Term Project Data Sources & Other Info

It is essential for a solid hand-on project that you find a good dataset & problem. The next most important thing is that you a) start early and b) talk to me and the TAs often. This project could end up being personally important: It is very common for people to ask during interviews about projects that you did at the University and actually get jobs based on those projects. This can become one of those cases.

Read the below material CAREFULLY.

Finding a good project:

1) It has to be a problem with some **ACTION** to take
2) The scenario has to make sense
3) Don't do something difficult-to-impossible (predicting stock prices or the weather; also doing so isn't *actionable* in itself - see Point 1)
4) It needs to be technically feasible within your group (you are welcome to simplify)
5) You need to be able to get the data (you need the data a.s.a.p.; think: no later than 2 weeks to get started with something)
6) The data has to satisfy certain criteria: enough features, enough instances
7) Identify something you can actually evaluate. This typically means that you can simulate your ACTION, or the basis for taking the action (e.g., a probability estimate) on predictions you make on a subset of the data -- and where you can argue how much better the overall results are over some 'less-intelligent' strategy that is not utilizing the predictions.

Below, there are some tips on finding a suitable dataset for your project. Typically projects using data supplied by one of the team members tend to do well (e.g., from someone's company).

Some examples of projects that tend to be difficult to pull off in terms of making a good business case:

- Predict Movie Revenue
- Predict Flight Delay
- Predict Yelp Ratings
- Anything with Genetic Data

Sources of data (these are examples from the past; I did not go recheck them all):

- YOURS – do you have data from work? – think about using it!
- Data Repositories such as
    - http://www.kdnuggets.com/datasets/index.html
    - http://www.sigkdd.org/kddcup/index.php
        - For example: the KDDCUP 1998 data is very realistic, doable, but not simple
        - check out some others..
    - http://kdd.ics.uci.edu/
- NYC has a great selection of really interesting data.
    - https://data.cityofnewyork.us
- DonorsChoose http://data.donorschoose.org/open-data/overview/
    - local NYC non profit is waiting for your help to get material to teachers in the area

    - This is REAL – they may actually use what you build!
    - They have more internal data than you will see online
- Kaggle http://www.kaggle.com/competitions
    - Many. You can also draw inspiration from there about your business case
- Sports data (NBA, MLB, etc.)
- Recommender system data
    - Movielens, Netflix, BookCrossing, and I think others
    - Can this be used for a recommender-like application?
        - https://github.com/sidooms/MovieTweetings
- Amazon Public Datasets
    - http://aws.amazon.com/datasets/
- An extensive list of publicly available datasets – see if you can dig up an interesting one!
    - https://github.com/caesar0301/awesome-public-datasets
- Google Dataset Search:
    - Google recently developed a service that helps with searching for datasets in particular.
        - https://toolbox.google.com/datasetsearch
    - You can also find interesting datasets if you search online.

Also, you are at liberty to make up certain details of the task. For instance, you can pretend not to have data that you have or pretend that the solution does not already exist. How would you approach the problem in that situation?

It is important to keep the following in mind

The technical details are important but a great project originates in a *good business case!* That means that you must tell a story such that the results of your analysis would be ACTIONABLE. It isn't about just applying some technique on some data that we see in the class.

For example, deriving "insights about consumer behavior that can be used by the strategy group to better target" is NOT ACTIONABLE. Using the predictions of your model to allocate marketing spend proportional to the predicted likelihood of leaving the company IS ACTIONABLE.

You need to evaluate the ACTION, *not the model*. This is one of the hardest parts and requires careful thinking. Obviously you do not know what will happen – but you need to try and 'simulate' with test data what is most likely going to happen.