Supervised Learning

Dr. Jianlin Cheng

University of Missouri, Columbia

Slides Adapted from Book and CMU, Stanford Machine Learning Courses

Fall, 2011

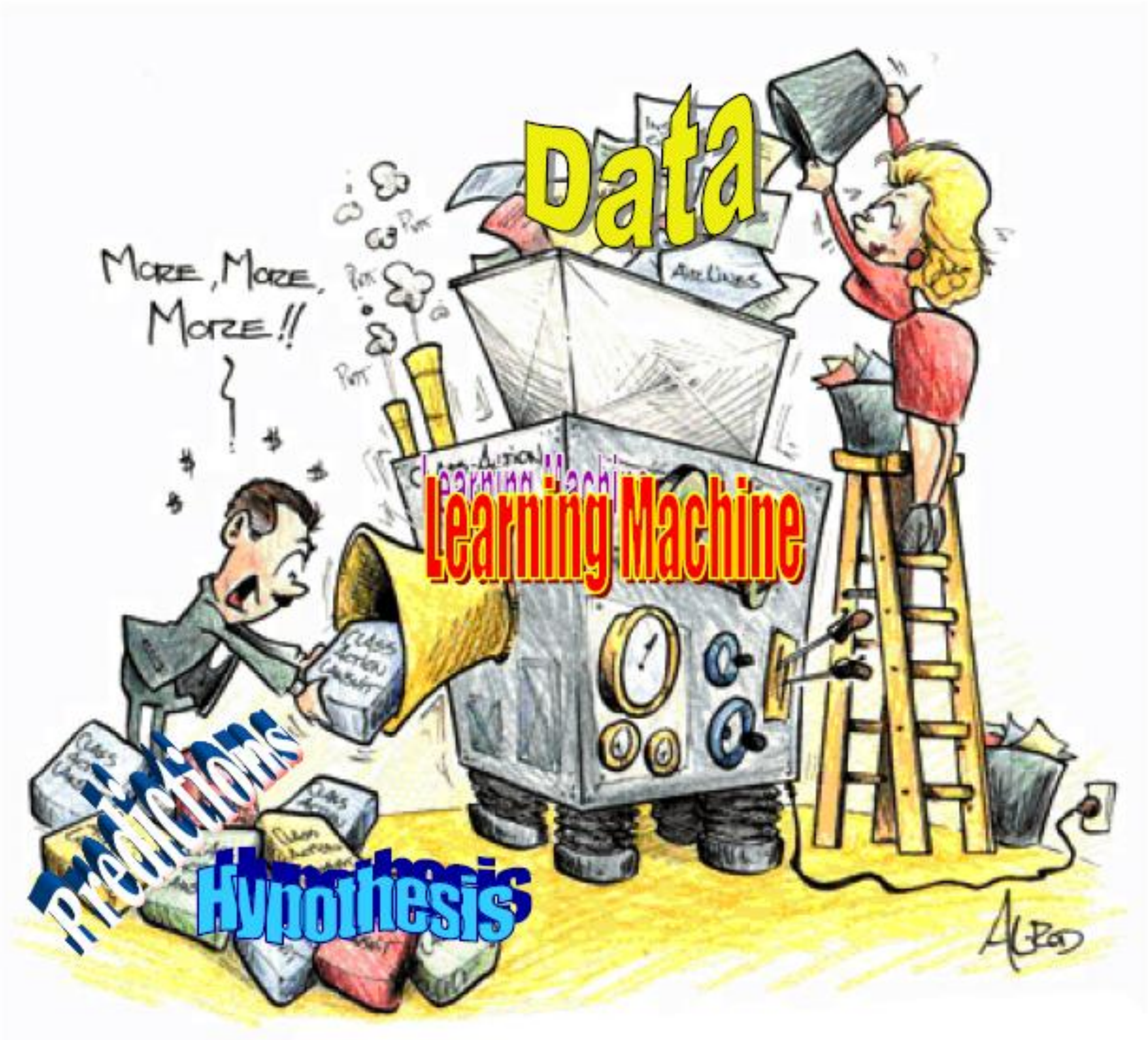
Syllabus

- **Course web site**: http://people.cs.missouri.edu/~chengji/supervised_learning/
- **Location**: EBW 355; **Time**: TuTh 4:00 pm - 5:15 pm; **Office Hours**: TuTh 3:00 pm - 4:00 pm
- **Text Book**: Pattern Recognition and Machine Learning, Christopher Bishop, Springer, 2007
- **Grading**: assignment (40%), group project report (30%), group project representation (30%); grade scale (A+, A, A-, B+, B, B-, C+, C, C-, and F)
- **Questions / Assignment submission**: mumachinelearning@gmail.com

Topics

- Introduction and Bayes optimal learning rule
- MLE and MAP (parametric learning)
- Generative VS discriminative methods
- Nonparametric methods
- Model selection
- Boosting and bagging
- Support vector machines
- Graphical models (emphasized)
- Semi-supervised learning

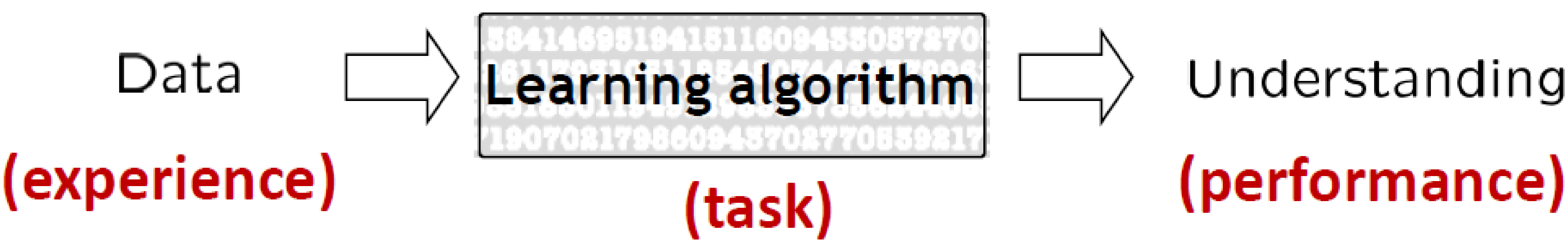
What is Machine Learning?



What is Machine Learning?

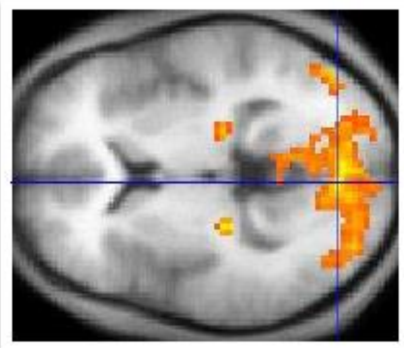
Study of algorithms that

- improve their performance
- at some task
- with experience



Machine Learning in Action

- Decoding thoughts from brain scans



Rob a bank …

Home » Health & Wellness

Brain Scans: Are You a Criminal?

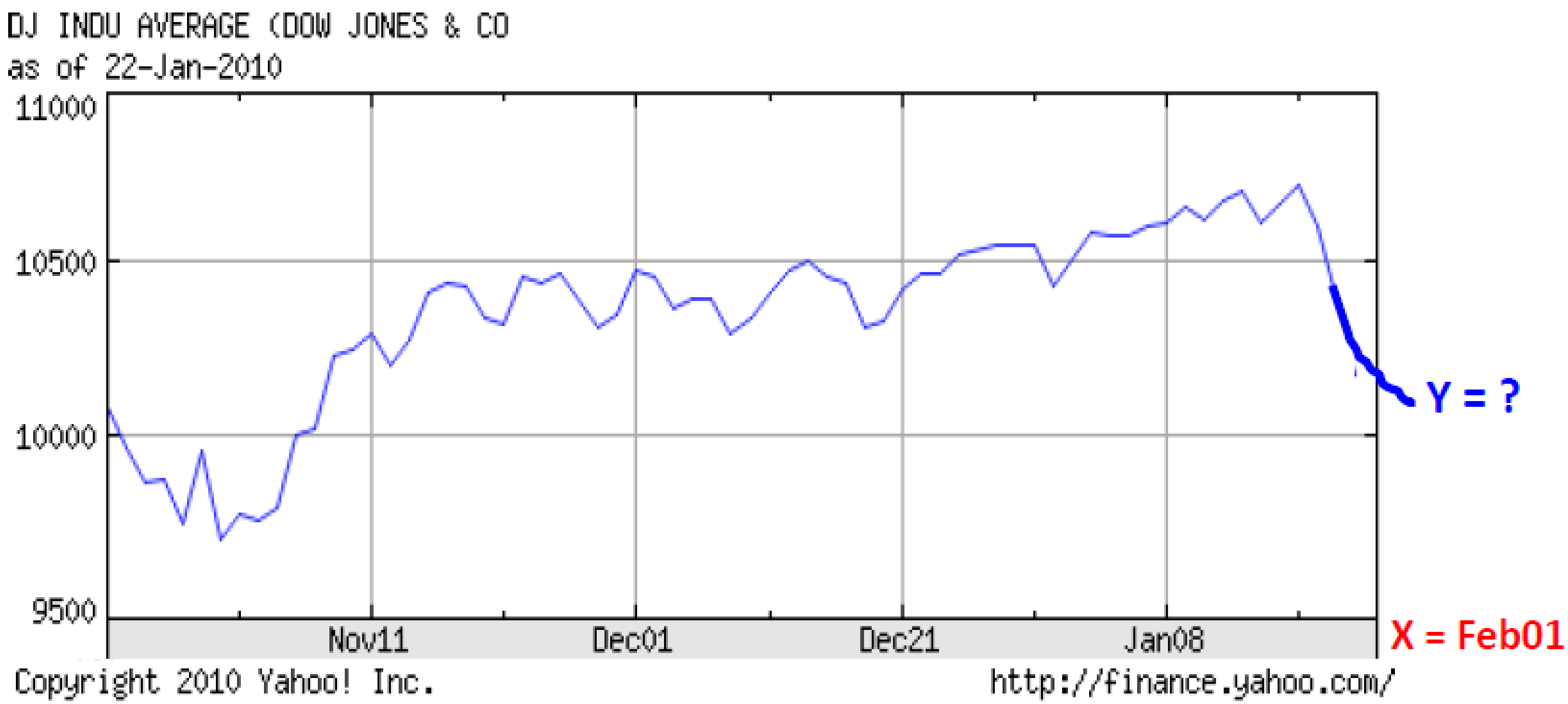
Published February 07, 2007 by:
Andrea Okrentowich
View Profile | Follow | Add to Favorites

More: Brain Scans | Brain Scan | Disposition | Defendant | Criminal Behavior

MRI Scans as Courtroom Evidence

Machine Learning in Action

- Stock Market Prediction



Machine Learning in Action

- Document classification

Sports
Science
News

Machine Learning in Action

- Spam filtering

Welcome to New Media Installation: Art that Learns

Hi everyone,

Welcome to New Media Installation:Art that Learns

The class will start tomorrow.
Make sure you attend the first class, even if you are on the Wait List.
The classes are held in Doherty Hall C316, and will be Tue, Thu 01:30-4:20 PM.

By now, you should be subscribed to our course mailing list: 10615-announce@cs.cmu.edu.

Spam/
Not spam

Natural _LoseWeight SuperFood Endorsed by Oprah Winfrey, Free Trial 1 bottle, pay only $5.95 for shipping mfw rlk Spam | X

=== Natural WeightL0SS Solution ===

Vital Acai is a natural WeightL0SS product that Enables people to lose wieght and cleansing their bodies faster than most other products on the market.

Here are some of the benefits of Vital Acai that You might not be aware of. These benefits have helped people who have been using Vital Acai daily to Achieve goals and reach new heights in there dieting that they never thought they could.

* Rapid WeightL0SS
* Increased metabolism - BurnFat & calories easily!

Machine Learning in Action

- Cars navigating on their own

Boss, the self-driving SUV
1st place in the DARPA Urban Challenge.
Photo courtesy of Tartan Racing.

Machine Learning in Action

- The **best** helicopter pilot is now a computer!
  - it runs a program that learns how to fly and make acrobatic maneuvers by itself!
  - no taped instructions, joysticks, or things like that …

[http://heli.stanford.edu/]

Machine Learning in Action

- Robot assistant?

[http://stair.stanford.edu/]

Natural language processing and speech recognition

- Now most pocket **Speech Recognizers** or **Translators** are running on some sort of learning device --- the more you play/use them, the smarter they become!

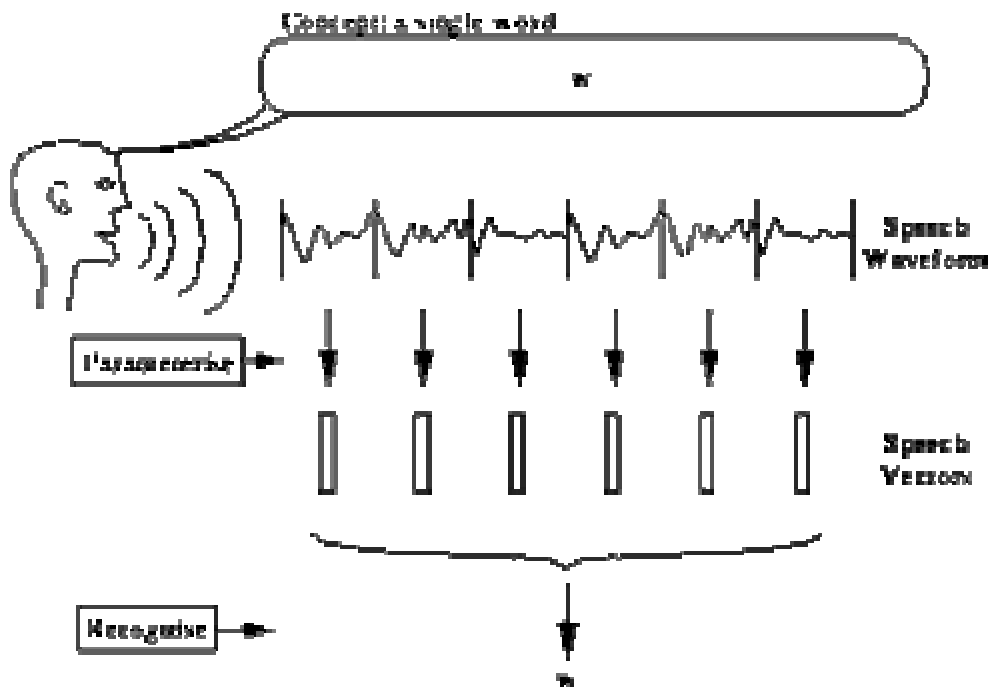


Fig. 1.2 Isolated Word Problem

Object Recognition

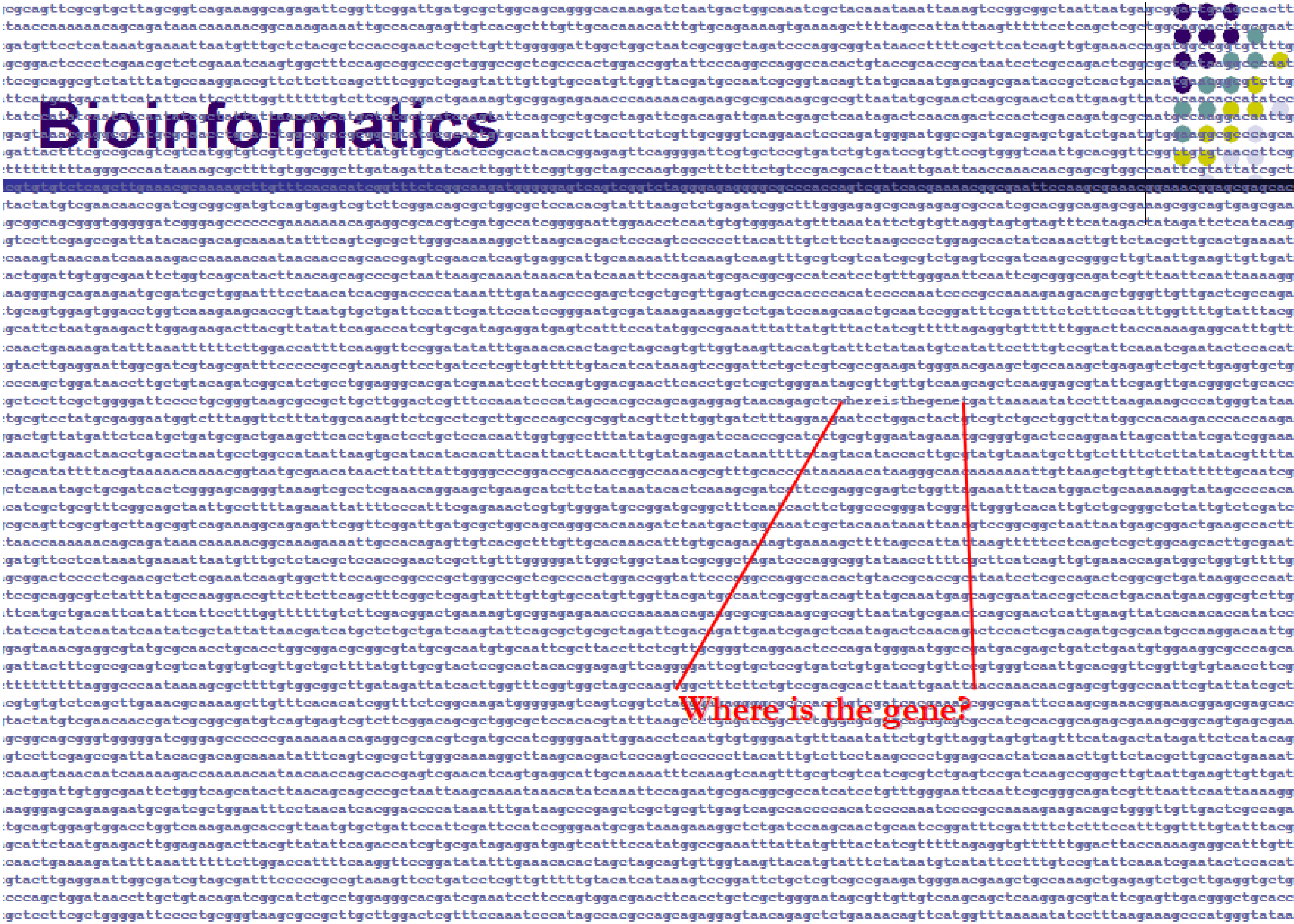
- Behind a security camera, most likely there is a computer that is learning and/or checking!

Face Recognition

Face Recognition




Bioinformatics



What this course is about

- Covers a wide range of Machine Learning techniques – from basic to state-of-the-art
- You will learn about the methods you heard about:
  - Naïve Bayes, logistic regression, nearest-neighbor, decision trees, boosting, neural nets, overfitting, regularization, dimensionality reduction, PCA, error bounds, VC dimension, SVMs, kernels, margin bounds, K-means, EM, mixture models, semi-supervised learning, HMMs, graphical models, active learning, reinforcement learning…
- Covers algorithms, theory and applications
- **It's going to be fun and hard work ☺**

Machine Learning Tasks

Broad categories -

- **Supervised learning**

  Classification, Regression

- **Unsupervised learning**

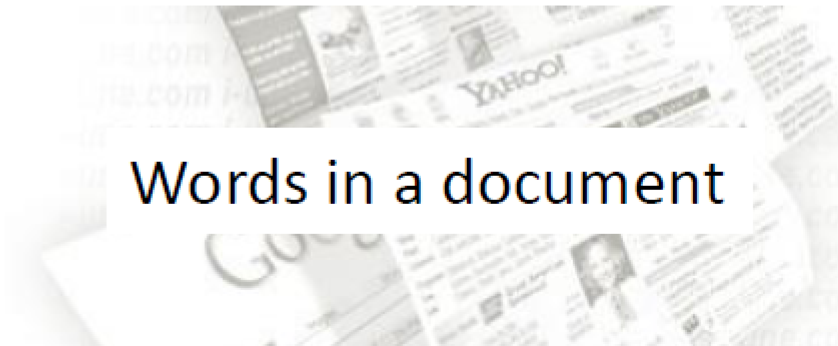
  Density estimation, Clustering, Dimensionality reduction

- Semi-supervised learning
- Active learning
- Reinforcement learning
- Many more …

Supervised Learning

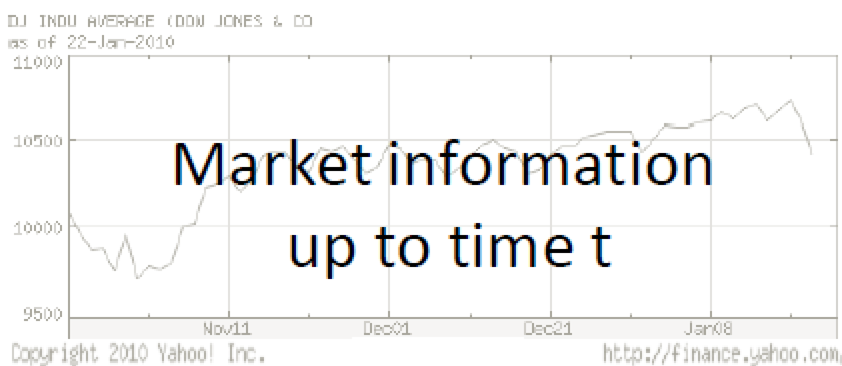
Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$



Words in a document

→ "Sports"
"News"
"Science"
...



Market information up to time t

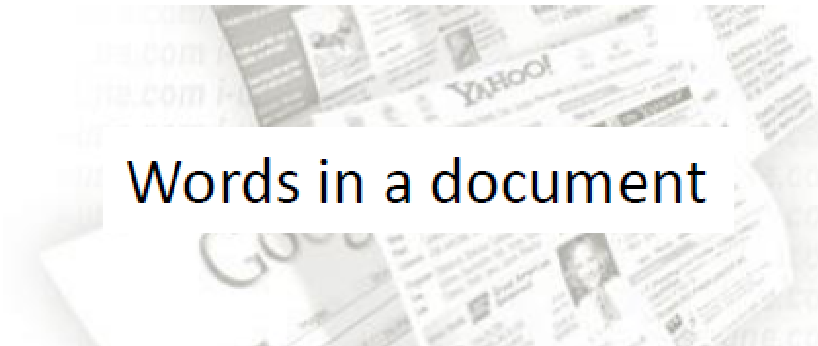
→ Share Price
"$ 24.50"

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

Supervised Learning - Classification

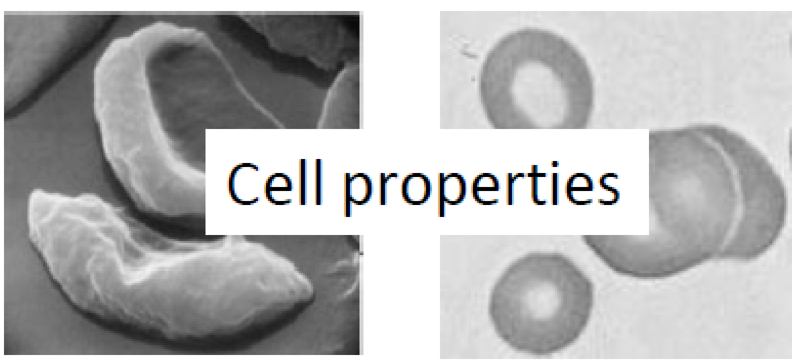
Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$



Words in a document

“Sports”
“News”
“Science”
…



Cell properties

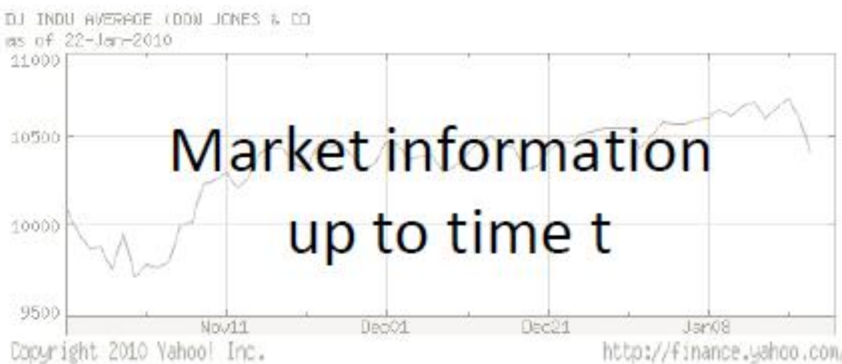
“Anemic cell”
“Healthy cell”

Discrete Labels

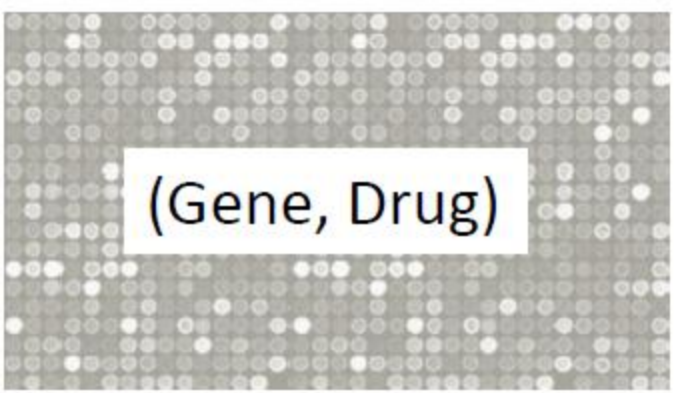
Supervised Learning - Regression

Feature Space $\mathcal{X}$

Label Space $\mathcal{Y}$



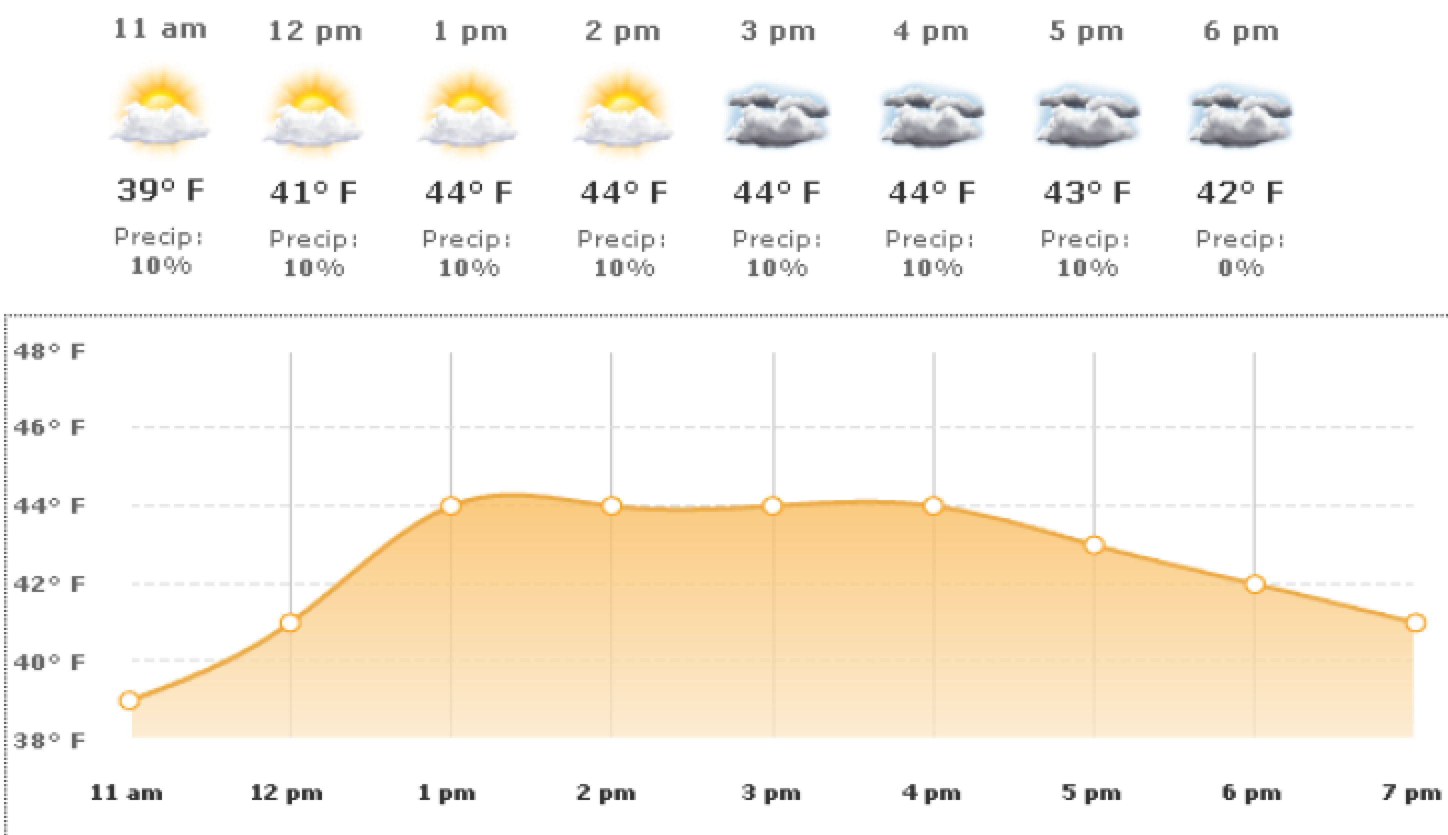
Share Price
"$ 24.50"



Expression level
"0.01"

Supervised Learning problems

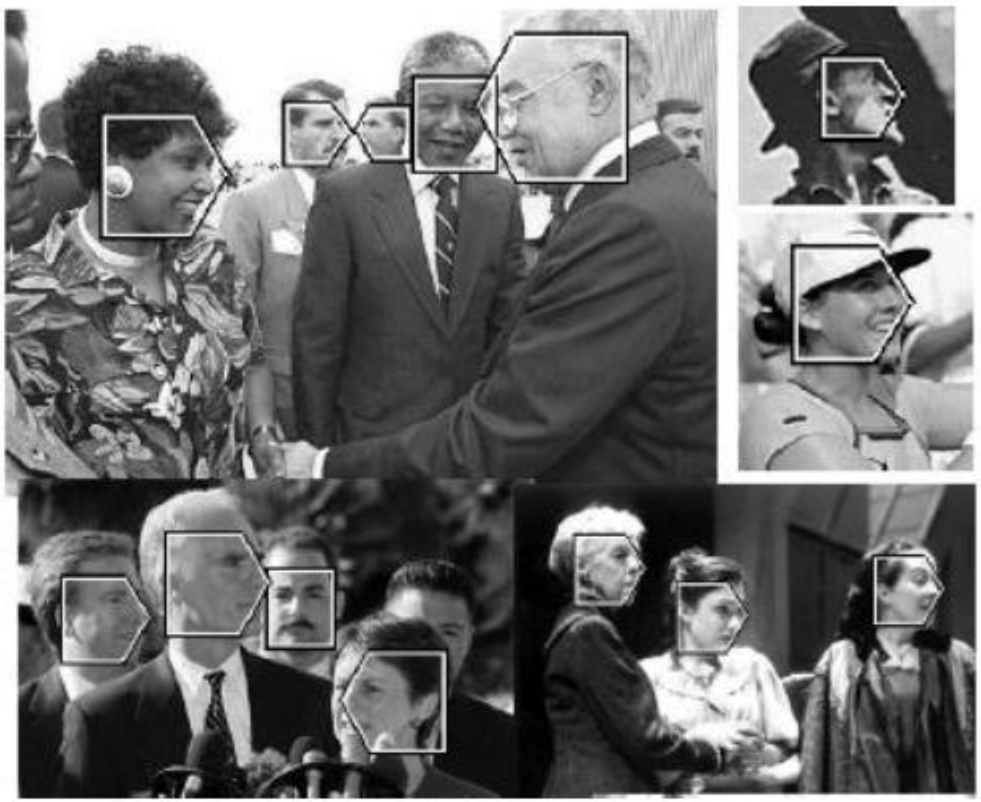
Features? Labels? Classification/Regression?



Temperature/Weather prediction

Supervised Learning problems

Features? **Labels?** **Classification/Regression?**



Face Detection

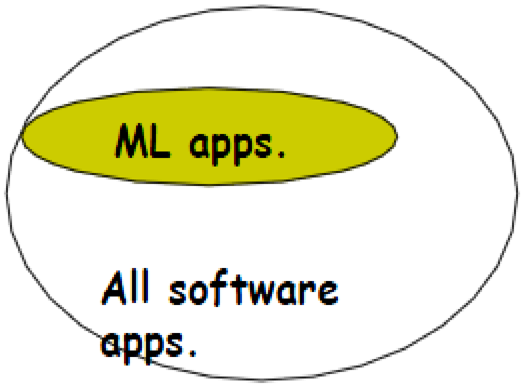
Supervised Learning problems

Features? **Labels?** **Classification/Regression?**

Environmental Mapping

Growth of Machine Learning

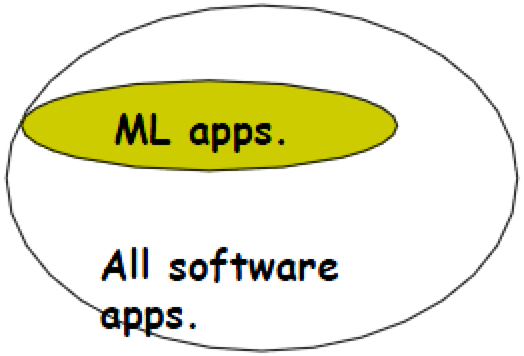
- Machine learning already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



- This ML niche is growing (why?)

Growth of Machine Learning

- Machine learning already the preferred approach to
  - Speech recognition, Natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - …



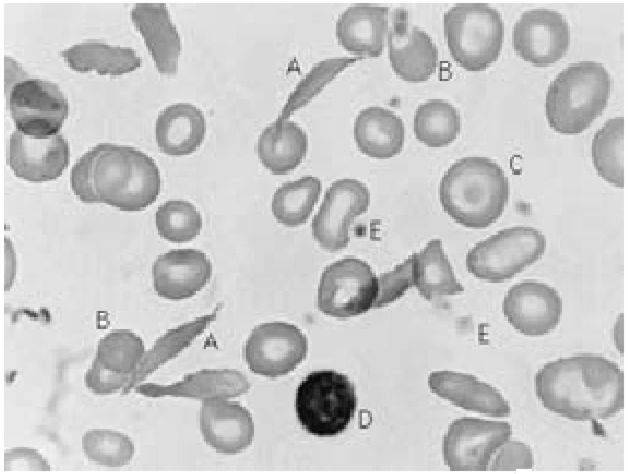
- This ML niche is growing
  - Improved machine learning algorithms
  - Increased data capture, networking
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

Function Approximation

- **Setting**:
  - Set of possible instances X
  - Unknown target function $f: X \rightarrow Y$
  - Set of function hypotheses $H = \{ h \mid h: X \rightarrow Y \}$

- **Given**:
  - Training examples $\{<x_i, y_i>\}$ of unknown target function f

- **Determine**:
  - Hypothesis $h \in H$ that best approximates f



Probably approximately correct learning

<u>PAC Learning Theory</u>
(supervised concept learning)

examples (*m*)

representational complexity (*H*)

error rate (ε)

failure probability (δ)

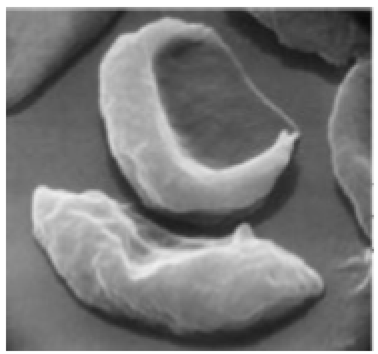
$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Occam's Razor – When everything is equal, a simple Solution is better.

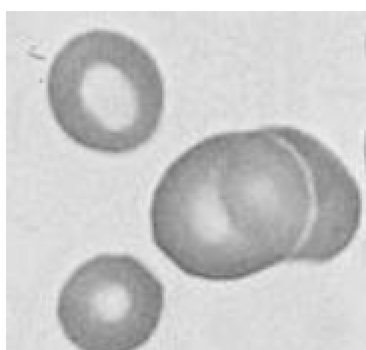
Supervised Learning Task

Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$. $\boxed{X \text{ - test data}}$

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \rightarrow \mathcal{Y}$



"Anemic cell (0)"

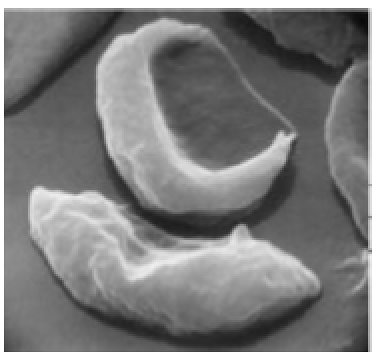


"Healthy cell (1)"

Performance Measures

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label Y and prediction $f(X)$

X	Y	$f(X)$	$\text{loss}(Y, f(X))$
	"Anemic cell"	"Anemic cell"	0
		"Healthy cell"	1

$$\boxed{\text{loss}(Y, f(X)) = 1_{\{f(X) \neq Y\}} \quad \textbf{0/1 loss}}$$

Performance Measures

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label *Y* and prediction *f*(*X*)

X	Share price, *Y*	*f*(*X*)	$\text{loss}(Y, f(X))$
Past performance, trade volume etc. as of Sept 8, 2010	"$24.50"	"$24.50"	0
		"$26.00"	1?
		"$26.10"	2?

$$\boxed{\text{loss}(Y, f(X)) = (f(X) - Y)^2 \quad \text{square loss}}$$

Performance Measures

Performance:

$\text{loss}(Y, f(X))$ - Measure of closeness between true label Y and prediction $f(X)$

Don't just want label of one test data (cell image), but any cell image $X \in \mathcal{X}$

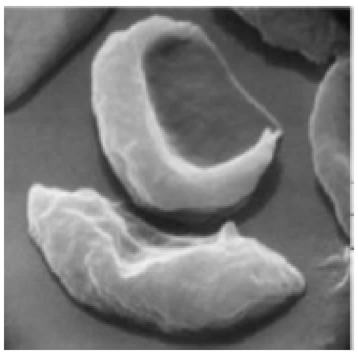
$$(X, Y) \sim P_{XY}$$

Given a cell image drawn randomly from the collection of all cell images, how well does the predictor perform on average?

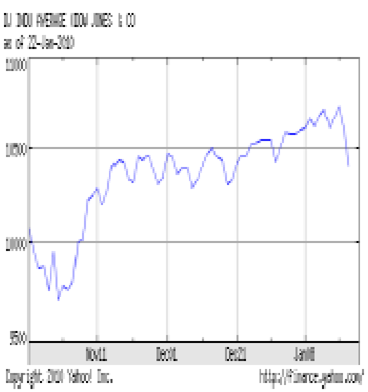
$$\boxed{\text{Risk } R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]}$$

Performance Measures

Performance: Risk $R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$



$\Rightarrow$ "Anemic cell"



$\Rightarrow$ Share Price "$ 24.50"

$\text{loss}(Y, f(X))$	Risk $R(f)$
$1_{\{f(X) \neq Y\}}$ **0/1 loss**	$P(f(X) \neq Y)$ **Probability of Error**
$(f(X) - Y)^2$ **square loss**	$\mathbb{E}[(f(X) - Y)^2]$ **Mean Square Error**

Bayes Optimal Rule

Ideal goal: Construct **prediction rule** $f^* : \mathcal{X} \to \mathcal{Y}$

$$f^* = \arg\min_f \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$$

Bayes optimal rule

Best possible performance:

Bayes Risk $\quad R(f^*) \leq R(f)$ for all f

BUT… Optimal rule is not computable - depends on unknown P_{XY} !

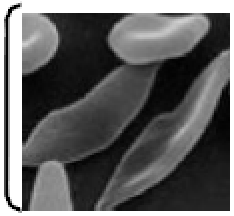
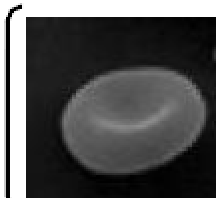
Experience - Training Data

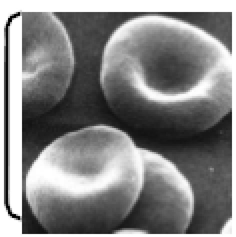
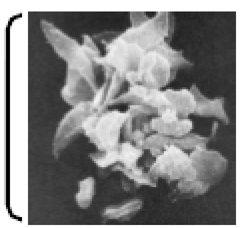
Can't minimize risk since P_{XY} unknown!

Training data (experience) provides a glimpse of P_{XY}

(observed) $\{(X_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P_{XY}$ **(unknown)**

independent, identically distributed

[ , Anemic cell] [ , Healthy cell]

[ , Healthy cell] [ , Anemic cell]

Provided by expert, measuring device, some experiment, ...

Supervised Learning

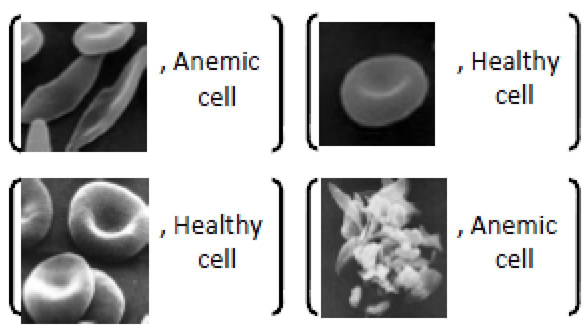
Task: Given $X \in \mathcal{X}$, predict $Y \in \mathcal{Y}$.

$\equiv$ Construct **prediction rule** $f : \mathcal{X} \to \mathcal{Y}$

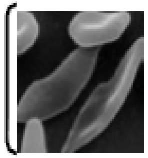
Performance: Risk $R(f) \equiv \mathbb{E}_{XY}\left[\text{loss}(Y, f(X))\right]$

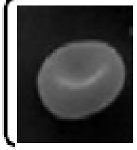
$$(X, Y) \sim P_{XY}$$

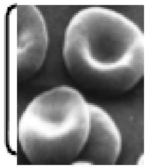
Experience: Training data $\{(X_i, Y_i)\}_{i=1}^{n} \overset{i.i.d.}{\sim} P_{XY}$ **(unknown)**



Machine Learning Algorithm


, Anemic cell

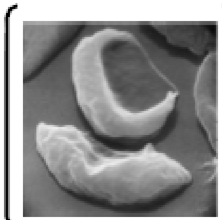

, Healthy cell


, Healthy cell


, Anemic cell

Training data $\{(X_i, Y_i)\}_{i=1}^{n}$ $\Rightarrow$ **Learning algorithm** $\Rightarrow$ $\widehat{f}_n$

$\widehat{f}_n$ is a mapping from $\mathcal{X} \rightarrow \mathcal{Y}$

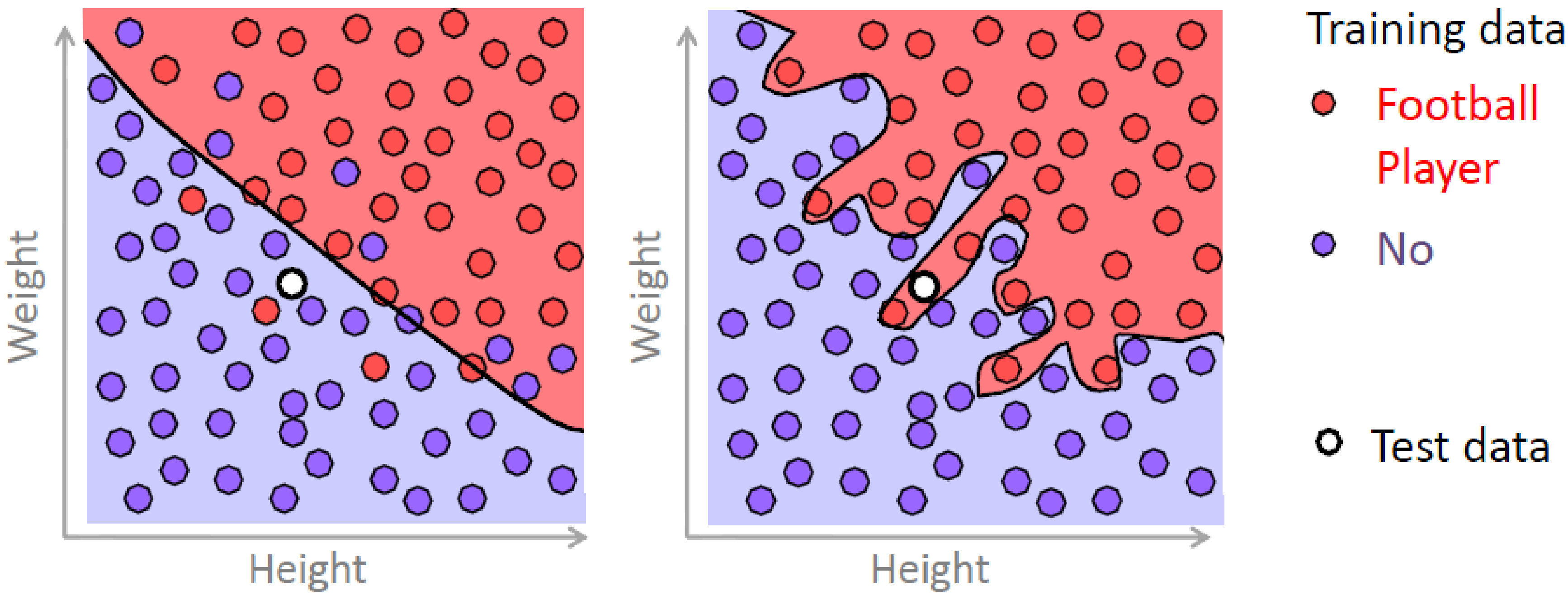
$\widehat{f}_n$ [] = "Anemic cell"

Test data X

Note: test data ≠ training data

Issues in ML

- A good machine learning algorithm
  - Does not **overfit** training data



  - **Generalizes** well to test data

More later …

How to sense Generalization Error?

- Can't compute generalization error. How can we get a sense of how well algorithm is performing in practice?
- One approach -
  - Split available data into two sets $\{(X_i, Y_i)\}_{i=1}^n$ $\{(X_i', Y_i')\}_{i=1}^n$
  - Training Data – used for training the algorithm

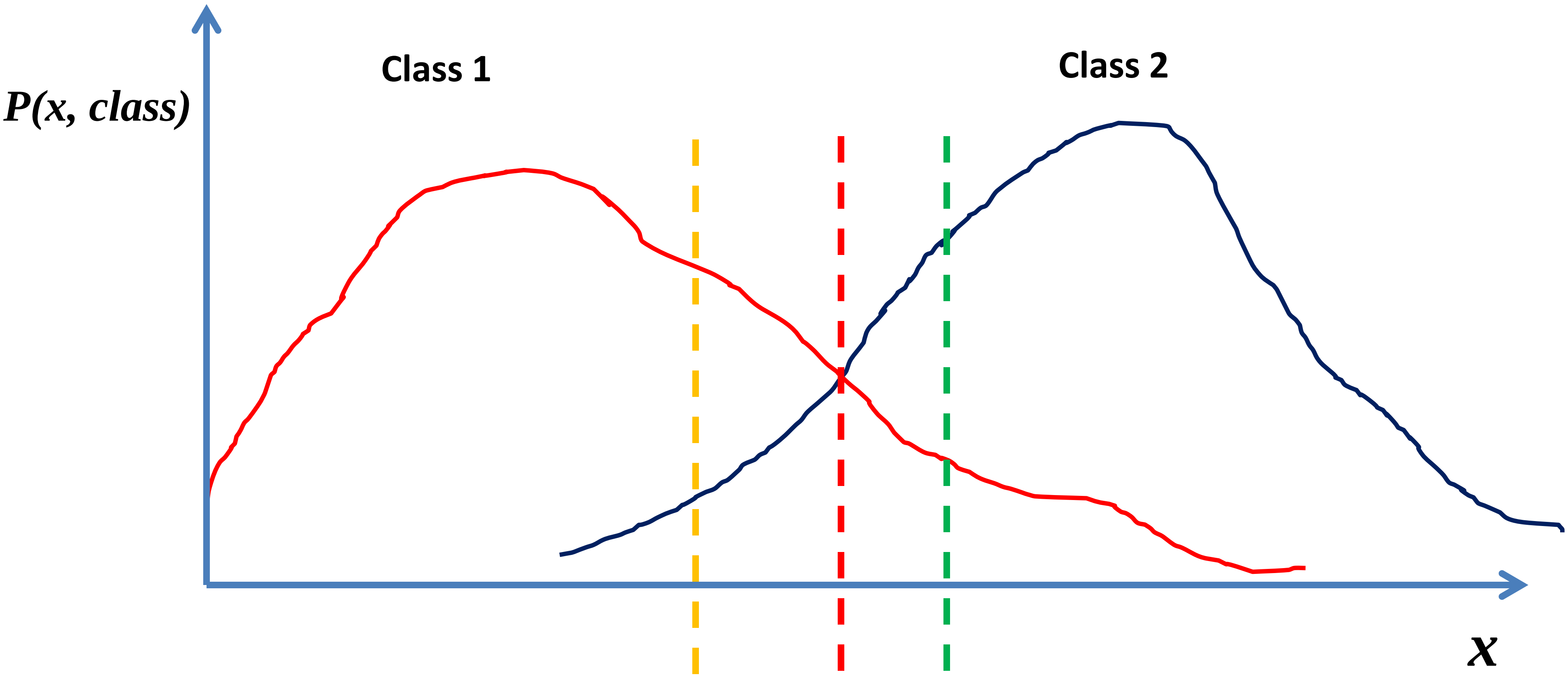
$$\{(X_i, Y_i)\}_{i=1}^n \Rightarrow \boxed{\textbf{Learning algorithm}} \Rightarrow \widehat{f}_n$$

  - Test Data (a.k.a. Validation Data, Hold-out Data) – provides estimate of generalization error

Test Error = $\frac{1}{n}\sum_{i=1}^{n}\left[\text{loss}(Y_i', \widehat{f}_n(X_i'))\right]$

Why not use Training Error?

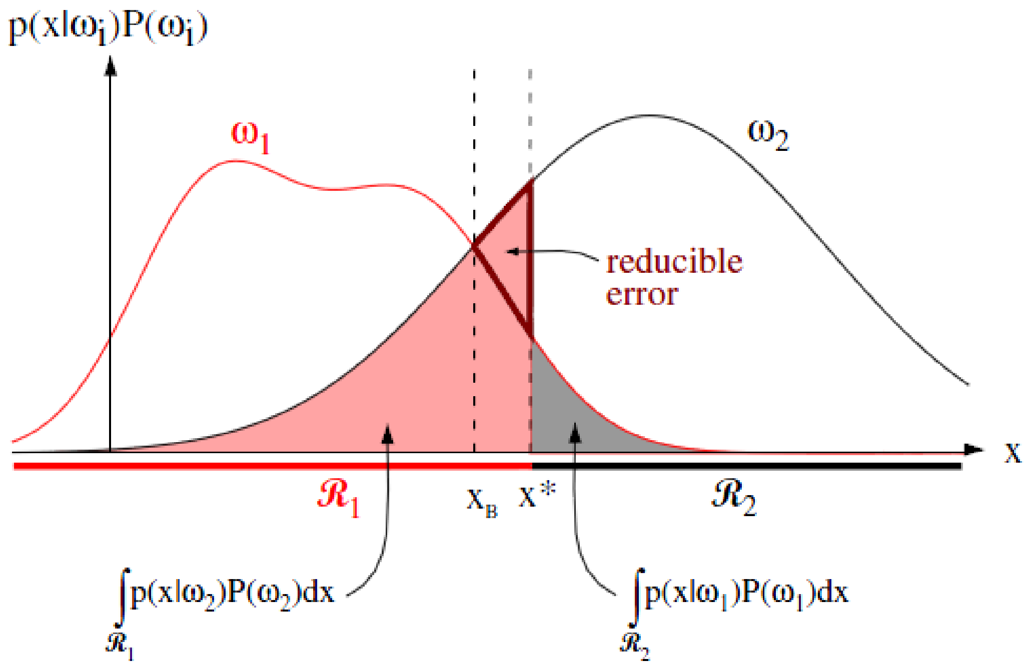
How to minimize errors?



Where to set a threshold on x to make classification in order to minimize classification errors?

Bayes Error

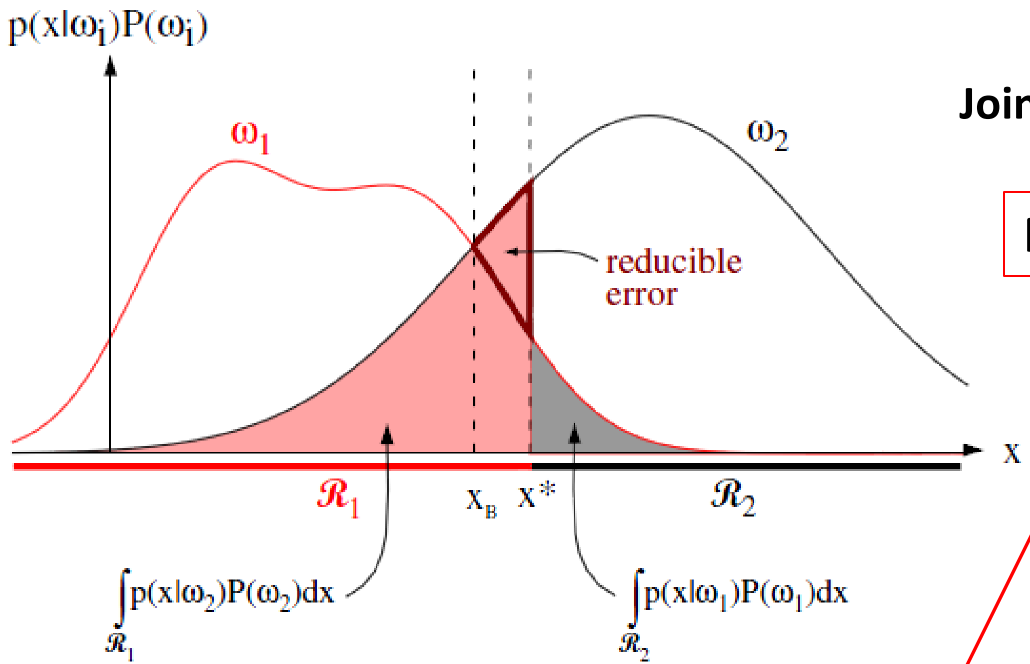
Calculate the probability of an error – Bayes error



$$
\begin{aligned}
P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\
&= P(\mathbf{x} \in \mathcal{R}_2|\omega_1)P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1|\omega_2)P(\omega_2) \\
&= \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)P(\omega_1)\, d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)P(\omega_2)\, d\mathbf{x}.
\end{aligned}
$$

Bayes Error

Calculate the probability of an error – Bayes error



Joint Prob = Con. Prob * Marginal Prob.

$$\boxed{P(x,y) = P(x|y)P(y)}$$

$$\begin{aligned} P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2|\omega_1)P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1|\omega_2)P(\omega_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)P(\omega_1)\, d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)P(\omega_2)\, d\mathbf{x}. \end{aligned}$$

Classifiers and Bayes error

- A classifier h is a mapping from feature vectors $\boldsymbol{x}$ to class labels $\{C_0, C_1\}$
- The Bayes error of h is the probability of a misclassification:

$$\boxed{\int_{\boldsymbol{x}\in H_0} P(C_1 \mid \boldsymbol{x})\, p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} + \int_{\boldsymbol{x}\in H_1} P(C_0 \mid \boldsymbol{x})\, p(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}}$$

Area that h classifies x as C_0

Bayes optimal classifiers

- Classifier that minimizes the Bayes error is called the Bayes optimal classifier:

- classify x as $\begin{cases} C_0 \text{ if } P(C_0 \mid x) > P(C_1 \mid x) \\ C_1 \text{ otherwise} \end{cases}$