Project Guidelines
Projects!

- Goal: apply machine learning to an interesting task

- Proposal (due next week): 1 pg
  - Who is in your group
  - Your task (and why is it interesting?)
  - Where did/will you get your data?
  - What’s your initial approach?
    - It’s okay if you can’t say much about algorithms yet
Important Rules of Thumb

- If possible – set aside test data now, don’t examine until end of course
- Allow time for iteration
- Understand your results
Meetings

- Status discussion
  - May 26/27

- Optional
- Sign-up procedure to appear on course page
How to do Machine Learning

1) Pick a feature representation for your task
2) Compile data
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the algorithm
6) Analyze the results
7) *Probably: go to (1)*
How to do Machine Learning

1) Pick a feature representation for your task
2) **Compile data**
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the algorithm
6) **Analyze the results**
7) *Probably: go to (1)*
How to do Machine Learning

1) Pick a feature representation for your task
2) Compile data
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the algorithm
6) Analyze the results
7) *Probably: go to (1)*
What’s the right task (for the class)?

- **Okay**: choose interesting, standard ML data set from UCI repository
- **Better**: use pre-existing but unique/important data set (e.g. Netflix prize, Google n-grams, Wikitables)
- **Best**: choose novel, important task and gather *new* data

Project **completion** is important
- Choose something interesting, but also something you can get done!

Things to consider:
- Availability of data
- “Munging” required
- Your knowledge of the domain
Examples (1 of 5)

- Something from your research
- The $ ones:
  - Price prediction (e.g. stock market)
  - Box office success
  - The “next big sound” see: nextbigsound.com
  - Sports contests
- UCI Repository
  - Tons of tasks, wines, mushrooms, text…
Examples (2 of 5)

- **More data sources**
  - Data.gov – US State data (agriculture, spending, etc.), census data
    - Also: NYC Big Apps
  - Customer reviews (summarization, deception detection…)
    - Other item attributes from review?
  - WikiData
  - [City of Chicago data portal](https://data.cityofchicago.org)
  - Twitter
Some of my favorites:

- Predicting blog “anger”
  - (I have a small data set for this)
- Politician sentiment on issues (from speech text)
- Compressing the Google n-grams data set
  - Unprecedented coverage, but takes 150G
  - Could a good ML approximation be much smaller?

Which lectures are good?
- I built a small data set for this last Spring

Other things people have done:
- Will you get into your target sorority? (based on income, major, activities, etc)
- SafeRide wait times
- Can you predict morphology in Arabic words based on semantics?
Examples (4 of 5)

- Generics in language

  Birds lay eggs
  Mosquitoes carry the West Nile Virus

  Horses are female
  Humans are seven feet tall

Can we build a predictor for this?
Examples (5 of 5)

- CTECs scores from text

- Ranking ungrad, grad programs in a particular field
  - Do a survey, build predictor of human rankings
  - Or mine Google scholar
Peer Review!

- Experiment in grading
  - Uses machine learning!
- You will review ~3 other project proposals/status reports/final reports
  - A (random) sample of your peer reviews will be graded for accuracy
- For your project grade
  - I will review at least one of your project deliverables
  - Your grade will be based on “an algorithm” that leverages my review and peer review scores
    - We will discuss the algorithm in class, later
Brainstorming project ideas

- What’s your second best project idea?
  - …that someone else could try