SNOOKER ANALYSIS

Sam Sitarski and ACES Snooker Club

Introduction:
The Current State of Snooker Analysis

In the lead up to the 2020 Betfred World Snooker Championship, Betfair.com released an article titled "*World Snooker Championship: Ten year trends point to...*" looking at which competitor had the best statistical fit of a champion. While the article is ultimately a non-offensive piece to get readers in the gambling mood - Betfair is an online betting site after all - the statistical analysis used to make predictions is lacking. "Working on the belief that the best player usually comes out on top no matter the circumstances," Betfair determined the champion-to-be most likely exhibited the following trends:

- **Is in his 30s**
- **Seeded in the top 10**
- **Played at the Crucible at least nine times**
- **Had reached the final before**
- **Didn't go past the quarter-final last year**
- **Has won a ranking event this season**

The players that fit the bill: (seeded #2) Neil Robertson, (7) Mark Selby, and (10) Shaun Murphy. Using the reasoning that Shaun Murphy's rank was lower than most champions and Neil Robertson had won a ranking event more recently than Mark Selby, Robertson was given the edge. Hindsight is 20/20, and we now know Shaun Murphy was upset in the last 32, and Mark Selby beat Neil Robertson on his run to the semi-final, where he'd lose to the soon to be sixth time world champion, Ronnie O'Sullivan.

The results of the 2020 World Snooker Championship shouldn't come as much of a surprise to anyone. O'Sullivan is arguably the greatest of all time, often playing as good as ever even at the age of 44, and can seemingly beat any other player at will. So why was he not captured by our statistical analysis?

Let's go through the stated trends one by one to determine their actual relevance.

Is in his 30s: The intuition behind this trend is that players get better with experience but deteriorate with age, so one would expect a player to be at their best when they have maximum experience before seeing age-based decline - in their 30s! Of course, this trend only holds at a higher level because different players decline at different rates. Players like Ronnie O'Sullivan and John Higgins have maintained their excellence well into their 40s. When comparing a random 30 year old to a random 40 year old, it is rational to expect the 30 year old to be better, but when comparing John Higgins to Judd Trump, age alone is a poor indicator of current level of play.

Seeded in the top 10: This trend is also intuitive in that one would expect that good players are higher seeds, but it is limited in the fact that it only differentiates between the top 10 and the bottom 22. If one says that a player in the top 10 will win, that still leaves roughly 30% of the players at the Crucible to choose from. Furthermore, it says nothing of a match up between two top 10 players. Ultimately, this trend selects players we can be confident are among the best, but it does not explain why they're good.

Played at the Crucible at least nine times: This trend is similar to the previous one in that it primarily captures good players while failing to acknowledge what makes one player better than another. Perennially good players are regularly seeded in the top 16 meaning they automatically reach the Crucible. If a player remains in the top 16 for a long enough period that they reach the Crucible at least nine times, we can be pretty confident they're a good player, but Judd Trump did not win the 2019 World Snooker Championship because it was his ninth Crucible appearance, rather, he won because he's a good player, arguably the best that year.

Had reached the final before: Perhaps you're picking up on a trend about these trends because this one is similar to the previous two. The old statistical adage is correlation does not equal causation. Reaching the final and being a good player are highly correlated, however, reaching the final does not make a player good.

Did not go past the quarter-final last year: This trend has a couple problems and very little actual information. First, this trend is a product of its sample. If we expand the sample to include all modern era (since 1977) world championships at the Crucible, we might find the opposite effect. Steve Davis and Stephen Hendry regularly made it past the quarter-final in the season before they won because they were repeat winners. Secondly, if one argues that today's game has more parity and thus fewer repeat winners, then this trend becomes a simple matter of probability. There are only 4 players who make it past the quarter-finals each year compared to 28 who don't. If today's game has more parity, then we could expect the following year's winner to come from the group that has more players.

Has won a ranking event this season: Once again this trend captures good players without actually determining why they're good. Good players are seeded. In order to become seeded, a player has to win ranking events. Therefore, good players win ranking events.

Beyond the six trends listed above, the Betfair article actually mentions one more trend to rule out the possibility of Judd Trump repeating as world champion: the fabled Crucible Curse. No first time winner has ever successfully defended their title the following year, often being upset earlier than expected. As far as trends go, this one is far more interesting as it is less reliant on intuition and more reliant on actual data, and it will be discussed in a later section.

In the end, the failure of the trends in the Betfair article is that they work under the assumption that good players win, so they focus too heavily on which players are good and not enough on why those players are good, leading to little distinction between good players.

Part One:
The Art of Break Building

In order to understand what makes one player better than another, we first have to consider the most basic aspects of the game.

What is the goal of a player?
To win a match.

How does one win a match?
By winning the majority of frames.

How does one win a frame?
By scoring more points than the opponent.

Snooker is a sequential game meaning the first player takes a turn at the table, followed by the second player. There are no interceptions or turnovers that would result in one player having a significantly different number of opportunities to score^. Using turns at the table we can continue:

How does a player score more points than the opponent?
By scoring more points per turn at the table.

How does a player score more points per turn at the table?
By maximizing player points per turn and minimizing opponent points per turn.

If we assume that there are a only 147 points on the table* then the maximum points per turn that a player can achieve is 147 by running a maximum on their first turn at the table. The minimum points per turn a player can achieve is 0 if the player fails to pot an object ball. In this case, maximizing player points per turn is also minimizing opponent points per turn. Finally, points per turn can be

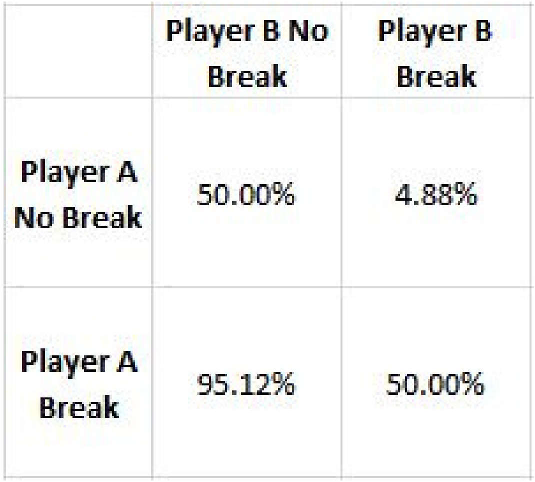
^We are counting having a player who just fouled and was forced to play again as a continuation of their turn at the table, not a separate turn.

*This assumption is not quite true since players can score through fouls without removing object balls from the table. However, the percentage of frames where the total score was greater than 147 is statistically insignificant.

rewritten as a more common phrase - break building - leading to the ultimate goal of the player:

What is the goal of a player?
To maximize break building.

Somewhere along the way of the bookkeeping of snooker the decision was made that breaks of less than 50 points were not worth recording. This, of course, defies logic because a difficult 49 clearance in the deciding frame is more noteworthy than a simple 50 break early in the first frame, but the record books say otherwise. Still, using the admittedly limited tournament, match, and frame data available from 1982 until the end of 2019^, we can break down frames based on whether a player achieved a break of 50+ or not.

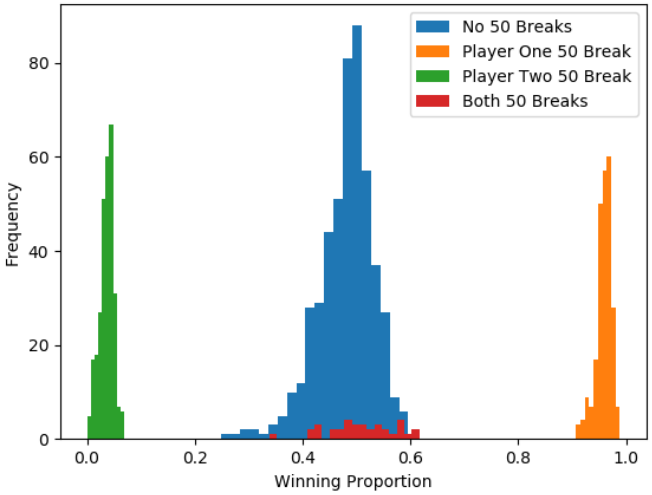


	Player B No Break	Player B Break
Player A No Break	50.00%	4.88%
Player A Break	95.12%	50.00%

The percentage listed is Player A's
winning percentage given the situation

Three observations to note. First, if a player gets a break of 50+, they are a near lock to win the frame. Subsequently, if a player allows a break of 50+ to their opponent, they are a near lock to lose the frame. Second, the diagonals of the table sum to 1. Conceptually this makes sense. Every frame has one winner and one loser. If neither player gets a 50+ break or both players get a 50+ break, the players end up in the same quadrant - 1 or 4. If only one player gets a 50+ break, the players end up in opposite quadrants - 2 or 3. Therefore, the diagonals capture every winner and loser from the matches that populate them. Third, the results of the table do not reflect an individual player's ability to perform in a given situation but rather the overall performance of the players as a whole. This point is better illustrated by the following graph which shows the distribution of individual player winning percentages based on different situations (based on a minimum of 100 situation frames per player).

^Data was collected from Kaggle user 'rusiano' and CueTracker.net



Here we can see that even the best players at winning when there are no 50+ breaks are nowhere near the worst players at winning when they get a 50+ break and their opponent doesn't. We can also see that there are very few frames in which both players achieve breaks of 50+, further evidence that maximizing player break building also minimizes opponent break building. Due to the relative obscurity of frames in which both players achieve breaks of 50+ and the fact that the distribution appears to mirror that of winning percentages when neither player achieves a 50+ break, we will be treating them as the same from here on out.

Individual player winning percentages present an interesting problem in that there is no guarantee that the winning percentages add up to 1. Consider a frame between Shaun Murphy and Mark Williams where neither player gets a break of 50+. In this case Shaun Murphy has a winning percentage of 53.93% and Mark Williams has a winning percentage of 58.95% for a total of 112.88%. Using the following formula we can find a player's percent chance to win given winning distributions:

$$FrameWinningPercentageA = \frac{WinningPercentageA - WinningPercentageA \times WinningPercentageB}{WinningPercentageA + WinningPercentageB - 2 \times WinningPercentageA \times WinningPercentageB}$$

Plugging in the numbers for Shaun Murphy and Mark Williams we find:

$$ShaunMurphyWinningPercentage = \frac{.5393 - .5393 \times .5895}{.5393 + .5895 - 2 \times .5393 \times .5895} = .4493$$

Therefore, Shaun Murphy, despite having a 53.93% overall chance to win a frame in which neither player achieves a break of 50+, will only have a 44.93% chance of winning such a frame against Mark Williams.

^Percentage of frames that won't have a 50+ break given both players have an average 50+ break percentage of 50% is calculated as (1 - .5) * (1 - .5) = .25.

In Judd Trump's 2019 - 2020 historically great season, he achieved 329 breaks of 50+ in 700 frames, meaning he had a 50+ break percentage of about 47%. In 2019 - 2020 Ronnie O'Sullivan had 203 breaks of 50+ in 422 frames for a 50+ break percentage of about 48%. As such, it would seem that the current upper bound for 50+ break percentage is roughly 50%. In a match between the best two break builders, we would about expect to see a frame without a 50+ break about once every four frames^. However, not every match is between the best break builders. In a match between two players with a 50+ break percentage of 30% - which is still extremely good - we would expect to see a frame without a 50+ break about once every two frames*.

So what makes a good player good? The most important feature of a good player is being able to win matches by maintaining a strong break rate. However, even among the world's best break builders there are still a large amount of frames where neither player achieves a break deemed worth recording. As such, a notable but less important feature of a good player is the ability to win frames that don't involve large breaks.

Part Two:
Why the Best Player Doesn't Always Win

Apophenia is a human tendency - commonly seen in gambling - to seek or perceive patterns among unrelated or random events. Upon seeing 2, 4, 6, and 8 win on consecutive rolls on a roulette wheel, a betting man may wager with absolute confidence that 10 will win next when the actual odds of such an occurrence remain at about 2.8%. While snooker is by no means as random as a roulette wheel, randomness is still part of the game. Events like a kick or a fluke are out of the player's control but can have a very real impact on the position of the balls or the result of the frame. A player can climb out of bed feeling stiff and not play as well as expected, or wake up feeling great and play the match of their life. If randomness never occurred, upsets would never happen. Yet from 2009 to 2019, the higher seed at the World Snooker Championship only won about 67% of the time.

Still, when a player lifts a trophy, we assume they were best player in the tournament, which can lead to some wacky mental gymnastics. The best players at the World Snooker Championship from 1990 to 1996 were Stephen Hendry, John Parrott, Stephen Hendry, Stephen Hendry, Stephen Hendry, Stephen Hendry, and Stephen Hendry respectively. John Parrott played phenomenally in the 1991 Snooker World Championship and was well deserving of the win. However, a wider perspective would suggest that Stephen Hendry was actually the best player in the tournament.

Consider the following thought experiment: Two players - A and B - both win 50% of the time when neither player or both players achieve a 50+ break, win 95% of the time when they achieve a 50+ break and their opponent doesn't, and win 5% of the time when they don't achieve a 50+ break and

*Percentage of frames that won't have a 50+ break given both players have an average 50+ break percentage of 30% is calculated as (1 - .3) * (1 - .3) = .49.

their opponent does. The two players have both done so while facing the same level of opponent, exactly league average. If players A and B were to repeatedly play a best of 35, how much better does player A's 50+ break percentage have to be before we can say player A will win at least 75% of the time?

Let's begin by considering the binomial distribution which is used to model repeated trials of situations where there can only be two outcomes (i.e. win or lose, heads or tails, true or false). In our hypothetical situation, there are 35 trials and we want to know how often player A will win 18 or more of them given different frame winning probabilities. This problem can be modeled as:

$$P(X \geq 18) = 1 - \frac{35!}{(35-17)! \times 17!} \times (1-p)^{17} \times (p)^{35-17}$$

Using the binomial distribution we can calculate that for player A to win a best of 35 around 75% of the time, they need to win about 55.55% of the frames against player B. Using the winning percentage formula above and algebra, we can calculate the difference in winning percentages as:

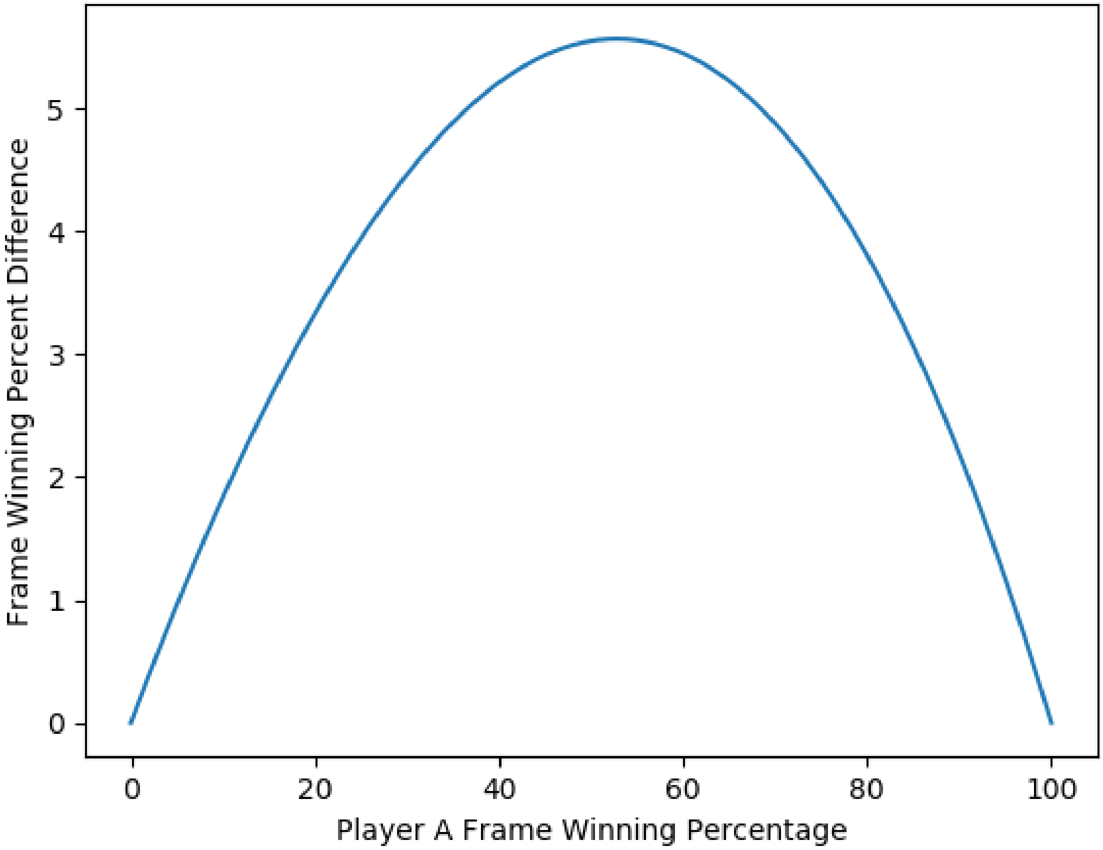
$$.5555 = \frac{A-(A)\times(A-X)}{A+(A-X)-2\times(A)\times(A-X)}$$

$$.5555 \times (2A - X - 2A^2 + 2AX) = A - A^2 + AX$$

$$X \times (.111A - .5555) = .111A^2 - .111A$$

$$X = \frac{.111A^2 - .111A}{.111A - .5555}$$

Where A is player A's winning percentage and X is the difference between player A's and player B's winning percentages. The following graph shows difference in frame winning percentage needed to win 75% of the time based on player A's frame winning percentage.



Reading the graph above, if player A has a frame winning percentage of 50%, then in order for player A to win 75% of best of 35s played against player B, player A's frame winning percentage needs to be at least 5.5% better than player B's. Next, we need to move from frame winning percentage to 50+ break percentages.

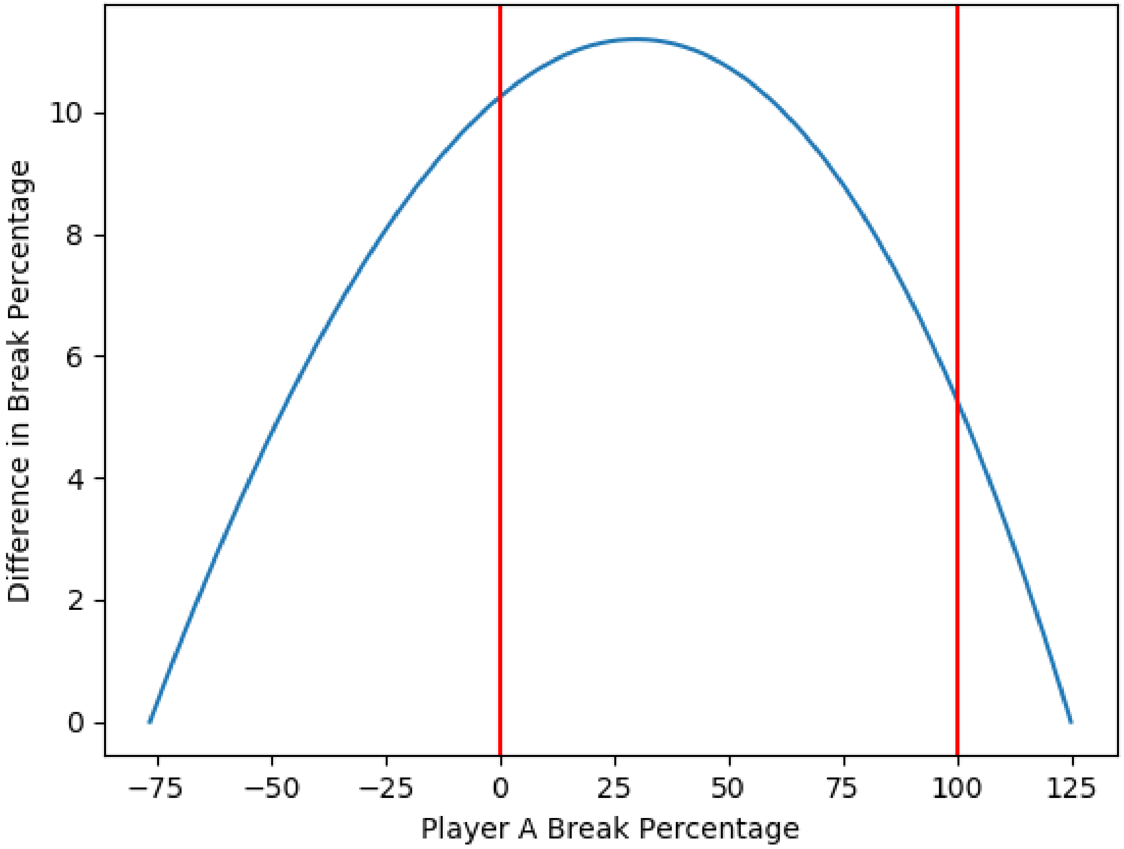
From the 2014 - 2015 season until the end of the 2018 - 2019 season the league average for 50+ break percentage was about 24.18%, and we'll say the league average is the same as the table listed in part one. Therefore, our hypothetical players would win 50% of the time when both or neither achieves a 50+ break, 99.73% of the time when they achieve a 50+ break and their opponent didn't, and 0.27% of the time when their opponent gets a 50+ break and they don't^. Now we get the following equations:

$$A = .5 \times (1 - Rate) \times (.7582) + .9973 \times (Rate) \times (.7582) + .0027 \times (1 - Rate) \times (.2418) + .5 \times (Rate) \times (.2418)$$

$$A - X = .5 \times (1 - Rate - x) \times (.7582) + .9973 \times (Rate - x) \times (.7582) + .0027 \times (1 - Rate - x) \times (.2418) + .5 \times (Rate - x) \times (.2418)$$

With a little more algebra we can calculate the difference in 50+ break percentage needed to win at least 75% of the time given player A's frame winning percentage. The following graph shows that:

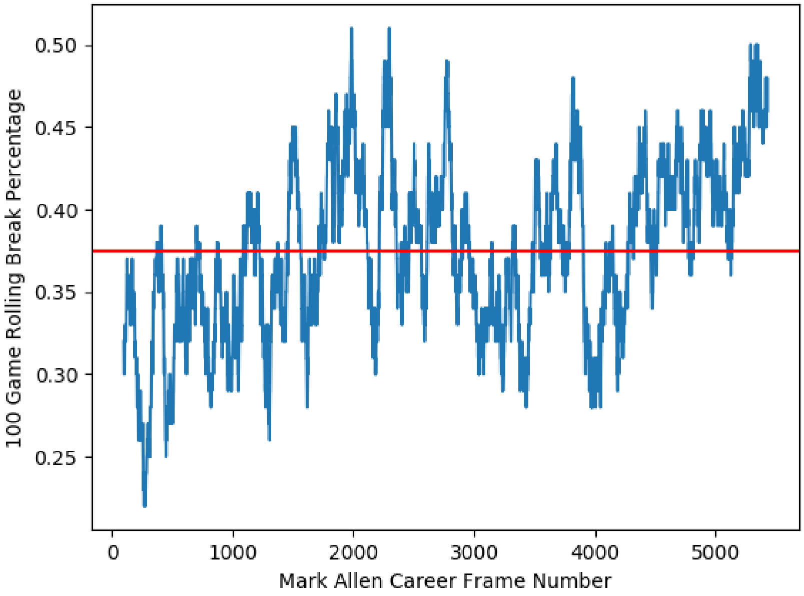
^The calculation to find 99.73% and 0.27% were done with the winning percentage formula in part one.



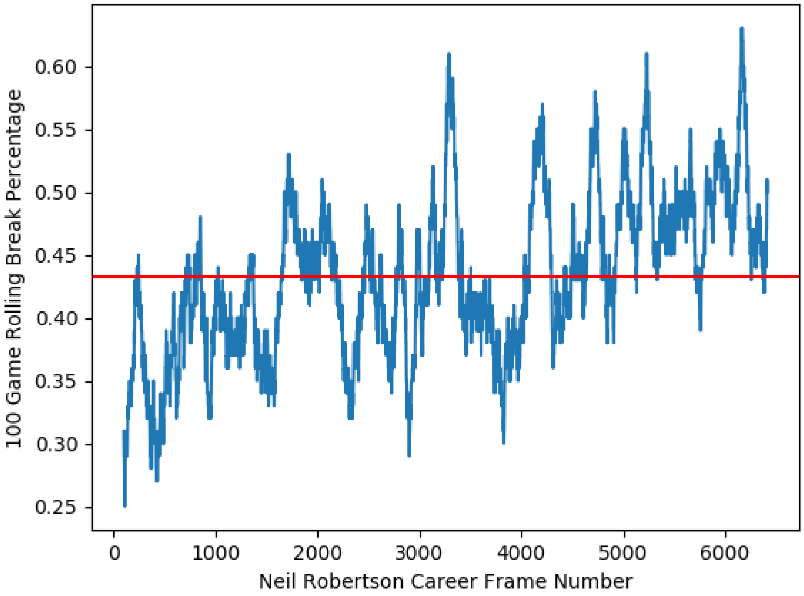
Reading the graph above, if player A has a 50+ break percentage of 50%, then in order for player for player A to win 75% of best of 35s played against player B, player A's 50+ break percentage needs to be at least 10% better than player B's.

There are a few important things to note. First, in this hypothetical, player A's frame winning percentage can't actually fall below 38% or rise above 87% or else their 50+ break percentage has to be below 0% or above 100%. Second, looking at the graph above, we can see that the floor for the difference in 50+ break percentage is around 5%. Considering Judd Trump and Ronnie O'Sullivan's 50+ break percentages of 47% and 48% respectively, and all else being equal, we can't be sure who's going to win the World Snooker Championship finals!

Of course, reality isn't a hypothetical thought experiment. However, there are aspects of reality that can lead to random variation, namely streakiness. During any given frame, either player can get "hot" or "cold" resulting in them playing above or below their expected level. Looking at the 100 frame rolling average graph of a player's break rate shows this clearly.



There may be some evidence to show that Mark Allen has improved his 50+ break percentage throughout his career, but the cyclical pattern is still obvious, fluctuating above and below his career mean. This phenomenon is not unique to just Mark Allen either. Let's take a look at Neil Robertson^:



^For both Mark Allen and Neil Robertson, frames from the 6-Reds World Championship were not included because of it is more difficult to run a recorded break of 50+ due to the tournament's alternate rule format.

In a match up between Neil Robertson and Mark Allen, which versions of the players would we see? The Neil Robertson with a 50+ break percentage of 60% and the Mark Allen with a 50+ break percentage of 30% or the Neil Robertson with a 50+ break percentage of 35% and the Mark Allen with a 50+ break percentage of 50%?

We already know how important having a large 50+ break percentage difference is to winning a match, so how does streakiness help us in this regard? The truth is not at all. The correlation between a player's last 100 frames and their 50+ break percentage in their next match is actually less than their 50+ break percentage for the season and their 50+ break percentage in their next match. Ultimately we're left knowing that all players have the ability to play above or below their career 50+ break percentage, but we have no way to know when they will do so.

Part Three:
A Simplistic Model and the Crucible Curse

Given what we know about what makes one player better than another - the ability to build breaks and the ability to win frames in which both or neither player gets a 50+ break - can we build a model that is both simple and accurately predicts which player will win?

To begin we want to define the predictors used in the model: `Break Rate Difference`, `Match Winning Percentage`, `No Breaks Frame Winning Percentage`, and `Points Per Frame Difference`.

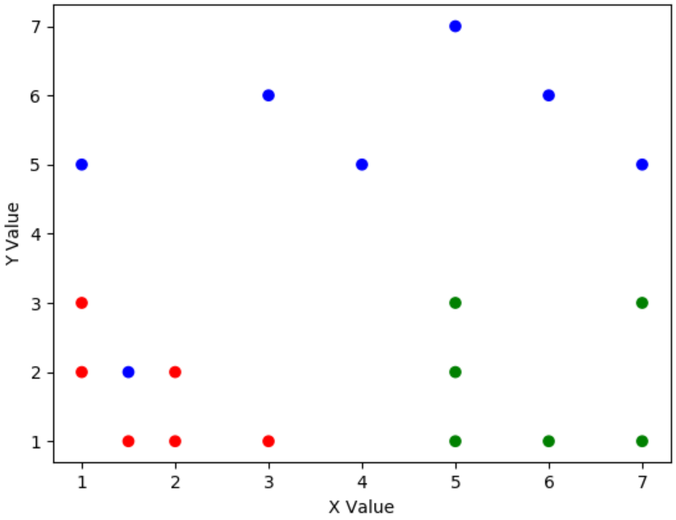
A player's break rate is calculated as the number of breaks achieved divided by number of frames played over the previous two seasons. `Break Rate Difference` is listed as TRUE if player A's break rate is higher than player B's break rate. A player's `Match Winning Percentage` is the number of matches won over the previous two seasons divided by the number of total matches played.

`Match Winning Percentage` is listed as TRUE if a player's winning percentage is higher than their opponent's winning percentage.

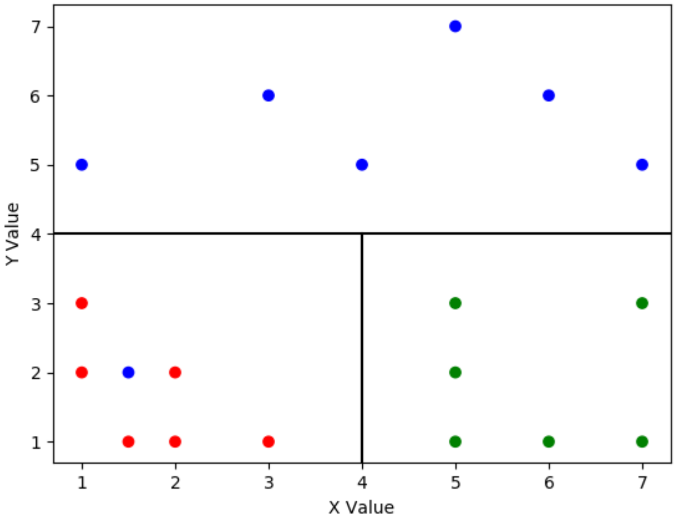
A player's `No Breaks Frame Winning Percentage` is a player's frame winning percentage in frames where neither gets a 50+ break over the previous five seasons weighted heavily towards the most recent season. `No Breaks Frame Winning Percentage` is also weighted for strength of opponent, so wins against weaker opponents don't count as much as wins against stronger opponents. `No Breaks Frame Winning Percentage` is listed as TRUE if a player's winning percentage is higher than their opponent's.

`Points Per Frame Difference` is the difference between a player's points per frame and their opponent's points per frame. `Points Per Frame Difference` is listed as TRUE if a player's points per frame is higher than their opponent's points per frame.

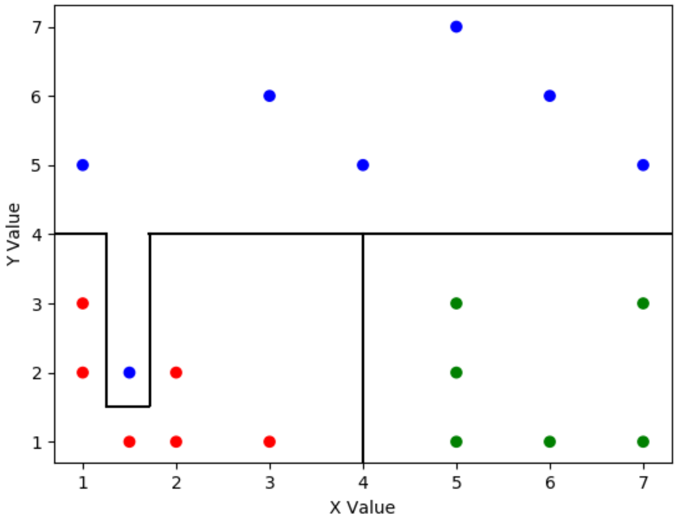
Using data from 1988 - 2019 and looking at matches that are at least best of 11 frames we can calculate these four predictors for every player-match in the sample. Next we split the data into two separate samples: train data and test data. The train data will be used to train the model to look for patterns between the predictors and the target (winning the match), while the test data will be data not previously seen by to model to avoid over-fitting the model to the training data. The model we'll be using is a decision tree, a simple but effective model that uses horizontal and vertical lines to partition the data into similar segments. We'll use a two predictor dummy data set with a target variable of `Color` to help better visualize how a decision tree works.



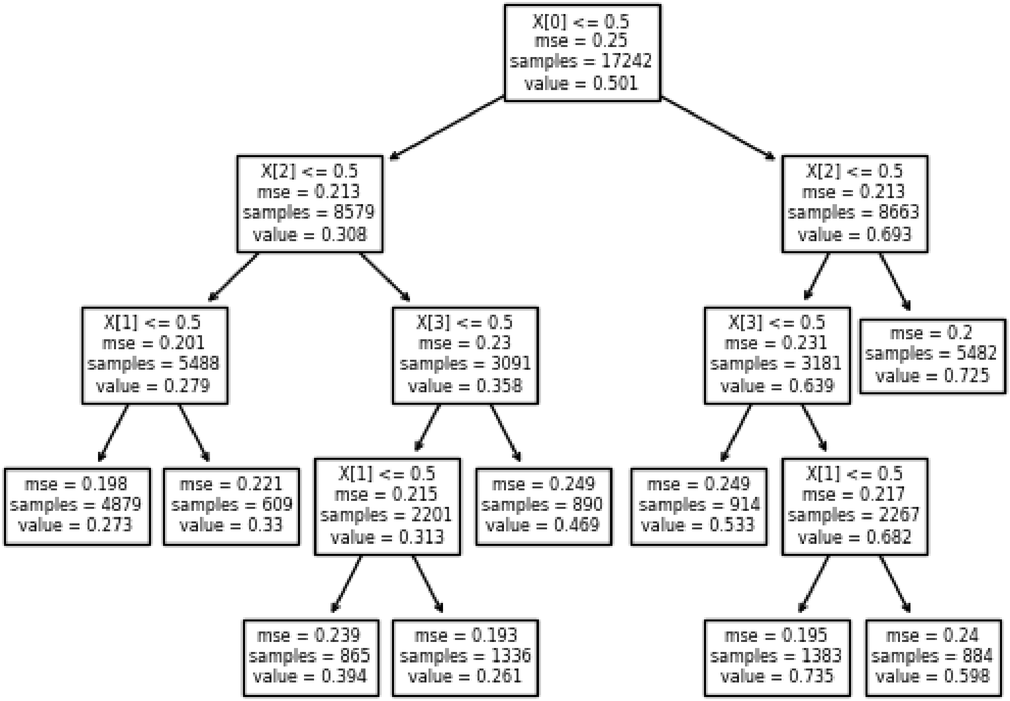
Using horizontal and vertical lines, the decision tree may partition the data as such to try to predict color. Dots that fall in the top section will be predicted as blue, bottom left will be predicted as red, and bottom right will be predicted as green:



However, an over-fit decision tree - which we want to avoid - may partition the data in this manner. Moving forward, we would not expect dots falling in the thin strip to be blue, but the model would predict them as such. Therefore, we need to use the test data to check that model has determined the overarching trends

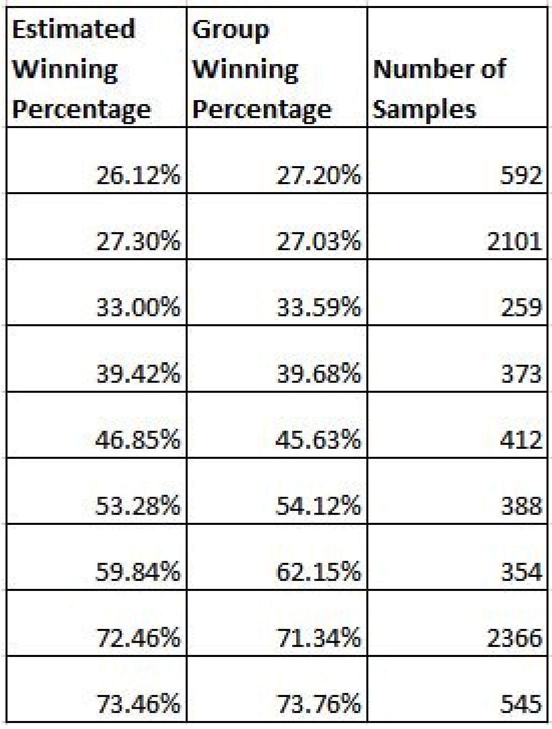


Though our snooker data can't easily be visualized because it's in four dimensions, the process is still the same. The results are the following decision tree which can be visualized as a flowchart:



Where X[0] is `Break Rate Difference`, X[1] is `Match Winning Percentage`, X[2] is `No Breaks Frame Winning Percentage, and X[3] is `Points Per Frame Difference`. Note that the computer interprets a TRUE variable as 1 and a FALSE variable as 0. Therefore, X[0] will be < .5 when `Break Rate Difference` is FALSE. Reading the tree is simple. A data point starts at the top, and moves right every time a variable is TRUE and left when a variable is FALSE. Consider the following data point where `Break Rate Difference` is TRUE, `Match Winning Percentage` is TRUE, `No Breaks Frame Winning Percentage` is FALSE, and `Points Per Frame Difference` is FALSE. Moving down the tree, we would initially move right, then left, then left again reaching a node with a winning percentage of 53.3%.

The model will predict a data point as one of nine different winning percentages based on which final node the data point ends up in. We can assign this prediction to a new variable which we will call `Estimated Winning Percentage`. Next, we can group the data based on `Estimated Winning Percentage`. There are nine groups for the nine final nodes in our model. Finally, we can calculate the actual match winning percentage of each of the groups to see how they compare to the `Estimated Winning Percentage`. The following table shows this:



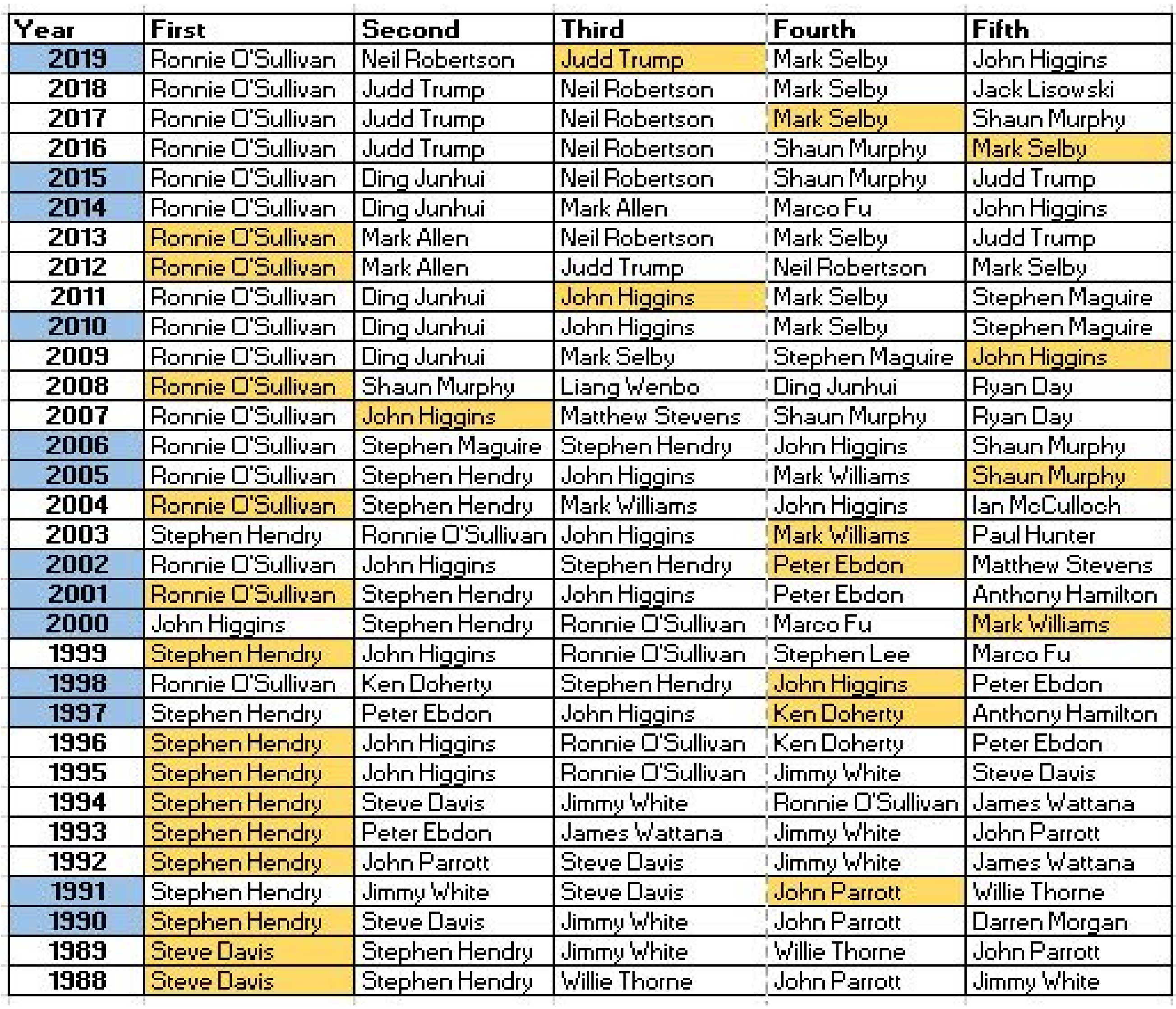
Estimated Winning Percentage	Group Winning Percentage	Number of Samples
26.12%	27.20%	592
27.30%	27.03%	2101
33.00%	33.59%	259
39.42%	39.68%	373
46.85%	45.63%	412
53.28%	54.12%	388
59.84%	62.15%	354
72.46%	71.34%	2366
73.46%	73.76%	545

By using four TRUE - FALSE predictors, it would appear that our model has pretty accurately captured what is happening on the table.

Now that we are confident that our model can make pretty accurate predictions, we can move onto looking at the World Snooker Championship. The probability that any given player will win it all is given by the following formula:

$$\sum(P(ReachFinal) \times P(OpponentReachFinal) \times P(BeatOpponent))$$

Where P(x) is shorthand for *the probability that event x will occur.* After we calculate every player's probability to win, we normalize the probabilities so that they sum to 100%. After this, we can sort the players based on probability to win to see which players are most likely to win. The following chart gives the favorites to win by year. Winners are highlighted in gold and years where there was a first time winner are highlighted in blue.



Year	First	Second	Third	Fourth	Fifth
2019	Ronnie O'Sullivan	Neil Robertson	Judd Trump	Mark Selby	John Higgins
2018	Ronnie O'Sullivan	Judd Trump	Neil Robertson	Mark Selby	Jack Lisowski
2017	Ronnie O'Sullivan	Judd Trump	Neil Robertson	Mark Selby	Shaun Murphy
2016	Ronnie O'Sullivan	Judd Trump	Neil Robertson	Shaun Murphy	Mark Selby
2015	Ronnie O'Sullivan	Ding Junhui	Neil Robertson	Shaun Murphy	Judd Trump
2014	Ronnie O'Sullivan	Ding Junhui	Mark Allen	Marco Fu	John Higgins
2013	Ronnie O'Sullivan	Mark Allen	Neil Robertson	Mark Selby	Judd Trump
2012	Ronnie O'Sullivan	Mark Allen	Judd Trump	Neil Robertson	Mark Selby
2011	Ronnie O'Sullivan	Ding Junhui	John Higgins	Mark Selby	Stephen Maguire
2010	Ronnie O'Sullivan	Ding Junhui	John Higgins	Mark Selby	Stephen Maguire
2009	Ronnie O'Sullivan	Ding Junhui	Mark Selby	Stephen Maguire	John Higgins
2008	Ronnie O'Sullivan	Shaun Murphy	Liang Wenbo	Ding Junhui	Ryan Day
2007	Ronnie O'Sullivan	John Higgins	Matthew Stevens	Shaun Murphy	Ryan Day
2006	Ronnie O'Sullivan	Stephen Maguire	Stephen Hendry	John Higgins	Shaun Murphy
2005	Ronnie O'Sullivan	Stephen Hendry	John Higgins	Mark Williams	Shaun Murphy
2004	Ronnie O'Sullivan	Stephen Hendry	Mark Williams	John Higgins	Ian McCulloch
2003	Stephen Hendry	Ronnie O'Sullivan	John Higgins	Mark Williams	Paul Hunter
2002	Ronnie O'Sullivan	John Higgins	Stephen Hendry	Peter Ebdon	Matthew Stevens
2001	Ronnie O'Sullivan	Stephen Hendry	John Higgins	Peter Ebdon	Anthony Hamilton
2000	John Higgins	Stephen Hendry	Ronnie O'Sullivan	Marco Fu	Mark Williams
1999	Stephen Hendry	John Higgins	Ronnie O'Sullivan	Stephen Lee	Marco Fu
1998	Ronnie O'Sullivan	Ken Doherty	Stephen Hendry	John Higgins	Peter Ebdon
1997	Stephen Hendry	Peter Ebdon	John Higgins	Ken Doherty	Anthony Hamilton
1996	Stephen Hendry	John Higgins	Ronnie O'Sullivan	Ken Doherty	Peter Ebdon
1995	Stephen Hendry	John Higgins	Ronnie O'Sullivan	Jimmy White	Steve Davis
1994	Stephen Hendry	Steve Davis	Jimmy White	Ronnie O'Sullivan	James Wattana
1993	Stephen Hendry	Peter Ebdon	James Wattana	Jimmy White	John Parrott
1992	Stephen Hendry	John Parrott	Steve Davis	Jimmy White	James Wattana
1991	Stephen Hendry	Jimmy White	Steve Davis	John Parrott	Willie Thorne
1990	Stephen Hendry	Steve Davis	Jimmy White	John Parrott	Darren Morgan
1989	Steve Davis	Stephen Hendry	Jimmy White	Willie Thorne	John Parrott
1988	Steve Davis	Stephen Hendry	Willie Thorne	John Parrott	Jimmy White

The yearly favorite to win should come as no surprise transitioning from Steve Davis to Stephen Hendry to Ronnie O'Sullivan with an appearance by John Higgins for good measure. The `Second` through `Fifth` columns have the remaining elite players who fall just short of the greatest of all time category - Judd Trump, Ding Junhui, Neil Robertson, Mark Selby, and Jimmy White among a few others - as well as a few outliers. Shaun Murphy's 2005 unranked championship run makes an appearance, and Liang Wenbo made some noise before losing to eventual champion Ronnie

O'Sullivan in 2008 which was expected due to his place in the top five that year. There are some misses as well. The model expected a lot from an unranked Marco Fu at the turn of the century only to see him lose in the first round, and a ranked Mark Allen should have done more in 2012 and 2013, but overall, the model predicts the winner in the top 5 in 27 of 32 years. Of the remaining five years, four were won by a first time winner, and the last was won by Mark Williams in 2018 after not participating in the tournament in 2017.

Of the 32 World Snooker Championships the model looked at, 13 were won by a first time winner. Only two of those first time winners were the favorite to win, Stephen Hendry and Ronnie O'Sullivan. The model looked at 12 tournaments in which the defending champion was a first time winner the previous year (The 2020 Betfred World Snooker Championship in which first time winner Judd Trump was the defending champion was not included). Only two of those first time defending champions were the favorite to repeat, Stephen Hendry and Ronnie O'Sullivan. The remaining 10 first time defending champions all required a little luck to win, luck we wouldn't expect to see the second time around^.

All of this is to say that the phenomenon of the Crucible Curse exists because we unjustly assign variables to the defending champion that are out of their control. Still, at this point, there have been 18 first time winners who have failed to defend their title. Should we expect that one of them would get lucky two years in a row?

In order to sort players from favorite to least favorite, our model assigns each player odds to win. The favorite to win is consistently around 1:4 odds to win and the top five typically fall between 1:4 and 1:12. In the 12 tournaments that our model looked at where the defending champion was a first time champion in the previous year, those players averaged about 1:9 odds or a 10% chance to win*. Assuming that the 12 samples are representative of the overall population of first time defending champions, we can use the binomial distribution to calculate that after 18 trials, no first time winner would successfully defend their title in the following year as 15%. While this is rather low, it's still not unexpected. Flipping a coin 3 times and getting three tails doesn't mean we're cursed never to see heads again despite lower probability of 12.5%. Ultimately, the Crucible Curse is a product of the wacky effects of a small sample size and should not be used as justification as to why a player won't win in any particular year.

^To clarify, saying a champion got lucky is not the same as saying a champion is undeserving. Luck is part of the game. The best players and the worst players are all equally subjected to it. Lucky players can still lose and unlucky players can still win.

*Stephen Hendry (1991) and Ronnie O'Sullivan (2002) both had the best odds to repeat at about 1:4 while Stuart Bingham (2016) had the worst odds to repeat at about 1:61.

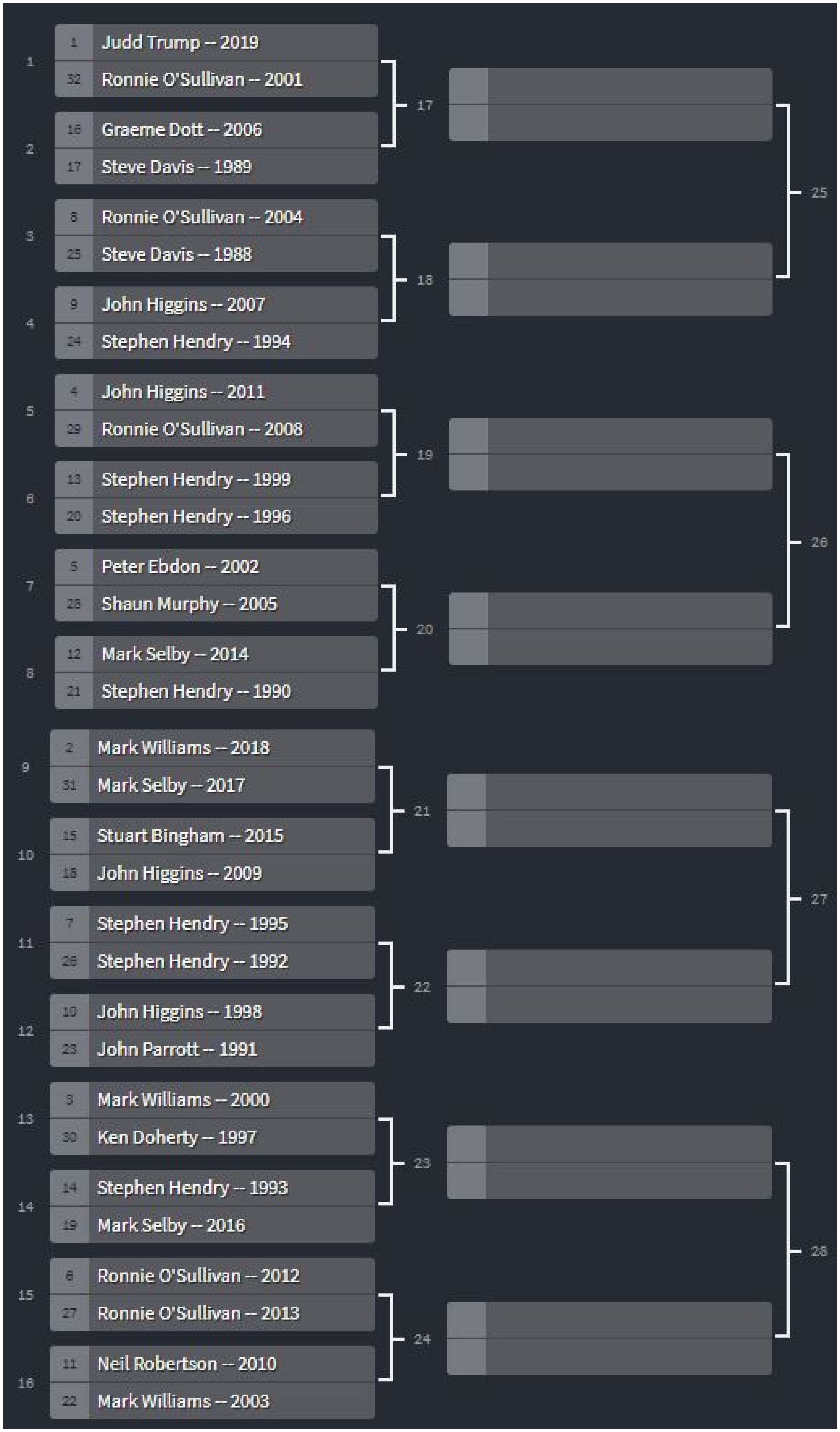
Conclusion:
A Super World Championship of Champions

A significant problem with working in snooker data is the disconnect between how the game is played and how the game is recorded. Consider the following example: two frames of a match between player A and player B are played. In frame one, player A barely misses a shot leaving player B in with an easy first red and a open table. Player B makes the first red, runs a break of 50, and wins the frame. In frame two, player B attempts a safety leaving reds on rail and a difficult first red. Player A pots the difficult red, moves the other reds off the rail into a better position, runs a break of 50, and wins the frame. Which player is the better player?

Because snooker is recorded at a frame level, there is no way to differentiate between these situations. Both players have won one frame and recorded one break of 50 in the process. However, when analyzing the players on a shot-by-shot basis, we would guess player A is better than player B because player B's break of 50 was the product of an easier table when he began while player A was forced to create his break of 50 by attempting difficult shots and moving balls off the rails.

Our analysis was done by trying to extract relevant information from frame level data. However, if shot-by-shot data were to become available, additional information may appear. For example, using frame data, our model can't predict that a player will win more than 73% of the time. If shot data were able to predict players that win 80% of the time, this information may change the World Snooker Championship odds calculated in Part Three.

One final note about the model is that break rates and points per frame were normalized by year before the difference was calculated. Because of this, the model can calculate the winning percentage of players from different years. In other words, we can simulate the first ever Super World Championship of Champions!



1
1 Judd Trump -- 2019
32 Ronnie O'Sullivan -- 2001

2
16 Graeme Dott -- 2006
17 Steve Davis -- 1989

3
8 Ronnie O'Sullivan -- 2004
25 Steve Davis -- 1988

4
9 John Higgins -- 2007
24 Stephen Hendry -- 1994

5
4 John Higgins -- 2011
29 Ronnie O'Sullivan -- 2008

6
13 Stephen Hendry -- 1999
20 Stephen Hendry -- 1996

7
5 Peter Ebdon -- 2002
28 Shaun Murphy -- 2005

8
12 Mark Selby -- 2014
21 Stephen Hendry -- 1990

9
2 Mark Williams -- 2018
31 Mark Selby -- 2017

10
15 Stuart Bingham -- 2015
18 John Higgins -- 2009

11
7 Stephen Hendry -- 1995
26 Stephen Hendry -- 1992

12
10 John Higgins -- 1998
23 John Parrott -- 1991

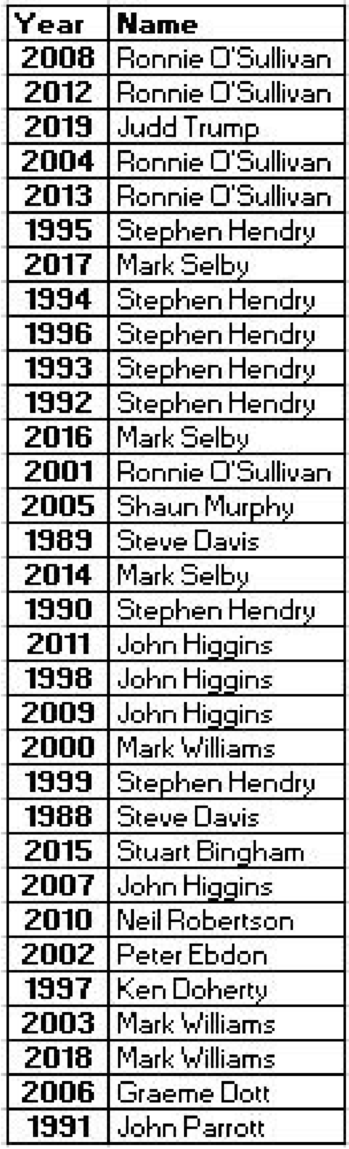
13
3 Mark Williams -- 2000
30 Ken Doherty -- 1997

14
14 Stephen Hendry -- 1993
19 Mark Selby -- 2016

15
6 Ronnie O'Sullivan -- 2012
27 Ronnie O'Sullivan -- 2013

16
11 Neil Robertson -- 2010
22 Mark Williams -- 2003

Seeds were randomly generated, which, unfortunately for Stephen Hendry means four of his seven "World Champion selves" have to square-off against "himselves" in the first round, killing his overall odds of winning the Super Championship. The players listed in order of most probable to least probable ultimate winner are:



Year	Name
2008	Ronnie O'Sullivan
2012	Ronnie O'Sullivan
2019	Judd Trump
2004	Ronnie O'Sullivan
2013	Ronnie O'Sullivan
1995	Stephen Hendry
2017	Mark Selby
1994	Stephen Hendry
1996	Stephen Hendry
1993	Stephen Hendry
1992	Stephen Hendry
2016	Mark Selby
2001	Ronnie O'Sullivan
2005	Shaun Murphy
1989	Steve Davis
2014	Mark Selby
1990	Stephen Hendry
2011	John Higgins
1998	John Higgins
2009	John Higgins
2000	Mark Williams
1999	Stephen Hendry
1988	Steve Davis
2015	Stuart Bingham
2007	John Higgins
2010	Neil Robertson
2002	Peter Ebdon
1997	Ken Doherty
2003	Mark Williams
2018	Mark Williams
2006	Graeme Dott
1991	John Parrott

Of course, these results are meant to be nothing except a fun little experiment. Clearly, to win the World Snooker Championship even a single time among a large field of very talented competitors, each of these players at the peaks of their careers should be considered elite and able to produce superb snooker on any given day. As such, many other variables would come into play when determining our Super Champion. Early match-ups are extremely important, and since seeding is random, a new simulation will likely produce a new champion.

Among snooker enthusiasts, arguments abound over which player is truly the greatest of all time. Though the final answer will always be speculation due to the impossibilty to transcend time, perhaps models such as this will lead to a more analytical discussion.

Betfair.com Article:
https://betting.betfair.com/betting/snooker/world-snooker-championship-tips-ten-year-trends-point-to-170720-719.html

Kaggle User 'rusiano':
https://www.kaggle.com/rusiano/snooker-data-19822020

CueTracker.net:
https://cuetracker.net/