Projects – Fall ‘11
Goal: To do some real machine learning

- A project *you* are interested in works better
- Data is often the hard part (get it in plenty of time)
- Usually involves additional reading
- Using software packages OK, but
  - “plug and play” an algorithm undesirable
  - Comparing algorithms is good
- Insightful survey/review paper OK (may involve talk)
- Theory problem OK, but high-risk
Requirements:

- **Project proposals** (due Fri. Oct 21)
  - One paragraph to one page per group
  - Problem, data, planned approach, success metric

- **Progress report** (due Wed Nov. 16)
  - 2 to 4 pages per group
  - Introduction, description of progress, problems, and expected scope of final project, disclosures

- **Final report** (due Dec. 5)
  - Body (8 to 15+ pages of text, 12 pt font) with references and citations. Can have appendices. Should have abstract and introduction (like a conference paper) and discuss related work
Some previous projects/ideas

• Exploring techniques for file pre-fetching
• Predicting senators’ votes based on campaign contributions
• Clustering of repetitive Genomic elements
• Comparing Naïve Bayes and Neural nets for text (topic) classification
• Author identification from text
• Road extraction from Aerial images
• Predicting success of major league baseball teams
• Exploring methods for handwritten digit recognition
• Wine cultivator recognition from chemical properties
• Kaggle competitions (Heritage Health Prize)
• Bird species ID from song (Matthew McKown)
Other comments

- Many data mining challenges/contests
  - Kaggle (including Heritage health)
  - can measure success
- Can experiment with method from research paper
- Recent work in:
  - Use of unlabeled data,
  - out of sample data
- Groups (2-4) are fine
- Double dipping OK