02450: Introduction to Machine Learning and Data Mining

Data Visualization and probability

DTU Compute
Department of Applied Mathematics and Computer Science

Reading Material

Reading material:
C5, C6
Feedback Groups of the day:

- Thomas Ørkild, Frederik Boe Hüttle, Frederik Warburg
- Alastair Ronald Main, Rebekka Vaarum Woldseth
- Andreas Toftegaard, Emilie McFall, Christian Bernitt
- Collin Cunningham, Mike Castro Lundin, Anna Hildigunnur Jonasdottir

Tue Herlau, Mikkel N. Schmidt and Morten Mørup

Introduction to Machine Learning and Data Mining

Course notes fall 2016, version 1

August 29, 2016

Technical University of Denmark

Lecture Schedule

1. Introduction
   30 August: C1

Data: Feature extraction, and visualization

2. Data and feature extraction
   6 September: C2, C3
3. Measures of similarity and summary statistics
   13 September: C4
4. **Data Visualization and probability**
   20 September: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   27 September: C7, C8 **(Project 1 due before 13:00)**
6. Overfitting and performance evaluation
   4 October: C9
7. Nearest Neighbor, Bayes and Naive Bayes
   11 October: TBA
8. Artificial Neural Networks and Bias/Variance
   25 October: TBA
9. AUC, ensemble methods and multi-class classifiers
   1 November: TBA

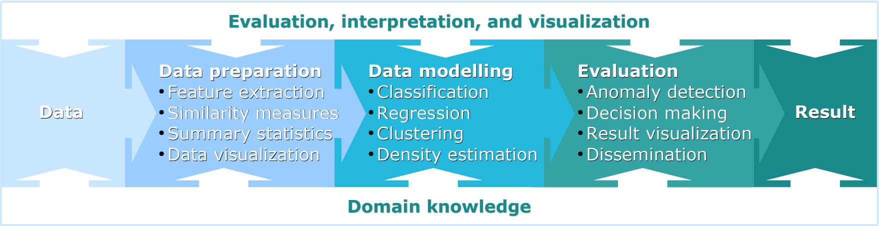
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    8 November: TBA **(Project 2 due before 13:00)**
11. Mixture models and association mining
    15 November: TBA
12. Density estimation and anomaly detection
    22 November: TBA

Recap

13. Recap and discussion of the exam
    29 November: TBA **(Project 3 due before 13:00)**

Data modeling framework



Todays learning objectives:
Explain and apply Bayes theorem
Be able to understand and apply a wide range of data visualization approaches
Understand good practice in plotting including Tufte's guidelines

Imagine you had to build an intelligent robot

- You want the robot to reason accurately
- So what does that mean?

Logical Reasoning

- We consider binary true/false propositions

$$A : My\ cycle\ is\ stolen$$

$$B : My\ cycle\ is\ not\ where\ I\ left\ it$$

$$\overline{A} : The\ negation\ of\ A,\ "Not\ A"$$

$$A \rightarrow B : If\ A\ is\ true\ then\ B\ is\ true$$

Aristotle, inventor of logic, ca. 350BCE

Logical Reasoning

- We consider binary true/false propositions

$$A: \textit{My cycle is stolen}$$

$$B: \textit{My cycle is not where I left it}$$

$$\overline{A}: \textit{The negation of A, "Not A"}$$

$$A \rightarrow B: \textit{If A is true then B is true}$$

- What conclusions can you draw if 'A implies B' and:
  - The bicycle is stolen (**A** is true)
  - The bicycle is not missing (**not B** is true)
  - The bicycle is missing (**B** is true)

Aristotle, inventor of logic, ca. 350BCE

$$\begin{array}{l} A \\ A \rightarrow B \\ \hline \therefore \end{array} \qquad \begin{array}{l} \overline{B} \\ A \rightarrow B \\ \hline \therefore \end{array} \qquad \begin{array}{l} B \\ A \rightarrow B \\ \hline \therefore \end{array}$$

Logical Reasoning

- We consider binary true/false propositions

$$A : My\ cycle\ is\ stolen$$

$$B : My\ cycle\ is\ not\ where\ I\ left\ it$$

$$\overline{A} : The\ negation\ of\ A,\ "Not\ A"$$

$$A \to B : If\ A\ is\ true\ then\ B\ is\ true$$

- What conclusions can you draw if 'A implies B' and:
  - The bicycle is stolen (**A** is true)
  - The bicycle is not missing (**not B** is true)
  - The bicycle is missing (**B** is true)

Aristotle, inventor of logic, ca. 350BCE

$$\begin{array}{l} A \\ A \to B \\ \hline \therefore B \end{array}$$

(If my cycle is stolen it is not where I left it)

$$\begin{array}{l} \overline{B} \\ A \to B \\ \hline \therefore \overline{A} \end{array}$$

(If my cycle is where I left it then it is not stolen)

$$\begin{array}{l} B \\ A \to B \\ \hline \therefore A \end{array}$$

(If my cycle is not where I left it then ...?)

- You can't logically conclude it is stolen, because it might be missing for another reason
- But it is *likely* stolen
- **Logic can't describe common-sense reasoning**

What is probabilities?

$$A: My\ cycle\ is\ stolen$$

$$B: My\ cycle\ is\ not\ where\ I\ left\ it$$

$$\overline{A}: The\ negation\ of\ A,\ "Not\ A"$$

R.T. Cox (1898 - 1991) and T. Bayes (1701-1761)

- The world is an uncertain place
- Suppose P(A|B) is the **degree-of-plausiability** A is true given that B is true

What is probabilities?

$$A : \textit{My cycle is stolen}$$
$$B : \textit{My cycle is not where I left it}$$
$$\overline{A} : \textit{The negation of A, "Not A"}$$

R.T. Cox (1898 - 1991) and T. Bayes (1701-1761)

- The world is an uncertain place
- Suppose P(A|B) is the **degree-of-plausiability** A is true given that B is true
- *"My cycle is more likely stolen if it is missing than if it is not missing"*

$$P(A|B) > P(A|\overline{B})$$

What is probabilities?

A : *My cycle is stolen*

B : *My cycle is not where I left it*

$\overline{A}$: *The negation of A, "Not A"*


R.T. Cox (1898 - 1991) and T. Bayes (1701-1761)

- The world is an uncertain place
- Suppose P(A|B) is the **degree-of-plausiability** A is true given that B is true
- *"My cycle is more likely stolen if it is missing than if it is not missing"*

$$P(A|B) > P(A|\overline{B})$$

- It can be proven that P must behave like a probability assignment (Cox's Theorem)

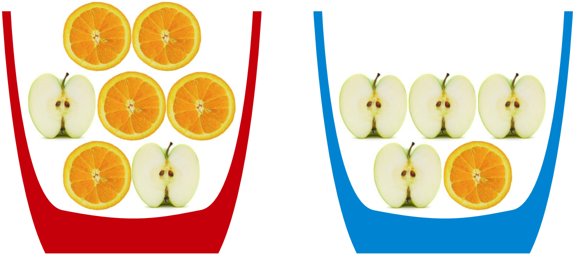
$$\textit{The sum rule:} \quad P(A|C) + P(\overline{A}|C) = 1$$

$$\textit{The product rule:} \quad P(AB|C) = P(B|AC)P(A|C)$$

- To do common-sense reasoning we should use probabilities

Example: Computing with probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?



Probabilities

- In more common notation we have
  - Sum rule

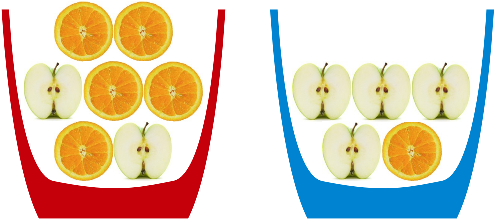
$$p(x) = p(x, y = 0) + p(x, y = 1)$$

  - Product rule

$$p(x, y) = p(x|y)p(y)$$

  - Bayes' rule

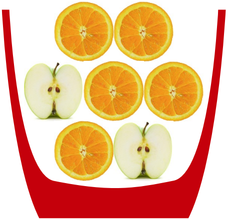
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

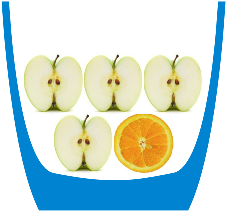


Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

		
	**2**	**5**
	**4**	**1**

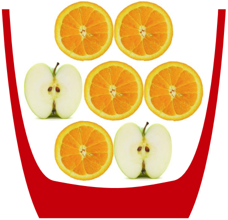


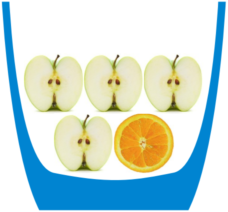


Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12





Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

Probabilities

- What is the probability of an **orange** if the bowl is **red**?
- What is the probability of the **red** bowl if the fruit is **orange**?

			
	2	5	7
	4	1	5
	6	6	12

$$p(o|r) = \frac{p(r,o)}{p(r)} = \frac{5/12}{7/12} = 5/7$$

$$p(r|o) = \frac{p(r,o)}{p(o)} = \frac{5/12}{6/12} = 5/6$$

$$= \frac{p(o|r)p(r)}{p(o)}$$

$$= \frac{5/7 \cdot 7/12}{6/12} = 5/6$$

Common sense

You go to the doctor and gets checked for a disease. The doctor informs you the test correctly gives a positive result 99% of the time. The test is positive.

Should you be worried?

You are doing jury duty and are asked to consider DNA evidence from a crime scene which supposedly match the accused. A world-renowed expert in DNA evidence informs you the test correctly gives a positive match 99% of the time, a claim the defence cannot contest. The test gave a positive match for the accused.

Do you think he is guilty?

Medical test continued

A medical test for a given disease

- Correctly identifies the disease 99% of the time (true positives), and
- Incorrectly turns out positive 2% of the time (false positives).

You know that

- 1% of the population suffers from the disease.

You go to the doctor to get tested, and the test turns out to be positive.

What is the probability you have the disease?

Hints:

- Identify from the text: *(x=Positive, y=0: no disease, y=1: Disease)*

  p(Positive|Disease)

  p(Positive|No Disease)

  p(Disease)

  p(No Disease)

- Use the basic rules of probability given to the right to find:

  p(Disease|Positive)

$$p(y) = \sum_x p(y,x)$$

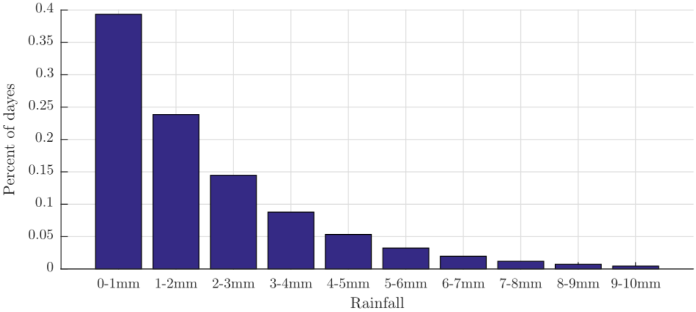
$$= p(y|x)p(x) + p(y|\overline{x})p(\overline{x})$$

$$p(x,y) = p(x|y)p(y)$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

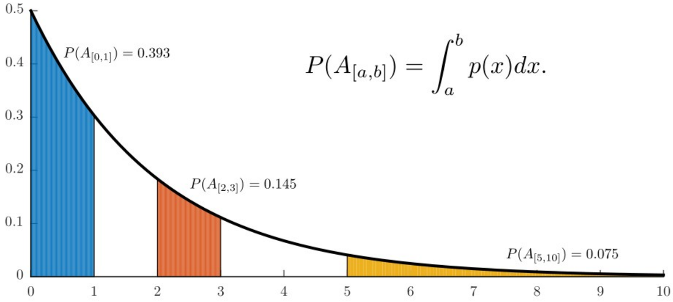
Probability vs. Density

- Suppose we consider the rainfall on an average day r
- **Can't** talk about the probability there will be **exactly** r=2.3 mm of rain, P(r=2.3mm)
- **Can** talk about the probability there will be **between** 1 and 2 mm of rain



Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**

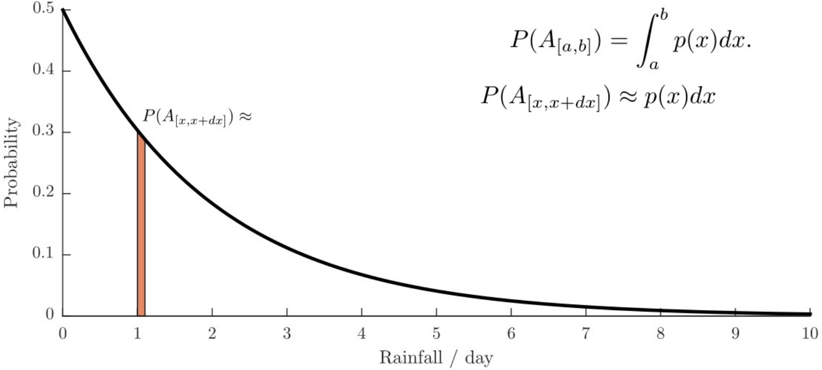


$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

$$A_{[a,b]} : \textit{There will be between a and b mm of rain}$$

Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?



$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

$$P(A_{[x,x+dx]}) \approx p(x)dx$$

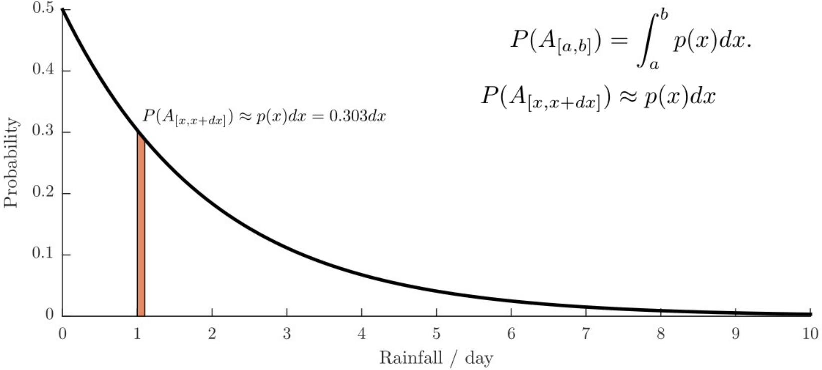
$A_{[a,b]}$: *There will be between a and b mm of rain*

Probability vs. Density

- These probabilities can be **represented** as integrals
- **Events** are **intervals**, the **probability** is the **integral**, the **curve** is the **density**
- What is the probability there will be between 1 and 1.1 mm of rain?

$$P(A_{[a,b]}) = \int_a^b p(x)dx.$$

$$P(A_{[x,x+dx]}) \approx p(x)dx$$



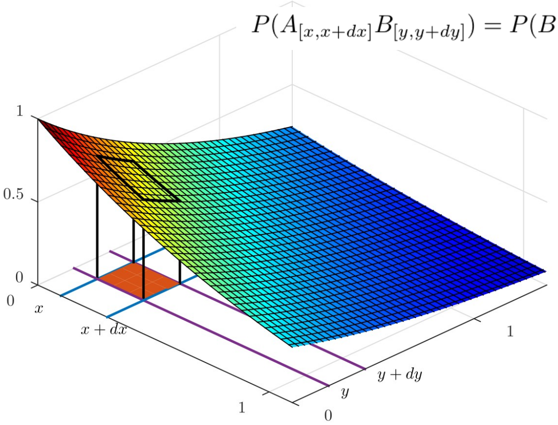
$A_{[a,b]}$: *There will be between a and b mm of rain*

Probability vs. Density

- For two variables x and y, the **probability** is an integral over an **area**

$$P((x,y) \in D) = \int_{(x,y)\in D} p(x,y)dxdy$$

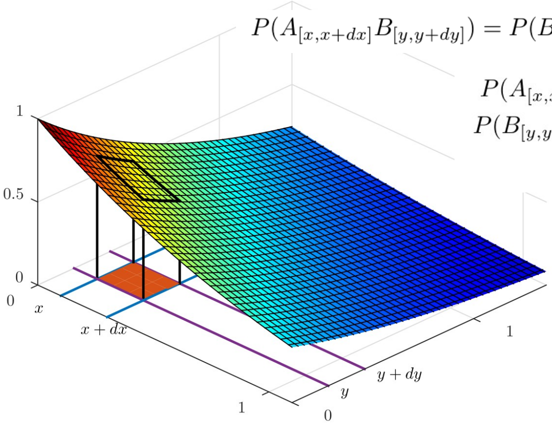
$$P(A_{[x,x+dx]}B_{[y,y+dy]}) = P(B_{[y,y+dy]}|A_{[x,x+dx]})P(A_{[x,x+dx]})$$



Probability vs. Density

- For two variables x and y, the **probability** is an integral over an **area**

$$P(A_{[x,x+dx]}) \approx p(x)dx \qquad\qquad P((x,y) \in D) = \int_{(x,y)\in D} p(x,y)dxdy$$



$$P(A_{[x,x+dx]}B_{[y,y+dy]}) = P(B_{[y,y+dy]}|A_{[x,x+dx]})P(A_{[x,x+dx]})$$

$$P(A_{[x,x+dx]}B_{[y,y+dy]}) \approx dxdyp(x,y)$$

$$P(B_{[y,y+dy]}|A_{[x,x+dx]}) \approx dyp(y|x)$$

$$P(A_{[x,x+dx]}) \approx dxp(x)$$

This implies:

$$p(x,y) = p(y|x)p(x)$$

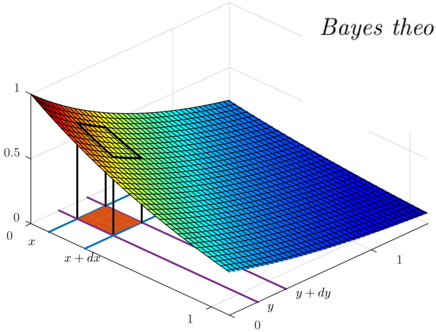
Probability vs. Density

- Thus, we have shown the rules of probability theory also holds for densities

$$\textit{The sum rule:} \qquad \int dx \; p(x|z) = 1$$

$$\textit{The product rule:} \qquad p(x,y|z) = p(y|x,z)p(x|z)$$

$$\textit{Bayes theorem:} \qquad p(x|y,z) = \frac{p(y|x,z)p(x|z)}{p(y|z)} = \frac{p(y|z)p(x|y,z)}{\int p(y|x',z)p(x'|z)dx'}.$$



Collecting all of this we obtain:

- Rules of probability for densities

$$\textit{Marginalization:} \quad \int dx\; p(x,y|z) = p(y|z)$$

$$\textit{The product rule:} \quad p(x,y|z) = p(y|x,z)p(x|z)$$

$$\textit{Bayes theorem:} \quad p(x|y,z) = \frac{p(y|z)p(x|y,z)}{\int p(y|x',z)p(x'|z)dx'}.$$

- Rules of probability for discrete variables

$$\textit{Marginalization:} \quad \sum_c p(x=c,y|z) = p(y|z)$$

$$\textit{The product rule:} \quad p(x,y|z) = p(y|x,z)p(x|z)$$

$$\textit{Bayes theorem:} \quad p(x|y,z) = \frac{p(y|z)p(x|y,z)}{\sum_c p(y|x=c,z)p(x=c|z)}.$$

Statistics and expected values

- **Discrete and continious expectations**

$$\mathbb{E}[g] = \sum_i g(x_i)p(x_i) \qquad \mathbb{E}[g] = \int_{-\infty}^{\infty} g(x)p(x)\mathrm{d}x$$

- **Mean**

$$\bar{x} = \mathbb{E}[x]$$

- **Covariance**

$$\mathrm{cov}(x, y) = \mathbb{E}[(x - \bar{x})(y - \bar{y})]$$

- **Variance**

$$\mathrm{var}(x) = \mathrm{cov}(x, x) = \mathbb{E}\left[(x - \bar{x})^2\right]$$

- **Standard deviation**

$$\mathrm{std}(x) = \sqrt{\mathrm{var}(x)}$$

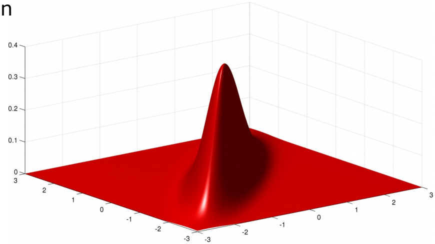
Multivariate Normal distribution

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

- Example: 2-dimensional Normal distribution

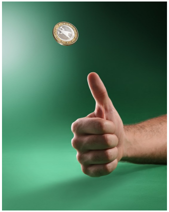
$$\mu = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{bmatrix}$$



Dogs and coins

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

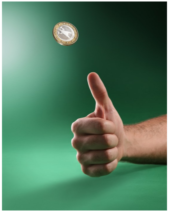


- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

Dogs and coins

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?



Intuition tells us the answers are different, but the situation seems similar…

The Bernouilli distribution

- A magic coin is one that comes up heads with probability θ
- Suppose $b = 1$ is the event the coin landed heads and $b = 0$ the coin landed tails
- The density is:

$$\text{Bernoulli distribution} : p(b|\theta) = \theta^b (1-\theta)^{1-b}$$

The Bernouilli distribution, repeated events

- A magic coin is one that comes up heads with probability θ
- Suppose $b = 1$ is the event the coin landed heads and $b = 0$ the coin landed tails
- The density is:

$$\text{Bernoulli distribution} : p(b|\theta) = \theta^b (1-\theta)^{1-b}$$

- Suppose someone gives you a sequence of N coin flips

$$\begin{aligned} p(b_1, \cdots, b_N|\theta) &= \prod_{i=1}^{N} p(b_i|\theta) \\ &= \theta^{\sum_{i=1}^{N} b_i}(1-\theta)^{N-\sum_{i=1}^{N} b_i} \\ &= \theta^m (1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N \end{aligned}$$

The Bernouilli distribution, learning

- A magic coin is one that comes up heads with probability θ
- Suppose $b = 1$ is the event the coin landed heads and $b = 0$ the coin landed tails
- The density is:

$$\text{Bernoulli distribution} : p(b|\theta) = \theta^b (1-\theta)^{1-b}$$

- Suppose someone gives you a sequence of N coin flips

$$p(b_1, \cdots, b_N|\theta) = \theta^m (1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N$$

- What is θ ?

The Bernouilli distribution, learning

- A magic coin is one that comes up heads with probability θ
- Suppose $b = 1$ is the event the coin landed heads and $b = 0$ the coin landed tails
- The density is:

$$\text{Bernoulli distribution} : p(b|\theta) = \theta^b (1-\theta)^{1-b}$$

- Suppose someone gives you a sequence of N coin flips

$$p(b_1, \cdots, b_N|\theta) = \theta^m (1-\theta)^{N-m}, \quad m = b_1 + b_2 + \cdots + b_N$$

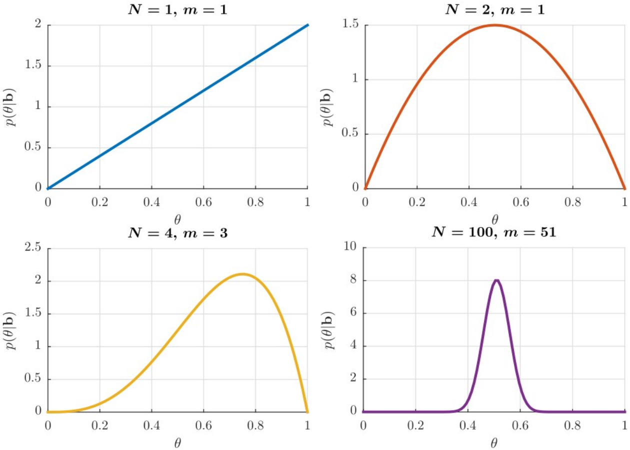
- What is θ ? Easy! Assume $p(\theta) = 1$ then

$$p(\theta|\boldsymbol{b}) = \frac{p(\boldsymbol{b}|\theta)p(\theta)}{\int_0^1 p(\boldsymbol{b}|\theta')p(\theta')d\theta'} = \frac{\theta^m (1-\theta)^{N-m}}{\int_0^1 \theta'^m (1-\theta')^{N-m} d\theta'}$$

$$= \frac{N!}{m!(N-m)!} \theta^m (1-\theta)^{N-m}$$

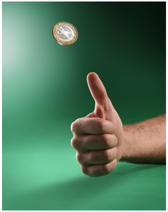
The Bernouilli distribution

$$p(\theta|\boldsymbol{b}) = \frac{N!}{m!(N-m)!}\theta^m(1-\theta)^{N-m}$$



The dog redux

- Your friend just got a dog.
- The dog can either be in the doghouse or outside
- The first four times you come by the dog is in the doghouse
- What is the chance the dog is in the doghouse tomorrow?

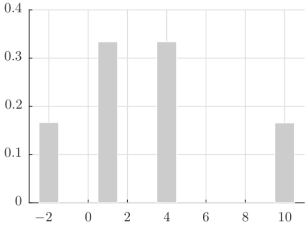


- Your friend buys a coin.
- The coin can either come up heads or tails
- The first four times you flip the coin it comes up tails
- What is the chance the coin comes up tails in the next flip?

$$p(\theta|\boldsymbol{b}) = \frac{p(\boldsymbol{b}|\theta)p(\theta)}{\int_0^1 p(\boldsymbol{b}|\theta')p(\theta')d\theta'} = \frac{\theta^m(1-\theta)^{N-m}p(\theta)}{\int_0^1 \theta'^m(1-\theta')^{N-m}p(\theta')d\theta'}$$

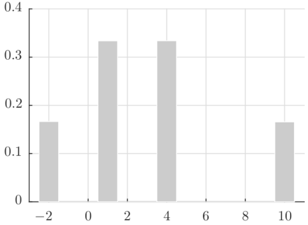
The central limit theorem

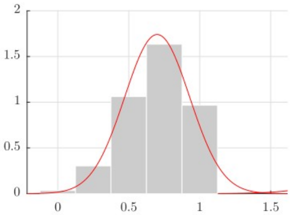
- Consider a “dice” with density:

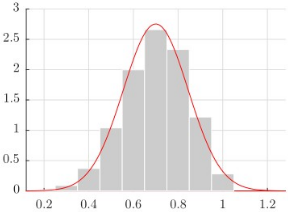


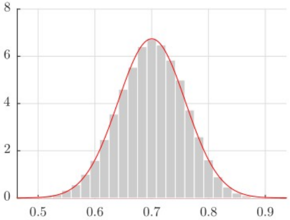
The central limit theorem

- Consider a “dice” with density:
- Suppose we roll the dice N times, and compute the average.
- What is the distribution of the average?



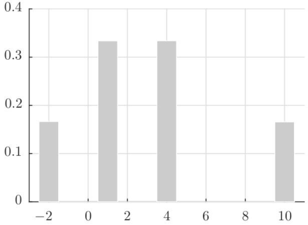


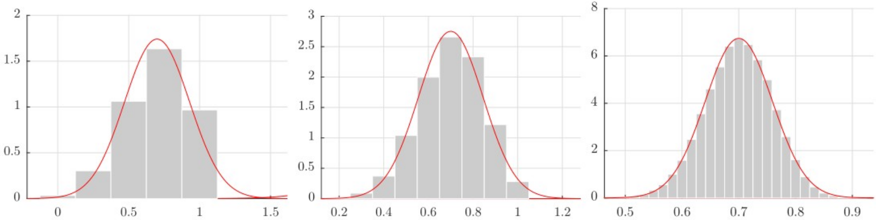




The central limit theorem

- Consider a "dice" with density:
- Suppose we roll the dice N times, and compute the average.
- What is the distribution of the average?





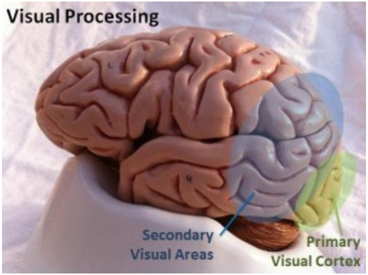
- The distribution becomes a normal distribution $N(x|\mu, \sigma^2)$ with

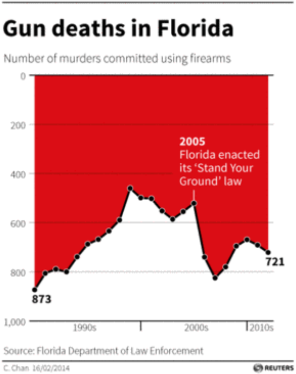
$$\mu = \{\text{Mean of a single dice}\} = 3$$

$$\sigma^2 = \frac{\{\text{Variance of a single dice}\}}{N} = \frac{14}{N}$$

Visualization

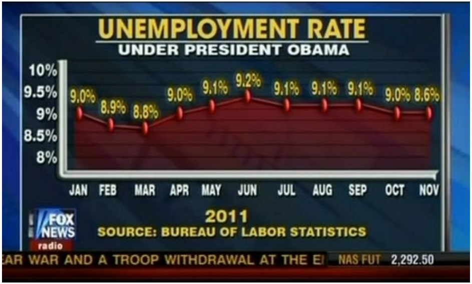
- A main function of the brain is to process visual information
- We can exploit this capacity using visualization of the data in order to:
  - Detect new patterns, i.e. exploratory data analysis)
  - **Dissiminate results, i.e. visualizations/plots in written work (today)**
- We should take into account how the brains visual system works

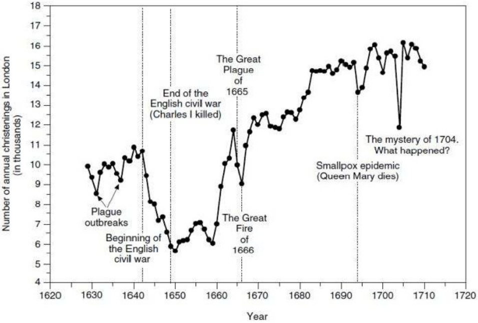




Visualization as technical writing

- Just as there are no recipe for writing, there is no definate recipe for graphing
- However there are guidelines, often more apparent for writing than for illustrations
  - The purpose of the text is to communicate an idea *(vs. plots has a purpose)*
  - Be grammatically correct *(vs. elementary "rules" of good plotting)*
  - Ensure the text is readable *(vs. labels, legends or lines nobody can read)*
  - Avoid long/complicated paragraph *(vs. plots that are overly complicated)*
  - Dont lie or exagerate. *(vs. distort data in a plot)*





Important choices for visualizations

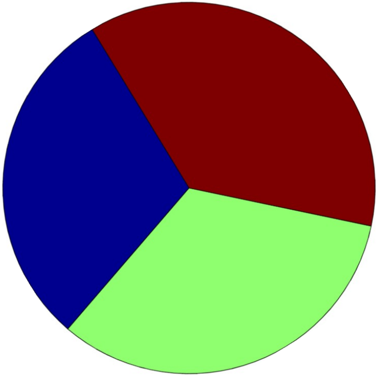
- **Representation:** How will you map objects, attributes, and relations to visual elements?
  - Positions, lengths, colors, areas, orientation

- **Arrangement:** How will you display the visual elements?
  - Viewpoint, transparency, separation, grouping

- **Selection:** How will you handle a large number of attributes and data objects?
  - Display a subset, focus on a region of interest, show summaries

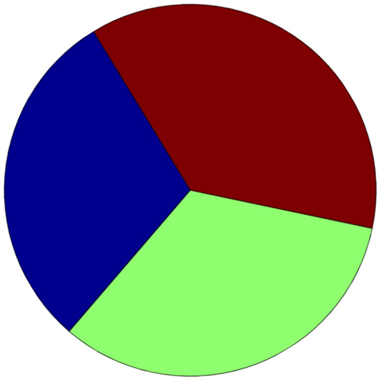
Representation

- **Area represents proportion**
  - Which is smallest, middle, and largest?
  - What are the proportions approximately?

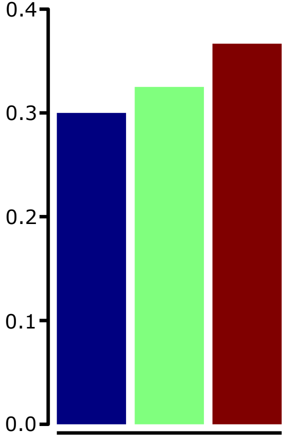


Representation

- **Area represents proportion**
  - Which is smallest, middle, and largest?
  - What are the proportions approximately?



- **Height represents proportion**



Arrangement

- **Placement of visual elements**
  - Can make a great difference in how easy it is to interpret data

- **Example**

	1	2	3	4	5	6
1	0	1	0	1	1	0
2	1	0	1	0	0	1
3	0	1	0	1	1	0
4	1	0	1	0	0	1
5	0	1	0	1	1	0
6	1	0	1	0	0	1
7	0	1	0	1	1	0
8	1	0	1	0	0	1
9	0	1	0	1	1	0

	6	1	3	2	5	4
4	1	1	1	0	0	0
2	1	1	1	0	0	0
6	1	1	1	0	0	0
8	1	1	1	0	0	0
5	0	0	0	1	1	1
3	0	0	0	1	1	1
9	0	0	0	1	1	1
1	0	0	0	1	1	1
7	0	0	0	1	1	1

Selection

- Elimination or de-emphasis of certain objects or attributes

- A subset of **attributes**
  - **Why?** A graph can only show so many attributes – focus on the relevant
  - **How?**
    - Dimensionality reduction
    - Plot pairs of attributes

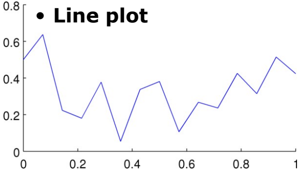
- A subset of **objects**
  - **Why?** A graph can only show so many objects – focus on the relevant
  - **How?**
    - Random sampling
    - Display of region of interest
    - Use density estimation

Types of plots

- **Distribution of a single attribute**
  - Histogram
  - Empirical cumulative distribution
  - Percentile plots
  - Box plot
- **Relation between attributes**
  - 2D histogram
  - Heat maps and contour plots
  - Scatter plots
- **Visualization of high-dimensional objects**
  - Matrix plots
  - Parallel coordinates

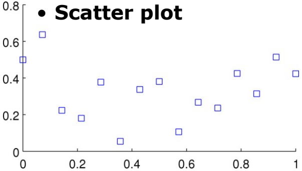
Basic plots

- **Line plot**



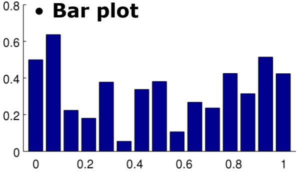

```
plot(x,y);
```

- **Scatter plot**




```
plot(x,y,'s');
scatter(x,y,'s')
```

- **Bar plot**




```
bar(x,y);
```

The iris data set

- **Three flowers**
  - 50 instances of each class, 150 in total

- **Attributes**
  - Sepal (outermost leaves)
    - length in cm
    - width in cm
  - Petal (innermost leaves)
    - length in cm
    - width in cm
  - Class of flower
    - Iris Setosa
    - Iris Versicolour
    - Iris Virginica

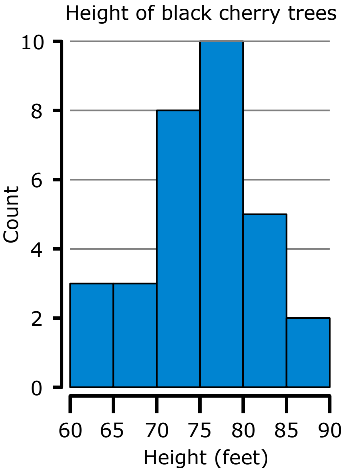
Fower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

X Observation x Attribute

Distribution of a single attribute

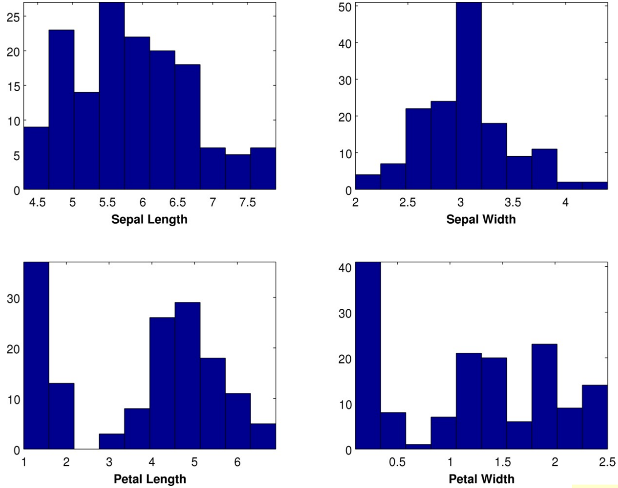
Histograms

- **Shows distribution of a single variable**
  - Divide the values into bins
  - Bar plot of the number of values in bin
  - Height indicates count of values
  - Shape determined by
    - Distribution of data
    - Number of bins / bin width



H = {60, 64, 64, 66, 67, 69, 71, 72, 72, 72, 72, 73, 74, 74, 75, 75, 76, 76, 76, 77, 77, 78, 78, 79, 80, 80, 81, 82, 84, 85, 85, 89}

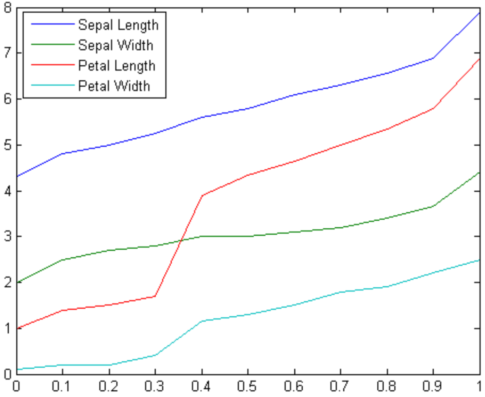
Histograms of the Iris data attributes




```
for m=1:M
    subplot(2,2,m)
    hist(X(:,m),10);
    axis tight;
    xlabel(attributeNames{m})
end
```

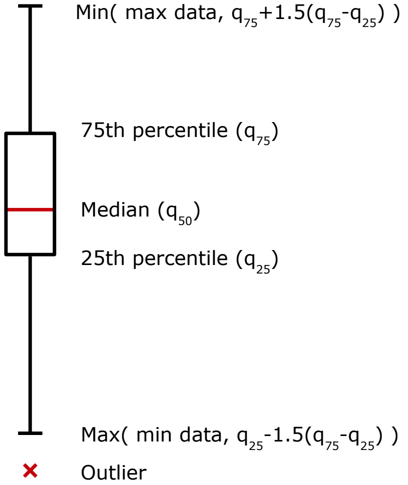
Percentile plots

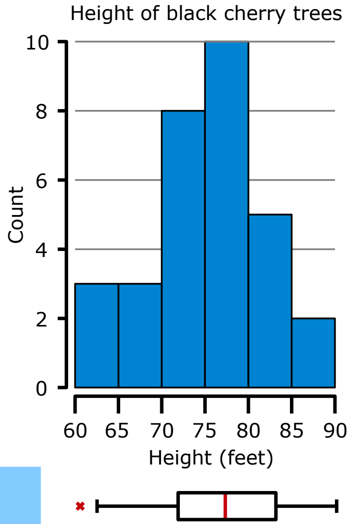
Percentiles: Given an ordinal or continuous attribute **x** and a number **p** between 0 and 100, the **p**th percentile is a value $\mathbf{x_p}$ of **x** such that **p** percent of the observed values of **x** are less than $\mathbf{x_p}$.




```
prctile = 0:0.1:1;
Y = quantile(X,prctile);
plot(prctile,Y);
legend(attributeNames);
```

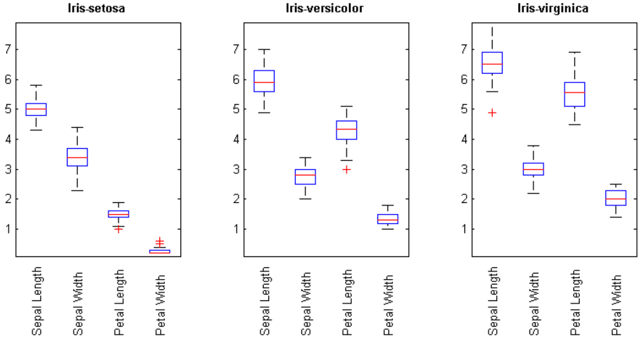
Box plots (revisited from last week)





The plotted whisker extends to the adjacent value, which is the most extreme data value that is not an outlier.

Box plots

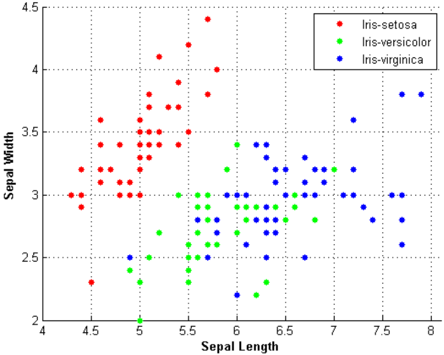



```
y_lim=[min(min(X)) max(max(X))];
for c = 1:C
    subplot(1,C,c)
    boxplot(X(y==c-1,:),'labels',attributeNames,'labelorientation','inline');
    axis([0.5 M+0.5,y_lim])
    title(classNames{c},'FontWeight','bold');
end
```

Relation between attributes

Scatter plots

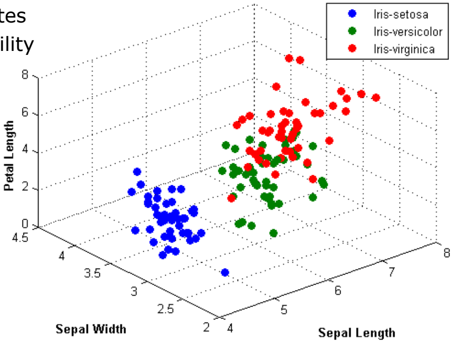
- Shows **relation** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability




```
gscatter(X(:,1),X(:,2),classNames(y+1));
grid on;
xlabel(measurementType(1),'fontweight','bold');
ylabel(measurementType(2),'fontweight','bold');
```

Scatter plots

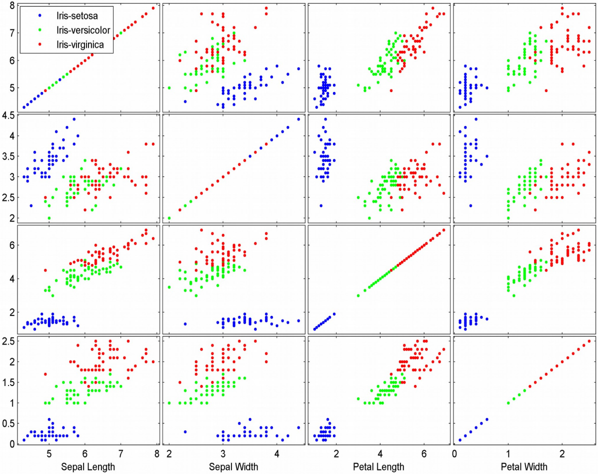
- Shows **relation** between attributes
  - Assess dependence between attributes
  - Used with classes to assess separability
  - 3D plots are often confusing; avoid if possible




```
hold all;
for c = 1:C
    scatter3(X(y==c-1,1),X(y==c-1,2),X(y==c-1,3),'filled');
end
grid on;
xlabel(attributeNames{1},'FontWeight','bold');
ylabel(attributeNames{2},'FontWeight','bold');
zlabel(attributeNames{3},'FontWeight','bold');
legend(classNames)
```

Scatter plots

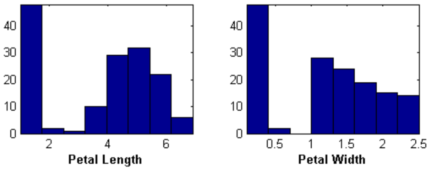
- Scatter plot matrix
  - All pairs of attributes



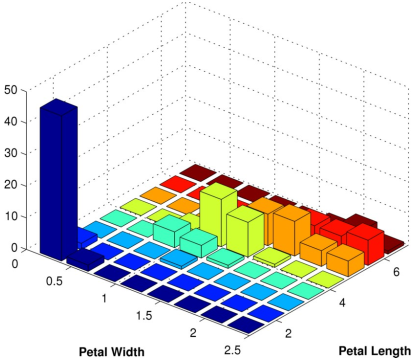

```
gplotmatrix(X,X,classNames(y+1),[],[],[],'on','',attributeNames);
```

Two-dimensional histograms

- Shows joint distribution of two variables



1D histogram of each attribute

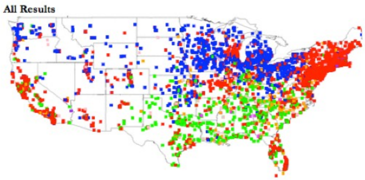


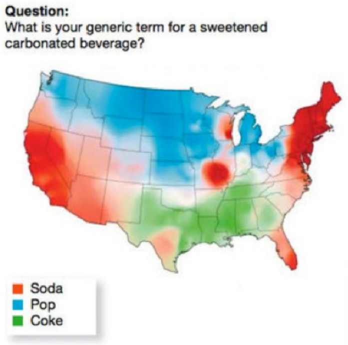
2D joint histogram

```
nBins = 8;
ind = [3 4];  % Indices of the two variables to create the 2D histogram from
[n,x,data] = hist2d(X(:,ind)',nBins);
bar3xy(x(1,:),x(2,:),n);
axis tight;
xlabel(attributeNames{ind(1)},'FontWeight','bold')
ylabel(attributeNames{ind(2)},'FontWeight','bold')
```

Other approaches to 2D density visualization

- 3-D plots should be avoided when possible. Consider instead **heat-maps**, **contour** plots or **images** instead of 2D histograms





- When using colors, choose a color scheme which use one color per class and let the intensity of the color indicate strength.

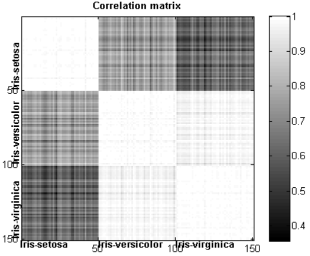
Visualization of high-dimensional objects

Matrix plots

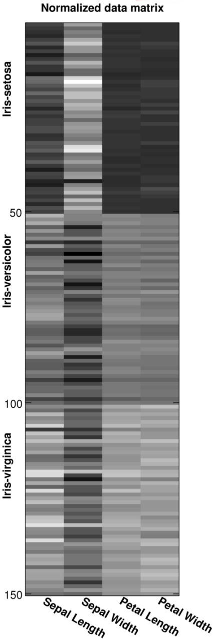
- **Plot of raw data matrix**
  - Useful when objects are sorted according to class
  - Typically, attributes are normalized

```
Z = zscore(X); % Standardize data
imagesc(Z);
colormap(gray)
title('Normalized data matrix');
```

- **Plots of similarity matrices**
  - Useful for visualizing the relation between objects

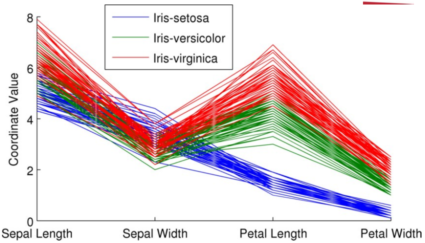



```
C = corrcoef(X');
h2 = imagesc(C);
colormap(gray);
axis equal;
axis tight;
colorbar;
title('Correlation matrix');
```



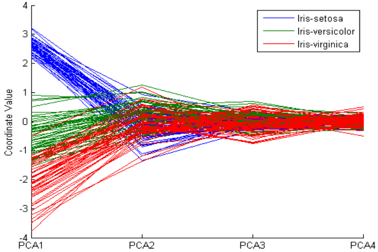
Parallel coordinates

- **Plot high-dimensional data**
- Instead of perpendicular axes
  - Use parallel axes
- Attribute values are plotted as a point
  - and the points are connected by a line
- Each object is represented as a line
- Lines representing a group of objects
  - Are similar in some sense
  - Ordering of attributes is important in seeing such groupings




```
parallelcoords(X,'group',classNames(y+1),...
'labels',attributeNames);
```

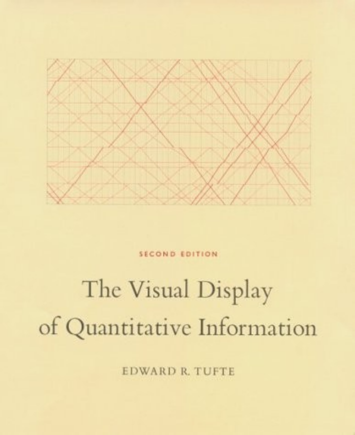
```
% Project data by PCA
Y = bsxfun(@minus, X, mean(X));
[U, S, V] = svd(Y);
Z = U*S;
% Plot data in space spanned by PCA
parallelcoords(Z,'group',classNames(y+1),...
'labels',{'PCA1','PCA2','PCA3','PCA4'});
```



Tufte's guidelines

- **Graphical excellence**
  - Well-designed presentation of interesting data – a matter of
    - substance, statistics, and design
  - Complex ideas communicated with
    - clarity, precision, and efficiency
  - Gives the viewer
    - the greatest number of ideas
    - in the shortest time
    - with the least ink
    - in the smallest place.
  - Nearly always multivariate
  - Requires telling the truth about the data
  - Maximise Data-ink ratio:

$$\text{Data-ink ratio} = \frac{\text{Data-ink}}{\text{Total ink used}}$$



Some guidelines I follow

- Everything should be vector graphics at all times (PDF or EPS; the zoom-in test)
- Print out the graphics and ensure it is readable
- Never use a piechart, 3D plots (heat-maps or contour plots are preferable)
- It payes to tweak colors, line widths, marker sizes, marker colors, etc
- Under no circumstances must you add a 3D effect to a plot
- Avoid excel for anything serious
- Always use a white background. Don't add colors unless it improves the plot
- The best plots are often **"plot(x,y,'.-');"**. Can the point be made like this?
- Turn of the box around the plot ("**box off;**" in matlab)
- Consider the scale/location of axis: "**axis equal;**" if axis are of the same type
- Automate as much as possible (labels, legends, axis). Avoid the Matlab plot editor
- For dataplots I use **Matlab2014a+matlabfrag+Latex+Tikz,** for other figures I usually use Incscape
- Use image captions. Add a *"mini-conclusion"* to the caption to tell a reader what he should/can take away from a plot
- Ask others about your plot without explaining them at first
- Be aware a typical reader won't read your main text
- **Use common sense**

Making good data visualizations is an art

For some interesting data visualizations see also

http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

http://www.informationisbeautiful.net/

http://www.junkcharts.typepad.com/