



Royal Institute of Technology

# DD2447 STAT. METH. IN CS HT 2014

★ Lecture 1-Intro  
“Chapter 1”



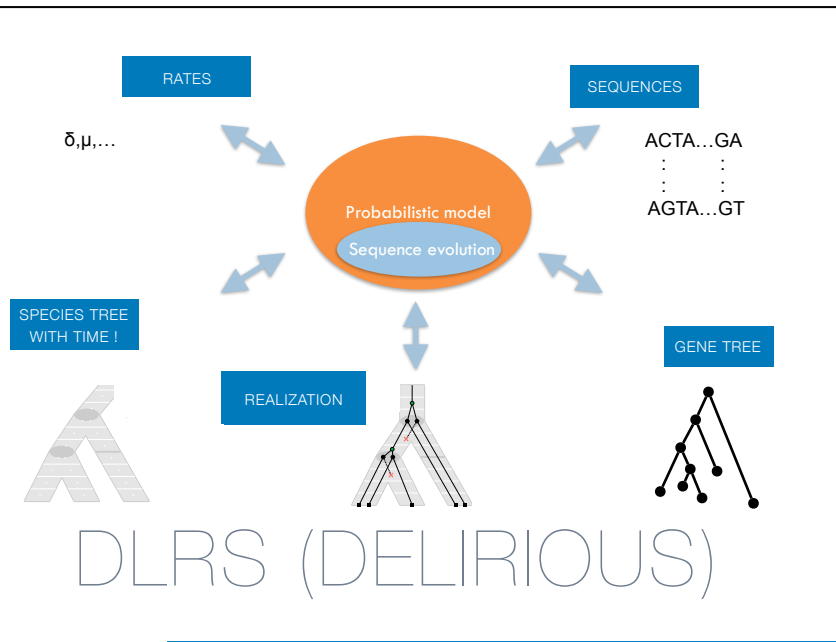
Royal Institute of Technology

# SciLifeLab

Computational Biology

Machine Learning – a main tool

Jens Lagergren



**THE TEXT BOOK**

New

Key machine learning researcher

Comprehensive

**Machine Learning**  
A Probabilistic Perspective

Kevin P. Murphy



- ★ Exercises – not graded
- ★ Homework (3) – Individual, handed in and graded
  - ★ out – nov 14, nov 28, dec 12
- ★ Project (1) – Individual, handed in and graded
  - ★ out – dec 12

## ADMINISTRATION - EXAMINATION

1. Use latex, you will have to use it later anyway (this is a recommendation)
2. Always include your name in the file with the solutions
3. Make each step in a derivation explicit

## KRISTOFFER'S COMMENTS

STATISTICAL METHODS IN APPLIED COMPUTER SCIENCE  
DD2447 | 6.0 CREDITS

Tools

Course overview

News feed

Schedule

General

Course plan etc

Honor code

Course wiki

statmet13

Reading list

Start, book etc.

KTH / COURSE WEB / STATISTICAL METHODS IN APPLIED COMPUTER SCIENCE

### Statistical Methods in Applied Computer Science

Selection: general and for Your course rounds/ groups

Welcome to the DD2447 course website!

The following information concerns statmet13 (i.e., the course given during fall semester 2013).

The course starts FRIDAY NOVEMBER 8. There will be NO lectures starting 8:00 in the morning any day. Please come to the first lecture since we will reschedule the wednesday lectures by voting on a better time and day.

Before the course starts a reading list, planned dates for assignments and laborations will be published under statmet13, see left menu. During the course slides will be provided in pdf format under statmet13 -> slides. As last year, we will be using Kevin P. Murphy's book "Machine Learning: a Probabilistic Perspective". According to Kårbokhandeln, they will have the book available at the course start.

## INTERACTION

- Lectures
- Solutions: mail, Scilife, or lectures
- KTH social [www.kth.se/social/course/DD2447/](http://www.kth.se/social/course/DD2447/)

## WHY MACHINE LEARNING?

- ★ The era of big data
- ★ Transaction data for large corporations
  - ★ Walmart has 2.5 petabytes ( $2.5 \cdot 10^{15}$ ) and handle 1M/hour
- ★ A human genome is 6 Gb
- ★ Meta-genomics
- ★ Baltic sea, hot-springs, your gut
- ★ A coke can can contain more microbes than there are north-americans

# STATISTICS, ML, DATA MINING?

- ★ Statistics — closed formulas
- ★ Statistical ML — computational methods
  - ★ they share models, probability
- ★ We will often apply a Bayesian approach
- ★ Data mining — less mathematical

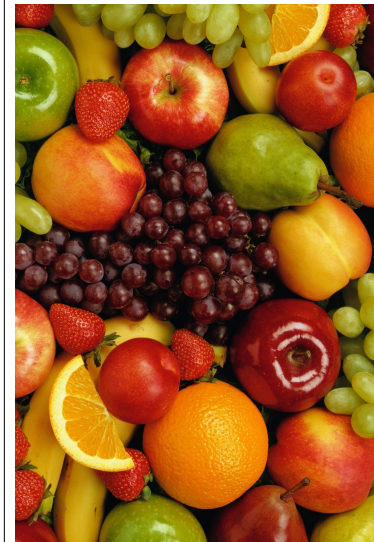
# SOME STUFF I EXPECT YOU TO KNOW

- ★ Supervised learning
- ★ Unsupervised learning
- ★ Training & testing



We have the answer ← means yes otherwise no

## SUPERVISED LEARNING



## UNSUPERVISED LEARNING

We do not have any correct answer

Find classes or groups

# SOME STUFF I EXPECT YOU TO KNOW

- ★ Supervised learning
  - ★  $D = \{(\mathbf{x}_i, y_i)\}$
  - ★  $y_i$  response variable (output variable)
  - ★  $\mathbf{x}_i$  features (input variables)
  - ★ classification & regression
- ★ Unsupervised learning
  - ★ find the right  $y_i$ 's, or
  - ★ find the right dependencies between the variables of  $\mathbf{x}_i$



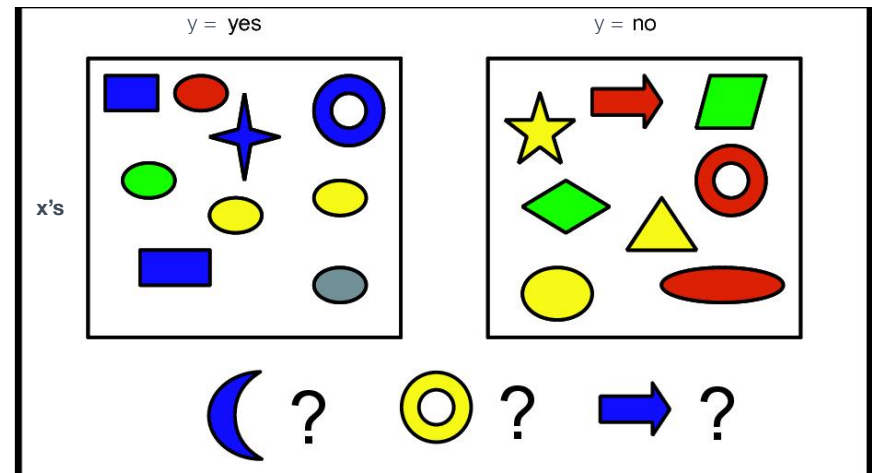
We have the answer ← means yes otherwise no

## BINARY CLASSIFICATION



★ ← means apple, ← means pear, otherwise other

## CATEGORICAL CLASSIFICATION

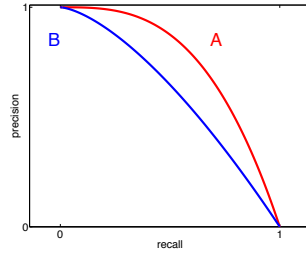


## CLASSIFICATION



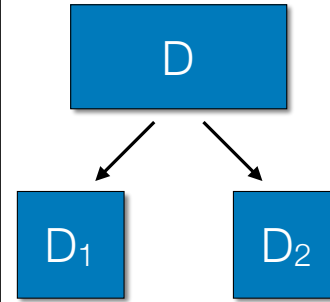
The probability of an answer

$$p(\text{yes} \mid \text{blue moon}, \mathcal{D})$$



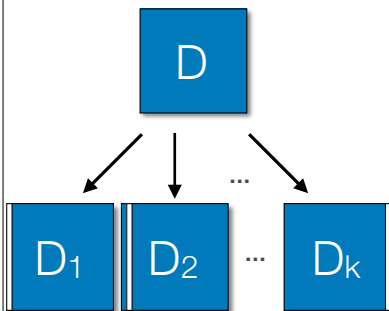
## PROBABILISTIC PREDICTION

## TRAINING – TEST

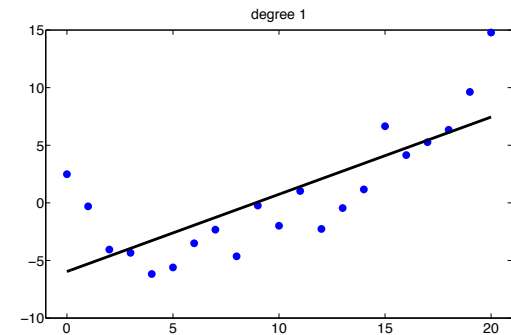
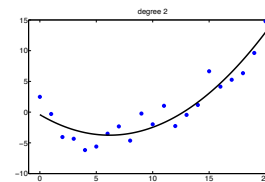


- Split data  $D$  into
  - training  $D_1$
  - test  $D_2$
- Use misclassification rate
- Problem: overfitting

## CROSS-VALIDATION



- Leave-one-out
  - let  $D_i = D \setminus x_i$
  - test on  $x_i$
- Use misclassification rate
- Redundancy, overlap



- ★ Size
- ★ Floor
- ★ Location

## REGRESSION

- ★ Googles smartass (ad selection system)
- ★ personalisation
- ★ Mail filter
- ★ Handwriting recognition
- ★ MNIST a dataset with 60000 training and 6.000 test images (of digits 0,..., 9)
- ★ Face recognition
- ★ Differentiate between setosa, versicolor, and virginica ???

## REAL WORLD APPLICATIONS

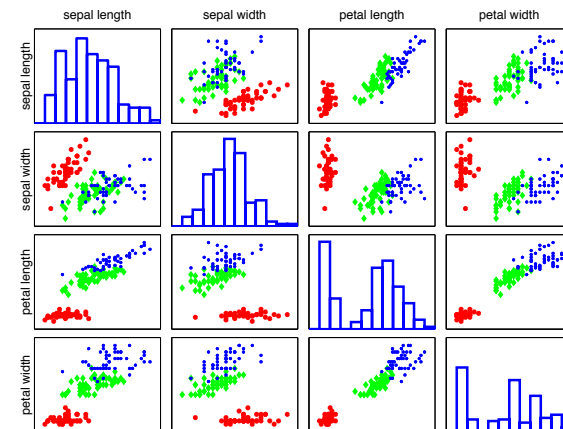
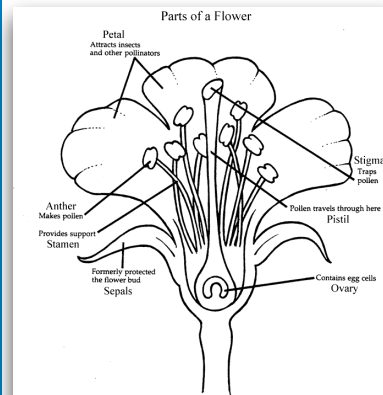


SETOSA, VERSICOLOR, AND VIRGINICA

## HIGH SCHOOL BIOLOGY

Petal – attracting insects and pollinators

Sepal – formerly protected the bud



SETOSA, VERSICOLOR AND VIRGINICA

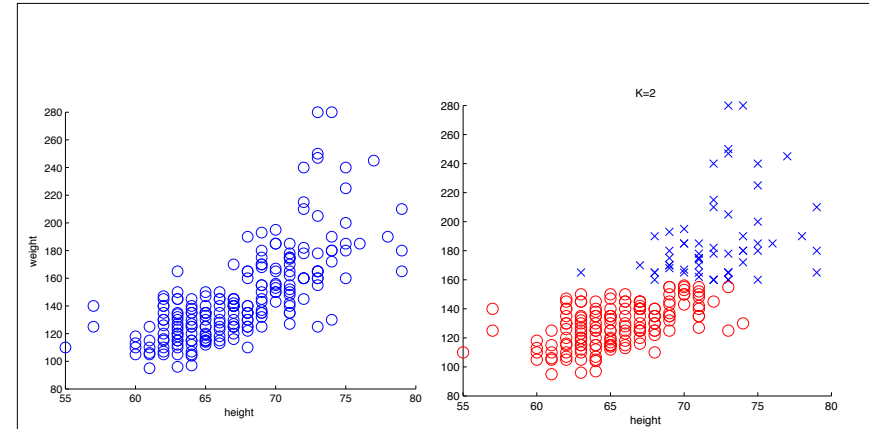
★ sepal length, sepal width, petal length, and petal width.



# UNSUPERVISED LEARNING

We do not have any correct answer

Find classes or groups



- \* Each subset should contain similar points
- \* Pairs of subsets should have dissimilar points.

# CLUSTERING

Molecular breast cancer data

5 subtypes



# HIERARCHICAL CLUSTERING

If you like **Arthur Russell**, try **The Clientele**.



**The Clientele**  
2,416 Followers

You listened to **Mark Kozelek** and **The Black Swans**. Here's an album you might like.



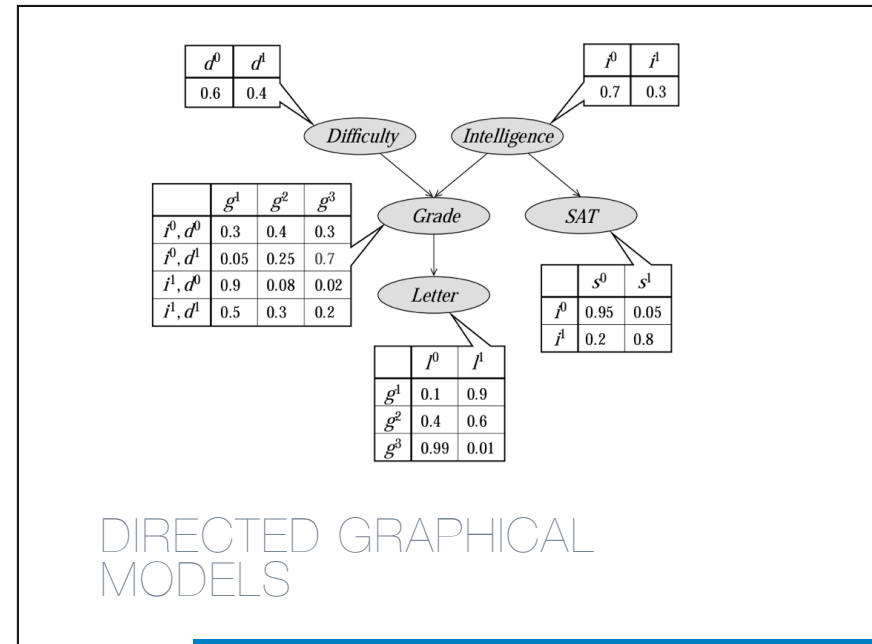
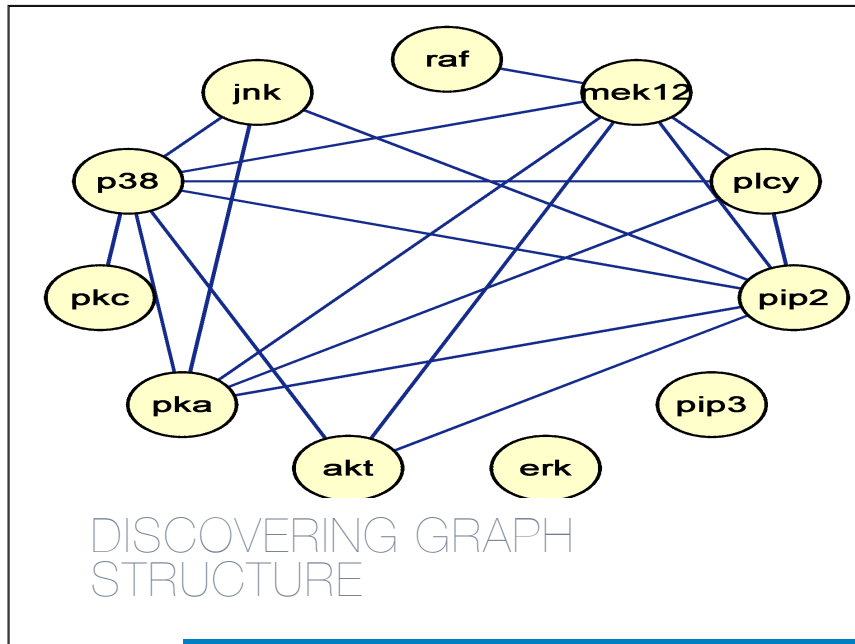
**Gentle Stream**  
The Amazing

You listened to **Emily and The Woods**. Check out **Peasant**.



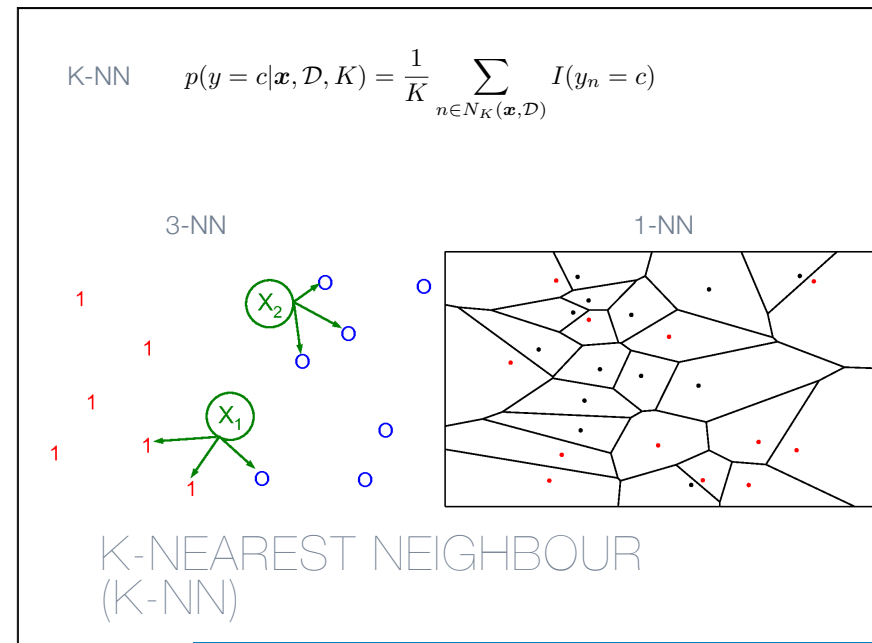
**Bound for Glory**  
Peasant

# COLLECTIVE FILTERING

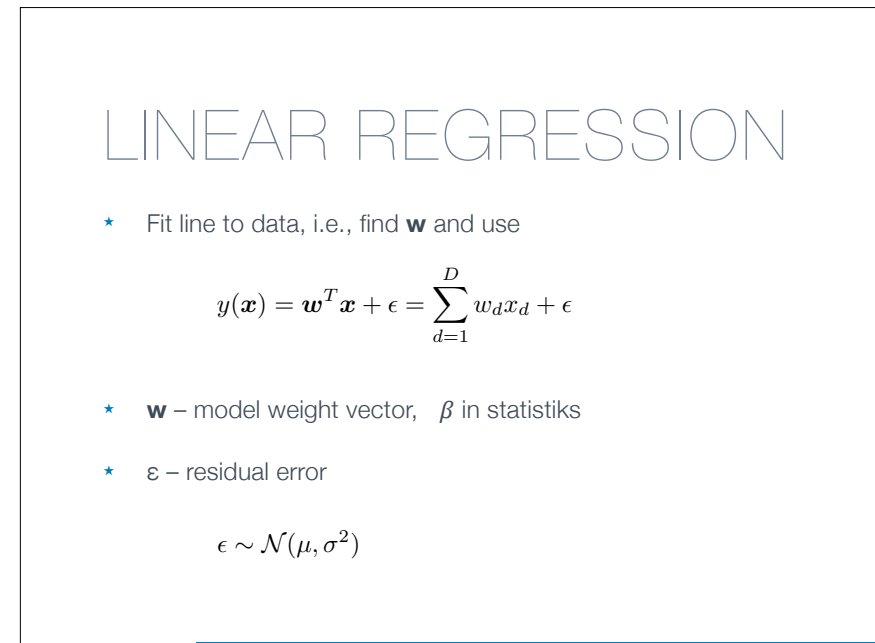
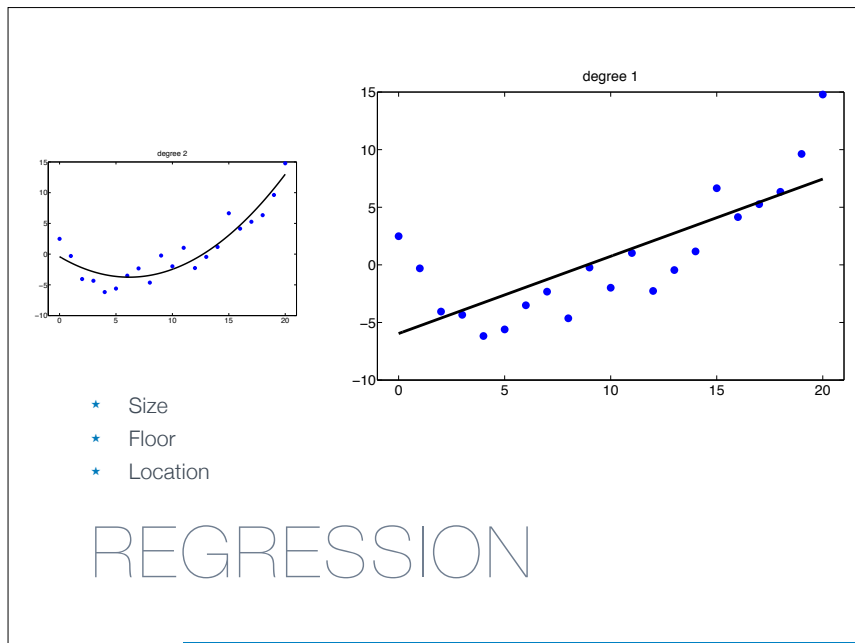
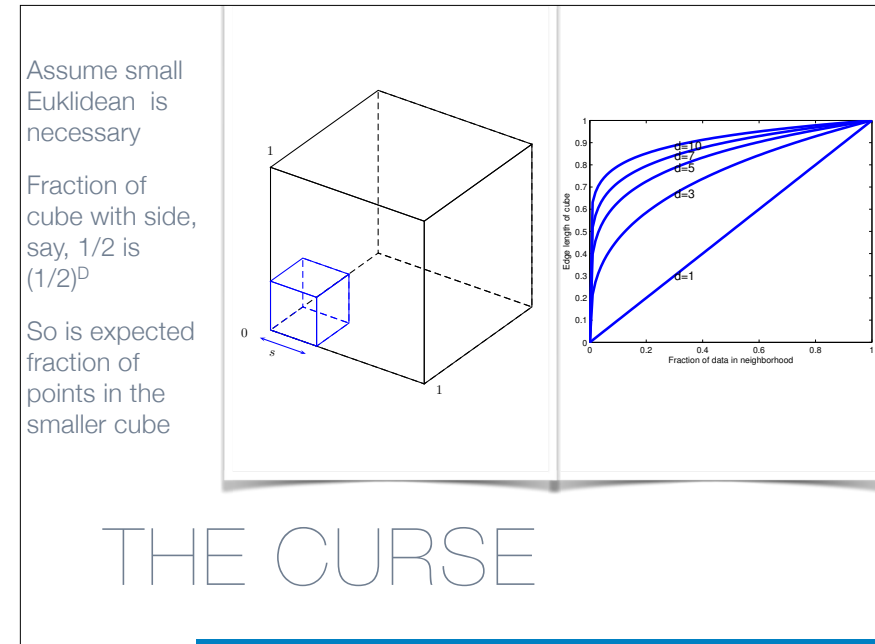
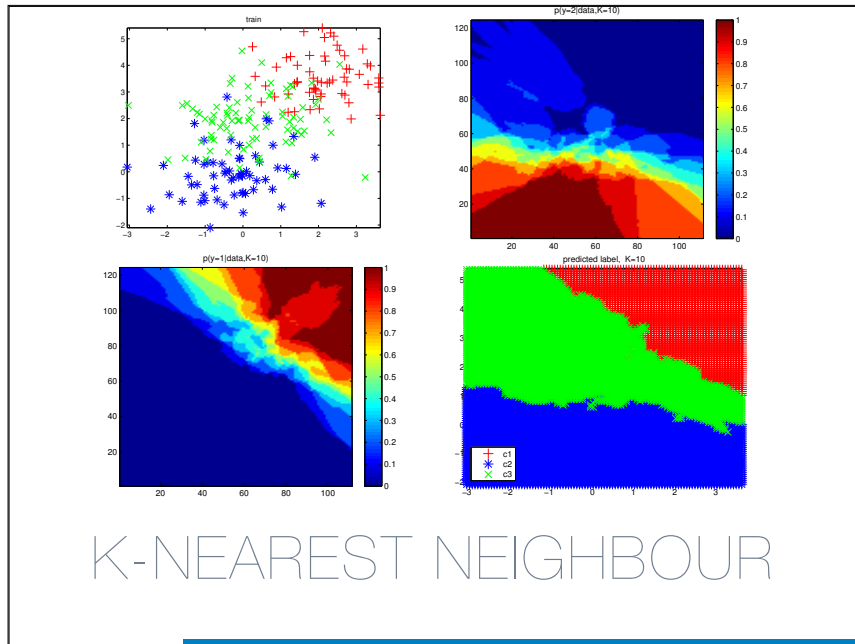


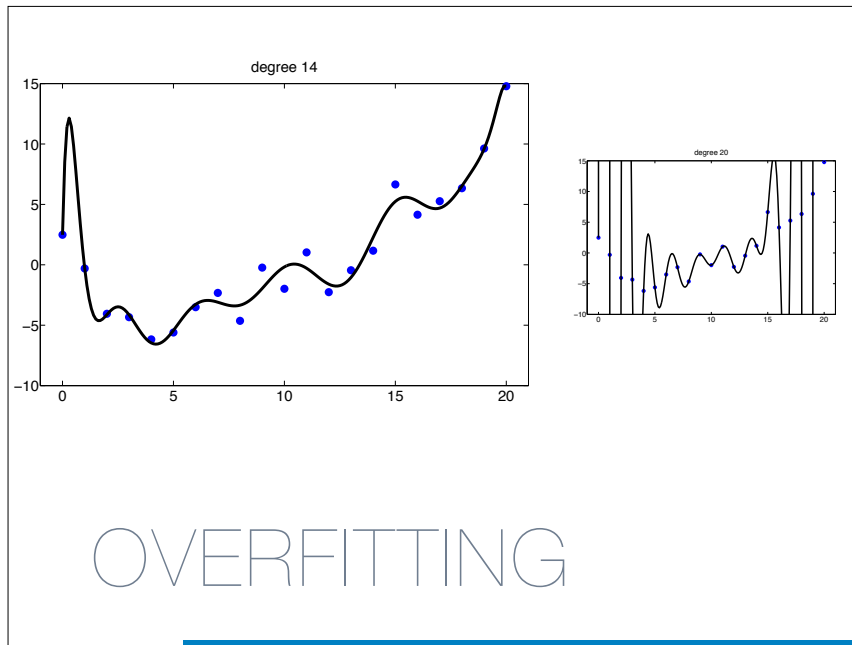
### PARAMETRIC VS NON-PARAMETRIC

- ★ Constant # parameters – parametric model (any distribution)
- ★ Representation grows with data – non-parametric model










# BAYESIAN


Fair (F)



$P(i|F) = \frac{1}{6}$

Used 99%

Biased/loaded (B)



$P(6|B) = \frac{1}{2}$

$P(6|B) = \frac{1}{10}, \forall i \in [5]$

Used 1%

- ★ A is the event 6,6,6
- ★ Bayesian
- ★ We get  $P(M|A) = \frac{P(A|M)P(M)}{P(A)}$  ← The same for F & B

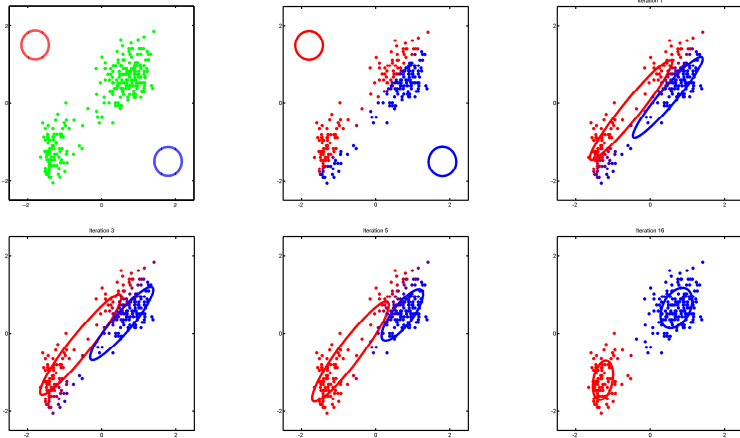
$$P(A|B)P(B) = \frac{1}{2}^3 * 0.01 < P(A|F)P(F) = \frac{1}{6}^3 * 0.99$$

- # SOME THOUGHTS ON MODELING
- ★ All models are wrong, but some are useful.
  - ★ Models are what we call the lies we are used to
  - ★ There are no model free approaches!
  - ★ use the term assumption instead
  - ★ Using models is a way to make assumptions explicit.
  - ★ Bayesian is a non-deterministic logic.

- # K-MEANS
- ★ Data vectors  $D = \{x_1, \dots, x_N\}$
  - ★ Randomly selected classes  $z_1, \dots, z_N$
  - ★ Iteratively do
 
$$\mu_c = \frac{1}{N_c} \sum_{n: z_n = c} x_n, \quad \text{where } N_c = |\{n : z_n = c\}|$$

$$z_n = \operatorname{argmin}_c \|x_n - \mu_c\|_2$$
  - ★ One step  $O(NKD)$ , can be improved

# EXAMPLE



## Expected complete; notation

$$\begin{aligned}
 \log p(\mathbf{x}_n | \theta') &= \log \sum_{z_n} p(\mathbf{x}_n, z_n | \theta') \\
 &= \log \sum_{z_n} p(z_n | \mathbf{x}_n, \theta) \frac{p(\mathbf{x}_n, z_n | \theta')}{p(z_n | \mathbf{x}_n, \theta)} \\
 &= \log E_{z_n} \left( \frac{p(\mathbf{x}_n, z_n | \theta')}{p(z_n | \mathbf{x}_n, \theta)} \mid \mathbf{x}_n, \theta \right) \\
 &\geq \text{Jensen} E_{z_n} \left( \log \frac{p(\mathbf{x}_n, z_n | \theta')}{p(z_n | \mathbf{x}_n, \theta)} \mid \mathbf{x}_n, \theta \right) \\
 &= \sum_{z_n} p(z_n | \mathbf{x}_n, \theta) \log \frac{p(\mathbf{x}_n, z_n | \theta')}{p(z_n | \mathbf{x}_n, \theta)} \\
 &= \sum_{z_n} p(z_n | \mathbf{x}_n, \theta) \log p(\mathbf{x}_n, z_n | \theta') - \sum_{z_n} p(z_n | \mathbf{x}_n, \theta) \log p(z_n | \mathbf{x}_n, \theta) \\
 &= Q_n(\theta'; \theta) - R_n(\theta; \theta)
 \end{aligned}$$

# COMBINING INFO FROM VARIOUS SENSORS



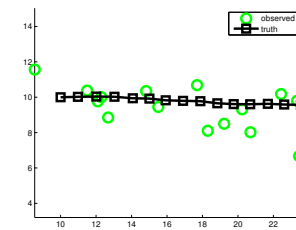
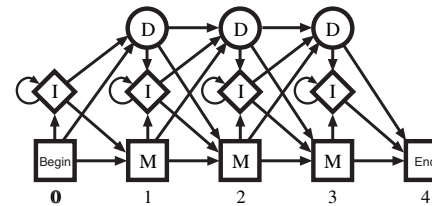
- Inertial sensors
- Camera
- Barometer

- Inertial sensors
- Radar
- Barometer
- Map

- Inertial sensors
- Cameras
- Radars
- Wheel speed sensors
- Steering wheel sensor

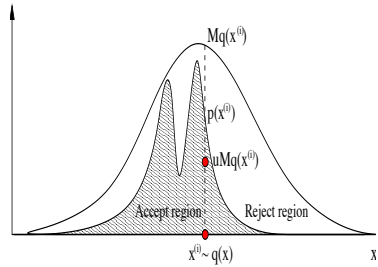
- Inertial sensors
- Ultra-wideband

Aim: Motion capture, find the motion (position, orientation, velocity and acceleration) of a person (or object) over time.



# HMMS ANS SSM

# REJECTION SAMPLING



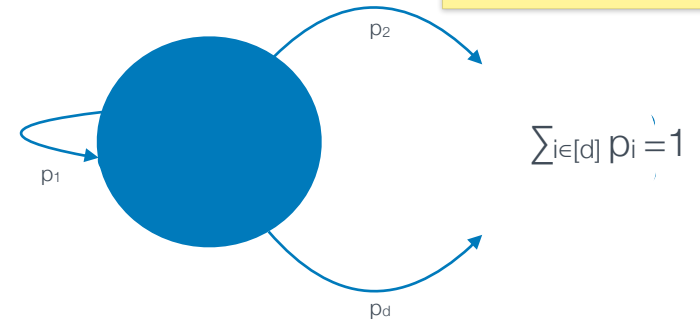
Algorithm

- ★ sample  $x \sim q(x)$
- ★ sample  $u \sim U(0, 1)$
- ★ if  $uMq(x) \leq p(x)$ , accept (and output  $x$ )

# MARKOV CHAIN (DISCRETE)

importans sampling  
kalman filter  
partical filter  
gibbs sampling

Probabilities on outgoing edges sum to one



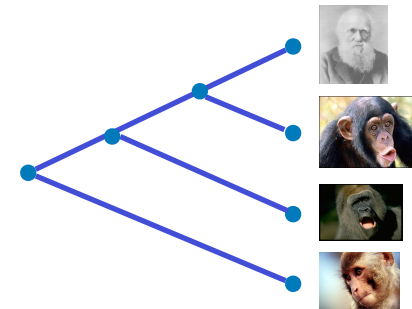
# IS THE CHIMP OUR CLOSEST RELATIVE?



# PHYLOGENY

Input: species

Output: tree where proximity correlates with similarity



# MR BAYES

**BIOINFORMATICS APPLICATIONS NOTE** Vol. 19 no. 12, 2003, pages 1572–1574  
DOI: 10.1093/bioinformatics/btg1190



## MrBayes 3: Bayesian phylogenetic inference under mixed models

Fredrik Ronquist<sup>1,\*</sup> and John P. Huelsenbeck<sup>2</sup>

<sup>1</sup>Department of Systematic Zoology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden and <sup>2</sup>Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093-0116, USA

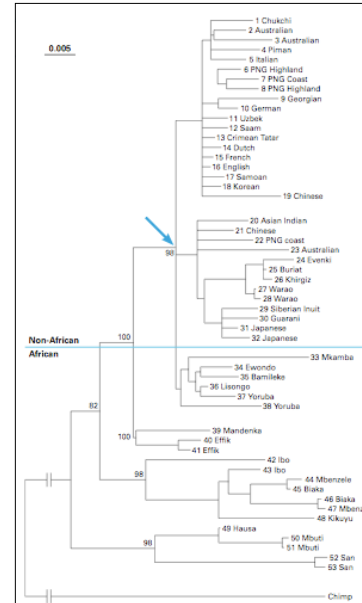
Received on December 20, 2002; revised on February 14, 2003; accepted on February 19, 2003



Prof. Entomology

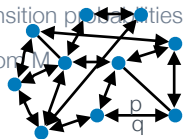


# OUR ORIGIN



# MCMC

- ★ In order to sample from  $p$ , set up a MC M
- ★ select transition probabilities so that  $p =$  stationary distribution
- ★ sample from  $M$



# CONTENT

- ★ Chapter 1: Introduction.
- ★ Chapter 2: Probability.
- ★ Chapter 3: Generative models for discrete data.
- ★ Chapter 4: Gaussian models.
- ★ Chapter 5: Bayesian statistics.
- ★ Chapter 10: Directed graphical models. Probably deeper.
- ★ Chapter 18: SSM (HMMs)
- ★ (Chapter 19. Undirected graphical models.)
- ★ Chapter 20 Exact inference for graphical models.
- ★ Chapter 23: Monte Carlo inference.
- ★ Chapter 24: Markov Chain Monte Carlo.
- ★ Chapter 25: Clustering
- ★ Chapter 26: Graphical Model Structure Learning
- ★ Particle MCMC