**Data**

**Acquiring Data**

In order to assess the trends of momentum over the course of a match, we required detailed point-by-point data of a large number of tennis matches. While tennis is iterative and governed by relatively simple scoring, certain unique conditions can only occur a few times in each match. A large number of matches permits us to test the effects of several of these circumstances.

Each year, there are hundreds of professional tennis tournaments awarding millions of dollars in prize money. Yet, the pinnacle of the tennis calendar is falls on four Grand Slam events. These four tournaments – the Australian Open (held in January), French Open (May), Wimbledon (July), and US Open (September) – offer players the most ranking points, publicity, prize money, and prestige. Players are always incentivized to exert effort towards victory, but the incentives are never higher than in the Grand Slams. Accordingly, players structure their complex schedules of training, recovery, and competition in order to maximize their performance at these four events. For these reasons, Grand Slam events serve as the ideal grounds for empirical studies of momentum. Grand Slam events are relatively rare and valuable, so all players should be incentivized to put forth substantial effort. Moreover, Grand Slam events possess significantly more resources and content than smaller tournaments. The Grand Slam events each partner with tracking services like IBM or Infosys. This allows for accurate and detailed information of each point.

We acquired point-by-point tennis data from Jeff Sackmann’s github page titled “Grand Slam Point-by-Point Data, 2011-present.” He scraped the data “from the four Grand Slam websites shortly after each event.” We used data from both men’s and women’s singles matches from all four Grand Slams. In 2018, the Australian Open and French Open shifted their partnership from IBM to Infosys. Infosys tracks points differently and has fewer features. This complicates the integration of the data, so we chose to only include matches from the Australian Open and French Open from 2011-2017. In total, this includes 7917 matches and over 1.4 million points.

IBM tracks several useful calculations for each point. Some of the important variables include elapsed time, serve speed, rally length, distance run, serve depth, and return depth. Not all courts, however, have been equipped to calculate all of these complex measurements. In general, the most recent tournaments and later rounds have much more complete measuring systems. Limit scope of research?

This point-by-point data for each match lacked important biological and competitive player level information. Each player carries unique tendencies, strategies, and skill levels into a match. These traits are difficult to quantify, but they can impact the changes of momentum in a match. Perhaps most importantly, a strong player could secure a long streak of victories over a weaker player simply because he or she is a superior player (Page & Coates, 2017). Thus, without an appropriate adjustment for skill level, an analysis of momentum can falsely attribute positive momentum to a player that is simply more skillful than his or her opponent. An adjacent page of Jeff Sackmann’s github begins to resolve this issue by presenting basic information of each player at the moment of every match. Among other factors, this page links a player’s ranking, height, handedness, and age at the time of each tournament to each match. A player’s ranking provides a rough- but often inaccurate- estimate of his or her skill level entering a match. A player’s ranking is composed of his or her success on the professional tour in the past twelve months. The ranking does not account for the proximity of success, the skill level of previous opponents, or the closeness of matches. In other words, a dominant victory against a top player two weeks prior can be weighted the same as a close victory over a weak player 11 months ago. The ranking can also be slow to react to developments like injuries and age. These inconsistencies allow for luck- or misfortune- to inhibit a players’ ranking from representing their skill level.

A more accurate estimate for the skill level of each player is the match’s betting odds. The match’s betting odds is a pre-match assessment of both player’s likelihood of victory. Sportsbooks have developed models and are incentivized to react to recent results and developments much faster than a player’s ranking. For strategic and personal reasons, certain players perform better on different tournaments. Betting odds can account for these nuances. We acquired the pre-match betting odds for each match from Tennis Data. Tennis Data features betting odds from two different sportsbooks: Bet365 and Pinnacles Sports. We added both to our data set.

**Major Transformations**

The point-by-point data came in the form of 106 csv files and the betting odds data came as 78 csv files. Much of the data wrangling and transformations with the data involved combining the files and synthesizing different measuring methods. We cleaned the variable names and joined all the data sets into one. In the end, we had one data set on a point level full with information about the match (i.e. betting odds) and information about each player (i.e. age, height, handedness, etc.). The dataset follows Hadley Wickham’s tidy data format.

It is worth noting that many of the player’s names in each file contained different spellings. These spellings often changed from year to year at the same source. Well known American tennis player Coco Gauff, for example, is also addressed as Cori Gauff in one file. Overall, the data included 110 name discrepancies. To resolve these problems, we synthesized the names into one spelling.

Each match included one or two extra empty observations at the beginning of each match. They did not include any extra information, so we removed them from our data. Moreover, the 2016 US Open calculated the players’ distance run in feet, while all other tournaments use meters. We converted the 2016 US Open into meters.

Second, we made several important variable transformations. Initially, the variables containing the score-line for each observation contained the score after the point had been played. We shifted these results back, so that each observations’ score-line now holds the score before the point is played. Here is a hypothetical example. After the second point of a match, the score is 30-0 and after the third point the score is 40-0. We transformed the data, so that the third point will display 30-0, because those are the scoring conditions in which the third point is played. At the exact moment that the players are competing in the third point, the score is still 30-0. This same processed was emulated for the game score and set score.

We also created several important indicator variables. The **tiebreak** indicator variable holds a value of 1 when a tiebreak is currently being played. This is important, because tiebreaks follow a very different structure than the rest of play. Tiebreaks are also found to have a disproportional effect on momentum (Page & Coates, 2017). The **retired** indicator variable holds a value of 1 when either player leaves a match early. Typically, this comes with an injury or disqualification. These matches are not complete, so they could provide tricky for our analysis. The retired indicator variable provides a mechanism for quickly identifying such matches. The **game\_victor** and **set\_victor** indicator variables hold values of 1 at the point where a player wins a game or set, respectively. Moments like these hold an outsized weight in a tennis match. They are both markers of progress and potential catalysts for momentum.

One of the most important moments in match are a player’s chances to win a game on his or her opponent’s serve (Klaasen & Magnus, 1998). The indicator variables **p1\_break\_pt**, **break\_converted** and **break\_saved** describe these chances. **P1\_break\_pt** holds a value of 1 when player 1 has a chance to win a game while player 2 serves. In other words, if player 1 wins that point, he will “break” player 2’s serve. **Break\_converted** holds a value of 1 when the returner breaks the opponent’s serve or “converts” the break point opportunity. Conversely, **break\_saved** holds a value of 1 when the server wins a break point or “saves” the break point. In general, break point chances are abnormally influential points. These indicator variables provide us the opportunity to analyze the impact of both converting or saving a break point.

Lastly, we derived several variables related to the time of a match. The variable **time\_diff** captures the number of seconds between the start of one point and the start of the previous point. This variable includes both the time playing the previous point and the time that players spend preparing for the current point. The **interruption** indicator variable holds a value of 1 for the point after any natural interruption of play. Players rest after the conclusion of a set, and after the third, fifth, seventh, ninth, and eleventh games of a set. The **change\_ends** indicator variable holds a value of 1 for the point after players switch ends of the court. Players change ends after every interruption and after the first game of the set. Players also change ends after the sixth, twelfth, eighteenth, etc. points in a tiebreak. The indicator variable **delay** captures long delays in play that are unnatural. These delays are typically caused by weather, but lengthy injury timeouts, player breaks, or other disruptions can also delay a match. **Delay** holds a value of 1 when play is stopped for longer than fifteen minutes.

**Faulty Data**

In general, IBM’s data collection methods produced very little noticeable missing or inaccurate observations. Yet, some courts were able to track more variables than others. In most cases, the data converted the missing variables into NA values. For select matches, however, the **distance\_run**, **rally\_count**, and **speed\_mph** variables held values of 0 instead of NA. This disrupts many important numerical calculations. We converted these 0 values into NAs.

Missing observations in the data set are rare, but they typically have a disproportional impact, because they disrupt the scoring of the entire match. In most cases, if the data skips an observation, then the measuring system’s scoring will continue to calculate the score as if the point never occurred. This can lead to odd mistakes, where the measuring system shows that two players have a completely different score line than in reality. Our data included 21 matches that contained a missing point that disrupted the scoring. We could have chosen to fix these scoring errors, but given the abundance of our data, we chose to remove them from the data set. 17 of the 21 faulty matches came from the French Open. Every French Open tournament from 2011 to 2015 included at least one faulty match.

Several other matches included missing observations in the middle of the match that did not disrupt the overall scoring. These matches had an observation that was full of NA values for every variable. After some thought, we chose to remove these matches as well. While usable, these matches contain a disruption in the analysis. There are enough other matches to account for these lost matches. Overall, 41 matches were lost an observation in the middle of the match. 38 of these matches came from the French Open. Each other tournament is represented once. 22 of the 41 faulty matches came from the 2014 French Open.

We also removed a match from the 2011 Australian Open, because a player retired after only seven points.

Occasionally, the measuring system stopped tracking a match’s data before the conclusion of the match. We verified that these matches were not the player retirements referenced earlier. 39 matches did not include the final point, and we flagged them as “incomplete.” We chose to retain both incomplete matches and matches with player retirements, because they still contain long uninterrupted sequences of points.

**Tennis Background**

* Hierarchical Scoring
* General Strategy
  + Serving

The rest is on the next section!