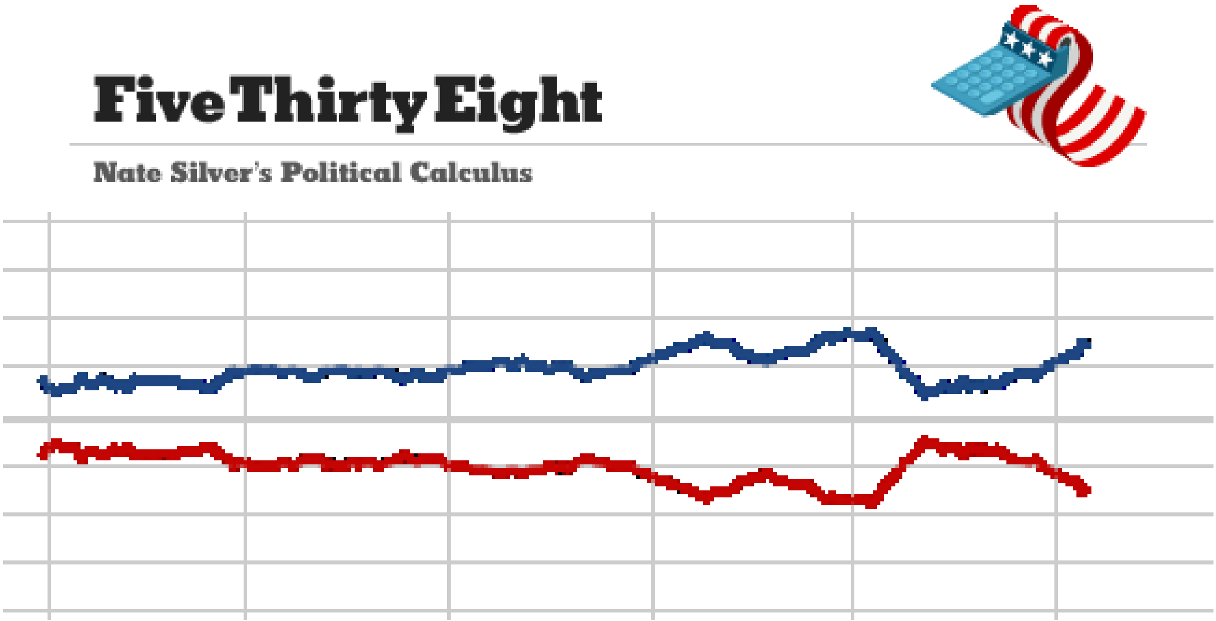
Stat315a: Statistical Learning

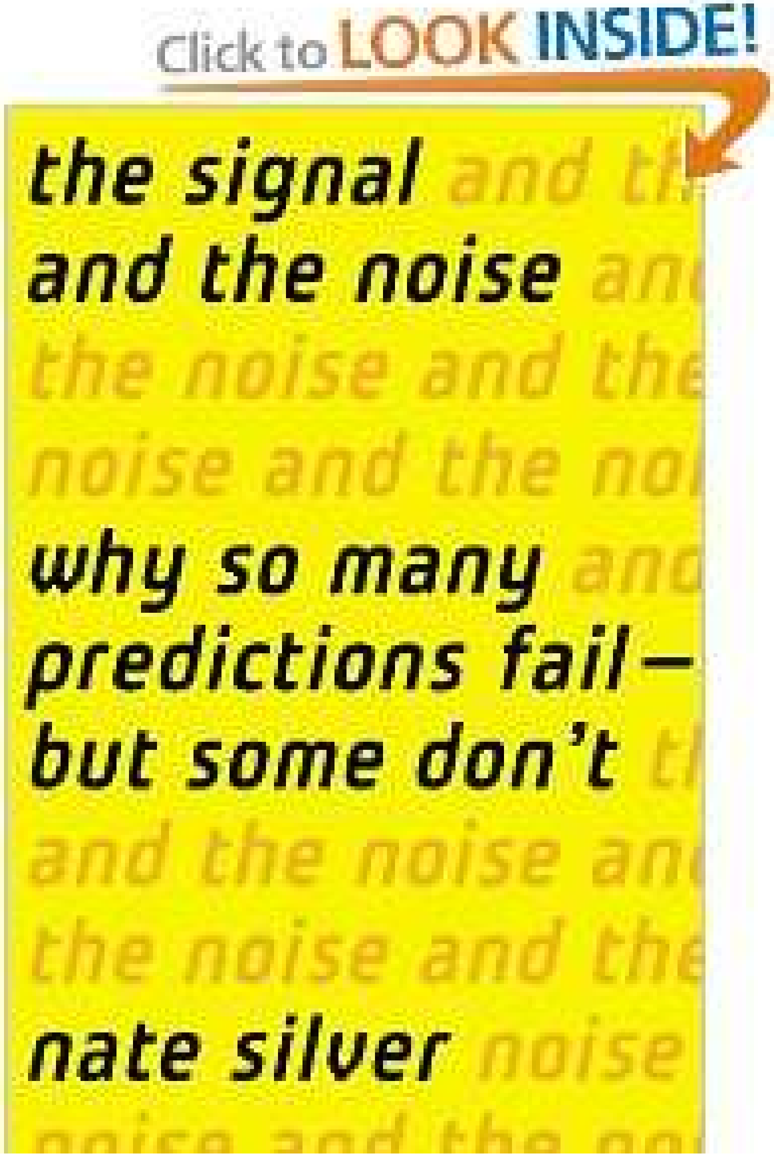
Statistics in the news

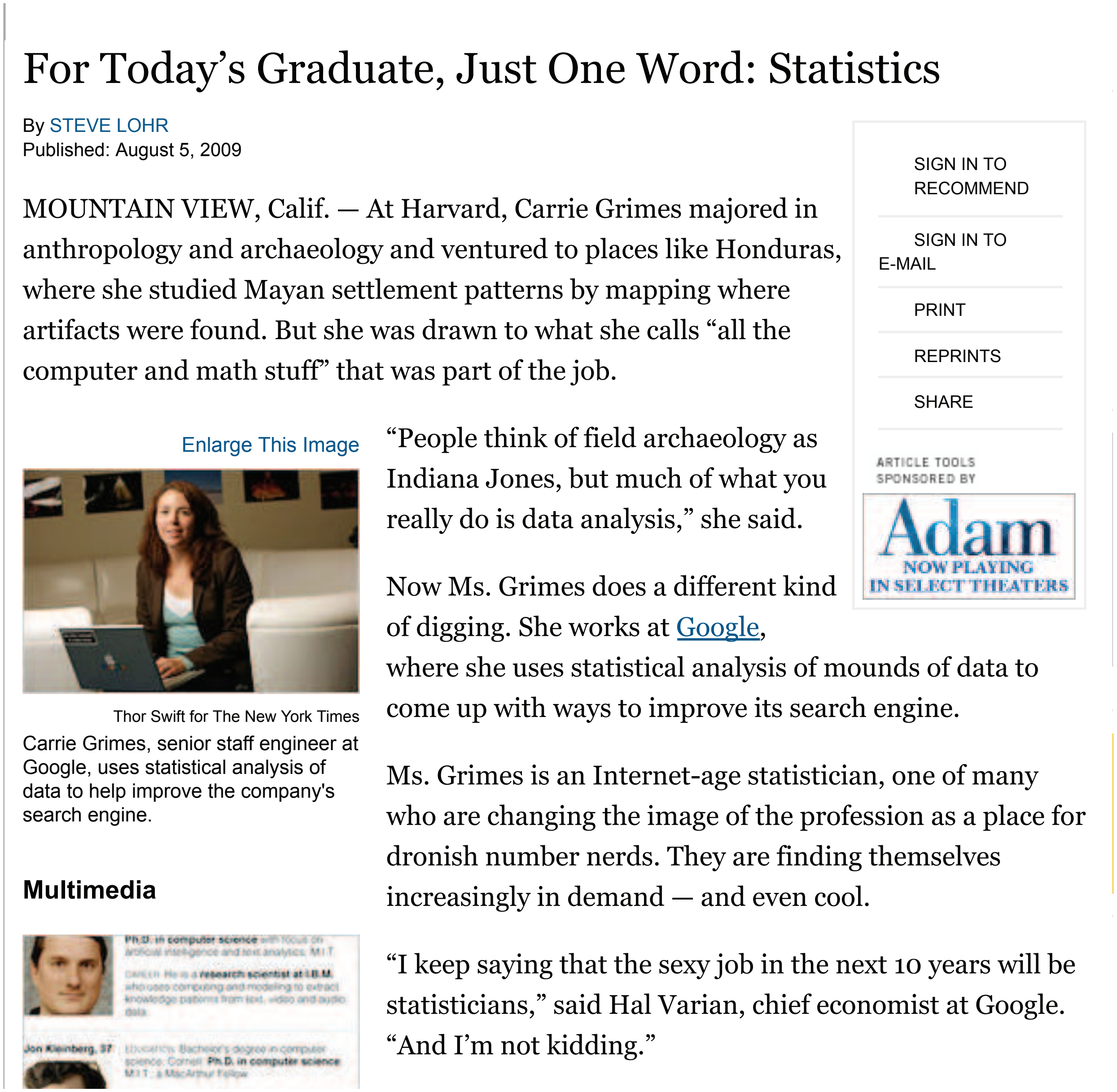
How IBM built Watson, its “Jeopardy”-playing supercomputer by Dawn Kawamoto DailyFinance 02/08/2011

Learning from its mistakes According to David Ferrucci (PI of Watson DeepQA technology for IBM Research), Watson’s software is wired for more that handling natural language processing.

“Its machine learning allows the computer to become smarter as it tries to answer questions — and to learn as it gets them right or wrong.”


FiveThirtyEight
Nate Silver's Political Calculus


Click to LOOK INSIDE!
the signal
and the noise
why so many
predictions fail—
but some don't
nate silver



For Today’s Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls “all the computer and math stuff” that was part of the job.

Enlarge This Image

Thor Swift for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

“People think of field archaeology as Indiana Jones, but much of what you really do is data analysis,” she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

“I keep saying that the sexy job in the next 10 years will be statisticians,” said Hal Varian, chief economist at Google. “And I’m not kidding.”

Multimedia

Ph.D. in computer science with focus on artificial intelligence and text analytics, M.I.T.

CAREER: He is a research scientist at I.B.M. who uses computing and modeling to extract knowledge patterns from text, video and audio data.

Jon Kleinberg, 37

EDUCATION: Bachelor's degree in computer science, Cornell; Ph.D. in computer science M.I.T.; a MacArthur Fellow

Quote of the Day, New York Times, August 5, 2009

”I keep saying that the sexy job in the next 10 years will be statisticians. And I’m not kidding.” - HAL VARIAN, chief economist at Google.

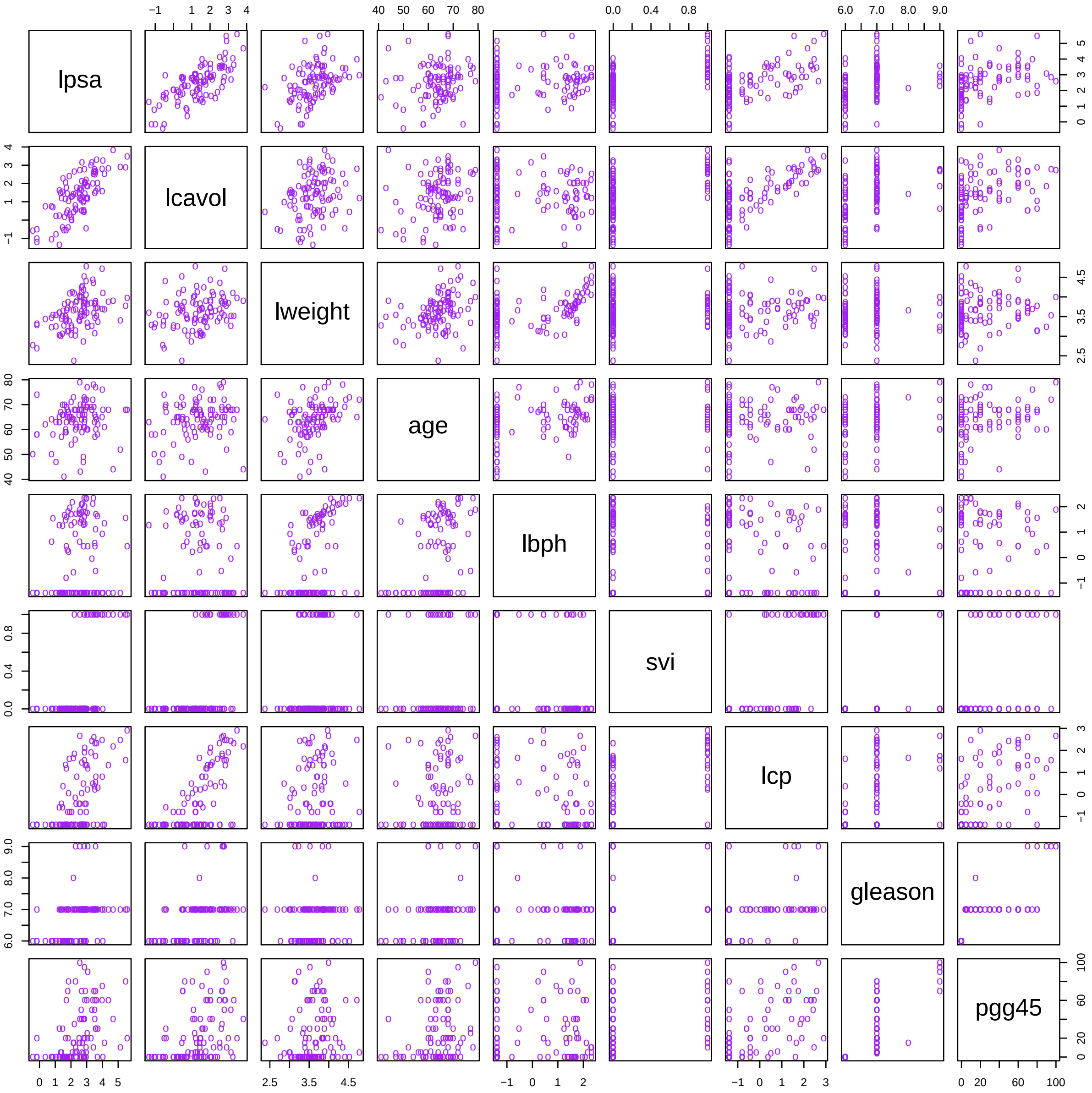
OLD

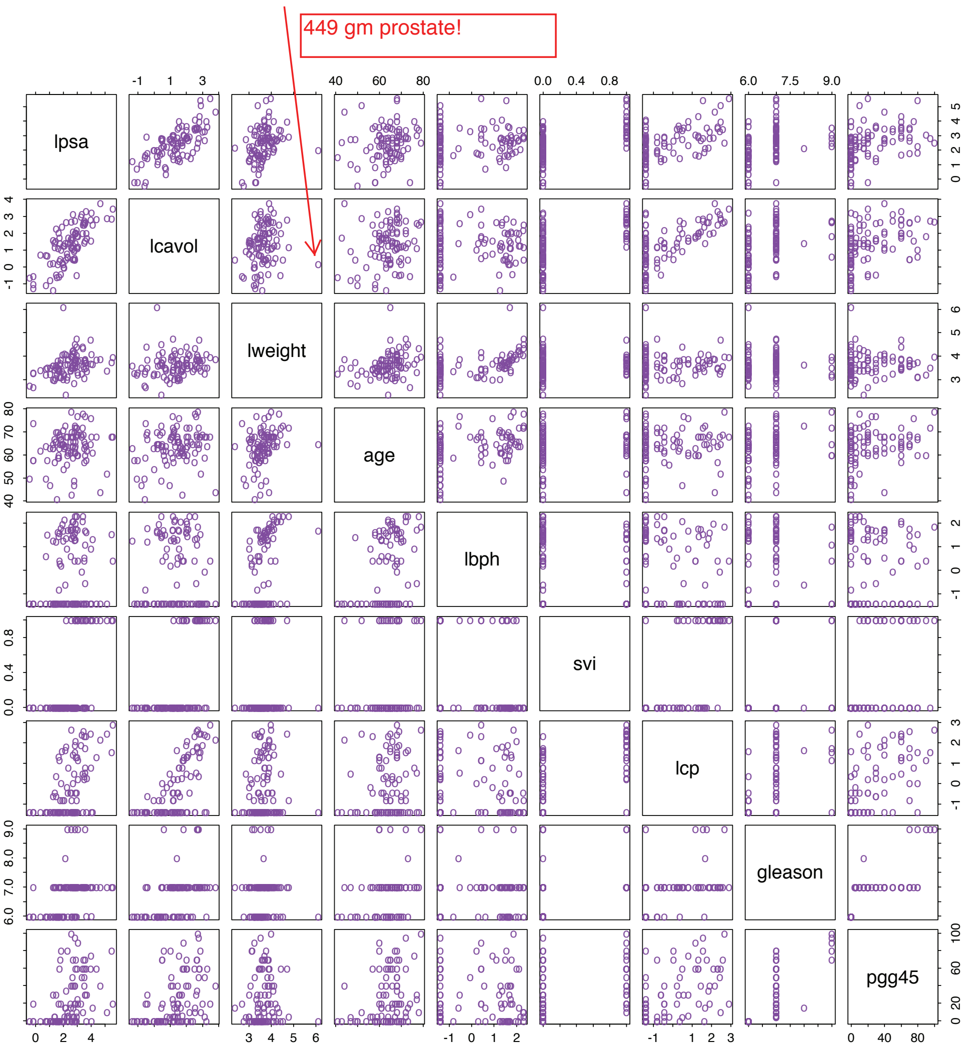
NEW

Statistical Learning Problems

- Identify the risk factors for prostate cancer (Fig 1.1).
- Classify a recorded phoneme (Fig 5.5) based on a log-periodogram.
- Predict whether someone will have a heart attack (Fig 4-12) on the basis of demographic, diet and clinical measurements
- Customize an email spam (Tab 1.1) detection system.
- Identify the numbers in a handwritten zip code (Fig 1.2), from a digitized image
- Classify a tissue sample into one of several cancer classes, based on a gene expression (Fig 1.3) profile
- Classify the pixels in a LANDSAT (Fig 13.6) image, by usage:

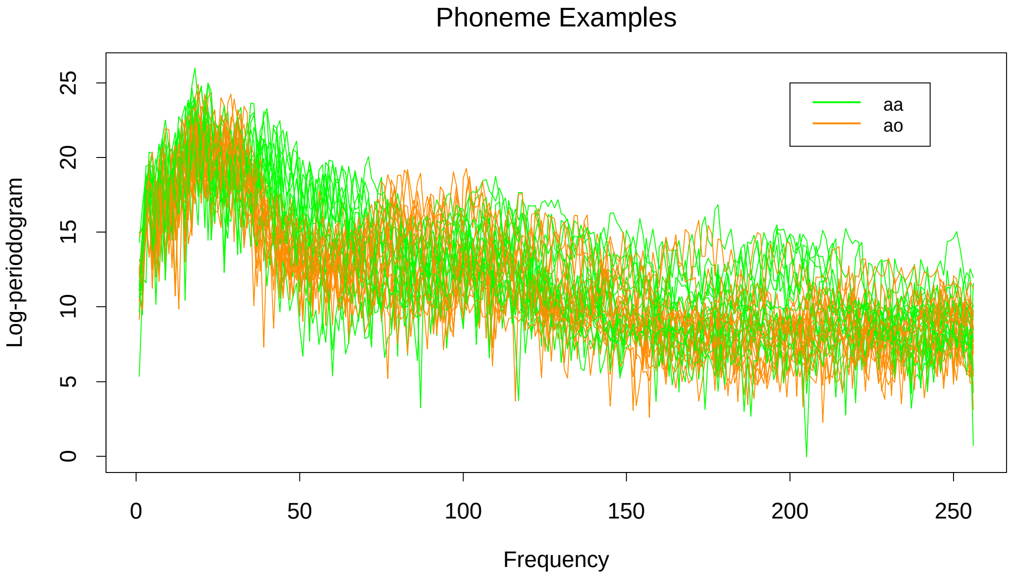
  {*red soil, cotton, vegetation stubble, mixture, gray soil, damp gray soil*}

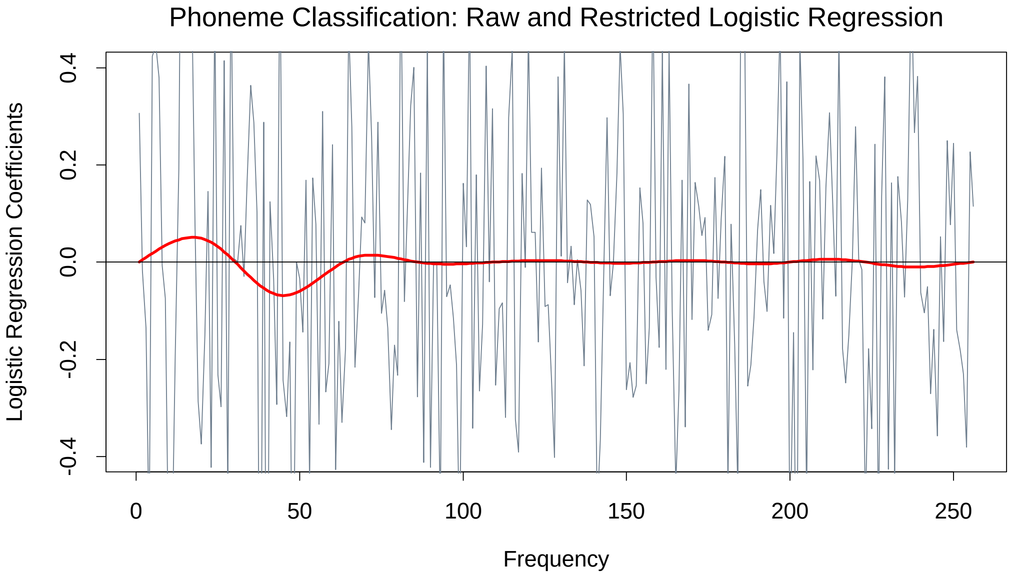

lpsa
lcavol
lweight
age
lbph
svi
lcp
gleason
pgg45
−1 1 2 3 4
40 50 60 70 80
0.0 0.4 0.8
6.0 7.0 8.0 9.0
0 1 2 3 4 5
2.5 3.5 4.5
−1 0 1 2
−1 0 1 2 3
0 20 60 100

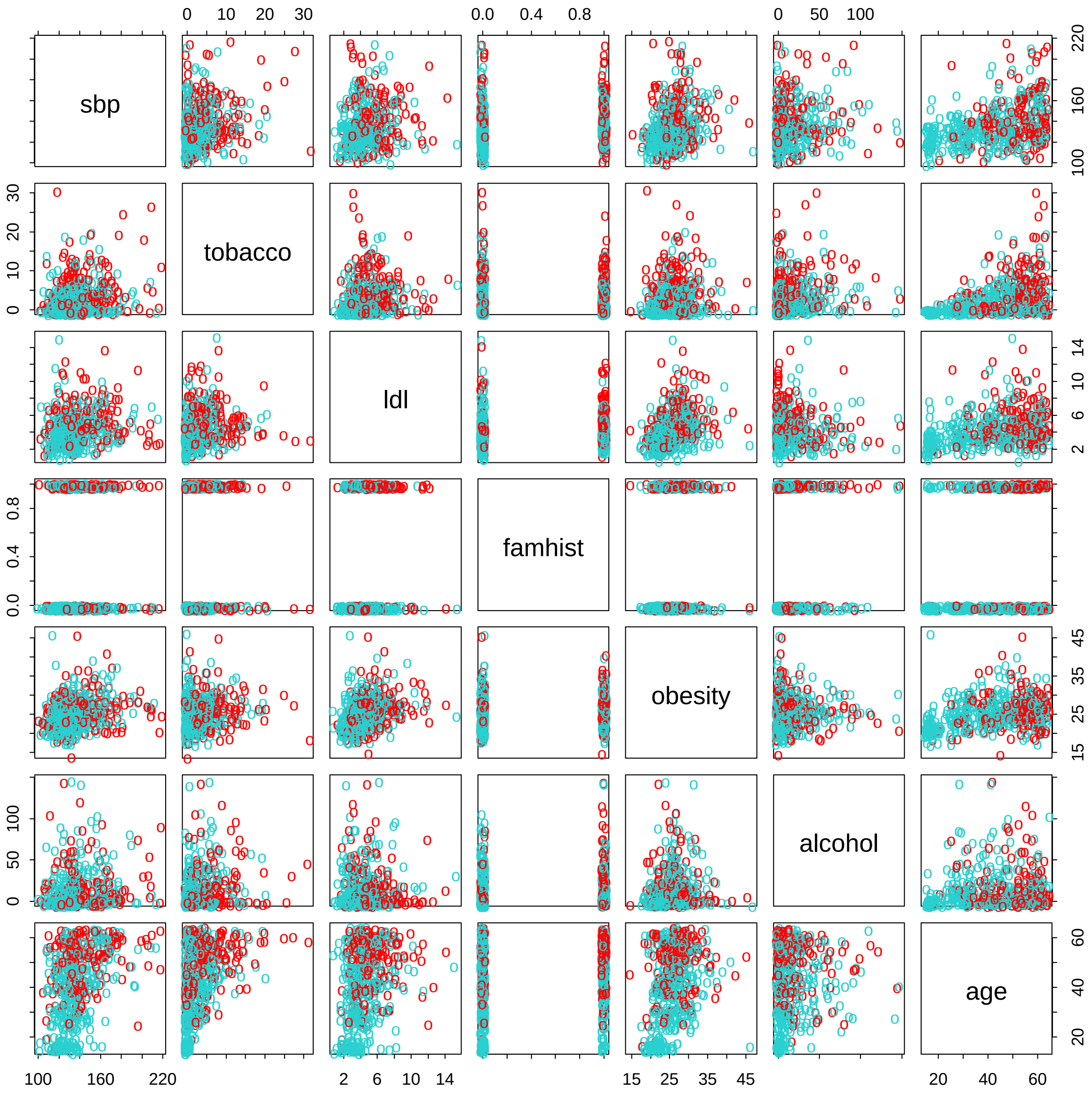


449 gm prostate!

Error pointed out by retired urologist Dr Stephen W. Link


Phoneme Examples
aa
ao
Log-periodogram
0
5
10
15
20
25
0
50
100
150
200
250
Frequency


Phoneme Classification: Raw and Restricted Logistic Regression
Logistic Regression Coefficients
-0.4
-0.2
0.0
0.2
0.4
0
50
100
150
200
250
Frequency

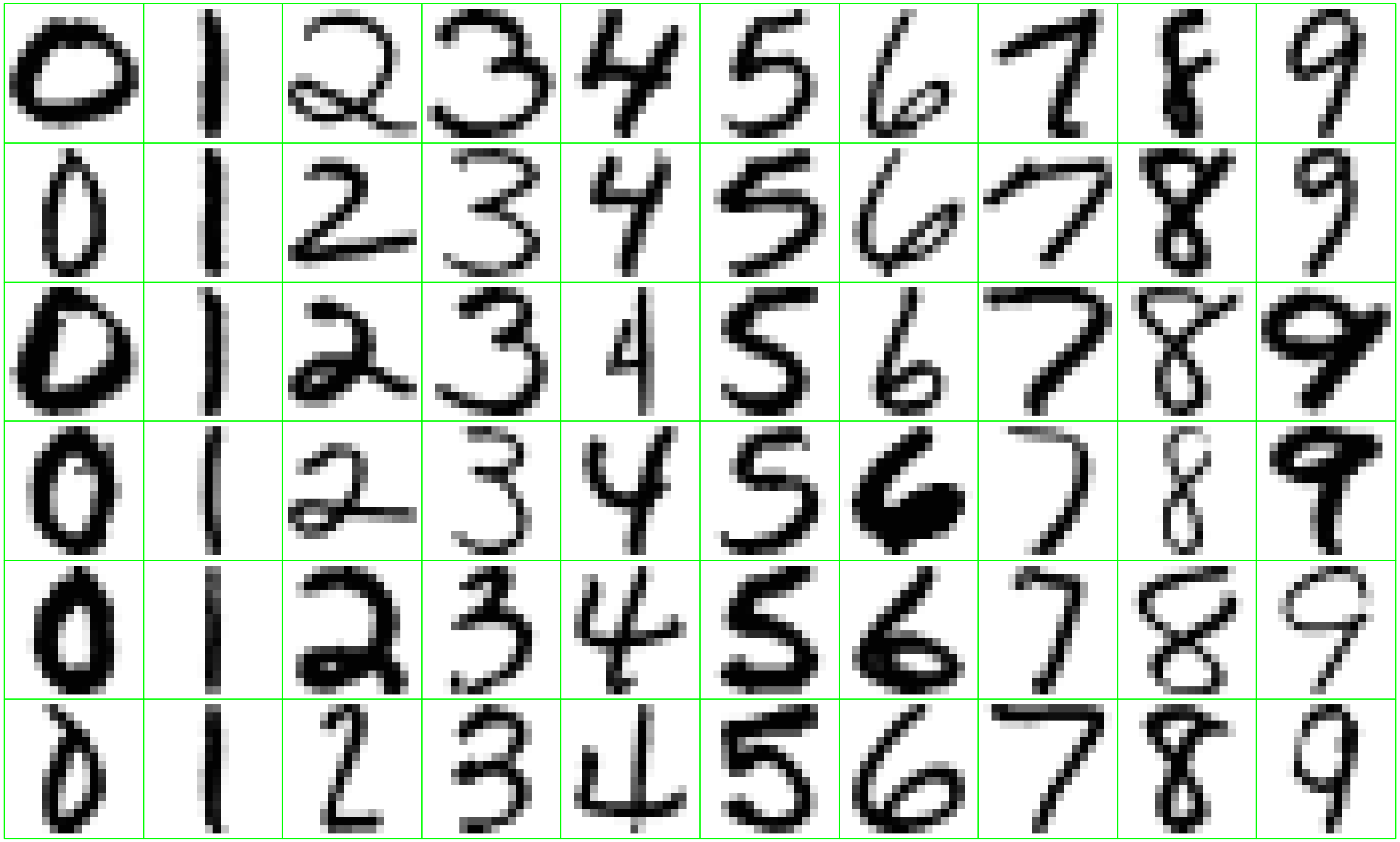

sbp
tobacco
ldl
famhist
obesity
alcohol
age
0 10 20 30
0.0 0.4 0.8
0 50 100
220 160 100
30 20 10 0
14 10 6 2
0.8 0.4 0.0
45 35 25 15
100 50 0
60 40 20
100 160 220
2 6 10 14
15 25 35 45
20 40 60

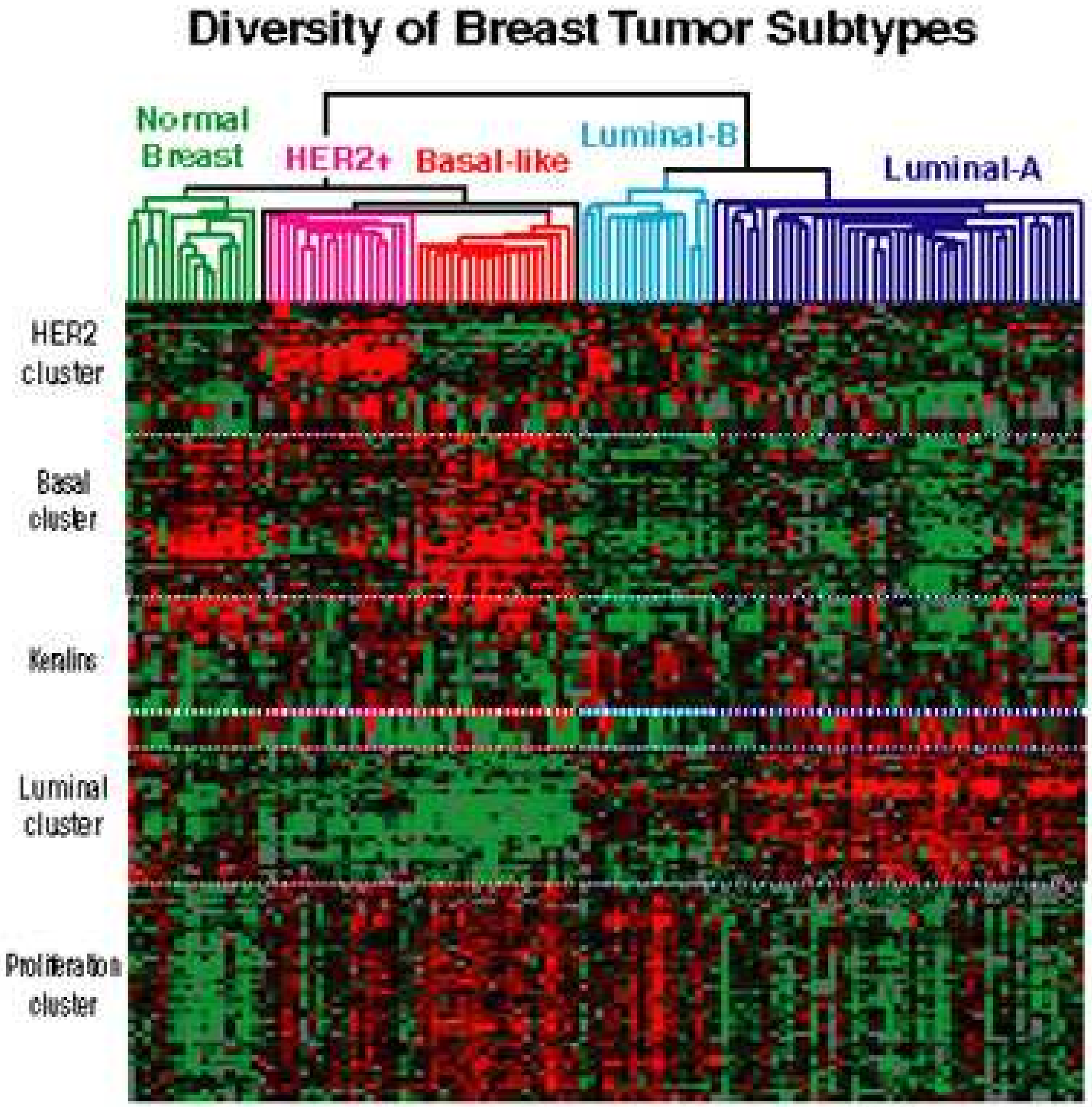
Spam detection

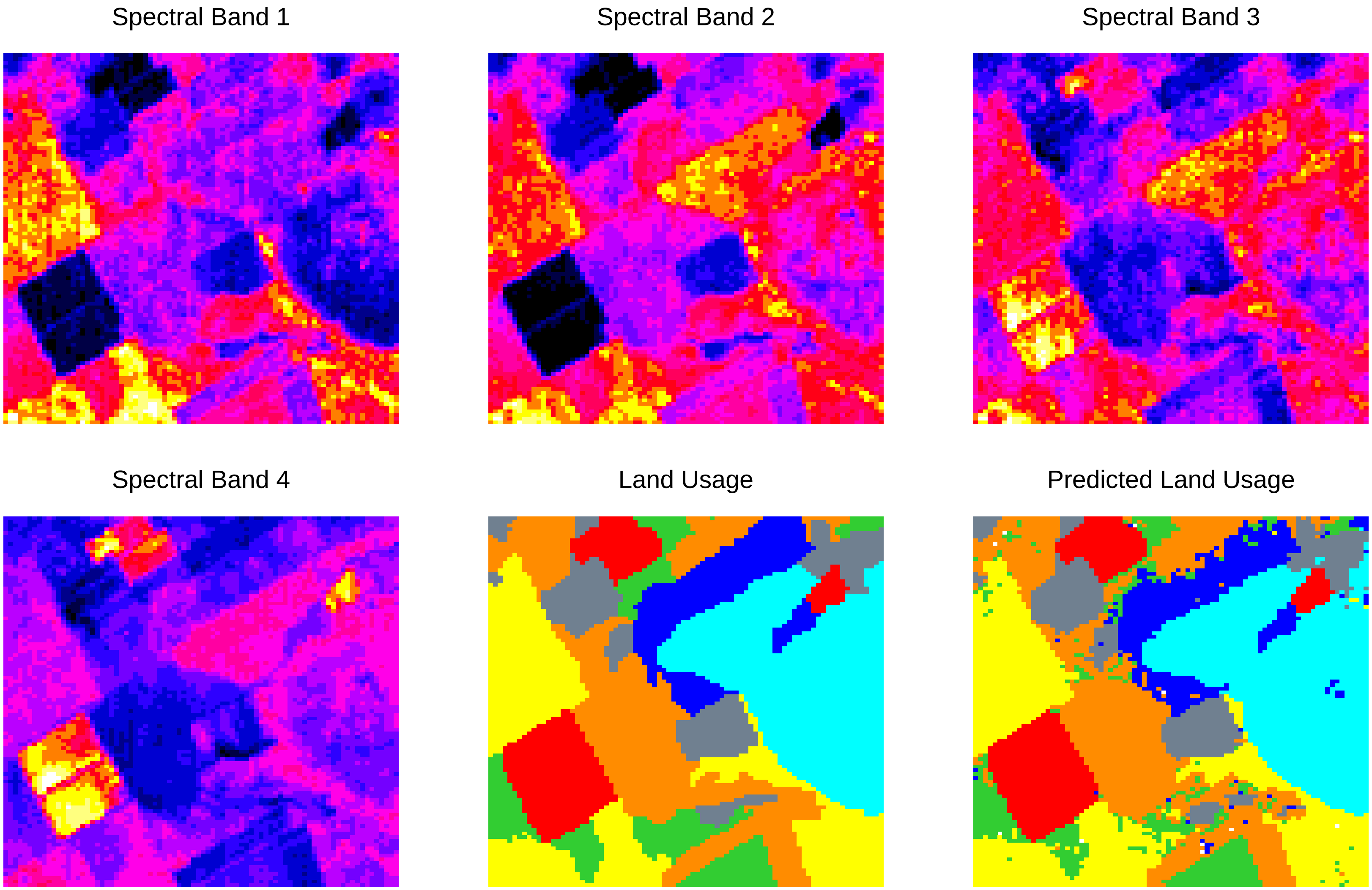
- data from 4601 emails sent to an individual (named George, at HP labs, before 2000). Each is labeled as *spam* or *email*.
- goal: build a customized spam filter.
- input features: relative frequencies of 57 of the most commonly occurring words and punctuation marks in these email messages.

	`george`	`you`	`hp`	`free`	`!`	`edu`	`remove`
`spam`	0.00	2.26	0.02	0.52	0.51	0.01	0.28
`email`	1.27	1.27	0.90	0.07	0.11	0.29	0.01

Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between `spam` *and* `email`*.*




Diversity of Breast Tumor Subtypes
Normal Breast
HER2+
Basal-like
Luminal-B
Luminal-A
HER2 cluster
Basal cluster
Keratins
Luminal cluster
Proliferation cluster


Spectral Band 1
Spectral Band 2
Spectral Band 3
Spectral Band 4
Land Usage
Predicted Land Usage

The Supervised Learning Problem

Starting point:

- Outcome measurement Y (also called dependent variable, response, target)
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables)
- In the *regression problem*, Y is quantitative (e.g price, blood pressure)
- In the *classification problem*, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample)
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Objectives

On the basis of the training data we would like to:

- Accurately predict unseen test cases
- Understand which inputs affect the outcome, and how
- Assess the quality of our predictions and inferences

Philosophy

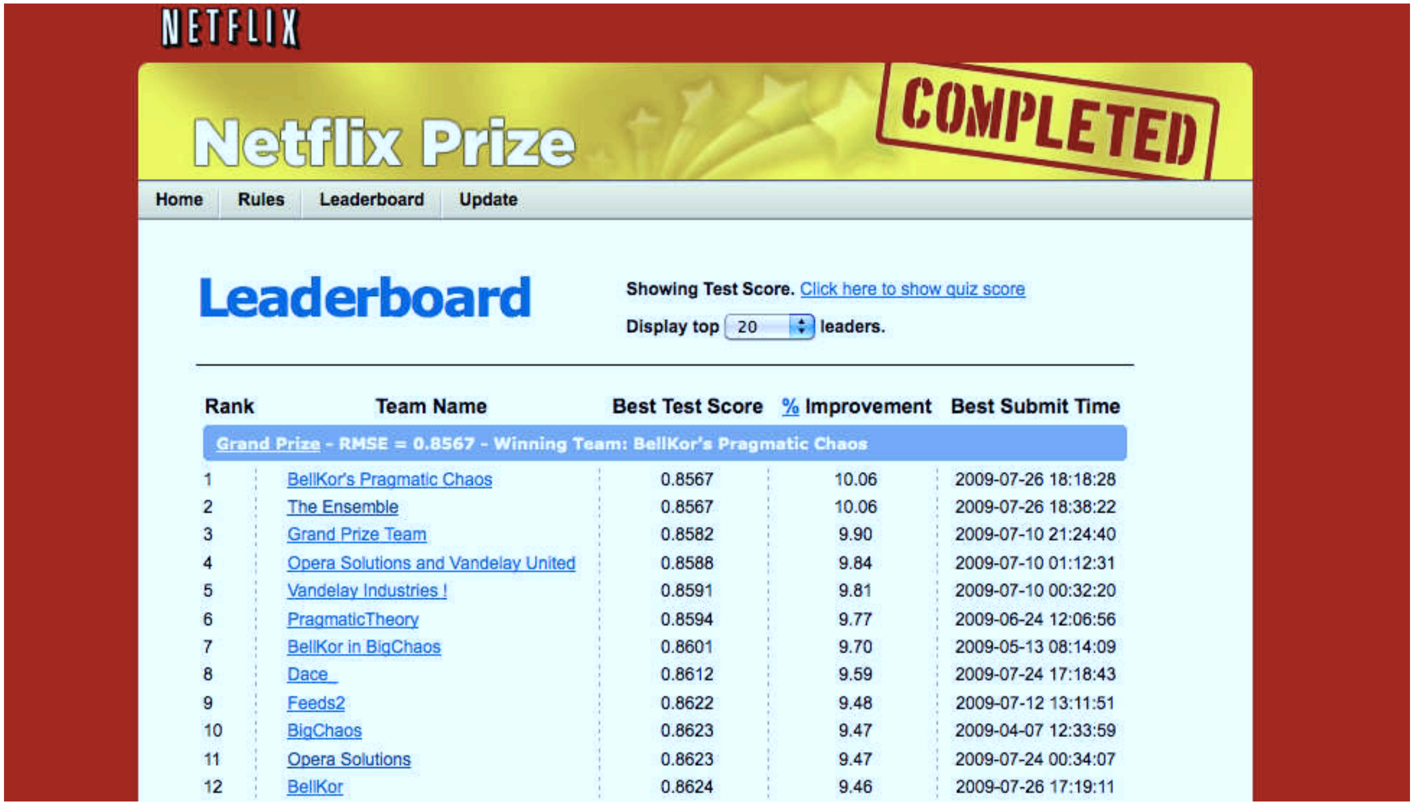
- It is important to understand the ideas behind the various techniques, in order to know how and when to use them.
- One has to understand the simpler methods first, in order to grasp the more sophisticated ones.
- It is important to accurately assess the performance of a method, to know how well or how badly it is working [simpler methods often perform as well as fancier ones!]
- This is an exciting research area, having important applications in science, industry and finance.

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- objective is more fuzzy — find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- difficult to know how well your are doing.
- different from supervised learning, but can be useful as a pre-processing step for supervised learning

The Netflix prize

- competition started in October 2006. Training data is ratings for $18,000$ movies by $400,000$ Netflix customers, each rating between 1 and 5
- training data is very sparse— about 98% missing
- objective is to predict the rating for a set of 1 million customer-movie pairs that are missing in the training data
- Netflix's original algorithm achieved a root MSE of 0.953. The first team to achieve a 10% improvement wins 1 million dollars.
- is this a supervised or unsupervised problem?



Leaderboard

Showing Test Score. Click here to show quiz score

Display top 20 **leaders.**

Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11