Introduction to Machine Learning

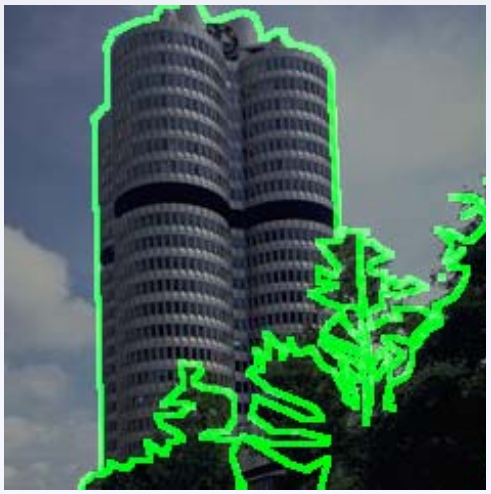
CSCI 1950-F

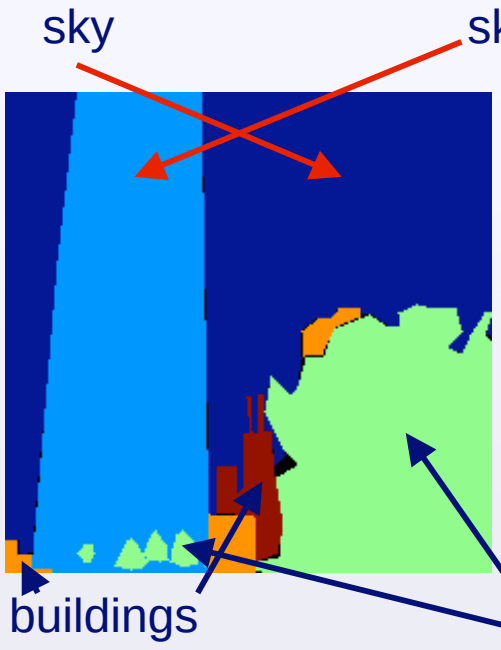
Instructors: Erik Sudderth & Mark Johnson

Graduate TA: Deqing Sun

Undergraduate TAs: Max Barrows & Evan Donahue

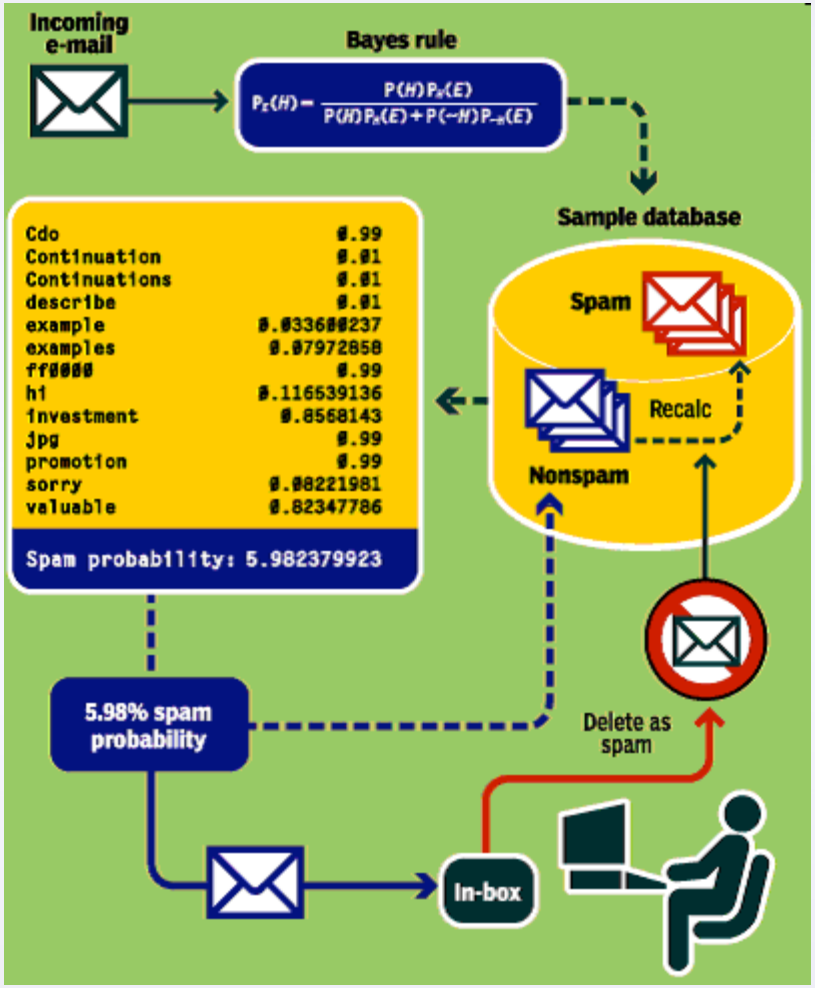
Visual Object Recognition

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

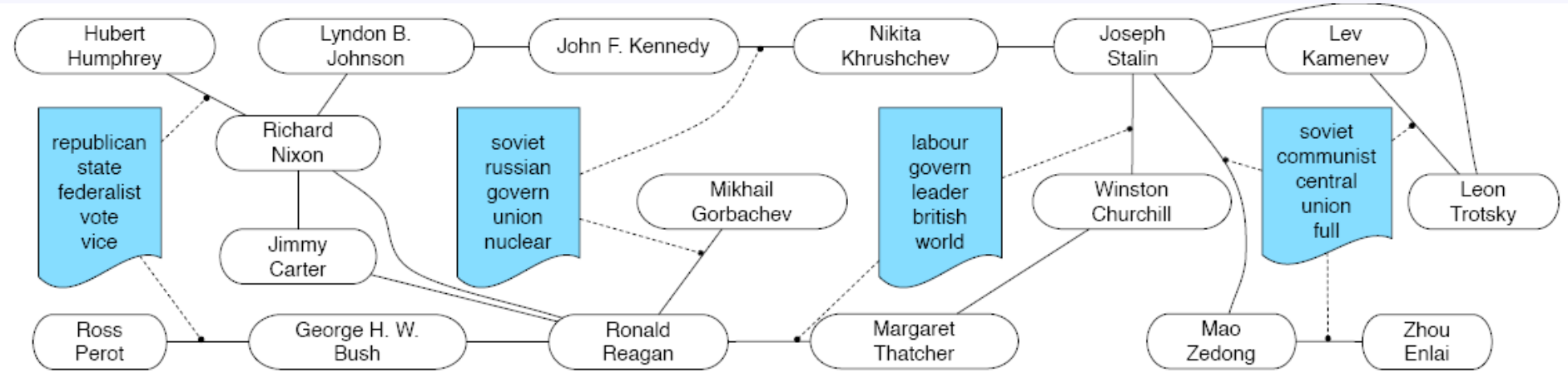
Leaderboard

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				

Social Network Analysis

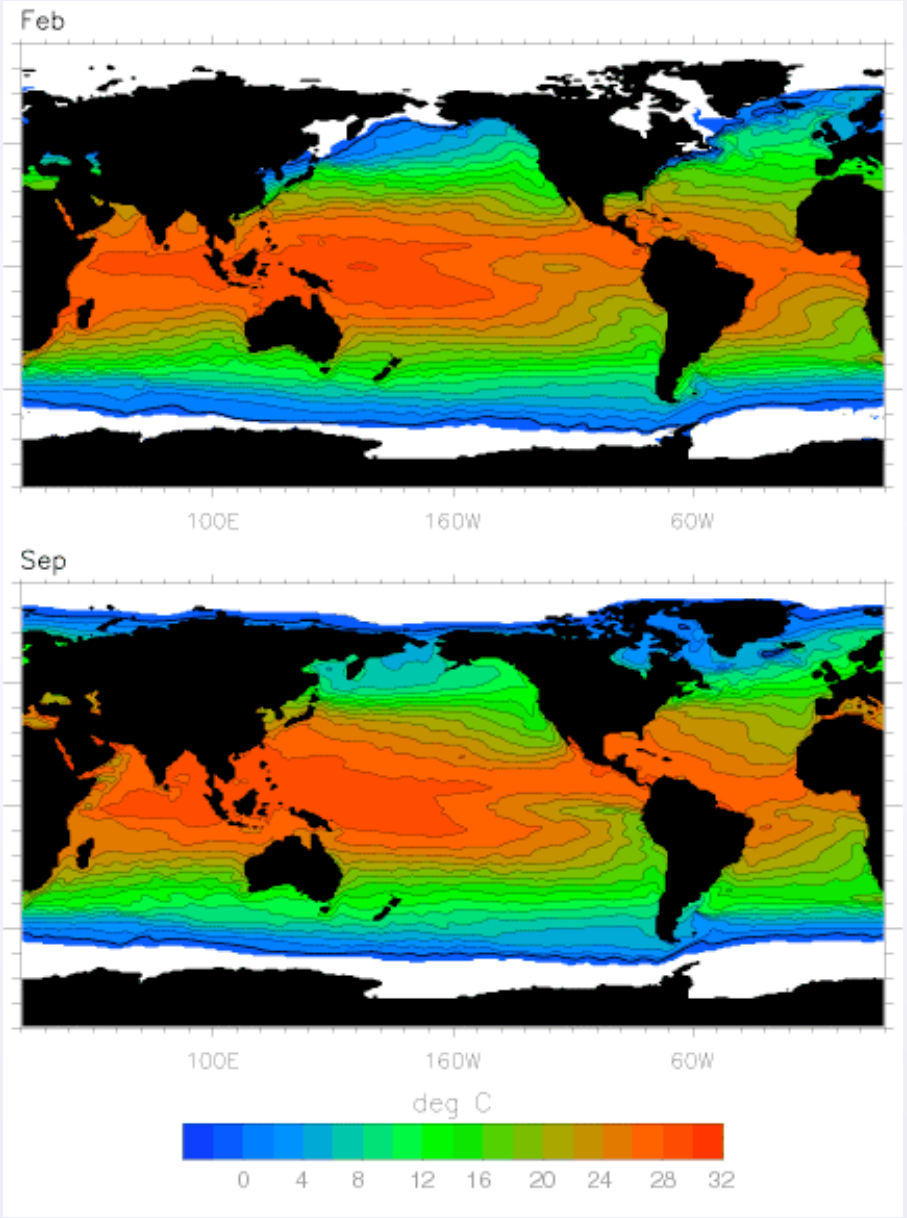
- Unsupervised discovery and visualization of relationships among people, companies, etc.
- Example: infer relationships among named entities directly from Wikipedia entries



Chang, Boyd-Graber, & Blei, KDD 2009

Climate Modeling

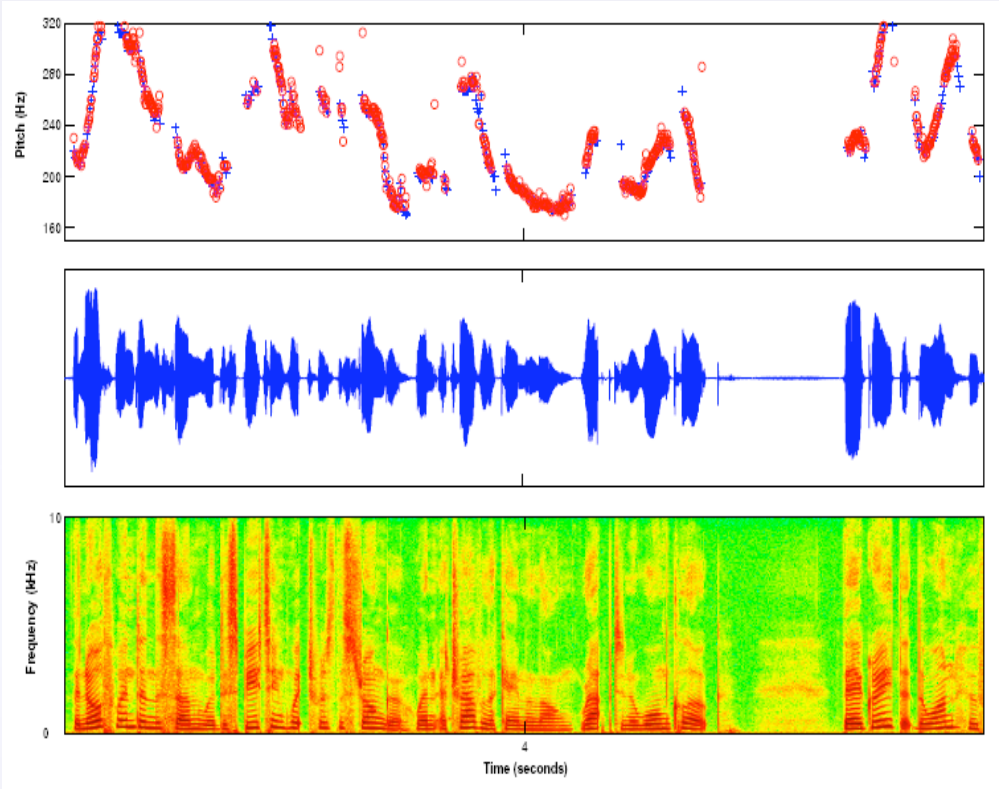
- Satellites measure sea-surface temperature at sparse locations
  - Partial coverage of ocean surface
  - Sometimes obscured by clouds, weather
- Would like to infer a dense temperature field, and track its evolution

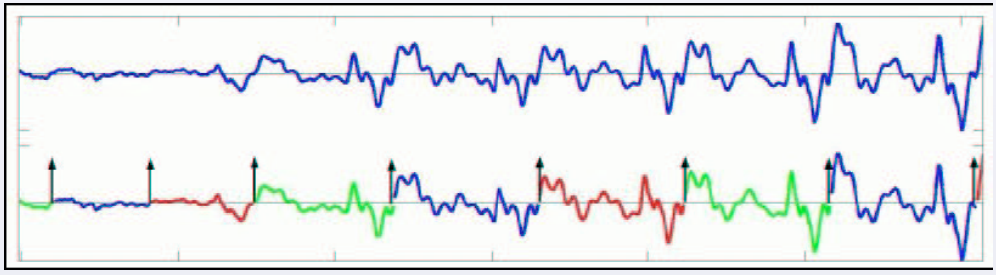


NASA Seasonal to Interannual Prediction Project
http://ct.gsfc.nasa.gov/annual.reports/ess98/nsipp.html

Speech Recognition

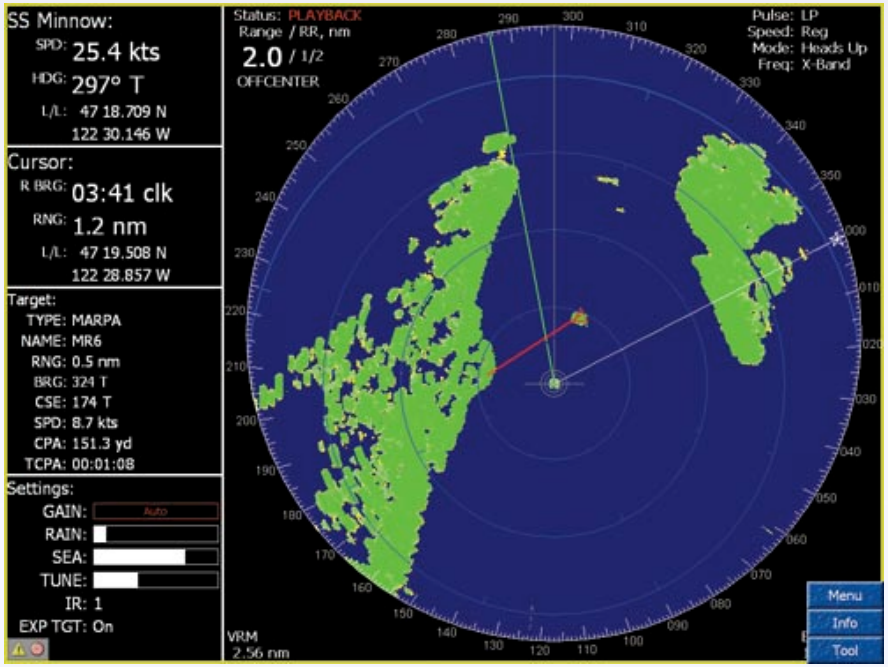
- Given an audio waveform, robustly extract & recognize any spoken words
- Statistical models can be used to
  - ➢ Provide greater robustness to noise
  - ➢ Adapt to accent of different speakers
  - ➢ Learn from training



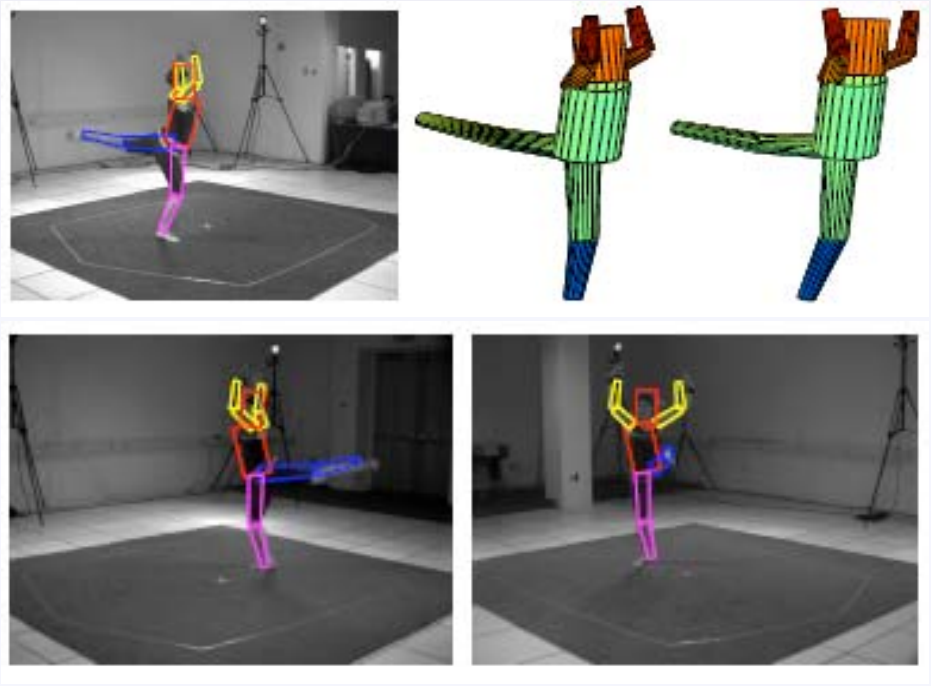


S. Roweis, 2004

Target Tracking



Radar-based tracking of multiple targets



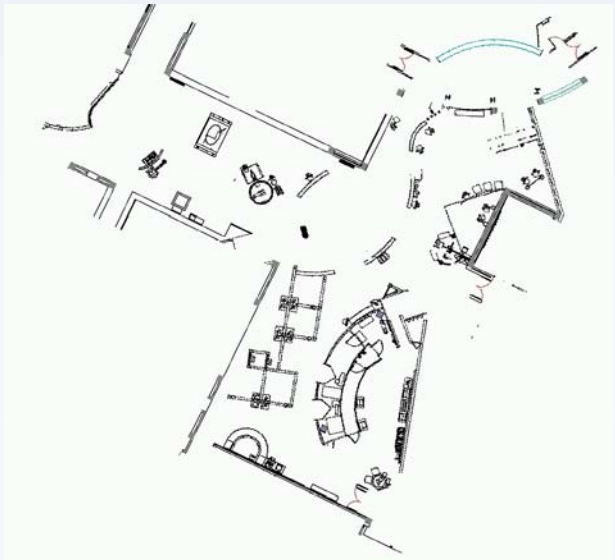
Visual tracking of articulated objects

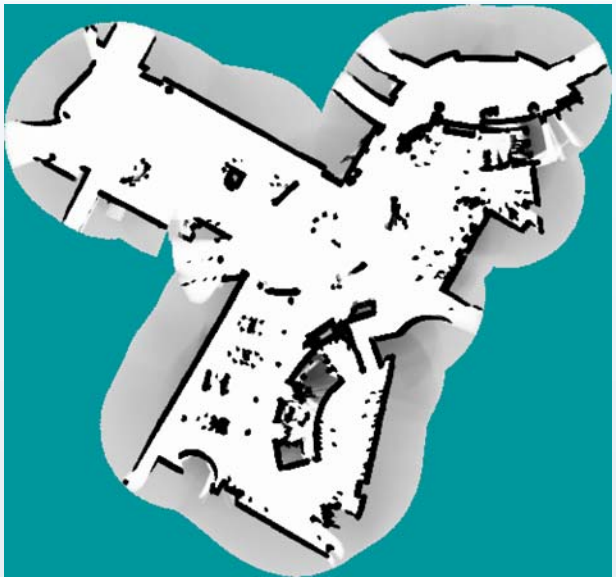
(L. Sigal et. al., 2006)

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

Robot Navigation: *SLAM*

Simultaneous Localization and Mapping



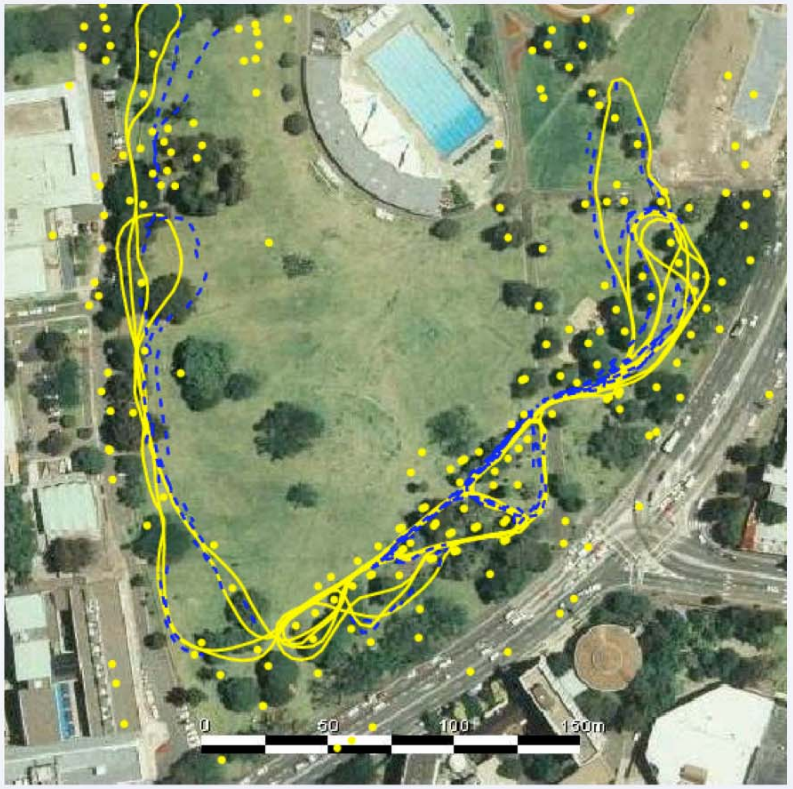


CAD Map

(S. Thrun, San Jose Tech Museum)

Estimated Map

Landmark SLAM
(E. Nebot, Victoria Park)

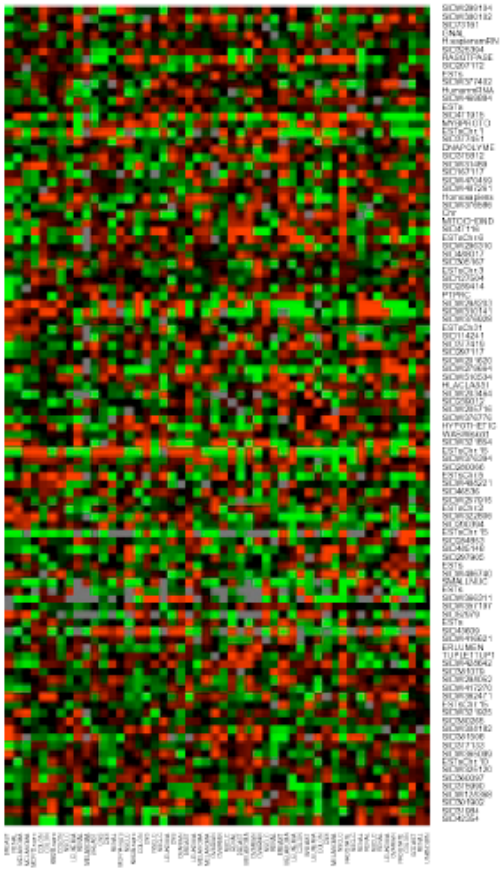


- As robot moves, estimate its pose & world geometry

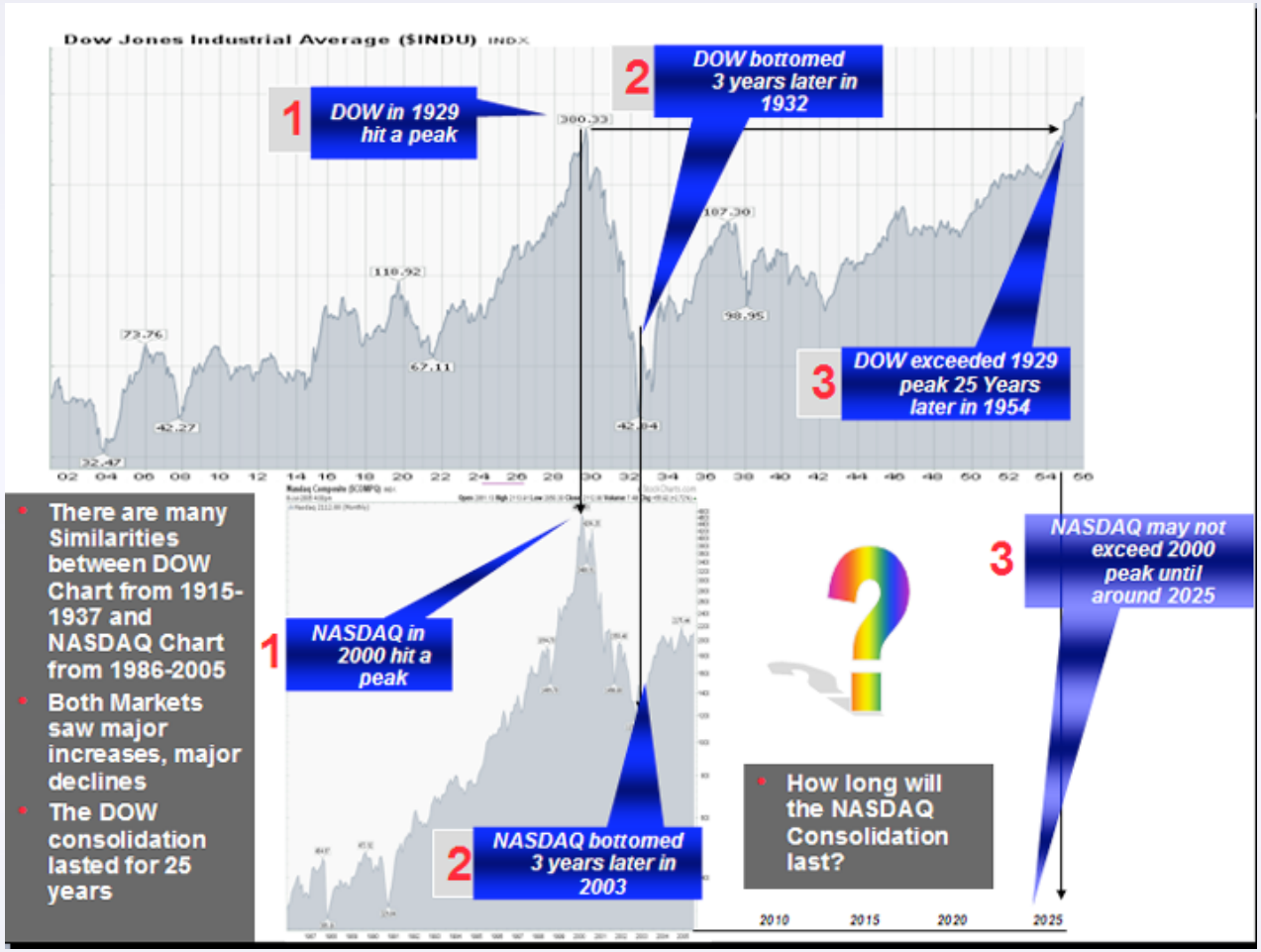
Human Tumor Microarray Data

- 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

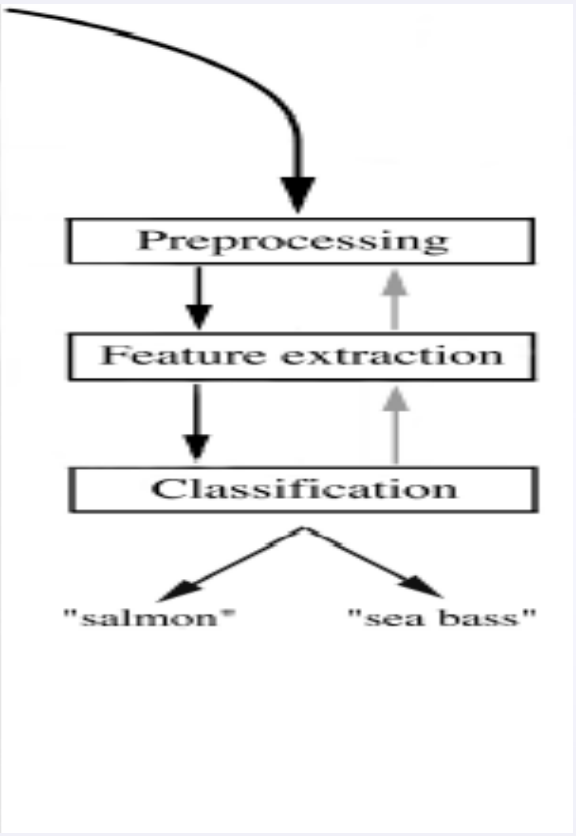
Administrative Details

- Prerequisites: comfort with basic
  - Programming
  - Calculus
  - Linear algebra
  - Probability
- Grading: undergraduate versus graduate
- Syllabus: subject to revision!

What is “machine learning”?

- Given a collection of examples (“training data”),
  predict something about novel examples
  - The novel examples are usually *incomplete*
- Example: sorting fish
  - Fish come off a conveyor belt in a fish factory
  - Your job: figure out what kind each fish is

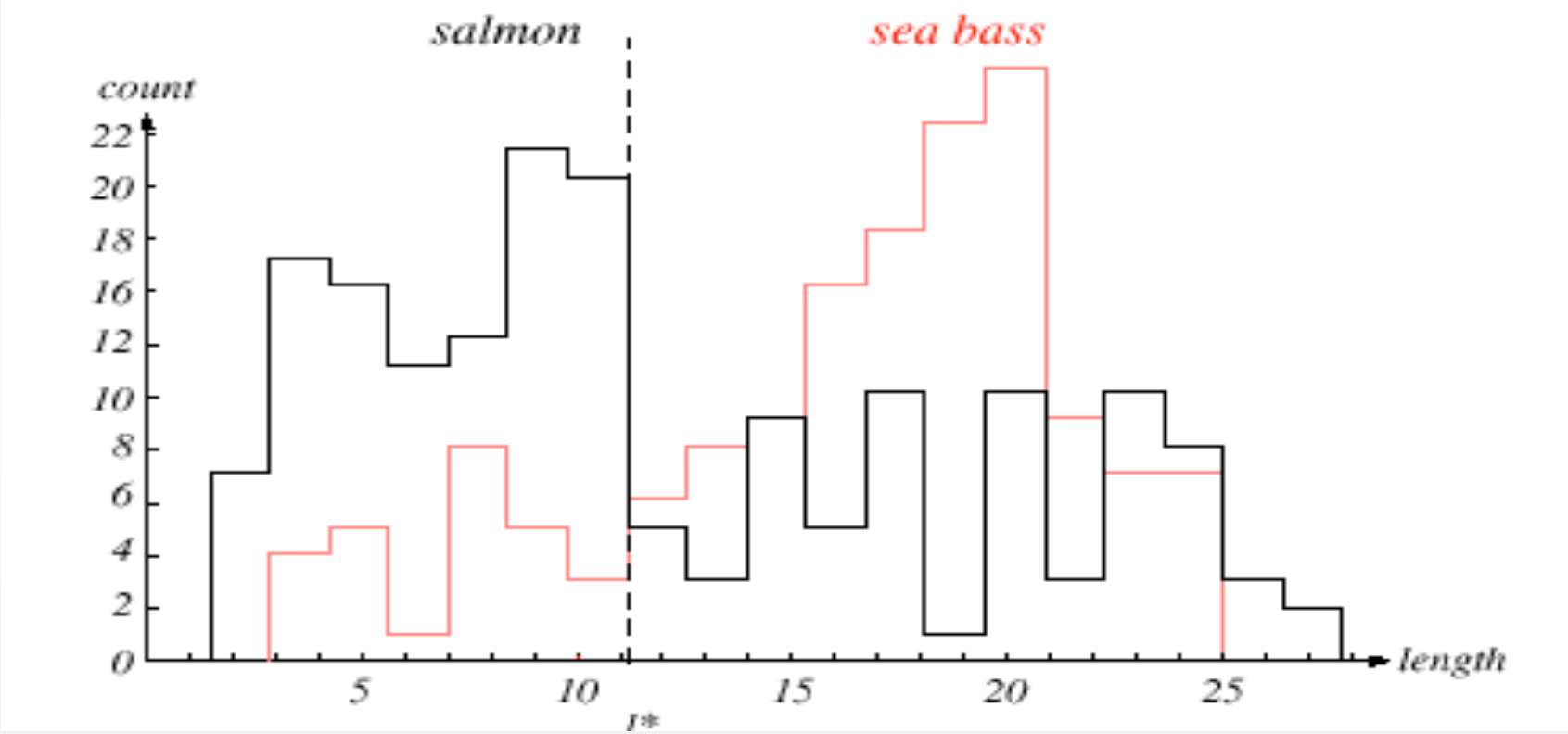
Automatically sorting fish



Sorting fish as a machine-learning problem

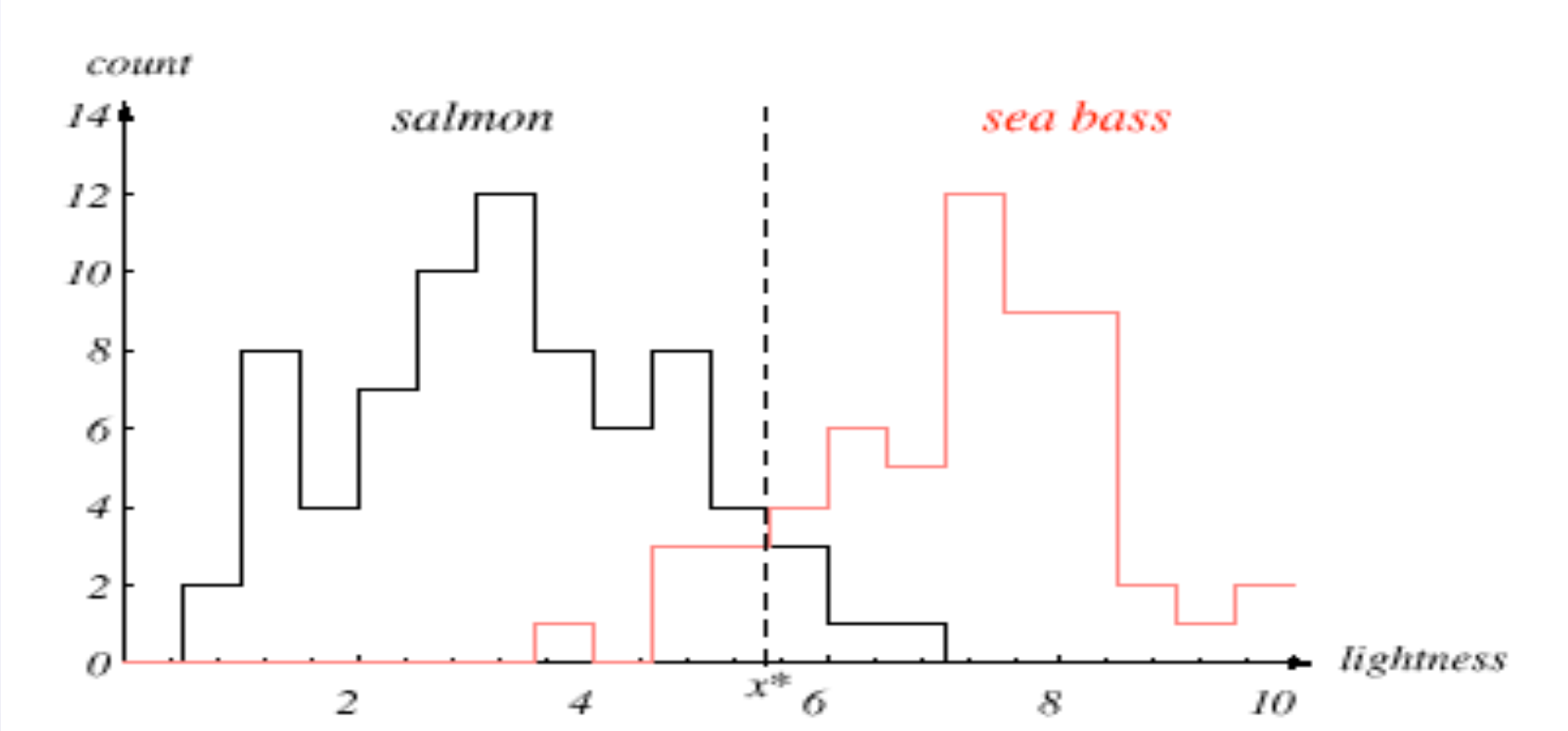
- Training data $D = ((x_1,y_1), ..., (x_n,y_n))$
  - A vector of measurements (*features*) x_i (e.g., weight, length, color) of each fish
  - A *label* y_i for each fish
- At run-time:
  - given a novel feature vector x
  - *predict* the corresponding label y

Length as a feature for classifying fish

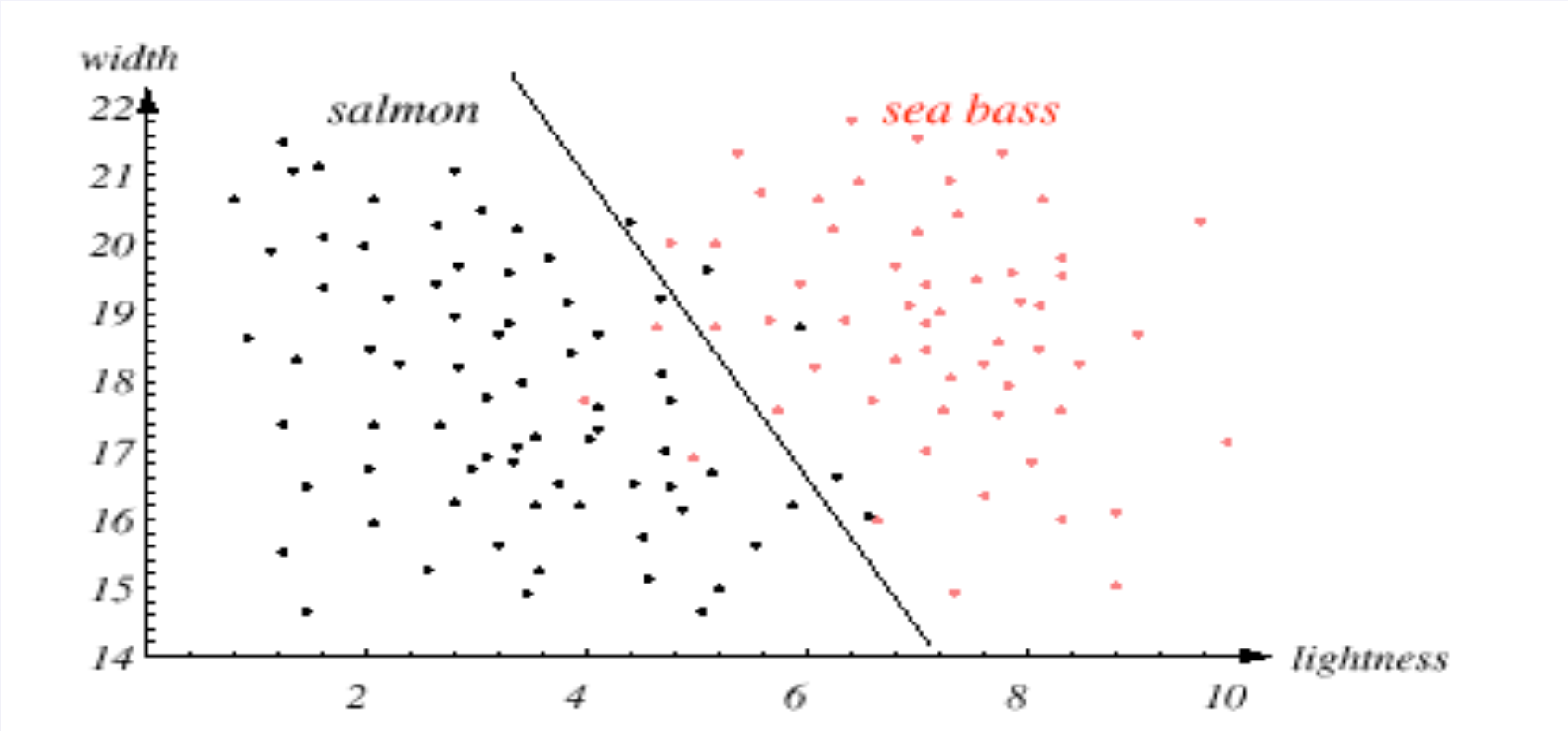


- Need to pick a *decision boundary*
  - Minimize *expected loss*

Lightness as a feature for classifying fish

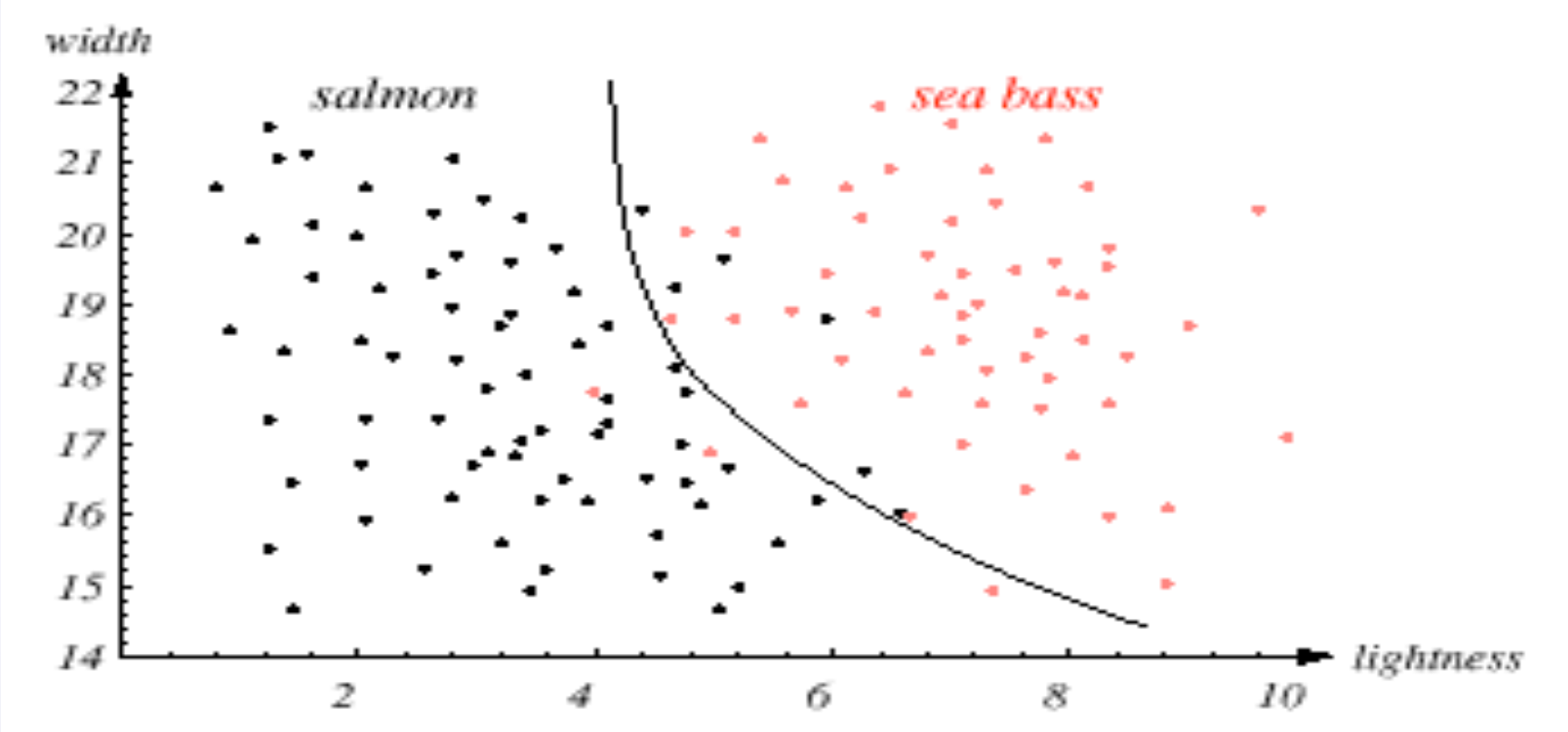


Length and lightness together as features

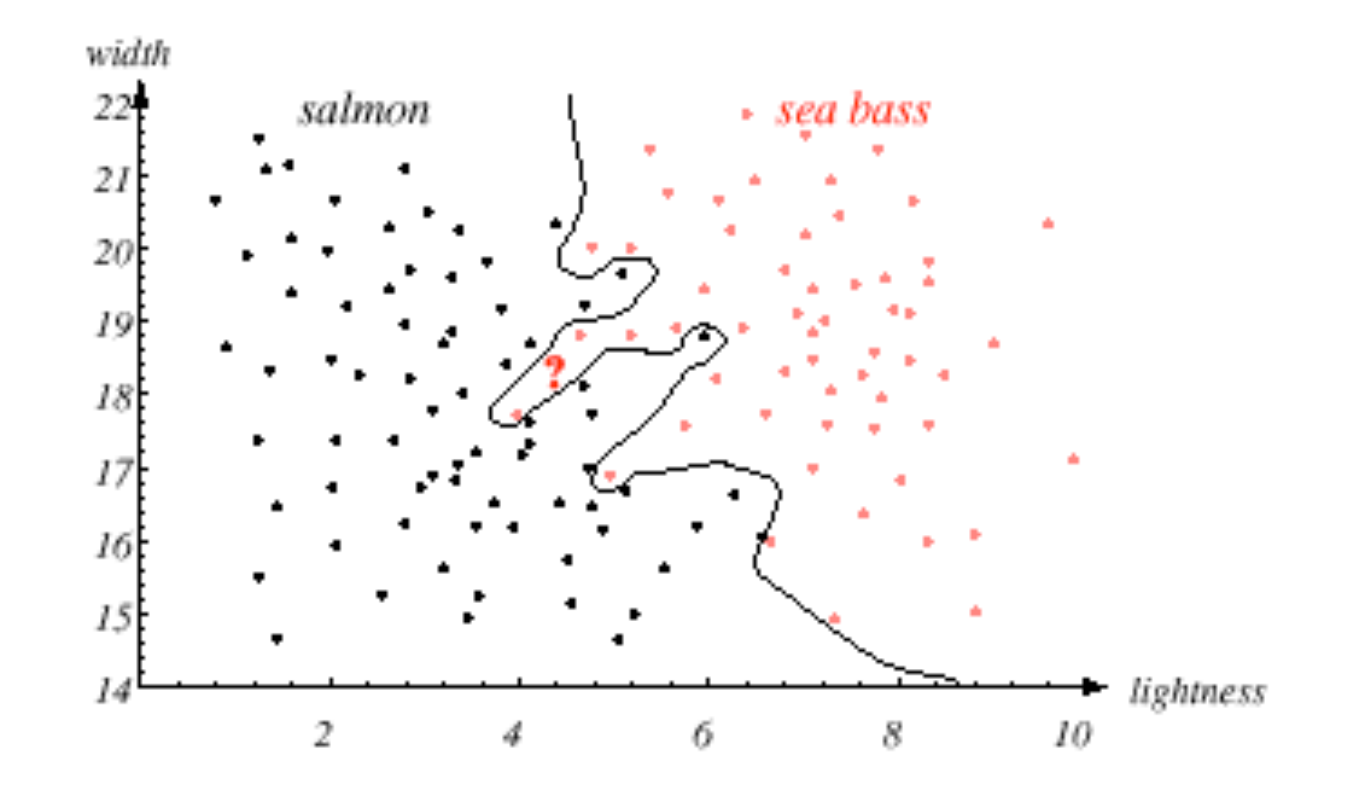


- Not unusual to have millions of features

More complex decision boundaries



Training set error ≠ test set error



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

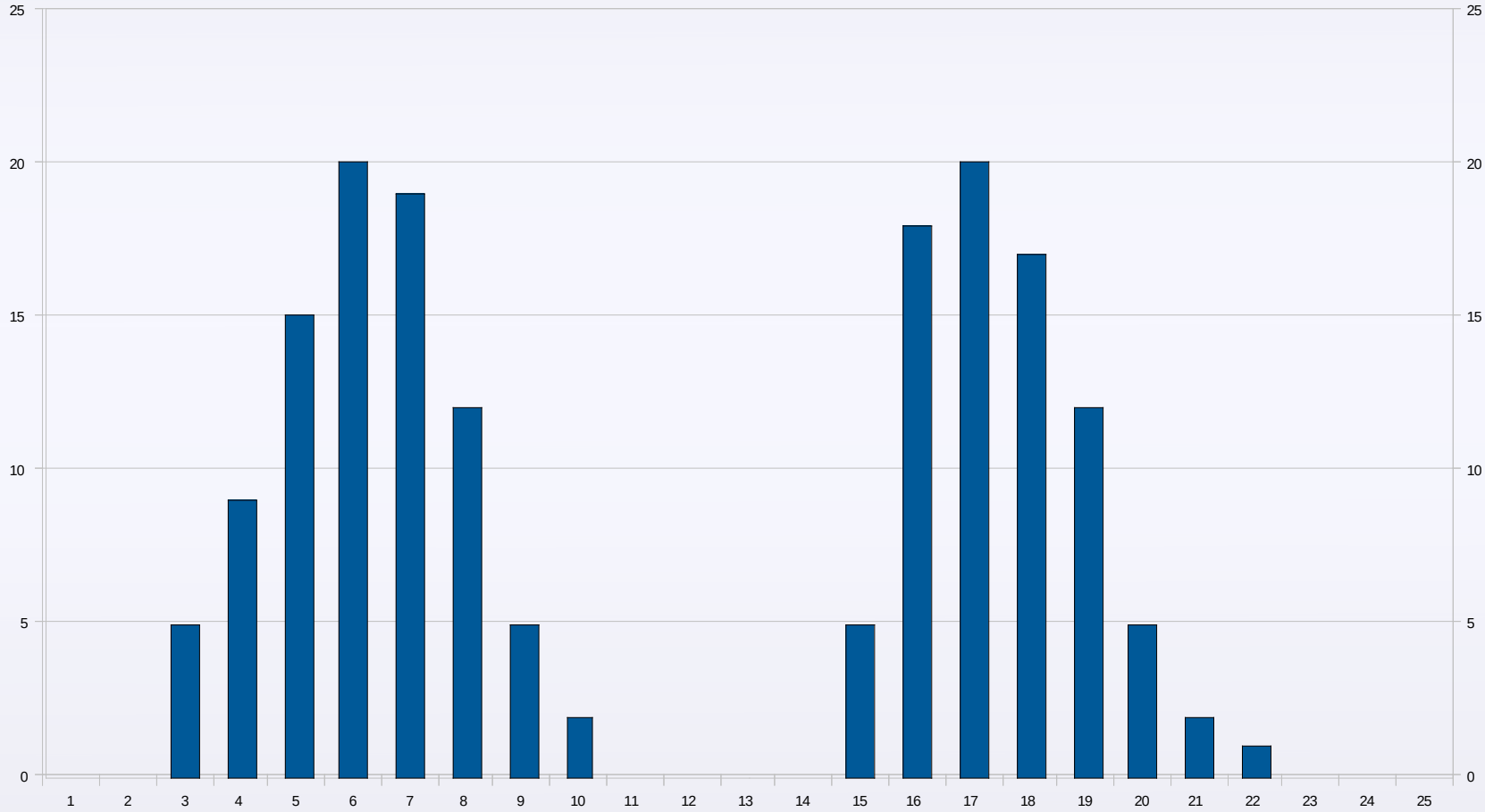
Recap: designing a fish classifier

- Choose the features
  - Usually *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 have to predict
    i.e., labels are *visible variables* in training data
- Unsupervised learning
  - Training data does not include labels
    i.e., labels are *hidden variables* in training data
- For classification problems, unsupervised learning is usually a kind of clustering

Unsupervised learning for classifying fish



Machine Learning Problems

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

Machine Learning Buzzwords

- Bayesian and frequentist estimation
- Model selection, cross-validation, overfitting
- Kernel methods: support vector machines (SVMs), Gaussian processes
- Graphical models: hidden Markov models, Markov random fields, belief propagation
- Expectation-Maximization (EM) algorithm
- Markov chain Monte Carlo (MCMC) methods