Computational Genomics

http://www.cs.cmu.edu/~02710

Introduction to probability, statistics and algorithms

(brief) intro to probability

Basic notations

- Random variable

  - referring to an element / event whose status is unknown:

  A = "gene g is increased 2 folds"

- Domain (usually denoted by Ω)

  - The set of values a random variable can take:

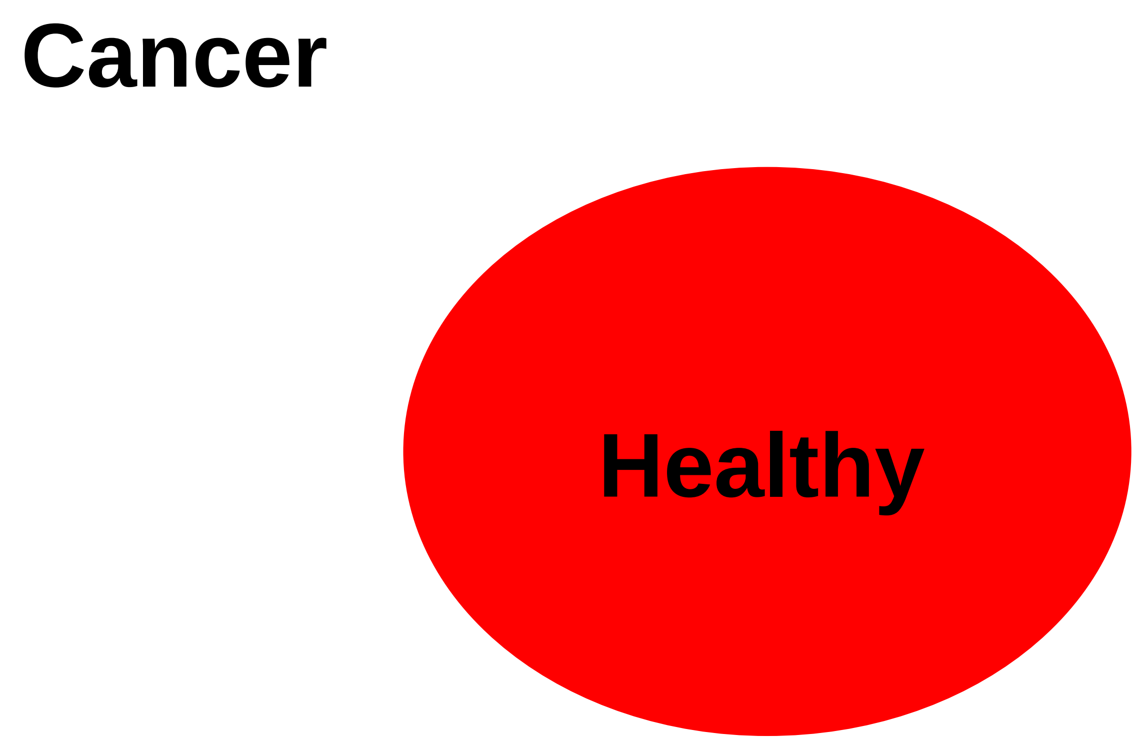
  - "A = Cancer?": Binary

  - "A = Protein family": Discrete

  - "A = Log ratio change in expression": Continuous

Priors

Degree of belief in an event in the absence of any other information



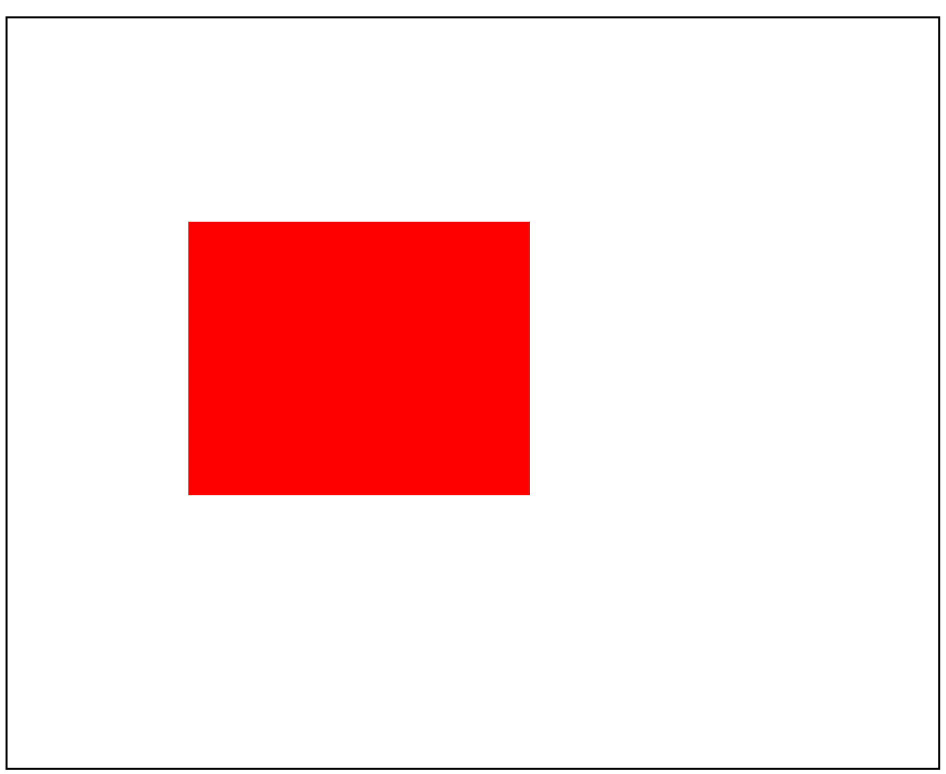
P(cancer) = 0.2

P(no cancer) = 0.8

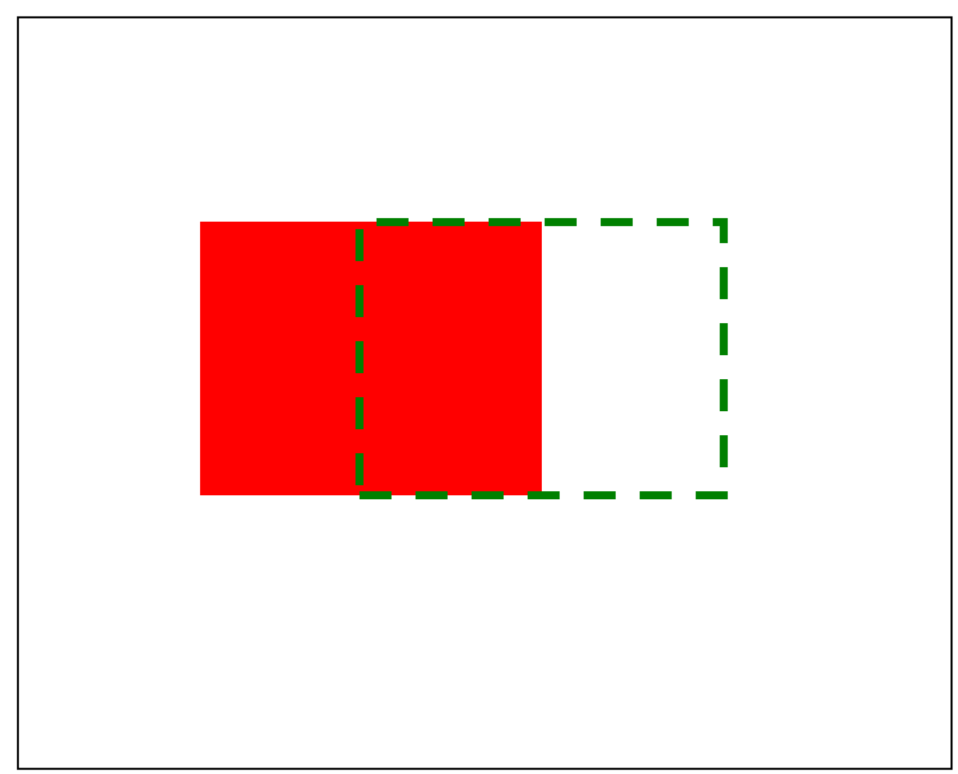
Conditional probability

- P(A = 1 | B = 1): The fraction of cases where A is true if B is true

P(A = 0.2)



P(A|B = 0.5)



Conditional probability

- In some cases, given knowledge of one or more random variables we can improve upon our prior belief of another random variable
- For example:

p(cancer) = 0.5

p(cancer | non smoker) = 1/4

p(cancer | smoker) = 3/4

Cancer	**Smoker**
1	1
0	0
1	0
1	1
0	1
1	1
0	0
0	0

Joint distributions

- The probability that a *set* of random variables will take a specific value is their joint distribution.
- Notation: $P(A \wedge B)$ or $P(A,B)$
- Example: P(cancer, smoking)

If we assume independence then

$$P(A,B)=P(A)P(B)$$

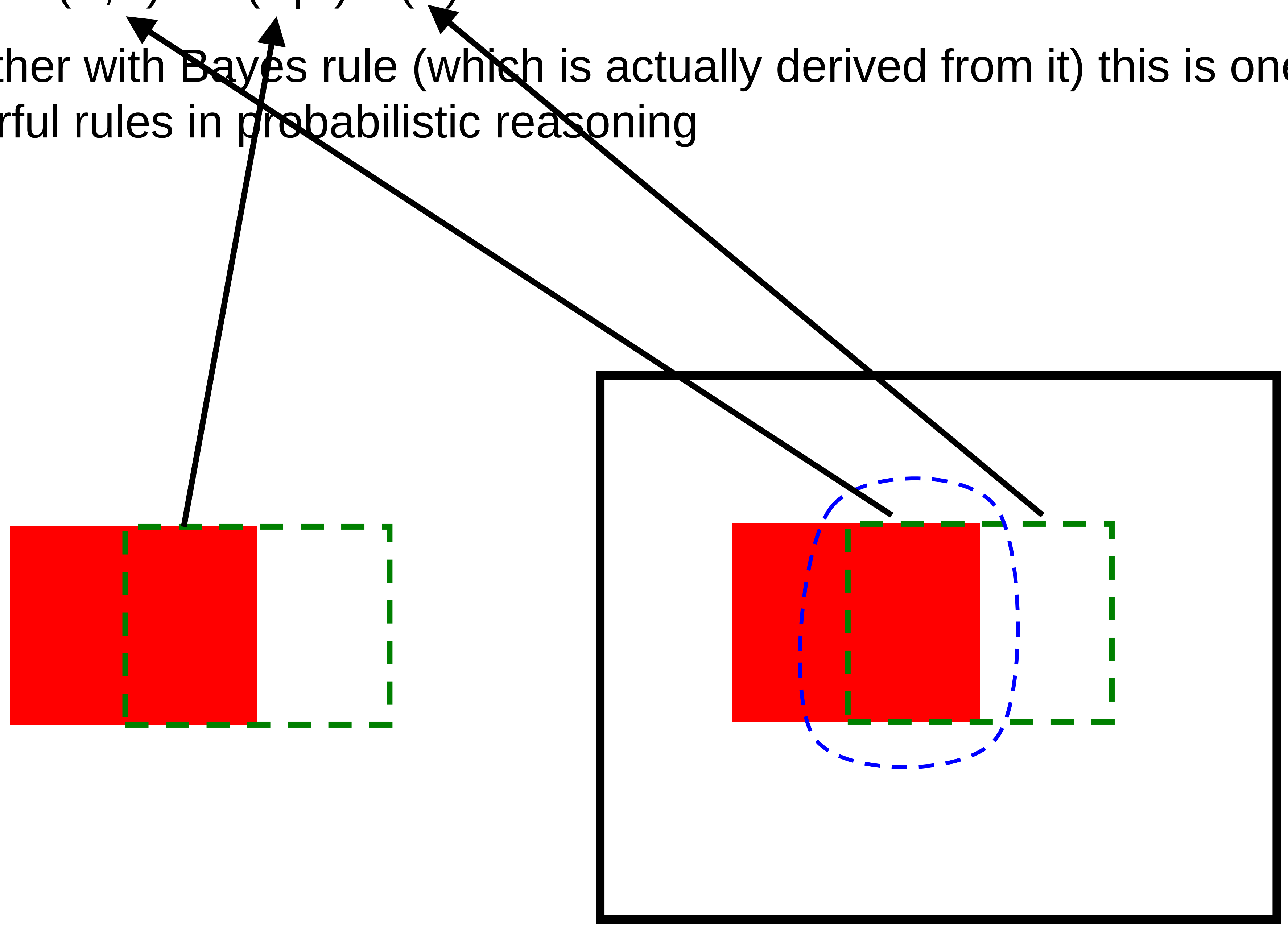
However, in many cases such an assumption maybe too strong (more later in the class)

Chain rule

- The joint distribution can be specified in terms of conditional probability:

  $P(A,B) = P(A|B)*P(B)$

- Together with Bayes rule (which is actually derived from it) this is one of the most powerful rules in probabilistic reasoning



Bayes rule

- One of the most important rules for this class.
- Derived from the chain rule:

  P(A,B) = P(A | B)P(B) = P(B | A)P(A)
- Thus,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

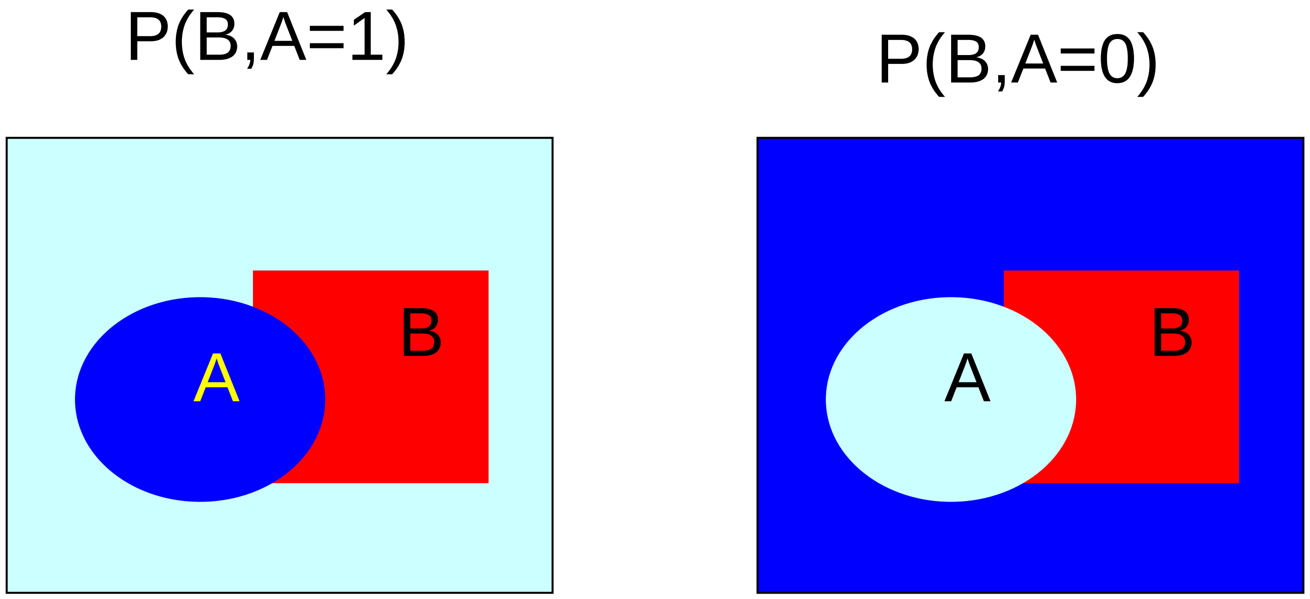
Thomas Bayes was an English clergyman who set out his theory of probability in 1764.

Bayes rule (cont)

Often it would be useful to derive the rule a bit further:

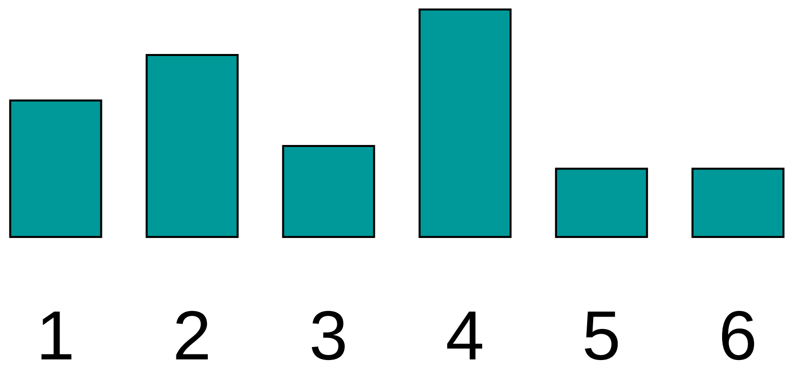
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_A P(B|A)P(A)}$$

This results from:
$P(B) = \sum_A P(B,A)$



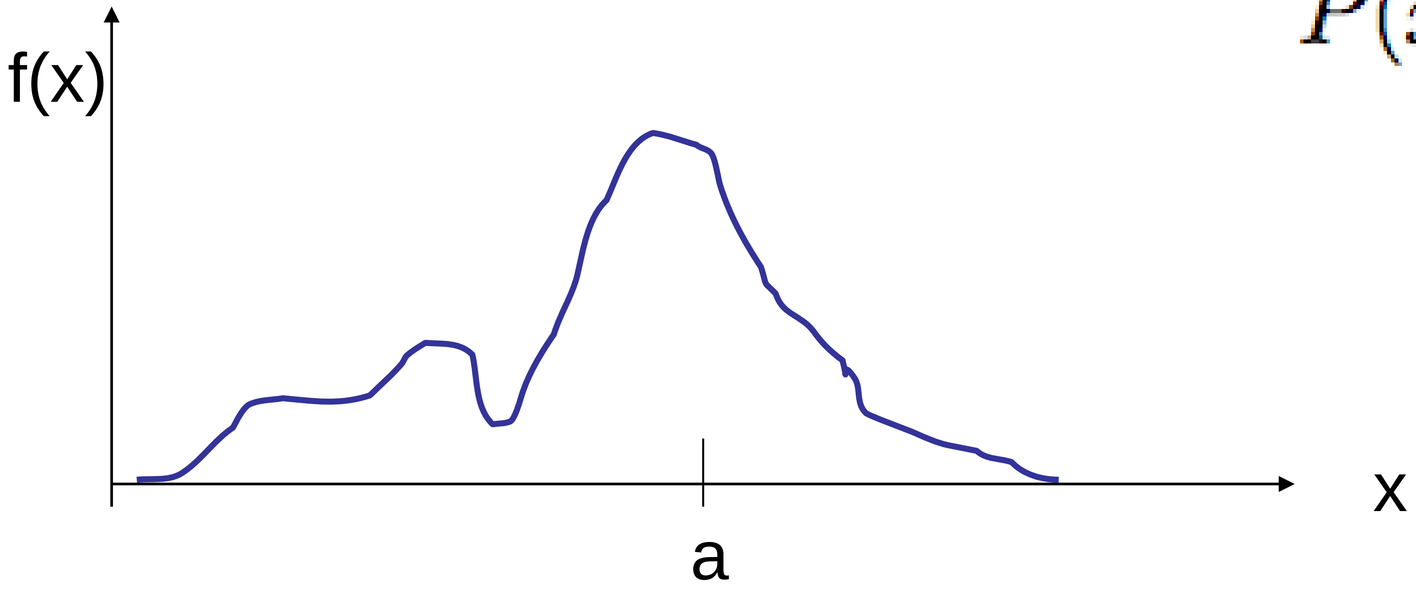
Probability Density Function

- Discrete distributions



$$\sum_i P(X = x_i) = 1$$

- Continuous: Cumulative Density Function (CDF): *F(a)*



$$P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau$$

Cumulative Density Functions

- Total probability

$$P(\Omega) = \int_{-\infty}^{\infty} f(x)dx = 1$$

- Probability Density Function (PDF)

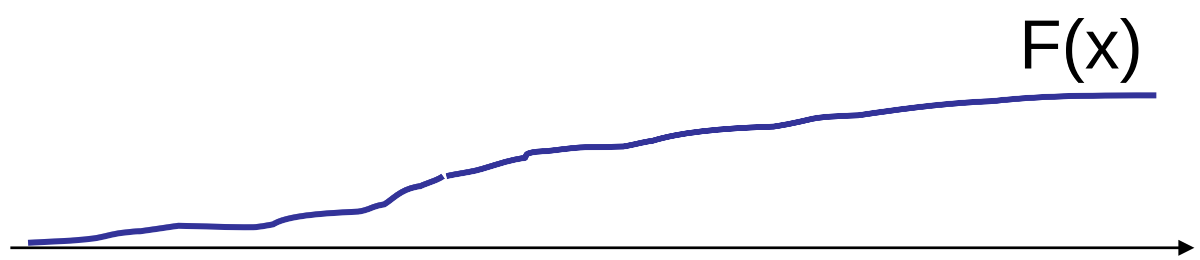
$$\frac{d}{dx}F(x) = f(x)$$

- Properties:

$$P(a \leq x \leq b) = \int_{b}^{a} f(x)dx = F(b) - F(a)$$

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$



$$F(a) \geq F(b) \ \forall a \geq b$$

Expectations

- Mean/Expected Value:

$$E[x] = \bar{x} = \int x f(x) dx$$

- Variance:

$$Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2$$

- In general:

$$E[x^2] = \int x^2 f(x) dx$$

$$E[g(x)] = \int g(x) f(x) dx$$

Multivariate

- Joint for (x,y)

$$P\left((x,y) \in A\right) = \int\int_A f(x,y)dxdy$$

- Marginal:

$$f(x) = \int f(x,y)dy$$

- Conditionals:

$$f(x|y) = \frac{f(x,y)}{f(y)}$$

- Chain rule:

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

Bayes Rule

- Standard form:

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

- Replacing the bottom:

$$f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

Binomial

- Distribution:

$$x \sim Binomial(p, n)$$

$$P(x = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Mean/Var:

$$E[x] = np$$

$$Var(x) = np(1-p)$$

Uniform

- Anything is equally likely in the region [a,b]

- Distribution:

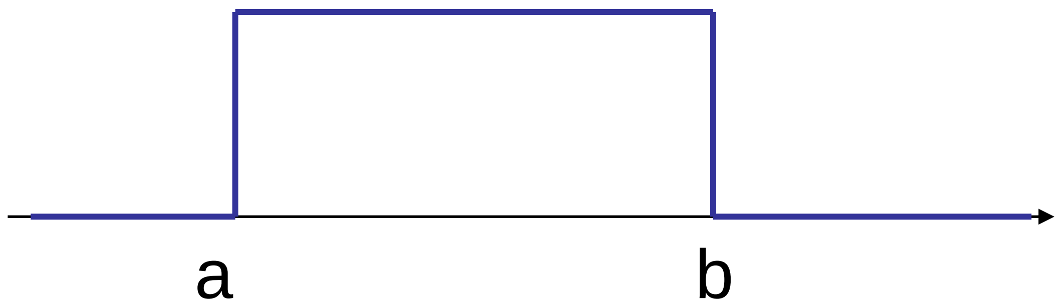
$$x \sim U(a,b)$$

- Mean/Var

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & otherwise \end{cases}$$

$$E[x] = \frac{a+b}{2}$$

$$Var(x) = \frac{a^2+ab+b^2}{3}$$

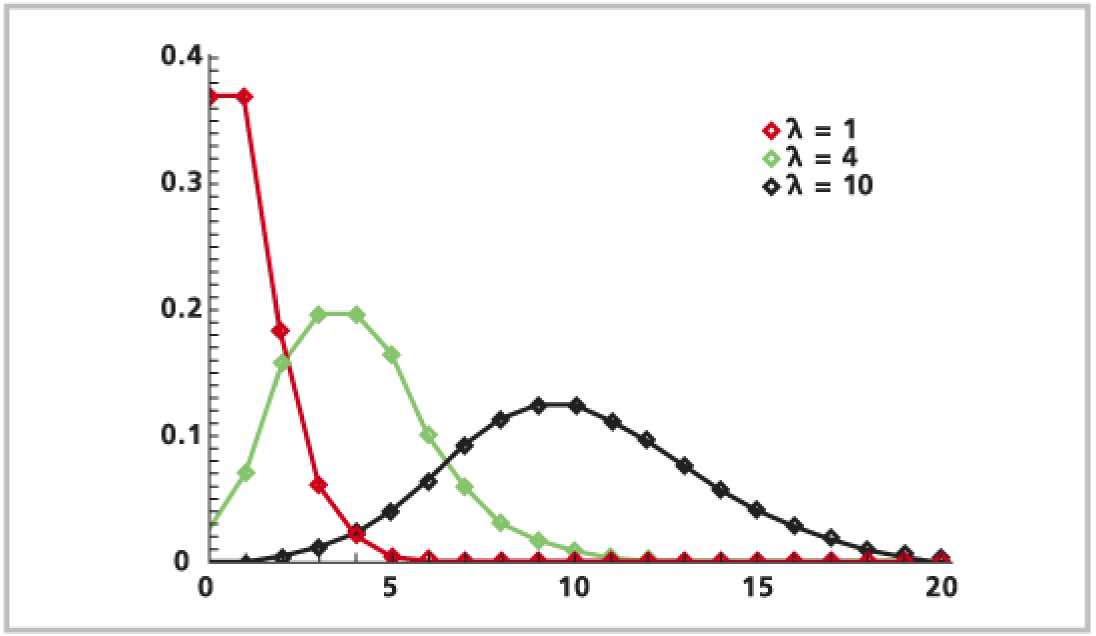


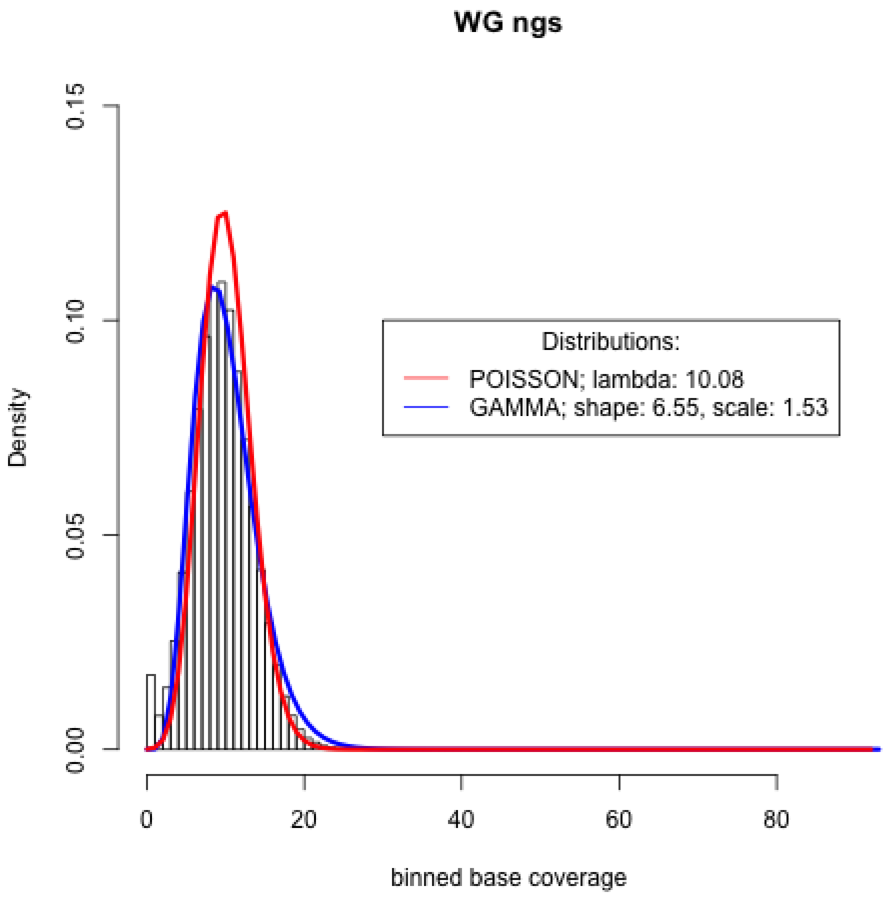
Poisson Distribution

$$p(x=k)=\frac{\lambda^k e^{-\lambda}}{k!}$$

- Discrete distribution
- Widely used in sequence analysis (read counts are discrete).
- λ is the expected value of x (the number of observations) and is also the variance:

$$E(x)=Var(x)=\lambda$$





Gaussian (Normal)

- If I look at the height of women in country xx, it will look approximately Gaussian
- Small random noise errors, look Gaussian/Normal

- Distribution:

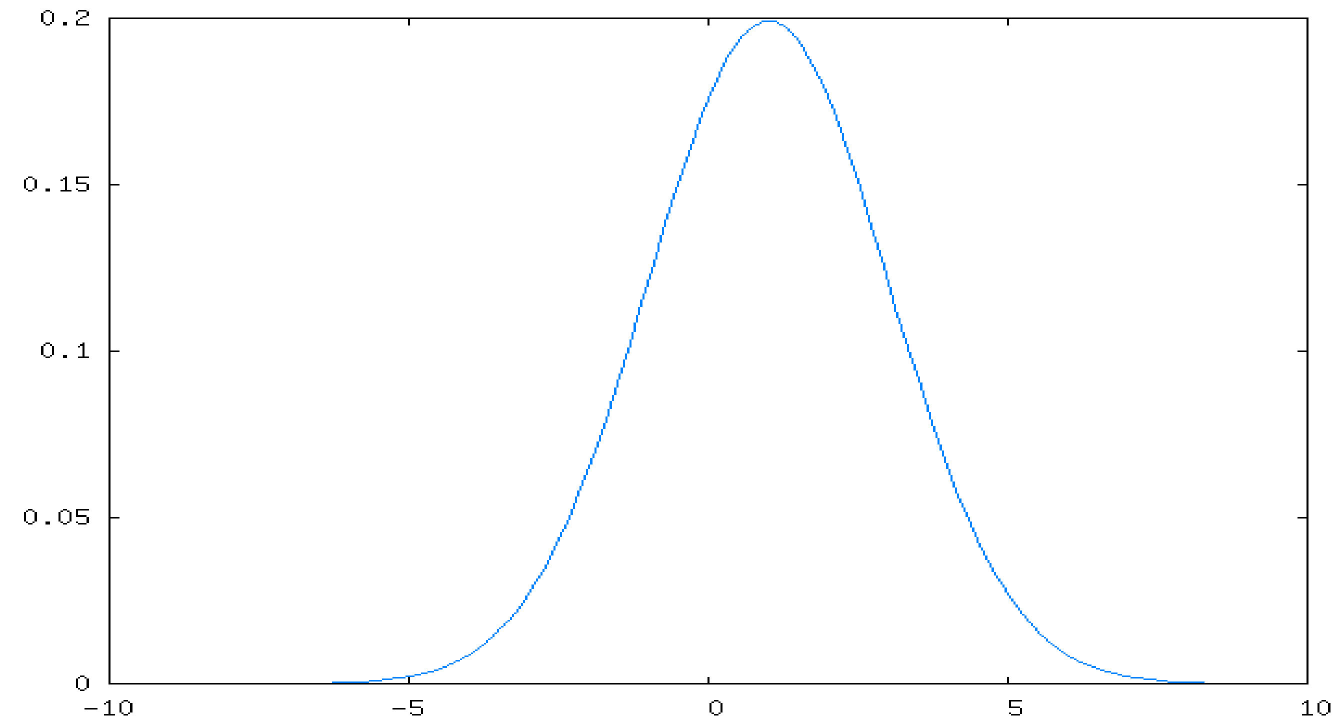
$$x \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Mean/var

$$E[x] = \mu$$

$$Var(x) = \sigma^2$$



Why Do People Use Gaussians

- Central Limit Theorem: (loosely)
  - Sum of a large number of IID random variables is approximately Gaussian

Multivariate Gaussians

- Distribution for vector x

$$x = (x_1, \ldots, x_N)^T, \quad x \sim N(\mu, \Sigma)$$

- PDF:

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

Multivariate Gaussians

$$f(x) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

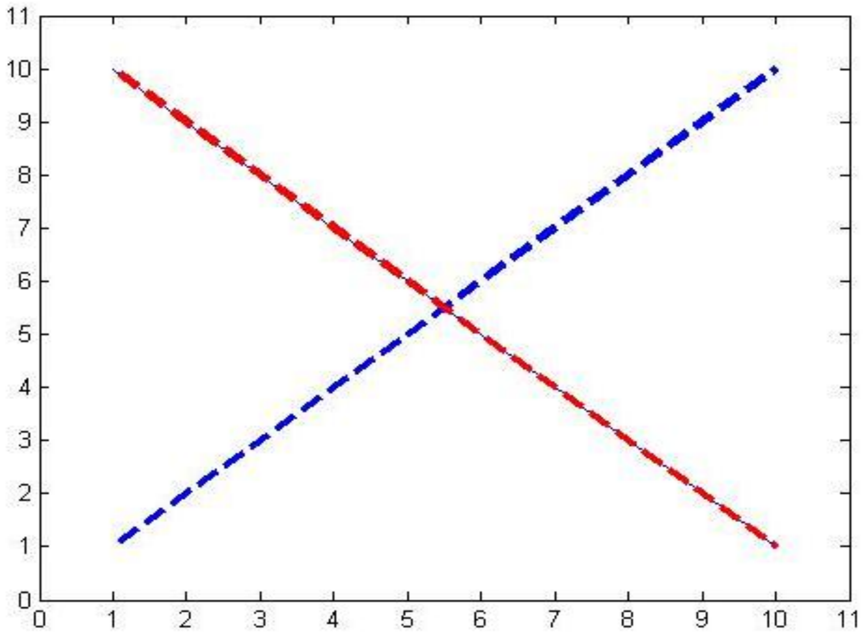
$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$Var(x) \rightarrow \Sigma = \begin{pmatrix} Var(x_1) & Cov(x_1, x_2) & \ldots & Cov(x_1, x_N) \\ Cov(x_2, x_1) & Var(x_2) & \ldots & Cov(x_2, x_N) \\ \vdots & & \ddots & \vdots \\ Cov(x_N, x_1) & Cov(x_N, x_2) & \ldots & Var(x_N) \end{pmatrix}$$

$$\mathrm{cov}(\chi_1, \chi_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - \mu_1)(x_{2,i} - \mu_2)$$

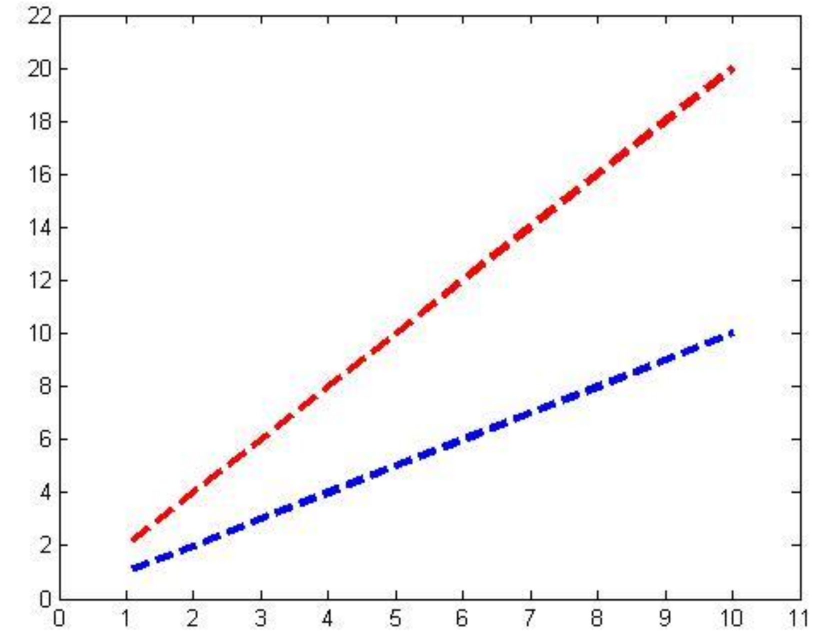
Covariance examples

Anti-correlated



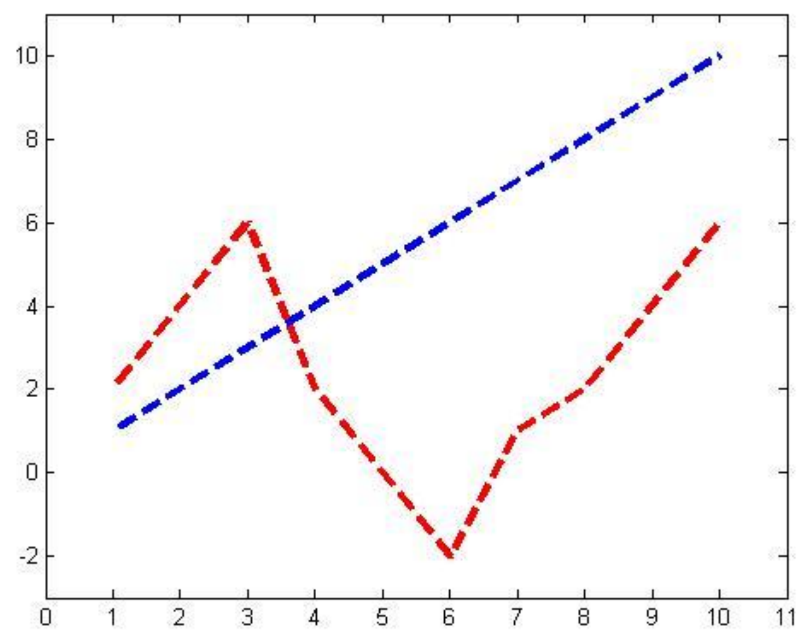
Covariance: -9.2

Correlated



Covariance: 18.33

Independent (almost)



Covariance: 0.6

(A few) key computational methods

Regression

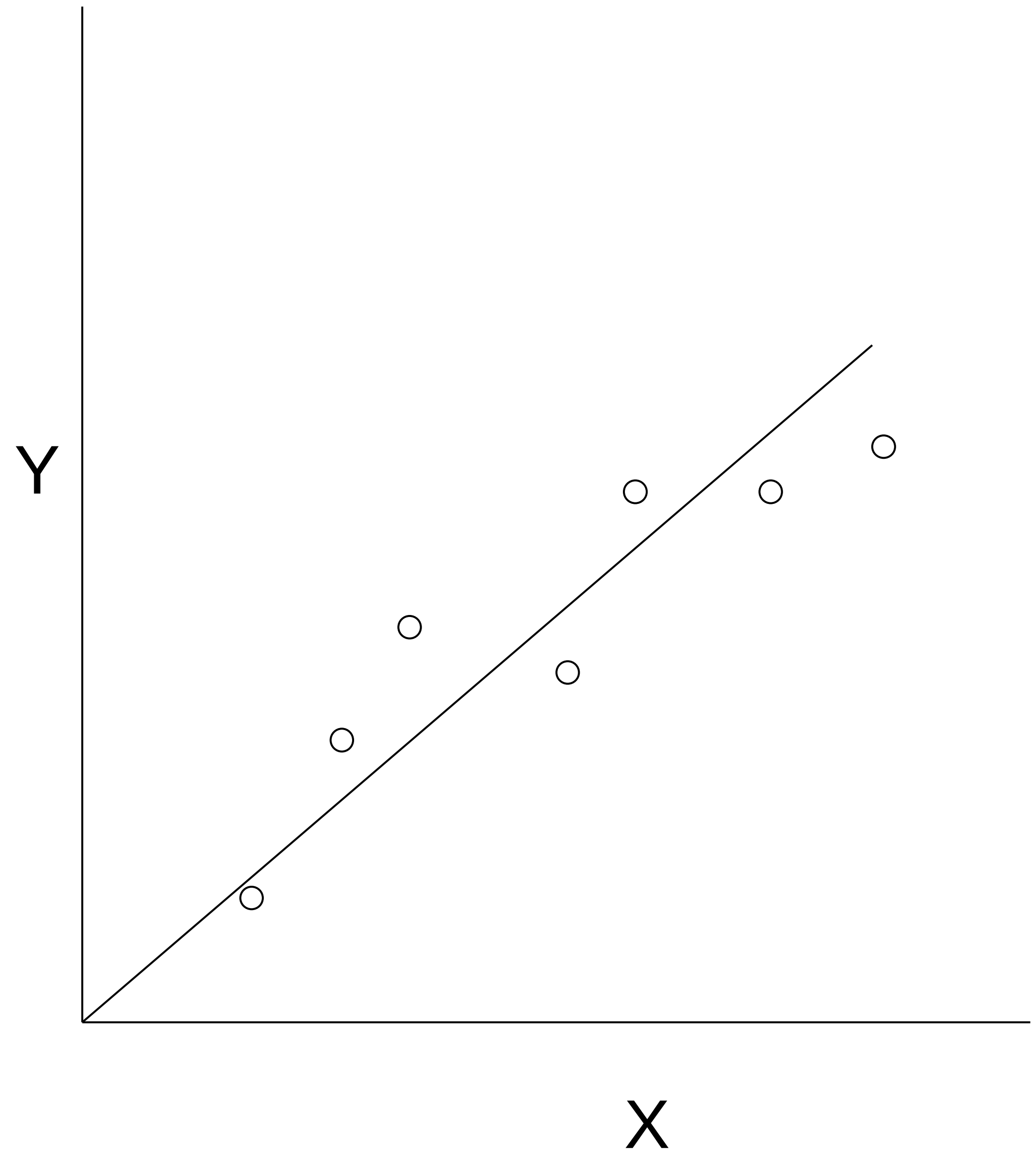
- Given an input x we would like to compute an output y
- In linear regression we assume that y and x are related with the following equation:

Observed values

What we are trying to predict

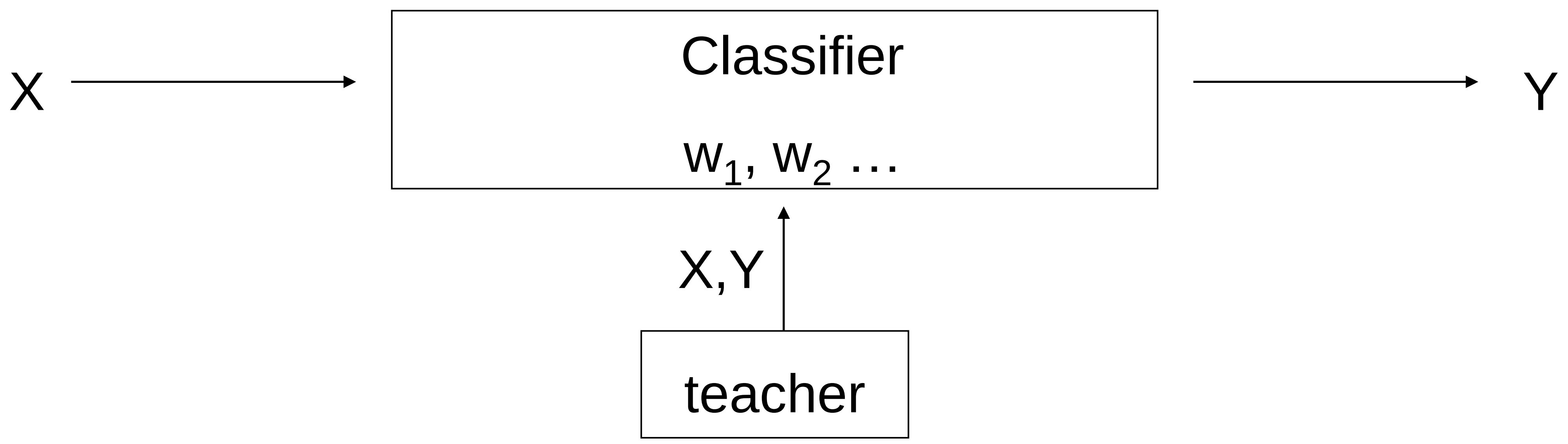
$$y = wx+\varepsilon$$

where w is a parameter and ε represents measurement or other noise



Supervised learning

- Classification is one of the key components of 'supervised learning'
- In supervised learning the teacher (us) provides the algorithm with the solutions to some of the instances and the goal is to generalize so that a model / method can be used to determine the labels of the unobserved samples



Types of classifiers

- We can divide the large variety of classification approaches into roughly two main types

1. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors

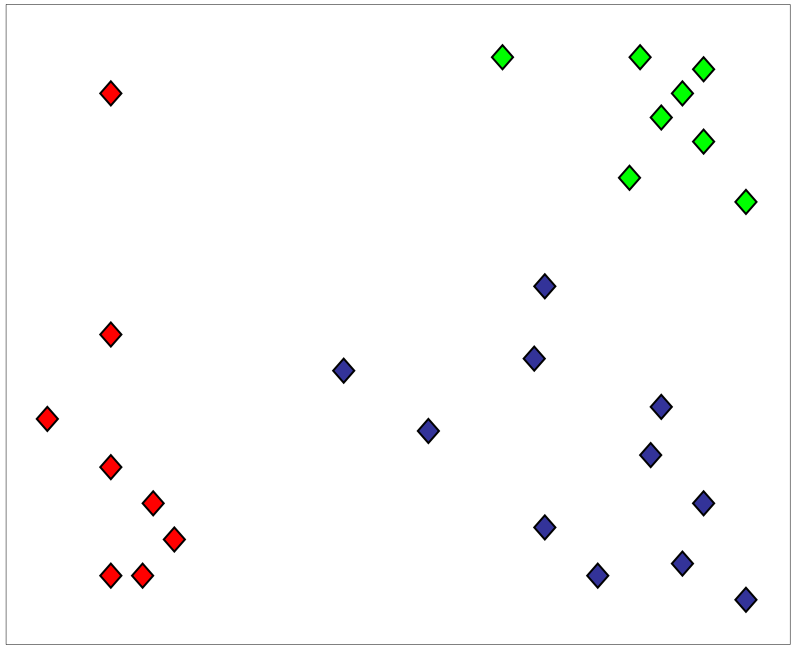
2. Generative:
   - build a generative statistical model
   - e.g., Naïve Bayes

3. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., decision tree, SVM

Unsupervised learning

We do not have a teacher that provides examples with their labels

- Goal: Organize data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects



Graphical models: Sparse methods for representing joint distributions

- Nodes represent random variables
- Edges represent conditional dependence
- Can be either directed (Bayesian networks, HMMs) or undirected (Markov Random Fields, Gaussian Random Fields)

Bayesian networks

Bayesian networks are directed acyclic graphs.