Statistics 151B: Modern Statistical Prediction and Machine Learning

Overview and introduction

Administrative information

- Homepage: `http://www.stat.berkeley.edu/~jon/stat-151b-spring-2012`
  - All announcements and materials here.
  - No printouts distributed in lecture.
- Instructor: Prof. J. McAuliffe
  - `jon@stat.berkeley.edu`
  - office hours: Thu 12:30 PM - 1:30 PM
- Graduate student instructor: Jeff Regier
  - `jeff@stat.berkeley.edu`
  - office hours: TBA

Administrative information

- Course texts:
  - T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed., 5th printing. Free pdf.
  - P. Dalgaard, *Introductory Statistics with R*, 2nd ed.
- Prerequisites:
  - basic linear algebra ($Ax = b$, $Ax = \lambda x$)
  - multivariate calculus ($\partial f / \partial x_i$, $\nabla_x f$)
  - statistical inference
- This is not a math course, but it uses math in an essential way to make ideas precise.

Computing

- Platform: the R statistical computing environment, `http://www.R-project.org .`
- Freely available, open source.
- Installation packages available for Windows, Mac, Linuxen.
- The de facto standard for statistical computing all over the world.
- You will become proficient in R: section will cover the Dalgaard book extensively (learning R and reviewing basic statistical inference).

Homework

- Four assignments over the course of the semester, due roughly every two weeks.
- Work in groups of three.
- Data analysis in R (submit a transcript of your R session, source files, brief written summaries), using techniques from the course.
- Problem-solving to check your understanding of theories and methods.
- No late submissions.
- 30% of the final grade.

Midterm

- Topics covered on the midterm will be clearly indicated. One of the weekly discussion sections will be devoted to review/Q&A.
- Thursday March 22nd, in class.
- 30% of the final grade.

Final project

- This year, something new: a competition on a contest dataset.
- Build the best prediction rule you can using ideas and methods from the course.
- Write a project report describing your analysis and results (guidelines provided beforehand).
- Work in the same group of three.
- Project grade depends on quality of work and report, *not* standing in the competition.
- I will open the contest on Kaggle in early April.
- Jeff and I will compete. . . prepare to get owned.
- 40% of the final grade.

Final project – alternative

- A substantial data analysis project.
- You pick
  - an area/topic/field/subject you are interested in,
  - a nontrivial dataset related to it,
  - a question to answer or theory to test.
- Address the question or theory using some of the ideas and methods from the course.
- Write a project report describing your analysis and results (guidelines provided beforehand).
- Work in same group of three.
- 40% of the final grade.

FINAL PROJECT DEADLINES

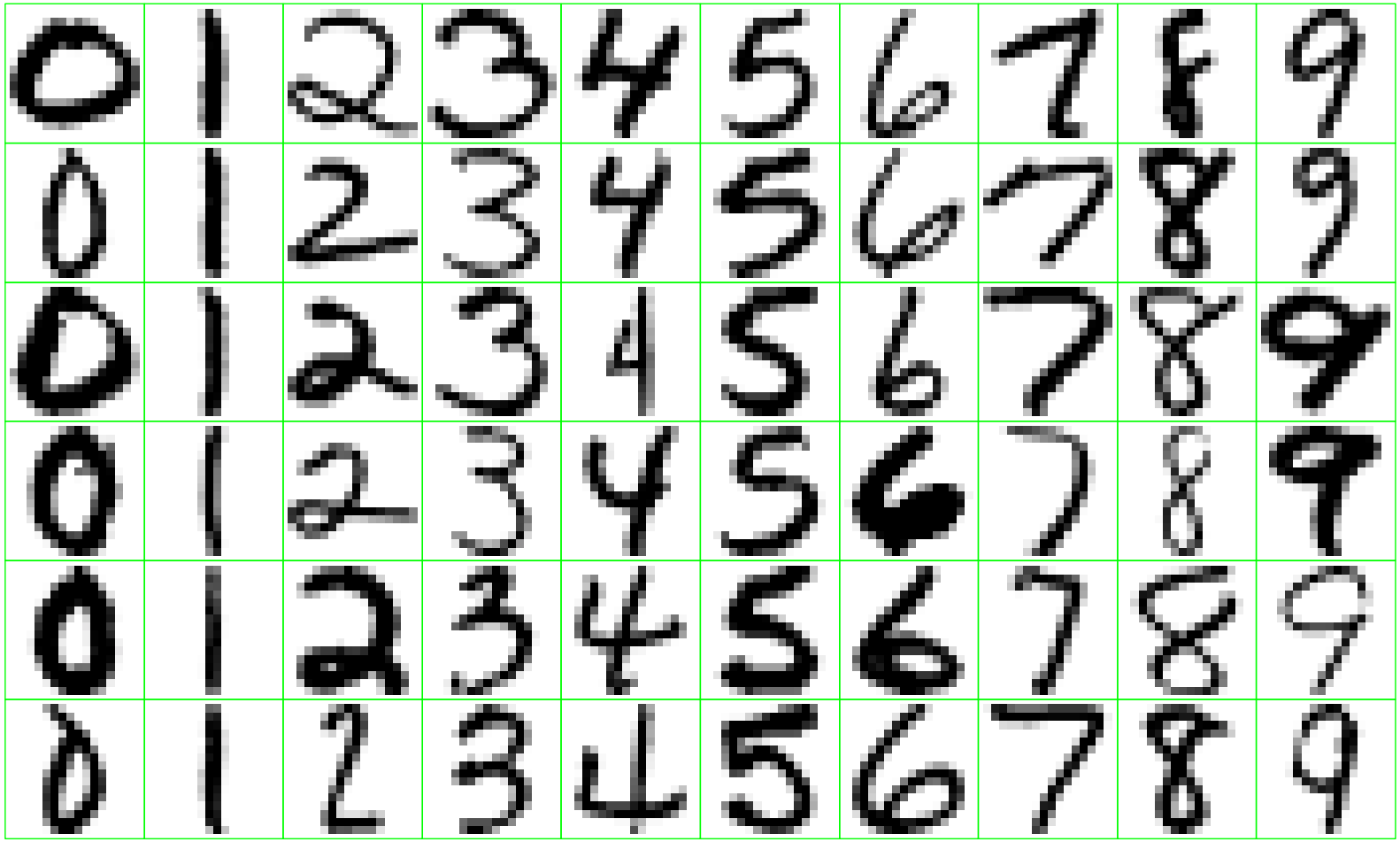
- Thursday April 5th: for those opting out of competition, final project proposal due **at the beginning of lecture**.
- Friday May 11th: final project report due **at 3pm**.
  - Last day permitted by the university. No extension is possible.
- You cannot pass the course without doing the final project.
- **Get in front of this project. Do not leave it to the last minute.**
- If you like, please come to my office hours with your ideas or questions about a candidate project.

Example: spam filtering

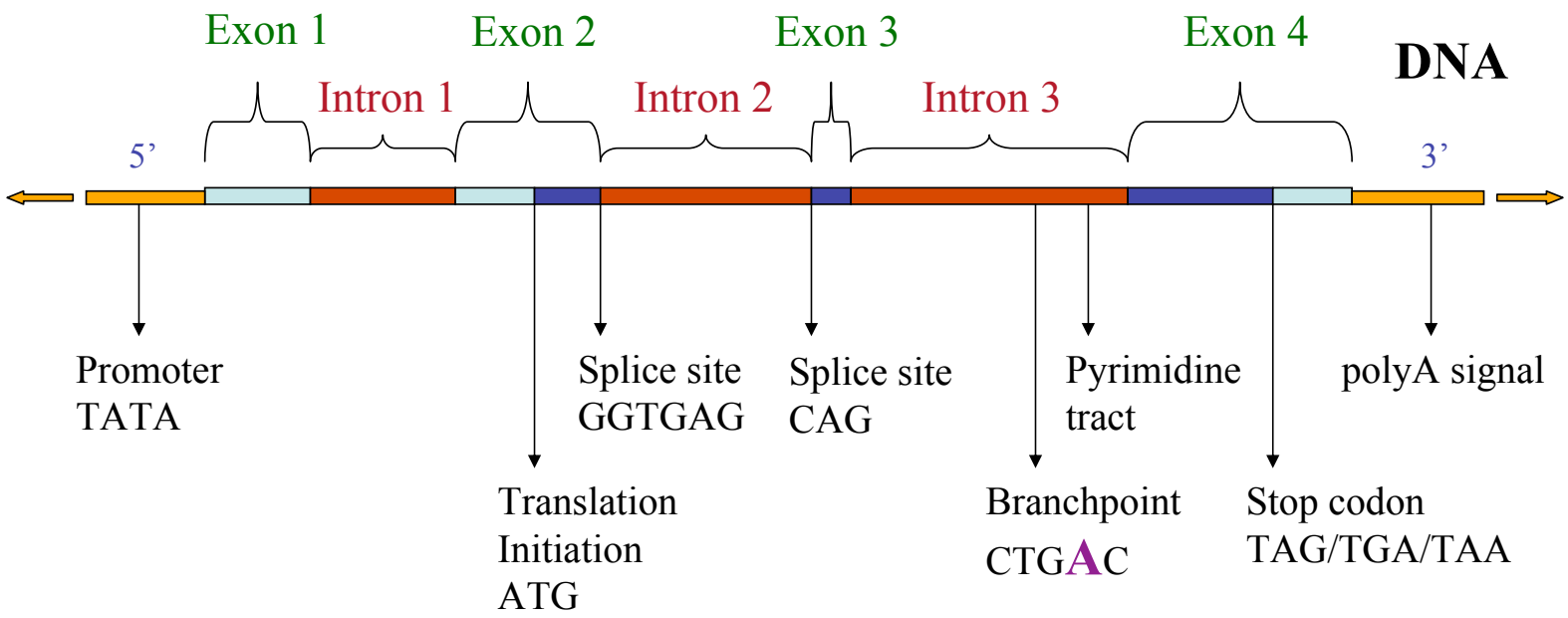
Table 1: *Average percentage of words or characters in an email message equal to the indicated word or character. We have chosen the words and characters showing the largest difference between* `spam` *and* `email`.

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

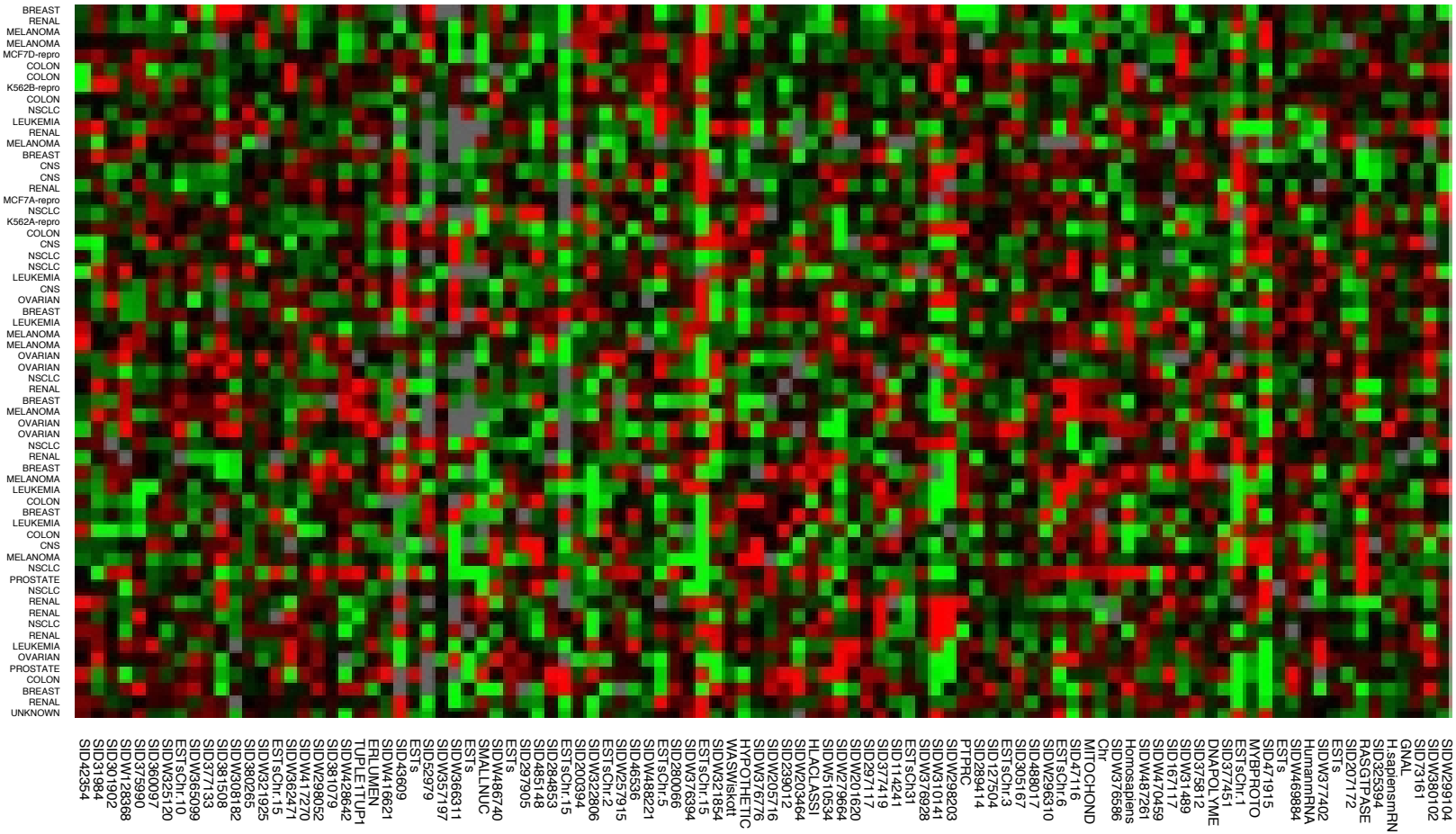
Example: handwritten-digit recognition



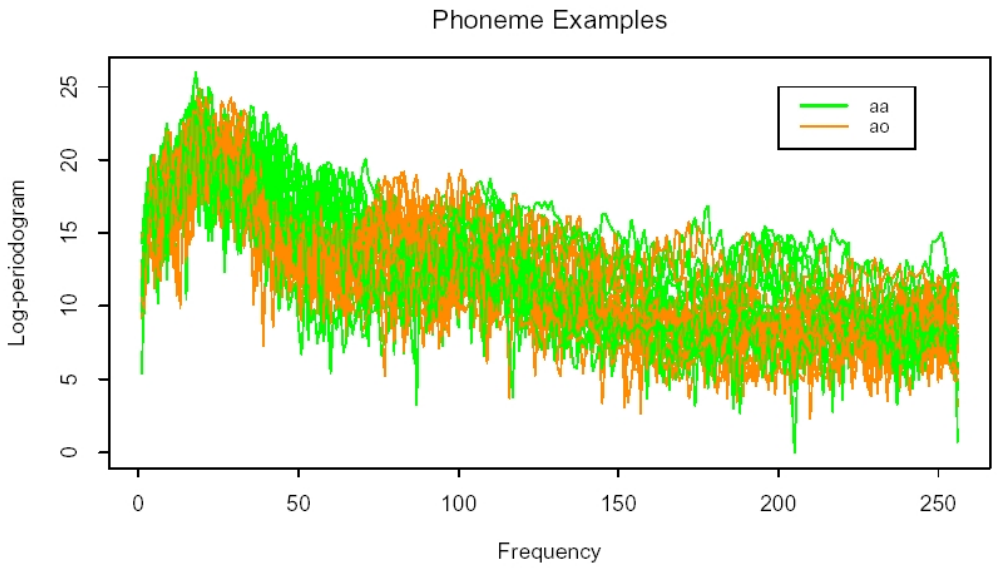
Example: gene finding

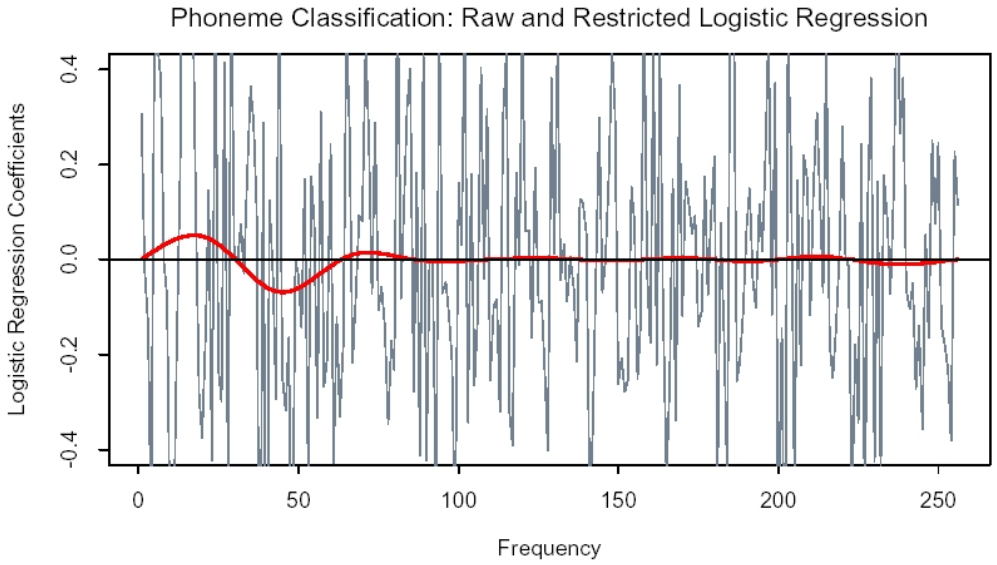


Example: typing cancer with microarrays

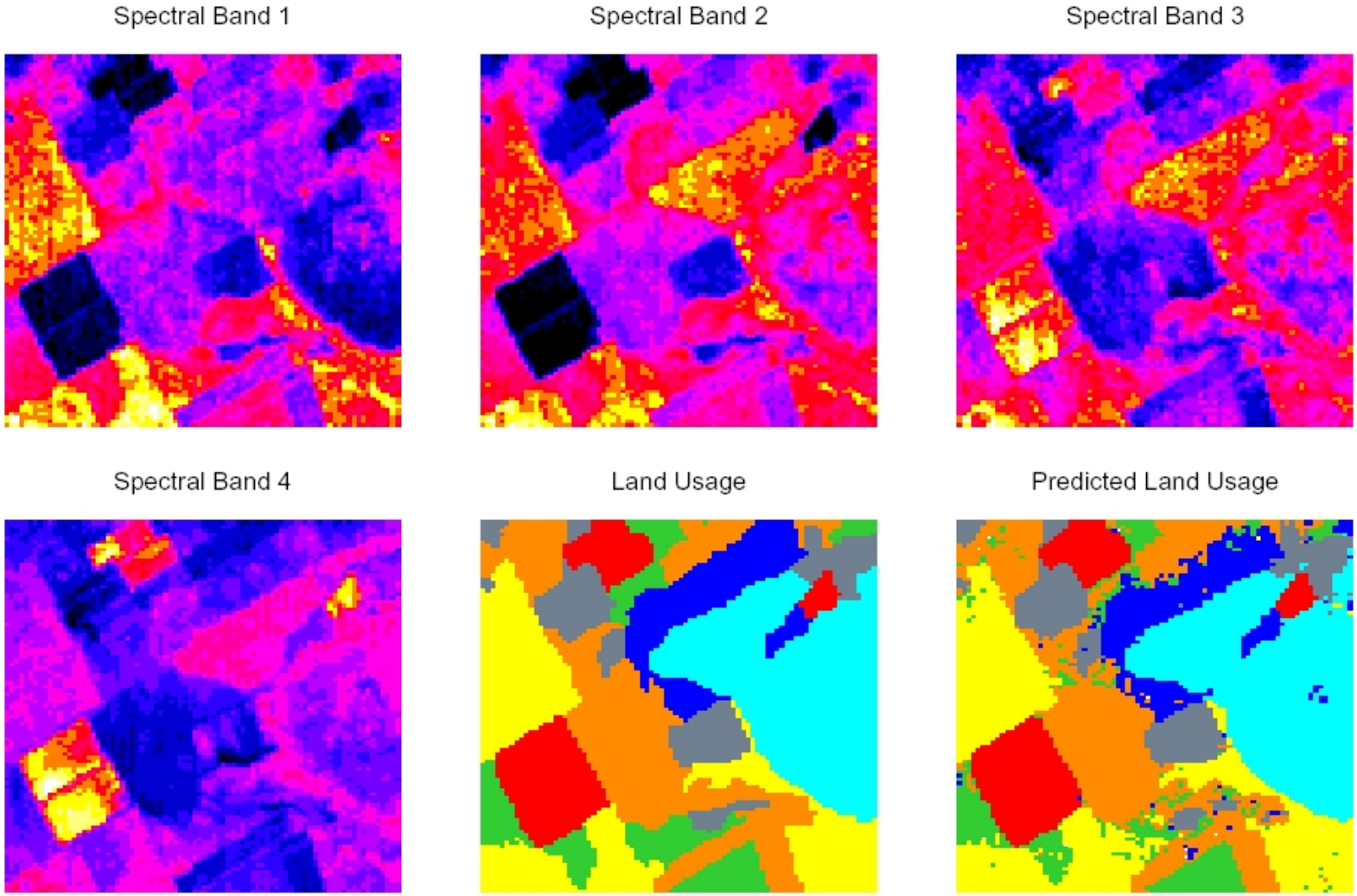


Example: phoneme detection



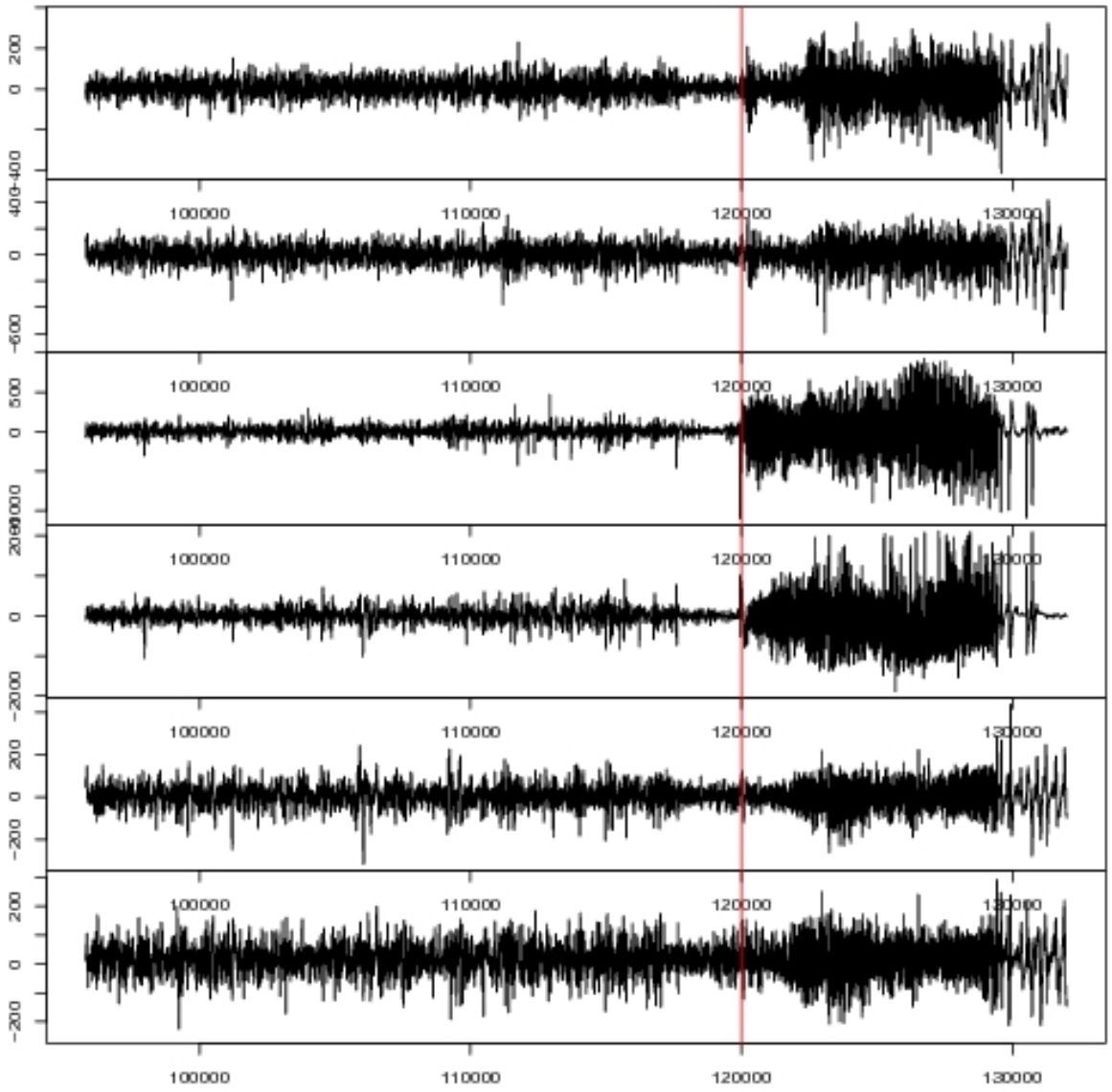


Example: Land usage from LANDSAT images

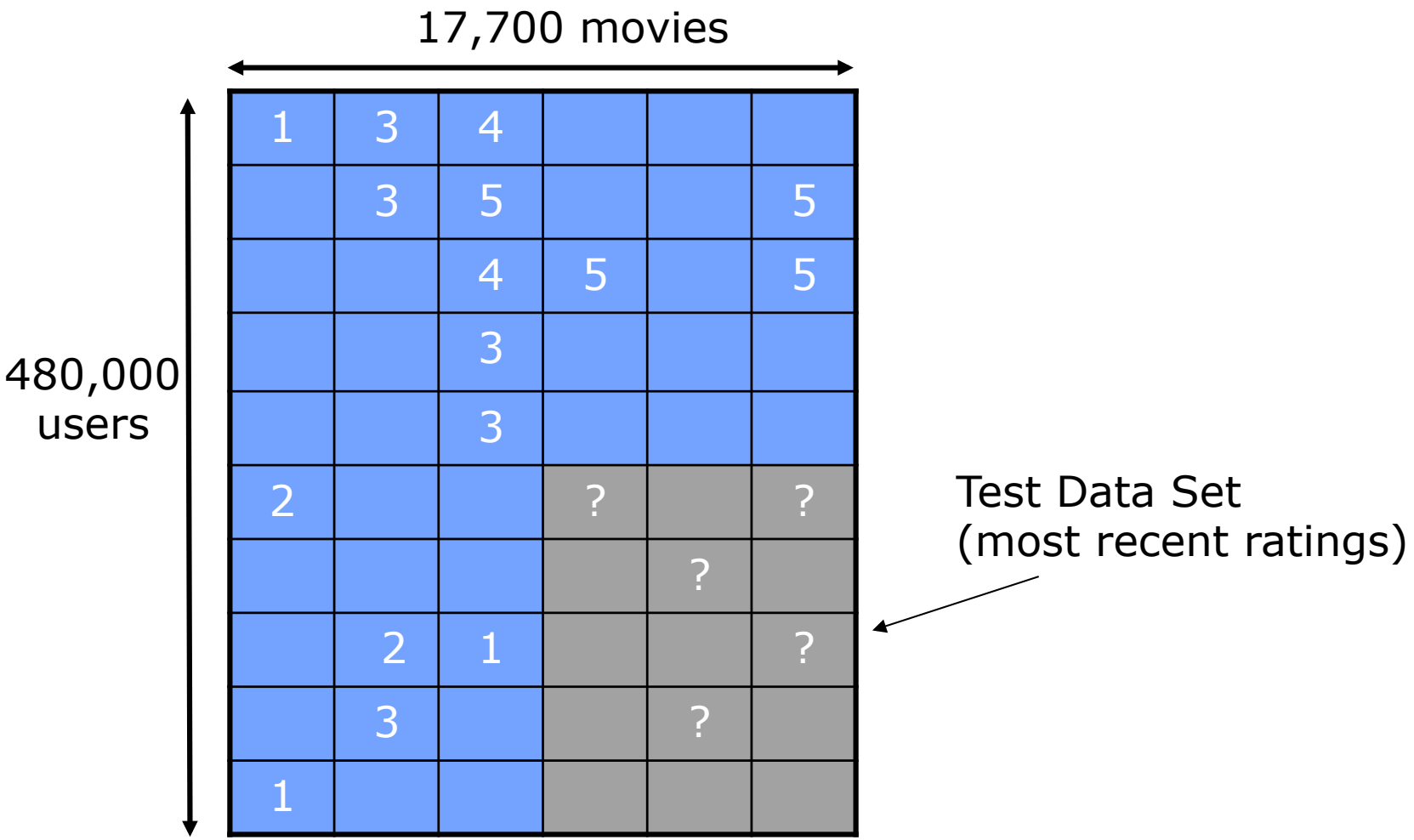


Example: face detection

Example: seizure prediction



Ex: “Collaborative filtering” (the Netflix prize)



Statistical prediction

- AKA “supervised learning” in the machine learning community. Ingredients of the general framework:
- An outcome Y (called the *response variable*, *dependent variable*, or *target*).
- A p-vector of observables X (called the *predictors*, *covariates*, *features*, *inputs*, *regressors*, or *independent variables*).
- *Regression*: Y is a continuous quantity (price, weight, concentration).
- *Classification*: Y takes values in a finite, unordered set:
  - clinical outcome: { lived, died }
  - consumer credit: { bankrupt in 2006, or not }
  - cancer type: { breast, melanoma, leukemia, . . . }

Objectives

- We have *training data* $\{(x_1, y_1), \ldots, (x_N, y_N)\}$. These are *observations*, *examples*, or *instances*. (Don't call them *samples*; the whole training data set is one *sample*.)
- Using the training data, the goals are:
  - Build a rule to predict y_{new} from x_{new}.
  - Understand which predictors are related to the response.
  - Estimate what the quality of future predictions and inferences will be.

Schedule of topics

[See syllabus]

A statistician's manifesto (Trevor Hastie)

- Understand the ideas behind the statistical methods, so you know how to use them, when to use them, when *not* to use them.
- Complicated methods build on simple methods. Understand the simple methods first.
- The results of a method are of little use without an assessment of how well or poorly it is doing.
  - *Corollary*: simple methods are sometimes just as good as complicated methods; "technology tends to overwhelm common sense" (D. Freedman)
- Statistical prediction is an active and exciting area of research, touching science, industry, and finance.