Team:

Rahul Bazaz

Howard Fan

Maulik Shah

Yuqing Zhang

Project:

1. (Difficulty \*\*\*\*\*\*, Preference \*\*\*\*) Data Science Bowl 2017, medical image processing, lung cancer detection

Reference: <https://www.kaggle.com/c/data-science-bowl-2017>

Preprocessing tutorial: <https://www.kaggle.com/gzuidhof/data-science-bowl-2017/full-preprocessing-tutorial/notebook>

Specific Aims:

* Process the data, and apply CNN or other ML techniques, build classifier for lung cancer, then evaluate performance

2. (Difficulty \*\*\*, Preference \*\*) StarPlus fMRI data

Reference: <http://www.cs.cmu.edu/~guestrin/Class/10701/projects.html#fmri>

Specific Aims:

* dimensional reduction and feature selection
* build classifier

3. (Difficulty \*\*\*\*\*, Preference \*\*\*\*) Image Captioning - Li Fei-Fei

Reference:<http://cs.stanford.edu/people/karpathy/cvpr2015.pdf>

Specific Aims:

* Combine method in Reference with Segmentation from Natural Language Expressions (<http://www.eccv2016.org/files/posters/S-1A-07.pdf>)? then see if it improves the performance in reference

4. (Difficulty \*\*\*, Preference \*\*\*) UC Irvine data sets - Amazon Book review dataset

Reference: [http://archive.ics.uci.edu/ml/datasets/Amazon+book+reviews#](http://archive.ics.uci.edu/ml/datasets/Amazon+book+reviews)

Specific Aims:

* Semtiment analysis, past research uses bag-of-words + n-gram to generate features, we could use methods like word2vec?
* Use features built from bag-of-words (or our new method) to build a classifier (or compare a bunch of classification methods, including neural network, random forest, etc.); compare with blog or compare between methods

A few other datasets that might be interesting:

* Pokemon: <https://www.kaggle.com/alopez247/pokemon>
* Trump’s world: <https://www.kaggle.com/buzzfeed/trumpworld>

(I don’t have specific ideas for these yet, just think they are interesting topics :) )