Aaron Brady (and Jack Lafond)

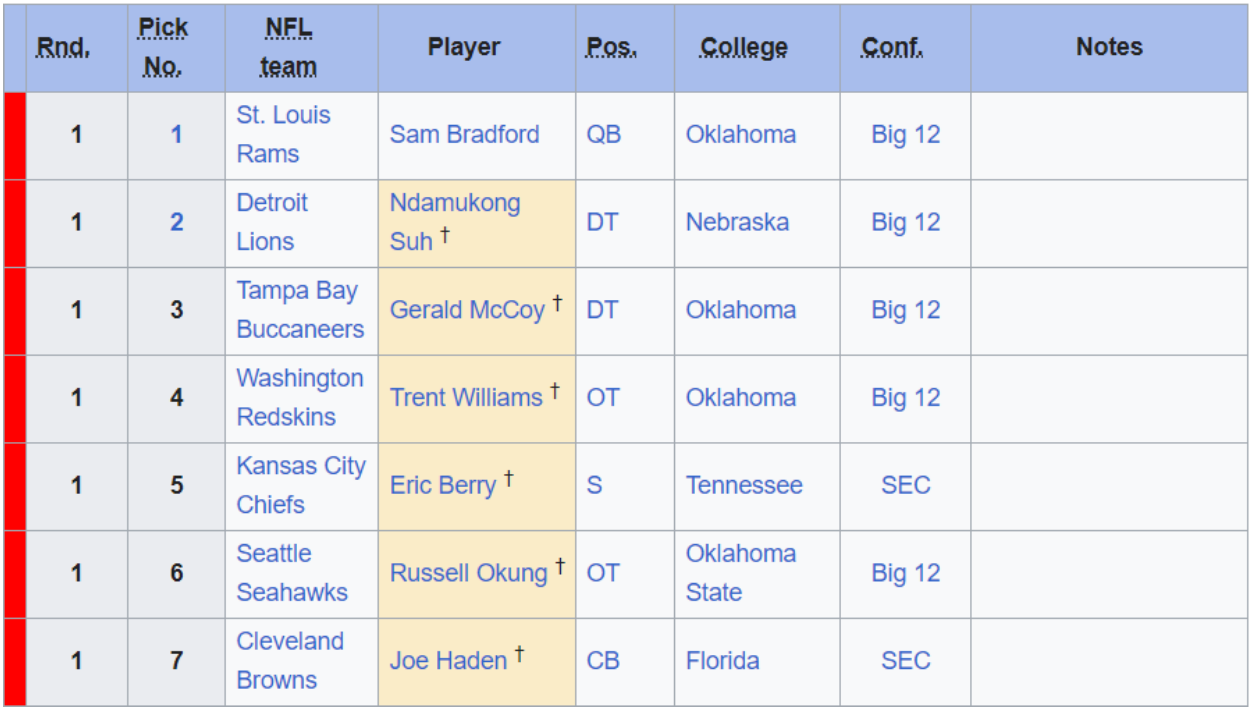
MA 4804 Process Book

### Overview, Motivation, and Questions

This project scrapes NFL draft and player data for the years 2010 to 2020 from Wikipedia and visualizes it in a Sankey diagram and stacked bar chart. The goal of this visual is to see the connections between college conferences, the NFL draft, and NFL divisions. The motivation of this project was to see if there were major pipelines from college to the NFL, and if certain colleges produce more early round draft picks. The questions we originally asked were how specific college teams connect to specific NFL teams? However, as we worked with the visual, we quickly realized it would become too crowded if we included all the specific teams. We then transitioned to more broad questions such as, which college conference has the most first round draft picks? Are there any pipelines from college conferences to NFL divisions? For example, are there a lot of picks made by NFC west teams from the PAC-12. While creating the visual a question that we had was how do the teams compare to each other within the division, are there teams that are making a lot more picks than others? For example, recently the 49ers have traded a lot of their picks for players, so we wanted to see if this has happened in the past.

### Data Source

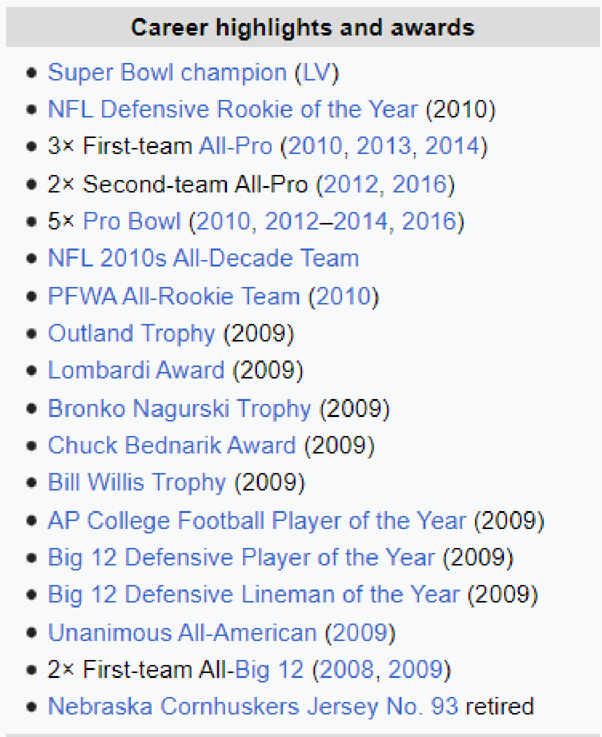
We created a data scraper using the bs4 python library to scrape the table, except for the notes column, pictured below for the years 2010 to 2020.



| | Rnd. | Pick No. | NFL team | Player | Pos. | College | Conf. | Notes |
|---|---|---|---|---|---|---|---|---|
| 🟥 | 1 | 1 | St. Louis Rams | Sam Bradford | QB | Oklahoma | Big 12 | |
| 🟥 | 1 | 2 | Detroit Lions | Ndamukong Suh † | DT | Nebraska | Big 12 | |
| 🟥 | 1 | 3 | Tampa Bay Buccaneers | Gerald McCoy † | DT | Oklahoma | Big 12 | |
| 🟥 | 1 | 4 | Washington Redskins | Trent Williams † | OT | Oklahoma | Big 12 | |
| 🟥 | 1 | 5 | Kansas City Chiefs | Eric Berry † | S | Tennessee | SEC | |
| 🟥 | 1 | 6 | Seattle Seahawks | Russell Okung † | OT | Oklahoma State | Big 12 | |
| 🟥 | 1 | 7 | Cleveland Browns | Joe Haden † | CB | Florida | SEC | |

https://en.wikipedia.org/wiki/2010_NFL_draft

The scraper also went into each player and scraped the number of MVP, SB MVP, SB WIN, OPOY, DPOY, OROY, DROY, First AP, Second AP, Pro Bowl the player has won. However, we did not end up visualizing this data, but could be done in an extension of this project.

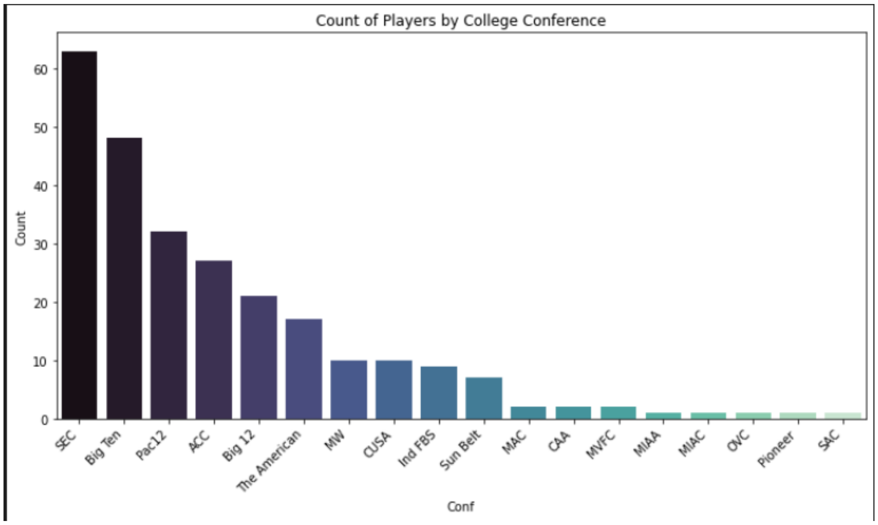


**Career highlights and awards**

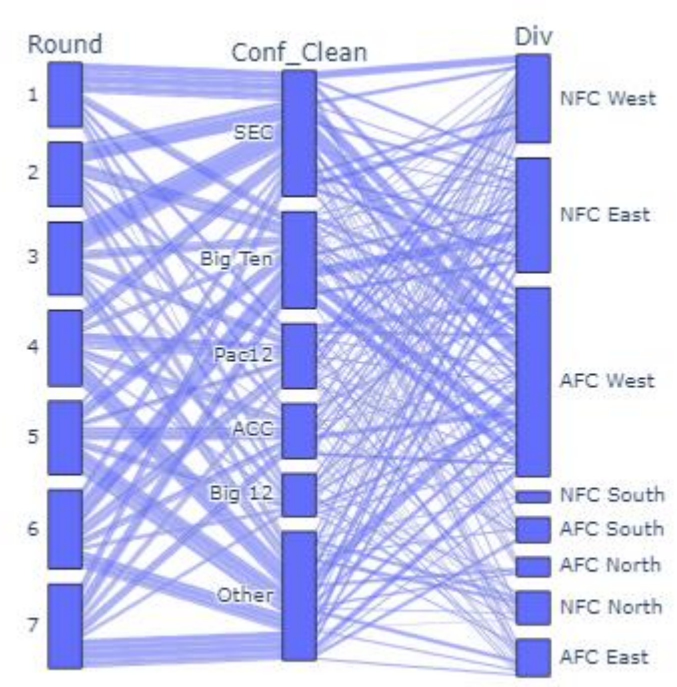
- Super Bowl champion (LV)
- NFL Defensive Rookie of the Year (2010)
- 3× First-team All-Pro (2010, 2013, 2014)
- 2× Second-team All-Pro (2012, 2016)
- 5× Pro Bowl (2010, 2012–2014, 2016)
- NFL 2010s All-Decade Team
- PFWA All-Rookie Team (2010)
- Outland Trophy (2009)
- Lombardi Award (2009)
- Bronko Nagurski Trophy (2009)
- Chuck Bednarik Award (2009)
- Bill Willis Trophy (2009)
- AP College Football Player of the Year (2009)
- Big 12 Defensive Player of the Year (2009)
- Big 12 Defensive Lineman of the Year (2009)
- Unanimous All-American (2009)
- 2× First-team All-Big 12 (2008, 2009)
- Nebraska Cornhuskers Jersey No. 93 retired

### Exploratory Data Analysis

For our exploratory analysis we examined the draft year of 2020, as this was the first data set we scraped. We graphed the count of players by different aspects of the data. We graphed the count of players drafted by NFL team, college team, college conference, and position. The graph provided below led us to create the cutoff for the college conferences to include in our final visual.

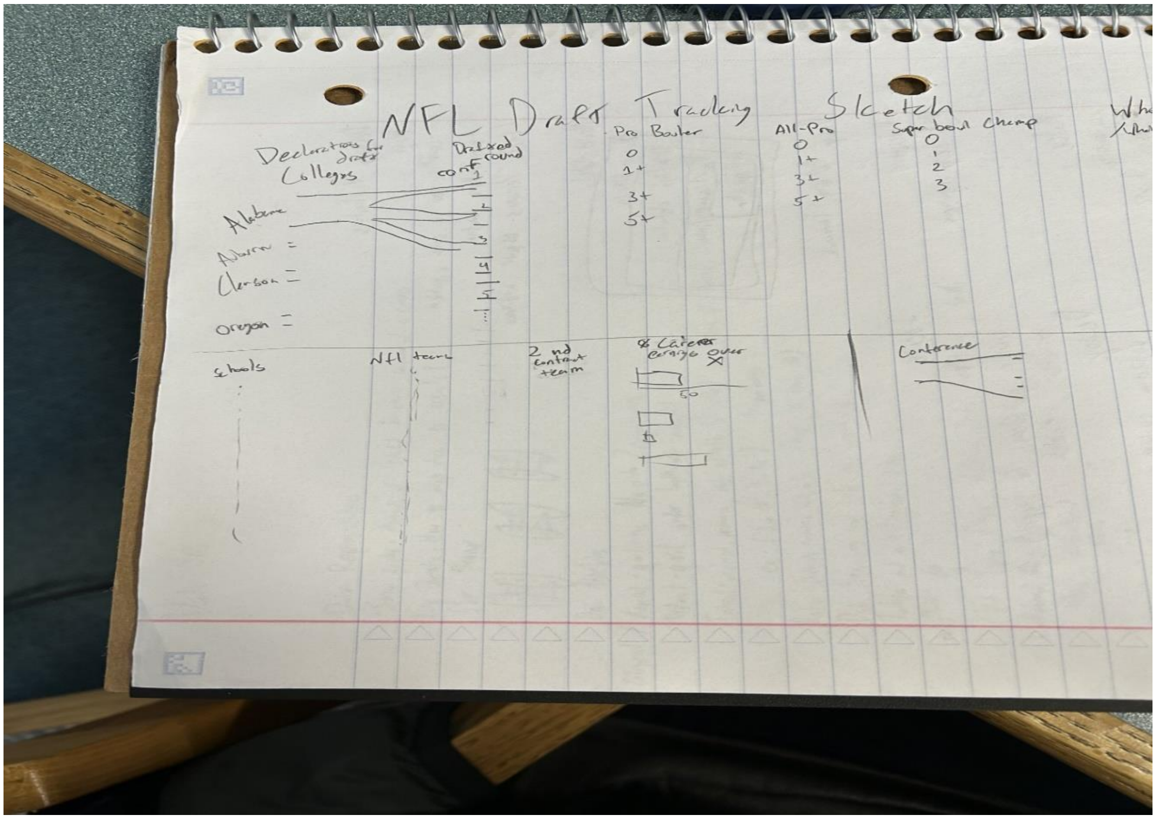


Also, during our EDA, we found many problems with our data quality. If a team had moved locations, they were still the same team but had a different name across the years (St. Louis Rams became the Los Angeles Rams). When mapping these changes to make them consistent, our code was incorrect and was changing all the team names for the players. This was throwing off counts and destroying the integrity of our data. We noticed this in our EDA and early on visuals. In this parallel category image we can see a heavy imbalance across the NFL divisions, which makes no sense.

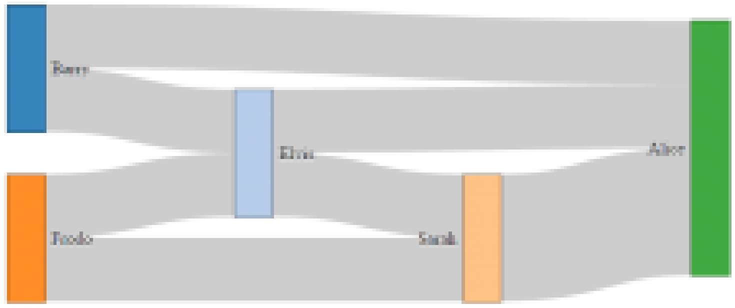


## Design Evolution

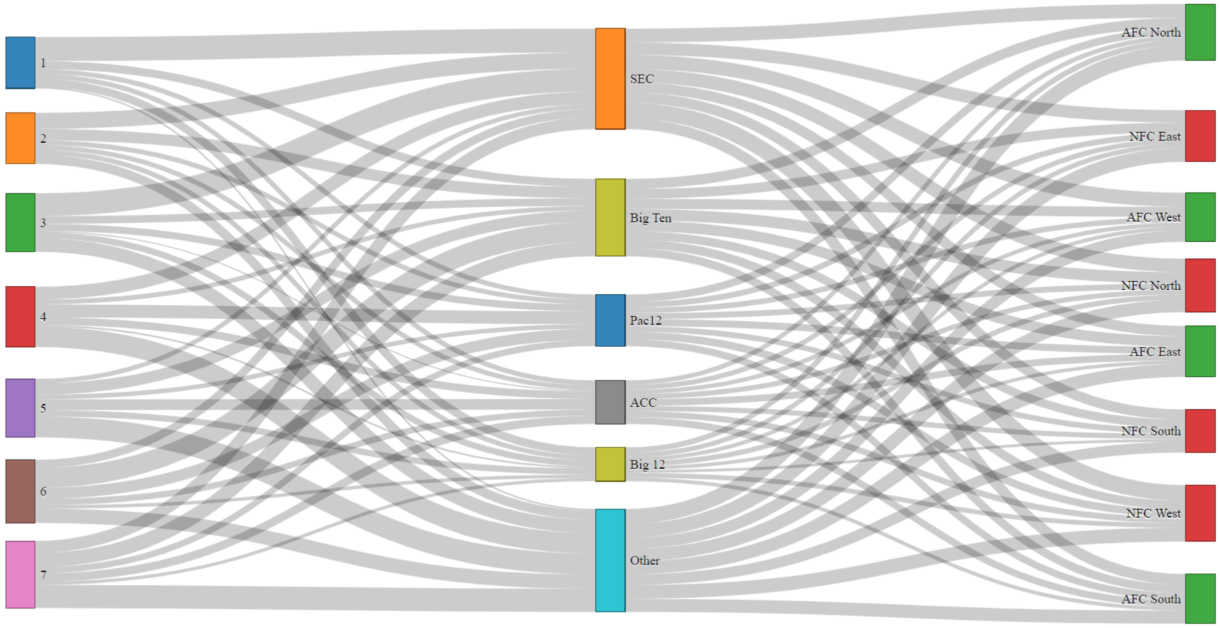
We originally proposed a Sankey diagram and that was the path we went down. Pictured below are early sketches on what we envisioned the Sankey looking like.



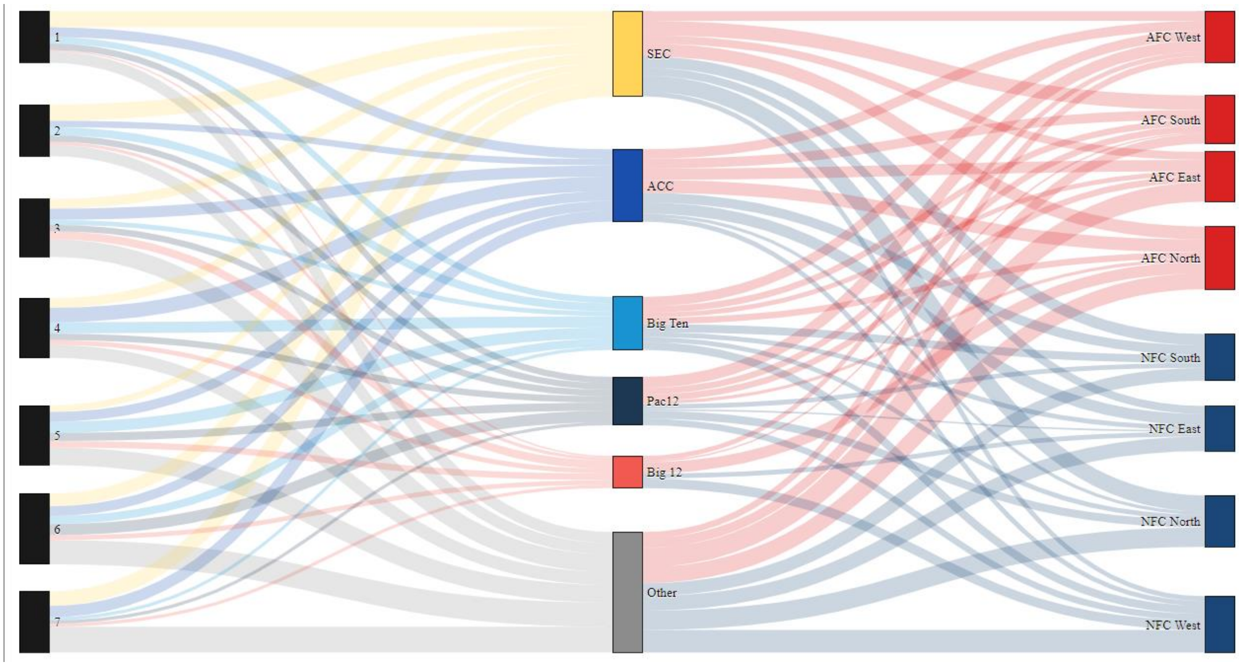
In the early stages we imagined the Sankey being able to link college to NFL to success, but we decided to just focus on college to NFL to not overwhelm the audience with all the lines. With all the connections at a super specific level of teams, the visual would not have been able to bring value to the audience. We know with this decision we were losing a lot of information, but through interactive filters we believe this information could be brought back in. We also thought many of the awards were too sparse to be able to look at in the big picture. We think that player stats can be visualized when filtered to specific divisions and would be something to implement as an extension of this project. So now that we had our EDA and rough sketches done, we decided to link three categories: 1. Round of Draft Selection 2. College Conference and 3. NFL Division. With this established we found starter code that created the following visual, which can be found here (https://gist.github.com/d3noob/06e72deea99e7b4859841f305f63ba85).



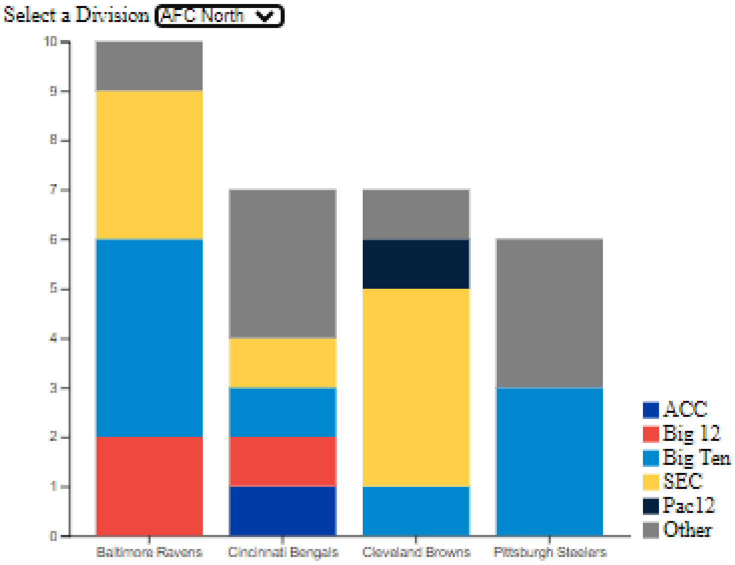
This visual is created using a source, target, value table, so we had to create these tables for each of our cleaned csvs. We used group by statements to group by the target and source and get a count. Changing the source csv to one of our csvs that was created by our cleaning code we created this:



We were pleased to see that the Sankey was successfully created, but a lot of work had to be done in terms of coloring and implementing the filter we wanted so the user can look at the different years. We changed the colors of the college conferences to the main color of the conferences logo and made that the color of the stroke that starts at a round and goes to the conference node. For the NFL divisions we made the AFC divisions the color of the AFC logo and for the NFC divisions made them the color of that logo. Our Sankey is pictured below before we implemented any interaction.



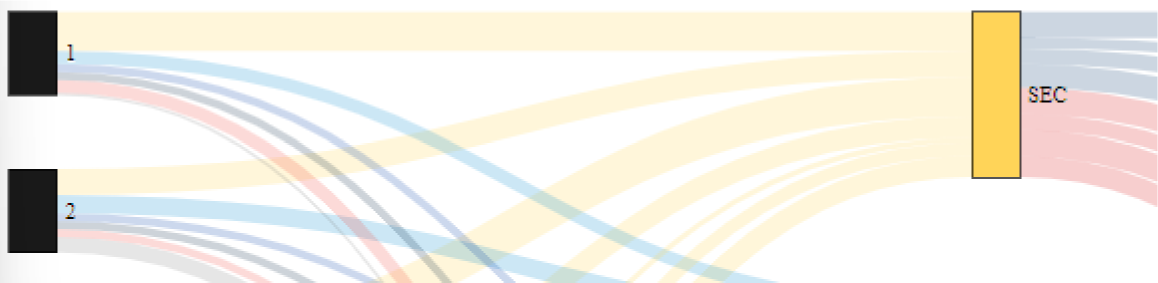
With this we thought we could take it a step further by comparing the college conferences NFL teams selected from. This would visualize the information lost when aggregating the NFL teams to divisions. We created a stacked bar chart. We chose a stacked bar chart because teams typically make 7 selections a draft, so the scale is small even when aggregated over 10 years, so it is still simple to make comparisons of size. This visual easily tells the audience what teams draft more from which conferences.


Select a Division AFC North

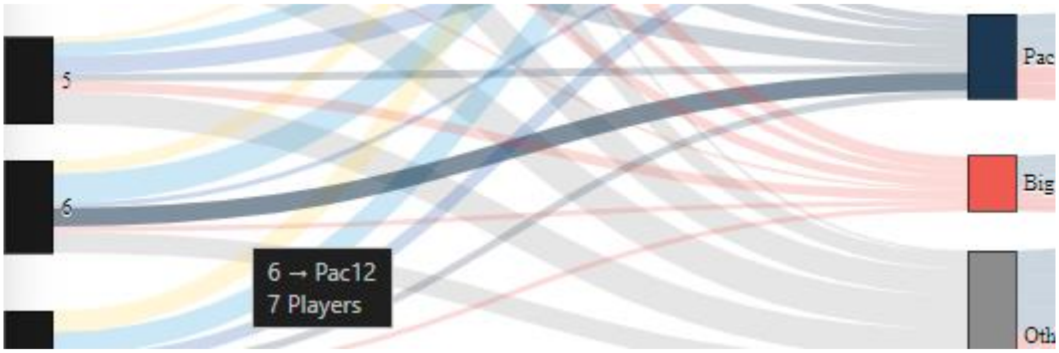
For this we originally thought of doing a chart to the left of each node of the divisions, but that would have been too crowded and made the visual so small that it would be impossible to read, so we pivoted to placing one graph under the Sankey and the user can pick the division. An alternative to the user filter would be to implement small multiples underneath the Sankey.

### Implementation

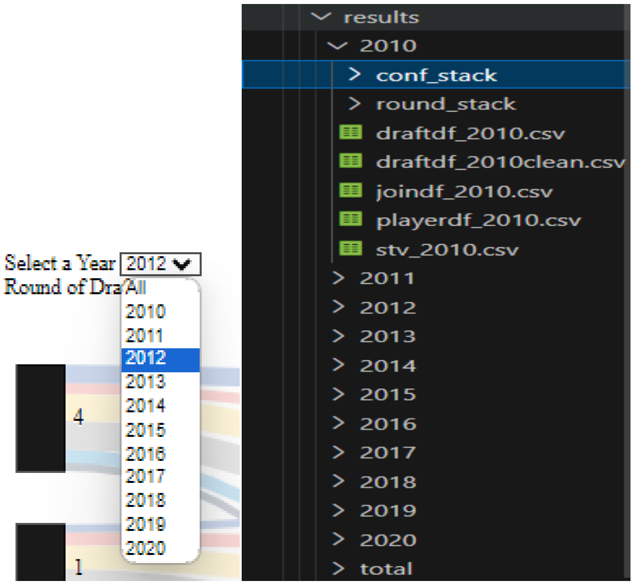
The base code had intractability by having the ability to drag the nodes vertically. This allows the user to set the order of the nodes vertically themselves. This is helpful if they are interested in specific relationships. For example, first round picks from the SEC, as they can line the two nodes up.



Now we can easily look at the round one node and see that nearly half the players picked in round one was from the SEC. The same idea can be applied to the NFL divisions, as pictured above I have all the AFC divisions below the NFC so we can compare how many SEC players went to the AFC versus NFC. The starting code also had tool tips implemented and hover highlighting. This gave the user exact measures of the target and source. The tool tip also appeared with the total count of players in that node when hovering over the node.

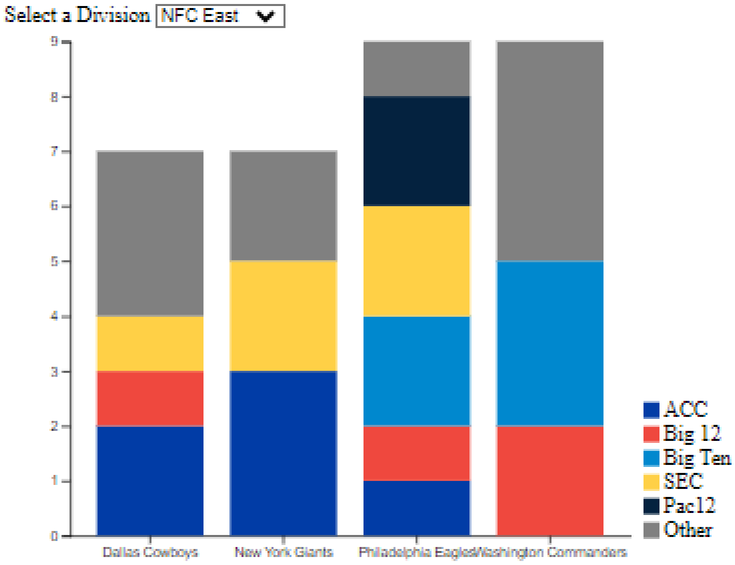


For intractability that we added was a year filter. We created a results folder that stored all the csvs that fed the visuals and broke it up by year. When the user selected a year, the visual would read from the corresponding folder and show the correct data. The year filter also applied to the stacked bar chart.



```
results
  2010
    conf_stack
    round_stack
    draftdf_2010.csv
    draftdf_2010clean.csv
    joindf_2010.csv
    playerdf_2010.csv
    stv_2010.csv
  2011
  2012
  2013
  2014
  2015
  2016
  2017
  2018
  2019
  2020
  total
```

STV stands for source, target, value which feeds the Sankey, and there are 8 csvs in the conf_stack which feeds the stacked bar chart. There are eight csvs because there is one for each team. The user can filter to what division they want to look at and that will be applied to the chart. Below is a picture of the chart filtered to 2012 and NFC East.



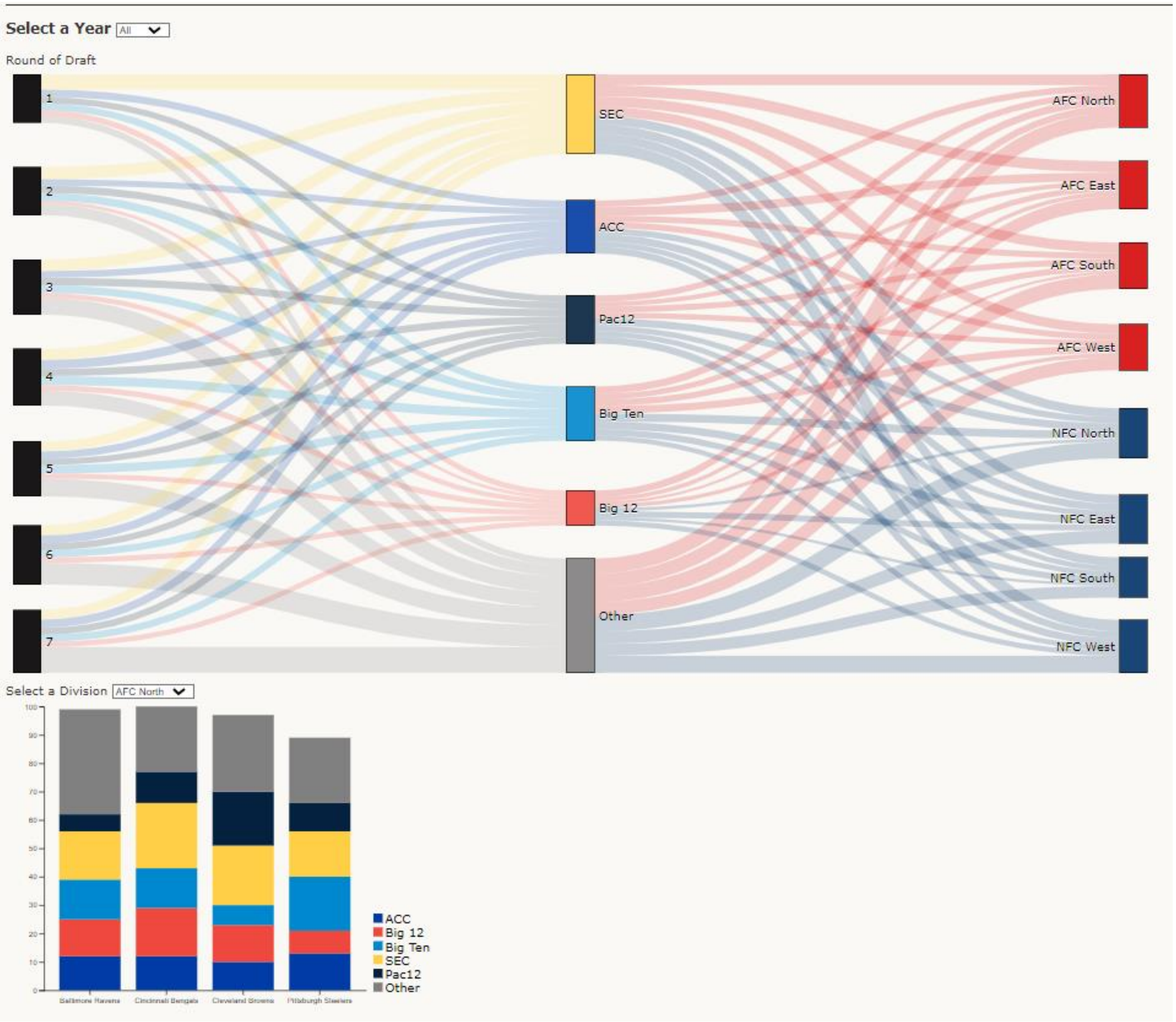
The division filter does not apply to the Sankey, but if we were to implement it to, we would switch the last column from NFL divisions to the teams in that division.

**Evaluation**

Through our visualization we learned that the SEC has the most first round selections over the last 10 years over any other college conference. And in recent years they have had a lot of early rounds (1-3) draft picks. Expanding on this, in general we see that the major conferences dominating the draft over the other node (there have been fewer draft picks from other divisions recently). We had expected to see that there could have been relationships between conferences and divisions but our visual does not show that any significant relationship exists. A question that is answered by the stacked bar chart is the difference in number of total draft picks between teams in a division. Additionally, from the stacked bar chart we learned that teams in the same division have similar college conference player breakdown, but there are teams that are much different from the rest of the division. For example, when looking at all for the year filter we see the Buffalo Bills selected a lot more ACC players than the rest of the division, and much less Big Ten players. I believe the visual works well at what it is designed to do. It shows the relationship between the NFL draft, college conferences, and NFL divisions and teams.

Improvements to this visual could be made like setting a pre-defined order for the nodes while keeping the drag feature. Additionally, we scraped player achievements but did not create a visual to show a player’s success on a team. I think this would be valuable to include to connect a player’s success to the NFL draft. Also, I think more interactive features such as applying the NFL division to the entire visual and make the right most nodes the teams in that division would be beneficial. A similar idea could be applied to the college conference. There are a lot of college teams in a conference so we would probably just show the top teams in the conference than another group.

**Final Visual**



Select a Year All

Select a Division AFC North