Machine Learning

Deng Cai (蔡登)

College of Computer Science
Zhejiang University

dengcai@gmail.com

Short Bio

- Dr. Deng Cai (蔡登)
  - dengcai@gmail.com, dengcai@cad.zju.edu.cn

- Professor at CS college (the state key lab of CAD&CG).
  - 紫金港校区蒙民伟楼508

- Research interests:
  - Machine learning
  - Data mining
  - Computer vision
  - …

- http://dengcai.zjulearning.org:8081/

Course Information

- Web: http://dengcai.zjulearning.org:8081/Courses/ML/
- Homework: http://assignment.zjulearning.org:8081/
  - 缺省用户名和密码：学号，登陆之后修改密码
- Time:
  - **Tuesday, 14:05 – 15:35**
  - **Thursday, 14:05 – 15:35**
- Place: Classroom 205, west Caoguangbiao Building, Yuquan Campus
- QQ group: ML_ZJU (494525143) (Apply with name and student ID)
- TA: 张永辉、胡津铭、冯昊

Course information (Cont'd)

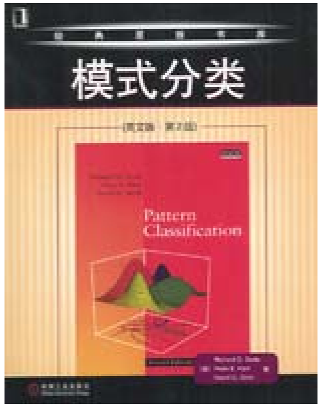
- Prerequisite:
  - Linear algebra, analysis, probability theory
  - Basic programming skills

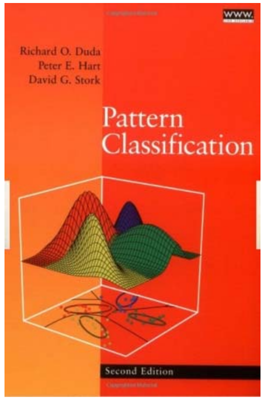
- Course textbook: No textbook is required. (Papers and other materials are available at the class web page)

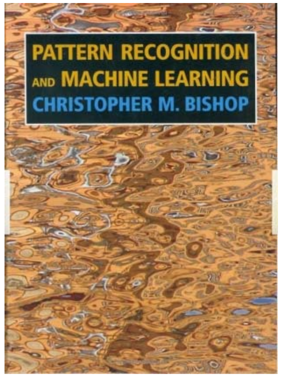
- Objective:
  - Basic understandings of some of the important machine learning methods.
  - Basic ability to use some machine learning techniques to solve real world problems.

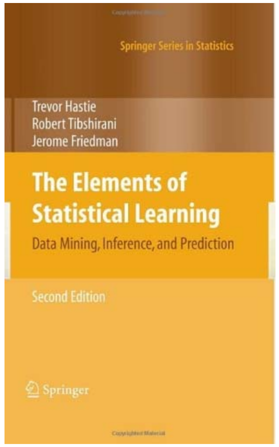
Reference Books

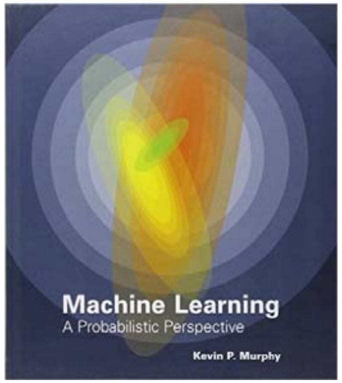
- R. Duda, P. Hart & D. Stork, ***Pattern Classification*** (2nd ed.), Wiley, 2000
- C. M. Bishop, **Pattern Recognition and Machine Learning**, Springer, 2006
- T. Hastie, R. Tibshirani & J. Friedman, **The Elements of Statistical Learning: Data Mining, Inference, and Prediction** (2nd ed.), Springer, 2009
- Kevin Murphy, **Machine Learning: A Probabilistic Perspective**, The MIT Press, 2012











Reference Books

- You can download all the books from the QQ group

Evaluation

- Quizzes (15%)
- Four assignments (10% each)
  - Everyone do it by himself
- Final exam (45%)

- Programming language:
  - Matlab
    - Tutorials
      - http://www.math.ufl.edu/help/matlab-tutorial/
      - http://www.math.mtu.edu/~msgocken/intro/node1.html
  - Python

Course Policies

- Class
  - No laptop, no cellphone.
- Cheating
  - No.
- Homework:
  - You have to write you own solution/program.
- Late Policy:
  - 0~24 hours: 90%
  - 24~48 hours: 50%
  - 48 hours ~: 25%
- Questions?

Why Take This Course?

- It is NOT
  - Easy course with high scores
  - Recommendation letter for US school application
    - Rank 1st

- You should
  - Work hard
  - Be honest

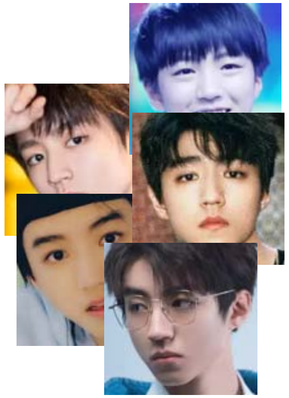
What is machine learning?

- Machine learning is the study of computer systems that improve their performance through experience.
  - Learn existing and known structures and rules.
  - Discover new findings and structures.
    - Face recognition
    - News summarization
- In machine learning, we study two types of problems

The first kind of problems

刘德华

章子怡

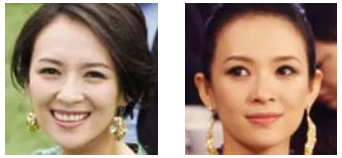


王俊凯

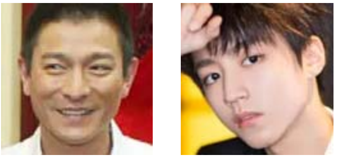
……

章子怡

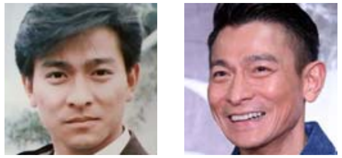
The first kind of problems



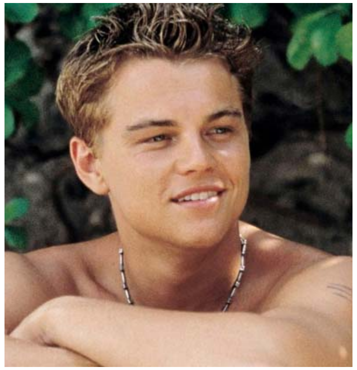
同一个人



不同人



同一个人



The first kind of problems

30岁

28岁

18岁

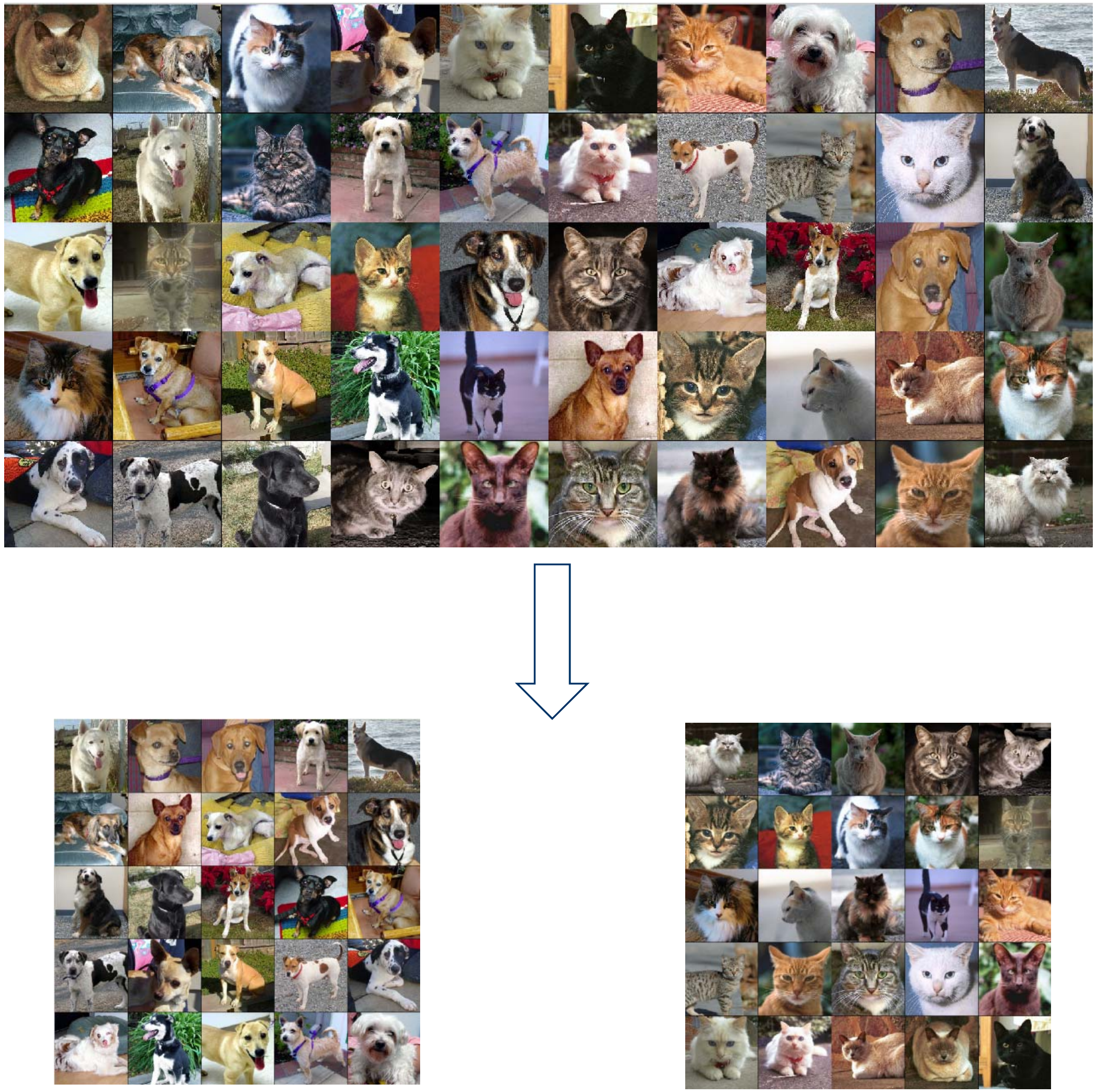
14岁

57岁

… …

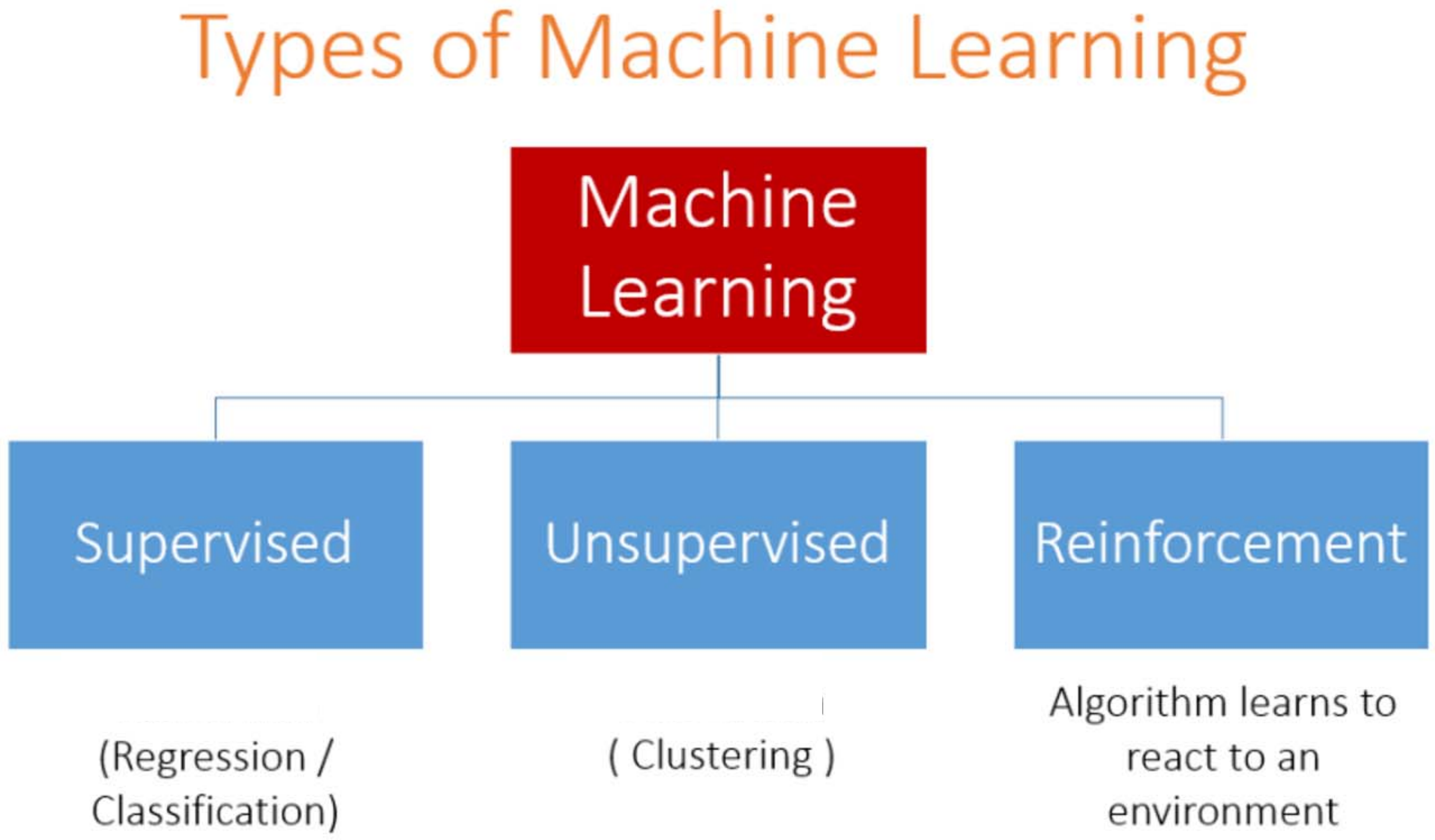
33岁

The second kind of problems



Two kinds of problems

- What are the differences?
- Supervised learning **vs.** Unsupervised learning



Types of Machine Learning

(Regression / Classification)

(Clustering)

Algorithm learns to react to an environment

Two kinds of problems

- What are the differences?
- Supervised learning **vs.** Unsupervised learning

- Supervised learning
  - Goal: learn a mapping from inputs $\boldsymbol{x}$ to outputs y
  - Training data: a labeled set of input-output pairs

  - Classification (Categorization, Decision making…)
    - y is a categorical variable
  - Regression
    - y is real-valued

Two kinds of problems

- What are the differences?
- Supervised learning **vs.** Unsupervised learning
- Unsupervised learning
  - We are only given inputs
  - Goal: find “interesting patterns”
  - Much less well-defined problem
  - Discovering clusters, Clustering
  - Discovering latent factors
    - Dimensionality reduction, Matrix factorization, Topic modeling

Two kinds of problems

- What are the differences?
- Supervised learning **vs.** Unsupervised learning
- Reinforcement learning
  - It is a supervised learning scenario
  - No desired category signal is given
  - The only teaching feedback is that the tentative category is right or wrong.
  - This is useful for learning how to act or behave when given occasional reward or punishment signals.

Focus of This Course

- What are the typical machine learning **problems**?
  - Supervised Learning
    - Classification (decision making)
    - Regression
  - Unsupervised Learning
    - Cluster analysis
    - Latent factor analysis
- What are the basic machine learning **tools (methods, algorithms)**?
- Matlab/Python programming

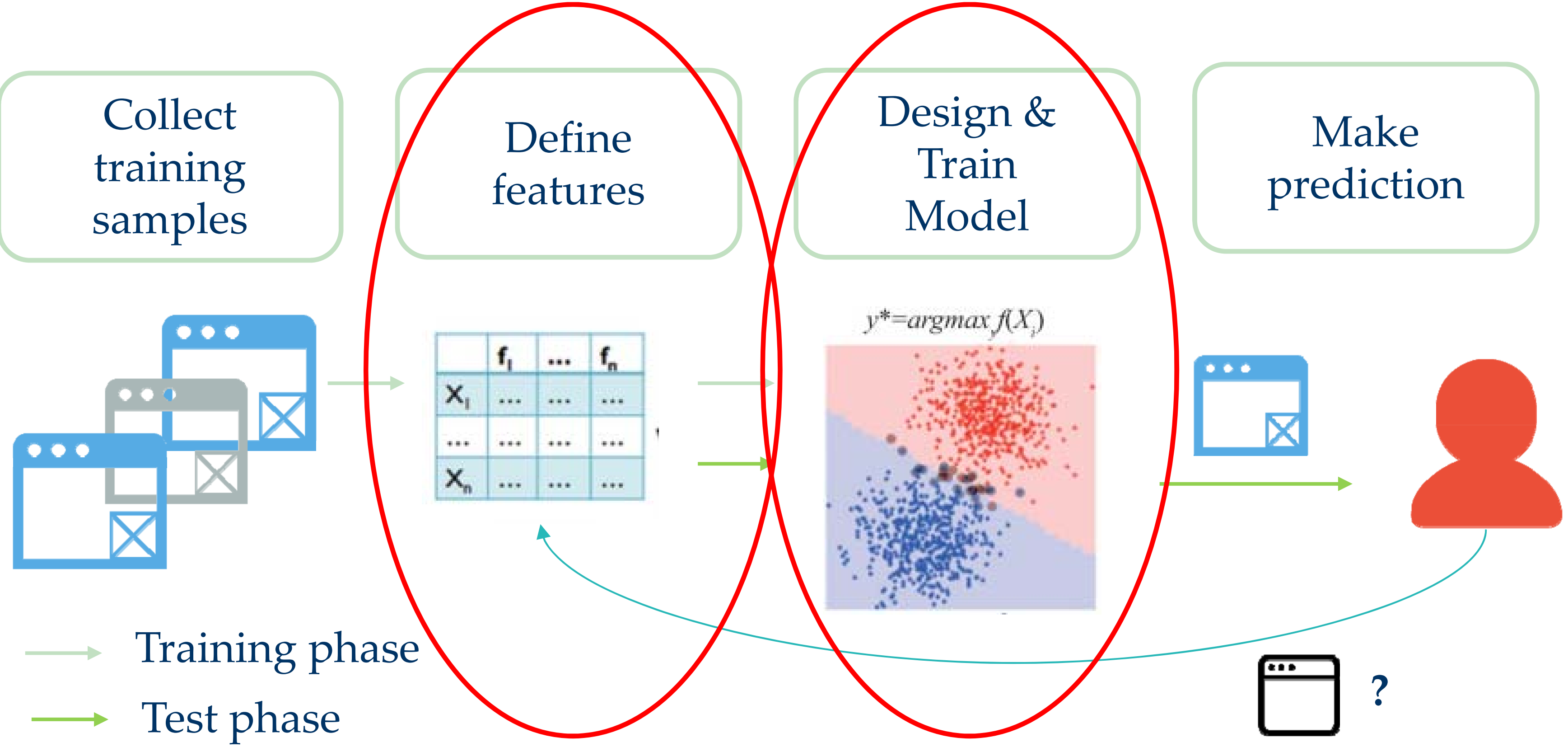
Basic Concepts of Supervised Learning

- Sample, example, pattern

- Features, predictors, independent variables
  - $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots \boldsymbol{x}_n$
- State of the nature, labels, pattern class, class, responses, dependent variables
  - $\omega_1, \omega_2, \cdots \omega_c$ or $y_1, y_2, \cdots y_c$ or $z_1, z_2, \cdots z_c$
- Training data
  - $(\boldsymbol{x}_1, \omega_1), (\boldsymbol{x}_2, \omega_2), \cdots (\boldsymbol{x}_n, \omega_n)$
- Model, statistical model, pattern class model, classifier
  - f
- Test data
- Training error & test error

Supervised Learning

Learning from **experience**(training data), and build **model** to **predict** the future



Collect training samples

Define features

Design & Train Model

Make prediction

Training phase

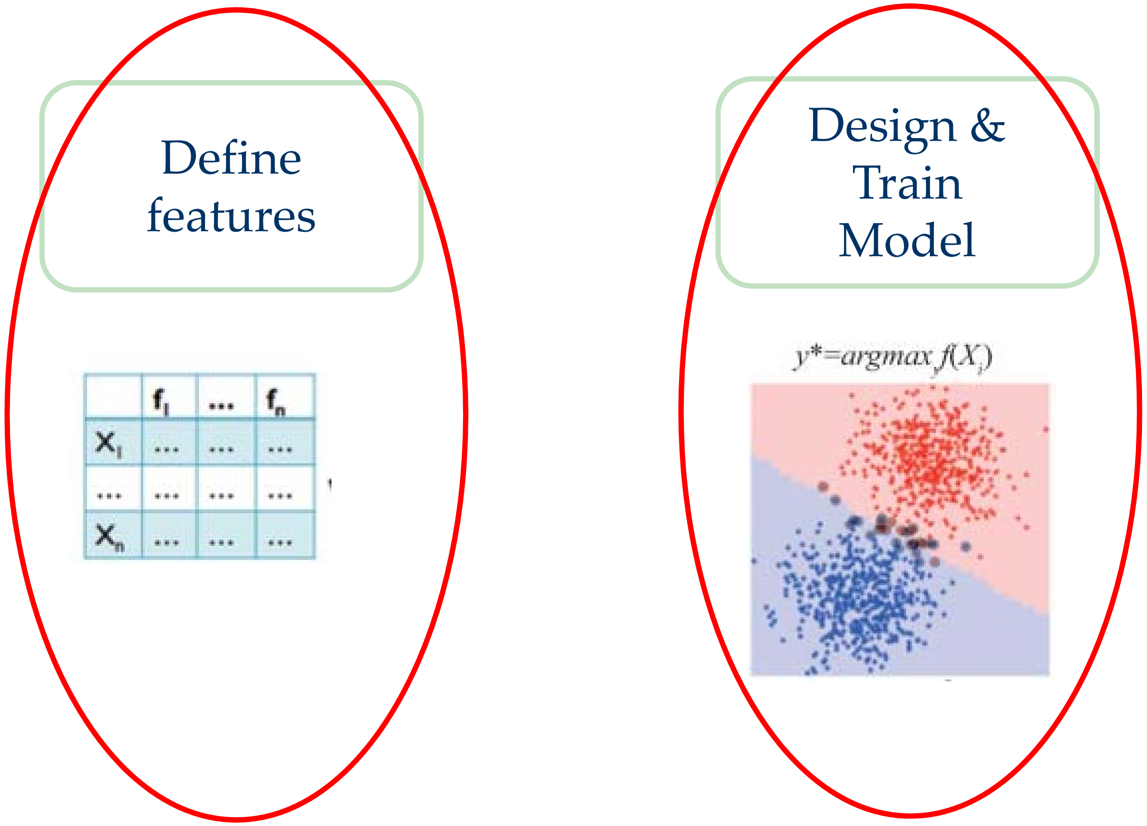
Test phase

Step 1

Step 2

Representation Learning

Supervised Learning



Define features

	f_1	…	f_n
X_1	…	…	…
…	…	…	…
X_n	…	…	…

Step 1

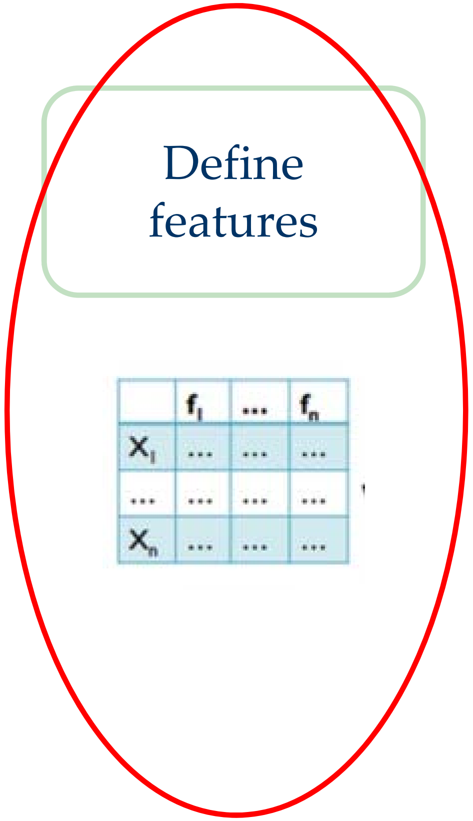
Design & Train Model

$y^{*}=\mathrm{argmax}\, f(X_i)$

Step 2

- Which step is more important in building a successful system?
- Which one is the focus of this course?

Why general classification hard?

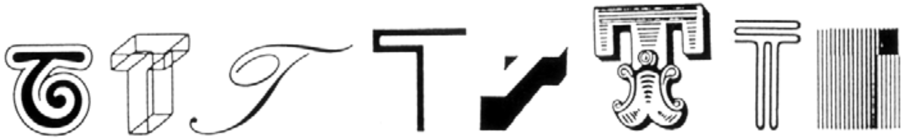


Define features

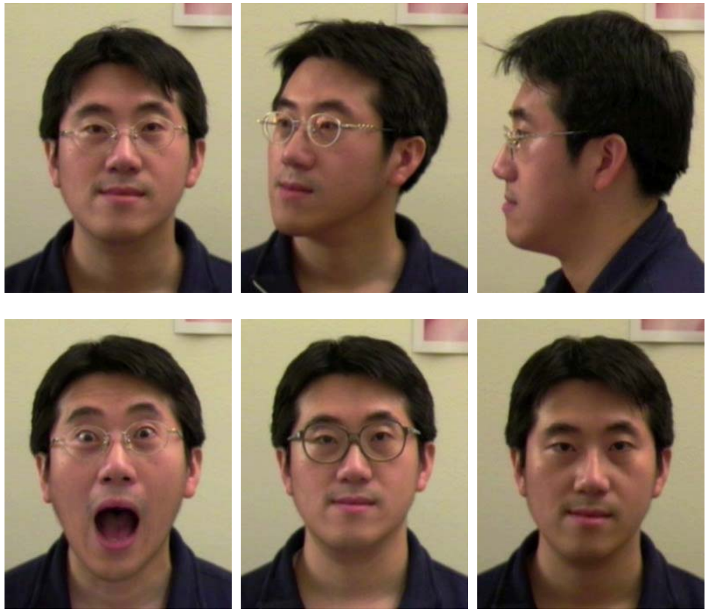
	f_1	...	f_n
x_1	...	...	...
...	...	...	...
x_n	...	...	...

Step 1 is not good enough

- Intra-class variability

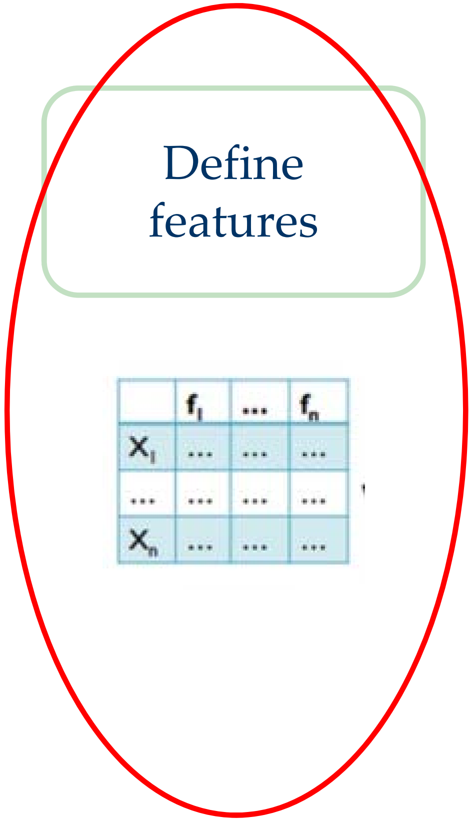


The letter "T" in different typefaces



Same face under different expression, pose, illumination

Why general classification hard?

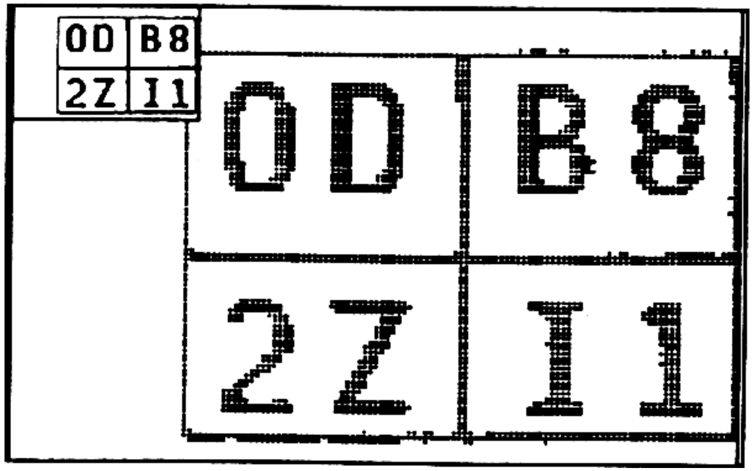


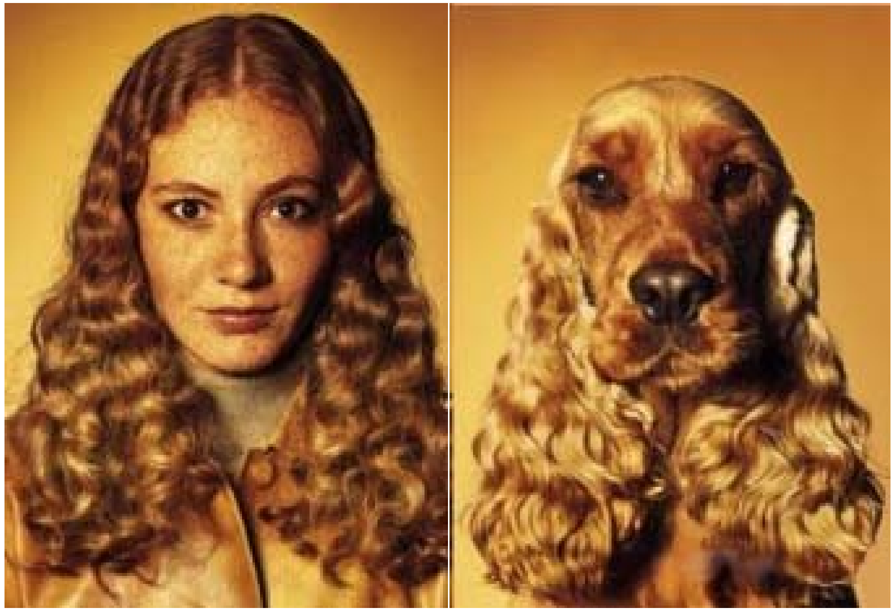
Define features

	f_1	...	f_n
x_1	...	...	...
...	...	...	...
x_n	...	...	...

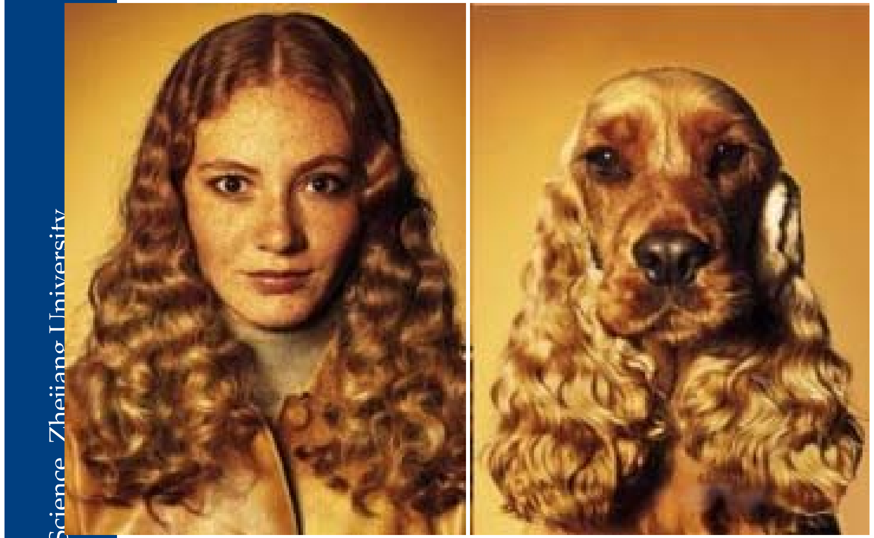
Step 1 is not good enough

- Inter-class similarity

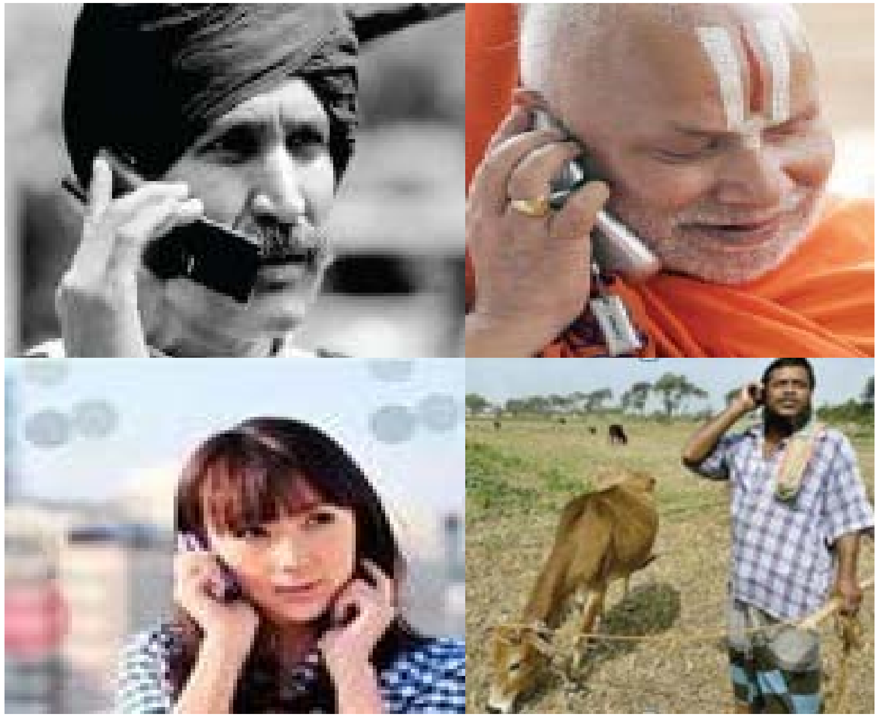




Semantic Gap



Looks similar
But semantically
different



Looks different
But semantically
the same

Representation: Features

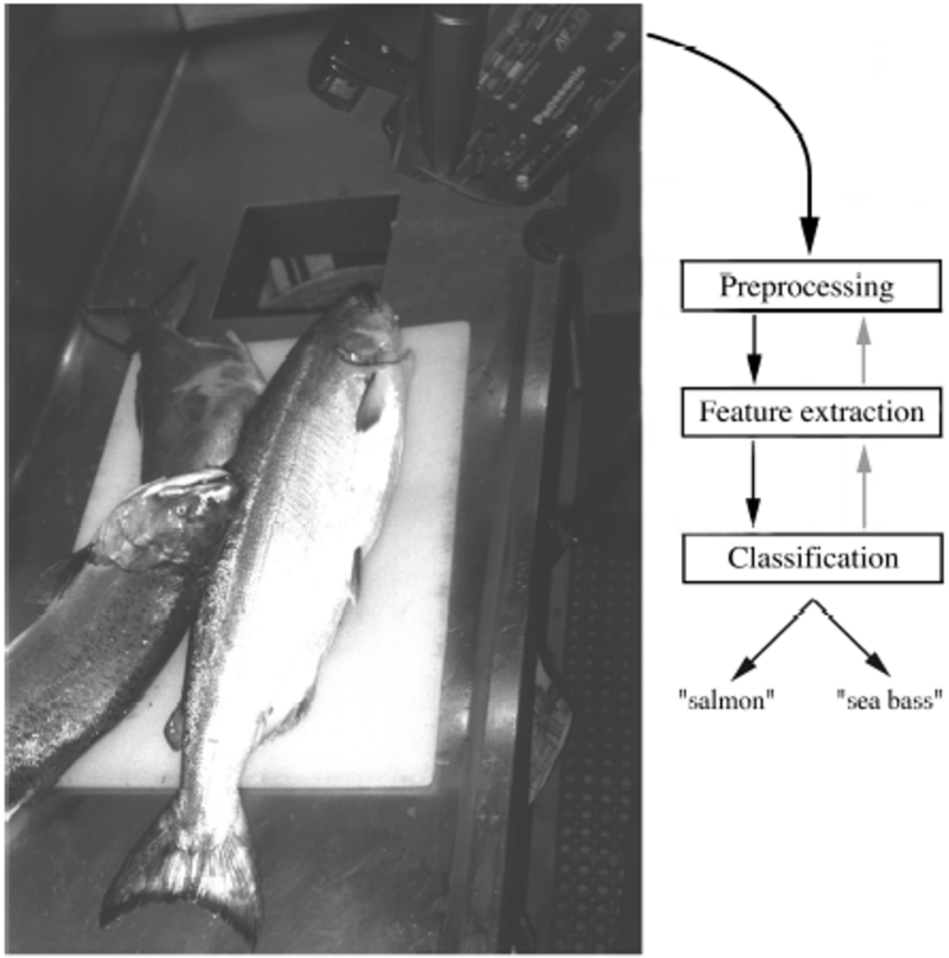
- Extract features to represent the samples
- Feature vector
- Good representation:
  - Low intra-class variability
  - Low inter-class similarity

Fish Classification: Salmon v. Sea Bass

Preprocessing involves image enhancement and segmentation;

(i) separate touching or occluding fishes and

(ii) extract fish contour

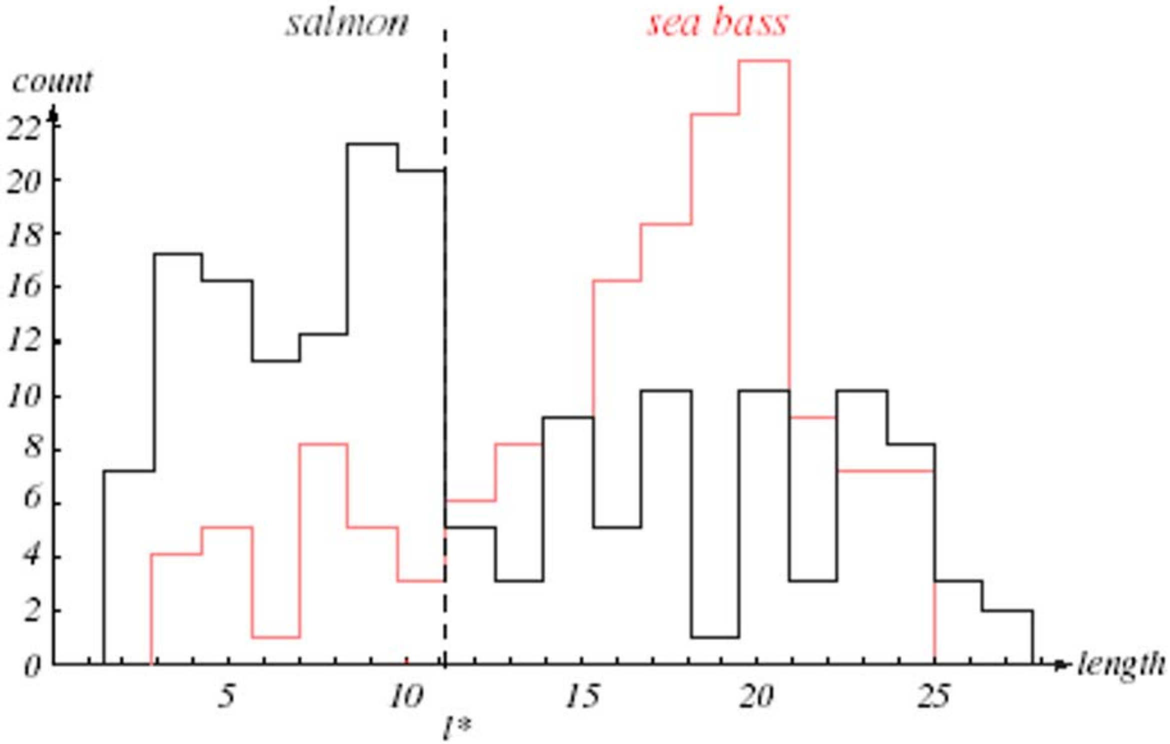


Representation: Fish Length As Feature

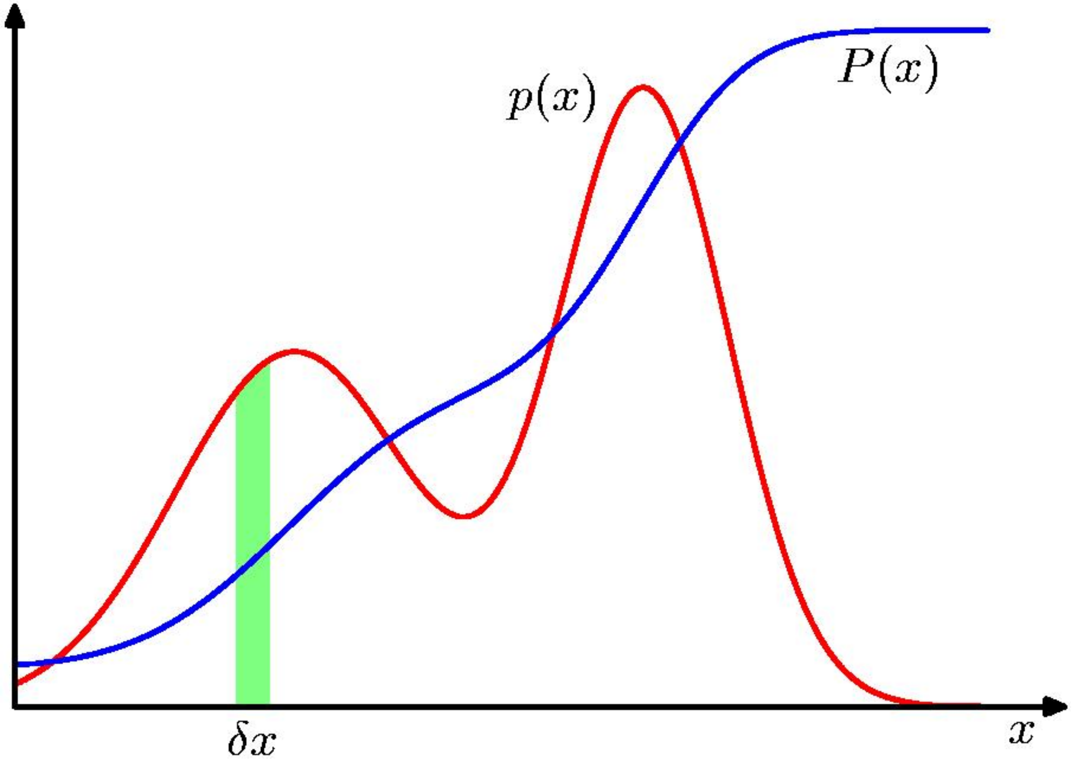
- How to design a classifier?

Representation: Fish Length As Feature

Training (design or learning) Samples



Probability Densities

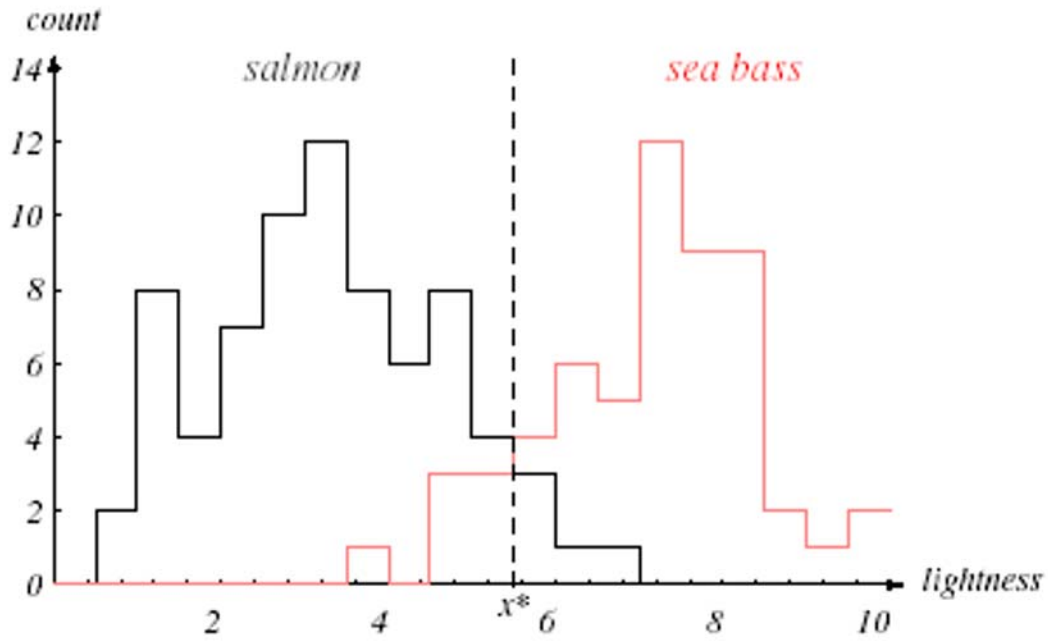


$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

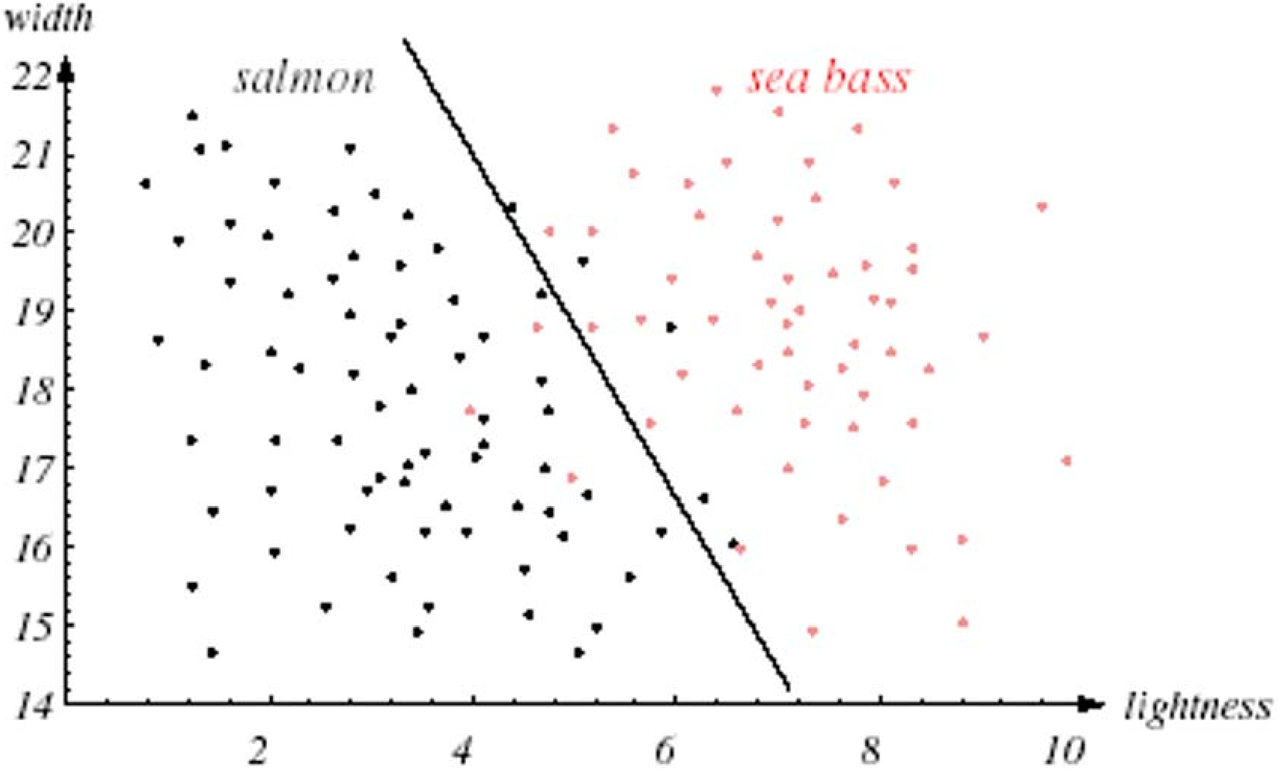
Fish Lightness As Feature



Overlap of these histograms is small compared to length feature

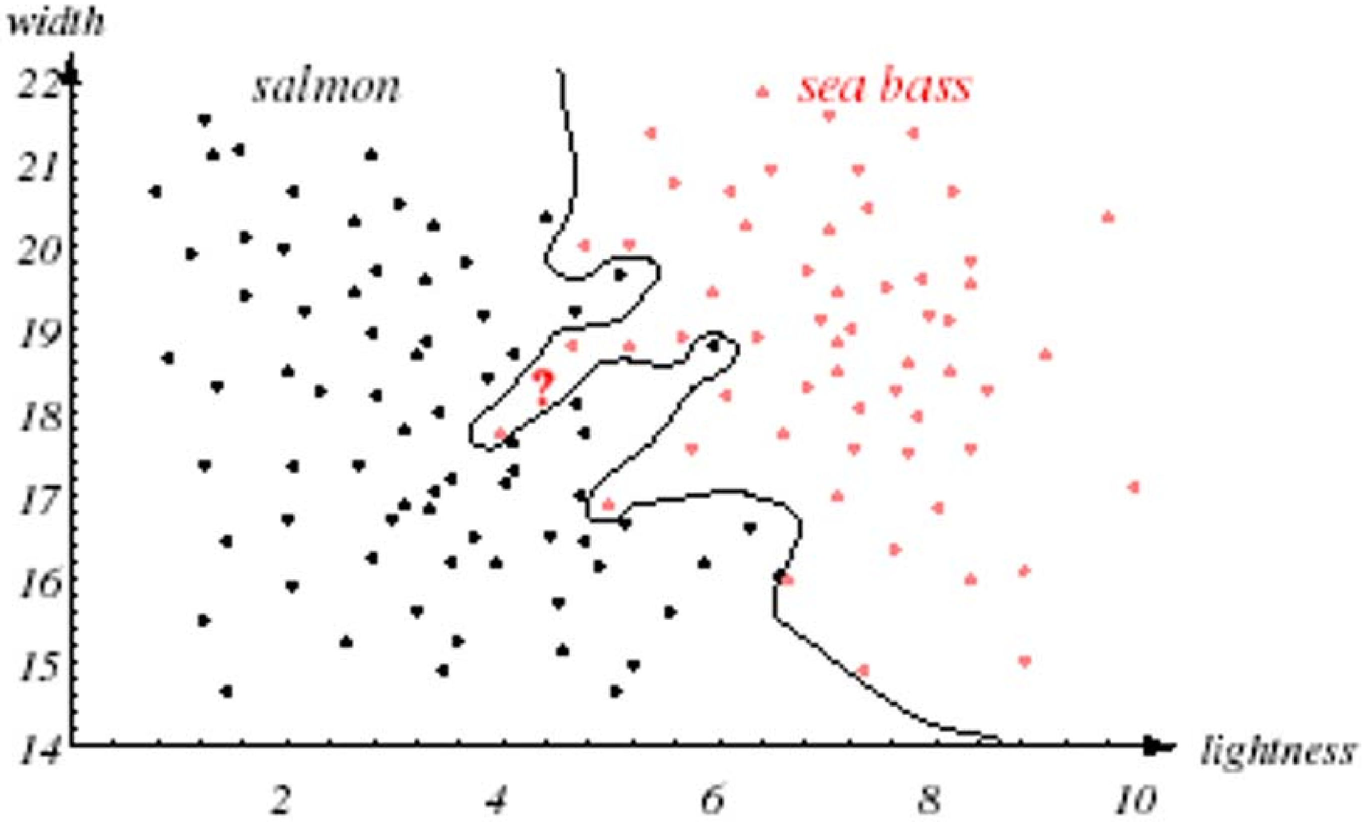
Two-dimensional Feature Space

Linear (simple) decision boundary

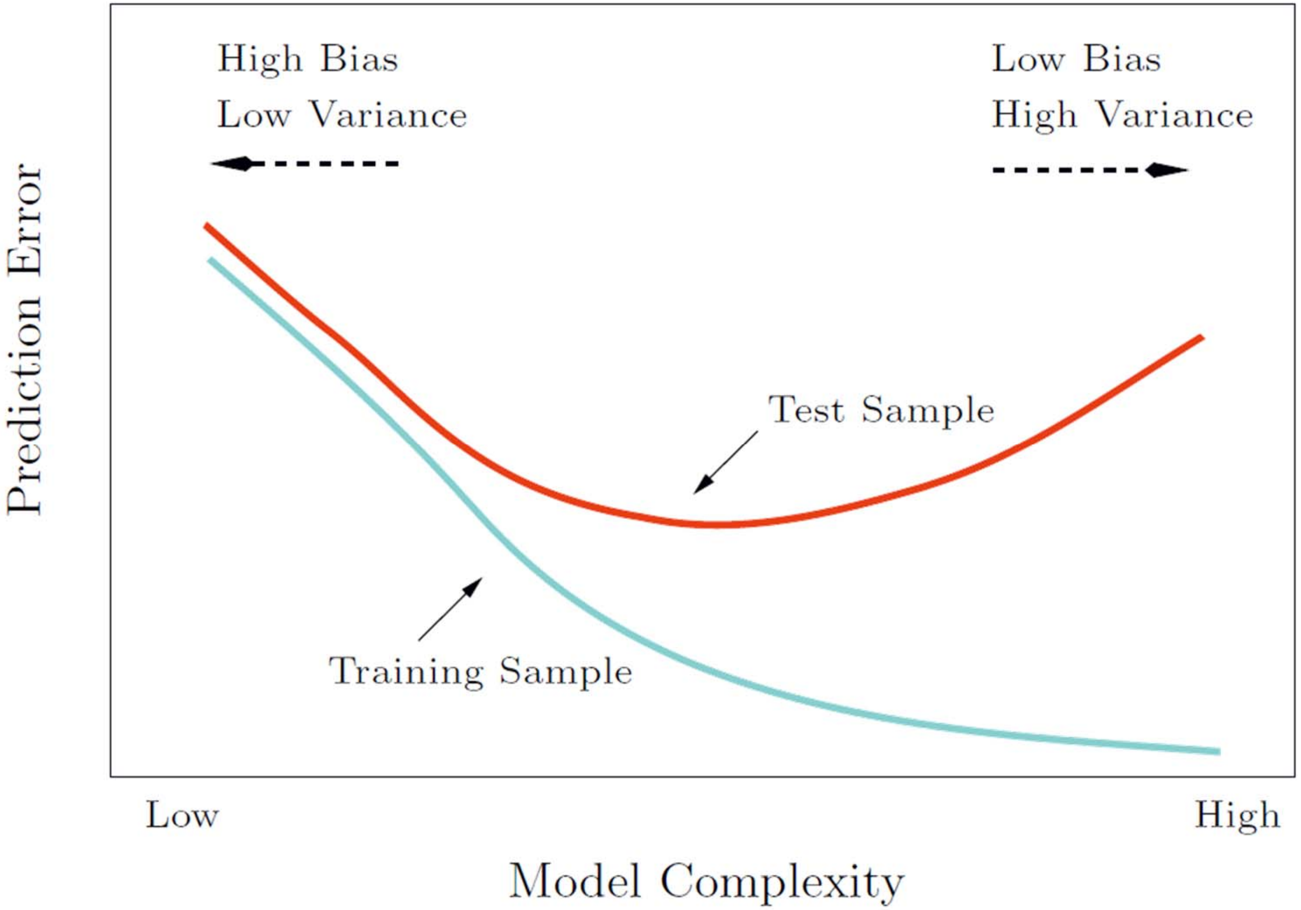


Two features together are better than individual features

Complex Decision Boundary



© Deng Cai, College of Computer Science, Zhejiang University


High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity

Generalization

- A generalization of a concept is an extension of the concept to less-specific criteria.
- Generalization of the classifier (model)
  - The performance of the classifier on test data.

- Training error:
- Simple model → large training error
- Complex model → less training error

- Test error:
- Simple model → ?
- Complex model → ?

Prerequisite Knowledge

- Probability:
  - Bayes theorem
- Analysis:
  - Gradient descent
- Linear Algebra
  - Linear space,
  - Matrix
    - Rank…
    - Positive definite matrix…
    - Eigenvector, eigenvalue
    - Singular vector, singular value