Introduction to Machine Learning

Brown University

CSCI 1950-F

Summer 2011

Instructor: Jeff Miller

http://www.dam.brown.edu/people/jmiller/ML/

What is “machine learning”?

Given a collection of examples,

predict something about novel examples.

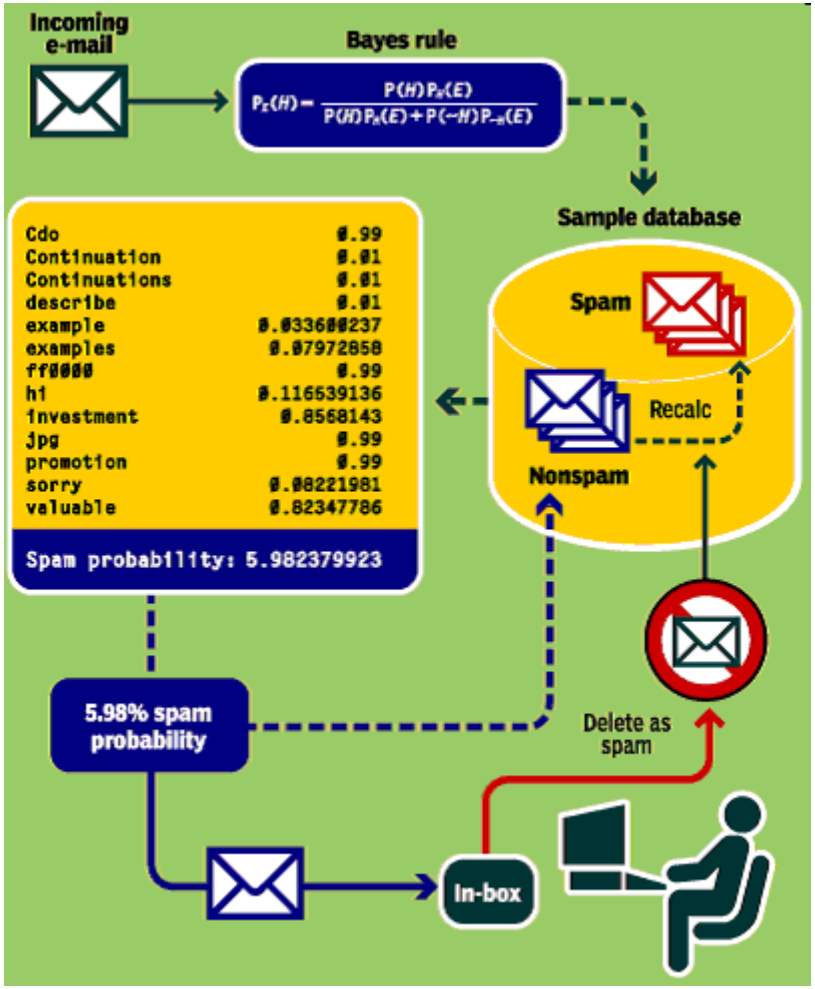
The examples are often incomplete.

“I keep saying the sexy job in the next 10 years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s? The ability to take data --- to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it --- that's going to be a hugely important skill in the next decades.”

- Hal Varian, Chief Economist at Google

Spam Filtering

- Binary classification problem: is this e-mail useful or spam?
- Noisy training data: messages previously marked as spam
- Wrinkle: spammers evolve to counter filter innovations



Spam Filter Express

http://www.spam-filter-express.com/

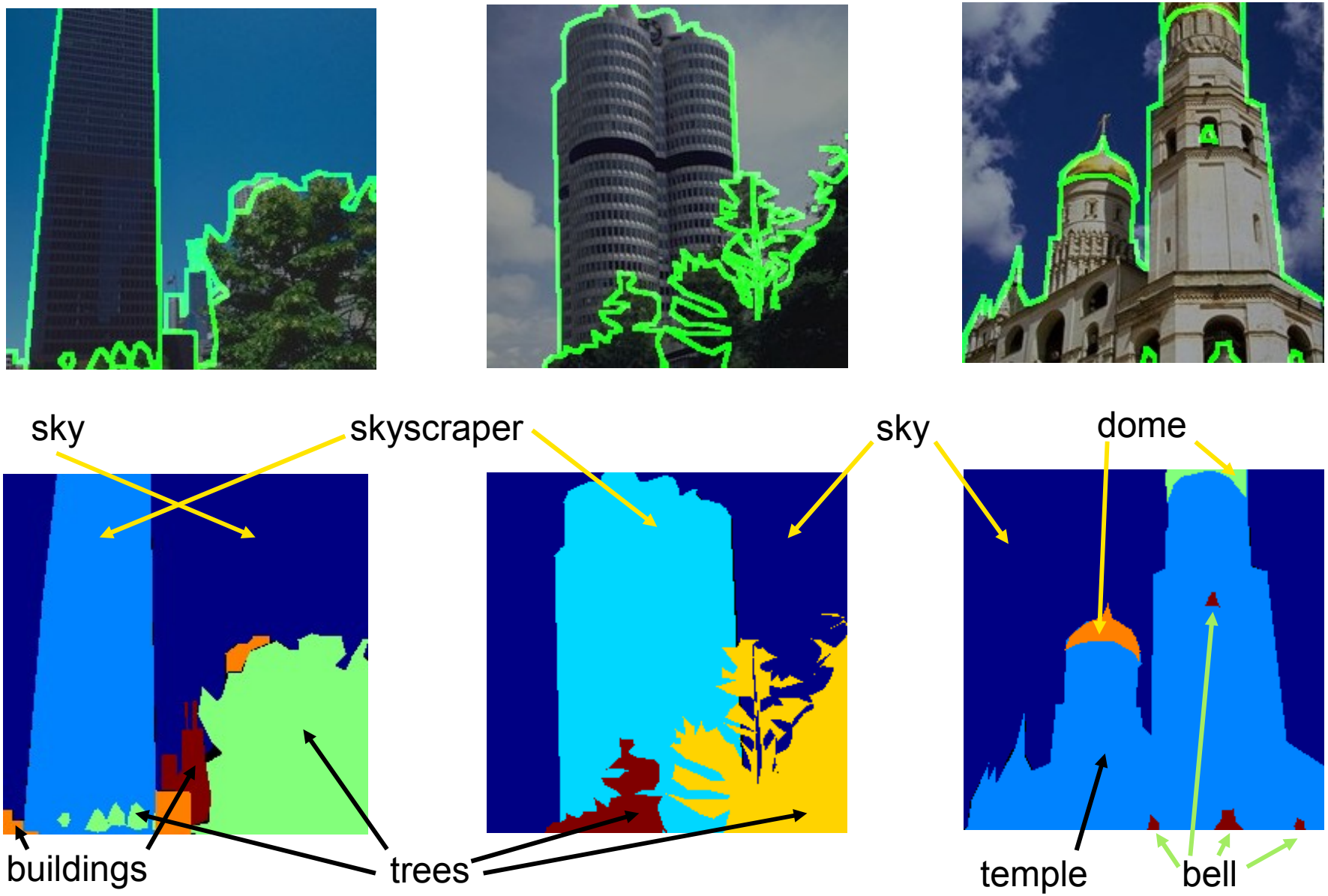
Collaborative Filtering

Leaderboard

Display top 20 leaders.

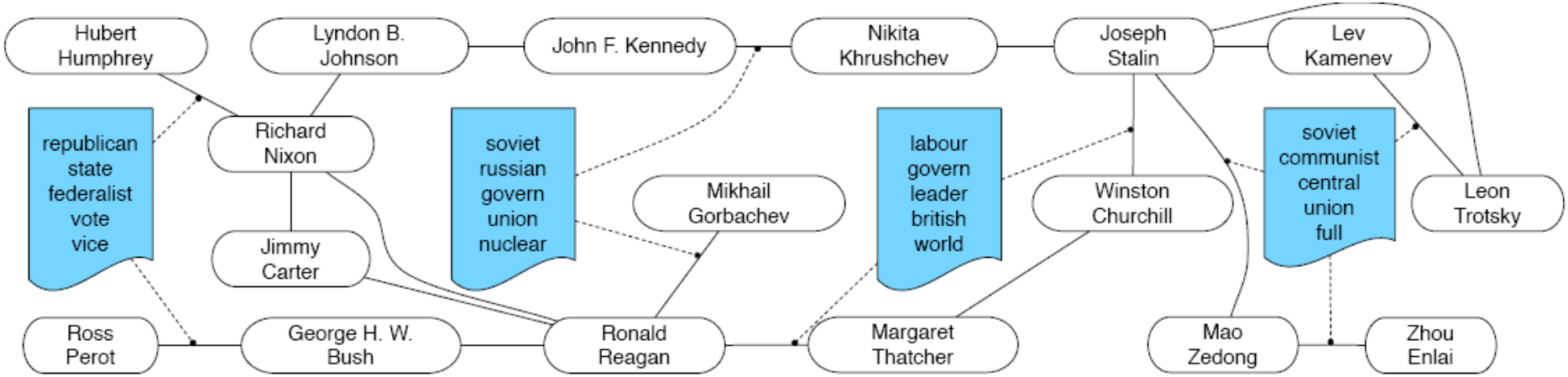
Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
Grand Prize - RMSE <= 0.8563				
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelay United	0.8573	9.89	2009-07-25 20:05:52
5	Vandelay Industries !	0.8579	9.83	2009-07-26 02:49:53
6	PragmaticTheory	0.8582	9.80	2009-07-12 15:09:53
7	BellKor in BigChaos	0.8590	9.71	2009-07-26 12:57:25
8	Dace	0.8603	9.58	2009-07-24 17:18:43
9	Opera Solutions	0.8611	9.49	2009-07-26 18:02:08
10	BellKor	0.8612	9.48	2009-07-26 17:19:11
11	BigChaos	0.8613	9.47	2009-06-23 23:06:52
12	Feeds2	0.8613	9.47	2009-07-24 20:06:46
Progress Prize 2008 - RMSE = 0.8616 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8633	9.26	2009-07-21 02:04:40
14	Gravity	0.8634	9.25	2009-07-26 15:58:34
15	Ces	0.8642	9.17	2009-07-25 17:42:38
16	Invisible Ideas	0.8644	9.14	2009-07-20 03:26:12
17	Just a guy in a garage	0.8650	9.08	2009-07-22 14:10:42
18	Craig Carmichael	0.8656	9.02	2009-07-25 16:00:54
19	J Dennis Su	0.8658	9.00	2009-03-11 09:41:54
20	acmehill	0.8659	8.99	2009-04-16 06:29:35
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
Cinematch score on quiz subset - RMSE = 0.9514				

Visual Object Recognition



Social Network Analysis

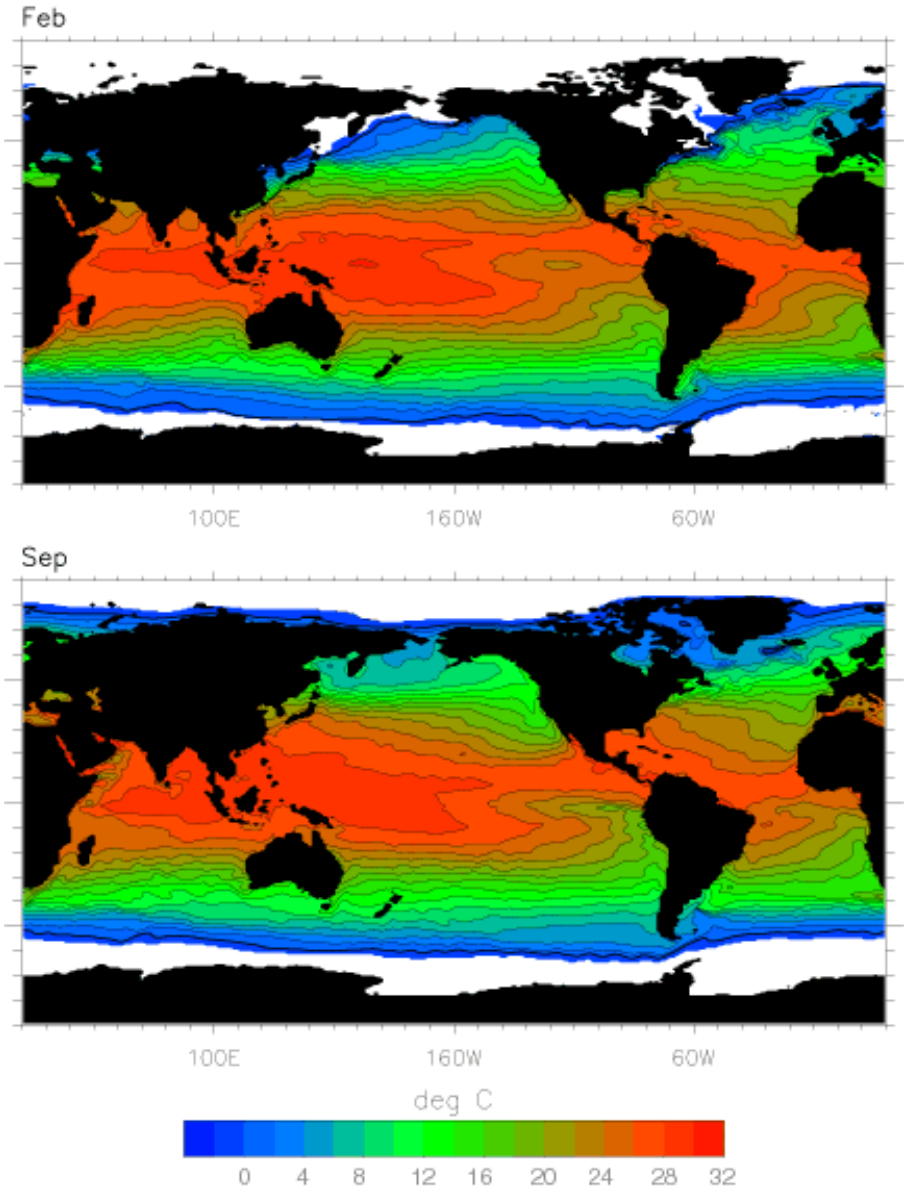
- Unsupervised discovery and visualization of relationships among people, companies, etc.
- Example: infer relationships among named entities directly from Wikipedia entries



Chang, Boyd-Graber, & Blei, KDD 2009

Climate Modeling

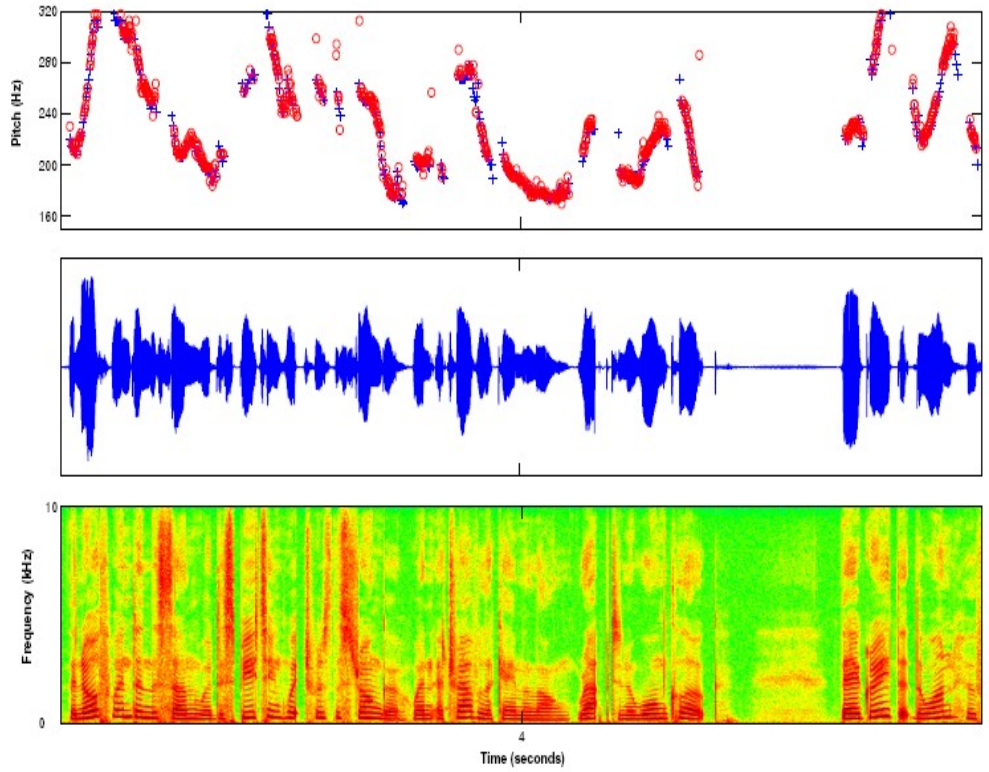
- Satellites measure sea-surface temperature at sparse locations
  - Partial coverage of ocean surface
  - Sometimes obscured by clouds, weather
- Would like to infer a dense temperature field, and track its evolution

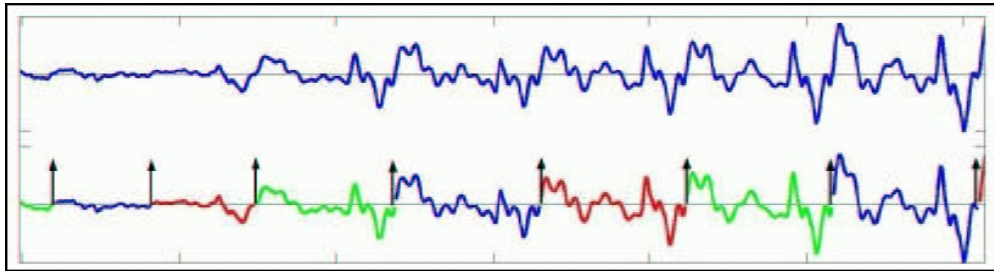


NASA Seasonal to Interannual Prediction Project
http://ct.gsfc.nasa.gov/annual.reports/ess98/nsipp.html

Speech Recognition

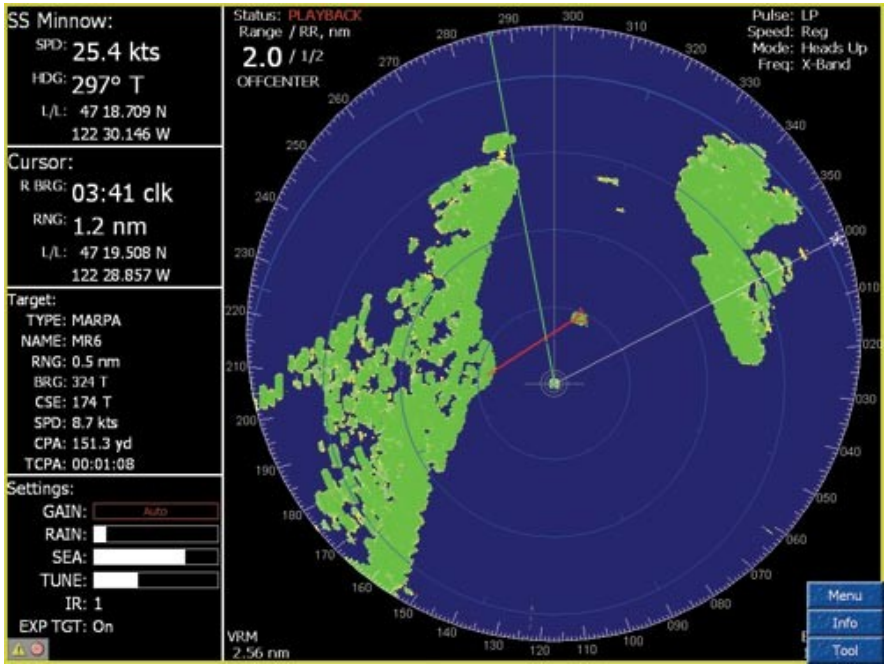
- Given an audio waveform, robustly extract & recognize any spoken words
- Statistical models can be used to
  - ➢ Provide greater robustness to noise
  - ➢ Adapt to accent of different speakers
  - ➢ Learn from training



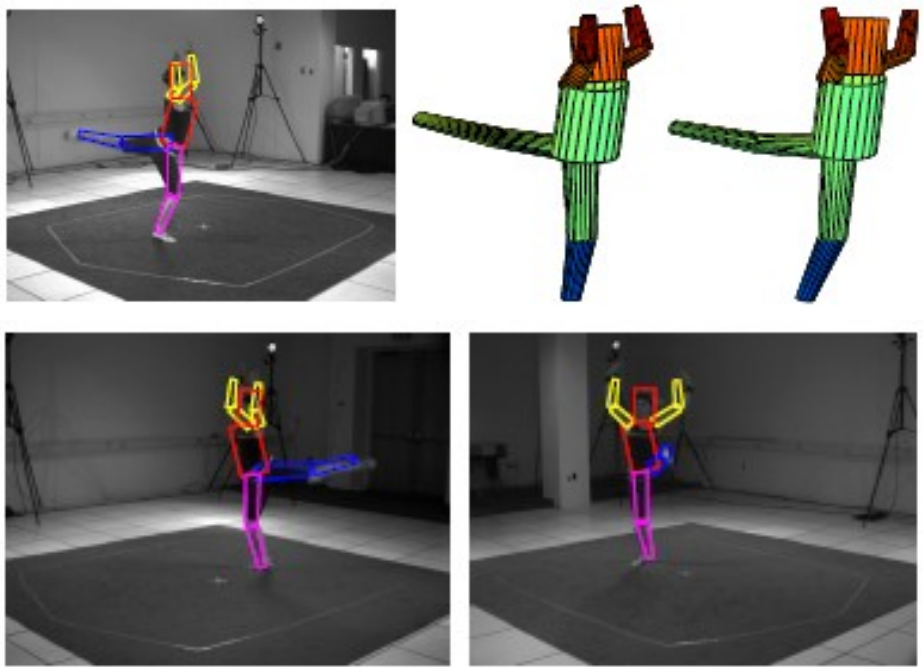


S. Roweis, 2004

Target Tracking



Radar-based tracking of multiple targets



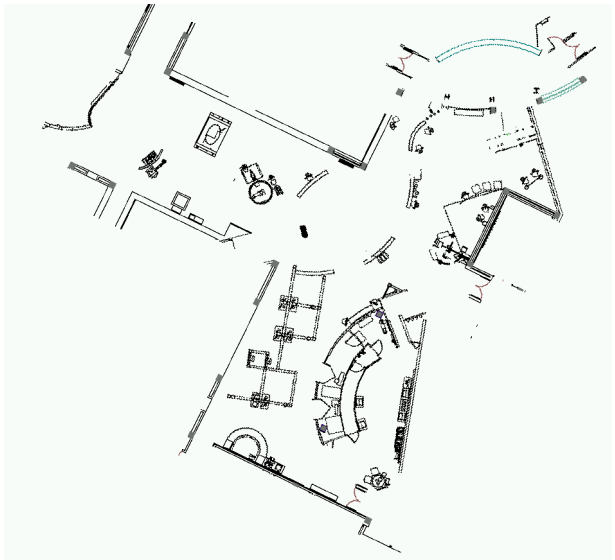
Visual tracking of articulated objects

(L. Sigal et. al., 2006)

- Estimate motion of targets in 3D world from indirect, potentially noisy measurements

Robot Navigation: *SLAM*

Simultaneous Localization and Mapping

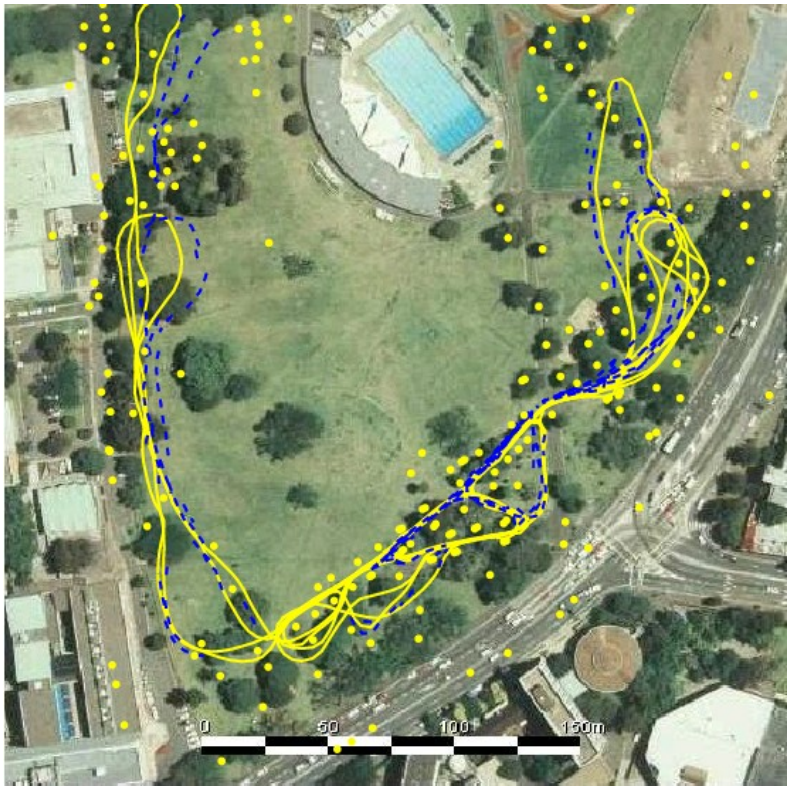


CAD Map

(S. Thrun, San Jose Tech Museum)

Estimated Map

Landmark SLAM
(E. Nebot, Victoria Park)



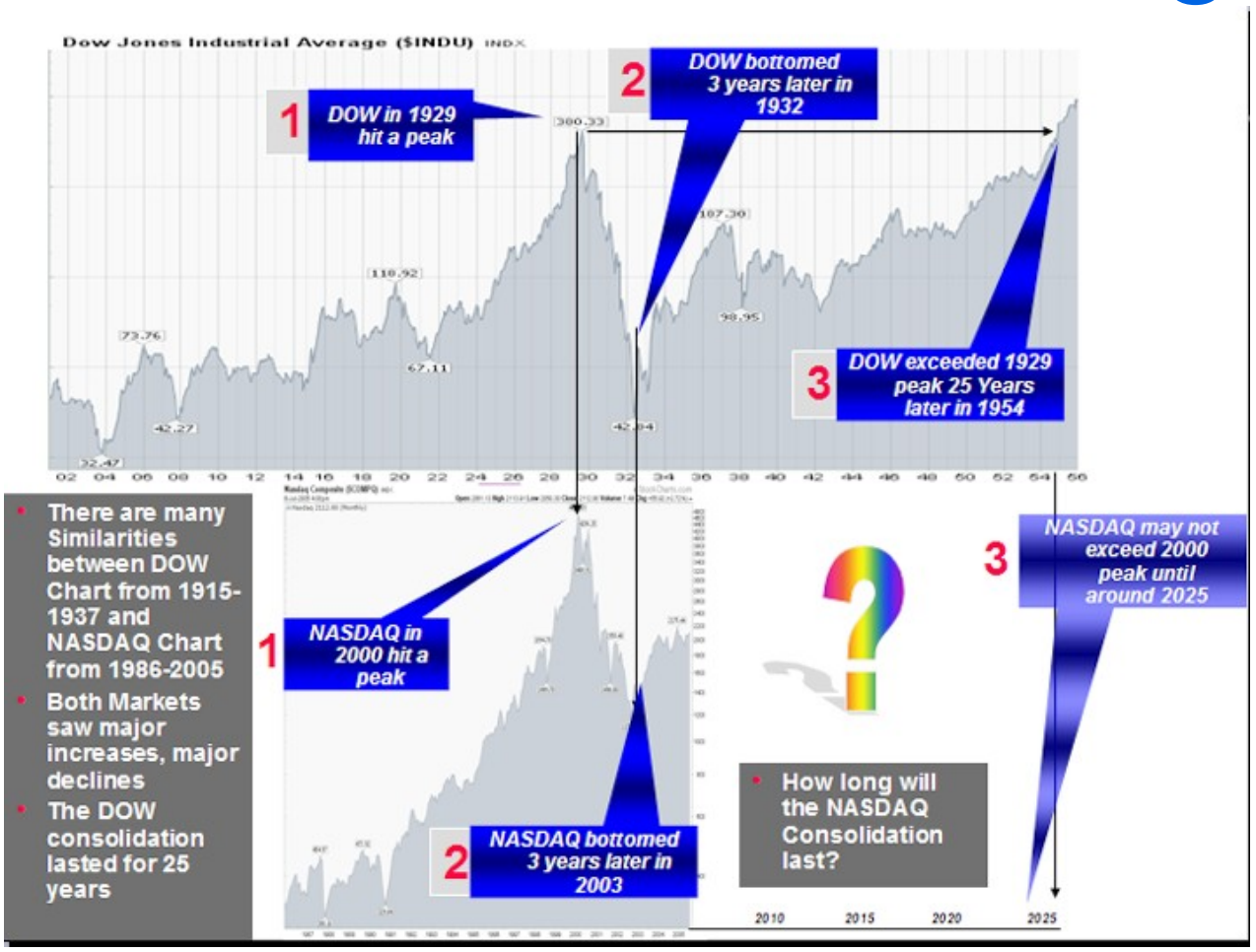
- As robot moves, estimate its pose & world geometry

Human Tumor Microarray Data

- 6830×64 matrix of real numbers.
- Rows correspond to genes, columns to tissue samples.
- Cluster rows (genes) can deduce functions of unknown genes from known genes with similar expression profiles.
- Cluster columns (samples) can identify disease profiles: tissues with similar disease should yield similar expression profiles.

Gene expression matrix

Financial Forecasting



http://www.steadfastinvestor.com/

- Predict future market behavior from historical data, news reports, expert opinions, …

Machine Learning Problems

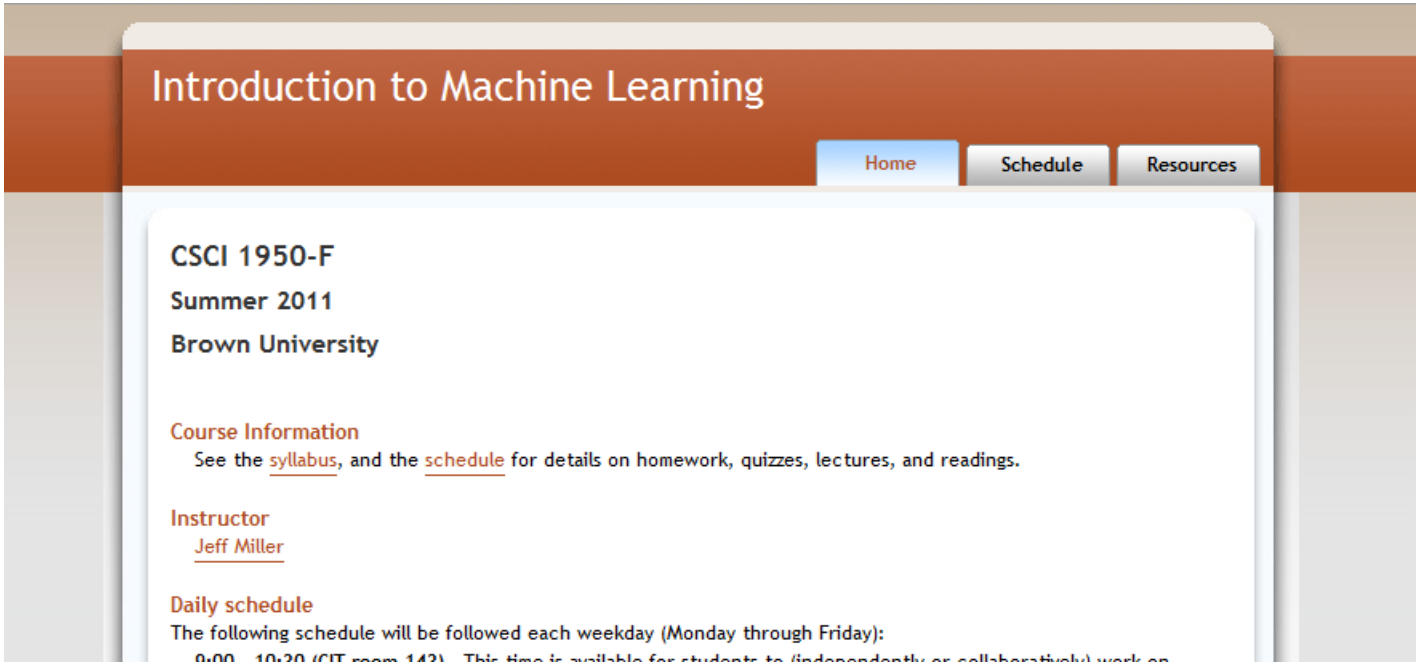
	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Course information

Course website

- Schedule, homeworks, etc.

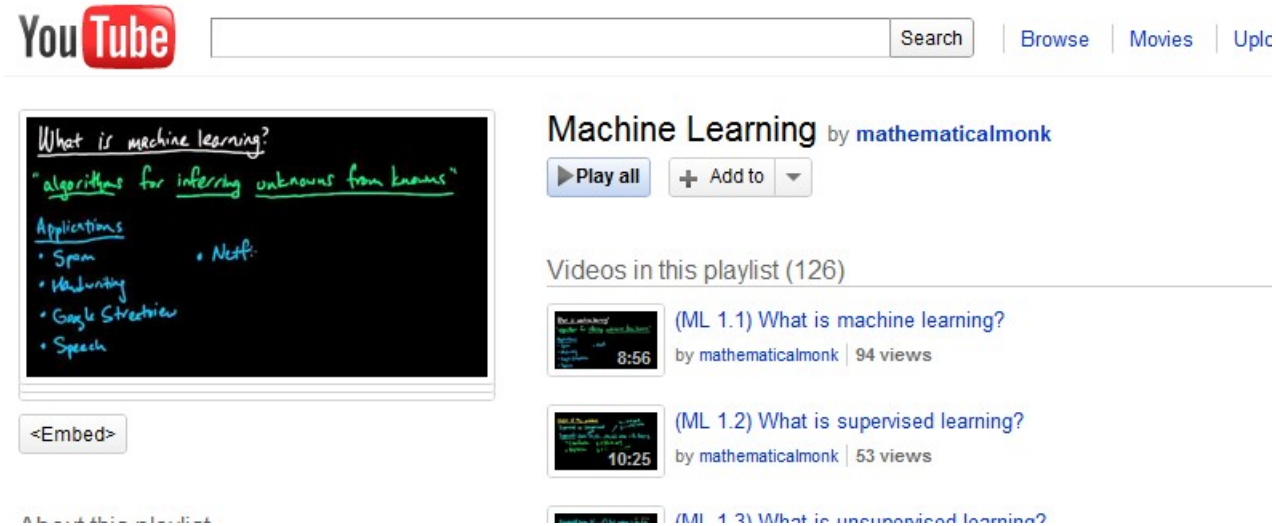
  http://www.dam.brown.edu/people/jmiller/ML/



Videos

- Lectures – intuition and understanding
- Videos – mathematical details
  - ➢ absorb concepts at your own pace
  - ➢ use face-to-face meeting time for discussion
  - ➢ (Note: You will need headphones)

mathematicalmonk on YouTube



Textbook

MACHINE LEARNING:
A PROBABILISTIC PERSPECTIVE

by Kevin Murphy

Available (for purchase) at the Metcalf Copy Center
(downstairs from the coffee shop in the Brown Bookstore)
Ask for the "course pack for CS 1950-F"

Daily schedule

- 9:00 - 10:30 (CIT room 143)
  - homework, videos, and/or reading
- 10:30 - 11:30 (CIT room 345)
  - group meeting for lecture/discussion
  - homework submission, quizzes
- 11:30 - 12:00 (CIT room 345 or 143)
  - Q&A, further discussion

    and/or
  - homework, videos, and/or reading

Grades

- 40% Homeworks (daily)
  - Mathematical exercises
  - Computer implementation of learning algorithms
  - Experimentation with real datasets
  - **Collaboration is encouraged**
- 40% Quizzes (daily)
  - Pencil and paper, focused on basic ideas
- 20% Class participation
  - Randomly selected person will discuss HW

Course Prerequisites

- Prerequisites: comfort with basic
  - Programming: Matlab for assignments
  - Calculus: simple integrals, partial derivatives
  - Linear algebra: matrix factorization, eigenvalues
  - Probability: discrete and continuous

- We will briefly review some Probability

Classification

and

K-nearest neighbor

What is “machine learning”?

Given a collection of examples,

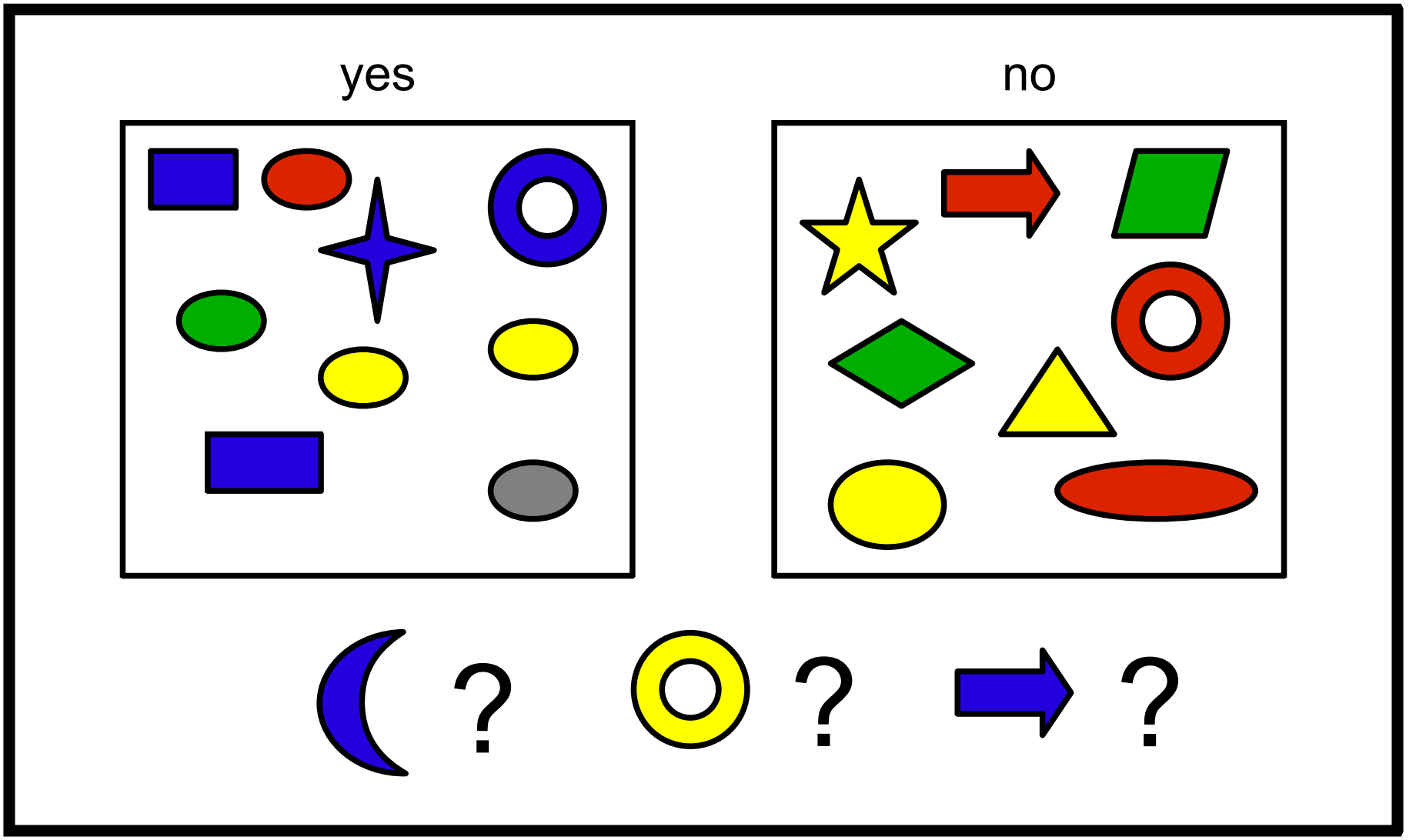
predict something about novel examples.

The examples are often incomplete.

Machine Learning Problems

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

Classification Problems



Classification Encoding

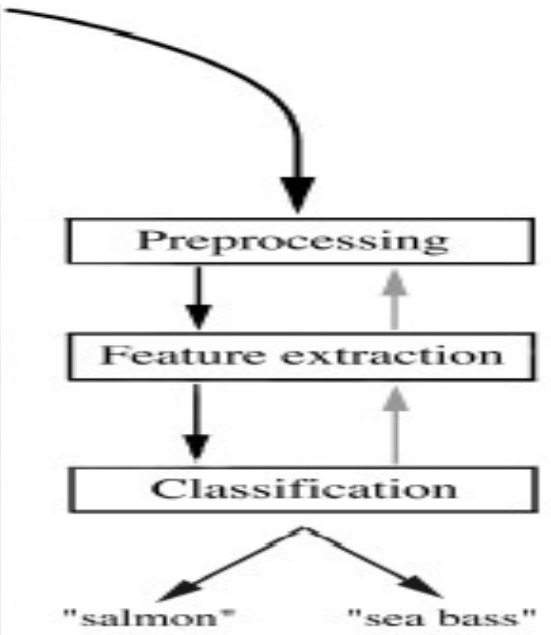
Color	Shape	Size (cm)
Blue	Square	10
Red	Ellipse	2.4
Red	Ellipse	20.7

Binary Label
1
1
0

Example:

Sort fish automatically

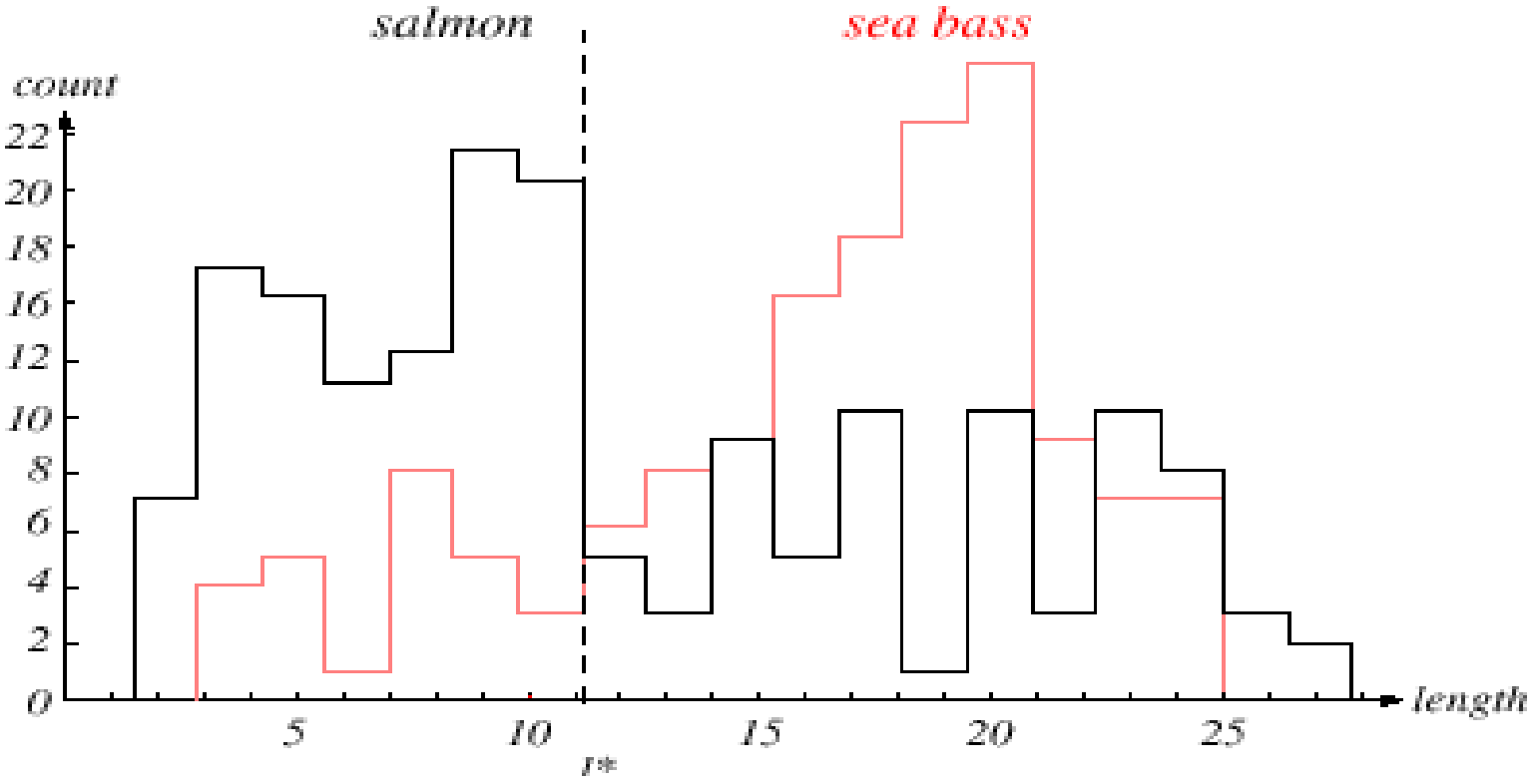
Automatically sorting fish



Sorting fish as a machine learning problem

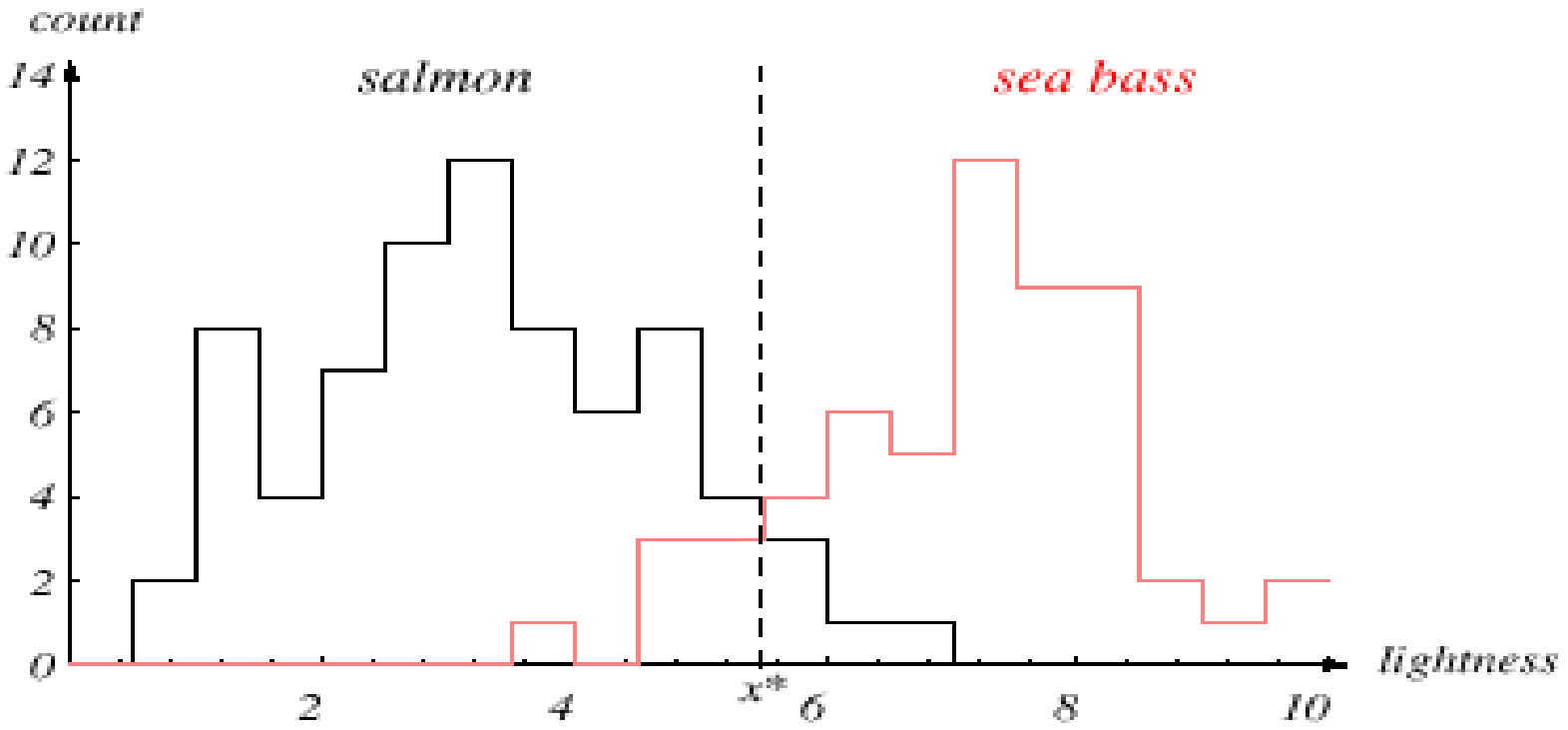
- Training data $D = ((x_1,y_1), \ldots, (x_n,y_n))$
  - A vector of measurements (*features*) x_i (e.g., weight, length, color) of each fish
  - A *label* y_i for each fish
- At run-time:
  - given a novel feature vector x
  - *predict* the corresponding label y

Length as a feature for classifying fish

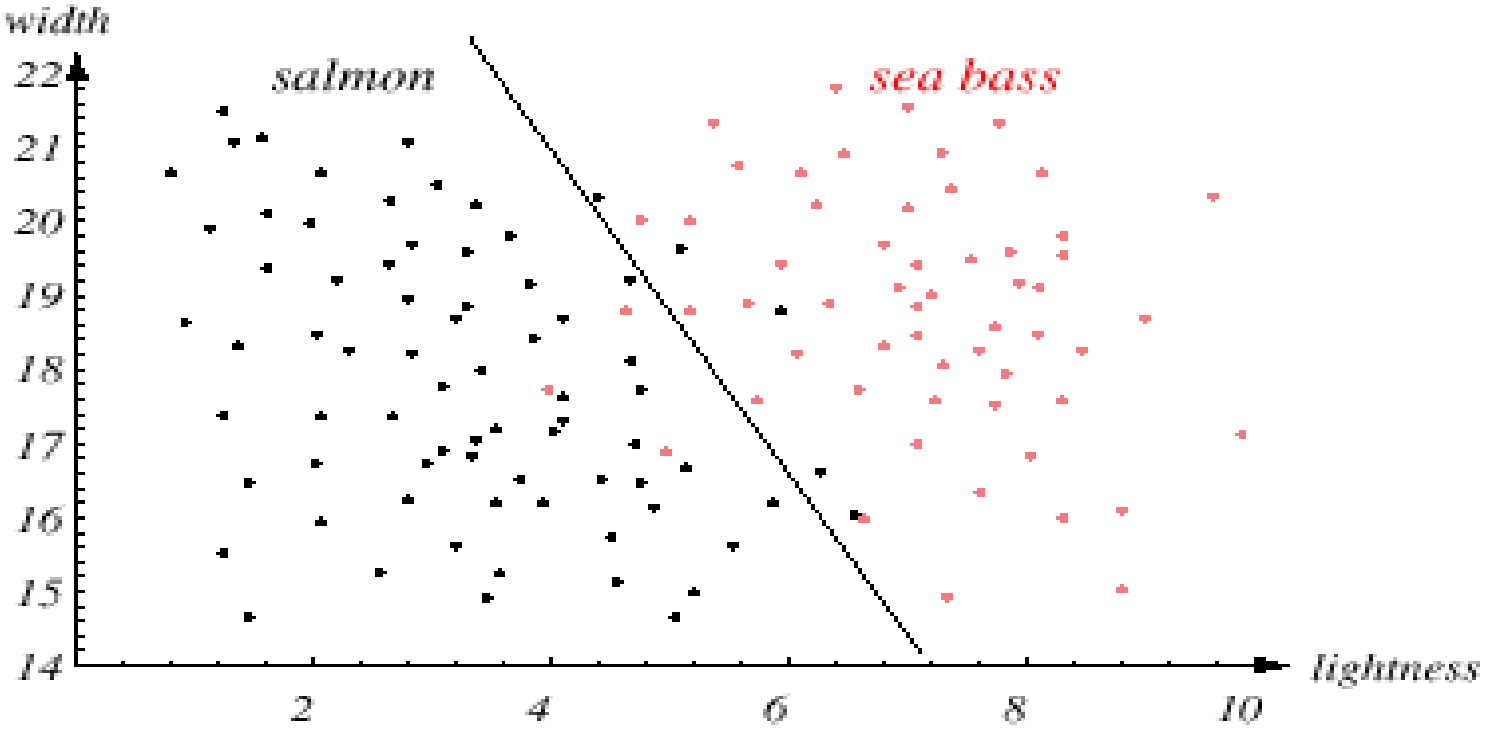


- Need to pick a *decision boundary*
  - Minimize *expected loss*

Lightness as a feature for classifying fish

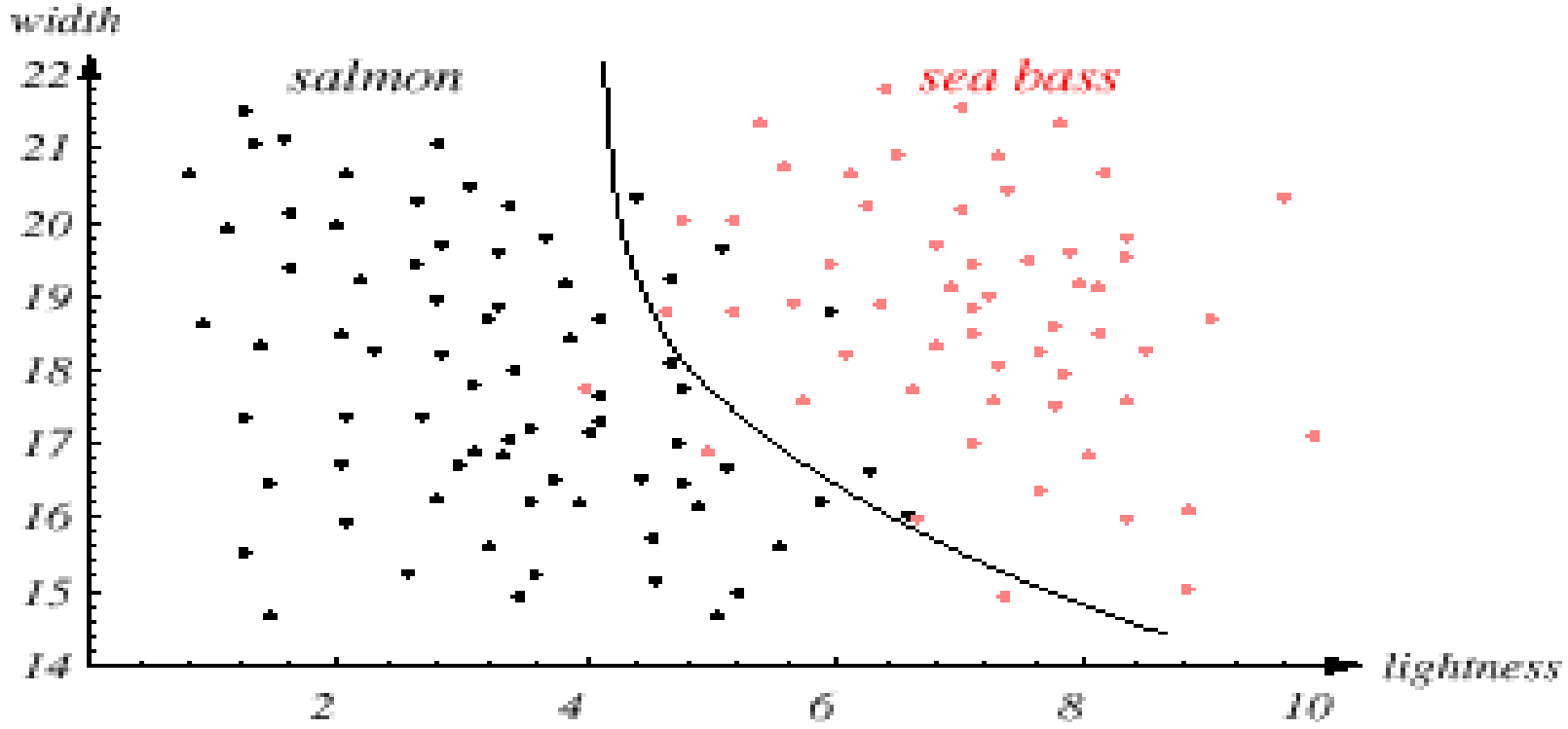


Length and lightness together as features

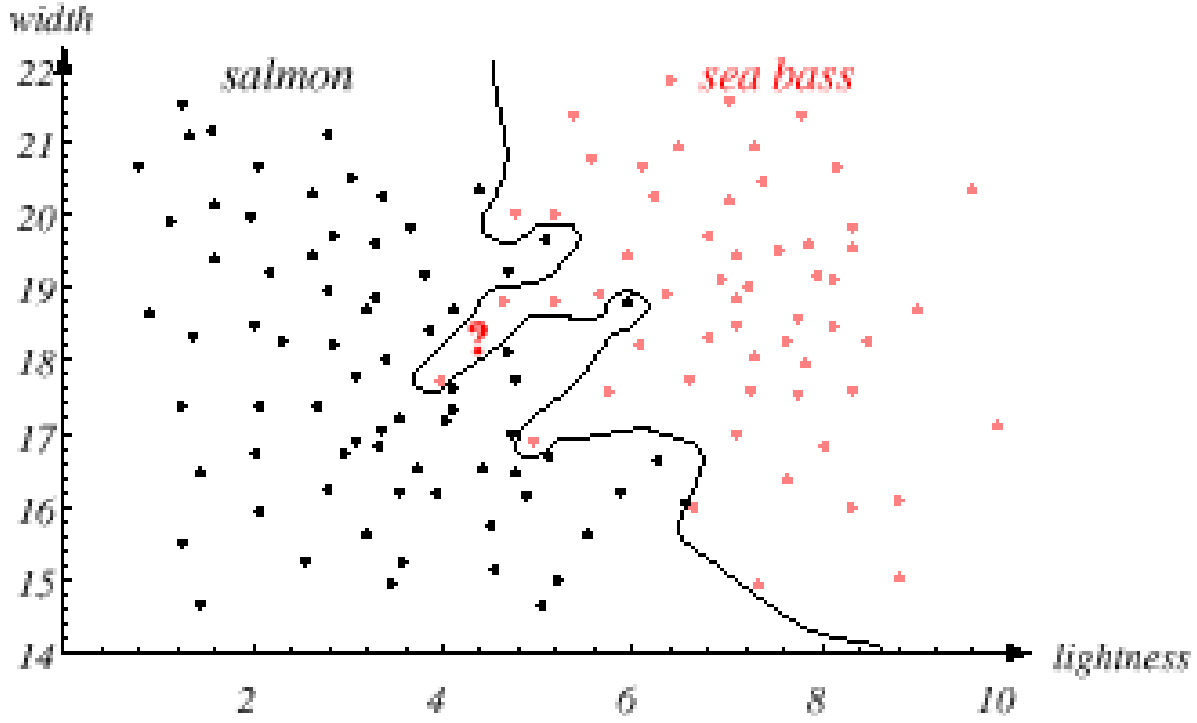


- Not unusual to have millions of features

More complex decision boundaries



Training set error ≠ test set error



- Occam's razor
- Bias-variance dilemma
  - More data!

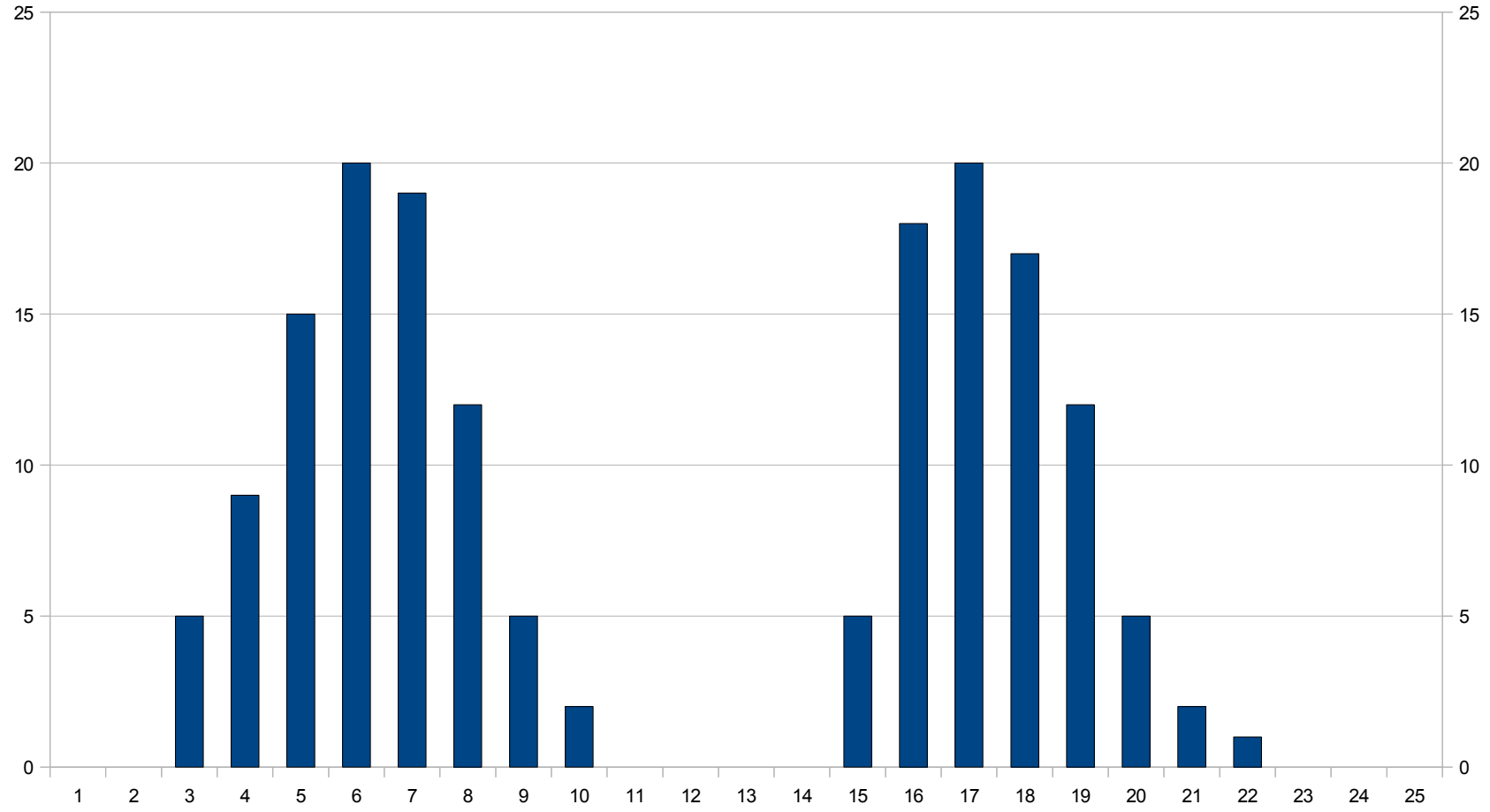
Recap: designing a fish classifier

- Choose the features
  - Can be *the most important step*!
- Collect training data
- Choose the model (e.g., shape of decision boundary)
- Estimate the model from training data
- Use the model to classify new examples
  - Machine learning is about last 3 steps

Supervised versus unsupervised learning

- Supervised learning
  - Training data includes labels we must predict: labels are *visible variables* in training data
- Unsupervised learning
  - Training data does not include labels: labels are *hidden variables* in training data
- For classification models, unsupervised learning usually becomes a kind of *clustering*

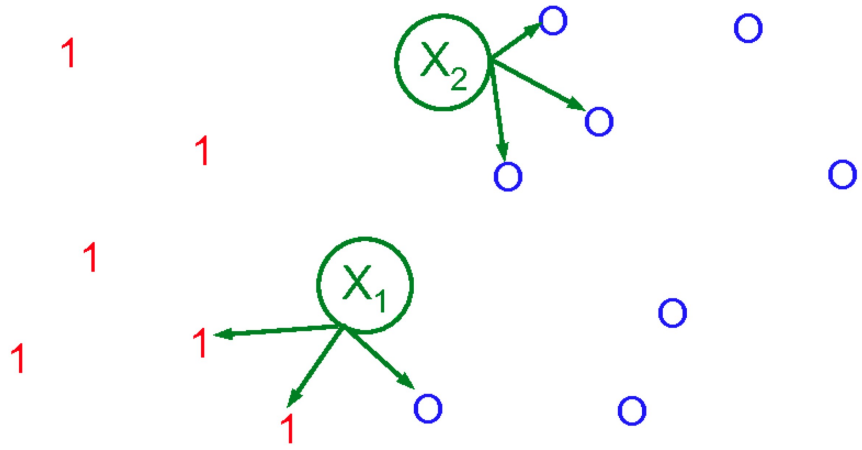
Unsupervised learning for classifying fish

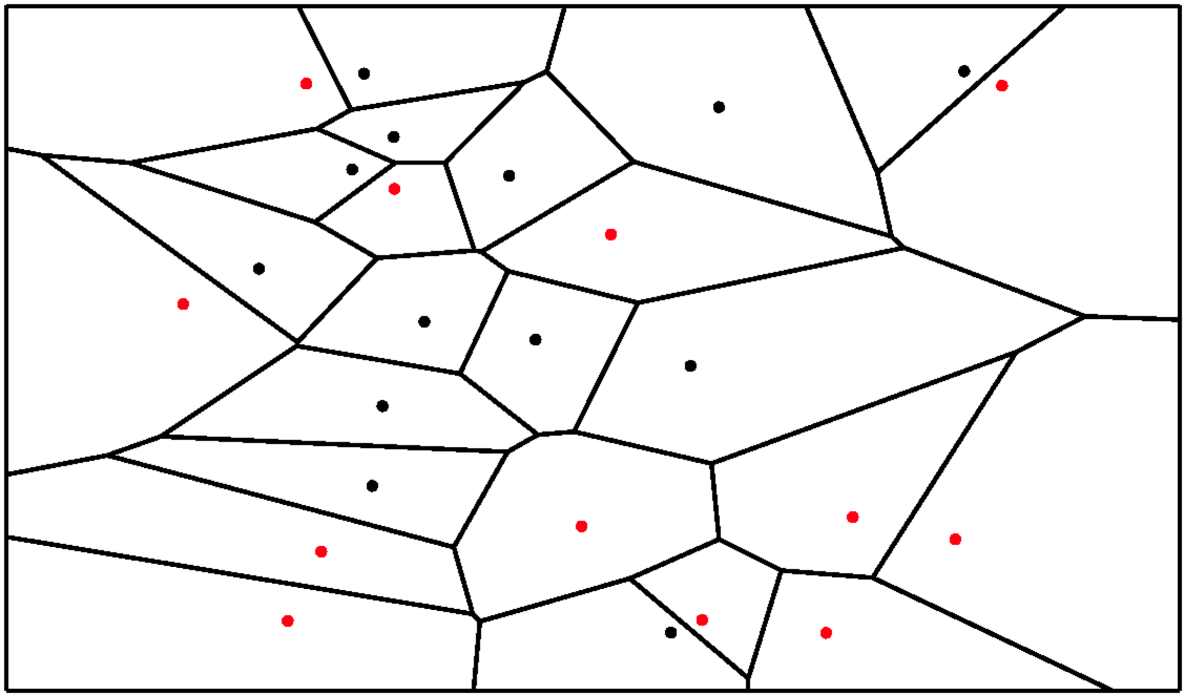


Salmon versus Sea Bass?

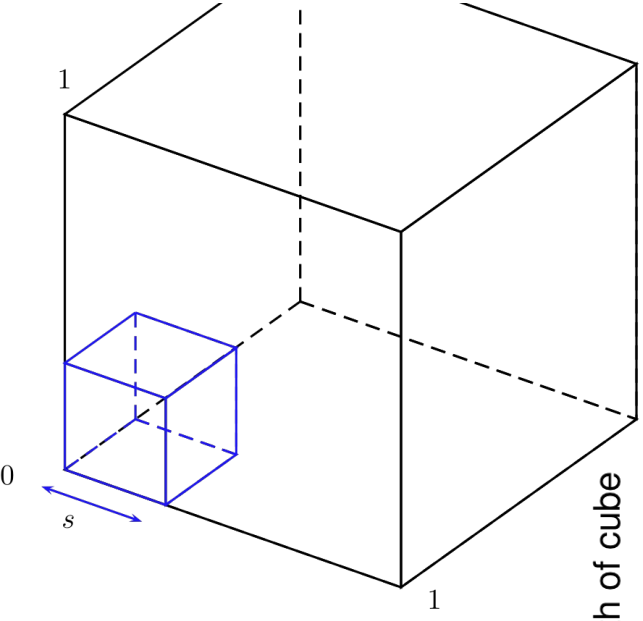
Adults versus juveniles?

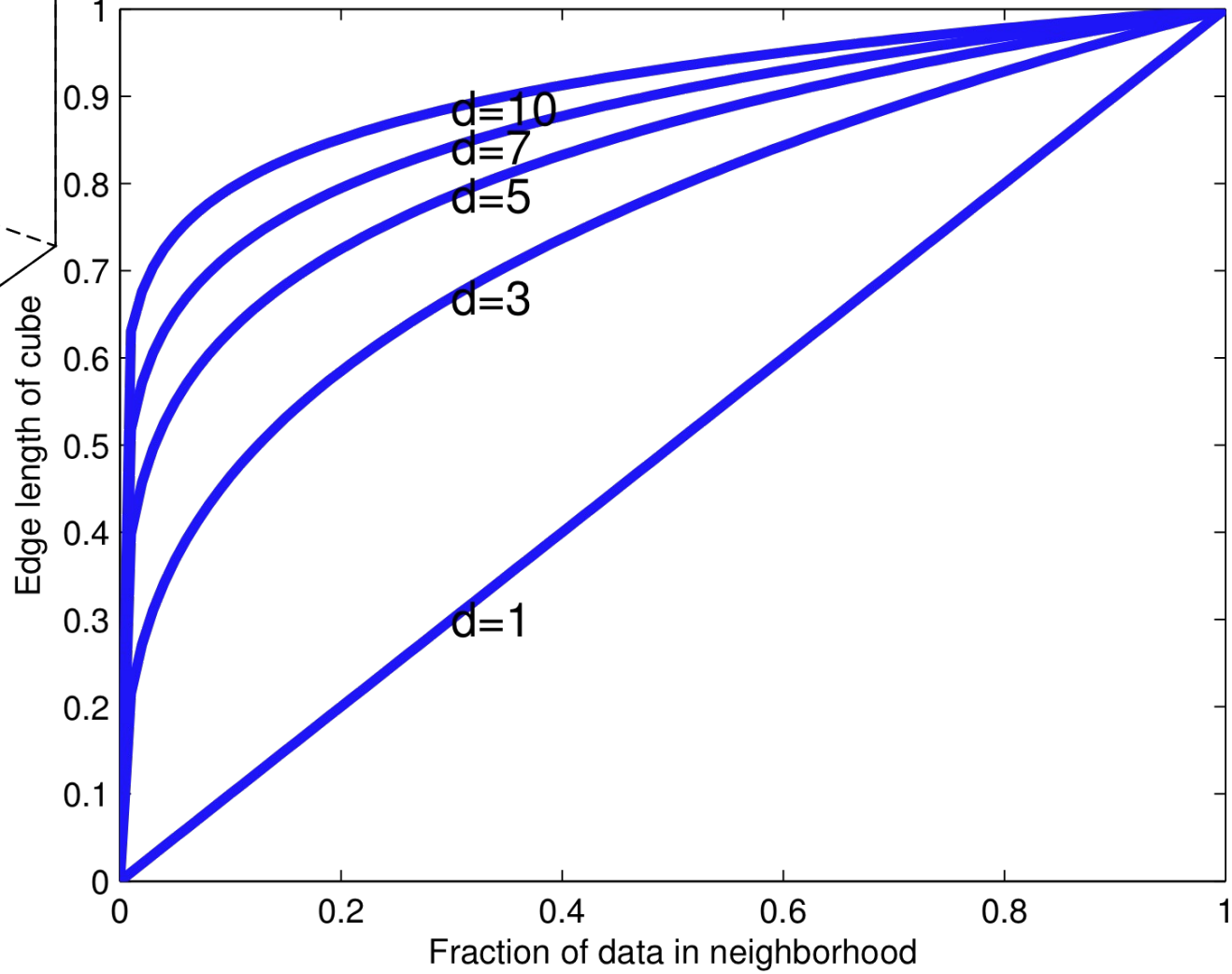
1-Nearest Neighbor





Curse of Dimensionality





What to do next

- Get a computer account (from me)

- Get headphones (if you don't have them)
- Get the textbook
- Visit the course website
- Watch videos (PP 1.S and 2.1-2.5) (~1 hour)

- Note: the 1st homework is due on Wednesday
- Note: the 1st quiz will be on Thursday