CS434a/541a: Pattern Recognition
Prof. Olga Veksler

Lecture 1

Outline of the lecture

- Syllabus
- Introduction to Pattern Recognition
- Review of Probability/Statistics

Syllabus

- Prerequisite
  - Analysis of algorithms (CS 340a/b)
  - First-year course in Calculus
  - Introductory Statistics (Stats 222a/b or equivalent)
  - Linear Algebra (040a/b)

will review

- Grading
  - Midterm 30%
  - Assignments 30%
  - Final Project 40%

Syllabus

- Assignments
  - bi-weekly
  - theoretical or programming in Matlab or C
  - no extensive programming
  - may include extra credit work
  - may discuss but work individually
  - due in the beginning of the class
- Midterm
  - open anything
  - roughly on November 8

Syllabus

- Final project
  - Choose from the list of topics or design your own
  - May work in group of 2, in which case it is expected to be more extensive
  - 5 to 8 page report
  - proposals due roughly November 1
  - due December 8

Intro to Pattern Recognition

- Outline
  - What is pattern recognition?
  - Some applications
  - Our toy example
  - Structure of a pattern recognition system
  - Design stages of a pattern recognition system

What is Pattern Recognition ?

- *Informally*
  - Recognize patterns in data
- *More formally*
  - Assign an object or an event to one of the several pre-specified categories (a category is usually called a class)

tea cup

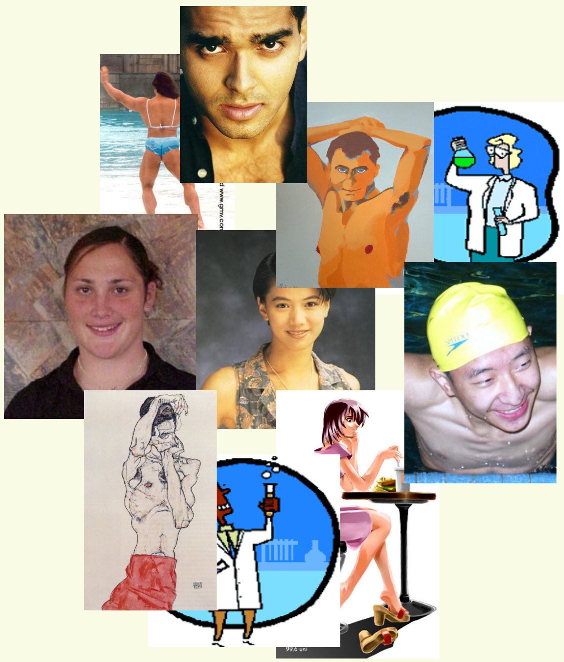
face

phone

Application: male or female?

classes

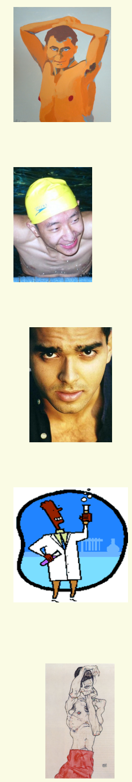
Objects (pictures)



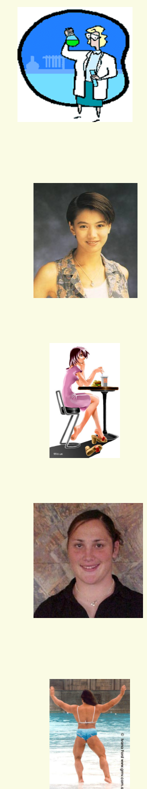
Perfect

PR system

male



female



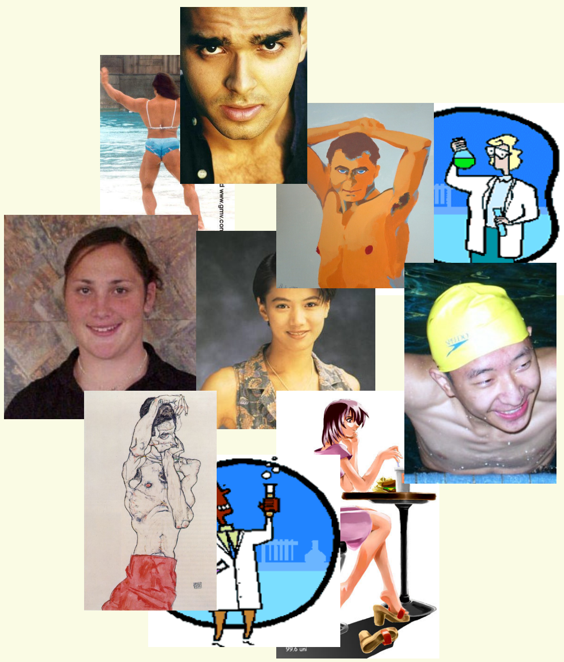
Application: photograph or not?

classes

photo

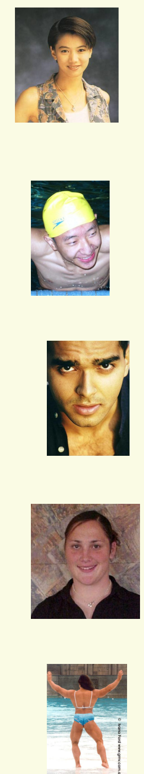
not photo

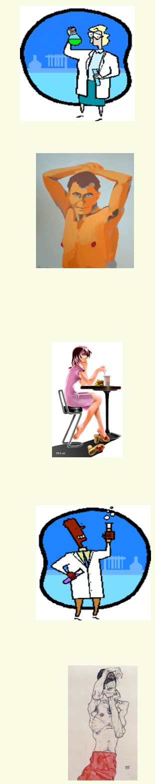
Objects (pictures)



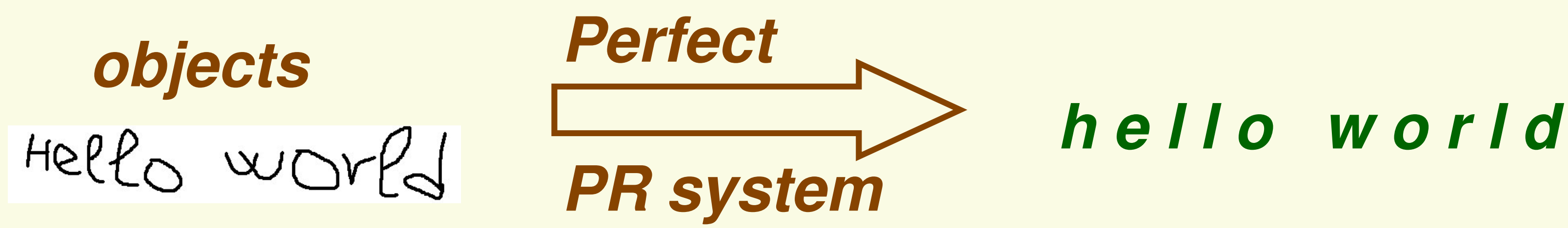
Perfect

PR system



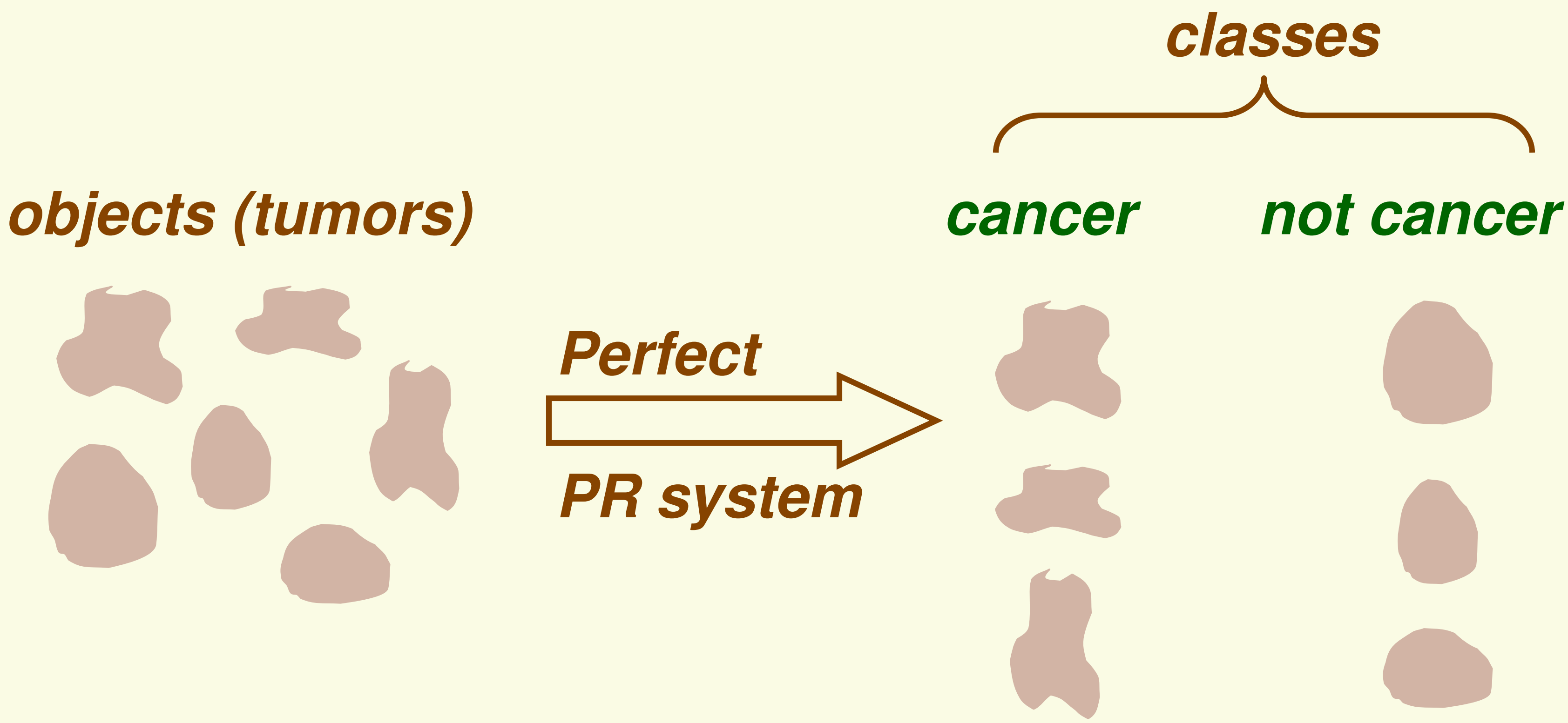


Application: Character Recognition



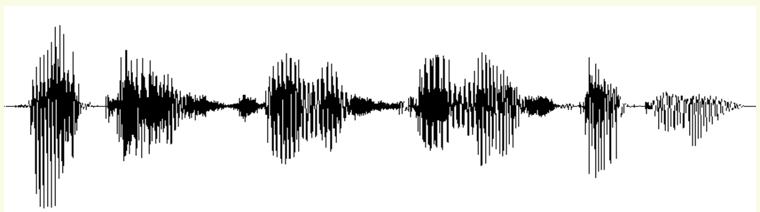
- In this case, the classes are all possible characters: **a, b, c,…., z**

Application: Medical diagnostics



Application: speech understanding

objects (acoustic signal) ***phonemes***



Perfect PR system → ***re-kig-'ni-sh&n***

- In this case, the classes are all phonemes

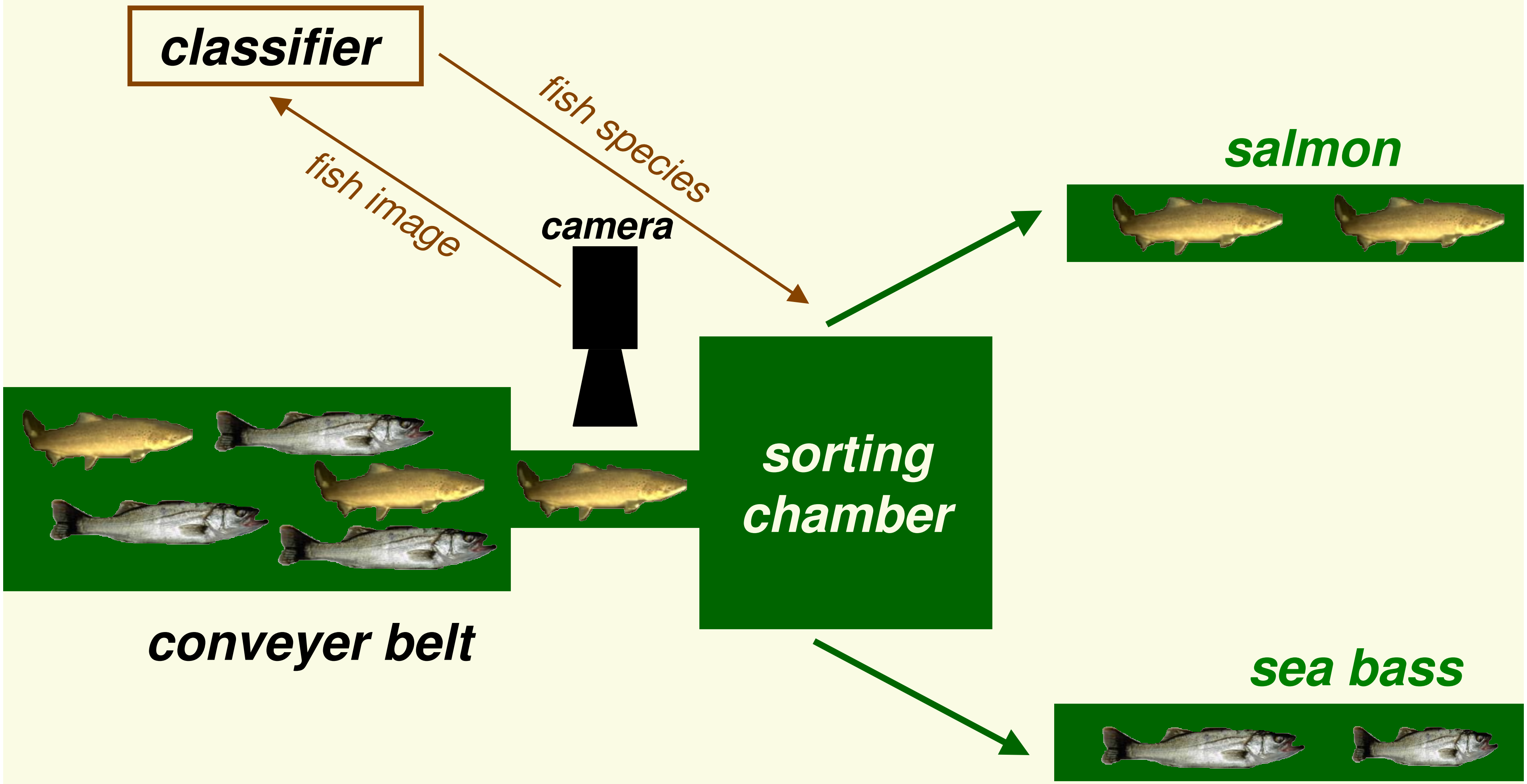
Application: Loan applications

classes

objects (people)

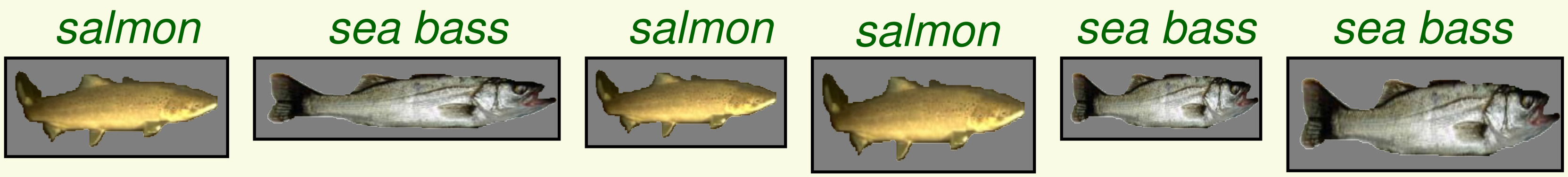
	income	debt	married	age	approve	deny
John Smith	200,000	0	yes	80		☑
Peter White	60,000	1,000	no	30	☑	
Ann Clark	100,000	10,000	yes	40	☑	
Susan Ho	0	20,000	no	25		☑

Our Toy Application: fish sorting



How to design a PR system?

- Collect data (training data) and classify by hand



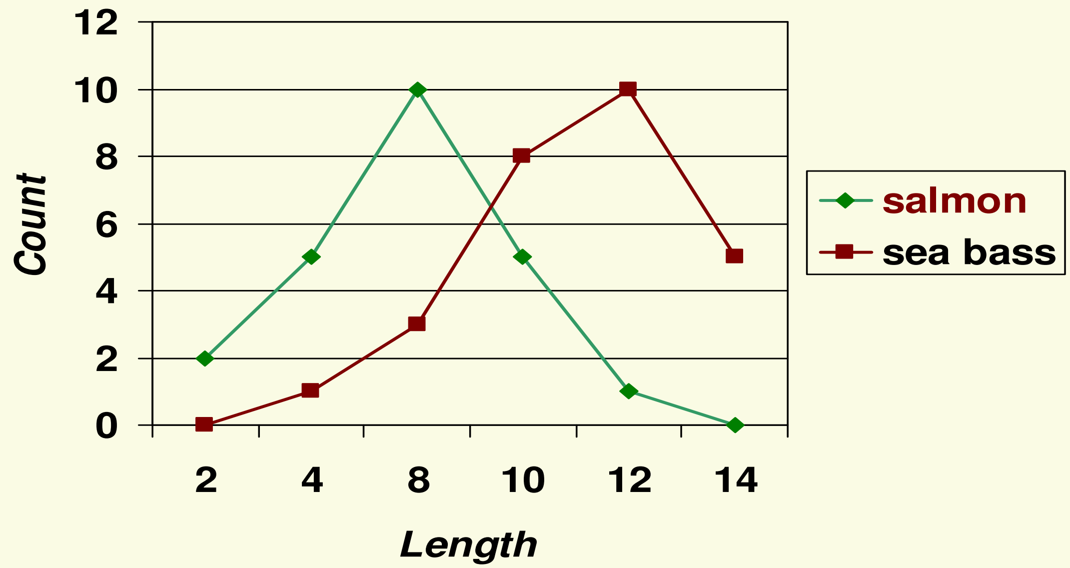
- Preprocess by segmenting fish from background

- Extract possibly discriminating features
  - length, lightness,width,number of fins,etc.
- Classifier design
  - Choose model
  - Train classifier on part of collected data (training data)
- Test classifier on the rest of collected data (test data) i.e. the data not used for training
  - Should classify new data (new fish images) well

Classifier design

- Notice salmon tends to be shorter than sea bass
- Use *fish length* as the discriminating feature
- Count number of bass and salmon of each length

	2	4	8	10	12	14
bass	0	1	3	8	10	5
salmon	2	5	10	5	1	0



Fish length as discriminating feature

- Find the best length *L* threshold

fish length < L → ***classify as salmon***

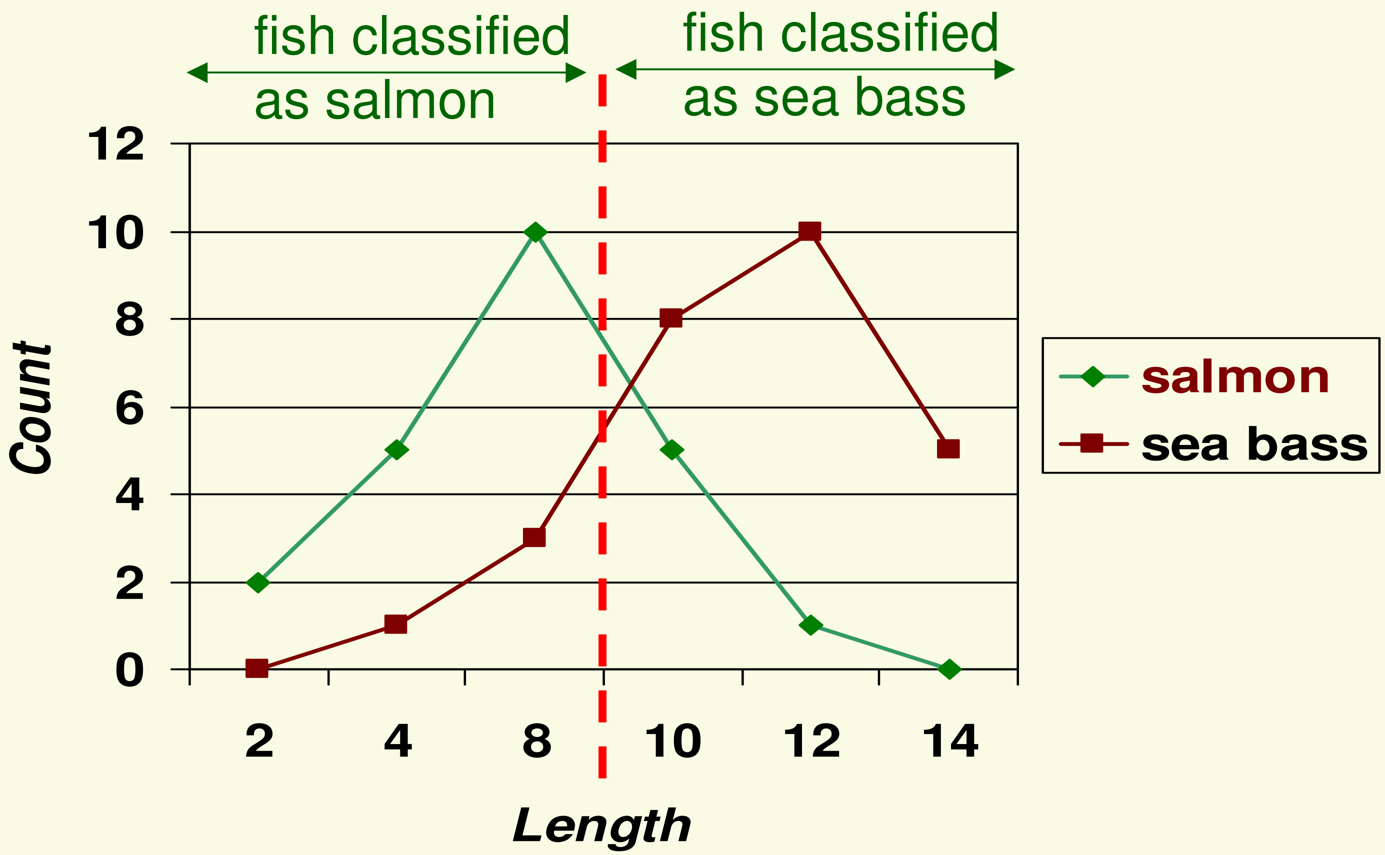
fish length > L → ***classify as sea bass***

- For example, at *L* = 5, misclassified:
  - 1 sea bass
  - 16 salmon

	2	4	8	10	12	14
bass	0	1	3	8	10	5
salmon	2	5	10	5	1	0

- Classification error (total error): $\frac{17}{50}$ **= *34%***

Fish Length as discriminating feature



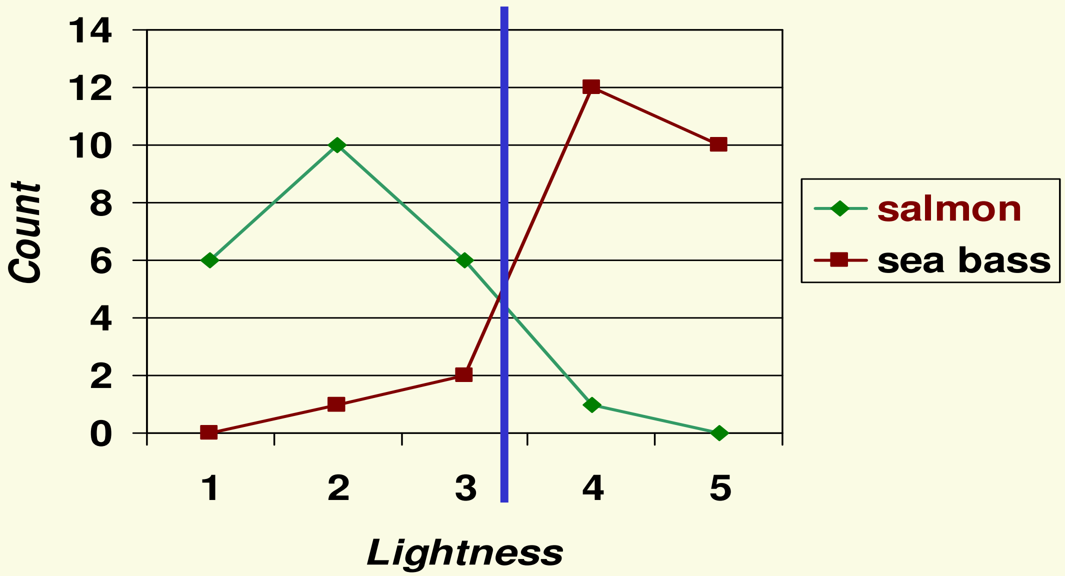
- After searching through all possible thresholds ***L***, the best ***L***= 9, and still 20% of fish is misclassified

Next Step

- Lesson learned:
  - Length is a poor feature alone!
- What to do?
  - Try another feature
  - Salmon tends to be lighter
  - Try average fish lightness

Fish lightness as discriminating feature

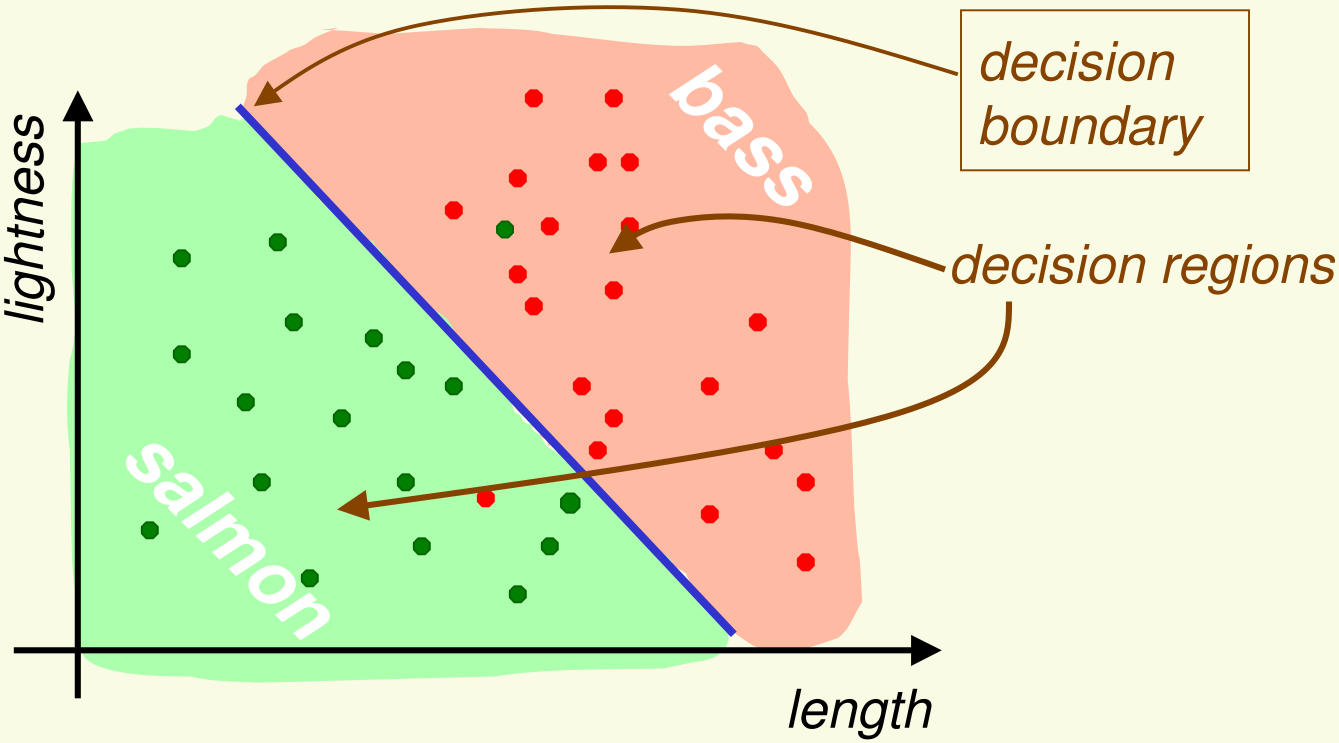
	1	2	3	4	5
bass	0	1	2	10	12
salmon	6	10	6	1	0



- Now fish are well separated at lightness threshold of 3.5 with classification error of 8%

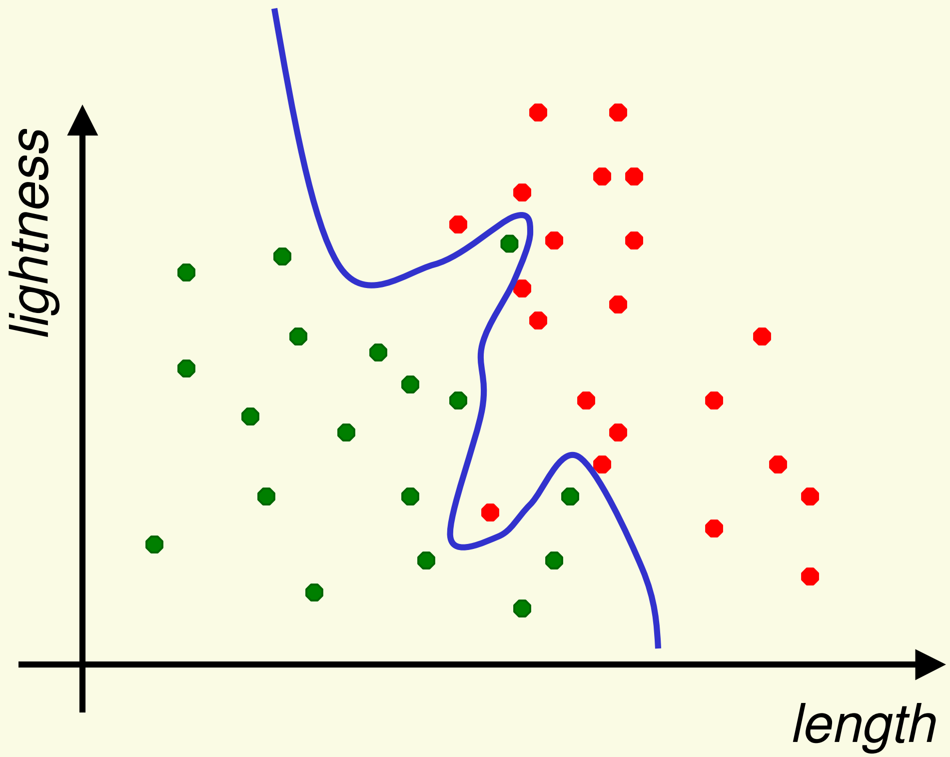
Can do even better by feature combining

- Use both *length* and *lightness* features
- Feature vector [*length*,*lightness*]



- Classification error 4%

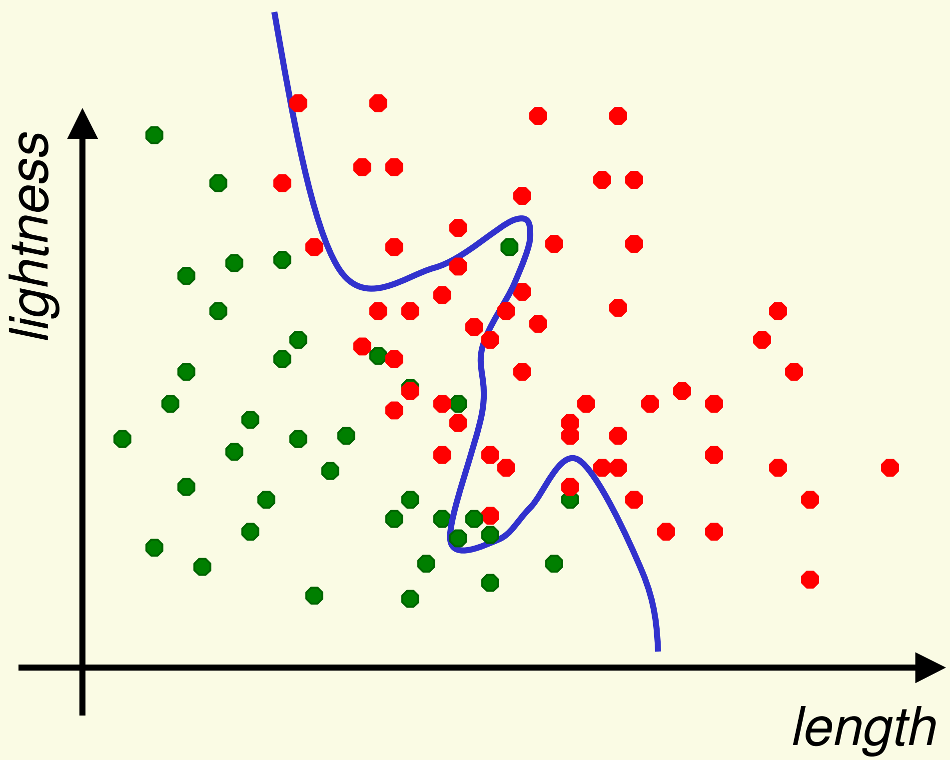
Better decision boundary



- Ideal decision boundary, 0% classification error

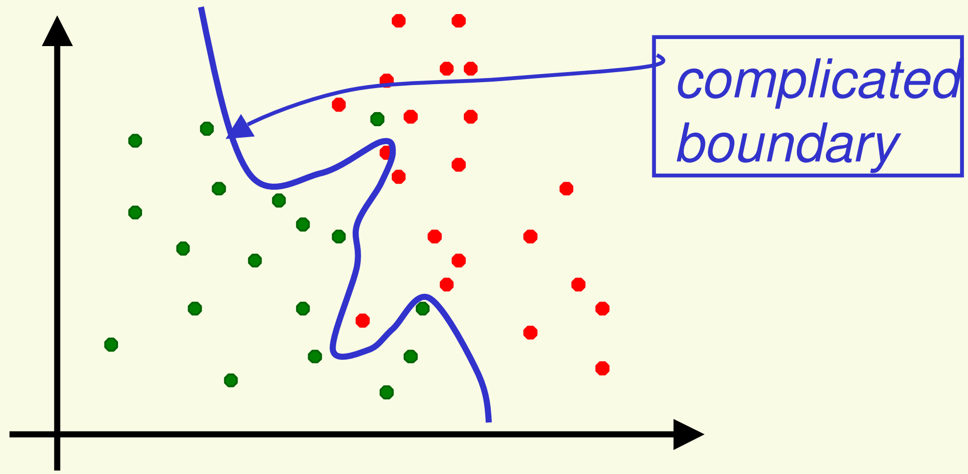
Test Classifier on New Data

- Classifier should perform well on new data
- Test “ideal” classifier on new data: 25% error



What Went Wrong?

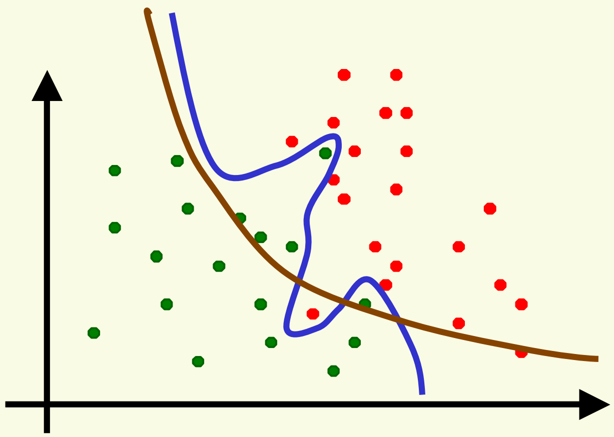
- Poor ***generalization***



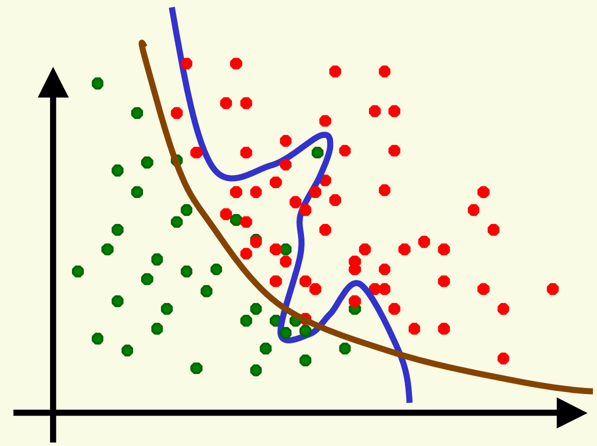
- Complicated boundaries do not generalize well to the new data, they are too "tuned" to the particular training data, rather than some true model which will separate salmon from sea bass well.
  - This is called overfitting the data

Generalization

training data



testing data



- Simpler decision boundary does not perform ideally on the training data but generalizes better on new data
- Favor simpler classifiers
  - William of Occam (1284-1347): “entities are not to be multiplied without necessity”

Pattern Recognition System Structure

domain dependent

camera, microphones, medical imaging devices, etc.

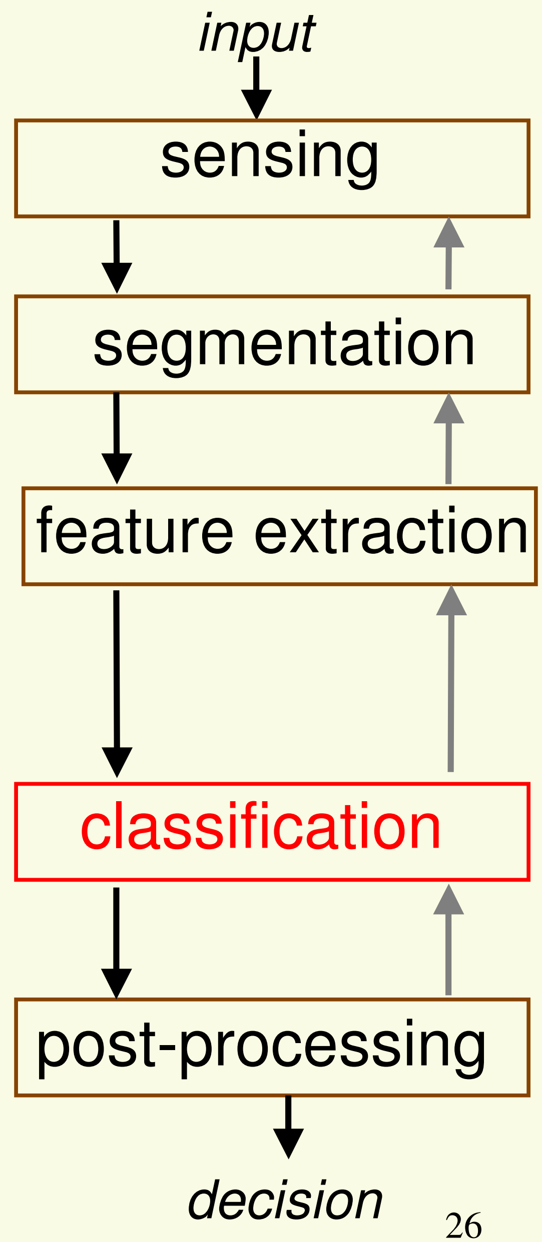
Patterns should be well separated and should not overlap.

Extract discriminating features. Good features make the work of classifier easy.

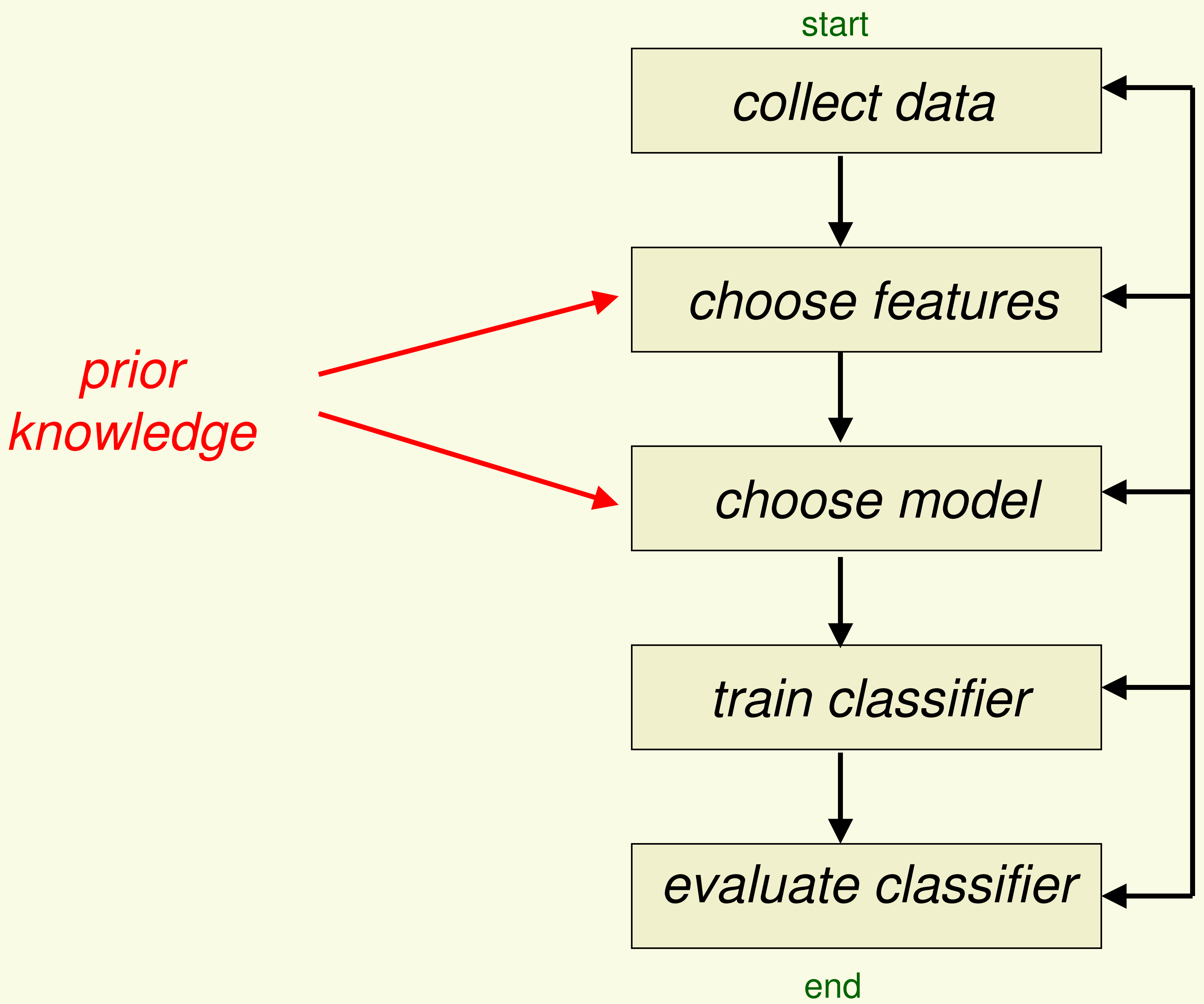
Use features to assign the object to a category. Better classifier makes feature extraction easier.
Our main topic in this course

Exploit context (input depending information) to improve system performance

Tne cat → ***The cat***

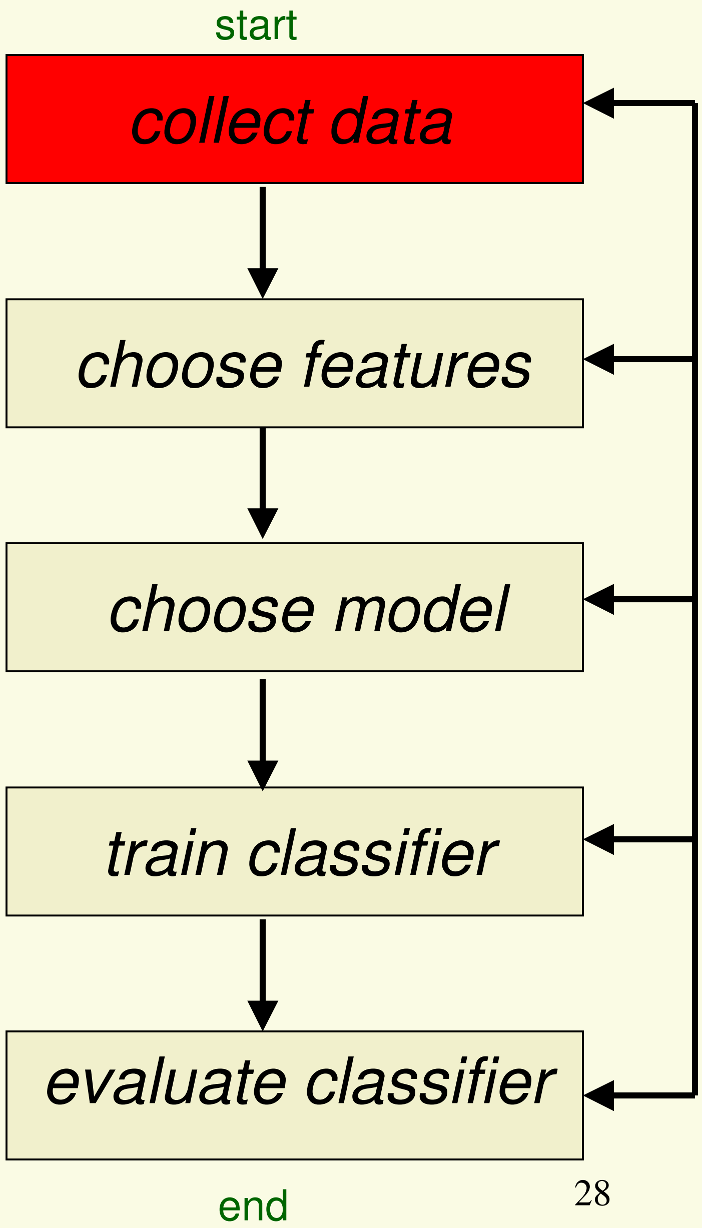


How to design a PR system?



Design Cycle cont.

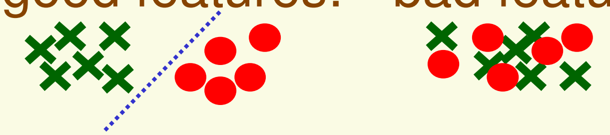
- Collect Data
  - Can be quite costly
  - How do we know when we have collected an adequately representative set of testing and training examples?



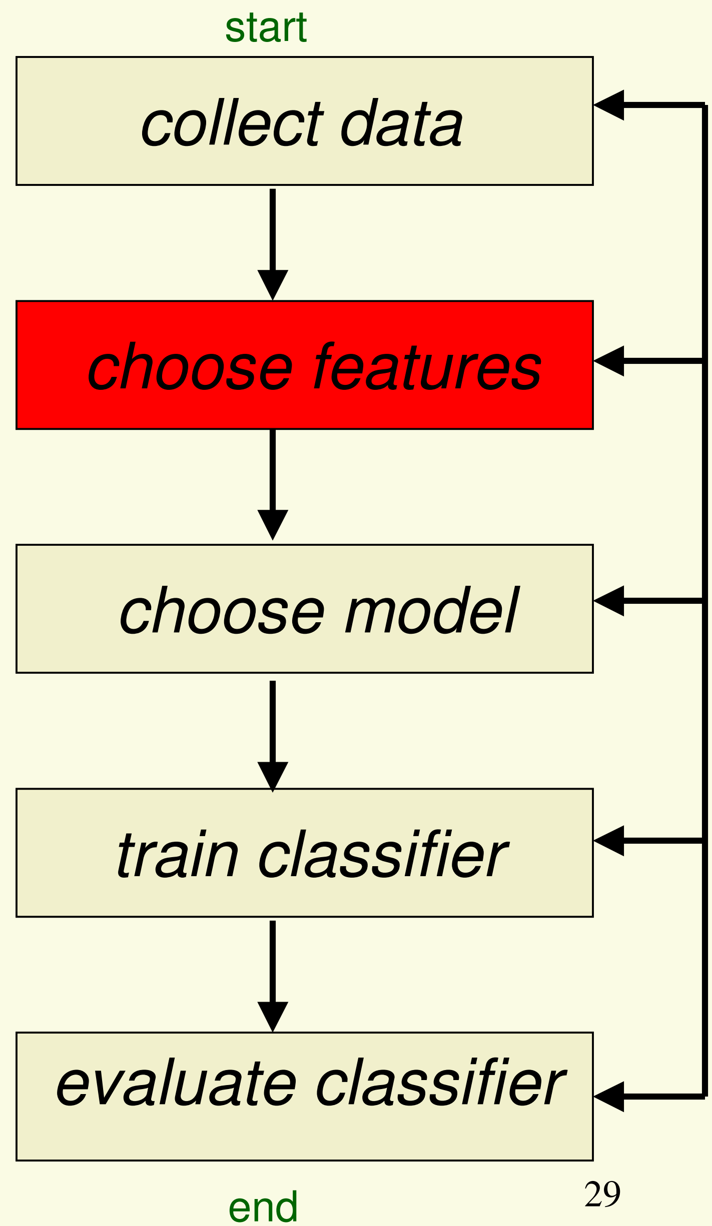
Design Cycle cont.

- Choose features
  - Should be discriminating, i.e. similar for objects in the same category, different for objects in different categories:

good features: bad features:

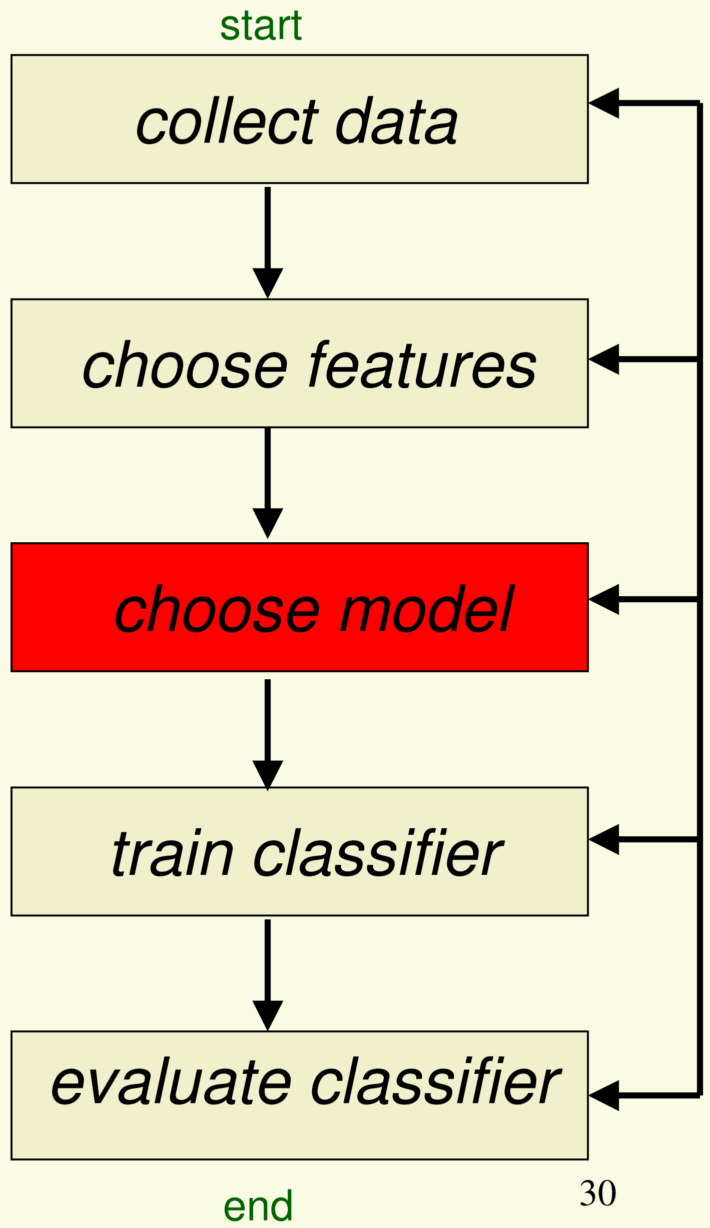


  - Prior knowledge plays a great role (domain dependent)
  - Easy to extract
  - Insensitive to noise and irrelevant transformations



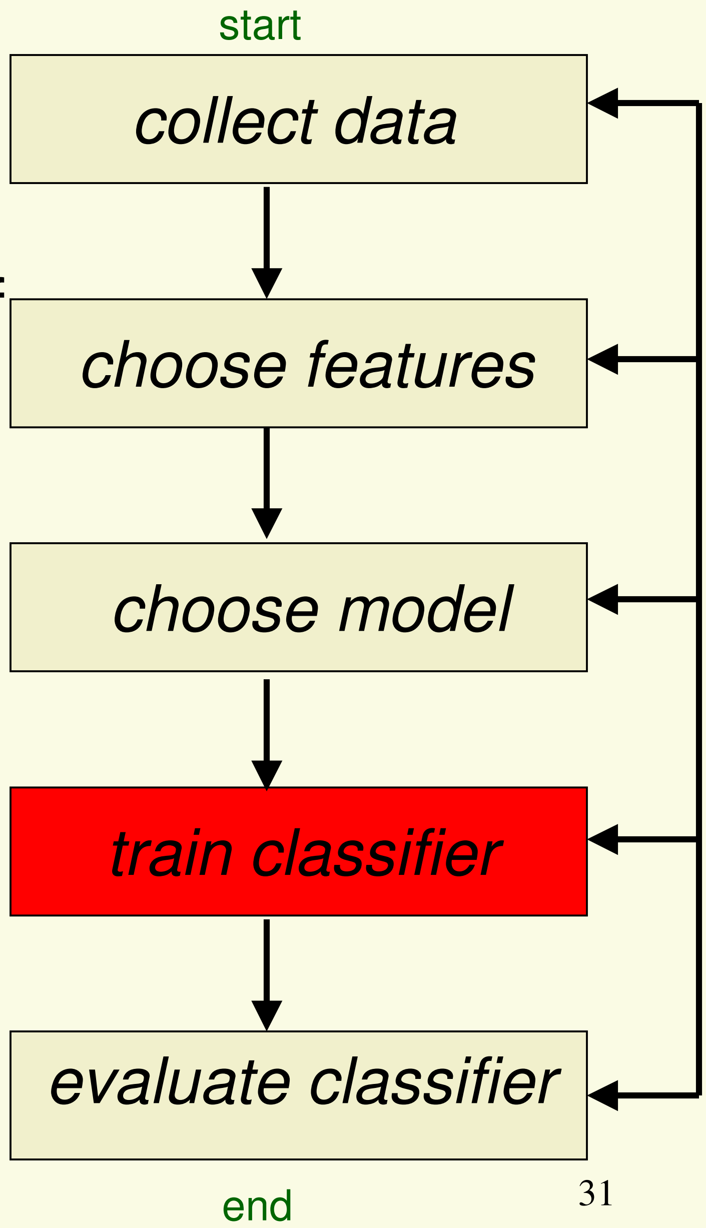
Design Cycle cont.

- Choose model
  - What type of classifier to use?
  - When should we try to reject one model and try another one?
  - What is the best classifier for the problem?



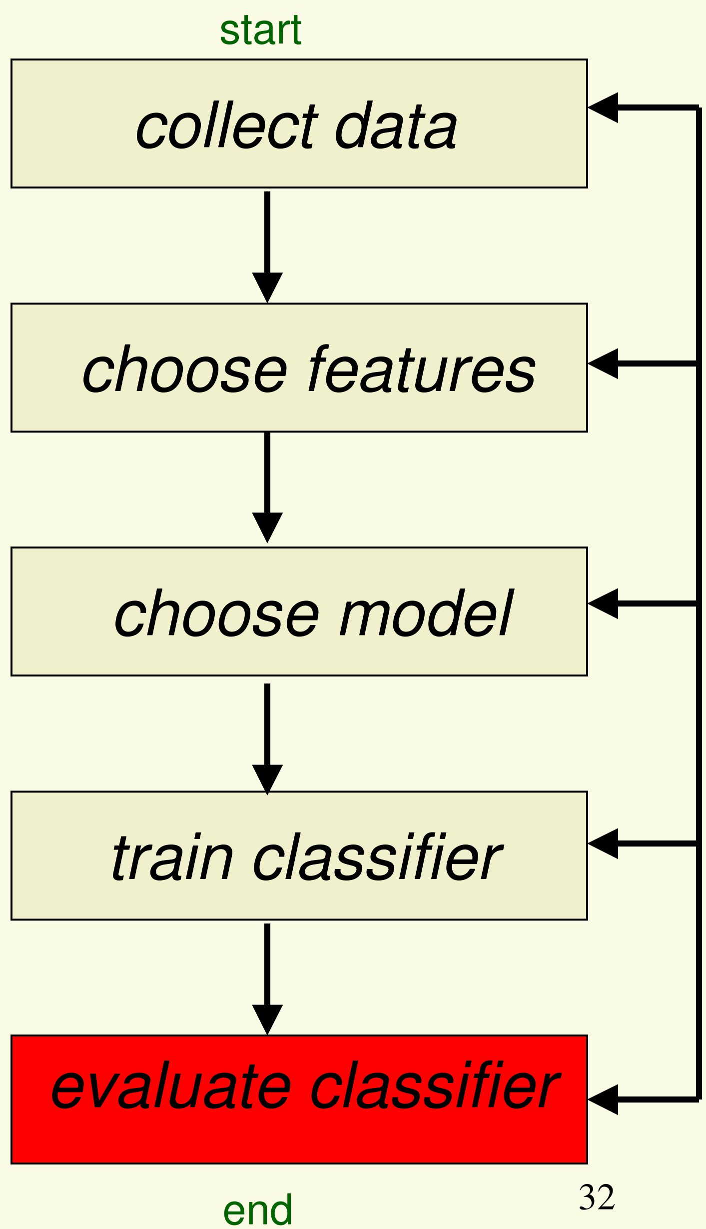
Design Cycle cont.

- Train classifier
  - Process of using data to determine the parameters of classifier
  - Change parameters of the chosen model so that the model fits the collected data
  - Many different procedures for training classifiers
  - Main scope of the course



Design Cycle cont.

- Evaluate Classifier
  - measure system performance
  - Identify the need for improvements in system components
  - How to adjust complexity of the model to avoid over-fitting? Any principled methods to do this?
  - Trade-off between computational complexity and performance



Conclusion

- **useful**
  - a lot of exciting and important applications
- **but hard**
  - must solve many issues for a successful pattern recognition system

Review: mostly probability and some statistics

Content

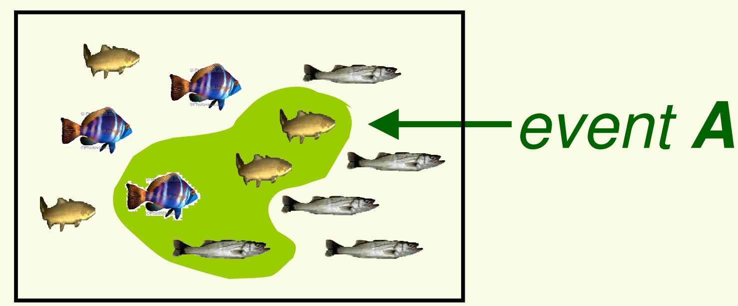
- Probability
  - Axioms and properties
  - Conditional probability and independence
  - Law of Total probability and Bayes theorem
- Random Variables
  - Discrete
  - Continuous
- Pairs of Random Variables
- Random Vectors
- Gaussian Random Variable

Basics

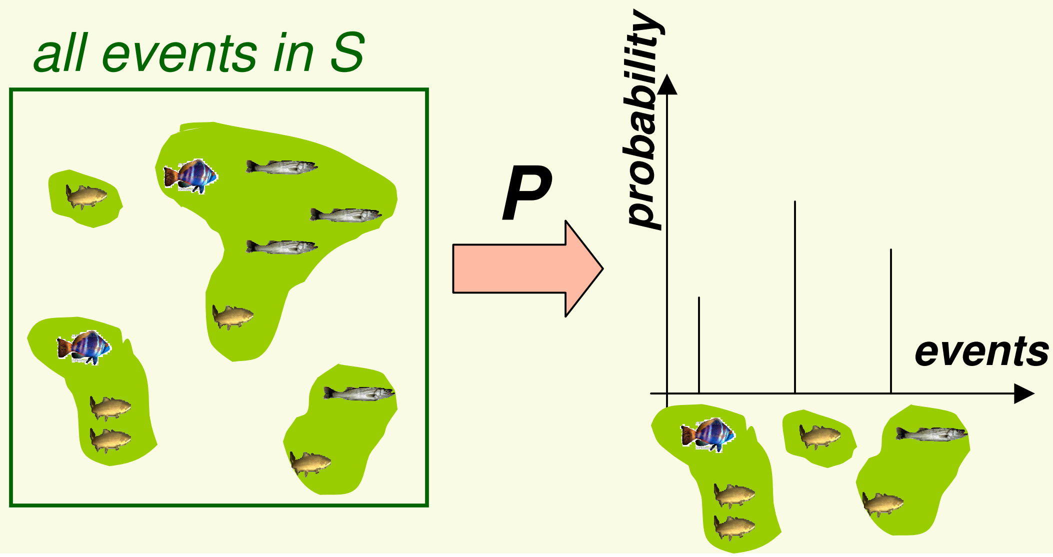
- We are performing a random experiment (catching one fish from the sea)
- Sample space *S*: the set of all possible outcomes
- An event *A:* a set of of possible outcomes of experiment, i.e. a subset of *S*
- *Probability law:a* rule *that* assigns probabilities to events in an experiment

$$A \longrightarrow \boldsymbol{P}(A)$$

S: *all fish in the sea*



total number of events: 2^{12}



Axioms of Probability

1. $P(A) \geq 0$
2. $P(S) = 1$
3. If $A \cap B = \varnothing$ then $P(A \cup B) = P(A) + P(B)$

Properties of Probability

$$P(\varnothing) = 0$$

$$P(A) \leq 1$$

$$P(A^c) = 1 - P(A)$$

$$A \subset B \;\Rightarrow\; P(A) < P(B)$$

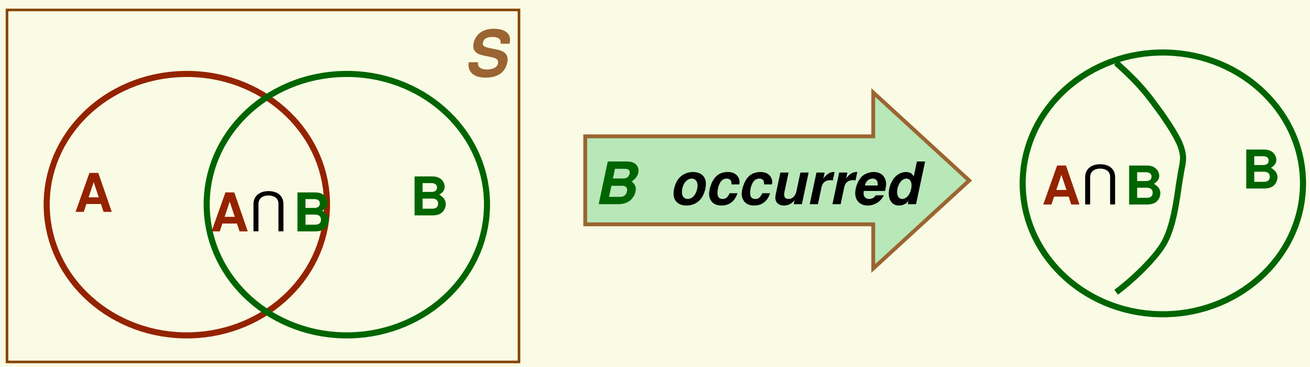
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$\{A_i \cap A_j = \varnothing, \forall i, j\} \Rightarrow P\left(\bigcup_{k=1}^{N} A_k\right) = \sum_{k=1}^{N} P(A_k)$$

Conditional Probability

- If A and B are two events, and we know that event B has occurred, then (if P(B)>0)

$$\boxed{P(A/B) = \frac{P(A \cap B)}{P(B)}}$$



the "new" sample space is **B,** the "new" **A** is old **A∩B**

- multiplication rule $\boxed{P(A \cap B) = P(A/B)\,P(B)}$

Independence

- A and B are independent events if

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

- By the law of conditional probability, if A and B are independent

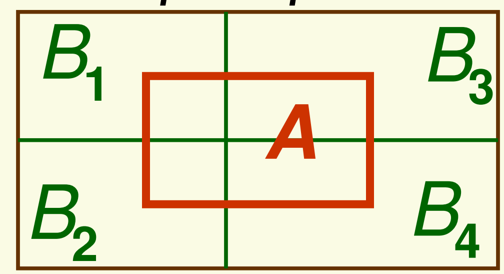
$$P(A|B) = \frac{P(A)\, P(B)}{P(B)} = P(A)$$

- If two events are not independent, then they are said to be dependent

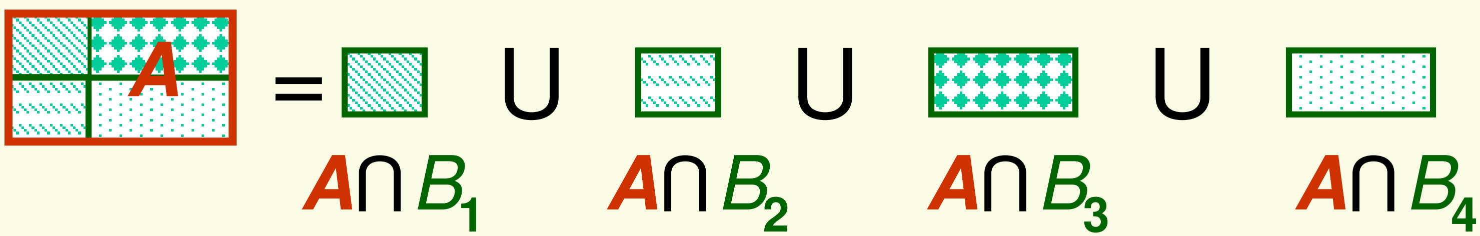
Law of Total Probability

- $B_1, B_2, \ldots, B_n$ partition S

sample space S



- Consider an event A



- Thus $P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) + P(A \cap B_4)$
- Or using multiplication rule:

$$P(A) = P(A/B_1)P(B_1) + \ldots + P(A/B_4)P(B_4)$$

$$\boxed{P(A) = \sum_{k=1}^{n} P(A|B_k)P(B_k)}$$

Bayes Theorem

- Let B_1, B_2, …, B_n, be a partition of the sample space S. Suppose event A occurs. What is the probability of event B_i ?
- **Answer:** **Bayes Rule**

from conditional probability

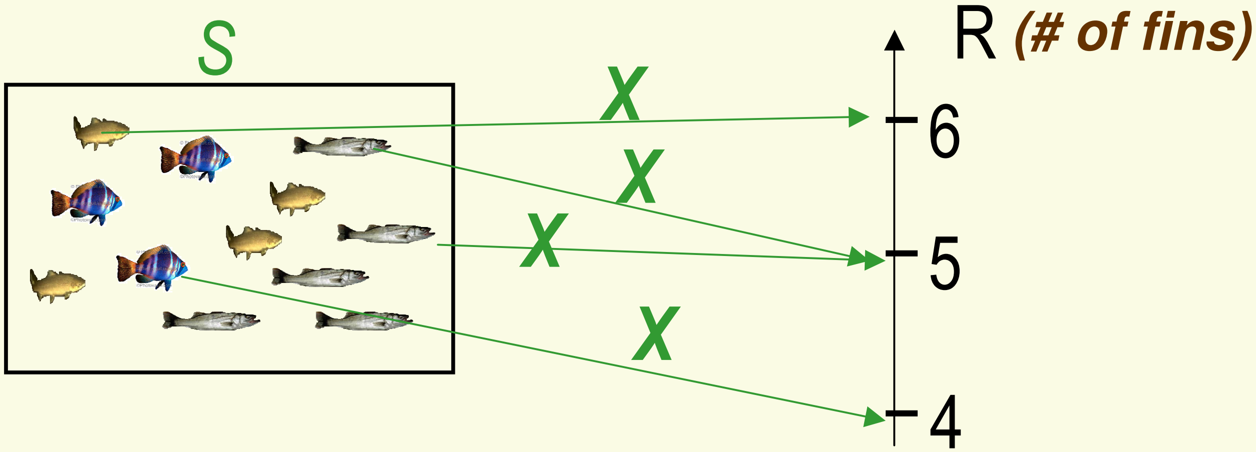
$$P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A \mid B_i)P(B_i)}{\sum_{k=1}^{n} P(A \mid B_k)P(B_k)}$$

from the law of total probability

- One of the most useful tools we are going to use

Random Variables

- In random experiment, usually assign some number to the outcome, for example, number of of fish fins
- A random variable ***X*** is a function from sample sample space ***S*** to a real number. ***X***: ***S*** $\rightarrow$ R



- ***X*** is random due to randomness of its argument
- $$P(X=a)=P(X(\omega)=a)=P(\omega\in\Omega\mid X(\omega)=a)$$

Two Types of Random Variables

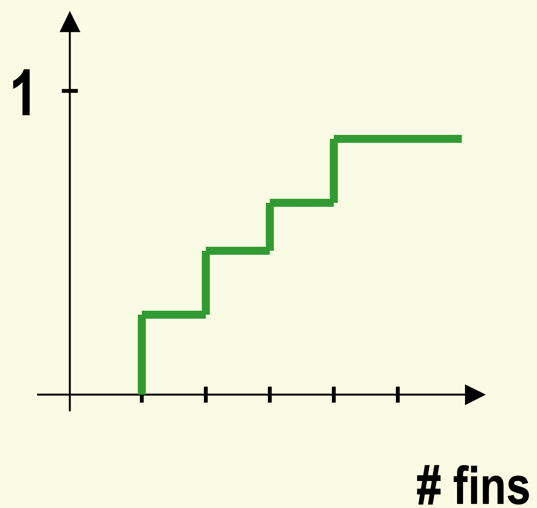
- ***Discrete*** random variable has countable number of values
  - number of fish fins (0,1,2,….,30)
- ***Continuous*** random variable has continuous number of values
  - fish weight (any real number between 0 and 100)

Cumulative Distribution Function

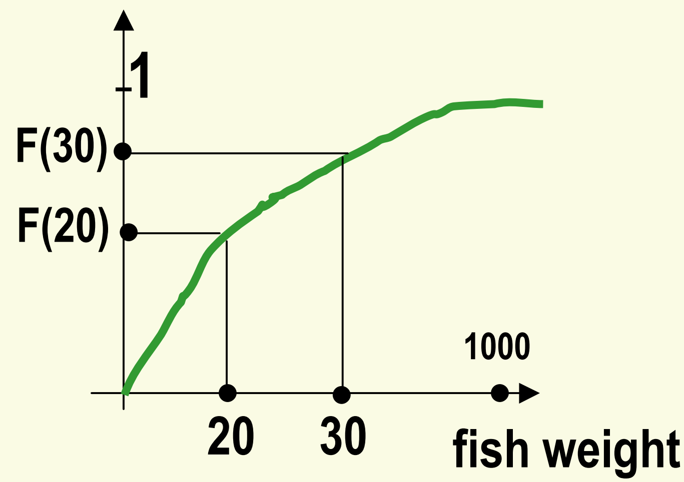
- Given a random variable ***X***, CDF is defined as

$$\boxed{F(a) = P(X \le a)}$$

CDF for discrete rv

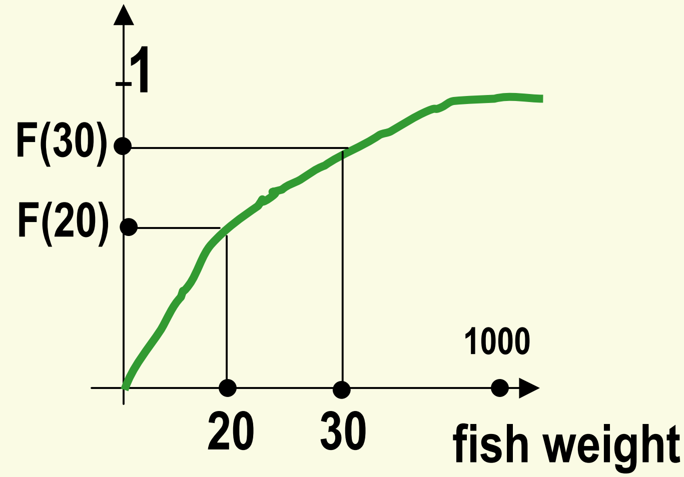


CDF for continuous rv



Properties of CDF $F(a) = P(X \leq a)$

CDF for continuous rv



1. $F(a)$ is non decreasing
2. $\lim_{b \to \infty} F(b) = 1$
3. $\lim_{b \to -\infty} F(b) = 0$

- Questions about ***X*** can be asked in terms of CDF

$$\boxed{P(a < X \leq b) = F(b) - F(a)}$$

Example:

P(fish weights between 20 and 30)=F(30)-F(20)

Discrete RV: Probability Mass Function

- Given a discrete random variable ***X***, we define the probability mass function as

$$\boxed{p(a) = P(X = a)}$$

- Satisfies all axioms of probability

- CDF in discrete case satisfies

$$F(a) = P(X \leq a) = \sum_{x \leq a} P(X = a) = \sum_{x \leq a} p(a)$$

Continuous RV: Probability Density Function

- Given a continuous RV ***X***, we say f(x) is its probability density function if
  - $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)\,dx$
  - and, more generally $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$

Properties of Probability Density Function

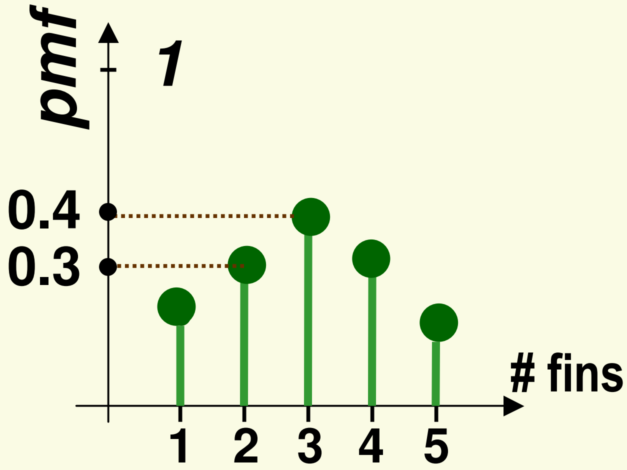
$$\frac{d}{dx}F(x) = f(x)$$

$$P(X = a) = \int_a^a f(x)\,dx = 0$$

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

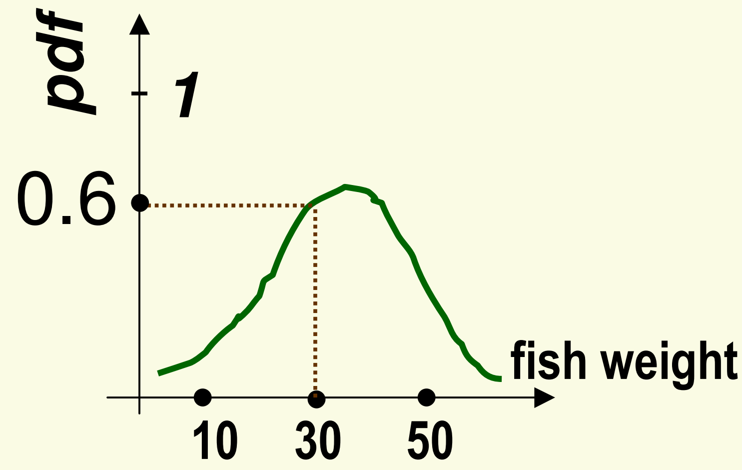
$$f(x) \ge 0$$

probability mass



- true probability
- P(fish has 2 or 3 fins)= =p(2)+p(3)=0.3+0.4
- take sums

probability density



- density, not probability
- P(fish weights 30kg) $\neq$ 0.6
- P(fish weights 30kg)=0
- P(fish weights between 29 and 31kg)= $\int_{29}^{31} f(x)dx$
- integrate

Expected Value

- Useful characterization of a r.v.
- Also known as mean, expectation, or first moment

discrete case: $\mu = E(X) = \sum_{\forall x} x\, p(x)$

continuous case: $\mu = E(X) = \int_{-\infty}^{\infty} x\, f(x)dx$

- Expectation can be thought of as the average or the center, or the expected average outcome over many experiments

Expected Value for Functions of X

- Let g(x) be a function of the r.v. X. Then

discrete case: $E[g(X)] = \sum_{\forall x} g(x)\, p(x)$

continuous case: $E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x) dx$

- An important function of X: $[X-E(X)]^2$
  - Variance $E[[X-E(X)]^2] = var(X) = \sigma^2$
  - Variance measures the spread around the mean
  - Standard deviation = $[Var(X)]^{1/2}$, has the same units as the r.v. X

Properties of Expectation

- If X is constant r.v. X=c, then E(X) = c

- If a and b are constants, E(aX+b)=aE(X)+b

- More generally,

$$E\left(\sum_{i=1}^{n}(a_i X_i + c_i)\right) = \sum_{i=1}^{n}(a_i E(X_i) + c_i)$$

- If a and b are constants, then
var(aX+b)= a^2var(X)

Pairs of Random Variables

- Say we have 2 random variables:
  - Fish weight ***X***
  - Fish lightness ***Y***
- Can define *joint* CDF

$$\boldsymbol{F(a,b) = P(X \le a, Y \le b) = P(\omega \in \Omega \,/\, X(\omega) \le a, Y(\omega) \le b)}$$

- Similar to single variable case, can define
  - discrete: joint probability mass function

$$p(a,b) = P(X = a, Y = b)$$

  - continuous: joint density function $f(x,y)$

$$P(a \le X \le b, c \le Y \le d) = \iint_{\substack{a \le x \le b \\ c \le y \le d}} f(x,y)\,dx\,dy$$

Marginal Distributions

- given joint mass function $p_{x,y}(a,b)$, marginal, i.e. probability mass function for r.v. X can be obtained from $p_{x,y}(a,b)$

$$\boxed{p_x(a) = \sum_{\forall y} p_{x,y}(a,y)} \qquad \boxed{p_y(b) = \sum_{\forall x} p_{x,y}(x,b)}$$

- marginal densities $f_x(x)$ and $f_y(y)$ are obtained from joint density $f_{x,y}$ (x,y) by integrating

$$\boxed{f_x(x) = \int_{y=-\infty}^{y=\infty} f_{x,y}(x,y)\,dy} \qquad \boxed{f_y(y) = \int_{x=-\infty}^{x=\infty} f_{x,y}(x,y)\,dx}$$

Independence of Random Variables

- r.v. X and Y are independent if

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

- *Theorem*: r.v. X and Y are independent if and only if

$$p_{x,y}(x,y) = p_y(y)p_x(x) \quad \textbf{\textit{(discrete)}}$$

$$f_{x,y}(x,y) = f_y(y)f_x(x) \quad \textbf{\textit{(continuous)}}$$

More on Independent RV's

- If X and Y are independent, then
  - E(XY)=E(X)E(Y)
  - Var(X+Y)=Var(X)+Var(Y)
  - G(X) and H(Y) are independent

Covariance

- Given r.v. X and Y, covariance is defined as:

$$\mathbf{cov}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{E}[(\boldsymbol{X}-\boldsymbol{E}(\boldsymbol{X}))(\boldsymbol{Y}-\boldsymbol{E}(\boldsymbol{Y}))] = \boldsymbol{E}(\boldsymbol{XY}) - \boldsymbol{E}(\boldsymbol{X})\boldsymbol{E}(\boldsymbol{Y})$$

- Covariance is useful for checking if features *X* and *Y* give similar information
- Covariance (from co-vary) indicates tendency of X and Y to vary together
  - If X and Y tend to increase together, Cov(X,Y) > 0
  - If X tends to decrease when Y increases, Cov(X,Y) < 0
  - If decrease (increase) in X does not predict behavior of Y, Cov(X,Y) is close to 0

Covariance Correlation

- If cov(X,Y) = 0, then X and Y are said to be uncorrelated (think unrelated). However X and Y are not necessarily independent.
- If X and Y are independent, cov(X,Y) = 0
- Can normalize covariance to get correlation

$$-1 \le cor(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} \le 1$$

Random Vectors

- Generalize from pairs of r.v. to vector of r.v. X= [X_1 X_2… X_3] (think multiple features)
- Joint CDF, PDF, PMF are defined similarly to the case of pair of r.v.'s

Example:

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n)$$

- All the properties of expectation, variance, covariance transfer with suitable modifications

Covariance Matrix

- characteristics summary of random vector
- $\text{cov}(X)=\text{cov}[X_1\ X_2 \ldots X_n] = \Sigma = E[(X-\mu)(X-\mu)^T] =$

$$\begin{bmatrix} E(X_1-\mu_1)(X_1-\mu_1) & \cdots & E(X_n-\mu_n)(X_1-\mu_1) \\ E(X_2-\mu_2)(X_1-\mu_1) & \cdots & E(X_n-\mu_n)(X_2-\mu_2) \\ \vdots & & \vdots \\ E(X_n-\mu_n)(X_1-\mu_1) & \cdots & E(X_n-\mu_n)(X_n-\mu_n) \end{bmatrix}$$

variances

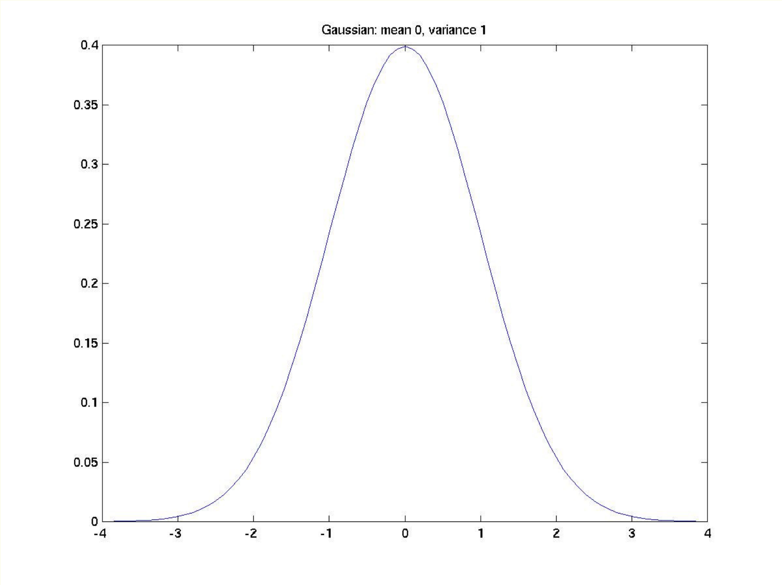
$$\begin{bmatrix} \sigma_1^2 & c_{12} & c_{13} \\ c_{21} & \sigma_2^2 & c_{23} \\ c_{31} & c_{32} & \sigma_3^2 \end{bmatrix}$$

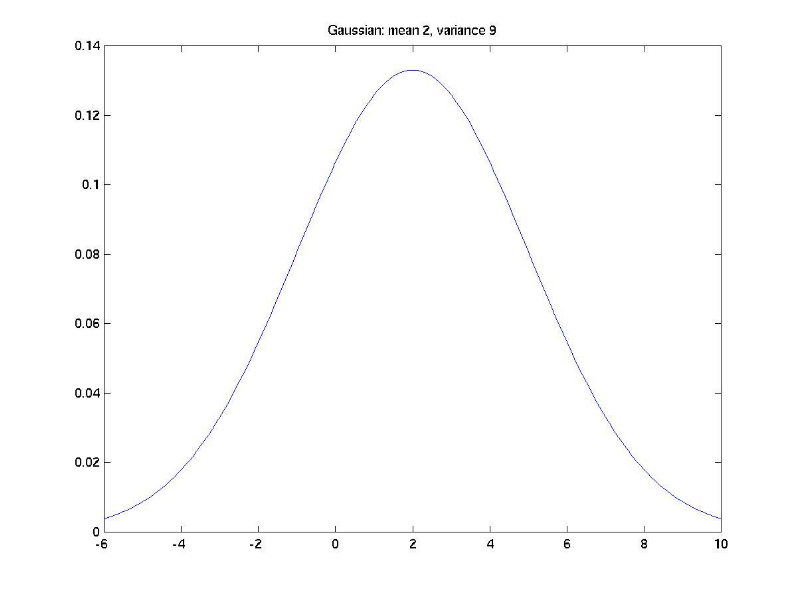
covariances

Normal or Gaussian Random Variable

- Has density $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

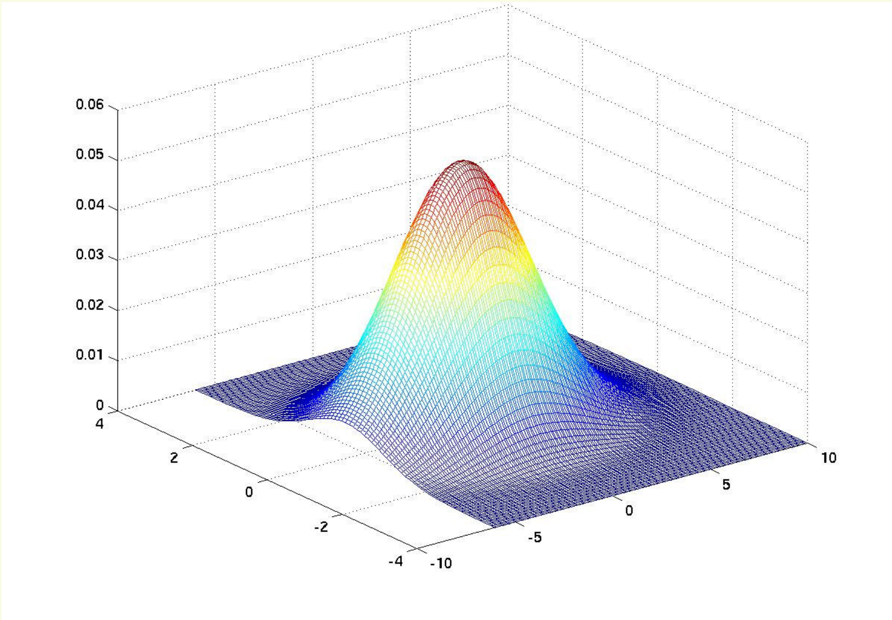
- Mean μ, and variance σ^2





Multivariate Gaussian

- has density $$f(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}\left[(x-\mu)\Sigma^{-1}(x-\mu)\right]}$$
- mean vector $\mu = \lfloor \mu_1, \ldots, \mu_n \rfloor$
- covariance matrix Σ



Why Gaussian?

- Frequently observed (Central limit theorem)
- Parameters μ and Σ are sufficient to characterize the distribution
- Nice to work with
  - Marginal and conditional distributions also are gaussians
  - If X_i's are uncorrelated then they are also independent

Summary

- Intro to Pattern Recognition
- Review of Probability and Statistics
- Next time will review linear algebra