Jia Wei Koh

Jwkoh6@student.monash.edu

The Big Four

An Interactive Visualisation of the Dominance of the Big Four

# Introduction

With the conclusion of the French Open 2019 on June 9th. Rafael Nadal has won a total of 12 French Open titles. Starting from the 2004 Wimbledon to the 2019 Australian Open, they have shared a total of 54 out of the last 60 men’s major singles titles and have appeared in all the major finals apart from the 2005 Australian Open and the 2014 US Open. They also have the top four prize money leaders of all time as well as a litany [other records](https://en.wikipedia.org/wiki/Big_Four_(tennis)#Combined_achievements). Throughout the history of tennis, no four have dominated tennis or achieved success the way the Big Four has. This presentation hence aims to visualise their rivalries with other players on tour, their individual dominance on both aspects of tennis, serves and returns. ("Big Four (tennis)," 27 April 2019; McGarry, 2019)

Table 1: Multiple Grand Slam Finals and Titles Won by Players (2005-2019 French Open)

| Rank | Player | Titles | Finals |
| --- | --- | --- | --- |
| 1 | [Switzerland](https://en.wikipedia.org/wiki/Switzerland) Roger Federer | 20 | 30 |
| 2 | [Spain](https://en.wikipedia.org/wiki/Spain)Rafael Nadal | 18 | 26 |
| 3 | [Serbia](https://en.wikipedia.org/wiki/Serbia)Novak Djokovic | 15 | 24 |
| 4 | [United Kingdom](https://en.wikipedia.org/wiki/United_Kingdom)Andy Murray | 3 | 11 |
| 5 | [Switzerland](https://en.wikipedia.org/wiki/Switzerland)[Stan Wawrinka](https://en.wikipedia.org/wiki/Stan_Wawrinka) | 3 | 4 |

With the recent expansion of [tennis analytics](https://www.atptour.com/en/stats/player-tendencies) by the Associations of Tennis Professionals (ATP), the governing authority of the men’s tour, interests in utilizing data to further analyse a player’s ability to play different styles (At the Net, From the Baseline), and ability to serve and return have increased. Given that [the dataset](https://github.com/JeffSackmann/tennis_slam_pointbypoint), which contains the details of point-by-point of Grand Slam tournaments from 2011 is available and released by Jeff Sackmann, the research here begs the main question:

*“What makes the Big Four so dominant in Grand Slams?”*

These questions can be broken down further into 2 separate questions:

1. Are they better at holding serve than most players?
2. Are they better at returning than most players?

The constructed visualisation aims to explain tennis in its simplest aspects: Serves and Returns. This visualisation targets to bring in a larger audience to appreciate tennis from a different viewpoint, as well as explain tennis in simple statistics to tennis fans, utilising dynamic visualisations with D3 to achieve this goal.

# Design:

The dataset was wrangled into 3 different documents, of which two were comma-delimiter separated files. These files were data on serves and returns of players in Grand Slam Level tournaments. The last file, a JavaScript Object Notation file (JSON) contained the major rivalries and how players performed against each other on the Grand Slam level from 2011 Australian Open.

In the first design sheet, several designs were considered on how to visualise player statistics regarding their individual serves and returns. These included a ranked bar chart, which would rank players from best to worst in each category. A scatter plot which would allow visualisation of a player’s statistic of such as serve speed, return points won on first serve and how it affects the serves and returns specifically. Finally, a bubble plot, which would include the number of matches that a player played on a particular court surface.

For visualising rivalries, the first iteration of the network graph was thought of as a graph which had 2 directions, with the thickness of the edges indicating how many wins a player had against another player while sizes indicated the number of matches played on a surface. The colours were then used to indicate the different surfaces.

The next 3 design sheets focused on finding better designs that may suit the visualisations better. The second design sheet and third design sheet focus particularly on how to improve the network graph as given that there are a lot of interconnectivity between the nodes and while plotting them the overlapping edges may cause an issue of too complex connectivity and too much clutter. Eventually, instead of a bi-directional graph, to simplify the graph design a one directional graph was used instead. The arrows outwards would indicate more wins against the player than losses, while length would indicate to scale how many matches were played against each other. The users are allowed to toggle between different surfaces.

The fourth design sheet, focuses on how to visualise how different variables would affect the win rate of both serves and returns across different surfaces. By adding interactivity and transitions, users are able to visualise changes in both axes without too much clutter. A scatter plot was considered at this point but rather, an interactive bubble chart would be used instead. The bubble size, proportionate to area rather than radius, would indicate the number of matches played across different surfaces as the user toggles between each surface. Area was used instead of radius as using radius would increase the Lie Factor, a term used by Edward Tufte to describe when the graphic proportions do not represent the data proportions accurately. The colours chosen were simple, all surfaces used a colour called light salmon, as it visualises effectively the different data points while not being too bright. Hard court uses the colour of the Australian Open court colour, Clay court uses the colours of the French open, while Grass courts uses the colours of grass. These are used to convey information effectively by evoking a familiarity of each surface colours to the audience, allowing for information to be processed effectively by reducing the load on the short term memory, utilising long term memory instead.

The fifth design sheet, combined the third and fourth design sheet to form a singular narrative, allowing for players to look at the

# Implementation:

# User Guide:

# Conclusion:

# Reflection and Limitations:

In hindsight, for this particular project, this dataset might have been much more complex than originally anticipated. The cleaning and wrangling process while time consuming, remains one of the more rewarding opportunities to learn methods of wrangling and cleaning data. The analysis method ought to be improved upon in tandem with the complexity of the dataset. The dataset provides more opportunities to create and analyse more complex measures than the one currently done. While outside the scope of this unit, logistic regression ought to be used to model and further improve upon the analysis in each aspect of how each individual point are won. For a sports fan and a budding data scientist, this project allows for the learning opportunity to look at tennis from a completely different perspective.

# References and Dataset:

Big Four (tennis). (27 April 2019). Retrieved from <https://en.wikipedia.org/w/index.php?title=Big_Four_(tennis)&oldid=894386723>

McGarry, A. (2019). Andy Murray is set to retire, and the likes of Roger Federer and Rafael Nadal might not be far behind. *Australian Broadcasting Corporation*. Retrieved from <https://www.abc.net.au/news/2019-01-16/get-ready-to-say-goodbye-to-the-big-four-era-in-tennis/10712254>

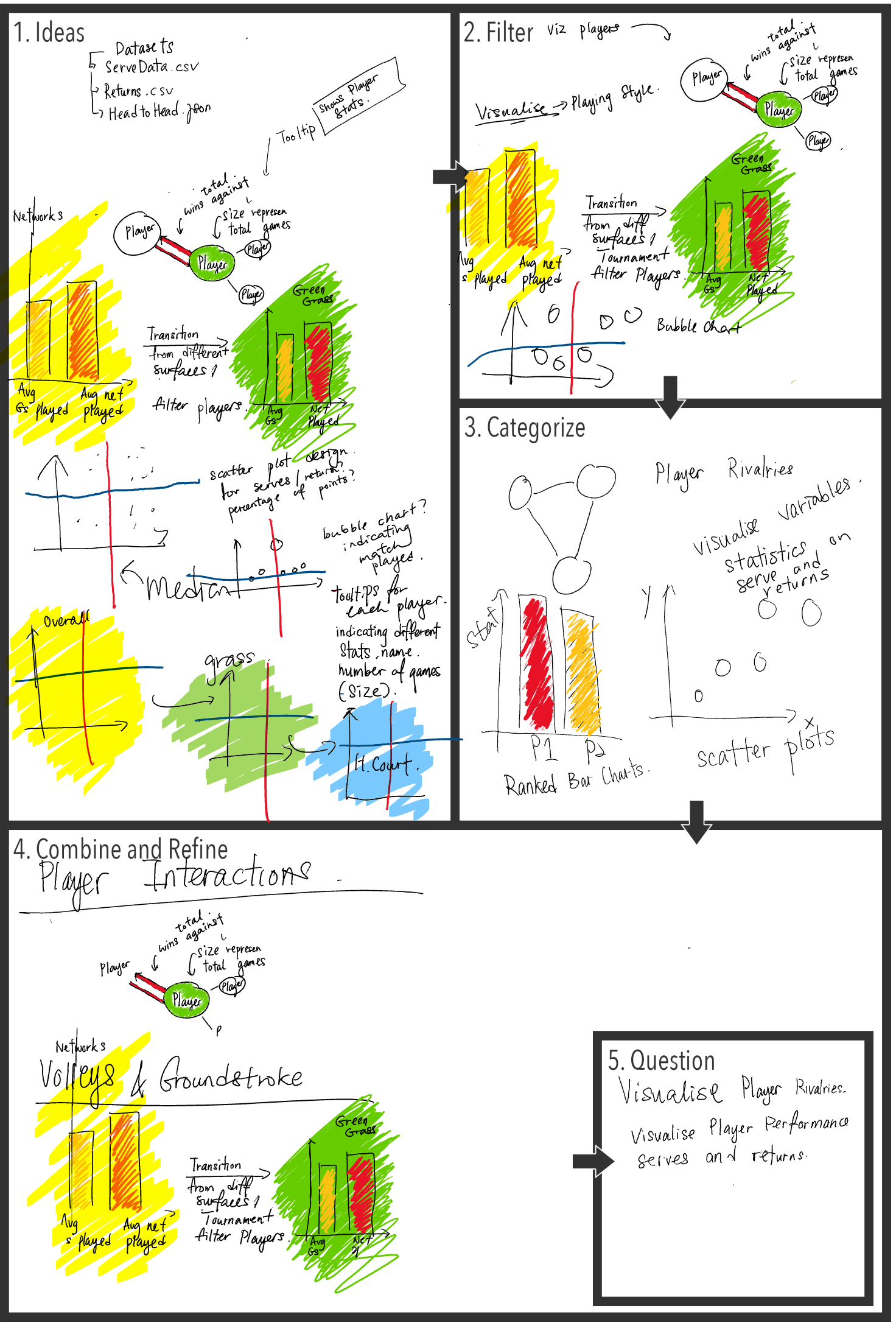
Jeff Sackmann <https://github.com/JeffSackmann/tennis_MatchChartingProject>

Jeff Sackmann <https://github.com/JeffSackmann/tennis_slam_pointbypoint>

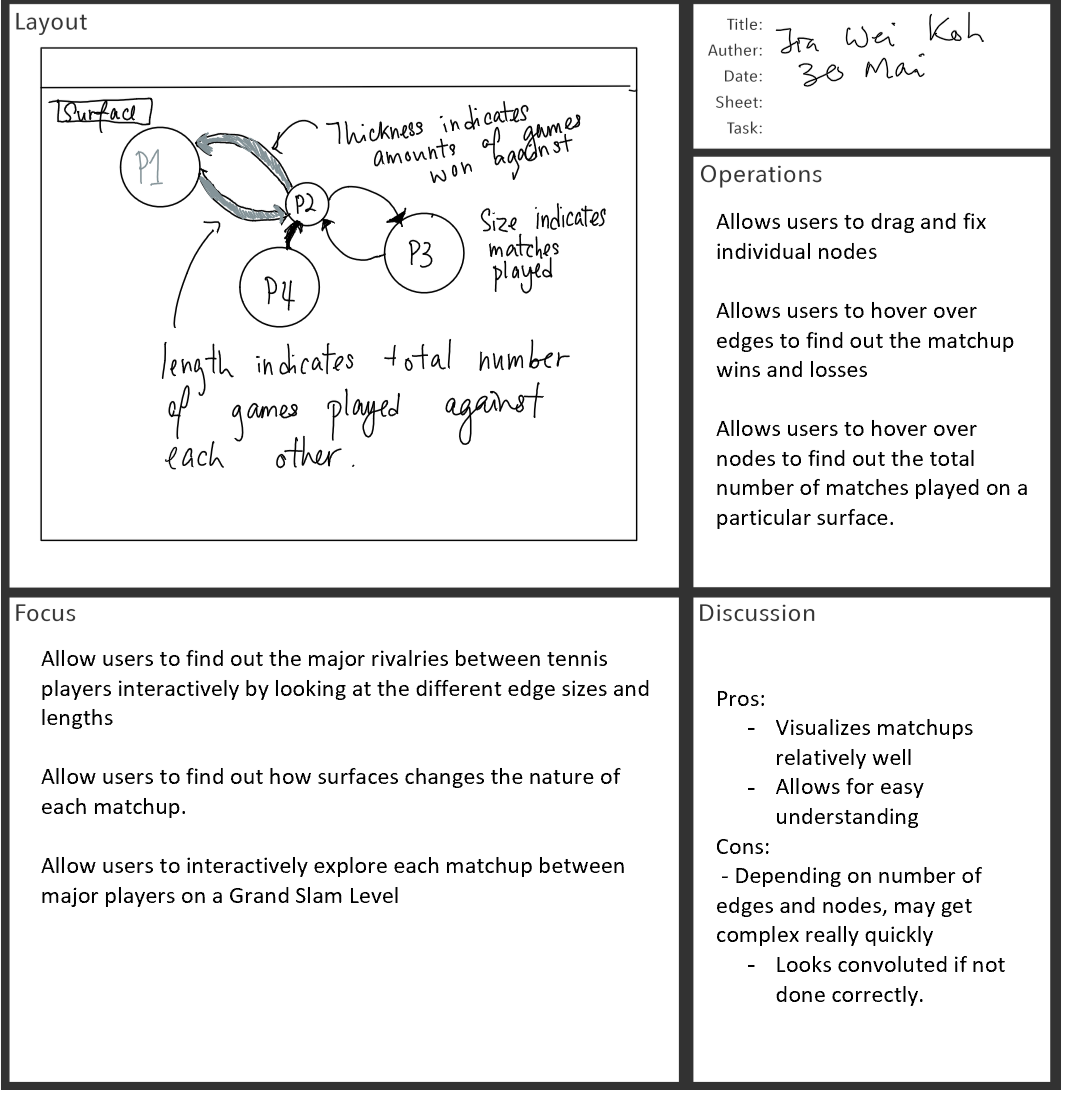
# 

# Appendix: Five Sheet Design

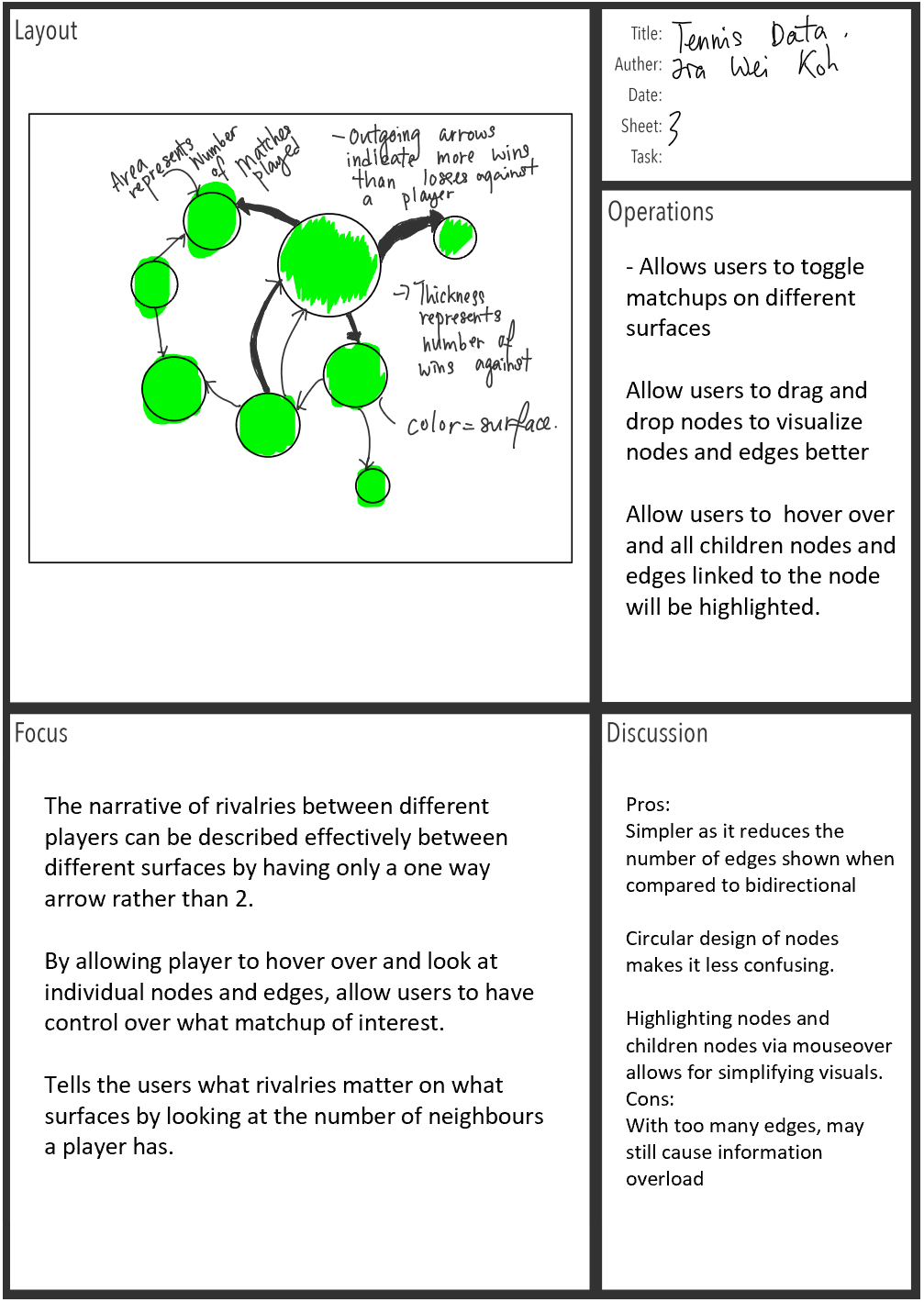
## Sheet 1:



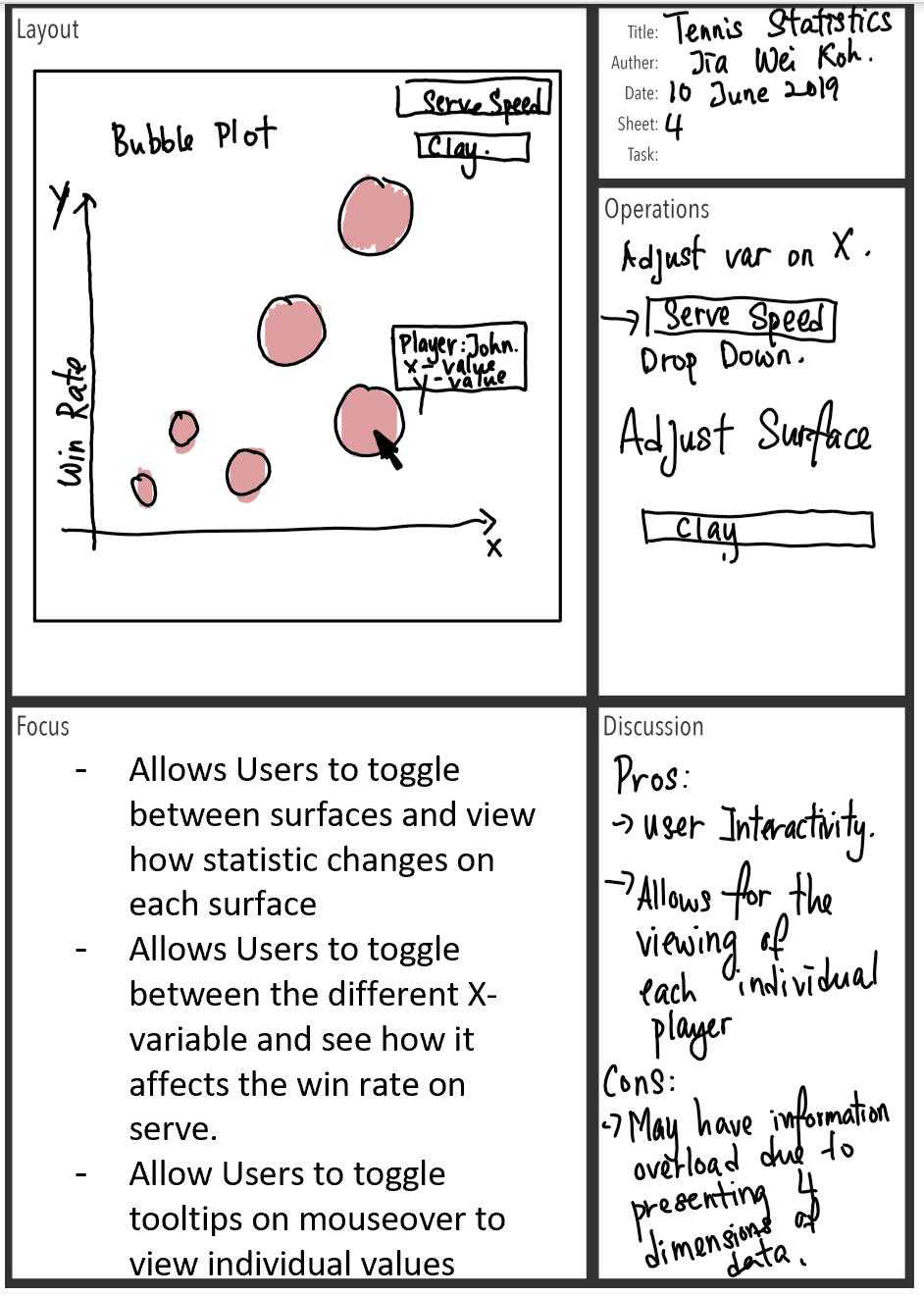
## Sheet 2:



## Sheet 3:



## Sheet 4:



## Sheet 5

