Final project

- Pick a group of 2-3 students – and a team name
- Pick a problem and dataset
- Look up related problems
- Run 3-5 methods, plus a baseline
  - Optimize hyperparameters
  - Show results in a table
- What can you do that is clever?
  - Usually taking advantage of the specifics of the problem
Final project deliverables

- 11/14 Project proposal
  - Give us enough information to give you feedback
- 11/28 Project checkpoint
  - Show that you are making progress
- 12/7-12/9 presentations in pods
- 12/12 Project report, code and notebook
Real World ML

- **Who cares? Why?**
  - Loss functions

- **Model form**
  - Feature engineering
  - Semi-supervised learning

- **Regularization**

- **Visualization/Interpretation**
  - Causality: “what if?”
Missing data

◆ **Real valued**
  - Replace the missing item with zero or average
  - Add a new variable indicating if it was missing

◆ **Categorical**
  - Treat it as a new category value
Categorical data

- Encode “one hot”
- Learn an embedding (semi-supervised)
- Use “mean encoding”
  - Replace the category variable with the average y-value for the corresponding category value.

<table>
<thead>
<tr>
<th>x_1</th>
<th>x_2</th>
<th>y</th>
<th>x_1'</th>
<th>x_2</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3.7</td>
<td>1</td>
<td>1.5</td>
<td>3.7</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>2.1</td>
<td>2</td>
<td>1.5</td>
<td>2.1</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>0.9</td>
<td>4</td>
<td>4.0</td>
<td>0.9</td>
<td>4</td>
</tr>
</tbody>
</table>
What you should know

◆ Think about the ‘true’ loss function (utility)
  ● Distinguish modeling from decision making

◆ Think about the features
  ● What do you have? What can you get?
  ● How should they be regularized (blocks)

◆ Think about what ML methods fit best
  ● Compare several