THE BIG FOUR

An Interactive Visualisation of the Dominance of
the Big Four

Jia Wei Koh
Jwkoh6@student.monash.edu

Contents

Introduction

In the world of sports, the term "Big Four" is used to describe the sheer dominance of 4 individuals, Roger Federer, Rafael Nadal, Novak Djokovic and Andy Murray in tennis. 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. (Table 1) They also have the top four prize money leaders of all time as well as a litany other records. 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	🇨🇭 Roger Federer	20	30
2	🇪🇸 Rafael Nadal	18	26
3	🇷🇸 Novak Djokovic	15	24
4	🇬🇧 Andy Murray	3	11
5	🇨🇭 Stan Wawrinka	3	4

With the recent expansion of tennis analytics 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, 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 reshaped 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 to improve the narrative of player rivalries 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. Hence, the more arrows outwards a player has, the better he is against all the other players that he played against.

The fourth design sheet focuses on how to visualise variables that 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 (Tufte, 2001). 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 colour 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 read through a short introduction to the Big Four's dominance in tennis, exploring player's rivalries and matchups against each other through the network chart, and finally analysing what makes a good player by analysing the serves and returns of each player. For each section, a short user guide is provided to allow users to explore the visualisations effectively, and a short summary for each section is used to introduce players to explore each visualisation.

Implementation:

The dataset had 2 million data points and was wrangled using Python and exported as the 3 datasets as stated above. The implementation is done using D3.js, a popular tool invented by Mike Bostock. The tool is selected primarily for its transitions and force simulations of network graphs, allowing for high amount of sophisticated user interactivity using Scalable Vector Images,

all the while allowing for everything to be ran on a simple web-browser. The smooth transitions of existing data allow for users to understand instantly the changes of value overall or for a particular datapoint.

For player rivalries, the data was imported using d3 and using a physics simulation built into d3(d3.force), allowed for nodes to be aligned approximately in a circle with heavier nodes weighted towards the centre. The visualisation allows for users to hover across nodes, which will highlight the edges and the children nodes that it is connected to. Nodes represents players, area represents number of matches played on the surface and edges represents their rivalries. Outgoing arrows means more wins against a player. Hovering over edges will tell you the Wins and Losses against the player. Furthermore, using d3.drag, users can drag and drop nodes to fix the position of nodes.

For Serves and Returns, utilising in built d3 functions and the combination of drop-down menus, transitions can be made using listening events for each data point. D3 allows for smooth transitions, which allow for users to instinctively interpret the number of matches played, the overall shift of values when changing from surfaces. Furthermore, tooltips allow for users to highlight players of interest while mouseover. These allows for users to explore the different variables that affect the data, while does not induce unnecessary clutter to the interface.

User Guide:

Accessing the File:

All ways of accessing requires access to the internet as to prevent cross-origin errors, data files are hosted online through a public repository on GitHub. There are 2 ways of accessing the visualisation presentation, first is through a [link](#) online, hosted on GitHub. The second way of accessing the data would be through the HTML file named index.html.

On opening the file, a short introduction of tennis and Big Four is provided. The user is then introduced to the rivalry section where users have the opportunity after a short introduction, interact with the network graph.

## Visualising Rivalries using the Network Chart
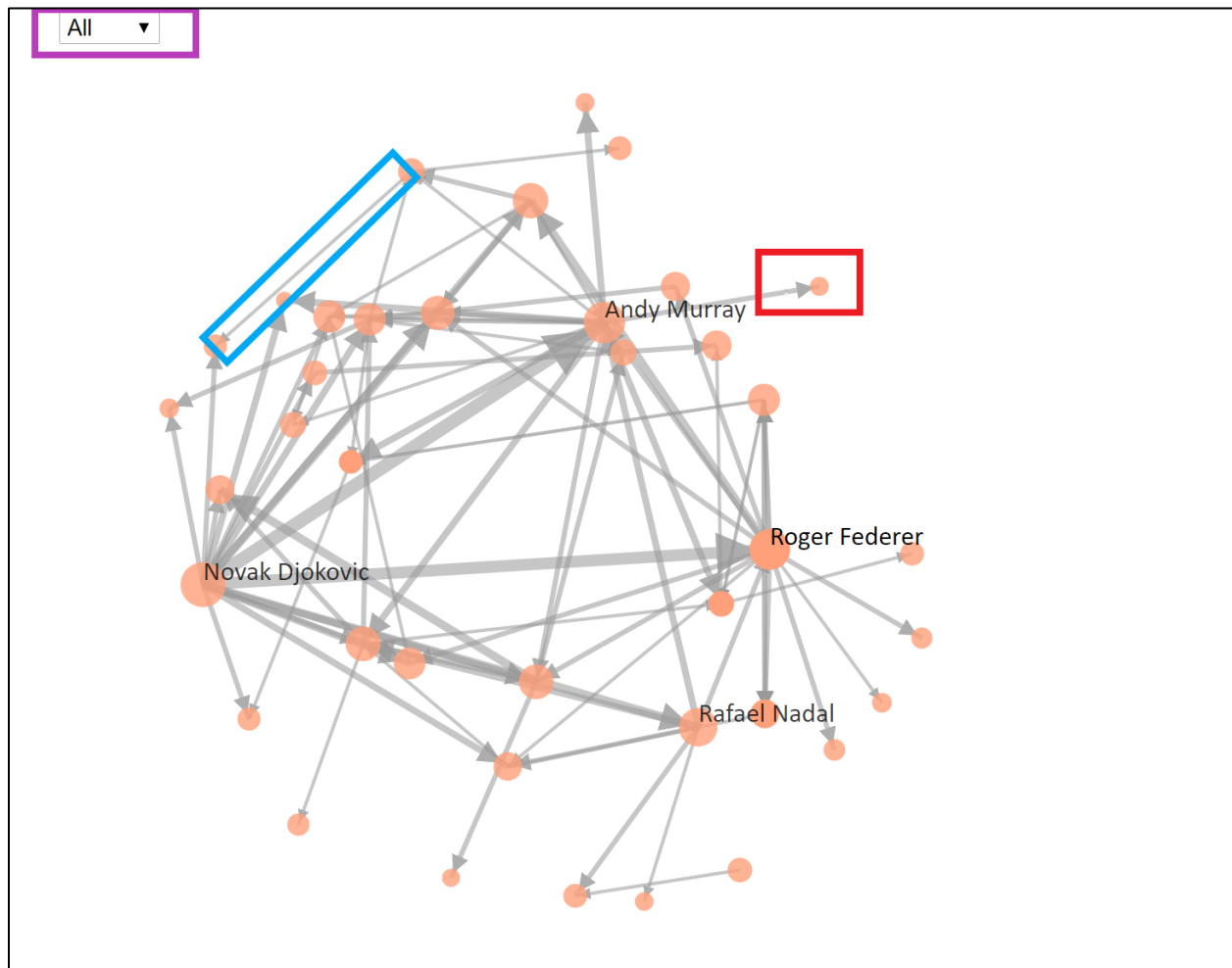
Purple Box: Drop-down box that allows users to select different surfaces. To restart the graph, select a different surface, and toggle back to the desired surface.

Blue Box: Mouseover to show the matchups represented by the edge.

Red Box: Mouseover to show the selected Player and edges as well as other players linked to this player. Drag and Drop to fix a node position

Physics Bug Event:
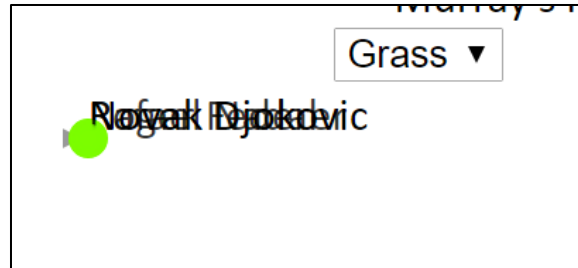
In an event that this happens, simply click on one of the dots and the animation will populate the other points in a circle, to restart the event, select a different surface, and toggle back to the desired surface.

Visualising the Serves and Returns:

Included on each of the bubble plots is a table which describes the selections available for the x-axis and a short description regarding the selected variable. To select a variable or surface simply select the drop-down menu. More instructions provided below.
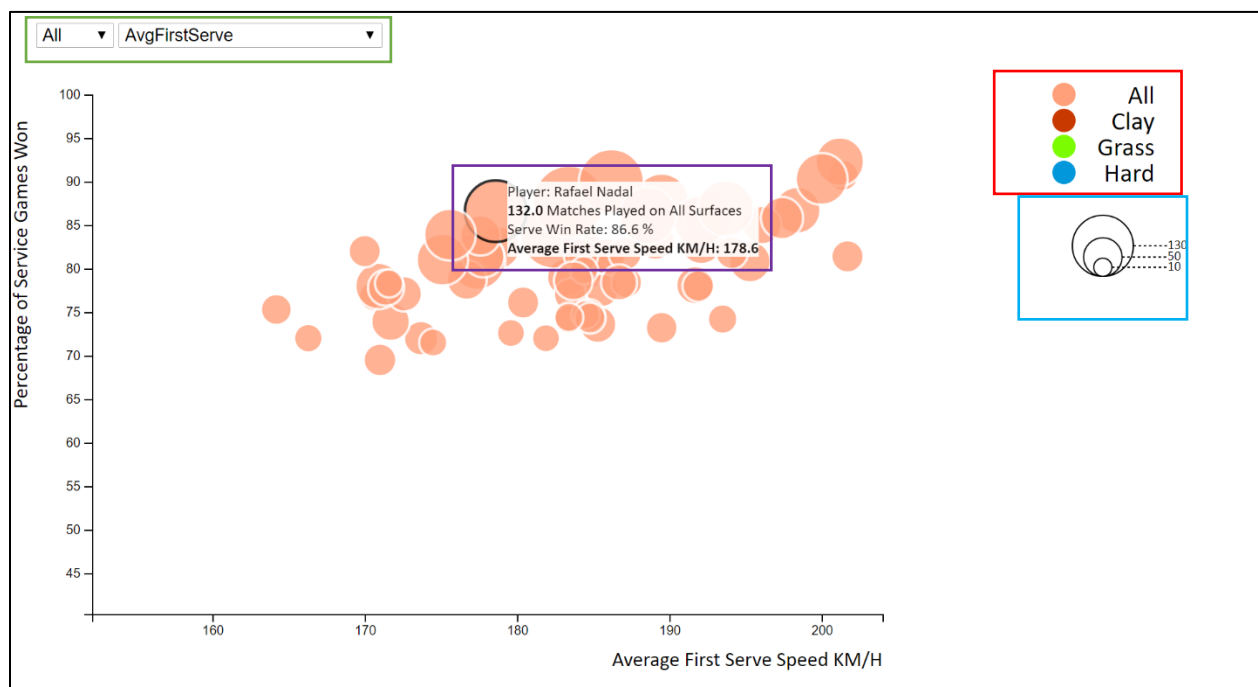
Purple Box: Mouseover shows a tooltip that shows the Player represented as well as the X and Y values.

Blue Box: Legend that shows the proportionate area to the number of matches played by each individual.

Red Box: Shows the different colours and their representation of different surfaces.

Green Box: Dropdown menus. The first allows for the user to select the surfaces, the second allows for the user to select the x-axis values that the user is interested in.

Conclusion:

The opportunity to provide a medium in which to communicate my favourite sport to audiences using statistics and simple but effective data visualisation has been nothing short of rewarding. The user interactions provided in these visualisations allows data to be communicated effectively to a user.

In hindsight, for this particular project, this dataset, alongside the inexperience in coding in JavaScript and d3 has resulted in a much more complex project 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. For this project, there were several limitations that could be improved upon. Firstly, the network graph could be improved by allowing tooltips to appear much faster upon hovering on the edges. The network graph could also be improved by fixing the physics bug as well as while hovering on the parent nodes, the children names also show up. The bubble chart could further be improved by allowing for users to plot variables on the y-axis, as well as allowing users to highlight and annotate player positions of whom they are interested in. Finally, the bubble chart has a problem in which the points overlap, these could be solved by using a jitter function or adding a force that repels the nodes slightly so that they do not overlap.

The dataset provides more opportunities to create and analyse more complex measures and more ways to communicate data to a wider audience . For a sports fan and a budding data scientist, this project allows for the learning opportunity to communicate tennis data to newcomers to the sport or for tennis enthusiast to look at tennis from a completely different perspective.
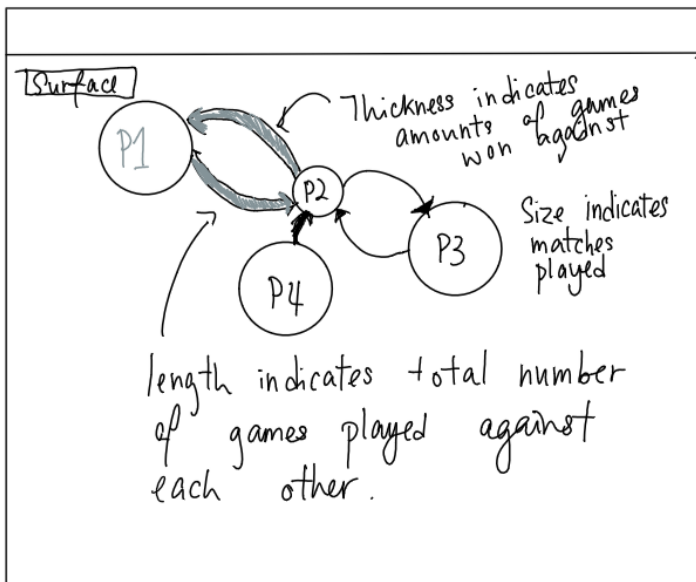
References and Dataset:

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

Tufte, E. R. (2001). *The visual display of quantitative information*: Second edition. Cheshire, Conn. : Graphics Press, [2001] ©2001.

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

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

Appendix: Five Sheet Design

Sheet 1:

Layout

Surface

P1 P2 P3 P4

Thickness indicates amounts of games won against

Size indicates matches played

length indicates total number of games played against each other.

Title:
Auther: Jia Wei Koh
Date: 30 Mai
Sheet:
Task:

Operations

Allows users to drag and fix individual nodes

Allows users to hover over edges to find out the matchup wins and losses

Allows users to hover over nodes to find out the total number of matches played on a particular surface.

Focus

Allow users to find out the major rivalries between tennis players interactively by looking at the different edge sizes and lengths

Allow users to find out how surfaces changes the nature of each matchup.

Allow users to interactively explore each matchup between major players on a Grand Slam Level
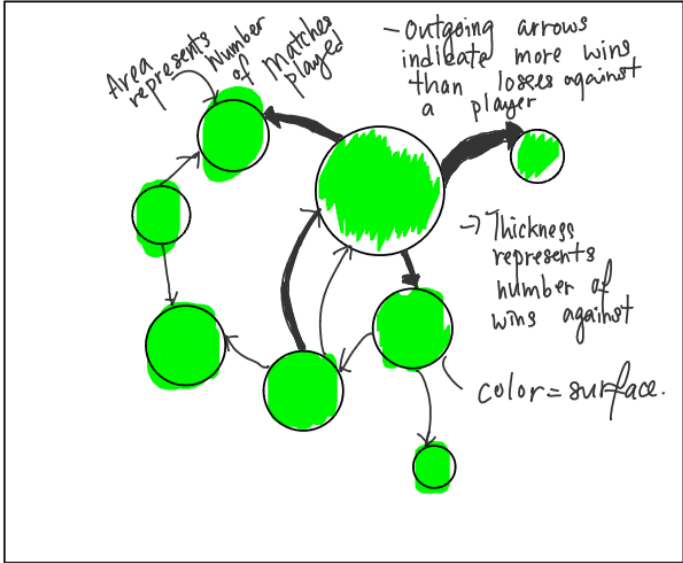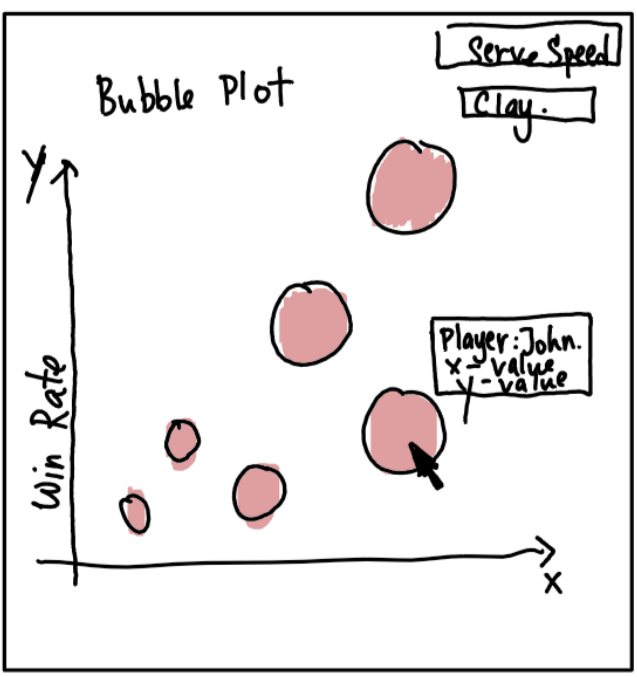
Discussion

Pros:
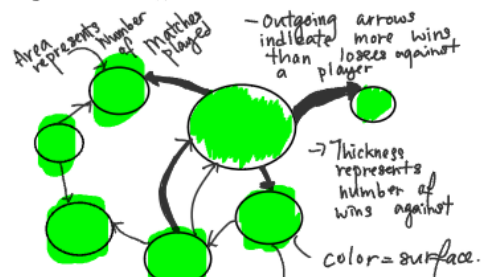- Visualizes matchups relatively well
- Allows for easy understanding

Cons:
- Depending on number of edges and nodes, may get complex really quickly
  - Looks convoluted if not done correctly.

Sheet 3:

Layout

Area represents number of matches played

— Outgoing arrows indicate more wins than losses against a player

→ Thickness represents number of wins against

color = surface.

Title: Tennis Data.
Auther: Jia Wei Koh
Date:
Sheet: 3
Task:

Operations

- Allows users to toggle matchups on different surfaces

Allow users to drag and drop nodes to visualize nodes and edges better

Allow users to hover over and all children nodes and edges linked to the node will be highlighted.

Focus

The narrative of rivalries between different players can be described effectively between different surfaces by having only a one way arrow rather than 2.

By allowing player to hover over and look at individual nodes and edges, allow users to have control over what matchup of interest.

Tells the users what rivalries matter on what surfaces by looking at the number of neighbours a player has.

Discussion

Pros:
Simpler as it reduces the number of edges shown when compared to bidirectional

Circular design of nodes makes it less confusing.

Highlighting nodes and children nodes via mouseover allows for simplifying visuals.
Cons:
With too many edges, may still cause information overload

Layout	Title: Tennis Statistics
	Auther: Jia Wei Koh.
	Date: 10 June 2019
	Sheet: 4
	Task:

Layout

Bubble Plot

Serve Speed
Clay.

Y axis — Win Rate
X axis

Player: John.
X → value
Y → value

Operations

Adjust var on X.

→ Serve Speed
Drop Down.

Adjust Surface

Clay

Focus

- Allows Users to toggle between surfaces and view how statistic changes on each surface
- Allows Users to toggle between the different X-variable and see how it affects the win rate on serve.
- Allow Users to toggle tooltips on mouseover to view individual values

Discussion

Pros:
→ user Interactivity.
→ Allows for the viewing of each individual player

Cons:
→ May have information overload due to presenting 4 dimensions of data.

13

Layout

Tennis Rivalries.

Area represents Number of matches played

— Outgoing arrows indicate more wins than losses against a player

→ Thickness represents number of wins against

color = surface.

Serves / Returns.
Bubble Plot

Serve Speed
Clay.

Y

Win Rate

Player: John.
x - value
y - value

x

Title: Tennis Data
Auther: Jra Wei Koh.
Date:
Sheet: 5
Task:

Operations

For Rivalries:
Drag and drop nodes that are arranged in a circle
Mouseover to show edges and parent and children nodes
Tooltips show matches played on nodes
Edges show matchups

Serve/Returns:
Tooltips show the x-values and the y values and player highlighted.
Dropdown menu toggles for surfaces and x variables.

Focus

Rivalries
- Allow users to focus on important rivalries when contending for Grand Slam titles
- Allow users to focus on Big Four and how their rivalries unfold.

Serves/Returns
- Allow users to focus on different players and which aspect of serves are their strengths and weaknesses
- Allow users to focus on performance in different surfaces for each individual player or overall.
- Allow users to focus on different metrics of each player.

Detail

Software:
    D3
    JavaScript
    Internet Browser.

Minimal Requirements

Per Edward Tufte,
Less is more!

14