Project WhoScored

Taylor Tillinger, Kongmeng Vang, John Vue, Alexander Hohl, Kevin VanEmmerick, Ajay Sadhu



Fig. 1: Emblems of all English Premier League Clubs

**Abstract**—Project WhoScored is a visualization website that create a fun experience for English Premier League fans and any other people who are interested in soccer statistics. WhoScored visualize all soccer players statistics of the 2014/2015 EPL season, and place the dataset into interactive graphs and map that lets users search for their favorite players or teams. This report will inform the readers on how project WhoScored was built and organized. We use related works as outside sources to help us build the project, explaining the design of the system where all dataset came from,how it was stored and then implement it to WhoScored. Learn how the graphs are created and how they worked, along with the evaluation process on creating WhoScored.

**Index Terms**—Visual Analytics, Crossfilter, Soccer, Interactive, Query

Introduction

The English Premier League is one of the best soccer leagues on this planet. It draws great viewership, filling stadiums locally and pubs, bars and living rooms globally with excited fans and spectators. Clubs like Manchester United, Chelsea, Liverpool, and Manchester City are known all over the world and draw especially excited crowds in Europe, Asia and Africa. Therefore, the amount of money drawn from sponsorship by Premier League clubs and its players is unprecedented, resulting in very lucrative contracts for involved companies and individuals.

Wherever people are excited in competitive sports, statistical data will usually be recorded by professional entities on multiple levels: leagues, clubs, players, games, seasons. Therefore, we face an enormous amount of data that has multiple dimensions, including space and time. In order to make sense out of the data, to analyse it for the identification of spatiotemporal patterns, we need tools that are easy to use, enable querying, and are viewer-friendly. While an array of such tools, visualization approaches and programming libraries already exist, they are mainly domain specific and fail to integrate a wider spectrum of data from various arenas.

* All authors are with the University of North Carolina at Charlotte
* Taylor Tillinger: ttilling@uncc.edu
* Kongmeng Vang: kvang20@uncc.edu
* John Vue: jvue2@uncc.edu
* Alexander Hohl: ahohl@uncc.edu
* Kevin VanEmmerick: kvanemme@uncc.edu
* Ajay Sadhu: asadhu@uncc.edu
* Manuscript received 8 Dec. 2014; accepted 8 Dec. 2014; date of publication
* 8 Dec 2014; date of current version 8 Dec 2014.

For information on obtaining reprints of this article, please send

e-mail to: *jing.yang@uncc.edu*.

Here, we developed a general framework for interactively analysing multidimensional data in order to identify spatiotemporal patterns that can be the starting point of further confirmatory investigation. We created a web-browser enabled application to query and visualize statistical data of the 2014/15 Premier League season on an individual player level, for instance goals scored, minutes played, or assists, using data provided by WhoSocred.com, a website that displays worldwide soccer results and that includes all big leagues in Europe and the Americas. WhoScored.com is created and maintained by a team of soccer enthusiasts and software developers based in London, UK. It provides live scores, statistics, ratings, team characterizations, previews and expert articles. Our goal was developing the capability to visually query the data using interactive histograms and bar charts and display the resulting subset using the Visual Analytics approach of multiple coordinated views, for instance a list of players and a geographical map of the country of their provenance.

# Related Work

As soccer is a very popular sport worldwide, there are multiple resources online that contain information and data. The Fédération Internationale de Football Association (FIFA) is the international governing body in soccer, and their website [1] contains articles and data about professional soccer. It also presents a history of soccer and results from previous world cups, which can also be accessed at other sources, e.g. [2]. More soccer statistics and records that we initially thought of using for data in our visualization can be found at soccerstats.com [3]. Such statistics were not collected until recently, and [4] goes into detail about how people would like to see the data and compares statistics from baseball to the lack of statistics in soccer.

There are multiple examples of existing visualization tools in sports: The probably most prominent example is SoccerStories [5], an application to analyze soccer data and communicate results to an interested audience. It provides an overview and a detail interface of game phases. These interfaces consist of multiple linked visualizations that are tailored towards analyzing various player and team actions, such as passing and goal attempt.

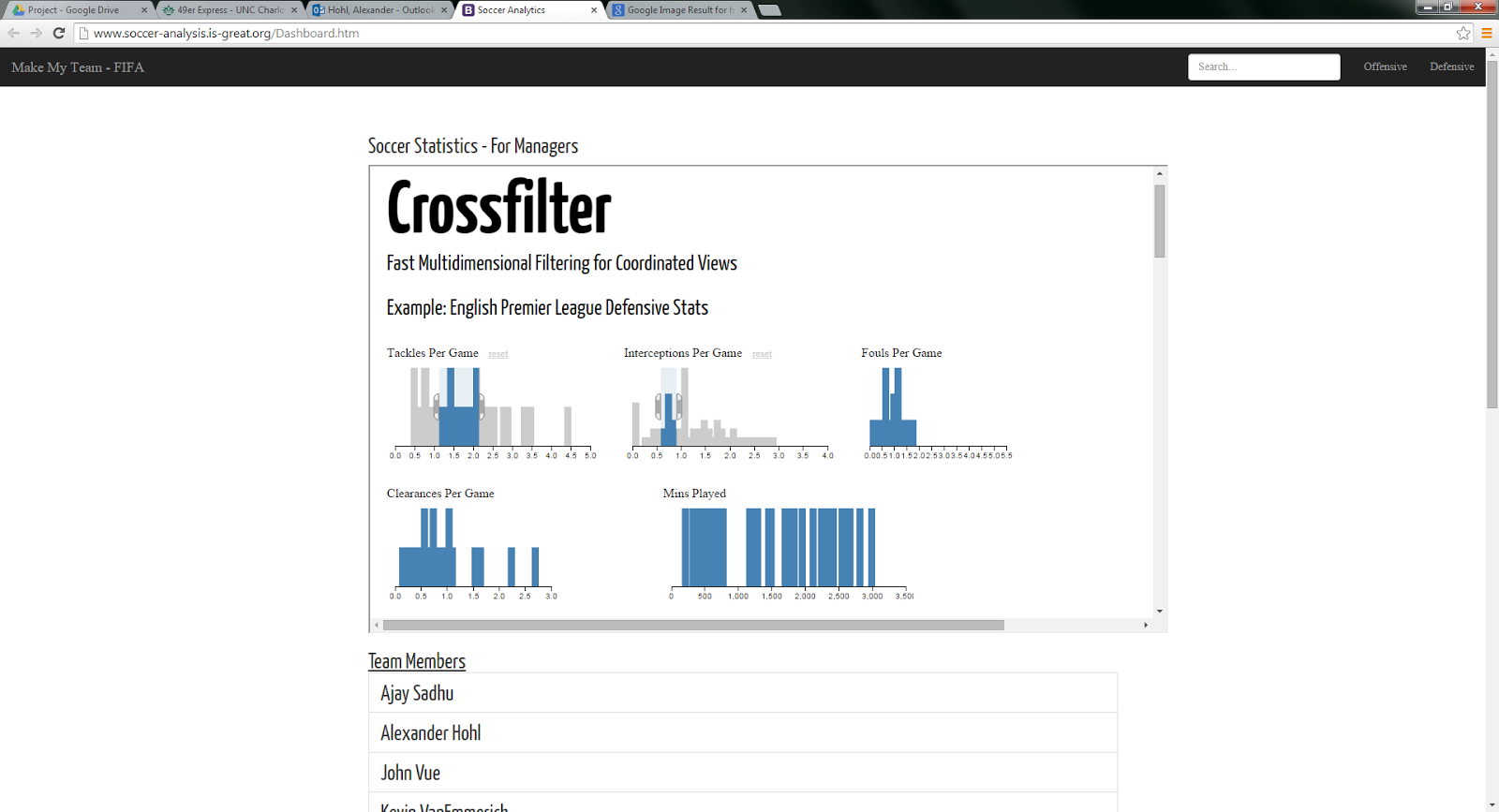
[6] introduce A Table!, a soccer ranking table that provides temporal navigation by combining two interaction techniques: DRAG-CELL, VIZ RANK. Drag-cell or viz rank are relevant to our

project because they provide similar data analyzing objective that we want our users to have. Drag-cell allows users to interact with the data itself in the database and Viz rank displays the differences and similarity for comparing different players’ attributes in a graphical display.

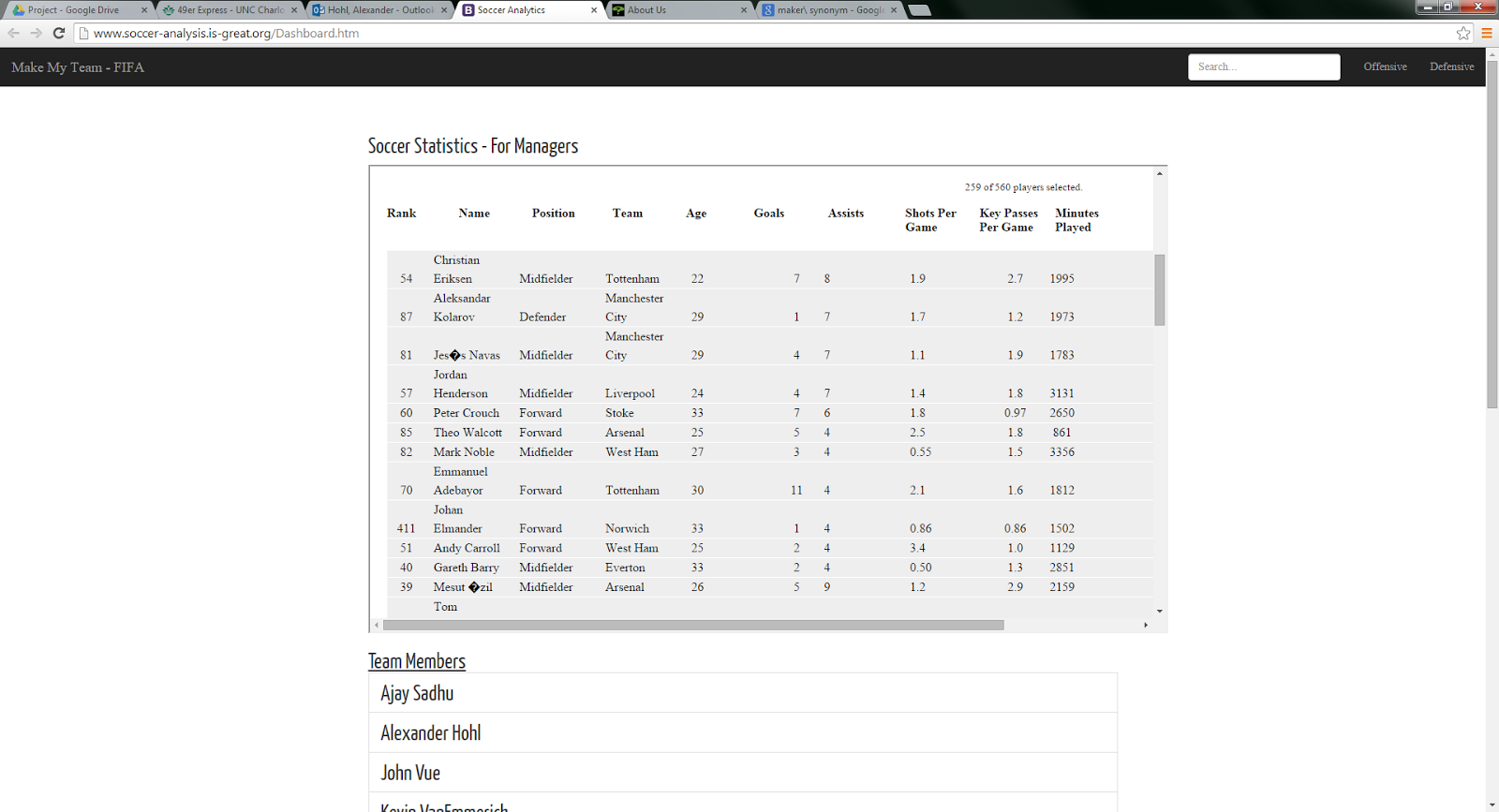
Soccer Matches [7] is designed to offer to a general audience a view of many attributes associated to the matches occurred in a soccer championship. It is a very interesting and good way of visually displaying data to a general audience because it uses colors and adjacency matrix to represent the overall data. Implementing the Soccer Matches visual display in order to work with teams in a league and track their performances against other teams or comparing the attributes of players might be interesting avenues for further research that go beyond the scope of our project.

CollaStar [8] uses a star glyph and linear walls. While an implementation of CollaStar in our project without modification does not seem feasible, we acknowledge the possibility of modifying the concept to, for instance, a star glyph for comparing the different statistics of the players or teams. We think that we could use something like this for showing the comparison of the overall stat or score of a player to other players.

Emphasizing the importance of visualization of sport data, a team of scientists at the Georgia Institute of Technology [9] present SportViz, which visualizes the Baseball data for a season in determining the performance of the team under certain circumstances by using filters and sorting. Using Principal Component Analysis, a methodology for grouping correlated variables into a set of values of linearly uncorrelated variables, the performance of players is measured on the Indian IPL cricket [10]. This paper shows how statistical analysis can be successfully leveraged for sports analysis and creates meaningful results effectively and efficiently.

Providing basic steps of today’s sports analysts routine, SnapShot [11] is an application which was developed to provide new customized visualization for Ice Hockey. SnapShot was well received among the community of Hockey analysts, especially because of its visualization capabilities that might have to potential to generate an interest among people who are traditionally resistant of statistics in Hockey analysis. Soccer Scoop [12] is an application for the manager to look at each individual player’s attributes before offering a contract. In this application, visualizations, glyphs, modified star plots, details on demand, color, and gestalt principles are used.

**Fig. 2:** Project WhoScored Corssfilter implementation for visual interactive query of English Premier League soccer statistics.



**Fig. 3:** Table output of query result.

In an example that is less domain-specific, [13] present a visual interface to display multiple queries of a document collection. They use a measure to rank similarity of each document to the query, and visualize the measure as distance between a query symbol and a document symbol along a line. Since there are multiple queries and documents, they place the query symbols in the center of multiple concentric circles, on which the document symbols are placed. The Circles are divided into sectors, in order to avoid cluttering while displaying all queries.

Since our project group will work with querying soccer data, I thought that this paper could give us hints about how to visualize them.

There is promising and very advanced work on the use of colors and as in our study, we make use of them for graphs and maps, guidance on the choice of color scheme is most welcome. ColorBrewer.org is an online tool for selecting color schemes for maps [14]. A good color scheme needs to be attractive, support the message of the map and match the nature of the data. The system suggests color schemes based on the following, user-specified parameters: number of data classes, kind of color scheme (sequential, diverging, or qualitative) and display environment (CRT, laptop, print, LCD projector). It also displays the chosen scheme in an example thematic map of U.S. counties, allowing the user to ‘test-drive’ the color scheme. Finally, ColorBrewer offers color schemes that “colorblind-safe”, which means that the user has the option of choosing schemes that avoid colors indistinguishable by people who suffer from the most common forms of color-blindness.

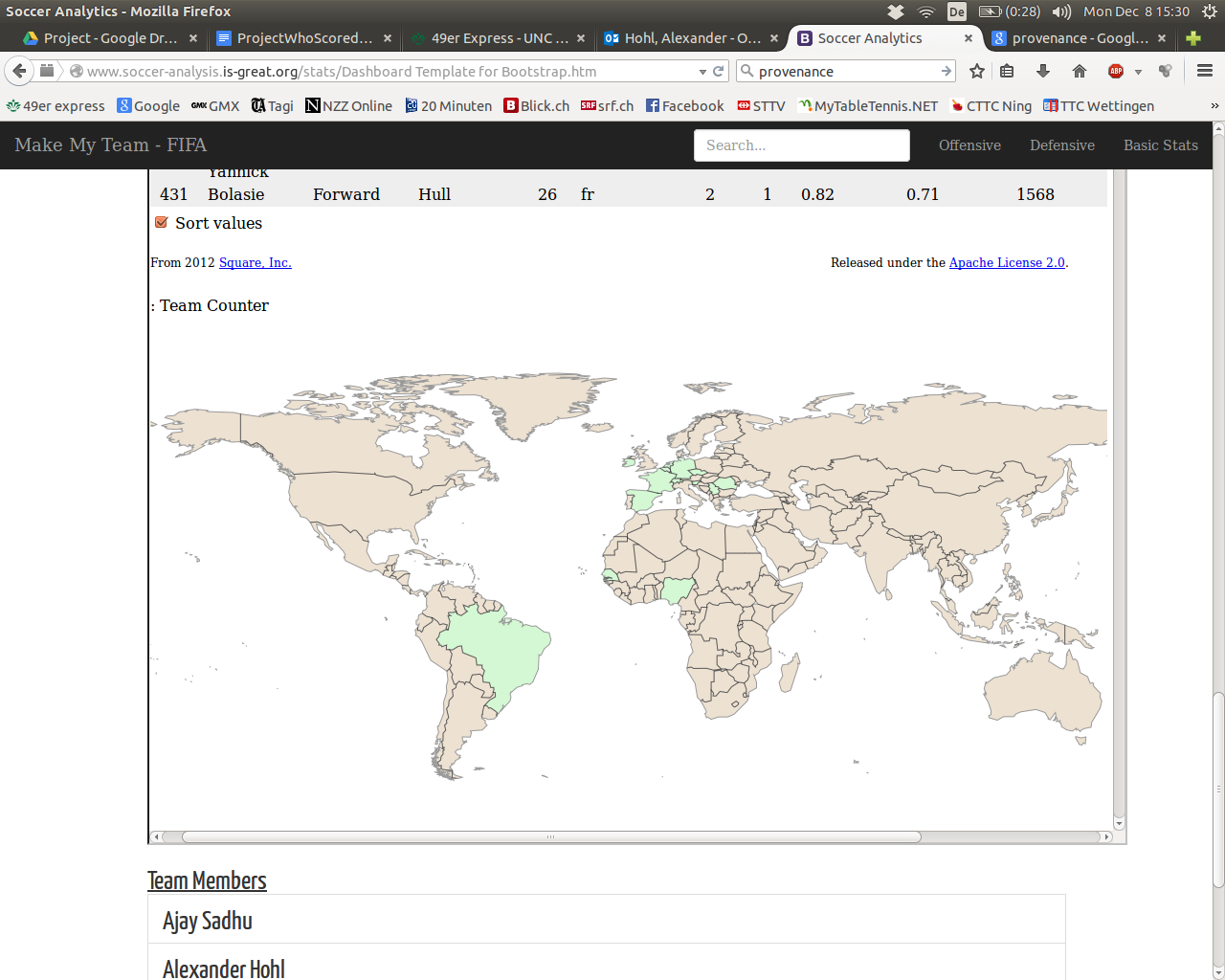
It is paramount to evaluate the visualization after successfully deploying an application, which is a step that is often missed in the process of designing new tools. Therefore, we find it important to keep in mind the guidelines found in the literature, e.g. [15], where the authors describe a research method called Multi-dimensional In-depth Long-term Case studies (MILCs) for assessing the creative activities in which users of information visualization systems engage in. They propose doing that by documenting 1) usage (interviews, observations, logging) and 2) the success of expert users in achieving their goals. The paper offers guidelines for conducting such studies for information visualization and even though we will not conduct a MILC during our project, we believe that the guidelines for assessing visualization systems are useful to us.

# System Design

For WhoScored, we are evaluating over 500 soccer players in the English Premier League. We used the following key player attributes: offensive (goals scored, assists, shots per game, key passes per game, minutes played) and defensive (tackles per game, interceptions per game, fouls per game, clearances per game). We downloaded the dataset from the website Whoscored.com. Even though the site offers additional datasets, we decided as a group to only use the key dataset to use for our project WhoScored.

We used Python to scrape all records of the 560 soccer players in the 2014/15 season into JSON file format, which we subsequently converted into CSV format for further use. In a next step, We implemented Crossfilter, a javascript library built on D3, developed by square inc. for data exploration within web-browsers, created in order to quickly filter datasets by multiple dimensions in a visually pleasing, interactive way. Using Crossfilter functionality, we created multiple histogram visualizations, each plotting the key player attribute dimensions mentioned above on the x-axis, against their frequencies on the y-axis (Fig. 2). These visualizations allow for interaction by offering the the capability to move sliders that are imposed on the histograms, letting the user specify the upper and lower boundaries or a range of values, which can be moved across the data range. That way, the multidimensional WhoScored dataset can be interactively queried within a visual interface. For instance, if the user wanted to know the players who scored five goals or more in this season, he would drag the buttons on the corresponding histogram visualization to highlight that specific bar and the chart below will filter itself to show only the players who scores five goals on the season.

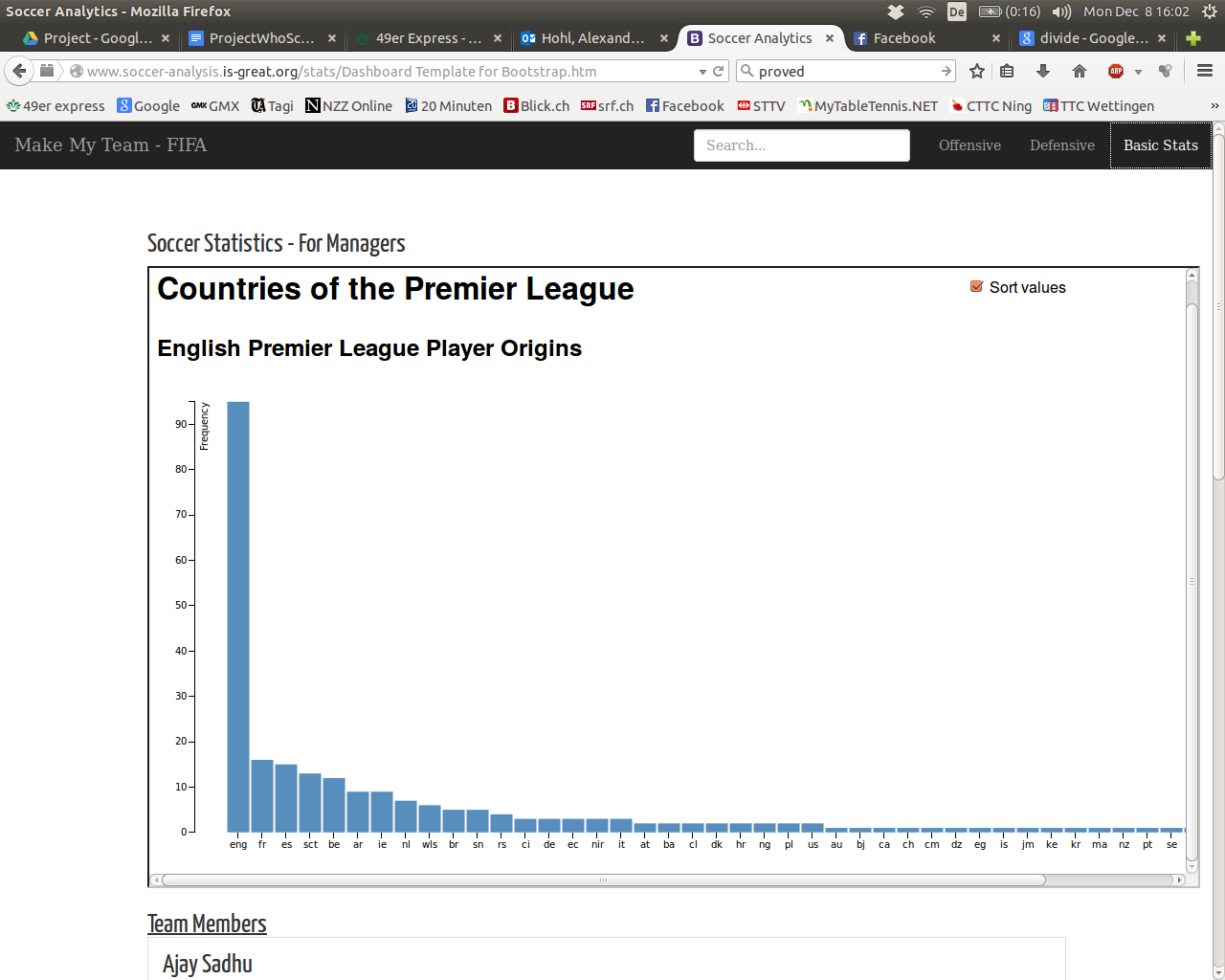
The system returns a count of selected players according to the user-defined query to the lower right of the histograms. In addition, the user can see a table of the resulting player records of the WhoScored dataset, containing the attributes rank, name, position, team, age in addition to the key player attributes mentioned above (Fig. 3). The rank attribute is computed by WhoScored.com and the creators of the site refuse to deliver a transparent explanation of how it is calculated. We restricted the maximum number of records shown in the table to 40, as a higher number would lead to endless scrolling through the list.

In addition, we display a map directly below the table that shows player provenance (home country) of the subset of players resulting from the query (Fig.4). The corresponding countries are highlighted by a visually pleasing green color and the map is automatically updated if the query changes.

**Fig. 4:** Map of player provenance.

In addition, we created a frequency distribution chart of player provenance (Fig. 5), breaking down the number of players by their nationalities. This visualization includes the functionality to toggle between sorting the bars (nationalities) by their height (number of players from the corresponding country) and sorting them alphabetically.

# Evaluation

We tested the application by formulating many different queries and it proved to work with all queries we formulated. In addition, we used mobile devices, such as a Samsung Galaxy S3 smartphone to see whether the interface performed equally well in a non-desktop environment. We did not identify any anomalies of the interface in this experiment, the application achieved an equally pleasing user experience compared to a desktop computer.

**Fig. 5:** Frequency distribution of player nationalities.

We presented our application to an independent domain expert, namely a Charlotte, NC-based soccer coach, who stated that the application is “very intuitive” and “easy to use”. He also expressed that “this system satisfies a need that has has been unmet to a satisfactory degree for quite some time” - namely the capability to interactively and visually formulate queries of soccer data that is available to everybody.

# Conclusion

Here, we presented Project WhoScored, in which we created an interactive visualization for the analysis of soccer data. We used data of the 2014/15 season of the British Premier League, which is provided by the popular website WhoScored.com, that hosts an array of soccer statistics from leagues all over the world.

Implementing functionality provided by D3, MongoDB and Crossfilter, we created an application capable of displaying and interactively querying multiple player statistics, such as goals score, assists, shots per game, and key passes per game. The visualization allows the user to identify the best scorers, assist givers, shooters, and passers in an interactive and visually pleasing way. The table output provides interesting additional information, such as club affiliation, age and rank.

Our efforts to evaluate the application yielded positive user feedback and the confirmation of cross-platform capability

Acknowledgments

The authors wish to thank Wayne Rooney, Sir Alex Ferguson, Sepp Blatter and Ronaldo Luís Nazário de Lima. This work was supported in part by a grant from the FIFA, the IOC and the NRA.

References

1. [1] http://www.fifa.com/index.htm
2. [2] http://www.historyofsoccer.info/
3. [3] http://www.soccerstats.com/
4. [4] http://grantland.com/the-triangle/the-adolescence-of-soccer-stats/
5. [5] Perin, C., R. Vuillemot, et al. (2013). "SoccerStories: A Kick-off for Visual Soccer Analysis." Visualization and Computer Graphics, IEEE Transactions on 19(12): 2506-2515.
6. [6] Perin, C., Vuillemot, R., & Fekete, J. D. (2014, April). A table!: improving temporal navigation in soccer ranking tables. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 887-896). ACME. Catmull. A tutorial on compensation tables. In Computer Graphics, volume 13, pages 1–7. ACM SIGGRAPH, 1979.
7. [7] Cava, R., & Freitas, C. D. S. Glyphs in Matrix Representation of Graphs for Displaying Soccer Games Results.
8. [8] Perin, C. (2013, November). Visualizing Multidimensional and Temporal Data at the Good Scale by Designing and Refining. In 25ème conférence francophone sur l'Interaction Homme-Machine, IHM'13.
9. [9] Cox, A., & Stasko, J. (2006). Sportsvis: Discovering meaning in sports statistics through information visualization. In Compendium of Symposium on Information Visualization (pp. 114-115).
10. [10] Manage, Ananda BW, and Stephen M. Scariano. "An Introductory Application of Principal Components to Cricket Data." Journal of Statistics Education 21.3 (2013).
11. [11] Pileggi, H., Stolper, C. D., Boyle, J. M., & Stasko, J. T. (2012). Snapshot: Visualization to propel ice hockey analytics. Visualization and Computer Graphics, IEEE Transactions on, 18(12), 2819-2828.
12. [12] Rusu, Adrian, et al. "Dynamic visualizations for soccer statistical analysis." Information Visualisation (IV), 2010 14th International Conference. IEEE, 2010.
13. [13] Havre, S., E. Hetzler, et al. (2001). Interactive visualization of multiple query results. Information Visualization, IEEE Symposium on, IEEE Computer Society.
14. [14] Harrower, M. and C. A. Brewer (2003). "Colorbrewer. org: an online tool for selecting colour schemes for maps." The Cartographic Journal 40(1): 27-37.
15. [15] Shneiderman, B. and C. Plaisant (2006). Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization, ACM