1) What is the problem you want to solve?

I want to help people find films that will: inspire them, challenge their view of the world, make them think and make them feel – 'Life Altering Films' (LAF).

2) Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

The clients are anyone who likes to watch films. Many of these people including myself and several friends and family care directly about this problem, because it can take a long time to find one's LAFs (through websites and other recommender systems) and many of these turn out to be 'false positives'. Additionally, many people will never find these LAFs, and perhaps miss out on opportunities they might have had to contribute to the world in ways they never new possible. In the later case this is even worse since people won't even know what they are missing out on. People will watch more LAFs that they wouldn't have otherwise had the opportunity to do so given time constraints for finding these films.

3) What data are you going to use for this? How will you acquire this data?

There are several public data sources available; the one primarily used will be from MovieLens.org (film survey data) which contains user film ratings, and potentially supplementing this with other movie data sets including: omdb.com (Open Movie Database) and imdb.com (Internet Movie Database). The data can be downloaded in csv files, in some cases, there may need to be some scrapers written (i.e. imdb) however I will explore csv information first so as to not blow-up the scope of the project.

4) In brief outline your approach to solving this problem (knowing that this might change later)

The leap of faith assumption is that different people will have different LAFs for different reasons (every person is unique). Films are a piece of art, which is subjective, meaning there aren't just 'great' films or LAFs for the masses, but the LAFs will be relative to the individual.

I aim to ask 10 or so of my friends/family to rate 100+ films from a list (that I will generate) of 5000 of the top rated films. Top-rated as per the results I will find from aggregating MovieLens data.

The next step will be to aggregate, and analyse the results to try to discover what patterns are emerging from these 10 users. For example some users might have a high proportion of highly rated films by Martin Scorsese, or some might have a high proportion of films in the year 1995 (which might have been a good year for films, or it might be because they were 15 years old at the time). Some people might rate films higher that have a 'romantic' genre within them.

An output of this analysis will be a profile for each user that will contain the predictor variables for that user. For example Bob might have predictor variables – 'Director', 'Genre', 'Year' whereas Alice might have 'Actors', 'Genre', 'Length'. A Summary of this profile will also be generated and given back to each user for feedback.

I will also attempt to create profiles of the users from MovieLens data (It may be a reduced subset of this data initially during experimentation). Then I will match these users with my users who have a similar profile, and compare the films and variables. The films from these MovieLens matched users will form the first version of the recommender system. Initially this will be done without machine learning (more like a constraint matching engine) and later I aim to hopefully build predictive models and move into machine learning- this will depend on the scope and results of the analysis work.

5) What are your deliverables? Typically, this would include code, along with a paper and/or slide deck.

There will be code for the following:

- Opening and setting up datasets to be used. This might include aggregation of ratings, and constructions and linking of dataframes between datasets. Linking by films titles may cause difficulties.
- Generating the list of films to be reviewed. A challenge will be to make sure the lists are likely to have many films the users has seen (i.e. will need to filter by year, language etc.)
- Aggregation and analysis of what are the patterns, and which variables are appearing most often if any.
- Generating profiles for each user.
- Creating matching engine and/or machine learning models to recommend films to users.

There will be a short paper explaining the results, along with some data stories for some of the users' profiles.