Data mining & Machine Learning

CS 373
Purdue University

Dan Goldwasser
dgoldwas@purdue.edu

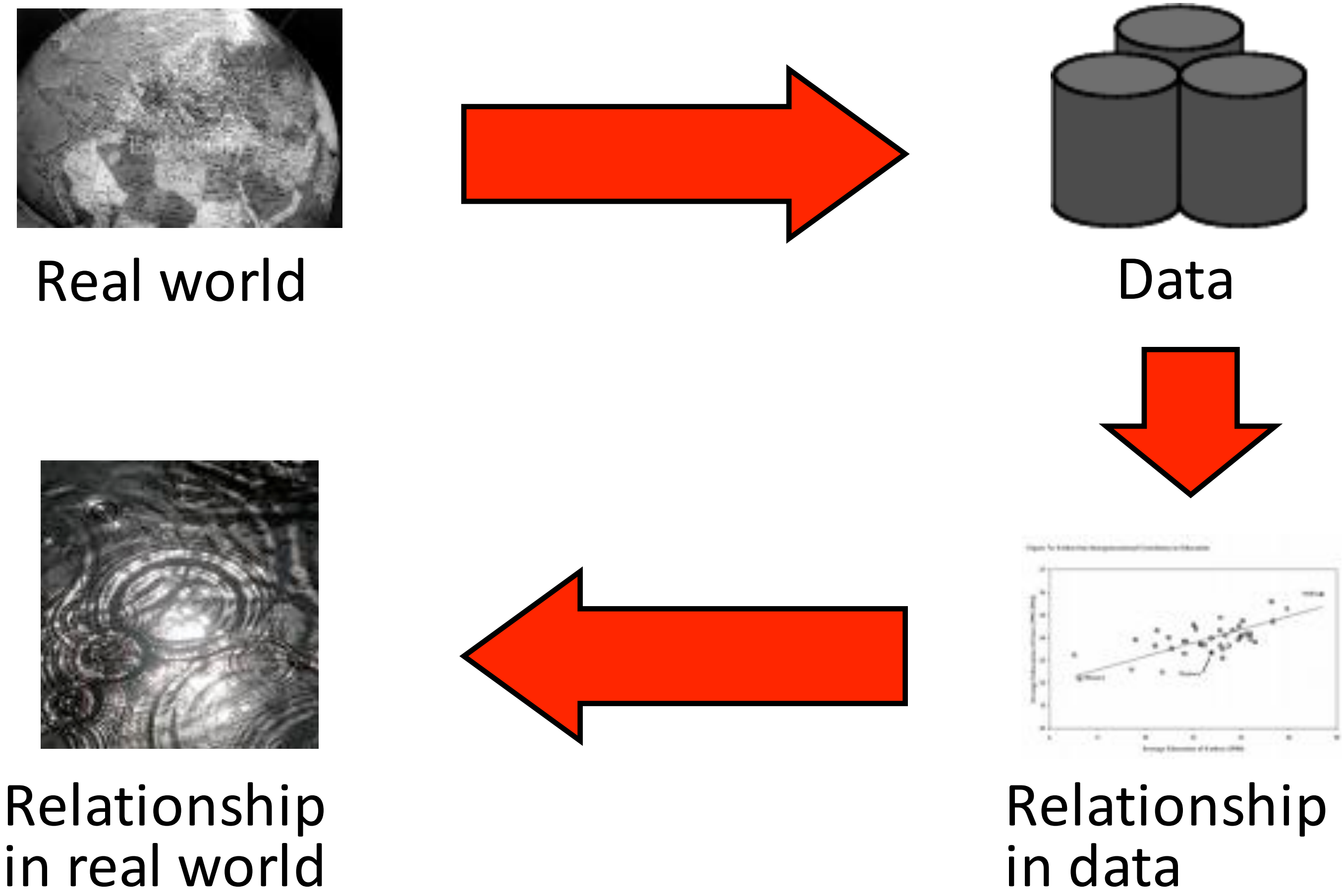
Today's Lecture

***DATA** is a big word*

- *What does it actually mean? What can we expect to find we collect data?*
- *In the age of big-data, how can we quickly "summarize" it?*
  - ***Find patterns, identify noise, etc.***

How can we answer questions using data?

Measurement



Real world

Data

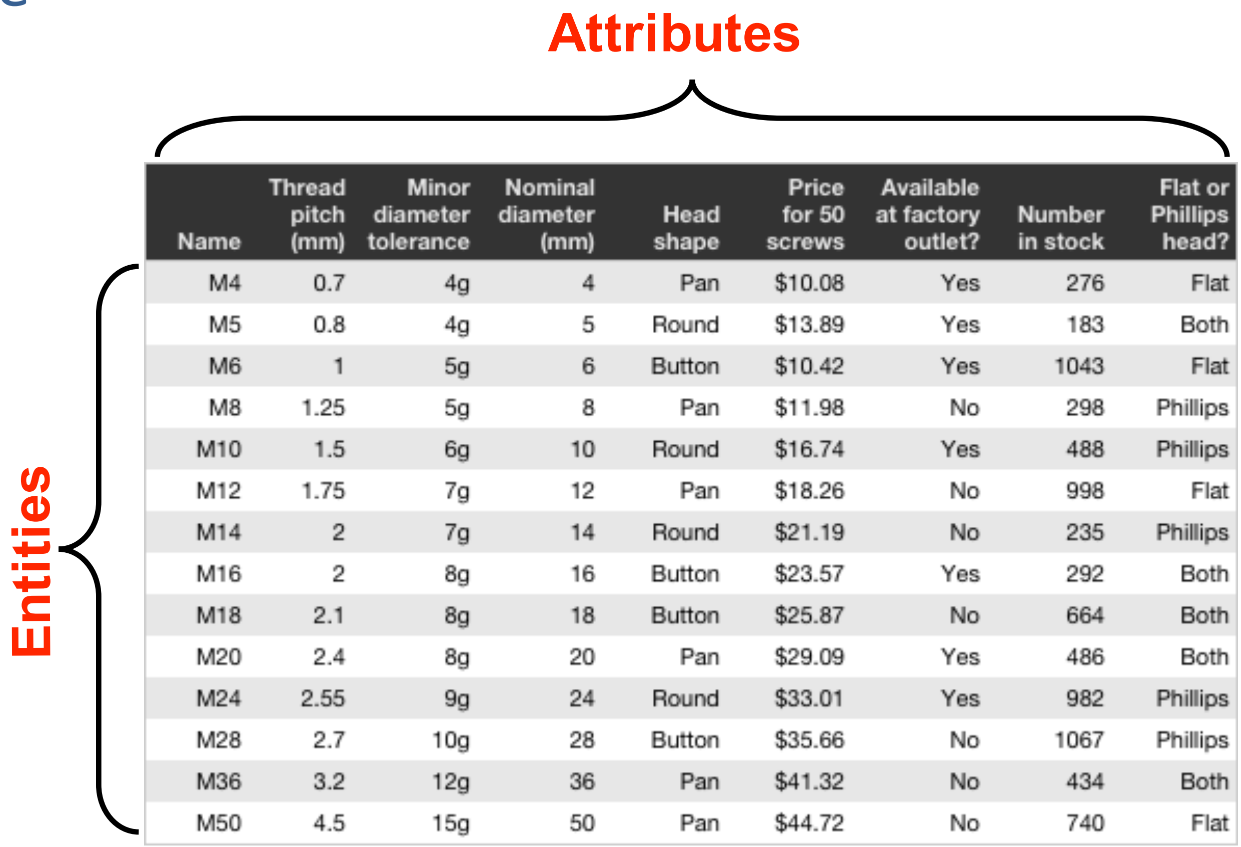
Relationship in data

Relationship in real world

Goal: map domain entities to symbolic representations

What is data?

- Collection of entities and their attributes
- **Attribute**: property or characteristic of an entity (e.g., eye color, temperature)
- **Entity**: collection of attributes
  Aka: record, point, case, sample, object, or instance



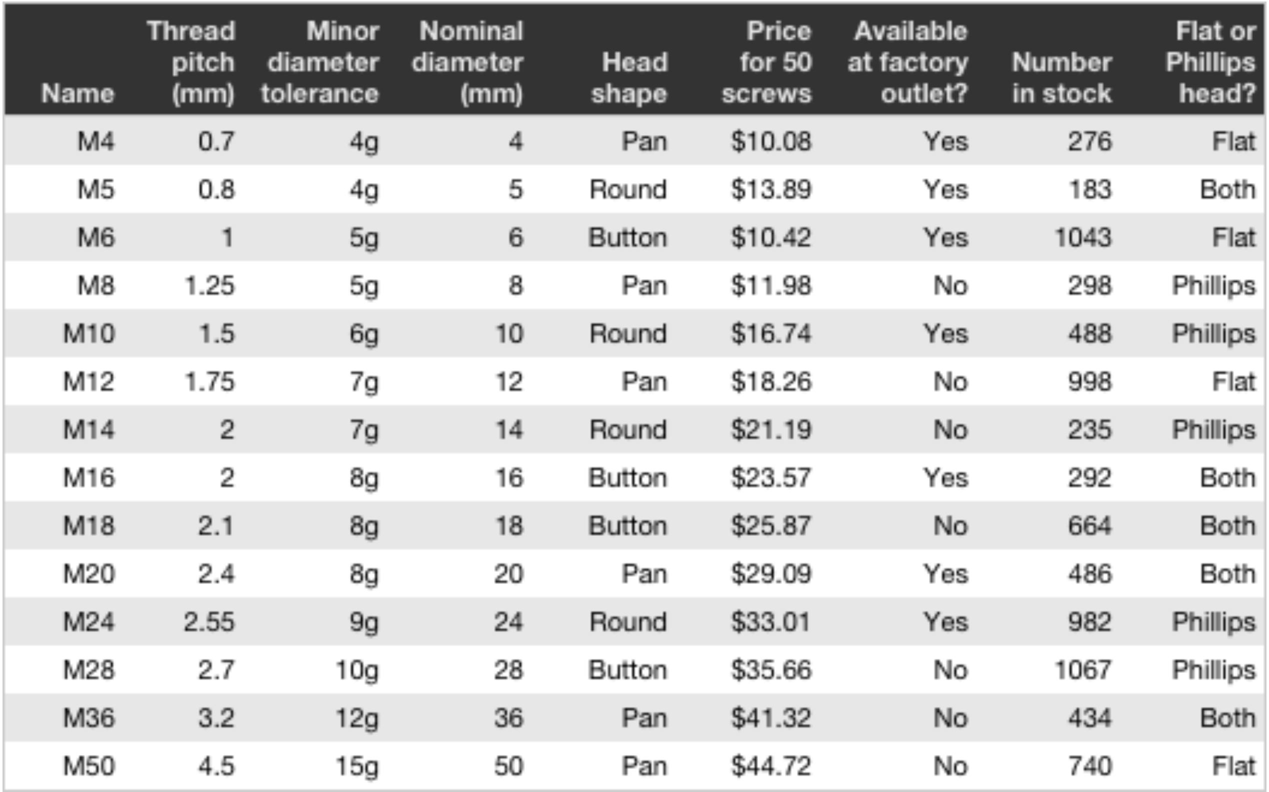
Attributes

Entities

Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Tabular data

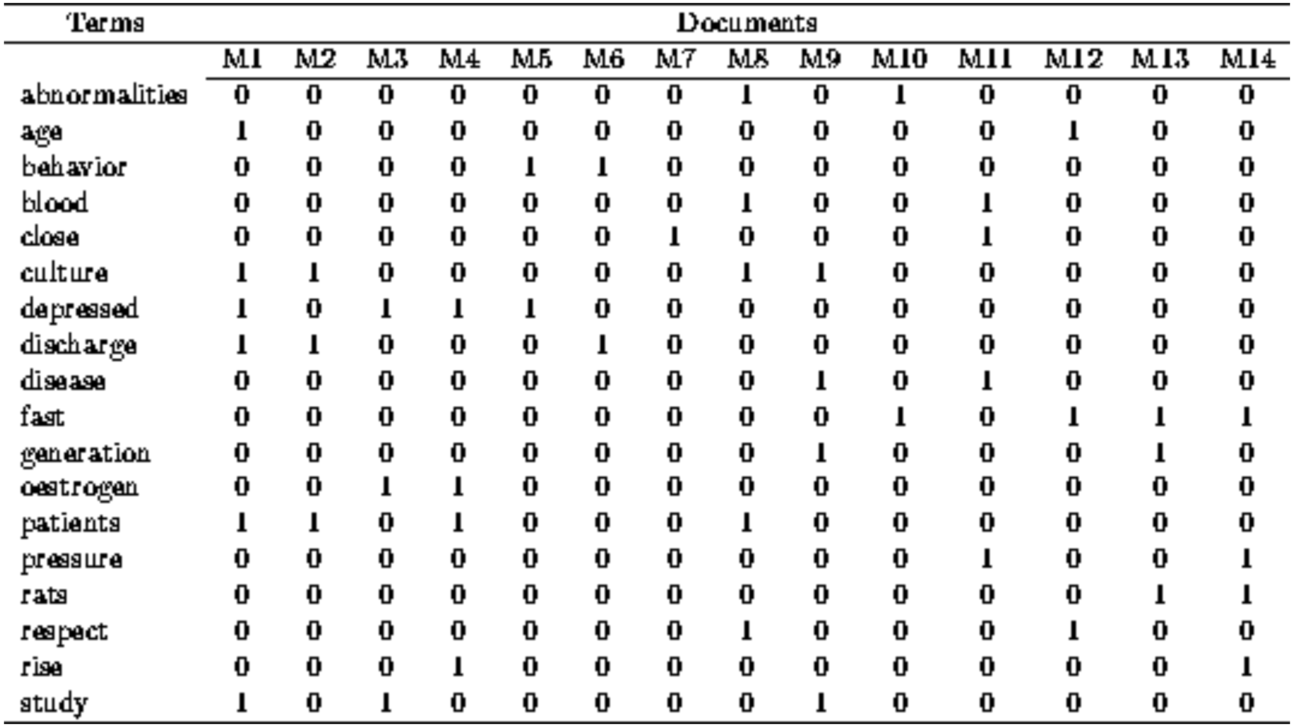
- Collection of records, each of which consists of a fixed set of attributes



Name	Thread pitch (mm)	Minor diameter tolerance	Nominal diameter (mm)	Head shape	Price for 50 screws	Available at factory outlet?	Number in stock	Flat or Phillips head?
M4	0.7	4g	4	Pan	$10.08	Yes	276	Flat
M5	0.8	4g	5	Round	$13.89	Yes	183	Both
M6	1	5g	6	Button	$10.42	Yes	1043	Flat
M8	1.25	5g	8	Pan	$11.98	No	298	Phillips
M10	1.5	6g	10	Round	$16.74	Yes	488	Phillips
M12	1.75	7g	12	Pan	$18.26	No	998	Flat
M14	2	7g	14	Round	$21.19	No	235	Phillips
M16	2	8g	16	Button	$23.57	Yes	292	Both
M18	2.1	8g	18	Button	$25.87	No	664	Both
M20	2.4	8g	20	Pan	$29.09	Yes	486	Both
M24	2.55	9g	24	Round	$33.01	Yes	982	Phillips
M28	2.7	10g	28	Button	$35.66	No	1067	Phillips
M36	3.2	12g	36	Pan	$41.32	No	434	Both
M50	4.5	15g	50	Pan	$44.72	No	740	Flat

Document data

- Each document is represented as a **term** vector, where each attribute records the number of times the term occurs in the document



Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Transaction data

- Each record corresponds to a transaction involving a set of items
- E.g., in a grocery store purchase, the set of products purchased by a customer constitute a transaction, while the individual products that were purchased are the items

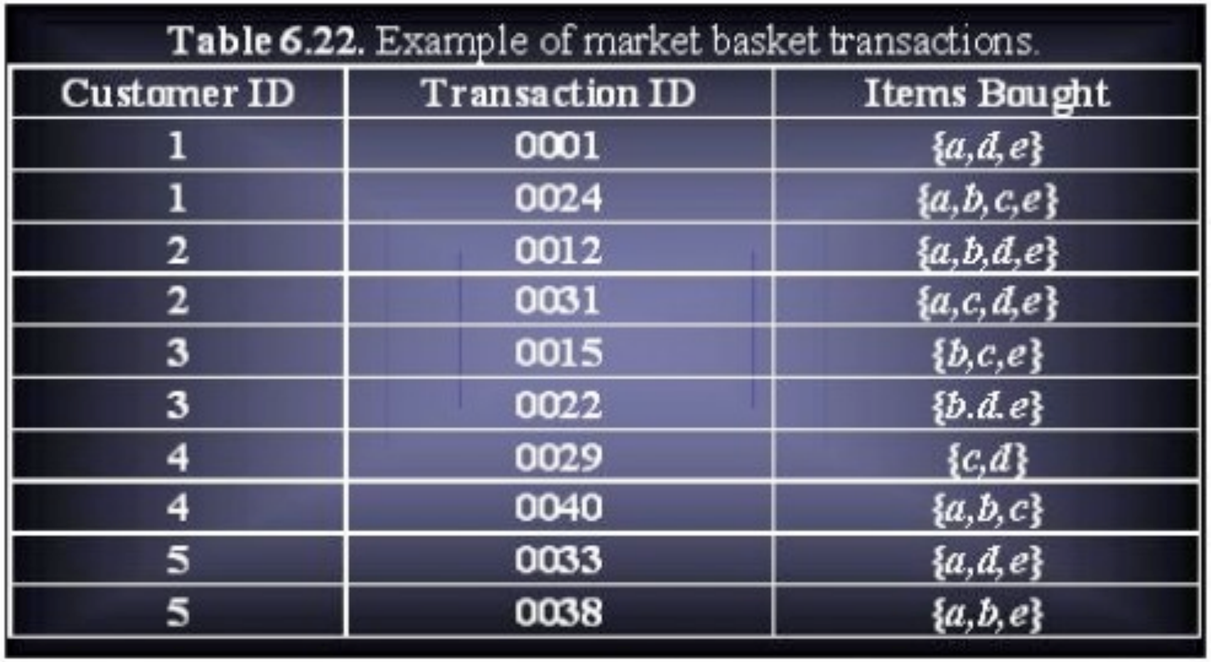


Table 6.22. Example of market basket transactions.

Customer ID	Transaction ID	Items Bought
1	0001	{a,d,e}
1	0024	{a,b,c,e}
2	0012	{a,b,d,e}
2	0031	{a,c,d,e}
3	0015	{b,c,e}
3	0022	{b.d.e}
4	0029	{c,d}
4	0040	{a,b,c}
5	0033	{a,d,e}
5	0038	{a,b,e}

Ordered data

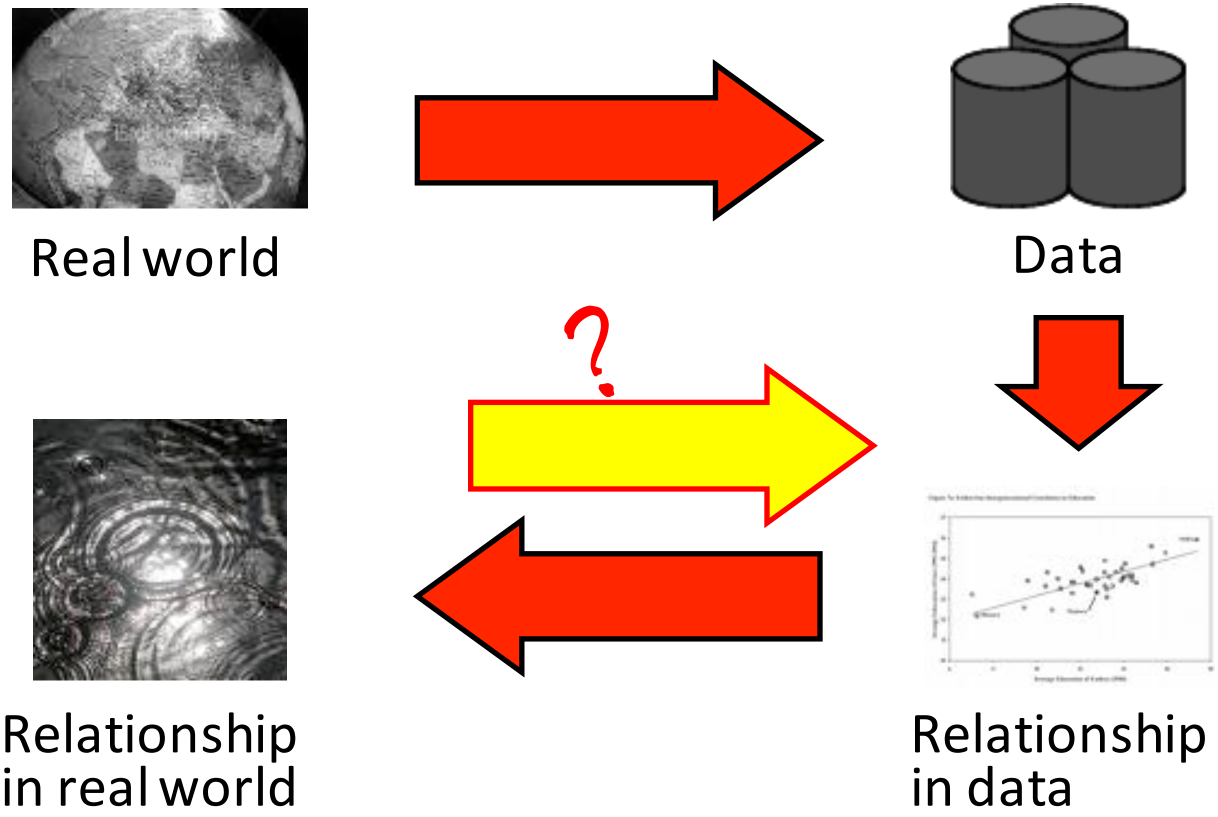
- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Graph data

- Nodes correspond to entities, edges correspond to relationships
- E.g.: Web graph with HTML links, molecules with atoms and bonds

Measurement



Real world

Data

?

Relationship in real world

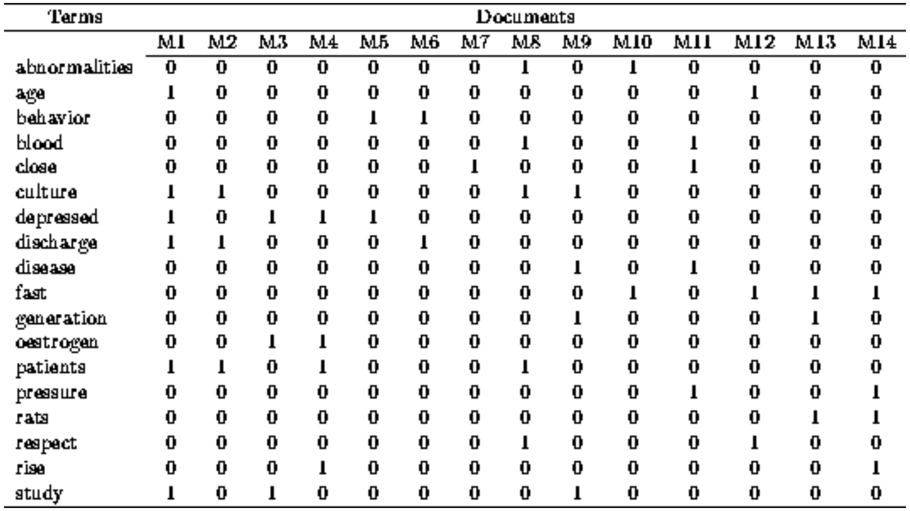
Relationship in data

Gc tations

Does the data representation provide the appropriate abstraction for answering questions about the real world?

Document Data

Document =
words frequencies



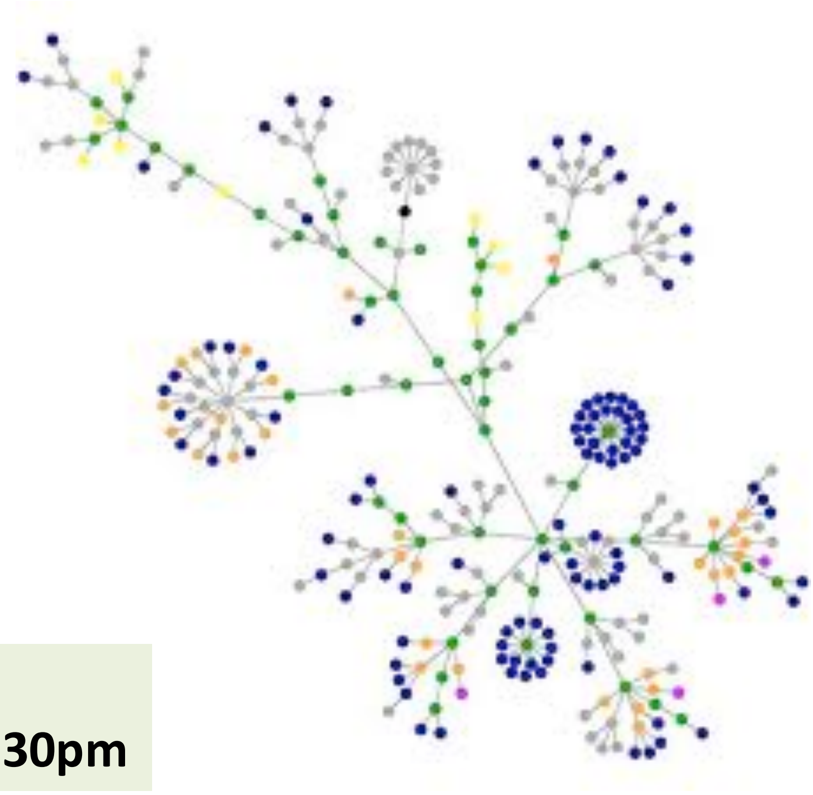
Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Task 1: based on this representation, identify the “hot topics” in the media, in the span of a month.

Task 2: based on this representation, identify the general sentiment about the new IPhone in the 12 hours after its release

Take Home Quiz

- Nodes are users in a social network, edges represent interactions between users.
- Edges are weighted as follows:
  - No interaction: no edge
  - Otherwise – edge weight: # interactions.



Will be posted on Piazza.
Please respond (privately) until **Monday 12:30pm**

Take Home Quiz

- You should:
  - *Find one example of a question (task) that can be answered using this representation, and one that cannot.*
  - *How would you modify the network representation to answer the second question?*
    - *What should you consider when changing the representation? What are the tradeoffs involved?*

Will be posted on Piazza.
Please respond (privately) until **Monday 12:30pm**

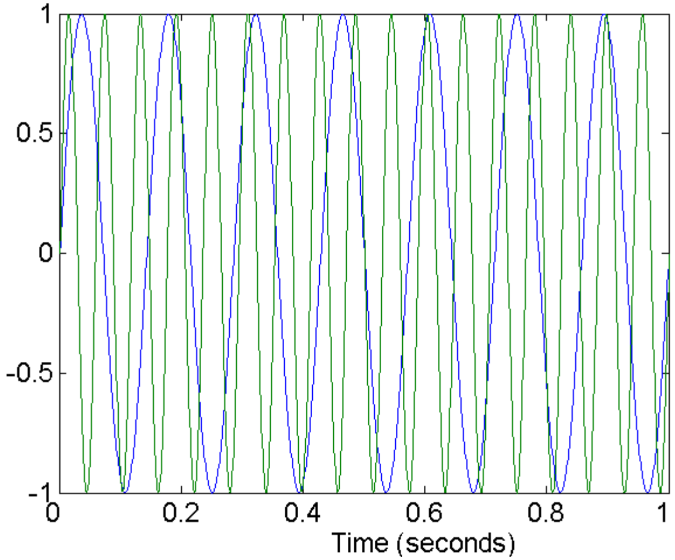
Data quality

- Examples of data quality problems:
  - Noise
  - Outliers
  - Missing values
  - Duplicate data

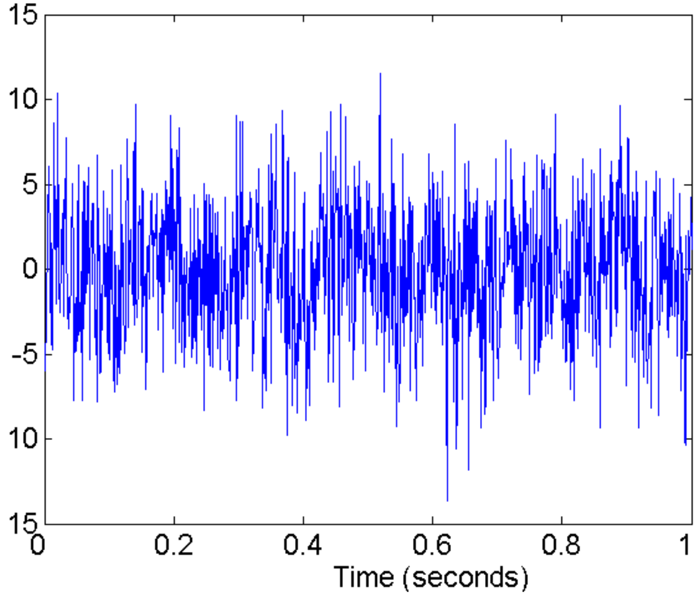
Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Noise

- Noise refers to measurement error in data values
  - Could be **random** error or **systematic** error



Two Sine Waves



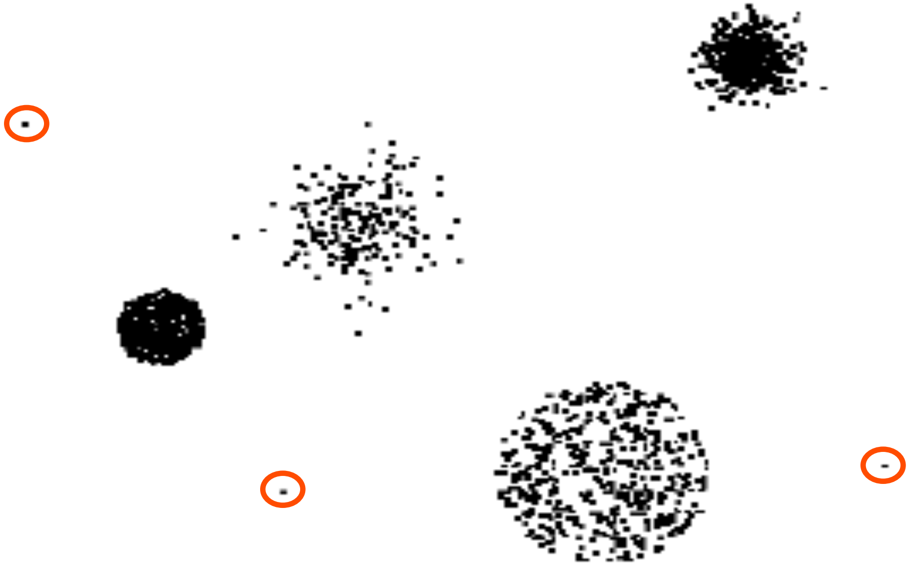
Two Sine Waves + Noise

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

- Could indicate "interesting" cases, or could indicate errors in the data
  - – **Should we care?**



Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Missing values

- **Reasons for missing values**
  - Information is not collected (e.g., decline to give their age)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- **Ways to handle missing values**
  - Eliminate entities with missing values
  - Estimate attributes with missing values
  - Ignore the missing values during analysis
  - Replace with all possible values (weighted by their probabilities)

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Duplicate data

- Data set may include data entities that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
  - Example: same person with multiple email addresses
- **Data cleaning**
  - Finding and dealing with duplicate entities
  - Finding and correcting measurement error
  - Dealing with missing values

Tan, Steinbach, Kumar. Introduction to Data Mining, 2004.

Other data preprocessing methods

- Sampling
- Attribute transformations
  - Discretization, distance calculations
  - **Feature construction**
- Dimensionality reduction and feature selection
- Recent trend: ***Representation Learning***

Data exploration and visualization

Exploratory data analysis

- Data analysis approach that employs a number of (mostly graphical) techniques to:
  - Maximize insight into data
  - Uncover underlying structure
  - Identify important variables
  - Detect outliers and anomalies
  - Test underlying modeling assumptions
  - Develop parsimonious models
  - **Generate hypotheses from data**

Visualization

- Human eye/brain have evolved powerful methods to detect structure in nature

- Display data in ways that exploit human pattern recognition abilities

- **Limitation**: Can be difficult to apply if data size (number of dimensions or instances) is large

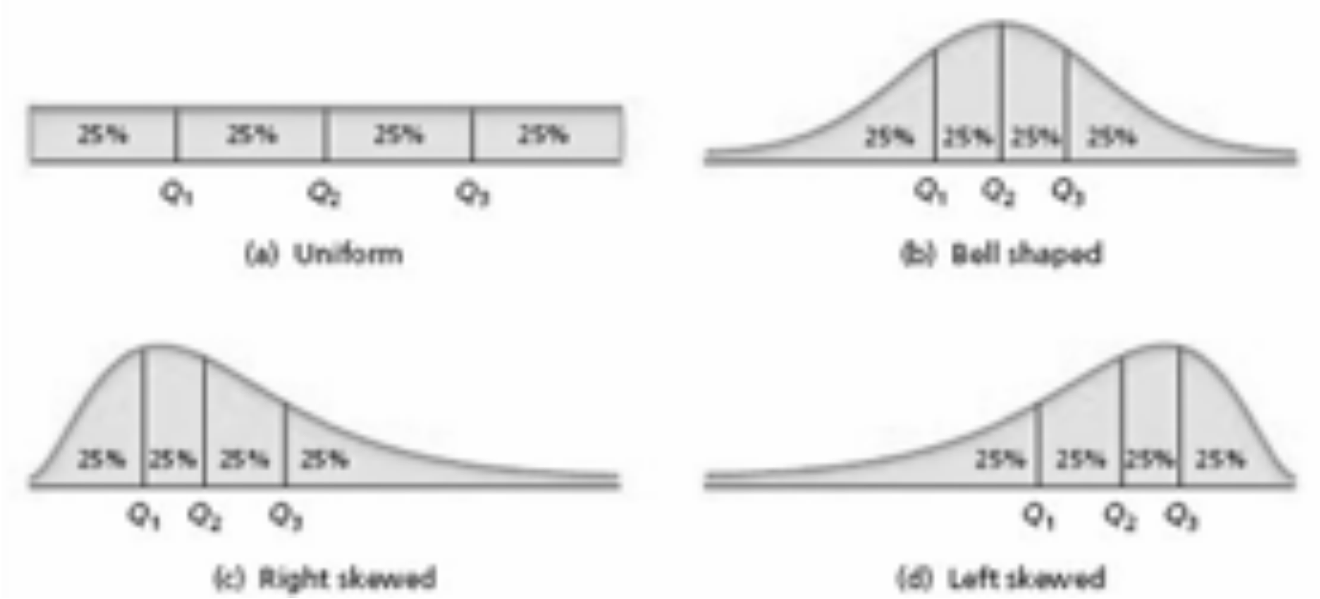
Visualizing/summarizing data

- **Low-dimensional data**
  - Summarizing data with simple statistics
  - Plotting raw data (1D, 2D, 3D)

- **Higher-dimensional data**
  - Principal component analysis
  - Multidimensional scaling

Data summarization

- **Measures of location**
  - **Mean**: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x(i)$
  - **Median**: value with 50% of points above and below
  - **Quartile**: value with 25% (75%) points above and below
  - **Mode**: most common value

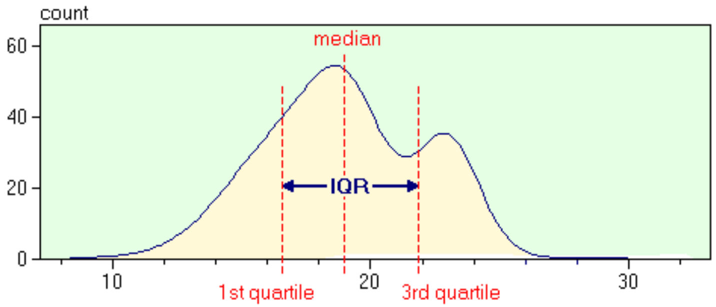


Data summarization

- Measures of dispersion or variability
  - **Variance**: $\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2$
  - **Standard deviation**:

$$\hat{\sigma}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x(i) - \mu)^2}$$

  - **Range**: difference between max and min point
  - **Interquartile range**: difference between 1st and 3rd Quartiles

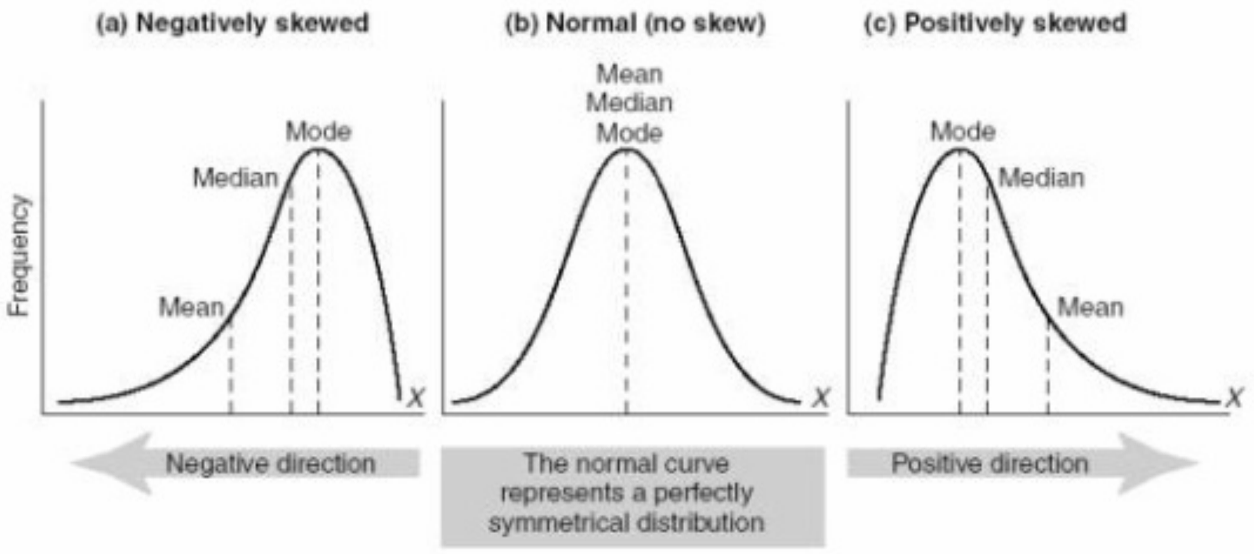


www.statistics4u.com

Data summarization

Skew: Measure of the asymmetry of a distribution

$$\mathbf{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\sum_{i=1}^{n}(x(i)-\hat{\mu})^3}{\left(\sum_{i=1}^{n}(x(i)-\hat{\mu})^2\right)^{\frac{3}{2}}}$$

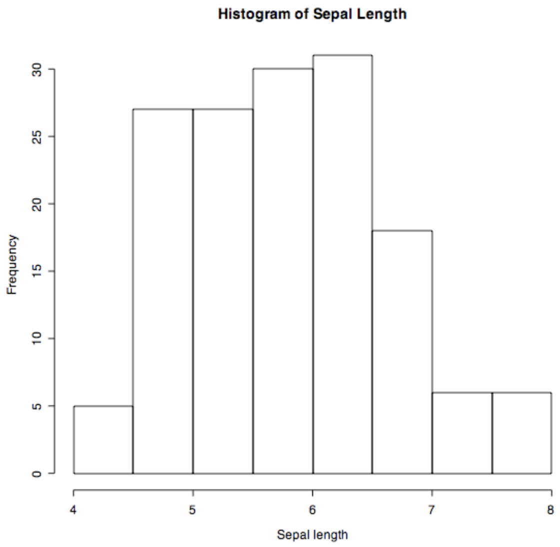


www.statistics4u.com

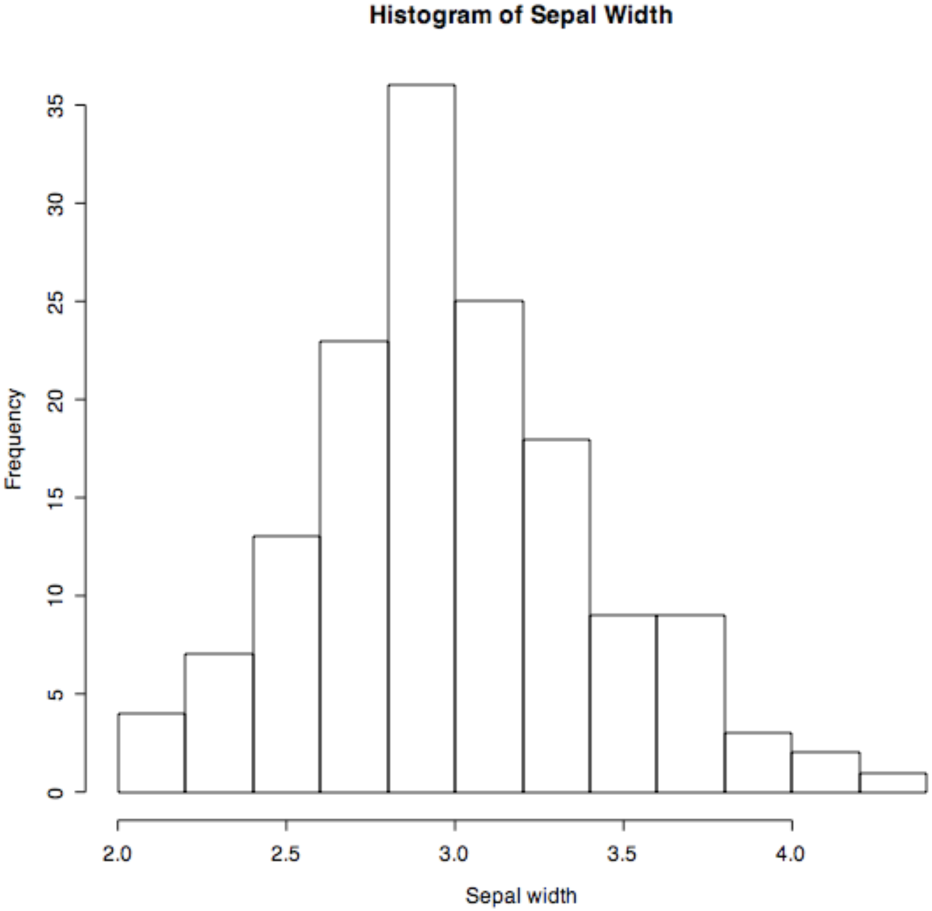
Histograms (1D)

- Most common plot for univariate data
- Split data range into equal-sized bins, count number of data points that fall into each bin

- **Graphically shows:**
  - Center (location)
  - Spread (scale)
  - Skew
  - Outliers
  - Multiple modes



Example histogram

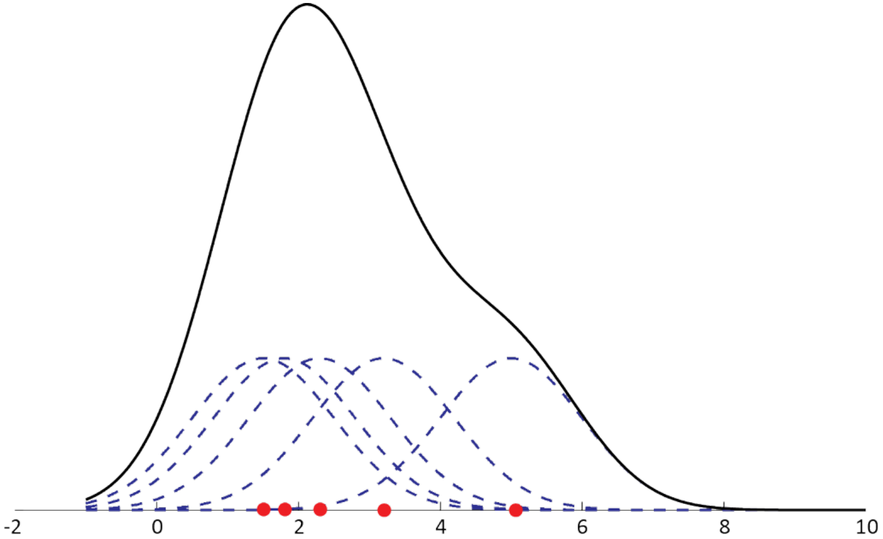


Fun fact!

normal distribution is the distribution that occurs most often in nature

Histogram limitations

- Histograms can be misleading for small datasets
  - Slight changes in the data or binning approach can result in different histograms
- **Solution**: *smoothed density plots*
  - Use kernel function to estimate density at each point x, pools information from neighboring points



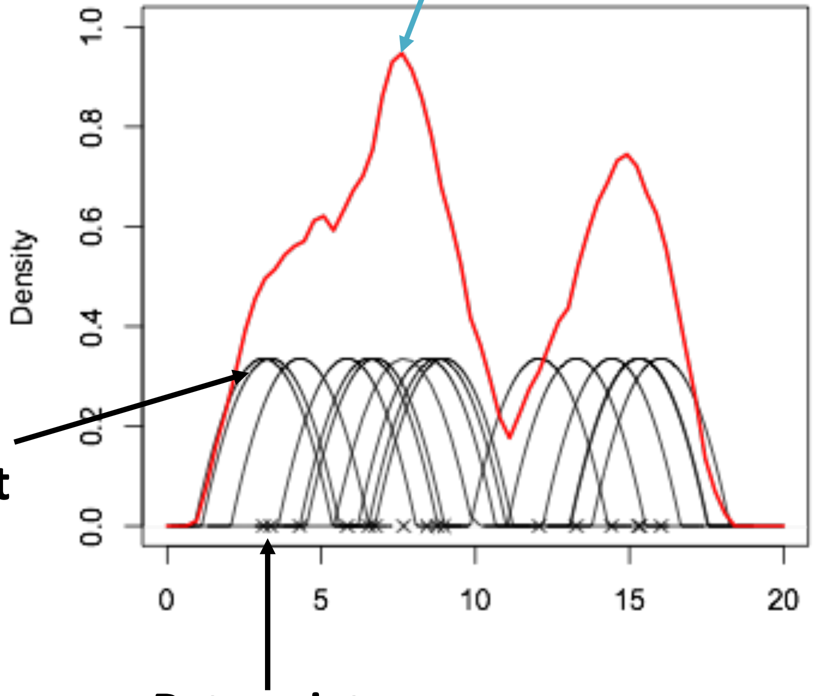
Density plots

- Estimated density is:

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

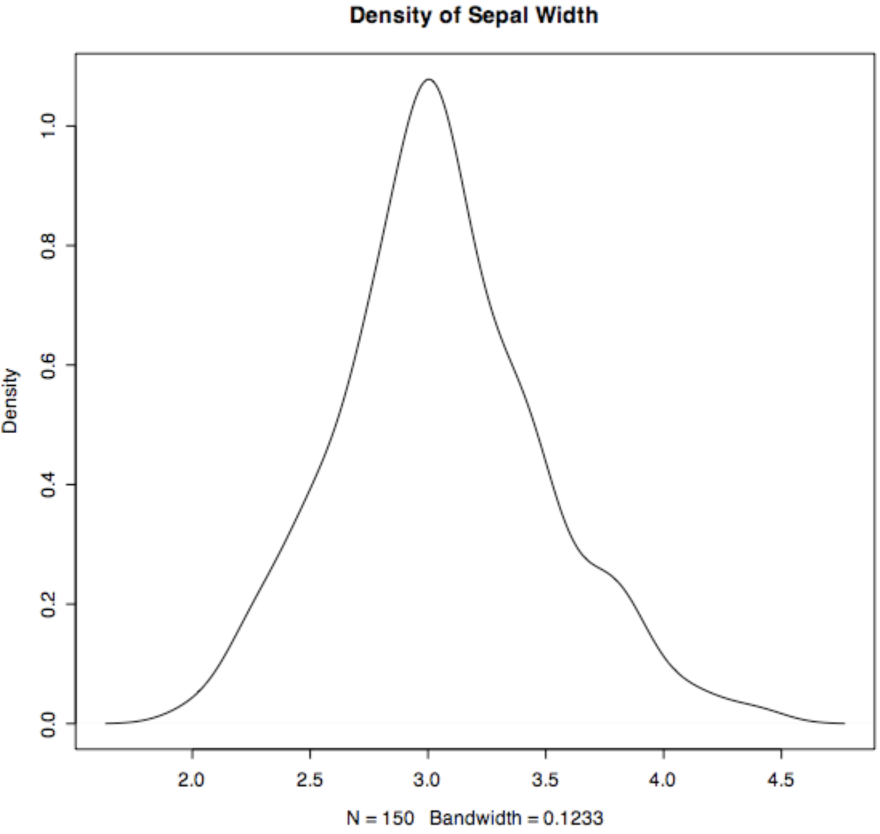
- **Two parameters:**
  - Kernel function K (e.g., Gaussian, Epanechnikov)
  - Bandwidth h

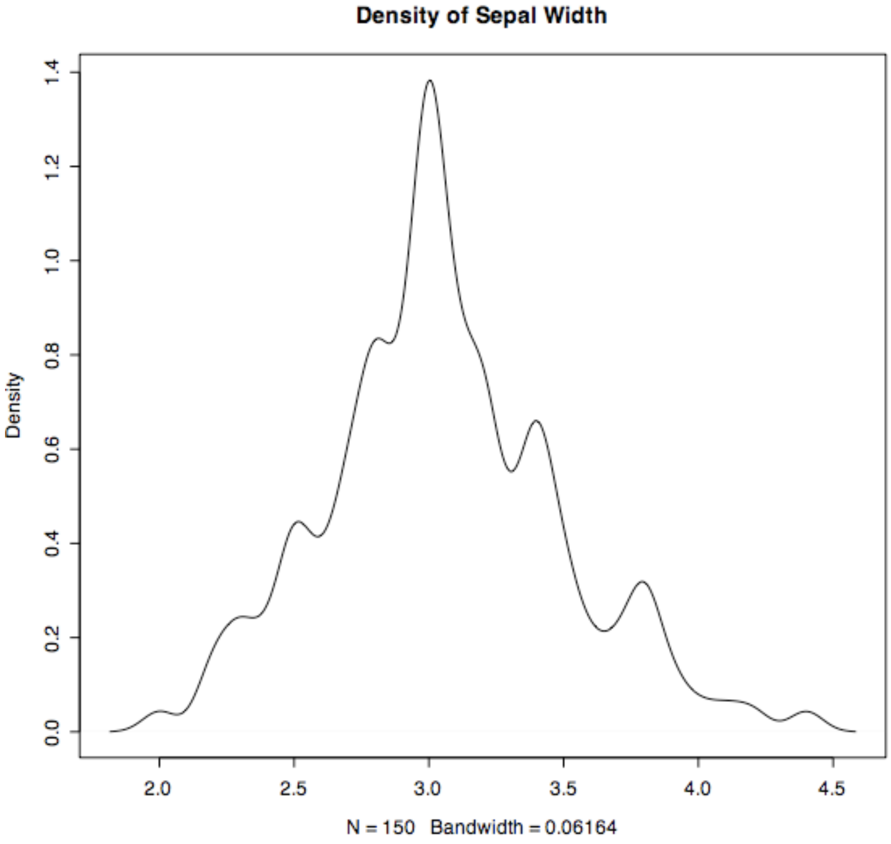
Density from sum of kernels



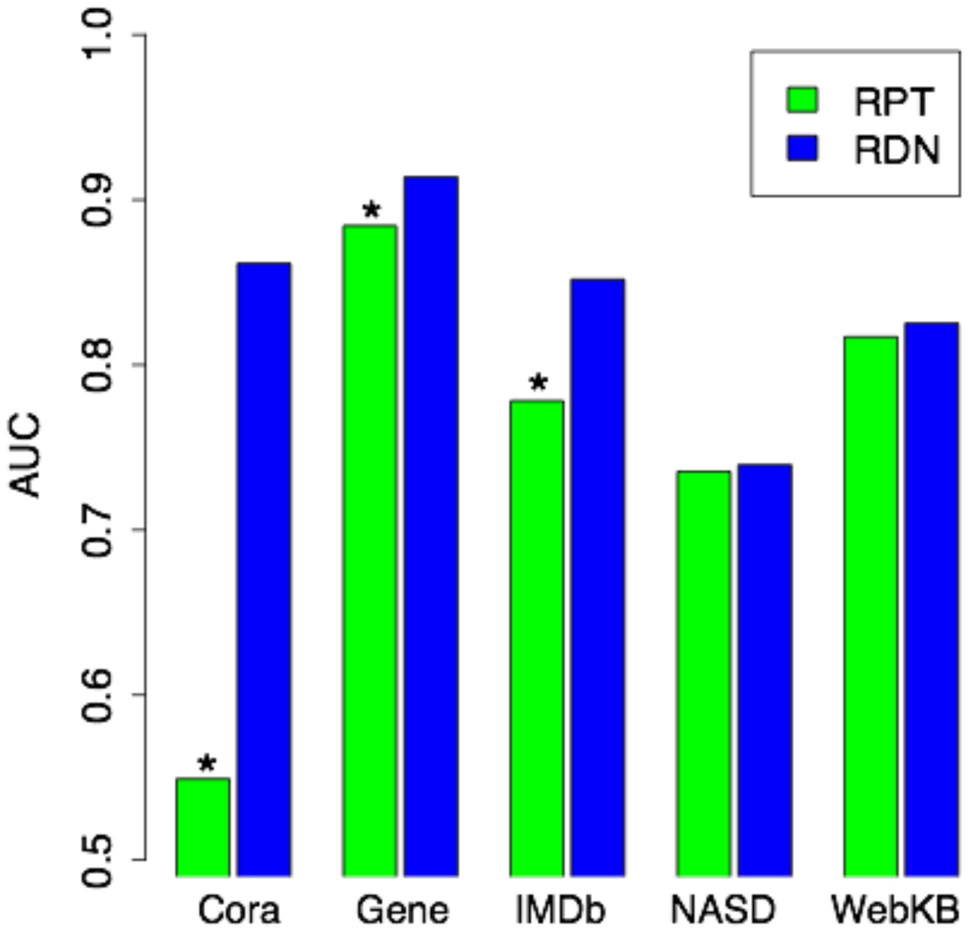
Kernel over data point

Data point


Density of Sepal Width
Density
0.0
0.2
0.4
0.6
0.8
1.0
2.0
2.5
3.0
3.5
4.0
4.5
N = 150   Bandwidth = 0.1233


Density of Sepal Width
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
2.0
2.5
3.0
3.5
4.0
4.5
N = 150   Bandwidth = 0.06164

Bar plots

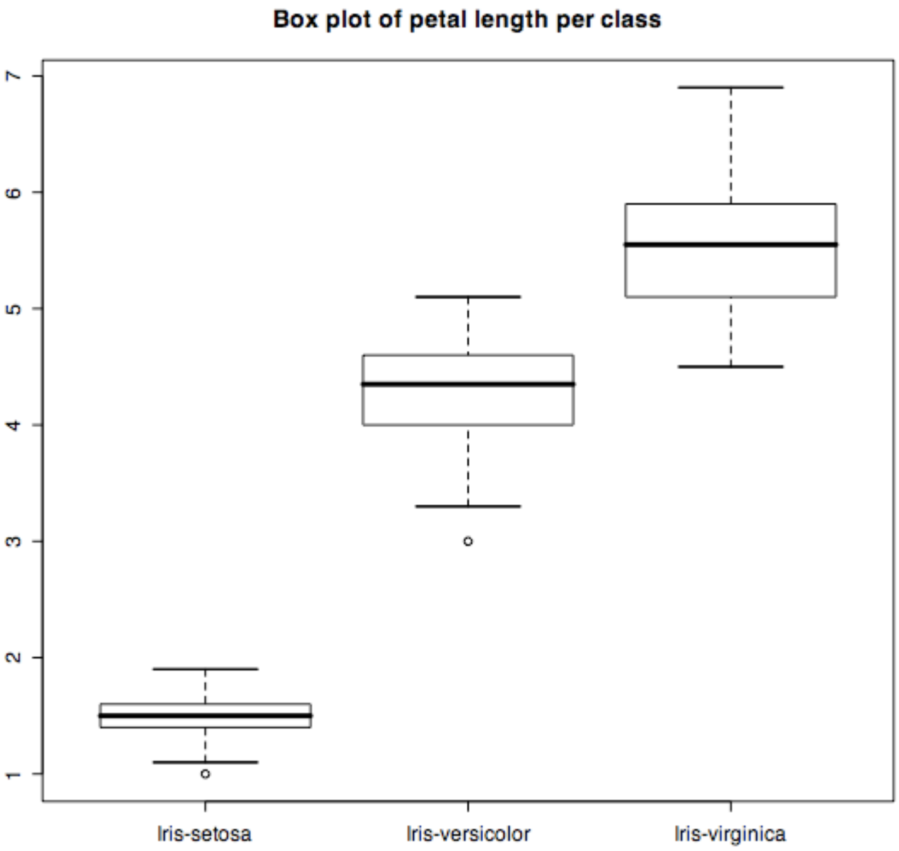


Box plot (2D)

- For each discrete value X, calculate quartiles and range of associated Y values

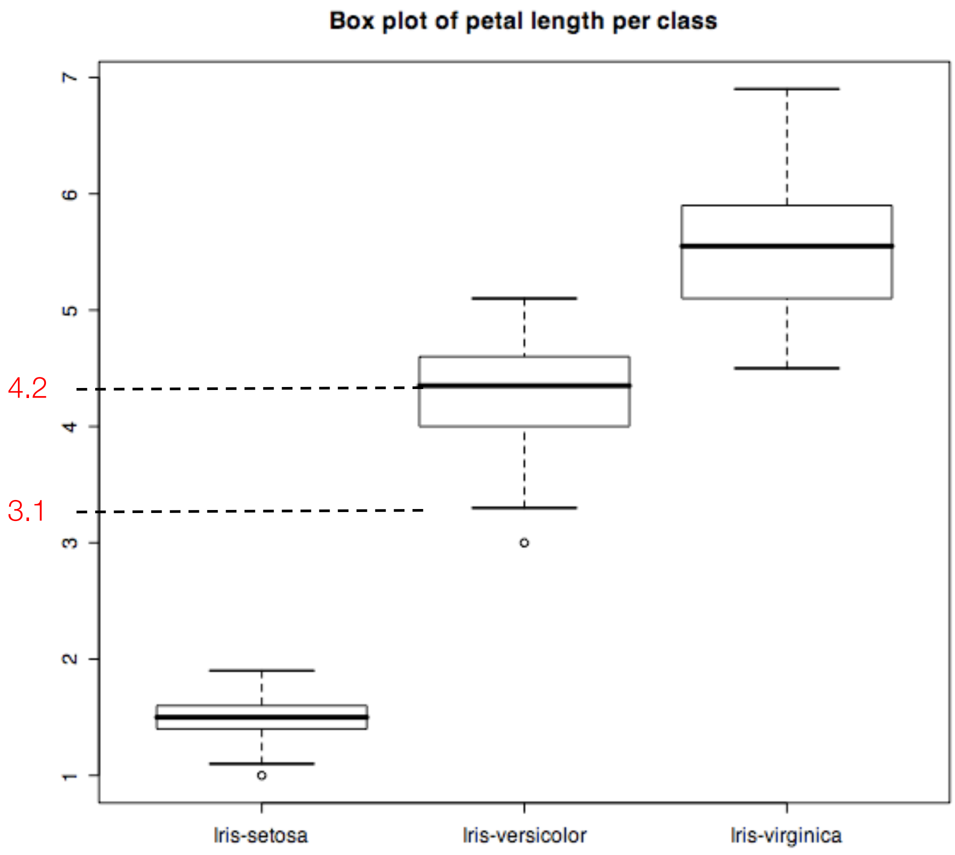
- **Data summary for**: *minimum, first quartile, median, third quartile, and maximum*

- Can also add outliers separately



Interpreting Box Plots

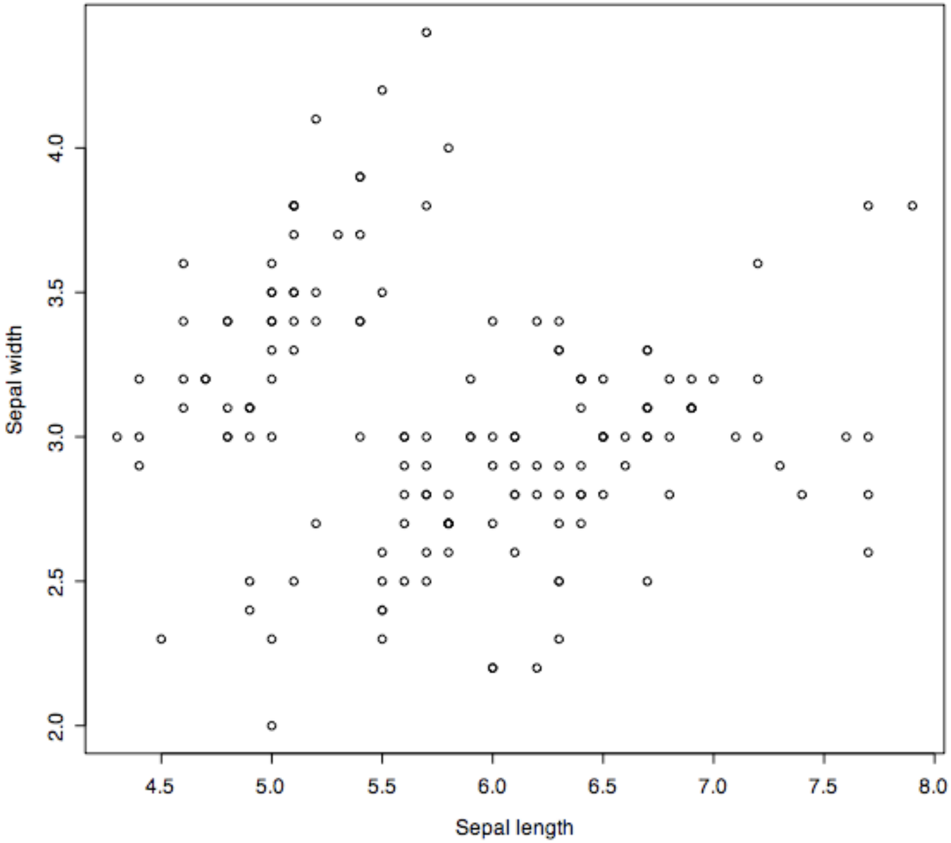
- Petals of Iris-Versicolor are:
  - Always longer than 4
  - At least 50% of the petals are longer than 4
  - There is exactly 1 petal that is 3.1 long, more than 1, at least 1?



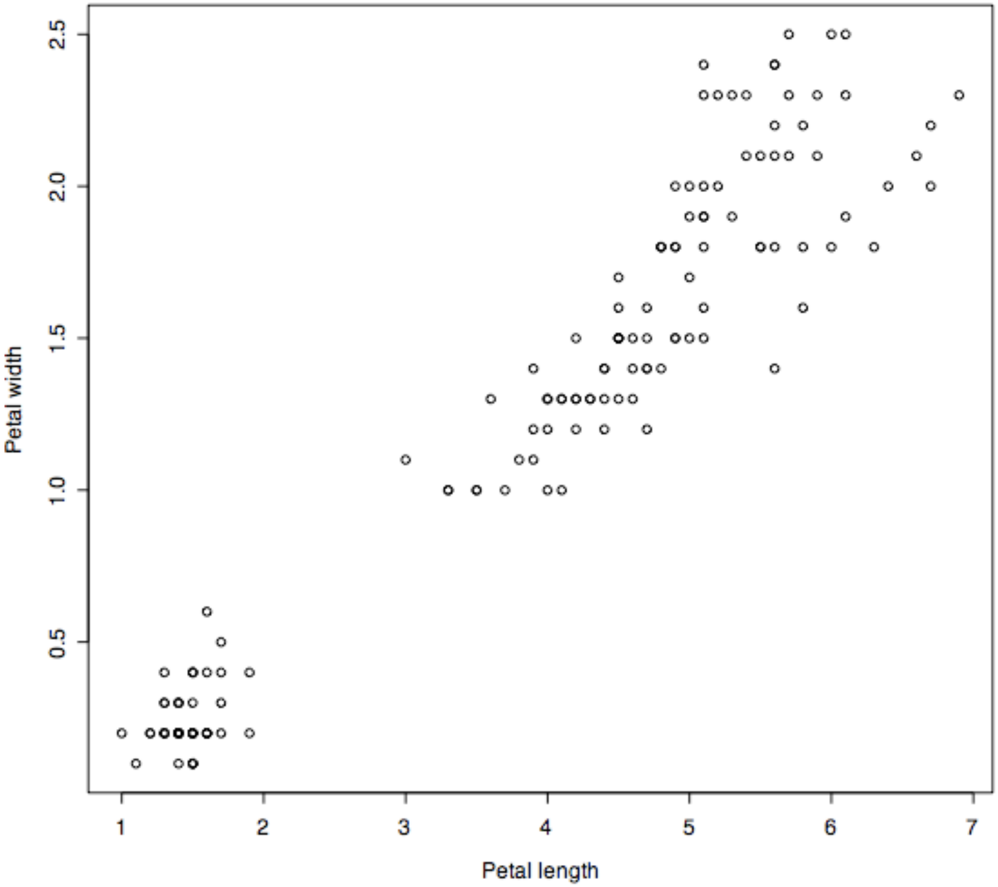
Scatter plot (2D)

- Most common plot for bivariate data
  - Horizontal X axis: the suspected **independent** variable
  - Vertical Y axis: the suspected **dependent** variable
- **Graphically shows:**
  - If X and Y are related
  - Linear or non-linear relationship
  - If the variation in Y depends on X
  - Outliers

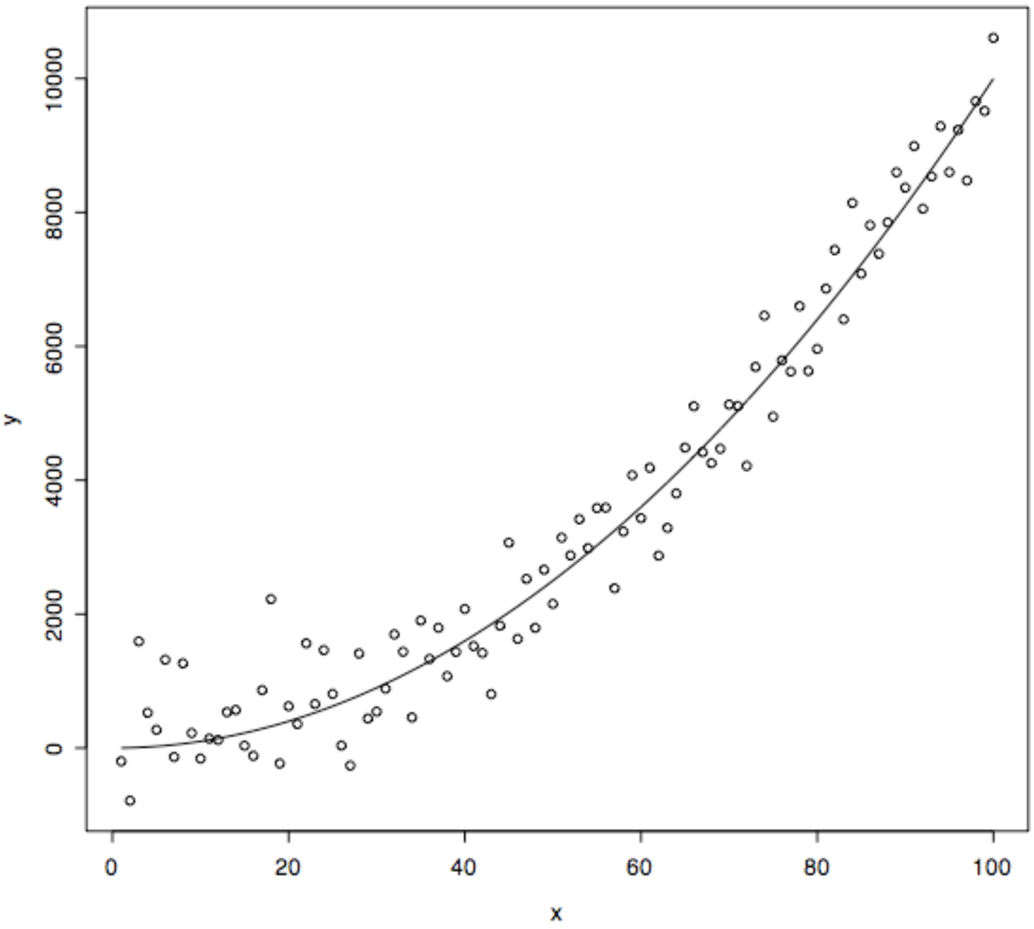
No relationship



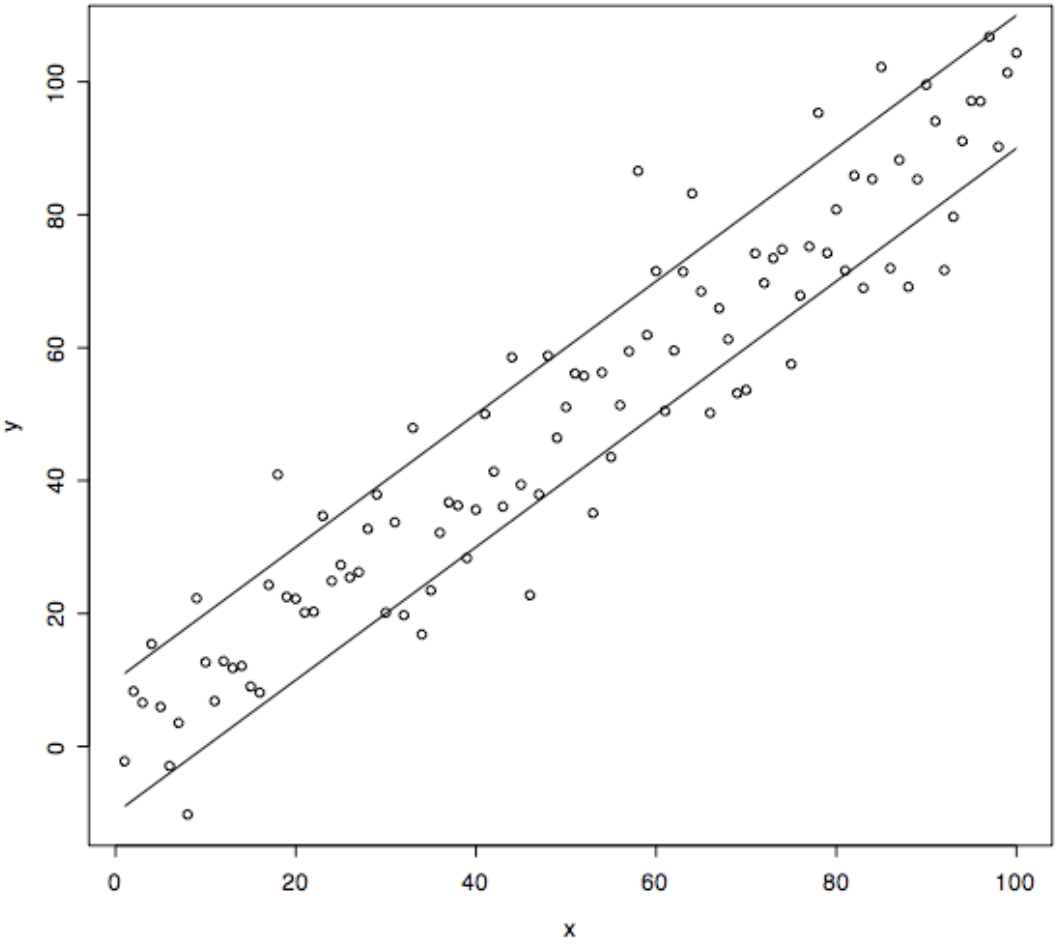
Linear relationship



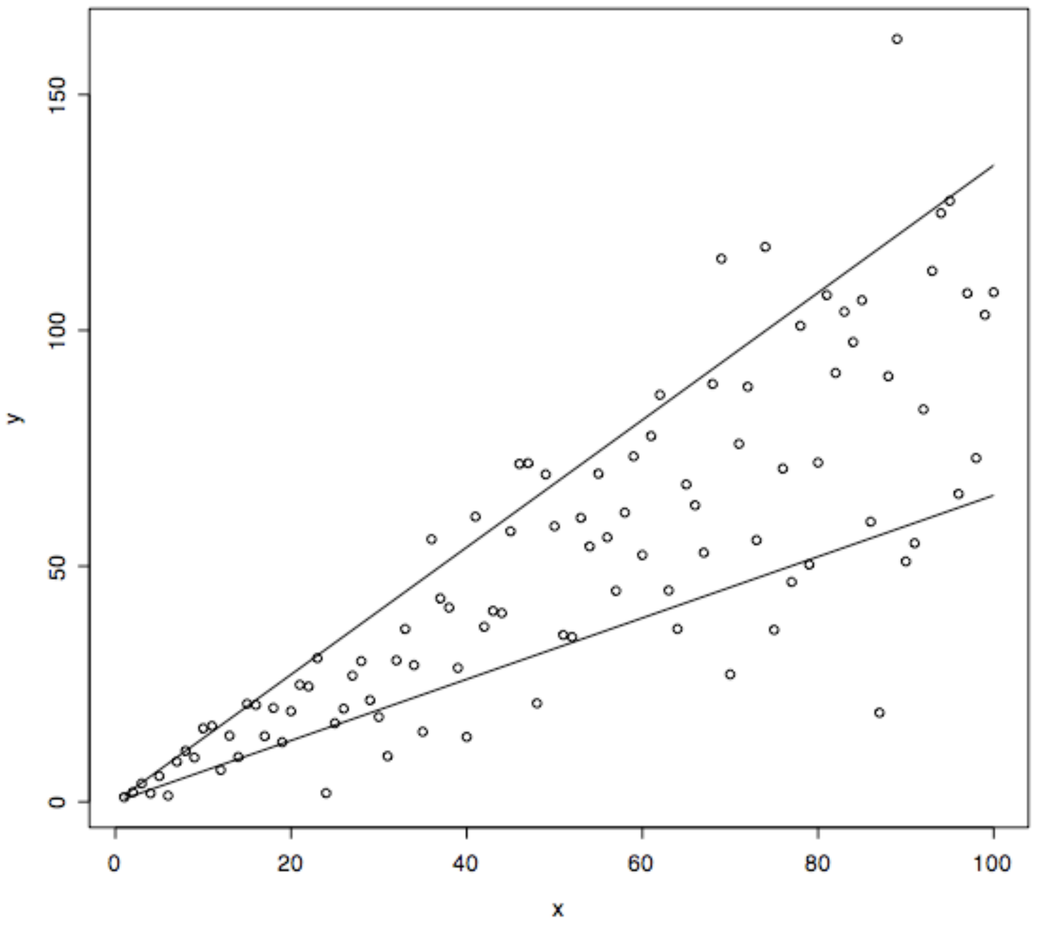
Non-linear relationship



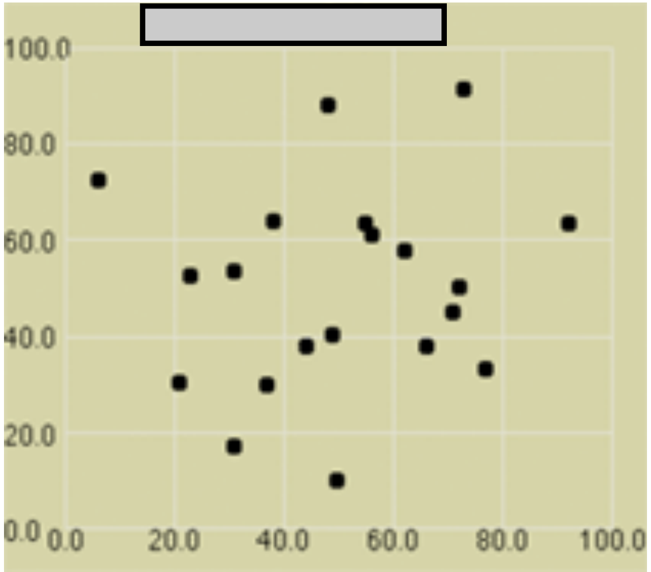
Homoskedastic

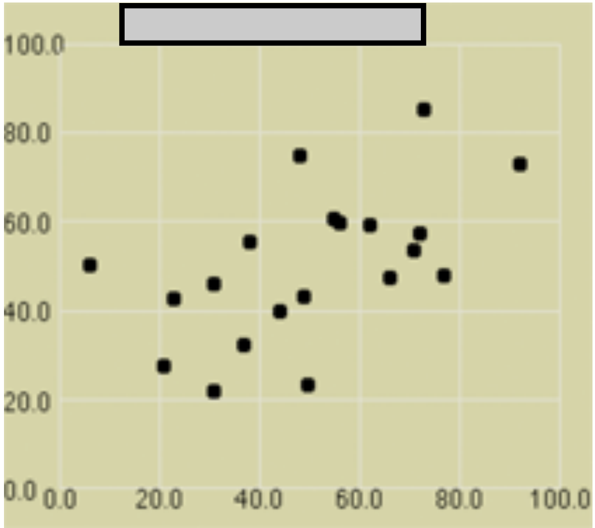


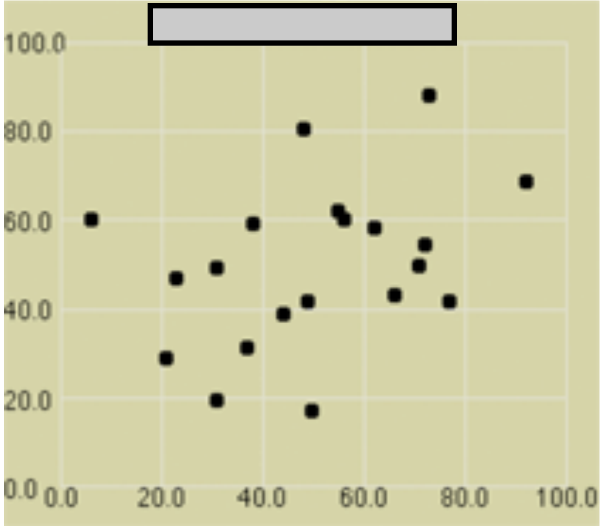
Heteroskedastic



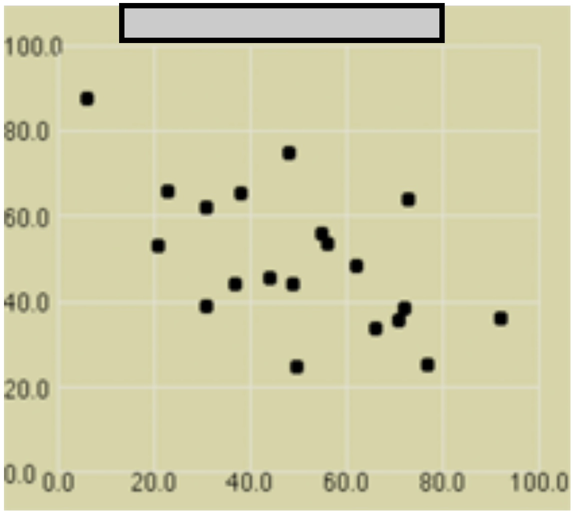
Which one of the plots shows the strongest/ weakest correlation?

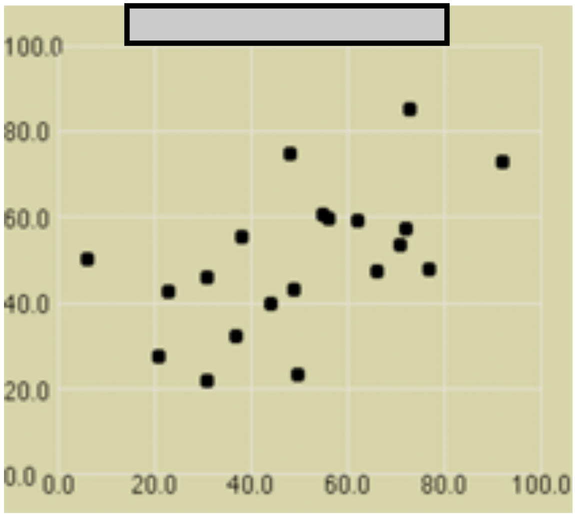


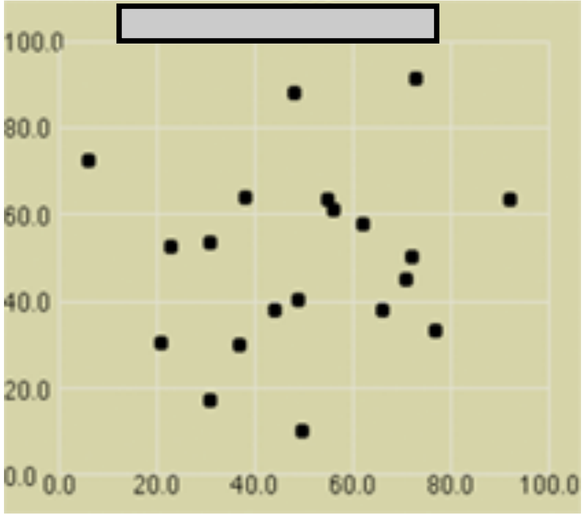




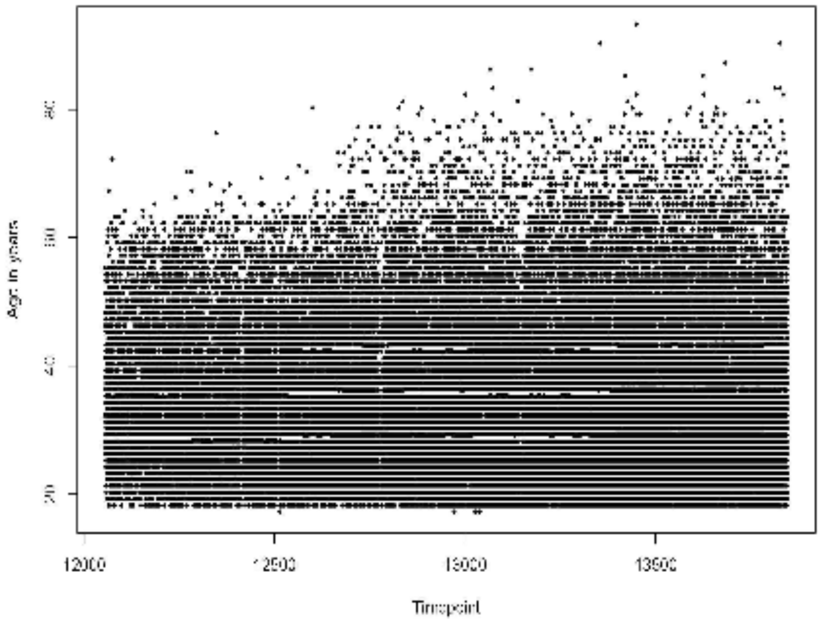
Which one of the plots describes a positive correlation?



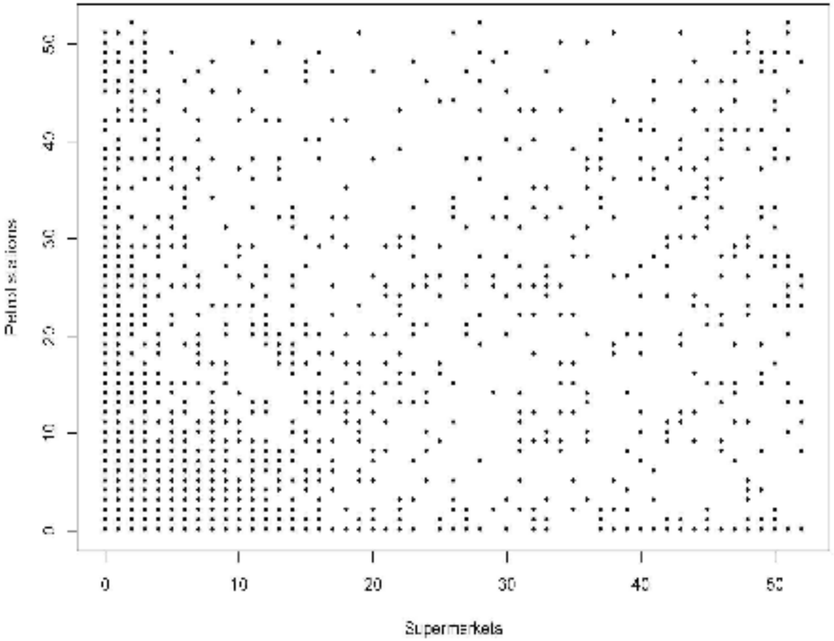




Scatterplot limitations

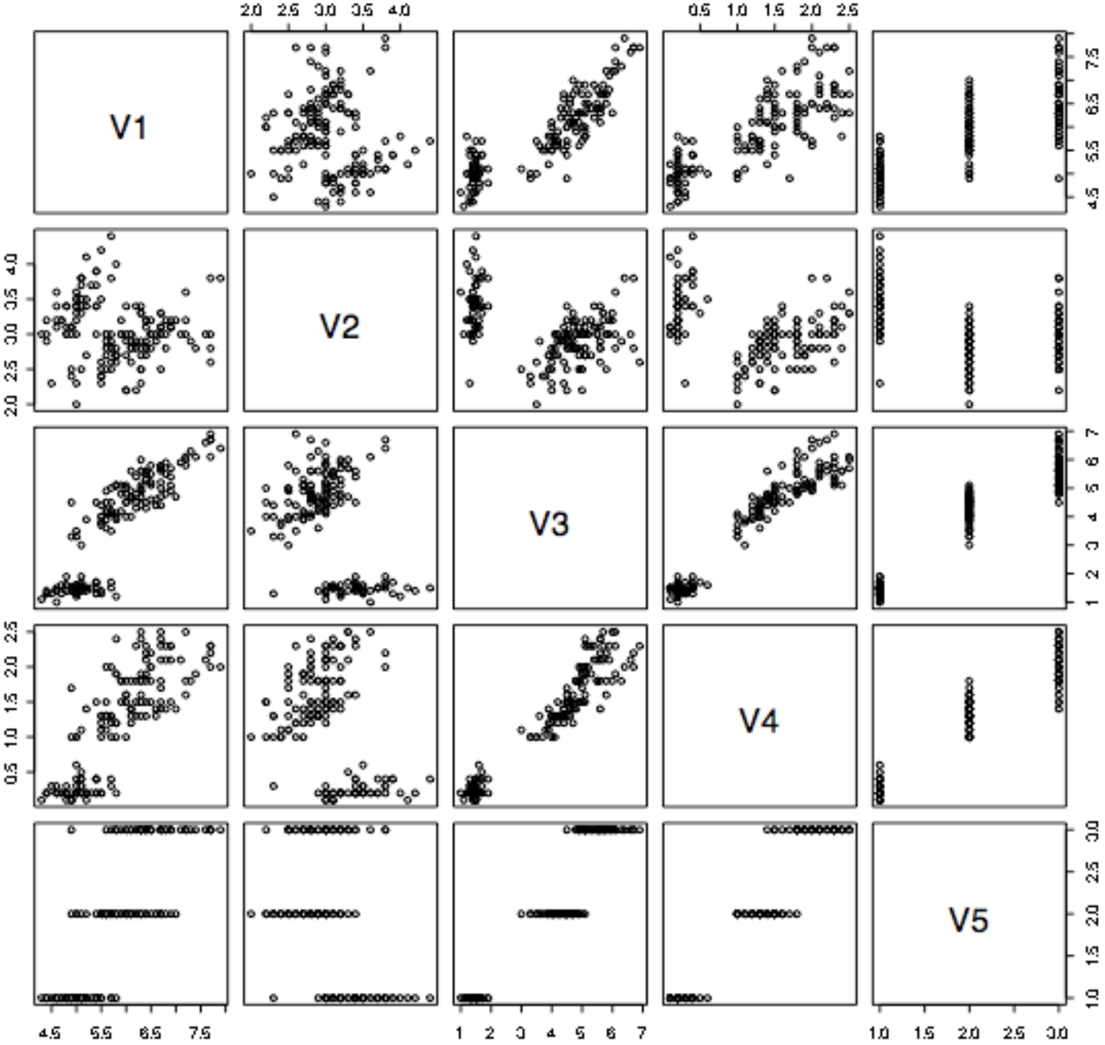


Too much data

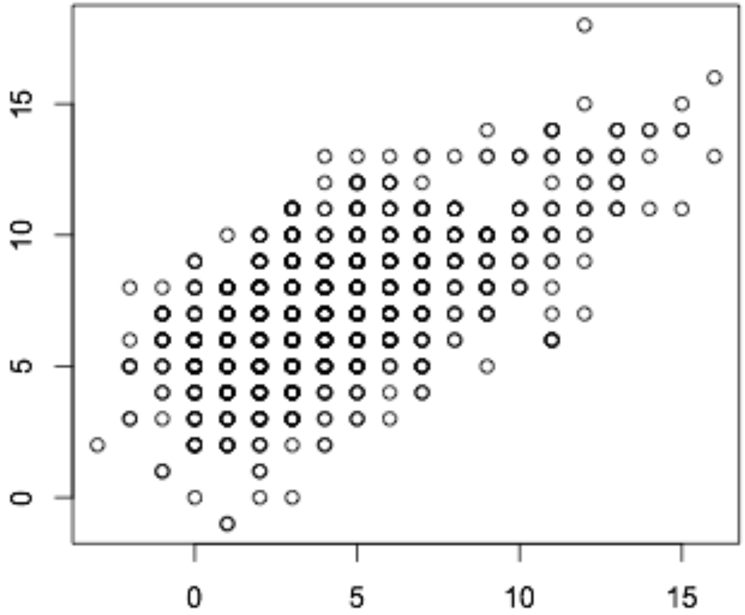


Overprinting

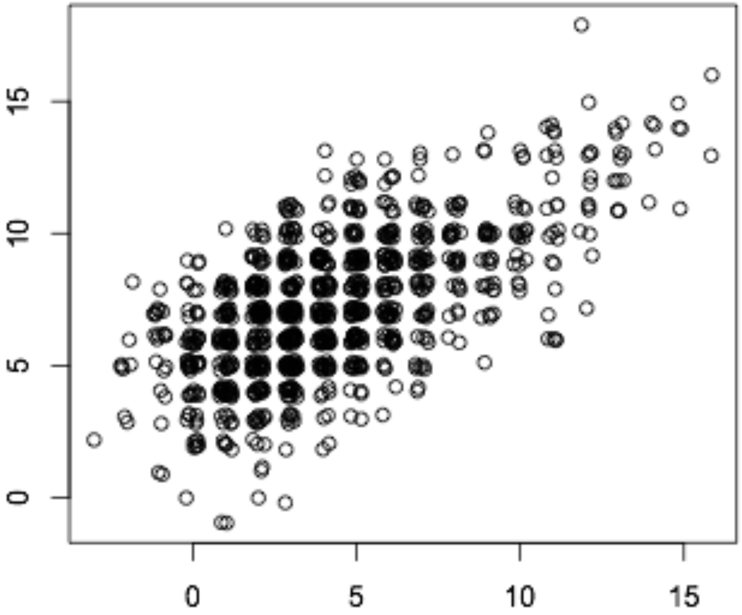
Scatterplot matrix



Scatterplot limitations



Overprinting

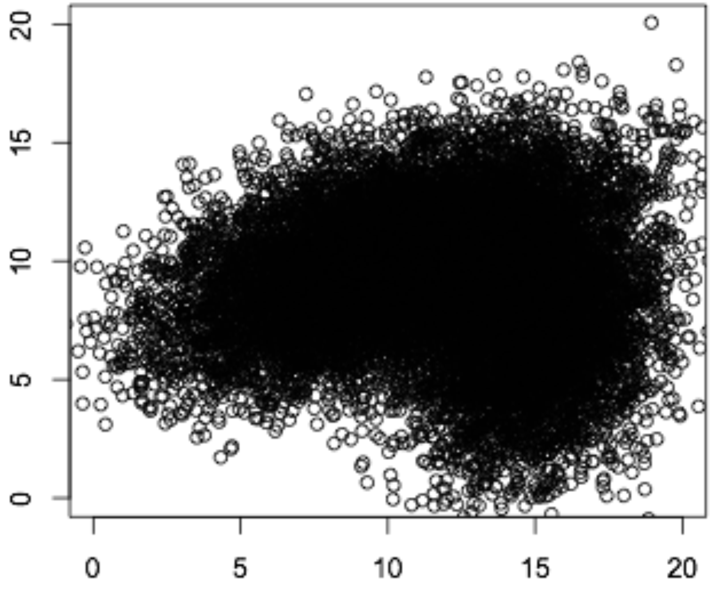


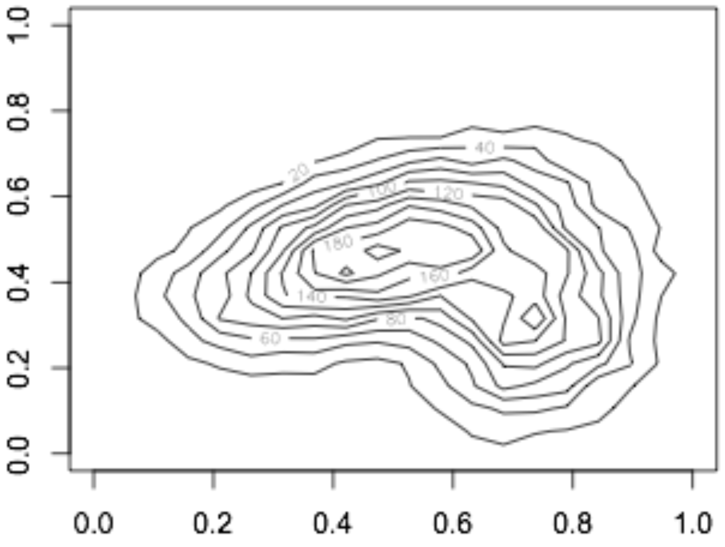
Solution: Jitter points

Contour plot (3D)

- Limitations of 2D scatterplot (e.g., when there is too much data to discern relationship)

- Solution: represent a 3D surface by plotting constant z slices (contours) in a 2D format





Introduction to R

- See: http://cran.r-project.org/