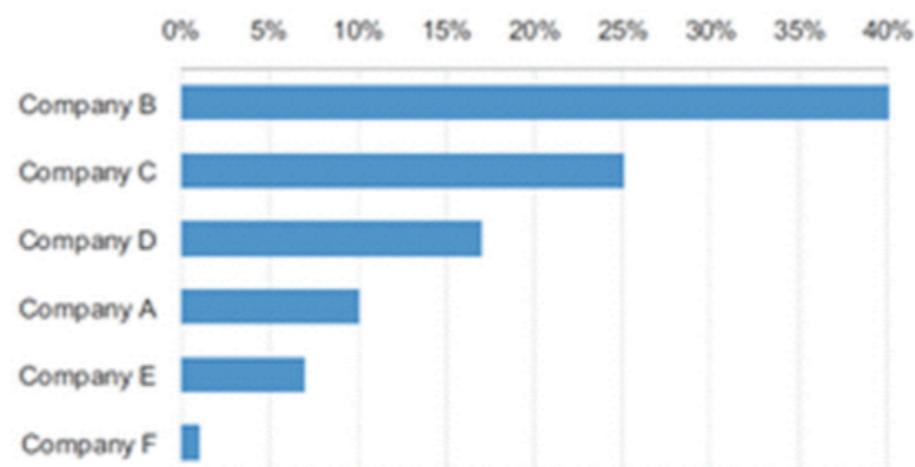


Pie vs. Bar Charts

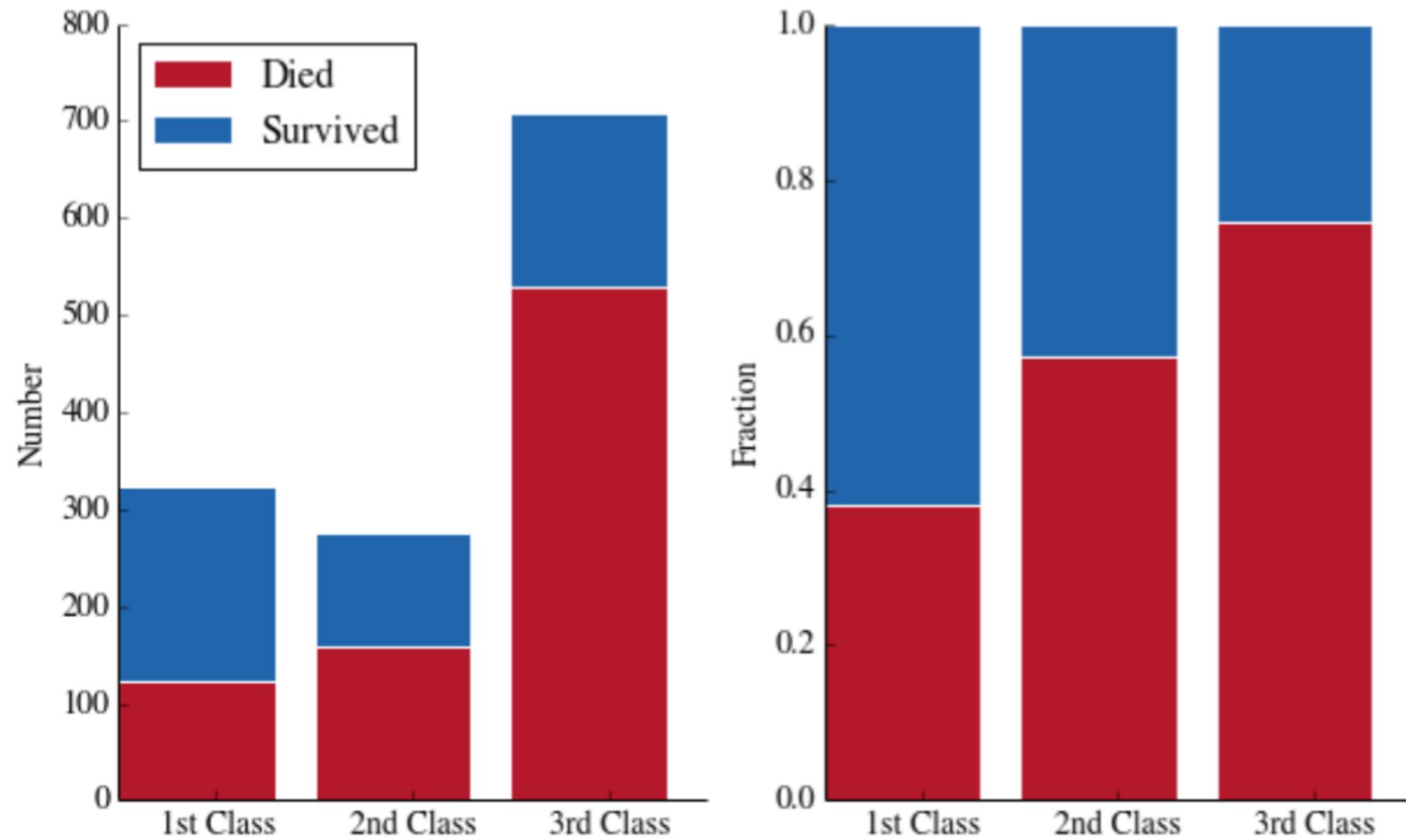
65% of the market is controlled by companies B and C



Company Percentages of Total Market Share



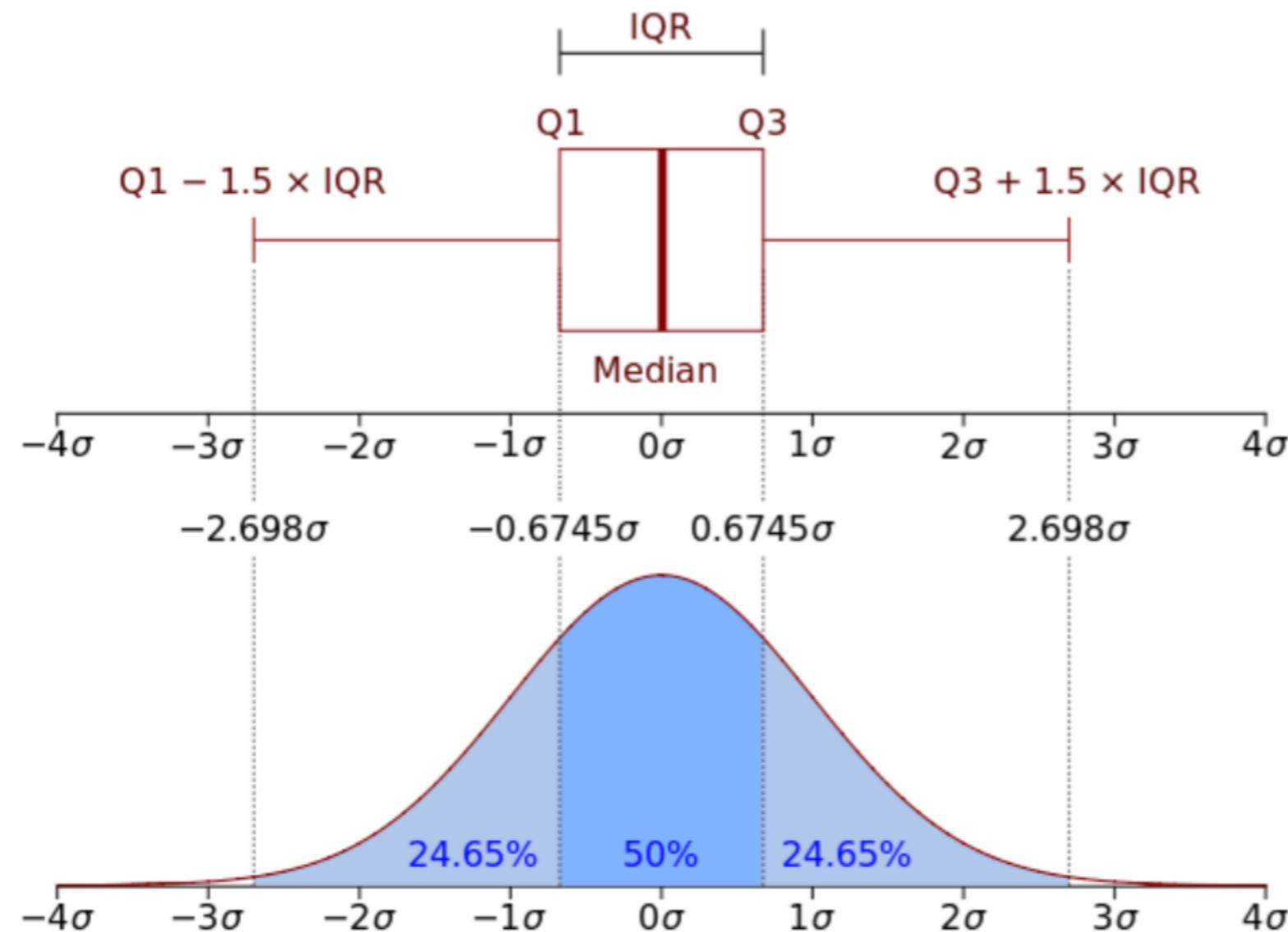
Stacked Bar Chart



Box plots...

Box Plots

aka Box-and-Whisker Plot



Wikipedia

MATPLOTLIB

<https://matplotlib.org>

SEABORN library is written ON Matplotlib and produce more complex visualisations in an easy way

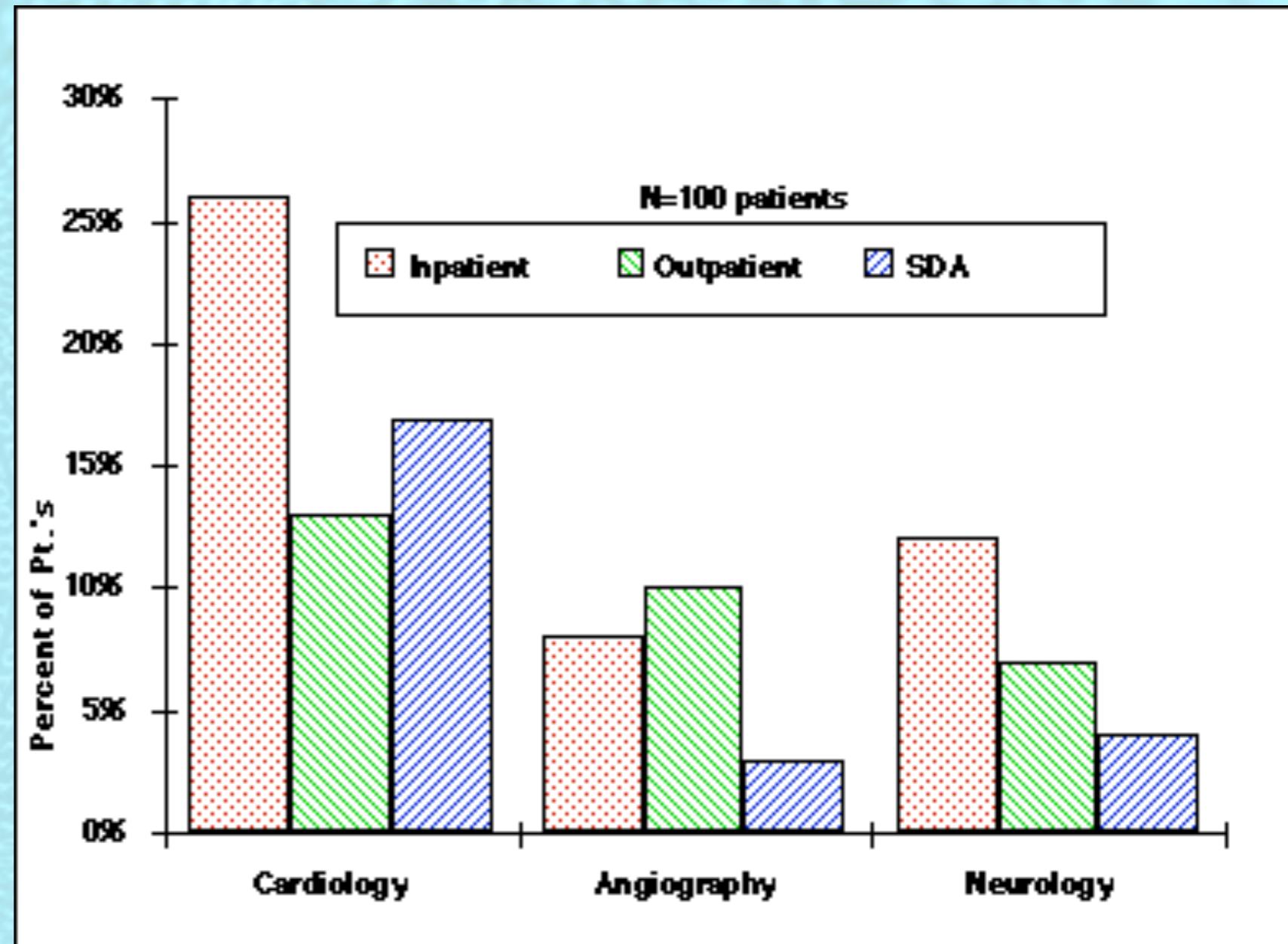
<https://seaborn.pydata.org>

BOKEH allow WEB plots and 3D visualizations in Python

<http://bokeh.pydata.org/en/latest/>

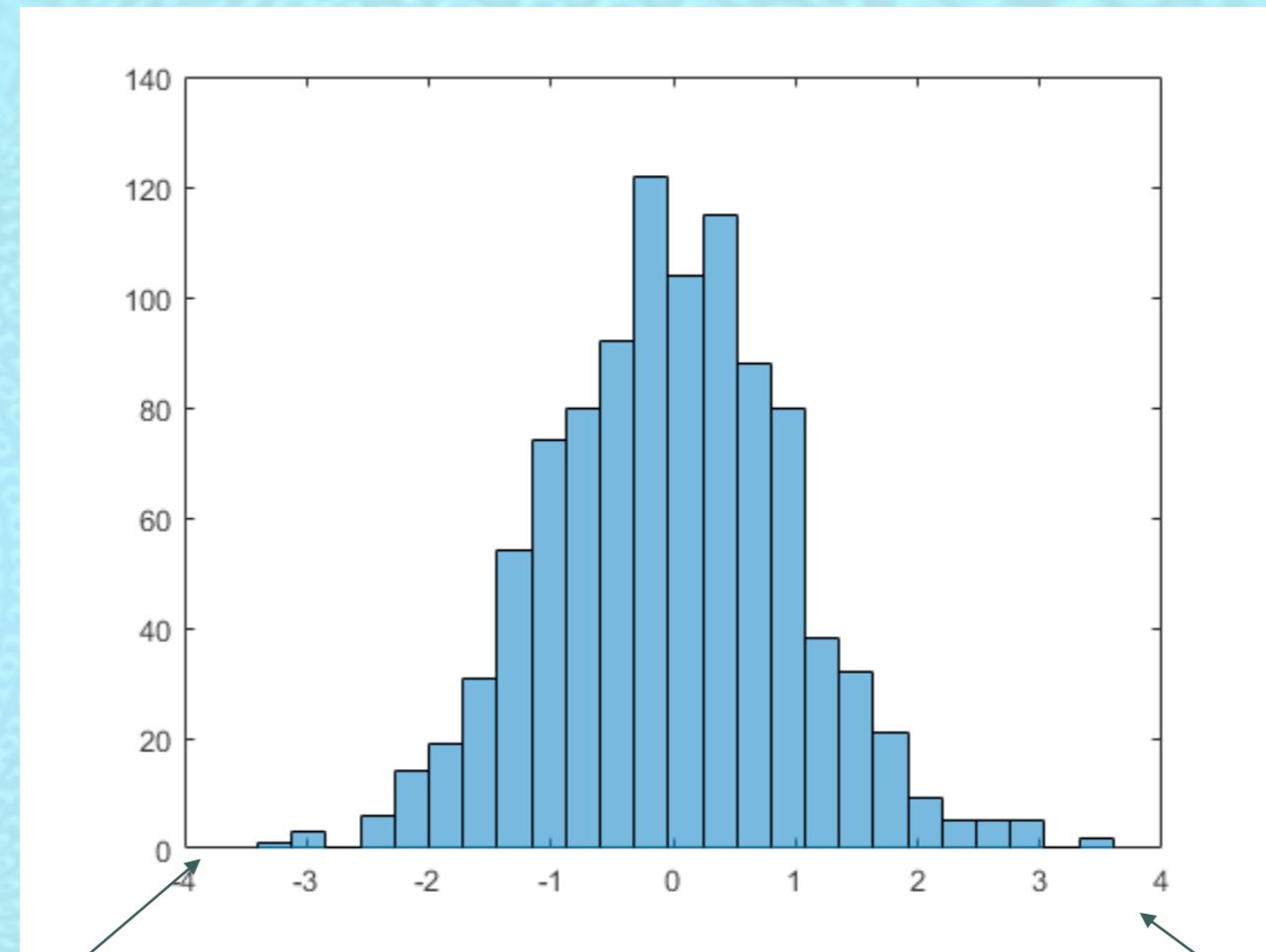
Summary statistics

*) In the worse case of a 1-D set of
o unordered categorical⁽¹⁾ values,
the best we can do is calculate
frequency:



$$\text{frequency}(\nu_i) = \frac{\text{number objects with attribute } \nu_i}{m}$$

*) In the happier case of a *numerical-set* of data, we can start playing with numbers



min

total size

where is centered...?

max

Measure of location: central tendency

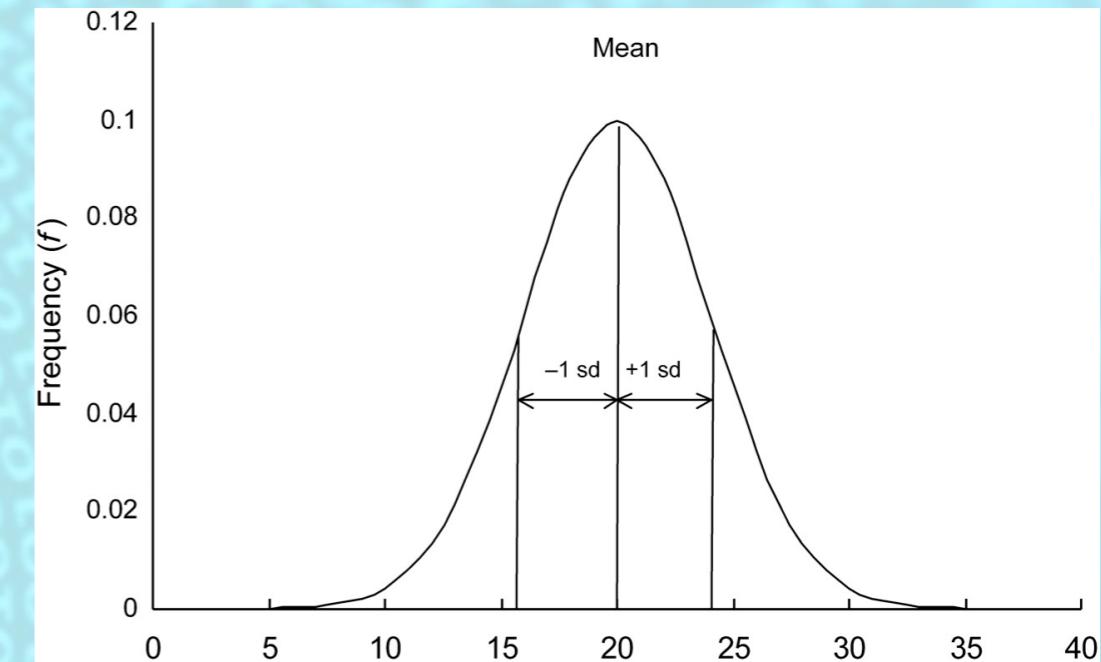
1) Mean

$$A = 2, 6, 7, 8, 9, 6, 4, 6, 2, 5, 8$$



$$\text{Mean} = (2+6+7+8+9+6+4+6+2+5+8) / 11 \sim 5.7$$

$$\mu = \bar{x} = \frac{x_1 + x_2 + x_3 \dots}{n} = \sum_i^n \frac{x_i}{n}$$



2) Median

$A = 2, 6, 7, 8, 9, 6, 4, 6, 2, 5, 8$ \longrightarrow central value

$$\text{sort}(A) = 2, 2, 4, 5, 6, \mathbf{6}, 6, 7, 8, 8, 9$$



Central tendency

Differences Mean and Median?

$A = 2,6,7,8,9,6,4,6,2,5,8000$

Mean = 732.27 Median = 6

Outliers

3) Mode

$A = 2,6,7,8,9,6,4,6,2,5,8$ \longrightarrow Most common value

$\text{sort}(A) = 2,2,4,5,6,6,6,7,8,8,9$ mode = median = 6

$B = 2,2,4,5,6,6,7,8,8,8,9$ \longrightarrow median = 6
mode = 8

NB “mode” is a special kind of frequency (the most frequent), so it can be applied to categorical

*) Generalization of median = **quantile** = $x_{y\%}$

The value less than which a certain percentile of data lies

$$\text{sort}(A) = \underline{2,2,4,5,6,6,6,7,8,8,9}$$

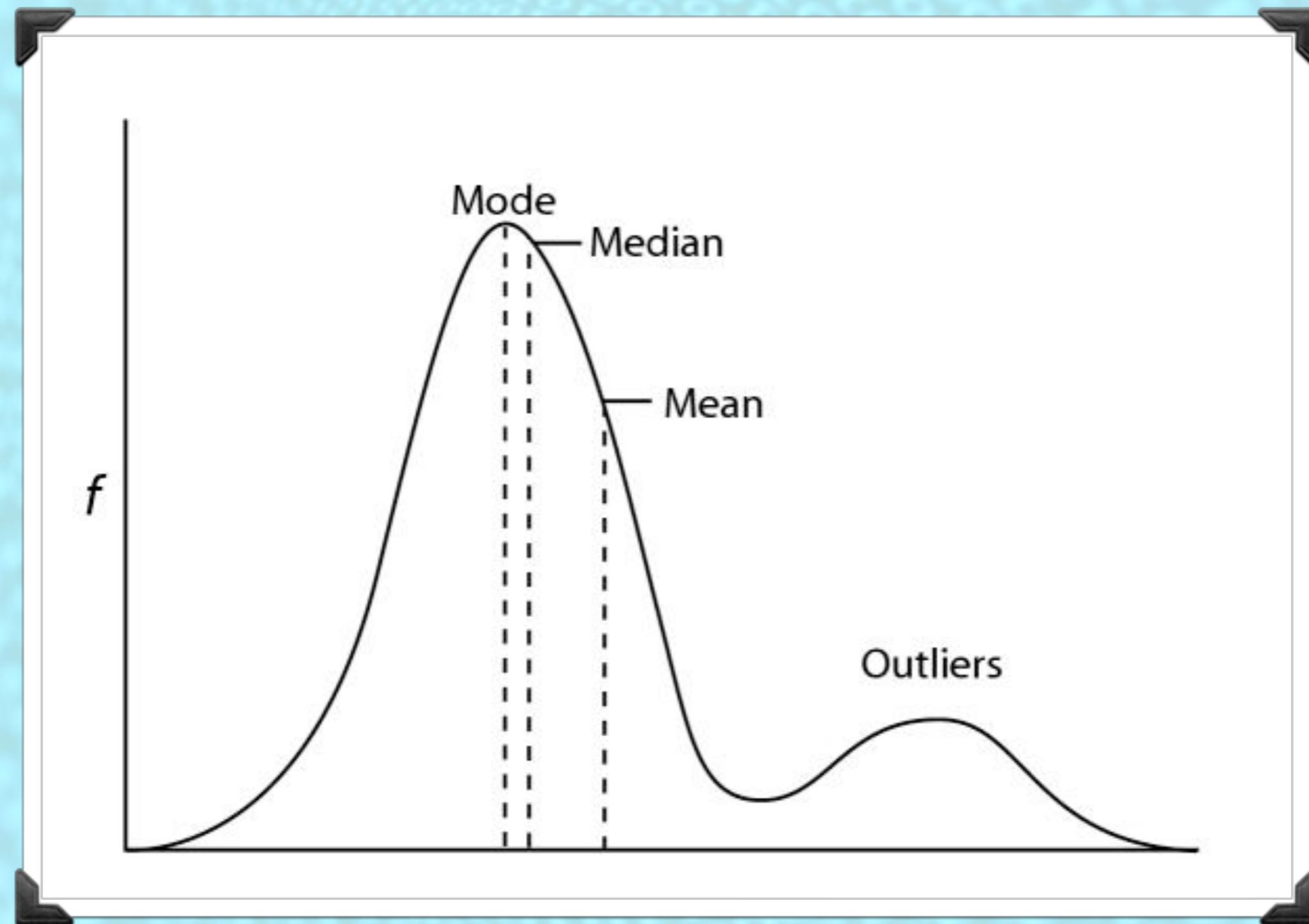
50%



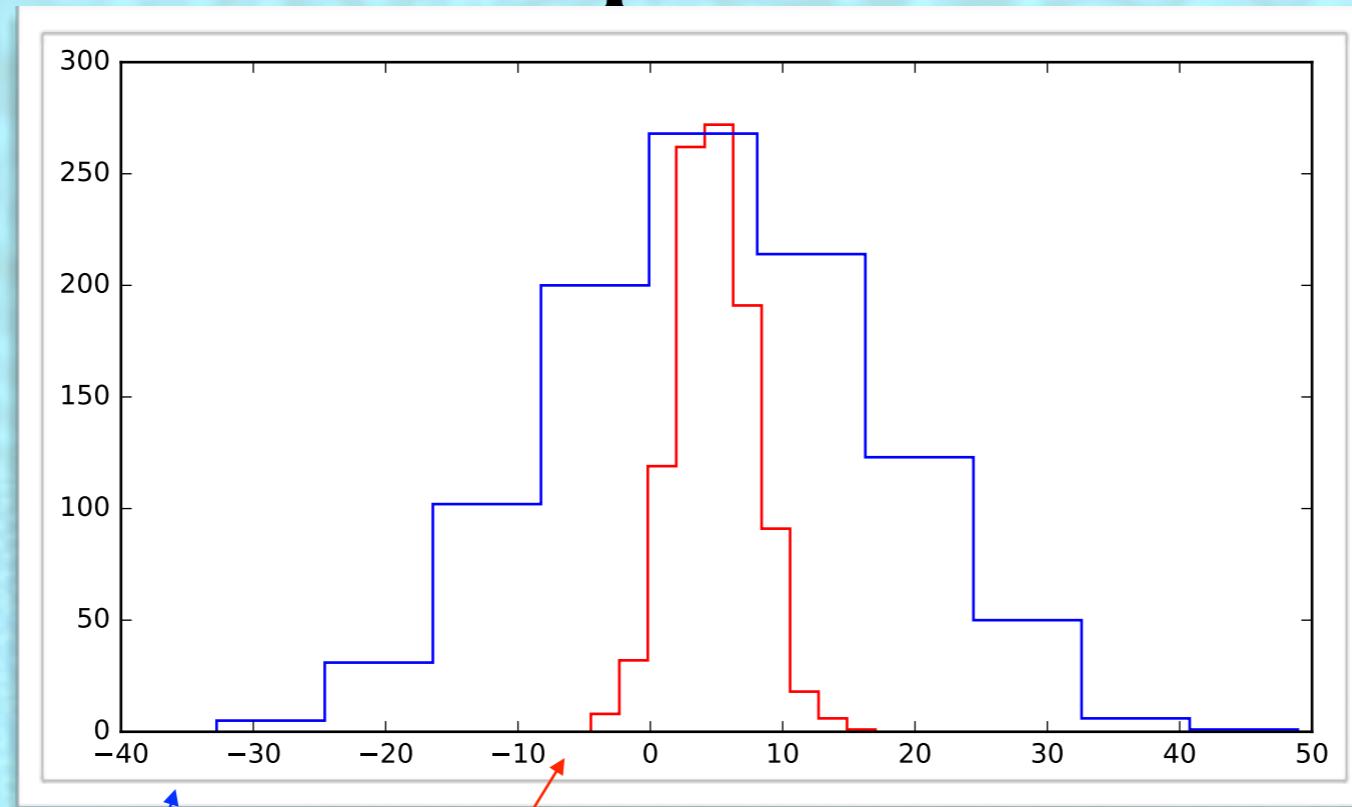
median = quantile of 50%

quantile of 30% ?

Behavior of mode, median and mean respect to outlier



Dispersion



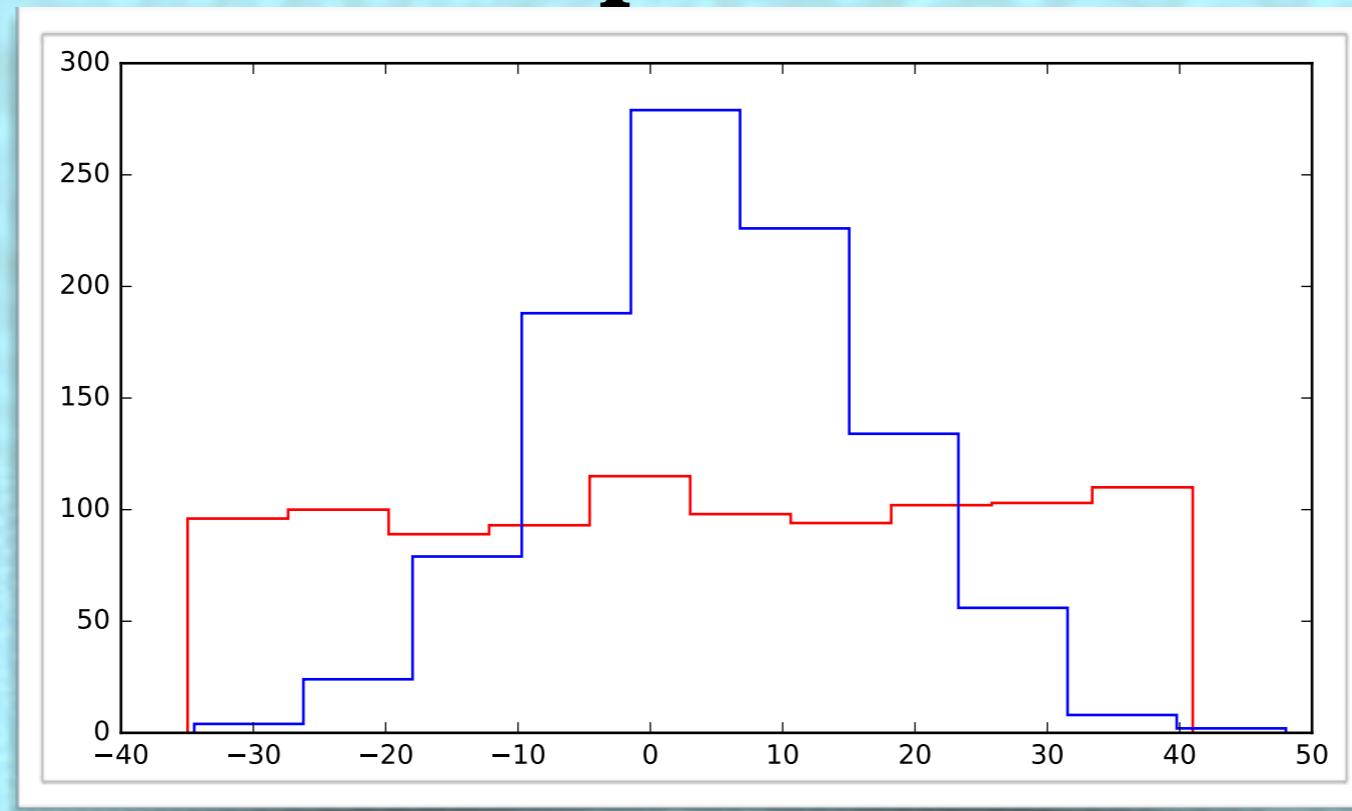
A measure of how spread the data is

1) Range

$$\max(A) - \min(A)$$

same mean different range

Dispersion



A measure of how spread the data is

1) Range

$$\max(A) - \min(A)$$

$$Var(x) = \sigma^2 = \frac{\sum_{i=1}^n (x - \bar{x})^2}{n}$$

variance

$$\sigma = \sqrt(Var(x))$$

standard deviation

$$Var_s(x) = \frac{\sum_{i=1}^n (x - \bar{x})^2}{n - 1}$$

variance from sample

$$\sigma_s = \sqrt(Var_s(x))$$

standard deviation from sample

3) More Robust estimators of spread

Both range and variance are quite affected by **outliers**

1) *interquartile range (IQR)*

$$IQR(x) = x_{75\%} - x_{25\%}$$

2) *Absolute average deviation (AAD)*

$$AAD(x) = \frac{1}{n} \sum_i^n |x_i - \bar{x}|$$

3) *Median absolute deviation (MAD)*

$$MAD(x) = median(|x_1 - \bar{x}|, \dots, |x_n - \bar{x}|)$$

Multivariate Summary statistics

*) measure of the location of data consisting of several attributes can be obtained by computing the mean and median separately

$$\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_n)$$

*) however the spread of the data it is most commonly measured by the

covariance matrix S

Multivariate Summary statistics

1) *covariance matrix*

$$S = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,n} \end{pmatrix}$$

$$x_{i,j} = \frac{1}{n-1} \sum_k^n (x_{k,i} - \bar{x}_i)(x_{k,j} - \bar{x}_j)$$

measure the degree to which two attributes vary together

if ~ 0 do not have a (linear) relation

but not direct measure degree of relationship

2 Correlation (or correlation coefficient r) matrix

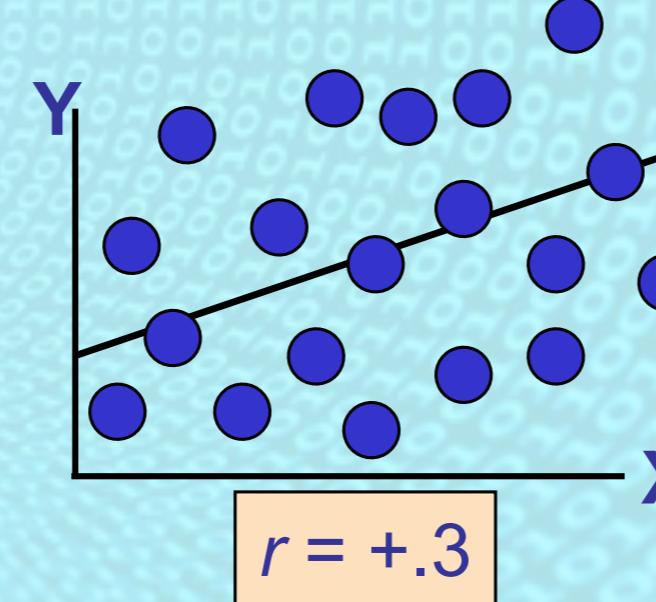
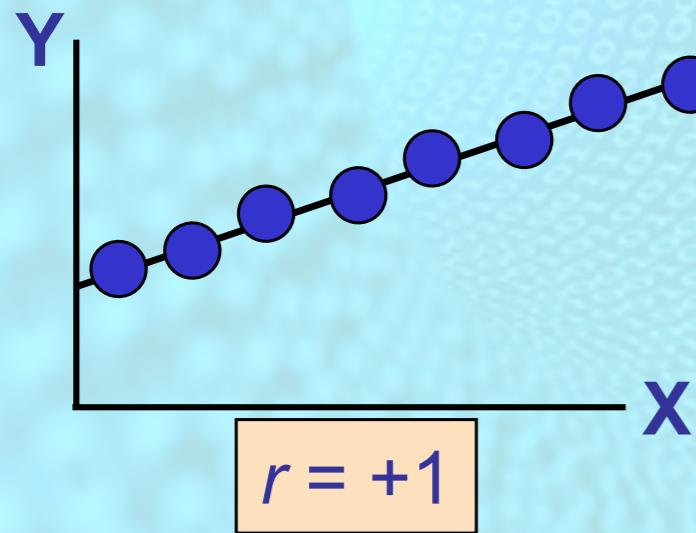
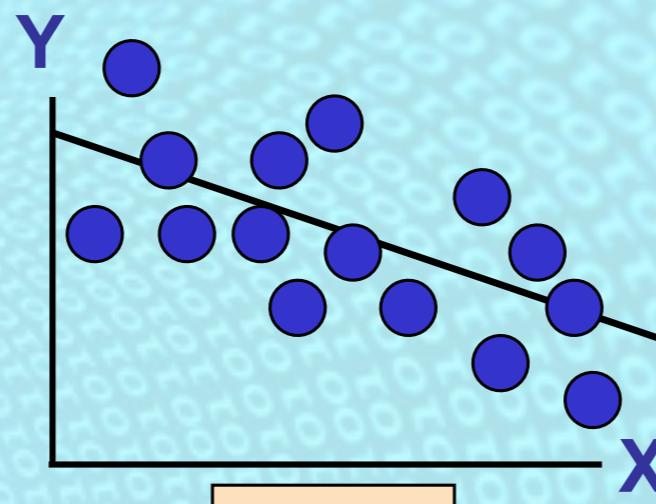
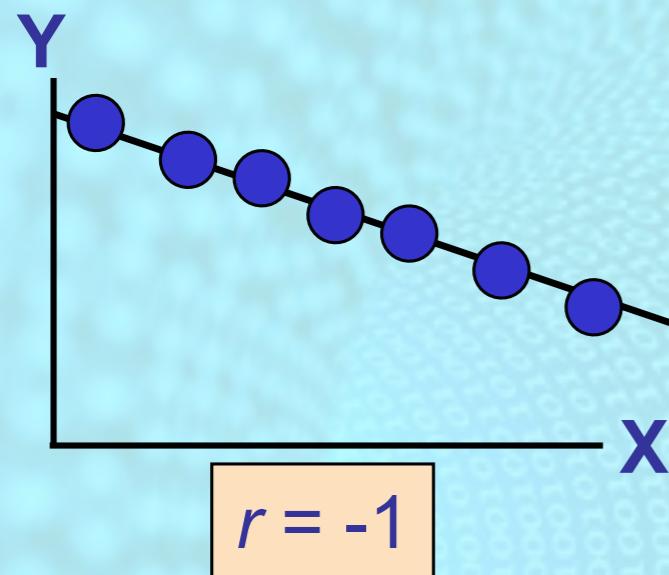
$$r_{i,j} = \frac{x_{i,j}}{\sigma_i \sigma_j}$$

gives immediate indication on how strongly two variables are (linearly) related

$$R = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & r_{2,2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,n} \end{pmatrix} \rightarrow \begin{pmatrix} 1 & r_{1,2} & \cdots & r_{1,n} \\ r_{2,1} & 1 & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & \end{pmatrix}$$

$$-1 < r_{ij} < 1$$

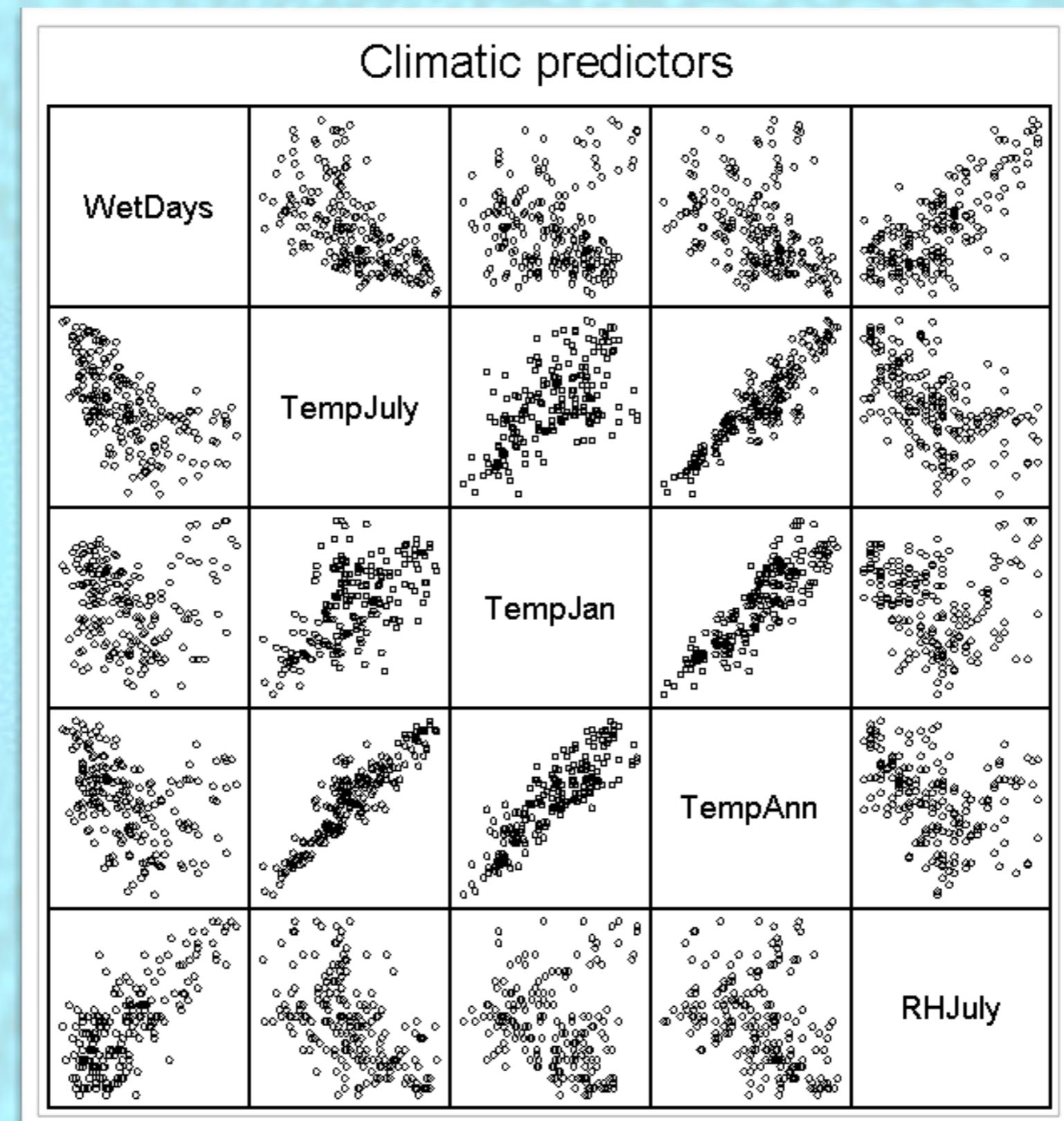
Correlation (r):



$r=0$ no correlation

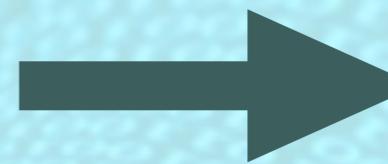
$r=1,-1$ max correlation

Multivariate analysis : correlation matrix



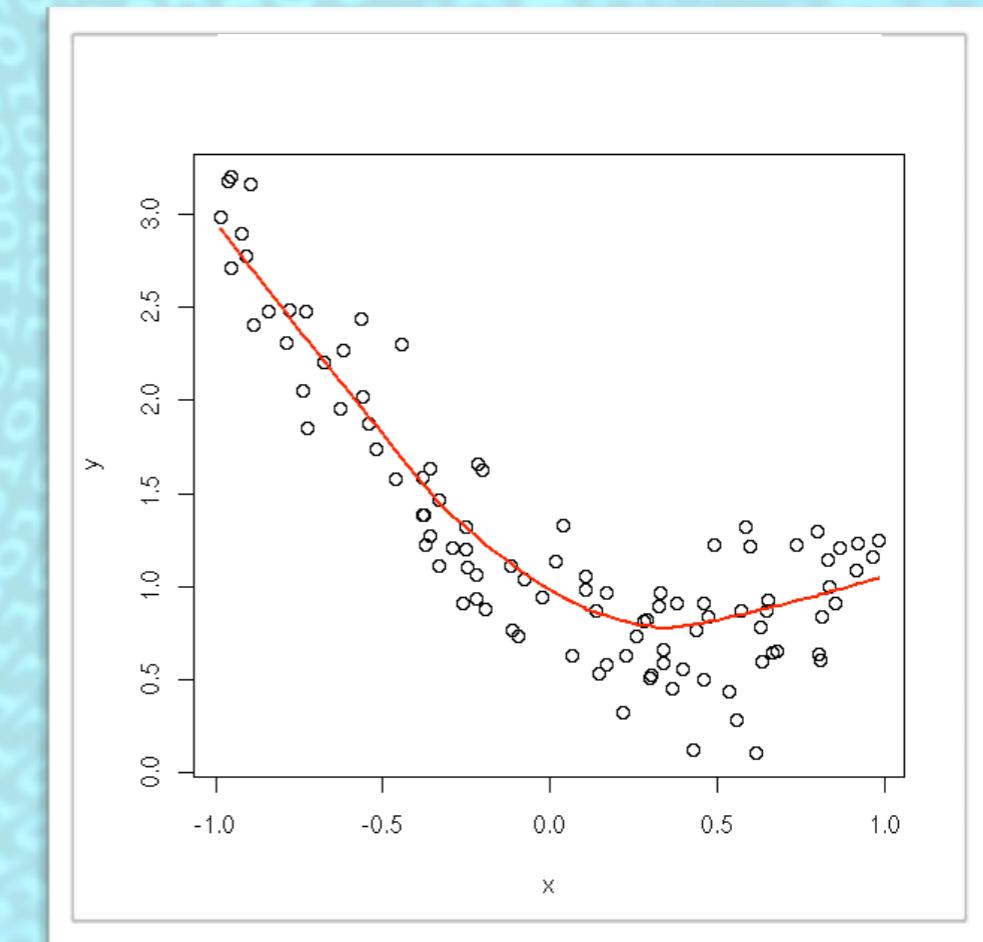
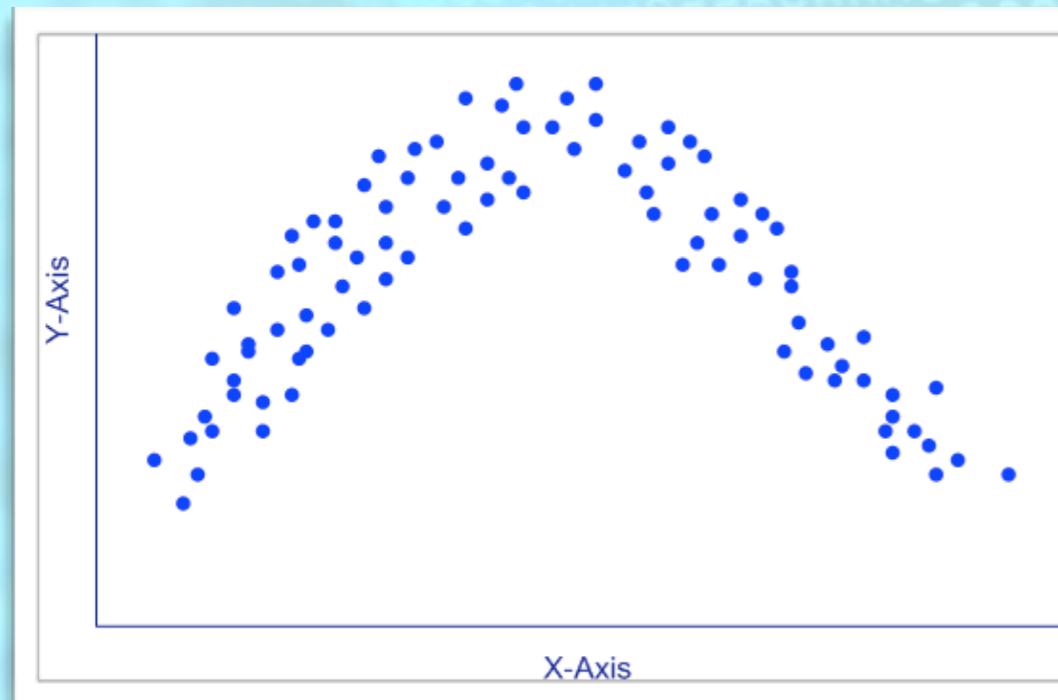
Pay attention:

correlation = 0



- 1) no linear correlation
- 2) does not imply causation

but you can have non-linear correlations...

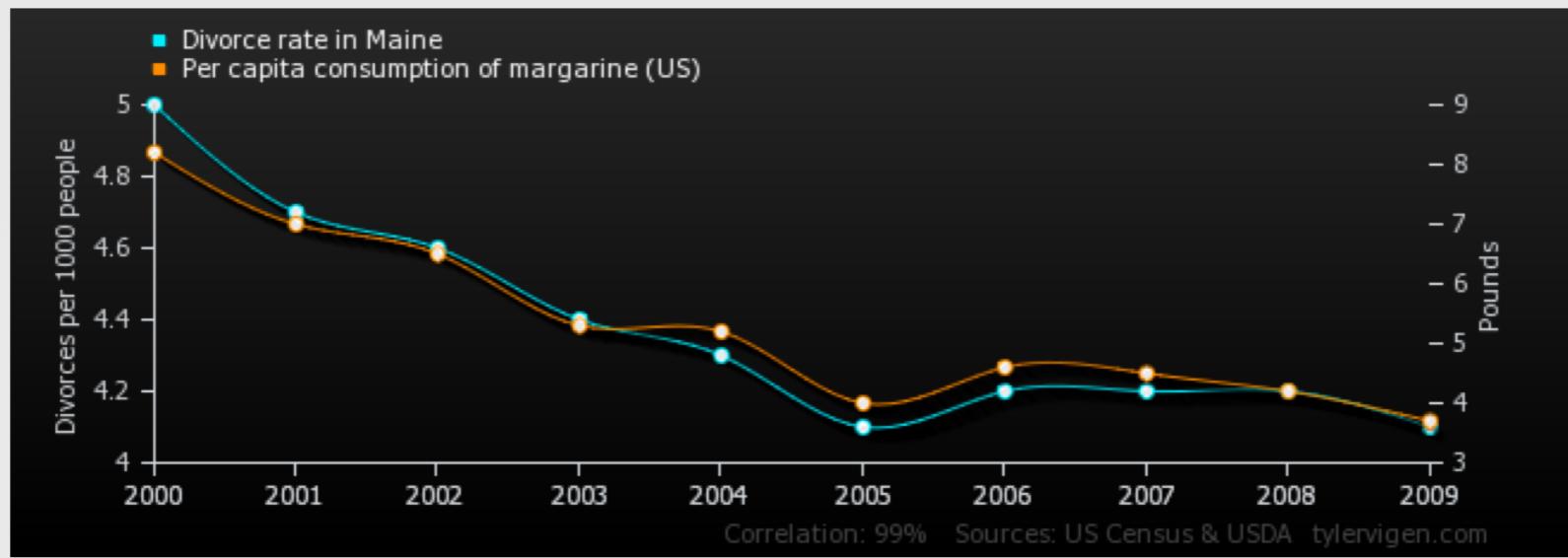


Correlation and causation

Divorce rate in Maine

correlates with

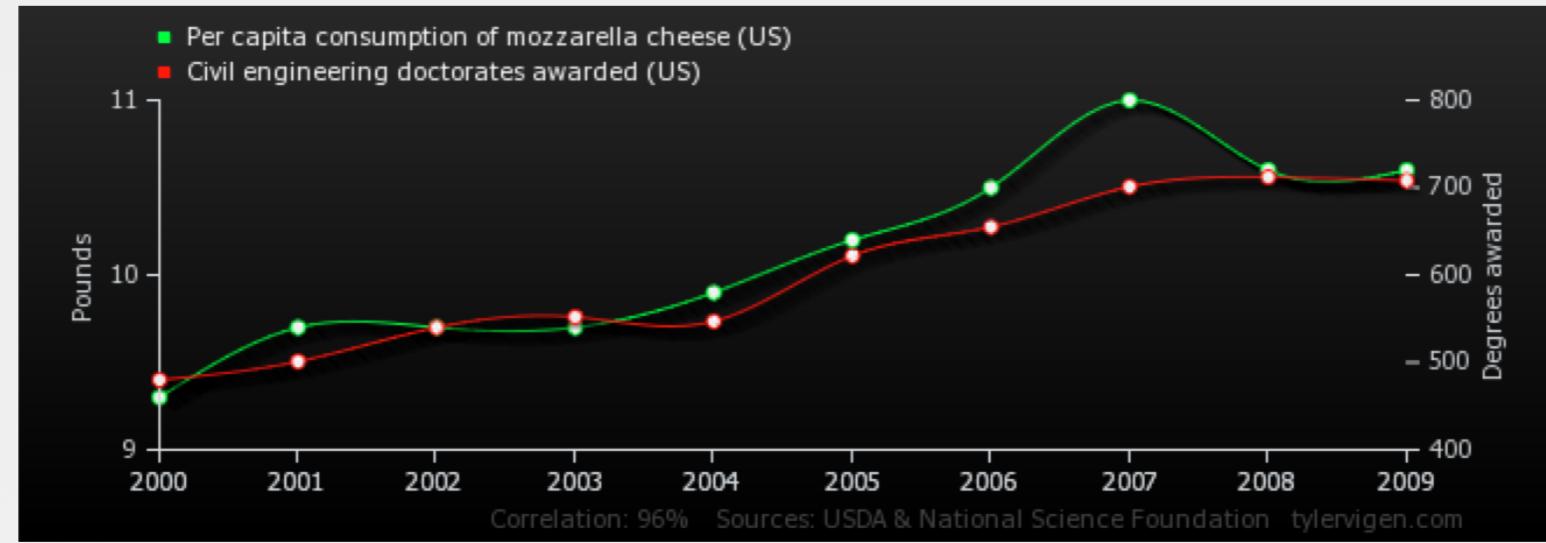
Per capita consumption of margarine (US)



Per capita consumption of mozzarella cheese (US)

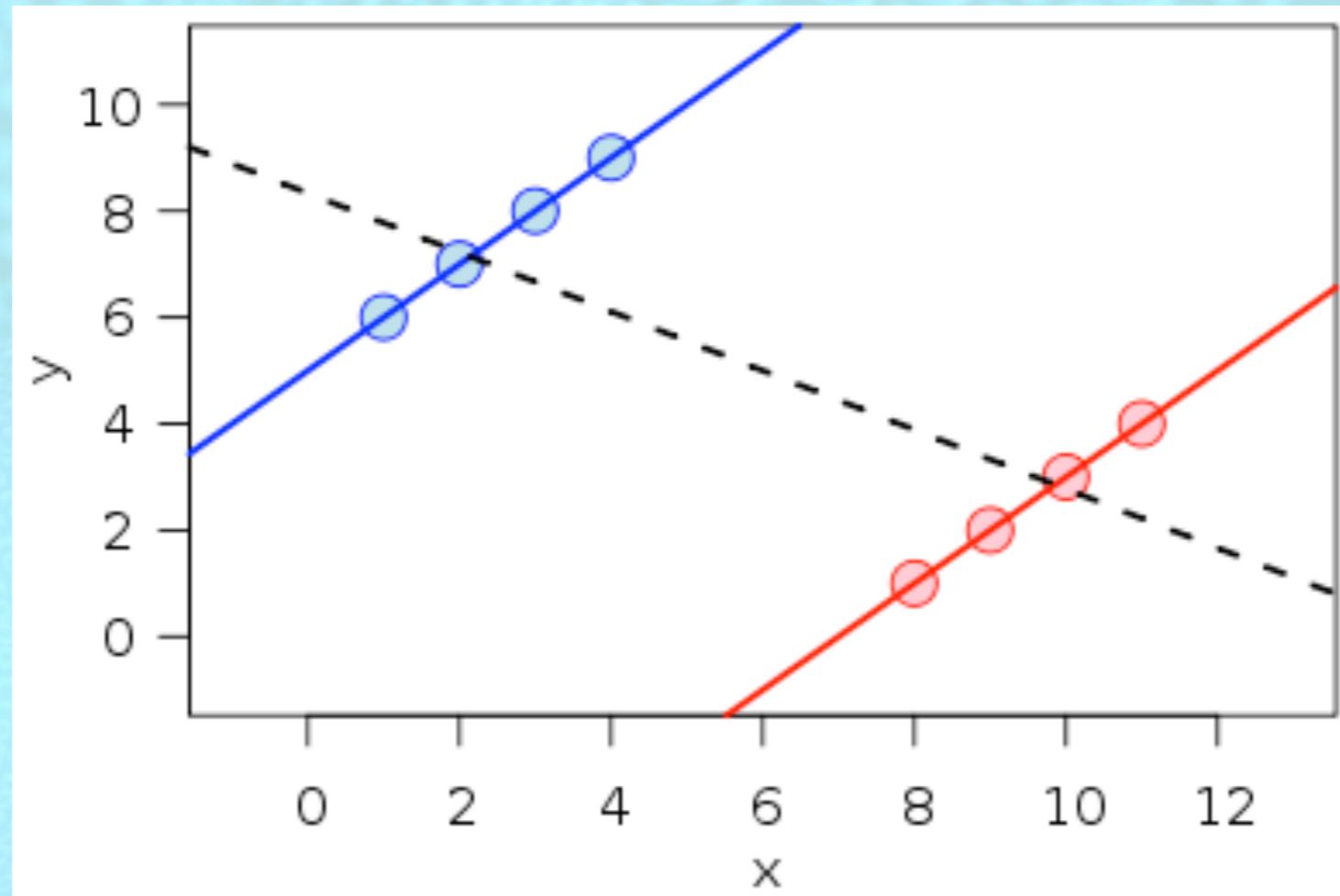
correlates with

Civil engineering doctorates awarded (US)



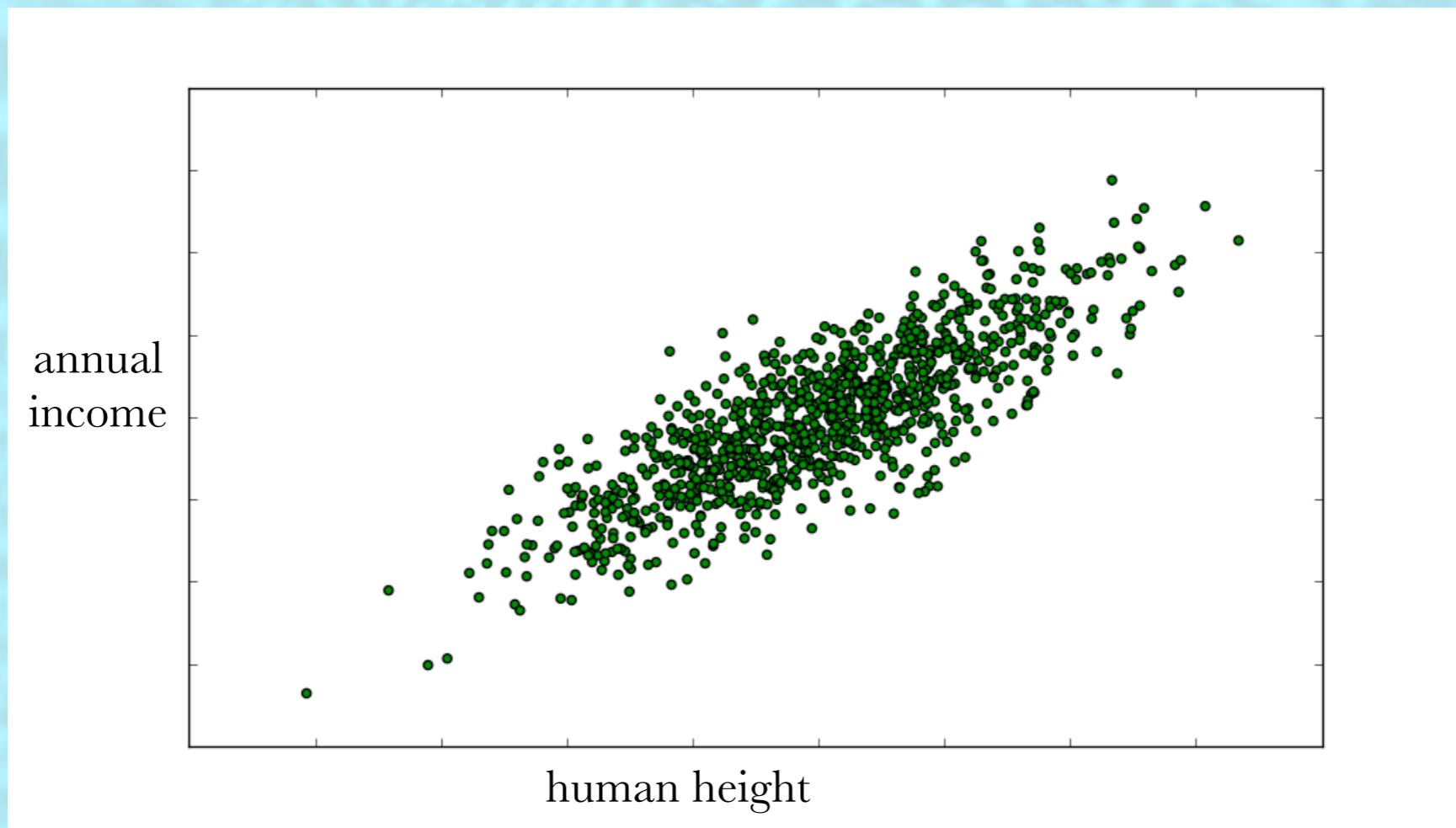
<http://tylervigen.com/page?page=1>

Simpson paradox

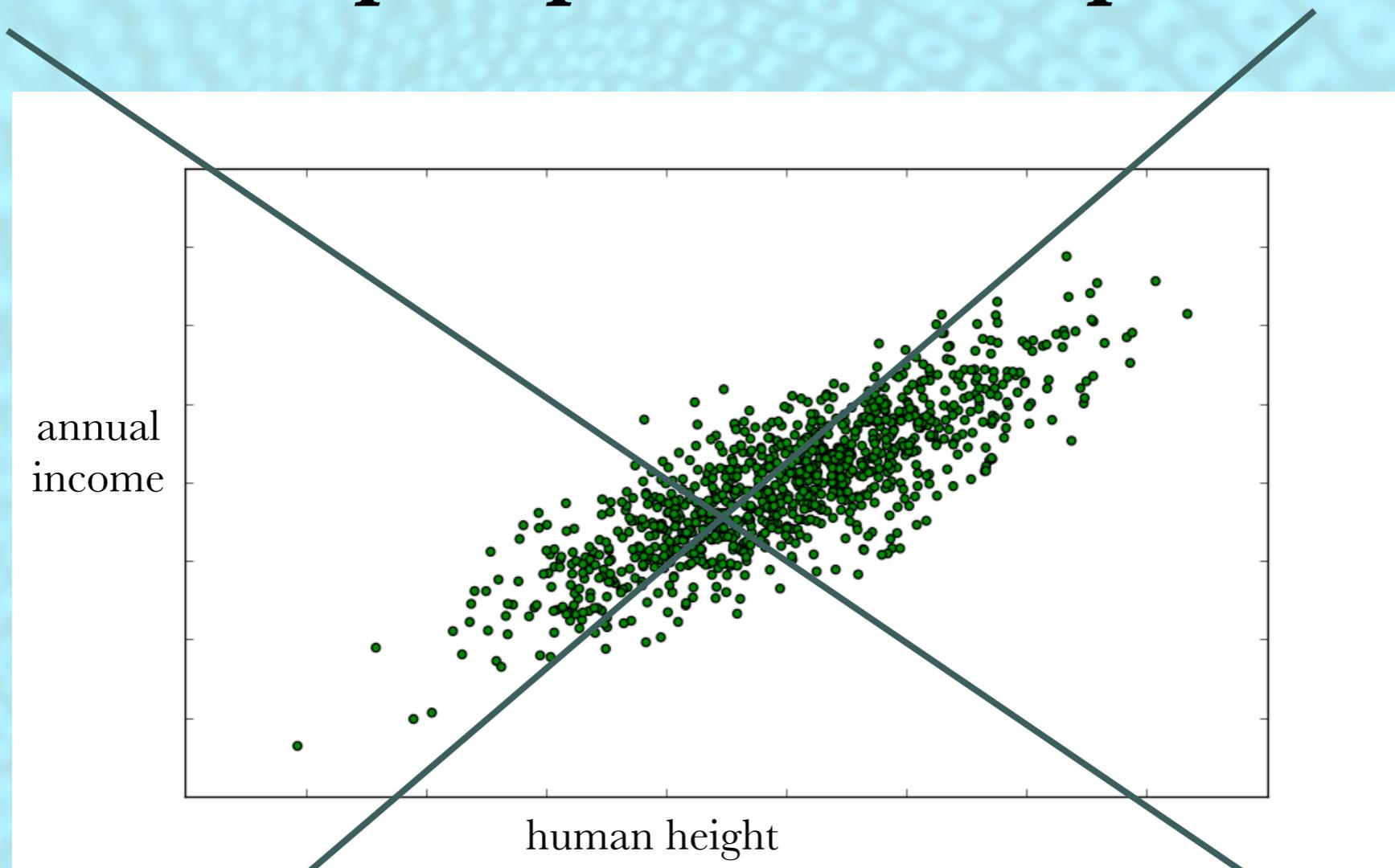


a trend appears in different groups of data but disappears or reverses when these groups are combined.

Simpson paradox: example

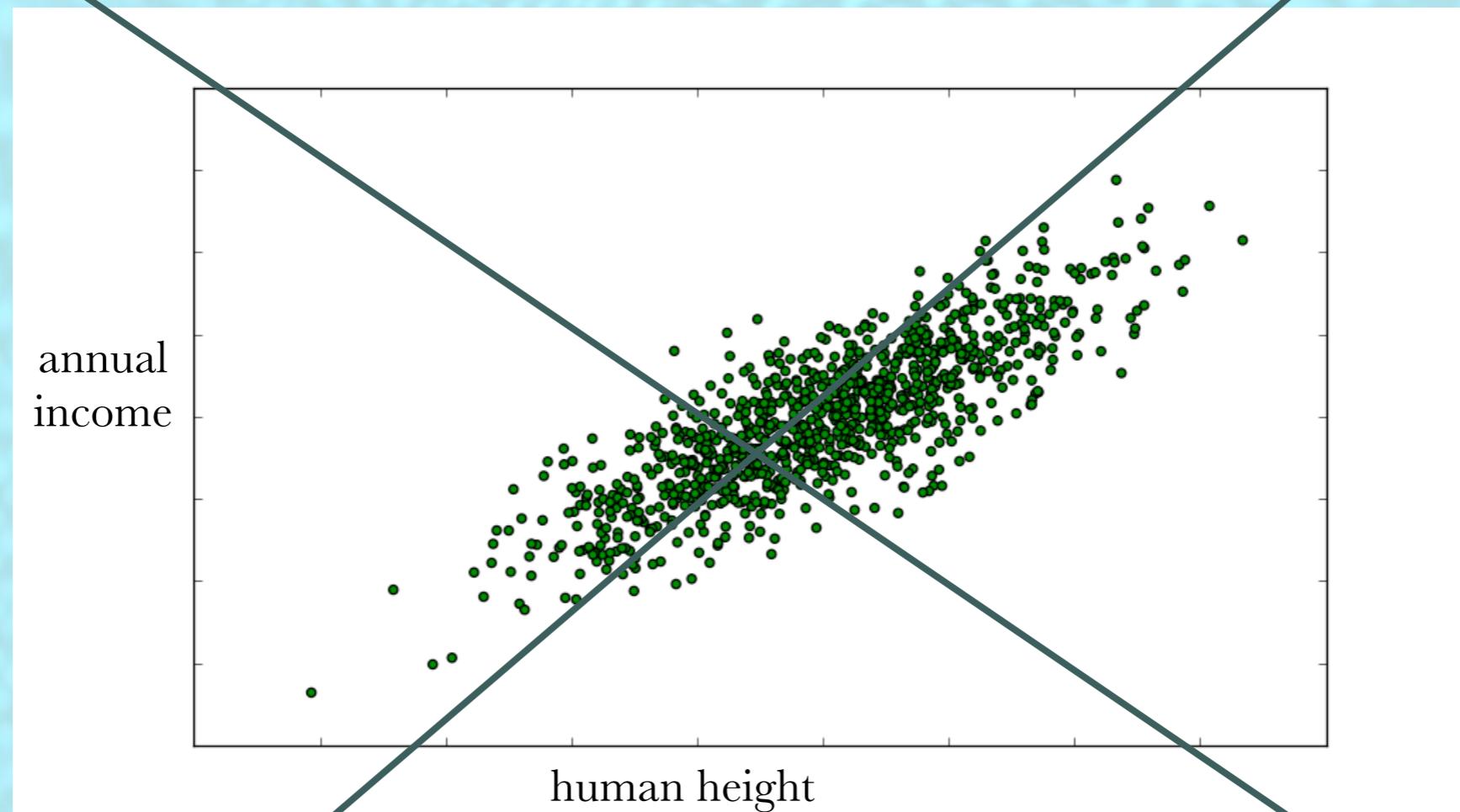


Simpson paradox: example



if I separate samples for country

Simpson paradox: example



*key issue: assumption in correlation of 2-variables,
all the others being equal*

solution? know your data

References

