**Project Title: The Sweet Science: Using XGBoost to predict the outcome of boxing matches**

**Application domain and datasets (150 words):**

The application domain will be focused on boxing, specifically the top 50 male boxers within each weight category (heavyweight to minimumweight) to identify the winner of upcoming boxing bouts (using XGBoost).

The data sets I am using I have scraped from [www.boxrec.com](http://www.boxrec.com), the largest boxing related database on the internet.

The first data set has details on 793 rows and 20 columns of data comprising of data such as win/loss/draw as well as information on debut date, division, division rating, and birthplace. Unfortunately, after scraping the data it became clear there are fields with missing values, however I have replaced these values where reasonable (such as missing reach data).

The second data set has also been scraped from [www.boxrec.com](http://www.boxrec.com) and contains information on every bout of the respective fighters in the first data set, from their first to their most recent bout, including data such as outcome (by method of victory or loss/round), bout date, and locations, and is just under 80,000 data points.

**Well-motivated analytical questions (150 words):**

* Are good “bigger” boxers better than good “smaller” boxers – a maxim in boxing that boxers in the same weight category with a height advantage will tend to win versus shorter opponents.
* What features are the most important (for this within my data set) when it comes to identifying the winner of a boxing out,
  + And do these features change when looked at on a per-division basis?
* Is there a ranking system by which we can evaluate a boxer’s record – i.e whether they have been “padded”, that is fought much weaker opponents, and how best to weigh those values/rankings.

**Plan (maximum 200 words):**

What is your plan? Include data processing, data derivation, model building and validation. The aim of your plan is to address your research questions.

1. Scrape data from www.boxrec.com (completed)
2. Clean the data to obtain the desired set, including filling null values if they exist (completed)
3. Use a random forest classifier and permutation to identify most important features within the data set
   1. Likely to be time since debut, height, reach.
4. Use XGBoost classifier to evaluate boxing bouts that have occurred in the time since the data sets were scraped.

**Specific questions to us**

1. Considering that a person’s reach is generally related to their height, is it reasonable that I am calculating the average difference between height and reach per division, and replacing missing reach values with their height + that average difference value?
   1. Similarly, I have calculated a mean height value by division and used this for missing values, however I think this could lead to inaccuracies.
2. Is the scope of the project too large for the coursework? If so, what might be a good way to narrow it down, or what area to focus on instead?
3. What other classification methods would you recommend checking out/try when looking to identifying potential winners of bouts, as well as for creating a ranking system for boxers with padded records.